

IBM SPSS Modeler

도움말 2

(소장용)

원문:

<https://www.ibm.com/docs/ko/spss-modeler/SaaS?topic=started-how-use-spss-modeler>

2022. 05.

국제개발연구소

<목차>

4) 그래프 노드	28
(1) 공통 그래프 노드 기능	28
① 모양, 오버레이, 패널 및 애니메이션	30
② 출력 탭 사용	31
③ 주석 탭 사용	32
④ 3 차원 그래프	32
(2) 그래프보드 노드	33
① 그래프보드 기본 탭	34
가. 필드(변수) 유형	36
② 그래프보드 세부사항 탭	38
가. 맵 시각화를 위한 맵 파일 선택	38
③ 사용 가능한 내장 그래프보드 시각화 유형	40
④ 맵 시각화 작성	46
⑤ 그래프보드 예제	46
가. 예제: 요약 통계가 포함된 막대형 차트	47
나. 예제: 요약 통계가 포함된 누적 막대형 차트	48
다. 예제: 패널링된 히스토그램	49
라. 예제: 패널링된 점도표	50
마. 예제: 상자도표	51
바. 예제: 원형 차트	52
사. 예제: 히트 맵	53
아. 예제: 산점도 행렬(SPLOM)	54
자. 예제: 합계의 코로플레스(색상 맵)	55
차. 예제: 맵 위의 막대형 차트	56
⑥ 그래프보드 모양 탭	57

⑦ 템플리트, 스타일시트 및 맵 위치 설정	59
가. IBM SPSS Collaboration and Deployment Services Repository 를 템플리트, 스타일시트 및 맵 파일 위치로 사용	59
⑧ 템플리트, 스타일시트 및 맵 파일 관리	60
(3) 맵 형태 파일 변환 및 배포	61
① 맵의 핵심 개념	62
② 맵 변환 유틸리티 사용	63
가. 1 단계 - 대상 및 소스 파일 선택	63
나. 2 단계 - 맵 키 선택	63
다. 3 단계 - 맵 편집	64
ㄱ. 맵 평활화	65
ㄴ. 지형 레이블 편집	65
• 외부 데이터 소스와 비교 대화 상자	66
ㄷ. 지형 합치기	67
• 합친 지형의 이름 지정 대화 상자	67
ㄹ. 지형 이동	68
ㅁ. 지형 삭제	68
ㅂ. 개별 요소 삭제	68
ㅅ. 투영법 설정	69
라. 4 단계 - 완료	69
③ 맵 파일 배포	70
(4) Plot 노드	70
① 구성 노드 탭	72
② 도표 옵션 탭	74
③ 도표 모양 탭	75
④ 도표 그래프 사용	76
(5) 다중 도표 노드	77

① 다중 도표 도표 탭	77
② 다중 도표 탭	79
③ 다중 도표 그래프 사용	80
(6) 시간 구성 노드	80
① 시간 구성 탭	81
② 시간 구성 모양 탭	82
③ 시간 도표 그래프 사용	83
(7) 분포 노드	84
① 분포 도표 탭	84
② 분포 모양 탭	85
③ 분포 노드 사용	86
(8) 히스토그램 노드	88
① 히스토그램 도표 탭	88
② 히스토그램 옵션 탭	88
③ 히스토그램 모양 탭	89
④ 히스토그램 사용	90
(9) 요약도표 노드	90
① 컬렉션 도표 탭	91
② 컬렉션 옵션 탭	91
③ 컬렉션 모양 탭	91
④ 컬렉션 그래프 사용	92
(10) 웹 노드	94
① 웹 구성 탭	95
② 웹 옵션 탭	96
③ 웹 모양 탭	97
④ 웹 그래프 사용	98
가. 웹 임계값 조정	101

나. 웹 요약 작성	103
(11) 평가 노드	103
① 평가 도표 탭	107
② 평가 옵션 탭	109
③ 평가 모양 탭	110
④ 모형 평가의 결과 읽기	110
⑤ 평가 차트 사용	112
(12) 맵 시각화 노드	113
① 맵 시각화 도표 탭	113
가. 맵 레이어 변경	114
② 맵 시각화 모양 탭	118
(13) t-SNE 노드	118
① t-SNE 노드 고급 옵션	119
② t-SNE 노드 출력 옵션	121
③ t-SNE 데이터 액세스 및 도표화	121
④ t-SNE 모델 너깃	123
(14) E-Plot(베타) 노드	124
① E-Plot(베타) 노드 도표 탭	124
② E-Plot(베타) 노드 옵션 탭	124
③ E-Plot(베타) 외형 탭	125
④ E-Plot 그래프 사용	125
(15) 그래프 출력에 대한 작업	127
(16) 그래프 탐색	128
① 밴드 사용	129
② 영역 사용	132
③ 표시된 요소 사용	134
④ 그래프에서 노드 생성	135

(17) 시각화 편집	138
① 시각화 편집 일반 규칙	139
② 텍스트 편집 및 형식화	141
③ 색상, 패턴, 대시 및 투명도 변경	141
④ 점 요소의 형태 및 가로 세로 비율 회전과 변경	142
⑤ 그래픽 요소의 크기 변경	143
⑥ 여백 및 패딩 지정	143
⑦ 숫자 형식 지정	144
⑧ 축 및 척도 설정 변경	145
⑨ 범주 편집	147
⑩ 패널 방향 변경	148
⑪ 좌표계 변환	149
⑫ 통계 및 그래픽 요소 변경	150
⑬ 범례 위치 변경	151
⑭ 시각화 및 시각화 데이터 복사	152
⑮ 그래프보드 편집기 키보드 단축키	152
⑯ 제목 및 꼬리말 추가	152
⑰ 그래프 스타일시트 사용	153
가. 스타일시트 적용	154
⑱ 그래프 인쇄, 저장, 복사 및 내보내기	155
가. 머리글 및 바닥글 환경 설정 설정	158
5) 출력 노드	159
(1) 출력 노드 개요	159
(2) 출력 관리	160
(3) 출력 보기	161
① 웹에 출판	162
가. 웹에 출력 출판	162

나. 웹을 통해 출판된 출력 보기	163
② HTML 브라우저에서 출력 보기	163
③ 출력 내보내기	164
④ 셀 및 열 선택	164
(4) 테이블 노드	165
① 테이블 노드 설정 탭	165
② 출력 노드 출력 탭	165
③ 테이블 브라우저	167
(5) 교차표 노드	168
① 교차표 노드 설정 탭	168
② 교차표 노드 모양 탭	169
③ 교차표 노드 출력 브라우저	170
(6) 분석 노드	171
① 분석 노드 분석 탭	172
② 분석 출력 브라우저	173
(7) 데이터 검토 노드	175
① 데이터 검토 노드 설정 탭	176
② 데이터 검토 품질 탭	177
③ 데이터 검토 출력 브라우저	177
가. 그래프 보기 및 생성	178
나. 통계 표시	179
다. 데이터 검토 브라우저 품질 탭	180
ㄱ. 결측값 대치	180
ㄴ. 이상값 및 극단값 처리	182
ㄷ. 결측 데이터로 필드 필터링	183
ㄹ. 결측 데이터가 있는 레코드 선택	183
ㅁ. 데이터 준비를 위해 기타 노드 생성	184

(8) 변환 노드	184
① 변환 노드 옵션 탭	185
② 변환 노드 출력 탭	186
③ 변환 노드 출력 뷰어	186
가. 변환을 위한 노드 생성	186
ㄱ. 그래프 생성	188
• 기타 조작	188
(9) 통계량 노드	188
① 통계량 노드 설정 탭	188
가. 상관관계 설정	189
② 통계량 출력 브라우저	190
가. 통계량에서 필터 노드 생성	191
(10) 평균 노드	191
① 독립 그룹에 대한 평균 비교	192
② 대응 필드 간 평균 비교	192
③ 평균 노드 옵션	192
④ 평균 노드 출력 브라우저	193
가. 필드 내 평균 출력 비교 그룹	193
나. 필드의 평균 출력 비교 쌍	194
(11) 보고서 노드	195
① 보고서 노드 템플릿 탭	195
② 보고서 노드 출력 브라우저	197
(12) 전역값 설정 노드	197
① 전역값 설정 노드 설정 탭	197
(13) 시뮬레이션 적합 노드	198
① 분포 적합	198
② 시뮬레이션 적합 노드 설정 탭	200

(14) 시뮬레이션 평가 노드	201
① 시뮬레이션 평가 노드 설정 탭	201
② 시뮬레이션 평가 노드 출력	204
가. 탐색 패널	204
나. 차트 출력	205
다. 차트 옵션	207
(15) 확장 출력 노드	208
① 확장 출력 노드 - 구문 탭	209
② 확장 출력 노드 - 콘솔 출력 탭	210
③ 확장 출력 노드 - 출력 탭	210
④ 확장 출력 브라우저	211
가. 확장 출력 브라우저 - 텍스트 출력 탭	211
나. 확장 출력 브라우저 - 그래프 출력 탭	211
(16) KDE 노드	211
① KDE 모델링 노드 및 KDE 시뮬레이션 노드 필드	212
② KDE 노드 작성 옵션	212
③ KDE 모델링 노드 및 KDE 시뮬레이션 노드 모델 옵션	214
6) 내보내기 노드	214
(1) 내보내기 노드의 개요	214
(2) 데이터베이스 내보내기 노드	216
① 데이터베이스 노드 내보내기 탭	216
② 데이터베이스 내보내기 병합 옵션	217
③ 데이터베이스 내보내기 스키마 옵션	219
가. SQL Server 에 대한 옵션	221
나. Oracle 에 대한 옵션	221
④ 데이터베이스 내보내기 인덱스 옵션	222
⑤ 데이터베이스 내보내기 고급 옵션	224

⑥ 벌크 로더 프로그래밍	226
가. IBM Db2 데이터베이스에 데이터 벌크 로드	227
나. IBM Netezza 데이터베이스에 데이터 벌크 로드	228
다. Oracle 데이터베이스에 데이터 벌크 로드	229
라. SQL Server 데이터베이스에 데이터 벌크 로드	230
마. Teradata 데이터베이스에 데이터 벌크 로드	231
바. 벌크 로더 프로그램 개발	232
사. 벌크 로더 프로그램 테스트	234
(3) 플랫폼 파일 내보내기 노드	235
① 플랫폼 파일 내보내기 탭	235
(4) 통계량 내보내기 노드	236
① 통계량 내보내기 노드 - 내보내기 탭	237
② IBM SPSS Statistics 에 대한 필드 이름 변경 또는 필터링	238
(5) Data Collection 내보내기 노드	238
(6) IBM Cognos 내보내기 노드	239
① Cognos 연결	240
② ODBC 연결	240
(7) IBM Cognos TM1 내보내기 노드	242
① 데이터를 내보낼 IBM Cognos TM1 큐브에 연결	243
② 내보낼 IBM Cognos TM1 데이터 맵핑	244
(8) SAS 내보내기 노드	245
① SAS 내보내기 노드 내보내기 탭	245
(9) Excel 내보내기 노드	245
① Excel 노드 내보내기 탭	246
(10) 확장 내보내기 노드	246
① 확장 내보내기 노드 - 명령문 탭	247
② 확장 내보내기 노드 - 콘솔 출력 탭	248

(11) XML 내보내기 노드	248
① XML 데이터 쓰기	249
② XML 레코드 매핑 옵션	249
③ XML 필드 매핑 옵션	250
④ XML 매핑 미리보기	250
(12) JSON 내보내기 노드	250
(13) 공통 내보내기 노드 탭	251
① 스트림 출판	252
7) 슈퍼노드	253
(1) 슈퍼노드 개요	253
(2) 슈퍼노드 유형	254
① 소스 슈퍼노드	254
② 프로세스 슈퍼노드	254
③ 터미널 슈퍼노드	255
(3) 슈퍼 노드 작성	255
① 슈퍼노드 중첩	256
(4) 슈퍼노드 잠금	256
① 슈퍼노드 잠금 및 잠금 해제	257
② 잠긴 슈퍼노드 편집	258
(5) 슈퍼노드 편집	258
① 슈퍼노드 유형 수정	258
② 슈퍼노드 주석(Annotation) 작성 및 이름 바꾸기	259
③ 슈퍼 노드 모수	260
가. 슈퍼노드 매개변수 정의	260
나. 슈퍼노드 매개변수의 값 설정	261
다. 슈퍼노드 모수를 사용하여 노드 특성 액세스	261
④ 슈퍼노드 및 캐싱	262

⑤ 슈퍼노드 및 스크립팅	263
(6) 슈퍼노드 저장 및 로드	263
3. 모델링 노드	264
1) 모델링 개요	264
(1) 모델링 노드의 개요	264
(2) 분할 모델 작성	270
① 분할 및 파티셔닝	271
② 분할 모델을 지원하는 모델링 노드	271
③ 분할 영향을 받는 기능	272
(3) 모델링 노드 필드 옵션	273
① 빈도 및 가중 필드 사용	275
(4) 모델링 노드 분석 옵션	277
① 성향 스코어	279
(5) 오분류 비용	280
(6) 모델 너깃	281
① 모델 링크	282
가. 모델 링크 정의 및 제거	282
나. 모델 링크 복사 및 붙여넣기	283
다. 모델 링크 및 슈퍼 노드	284
② 모델 교체	284
③ 모델 팔레트	285
④ 모델 너깃 찾아보기	286
⑤ 모델 너깃 요약/정보	288
⑥ 예측변수 중요도	288
가. 중요도에 기반하여 변수 필터링	290
⑦ 앙상블 뷰어	290
가. 앙상블 모델	290

ㄱ. 모델 요약(양상블 뷰어)	291
ㄴ. 예측변수 중요도(양상블 뷰어)	291
ㄷ. 예측자 빈도(양상블 뷰어)	292
ㄹ. 구성요소 모델 정확도(양상블 뷰어)	292
ㅁ. 구성요소 모델 세부사항 (양상블 뷰어)	292
ㅂ. 자동 데이터 준비 (양상블 뷰어)	293
⑧ 분할 모델의 모델 너깃	293
가. 분할 모델 뷰어	293
⑨ 스트림에서 모델 너깃 사용	294
⑩ 모델링 노드 재생성	295
⑪ 모델을 PMML 로 가져오기 및 내보내기	296
가. PMML 을 지원하는 모델 유형	297
⑫ 스코어링 어댑터에 대한 모델 게시	298
⑬ 세분화되지 않은 모델	299
(7) 생성된 통계 모형 고급 출력	299
(8) 군집 모델 모델 탭	299
(9) 규칙 세트/의사결정 트리 모형 탭 생성	300
2) 선별 모델	300
(1) 필드 및 레코드 선별	300
(2) 필드선택 노드	301
① 필드선택 모델 설정	301
② 필드선택 옵션	302
(3) 필드선택 모델 너깃	303
① 필드선택 모델 결과	304
② 중요도에 따라 필드 선택	304
③ 필드선택 모델에서 필터 생성	305
(4) 이상 항목 발견 노드	305

① 이상 항목 발견 모델 옵션	306
② 이상 항목 발견 고급 옵션	307
(5) 이상 항목 발견 모델 너깃	308
① 이상 항목 발견 모델 세부사항	309
② 이상 항목 발견 모델 요약	309
③ 이상 항목 발견 모델 설정	310
3) 자동화된 모델링 노드	310
(1) 자동화된 모델링 노드 알고리즘 설정	312
(2) 자동화된 모델링 노드 중지 규칙	312
(3) 실행 피드백	313
(4) 자동 분류자 노드	313
① 자동 분류자 노드 모델 옵션	314
② 자동 분류자 노드 고급 옵션	316
③ 자동 분류자 노드 삭제 옵션	319
④ 자동 분류자 노드 설정 옵션	320
(5) 자동 숫자 노드	320
① 자동 숫자 노드 모델 옵션	322
② 자동 숫자 노드 고급 옵션	323
③ 자동 숫자 노드 설정 옵션	326
(6) 자동 군집 노드	326
① 자동 군집 노드 모델 옵션	327
② 자동 군집 노드 고급 옵션	328
③ 자동 군집 노드 삭제 옵션	329
(7) 자동화된 모델 너깃	330
① 노드 및 모델 생성	331
② 평가 차트 생성	332
③ 평가 그래프	332

④ 자동화된 모델 너깃 요약	333
⑤ 연속 기계 학습	333
4) 의사결정 트리	342
(1) 의사결정 트리 모형	342
(2) 대화형 트리 작성기	344
① 트리 성장 및 가지치기	345
② 사용자 정의 분할 정의	346
가. 예측자 세부사항 보기	347
③ 분할 세부사항 및 대응	347
④ 트리 보기 사용자 정의	348
⑤ 이득	348
가. 분류 이익	349
나. 분류 이익 및 ROI	350
다. 회귀분석 이익	351
라. Gains 차트	351
마. 이익 기반 선택	352
⑥ 위험	353
⑦ 트리 모델 및 결과 저장	353
가. 트리 작성기에서 모델 생성	354
나. 트리 성장 지시문	354
다. 트리 지시문 업데이트	356
라. 모델, 이익, 위험 정보 내보내기	356
⑧ 필터 및 선택 노드 생성	357
⑨ 의사결정 트리에서 규칙 세트 생성	357
(3) 직접 트리 모델 작성	358
(4) 의사결정 트리 노드	359
① C&R 트리 노드	360

② CHAID 노드	361
③ QUEST 노드	362
④ 의사결정 트리 노드 필드 옵션	363
⑤ 의사결정 트리 노드 작성 옵션	363
가. 의사결정 트리 노드 - 목적	363
나. 의사결정 트리 노드 - 기본	365
다. 의사결정 트리 노드 - 중지 규칙	366
라. 의사결정 트리 노드 - 앙상블	366
마. C&R 트리 및 QUEST 노드 - 비용 및 사전	367
바. CHAID 노드 - 비용	368
사. C&R 트리 노드 - 고급	368
아. QUEST 노드 - 고급	369
자. CHAID 노드 - 고급	370
⑥ 의사결정 트리 노드 모델 옵션	371
(5) C5.0 노드	372
① C5.0 노드 모델 옵션	373
(6) Tree-AS 노드	375
① Tree-AS 노드 필드 옵션	376
② Tree-AS 노드 작성 옵션	376
가. Tree-AS 노드 - 기본	376
나. Tree-AS 노드 - 성장	377
다. Tree-AS 노드 - 중지 규칙	378
라. Tree-AS 노드 - 비용	378
③ Tree-AS 노드 모델 옵션	379
④ Tree-AS 모델 너깃	379
가. Tree-AS 모델 너깃 출력	379
나. Tree-AS 모델 너깃 설정	381

(7) 랜덤 트리 노드	382
① 랜덤 트리 노드 필드 옵션	383
② 랜덤 트리 노드 작성 옵션	383
가. 랜덤 트리 노드 - 기본	384
나. 랜덤 트리 노드 - 비용	385
다. 랜덤 트리 노드 - 고급	385
③ 랜덤 트리 노드 모델 옵션	386
④ 랜덤 트리 모델 너깃	386
가. 랜덤 트리 모델 너깃 출력	386
나. 랜덤 트리 모델 너깃 설정	388
(8) C&R 트리, CHAID, QUEST, C5.0 의사결정 트리 모형 너깃	389
① 단일 트리 모델 너깃	390
가. 의사결정 트리 모형 규칙	391
나. 의사결정 트리 모형 뷰어	393
다. 의사결정 트리/규칙 세트 모델 너깃 설정	394
라. 부스팅 C5.0 모델	395
마. 그래프 생성	396
② 부스팅, 배깅 및 매우 큰 데이터 세트의 모델 너깃	396
(9) C&R 트리, CHAID, QUEST, C5.0, Apriori 규칙 세트 모델 너깃	397
① 규칙 세트 모델 탭	399
(10) AnswerTree 3.0 에서 프로젝트 가져오기	399
5) 베이지안 신경망 모델	400
(1) 베이지안 네트워크 노드	400
① 베이지안 네트워크 노드 모델 옵션	402
② 베이지안 네트워크 노드 고급 옵션	403
(2) 베이지안 신경망 모델 너깃	405
① 베이지안 신경망 모델 설정	406

② 베이지안 신경망 모델 요약	406
6) 신경망	407
(1) 신경망 모델	408
(2) 레거시 스트림이 있는 신경망 사용	409
(3) 목적 (신경망)	409
(4) 기본 (신경망)	411
(5) 중지 규칙(신경망)	412
(6) 앙상블 (신경망)	413
(7) 고급 (신경망)	414
(8) 모델 옵션(신경망)	415
(9) 모델 요약 (신경망)	416
(10) 예측변수 중요도 (신경망)	417
(11) 관측값 별 예측값 (신경망)	418
(12) 분류 (신경망)	419
(13) 네트워크 (신경망)	420
(14) 설정 (신경망)	421
7) 의사결정 목록	422
(1) 의사결정 목록 모델 옵션	424
(2) 의사결정 목록 노드 고급 옵션	425
(3) 의사결정 목록 모델 너깃	426
① 의사결정 목록 모델 너깃 설정	426
(4) 의사결정 목록 뷰어	427
① 작업 모델 분할창	427
② 대안 탭	429
③ 스냅샷 탭	430
④ 의사결정 목록 뷰어에 대한 작업	431
가. 마이닝 작업	431

ㄱ. 마이닝 작업 실행	431
ㄴ. 마이닝 작업 작성 및 편집	432
• 새 설정	434
• 고급 매개변수 편집	434
• 사용 가능 필드 사용자 정의	434
ㄷ. 데이터 선택 구성	435
• 선택 조건 지정	435
♦ 값 삽입	436
나. 실행 취소 및 다시 실행 조치	436
다. 세그먼트 규칙	436
ㄱ. 세그먼트 삽입	437
ㄴ. 세그먼트 규칙 편집	437
• 조건 삽입/편집	438
• 세그먼트 규칙 조건 삭제	438
ㄷ. 세그먼트 복사	439
ㄹ. 대체 모델	439
라. 모델 사용자 정의	440
ㄱ. 세그먼트 우선 순위 지정	440
ㄴ. 세그먼트 삭제	440
ㄷ. 세그먼트 제외	440
ㄹ. 목표 값 변경	441
마. 새 모델 생성	441
바. 모델 평가	442
ㄱ. 모델 측도 구성	442
• 측도 새로 고침	443
ㄴ. Excel로 평가	443
• 사용자 정의 측도에 대한 입력 선택	444

• MS Excel 통합 설정	444
• 모델 속도 변경	445
사. 모델 시각화	447
ㄱ. Gains 차트	447
• 차트 옵션	448
8) 통계 모델	448
(1) 선형 노드	449
① 선형 모델	450
가. 목적 (선형 모델)	450
나. 기본 (선형 모델)	451
다. 모델 선택(선형 모델)	452
라. 앙상블 (선형 모델)	453
마. 고급 (선형 모델)	454
바. 모델 옵션(선형 모델)	454
사. 모델 요약 (선형 모델)	454
아. 자동 데이터 준비 (선형 모델)	454
자. 예측자 중요도 (선형 모델)	455
차. 관측값 별 예측값 (선형 모델)	455
카. 잔차 (선형 모델)	455
타. 이상치 (선형 모델)	456
파. 효과 (선형 모델)	456
하. 계수 (선형 모델)	457
거. 추정 평균(선형 모델)	458
너. 모델 작성 요약 (선형 모델)	458
더. 설정 (선형 모델)	458
(2) Linear-AS 노드	459
① Linear-AS 모델	459

가. 기본(linear-AS 모델)	460
나. 모델 선택(linear-AS 모델)	460
다. 모델 옵션(Linear-AS 모델)	461
라. 대화형 출력(linear-AS 모델)	462
마. 설정(linear-AS 모델)	463
(3) 로지스틱 노드	463
① 로지스틱 노드 모델 옵션	464
② 로지스틱 회귀 모형에 항 추가	468
③ 로지스틱 노드 고급 옵션	469
④ 로지스틱 회귀분석 수렴 옵션	470
⑤ 로지스틱 회귀분석 고급 출력	470
⑥ 로지스틱 회귀분석 단계별 옵션	471
(4) 로지스틱 모델 너깃	472
① 로지스틱 너깃 모델 세부사항	473
② 로지스틱 모델 너깃 요약	474
③ 로지스틱 모델 너깃 설정	474
④ 로지스틱 모델 너깃 고급 출력	476
(5) PCA/요인 노드	477
① PCA/요인 노드 모델 옵션	477
② PCA/요인 노드 고급 옵션	478
③ PCA/요인 노드 회전 옵션	479
(6) PCA/요인 모델 너깃	480
① PCA/요인 모델 너깃 방정식	480
② PCA/요인 모델 너깃 요약	480
③ PCA/요인 모델 너깃 고급 출력	480
(7) 판별 노드	481
① 판별 노드 모델 옵션	482

② 판별 노드 고급 옵션	482
③ 판별 노드 출력 옵션	483
④ 판별 노드 단계 옵션	484
⑤ 판별 모델 너깃	485
가. 판별 모델 너깃 고급 출력	485
나. 판별 모델 너깃 설정	486
다. 판별 모델 너깃 요약	486
(8) GenLin 노드	486
① GenLin 노드 필드 옵션	487
② GenLin 노드 모델 옵션	488
③ GenLin 노드 고급 옵션	489
④ 일반화 선형 모델 반복	491
⑤ 일반화 선형 모델 고급 출력	492
⑥ GenLin 모델 너깃	493
가. GenLin 모델 너깃 고급 출력	494
나. GenLin 모델 너깃 설정	494
다. GenLin 모델 너깃 요약	494
(9) 일반화 선형 혼합 모델	495
① GLMM 노드	495
가. 일반화 선형 혼합 모델	495
ㄱ. 목표 (일반화 선형 혼합 모델)	497
ㄴ. 고정 효과 (일반화 선형 혼합 모델)	499
• 사용자 정의 항 추가 (일반화 선형 혼합 모델)	500
ㄷ. 랜덤효과 (일반화 선형 혼합 모델)	501
• 랜덤 효과 블록 (일반화 선형 혼합 모델)	501
ㄹ. 가중치 및 변위 (일반화 선형 혼합 모델)	503
ㅁ. 일반 작성 옵션 (일반화 선형 혼합 모델)	503

ㄴ. 추정 (일반화 선형 혼합 모델)	504
ㄷ. 일반 (일반화 선형 혼합 모델)	505
ㄹ. 평균 추정 (일반화 선형 혼합 모델)	506
ㅈ. 모델 보기 (일반화 선형 혼합 모델)	507
• 모델 요약 (일반화 선형 혼합 모델)	507
• 데이터 구조 (일반화 선형 혼합 모델)	508
• 관측값 별 예측값 (일반화 선형 혼합 모델)	508
• 분류 (일반화 선형 혼합 모델)	508
• 고정 효과 (일반화 선형 혼합 모델)	509
• 고정 계수 (일반화 선형 혼합 모델)	509
• 랜덤효과 공분산 (일반화 선형 혼합 모델)	510
• 공분산 모수 (일반화 선형 혼합 모델)	510
• 평균 추정: 유의한 효과 (일반화 선형 혼합 모델)	511
• 평균 추정: 사용자 정의 효과 (일반화 선형 혼합 모델)	511
• 설정 (일반화 선형 혼합 모델)	512
(10) GLE 노드	513
① 목표(GLE 모델)	514
② 모델 효과(GLE 모델)	516
가. 사용자 정의 항 추가(GLE 모델)	517
③ 가중치 및 오프셋(GLE 모델)	518
④ 작성 옵션(GLE 모델)	518
⑤ 추정(GLE 모델)	519
⑥ 모델 선택(GLE 모델)	520
⑦ 모델 옵션(GLE 모델)	521
⑧ GLE 모델 너깃	522
가. GLE 모델 너깃 출력	522
나. GLE 모델 너깃 설정	523

(11) Cox 노드	523
① Cox 노드 필드 옵션	524
② Cox 노드 모델 옵션	524
가. Cox 회귀 모형에 항 추가	526
③ Cox 노드 고급 옵션	526
가. Cox 노드 수렴 기준	527
나. Cox 노드 고급 출력 옵션	527
다. Cox 노드 단계별 기준	528
④ Cox 노드 설정 옵션	528
⑤ Cox 모델 너깃	529
가. Cox 회귀분석 출력 설정	529
나. Cox 회귀분석 고급 출력	530
9) 군집 모델	530
(1) 코호넨 노드	531
① 코호넨 노드 모델 옵션	532
② 코호넨 노드 고급 옵션	534
(2) 코호넨 모델 너깃	534
① 코호넨 모델 요약	535
(3) K-평균 노드	535
① K-평균 노드 모델 옵션	536
② K-평균 노드 고급 옵션	536
(4) K-평균 모델 너깃	537
① K-평균 모델 요약	537
(5) 이단계 군집 노드	538
① TwoStep 군집 노드 모델 옵션	538
(6) TwoStep 군집 모델 너깃	540
① 이단계 모델 요약	540

(7) TwoStep-AS 군집 노드	540
① Twostep-AS 군집분석	540
가. 필드 탭 (Twostep-AS 군집)	541
나. 기본 (Twostep-AS 군집)	541
다. 기능 트리 기준(Twostep-AS 군집)	542
라. 표준화	544
마. 필드선택	544
바. 모델 출력	545
사. 모델 옵션	547
(8) TwoStep-AS 군집 모델 너깃	547
① TwoStep-AS 군집 모델 너깃 설정	547
(9) K-평균-AS 노드	548
① K-Means-AS 노드 필드	548
② K-평균-AS 노드 작성 옵션	548
(10) 군집 뷰어	550
① 군집 뷰어 - 모델 탭	551
가. 모델 요약 보기	551
나. 군집 보기	552
ㄱ. 군집 및 변수 전치	553
ㄴ. 변수 정렬	553
ㄷ. 군집 정렬	553
ㄹ. 셀 콘텐츠	553
다. 군집 예측자 중요도 보기	554
라. 군집 크기 보기	554
마. 셀 분포 보기	554
바. 군집 비교 보기	554
② 군집 뷰어 탐색	555

③ 군집 모델에서 그래프 생성	557
10) 연관 규칙	558
(1) 테이블 대 트랜잭션 데이터	560
(2) Apriori 노드	561
① Apriori 노드 모델 옵션	562
② Apriori 노드 고급 옵션	563
(3) CARMA 노드	564
① CARMA 노드 필드 옵션	565
② CARMA 노드 모델 옵션	566
③ CARMA 노드 고급 옵션	567
(4) 연관 규칙 모델 너깃	567
① 연관 규칙 모델 너깃 세부사항	568
가. 규칙의 필터 지정	571
나. 규칙의 그래프 생성	572
② 연관 규칙 모델 너깃 설정	572
③ 연관 규칙 모델 너깃 요약	574
④ 연관 모델 너깃에서 규칙 세트 생성	574
⑤ 필터링된 모델 생성	574
⑥ 연관 규칙 스코어링	575
⑦ 연관 모델 배포	577
(5) 시퀀스 노드	579
① 시퀀스 노드 필드 옵션	580
② 시퀀스 노드 모델 옵션	581
③ 시퀀스 노드 고급 옵션	581
④ 시퀀스 모델 너깃	583
가. 시퀀스 모델 너깃 세부사항	585
나. 시퀀스 모델 너깃 설정	587

다. 시퀀스 모델 너깃 요약	587
라. 시퀀스 모델 너깃에서 규칙 수퍼 노드 생성	588
(6) 연관 규칙 노드	589
① 연관 규칙 - 필드 옵션	589
② 연관 규칙 - 규칙 작성	590
③ 연관 규칙 - 변환	592
④ 연관 규칙 - 출력	592
⑤ 연관 규칙 - 모델 옵션	594
⑥ 연관성 규칙 모델 너깃	595
가. 연관 규칙 모델 너깃 세부사항	596
나. 연관 규칙 모델 너깃 설정	596

4) 그래프 노드

(1) 공통 그래프 노드 기능

데이터 마이닝 프로세스의 여러 단계에서는 그래프 및 차트를 사용하여 IBM® SPSS® Modeler로 가져온 데이터를 탐색합니다. 예를 들어, 도표 또는 분포 노드를 데이터 소스에 연결하여 데이터 유형 및 분포를 살펴볼 수 있습니다. 그런 다음 레코드 및 필드 조작을 수행하여 다운스트림 모델링 조작을 위해 데이터를 준비할 수 있습니다. 그래프의 또 다른 공통 사용법은 새로 파생된 필드 간 분포 및 관계를 확인하는 것입니다.

그래프 팔레트에는 다음과 같은 노드가 포함되어 있습니다.

-  그래프보드 노드는 하나의 단일 노드에 있는 여러 가지 유형의 많은 그래프를 제공합니다. 이 노드를 사용하여 탐색하려는 데이터 필드를 선택하고 선택된 데이터에 대해 사용 가능한 것 중에서 그래프를 선택할 수 있습니다. 이 노드는 필드 선택사항에 대해 작업하지 않는 모든 그래프 유형을 자동으로 필터링합니다.
-  Plot 노드는 수치 필드 사이의 관계를 보여줍니다. 포인트(산점도) 또는 선을 사용하여 도표를 작성할 수 있습니다.
-  분포 노드는 대출 유형이나 성별 같은 기호적(범주형) 값의 발생을 보여줍니다. 일반적으로, 데이터의 불균형을 표시하기 위해 분포 노드를 사용하는 경우 모델을 작성하기 전에 균형 노드를 사용하여 교정할 수 있습니다.
-  히스토그램 노드는 수치 필드에 대한 값의 발생을 표시합니다. 보통 조작 및 모델 작성 전에 데이터를 탐색하는 데 사용합니다. 분포 노드와 비슷하게, 히스토그램 노드는 자주 데이터의 불균형을 드러내 보입니다.
-  요약도표 노드는 다른 필드의 값에 상대적으로 하나의 숫자 필드의 값의 분포를 표시합니다. (히스토그램과 유사한 그래프를 작성합니다.) 값이 시간에 따라 변하는 변수 또는 필드를 설명하는 데 유용합니다. 3-D 그래프를 사용하여 범주별 분포를 표시하는 기호 축을 포함할 수도 있습니다.
-  다중 도표 노드는 단일 X 필드 위에 다중 Y 필드를 표시하는 도표를 작성합니다. Y 필드는 색상이 지정된 선으로 도표됩니다. 각각은 스타일이 Line으로 설정되고 X 모드가 Sort로 설정된 Plot 노드와 동등합니다. 다중 도표는 시간에 따라서 여러 변수의 변동을 탐색하기 원할 때 유용합니다.



웹 노드는 둘 이상의 기호(범주형) 필드의 값 사이의 관계의 강도를 설명합니다. 그래프는 다양한 너비의 선을 사용하여 연결 강도를 표시합니다. 예를 들어 웹 노드를 사용하여 전자상거래 사이트에 있는 항목 세트의 구매 사이의 관계를 탐색할 수 있습니다.



시간 구성 노드는 하나 이상의 시계열 데이터 세트를 표시합니다. 일반적으로, 먼저 시간 간격 노드를 사용하여 *TimeLabel* 필드를 작성하는데, 이것이 x축을 레이블하는 데 사용합니다.



평가 노드는 예측 모델을 평가하고 비교하는 데 도움이 됩니다. 평가 차트는 모델이 특정 결과를 얼마나 잘 예측하는지를 보여줍니다. 예측값과 예측의 신뢰도를 바탕으로 레코드를 정렬합니다. 레코드를 동일한 크기의 그룹(분위수)으로 분할한 후 각 분위수에 대한 비즈니스 기준의 값을 가장 높은 값부터 가장 낮은 값으로 도표를 그립니다. 다중 모델이 도표에 선구분 변수로 표시됩니다.



맵 시각화 노드는 다중 입력 연결을 승인하고 지리 공간적 데이터를 맵에 일련의 레이어로 표시할 수 있습니다. 각각의 레이어는 하나의 지리 공간적 필드입니다. 예를 들어, 기준 레이어가 한 국가의 맵이고 그 위에 도로에 대한 레이어 하나, 강에 대한 레이어 하나, 도시에 대한 레이어 하나가 있을 수 있습니다.



E-Plot(베타) 노드는 수치 필드 사이의 관계를 보여줍니다. 이는 Plot 노드와 유사하나 옵션이 다르며 출력이 이 노드에 한정된 새 그래프 인터페이스를 사용합니다. 베타 레벨 노드를 사용하여 새 그래프 기능을 활용할 수 있습니다.



t-SNE(t-Distributed Stochastic Neighbor Embedding)는 고차원 데이터를 시각화하기 위한 도구입니다. 이는 데이터 점의 연관관계를 확률로 변환합니다. SPSS Modeler에서 t-SNE 노드는 Python으로 구현되며 scikit-learn© Python 라이브러리가 필요합니다.

그래프 노드를 스트림에 추가한 경우에는 해당 노드를 두 번 클릭하여 옵션을 지정하기 위한 대화 상자를 열 수 있습니다. 대부분의 그래프에는 하나 이상의 탭에 제공된 다수의 고유 옵션이 포함되어 있습니다. 모든 그래프에 공통인 몇몇 탭 옵션도 있습니다. 다음의 주제 예는 이 공통 옵션에 대한 자세한 정보가 포함되어 있습니다.

그래프 노드에 대한 옵션을 구성한 경우에는 대화 상자 내에서 또는 스트림의 일부로 해당 옵션을 실행할 수 있습니다. 생성된 그래프 창에서는 데이터의 영역 또는 선택사항을 기반으로 파생(세트 및 플래그) 및 선택 노드를 생성하여 사실상 데이터의 "서브세트를 작성"할 수 있습니다. 예를 들어, 이 강력한 기능을 사용하여 이상치를 식별하고 제외할 수 있습니다.

① 모양, 오버레이, 패널 및 애니메이션

오버레이 및 모양

모양(및 오버레이)은 시각화에 차원성을 추가합니다. 모양(그룹화, 군집화 또는 누적)의 효과는 시각화 유형, 필드(변수) 유형, 그래픽 요소 유형 및 통계에 따라 다릅니다. 예를 들어, 색상에 대한 범주형 필드는 산점도에서 점을 그룹화하거나 누적 막대형 차트에서 누적을 작성하는 데 사용할 수 있습니다. 색상에 대한 연속형 숫자 범위는 산점도에 있는 각 점의 범위 값을 표시하는 데 사용할 수 있습니다.

요구에 맞는 모양과 오버레이를 찾으려면 여러 모양과 오버레이를 사용하여 실험해 보아야 합니다. 다음 설명은 적합한 모양과 오버레이를 선택하는 데 도움이 될 수 있습니다.

참고: 모든 모양 또는 오버레이를 모든 시각화 유형에 사용할 수 있는 것은 아닙니다.

- **색상.** 색상이 범주형 필드에 의해 정의되면 개별 범주를 기준으로 각 범주에 한 색상씩 시각화를 분할합니다. 색상이 연속형 숫자 범위일 때는 범위 필드의 값에 따라 색상이 달라집니다. 그래픽 요소(예: 막대 또는 선택란)가 둘 이상의 레코드/케이스를 나타내고 범위 필드가 색상에 사용되는 경우 범위 필드의 **평균**에 따라 색상이 달라집니다.
- **형태.** 형태는 시각화를 각 범주에 하나씩 서로 다른 여러 형태의 요소로 분할하는 범주형 필드에 의해 정의됩니다.
- **투명도.** 투명도가 범주형 필드에 의해 정의되는 경우 개별 범주를 기준으로 각 범주에 한 투명도 수준씩 시각화를 분할합니다. 투명도가 연속형 숫자 범위일 때는 범위 필드의 값에 따라 투명도가 달라집니다. 그래픽 요소(예: 막대 또는 선택란)가 둘 이상의 레코드/케이스를 나타내고 범위 필드가 투명도에 사용되는 경우 범위 필드의 **평균**에 따라 색상이 달라집니다. 가장 큰 값에서 그래픽 요소는 완전 투명합니다. 가장 작은 값에서는 완전 불투명합니다.
- **데이터 레이블.** 데이터 레이블은 해당 값이 그래픽 요소에 연결되는 레이블을 작성하는 데 사용되는 필드 유형에 의해 정의됩니다.
- **크기.** 크기가 범주형 필드에 의해 정의되면 개별 범주를 기준으로 각 범주에 한 크기씩 시각화를 분할합니다. 크기가 연속형 숫자 범위일 때는 범위 필드의 값에 따라 크기가 달라집니다. 그래픽 요소(예: 막대 또는 선택란)가 둘 이상의 레코드/케이스를 나타내고 범위 필드가 크기에 사용되는 경우 범위 필드의 **평균**에 따라 크기가 달라집니다.

패널링 및 애니메이션

패널링. 패널링(면 작성이라고도 함)은 그래프 테이블을 작성합니다. 패널링 필드에 범주당 하나의 그래프가 생성되지만 모든 패널이 동시에 표시됩니다. 패널링은 패널링 필드의 조건이 시각화에 영향을 미치는지 여부를 확인하는 데 유용합니다. 예를 들어, 빈도 분포가 남성과 여성 간에 동일한지 판별하기 위해 성별로 히스토그램을 패널링할 수 있습니다. 즉, 성별 차이가 급여에 영향을 미치는지 여부를 확인할 수 있습니다. 패널링할 범주형 필드를 선택하십시오.

애니메이션. 애니메이션은 애니메이션 필드의 값으로 여러 그래프가 작성된다는 점에서 패널링과 비슷하지만 이러한 그래프는 함께 표시되지 않습니다. 오히려 사용자가 탐색 모드의 제어를 사용하여 출력을 애니메이션하고 개별 그래프를 시퀀스대로 표시합니다. 또한 패널링과 달리 애니메이션은 범주형 필드를 필요로 하지 않습니다. 값이 자동으로 범위로 분할되는 연속형 필드를 지정할 수 있습니다. 탐색 모드에서 애니메이션 제어로 범위의 크기를 바꿀 수 있습니다. 모든 시각화에서 애니메이션을 제공하는 것은 아닙니다.

② 출력 탭 사용

모든 그래프 유형에 대해 생성된 그래프의 표시 및 파일 이름에 대해 다음과 같은 옵션을 지정할 수 있습니다.

 **참고:** 분포 노드 그래프에는 추가적인 설정이 있습니다.

출력 이름. 노드가 실행될 때 생성되는 그래프의 이름을 지정합니다. **자동**은 출력을 생성하는 노드를 기반으로 이름을 선택합니다. 선택적으로 **사용자 정의를** 선택하여 다른 이름을 지정할 수 있습니다.

화면으로 출력. 새 창에서 그래프를 생성하고 표시하려면 선택하십시오.

파일로 출력. 출력을 파일로 저장하려면 선택하십시오.

- **그래프 출력.** 그래프 형식으로 출력을 생성하려면 선택하십시오. 분포 노드에서만 사용할 수 있습니다.
- **테이블 출력.** 테이블 형식으로 출력을 생성하려면 선택하십시오. 분포 노드에서만 사용할 수 있습니다.
- **파일 이름.** 생성되는 그래프 또는 테이블에 사용되는 파일 이름을 지정하십시오. 생략 기호 단추(..)를 사용하여 특정 파일 및 위치를 지정하십시오.
- **파일 유형.** 드롭 다운 목록에서 파일 유형을 지정하십시오. **테이블 출력** 옵션을 가진 분포 노드를 제외한 모든 그래프 노드에 대해 사용 가능한 그래프 파일 유형은 다음과 같습니다.

- 비트맵(.bmp)

- PNG(.png)

- 출력 오브젝트(.cou)

- JPEG(.jpg)

- HTML(.html)

- 다른 IBM® SPSS® Statistics 애플리케이션에서 사용하기 위한 ViZml 문서(.xml)

분포 노드의 **테이블 출력** 옵션에 대해 사용 가능한 파일 유형은 다음과 같습니다.

- 탭으로 구분된 데이터(.tab)
- 쉼표로 구분된 데이터(.csv)
- HTML(.html)
- 출력 오브젝트(.cou)

출력 페이지 번호 매기기. 출력을 HTML로 저장하면 이 옵션이 사용으로 설정되어 각 HTML 페이지의 크기를 제어할 수 있습니다. (분포 노드에만 적용됩니다.)

페이지당 선. 출력 페이지 번호 매기기가 선택되면 이 옵션이 사용으로 설정되어 각 HTML 페이지의 길이를 판별할 수 있습니다. 기본 설정은 400개의 행입니다. (분포 노드에만 적용됩니다.)

③ 주석 탭 사용

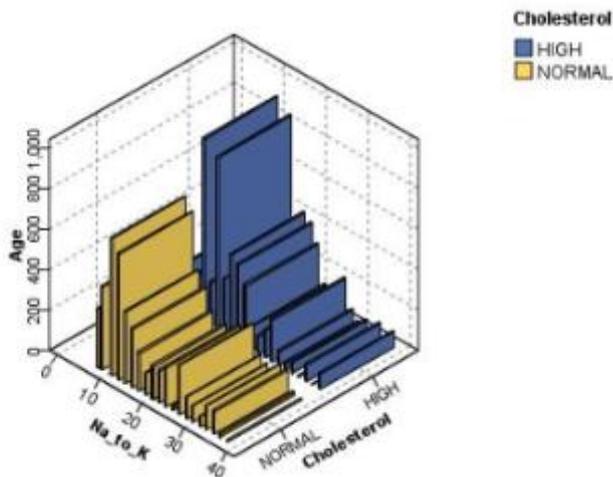
이 탭은 모든 노드에 사용되며 노드의 이름을 바꾸고 사용자 맞춤 도구팁을 제공하며 긴 주석을 저장하기 위한 옵션을 제공합니다.

④ 3차원 그래프

IBM® SPSS® Modeler의 도표 및 컬렉션 그래프는 세 번째 축에 정보를 표시할 수 있습니다. 이를 통해 서브셋을 선택하기 위해 데이터를 시각화하고 모델링을 위해 새 필드를 파생시킬 때 추가적인 유연성이 제공됩니다.

3차원 그래프를 작성한 후에는 해당 그래프를 클릭한 후 마우스를 끌어서 회전시켜 임의의 각에서 볼 수 있습니다.

그림 1. x, y 및 z축이 있는 컬렉션 그래프



IBM SPSS Modeler에서는 세 번째 축에서 정보를 도표화(진정한 3차원 그래프)하고 3차원 효과로 그래프를 표시하는 두 가지 방식으로 3차원 그래프를 작성할 수 있습니다. 두 방법 모두 도표 및 콜렉션에 사용할 수 있습니다.

세 번째 축에서 정보를 도표화하려면 다음을 수행하십시오.

1. 그래프 노드 대화 상자에서 **도표** 탭을 클릭하십시오.
2. 3차원 단추를 클릭하여 z축에 대한 옵션을 사용으로 설정하십시오.
3. 필드 선택기 단추를 사용하여 z축에 대한 필드를 선택하십시오. 일부 경우에는 기호 필드만 여기서 허용됩니다. 필드 선택기는 적절한 필드를 표시합니다.

3차원 효과를 그래프에 추가하려면 다음을 수행하십시오.

1. 그래프를 작성하고 나면 출력 창에서 **그래프** 탭을 클릭하십시오.
2. 3차원 단추를 클릭하여 보기를 3차원 그래프로 전환하십시오.

(2) 그래프보드 노드

그래프보드 노드를 사용하면 단일 노드에서 다양한 그래프 출력(막대형 차트, 원형 차트, 히스토그램, 산점도, 히트 맵 등) 중에서 선택할 수 있습니다. 첫 번째 탭에서 탐색할 데이터 필드를 선택하여 시작하면 노드에서 데이터에 대해 작동하는 그래프 유형을 선택할 수 있습니다. 이 노드는 필드 선택사항에 대해 작업하지 않는 모든 그래프 유형을 자동으로 필터링합니다. 세부사항 탭에서 상세 또는 고급 그래프 옵션을 정의할 수 있습니다.

 **참고:** 노드를 편집하거나 그래프 유형을 선택하기 위해 그래프보드 노드를 데이터가 있는 스트림에 연결해야 합니다.

사용 가능한 시각화 템플릿(및 스타일 시트 및 맵)을 제어할 수 있게 하는 두 개의 단추가 있습니다.

관리. 컴퓨터에서 시각화 템플릿, 스타일시트 및 맵을 관리합니다. 로컬 시스템에서 시각화 템플릿, 스타일시트 및 맵을 가져오고, 내보내고, 이름을 바꾸고, 삭제할 수 있습니다. 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리의 내용을 참조하십시오.

위치. 시각화 템플릿, 스타일시트 및 맵이 저장된 위치를 변경합니다. 현재 위치는 단추의 오른쪽에 표시됩니다. 자세한 정보는 템플릿, 스타일시트 및 맵 위치 설정의 내용을 참조하십시오.

① 그래프보드 기본 탭

어느 시각화 유형이 데이터를 가장 잘 표현하는지 확신할 수 없는 경우에는 기본 탭을 사용하십시오. 데이터를 선택하면 해당 데이터에 적합한 시각화 유형 서브세트가 표시됩니다. 예를 들어, 그래프보드 예제의 내용을 참조하십시오.

1. 목록에서 하나 이상의 필드(변수)를 선택하십시오. 여러 개의 필드를 선택하려면 Ctrl+클릭을 사용하십시오.
필드의 측정 수준은 사용 가능한 시각화 유형을 결정합니다. 목록에서 필드를 마우스 오른쪽 단추로 클릭하고 옵션을 선택하여 측정 수준을 변경할 수 있습니다. 사용 가능한 측정 수준 유형에 대한 자세한 정보는 필드(변수) 유형의 내용을 참조하십시오.
2. 시각화 유형을 선택하십시오. 사용 가능한 유형에 대한 설명은 사용 가능한 내장 그래프보드 시각화 유형의 내용을 참조하십시오.
3. 특정 시각화의 경우 요약 통계를 선택할 수 있습니다. 통계가 개수 기반 통계인지 또는 연속형 필드에서 계산되는지 여부에 따라 사용 가능한 통계 서브세트가 다릅니다. 또한 템플릿 자체에 따라 사용 가능한 통계가 다릅니다. 다음 단계 뒤에 사용 가능한 전체 통계 목록을 제공합니다.
4. 추가 옵션(예: 선택적 모양 및 패널 필드)을 정의하려면 **세부사항**을 클릭하십시오. 자세한 정보는 그래프보드 세부사항 탭의 내용을 참조하십시오.

연속형 필드에서 계산한 요약 통계

- **평균(Mean)**. 중심 경향에 대한 척도입니다. 합계를 케이스 수로 나눈 산술 평균 값입니다.
- **중앙값(Median)**. 전체 케이스의 절반이 위 아래에 해당되는 값으로 제50 백분위수입니다. 케이스 수가 짝수인 경우 중앙값은 케이스를 오름차순이나 내림차순으로 정렬했을 때 중간에 있는 두 개의 케이스의 평균입니다. 중앙값은 평균과 달리 중심을 벗어난 값에는 영향을 받지 않는 중심 경향 척도이며, 상한 극단값 또는 하한 극단값에 따라 달라질 수 있습니다.
- **최빈값(Mode)**. 가장 자주 발생하는 값입니다. 여러 값에서 최대 발생 빈도를 공유하는 경우 각각을 최빈값이라고 합니다.
- **최소값(Minimum)**. 숫자변수의 가장 작은 값입니다.
- **최대값(Maximum)**. 숫자변수의 가장 큰 값입니다.
- **범위**. 최소값과 최대값의 차이입니다.
- **중간 범위**. 범위의 중간 값으로 최소값과의 차이와 최대값과의 차이가 같은 값입니다.
- **합계(Sum)**. 비결측값을 갖는 전체 케이스 값의 총계입니다.
- **누적 합계**. 값의 누적 합계입니다. 각 그래픽 요소는 하나의 하위 그룹 합계와 이전의 모든 그룹의 총 합계를 더한 값을 표시합니다.
- **퍼센트 합계**. 모든 그룹의 합계에 대비되는 합산 필드 기준의 각 하위 그룹 내 백분율입니다.
- **누적 퍼센트 합계**. 모든 그룹의 합계에 대비되는 합산 필드 기준의 각 하위 그룹 내 누적 백분율입니다. 각 그래픽 요소는 하나의 하위 그룹의 백분율과 이전의 모든 그룹의 전체 백분율을 더한 값을 표시합니다.
- **분산(Variance)**. 평균에 대한 산포 척도로, 평균으로부터의 제곱합 편차를 케이스 수에서 1을 뺀 값으로 나눈 값과 같습니다. 분산은 변수 자체의 제곱 단위로 측정됩니다.

- **표준 편차(Standard Deviation)**. 평균에 대한 산포 측도입니다. 정규 분포에서 케이스의 68%는 평균의 표준 편차 내에 있으며 케이스의 95%는 2배 표준 편차 내에 있습니다. 예를 들어, 평균 연령이 45세이고 표준 편차가 10인 경우 정규 분포 내에서 95% 케이스는 25세와 65세 사이에 있습니다.
- **표준 오차(Standard Error)**. 검정 통계량 값이 표본마다 얼마나 달라지는지에 대한 측도입니다. 이 항목은 통계에 대한 표본 분포의 표준 편차가 됩니다. 예를 들어, 평균의 표준 오차는 표본 평균의 표준 편차입니다.
- **첨도(Kurtosis)**. 이상치가 있는 정도에 대한 측도입니다. 정규 분포의 경우 첨도 통계 값은 0입니다. 양(+)의 첨도는 데이터가 정규 분포보다 더 극단적인 이상치를 나타냄을 표시합니다. 음의 첨도는 데이터가 정규 분포보다 극단적인 이상치를 나타냄을 표시합니다.
- **왜도(Skewness)**. 분포의 비대칭성에 대한 측도입니다. 정규 분포는 대칭이므로 왜도 값이 0입니다. 양의 왜도가 많은 분포는 오른쪽이 깎입니다. 유의한 음의 왜도를 가지는 분포에는 왼쪽으로 긴 꼬리가 나타납니다. 왜도값이 표준 오차의 두 배를 넘는 것은 대칭에서 벗어난 정도를 나타냅니다.

다음과 같은 지역 통계는 하위 그룹당 둘 이상의 그래픽 요소를 생성할 수 있습니다. 구간, 영역 또는 가장자리 그래픽 요소를 사용하는 경우 지역 통계는 범위를 표시하는 하나의 그래픽 요소를 생성합니다. 다른 모든 그래픽 요소는 두 개의 개별 요소, 즉 범위의 시작을 표시하는 요소와 범위의 끝을 표시하는 요소를 생성합니다.

- **지역: 범위**. 최소값과 최대값 사이의 값 범위입니다.
- **지역: 평균의 95% 신뢰구간**. 모집단 평균을 포함할 가능성이 95%인 값 범위입니다.
- **지역: 개별의 95% 신뢰구간**. 주어진 개별 케이스의 예측값을 포함할 가능성이 95%인 값 범위입니다.
- **지역: 평균 이상/이하의 1배 표준 편차**. 평균의 이상 및 이하로 1배 표준 편차만큼 떨어져 있는 값 사이의 범위입니다.
- **지역: 평균 이상/이하의 1배 표준 오차**. 평균의 이상 및 이하로 1배 표준 오차만큼 떨어져 있는 값 사이의 범위입니다.

개수 기준 요약 통계

- **개수**. 행/케이스의 수입입니다.
- **누적 개수**. 행/케이스의 수입입니다. 각 그래픽 요소는 하나의 하위 그룹의 개수와 이전의 모든 그룹의 총계를 더한 값을 표시합니다.
- **개수 퍼센트**. 행/케이스의 총 수에 대비되는 각 하위 그룹의 행/케이스 백분율입니다.
- **누적 개수 퍼센트**. 행/케이스의 총 수에 대비되는 각 하위 그룹의 행/케이스 누적 백분율입니다. 각 그래픽 요소는 하나의 하위 그룹의 백분율과 이전의 모든 그룹의 전체 백분율을 더한 값을 표시합니다.

가. 필드(변수) 유형

아이콘은 필드 목록에서 필드 옆에 표시되며 필드 유형과 데이터 유형을 나타냅니다. 아이콘은 또한 다중 응답 세트를 식별합니다.

표 1. 측정 수준 아이콘

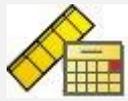
측정 수준	숫자	문자열	날짜	시간
연속		해당사항없음		
순서형 세트				
세트				

표 2. 다중 응답 세트 아이콘

다중반응세트 유형	아이콘
다중 응답 세트, 다중 범주	
다중 응답 세트, 다중 이분형	

측정 수준

필드의 측정 수준은 시각화를 만들 때 중요한 역할을 합니다. 다음은 측정 수준에 대한 설명입니다. 필드 목록에서 필드를 마우스 오른쪽 단추로 클릭하고 옵션을 선택하여 측정 수준을 임시로 변경할 수 있습니다. 대부분의 경우 가장 광범위한 필드 분류인 범주형과 연속형 두 가지만 고려해야 합니다.

범주형. 고유한 값 또는 범주의 수가 제한된 데이터(예: 성별 또는 종교)입니다. 문자열(영숫자) 필드 또는 숫자 코드를 사용하여 범주를 나타내는 숫자 필드(예: 0 = 남성, 1 = 여성)가 범주형 필드에 해당될 수 있습니다. 질적 데이터라고도 합니다. 세트, 순서형 세트 및 플래그는 모두 범주형 필드입니다.

- **세트.** 해당 값이 고유한 순위가 없는 범주를 나타내는 필드/변수입니다(예: 직원이 근무하는 회사의 부서). 명목 변수의 예는 지역, 우편번호 및 종교입니다. 명목 변수라고도 합니다.

- **순서형 세트.** 해당 값이 고유한 순위가 있는 범주를 나타내는 필드/변수입니다(예: 매우 불만족에서 매우 만족에 이르는 서비스 만족도 수준). 순서형 세트의 예로는 만족도나 신뢰도를 나타내는 태도 스코어 및 선호도 등급 스코어가 있습니다. 순서 변수라고도 합니다.
- **플래그.** 두 개의 고유한 값(예: 예와 아니오 또는 1과 2)이 있는 필드/변수입니다. 이분형 변수라고도 합니다.

연속형. 구간 또는 비율 척도로 측정된 데이터이며 데이터 값은 값의 순서와 값 사이의 차이를 모두 나타냅니다. 예를 들어, 급여 \$72,195는 급여 \$52,398보다 높으며 두 값 사이의 차이는 \$19,797입니다. 양적, 척도 또는 숫자 범위 데이터라고도 합니다.

범주형 필드는 일반적으로 별도의 그래픽 요소를 그리거나 그래픽 요소를 그룹화하기 위해 시각화에서 범주를 정의합니다. 연속형 필드는 대개 범주형 필드의 범주 내에 요약됩니다. 예를 들어, 성별 범주에 대한 수입의 기본 시각화에는 남자의 평균 수입과 여자의 평균 수입이 표시됩니다. 산점도에서와 마찬가지로 연속형 필드의 원래 값을 도표화할 수도 있습니다. 예를 들어, 산점도는 각 케이스의 현재 급여와 시작 급여를 표시할 수 있습니다. 범주형 필드를 사용하여 케이스를 성별로 그룹화할 수 있습니다.

데이터 유형

측정 수준이 필드 유형을 결정하는 필드의 유일한 특성인 것은 아닙니다. 필드는 특정 데이터 유형으로도 저장됩니다. 가능한 데이터 유형은 문자열(문자와 같이 숫자가 아닌 데이터), 숫자 값(실수) 및 날짜입니다. 필드의 데이터 유형은 측정 수준과 달리 임시로 변경할 수 없습니다. 데이터가 원래 데이터 세트에 저장되는 방식을 변경해야 합니다.

다중 응답 세트

일부 데이터 파일에서는 **다중 응답 세트**라는 특수한 종류의 "필드"도 지원합니다. 다중 응답 세트는 일반적인 의미에서는 실제로 "필드"가 아닙니다. 다중 응답 세트는 반응자가 둘 이상의 응답을 제공할 수 있는 질문에 대해 다중 필드를 사용하여 응답을 기록합니다. 다중 응답 세트는 범주형 필드처럼 처리되며 범주형 필드로 수행할 수 있는 대부분의 작업은 다중 응답 세트로도 수행할 수 있습니다.

다중 응답 세트는 다중 이분형 세트 또는 다중 범주 세트일 수 있습니다.

다중 이분형 세트. 다중 이분형 세트는 일반적으로 예/아니오, 유/무, 선택함/선택 안 함 등과 같이 값을 두 개만 가질 수 있는 다중 이분형 필드로 구성됩니다. 필드가 엄격하게는 이분형이 아닐 수도 있지만 변수 세트의 모든 필드가 같은 방식으로 코딩됩니다.

예를 들어, 설문조사에서는 "다음 중 주로 어디에서 뉴스를 보십니까?"라는 질문에 대해 다섯 가지의 선택 가능한 응답을 제공할 수 있습니다. 반응자는 각 선택사항 옆에 있는 선택란을 선택하여 복수의 선택사항을 표시할 수 있습니다. 다섯 가지 응답은 데이터 파일에서 다섯 개의 필드가 되며 *아니오*(선택 안 함)는 0으로, *예*(선택함)는 1로 코딩됩니다.

다중 범주 세트. 다중 범주 세트는 대개 선택 가능한 다수의 반응 범주를 포함하고 모두 같은 방법으로 코딩되는 다중 필드로 구성됩니다. 예를 들어, "여러분의 민족 전통을 가장 잘 설명하는 민족성을 세 개까지 나열해보십시오."라는 설문조사 항목이 있습니다. 수백 개의 선택 가능한 응답이 있지만 코딩을 위해 목록에는 가장 일반적인 민족성 40개만 나열하고 그 외의 나머지는 "기타" 범주로 분류합니다. 데이터 파일에서 세 개의 선택사항은 세 개의 필드가 되고 각 필드에는 41개의 범주(40개의 코딩된 민족성과 하나의 "기타" 범주)가 포함됩니다.

② 그래프보드 세부사항 탭

작성할 시각화 유형을 알고 있거나 시각화에 선택적 모양, 패널 및/또는 애니메이션을 추가하려는 경우에는 세부사항 탭을 사용하십시오. 예를 들어, 그래프보드 예제의 내용을 참조하십시오.

1. 기본 탭에서 시각화 유형을 선택했으면 해당 유형이 표시됩니다. 그렇지 않은 경우에는 드롭다운 목록에서 시각화 유형을 선택하십시오. 시각화 유형에 대한 정보는 사용 가능한 내장 그래프보드 시각화 유형의 내용을 참조하십시오.
2. 시각화 썸네일 이미지 바로 오른쪽에는 시각화 유형에 필요한 필드(변수)를 지정하는 제어가 있습니다. 이러한 필드를 모두 지정해야 합니다.
3. 특정 시각화의 경우 요약 통계를 선택할 수 있습니다. 일부의 경우(예: 막대형 차트) 투명도 모양에 이러한 요약 옵션 중 하나를 사용할 수 있습니다. 요약 통계에 대한 설명은 그래프보드 기본 탭의 내용을 참조하십시오.
4. 선택적 모양을 하나 이상 선택할 수 있습니다. 이 경우 시각화에 다른 필드를 포함시킬 수 있으므로 차원을 추가할 수 있습니다. 예를 들어, 필드를 사용하여 산점도에 있는 점의 크기에 변화를 줄 수 있습니다. 선택적 모양에 대한 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오. 스크립팅을 통해서도 투명도 모양이 지원되지 않습니다.
5. 맵 시각화를 작성하는 경우 **맵 파일** 그룹은 사용할 맵 파일을 표시합니다. 기본 맵 파일이 있으면 이 파일이 표시됩니다. 맵 파일을 변경하려면 **맵 파일 선택**을 클릭하여 맵 선택 대화 상자를 표시하십시오. 이 대화 상자에서 기본 맵 파일을 지정할 수도 있습니다. 자세한 정보는 맵 시각화를 위한 맵 파일 선택의 내용을 참조하십시오.
6. 패널링 또는 애니메이션 옵션 중 하나 이상 선택할 수 있습니다. 패널링 및 애니메이션 옵션에 대한 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

가. 맵 시각화를 위한 맵 파일 선택

맵 시각화 템플릿을 선택하는 경우 맵을 그리기 위한 지리적 정보를 정의하는 맵 파일이 필요합니다. 기본 맵 파일이 있으면 이 파일이 맵 시각화에 사용됩니다. 다른 맵 파일을 선택하려면 세부사항 탭에서 **맵 파일 선택**을 클릭하여 맵 선택 대화 상자를 표시하십시오.

맵 선택 대화 상자에서는 기본 맵 파일과 참조 맵 파일을 선택할 수 있습니다. 맵 파일은 맵을 그리기 위한 지리적 정보를 정의합니다. 애플리케이션은 표준 맵 파일과 함께 설치됩니다. 사용

하려는 다른 ESRI 형태 파일이 있는 경우 먼저 형태 파일을 SMZ 파일로 변환해야 합니다. 자세한 정보는 맵 형태 파일 변환 및 배포의 내용을 참조하십시오. 맵을 변환한 후에는 템플릿 선택기 대화 상자에서 **관리...**를 클릭하여 맵을 관리 시스템으로 가져오십시오. 그러면 맵 선택 대화 상자에서 해당 맵을 사용할 수 있습니다.

다음은 맵 파일을 지정할 때 고려해야 할 사항입니다.

- 모든 맵 템플릿은 하나 이상의 맵 파일이 필요합니다.
- 맵 파일은 일반적으로 맵 키 속성을 데이터 키에 연결합니다.
- 템플릿에 데이터 키에 연결하는 맵 키가 필요 없는 경우에는 참조 맵 파일과 참조 맵에 요소를 그리기 위한 좌표(예: 경도 및 위도)를 지정하는 필드가 필요합니다.
- 오버레이 맵 템플릿은 두 개의 맵, 즉 기본 맵 파일과 참조 맵 파일이 필요합니다. 참조 맵이 기본 맵 파일 뒤에 있도록 참조 맵이 먼저 그려집니다.

속성 및 지형과 같은 맵 용어에 대한 정보는 맵의 핵심 개념의 내용을 참조하십시오.

맵 파일. 관리 시스템에 있는 어떤 맵 파일이든 선택할 수 있습니다. 여기에는 사전 설치된 맵 파일 및 가져온 맵 파일도 포함됩니다. 맵 파일 관리에 대한 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리의 내용을 참조하십시오.

맵 키. 맵 파일을 데이터 키에 연결하는 키로 사용할 속성을 지정하십시오.

이 맵 파일 및 설정을 기본값으로 저장. 선택한 맵 파일을 기본값으로 사용하려면 이 선택란을 선택하십시오. 기본 맵 파일을 지정한 경우에는 맵 시각화를 작성할 때마다 맵 파일을 지정하지 않아도 됩니다.

데이터 키. 이 제어는 템플릿 선택기 세부사항 탭에 표시되는 것과 동일한 값을 나열합니다. 여기서는 선택하는 특정 맵 파일로 인해 키를 변경해야 하는 경우 편의를 위해 제공됩니다.

시각화에 모든 맵 지형 표시. 이 옵션을 선택하면 일치하는 데이터 키 값이 없는 경우에도 시각화에 맵의 모든 지형이 렌더링됩니다. 데이터가 있는 지형만 보려면 이 옵션을 선택 취소하십시오. **일치하지 않는 맵 키** 목록에 표시된 맵 키로 식별되는 지형이 시각화에 렌더링되지 않습니다.

맵과 데이터 값 비교. 맵 키와 데이터 키는 서로 연결되어 맵 시각화를 작성합니다. 맵 키 및 데이터 키는 동일한 도메인(예: 국가 및 지역)에서 그려야 합니다. 데이터 키 및 맵 키 값이 일치하는지 테스트하려면 **비교**를 클릭하십시오. 표시되는 아이콘은 비교 상태를 알려줍니다. 아래에 이러한 아이콘에 대한 설명이 있습니다. 비교를 수행한 후 일치하는 맵 키 값이 없는 데이터 키 값이 있으면 해당 데이터 키 값이 **일치하지 않는 데이터 키** 목록에 표시됩니다. **일치하지 않는 맵 키** 목록에는 일치하는 데이터 키 값이 없는 맵 키 값이 표시됩니다. **시각화에 모든 맵 지형 표시**를 선택하지 않은 경우, 이러한 맵 키 값으로 식별되는 지형은 렌더링되지 않습니다.

표 1. 비교 아이콘

아이콘	설명
	비교를 수행하지 않았습니다. 비교 를 클릭하기 전의 기본 상태입니다. 데이터 키와 맵 키의 값이 일치하는지 모르므로 주의해서 진행해야 합니다.
	비교를 수행했으며 데이터 키와 맵 키의 값이 완전히 일치합니다. 데이터 키 값마다 맵 키로 식별되는 일치하는 지형이 있습니다.
	비교를 수행했으며 일부 데이터 키와 맵 키의 값이 일치하지 않습니다. 일부 데이터 키 값의 경우 맵 키로 식별되는 일치하는 지형이 없습니다. 주의해서 진행해야 합니다. 진행하는 경우 맵 시각화에 일부 데이터 값이 포함되지 않습니다.
	비교를 수행했으며 데이터 키 값과 맵 키 값이 일치하지 않습니다. 진행할 경우 맵이 렌더링되지 않으므로 다른 데이터 키 또는 맵 키를 선택해야 합니다.

③ 사용 가능한 내장 그래프보드 시각화 유형

다양한 여러 시각화 유형을 작성할 수 있습니다. 기본 및 세부사항 탭에서 다음에 나열된 내장 유형을 모두 사용할 수 있습니다. 템플릿(특히 맵 템플릿)에 대한 일부 설명은 **특수 텍스트**를 사용하여 세부사항 탭에 지정된 필드(변수)를 식별합니다.

표 1. 사용 가능한 그래프 유형

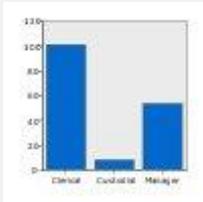
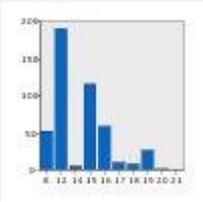
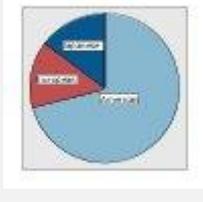
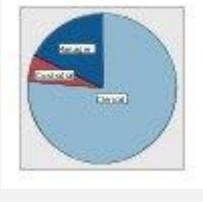
차트 아이콘	설명	차트 아이콘	설명
	막대 연속형 숫자 필드 에 대한 요약 통계를 계산하고 범주형 필드의 각 범주에 대한 결과를 막대로 표시합니다. 필수: 범주형 필드 및 연속형 필드.		개수 막대형 차트 범주형 필드 의 각 범주에 있는 행/케이스의 비율을 막대로 표시합니다. 분포 그래프 노드를 사용하여 이 그래프를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 분포 노드 주제를 참조하십시오. 필수: 하나의 범주형 필드.
	원 연속형 숫자 필드 의 합계를 계산하고 범주형 필드의 각 범주에 분포된 연속형 숫자 필드 합 비율을 원의 조각으로 표시합니다. 필수: 범주형 필드 및 연속형 필드.		개수 원형 차트 범주형 필드의 각 범주에 있는 행/케이스의 비율을 원의 조각으로 표시합니다. 필수: 하나의 범주형 필드.

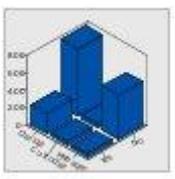
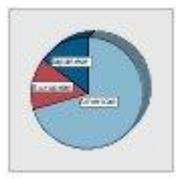
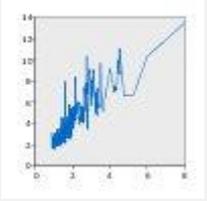
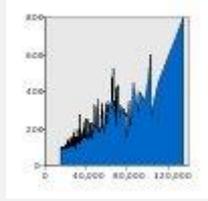
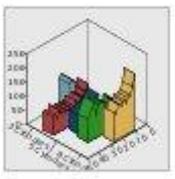
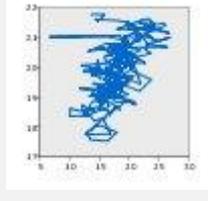
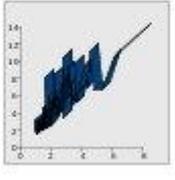
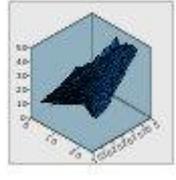
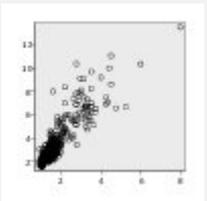
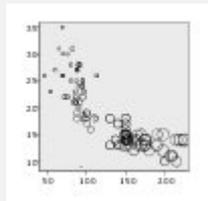
차트 아이콘	설명	차트 아이콘	설명
	<p>3차원 막대형 차트 연속형 숫자 필드에 대한 요약 통계를 계산하고 두 범주형 필드의 범주가 교차하는 지점의 결과를 표시합니다.</p> <p>필수: 범주형 필드 쌍 및 연속형 필드.</p>		<p>3차원 원형 차트 이 차트는 추가된 3차원 효과를 제외하면 원형 차트와 동일합니다.</p> <p>필수: 범주형 필드 및 연속형 필드.</p>
	<p>선 한 필드의 각 값에 대한 다른 필드의 요약 통계를 계산하고 값을 연결하는 선을 그립니다. 도표 그래프 노드를 사용하여 선 도표 그래프를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 Plot 노드 주제를 참조하십시오.</p> <p>필수: 임의 유형의 필드 쌍.</p>		<p>영역 한 필드의 각 값에 대한 다른 필드의 요약 통계를 계산하고 값을 연결하는 영역을 그립니다. 영역은 아래 공간이 채색된 선과 유사하므로 선과 영역 간의 차이는 매우 작습니다. 그러나 색상 모양을 사용할 경우 선이 단순하게 분할되고 영역이 누적됩니다.</p> <p>필수: 임의 유형의 필드 쌍.</p>
	<p>3차원 영역 한 필드의 값에 대해 구성되고 범주형 필드에 의해 분할된 다른 필드의 값을 표시합니다. 범주마다 영역 요소가 그려집니다.</p> <p>필수: 범주형 필드 및 임의 유형의 필드 쌍.</p>		<p>경로 한 필드의 값에 대해 구성된 다른 필드의 값을 표시하고 이러한 값을 원래 데이터 세트에 표시된 순서대로 하나의 선으로 연결합니다. 순서 지정이 경로와 선의 주된 차이점입니다.</p> <p>필수: 임의 유형의 필드 쌍.</p>
	<p>리본도표 한 필드의 각 값에 대한 다른 필드의 요약 통계를 계산하고 값을 연결하는 리본을 그립니다. 리본은 본질적으로 3차원 효과를 갖는 선입니다. 진정한 3차원 그래프는 아닙니다.</p> <p>필수: 임의 유형의 필드 쌍.</p>		<p>Surface 서로의 값에 대해 구성된 세 필드의 값을 표시하고 이러한 값을 하나의 표면으로 연결합니다.</p> <p>필수: 임의 유형의 세 개의 필드.</p>
	<p>산점도 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. 이 그래프는 필드 사이의 관계를 강조 표시할 수 있습니다(관계가 있는 경우). 도표 그래프 노드를 사용하여 산점도를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 Plot 노드 주제를 참조하십시오.</p> <p>필수: 임의 유형의 필드 쌍.</p>		<p>거품 도표 기본 산점도와 마찬가지로 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. 차이점은 세 번째 필드의 값이 개별 점의 크기에 변화를 주는 데 사용된다는 점입니다.</p> <p>필수: 임의 유형의 세 개의 필드.</p>

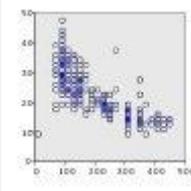
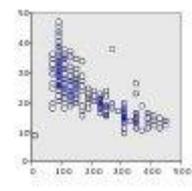
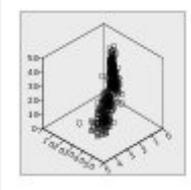
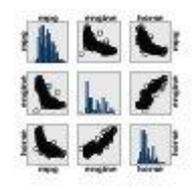
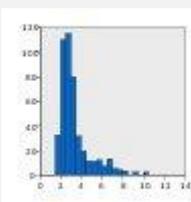
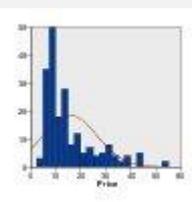
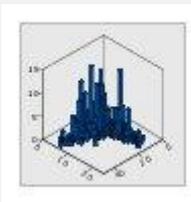
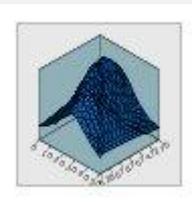
차트 아이콘	설명	차트 아이콘	설명
	<p>구간화된 산점도 기본 산점도와 마찬가지로 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. 차이점은 유사한 값이 그룹으로 구간화되고 색상 또는 크기 모양이 각 구간의 케이스 수를 표시하는 데 사용된다는 점입니다.</p> <p>필수 : 연속형 필드 쌍.</p>		<p>육각형 구간화된 산점도 구간화된 산점도 설명을 참조하십시오. 차이점은 기본 구간의 형태이며 구간이 원이 아니라 육각형과 비슷합니다. 결과로 생성되는 육각형 구간화된 산점도는 구간화된 산점도와 유사합니다. 그러나 기본 구간의 형태가 다르기 때문에 두 그래프 간에 각 구간의 값 수가 다릅니다.</p> <p>필수 : 연속형 필드 쌍.</p>
	<p>3차원 산점도 (3-D Scatterplot) 서로에 대해 구성된 세 필드의 값을 표시합니다. 이 그래프는 필드 사이의 관계를 강조 표시할 수 있습니다(관계가 있는 경우). 도표 그래프 노드를 사용하여 3차원 산점도를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 Plot 노드 주제를 참조하십시오.</p> <p>필수: 임의 유형의 세 개의 필드.</p>		<p>산점도 행렬(SPLOM) 필드마다 한 필드의 값에 대해 구성된 다른 필드의 값을 표시합니다. SPLOM은 산점도 테이블과 유사합니다. SPLOM에도 각 필드의 히스토그램이 포함됩니다.</p> <p>필수 : 두 개 이상의 연속형 필드.</p>
	<p>히스토그램 필드의 빈도 분포를 표시합니다. 히스토그램은 분포 유형을 판별하고 분포가 비대칭인지 여부를 확인하는데 도움이 될 수 있습니다. 히스토그램 그래프 노드를 사용하여 이 그래프를 생성할 수도 있습니다. 이 노드는 일부 추가 옵션을 제공합니다. 자세한 정보는 히스토그램 도표 탭 주제를 참조하십시오.</p> <p>필수: 임의 유형의 하나의 필드.</p>		<p>정규 분포 히스토그램 정규 분포 곡선이 겹쳐진 연속형 필드의 빈도 분포를 표시합니다.</p> <p>필수: 하나의 연속성 필드.</p>
	<p>3차원 히스토그램 연속형 필드 쌍의 빈도 분포를 표시합니다.</p> <p>필수 : 연속형 필드 쌍.</p>		<p>3차원 밀도 연속형 필드 쌍의 빈도 분포를 표시합니다. 3차원 히스토그램과 유사하며 유일한 차이점은 분포를 표시하는 데 막대 대신 표면이 사용된다는 점입니다.</p> <p>필수 : 연속형 필드 쌍.</p>

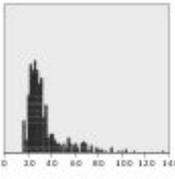
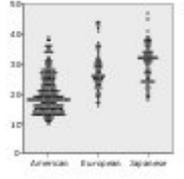
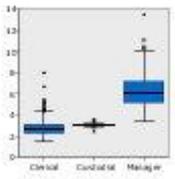
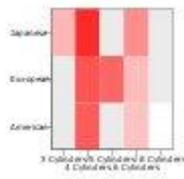
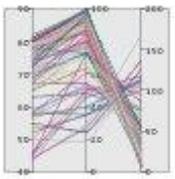
차트 아이콘	설명	차트 아이콘	설명
	<p>점도표 개별 케이스/행을 표시하고 x축의 고유 데이터 포인트에서 케이스/행을 누적시킵니다. 이 그래프는 데이터의 분포를 표시하는 점에서 히스토그램과 유사하지만 특정 구간(값 범위)의 집계된 개수가 아니라 각 케이스/행을 표시합니다. 필수: 임의 유형의 하나의 필드.</p>		<p>2차원 점도표 범주형 필드의 범주마다 개별 케이스/행을 표시하고 y축의 고유 데이터 포인트에서 케이스/행을 누적시킵니다. 필수: 범주형 필드 및 연속형 필드.</p>
	<p>상자도표 범주형 필드의 범주마다 연속형 필드의 5가지 통계(최소값, 첫 번째 사분위수, 중앙값, 세 번째 사분위수 및 최대값)를 계산합니다. 결과가 상자도표/스키마 요소로 표시됩니다. 상자도표를 통해 범주 간에 연속형 데이터의 분포가 얼마나 다른지 알 수 있습니다. 필수: 범주형 필드 및 연속형 필드.</p>		<p>히트 맵 두 범주형 필드 사이에서 범주가 교차하는 지점의 연속형 필드 평균을 계산합니다. 필수: 범주형 필드 쌍 및 연속형 필드.</p>
	<p>동형 각 필드에 대한 평행 축을 만들고 데이터의 케이스/행마다 필드 값을 지나는 선을 그립니다. 필수: 두 개 이상의 연속형 필드.</p>		<p>개수의 코로플레스 범주형 필드(데이터 키)의 각 범주에 대한 개수를 계산하고 범주에 해당하는 맵 지형에서 색포화도를 사용하여 개수를 나타내는 맵을 그립니다. 필수: 범주형 필드. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>
	<p>평균/중앙값/합계의 코로플레스 범주형 필드(데이터 키)의 각 범주에 대한 연속형 필드(색상)의 평균, 중앙값 또는 합계를 계산하고 범주에 해당하는 맵 지형에서 색포화도를 사용하여 계산된 통계를 나타내는 맵을 그립니다. 필수: 범주형 필드 및 연속형 필드. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>		<p>깊의 코로플레스 하나의 범주형 필드(데이터 키)로 정의된 값에 해당하는 맵 지형에 대한 다른 범주형 필드(색상)의 값을 색상을 사용하여 나타내는 맵을 그립니다. 각 지형에 대한 색상 필드의 범주형 값이 여러 개인 경우 모달 값이 사용됩니다. 필수: 범주형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>

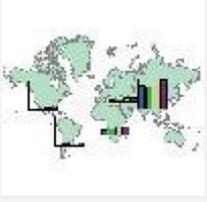
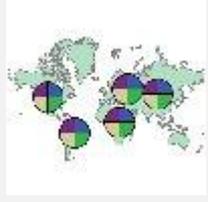
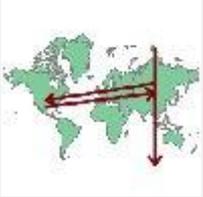
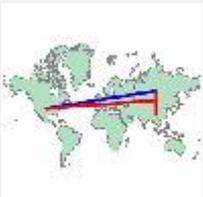
차트 아이콘	설명	차트 아이콘	설명
	<p>개수의 코로플레스 위의 좌표 코로플레스 맵에 점을 그리기 위한 좌표를 식별하는 두 개의 추가적 연속형 필드(경도 및 위도)가 있다는 점을 제외하곤 개수의 코로플레스와 유사합니다.</p> <p>필수: 범주형 필드와 연속형 필드의 쌍. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>		<p>평균/중앙값/합계의 코로플레스 위의 좌표 코로플레스 맵에 점을 그리기 위한 좌표를 식별하는 두 개의 추가적 연속형 필드(경도 및 위도)가 있다는 점을 제외하곤 평균/중앙값/합계의 코로플레스와 유사합니다.</p> <p>필수: 범주형 필드 및 세 개의 연속형 필드. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>
	<p>값의 코로플레스 위의 좌표 코로플레스 맵에 점을 그리기 위한 좌표를 식별하는 두 개의 추가적 연속형 필드(경도 및 위도)가 있다는 점을 제외하곤 값의 코로플레스와 유사합니다.</p> <p>필수: 범주형 필드 쌍과 연속형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>		<p>맵 위의 개수 막대형 차트 각 맵 지형(데이터 키)에 대해 범주형 필드(범주)의 각 범주에 있는 행/케이스의 비율을 계산하여 맵과 함께 각 맵 지형의 중앙에 막대형 차트를 그립니다.</p> <p>필수: 범주형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>
	<p>맵 위의 막대형 차트 연속형 필드(값)의 요약 통계를 계산하고 각 맵 지형(데이터 키)에 대한 범주형 필드(범주)의 각 범주 결과를 각 맵 지형의 중앙에 위치한 막대형 차트로 표시합니다.</p> <p>필수: 범주형 필드 쌍 및 연속형 필드. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>		<p>맵 위의 개수 원형 차트 각 맵 지형(데이터 키)에 대해 범주형 필드(범주)의 각 범주에 있는 행/케이스의 비율을 표시하고 맵과 함께 각 맵 지형의 중앙에 비율을 원형 차트의 조각으로 그립니다.</p> <p>필수: 범주형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>
	<p>맵 위의 원형 차트 각 맵 지형(데이터 키)에 대해 범주형 필드(범주)의 각 범주에 있는 연속형 필드(값)의 합계를 계산하고 맵과 함께 각 맵 지형의 중앙에 합계를 원형 차트의 조각으로 그립니다.</p> <p>필수: 범주형 필드 쌍 및 연속형 필드. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>		<p>맵 위의 선형 차트 각 맵 지형(데이터 키)에 대해 한 필드(X)의 각 값에 대한 다른 연속형 필드(Y)의 요약 통계를 계산하고 맵과 함께 각 맵 지형의 중앙에 값을 연결하는 선형 차트를 그립니다.</p> <p>필수: 범주형 필드 및 임의 유형의 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 맵 파일.</p>

차트 아이콘	설명	차트 아이콘	설명
	<p>참조 맵 위의 좌표 점의 좌표를 식별하는 연속형 필드(경도 및 위도)를 사용하여 맵과 점을 그립니다.</p> <p>필수: 범위 필드 쌍. 맵 파일.</p>		<p>참조 맵 위의 화살표 각 화살표의 시작점(시작 경도 및 시작 위도)과 종료점(종료 경도 및 종료 위도)을 식별하는 연속형 필드를 사용하여 맵과 화살표를 그립니다. 데이터의 각 레코드/케이스가 맵에서 하나의 화살표를 생성합니다.</p> <p>필수: 네 개의 연속형 필드. 맵 파일.</p>
	<p>점 오버레이 맵 참조 맵을 그리고 그 위에 점 지형이 범주형 필드(색상)로 채색된 다른 점 맵을 겹칩니다.</p> <p>필수: 범주형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 점 맵 파일. 참조 맵 파일.</p>		<p>다각형 오버레이 맵 참조 맵을 그리고 그 위에 다각형 지형이 범주형 필드(색상)로 채색된 다른 다각형 맵을 겹칩니다.</p> <p>필수: 범주형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 다각형 맵 파일. 참조 맵 파일.</p>
	<p>선 오버레이 맵 참조 맵을 그리고 그 위에 선 지형이 범주형 필드(색상)로 채색된 다른 선 맵을 겹칩니다.</p> <p>필수: 범주형 필드 쌍. 해당 키가 데이터 키 범주와 일치하는 선 맵 파일. 참조 맵 파일.</p>		

④ 맵 시각화 작성

다수의 시각화는 관심 필드(변수)와 이러한 필드를 시각화할 템플릿이 두 가지만 선택하면 됩니다. 추가적 선택이나 조치가 필요하지 않습니다. 그러나 맵 시각화를 작성하려면 최소한 하나의 추가 단계가 필요합니다. 즉, 맵 시각화를 위한 지리적 정보를 정의하는 맵 파일을 선택해야 합니다.

단순한 맵을 작성하는 기본 단계는 다음과 같습니다.

1. 기본 탭에서 관심 필드를 선택하십시오. 다양한 맵 시각화에 필요한 필드 유형 및 수에 대한 정보는 사용 가능한 내장 그래프보드 시각화 유형의 내용을 참조하십시오.
2. 맵 템플릿을 선택하십시오.
3. 세부사항 탭을 클릭하십시오.
4. **데이터 키** 및 기타 필요한 드롭 다운 목록이 올바른 필드로 설정되었는지 확인하십시오.
5. 맵 파일 그룹에서 **맵 파일 선택**을 클릭하십시오.
6. 맵 선택 대화 상자를 사용하여 맵 파일 및 맵 키를 선택하십시오. 맵 키의 값은 **데이터 키**로 지정한 필드의 값과 일치해야 합니다. **비교** 단추를 사용하여 이러한 값을 비교할 수 있습니다. 오버레이 맵 템플릿을 선택하는 경우에는 참조 맵도 선택해야 합니다. 참조 맵은 데이터에 맞추어져 있지 않습니다. 참조 맵은 주요 맵의 배경으로 사용됩니다. 맵 선택 대화 상자에 대한 자세한 정보는 맵 시각화를 위한 맵 파일 선택의 내용을 참조하십시오.
7. **확인**을 클릭하여 맵 선택 대화 상자를 닫으십시오.
8. 그래프보드 템플릿 선택기에서 **실행**을 클릭하여 맵 시각화를 작성하십시오.

⑤ 그래프보드 예제

이 절에는 사용 가능한 옵션을 설명하는 서로 다른 여러 예제가 있습니다. 이러한 예제에서는 또한 시각화 결과물에 대한 해석 정보를 제공합니다.

이 예제에서는 *graphboard.str*이라는 스트림을 사용하고 이 스트림은 *employee_data.sav*, *customer_subset.sav* 및 *worldsales.sav*라는 데이터 파일을 참조합니다. 이러한 파일은 IBM® SPSS® Modeler 클라이언트 설치의 *데모* 폴더에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *graphboard.str* 파일은 *스트림* 폴더에 있습니다. 표시된 순서대로 예제를 읽는 것이 좋습니다. 이어지는 예제는 이전 예제를 기반으로 작성됩니다.

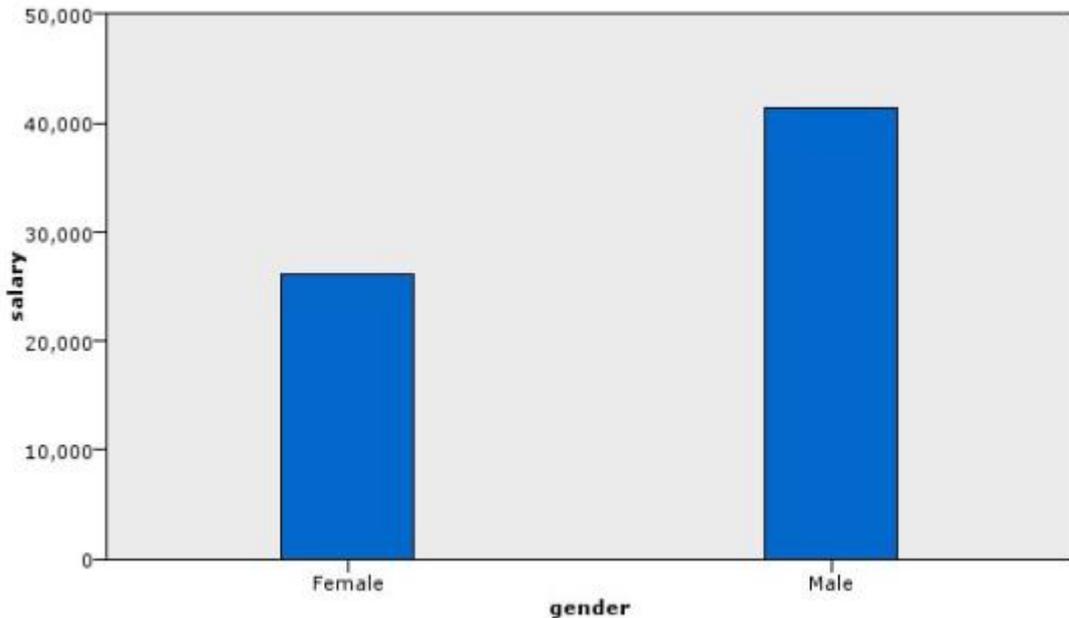
가. 예제: 요약 통계가 포함된 막대형 차트

세트/범주형 변수의 각 범주에 대해 연속형 숫자 필드/변수를 요약한 막대형 차트를 작성합니다. 구체적으로 남여 평균 급여를 보여주는 막대형 차트를 작성합니다.

이 예제와 다음 예제 중 일부에서는 회사의 직원에 대한 정보가 포함된 가설 데이터 세트인 *직원 데이터*를 사용합니다.

1. *employee_data.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.
2. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
3. 기본 탭에서 *성별* 및 *현재 급여*를 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
4. **막대**를 선택하십시오.
5. 요약 드롭 다운 목록에서 **평균**을 선택하십시오.
6. **실행**을 클릭하십시오.
7. 결과로 표시되는 화면에서 "필드 및 값 레이블 표시" 도구 모음 단추(도구 모음 가운데 있는 두 개의 단추 중 두 번째)를 클릭하십시오.

그림 1. 요약 통계가 포함된 막대형 차트



다음을 관측할 수 있습니다.

- 막대의 높이를 볼 때 남자의 평균 급여가 여자의 평균 급여보다 높습니다.

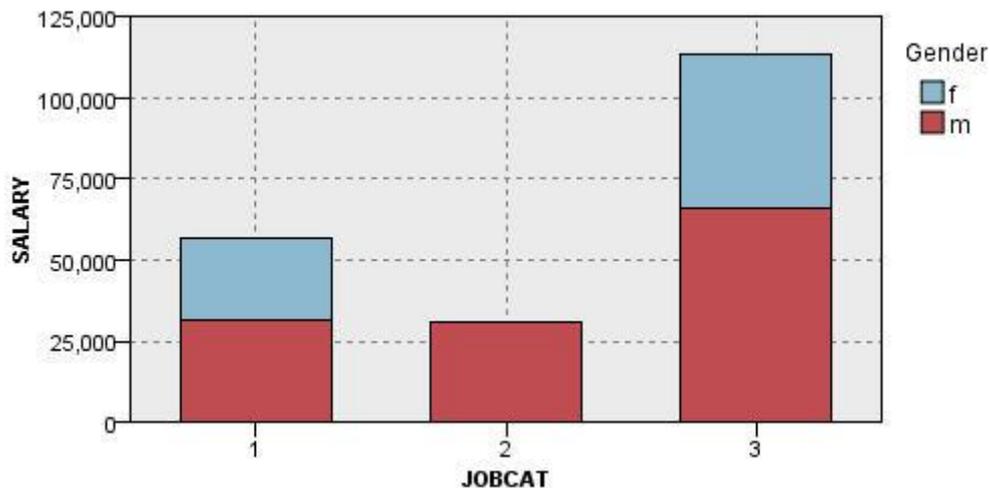
나. 예제: 요약 통계가 포함된 누적 막대형 차트

이제 누적 막대형 차트를 만들어 남녀 평균 급여의 차이가 직업 유형과 관련이 있는지 확인할 수 있습니다. 특정 직업 유형에서는 여자의 평균 급여가 남자보다 높을 수 있습니다.

 **참고:** 이 예제에서는 직원 데이터를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 **직원 범주** 및 **현재 급여**를 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **막대**를 선택하십시오.
4. 요약 목록에서 **평균**을 선택하십시오.
5. 세부사항 탭을 클릭하십시오. 이전 탭에서의 선택이 여기에 반영됩니다.
6. 선택적 모양 그룹의 색상 드롭 다운 목록에서 **성별**을 선택하십시오.
7. **실행**을 클릭하십시오.

그림 1. 누적 막대형 차트



다음을 관측할 수 있습니다.

- 각 직업 유형에 따른 평균 급여 차이는 모든 남녀의 평균 급여를 비교한 막대형 차트만큼 커 보이지 않습니다. 그룹에 따라 남녀 수가 다를 수 있습니다. 개수 막대형 차트를 만들어 이를 확인할 수 있습니다.
- 직업 유형에 관계없이 남자의 평균 급여가 항상 여자의 평균 급여보다 높습니다.

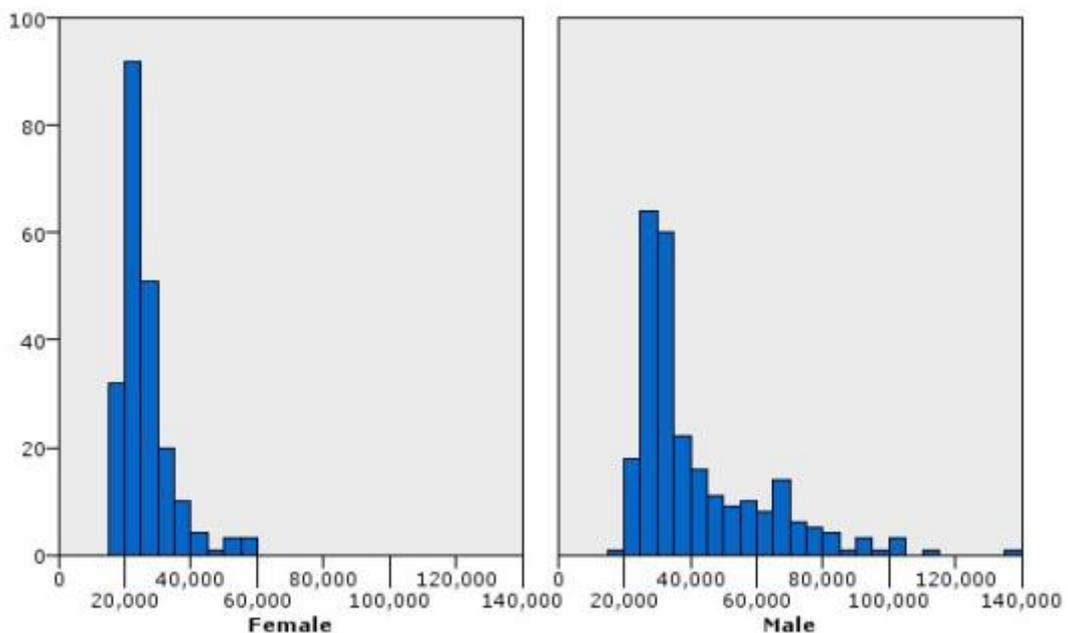
다. 예제: 패널링된 히스토그램

남녀 급여의 빈도 분포를 비교할 수 있도록 성별로 패널링된 히스토그램을 만듭니다. 빈도 분포는 특정 급여 범위 내에 얼마나 많은 케이스/행이 포함되는지 보여줍니다. 패널링된 히스토그램을 사용하여 성별에 따른 급여 차이를 더 자세히 분석할 수 있습니다.

참고: 이 예제는 직원 데이터를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 **현재 급여**를 선택하십시오.
3. **히스토그램**을 선택하십시오.
4. 세부사항 탭을 클릭하십시오.
5. 패널 및 애니메이션 그룹의 패널 전체 드롭 다운 목록에서 **성별**을 선택하십시오.
6. **실행**을 클릭하십시오.

그림 1. 패널링된 히스토그램



다음을 관측할 수 있습니다.

- 두 빈도 분포 모두 정규 분포가 아닙니다. 즉, 이 두 히스토그램은 데이터가 정규 분포일 때 보이는 종 곡선과 유사하지 않습니다.
- 더 높은 막대가 각 그래프의 왼쪽에 있습니다. 따라서 남녀 모두 더 높은 급여가 아니라 더 낮은 급여를 더 작성해야 합니다.
- 남자와 여자의 급여 빈도 분포가 서로 같지 않습니다. 히스토그램의 모양에 주의하십시오. 높은 급여를 받는 사람은 남자가 여자보다 많습니다.

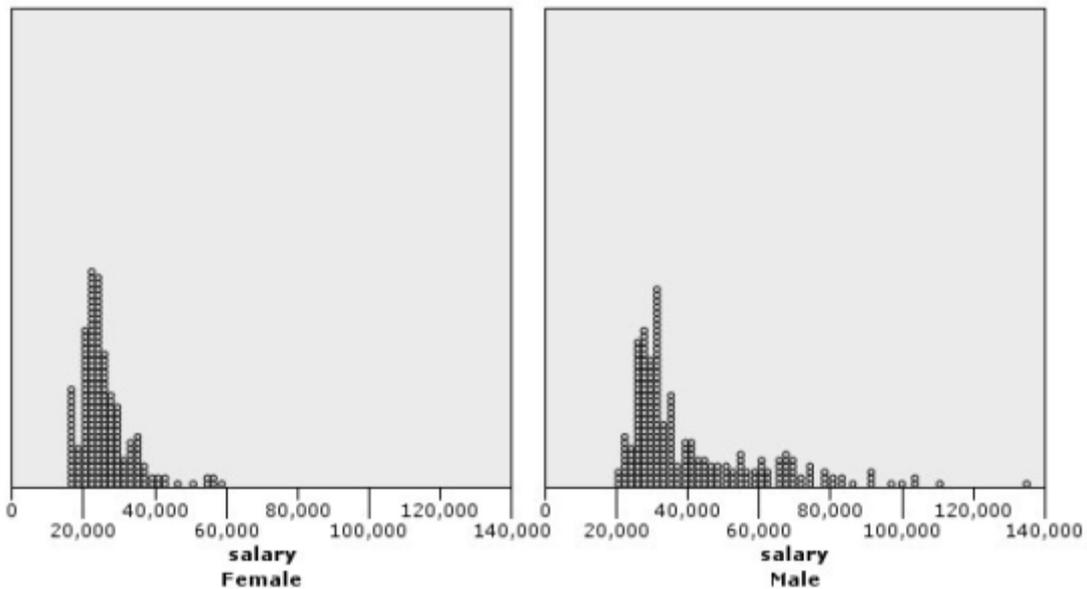
라. 예제: 패널링된 점도표

히스토그램처럼 점도표는 연속형 숫자 범위의 분포를 보여줍니다. 구간화된 데이터 범위에 대한 개수를 보여주는 히스토그램과는 달리 점도표는 데이터에 있는 모든 행/케이스를 보여줍니다. 따라서 점도표는 히스토그램에 비해 높은 세분성(granularity)을 제공합니다. 실제로 빈도 분포를 분석할 때 시작점으로 점도표를 더 선호할 수 있습니다.

참고: 이 예제는 직원 데이터를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *현재 급여*를 선택하십시오.
3. **점도표**를 선택하십시오.
4. 세부사항 탭을 클릭하십시오.
5. 패널 및 애니메이션 그룹의 패널 전체 드롭 다운 목록에서 *성별*을 선택하십시오.
6. **실행**을 클릭하십시오.
7. 결과로 표시되는 출력 창을 최대화하여 도표를 더욱 분명하게 볼 수 있습니다.

그림 1. 패널링된 점도표



히스토그램(예제: 패널링된 히스토그램 참조)과 비교하여 다음을 관측할 수 있습니다.

- 여자 히스토그램에 표시된 피크 20,000이 점도표에서는 그만큼 급격한 증가로 표시되지 않습니다. 다수의 케이스/행이 그 값 주위에 집중되어 있지만 대부분의 값은 25,000에 더 가깝습니다. 이러한 단위 수준은 히스토그램에서 표시되지 않습니다.
- 남자 히스토그램에서는 남자 평균 급여가 40,000에 이르러 점차 하강하지만 점도표에서는 이 값 이후부터 80,000까지 꽤 일정한 분포를 보여줍니다. 해당 범위 내의 특정 급여 값에서 세 명 이상의 남자가 특별한 급여를 받고 있습니다.

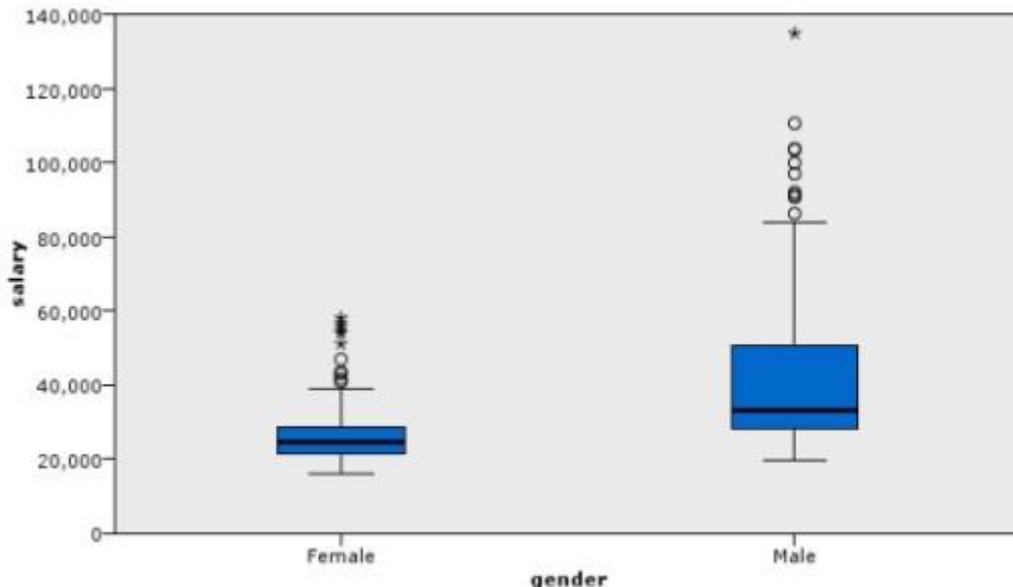
마. 예제: 상자도표

상자도표는 데이터의 분포 상태를 표시하는 또하나의 유용한 시각화입니다. 상자도표에는 시각화 작성 후 탐색하는 여러 통계 측도가 포함됩니다.

참고: 이 예제는 직원 데이터를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 **성별** 및 **현재 급여**를 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **상자도표**를 선택하십시오.
4. **실행**을 클릭하십시오.

그림 1. 상자도표



다음은 상자도표의 다양한 부분에 대한 설명입니다.

- 상자 가운데 있는 진한 선은 급여의 중앙값입니다. 케이스/행의 반은 중앙값보다 큰 값을 가지고 나머지 반은 낮은 값을 가집니다. 평균과 마찬가지로 중앙값은 중심 경향의 측도입니다. 평균과 달리 중앙값은 극단값을 갖는 케이스/행에 덜 영향을 받습니다. 이 예제에서는 중앙값이 평균보다 낮습니다(예제: 요약 통계가 포함된 막대형 차트와 비교할 때). 평균과 중앙값의 차이는 일부 케이스/행이 평균을 높이는 극단값을 가짐을 의미합니다. 즉, 일부 직원이 많은 급여를 받습니다.
- 상자의 맨 아래는 25번째 백분위수를 표시합니다. 케이스/행의 25%는 25번째 백분위수보다 낮은 값을 가집니다. 상자의 맨 위는 75번째 백분위수를 표시합니다. 케이스/행의 25%는 75번째 백분위수보다 높은 값을 가집니다. 이는 케이스/행의 50%가 상자 내에 있음을 의미합니다. 여자의 상자가 남자의 상자보다 훨씬 짧습니다. 이로써 **급여** 범위가 남자보다 여자가 더 작다는 결론이 도출됩니다. 상자의 맨 위와 맨 아래를 종종 **히지**라고 합니다.

- 상자에서 확장된 T형 막대는 **내부 펜스** 또는 **수염 도표**라고 합니다. T형 막대는 상자 높이의 1.5배까지 확장되거나, 그러한 범위 내의 값을 가진 케이스/행이 없는 경우, 최소값 또는 최대값까지 확장됩니다. 데이터가 정규 분포되어 있는 경우 데이터의 약 95%가 내부 펜스 사이에 있을 것으로 예상됩니다. 이 예제에서는 여자의 내부 펜스가 남자보다 더 적게 확장됩니다. 이는 **굽어** 범위가 남자보다 여자가 더 작다는 의미도 내포하고 있습니다.
- 점은 **이상값**입니다. 이상값은 내부 펜스에 해당하지 않는 값으로 정의됩니다. 이상값은 극단값입니다. 별표는 **극단적인 이상값**입니다. 극단적인 이상값은 상자 높이의 세 배보다 많은 값을 갖는 케이스/행을 나타냅니다. 남녀 모두 여러 개의 이상값이 있습니다. 평균이 중앙값보다 더 큼니다. 평균이 더 큰 이유는 이러한 이상값 때문입니다.

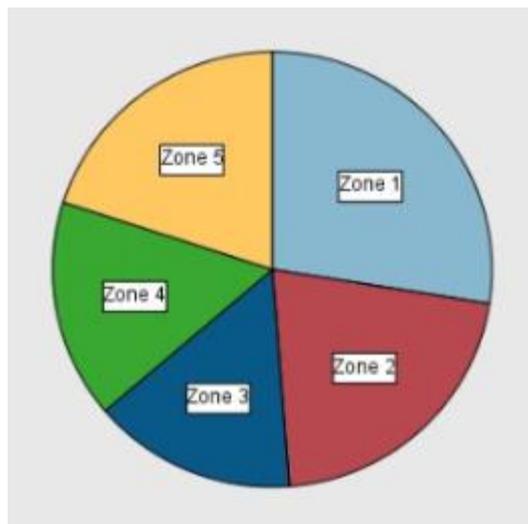
바. 예제: 원형 차트

이제 다른 데이터 세트를 사용하여 일부 다른 시각화 유형을 탐색합니다. 데이터 세트는 고객에 대한 정보가 포함된 가설 데이터 파일인 *customer_subset*입니다.

먼저 원형 차트를 만들어 서로 다른 지역의 고객 비율을 확인합니다.

1. *customer_subset.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.
2. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
3. 기본 탭에서 **지리 표시기**를 선택하십시오.
4. **개수 원형 차트**를 선택하십시오.
5. **실행**을 클릭하십시오.

그림 1. 원형 차트



다음을 관측할 수 있습니다.

- 다른 지역보다 Zone 1에 더 많은 고객이 있습니다.
- 나머지 지역에서는 고객이 균등하게 분포되어 있습니다.

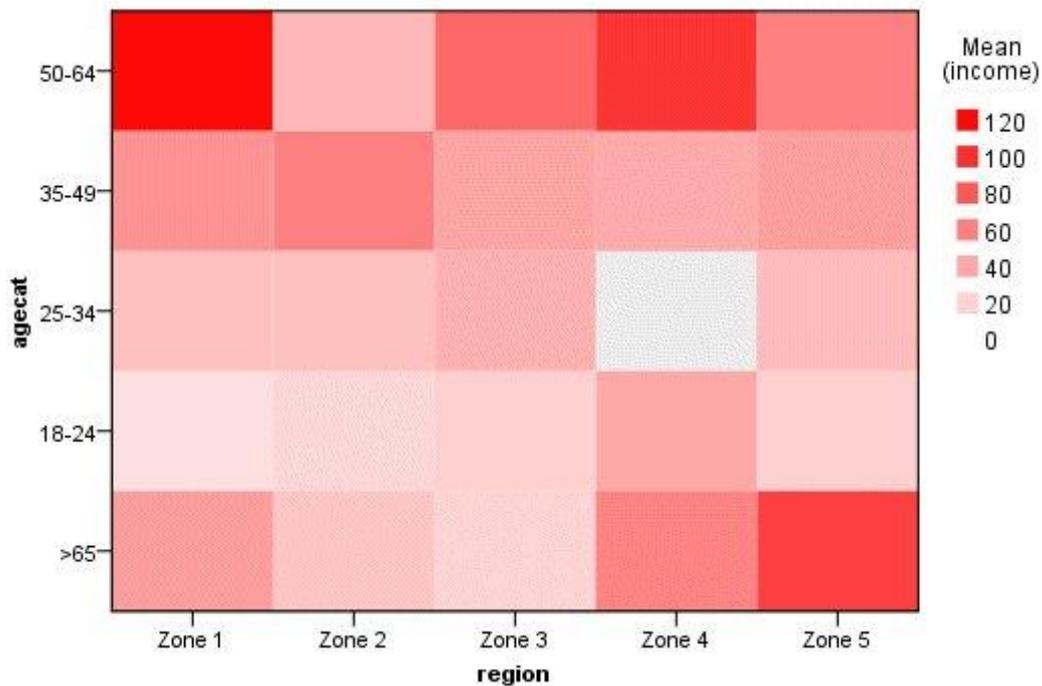
사. 예제: 히트 맵

이제 서로 다른 지역 및 연령 그룹에 속하는 고객의 평균 소득을 확인하기 위한 범주형 히트 맵을 만듭니다.

참고: 이 예제에서는 *customer_subset*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *지형 표시기*, *연령 범주* 및 *가구소득(천단위)* 순으로 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **히트 맵**을 선택하십시오.
4. **실행**을 클릭하십시오.
5. 결과로 표시되는 출력 화면에서 "필드 및 값 레이블 표시" 도구 모음 단추(도구 모음 가운데 있는 두 개 중 오른쪽 단추)를 클릭하십시오.

그림 1. 범주형 히트 맵



다음을 관측할 수 있습니다.

- 히트 맵은 숫자 대신 색상을 사용하여 셀 값을 표시하는 테이블과 비슷합니다. 밝고 짙은 빨강은 가장 높은 값을 표시하고 회색은 낮은 값을 표시합니다. 각 셀의 값은 각 범주 쌍에 대한 연속형 필드/변수의 평균입니다.
- Zone 2 및 Zone 5를 제외하고, 연령이 50세에서 64세 사이인 고객 그룹이 다른 그룹보다 평균 가구소득이 더 많습니다.
- Zone 4에는 25세에서 34세 사이의 고객이 없습니다.

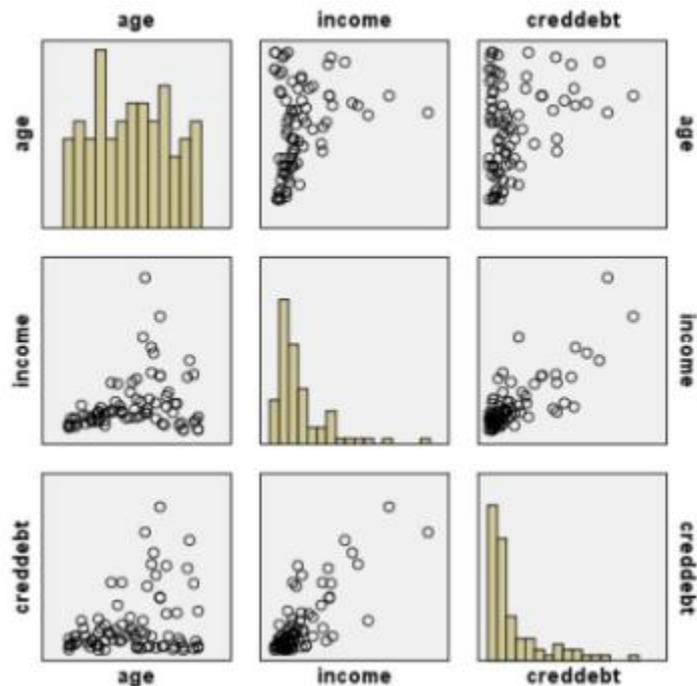
아. 예제: 산점도 행렬(SPLOM)

서로 다른 여러 변수에 대한 산점도 행렬을 작성하여 데이터 세트의 변수 간에 관계가 있는지 판별합니다.

참고: 이 예제에서는 customer_subset를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *연령*, *가구소득(천단위)* 및 *신용카드 대출(천단위)* 을 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **SPLOM**을 선택하십시오.
4. **실행**을 클릭하십시오.
5. 출력 창을 최대화하여 행렬을 더욱 분명하게 볼 수 있습니다.

그림 1. 산점도 행렬(SPLOM)



다음을 관측할 수 있습니다.

- 대각선으로 표시되는 히스토그램은 SPLOM에서 각 변수의 분포를 보여줍니다. *연령*에 대한 히스토그램은 상단 왼쪽 셀에 표시되고 *소득*에 대한 히스토그램은 중앙 셀에 표시되며 *신용대출*에 대한 히스토그램은 하단 오른쪽 셀에 표시됩니다. 정규 분포로 표시되는 변수가 없습니다. 즉, 종 곡선과 유사한 히스토그램이 없습니다. 또한 *소득* 및 *신용대출*에 대한 히스토그램은 정적으로 비대칭됩니다.
- *연령*과 다른 변수 사이에 아무런 관계가 없어 보입니다.
- *소득*과 *신용대출* 사이에는 선형 관계가 있습니다. 즉, *소득*이 증가할수록 *신용대출*이 증가합니다. 이러한 변수와 다른 관련 변수에 대한 개별 산점도를 만들어 관계를 더욱 자세히 탐색할 수 있습니다.

자. 예제: 합계의 코로플레스(색상 맵)

이제 맵 시각화를 작성합니다. 그런 후에 다음 예제에서 이 시각화의 변형을 작성할 것입니다. 데이터 세트는 대륙 및 제품별 판매 수입이 포함된 가설 데이터 파일인 *worldsales*입니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 **대륙** 및 **수입**을 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **합계의 코로플레스**를 선택하십시오.
4. 세부사항 탭을 클릭하십시오.
5. 선택적 모양 그룹의 데이터 레이블 드롭 다운 목록에서 **대륙**을 선택하십시오.
6. 맵 파일 그룹에서 **맵 파일 선택**을 클릭하십시오.
7. 맵 선택 대화 상자에서 **맵**이 **대륙**으로 설정되고 **맵 키**가 **CONTINENT**로 설정되었는지 확인하십시오.
8. 맵과 데이터 값 비교 그룹에서 **비교**를 클릭하여 맵 키가 데이터 키와 일치하는지 확인하십시오. 이 예제에서는 모든 데이터 키 값이 맵 키 및 지형과 일치합니다. 오세아니아에 대한 데이터가 없음도 알 수 있습니다.
9. 맵 선택 대화 상자에서 **확인**을 클릭하십시오.
10. **실행**을 클릭하십시오.

그림 1. 합계의 코로플레스



이 맵 시각화에서는 북미의 수입이 가장 높고 남미와 아프리카의 수입이 가장 낮다는 것을 쉽게 알 수 있습니다. 데이터 레이블 모양에 **대륙**을 사용했기 때문에 각 대륙의 레이블이 지정되어 있습니다.

차. 예제: 맵 위의 막대형 차트

이 예제는 각 대륙에서 제품별로 수입이 어떻게 나뉘는지 보여줍니다.

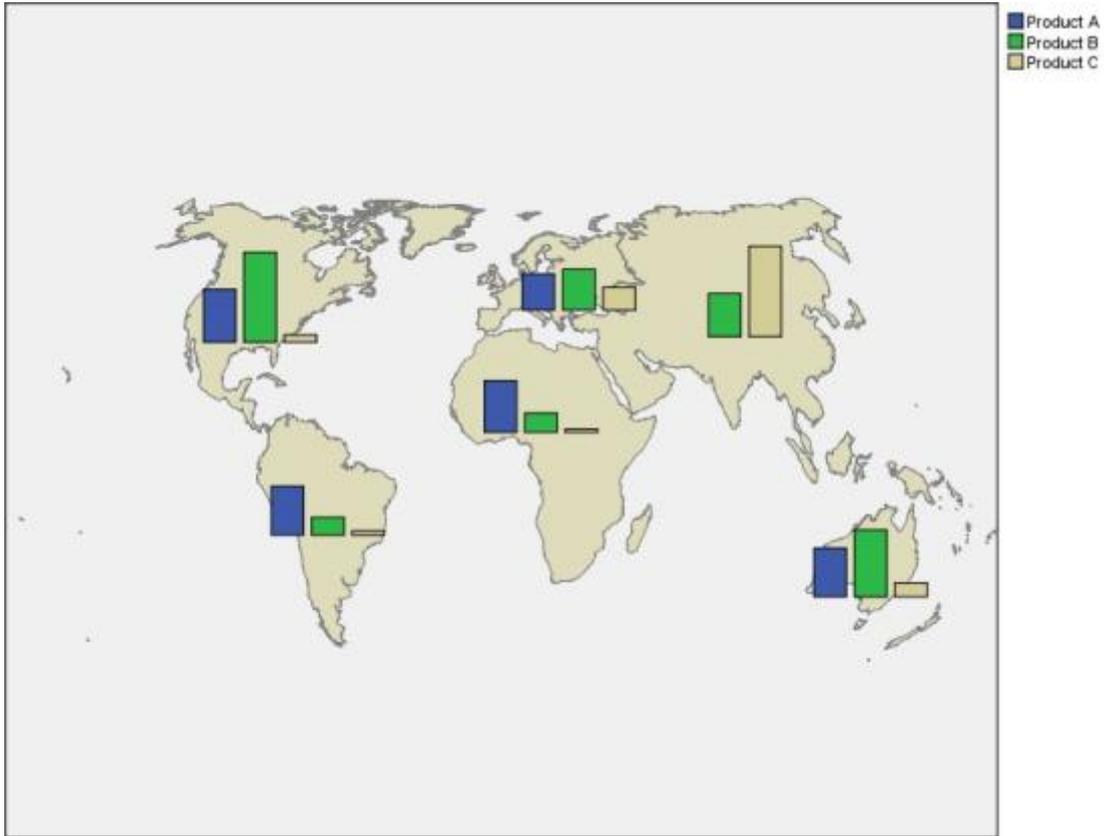
참고: 이 예제에서는 *worldsales*를 사용합니다.

1. 그래프보드 노드를 추가하여 편집을 위해 여십시오.
2. 기본 탭에서 *대륙*, *제품* 및 *수입*을 선택하십시오. (여러 필드/변수를 선택하려면 Ctrl+클릭을 사용하십시오.)
3. **맵 위의 막대형 차트**를 선택하십시오.
4. 세부사항 탭을 클릭하십시오.
특정 유형의 필드를 둘 이상 사용하는 경우 각 필드가 올바른 슬롯에 지정되었는지 확인하는 것이 중요합니다.
5. 범주 드롭 다운 목록에서 *제품*을 선택하십시오.
6. 값 드롭 다운 목록에서 *수입*을 선택하십시오.
7. 데이터 키 드롭 다운 목록에서 *대륙*을 선택하십시오.
8. 요약 드롭 다운 목록에서 *합계*를 선택하십시오.
9. 맵 파일 그룹에서 **맵 파일 선택**을 클릭하십시오.
10. 맵 선택 대화 상자에서 **맵**이 *대륙*으로 설정되고 **맵 키**가 *CONTINENT*로 설정되었는지 확인하십시오.
11. 맵과 데이터 값 비교 그룹에서 **비교**를 클릭하여 맵 키가 데이터 키와 일치하는지 확인하십시오. 이 예제에서는 모든 데이터 키 값이 맵 키 및 지형과 일치합니다. 오세아니아에 대한 데이터가 없음도 알 수 있습니다.
12. 맵 선택 대화 상자에서 **확인**을 클릭하십시오.
13. **실행**을 클릭하십시오.
14. 결과로 표시되는 출력 창을 최대화하여 화면을 더욱 분명하게 볼 수 있습니다.

다음을 관측할 수 있습니다.

- 남미와 아프리카에서 전체 제품의 총 수입 분포가 유사합니다.
- *제품 C*는 아시아를 제외한 모든 곳에서 수입이 가장 적습니다.
- 아시아에서는 *제품 A*로 인한 수입이 없거나 최소입니다.

그림 1. 맵 위의 막대형 차트



⑥ 그래프보드 모양 탭

그래프보드 모양 탭 그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

일반 모양 옵션

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

표본추출. 큰 데이터 세트에 대한 방법을 지정하십시오. 최대 데이터 세트 크기를 지정하거나 기본 레코드 수를 사용할 수 있습니다. **표본** 옵션을 선택하면 큰 데이터 세트에 대한 성능이 개선됩니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 점을 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격히 저하될 수 있다는 점에 유의해야 합니다.

스타일시트 모양 옵션

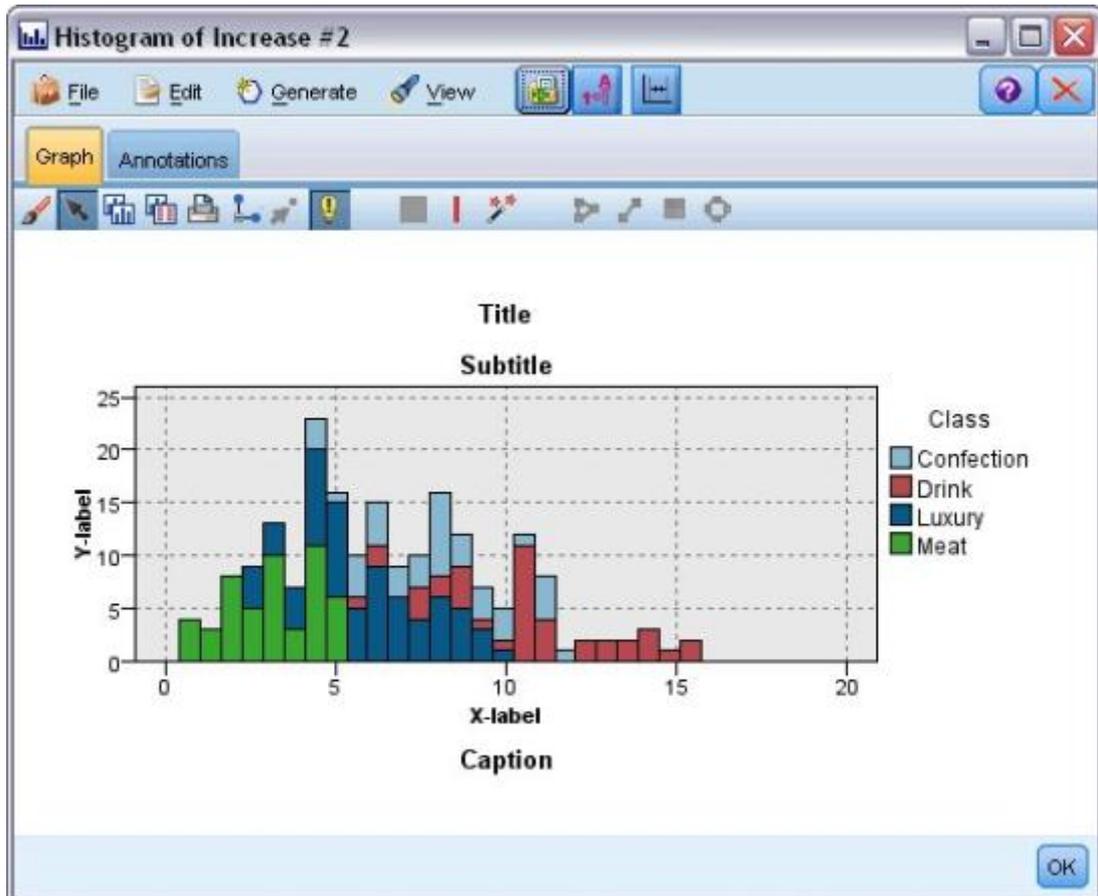
사용 가능한 시각화 템플릿(및 스타일 시트 및 맵)을 제어할 수 있게 하는 두 개의 단추가 있습니다.

관리. 컴퓨터에서 시각화 템플릿, 스타일시트 및 맵을 관리합니다. 로컬 시스템에서 시각화 템플릿, 스타일시트 및 맵을 가져오고, 내보내고, 이름을 바꾸고, 삭제할 수 있습니다. 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리의 내용을 참조하십시오.

위치. 시각화 템플릿, 스타일시트 및 맵이 저장된 위치를 변경합니다. 현재 위치는 단추의 오른쪽 쪽에 표시됩니다. 자세한 정보는 템플릿, 스타일시트 및 맵 위치 설정의 내용을 참조하십시오.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. 일부 그래프에서는 이 모든 옵션이 사용되지 않습니다.

그림 1. 다양한 그래프 모양 옵션의 위치



⑦ 템플리트, 스타일시트 및 맵 위치 설정

시각화 템플리트, 시각화 스타일시트 및 맵 파일은 특정 로컬 폴더 또는 IBM® SPSS® Collaboration and Deployment Services Repository에 저장됩니다. 템플리트, 스타일시트 및 맵을 선택하면 이 위치에 내장된 것만 표시됩니다. 모든 템플리트, 스타일시트 및 맵 파일을 한 곳에 보관하면 IBM SPSS 애플리케이션이 이러한 템플리트, 스타일시트 및 맵 파일에 쉽게 액세스할 수 있습니다. 이 위치에 템플리트, 스타일시트 및 맵 파일을 추가하는 방법에 대해서는 템플리트, 스타일시트 및 맵 파일 관리의 내용을 참조하십시오.

템플리트, 스타일시트 및 맵 파일 위치 설정 방법

1. 템플리트 또는 스타일시트 대화 상자에서 **위치...**를 클릭하여 템플리트, 스타일시트 및 맵 대화 상자를 표시하십시오.

2. 템플리트, 스타일시트 및 맵 파일의 기본 위치 옵션을 선택하십시오.

로컬 시스템. 템플리트, 스타일시트 및 맵 파일이 로컬 컴퓨터의 특정 폴더에 있습니다. Windows XP에서 이 폴더의 위치는 *C:\#Documents and Settings\#<user>\#Application Data\#SPSSInc\#Graphboard*입니다. 폴더를 변경할 수 없습니다.

IBM SPSS Collaboration and Deployment Services Repository. 템플리트, 스타일시트 및 맵 파일이 IBM SPSS Collaboration and Deployment Services Repository의 사용자 지정 폴더에 있습니다. 특정 폴더를 지정하려면 **폴더**를 클릭하십시오. 추가 정보는 IBM SPSS Collaboration and Deployment Services Repository를 템플리트, 스타일시트 및 맵 파일 위치로 사용의 내용을 참조하십시오.

3. 확인을 클릭하십시오.

가. IBM SPSS Collaboration and Deployment Services Repository를 템플리트, 스타일시트 및 맵 파일 위치로 사용

시각화 템플리트 및 스타일시트는 IBM® SPSS® Collaboration and Deployment Services Repository에 저장할 수 있습니다. 이 위치는 IBM SPSS Collaboration and Deployment Services Repository에 있는 특정 폴더입니다. 이 위치를 기본 위치로 설정하면 이 위치에 있는 모든 템플리트, 스타일시트 및 맵 파일을 선택해 사용할 수 있습니다.

이 기능에는 통계 어댑터 옵션이 필요합니다.

IBM SPSS Collaboration and Deployment Services Repository의 폴더를 템플리트, 스타일시트 및 맵 파일 위치로 설정하는 방법

1. 위치 단추가 있는 대화 상자에서 **위치...**를 클릭하십시오.

2. IBM SPSS Collaboration and Deployment Services Repository를 선택하십시오.

3. 폴더를 클릭하십시오.

참고: IBM SPSS Collaboration and Deployment Services Repository에 아직 연결되지 않은 경우 연결 정보 입력을 요구하는 프롬프트가 표시됩니다.

4. 폴더 선택 대화 상자에서 템플리트, 스타일시트 및 맵 파일이 저장된 폴더를 선택하십시오.

5. 선택적으로 레이블 검색에서 레이블을 선택할 수 있습니다. 해당 레이블을 가진 템플리트, 스타일시트 및 맵 파일만 표시됩니다.

6. 특정 템플리트 또는 스타일시트가 포함된 폴더를 찾으려면 검색 탭에서 템플리트, 스타일시트 또는 맵 파일을 검색할 수 있습니다. 폴더 선택 대화 상자는 찾은 템플리트, 스타일시트 또는 맵 파일이 있는 폴더를 자동으로 선택합니다.

7. 폴더 선택을 클릭하십시오.

⑧ 템플리트, 스타일시트 및 맵 파일 관리

템플리트, 스타일시트 및 맵 파일 대화 상자를 사용하여 컴퓨터의 로컬 위치에 있는 템플리트, 스타일시트 및 맵 파일을 관리할 수 있습니다. 이 대화 상자를 사용하여 컴퓨터의 로컬 위치에 있는 시각화 템플리트, 스타일시트 및 맵 파일에 대해 가져오기, 내보내기, 이름 바꾸기 및 삭제를 수행할 수 있습니다.

템플리트, 스타일시트 또는 맵을 관리하는 대화 상자 중 하나에서 **관리...**를 클릭하십시오.

템플리트, 스타일시트 또는 맵 관리 대화 상자

템플리트 탭은 모든 로컬 템플리트를 나열합니다. 스타일시트 탭은 로컬 스타일시트를 나열하고 표본 데이터가 포함된 표본 시각화를 표시합니다. 스타일시트 중 하나를 선택하여 해당 스타일을 시각화 예에 적용할 수 있습니다. 자세한 정보는 스타일시트 적용 주제를 참조하십시오. 맵 탭은 모든 로컬 맵 파일을 나열합니다. 이 탭은 또한 맵의 미리보기, 주석(맵을 작성할 때 주석을 제공한 경우) 및 표본 값이 포함된 맵 키도 표시합니다.

현재 활성화된 탭에서 작동하는 단추는 다음과 같습니다.

가져오기. 파일 시스템에서 시각화 템플리트, 스타일시트 또는 맵 파일을 가져옵니다. 템플리트, 스타일시트 또는 맵 파일을 가져오면 IBM® SPSS® 애플리케이션에서 이러한 파일을 사용할 수 있습니다. 다른 사용자가 템플리트, 스타일시트 또는 맵 파일을 보내온 경우에는 파일을 애플리케이션으로 가져온 후에 사용합니다.

내보내기. 파일 시스템으로 시각화 템플리트, 스타일시트 또는 맵 파일을 내보냅니다. 다른 사용자에게 템플리트, 스타일시트 또는 맵 파일을 보내려면 해당 파일을 내보내십시오.

이름 바꾸기. 선택한 시각화 템플리트, 스타일시트 또는 맵 파일의 이름을 바꿉니다. 이름을 이미 사용되는 이름으로 변경할 수 없습니다.

맵 키 내보내기. 맵 키를 CSV(쉼표로 구분된 값) 파일로 내보냅니다. 이 단추는 맵 탭에서만 사용됩니다.

삭제. 선택한 시각화 템플릿, 스타일시트 또는 맵 파일을 삭제합니다. Ctrl-클릭을 사용하여 여러 템플릿, 스타일시트 또는 맵 파일을 선택할 수 있습니다. 삭제에 대한 실행 취소 조치가 없으므로 주의해서 사용하십시오.

(3) 맵 형태 파일 변환 및 배포

그래프보드 템플릿 선택기에서는 시각화 템플릿과 SMZ 파일을 조합하여 맵 시각화를 작성할 수 있습니다. SMZ 파일은 맵을 그리기 위한 지리적 정보(예: 국경)를 포함하는 점에서 ESRI 형태 파일(SHP 파일 형식)과 유사하지만 맵 시각화를 위해 최적화되어 있습니다. 그래프보드 템플릿 선택기는 선택한 수의 SMZ 파일과 함께 사전 설치되어 있습니다. 맵 시각화에 사용하려는 기존 ESRI 형태 파일이 있는 경우, 먼저 맵 변환 유틸리티를 사용하여 형태 파일을 SMZ 파일로 변환해야 합니다. 맵 변환 유틸리티는 점, 폴리라인 또는 단일 레이어를 포함한 다각형(형태 유형 1, 3 및 5) ESRI 형태 파일을 지원합니다.

맵 변환 유틸리티에서는 ESRI 형태 파일을 변환할 수 있을 뿐만 아니라 맵의 세부사항 수준을 수정하고 지형 레이블을 변경하며 지형을 합치고 지형을 이동시킬 수 있습니다. 기존 SMZ 파일(사전 설치된 SMZ 파일 포함)을 수정하는 데에도 맵 변환 유틸리티를 사용할 수 있습니다.

사전 설치된 SMZ 파일 편집

1. 관리 시스템에서 SMZ 파일을 내보내십시오. 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리 주제를 참조하십시오.
2. 맵 변환 유틸리티를 사용하여 내보낸 SMZ 파일을 열고 편집하십시오. 파일을 다른 이름으로 저장하는 것이 좋습니다. 자세한 정보는 맵 변환 유틸리티 사용 주제를 참조하십시오.
3. 수정된 SMZ 파일을 관리 시스템으로 가져오십시오. 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리 주제를 참조하십시오.

맵 파일에 대한 추가 자원

맵핑 요구사항을 지원하는 데 사용할 수 있는 SHP 파일 형식의 지리 공간적 데이터는 다양한 개인용 및 공용 소스로부터 얻을 수 있습니다. 무료 데이터를 찾는 경우에는 지역 정부 웹사이트를 확인하십시오. 이 제품에 포함된 다수의 템플릿은 GeoCommons(<http://www.geocommons.com>) 및 미국 통계국(<http://www.census.gov>)에서 제공하는 공개적으로 사용 가능한 데이터를 기반으로 합니다.

중요 주의사항: 비IBM 제품에 관한 정보는 해당 제품의 공급업체, 공개 자료 또는 기타 범용 소스로부터 얻은 것입니다. IBM에서는 이러한 제품들을 테스트하지 않았으므로, 비IBM 제품과 관련된 성능의 정확성, 호환성 또는 기타 청구에 대해서는 확신할 수 없습니다. 비IBM 제품의 성능에 대한 의문사항은 해당 제품의 공급업체에 문의하십시오. 이 정보에서 언급되는 비IBM의 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 이 IBM 프로그램과 함께 제공되는 고지 파일에 표시하지 않는 한, 해당 웹 사이트의 자료는 본 IBM 프로그램 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

① 맵의 핵심 개념

형태 파일과 관련된 몇 가지 핵심 개념을 이해하면 맵 변환 유틸리티를 효과적으로 사용하는 데 도움이 됩니다.

형태 파일은 맵을 그리기 위한 지리적 정보를 제공합니다. 맵 변환 유틸리티는 다음의 세 가지 유형의 형태 파일을 지원합니다.

- **점.** 이 형태 파일은 점의 위치(예: 도시)를 식별합니다.
- **폴리라인.** 이 형태 파일은 경로 및 경로의 위치(예: 강)를 식별합니다.
- **다각형.** 이 형태 파일은 경계가 있는 지역 및 이러한 지역의 위치(예: 국가)를 식별합니다.

대부분 다각형 형태 파일을 가장 많이 사용하게 됩니다. 코로플레스 맵은 다각형 형태 파일에서 작성됩니다. 코로플레스 맵은 색상을 사용하여 개별 다각형(지역) 내의 값을 나타냅니다. 점 및 폴리라인 형태 파일은 일반적으로 다각형 형태 파일 위에 오버레이됩니다. 미국 주의 다각형 형태 파일에 오버레이된 미국 도시의 점 형태 파일이 한 예입니다.

형태 파일은 **지형**으로 구성됩니다. 지형은 개별 지리적 엔티티입니다. 예를 들어, 지형은 국가, 주, 도시 등일 수 있습니다. 형태 파일은 지형에 대한 데이터도 포함합니다. 이러한 데이터는 속성에 저장됩니다. 속성은 데이터 파일의 필드 또는 변수와 유사합니다. 지형의 **맵 키**인 속성이 하나 이상 있습니다. 맵 키는 국가 이름 또는 주 이름과 같은 레이블일 수 있습니다. 맵 키는 맵 시각화를 작성하기 위해 데이터 파일의 변수/필드에 연결하는 것입니다.

SMZ 파일에서는 키 속성만 유지할 수 있습니다. 맵 변환 유틸리티는 추가 속성 저장을 지원하지 않습니다. 따라서 서로 다른 수준에서 통합하려면 여러 개의 SMZ 파일을 작성해야 합니다. 예를 들어, 미국 주와 지역을 통합하려면 주를 식별하는 키가 있는 SMZ 파일과 지역을 식별하는 키가 있는 SMZ 파일이 개별적으로 필요합니다.

② 맵 변환 유틸리티 사용

맵 변환 유틸리티 시작 방법

메뉴에서 다음을 선택하십시오.

도구 > 맵 변환 유틸리티

맵 변환 유틸리티에는 네 개의 주 화면(단계)이 있습니다. 단계 중 하나에는 맵 파일 편집과 관련된 보다 세부적인 제어를 위한 하위 단계도 포함됩니다.

가. 1단계 - 대상 및 소스 파일 선택

먼저 변환되는 맵 파일의 소스 맵 파일과 대상을 선택해야 합니다. 형태 파일에는 *.shp* 및 *.dbf* 파일이 모두 필요합니다.

변환을 위해 *.shp*(ESRI) 또는 *.smz* 파일 선택. 찾아보기로 컴퓨터에 있는 기존 맵 파일을 찾으십시오. 이 파일은 SMZ 파일로 변환하고 저장할 파일입니다. 형태 파일을 위한 *.dbf* 파일은 반드시 *.shp* 파일과 일치하는 기존 파일 이름과 동일한 위치에 저장해야 합니다. *.dbf* 파일이 필요한 이유는 이 파일에 *.shp* 파일의 속성 정보가 있기 때문입니다.

변환되는 맵 파일의 대상 및 파일 이름 설정. 원래 맵 소스에서 작성할 SMZ 파일의 경로 및 파일 이름을 입력하십시오.

- **템플릿 선택기로 가져오기.** 파일 시스템에 파일을 저장할 수 있을 뿐만 아니라 선택적으로 템플릿 선택기의 관리 목록에 맵을 추가할 수 있습니다. 이 옵션을 선택하면 컴퓨터에 설치된 IBM® SPSS® 제품의 템플릿 선택기에서 자동으로 해당 맵을 사용할 수 있습니다. 지금 템플릿 선택기로 가져오지 않는 경우 나중에 수동으로 가져와야 합니다. 템플릿 선택기 관리 시스템에 맵을 가져오는 방법에 대한 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리의 내용을 참조하십시오.

나. 2단계 - 맵 키 선택

이제 SMZ 파일에 포함시킬 맵 키를 선택합니다. 그런 다음 맵 렌더링에 영향을 주는 일부 옵션을 변경할 수 있습니다. 맵 변환 유틸리티에서의 이후 단계에는 맵의 미리보기가 포함됩니다. 선택하는 렌더링 옵션은 맵 미리보기를 생성하는 데 사용됩니다.

기본 맵 키 선택. 맵의 지형을 식별하고 레이블을 지정하는 기본 키인 속성을 선택하십시오. 예를 들어, 세계 맵의 기본 키는 국가 이름을 식별하는 속성일 수 있습니다. 기본 키는 또한 데이

터를 맵 지형에 연결하므로 선택하는 속성의 값(레이블)이 데이터의 값과 일치해야 합니다. 속성을 선택하면 레이블 예가 표시됩니다. 이러한 레이블을 변경해야 하는 경우 나중 단계에서 이를 수행할 수 있습니다.

포함시킬 추가 키 선택. 기본 맵 키 이외에, 생성된 SMZ 파일에 포함시킬 다른 키 속성을 선택하십시오. 예를 들어, 일부 속성에 변환된 레이블이 있을 수 있습니다. 다른 언어로 코딩된 데이터를 예상하는 경우 이러한 속성을 유지하려 할 수 있습니다. 기본 키와 동일한 지형을 나타내는 추가 키만 선택할 수 있습니다. 예를 들어, 기본 키가 미국 주의 전체 이름인 경우 미국 주를 나타내는 대체 키(예: 주 약어)만 선택할 수 있습니다.

자동으로 맵 평활화. 다각형을 포함하는 형태 파일에는 일반적으로 통계 맵 시각화를 위한 너무 많은 데이터 점과 너무 많은 세부사항이 있습니다. 세부사항이 지나치게 많으면 산만하고 성능에 부정적인 영향을 미칠 수 있습니다. 세부사항 수준을 낮추고 평활화로 맵을 일반화할 수 있습니다. 이렇게 하면 맵이 더 단정해 보이고 더 빨리 렌더링됩니다. 맵이 자동으로 평활화되는 경우 최대 각은 15도이고 유지할 백분율은 99입니다. 이러한 설정에 대한 정보는 맵 평활화의 내용을 참조하십시오. 나중에 다른 단계에서 평활화를 추가로 적용할 기회가 있습니다.

같은 지형에서 맞닿은 다각형 사이의 경계 제거. 일부 지형은 주 관심 지형 내부에 경계가 있는 하위 지형을 포함할 수 있습니다. 예를 들어, 세계 대륙 맵에는 각 대륙에 포함된 국가의 내부 경계가 포함될 수 있습니다. 이 옵션을 선택하면 맵에 내부 경계가 표시되지 않습니다. 세계 대륙 맵 예에서 이 옵션을 선택하면 대륙 경계는 유지되지만 국가 경계가 제거됩니다.

다. 3단계 - 맵 편집

이제 맵에 대한 기본 옵션을 지정했으므로 보다 구체적인 옵션을 편집할 수 있습니다. 이러한 수정은 선택사항입니다. 맵 변환 유틸리티의 이 단계에서는 연관된 작업의 수행 과정을 안내하고 변경사항을 확인할 수 있도록 맵의 미리보기를 표시합니다. 형태 파일 유형(점, 폴리라인 또는 다각형) 및 좌표계에 따라 일부 작업이 사용 불가능할 수도 있습니다.

모든 작업은 맵 변환 유틸리티의 왼쪽에 다음과 같은 공통 제어가 있습니다.

맵에 레이블 표시. 기본적으로 미리보기에는 지형 레이블이 표시되지 않습니다. 이러한 레이블을 표시하도록 선택할 수 있습니다. 지형 레이블은 지형을 식별하는 데 도움이 될 수 있지만 미리보기 맵에서 직접 선택하는 데 방해가 될 수 있습니다. 필요한 경우, 예를 들어, 지형 레이블을 편집하는 경우에는 이 옵션을 끄십시오.

맵 미리보기 채색. 기본적으로 맵 미리보기는 영역을 단색으로 표시합니다. 모든 지형의 색상이 동일합니다. 개별 맵 지형에 다양한 색상이 지정되도록 선택할 수 있습니다. 이 옵션은 맵에서 서로 다른 지형을 구별하는 데 유용할 수 있습니다. 특히 지형을 합칠 때 미리보기에서 새 지형이 어떻게 표시되는지 확인하려는 경우에 유용합니다.

모든 작업은 또한 맵 변환 유틸리티의 오른쪽에 다음과 같은 공통 제어가 있습니다.

실행 취소. 이전 상태로 되돌아가려면 **실행 취소**를 클릭하십시오. 최대 100번의 변경을 실행 취소할 수 있습니다.

ㄱ. 맵 평활화

다각형을 포함하는 형태 파일에는 일반적으로 통계 맵 시각화를 위한 너무 많은 데이터 점과 너무 많은 세부사항이 있습니다. 세부사항이 지나치게 많으면 산만하고 성능에 부정적인 영향을 미칠 수 있습니다. 세부사항 수준을 낮추고 평활화로 맵을 일반화할 수 있습니다. 이렇게 하면 맵이 더 단정해 보이고 더 빨리 렌더링됩니다. 점 및 폴리라인 맵에는 이 옵션을 사용할 수 없습니다.

최대 각. 최대 각은 1과 20 사이의 값이어야 하며 거의 선형인 일련의 점을 평활화하기 위한 허용 오차를 지정합니다. 값이 클수록 선형 평활화에 대한 허용 오차가 커지고 그에 따라 더 많은 점이 삭제되어 좀 더 일반화된 맵이 됩니다. 선형 평활화를 적용하기 위해 맵 변환 유틸리티는 맵에서 세 개의 점으로 구성된 각각의 세트가 이루는 내각을 확인합니다. 180에서 내각을 뺀 값이 지정한 값보다 작으면 맵 변환 유틸리티는 가운데 점을 삭제합니다. 즉, 맵 변환 유틸리티는 세 개의 점으로 형성된 선이 거의 직선인지 여부를 확인합니다. 그러한 경우 맵 변환 유틸리티는 해당 선을 엔드포인트 사이의 직선으로 처리하여 중간 점을 삭제합니다.

유지할 퍼센트. 유지할 백분율은 90과 100 사이의 값이어야 하며 맵을 평활화할 때 유지할 땅 영역의 양을 결정합니다. 이 옵션은 여러 다각형을 포함하는 지형(예: 지형에 섬이 포함된 경우)에만 영향을 줍니다. 지형의 총 영역에서 다각형을 뺀 값이 원래 영역의 지정된 백분율보다 큰 경우 맵 변환 유틸리티는 맵에서 해당 다각형을 삭제합니다. 맵 변환 유틸리티는 지형의 모든 다각형을 제거하지는 않습니다. 즉, 적용되는 평활화 양과 무관하게 지형에는 항상 하나 이상의 다각형이 있습니다.

최대 각과 유지할 백분율을 선택한 후에는 **적용**을 클릭하십시오. 미리보기가 평활화 변경사항으로 업데이트됩니다. 맵을 다시 평활화해야 하는 경우 원하는 평활 수준이 될 때까지 반복하십시오. 평활화에는 한계가 있습니다. 반복해서 평활화하면 맵에 추가 평활화를 적용할 수 없는 지점에 이르게 됩니다.

ㄴ. 지형 레이블 편집

필요에 따라(예: 예상 데이터와 일치시키기 위해) 지형 레이블을 편집하고 맵에서 레이블의 위치를 바꿀 수 있습니다. 레이블을 변경할 필요가 없다고 생각하는 경우에도 맵에서 시각화를 작성하기 전에 레이블을 검토해야 합니다. 미리보기에는 기본적으로 레이블이 표시되지 않으므로 **맵에 레이블 표시**를 선택하여 레이블을 표시할 수도 있습니다.

키. 검토하거나 편집할 지형 레이블이 포함된 키를 선택하십시오.

변수. 이 목록은 선택한 키에 포함된 지형 레이블을 표시합니다. 레이블을 편집하려면 목록에서 레이블을 두 번 클릭하십시오. 맵에 레이블이 표시되는 경우 맵 미리보기에서 직접 지형 레이블을 두 번 클릭할 수도 있습니다. 레이블을 실제 데이터 파일과 비교하려면 **비교**를 클릭하십시오.

X/Y. 이 텍스트 상자는 맵에서 선택한 지형 레이블의 현재 중심점을 나열합니다. 단위는 맵의 좌표에 표시됩니다. 좌표는 로컬 데카르트 좌표(예: 미국 평면 좌표계) 또는 지리적 좌표(X는 경도이고 Y는 위도인 좌표)일 수 있습니다. 레이블의 새 위치에 대한 좌표를 입력하십시오. 레이블이 표시되는 경우 맵에서 레이블을 클릭하여 끌기 조작으로 이동시킬 수 있습니다. 텍스트 상자가 새 위치로 업데이트됩니다.

비교. 특정 키의 지형 레이블과 비교할 데이터 값이 포함된 데이터 파일이 있는 경우 **비교**를 클릭하여 외부 데이터 소스와 비교 대화 상자를 표시하십시오. 이 대화 상자에서 데이터 파일을 열고 해당 값을 맵 키의 지형 레이블에 있는 값과 직접 비교할 수 있습니다.

- **외부 데이터 소스와 비교 대화 상자**

외부 데이터 소스와 비교 대화 상자에서는 탭으로 구분된 값 파일(.txt 확장자를 가짐), 쉼표로 구분된 값 파일(.csv 확장자를 가짐) 또는 IBM® SPSS® Statistics에 맞게 형식화된 데이터 파일(.sav 확장자를 가짐)을 열 수 있습니다. 파일이 열리면 데이터 파일에서 필드를 선택하여 특정 맵 키의 지형 레이블과 비교할 수 있습니다. 그런 다음 맵 파일에서 일치하지 않는 부분을 정정할 수 있습니다.

데이터 파일의 필드. 지형 레이블과 값을 비교할 필드를 선택하십시오. .txt 또는 .csv 파일의 첫 번째 행에 각 필드의 설명 레이블이 있으면 **첫 번째 행을 열 레이블로 사용**을 선택하십시오. 그렇지 않은 경우 각 필드는 데이터 파일에서 해당 위치로 식별됩니다(예: "열 1", "열 2" 등).

비교할 키. 데이터 파일 필드 값과 지형 레이블을 비교할 맵 키를 선택하십시오.

비교. 값을 비교할 준비가 되었을 때 클릭하십시오.

비교 결과. 기본적으로 비교 결과 테이블은 데이터 파일에서 일치하지 않는 필드 값만 나열합니다. 애플리케이션은 주로 삽입되었거나 누락된 공간이 있는지 확인함으로써 관련된 지형 레이블을 찾으려고 합니다. **맵 레이블** 열에서 드롭 다운 목록을 클릭하여 맵 파일의 지형 레이블을 표시된 필드 값과 일치시키십시오. 맵 파일에 일치하는 지형 레이블이 없으면 **일치하지 않은 상태**로 두기를 선택하십시오. 이미 지형 레이블과 일치하는 필드 값을 포함하여 모든 필드 값을 보려면 **일치하지 않은 케이스만 표시**를 선택 취소하십시오. 하나 이상의 일치를 대체하려는 경우 이를 수행할 수 있습니다.

각 지형을 한 번만 사용하여 필드 값에 일치시킬 수 있습니다. 여러 지형을 하나의 필드 값에 일치시키려는 경우 지형을 합친 후 합쳐진 새 지형을 필드 값에 일치시킬 수 있습니다. 지형 합치기에 대한 자세한 정보는 지형 합치기의 내용을 참조하십시오.

ㄷ. 지형 합치기

지형 합치기는 맵에서 더 큰 지역을 작성하는 데 유용합니다. 예를 들어, 주 맵을 변환하는 경우 주(이 예제의 지형)를 더 큰 북부, 남부, 동부 및 서부 지역으로 합칠 수 있습니다.

키. 합칠 지형을 식별하는 데 도움이 될 지형 레이블이 포함된 맵 키를 선택하십시오.

변수. 합칠 첫 번째 지형을 클릭하십시오. Ctrl-클릭을 사용하여 합칠 다른 지형을 선택하십시오. 지형은 맵 미리보기에서도 선택됩니다. 목록에서 지형을 선택할 수 있을 뿐만 아니라 맵 미리보기에서 직접 지형을 클릭 및 Ctrl-클릭할 수 있습니다.

합칠 지형을 선택한 후에는 **합치기**를 클릭하여 합친 지형의 이름 지정 대화 상자를 표시하십시오. 여기서 새 지형에 레이블을 적용할 수 있습니다. 지형을 합친 후 결과가 예상과 일치하는지 확인하기 위해 **맵 미리보기 채색**을 선택할 수 있습니다.

지형을 합친 후 새 지형의 레이블을 이동시킬 수도 있습니다. *지형 레이블 편집* 작업에서 이를 수행할 수 있습니다. 자세한 정보는 지형 레이블 편집의 내용을 참조하십시오.

- **합친 지형의 이름 지정 대화 상자**

합친 지형의 이름 지정 대화 상자에서는 합친 새 지형에 레이블을 지정할 수 있습니다.

레이블 테이블은 맵 파일에 있는 각 키에 대한 정보를 표시하며 이 테이블에서 각 키에 레이블을 지정할 수 있습니다.

새 레이블. 특정 맵 키에 지정할 합친 지형의 새 레이블을 입력하십시오.

키. 새 레이블을 지정할 맵 키입니다.

이전 레이블. 새 지형으로 합칠 지형의 레이블입니다.

맞닿은 다각형 사이의 경계 제거. 합쳐진 지형에서 경계를 제거하려면 이 옵션을 선택하십시오. 예를 들어, 주를 지리적 지역으로 합친 경우 이 옵션은 개별 주의 경계를 제거합니다.

ㄹ. 지형 이동

맵에서 지형을 이동시킬 수 있습니다. 이는 본토와 딸린 섬처럼 여러 지형을 한 데 모으려는 경우에 유용할 수 있습니다.

키. 이동시킬 지형을 식별하는 데 도움이 될 지형 레이블이 포함된 맵 키를 선택하십시오.

변수. 이동시킬 지형을 클릭하십시오. 지형은 맵 미리보기에서 선택됩니다. 맵 미리보기에서 직접 지형을 클릭할 수도 있습니다.

X/Y. 이 텍스트 상자는 맵에서 지형의 현재 중심점을 나열합니다. 단위는 맵의 좌표에 표시됩니다. 좌표는 로컬 데카르트 좌표(예: 미국 평면 좌표계) 또는 지리적 좌표(X는 경도이고 Y는 위도인 좌표)일 수 있습니다. 지형의 새 위치에 대한 좌표를 입력하십시오. 맵에서 지형을 클릭하여 끌기 조작으로 이동시킬 수도 있습니다. 텍스트 상자가 새 위치로 업데이트됩니다.

ㄻ. 지형 삭제

맵에서 원하지 않는 지형을 삭제할 수 있습니다. 이는 맵 시각화에서 관련이 없는 지형을 삭제하여 복잡성을 제거하려는 경우에 유용할 수 있습니다.

키. 삭제할 지형을 식별하는 데 도움이 될 지형 레이블이 포함된 맵 키를 선택하십시오.

변수. 삭제할 지형을 클릭하십시오. 동시에 여러 지형을 삭제하려면 Ctrl-클릭을 사용하여 지형을 추가로 선택하십시오. 지형은 맵 미리보기에서도 선택됩니다. 목록에서 지형을 선택할 수 있을 뿐만 아니라 맵 미리보기에서 직접 지형을 클릭 및 Ctrl-클릭할 수 있습니다.

ㄼ. 개별 요소 삭제

전체 지형을 삭제할 수 있을 뿐만 아니라 지형을 구성하는 일부 개별 요소(예: 호수 및 작은 섬)를 삭제할 수도 있습니다. 점 맵에는 이 옵션을 사용할 수 없습니다.

요소. 삭제할 요소를 클릭하십시오. 동시에 여러 요소를 삭제하려면 Ctrl-클릭을 사용하여 요소를 추가로 선택하십시오. 요소는 맵 미리보기에서도 선택됩니다. 목록에서 요소를 선택할 수 있을 뿐만 아니라 맵 미리보기에서 직접 요소를 클릭 및 Ctrl-클릭할 수 있습니다. 요소 이름 목록은 설명적이지 않으므로(각 요소는 지형 내에서 번호로 지정됨) 맵 미리보기에서 원하는 요소를 선택했는지 확인해야 합니다.

스. 투영법 설정

맵 투영법은 3차원의 지구를 2차원으로 나타내는 방법을 지정합니다. 모든 투영법은 왜곡을 야기시킵니다. 그러나 구형 맵을 보는지 또는 좀 더 국지적인 맵을 보는지에 따라 더 적합한 투영법이 있습니다. 또한 일부 투영법은 원래 지형의 형태를 유지합니다. 형태를 유지하는 투영법은 정각 투영법입니다. 이 옵션은 지리적 좌표(경도 및 위도)가 있는 맵에만 사용할 수 있습니다.

맵 변환 유틸리티의 다른 옵션과 달리 투영법은 맵 시각화 작성 후에 변경할 수 있습니다.

투영법. 맵 투영법을 선택하십시오. 구형 또는 반구형 맵을 작성하는 경우에는 *국지*, *메르카토르* 또는 *빈켈 트리펠* 투영법을 사용하십시오. 더 작은 영역에는 *국지*, *람베르트 정각원추* 또는 *횡축 메르카토르* 투영법을 사용하십시오. 모든 투영법은 데이터에 WGS83 타원체를 사용합니다.

- **국지** 투영법은 항상 맵이 국지 좌표계(예: 미국 평면 좌표계)로 작성된 경우에 사용됩니다. 이러한 좌표계는 지리적 좌표(경도 및 위도)가 아닌 데카르트 좌표에 의해 정의됩니다. 국지 투영법에서는 데카르트 좌표계의 수평선과 수직선의 간격이 동일합니다. 국지 투영법은 정각 투영법이 아닙니다.
- **메르카토르** 투영법은 구형 맵을 위한 정각 투영법입니다. 수평선과 수직선이 일직선이며 항상 서로 수직을 이룹니다. 메르카토르 투영법은 북극과 남극에 접근함에 따라 무한대로 확장되므로 북극 또는 남극을 포함하는 맵에는 사용할 수 없습니다. 맵이 이러한 한계에 접근할 때 가장 크게 왜곡됩니다.
- **빈켈 트리펠** 투영법은 구형 맵을 위한 비정각 투영법입니다. 정각 투영법은 아니지만 형태와 크기 사이에 적절한 균형을 제공합니다. 적도와 본초자오선을 제외한 모든 선이 곡선입니다. 구형 맵에 북극 또는 남극이 포함되는 경우 이 투영법을 선택하는 것이 좋습니다.
- 이름에서 알 수 있듯이 **람베르트 정각원추** 투영법은 정각 투영법이며 북쪽과 남쪽에 비해 동쪽과 서쪽이 더 긴 대륙 또는 그보다 작은 육지의 맵에 사용됩니다.
- **횡축 메르카토르**는 대륙 또는 그보다 작은 육지의 맵을 위한 또하나의 정각 투영법입니다. 동쪽과 서쪽에 비해 북쪽과 남쪽이 더 긴 육지에 이 투영법을 사용하십시오.

라. 4단계 - 완료

이 시점에서 맵 파일을 설명하는 주석을 추가하고 맵 키에서 표본 데이터 파일을 작성할 수 있습니다.

맵 키. 맵 파일에 여러 키가 있으면 미리보기에 지형 레이블을 표시할 맵 키를 선택하십시오. 맵에서 데이터 파일을 작성하는 경우 이러한 레이블이 데이터 값에 사용됩니다.

주석. 맵을 설명하거나 사용자와 관련이 있을 수 있는 추가 정보(예: 원래 형태 파일의 소스)를 제공하는 주석을 입력하십시오. 주석은 그래프보드 템플릿 선택기의 관리 시스템에 표시됩니다.

지형 레이블에서 데이터 세트 작성. 표시된 지형 레이블에서 데이터 파일을 작성하려는 경우 이 옵션을 선택하십시오. **찾아보기...**를 클릭하면 위치 및 파일 이름을 지정할 수 있습니다. **.txt** 확장자를 추가하면 파일이 탭으로 구분된 값 파일로 저장됩니다. **.csv** 확장자를 추가하면 파일이 쉼표로 구분된 값 파일로 저장됩니다. **.sav** 확장자를 추가하면 파일이 IBM® SPSS® Statistics 형식으로 저장됩니다. 확장자를 지정하지 않으면 SAV가 기본값입니다.

③ 맵 파일 배포

맵 변환 유틸리티의 첫 번째 단계에서 변환된 SMZ 파일을 저장할 위치를 선택했습니다. 또한 그래프보드 템플릿 선택기의 관리 시스템에 맵을 추가하도록 선택했을 수도 있습니다. 관리 시스템에 저장하도록 선택한 경우, 동일한 컴퓨터에서 실행하는 모든 IBM® SPSS® 제품에서 해당 맵을 사용할 수 있습니다.

맵을 다른 사용자에게 배포하려면 맵을 배포할 사용자에게 SMZ를 보내야 합니다. 그러면 해당 사용자가 관리 시스템을 사용하여 맵을 가져올 수 있습니다. 1단계에서 위치를 지정한 파일은 보내기만 하면 됩니다. 관리 시스템에 있는 파일을 보내려면 먼저 파일을 내보내야 합니다.

1. 템플릿 선택기에서 관리...를 클릭하십시오.
2. 맵 탭을 클릭하십시오.
3. 배포할 맵을 선택하십시오.
4. **내보내기...**를 클릭하고 파일을 저장할 위치를 선택하십시오.

이제 실제 맵 파일을 다른 사용자에게 보낼 수 있습니다. 사용자는 이 프로세스를 역으로 수행하여 맵을 관리 시스템으로 가져와야 합니다.

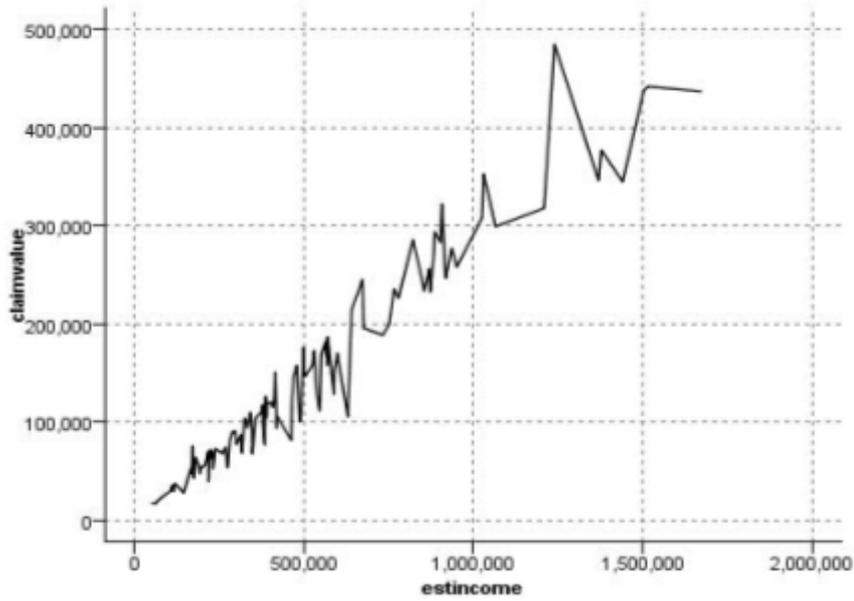
(4) Plot 노드

Plot 노드는 수치 필드 사이의 관계를 보여줍니다. 포인트(산점도) 또는 선을 사용하여 도표를 작성할 수 있습니다. 대화 상자에서 X 모드를 지정하여 세 가지 유형의 선 도표를 작성할 수 있습니다.

X 모드 = 정렬

X 모드를 **정렬**로 설정하면 데이터가 x 축에 구성된 필드에 대한 값으로 정렬됩니다. 그러면 그래프의 왼쪽에서 오른쪽으로 진행되는 단일 선이 생성됩니다. 명목 필드를 오버레이로 사용하면 그래프에서 왼쪽에서 오른쪽으로 진행되는 다양한 색상의 다중 선이 생성됩니다.

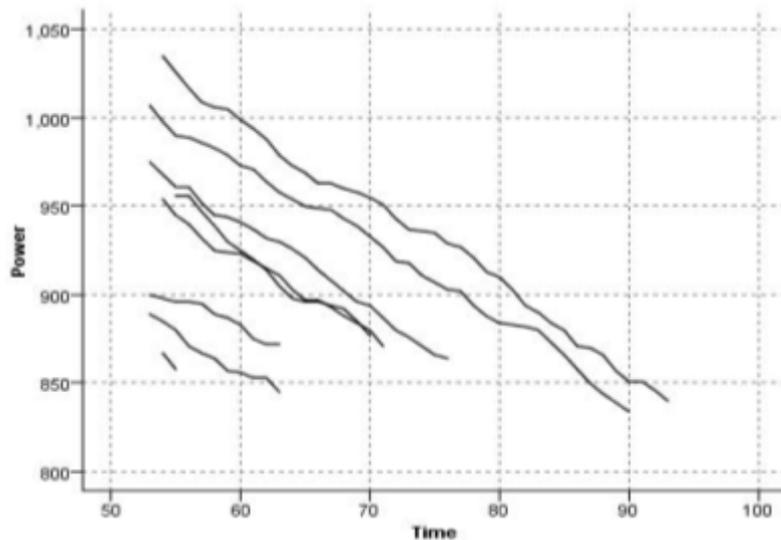
그림 1. X 모드가 정렬로 설정된 선 도표



X 모드 = 오버레이

X 모드를 **오버레이**로 설정하면 동일한 그래프에서 다중 선 도표가 작성됩니다. x축의 값이 증가하는 한 데이터가 단일 선에 구성되며 데이터가 오버레이 도표에 대해 정렬되지 않습니다. 값이 감소하면 새 선이 시작됩니다. 예를 들어, x가 0에서 100으로 이동하면 y값이 단일 선에 구성될 것입니다. x가 100 아래가 되면 첫 번째 선 외에 새 선이 구성될 것입니다. 완료된 도표는 일련의 y값을 비교하는 데 유용한 수많은 도표를 갖게 될 것입니다. 이 유형의 도표는 정기적인 시간 구성요소(연속되는 24시간 동안의 전기 요구량 등)가 있는 데이터에 유용합니다.

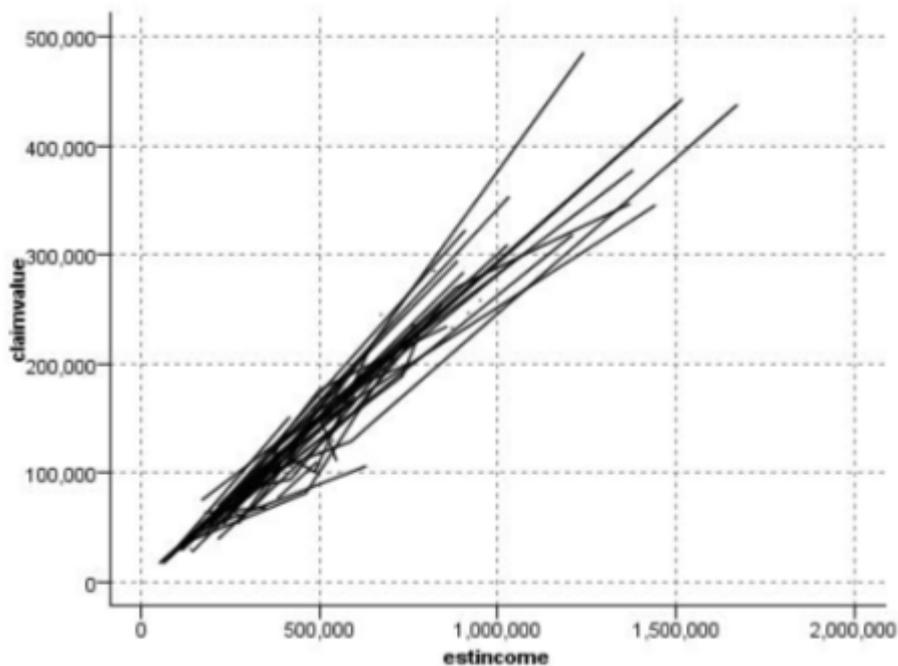
그림 2. X 모드가 오버레이로 설정된 선 도표



X 모드 = 읽은 대로

X 모드를 읽은 대로 도표로 설정하면 x 및 y 값을 데이터 소스에서 읽은 대로 구성합니다. 이 옵션은 사용자가 데이터의 순서에 따른 추세 또는 패턴에 관심이 있는 경우에 시계열 구성요소가 있는 데이터에 유용합니다. 이 유형의 도표를 작성하기 전에 데이터를 정렬해야 합니다. 또한 패턴이 정렬에 의존하는 정도를 판별하기 위해 X 모드가 정렬 및 읽은 대로로 설정된 두 개의 유사한 도표를 비교하는 데 유용합니다.

그림 3. 이전에 정렬로 표시된 선 도표, X 모드를 읽은 대로로 설정하고 다시 실행



또한 그래프보드 노드를 사용하는 방법으로도 산점도 및 선 도표를 생성할 수 있습니다. 하지만 이 노드에서 더 많은 옵션을 선택할 수 있습니다. 자세한 정보는 사용 가능한 내장 그래프보드 시각화 유형의 내용을 참조하십시오.

① 구성 노드 탭

도표는 Y 필드의 값 대 X 필드의 값을 표시합니다. 종종 이러한 필드는 각각 종속변수 및 독립 변수에 해당됩니다.

X 필드. 목록에서 수평 x 축에 표시할 필드를 선택하십시오.

Y 필드. 목록에서 수직 y 축으로 표시할 필드를 선택합니다.

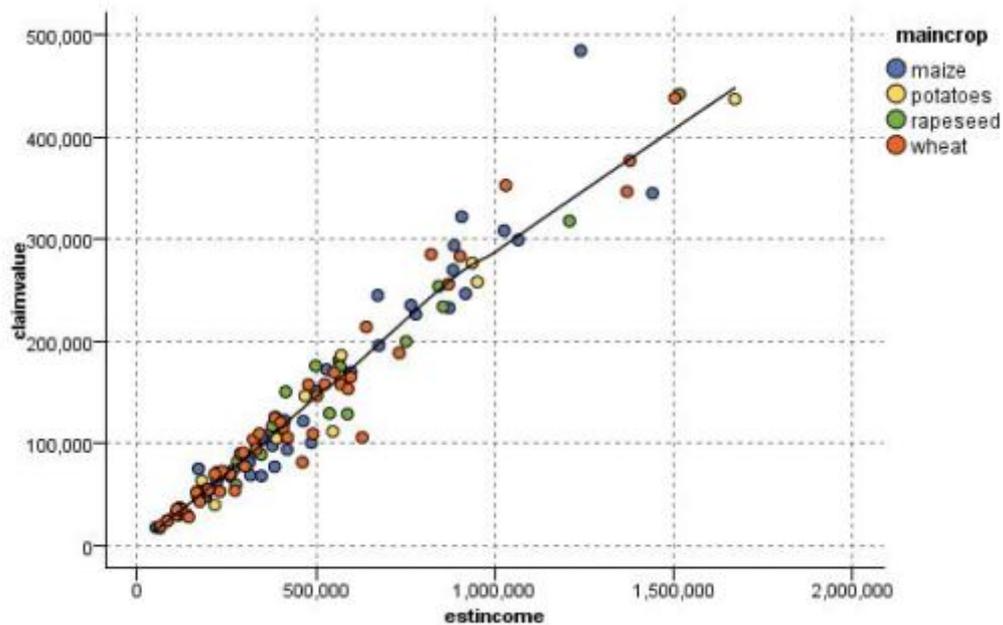
Z 필드. 3D 차트 단추를 클릭하면 목록에서 z 축을 표시할 필드를 선택할 수 있습니다.

오버레이. 여러 가지 방식으로 데이터 값에 대한 범주를 표시할 수 있습니다. 예를 들어, *maincrop*를 색상 오버레이로 사용하여 클레임 지원자가 키운 주요 작물에 대한 *estincome* 및 *claimvalue* 값을 표시할 수 있습니다. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

오버레이 유형. 오버레이 함수나 다듬기가 표시되는지 여부를 지정합니다. 다듬기 및 오버레이 함수는 항상 y 함수로 계산됩니다.

- **없음.** 오버레이가 표시되지 않습니다.
- **다듬기.** LOESS(locally weighted iterative robust least squares regression)를 사용하여 계산된 다듬은 회귀선 적합을 표시합니다. 이 방법은 각각 도표 내의 작은 영역에 집중하여 일련의 회귀분석을 효과적으로 계산합니다. 이로 인해 평활 곡선을 작성하기 위해 결합된 일련의 "로컬" 회귀선이 생성됩니다.

그림 1. LOESS 다듬기 오버레이로 구성



- **함수.** 실제 값과 비교할 알려진 함수를 지정하려면 선택하십시오. 예를 들어, 실제 값과 예측 값을 비교하려면 $y = x$ 함수를 오버레이로 도표화하십시오. 텍스트 상자에서 $y =$ 에 대한 함수를 지정하십시오. 기본 함수는 $y = x$ 이지만 x 축에서 모든 종류의 함수(이차 함수 또는 임의의 표현식 등)를 지정할 수 있습니다.

참고: 오버레이 함수는 패널 또는 애니메이션 그래프에 대해 사용할 수 없습니다.

일단 도표에 대한 옵션을 설정한 후에는 **실행**을 클릭하여 대화 상자에서 직접 도표를 실행할 수 있습니다. 단, 구간화, X 모드 및 유형 등의 추가 지정 사항에 대한 옵션을 사용해야 하는 경우도 있습니다.

② 도표 옵션 탭

스타일. 도표 스타일에 대해 **점** 또는 **선**을 선택하십시오. **선**을 선택하면 **X 모드** 제어가 활성화됩니다. **점**을 선택하면 더하기 기호(+)가 기본 점 모양으로 사용됩니다. 그래프가 작성되고 나면 점 모양을 변경하고 해당 크기를 변경할 수 있습니다.

X 모드. 선 도표의 경우 X 모드를 선택하여 선 도표의 스타일을 정의해야 합니다. **정렬**, **오버레이** 또는 **읽은 대로**를 선택하십시오. 자세한 정보는 Plot 노드 주제를 참조하십시오. **오버레이** 또는 **읽은 대로**의 경우에는 처음 *n*개 레코드의 표본을 추출하는 데 사용되는 최대 데이터 세트 크기를 지정해야 합니다. 그렇지 않으면 기본값인 2,000개의 레코드가 사용됩니다.

자동 X 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

자동 Y 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

자동 Z 범위. 도표 탭에서 3차원 그래프가 지정되는 경우에만 사용됩니다. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브세트를 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

지터. **변동**으로도 알려져 있는 지터는 다수의 값이 반복되는 데이터 세트의 포인트 도표에 유용합니다. 더 명확한 값 분포를 보기 위해 지터를 사용하여 실제 값 주위에 무작위로 점을 분포시킬 수 있습니다.

IBM® SPSS® Modeler의 이전 버전 사용자에게 대한 참고: 도표에서 사용되는 지터 값은 이 IBM SPSS Modeler 릴리스에서 다른 메트릭을 사용합니다. 이전 버전에서는 값이 실제 숫자였지만 이제는 프레임 크기의 비율입니다. 이는 이전 스트림의 변동 값이 지나치게 커질 수 있음을 의미합니다. 이 릴리스의 경우 0(영)이 아닌 변동 값은 값 0.2로 변환됩니다.

도표화할 최대 레코드 수. 큰 데이터 세트를 도표화하는 방법을 지정하십시오. 최대 데이터 세트 크기를 지정하거나 기본값인 2,000개의 레코드를 사용할 수 있습니다. **구간** 또는 **표본** 옵션을 선택하면 큰 데이터 세트에 대해 성능이 개선됩니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 포인트를 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격하게 저하될 수 있다는 점에 유의해야 합니다.

참고: X 모드가 **오버레이** 또는 **읽은 대로**로 설정된 경우 이 옵션은 사용 안함으로 설정되고 처음 *n*개 레코드만 사용됩니다.

- **구간.** 데이터 세트에 지정된 수의 레코드보다 많은 레코드가 포함되어 있는 경우 구간화를 사용하여 설정하려면 선택하십시오. 구간화는 실제로 도표화하기 전에 그래프를 세분화된 눈금으로 나누고 각각의 눈금 셀에 표시되는 점의 수를 계수합니다. 최종 그래프에서는 구간 중심값(구간에 있는 모든 점 위치의 평균)에서 셀당 하나의 점이 도표화됩니다. 도표화된 기호의 크기는 해당 영역에 있는 점의 수를 표시합니다(크기를 오버레이로 사용하지 않은 경우). 중심값 및 크기를 사용하여 점의 수를 나타내면 밀집된 영역(구별되지 않는 색상 덩어리)에서 도표가 겹치는 걸 방지하고 기호 아티팩트(밀도의 인위적인 패턴)를 줄이므로 구간화된 도표는 큰 데이터 세트를 나타내는 우수한 방법이 됩니다. 기호 아티팩트는 원시 데이터에 없는 밀집된 영역을 생성하는 방식으로 특정 기호(특히, 더하기 기호 [+])가 충돌할 때 발생합니다.
- **표본.** 텍스트 필드에서 입력한 레코드 수까지 무작위로 데이터의 표본을 추출하려면 선택하십시오. 기본값은 2,000입니다.

③ 도표 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 x 축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

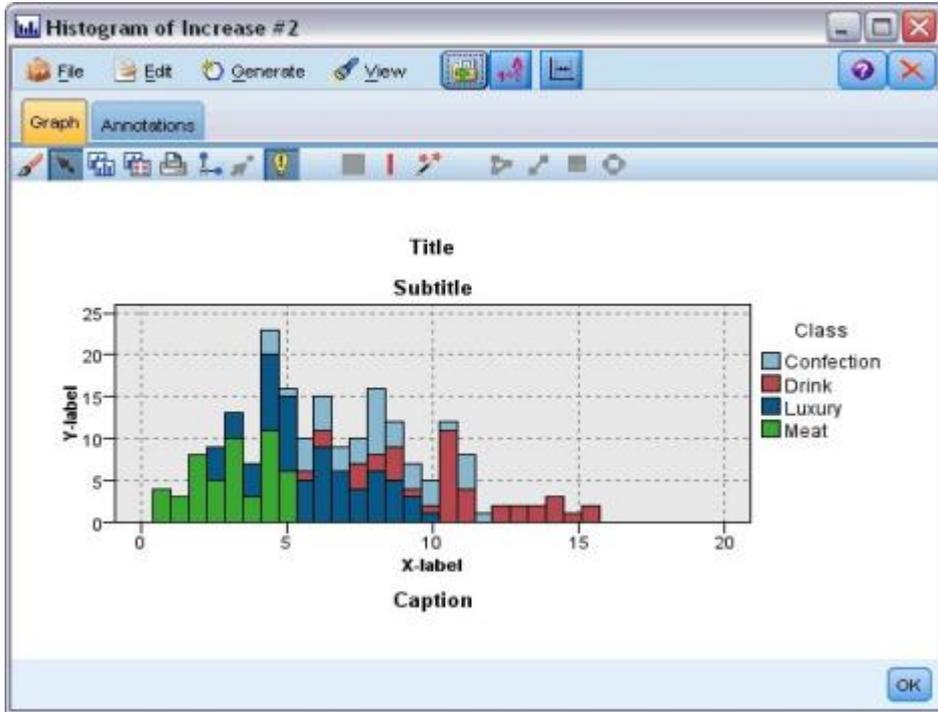
Y 레이블. 자동으로 생성된 y 축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Z 레이블. 3-D 그래프에만 사용할 수 있습니다(자동으로 생성된 z 축 레이블의 경우 또는 **사용자 정의**를 선택하여 사용자 정의 레이블을 지정하는 경우는 제외).

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. 일부 그래프에서는 이 모든 옵션이 사용되지 않습니다.

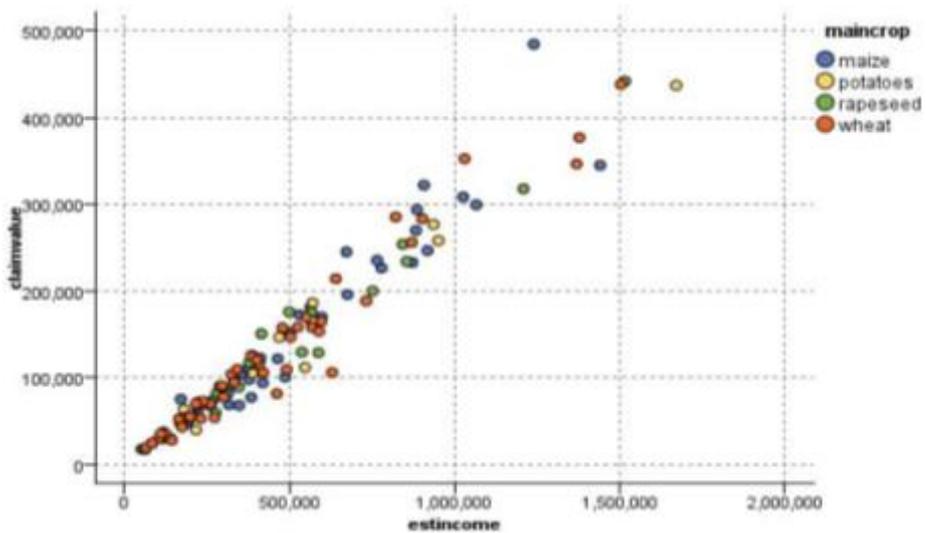
그림 1. 다양한 그래프 모양 옵션의 위치



④ 도표 그래프 사용

도표 및 다중 도표는 본질적으로 Y 에 대한 X 의 도표입니다. 예를 들어, 농업 허가 신청에서 잠재적 사기를 조사하는 경우에는 신경망에 의해 추정된 수입에 대해 해당 신청에서 청구된 수입을 도표화할 수 있습니다. 작물 유형 등의 오버레이를 사용하면 청구(값 또는 번호)와 작물 유형 사이에 관계가 있는 여부가 표시됩니다.

그림 1. 기본 작물 유형을 오버레이로 가진 추정된 수입과 청구 값 간 관계의 도표



도표, 다중 도표 및 평가 차트는 X 에 대한 Y 의 2차원 표시이므로 영역을 정의하거나 요소를 표시하거나 밴드를 그려서 이들과 쉽게 상호작용할 수 있습니다. 해당 영역, 밴드 또는 요소로 표시된 데이터에 대한 노트도 생성할 수 있습니다. 자세한 정보는 그래프 탐색의 내용을 참조하십시오.

(5) 다중 도표 노트

다중 도표는 단일 X 필드 위에 다중 Y 필드를 표시하는 특수 유형의 도표입니다. Y 필드는 색상 지정된 선으로 도표화되며 각각 스타일이 선으로 설정되고 X 모드가 정렬로 설정된 구성 노트와 동등합니다. 다중 도표는 시간 시퀀스 데이터를 가지고 있을 때 시간 경과에 따른 여러 변수의 변동을 탐색하려는 경우에 유용합니다.

① 다중 도표 도표 탭

다중 도표는 단일 X 필드 위에 다중 Y 필드를 표시하는 특수한 유형의 도표입니다.

X 필드. 목록에서 수평 x 축에 표시할 필드를 선택하십시오.

Y 필드. X 필드 값의 범위에 대해 표시할 하나 이상의 필드를 목록에서 선택하십시오. 다중 필드를 선택하려면 필드 선택기 단추를 사용하십시오. 목록에서 필드를 제거하려면 삭제 단추를 클릭하십시오.

오버레이. 여러 가지 방식으로 데이터 값에 대한 범주를 표시할 수 있습니다. 예를 들어, 애니메이션 오버레이를 사용하여 데이터의 값 각각에 대한 다중 도표를 표시할 수 있습니다. 이는 10개보다 많은 범주가 포함된 세트에 대해 유용합니다. 15개보다 많은 범주를 가진 세트에 사용된 경우에는 성능이 저하될 수 있습니다. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

정규화. 그래프에 표시되도록 모든 Y 값을 범위 0-1로 스케일링하려면 이 옵션을 선택하십시오. 정규화는 각 시리즈에 대한 값의 범위에 있는 차이로 인해 모호할 수 있는 선 사이의 관계를 탐색하는 데 도움이 되며 동일한 그래프에서 여러 선을 도표화하거나 패널에서 나란히 도표를 비교할 때 권장됩니다. (모든 데이터 값이 유사한 범위에 속하는 경우에는 정규화가 필요 없습니다.)

그림 1. 시간 경과에 따른 발전소 변동을 표시하는 표준 다중 도표(정규화를 수행하지 않으면 전압에 대한 도표를 볼 수 없음)

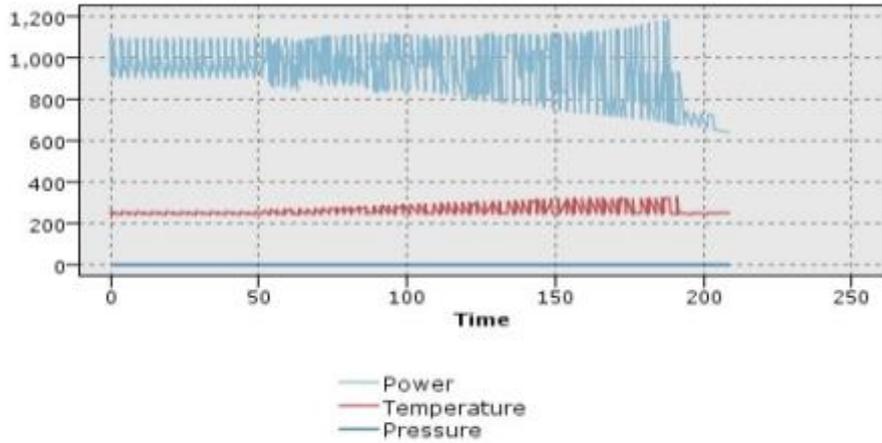
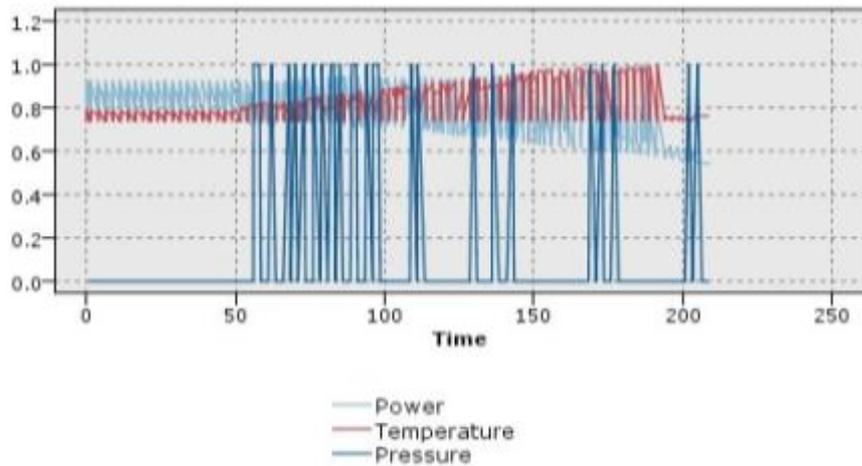


그림 2. 전압에 대한 도표를 표시하는 정규화된 다중 도표



오버레이 함수. 실제 값과 비교할 알려진 함수를 지정하려면 선택하십시오. 예를 들어, 실제 값과 예측값을 비교하려면 $y = x$ 함수를 오버레이로 도표화하십시오. 텍스트 상자에서 $y=$ 에 대한 함수를 지정하십시오. 기본 함수는 $y = x$ 이지만 x 축에서 모든 종류의 함수(이차 함수 또는 임의의 표현식 등)를 지정할 수 있습니다.

참고: 오버레이 함수는 패널 또는 애니메이션 그래프에 대해 사용할 수 없습니다.

레코드 수가 다음보다 많은 경우. 큰 데이터 세트를 도표화하는 방법을 지정하십시오. 최대 데이터 세트 크기를 지정하거나 기본값인 2,000개의 점을 사용할 수 있습니다. 구간 또는 표본 옵션을 선택하면 큰 데이터 세트에 대해 성능이 개선됩니다. 또는 모든 데이터 사용을 선택하여 모든 데이터 포인트를 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격하게 저하될 수 있다는 점에 유의해야 합니다.

참고: X 모드가 오버레이 또는 읽은 대로로 설정된 경우 이 옵션은 사용 안함으로 설정되고 처음 n 개 레코드만 사용됩니다.

- **구간.** 데이터 세트에 지정된 수의 레코드보다 많은 레코드가 포함되어 있는 경우 구간화를 사용하여 설정하려면 선택하십시오. 구간화는 실제로 도표화하기 전에 그래프를 세분화된 눈금으로 나누고 각각의 눈금 셀에 표시되는 연결의 수를 계수합니다. 최종 그래프에서는 구간 중심 값(구간에 있는 모든 연결 점의 평균)에서 셀당 하나의 연결이 사용됩니다.
- **표본.** 지정된 수의 레코드로 데이터에서 무작위로 표본을 추출하려면 선택하십시오.

② 다중 도표 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

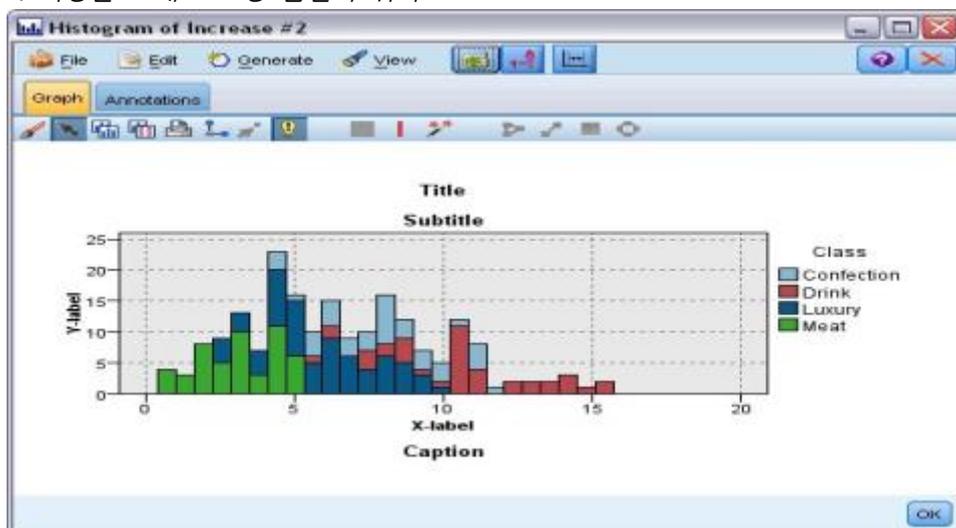
X 레이블. 자동으로 생성된 x축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 y축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. 일부 그래프에서는 이 모든 옵션이 사용되지 않습니다.

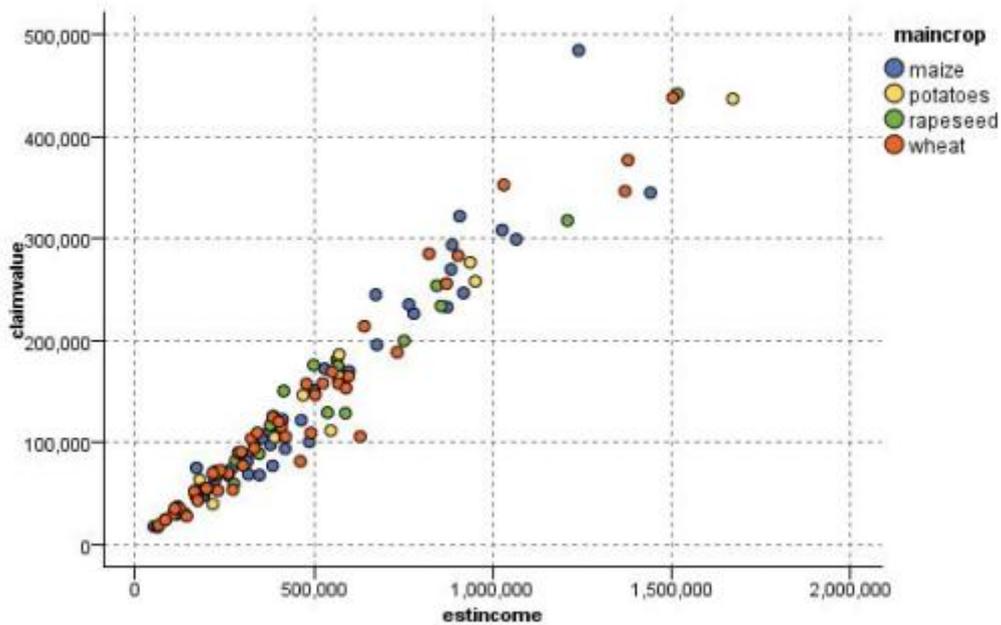
그림 1. 다양한 그래프 모양 옵션의 위치



③ 다중 도표 그래프 사용

도표 및 다중 도표는 본질적으로 Y에 대한 X의 도표입니다. 예를 들어, 농업 허가 신청에서 잠재적 사기를 조사하는 경우에는 신경망에 의해 추정된 수입에 대해 해당 신청에서 청구된 수입을 도표화할 수 있습니다. 작물 유형 등의 오버레이를 사용하면 청구(값 또는 번호)와 작물 유형 사이에 관계가 있는 여부가 표시됩니다.

그림 1. 기본 작물 유형을 오버레이로 가진 추정된 수입과 청구 값 간 관계의 도표



도표, 다중 도표 및 평가 차트는 X에 대한 Y의 2차원 표시이므로 영역을 정의하거나 요소를 표시하거나 밴드를 그려서 이들과 쉽게 상호작용할 수 있습니다. 해당 영역, 밴드 또는 요소로 표시된 데이터에 대한 노트도 생성할 수 있습니다. 자세한 정보는 그래프 탐색의 내용을 참조하십시오.

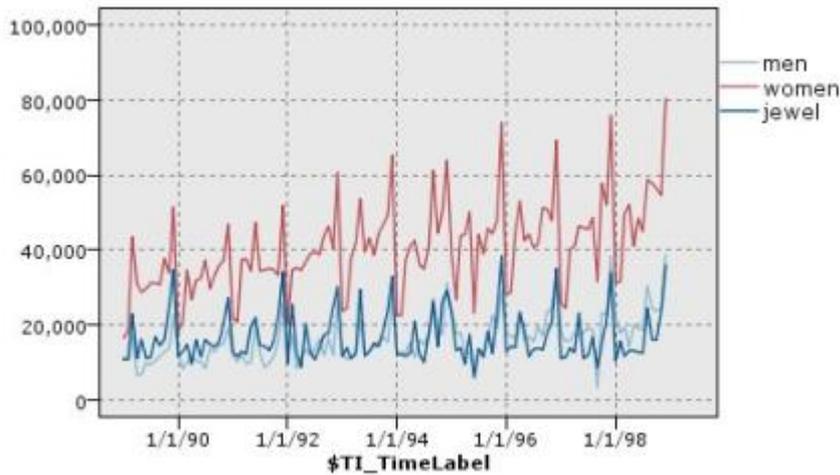
(6) 시간 구성 노트

시간 구성 노트에서는 시간에 따라 구성된 하나 이상의 시계열을 볼 수 있습니다. 구성하는 계열은 숫자 값을 포함해야 하며 주기가 균일한 시간 범위에서 발생한다고 가정합니다.

SPSS® Modeler 버전 17.1 이전에서 일반적으로, 시간 구성 노트 전에 시간 간격 노트를 사용하여 *TimeLabel* 필드를 작성합니다. 이 필드는 그래프에 있는 x축의 레이블을 지정하는 데 기본적으로 사용됩니다.

추가 정보는 시간 구간 노트(더 이상 사용되지 않음)의 내용을 참조하십시오.

그림 1. 시간에 따른 남성 및 여성용 의류와 장신구 판매량 구성



개입 및 이벤트 작성

컨텍스트 메뉴에서 파생(플래그 또는 명목) 노드를 생성하여 시간 구성에서 이벤트 및 개입 필드를 작성할 수 있습니다. 예를 들어, 철도 파업의 경우 이벤트 필드를 작성할 수 있으며, 이벤트가 발생했으면 드라이브 상태는 True이고 그렇지 않으면 False입니다. 개입 필드의 경우, 가격 상승을 예로 들면, 파생 개수를 사용하여 상승 날짜를 식별할 수 있습니다(이전 가격에는 0, 새 가격에는 1 사용). 자세한 정보는 파생 노드의 내용을 참조하십시오.

① 시간 구성 탭

도표. 시계열 데이터를 구성하는 방법을 선택할 수 있습니다.

- **선택된 계열.** 선택된 시계열의 값을 구성합니다. 신뢰구간을 구성할 때 이 옵션을 선택하는 경우 정규화 선택란을 선택 취소하십시오.
- **선택된 시계열 모델.** 시계열 모델과 함께 사용되는 경우, 이 옵션은 하나 이상의 선택된 시계열에 대한 모든 관련 필드(실제 및 예측 값과 신뢰구간)를 구성합니다. 이 옵션을 사용하면 대화 상자의 다른 옵션 중 일부가 사용되지 않습니다. 이 옵션은 신뢰구간을 구성하는 경우에 선호되는 옵션입니다.

계열. 구성할 시계열 데이터를 포함하는 하나 이상의 필드를 선택하십시오. 데이터는 숫자여야 합니다.

X축 레이블. 도표에서 x축의 레이블로 사용할 단일 필드 또는 기본 레이블을 선택하십시오. 기본값을 선택하는 경우, 업스트림 시간 구간 노드가 없으면 시스템은 업스트림(SPSS® Modeler

버전 17.1 이하에서 작성된 스트림) 시간 구간 노드에서 작성된 TimeLabel 필드 또는 순차 정수를 사용합니다.

추가 정보는 시간 구간 노드(더 이상 사용되지 않음)의 내용을 참조하십시오.

개별 패널에 계열 표시. 각 계열을 개별 패널에 표시할지 여부를 지정합니다. 각 계열을 개별 패널에 표시하지 않는 경우, 모든 시계열이 동일한 그래프에서 구성되고 평활기를 사용할 수 없습니다. 동일한 그래프에서 모든 시계열을 구성하면 각 계열이 다른 색상으로 표시됩니다.

정규화. 그래프에 표시되도록 모든 Y 값을 범위 0-1로 스케일링하려면 이 옵션을 선택하십시오. 정규화는 각 시리즈에 대한 값의 범위에 있는 차이로 인해 모호할 수 있는 선 사이의 관계를 탐색하는 데 도움이 되며 동일한 그래프에서 여러 선을 도표화하거나 패널에서 나란히 도표를 비교할 때 권장됩니다. (모든 데이터 값이 유사한 범위에 속하는 경우에는 정규화가 필요 없습니다.)

표시. 도표에 표시할 하나 이상의 요소를 선택하십시오. 선, 점 및 (LOESS) 평활기 중에서 선택할 수 있습니다. 평활기는 계열을 개별 패널에 표시하는 경우에만 사용할 수 있습니다. 기본적으로 선 요소가 선택됩니다. 그래프 노드를 실행하기 전에 하나 이상의 도표 요소를 선택해야 합니다. 그러지 않을 경우, 시스템에서 구성할 요소를 선택하지 않았음을 알리는 오류를 리턴합니다.

레코드 제한. 구성할 레코드 수를 제한하려면 이 옵션을 선택하십시오. **구성할 최대 레코드 수** 옵션에서 구성할 레코드 수(데이터 파일의 시작 부분에서 읽혀지는)를 지정하십시오. 기본적으로 이 수는 2,000으로 설정됩니다. 데이터 파일에 마지막 n 개의 레코드를 구성하려면 이 노드 전에 정렬 노드를 사용하여 레코드를 시간을 기준으로 내림차순으로 배열할 수 있습니다.

② 시간 구성 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

X 레이블. 자동으로 생성된 x축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

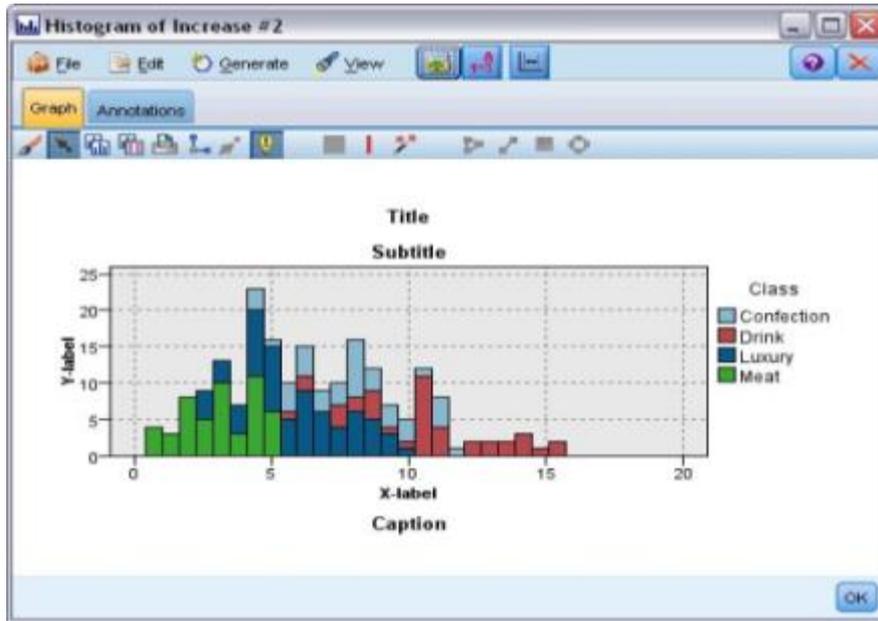
Y 레이블. 자동으로 생성된 y축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

레이아웃. 시간 도표의 경우에만 시간 값이 가로 축과 세로 축 중 어느 축을 따라 도표화되는지를 지정할 수 있습니다.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. 일부 그래프에서는 이 모든 옵션이 사용되지 않습니다.

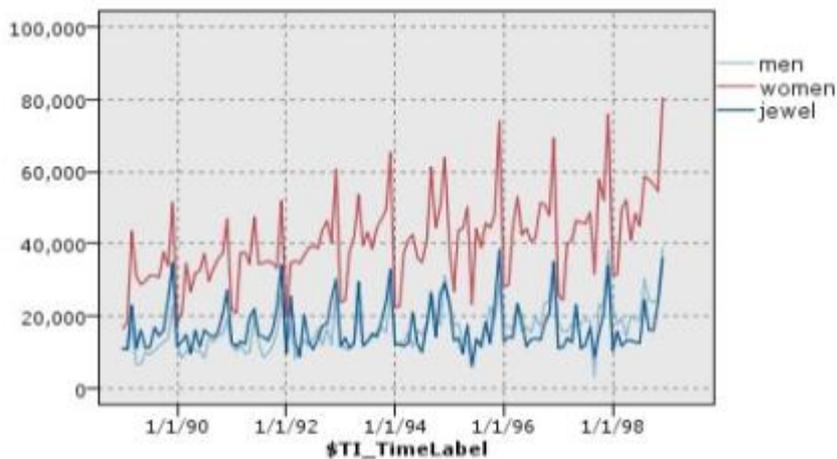
그림 1. 다양한 그래프 모양 옵션의 위치



③ 시간 도표 그래프 사용

시간 도표 그래프를 작성하면 그래프 표시를 조정하고 추가 분석을 위해 노드를 생성하는 여러 옵션을 사용할 수 있습니다. 자세한 정보는 그래프 탐색의 내용을 참조하십시오.

그림 1. 시간에 따른 남성 및 여성용 의류와 장신구 판매량 구성



시간 구성을 작성하고 밴드를 정의하며 결과를 검토한 후에는 메뉴 생성 및 컨텍스트 메뉴의 옵션을 사용하여 선택 또는 파생 노드를 작성할 수 있습니다. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

(7) 분포 노드

분포 그래프 또는 테이블은 데이터 세트에서 담보 유형 또는 성별 등의 숫자가 아닌 기호 값의 발생을 표시합니다. 분포 노드는 일반적으로 모델을 작성하기 전에 균형 노드를 사용하여 수정할 수 있는 데이터의 불균형을 표시하는 데 사용됩니다. 분포 그래프 또는 테이블 창의 생성 메뉴를 사용하여 균형 노드를 자동으로 생성할 수 있습니다.

그래프보드 노드를 사용하여 개수 그래프의 막대를 생성할 수도 있습니다. 그러나 이 노드에서 선택할 수 있는 옵션이 더 많습니다. 자세한 정보는 사용 가능한 내장 그래프보드 시각화 유형의 내용을 참조하십시오.

참고: 숫자 값의 발생을 표시하려면 히스토그램 노드를 사용해야 합니다.

① 분포 도표 탭

도표. 분포 유형을 선택하십시오. 선택된 필드의 분포를 표시하려면 **선택된 필드**를 선택하십시오. 데이터 세트에서 플래그 필드에 대한 true 값의 분포를 표시하려면 **모든 플래그(true 값)**를 선택하십시오.

필드. 값의 분포를 표시할 명목 또는 플래그 필드를 선택하십시오. 명시적으로 숫자로 설정되지 않은 필드만 목록에 표시됩니다.

오버레이. 지정된 필드의 각 값에서 해당 값의 분포를 보여주는 색상 오버레이로 사용할 명목 또는 플래그 필드를 선택하십시오. 예를 들어, 마케팅 캠페인 반응(*rep*)을 하위의 수에 대한 오버레이(*children*)로 사용하여 패밀리 크기별 응답성을 보여줄 수 있습니다. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

색상별 정규화. 모든 막대가 그래프의 전체 너비를 차지하도록 막대의 축척을 지정하려면 선택하십시오. 오버레이 값은 각 막대의 비율과 동일하므로 더 쉽게 범주에서 비교를 수행할 수 있습니다.

정렬. 분포 그래프에서 값을 표시하는 데 사용되는 방법을 선택하십시오. 알파벳순을 사용하려면 **알파벳**을 선택하고 발생의 내림차순으로 값을 나열하려면 **개수별**을 선택하십시오.

비례 척도. 개수가 가장 큰 값이 도표의 전체 너비를 채우도록 값 분포의 축척을 지정하려면 선택하십시오. 다른 모든 막대는 이 값에 대해 축척이 지정됩니다. 이 옵션을 선택 취소하면 각 값의 총 수에 따라 막대의 축척이 지정됩니다.

② 분포 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

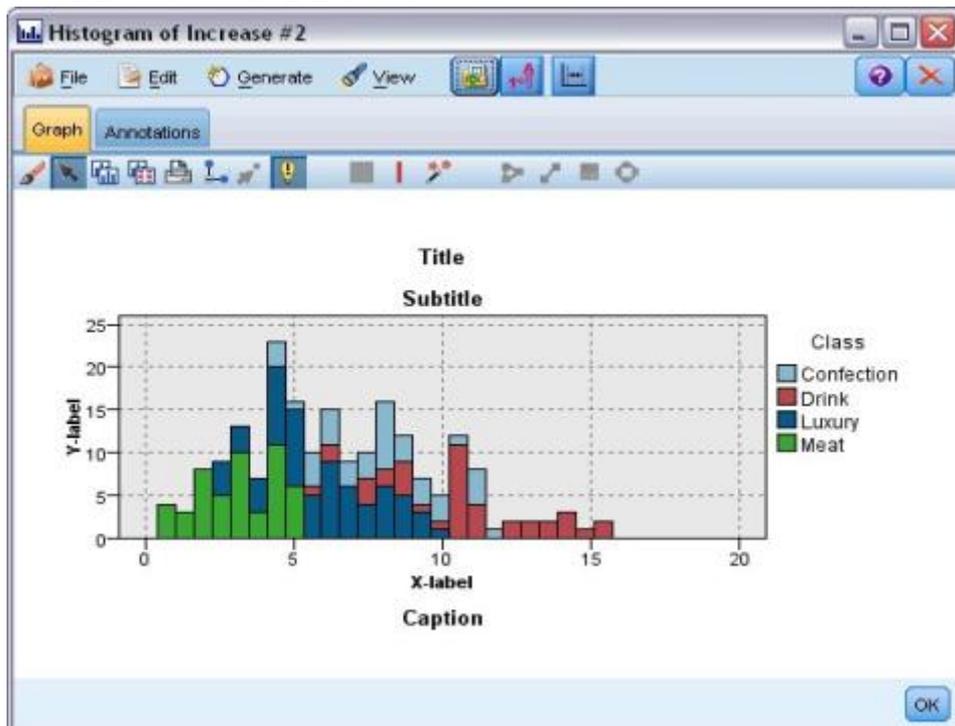
X 레이블. 자동으로 생성된 x축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 y축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. 일부 그래프에서는 이 모든 옵션이 사용되지 않습니다.

그림 1. 다양한 그래프 모양 옵션의 위치



③ 분포 노드 사용

분포 노드는 데이터 세트에서 기호 값의 분포를 표시하는 데 사용됩니다. 분포 노드는 데이터를 탐색하고 불균형을 정정하기 위해 조작 노드 전에 자주 사용됩니다. 예를 들어, 자녀가 없는 응답자의 인스턴스가 다른 유형의 응답자보다 훨씬 자주 발생하는 경우에는 향후 데이터 마이닝 조작에서 더 유용한 규칙이 생성될 수 있도록 이 인스턴스를 줄이길 원할 수 있습니다. 분포 노드를 사용하면 이러한 불균형에 대한 의사결정을 검토하고 작성하는 데 도움이 됩니다.

분포 노드는 데이터를 분석하는 데 필요한 그래프와 테이블을 모두 생성한다는 점에서 특이합니다.

그림 1. 마케팅 캠페인에 응답한 자녀가 있거나 없는 사람의 수를 표시하는 분포 그래프

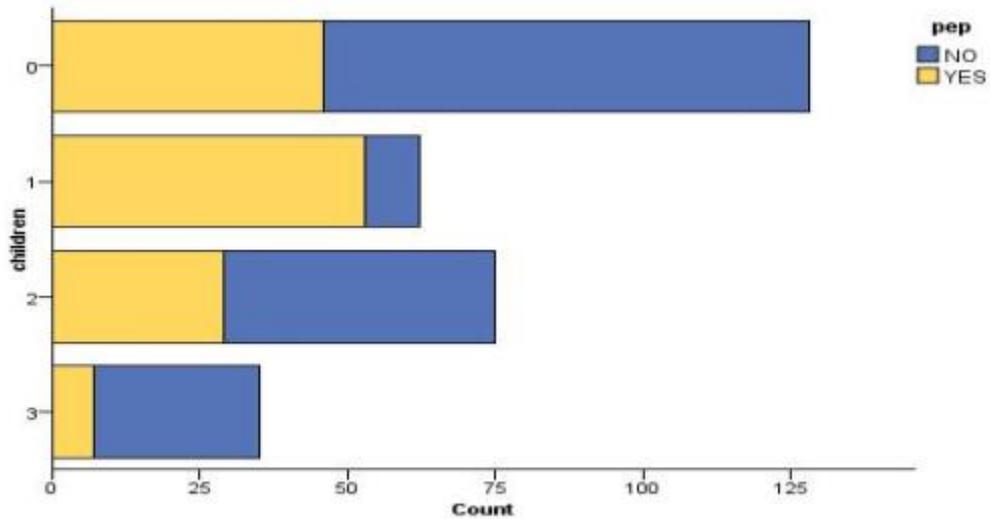
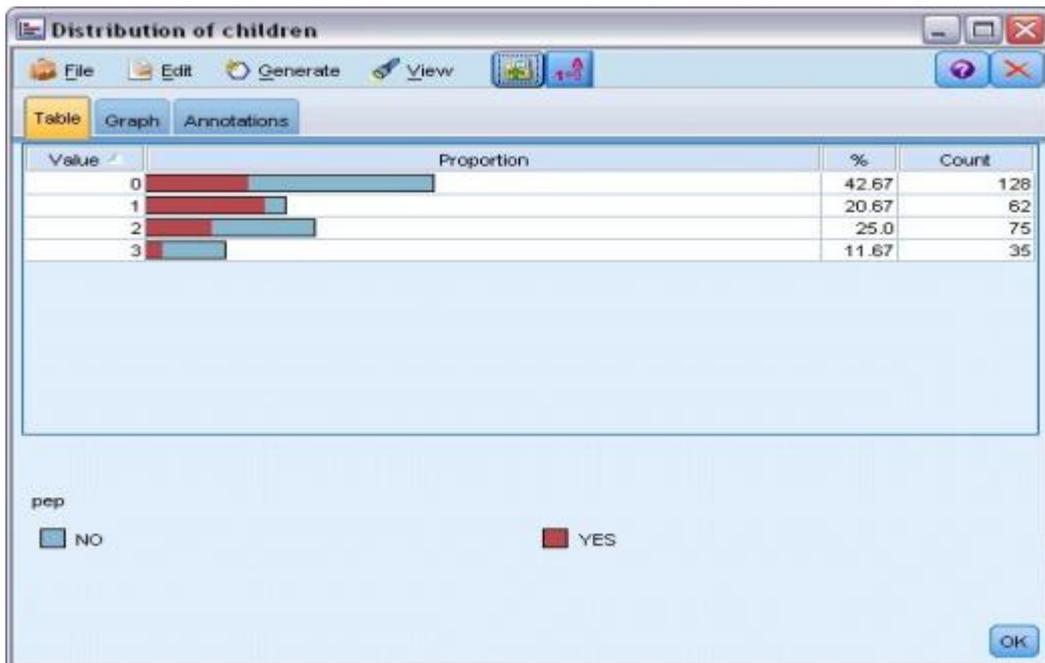


그림 2. 마케팅 캠페인에 응답한 자녀가 있거나 없는 사람의 비율을 표시하는 분포 테이블



분포 테이블 및 그래프를 작성하고 결과를 검토한 후에는 메뉴의 옵션을 사용하여 값을 그룹화하고 값을 복사하고 데이터 준비를 위해 다수의 노드를 생성할 수 있습니다. 또한 MS Word 또는 MS PowerPoint 등의 기타 애플리케이션에서 사용하기 위해 그래프 및 테이블 정보를 복사하거나 내보낼 수 있습니다. 자세한 정보는 그래프 인쇄, 저장, 복사 및 내보내기의 내용을 참조하십시오.

분포 테이블에서 값을 선택하고 복사하려면 다음을 수행하십시오.

1. 마우스 단추를 클릭한 상태로 행 위로 끌어서 값 세트를 선택하십시오. 편집 메뉴를 사용하여 값을 **모두 선택**할 수도 있습니다.
2. 편집 메뉴에서 **테이블 복사** 또는 **테이블 복사(필드 이름 포함)**를 선택하십시오.
3. 클립보드 또는 원하는 애플리케이션에 붙여넣으십시오.
참고: 막대는 직접 복사되지 않습니다. 대신 테이블 값이 복사됩니다. 이는 오버레이된 값은 복사된 테이블에 표시되지 않음을 의미합니다.

분포 테이블에서 값을 그룹화하려면 다음을 수행하십시오.

1. Ctrl+클릭 방법을 사용하여 그룹화할 값을 선택하십시오.
2. 편집 메뉴에서 그룹을 선택하십시오.

참고: 값을 그룹화하고 그룹 해제하면 그래프 탭의 그래프가 자동으로 다시 그려져 변경사항이 표시됩니다.

다음과 같은 작업도 수행할 수 있습니다.

- 분포 목록에서 그룹 이름을 선택한 후 편집 메뉴에서 **그룹 해제**를 선택하여 그룹 해제
- 분포 목록에서 그룹 이름을 선택한 후 편집 메뉴에서 **그룹 편집**을 선택하여 그룹 편집. 이 작업을 수행하면 값을 그룹으로 전환하거나 그룹에서 전환할 수 있는 대화 상자가 열립니다.

메뉴 생성 옵션

생성 메뉴의 옵션을 사용하여 데이터의 서브세트를 선택하거나 플래그 필드를 파생시키거나 값을 재그룹화하거나 값을 재분류하거나 그래프 또는 테이블에서 데이터의 균형을 맞출 수 있습니다. 이 조작은 데이터 준비 노드를 생성하여 스트림 캔버스에 배치합니다. 생성된 노드를 사용하려면 해당 노드를 기존 스트림에 연결하십시오. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

(8) 히스토그램 노드

히스토그램 노드는 숫자 필드에 대한 값의 발생을 표시합니다. 히스토그램 노드는 조작 및 모델 작성 전에 데이터를 탐색하는 데 사용될 수도 있습니다. 분포 노드와 마찬가지로 히스토그램 노드는 데이터의 불균형을 표시하는 데 자주 사용됩니다. 그래프보드 노드를 사용하여 히스토그램을 생성할 수도 있지만 이 노드에서 더 많은 옵션을 선택할 수 있습니다. 자세한 정보는 가능한 내장 그래프보드 시각화 유형의 내용을 참조하십시오.

참고: 기호 필드에 대한 값의 발생을 표시하려면 분포 노드를 사용해야 합니다.

① 히스토그램 도표 탭

필드. 값의 분포를 표시할 숫자 필드를 선택하십시오. 명시적으로 기호(범주형)로 정의되지 않은 필드만 나열됩니다.

오버레이. 지정된 필드에 대한 값의 범주를 표시할 기호 필드를 선택하십시오. 오버레이 필드를 선택하면 히스토그램이 오버레이 필드의 다양한 범주를 나타내는 데 사용되는 색상을 가진 누적 차트로 변환됩니다. 히스토그램 노드를 사용하는 경우에는 세 가지 유형의 오버레이(색상, 패널, 애니메이션)가 있습니다. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

② 히스토그램 옵션 탭

자동 X 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브셋을 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

구간. 숫자별 또는 너비별을 선택하십시오.

- 지정된 구간의 수 및 범위에 따라 너비가 결정되는 고정된 수의 막대를 표시하려면 **숫자별**을 선택하십시오. **구간 수** 옵션에서 그래프에 사용할 구간 수를 표시하십시오. 화살표를 사용하여 수를 조정하십시오.
- 고정된 너비의 구간을 사용하여 그래프를 작성하려면 **너비별**을 선택하십시오. 구간 수는 값의 범위 및 지정된 너비에 따라 다릅니다. **구간 너비** 옵션에서 막대의 너비를 표시하십시오.

색상별 정규화. 모든 막대를 동일한 높이로 조정하여 오버레이된 값을 각 막대에서 전체 케이스의 백분율로 표시하려면 선택하십시오.

정규 곡선 표시. 데이터의 평균 및 분산을 표시하는 정규 곡선을 그래프에 추가하려면 선택하십시오.

각 색상별로 밴드 분리. 각각의 오버레이된 값을 그래프에서 별도의 밴드로 표시하려면 선택하십시오.

③ 히스토그램 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

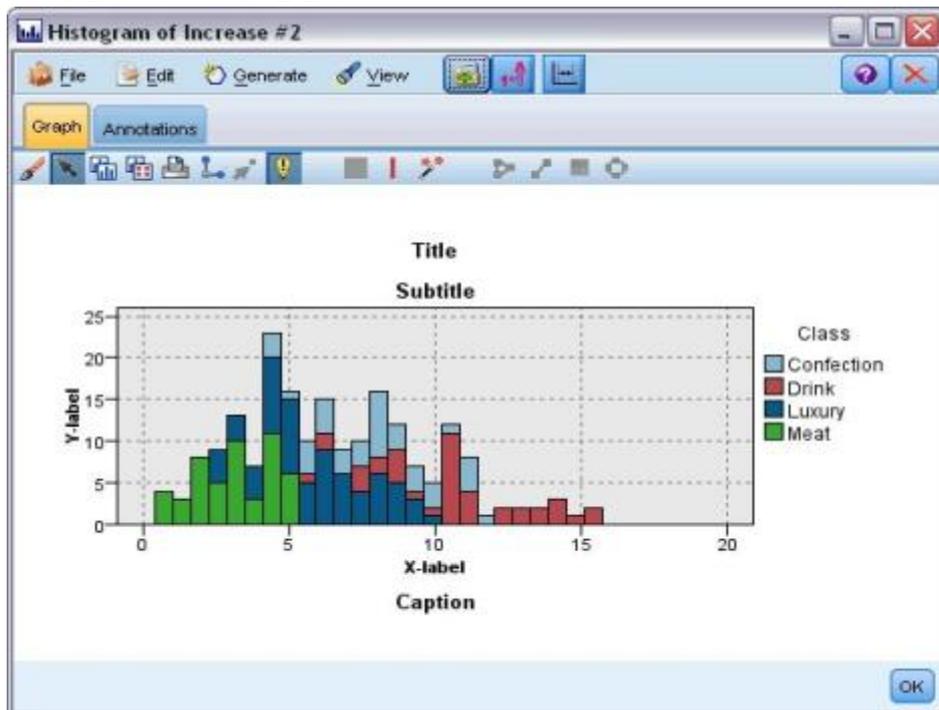
X 레이블. 자동으로 생성된 x축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 y축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다음 예제에서는 그래프에서 모양 옵션이 배치되는 위치를 보여줍니다. 일부 그래프에서는 이 모든 옵션이 사용되지 않습니다.

그림 1. 다양한 그래프 모양 옵션의 위치



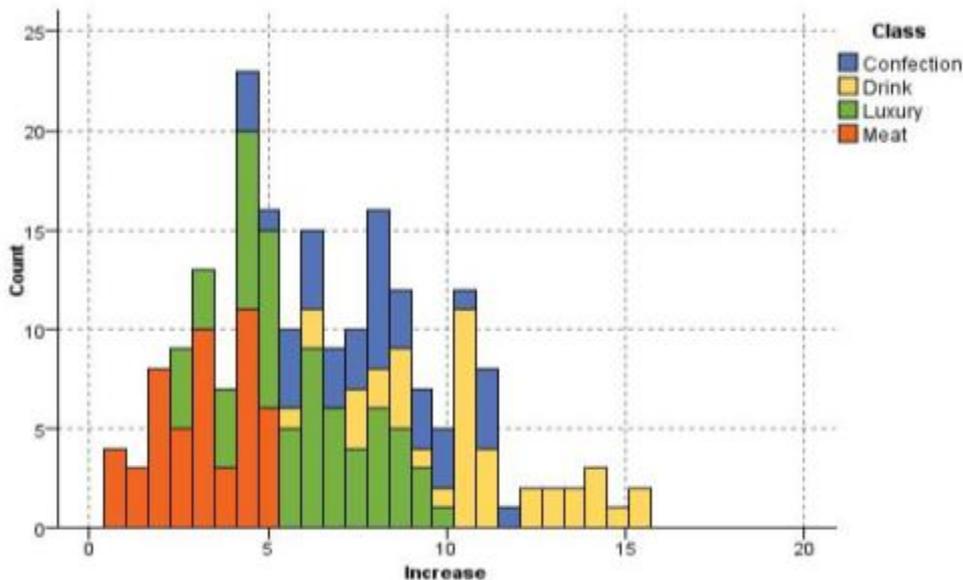
④ 히스토그램 사용

히스토그램은 값 범위가 x 축을 따라 분포하는 숫자 필드에서 값의 분포를 보여줍니다. 히스토그램은 컬렉션 그래프와 비슷하게 작동합니다. 컬렉션은 단일 필드에 대한 값의 발생 대신 *다른 숫자 필드의 값에 대해 상대적인 하나의 숫자 필드의 값 분포*를 표시합니다.

그래프를 작성하고 나면 결과를 검토하고 밴드를 정의하여 x 축을 따라 값을 분할하거나 영역을 정의할 수 있습니다. 그래프 내에서 요소를 표시할 수도 있습니다. 자세한 정보는 그래프 탐색의 내용을 참조하십시오.

생성 메뉴의 옵션을 통해 그래프의 데이터 또는 더 구체적으로 밴드, 영역 또는 표시된 요소 내의 데이터를 사용하여 균형, 선택 또는 파생 노드를 작성할 수 있습니다. 이 유형의 그래프는 스트림에서 사용할 그래프에서 균형 노드를 생성하여 불균형을 정정하고 데이터를 탐색하기 위해 조작 노드 앞에서 자주 사용됩니다. 파생 플래그 노드를 생성하여 각 레코드가 속하는 밴드를 표시하는 필드를 추가하거나 선택 노드를 생성하여 특정 값 범위 또는 세트 내 모든 레코드를 선택할 수도 있습니다. 이러한 조작을 통해 데이터의 특정 서브세트에 초점을 두고 추가적으로 탐색할 수 있습니다. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

그림 1. 프로모션으로 인한 범주별 구매 증가 분포를 보여주는 히스토그램



(9) 요약도표 노드

요약도표는 단일 필드에 대한 값의 발생 대신 다른 필드의 값에 대해 상대적인 하나의 숫자 필드에 대한 값의 분포를 표시한다는 점을 제외하고 히스토그램과 비슷합니다. 요약도표는 시간 경과에 따라 값이 변경되는 변수 또는 필드를 보여주는 데 유용합니다. 3-D 그래프를 사용하여 범주별 분포를 표시하는 기호 축을 포함할 수도 있습니다. 2차원 요약도표는 사용된 오버레이와 함께 누적 막대형 차트로 표시됩니다. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

① 컬렉션 도표 탭

수집. 값을 기간에서 지정된 필드에 대한 값의 범위 동안 수집하고 표시할 필드를 선택하십시오. 기호로 정의되지 않은 필드만 나열됩니다.

기간. 값을 수집에서 지정된 필드를 표시하는 데 사용할 필드를 선택하십시오.

기준. 3차원 그래프 작성 시 사용으로 설정되면 이 옵션을 사용하여 범주별로 컬렉션 필드를 표시하는 데 사용되는 명목 또는 플래그 필드를 선택할 수 있습니다.

연산. 컬렉션 그래프의 각 막대가 표시하는 사항을 선택하십시오. 옵션으로는 **합계**, **평균**, **최대 값**, **최소값**, **표준 편차**가 있습니다.

오버레이. 선택한 필드에 대한 값의 범주를 표시할 기호 필드를 선택하십시오. 오버레이 필드를 선택하면 컬렉션이 변환되고 각 범주에 대해 다양한 색상의 여러 막대가 작성됩니다. 이 노트에는 색상, 패널, 애니메이션이라는 세 가지 유형의 오버레이가 있습니다. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

② 컬렉션 옵션 탭

자동 X 범위. 이 축을 따르는 데이터의 전체 값 범위를 사용하려면 선택하십시오. 지정된 **최소** 및 **최대** 값을 기반으로 값의 명시적 서브셋을 사용하려면 선택 취소하십시오. 값을 입력하거나 화살표를 사용하십시오. 빠르게 그래프를 작성할 수 있도록 기본적으로 자동 범위가 선택됩니다.

구간. 숫자별 또는 **너비별**을 선택하십시오.

- 지정된 구간의 수 및 범위에 따라 너비가 결정되는 고정된 수의 막대를 표시하려면 **숫자별**을 선택하십시오. **구간 수** 옵션에서 그래프에 사용할 구간 수를 표시하십시오. 화살표를 사용하여 수를 조정하십시오.
- 고정된 너비의 구간을 사용하여 그래프를 작성하려면 **너비별**을 선택하십시오. 구간 수는 값의 범위 및 지정된 너비에 따라 다릅니다. **구간 너비** 옵션에서 막대의 너비를 표시하십시오.

③ 컬렉션 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

기간 레이블. 자동으로 생성된 레이블을 승인하거나 사용자 정의를 선택하여 레이블을 지정하십시오.

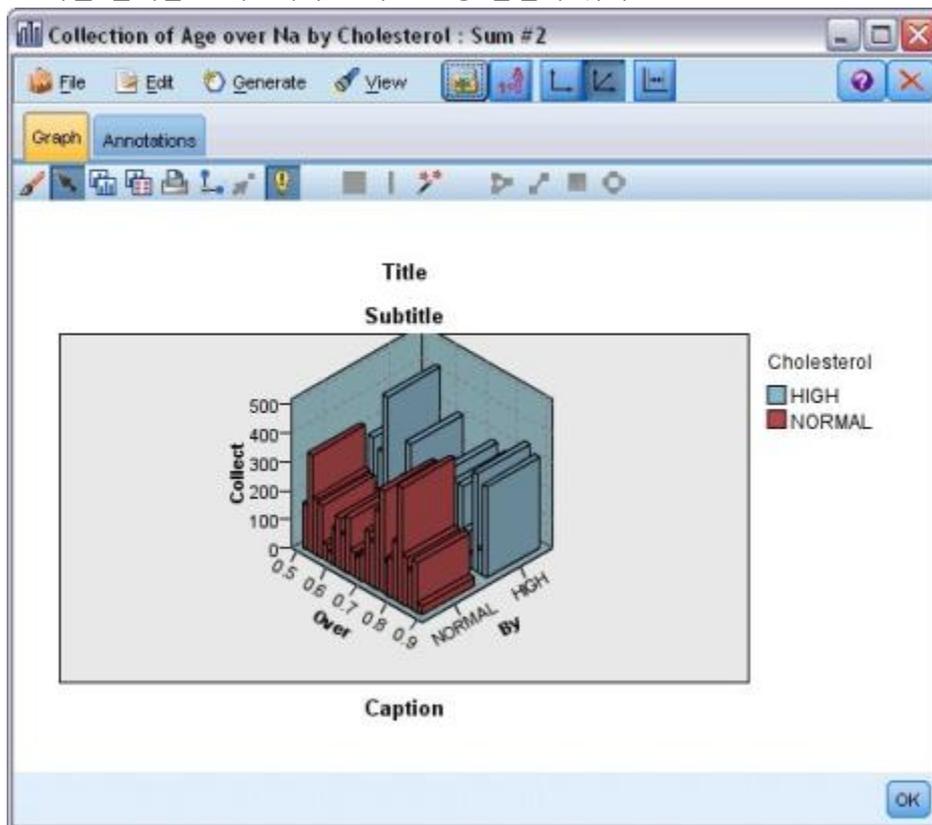
수집 레이블. 자동으로 생성된 레이블을 승인하거나 사용자 정의를 선택하여 레이블을 지정하십시오.

기준 레이블. 자동으로 생성된 레이블을 승인하거나 사용자 정의를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

다음 예제에서는 3차원 버전의 그래프에서 모양 옵션의 위치를 보여줍니다.

그림 1. 3차원 콜렉션 그래프에서 그래프 모양 옵션의 위치



④ 콜렉션 그래프 사용

콜렉션은 단일 필드에 대한 값의 발생 대신 다른 숫자 필드의 값에 대해 상대적인 하나의 숫자 필드의 값 분포를 표시합니다. 히스토그램은 콜렉션 그래프와 비슷하게 작동합니다. 히스토그램은 값 범위가 x축을 따라 분포하는 숫자 필드에서 값의 분포를 보여줍니다.

그래프를 작성하고 나면 결과를 검토하고 밴드를 정의하여 x축을 따라 값을 분할하거나 영역을 정의할 수 있습니다. 그래프 내에서 요소를 표시할 수도 있습니다. 자세한 정보는 그래프 탐색의 내용을 참조하십시오.

생성 메뉴의 옵션을 통해 그래프의 데이터 또는 더 구체적으로 밴드, 영역 또는 표시된 요소 내의 데이터를 사용하여 균형, 선택 또는 파생 노드를 작성할 수 있습니다. 이 유형의 그래프는 스트림에서 사용할 그래프에서 균형 노드를 생성하여 불균형을 경정하고 데이터를 탐색하기 위해 조작 노드 앞에서 자주 사용됩니다. 파생 플래그 노드를 생성하여 각 레코드가 속하는 밴드를 표시하는 필드를 추가하거나 선택 노드를 생성하여 특정 값 범위 또는 세트 내 모든 레코드를 선택할 수도 있습니다. 이러한 조작을 통해 데이터의 특정 서브세트에 초점을 두고 추가적으로 탐색할 수 있습니다. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

그림 1. 콜레스테롤 수준 높음 및 정상에 대해 연령에 대한 Na_to_K의 합계를 보여주는 3차원 컬렉션 그래프

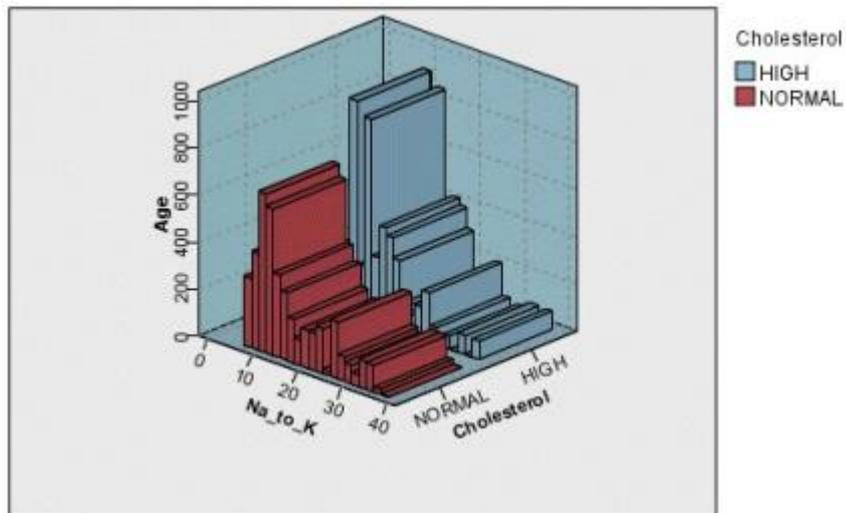
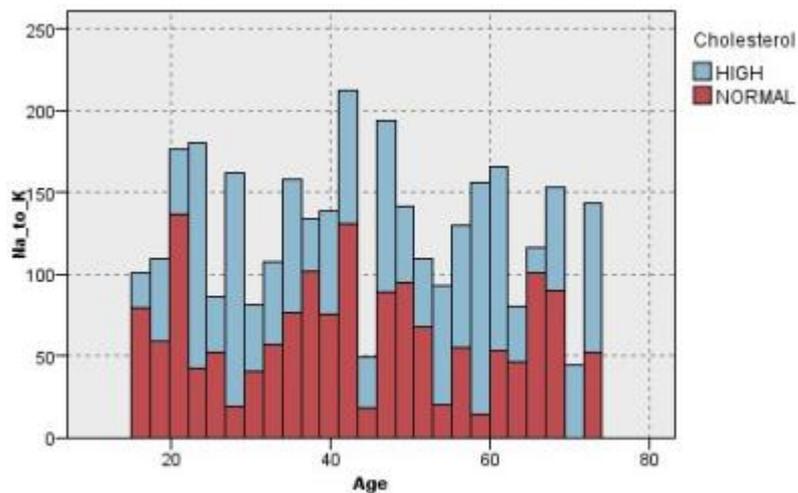


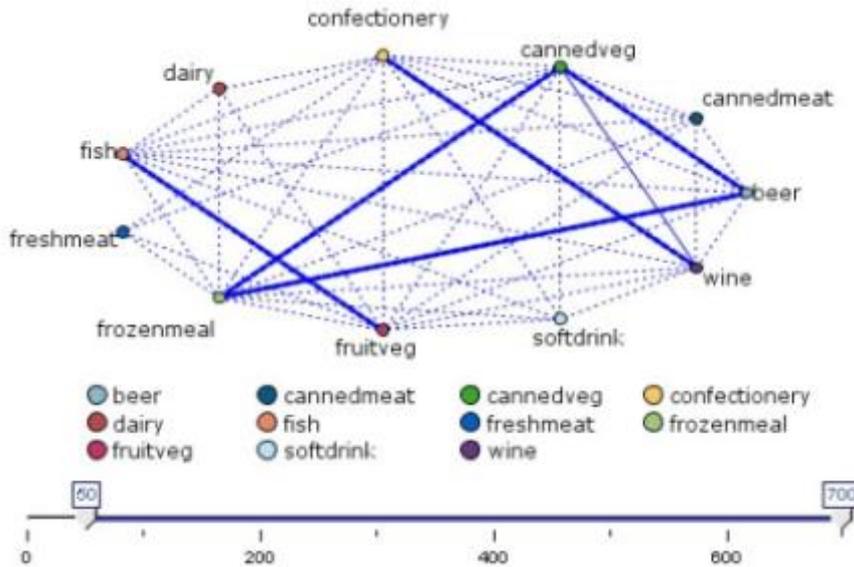
그림 2. z축은 표시되지 않지만 콜레스테롤을 색상 오버레이로 가진 컬렉션 그래프



(10) 웹 노드

웹 노드는 둘 이상의 기호 필드의 값 사이의 관계 강도를 표시합니다. 그래프는 연결 강도를 나타내는 다양한 유형의 선을 사용하여 연결을 표시합니다. 예를 들어, 웹 노드를 사용하여 전자상거래 사이트 또는 일반 소매판매점에서의 다양한 항목 구매 간의 관계를 탐색할 수 있습니다.

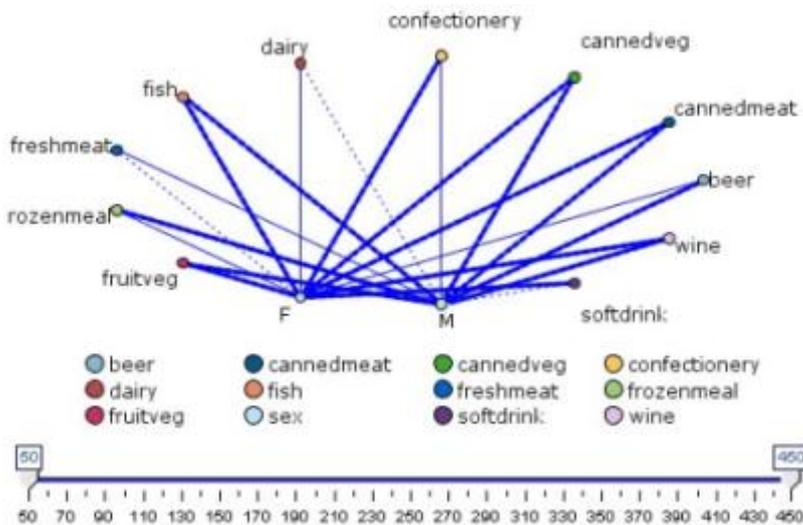
그림 1. 식품품 항목 구매 간의 관계를 보여주는 웹 그래프



방향이 있는 웹

방향이 있는 웹 노드는 기호 필드 사이의 관계 강도를 표시한다는 점에서 웹 노드와 유사합니다. 그러나, 방향이 있는 웹 그래프는 하나 이상의 시작 필드에서 하나의 대상 필드로의 연결만 표시합니다. 한 방향으로만 연결된다는 의미에서 연결은 단방향 연결입니다.

그림 2. 식품품 항목의 구매와 성별 간의 관계를 보여주는 방향이 있는 웹 그래프



웹 노드와 마찬가지로, 그래프는 연결 강도를 나타내는 다양한 유형의 선을 사용하여 연결을 표시합니다. 예를 들어, 방향이 있는 웹 노드를 사용하여 특정 구매 항목에 대한 성향과 성별 간의 관계를 탐색할 수 있습니다.

① 웹 구성 탭

웹. 지정된 모든 필드 간의 관계 강도를 보여주는 웹 그래프를 작성하려면 선택하십시오.

방향이 있는 웹. 하나의 필드(예: 성별 또는 종교)의 값과 여러 필드 간의 관계 강도를 보여주는 방향이 있는 웹 그래프를 작성하려면 선택하십시오. 이 옵션을 선택하면 대상 필드가 활성화되고 아래의 필드 제어가 보다 명확하게 알 수 있도록 시작 필드로 이름이 바뀝니다.

대상 필드(방향이 있는 웹에만 해당). 방향이 있는 웹에 사용되는 플래그 또는 명목 필드를 선택하십시오. 명시적으로 숫자로 설정되지 않은 필드만 나열됩니다.

필드/시작 필드. 웹 그래프를 작성할 필드를 선택하십시오. 명시적으로 숫자로 설정되지 않은 필드만 나열됩니다. 필드 선택기 단추를 사용하여 여러 필드를 선택하거나 유형별로 필드를 선택하십시오.

참고: 방향이 있는 웹의 경우, 이 제어는 시작 필드를 선택하는 데 사용됩니다.

참 플래그만 표시. 플래그 필드의 참 플래그만 표시하려면 선택하십시오. 이 옵션은 웹 디스플레이를 단순화하며 가끔 양의 값의 발생이 특히 중요한 데이터에 사용됩니다.

선 값. 드롭 다운 목록에서 임계값 유형을 선택하십시오.

- **절대값:** 각 값 쌍을 갖는 레코드 수를 기반으로 임계값을 설정합니다.
- **전체 백분율:** 링크가 웹 그래프에서 제공되는 각 값 쌍의 모든 발생의 비율로서 제공하는 절대 케이스 수를 표시합니다.
- **더 작은 필드/값의 백분율 및 더 큰 필드/값의 백분율:** 백분율을 평가하는 데 사용할 필드/값을 나타냅니다. 예를 들어, 100개의 레코드가 *약제* 필드에 *drugY* 값을 갖고 10개만 *BP* 필드에 낮은 값을 갖는다고 가정하십시오. 7개의 레코드가 *drugY* 및 낮은 값을 모두 갖는 경우, 이 백분율은 참조하는 필드(더 작은 필드(*BP*) 또는 더 큰 필드(*약제*))에 따라 70% 또는 7%입니다.

참고: 방향이 있는 웹 그래프의 경우, 위의 세 번째 및 네 번째 옵션을 사용할 수 없습니다. 대신, "대상" 필드/값의 백분율 및 "시작" 필드/값의 백분율을 선택할 수 있습니다.

강한 링크가 더 굵음. 기본적으로 선택됩니다. 필드 사이의 링크를 표시하는 표준 방법입니다.

약한 링크가 더 굵음. 굵은 선으로 표시되는 링크의 의미를 정반대로 바꾸려면 선택하십시오. 이 옵션은 부정행위를 발견하거나 이상치를 조사하는 데 자주 사용됩니다.

② 웹 옵션 탭

웹 노드의 옵션 탭에는 출력 그래프를 사용자 정의하는 다수의 추가 옵션이 있습니다.

링크 수. 다음 옵션은 출력 그래프에 표시되는 링크 수를 제어하는 데 사용됩니다. 이 옵션 중 일부(예: **약한 링크 상한** 및 **강한 링크 하한**)는 출력 그래프 창에서도 사용 가능합니다. 또한 최종 그래프에 있는 슬라이더 제어를 사용하여 표시되는 링크 수를 조정할 수도 있습니다.

- **표시할 최대 링크 수.** 출력 그래프에 표시할 최대 링크 수를 나타내는 숫자를 지정하십시오. 화살표를 사용하여 값을 조정하십시오.
- **다음 이상의 링크만 표시.** 웹의 연결을 표시할 최소값을 나타내는 숫자를 지정하십시오. 화살표를 사용하여 값을 조정하십시오.
- **모든 링크 표시.** 최소 또는 최대 값과 상관없이 모든 링크를 표시하려면 지정하십시오. 이 옵션을 선택하면 많은 수의 필드가 있는 경우 처리 시간이 늘어날 수 있습니다.

레코드 수가 매우 적은 경우 삭제. 지원하는 레코드 수가 너무 적은 연결을 무시하려면 선택하십시오. **Min. records/line**에 숫자를 입력하여 이 옵션의 임계값을 설정하십시오.

레코드 수가 매우 많은 경우 삭제. 강력하게 지원되는 연결을 무시하려면 선택하십시오. **Max. records/line**에 숫자를 입력하십시오.

약한 링크 상한. 약한 연결(점선)과 보통 연결(실선)의 임계값을 나타내는 숫자를 지정하십시오. 이 값 아래의 모든 연결은 약한 연결로 간주됩니다.

강한 링크 하한. 강한 연결(굵은 선)과 보통 연결(실선)의 임계값을 지정하십시오. 이 값 위의 모든 연결은 강한 연결로 간주됩니다.

링크 크기. 링크 크기를 제어하는 옵션을 지정하십시오.

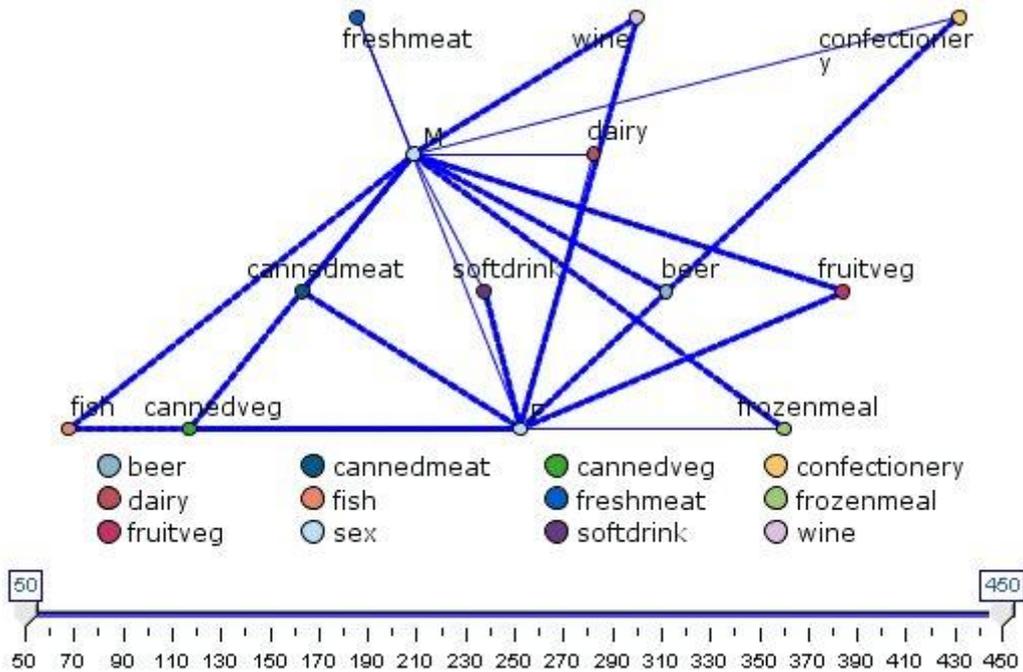
- **링크 크기가 계속 변화.** 실제 데이터 값을 기반으로 연결 강도의 변화를 반영하는 링크 크기 범위를 표시하려면 선택하십시오.
- **링크 크기가 강한/보통/약한 범주 표시.** 연결의 세 가지 강도(강함, 보통 및 약함)를 표시하려면 선택하십시오. 최종 그래프에서뿐만 아니라 위에서도 이러한 범주의 분별점을 지정할 수 있습니다.

웹 디스플레이. 웹 디스플레이의 유형을 선택하십시오.

- **원 레이아웃.** 표준 웹 디스플레이를 사용하려면 선택하십시오.
- **네트워크 레이아웃.** 가장 강한 링크를 함께 모으는 알고리즘을 사용하려면 선택하십시오. 굵은 선뿐만 아니라 공간 분화를 사용하여 강한 링크를 강조표시하는 데 사용됩니다.
- **방향이 있는 레이아웃.** 방향이 있는 웹 디스플레이를 작성하려면 선택하십시오. 방향이 있는 웹 디스플레이는 방향의 초점으로 구성 탭의 **대상 필드** 선택항목을 사용합니다.

- **눈금 레이아웃**. 간격이 일정한 눈금 패턴으로 레이아웃된 웹 디스플레이를 작성하려면 선택하십시오.

그림 1. 냉동식품 및 통조림 야채에서 다른 식품 항목으로의 강한 연결을 보여주는 웹 그래프



① **참고:** 표시된 링크를 필터링하면(웹 그래프의 슬라이더를 사용하거나 웹 노드의 옵션 탭에 위의 링크만 표시 제어 사용) 표시된 상태로 남아 있는 모든 링크가 단일 값인 상황이 발생합니다. 즉, 웹 노드의 옵션 탭에서 아래 약한 링크 및 위의 강한 링크 제어에 정의된 대로 모두 약한 링크, 모두 중간 링크 또는 모두 강한 링크입니다. 이 경우 웹 그래프 출력에서 모든 링크가 중간 너비 라인으로 표시됩니다.

③ 웹 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

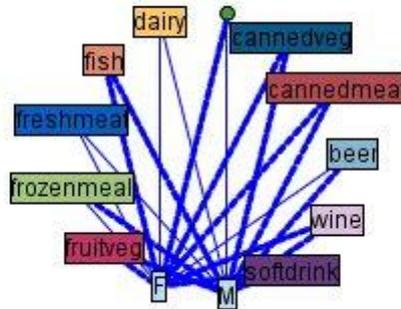
캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

범례 표시. 범례가 표시되는지 여부를 지정할 수 있습니다. 많은 수의 필드를 가진 도표의 경우 범례를 숨기면 도표의 모양이 개선될 수 있습니다.

레이블을 노드로 사용. 인접 레이블을 표시하는 대신 각 노드 내부에 레이블 텍스트를 포함할 수 있습니다. 적은 수의 필드를 가진 도표의 경우에는 이를 통해 차트의 가독성이 향상될 수 있습니다.

그림 1. 레이블을 노드로 표시하는 웹 그래프

Relationship between gender and grocery purchases



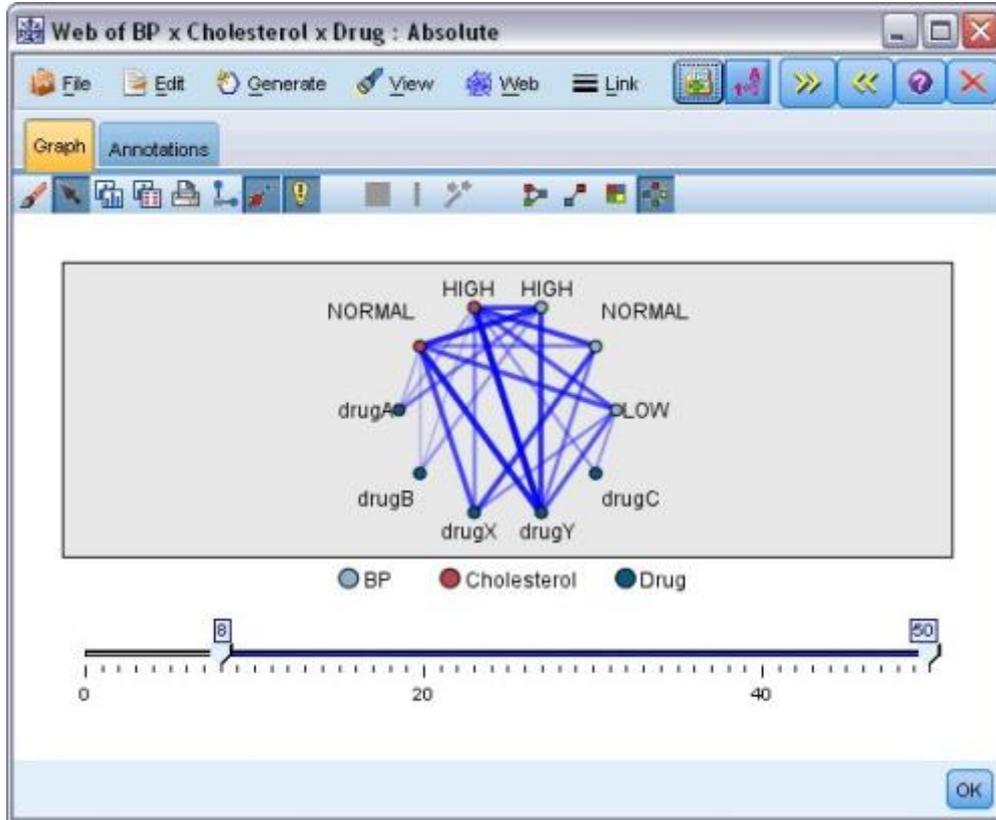
④ 웹 그래프 사용

웹 노드는 둘 이상의 기호 필드의 값 사이의 관계 강도를 표시하는 데 사용됩니다. 연결은 그래프에서 연결 강도를 나타내는 다양한 유형의 선으로 표시됩니다. 웹 노드를 사용하여, 예를 들어, 콜레스테롤 수준과 혈압 그리고 환자의 질병을 치료하는 데 효과적이었던 약제 사이의 관계를 탐색할 수 있습니다.

- 강한 연결은 굵은 선으로 표시됩니다. 이는 두 값이 강하게 관련되어 있고 추가 탐색이 필요함을 표시합니다.
- 중간 연결은 보통 굵기의 선으로 표시됩니다.
- 약한 연결은 점선으로 표시됩니다.
- 두 값 사이에 선이 표시되지 않는 경우, 이는 두 값이 결코 동일한 레코드에서 발생하지 않거나 이러한 조합이 웹 노드 대화 상자에 지정된 임계값 아래의 다수의 레코드에서 발생함을 의미합니다.

웹 노드를 작성하면 그래프 표시를 조정하고 추가 분석을 위해 노드를 생성하는 여러 옵션을 사용할 수 있습니다.

그림 1. 다수의 강한 관계(예: 정상 혈압과 DrugX 및 고콜레스테롤과 DrugY)를 나타내는 웹 그래프



웹 노드 및 방향이 있는 웹 노드 둘 다에 대해 다음을 수행할 수 있습니다.

- 웹 디스플레이의 레이아웃을 변경합니다.
- 점을 숨겨 디스플레이를 단순화합니다.
- 선 스타일을 제어하는 임계값을 변경합니다.
- 값 사이의 선을 강조표시하여 "선택된" 관계를 나타냅니다.
- 하나 이상의 "선택된" 레코드에 대한 선택 노드를 생성하거나 웹에 있는 하나 이상의 관계와 연관된 파생 플래그 노드를 생성합니다.

점 조정

- **이동:** 점에서 마우스를 클릭하여 새 위치로 끌어와 점을 이동시킵니다. 새 위치를 반영하도록 웹이 다시 그려집니다.
- **숨기기:** 웹의 점을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **숨기기** 또는 **숨기기 및 다시 계획**을 선택하여 점을 숨깁니다. **숨기기**는 단지 선택된 점 및 그와 연관된 선을 숨기기만 합니다. **숨기기 및 다시 계획**은 수행한 변경에 맞게 웹을 다시 그립니다. 수동 이동이 수행되지 않습니다.

- 표시: 그래프 창의 웹 메뉴에서 **모두 표시** 또는 **모두 표시 및 다시 계획**을 선택하여 숨겨진 모든 점을 표시합니다. **모두 표시 및 다시 계획**을 선택하면 이전에 숨겨진 모든 점과 해당 연결을 포함하도록 웹이 다시 그려집니다.

선 선택 또는 "강조표시"

선택된 선은 빨간색으로 강조표시됩니다.

1. 한 개의 선을 선택하려면 선을 마우스 왼쪽 단추로 클릭하십시오.
2. 여러 개의 선을 선택하려면 다음 중 하나를 수행하십시오.

- 커서를 사용하여 해당 선을 선택할 점 주위에 원을 그리십시오.
- Ctrl 키를 누른 상태에서 선택할 개별 선을 마우스 왼쪽 단추로 클릭하십시오.

그래프 배경을 클릭하거나 그래프 창의 웹 메뉴에서 **선택항목 지우기**를 선택하여 선택된 모든 행을 선택 취소할 수 있습니다.

다른 레이아웃을 사용하여 웹 보기

웹 메뉴에서 **원 레이아웃**, **네트워크 레이아웃**, **방향이 있는 레이아웃** 또는 **눈금 레이아웃**을 선택하여 그래프의 레이아웃을 변경하십시오.

링크 슬라이더를 켜거나 끄기

보기 메뉴에서 **링크 슬라이더**를 선택하십시오.

단일 관계의 레코드를 선택하거나 플래그 지정

1. 관심이 있는 관계를 나타내는 선을 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 **링크에 대한 선택 노드 생성** 또는 **링크에 대한 파생 노드 생성**을 선택하십시오.

선택 노드 또는 파생 노드는 해당 옵션 및 조건이 지정된 상태로 스트림 캔버스에 자동으로 추가됩니다.

- 선택 노드는 지정된 관계의 모든 레코드를 선택합니다.

- 파생 노드는 선택된 관계가 전체 데이터 세트의 레코드에 대해 참인지 여부를 나타내는 플래그를 생성합니다. 플래그 필드의 이름은 관계를 갖는 두 값을 밑줄로 결합하여 지정합니다(예: LOW_drugC 또는 drugC_LOW).

관계 그룹의 플래그 레코드 선택

1. 웹 디스플레이에서 관심이 있는 관계를 나타내는 선을 선택하십시오.
 2. 그래프 창의 생성 메뉴에서 **Select Node ("And")**, **Select Node ("Or")**, **Derive Node ("And")** 또는 **Derive Node ("Or")**를 선택하십시오.
- "Or" 노드는 조건의 이접성을 제공합니다. 이는 선택된 관계 중 임의 관계가 참인 레코드에 노드가 적용됨을 의미합니다.
 - "And" 노드는 조건의 연접성을 제공합니다. 이는 선택된 모든 관계가 참인 레코드에만 노드가 적용됨을 의미합니다. 선택된 관계 중 상호 배타적인 관계가 있으면 오류가 발생합니다.

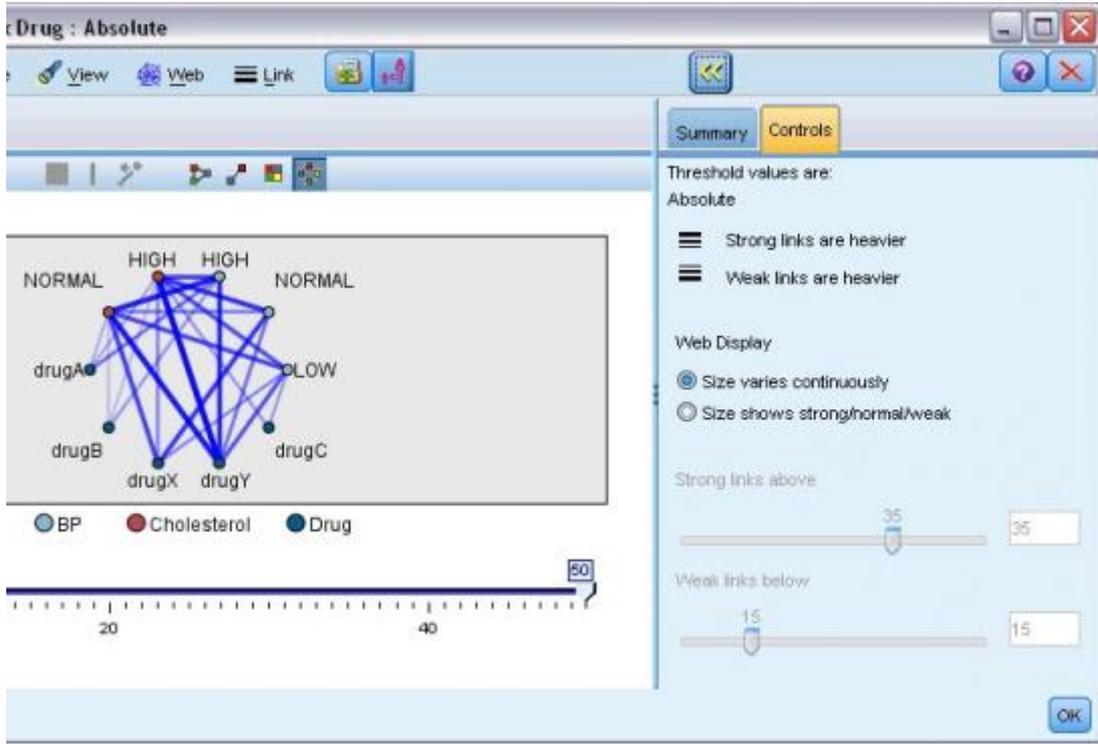
선택을 완료하면 선택 노드 또는 파생 노드가 해당 옵션 및 조건이 지정된 상태로 스트림 캔버스에 자동으로 추가됩니다.

참고: 표시된 링크를 필터링하면(웹 그래프의 슬라이더를 사용하거나 웹 노드의 옵션 탭에 위의 링크만 표시 제어 사용) 표시된 상태로 남아 있는 모든 링크가 단일 값인 상황이 발생합니다. 즉, 웹 노드의 옵션 탭에서 **아래 약한 링크** 및 **위의 강한 링크** 제어에 정의된 대로 모두 약한 링크, 모두 중간 링크 또는 모두 강한 링크입니다. 이 경우 웹 그래프 출력에서 모든 링크가 중간 너비 라인으로 표시됩니다.

가. 웹 임계값 조정

웹 그래프를 작성한 후에는 최소 가시성을 변경하는 도구 모음 슬라이더를 사용하여 선 스타일을 제어하는 임계값을 조정할 수 있습니다. 또한 도구 모음에서 노란색 이중 화살표 단추를 클릭하여 웹 그래프 창을 펼친 후 추가 임계값 옵션을 볼 수 있습니다. **제어** 탭을 클릭하면 추가 옵션이 표시됩니다.

그림 1. 디스플레이 및 임계값 옵션이 있는 펼친 창



임계값. 웹 노드 대화 상자에서 작성 중에 선택된 임계값 유형을 표시합니다.

강한 링크가 더 굵음. 기본적으로 선택됩니다. 필드 사이의 링크를 표시하는 표준 방법입니다.

약한 링크가 더 굵음. 굵은 선으로 표시되는 링크의 의미를 정반대로 바꾸려면 선택하십시오. 이 옵션은 부정행위를 발견하거나 이상값을 조사하는 데 자주 사용됩니다.

웹 디스플레이. 출력 그래프에서 링크 크기를 제어하는 옵션을 지정하십시오.

- 크기가 계속 변화. 실제 데이터 값을 기반으로 연결 강도의 변화를 반영하는 링크 크기 범위를 표시하려면 선택하십시오.
- 크기가 강함/보통/약함 표시. 연결의 세 가지 강도(강함, 보통 및 약함)를 표시하려면 선택하십시오. 최종 그래프에서만 아니라 위에서도 이러한 범주의 분별점을 지정할 수 있습니다.

강한 링크 하한. 강한 연결(굵은 선)과 보통 연결(실선)의 임계값을 지정하십시오. 이 값 위의 모든 연결은 강한 연결로 간주됩니다. 슬라이더를 사용하여 값을 조정하거나 필드에 숫자를 입력하십시오.

약한 링크 상한. 약한 연결(점선)과 보통 연결(실선)의 임계값을 나타내는 숫자를 지정하십시오. 이 값 아래의 모든 연결은 약한 연결로 간주됩니다. 슬라이더를 사용하여 값을 조정하거나 필드에 숫자를 입력하십시오.

웹의 임계값을 조정한 후에는 웹 그래프 도구 모음에 있는 웹 메뉴를 통해 새 임계값으로 웹 디스플레이를 다시 계획하거나 다시 그릴 수 있습니다. 가장 의미있는 패턴을 표시하는 설정을 알았으면, 그래프 창의 웹 메뉴에서 **상위 노드 업데이트**를 선택하여 웹 노드(상위 웹 노드라고도 함)의 원래 설정을 업데이트할 수 있습니다.

나. 웹 요약 작성

도구 모음에서 노란색 이중 화살표 단추를 클릭하여 웹 그래프 창을 펼친 후 강한 링크, 중간 링크 및 약한 링크를 나열하는 웹 요약 문서를 작성할 수 있습니다. **요약** 탭을 클릭하면 각 링크 유형의 테이블이 표시됩니다. 각각의 토크 단추를 사용하여 테이블을 펼치고 접을 수 있습니다.

요약을 인쇄하려면 웹 그래프 창의 메뉴에서 다음을 선택하십시오.

파일 > 요약 인쇄

(11) 평가 노드

평가 노드는 애플리케이션에 대해 최적 모델을 선택하기 위해 예측 모형을 평가하고 비교하는 쉬운 방법을 제공합니다. 평가 차트는 특정 결과 예측 시 모델이 작동하는 방식을 보여줍니다. 평가 차트는 예측의 신뢰도 및 예측값을 기반으로 레코드를 정렬하고 레코드를 동일한 크기의 그룹(분위수)으로 분할한 후 각 분위수에 대한 비즈니스 기준의 값을 내림차순으로 도표로 작성하여 작동합니다. 다중 모델이 도표에 선구분 변수로 표시됩니다.

결과는 특정 값 또는 값 범위를 **적중**으로 정의하여 처리됩니다. 적중은 일반적으로 관심 있는 이벤트(예: 특정 의료 진단) 또는 일부 정렬(예: 고객에 대한 판매)의 성공을 표시합니다. 대화 상자의 옵션 탭에서 적중 기준을 정의하거나 다음과 같이 기본 적중 기준을 사용할 수 있습니다.

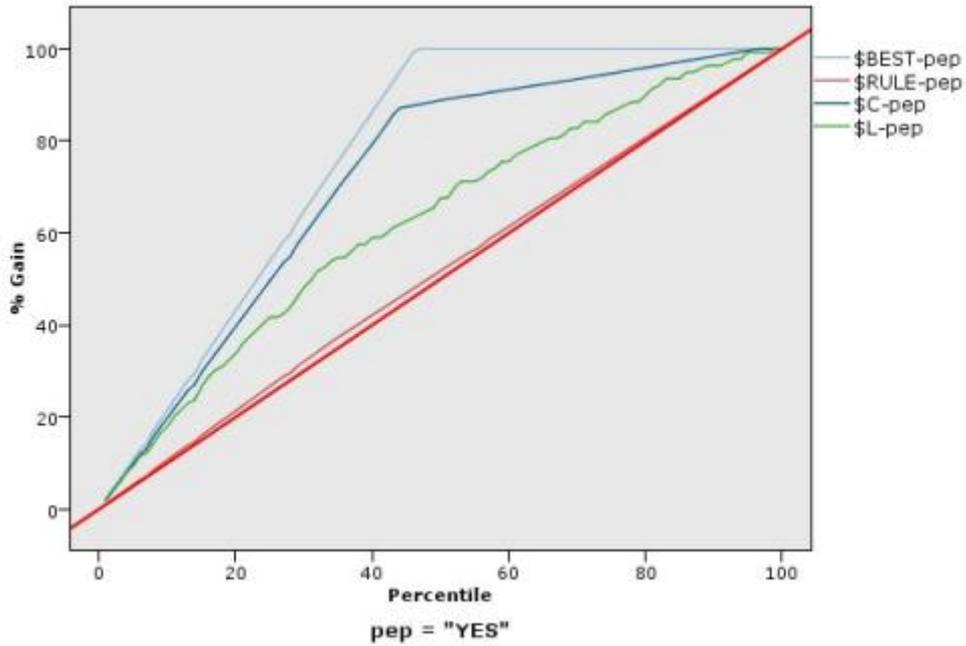
- **플래그** 출력 필드는 직설적이어서 적중은 **참** 값에 해당합니다.
- **명목** 출력 필드의 경우 세트의 첫 번째 값이 적중을 정의합니다.
- **연속형** 출력 필드의 경우 적중은 필드 범위의 중심점보다 큰 값과 동일합니다.

여섯 가지 유형의 평가 차트가 있으며 각각의 차트는 다른 평가 기준을 강조합니다.

Gains 차트

Gains는 각 분위수에서 발생하는 적중 총계의 비율로 정의됩니다. Gains는 (분위수의 적중 수 / 적중 수 총계) × 100%로 계산됩니다.

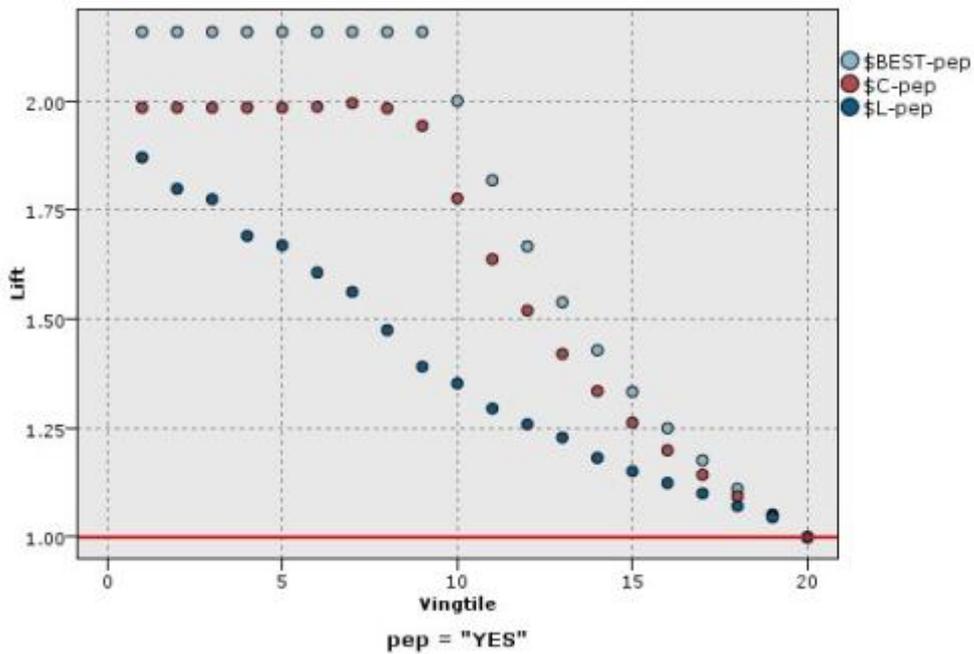
그림 1. 기준선, 최적 예측선 및 비즈니스 규칙이 표시된 Gains 차트(누적)



리프트 도표

리프트는 적중인 각 분위수의 레코드 백분율을 학습 데이터의 전체 적중 백분율과 비교합니다. 리프트는 (분위수의 적중 수 / 분위수의 레코드 수) / (적중 총계 / 레코드 총계)로 계산됩니다.

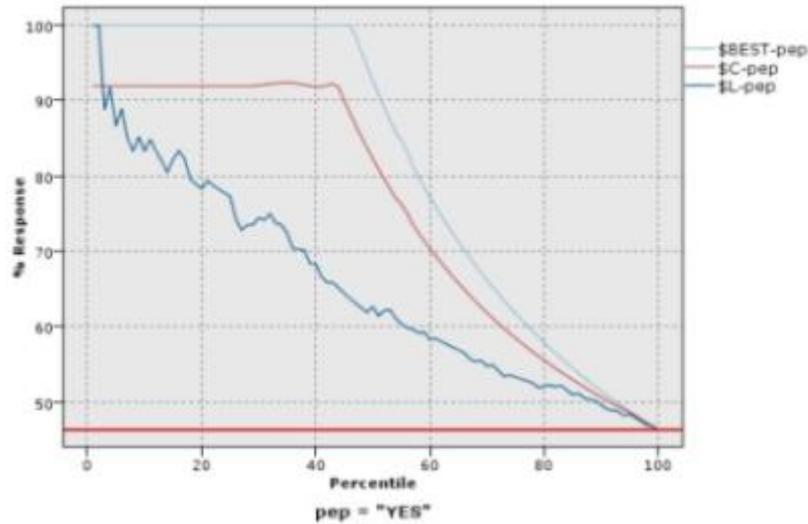
그림 2. 점 및 최적 예측선을 사용하는 리프트 도표(누적)



반응 차트

반응은 단순히 적중인 분위수의 레코드 백분율입니다. 반응은 (분위수의 적중 수 / 분위수의 레코드 수) × 100%로 계산됩니다.

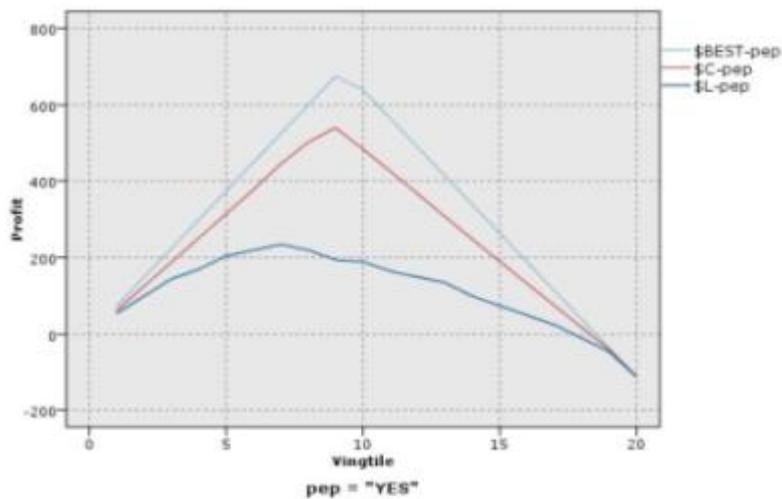
그림 3. 최적 예측선을 사용하는 반응 차트(누적)



이익 차트

이익은 각 레코드에 대한 수입에서 해당 레코드에 대한 비용을 뺀 값입니다. 분위수의 이익은 단순히 분위수의 전체 레코드 이익 합계입니다. 수입은 적중에만 적용되는 것으로 가정되지만 비용은 모든 레코드에 적용됩니다. 이익 및 비용은 고정이거나 데이터의 필드에 의해 정의될 수 있습니다. 이익은 (분위수의 레코드에 대한 수입의 합계 - 분위수의 레코드에 대한 비용의 합계)로 계산됩니다.

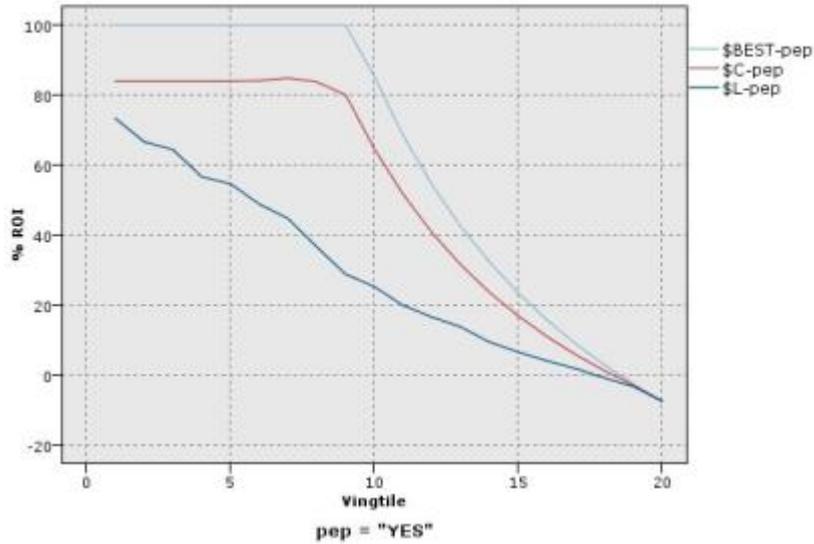
그림 4. 최적 예측선을 사용하는 이익 차트(누적)



ROI 차트

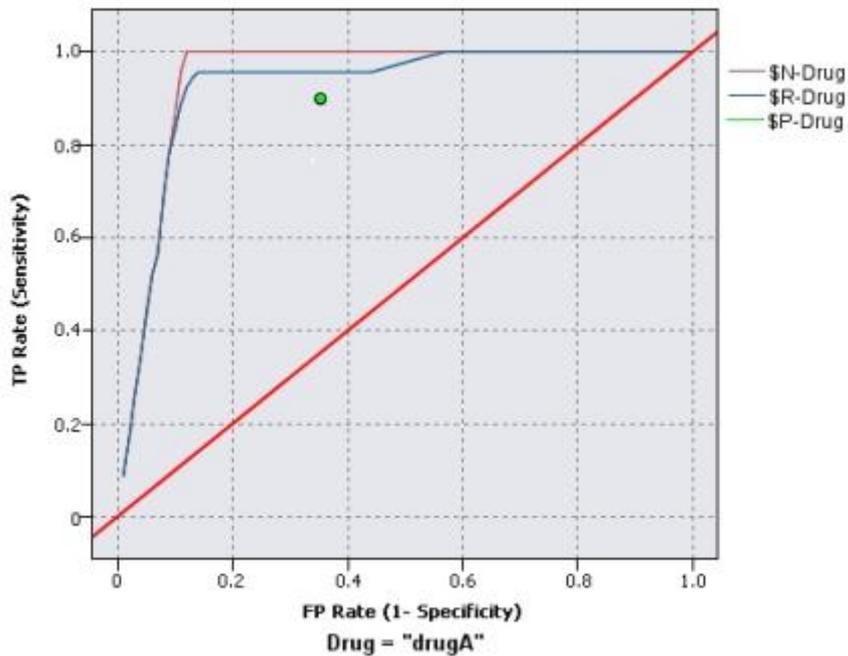
ROI(Return On Investment)는 수입 및 비용 정의를 포함한다는 점에서 이익과 비슷합니다. ROI는 분위수에 대한 비용과 이익을 비교합니다. ROI는 (분위수에 대한 이익 / 분위수에 대한 비용) × 100%로 계산됩니다.

그림 5. 최적 예측선을 사용하는 ROI 차트(누적)



ROC 차트

그림 6. 최적 예측선을 사용하는 ROC 차트



ROC(Receiver Operator Characteristic)는 이분형 분류자와 함께만 사용할 수 있습니다. ROC는 분류자의 성능을 기반으로 분류자를 시각화하고 구성하고 선택하는 데 사용할 수 있습니다. ROC 차트는 분류자의 거짓 긍정 비율에 대해 참 긍정 비율(민감도)을 도표화합니다. ROC 차트는 이익(참 긍정)과 비용(거짓 긍정) 간 상대적인 균형을 보여줍니다. 참 긍정은 적중인 인스턴스이며 적중으로 분류됩니다. 따라서 참 긍정 비율은 참 긍정 수를 실제로 적중인 인스턴스 수로 나눠서 계산됩니다. 거짓 긍정은 빗나감인 인스턴스이며 적중으로 분류됩니다. 따라서 거짓 긍정 비율은 거짓 긍정 수를 실제로는 빗나감인 인스턴스 수로 나눠서 계산됩니다.

각각의 점이 해당 분위수에 대한 값에 더 높은 모든 분위수를 더한 값과 동일하도록 평가 차트는 누적일 수도 있습니다. 누적 차트가 일반적으로 모델의 전체 성능을 더 잘 전달하지만 비누적 차트가 모델에 대한 특정 문제점 영역 표시에서 뛰어날 수도 있습니다.

참고: 평가 노드에서는 필드 이름에 심표 사용이 지원되지 않습니다. 필드 이름에 심표가 포함된 경우 심표를 제거하거나, 필드 이름을 따옴표로 묶어야 합니다.

① 평가 도표 탭

차트 유형. 이익(Gains), 반응, 리프트, 이익(Profit), 투자수익률(ROI) 또는 ROC(Receiver Operator Characteristic) 유형 중 하나를 선택하십시오.

누적 도표. 누적 차트를 작성하려면 선택하십시오. 누적 차트에 각 분위수 및 더 높은 분위수에 대해 값이 표시됩니다. (ROC 차트에는 **누적 도표**를 사용할 수 없습니다.)

기준선 포함. 도표에 기준선을 포함하려면 선택하십시오. 이 기준선은 신뢰도가 관련이 없어지는, 적중 수에 대한 완전한 임의 분포를 표시합니다. (이익 및 ROI 차트에는 **기준선 포함**을 사용할 수 없습니다.)

최적 예측선 포함. 도표에 최적 예측선을 포함하려면 선택하십시오. 이 최적 예측선은 완벽한 신뢰도(적중 수 = 케이스 중 100%)를 표시합니다. (ROC 차트에는 **최적 예측선**을 사용할 수 없습니다.)

모든 차트 유형에 이익 기준 사용. 정규 적중 수 대신 평가 측도를 계산할 때 이익 기준(비용, 수입, 가중치)을 사용하려면 선택하십시오. 특정 숫자 대상을 포함한 모델의 경우(예: 제안에 응해 고객으로부터 얻은 수입을 예측하는 모델), 목표 필드 값은 적중 수보다 나은 모델 성능 측도를 제공합니다. 이 옵션을 선택하면 이익, 응답, 리프트 차트에 **비용, 수익, 가중치** 필드를 사용할 수 있습니다. 이 세 가지 차트 유형에 이익 기준을 사용하려면 **수입**을 목표 필드로, **비용**을 0.0으로 설정하여 이익이 수입과 같도록 해야 하고 모든 레코드가 적중 수로 계수되도록 사용자 정의된 적중 조건을 "참"으로 지정해야 합니다. (ROC 차트에는 **모든 차트 유형에 이익 기준 사용**을 사용할 수 없습니다.)

예측/예측자 필드를 찾을 때 사용. 해당 메타데이터를 사용하여 그래프에서 예측 필드를 검색하려면 **모델 출력 필드 메타데이터**를 선택하고, 이름을 기준으로 검색하려면 **필드 이름 형식**을 선택하십시오.

도표 스코어 필드. 스코어 필드 선택기를 사용하려면 이 선택란을 선택하십시오. 그런 다음 하나 이상의 범위 또는 연속 스코어 필드, 즉 엄격한 예측 모형은 아니지만 적중 성향이라는 점에서 레코드의 순위를 매기는 데 유용한 필드를 선택하십시오. 평가 노드는 하나 이상의 스코어 필드의 조합을 하나 이상의 예측 모형과 비교할 수 있습니다. 일반 예에서는 여러 RFM 필드를 최적 예측 모형과 비교합니다.

목표. 필드 선택기를 사용하여 목표 필드를 선택하십시오. 둘 이상의 값을 가진 명목 필드 또는 인스턴스화된 플래그를 선택하십시오.

참고: 이 목표 필드는 스코어 필드에만 적용 가능하며(예측 모형이 고유 대상 정의) 사용자 정의 적중 기준이 옵션 탭에 설정되어 있는 경우 무시됩니다.

파티션별 분할. 파티션 필드를 사용하여 레코드를 학습, 검정, 검증 표본으로 분할하는 경우, 이 옵션을 선택하여 각 파티션에 대해 개별 평가 차트를 표시하십시오. 자세한 정보는 파티션 노드의 내용을 참조하십시오.

참고: 파티션별로 분할하는 경우 파티션 필드에 널값이 있는 레코드가 평가에서 제외됩니다. 파티션 노드는 널값을 생성하지 않으므로 파티션 노드가 사용되는 경우 이는 문제가 되지 않습니다.

도표. 드롭 다운 목록에서 차트에 표시할 분위수의 크기를 선택하십시오. 옵션에는 **사분위수, 오분위수, 십분위수, 이십분위수, 백분위수, 천분위수**가 있습니다. (ROC 차트에는 **도표**를 사용할 수 없습니다.)

스타일. 선 또는 점을 선택하십시오.

ROC 차트를 제외한 모든 차트 유형의 경우 추가 제어를 사용하면 비용, 수입, 가중치를 지정할 수 있습니다.

- **비용.** 각 레코드와 연관된 비용을 지정합니다. **고정** 또는 **가변** 비용을 선택할 수 있습니다. 고정 비용의 경우 비용 값을 지정하십시오. 가변 비용의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 비용 필드로 선택하십시오. (ROC 차트에는 **비용**을 사용할 수 없습니다.)
- **수입.** 적중을 나타내는 각 레코드와 연관된 수입을 지정합니다. **고정** 또는 **가변** 비용을 선택할 수 있습니다. 고정 수입의 경우 수입 값을 지정하십시오. 가변 수입의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 수입 필드로 선택하십시오. (ROC 차트에는 **수입**을 사용할 수 없습니다.)

- **가중치.** 데이터의 레코드가 둘 이상의 단위를 표시하는 경우 빈도 가중치를 사용하여 결과를 조정할 수 있습니다. **고정** 또는 **가변** 가중치를 사용하여 각 레코드와 연관된 가중치를 지정하십시오. 고정 가중치의 경우 가중값(레코드별 노드 수)을 지정하십시오. 가변 가중치의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 가중 필드로 선택하십시오. (ROC 차트에는 **가중치**를 사용할 수 없습니다.)

② 평가 옵션 탭

평가 차트에 대한 옵션 탭은 차트에 표시되는 적중, 스코어링 기준 및 비즈니스 규칙 정의 시 유연성을 제공합니다. 모형 평가 결과를 내보내기 위한 옵션도 설정할 수 있습니다.

사용자 정의 적중. 적중을 표시하는 데 사용되는 사용자 정의 조건을 지정하려면 선택하십시오. 이 옵션은 값의 순서 및 목표 필드의 유형에서 관심 있는 결과를 추론할 때보다 관심 있는 결과를 정의하는 경우에 유용합니다.

조건. 위에서 **사용자 정의 적중**이 선택되면 적중 조건에 대해 CLEM 표현식을 지정해야 합니다. 예를 들어, @TARGET = "YES"는 목표 필드에 대한 Yes 값이 평가에서 적중으로 계수됨을 나타내는 유효한 조건입니다. 지정된 조건은 모든 목표 필드에 사용됩니다. 조건을 작성하려면 필드를 입력하거나 표현식 작성기를 사용하여 조건식을 생성하십시오. 데이터가 인스턴스화되는 경우에는 표현식 작성기에서 직접 값을 삽입할 수 있습니다.

사용자 정의 스코어. 스코어링 케이스를 분위수에 지정하기 전에 스코어링 케이스에 사용되는 조건을 지정하려면 선택하십시오. 기본 스코어는 예측값 및 신뢰도로부터 계산됩니다. 표현식 필드를 사용하여 사용자 정의 스코어링 표현식을 작성하십시오.

- **표현식.** 스코어링에 사용되는 CLEM 표현식을 지정하십시오. 예를 들어, 0-1 범위의 숫자 출력이 낮은 값이 높은 값보다 나은 것으로 정렬되는 경우 적중을 @TARGET < 0.5로 정의하고 연관된 스코어를 1 - @PREDICTED로 정의할 수 있습니다. 스코어 표현식에서는 숫자 값을 생성해야 합니다. 조건을 작성하려면 필드를 입력하거나 표현식 작성기를 사용하여 조건식을 생성하십시오.

비즈니스 규칙 포함. 관심 있는 기준을 반영하는 규칙 조건을 지정하려면 선택하십시오. 예를 들어, mortgage = "Y" and income >= 33000인 모든 케이스에 대해 규칙을 표시하길 원할 수 있습니다. 비즈니스 규칙이 차트에서 그려지고 키에서 규칙으로 레이블 지정됩니다. (**비즈니스 규칙 포함**은 REC 차트에 대해서는 지원되지 않습니다.)

- **조건.** 출력 차트에서 비즈니스 규칙을 정의하는 데 사용되는 CLEM 표현식을 지정하십시오. 단순히 필드를 입력하거나 표현식 작성기를 사용하여 조건식을 생성하십시오. 데이터가 인스턴스화되는 경우에는 표현식 작성기에서 직접 값을 삽입할 수 있습니다.

파일로 결과 내보내기. 모형 평가 결과를 구분된 텍스트 파일로 내보내려면 선택하십시오. 이 파일을 읽고 계산된 값에 대해 특수 분석을 수행할 수 있습니다. 내보내기를 위해 다음과 같은 옵션을 설정하십시오.

- **파일 이름.** 출력 파일의 파일 이름을 입력하십시오. 생략 기호 단추(...)를 사용하여 원하는 폴더를 찾아보십시오.
- **구분자.** 필드 구분자로 사용할 문자(예: 쉼표 또는 공백)를 입력하십시오.

필드 이름 포함. 필드 이름을 출력 파일의 첫 번째 행으로 포함하려면 이 옵션을 선택하십시오.

각 레코드 다음에 줄 바꾸기. 새로운 행에서 각각의 레코드를 시작하려면 이 옵션을 선택하십시오.

③ 평가 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

텍스트. 자동으로 생성된 텍스트 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

X 레이블. 자동으로 생성된 x 축(가로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

Y 레이블. 자동으로 생성된 y 축(세로) 레이블을 승인하거나 **사용자 정의**를 선택하여 레이블을 지정하십시오.

눈금선 표시. 기본적으로 선택되는 이 옵션은 더 쉽게 영역 및 밴드 절사 지점을 결정할 수 있게 하는 눈금선을 도표 또는 그래프 뒤에 표시합니다. 그래프 배경이 흰색인 경우가 아니면 눈금선은 항상 흰색으로 표시됩니다. 그래프 배경이 흰색이면 눈금선은 회색으로 표시됩니다.

④ 모형 평가의 결과 읽기

평가 차트의 해석은 차트 유형에 대한 특정 범위에 따라 다르지만 모든 평가 차트에 공통인 몇몇 특성이 있습니다. 누적 차트의 경우 더 높고 있는 선은 더 나은 모델을 표시합니다(특히 차트의 왼쪽에서). 많은 경우 여러 모델을 비교하면 선이 겹쳐 한 모델이 차트의 한 부분에서 더 높고 다른 모델이 차트의 다른 부분에서 더 높습니다. 이 경우에는 선택할 모델을 결정할 때 원하는 표본의 부분(x 축에서의 위치를 정의함)을 고려해야 합니다.

대부분의 비누적 차트는 매우 비슷합니다. 양호한 모델의 경우 비누적 차트는 차트의 왼쪽에서 더 높고 차트의 오른쪽에서 낮아야 합니다. (비누적 차트에 톱니 패턴이 표시되는 경우에는 분위수의 수를 줄여 그래프를 도표화하고 재실행하여 평탄하게 할 수 있습니다.) 차트 왼쪽의 내려간 부분 또는 오른쪽의 올라간 부분은 모델의 예측이 양호하지 않은 영역을 표시할 수 있습니다. 전체 그래프에서 평평한 선은 본질적으로 정보를 제공하지 않는 모델을 표시합니다.

Gains 차트. 누적 Gains 차트는 항상 0%에서 시작하여 왼쪽에서 오른쪽으로 이동하면서 100%에서 끝납니다. 양호한 모델의 경우 Gains 차트는 100%를 향해 가파르게 상승한 후 수평을 유지합니다. 정보를 제공하지 않는 모델은 왼쪽 하단에서 오른쪽 상단으로 대각선으로 진행합니다(기준선 포함이 선택된 경우 차트에 표시됨).

리프트 도표. 누적 리프트 도표는 1.0 이상에서 시작하여 왼쪽에서 오른쪽으로 이동함에 따라 1.0에 도달할 때까지 점진적으로 내려갑니다. 차트의 오른쪽 가장자리는 전체 데이터 세트를 나타내므로 데이터의 적중 수에 대한 누적 분위수의 적중 수 비율은 1.0입니다. 양호한 모델의 경우 리프트는 왼쪽에서 1.0보다 훨씬 위에서 시작하여 오른쪽으로 이동할 때 높은 위치에서 안정 상태를 유지한 후 차트 오른쪽에서 1.0을 향해 급격하게 하강해야 합니다. 정보를 제공하지 않는 모델의 경우에는 전체 그래프에 대해 선이 1.0 주위를 맴돕니다. (기준선 포함이 선택되면 참조를 위해 1.0에서 가로 선이 차트에 표시됩니다.)

반응 차트. 누적 응답 차트는 척도화를 제외하고 리프트 도표와 매우 비슷합니다. 반응 차트는 일반적으로 100% 근처에서 시작하여 차트의 오른쪽 가장자리에서 전체 응답률(적중 총계/레코드 총계)에 도달할 때까지 점진적으로 내려갑니다. 양호한 모델의 경우 선은 왼쪽에서 100% 또는 이에 근접한 값에서 시작하여 오른쪽으로 이동함에 따라 높은 위치에서 안정 상태를 유지한 후 차트 오른쪽에서 전체 반응률을 향해 급격하게 하강합니다. 정보를 제공하지 않는 모델의 경우에는 전체 그래프에 대해 선이 전체 반응률 주위를 맴돕니다. (기준선 포함이 선택되면 참조를 위해 전체 반응률에서 가로 선이 차트에 표시됩니다.)

이익 차트. 누적 이익 차트는 선택된 표본의 크기를 늘리면서 왼쪽에서 오른쪽으로 이동할 때 이익의 합계를 표시합니다. 이익 차트는 일반적으로 0 근처에서 시작하고 가운데에서 최대치 또는 높은 위치의 안정 상태에 도달할 때까지 오른쪽으로 이동하면서 점진적으로 증가한 후 차트의 오른쪽 가장자리를 향해 감소합니다. 양호한 모델의 경우 이익은 차트의 가운데 쪽에 잘 정의된 최대치를 표시합니다. 정보를 제공하지 않는 모델의 경우에는 선이 상대적으로 직선이며 적용되는 비용/수입 구조에 따라 증가하거나 감소하거나 수평을 유지할 수 있습니다.

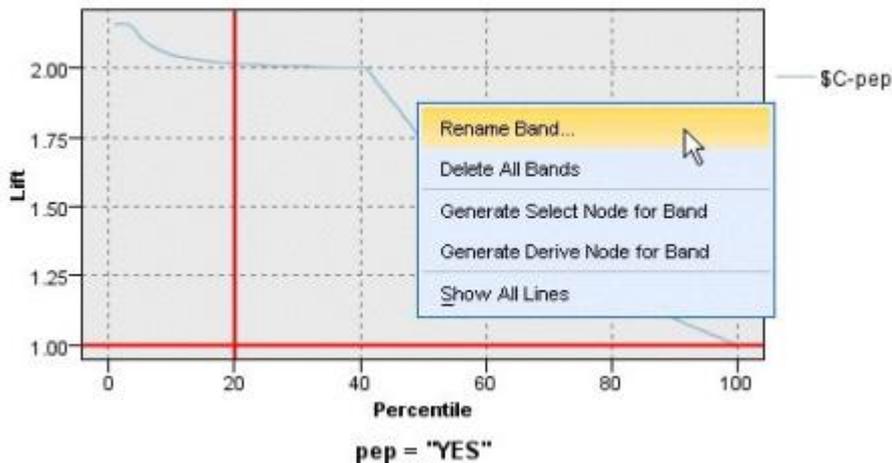
ROI 차트. 누적 ROI(Return On Investment) 차트는 척도화를 제외하고 반응 차트 및 리프트 도표와 비슷합니다. ROI 차트는 일반적으로 0% 이상에서 시작한 후 전체 데이터 세트에 대한 전체 ROI(음수가 될 수 있음)에 도달할 때까지 점진적으로 내려갑니다. 양호한 모델의 경우 선은 0%보다 훨씬 위에서 시작하고 오른쪽으로 이동함에 따라 높은 위치에서 안정 상태를 유지한 후 차트 오른쪽의 전체 ROI를 향해 급격하게 하강해야 합니다. 정보를 제공하지 않는 모델의 경우에는 선이 전체 ROI 값 주위를 맴돌아야 합니다.

ROC 차트. ROC 곡선은 일반적으로 누적 Gains 차트 모양을 가지고 있습니다. 곡선은 (0,0) 좌표에서 시작한 후 왼쪽에서 오른쪽으로 이동하면서 (1,1) 좌표에서 끝납니다. (0,1) 좌표를 향해 급격하게 상승한 후 수평을 유지하는 차트는 양호한 분류자를 표시합니다. 인스턴스를 무작위로 적중 또는 빗나감으로 분류하는 모델은 왼쪽 하단에서 오른쪽 상단으로 대각선으로 진행합니다 (기준선 포함이 선택된 경우 차트에 표시됨). 모델에 대해 신뢰도 필드가 제공되지 않으면 해당 모델은 단일 점으로 도표화됩니다. 분류의 최적 임계값을 가진 분류자는 차트의 (0,1) 좌표(또는 왼쪽 상단)에 가장 가까운 위치에 있습니다. 이 위치는 적중으로 올바르게 분류되는 많은 수의 인스턴스와 적중으로 잘못 분류되는 적은 수의 인스턴스를 나타냅니다. 대각선 위의 점은 양호한 분류 결과를 나타냅니다. 대각선 아래의 점은 인스턴스가 무작위로 분류된 경우보다 나쁜 양호하지 않은 분류 결과를 나타냅니다.

⑤ 평가 차트 사용

마우스를 사용하여 평가 차트를 탐색하는 것은 히스토그램 또는 컬렉션 그래프를 사용하는 것과 비슷합니다. x축은 이십분위수 또는 십분위수 등의 지정된 분위수에서 모델 스코어를 나타냅니다.

그림 1. 평가 차트에 대한 작업



분할자 아이콘을 사용하여 x축을 동등한 밴드로 자동으로 분할하는 옵션을 표시하여 히스토그램의 경우와 마찬가지로 x축을 밴드로 파티셔닝할 수 있습니다. 자세한 정보는 그래프 탐색의 내용을 참조하십시오. 편집 메뉴에서 **그래프 밴드**를 선택하여 밴드의 경계를 수동으로 편집할 수 있습니다.

평가 차트를 작성하고 밴드를 정의하고 결과를 검토한 후에는 컨텍스트 메뉴 및 생성 메뉴의 옵션을 사용하여 그래프의 선택사항을 기반으로 자동으로 노드를 작성할 수 있습니다. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

평가 차트에서 노드를 생성할 때 차트에서 사용 가능한 모든 모델 중에서 하나의 모델을 선택하라는 프롬프트가 표시됩니다.

모델을 선택한 후 **확인**을 클릭하여 새 노드를 스트림 캔버스에 생성하십시오.

(12) 맵 시각화 노드

맵 시각화 노드는 다중 입력 연결을 승인하고 지리 공간적 데이터를 맵에 일련의 레이어로 표시할 수 있습니다. 각각의 레이어는 하나의 지리 공간적 필드입니다. 예를 들어, 기존 레이어가 한 국가의 맵이고 그 위에 도로에 대한 레이어 하나, 강에 대한 레이어 하나, 도시에 대한 레이어 하나가 있을 수 있습니다.

대부분의 지리 공간적 데이터 세트는 일반적으로 하나의 지리 공간적 필드를 포함하고 있지만 하나의 입력에 여러 지리 공간적 필드가 있으면 표시할 필드를 선택할 수 있습니다. 동일한 입력 연결의 두 필드는 동시에 표시할 수 없습니다. 하지만 수신 연결을 복사하여 붙여넣고 각각으로부터 다른 필드를 표시할 수 있습니다.

① 맵 시각화 도표 탭

레이어

이 테이블에는 맵 노드에 대한 입력에 관한 정보가 표시됩니다. 레이어의 순서는 노드가 실행될 때 맵 미리보기와 시각적 출력 모두에서 레이어가 표시되는 순서를 지시합니다. 테이블의 맨 위 행이 '맨 위' 레이어이고 맨 아래 행이 '맨 아래' 레이어입니다. 즉, 각각의 레이어는 맵의 테이블에서 바로 아래에 있는 레이어 앞에 표시됩니다.

참고: 테이블의 레이어에 3차원 지리 공간적 필드가 포함되어 있으면 x축 및 y축만 도표화됩니다. z축은 무시됩니다.

이름

이름은 각 레이어에 대해 자동으로 작성되며 tag[source node:connected node] 형식을 사용하여 구성됩니다. 기본적으로 태그는 숫자로 표시되며 1은 연결되는 첫 번째 입력을 나타내고 2는 두 번째 입력을 나타내는 방식으로 표시됩니다. 필요한 경우 **레이어 편집** 단추를 눌러 맵 레이어 옵션 변경 대화 상자에서 태그를 변경하십시오. 예를 들어, 태그가 "도로" 또는 "구/군/시"가 되도록 변경하여 데이터 입력을 반영할 수 있습니다.

유형

레이어로 선택되는 지리 공간적 필드의 측정 유형 아이콘을 표시합니다. 입력 데이터에 지리 공간적 측정 유형을 가진 여러 필드가 포함되어 있는 경우 기본 선택사항에서는 다음 정렬 순서를 사용합니다.

1. 점
2. 선 스트링
3. 다각형
4. 다중 점
5. 다중 선 스트링
6. 복수 다각형

참고: 동일한 측정 유형을 가진 두 개의 필드가 있으면 첫 번째 필드(이름별 알파벳순)가 기본적으로 선택됩니다.

기호

참고: 이 열은 점 및 다중 점 필드의 경우에만 완료됩니다.

점 또는 다중 점 필드에 사용되는 기호를 표시합니다. 필요한 경우 **레이어 편집** 단추를 눌러 맵 레이어 옵션 변경 대화 상자에서 기호를 변경하십시오.

색상

맵에서 레이어를 나타내는 데 사용되는 색상을 표시합니다. 필요한 경우 **레이어 편집** 단추를 눌러 맵 레이어 옵션 변경 대화 상자에서 색상을 변경하십시오. 색상은 측정 유형에 따라 다양한 항목에 적용됩니다.

- 점 또는 다중 점의 경우 색상은 레이어에 대한 기호에 적용됩니다.
- 선 스트링 및 다각형의 경우 색상은 전체 모양에 적용됩니다. 다각형은 항상 검은색 윤곽선을 가지고 있습니다. 열에 표시되는 색상은 모양을 채우는 데 사용되는 색상입니다.

미리보기

이 분할창에는 **레이어** 테이블에서의 현재 입력 선택사항에 대한 미리보기가 표시됩니다. 미리보기는 레이어의 순서, 기호, 색상 및 레이어와 연관된 기타 표시 설정을 고려하며 가능한 경우 설정이 변경될 때마다 표시를 업데이트합니다. 스트림의 다른 위치(예: 레이어로 사용할 지리 공간적 필드)에서 세부사항을 변경하거나 연관된 집계 함수 등의 세부사항을 수정하는 경우에는 **데이터 새로 고치기** 단추를 클릭하여 미리보기를 업데이트해야 할 수 있습니다.

스트림을 실행하기 전에 **미리보기**를 사용하여 표시 설정을 설정하십시오. 큰 데이터 세트를 사용할 때 발생할 수 있는 시간 지연을 방지하기 위해 미리보기에서는 각각의 레이어에 대한 표본을 추출하고 처음 100개 레코드로부터 표시를 작성합니다.

가. 맵 레이어 변경

맵 레이어 옵션 변경 대화 상자를 사용하여 시각화 노드의 **도표** 탭에 표시되는 레이어의 다양한 세부사항을 수정할 수 있습니다.

입력 세부사항

태그

기본적으로 태그는 숫자입니다. 이 숫자를 더 의미 있는 태그로 바꿔 맵에서 레이어 식별을 지원할 수 있습니다. 예를 들어, 태그는 데이터 입력의 이름일 수 있습니다(예: "구/군/시").

레이어 필드

입력 데이터에 둘 이상의 지리 공간적 필드가 있는 경우 이 옵션을 사용하여 맵에 레이어로 표시할 필드를 선택하십시오.

기본적으로 선택할 수 있는 레이어는 다음과 같은 순서로 정렬되어 있습니다.

- 점
- 선 스트링
- 다각형
- 다중 점
- 다중 선 스트링
- 복수 다각형

표시 설정

육각형 구간화 사용

 **참고:** 이 옵션은 점 및 다중 점 필드에만 영향을 미칩니다.

육각형 구간화에서는 x 및 y 좌표를 기반으로 인접한 점을 단일 점으로 결합하여 맵에 표시합니다. 단일 점은 육각형으로 표시되지만 사실상 다각형으로 렌더링됩니다.

육각형은 다각형으로 렌더링되므로 육각형 구간화가 켜진 점 필드는 모두 다각형으로 처리됩니다. 이는 맵 노드 대화 상자에서 **유형별 정렬**을 선택하면 육각형 구간화가 적용된 점 레이어는 모두 다각형 레이어 위와 선 스트링 및 점 레이어 아래에 렌더링됨을 의미합니다.

다중 점 필드에 대해 육각형 구간화를 사용하는 경우에는 먼저 중심 점을 계산하기 위해 다중 점 값을 구간화하여 해당 필드가 점 필드로 변환됩니다. 중심 점은 육각형 구간을 계산하는 데 사용됩니다.

통합

 **참고:** 이 열은 **육각형 구간화 사용** 선택란을 선택하고 **오버레이**도 선택하는 경우에만 사용할 수 있습니다.

육각형 구간화를 사용하는 점 레이어에 대해 **오버레이** 필드를 선택하는 경우에는 육각형 내 모든 점에 대해 해당 필드에 있는 모든 값을 통합해야 합니다. 맵에 적용할 오버레이 필드에 대한 집계 함수를 지정하십시오. 사용 가능한 집계 함수는 측정 유형에 따라 다릅니다.

- 실수 또는 정수 저장 공간을 가진 연속형 측정 유형에 대한 집계 함수:
 - 합계
 - 평균
 - 최소값

- 최대값
 - 중앙값
 - 첫 번째 사분위수
 - 세 번째 사분위수
- 시간, 날짜 또는 시간소인 저장 공간을 가진 연속형 측정 유형에 대한 집계 함수:
- 평균
 - 최소값
 - 최대값
- 명목 또는 범주형 측정 유형에 대한 집계 함수:
- 모드
 - 최소값
 - 최대값
- 플래그 측정 유형에 대한 집계 함수:
- 참(참인 항목이 있는 경우)
 - 거짓(거짓인 항목이 있는 경우)

색상

데이터에 있는 다른 필드의 값을 기반으로 기능에 색상을 지정하는 오버레이 필드 또는 지리 공간적 필드의 모든 기능에 적용할 표준 색상을 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 사용자 옵션 대화 상자의 표시 탭에 있는 **차트 범주 색상 순서** 분할창에 표시되는 색상의 팔레트에서 색상을 선택할 수 있습니다.

추가 정보는 표시 옵션 설정의 내용을 참조하십시오.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 필드를 선택할 수 있습니다.

- 명목 또는 범주형 오버레이 필드의 경우 선택할 수 있는 색상 팔레트는 **표준** 색상 옵션에 대해 표시되는 것과 동일합니다.
- 연속형 및 순서 오버레이 필드의 경우에는 두 번째 드롭 다운 목록이 표시되고 여기서 색상을 선택합니다. 색상을 선택하면 연속형 또는 순서 필드의 값에 따라 해당 색상의 채도가 변경되어 오버레이가 적용됩니다. 가장 높은 값은 드롭 다운 목록에서 선택된 색상을 사용하고 더 낮은 값은 더 낮은 채도로 표시됩니다.

기호

 **참고:** 점 및 다중 점 측정 유형에 대해서만 사용으로 설정됨.

지리 공간적 필드의 모든 레코드에 적용되는 **표준** 기호를 사용할지 아니면 데이터에 있는 다른 필드의 값을 기반으로 점에 대한 기호 아이콘을 변경하는 **오버레이** 기호를 사용할지 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 드롭 다운 목록에서 기본 기호 중 하나를 선택하여 맵에 점 데이터를 나타낼 수 있습니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 명목, 순서 또는 범주형 필드를 선택할 수 있습니다. 오버레이 필드의 각각의 값에 대해 다른 기호가 맵에 표시됩니다.

예를 들어, 데이터에 상점의 위치를 나타내는 점 필드가 포함되어 있고 오버레이가 상점 유형 필드일 수 있습니다. 이 예제에서 모든 식품 상점은 맵에서 십자 기호로 식별되고 모든 전자제품 상점은 사각형 기호로 식별될 수 있습니다.

크기

 **참고:** 점, 다중 점, 선 스트링 및 다중 선 스트링 측정 유형에 대해서만 사용으로 설정됨.

지리 공간적 필드의 모든 레코드에 적용되는 **표준** 크기를 사용할지 아니면 데이터에 있는 다른 필드의 값을 기반으로 선 굵기 또는 기호 아이콘의 크기를 변경하는 **오버레이** 크기를 사용할지 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 픽셀 너비 값을 선택할 수 있습니다. 사용 가능한 옵션은 1, 2, 3, 4, 5, 10, 20 또는 30입니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 필드를 선택할 수 있습니다. 점 또는 선의 굵기는 선택한 필드의 값에 따라 다릅니다.

투명도

지리 공간적 필드의 모든 레코드에 적용되는 **표준** 투명도를 사용할지 아니면 데이터에 있는 다른 필드의 값을 기반으로 기호, 선 또는 다각형의 투명도를 변경하는 **오버레이** 투명도를 사용할지 선택하려면 이 옵션을 사용하십시오.

표준을 선택하는 경우에는 0%(불투명)에서 시작하여 10%씩 증분되어 100%(투명)까지 증가하는 투명도 수준 선택사항 중에서 선택할 수 있습니다.

오버레이를 선택하는 경우에는 **레이어 필드**로 선택된 지리 공간적 필드가 포함된 데이터 소스에서 필드를 선택할 수 있습니다. 오버레이 필드의 각각의 값에 대해 다른 투명도 수준이 맵에 표시됩니다. 투명도는 점, 선 또는 다각형에 대해 색상 드롭 다운 목록에서 선택한 색상에 적용됩니다.

데이터 레이블

참고: 육각형 구간화 사용 선택란을 선택하는 경우 이 옵션은 사용할 수 없습니다.

맵에서 데이터 레이블로 사용할 필드를 선택하려면 이 옵션을 사용하십시오. 예를 들어, 다각형 레이어에 적용된 경우 데이터 레이블은 각 다각형의 이름이 포함된 이름 필드일 수 있습니다. 이름 필드를 선택하면 해당 이름이 맵에 표시됩니다.

② 맵 시각화 모양 탭

그래프를 작성하기 전에 모양 옵션을 지정할 수 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

캡션. 그래프 캡션에 사용할 텍스트를 입력하십시오.

(13) t-SNE 노드

t-SNE(t-Distributed Stochastic Neighbor Embedding)는 고차원 데이터를 시각화하기 위한 도구입니다. 이는 데이터 점의 연관관계를 확률로 변환합니다. 원래 공간의 연관관계가 가우스 결합 확률에 의해 표현되고 임베드된 공간의 연관관계가 스튜던트 T-분산에 의해 표현됩니다. 이로 인해 t-SNE가 로컬 구조에 특히 민감할 수 있으며 기존 기술에 비해 몇 가지 기타 장점을 갖게 됩니다.¹⁾

- 단일 맵의 많은 척도에서 구조 표시
- 다중, 이형, 매니폴드 또는 군집에 있는 데이터 표시
- 중심에서 함께 복잡한 포인트로 경향성 저하

SPSS® Modeler에서 t-SNE 노드는 Python으로 구현되며 scikit-learn® Python 라이브러리가 필요합니다. t-SNE 및 scikit-learn 라이브러리에 대한 세부사항은 다음을 참조하십시오.

- <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE>
- <https://scikit-learn.org/stable/modules/manifold.html#t-sne>

노드 팔레트의 Python 탭은 이 노드와 다른 Python 노드로 구성됩니다. t-SNE 노드는 그래프 탭에서도 사용할 수 있습니다.

1) 참조:

van der Maaten, L.J.P.; Hinton, G. "Visualizing High-Dimensional Data using t-SNE." Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. "t-Distributed Stochastic Neighbor Embedding."

van der Maaten, L.J.P. "Accelerating t-SNE using Tree-Based Algorithms." Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

① t-SNE 노드 고급 옵션

t-SNE 노드에 설정할 옵션에 따라 **단순** 모드 또는 **고급** 모드를 선택하십시오.

시각화 유형. 그래프를 2차원 또는 3차원으로 그릴지 지정하려면 **2D** 또는 **3D**를 선택하십시오.

방법. Barnes Hut 또는 **정확**을 선택하십시오. 기본적으로 기울기 계산 알고리즘은 Barnes-Hut 근사값을 사용하며 이 방법은 정확 방법보다 더 빠릅니다. Barnes-Hut 근사값을 사용하면 t-SNE 기술을 실제의 대규모 데이터 세트에 적용할 수 있습니다. 정확 알고리즘은 가장 가까운 이웃 항목을 피하는 더 나은 작업을 수행합니다.

초기화. 임베딩의 초기화에 대해 **난수** 또는 **PCA**를 선택하십시오.

대상 필드. 출력 그래프에 색상표로 표시할 대상 필드를 선택하십시오. 여기서 대상 필드를 지정하지 않을 경우 그래프에서 단색을 사용합니다.

최적화

혼란. 혼란은 다른 매니폴드 학습 알고리즘에서 사용되는 가장 가까운 이웃 항목 수와 관련되어 있습니다. 대개 데이터 세트가 더 클수록 더 큰 혼란이 필요합니다. 5에서 50 사이의 값을 선택하는 것이 좋습니다. 기본값은 30이고, 범위는 2 - 9999999입니다.

조기 과장. 이 설정은 원래 공간의 기본 군집이 임베드 공간에 얼마나 조밀하게 있고 그 사이에 얼마나 많은 공간이 있는지를 제어합니다. 기본값은 12이고, 범위는 2 - 9999999입니다.

학습률. 학습률이 너무 높으면 데이터는 모든 포인트가 최근접 이웃에서 대략적으로 같은 거리에 있는 "볼"로 보입니다. 학습률이 너무 낮으면, 대부분의 포인트는 이상값이 거의 없는 낮은 밀도의 구름으로 압축되어 보일 수 있습니다. 비용 함수가 잘못된 로컬 최소값에서 막히면 학습률을 늘리는 것이 도움이 될 수 있습니다. 기본값은 200이고, 범위는 0 - 9999999입니다.

최대 반복 수. 최적화를 위한 최대 반복 수입니다. 기본값은 1000이고, 범위는 250 - 9999999입니다.

각도 크기. 한 점에서 측정된 멀리 떨어진 노드의 각도 크기입니다. 0과 1 사이의 값을 입력하십시오. 기본값은 0.5입니다.

난수 시드

난수 시드 설정. 난수 생성기에서 사용한 시드를 생성하려면 이 옵션을 선택하고 **생성**을 클릭하십시오.

최적화 중단 조건

진행률 없는 최대 반복 수. 최적화를 중지하기 전에 진행하지 않은 최대 반복 수로서, 초기 과장에서 250번 초기 반복 후 사용됩니다. 진행률은 50번 반복할 때마다 확인하므로, 이 값은 다음 50의 배수로 반올림됩니다. 기본값은 300이고, 범위는 0 - 9999999입니다.

최소 기울기 노름. 기울기 노름이 이 최소 임계값 이하일 경우 최적화가 중지됩니다. 기본값은 1.0E-7입니다.

메트릭. 기능 배열에서 인스턴스 사이의 거리를 계산할 때 사용할 메트릭입니다. 메트릭이 문자열이면 메트릭 모수 또는 pairwise.PAIRWISE_DISTANCE_FUNCTIONS에 나열된 메트릭의 경우 scipy.spatial.distance.pdist에 허용된 옵션 중 하나여야 합니다. 사용 가능한 메트릭 유형 중 하나를 선택하십시오. 기본값은 유클리디안입니다.

레코드 수가 다음보다 많은 경우. 큰 데이터 세트를 도표화하는 방법을 지정하십시오. 최대 데이터 세트 크기를 지정하거나 기본값인 2,000개의 점을 사용할 수 있습니다. 구간 또는 표본 옵션을 선택하면 큰 데이터 세트에 대해 성능이 개선됩니다. 또는 모든 데이터 사용을 선택하여 모든 데이터 포인트를 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격하게 저하될 수 있다는 점에 유의해야 합니다.

- **구간.** 데이터 세트에 지정된 수의 레코드보다 많은 레코드가 포함되어 있는 경우 구간화를 사용하여 설정하려면 선택하십시오. 구간화는 실제로 도표화하기 전에 그래프를 세분화된 눈금으로 나누고 각각의 눈금 셀에 표시되는 연결의 수를 계수합니다. 최종 그래프에서는 구간 중심 값(구간에 있는 모든 연결 점의 평균)에서 셀당 하나의 연결이 사용됩니다.
- **표본** 지정된 수의 레코드로 데이터에서 무작위로 표본을 추출하려면 선택하십시오.

다음 표에서는 SPSS® Modeler t-SNE 노드 대화 상자의 고급 탭에 있는 설정과 Python t-SNE 라이브러리 모수 간의 관계를 보여줍니다.

표 1. Python 라이브러리 모수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	Python t-SNE모수
모드	mode_type	
시각화 유형	n_components	n_components
방법	method	method
임베드 초기화	init	init
목표	target_field	target_field
혼란	perplexity	perplexity
조기 과장	early_exaggeration	early_exaggeration

SPSS Modeler 설정	스크립트 이름(특성 이름)	Python t-SNE모수
학습률	learning_rate	learning_rate
최대 반복 수	n_iter	n_iter
각도 크기	angle	angle
난수 시드 설정	enable_random_seed	
난수 시드	random_seed	random_state
진행률 없는 최대 반복	n_iter_without_progress	n_iter_without_progress
최소 기울기 노름	min_grad_norm	min_grad_norm
다중 혼란을 사용하여 t-SNE 수행	isGridSearch	

② t-SNE 노드 출력 옵션

출력 탭에서 t-SNE 노드 출력에 대한 옵션을 지정하십시오.

출력 이름. 노드가 실행될 때 생성되는 출력의 이름을 지정합니다. **자동**을 선택하면 출력의 이름이 자동으로 설정됩니다.

화면으로 출력. 새 창에서 출력을 생성하고 표시하려면 이 옵션을 선택하십시오. 또한 출력이 출력 관리자에 추가됩니다.

파일로 출력. 출력을 파일에 저장하려면 이 옵션을 선택하십시오. 그러면 **파일 이름** 및 **파일 유형** 필드를 사용할 수 있습니다. 비교 목적으로 기타 필드를 사용하여 도표를 작성하거나 분류 또는 회귀 모델 내의 예측자로서 출력을 사용하려면 t-SNE 노드에 이 출력 파일에 대한 액세스가 필요합니다. t-SNE 모델은 고정된 파일 소스 노드를 사용하여 가장 쉽게 액세스할 수 있는 x, y (및 z) 좌표 필드의 결과 파일을 작성합니다.

③ t-SNE 데이터 액세스 및 도표화

파일에 출력 옵션을 사용하여 파일에 t-SNE 출력을 저장하는 경우, 비교 목적으로 기타 필드를 사용하여 도표를 작성하거나 분류 또는 회귀 모델 내의 예측자로서 출력을 사용할 수 있습니다. t-SNE 모델은 고정된 파일 소스 노드를 사용하여 가장 쉽게 액세스할 수 있는 x, y (및 z) 좌표 필드의 결과 파일을 작성합니다. 이 절에서는 예 정보를 제공합니다.

1. t-SNE 노드 대화 상자에서 **출력** 탭을 여십시오.
2. **파일에 출력**을 선택하고 파일 이름을 입력하십시오. 기본 HTML 파일 유형을 사용하십시오. 모델을 실행할 때 사용자의 결과 위치에 세 개의 출력 파일을 생성합니다.

- 텍스트 파일(result_xxxxxx.txt)
- HTML 파일(사용자가 지정한 파일 이름)
- PNG 파일(tsne_chart_yyyyyy.png)

텍스트 파일은 사용자가 필요로 하는 데이터를 포함할 수 있으나 기술적인 이유로 표준 또는 지수표기 형식일 수 있습니다. 지수표기 형식(1.11111111e+01)인 경우, 형식을 인식하는 새 스트림을 작성해야 합니다.

텍스트 파일이 지수표기 숫자 형식인 경우의 t-SNE 도표 데이터 액세스

1. 새 스트림(파일 > 새 스트림)을 작성하십시오.
2. 도구 > 스트림 특성 > 옵션으로 이동하여 숫자형식을 선택하고 숫자 표시 형식으로 지수표기 (#.###E+##)를 선택하십시오.
3. 고정된 파일 소스 노드를 캔버스에 추가하고 파일 탭에서 다음 설정을 사용하십시오.
 - 헤더 행 건너뛰기: 1
 - 레코드 길이: 54
 - tSNE_x 시작: 3, 길이 16
 - tSNE_y 시작: 20, 길이: 16
 - tSNE_z 시작: 36, 길이: 16
4. 유형 탭에서 숫자가 실수로 인식되어야 합니다. 읽기 값을 클릭하면 다음과 유사한 필드 값이 표시되어야 합니다.

표 1. 예 필드 값

필드	측정	값
tSNE_x	연속	[-7.07176703,7.14338837]
tSNE_y	연속	[-9.2188112,8.89647667]
tSNE_x	연속	[-9.95892882,9.95742482]

5. 스트림에 선택 노드를 추가하여 널리 읽히는 파일 내 텍스트의 다음 아래쪽 두 행을 삭제할 수 있습니다.

```
*****
Perform t-SNE (total time 9.5s)
```

선택 노드의 설정 탭에서 모드에 대해 삭제를 선택하고 @NULL(tSNE_x) 조건을 사용하여 행을 삭제하십시오.

6. 유형 노드 및 플랫 파일 내보내기 노드를 스트림에 추가하여 원래 스트림에 복사하여 붙여넣기될 Var. 파일을 작성하십시오.

텍스트 파일이 표준 숫자 형식인 경우의 t-SNE 도표 데이터 액세스

1. 새 스트림(파일 > 새 스트림)을 작성하십시오.
2. 고정된 파일 소스 노드를 캔버스에 추가하십시오. 다음과 같은 세 가지 노드가 t-SNE 데이터에 액세스하는 데 모두 필요합니다.

그림 1. Stream for accessing t-SNE plot data in standard numeric format



3. 고정된 파일 소스 노드의 파일 탭에서 다음 설정을 사용하십시오.
 - 헤더 행 건너뛰기: 1
 - 레코드 길이: 29
 - tSNE_x 시작: 3, 길이 12
 - tSNE_y 시작: 16, 길이: 12
4. 필터 탭에서 field1 및 field2를 tsneX 및 tsneY로 이름을 변경할 수 있습니다.
5. 순서 합치기 방법을 사용하여 합치기 노드를 추가하여 이를 스트림에 연결하십시오.
6. 이제 plot 노드를 사용하여 tsneX 대 tsneY를 도표로 작성하고 조사 중인 필드를 사용하여 색상을 지정할 수 있습니다.

④ t-SNE 모델 너깃

t-SNE 모델 너깃은 t-SNE 모델이 캡처한 모든 정보를 포함합니다. 다음 탭을 사용할 수 있습니다.

그래프

그래프 탭은 t-SNE 노드에 대한 차트 출력을 표시합니다. pyplot 산점도 차트는 최저 차원 결과를 표시합니다. t-SNE 노드의 고급 탭에서 **다중 당혹도를 사용하여 t-SNE 수행** 옵션을 선택하지 않은 경우, 당혹도가 다른 여섯 개의 그래프가 아니라 한 개의 그래프만 포함됩니다.

텍스트 출력

텍스트 출력 탭은 t-SNE 알고리즘의 결과를 표시합니다. t-SNE 노드의 고급 탭에서 **2D 시각화**를 선택한 경우, 여기의 결과가 2차원의 포인트 값입니다. **3D**를 선택한 경우, 결과가 3차원의 포인트 값입니다.

(14) E-Plot(베타) 노드

E-Plot(베타) 노드는 수치 필드 사이의 관계를 보여줍니다. E-Plot(베타) 노드는 Plot 노드와 유사하나 옵션이 다르며 새 그래프 기능을 사용합니다. SPSS® Modeler에서 노드를 사용하여 새 그래프 기능을 활용할 수 있습니다.

E-Plot(베타) 노드는 산점도, 선도표 및 막대형 차트를 제공하여 수치 필드 사이의 관계를 설명합니다. 이 노드의 새 그래프 인터페이스는 직관적이며 현대적이며 높은 수준의 사용자 정의가 가능하며 데이터 도표가 대화형입니다. 추가 정보는 E-Plot 그래프 사용의 내용을 참조하십시오.

① E-Plot(베타) 노드 도표 탭

도표는 Y 필드의 값 대 X 필드의 값을 표시합니다. 종종 이러한 필드는 각각 종속변수 및 독립 변수에 해당됩니다.

X 필드. 목록에서 수평 x축으로 표시할 필드를 선택합니다.

Y 필드. 목록에서 수직 y축으로 표시할 필드를 선택합니다.

오버레이. 여러 가지 방식으로 데이터 값에 대한 범주를 표시할 수 있습니다. 예를 들어, *maincrop* 필드를 색상 오버레이로 사용하여 클레임 지원자가 키운 주요 작물에 대한 *estincome* 및 *claimvalue* 값을 표시할 수 있습니다. 출력에서 색상 매핑, 크기 매핑 및 모양 매핑에 대한 필드를 선택하십시오. 또한 대화형 출력에 포함할 기타 관심 있는 필드로 선택하십시오. 자세한 정보는 모양, 오버레이, 패널 및 애니메이션의 내용을 참조하십시오.

E-Plot에 옵션을 설정한 후에는 **실행**을 클릭하여 대화 상자에서 직접 도표를 실행할 수 있습니다. 추가 지정 사항에 대한 옵션 탭을 사용해야 하는 경우도 있습니다.

② E-Plot(베타) 노드 옵션 탭

도표화할 최대 레코드 수. 큰 데이터 세트를 도표화하는 방법을 지정하십시오. 최대 데이터 세트 크기를 지정하거나 기본값인 2,000개의 레코드를 사용할 수 있습니다. **표본** 옵션을 선택하면 큰 데이터 세트에 대한 성능이 개선됩니다. 샘플 옵션은 텍스트 필드에서 입력한 레코드 수까지 무작위로 데이터의 표본을 추출합니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 점을 도표화하도록 선택할 수 있지만 소프트웨어의 성능이 급격하게 저하될 수 있다는 점에 유의해야 합니다.

③ E-Plot(베타) 외형 탭

원하는 경우, 그래프 작성 전에 제목 및 부제목을 지정할 수 있습니다. 이러한 옵션은 그래프 작성 후에 지정하거나 변경할 수도 있습니다.

제목. 그래프 제목에 사용할 텍스트를 입력하십시오.

부제목. 그래프 부제목에 사용할 텍스트를 입력하십시오.

④ E-Plot 그래프 사용

E-Plot(베타) 노드는 산점도, 선도표 및 막대형 차트를 제공하여 수치 필드 사이의 관계를 설명합니다. 이 베타 노드에 도입된 새 그래프 인터페이스는 다양한 새 기능 및 개선된 기능을 포함합니다.

그림 1. E-Plot (Beta) scatterplot graph



그래프 탭의 왼쪽 상단 모서리에서 차트의 특정 섹션을 확대하거나 초기의 전체 보기로 다시 돌아가도록 축소하거나 차트를 외부에서 사용하도록 저장할 수 있는 도구 모음을 제공합니다.

그림 2. Toolbar



창의 아래쪽에서 슬라이더를 사용하여 차트의 특정 섹션을 확대할 수 있습니다. 작은 직사각형 제어를 오른쪽 및 왼쪽으로 이동하여 확대/축소할 수 있습니다. 이 슬라이더를 사용하려면 먼저 도구 상자 옵션 영역에서 이를 켜야 합니다.

그림 3. Zoom slider



창의 왼쪽 방향은 표시되는 값의 변위를 변경할 수 있는 제어를 제공합니다. 해당 제어를 사용하려면 먼저 데이터 맵핑 옵션 영역에서 옵션을 지정해야 합니다. 아래 예에서는 색상 맵핑용으로 PM25라는 필드가 선택되고 크기 맵핑용으로 PM10이라는 필드가 선택되고 모양 맵핑용으로 City라는 필드가 선택됩니다. 수직 색상 막대 위로 마우스를 이동하여 그래프의 해당 영역을 강조 표시하거나 삼각형을 위 또는 아래로 미십시오.

그림 4. Range controls



창의 오른쪽에서 일련의 확장 가능한 옵션을 사용하여 데이터와 상호작용하고 실시간으로 차트의 외형을 변경할 수 있습니다.

그림 5. Expandable options



기본 옵션

그림 6. Basic options

Theme	Dark
Title	My Title
Subtitle	My Subtitle
Chart Type	Scatter
Series	AQIndex
Basic	

어둡거나 밝은 테마를 선택하고 제목 및 부제목을 지정하고 도표 유형(산점도, 선형 또는 막대)을 선택하고 Y축에 표시되는 계열을 선택하십시오. 선도표를 선택하는 경우, Y축의 필드만 표시되고 색상 맵핑 및 크기 맵핑에 대한 데이터 맵핑 옵션에서 Y축의 필드만 사용 가능합니다. 막대형 차트를 선택하는 경우, 데이터 맵핑 옵션에서 색상 맵핑 옵션만 사용 가능합니다. 계열의 경우, E-Plot 노드의 도표 탭에서 선택한 모든 관심 있는 필드를 여기서 사용할 수 있습니다.

데이터 맵핑 옵션

그림 7. Data map options



색상 맵핑에 대해 연속형 필드 또는 범주형 필드를 선택하십시오. 연속형 필드가 선택되면 녹색에서 빨간색까지의 색상이 표시됩니다. 값을 낮추면 색상이 빨간색에 가까워지고 값을 높이면 색상이 녹색에 가까워집니다. 범주형 필드를 선택하면 정의된 색상표에 따라 필드 색상이 표시됩니다.

크기 맵핑은 연속형 필드만 지원합니다. 차트에서 값을 낮출수록 도표 크기가 작아집니다.

모양 맵핑은 범주형 필드만 지원합니다. 맵에 표시되는 모양은 시각화를 각 범주에 하나씩 다른 모양의 요소로 분할하는 범주형 필드에 의해 정의됩니다.

팔레트 옵션

그림 8. Palette options



제목 및 계열에 사용되는 색상을 사용자 정의하려면 팔레트를 사용하십시오. 드롭 다운에서 제목 또는 계열을 선택하고 사전정의된 색상 편집을 클릭한 다음 기타를 클릭하여 색상을 선택하십시오. 또는 RGB 또는 16진수 필드를 사용하여 정확한 색상을 선택할 수 있습니다.

도구 상자 옵션

그림 9. Toolbox options



도구 상자 옵션을 사용하여 확대/축소 슬라이더를 켜거나 끄고 눈금선 특성을 설정하고 마우스 추적을 켜거나 끄십시오. 마우스 추적은 차트 위로 마우스를 이동할 때 정확한 도표 위치를 표시합니다.

(15) 그래프 출력에 대한 작업

스트림을 실행하고 그래프를 확보할 때 출력 창에서 직접 이 그래프와 상호작용하고 이 그래프를 편집할 수 있습니다. 예를 들어, 다음을 수행할 수 있습니다.

- 데이터를 분석하고 값을 식별하여 그래프 탐색. 밴드 및 영역을 그리거나 요소를 표시하여 선택, 파생 또는 균형 노드를 생성할 수 있습니다.

- 그래프의 레이아웃 및 모양을 변경하기 위해 그래프 편집
- 제목, 각주 또는 축 레이블 적용
- 스타일시트 업데이트 또는 변경
- 그래프 인쇄, 저장, 복사 또는 내보내기

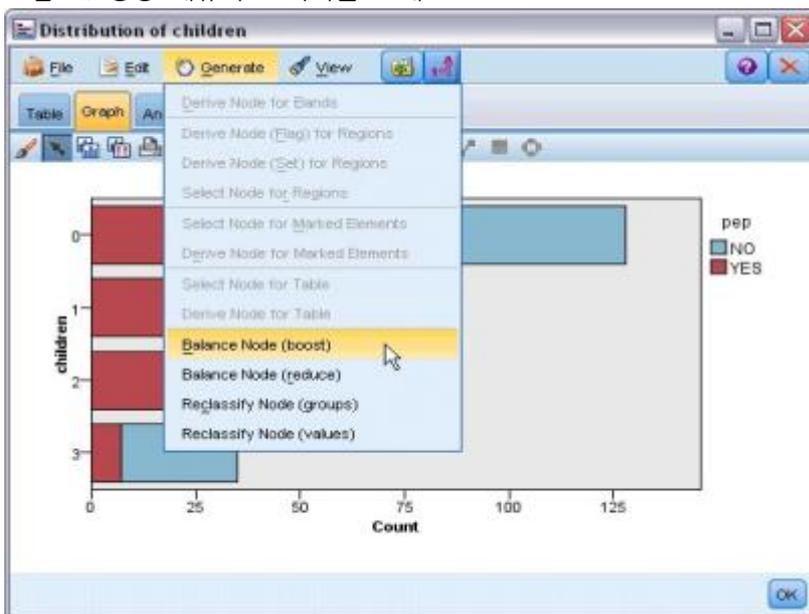
(16) 그래프 탐색

편집 모드를 사용하면 그래프의 레이아웃 및 모양을 편집할 수 있지만 탐색 모드를 사용하면 그래프로 표시되는 데이터 및 값을 분석적으로 탐색할 수 있습니다. 탐색의 주요 목표는 데이터를 분석한 후 밴드, 영역 및 표시를 통해 값을 식별하여 선택, 파생 또는 균형 노드를 생성하는 것입니다. 이 모드를 선택하려면 도구 모음 아이콘을 클릭하거나 메뉴에서 보기 > 탐색 모드를 선택하십시오.

일부 그래프는 모든 탐색 도구를 사용할 수 있지만 다른 그래프는 하나만 승인합니다. 탐색 모드는 다음을 포함합니다.

- 척도 x축을 따라 값을 분할하는 데 사용되는 밴드 정의 및 편집. 자세한 정보는 밴드 사용의 내용을 참조하십시오.
- 직사각형 영역에서 값 그룹을 식별하는 데 사용되는 영역 정의 및 편집. 자세한 정보는 영역 사용의 내용을 참조하십시오.
- 선택 또는 파생 노드를 생성하는 데 사용할 수 있는 값을 직접 선택하기 위해 요소 표시 또는 표시 해제. 자세한 정보는 표시된 요소 사용의 내용을 참조하십시오.
- 스트림에서 사용할 밴드, 영역, 표시된 요소 및 웹 링크에 의해 식별된 값을 사용하여 노드 생성. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

그림 1. 생성 메뉴가 표시되는 그래프



① 밴드 사용

x축에 척도 필드가 있는 그래프에서는 세로 밴드 라인을 그려 x축에서 값의 범위를 분할할 수 있습니다. 그래프에 패널이 여럿 있는 경우에는 한 패널에서 그려진 밴드 라인이 다른 패널에도 표시됩니다.

일부 그래프는 밴드를 승인하지 않습니다. 밴드를 가질 수 있는 그래프로는 히스토그램, 막대형 차트 및 분포, 도표(선, 산점도, 시간 등), 컬렉션 및 평가 차트 등이 있습니다. 패널이 있는 그래프에서는 밴드가 모든 패널에 표시됩니다. SPLOM에서는 필드/변수 밴드가 그려진 축이 플립 되었기 때문에 가로 밴드 라인이 표시됩니다.

그림 1. 3개의 밴드가 있는 그래프

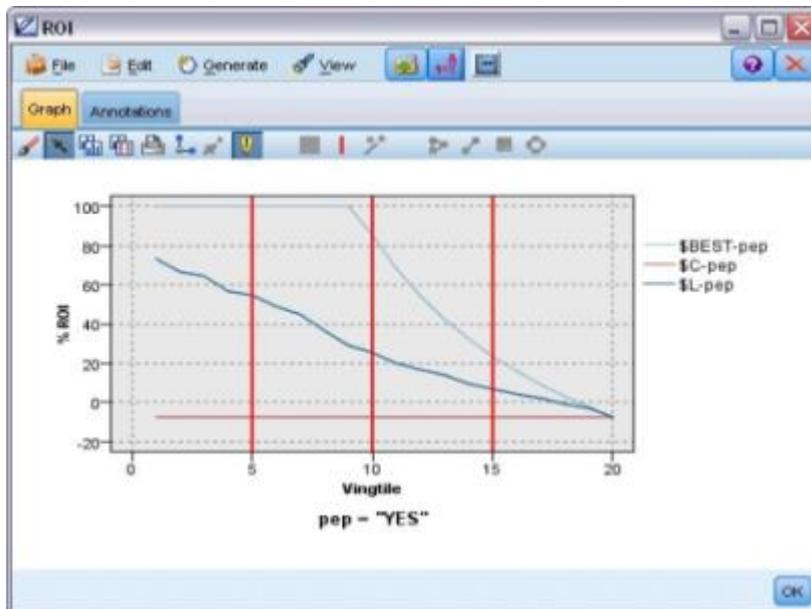
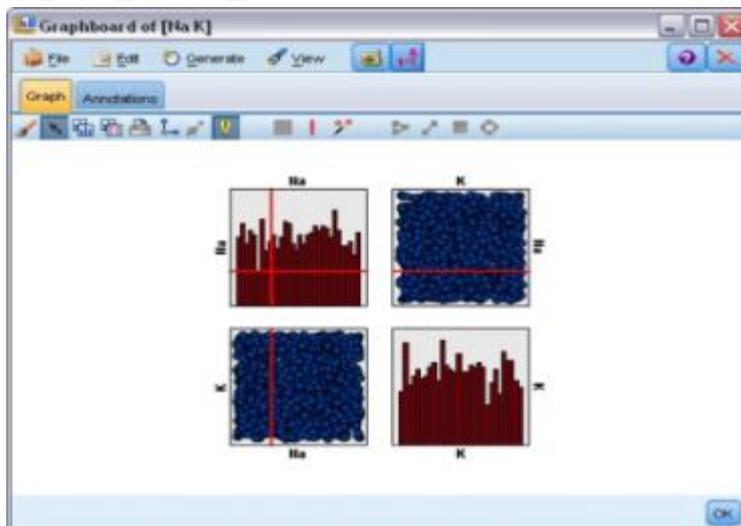


그림 2. 밴드가 있는 SPLOM



밴드 정의

밴드가 없는 그래프에서 밴드 라인을 추가하면 그래프가 2개의 밴드로 분할됩니다. 밴드 라인 값은 왼쪽에서 오른쪽으로 그래프를 읽을 때 두 번째 밴드의 시작점(하한이라고도 함)을 나타냅니다. 마찬가지로 2개의 밴드가 있는 그래프에서 밴드 라인을 추가하면 두 밴드 중 하나가 둘로 분할되어 3개의 밴드가 있게 됩니다. 기본적으로 밴드는 *bandN*으로 이름이 지정됩니다. 여기서 *N*은 *x*축에서 왼쪽부터 오른쪽까지의 밴드 수와 동일합니다.

밴드를 정의하고 나면 밴드를 끌어서 놓아 *x*축에서 밴드의 위치를 재설정할 수 있습니다. 밴드 내부를 마우스 오른쪽 단추로 클릭하여 해당 특정 밴드에 대한 노드 이름 바꾸기, 삭제 또는 생성 등의 작업에 대한 더 많은 단축키를 볼 수 있습니다.

밴드를 정의하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 보기 > 탐색 모드를 선택하십시오.
2. 탐색 모드 도구 모음에서 밴드 그리기 단추를 클릭하십시오.

그림 3. 밴드 그리기 도구 모음 단추



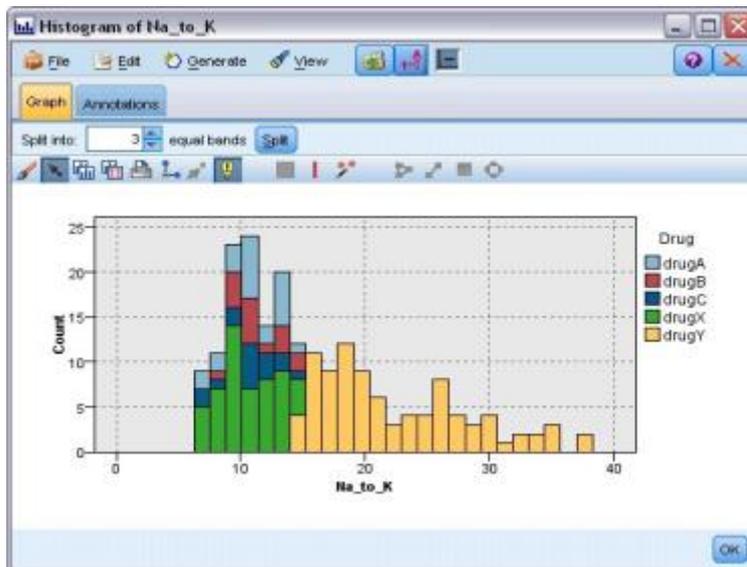
3. 밴드를 승인하는 그래프에서 밴드 라인을 정의할 *x*축 값 지점을 클릭하십시오.

참고: 그래프를 밴드로 분할 도구 모음 아이콘을 클릭하고 원하는 동등한 밴드의 수를 입력한 후 분할을 클릭할 수도 있습니다.

그림 4. 밴드로 분할하기 위한 옵션이 포함된 도구 모음을 펼치는 데 사용되는 분할자 아이콘



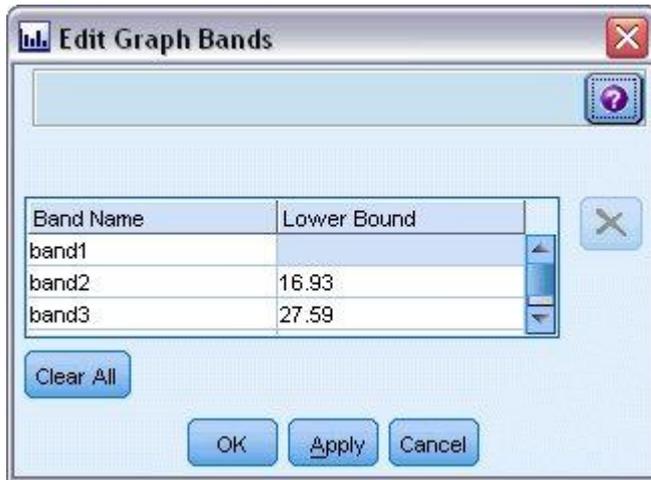
그림 5. 밴드가 사용으로 설정된 동등한 밴드 도구 모음 작성



밴드 편집, 이름 바꾸기 및 삭제

그래프 밴드 편집 대화 상자에서 또는 그래프 자체의 컨텍스트 메뉴를 통해 기존 밴드의 특성을 편집할 수 있습니다.

그림 6. 그래프 밴드 편집 대화 상자



밴드를 편집하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 밴드 그리기 단추를 클릭하십시오.
3. 메뉴에서 **편집 > 그래프 밴드**를 선택하십시오. 그래프 밴드 편집 대화 상자가 열립니다.
4. 그래프에 필드가 여럿 있는 경우(예: SPLOM 그래프)에는 드롭 다운 목록에서 원하는 필드를 선택할 수 있습니다.
5. 이름 및 하한을 입력하여 새 밴드를 추가하십시오. Enter 키를 눌러 새 행을 시작하십시오.
6. 하한 값을 조정하여 밴드의 경계를 편집하십시오.
7. 새 밴드 이름을 입력하여 밴드의 이름을 바꾸십시오.
8. 테이블에서 라인을 선택한 후 삭제 단추를 클릭하여 밴드를 삭제하십시오.
9. **확인**을 클릭하여 변경사항을 적용하고 대화 상자를 닫으십시오.

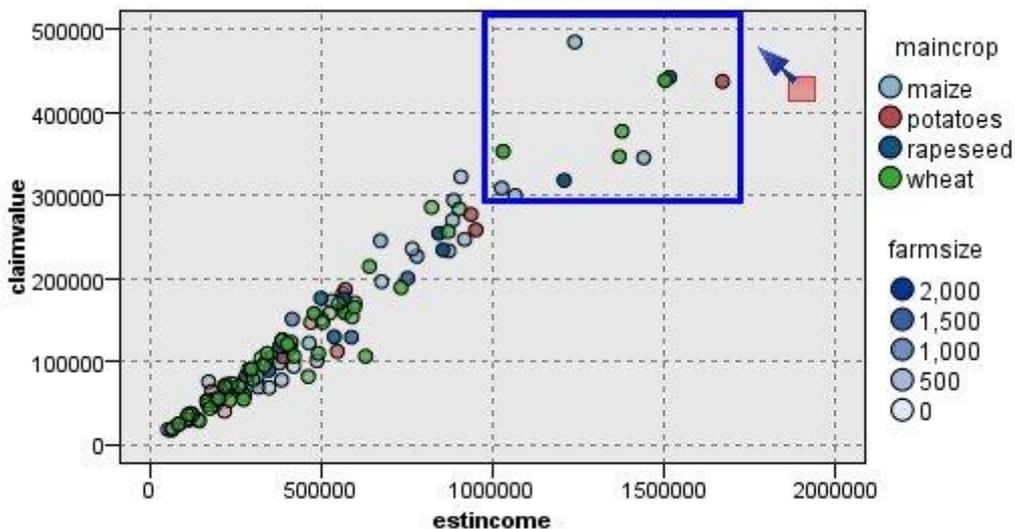
참고: 밴드의 라인을 마우스 오른쪽 단추로 클릭한 후 컨텍스트 메뉴에서 원하는 옵션을 선택하여 그래프에서 직접 밴드를 삭제하고 밴드의 이름을 바꿀 수도 있습니다.

② 영역 사용

두 개의 척도(또는 범위) 축이 있는 그래프에서는 영역을 그려서 그리는 직사각형 영역(영역이라고 함) 내에서 값을 그룹화할 수 있습니다. 영역은 최소 및 최대 X 및 Y 값으로 설명되는 그래프의 영역입니다. 그래프에 분할창이 여럿 있으면 한 패널에서 그려지는 영역이 다른 패널에도 표시됩니다.

일부 그래프는 영역을 승인하지 않습니다. 영역을 승인하는 그래프로는 도표(선, 산점도, 거품, 시간 등), SPLOM, 콜렉션 등이 있습니다. 이 영역은 X, Y 공간에서 그려지므로 1차원, 3차원 또는 애니메이션 도표에서 정의할 수 없습니다. 패널이 있는 그래프에서는 영역이 모든 패널에 표시됩니다. 산점도 교차표(SPLOM)를 사용하는 경우 대각선 도표는 하나의 척도 필드만 표시하기 때문에 대각선 도표가 아니라 해당 상단 도표에 해당 영역이 표시됩니다.

그림 1. 높은 클레임 값의 영역 정의



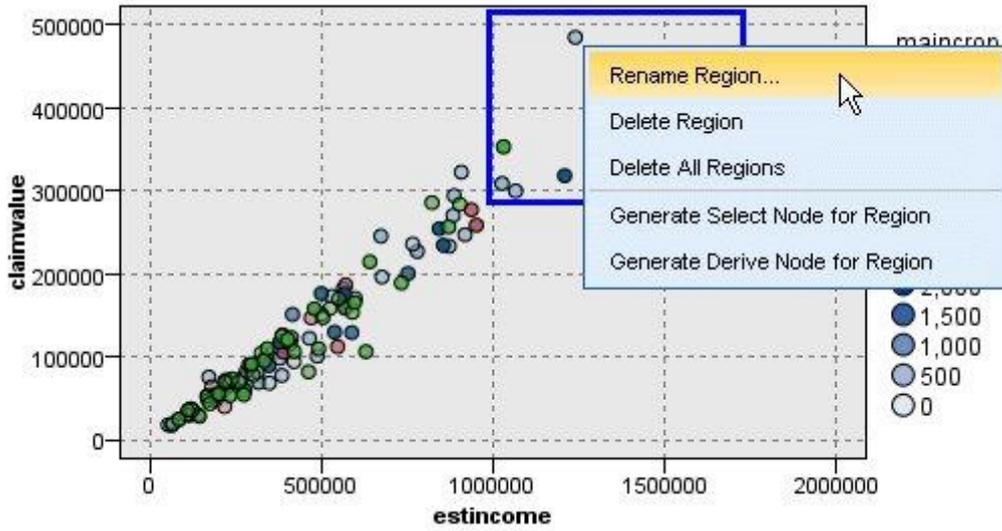
영역 정의

영역을 정의하는 모든 위치에서 값 그룹을 작성합니다. 기본적으로 각각의 새 영역을 *Region<N>*이라고 합니다. 여기서 N 은 이미 작성된 영역의 수에 해당합니다.

정의된 영역이 있으면 영역 라인을 마우스 오른쪽 단추로 클릭하여 일부 기본 단축키를 가져올 수 있습니다. 하지만 라인 위가 아니라 영역 내부를 마우스 오른쪽 단추로 클릭하여 해당 특정 영역을 위한 이름 바꾸기, 삭제, 선택 및 파생 노드 생성 등의 작업에 대한 많은 다른 단축키를 볼 수 있습니다.

특정 영역 또는 여러 영역 중 하나에 포함되어 있으면 레코드의 서브세트를 선택할 수 있습니다. 영역에 포함되어 있는지 여부에 따라 플래그 레코드에 대한 파생 노드를 생성하여 레코드에 대한 영역 정보도 통합할 수 있습니다. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

그림 2. 높은 클레임 값의 영역 탐색



영역을 정의하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 보기 > 탐색 모드를 선택하십시오.
2. 탐색 모드 도구 모음에서 영역 그리기 단추를 클릭하십시오.

그림 3. 영역 그리기 도구 모음 단추



3. 영역을 승인하는 그래프에서 마우스를 클릭한 후 끌어서 직사각형 영역을 그리십시오.

영역 편집, 이름 바꾸기 및 삭제

그래프 영역 편집 대화 상자에서 또는 그래프 자체의 컨텍스트 메뉴를 통해 기존 영역의 특성을 편집할 수 있습니다.

그림 4. 정의된 영역에 대한 특성 지정



영역을 편집하려면 다음을 수행하십시오.

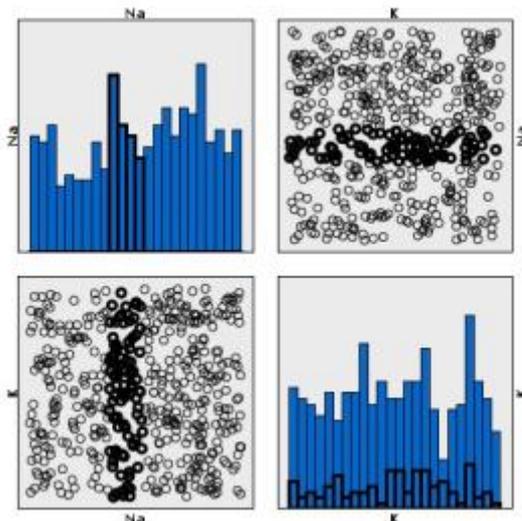
1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 영역 그리기 단추를 클릭하십시오.
3. 메뉴에서 **편집 > 그래프 영역**을 선택하십시오. 그래프 영역 편집 대화 상자가 열립니다.
4. 그래프에 필드가 여럿 있는 경우(예: SPLOM 그래프)에는 *필드 A* 및 *필드 B* 열에서 영역에 대한 필드를 정의해야 합니다.
5. 이름을 입력하고 필드 이름을 선택(해당되는 경우)하고 각 필드의 최대 및 최소 경계를 정의하여 새 라인에서 새 영역을 추가하십시오. Enter 키를 눌러 새 행을 시작하십시오.
6. *A* 및 *B*에 대한 **최소** 및 **최대** 값을 조정하여 기존 영역 경계를 편집하십시오.
7. 테이블에서 영역 이름을 변경하여 영역의 이름을 바꾸십시오.
8. 테이블의 라인을 선택한 후 삭제 단추를 클릭하여 영역을 삭제하십시오.
9. **확인**을 클릭하여 변경사항을 적용하고 대화 상자를 닫으십시오.

참고: 또는 영역의 라인을 마우스 오른쪽 단추로 클릭한 후 컨텍스트 메뉴에서 원하는 옵션을 선택하여 그래프에서 직접 영역을 삭제하고 영역의 이름을 바꿀 수 있습니다.

③ 표시된 요소 사용

모든 그래프에서 막대, 조각 및 점 등의 요소를 표시할 수 있습니다. 선은 해당 케이스의 필드를 나타내므로 시간 도표, 다중 도표 및 평가 그래프 이외의 그래프에서는 선, 영역 및 면을 표시할 수 없습니다. 요소를 표시할 때마다 반드시 해당 요소가 나타내는 모든 데이터를 강조표시합니다. 동일한 케이스가 둘 이상의 위치에서 표시되는 그래프(예: SPLOM)에서는 표시가 브러싱과 동의어입니다. 그래프에서 요소를 표시할 수 있으며 밴드 및 영역 내에서도 표시할 수 있습니다. 요소를 표시한 후 편집 모드로 돌아갈 때마다 표시는 계속 표시됩니다.

그림 1. SPLOM에서 요소 표시



그래프에서 요소를 클릭하여 요소를 표시하고 표시 해제할 수 있습니다. 처음으로 요소를 클릭하여 표시하면 표시되었음을 나타내는 굵은 테두리 색상과 함께 요소가 표시됩니다. 요소를 다시 클릭하면 테두리가 사라지고 요소가 더 이상 표시되지 않습니다. 여러 요소를 표시하려면 요소를 클릭하는 동안 Ctrl 키를 누르고 있거나 "마술 지팡이"를 사용하여 표시할 각 요소 주위에서 마우스를 끄십시오. Ctrl 키를 누르지 않고 다른 영역 또는 요소를 클릭하면 이전에 표시된 모든 요소가 선택 취소된다는 점을 기억하십시오.

그래프의 표시된 요소에서 선택 및 파생 노드를 생성할 수 있습니다. 자세한 정보는 그래프에서 노드 생성의 내용을 참조하십시오.

요소를 표시하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 요소 표시 단추를 클릭하십시오.
그림 2. 요소 표시 도구 모음 단추

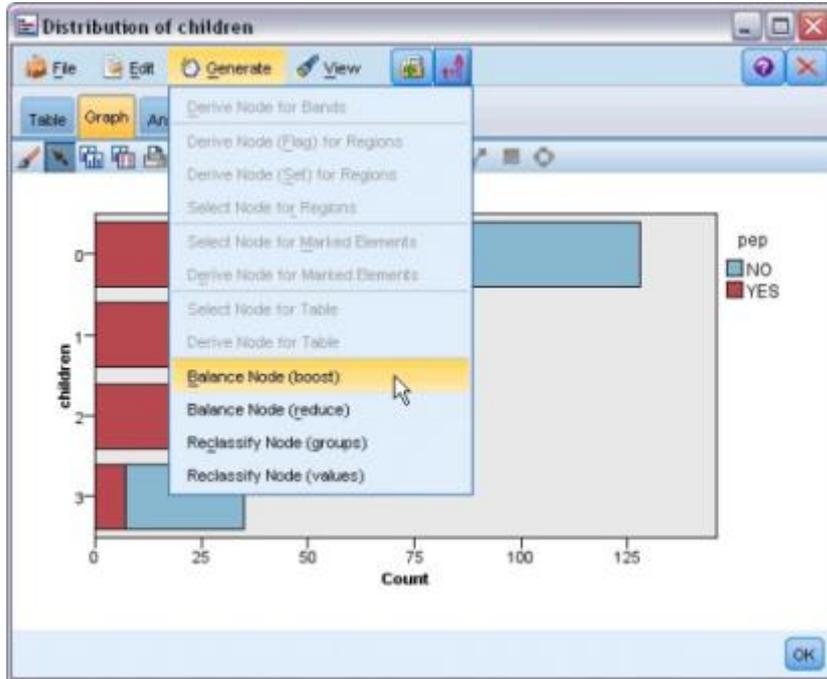
3. 필요한 요소를 클릭하거나 마우스를 클릭한 후 끌어서 여러 요소가 포함된 영역 주위에 선을 그리십시오.

④ 그래프에서 노드 생성

IBM® SPSS® Modeler 그래프가 제공하는 가장 강력한 기능 중 하나는 그래프 또는 그래프 내 선택사항으로부터 노드를 생성하는 기능입니다. 예를 들어, 시간 도표 그래프에서 데이터의 영역 또는 선택사항을 기반으로 파생 및 선택 노드를 생성하여 사실상 데이터의 "서브셋을 작성"할 수 있습니다. 예를 들어, 이 강력한 기능을 사용하여 이상치를 식별하여 제외할 수 있습니다.

밴드를 그릴 수 있을 때마다 파생 노드도 생성할 수 있습니다. 두 개의 척도 축을 가진 그래프에서는 그래프에서 그려진 영역에서 파생 또는 선택 노드를 생성할 수 있습니다. 표시된 요소를 가진 그래프에서는 파생 노드 및 선택 노드를 생성할 수 있으며 일부 경우에는 이 요소에서 필터 노드를 생성할 수 있습니다. 균형 노드 생성은 개수 분포를 표시하는 그래프에 대해 사용으로 설정됩니다.

그림 1. 생성 메뉴가 표시되는 그래프



노드를 생성할 때마다 노드를 기존 스트림에 연결할 수 있도록 노드가 스트림 캔버스에 직접 배치됩니다. 그래프에서 선택, 파생, 균형, 필터 및 재분류 노드를 생성할 수 있습니다.

선택 노드

선택 노드는 영역 내 레코드 포함 및 영역 외부의 모든 레코드 제외(다운스트림 처리의 경우 그 반대)에 대해 검증하기 위해 선택 노드를 생성할 수 있습니다.

- **밴드의 경우.** 해당 밴드 내 레코드를 포함하거나 제외하는 선택 노드를 생성할 수 있습니다. 선택 노드에서 사용할 밴드를 선택해야 하므로 **밴드에 대한 선택 노드**만은 컨텍스트 메뉴를 통해서만 사용 가능합니다.
- **영역의 경우.** 영역 내 레코드를 포함하거나 제외하는 선택 노드를 생성할 수 있습니다.
- **표시된 요소의 경우.** 표시된 요소 또는 웹 그래프 링크에 해당하는 레코드를 캡처하는 선택 노드를 생성할 수 있습니다.

파생 노드

파생 노드는 영역, 밴드 및 표시된 요소에서 생성될 수 있습니다. 모든 그래프는 파생 노드를 생성할 수 있습니다. 평가 차트의 경우에는 모델 선택을 위한 대화 상자가 표시됩니다. 웹 그래프의 경우에는 **파생 노드("And")** 및 **파생 노드("Or")**를 사용할 수 있습니다.

- **밴드의 경우.** 밴드 편집 대화 상자에 나열되는 밴드 이름을 범주 이름으로 사용하여 축에 표시된 각각의 간격에 대해 하나의 범주를 생성하는 파생 노드를 생성할 수 있습니다.
- **영역의 경우.** 플래그가 영역 내 레코드에 대해 T로 설정되고 모든 영역 외부의 레코드에 대해

F로 설정되는 *in_region*이라는 플래그 필드를 작성하는 파생 노드(플래그로 파생)를 생성할 수 있습니다. 레코드가 속하는 영역의 이름을 값으로 사용하는 각 레코드에 대해 *region*이라는 새 필드를 가진 각각의 영역에 대한 값을 가진 세트를 생성하는 파생 노드(세트로 파생)도 생성할 수 있습니다. 모든 영역 외부의 레코드는 기본 영역의 이름을 수신합니다. 값 이름은 영역 편집 대화 상자에 나열되는 영역 이름이 됩니다.

- **표시된 요소의 경우.** 모든 표시된 요소에 대해 **참고**이고 모든 기타 레코드에 대해 **거짓**인 플래그를 계산하는 파생 노드를 생성할 수 있습니다.

균형 노드

균형 노드는 데이터에서 불균형을 정정하기 위해 생성될 수 있습니다(예: 공통 값의 빈도 감소(균형 노드(감소) 메뉴 옵션 사용) 또는 빈도가 낮은 값의 발생 부스팅(균형 노드(부스트) 메뉴 옵션 사용)). 균형 노드 생성은 개수의 분포를 표시하는 그래프에 대해 사용으로 설정됩니다(예: 히스토그램, 점, 컬렉션, 개수의 막대형, 개수의 원형, 다중 도표).

필터 노드

필터 노드는 그래프에서 표시된 노드 또는 선을 기반으로 필드의 이름을 바꾸고 필드를 필터링하기 위해 생성될 수 있습니다. 평가 차트의 경우 최적 맞춤 선이 필터 노드를 생성하지 않습니다.

재분류 노드

재분류 노드는 값의 코딩을 변경하기 위해 생성될 수 있습니다. 이 옵션은 분포 그래프에 사용됩니다. 그룹에 포함되는지 여부에 따라 표시된 필드의 특정 값의 코딩을 변경하기 위해 그룹에 대해 재분류 노드를 생성할 수 있습니다(테이블 탭에서 Ctrl+클릭을 사용하여 그룹 선택). 수많은 값의 기존 세트로 데이터의 코딩을 변경하기 위해 값에 대해 재분류 노드를 생성할 수도 있습니다(예: 분석을 위해 다양한 회사의 재무 데이터를 병합하기 위해 데이터를 표준 값 세트로 재분류).

참고: 값이 사전 정의되어 있는 경우에는 해당 값을 플랫폼 파일로서 IBM SPSS Modeler로 읽어오고 분포를 사용하여 모든 값을 표시할 수 있습니다. 그런 다음 차트에서 직접 이 필드에 대한 재분류(값) 노드를 생성하십시오. 그러면 재분류 노드의 새 값 열(드롭 다운 목록)에 모든 목표 값이 배치됩니다.

재분류 노드에 대한 옵션을 설정할 때 테이블을 사용하면 이전 세트 값으로부터 사용자가 지정하는 새 값에 대한 명확한 매핑을 사용할 수 있습니다.

- **원래 값.** 이 열에는 선택 필드의 기존 값이 나열됩니다.
- **새로운 값.** 이 열을 사용하여 새 범주 값을 입력하거나 드롭 다운 목록에서 값을 선택하십시오. 분포 차트의 값을 사용하여 재분류 노드를 자동으로 생성하는 경우 이 값은 드롭 다운 목록에 포함되어 있습니다. 이를 통해 기존 값을 알려진 값 세트에 신속하게 매핑할 수 있습니다

다. 예를 들어, 의료 기관에서 네트워크 또는 로케일을 기반으로 진단을 다르게 그룹화하는 경우가 있습니다. 합병 또는 인수 이후 모든 당사자는 일관된 방식으로 새 데이터 또는 기존 데이터를 재분류해야 합니다. 긴 목록으로부터 각각의 목표 값을 수동으로 입력하는 대신 값의 마스터 목록을 IBM SPSS Modeler로 읽어오고 **진단** 필드에 대한 분포 차트를 실행하고 이 차트에서 직접 이 필드에 대한 재분류(값) 노드를 생성할 수 있습니다. 이 프로세스를 수행하면 모든 목표 진단 값을 새 값 드롭 다운 목록에서 사용할 수 있습니다.

재분류 노드에 대한 자세한 정보는 재분류 노드에 대한 옵션 설정의 내용을 참조하십시오.

그래프에서 노드 생성

그래프 출력 창의 생성 메뉴를 사용하여 노드를 생성할 수 있습니다. 생성된 노드는 스트림 캔버스에 배치됩니다. 해당 노드를 사용하려면 해당 노드를 기존 스트림에 연결하십시오.

그래프에서 노드를 생성하려면 다음을 수행하십시오.

1. 탐색 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 탐색 모드**를 선택하십시오.
2. 탐색 모드 도구 모음에서 영역 단추를 클릭하십시오.
3. 노드를 생성하기 위해 필요한 밴드, 영역 또는 표시된 요소를 정의하십시오.
4. 생성 메뉴에서 생성할 노드의 유형을 선택하십시오. 가능한 유형만 사용으로 설정됩니다.

참고: 마우스 오른쪽 단추를 클릭한 후 컨텍스트 메뉴에서 원하는 생성 옵션을 선택하여 그래프에서 직접 노드를 생성할 수도 있습니다.

(17) 시각화 편집

탐색 모드에서는 시각화로 표시되는 데이터 및 값을 분석적으로 탐색할 수 있는 반면, 편집 모드에서는 시각화의 레이아웃 및 모양을 변경할 수 있습니다. 예를 들어, 조직의 스타일 가이드에 맞게 글꼴 및 색상을 변경할 수 있습니다. 이 모드를 선택하려면 메뉴에서 **보기 > 편집 모드**를 선택하십시오(또는 도구 모음 아이콘 클릭).

편집 모드에서는 시각화 레이아웃의 다양한 측면에 영향을 주는 여러 도구 모음이 제공됩니다. 사용하지 않는 도구 모음이 있는 경우 이러한 도구 모음을 숨겨 대화 상자에서 그래프가 표시되는 공간을 늘릴 수 있습니다. 도구 모음을 선택하거나 선택 취소하려면 보기 메뉴에서 관련 도구 모음 이름을 클릭하십시오.

참고: 시각화에 세부사항을 추가하기 위해 제목, 각주 및 축 레이블을 적용할 수 있습니다. 자세한 정보는 제목 및 꼬리말 추가 주제를 참조하십시오.

편집 모드에서는 시각화를 편집하는 몇 가지 옵션이 제공됩니다. 다음을 수행할 수 있습니다.

- 텍스트를 편집하고 형식화합니다.
- 프레임 및 그래픽 요소의 채움 색상, 투명도 및 패턴을 변경합니다.
- 경계와 선의 색상 및 대시를 변경합니다.
- 점 요소의 형태와 가로 세로 비율을 회전시키고 변경합니다.
- 그래픽 요소(예: 막대 및 점)의 크기를 변경합니다.
- 여백 및 패딩을 사용하여 항목 주위의 공간을 조정합니다.
- 숫자에 대한 형식화를 지정합니다.
- 축 및 척도 설정을 변경합니다.
- 범주 축에서 범주를 정렬하고 제외시키며 합칩니다.
- 패널의 방향을 설정합니다.
- 좌표계에 변환을 적용합니다.
- 통계, 그래픽 요소 유형 및 충돌 한정자를 변경합니다.
- 범례의 위치를 변경합니다.
- 시각화 스타일시트를 적용합니다.

다음 주제에서는 이러한 다양한 작업을 수행하는 방법에 대해 설명합니다. 그래프를 편집하는 일반 규칙도 읽을 것을 권장합니다.

편집 모드로 전환하는 방법

메뉴에서 다음을 선택하십시오.

보기 > 편집 모드

① 시각화 편집 일반 규칙

편집 모드

모든 편집은 편집 모드에서 수행됩니다. 편집 모드를 사용하려면 메뉴에서 다음을 선택하십시오.

보기 > 편집 모드

선택

편집에 사용 가능한 옵션은 선택에 따라 다릅니다. 선택하는 항목에 따라 다른 도구 모음 및 특성 팔레트 옵션이 사용됩니다. 사용되는 항목만 현재 선택에 적용됩니다. 예를 들어, 축을 선택하는 경우 특성 팔레트에서 척도, 주 눈금 및 보조 눈금 탭이 사용 가능합니다.

다음은 시각화에서 항목을 선택하는 데 유용한 몇 가지 팁입니다.

- 항목을 클릭하면 항목이 선택됩니다.
- 그래픽 요소(예: 산점도의 점 또는 막대형 차트의 막대)는 한 번 클릭하여 선택합니다. 첫 번째 선택 후 다시 클릭하면 선택 범위가 그래픽 요소 그룹 또는 하나의 그래픽 요소로 좁혀집니다.
- 모든 것을 선택 취소하려면 Esc를 누르십시오.

팔레트

시각화에서 항목을 선택하면 다양한 팔레트가 업데이트되어 선택을 반영합니다. 팔레트에는 선택을 편집할 수 있는 제어가 있습니다. 팔레트는 여러 제어 및 탭이 포함된 패널이거나 도구 모음일 수 있습니다. 팔레트가 숨겨져 있을 수 있으므로 편집에 필요한 팔레트가 표시되었는지 확인하십시오. 보기 메뉴에 현재 표시된 팔레트가 있는지 확인하십시오.

도구 모음 팔레트의 빈 공간 또는 다른 팔레트의 왼쪽을 클릭한 후 끌어와 팔레트의 위치를 바꿀 수 있습니다. 시각적 피드백을 통해 팔레트를 고정할 위치를 알 수 있습니다. 도구 모음이 아닌 팔레트의 경우, 단기 단추를 클릭하여 팔레트를 숨기고 고정 해제 단추를 클릭하여 팔레트를 별도의 창에 표시할 수도 있습니다. 특정 팔레트에 대한 도움말을 표시하려면 도움말 단추를 클릭하십시오.

자동 설정

일부 설정은 **-자동-** 옵션을 제공합니다. 이는 자동 값이 적용됨을 의미합니다. 사용되는 자동 설정은 고유한 시각화 및 데이터 값에 따라 다릅니다. 값을 입력하여 자동 설정을 대체할 수 있습니다. 자동 설정을 복원하려면 현재 값을 삭제하고 Enter를 누르십시오. 설정이 **-자동-**을 다시 표시합니다.

항목 제거/숨기기

시각화에서 다양한 항목을 제거하거나 숨길 수 있습니다. 예를 들어, 범례 또는 축 레이블을 숨길 수 있습니다. 항목을 삭제하려면 항목을 선택하고 삭제를 누르십시오. 항목이 삭제를 허용하지 않는 경우에는 항목이 삭제되지 않습니다. 실수로 항목을 삭제한 경우에는 Ctrl+Z를 눌러 삭제를 실행 취소하십시오.

시/도

일부 도구 모음은 현재 선택 상태를 반영하지만 일부는 그렇지 않습니다. 특성 팔레트는 항상 상태를 반영합니다. 도구 모음이 상태를 반영하지 않는 경우 해당 도구 모음을 설명하는 주제에서 이에 대해 설명합니다.

② 텍스트 편집 및 형식화

기존 텍스트를 편집하고 전체 텍스트 블록의 형식화를 변경할 수 있습니다. 데이터 값에 직접 연결된 텍스트는 편집할 수 없습니다. 예를 들어, 눈금 레이블은 해당 콘텐츠가 기본 데이터에서 파생되기 때문에 편집할 수 없습니다. 그러나 시각화의 모든 텍스트를 형식화할 수 있습니다.

기존 텍스트 편집 방법

1. 텍스트 블록을 두 번 클릭하십시오. 이 조치는 모든 텍스트를 선택합니다. 텍스트를 편집하는 동안에는 시각화의 다른 부분을 변경할 수 없으므로 모든 도구 모음이 사용되지 않습니다.
2. 텍스트를 입력하여 기존 텍스트를 대체하십시오. 텍스트를 다시 클릭하면 커서가 표시됩니다. 원하는 위치에 커서를 놓고 추가 텍스트를 입력하십시오.

텍스트 형식화 방법

1. 텍스트를 포함하는 프레임을 선택하십시오. 텍스트를 두 번 클릭하지 마십시오.
2. 글꼴 도구 모음을 사용하여 텍스트를 형식화하십시오. 이 도구 모음이 사용되지 않는 경우 텍스트를 포함하는 프레임만 선택되었는지 확인하십시오. 텍스트 자체가 선택된 경우에는 이 도구 모음이 사용되지 않습니다.

글꼴과 관련하여 다음을 변경할 수 있습니다.

- 색상
- 글자체(예: Arial 또는 Verdana)
- 크기(다른 단위(예: pc)를 표시하지 않는 한 단위는 pt임)
- 가중치
- 텍스트 프레임과 관련한 맞추기

형식화는 프레임 내의 모든 텍스트에 적용됩니다. 텍스트의 특정 블록에 있는 개별 문자 또는 단어의 형식화는 변경할 수 없습니다.

③ 색상, 패턴, 대시 및 투명도 변경

시각화에 있는 다양한 항목에는 채움과 경계가 있습니다. 가장 명확한 예는 막대형 차트의 막대입니다. 막대의 색상은 채움 색상입니다. 또한 막대 주위에 검은색 실선 경계가 있을 수도 있습니다.

시각화에는 좀 덜 명확하지만 역시 채움 색상이 있는 다른 항목들이 있습니다. 채움 색상이 투명하면 채움이 있는지 모를 수도 있습니다. 예를 들어, 축 레이블에 있는 텍스트를 고려하십시오. 이 텍스트가 "떠 있는" 텍스트처럼 보이지만 실제로는 투명한 채움 색상을 갖는 프레임 안에 표시됩니다. 축 레이블을 선택하여 프레임을 볼 수 있습니다.

전체 시각화를 두르는 프레임을 비롯하여 시각화 내의 모든 프레임은 채움 및 경계 스타일을 가질 수 있습니다. 또한 모든 채움은 조정 가능한 불투명도/투명도 수준과 연관됩니다.

색상, 패턴, 대시 및 투명도 변경 방법

1. 형식화할 항목을 선택하십시오. 예를 들어, 막대형 차트의 막대 또는 텍스트를 포함하는 프레임 선택하십시오. 시각화가 범주형 변수 또는 필드로 분할되는 경우에는 개별 범주에 해당되는 그룹을 선택할 수도 있습니다. 그러면 해당 그룹에 지정된 기본 모양을 변경할 수 있습니다. 예를 들어, 누적 막대형 차트에서 누적 그룹 중 하나의 색상을 변경할 수 있습니다.
2. 채움 색상, 경계 색상 또는 채움 패턴을 변경하려면 색상 도구 모음을 사용하십시오.
참고: 이 도구 모음은 현재 선택 상태를 반영하지 않습니다.

색상 또는 채움을 변경하려면 단추를 클릭하여 표시된 옵션을 선택하거나 드롭 다운 화살표를 클릭하여 다른 옵션을 선택하십시오. 색상의 경우, 빨간색 대각선이 그어진 흰색처럼 보이는 하나의 색상이 있습니다. 이것은 투명한 색상입니다. 예를 들어, 막대의 경계를 숨기는 데 이 색상을 사용할 수 있습니다.

- 첫 번째 단추는 채움 색상을 제어합니다. 색상이 연속형 또는 순서 필드와 연관되어 있는 경우, 이 단추는 데이터의 가장 높은 값과 연관된 색상의 채움 색상을 변경합니다. 특성 팔레트의 색상 탭을 사용하여 가장 낮은 값 및 결측 데이터와 연관된 색상을 변경할 수 있습니다. 요소의 색상은 기본 데이터의 값이 증가함에 따라 낮은 색상에서 높은 색상으로 증분식으로 변경됩니다.
 - 두 번째 단추는 경계 색상을 제어합니다.
 - 세 번째 단추는 채움 패턴을 제어합니다. 채움 패턴은 경계 색상을 사용합니다. 따라서 채움 패턴은 표시되는 경계 색상이 있는 경우에만 표시됩니다.
 - 네 번째 제어는 채움 색상 및 패턴의 불투명도를 제어하는 슬라이더 및 텍스트 상자입니다. 백분율이 낮을수록 불투명도가 낮고 투명도가 높습니다. 100%는 완전히 불투명한 상태입니다(투명도 없음).
3. 경계 또는 선의 대시를 변경하려면 선 도구 모음을 사용하십시오.
참고: 이 도구 모음은 현재 선택 상태를 반영하지 않습니다.

다른 도구 모음과 마찬가지로 단추를 클릭하여 표시된 옵션을 선택하거나 드롭다운 화살표를 클릭하여 다른 옵션을 선택하십시오.

④ 점 요소의 형태 및 가로 세로 비율 회전과 변경

점 요소를 회전시키거나 사전 정의된 다른 형태를 지정하거나 가로 세로 비율(너비 대 높이 비율)을 변경할 수 있습니다.

점 요소 수정 방법

1. 점 요소를 선택하십시오. 개별 점 요소의 형태 및 가로 세로 비율은 회전시키거나 변경할 수 없습니다.
2. 기호 도구 모음을 사용하여 점을 수정하십시오.
 - 첫 번째 단추를 사용하여 점의 형태를 변경할 수 있습니다. 드롭 다운 화살표를 클릭하고 사전 정의된 형태를 선택하십시오.
 - 두 번째 단추를 사용하여 점을 특정 컴퍼스 위치로 회전시킬 수 있습니다. 드롭 다운 화살표를 클릭한 후 바늘을 원하는 위치로 끌어오십시오.
 - 세 번째 단추를 사용하여 가로 세로 비율을 변경할 수 있습니다. 드롭 다운 화살표를 클릭한 후 표시되는 직사각형을 끌어오십시오. 직사각형의 형태는 가로 세로 비율을 나타냅니다.

⑤ 그래픽 요소의 크기 변경

시각화에 있는 그래픽 요소의 크기를 변경할 수 있습니다. 여기에는 막대, 선 및 점이 포함됩니다. 변수 또는 필드로 그래픽 요소의 크기가 지정되는 경우 지정된 크기는 **최소** 크기입니다.

그래픽 요소의 크기 변경 방법

1. 크기 조정할 그래픽 요소를 선택하십시오.
2. 슬라이더를 사용하여 크기를 변경하십시오.

⑥ 여백 및 패딩 지정

시각화에서 프레임 주위 또는 내부에 공간이 너무 많거나 너무 적은 경우 여백 및 패딩 설정을 변경할 수 있습니다. **여백**은 프레임과 프레임 주변에 있는 다른 항목들 사이에 있는 공간의 양입니다. **패딩**은 프레임의 경계와 프레임의 **컨텐츠** 사이에 있는 공간의 양입니다.

여백 및 패딩 지정 방법

1. 여백 및 패딩을 지정할 프레임을 선택하십시오. 이 프레임은 텍스트 프레임, 범례를 두르는 프레임 또는 그래픽 요소(예: 막대 및 점)를 표시하는 데이터 프레임일 수 있습니다.
2. 특성 팔레트의 여백 탭을 사용하여 설정을 지정하십시오. 다른 단위(예: cm 또는 in)를 표시하지 않는 한 모든 크기의 단위는 픽셀입니다.

⑦ 숫자 형식 지정

연속형 축의 눈금 레이블 또는 숫자를 표시하는 데이터 값 레이블에 표시되는 숫자의 형식을 지정할 수 있습니다. 예를 들어, 눈금 레이블에 표시되는 숫자가 천단위로 표시되도록 지정할 수 있습니다.

숫자 형식 지정 방법

1. 연속형 축 눈금 레이블 또는 데이터 값 레이블(숫자를 포함하는 경우)을 선택하십시오.
2. 특성 팔레트에서 **형식** 탭을 클릭하십시오.
3. 원하는 숫자 형식 지정 옵션을 선택하십시오.

접두부. 숫자 시작 부분에 표시할 문자입니다. 예를 들어, 숫자가 미국 달러 단위의 금액이면 달러 부호(\$)를 입력하십시오.

접미부. 숫자 끝 부분에 표시할 문자입니다. 예를 들어, 숫자가 백분율이면 백분율 부호(%)를 입력하십시오.

최소 정수 자릿수. 십진 표시의 정수 부분에 표시할 최소 자릿수입니다. 실제값에 최소 자릿수가 포함되지 않는 경우 값의 정수 부분은 0으로 채워집니다.

최대 정수 자릿수. 십진 표시의 정수 부분에 표시할 최대 자릿수입니다. 실제값이 최대 자릿수를 초과하는 경우 값의 정수 부분은 별표로 대체됩니다.

최소 소수 자릿수. 십진 또는 지수 표시의 소수 부분에 표시할 최소 자릿수입니다. 실제값에 최소 자릿수가 포함되지 않는 경우 값의 소수 부분은 0으로 채워집니다.

최대 소수 자릿수. 십진 또는 지수 표시의 소수 부분에 표시할 최대 자릿수입니다. 실제값이 최대 자릿수를 초과하는 경우 소수는 해당 자릿수로 반올림됩니다.

지수 표기법. 숫자를 지수 표기법으로 표시할지 여부입니다. 지수 표기법은 매우 크거나 작은 숫자에 유용합니다. **-자동-**을 선택하면 애플리케이션에서 지수 표기법이 적합한 시점을 결정합니다.

스케일링. 척도 요인이며 원래 값을 나누는 하나의 숫자입니다. 숫자가 크지만 숫자를 수용하기 위해 레이블이 너무 많이 확장되는 것을 원하지 않는 경우에는 척도 요인을 사용하십시오. 눈금 레이블의 숫자 형식을 변경하는 경우 축 제목을 편집하여 숫자를 해석하는 방법을 표시하십시오. 예를 들어, 척도 축이 금액을 표시하고 레이블이 30,000, 50,000 및 70,000이라고 가정하십시오. 30, 50 및 70을 표시하기 위해 척도 요인 1000을 입력할 수 있습니다. 그런 다음에는 천단위 텍스트를 포함하도록 척도 축 제목을 편집해야 합니다.

괄호(음의 값). 음의 값에 괄호를 사용하는지 여부를 지정합니다.

숫자 분리. 숫자 그룹 사이에 문자를 표시하는지 여부입니다. 사용하는 컴퓨터의 현재 로케일이 숫자 분리에 사용되는 문자를 결정합니다.

⑧ 축 및 척도 설정 변경

축과 척도를 수정하는 데 사용하는 몇 가지 옵션이 있습니다.

축 및 척도 설정 변경 방법

1. 축의 특정 부분을 선택하십시오(예: 축 레이블 또는 눈금 레이블).
2. 특성 팔레트에 있는 척도, 주 눈금 및 보조 눈금 탭을 사용하여 축 및 척도 설정을 변경하십시오.

척도 탭

참고: 데이터가 미리 수집되는 그래프(예: 히스토그램)의 경우 척도 탭이 표시되지 않습니다.

유형. 척도가 선형 척도인지 또는 변환 척도인지 여부를 지정합니다. 척도 변환을 통해 더 쉽게 데이터를 이해하고 통계 추론에 필요한 가정을 세울 수 있습니다. 산점도에서는 독립변수(또는 필드)와 종속변수(또는 필드) 간의 관계가 비선형인 경우 변환된 척도를 사용할 수 있습니다. 비대칭 히스토그램을 정규 분포와 유사하게 보이도록 대칭적으로 만드는 데에도 척도 변환을 사용할 수 있습니다. 데이터가 표시되는 척도만 변환하고 실제 데이터는 변환하지 않습니다.

- **선형.** 변환되지 않은 선형 척도를 지정합니다.
- **로그.** 기본-10로그 변환 척도를 지정합니다. 0과 음의 값을 수용하기 위해 이 변환에서는 수정된 로그 함수 버전을 사용합니다. 이 "안전 로그" 함수는 $\text{sign}(x) * \log(1 + \text{abs}(x))$ 로 정의됩니다. 따라서 $\text{safeLog}(-99)$ 는 다음과 같습니다.

$$\text{sign}(-99) * \log(1 + \text{abs}(-99)) = -1 * \log(1 + 99) = -1 * 2 = -2$$

- **거듭제곱.** 지수 0.5를 사용하여 거듭제곱 변환 척도를 지정합니다. 음의 값을 수용하기 위해 이 변환에서는 수정된 거듭제곱 함수 버전을 사용합니다. 이 "안전 거듭제곱" 함수는 $\text{sign}(x) * \text{pow}(\text{abs}(x), 0.5)$ 로 정의됩니다. 따라서 $\text{safePower}(-100)$ 은 다음과 같습니다.

$$\text{sign}(-100) * \text{pow}(\text{abs}(-100), 0.5) = -1 * \text{pow}(100, 0.5) = -1 * 10 = -10$$

최소값/최대값/적당히 낮음/적당히 높음. 척도의 범위를 지정합니다. **적당히 낮음**과 **적당히 높음**을 선택하면 애플리케이션이 데이터를 기반으로 적절한 척도를 선택합니다. 일반적으로 최소값 및 최대값은 최대 및 최소 데이터 값보다 크거나 작은 전체 값이므로 "적당한" 값입니다. 예를 들어, 데이터 범위가 4-92인 경우 척도의 적당히 낮은 값과 높은 값은 실제 데이터 최소값 및 최대값이 아닌 0과 100이 될 수 있습니다. 너무 작은 범위를 설정해 중요한 항목이 숨겨지지 않도록 주의해야 합니다. 또한 **0 포함** 옵션이 선택된 경우 명시적 최소값과 최대값을 설정할 수 없습니다.

낮은 여백/높은 여백. 낮거나 높은 축 끝에 여백을 만듭니다. 여백은 선택된 축과 직각으로 표시됩니다. 단위는 다른 단위(cm 또는 in)를 표시하지 않는 한 픽셀이 됩니다. 예를 들어, 수직축에 대해 **높은 여백**을 5로 설정하면 데이터 프레임 맨 위를 따라 5px의 수평 여백이 만들어집니다.

반전. 척도가 반전되는지 여부를 지정합니다.

0 포함. 척도에 0이 포함됨을 표시합니다. 이 옵션은 일반적으로 막대형 차트에서 가장 작은 막대의 높이에 가까운 값이 아닌 0에서 막대가 시작되도록 하는 데 사용됩니다. 이 옵션을 선택하면 척도 범위에 대해 사용자 정의 최소값 및 최대값을 설정할 수 없으므로 **최소값** 및 **최대값**이 사용되지 않습니다.

주 눈금/보조 눈금 탭

눈금 또는 눈금 표시는 축 위에 표시되는 선입니다. 이러한 눈금은 특정 구간 또는 범주에서 값을 표시합니다. **주 눈금**은 레이블이 있는 눈금 표시입니다. 주 눈금은 또한 다른 눈금 표시보다 깊습니다. **보조 눈금**은 주 눈금 표시 사이에 표시되는 눈금 표시입니다. 눈금 유형에 고유한 옵션도 있지만 대부분의 옵션은 주 눈금 및 보조 눈금에 사용할 수 있습니다.

눈금 표시. 그래프에 주 눈금을 표시할지 보조 눈금을 표시할지 여부를 지정합니다.

눈금선 표시. 눈금선을 주 눈금에서 표시할지 보조 눈금에서 표시할지 여부를 지정합니다. **눈금선**은 축에서 축까지 전체 그래프를 가로지르는 선입니다.

위치. 축과 관련된 눈금 표시의 위치를 지정합니다.

길이. 눈금 표시의 길이를 지정합니다. 단위는 다른 단위(cm 또는 in)를 표시하지 않는 한 픽셀이 됩니다.

기준. *주 눈금에만 적용합니다.* 첫 번째 주 눈금이 표시되는 지점의 값을 지정합니다.

델타. *주 눈금에만 적용합니다.* 주 눈금 사이의 간격을 지정합니다. 즉, 주 눈금은 n 번째 값마다 표시되며 여기서 n 은 델타 값입니다.

구획. *보조 눈금에만 적용합니다.* 주 눈금 사이의 보조 눈금 구획 수를 지정합니다. 보조 눈금 수는 구획 수보다 하나가 적습니다. 예를 들어, 0과 100에 주 눈금이 있다고 가정하십시오. 보조 눈금 구획 수로 2를 입력하면 50에 한 개의 보조 눈금이 생기고 0-100 범위를 나누어 두 개의 구획이 작성됩니다.

⑨ 범주 편집

범주 축에서 다음과 같은 방법으로 범주를 편집할 수 있습니다.

- 범주를 표시하는 정렬 순서를 변경합니다.
- 특정 범주를 제외시킵니다.
- 데이터 세트에 나타나지 않는 범주를 추가합니다.
- 작은 범주들을 하나의 범주로 합칩니다.

범주 정렬 순서 변경 방법

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.
참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM® SPSS® Modeler의 보기 메뉴에서 **범주**를 선택하십시오.

2. 범주 팔레트의 드롭 다운 목록에서 정렬 옵션을 선택하십시오.
사용자 정의. 팔레트에 표시되는 순서를 기준으로 범주를 정렬합니다. 화살표 단추를 사용하여 범주를 목록 맨 위, 위, 아래 또는 맨 아래로 이동시키십시오.

데이터. 데이터 세트의 범주 순서를 기반으로 범주를 정렬합니다.

이름. 팔레트에 표시되는 이름을 사용하여 알파벳순으로 범주를 정렬합니다. 값 및 레이블을 표시하는 도구 모음 단추가 선택되었는지 여부에 따라 값이거나 레이블일 수 있습니다.

값. 팔레트에서 괄호 안에 표시되는 값을 사용하여 기본 데이터 값을 기준으로 범주를 정렬합니다. 메타데이터가 있는 데이터 소스(예: IBM SPSS Statistics 데이터 파일)만 이 옵션을 지원합니다.

통계. 각 범주에 대해 계산된 통계를 기준으로 범주를 정렬합니다. 통계의 예로는 개수, 퍼센트, 평균 등을 들 수 있습니다. 이 옵션은 그래프에서 통계가 사용되는 경우에만 사용할 수 있습니다.

범주 추가 방법

기본적으로 데이터 세트에 나타나는 범주만 사용할 수 있습니다. 필요한 경우 범주를 시각화에 추가할 수 있습니다.

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.
참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.

2. 범주 팔레트에서 범주 추가 단추를 클릭하십시오.

그림 1. 범주 추가 단추



3. 새 범주 추가 대화 상자에서 범주 이름을 입력하십시오.
4. **확인**을 클릭하십시오.

특정 범주 제외 방법

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.
참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.
2. 범주 팔레트에서 포함 목록에 있는 범주 이름을 선택한 다음 X 단추를 클릭하십시오. 범주를 다시 옮기려면 제외 목록에서 범주 이름을 선택한 다음 목록의 오른쪽에 있는 화살표를 클릭하십시오.

작은 범주를 합치는 방법

너무 작아 개별적으로 표시할 필요가 없는 범주를 결합할 수 있습니다. 예를 들어, 범주가 많은 원형 차트가 있는 경우 백분율이 10 미만인 범주를 합칠 수 있습니다. 가산적 통계인 경우에만 합칠 수 있습니다. 예를 들어, 평균은 가산적이지 않으므로 합칠 수 없습니다. 따라서 평균을 사용한 범주 합치기는 사용할 수 없습니다.

1. 범주 축을 선택하십시오. 범주 팔레트가 축의 범주를 표시합니다.
참고: 팔레트가 표시되지 않으면 팔레트를 사용하도록 설정했는지 확인하십시오. IBM SPSS Modeler의 보기 메뉴에서 **범주**를 선택하십시오.
2. 범주 팔레트에서 **합치기**를 선택하고 백분율을 지정하십시오. 총 백분율이 지정된 수보다 작은 범주가 하나의 범주로 결합됩니다. 백분율은 차트에 표시된 통계를 기반으로 합니다. 합치기는 개수 기반 및 합계(합) 통계에만 사용할 수 있습니다.

⑩ 패널 방향 변경

시각화에서 패널을 사용하는 경우 패널 방향을 변경할 수 있습니다.

패널 방향 변경 방법

1. 시각화의 임의 부분을 선택하십시오.
2. 특성 팔레트에서 **패널** 탭을 클릭하십시오.
3. **레이아웃**에서 옵션을 선택하십시오.

테이블. 각 개별 값에 지정된 행 또는 열이 있다는 점에서 패널의 레이아웃이 테이블과 유사합니다.

전치. 패널의 레이아웃이 테이블과 유사할 뿐만 아니라 원래 행과 열이 스왑됩니다. 이 옵션은 그래프 자체를 전치시키는 것과는 다릅니다. 이 옵션을 선택할 때 x 축 및 y 축은 변경되지 않습니다.

목록. 각 셀이 값 조합을 나타낸다는 점에서 패널의 레이아웃이 목록과 유사합니다. 열 및 행이 더 이상 개별 값에 지정되지 않습니다. 이 옵션을 사용하면 필요한 경우 패널이 랩핑됩니다.

⑪ 좌표계 변환

다수의 시각화가 평면 직교 좌표계에 표시됩니다. 필요에 따라 좌표계를 변환할 수 있습니다. 예를 들어, 좌표계에 극 변환을 적용하고 기울기 아래 그림자 효과를 추가하며 축을 전치시킬 수 있습니다. 또한 현재 시각화에 변환이 이미 적용된 경우 이러한 변환을 취소할 수 있습니다. 예를 들어, 극 좌표계에 원형 차트가 그려집니다. 극 변환을 실행 취소하고 원형 차트를 직교 좌표계에서 하나의 누적 막대형 차트로 표시할 수 있습니다.

좌표계 변환 방법

1. 변환할 좌표계를 선택하십시오. 개별 그래프의 프레임을 선택하여 좌표계를 선택합니다.
2. 특성 팔레트에서 **좌표** 탭을 클릭하십시오.
3. 좌표계에 적용할 변환을 선택하십시오. 변환을 선택 취소하여 실행을 취소할 수도 있습니다.
전치. 축 방향을 변경하는 것을 전치라고 합니다. 전치는 2차원 시각화에서 수직 축과 수평 축을 스와핑하는 것과 유사합니다.

극. 극 변환은 그래프 중심으로부터의 특정 거리와 특정 각도에서 그래픽 요소를 그립니다. 원형 차트는 특정 각도에서 개별 막대를 그리는 극 변환이 적용된 1차원 시각화입니다. 방사형 차트는 그래프 중심으로부터의 특정 거리와 특정 각도에서 그래픽 요소를 그리는 극 변환이 적용된 2차원 시각화입니다. 3차원 시각화에는 깊이 차원이 추가로 포함됩니다.

기울기. 기울기 변환은 그래픽 요소에 3차원 효과를 추가합니다. 이 변환은 그래픽 요소에 깊이를 추가하지만 깊이는 단순한 장식 차원에 불과합니다. 깊이는 특정 데이터 값에 영향을 받지 않습니다.

동일한 비율. 동일한 비율을 적용하는 경우 각 척도에서 동일한 거리는 데이터 값에서의 차이가 동일함을 나타냅니다. 예를 들어, 두 척도 모두 2cm는 차이값 1000을 나타냅니다.

변환 전 여백 %. 변환 후 축이 클리핑되는 경우 변환을 적용하기 전에 그래프에 여백을 추가

하러 할 수 있습니다. 여백은 좌표계에 변환이 적용되기 전에 차원을 특정 백분율만큼 축소시킵니다. 아래쪽 x , 위쪽 x , 아래쪽 y 및 위쪽 y 차원을 나열된 순서대로 제어할 수 있습니다.

변환 후 여백 %. 그래프의 가로 세로 비율을 변경하려는 경우 변환을 적용한 후에 그래프에 여백을 추가할 수 있습니다. 여백은 좌표계에 변환이 적용된 후 차원을 특정 백분율만큼 축소시킵니다. 그래프에 변환이 적용되지 않는 경우에도 이러한 여백을 적용할 수 있습니다. 아래쪽 x , 위쪽 x , 아래쪽 y 및 위쪽 y 차원을 나열된 순서대로 제어할 수 있습니다.

⑫ 통계 및 그래픽 요소 변경

그래픽 요소를 다른 유형으로 변환하고 그래픽 요소를 그리는 데 사용하는 통계를 변경하며 그래픽 요소가 겹쳐질 때 발생하는 상황을 결정하는 충돌 한정자를 지정할 수 있습니다.

그래픽 요소 변환 방법

1. 변환할 그래픽 요소를 선택하십시오.
2. 특성 팔레트에서 **요소** 탭을 클릭하십시오.
3. 유형 목록에서 새 그래픽 요소 유형을 선택하십시오.

표 1. 그래픽 요소 유형

그래픽 요소 유형	설명
점	특정 데이터 점을 식별하는 마커입니다. 점 요소는 산점도 및 다른 관련 시각화에서 사용됩니다.
구간	특정 데이터 값에서 그려지고 원점과 다른 데이터 값 사이의 공간을 채우는 직사각형 형태입니다. 구간 요소는 막대형 차트와 히스토그램에서 사용됩니다.
선	데이터 값을 연결하는 선입니다.
경로	데이터 세트에 나타나는 순서로 데이터 값을 연결하는 선입니다.
영역	데이터 요소를 연결하는 선이며 선과 원점 사이의 영역이 채워집니다.
다각형	데이터 영역을 둘러싼 여러 면으로 구성된 도형입니다. 다각형 요소는 구간화된 산점도 또는 맵에서 사용할 수 있습니다.
스키마	이상치를 나타내는 수염도표 및 마커가 있는 하나의 상자로 구성된 요소입니다. 스키마 요소는 상자 도표에 사용됩니다.

통계 변경 방법

1. 통계를 변경할 그래픽 요소를 선택하십시오.
2. 특성 팔레트에서 **요소** 탭을 클릭하십시오.

충돌 한정자 지정 방법

충돌 한정자는 그래픽 요소가 겹쳐질 때 발생하는 상황을 결정합니다.

1. 충돌 한정자를 지정할 그래픽 요소를 선택하십시오.
2. 특성 팔레트에서 요소 탭을 클릭하십시오.
3. 수정자 드롭 다운 목록에서 충돌 한정자를 선택하십시오. **-자동-**을 선택하면 애플리케이션이 그래픽 요소 유형 및 통계에 적합한 충돌 한정자를 결정합니다.
오버레이. 값이 동일하면 서로의 위에 그래픽 요소를 그립니다.

누적. 데이터 값이 동일한 경우 일반적으로 겹쳐지는 그래픽 요소를 누적시킵니다.

닷지. 같은 값에서 표시되는 다른 그래픽 요소 위에 그래픽 요소를 겹치는 대신 그 옆으로 이동시킵니다. 그래픽 요소가 대칭적으로 배열됩니다. 즉, 그래픽 요소가 중앙 위치의 반대편으로 이동합니다. 닷지는 균집화와 유사합니다.

적재. 같은 값에서 표시되는 다른 그래픽 요소 위에 그래픽 요소를 겹치는 대신 그 옆으로 이동시킵니다. 그래픽 요소가 비대칭적으로 배열됩니다. 즉, 맨 아래의 그래픽 요소가 척도의 특정 값에 위치하고 그래픽 요소가 서로의 위에 적재됩니다.

지터(정규). 정규 분포를 사용하여 동일한 데이터 값에 있는 그래픽 요소의 위치를 무작위로 바꿉니다.

지터(균등). 균등 분포를 사용하여 동일한 데이터 값에 있는 그래픽 요소의 위치를 무작위로 바꿉니다.

⑬ 범례 위치 변경

그래프에 범례가 포함되는 경우 범례는 일반적으로 그래프의 오른쪽에 표시됩니다. 필요한 경우 이 위치를 변경할 수 있습니다.

범례 위치 변경 방법

1. 범례를 선택하십시오.
2. 특성 팔레트에서 **범례** 탭을 클릭하십시오.
3. 위치를 선택하십시오.

⑭ 시각화 및 시각화 데이터 복사

일반 팔레트에는 시각화와 해당 데이터를 복사하는 단추가 있습니다.

그림 1. 시각화 복사 단추



시각화 복사. 이 조치는 시각화를 클립보드에 이미지로 복사합니다. 여러 이미지 형식을 사용할 수 있습니다. 이미지를 다른 애플리케이션에 붙여넣을 때는 "선택하여 붙여넣기" 옵션을 선택하여 사용 가능한 이미지 형식 중 하나를 선택할 수 있습니다.

그림 2. 시각화 데이터 복사 단추



시각화 데이터 복사. 이 조치는 시각화를 작성하는 데 사용되는 기본 데이터를 복사합니다. 데이터를 일반 텍스트 또는 HTML 형식의 텍스트로 클립보드에 복사합니다. 데이터를 다른 애플리케이션에 붙여넣을 때는 "선택하여 붙여넣기" 옵션을 선택하여 이러한 형식 중 하나를 선택할 수 있습니다.

⑮ 그래프보드 편집기 키보드 단축키

표 1. 키보드 단축키

단축키	기능
Ctrl+Space	탐색 모드와 편집 모드 간 전환
Delete	시각화 항목 삭제
Ctrl+Z	실행 취소
Ctrl+Y	다시 실행
F2	그래프에서 항목을 선택하기 위한 아웃라인 표시

⑯ 제목 및 꼬리말 추가

모든 그래프 유형에 대해 그래프에 표시되는 항목의 식별을 돕기 위해 고유 제목, 꼬리말 또는 축 레이블을 추가할 수 있습니다.

그래프에 제목 추가

1. 메뉴에서 **편집 > 그래프 제목 추가**를 선택하십시오. <TITLE>이 포함된 텍스트 상자가 그래프 위에 표시됩니다.
2. 편집 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 편집 모드**를 선택하십시오.
3. <TITLE> 텍스트를 두 번 클릭하십시오.
4. 필요한 제목을 입력한 후 Return을 누르십시오.

그래프에 꼬리말 추가

1. 메뉴에서 **편집 > 그래프 꼬리말 추가**를 선택하십시오. <FOOTNOTE>가 포함된 텍스트 상자가 그래프 아래에 표시됩니다.
2. 편집 모드에 있는지 확인하십시오. 메뉴에서 **보기 > 편집 모드**를 선택하십시오.
3. <FOOTNOTE> 텍스트를 두 번 클릭하십시오.
4. 필요한 제목을 입력한 후 Return을 누르십시오.

⑰ 그래프 스타일시트 사용

색상, 글꼴, 기호, 선 굵기 등의 기본 그래프 표시 정보는 스타일시트에 의해 제어됩니다. IBM® SPSS® Modeler와 함께 제공되는 기본 스타일시트가 있지만 필요한 경우 변경할 수 있습니다. 예를 들어, 그래프에서 사용할 프리젠테이션에 대한 공동 색상 구성표를 가질 수 있습니다. 자세한 정보는 시각화 편집의 내용을 참조하십시오.

그래프 노드에서 편집 모드를 사용하여 그래프 모양에 대해 스타일 변경사항을 작성할 수 있습니다. 그런 다음 **편집 > 스타일** 메뉴를 사용하여 변경사항을 현재 그래프 노드에서 이후에 생성되는 모든 그래프에 적용할 스타일시트로 저장하거나 IBM SPSS Modeler를 사용하여 생성하는 모든 그래프에 대한 새로운 기본 스타일시트로 저장할 수 있습니다.

편집 메뉴의 **스타일** 옵션에서는 다섯 가지 스타일시트 옵션을 사용할 수 있습니다.

- **스타일시트 전환.** 그래프 모양을 변경하기 위해 선택할 수 있는 저장된 다양한 스타일시트의 목록을 표시합니다. 자세한 정보는 스타일시트 적용의 내용을 참조하십시오.
- **노드에서 스타일 저장.** 현재 스트림의 동일한 그래프 노드에서 작성되는 향후 그래프에 적용되도록 수정사항을 선택된 그래프의 스타일에 저장합니다.
- **스타일을 기본값으로 저장.** 모든 스트림의 모든 그래프 노드에서 작성되는 모든 향후 그래프에 적용되도록 수정사항을 선택된 그래프의 스타일에 저장합니다. 이 옵션을 선택한 후에는 **기본 스타일 적용**을 사용하여 동일한 스타일을 사용하도록 다른 기존 그래프를 변경할 수 있습니다.
- **기본 스타일 적용.** 선택된 그래프의 스타일을 현재 기본 스타일로 저장되는 스타일로 변경합니다.
- **원본 스타일 적용.** 그래프의 스타일을 다시 원래 기본값으로 제공된 스타일로 변경합니다.

가. 스타일시트 적용

시각화의 스타일 특성을 지정하는 시각화 스타일시트를 적용할 수 있습니다. 예를 들어, 스타일시트는 글꼴, 대시 및 색상을 정의할 수 있습니다. 스타일시트를 사용하면 어느 정도까지는 수동으로 수행해야 할 편집을 쉽게 수행할 수 있습니다. 그러나 스타일시트는 *스타일* 변경에 한정됩니다. 범례 위치 또는 척도 범위와 같은 다른 변경은 스타일시트에 저장되지 않습니다.

스타일시트 적용 방법

1. 메뉴에서 다음을 선택하십시오.

편집 > 스타일 > 스타일시트 전환

2. 스타일시트 전환 대화 상자를 사용하여 스타일시트를 선택하십시오.

3. 대화 상자를 닫지 않고 시각화에 스타일시트를 적용하려면 **적용**을 클릭하십시오. 스타일시트를 적용하고 대화 상자를 닫으려면 **확인**을 클릭하십시오.

스타일시트 전환/선택 대화 상자

대화 상자의 맨 위에 있는 테이블은 현재 사용 가능한 모든 시각화 스타일시트를 나열합니다. 일부 스타일시트는 사전 설치되었고 나머지 스타일시트는 IBM® SPSS® Visualization Designer (별매품)에서 작성되었을 수 있습니다.

대화 상자 맨 아래에서는 표본 데이터를 사용한 시각화의 예를 표시합니다. 스타일시트 중 하나를 선택하여 해당 스타일을 시각화 예에 적용하십시오. 이러한 예는 스타일시트가 어떻게 실제 시각화에 영향을 주는지 판별하는 데 도움이 됩니다.

대화 상자는 또한 다음과 같은 옵션을 제공합니다.

기존 스타일. 기본적으로 스타일시트는 시각화의 모든 스타일을 겹쳐쓸 수 있습니다. 이 작동을 변경할 수 있습니다.

- **모든 스타일 겹쳐쓰기.** 스타일시트를 적용할 때 현재 편집 세션 동안 시각화에서 수정된 스타일을 포함하여 시각화의 모든 스타일을 겹쳐씹니다.
- **수정된 스타일 유지.** 스타일시트를 적용할 때 현재 편집 세션 동안 시각화에서 수정되지 *않은* 스타일만 겹쳐씹니다. 현재 편집 세션 동안 수정된 스타일은 유지됩니다.

관리. 컴퓨터에서 시각화 템플릿, 스타일시트 및 맵을 관리합니다. 로컬 시스템에서 시각화 템플릿, 스타일시트 및 맵을 가져오고, 내보내고, 이름을 바꾸고, 삭제할 수 있습니다. 자세한 정보는 템플릿, 스타일시트 및 맵 파일 관리의 내용을 참조하십시오.

위치. 시각화 템플릿, 스타일시트 및 맵이 저장된 위치를 변경합니다. 현재 위치는 단추의 오른쪽에 표시됩니다. 자세한 정보는 템플릿, 스타일시트 및 맵 위치 설정의 내용을 참조하십시오.

⑧ 그래프 인쇄, 저장, 복사 및 내보내기

각각의 그래프에는 그래프를 저장하거나 인쇄하거나 다른 형식으로 내보낼 수 있게 하는 다수의 옵션이 있습니다. 이 옵션은 대부분 파일 메뉴에서 사용할 수 있습니다. 또한 편집 메뉴에서 다른 애플리케이션에서 사용하기 위해 그래프, 그래프 내 데이터 또는 Microsoft Office Drawing Object를 복사하도록 선택할 수 있습니다.

인쇄

그래프를 인쇄하려면 **인쇄** 메뉴 항목 또는 단추를 사용하십시오. 인쇄하기 전에 **페이지 설정 및 인쇄 미리보기**를 사용하여 인쇄 옵션을 설정하고 출력을 미리 볼 수 있습니다. 페이지 머리글 및 바닥글 구성에 대한 자세한 정보는 머리글 및 바닥글 환경 설정 설정의 내용을 참조하십시오.

그래프 저장

그래프를 IBM® SPSS® Modeler 출력 파일(*.cou)에 저장하려면 메뉴에서 **파일 > 저장** 또는 **파일 > 다른 이름으로 저장**을 선택하십시오.

또는

그래프를 리포지토리에 저장하려면 메뉴에서 **파일 > 출력 저장**을 선택하십시오.

그래프 복사

MS Word 또는 MS PowerPoint 등의 다른 애플리케이션에서 사용하기 위해 그래프를 복사하려면 메뉴에서 **편집 > 그래프 복사**를 선택하십시오.

데이터 복사

MS Excel 또는 MS Word 등의 다른 애플리케이션에서 사용하기 위해 데이터를 복사하려면 메뉴에서 **편집 > 데이터 복사**를 선택하십시오. 기본적으로 데이터의 형식은 HTML로 지정됩니다. 붙여넣을 때 다른 형식 옵션을 보려면 다른 애플리케이션에서 **선택하여 붙여넣기**를 사용하십시오.

Microsoft Office Graphic Object 복사

그래프를 Microsoft Office Graphic Object로 복사하고 Excel 또는 PowerPoint와 같은 Microsoft Office 애플리케이션에서 사용할 수 있습니다. 그래프를 복사하려면 메뉴에서 **편집 > Microsoft Office Graphic Object 복사**를 선택하십시오. 콘텐츠가 클립보드에 복사되고 기본적으로 2진 형식이 됩니다. 붙여넣을 때 다른 형식 옵션을 지정하려면 Microsoft Office 애플리케이션에서 **선택하여 붙여넣기**를 사용하십시오.

일부 콘텐츠에서 이 기능을 지원하지 않을 수 있습니다. 이 경우 **Microsoft Office Graphic Object 복사** 메뉴 옵션을 사용하지 않습니다. Office 애플리케이션에 붙여넣기 후에는 그래프의 모양이 달라질 수 있지만 그래프 데이터는 동일합니다.

Excel에 복사하여 붙여넣을 수 있는 그래프 출력의 유형은 단순 막대형, 누적 막대형, 단순 상자도표, 군집 상자도표, 단순 분산형 및 그룹화 분산형의 6가지가 있습니다. 이러한 그래프 유형의 패널 및 애니메이션 옵션을 사용하는 경우 **Microsoft Office Graphic Object 복사** 옵션이 SPSS Modeler에서 사용되지 않습니다. 선택적 모양 또는 오버레이 등의 다른 설정에서는 옵션이 부분적으로 지원됩니다. 세부사항은 다음 표를 참조하십시오.

표 1. Microsoft Graphic Object 복사 지원

그래프 출력 템플릿	Modeler 그래프 노드	Modeler 그래프 유형	기본설정	선택적 모양	오버레이	Microsoft Graphic Object 복사 지원	설명
단순 막대형	그래프보드	막대	예	아니오	해당사항 없음	예	
		개수 막대	예	아니오	해당사항 없음	예	
	분포	막대	예	해당사항 없음	아니오	예	
누적 막대형	그래프보드	막대	예	예	해당사항 없음	예(제한 있음)	선택적 모양의 범주형 변수에만 예
		개수 막대	예	예	해당사항 없음	예(제한 있음)	선택적 모양의 범주형 변수에만 예
	분포	막대	예	해당사항 없음	예	예	

그래프 출력 템플릿	Modeler 그래프 노드	Modeler 그래프 유형	기본설정	선택적 모양	오버레이	Microsoft Graphic Object 복사 지원	설명
상자도표	그래프보 드	상자도표	예	아니오	해당사항 없음	예(제한 있음)	Windows에 서만 예
		상자도표	예	예	해당사항 없음	아니오	
군집 상자도표	그래프보 드	군집 상자도표	예	아니오	해당사항 없음	예(제한 있음)	Windows에 서만 예
		군집 상자도표	예	예	해당사항 없음	아니오	
단순 분산형	그래프보 드	거품 도표	예	아니오	해당사항 없음	예(제한 있음)	X와 Y 필드 들 다의 연속형 변수와 크기의 범주형 변수에만 예
		산점도	예	아니오	해당사항 없음	예(제한 있음)	X와 Y 필드 들 다의 연속형 변수에만 예
	도표	점	예	해당사항 없음	아니오	예(제한 있음)	X와 Y 필드 들 다의 연속형 변수에만 예
그룹화 분산형	그래프보 드	거품 도표	예	예	해당사항 없음	아니오	
		산점도	예	예	해당사항 없음	예(제한 있음)	X와 Y 필드 들 다의 연속형 변수와 선택적 모양의 범주 변수에만 예
	도표	점	예	해당사항 없음	아니오	예(제한 있음)	X와 Y 필드 들 다의 연속형 변수와 오버레이의 범주형 변수에만 예

그래프 내보내기

그래프 내보내기 옵션을 사용하면 다른 애플리케이션에서 사용하기 위해 비트맵(.bmp), JPEG(.jpg), PNG(.png), HTML(.html), PDF(.pdf) 또는 ViZml 문서(.xml) 형식 중 하나로 그래프를 내보낼 수 있습니다.

 **참고:** PDF 옵션이 선택되면 그래픽의 크기로 다듬어진 고해상도 PDF 파일로 그래프를 내보냅니다.

그래프를 내보내려면 메뉴에서 **파일 > 그래프 내보내기**를 선택한 후 형식을 선택하십시오.

테이블 내보내기

테이블 내보내기 옵션을 사용하면 탭 구분(.tab), 심표 구분(.csv) 또는 HTML(.html) 형식 중 하나로 테이블을 내보낼 수 있습니다.

테이블을 내보내려면 메뉴에서 **파일 > 테이블 내보내기**를 선택한 후 형식을 선택하십시오.

이 절의 나머지 부분에서는 그래프를 작성하여 해당 출력 창에서 사용하기 위한 특정 옵션에 초점을 둡니다.

가. 머리글 및 바닥글 환경 설정 설정

그래프 및 기타 유형의 출력의 경우 페이지가 인쇄될 때 머리글 및 바닥글에 대한 옵션을 지정할 수 있습니다.

페이지 머리글. 인쇄 및 내보내기를 위해 페이지 머리글의 유형 및 위치를 선택하십시오.

페이지 바닥글. 인쇄 및 내보내기를 위해 페이지 바닥글의 유형 및 위치를 선택하십시오.

글꼴 옵션. 드롭 다운 목록에서 글꼴 및 해당 포인트 크기를 선택하십시오. 굵은체 또는 기울임꼴 단추를 클릭하여 이 효과를 추가하십시오.

머리글 및 바닥글 주위에 테두리 그리기. 얇은 직사각형 테두리로 머리글 및 바닥글을 둘러싸려면 선택하십시오.

5) 출력 노드

(1) 출력 노드 개요

출력 노드는 데이터 및 모형에 대한 정보를 얻을 수 있는 방법을 제공합니다. 또한 기타 소프트웨어 도구와 접속할 수 있도록 데이터를 다양한 형식으로 내보낼 수 있는 메커니즘을 제공합니다.

다음 출력 노드를 사용할 수 있습니다.



테이블 노드는 데이터를 표 형식으로 표시하는데, 이것을 파일에 쓸 수도 있습니다. 이것은 쉽게 읽을 수 있는 양식으로 데이터 값을 조사하거나 내보내야 할 때 유용합니다.



행렬 노드는 필드 사이의 관계를 표시하는 테이블을 작성합니다. 두 기호 필드 사이의 관계를 표시하기 위해 가장 일반적으로 사용하지만, 플래그 필드나 수치 필드 사이의 관계도 표시할 수 있습니다.



분석 노드는 정확한 예측을 생성하기 위한 예측 모델의 능력을 평가합니다. 분석 노드는 하나 이상의 모델 너깃에 대해 예측값과 실제 값 사이의 다양한 비교를 수행합니다. 또한 예측 모델을 서로 비교할 수도 있습니다.



데이터 검토 노드는 요약 통계량, 각 필드에 대한 히스토그램과 분포뿐 아니라 이상값, 결측값, 극단값에 대한 정보를 포함하여 데이터에 대한 포괄적인 정보를 간략하게 제공합니다. 결과는 전체 크기 그래프 및 데이터 준비 노드를 생성하기 위해 정렬하고 사용할 수 있는 읽기 쉬운 행렬로 표시됩니다.



변환 노드를 사용하면 선택된 필드에 적용하기 전에 변환 결과를 선택하고 시각적으로 미리볼 수 있습니다.



통계량 노드는 수치 필드에 관한 기본 요약 정보를 제공합니다. 개별 필드에 대한 요약 통계량 및 필드 사이의 상관계수를 계산합니다.



평균 노드는 독립 집단 사이 또는 관련된 필드의 쌍 사이의 평균을 비교하여 상당한 차이가 존재하는지 여부를 검정합니다. 예를 들어, 프로모션을 실행하기 전후의 평균 수익을 비교하거나 프로모션을 받지 않은 고객과 받은 고객으로부터의 수익을 비교할 수 있습니다.



보고서 노드는 고정 텍스트뿐 아니라 데이터 및 데이터로부터 파생된 기타 표현식을 포함한 형식화된 보고서를 작성합니다. 텍스트 템플릿을 사용하여 보고서의 형식을 지정하여 고정 텍스트 및 데이터 출력 생성을 정의합니다. 템플릿에서 HTML 태그를 사용하고 출력 탭에서 옵션을 설정하여 사용자 정의 텍스트 형식화를 제공할 수 있습니다. 템플릿에서 CLEM 표현식을 사용하여 데이터 값과 기타 조건부 출력을 포함할 수 있습니다.



전역값 설정 노드는 데이터를 스캔하고 CLEM 표현식에서 사용할 수 있는 요약 값을 계산합니다. 예를 들어, 이 노드를 사용하여 age라는 필드에 대한 통계량을 계산한 후 @GLOBAL_MEAN(age) 함수를 삽입하여 CLEM 표현식에서 age의 전체 평균을 사용할 수 있습니다.



시뮬레이션 적합 노드는 각 필드에 있는 데이터의 통계 분포를 분석하고 각 필드에 최상의 적합 분포가 지정된 시뮬레이션 생성 노드를 생성(또는 업데이트)합니다. 시뮬레이션 생성 노드를 사용하여 시뮬레이션된 데이터를 생성할 수 있습니다.



시뮬레이션 평가 노드는 지정된 예측 대상 필드를 평가하고 대상 필드에 관한 분포 및 상관관계 정보를 제공합니다.

(2) 출력 관리

출력 관리자는 IBM® SPSS® Modeler 세션 동안 생성된 차트, 그래프 및 테이블을 표시합니다. 출력 관리자에서 출력을 두 번 클릭하여 항상 출력을 다시 열 수 있습니다. 해당 스트림 또는 노드를 재실행하지 않아도 됩니다.

출력 관리자 보기

보기 메뉴를 열고 **관리자**를 선택하십시오. **출력** 탭을 클릭하십시오.

출력 관리자에서 다음을 수행할 수 있습니다.

- 히스토그램, 평가 차트, 테이블 등의 기존 출력 오브젝트 표시
- 출력 오브젝트의 이름 바꾸기
- 디스크 또는 IBM SPSS Collaboration and Deployment Services Repository에 출력 오브젝트 저장(사용 가능한 경우).

- 현재 프로젝트에 출력 파일 추가
- 현재 세션에서 저장되지 않은 출력 오브젝트 삭제
- 저장된 출력 오브젝트 열기 또는 IBM SPSS Collaboration and Deployment Services Repository에서 저장된 출력 오브젝트 검색(사용 가능한 경우)

이 옵션에 액세스하려면 출력 탭을 마우스 오른쪽 단추로 클릭하십시오.

(3) 출력 보기

화면 출력은 출력 브라우저 창에 표시됩니다. 출력 브라우저 창에는 출력을 인쇄 또는 저장하거나 다른 형식으로 내보낼 수 있는 메뉴 세트가 있습니다. 출력 유형에 따라 특정 옵션은 다를 수 있습니다.

데이터 인쇄, 저장 및 내보내기. 다음과 같이 자세한 정보를 사용할 수 있습니다.

- 출력을 인쇄하려면 인쇄 메뉴 옵션 또는 단추를 사용하십시오. 인쇄하기 전에 **페이지 설정 및 인쇄 미리보기**를 사용하여 인쇄 옵션을 설정하고 출력을 미리 볼 수 있습니다. 페이지 머리글 및 바닥글 구성에 대한 자세한 정보는 머리글 및 바닥글 환경 설정 설정의 내용을 참조하십시오.
- 출력을 IBM® SPSS® Modeler 출력 파일(.cou)에 저장하려면 파일 메뉴에서 **저장** 또는 **다른 이름으로 저장**을 선택하십시오.
- 텍스트 또는 HTML 등의 다른 형식으로 출력을 저장하려면 파일 메뉴에서 **내보내기**를 선택하십시오. 자세한 정보는 출력 내보내기 주제를 참조하십시오. 출력에 해당 형식으로 내보내기에 적합한 데이터가 포함된 경우에만 해당 형식을 선택할 수 있습니다. 예를 들어, 의사결정 트리의 콘텐츠는 텍스트로 내보낼 수 있으나 K-평균 모델의 콘텐츠는 텍스트로는 의미가 전달되지 않습니다.
- 다른 사용자가 IBM SPSS Collaboration and Deployment Services Deployment Portal을 사용하여 출력을 볼 수 있도록 공유 리포지토리에 출력을 저장하려면 파일 메뉴에서 **웹에 출판**을 선택하십시오. 이 옵션을 사용하려면 IBM SPSS Collaboration and Deployment Services에 대한 별도의 라이선스가 필요합니다.

셀 및 열 선택. 편집 메뉴에는 현재 출력 유형에 맞게 셀 및 열을 선택, 선택 취소 및 복사하기 위한 다양한 옵션이 포함되어 있습니다. 자세한 정보는 셀 및 열 선택을 참조하십시오.

새 노트 생성. 생성 메뉴를 사용하면 출력 브라우저의 콘텐츠를 기반으로 하여 새 노트를 생성할 수 있습니다. 옵션은 출력 유형 및 현재 선택된 출력 내의 항목에 따라 다릅니다. 특정 유형의 출력에 대한 노트 생성 옵션에 대한 세부사항은 해당 출력에 대한 문서를 참조하십시오.

① 웹에 출판

웹에 출판 기능을 사용하면 특정 유형의 스트림 출력을 IBM® SPSS® Collaboration and Deployment Services의 기초가 되는 중앙 공유 IBM SPSS Collaboration and Deployment Services Repository에 출판할 수 있습니다. 이 옵션을 사용하면 이 출력을 볼 필요가 있는 다른 사용자가 IBM SPSS Modeler를 설치할 필요 없이 인터넷 액세스 및 IBM SPSS Collaboration and Deployment Services 계정을 사용하여 출력을 볼 수 있습니다.

다음은 웹에 출판 기능을 지원하는 IBM SPSS Modeler 노드를 나열한 표입니다. 이러한 노드의 출력은 출력 오브젝트(.cou) 형식으로 IBM SPSS Collaboration and Deployment Services Repository에 저장되며 IBM SPSS Collaboration and Deployment Services Deployment Portal에서 직접 볼 수 있습니다.

기타 유형의 출력은 사용자의 시스템에 관련 애플리케이션(예를 들어, 스트림 오브젝트의 경우 IBM SPSS Modeler)이 설치된 경우에만 볼 수 있습니다.

표 1. 웹에 출판을 지원하는 노드

노드 유형	노드
그래프	모두
출력	테이블
	교차표
	데이터 검토
	변환
	평균
	분석
	통계량
	보고서(HTML)
	IBM SPSS Statistics

가. 웹에 출력 출판

웹에 출력을 출판하려면 다음을 수행하십시오.

1. IBM® SPSS® Modeler 스트림에서 표에 나열된 노드 중 하나를 실행하십시오. 그러면 새 창에 출력 오브젝트(표, 교차표 또는 보고서 오브젝트 등)가 작성됩니다.

2. 출력 오브젝트 창에서 다음을 선택하십시오.

파일 > 웹에 출판

참고: 표준 웹 브라우저와 함께 사용하도록 단순 HTML 파일만 내보내려면 파일 메뉴에서 **내 보내기**를 선택하고 **HTML**을 선택하십시오.

3. IBM SPSS Collaboration and Deployment Services Repository에 연결하십시오.
연결되면 여러 가지 저장 공간 옵션을 제공하는 리포지토리: 저장 대화 상자가 표시됩니다.
4. 원하는 저장 공간 옵션을 선택하였으면 **저장**을 클릭하십시오.

나. 웹을 통해 출판된 출력 보기

이 기능을 사용하려면 IBM® SPSS® Collaboration and Deployment Services 계정이 설정되어 있어야 합니다. 볼 오브젝트 유형과 관련된 애플리케이션(예: IBM SPSS Modeler 또는 IBM SPSS Statistics)이 설치되어 있는 경우에는 브라우저가 아니라 애플리케이션 자체에 출력이 표시됩니다.

웹을 통해 출판된 출력을 보려면 다음을 수행하십시오.

1. 브라우저에서 `http://<repos_host>:<repos_port>/peb`으로 이동하십시오.
여기서, `repos_host` 및 `repos_port`는 IBM SPSS Collaboration and Deployment Services 호스트의 호스트 이름 및 포트 번호입니다.
2. IBM SPSS Collaboration and Deployment Services 계정에 대한 로그인 세부사항을 입력하십시오.
3. **컨텐츠 리포지토리**를 클릭하십시오.
4. 볼 오브젝트로 이동하거나 검색하십시오.
5. 오브젝트 이름을 클릭하십시오. 그래프 등의 일부 오브젝트 유형에 대해서는 오브젝트가 브라우저에서 렌더링되는 동안 지연될 수 있습니다.

② HTML 브라우저에서 출력 보기

선형, 로지스틱 및 PCA/Factor 모델 너깃의 고급 탭에서 Internet Explorer와 같은 별도의 브라우저에서 표시되는 정보를 볼 수 있습니다. 정보는 회사 인트라넷 또는 인터넷 사이트와 같이 어디서나 저장하고 재사용할 수 있는 HTML 등의 출력입니다.

브라우저에 정보를 표시하려면 모델 너깃의 고급 탭의 왼쪽 상단에 있는 모델 아이콘 아래의 시작 단추를 클릭하십시오.

③ 출력 내보내기

출력 브라우저 창에서 출력을 텍스트 또는 HTML과 같은 다른 형식으로 내보내도록 선택할 수 있습니다. 내보내기 형식은 출력의 유형에 따라 다르지만 일반적으로 출력을 생성하기 위해 사용한 노드에서 **파일에 저장**을 선택할 때 사용할 수 있는 파일 유형 옵션과 유사합니다.

참고: 출력에 해당 형식으로 내보내기에 적합한 데이터가 포함된 경우에만 해당 형식을 선택할 수 있습니다. 예를 들어, 의사결정 트리의 콘텐츠는 텍스트로 내보낼 수 있으나 K-평균 모델의 콘텐츠는 텍스트로는 의미가 전달되지 않습니다.

출력을 내보내려면 다음을 수행하십시오.

1. 출력 브라우저에서 파일 메뉴를 열고 내보내기를 선택하십시오. 그런 다음 생성할 파일 유형을 선택하십시오.
 - **탭 구분 데이터(*.tab).** 이 옵션은 데이터 값을 포함하는 형식화된 텍스트 파일을 생성합니다. 이 스타일은 정보를 다른 애플리케이션으로 가져올 수 있는 일반 텍스트 표시를 생성하는 경우에 유용할 때가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
 - **coma 구분 데이터(*.dat).** 이 옵션은 데이터 값을 포함하는 coma 구분 텍스트 파일을 생성합니다. 이 유형은 스프레드시트 또는 기타 데이터 분석 애플리케이션으로 가져올 수 있는 데이터 파일을 생성하는 빠른 방법으로 유용한 경우가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
 - **전치된 탭 구분 데이터(*.tab).** 이 옵션은 탭 구분 데이터와 동일하나 데이터가 전치되어 행이 필드를 나타내고 열이 레코드를 나타낸다는 점만 다릅니다.
 - **전치된 coma 구분 데이터(*.dat).** 이 옵션은 coma 구분 데이터와 동일하나 데이터가 전치되어 행이 필드를 나타내고 열이 레코드를 나타낸다는 점만 다릅니다.
 - **HTML (*.html).** 이 옵션은 파일에 HTML 형식의 출력을 씁니다.

④ 셀 및 열 선택

표 노드, 교차표 노드 및 평균 노드를 포함하여 많은 노드가 표 형식의 출력을 생성합니다. 이러한 출력 표는 셀 선택, 클립보드에 표의 전부 또는 일부 복사, 현재 선택을 기반으로 하여 새 노드 생성, 표 저장 및 인쇄 등을 포함하여 유사한 방법으로 보고 조작할 수 있습니다.

셀 선택. 셀을 선택하려면 해당 셀을 클릭하십시오. 직사각형 범위의 셀을 선택하려면 원하는 범위의 한 코너를 클릭하고 마우스를 해당 범위의 다른 코너로 끈 다음 마우스 단추를 놓으십시오. 전체 열을 선택하려면 열 머리말을 클릭하십시오. 다중 열을 선택하려면 열 머리말에서 Shift-클릭 또는 Ctrl-클릭을 사용하십시오.

새로 선택하면 이전 선택이 지워집니다. 선택하는 동안 Ctrl 키를 아래로 누르고 있으면 이전 선택을 지우지 않고 새 선택을 기존 선택에 추가할 수 있습니다. 이 방법을 사용하여 연속되지 않은 다중 표 영역을 선택할 수 있습니다. 편집 메뉴에는 **모두 선택** 및 **선택 지우기** 옵션도 포함됩니다.

열 다시 정렬. 테이블 노드 및 평균 노드 출력 브라우저를 사용하면 열 머리말을 클릭하고 이를 원하는 위치에 끌어다 놓음으로써 표에서 열을 이동할 수 있습니다. 한 번에 한 열만 이동할 수 있습니다.

(4) 테이블 노드

테이블 노드는 데이터의 값을 나열하는 테이블을 작성합니다. 스트림의 모든 필드 및 모든 값이 포함되므로 읽기 쉬운 양식으로 데이터 값을 조사하거나 내보내는 데 유용합니다. 선택적으로, 특정 조건을 충족시키는 레코드를 강조표시할 수 있습니다.

참고: 작업 중인 데이터 세트가 작지 않은 경우에는 테이블 노드에 전달할 데이터의 서브 세트를 선택하는 것이 좋습니다. 레코드 수가 표시 구조에 포함될 수 있는 크기(예: 1억 행)를 초과하면 테이블 노드는 데이터를 적절히 표시할 수 없습니다.

① 테이블 노드 설정 탭

레코드 강조표시 조건. 레코드 강조표시 조건을 충족시키는 CLEM 표현식을 입력하여 테이블에서 레코드를 강조표시할 수 있습니다. 이 옵션은 **화면에 출력**을 선택한 경우에만 사용됩니다.

② 출력 노드 출력 탭

표 유형의 출력을 생성하는 노드의 경우, 출력 탭을 사용하면 결과의 형식 및 위치를 지정할 수 있습니다.

출력 이름. 노드가 실행될 때 생성되는 출력의 이름을 지정합니다. **자동**은 출력을 생성하는 노드를 기반으로 이름을 선택합니다. 필요에 따라 **사용자 정의**를 선택하여 다른 이름을 지정할 수 있습니다.

화면에 출력(기본값). 온라인으로 볼 출력 오브젝트를 생성합니다. 출력 오브젝트는 출력 노드가 실행될 때 관리자 창의 출력 탭에 표시됩니다.

파일로 출력. 노드가 실행될 때 출력을 파일에 저장합니다. 이 옵션을 선택하는 경우, 파일 이름을 입력하거나 디렉토리로 이동하여 파일 선택자 단추를 사용하여 파일 이름을 지정한 다음 파일 유형을 선택하십시오. 일부 파일 유형은 특정 유형의 출력에 대해 사용 불가능합니다.

참고:

출력 노드의 출력 데이터는 다음 규칙에 따라 인코딩됩니다.

- 출력 노드를 실행할 때 스트림 인코딩 값(스트림 옵션 탭에 설정됨)이 출력으로 설정됩니다.
- 출력이 생성된 후에는 스트림 인코딩이 변경되어도 인코딩은 변경되지 않습니다.
- 출력 노드 출력을 내보낼 때 출력 파일은 현재 정의된 스트림 인코딩을 사용하여 내보냅니다. 출력이 작성된 후에는 스트림 인코딩을 변경하더라도 생성된 출력에 영향을 미치지 않습니다.

이러한 규칙에는 다음과 같은 예외가 적용됩니다.

- 모든 HTML 내보내기는 UTF-8 형식으로 인코딩됩니다.
- 확장 출력 노드의 출력은 사용자 정의 사용자 스크립트를 통해 생성됩니다. 따라서 인코딩은 스크립트를 통해 제어됩니다.

파일에 출력을 저장하는 데 다음 옵션을 사용할 수 있습니다.

- **데이터(탭 구분 데이터)(*tab)**. 이 옵션은 데이터 값을 포함하는 형식화된 텍스트 파일을 생성합니다. 이 스타일은 정보를 다른 애플리케이션으로 가져올 수 있는 일반 텍스트 표시를 생성하는 경우에 유용할 때가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
- **데이터(coma 구분 데이터)(*dat)**. 이 옵션은 데이터 값을 포함하는 coma 구분 텍스트 파일을 생성합니다. 이 유형은 스프레드시트 또는 기타 데이터 분석 애플리케이션으로 가져올 수 있는 데이터 파일을 생성하는 빠른 방법으로 유용한 경우가 많습니다. 이 옵션은 표, 교차표 및 평균 노드에 대해 사용할 수 있습니다.
- **HTML (*.html)**. 이 옵션은 파일에 HTML 형식의 출력을 씁니다. (표, 교차표 또는 평균 노드의) 표 형식의 출력인 경우, HTML 파일 세트에 HTML 표 내에 필드 이름 및 데이터를 나열하는 콘텐츠 패널이 포함됩니다. 표의 행의 수가 **페이지당 선** 지정 사항을 초과하면 표가 다중 HTML 파일로 분할될 수 있습니다. 이 경우, 콘텐츠 패널에 모든 표 페이지에 대한 링크가 포함되며 표를 탐색할 수 있는 방법이 제공됩니다. 표가 아닌 출력의 경우, 노드 결과를 포함하는 단일 HTML 파일이 생성됩니다.

참고: HTML 출력에 첫 번째 페이지에 대한 형식화만 포함된 경우, **출력 페이지 번호 매기기를** 선택하고 모든 출력이 단일 페이지에 포함되도록 **페이지당 선** 지정 사항을 조정하십시오. 또는 보고서 노드와 같이 노드에 대한 출력 템플릿이 사용자 정의 HTML 태그를 포함하는 경우에는 형식 유형으로 **사용자 정의**를 지정해야 합니다.

- **텍스트 파일(*.txt)**. 이 옵션은 출력을 포함하는 텍스트 파일을 생성합니다. 이 스타일은 워드 프로세서 또는 프리젠테이션 소프트웨어 등의 다른 애플리케이션으로 가져올 수 있는 출력을 생성하는 경우에 유용할 때가 많습니다. 이 옵션은 일부 노드에는 사용할 수 없습니다.
- **출력 오브젝트(*.cou)**. 이 형식으로 저장되는 출력 오브젝트는 IBM® SPSS® Modeler에서 열고 볼 수 있으며 프로젝트에 추가할 수 있으며 IBM SPSS Collaboration and Deployment Services Repository를 사용하여 공개하고 추적할 수 있습니다.

출력 보기. 평균 노드에 대해서 기본적으로 단순 또는 고급 출력이 표시되도록 지정할 수 있습니다. 또한 생성된 출력을 찾아볼 때 이러한 보기 사이를 토글할 수 있습니다. 자세한 정보는 평균 노드 출력 브라우저 주제를 참조하십시오.

형식. 보고서 노드의 경우, 출력이 자동으로 형식화되거나 템플릿에 포함된 HTML을 사용하여 형식화되도록 선택할 수 있습니다. 템플릿에서 HTML 형식화를 허용하도록 **사용자 정의**를 선택하십시오.

제목. 보고서 노드의 경우, 보고서 출력의 맨 위에 표시될 선택적 제목 텍스트를 지정할 수 있습니다.

삽입된 텍스트 강조표시. 보고서 노드의 경우, 보고서 템플릿의 CLEM 표현식을 사용하여 생성된 텍스트를 강조표시하려면 이 옵션을 선택하십시오. 자세한 정보는 보고서 노드 템플릿 탭의 내용을 참조하십시오. **사용자 정의 형식화**를 사용하는 경우에는 이 옵션을 권장하지 않습니다.

페이지당 선. 보고서 노드의 경우, 출력 보고서의 자동 형식화 동안 각 페이지에 포함할 선 수를 지정합니다.

데이터 전치. 이 옵션은 행이 필드를 나타내고 열이 레코드를 나타내도록 데이터를 내보내기 전에 전치합니다.

참고: 큰 표의 경우, 특히 원격 서버로 작업하는 경우 위 옵션이 비능률적일 수 있습니다. 이런 경우, 파일 출력 노드를 사용하면 성능이 훨씬 개선됩니다. 자세한 정보는 플랫폼 파일 내보내기 노드 주제를 참조하십시오.

③ 테이블 브라우저

테이블 브라우저는 표 형식 데이터를 표시하며 여기서 셀 선택 및 복사, 열 재정렬, 테이블 저장 및 인쇄를 포함한 표준 조작을 수행할 수 있습니다. 자세한 정보는 셀 및 열 선택의 내용을 참조하십시오. 이러한 조작은 노드에서 데이터를 미리볼 때 수행할 수 있는 조작과 동일합니다.

테이블 데이터 내보내기. 다음을 선택하여 테이블 브라우저에서 데이터를 내보낼 수 있습니다.

파일 > 내보내기

자세한 정보는 출력 내보내기의 내용을 참조하십시오.

Windows 제어판에 지정되어 있거나 분산 모드에서 실행 중인 경우 서버 컴퓨터에 지정된 시스템 기본 인코딩 형식으로 데이터를 내보냅니다.

테이블 검색. 주 도구 모음의 검색 단추(쌍안경 아이콘 포함)가 검색 도구 모음을 활성화하며 이 검색 도구를 사용하여 테이블에서 특정 값을 검색할 수 있습니다. 테이블에서 앞 또는 뒤로 검색할 수 있고 대소문자 구분 검색(Aa 단추)을 지정할 수 있으며 검색 중단 단추로 진행 중인 검색을 중단할 수 있습니다.

새 노드 생성. 생성 메뉴에는 노드 생성 작업이 포함됩니다.

- **Select Node ("Records").** 테이블의 셀이 선택된 레코드를 선택하는 선택 노드를 생성합니다.
- **Select ("And").** 테이블에서 선택된 모든 값을 포함하는 레코드를 선택하는 선택 노드를 생성합니다.
- **Select ("Or").** 테이블에서 선택된 값 중 임의 값을 포함하는 레코드를 선택하는 선택 노드를 생성합니다.
- **Derive ("Records").** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드는 테이블의 임의 셀이 선택된 레코드의 경우 *T*를 포함하고 나머지 레코드의 경우 *F*를 포함합니다.
- **Derive ("And").** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드는 테이블에서 선택된 모든 값을 포함하는 레코드의 경우 *T*를 포함하고 나머지 레코드의 경우 *F*를 포함합니다.
- **Derive ("Or").** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드는 테이블에서 선택된 값 중 임의 값을 포함하는 레코드의 경우 *T*를 포함하고 나머지 레코드의 경우 *F*를 포함합니다.

(5) 교차표 노드

교차표 노드를 사용하면 필드 간 관계를 표시하는 테이블을 작성할 수 있습니다. 이는 두 범주형 필드(플래그, 명목 또는 순서) 간 관계를 표시하는 데 가장 일반적으로 사용되지만 연속형(숫자 범위) 필드 간 관계를 표시하는 데도 사용될 수 있습니다.

① 교차표 노드 설정 탭

설정 탭에서는 교차표의 구조에 대한 옵션을 지정할 수 있습니다.

필드. 다음 옵션에서 필드 선택 유형을 선택하십시오.

- **선택.** 이 옵션을 사용하면 행에 대한 범주형 필드 하나와 교차표의 열에 대한 범주형 필드 하나를 선택할 수 있습니다. 교차표의 행 및 열은 선택된 범주형 필드에 대한 값의 목록에 의해 정의됩니다. 교차표의 셀에는 아래에서 선택된 요약 통계가 포함되어 있습니다.
- **모든 플래그(참 값).** 이 옵션은 데이터의 각 플래그 필드에 대해 하나의 행 및 열을 가진 교차표를 요청합니다. 교차표의 셀에는 각 플래그 조합에 대한 이중 긍정의 개수가 포함되어 있습니다. 즉, **빵 구입**에 해당하는 행과 **치즈 구입**에 해당하는 열의 경우 해당 행 및 열의 교차점에 있는 셀에는 **빵 구입**과 **치즈 구입**이 모두 참인 레코드의 수가 포함되어 있습니다.

- **모든 숫자.** 이 옵션은 각 숫자 필드에 대해 하나의 행 및 열을 가진 교차표를 요청합니다. 교차표의 셀은 해당 필드 쌍에 대한 교차곱의 합계를 나타냅니다. 즉, 교차표의 각 셀에 대해 행 필드 및 열 필드의 값이 각 레코드에 대해 곱해진 후 레코드에 대해 합계가 계산됩니다.

결측값 포함. 사용자 결측(공백) 및 시스템 결측(\$null\$) 값을 행 및 열 출력에 포함합니다. 예를 들어, N/A 값이 선택된 열 필드에 대해 사용자 결측으로 정의된 경우에는 다른 범주와 마찬가지로 N/A라는 별도의 열이 테이블에 포함됩니다(이 값이 실제로 데이터에서 발생한다고 가정함). 이 옵션이 선택 취소되면 발생 빈도에 관계없이 N/A 열은 제외됩니다.

참고: 결측값을 포함하는 옵션은 선택된 필드가 교차 분석표인 경우에만 적용됩니다. 공백값은 \$null\$에 매핑되며 모드가 **선택됨**이고 콘텐츠가 **함수**로 설정된 경우 함수 필드에 대한 통합에서 제외되고 모드가 **모든 숫자**로 설정된 경우 모든 숫자 필드에 대한 통합에서 제외됩니다.

셀 콘텐츠. 위에서 **선택** 필드를 선택한 경우에는 교차표의 셀에서 사용할 통계를 지정할 수 있습니다. 개수 기반 통계를 선택하거나 오버레이 필드를 선택하여 행 및 열 필드의 값을 기반으로 숫자 필드의 값을 요약하십시오.

- **교차 분석표.** 셀 값은 해당 값 조합을 가진 레코드 수의 백분율 및/또는 개수입니다. 모양 탭의 옵션을 사용하여 원하는 교차 분석표 요약을 지정할 수 있습니다. 전역값 카이제곱 값도 유의수준과 함께 표시됩니다. 자세한 정보는 교차표 노드 출력 브라우저의 내용을 참조하십시오.
- **함수.** 요약 함수를 선택하는 경우 셀 값은 적절한 행 및 열 값을 가진 케이스에 대해 선택된 오버레이 필드 값의 함수입니다. 예를 들어, 행 필드가 **지역**이고 열 필드가 **제품**인 경우 오버레이 필드가 **수입**이면 **북동 지역** 행의 셀과 **위젯** 열은 북동 지역에서 판매된 위젯에 대한 수입의 합계(또는 평균, 최소값, 최대값)를 포함합니다. 기본 요약 함수는 **평균**입니다. 함수 필드를 요약하기 위해 다른 함수를 선택할 수 있습니다. 옵션으로는 **평균**, **합계**, **SDev(표준 편차)**, **최대값** 및 **최소값**이 있습니다.

② 교차표 노드 모양 탭

모양 탭에서는 교차 분석표 교차표에 대해 제공되는 통계와 교차표에 대한 정렬 및 강조표시 옵션을 제거할 수 있습니다.

행 및 열. 교차표에서 행 및 열 표제의 정렬을 제어합니다. 기본값은 **정렬되지 않음**입니다. **오름차순** 또는 **내림차순**을 선택하여 지정된 방향으로 행 및 열 표제를 정렬하십시오.

오버레이. 교차표에서 극단값을 강조표시할 수 있게 합니다. 값은 셀 개수(교차 분석표 교차표의 경우) 또는 계산된 값(함수 교차표의 경우)을 기반으로 강조표시됩니다.

- **맨 위 강조표시.** 교차표에서 가장 높은 값을 빨간색으로 강조표시하도록 요청할 수 있습니다. 강조표시할 값 수를 지정하십시오.

- **맨 아래 강조표시.** 교차표에서 가장 낮은 값을 녹색으로 강조표시하도록 요청할 수도 있습니다. 강조표시할 값 수를 지정하십시오.

참고: 두 강조표시 옵션에 대해 동률을 사용하면 요청한 것보다 많은 값을 강조표시할 수 있습니다. 예를 들어, 셀 사이에 6개의 0(영)이 있는 교차표가 있을 때 **맨 아래 5개 강조표시**를 요청하면 6개의 0(영)이 모두 강조표시됩니다.

교차 분석표 셀 내용. 교차 분석표의 경우 교차 분석표 교차표에 대해 교차표에 포함된 요약 통계를 지정할 수 있습니다. 이 옵션은 설정 탭에서 **모든 숫자** 또는 **함수** 옵션을 선택하는 경우 사용할 수 없습니다.

- **개수.** 셀에는 해당 열 값을 가진 행 값이 포함된 레코드의 수가 포함되어 있습니다. 이는 유일한 기본 셀 내용입니다.
- **기대값.** 행과 열 사이에 관계가 없다고 가정했을 때 셀에 있는 레코드의 수에 대한 기대값입니다. 기대값은 다음 수식을 기반으로 합니다.

$$p(\text{row value}) * p(\text{column value}) * \text{total number of records}$$

- **잔차.** 관측값과 기대값 사이의 차이입니다.
- **행의 백분율.** 해당 열 값을 가진 행 값이 포함된 모든 레코드의 백분율입니다. 행 내에서 백분율을 합계는 100입니다.
- **열의 백분율.** 해당 행 값을 가진 열 값이 포함된 모든 레코드의 백분율입니다. 열 내에서 백분율을 합계는 100입니다.
- **총계의 백분율.** 열 값과 행 값의 조합을 가진 모든 레코드의 백분율입니다. 전체 교차표에서 백분율 합계는 100입니다.
- **행 및 열 총계 포함.** 열 및 행 총계에 대한 교차표에 행 및 열을 추가합니다.
- **설정 적용.** (출력 브라우저 전용) 출력 브라우저를 닫은 후 다시 열지 않고 교차표 노드 출력의 모양을 변경할 수 있게 합니다. 출력 브라우저의 이 탭에서 변경사항을 작성하고 이 단추를 클릭한 후 교차표 탭을 선택하여 변경사항의 영향을 확인하십시오.

③ 교차표 노드 출력 브라우저

교차표 브라우저는 교차 분석표 데이터를 표시하며 교차표에 대해 셀 선택, 교차표의 전부 또는 일부를 클립보드에 복사, 교차표 선택사항을 기반으로 새 노드 생성, 교차표 저장 및 인쇄를 포함한 조작을 수행할 수 있게 합니다. 교차표 브라우저는 Oracle의 Naive Bayes 모델 등의 특정 모델의 출력을 표시하는 데도 사용할 수 있습니다.

파일 및 편집 메뉴는 출력 인쇄, 저장 및 내보내기와 데이터 선택 및 복사를 위한 일반적인 옵션을 제공합니다. 자세한 정보는 출력 보기 주제를 참조하십시오.

카이제곱 두 범주형 필드의 교차 분석표에 대해 전역값 Pearson의 카이제곱도 테이블 아래에 표시됩니다. 이 검정은 관계가 존재하지 않는 경우 예상하는 개수와 관측개수 간 차이를 기반으로 두 필드가 관련되지 않을 확률을 표시합니다. 예를 들어, 고객 만족도와 상점 위치 사이에 관계가 없으면 모든 상점에 대해 비슷한 만족도를 예상합니다. 하지만 특정 상점의 고객이 다른 고객보다 높은 비율을 지속적으로 보고하는 경우에는 우연의 일치라고 의심할 수 있습니다. 차이가 클수록 우연 표본추출 오류만의 결과였을 확률이 더 낮습니다.

- 카이제곱 검정은 두 필드가 관계가 없을 확률을 표시하며 이 경우 관측빈도와 기대빈도 간 차이는 우연만의 결과입니다. 이 확률이 매우 낮은 경우(일반적으로 5% 미만) 두 필드 간 관계를 유의하다고 합니다.
- 하나의 열 또는 하나의 행만 있는 경우(일원 카이제곱 검정) 자유도는 셀의 수에서 1을 뺀 값입니다. 이원 카이제곱의 경우 자유도는 행의 수에서 1을 뺀 값에 열의 수에서 1을 뺀 값을 곱한 값입니다.
- 셀 기대빈도가 5 미만인 경우에는 카이제곱 통계량 해석 시 주의하십시오.
- 카이제곱 검정은 두 필드의 교차 분석표에만 사용할 수 있습니다. (설정 탭에서 **모든 플래그** 또는 **모든 숫자**가 선택되면 이 검정이 표시되지 않습니다.)

생성 메뉴. 생성 메뉴에는 노드 생성 작업이 포함됩니다. 이 조작은 교차 분석표 교차표에만 사용할 수 있으므로 교차표에서 하나 이상의 셀이 선택되어 있어야 합니다.

- **선택 노드.** 교차표에서 선택된 셀과 일치하는 레코드를 선택하는 선택 노드를 생성합니다.
- **파생 노드(플래그).** 새 플래그 필드를 작성하는 파생 노드를 생성합니다. 플래그 필드에는 T (교차표에서 선택된 셀과 일치하는 레코드의 경우) 및 F (나머지 레코드의 경우)가 포함되어 있습니다.
- **파생 노드(세트).** 새 명목 필드를 작성하기 위해 파생 노드를 생성합니다. 명목 필드에는 교차표에서 선택된 셀의 연속 세트 각각에 대해 하나의 범주가 포함되어 있습니다.

(6) 분석 노드

분석 노드를 통해 정확한 예측을 생성하기 위한 모델의 능력을 평가할 수 있습니다. 분석 노드는 하나 이상의 모델 너깃에 대한 예측 값과 실제 값(목표 필드)의 다양한 비교를 수행합니다. 분석 노드를 사용하여 예측 모형을 다른 예측 모형과 비교할 수도 있습니다.

분석 노드를 실행하는 경우 분석 결과의 요약이 실행된 스트림의 각 모델 너깃에 대한 요약 탭의 분석 섹션에 자동으로 추가됩니다. 자세한 분석 결과는 관리자 창의 출력 탭에 표시되거나 파일에 직접 기록될 수 있습니다.

 **참고:** 분석 노드가 예측 값을 실제 값과 비교하므로 감독 모델(목표 필드가 필요함)에서만 유용합니다. 클러스터링 알고리즘과 같은 무감독 모델의 경우 비교의 기준으로 사용할 수 있는 실제 결과가 없습니다.

① 분석 노드 분석 탭

분석 탭을 사용하면 분석의 세부사항을 지정할 수 있습니다.

일치 교차표(기호 또는 범주형 대상의 경우). 생성(예측)된 각 필드와 범주형 대상의 목표 필드(플래그, 명목 또는 순서) 간 일치 패턴을 표시합니다. 실제 값으로 정의된 행과 예측 값으로 정의된 열이 있으며 각 셀에 해당 패턴이 있는 레코드 수가 있는 테이블이 표시됩니다. 예측에서 계통 오류를 식별하는 데 유용합니다. 두 개 이상의 생성 필드가 동일한 출력 필드와 관련되어 있지만 다른 모델에서 생성한 경우 이러한 필드가 동의하고 거부하는 케이스가 계산되고 총계가 표시됩니다. 동의하는 케이스의 경우 다른 올바른/잘못된 통계 세트가 표시됩니다.

성능 평가. 범주형 출력이 있는 모델의 성능 평가 통계를 표시합니다. 출력 필드의 각 범주에 대해 보고되는 이 통계는 해당 범주에 속하는 레코드를 예측하기 위한 모델의 평균 정보 콘텐츠의 측도(비트 단위)입니다. 분류 문제의 어려움을 고려하므로 드문 범주의 정확한 예측은 공통 범주의 정확한 예측보다 높은 성능 평가 지수를 얻습니다. 모델이 범주 추측에 지나지 않는 경우 해당 범주의 성능 평가 지수는 0이 됩니다.

평가 메트릭(AUC & Gini, 2진 분류자만). 2진 분류자의 경우 이 옵션은 AUC(Area Under Curve) 및 Gini 계수 평가 메트릭을 보고합니다. 각 2진 모델에 대해 이러한 두 개의 평가 메트릭을 함께 계산합니다. 메트릭의 값은 분석 출력 브라우저에 테이블로 보고됩니다.

AUC 평가 메트릭은 ROC(Receiver Operator Characteristic) 곡선 아래의 면적으로 계산되며 분류자의 예상 성능에 대한 스칼라 표시입니다. AUC는 항상 0과 1 사이이며 높은 수는 좋은 분류자를 나타냅니다. 좌표 (0,0)과 (1,1) 사이의 대각선 ROC 곡선은 무작위 분류자를 나타내며 AUC가 0.5입니다. 따라서 실제 분류자에는 0.5 미만의 AUC가 없습니다.

Gini 계수 평가 메트릭은 AUC에 대한 대체 평가 메트릭으로 사용되는 경우가 있으며 두 개의 측도는 밀접하게 관련되어 있습니다. Gini 계수는 ROC 곡선과 대각선 간 면적의 2배로 계산되거나 $Gini = 2AUC - 1$ 로 계산됩니다. Gini 계수는 항상 0과 1 사이이며 높은 수는 좋은 분류자를 나타냅니다. Gini 계수는 가능성은 없지만 ROC 곡선이 대각선 아래에 있는 경우에 음수입니다.

신뢰도 수치(사용 가능한 경우). 신뢰도 필드를 생성하는 모델의 경우 이 옵션은 신뢰도 값의 통계와 예측에 대한 관계를 보고합니다. 이 옵션에 대해 두 개의 설정이 있습니다.

- **해당 임계값.** 정확도가 지정된 백분율이 되는 신뢰수준 하한을 보고합니다.
- **정확도 향상.** 정확도가 지정된 요인만큼 향상되는 신뢰수준 하한을 보고합니다. 예를 들어, 전체 정확도가 90%이며 이 옵션이 2.0으로 설정된 경우 보고된 값은 95% 정확도에 필요한 신뢰도입니다.

예측/예측자 필드를 찾을 때 사용. 예측 필드가 원래의 목표 필드와 일치하는 정도를 판별합니다.

- **모델 출력 필드 메타데이터.** 예측 필드를 모델 필드 정보에 기반하여 대상과 일치시키므로 예측 필드의 이름이 변경된 경우에도 일치가 허용됩니다. 유형 노드를 사용하여 값 대화 상자에서 예측 필드의 모델 필드 정보에도 액세스할 수 있습니다. 자세한 정보는 값 대화 상자 사용의 내용을 참조하십시오.
- **필드 이름 형식.** 이름 지정 규칙에 기반하여 필드를 일치시킵니다. 예를 들어, *response*라는 대상의 C5.0 모델 너깃에서 생성한 예측 값은 *\$C-response*라는 필드에 있어야 합니다.

파티션별 구분. 파티션 필드를 사용하여 레코드를 학습, 테스트, 검증 샘플로 분할하는 경우 이 옵션을 선택하여 각 파티션에 대해 개별적으로 결과를 표시하십시오. 자세한 정보는 파티션 노드의 내용을 참조하십시오.

참고: 파티션별로 구분하는 경우 파티션 필드에 널값이 있는 레코드가 분석에서 제외됩니다. 파티션 노드는 널값을 생성하지 않으므로 파티션 노드가 사용되는 경우 이는 문제가 되지 않습니다.

사용자 정의 분석. 모델을 평가하는 데 사용할 사용자의 분석 계산을 지정할 수 있습니다. CLEM 표현식을 사용하여 각 레코드에 대해 계산해야 하는 값과 레코드 수준 스코어를 전체 스코어로 결합하는 방법을 지정하십시오. 함수 **@TARGET** 및 **@PREDICTED**를 사용하여 각각 대상 (실제 출력) 값과 예측 값을 참조하십시오.

- **If.** 일부 조건에 따라 다른 계산을 사용해야 하는 경우 조건식을 지정하십시오.
- **Then.** If 조건이 true인 경우 계산을 지정하십시오.
- **Else.** If 조건이 false인 경우 계산을 지정하십시오.
- **사용.** 개별 스코어에서 전체 스코어를 계산하기 위한 통계를 선택하십시오.

분석을 필드별로 구분. 분석을 구분하는 데 사용 가능한 범주형 필드를 표시합니다. 전체 분석 외에 각 구분 필드의 각 범주에 대한 개별 분석이 보고됩니다.

② 분석 출력 브라우저

분석 출력 브라우저는 분석 노드를 실행한 결과를 표시합니다. 일반 저장, 내보내기, 인쇄 옵션은 파일 메뉴에서 사용할 수 있습니다. 자세한 정보는 출력 보기의 내용을 참조하십시오.

분석 출력을 처음 찾아볼 때 결과가 펼쳐집니다. 결과를 본 후에 숨기려면 항목의 왼쪽에 있는 펼치기 제어를 사용하여 숨길 특정 결과를 접거나 **모두 접기** 단추를 클릭하여 모든 결과를 접으십시오. 결과를 접은 후에 다시 보려면 항목 왼쪽에 있는 펼치기 제어를 사용하여 결과를 표시하거나 **모두 펼치기** 단추를 클릭하여 모든 결과를 표시하십시오.

출력 필드의 결과. 분석 출력에는 생성된 모델에 의해 작성되는 해당 예측 필드가 있는 각 출력 필드의 섹션이 있습니다.

비교. 출력 필드 섹션에는 해당 출력 필드와 연관된 각 예측 필드의 서브섹션이 있습니다. 범주형 출력 필드의 경우 이 섹션의 최상위 수준에는 올바른 예측과 잘못된 예측의 수 및 백분율과 스트림에 있는 총 레코드 수를 표시하는 테이블이 있습니다. 숫자 출력 필드의 경우 이 섹션은 다음 정보를 표시합니다.

- **최소 오차.** 최소 오차(관측 값과 예측 값의 차이)를 표시합니다.
- **최대 오차.** 최대 오차를 표시합니다.
- **평균 오차.** 모든 레코드에서 오차의 평균을 표시합니다. 모델에 계통 **편향**이 있는지 여부(과소 평가보다 과대평가 경향이 강하거나 반대의 경우)를 표시합니다.
- **평균 절대 오차.** 모든 레코드에서 오차의 절대값 평균을 표시합니다. 방향에 관계없이 오차의 평균 크기를 표시합니다.
- **표준 편차.** 오차의 표준 편차를 표시합니다.
- **선형 상관.** 예측 값과 실제 값의 선형 상관을 표시합니다. 이 통계는 -1.0과 1.0 사이입니다. +1.0에 가까운 값은 강하게 긍정적인 연관을 표시하며 높은 예측 값이 높은 실제 값과 연관되어 있고 낮은 예측 값이 낮은 실제 값과 연관되어 있습니다. -1.0에 가까운 값은 강하게 부정적인 연관을 표시하며 높은 예측 값이 낮은 실제 값과 연관되어 있고 반대의 경우도 같습니다. 0.0에 가까운 값은 약한 연관을 표시하며 예측 값이 실제 값에 관계없이 크거나 작습니다. **참고:** 여기에서 공백 항목은 실제 또는 예측 값이 상수이므로 이 경우 선형 상관을 계산할 수 없음을 표시합니다.
- **발생.** 분석에서 사용되는 레코드 수를 표시합니다.

일치 교차표. 범주형 출력 필드의 경우 분석 옵션에서 일치 교차표를 요청하면 교차표가 포함된 서브섹션이 여기에 표시됩니다. 행은 실제 관측 값을 표시하며 열은 예측 값을 표시합니다. 테이블의 셀은 예측 값과 실제 값의 각 조합에 대한 레코드 수를 표시합니다.

성능 평가. 범주형 출력 필드의 경우 분석 옵션에서 성능 평가 통계를 요청하면 성능 평가 결과가 여기에 표시됩니다. 각 출력 범주가 해당 성능 평가 통계와 함께 나열됩니다.

신뢰도 보고서. 범주형 출력 필드의 경우 분석 옵션에서 신뢰도를 요청하면 값이 여기에 표시됩니다. 모델 신뢰도에 대해 다음 통계가 보고됩니다.

- **범위.** 스트림 데이터에 있는 레코드에 대한 신뢰도의 범위(최소값 및 최대값)를 표시합니다.
- **올바른 평균.** 올바르게 분류된 레코드의 평균 신뢰도를 표시합니다.
- **잘못된 평균.** 잘못 분류된 레코드의 평균 신뢰도를 표시합니다.
- **항상 올바른 하한.** 예측이 항상 올바른 신뢰도 임계값 하한을 표시하고 이 기준을 충족시키는 케이스의 백분율을 표시합니다.
- **항상 잘못된 상한.** 예측이 항상 잘못된 신뢰도 임계값 상한을 표시하고 이 기준을 충족시키는 경우의 백분율을 표시합니다.
- **X% 정확도 하한.** 정확도가 X%인 신뢰수준을 표시합니다. X는 분석 옵션에서 **해당 임계값**에 지정된 대략적인 값입니다. 일부 모델 및 데이터 세트의 경우 옵션에 지정된 정확한 임계값을 제공하는 신뢰도를 선택할 수 없습니다(일반적으로 임계값에 가까운 동일한 신뢰도가 있는 비

스한 케이스의 군집으로 인해). 보고된 임계값은 단일 신뢰도 임계값과 함께 얻을 수 있는 지정된 정확도 기준에 가장 가까운 값입니다.

- **올바른 X 중첩 하한.** 정확도가 전체 데이터 세트의 경우보다 X배 양호한 신뢰도를 표시합니다. X는 분석 옵션에서 **정확도 향상**에 지정된 값입니다.

상호 동의. 동일한 출력 필드를 예측하는 두 개 이상의 생성된 모델이 스트림에 포함된 경우 모델에서 생성한 예측 간 **동의**에서 통계를 확인할 수도 있습니다. 여기에는 예측이 동의하는 레코드의 수와 백분율(범주형 출력 필드의 경우) 또는 오차 요약 통계(연속형 출력 필드의 경우)가 포함됩니다. 범주형 필드의 경우 여기에는 모델이 동의하는(동일한 예측 값을 생성함) 레코드의 서브세트에 대한 실제 값과 비교한 예측의 분석이 포함됩니다.

평가 메트릭. 2진 분류자의 경우 분석 옵션에서 평가 메트릭을 요청하면 AUC 및 Gini 계수 평가 메트릭의 값이 이 섹션의 테이블에 표시됩니다. 테이블에는 각 2진 분류자 모델에 대한 하나의 행이 있습니다. 각 모델이 아닌 각 출력 필드에 대한 평가 메트릭 테이블이 표시됩니다.

(7) 데이터 검토 노드

데이터 검토 노드는 전체 크기 그래프 및 다양한 데이터 준비 노드를 생성하기 위해 정렬하고 사용할 수 있는 읽기 쉬운 교차표에 제공되는 IBM® SPSS® Modeler로 가져오는 데이터에 대한 포괄적인 정보를 간략하게 제공합니다.

- 감사 탭은 데이터에 대한 사전 이해를 얻는 데 유용할 수 있는 요약 통계, 히스토그램 및 분포 그래프를 제공하는 보고서를 표시합니다. 보고서는 필드 이름 앞에 저장 공간 아이콘도 표시합니다.
- 감사 보고서의 품질 탭은 이상치, 극단값 및 결측값에 대한 정보를 표시하고 이 값을 처리하는 데 필요한 도구를 제공합니다.

데이터 검토 노드 사용

데이터 검토 노드는 소스 노드에 직접 연결되거나 인스턴스화된 유형 노드로부터 다운스트림으로 연결될 수 있습니다. 결과를 기반으로 다수의 데이터 준비 노드도 생성할 수 있습니다. 예를 들어, 모델링 시 유용하도록 결측값이 지나치게 많이 포함된 필드를 제외하는 필터 노드를 생성하고 나머지 모든 필드에 대한 결측값을 대치하는 SuperNode를 생성할 수 있습니다. 여기서 감사의 실제 효과가 적용되어 데이터의 현재 상태를 평가할 수 있을 뿐만 아니라 평가를 기반으로 조치를 취할 수도 있습니다.

데이터 선별 또는 표본추출. 초기 감사는 큰 데이터를 처리할 때 특히 효과적이므로 표본 노드를 사용하면 레코드의 서브세트만 선택하여 초기 탐색 중 처리 시간을 줄일 수 있습니다. 데이터 검토 노드는 분석의 탐색 단계에서 필드선택 및 이상 항목 발견 등의 노드와 조합하여 사용할 수도 있습니다.

① 데이터 검토 노드 설정 탭

설정 탭에서는 감사에 대한 기본 매개변수를 지정할 수 있습니다.

기본값. 다음과 같이 단순히 노드를 스트림에 연결하고 실행을 클릭하여 기본 설정을 기반으로 모든 필드에 대한 감사 보고서를 생성할 수 있습니다.

- 유형 노드 설정이 없으면 모든 필드가 보고서에 포함됩니다.
- 유형 설정(인스턴스화되었는지 여부는 관계 없음)이 있으면 모든 **입력**, **목표** 및 **둘 다** 필드가 표시에 포함됩니다. 하나의 **목표** 필드가 있는 경우에는 해당 필드를 오버레이 필드로 사용하십시오. 둘 이상의 **목표** 필드가 지정된 경우에는 기본 오버레이가 지정되지 않습니다.

사용자 정의 필드 사용. 수동으로 필드를 선택하려면 이 옵션을 선택하십시오. 오른쪽의 필드 선택기 단추를 사용하여 개별적으로 또는 유형별로 필드를 선택하십시오.

오버레이 필드. 오버레이 필드는 감사 보고서에 표시된 썸네일 그래프를 그리는 데 사용됩니다. 연속형(숫자 범위) 필드의 경우 이변량 통계(공분산 및 상관)도 계산됩니다. 유형 노드 설정을 기반으로 단일 **목표** 필드가 있는 경우 해당 필드는 위에 설명된 대로 기본 오버레이 필드로 사용됩니다. 또는 오버레이를 지정하기 위해 **사용자 정의 필드 사용**을 선택할 수 있습니다.

표시. 출력에서 그래프를 사용할 수 있는지 여부를 지정하고 기본적으로 표시되는 통계를 선택할 수 있게 합니다.

- **그래프.** 각각의 선택된 필드에 대한 그래프를 표시합니다(데이터에 적합한 대로 분산(막대형) 그래프, 히스토그램 또는 산점도). 그래프는 초기 보고서에서 썸네일로 표시되지만 전체 크기 그래프 및 그래프 노드도 생성될 수 있습니다. 자세한 정보는 데이터 검토 출력 브라우저의 내용을 참조하십시오.
- **기본/고급 통계.** 기본적으로 출력에 표시되는 통계의 수준을 지정합니다. 이 설정이 초기 표시를 결정하는 동안 이 설정과 관계없이 출력에서 모든 통계를 사용할 수 있습니다. 자세한 정보는 통계 표시의 내용을 참조하십시오.

중앙값 및 모드. 보고서에서 모든 필드에 대한 중앙값 및 모드를 계산합니다. 큰 데이터 세트를 사용하는 경우 이 통계는 다른 통계보다 계산하는 데 시간이 오래 걸리므로 처리 시간이 늘어날 수 있습니다. 중앙값만 계산하는 경우 보고된 값은 일부 경우 전체 데이터 세트 대신 2000개의 레코드를 가진 표본을 기반으로 할 수 있습니다. 이 표본추출은 이를 수행하지 않으면 메모리 제한이 초과되는 경우에 필드별로 수행됩니다. 표본추출이 적용되는 경우 이에 따라 출력에서 결과에 레이블이 지정됩니다(단순히 **중앙값**보다는 **표본 중앙값**이 지정됨). 중앙값 이외의 모든 통계는 항상 전체 데이터 세트를 사용하여 계산됩니다.

비어 있거나 유형 없는 필드. 인스턴스화된 데이터와 함께 사용되는 경우 유형 없는 필드는 감사 보고서에 포함되지 않습니다. 유형 없는 필드(비어 있는 필드 포함)를 포함하려면 업스트림

유형 노드에서 **모든 값 지우기**를 선택하십시오. 그러면 데이터가 인스턴스화되지 않아 모든 필드가 보고서에 포함됩니다. 예를 들어, 모든 필드의 전체 목록을 얻으려고 하거나 비어 있는 필드를 제외할 필터 노드를 생성하려는 경우 이것이 유용할 수 있습니다. 자세한 정보는 **결측 데이터로 필드 필터링**의 내용을 참조하십시오.

② 데이터 검토 품질 탭

데이터 검토 노드의 품질 탭은 결측값, 이상치 및 극단값을 처리하기 위한 옵션을 제공합니다.

결측값

- **유효한 값을 가진 레코드 수.** 각각의 평가된 필드에 대해 유효한 값을 가진 레코드의 수를 표시하려면 이 옵션을 선택하십시오. 널(정의되지 않음) 값, 공백값, 공백 및 빈 문자열은 항상 유효하지 않은 값으로 처리됩니다.
- **유효하지 않은 값을 가진 레코드 수 분석.** 각 필드에 대해 각 유형의 유효하지 않은 값을 가진 레코드의 수를 표시하려면 이 옵션을 선택하십시오.

이상치 및 극단값

이상치 및 극단값에 대한 발견 방법. 두 가지 방법이 지원됩니다.

평균으로부터의 표준 편차. 평균으로부터의 표준 편차 수를 기반으로 이상치 및 극단값을 발견합니다. 예를 들어, 평균이 100이고 표준 편차가 10인 필드가 있는 경우 3.0을 지정하여 70 미만 또는 130 이상의 값은 이상치로 처리됨을 나타낼 수 있습니다.

사분위수 범위. 두 개의 중심 사분위수가 속하는 범위(25번째 백분위수와 75번째 백분위수 사이)인 사분위수 범위를 기반으로 이상치 및 극단값을 발견합니다. 예를 들어, 기본 설정인 1.5를 기반으로 하면 이상치의 하한 임계값은 $Q1 - 1.5 * IQR$ 이고 상한 임계값은 $Q3 + 1.5 * IQR$ 입니다. 이 옵션을 사용하면 큰 데이터 세트에서 성능이 저하될 수 있습니다.

③ 데이터 검토 출력 브라우저

데이터 검토 브라우저는 데이터에 대한 개요를 얻기 위한 강력한 도구입니다. 감사 탭에는 모든 필드에 대한 썸네일 그래프, 저장 공간 아이콘 및 통계가 표시되지만 품질 탭에는 이상치, 극단값 및 결측값에 대한 정보가 표시됩니다. 초기 그래프 및 요약 통계를 기반으로 숫자 필드의 코딩을 변경하거나 새 필드를 파생시키거나 명목 필드의 값을 재분류하도록 결정할 수 있습니다. 또는 더 정교한 시각화를 사용하여 추가로 탐색할 수 있습니다. 데이터를 변환하거나 시각화하는

데 사용할 수 있는 임의의 수의 노드를 작성하기 위해 생성 메뉴를 사용하여 감사 보고서 브라우저에서 이를 수행할 수 있습니다.

- 열 헤더를 클릭하여 열을 정렬하거나 끌어서 놓기를 사용하여 열을 다시 정렬하십시오. 대부분의 표준 출력 조작도 지원됩니다. 자세한 정보는 출력 보기 주제를 참조하십시오.
- 측정 또는 고유 열에서 필드를 두 번 클릭하여 필드에 대한 값 및 범위를 보십시오.
- 도구 모음 또는 편집 메뉴를 사용하여 값 레이블을 표시하거나 숨기거나 표시할 통계를 선택할 수 있습니다. 자세한 정보는 통계 표시의 내용을 참조하십시오.
- 필드 이름 왼쪽의 저장 공간 아이콘을 확인하십시오. 저장 공간은 데이터를 필드에 저장하는 방식을 설명합니다. 예를 들어, 값이 1 및 0인 필드는 경수 데이터를 저장합니다. 이는 데이터 사용에 대해 설명하고 저장 공간에 영향을 미치지 않는 측정 수준과 구별됩니다. 자세한 정보는 필드 저장 공간 및 형식화 설정의 내용을 참조하십시오.

가. 그래프 보기 및 생성

오버레이가 선택되지 않은 경우 감사 탭에는 막대형 차트(명목 또는 플래그 필드용) 또는 히스토그램(연속형 필드)이 표시됩니다.

명목 또는 플래그 필드 오버레이의 경우 그래프는 오버레이의 값을 기준으로 색상이 지정됩니다.

연속형 필드 오버레이의 경우에는 1차원 막대 또는 히스토그램이 아니라 2차원 산점도가 생성됩니다. 이 경우 x축은 오버레이 필드에 맵핑되어 테이블을 아래로 읽을 때 모든 x축에서 동일한 척도를 볼 수 있습니다.

- 플래그 또는 명목 필드의 경우 막대 위에 마우스 커서를 두면 도구 팁에 기본 값 또는 레이블이 표시됩니다.
- 플래그 또는 명목 필드의 경우 도구 모음을 사용하여 썸네일 그래프의 방향을 가로에서 세로로 토글하십시오.
- 썸네일로부터 전체 크기 그래프를 생성하려면 썸네일을 두 번 클릭하거나 썸네일을 선택한 후 생성 메뉴에서 **그래프 출력**을 선택하십시오. **참고:** 썸네일 그래프가 표본 추출된 데이터를 기반으로 한 경우 생성되는 그래프는 원래 데이터 스트림이 계속 열려 있으면 모든 케이스를 포함합니다.
출력을 작성한 데이터 검토 노드가 스트림에 연결되어 있는 경우에만 그래프를 생성할 수 있습니다.

- 일치하는 그래프 노드를 생성하려면 감사 탭에서 하나 이상의 필드를 선택한 후 생성 메뉴에서 **그래프 노드**를 선택하십시오. 결과 노드가 스트림 캔버스에 추가되며 스트림이 실행될 때마다 그래프를 다시 작성하는 데 사용될 수 있습니다.
- 오버레이 세트에 100개를 초과하는 값이 포함되어 있으면 경고가 발생하고 해당 오버레이는 포함되지 않습니다.

나. 통계 표시

통계 표시 대화 상자에서는 감사 탭에 표시되는 통계를 선택할 수 있습니다. 초기 설정은 데이터 검토 노드에서 지정됩니다. 자세한 정보는 데이터 검토 노드 설정 탭의 내용을 참조하십시오.

최소값(Minimum). 숫자변수의 가장 작은 값입니다.

최대값(Maximum). 숫자변수의 가장 큰 값입니다.

합계(Sum). 비결측값을 갖는 전체 케이스 값의 총계입니다.

범위(Range). 숫자변수의 가장 큰 값과 가장 작은 값의 차이로 최대값에서 최소값을 뺀 값을 의미합니다.

평균(Mean). 중심 경향에 대한 척도입니다. 합계를 케이스 수로 나눈 산술 평균 값입니다.

평균의 표준 오차(Standard Error of Mean). 동일 분포로부터 선택한 표본 간에 발생할 수 있는 평균값의 차이에 대한 척도입니다. 이 값을 사용하여 관측 평균과 가설 값을 간략하게 비교할 수 있습니다. 즉, 표준 오차에 대한 차이 비율이 ± 2 보다 작거나 ± 2 보다 큰 경우 두 값이 다르다고 판단할 수 있습니다.

표준 편차(standard deviation). 평균 주위의 산포 척도이며 분산의 제곱근과 같습니다. 표준 편차는 원래 변수와 같은 단위로 측정됩니다.

분산(Variance). 평균에 대한 산포 척도로, 평균으로부터의 제곱합 편차를 케이스 수에서 1을 뺀 값으로 나눈 값과 같습니다. 분산은 변수 자체의 제곱 단위로 측정됩니다.

왜도(Skewness). 분포의 비대칭성에 대한 척도입니다. 정규 분포는 대칭이므로 왜도 값이 0입니다. 양의 왜도가 많은 분포는 오른쪽이 길다. 유의한 음의 왜도를 가지는 분포에는 왼쪽으로 긴 꼬리가 나타납니다. 왜도값이 표준 오차의 두 배를 넘는 것은 대칭에서 벗어난 정도를 나타냅니다.

왜도의 표준 오차(Standard Error of Skewness). 표준 오차에 대한 왜도의 비율을 정규성 검정에 사용할 수 있습니다. 즉, 비율이 -2 보다 작거나 $+2$ 보다 큰 경우 정규성을 거부할 수 있습니다. 왜도가 큰 양의 값인 경우 오른쪽이 길어지고 큰 음의 값인 경우 왼쪽이 길어집니다.

첨도(Kurtosis). 이상치가 있는 정도에 대한 척도입니다. 정규 분포의 경우 첨도 통계 값은 0입니다. 양(+)의 첨도는 데이터가 정규 분포보다 더 극단적인 이상치를 나타냄을 표시합니다. 음의 첨도는 데이터가 정규 분포보다 극단적인 이상치를 나타냄을 표시합니다.

첨도의 표준 오차(Standard Error of Kurtosis). 표준 오차에 대한 첨도의 비율을 정규성 검증에 사용할 수 있습니다. 즉, 비율이 -2보다 작거나 +2보다 큰 경우 정규성을 거부할 수 있습니다. 첨도가 높은 양의 값인 경우 분포의 양끝이 정규 분포의 양끝보다 길어지고 음의 값인 경우 양끝이 짧아집니다(상자 형태 균일 분포와 유사).

고유(Unique). 모든 효과를 동시에 평가하고 유형에 관계없이 다른 모든 효과에 대해 각 효과를 조정합니다.

유효함(Valid). 시스템 결측값 또는 사용자 결측값이 지정되어 있지 않은 케이스가 유효 케이스입니다. 널(정의되지 않은) 값, 공백값, 공백 및 빈 문자열은 항상 유효하지 않은 값으로 처리됩니다.

중앙값(Median). 전체 케이스의 절반이 위 아래에 해당되는 값으로 제50 백분위수입니다. 케이스 수가 짝수인 경우 중앙값은 케이스를 오름차순이나 내림차순으로 정렬했을 때 중간에 있는 두 개의 케이스의 평균입니다. 중앙값은 평균과 달리 중심을 벗어난 값에는 영향을 받지 않는 중심 경향 측도이며, 상한 극단값 또는 하한 극단값에 따라 달라질 수 있습니다.

최빈값(Mode). 가장 자주 발생하는 값입니다. 여러 값에서 최대 발생 빈도를 공유하는 경우 각각을 최빈값이라고 합니다.

중앙값 및 모드는 성능 향상을 위해 기본적으로 표시되지 않지만 데이터 검토 노드의 설정 탭에서 선택할 수 있습니다. 자세한 정보는 데이터 검토 노드 설정 탭의 내용을 참조하십시오.

오버레이에 대한 통계

연속형(숫자 범위) 오버레이 필드가 사용 중인 경우에는 다음과 같은 통계도 사용할 수 있습니다.

공분산(Covariance). 두 변수 간 연관성을 표준화하지 않은 측도로서, N-1로 나눈 교차곱 편차와 같습니다.

다. 데이터 검토 브라우저 품질 탭

데이터 검토 브라우저의 품질 탭에는 데이터 품질 분석의 결과가 표시되며 사용자가 이상치, 극단값 및 결측값에 대한 처리를 지정할 수 있습니다.

ㄱ. 결측값 대체

감사 보고서에는 유효한 값, 널 값 및 공백 값의 수와 함께 각 필드에 대한 완료 레코드의 퍼센트가 나열됩니다. 특정 필드에 대한 결측값 대체를 선택하여 이러한 변환을 적용할 수퍼노드를 생성할 수 있습니다.

1. **결측값 대치** 열에서 대치할 값의 유형을 지정하십시오. 단, 있는 경우에 한합니다. 공백 또는 널 또는 둘 다 대치하도록 선택하거나 대치할 값을 선택하는 사용자 정의 조건 또는 표현식을 지정할 수 있습니다.

IBM® SPSS® Modeler에 의해 인지되는 결측값에는 몇 가지 유형이 있습니다.

- **널 또는 시스템 결측값.** 이들은 데이터베이스나 소스 파일에 공백으로 남겨졌고 소스 또는 유형 노드에서 "결측"으로 정의되지 않은 문자열이 아닌 값입니다. 시스템 결측값은 \$null \$로 표시됩니다. 빈 문자열은 특정 데이터베이스에 의해 널로 처리되더라도 IBM SPSS Modeler에서는 널로 간주되지 않음을 유의하십시오.
- **빈 문자열 및 공백.** 빈 문자열 값과 공백(눈에 보이는 문자가 없는 문자열)은 널값과는 별개로 처리됩니다. 빈 문자열은 대부분의 경우에서 공백과 동일하게 처리됩니다. 예를 들어, 소스나 유형 노드에서 공백을 공란으로 처리하는 옵션을 선택한 경우 이 설정은 빈 문자열에도 적용됩니다.
- **공백 또는 사용자 정의 결측값.** 이들은 소스 노드 또는 유형 노드에서 결측으로 명백하게 정의되어 있는 unknown, 99 또는 -1 등과 같은 값입니다. 또는 널과 공백을 공란으로 처리하기로 선택할 수도 있는데 그러면 이들은 특수 처리용으로 플래그가 지정되고 대부분의 계산에서 제외됩니다. 예를 들어, @BLANK 함수를 사용하여 이들 값 및 다른 유형의 결측값을 공란으로 처리할 수 있습니다.

2. **방법** 열에서 사용할 방법을 지정하십시오.

결측값을 대치하기 위해 다음 방법을 사용할 수 있습니다.

고정됨 고정된 값을 대체합니다(필드 평균, 범위의 중심점 또는 사용자가 지정하는 상수).

임의 보통 또는 균일 분포를 기반으로 변량 값을 대체합니다.

표현식. 사용자 정의 표현식을 지정할 수 있습니다. 예를 들어, 전역값 설정 노드에 의해 생성된 글로벌 변수로 값을 대체할 수 있습니다.

알고리즘. C&RT 알고리즘을 기반으로 모델에 의해 예측된 값을 대체합니다. 이 방법을 사용하여 대체된 각 필드의 경우, 공백과 널을 모델에 의해 예측된 값으로 대체하는 채움 노드와 함께 별도의 C&RT 모델이 있습니다. 그러면 필터 노드가 모델에 의해 생성된 예측 필드를 제거하는 데 사용됩니다.

3. 결측값 슈퍼노드를 생성하려면 메뉴에서 다음을 선택하십시오.

생성 > 결측값 슈퍼노드

결측값 슈퍼노드 대화 상자가 표시됩니다.

4. **모든 필드** 또는 **선택된 필드**만을 선택하고 필요에 따라 표본 크기를 지정하십시오. (지정된 표본은 퍼센트이며 기본적으로 모든 레코드의 10%가 표본화됩니다.)

5. 생성된 슈퍼노드를 스트림 캔버스에 추가하려면 확인을 클릭하십시오.
6. 슈퍼노드를 스트림에 첨부하여 변환을 적용하십시오.

슈퍼노드 내에서 모델 너짓, 채움 및 필터 노드의 조합이 적절히 사용됩니다. 슈퍼노드를 편집하고 **확대**를 클릭하여 슈퍼노드 내의 특정 노드를 추가, 편집 또는 제거하여 작동을 세분화함으로써 작동 방법을 이해할 수 있습니다.

ㄴ. 이상값 및 극단값 처리

감사 보고서에는 이상값 수가 나열되고 데이터 검토 노드에서 지정된 발견 옵션을 기반으로 하여 각 필드에 대한 극단값이 나열됩니다. 자세한 정보는 데이터 검토 품질 탭의 내용을 참조하십시오. 특정 필드에 대해 이러한 값 강제 변환, 삭제 또는 무효화를 필요에 따라 선택한 다음 변환을 적용할 슈퍼노드를 생성할 수 있습니다.

1. 동작 열에서 특정 필드에 대한 이상값 및 극단값 처리를 지정하십시오.
이상값 및 극단값에 대해 사용 가능한 동작은 다음과 같습니다.
 - **강제 적용.** 이상값 및 극단값을 극단값으로 간주되지 않는 가장 가까운 값으로 대체합니다. 예를 들어, 이상값이 세 표준편차 위 또는 아래의 값으로 정의된 경우, 모든 이상값이 해당 범위 내의 최대값 또는 최저값으로 대체됩니다.
 - **삭제.** 지정된 필드에 대한 이상값 및 극단값을 삭제합니다.
 - **무효화.** 널이거나 시스템 결측값인 이상값 및 극단값을 바꿉니다.
 - **이상값 강제 변환/극단값 삭제.** 극단값만 삭제합니다.
 - **이상값 강제 변환/극단값 무효화.** 극단값만 무효화합니다.
2. 슈퍼노드를 생성하려면 메뉴에서 다음을 선택하십시오.
생성 > 이상치 및 극단값 슈퍼노드

이상값 슈퍼노드 대화 상자가 표시됩니다.
3. **모든 필드** 또는 **선택된 필드만**을 선택하고 **확인**을 클릭하여 생성된 슈퍼노드를 스트림 캔버스에 추가하십시오.
4. 슈퍼노드를 스트림에 첨부하여 변환을 적용하십시오.

선택적으로 슈퍼노드를 편집하고 확대하여 찾아보거나 변경할 수 있습니다. 슈퍼노드 내에서 필요에 따라 일련의 선택 및/또는 채움 노드를 사용하여 값이 삭제, 강제 변환 또는 무효화됩니다.

㉔. 결측 데이터로 필드 필터링

데이터 검토 브라우저를 통해, 품질에서 필터 생성 대화 상자를 사용하여 품질 분석의 결과를 기반으로 하여 새 필터 노드를 생성할 수 있습니다.

모드. 지정된 필드에 대해 원하는 작업, 즉, **포함** 또는 **제외**를 선택하십시오.

- **선택된 필드.** 필터 노드가 품질 탭에서 선택된 필드를 포함/제외합니다. 예를 들어, **% 완료** 열에서 테이블을 정렬할 수 있으며 Shift-클릭을 선택하여 가장 적게 완료된 필드를 선택한 다음 해당 필드를 제외하는 필터 노드를 생성할 수 있습니다.
- **다음보다 큰 품질 퍼센트 필드.** 필터 노드가 레코드 완료 퍼센트가 지정된 임계값보다 큰 필드를 포함/제외합니다. 기본 임계값은 50%입니다.

비어 있거나 유형이 없는 필드 필터링

데이터값이 인스턴스화된 후에 유형이 없거나 비어 있는 필드가 감사 결과 및 IBM® SPSS® Modeler의 대부분의 기타 출력에서 제외됩니다. 이러한 필드는 모델링 목적으로는 무시되나 데이터를 과장하거나 산만하게 만들 수 있습니다. 그런 경우, 데이터 검토 브라우저를 사용하여 이러한 필드를 스트림에서 제거하는 필터 노드를 생성할 수 있습니다.

1. 비어 있거나 유형이 없는 필드를 포함하여 모든 필드가 감사에 포함되도록 하려면 업스트림 소스 또는 유형 노드에서 **모든 값 지우기**를 클릭하거나 모든 필드에 대해 <Pass>로 값을 설정하십시오.
2. 데이터 검토 브라우저에서 **% 완료** 열을 정렬하고 0 개의 유효한 값이 있는 필드 또는 기타 임계값이 적용된 필드를 선택하고 생성 메뉴를 사용하여 스트림에 추가될 수 있는 필터 노드를 생성하십시오.

㉕. 결측 데이터가 있는 레코드 선택

데이터 검토 브라우저를 통해, 품질 분석의 결과를 기반으로 하여 새 선택 노드를 생성할 수 있습니다.

1. 데이터 검토 브라우저에서 품질 탭을 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.

생성 > 결측값 선택 노드

선택 노드 생성 대화 상자가 표시됩니다.

레코드의 선택 조건. 레코드가 **유효** 또는 **유효하지 않음**일 때 유지하는지 여부를 지정합니다.

유효하지 않은 값 검색. 유효하지 않은 값 검사 여부를 지정합니다.

- 모든 필드. 선택 노드가 모든 필드의 유효하지 않은 값을 검사합니다.
- 테이블에서 선택된 필드. 선택 노드가 품질 출력 테이블에서 현재 선택된 필드만 검사합니다.
- 다음보다 큰 품질 퍼센트 필드. 선택 노드가 레코드 완료 퍼센트가 지정된 임계값보다 큰 필드를 검사합니다. 기본 임계값은 50%입니다.

유효하지 않은 값이 다음 위치에서 발견되는 경우 레코드를 유효하지 않은 것으로 간주. 레코드를 유효하지 않은 것으로 식별하는 조건을 지정합니다.

- 위 필드 중 임의의 필드. 지정된 위 필드 중 임의의 필드에 해당 레코드에 대해 유효하지 않은 값이 포함되면 선택 노드가 레코드를 유효하지 않은 것으로 간주합니다.
- 위 필드 중 모든 필드. 지정된 위 필드 중 모든 필드에 해당 레코드에 대해 유효하지 않은 값이 포함되면 선택 노드가 레코드를 유효하지 않은 것으로 간주합니다.

ㅁ. 데이터 준비를 위해 기타 노드 생성

데이터 준비에서 사용되는 다양한 노드를 데이터 검토 브라우저에서 직접 생성할 수 있습니다 (재분류, 구간화 및 파생 노드 포함). 예:

- 감사 보고서에서 *claimvalue* 값과 *farmincome* 값을 모두 선택한 후 생성 메뉴에서 **파생**을 선택하여 이 두 값을 기반으로 새 필드를 파생시킬 수 있습니다. 새 노드는 스트림 캔버스에 추가됩니다.
- 마찬가지로 감사 결과에 따라 *farmincome*을 백분위수 기반 구간으로 코딩을 변경하면 더 집중된 분석이 제공되는지 판별할 수 있습니다. 구간화 노드를 생성하려면 표시에서 필드 행을 선택하고 생성 메뉴에서 **구간화**를 선택하십시오.

노드가 생성되어 스트림 캔버스에 추가된 후에는 해당 노드를 스트림에 연결하고 해당 노드를 열어서 선택된 필드에 대한 옵션을 지정해야 합니다.

(8) 변환 노드

회귀분석, 로지스틱 회귀분석 및 판별 분석과 같은 일반 스코어링 기술을 사용하기 전에 수행해야 하는 중요한 단계 중 하나는 입력 필드를 정규화하는 것입니다. 이러한 기술은 다수의 원시 데이터 파일에 적용되지 않을 수 있는 데이터의 정규 분포에 대한 가정을 수행합니다. 실세계 데이터를 처리하는 한 가지 방법은 원시 데이터 요소를 정규성이 더 높은 정규 분포 쪽으로 이동시키는 변환을 적용하는 것입니다. 또한 정규화된 필드는 서로 쉽게 비교할 수 있습니다. 예를 들어, 원시 데이터 파일에서 수입과 연령은 척도가 완전히 다르지만 정규화하는 경우 각각의 상대적 영향력을 쉽게 해석할 수 있습니다.

변환 노드에서는 사용할 가장 좋은 변환을 시각적으로 신속하게 평가할 수 있는 출력 뷰어를 제공합니다. 변수가 정상적으로 분포되는지 한 눈에 알 수 있고 필요한 경우 원하는 변환을 선택하여 적용할 수 있습니다. 여러 필드를 선택하여 필드당 하나의 변환을 수행할 수 있습니다.

필드에 대해 선호하는 변환을 선택한 후에는 변환을 수행하는 파생 또는 채움 노드를 생성하여 이들을 스트림에 연결할 수 있습니다. 파생 노드는 새 필드를 작성하고 채움 노드는 기존 필드를 변환합니다. 자세한 정보는 그래프 생성의 내용을 참조하십시오.

변환 노드 필드 탭

필드 탭에서는 가능한 변환을 보고 적용하는 데 사용할 데이터 필드를 지정할 수 있습니다. 숫자 필드만 변환할 수 있습니다. 필드 선택기 단추를 클릭하고 표시된 목록에서 하나 이상의 숫자 필드를 선택하십시오.

① 변환 노드 옵션 탭

옵션 탭에서는 포함시키려는 변환의 유형을 지정할 수 있습니다. 사용 가능한 모든 변환을 포함시키거나 변환을 개별적으로 선택할 수 있습니다.

후자의 경우, 역변환 및 로그 변환을 위해 데이터를 오프셋하기 위한 숫자를 입력할 수도 있습니다. 이는 데이터에 있는 다수의 0으로 인해 평균 및 표준 편차 결과가 편향되는 경우에 유용합니다.

예를 들어, 몇 개의 0 값이 있는 *BALANCE* 필드가 있고 이 필드에서 역변환을 사용하려 합니다. 원하지 않는 편향을 피하기 위해 **역(1/x)**을 선택하고 **데이터 오프셋 사용** 필드에 1을 입력합니다. (이 오프셋은 IBM® SPSS® Modeler에서 @OFFSET 시퀀스 함수에 의해 수행된 오프셋과 관련이 없습니다.)

모든 공식. 사용 가능한 모든 변환이 계산되고 출력에 표시되어야 함을 나타냅니다.

공식 선택. 계산하고 출력에 표시할 다양한 변환을 선택할 수 있습니다.

- **역(1/x).** 역변환이 출력에 표시되어야 함을 나타냅니다.
- **로그(로그 n).** 로그_n 변환이 출력에 표시되어야 함을 나타냅니다.
- **로그(로그 10).** 로그₁₀ 변환이 출력에 표시되어야 함을 나타냅니다.
- **지수.** 지수 변환(e^x)이 출력에 표시되어야 함을 나타냅니다.
- **제곱근.** 제곱근 변환이 출력에 표시되어야 함을 나타냅니다.

② 변환 노드 출력 탭

출력 탭을 사용하여 출력의 형식 및 위치를 지정할 수 있습니다. 결과를 화면에 표시하거나 표준 파일 유형 중 하나로도 보낼 수 있습니다. 자세한 정보는 출력 노드 출력 탭 주제를 참조하십시오.

③ 변환 노드 출력 뷰어

출력 뷰어를 사용하여 변환 노드의 실행 결과를 볼 수 있습니다. 뷰어는 변환의 썸네일 보기에 필드당 여러 변환을 표시하는 강력한 도구이므로 뷰어를 통해 필드를 신속하게 비교할 수 있습니다. 해당 파일 메뉴의 옵션을 사용하여 출력을 저장, 내보내기 또는 인쇄할 수 있습니다. 자세한 정보는 출력 보기 주제를 참조하십시오.

변환마다(선택된 변환 제외) 아래에 다음 형식의 범례가 표시됩니다.

Mean (Standard deviation)

가. 변환을 위한 노드 생성

출력 뷰어는 데이터 준비에 유용한 시작점을 제공합니다. 예를 들어, 정규 분포를 가정하는 스코어링 기술(예: 로지스틱 회귀분석 또는 판별 분석)을 사용할 수 있도록 AGE 필드를 정규화하려 할 수 있습니다. 초기 그래프와 요약 통계를 기반으로, 특정 분포(예: 로그)에 따라 AGE 필드를 변환하기로 할 수 있습니다. 선호하는 분포를 선택한 후에는 스코어링에 사용할 표준화된 변환이 있는 파생 노드를 생성할 수 있습니다.

출력 뷰어에서 다음 필드 작업 노드를 생성할 수 있습니다.

- 파생
- 채움

파생 노드는 원하는 변환으로 새 필드를 작성하는 반면, 채움 노드는 기존 필드를 변환합니다. 노드는 수퍼노드 양식으로 캔버스에 배치됩니다.

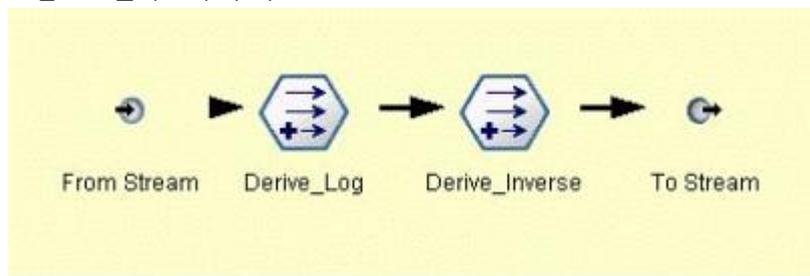
서로 다른 필드에 대해 동일한 변환을 선택하는 경우, 파생 또는 채움 노드는 해당 변환이 적용되는 모든 필드에 대한 해당 변환 유형의 공식을 포함합니다. 예를 들어, 파생 노드를 생성하기 위해 다음 테이블에 표시된 필드 및 변환을 선택했다고 가정하십시오.

표 1. 파생 노드 생성 예

필드	변환
AGE	현재 분포
INCOME	로그
OPEN_BAL	역
BALANCE	역

수퍼노드에는 다음 노드가 포함됩니다.

그림 1. 캔버스의 수퍼노드



이 예에서, Derive_Log 노드에는 *INCOME* 필드에 대한 로그 공식이 있고 Derive_Inverse 노드에는 *OPEN_BAL* 및 *BALANCE* 필드에 대한 역 공식이 있습니다.

노드를 생성하려면 다음을 수행하십시오.

1. 출력 뷰어의 필드마다 원하는 변환을 선택하십시오.
2. 생성 메뉴에서 **파생 노드** 또는 **채움 노드**를 선택하십시오.
그러면 파생 노드 생성 또는 채움 노드 생성 대화 상자가 표시됩니다.

비표준화 변환 또는 **표준화 변환(z-스코어)**을 선택하십시오. 두 번째 옵션은 변환에 *z* 스코어를 적용합니다. *z* 스코어는 값을 표준 편차에서 변수 평균으로부터의 거리 함수로 표현합니다. 예를 들어, *AGE* 필드에 로그 변환을 적용하고 표준 변환을 선택하는 경우, 생성되는 노드에 대한 최종 등식은 다음과 같습니다.

$$(\log(\text{AGE}) - \text{Mean}) / \text{SD}$$

노드가 생성되고 스트림 캔버스에 표시되면 다음을 수행하십시오.

1. 노드를 스트림에 연결하십시오.
2. 수퍼노드의 경우, 노드를 두 번 클릭하여 해당 콘텐츠를 보십시오(선택사항).
3. 파생 또는 채움 노드를 두 번 클릭하여 선택된 필드의 옵션을 수정하십시오(선택사항).

ㄱ. 그래프 생성

출력 뷰어에서 썸네일 히스토그램으로부터 전체 크기 히스토그램 출력을 생성할 수 있습니다.

그래프를 생성하려면 다음을 수행하십시오.

1. 출력 뷰어에서 썸네일 그래프를 두 번 클릭하십시오.

or

출력 뷰어에서 썸네일 그래프를 선택하십시오.

2. 생성 메뉴에서 **그래프 출력**을 선택하십시오.

그러면 정규 분포 곡선이 오버레이된 히스토그램이 표시됩니다. 이를 통해 사용 가능한 변환이 정규 분포와 얼마나 일치하는지 비교할 수 있습니다.

참고: 출력을 작성한 변환 노드가 스트림에 연결되어 있는 경우에만 그래프를 생성할 수 있습니다.

• 기타 조작

출력 뷰어에서 다음 조작도 수행할 수 있습니다.

- 필드 열을 기준으로 출력 눈금을 정렬합니다.
- 출력을 HTML 파일로 내보냅니다. 자세한 정보는 출력 내보내기 주제를 참조하십시오.

(9) 통계량 노드

통계량 노드는 수치 필드에 관한 기본 요약 정보를 제공합니다. 개별 필드에 대한 요약 통계량 및 필드 사이의 상관계수를 얻을 수 있습니다.

① 통계량 노드 설정 탭

탐색. 개별 요약 통계량이 필요한 필드를 선택하십시오. 다중 필드를 선택할 수 있습니다.

통계량. 보고할 통계량을 선택하십시오. 사용 가능한 옵션은 **개수**, **평균**, **합계**, **최소**, **최대**, **범위**, **분산**, **표준편차**, **평균의 표준오차**, **중앙값** 및 **최빈값**입니다.

상관분석. 상관분석할 필드를 선택하십시오. 다중 필드를 선택할 수 있습니다. 상관관계 필드가 선택되면 각 탐색 필드 및 상관관계 필드 사이의 상관관계가 출력에 나열됩니다.

상관관계 설정. 출력에 상관관계의 강도를 표시하기 위한 옵션을 지정할 수 있습니다. 자세한 정보는 상관관계 설정 주제를 참조하십시오.

가. 상관관계 설정

IBM® SPSS® Modeler는 중요한 관계를 강조표시하는 것을 돕기 위해 기술통계 레이블을 사용하여 상관관계 특성을 지정할 수 있습니다. **상관관계**는 두 연속형(수치 범위) 필드 사이의 관계의 강도를 측정합니다. -1.0에서 1.0 사이의 값을 사용합니다. +1.0에 가까운 값은 강력한 양의 연관을 표시하므로 한 필드의 높은 값은 다른 필드의 높은 값과 연관되며 낮은 값은 낮은 값과 연관됩니다. -1.0에 가까운 값은 강력한 음의 연관을 표시하므로 한 필드의 높은 값은 다른 필드의 낮은 값과 연관되며 반대의 경우도 마찬가지입니다. 0.0에 가까운 값은 약한 연관을 표시하므로 두 필드의 값이 다소 독립적입니다.

상관관계 설정 대화 상자를 사용하여 상관관계 레이블의 표시를 제어하고 범주를 정의하는 임계값을 변경하고 각 범위에 사용되는 레이블을 변경할 수 있습니다. 상관관계 값의 특성을 지정하는 방법은 문제점 도메인에 크게 종속되므로 특정 상황에 맞게 범위 및 레이블을 사용자 정의해야 합니다.

출력에 상관관계 강도 레이블 표시. 이 옵션은 기본적으로 선택됩니다. 이 옵션을 선택 취소하면 출력에서 기술통계 레이블이 생략됩니다.

상관관계 강도. 상관관계 강도를 정의하고 레이블을 지정하기 위한 두 개의 옵션이 있습니다.

- **중요도 기준으로 상관관계 강도 정의(1-p).** 평균과의 차분을 단독 가능성으로 설명할 수 있는 1 빼기 유의수준 또는 1 빼기 확률로 정의되는 중요도를 기준으로 하여 상관관계 레이블을 지정합니다. 이 값이 1에 가까울수록 두 필드가 독립적이지 않을 가능성이 커집니다. 즉, 둘 사이에 어떠한 연관이 존재할 가능성이 커집니다. 일반적으로 절대값보다는 중요도를 기준으로 하여 상관관계 레이블을 지정하는 것을 권장합니다. 이 방법이 데이터의 변동을 고려하기 때문입니다. 예를 들어, 계수 0.6이 한 데이터 세트에서는 매우 유의적이거나 다른 데이터 세트에서는 유의적이지 않을 수 있습니다. 기본적으로 0.0에서 0.9 사이의 중요도 값은 *약함*, 0.9에서 0.95 사이는 *중간*, 0.95에서 1.0 사이는 *강함*으로 레이블이 지정됩니다.
- **절대값 기준으로 상관관계 강도 정의.** 위에서 설명한 대로 -1.0에서 1.0 사이의 값을 사용하여 Pearson 상관 계수의 절대값을 기준으로 하여 상관관계 레이블을 지정합니다. 이 측도의 절대값이 1에 가까울수록 상관관계가 강한 것입니다. 기본적으로 (절대값에서) 0.0에서 0.3333 사이의 상관관계는 *약함*, 0.3333에서 0.6666 사이는 *중간*, 0.6666에서 1.0 사이는 *강함*으로 레이블이 지정됩니다. 단, 지정된 값의 유의수준은 한 데이터 세트에서 다른 데이터 세트로 일반화하기 어렵습니다. 이런 이유로 인해 대부분의 경우에 절대값이 아니라 확률을 기준으로 하는 상관관계 정의가 사용됩니다.

② 통계량 출력 브라우저

통계량 노드 출력 브라우저는 통계 분석의 결과를 표시하며 필드 선택, 선택을 기반으로 하여 새 노드 생성 및 결과 저장 및 인쇄 등의 작업을 수행할 수 있도록 해줍니다. 파일 메뉴에서 일반적인 저장, 내보내기 및 인쇄 옵션을 사용할 수 있으며 일반적인 편집 옵션은 편집 메뉴에서 사용할 수 있습니다. 자세한 정보는 출력 보기 주제를 참조하십시오.

처음 통계량 출력을 찾으면 결과가 펼쳐집니다. 결과를 본 후에 숨기려면 항목의 왼쪽에 있는 펼치기 제어를 사용하여 숨길 특정 결과를 접거나 모두 접기 단추를 클릭하여 모든 결과를 접으십시오. 결과를 접은 후에 다시 보려면 항목 왼쪽에 있는 펼치기 제어를 사용하여 결과를 표시하거나 모두 펼치기 단추를 클릭하여 모든 결과를 표시하십시오.

출력에는 요청된 통계량 표를 포함하여 각 탐색 필드에 대한 섹션이 포함됩니다.

- **개수.** 필드에 대한 유효한 값이 있는 레코드 수입니다.
- **평균.** 모든 레코드 전체에 걸친 필드의 평균 값입니다.
- **합계.** 모든 레코드 전체에 걸친 필드의 값의 합계입니다.
- **최소.** 필드의 최소값입니다.
- **최대.** 필드의 최대값입니다.
- **범위.** 최소값 및 최대값의 차이입니다.
- **분산.** 필드의 값에서 변동의 척도입니다. 각 값 및 전체 평균 사이의 차분을 구하고 이를 제곱하여 전체 값을 합한 다음 레코드 수로 나누는 방법으로 계산합니다.
- **표준 편차.** 필드 값에서의 또 다른 변동 척도이며 분산의 제곱근으로 계산됩니다.
- **평균의 표준오차.** 평균을 새 데이터에 적용하는 것으로 가정할 때 필드의 평균 추정값의 불확실성 척도입니다.
- **중앙값.** 필드의 "중간" 값입니다. 즉, 필드의 값을 기준으로 하여 데이터를 상한 절반과 하한 절반으로 나누는 값입니다.
- **모드.** 데이터에서 가장 흔한 단일값입니다.

상관계수. 상관분석 필드를 지정하면 출력에도 탐색 필드와 각 상관분석 필드 사이의 Pearson 상관을 나열하는 섹션 및 상관관계 값에 대한 선택적 기술통계 레이블이 포함됩니다. 자세한 정보는 상관관계 설정 주제를 참조하십시오.

생성 메뉴. 생성 메뉴에는 노드 생성 작업이 포함됩니다.

- **필터.** 다른 필드와 상관관계가 없거나 약한 필드를 필터링하기 위해 필터 노드를 생성합니다. 자세한 정보는 통계량에서 필터 노드 생성 주제를 참조하십시오.

가. 통계량에서 필터 노드 생성

통계량 출력 브라우저에서 생성된 필터 노드는 다른 필드와의 상관관계를 기준으로 하여 필드를 필터링합니다. 절대값 순서로 상관관계를 정렬하고 (통계량 대화 상자의 생성 필터의 기준 설정에 따라) 가장 큰 상관관계를 구하고 이러한 큰 상관관계에 표시되는 모든 필드를 전달하는 필터를 작성하는 방법으로 작업합니다.

모드. 상관관계를 선택하는 방법을 결정합니다. **포함**을 선택하면 지정된 상관관계에 표시되는 필드가 유지됩니다. **제외**를 선택하면 필드가 필터링됩니다.

표시되는 포함/제외 필드. 상관관계 선택의 기준을 정의합니다.

- **최대 상관관계 수.** 지정된 수의 상관관계 및 해당 상관관계에 표시되는 포함/제외 필드 수를 선택합니다.
- **최대 상관관계 퍼센트(%).** 지정된 퍼센트(n%)의 상관관계 및 해당 상관관계에 표시되는 포함/제외 필드 수를 선택합니다.
- **다음보다 큰 상관관계.** 지정된 임계값보다 절대값이 큰 상관관계를 선택합니다.

(10) 평균 노드

평균 노드는 독립 그룹 사이 또는 관련된 필드의 쌍 사이의 평균을 비교하여 상당한 차이가 존재하는지 여부를 검정합니다. 예를 들어, 프로모션 실행 전후의 평균 수입을 비교하거나 프로모션을 받지 않은 고객으로부터의 수입을 프로모션을 받은 고객으로부터의 수입과 비교할 수 있습니다.

데이터에 따라 두 가지 다른 방식으로 평균을 비교할 수 있습니다.

- **필드 내 그룹 사이.** 독립 그룹을 비교하려면 검정 필드 및 그룹화 필드를 선택하십시오. 예를 들어, 프로모션을 보낼 때 "검증용" 고객의 표본을 제외하고 검증용 그룹에 대한 평균 수입을 모든 다른 그룹과 비교할 수 있습니다. 이 경우 고객이 제안을 받았는지를 표시하는 플래그 또는 명목 필드를 사용하여 각 고객에 대한 수입을 표시하는 단일 검정 필드를 지정할 수 있습니다. 표본은 각 레코드가 하나의 그룹 또는 다른 그룹에 지정된다는 점에서 독립적이며 한 그룹의 특정 멤버를 다른 그룹의 특정 멤버에 링크할 수 있는 방법이 없습니다. 여러 그룹에 대한 평균을 비교하기 위해 셋 이상의 값을 가진 명목 필드를 지정할 수도 있습니다. 실행되면 노드는 선택된 필드에서 일원 ANOVA 검정을 계산합니다. 두 개의 필드 그룹만 있는 경우 일원 ANOVA 결과는 본질적으로 독립 표본 t 검정과 동일합니다. 자세한 정보는 독립 그룹에 대한 평균 비교 주제를 참조하십시오.
- **필드 쌍 사이.** 두 관련 필드에 대한 평균을 비교하는 경우 결과가 의미를 가지려면 어떤 방식으로든 그룹이 쌍을 이루어야 합니다. 예를 들어, 프로모션 실행 전후 동일한 고객 그룹으로부터의 평균 수입을 비교하거나 남편-아내 쌍 간 서비스 사용 요금을 비교하여 차이가 있는지

확인할 수 있습니다. 각각의 레코드에는 의미 있게 비교할 수 있는 두 개의 독립되어 있지만 관련된 측도가 포함되어 있습니다. 실행되면 노드는 선택된 각각의 필드 쌍에서 대응 표본 t 검정을 계산합니다. 자세한 정보는 대응 필드 간 평균 비교 주제를 참조하십시오.

① 독립 그룹에 대한 평균 비교

평균 노드에서 **필드 내 그룹 사이**를 선택하여 둘 이상의 독립 그룹에 대한 평균을 비교하십시오.

그룹화 필드. 레코드를 비교할 그룹(예: 제안을 받은 그룹과 제안을 받지 않은 그룹)으로 나누는 둘 이상의 고유 값을 가진 숫자 플래그 또는 명목 필드를 선택하십시오. 검정 필드 수에 관계없이 하나의 그룹화 필드만 선택할 수 있습니다.

검정 필드. 검정할 측도가 포함된 하나 이상의 숫자 필드를 선택하십시오. 선택하는 각각의 필드에 대해 별도의 검정이 수행됩니다. 예를 들어, 사용, 수입 및 이탈에 대한 지정된 프로모션의 영향을 검정할 수 있습니다.

② 대응 필드 간 평균 비교

평균 노드에서 **필드 쌍 사이**를 선택하여 별도의 필드 간 평균을 비교하십시오. 결과가 의미를 가지려면 이 필드가 어떤 방식으로든 관련되어 있어야 합니다(예: 프로모션 전후의 수입). 다중 필드 쌍도 선택할 수 있습니다.

필드 1. 비교할 첫 번째 측도가 포함된 숫자 필드를 선택하십시오. 전후 연구에서 이는 "전" 필드입니다.

필드 2. 비교할 두 번째 필드를 선택하십시오.

추가. 선택한 쌍을 검정 필드 쌍 목록에 추가합니다.

필요에 따라 필드 선택을 반복하여 여러 쌍을 목록에 추가하십시오.

상관관계 설정. 상관관계의 강도에 레이블을 지정하는 옵션을 지정할 수 있게 합니다. 자세한 정보는 상관관계 설정 주제를 참조하십시오.

③ 평균 노드 옵션

옵션 탭에서는 결과에 레이블을 지정하는 데 사용된 임계값 p 값을 중요, 보통 또는 중요하지 않음으로 설정할 수 있습니다. 각 순위화에 대한 레이블도 편집할 수 있습니다. 중요도는 백분율

척도에 따라 측정되며 우연에 의해서만 관측된 결과만큼 또는 그 이상 극단적인 결과를 얻는 확률(예: 두 필드 간 평균 차이)을 1에서 뺀 것으로 광범위하게 정의될 수 있습니다. 예를 들어, p 값이 0.95보다 크면 우연에 의해서만 결과를 설명할 수 있는 가능성이 5% 미만임을 나타냅니다.

중요도 레이블. 출력의 각 필드 쌍 또는 그룹에 레이블을 지정하는 데 사용되는 레이블을 편집할 수 있습니다. 기본 레이블은 *중요, 보통, 중요하지 않음*입니다.

절사 값. 각 순위에 대한 임계값을 지정합니다. 일반적으로 p 값이 0.95보다 크면 중요로 순위화되고 이 값이 0.9보다 작으면 중요하지 않음으로 순위화되지만 이 임계값은 필요에 따라 조정할 수 있습니다.

참고: 중요도 척도는 다수의 노드에서 사용할 수 있습니다. 구체적인 계산은 노드와 사용된 목표 및 입력 필드의 유형에 따라 다르지만 값은 모두 백분을 척도로 측정되기 때문에 계속 비교할 수 있습니다.

④ 평균 노드 출력 브라우저

평균 출력 브라우저는 교차 분석표 데이터를 표시하며 한 번에 한 행씩 테이블을 선택하여 복사하고 열별로 정렬하고 테이블을 저장 및 인쇄하는 등의 표준 조작을 수행할 수 있게 합니다. 자세한 정보는 출력 보기 주제를 참조하십시오.

테이블의 구체적인 정보는 비교 유형(별도의 필드 또는 하나의 필드에 있는 그룹)에 따라 다릅니다.

정렬 기준. 특정 열을 기준으로 출력을 정렬할 수 있게 합니다. 위로 또는 아래로 화살표를 클릭하여 정렬 방향을 변경하십시오. 또는 열 표제를 클릭하여 해당 열을 기준으로 정렬할 수 있습니다. (열에서 정렬 방향을 변경하려면 다시 클릭하십시오.)

보기. **단순** 또는 **고급**을 선택하여 표시의 세부사항 수준을 제어할 수 있습니다. 고급 보기는 단순 보기의 모든 정보가 포함되어 있으며 추가적인 세부사항도 제공합니다.

가. 필드 내 평균 출력 비교 그룹

필드 내 그룹을 비교하면 그룹화 필드의 이름이 출력 테이블 위에 표시되고 평균 및 관련 통계가 각 그룹에 대해 별도로 보고됩니다. 해당 테이블에는 각 검정 필드에 대한 별도의 행이 포함되어 있습니다.

다음의 열이 표시됩니다.

- **필드**. 선택된 검정 필드의 이름을 나열합니다.
- **그룹별 평균**. 그룹화 필드의 각 범주에 대한 평균을 표시합니다. 예를 들어, 특별 판매 제안 (새 프로모션)을 받은 사용자를 해당 제안을 받지 않은 사용자(표준)와 비교할 수 있습니다. 고급 보기에는 표준 편차, 표준 오차 및 개수도 표시됩니다.
- **중요도**. 중요도 값 및 레이블을 표시합니다. 자세한 정보는 평균 노드 옵션 주제를 참조하십시오.

고급 출력

고급 보기에는 다음과 같은 추가적인 열이 표시됩니다.

- **F 검정**. 이 검정은 그룹과 각 그룹 내 분산 사이의 분산 비율을 기반으로 합니다. 모든 그룹에 대해 평균이 동일하면 둘 다 동일한 모그룹 분산의 추정값이므로 F 비가 1에 가까울 것으로 예상합니다. 이 비율이 클수록 그룹 간 변동이 더 크고 유의차가 존재할 가능성이 더 높습니다.
- **자유도**. 자유도를 표시합니다.

나. 필드의 평균 출력 비교 쌍

개별 필드를 비교하면 출력 테이블에 선택된 각 필드 쌍에 대해 하나의 행이 포함됩니다.

- **필드 1/2**. 각 쌍에서 첫 번째 및 두 번째 필드의 이름을 표시합니다. 고급 보기에는 표준 편차, 표준 오차 및 개수도 표시됩니다.
- **평균 1/2**. 각 필드에 대한 평균을 각각 표시합니다.
- **상관관계**. 두 연속형(숫자 범위) 필드 간 관계의 강도를 측정합니다. +1.0에 가까운 값은 강한 긍정적인 연관을 표시하고 -1.0에 가까운 값은 강한 부정적인 연관을 표시합니다. 자세한 정보는 상관관계 설정 주제를 참조하십시오.
- **평균 차이**. 두 필드 평균 간 차이를 표시합니다.
- **중요도**. 중요도 값 및 레이블을 표시합니다. 자세한 정보는 평균 노드 옵션 주제를 참조하십시오.

고급 출력

고급 출력은 다음의 열을 추가합니다.

95% 신뢰구간. 참 평균이 이 모집단에서 이 크기의 가능한 모든 표본 중 95%에 속할 수 있는 범위의 상한 및 하한입니다.

T 검정. t 통계는 평균 차이를 해당 표준 오차로 나눠서 얻습니다. 이 통계의 절대값이 클수록 평균이 동일하지 않을 확률이 높습니다.

자유도. 통계의 자유도를 표시합니다.

(11) 보고서 노드

보고서 노드를 사용하면 고정 텍스트뿐 아니라 데이터 및 데이터로부터 파생된 기타 표현식을 포함한 형식화된 보고서를 작성할 수 있습니다. 텍스트 템플릿을 사용하여 보고서의 형식을 지정하여 고정 텍스트 및 데이터 출력 생성을 정의합니다. 템플릿에서 HTML 태그를 사용하고 출력 탭에서 옵션을 설정하여 사용자 정의 텍스트 형식화를 제공할 수 있습니다. 템플릿에서 CLEM 표현식을 사용하면 데이터 값과 기타 조건부 출력이 보고서에 포함됩니다.

보고서 노드에 대한 대안

보고서 노드는 특정 조건을 충족하는 모든 레코드와 같이 스트림의 레코드 또는 케이스 출력을 나열하는 데 가장 일반적으로 사용됩니다. 따라서 테이블 노드에 대한 덜 구조화된 대안으로 생각될 수 있습니다.

- 유형 노드에서 지정된 필드 정의와 같은 데이터 자체가 아니라 필드 정보 또는 스트림에서 정의된 사항 등이 나열된 보고서를 원하는 경우에는 스크립트를 대신 사용할 수 있습니다.
- 다중 출력 오브젝트(하나 이상의 스트림에 의해 생성된 모델, 테이블 및 그래프 컬렉션 등)를 포함하며 다중 형식(텍스트, HTML 및 Microsoft Word/Office 등)의 출력으로 사용할 수 있는 보고서를 원하는 경우에는 IBM® SPSS® Modeler 프로젝트를 사용하십시오.
- 스크립트를 사용하지 않고 필드 이름 목록을 생성하려면 모든 레코드를 삭제하는 표본 노드를 먼저 사용하고 테이블 노드를 사용할 수 있습니다. 그러면 행이 없는 테이블이 생성되어 단일 열에 필드 이름 목록을 생성하도록 내보낼 때 전치될 수 있습니다. (수행하려면 테이블 노드의 출력 탭에서 **데이터 전치**를 선택하십시오.)

① 보고서 노드 템플릿 탭

템플릿 작성. 보고서의 콘텐츠를 정의하기 위해 보고서 노드 템플릿 탭에서 템플릿을 작성할 수 있습니다. 템플릿은 각 줄이 보고서의 콘텐츠를 지정하는 텍스트, 콘텐츠 줄의 범위를 표시하는 데 사용되는 특수 태그 줄로 구성됩니다. 각 콘텐츠 줄 내의 대괄호([])로 묶인 CLEM 표현식은 해당 줄을 보고서에 보내기 전에 평가됩니다. 템플릿 내의 줄에 대해 가능한 범위는 세 가지입니다.

고정됨. 달리 표시되지 않은 줄은 고정된 것으로 간주됩니다. 고정된 줄은 포함된 표현식이 모두 평가된 후에 보고서에 한 번만 복사됩니다. 예를 들어,

This is my report, printed on [@TODAY]

줄은 텍스트 및 현재 날짜를 포함하는 단일 줄을 보고서에 복사합니다.

글로벌(반복계산 ALL). 특수 태그 #ALL 및 # 사이에 포함된 줄은 입력 데이터의 각 레코드에 대해 한 번씩 보고서에 복사됩니다. 대괄호로 묶인 CLEM 표현식은 각 출력 줄의 현재 레코드를 기반으로 하여 평가됩니다. 예를 들어,

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

줄은 각 레코드에 대해 레코드 번호 및 나이를 표시하는 한 줄을 포함합니다.

모든 레코드의 목록을 생성하려면 다음을 수행하십시오.

```
#ALL
[Age] [Sex] [Cholesterol] [BP]
#
```

조건부(반복계산 WHERE). 특수 태그 #WHERE <condition> 및 # 사이에 포함된 줄은 지정된 조건이 참인 경우에 각 레코드에 대해 한 번씩 보고서에 복사됩니다. 조건은 CLEM 표현식입니다. (WHERE 조건에서 대괄호는 선택적입니다.) 예를 들어,

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

줄은 성별에 대해 M 값을 갖는 각 레코드에 대해 파일에 한 줄을 작성합니다. 완전한 보고서는 템플리트를 입력 데이터에 적용하여 정의된 고정, 글로벌 및 조건부 줄을 포함합니다.

출력 탭을 사용하여 결과를 표시하거나 저장하기 위해 다양한 유형의 출력 노드에 대해 공통적인 옵션을 지정할 수 있습니다. 자세한 정보는 출력 노드 출력 탭 주제를 참조하십시오.

HTML 또는 XML 형식의 출력 데이터

이러한 형식 중 하나로 보고서를 쓰기 위해 템플리트에 직접 HTML 또는 XML 태그를 포함할 수 있습니다. 예를 들어, 다음 템플리트는 HTML 테이블을 생성합니다.

```
This report is written in HTML.
Only records where Age is above 60 are included.

<HTML>
  <TABLE border="2">
    <TR>
      <TD>Age</TD>
      <TD>BP</TD>
      <TD>Cholesterol</TD>
      <TD>Drug</TD>
    </TR>

    #WHERE Age > 60
    <TR>
      <TD>[Age]</TD>
      <TD>[BP]</TD>
      <TD>[Cholesterol]</TD>
      <TD>[Drug]</TD>
    </TR>

  #
</TABLE>
</HTML>
```

② 보고서 노드 출력 브라우저

보고서 브라우저는 사용자에게 생성된 보고서의 콘텐츠를 표시합니다. 파일 메뉴에서 일반적인 저장, 내보내기 및 인쇄 옵션을 사용할 수 있으며 일반적인 편집 옵션은 편집 메뉴에서 사용할 수 있습니다. 자세한 정보는 출력 보기 주제를 참조하십시오.

(12) 전역값 설정 노드

전역값 설정 노드는 데이터를 스캔하고 CLEM 표현식에서 사용할 수 있는 요약 값을 계산합니다. 예를 들어, 전역값 설정 노드를 사용하여 *age*라는 필드에 대한 통계량을 계산한 후 @GLOBAL_MEAN(*age*) 함수를 삽입하여 CLEM 표현식에서 *age*의 전체 평균을 사용할 수 있습니다.

① 전역값 설정 노드 설정 탭

전역값 작성. 전역값을 사용 가능하도록 만들 필드를 선택합니다. 다중 필드를 선택할 수 있습니다. 각 필드에 대해 필드 이름 옆의 열에 원하는 통계가 선택되었는지 확인하여 계산할 통계를 지정하십시오.

- **평균.** 모든 레코드 전체에 걸친 필드의 평균 값입니다.
- **합계.** 모든 레코드 전체에 걸친 필드의 값의 합계입니다.
- **최소.** 필드의 최소값입니다.
- **최대.** 필드의 최대값입니다.
- **표준편차.** 표준 편차입니다. 필드 값에서의 변동 측도이며 분산의 제곱근으로 계산됩니다.

기본 작업. 새 필드가 위의 전역값 목록에 추가될 때 여기서 선택되는 옵션이 사용됩니다. 기본 통계 집합을 변경하려면 적절히 통계량을 선택 또는 선택 취소하십시오. 또한 **적용** 단추를 사용하여 기본 작업을 목록 내의 모든 필드에 적용할 수 있습니다.

 **참고:** 일부 작업은 숫자가 아닌 필드(날짜/시간 필드의 합계 등)에는 적용할 수 없습니다. 선택된 필드와 함께 사용할 수 없는 작업은 사용할 수 없도록 설정됩니다.

실행 전에 모든 전역값 선택 취소. 새로운 값을 계산하기 전에 모든 전역값을 제거하려면 이 옵션을 선택하십시오. 이 옵션을 선택하지 않으면 새로 계산된 값이 기존값을 대체하나 다시 계산되지 않은 전역값은 사용 가능한 상태로 유지됩니다.

실행 후에 작성된 전역값 미리보기 표시. 이 옵션을 선택하면 계산된 전역값을 표시하기 위해 실행 후에 스트림 특성 대화 상자의 전역값 탭이 표시됩니다.

(13) 시뮬레이션 적합 노드

시뮬레이션 적합 노드는 후보 통계 분포 집합을 데이터 내의 각 필드에 맞춥니다. 필드에 대한 각 분포의 적합도는 적합도 기준을 사용하여 평가됩니다. 시뮬레이션 적합 노드가 실행될 때 시뮬레이션 생성 노드가 작성되거나 기존 노드가 업데이트됩니다. 각 필드에 가장 적합한 분포가 지정됩니다. 시뮬레이션 생성 노드를 사용하여 각 필드에 대한 시뮬레이션된 데이터를 생성할 수 있습니다.

시뮬레이션 적합 노드가 터미널 노드이더라도 생성된 모형 팔레트에 모형을 추가하거나 출력 탭에 출력 또는 도표를 추가하거나 데이터를 내보내지 않습니다.

참고: 히스토리 데이터가 희박한 경우, 즉, 결측값이 많은 경우, 분포를 데이터에 맞추기에 충분한 유효한 값을 찾기 위해 구성요소를 맞추는 데 어려움이 있을 수 있습니다. 데이터가 희박한 경우, 맞춤 수행 전에 희박한 데이터가 필수가 아니면 필드를 제거하거나 결측값을 대체해야 합니다. 데이터 검토 노드의 **품질** 탭에서 옵션을 사용하여 완료된 레코드의 수를 보고 희박한 필드를 식별하고 대체 방법을 선택할 수 있습니다. 레코드 수가 분포 맞춤에 대해 충분하지 않으면 균형 노드를 사용하여 레코드 수를 늘릴 수 있습니다.

시뮬레이션 생성 노드를 자동으로 작성하기 위해 시뮬레이션 적합 노드 사용

시뮬레이션 적합 노드가 처음 실행될 때 시뮬레이션 적합 노드에 대한 업데이트 링크가 있는 시뮬레이션 생성 노드가 작성됩니다. 시뮬레이션 적합 노드가 다시 실행될 때 업데이트 링크가 제거된 경우에만 새 시뮬레이션 생성 노드가 작성됩니다. 시뮬레이션 적합 노드는 연결된 시뮬레이션 생성 노드를 업데이트하는 데에도 사용될 수 있습니다. 결과는 동일한 필드가 두 노드에 모두 존재하는지, 필드가 시뮬레이션 생성 노드에서 잠겨있지 않은지 여부에 따라 다릅니다. 자세한 정보는 시뮬레이션 생성 노드의 내용을 참조하십시오.

시뮬레이션 적합 노드는 시뮬레이션 생성 노드에 대한 업데이트 링크만 가질 수 있습니다. 시뮬레이션 생성 노드에 대한 업데이트를 정의하려면 다음 단계를 따르십시오.

1. 시뮬레이션 적합 노드를 마우스 오른쪽 단추로 클릭하십시오.
2. 메뉴에서 **업데이트 링크 정의**를 선택하십시오.
3. 업데이트 링크를 정의할 시뮬레이션 생성 노드를 클릭하십시오.

시뮬레이션 적합 노드 및 시뮬레이션 생성 노드 사이의 업데이트 링크를 제거하려면 마우스 오른쪽 단추로 업데이트 링크를 클릭하고 **링크 제거**를 선택하십시오.

① 분포 적합

통계 분포는 변수가 취할 수 있는 발생 값의 이론적 빈도입니다. 시뮬레이션 적합 노드에서 이론적 통계 분포 집합은 데이터의 각 필드와 비교됩니다. 맞춤에 대해 사용 가능한 분포에 대해

서는 분포 제목에서 설명합니다. 이론적 분포의 모수는 적합도(Anderson-Darling 기준 또는 Kolmogorov-Smirnov 기준)에 따라 데이터에 최적 맞춤될 수 있도록 조정됩니다. 시뮬레이션 적합 노드에 의한 분포 맞춤의 결과는 어떤 분포가 적합한지, 각 분포에 대한 모수의 최적 추정 값 및 각 분포가 데이터에 적합한 정도 등을 표시합니다. 분포 적합 작업 중에 수치 저장 유형이 있는 필드 사이의 상관계수 및 범주형 분포가 있는 필드 사이의 우연성도 계산됩니다. 분포 맞춤의 결과는 시뮬레이션 생성 노드를 작성하는 데 사용됩니다.

분포가 데이터에 맞춰지기 전에 첫 번째 1000 개의 레코드에서 결측값을 검사합니다. 결측값이 너무 많으면 분포 맞춤이 가능하지 않습니다. 그런 경우, 다음 옵션 중 적합한 옵션을 선택해야 합니다.

- 결측값이 있는 레코드를 제거하기 위해 업스트림 노드 사용.
- 결측값에 대한 레코드를 대체하기 위해 업스트림 노드 사용.

분포 맞춤은 사용자 결측값을 제외하지 않습니다. 데이터에 사용자 결측값이 있으며 이러한 값을 분포 맞춤에서 제외하려면 해당 값을 시스템 결측값으로 설정해야 합니다.

분포가 맞춰질 때 필드의 역할은 고려되지 않습니다. 예를 들어, 역할이 **목표인** 필드는 역할이 **입력, 없음, 모두, 파티션, 분할, 빈도 및 ID**인 필드와 동일하게 처리됩니다.

필드는 저장 유형 및 측정 수준에 따라 분포 적합 작업 중에 다르게 처리됩니다. 분포 적합 작업 중의 필드 처리에 대해서는 다음 표에서 설명합니다.

표 1. 필드의 저장 유형 및 측정 수준에 따른 분포 적합

저장 유형	측정 수준					
	연속	범주형	플래그	명목	순서	유형 없음
문자열	불가능		범주형, Dice 및 고정 분포가 맞춰집니다.			
정수						
실수						
시간	모두 분포가 맞춰집니다. 상관관계 및 우연성이 계산됩니다.		범주형 분포가 맞춰집니다. 상관관계는 계산되지 않습니다.		이항, 음이항 및 포아송 분포가 맞춰지고 상관관계가 계산됩니다.	필드가 무시되어 시뮬레이션 생성 노드로 전달되지 않습니다.
날짜						
시간소인						
알 수 없음			적절한 저장 유형이 데이터에서 결정됩니다.			

측정 수준 순서가 있는 필드는 연속형 필드처럼 처리되고 시뮬레이션 생성 노드의 상관관계 표에 포함됩니다. 이항, 음이항 또는 포아송 외의 분포를 순서 필드에 맞추려면 필드의 측정 수준을 연속형으로 변경해야 합니다. 순서 필드의 각 값에 대해 앞에서 레이블을 정의한 경우, 측정 수준을 연속형으로 변경하면 레이블이 손실됩니다.

단일값을 가진 필드는 분포 적합 작업 중에 다중 값을 가진 필드와 다르게 처리되지 않습니다. 저장 유형 시간, 날짜 또는 시간소인을 가진 필드는 수치로 처리됩니다.

분할 필드에 분포 적합

데이터에 분할 필드가 포함되며 각 분할에 대해 분포 맞춤을 별도로 수행하려면 업스트림 구조 변환 노드를 사용하여 데이터를 변환해야 합니다. 구조변환 노드를 사용하여 분할 필드의 각 값에 대해 새 필드를 생성하십시오. 그러면 이 구조변환된 데이터를 시뮬레이션 적합 노드에서 분포 적합 작업 중에 사용할 수 있습니다.

② 시뮬레이션 적합 노드 설정 탭

소스 노드 이름. **자동**을 선택하여 자동으로 생성되거나 업데이트된 시뮬레이션 노드의 이름을 생성할 수 있습니다. 자동으로 생성되는 이름은 사용자 정의 이름이 지정된 경우에는 시뮬레이션 적합 노드에서 지정된 이름이며 시뮬레이션 적합 노드에서 사용자 정의 이름이 지정되지 않은 경우에는 Sim Gen입니다. 인접한 텍스트 필드에서 사용자 정의 이름을 지정하려면 **사용자 정의**를 선택하십시오. 텍스트 필드를 편집하지 않은 한 기본 사용자 정의 이름은 Sim Gen입니다.

맞춤 옵션 이러한 옵션을 사용하여 분포가 필드에 맞춰지는 방법 및 분포의 맞춤이 평가되는 방법을 지정할 수 있습니다.

- **표본추출할 케이스 수.** 데이터 세트 내의 필드로 분포를 맞춤 때 사용할 케이스 수를 지정합니다. 데이터 내의 모든 레코드에 분포를 맞추려면 **모두**를 선택하십시오. 데이터 세트가 매우 크면 분포 맞춤에 사용할 케이스 수를 제한하는 방법을 고려할 수도 있습니다. 첫 N 케이스만 사용하려면 **첫 N 케이스로 제한**을 선택하십시오. 사용할 케이스 수를 지정하려면 화살표를 클릭하십시오. 또는 업스트림 노드를 사용하여 분포 맞춤에 대한 레코드의 임의 표본을 사용할 수 있습니다.
- **기준 적합도(연속형 필드 전용).** 연속형 필드의 경우, 필드에 대한 분포를 맞춤 때 Anderson-Darling 검정 또는 Kolmogorov-Smirnoff 검정 적합도를 선택하여 분포 순위를 매기십시오. Anderson-Darling 검정이 기본적으로 선택되며 끝 영역에서 가장 적합한 맞춤을 사용하려면 특히 권장됩니다. 모든 후보 분포에 대해 모든 통계량이 계산되나 분포 정렬 및 가장 적합한 분포 적합 판별에는 선택된 통계만 사용됩니다.
- **구간(경험적 분포 전용).** 연속형 필드의 경우, 경험적 분포는 히스토리 데이터의 누적 분포 함수입니다. 각 값 또는 값 범위의 확률이며 데이터에서 직접 파생됩니다. 화살표를 클릭하여 연속형 필드에 대한 경험적 분포를 계산하는 데 사용되는 구간 수를 지정할 수 있습니다. 기본 값은 100이며 최대값은 1000입니다.

- **가중 필드(선택적).** 데이터 세트에 가중 필드가 포함된 경우, 필드 선택도구 아이콘을 클릭하여 목록에서 가중 필드를 선택하십시오. 그러면 분포 적합 프로세스에서 가중 필드가 제외됩니다. 목록은 측정 수준이 연속적인 데이터 세트 내의 모든 필드를 표시합니다. 가중 필드는 한 개만 선택할 수 있습니다.

(14) 시뮬레이션 평가 노드

시뮬레이션 평가 노드는 지정된 필드를 평가하고 필드의 분포를 제공하고 분포 및 상관계수 도표를 생성하는 터미널 노드입니다. 이 노드는 주로 연속형 필드를 평가하는 데 사용됩니다. 따라서 평가 노드에 의해 생성되는 평가 차트를 보완하며 이산형 필드 평가에 유용합니다. 또 다른 차이점은 시뮬레이션 평가 노드는 여러 반복에 걸친 단일 예측을 평가하는 반면 평가 노드는 단일 반복이 있는 다중 평가를 각각 평가한다는 점입니다. 시뮬레이션 생성 노드의 분포모수에 대해 둘 이상의 값이 지정된 경우에 반복이 생성됩니다. 자세한 정보는 반복 주제를 참조하십시오.

시뮬레이션 평가 노드는 시뮬레이션 적합 및 시뮬레이션 생성 노드에서 얻은 데이터를 사용하여 계획됩니다. 단, 노드는 다른 임의의 노드와 함께 사용될 수 있습니다. 시뮬레이션 생성 노드 및 시뮬레이션 평가 노드 사이에 수에 관계없이 처리 단계를 배치할 수 있습니다.

❖ **중요사항:** 시뮬레이션 평가 노드에는 목표 필드에 대한 유효한 값이 있는 1000개 이상의 레코드가 필요합니다.

① 시뮬레이션 평가 노드 설정 탭

시뮬레이션 평가 노드의 설정 탭에서 데이터 세트 내의 각 필드의 역할을 지정하고 시뮬레이션에 의해 생성되는 출력을 사용자 정의할 수 있습니다.

항목 선택. 시뮬레이션 평가 노드의 세 가지 뷰(필드, 밀도 함수 및 출력) 사이에서 전환할 수 있습니다.

필드 보기

대상 필드. 필수 필드입니다. 드롭 다운 목록에서 데이터 세트의 목표 필드를 선택하려면 화살표를 클릭하십시오. 선택된 필드는 연속형, 순서 또는 명목 측정 수준은 가질 수 있으나 날짜 또는 지정되지 않은 측정 수준은 가질 수 없습니다.

반복 필드(선택적). 데이터에 데이터 내의 각 레코드가 속한 반복을 표시하는 반복 필드가 있으면 여기서 선택해야 합니다. 즉, 각 반복이 별도로 평가됩니다. 연속형, 순서 또는 명목 측정 수준의 필드만 선택할 수 있습니다.

입력 데이터가 이미 반복에 의해 정렬되어 있음. 반복 필드가 **반복 필드(선택적)** 필드에서 지정되는 경우에만 사용 가능합니다. 입력 데이터가 **반복 필드(선택적)**에서 지정된 반복 필드에 의해 이미 정렬되었음을 확인하는 경우에만 이 옵션을 선택하십시오.

구성할 최대 반복 수. 반복 필드가 **반복 필드(선택적)** 필드에서 지정되는 경우에만 사용 가능합니다. 구성할 반복계산 수를 지정하려면 화살표를 클릭하십시오. 이 번호를 지정하면 단일 도표에 너무 많은 반복을 구성하여 도표 해석을 어렵게 만드는 것을 방지할 수 있습니다. 설정 가능한 최저 수준의 최대반복수는 2입니다. 최고 수준은 50입니다. 구성할 최대반복수는 처음에는 10으로 설정됩니다.

상관관계 토네이도에 대한 입력 필드. 상관관계 토네이도 도표는 지정된 목표 및 지정된 각 입력 사이의 상관계수를 표시하는 막대형 차트입니다. 사용가능한 시뮬레이션한 입력 목록에서 필드 선택도구 아이콘을 클릭하여 토네이도 도표에 포함할 입력 필드를 선택하십시오. 연속형 및 순서형 측정 수준이 있는 입력 필드만 선택할 수 있습니다. 명목형, 유형 없음 및 날짜 입력 필드는 목록에 사용할 수 없으며 선택할 수 없습니다.

밀도 함수 보기

이 보기의 옵션을 사용하면 범주형 대상에 대한 예측값의 막대형 차트와 마찬가지로 연속형 대상에 대한 확률 밀도 함수 및 누적 분포 함수의 출력을 사용자 정의할 수 있습니다.

밀도함수. 밀도함수는 시뮬레이션에서 결과 세트를 조사하는 기본적인 방법입니다.

- **확률 밀도 함수(PDF).** 목표 필드에 대한 확률 밀도 함수를 생성하려면 이 옵션을 선택하십시오. 확률 밀도 함수는 목표 값의 분포를 표시합니다. 확률 밀도 함수를 사용하면 목표가 특정 영역 내에 있을 확률을 판별할 수 있습니다. 범주형 대상(측정 수준이 명목형 또는 순서인 경우)의 경우, 각 범주의 대상이 해당되는 케이스 퍼센트를 표시하는 막대형 차트가 생성됩니다.
- **누적 분포 함수(CDF).** 목표 필드에 대한 누적 분포 함수를 생성하려면 이 옵션을 선택하십시오. 누적 분포 함수는 대상의 값이 지정된 값 이하인 확률을 표시합니다. 연속형 대상에만 사용 가능합니다.

참조선(연속형). 이러한 옵션은 **확률 밀도 함수(PDF)** 또는 **누적 분포 함수(CDF)** 또는 둘 다 선택된 경우에만 사용 가능합니다. 해당 옵션을 사용하면 확률 밀도 함수 및 누적 분포 함수에 다양한 고정된 수직 참조선을 추가할 수 있습니다.

- **평균.** 목표 필드의 평균 값에 참조선을 추가하려면 이 옵션을 선택하십시오.
- **중앙값.** 목표 필드의 중앙값에 참조선을 추가하려면 이 옵션을 선택하십시오.
- **표준 편차.** 목표 필드의 평균 값에서부터 지정된 수의 표준편차 더하기 및 빼기에 참조선을 추가하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 인접한 **숫자** 필드를 사용할 수 있습니다. 표준편차의 수를 지정하려면 화살표를 클릭하십시오. 최소 표준편차 수는 1이며 최대수는 10입니다. 표준편차의 수는 처음에 3으로 설정됩니다.
- **백분위수.** 목표 필드의 분포의 두 개의 백분위수 값에 참조선을 추가하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 인접한 **아래쪽** 및 **위쪽** 텍스트 필드를 사용할 수 있습니다. 예를 들어, **위쪽** 텍스트 필드에 90 값을 입력하면 목표의 90번째 백분위수에 참조선이 추가됩니다.

이 값은 관측값의 90% 아래에 해당하는 값입니다. 이와 유사하게 **아래쪽** 텍스트 필드의 10 값은 목표의 열 번째 백분위수를 나타내며 관측값의 10% 아래에 해당되는 값입니다.

- **사용자 정의 참조선.** 수평축 변수와 함께 지정된 값에 참조선을 추가하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 인접한 **값** 테이블을 사용할 수 있습니다. 유효한 숫자를 **값** 테이블에 입력할 때마다 비어 있는 새 행이 테이블의 아래 쪽에 추가됩니다. 유효한 수는 목표 필드의 값 범위 내의 수입입니다.

참고: (다중 반복으로부터) 다중 밀도함수 또는 분포 함수가 단일 차트에 표시되는 경우, 사용자 정의 선이 아니라 참조선이 별도로 각 함수에 적용됩니다.

범주형 목표(PDF만 해당). 이러한 옵션은 **확률 밀도 함수(PDF)**가 선택된 경우에만 사용 가능합니다.

- **보고할 범주 값.** 범주형 목표 필드가 있는 모형의 경우, 모형의 결과는 목표 값이 각 범주에 속하는 각 범주에 대해 하나씩 존재하는 예측 확률의 집합입니다. 가장 높은 확률의 범주가 예측 범주가 되고 확률 밀도 함수에 대한 막대형 차트를 생성하는 데 사용됩니다. 막대형 차트를 생성하려면 **예측 범주**를 선택하십시오. 목표 필드의 각 범주에 대한 예측 확률 분포 히스토그램을 생성하려면 **예측 확률**을 선택하십시오. 또한 두 유형의 차트를 모두 생성하기 위해 모두를 선택할 수 있습니다.
- **민감도 분석의 그룹화.** 민감도 분석 반복계산이 포함된 시뮬레이션은 분석에 의해 정의되는 각 반복에 대해 독립적 목표 필드(또는 모델의 예측 목표 필드)를 생성합니다. 변형되는 분포 모수의 각 값에 대해 하나의 반복이 있습니다. 반복이 있으면 범주형 목표 필드의 예측 범주의 막대형 차트가 모든 반복의 결과를 포함하는 수평배열 막대도표로 표시됩니다. **범주를 함께 그룹화** 또는 **반복계산을 함께 그룹화**를 선택하십시오.

출력 보기

목표 분포의 백분위수 값. 이 옵션을 사용하면 목표 분포의 백분위수 값의 표를 만들고 표시할 백분위수를 지정할 수 있습니다.

백분위수 값의 표 만들기. 연속형 목표 필드의 경우, 목표 분포의 지정된 백분위수 표를 얻으려면 이 옵션을 선택하십시오. 다음 옵션 중 하나를 선택하여 백분위수를 지정하십시오.

- **사분위수.** 사분위수는 목표 필드 분포의 25번째, 50번째, 75번째 백분위수입니다. 관측값은 네 그룹의 동일한 크기로 나뉩니다.
- **구간.** 네 개가 아닌 동일한 수의 그룹이 필요하면 구간을 선택하십시오. 이 옵션을 선택하면 인접한 **숫자** 필드를 사용할 수 있습니다. 구간 수를 지정하려면 화살표를 클릭하십시오. 최소 구간 수는 2이며 최대수는 100입니다. 구간 수는 처음에 10으로 설정됩니다.
- **사용자 정의 백분위수.** 개별 백분위수(예를 들어, 99번째 백분위수)를 지정하려면 **사용자 정의 백분위수**를 선택하십시오. 이 옵션을 선택하면 인접한 **값** 테이블을 사용할 수 있습니다. 1에서 100 사이의 유효한 숫자를 **값** 테이블에 입력할 때마다 비어 있는 새 행이 테이블의 아래 쪽에 추가됩니다.

② 시뮬레이션 평가 노드 출력

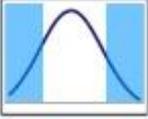
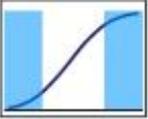
시뮬레이션 평가 노드가 실행될 때 출력이 출력 관리자에 추가됩니다. 시뮬레이션 평가 출력 브라우저는 시뮬레이션 평가 노드의 실행 결과를 표시합니다. **파일** 메뉴에서 일반적인 저장, 내보내기 및 인쇄 옵션을 사용할 수 있으며 일반적인 편집 옵션은 **편집** 메뉴에서 사용할 수 있습니다. 자세한 정보는 출력 보기 주제를 참조하십시오. 도표 중 하나를 선택하면 **보기** 메뉴만 사용 가능합니다. 분포 테이블 또는 정보 출력에는 사용할 수 없습니다. **보기** 메뉴에서 **편집 모드**를 선택하여 도표의 레이아웃 및 모양을 변경하거나 **탐색 모드**를 선택하여 도표에 표시되는 데이터 및 값을 탐색할 수 있습니다. 정적 모드는 도표 참조선 및 슬라이더를 이동할 수 없도록 현재 위치에서 고정합니다. 정적 모드는 참조선이 있는 도표를 복사, 인쇄 또는 내보낼 수 있는 유일한 모드입니다. 이 모드를 선택하려면 **보기** 메뉴에서 **정적 모드**를 클릭하십시오.

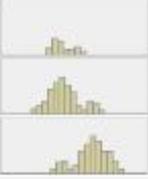
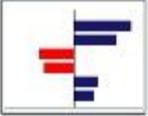
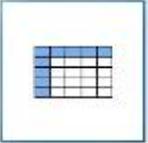
시뮬레이션 평가 출력 브라우저 창은 두 개의 패널로 구성됩니다. 창의 왼쪽에는 시뮬레이션 평가 노드가 실행될 때 생성된 도표의 썸네일 표시가 표시되는 탐색 패널이 있습니다. 썸네일을 선택하면 창의 오른쪽에 있는 패널에 도표 출력이 표시됩니다.

가. 탐색 패널

출력 브라우저의 탐색 패널에는 시뮬레이션에서 생성된 도표의 썸네일이 포함됩니다. 탐색 패널에 표시되는 썸네일은 목표 필드의 측정 수준 및 시뮬레이션 평가 노드 대화 상자에서 선택된 옵션에 따라 다릅니다. 썸네일에 대한 설명은 다음 표에서 제공합니다.

표 1. 탐색 패널 썸네일

썸네일	설명	설명
	확률 밀도 함수	이 썸네일은 목표 필드의 측정 수준이 연속형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 확률 밀도 함수(PDF) 가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 범주형이면 이 썸네일이 표시되지 않습니다.
	누적 분포 함수	이 썸네일은 목표 필드의 측정 수준이 연속형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 누적 분포 함수(CDF) 가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 범주형이면 이 썸네일이 표시되지 않습니다.
	예측된 범주 값	이 썸네일은 목표 필드의 측정 수준이 범주형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 확률 밀도 함수(PDF) 가 선택되고 보고할 범주 값 영역에서 예측된 범주 또는 모두 가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 연속형이면 이 썸네일이 표시되지 않습니다.

썸네일	설명	설명
	예측된 범주 확률	이 썸네일은 목표 필드의 측정 수준이 범주형이고 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기에서 확률 밀도 함수(PDF) 가 선택되고 보고할 범주 값 영역에서 예측된 확률 또는 모두 가 선택된 경우에만 표시됩니다. 목표 필드의 측정 수준이 연속형이면 이 썸네일이 표시되지 않습니다.
	토네이도 도표	이 썸네일은 시뮬레이션 평가 노드 대화 상자의 필드 보기의 상관관계 토네이도의 입력 필드 필드에서 하나 이상의 입력 필드가 선택된 경우에만 표시됩니다.
	분포 표	이 썸네일은 목표 필드의 측정 수준이 연속형이고 시뮬레이션 평가 노드 대화 상자의 출력 보기에서 백분위수 값의 표 만들기 가 선택된 경우에만 표시됩니다. 이 도표에는 보기 메뉴를 사용할 수 없습니다. 목표 필드의 측정 수준이 범주형이면 이 썸네일이 표시되지 않습니다.
	정보	이 썸네일은 항상 표시됩니다. 이 출력에는 보기 메뉴를 사용할 수 없습니다.

나. 차트 출력

사용 가능한 출력 도표의 유형은 목표 필드의 측정 수준, 반복 필드 사용 여부 및 시뮬레이션 평가 노드 대화 상자에서 선택된 옵션에 따라 다릅니다. 시뮬레이션에서 생성된 수많은 도표는 표시를 사용자 정의할 수 있는 대화형 기능을 갖고 있습니다. 대화형 기능은 **도표 옵션**을 클릭하여 사용 가능합니다. 모든 시뮬레이션 도표는 그래프 보드로 시각화됩니다.

연속형 대상의 확률 밀도 함수 차트. 이 도표는 확률 및 빈도를 모두 표시하며 왼쪽 수직 축에 확률 척도가 있으며 오른쪽 수직 축에 빈도 척도가 있습니다. 도표는 두 개의 슬라이딩 수직 참조선을 사용하여 별도의 영역으로 구분됩니다. 도표 아래 표는 각 영역 내의 분포 퍼센트를 표시합니다. 반복으로 인해 동일한 도표에 다중 밀도함수가 표시되는 경우, 표에 각 밀도 함수와 연관된 확률에 대한 별도의 행, 반복 이름을 포함하는 추가 열 및 각 밀도 함수와 연관된 색상이 있을 수 있습니다. 반복은 반복 레이블에 따라 표에 문자순으로 표시됩니다. 반복 레이블을 사용할 수 없으면 반복 값이 대신 사용됩니다. 표는 편집할 수 없습니다.

각 참조선에는 선을 쉽게 이동할 수 있는 슬라이더(역삼각형)가 있습니다. 각 슬라이더에는 현재 위치를 표시하는 레이블이 있습니다. 기본적으로 슬라이더는 분포의 5번째 및 95번째 백분위수에 위치합니다. 다중 반복이 있는 경우, 표에 나열된 첫 번째 반복의 5번째 및 95번째 백분위수에 슬라이더가 위치합니다. 선을 서로 교차하여 이동할 수 없습니다.

수많은 추가 기능은 **도표 옵션**을 클릭하여 사용 가능합니다. 특히, 슬라이더의 위치를 명시적으로 설정하고 고정 참조선을 추가하고 도표 보기를 연속형 곡선에서 히스토그램으로 변경할 수 있습니다. 자세한 정보는 차트 옵션 주제를 참조하십시오. 도표를 복사하거나 내보내려면 마우스 오른쪽 단추로 도표를 클릭하십시오.

연속형 대상의 누적 분포 함수. 이 도표에는 두 개의 동일한 이동 가능한 수직 참조선 및 확률 밀도 함수 도표에 대해 설명하는 연관된 표가 있습니다. 슬라이더 제어 및 표는 다중 반복이 있는 경우에 확률 밀도 함수와 동일하게 작동합니다. 각 반복에 속한 밀도함수를 식별하기 위해 사용되는 것과 동일한 색상이 분포 함수에 사용됩니다.

또한 이 도표는 슬라이더의 위치를 명시적으로 설정하고 고정 참조선을 추가하고 누적 분포 함수가 증가 함수(기본값) 또는 감소 함수로 표시되는지 여부를 지정하는 데 사용할 수 있는 도표 옵션 대화 상자에 대한 액세스를 제공합니다. 자세한 정보는 차트 옵션 주제를 참조하십시오. 도표를 복사하거나 내보내거나 편집하려면 마우스 오른쪽 단추로 도표를 클릭하십시오. **편집**을 선택하면 Float 그래프보드 편집기 창에 도표가 열립니다.

범주형 대상의 예측 범주 값 도표. 범주형 목표 필드의 경우, 막대형 차트가 예측값을 표시합니다. 예측값은 각 범주에 해당될 것으로 예측되는 목표 필드의 퍼센트로 표시됩니다. 민감도 분석 반복계산이 있는 범주형 목표 필드의 경우, 예측 목표 범주의 결과가 모든 반복의 결과를 포함하는 수평배열 막대도표로 표시됩니다. 도표는 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기의 **민감도 분석의 그룹화** 영역에서 선택한 옵션에 따라 범주 또는 반복에 의해 수평배열됩니다. 도표를 복사하거나 내보내거나 편집하려면 마우스 오른쪽 단추로 도표를 클릭하십시오. **편집**을 선택하면 Float 그래프보드 편집기 창에 도표가 열립니다.

범주형 대상의 예측 범주 확률 도표. 범주형 목표 필드의 경우, 히스토그램은 대상의 각 범주에 대한 예측 확률의 분포를 표시합니다. 민감도 분석 반복계산이 있는 범주형 목표 필드의 경우, 시뮬레이션 평가 노드 대화 상자의 밀도 함수 보기의 **민감도 분석의 그룹화** 영역에서 선택한 옵션에 따라 범주 또는 반복 기준으로 히스토그램이 표시됩니다. 이 히스토그램은 범주 기준으로 그룹화되고 반복 레이블을 포함하는 드롭 다운 목록을 사용하면 표시할 반복을 선택할 수 있습니다. 또한 마우스 오른쪽 단추로 도표를 클릭하고 **반복** 하위 메뉴에서 반복을 선택함으로써 표시할 반복을 선택할 수 있습니다. 이 히스토그램은 반복 기준으로 그룹화되고 범주 이름을 포함하는 드롭 다운 목록을 사용하면 표시할 범주를 선택할 수 있습니다. 또한 마우스 오른쪽 단추로 도표를 클릭하고 **범주** 하위 메뉴에서 범주를 선택함으로써 표시할 범주를 선택할 수 있습니다.

이 도표는 모델의 서브세트에서만 사용 가능하며 모델 너깃에서 모든 그룹 확률을 생성하기 위한 옵션을 선택해야 합니다. 예를 들어, 로지스틱 모델 너깃에서 **모든 확률 추가**를 선택해야 합니다. 다음 모델 너깃은 이 옵션을 지원합니다.

- 로지스틱, SVM, Bayes, 신경망 및 KNN
- 로지스틱 회귀분석, 의사결정 트리 및 Naive Bayes에 대한 Db2/ISW In-Database 마이닝 모형

기본적으로 모든 그룹 확률을 생성하기 위한 옵션은 이러한 모델 너깅에서 선택되지 않습니다. **토네이도 도표.** 토네이도 도표는 각 지정된 입력에 대한 목표 필드의 민감도를 표시하는 막대형 차트입니다. 민감도는 목표와 각 입력의 상관관계에 의해 측정됩니다. 도표의 제목에는 목표 필드의 이름이 포함됩니다. 도표의 각 막대는 목표 필드 및 입력 필드 사이의 상관관계를 나타냅니다. 도표에 포함되는 시뮬레이션한 입력은 시뮬레이션 평가 노드 대화 상자의 필드 보기의 **상관관계 토네이도의 입력 필드** 필드에서 선택된 입력입니다. 각 막대는 상관관계 값으로 레이블이 붙여집니다. 막대는 가장 큰 값에서 가장 작은 값까지 상관관계수의 절대값에 의해 순서가 지정됩니다. 반복이 있는 경우에는 각 반복에 대해 별도의 차트가 생성됩니다. 각 도표에는 반복의 이름을 포함하는 부제목이 있습니다.

분포 표. 이 표에는 목표 필드의 값이 포함되며 해당 값 아래에 관측값의 지정된 퍼센트가 포함됩니다. 표에는 시뮬레이션 평가 노드 대화 상자의 출력 보기에서 지정된 각 백분위수 값에 대한 행이 포함됩니다. 백분위수 값은 사분위수, 동등하게 간격이 지정된 수가 다른 백분위수, 개별적으로 지정된 백분위수 등이 될 수 있습니다. 분포 표에는 각 반복에 대한 열이 포함됩니다.

정보. 이 절에서는 평가에 사용되는 필드 및 레코드의 전체 요약을 제공합니다. 또한 각 반복으로 구분된 입력 필드 및 레코드 빈도를 표시합니다.

다. 차트 옵션

도표 옵션 대화 상자에서는 시뮬레이션에서 생성된 확률 밀도 함수 및 누적 분포 함수의 활성화도 표 표시를 사용자 정의할 수 있습니다.

보기. 보기 드롭 다운 목록은 확률 밀도 함수 차트에만 적용됩니다. 이를 사용하여 연속형 곡선에서 히스토그램으로 차트 보기를 토글할 수 있습니다. 이 기능은 다중 반복의 다중 밀도함수가 동일한 차트에 표시되는 경우에는 사용할 수 없습니다. 다중 밀도함수가 있으면 다중 밀도함수를 연속형 곡선으로만 볼 수 있습니다.

순서. 순서 드롭 다운 목록은 누적 분포 함수 차트에만 적용됩니다. 이는 누적 분포 함수가 오름차순 함수(기본값) 또는 내림차순 함수로 표시되는지 지정합니다. 내림차순 함수로 표시되면 수평축 변수의 지정된 포인트에서 함수의 값이 해당 포인트의 오른쪽에 목표 필드가 놓이는 확률이 됩니다.

슬라이더 위치. 상한 텍스트 필드에는 오른쪽 슬라이딩 참조선의 현재 위치가 포함됩니다. 하한 텍스트 필드에는 왼쪽 슬라이딩 참조선의 현재 위치가 포함됩니다. 상한 및 하한 텍스트 필드에 값을 입력하여 슬라이더의 위치를 명시적으로 설정할 수 있습니다. 하한 텍스트 필드의 값은 반드시 상한 텍스트 필드의 값 미만이어야 합니다. -무한을 선택하여 왼쪽 참조선을 제거하면 위치를 효과적으로 음의 무한대로 설정할 수 있습니다. 이 조치를 수행하면 하한 텍스트 필드를 사용할 수 없습니다. -무한을 선택하여 오른쪽 참조선을 제거하면 위치를 효과적으로 무한대로 설정할 수 있습니다. 이 조치를 수행하면 상한 텍스트 필드를 사용할 수 없습니다. 두 참조선을 모두 제거할 수는 없습니다. -무한을 선택하면 무한대 선택란을 선택할 수 없으며 반대의 경우도 마찬가지입니다.

참조선. 확률 밀도 함수 및 누적 분포 함수에 다양한 고정된 수직 참조선을 추가할 수 있습니다.

- **평균.** 목표 필드의 평균에 참조선을 추가할 수 있습니다.
- **중앙값.** 목표 필드의 중앙값에 참조선을 추가할 수 있습니다.
- **표준 편차.** 목표 필드의 평균 값에서부터 지정된 수의 표준편차 더하기 및 빼기에 참조선을 추가할 수 있습니다. 인접한 텍스트 필드에서 사용할 표준편차의 수를 입력할 수 있습니다. 최소 표준편차 수는 1이며 최대수는 10입니다. 표준편차의 수는 처음에 3으로 설정됩니다.
- **백분위수.** 아래쪽 및 위쪽 텍스트 필드에 값을 입력하여 목표 필드에 대한 분포의 한 개 또는 두 개의 백분위수 값에 참조선을 추가할 수 있습니다. 예를 들어, 위쪽 텍스트 필드의 95 값은 95번째 백분위수를 나타내며 관측값의 95% 아래에 해당되는 값입니다. 이와 유사하게 아래쪽 텍스트 필드의 5 값은 다섯 번째 백분위수를 나타내며 관측값의 5% 아래에 해당되는 값입니다. 아래쪽 텍스트 필드의 경우, 최소 백분위수 값은 0이며 최대수는 49입니다. 위쪽 텍스트 필드의 경우, 최소 백분위수 값은 50이며 최대수는 100입니다.
- **사용자 정의 위치.** 수평축 변수와 함께 지정된 값에 참조선을 추가할 수 있습니다. 눈금에서 항목을 삭제하여 사용자 정의 참조선을 제거할 수 있습니다.

확인을 클릭하면 도표 옵션 대화 상자에서 선택된 옵션을 반영하기 위해 슬라이더, 슬라이더 위의 레이블, 참조선 및 도표 아래의 표가 업데이트됩니다. 변경하지 않고 대화 상자를 닫으려면 **취소**를 클릭하십시오. 참조선은 도표 옵션 대화 상자에서 연관된 선택을 선택 취소하고 **확인**을 클릭하여 제거할 수 있습니다.

참고: 민감도 분석 반복계산의 결과로 인해 다중 밀도함수 또는 분포 함수가 단일 도표에 표시되는 경우, 사용자 정의 선이 아니라 참조선이 각 함수에 별도로 적용됩니다. 첫 번째 반복에 대한 참조선만 표시됩니다. 참조선 레이블에는 반복 레이블이 포함됩니다. 반복 레이블은 업스트림(일반적으로 시뮬레이션 생성 노드)에서 파생됩니다. 반복 레이블을 사용할 수 없으면 반복 값이 대신 사용됩니다. **평균, 중앙값, 표준 편차 및 백분위수** 옵션은 다중 반복이 있는 누적 분포 함수에는 사용할 수 없습니다.

(15) 확장 출력 노드

확장 출력 노드 대화 상자의 출력 탭에서 **화면에 출력**을 선택하면 화면 출력이 출력 브라우저 창에 표시됩니다. 또한 출력이 출력 관리자에 추가됩니다. 출력 브라우저 창에는 출력을 인쇄 또는 저장하거나 다른 형식으로 내보낼 수 있는 메뉴 세트가 있습니다. **편집** 메뉴에는 **복사** 옵션만 있습니다. 확장 출력 노드의 출력 브라우저에는 두 개의 탭, 즉, **텍스트 출력**을 표시하는 텍스트 출력 탭과 그래프 및 차트를 표시하는 **그래프 출력** 탭이 있습니다.

확장 출력 노드 대화 상자의 출력 탭에서 **파일에 출력**을 선택하면 확장 출력 노드가 성공적으로 실행될 때 출력 브라우저 창이 표시되지 않습니다.

① 확장 출력 노드 - 구문 탭

구문 유형(R 또는 Python for Spark)을 선택하십시오. 자세한 정보는 다음 섹션을 참조하십시오. 구문이 준비되면 **실행**을 클릭하여 확장 출력 노드를 실행할 수 있습니다. 출력 개체가 출력 관리자에 추가되거나 선택적으로 출력 탭의 **파일 이름** 필드에서 지정된 파일에 추가됩니다.

R 구문

R 구문. 데이터 분석을 위해 R 스크립트 구문을 이 필드에 입력, 붙여넣기 또는 사용자 정의할 수 있습니다.

플래그 필드 변환. 플래그 필드를 처리하는 방법을 지정합니다. **문자열에서 요인으로**, 정수 및 실수에서 **double**로 및 논리 값(True, False)이라는 두 가지 옵션이 있습니다. 논리 값(True, False)을 선택하면 플래그 필드의 원래 값이 손실됩니다. 예를 들어, 필드에 남성 및 여성 값이 있는 경우, 해당 값이 True 및 False로 변경됩니다.

결측값을 R '사용할 수 없음' 값(NA)으로 변환. 선택하면 모든 결측값이 R NA 값으로 변환됩니다. NA 값은 결측값을 식별하기 위해 R에서 사용됩니다. 사용하는 일부 R 함수에는 데이터에 NA가 포함된 경우에 함수가 작동하는 방식을 제어하는 데 사용되는 인수가 있습니다. 예를 들어, 함수에서 NA를 포함하는 레코드를 자동으로 제외하도록 선택할 수 있습니다. 이 옵션을 선택하지 않으면 모든 결측값이 변경되지 않은 상태로 R에 전달되고 R 스크립트가 실행될 때 오류가 발생할 수 있습니다.

날짜/시간 필드를 시간대의 특수 제어가 있는 R 클래스로 변환. 이 옵션을 선택하면 날짜 또는 날짜/시간 형식의 변수가 R 날짜/시간 개체로 변환됩니다. 다음 옵션 중 하나를 선택해야 합니다.

- R POSIXct. 날짜 또는 날짜/시간 형식의 변수가 R POSIXct 개체로 변환됩니다.
- R POSIXlt (목록). 날짜 또는 날짜/시간 형식의 변수가 R POSIXlt 개체로 변환됩니다.

참고: POSIX 형식은 고급 옵션입니다. R 스크립트에서 날짜/시간 필드가 해당 형식이 필요한 방식으로 처리되도록 지정된 경우에만 이 옵션을 사용하십시오. POSIX 형식은 시간 형식이 있는 변수에 적용되지 않습니다.

Python 구문(S)

Python 구문. 이 필드에 데이터 분석을 위한 Python 스크립팅 구문을 입력하거나 붙여넣거나 사용자 정의할 수 있습니다. Python for Spark에 대한 자세한 정보는 Python for Spark 및 Python for Spark로 스크립팅의 내용을 참조하십시오.

② 확장 출력 노드 - 콘솔 출력 탭

콘솔 출력 탭에는 명령문 탭에서 R 스크립트 또는 Python for Spark 스크립트가 실행될 때 수신된 모든 출력이 포함됩니다. 예를 들어, R 스크립트를 사용하는 경우, **명령문** 탭의 **R 명령문** 필드의 R 스크립트가 실행될 때 R 콘솔에서 수신된 출력을 표시합니다. 이 출력에는 R 또는 Python 스크립트가 실행될 때 생성되는 R 또는 Python 오류 메시지 또는 경고가 포함됩니다. 출력은 주로 스크립트를 디버그하는 데 사용될 수 있습니다. 콘솔 출력 탭에는 **R 명령문** 또는 **Python 명령문** 필드의 스크립트도 포함됩니다.

확장 출력 스크립트가 실행될 때마다 R 콘솔 또는 Python for Spark에서 수신된 출력이 **콘솔 출력** 탭의 내용을 덮어씁니다. 출력은 편집할 수 없습니다.

③ 확장 출력 노드 - 출력 탭

출력 이름. 노드가 실행될 때 생성되는 출력의 이름을 지정합니다. **자동**을 선택하면 출력의 이름이 스크립트 유형에 따라 자동으로 "R Output" 또는 "Python Output"으로 설정됩니다. 선택적으로 **사용자 정의**를 선택하여 다른 이름을 지정할 수 있습니다.

화면으로 출력. 새 창에서 출력을 생성하고 표시하려면 이 옵션을 선택하십시오. 또한 출력이 출력 관리자에 추가됩니다.

파일로 출력. 출력을 파일에 저장하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 **출력 그래프** 및 **출력 파일** 단일 선택 단추를 사용할 수 있습니다.

그래프 출력. **파일로 출력**이 선택된 경우에만 사용 가능합니다. 확장 출력 노드를 실행하여 발생한 모든 그래프를 파일에 저장하려면 이 옵션을 선택하십시오. 생성된 출력에 대해 사용할 파일 이름을 **파일 이름** 필드에서 지정하십시오. 생략 기호(...)를 클릭하여 특정 파일 및 위치를 선택하십시오. **파일 유형** 드롭 다운 목록에서 파일 유형을 지정하십시오. 다음 파일 유형이 사용 가능합니다.

- 출력 오브젝트(.cou)
- HTML(.html)

출력 텍스트. **파일로 출력**이 선택된 경우에만 사용 가능합니다. 확장 출력 노드를 실행하여 발생한 모든 텍스트 출력을 파일에 저장하려면 이 옵션을 선택하십시오. 생성된 출력에 대해 사용할 파일 이름을 **파일 이름** 필드에서 지정하십시오. 생략 기호(...)를 클릭하여 특정 파일 및 위치를 지정하십시오. **파일 유형** 드롭 다운 목록에서 파일 유형을 지정하십시오. 다음 파일 유형이 사용 가능합니다.

- HTML(.html)
- 출력 오브젝트(.cou)
- 텍스트 문서(.txt)

④ 확장 출력 브라우저

확장 출력 노드 대화 상자의 **출력** 탭에서 **화면에 출력**을 선택하면 화면 출력이 출력 브라우저 창에 표시됩니다. 또한 출력이 출력 관리자에 추가됩니다. 출력 브라우저 창에는 출력을 인쇄 또는 저장하거나 다른 형식으로 내보낼 수 있는 메뉴 세트가 있습니다. **편집** 메뉴에는 **복사** 옵션만 있습니다. 확장 출력 노드의 출력 브라우저에는 두 개의 탭이 있습니다.

- 텍스트 출력 탭은 텍스트 출력을 표시합니다.
- 그래프 출력 탭은 그래프 및 도표를 표시합니다.

화면에 출력 대신 확장 출력 노드 대화 상자의 **출력** 탭에서 **파일에 출력**을 선택하면 확장 출력 노드가 성공적으로 실행될 때 출력 브라우저 창이 표시되지 않습니다.

가. 확장 출력 브라우저 - 텍스트 출력 탭

텍스트 출력 탭에는 확장 출력 노드의 **명령문** 탭에 있는 R 스크립트 또는 Python for Spark 스크립트를 실행할 때 생성되는 모든 텍스트 출력이 표시됩니다.

 **참고:** 확장 출력 스크립트를 실행하여 그 결과로 발생하는 R 또는 Python for Spark 오류 메시지 또는 경고 또한 항상 확장 출력 노드의 **콘솔 출력** 탭에 표시됩니다.

나. 확장 출력 브라우저 - 그래프 출력 탭

그래프 출력 탭에는 확장 출력 노드의 **명령문** 탭에 있는 R 스크립트 또는 Python for Spark 스크립트를 실행할 때 생성되는 모든 그래프 또는 차트가 표시됩니다. 예를 들어, R 스크립트에 R plot 함수에 대한 호출이 포함된 경우, 결과 그래프가 이 탭에 표시됩니다.

(16) KDE 노드

KDE(Kernel Density Estimation)²⁾는 효과적인 쿼리를 위해 볼 트리 또는 KD 트리 알고리즘을 사용하며, 자율 학습, 기능 엔지니어링 및 데이터 모델링을 따릅니다. KDE와 같은 이웃 기반의 접근법은 가장 인기 있고 유용한 밀도 추정 기법의 일부입니다. KDE는 모든 차원으로 수행할 수 있지만, 차원이 높을 경우 성능이 저하될 수 있습니다. SPSS® Modeler의 KDE 모델링 및 KDE 시뮬레이션 노드에는 KDE 라이브러리의 핵심 기능과 일반적으로 사용되는 매개변수가 표시됩니다. 이 노드는 Python으로 구현됩니다.²⁾

KDE 노드를 사용하려면 업스트림 유형 노드를 설정해야 합니다. KDE 노드는 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 입력 값을 읽습니다.

2) "User Guide." Kernel Density Estimation. Web. © 2007-2018, scikit-learn 개발자.

KDE 모델링 노드는 SPSS Modeler의 모델링 탭 및 Python 탭에서 사용 가능합니다. KDE 모델링 노드는 모델 너깃을 생성하며, 너깃의 스코어 값은 입력 데이터의 커널 밀도 값입니다.

KDE 시뮬레이션 노드는 출력 탭 및 Python 탭에서 사용 가능합니다. KDE 시뮬레이션 노드는 KDE 생성 소스 노드를 생성하며, 이 노드는 입력 데이터와 동일한 분포를 갖는 일부 레코드를 생성할 수 있습니다. KDE 생성 노드에는 노드가 생성할 레코드 수(기본값: 1)를 지정하고 난수 시드를 생성할 수 있는 설정 탭이 있습니다.

예제를 포함한 KDE에 대한 자세한 정보는 KDE 문서 (<http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation>)를 참조하십시오.³⁾

① KDE 모델링 노드 및 KDE 시뮬레이션 노드 필드

필드 탭은 분석에 사용되는 필드를 지정합니다.

사전 정의된 역할 사용: 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 입력 설정을 사용합니다.

사용자 정의 필드 할당 사용: 입력을 수동으로 할당하려면 이 옵션을 선택하십시오.

필드. 이 목록에서 화면의 오른쪽에 있는 입력 목록으로 수동으로 항목을 할당하려면 화살표 단추를 사용하십시오. 아이콘은 각 필드에 대한 유효한 측정 수준을 나타냅니다. 목록의 모든 필드를 선택하려면 모두 단추를 클릭하거나 개별 측정 수준 단추를 클릭하여 이 측정 수준의 모든 필드를 선택하십시오.

입력 군집에 대한 입력으로 하나 이상의 필드를 선택하십시오. KDE는 연속형 필드만 처리할 수 있습니다.

② KDE 노드 작성 옵션

작성 옵션 탭에서는 커널 밀도 매개변수 및 군집 레이블에 대한 **기본 옵션**과 **고급 옵션**(예: 공차, 리프 크기, 가로 먼저 접근법의 사용 여부)을 포함하여 KDE 노드에 대한 작성 옵션을 지정할 수 있습니다. 이러한 옵션에 대한 추가 정보는 다음 온라인 자원을 참조하십시오.

- Kernel Density Estimation Python API Parameter Reference⁴⁾
- Kernel Density Estimation User Guide⁵⁾

3) "User Guide." Kernel Density Estimation. Web. © 2007-2018, scikit-learn 개발자.

4) "API Reference." sklearn.neighbors.KernelDensity. Web. © 2007-2018, scikit-learn 개발자.

5) "User Guide." Kernel Density Estimation. Web. © 2007-2018, scikit-learn 개발자.

기본

대역폭. 커널 대역폭을 지정하십시오.

커널. 사용할 커널을 선택하십시오. KDE 모델링 노드에 대한 사용 가능 커널은 **가우스, Tophat, Epanechnikov, Eponential, 선형, 코사인**입니다. KDE 시뮬레이션 노드에 대한 사용 가능 커널은 **가우스** 또는 **Tophat**입니다. 이러한 사용 가능 커널에 대한 세부사항은 Kernel Density Estimation User Guide를 참조하십시오.

알고리즘. 사용할 트리 알고리즘에 대해 **자동, 볼 트리, KD 트리** 중 하나를 선택하십시오. 자세한 정보는 Ball Tree⁶⁾ 및 KD Tree를 참조하십시오.⁷⁾

메트릭. 거리 메트릭을 선택하십시오. 사용 가능 메트릭은 **유클리디안, Braycurtis, 체비셰프, 캔버라, 도시 블록, 다이스, 해밍, 무한, 자카드, L1, L2, 매칭, 맨해튼, P, Rogerstanimoto, Russellrao, Sokalmichener, Sokalsneath, Kulsinski, 민코스키**입니다. **민코스키**를 선택한 경우 **P** 값을 원하는 값으로 설정하십시오.

이 드롭 다운에서 사용 가능한 메트릭은 선택한 알고리즘에 따라 달라집니다. 또한 밀도 출력의 정규화는 유클리디안 거리 메트릭의 경우에만 정확합니다.

고급

절대 공차. 결과의 원하는 절대 공차를 지정하십시오. 일반적으로 공차가 클수록 실행 시간이 빨라집니다. 기본값은 **0.0**입니다.

상대 공차. 결과의 원하는 상대 공차를 지정하십시오. 일반적으로 공차가 클수록 실행 시간이 빨라집니다. 기본값은 **1E-8**입니다.

리프 크기. 기반 트리의 리프 크기를 지정하십시오. 기본값은 **40**입니다. 리프 크기를 변경하면 성능과 필요한 메모리에 상당한 영향을 미칠 수 있습니다. 볼 트리 및 KD 트리 알고리즘에 대한 자세한 정보는 Ball Tree 및 KD Tree를 참조하십시오.

가로 먼저. 가로 먼저 접근법을 사용하려면 **True**를 선택하고, 깊이 먼저 접근법을 사용하려면 **False**를 선택하십시오.

다음 표는 SPSS® Modeler KDE 노드 대화 상자의 설정과 Python KDE 라이브러리 매개변수 간의 관계를 보여줍니다.

6) "Ball Tree." Five balltree construction algorithms. © 1989, Omohundro, S.M., International Computer Science Institute Technical Report.

7) "K-D Tree." Multidimensional binary search trees used for associative searching. © 1975, Bentley, J.L., Communications of the ACM.

표 1. Python 라이브러리 모수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	KDE 매개변수
입력	inputs	
대역폭	bandwidth	bandwidth
커널	kernel	kernel
알고리즘	algorithm	algorithm
메트릭	metric	metric
P 값	pValue	pValue
절대 공차	atol	atol
상대 공차	rtol	Rtol
리프 크기	leafSize	leafSize
가로 먼저	breadthFirst	breadthFirst

③ KDE 모델링 노드 및 KDE 시뮬레이션 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

6) 내보내기 노드

(1) 내보내기 노드의 개요

내보내기 노드는 다양한 형식으로 데이터를 내보내 다른 소프트웨어 도구와 인터페이스로 접속하는 메커니즘을 제공합니다.

사용 가능한 내보내기 노드는 다음과 같습니다.



데이터베이스 내보내기 노드는 데이터를 ODBC 준수 관계형 데이터 소스에 기록합니다. ODBC 데이터 소스에 쓰기 위해 데이터 소스가 존재하고 사용자에게 쓰기 권한이 있어야 합니다.



플랫 파일 내보내기 노드는 데이터를 구분된 텍스트 파일로 출력합니다. 다른 분석 또는 스프레드시트 소프트웨어가 읽을 수 있는 데이터 내보내기에 유용합니다.



통계량 내보내기 노드는 IBM® SPSS® Statistics .sav 또는 .zsav 형식으로 데이터를 출력합니다. .sav 또는 .zsav 파일은 IBM SPSS Statistics Base 및 기타 제품에서 읽을 수 있습니다. 이것은 또한 IBM SPSS Modeler의 캐시 파일에 사용하는 형식입니다.



Data Collection 내보내기 노드는 Data Collection 시장 조사 소프트웨어에서 사용하는 형식으로 데이터를 출력합니다. 이 노드를 사용하려면 Data Collection 데이터 라이브러리가 설치되어야 합니다.



IBM Cognos 내보내기 노드는 Cognos 데이터베이스가 읽을 수 있는 형식으로 데이터를 내보냅니다.



IBM Cognos TM1 내보내기 노드는 Cognos TM1 데이터베이스가 읽을 수 있는 형식으로 데이터를 내보냅니다.



SAS 내보내기 노드는 SAS 또는 SAS 호환 가능한 소프트웨어 패키지로 읽어들이기 위해 데이터를 SAS 형식으로 출력합니다. SAS for Windows/OS2, SAS for UNIX 또는 SAS 버전 7/8의 세 가지 SAS 파일 형식이 사용 가능합니다.



Excel 내보내기 노드는 데이터를 Microsoft Excel .xlsx 파일 형식으로 출력합니다. (선택사항)노드가 실행될 때 Excel을 자동으로 시작하고 내보내진 파일을 열도록 선택할 수 있습니다.



XML 내보내기 노드는 데이터를 XML 형식의 파일로 출력합니다. 선택적으로 XML 소스 노드를 작성하여 내보내진 데이터를 다시 스트림으로 읽을 수 있습니다.



JSON 내보내기 노드에서는 데이터를 JSON 형식으로 출력합니다. 자세한 정보는 JSON 내보내기 노드의 내용을 참조하십시오.

(2) 데이터베이스 내보내기 노드

데이터베이스 노드를 사용하여 ODBC 준수 관계형 데이터 소스에 데이터를 쓸 수 있습니다. 여기에 대해서는 데이터베이스 소스 노드에 대한 설명을 참조하십시오. 자세한 정보는 데이터베이스 소스 노드 주제를 참조하십시오.

데이터베이스에 데이터를 쓰려면 다음과 같은 일반 단계를 사용하십시오.

1. ODBC 드라이버를 설치하고 원하는 데이터베이스에 데이터 소스를 구성하십시오.
2. 데이터베이스 노드 내보내기 탭에서 쓸 데이터 소스 및 테이블을 지정하십시오. 새 테이블을 작성하거나 데이터를 기존 테이블에 삽입할 수 있습니다.
3. 필요에 따라 추가 옵션을 지정하십시오.

이러한 단계에 대해서는 다음 몇 가지 주제에서 더 자세히 설명합니다.

① 데이터베이스 노드 내보내기 탭

참고: 내보낼 수 있는 일부 데이터베이스는 길이가 30자를 초과하는 열 이름을 테이블에서 지원하지 않을 수 있습니다. 테이블에 올바르지 않은 열 이름이 있다는 오류 메시지가 표시되면 30자 미만으로 해당 이름의 크기를 줄이십시오.

데이터 소스. 선택된 데이터 소스를 표시합니다. 이름을 입력하거나 드롭 다운 목록에서 이름을 선택하십시오. 목록에 원하는 데이터베이스가 표시되지 않으면 **새 데이터베이스 연결 추가**를 선택하고 데이터베이스 연결 대화 상자에서 데이터베이스를 찾으십시오. 자세한 정보는 데이터베이스 연결 추가의 내용을 참조하십시오.

테이블 이름. 데이터를 전송할 테이블의 이름을 입력하십시오. **테이블에 삽입** 옵션을 선택하는 경우에는 **선택** 단추를 클릭하여 데이터베이스에서 기존 테이블을 선택할 수 있습니다.

테이블 작성. 새 데이터베이스 테이블을 작성하거나 기존 데이터베이스 테이블을 겹쳐쓰려면 이 옵션을 선택하십시오.

테이블에 삽입. 기존 데이터베이스 테이블에서 새 행으로 데이터를 삽입하려면 이 옵션을 선택하십시오.

테이블 병합. (사용 가능한 경우) 선택된 데이터베이스 열을 해당 소스 데이터 필드의 값으로 업데이트하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 소스 데이터 필드를 데이터베이스 열에 매핑할 수 있는 대화 상자를 표시하는 **병합** 단추를 사용할 수 있습니다.

기존 테이블 삭제. 새 테이블 작성 시 동일한 이름의 기존 테이블을 삭제하려면 이 옵션을 선택하십시오.

기존 행 삭제. 테이블에 삽입 시 내보내기 전에 테이블에서 기존 행을 삭제하려면 이 옵션을 선택하십시오.

참고: 위 두 옵션 중 하나를 선택할 경우 노드를 실행할 때 **겹쳐쓰기 경고** 메시지가 수신됩니다. 경고를 표시하지 않으려면 사용자 옵션 대화 상자의 알림 탭에서 **노드가 데이터베이스 테이블을 겹쳐쓸 때 경고를 선택 취소**하십시오.

기본 문자열 크기. 업스트림 유형 노드에서 유형 없음으로 표시한 필드는 데이터베이스에 문자열 필드로 작성됩니다. 유형 없는 필드에 사용할 문자열의 크기를 지정하십시오.

스키마를 클릭하여 다양한 내보내기 옵션을 설정(이 기능을 지원하는 데이터베이스의 경우)하고 필드에 대해 SQL 데이터 유형을 설정하고 데이터베이스 인덱싱을 위해 기본 키를 지정할 수 있는 대화 상자를 여십시오. 자세한 정보는 데이터베이스 내보내기 스키마 옵션의 내용을 참조하십시오.

인덱스를 클릭하여 데이터베이스 성능을 향상시키기 위해 내보낸 테이블을 인덱싱하는 데 필요한 옵션을 지정하십시오. 자세한 정보는 데이터베이스 내보내기 인덱스 옵션의 내용을 참조하십시오.

고급을 클릭하여 벌크 로드 및 데이터베이스 커밋 옵션을 지정하십시오. 자세한 정보는 데이터베이스 내보내기 고급 옵션의 내용을 참조하십시오.

테이블 및 열 이름 따옴표로 묶기. CREATE TABLE문을 데이터베이스에 전송할 때 사용되는 옵션을 선택하십시오. 공백 또는 비표준 문자가 포함된 테이블 또는 열은 따옴표로 묶어야 합니다.

- **필요에 따라.** IBM® SPSS® Modeler가 개별적으로 따옴표가 필요한 시기를 자동으로 판별할 수 있게 하려면 선택하십시오.
- **항상.** 테이블 및 열 이름을 항상 따옴표로 묶으려면 선택하십시오.
- **사용 안 함.** 따옴표를 사용하지 않으려면 선택하십시오.

현재 데이터의 입력 노드 생성. 지정된 데이터 소스 및 테이블로 내보낸 대로 데이터에 대한 데이터베이스 소스 노드를 생성하려면 선택하십시오. 실행 시 이 노드는 스트림 캔버스에 추가됩니다.

② 데이터베이스 내보내기 병합 옵션

이 대화 상자에서는 소스 데이터의 필드를 목표 데이터베이스 테이블의 열에 맵핑할 수 있습니다. 소스 데이터 필드가 데이터베이스 열에 맵핑되는 경우에는 스트림이 실행될 때 열 값이 소스 데이터 값으로 바뀝니다. 맵핑되지 않은 소스 필드는 데이터베이스에서 변경되지 않고 유지됩니다.

맵 필드. 소스 데이터 필드와 데이터베이스 열 사이의 매핑을 지정하는 위치입니다. 데이터베이스의 열과 동일한 이름을 가진 소스 데이터 필드는 자동으로 매핑됩니다.

- **맵핑.** 단추 왼쪽의 필드 목록에서 선택된 소스 데이터 필드를 오른쪽 목록에서 선택된 데이터베이스 열에 매핑합니다. 한 번에 둘 이상의 필드를 매핑할 수 있지만 두 목록에서 선택된 항목 수는 동일해야 합니다.
- **맵핑 해제.** 하나 이상의 선택된 데이터베이스 열에 대한 매핑을 제거합니다. 이 단추는 대화 상자 오른쪽의 테이블에서 필드 또는 데이터베이스 열을 선택하면 활성화됩니다.
- **추가.** 단추 왼쪽의 필드 목록에서 선택된 하나 이상의 소스 데이터 필드를 매핑 준비가 된 오른쪽의 목록에 추가합니다. 이 단추는 왼쪽의 목록에서 필드를 선택했을 때 해당 이름을 가진 필드가 오른쪽의 목록에 없는 경우 활성화됩니다. 이 단추를 클릭하면 선택된 필드가 동일한 이름의 새 데이터베이스 열에 매핑됩니다. <NEW>라는 단어가 데이터베이스 열 이름 뒤에 표시되어 이 필드가 새 필드임을 표시합니다.

행 병합. 키 필드(예: 트랜잭션 ID)를 사용하여 키 필드에서 동일한 값을 가진 레코드를 병합합니다. 이는 데이터베이스 "일치 결합"과 동등합니다. 키 값은 기본 키의 값과 동일해야 합니다. 즉, 고유해야 하며 널값을 포함할 수 없습니다.

- **가능한 키.** 모든 입력 데이터 소스에서 발견된 모든 필드를 나열합니다. 이 목록에서 하나 이상의 필드를 선택한 후 화살표 단추를 사용하여 레코드 병합을 위해 키 필드로 추가하십시오. 해당 매핑된 데이터베이스 열을 가진 맵 필드를 모두 키로 사용할 수 있습니다(이름 뒤에 <NEW>가 표시된 새 데이터베이스 열로 추가된 필드는 사용할 수 없음).
- **병합을 위한 키.** 키 필드의 값을 기반으로 모든 입력 데이터 소스의 레코드를 병합하는 데 사용되는 모든 필드를 나열합니다. 목록에서 키를 제거하려면 하나의 키를 선택한 후 화살표 단추를 사용하여 가능한 키 목록에 반환하십시오. 둘 이상의 키 필드가 선택되면 아래의 옵션을 사용할 수 있습니다.
- **데이터베이스에 있는 레코드만 포함.** 부분 결합을 수행합니다. 레코드가 데이터베이스 및 스트림에 있는 경우에는 매핑된 필드가 업데이트됩니다.
- **데이터베이스에 레코드 추가.** 외부 결합을 수행합니다. 스트림의 모든 레코드가 병합되거나(동일한 레코드가 데이터베이스에 있는 경우) 추가됩니다(레코드가 아직 데이터베이스에 없는 경우).

새 데이터베이스 열에 소스 데이터 필드를 매핑하려면 다음을 수행하십시오.

1. 왼쪽 목록의 **맵 필드** 아래에서 소스 필드 이름을 클릭하십시오.
2. **추가** 단추를 클릭하여 매핑을 완료하십시오.

기존 데이터베이스 열에 소스 데이터 필드를 매핑하려면 다음을 수행하십시오.

1. 왼쪽 목록의 **맵 필드** 아래에서 소스 필드 이름을 클릭하십시오.
2. 오른쪽의 **데이터베이스 열** 아래에서 열 이름을 클릭하십시오.
3. **맵** 단추를 클릭하여 매핑을 완료하십시오.

맵핑을 제거하려면 다음을 수행하십시오.

1. 오른쪽 목록의 필드 아래에서 맵핑을 제거할 필드의 이름을 클릭하십시오.
2. **맵핑 해제** 단추를 클릭하십시오.

목록에서 필드를 선택 취소하려면 다음을 수행하십시오.

CTRL 키를 누른 상태로 필드 이름을 클릭하십시오.

③ 데이터베이스 내보내기 스키마 옵션

데이터베이스 내보내기 스키마 대화 상자에서는 데이터베이스 내보내기를 위한 옵션을 설정하고 (이 옵션을 지원하는 데이터베이스의 경우) 필드에 대한 SQL 데이터 유형을 설정하고 기본 키인 필드를 지정하고 내보낼 때 생성되는 CREATE TABLE문을 사용자 정의할 수 있습니다.

이 대화 상자에는 여러 파트가 있습니다.

- 맨 위의 섹션(표시된 경우)에는 이 옵션을 지원하는 데이터베이스에 내보내기 위한 옵션이 포함되어 있습니다. 해당 데이터베이스에 연결되지 않은 경우에는 이 섹션이 표시되지 않습니다.
- 가운데의 텍스트 필드에는 기본적으로 다음 형식을 따르는 CREATE TABLE 명령을 생성하는 데 사용되는 템플릿이 표시됩니다.

```
CREATE TABLE <table-name> <(table columns)>
```

- 아래쪽의 테이블에서는 각 필드에 대한 SQL 데이터 유형을 지정하고 아래에 설명된 대로 기본 키인 필드를 표시할 수 있습니다. 이 대화 상자는 테이블에서 지정하는 사항에 따라 <table-name> 및 <(table columns)> 모수의 값을 자동으로 생성합니다.

데이터베이스 내보내기 옵션 설정

이 섹션이 표시되는 경우에는 데이터베이스에 내보내는 데 필요한 다수의 설정을 지정할 수 있습니다. 이 기능을 지원하는 데이터베이스 유형은 다음과 같습니다.

- SQL Server Enterprise 및 Developer Edition. 자세한 정보는 SQL Server에 대한 옵션의 내용을 참조하십시오.
- Oracle Enterprise 또는 Personal Edition. 자세한 정보는 Oracle에 대한 옵션의 내용을 참조하십시오.

CREATE TABLE문 사용자 정의

이 대화 상자의 텍스트 필드 부분을 사용하여 CREATE TABLE문에 데이터베이스별 옵션을 추가할 수 있습니다.

1. **CREATE TABLE 명령 사용자 정의** 선택란을 선택하여 텍스트 창을 활성화하십시오.
2. 명령문에 데이터베이스별 옵션을 추가하십시오. 텍스트 <table-name> 및 (<table-columns>) 모수는 IBM® SPSS® Modeler에 의해 실제 테이블 이름 및 열 정의에 대해 대체되므로 이들 모수는 보존해야 합니다.

SQL 데이터 유형 설정

기본적으로 IBM SPSS Modeler를 사용하면 데이터베이스 서버가 SQL 데이터 유형을 자동으로 지정할 수 있습니다. 필드에 대한 자동 유형을 대체하려면 해당 필드에 해당하는 행을 찾은 후 스키마 테이블의 **유형 열**에 있는 드롭 다운 목록에서 원하는 유형을 선택하십시오. Shift+클릭을 사용하여 둘 이상의 행을 선택할 수 있습니다.

길이, 정밀도 또는 척도 인수(BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC 및 NUMBER)를 사용하는 유형의 경우 데이터베이스 서버에 자동 길이 지정을 허용하는 대신 길이를 지정해야 합니다. 예를 들어, 길이에 대해 상당한 값(예: VARCHAR(25))을 지정하면 사용자가 의도한 경우 IBM SPSS Modeler에서의 저장 유형이 겹쳐집니다. 자동 지정을 대체하려면 유형 드롭 다운 목록에서 **지정**을 선택하고 유형 정의를 원하는 SQL 유형 정의 명령문으로 바꾸십시오.

이를 수행하는 가장 쉬운 방법은 먼저 원하는 유형 정의에 가장 근접한 유형을 선택한 후 **지정**을 선택하여 해당 정의를 편집하는 것입니다. 예를 들어, SQL 데이터 유형을 VARCHAR(25)로 설정하려면 먼저 유형 드롭 다운 목록에서 유형을 **VARCHAR(길이)**로 설정한 후 **지정**을 선택하고 텍스트 길이를 값 25로 바꾸십시오.

기본 키

내보낸 테이블의 열 중 하나 이상이 모든 행에 고유 값 또는 값 조합을 가져야 하는 경우에는 적용되는 각 필드에 대해 **기본 키** 선택란을 선택하여 이를 표시할 수 있습니다. 대부분의 데이터베이스는 기본 키 제한조건을 무효화하는 방식으로 테이블을 수정할 수 없게 하며 이 제한을 적용하기 위해 기본 키에 대한 인덱스를 자동으로 작성합니다. (선택적으로 인덱스 대화 상자에서 기타 필드에 대한 인덱스를 작성할 수 있습니다. 자세한 정보는 데이터베이스 내보내기 인덱스 옵션의 내용을 참조하십시오.)

가. SQL Server에 대한 옵션

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **행.** 행 수준 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW);와 동등).
- **페이지.** 페이지 수준 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE);).

나. Oracle에 대한 옵션

Oracle 설정 - 기본 옵션

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **기본값.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS;). 이 케이스에서 이는 기본 옵션과 동일한 효과를 가집니다.
- **기본.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS BASIC;).

Oracle 설정 - 고급 옵션

압축 사용. 선택된 경우 압축을 사용하여 내보낼 테이블을 작성합니다.

압축. 압축의 수준을 선택하십시오.

- **기본값.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS;). 이 케이스에서 이는 기본 옵션과 동일한 효과를 가집니다.
- **기본.** 기본 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...) COMPRESS BASIC;).
- **OLTP.** OLTP 압축을 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE (...)COMPRESS FOR OLTP;).

- **쿼리 낮음/높음.** (Exadata 서버 전용) 쿼리에 대해 HCC(Hybrid Columnar Compression)를 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; 또는 CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH;). 쿼리에 대한 압축은 데이터 웨어하우징 환경에서 유용합니다. HIGH는 LOW보다 높은 압축 비율을 제공합니다.
- **아카이브 낮음/높음.** (Exadata 서버 전용) 아카이브에 대해 HCC(Hybrid Columnar Compression)를 사용으로 설정합니다(예: SQL의 CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; 또는 CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH;). 아카이브에 대한 압축은 장기간 저장될 데이터를 압축하는 경우에 유용합니다. HIGH는 LOW보다 높은 압축 비율을 제공합니다.

④ 데이터베이스 내보내기 인덱스 옵션

인덱스 대화 상자를 사용하면 IBM® SPSS® Modeler에서 내보낸 데이터베이스 테이블에서 인덱스를 작성할 수 있습니다. 필요에 따라 포함할 필드 세트를 지정하고 CREATE INDEX 명령을 사용자 정의할 수 있습니다.

이 대화 상자는 두 개의 파트로 구성됩니다.

- 위쪽 텍스트 필드에는 하나 이상의 CREATE INDEX 명령을 생성하는 데 사용할 수 있는 템플릿이 표시되며 기본적으로 형식은 다음과 같습니다.

```
CREATE INDEX <index-name> ON <table-name>
```

- 대화 상자 아래쪽의 테이블에서는 작성할 각각의 인덱스에 대한 사양을 추가할 수 있습니다. 각각의 인덱스에 대해 포함할 필드 또는 열 및 인덱스 이름을 지정하십시오. 이 대화 상자에서는 자동으로 <index-name> 및 <table-name> 매개변수의 값을 적절하게 생성합니다. 예를 들어, *empid* 및 *deptid* 필드의 단일 인덱스에 대해 생성된 SQL의 모양은 다음과 같습니다.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)
```

여러 행을 추가하여 여러 인덱스를 작성할 수 있습니다. 각각의 행에 대해 별도의 CREATE INDEX 명령이 생성됩니다.

CREATE INDEX 명령 사용자 정의

선택적으로 모든 인덱스 또는 특정 인덱스에 대해 CREATE INDEX 명령을 사용자 정의할 수 있습니다. 이를 통해 필요에 따라 특정 데이터베이스 요구사항 또는 옵션을 수용하고 모든 인덱스 또는 특정 인덱스에 사용자 정의를 적용할 수 있는 유연성이 제공됩니다.

- 위쪽 대화 상자에서 **CREATE INDEX 명령 사용자 정의**를 선택하여 이후에 추가된 모든 인덱스에 사용되는 템플릿을 수정하십시오. 테이블에 이미 추가된 인덱스에는 변경사항이 자동으로 적용되지 않습니다.
- 테이블에서 하나 이상의 행을 선택한 후 대화 상자 위쪽의 **선택된 인덱스 업데이트**를 클릭하여 선택된 모든 행에 현재 사용자 정의를 적용하십시오.
- 각각의 행에서 **사용자 정의** 선택란을 선택하여 해당 인덱스에 대한 명령 템플릿만 수정하십시오.

<index-name> 및 <table-name> 매개변수의 값은 테이블 사양을 기반으로 대화 상자에 의해 자동으로 생성되며 직접 편집할 수 없습니다.

BITMAP 키워드. Oracle 데이터베이스를 사용하는 경우에는 다음과 같이 표준 인덱스 대신 비트맵 인덱스를 작성하도록 템플릿을 사용자 정의할 수 있습니다.

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

비트맵 인덱스는 고유 값 수가 적은 열을 인덱싱하는 경우에 유용할 수 있습니다. 결과 SQL의 모양은 다음과 같습니다.

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

UNIQUE 키워드. 대부분의 데이터베이스는 CREATE INDEX 명령에서 UNIQUE 키워드를 지원합니다. 이 키워드는 기본 테이블에 대한 기본 키 제한조건과 비슷한 고유성 제한조건을 적용합니다.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

실제로 기본 키로 지정된 필드의 경우 이 사양은 필요하지 않습니다. 대부분의 데이터베이스는 CREATE TABLE 명령에서 기본 키 필드로 지정된 필드에 대해 자동으로 인덱스를 작성하므로 이 필드에서 명시적으로 인덱스를 작성하지 않아도 됩니다. 자세한 정보는 데이터베이스 내보내기 스키마 옵션의 내용을 참조하십시오.

FILLFACTOR 키워드. 인덱스에 대한 일부 실제 매개변수를 미세 조정할 수 있습니다. 예를 들어, SQL Server를 사용하면 테이블에 대해 향후 변경사항이 작성될 때 사용자가 유지보수 비용과 인덱스 크기(초기 작성 후)의 균형을 맞출 수 있습니다.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

기타 주석

- 지정된 이름을 가진 인덱스가 이미 존재하는 경우 인덱스 작성은 실패합니다. 모든 실패는 초

기에 경고로 처리되어 후속 인덱스가 작성된 다음 모든 인덱스가 시도된 후 메시지 로그에 오류로 다시 보고될 수 있게 합니다.

- 최상의 성능을 위해 데이터가 테이블에 로드된 후 인덱스를 작성해야 합니다. 인덱스는 하나 이상의 열을 포함하고 있어야 합니다.
- 노드를 실행하기 전에 메시지 로그에서 생성된 SQL을 미리 볼 수 있습니다.
- 데이터베이스에 작성된 임시 테이블의 경우(즉, 노드 캐싱이 사용으로 설정된 경우) 기본 키 및 인덱스를 지정하는 옵션을 사용할 수 없습니다. 하지만 시스템에서는 다운스트림 노드에서 데이터가 사용되는 방식에 따라 적절하게 임시 테이블에서 인덱스를 작성할 수 있습니다. 예를 들어, 캐싱된 데이터가 이후에 DEPT 열에 의해 결합되는 경우에는 이 열에서 캐싱된 테이블을 인덱싱하는 것이 합리적입니다.

인덱스 및 쿼리 최적화

일부 데이터베이스 관리 시스템에서는 데이터베이스 테이블이 작성되고 로드되고 인덱싱되고 난 후 최적화 프로그램이 새 테이블에서 쿼리 실행의 속도를 높이기 위해 인덱스를 이용하려면 먼저 추가적인 단계가 필요합니다. 예를 들어, Oracle에서 비용 기반 쿼리 최적화 프로그램은 쿼리 최적화에서 인덱스를 사용하려면 먼저 테이블을 분석해야 합니다. Oracle에 대한 내부 ODBC 특성 파일(사용자에게 표시되지 않음)에는 다음과 같이 이를 수행하는 옵션이 포함되어 있습니다.

```
# Defines SQL to be executed after a table and any associated indexes  
# have been created and populated  
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

기본 키와 인덱스 중 어느 것이 정의되는지 여부에 관계없이 Oracle에서 테이블이 작성될 때마다 이 단계가 실행됩니다. 필요한 경우 추가적인 데이터베이스에 대한 ODBC 특성 파일을 비슷한 방식으로 사용자 정의할 수 있습니다. 지원 부서에 문의하여 지원을 받으십시오.

⑤ 데이터베이스 내보내기 고급 옵션

데이터베이스 내보내기 노드 대화 상자에서 고급 단추를 클릭하면 데이터베이스에 결과 내보내기에 대한 기술 세부사항을 지정할 수 있는 새 대화 상자가 표시됩니다.

일괄 커밋 사용. 데이터베이스에 대한 행별 커밋을 끄려면 선택하십시오.

일괄처리 크기. 메모리로 커밋하기 전에 데이터베이스로 보낼 레코드 수를 지정합니다. 이 숫자를 낮추면 전송 속도는 느려지지만 데이터 무결성이 향상됩니다. 데이터베이스의 최적 성능을 위해 이 숫자를 미세 조정할 수 있습니다.

벌크 로드 사용. IBM® SPSS® Modeler에서 직접 데이터베이스에 데이터를 벌크 로드하는 방법을 지정합니다. 특정 시나리오에 적합한 벌크 로드 옵션을 선택하기 위해 일부 실험이 필요할 수 있습니다.

- **ODBC를 통해.** 일반적인 데이터베이스에 내보내기보다 효율적으로 다중 행 삽입을 실행하기 위해 ODBC API를 사용하려면 선택하십시오. 아래 옵션에서 행 방식 바인딩과 열 방식 바인딩 중에서 선택하십시오.
- **외부 로더를 통해.** 데이터베이스에 고유한 사용자 정의 벌크 로더 프로그램을 사용하려면 선택하십시오. 이 옵션을 선택하면 아래의 다양한 옵션이 활성화됩니다.

고급 ODBC 옵션. 이 옵션은 ODBC를 통제가 선택된 경우에만 사용할 수 있습니다. 이 기능은 일부 ODBC 드라이버에서 지원하지 않을 수 있습니다.

- **행 방식.** 데이터베이스에 데이터를 로드하기 위해 SQLBulkOperations 호출을 사용하려면 행 방식 바인딩을 선택하십시오. 행 방식 바인딩은 레코드별로 데이터를 삽입하는 매개변수화된 삽입을 사용하는 경우보다 일반적으로 속도가 향상됩니다.
- **열 방식.** 데이터베이스에 데이터를 로드하기 위해 열 방식 바인딩을 사용하려면 선택하십시오. 열 방식 바인딩은 매개변수화된 INSERT문에서 각각의 데이터베이스 열을 N 개 값의 배열에 바인딩하여 성능을 향상시킵니다. INSERT문을 한 번 실행하면 N 개의 행이 데이터베이스에 삽입됩니다. 이 방법은 성능을 상당히 향상시킬 수 있습니다.

외부 로더 옵션. 외부 로더를 통제가 지정되면 파일에 데이터 세트를 내보내고 해당 파일의 데이터를 데이터베이스에 로드하기 위해 사용자 정의 로더 프로그램을 지정 및 실행하는 데 필요한 다양한 옵션이 표시됩니다. IBM SPSS Modeler는 다수의 인기 있는 데이터베이스 시스템에 대한 외부 로더와 인터페이스로 접속할 수 있습니다. 여러 스크립트가 소프트웨어와 함께 포함되었으며 scripts 서브디렉토리 아래의 기술 문서와 함께 사용 가능합니다. 이 기능을 사용하려면 Python 2.7이 IBM SPSS Modeler 또는 IBM SPSS Modeler Server와 동일한 시스템에 설치되어 있어야 하며 python_exe_path 매개변수가 options.cfg 파일에서 설정되어 있어야 합니다. 자세한 정보는 벌크 로더 프로그래밍의 내용을 참조하십시오.

- **구분자 사용.** 내보낸 파일에서 사용해야 하는 구분 문자를 지정합니다. 탭으로 구분하려면 탭을 선택하고 공백으로 구분하려면 공백을 선택하십시오. 쉼표(,) 등의 기타 문자를 지정하려면 기타를 선택하십시오.
- **데이터 파일 지정.** 벌크 로드 수행 중에 작성된 데이터 파일에 사용할 경로를 입력하려면 선택하십시오. 기본적으로 서버의 temp 디렉토리에서 임시 파일이 작성됩니다.
- **로더 프로그램 지정.** 벌크 로드 프로그램을 지정하려면 선택하십시오. 기본적으로 소프트웨어는 IBM SPSS Modeler 설치의 scripts 서브디렉토리에서 지정된 데이터베이스에 대해 실행할 Python 스크립트를 검색합니다. 여러 스크립트가 소프트웨어와 함께 포함되었으며 scripts 서브디렉토리 아래의 기술 문서와 함께 사용 가능합니다.
- **로그 생성.** 지정된 디렉토리에 로그 파일을 생성하려면 선택하십시오. 로그 파일은 오류 정보를 포함하고 있으며 벌크 로드 조작이 실패하는 경우에 유용합니다.
- **테이블 크기 확인.** 테이블 크기 증가가 IBM SPSS Modeler에서 내보낸 행의 수와 일치하는지 확인하는 테이블 확인을 수행하려면 선택하십시오.
- **추가 로더 옵션.** 로더 프로그램에 대한 추가적인 인수를 지정합니다. 공백이 포함된 인수의 경우에는 큰따옴표를 사용하십시오.

큰따옴표는 백슬래시로 이스케이프하여 선택적 인수에 포함됩니다. 예를 들어, -comment "This is a ₩"comment₩"로 지정된 옵션은 -comment 플래그와 This is a "comment"로 렌더링되는 주석 자체를 모두 포함합니다.

단일 백슬래시는 또다른 백슬래시로 이스케이프하여 포함될 수 있습니다. 예를 들어, -specialdir "C:₩₩Test Scripts₩₩"로 지정된 옵션은 -specialdir 플래그와 C:₩Test Scripts₩로 렌더링된 디렉토리를 포함합니다.

⑥ 벌크 로더 프로그래밍

데이터베이스 내보내기 노드에는 고급 옵션 대화 상자에서 벌크 로드를 위한 옵션이 있습니다. 벌크 로더 프로그램을 사용하면 텍스트 파일에서 데이터베이스로 데이터를 로드할 수 있습니다.

벌크 로드 사용 - 외부 로더를 통해 옵션은 다음 세 가지 작업을 수행하도록 IBM® SPSS® Modeler를 구성합니다.

- 필요한 데이터베이스 테이블 작성.
- 텍스트 파일에 데이터 내보내기.
- 이 파일에서 데이터베이스 테이블로 데이터를 로드하기 위해 벌크 로더 프로그램 호출.

일반적으로 벌크 로더 프로그램은 데이터베이스 로드 유틸리티 자체(예: Oracle의 sqldr 유틸리티)가 아니지만 올바른 인수를 구성하는 작은 스크립트 또는 프로그램이 데이터베이스 특정 보조 파일(예: 제어 파일)을 작성한 후 데이터베이스 로드 유틸리티를 호출합니다. 다음 절의 정보는 기존 벌크 로더를 편집하는 데 유용합니다.

또는 벌크 로드를 위해 사용자의 프로그램을 작성할 수 있습니다. 자세한 정보는 벌크 로더 프로그램 개발의 내용을 참조하십시오. 표준 기술 지원 계약에는 이 사항이 포함되어 있지 않으며 지원이 필요한 경우 IBM 서비스 담당자에게 문의해야 합니다.

벌크 로드를 위한 스크립트

IBM SPSS Modeler에는 Python 스크립트를 사용하여 구현되는 여러 가지 다른 데이터베이스를 위한 여러 개의 벌크 로더 프로그램이 제공됩니다. **외부 로더를 통해** 옵션을 선택하여 데이터베이스 내보내기 노드가 포함된 스트림을 실행하는 경우 IBM SPSS Modeler는 ODBC를 통해 데이터베이스 테이블을 작성하고(필요한 경우) IBM SPSS Modeler Server를 실행 중인 호스트에서 임시 파일에 데이터를 내보낸 후 벌크 로드 스크립트를 호출합니다. 그런 다음, 이 스크립트는 DBMS 벤더에서 제공하는 유틸리티를 실행하여 임시 파일에서 데이터베이스로 데이터를 업로드합니다.

참고: IBM SPSS Modeler 설치에는 Python 런타임 해석기가 포함되지 않으므로 Python을 별도로 설치해야 합니다. 자세한 정보는 데이터베이스 내보내기 고급 옵션의 내용을 참조하십시오.

다음 표에 나열된 데이터베이스에 스크립트가 제공됩니다(IBM SPSS Modeler 설치 디렉토리의 `₩scripts` 폴더에).

표 1. 제공되는 벌크 로더 스크립트		
데이터베이스	스크립트 이름	추가정보
IBM Db2	db2_loader.py	자세한 정보는 IBM Db2 데이터베이스에 데이터 벌크 로드의 내용을 참조하십시오.
IBM Netezza	netezza_loader.py	자세한 정보는 IBM Netezza 데이터베이스에 데이터 벌크 로드의 내용을 참조하십시오.
Oracle	oracle_loader.py	자세한 정보는 Oracle 데이터베이스에 데이터 벌크 로드의 내용을 참조하십시오.
SQL Server	mssql_loader.py	자세한 정보는 SQL Server 데이터베이스에 데이터 벌크 로드의 내용을 참조하십시오.
Teradata	teradata_loader.py	자세한 정보는 Teradata 데이터베이스에 데이터 벌크 로드의 내용을 참조하십시오.

가. IBM Db2 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM® SPSS® Modeler에서 IBM Db2 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Db2 명령행 프로세서(CLP) 유틸리티가 설치되어 있는지 여부 확인

db2_loader.py 스크립트는 Db2 LOAD 명령을 호출합니다. 명령행 프로세서(UNIX의 db2, Windows의 db2cmd)가 db2_loader.py를 실행할 서버(일반적으로 IBM SPSS Modeler Server를 실행 중인 호스트)에 설치되어 있는지 확인하십시오.

로컬 데이터베이스 별명이 실제 데이터베이스 이름과 동일한지 확인하십시오.

Db2 로컬 데이터베이스 별명은 로컬 또는 원격 Db2 인스턴스에서 데이터베이스를 참조하기 위해 Db2 클라이언트 소프트웨어에서 사용하는 이름입니다. 로컬 데이터베이스 별명이 원격 데이터베이스의 이름과 다른 경우 추가 로더 옵션을 제공하십시오.

```
-alias <local_database_alias>
```

예를 들어, 원격 데이터베이스의 이름이 호스트 GALAXY에서 STARS로 지정되었지만 IBM SPSS Modeler Server를 실행 중인 호스트의 Db2 로컬 데이터베이스 별명이 STARS_GALAXY입니다. 추가 로더 옵션을 사용하십시오.

```
-alias STARS_GALAXY
```

비ASCII 문자 데이터 인코딩

ASCII 형식이 아닌 데이터를 벌크 로드하는 경우 db2_loader.py의 구성 섹션에 있는 코드 페이지 변수가 사용자의 시스템에서 올바르게 설정되었는지 확인하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

나. IBM Netezza 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM® SPSS® Modeler에서 IBM Netezza 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Netezza nzload 유틸리티가 설치되었는지 확인

netezza_loader.py 스크립트는 Netezza 유틸리티 *nzload*를 호출합니다. *netezza_loader.py*를 실행할 서버에 *nzload*가 설치되었으며 올바르게 구성되었는지 확인하십시오.

비ASCII 데이터 내보내기

내보내기에 ASCII 형식이 아닌 데이터가 포함된 경우 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 `-encoding UTF8`을 추가해야 할 수도 있습니다. 이 경우 비ASCII 데이터가 올바르게 업로드되었는지 확인해야 합니다.

날짜, 시간, 시간소인 형식 데이터

스트림 특성에서 데이터 형식을 **DD-MM-YYYY**로 설정하고 시간 형식을 **HH:MM:SS**로 설정하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

기존 테이블에 데이터를 삽입할 때 스트림 및 대상 테이블에 있는 다른 순서의 열

스트림에 있는 열의 순서가 대상 테이블에 있는 열과 다른 경우 데이터 값이 잘못된 열에 삽입됩니다. 필드 재정렬 노드를 사용하여 스트림에 있는 열의 순서가 대상 테이블에 있는 순서와 일치하는지 확인하십시오. 자세한 정보는 필드 다시 정렬 노드의 내용을 참조하십시오.

`nzload` 진행 상태 추적

로컬 모드에서 IBM SPSS Modeler를 실행하는 경우 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 `-sts`를 추가하여 `nzload` 유틸리티로 여는 명령 창에서 10000행마다 상태 메시지를 보십시오.

다. Oracle 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM® SPSS® Modeler에서 Oracle 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Oracle `sqlldr` 유틸리티가 설치되었는지 확인

`oracle_loader.py` 스크립트는 Oracle 유틸리티 `sqlldr`을 호출합니다. `sqlldr`이 Oracle 클라이언트에 자동으로 포함되지 않는다는 점을 참고하십시오. `oracle_loader.py`를 실행할 서버에 `sqlldr`이 설치되었는지 확인하십시오.

데이터베이스 SID 또는 서비스 이름 지정

비로컬 Oracle 서버에 데이터를 내보내거나 로컬 Oracle 서버에 여러 데이터베이스가 있는 경우 SID 또는 서비스 이름을 전달하기 위해 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 다음을 지정해야 합니다.

```
-database <SID>
```

`oracle_loader.py`에서 구성 섹션 편집

UNIX(및 선택적으로 Windows) 시스템에서 `oracle_loader.py` 스크립트의 처음에 있는 구성 섹션을 편집하십시오. 여기에서 `ORACLE_SID`, `NLS_LANG`, `TNS_ADMIN`, `ORACLE_HOME` 환경 변수의 값을 적절하게 지정하고 `sqlldr` 유틸리티의 전체 경로를 지정할 수 있습니다.

날짜, 시간, 시간소인 형식 데이터

스트림 특성에서 일반적으로 날짜 형식을 `YYYY-MM-DD`로 설정하고 시간 형식을 `HH:MM:SS`로 설정해야 합니다.

위와 다른 날짜 및 시간 형식을 사용해야 하는 경우 Oracle 문서를 참조하고 *oracle_loader.py* 스크립트 파일을 편집하십시오.

비ASCII 문자 데이터 인코딩

ASCII 형식이 아닌 데이터를 벌크 로드하는 경우 시스템에서 환경 변수 NLS_LANG이 올바르게 설정되었는지 확인해야 합니다. 이는 Oracle 로더 유틸리티 *sqlldr*에서 읽습니다. 예를 들어, Windows에서 Shift-JIS의 NLS_LANG에 올바른 값은 Japanese_Japan.JA16SJIS입니다. NLS_LANG에 대한 자세한 정보는 Oracle 문서를 확인하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

라. SQL Server 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM® SPSS® Modeler에서 SQL Server 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

SQL Server bcp.exe 유틸리티가 설치되었는지 확인

mssql_loader.py 스크립트는 SQL Server 유틸리티 *bcp.exe*를 호출합니다. *mssql_loader.py*를 실행할 서버에 *bcp.exe*가 설치되었는지 확인하십시오.

구분자로 공백 사용이 작동되지 않음

DB 내보내기 고급 옵션 대화 상자에서 구분자로 공백을 선택하지 마십시오.

테이블 크기 확인 옵션 권장

DB 내보내기 고급 옵션 대화 상자에서 **테이블 크기 확인** 옵션을 사용하도록 권장됩니다. 벌크 로드 프로세스의 실패는 항상 발견되지 않으며 이 옵션을 사용하면 올바른 수의 행이 로드되었는지 추가 확인을 수행합니다.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

완전한 SQL Server 이름 지정 인스턴스 지정

SPSS Modeler가 규정되지 않은 호스트 이름으로 인해 SQL Server에 액세스할 수 없는 경우가 있을 수 있으며 다음 오류를 표시합니다.

외부 벌크 로더를 실행하는 중 오류가 발생했습니다.

로그 파일이 자세한 정보를 제공할 수 있습니다.

이 오류를 정정하려면 **추가 로더 옵션** 필드에 큰따옴표를 포함하여 다음 문자열을 추가하십시오.

```
"-S mhreboot.spss.com\SQLSERVER"
```

마. Teradata 데이터베이스에 데이터 벌크 로드

다음 사항은 DB 내보내기 고급 옵션 대화 상자의 외부 로더 옵션을 사용하여 IBM® SPSS® Modeler에서 Teradata 데이터베이스로 벌크 로드하도록 구성하는 데 유용할 수 있습니다.

Teradata fastload 유틸리티가 설치되었는지 확인

teradata_loader.py 스크립트는 Teradata 유틸리티 *fastload*를 호출합니다. *teradata_loader.py*를 실행할 서버에 *fastload*가 설치되었으며 올바르게 구성되었는지 확인하십시오.

데이터를 비어 있는 테이블에만 벌크 로드할 수 있음

벌크 로드의 대상으로 비어 있는 테이블만을 사용할 수 있습니다. 대상 테이블에 벌크 로드 이전의 데이터가 있는 경우 작업이 실패합니다.

날짜, 시간, 시간소인 형식 데이터

스트림 특성에서 날짜 형식을 YYYY-MM-DD로 설정하고 시간 형식을 HH:MM:SS로 설정하십시오.

공백 문자열

공백 문자열을 널값으로 데이터베이스에 내보냅니다.

Teradata 프로세스 ID(tpid)

기본적으로 *fastload*는 tpid=dbc를 사용하여 Teradata 시스템에 데이터를 내보냅니다. 일반적으로 dbccop1을 Teradata 서버의 IP 주소와 연관시키는 HOSTS 파일의 항목이 있습니다. 다른 서버를 사용하려면 이 서버의 tpid를 전달하기 위해 DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 다음을 지정하십시오.

```
-tpid <id>
```

테이블 및 열 이름의 공백

테이블 또는 열 이름에 공백이 포함된 경우 벌크 로드 작업이 실패합니다. 가능한 경우 공백을 제거하여 테이블 또는 열 이름을 변경하십시오.

바. 벌크 로더 프로그램 개발

이 주제에서는 텍스트 파일에서 데이터베이스로 데이터를 로드하기 위해 IBM® SPSS® Modeler에서 실행할 수 있는 벌크 로더 프로그램을 개발하는 방법을 설명합니다. 표준 기술 지원 계약에는 이 사항이 포함되어 있지 않으며 지원이 필요한 경우 IBM 서비스 담당자에게 문의해야 합니다.

Python을 사용한 벌크 로더 프로그램 작성

기본적으로 IBM SPSS Modeler는 데이터베이스 유형에 기반하여 기본 벌크 로더 프로그램을 검색합니다. 표 1의 내용을 참조하십시오.

일괄처리 로더 프로그램을 개발하는 데 도움이 되는 test_loader.py 스크립트를 사용할 수 있습니다. 자세한 정보는 벌크 로더 프로그램 테스트의 내용을 참조하십시오.

벌크 로더 프로그램에 전달된 오브젝트

IBM SPSS Modeler는 벌크 로더 프로그램에 전달되는 두 개의 파일을 작성합니다.

- **데이터 파일.** 이 파일에는 텍스트 형식으로 로드할 데이터가 있습니다.
- **스키마 파일.** 이 파일은 열의 이름과 유형에 대해 설명하고 데이터 파일을 형식화하는 방법 (예: 필드 간 구분자로 사용되는 문자)에 대한 정보를 제공하는 XML 파일입니다.

또한 IBM SPSS Modeler는 벌크 로드 프로그램을 호출할 때 테이블 이름, 사용자 이름 및 비밀번호와 같은 기타 정보를 인수로 전달합니다.

참고: IBM SPSS Modeler에 성공적으로 완료되었음을 알리기 위해 벌크 로더 프로그램은 스키마 파일을 삭제해야 합니다.

벌크 로더 프로그램에 전달되는 인수

프로그램에 전달되는 인수는 다음 표와 같습니다.

표 1. 벌크 로더에 전달되는 인수	
인수	설명
schemafilename	스키마 파일의 경로입니다.
datafile	데이터 파일의 경로입니다.
servername	DBMS 서버의 이름이며 공백일 수 있습니다.
databasename	DBMS 서버에 있는 데이터베이스의 이름이며 공백일 수 있습니다.

인수	설명
username	데이터베이스에 로그인하는 데 사용하는 사용자 이름입니다.
password	데이터베이스에 로그인하는 데 사용하는 비밀번호입니다.
tablename	로드할 테이블의 이름입니다.
ownername	테이블 소유자의 이름입니다(스키마 이름이라고도 함).

DB 내보내기 고급 옵션 대화 상자의 **추가 로더 옵션** 필드에 지정된 옵션은 이러한 표준 인수 이후에 벌크 로더 프로그램에 전달됩니다.

데이터 파일의 형식

데이터는 텍스트 형식으로 데이터 파일에 기록되며 각 필드는 DB 내보내기 고급 옵션 대화 상자에 지정된 구분 문자로 구분됩니다. 다음은 탭 구분 데이터 파일이 표시되는 방식의 예입니다.

48	F	HIGH	NORMAL	0.692623	0.055369	drugA
15	M	NORMAL	HIGH	0.678247	0.040851	drugY
37	M	HIGH	NORMAL	0.538192	0.069780	drugA
35	F	HIGH	HIGH	0.635680	0.068481	drugA

파일은 IBM SPSS Modeler Server(또는 IBM SPSS Modeler Server에 연결되지 않은 경우 IBM SPSS Modeler)에서 사용하는 로컬 인코딩으로 작성됩니다. 일부 형식화는 IBM SPSS Modeler 스트림 설정을 통해 제어됩니다.

스키마 파일의 형식

스키마 파일은 데이터 파일에 대해 설명하는 XML 파일입니다. 다음은 이전 데이터 파일에 함께 제공되는 예입니다.

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
  <table delimiter="wt" commit_every="10000" date_format="YYYY-MM-DD"
time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C"
type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
  </table>
</DBSCHEMA>
```

다음 두 개의 표는 스키마 파일의 <table> 및 <column> 요소에 대한 속성을 나열합니다.

표 2. <table> 요소의 속성

속성	설명
delimiter	필드 구분 문자입니다(TAB이 \t로 표시됨).
commit_every	일괄처리 크기 간격입니다(DB 내보내기 고급 옵션 대화 상자에 있음).
date_format	날짜를 표시하는 데 사용되는 형식입니다.
time_format	시간을 표시하는 데 사용되는 형식입니다.
append_existing	로드할 테이블에 데이터가 이미 있으면 true이고 그렇지 않으면 false입니다.
delete_datafile	벌크 로드 프로그램이 로드 완료 시 데이터 파일을 삭제해야 하는 경우 true입니다.

표 3. <column> 요소의 속성

속성	설명
name	열 이름입니다.
encoded_name	데이터 파일과 동일한 인코딩으로 변환되고 일련의 2자리 16진수로 출력되는 열 이름입니다.
type	열의 데이터 유형이며 integer, real, char, time, date, datetime 중 하나입니다.
size	char 데이터 유형의 경우 문자 수로 나타내는 열의 최대 너비입니다.

사. 벌크 로더 프로그램 테스트

IBM® SPSS® Modeler 설치 디렉토리의 `\scripts` 폴더에 포함된 테스트 스크립트 `test_loader.py`를 사용하여 벌크 로드를 테스트할 수 있습니다. 이와 같이 테스트하면 IBM SPSS Modeler에서 사용하도록 벌크 로드 프로그램 또는 스크립트를 개발, 디버그 또는 문제점 해결할 때 유용합니다.

테스트 스크립트를 사용하려면 다음과 같이 계속하십시오.

1. `test_loader.py` 스크립트를 실행하여 스키마 및 데이터 파일을 `schema.xml` 및 `data.txt` 파일에 복사하고 Windows 일괄처리 파일(`test.bat`)을 작성하십시오.

2. *test.bat* 파일을 편집하여 테스트할 벌크 로더 프로그램 또는 스크립트를 선택하십시오.
3. 명령 셸에서 *test.bat*를 실행하여 선택한 벌크 로드 프로그램 또는 스크립트를 테스트하십시오.

참고: *test.bat*를 실행하면 데이터가 실제로 데이터베이스에 로드되지 않습니다.

(3) 플랫 파일 내보내기 노드

플랫 파일 내보내기 노드를 사용하면 데이터를 구분된 텍스트 파일로 쓸 수 있습니다. 이 방법은 다른 분석 또는 스프레드시트 소프트웨어가 읽을 수 있는 데이터 내보내기에 유용합니다.

데이터에 지리 공간적 정보가 포함되어 있으면 이를 플랫 파일로 내보낼 수 있으며 동일한 스트림 내에서 사용하기 위해 가변파일 소스 노드를 생성하는 경우에는 모든 저장 공간, 측정 및 지리 공간적 메타데이터가 새 소스 노드에서 세분화됩니다. 그러나 데이터를 내보낸 다음 이를 다른 스트림으로 가져오는 경우에는 새 소스 노드에서 지리 공간적 메타데이터를 설정하려면 몇 가지 추가 단계를 수행해야 합니다. 자세한 정보는 가변파일 노드의 내용을 참조하십시오.

참고: IBM® SPSS® Modeler에서 더 이상 캐시 파일에 대해 이전 캐시 형식을 사용하지 않으므로 파일을 해당 형식으로 쓸 수 없습니다. IBM SPSS Modeler 캐시 파일은 이제 IBM SPSS Statistics .sav 형식으로 저장되며 이 형식은 Statistics 내보내기 노드를 사용하여 작성할 수 있습니다. 자세한 정보는 통계량 내보내기 노드의 내용을 참조하십시오.

① 플랫 파일 내보내기 탭

파일 내보내기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일 위치를 찾아보십시오.

쓰기 모드. **겹쳐쓰기**가 선택되면 지정된 파일의 기존 데이터가 모두 겹쳐써집니다. **추가**가 선택되면 출력이 기존 파일의 끝에 추가되어 포함된 모든 데이터를 유지합니다.

- **필드 이름 포함.** 이 옵션이 선택되면 출력 파일의 첫 번째 행에 필드 이름이 기록됩니다. 이 옵션은 **겹쳐쓰기** 쓰기 모드에만 사용할 수 있습니다.

각 레코드 다음에 줄 바꾸기. 이 옵션이 선택된 경우에는 각각의 레코드가 출력 파일의 새로운 행에서 작성됩니다.

필드 구분 문자: 생성된 텍스트 파일의 필드 값 사이에 삽입할 문자를 지정합니다. 옵션은 **쉼표**, **탭**, **공백** 및 **기타**입니다. **기타**를 선택하는 경우에는 원하는 구분 문자를 텍스트 상자에 입력하십시오.

따옴표 기호. 기호 필드의 값에 사용할 따옴표의 유형을 지정합니다. 옵션은 없음(값을 따옴표로 묶지 않음), 작은따옴표('), 큰따옴표(") 및 기타입니다. 기타를 선택하는 경우에는 원하는 따옴표 문자를 텍스트 상자에 입력하십시오.

인코딩. 사용되는 텍스트 인코딩 방법을 지정합니다. 시스템 기본값, 스트림 기본값 또는 UTF-8 중에서 선택할 수 있습니다.

- 시스템 기본값은 Windows 제어판에 지정되어 있거나 분산 모드에서 실행 중인 경우 서버 컴퓨터에 지정되어 있습니다.
- 스트림 기본값은 스트림 특성 대화 상자에서 지정됩니다.

소수점 기호. 데이터에서 소수점 표시 방법을 지정합니다.

- **스트림 기본값.** 현재 스트림의 기본 설정에 의해 정의된 소수점 구분 문자가 사용됩니다. 이는 일반적으로 컴퓨터의 로케일 설정에 의해 정의된 소수점 구분 문자입니다.
- **마침표(.).** 마침표가 소수점 구분 문자로 사용됩니다.
- **쉼표(,).** 쉼표가 소수점 구분 문자로 사용됩니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 가변파일 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 가변파일 노드의 내용을 참조하십시오.

(4) 통계량 내보내기 노드

통계량 내보내기 노드를 사용하면 IBM® SPSS® Statistics .sav 형식으로 데이터를 내보낼 수 있습니다. IBM SPSS Statistics .sav 파일은 IBM SPSS Statistics 기본 및 기타 모듈에서 읽을 수 있습니다. 이 형식은 IBM SPSS Modeler 캐시 파일에도 사용되는 형식입니다.

IBM SPSS Statistics 변수 이름은 64자로 제한되어 있으며 특정 문자(공백, 달러 기호(\$), 대시(-) 등)를 포함할 수 없으므로 IBM SPSS Modeler 필드 이름을 IBM SPSS Statistics 변수 이름에 맵핑하면 오류가 발생할 수 있습니다. 이러한 제한에 맞게 조정하는 두 가지 방법이 있습니다.

- 필터 탭을 클릭하여 IBM SPSS Statistics 변수 이름에 요구 사항에 맞게 필드의 이름을 변경할 수 있습니다. 자세한 정보는 IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링 주제를 참조하십시오.
- IBM SPSS Modeler에서 필드 이름 및 레이블을 모두 내보내려면 선택하십시오.

참고: IBM SPSS Modeler는 유니코드 UTF-8 형식으로 .sav 파일을 작성합니다. IBM SPSS Statistics는 릴리스 16.0부터는 유니코드 UTF-8 형식의 파일만 지원합니다. 데이터 손상을 방지하기 위해 유니코드 인코딩으로 저장된 .sav 파일을 IBM SPSS Statistics 16.0 이전 버전에서 사용하지 마십시오. 자세한 정보는 IBM SPSS Statistics 도움말을 참조하십시오.

다중 응답 세트. 파일을 내보낼 때 스트림에서 정의한 모든 다중 응답 세트는 자동으로 유지됩니다. 필터 탭의 모든 노드에서 다중 응답 세트를 보고 편집할 수 있습니다. 자세한 정보는 다중 응답 세트 편집 주제를 참조하십시오.

① 통계량 내보내기 노드 - 내보내기 탭

파일 내보내기 파일 이름을 지정합니다. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일 위치를 찾아보십시오.

파일 유형 파일이 일반적인 *.sav* 또는 압축된 *.zsav* 형식으로 저장된 경우에 선택하십시오.

비밀번호로 파일 암호화 비밀번호로 파일을 보호하려면 이 선택란을 선택하십시오. 별도의 대화 상자에 **비밀번호**를 입력하고 확인하도록 프롬프트됩니다.

 **참고:** 비밀번호로 보호된 파일은 SPSS® Modeler 버전 16 이상 또는 SPSS Statistics 버전 21 이상에서만 열 수 있습니다.

필드 이름 내보내기 SPSS Modeler에서 SPSS Statistics *.sav* 또는 *.zsav* 파일로 내보낼 때 변수 이름 및 레이블을 처리하는 방법을 지정합니다.

- **이름 및 변수 레이블** SPSS Modeler 필드 이름 및 레이블을 모두 내보내려면 선택하십시오. 이름은 SPSS Statistics 변수이름으로 내보내는 반면 레이블은 SPSS Statistics 변수 레이블로 내보냅니다.
- **이름을 변수 레이블로 사용** SPSS Statistics에서 변수 레이블로 SPSS Modeler 필드 이름을 사용하려면 선택하십시오. SPSS Modeler를 사용하면 SPSS Statistics 변수 이름에서는 유효하지 않은 문자를 필드 이름에서는 허용합니다. 유효하지 않은 SPSS Statistics 이름이 작성될 가능성을 방지하려면 대신 **이름을 변수 레이블로 사용**을 선택하십시오. 또는 필터 탭을 사용하여 필드 이름을 조정하십시오.

애플리케이션 시작 SPSS Statistics가 사용자의 컴퓨터에 설치되어 있으면 이 옵션을 선택하여 저장된 데이터 파일에 대해 직접 애플리케이션을 호출할 수 있습니다. 애플리케이션 시작에 필요한 옵션이 헬퍼 애플리케이션 대화 상자에서 지정되어야 합니다. 자세한 정보는 헬퍼 애플리케이션의 내용을 참조하십시오. 외부 프로그램을 열지 않고 간단하게 SPSS Statistics *.sav* 또는 *.zsav* 파일을 작성하려면 이 옵션을 선택 취소하십시오.

 **참고:** SPSS Modeler 및 SPSS Statistics를 서버(분산) 모드에서 함께 실행할 때 데이터를 쓰고 SPSS Statistics 세션을 시작하면 SPSS Statistics 클라이언트가 자동으로 열려 활성 데이터 세트에서 읽는 데이터 세트를 표시합니다. SPSS Statistics 클라이언트가 시작된 후 여기에서 데이터 파일을 수동으로 열면 임시로 해결될 수 있습니다.

이 데이터의 가져오기 노드 생성 내보낸 데이터 파일을 읽을 통계량 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 통계량 파일 노드의 내용을 참조하십시오.

② IBM SPSS Statistics에 대한 필드 이름 변경 또는 필터링

IBM® SPSS® Modeler에서 IBM SPSS Statistics 등의 외부 애플리케이션으로 데이터를 내보내거나 배포하기 전에 필드 이름을 변경하거나 조정해야 하는 경우가 있습니다. 통계량 변환, 통계량 출력 및 통계량 내보내기 대화 상자에는 이 프로세스를 활용하는 데 필요한 필터 탭이 포함되어 있습니다.

필터 탭 기능에 대한 기본적인 설명은 다른 위치에서 설명합니다. 자세한 정보는 필터링 옵션 설정 주제를 참조하십시오.

IBM SPSS Statistics 이름 지정 규칙을 준수하도록 필드 이름을 조정하려면 다음을 수행하십시오.

1. 필터 탭에서 필터 옵션 메뉴 도구 모음 단추(도구 모음 중 첫 번째 도구)를 클릭하십시오.
2. IBM SPSS Statistics에 대한 이름 변경을 선택하십시오.
3. IBM SPSS Statistics에 대한 이름 변경 대화 상자에서 파일 이름의 유효하지 않은 문자를 **해시(#)** 문자 또는 **밑줄(_)**로 바꿀 수 있습니다.

다중 응답 세트 이름 변경. 통계량 파일 소스 노드를 사용하여 IBM SPSS Modeler로 가져올 수 있는 다중 응답 세트의 이름을 조정하려면 이 옵션을 선택하십시오. 설문조사 응답과 같이 각 케이스에 대해 둘 이상의 값을 가질 수 있는 데이터를 기록하는 데 사용됩니다.

(5) Data Collection 내보내기 노드

Data Collection 내보내기 노드는 Data Collection 데이터 모델을 기반으로 Data Collection 시장 조사 소프트웨어에서 사용하는 형식으로 데이터를 저장합니다. 이 형식은 케이스 데이터(설문조사 중에 수집된 질문에 대한 실제 응답)를 케이스 데이터의 수집 및 구성 방식에 대해 설명하는 메타데이터와 구별합니다. 메타데이터는 케이스 데이터의 구조 정의, 질문 텍스트, 변수 이름 및 설명, 다중 응답 세트, 다양한 텍스트의 변환 등의 정보로 구성됩니다. 자세한 정보는 Data Collection 노드의 내용을 참조하십시오.

메타데이터 파일. 내보낸 메타데이터가 저장될 질문지 정의 파일(.*mda*)의 이름을 지정합니다. 기본 질문지는 필드 유형 정보를 기반으로 작성됩니다. 예를 들어, 명목(세트) 필드는 각각의 정의된 값에 대해 별도의 선택란 및 질문 텍스트로 사용되는 필드 설명이 포함된 단일 질문으로 표시될 수 있습니다.

메타데이터 병합. 메타데이터가 기존 버전을 겹쳐쓸지 아니면 기존 메타데이터와 병합될지를 지정합니다. 병합 옵션이 선택되면 스트림이 실행될 때마다 새 버전이 작성됩니다. 이를 통해 질문지가 변경될 때 질문지의 버전을 추적할 수 있습니다. 각각의 버전은 특정 케이스 데이터 세트를 수집하는 데 사용되는 메타데이터의 스냅샷으로 간주될 수 있습니다.

시스템 변수 사용. 시스템 변수가 내보낸 .mdd 파일에 포함되는지 여부를 지정합니다. 여기에는 *Respondent.Serial*, *Respondent.Origin*, *DataCollection.StartTime* 등의 변수가 포함됩니다.

케이스 데이터 설정. 케이스 데이터를 내보내는 IBM® SPSS® Statistics 데이터(.sav) 파일을 지정합니다. 변수 및 값 이름에 대한 모든 제한사항이 여기서 적용되므로 예를 들어, 필터 탭으로 전환한 후 필터 옵션 메뉴의 "IBM SPSS Statistics에 대해 이름 바꾸기" 옵션을 사용하여 필드 이름에서 유효하지 않은 문자를 정정해야 할 수 있습니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 Data Collection 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오.

다중 응답 세트. 파일을 내보낼 때 스트림에서 정의한 모든 다중 응답 세트는 자동으로 유지됩니다. 필터 탭의 모든 노드에서 다중 응답 세트를 보고 편집할 수 있습니다. 자세한 정보는 다중 응답 세트 편집 주제를 참조하십시오.

(6) IBM Cognos 내보내기 노드

IBM Cognos 내보내기 노드에서는 IBM® SPSS® Modeler 스트림의 데이터를 UTF-8 형식으로 Cognos Analytics로 내보낼 수 있습니다. 이러한 방식으로 Cognos는 IBM SPSS Modeler에서 변환된 데이터 또는 스코어링된 데이터를 사용할 수 있습니다. 예를 들어, Cognos Report Studio를 사용하여 예측 및 신뢰도를 포함한 내보낸 데이터를 기반으로 보고서를 작성할 수 있습니다. 그런 다음 보고서를 Cognos 서버에 저장하고 Cognos 사용자에게 배포할 수 있습니다.

 **참고:** 관계형 데이터만 내보낼 수 있으며 OLAP 데이터는 내보낼 수 없습니다.

Cognos로 데이터를 내보내려면 다음을 지정해야 합니다.

- Cognos 연결 - Cognos Analytics 서버에 대한 연결
- ODBC 연결 - Cognos 서버가 사용하는 Cognos 데이터 서버에 대한 연결

Cognos 연결 내에서 사용할 Cognos 데이터 소스를 지정합니다. 이 데이터 소스는 ODBC 데이터 소스와 동일한 로그인을 사용해야 합니다.

실제 스트림 데이터는 데이터 서버로 내보내고 패키지 메타데이터는 Cognos 서버로 내보내십시오.

다른 내보내기 노드에서와 마찬가지로, 노드 대화 상자의 출판 탭을 사용하여 IBM SPSS Modeler Solution Publisher를 통해 배포할 스트림을 출판할 수도 있습니다.

 **참고:** Cognos 소스 노드는 Cognos CQM 패키지만 지원합니다. DQM 패키지는 지원되지 않습니다.

① Cognos 연결

여기서는 내보내기를 위해 사용할 Cognos Analytics 서버에 대한 연결을 지정합니다. 이 프로 시저에는 Cognos 서버의 새 패키지로 메타데이터를 내보내는 것이 포함되지만 스트림 데이터는 Cognos 데이터 서버로 내보냅니다.

연결. 편집 단추를 클릭하여 데이터를 내보낼 Cognos 서버의 기타 세부사항 및 URL을 정의할 수 있는 대화 상자를 표시하십시오. IBM® SPSS® Modeler를 통해 이미 Cognos 서버에 로그인한 경우에는 현재 연결의 세부사항도 편집할 수 있습니다. 자세한 정보는 Cognos 연결의 내용을 참조하십시오.

데이터 소스. 데이터를 내보내는 Cognos 데이터 소스(일반적으로 데이터베이스)의 이름입니다. 드롭 다운 목록에는 현재 연결에서 액세스할 수 있는 모든 Cognos 데이터 소스가 표시됩니다. **새로 고치기** 단추를 클릭하여 목록을 업데이트하십시오.

폴더. 내보내기 패키지를 작성할 Cognos 서버의 폴더 이름과 경로입니다.

패키지 이름. 내보낸 메타데이터를 포함할 지정된 폴더의 패키지 이름입니다. 이는 단일 쿼리 제목을 가진 새 패키지여야 하며 기존 패키지에 내보낼 수 없습니다.

모드 내보내기 수행 방법을 지정합니다.

- **지금 패키지 게시.** (기본값) 실행을 클릭하는 즉시 내보내기 작업을 수행합니다.
- **조치 스크립트 내보내기.** 예를 들어, Framework Manager를 사용하여 나중에 실행할 수 있는 XML 스크립트를 작성하여 내보내기를 수행합니다. **파일** 필드에서 스크립트의 경로 및 파일 이름을 입력하거나 **편집** 단추를 사용하여 스크립트 파일의 이름 및 위치를 지정하십시오.

현재 데이터의 입력 노드 생성. 지정된 데이터 소스 및 테이블로 내보낸 대로 데이터에 대한 소스 노드를 생성하려면 선택하십시오. **실행**을 클릭하면 이 노드가 스트림 캔버스에 추가됩니다.

② ODBC 연결

여기서는 스트림 데이터를 내보낼 Cognos 데이터 서버(즉, 데이터베이스)에 대한 연결을 지정합니다.

참고: 여기서 지정하는 데이터 소스가 **Cognos 연결** 패널에 지정된 것과 동일한 데이터 소스를 가리키는지 확인해야 합니다. Cognos 연결 데이터 소스가 ODBC 데이터 소스와 동일한 로그인을 사용하는지도 확인해야 합니다.

데이터 소스. 선택된 데이터 소스를 표시합니다. 이름을 입력하거나 드롭 다운 목록에서 이름을 선택하십시오. 목록에 원하는 데이터베이스가 표시되지 않으면 **새 데이터베이스 연결 추가**를 선택하고 데이터베이스 연결 대화 상자에서 데이터베이스를 찾으십시오. 자세한 정보는 데이터베이스 연결 추가의 내용을 참조하십시오.

테이블 이름. 데이터를 전송할 테이블의 이름을 입력하십시오. **테이블에 삽입** 옵션을 선택하는 경우에는 **선택** 단추를 클릭하여 데이터베이스에서 기존 테이블을 선택할 수 있습니다.

테이블 작성. 새 데이터베이스 테이블을 작성하거나 기존 데이터베이스 테이블을 겹쳐쓰려면 이 옵션을 선택하십시오.

테이블에 삽입. 기존 데이터베이스 테이블에서 새 행으로 데이터를 삽입하려면 이 옵션을 선택하십시오.

테이블 병합. (사용 가능한 경우) 선택된 데이터베이스 열을 해당 소스 데이터 필드의 값으로 업데이트하려면 이 옵션을 선택하십시오. 이 옵션을 선택하면 소스 데이터 필드를 데이터베이스 열에 매핑할 수 있는 대화 상자를 표시하는 **병합** 단추를 사용할 수 있습니다.

기존 테이블 삭제. 새 테이블 작성 시 동일한 이름의 기존 테이블을 삭제하려면 이 옵션을 선택하십시오.

기존 행 삭제. 테이블에 삽입 시 내보내기 전에 테이블에서 기존 행을 삭제하려면 이 옵션을 선택하십시오.

 **참고:** 위 두 옵션 중 하나를 선택할 경우 노드를 실행할 때 **겹쳐쓰기 경고** 메시지가 수신됩니다. 경고를 표시하지 않으려면 사용자 옵션 대화 상자의 알림 탭에서 **노드가 데이터베이스 테이블을 겹쳐쓸 때 경고를 선택 취소**하십시오.

기본 문자열 크기. 업스트림 유형 노드에서 유형 없음으로 표시한 필드는 데이터베이스에 문자열 필드로 작성됩니다. 유형 없는 필드에 사용할 문자열의 크기를 지정하십시오.

스키마를 클릭하여 다양한 내보내기 옵션을 설정(이 기능을 지원하는 데이터베이스의 경우)하고 필드에 대해 SQL 데이터 유형을 설정하고 데이터베이스 인덱싱을 위해 기본 키를 지정할 수 있는 대화 상자를 여십시오. 자세한 정보는 데이터베이스 내보내기 스키마 옵션의 내용을 참조하십시오.

인덱스를 클릭하여 데이터베이스 성능을 향상시키기 위해 내보낸 테이블을 인덱싱하는 데 필요한 옵션을 지정하십시오. 자세한 정보는 데이터베이스 내보내기 인덱스 옵션의 내용을 참조하십시오.

고급을 클릭하여 벌크 로드 및 데이터베이스 커밋 옵션을 지정하십시오. 자세한 정보는 데이터베이스 내보내기 고급 옵션의 내용을 참조하십시오.

테이블 및 열 이름 따옴표로 묶기. CREATE TABLE문을 데이터베이스에 전송할 때 사용되는 옵션을 선택하십시오. 공백 또는 비표준 문자가 포함된 테이블 또는 열은 따옴표로 묶어야 합니다.

- 필요에 따라. IBM® SPSS® Modeler가 개별적으로 따옴표가 필요한 시기를 자동으로 판별할 수 있게 하려면 선택하십시오.
- 항상. 테이블 및 열 이름을 항상 따옴표로 묶으려면 선택하십시오.
- 사용 안 함. 따옴표를 사용하지 않으려면 선택하십시오.

현재 데이터의 입력 노드 생성: 지정된 데이터 소스 및 테이블로 내보낸 대로 데이터에 대한 소스 노드를 생성하려면 선택하십시오. 실행을 클릭하면 이 노드가 스트림 캔버스에 추가됩니다.

(7) IBM Cognos TM1 내보내기 노드

IBM Cognos 내보내기 노드에서는 SPSS® Modeler 스트림의 데이터를 Cognos TM1로 내보낼 수 있습니다. 이러한 방식으로 Cognos Analytics는 SPSS Modeler에서 변환된 데이터 또는 스코어링된 데이터를 사용할 수 있습니다.

참고: 축도만 내보낼 수 있습니다(컨텍스트 차원 데이터는 내보낼 수 없음). 또는 큐브에 새 요소를 추가할 수 있습니다.

데이터를 Cognos Analytics로 내보내려면 다음을 지정해야 합니다.

- Cognos TM1 서버로의 연결
- 데이터를 내보내는 큐브
- SPSS 데이터 이름에서 동등한 TM1 차원 및 축도로의 맵핑

참고: TM1 사용자에게는 큐브 쓰기 권한, 차원 읽기 권한 및 차원 요소 쓰기 권한이 필요합니다. 또한, IBM Cognos TM1 10.2 수정팩 3 이상이 있어야 SPSS Modeler에서 Cognos TM1 데이터를 가져오고 내보낼 수 있습니다. 이전 버전을 기반으로 하는 기존 스트림은 여전히 작동합니다.

이 노드에는 관리자 신임 정보가 필요하지 않습니다. 이전 레거시 17.1 이전의 TM1 노드를 여전히 사용하는 경우 관리자 신임 정보가 필요합니다.

SPSS Modeler는 IntegratedSecurityMode 1, 4, 5를 통한 Cognos TM1 작업만 지원합니다.

다른 내보내기 노드에서와 마찬가지로, 노드 대화 상자의 출판 탭을 사용하여 IBM® SPSS Modeler Solution Publisher를 통해 배포할 스트림을 출판할 수도 있습니다.

참고: SPSS Modeler에서 TM1 소스 또는 내보내기 노드를 사용하려면 먼저 tm1s.cfg 파일에서 일부 설정을 유효화해야 합니다. 이 파일은 TM1 서버의 루트 디렉토리에 있는 TM1 서버 구성 파일입니다.

- HTTPPortNumber - 유효한 포트 번호를 설정합니다. 일반적으로 1 - 65535입니다. 이 번호는 나중에 노드의 연결에 지정한 포트 번호가 아닙니다. 이 포트는 기본적으로 사용되지 않는 TM1에서 사용하는 내부 포트입니다. 필요하다면 TM1 관리자에게 문의하여 이 포트의 올바른 설정을 확인하십시오.
- UseSSL - 참으로 설정하면 HTTPS가 전송 프로토콜로 사용됩니다. 이 경우 TM1 인증을 SPSS Modeler Server JRE로 가져와야 합니다.

① 데이터를 내보낼 IBM Cognos TM1 큐브에 연결

데이터를 IBM Cognos TM1 데이터베이스로 내보내려면 IBM Cognos TM1 대화 상자의 **연결** 탭에서 서버 연결 세부사항을 지정하고 연관된 큐브 및 데이터 세부사항을 선택하십시오.

참고: TM1에 데이터를 내보낼 때 실제 "널" 값만 삭제됩니다. 영(0) 값은 유효한 값으로 내보내집니다. 또한 맵핑 탭에서는 저장 유형이 문자열인 필드만 차원으로 맵핑할 수 있습니다. TM1로 내보내기 전, IBM® SPSS® Modeler 클라이언트를 사용하여 문자열이 아닌 데이터 유형을 문자열로 변환해야 합니다.

연결 유형, 관리 서버 또는 TM1 서버를 선택하십시오. 관리 서버는 Planning Analytics on Cloud에서 제거되었으므로, 이전 관리 서버에 연결되는 이전 스트림이 있는 경우 해당 스트림이 Planning Analytics on Cloud를 대신 가리키도록 수정할 수 있습니다. 여기서 **관리 서버**를 선택할 경우 서버 URL(REST API의 **호스트** 이름)와 서버 이름을 입력해야 합니다. **TM1 서버**를 선택할 경우 다음 절로 이동하십시오.

TM1 서버 URL. 연결할 TM1 서버가 설치된 관리 호스트의 URL을 입력하십시오. 관리 호스트는 모든 TM1 서버에 대한 단일 URL로 정의됩니다. 이 URL에서, 사용하는 환경에 설치되어 실행 중인 모든 IBM Cognos TM1 서버를 검색하고 액세스할 수 있습니다. 로그인을 클릭하십시오. 이전에 이 서버에 연결한 적이 없는 경우, **사용자 이름** 및 **비밀번호** 입력을 요구하는 프롬프트가 표시됩니다. 또는 이전에 입력하여 **저장된 신임 정보**로 저장한 로그인 세부사항을 검색할 수 있습니다.

내보낼 TM1 큐브 선택 데이터를 내보낼 수 있는 TM1 서버 내의 큐브 이름을 표시합니다.

내보낼 데이터를 선택하려면, 큐브를 선택하고 오른쪽 화살표를 클릭하여 큐브를 **큐브로 내보내기** 필드로 이동시키십시오. 큐브를 선택했으면 **맵핑** 탭을 사용하여 TM1 차원 및 축도를 관련 SPSS 필드 또는 고정값(**선택** 조작)에 맵핑하십시오.

② 내보낼 IBM Cognos TM1 데이터 맵핑

TM1 관리 호스트 및 연관된 TM1 서버와 큐브를 선택한 후, IBM Cognos TM1 내보내기 대화 상자의 맵핑 탭을 사용하여 TM1 차원 및 측도를 SPSS 필드에 맵핑하거나 TM1 차원을 고정값으로 설정하십시오.

 **참고:** 저장 유형이 문자열인 필드만 차원으로 맵핑할 수 있습니다. TM1로 내보내기 전, IBM® SPSS® Modeler 클라이언트를 사용하여 문자열이 아닌 데이터 유형을 문자열로 변환해야 합니다.

필드 내보내기에 사용할 수 있는 SPSS 데이터 파일의 데이터 필드 이름을 나열합니다.

TM1 차원 연결 탭에서 선택된 TM1 큐브를 해당 정규 차원, 측도 차원 및 선택된 측도 차원의 요소와 함께 표시합니다. SPSS 데이터 필드로 맵핑하려면 TM1 차원 또는 측도의 이름을 선택하십시오.

맵핑 탭에서는 다음 옵션을 사용할 수 있습니다.

측도 차원 선택 선택된 큐브의 차원 목록에서 측도 차원이 될 차원을 선택하십시오.

측도 차원을 제외한 차원을 선택하고 **선택**을 클릭하면 선택된 차원의 리프 요소를 표시하는 대화 상자가 표시됩니다. 리프 요소만 선택할 수 있습니다. 선택된 요소는 **S**로 레이블이 지정됩니다.

맵핑 선택된 SPSS 데이터 필드를 선택된 TM1 차원 또는 측도(정규 차원, 측도 차원의 특정 측도 또는 요소)로 맵핑합니다. 맵핑된 필드는 **M**으로 레이블 지정됩니다.

맵핑 해제 선택된 TM1 차원 또는 측도에서 선택된 SPSS 데이터 필드를 맵핑 해제합니다. 한 번에 하나의 맵핑만 맵핑 해제할 수 있습니다. 맵핑 해제된 SPSS 데이터 필드는 다시 왼쪽 열로 이동합니다.

새로 작성 TM1 측도 차원에서 측도를 새로 작성합니다. 새 **TM1 측도 이름**을 입력하는 대화 상자가 표시됩니다. 이 옵션은 측도 차원에만 사용할 수 있고 정규 차원에는 사용할 수 없습니다.

TM1에 대한 자세한 정보는 IBM Cognos TM1 문서(http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html)를 참조하십시오.

(8) SAS 내보내기 노드

이 기능은 SPSS® Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

SAS 내보내기 노드를 사용하면 SAS 또는 SAS 호환 가능한 소프트웨어 패키지로 읽어들이기 위해 데이터를 SAS 형식으로 쓸 수 있습니다. SAS for Windows/OS2, SAS for UNIX 또는 SAS의 세 가지 SAS 파일 형식이 사용 가능합니다.

① SAS 내보내기 노드 내보내기 탭

이 기능은 SPSS® Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

파일 내보내기. 파일의 이름을 지정합니다. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일 위치를 찾아보십시오.

내보내기. 파일 내보내기 형식을 지정합니다. 옵션은 **SAS for Windows/OS2**, **SAS for UNIX** 또는 **SAS 버전 7/8/9**입니다.

필드 이름 내보내기. SAS와 함께 사용하기 위해 IBM® SPSS Modeler에서 필드 이름 및 레이블을 내보내려면 이 옵션을 선택하십시오.

- **이름 및 변수 레이블.** IBM SPSS Modeler 필드 이름 및 레이블을 모두 내보내려면 선택하십시오. 이름은 SAS 변수이름으로 내보내는 반면 레이블은 SAS 변수 레이블로 내보냅니다.
- **이름을 변수 레이블로 사용.** SAS에서 변수 레이블로 IBM SPSS Modeler 필드 이름을 사용하려면 선택하십시오. IBM SPSS Modeler를 사용하면 SAS 변수 이름에서는 유효하지 않은 문자를 필드 이름에서는 허용합니다. 유효하지 않은 SAS 이름이 작성될 가능성을 방지하려면 **대신 이름 및 변수 레이블**을 선택하십시오.

현재 데이터의 입력 노드 생성: 내보낸 데이터 파일을 읽을 SAS 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 SAS 소스 노드 주제를 참조하십시오.

 **참고:** 허용되는 최대 문자열 길이는 255바이트입니다. 문자열이 255바이트를 넘으면 내보낼 때 잘립니다.

(9) Excel 내보내기 노드

Excel 내보내기 노드는 Microsoft Excel .xlsx 형식으로 데이터를 출력합니다. 선택적으로 자동으로 Excel을 시작한 후 노드가 실행될 때 내보낸 파일을 열도록 선택할 수 있습니다.

① Excel 노드 내보내기 탭

파일 이름. 파일 이름을 입력하거나 파일 선택기 단추를 클릭하여 파일의 위치로 이동하십시오. 기본 파일 이름은 `excelxp.xlsx`입니다.

파일 유형. Excel `.xlsx` 파일 형식이 지원됩니다.

새 파일 작성. 새 Excel 파일을 작성합니다.

기존 파일에 삽입. 콘텐츠는 **셀에서 시작** 필드에 의해 지정된 셀에서 시작하여 대체됩니다. 스프레드시트의 기타 셀은 원래 콘텐츠로 남아 있습니다.

필드 이름 포함. 필드 이름이 워크시트의 첫 번째 행에 포함되는지 여부를 지정합니다.

셀에서 시작. 첫 번째 내보내기 레코드(**필드 이름 포함**이 선택된 경우에는 첫 번째 필드 이름)에 사용되는 셀 위치입니다. 데이터는 오른쪽까지 이 초기 셀에서 아래로 채워집니다.

워크시트 선택. 데이터를 내보낼 워크시트를 지정합니다. 인덱스별 또는 이름별로 워크시트를 식별할 수 있습니다.

- **인덱스별.** 새 파일을 작성하는 경우 0에서 9까지의 숫자를 지정하여 내보낼 워크시트를 식별하십시오(첫 번째 워크시트의 경우 0으로 시작하고 두 번째 워크시트의 경우 1로 시작하는 방식임). 워크시트가 이미 이 위치에 있는 경우에만 10 이상의 값을 사용할 수 있습니다.
- **이름별.** 새 파일을 작성하는 경우 워크시트에 사용되는 이름을 지정하십시오. 기존 파일에 삽입하는 경우 이 워크시트가 있으면 이 워크시트에 데이터가 삽입되고 이 워크시트가 없으면 이 이름을 가진 새 워크시트가 작성됩니다.

Excel 시작. 노드가 실행될 때 내보낸 파일에 대해 Excel이 자동으로 시작되는지를 지정합니다. IBM® SPSS® Modeler Server에 대해 분산 모드에서 실행 중인 경우 출력은 서버 파일 시스템에 저장되고 Excel은 내보낸 파일의 사본을 사용하여 클라이언트에서 시작됩니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 읽을 Excel 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 Excel 소스 노드의 내용을 참조하십시오.

(10) 확장 내보내기 노드

확장 내보내기 노드를 사용하면 R 또는 Python for Spark 스크립트를 실행하여 데이터를 내보낼 수 있습니다.

① 확장 내보내기 노드 - 명령문 탭

구문 유형(R 또는 Python for Spark)을 선택하십시오. 자세한 정보는 다음 섹션을 참조하십시오. 명령문이 준비되면 **실행**을 클릭하여 확장 내보내기 노드를 실행할 수 있습니다.

R 구문

R 구문. 데이터 분석을 위해 R 스크립트 구문을 이 필드에 입력, 붙여넣기 또는 사용자 정의할 수 있습니다.

플래그 필드 변환. 플래그 필드를 처리하는 방법을 지정합니다. **문자열에서 요인으로**, 정수 및 실수에서 **double**로 및 논리 값(True, False)이라는 두 가지 옵션이 있습니다. 논리 값(True, False)을 선택하면 플래그 필드의 원래 값이 손실됩니다. 예를 들어, 필드에 남성 및 여성 값이 있는 경우, 해당 값이 True 및 False로 변경됩니다.

결측값을 R '사용할 수 없음' 값(NA)으로 변환. 선택하면 모든 결측값이 R NA 값으로 변환됩니다. NA 값은 결측값을 식별하기 위해 R에서 사용됩니다. 사용하는 일부 R 함수에는 데이터에 NA가 포함된 경우에 함수가 작동하는 방식을 제어하는 데 사용되는 인수があります. 예를 들어, 함수에서 NA를 포함하는 레코드를 자동으로 제외하도록 선택할 수 있습니다. 이 옵션을 선택하지 않으면 모든 결측값이 변경되지 않은 상태로 R에 전달되고 R 스크립트가 실행될 때 오류가 발생할 수 있습니다.

날짜/시간 필드를 시간대의 특수 제어가 있는 R 클래스로 변환. 이 옵션을 선택하면 날짜 또는 날짜/시간 형식의 변수가 R 날짜/시간 개체로 변환됩니다. 다음 옵션 중 하나를 선택해야 합니다.

- R POSIXct. 날짜 또는 날짜/시간 형식의 변수가 R POSIXct 개체로 변환됩니다.
- R POSIXlt (목록). 날짜 또는 날짜/시간 형식의 변수가 R POSIXlt 개체로 변환됩니다.

 **참고:** POSIX 형식은 고급 옵션입니다. R 스크립트에서 날짜/시간 필드가 해당 형식이 필요한 방식으로 처리되도록 지정된 경우에만 이 옵션을 사용하십시오. POSIX 형식은 시간 형식이 있는 변수에 적용되지 않습니다.

Python 구문(S)

Python 구문. 이 필드에 데이터 분석을 위한 Python 스크립팅 구문을 입력하거나 붙여넣거나 사용자 정의할 수 있습니다. Python for Spark에 대한 자세한 정보는 Python for Spark 및 Python for Spark로 스크립팅의 내용을 참조하십시오.

② 확장 내보내기 노드 - 콘솔 출력 탭

콘솔 출력 탭에는 명령문 탭에서 R 스크립트 또는 Python for Spark 스크립트가 실행될 때 수신된 모든 출력이 포함됩니다. 예를 들어, R 스크립트를 사용하는 경우, **명령문** 탭의 **R 명령문** 필드의 R 스크립트가 실행될 때 R 콘솔에서 수신된 출력을 표시합니다. 이 출력에는 R 또는 Python 스크립트가 실행될 때 생성되는 R 또는 Python 오류 메시지 또는 경고가 포함됩니다. 출력은 주로 스크립트를 디버그하는 데 사용될 수 있습니다. **콘솔 출력** 탭에는 **R 명령문** 또는 **Python 명령문** 필드의 스크립트도 포함됩니다.

확장 내보내기 스크립트가 실행될 때마다 R 콘솔 또는 Python for Spark에서 수신된 출력이 **콘솔 출력** 탭의 내용을 덮어씁니다. 출력은 편집할 수 없습니다.

(11) XML 내보내기 노드

XML 내보내기 노드에서는 UTF-8 인코딩을 사용하여 XML 형식의 데이터를 출력할 수 있습니다. 선택적으로 XML 소스 노드를 작성하여 내보내진 데이터를 다시 스트림으로 읽을 수 있습니다.

XML 내보내기 파일. 데이터를 내보낼 XML 파일의 전체 경로 및 파일 이름입니다.

XML 스키마 사용. 스키마 또는 DTD를 사용하여 내보내는 데이터의 구조를 제어하려면 이 선택란을 선택하십시오. 그러면 아래에 설명된 **맵핑** 단추가 활성화됩니다.

스키마 또는 DTD를 사용하지 않는 경우에는 내보내는 데이터에 다음과 같은 기본 구조가 사용됩니다.

```
<records>
  <record>
    <fieldname1>value</fieldname1>
    <fieldname2>value</fieldname2>
    :
    <fieldnameN>value</fieldnameN>
  </record>
  <record>
    :
    :
  </record>
  :
  :
</records>
```

필드 이름에 있는 공백은 밑줄로 대체됩니다. 예를 들어, "My Field"는 <My_Field>가 됩니다.

맵핑. XML 스키마를 사용하기로 선택한 경우, 이 단추는 각각의 새 레코드를 시작하는 데 사용할 XML 구조 파트를 지정할 수 있는 대화 상자를 엽니다. 자세한 정보는 XML 레코드 맵핑 옵션 주제를 참조하십시오.

맵핑된 필드. 맵핑된 필드 수를 표시합니다.

현재 데이터의 입력 노드 생성. 내보낸 데이터 파일을 스트림으로 다시 읽어오는 XML 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 자세한 정보는 XML 소스 노드 주제를 참조하십시오.

① XML 데이터 쓰기

XML 요소를 지정하면 요소 태그 안에 해당 필드 값이 배치됩니다.

```
<element>value</element>
```

속성을 맵핑하면 해당 필드 값은 속성의 값으로서 배치됩니다.

```
<element attribute="value">
```

필드를 <records> 요소 위에 있는 요소에 맵핑하는 경우, 해당 필드는 한 번만 쓰여지고 모든 레코드에 대한 하나의 상수가 됩니다. 이 요소의 값은 첫 번째 레코드에서 비롯됩니다.

빈 콘텐츠를 지정하면 널값이 쓰여집니다. 요소의 경우 다음과 같습니다.

```
<element></element>
```

속성의 경우 다음과 같습니다.

```
<element attribute="">
```

② XML 레코드 맵핑 옵션

레코드 탭에서 각각의 새 레코드를 시작하는 데 사용할 XML 구조 파트를 지정할 수 있습니다. 스키마로 올바르게 맵핑하려면 레코드 구분자를 지정해야 합니다.

XML 구조. 이전 화면에서 지정한 XML 스키마의 구조를 보여주는 계층 구조 트리입니다.

레코드(XPath 표현식). 레코드 구분자를 설정하려면 XML 구조에서 요소를 선택하고 오른쪽 화살표 단추를 클릭하십시오. 소스 데이터에서 이 요소가 발견될 때마다 출력 파일에 새 레코드가 작성됩니다.

참고: XML 구조의 루트 요소를 선택하는 경우, 하나의 레코드만 쓸 수 있고 기타 모든 레코드는 건너됩니다.

③ XML 필드 매핑 옵션

필드 탭은 스키마 파일이 사용될 때 데이터 세트의 필드를 XML 구조의 요소 또는 속성으로 매핑하는 데 사용됩니다.

요소 또는 속성 이름과 일치하는 필드 이름은 해당 요소 또는 속성 이름이 고유한 경우 자동으로 매핑됩니다. 따라서 이름이 field1인 요소 및 속성이 모두 있으면 자동 매핑이 수행되지 않습니다. 구조에 이름이 field1인 항목이 하나만 있으면 스트림에서 해당 이름을 갖는 필드는 자동으로 매핑됩니다.

필드. 모델의 필드 목록입니다. 매핑의 소스 파트로 하나 이상의 필드를 선택하십시오. 목록 맨 아래에 있는 단추를 사용하여 모든 필드를 선택하거나 특정 측정 수준을 갖는 모든 필드를 선택할 수 있습니다.

XML 구조. 매핑 대상으로 사용할 XML 구조의 요소를 선택하십시오. 매핑을 작성하려면 매핑을 클릭하십시오. 그러면 매핑이 표시됩니다. 이러한 방식으로 매핑된 필드의 수가 이 목록 아래에 표시됩니다.

매핑을 제거하려면 XML 구조 목록에서 해당 항목을 선택하고 **매핑 해제**를 클릭하십시오.

속성 표시. XML 구조에 있는 XML 요소의 속성(있는 경우)을 표시하거나 숨깁니다.

④ XML 매핑 미리보기

미리보기 탭에서 **업데이트**를 클릭하면 작성될 XML의 미리보기를 볼 수 있습니다.

매핑이 올바르지 않은 경우, 레코드 또는 필드 탭으로 돌아가 오류를 정정하고 다시 **업데이트**를 클릭하여 결과를 확인하십시오.

(12) JSON 내보내기 노드

JSON 내보내기 노드에서는 UTF-8 인코딩을 사용하여 데이터를 JSON 형식으로 출력할 수 있습니다. 선택적으로 JSON 소스 노드를 작성하여 내보낸 데이터를 다시 스트림으로 읽어들이 수도 있습니다.

SPSS® Modeler가 데이터를 JSON 내보내기 파일에 쓸 때 다음과 같은 변환이 수행됩니다.

표 1. JSON 데이터 내보내기 변환

SPSS Modeler 데이터 저장 공간	JSON값
문자열	문자열
정수	number(int)
실수	number(real)
날짜	문자열
시간	문자열
시간소인	문자열
목록	지원되지 않습니다. 목록 필드는 제외됩니다.
결측값	null

JSON 내보내기 파일: 데이터를 내보낼 JSON 파일의 전체 경로 및 파일 이름입니다.

JSON 문자열 형식: JSON 문자열의 형식을 지정하십시오. JSON 내보내기 노드가 이름 및 값 쌍 컬렉션을 출력하도록 하려면 **레코드**를 선택하십시오. (이름 없이) 값만 내보내려는 경우 **값**을 선택하십시오.

JSON 문자열 형식: JSON 문자열의 형식을 지정하십시오. JSON 내보내기 노드가 이름 및 값 쌍 컬렉션을 출력하도록 하려면 **레코드**를 선택하십시오. (이름 없이) 값만 내보내려는 경우 **값**을 선택하십시오.

현재 데이터의 입력 노드 생성: 내보낸 데이터 파일을 다시 스트림으로 읽어들이 JSON 소스 노드를 자동으로 생성하려면 이 옵션을 선택하십시오. 추가 정보는 JSON 소스 노드의 내용을 참조하십시오.

(13) 공통 내보내기 노드 탭

다음은 해당되는 탭을 클릭하여 모든 내보내기 노드에 대해 지정할 수 있는 옵션입니다.

- **출판 탭.** 스트림 결과를 출판하는 데 사용됩니다.
- **주석(Annotation) 탭.** 이 탭은 모든 노드에 사용되며 노드의 이름을 바꾸고 사용자 맞춤 도구 팁을 제공하며 긴 주석(Annotation)을 저장하기 위한 옵션을 제공합니다.

① 스트림 출판

스트림 출판은 데이터베이스, 플랫 파일, Statistics 내보내기, 확장 내보내기, 데이터 콜렉션 내보내기, SAS 내보내기, Excel 및 XML 내보내기 노드 등의 표준 내보내기 노드를 사용하여 IBM® SPSS® Modeler에서 직접 수행됩니다. 내보내기 노드의 유형에 따라 출판된 스트림이 IBM SPSS Modeler Solution Publisher Runtime 또는 외부 애플리케이션을 사용하여 실행될 때마다 기록될 결과의 형식이 결정됩니다. 예를 들어, 출판된 스트림이 실행될 때마다 결과를 데이터베이스에 기록하려면 데이터베이스 내보내기 노드를 사용하십시오.

스트림 출판

1. 일반 방식으로 스트림을 열거나 작성하고 내보내기 노드를 끝에 첨부하십시오.
2. 내보내기 노드의 출판 탭에서, 출판된 파일의 루트 이름(즉, .pim, .par, .xml 등의 다양한 확장자를 붙여쓸 파일 이름)을 지정하십시오.
3. 스트림을 출판하려면 **출판**을 클릭하고, 노드가 실행될 때마다 스트림을 출판하려면 **스트림 출판**을 선택하십시오.

출판된 이름 - 출판된 이미지 및 모수 파일에 대한 루트 이름을 지정하십시오.

- **이미지 파일(*.pim)**은 Runtime이 내보내기 당시와 똑같이 출판된 스트림을 실행하는 데 필요한 모든 정보를 제공합니다. 스트림의 설정(예: 입력 데이터 소스 또는 출력 데이터 파일)을 변경하지 않아도 된다면 이미지 파일만 배포할 수 있습니다.
- **모수 파일(*.par)**에는 데이터 소스, 출력 파일 및 실행 옵션에 대한 구성 가능 정보가 포함되어 있습니다. 스트림을 다시 출판하지 않고도 스트림의 입력 또는 출력을 제어할 수 있으려면 이미지 파일뿐 아니라 모수 파일도 필요합니다.
- **메타데이터 파일(*.xml)**은 이미지 및 해당 데이터 모델의 입력 및 출력을 설명합니다. 이 파일은 런타임 라이브러리를 임베드하고 입력 및 출력 데이터의 구조를 알아야 하는 애플리케이션에서 사용하도록 설계되었습니다.

 **참고:** 이 파일은 **메타데이터 출판** 옵션을 선택한 경우에만 생성됩니다.

모수 출판 - 필요한 경우, 스트림 모수를 *.par 파일에 포함시킬 수 있습니다. *.par 파일을 편집하거나 런타임 API를 통해, 이미지를 실행할 때 이러한 스트림 모수값을 변경할 수 있습니다.

이 옵션을 선택하면 **모수 단추**가 활성화됩니다. 이 단추를 클릭하면 모수 출판 대화 상자가 표시됩니다.

출판 열에서 관련 옵션을 선택하여, 출판된 이미지에 포함될 모수를 선택하십시오.

스트림 실행 시 - 노드가 실행될 때 스트림이 자동으로 출판되는지 여부를 지정합니다.

- **데이터 내보내기** - 스트림을 출판하지 않고 표준 방식으로 내보내기 노드를 실행합니다. (기본적으로 이 노드는 IBM SPSS Modeler Solution Publisher를 사용할 수 없는 경우와 동일한 방식으로 IBM SPSS Modeler에서 실행됩니다.) 이 옵션을 선택하면 내보내기 노드 대화 상자에서 출판을 클릭하여 명시적으로 출판하지 않는 한 스트림이 출판되지 않습니다. 또는 도구 모음의 출판 도구를 사용하거나 스크립트를 사용하여 현재 스트림을 출판할 수도 있습니다.
- **스트림 출판** - IBM SPSS Modeler Solution Publisher를 사용하여 배포할 스트림을 출판합니다. 스트림이 실행될 때마다 스트림을 자동으로 출판하려면 이 옵션을 선택하십시오.

참고:

- 출판된 스트림을 새 데이터 또는 업데이트된 데이터로 실행할 계획인 경우, 입력 파일의 필드 순서는 출판된 스트림에 지정된 소스 노드 입력 파일의 필드 순서와 동일해야 합니다.
- 외부 애플리케이션에 출판할 때는 관계없는 필드를 필터링하거나 입력 요구사항에 맞게 필드 이름을 변경할 것을 고려하십시오. 두 작업 모두, 내보내기 노드 전에 필터 노드를 사용하여 수행할 수 있습니다.

7) 슈퍼노드

(1) 슈퍼노드 개요

IBM® SPSS® Modeler 비주얼 프로그래밍 인터페이스가 배우기 쉬운 이유 중 하나는 각 노드가 명확하게 정의된 기능을 가진다는 것입니다. 그러나 복잡한 처리의 경우 긴 노드 시퀀스가 필요할 수도 있습니다. 이로 인해 결국 스트림 캔버스가 어수선해지고 스트림 다이어그램을 이해하기 어려워질 수 있습니다. 두 가지 방법을 사용하여 길고 복잡한 스트림으로 인한 혼잡을 피할 수 있습니다.

- 처리 시퀀스를 여러 스트림(하나를 다른 하나에 공급)으로 분할할 수 있습니다. 예를 들어, 첫 번째 스트림은 두 번째 스트림이 입력으로 사용하는 데이터 파일을 작성합니다. 두 번째 스트림은 세 번째 스트림이 입력으로 사용하는 파일을 작성하며, 계속해서 이와 같이 반복됩니다. 이러한 여러 스트림을 하나의 **프로젝트**에 저장하여 이들을 관리할 수 있습니다. 프로젝트는 여러 스트림과 해당 출력을 위한 조직을 제공합니다. 그러나 프로젝트 파일에는 포함되는 오브젝트에 대한 참조만 포함되며, 따라서 여전히 관리해야 할 다수의 스트림 파일이 있습니다.
- 복잡한 스트림 프로세스에 대해 작업할 때 보다 간소화된 대안으로서 **슈퍼노드**를 작성할 수 있습니다.

슈퍼노드는 데이터 스트림의 섹션을 캡슐화하여 여러 노드를 하나의 노드로 그룹화합니다. 이는 데이터 마이너에게 여러 가지 이점을 제공합니다.

- 스트림이 더 깔끔하고 관리하기가 더 쉽습니다.
- 노드를 하나의 비즈니스별 슈퍼노드로 결합할 수 있습니다.
- 여러 데이터 마이닝 프로젝트에서 재사용할 수 있도록 슈퍼노드를 라이브러리로 내보낼 수 있습니다.

(2) 슈퍼노드 유형

슈퍼노드는 데이터 스트림에서 별 아이콘으로 표시됩니다. 아이콘은 음영 처리되어 슈퍼노드의 유형과 스트림이 흐르는 방향을 나타냅니다.

세 가지 유형의 슈퍼노드가 있습니다.

- 소스 슈퍼노드
- 프로세스 슈퍼노드
- 터미널 슈퍼노드

① 소스 슈퍼노드

소스 슈퍼노드는 보통 소스 노드처럼 데이터 소스를 포함하며, 보통 소스 노드를 사용할 수 있는 곳이면 어디서나 사용할 수 있습니다. 소스 슈퍼노드의 왼쪽이 음영 처리되어 왼쪽이 "달려" 있고 슈퍼노드에서 아래로 데이터가 흘러야 함을 표시합니다.

소스 슈퍼노드는 연결점이 오른쪽에 하나만 있으며, 이는 데이터가 해당 슈퍼노드에서 나와 스트림으로 흐름을 표시합니다.

② 프로세스 슈퍼노드

프로세스 슈퍼노드는 프로세스 노드만 포함하며, 이 유형의 슈퍼노드로 데이터가 들어올 수 있을 뿐만 아니라 이 유형의 슈퍼노드에서 데이터가 나갈 수 있음을 표시하기 위해 음영 처리되지 않습니다.

프로세스 슈퍼노드는 왼쪽과 오른쪽 모두에 연결점이 있어서 데이터가 슈퍼노드로 들어가고 여기서 나와 다시 스트림으로 흐름을 보여줍니다. 슈퍼노드는 추가 스트림 단편과 추가 스트림까지 포함할 수 있지만, 양 연결점은 시작 스트림과 끝 스트림 지점을 연결하는 단일 경로를 통해 흘러야 합니다.

참고: 프로세스 슈퍼노드는 조작 슈퍼노드라고도 합니다.

③ 터미널 슈퍼노드

터미널 슈퍼노드는 하나 이상의 터미널 노드(plot, 테이블 등)를 포함하며 터미널 노드와 동일한 방식으로 사용할 수 있습니다. 터미널 슈퍼노드는 오른쪽이 음영 처리되어 오른쪽이 "달려" 있고 터미널 슈퍼노드만 데이터가 흐를 수 있음을 표시합니다.

터미널 슈퍼노드는 연결점이 왼쪽에 하나만 있으며, 이는 데이터가 스트림에서 슈퍼노드로 들어가 해당 슈퍼노드 내에서 종료됨을 표시합니다.

터미널 슈퍼노드에는 슈퍼노드 내에 있는 모든 터미널 노드의 실행 순서를 지정하는 데 사용되는 스크립트도 포함될 수 있습니다. 자세한 정보는 슈퍼노드 및 스크립팅 주제를 참조하십시오.

(3) 슈퍼 노드 작성

슈퍼노드를 작성하면 여러 노드가 하나의 노드로 캡슐화되므로 데이터 스트림이 "수축"됩니다. 캔버스에서 스트림을 작성하거나 로드한 후 여러 가지 방식으로 슈퍼노드를 작성할 수 있습니다.

다중 선택

슈퍼노드를 작성하는 가장 쉬운 방법은 캡슐화할 노드를 모두 선택하는 것입니다.

1. 마우스를 사용하여 스트림 캔버스에서 여러 노드를 선택하십시오. Shift-클릭을 사용하여 스트림 또는 스트림 섹션을 선택할 수도 있습니다.

참고: 선택하는 노드는 연속 또는 갈라진 스트림의 노드여야 합니다. 인접하지 않거나 어떤 방식으로든 연결되지 않은 노드는 선택할 수 없습니다.

2. 다음 세 가지 방법 중 하나를 사용하여 선택된 노드를 캡슐화하십시오.
 - 도구 모음에서 슈퍼노드 아이콘(별 모양과 유사)을 클릭하십시오.
 - 슈퍼노드를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 다음을 선택하십시오.
슈퍼노드 작성 > 선택항목에서
 - 슈퍼노드 메뉴에서 다음을 선택하십시오.
슈퍼노드 작성 > 선택항목에서

이 세 옵션 모두 노드를 하나의 슈퍼노드로 캡슐화하며, 슈퍼노드는 해당 콘텐츠를 기반으로 해당 유형(소스, 프로세스 또는 터미널)을 반영하도록 음영 처리됩니다.

단일 선택

단일 노드를 선택하고 메뉴 옵션으로 슈퍼노드의 시작 및 끝을 정하거나 선택된 노드의 모든 다운스트림 노드를 캡슐화하여 슈퍼노드를 작성할 수도 있습니다.

1. 캡슐화의 시작을 결정하는 노드를 클릭하십시오.
2. 슈퍼노드 메뉴에서 다음을 선택하십시오.

슈퍼노드 작성 > 여기에서

스트림 섹션의 시작 및 끝을 선택하여 노드를 캡슐화하면 보다 대화식으로 슈퍼노드를 작성할 수 있습니다.

1. 슈퍼노드에 포함시킬 첫 번째 또는 마지막 노드를 클릭하십시오.
2. 슈퍼노드 메뉴에서 다음을 선택하십시오.

슈퍼노드 작성 > ...선택

3. 또는 원하는 노드를 마우스 오른쪽 단추로 클릭하여 컨텍스트 메뉴 옵션을 사용할 수 있습니다.
4. 커서가 슈퍼노드 아이콘으로 바뀌어 스트림의 다른 지점을 선택해야 함을 표시합니다. 위 또는 아래로 움직여 슈퍼노드 단편의 "다른 쪽 끝"으로 이동한 후 노드를 클릭하십시오. 그러면 그 사이에 있는 모든 노드가 슈퍼노드 별 아이콘으로 바뀝니다.

참고: 선택하는 노드는 연속 또는 갈라진 스트림의 노드여야 합니다. 인접하지 않거나 어떤 방식으로든 연결되지 않은 노드는 선택할 수 없습니다.

① 슈퍼노드 중첩

슈퍼노드는 다른 슈퍼노드 내에 중첩시킬 수 있습니다. 중첩된 슈퍼노드에는 각 슈퍼노드 유형(소스, 프로세스 및 터미널)에 적용되는 것과 동일한 규칙이 적용됩니다. 예를 들어, 중첩을 포함하는 프로세스 슈퍼노드가 프로세스 슈퍼노드로 유지되려면 중첩된 모든 슈퍼노드를 통과하는 연속된 데이터 플로우가 있어야 합니다. 중첩된 슈퍼노드 중 하나가 터미널이면 데이터는 더 이상 해당 계층 구조를 통해 흐르지 않습니다.

터미널 및 소스 슈퍼노드는 다른 유형의 중첩된 슈퍼노드를 포함할 수 있지만, 슈퍼노드 작성에 적용되는 기본 규칙과 동일한 규칙이 적용됩니다.

(4) 슈퍼노드 잠금

슈퍼노드를 작성한 후에는 슈퍼노드가 수정되지 않도록 비밀번호를 사용하여 슈퍼노드를 잠글 수 있습니다. 예를 들어, IBM® SPSS® Modeler 인콰이어리 설정 경험이 적은 조직의 다른 사

용자가 사용할 수 있도록 고정값 템플릿로서 스트림 또는 스트림 파트를 작성하는 경우에 이를 수행할 수 있습니다.

수퍼노드가 잠긴 경우에도 사용자는 계속 매개변수 탭에서 정의된 매개변수의 값을 입력할 수 있으며, 비밀번호를 입력하지 않고 잠긴 수퍼노드를 실행할 수 있습니다.

 **참고:** 스크립트를 사용하여 잠금 및 잠금 해제를 수행할 수 없습니다.

① 수퍼노드 잠금 및 잠금 해제

 **경고:** 분실한 비밀번호는 복구할 수 없습니다.

세 탭 중 하나에서 수퍼노드를 잠그거나 잠금 해제할 수 있습니다.

1. **노드 잠금**을 클릭하십시오.
2. 비밀번호를 입력하고 확인하십시오.
3. **확인**을 클릭하십시오.

비밀번호가 보호되는 수퍼노드는 스트림 캔버스에서 수퍼노드 아이콘의 맨 위 왼쪽에 작은 자물쇠 기호로 식별됩니다.

수퍼노드 잠금 해제

1. 비밀번호 보호를 영구적으로 제거하려면 **노드 잠금 해제**를 클릭하십시오. 비밀번호를 입력하도록 프롬프트됩니다.
2. 비밀번호를 입력하고 **확인**을 클릭하십시오. 해당 수퍼노드의 비밀번호가 더 이상 보호되지 않고 스트림에서 해당 아이콘 옆에 자물쇠 기호가 더 이상 표시되지 않습니다.

잠긴 수퍼노드가 포함된 스트림이 SPSS® Modeler 버전 16 - 17.0에 저장된 경우, SPSS Modeler에서 설치한 JRE가 다를 때 IBM® SPSS Collaboration and Deployment Services 또는 Mac과 같은 서로 다른 환경에서 스트림을 열 경우 먼저 스트림을 열고 잠금을 해제한 후 스트림이 마지막으로 저장된 이전 환경에서 버전 17.1 이상을 사용하여 다시 저장해야 합니다.

일부 경우 버전 18 이전의 스트림에서 수퍼노드를 잠금 해제할 경우 잘못된 비밀번호 오류가 표시됩니다. 이 문제를 해결하려면 노드를 마지막으로 저장할 당시와 동일한 시스템 로컬 설정을 사용하는 동일한 플랫폼에서 정확한 IBM SPSS Modeler 버전(또는 최신 버전)을 사용하여 노드를 다시 열고 잠금 해제하십시오. 그런 다음 버전 18 이상에서 노드를 열고 노드를 잠금 후 스트림을 다시 저장하십시오.

② 잠긴 수퍼노드 편집

매개변수를 정의하거나 확대하여 잠긴 수퍼노드를 표시하려 하면 비밀번호 입력을 요구하는 프롬프트가 표시됩니다.

비밀번호를 입력하고 **확인**을 클릭하십시오.

이제 해당 수퍼노드가 있는 스트림을 닫을 때까지 필요할 때마다 매개변수 정의를 편집하고 수퍼노드를 확대/축소할 수 있습니다.

이 조치로 인해 비밀번호 보호가 제거되지는 않습니다. 단지 수퍼노드에 액세스하여 관련 작업을 수행할 수만 있습니다. 자세한 정보는 수퍼노드 잠금 및 잠금 해제 주제를 참조하십시오.

(5) 수퍼노드 편집

수퍼노드를 작성한 후에는 수퍼노드를 확대하여 보다 자세하게 검토할 수 있습니다. 수퍼노드가 잠겨 있으면 비밀번호 입력을 요구하는 프롬프트가 표시됩니다. 자세한 정보는 잠긴 수퍼노드 편집 주제를 참조하십시오.

수퍼노드의 콘텐츠를 보려면 IBM® SPSS® Modeler 도구 모음의 확대 아이콘을 사용하거나 다음 방법을 사용할 수 있습니다.

1. 수퍼노드를 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 **확대**를 선택하십시오.

조금 다른 IBM SPSS Modeler 환경에서 선택한 수퍼노드의 콘텐츠가 표시되며, 스트림 또는 스트림 단편을 통한 데이터의 흐름을 나타내는 커넥터도 함께 표시됩니다. 이 수준에서는 스트림 캔버스에서 다음 태스크를 수행할 수 있습니다.

- 수퍼노드 유형(소스, 프로세스 또는 터미널)을 수정할 수 있습니다.
- 매개변수를 작성하거나 매개변수의 값을 편집할 수 있습니다. 매개변수는 스크립팅과 CLEM 표현식에서 사용됩니다.
- 수퍼노드와 해당 하위 노드에 대해 캐싱 옵션을 지정할 수 있습니다.
- 수퍼노드 스크립트를 작성하거나 수정할 수 있습니다(터미널 수퍼노드에만 해당).

① 수퍼노드 유형 수정

일부 환경에서는 수퍼노드의 유형을 변경하는 것이 유용합니다. 이 옵션은 수퍼노드를 확대한 경

우에만 사용할 수 있으며, 해당 수준에서는 이 옵션이 해당 슈퍼노드에만 적용됩니다. 다음 표에서 세 가지 유형의 슈퍼노드를 설명합니다.

표 1. 슈퍼노드 유형	
슈퍼노드 유형	설명
소스 슈퍼노드	나가는 하나의 연결
프로세스 슈퍼노드	두 개의 연결: 들어오는 연결과 나가는 연결
터미널 슈퍼노드	들어오는 하나의 연결

슈퍼노드의 유형 변경

1. 슈퍼노드를 확대해야 합니다.
2. 슈퍼노드 메뉴에서 **슈퍼노드 유형**을 선택한 후 유형을 선택하십시오.

② 슈퍼노드 주석(Annotation) 작성 및 이름 바꾸기

스트림에서 표시되는 슈퍼노드의 이름을 바꾸고 프로젝트 또는 보고서에서 사용되는 주석(Annotation)을 작성할 수 있습니다. 이러한 특성에 액세스하려면 다음을 수행하십시오.

- 슈퍼노드를 마우스 오른쪽 단추로 클릭하고(축소) **이름 변경 및 주석달기**를 선택하십시오.
- 또는 슈퍼노드 메뉴에서 **이름 변경 및 주석달기**를 선택하십시오. 이 옵션은 확대 및 축소 모드 둘 다에서 사용할 수 있습니다.

두 경우 모두 주석(Annotation) 탭이 선택된 대화 상자가 열립니다. 여기에 있는 옵션을 사용하여 스트림 캔버스에 표시되는 이름을 사용자 정의하고 슈퍼노드 조작에 관한 문서를 제공하십시오.

슈퍼노드에서 주석 사용

주석이 달린 노드 또는 너깃에서 슈퍼노드를 작성하는 경우, 슈퍼노드에 주석이 표시되도록 하려면 슈퍼노드를 작성하기 위한 선택항목에 주석을 포함시켜야 합니다. 선택항목에서 주석을 생략하면 슈퍼노드를 작성할 때 주석이 스트림에 남지 않습니다.

주석을 포함한 슈퍼노드를 펼치면 주석은 슈퍼노드를 작성하기 전에 있었던 위치로 복귀합니다.

주석이 달린 오브젝트가 포함된 슈퍼노드를 펼치는 경우, 주석이 슈퍼노드에 포함되지 않았으면 오브젝트는 원래 위치로 복귀하지만 주석은 다시 첨부되지 않습니다.

③ 수퍼 노드 모수

IBM® SPSS® Modeler에서는 사용자 정의 변수(예: Minvalue)를 설정할 수 있으며, 스크립팅 또는 CLEM 표현식에서 사용할 때 해당 값을 지정할 수 있습니다. 이러한 변수를 **매개변수**라고 합니다. 스트림, 세션 및 수퍼노드의 매개변수를 설정할 수 있습니다. 수퍼노드에 대해 설정된 매개변수는 해당 수퍼노드 또는 중첩된 노드에서 CLEM 표현식을 작성할 때 사용할 수 있습니다. 중첩된 수퍼노드에 대해 설정된 매개변수는 해당 상위 수퍼노드에서 사용할 수 없습니다.

두 단계로 수퍼노드의 매개변수를 작성하고 설정할 수 있습니다.

1. 수퍼노드의 매개변수를 정의합니다.
2. 그런 다음, 수퍼노드의 각 매개변수의 값을 지정합니다.

그러면 캡슐화 노드의 CLEM 표현식에서 이러한 매개변수를 사용할 수 있습니다.

가. 수퍼노드 매개변수 정의

확대 및 축소 모드 둘 다에서 수퍼노드의 매개변수를 정의할 수 있습니다. 정의된 매개변수는 모든 캡슐화 노드에 적용됩니다. 수퍼노드의 매개변수를 정의하려면 먼저 수퍼노드 대화 상자의 매개변수 탭에 액세스해야 합니다. 다음 방법 중 하나를 사용하여 대화 상자를 여십시오.

- 스트림에서 수퍼노드를 두 번 클릭하십시오.
- 수퍼노드 메뉴에서 **매개변수 설정**을 선택하십시오.
- 또는 수퍼노드를 확대한 경우 컨텍스트 메뉴에서 **매개변수 설정**을 선택하십시오.

대화 상자를 열면 매개변수 탭이 이전에 정의한 매개변수와 함께 표시됩니다.

새 매개변수 정의

매개변수 정의 단추를 클릭하여 대화 상자를 여십시오.

이름 모수 이름이 여기 나열됩니다. 이 필드에 이름을 입력하여 새 모수를 작성할 수 있습니다. 예를 들어, 최저 기온에 대한 모수를 작성하려면 minvalue를 입력할 수 있습니다. CLEM 표현식에서 모수를 나타내는 \$P- 접두문자를 포함시키지 마십시오. 이 이름은 CLEM 표현식 작성기에 표시하는 데도 사용됩니다.

긴 이름. 작성된 각 모수에 대한 설명 이름을 나열합니다.

저장 공간 목록에서 저장 유형을 선택하십시오. 저장 공간은 모수에 데이터 값이 저장되는 방법을 표시합니다. 예를 들어, 유지할 선행 0이 포함된 값(예: 008)에 대한 작업 시 저장 유형으로

문자열을 선택해야 합니다. 그렇지 않으면, 값에서 0이 제거됩니다. 사용 가능한 저장 유형은 문자열, 정수, 실수, 시간, 날짜, 시간소인입니다. 날짜 모수의 경우, 다음 단락에 표시된 대로 ISO 표준 표기법을 사용하여 값을 지정해야 합니다.

값 각 모수의 현재 값을 나열합니다. 필요에 따라 모수를 조정하십시오. 데이터 모수의 경우, ISO 표준 표기법(즉, YYYY-MM-DD)으로 값을 지정해야 합니다. 다른 형식으로 지정된 날짜는 허용되지 않습니다.

유형(선택사항). 외부 애플리케이션에 스트림을 배포할 계획이면 목록에서 측정 수준을 선택하십시오. 그렇지 않으면 유형 열을 있는 그대로 두는 것이 바람직합니다. 모수의 값 제한조건(예: 숫자 범위의 상한 및 하한)을 지정하려면 목록에서 **지정**을 선택하십시오.

사용자 인터페이스를 통해서만 모수에 대해 긴 이름, 저장 공간, 유형 옵션을 설정할 수 있습니다. 이러한 옵션은 스크립트를 사용하여 설정할 수 없습니다.

선택된 모수를 사용 가능한 모수 목록 위, 아래로 추가로 이동하려면 오른쪽에 있는 화살표를 선택하십시오. 선택된 모수를 제거하려면 삭제 단추(X로 표시)를 사용하십시오.

나. 슈퍼노드 매개변수의 값 설정

슈퍼노드의 매개변수를 정의한 후에는 CLEM 표현식 또는 스크립트에서 매개변수를 사용하여 값을 지정할 수 있습니다.

슈퍼노드의 매개변수를 지정하려면 다음을 수행하십시오.

1. 슈퍼노드 아이콘을 두 번 클릭하여 슈퍼노드 대화 상자를 여십시오.
2. 또는 슈퍼노드 메뉴에서 **매개변수 설정**을 선택하십시오.
3. **매개변수** 탭을 클릭하십시오. 참고: 이 대화 상자의 필드는 이 탭에서 **매개변수 정의** 단추를 클릭하여 정의한 필드입니다.
4. 작성한 각 매개변수의 텍스트 상자에 값을 입력하십시오. 예를 들어, *minvalue* 값을 관심이 있는 특정 임계값으로 설정할 수 있습니다. 그러면 다수의 조작(예: 향후 탐색을 위해 이 임계값보다 높거나 낮은 레코드 선택)에서 이 매개변수를 사용할 수 있습니다.

다. 슈퍼노드 모수를 사용하여 노드 특성 액세스

슈퍼노드 모수를 사용하여 캡슐화 노드의 노드 특성(슬롯 모수라고도 함)을 정의할 수도 있습니다. 예를 들어, 슈퍼노드가 사용 가능한 데이터의 무작위 표본을 사용하여 특정 시간 동안 캡슐화 신경망 노드를 학습시키도록 지정한다고 가정하십시오. 모수를 사용하여 시간 길이 및 백분율 표본에 대한 값을 지정할 수 있습니다.

예제 슈퍼노드에 *표본*이라는 표본 노드와 *학습*이라는 신경망 노드가 포함된다고 가정하십시오. 노드 대화 상자를 사용하여 표본 노드의 **표본** 설정을 **무작위 %**로 설정하고 신경망 노드의 **중지 시점** 설정을 **시간**으로 지정할 수 있습니다. 이러한 옵션을 지정하면 모수를 사용하여 노드 특성에 액세스하고 슈퍼노드에 고유한 값을 지정할 수 있습니다. 슈퍼노드 대화 상자에서 **모수 정의**를 클릭하고 다음 표에 표시된 모수를 작성하십시오.

표 1. 작성할 모수		
매개변수	값	긴 이름
Train.time	5	학습 시간(분)
Sample.random	10	백분율 무작위 표본

참고: *Sample.random*과 같은 모수 이름은 노드 특성을 참조하는 데 올바른 구문을 사용하며, 여기서 *Sample*은 노드의 이름을 나타내고 *random*은 노드 특성입니다.

이러한 모수를 정의한 후에는 각 대화 상자를 다시 열지 않고 표본 및 신경망 노드 특성의 값을 쉽게 수정할 수 있습니다. 슈퍼노드 메뉴에서 **모수 설정**을 선택하여 슈퍼노드 대화 상자의 모수 탭에 액세스하고 여기서 **무작위 %** 및 **시간**에 대해 새 값을 지정할 수 있습니다. 이는 특히 모델 작성을 여러 번 반복할 때 데이터를 탐색하는 데 유용합니다.

④ 슈퍼노드 및 캐싱

슈퍼노드 내에서 터미널 노드를 제외한 모든 노드를 캐싱할 수 있습니다. 노드를 마우스 오른쪽 단추로 클릭하고 캐시 컨텍스트 메뉴에서 여러 옵션 중 하나를 선택하여 캐싱을 제어합니다. 이 메뉴 옵션은 슈퍼노드 외부에서 사용 가능하고 슈퍼노드 내에 캡슐화된 노드에 사용할 수 있습니다.

슈퍼노드 캐시에 대한 몇 가지 지침은 다음과 같습니다.

- 슈퍼노드 내에 캡슐화된 노드 중 캐싱이 사용되는 노드가 있으면 해당 슈퍼노드 또한 캐싱이 사용됩니다.
- 슈퍼노드에서 캐시를 사용되지 않도록 설정하면 모든 캡슐화 노드에 대해서도 캐시가 사용되지 않습니다.
- 슈퍼노드에서 캐싱을 사용하면 실제로 캐싱 가능한 마지막 하위 노드에서 캐시가 사용됩니다. 즉, 마지막 하위 노드가 선택 노드인 경우 해당 선택 노드에 캐시가 사용됩니다. 마지막 하위 노드가 터미널 노드(캐싱을 허용하지 않음)이면 캐싱을 지원하는 그 다음 업스트림 노드에 캐시가 사용됩니다.
- 슈퍼노드의 하위 노드에 대해 캐시를 설정하면, 캐싱되는 노드에서 업스트림인 활동(예: 노드 추가 또는 편집)이 캐시를 비웁니다.

⑤ 슈퍼노드 및 스크립팅

SPSS® Modeler 스크립팅 언어를 사용하여 터미널 슈퍼노드의 콘텐츠를 조작하고 실행하는 단순 프로그램을 작성할 수 있습니다. 예를 들어, 복잡한 스트림의 실행 순서를 지정하려 할 수 있습니다. 슈퍼노드에 구성 노드 전에 실행해야 하는 전역값 설정 노드가 포함되는 경우, 전역값 설정 노드를 먼저 실행하는 스크립트를 작성할 수 있습니다. 평균이나 표준 편차 같이 이 노드가 계산하는 값을 구성 노드가 실행될 때 사용할 수 있습니다.

슈퍼노드 대화 상자의 스크립트 탭은 터미널 슈퍼노드에만 사용할 수 있습니다.

터미널 슈퍼노드에 대한 스크립팅 대화 상자를 열려면 다음을 수행하십시오.

- 슈퍼노드 캔버스를 마우스 오른쪽 단추로 클릭하고 **슈퍼노드 스크립트**를 선택하십시오.
- 또는 확대 및 축소 모드 둘 다에서 슈퍼노드 메뉴로부터 **슈퍼노드 스크립트**를 선택할 수 있습니다.

참고: 슈퍼노드 스크립트는 대화 상자에서 **현재 스크립트 실행**을 선택한 경우에 해당 스트림 및 슈퍼노드에서만 실행됩니다.

SPSS Modeler에서 스크립트를 작성하고 사용하는 데 필요한 고유 옵션에 대해서는 제품 다운로드에서 PDF 파일로 제공되는 *스크립팅 및 자동화 안내서*를 참조하십시오.

(6) 슈퍼노드 저장 및 로드

슈퍼노드의 장점 중 하나는 슈퍼노드를 저장하여 다른 스트림에서 재사용할 수 있는 것입니다. 슈퍼노드를 저장하고 로드할 때 슈퍼노드는 .slb 확장자를 사용합니다.

슈퍼노드 저장

1. 슈퍼노드를 확대하십시오.
2. 슈퍼노드 메뉴에서 **슈퍼노드 저장**을 선택하십시오.
3. 대화 상자에서 파일 이름 및 디렉토리를 지정하십시오.
4. 저장된 슈퍼노드를 현재 프로젝트에 추가할지 여부를 선택하십시오.
5. **저장**을 클릭합니다.

슈퍼노드 로드

1. IBM® SPSS® Modeler 창의 삽입 메뉴에서 **슈퍼노드**를 선택하십시오.
2. 현재 디렉토리의 슈퍼노드 파일(.slb)을 선택하거나 찾아보기를 사용하여 다른 디렉토리의 슈퍼노드 파일을 찾으십시오.
3. **로드**를 클릭하십시오.

참고: 가져온 슈퍼노드의 매개변수는 모두 기본값을 가집니다. 매개변수를 변경하려면 스트림 캔버스에서 슈퍼노드를 두 번 클릭하십시오.

3. 모델링 노드

1) 모델링 개요

(1) 모델링 노드의 개요

IBM® SPSS® Modeler는 기계 학습, 인공지능 및 통계로부터 취한 다양한 모델링 방법을 제공합니다. 모델링 팔레트에서 사용할 수 있는 이러한 방법을 통해 데이터로부터 새로운 정보를 얻어서 예측 모형을 개발할 수 있습니다. 각각의 방법은 그것만의 장점이 있으며 특정한 문제점 유형에 가장 적합합니다.

*IBM SPSS Modeler 애플리케이션 안내서*에서는 모델링 프로세스에 대한 일반적인 소개와 함께 이러한 여러 방법의 예제를 제공합니다. 이 안내서는 온라인 자습서로 사용 가능합니다. 자세한 정보.

모델링 방법은 다음 범주로 나뉩니다.

- 감독
- 연관
- 세분화

감독 모델

*감독 모델*은 하나 이상의 출력 또는 **목표** 필드를 예측하기 위해 하나 이상의 **입력** 필드 값을 사용합니다. 이러한 기술의 일부 예는 다음과 같습니다. 의사결정 트리(C&R 트리, QUEST, CHAID 및 C5.0 알고리즘), 회귀분석(1차, 로지스틱, 일반화 선형 및 Cox 회귀 알고리즘), 신경망, 지원 벡터 머신 및 베이지안 네트워크입니다.

감독 모델은 조직이 예를 들어, 고객이 구매할지 또는 떠날지 여부 또는 트랜잭션이 알려진 사기 패턴과 매치하는지 여부 등과 같이 알려진 결과를 예측하는 데 도움을 줍니다. 모델링 기법은 시스템 학습, 규칙 귀납, 하위 그룹 식별, 통계 방법 및 다중 모델 생성을 포함합니다.

감독 노드



자동 분류자 노드는 이분형 결과(예 또는 아니오, 이탈 또는 이탈 안함 등)에 대해 다수의 여러 모델을 작성하고 비교하여 주어진 분석을 위한 최상의 접근 방식을 선택할 수 있게 합니다. 많은 모델링 알고리즘이 지원되어 사용할 방법, 각각에 대한 특정 옵션, 결과 비교 기준을 선택할 수 있습니다. 이 노드는 지정된 옵션을 기반으로 모델 세트를 생성하고 사용자가 지정하는 기준에 따라 최상의 후보를 순위화합니다.



자동 수치 노드는 수많은 방법을 사용하여 연속적 수치 범위 결과의 모델을 추정하고 비교합니다. 이 노드는 자동 분류자 노드에서와 같은 방식으로 작동하므로 사용할 알고리즘을 선택하고 단일 모델링 전달에서 여러 옵션의 조합을 실험할 수 있습니다. 지원되는 알고리즘에는 신경망, C&R 트리, CHAID, 선형 회귀, 일반화 선형 회귀 및 지원 벡터 머신(SVM)이 있습니다. 모델은 상관관계, 상대 오차 또는 사용된 변수의 수를 기반으로 비교할 수 있습니다.



분류 및 회귀(C&R) 트리 노드는 추가 관측값을 예측하거나 분류할 수 있게 하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 "순수"로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).



QUEST 노드는 의사결정 트리를 작성하기 위한 이분형 분류 방법을 제공하며, 대형 C&R 트리 분석에 필요한 처리 시간을 줄이는 동시에 분류 트리 방법에서 찾아낸 경향을 줄여 더 많은 분할을 허용하는 입력을 선호하도록 설계되었습니다. 입력 필드는 숫자 범위(연속)일 수 있지만 대상 필드는 범주형이어야 합니다. 모든 분할은 이분형입니다.



CHAID 노드는 최적 분할을 식별하기 위해 카이제곱 통계량을 사용하여 의사결정 트리를 생성합니다. C&R 트리 및 QUEST 노드와 달리, CHAID는 비이분형 트리를 생성할 수 있으며 이는 일부 분할에 둘 이상의 분기가 있음을 의미합니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



C5.0 노드는 의사결정 트리 또는 규칙 세트를 작성합니다. 모델은 각 수준에서 최대 정보 이익을 제공하는 필드를 기반으로 샘플을 분할하여 작동합니다. 대상 필드는 범주형이어야 합니다. 세 개 이상의 부집단으로의 다중 분할이 허용됩니다.



의사결정 목록 노드는 전체 채우기에 상대적인 주어진 이분형 결과의 상위 또는 하위 우도를 표시하는 부집단 또는 세그먼트를 식별합니다. 예를 들어, 캠페인을 이탈할 가능성이 없거나 우호적으로 응답할 가능성이 가장 많은 고객을 찾고 있습니다. 자체 사용자 정의 세그먼트를 추가하고 대체 모델을 나란히 미리보기하여 결과를 비교함으로써 비즈니스 지식을 모델에 통합할 수 있습니다. 의사결정 목록 모델은 각 규칙에 조건과 결과가 있는 규칙 목록으로 구성됩니다. 규칙은 순서대로 적용되며 매치하는 첫 번째 규칙이 결과를 결정합니다.



선형 회귀 모형은 목표와 하나 이상의 예측변수 간의 선형 관계를 기반으로 연속형 목표를 예측합니다.



PCA/요인 노드에서는 강력한 데이터 축소 기법을 제공하여 데이터의 복잡도를 줄입니다. 비선형 주성분분석(PCA)은 구성요소가 서로 직각(수직)인 전체 필드 세트에서 변동을 캡처하는 입력 필드의 선형 조합을 찾습니다. 요인 분석은 관측된 필드 세트 내에서 상관관계 패턴을 설명하는 기본 요인을 식별하려고 시도합니다. 두 접근 방식 모두 목표는 원래 필드 세트의 정보를 효과적으로 요약하는 적은 수의 파생 필드를 찾는 것입니다.



필드선택 노드는 기준(예: 결측값의 퍼센트) 세트를 기반으로 제거용 입력 필드를 차단합니다. 그런 다음 지정된 대상에 상대적인 남아 있는 입력의 중요도에 대해 순위를 매깁니다. 예를 들어, 수백 개의 잠재 입력이 있는 데이터 세트가 있다면 환자 결과 모델링 시 어느 것이 가장 유용합니까?



판별 분석은 로지스틱 회귀분석보다 엄격한 가정을 하지만 해당 가정이 충족되면 로지스틱 회귀 분석의 귀중한 대안 또는 보조물이 될 수 있습니다.



로지스틱 회귀분석은 입력 필드 값을 기반으로 레코드를 분류하는 통계 기법입니다. 선형 회귀와 유사하지만 숫자 범위 대신 범주형 대상 필드를 사용합니다.



일반화 선형 모델은 종속변수가 요인과 선형적으로 관련되고 지정된 연결함수를 통해 공변되도록 일반 선형 모델을 확장합니다. 더욱이 모델을 사용하면 종속변수가 비정규 분포를 가질 수 있습니다. 선형 회귀, 로지스틱 회귀분석, 카운트 데이터에 대한 로그선형 모델, 간격 중도절단 생존 모델을 포함하여 상당수 통계 모델의 기능을 포함합니다.



일반화 선형 혼합 모델(GLMM)은 목표가 비정규 분포를 가질 수 있고 지정된 연결함수를 통해 요인 및 공변량과 선형적으로 관련되며 관측값을 상관시킬 수 있도록 선형 모델을 확장합니다. 일반화 선형 혼합 모델은 단순 선형 회귀에서 비정규 장기적인 데이터에 대한 복합 다중 수준 모델에 이르기까지 다양한 모델을 포함합니다.



Cox 회귀 노드를 통해 중도절단된 레코드가 있는 데서 시간 대 이벤트 데이터에 대한 생존 모델을 작성할 수 있습니다. 이 모델은 주어진 입력 변수 값에 대해 주어진 시간(t)에 흥미있는 이벤트가 발생한 확률을 예측하는 생존함수를 생성합니다.



지원 벡터 머신(SVM) 노드를 사용하면 데이터를 과적합 없이 두 개의 그룹 중 하나로 분류할 수 있습니다. SVM은 다수의 입력 필드가 있는 데이터 세트 등과 같은 광범위한 데이터 세트와 잘 작동합니다.



베이지안 네트워크 노드를 통해 관측 및 레코드된 증거를 실세계 지식과 조합하여 발생 우도를 확립함으로써 확률 모델을 작성할 수 있습니다. 이 노드는 주로 분류에 사용하는 TAN(Tree Augmented Naïve Bayes) 및 Markov Blanket 네트워크에 초점을 맞춥니다.



SLRM(Self-Learning Response Model) 노드를 사용하면 하나의 새 케이스 또는 소수의 새 케이스를 사용하여 모든 데이터를 사용하는 모델을 다시 학습시킬 필요 없이 모델을 재평가할 수 있는 모델을 작성할 수 있습니다.



시계열 노드는 시계열 데이터에 대한 지수평활, 일변량 자기회귀 통합 이동 평균 (ARIMA), 다변량 ARIMA(또는 전이 함수) 모델을 추정하고 미래 성능을 위한 예측값을 생성합니다. 이 시계열 노드는 SPSS Modeler 버전 18에서 더 이상 사용되지 않는 이전의 시계열 노드와 유사합니다. 그러나 이 새 시계열 노드는 IBM SPSS Analytic Server의 기능을 이용하여 빅 데이터를 처리해서 SPSS Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시하도록 설계되었습니다.



KNN(k-Nearest Neighbor) 노드는 새 케이스를 k 가 정수인 예측자 공간에서 가장 가까이 있는 k 오브젝트의 범주 또는 값과 연관시킵니다. 유사한 케이스는 서로 가까이 있고 유사하지 않은 케이스는 서로 멀리 떨어져 있습니다.



STP(Spatio-Temporal Prediction) 노드는 위치 데이터, 예측(예측자)을 위한 입력 필드, 시간 필드 및 대상 필드를 포함하는 데이터를 사용합니다. 각 위치에는 각 측정 시간에 각 예측변수의 값을 나타내는 데이터에 여러 행이 있습니다. 데이터가 분석된 후에는 분석에 사용된 모양 데이터 내에서 어떤 위치에서든 목표 값을 예측하는 데 사용할 수 있습니다.

연관 모델

*연관 모델*은 하나 이상의 엔티티(예: 이벤트, 구매 또는 속성)가 하나 이상의 다른 엔티티와 연관되어 있는 데이터에서 패턴을 발견합니다. 모델은 이러한 관계를 정의하는 규칙 세트를 구성합니다. 여기에서 데이터 내의 필드는 입력과 목표 둘 모두의 역할을 할 수 있습니다. 이러한 연관을 수동으로 찾을 수 있지만 연관 규칙 알고리즘은 이를 보다 신속하게 수행하므로 더 복잡한 패턴을 탐색할 수 있습니다. Apriori 및 Carma 모델은 이러한 알고리즘 사용의 예입니다. 연관 모델의 또 다른 유형은 순차 발견 모델이며 이는 시간 구조 데이터에서 순차 패턴을 발견합니다.

연관 모델은 다중 결과를 예측할 때 가장 유용합니다(예: 제품 X를 구매한 고객이 Y와 Z도 구매함). 연관 모델은 특정 결론(예: 구매 결정)을 조건 세트와 연관시킵니다. 다른 표준 의사결정 트리 알고리즘(C5.0 및 C&RT)에 비해 연관 규칙 알고리즘의 장점은 어떤 속성 사이에도 연관성이 있을 수 있다는 점입니다. 의사결정 트리 알고리즘은 단일 결론만 포함하는 규칙을 작성하지만, 연관 알고리즘은 각각 다른 결론을 보유할 수 있는 많은 규칙을 찾으려고 합니다.

연관 노드



Apriori 노드는 데이터에서 규칙 세트를 추출하고 정보 내용이 가장 많은 규칙을 꺼냅니다. Apriori는 규칙을 선택하는 5개의 서로 다른 방법을 제공하며 정교한 색인화 스킴을 사용하여 대형 데이터 세트를 효율적으로 처리합니다. 큰 문제점의 경우, Apriori는 일반적으로 학습 속도가 빠릅니다. 보유할 수 있는 규칙 수에 임의의 제한이 없으며 최대 32개의 전제조건을 가진 규칙을 처리할 수 있습니다. Apriori에서는 입력 및 출력 필드가 모두 범주형이어야 하지만 이런 유형의 데이터에 최적화되어 있기 때문에 우수한 성능을 제공합니다.



CARMA 모델은 입력 또는 대상 필드를 지정하지 않아도 데이터에서 규칙 세트를 추출합니다. Apriori와 대조적으로 CARMA 노드는 단지 전향 지원이 아니라 규칙 지원(전향 및 후향 둘 다에 대한 지원)을 위한 작성 설정을 제공합니다. 이는 생성된 규칙을 보다 다양한 애플리케이션에 사용하여, 예를 들어 후향이 이번 휴가철에 홍보할 항목인 제품 또는 서비스 목록을 찾을 수 있음을 의미합니다.



순차규칙 노드는 순차 또는 시간 지향 데이터에서 연관 규칙을 발견합니다. 순차규칙은 예측 가능한 순서로 발생하는 경향이 있는 항목 세트 목록입니다. 예를 들어, 면도기와 애프터쉐이브 로션을 구매하는 고객은 다음 번 구매 시에 면도용 크림을 구매할 수 있습니다. 순차규칙 노드는 순차규칙을 찾는 데 효율적인 2패스 방법을 사용하는 CARMA 연관 규칙 알고리즘을 기반으로 합니다.



연관 규칙 노드는 Apriori 노드와 유사합니다. 그러나 Apriori와는 달리, 연관 규칙 노드는 목록 데이터를 처리할 수 있습니다. 또한, 연관 규칙 노드는 빅 데이터를 처리하고 더 빠른 병렬 처리를 사용하기 위해 IBM SPSS Analytic Server와 함께 사용할 수 있습니다.

세분화 모델

*세분화 모델*은 데이터를 입력 필드의 패턴이 유사한 레코드의 세그먼트 또는 군집으로 나눕니다. 이들은 입력 필드에만 관심이 있으므로 세분화 모델에는 출력이나 목표 필드와 같은 개념이 없습니다. 세분화 모델의 예는 코호넨 네트워크, K-평균 군집, 이단계 군집 및 이상 항목 발견입니다.

세분화 모델("군집 모델"로도 알려짐)은 특정 결과가 알려지지 않은 경우에 유용합니다. 예를 들어, 새로운 사기 패턴을 식별할 때나, 고객 기반에 있는 관심 그룹을 식별할 때입니다. 군집 모

델은 유사한 레코드 그룹을 식별하고 레코드에 이들이 속하는 그룹에 따라 레이블을 붙이는 데 초점을 둡니다. 이는 그룹과 그룹의 특성에 대한 사전 지식 없이도 수행되고, 예측할 모델에 대한 사전에 정의된 출력이나 목표 필드가 없다는 면에서 군집 모델을 다른 모델링 기법과 구별해 줍니다. 이러한 모델에는 옳고 그른 응답이 없습니다. 이들 값은 데이터에서 관심 그룹을 캡처하는 기능에 의해 결정되고 이러한 그룹에 대한 유용한 설명을 제공합니다. 군집 모델은 종종 군집이나 세그먼트를 작성하는 데 사용하고 이러한 군집이나 세그먼트는 이후의 분석에서 입력으로 사용됩니다(예: 잠재 고객을 동종 하위그룹으로 분할하는 방법으로).

세분화 노드



자동 군집 노드는 유사한 특성을 가진 레코드 그룹을 식별하는 군집 모델을 추정하고 비교합니다. 이 노드는 다른 자동 모델링 노드와 동일한 방법으로 작동하여 단일 모델링 패스에서 다중 옵션 조합을 실험할 수 있습니다. 군집 모델의 유용성을 필터링하고 순위화하며 특정 필드의 중요성을 기반으로 축도를 제공하려고 시도하는 기본 축도를 사용하여 모델을 비교할 수 있습니다.



K-평균 노드는 데이터 세트를 고유 그룹(또는 군집)으로 군집화합니다. 이 방법은 고정된 수의 군집을 정의하고 반복적으로 레코드를 군집에 지정하며, 추가 세분화가 더 이상 모델을 향상시킬 수 없을 때까지 군집중심을 조정합니다. 결과를 예상하는 대신 k-평균은 자율 학습으로 알려진 프로세스를 사용하여 입력 필드 세트의 패턴을 찾아냅니다.



코호넨 노드는 데이터 세트를 고유 그룹으로 군집화하는 데 사용할 수 있는 신경망 유형을 생성합니다. 네트워크가 완전히 숙달되면, 유사 레코드는 출력 맵 가까이 있지만, 다른 레코드는 멀리 떨어져 있을 것입니다. 모델 너깃에서 각 단위별로 캡처된 관측값을 살펴 강한 단위를 식별할 수 있습니다. 이것은 적당한 군집 수에 대한 감각을 제공할 것입니다.



이단계 노드는 2단계 군집방법을 사용합니다. 첫 번째 단계는 원시 입력 데이터를 관리 가능한 하위 군집 세트로 압축하기 위해 데이터를 통한 단일 전달을 수행합니다. 두 번째 단계는 계층적 군집 방법을 사용하여 하위 군집을 점점 더 큰 군집으로 계속해서 병합하는 것입니다. 이단계는 학습 데이터에 대한 최적 군집 수를 자동으로 평가하는 장점이 있습니다. 혼합 필드 유형과 대형 데이터 세트를 효율적으로 처리할 수 있습니다.



이상 항목 발견 노드는 "정상" 데이터 패턴을 따르지 않는 특이 케이스 또는 이상값을 식별합니다. 이 노드를 사용하면 이전에 알려진 패턴에 적합하지 않고, 찾고 있는 패턴을 정확하게 모르더라도 이상값을 식별할 수 있습니다.

In-Database 마이닝 모델

IBM SPSS Modeler에서 Oracle Data Miner 및 Microsoft Analysis Services를 포함하여 데이터베이스 벤더에서 제공하는 데이터 마이닝과 모델링 도구의 통합을 지원합니다. 모두 IBM SPSS Modeler 애플리케이션 내에서 시작하여 데이터베이스 내에 모델을 작성하고, 스코어링하고, 저장할 수 있습니다. 자세한 정보는 *IBM SPSS Modeler In-Database 마이닝 안내서*를 참조하십시오.

IBM SPSS Statistics 모델

컴퓨터에 IBM SPSS Statistics 사본이 설치되고 사용이 허가된 경우에는 IBM SPSS Modeler 내에서 특정 IBM SPSS Statistics 루틴에 액세스하고 실행하여 모델을 작성하고 스코어링할 수 있습니다.

(2) 분할 모델 작성

분할 모델링을 사용하면 단일 스트림을 사용하여 플래그, 명목 또는 연속형 입력 필드의 가능한 각 값에 대해 별도의 모델을 작성할 수 있습니다. 결과로 생성되는 모델은 모두 단일 모델 너짓에서 액세스할 수 있습니다. 입력 필드의 가능한 값은 모델에 매우 다른 효과를 줄 수 있습니다. 분할 모델링을 사용하면 스트림의 단일 실행으로 가능한 각 필드의 최적 적합 모델을 쉽게 작성할 수 있습니다.

대화형 모델링 세션은 분할을 사용할 수 없습니다. 대화형 모델링에서는 각 모델을 개별적으로 지정하므로, 분할 사용 시 장점은 없으며, 여러 모델이 자동으로 작성됩니다.

특정 입력 필드를 분할 필드로 지정하여 모델링 작업을 분할합니다. 유형 지정에서 필드 역할을 **분할**로 설정하여 이를 수행할 수 있습니다.

측정 수준이 **플래그**, **명목**, **순서** 또는 **연속**인 필드만 분할 필드로 지정할 수 있습니다.

분할 필드로 둘 이상의 입력 필드를 지정할 수 있습니다. 그러나 이 경우 작성된 모델 수가 약간 증가할 수 있습니다. 모델은 선택한 분할 필드 값의 모든 가능한 결합에 대해 작성됩니다. 예를 들어, 각각 3개의 가능한 값을 포함하는 3개의 입력 필드가 분할 필드로 지정된 경우 이로 인해 27개의 서로 다른 모델이 작성될 수 있습니다.

분할 필드로 하나 이상의 필드를 지정한 후에도 계속해서 모델링 노드 대화 상자의 선택란 설정을 통해 단일 모델을 작성할 것인지, 아니면 분할 모델을 작성할 것인지 선택할 수 있습니다.

분할 필드가 정의되었지만, 선택란이 선택되지 않은 경우 단일 모델만 생성됩니다. 마찬가지로 선택란을 선택했지만, 분할 필드가 정의되지 않은 경우 분할은 무시되고 단일 모델이 생성됩니다.

스트림을 실행하면 하나 이상의 분할 필드의 가능한 각 값에 대해 이면에서 별도의 모델이 작성되지만, 모델 팔레트와 스트림 캔버스에는 단일 모델 너깃만 배치됩니다. 분할 모델 너깃은 분할 기호로 표시됩니다. 이 기호는 너깃 이미지에 2개의 회색 사각형이 오버레이되어 나타납니다.

분할 모델 너깃을 찾아보는 경우 작성된 모든 개별 모델의 목록이 나타납니다.

뷰어에서 해당 너깃 아이콘을 두 번 클릭하여 목록에서 개별 모델을 조사할 수 있습니다. 그러면 개별 모델에 대해 표준 브라우저 창이 열립니다. 너깃이 캔버스에 있을 때 그래프 썸네일을 두 번 클릭하면 전체 크기 그래프가 열립니다. 자세한 정보는 분할 모델 뷰어 주제를 참조하십시오.

모델을 분할 모델로 작성하면 여기에서 분할 처리를 제거할 수 없으며, 분할 모델링 노드 또는 너깃에서 다운스트림을 추가로 분할하는 작업을 실행 취소할 수 없습니다.

예. 국내 소매업체에서 자국 주변의 각 매장에서 제품 범주별로 판매를 추정하려고 합니다. 분할 모델링을 사용하여 입력 데이터의 매장 필드를 분할 필드를 지정하고, 이를 통해 단일 작업으로 각 매장에서 각 범주에 대해 별도의 모델을 작성할 수 있습니다. 그러면 결과로 생성된 정보를 사용하여 단일 모델에서 얻을 수 있는 것보다 훨씬 더 정확하게 재고 수준을 제어할 수 있습니다.

① 분할 및 파티셔닝

분할에는 파티셔닝에 공통된 몇 가지 기능이 있지만, 두 개는 서로 다른 방식으로 사용됩니다.

파티셔닝은 데이터 세트를 무작위로 두 개 또는 세 개의 파트(훈련, 검정, 선택적으로 검증)로 구분하며, 단일 모델의 성능을 검정하는 데 사용됩니다.

분할은 분할 필드의 가능한 값이 있는 만큼 많은 파트로 데이터 세트를 구분하며, 다중 모델을 작성하는 데 사용됩니다.

파티셔닝 및 분할은 서로 완전히 독립적으로 작동합니다. 모델링 노드에서 둘 다 선택하거나 둘 다 선택하지 않을 수 있습니다.

② 분할 모델을 지원하는 모델링 노드

많은 모델링 노드에서 분할 모델을 작성할 수 있습니다. 예외로는, 자동 군집, PCA/요인, 필드선택, SLRM, 임의 트리, Tree-AS, Linear-AS, LSVM, 연관 모델(Apriori, Carma, 시퀀스), 군집 모델(K-평균, 코호넨, 이단계, 이상 항목), Statistics 모델, In-Database 모델링에 사용된 노드가 있습니다.

분할 모델링을 지원하는 모델링 노드는 다음과 같습니다.

	C&R트리		BayesNet		선형
	QUEST		GenLin		GLMM
	CHAID		KNN		STP
	C5.0		Cox		One-ClassSVM
	신경망		자동분류자		XGBoostTree
	의사결정목록		자동숫자		XGBoostLinear
	회귀분석		로지스틱		HDBSCAN
	판별		SVM		시계열

③ 분할 영향을 받는 기능

분할 모델의 사용은 다양한 방식으로 여러 IBM® SPSS® Modeler 기능에 영향을 줍니다. 이 절에서는 스트림의 다른 노드와 함께 분할 모델을 사용하는 방법에 대한 지침을 제공합니다.

레코드 Ops 노드

표본 노드를 포함하는 스트림에서 분할 모델을 사용하는 경우 레코드의 고른 표본추출을 달성하기 위해 분할 필드로 레코드를 층화합니다. 이 옵션은 표본추출 방법으로 복잡을 선택한 경우에 사용 가능합니다.

스트림이 균형 노드를 포함하면 균형이 분할 안에 있는 레코드의 서브셋이 아닌, 입력 레코드의 전체 세트에 적용됩니다.

통합 노드를 통해 레코드를 통합하는 경우 각 분할에 대한 통합을 계산하려면 분할 필드를 키 필드로 설정하십시오.

필드 Ops 노드

유형 노드는 분할 노드로 사용할 하나 이상의 필드를 지정하는 위치입니다.

참고: 앙상블 노드는 둘 이상의 모델 너깃을 결합하는 데 사용되지만, 분할 모델은 단일 모델 너깃에 포함되므로 분할 조치를 반전시키는 데 앙상블 노드를 사용할 수 없습니다.

모델링 노드

분할 모델은 예측자 중요도의 계산(모델 추정 시 예측자 입력 필드의 상대적 중요도)을 지원하지 않습니다. 예측자 중요도 설정은 분할 모델을 작성할 때 무시됩니다.

참고: 수정된 성향 스코어 설정은 분할 모델을 사용할 때 무시됩니다.

KNN(최근접 이웃) 노드는 목표 필드를 예측하도록 설정된 경우에만 분할 모델을 지원합니다. 대체 설정(최근접 이웃만 식별)은 모델을 작성하지 않습니다. **k 자동 선택** 옵션이 선택된 경우 각 분할 모델에는 서로 다른 수의 최근접 이웃이 포함될 수 있습니다. 따라서 전체 모델은 모든 분할 모델에서 찾은 가장 많은 최근접 이웃 수와 동일한 수의 생성된 열을 포함합니다. 최근접 이웃 수가 이 최대값보다 적은 분할 모델에서는 대응하는 수만큼의 열이 \$null 값으로 채워집니다. 자세한 정보는 KNN 노드 주제를 참조하십시오.

데이터베이스 모델링 노드

In-Database 모델링 노드는 분할 모델을 지원하지 않습니다.

모델 너깃

분할 모델 너깃에서 **PMML로 내보내기**는 너깃이 다중 모델을 포함하고 PMML이 패키지 등을 지원하지 않으므로 불가능합니다. 텍스트 또는 HTML로 내보낼 수 있습니다.

(3) 모델링 노드 필드 옵션

모든 모델링 노드에는 필드 탭이 있으며, 여기에서 모델 작성 시 사용할 필드를 지정할 수 있습니다.

모델을 작성하려면 먼저 목표 및 입력으로 사용할 필드를 지정해야 합니다. 몇 가지 예외가 있지만, 모든 모델링 노드는 업스트림 유형 노드에서 필드 정보를 사용합니다. 유형 노드를 사용하여 입력 및 목표 필드를 선택하는 경우 이 탭의 내용을 변경하지 않아도 됩니다. (예외로는, 모델링 노드에 필드 설정을 지정해야 하는 시퀀스 노드 및 텍스트 추출 노드가 포함됩니다.)

유형 노드 설정 사용. 이 옵션은 업스트림 유형 노드의 필드 정보를 사용하도록 노드에 지시합니다. 이는 기본값입니다.

사용자 정의 설정 사용. 이 옵션에서는 업스트림 유형 노드에 지정된 항목 대신, 여기에 지정된 필드 정보를 사용하도록 노드에 지시합니다. 이 옵션을 선택한 후 필요하면 아래 필드를 지정합니다.

참고: 모든 노드에서 모든 필드가 표시되지는 않습니다.

- **트랜잭션 형식 사용(Apriori, CARMA, MS 연관 규칙, Oracle Apriori 노드만 해당).** 소스 데이터가 **트랜잭션 형식**인 경우 이 선택란을 선택합니다. 이 형식의 레코드는 2개 필드(ID와 컨테츠에 대해 각각 하나씩)를 포함합니다. 각 레코드는 단일 트랜잭션 또는 항목을 나타내고, 연관된 품목은 동일한 ID를 보유하여 링크됩니다. 데이터가 **표 형식**인 경우 이 상자를 선택 취소합니다. 이 경우 항목은 별도의 플래그로 표시되며, 각 플래그 필드는 특정 항목의 존재 여부를 나타내고, 각 레코드는 연관된 항목의 전체 세트를 나타냅니다. 자세한 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

- **ID.** 트랜잭션 데이터의 경우 목록에서 ID 필드를 선택합니다. 숫자 또는 기호 필드를 ID 필드로 사용할 수 있습니다. 이 필드의 각 고유한 값은 분석의 특정 단위를 표시해야 합니다. 예를 들어, 장바구니 애플리케이션에서 각 ID는 단일 고객을 나타낼 수 있습니다. 웹 로그 분석 애플리케이션의 경우 각 ID는 컴퓨터(IP 주소별) 또는 사용자(로그인 데이터별)를 나타낼 수 있습니다.
- **연속적 ID.** (Apriori 및 CARMA 노드만) ID가 동일한 모든 레코드를 데이터 스트림에서 함께 그룹화하도록 데이터를 사전 정렬한 경우 이 옵션을 선택하여 처리 속도를 높입니다. 데이터가 사전 정렬되지 않았거나 정렬 여부가 확실하지 않은 경우 이 옵션을 선택하지 않은 상태로 두면 노드가 데이터를 자동으로 정렬합니다.

 **참고:** 데이터가 정렬되지 않은 상태에서 이 옵션을 선택하면 모델에서 유효하지 않은 결과가 발생할 수 있습니다.

- **내용.** 모델의 내용 필드를 지정합니다. 이 필드는 연관 모델링에서 관심이 있는 항목을 포함합니다. 다중 플래그 필드(데이터가 표 형식인 경우) 또는 단일 명목 필드(데이터가 트랜잭션 형식인 경우)를 지정할 수 있습니다.
- **목표.** 하나 이상의 목표 필드가 필요한 모델의 경우 하나 이상의 목표 필드를 선택합니다. 유형 노드에서 필드 역할을 목표로 설정하는 것과 유사합니다.
- **평가.** (자동 군집 모델만 해당.) 군집 모델에는 목표가 지정되지 않지만, 평가 필드를 선택하여 해당 중요도 수준을 식별할 수 있습니다. 또한 군집이 이 필드의 값을 구별하는 정도를 평가할 수 있습니다. 그러면 차례로 이 필드를 예측하는 데 군집을 사용할 수 있는지 여부를 표시합니다. 참고 평가 필드는 둘 이상의 값을 포함하는 문자열이어야 합니다.
 - **입력** 하나 이상의 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 **입력**으로 설정하는 것과 유사합니다.

- **파티션.** 이 필드에서는 모델 작성에 대한 훈련, 검정, 검증 단계에 대한 별도의 샘플로 데이터를 파티션하는 데 사용되는 필드를 지정할 수 있습니다. 한 표본을 사용하여 모델을 생성하고 다른 표본을 사용하여 이를 검증함으로써 현재 데이터에 유사한 보다 큰 데이터 세트로 모델이 일반화되는 정도를 잘 표시할 수 있습니다. 유형 또는 파티션 노드를 사용하여 다중 파티션 필드를 정의한 경우 파티셔닝을 사용하는 각 모델링 노드에 있는 필드 탭에서 단일 파티션 필드를 선택해야 합니다. (하나의 파티션만 존재하는 경우 파티션이 사용될 때마다 자동으로 사용됩니다.) 또한 분석에서 선택한 파티션을 적용하려면 노드의 모델 옵션 탭에서 파티셔닝을 사용 가능하게 해야 합니다. (이 옵션을 선택 취소하면 필드 설정을 변경하지 않고도, 파티셔닝을 사용하지 않을 수 있습니다.)
- **분할.** 분할 모델의 경우 단일 또는 복수 분할 필드를 선택하십시오. 이는 유형 노드에서 필드 역할을 *분할*로 설정하는 것과 유사합니다. 측정 수준이 **플래그**, **명목**, **순서** 또는 **연속**인 필드만 분할 필드로 지정할 수 있습니다. 분할 필드로 선택된 필드는 목표, 입력, 파티션, 빈도 또는 가중 필드로 사용할 수 없습니다. 자세한 정보는 분할 모델 작성 주제를 참조하십시오.
- **빈도 필드 사용.** 이 옵션에서는 빈도 가중치로 필드를 선택할 수 있습니다. 훈련 데이터의 레코드가 각각 둘 이상의 단위를 나타내는 경우(예: 통합 데이터를 사용하는 경우) 이를 사용합니다. 필드 값은 각 레코드로 나타낸 노드 수여야 합니다. 자세한 정보는 빈도 및 가중 필드 사용 주제를 참조하십시오.

참고: 메타데이터가 유효하지 않음(입력/출력 필드에서) 오류 메시지가 나타나면 필요한 모든 필드(예: 빈도 필드)를 지정했는지 확인합니다.

- **가중 필드 사용.** 이 옵션에서는 케이스 가중치로 필드를 선택할 수 있습니다. 케이스 가중치는 출력 필드의 수준에서 분산의 차이를 설명하는 데 사용됩니다. 자세한 정보는 빈도 및 가중 필드 사용 주제를 참조하십시오.
- **후향.** 규칙 귀납 노드(Apriori)의 경우 결과로 생성된 규칙 세트에서 후향으로 사용할 필드를 선택합니다. (이는 유형 노드에서 역할이 *목표* 또는 *둘 다*인 필드에 대응합니다.)
- **전향.** 규칙 귀납 노드(Apriori)의 경우 결과로 생성된 규칙 세트에서 전향으로 사용할 필드를 선택합니다. (이는 유형 노드에서 역할이 *입력* 또는 *둘 다*인 필드에 대응합니다.)

일부 모델은 이 절에서 설명한 것과 다른 필드 탭이 있기도 합니다.

- 자세한 정보는 시퀀스 노드 필드 옵션 주제를 참조하십시오.
- 자세한 정보는 CARMA 노드 필드 옵션 주제를 참조하십시오.

① 빈도 및 가중 필드 사용

빈도 및 가중 필드는 일부 레코드를 다른 레코드와 비교했을 때 추가 중요도를 부과하는 데 사용됩니다. 예를 들어, 모집단의 한 섹션에서 훈련 데이터의 표본이 부족하다는 점(가중치)을 알고 있거나 한 레코드가 여러 동일 케이스를 나타내는 경우(빈도)가 이에 해당합니다.

- 빈도 필드의 값은 양의 정수여야 합니다. 케이스 빈도가 음수 또는 0인 레코드는 분석에서 제외됩니다. 정수가 아닌 빈도 가중치는 가장 근사한 정수로 수정됩니다(올림/내림).
- 케이스 가중치 값은 양수여야 하지만 정수 값은 아니어도 됩니다. 케이스 가중치가 음수 또는 0인 레코드는 분석에서 제외됩니다.

빈도 및 가중 필드 스코어링

빈도 및 가중 필드는 훈련 모델에 사용되지만, 스코어링에는 사용되지 않습니다. 각 레코드의 스코어는 나타내는 케이스 수에 상관없이 해당 특성에 기반하기 때문입니다. 예를 들어, 다음 표에 데이터가 있습니다.

표 1. 데이터 예

기혼	응답됨
예	예
예	예
예	예
예	아니오
아니오	예
아니오	아니오
아니오	아니오

이에 기반하여 4명의 기혼자 중 3명이 프로모션에 반응했으며, 3명의 미혼자 중 2명은 반응하지 않았다고 결론을 내릴 수 있습니다. 따라서 다음 표에 표시된 대로, 새 레코드 스코어를 적절히 계산할 수 있습니다.

표 2. 스코어 계산된 레코드 예

기혼	\$-Responded	\$RP-Responded
예	예	0.75(3/4)
아니오	아니오	0.67(2/3)

또는 다음 표에 표시된 대로 빈도 필드를 사용하여 훈련 데이터를 더 축약해서 저장할 수 있습니다.

표 3. 스코어 계산된 레코드 대체 예		
기혼	응답됨	빈도
예	예	3
예	아니오	1
아니오	예	1
아니오	아니오	2

이는 정확히 동일한 데이터 세트를 나타내므로, 결혼상태에만 기반하여 동일한 모델을 작성하고 반응을 예측합니다. 스코어링 데이터에서 기혼인 사람이 10명인 경우 별도의 10개 레코드로 표시되거나, 빈도 값인 10인 하나의 레코드로 표시되는지 여부에 상관없이 각각에 예의 반응을 예측합니다. 가중치(일반적으로 정수가 아님)는 마찬가지로 레코드 중요도를 표시하는 항목으로 간주할 수 있습니다. 그래서 레코드 스코어링에서 빈도 및 가중 필드를 사용하지 않는 것입니다.

모델 평가 및 비교

일부 모델 유형은 빈도 필드를 지원하고, 일부는 가중 필드를 지원하고, 일부는 둘 다 지원합니다. 그러나 이들이 적용되는 모든 케이스에서 이들은 모델을 작성할 때만 사용되며, 평가 노드 또는 분석 노드를 사용하여 모델을 평가하거나 자동 분류자 및 자동 숫자 노드에서 지원하는 대부분의 방법을 사용하여 모델을 순위화할 때 고려되지 않습니다.

- 예를 들어, 평가 차트로 모델을 비교할 때 빈도와 가중값은 무시됩니다. 이를 통해 이러한 필드를 사용하는 모델과 사용하지 않는 모델 사이에서 수준을 비교합니다. 그러나 정확한 평가를 위해 빈도 또는 가중 필드에 의존하지 않고도 모집단을 정확히 나타내는 데이터 세트를 사용해야 함을 의미합니다. 실제로 빈도 또는 가중 필드의 값이 항상 널 또는 1인 검정 표본을 사용하여 모델을 평가하도록 보장하여 이를 수행할 수 있습니다. (이러한 제한은 모델을 평가할 때만 적용됩니다. 빈도 또는 가중값이 훈련 및 검정 표본 모두에서 항상 1이면 처음부터 이러한 필드를 사용할 이유가 없습니다.)
- 자동 분류자를 사용하는 경우 이익에 기반하여 모델을 순위화할 때 빈도를 고려할 수 있으므로, 이 방법이 이 케이스에 권장됩니다.
- 필요한 경우 파티션 노드를 사용하여 데이터를 훈련 및 검정 표본으로 분할할 수 있습니다.

(4) 모델링 노드 분석 옵션

많은 모델링 노드가 원래 및 수정된 성향 스코어와 함께 예측자 중요도 정보를 확보할 수 있는 분석 탭을 포함합니다.

모형 평가

예측자 중요도 계산. 적절한 중요도 측도를 생성하는 모델의 경우 모델 측정 시 각 예측자의 상대적 중요도를 나타내는 차트를 표시할 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 예측자 중요도는 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 몇몇 모델은 기본적으로 해제되어 있음에 유의하십시오. 예측자 중요도는 의사결정 목록 모델에 사용할 수 없습니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

성향 스코어

성향 스코어는 모델링 노드 및 모델 너깃의 설정 탭에서 사용 가능합니다. 이 기능은 선택한 목표가 플래그 필드인 경우에만 사용 가능합니다. 자세한 정보는 성향 스코어의 내용을 참조하십시오.

원시 성향 스코어 계산. 원시 성향 스코어는 학습 데이터에만 기반하여 모델에서 파생됩니다. 모델이 참 값(응답함)을 예측하면 성향은 P 와 동일합니다. 여기서 P 는 예측 확률입니다. 모델이 거짓 값을 예측하면 성향은 $(1 - P)$ 로 계산됩니다.

- 모델 작성 시 이 옵션을 선택한 경우 기본적으로 모델 너깃에서 성향 스코어가 사용 가능합니다. 그러나 모델링 노드에서 선택 여부에 상관없이 언제나 모델 너깃에서 원시 성향 스코어를 사용하도록 선택할 수 있습니다.
- 모델 스코어링 시 원시 성향 스코어는 표준 접두문자에 문자 RP 가 추가된 필드에 추가됩니다. 예를 들어 예측이 이름이 $\$R-churn$ 인 필드에 있는 경우 성향 스코어 필드 이름은 $\$RRP-churn$ 입니다.

수정된 성향 스코어 계산. 원시 성향은 모델에서 제공된 추정값에만 기반하며, 과적합할 경우 성향의 지나친 낙관적 추정값으로 이어질 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 수행 방법을 보고 적절히 더 나은 추정값을 제공하도록 성향을 조정하여 보완하려고 합니다.

- 이 설정에서는 유효한 파티션 필드가 스트림에 존재해야 합니다.
- 원시 신뢰도 스코어와 달리, 수정된 성향 스코어는 모델 작성 시 계산해야 합니다. 그렇지 않으면 모델 너깃 스코어링에서 사용 불가능합니다.
- 모델 스코어링 시 수정된 성향 스코어는 표준 접두문자에 문자 AP 가 추가된 필드에 추가됩니다. 예를 들어 예측이 이름이 $\$R-churn$ 인 필드에 있는 경우 성향 스코어 필드 이름은 $\$RAP-churn$ 입니다. 수정된 성향 스코어는 로지스틱 회귀분석 모델에서 사용할 수 없습니다.
- 수정된 성향 스코어를 계산할 때 계산에 사용된 검정 또는 검증 파티션은 균형을 맞출 수 없습니다. 이를 방지하려면 업스트림 균형 노드에서 **균형 학습 데이터만** 옵션을 선택해야 합니다. 또한 복잡한 샘플에서 업스트림을 사용하는 경우 이는 수정된 성향 스코어를 무효화합니다.
- 수정된 성향 스코어는 "증폭된" 트리 및 규칙 세트 모델에서 사용할 수 없습니다. 자세한 정보는 부스팅 C5.0 모델의 내용을 참조하십시오.

다음은 기준. 수정된 성향 스코어를 계산할 경우 파티션 필드가 스트림에 존재해야 합니다. 이 계산에서 검정 또는 검증 파티션 중 사용할 항목을 지정할 수 있습니다. 최상의 결과를 얻으려면 검정 또는 검증 파티션은 원래 모델을 학습시키는 데 사용되는 파티션만큼 많은 레코드를 최소한으로 포함해야 합니다.

① 성향 스코어

예 또는 아니오 예측을 리턴하는 모델의 경우 표준 예측 및 신뢰도 값 외에도 성향 스코어를 요청할 수 있습니다. 성향 스코어는 특정 결과 또는 반응의 우도를 나타냅니다. 다음 표에는 예가 포함되어 있습니다.

표 1. 성향 스코어	
고객	응답 성향
Joe Smith	35%
Jane Smith	15%

성향 스코어는 플래그 목표를 포함하는 모델에서만 사용 가능하고, 소스 또는 유형 노드에 지정된 대로, 필드에 정의된 *참* 값의 우도를 나타냅니다.

성향 스코어 대 신뢰도 스코어

성향 스코어는 예 또는 아니오로 현재 예측에 적용되는 신뢰도 스코어와는 다릅니다. 예측이 *아니오*인 경우 예를 들어, 신뢰도가 높으면 실제로 반응하지 *않을* 우도가 높습니다. 성향 스코어는 모든 레코드에서 더 쉽게 비교할 수 있도록 이 제한을 우회합니다. 예를 들어 신뢰도가 0.85인 *아니오* 예측은 0.15의 원시 성향(또는 $1 - 0.85$)으로 변환됩니다.

표 2. 신뢰도 스코어		
고객	예측	신뢰도
Joe Smith	응답함	0.35
Jane Smith	응답하지 않음	0.85

성향 스코어 확보

- 성향 스코어는 모델링 노드의 분석 탭 또는 모델 너깃의 설정 탭에서 사용 가능합니다. 이 기능은 선택한 목표가 플래그 필드인 경우에만 사용 가능합니다. 자세한 정보는 모델링 노드 분석 옵션의 내용을 참조하십시오.
- 성향 스코어는 사용된 앙상블 방법에 따라 앙상블 노드에서도 계산할 수 있습니다.

수정된 성향 스코어 계산

수정된 성향 스코어는 모델 작성 프로세스 중에 계산되며, 그 외의 경우에는 사용할 수 없습니다. 모델을 작성하면 검정 또는 검증 파티션의 데이터를 사용하여 스코어를 계산하고, 해당 파티션에서 원래 모델의 성능을 분석하여 수정된 성향 스코어를 전달하는 새 모델을 구성합니다. 모델 유형에 따라 두 개 방법 중 하나를 사용하여 수정된 성향 스코어를 계산할 수 있습니다.

- 규칙 세트 및 트리 모델의 경우 수정된 성향 스코어는 트리 모델인 경우 각 트리 노드 또는 규칙 세트 모델인 경우 각 규칙의 지원 및 신뢰도에서 각 범주의 빈도를 재계산하여 생성됩니다. 그러면 원래 모델에 저장되는 새 규칙 세트 또는 트리 모델이 수정된 성향 스코어를 요청할 때마다 사용됩니다. 원래 모델을 새 데이터에 적용할 때마다 새 모델은 후속으로 원시 성향 스코어에 적용되어 조정된 스코어를 생성할 수 있습니다.
- 다른 모델의 경우 검정 또는 검증 파티션에서 원래 모델 스코어를 계산하여 생성된 레코드는 원시 성향 스코어로 구간화됩니다. 그런 다음, 각 구간의 평균 원시 성향에서 동일한 구간의 관측된 평균 성향으로 매핑되는 비선형 함수를 정의하는 신경망 모델이 훈련됩니다. 트리 모델에 대해 앞서 언급한 대로, 결과로 생성되는 신경망 모델은 원래 모델에 저장되고 수정된 성향 스코어를 요청할 때마다 원시 성향 스코어에 적용할 수 있습니다.

검정 분할에서 결측값 관련 주의. 검정/검증 파티션에서 결측 입력값을 처리 방법은 모델에 따라 달라집니다(자세한 정보는 개별 모델 스코어링 알고리즘 참조). C5 모델은 결측 입력이 있는 경우 조정된 성향을 계산할 수 없습니다.

(5) 오분류 비용

일부 컨텍스트에서는 특정 오차 유형이 다른 유형에 비해 더 값비쌉니다. 예를 들어, 고위험 신용 거래 신청자를 저위험(오차의 한 유형)으로 분류하는 것이 저위험 신청자를 고위험(다른 유형의 오차)으로 분류하는 것보다 비용이 더 나갈 수 있습니다. 오분류 비용을 통해 여러 다른 유형의 예측 오차의 상대적 중요도를 지정할 수 있습니다.

오분류 비용은 기본적으로 특정 결과에 적용된 가중값입니다. 이 가중값은 모델에 영향을 미치는 요인이 되어 실제로 예측변수를 변경합니다(값비싼 실수에 대비하는 보호의 한 방법으로).

C5.0 모델을 제외하고, 오분류 비용은 모델을 스코어링할 때 적용되지 않으며 자동 분류자 노드, 평가 노드 또는 분석 노드를 사용하여 모델을 순위화하거나 비교할 때 고려되지 않습니다. 비용을 포함한 모델은 포함하지 않은 모델보다 적은 오차를 생성하지 않고 전체 정확도 측면에서 더 높게 순위화되지 않을 수는 있지만, *보다 저렴* 오차를 선호하는 편향이 내재되어 있어서 실제적으로 성능이 더 좋을 수 있습니다.

비용 교차표는 예측 범주와 실제 범주의 가능한 각 조합의 비용을 표시합니다. 기본적으로 모든 오분류 비용은 1.0으로 설정됩니다. 사용자 정의 비용 값을 입력하려면 **오분류 비용 사용**을 선택하고 비용 교차표에 사용자 정의 값을 입력하십시오.

오분류 비용을 변경하려면 예측 및 실제 값의 원하는 조합에 해당하는 셀을 선택하고 셀의 기존 콘텐츠를 삭제한 후 셀의 원하는 비용을 입력하십시오. 비용은 자동으로 대칭되지 않습니다. 예를 들어, A를 B로 오분류한 비용을 2.0으로 설정할 경우 이 설정을 명시적으로 변경하지 않으면 B를 A로 오분류하는 비용의 기본값은 여전히 1.0입니다.

참고: 의사결정 트리 모형에서만 작성 시 비용을 지정할 수 있습니다.

(6) 모델 너깃

그림 1. 모델 너깃



모델 너깃은 모델의 컨테이너입니다. 즉, SPSS® Modeler에서 모델 작성 작업의 결과를 나타내는 규칙, 수식 또는 방정식의 세트입니다. 너깃의 주 목적은 예측을 생성하거나 모델 특성의 추가 분석을 사용 가능하게 하기 위해 데이터 스코어를 계산하는 것입니다. 화면에서 모델 너깃을 열면 모델에 대한 다양한 세부사항(예: 모델 작성 시 입력 필드의 상대적 중요도)을 확인할 수 있습니다. 예측을 보려면 추가 프로세스 또는 출력 노드를 첨부하고 실행해야 합니다. 자세한 정보는 스트림에서 모델 너깃 사용의 내용을 참조하십시오.

그림 2. 모델링 노드에서 모델 너깃으로의 모델 링크



모델링 노드를 성공적으로 실행하면 대응하는 모델 너깃이 스트림 캔버스에 배치됩니다. 이는 다이아몬드 형태의 금색 아이콘(그래서 "너깃"이라고 함)으로 표시됩니다. 스트림 캔버스에서 너깃은 모델링 노드 이전에 가장 근접한 적절한 노드에 대한 연결(실선)과 모델링 노드 자체에 대한 링크(점선)로 표시됩니다.

너깃은 IBM® SPSS Modeler 창의 오른쪽 상단 코너에 있는 모델 팔레트에도 배치됩니다. 두 위치 어디에서든 모델의 세부사항을 보기 위해 너깃을 선택하고 찾아볼 수 있습니다.

너깃은 모델링 노드가 성공적으로 실행되면 항상 모델 팔레트에 배치됩니다. 스트림 캔버스에 추가로 너깃을 배치할지 여부를 제어하도록 사용자 옵션을 설정할 수 있습니다.

다음 주제에서는 IBM SPSS Modeler에서 모델 너깃 사용에 대한 정보를 제공합니다. 사용된 알고리즘을 자세히 이해하려면 제품 다운로드에서 PDF 파일로 제공되는 *IBM SPSS Modeler 알고리즘 안내서*를 참조하십시오.

① 모델 링크

그림 1. 모델링 노드에서 모델 너깃으로의 모델 링크



기본적으로 너깃은 이를 작성한 모델링 노드에 대한 링크를 포함하는 캔버스에 표시됩니다. 특히, 각 모델링 노드에서 업데이트되는 너깃을 식별하여, 여러 너깃을 포함하는 복잡한 스트림에 유용합니다. 각 링크는 모델링 노드를 실행할 때 모델을 바꿀 것인지 표시하기 위해 기호를 포함합니다. 자세한 정보는 모델 교체의 내용을 참조하십시오.

가. 모델 링크 정의 및 제거

캔버스에서 수동으로 링크를 정의 및 제거할 수 있습니다. 새 링크를 정의하는 경우 커서가 링크 커서로 변경됩니다.

그림 1. 링크 커서



새 링크 정의(컨텍스트 메뉴)

1. 링크를 시작하려는 모델링 노드에서 마우스 오른쪽 단추를 클릭하십시오.
2. 컨텍스트 메뉴에서 **모델 링크 정의**를 선택하십시오.
3. 링크를 종료할 너깃을 클릭하십시오.

새 링크 정의(주 메뉴)

1. 링크를 시작하려는 모델링 노드를 클릭하십시오.
2. 주 메뉴에서 다음을 선택하십시오.
편집 > 노드 > 모델 링크 정의

3. 링크를 종료할 너깃을 클릭하십시오.

기존 링크 제거(컨텍스트 메뉴)

1. 링크의 끝에 있는 너깃을 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 **모델 링크 제거**를 선택하십시오.

또는,

1. 링크의 가운데에 있는 기호를 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 링크 제거를 선택하십시오.

기존 링크 제거(주 메뉴)

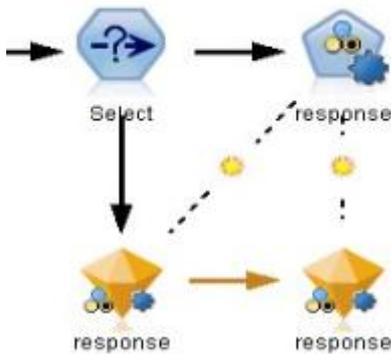
1. 링크를 제거하려는 모델링 노드 또는 너깃을 클릭하십시오.
2. 주 메뉴에서 다음을 선택하십시오.

편집 > 노드 > 모델 링크 제거

나. 모델 링크 복사 및 붙여넣기

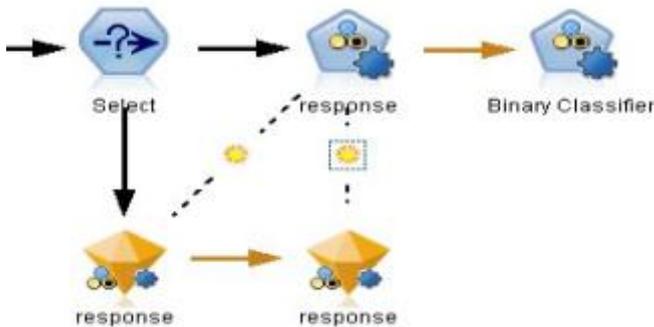
모델링 노드 없이 링크된 너깃을 복사하고 동일한 스트림에 붙여넣는 경우 모델링 노드에 대한 링크와 함께 너깃을 붙여넣습니다. 새 링크는 원래 링크와 동일한 모델 바꾸기 상태(모델 교체 참조)를 보유합니다.

그림 1. 링크된 너깃 복사 및 붙여넣기



링크된 모델링 노드와 함께 너깃을 복사하고 붙여넣는 경우 오브젝트를 붙여넣는 위치(동일한 스트림 또는 새 스트림)에 상관없이 링크가 유지됩니다.

그림 2. 링크된 너깃 복사 및 붙여넣기



참고: 모델링 노드 없이 링크된 너깃을 복사하고 너깃을 새 스트림 또는 모델링 노드를 포함하지 않는 슈퍼 노드에 붙여넣는 경우 링크가 끊어지고 너깃만 붙여넣습니다.

다. 모델 링크 및 슈퍼 노드

모델링 노드 또는 링크된 모델(둘 중 하나만)의 모델 너깃을 포함하도록 슈퍼 노드를 정의하는 경우 링크가 끊어집니다. 슈퍼 노드를 확장해도 링크는 복원되지 않습니다. 슈퍼 노드 작성을 실행 취소해야만 복원할 수 있습니다.

② 모델 교체

너깃을 작성한 모델링 노드의 재실행 시 기존 너깃을 교체(즉, 업데이트)할 것인지 여부를 선택할 수 있습니다. 교체 옵션을 끄면 모델링 노드를 재실행할 때 새 너깃이 작성됩니다.

모델링 노드에서 너깃까지의 각 링크는 모델링 노드를 재실행할 때 모델의 교체 여부를 표시하도록 기호를 포함합니다.

그림 1. 모델 교체가 설정된 모델 링크



링크는 모델 교체가 설정된 상태(링크의 작은 햇살 기호로 표시됨)로 처음에 표시됩니다. 이 상태에서 링크의 한쪽 끝에 있는 모델링 노드를 재실행하기만 하면 다른 끝에서 너깃이 업데이트됩니다.

그림 2. 모델 교체가 해제된 모델 링크



모델 교체를 끄면 링크 기호가 회색 점으로 바뀝니다. 이 상태에서 링크의 한쪽 끝에 있는 모델링 노드를 재실행하면 캔버스에 너깃의 새로 업데이트된 버전이 추가됩니다.

이 경우 모델 팔레트에서 **이전 모델 교체** 시스템 옵션 설정에 따라 기존 너깃이 업데이트되거나 새 너깃이 추가됩니다.

실행 순서

모델 너깃을 포함하는 다중 분기가 있는 스트림을 실행하는 경우 결과로 생성된 모델 너깃을 사용하는 분기보다 모델 교체가 설정된 분기를 실행하도록 스트림을 먼저 평가합니다.

요구 사항이 더 복잡한 경우 스크립트를 통해 실행 순서를 수동으로 설정할 수 있습니다.

모델 교체 설정 변경

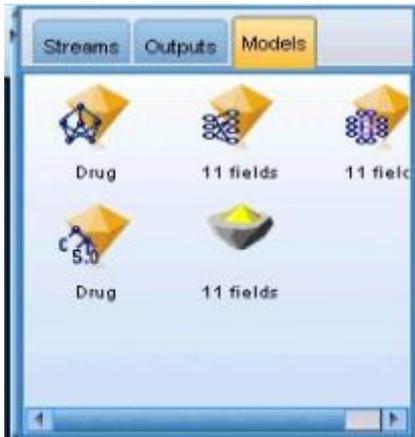
1. 링크에서 기호를 마우스 오른쪽 단추로 클릭하십시오.
2. 원하는 경우 **모델 교체 설정/해제**를 선택하십시오.

참고: 모델 링크에서 모델 교체 설정은 사용자 옵션 대화 상자의 알림 탭(도구 > 옵션 > 사용자 옵션)의 설정을 대체합니다.

③ 모델 팔레트

모델 팔레트(관리자 창의 모델 탭에 있음)에서는 다양한 방식으로 모델 너깃을 사용, 검사, 수정할 수 있습니다.

그림 1. 모델 팔레트



모델 팔레트에서 모델 너깃을 마우스 오른쪽 단추로 클릭하면 다음 옵션을 포함하는 컨텍스트 메뉴가 열립니다.

- **스트림에 추가.** 현재 활성 스트림에 모델 너깃을 추가합니다. 스트림에 선택한 노드가 있으면 해당 연결이 가능한 경우 모델 너깃은 선택한 노드에 연결되고, 그렇지 않으면 가능한 최근접 노드에 연결됩니다. 너깃은 스트림에 모델을 작성한 모델링 노드가 있는 경우 해당 노드에 대한 링크와 함께 표시됩니다.
- **찾아보기.** 너깃에 대한 모델 브라우저를 엽니다.
- **이름 변경 및 주석 작성.** 모델 너깃 이름을 변경하거나 너깃의 주석을 수정할 수 있습니다.
- **모델링 노드 생성.** 수정 또는 업데이트하려는 모델 너깃이 있고, 모델을 작성하는 데 사용된 스트림을 사용할 수 없는 경우 이 옵션을 사용하여 원래 모델을 작성하는 데 사용한 동일한 옵션으로 모델링 노드를 재생성할 수 있습니다.
- **모델 저장, 다른 이름으로 모델 저장.** 외부에서 생성된 모델(.gm) 이분형 파일에 모델 너깃을 저장합니다.
- **모델 보관.** 모델 너깃을 IBM® SPSS® Collaboration and Deployment Services Repository에 보관합니다.

- **PMML 내보내기.** 모델 너깃을 IBM SPSS Modeler 외부의 새 데이터 스코어링에 사용할 수 있는 예측 모델 마크업 언어(PMML)로 내보냅니다. **PMML 내보내기**는 생성된 모든 모델 노드에서 사용 가능합니다.
- **프로젝트에 추가.** 모델 너깃을 저장하고 현재 프로젝트에 추가합니다. 클래스 탭에서 너깃이 생성된 모델 폴더에 추가됩니다. CRISP-DM 탭에서는 기본 프로젝트 단계에 추가됩니다.
- **삭제.** 팔레트에서 모델 너깃을 삭제합니다.

모델 팔레트에서 차지하지 않은 영역을 마우스 오른쪽 단추로 클릭하면 다음 옵션을 포함하는 컨텍스트 메뉴가 열립니다.

- **모델 열기.** 이전에 IBM SPSS Modeler에서 작성한 모델 너깃을 로드합니다.
- **모델 검색.** IBM SPSS Collaboration and Deployment Services 리포지토리에서 저장된 모델을 검색합니다.
- **팔레트 로드.** 외부 파일에서 저장된 모델 팔레트를 로드합니다.
- **팔레트 검색.** IBM SPSS Collaboration and Deployment Services 리포지토리에서 저장된 모델 팔레트를 검색합니다.
- **팔레트 저장.** 외부에서 생성된 모델 팔레트(.gen) 파일에 모델 팔레트의 전체 내용을 저장합니다.
- **팔레트 보관.** 모델 팔레트의 전체 내용을 IBM SPSS Collaboration and Deployment Services 리포지토리에 보관합니다.
- **팔레트 지우기.** 팔레트에서 모든 너깃을 삭제합니다.
- **프로젝트에 팔레트 추가.** 모델 팔레트를 저장하고 현재 프로젝트에 추가합니다. 클래스 탭에서 너깃이 생성된 모델 폴더에 추가됩니다. CRISP-DM 탭에서는 기본 프로젝트 단계에 추가됩니다.
- **PMML 가져오기.** 외부 파일에서 모델을 로드합니다. IBM SPSS Statistics 또는 이 형식을 지원하는 기타 애플리케이션에서 작성된 PMML 모델을 열고, 찾아보고, 스코어를 계산할 수 있습니다. 자세한 정보는 모델을 PMML로 가져오기 및 내보내기의 내용을 참조하십시오.

④ 모델 너깃 찾아보기

모델 너깃 브라우저에서는 모델의 결과를 탐색하고 사용할 수 있습니다. 브라우저에서는 생성된 모델을 저장 또는 인쇄하거나 내보내고, 모델 요약을 탐색하고 모델의 주석을 보거나 편집할 수 있습니다. 일부 유형의 모델 너깃에서는 필터 노드 또는 규칙 세트 노드와 같은 새 노드를 생성할 수도 있습니다. 일부 모델의 경우 규칙 또는 군집 중심과 같은 모델 모수도 볼 수 있습니다. 일부 모델 유형의 경우(트리 기반 모델 및 군집 모델) 모델 구조에 대한 그래픽 표현을 볼 수도 있습니다. 모델 너깃 브라우저를 사용할 때 제어는 아래에서 설명합니다.

메뉴

파일 메뉴. 모든 모델 너깃에는 다음 옵션 중 일부 서브세트를 포함하는 파일 메뉴가 있습니다.

- **노드 저장.** 모델 너깃을 노드(nod) 파일에 저장합니다.
- **노드 보관.** 모델 너깃을 IBM® SPSS® Collaboration and Deployment Services 리포지토리에 보관합니다.
- **머리글 및 바닥글.** 너깃에서 인쇄할 때 페이지 머리글 및 바닥글을 편집할 수 있습니다.
- **페이지 설정.** 너깃에서 인쇄할 때 페이지 설정을 편집할 수 있습니다.
- **인쇄 미리보기.** 인쇄할 때 너깃의 표시 방법에 대한 미리보기를 표시합니다. 하위 메뉴에서 미리 보려는 정보를 선택합니다.
- **인쇄.** 너깃의 콘텐츠를 인쇄합니다. 하위 메뉴에서 인쇄하려는 정보를 선택합니다.
- **인쇄 보기.** 현재 보기 또는 모든 보기를 인쇄합니다.
- **텍스트 내보내기.** 너깃의 콘텐츠를 텍스트 파일로 내보냅니다. 하위 메뉴에서 내보내려는 정보를 선택합니다.
- **HTML 내보내기.** 너깃의 콘텐츠를 HTML 파일로 내보냅니다. 하위 메뉴에서 내보내려는 정보를 선택합니다.
- **PMML 내보내기.** 모델을 예측 모델 마크업 언어(PMML)로 내보냅니다. 그러면 다른 PMML 호환 소프트웨어에서 사용할 수 있습니다. 자세한 정보는 모델을 PMML로 가져오기 및 내보내기의 내용을 참조하십시오.
- **SQL 내보내기.** SQL(Structured Query Language)로 모델을 내보냅니다. 그러면 다른 데이터베이스에서 편집하고 사용할 수 있습니다.

 **참고:** SQL 내보내기는 다음 모델에서만 사용 가능합니다. C5, C&RT, CHAID, QUEST, 선형 회귀, 로지스틱 회귀분석, 신경망, PCA/요인, 의사결정 목록 모델.

- **UDF로 게시.** 설치된 스코어링 어댑터가 있는 데이터베이스로 모델을 게시합니다. 그러면 데이터베이스 내에서 모델 스코어링을 수행할 수 있습니다. 자세한 정보는 스코어링 어댑터에 대한 모델 게시의 내용을 참조하십시오.

생성 메뉴. 대부분의 모델 너깃에는 모델 너깃에 기반하여 새 노드를 생성할 수 있는 생성 메뉴도 있습니다. 이 메뉴에서 사용 가능한 옵션은 찾아보는 모델 유형에 따라 달라집니다. 특정 모델에서 생성할 수 있는 항목에 대한 세부사항을 보려면 특정 모델 너깃 유형을 참조하십시오.

보기 메뉴. 너깃의 모델 탭에서 이 메뉴를 사용하면 현재 모드에서 사용 가능한 다양한 시각화 도구 모음을 표시하거나 숨길 수 있습니다. 도구 모음의 전체 세트를 사용 가능하게 하려면 일반 도구 모음에서 편집 모드(붓 아이콘)를 선택합니다.

미리보기 단추. 일부 모델 너깃에는 미리보기 단추가 있습니다. 이를 통해 모델링 프로세스에서 작성된 추가 필드를 포함하여 모델 데이터 표본을 볼 수 있습니다. 표시되는 기본 행 수는 10입니다. 그러나 스트림 특성에서 이를 변경할 수 있습니다.

현재 프로젝트에 추가 단추. 모델 너깃을 저장하고 현재 프로젝트에 추가합니다. 클래스 탭에서 너깃이 생성된 모델 폴더에 추가됩니다. CRISP-DM 탭에서는 기본 프로젝트 단계에 추가됩니다.

⑤ 모델 너깃 요약/정보

모델 너깃의 요약 탭 또는 정보 보기에서는 필드, 작성 설정, 모델 추정 프로세스에 대한 정보를 표시합니다. 결과는 특정 항목을 클릭하여 펼치거나 접을 수 있는 트리 보기로 표시됩니다.

분석. 모델에 대한 정보를 표시합니다. 특정 세부사항은 모델 유형에 따라 달라지며, 각 모델 너깃의 섹션에서 다룹니다. 또한 이 모델링 노드에 첨부된 분석 노드를 실행하면 해당 분석의 정보도 이 섹션에 표시됩니다.

필드. 모델을 작성할 때 대상과 입력으로 사용되는 필드를 나열합니다. 분할 모델의 경우 분할을 판별하는 필드도 나열합니다.

 **참고:** 신경망 모델, 선형 모델 및 부스팅 또는 배경 모드를 사용하는 기타 모델에 대한 정보 보기에서는 유형이 플래그, 명목형 또는 순서인지 여부에 관계없이 표시되는 아이콘이 동일합니다.

작성 설정/옵션. 모델을 작성할 때 사용되는 설정에 대한 정보가 포함되어 있습니다.

훈련 요약. 모델 유형, 모델을 작성하는 데 사용되는 스트림, 모델을 작성한 사용자, 모델 작성 시점, 모델 작성 시 경과 시간을 표시합니다. 모델을 작성할 때 경과 시간은 정보 보기가 아니라 요약 탭에서만 사용 가능합니다.

⑥ 예측변수 중요도

일반적으로, 가장 중요한 예측자 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하기를 원합니다. 예측자 중요도 차트를 사용하면 모델 추정 시 각 예측자의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측자에 대한 값의 합은 1.0이 됩니다. 예측자 중요도는 모형 정확도와는 관련이 없습니다. 단지 예측 시 각 예측자의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

예측자 중요도는 신경망, 의사결정 트리(C&R 트리, C5.0, CHAID, QUEST), 베이지안 네트워크, 판별분석, SVM, SLRM 모델, 선형 및 로지스틱 회귀분석, 일반화 선형, 최근접 이웃(KNN) 모델을 포함하여 중요도의 적절한 통계 측도를 생성하는 모델에서 사용 가능합니다. 이러한 모델 대부분에서 예측자 중요도는 모델링 노드의 분석 탭에서 사용할 수 있습니다. 자세한 정보는 모델링 노드 분석 옵션의 내용을 참조하십시오. KNN 모델의 경우 이웃의 내용을 참조하십시오.

참고: 예측자 중요도는 분할 모델에서 지원되지 않습니다. 예측자 중요도 설정은 분할 모델을 작성할 때 무시됩니다. 자세한 정보는 분할 모델 작성 주제를 참조하십시오.

예측자 중요도 계산은 특히 대형 데이터 세트를 사용할 때 모델 작성보다 시간이 더 오래 걸릴 수 있습니다. SVM 및 로지스틱 회귀분석의 경우 다른 모델보다 계산이 더 오래 걸리고, 기본적

으로 이 모델에서는 사용되지 않습니다. 많은 예측자를 포함하는 데이터 세트를 사용하는 경우 필드선택 노드를 사용하는 초기 선별은 더 빠른 결과를 제공할 수 있습니다(아래 참조).

- 예측자 중요도는 사용 가능한 경우 검정 파티션에서 계산됩니다. 그렇지 않으면 훈련 데이터가 사용됩니다.
- SLRM 모델의 경우 예측자 중요도가 사용 가능하지만, SLRM 알고리즘에 의해 계산됩니다. 자세한 정보는 SLRM 모델 너깃 주제를 참조하십시오.
- IBM® SPSS® Modeler의 그래프 도구를 사용하여 그래프와 상호작용하고 그래프를 편집 및 저장할 수 있습니다.
- 선택적으로 예측변수 중요도 차트의 정보에 기반하여 필터 노드를 생성할 수 있습니다. 자세한 정보는 중요도에 기반하여 변수 필터링 주제를 참조하십시오.

예측자 중요도 및 필드선택

모델 너깃에 표시된 예측자 중요도 차트는 일부 경우에 필드선택 노드와 유사한 결과를 제공할 수도 있습니다. 필드선택이 지정된 목표에 대한 관계의 강도에 기반하여 다른 입력에 독립적으로 각 입력 필드를 순위화하는 반면, 예측자 중요도 차트는 이 특정 모델에 대한 각 입력의 상대적 중요도를 표시합니다. 따라서 필드선택은 선별 입력보다 보수적입니다. 예를 들어, *직위* 및 *작업 범주*가 모두 급여와 긴밀히 관련되어 있는 경우 필드선택은 둘 다 중요한 항목임을 표시합니다. 그러나 모델링에서 상호작용 및 상관관계도 고려합니다. 따라서 둘 다 동일한 정보의 많은 부분을 복제하는 경우 두 개의 입력 중 하나만 사용함을 알 수 있습니다. 실제로, 필드선택은 예비 선별에서, 특히 변수가 많은 큰 데이터 세트를 처리할 때 가장 유용하며, 예측자 중요도는 모델을 미세 조정할 때 가장 유용합니다.

단일 모델과 자동화된 모델링 노드 사이의 예측자 중요도 차이

개별 노드에서 단일 모델을 작성하는지, 아니면 자동화된 모델링 노드를 사용하여 결과를 생성하는지에 따라 예측자 중요도에서 약간의 차이가 발생할 수 있습니다. 구현에서 이러한 차이는 일부 엔지니어링 제한 때문에 발생합니다.

예를 들어, CHAID와 같은 단일 분류자를 사용하는 경우 계산은 중지 규칙을 적용하고 중요도 값을 계산할 때 확률 값을 사용합니다. 대조적으로 자동 분류자는 중지 규칙을 사용하지 않으며, 계산에서 예측 레이블을 직접 사용합니다. 이러한 차이는 자동 분류자를 사용하여 단일 모델을 생성하는 경우 단일 분류자에 대해 계산된 항목과 비교했을 때 중요도 값이 대략적인 추정으로 간주될 수 있음을 의미합니다. 보다 정확한 예측자 중요도 값을 얻기 위해 자동화된 모델링 노드 대신 단일 노드를 사용하도록 제안합니다.

가. 중요도에 기반하여 변수 필터링

선택적으로 예측변수 중요도 차트의 정보에 기반하여 필터 노드를 생성할 수 있습니다.

해당되는 경우 차트에서 포함할 예측변수를 표시하고 메뉴에서 다음을 선택하십시오.

생성 > 필터 노드(예측자 중요도)

또는

> 필드 선택(예측자 중요도)

최상위 변수. 가장 중요한 예측변수를 지정된 수까지 포함하거나 제외합니다.

다음보다 큰 중요도. 상대적 중요도가 지정된 값보다 큰 모든 예측변수를 포함하거나 제외합니다.

⑦ 앙상블 뷰어

가. 앙상블 모델

앙상블 모델은 앙상블의 구성요소 모델 및 전체로서 앙상블의 성능에 대한 정보를 제공합니다.

기본(보기 독립적) 도구 모음에서 스코어링에 대해 앙상블 모델을 사용할지 참조 모델을 사용할지 선택할 수 있습니다. 스코어링에 대해 앙상블이 사용되는 경우 결합 규칙도 선택할 수 있습니다. 이러한 변경 사항은 모델 재실행을 요구하지 않지만, 선택 사항이 스코어링 및/또는 다운스트림 모델 평가에 대해 모델 (덩어리)에 저장됩니다. 또한 앙상블 뷰어에서 내보낸 PMML에 영향을 미칩니다.

결합 규칙. 앙상블을 스코어링할 때 앙상블 스코어값을 계산하기 위해 기본 모델에서 예측값을 조합하는 데 사용되는 규칙입니다.

- **범주형** 목표에 대한 앙상블 예측값은 투표, 최고 확률 또는 최고 평균 확률을 사용하여 조합될 수 있습니다. **투표**는 기본 모델에서 최고 확률을 가지는 범주를 선택합니다. **최고 확률**은 모든 기본 모델에서 단일 최고 확률을 획득하는 범주를 선택합니다. **최고 평균 확률**은 범주 확률이 기본 모델에서 평균이 될 때 최고값이 있는 범주를 선택합니다.
- **연속형** 목표에 대한 앙상블 예측값은 기본 모델의 예측값의 평균 또는 중앙값을 사용하여 조합될 수 있습니다.

기본값은 모델 작성 동안 만들어진 지정 사항에서 가져옵니다. 결합 규칙을 변경하면 모델 정확도가 다시 계산되고 모델 정확도의 모든 보기가 업데이트됩니다. 예측변수 중요도 차트도 업데이트됩니다. 스코어링에 대해 참조 모델을 선택한 경우 이 제어는 비활성화됩니다.

모든 결합 규칙 표시. 선택하는 경우, 모델 품질 차트에 사용 가능한 모든 결합 규칙의 결과가 표시됩니다. 또한 구성요소 모델 정확도 차트가 업데이트되어 각 투표 방법에 대한 참조선을 표시합니다.

ㄱ. 모델 요약(양상블 뷰어)

모델 요약 보기는 양상블 품질 및 다양성을 한 눈에 파악할 수 있도록 요약한 스냅샷입니다.

품질. 차트에 참조 모델 및 단순 모델과 비교하여 최종 모델의 정확도가 표시됩니다. 크게 표시되는 정확도가 더 나은 형식이며 "최상의" 모델이 최고 정확도를 가집니다. 범주형 목표의 경우, 정확도는 단순히 예측값이 관측값과 일치하는 레코드의 백분율입니다. 연속형 목표의 경우, 정확도는 예측의 절대 평균 오차와 예측값 범위(예측값의 절대값 평균 빼기 관측값)의 비율(최대 예측값 빼기 최소 예측값)을 1에서 뺀 값입니다.

배깅 양상블의 경우, 참조 모델은 전체 학습 파티션에 작성된 표준 모델입니다. 부스팅된 양상블의 경우, 참조 모델은 첫 번째 구성요소 모델입니다.

단순 모델은 모델이 작성되지 않은 경우 정확도를 나타내며 전형 범주에 모든 레코드를 할당합니다. 단순 모델은 연속형 목표에 대해 계산되지 않습니다.

다양성. 차트에 양상블을 작성하는 데 사용된 구성요소 모델 중에서 "의견의 다양성"이 표시됩니다. 크게 표시되는 것이 더 다양한 형식입니다. 이것은 기본 모델에서 예측이 얼마나 다양한지에 대한 척도입니다. 다양성은 부스팅된 양상블 모델에 대해 사용할 수 없으며 연속형 목표에 대해 표시되지 않습니다.

ㄴ. 예측변수 중요도(양상블 뷰어)

일반적으로, 가장 중요한 예측자 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하기를 원합니다. 예측자 중요도 차트를 사용하면 모델 추정 시 각 예측자의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측자에 대한 값의 합은 1.0이 됩니다. 예측자 중요도는 모형 정확도와는 관련이 없습니다. 단지 예측 시 각 예측자의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

예측변수 중요도는 모든 양상블 모델에 사용할 수 있는 것은 아닙니다. 예측변수 세트는 구성요소 모델에서 다양할 수 있지만, 적어도 하나의 구성요소 모델에서 사용된 예측변수에 대해 중요도가 계산될 수 있습니다.

㉔. 예측자 빈도(양상블 뷰어)

예측변수 세트는 모델링 방법의 선택 또는 예측변수 선택으로 인해 구성요소 모델에서 다양할 수 있습니다. 예측변수 빈도 그림은 양상블의 구성요소 모델에서 예측변수의 분포를 표시하는 점도표입니다. 각 점은 예측변수를 포함하는 하나 이상의 구성요소 모델을 나타냅니다. 예측변수는 Y축에 표시되며 빈도의 내림차순으로 정렬됩니다. 그러므로 맨 위의 예측변수는 구성요소 모델의 최대 수에 사용되는 예측변수이고 맨 아래의 예측변수는 최소 수에 사용되는 예측변수입니다. 상위 10개 예측변수가 표시됩니다.

가장 자주 나타나는 예측변수가 일반적으로 가장 중요합니다. 이 그림은 예측변수 세트가 구성요소 모델에서 다양하지 못한 방법에서는 유용하지 않습니다.

㉕. 구성요소 모델 정확도(양상블 뷰어)

차트는 구성요소 모델의 예측 정확도에 대한 점도표입니다. 각 점은 Y축에 그려진 정확도 수준과 함께 하나 이상의 구성요소 모델을 나타냅니다. 해당하는 개별 구성요소 모델에 대한 정보를 얻으려면 점을 가리키십시오.

참조선. 도표는 참조 모델 및 naïve 모델과 양상블에 대한 색상 코드화된 선을 표시합니다. 스코어링에 사용될 모델에 해당하는 선 옆에 체크 표시가 나타납니다.

상호작용성. 결합 규칙을 변경하면 차트가 업데이트됩니다.

부스팅된 양상블. 부스팅된 양상블에 대해 선형 차트가 표시됩니다.

㉖. 구성요소 모델 세부사항 (양상블 뷰어)

테이블에 행별로 나열된 구성요소 모델에 대한 정보가 표시됩니다. 기본적으로 구성요소 모델은 오름차순 모델 번호 순으로 정렬됩니다. 열 값에 따라 행을 오름차순 또는 내림차순으로 정렬할 수 있습니다.

모델. 구성요소 모델이 만들어진 연속 순서를 나타내는 번호입니다.

정확도. 퍼센트로 형식화된 전체 정확도입니다.

방법. 모델링 방법입니다.

예측변수. 구성요소 모델에 사용된 예측변수의 수입니다.

모델 크기. 모델 크기는 모델링 방법에 따라 다릅니다. 트리의 경우 트리의 노드 수이고, 선형 모델의 경우 계수이며, 신경망의 경우 시냅스 수가 됩니다.

레코드. 학습 표본의 가중치를 부여한 입력 레코드 수입니다.

ㄴ. 자동 데이터 준비 (양상블 뷰어)

이 보기에는 제외되는 필드 및 변환된 필드가 어떻게 자동 데이터 준비(ADP) 단계에서 유도되었는지에 대한 정보가 표시됩니다. 변환되었거나 제외된 각 필드에 대해 테이블에 필드 이름, 분석에서 역할, ADP 단계에서 실행한 작업이 나열됩니다. 필드는 필드 이름의 알파벳 오름차순으로 정렬됩니다.

이상치 자르기 작업은 표시되는 경우, 절사 값을 넘는 연속형 예측자 값(평균에서 3배 표준 편차)이 절사 값으로 설정되었음을 나타냅니다.

㉔ 분할 모델의 모델 너깃

분할 모델의 모델 너깃에서는 분할로 작성된 모든 개별 모델에 대한 액세스를 제공합니다.

분할 모델 너깃은 다음을 포함합니다.

- 각 모델에 대한 통계량 집합과 함께 작성된 모든 분할 모델의 목록
- 전체 모델에 대한 정보

분할 모델의 목록에서 개별 모델을 열어 추가로 탐색할 수 있습니다.

가. 분할 모델 뷰어

모델 탭에서는 너깃에 포함된 모든 모델을 나열하고 분할 모델에 대한 다양한 양식의 통계를 제공합니다. 모델링 노드에 따라 두 가지 일반 양식이 있습니다.

정렬기준. 이 목록을 사용하여 모델을 나열하는 순서를 선택합니다. 표시 열 값에 따라 목록을 오름차순 또는 내림차순으로 정렬할 수 있습니다. 또는 열 머리말을 클릭하여 해당 열로 목록을 정렬합니다. 기본값은 전체 정확도의 내림차순입니다.

열 메뉴 표시/숨기기. 표시하거나 숨길 개별 열을 선택할 수 있는 메뉴를 표시하려면 이 단추를 클릭합니다.

보기. 파티셔닝을 사용하는 경우 훈련 데이터 또는 검정 데이터에 대한 결과를 보도록 선택할 수 있습니다.

각 분할에서 표시되는 세부사항은 다음과 같습니다.

그래프. 이 모델의 데이터 분포를 나타내는 썸네일. 너깃이 캔버스에 있을 때 전체 크기로 그래프를 열려면 썸네일을 두 번 클릭합니다.

모델. 모델 유형의 아이콘. 이 특정 분할에 대한 모델 너깃을 열려면 아이콘을 두 번 클릭합니다.

분할 필드. 모델링 노드에서 분할 필드로 지정된 필드로, 여러 가능한 값을 포함합니다.

분할의 레코드 수. 이 특정 분할과 관련된 레코드 수.

사용된 필드 수. 사용된 입력 필드 수에 따라 분할 모델을 순위화합니다.

전체 정확도(%). 분할 모델의 총 레코드 수와 해당 분할 모델에서 올바르게 예측된 레코드의 퍼센트.

분할. 열 머리말에 분할을 만드는 데 사용하는 필드가 표시되며, 셀은 분할 값입니다. 해당 분할에 대해 작성된 모델에 대해 모델 뷰어를 열려면 분할을 두 번 클릭하십시오.

정확도. 퍼센트로 형식화된 전체 정확도입니다.

모델 크기. 모델 크기는 모델링 방법에 따라 다릅니다. 트리의 경우 트리의 노드 수이고, 선형 모델의 경우 계수이며, 신경망의 경우 시냅스 수가 됩니다.

레코드. 훈련 표본의 가중치를 부여한 입력 레코드 수입니다.

⑨ 스트림에서 모델 너깃 사용

모델 너깃은 새 데이터 스코어를 계산하도록 스트림에 배치되고 새 노드를 생성합니다. 데이터 스코어링을 통해 새 레코드에 대한 예측을 작성하기 위해 모델 작성으로 얻은 정보를 사용할 수 있습니다. 스코어링 결과를 보려면 너깃에 터미널 노드(즉, 처리 또는 출력 노드)를 첨부하고 터미널 노드를 실행해야 합니다.

일부 모델에서 모델 너깃은 신뢰도 값 또는 군집 중심으로부터의 거리와 같은 예측 품질에 대한 추가 정보를 제공할 수도 있습니다. 새 노드를 생성하면 생성된 모델 구조에 기반하여 새 노드를 쉽게 작성할 수 있습니다. 예를 들어, 입력 필드 선택을 수행하는 대부분의 모델에서는 모델이 중요한 것으로 식별한 입력 필드만 전달하는 필터 노드를 생성할 수 있습니다.

 **참고:** IBM® SPSS® Modeler의 다른 버전에서 실행하는 경우 주어진 모델에서 주어진 케이스에 지정된 스코어가 조금 다를 수 있습니다. 일반적으로 버전 간 소프트웨어를 개선한 결과입니다.

데이터 스코어링을 위해 모델 너깃을 사용하려면

1. 데이터가 전달되는 데이터 소스 또는 스트림에 모델 너깃을 연결하십시오.
2. 모델 너깃에 하나 이상의 처리 또는 출력 노드(예: 테이블 또는 분석 노드)를 추가하거나 연결하십시오.
3. 모델 너깃에서 노드 다운스트림 중 하나를 실행하십시오.

참고: 데이터 스코어링을 위해 세분화되지 않은 규칙 노드를 사용할 수 없습니다. 연관 규칙 모델에 기반하여 데이터 스코어를 계산하려면 세분화되지 않은 규칙 노드를 사용하여 규칙 세트 너깃을 생성하거나 스코어링을 위해 규칙 세트 너깃을 사용하십시오. 자세한 정보는 연관 모델 너깃에서 규칙 세트 생성의 내용을 참조하십시오.

처리 노드 생성을 위해 모델 너깃을 사용하려면

1. 팔레트에서 모델을 찾아보거나 스트림 캔버스에서 모델을 편집하십시오.
2. 모델 너깃 브라우저 창의 생성 메뉴에서 원하는 노드 유형을 선택하십시오. 사용 가능한 옵션은 모델 너깃 유형에 따라 달라집니다. 특정 모델에서 생성할 수 있는 항목에 대한 세부사항을 보려면 특정 모델 너깃 유형을 참조하십시오.

⑩ 모델링 노드 재생성

수정 또는 업데이트하려는 모델 너깃이 있고, 모델을 작성하는 데 사용된 스트림을 사용할 수 없는 경우 원래 모델을 작성하는 데 사용한 동일한 옵션으로 모델링 노드를 재생성할 수 있습니다.

모델을 재작성하려면 모델 팔레트에서 모델을 마우스 오른쪽 단추로 클릭하고 **모델링 노드 생성**을 선택하십시오.

또는 모델을 찾아볼 때 생성 메뉴에서 **모델링 노드 생성**을 선택하십시오.

재생성된 모델링 노드의 기능은 대부분의 경우 원래 모델을 작성하는 데 사용된 항목과 동일합니다.

- 의사결정 트리 모형의 경우 대화형 세션 중에 지정된 추가 설정을 노드에 저장할 수 있으며, 재생성된 모델링 노드에서 **트리 지시문 사용** 옵션을 사용할 수 있습니다.
- 의사결정 목록 모델에서 **저장된 대화형 세션 정보 사용** 옵션이 사용 가능합니다. 자세한 정보는 의사결정 목록 모델 옵션의 내용을 참조하십시오.
- 시계열 모델의 경우 **기존 모델을 사용하여 추정 계속** 옵션이 사용 가능하므로, 현재 데이터로 이전 모델을 재생성할 수 있습니다. 자세한 정보는 시계열 모델 옵션 주제를 참조하십시오.

⑪ 모델을 PMML로 가져오기 및 내보내기

PMML 또는 예측 모델 마크업 언어는 모델에 대한 입력, 데이터를 데이터 마이닝을 위해 준비하는 데 사용하는 변환 및 모델 자체를 정의하는 모수를 포함하여 데이터 마이닝과 통계 모델을 설명하기 위한 XML 형식입니다. IBM® SPSS® Modeler에서는 PMML을 가져오고 내보내고, IBM SPSS Statistics 등과 같이 이 형식을 지원하는 다른 애플리케이션과 모델을 공유할 수 있게 만들 수 있습니다.

PMML에 대한 자세한 정보는 데이터 마이닝 그룹 웹 사이트(<http://www.dmg.org>)를 참조하십시오.

모델 내보내기

PMML 내보내기는 IBM SPSS Modeler에서 생성되는 대부분의 모델 유형에 지원됩니다. 자세한 정보는 PMML을 지원하는 모델 유형의 내용을 참조하십시오.

1. 모델 팔레트에서 모델 너깃을 마우스 오른쪽 단추로 클릭하십시오. (또는 캔버스에서 모델 너깃을 두 번 클릭하고 파일 메뉴를 선택하십시오.)
2. 메뉴에서 **PMML 내보내기**를 클릭하십시오.
3. 내보내기(또는 저장) 대화 상자에서 대상 디렉토리 및 모델의 고유 이름을 지정하십시오.



참고:

PMML 내보내기 옵션은 사용자 옵션 대화 상자에서 변경할 수 있습니다. 기본 메뉴에서 다음을 클릭하십시오.

도구 > 옵션 > 사용자 옵션

PMML 탭을 클릭하십시오.

자세한 정보는 PMML 내보내기 옵션 설정의 내용을 참조하십시오.

PMML로 저장된 모델 가져오기

IBM SPSS Modeler 또는 또 다른 애플리케이션에서 PMML로서 내보낸 모델은 모델 팔레트로 가져올 수 있습니다. 자세한 정보는 PMML을 지원하는 모델 유형의 내용을 참조하십시오.

1. 모델 팔레트에서 팔레트를 마우스 오른쪽 단추로 클릭하고 메뉴에서 **PMML 가져오기**를 선택하십시오.
2. 가져올 파일을 선택하고 필요에 따라 변수 레이블의 옵션을 지정하십시오.
3. 열기를 클릭하십시오.

모델에 있는 경우 변수 레이블을 사용하십시오. PMML은 데이터 사전에서 변수에 변수 이름과 변수 레이블(예: *RefID*의 경우 Referrer ID) 둘 모두를 지정할 수도 있습니다. 변수 레이블이 원래 내보낸 PMML에 있는 경우 이를 사용하려면 이 옵션을 선택하십시오.

변수 레이블 옵션을 선택했지만 PMML에 변수 레이블이 없는 경우에는 변수 이름을 통상적으로 사용합니다.

가. PMML을 지원하는 모델 유형

PMML 내보내기

IBM SPSS Modeler 모델. IBM® SPSS® Modeler에서 작성된 다음 모델은 PMML 4.0으로서 내보낼 수 있습니다.

- C&R 트리
- QUEST
- CHAID
- 신경망
- C5.0
- 로지스틱 회귀분석
- Genlin
- SVM
- Apriori
- Carma
- K-평균
- 코호넨
- TwoStep
- TwoStep-AS
- GLMM(PMML은 모든 GLMM 모델용으로 내보낼 수 있지만, 고정 효과만 있습니다.)
- 의사결정 목록
- Cox
- 순차규칙(순차규칙 PMML 모델에 대한 스코어링은 지원되지 않음)
- 임의 트리
- Tree-AS
- 선형
- Linear-AS
- 회귀분석
- 로지스틱
- GLE
- LSVM
- 이상 항목 발견
- KNN
- 연관 규칙

데이터베이스 원시 모델. 데이터베이스 원시 알고리즘을 사용하여 생성된 모델의 경우 PMML 내보내기를 사용할 수 없음. Microsoft 또는 Oracle Data Miner에서 분석 서비스를 사용하여 작성된 모델은 내보낼 수 없습니다.

PMML 가져오기

IBM SPSS Modeler는 IBM SPSS Modeler에서 내보낸 모델뿐만 아니라 IBM SPSS Statistics 17.0 이상에 의해 생성된 모델 또는 변환 PMML을 포함하여 모든 IBM SPSS Statistics 제품의 현재 버전에 의해 생성된 PMML 모델을 가져오고 스코어링할 수 있습니다. 본질적으로, 이는 스코어링 엔진이 스코어링할 수 있는 모든 PMML을 의미하며 다음과 같은 예외가 있습니다.

- Apriori, CARMA, 이상 항목 발견, 순차규칙 및 연관성규칙모델은 가져올 수 없습니다.
- PMML 모델은 스코어링에 사용할 수 있더라도 IBM SPSS Modeler에 가져온 후에는 찾아볼 수 없을 수 있습니다. (여기에는 우선 IBM SPSS Modeler에서 내보낸 모델이 포함됨을 유의하십시오. 이 제한을 피하려면 모델을 PMML이 아니라 생성된 모델 파일[*gm]로서 내보내십시오.)
- 가져올 때는 제한된 검증이 발생하지만 모델을 스코어링하려고 시도할 때 전체 검증이 수행됩니다. 따라서 가져오기가 성공할 수는 있지만 스코어링은 실패하거나 부정확한 결과를 낼 수 있습니다.

참고: IBM SPSS Modeler에 가져온 타사 PMML의 경우, IBM SPSS Modeler는 인지되고 스코어된 유효한 PMML을 스코어링하려고 시도합니다. 모든 PMML이 스코어할지 또는 이를 생성한 애플리케이션과 같은 방식으로 스코어할지는 보장이 되지 않습니다.

⑫ 스코어링 어댑터에 대한 모델 게시

스코어링 어댑터가 설치된 데이터베이스 서버에 모델을 게시할 수 있습니다. 스코어링 어댑터를 사용하면 데이터베이스의 사용자 정의 함수(UDF) 기능을 사용하여 데이터베이스에서 모델 스코어링을 수행할 수 있습니다. 데이터베이스에서 스코어링을 수행하면 스코어링 전에 데이터를 추출하지 않아도 됩니다. 스코어링 어댑터에 게시하면 UDF를 실행할 예제 SQL도 생성됩니다.

스코어링 어댑터를 게시하는 방법

1. 모델 너깃을 두 번 클릭하여 여십시오.
2. 모델 너깃 메뉴에서 다음을 선택하십시오.
파일 > UDF로 게시
3. 대화 상자에 관련 필드를 채우고 **확인**을 클릭하십시오.

데이터베이스 연결. 모델에서 사용하려는 데이터베이스에 대한 연결 세부사항.

게시 ID. (z/OS용 Db2 데이터베이스만 해당) 모델의 식별자. 동일한 모델을 다시 작성하고 동일한 게시 ID를 사용하는 경우 생성된 SQL은 동일합니다. 따라서 이전에 생성된 SQL을 사용하는 애플리케이션을 변경하지 않고도 모델을 다시 작성할 수 있습니다. (다른 데이터베이스의 경우 생성된 SQL은 모델에 고유합니다.)

예제 SQL 생성. 이 옵션을 선택하면 파일 필드에 지정한 파일에서 예제 SQL을 생성합니다.

⑬ 세분화되지 않은 모델

세분화되지 않은 모델은 데이터에서 추출된 정보를 포함하지만, 예측을 직접 생성하는 목적을 위해 설계되지는 않았습니다. 즉, 스트림에 추가할 수 없음을 의미합니다. 세분화되지 않은 모델은 생성된 모델 팔레트에서 "정제되지 않은 다이아몬드"로 표시됩니다.

그림 1. 세분화되지 않은 모델 아이콘



세분화되지 않은 규칙 모델에 대한 정보를 확인하려면 모델을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **찾아보기**를 선택하십시오. IBM® SPSS® Modeler에서 생성된 다른 모델과 마찬가지로, 다양한 탭에서 작성된 모델에 대한 요약 및 규칙 정보를 제공합니다.

노드 생성. 생성 메뉴를 사용하면 규칙을 기반으로 새 노드를 작성할 수 있습니다.

- **선택 노드.** 현재 선택된 규칙이 적용되는 레코드를 선택할 선택 노드를 생성합니다. 규칙이 선택되지 않으면 이 옵션은 사용할 수 없습니다.
- **규칙 세트.** 단일 대상 필드의 값을 예측할 규칙 세트 노드를 생성합니다. 자세한 정보는 연관 모델 너깃에서 규칙 세트 생성의 내용을 참조하십시오.

(7) 생성된 통계 모형 고급 출력

선형 회귀, 로지스틱 회귀분석 및 요인/PCA 모델 너깃 브라우저의 고급 탭은 모델에 대한 자세한 통계를 표시합니다. 이러한 모형 통계량에 해석에 대한 자세한 정보를 보려면 아래의 특정 모델 유형에 대한 링크를 선택하십시오.

(8) 군집 모델 모델 탭

군집 모델(TwoStep, K-평균 및 코호넨)에 대한 모델 탭에는 모델에 의해 정의된 군집에 대한 자세한 정보가 포함되어 있습니다.

처음으로 군집 모델 너깃을 찾아보는 경우에는 모델 탭 결과가 접혀 있습니다. 관심 있는 결과를 보려면 항목의 왼쪽에 있는 펼치기 제어를 사용하여 결과를 펼치거나 **모두 펼치기** 단추를 클릭하여 모든 결과를 표시하십시오. 결과 보기를 완료했을 때 결과를 숨기려면 펼치기 제어를 사용하여 숨기려는 특정 결과를 접거나 **모두 접기** 단추를 클릭하여 모든 결과를 접으십시오.

군집. 군집에는 레이블이 지정되며 각각의 군집에 지정된 레코드 수가 표시됩니다. 각각의 군집은 군집의 **프로토타입**으로 간주할 수 있는 중심에 의해 설명됩니다. 군집에 사용 가능한 세부사항은 군집 모델의 유형에 따라 다릅니다. 자세한 정보를 보려면 아래에서 군집 모델의 특정 유형을 선택하십시오.

(9) 규칙 세트/의사결정 트리 모형 탭 생성

생성된 규칙 세트 노드 모델 탭에는 알고리즘을 통해 데이터에서 추출된 규칙 목록이 표시됩니다. 규칙의 형식은 생성된 모델의 유형에 따라 다릅니다. 세부사항을 보려면 아래의 특정 모델 유형을 선택하십시오.

2) 선별 모델

(1) 필드 및 레코드 선별

분석의 예비 단계에서 여러 모델링 노드를 사용하여 모델링과 가장 관련된 필드 및 레코드를 찾을 수 있습니다. 필드선택 노드를 사용하여 중요도 기준으로 필드를 선별 순위화하고 이상 항목 발견 노드를 사용하여 "정규" 데이터의 알려진 패턴을 준수하지 않는 이상 레코드를 찾을 수 있습니다.



필드선택 노드는 기준(예: 결측값의 퍼센트) 세트를 기반으로 제거용 입력 필드를 차단합니다. 그런 다음 지정된 대상에 상대적인 남아 있는 입력의 중요도에 대해 순위를 매깁니다. 예를 들어, 수백 개의 잠재 입력이 있는 데이터 세트가 있다면 환자 결과 모델링 시 어느 것이 가장 유용합니까?



이상 항목 발견 노드는 "정상" 데이터 패턴을 따르지 않는 특이 케이스 또는 이상값을 식별합니다. 이 노드를 사용하면 이전에 알려진 패턴에 적합하지 않고, 찾고 있는 패턴을 정확하게 모르더라도 이상값을 식별할 수 있습니다.

이상 항목 발견은 특정 목표(종속) 필드를 고려하지 않고 해당 필드가 예측하려고 하는 패턴에 관련되는지 여부에 관계없이 모델에서 선택된 필드 세트를 기반으로 군집분석을 통해 특수 레코드 또는 케이스를 식별한다는 점에 유의하십시오. 이러한 이유로, 필드선택 또는 필드 선별 및 순위화를 위한 다른 기법과 함께 이상 항목 발견을 사용하고자 할 수 있습니다. 예를 들어, 필드 선택을 사용하여 특정 목표와 관련된 가장 중요한 필드를 식별한 후 이상 항목 발견을 사용하여

해당 필드와 관련된 가장 특이한 레코드를 찾을 수 있습니다. (대체 접근 방식으로, 의사결정 트리 모형을 작성하고 잠재적 이상 항목으로 오분류된 레코드를 탐색할 수 있습니다. 그러나, 이 방법은 대규모로 복제하거나 자동화하기에 어렵습니다.)

(2) 필드선택 노드

데이터 마이닝 문제점은 잠재적으로 입력으로 사용할 수 있는 수백 또는 심지어 수천 개의 필드입니다. 결과적으로 모델에 포함시킬 필드나 변수를 검토하는 데 상당한 시간과 노력이 소모될 수 있습니다. 이 선택의 범위를 좁히려면 필드선택 알고리즘을 사용하여 주어진 분석에 가장 중요한 필드를 식별할 수 있습니다. 예를 들어, 요인 수를 기준으로 하여 환자 결과를 예측하려 시도하는 경우 어느 요인이 가장 중요합니까?

필드선택은 다음 세 가지 단계로 이루어집니다.

- **선별.** 중요하지 않고 문제가 되는 입력 및 레코드나 예측값이 너무 많거나 유용한 변화가 너무 많거나 적은 입력 필드와 같은 케이스를 제거합니다.
- **순위화.** 중요도를 기준으로 하여 나머지 입력을 정렬하고 순위를 지정합니다.
- **선택.** 예를 들어 가장 중요한 입력만 보존하고 다른 모든 입력은 필터링 또는 제외해서 후속 모델에 사용할 변수의 서브세트를 식별합니다.

많은 조직이 너무 많은 데이터로 과부하된 경우에는 모델링 프로세스를 단순화하고 가속화할 때 필드선택이 실질적으로 유용할 수 있습니다. 가장 중요한 필드에 빠르게 집중함으로써 필요한 계산량을 줄일 수 있습니다. 간과할 수 있는 작지만 중요한 관계를 보다 간편하게 찾고 궁극적으로는 더 단순 및 정확하고 쉽게 설명할 수 있는 모델을 확보합니다. 모델에 사용하는 필드 수를 줄임으로써 미래 반복에서 수집되는 데이터의 양과 스코어링 시간을 줄이는 것이 가능함을 알 수 있습니다.

예제. 통신회사에 회사 고객 5,000명이 특별 프로모션에 보인 반응에 대한 정보를 포함하는 데이터 웨어하우스가 있습니다. 이때 데이터에는 고객의 나이, 고용, 수입, 통신 사용 통계량을 포함하는 여러 필드가 있습니다. 세 개의 대상 필드는 세 가지 각 제안에 고객이 반응하는지 여부를 보여줍니다. 회사는 이 데이터를 사용하여 고객이 향후에 유사한 제안에 반응할 가능성을 예측할 수 있습니다.

요구사항. 단일 대상 필드(해당 역할이 목표로 설정된 필드)와 목표와 관련하여 선별 또는 순위화할 다중 입력 필드. 목표 및 입력 필드 모두 연속형(숫자 범위) 또는 범주형의 측정 수준을 포함할 수 있습니다.

① 필드선택 모델 설정

모델 탭의 설정에는 입력 필드 선별 기준을 미세 조정할 수 있는 설정과 함께 표준 모델 옵션이 포함되어 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

입력 필드 선별

선별은 입력/목표 관계와 관련하여 유용한 정보를 추가하지 않는 입력 또는 케이스 제거 작업을 포함합니다. 선별 옵션은 선택한 목표 필드와 관련하여 예측력과 상관없이 문제가 되는 필드의 속성에 기반합니다. 선별된 필드는 입력 순위화에 사용되는 계산에서 제외되며 선택적으로 모델링에 사용되는 데이터에서 필터링 또는 제거될 수 있습니다.

다음 기준에 기반하여 필드를 선별할 수 있습니다.

- **결측값의 최대 퍼센트.** 결측값이 너무 많은 필드(총 레코드 수의 퍼센트로 표시됨)를 선별합니다. 결측값 퍼센트가 높은 필드는 예측 정보를 거의 제공하지 않습니다.
- **단일 범주의 최대 레코드 퍼센트.** 총 레코드 수에 비해 동일한 범주에 속하는 레코드가 너무 많은 필드를 선별합니다. 예를 들어 데이터베이스에서 고객 중 95%가 동일한 차량을 운전하는 경우 이 정보를 포함해도 고객을 서로 구별하는 데 유용하지 않습니다. 그래서 지정된 최대값을 초과하는 필드가 선별됩니다. 이 옵션은 범주형 필드에만 적용됩니다.
- **최대 범주 수를 레코드의 백분율로.** 총 레코드 수와 관련된 범주가 너무 많은 필드를 선별합니다. 범주의 높은 퍼센트가 단일 케이스만 포함하는 경우 필드 사용이 제한될 수 있습니다. 예를 들어, 모든 고객이 서로 다른 모자를 착용한 경우 이 정보는 동작의 모델링 패턴에 별로 유용하지 않습니다. 이 옵션은 범주형 필드에만 적용됩니다.
- **최소 변동계수.** 변동계수가 지정된 최소값 이하인 필드를 선별합니다. 이 측도는 입력 필드 표준 편차를 입력 필드의 평균값으로 나눈 비율입니다. 이 값이 0에 가까우면 변수 값에서 변동은 많지 않습니다. 이 옵션은 연속형 필드(숫자 범위)에만 적용됩니다.
- **최소 표준 편차.** 표준 편차가 지정된 최소값 이하인 필드를 선별합니다. 이 옵션은 연속형 필드(숫자 범위)에만 적용됩니다.

결측 데이터를 포함하는 레코드. 목표 필드에 대한 결측값 또는 모든 입력에 대한 결측값이 있는 레코드나 케이스는 순위화에 사용되는 모든 계산에서 자동으로 제외됩니다.

② 필드선택 옵션

옵션 탭으로 모델 너깃에 입력 필드를 선택하거나 제외시키기 위한 기본 설정을 지정할 수 있습니다. 그런 다음 모델을 스트림에 추가하여 후속 모델 작성 노력에 사용할 필드의 서브세트를 선택할 수 있습니다. 또는 모델을 생성한 후 모델 브라우저에서 추가 필드를 선택 또는 선택 취소하여 이 설정을 대체할 수 있습니다. 하지만 기본 설정은 추가로 변경하지 않고도 모델 너깃을 적용할 수 있어서 특히 스크립팅 용도에 유용할 수 있습니다.

자세한 정보는 필드선택 모델 결과의 내용을 참조하십시오.

다음 옵션을 사용할 수 있습니다.

순위 지정된 모든 필드. 중요, 보통 또는 중요하지 않음과 같은 순위를 기준으로 하여 필드를 선택합니다. 한 순위 또는 또 다른 순위에 레코드를 지정하는 데 사용하는 절사 값 외에 각 순위의 레이블을 편집할 수 있습니다.

최대 필드 수. 중요도에 따라 상위 n 개의 필드를 선택합니다.

다음보다 큰 중요도. 중요도가 지정된 값보다 큰 모든 필드를 선택합니다.

목표 필드는 선택과 상관 없이 항상 보존됩니다.

중요도 순위화 옵션

모든 범주형. 모든 입력 및 목표가 범주형인 경우 네 개의 척도에 따라 중요도를 순위화할 수 있습니다.

- **Pearson 카이제곱.** 기존 관계의 강도 또는 방향을 나타내지 않고 목표 및 입력의 독립성을 검정합니다.
- **우도비 카이제곱.** Pearson의 카이제곱과 비슷하지만 목표-입력 독립성도 검정합니다.
- **Cramer의 V** Pearson의 카이제곱 통계에 기반한 연관성 척도. 값은 0(연관 없음)부터 1(완벽한 연관)까지 범위입니다.
- **람다.** 목표 값을 예측하기 위해 변수를 사용하는 경우 오차 내 비례 축소를 반영하는 연관성 척도. 1의 값은 입력 필드가 목표를 완벽하게 예측함을 나타내고, 0의 값은 입력이 목표에 대한 유용한 정보를 제공하지 않음을 나타냅니다.

일부 범주형. 모두가 아닌 일부 입력이 범주형이고 목표도 범주형인 경우 중요도는 Pearson 또는 우도비 카이제곱에 기반하여 순위화할 수 있습니다. (Cramer의 V 및 람다는 모든 입력이 범주형이 아닌 경우 사용할 수 없습니다.)

범주형 대 연속형. 연속형 목표에서 범주형 입력을 순위화하거나 반대의 경우(둘 중 하나가 범주형으로 둘 다 범주형은 아닌 경우) F 통계량이 사용됩니다.

모두 연속형. 연속형 목표에서 연속 입력을 순위화할 때 상관계수에 기반한 t 통계량이 사용됩니다.

(3) 필드선택 모델 너깃

필드선택 모델 너깃에서는 필드선택 노드에서 순위화한 대로, 선택한 목표와 관련된 각 입력의 중요도를 표시합니다. 순위화 전에 선별된 필드도 나열됩니다. 자세한 정보는 필드선택 노드의 내용을 참조하십시오.

필드선택 모델 너깃을 포함하는 스트림을 실행할 때 모델은 모델 탭에서 현재 선택이 표시한 대로, 선택한 입력만 유지하는 필터 역할을 합니다. 예를 들어, 중요도 순위화된 모든 필드를 선택하거나(기본 옵션 중 하나) 모델 탭에서 필드의 서브세트를 수동으로 선택할 수 있습니다. 목표 필드도 선택에 상관없이 유지됩니다. 다른 모든 필드는 제외됩니다.

필터링은 필드 이름에만 기반합니다. 예를 들어, 나이 및 소득을 선택한 경우 이 이름 중 하나와 일치하는 필드가 유지됩니다. 모델은 새 데이터에 기반하여 필드 순위를 업데이트하지 않습니다. 선택한 이름에 따라서만 필드를 필터링합니다. 따라서 새 데이터 또는 업데이트된 데이터에 모델을 적용할 경우 신중해야 합니다. 문제가 의심되면 모델을 재생성하는 것이 좋습니다.

① 필드선택 모델 결과

필드선택 모델 너깃의 모델 탭에서는 상위 분할창에 있는 모든 입력의 순위 및 중요도를 표시하고, 왼쪽에 있는 열의 선택란을 사용하여 필터링을 위해 필드를 선택할 수 있습니다. 스트림을 실행할 때 선택된 필드만 유지되고, 다른 필드는 제거됩니다. 기본 선택은 모델 작성 노드에 지정된 옵션에 기반하지만, 필요에 따라 추가 필드를 선택하거나 선택 취소할 수 있습니다.

아래 분할창에서는 결측값의 퍼센트 또는 모델링 노드에 지정된 다른 기준에 따라 순위에서 제외된 입력을 나열합니다. 순위화된 필드와 마찬가지로 왼쪽에 있는 열의 선택란을 사용하여 이러한 필드를 포함하거나 삭제할 수 있습니다. 자세한 정보는 필드선택 모델 설정의 내용을 참조하십시오.

- 순위, 필드 이름, 중요도 또는 기타 표시된 열로 목록을 정렬하려면 열 헤더를 클릭하십시오. 또는 도구 모음을 사용하려면 정렬 기준 목록에서 원하는 항목을 선택하고 위로 및 아래로 화살표를 사용하여 정렬 방향을 변경하십시오.
- 도구 모음을 사용하여 모든 필드를 선택 또는 선택 취소하고 필드 확인 대화 상자에 액세스할 수 있습니다. 이 대화 상자에서는 순위 또는 중요도를 기준으로 필드를 선택할 수 있습니다. 또한 Shift 및 Ctrl 키를 누른 상태로 필드를 클릭하여 선택을 확장하고 스페이스바를 사용하여 선택한 필드 그룹을 설정하거나 해제할 수 있습니다. 자세한 정보는 중요도에 따라 필드 선택의 내용을 참조하십시오.
- 중요, 주변 또는 중요하지 않음으로 입력을 순위화할 때 임계값은 테이블 아래 범례에 표시됩니다. 이러한 값은 모델링 노드에 지정됩니다. 자세한 정보는 필드선택 옵션의 내용을 참조하십시오.

② 중요도에 따라 필드 선택

필드선택 모델 너깃을 사용하여 데이터를 스코어링하는 경우 순위화 또는 선별된 필드(왼쪽 열의 선택란으로 표시됨) 목록에서 선택된 모든 필드가 유지됩니다. 기타 필드는 삭제됩니다. 선택을 변경하려면 도구 모음을 사용하여 필드 선택 대화 상자에 액세스할 수 있습니다. 여기서 순위 또는 중요도로 필드를 선택할 수 있습니다.

표시된 모든 필드. 중요, 주변 또는 중요하지 않음으로 표시된 모든 필드를 선택합니다.

최대 필드 수. 중요도에 따라 상위 n 개 필드를 선택할 수 있습니다.

다음보다 큰 중요도. 지정된 임계값보다 큰 중요도의 모든 필드를 선택합니다.

③ 필드선택 모델에서 필터 생성

필드선택 모델의 결과에 따라 기능에서 필터 생성 대화 상자를 사용하여 지정된 목표와 관련된 중요도에 따라 필드의 서브세트를 포함 또는 제외하는 하나 이상의 필터 노드를 생성할 수 있습니다. 모델 너기도 필터로 사용할 수 있습니다. 그러면 모델을 복사 또는 수정하지 않고도 탄력적으로 필드의 다른 서브세트를 실험할 수 있습니다. 목표 필드는 포함 또는 제외의 선택 여부에 상관없이 항상 필터로 유지됩니다.

포함/제외. 필드를 포함하거나 제외하도록 선택할 수 있습니다. 예를 들어, 상위 10개 필드를 포함하거나 중요하지 않음으로 표시된 모든 필드를 제외할 수 있습니다.

선택된 필드. 현재 테이블에서 선택된 모든 필드를 포함하거나 제외합니다.

표시된 모든 필드. 중요, 주변 또는 중요하지 않음으로 표시된 모든 필드를 선택합니다.

최대 필드 수. 중요도에 따라 상위 n 개 필드를 선택할 수 있습니다.

다음보다 큰 중요도. 지정된 임계값보다 큰 중요도의 모든 필드를 선택합니다.

(4) 이상 항목 발견 노드

이상 항목 발견 모델은 데이터에서 이상치 또는 특수 케이스를 식별하기 위해 사용됩니다. 특수 케이스에 대한 규칙을 저장하는 다른 모델링 방법과 달리, 이상 항목 발견 모델은 유사하게 보이는 보통의 작동에 대한 정보를 저장합니다. 그러면 이상치가 알려진 패턴을 따르지 않을 경우에도 이상치를 식별할 수 있고, 특히 새 패턴이 끊임없이 새로 생성될 수 있는 부정 수단 발견과 같은 애플리케이션에서 유용할 수 있습니다. 이상 항목 발견은 비감독 방법으로, 시작점으로 사용할 부정 수단의 알려진 케이스를 포함하는 훈련 데이터 세트가 필요하지 않습니다.

이상치를 식별하는 전형적인 방법에서는 일반적으로 한 번에 하나 또는 두 개의 변수를 검색하지만, 이상 항목 발견은 유사한 레코드를 놓을 군집 또는 피어 그룹을 식별하기 위해 많은 필드수를 조사할 수 있습니다. 각 레코드는 해당 피어 그룹에 다른 레코드와 비교되어 가능한 이상 항목을 식별할 수 있습니다. 케이스가 보통의 중심에서 멀어질수록 한층 특수하게 됩니다. 예를 들어, 알고리즘은 레코드를 세 개의 별도의 군집으로 묶고 하나의 군집 중심에서 멀리 있는 레코드에 플래그를 지정할 수 있습니다.

각 레코드에는 케이스가 속하는 군집에서 해당 평균에 대한 그룹 편차 지수의 비율인 이상 항목 지수가 지정됩니다. 이 지수의 값이 클수록 케이스의 편차는 평균보다 커집니다. 일반적인 상황에서, 이상 항목 지수 값이 1 또는 1.5보다 작은 케이스는 이상 항목으로 간주되지 않습니다. 편차가 평균과 같거나 약간 크기 때문입니다. 그러나 지수 값이 2보다 큰 케이스는 좋은 이상 항목 후보가 될 수 있습니다. 편차가 최소 평균의 두 배이기 때문입니다.

이상 항목 발견은 추가 분석에 대해 후보여야 하는 특수 케이스 또는 레코드의 빠른 발견을 위해 설계된 탐색 방법입니다. 이러한 항목은 *의심이 가는* 이상 항목(엄밀한 검사에서 실제로 밝혀지거나 그렇지 않을 수 있는)으로 간주해야 합니다. 레코드가 완전히 유효하다는 것을 알 수 있지만, 모델 작성 목적을 위해 데이터로부터 선별하기 위해 선택할 수 있습니다. 또는, 알고리즘이 반복적으로 거짓 이상 항목을 나타내면, 이는 데이터 수집 프로세스에서의 오류 또는 아티팩트를 가리킬 수 있습니다.

이상 항목 발견은 특정 목표(중속) 필드를 고려하지 않고 해당 필드가 예측하려고 하는 패턴에 관련되는지 여부에 관계없이 모델에서 선택된 필드 세트를 기반으로 군집분석을 통해 특수 레코드 또는 케이스를 식별한다는 점에 유의하십시오. 이러한 이유로, 필드선택 또는 필드 선별 및 순위화를 위한 다른 기법과 함께 이상 항목 발견을 사용하고자 할 수 있습니다. 예를 들어, 필드 선택을 사용하여 특정 목표와 관련된 가장 중요한 필드를 식별한 후 이상 항목 발견을 사용하여 해당 필드와 관련된 가장 특이한 레코드를 찾을 수 있습니다. (대체 접근 방식으로, 의사결정 트리 모형을 작성하고 잠재적 이상 항목으로 오분류된 레코드를 탐색할 수 있습니다. 그러나, 이 방법은 대규모로 복제하거나 자동화하기에 어렵습니다.)

예제. 농업 개발 기금의 가능한 부정 행위 선별 심사에서, 이상 항목 발견을 사용하여 표준 편차를 발견함으로써 이상 항목으로 추후 조사할 가치가 있는 레코드를 강조할 수 있습니다. 특히 농장의 유형과 규모에 비해 너무 많이(또는 너무 적게) 클레임하는 것으로 보이는 기금 애플리케이션에 관심이 있습니다.

요구사항. 하나 이상의 입력 필드. 소스 또는 유형 노드를 사용하여 역할이 입력으로 설정된 필드만 입력으로 사용할 수 있음에 유의하십시오. 목표 필드(목표 또는 둘 다에 설정된 역할)는 무시됩니다.

강도. 알려진 규칙 세트를 준수하지 않는 케이스에 플래그를 지정해서 이상 항목 발견 모델은 심지어 이전에 알려진 패턴을 따르지 않는 특이 케이스를 식별할 수 있습니다. 필드선택과 함께 사용하여 이상 항목 발견은 많은 양의 데이터를 선별해서 상대적으로 가장 관심이 있는 레코드를 빠르게 식별할 수 있습니다.

① 이상 항목 발견 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

이상 항목의 절사 값 판별 기준. 플래그가 지정된 이상 항목의 절사 값을 판별하는 데 사용되는 방법을 지정합니다. 다음 옵션을 사용할 수 있습니다.

- **최소 이상 항목 지수 수준.** 플래그 지정 이상 항목의 최소 절사 값을 지정합니다. 이 임계값을 충족시키거나 초과한 레코드에 플래그가 지정됩니다.
- **훈련 데이터의 최고 이상 항목 레코드 백분율.** 훈련 데이터에서 지정된 백분율의 레코드에 플래그를 지정하는 수준에서 자동으로 임계값을 설정합니다. 결과적인 절사는 모델에 모수로 포함됩니다. 이 옵션은 스코어링 중에 플래그를 지정할 레코드의 실제 백분율이 *아닌* 절사 값을 설정할 방법을 판별함에 유의하십시오. 실제 스코어링 결과는 데이터에 따라 다를 수 있습니다.
- **훈련 데이터의 최고 이상 항목 레코드 수.** 훈련 데이터에서 지정된 수의 레코드에 플래그를 지정하는 수준에서 자동으로 임계값을 설정합니다. 결과적인 임계값은 모델에 모수로 포함됩니다. 이 옵션은 스코어링 중에 플래그를 지정할 레코드의 특정 수가 *아닌* 절사 값을 설정할 방법을 판별함에 유의하십시오. 실제 스코어링 결과는 데이터에 따라 다를 수 있습니다.

참고: 절사 값을 판별하는 방식과 무관하게 절사 값은 각 레코드의 보고된 기본 이상 항목 지수 값에 영향을 미치지 않습니다. 단순히 모델을 추정하거나 스코어링할 때 레코드에 이상 항목으로 플래그를 지정하기 위한 임계값을 지정할 뿐입니다. 나중에 보다 많거나 적은 수의 레코드를 검토하려는 경우에는 선택 노드를 사용하여 이상 항목 지수 값($\$O\text{-AnomalyIndex} > X$)을 기준으로 레코드의 서브세트를 식별할 수 있습니다.

보고할 이상 항목 필드 수. 특정 레코드가 이상 항목으로 플래그 지정된 이유에 대한 표시로 보고할 필드 수를 지정합니다. 레코드가 할당된 군집의 필드 표준에서 최대 편차를 표시한다고 정의된 최고 이상 항목 필드가 보고됩니다.

② 이상 항목 발견 고급 옵션

결측값 및 기타 설정에 대한 옵션을 지정하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

조정 계수. 거리 계산에서 연속형(수치 범위) 및 범주형 필드에 주어진 상대값 가중치의 균형을 잡는 데 사용되는 값입니다. 이 값이 크면 연속형 필드의 영향력이 증가합니다. 0이 아닌 값이어야 합니다.

자동으로 피어 그룹 수 계산. 이상 항목 발견을 사용하여 훈련 데이터에 대한 최적의 수의 피어 그룹을 선택하기 위한 여러 가능한 솔루션을 빠르게 분석할 수 있습니다. 피어 그룹의 최대 수와 최소 수를 설정해서 범위를 넓히거나 좁힐 수 있습니다. 값이 크면 시스템을 가능한 솔루션을 보다 광범위하게 탐색하지만 처리 시간이 늘어납니다.

피어 그룹 수 지정. 모델에 포함시킬 군집 수를 알고 있으면 이 옵션을 선택하고 피어 그룹 수를 입력하십시오. 일반적으로 이 옵션을 선택하면 성능이 개선됩니다.

잡음 수준 및 비율. 이 설정은 두 단계 군집화 중 이상치의 처리 방식을 판별합니다. 첫 번째 단계에서는 아주 많은 수의 개별 레코드 데이터를 관리 가능한 수의 군집으로 압축하기 위해 군집 기능(CF) 트리를 사용합니다. 트리는 유사성 측도를 기준으로 작성되며 트리의 노드에 레코드가 너무 많아지면 하위 노드로 레코드를 분할합니다. 두 번째 단계는, CF 트리의 터미널 노드에서 계층적 군집이 시작됩니다. 첫 번째 데이터 전달 시 잡음 처리가 켜져서 두 번째 데이터 전달 시에 꺼집니다. 첫 번째 데이터 전달의 잡음 군집 케이스가 두 번째 데이터 전달의 일반 군집에 지정됩니다.

- **잡음 수준.** 0과 0.5 사이의 값을 지정하십시오. 이 설정은 성장 단계 동안 CF 트리가 채워지는 경우에만 관련되며, 이는 리프 노드에 더 이상의 케이스를 허용할 수 없고 리프 노드를 분할할 수 없음을 의미합니다.

CF 트리가 채워지고 잡음 수준이 0으로 설정되면 임계값이 증가하여 CF 트리가 모든 케이스로 다시 성장합니다. 최종 군집화 후 군집에 할당할 수 없는 값에는 이상치 레이블이 붙습니다. 이상치 군집에는 식별 번호 -1이 지정됩니다. 이상치 군집은 군집 개수에 포함되지 않습니다. 즉, n 개의 군집 및 잡음 처리를 지정하는 경우 알고리즘은 n 개의 군집과 하나의 잡음 군집을 출력합니다. 실질적으로, 이 값을 늘리면 알고리즘이 이상 레코드를 별도의 이상치 군집에 할당하지 않고 보다 자유롭게 트리에 맞춥니다.

CF 트리가 채워지고 잡음 수준이 0보다 큰 경우에는 희박한 리프의 데이터를 자체 잡음 리프에 배치한 후 CF 트리가 재성장합니다. 희박한 리프의 케이스 수 대 가장 큰 리프의 케이스 수 비율이 잡음 수준 미만인 경우 리프가 희박하다고 간주됩니다. 트리가 성장하고 난 후 가능하면 CF 트리에 이상치가 배치됩니다. 그렇지 않은 경우에는 군집화의 두 번째 단계 중에 이상치가 삭제됩니다.

- **잡음 비율.** 잡음 버퍼링에 사용해야 하는 구성요소에 할당되는 메모리 부분을 지정합니다. 이 값의 범위는 0.0 - 0.5입니다. 특정 케이스를 트리의 리프에 삽입하여 리프가 임계값 미만으로 조밀해질 경우 리프가 분할되지 않습니다. 조밀도가 임계값을 초과하면 리프가 분할되어 CF 트리에 또 다른 작은 군집이 추가됩니다. 실질적으로, 이 설정값을 늘리면 알고리즘은 자연스럽게 더 단순한 트리를 추구하는 쪽으로 신속히 나아가게 됩니다.

결측값 대치. 연속형 필드의 경우 결측값 대신 필드 평균을 대치하십시오. 범주형 필드의 경우에는 결측 범주가 결합되어 유효 범주로 처리됩니다. 이 옵션을 선택 취소하면 결측값이 있는 레코드가 분석에서 제외됩니다.

(5) 이상 항목 발견 모델 너깃

이상 항목 발견 모델 너깃은 훈련 데이터 및 추정 프로세스에 대한 정보와 이상 항목 발견 모델이 캡처한 모든 정보를 포함합니다.

이상 항목 발견 모델 너깃을 포함한 스트림을 실행할 때 모델 너깃의 설정 탭에서 선택한 사항

에 따라 판별된 많은 새 필드가 스트림에 추가됩니다. 자세한 정보는 이상 항목 발견 모델 설정의 내용을 참조하십시오. 새 필드 이름은 다음 테이블에 요약된 것처럼 모델 이름을 기준으로 지정되고 \$O 접두문자가 붙습니다.

표 1. 새 필드 이름 생성	
필드 이름	설명
\$O-Anomaly	레코드가 이상 항목인지 여부를 표시하는 플래그 필드입니다.
\$O-AnomalyIndex	레코드의 이상 항목 지수 값입니다.
\$O-PeerGroup	레코드가 할당되는 피어 그룹을 지정합니다.
\$O-Field-n	군집 표준 편차에서 최고 이상값이 n번째인 이상 항목 필드의 이름입니다.
\$O-FieldImpact-n	필드의 변수 편차 지수입니다. 이 값은 레코드가 할당된 군집의 필드 표준에서 편차를 측정합니다.

선택적으로 결과를 보다 쉽게 읽을 수 있도록 정상 레코드의 스코어를 억제할 수 있습니다. 자세한 정보는 이상 항목 발견 모델 설정의 내용을 참조하십시오.

① 이상 항목 발견 모델 세부사항

생성된 이상 항목 발견 모델의 모델 탭은 모델의 피어 그룹에 대한 정보를 표시합니다.

피어 그룹 크기 및 통계는 학습 데이터를 기준으로 한 추정값으로, 동일한 데이터로 실행하더라도 실제 스코어링 결과와 약간 다를 수 있음에 유의하십시오.

보고되지 않은 이유의 잔차는 이상 항목으로 식별된 레코드의 이상 항목 열 각각에 대해 1에서 평균 이상 항목 지수 값의 합계를 뺀 값입니다. 이 퍼센트는 보고된 필드로 이상 항목이 어느 정도 설명되는지를 나타냅니다. 이는 보고할 이상 항목 필드를 판별하는 데 도움이 될 수 있습니다.

② 이상 항목 발견 모델 요약

이상 항목 발견 모델 너깃의 요약 탭은 필드, 작성 설정, 추정 프로세스에 대한 정보를 표시합니다. 레코드에 이상 항목 플래그를 지정하는 데 사용되는 절사 값과 함께 피어 그룹 수도 표시됩니다.

③ 이상 항목 발견 모델 설정

설정 탭을 사용하여 모델 너깃 스코어링에 대한 옵션을 지정하십시오.

이상 항목 레코드 처리 방식 표시 출력에서 이상 항목 레코드를 처리할 방식을 지정합니다.

- 플래그 및 지수 모델에 포함된 절사 값을 초과한 모든 레코드에 참으로 설정되는 플래그 필드를 작성합니다. 각 레코드에 대한 이상 항목 지수도 개별 필드에 보고됩니다. 자세한 정보는 이상 항목 발견 모델 옵션의 내용을 참조하십시오.
- 플래그만 각 레코드의 이상 항목 지수를 보고하지 않고 플래그 필드를 작성합니다.
- 지수만 플래그 필드를 작성하지 않고 이상 항목 지수를 보고합니다.

보고할 이상 항목 필드 수 특정 레코드가 이상 항목으로 플래그 지정된 이유에 대한 표시로 보고할 필드 수를 지정합니다. 레코드가 할당된 군집의 필드 표준에서 최대 편차를 표시한다고 정의된 최고 이상 항목 필드가 보고됩니다.

레코드 삭제 다운스트림 노드의 잠재적 이상 항목에 보다 쉽게 초점을 맞출 수 있도록 스트림에서 정상 레코드를 모두 삭제하려면 이 옵션을 선택하십시오. 또는 모델에 따라 잠재적 이상 항목이라 플래그가 지정되지 않은 레코드로 후속 분석을 제한하기 위해 모든 이상 항목 레코드를 버리도록 선택할 수도 있습니다.

참고: 약간의 반올림 차이로 인해 동일한 데이터로 실행하더라도 스코어링 중에 플래그가 지정된 실제 레코드 수와 모델 훈련 중에 플래그가 지정된 레코드 수가 다를 수 있습니다.

이 유형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- 기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- 데이터베이스 외부 스코어 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

3) 자동화된 모델링 노드

자동화된 모델링 노드는 여러 다른 모델링 방법을 추정하고 비교해서 단일 모델링 실행에 광범위한 접근법을 시도할 수 있게 합니다. 사용할 모델링 알고리즘 및 조합을 포함한(그렇지 않을

경우 상호 배타적임) 각각의 특정 옵션을 선택할 수 있습니다. 예를 들어, 신경망의 신속, 동적 또는 가지치기 방법 중에서 선택하기 보다는 이 방법을 모두 시도할 수 있습니다. 노드는 가능한 모든 옵션 조합을 탐색하고, 사용자가 지정한 척도에 기반하여 각 후보 모델을 순위화하고 추가 분석 또는 스코어링에 사용할 때 가장 효과적인 항목을 저장합니다.

분석 필요에 따라 세 가지 자동화된 모델링 노드 중에서 선택할 수 있습니다.



자동 분류자 노드는 이분형 결과(예 또는 아니오, 이탈 또는 이탈 안함 등)에 대해 다수의 여러 모델을 작성하고 비교하여 주어진 분석을 위한 최상의 접근 방식을 선택할 수 있게 합니다. 많은 모델링 알고리즘이 지원되어 사용할 방법, 각각에 대한 특정 옵션, 결과 비교 기준을 선택할 수 있습니다. 이 노드는 지정된 옵션을 기반으로 모델 세트를 생성하고 사용자가 지정하는 기준에 따라 최상의 후보를 순위화합니다.



자동 수치 노드는 수많은 방법을 사용하여 연속적 수치 범위 결과의 모델을 추정하고 비교합니다. 이 노드는 자동 분류자 노드에서와 같은 방식으로 작동하므로 사용할 알고리즘을 선택하고 단일 모델링 전달에서 여러 옵션의 조합을 실험할 수 있습니다. 지원되는 알고리즘에는 신경망, C&R 트리, CHAID, 선형 회귀, 일반화 선형 회귀 및 지원 벡터 머신(SVM)이 있습니다. 모델은 상관관계, 상대 오차 또는 사용된 변수의 수를 기반으로 비교할 수 있습니다.



자동 군집 노드는 유사한 특성을 가진 레코드 그룹을 식별하는 군집 모델을 추정하고 비교합니다. 이 노드는 다른 자동 모델링 노드와 동일한 방법으로 작동하여 단일 모델링 패스에서 다중 옵션 조합을 실험할 수 있습니다. 군집 모델의 유용성을 필터링하고 순위화하며 특정 필드의 중요성을 기반으로 척도를 제공하려고 시도하는 기본 척도를 사용하여 모델을 비교할 수 있습니다.

최상의 모델은 단일 복합 모델 너깃에 저장되어 찾아보기 및 비교가 가능하고 스코어링에 사용할 모델을 선택할 수 있습니다.

- 이분형, 명목, 수치 목표의 경우에만 여러 스코어링 모델을 선택하고 단일 모델 앙상블에 스코어를 결합할 수 있습니다. 여러 모델로부터 예측을 결합함으로써 개별 모델의 제한사항을 피할 수 있으며 이를 통해 종종 모델 중 하나에서 확보할 수 있는 것보다 전반적인 정확도가 높아집니다.
- 선택적으로 결과에 드릴다운하고 추가로 사용 및 탐색하려는 개별 모델에 대한 모델 너깃 또는 모델링 노드를 생성하도록 선택할 수 있습니다.

모델 및 실행 시간

데이터 세트 및 모델 수에 따라 자동화된 모델링 노드는 실행하는 데 몇 시간 또는 그 이상이 소요될 수 있습니다. 옵션을 선택할 때 생성되는 모델 수에 주의하십시오. 실현 가능한 경우 시스템 자원 수요가 적은 밤이나 주말에 모델링 실행을 스케줄할 수 있습니다.

- 필요에 따라 파티션 또는 표본 노드를 사용하여 초기 훈련 전달에 포함되는 레코드 수를 줄일 수 있습니다. 몇 개의 후보 모델로 선택사항 범위를 좁히면 전체 데이터 세트가 복원될 수 있습니다.
- 입력 필드 수를 줄이려면 필드선택을 사용하십시오. 자세한 정보는 필드선택 노드의 내용을 참조하십시오. 또는 초기 모델링 실행을 사용하여 추가로 탐색할 가치가 있는 필드 및 옵션을 식별할 수 있습니다. 예를 들어, 가장 우수한 모델이 모두 동일한 세 가지 필드를 사용하는 것 같으면 이는 이러한 필드를 유지할 가치가 있다는 강력한 표시입니다.
- 선택적으로 하나의 모델을 추정하는 데 걸리는 시간을 제한하고 모델 선별 및 순위 지정에 사용되는 평가 측도를 지정할 수 있습니다.

(1) 자동화된 모델링 노드 알고리즘 설정

각 모델 유형마다 기본 설정을 사용하거나 각 모델 유형에 맞는 옵션을 선택할 수 있습니다. 특정 옵션은 각 설정마다 선택하지 않고 대부분의 경우 적용하려는 만큼의 수를 선택할 수 있다는 점을 제외하면, 개별 모델링 노드에서 사용 가능한 옵션과 유사합니다. 예를 들어, 신경망 모델을 비교하는 경우 여러 다른 학습 방법을 선택하고 난수 시드 없이 또는 난수 시드를 포함하여 각 방법을 시도할 수 있습니다. 단일 전달에서 여러 다른 모델을 쉽게 생성할 수 있도록 선택된 옵션의 가능한 모든 조합이 사용됩니다. 하지만 많은 설정을 선택하면 모델 수가 아주 빠르게 증가할 수 있으므로 주의해서 사용하십시오.

각 모델 유형에 맞는 옵션을 선택하려면

1. 자동화된 모델링 노드에서 **고급** 탭을 선택하십시오.
2. 모델 유형의 **모델 모수** 열을 클릭하십시오.
3. 드롭 다운 메뉴에서 **지정**을 선택하십시오.
4. **알고리즘 설정** 대화 상자의 **옵션** 열에서 옵션을 선택하십시오.

참고: 추가 옵션은 알고리즘 설정 대화 상자의 고급 탭에서 사용 가능합니다.

(2) 자동화된 모델링 노드 중지 규칙

자동화된 모델링 노드에 지정된 중지 규칙은 노드가 작성한 개별 모델의 정지가 아닌 전체 노드 실행에 관련됩니다.

전체 실행 시간 제한. (신경망, K-평균, 코호넨, 이단계, SVM, KNN, Bayes 넷, C&R 트리 모델만) 지정된 시간 후에 실행을 중지합니다. 해당 시점까지 생성된 모든 모델이 모델 너깃에 포함되고 더 이상의 모델이 생성되지 않습니다.

유효 모델이 생성되는 즉시 정지. 모델이 삭제 탭(자동 분류자 또는 자동 군집 노드의 경우) 또는 모델 탭(자동 수치 노드의 경우)에 지정된 모든 기준을 전달할 때 실행을 중지합니다. 자세한 정보는 자동 분류자 노드 삭제 옵션의 내용을 참조하십시오. 자세한 정보는 자동 군집 노드 삭제 옵션의 내용을 참조하십시오.

(3) 실행 피드백

스트림을 실행하는 데 3초 이상 걸리는 경우 표준 실행 피드백 대화 상자를 표시하는 것 외에 IBM® SPSS® Modeler는 관련된 모델 수에 대한 정보를 표시합니다.

(4) 자동 분류자 노드

자동 분류자 노드는 단일 모델링 실행에서 다양한 접근법을 시도할 수 있도록 여러 다른 방법을 사용하여 명목(변수군)또는 이분형(예/아니오) 목표에 대해 모델을 추정하고 비교합니다. 사용할 알고리즘을 선택하고 여러 옵션을 조합하여 실행할 수 있습니다. 예를 들어, SVM의 방사형 기본 함수, 다항, 시그모이드 또는 선형 방법 중에서 선택하지 않고 이 방법을 모두 시도할 수 있습니다. 노드는 가능한 모든 옵션 조합을 탐색하고, 사용자가 지정한 측도에 기반하여 각 후보 모델을 순위화하고 추가 분석 또는 스코어링에 사용할 때 가장 효과적인 항목을 저장합니다. 추가 정보는 자동화된 모델링 노드의 내용을 참조하십시오.

예제

한 소매업체에는 지난 캠페인의 특정 고객에 대한 오퍼를 추적하는 히스토리 데이터가 있습니다. 이 회사는 현재 각 고객에 올바른 오퍼를 매치해서 보다 수익성이 좋은 결과를 산출하고자 합니다.

요구 사항

측정 수준이 *명목* 또는 *플래그*인 하나의 목표 필드(역할 세트가 **목표**로 설정된) 및 최소 하나의 입력 필드(역할 세트가 **입력**으로 설정된). 플래그 필드의 경우 이익, 리프트, 관련 통계를 계산할 때 적중을 표시하기 위해 목표에 정의된 **참** 값이 사용됩니다. 입력 필드의 가능한 측정 수준은 *연속형* 또는 *범주형*이며, 일부 입력이 몇 가지 모델 유형에 적합하지 않을 수 있다는 제한사항이 있습니다. 예를 들어, C&R 트리, CHAID, QUEST 모델의 입력으로 사용된 순서 필드는 수치 저장 공간(문자열이 아닌)이 있어야 하며 다르게 지정될 경우 이러한 모델에서 무시됩니다. 마찬가지로, 연속형 입력 필드는 일부 경우 구간화될 수 있습니다. 요구 사항은 개별 모델링 노드를 사용할 때와 동일합니다. 예를 들어, Bayes Net 모델은 Bayes Net 노드에서 생성되든 또는 자동 분류자 노드에서 생성되든 이에 상관없이 동일하게 작동합니다.

빈도 및 가중치 필드

빈도 및 가중치는 다른 레코드에 비해 일부 레코드에 추가 중요도를 부여하는 용도로 사용되며, 이는 예를 들어, 작성 데이터 세트가 상위 모집단 섹션을 실제보다 낮게 표시(가중치)함을 사용자가 알고 있거나 한 레코드가 많은 동일한 케이스를 표시(빈도)하기 때문입니다. 빈도 필드는 지정된 경우 C&R 트리, CHAID, QUEST, 의사결정 목록, Bayes Net 모델에 사용될 수 있습니다. 가중치 필드는 C&RT, CHAID, C5.0 모델에 사용될 수 있습니다. 다른 모델 유형은 이러한 필드를 무시하고 모델을 작성합니다. 빈도 및 가중 필드는 모델 작성에만 사용되며 모델 평가 또는 스코어링 시에는 고려되지 않습니다. 추가 정보는 빈도 및 가중 필드 사용의 내용을 참조하십시오.

접두문자

자동 분류자 노드의 너깃에 테이블 노드를 첨부할 경우 \$ 접두문자로 시작하는 이름의 테이블에 새 변수가 여러 개 있습니다.

스코어링 중에 생성된 필드 이름은 표준 접두문자가 아닌 목표 필드를 기반으로 합니다. 서로 다른 모델 유형은 서로 다른 접두문자 집합을 사용합니다.

예를 들어 접두문자 \$G, \$R, \$C는 각각 일반화 선형 모델, CHAID 모델, C5.0 모델을 통해 생성되는 예측에 대한 접두문자로 사용됩니다. \$X는 일반적으로 앙상블을 사용하여 생성되고, \$XR, \$XS, \$XF는 목표 필드가 연속형, 범주형 또는 플래그 필드인 경우에 각각 접두문자로 사용됩니다.

\$.C 접두문자는 범주형 또는 플래그 대상의 예측 신뢰도에 사용됩니다. 예를 들어 \$XFC는 앙상블 플래그 예측 신뢰도에 대한 접두문자로 사용됩니다. \$RC 및 \$CC는 각각 CHAID 모델 및 C5.0 모델의 단일 예측 신뢰도에 대한 접두문자입니다.

지원되는 모델 유형

지원되는 모델 유형으로는 신경망, C&R 트리, QUEST, CHAID, C5.0, 로지스틱 회귀분석, 의사결정 목록, Bayes Net, 판별, 최근접 이웃, SVM, XGBoost Tree 및 XGBoost-AS가 있습니다. 자세한 정보는 자동 분류자 노드 고급 옵션의 내용을 참조하십시오.

연속 기계 학습

모델링의 불편한 점은 시간이 지남에 따라 데이터가 변경되어 모델이 구식이 된다는 것입니다. 이를 일반적으로 *모델 드리프트* 또는 *개념 드리프트*라고 합니다. 모델 드리프트를 효과적으로 극복하기 위해 SPSS Modeler는 연속 자동 기계 학습을 제공합니다. 이 기능은 자동 분류자 노드 및 자동 숫자 노드 모델 너깃에 사용할 수 있습니다. 자세한 정보는 연속 기계 학습의 내용을 참조하십시오.

① 자동 분류자 노드 모델 옵션

자동 분류자 노드의 모델 탭으로 모델을 비교하는 데 사용되는 기준과 함께 작성할 모델 수를 지정할 수 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

교차 검증. 교차 검증을 수행하면 학습을 실행하는 *알려진 데이터*의 데이터 세트(학습 데이터 세트) 및 모델을 테스트하는 대상이 되는 *알려지지 않은 데이터*의 데이터 세트(검증 데이터 세트 또는 테스트 세트)를 모델링합니다. 교차 검증은 모델이 과적합 또는 선택 편향과 같은 문제점에 플래그를 지정하기 위해 예측에 사용되지 않는 새로운 데이터를 예측하는 기능을 테스트하기 위해 사용합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

모델 순위화 기준. 모델을 비교하고 순위화하는 데 사용되는 기준을 지정합니다. 옵션으로는 전체 정확도, ROC 곡선 아래 영역, 이익, 리프트, 필드 수가 있습니다. 여기에 선택된 사항과 상관 없이 요약 보고서에서 모든 측도가 사용 가능함에 유의하십시오.

 **참고:** 명목(변수군) 목표의 경우 순위화가 **전체 정확도** 또는 **필드 수**로 제한됩니다.

이익, 리프트, 관련 통계를 계산할 때 적중을 표시하기 위해 목표에 정의된 **참** 값이 사용됩니다.

- **전체 정확도.** 총 레코드 수를 기준으로 모델에서 올바르게 예측한 레코드의 퍼센트입니다.
- **ROC 곡선 아래 영역.** ROC 곡선은 모델 성능에 대한 지수를 제공합니다. 곡선이 참조선보다 위에 있을수록 검정이 더 정확합니다.
- **이익(누적).** 지정된 비용, 수입, 가중치를 기준으로 계산한 누적 백분위수의 이익 합계(예측의 신뢰도 측면에서 정렬됨)입니다. 일반적으로 이익은 최상위 백분위수로 거의 0에서 시작해서 꾸준히 증가한 후에 감소합니다. 우수한 모델의 경우 이익은 발생하는 백분위수와 함께 보고되는 잘 정의된 최대치를 표시합니다. 정보를 제공하지 않는 모델의 경우에는 이익 곡선이 상대적으로 직선이며 적용되는 비용/수입 구조에 따라 증가 또는 감소하거나 동일한 수준을 유지할 수 있습니다.
- **리프트(누적).** 전체 표본에 상대적인 누적 분위수의 적중률입니다(분위수는 예측의 신뢰도 측면에서 정렬됨). 예를 들어, 최고 분위수의 리프트 값 3은 표본 전반에서 3배 높은 적중률을 나타냅니다. 우수한 모델의 경우에는 리프트가 최고 분위수의 1.0 위에서 시작한 후 더 낮은 분위수의 1.0을 향해 급격하게 감소해야 합니다. 어떤 정보도 제공하지 않는 모델은 리프트가 1.0 주위에서 머뭅니다.
- **필드 수.** 사용된 입력 필드 수를 기준으로 모델의 순위를 정합니다.

모델 순위화 사용. 파티션이 사용 중인 경우 순위가 학습 데이터 세트 또는 검정 세트를 기준으로 하는지 여부를 지정할 수 있습니다. 큰 데이터 세트의 경우 모델의 예비 심사에 파티션을 사용하면 성능이 상당히 개선될 수 있습니다.

사용할 모델 수. 노드가 생성한 모델 너깃에 나열할 모델의 최대 수를 지정합니다. 지정된 순위화 기준에 따라 가장 높은 순위의 모델이 나열됩니다. 이 한계를 높이면 성능이 저하될 수 있음에 유의하십시오. 허용 가능한 최대값은 100입니다.

예측자 중요도 계산. 적절한 중요도 측도를 생성하는 모델의 경우 모델 측정 시 각 예측변수의 상대적 중요도를 나타내는 차트를 표시할 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델

링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 예측변수 중요도는 일부 모델을 계산하는 데 필요한 시간을 연장할 수 있으며 단순히 여러 다른 모델을 광범위하게 비교하려는 경우 권장되지 않습니다. 보다 자세하게 탐색하려는 모델로 분석 범위를 좁히는 경우 더 유용합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

이익 기준. 플래그 목표만 해당됩니다. 이익은 각 레코드의 수입에서 레코드의 비용을 뺀 값입니다. 분위수의 이익은 단순히 분위수의 전체 레코드 이익 합계입니다. 이익은 적중에만 적용됩니다. 추측하지만 비용은 모든 레코드에 적용됩니다.

- **비용.** 각 레코드와 연관된 비용을 지정합니다. **고정** 또는 **가변** 비용을 선택할 수 있습니다. 고정 비용의 경우 비용 값을 지정하십시오. 가변 비용의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 비용 필드로 선택하십시오. (ROC 차트에는 **비용**을 사용할 수 없습니다.)
- **수입.** 적중을 나타내는 각 레코드와 연관된 수입을 지정합니다. **고정** 또는 **가변** 비용을 선택할 수 있습니다. 고정 수입의 경우 수입 값을 지정하십시오. 가변 수입의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 수입 필드로 선택하십시오. (ROC 차트에는 **수입**을 사용할 수 없습니다.)
- **가중치.** 데이터의 레코드가 둘 이상의 단위를 표시하는 경우 빈도 가중치를 사용하여 결과를 조정할 수 있습니다. **고정** 또는 **가변** 가중치를 사용하여 각 레코드와 연관된 가중치를 지정하십시오. 고정 가중치의 경우 가중값(레코드별 노드 수)을 지정하십시오. 가변 가중치의 경우에는 필드 선택기 단추를 클릭하여 한 필드를 가중 필드로 선택하십시오. (ROC 차트에는 **가중치**를 사용할 수 없습니다.)

리프트 기준. 플래그 목표만 해당됩니다. 리프트 계산에 사용할 백분위수를 지정합니다. 결과를 비교할 때 이 값도 변경될 수 있음에 유의하십시오. 자세한 정보는 자동화된 모델 너깃의 내용을 참조하십시오.

② 자동 분류자 노드 고급 옵션

자동 분류자 노드의 고급 탭으로 파티션을 적용하고(가능한 경우), 사용할 알고리즘을 선택하고, 중지 규칙을 지정할 수 있습니다.

사용한 모델. 왼쪽 열의 선택란을 사용하여 비교에 포함할 모델 유형(알고리즘)을 선택합니다. 더 많은 유형을 선택할수록 모델이 더 많이 작성되고 처리 시간은 더 오래 걸립니다.

모델 유형. 사용 가능한 알고리즘을 나열합니다(아래 참조).

모델 매개변수. 각 모델 유형마다 기본 설정을 사용하거나 **지정**을 선택하여 각 모델 유형에 맞는 옵션을 선택할 수 있습니다. 특정 옵션은 여러 옵션 또는 조합을 선택할 수 있다는 점을 제외하면, 개별 모델링 노드에서 사용 가능한 옵션과 유사합니다. 예를 들어, 신경망 모델을 비교하는 경우 여섯 개의 학습 방법 중에서 하나를 선택하기 보다는 모두를 선택하여 단일 전달에서 여섯 개의 모델을 학습시킬 수 있습니다.

모델 수. 현재 설정을 기준으로 하여 각 알고리즘에 대해 생성되는 모델 수를 나열합니다. 옵션을 조합할 경우 모델 수가 빠르게 증가하므로 특히 큰 데이터 세트를 사용할 때에는 이 수에 세심한 주의를 기울일 것을 강력히 권장합니다.

단일 모델 작성에 소요되는 최대 시간 제한. (K-평균, 코호넨, TwoStep, SVM, KNN, Bayes Net, 의사결정 목록 모델만) 한 모델의 최대 시간 제한을 설정합니다. 예를 들어, 일부 복잡한 상호작용으로 인해 특정 모델의 학습에 예기치 않게 오랜 시간이 필요할 경우 전체 모델링 실행을 지탱하기 위해 이를 원하지 않을 것입니다.

 **참고:** 목표가 명목(변수군) 필드인 경우 의사결정 목록 옵션이 사용 불가능합니다.

지원되는 알고리즘



지원 벡터 머신(SVM) 노드를 사용하면 데이터를 과적합 없이 두 개의 그룹 중 하나로 분류할 수 있습니다. SVM은 다수의 입력 필드가 있는 데이터 세트 등과 같은 광범위한 데이터 세트와 잘 작동합니다.



KNN(k-Nearest Neighbor) 노드는 새 케이스를 k 가 정수인 예측자 공간에서 가장 가까이 있는 k 오브젝트의 범주 또는 값과 연관시킵니다. 유사한 케이스는 서로 가까이 있고 유사하지 않은 케이스는 서로 멀리 떨어져 있습니다.



판별 분석은 로지스틱 회귀분석보다 엄격한 가정을 하지만 해당 가정이 충족되면 로지스틱 회귀 분석의 귀중한 대안 또는 보조물이 될 수 있습니다.



베이지안 네트워크 노드를 통해 관측 및 레코드된 증거를 실세계 지식과 조합하여 발생 우도를 확립함으로써 확률 모델을 작성할 수 있습니다. 이 노드는 주로 분류에 사용하는 TAN(Tree Augmented Naïve Bayes) 및 Markov Blanket 네트워크에 초점을 맞춥니다.



의사결정 목록 노드는 전체 채우기에 상대적인 주어진 이분형 결과의 상위 또는 하위 우도를 표시하는 부집단 또는 세그먼트를 식별합니다. 예를 들어, 캠페인을 이탈할 가능성이 없거나 우호적으로 응답할 가능성이 가장 많은 고객을 찾고 있습니다. 자체 사용자 정의 세그먼트를 추가하고 대체 모델을 나란히 미리보기하여 결과를 비교함으로써 비즈니스 지식을 모델에 통합할 수 있습니다. 의사결정 목록 모델은 각 규칙에 조건과 결과가 있는 규칙 목록으로 구성됩니다. 규칙은 순서대로 적용되며 매치하는 첫 번째 규칙이 결과를 결정합니다.



로지스틱 회귀분석은 입력 필드 값을 기반으로 레코드를 분류하는 통계 기법입니다. 선형 회귀와 유사하지만 숫자 범위 대신 범주형 대상 필드를 사용합니다.



CHAID 노드는 최적 분할을 식별하기 위해 카이제곱 통계량을 사용하여 의사결정 트리를 생성합니다. C&R 트리 및 QUEST 노드와 달리, CHAID는 비이분형 트리를 생성할 수 있으며 이는 일부 분할에 둘 이상의 분기가 있음을 의미합니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



QUEST 노드는 의사결정 트리를 작성하기 위한 이분형 분류 방법을 제공하며, 대형 C&R 트리 분석에 필요한 처리 시간을 줄이는 동시에 분류 트리 방법에서 찾아낸 경향을 줄여 더 많은 분할을 허용하는 입력을 선호하도록 설계되었습니다. 입력 필드는 숫자 범위(연속)일 수 있지만 대상 필드는 범주형이어야 합니다. 모든 분할은 이분형입니다.



분류 및 회귀(C&R) 트리 노드는 추가 관측값을 예측하거나 분류할 수 있게 하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 "순수"로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).



C5.0 노드는 의사결정 트리 또는 규칙 세트를 작성합니다. 모델은 각 수준에서 최대 정보 이익을 제공하는 필드를 기반으로 샘플을 분할하여 작동합니다. 대상 필드는 범주형이어야 합니다. 세 개 이상의 부집단으로의 다중 분할이 허용됩니다.



신경망 노드는 인간 두뇌가 정보를 처리하는 방법의 단순화된 모델을 사용합니다. 뉴런의 추상 버전을 닮은 상호연결된 많은 수의 단순 처리 장치를 시뮬레이션하여 작업합니다. 신경망은 강력한 범용 함수 추정량이며 학습하거나 적용하기 위해 약간의 통계 또는 수학적 지식이 필요합니다.



선형 회귀 모형은 목표와 하나 이상의 예측변수 간의 선형 관계를 기반으로 연속형 목표를 예측합니다.



선형 지원 벡터 머신(LSVM) 노드를 사용하면 과적합 없이 두 개의 그룹 중 하나로 데이터를 분류할 수 있습니다. LSVM은 선형이며, 다수의 레코드가 있는 데이터 세트와 같은 광범위한 데이터 세트와 함께 잘 작동합니다.



이 랜덤 트리 노드는 기존 C&RT 노드와 유사하지만, 빅 데이터를 처리하여 단일 트리를 작성하도록 설계되었으며 SPSS® Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시합니다. 랜덤 트리 노드는 추가 관측값을 예측하거나 분류하는 데 사용하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 순수로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).



Tree-AS 노드는 기존 CHAID 노드와 유사하지만, 빅 데이터를 처리하여 단일 트리를 작성하도록 설계되었으며 SPSS Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시합니다. 이 노드는 최적 분할을 식별하기 위해 카이제곱 통계량(CHAID)을 사용하여 의사결정 트리를 생성합니다. 이 CHAID의 사용은 일부 분할이 셋 이상의 분기를 가짐을 의미하는 비2진 트리를 생성할 수 있습니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



XGBoost Tree[®]는 트리 모델을 기본 모델로 사용하는 기울기 부스팅 알고리즘의 고급 구현입니다. 부스팅 알고리즘은 약한 분류자를 반복적으로 학습한 다음 이를 강한 최종 분류자에 추가합니다. XGBoost Tree는 유연성이 매우 뛰어나 대부분의 사용자에게 유용한 여러 매개변수를 제공하므로, SPSS Modeler의 XGBoost Tree 노드에는 핵심 기능과 일반적으로 사용되는 매개변수가 표시됩니다. 이 노드는 Python으로 구현됩니다.



XGBoost[®]는 기울기 부스팅 알고리즘의 고급 구현입니다. 부스팅 알고리즘은 약한 분류자를 반복적으로 학습한 다음 이를 강한 최종 분류자에 추가합니다. XGBoost는 유연성이 매우 뛰어나 대부분의 사용자에게 유용한 여러 매개변수를 제공하므로, SPSS Modeler의 XGBoost-AS 노드에는 핵심 기능과 일반적으로 사용되는 매개변수가 표시됩니다. XGBoost-AS 노드는 Spark로 구현됩니다.

참고: Analytic Server에서 실행할 Tree-AS를 선택하면 파티션 노드 업스트림이 있을 때 모델을 작성하지 못합니다. 이 경우 자동 분류자가 Analytic Server의 다른 모델링 노드와 작동하게 하려면 Tree-AS 모델 유형을 선택 취소하십시오.

③ 자동 분류자 노드 삭제 옵션

자동 분류자 노드의 삭제 탭을 사용하여 일정한 기준에 일치하지 않는 모델을 자동으로 삭제할 수 있습니다. 이 모델은 요약 보고서에 나열되지 않습니다.

전체 정확도의 최소 임계값 및 모델에 사용된 변수 수의 최대 임계값을 지정할 수 있습니다. 또한 플래그 목표의 경우 리프트, 이익, 곡선 아래 영역의 최소 임계값을 지정할 수 있습니다. 리프트 및 이익은 모델 탭에 지정된 대로 판별됩니다. 자세한 정보는 자동 분류자 노드 모델 옵션의 내용을 참조하십시오.

선택적으로 지정된 모든 기준을 만족하는 모델을 처음 생성할 때 실행을 중지하도록 노드를 구성할 수 있습니다. 자세한 정보는 자동화된 모델링 노드 중지 규칙의 내용을 참조하십시오.

④ 자동 분류자 노드 설정 옵션

자동 분류자 노드의 설정 탭에서 너깃에 사용 가능한 스코어 시간 옵션을 미리 구성할 수 있습니다.

양상블 모델이 생성한 필드 필터링. 양상블 노드에 반영되는 개별 모델이 생성한 모든 추가 필드의 출력에서 제거합니다. 모든 입력 모델에서 결합된 스코어에만 관심이 있는 경우 이 선택란을 선택합니다. 예를 들어, 분석 노드 또는 평가 노드를 사용하여 각 개별 입력 모델의 정확도와 결합된 스코어의 정확도를 비교하려는 경우에는 이 옵션을 선택 취소해야 합니다.

(5) 자동 숫자 노드

자동 숫자 노드는 여러 많은 방법을 사용하여 연속 숫자 범위 결과에 대한 모델을 추정하고 비교합니다. 이를 통해 단일 모델링 실행에서 다양한 접근 방식을 시도할 수 있습니다. 사용할 알고리즘을 선택하고 여러 옵션을 조합하여 실험할 수 있습니다. 예를 들어, 신경망, 선형 회귀, C&RT, CHAID 모델을 통해 하우스 값 예측하여 가장 효과적으로 수행되는 항목을 확인하고, 단계 선택, 전진, 후진 회귀분석 방법을 다양하게 조합해볼 수 있습니다. 노드는 가능한 모든 옵션 조합을 탐색하고, 사용자가 지정한 측도에 기반하여 각 후보 모델을 순위화하고 추가 분석 또는 스코어링에 사용할 때 가장 효과적인 항목을 저장합니다. 자세한 정보는 자동화된 모델링 노드 주제를 참조하십시오.

예제

지방 자치 단체에서 부동산 세금을 보다 정확히 추정하고 모든 자산을 검사하지 않고도 필요한 특정 특성의 값을 조정하고자 합니다. 자동 숫자 노드를 사용하면 분석가가 건물 유형, 이웃, 크기, 기타 알려진 요인에 기반하여 자산 가치를 예측하는 여러 모델을 생성하고 비교할 수 있습니다.

요구 사항

단일 목표 필드(역할이 **목표**로 설정됨)와 하나 이상의 입력 필드(역할이 **입력**으로 설정됨). 목표는 연속형(숫자 범위, 예: 나이 또는 소득) 필드여야 합니다. 입력 필드는 연속형 또는 범주형으로, 일부 입력은 일부 모델 유형에 적합하지 않다는 제한사항이 있습니다. 예를 들어, C&R 트리 모델은 입력으로 범주형 문자열 필드를 사용하지만 선형 회귀 모형은 이 필드를 사용할 수 없으며 지정된 경우 해당 필드를 무시합니다. 요구 사항은 개별 모델링 노드

를 사용할 때와 동일합니다. 예를 들어, CHAID 모델은 생성 위치(CHAID 노드 또는 자동 숫자 노드)에 상관없이 동일하게 작동합니다.

빈도 및 가중치 필드

빈도 및 가중치는 다른 레코드에 비해 일부 레코드에 추가 중요도를 부여하는 용도로 사용되며, 이는 예를 들어, 작성 데이터 세트가 상위 모집단 섹션을 실제보다 낮게 표시(가중치)함을 사용자가 알고 있거나 한 레코드가 많은 동일한 케이스를 표시(빈도)하기 때문입니다. 이를 지정한 경우 빈도 필드는 C&R 트리 및 CHAID 알고리즘에서 사용할 수 있습니다. 가중 필드는 C&RT, CHAID, 회귀분석, GenLin 알고리즘에서 사용할 수 있습니다. 다른 모델 유형은 이러한 필드를 무시하고 모델을 작성합니다. 빈도 및 가중 필드는 모델 작성에만 사용되고, 모델 평가 또는 스코어링에서는 고려되지 않습니다. 자세한 정보는 빈도 및 가중 필드 사용의 내용을 참조하십시오.

접두문자

자동 숫자 노드의 너깃에 테이블 노드를 첨부할 경우 \$ 접두문자로 시작하는 이름의 테이블에 새 변수가 여러 개 있습니다.

스코어링 중에 생성된 필드 이름은 표준 접두문자가 아닌 목표 필드를 기반으로 합니다. 서로 다른 모델 유형은 서로 다른 접두문자 집합을 사용합니다.

예를 들어 접두문자 \$G, \$R, \$C는 각각 일반화 선형 모델, CHAID 모델, C5.0 모델을 통해 생성되는 예측에 대한 접두문자로 사용됩니다. \$X는 일반적으로 앙상블을 사용하여 생성되고, \$XR, \$XS, \$XF는 목표 필드가 연속형, 범주형 또는 플래그 필드인 경우에 각각 접두문자로 사용됩니다.

\$.E 접두문자는 연속형 대상의 예측 신뢰도에 사용됩니다. 예를 들어 \$XRE는 앙상블 연속형 예측 신뢰도에 대한 접두문자로 사용됩니다. \$RC 및 \$GE는 일반화 선형 모델의 단일 예측 신뢰도에 대한 접두문자입니다.

지원되는 모델 유형

지원되는 모델 유형으로는, 신경망, C&R 트리, CHAID, 회귀분석, GenLin, 최근접 이웃, SVM, XGBoost Linear, GLE 및 XGBoost-AS를 포함합니다. 추가 정보는 자동 숫자 노드 고급 옵션의 내용을 참조하십시오.

연속 기계 학습

모델링의 불편한 점은 시간이 지남에 따라 데이터가 변경되어 모델이 구식이 된다는 것입니다. 이를 일반적으로 *모델 드리프트* 또는 *개념 드리프트*라고 합니다. 모델 드리프트를 효과적으로 극복하기 위해 SPSS Modeler는 연속 자동 기계 학습을 제공합니다. 이 기능은 자동 분류자 노드 및 자동 숫자 노드 모델 너깃에 사용할 수 있습니다. 자세한 정보는 연속 기계 학습의 내용을 참조하십시오.

① 자동 숫자 노드 모델 옵션

자동 숫자 노드의 모델 탭에서는 모델 비교에 사용되는 기준과 함께 저장할 모델 수를 지정할 수 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

교차 검증. 교차 검증을 수행하면 학습을 실행하는 *알려진 데이터*의 데이터 세트(학습 데이터 세트) 및 모델을 테스트하는 대상이 되는 *알려지지 않은 데이터*의 데이터 세트(검증 데이터 세트 또는 테스트 세트)를 모델링합니다. 교차 검증은 모델이 과적합 또는 선택 편향과 같은 문제점에 플래그를 지정하기 위해 예측에 사용되지 않는 새로운 데이터를 예측하는 기능을 테스트하기 위해 사용합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

모델 순위화 기준. 모델을 비교하는 데 사용되는 기준을 지정합니다.

- **상관관계.** 각 레코드의 관측값과 모델에서 예측한 값 사이의 Pearson 상관. 상관관계는 두 변수 사이의 선형 상관관계에 대한 척도로, 값이 1에 가까울수록 더 강한 관계임을 나타냅니다. (상관 관계 값 범위는 -1(완벽한 음의 관계를 나타냄)에서 +1(완벽한 양의 관계를 나타냄) 사이입니다. 0의 값은 선형 관계가 아님을 나타내지만, 음의 상관관계인 모델은 순위가 가장 낮습니다.)
- **필드 수.** 모델에서 예측변수로 사용되는 필드 수. 필드가 더 적은 모델을 선택하면 데이터 준비를 간소화하고 일부 경우에 성능을 향상시킬 수 있습니다.
- **상대 오차.** 상대 오차는 모델에서 예측한 항목에서 관측값 분산을 평균에서 관측값의 분산으로 나눈 비율입니다. 실제로, 이 경우 예측으로 대상 필드의 평균값을 리턴하는 널 또는 절편 모델과 비교했을 때 이 모델의 상대적인 성능을 비교합니다. 좋은 모델인 경우 이 값은 1 미만이어야 합니다. 이는 모델이 널 모델보다 정확함을 의미합니다. 상대 오차가 1보다 큰 모델은 널 모델보다 덜 정확하므로 유용하지 않습니다. 선형 회귀 모형의 경우 상대 오차는 상관관계의 제곱과 동일하고 새 정보를 추가하지 않습니다. 비선형 모델의 경우 상대 오차는 상관관계와 무관하며, 모델 성능 평가 시 추가 척도를 제공합니다.

모델 순위화 사용. 파티션이 사용 중인 경우 순위의 기준(학습 파티션 또는 검정 분할)을 지정할 수 있습니다. 큰 데이터 세트의 경우 모델의 예비 심사에 파티션을 사용하면 성능이 상당히 개선될 수 있습니다.

사용할 모델 수. 노드에서 생성한 모델 너깃에 표시할 최대 모델 수를 지정합니다. 지정된 순위화 기준에 따라 가장 높은 순위의 모델이 나열됩니다. 이 한계를 늘리면 더 많은 모델에 대한 결과를 비교할 수 있지만 성능이 느려질 수 있습니다. 허용 가능한 최대값은 100입니다.

예측자 중요도 계산. 적절한 중요도 축도를 생성하는 모델의 경우 모델 측정 시 각 예측변수의 상대적 중요도를 나타내는 차트를 표시할 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 예측변수 중요도는 일부 모델을 계산하는 데 필요한 시간을 연장할 수 있으며 단순히 여러 다른 모델을 광범위하게 비교하려는 경우 권장되지 않습니다. 보다 자세하게 탐색하려는 모델로 분석 범위를 좁히는 경우 더 유용합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

다음 경우에는 모델을 유지하지 않음. 상관관계, 상대 오차, 사용된 필드 수의 임계값을 지정합니다. 이 기준을 만족하는데 실패한 모델은 삭제되고 요약 보고서에 나열되지 않습니다.

- **상관관계 미만 기준.** 요약 보고서에 포함할 모델의 최소 상관관계(절대값 기준).
- **사용된 필드 수의 초과 기준.** 포함할 모델에서 사용할 최대 필드 수.
- **상대 오차의 초과 기준.** 포함할 모델의 최대 상대 오차입니다.

선택적으로 지정된 모든 기준을 만족하는 모델을 처음 생성할 때 실행을 중지하도록 노드를 구성할 수 있습니다. 자세한 정보는 자동화된 모델링 노드 중지 규칙의 내용을 참조하십시오.

② 자동 숫자 노드 고급 옵션

자동 숫자 노드의 고급 탭에서는 중지 규칙을 사용 및 지정할 알고리즘과 옵션을 선택할 수 있습니다.

사용한 모델. 왼쪽 열의 선택란을 사용하여 비교에 포함할 모델 유형(알고리즘)을 선택합니다. 더 많은 유형을 선택할수록 모델이 더 많이 작성되고 처리 시간은 더 오래 걸립니다.

모델 유형. 사용 가능한 알고리즘을 나열합니다(아래 참조).

모델 매개변수. 각 모델 유형마다 기본 설정을 사용하거나 **지정**을 선택하여 각 모델 유형에 맞는 옵션을 선택할 수 있습니다. 특정 옵션은 여러 옵션 또는 조합을 선택할 수 있다는 점을 제외하면, 개별 모델링 노드에서 사용 가능한 옵션과 유사합니다. 예를 들어, 신경망 모델을 비교하는 경우 여섯 개의 학습 방법 중에서 하나를 선택하기 보다는 모두를 선택하여 단일 전달에서 여섯 개의 모델을 학습시킬 수 있습니다.

모델 수. 현재 설정을 기준으로 하여 각 알고리즘에 대해 생성되는 모델 수를 나열합니다. 옵션을 조합할 경우 모델 수가 빠르게 증가하므로 특히 큰 데이터 세트를 사용할 때에는 이 수에 세심한 주의를 기울일 것을 강력히 권장합니다.

단일 모델 작성에 소요되는 최대 시간 제한. (K-평균, 코호넨, TwoStep, SVM, KNN, Bayes Net, 의사결정 목록 모델만) 한 모델의 최대 시간 제한을 설정합니다. 예를 들어, 일부 복잡한 상호작용으로 인해 특정 모델의 학습에 예기치 않게 오랜 시간이 필요할 경우 전체 모델링 실행을 지탱하기 위해 이를 원하지 않을 것입니다.

지원되는 알고리즘



신경망 노드는 인간 두뇌가 정보를 처리하는 방법의 단순화된 모델을 사용합니다. 뉴런의 추상 버전을 닮은 상호연결된 많은 수의 단순 처리 장치를 시뮬레이션하여 작업합니다. 신경망은 강력한 범용 함수 추정량이며 학습하거나 적용하기 위해 약간의 통계 또는 수학적 지식이 필요합니다.



분류 및 회귀(C&R) 트리 노드는 추가 관측값을 예측하거나 분류할 수 있게 하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 "순수"로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).



CHAID 노드는 최적 분할을 식별하기 위해 카이제곱 통계량을 사용하여 의사결정 트리를 생성합니다. C&R 트리 및 QUEST 노드와 달리, CHAID는 비이분형 트리를 생성할 수 있으며 이는 일부 분할에 둘 이상의 분기가 있음을 의미합니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



선형 회귀는 데이터를 요약통계하고 예측 및 실제 출력 값 사이의 불일치를 최소화하는 직선이나 표면에 적합하게 하여 예측하기 위한 일반적인 통계 기법입니다.



일반화 선형 모델은 종속변수가 요인과 선형적으로 관련되고 지정된 연결함수를 통해 공변되도록 일반 선형 모델을 확장합니다. 더욱이 모델을 사용하면 종속변수가 비정규 분포를 가질 수 있습니다. 선형 회귀, 로지스틱 회귀분석, 카운트 데이터에 대한 로그선형 모델, 간격 중도절단 생존 모델을 포함하여 상당수 통계 모델의 기능을 포함합니다.



KNN(k -Nearest Neighbor) 노드는 새 케이스를 k 가 정수인 예측자 공간에서 가장 가까이에 있는 k 오브젝트의 범주 또는 값과 연관시킵니다. 유사한 케이스는 서로 가까이에 있고 유사하지 않은 케이스는 서로 멀리 떨어져 있습니다.



지원 벡터 머신(SVM) 노드를 사용하면 데이터를 과적합 없이 두 개의 그룹 중 하나로 분류할 수 있습니다. SVM은 다수의 입력 필드가 있는 데이터 세트 등과 같은 광범위한 데이터 세트와 잘 작동합니다.



선형 회귀 모형은 목표와 하나 이상의 예측변수 간의 선형 관계를 기반으로 연속형 목표를 예측합니다.



선형 지원 벡터 머신(LSVM) 노드를 사용하면 과적합 없이 두 개의 그룹 중 하나로 데이터를 분류할 수 있습니다. LSVM은 선형이며, 다수의 레코드가 있는 데이터 세트와 같은 광범위한 데이터 세트와 함께 잘 작동합니다.



이 랜덤 트리 노드는 기존 C&RT 노드와 유사하지만, 빅 데이터를 처리하여 단일 트리를 작성하도록 설계되었으며 SPSS® Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시합니다. 랜덤 트리 노드는 추가 관측값을 예측하거나 분류하는 데 사용하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 순수로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).



Tree-AS 노드는 기존 CHAID 노드와 유사하지만, 빅 데이터를 처리하여 단일 트리를 작성하도록 설계되었으며 SPSS Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시합니다. 이 노드는 최적 분할을 식별하기 위해 카이제곱 통계량(CHAID)을 사용하여 의사결정 트리를 생성합니다. 이 CHAID의 사용은 일부 분할이 셋 이상의 분기를 가짐을 의미하는 비2진 트리를 생성할 수 있습니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



XGBoost Linear©는 선형 모델을 기본 모델로 사용하는 기울기 부스팅 알고리즘의 고급 구현입니다. 부스팅 알고리즘은 약한 분류자를 반복적으로 학습한 다음 이를 강한 최종 분류자에 추가합니다. SPSS Modeler의 XGBoost Linear 노드는 Python으로 구현됩니다.



GLE는 목표가 비정규 분포를 가질 수 있고 지정된 연결함수를 통해 요인 및 공변량과 선형적으로 관련되며 관측값을 상관시킬 수 있도록 선형 모델을 확장합니다. 일반화 선형 혼합 모델은 단순 선형 회귀에서 비정규 장기적인 데이터에 대한 복합 다중 수준 모델에 이르기까지 다양한 모델을 포함합니다.



XGBoost©는 기울기 부스팅 알고리즘의 고급 구현입니다. 부스팅 알고리즘은 약한 분류자를 반복적으로 학습한 다음 이를 강한 최종 분류자에 추가합니다. XGBoost는 유연성이 매우 뛰어나 대부분의 사용자에게 유용한 여러 매개변수를 제공하므로, SPSS Modeler의 XGBoost-AS 노드에는 핵심 기능과 일반적으로 사용되는 매개변수가 표시됩니다. XGBoost-AS 노드는 Spark로 구현됩니다.

③ 자동 숫자 노드 설정 옵션

자동 숫자 노드의 설정 탭에서는 너깃에서 사용 가능한 스코어-시간 옵션을 사전 구성할 수 있습니다.

양상블 모델이 생성한 필드 필터링. 양상블 노드에 반영되는 개별 모델이 생성한 모든 추가 필드의 출력에서 제거합니다. 모든 입력 모델에서 결합된 스코어에만 관심이 있는 경우 이 선택란을 선택합니다. 예를 들어, 분석 노드 또는 평가 노드를 사용하여 각 개별 입력 모델의 정확도와 결합된 스코어의 정확도를 비교하려는 경우에는 이 옵션을 선택 취소해야 합니다.

표준 오차 계산. 연속형(숫자 범위) 목표인 경우 표준 오차 계산은 기본적으로 측정 또는 추정된 값과 참 값 사이의 차이를 계산하고 이러한 추정이 얼마나 일치하는지 표시하는 방법으로 실행됩니다.

(6) 자동 군집 노드

자동 군집 노드는 특성이 유사한 레코드 그룹을 식별하는 군집 모델을 추정하고 비교합니다. 노드는 다른 자동화된 모델링 노드와 동일한 방식으로 작동하며 단일 모델링 전달에서 여러 옵션 조합으로 실험할 수 있게 합니다. 군집 모델의 유용성을 필터링하고 순위화하며 특정 필드의 중요성을 기반으로 측도를 제공하려고 시도하는 기본 측도를 사용하여 모델을 비교할 수 있습니다.

군집 모델은 종종 후속 분석의 입력으로 사용할 수 있는 그룹을 식별하는 데 사용됩니다. 예를 들어, 소득과 같은 인구 통계적 특성이나 과거에 구매한 서비스를 기준으로 하여 고객 그룹을 목표화할 수 있습니다. 그룹 및 특성에 대한 사전 지식 없이도(검색할 그룹 수나 그룹을 정의하는 데 사용할 기능을 몰라도 됨) 이를 수행할 수 있습니다. 군집 모델은 목표 필드를 사용하지 않고 참 또는 거짓으로 평가할 수 있는 특정 예측을 리턴하지 않기 때문에 종종 자율 학습 모델이라 부릅니다. 군집 모델의 값은 데이터에서 관심 있는 그룹을 캡처하고 이 그룹에 대한 유용한 설명을 제공하는 기능으로 판별됩니다. 자세한 정보는 군집 모델의 내용을 참조하십시오.

요구사항. 관심 있는 특성을 정의하는 하나 이상의 필드입니다. 군집 모델은 참 또는 거짓으로 평가할 수 있는 특정 예측을 수행하지 않기 때문에 다른 모델과 동일한 방식으로 목표 필드를 사용하지 않습니다. 대신에 관련될 수 있는 케이스 그룹을 식별하는 데 사용됩니다. 예를 들어, 주어진 컴퓨터가 오퍼에 이탈 또는 응답할지 여부를 예측하기 위해 군집 모델을 사용할 수 없습니다. 하지만 고객의 경향을 기준으로 하여 그룹에 고객을 지정하기 위해 군집 모델을 사용할 수는 있습니다. 가중치 및 빈도 필드는 사용하지 않습니다.

평가 필드. 목표가 사용되지 않을 때 선택적으로 모델 비교에 사용할 하나 이상의 평가 필드를 지정할 수 있습니다. 군집 모델의 유용성은 군집이 이러한 필드를 구별하는 정도를 측정하여 평가할 수 있습니다.

지원되는 모델 유형

지원되는 모델 유형으로는 이단계, K-평균, 코호넨, One-Class SVM 및 K-Means-AS가 있습니다.

① 자동 군집 노드 모델 옵션

자동 군집 노드의 모델 탭으로 모델을 비교하는 데 사용되는 기준과 함께 저장할 모델 수를 지정할 수 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

모델 순위화 기준. 모델을 비교하고 순위화하는 데 사용되는 기준을 지정합니다.

- **실루엣.** 군집 결합 및 분리를 모두 측정하는 지수입니다. 자세한 정보는 아래의 **실루엣 순위화 척도**를 참조하십시오.
- **군집 수.** 모델의 군집 수입니다.
- **가장 작은 군집 크기.** 최소 군집 크기입니다.
- **가장 큰 군집 크기.** 최대 군집 크기입니다.
- **최소 / 최대 군집.** 가장 작은 군집의 크기 대 가장 큰 군집의 크기 비율입니다.
- **중요도.** 필드 탭의 평가 필드 중요도입니다. 중요도는 평가 필드가 지정된 경우에만 계산할 수 있음에 유의하십시오.

모델 순위화 사용. 파티션이 사용 중인 경우 순위가 학습 데이터 세트 또는 검정 세트를 기준으로 하는지 여부를 지정할 수 있습니다. 큰 데이터 세트의 경우 모델의 예비 심사에 파티션을 사용하면 성능이 상당히 개선될 수 있습니다.

유지할 모델 수. 노드가 생성한 너깃에 나열할 모델의 최대 수를 지정합니다. 지정된 순위화 기준에 따라 가장 높은 순위의 모델이 나열됩니다. 이 한계를 높이면 성능이 저하될 수 있음에 유의하십시오. 허용 가능한 최대값은 100입니다.

실루엣 순위화 척도

기본 순위화 척도, 실루엣의 기본값은 0입니다. 0 미만(즉, 음수)의 값은 지정된 군집의 포인트와 케이스 간 평균 거리가 또 다른 군집의 포인트에 대한 최소 평균 거리를 초과함을 나타내기 때문입니다. 따라서 실루엣이 음수인 모델은 삭제해도 안전합니다.

순위화 측도는 실제로 군집 결합(조밀하게 결합된 군집을 포함한 모델 선호) 및 군집 분리(멀리 떨어진 군집을 포함한 모델 선호) 개념을 조합하는 수정된 실루엣 계수입니다. 평균 실루엣 계수는 단순히 각 개별 케이스마다 다음 계산의 모든 케이스에 대한 평균입니다.

$$(B - A) / \max(A, B)$$

여기서, A 는 케이스로부터 케이스가 속한 군집의 중심값까지의 거리이고 B 는 케이스로부터 다른 모든 군집의 중심값까지의 최소 거리입니다.

실루엣 계수(및 평균) 범위는 -1(매우 빈약한 모델을 나타냄) ~ 1(우수한 모델을 나타냄)입니다. 총 케이스 수준(총 실루엣을 생성함) 또는 군집 수준(군집 실루엣을 생성함)에서 평균을 구할 수 있습니다. 거리는 유클리디안 거리를 사용하여 계산할 수 있습니다.

② 자동 군집 노드 고급 옵션

자동 군집 노드의 고급 탭으로 파티션을 적용하고(가능한 경우), 사용할 알고리즘을 선택하고, 중지 규칙을 지정할 수 있습니다.

사용한 모델. 왼쪽 열의 선택란을 사용하여 비교에 포함할 모델 유형(알고리즘)을 선택합니다. 더 많은 유형을 선택할수록 모델이 더 많이 작성되고 처리 시간은 더 오래 걸립니다.

모델 유형. 사용 가능한 알고리즘을 나열합니다(아래 참조).

모델 매개변수. 각 모델 유형마다 기본 설정을 사용하거나 지정을 선택하여 각 모델 유형에 맞는 옵션을 선택할 수 있습니다. 특정 옵션은 여러 옵션 또는 조합을 선택할 수 있다는 점을 제외하면, 개별 모델링 노드에서 사용 가능한 옵션과 유사합니다. 예를 들어, 신경망 모델을 비교하는 경우 여섯 개의 학습 방법 중에서 하나를 선택하기 보다는 모두를 선택하여 단일 전달에서 여섯 개의 모델을 학습시킬 수 있습니다.

모델 수. 현재 설정을 기준으로 하여 각 알고리즘에 대해 생성되는 모델 수를 나열합니다. 옵션을 조합할 경우 모델 수가 빠르게 증가하므로 특히 큰 데이터 세트를 사용할 때에는 이 수에 세심한 주의를 기울일 것을 강력히 권장합니다.

단일 모델 작성에 소요되는 최대 시간 제한. (K-평균, 코호넨, TwoStep, SVM, KNN, Bayes Net, 의사결정 목록 모델만) 한 모델의 최대 시간 제한을 설정합니다. 예를 들어, 일부 복잡한 상호작용으로 인해 특정 모델의 학습에 예기치 않게 오랜 시간이 필요할 경우 전체 모델링 실행을 지탱하기 위해 이를 원하지 않을 것입니다.

지원되는 알고리즘



K-평균 노드는 데이터 세트를 고유 그룹(또는 군집)으로 군집화합니다. 이 방법은 고정된 수의 군집을 정의하고 반복적으로 레코드를 군집에 지정하며, 추가 세분화가 더 이상 모델을 향상시킬 수 없을 때까지 군집중심을 조정합니다. 결과를 예상하는 대신 k -평균은 자율 학습으로 알려진 프로세스를 사용하여 입력 필드 세트의 패턴을 찾아냅니다.



코호넨 노드는 데이터 세트를 고유 그룹으로 군집화하는 데 사용할 수 있는 신경망 유형을 생성합니다. 네트워크가 완전히 숙달되면, 유사 레코드는 출력 맵 가까이 있지만, 다른 레코드는 멀리 떨어져 있을 것입니다. 모델 너깅에서 각 단위별로 캡처된 관측값을 살펴 강한 단위를 식별할 수 있습니다. 이것은 적당한 군집 수에 대한 감각을 제공할 것입니다.



이단계 노드는 2단계 군집방법을 사용합니다. 첫 번째 단계는 원시 입력 데이터를 관리 가능한 하위 군집 세트로 압축하기 위해 데이터를 통한 단일 전달을 수행합니다. 두 번째 단계는 계층적 군집 방법을 사용하여 하위 군집을 점점 더 큰 군집으로 계속해서 병합하는 것입니다. 이단계는 학습 데이터에 대한 최적 군집 수를 자동으로 평가하는 장점이 있습니다. 혼합 필드 유형과 대형 데이터 세트를 효율적으로 처리할 수 있습니다.



One-Class SVM 노드에는 자율 학습 알고리즘이 사용됩니다. 이 노드는 이상 탐지에 사용할 수 있습니다. 주어진 표본 세트의 소프트 경계를 탐지하여 새 포인트를 해당 세트에 속하거나 속하지 않는 것으로 분류합니다. SPSS® Modeler의 이 One-Class SVM 모델링 노드는 Python으로 구현되며, scikit-learn© Python 라이브러리가 필요합니다.



K-평균은 일반적으로 가장 많이 사용되는 군집 알고리즘 중 하나입니다. 이는 데이터 점을 사전정의된 수의 군집으로 군집화합니다. SPSS Modeler의 K-Means-AS 노드는 Spark로 구현됩니다. K-평균 알고리즘에 대한 세부사항은 <https://spark.apache.org/docs/2.2.0/ml-clustering.html>의 내용을 참조하십시오. K-Means-AS 노드는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

③ 자동 군집 노드 삭제 옵션

자동 군집 노드의 삭제 탭을 사용하여 일정한 기준에 일치하지 않는 모델을 자동으로 삭제할 수 있습니다. 이 모델은 모델 너깅에 나열되지 않습니다.

최소 실루엣 값, 군집 수, 군집 크기, 모델에 사용된 평가 필드의 중요도를 지정할 수 있습니다. 군집 수와 크기 및 실루엣은 모델링 노드에 지정된 대로 판별됩니다. 자세한 정보는 자동 군집 노드 모델 옵션의 내용을 참조하십시오.

선택적으로 지정된 모든 기준을 만족하는 모델을 처음 생성할 때 실행을 중지하도록 노드를 구성할 수 있습니다. 자세한 정보는 자동화된 모델링 노드 중지 규칙의 내용을 참조하십시오.

(7) 자동화된 모델 너깃

자동화된 모델링 노드를 실행하면 노드는 가능한 모든 옵션 조합에 대해 후보 모델을 추정하고, 사용자가 지정한 척도에 따라 각 후보 모델을 순위화하고, 자동화된 복합 모델 너깃에서 최상의 모델을 저장합니다. 이 모델 너깃은 실제로 노드에서 생성된 하나 이상의 모델 세트를 포함합니다. 이러한 모델은 스코어링에 사용하기 위해 개별적으로 찾아보거나 선택할 수 있습니다. 각 모델에 대해 모델 유형 및 작성 시간과 모델 유형에 적절한 경우 기타 여러 척도가 나열됩니다. 이 열을 기준으로 테이블을 정렬하여 관심이 가장 많은 모델을 빠르게 식별할 수 있습니다.

- 개별 모델 너깃을 찾아보려면 너깃 아이콘을 두 번 클릭하십시오. 여기에서 스트림 캔버스로 해당 모델의 모델링 노드를 생성하거나 모델 팔레트에 모델 너깃의 사본을 생성할 수 있습니다.
- 썸네일 그래프는 아래 요약된 대로, 각 모델 유형에 대한 빠른 시각적 평가를 제공합니다. 썸네일을 두 번 클릭하면 전체 크기 그래프를 생성할 수 있습니다. 전체 크기 도표는 최대 1000개의 포인트를 표시하며, 데이터 세트가 추가 항목을 포함하는 경우 표본에 기반합니다. (산점도인 경우에만 표시될 때마다 그래프가 재생성되므로, 업스트림 데이터(예: 난수 표본 또는 난수 시드 설정이 선택되지 않은 경우 파티션의 업데이트)의 변경은 산점도를 다시 그릴 때마다 반영될 수 있습니다.
- 도구 모음을 사용하여 모델 탭에서 특정 열의 표시 또는 숨기기를 수행하거나 테이블을 정렬하는 데 사용되는 열을 변경하십시오. (또한 열 헤더를 클릭하여 정렬을 변경할 수도 있습니다.)
- 삭제 단추를 사용하여 사용되지 않는 모델을 영구적으로 제거하십시오.
- 열을 다시 정렬하려면 열 헤더를 클릭하고 열을 원하는 위치로 끄십시오.
- 파티션이 사용 중인 경우 해당되는 경우 훈련 또는 검정 분할에 대한 결과를 볼 수 있습니다.

특정 열은 아래 설명한 대로, 비교할 모델 유형에 따라 달라집니다.

이분형 목표

- 이분형 모델에서 썸네일 그래프는 예측 값에 오버레이된 형식으로 실제 값의 분포를 표시하여 각 범주에서 레코드가 올바르게 예측된 정보를 시각적으로 빠르게 표시할 수 있습니다.
- 순위화 기준은 자동 분류자 모델링 노드의 옵션과 매치됩니다. 자세한 정보는 자동 분류자 노드 모델 옵션 주제를 참조하십시오.

- 최대 수익을 위해 최대값이 나타나는 백분위수도 보고됩니다.
- 누적 리프트의 경우 도구 모음을 사용하여 선택된 백분위수를 변경할 수 있습니다.

명목 목표

- 명목(세트) 모델에서 썸네일 그래프는 예측 값에 오버레이된 형식으로 실제 값의 분포를 표시하여 각 범주에서 레코드가 올바르게 예측된 정보를 시각적으로 빠르게 표시할 수 있습니다.
- 순위화 기준은 자동 분류자 모델링 노드의 옵션과 매치됩니다. 자세한 정보는 자동 분류자 노드 모델 옵션 주제를 참조하십시오.

연속형 목표

- 연속(숫자 범위) 모델의 경우 그래프는 각 모델의 관측 값에 대한 예측을 구성하여, 둘 사이의 상관관계에 대한 빠른 시각적 표시를 제공합니다. 좋은 모델인 경우 포인트는 그래프에서 무작위로 퍼져있는 대신, 대각선으로 군집되는 경향이 있습니다.
- 순위화 기준은 자동 수치 모델링 노드의 옵션과 매치됩니다. 자세한 정보는 자동 숫자 노드 모델 옵션 주제를 참조하십시오.

군집 목표

- 군집 모델의 경우 그래프는 각 모델의 군집에 대한 빈도를 구성하여, 군집 분포의 빠른 시각적 표시를 제공합니다.
- 순위화 기준은 자동 군집 모델링 노드의 옵션과 매치됩니다. 자세한 정보는 자동 군집 노드 모델 옵션 주제를 참조하십시오.

스코어링을 위해 모델 선택

사용? 열에서는 스코어링에 사용할 모델을 선택할 수 있습니다.

- 이분형, 명목, 숫자 목표의 경우 여러 스코어링 모델을 선택하고 하나의 앙상블 모델 너깃에 스코어를 결합할 수 있습니다. 여러 모델로부터 예측을 결합함으로써 개별 모델의 제한사항을 피할 수 있으며 이를 통해 종종 모델 중 하나에서 확보할 수 있는 것보다 전반적인 정확도가 높아집니다.
- 군집 모델의 경우 한 번에 하나의 스코어링 모델만 선택할 수 있습니다. 기본적으로 상위 순위 항목이 먼저 선택됩니다.

① 노드 및 모델 생성

작성된 자동화된 모델링 노드 또는 복합 자동화된 모델 너깃의 사본을 생성할 수 있습니다. 예를 들어 자동화된 모델 너깃이 작성된 원래 스트림이 없는 경우 유용할 수 있습니다. 또는 자동화된 모델 너깃에 나열된 개별 모델에 대한 너깃 또는 모델링 노드를 생성할 수 있습니다.

자동화된 모델링 너깃

생성 메뉴에서 **모델을 팔레트로**를 선택하여 자동화된 모델 너깃을 모델 팔레트에 추가합니다. 생성된 모델은 스트림을 재실행하지 않고도 그대로 저장 또는 사용될 수 있습니다.

또는 생성 메뉴에서 **모델링 노드 생성**을 선택하여 스트림 캔버스에 모델링 노드를 추가할 수 있습니다. 이 노드는 전체 모델링 실행을 반복하지 않고도 선택한 모델을 재평가하는 데 사용할 수 있습니다.

개별 모델링 너깃

1. **모델** 메뉴에서 필요한 개별 너깃을 두 번 클릭하십시오. 새 대화 상자에서 해당 너깃의 사본이 열립니다.
2. 새 대화 상자의 생성 메뉴에서 **모델을 팔레트로**를 선택하여 개별 모델링 너깃을 모델 팔레트에 추가하십시오.
3. 또는 새 대화 상자의 생성 메뉴에서 **모델링 노드 생성**을 선택하여 스트림 캔버스에 개별 모델링 노드를 추가할 수 있습니다.

② 평가 차트 생성

이분형 모델에서만 각 모델의 성능을 평가 및 비교하는 시각적 방법을 제공하는 평가 차트를 생성할 수 있습니다. 평가 차트는 자동 숫자 또는 자동 군집 노드에서 생성된 모델에서는 사용할 수 없습니다.

1. 자동 분류자 자동화된 모델 너깃의 *사용?* 열 아래에서 평가할 모델을 선택하십시오.
2. 생성 메뉴에서 **평가 차트**를 선택하십시오. 평가 차트 대화 상자가 표시됩니다.
3. 차트 유형 및 원하는 경우 기타 옵션을 선택하십시오.

③ 평가 그래프

자동화된 모델 너깃의 모델 탭에서는 드릴다운하여 표시된 각 모델의 개별 그래프를 표시할 수 있습니다. 자동 분류자 및 자동 숫자 너깃의 경우 그래프 탭에서는 결합된 모든 모델의 결과를 반영하는 그래프 및 예측자 중요도를 모두 표시합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

자동 분류자의 경우 분포 그래프가 표시되지만, 자동 숫자의 경우 다중 도표(산점도라고도 함)가 표시됩니다.

④ 자동화된 모델 너깃 요약

자동화된 모델 너깃의 요약 탭에는 각 모델에 대한 작성 설정과 함께 사용된 필드, 모든 모델을 작성하기 위해 경과된 총 시간, 사용된 모델 순위화 기준이 나열됩니다.

⑤ 연속 기계 학습

IBM의 연구 결과, 생물학의 자연 선택에서 영감을 받은 *연속 기계 학습*을 자동 분류자 노드와 자동 숫자 노드에 제공합니다.

모델링의 불편한 점은 시간이 지남에 따라 데이터가 변경되어 모델이 구식이 된다는 것입니다. 이를 일반적으로 *모델 드리프트* 또는 *개념 드리프트*라고 합니다. 모델 드리프트를 효과적으로 극복하기 위해 SPSS Modeler는 연속 자동 기계 학습을 제공합니다.

모델 드리프트란 무엇일까요? 히스토리 데이터를 기반으로 모델을 빌드하면 모델이 정체될 수 있습니다. 대부분의 경우 새로운 변형, 새로운 패턴, 새로운 추세 등 이전 히스토리 데이터가 포착하지 못하는 새로운 데이터가 항상 유입되고 있습니다. 이 문제를 해결하기 위해 IBM은 종의 자연 선택이라는 유명한 생물학 현상에서 영감을 받았습니다. 모델을 종으로 생각하고 데이터를 자연으로 생각하십시오. 자연이 종을 선택하는 것처럼 데이터가 모델을 선택하도록 해야 합니다. 모델과 종 사이에는 한 가지 큰 차이가 있습니다. 종은 진화할 수 있지만 모델은 일단 작성되면 정적입니다.

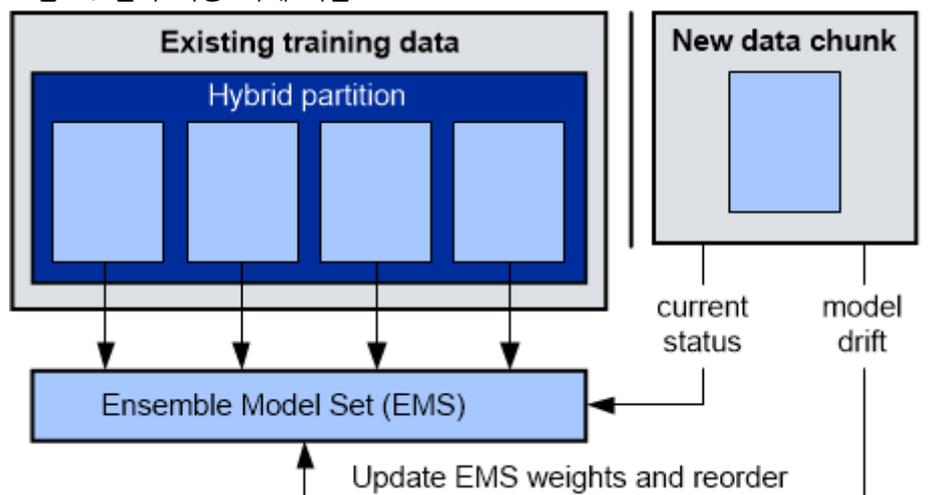
종이 진화하기 위해서는 두 가지 전제조건이 있습니다. 첫 번째는 유전자 돌연변이이고 두 번째는 인구입니다. 이제 모델링 관점에서 첫 번째 전제조건(유전자 돌연변이)을 충족하려면 기존 모델에 새로운 데이터 변경사항을 도입해야 합니다. 두 번째 전제조건(인구)을 충족하려면 단 하나의 모델이 아닌 여러 개의 모델을 사용해야 합니다. 여러 모델을 나타낼 수 있는 것은 무엇일까요? 바로 앙상블 모델 세트(EMS)입니다!

다음 그림은 EMS가 어떻게 진화할 수 있는지를 보여줍니다. 그림의 왼쪽 상단 부분은 하이브리드 파티션이 있는 히스토리 데이터를 나타냅니다. 하이브리드 파티션은 초기 EMS를 풍부하게 만듭니다. 그림의 오른쪽 상단은 사용 가능해진 새 데이터 청크를 나타내며, 양쪽에 세로 막대가 있습니다. 왼쪽 세로 막대는 현재 상태를 나타내며, 오른쪽 세로 막대는 모델 드리프트 위험이 있는 상태를 나타냅니다. 새로운 연속 기계 학습 라운드마다 모델을 진화시키고 모델 드리프트를 방지하기 위해 두 개의 단계가 수행됩니다.

먼저 기존 학습 데이터를 사용하여 앙상블 모델 세트(EMS)를 구성합니다. 그 후 새 데이터 청크를 사용할 수 있게 되면 새 데이터에 대해 새 모델이 작성되어 EMS에 구성요소 모델로 추가됩니다. 새 데이터를 사용하여 EMS에 있는 기존 구성요소 모델의 가중값이 재평가됩니다. 이러한 재평가의 결과로, 더 높은 가중값을 갖는 구성요소 모델이 현재 예측을 위해 선택되고, 더 낮은 가중값을 갖는 구성요소 모델은 EMS에서 삭제될 수 있습니다. 이 프로세스는 모델 가중값과 모

델 인스턴스 모두에 대해 EMS를 새로 고치므로, 시간 경과에 따른 불가피한 데이터 변경사항을 해결할 수 있는 유연하고 효율적인 방식으로 진화합니다.

그림 1. 연속 자동 기계 학습



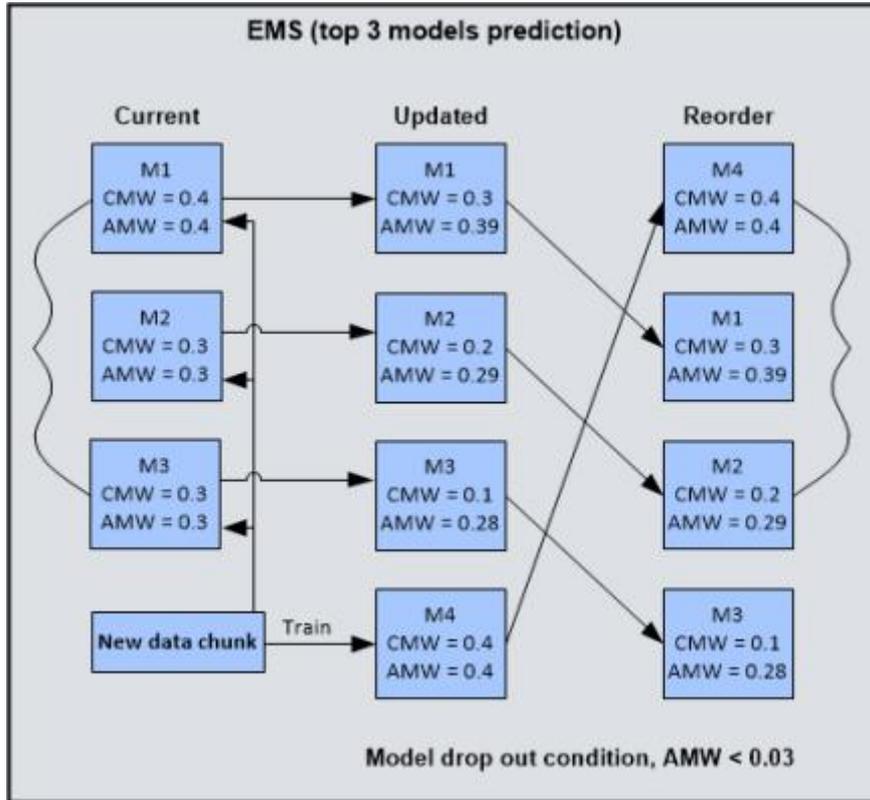
양상블 모델 세트(EMS)는 자동으로 생성되는 모델 너깃이며, 자동 모델링 노드와 생성된 자동 모델 너깃 사이에는 새로 고침 관계를 정의하는 새로 고침 링크가 있습니다. 연속 자동 기계 학습을 활성화하면 새 데이터 자산이 자동 모델링 노드에 연속으로 공급되어 새 구성요소 모델을 생성합니다. 모델 너깃은 대체되는 대신 업데이트됩니다.

다음 그림은 연속 기계 학습 시나리오에서 EMS의 내부 구조 예를 제공합니다. 현재 예측을 위해 상위 3개 구성요소 모델만 선택됩니다. 구성요소 모델(M1, M2, M3로 표시됨)마다 두 종류의 가중값이 유지됩니다. 현재 모델 가중값(CMW)은 새 데이터 청크에 대한 구성요소 모델의 성능을 설명하고, 누적 모델 가중값(AMW)은 최근 데이터 청크에 대한 구성요소 모델의 종합적인 성능을 설명합니다. AMW는 CMW와 이전 값을 통해 반복적으로 계산되며, 이들 사이의 균형을 맞추기 위한 하이퍼 매개변수 베타가 있습니다. AMW 계산 수식을 *지수 이동 평균*이라고 합니다.

새 데이터 청크가 사용 가능해지면 SPSS Modeler는 먼저 이를 사용하여 몇 개의 새 구성요소 모델을 작성합니다. 이 예제 그림에서 모델 4(M4)는 초기 모델 작성 프로세스 중에 계산된 CMW 및 AMW를 사용하여 작성됩니다. 그런 다음 SPSS Modeler는 새 데이터 청크를 사용하여 기존 구성요소 모델(M1, M2, M3)의 측도를 재평가하고 재평가 결과에 따라 CMW 및 AMW를 업데이트합니다. 끝으로, SPSS Modeler는 CMW 또는 AMW를 기준으로 구성요소 모델을 다시 정렬하고 그에 따라 상위 3개 구성요소 모델을 선택할 수 있습니다.

이 그림에서 CMW는 정규화된 값(합계 = 1)을 사용하여 설명되며, AMW는 CMW를 기준으로 계산됩니다. SPSS Modeler에서는 CMW와 AMW를 단순하게 나타내기 위해 절대값(선택된 평가 가중 측도와 동일함 - 예: 정확도)이 선택됩니다.

그림 2. EMS 구조



아래에 표시된 바와 같이 EMS 구성요소 모델마다 두 가지 유형의 가중값이 정의되어 있으며, 둘 다 상위 N개 모델 및 구성요소 모델 드롭아웃을 선택하는 데 사용할 수 있습니다.

- 현재 모델 가중값(CMW)은 새 데이터 청크에 대한 평가를 통해 계산됩니다(예: 새 데이터 청크에 대한 평가 정확도).
- 누적 모델 가중값(AMW)은 CMW와 기존 AMW를 결합하여 계산됩니다(예: 지수 가중 이동 평균(EWMA)).

AMW 계산을 위한 지수 이동 평균 수식:

$$AMW = \beta * AMW + (1 - \beta) * CMW, \quad \text{suggested } \beta > 0.9$$

SPSS Modeler에서 자동 분류자 노드를 실행하여 모델 너깃을 생성한 후 연속 기계 학습에 다음 모델 옵션을 사용할 수 있습니다.

- 모델 새로 고침 중 연속 자동 기계 학습 사용. 연속 기계 학습을 사용으로 설정하려면 이 옵션을 선택하십시오. 참고로, 연속 자동 모델을 훈련시키려면 일관된 메타데이터(데이터 모델)를 사용해야 합니다. 이 옵션을 선택하면 아래의 다른 옵션도 사용으로 설정됩니다.
- 자동 모델 가중값 재평가 사용. 이 옵션은 모델 새로 고침 중 평가 속도(예: 정확도)를 계산하고 업데이트할지 여부를 제어합니다. 이 옵션을 선택하면 EMS 후 (모델 새로 고침 중) 자동 평가 프로세스가 실행됩니다. 이는 일반적으로 새 데이터로 기존 구성요소 모델을 재평가하여 데이터의 현재 상태를 반영해야 하기 때문입니다. 그런 다음 재평가 결과에 따라 EMS 구성요소 모델의 가중값이 지정되고, 구성요소 모델이 최종 앙상블 예측에 기여하는 비율을 결정하는 데 가중값이 사용됩니다. 이 옵션은 기본적으로 선택되어 있습니다.

그림 3. 모델 설정



그림 4. 플래그 대상

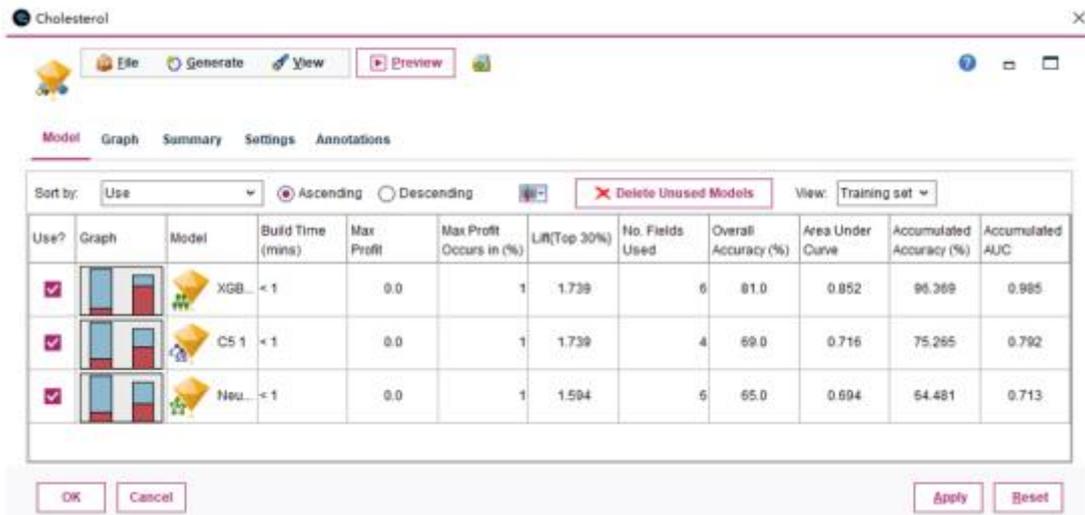


그림 5. 변수군 대상



다음은 자동 분류자 노드에 대해 지원되는 CMW 및 AMW입니다.

표 1. 지원되는 CMW 및 AMW

대상 유형	CMW	AMW
플래그 대상	전체 정확도 곡선 아래 영역(AUC)	누적 정확도 누적 AUC
변수군 대상	전체 정확도	누적 정확도

다음 세 가지 옵션은 최근 데이터 체크 기간 동안 구성요소 모델의 성능을 평가하는 데 사용되는 AMW와 관련이 있습니다.

- **모델 가중값 재평가 중 누적 요인 사용.** 이 옵션을 선택하면 모델 가중값 재평가 중에 AMW 계산이 사용으로 설정됩니다. AMW는 최근 데이터 체크 기간 동안 EMS 구성요소 모델의 종합적인 성능을 나타내고, 위의 AMW 수식에 정의된 누적 요인 β 와 관련이 있으며, 노드 특성에서 조정할 수 있습니다. 이 옵션을 선택하지 않으면 CMW만 계산됩니다. 이 옵션은 기본적으로 선택되어 있습니다.
- **모델 새로 고침 중 누적 한계를 기준으로 모델 축소 수행.** 모델 새로 고침 중 AMW 값이 지정된 한계 미만인 구성요소 모델을 자동 모델 EMS에서 제거하려면 이 옵션을 선택하십시오. 그러면 쓸모 없는 구성요소 모델을 삭제하여 자동 모델 EMS가 너무 무거워지는 것을 방지하는 데 도움이 될 수 있습니다.
누적 한계 값 평가는 **평가 가중 투표**를 앙상블 방법으로 선택할 때 사용되는 가중 측도와 관련됩니다. 아래 사항을 참조하십시오.

그림 6. 대상

The image shows two configuration panels. The top panel, titled 'Flag Target', has a dropdown menu for 'Ensemble method' set to 'Evaluation-weighted voting'. Below it, 'If voting is tied, select value using:' has three radio buttons: 'Random selection' (selected), 'Highest confidence', and 'Raw propensity'. Underneath, 'Select evaluation-weighted measure:' has two radio buttons: 'Model Accuracy' (selected) and 'Area under curve'. The bottom panel, titled 'Set Target', has a dropdown menu for 'Ensemble method' set to 'Confidence-weighted voting'. Below it, 'If voting is tied, select value using:' has two radio buttons: 'Random selection' (selected) and 'Highest confidence'.

평가 가중 측도로 **모델 정확도**를 선택할 경우 누적 정확도가 지정된 한계 미만인 모델이 삭제됩니다. 평가 가중 측도로 **곡선 아래 영역**을 선택할 경우 누적 AUC가 지정된 한계 미만인 모델이 삭제됩니다.

기본적으로 **모델 정확도**는 자동 분류자 노드의 평가 가중 측도로 사용되며, 플래그 대상의 경우 선택적 AUC ROC 측도가 있습니다.

- 누적 평가 가중 투표 사용. 현재 스코어링/예측에 AMW를 사용하려면 이 옵션을 선택하십시오. 그렇지 않으면 CMW가 기본적으로 사용됩니다. 이 옵션은 **평가 가중 투표**를 앙상블 방법으로 선택한 경우에 사용으로 설정됩니다.

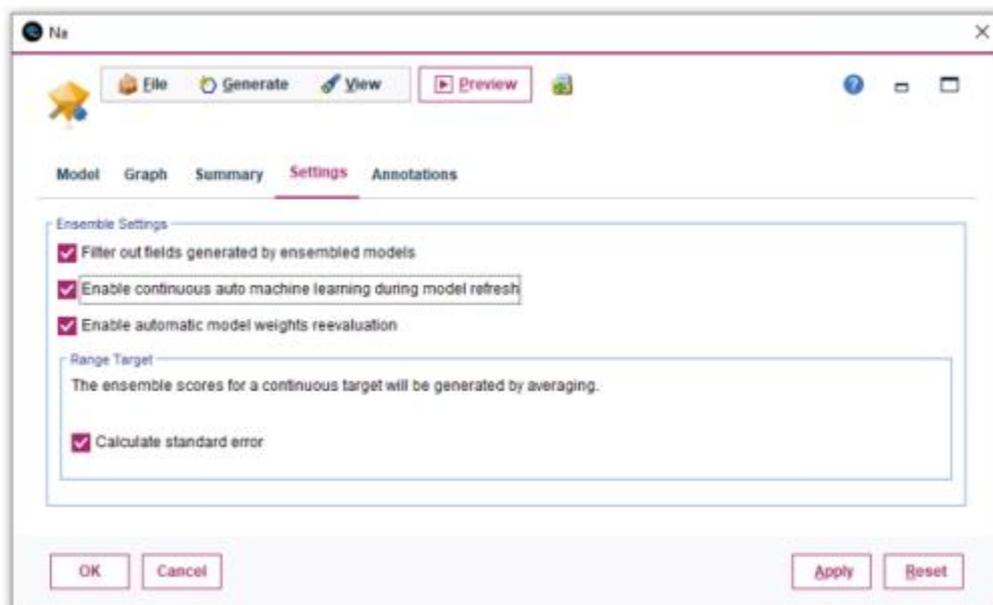
플래그 대상의 경우, 이 옵션을 선택하고 **모델 정확도**를 평가 가중치 측도로 선택하면 **누적 정확도**가 AMW로 사용되어 현재 스코어링을 수행합니다. 또는 **곡선 아래 영역**을 평가 가중치 측도로 선택할 경우 **누적 AUC**가 AMW로 사용되어 현재 스코어링을 수행합니다. 이 옵션을 선택하지 않고 **모델 정확도**를 평가 가중치 측도로 선택하면 **전체 정확도**가 CMW로 사용되어 현재 스코어링을 수행합니다. **곡선 아래 영역**을 선택할 경우 **곡선 아래 영역**이 CMW로 사용되어 현재 스코어링을 수행합니다.

변수군 대상의 경우, **누적 평가 가중 투표 사용** 옵션을 사용하면 **누적 정확도**가 AMW로 사용되어 현재 스코어링을 수행합니다. 그렇지 않은 경우 **전체 정확도**가 CMW로 사용되어 현재 스코어링을 수행합니다.

연속 자동 기계 학습을 사용하면 자동 모델을 다시 작성하여 자동 모델 너깃이 항상 진화하므로, 데이터의 현재 상태를 반영하는 최신 버전을 얻을 수 있습니다. SPSS Modeler는 EMS에 있는 다양한 상위 N개 구성요소 모델을 현재 가중값에 따라 선택할 수 있는 유연성을 제공하므로, 다양한 기간 동안 다양한 데이터와 보조를 맞출 수 있습니다.

① **참고:** 자동 숫자 노드는 훨씬 간단한 경우로서, 자동 분류자 노드에 옵션의 서브세트를 제공합니다.

그림 7. 자동 숫자 노드



예

이 예에서는 통신 산업에 연속 기계 학습을 사용하여 행동을 예측하고 고객을 유지합니다.

다음 플로우에서 데이터 자산에는 지난달에 탈퇴한 고객에 대한 정보(Churn 열)가 포함되어 있습니다. 매월 새로운 데이터가 제공되므로 이 시나리오는 연속 기계 학습에 적합합니다. 이 예에서 1월(Jan) 데이터는 초기 자동 모델을 구성하는 데 사용되고, 2월(Feb) 데이터는 연속 기계 학습을 통해 자동 모델을 개선하는 데 사용됩니다.

그림 8. 통신 예



플로우의 위쪽 분기에서 데이터 자산 노드 뒤에는 중요하지 않은 필드를 필터링하는 필터 노드가 있습니다. 분기의 끝에는 터미널 자동 분류자 모델링 노드가 있습니다. 노드의 전문가 설정에서 훈련 프로세스에 사용할 알고리즘을 선택합니다. 이 예에서는 로지스틱 회귀분석, 베이지안 네트워크 및 신경망의 세 가지 알고리즘을 선택합니다. 그런 다음 플로우를 실행하여 자동 모델 너깃을 생성합니다.

그림 9. 평가 측도

The screenshot shows the 'churn' model evaluation window. It features a table with columns for 'Us...', 'Graph', 'Model', 'Build Time (mins)', 'Max Profit', 'Max Profit Occurs in', 'Lift (Top 3...', 'No. Fields Used', 'Overall Accuracy', 'Area Under Curve', 'Accumulated Accuracy (%)', and 'Accumulated AUC'. The table contains three rows of model results, each with a checked checkbox in the 'Us...' column.

Us...	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in	Lift (Top 3...	No. Fields Used	Overall Accuracy	Area Under Curve	Accumulated Accuracy (%)	Accumulated AUC
<input checked="" type="checkbox"/>		Ba...	<1	205.0	13	2.222	10	80.6	0.852	80.6	0.852
<input checked="" type="checkbox"/>		Lo...	<1	145.0	13	2.171	10	79.0	0.794	79.0	0.794
<input checked="" type="checkbox"/>		N...	<1	145.0	11	2.041	10	79.2	0.791	79.2	0.791

이제 자동 모델 너깃 내부에 무엇이 있는지 살펴보겠습니다. 선택한 세 가지 알고리즘에 대한 세 가지 구성요소 모델이 포함되어 있음을 알 수 있습니다. 구성요소 모델마다 몇 가지 평가 측도가 생성됩니다(예: 정확도 및 곡선 아래 영역). 이러한 평가 측도는 학습 데이터(1월 데이터 세트)에 대한 구성요소 모델의 성능을 설명합니다. 현재 앙상블 예측에 사용할 구성요소 모델을 선택할 수 있습니다.

플로우의 위쪽 분기에서 데이터 자산 노드 뒤에는 중요하지 않은 필드를 필터링하는 필터 노드가 있습니다. 분기의 끝에는 터미널 자동 분류자 모델링 노드가 있습니다. 노드의 전문가 설정에서 훈련 프로세스에 사용할 알고리즘을 선택합니다. 이 예에서는 로지스틱 회귀분석, 베이지안 네트워크 및 신경망의 세 가지 알고리즘을 선택합니다. 그런 다음 플로우를 실행하여 자동 모델 너깃을 생성합니다.

누적 평가 측도도 볼 수 있습니다. 이러한 누적 측도는 연속 기계 학습용으로, 최근 데이터 변경 사항에 대한 구성요소 모델의 성능을 설명하므로 일정 기간 동안 모델의 종합적인 성능을 파악할 수 있습니다. 이것은 초기 자동 모델이므로 누적 측도에 대한 초기 값이 관련 현재 측도와 동일합니다. 기본적으로 평가 측도는 학습 데이터에 대해 계산되므로 어느 정도의 과적합이 있을 수 있습니다. 이를 방지하기 위해 자동 분류자 노드는 교차 검증을 통해 보다 안정적인 평가 측도를 계산하는 작성 옵션을 제공합니다.

다음으로, 최종 앙상블 예측이 어떻게 생성되는지 살펴보겠습니다. 자동 모델의 특성을 열면 **앙상블 플래그 대상** 아래에 있는 훈련 대상 이탈 필드는 예/아니오 플래그 대상입니다. **앙상블 변수군 대상**(값이 세 개 이상 포함된 변수군 대상 필드의 경우) 아래에는 **앙상블 방법** 드롭 다운이 있습니다. 드롭 다운에서 여러 옵션을 사용할 수 있습니다. 예를 들어 **다수결 투표**는 각 구성요소 모델이 한 개의 투표 티켓을 보유하고 있음을 의미하며, **신뢰 가중 투표**는 각 구성요소 모델의 예측에 대한 신뢰도 필드가 투표 가중값으로 사용됨을 의미합니다. 신뢰도가 높을수록 최종 앙상블 예측에 더 많은 영향을 미칩니다. 마찬가지로, 연속 기계 학습에 대한 더 나은 지원을 제공하기 위해 **평가 가중 투표**를 사용할 수 있습니다. 그러면 구성요소 모델의 평가 측도(예: 모델 정확도 또는 곡선 아래 영역)가 투표 가중값으로 사용됩니다. 플래그 대상의 경우 **평가 가중 투표**를 사용할 때 특정 평가 측도를 투표 가중값으로 선택하는 옵션도 있습니다. 변수군 대상의 경우 **정확도**만 현재 지원됩니다.

그림 10. 변수군 및 플래그 대상

The image shows two configuration panels for ensemble methods. The top panel, titled 'Flag Target', has 'Ensemble method' set to 'Evaluation-weighted voting'. Under 'If voting is tied, select value using:', 'Random selection' is selected. Under 'Select evaluation-weighted measure:', 'Model Accuracy' is selected. The bottom panel, titled 'Set Target', has 'Ensemble method' set to 'Confidence-weighted voting'. Under 'If voting is tied, select value using:', 'Random selection' is selected.

양상블 일반 설정 아래에서 연속 기계 학습을 켤 수 있습니다. 그런 다음 2월 데이터를 사용하여 어떤 상황이 발생하는지 확인할 수 있습니다. 기존 구성요소 모델 알고리즘을 구별하기 위해 두 가지 다른 알고리즘을 선택할 수 있습니다. 그런 다음 플로우를 다시 작성하고 자동 모델의 콘텐츠를 확인하면 두 개의 새 구성요소 모델(C5 및 CRT)이 추가된 것을 알 수 있습니다. 또한 기존 구성요소 모델에 대한 평가 측도도 재계산되었습니다. CMW 측도와 AMW 측도가 모두 이전과 다릅니다. 이제 두 측도를 원래 자동 모델의 해당 측도와 비교할 수 있습니다.

그림 11. 평가 측도

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift (Top 30%)	No. Fields Used	Overall Accuracy (%)	Area Under Curve	Accumulated Accuracy (%)	Accumulated AUC
<input checked="" type="checkbox"/>		Bayesian N.	< 1	0.0	1	2.115	10	78.4	0.85	78.4	0.85
<input checked="" type="checkbox"/>		Logistic regr.	< 1	0.0	1	2.0	10	75.6	0.805	75.6	0.805
<input checked="" type="checkbox"/>		Neural Net 1	< 1	0.0	1	1.908	10	74.8	0.801	74.8	0.801
<input checked="" type="checkbox"/>		Neural Net 1	< 1	0.0	1	1.892	10	73.4	0.769	78.62	0.789
<input checked="" type="checkbox"/>		Logistic regr.	< 1	0.0	1	1.815	10	73.4	0.758	78.44	0.791
<input checked="" type="checkbox"/>		Bayesian N.	< 1	0.0	1	1.77	10	73.4	0.718	79.88	0.838

결과 향상된 자동 모델을 사용할 경우 우선순위가 지정된 평가 측도를 선택하고 해당 측도를 기준으로 정렬된 상위 N개 구성요소 모델을 얻을 수 있습니다. 그런 다음 상위 N개 구성요소 모델을 사용하여 수신되는 예측 분석 요청에 대한 최종 양상블 예측에 참여할 수 있습니다. 평가 가중 투표를 양상블 방법으로 선택한 경우 양상블 일반 설정에서 누적 평가 가중 투표 사용 옵션을 선택하기만 하면 누적 측도를 투표 가중값으로 사용할 수 있습니다. 선택 취소할 경우 CMW 측도가 평가 가중 투표에 기본적으로 사용됩니다.

연속 기계 학습을 통해 자동 모델은 새 데이터 청크에 대해 지속적으로 다시 작성되면서 항상 진화하므로, 모델이 데이터의 현재 상태를 반영하는 최신 버전이 될 수 있습니다. 따라서 EMS에 있는 다양한 상위 N개 구성요소 모델을 현재 또는 누적 평가 측도에 따라 유연하게 선택할 수 있으므로, 다양한 기간 동안 다양한 데이터와 보조를 맞출 수 있습니다.

4) 의사결정 트리

(1) 의사결정 트리 모형

의사결정 트리 모형을 사용하여 의사결정 규칙 세트를 기준으로 하여 향후 관측값을 예측 또는 분류하는 분류 시스템을 개발하십시오. 데이터를 관심 있는 클래스로 나눈 경우(예를 들어, 고위험 대 저위험 대출, 가입자 대 비가입자, 유권자 대 비유권자 또는 박테리아 유형) 데이터를 사용하여 오래된 케이스나 새 케이스를 최대 정확도로 분류하는 데 사용할 수 있는 규칙을 작성할 수 있습니다. 예를 들어, 나이 및 기타 요인을 기준으로 하여 신용 거래 위험 또는 구매 의향을 분류하는 트리를 작성할 수 있습니다.

때로 *규칙 귀납*이라 부르는 이 접근법은 여러 가지 장점이 있습니다. 첫째, 트리를 찾아볼 때 모델 배후의 추론 프로세스가 명확합니다. 이는 내부 로직을 이해하기 어려운 기타 *블랙박스* 모델링 기법과 대조됩니다.

두 번째로, 프로세스는 실제로 의사결정에서 중요한 속성만을 자동으로 규칙에 포함시킵니다. 트리 정확도에 기여하지 않는 속성은 무시합니다. 이러한 방식은 데이터에 대한 매우 유용한 정보를 산출하며 신경망과 같은 다른 학습 기법을 학습시키기 전에 관련 필드로 데이터를 축소하는 데 사용할 수 있습니다.

의사결정 트리 모형 너그은 if-then 규칙 컬렉션(*규칙 세트*)으로 변환할 수 있으며 많은 경우, 보다 이해하기 쉬운 형태로 정보를 표시합니다. 의사결정 트리 프리젠테이션은 데이터의 속성이 문제에 관련된 서브세트로 모집단을 *분할* 또는 *파티셔닝*하는 방식을 확인하려는 경우에 유용합니다. 트리-AS 노드 출력은 규칙 세트를 작성할 필요 없이 너그에 직접 규칙 목록을 포함시킴으로 기타 의사결정 트리 노드와 차이가 있습니다. 규칙 세트 프리젠테이션은 특정 항목 그룹이 특정 결론에 관련되는 방식을 보려는 경우에 유용합니다. 예를 들어, 다음 규칙은 구매할 가치가 있는 자동차 그룹에 대한 *프로파일*을 제공합니다.

```
IF tested = 'yes'  
AND mileage = 'low'  
THEN -> 'BUY'.
```

트리 작성 알고리즘

분류 및 세분화 분석을 수행하는 데 여러 알고리즘을 사용할 수 있습니다. 이 알고리즘은 모두 기본적으로 동일한 사항을 수행합니다. 데이터 세트의 모든 필드를 검사하고 데이터를 하위 그룹으로 분할해서 최상의 분류 또는 예측을 제공하는 필드를 찾습니다. 트리가 완료될 때까지(특정 중지 기준에 정의된 대로) 하위 그룹을 더 작은 단위로 분할하면서 프로세스가 반복해서 적용됩니다. 트리 작성에 사용된 목표 및 입력 필드는 사용한 알고리즘에 따라 연속형(수치 범위) 또는 범주형이 가능합니다. 연속형 목표를 사용하는 경우 회귀분석 트리가 생성되고 범주형 목표를 사용하면 분류 트리가 생성됩니다.



분류 및 회귀(C&R) 트리 노드는 추가 관측값을 예측하거나 분류할 수 있게 하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 "순수"로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).



CHAID 노드는 최적 분할을 식별하기 위해 카이제곱 통계량을 사용하여 의사결정 트리를 생성합니다. C&R 트리 및 QUEST 노드와 달리, CHAID는 비이분형 트리를 생성할 수 있으며 이는 일부 분할에 둘 이상의 분기가 있음을 의미합니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



QUEST 노드는 의사결정 트리를 작성하기 위한 이분형 분류 방법을 제공하며, 대형 C&R 트리 분석에 필요한 처리 시간을 줄이는 동시에 분류 트리 방법에서 찾은 경향을 줄여 더 많은 분할을 허용하는 입력을 선호하도록 설계되었습니다. 입력 필드는 숫자 범위(연속)일 수 있지만 대상 필드는 범주형이어야 합니다. 모든 분할은 이분형입니다.



C5.0 노드는 의사결정 트리 또는 규칙 세트를 작성합니다. 모델은 각 수준에서 최대 정보 이익을 제공하는 필드를 기반으로 샘플을 분할하여 작동합니다. 대상 필드는 범주형이어야 합니다. 세 개 이상의 부집단으로의 다중 분할이 허용됩니다.



Tree-AS 노드는 기존 CHAID 노드와 유사하지만, 빅 데이터를 처리하여 단일 트리를 작성하도록 설계되었으며 SPSS® Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시합니다. 이 노드는 최적 분할을 식별하기 위해 카이제곱 통계량(CHAID)을 사용하여 의사결정 트리를 생성합니다. 이 CHAID의 사용은 일부 분할이 셋 이상의 분기를 가짐을 의미하는 비2진 트리를 생성할 수 있습니다. 목표 및 입력 필드는 숫자 범위(연속형) 또는 범주형입니다. Exhaustive CHAID는 가능한 모든 분할을 탐색하는 보다 철저한 작업을 수행하지만 계산하는 데 시간이 더 걸리는 변형 CHAID입니다.



이 랜덤 트리 노드는 기존 C&RT 노드와 유사하지만, 빅 데이터를 처리하여 단일 트리를 작성하도록 설계되었으며 SPSS Modeler 버전 17에 추가된 출력 뷰어에 결과 모델을 표시합니다. 랜덤 트리 노드는 추가 관측값을 예측하거나 분류하는 데 사용하는 의사결정 트리를 생성합니다. 이 방법은 재귀적 파티셔닝을 사용하여 각 단계마다 불순도를 최소화하여 학습 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 대상 필드의 특정 범주에 속하면 순수로 간주됩니다. 목표 및 입력 필드는 숫자 범위 또는 범주형(명목형, 순서형 또는 플래그)입니다. 모든 분할은 이분형입니다(오직 두 개의 부집단).

트리 기반 분석의 일반 용도

다음은 트리 기반 분석의 몇 가지 일반 용도입니다.

세분화: 특정 클래스의 멤버일 수 있는 개인을 식별합니다.

층화: 고, 중, 저위험 그룹과 같은 여러 범주 중 하나로 케이스를 지정합니다.

예측: 규칙을 작성하고 사용하여 미래 이벤트를 예측합니다. 예측은 연속형 변수의 값에 예측 속성을 관련시키려는 시도를 의미할 수도 있습니다.

데이터 축소 및 변수 선별: 정규 모수 모델을 작성하는 데 사용할 큰 변수 세트에서 유용한 예측변수 서브세트를 선택하십시오.

상호작용 식별: 특정 하위 그룹에만 관련된 관계를 식별하고 정규 모수 모델에 이를 지정합니다.

범주 병합 및 연속형 변수 배당: 최소의 정보 손실로 그룹 예측변수 범주 및 연속형 변수를 다시 코딩합니다.

(2) 대화형 트리 작성기

트리 모델을 자동으로 생성하거나(이 경우 알고리즘은 각 수준에서 최상의 분할을 결정함), 대화형 트리 작성기를 사용하여 제어할 수 있습니다(이 경우 모델 너깅을 저장하기 전에 트리를 세분화 또는 단순화할 비즈니스 지식을 적용함).

1. 스트림을 작성하고 의사결정 트리 노드 C&R 트리, CHAID 또는 QUEST 중 하나를 추가하십시오.

 **참고:** 대화형 트리 작성은 Tree-AS 또는 C5.0 트리에서 지원되지 않습니다.

2. 노드를 열고 필드 탭에서 목표 및 예측자 필드를 선택하고 필요한 경우 추가 모델 옵션을 지정하십시오. 특정 지시사항의 경우 각 트리 작성 노드에 대한 문서를 참조하십시오.
3. 작성 옵션 탭의 목표 패널에서 **대화형 세션 시작**을 선택하십시오.
4. **실행**을 클릭하여 트리 작성기를 시작하십시오.

루트 노드부터 시작하여 현재 트리가 표시됩니다. 수준별로 트리를 편집 및 가지치기하고 하나 이상의 모델을 생성하기 전에 이익, 위험, 관련 정보에 액세스할 수 있습니다.

설명

- C&R 트리, CHAID, QUEST 노드에서 모델에 사용된 순서 필드는 숫자 저장 공간(문자열이 아님)을 포함해야 합니다. 필요한 경우 재분류 노드를 사용하여 변환할 수 있습니다.
- 선택적으로 파티션 필드를 사용하여 데이터를 훈련 및 검정 표본으로 구분할 수 있습니다.
- 트리 작성기를 사용하는 대신, 다른 IBM® SPSS® Modeler 모델과 같이 모델링 노드에서 모델을 직접 생성할 수 있습니다. 자세한 정보는 직접 트리 모델 작성 주제를 참조하십시오.

① 트리 성장 및 가지치기

트리 작성기를 시작하려면 C&R 트리, QUEST 또는 CHAID 노드를 포함하는 스트림을 실행하십시오. 이때, 작성 옵션 탭의 목적 패널에서 **대화형 세션 시작** 옵션을 선택해야 합니다.

트리 작성기의 뷰어 탭에서는 루트 노드부터 시작하여 현재 트리를 볼 수 있습니다.

1. 트리를 성장시키려면 메뉴에서 다음을 선택하십시오.

트리 > 트리 성장

시스템은 하나 이상의 중지 기준을 만족할 때까지 각 분기를 반복적으로 분할하여 트리를 작성합니다. 각 분할에서 사용된 모델링 방법에 따라 최상의 예측변수가 자동으로 선택됩니다.

2. 또는 **트리 한 수준 성장**을 선택하여 단일 수준을 추가하십시오.
3. 특정 노드 아래 분기를 추가하려면 노드를 선택하고 **분기 성장**을 선택하십시오.
4. 분기에 사용된 예측변수를 선택하려면 원하는 노드를 선택하고 **사용자 정의 분할로 분기 성장**을 선택하십시오. 자세한 정보는 사용자 정의 분할 정의의 내용을 참조하십시오.
5. 분기를 가지치기하려면 노드를 선택하고 **분기 제거**를 선택하여 선택한 노드를 지우십시오.
6. 트리에서 아래쪽 수준을 제거하려면 **한 수준 제거**를 선택하십시오.
7. C&R 트리 및 QUEST 트리의 경우에만 **트리 성장 및 가지치기**를 선택하여 터미널 노드 수에 기반하여 위험 추정값을 조정하는 비용-복잡도 알고리즘에 따라 가지치기를 수행하십시오. 그러면 일반적으로 더 단순한 트리가 생성됩니다. 자세한 정보는 C&R 트리 노드의 내용을 참조하십시오.

뷰어 탭에서 분할 규칙 읽기

뷰어 탭에서 분할 규칙을 보는 경우 꺾쇠 괄호는 인접한 값이 범위에 포함됨을 의미하지만, 소괄호는 인접한 값이 범위에서 제외됨을 의미합니다. 따라서 표현식 (23,37]은 23(제외)에서 37(포함) 사이의 범위(즉, 24부터 37까지)를 의미합니다. 모델 탭에서도 다음과 같이 동일한 조건이 표시됩니다.

Age > 23 and Age <= 37

트리 성장 중단. 트리 성장 작업을 중단하려면(예를 들어, 예상보다 오래 걸리는 경우) 도구 모음에서 실행 중지 단추를 클릭하십시오.

그림 1. 실행 중지 단추



단추는 트리 성장 중에만 사용 가능합니다. 현재 포인트에서 현재 성장 작업을 중지하며, 변경사항을 저장하거나 창을 닫지 않은 상태로 이미 추가된 노드는 남겨둡니다. 트리 작성기는 열려 있으며, 여기에서 모델을 생성하거나 지시문을 업데이트하거나 필요한 경우 적절한 형식으로 출력을 내보낼 수 있습니다.

② 사용자 정의 분할 정의

분할 정의 대화 상자에서는 예측변수를 선택하고 각 분할에 대한 조건을 지정할 수 있습니다.

1. 트리 작성기의 뷰어 탭에서 노드를 선택하고 메뉴에서 다음을 선택하십시오.

트리 > 사용자 정의 분할로 분기 성장

2. 드롭 다운 목록에서 원하는 예측변수를 선택하거나 **예측변수** 단추를 클릭하여 각 예측변수의 세부사항을 보십시오. 자세한 정보는 예측자 세부사항 보기의 내용을 참조하십시오.
3. 각 분할에 대한 기본 조건을 수락하거나 **사용자 정의를** 선택하여 분할에 대한 조건을 적절히 지정할 수 있습니다.

- 연속형(숫자 범위) 예측변수의 경우 **범위 값 편집 필드**를 사용하여 각 새 노드에 포함되는 값의 범위를 지정할 수 있습니다.
- 범주형 예측변수의 경우 **세트 값 편집** 또는 **순서 값 편집 필드**를 사용하여 각 새 노드에 맵핑되는 특정 값(또는 순서 예측변수의 경우 값의 범위)을 지정할 수 있습니다.

4. **성장**을 선택하여 선택한 예측변수를 통해 분기를 재성장시키십시오.

일반적으로 트리는 중지 규칙에 상관없이 예측변수를 사용하여 분할할 수 있습니다. 유일한 예외는 노드가 순수하거나(즉, 케이스 전부가 동일한 목표 클래스에 포함되어 분할할 항목이 없음) 선택한 예측변수가 일관되는 경우(분할할 목표가 없음)입니다.

결측값 입력. CHAID 트리인 경우에만, 지정된 예측변수에서 결측값이 사용 가능하면 특정 하위 노드에 이를 지정하도록 사용자 정의 분할을 정의할 때 옵션이 제공됩니다. (C&R 트리 및 QUEST의 경우 결측값은 알고리즘에 정의된 대용을 사용하여 처리됩니다. 자세한 정보는 분할 세부사항 및 대용의 내용을 참조하십시오.)

가. 예측자 세부사항 보기

예측자 선택 대화 상자에서는 현재 분할에 사용할 수 있는 사용 가능한 예측자(또는 때때로 "경쟁자"라고도 함)의 통계를 표시합니다.

- CHAID 및 exhaustive CHAID의 카이제곱 통계량은 각 범주형 예측자에서 나열됩니다. 예측자가 숫자 범위인 경우 F 통계량이 표시됩니다. 카이제곱 통계량은 분할 필드에서 목표 필드가 얼마나 독립되어 있는지 정도의 척도입니다. 높은 카이제곱 통계량은 일반적으로 더 낮은 확률과 연관됩니다. 즉, 두 개 필드가 서로 독립될 가능성이 낮으며, 분할이 바람직한 분할임을 표시합니다. 또한 자유도도 포함됩니다. 이 방법이 삼원 분할의 경우 이원 분할보다 큰 통계와 작은 확률을 보유하기 쉽다는 사실을 고려하기 때문입니다.
- C&R 트리 및 QUEST의 경우 각 예측자의 개선도가 표시됩니다. 개선도가 클수록 예측자가 사용된 경우 상위와 하위 노드 사이의 불순도가 더 많이 감소합니다. (순수한 노드는 모든 케이스가 단일 목표 범주에 속하는 노드입니다. 트리에서 불순도가 낮을수록 모형이 데이터에 더 적합합니다.) 즉, 일반적으로 개선도가 높은 그림은 이 트리 유형에서 유용한 분할을 표시합니다. 사용되는 불순도 척도는 트리 작성 노드에서 지정됩니다.

③ 분할 세부사항 및 대응

뷰어 탭에서 노드를 선택하고 도구 모음 오른쪽에 있는 분할 정보 단추를 선택하여 해당 노드의 분할에 대한 세부사항을 볼 수 있습니다. 관련 통계와 함께 사용되는 분할 규칙이 표시됩니다. C&R 트리 범주형 트리의 경우 개선도와 연관도 표시됩니다. 연관은 대응 및 1차 분할 필드 사이의 대응에 대한 척도이며, 일반적으로 분할 필드와 가장 비슷한 항목이 "최상"의 대응입니다. C&R 트리 및 QUEST의 경우 1차 예측자 대신 사용되는 대응도 함께 나열됩니다.

선택한 노드의 분할을 편집하려면 대응 패널 왼쪽에 있는 아이콘을 클릭하여 분할 정의 대화 상자를 열면 됩니다. (단축 아이콘으로, 아이콘을 클릭하여 1차 분할 필드로 선택하기 전에 목록에서 대응을 선택할 수 있습니다.)

대응. 적용 가능한 경우 선택한 노드에 대한 기본 분할 필드의 대응이 표시됩니다. 대응은 주어진 레코드의 기본 예측자 값이 결측된 경우에 사용되는 대체 필드입니다. 주어진 분할의 허용된 최대 대응 수는 트리 작성 노드에 지정되지만 실제 수는 훈련 데이터에 따라 다릅니다. 일반적으로 결측 데이터가 많을수록 더 많은 대응이 사용될 수 있습니다. 기타 의사결정 트리 모형의 경우에는 이 탭이 비어 있습니다.

참고: 모델에 포함하려면 훈련 단계 중에 대응을 식별해야 합니다. 훈련 표본에 결측값이 없으면 대응이 식별되지 않으며, 검정 또는 스코어링 중에 발견된 결측값이 있는 레코드는 자동으로 레코드 수가 가장 많은 하위 노드로 들어갑니다. 검정 또는 스코어링 중에 결측값이 예상되는 경우 반드시 훈련 표본에서도 값이 결측되었는지 확인하십시오. CHAID 트리에는 대응을 사용할 수 없습니다.

대응은 CHAID 트리에서 사용되지 않지만, 사용자 정의 분할을 정의할 때 특정 하위 노드에 이를 지정하는 옵션이 제공됩니다. 자세한 정보는 사용자 정의 분할 정의의 내용을 참조하십시오.

④ 트리 보기 사용자 정의

트리 작성기의 뷰어 탭에서는 현재 트리를 표시합니다. 기본적으로 트리의 모든 분기는 펼쳐져 있지만, 필요한 경우 분기를 펼치거나 접고, 다른 설정을 사용자 정의할 수도 있습니다.

- 상위 노드의 맨 아래 오른쪽에 있는 빼기 부호(-)를 클릭하여 해당 하위 노드를 모두 숨기십시오. 상위 노드의 맨 아래 오른쪽에 있는 더하기 부호(+)를 클릭하여 해당 하위 노드를 표시하십시오.
- 보기 메뉴 또는 도구 모음을 사용하여 트리 방향을 변경하십시오(위에서 아래로, 왼쪽에서 오른쪽 또는 오른쪽에서 왼쪽).
- 주 도구 모음에서 "필드 및 값 레이블 표시" 단추를 클릭하여 필드 및 값 레이블을 표시하거나 숨기십시오.
- 돋보기 단추를 사용하여 보기를 축소/확대하거나 도구 모음 오른쪽에 있는 트리 맵 단추를 사용하여 전체 트리의 다이어그램을 보십시오.
- 파티션 필드가 사용 중이면 훈련 및 검정 분할(보기 > 파티션) 사이에서 트리 보기를 전환할 수 있습니다. 검정 표본이 표시되면 트리는 볼 수 있어도 편집은 불가능합니다. (현재 파티션은 창의 오른쪽 하단 코너에 있는 상태 표시줄에 표시됩니다.)
- 분할 정보 단추(도구 모음에서 맨 오른쪽에 있는 "i" 단추)를 클릭하여 현재 분할에 대한 세부 사항을 보십시오. 자세한 정보는 분할 세부사항 및 대용 주제를 참조하십시오.
- 각 노드 내 통계, 그래프 또는 둘 다를 표시하십시오(아래 참조).

통계 및 그래프 표시

노드 통계. 범주형 목표 필드의 경우 각 노드의 테이블은 각 범주의 레코드 수 및 퍼센트와 노드가 나타내는 전체 샘플의 퍼센트를 표시합니다. 연속형 목표 필드(숫자 범위)의 경우 테이블은 평균, 표준 편차, 레코드 수, 목표 필드의 예측값을 표시합니다.

노드 그래프. 범주형 목표 필드의 경우 그래프는 목표 필드의 각 범주에서 퍼센트를 나타내는 막대형 차트입니다. 테이블에서 각 행 앞에는 노드의 그래프에서 각 목표 필드 범주를 나타내는 색에 대응하는 색상 견본이 나옵니다. 연속형 목표 필드(숫자 범위)의 경우 그래프는 노드에 있는 레코드에 대한 목표 필드의 히스토그램을 표시합니다.

⑤ 이득

이익 탭에서는 트리의 모든 터미널 노드에 대한 통계를 표시합니다. 이익은 지정된 노드에서 평균 또는 비율이 전체 평균과 얼마나 다른지의 측도를 제공합니다. 일반적으로 이 차이가 클수록 의사결정을 내리는 도구로서 트리의 유용성이 높아집니다. 예를 들어, 노드에서 지수 또는 "리프트" 값이 148%인 경우 노드의 레코드가 전반적으로 데이터 세트에 비해 목표 범주에 포함될 가능성이 1.5배 정도임을 의미합니다.

과적합 방지 세트가 지정된 C&R 트리 및 QUEST 노드의 경우 다음과 같이 통계의 두 개 세트가 표시됩니다.

- 트리 성장 세트 - 과적합 방지 세트가 제거된 훈련 표본
- 과적합 방지 세트

기타 C&R 트리 및 QUEST 대화형 트리와 모든 CHAID 대화형 트리의 경우 트리 성장 세트 통계만 표시됩니다.

이익 탭에서는 다음을 수행할 수 있습니다.

- 노드별, 누적 또는 분위수 통계를 표시합니다.
- 이익 또는 수익을 표시합니다.
- 테이블 및 차트 간 보기를 전환합니다.
- 목표 범주(범주형 목표만 해당)를 선택합니다.
- 지수 퍼센트에 따라 오름차순 또는 내림차순으로 테이블을 정렬합니다. 다중 파티션에 대한 통계가 표시되는 경우 항상 검정 표본이 아닌 훈련 표본에 정렬이 적용됩니다.

일반적으로 이익 테이블에서 선택한 내용은 트리 보기에서 업데이트되며, 반대의 상황도 마찬가지입니다. 예를 들어, 테이블에서 행을 선택하면 대응하는 노드가 트리에서 선택됩니다.

가. 분류 이익

분류 트리(범주형 목표 변수가 있는 트리)의 경우 이익 지수 퍼센트는 각 노드에서 주어진 목표 범주의 비율이 전반적인 비율과 얼마나 다른지를 알려줍니다.

노드별 통계

이 보기에서 테이블은 터미널 노드마다 한 개 행을 표시합니다. 예를 들어, DM 캠페인에 대한 전반적인 반응이 10%지만, 노드 X에 속하는 레코드 중 20%가 긍정적으로 반응한 경우 노드의 지수 퍼센트는 200%이며, 이는 이 그룹의 반응자가 전반적인 인구에 비해 구매할 확률이 2 배임을 의미합니다.

과적합 방지 세트가 지정된 C&R 트리 및 QUEST 노드의 경우 다음과 같이 통계의 두 개 세트가 표시됩니다.

- 트리 성장 세트 - 과적합 방지 세트가 제거된 훈련 표본
- 과적합 방지 세트

기타 C&R 트리 및 QUEST 대화형 트리와 모든 CHAID 대화형 트리의 경우 트리 성장 세트 통계만 표시됩니다.

노드. 현재 노드의 ID(뷰어 탭에 표시됨).

노드: n. 해당 노드에 있는 총 레코드 수.

노드(%). 이 노드에 속하는 데이터 세트의 모든 레코드 퍼센트.

이익: n. 이 노드에 포함되는 선택된 목표 범주를 포함하는 레코드 수. 즉, 목표 범주에 포함되는 데이터 세트의 모든 레코드 중에서 이 노드에는 몇 개나 있습니까?

이익(%). 전체 데이터 세트 중 이 노드에 속하며 목표 범주에 있는 모든 레코드의 퍼센트.

반응(%). 목표 범주에 포함되는 현재 노드에 있는 레코드의 퍼센트. 이 컨텍스트에서 반응은 때때로 "적중"이라고도 합니다.

지수(%). 전체 데이터 세트에 대한 반응 퍼센트의 비율로 표현되는 현재 노드의 반응 퍼센트. 예를 들어, 지수 값이 300%인 경우 이 노드의 레코드가 전반적으로 데이터 세트에 비해 목표 범주에 포함될 가능성이 3배 정도임을 의미합니다.

누적 통계

누적 보기에서 테이블은 해당 하나의 노드를 표시하지만, 통계는 누적으로, 지수 퍼센트의 오름차순 또는 내림차순으로 정렬됩니다. 예를 들어, 내림차순 정렬이 적용된 경우 지수 퍼센트가 가장 높은 노드가 처음 나열되고, 다음에 나오는 행의 통계는 해당 행 이상에서 누적됩니다.

누적 지수 퍼센트는 반응 퍼센트가 더 낮은 노드가 추가될 때 행 단위로 감소합니다. 마지막 행의 누적 지수는 항상 100%입니다. 이 포인트에서 전체 데이터 세트가 포함되기 때문입니다.

사분위수

이 보기에서 테이블의 각 행은 노드보다 분위수를 나타냅니다. 분위수는 사분위수, 5분위수(1/5), 십분위수(1/10), 20분위수(1/20) 또는 백분위수(1/100)입니다. 해당 퍼센트를 구성하는데 둘 이상의 노드가 필요한 경우 단일 분위수에 다중 노드를 나열할 수 있습니다(예: 사분위수가 표시되지만 상위 2개 노드가 모든 케이스의 50% 미만을 포함하는 경우). 나머지 테이블은 누적이며, 누적 보기와 동일한 방식으로 해석할 수 있습니다.

나. 분류 이익 및 ROI

분류 트리의 경우 이익 통계는 이익 및 투자수익률(ROI)의 관점에서 표시할 수도 있습니다. 이익 정의 대화 상자에서는 각 범주의 수입 및 비용을 지정할 수 있습니다.

1. 이익 탭에서 도구 모음의 이익 단추(레이블이 \$/\$임)를 클릭하여 대화 상자에 액세스하십시오.
2. 목표 필드의 각 범주에 대한 수입 및 비용 값을 입력하십시오.

예를 들어, 각 고객에게 제안을 메일로 보내는 데 \$0.48의 비용이 들고, 긍정적인 반응으로부터 얻는 수입이 3개월 구독의 경우 \$9.95인 경우 각 no 반응은 \$0.48의 비용이 들고 각 yes는 \$9.47의 수입을 가져다 줍니다(9.95-0.48로 계산).

이익 테이블에서 **이익**은 터미널 노드에 있는 각 레코드에 대해 수입 합계에서 지출을 뺀 값으로 계산됩니다. ROI는 노드에서 총 이익을 총 지출로 나눈 값입니다.

설명

- 이익 값은 핵심에 더 근접한 관점에서 통계를 조회하는 방법으로 이익 테이블에 표시되는 평균 이익 및 ROI 값에만 영향을 줍니다. 기본 트리 모델 구조에는 영향을 주지 않습니다. 이익은 오분류 비용(트리 작성 노드에서 지정되며, 비용상의 실수를 막기 위해 모델로 포함됨)과 혼동해서는 안 됩니다.
- 이익 지정은 한 대화형 트리 작성 세션과 다음 세션 사이에서 지속되지 않습니다.

다. 회귀분석 이익

회귀 트리의 경우 노드별, 누적 노드별, 분위수 보기 사이에서 선택할 수 있습니다. 테이블에는 평균값이 표시됩니다. 차트는 수량에서만 사용할 수 있습니다.

라. Gains 차트

차트는 테이블의 대체 항목으로 이익 탭에 표시할 수 있습니다.

1. 이익 탭에서 사분위수 아이콘(도구 모음의 왼쪽에서 세 번째)을 선택하십시오. (차트는 노드별 또는 누적 통계에서 사용할 수 없습니다.)
2. 차트 아이콘을 선택하십시오.
3. 원하는 경우 드롭 다운 목록에서 표시된 단위(백분위수, 십분위수 등)를 선택하십시오.
4. **이익**, **반응** 또는 **리프트**를 선택하여 표시되는 축도를 변경하십시오.

Gains 차트

Gains 차트는 테이블에서 *이익(%)* 열에 있는 값을 구성합니다. 이익은 다음 방정식을 사용하여 트리에 있는 총 적중 수에 상대적인 각 증분의 적중 비율로 정의됩니다.

$$(\text{증분의 적중 수} / \text{총 적중 수}) \times 100\%$$

차트는 트리에 있는 모든 적중의 주어진 퍼센트를 캡처하기 위해 포함시켜야 하는 범위를 효과적으로 보여줍니다. 대각선은 모델을 사용하지 않는 경우 전체 샘플의 기대 반응을 구성합니다.

이 경우 한 사람이 다른 항목에 응답하는 것과 같기 때문에 반응률은 일정합니다. 두 배로 산출하려면 두 배 더 많은 사람들에게 질문해야 합니다. 곡선은 이익에 기반하여 더 높은 백분위수에 위치한 사람만 포함하여 반응을 얼마나 개선시킬 수 있는지 표시합니다. 예를 들어, 상위 50%만 포함하면 70% 이상의 긍정적인 반응이 돌아옵니다. 곡선이 가파를수록 이익이 높아집니다.

리프트 도표

리프트 도표는 테이블에서 *지수(%)* 열에 있는 값을 구성합니다. 이 차트는 다음 방정식을 사용하여, 학습 데이터 세트에 있는 전체 적중 퍼센트와 적중에 해당하는 각 증분에 있는 레코드 퍼센트를 비교합니다.

$$(\text{증분의 적중 수} / \text{증분의 레코드 수}) / (\text{총 적중 수} / \text{총 레코드 수})$$

반응 차트

반응 차트는 테이블의 *반응(%)* 열에 있는 값을 구성합니다. 반응은 다음 방정식을 사용하여 계산된, 적중에 해당하는 증분에 있는 레코드의 퍼센트입니다.

$$(\text{증분의 반응 수} / \text{증분의 레코드 수}) \times 100\%$$

마. 이익 기반 선택

이익 기반 선택 대화 상자에서는 지정된 규칙 또는 임계값에 따라 최상 또는 최저 이익을 포함하는 터미널 노드를 자동으로 선택할 수 있습니다. 그러면 선택에 따라 선택 노드를 생성할 수 있습니다.

1. 이익 탭에서 노드별 또는 누적 보기로 선택하고 선택의 기준으로 정할 목표 범주를 선택하십시오. (선택은 현재 테이블 표시에 기반하며 사분위수에서는 사용할 수 없습니다.)
2. 이익 탭의 메뉴에서 다음을 선택하십시오.

편집 > 터미널 노드 선택 > 이익 기반 선택

선택된 항목만. 매치 노드 또는 비매치 노드를 선택할 수 있습니다(예: 상위 100개 레코드 외 모두 선택).

이익 정보로 매치. 다음을 포함하여 현재 목표 범주의 이익 통계에 기반하는 매치 노드.

- 이익, 반응 또는 리프트(지수)가 지정된 임계값(예: 반응이 50% 이상)과 일치하는 노드.
- 목표 범주의 이익에 기반하는 상위 n개 노드.
- 지정된 레코드 수까지 상위 노드.
- 학습 데이터의 지정된 퍼센트까지 상위 노드.

3. **확인**을 클릭하여 뷰어 탭에서 선택을 업데이트하십시오.
4. 뷰어 탭에서 현재 선택에 기반하여 새 선택 노드를 작성하려면 생성 메뉴에서 **선택 노드**를 선택하십시오. 자세한 정보는 필터 및 선택 노드 생성의 내용을 참조하십시오.

참고: 실제로 레코드나 퍼센트가 아닌 노드를 선택하므로, 선택 기준과의 완벽한 매치는 항상 달성하지 못할 수도 있습니다. 시스템은 *최대* 지정된 수준까지 전체 노드를 선택합니다. 예를 들어, 상위 12개 케이스를 선택하고 처음 노드에 10개가 있고 두 번째 노드에 2개가 있으면, 처음 노드만 선택됩니다.

⑥ 위험

위험은 모든 수준에서 오분류의 확률을 알려줍니다. 위험 탭에서는 포인트 위험 추정값과 오분류 표(범주형 출력의 경우)를 표시합니다.

- 숫자 예측의 경우 위험은 각 터미널 노드에서 분산의 통합 추정값입니다.
- 범주형 예측의 경우 위험은 사전 또는 오분류 비용에 맞게 수정되었고, 잘못 분류된 사례의 비율입니다.

⑦ 트리 모델 및 결과 저장

다음은 포함하여 여러 방법으로 대화형 트리 작성 세션의 결과를 저장하거나 내보낼 수 있습니다.

- 현재 트리에 기반하여 모델을 생성하십시오(**생성 > 모델 생성**).
- 현재 트리를 성장시키는데 사용된 지시문을 저장하십시오. 다음에 트리 작성 노드를 실행할 때 사용자가 정의한 사용자 정의 분할을 포함하여 현재 트리가 자동으로 재성장됩니다.
- 모델, 이익, 위험 정보를 내보내십시오. 자세한 정보는 모델, 이익, 위험 정보 내보내기 주제를 참조하십시오.

트리 작성기 또는 트리 모델 너깃에서 다음을 수행할 수 있습니다.

- 현재 트리를 기반으로 필터 또는 선택 노드를 생성합니다. 자세한 정보는 필터 및 선택 노드 생성의 내용을 참조하십시오.
- 트리의 터미널 분기를 정의하는 규칙 세트로 트리 구조를 표시하는 규칙 세트 너깃을 생성합니다. 자세한 정보는 의사결정 트리에서 규칙 세트 생성의 내용을 참조하십시오.
- 또한 트리 모델 너깃의 경우에만 모델을 PMML 형식으로 내보낼 수 있습니다. 자세한 정보는 모델 팔레트의 내용을 참조하십시오. 모델에 사용자 정의 분할이 포함된 경우 이 정보는 내보낸 PMML에서 보존되지 않습니다. (분할은 보존되지만 알고리즘을 통해 선택된 것이 아니라 사용자 정의되었다는 사실은 그렇지 않습니다.)

- 현재 트리의 선택된 부분을 기반으로 그래프를 생성합니다. 스트림의 다른 노드에 연결되어 있을 경우에는 너깃에 대해서만 작동합니다. 자세한 정보는 그래프 생성의 내용을 참조하십시오.

참고: 대화형 트리 자체는 저장할 수 없습니다. 작업을 유실하지 않으려면 트리 작성기 창을 닫기 전에 모델을 생성하고/하거나 트리 지시문을 업데이트하십시오.

가. 트리 작성기에서 모델 생성

현재 트리에 기반한 모델을 생성하려면 트리 작성기 메뉴에서 다음을 선택하십시오.

생성 > 모델

새 모델 생성 대화 상자에 있는 다음 옵션 중에서 선택할 수 있습니다.

모델 이름. 사용자 정의 이름을 지정하거나 모델링 노드의 이름에 기반하여 자동으로 이름을 생성할 수 있습니다.

노드 작성 위치. 캔버스, GM 팔레트 또는 모두에서 노드를 추가할 수 있습니다.

트리 지시문 포함. 생성된 모델의 현재 트리에서 지시문을 포함하려면 이 상자를 선택합니다. 이를 통해 필요한 경우 트리를 재생성할 수 있습니다. 자세한 정보는 트리 성장 지시문의 내용을 참조하십시오.

나. 트리 성장 지시문

C&R 트리, CHAID, QUEST 모델의 경우 트리 지시문은 한 번에 한 개 수준씩 트리 성장 조건을 지정합니다. 지시문은 대화형 트리 작성기를 노드에서 실행할 때마다 적용됩니다.

- 지시문은 이전 대화형 세션 중에 작성된 트리를 재생성하는 방식으로 가장 안전하게 사용됩니다. 자세한 정보는 트리 지시문 업데이트 주제를 참조하십시오. 또한 지시문을 수동으로 편집할 수도 있지만, 신중을 기해야 합니다.
- 지시문은 지시문에서 설명하는 트리 구조에 특정합니다. 따라서 기본 데이터 또는 모델링 옵션을 변경하면 이전에 올바른 지시문 세트에서 문제가 발생할 수 있습니다. 예를 들어, CHAID 알고리즘이 업데이트된 데이터에 기반하여 이원 분할을 삼원 분할로 변경한 경우 이전 이원 분할에 기반한 지시문은 실패합니다.

참고: 직접 모델을 생성하려는 경우(트리 작성기를 사용하지 않음) 트리 지시문은 무시됩니다.

지시문 편집

1. 저장된 지시문을 보거나 편집하려면 트리 작성 노드를 열고 작성 옵션 탭의 목표 패널을 선택하십시오.
2. 대화형 세션 시작을 선택하여 제어를 사용 가능하게 하고 트리 지시문 사용을 선택하고 지시문을 클릭하십시오.

지시문 명령문

지시문은 루트 노드부터 시작하여 트리를 성장시키는 조건을 지정합니다. 예를 들어, 트리를 한 수준 성장시키려면:

```
Grow Node Index 0 Children 1 2
```

예측자를 지정하지 않으면 알고리즘은 최상의 분할을 선택합니다.

첫 번째 분할은 항상 루트 노드(Index 0)에 있어야 하며 두 하위의 지수 값을 지정해야 합니다 (이 경우 1 및 2). Node 2를 작성한 루트를 처음 성장시키는 경우가 아니라면 Grow Node Index 2 Children 3 4를 지정하는 구문은 유효하지 않습니다.

트리를 성장시키려면:

```
트리 성장
```

트리 성장 및 가지치기를 수행하려면(C&R 트리만 해당):

```
Grow_And_Prune Tree
```

연속형 예측자에 대한 사용자 정의 분할을 지정하려면:

```
Grow Node Index 0 Children 1 2 Spliton  
( "EDUCATE", Interval ( NegativeInfinity, 12.5)  
Interval ( 12.5, Infinity ) )
```

값이 2개인 명목 예측자에서 분할하려면:

```
Grow Node Index 2 Children 3 4 Spliton  
( "GENDER", Group( "0.0" )Group( "1.0" ) )
```

값이 여러 개인 명목 예측자에서 분할하려면:

```
Grow Node Index 6 Children 7 8 Spliton  
( "ORGS", Group( "2.0","4.0" )  
Group( "0.0","1.0","3.0","6.0" ) )
```

순서 예측자에서 분할하려면:

```
Grow Node Index 4 Children 5 6 Spliton  
( "CHILDS", Interval ( NegativeInfinity, 1.0)  
Interval ( 1.0, Infinity ))
```

참고: 사용자 정의 분할을 지정하면 필드 이름 및 값(EDUCATE, GENDER, CHILDS 등)은 대소문자를 구분합니다.

CHAID 트리에 대한 지시문

CHAID 트리의 지시문은 특히, 데이터 또는 모델에서의 변경에 민감합니다. C&R 트리 및 QUEST와 달리 이분형 분할 사용이 제한되지 않기 때문입니다. 예를 들어, 다음 구문은 완벽하게 유효해 보이지만, 알고리즘이 루트 노드를 셋 이상의 하위로 분할할 경우 실패합니다.

```
Grow Node Index 0 Children 1 2  
Grow Node Index 1 Children 3 4
```

CHAID에서는 Node 0이 3개 또는 4개의 하위를 포함할 수 있으며, 이로 인해 구문의 두 번째 줄이 실패할 수 있습니다.

스크립트에서 지시문 사용

또한 지시문은 삼중 따옴표를 사용하여 스크립트에 임베드될 수 있습니다.

다. 트리 지시문 업데이트

대화형 트리 작성 세션에서 작업을 유지하기 위해 현재 트리를 생성하는 데 사용된 지시문을 저장할 수 있습니다. 추가로 편집할 수 없는 모델 너깃 저장과는 달리, 이를 통해 추가로 편집하도록 현재 상태에서 트리를 재생성할 수 있습니다.

지시문을 업데이트하려면 트리 작성기 메뉴에서 다음을 선택하십시오.

파일 > 지시문 업데이트

지시문은 트리(C&R 트리, QUEST 또는 CHAID)를 작성하는 데 사용된 모델링 노드에 저장되고 이를 사용하여 현재 트리를 재생성할 수 있습니다. 자세한 정보는 트리 성장 지시문 주제를 참조하십시오.

라. 모델, 이익, 위험 정보 내보내기

트리 작성기에서 모델, 이익, 위험 통계를 텍스트, HTML 또는 이미지 형식으로 적절히 내보낼 수 있습니다.

1. 트리 작성기 창에서 내보내려는 탭 또는 보기를 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.

파일 > 내보내기

3. 텍스트, HTML 또는 그래프를 적절히 선택하고 하위 메뉴에서 내보낼 특정 항목을 선택하십시오.

해당되는 경우 내보내기는 현재 선택에 기반합니다.

텍스트 또는 HTML 형식 내보내기. 학습 또는 검정 분할(정의된 경우) 이익 또는 위험 통계를 내보낼 수 있습니다. 내보내기는 이익 탭의 현재 선택에 기반합니다. 예를 들어, 노드별, 누적 또는 분위수 통계를 선택할 수 있습니다.

그래픽 내보내기. 뷰어 탭에 표시된 대로 현재 트리를 내보내거나 학습 또는 검정 분할(정의된 경우)에 대한 Gains 차트를 내보낼 수 있습니다. 사용 가능한 형식으로는 *.JPEG*, *.PNG*, *.BMP*가 있습니다. 이익의 경우 내보내기는 이익 탭(차트가 표시되는 경우에만 사용 가능함)에서 현재 선택에 기반합니다.

⑧ 필터 및 선택 노드 생성

트리 작성기 창에서 또는 의사결정 트리 모형 너깃을 찾아볼 때 메뉴에서 다음을 선택하십시오.

생성 > 필터 노드

or

> 선택 노드

필터 노드. 현재 트리에서 사용하지 않는 필드를 필터링하는 노드를 생성합니다. 알고리즘에서 중요한 항목으로 선택된 해당 필드만 포함하도록 데이터 세트를 줄이는 가장 빠른 방법입니다. 이 의사결정 트리 노드에서 유형 노드 업스트림이 있으면 역할이 목표인 모든 필드가 필터 모델 너깃에서 전달됩니다.

선택 노드. 현재 노드에 포함되는 모든 레코드를 선택하는 노드를 생성합니다. 이 옵션에서는 뷰어 탭에서 하나 이상의 트리 분기를 선택해야 합니다.

모델 너깃은 스트림 캔버스에 배치됩니다.

⑨ 의사결정 트리에서 규칙 세트 생성

트리의 터미널 분기를 정의하는 규칙 세트로 트리 구조를 나타내는 규칙 세트 모델 너깃을 생성할 수 있습니다. 규칙 세트는 종종 전체 의사결정 트리(단, 보다 덜 복잡한 모델 포함)에서 대부분

분의 중요한 정보를 보유할 수 있습니다. 가장 중요한 차이는 규칙 세트를 포함하는 경우 둘 이상의 규칙이 특정 레코드에 적용되거나 규칙이 전혀 적용되지 않다는 점입니다. 예를 들어, *no* 결과와 뒤에 *yes*를 예측하는 모든 규칙이 나오는 모든 규칙을 확인할 수 있습니다. 다중 규칙이 적용되는 경우 각 규칙은 해당 규칙과 연관된 신뢰도에 기반하여 가중된 "투표"를 확보하고 최종 예측은 문제가 되는 레코드에 적용되는 모든 규칙의 가중된 투표를 결합하여 결정됩니다. 적용된 규칙이 없으면 기본 예측이 레코드에 지정됩니다.

참고: 규칙 세트 스코어를 계산할 때 트리에서의 스코어링과 비교했을 때 스코어링의 차이를 확인할 수 있습니다. 트리의 각 터미널 분기에서 독립적으로 스코어가 계산되기 때문입니다. 이 차이가 눈에 띄는 만큼 큰 영역은 데이터에 결측값이 있는 경우입니다.

규칙 세트는 범주형 대상 필드(회귀분석 트리 없음)를 포함하는 트리에서만 생성할 수 있습니다.

트리 작성기 창에서 또는 의사결정 트리 모형 너깃을 찾아볼 때 메뉴에서 다음을 선택하십시오.

생성 > 규칙 세트

규칙 세트 이름 새 규칙 세트 모델 너깃 이름을 지정합니다.

노드 작성 위치 새 규칙 세트 모델 너깃의 위치를 제어합니다. **캔버스**, **GM 팔레트** 또는 **모두**를 선택하십시오.

최소 인스턴스 규칙 세트 모델 너깃에서 보존할 최소 인스턴스 수(규칙이 적용되는 레코드 수)를 지정합니다. 지원이 지정된 값보다 적은 규칙은 새 규칙 세트에 포함되지 않습니다.

최소 신뢰도 규칙 세트 모델 너깃에서 유지할 규칙의 최소 신뢰도를 지정합니다. 신뢰도가 지정된 값보다 적은 규칙은 새 규칙 세트에 포함되지 않습니다.

(3) 직접 트리 모델 작성

대화형 트리 작성기 사용의 대안으로, 스트림을 실행할 때 노드에서 직접 의사결정 트리 모형을 작성할 수 있습니다. 이는 대부분의 다른 모델 작성 노드에서도 일관됩니다. C5.0 트리 및 Tree-AS 모델의 경우(대화형 트리 작성기에서 지원하지 않음) 이는 사용할 수 있는 유일한 방법입니다.

1. 스트림을 작성하고 의사결정 트리 노드, C&R 트리, CHAID 또는 QUEST, C5.0, 또는 Tree-AS 중 하나를 추가하십시오.
2. C&R 트리, QUEST 또는 CHAID의 경우 작성 옵션 탭의 목표 패널에서 주 목표 중 하나를 선택하십시오. 단일 트리 작성을 선택한 경우 **모드를 모델 생성**으로 설정해야 합니다. C5.0의 경우 모델 탭에서 **출력 유형을 의사결정 트리**로 설정하십시오.

Tree-AS의 경우 작성 옵션 탭의 기본 패널에서 **트리 성장 알고리즘 유형**을 선택하십시오.

3. 목표 및 예측자 필드를 선택하고 필요한 경우 추가 모델 옵션을 지정하십시오. 특정 지시사항의 경우 각 트리 작성 노드에 대한 문서를 참조하십시오.

4. 스트림을 실행하여 모델을 생성합니다.

트리 작성에 대한 설명

- 이 방법을 사용하여 트리를 생성하는 경우 트리 성장 지시문은 무시됩니다.
- 대화형인지 직접인지에 상관없이 의사결정 트리를 작성하는 두 방법은 궁극적으로 유사한 모델을 생성합니다. 제어 범위가 달라질 뿐입니다.

(4) 의사결정 트리 노드

IBM® SPSS® Modeler의 의사결정 트리 노드는 다음 트리 작성 알고리즘에 대한 액세스를 제공합니다.

- | | | |
|----------|---------|-----------|
| - C&R 트리 | - CHAID | - Tree-AS |
| - QUEST | - C5.0 | - 임의 트리 |

자세한 정보는 의사결정 트리 모형 주제를 참조하십시오.

알고리즘은 데이터를 작은 하위 그룹으로 분할하여 반복적으로 의사결정 트리를 구성할 수 있다는 점에서 유사합니다. 그러나 일부 중요한 차이가 있습니다.

입력 필드. 입력 필드(예측자)는 연속형, 범주형, 플래그, 명목형 또는 순서와 같은 유형(측정 수준)이 될 수 있습니다.

목표 필드. 목표 필드는 하나만 지정할 수 있습니다. C&R 트리, CHAID, Tree-AS, 랜덤 트리의 경우, 대상은 연속형, 범주형, 플래그, 명목형 또는 순서일 수 있습니다. QUEST의 경우, 범주형, 플래그 또는 명목형이 될 수 있습니다. C5.0의 경우, 목표는 플래그, 명목형 또는 순서가 될 수 있습니다.

분할 유형. C&R 트리, QUEST, 랜덤 트리는 이분형 분할만 지원합니다(즉, 트리의 각 노드는 두 개 이하의 분기로만 분할할 수 있습니다). 반대로, CHAID, C5.0 및 Tree-AS는 한 번에 세 개 이상의 분기로의 분할을 지원합니다.

분할에 사용되는 방법. 알고리즘은 분할을 결정하기 위해 사용되는 기준에서 다릅니다. C&R 트리가 범주형 출력을 예측할 경우, 산포도 측도가 사용됩니다(기본값은 Gini 계수이며, 변경할 수 있습니다). 연속형 목표의 경우, 최소 편차 제공 방법이 사용됩니다. CHAID 및 Tree-AS는 카이제곱 검정을 사용합니다. QUEST는 범주형 예측자에 대해 카이제곱 검정을 사용하고 연속형 입력에 대해 공차 분석을 사용합니다. C5.0의 경우 정보 이론 측도가 사용됩니다(정보 이익 비율).

결측값 처리. 모든 알고리즘은 예측자 필드에 대한 결측값을 허용합니다. 알고리즘은 여러 방법으로 결측값을 처리합니다. C&R 트리 및 QUEST는 대체 예측 필드를 사용하여(필요한 경우) 훈련 동안 트리를 통해 결측값이 있는 레코드로 진행합니다. CHAID는 결측값을 별도의 범주를 작성하고 트리 작성에서 사용되도록 합니다. C5.0은 결측값이 있는 필드를 기반으로 분할이 이뤄지는 노드에서 트리의 각 분기로 레코드의 일부를 전달하는 비율(fractioning) 방법을 사용합니다.

가지치기. C&R 트리, QUEST 및 C5.0은 트리를 완전하게 증가시키기 위한 옵션을 제공하고 트리 정확도에 유의적으로 기여하지 않는 하위 수준 분할을 제거하여 다시 가지치기를 합니다. 그러나 모든 의사결정 트리 알고리즘은 몇 개의 데이터 레코드만 있는 분기를 피할 수 있도록 최소 하위 그룹 크기를 제어할 수 있도록 허용합니다.

대화형 트리 작성. C&R 트리, QUEST 및 CHAID는 대화형 세션을 실행하기 위한 옵션을 제공합니다. 그러면 한 번에 한 수준씩 트리를 작성하고, 분할을 편집하며, 모델 작성 전에 트리를 가지치기할 수 있습니다. C5.0, Tree-AS, 랜덤 트리에는 대화형 옵션이 없습니다.

사전 확률. C&R 트리 및 QUEST는 범주형 목표 필드를 예측할 때 범주에 대한 사전 확률의 지정을 지원합니다. 사전 확률은 훈련 데이터가 그려지는 모집단에서 각 목표 범주에 대한 전체 상대 빈도의 추정값입니다. 즉, 예측자 값에 대한 어떤 사항을 알기 전에 각각의 가능한 목표 값에 대해 추정하는 확률 추정값입니다. CHAID, C5.0, Tree-AS, 랜덤 트리는 사전확률 지정을 지원하지 않습니다.

규칙 세트. Tree-AS 또는 랜덤 트리에 사용할 수 없습니다. 범주형 목표 필드가 있는 모델의 경우, 의사결정 트리 노드는 간혹 복잡한 의사결정 트리보다 해석하기 쉬울 수 있는 규칙 세트 양식으로 모델을 작성할 수 있는 옵션을 제공합니다. C&R 트리, QUEST 및 CHAID의 경우 대화형 세션을 통해 규칙 세트를 생성하고, C5.0의 경우 모델링 노드에서 이 옵션을 지정할 수 있습니다. 또한 모든 의사결정 트리 모형은 모델 너깅에서 설정된 규칙을 생성할 수 있도록 합니다. 자세한 정보는 의사결정 트리에서 규칙 세트 생성 주제를 참조하십시오.

① C&R 트리 노드

분류 및 회귀분석(C&R) 트리 노드는 트리 기반의 분류 및 예측 방법입니다. C5.0과 마찬가지로, 이 방법은 재귀적 분할을 사용하여 훈련 레코드를 출력 필드 값이 유사한 세그먼트로 분할합니다. C&R 트리 노드는 분할로 인한 불순도 지수를 줄여서 측정되는 최상의 분할을 찾기 위해 입력 필드를 검토하는 것으로 시작합니다. 분할이 두 개의 하위 그룹을 정의하고, 중지 기준 중 하나가 트리거될 때까지 각 그룹은 계속해서 두 개의 추가 하위 그룹으로 분할되는 식입니다. 모든 분할은 이분형(하위 그룹을 두 개만)입니다.

가지치기

C&R 트리는 처음에 트리를 성장시킨 후 터미널 노드 수에 따라 위험 추정값을 조정하는 비용

복잡도 알고리즘을 기준으로 하여 가지치기를 수행할 옵션을 제공합니다. 보다 복잡한 기준에 따라 가지치기를 수행하기 전에 트리를 성장시키는 이 방법으로 트리가 더 작아지고 교차 검증 특성은 개선될 수 있습니다. 터미널 노드 수를 늘리면 일반적으로 현재(훈련) 데이터의 위험이 감소하지만 모델이 보이지 않는 데이터로 일반화될 때 실제 위험이 더 커질 수 있습니다. 극단적인 경우 훈련 세트에서 각 레코드마다 별도의 터미널 노드가 있다고 가정하십시오. 모든 레코드가 자신의 노드이므로 위험 추정값이 0%이지만 보이지 않는(검정) 데이터의 오분류 위험은 거의 확실하게 0보다 큼니다. 비용 복잡도 측도가 이를 보완하려 시도합니다.

예. 케이블 TV 회사는 어느 고객이 케이블을 통해 대화형 뉴스 서비스에 등록하는지 판별하기 위해 마케팅 연구를 의뢰했습니다. 연구 데이터를 사용하여 목표 필드가 등록을 구매할 의도이고 예측자 필드가 나이, 성별, 교육, 수입 범주, 매일 TV 시청에 소모하는 시간, 자녀 수를 포함하는 스트림을 작성할 수 있습니다. C&R 트리 노드를 스트림에 적용하여 캠페인의 최고 반응률을 얻도록 반응을 예측 및 분류할 수 있습니다.

요구사항. C&R 트리 모델을 훈련하려면 하나 이상의 입력 필드 및 목표 필드가 정확히 하나 필요합니다. 목표 및 입력 필드는 연속형(수치 범위) 또는 범주형이 가능합니다. 둘 다 또는 없음으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 하고 모델에 사용된 순서(정렬된 세트) 필드에는 수치 저장 공간(문자열이 아닌)이 있어야 합니다. 필요한 경우 재분류 노드를 사용하여 변환할 수 있습니다.

강도. C&R 트리 모델은 데이터 누락이나 많은 수의 필드와 같은 문제가 발생할 때 상당히 강건합니다. 일반적으로 추정하기 위해 긴 학습 시간이 필요하지 않습니다. 또한 C&R 트리 모델은 모델에서 파생된 규칙의 해석이 매우 직설적이어서 다른 모델 유형보다 이해하기 쉽습니다. C5.0와 달리, C&R 트리는 연속형 및 범주형 출력 필드를 수용할 수 있습니다.

② CHAID 노드

CHAID 또는 카이제곱 자동 상호작용 발견은 카이제곱 통계량을 사용하여 최적의 분할을 식별해서 의사결정 트리를 작성하기 위한 분류 방법입니다.

먼저 CHAID는 각 입력 필드와 출력 사이의 교차 분석표를 탐색하고 카이제곱 독립 검정을 사용하여 유의수준을 검정합니다. 둘 이상의 관계가 통계적으로 유의적이면 CHAID는 가장 유의적인(최소 p 값) 입력 필드를 선택합니다. 입력에 둘 이상의 범주가 있는 경우에는 이 범주를 비교하고 결과에 차이가 없는 범주는 함께 접습니다. 최소유의차를 표시하는 범주 쌍을 연속으로 결합해서 이를 수행합니다. 나머지 모든 범주가 지정된 검정 수준에서 서로 다르면 이 범주 병합 프로세스는 중지됩니다. 명목 입력 필드의 경우 범주가 병합될 수 있으며 순서 세트의 경우에는 연속형 범주만 병합될 수 있습니다.

Exhaustive CHAID는 각 예측자에 대한 모든 가능한 분할을 탐색하는 보다 전반적인 작업을 수행하지만 계산 시간이 오래 걸리는 CHAID의 수정 모델입니다.

요구사항. 목표 및 입력 필드는 연속형 또는 범주형이 가능하고 노드는 각 수준에서 둘 이상의 하위 그룹으로 분할될 수 있습니다. 모델에 사용된 순서 필드에 숫자 저장 공간(문자열이 아님)이 있어야 합니다. 필요한 경우 재분류 노드를 사용하여 변환할 수 있습니다.

강도. C&R 트리 및 QUEST 노드와 달리, CHAID는 비이분형 트리를 생성할 수 있으며 이는 일부 분할에 둘 이상의 분기가 있음을 의미합니다. 따라서 이 노드는 이분형 성장 방법보다 광범위한 트리를 작성하는 경향이 있습니다. CHAID는 모든 유형의 입력에 작용하며 케이스 가중치 및 빈도 변수를 모두 허용합니다.

③ QUEST 노드

QUEST(또는 Quick, Unbiased, Efficient Statistical Tree)는 의사결정 트리를 작성하는 이분형 분류 방법입니다. 해당 개발에서 주요 동기는 많은 변수나 많은 케이스를 포함하는 대형 C&R 트리 분석에 필요한 처리 시간을 줄이는 데 있습니다. QUEST의 두 번째 목표는 더 많은 분할을 허용하는 입력, 즉 연속적인(수적 범위) 입력 필드 또는 많은 범주의 입력 필드를 위해 분류 트리 방법에서 찾은 경향을 줄이는 것입니다.

- QUEST는 노드에서 입력 필드를 평가하기 위해 유의수준 검정에 기반하여 일련의 규칙을 사용합니다. 선택 목적으로 노드의 각 입력에서 최소 단일 검정을 수행해야 할 수도 있습니다. C&R 트리와 달리, 모든 분할을 탐색하지 않습니다. 또한 C&R 트리 및 CHAID와 달리, 선택을 위해 입력 필드를 평가할 때 범주형 조합을 검정하지 않습니다. 그러면 분석 속도가 빨라집니다.
- 분할은 목표 범주에서 구성된 그룹에서 선택한 입력을 통해 2차 판별 분석을 실행하여 판별됩니다. 이 방법은 최적의 분할을 판별하기 위해 다시 소모적 검색(C&R 트리)에서 속도를 향상시킵니다.

요구사항. 입력 필드는 연속형(숫자 범위)일 수 있지만, 목표 필드는 범주형이어야 합니다. 모든 분할은 이분형입니다. 가중 필드는 사용할 수 없습니다. 모델에 사용된 순서 필드(정렬된 세트)에 숫자 저장 공간(문자열이 아님)이 있어야 합니다. 필요한 경우 재분류 노드를 사용하여 변환할 수 있습니다.

강도. CHAID와는 비슷하지만, C&R 트리와는 달리, QUEST는 입력 필드의 사용 여부를 결정하기 위해 통계 검정을 사용합니다. 또한 입력 선택과 분할의 문제를 구분하여 각각에 서로 다른 기준을 적용합니다. 이는, 변수 선택을 판별하는 통계 검정 결과가 분할도 생성하는 CHAID와는 대비됩니다. 마찬가지로, C&R 트리는 불순도-변경 측도를 사용하여 입력 필드를 선택하고 분할을 판별합니다.

④ 의사결정 트리 노드 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 역할 설정(목표, 예측자 등)을 사용합니다.

사용자 정의 필드 할당 사용: 수동으로 대상, 예측자 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

필드. 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

목록의 모든 필드를 선택하려면 **모두** 단추를 클릭하거나 개별 측정 수준 단추를 클릭하여 이 측정 수준의 모든 필드를 선택하십시오.

목표: 예측에 대한 목표로 하나의 필드를 선택합니다.

예측변수(입력). 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

분석 가중값. (CHAID, C&RT, Trees-AS만) 케이스 가중치로 필드를 사용하려면 여기에서 필드를 지정합니다. 케이스 가중치는 출력 필드의 수준에서 분산의 차이를 설명하는 데 사용됩니다. 자세한 정보는 빈도 및 가중 필드 사용 주제를 참조하십시오.

⑤ 의사결정 트리 노드 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, **실행** 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

탭은 모델에 특정한 사용자 정의를 설정하는 여러 창을 포함합니다.

가. 의사결정 트리 노드 - 목적

C&R 트리, QUEST 및 CHAID 노드의 경우, 작성 옵션 탭의 목적 분할창에서 새 모델을 작성하거나 기존 모델을 업데이트할 것을 선택할 수 있습니다. 또한 노드의 주 목적을 설정할 수도 있습니다(표준 모델을 작성하거나, 고급 정확도 또는 안정성을 사용하여 작성하거나, 매우 큰 데이터 세트와 함께 사용하기 위해 작성하기 위해).

원하는 작업

새 모델 작성. (기본값) 이 모델링 노드를 포함하는 스트림을 실행할 때마다 새 모델을 완전하게 작성합니다.

기존 모델 훈련 계속. 기본적으로 모델링 노드가 실행될 때마다 완전한 새 모델이 작성됩니다. 이 옵션을 선택할 경우 노드가 정상적으로 생성한 마지막 모델로 훈련을 계속합니다. 이를 통해 원래 데이터에 액세스할 필요 없이 기존 모델을 업데이트하거나 새로 고칠 수 있어서 오직 새 레코드 또는 업데이트된 레코드만 스트림에 입력되므로 상당히 빠르게 작업을 수행할 수 있습니다. 이전 모델에 대한 세부 사항이 모델링 노드와 함께 저장되어 스트림 또는 모델 팔레트에서 이전 모델 너깃을 더 이상 사용할 수 없는 경우에도 이 옵션을 사용할 수 있습니다.

참고: 이 옵션은 **단일 트리 작성**(C&R 트리, CHAID 및 QUEST의 경우), **표준 모델 작성**(신경망 및 선형의 경우) 또는 **매우 큰 데이터 세트에 대한 모델 작성**을 목적으로 선택하는 경우에만 활성화됩니다.

원하는 기본 목적

- **단일 트리 작성.** 단일의 표준 의사결정 트리 모형을 작성합니다. 표준 모델은 일반적으로 해석하기 쉬우므로, 다른 목적 옵션을 사용하여 작성되는 모델보다 스코어링이 더 빠를 수 있습니다.

참고: 분할 모델의 경우, **기존 모델 훈련 계속**과 함께 이 옵션을 사용하려면 Analytic Server에 연결되어 있어야 합니다.

모드 모델 작성에 사용되는 방법을 지정합니다. **모델 생성**은 스트림이 실행될 때 자동으로 모델을 작성합니다. **대화형 세션 시작**은 트리 작성기를 엽니다. 이 작성기에서는 한 번에 하나의 수준에서 트리를 작성하고, 분할을 편집하며, 모델 너깃 작성 전에 원하는 대로 가지칠 수 있습니다.

트리 지시문 사용. 노드로부터 대화식 트리를 생성할 때 적용할 지시문을 지정하려면 이 옵션을 선택하십시오. 예를 들어, 첫 번째 및 두 번째 수준 분할을 지정할 수 있고, 이 분할은 트리 작성기가 실행될 때 자동으로 분할됩니다. 또한 나중 날짜에 트리를 다시 작성하기 위해 대화식 트리 작성 세션에서 지시문을 저장할 수도 있습니다. 자세한 정보는 트리 지시문 업데이트 주제를 참조하십시오.

- **모형 정확도(부스팅) 개선.** 모형 정확도 비율을 향상시키기 위해 **부스팅**이라고 하는 특수 방법을 사용하려는 경우 이 옵션을 선택하십시오. 부스팅은 여러 모델을 순차적으로 작성하는 방식으로 작동합니다. 첫 번째 모델은 일반적인 방법으로 작성됩니다. 그런 다음 두 번째 모델은 첫 번째 모델이 잘못 분류한 레코드에 초점을 맞추는 방식으로 작성됩니다. 그리고 나서 세 번째 모델은 두 번째 모델의 오류에 초점을 맞추기 위해 작성됩니다. 그 다음도 마찬가지입니다. 마지막으로 전체 모델 세트를 케이스에 적용하고 가중 투표 프로시저를 사용하여 개별 예측을 하나의 전체 예측으로 결합해서 케이스를 분류합니다. 부스팅은 의사결정 트리 모형의 정확도를 유의미하게 개선할 수 있지만, 더 오랜 훈련이 필요합니다.

- **모델 안정성(배깅) 개선.** 모델 안정성을 개선하고 과적합을 피하기 위해 **배깅**(붓스트랩 통합) 이라고 하는 특수 방법을 사용하려는 경우 이 옵션을 선택하십시오. 이 옵션은 한층 신뢰할 만한 예측을 확보하기 위해 여러 모델을 작성하여 조합합니다. 이 옵션 사용으로 확보되는 모델은 표준 모델보다 작성 및 스코어링에 긴 시간이 소요될 수 있습니다.
- **매우 큰 데이터 세트를 위한 모델 작성.** 너무 커서 다른 목적 옵션을 사용하여 모델을 작성할 수 없는 데이터 세트에 대해 작업할 때 이 옵션을 선택하십시오. 이 옵션은 데이터를 더 작은 데이터 블록으로 나누고, 각각의 블록에서 모델을 작성합니다. 가장 정확한 모델은 자동으로 선택되어 단일 모델 너깃에 결합됩니다. 이 화면에서 **기존 모델 훈련 계속** 옵션을 선택하면 점증적 모델 업데이트를 수행할 수 있습니다.

 **참고:** 이 대형 데이터 세트 옵션에는 IBM® SPSS® Modeler Server에 대한 연결이 필요합니다.

나. 의사결정 트리 노드 - 기본

의사결정 트리 작성 방법에 대한 기본 옵션을 지정하십시오.

트리 성장 알고리즘 (CHAID 및 Tree-AS만 해당) 사용하려는 **CHAID** 알고리즘 유형을 선택합니다. **Exhaustive CHAID**는 각 예측자에 대한 모든 가능한 분할을 탐색하는 보다 전반적인 작업을 수행하지만 계산 시간이 오래 걸리는 CHAID의 수정 모델입니다.

최대 트리 깊이 루트 노드 아래의 최대 수준 수를 지정합니다(표본이 반복적으로 분할된 횟수). 기본값은 5입니다. **사용자 정의를** 선택하고 값을 입력하여 여러 수준 수를 지정합니다.

가지치기(C&RT 및 QUEST만 해당)

과적합을 방지하기 위해 트리 가지치기 가지치기는 트리 정확도에 거의 기여하지 않는 아래쪽 수준의 분할을 제거하는 작업으로 구성됩니다. 가지치기는 트리를 단순화시켜 더 쉽게 해석하고 종종 일반화를 향상시키기도 합니다 가지치기 없이 전체 트리를 원하는 경우 이 옵션을 선택 취소한 상태로 두십시오.

- **표준 오차의 최대 위험차 설정** 더 자유로운 가지치기 규칙을 지정할 수 있습니다. 표준 오차 규칙에서는 알고리즘에서 위험 추정값이 가장 작은 위험을 포함하는 서브트리의 값에 근사한 (클 수도 있음) 가장 단순한 트리를 선택할 수 있습니다. 이때 값은 가지치기한 트리과 위험 추정값 측면에서 가장 작은 위험을 지닌 트리 사이에서 허용 가능한 위험 추정값 차이의 크기를 나타냅니다. 예를 들어, 2를 지정하면 위험 추정값이 전체 트리의 위험 추정값보다 큰(2 × 표준 오차) 트리가 선택될 수 있습니다.

최대 대응. 대응은 결측값을 처리하기 위한 방법입니다. 트리의 각 분할에서 알고리즘은 선택한 분할 필드와 가장 유사한 입력 필드를 식별합니다. 이러한 필드를 해당 분할의 *d//용*이라고 합니다. 레코드를 분류해야 하지만 분할 필드에 결측값이 있으면 대응 필드의 해당 값을 사용하여 분할을 수행할 수 있습니다. 이 설정을 늘리면 결측값을 보다 탄력적으로 처리할 수 있지만, 메모리 사용량이 늘어나고 훈련 시간이 더 길어질 수 있습니다.

다. 의사결정 트리 노드 - 중지 규칙

이 옵션은 트리 구성 방법을 제어합니다. 중지 규칙은 트리의 분할 특정 분기를 중지하는 시점을 판별합니다. 매우 작은 하위 그룹을 작성하는 분할을 방지하도록 최소 분기 크기를 설정합니다. **부모마디 최소 레코드 수**는 분할할 노드(상위)의 레코드 수가 지정된 값 미만인 경우 분할을 방지합니다. **자식마디 최소 레코드 수**는 분할로 작성된 분기(하위)의 레코드 수가 지정된 값 미만인 경우 분할을 방지합니다.

- **퍼센트 사용** 전체 훈련 데이터의 퍼센트 관점에서 크기를 지정합니다.
- **절대값 사용** 레코드의 절대값으로 크기를 지정합니다.

라. 의사결정 트리 노드 - 앙상블

이 설정은 부스팅, 배깅 또는 아주 큰 데이터 세트가 목표에서 요청될 때 발생하는 앙상블 동작을 결정합니다. 선택한 목표에 해당하지 않는 옵션은 무시됩니다.

배깅 및 아주 큰 데이터 세트. 앙상블을 스코어링할 때 앙상블 스코어값을 계산하기 위해 기본 모델에서 예측값을 조합하는 데 사용되는 규칙입니다.

- **범주형 목표의 기본 결합 규칙.** 범주형 목표에 대한 앙상블 예측값은 투표, 최고 확률 또는 최고 평균 확률을 사용하여 조합될 수 있습니다. **투표**는 기본 모델에서 최고 확률을 가지는 범주를 선택합니다. **최고 확률**은 모든 기본 모델에서 단일 최고 확률을 획득하는 범주를 선택합니다. **최고 평균 확률**은 범주 확률이 기본 모델에서 평균이 될 때 최고값이 있는 범주를 선택합니다.
- **연속형 목표의 기본 결합 규칙.** 연속형 목표에 대한 앙상블 예측값은 기본 모델의 예측값 평균 또는 중앙값을 사용하여 조합될 수 있습니다.

모형 정확도를 향상시키는 것이 목표인 경우 결합 규칙 선택이 무시됨에 유의하십시오. 부스팅은 항상 가중 다수 투표를 사용하여 범주형 목표를 스코어링하고 가중 중앙값을 사용하여 연속형 목표를 스코어링합니다.

부스팅 및 배깅. 모형 정확도 또는 안정성을 향상시키는 것이 목표일 때 작성할 기본 모형 수를 지정하십시오. 배깅의 경우, 붓스트랩 표본의 수입니다. 양의 정수여야 합니다.

마. C&R 트리 및 QUEST 노드 - 비용 및 사전

오분류 비용

일부 컨텍스트에서는 특정 오차 유형이 다른 유형에 비해 더 값비쌉니다. 예를 들어, 고위험 신용 거래 신청자를 저위험(오차의 한 유형)으로 분류하는 것이 저위험 신청자를 고위험(다른 유형의 오차)으로 분류하는 것보다 비용이 더 나갈 수 있습니다. 오분류 비용을 통해 여러 다른 유형의 예측 오차의 상대적 중요도를 지정할 수 있습니다.

오분류 비용은 기본적으로 특정 결과에 적용된 가중값입니다. 이 가중값은 모델에 영향을 미치는 요인이 되어 실제로 예측변수를 변경합니다(값비싼 실수에 대비하는 보호의 한 방법으로).

C5.0 모델을 제외하고, 오분류 비용은 모델을 스코어링할 때 적용되지 않으며 자동 분류자 노드, 평가 노드 또는 분석 노드를 사용하여 모델을 순위화하거나 비교할 때 고려되지 않습니다. 비용을 포함한 모델은 포함하지 않은 모델보다 적은 오차를 생성하지 않고 전체 정확도 측면에서 더 높게 순위화되지 않을 수는 있지만, *보다 저렴* 오차를 선호하는 편향이 내재되어 있어서 실제적으로 성능이 더 좋을 수 있습니다.

비용 교차표는 예측 범주와 실제 범주의 가능한 각 조합의 비용을 표시합니다. 기본적으로 모든 오분류 비용은 1.0으로 설정됩니다. 사용자 정의 비용 값을 입력하려면 **오분류 비용 사용**을 선택하고 비용 교차표에 사용자 정의 값을 입력하십시오.

오분류 비용을 변경하려면 예측 및 실제 값의 원하는 조합에 해당하는 셀을 선택하고 셀의 기존 콘텐츠를 삭제한 후 셀의 원하는 비용을 입력하십시오. 비용은 자동으로 대칭되지 않습니다. 예를 들어, A를 B로 오분류한 비용을 2.0으로 설정할 경우 이 설정을 명시적으로 변경하지 않으면 B를 A로 오분류하는 비용의 기본값은 여전히 1.0입니다.

사전

이 옵션으로 범주형 목표 필드를 예상할 때 범주에 대한 사전 확률을 지정할 수 있습니다. **사전 확률**은 훈련 데이터를 그리는 모집단의 각 목표 범주에 대한 전체 상대 빈도의 추정값입니다. 즉, 예측자 값에 대한 무언가를 알기 *이전의* 가능한 각 목표 값에 대한 확률 추정값입니다. 사전 확률을 설정하는 세 가지 방법이 있습니다.

- **훈련 데이터 기준.** 이는 기본값입니다. 사전 확률은 훈련 데이터에서 범주의 상대 빈도를 기반으로 합니다.
- **모든 클래스에 대해 동등함.** 모든 범주의 사전 확률이 $1/k$ 로 정의되며, k 는 목표 범주의 수입니다.
- **사용자 정의.** 사용자가 직접 사전 확률을 지정할 수 있습니다. 사전 확률의 시작값은 모든 클래스에 대해 동등함으로 설정됩니다. 사용자 정의 값에 대한 개별 범주의 확률을 조정할 수 있습니다. 특정 범주의 확률을 조정하려면 원하는 범주에 해당하는 테이블의 확률 셀을 선택하고 셀의 콘텐츠를 삭제한 후 원하는 값을 입력하십시오.

모든 범주의 사전 확률 합계는 1.0이어야 합니다(**확률 제한조건**). 합계가 1.0이 아니면 값을 자동으로 표준화하는 옵션과 함께 경고 메시지가 표시됩니다. 이 자동 조정은 확률 제한조건을 시행하면서 범주 전체에서 비율을 유지합니다. 언제든지 **표준화** 단추를 클릭해서 이 조정을 수행할 수 있습니다. 모든 범주에 동일한 값으로 테이블을 재설정하려면 **평균화** 단추를 클릭하십시오.

오분류 비용을 사용하여 사전 확률 조정. 이 옵션을 사용하면 오분류 비용(비용 탭에 지정됨)에 기반하여 사전 확률을 조정할 수 있습니다. 이를 통해 투입 불순도 측도를 사용하는 트리의 트리 성장 프로세스로 비용 정보를 직접 통합할 수 있습니다. (이 옵션을 선택하지 않으면 비용 정보는 투입 측도에 기반하여 레코드를 분류하고 트리의 위험 추정값을 계산하는 데만 사용됩니다.)

바. CHAID 노드 - 비용

일부 컨텍스트에서는 특정 오차 유형이 다른 유형에 비해 더 값비쌉니다. 예를 들어, 고위험 신용 거래 신청자를 저위험(오차의 한 유형)으로 분류하는 것이 저위험 신청자를 고위험(다른 유형의 오차)으로 분류하는 것보다 비용이 더 나갈 수 있습니다. 오분류 비용을 통해 여러 다른 유형의 예측 오차의 상대적 중요도를 지정할 수 있습니다.

오분류 비용은 기본적으로 특정 결과에 적용된 가중값입니다. 이 가중값은 모델에 영향을 미치는 요인이 되어 실제로 예측변수를 변경합니다(값비싼 실수에 대비하는 보호의 한 방법으로).

C5.0 모델을 제외하고, 오분류 비용은 모델을 스코어링할 때 적용되지 않으며 자동 분류자 노드, 평가 노드 또는 분석 노드를 사용하여 모델을 순위화하거나 비교할 때 고려되지 않습니다. 비용을 포함한 모델은 포함하지 않은 모델보다 적은 오차를 생성하지 않고 전체 정확도 측면에서 더 높게 순위화되지 않을 수는 있지만, *보다 저렴* 오차를 선호하는 편향이 내재되어 있어서 실제적으로 성능이 더 좋을 수 있습니다.

비용 교차표는 예측 범주와 실제 범주의 가능한 각 조합의 비용을 표시합니다. 기본적으로 모든 오분류 비용은 1.0으로 설정됩니다. 사용자 정의 비용 값을 입력하려면 **오분류 비용 사용**을 선택하고 비용 교차표에 사용자 정의 값을 입력하십시오.

오분류 비용을 변경하려면 예측 및 실제 값의 원하는 조합에 해당하는 셀을 선택하고 셀의 기존 콘텐츠를 삭제한 후 셀의 원하는 비용을 입력하십시오. 비용은 자동으로 대칭되지 않습니다. 예를 들어, *A*를 *B*로 오분류한 비용을 2.0으로 설정할 경우 이 설정을 명시적으로 변경하지 않으면 *B*를 *A*로 오분류하는 비용의 기본값은 여전히 1.0입니다.

사. C&R 트리 노드 - 고급

고급 옵션을 통해 트리 작성 프로세스를 미세 조정할 수 있습니다.

불순도의 최소 변화. 불순도의 최소 변화를 지정하여 트리에서 새 분할을 작성하십시오. 불순도는 각 그룹 내 광범위한 출력 필드 값 범위를 보유하는 트리에서 정의된 부그룹의 범위를 말합니다. 범주형 목표의 경우 노드에 있는 케이스의 100%가 목표 필드의 특정 필드에 속하는 경우, 노드는 "순수"하다고 간주됩니다. 트리 작성의 목표는 유사한 출력 값을 포함하는 부그룹을 작성하는 것입니다(즉, 각 노드에서 불순도를 최소화함). 분기에 대한 최상의 분할이 지정된 수치 미만으로 불순도를 감소시키는 경우 분할은 수행되지 않습니다.

범주형 목표에 대한 불순도 측도. 범주형 목표 필드의 경우 트리 불순도를 측정하는 데 사용되는 방법을 지정합니다. (연속형 목표의 경우 이 옵션은 무시되고 가장 낮은 제곱 편차 불순도 측도가 항상 사용됩니다.)

- Gini는 분기에 대한 범주 소속 확률에 기반한 일반적인 불순도 측도입니다.
- 투잉은 이분형 분할을 강조하는 불순도 측도로, 분할에서 대략적으로 균등한 크기의 분기를 생성할 수 있습니다.
- 정렬은 순서 목표에만 적용 가능하므로 연속형 목표 클래스만 그룹화할 수 있다는 추가적인 제한조건을 추가합니다. 명목 목표에서 이 옵션을 선택한 경우 표준 투잉 측도가 기본적으로 사용됩니다.

과적합 방지 세트. 내부적으로 알고리즘은 레코드를 모델 작성 세트 및 과적합 방지 세트로 분할합니다. 이 세트는 해당 방법이 데이터에서 모델링 우연 변동을 일으키지 않도록 훈련 중에 오차를 추적하는 데 사용되는 데이터 레코드의 독립된 세트입니다. 레코드 퍼센트를 지정합니다. 기본값은 30입니다.

결과 복제. 난수 시드를 설정하면 분석을 복제할 수 있습니다. 정수를 지정하거나, 생성을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 유사 난수 정수를 작성합니다.

아. QUEST 노드 - 고급

고급 옵션을 통해 트리 작성 프로세스를 미세 조정할 수 있습니다.

분할 유의 수준 노드 분할에 대한 유의 수준(알파)을 지정합니다. 값은 0과 1 사이여야 합니다. 값이 낮을수록 노드 수가 적은 트리를 생성합니다.

과적합 방지 세트. 내부적으로 알고리즘은 레코드를 모델 작성 세트 및 과적합 방지 세트로 분할합니다. 이 세트는 해당 방법이 데이터에서 모델링 우연 변동을 일으키지 않도록 훈련 중에 오차를 추적하는 데 사용되는 데이터 레코드의 독립된 세트입니다. 레코드 퍼센트를 지정합니다. 기본값은 30입니다.

결과 복제. 난수 시드를 설정하면 분석을 복제할 수 있습니다. 정수를 지정하거나, 생성을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 유사 난수 정수를 작성합니다.

자. CHAID 노드 - 고급

고급 옵션을 통해 트리 작성 프로세스를 미세 조정할 수 있습니다.

분할 유의 수준 노드 분할에 대한 유의 수준(알파)을 지정합니다. 값은 0과 1 사이여야 합니다. 값이 낮을수록 노드 수가 적은 트리를 생성합니다.

병합 유의 수준. 범주 병합에 대한 유의 수준(알파)을 지정합니다. 값은 0보다 크고 1보다 작거나 같아야 합니다. 범주가 병합되지 않도록 하려면 1 값을 지정하십시오. 연속형 목표의 경우, 이는 최종 트리에서 변수에 대한 범주 수가 지정된 구간 수와 일치함을 의미합니다. 이 옵션은 Exhaustive CHAID에 사용할 수 없습니다.

Bonferroni 방법을 사용하여 유의수준 조정. 예측자의 다양한 범주 조합을 검정할 때 유의수준 값을 조정합니다. 값은 범주 수와 예측자의 측정 수준에 직접 관련되는 검정 수를 기반으로 조정됩니다. 이 방법은 일반적으로 거짓 양성 오차율을 더 효율적으로 제어하기 때문에 더 바람직합니다. 이 옵션을 사용하지 않도록 설정하면 참인 차이를 찾기 위해 분석 능력이 증가되지만 허위 긍정 비율이 증가될 수 있습니다. 특히 작은 표본에서는 이 옵션을 사용하지 않는 것이 좋습니다.

노드 내에서 병합된 범주의 재분할 허용. CHAID 알고리즘은 모델을 설명하는 가장 단순한 트리를 생성하기 위해 범주를 병합하려고 시도합니다. 이 옵션을 선택하면 더 나은 솔루션을 생성하는 경우 병합된 범주를 다시 분할할 수 있습니다.

범주형 목표에 대한 카이제곱. 범주형 목표의 경우, 카이제곱 통계량을 계산하기 위해 사용되는 방법을 지정할 수 있습니다.

- **Pearson.** 이 방법은 빠른 계산이 가능하지만 작은 표본에서는 주의하여 사용해야 합니다.
- **우도비.** 이 방법은 피어슨보다 강력하지만, 계산하는데 더 오래 걸립니다. 작은 표본의 경우 이 방법을 사용하는 것이 좋습니다. 연속형 목표인 경우 항상 이 방법을 사용합니다.

셀 기대빈도의 최소 변화량. 셀 빈도를 추정할 때(명목형 모델과 행 효과 순서 모델 둘 다에 대해), 반복 프로시저(엡실론)는 특정 분할에 대한 카이제곱 검정에 사용되는 최적 추정에 대한 수렴에 사용됩니다. 엡실론은 반복이 계속되기 위해 발생해야 하는 변화량을 판별합니다. 마지막 반복으로부터의 변화가 지정된 값보다 작은 경우 반복이 중지됩니다. 수렴되지 않는 알고리즘의 문 제점이 발생한 경우 이 값을 늘리거나 수렴이 발생할 때까지 최대 반복 수를 늘릴 수 있습니다.

수렴을 위한 최대 반복. 수렴 발생 여부에 관계없이, 중지 이전의 최대 반복 수를 지정합니다.

과적합 방지 세트. (이 옵션은 대화형 트리 작성기를 사용할 경우에만 사용 가능합니다.) 내부적으로 알고리즘은 레코드를 모델 작성 세트 및 과적합 방지 세트로 분할합니다. 이 세트는 해당 방법이 데이터에서 모델링 우연 변동을 일으키지 않도록 훈련 중에 오차를 추적하는 데 사용되는 데이터 레코드의 독립된 세트입니다. 레코드 퍼센트를 지정합니다. 기본값은 30입니다.

결과 복제. 난수 시드를 설정하면 분석을 복제할 수 있습니다. 정수를 지정하거나, 생성을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 유사 난수 정수를 작성합니다.

⑥ 의사결정 트리 노드 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 예측자 중요도 정보 및 플래그 목표의 원래 및 수정된 성향 스코어를 확보할 수도 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

모형 평가

예측자 중요도 계산. 적절한 중요도 측도를 생성하는 모델의 경우 모델 측정 시 각 예측자의 상대적 중요도를 나타내는 차트를 표시할 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 예측자 중요도는 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 몇몇 모델은 기본적으로 해제되어 있음에 유의하십시오. 예측자 중요도는 의사결정 목록 모델에 사용할 수 없습니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

성향 스코어

성향 스코어는 모델링 노드 및 모델 너깃의 설정 탭에서 사용 가능합니다. 이 기능은 선택한 목표가 플래그 필드인 경우에만 사용 가능합니다. 자세한 정보는 성향 스코어의 내용을 참조하십시오.

원시 성향 스코어 계산. 원시 성향 스코어는 학습 데이터에만 기반하여 모델에서 파생됩니다. 모델이 참 값(응답함)을 예측하면 성향은 P와 동일합니다. 여기서 P는 예측 확률입니다. 모델이 거짓 값을 예측하면 성향은 $(1 - P)$ 로 계산됩니다.

- 모델 작성 시 이 옵션을 선택한 경우 기본적으로 모델 너깃에서 성향 스코어가 사용 가능합니다. 그러나 모델링 노드에서 선택 여부에 상관없이 언제나 모델 너깃에서 원시 성향 스코어를 사용하도록 선택할 수 있습니다.
- 모델 스코어링 시 원시 성향 스코어는 표준 접두문자에 문자 *RP*가 추가된 필드에 추가됩니다. 예를 들어 예측이 이름이 *\$R-churn*인 필드에 있는 경우 성향 스코어 필드 이름은 *\$RRP-churn*입니다.

수정된 성향 스코어 계산. 원시 성향은 모델에서 제공된 추정값에만 기반하며, 과적합할 경우 성향의 지나친 낙관적 추정값으로 이어질 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 수행 방법을 보고 적절히 더 나은 추정값을 제공하도록 성향을 조정하여 보완하려고 합니다.

- 이 설정에서는 유효한 파티션 필드가 스트림에 존재해야 합니다.
- 원시 신뢰도 스코어와 달리, 수정된 성향 스코어는 모델 작성 시 계산해야 합니다. 그렇지 않으면 모델 너짓 스코어링에서 사용 불가능합니다.
- 모델 스코어링 시 수정된 성향 스코어는 표준 접두문자에 문자 *AP*가 추가된 필드에 추가됩니다. 예를 들어 예측이 이름이 *\$R-churn*인 필드에 있는 경우 성향 스코어 필드 이름은 *\$RAP-churn*입니다. 수정된 성향 스코어는 로지스틱 회귀분석 모델에서 사용할 수 없습니다.
- 수정된 성향 스코어를 계산할 때 계산에 사용된 검정 또는 검증 파티션은 균형을 맞출 수 없습니다. 이를 방지하려면 업스트림 균형 노드에서 균형 학습 데이터만 옵션을 선택해야 합니다. 또한 복잡한 샘플에서 업스트림을 사용하는 경우 이는 수정된 성향 스코어를 무효화합니다.
- 수정된 성향 스코어는 "증폭된" 트리 및 규칙 세트 모델에서 사용할 수 없습니다. 자세한 정보는 부스팅 C5.0 모델의 내용을 참조하십시오.

다음은 기준. 수정된 성향 스코어를 계산할 경우 파티션 필드가 스트림에 존재해야 합니다. 이 계산에서 검정 또는 검증 파티션 중 사용할 항목을 지정할 수 있습니다. 최상의 결과를 얻으려면 검정 또는 검증 파티션은 원래 모델을 학습시키는 데 사용되는 파티션만큼 많은 레코드를 최소한으로 포함해야 합니다.

(5) C5.0 노드

이 기능은 SPSS® Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

이 노드는 C5.0 알고리즘을 사용하여 **의사결정 트리** 또는 **규칙 세트**를 작성합니다. C5.0 모델은 최대 **정보 이득**을 제공하는 필드를 기준으로 하여 표본을 분할하는 방식으로 작동합니다. 첫 번째 분할을 통해 정의된 각 부표본이 일반적으로 다른 필드를 기준으로 하여 다시 분할되고, 부표본을 더 이상 분할할 수 없게 될 때까지 프로세스가 반복됩니다. 마지막으로 최저 수준의 분할을 재검토해서 모델 값에 상당히 기여하지 않는 분할은 제거 또는 **가지치기**됩니다.

참고: C5.0 노드는 범주형 목표만 예측할 수 있습니다. 범주형(명목 또는 순서) 필드가 있는 데이터 분석하는 경우 노드는 릴리스 11.0 이전의 C5.0 버전보다 범주를 그룹화할 가능성이 있습니다.

C5.0는 두 종류의 모델을 생성할 수 있습니다. **의사결정 트리**는 알고리즘이 찾는 분할을 직선적으로 설명합니다. 각 터미널(또는 "리프") 노드는 학습 데이터의 특정 서브세트를 설명하고 학습 데이터의 각 케이스는 트리의 정확히 한 터미널 노드에 속합니다. 즉, 의사결정 트리에 표시된 특정 데이터 레코드에 정확히 하나의 예측이 가능합니다.

이와 반대로, **규칙 세트**는 개별 레코드를 예측하려 시도하는 규칙 세트입니다. 규칙 세트는 의사결정 트리에서 파생되며 어느 정도는 의사결정 트리에 있는 정보의 단순화된 또는 엄선된 버전을 나타냅니다. 규칙 세트는 종종 전체 의사결정 트리(단, 보다 덜 복잡한 모델 포함)에서 대부분의 중요한 정보를 보유할 수 있습니다. 규칙 세트는 작동 방식으로 인해 의사결정 트리와 특

성이 동일하지 않습니다. 가장 중요한 차이는 규칙 세트의 경우 특정 레코드에 둘 이상의 규칙이 적용되거나 규칙이 전혀 적용되지 않을 수도 있다는 점입니다. 여러 규칙이 적용되는 경우 각 규칙은 규칙과 연관된 신뢰도를 기준으로 하여 가중된 "투표"를 얻고, 논의되는 레코드에 적용되는 모든 규칙의 가중된 투표를 조합해서 최종 예측이 결정됩니다. 적용된 규칙이 없으면 기본 예측이 레코드에 지정됩니다.

예. 한 의료 연구원은 모두 동일한 질병을 앓고 있는 일련의 환자에 대한 데이터를 수집해왔습니다. 치료 과정 중에 각 환자는 다섯 가지 약물 치료 중 하나에 반응했습니다. C5.0 모델을 다른 노드와 함께 사용하여 동일한 질병을 앓는 미래의 환자에게 어느 약품이 적합한지 찾을 수 있습니다.

요구사항. C5.0 모델을 학습시키려면 범주형(즉, 명목 또는 순서) 목표 필드 하나와 임의의 유형의 입력 필드 하나 이상이 있어야 합니다. 둘 다 또는 없음으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다. 가중 필드도 지정할 수 있습니다.

강도. C5.0 모델은 데이터 누락이나 많은 수의 입력 필드와 같은 문제가 발생할 때 상당히 강건합니다. 일반적으로 추정하기 위해 긴 학습 시간이 필요하지 않습니다. 또한 C5.0 모델은 모델에서 파생된 규칙의 해석이 매우 직설적이어서 다른 모델 유형보다 이해하기 쉽습니다. C5.0은 분류 정확도를 높이는 강력한 부스팅 방법도 제공합니다.

 **참고:** 병렬 처리를 사용할 경우 C5.0 모델 작성 속도가 개선될 수 있습니다.

① C5.0 노드 모델 옵션

이 기능은 SPSS® Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

모델 이름. 생성할 모델의 이름을 지정합니다.

- **자동.** 이 옵션이 선택되면 대상 필드 이름을 기준으로 하여 모델 이름이 자동으로 생성됩니다. 이는 기본값입니다.
- **사용자 정의.** 이 노드가 작성할 모델 너깃에 직접 이름을 지정하려면 이 옵션을 선택하십시오.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

출력 유형. 결과적인 모델 너깃 유형을 **의사결정 트리** 또는 **규칙 세트**로 할지 여부를 지정합니다.

그룹 기호. 이 옵션이 선택되면 C5.0은 출력 필드에 대해 패턴이 유사한 기호 값을 결합하려 시도합니다. 이 옵션을 선택하지 않을 경우 C5.0은 상위 노드를 분할하는 데 사용된 기호 필드의 모든 값에 대한 하위 노드를 작성합니다. 예를 들어, C5.0은 COLOR 필드(값은 RED, GREEN, BLUE)에서 분할할 경우 기본적으로 세 가지 분할을 작성합니다. 하지만 이 옵션이 선택되고 COLOR = RED인 레코드가 COLOR = BLUE인 레코드와 매우 유사한 경우에는 두 개의 분할 즉, 한 그룹에 GREEN을 작성하고 다른 그룹에는 BLUE 및 RED를 함께 작성합니다.

부스팅 사용. C5.0 알고리즘에는 정확도 비율을 개선하기 위한 부스팅이라는 특수 방법이 있습니다. 여러 모델을 한 시퀀스에 작성하는 방식으로 작동합니다. 첫 번째 모델은 일반적인 방법으로 작성됩니다. 그런 다음 두 번째 모델은 첫 번째 모델이 잘못 분류한 레코드에 초점을 맞추는 방식으로 작성됩니다. 그리고 나서 세 번째 모델은 두 번째 모델의 오류에 초점을 맞추기 위해 작성됩니다. 그 다음도 마찬가지입니다. 마지막으로 전체 모델 세트를 케이스에 적용하고 가중 투표 프로시저를 사용하여 개별 예측을 하나의 전체 예측으로 결합해서 케이스를 분류합니다. 부스팅은 C5.0 모델의 정확도를 상당히 개선할 수 있지만 더 오래 학습해야 합니다. **시행 수** 옵션으로 부스팅 모델에 사용되는 모델 수를 제어할 수 있습니다.

교차 검증. 이 옵션이 선택되면 C5.0은 학습 데이터의 서브세트에 작성된 모델 세트를 사용하여 전체 데이터 세트에 작성된 모델의 정확도를 추정합니다. 이 옵션은 데이터 세트가 너무 작아서 일반 학습 및 검정 세트로 분할할 수 없는 경우에 유용합니다. 교차 검증 모델은 정확도 추정값이 계산되고 나면 삭제됩니다. **중첩 수** 또는 교차 검증에 사용되는 모델 수를 지정할 수 있습니다. IBM® SPSS Modeler의 이전 버전에서는 모델의 작성 및 교차 검증이 두 개의 개별 작업이었음에 유의하십시오. 현재 버전은 개별 모델 작성 단계가 필요하지 않습니다. 모델 작성 및 교차 검증이 동시에 수행됩니다.

모드. 단순 학습의 경우 대부분의 C5.0 모수가 자동으로 설정됩니다. 고급 학습에서는 학습 모수를 보다 직접적으로 제어할 수 있습니다.

단순 모드 옵션

선호. 기본적으로 C5.0은 가능한 가장 정확한 트리를 생성하려 시도합니다. 일부 인스턴스에서는 이 모델이 새 데이터에 적용될 때 성능을 저하시킬 수 있는 과적합을 유발할 수 있습니다. 이 문제의 영향을 덜 받는 알고리즘 설정을 사용하려면 **범용성**을 선택하십시오.

참고: **Generality** 옵션을 선택한 채 작성된 모델이 다른 모델보다 일반화된다고 보장되지는 않습니다. 범용성이 중요 문제이면 항상 남겨진 검정 표본에 대해 모델을 검증하십시오.

예상 잡음(%). 학습 세트에서 불량 또는 오류 데이터의 예상 비율을 지정합니다.

고급 모드 옵션

가지치기 심각도. 의사결정 트리 또는 규칙 세트를 가지치기할 범위를 판별합니다. 더 작고 보다 간결한 트리를 원하는 경우 이 값을 늘리십시오. 보다 정확한 트리를 원하면 값을 줄이십시오. 이 설정은 로컬 가지치기에만 영향을 미칩니다(아래의 "글로벌 가지치기 사용" 참조).

하위 분기별 최소 레코드. 하위 그룹의 크기를 사용하여 트리 분기의 분할 수를 제한할 수 있습니다. 트리의 분기는 결과적인 하위 분기 중 둘 이상에 학습 세트에서 최소 이 수만큼의 레코드가 포함된 경우에만 분할됩니다. 기본값은 2입니다. 불량 데이터의 **과도한 학습**을 방지하려면 이 값을 늘리십시오.

글로벌 가지치기 사용. 트리는 두 단계로 가지치기됩니다. 첫 번째는 로컬 가지치기 단계로, 하위 트리를 검토하고 모델의 정확도를 높이기 위해 분기를 접습니다. 두 번째인 글로벌 가지치기 단계는 트리를 전체적으로 고려합니다. 약한 하위 트리가 접힐 수 있습니다. 기본적으로 글로벌 가지치기가 수행됩니다. 글로벌 가지치기 단계를 생략하려면 이 옵션을 선택 취소하십시오.

필드유용성 사전조사. 이 옵션을 선택하면 C5.0이 모델 작성을 시작하기 전에 예측변수의 유용성을 검토합니다. 관련이 없는 것으로 밝혀진 예측변수는 모델 작성 프로세스에서 제외됩니다. 이 옵션은 많은 예측변수 필드가 있는 모델에 유용할 수 있으며 과적합을 차단하는 데 도움이 됩니다.

 **참고:** 병렬 처리를 사용할 경우 C5.0 모델 작성 속도가 개선될 수 있습니다.

(6) Tree-AS 노드

Tree-AS 노드는 분산 환경의 데이터와 함께 사용할 수 있습니다. 이 노드에서는 CHAID 또는 Exhaustive CHAID 모델을 사용하여 의사결정 트리를 작성할 수도 있습니다.

CHAID 또는 카이제곱 자동 상호작용 발견은 카이제곱 통계량을 사용하여 최적의 분할을 식별해서 의사결정 트리를 작성하기 위한 분류 방법입니다.

먼저 CHAID는 각 입력 필드와 출력 사이의 교차 분석표를 탐색하고 카이제곱 독립 검정을 사용하여 유의수준을 검정합니다. 둘 이상의 관계가 통계적으로 유의적이면 CHAID는 가장 유의적인 (최소 p 값) 입력 필드를 선택합니다. 입력에 둘 이상의 범주가 있는 경우에는 이 범주를 비교하고 결과에 차이가 없는 범주는 함께 접습니다. 최소유의차를 표시하는 범주 쌍을 연속으로 결합해서 이를 수행합니다. 나머지 모든 범주가 지정된 검정 수준에서 서로 다르다면 이 범주 병합 프로세스는 중지됩니다. 명목 입력 필드의 경우 범주가 병합될 수 있으며 순서 세트의 경우에는 연속형 범주만 병합될 수 있습니다.

Exhaustive CHAID는 각 예측자에 대한 모든 가능한 분할을 탐색하는 보다 전반적인 작업을 수행하지만 계산 시간이 오래 걸리는 CHAID의 수정 모델입니다.

요구사항. 목표 및 입력 필드는 연속형 또는 범주형이 가능하고 노드는 각 수준에서 둘 이상의 하위 그룹으로 분할될 수 있습니다. 모델에 사용된 순서 필드에 숫자 저장 공간(문자열이 아님)이 있어야 합니다. 필요한 경우 재분류 노드를 사용하여 변환하십시오.

강도. CHAID에서는 비이분형 트리를 생성하여, 일부 분할이 세 개 이상의 분기를 포함할 수 있음을 의미합니다. 따라서 이 노드는 이분형 성장 방법보다 광범위한 트리를 작성하는 경향이 있습니다. CHAID는 모든 유형의 입력에 작용하며 케이스 가중치 및 빈도 변수를 모두 허용합니다.

① Tree-AS 노드 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드를 지정할 수 있습니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 역할 설정(목표, 예측자 등)을 사용합니다.

사용자 정의 필드 할당 사용: 수동으로 대상, 예측자 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

필드. 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

목록의 모든 필드를 선택하려면 **모두** 단추를 클릭하거나 개별 측정 수준 단추를 클릭하여 이 측정 수준의 모든 필드를 선택하십시오.

목표: 예측에 대한 목표로 하나의 필드를 선택합니다.

예측자 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

분석 가중값 케이스 가중치로 필드를 사용하려면 여기에서 필드를 지정합니다. 케이스 가중치는 출력 필드의 수준에서 분산의 차이를 설명하는 데 사용됩니다. 추가 정보는 빈도 및 가중 필드 사용의 내용을 참조하십시오.

② Tree-AS 노드 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, **실행** 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

탭은 모델에 특정한 사용자 정의를 설정하는 여러 창을 포함합니다.

가. Tree-AS 노드 - 기본

의사결정 트리 작성 방법에 대한 기본 옵션을 지정하십시오.

트리 성장 알고리즘 사용하려는 CHAID 알고리즘 유형을 선택합니다. Exhaustive CHAID는 각 예측자에 대한 모든 가능한 분할을 탐색하는 보다 전반적인 작업을 수행하지만 계산 시간이 오래 걸리는 CHAID의 수정 모델입니다.

최대 트리 깊이 루트 노드 아래의 최대 수준 수를 지정합니다(표본이 반복적으로 분할된 횟수). 기본값은 5입니다. 최대 수준(노드라고도 함)은 50,000입니다.

구간화 연속 데이터를 사용하는 경우 입력을 구간화해야 합니다. 선행 노드에서 이를 수행할 수 있습니다. 그러나 Tree-AS 노드는 연속 입력을 자동으로 구간화합니다. Tree-AS 노드를 사용하여 데이터를 자동으로 구간화하는 경우 입력을 구분할 **구간 수**를 선택합니다. 데이터는 동일한 빈도로 구간으로 구분됩니다. 사용 가능한 옵션은 2, 4, 5, 10, 20, 25, 50 또는 100입니다.

나. Tree-AS 노드 - 성장

성장 옵션을 사용하여 트리 작성 프로세스를 미세 조정하십시오.

P-값에서 효과 크기로 전환하기 위한 레코드 임계값 트리 작성 시 모델이 P-값 설정 사용을 유효 크기 설정으로 전환하는 레코드 수를 지정합니다. 기본값은 1,000,000입니다.

분할 유의 수준 노드 분할에 대한 유의 수준(알파)을 지정합니다. 값은 0.01과 0.99 사이여야 합니다. 값이 낮을수록 노드 수가 적은 트리를 생성합니다.

병합 유의 수준 범주 병합에 대한 유의 수준(알파)을 지정합니다. 값은 0.01과 0.99 사이여야 합니다. 이 옵션은 Exhaustive CHAID에 사용할 수 없습니다.

Bonferroni 방법을 사용하여 유의수준 값 조정 예측자의 다양한 범주 조합을 검정할 때 유의수준 값을 조정합니다. 값은 범주 수와 예측자의 측정 수준에 직접 관련되는 검정 수를 기반으로 조정됩니다. 이 방법은 일반적으로 거짓 양성 오차율을 더 효율적으로 제어하기 때문에 더 바람직합니다. 이 옵션을 사용하지 않으면 참의 차이를 찾는 분석 기능이 향상되지만, 거짓 양성 비율이 늘어납니다. 특히 작은 표본에서는 이 옵션을 사용하지 않는 것이 좋습니다.

효과 크기 임계값(연속형 목표만) 연속형 목표 사용 시 노드를 분할하고 범주를 병합할 때 사용할 효과 크기 임계값을 설정합니다. 값은 0.01과 0.99 사이에 있어야 합니다.

효과 크기 임계값(범주형 목표만) 범주형 목표 사용 시 노드를 분할하고 범주를 병합할 때 사용할 효과 크기 임계값을 설정합니다. 값은 0.01과 0.99 사이에 있어야 합니다.

노드 내에서 병합된 범주 재분할 허용 CHAID 알고리즘은 모델을 설명하는 가장 단순한 트리를 생성하기 위해 범주를 병합하려고 합니다. 이 옵션을 선택하면 더 나은 솔루션을 생성하는 경우 병합된 범주를 다시 분할할 수 있습니다.

리프 노드 그룹화에 대한 유의 수준 리프 노드 그룹을 형성하는 방법 또는 특이한 리프 노드를 식별하는 방법을 판별하는 유의 수준을 지정합니다.

범주형 목표에 대한 카이제곱 범주형 목표의 경우 카이제곱 통계량을 계산하는 데 사용되는 방법을 지정할 수 있습니다.

- Pearson 이 방법은 빠른 계산이 가능하지만 작은 표본에서는 주의하여 사용해야 합니다.
- 우도비 이 방법은 Pearson보다 강력하지만, 계산하는 데 더 오래 걸립니다. 작은 표본의 경우 이 방법을 사용하는 것이 좋습니다. 연속형 목표인 경우 항상 이 방법을 사용합니다.

다. Tree-AS 노드 - 중지 규칙

이 옵션은 트리 구성 방법을 제어합니다. 중지 규칙은 트리의 분할 특정 분기를 중지하는 시점을 판별합니다. 매우 작은 하위 그룹을 작성하는 분할을 방지하도록 최소 분기 크기를 설정합니다. 부모마디 최소 레코드 수는 분할할 노드(상위)의 레코드 수가 지정된 값 미만인 경우 분할을 방지합니다. 자식마디 최소 레코드 수는 분할로 작성된 분기(하위)의 레코드 수가 지정된 값 미만인 경우 분할을 방지합니다.

- 퍼센트 사용 전체 훈련 데이터의 퍼센트 관점에서 크기를 지정합니다.
- 절대값 사용 레코드의 절대값으로 크기를 지정합니다.

셀 기대빈도의 최소 변화 셀 빈도를 추정할 때(명목 모델 및 행 효과 순서 모델 모두에서) 대체 프로시저(엡실론)를 사용하여 특정 분할에 대한 카이제곱 검정에 사용된 최적의 추정값으로 수렴합니다. 엡실론은 반복이 계속되기 위해 발생해야 하는 변화량을 판별합니다. 마지막 반복으로부터의 변화가 지정된 값보다 작은 경우 반복이 중지됩니다. 수렴되지 않는 알고리즘의 문제점이 발생한 경우 이 값을 늘리거나 수렴이 발생할 때까지 최대 반복 수를 늘릴 수 있습니다.

수렴을 위한 최대 반복 수렴이 발생하는지에 상관없이 중지하기 전에 최대 반복 수를 지정합니다.

라. Tree-AS 노드 - 비용

일부 컨텍스트에서는 특정 오차 유형이 다른 유형에 비해 더 값비쌉니다. 예를 들어, 고위험 신용 거래 신청자를 저위험(오차의 한 유형)으로 분류하는 것이 저위험 신청자를 고위험(다른 유형의 오차)으로 분류하는 것보다 비용이 더 나갈 수 있습니다. 오분류 비용을 통해 여러 다른 유형의 예측 오차의 상대적 중요도를 지정할 수 있습니다.

오분류 비용은 기본적으로 특정 결과에 적용된 가중값입니다. 이 가중값은 모델에 영향을 미치는 요인이 되어 실제로 예측변수를 변경합니다(값비싼 실수에 대비하는 보호의 한 방법으로).

비용을 포함하는 모델은 그렇지 않은 항목보다 적은 오류를 생성하지 않으며, 전반적인 정확도 면에서 순위가 더 높지 않을 수도 있지만, 비용이 더 적게 드는 오류를 위해 기본 성향을 가지고 있으므로 실질적인 면에서 성능이 더 좋습니다.

비용 교차표는 예측 범주와 실제 범주의 가능한 각 조합의 비용을 표시합니다. 기본적으로 모든 오분류 비용은 1.0으로 설정됩니다. 사용자 정의 비용 값을 입력하려면 **오분류 비용 사용**을 선택하고 비용 교차표에 사용자 정의 값을 입력하십시오.

오분류 비용을 변경하려면 예측 및 실제 값의 원하는 조합에 해당하는 셀을 선택하고 셀의 기존 콘텐츠를 삭제한 후 셀의 원하는 비용을 입력하십시오. 비용은 자동으로 대칭되지 않습니다. 예를 들어, A를 B로 오분류한 비용을 2.0으로 설정할 경우 이 설정을 명시적으로 변경하지 않으면 B를 A로 오분류하는 비용의 기본값은 여전히 1.0입니다.

순서 목표의 경우에만 **순서 목표의 기본 비용 증가**를 선택하고 비용 행렬에서 기본값을 설정할 수 있습니다. 사용 가능한 옵션은 다음 목록에서 설명합니다.

- **증가 없음** - 올바른 모든 예측에 대한 기본값, 1.0.
- **선형** - 연속된 잘못된 각 예측은 비용을 1씩 증가시킵니다.
- **제곱** - 연속된 잘못된 각 예측은 선형 값의 제곱입니다. 이 경우 값은 1, 4, 9 등과 같습니다.
- **사용자 정의** - 테이블에서 값을 수동으로 편집하면 드롭 다운 옵션이 자동으로 **사용자 정의**로 변경됩니다. 드롭 다운 선택을 다른 옵션으로 변경하면 편집된 값을 선택한 옵션의 값으로 바꿉니다.

③ Tree-AS 노드 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 모델 스코어링 중에 신뢰도 값을 계산하고 식별 ID를 추가할 수도 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

신뢰도 계산 모델 스코어링 중에 신뢰도 필드를 추가하려면 이 선택란을 선택합니다.

규칙 식별자 레코드가 지정된 리프 노드의 ID를 포함하는 모델의 스코어링 중에 필드를 추가하려면 이 선택란을 선택합니다.

④ Tree-AS 모델 너깃

가. Tree-AS 모델 너깃 출력

Tree-AS 모델을 작성한 후 출력 뷰어에서 다음 정보를 사용할 수 있습니다.

모델 정보 테이블

모델 정보 테이블에서는 모델에 대한 주요 정보를 제공합니다. 테이블은 다음과 같은 일부 상위 수준 모델 설정을 식별합니다.

- 사용된 알고리즘 유형(CHAID 또는 Exhaustive CHAID).
- 유형 노드 또는 Tree-AS 노드 필드 탭에서 선택된 목표 필드 이름.
- 유형 노드 또는 Tree-AS 노드 필드 탭에서 예측자로 선택된 필드 이름.
- 데이터에 있는 레코드 수. 빈도 가중치로 모델을 작성하는 경우, 이 값은 가중된 유효한 개수가 되며 트리의 기반이 되는 레코드 수를 나타냅니다.
- 생성된 트리에 있는 리프 노드 수.
- 트리에서 수준 수(즉, 트리 깊이).

예측변수 중요도

예측자 중요도 그래프는 막대형 차트로 모델에 있는 상위 10개 입력(예측자)의 중요도를 표시합니다.

차트에 필드가 10개가 넘으면 차트 아래 슬라이더를 사용하여 차트에 포함되는 예측자의 선택을 변경할 수 있습니다. 슬라이더의 표시기 마크는 고정된 너비이며, 슬라이더의 각 마크는 10개 필드를 나타냅니다. 슬라이더와 함께 표시기 마크를 이동하여 예측자 중요도로 정렬된 다음 또는 이전 10개 필드를 표시할 수 있습니다.

차트를 두 번 클릭하면 그래프 설정을 편집할 수 있는 별도의 대화 상자를 열 수 있습니다. 예를 들어, 그래프 크기, 사용된 글꼴의 크기와 색상과 같은 항목을 수정할 수 있습니다. 별도의 이 편집 대화 상자를 닫으면 출력 탭에 표시된 차트에 변경이 적용됩니다.

상위 의사결정 규칙 테이블

기본적으로 이 대화형 테이블은 리프 노드 내 포함된 총 레코드의 퍼센트에 기반하여 출력에서 상위 5개 리프 노드에 대한 규칙 통계를 표시합니다.

테이블을 두 번 클릭하면 테이블에 표시된 규칙 정보를 편집할 수 있는 별도의 대화 상자를 열 수 있습니다. 표시되는 정보와 대화 상자에서 사용 가능한 옵션은 목표의 데이터 유형(예: 범주형 또는 연속형)에 따라 달라집니다.

다음 규칙 정보가 테이블에 표시됩니다.

- 규칙 ID
- 규칙이 적용되고 구성되는 방법에 대한 세부사항
- 각 규칙의 레코드 개수. 빈도 가중치로 모델을 작성하는 경우, 이 값은 가중된 유효한 개수가 되며 트리의 기반이 되는 레코드 수를 나타냅니다.
- 각 규칙에서 레코드 퍼센트

또한 연속형 목표의 경우 테이블의 추가 열은 각 규칙에 대한 평균값을 표시합니다.

다음 테이블 콘텐츠 옵션을 사용하여 규칙 테이블 레이아웃을 변경할 수 있습니다.

- 상위 의사결정 규칙 상위 5개 의사결정 규칙은 리프 노드 내 포함된 총 레코드의 퍼센트로 정렬됩니다.
- 모든 규칙 테이블에는 모델에서 생성한 모든 리프 노드가 포함되지만 페이지당 20개의 규칙만 표시합니다. 이 레이아웃을 선택하면 ID로 규칙 찾기 및 페이지의 추가 옵션을 사용하여 규칙을 검색할 수 있습니다.

또한 범주형 목표인 경우 범주별 상위 규칙 옵션을 사용하여 규칙 테이블 레이아웃을 대체할 수 있습니다. 상위 5개 의사결정 규칙은 사용자가 선택한 목표 범주에 대한 총 레코드의 퍼센트로 정렬됩니다.

규칙 테이블의 레이아웃을 변경하는 경우 대화 상자 왼쪽 상단에 있는 뷰어로 복사 단추를 클릭하여 수정된 규칙 테이블을 출력 뷰어로 다시 복사할 수 있습니다.

나. Tree-AS 모델 너깃 설정

Tree-AS 모델 너깃의 설정 탭에서 모델 스코어링 중 신뢰도 및 SQL 생성에 대한 옵션을 지정합니다. 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능합니다.

신뢰도 계산 스코어링 작업에 신뢰도를 포함하려면 이 선택란을 선택합니다. 데이터베이스에서 모델 스코어를 계산할 때 신뢰도를 제외하면 보다 효율적인 SQL을 생성할 수 있음을 의미합니다. 회귀 트리에서는 신뢰도를 지정하지 않습니다.

규칙 식별자 각 레코드가 지정된 터미널 노드의 ID를 표시하는 스코어링 출력에서 필드를 추가하려면 이 선택란을 선택합니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

(7) 랜덤 트리 노드

랜덤 트리 노드는 분산 환경의 데이터와 함께 사용할 수 있습니다. 이 노드에 다중 의사결정 트리로 구성된 앙상블 모형을 작성하십시오.

랜덤 트리 노드는 분류 및 회귀분석 트리를 토대로 작성된 트리 기반의 분류 및 예측 방법입니다. C&R 트리와 마찬가지로, 이 예측 방법은 재귀적 파티셔닝을 사용하여 학습 레코드를 출력 필드 값이 유사한 세그먼트로 분할합니다. 이 노드는 먼저 분할로 인한 불순도 지수를 줄여서 측정되는 최상의 분할을 찾기 위해 사용 가능한 입력 필드를 검토합니다. 그런 다음 분할이 두 개의 하위 그룹을 정의하고, 중지 기준 중 하나가 트리거될 때까지 각 그룹은 계속해서 두 개의 하위 그룹으로 추가 분할되는 식입니다. 모든 분할은 이분형(하위 그룹을 두 개만)입니다.

랜덤 트리 노드는 복원 붓스트랩 표본추출을 사용하여 표본 데이터를 생성합니다. 표본 데이터를 사용하여 트리 모델이 성장합니다. 트리 성장 중에는 랜덤 트리에서 데이터를 다시 표본 추출하지 않습니다. 대신 예측변수의 일부를 무작위로 선택하고 최적의 예측변수를 사용하여 트리 노드를 분할합니다. 이 프로세스는 각 트리 노드를 분할할 때 반복됩니다. 이것이 랜덤 포리스트에서 트리가 성장하는 기본 개념입니다.

랜덤 트리에서는 C&R 트리와 유사한 트리를 사용합니다. 이러한 트리는 이분형이므로 각 필드를 분할하면 두 개의 분기가 생성됩니다. 범주가 여러 개인 범주형 필드의 경우 내부 분할 기준에 따라 범주가 두 개의 그룹으로 그룹화됩니다. 각 트리는 최대 범위까지 성장합니다(가지치기 없음). 스코어링 시 랜덤 트리에서는 다수 투표(분류용) 또는 평균(회귀분석용)에 따라 개별 트리 스코어를 결합합니다.

랜덤 트리와 C&R 트리의 차이점은 다음과 같습니다.

- 랜덤 트리 노드에서는 지정된 수의 예측변수를 무작위로 선택하고 선택사항 중 최적의 예측변수를 사용하여 노드를 분할합니다. 반대로, C&R 트리에서는 모든 예측변수 중에서 최적의 예측변수를 찾습니다.
- 일반적으로 각 리프 노드에 단일 레코드가 포함될 때까지 랜덤 트리의 각 트리가 완전히 성장합니다. 따라서 트리 깊이가 매우 커질 수 있습니다. 그러나 표준 C&R 트리에서는 트리 성장에 여러 다른 정지규칙을 사용하므로, 일반적으로 훨씬 깊이가 낮은 트리가 생성됩니다.

랜덤 트리는 C&R 트리와 비교했을 때 두 개의 기능이 추가되었습니다.

- 첫 번째 기능은 원래 데이터 세트에서 복원 표본추출하여 학습 데이터 세트의 복제본을 작성하는 *배깅*입니다. 이 동작을 수행하면 원래 데이터 세트와 동일한 크기의 붓스트랩 표본이 작성된 다음 *구성요소 모델*이 각 복제본에 작성됩니다. 이 구성요소 모델들은 함께 앙상블 모형을 형성합니다.
- 두 번째 기능은 트리의 각 분할에서 불순도 측도에 대해 입력 필드의 표본추출만 고려하는 것입니다.

요구사항. 랜덤 트리 모델을 학습하려면 하나 이상의 입력 필드와 하나의 *대상* 필드가 필요합니다. 목표 및 입력 필드는 연속형(수치 범위) 또는 범주형이 가능합니다. *모두* 또는 *없음*으로 설정되는 필드는 무시됩니다. 모델에 사용된 필드의 유형은 완전히 인스턴스화되어 있어야 하고, 모델에 사용된 순서(정렬된 세트) 필드에는 수치 저장 공간(문자열이 아닌)이 있어야 합니다. 필요한 경우 재분류 노드를 사용하여 변환할 수 있습니다.

강도. 랜덤 트리 모델은 대형 데이터 세트 및 많은 수의 필드를 처리할 때 강력합니다. 또한 배경 및 필드 표본추출 사용으로 인해 과적합이 발생할 가능성이 훨씬 줄어들어 새 데이터를 사용할 때 검정에 표시되는 결과가 반복될 가능성이 더 높아집니다.

① 랜덤 트리 노드 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드를 지정할 수 있습니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 역할 설정(목표, 예측자 등)을 사용합니다.

사용자 정의 필드 할당 사용: 수동으로 대상, 예측자 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

필드. 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

목록의 모든 필드를 선택하려면 **모두** 단추를 클릭하거나 개별 측정 수준 단추를 클릭하여 이 측정 수준의 모든 필드를 선택하십시오.

목표: 예측에 대한 목표로 하나의 필드를 선택합니다.

예측자 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

분석 가중값 케이스 가중치로 필드를 사용하려면 여기에서 필드를 지정합니다. 케이스 가중치는 출력 필드의 수준에서 분산의 차이를 설명하는 데 사용됩니다. 추가 정보는 빈도 및 가중 필드 사용의 내용을 참조하십시오.

② 랜덤 트리 노드 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, **실행** 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

탭은 모델에 특정한 사용자 정의를 설정하는 여러 창을 포함합니다.

가. 랜덤 트리 노드 - 기본

의사결정 트리 작성 방법의 기본 옵션을 지정하십시오.

작성할 모델 수. 노드가 작성할 수 있는 최대 트리 수를 지정하십시오.

표본 크기 기본적으로 붓스트랩 표본의 크기는 원본 학습 데이터와 동일합니다. 대형 데이터 세트를 처리하는 경우 표본 크기를 줄이면 성능이 높아질 수 있습니다. 0 - 1의 비율입니다. 예를 들어 표본 크기를 0.6으로 설정하여 원래 학습 데이터 크기의 60%로 줄이십시오.

불균형한 데이터 처리. 모델의 대상이 플래그 결과이고(예: 구매 또는 구매 안 함) 원하는 결과에 원하지 않는 결과의 비율이 매우 작으면 데이터의 균형이 맞지 않고 모델이 수행하는 붓스트랩 표본추출이 모델 정확도에 영향을 줄 수 있습니다. 정확도를 향상시키려면 이 선택란을 선택하십시오. 그러면 모델이 원하는 결과의 더 많은 부분을 캡처하고 더 나은 모델을 생성합니다.

변수 선택에 가중된 표본추출 사용. 기본적으로 각 리프 노드의 변수는 동일한 확률로 임의 선택됩니다. 가중치를 변수에 적용하고 선택 프로세스를 향상시키려면 이 선택란을 선택하십시오. 가중치는 랜덤 트리 노드 자체에서 계산합니다. 중요한 필드(가중치가 높음)일수록 예측자로 선택될 가능성이 큼니다.

최대 노드 수. 개별 트리에 허용되는 리프 노드의 최대 수를 지정하십시오. 다음 분할에서 수가 초과하면 분할이 발생하기 전에 트리 성장이 중지됩니다.

최대 트리 깊이. 루트 노드 아래에 리프 노드의 최대 수준 수(즉, 표본을 반복적으로 분할하는 횟수)를 지정하십시오.

최소 하위 노드 크기. 상위 노드를 분할한 후 하위 노드에 포함해야 하는 최소 레코드 수를 지정하십시오. 하위 노드에 입력한 수보다 적은 레코드가 포함되면 상위 노드가 분할됩니다.

분할에 사용할 예측자 수 지정. 분할 모델을 작성하는 경우, 각 분할 작성에 사용할 최소 예측자 수를 설정하십시오. 그러면 분할로 인해 과도하게 작은 부집단이 작성되는 것을 방지할 수 있습니다. 이 옵션을 선택하지 않으면 기본값은 분류의 경우 $\lfloor \sqrt{M} \rfloor$ 이고 회귀분석의 경우 $\lfloor M/3 \rfloor$ 입니다. 여기서 M은 예측자 변수의 총 수입니다. 이 옵션을 선택하면 지정된 수의 예측자가 사용됩니다.

 **참고:** 분할에 대한 예측자 수는 데이터의 총 예측자 수보다 클 수 없습니다.

더 이상 정확도를 개선할 수 없는 경우에 작성 중단. 랜덤 트리에서는 학습 중지 시기를 결정할 때 특정 프로시저를 사용합니다. 특히 현재 앙상블 정확도가 지정된 임계값보다 적게 개선되면 새 트리 추가를 중지합니다. 그러면 **작성할 모델 수** 옵션에 지정된 값보다 트리 수가 적은 모델이 생성될 수 있습니다.

나. 랜덤 트리 노드 - 비용

일부 컨텍스트에서는 특정 오차 유형이 다른 유형에 비해 더 값비쌉니다. 예를 들어, 고위험 신용 거래 신청자를 저위험(오차의 한 유형)으로 분류하는 것이 저위험 신청자를 고위험(다른 유형의 오차)으로 분류하는 것보다 비용이 더 나갈 수 있습니다. 오분류 비용을 통해 여러 다른 유형의 예측 오차의 상대적 중요도를 지정할 수 있습니다.

오분류 비용은 기본적으로 특정 결과에 적용된 가중값입니다. 이 가중값은 모델에 영향을 미치는 요인이 되어 실제로 예측변수를 변경합니다(값비싼 실수에 대비하는 보호의 한 방법으로).

비용을 포함하는 모델은 그렇지 않은 항목보다 적은 오류를 생성하지 않으며, 전반적인 정확도 면에서 순위가 더 높지 않을 수도 있지만, 비용이 더 적게 드는 오류를 위해 기본 성향을 가지고 있으므로 실질적인 면에서 성능이 더 좋습니다.

비용 교차표는 예측 범주와 실제 범주의 가능한 각 조합의 비용을 표시합니다. 기본적으로 모든 오분류 비용은 1.0으로 설정됩니다. 사용자 정의 비용 값을 입력하려면 **오분류 비용 사용**을 선택하고 비용 교차표에 사용자 정의 값을 입력하십시오.

오분류 비용을 변경하려면 예측 및 실제 값의 원하는 조합에 해당하는 셀을 선택하고 셀의 기존 콘텐츠를 삭제한 후 셀의 원하는 비용을 입력하십시오. 비용은 자동으로 대칭되지 않습니다. 예를 들어, A를 B로 오분류한 비용을 2.0으로 설정할 경우 이 설정을 명시적으로 변경하지 않으면 B를 A로 오분류하는 비용의 기본값은 여전히 1.0입니다.

순서 목표의 경우에만 **순서 목표의 기본 비용 증가**를 선택하고 비용 행렬에서 기본값을 설정할 수 있습니다. 사용 가능한 옵션은 다음 목록에서 설명합니다.

- **증가 없음** - 잘못된 모든 예측에 대한 기본값은 1.0입니다.
- **선형** - 연속된 잘못된 각 예측은 비용을 1씩 증가시킵니다.
- **제공** - 연속된 잘못된 각 예측은 선형 값의 제공입니다. 이 경우 값은 1, 4, 9 등과 같습니다.
- **사용자 정의** - 테이블에서 값을 수동으로 편집하면 드롭다운 옵션이 자동으로 **사용자 정의**로 변경됩니다. 드롭다운 선택을 다른 옵션으로 변경하면 편집된 값을 선택한 옵션의 값으로 바꿉니다.

다. 랜덤 트리 노드 - 고급

의사결정 트리 작성 방법의 고급 옵션을 지정하십시오.

결측값의 최대 백분율. 입력에서 허용되는 결측값의 최대 백분율을 지정하십시오. 퍼센트가 이 수를 초과하면 모델 작성에서 입력이 제외됩니다.

단일 범주 다수가 있는 필드 제외. 필드 내에서 단일 범주에 속하는 레코드의 최대 퍼센트를 지정하십시오. 범주 값이 지정된 퍼센트보다 높은 레코드 퍼센트를 나타내면 전체 필드가 모델 작성에서 제외됩니다.

최대 필드 범주 수. 필드 내에 포함되는 최대 범주 수를 지정하십시오. 범주 수가 이 수를 초과하면 이 필드가 모델 작성에서 제외됩니다.

최소 필드 변동. 연속형 필드의 변동계수가 여기에 지정한 값보다 작을 경우(즉, 필드가 거의 일정할 경우) 이 필드는 모델 작성에서 제외됩니다.

구간 수. 연속 입력에 사용할 동일한 빈도 구간 수를 지정하십시오. 사용가능 옵션은 2, 4, 5, 10, 20, 25 ,50 또는 100입니다.

보고할 흥미로운 규칙 수. 보고할 규칙 수를 지정하십시오(최소값 1, 최대값 1000, 기본값은 50).

③ 랜덤 트리 노드 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 모델 스코어링 중에 예측자의 중요도를 계산하도록 선택할 수도 있습니다.

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

④ 랜덤 트리 모델 너깃

가. 랜덤 트리 모델 너깃 출력

랜덤 트리 모델을 작성하면 출력 뷰어에 다음 정보가 제공됩니다.

모델 정보 테이블

모델 정보 테이블은 모델에 대한 주요 정보를 제공합니다. 이 테이블에는 항상 다음과 같은 고급 모델 설정이 포함되어 있습니다.

- 유형 노드 또는 랜덤 트리 노드 필드 탭에서 선택된 목표 필드의 이름
- 모델 작성 방법 - 랜덤 트리
- 모델에 입력된 예측자 수

테이블에 표시되는 추가 세부사항은 분류 모델 또는 회귀 모델을 작성하는지 여부 및 불균형 데이터를 처리하기 위해 모델이 작성되었는지 여부에 따라 다릅니다.

- 분류 모델(기본 설정)
 - 모형 정확도
 - 오분류 규칙

- 분류 모델(불균형 데이터 처리 선택)
 - Gmean
 - 참 긍정 비율(클래스로 세분화됨)
- 회귀 모형
 - 제공된 평균제곱오차
 - 상대 오차
 - 설명된 분산

레코드 요약

요약에는 모델 적합에 사용된 레코드 수 및 제외된 레코드 수가 표시됩니다. 레코드 수와 정수의 퍼센트가 표시됩니다. 모델이 빈도 가중치를 포함하도록 작성된 경우 포함 및 제외된 가중되지 않은 레코드 수도 표시됩니다.

예측변수 중요도

예측자 중요도 그래프는 막대형 차트로 모델에 있는 상위 10개 입력(예측자)의 중요도를 표시합니다.

차트에 필드가 10개가 넘으면 차트 아래 슬라이더를 사용하여 차트에 포함되는 예측자의 선택을 변경할 수 있습니다. 슬라이더의 표시기 마크는 고정된 너비이며, 슬라이더의 각 마크는 10개 필드를 나타냅니다. 슬라이더와 함께 표시기 마크를 이동하여 예측자 중요도로 정렬된 다음 또는 이전 10개 필드를 표시할 수 있습니다.

차트를 두 번 클릭하면 그래프 크기를 편집할 수 있는 별도의 대화 상자를 열 수 있습니다. 별도의 이 편집 대화 상자를 닫으면 출력 탭에 표시된 차트에 변경이 적용됩니다.

상위 의사결정 규칙 테이블

기본적으로 이 대화형 테이블에는 상위 규칙의 통계가 표시되며, 이 통계는 흥미도를 기준으로 정렬됩니다.

테이블을 두 번 클릭하면 테이블에 표시된 규칙 정보를 편집할 수 있는 별도의 대화 상자를 열 수 있습니다. 표시되는 정보와 대화 상자에서 사용 가능한 옵션은 목표의 데이터 유형(예: 범주형 또는 연속형)에 따라 달라집니다.

다음 규칙 정보가 테이블에 표시됩니다.

- 규칙이 적용되고 구성되는 방법에 대한 세부사항
- 결과가 가장 빈도가 많은 범주에 있는지 여부
- 규칙 정확도
- 트리 정확도
- 흥미 지수

흥미 지수는 다음 수식을 사용하여 계산됩니다.

$$I_{index}(t) = P(A(t)) * P(B(t)) * (P(B(t)|A(t)) + P(\bar{B}(t)|\bar{A}(t)))$$

이 수식에서 각 요소는 다음과 같습니다.

- P(A(t))는 트리 정확도입니다.
- P(B(t))는 규칙 정확도입니다.
- P(B(t)|A(t))는 트리 및 노드별 정확한 예측을 나타냅니다.
- 나머지 수식은 트리 및 노드별 부정확 예측을 나타냅니다.

규칙 테이블 레이아웃은 다음 **테이블 콘텐츠** 옵션을 사용하여 변경할 수 있습니다.

- **상위 의사결정 규칙** - 흥미 지수를 기준으로 정렬되는 상위 5개 의사결정 규칙입니다.
- **모든 규칙** - 이 테이블에는 모델에서 생성한 모든 규칙이 포함되지만 페이지당 20개의 규칙만 표시됩니다. 이 레이아웃을 선택하면 **ID로 규칙 찾기** 및 **페이지**의 추가 옵션을 사용하여 규칙을 검색할 수 있습니다.

또한 범주형 대상의 경우 **범주별 상위 규칙** 옵션을 사용하여 규칙 테이블 레이아웃을 변경할 수 있습니다. 상위 5개 의사결정 규칙은 사용자가 선택한 **목표 범주**에 대한 총 레코드의 퍼센트로 정렬됩니다.

참고: 범주형 대상의 경우 작성 옵션의 기본 탭에서 **불균형 데이터 처리**를 선택하지 않은 경우에만 이 테이블을 사용할 수 있습니다.

규칙 테이블의 레이아웃을 변경하는 경우 대화 상자 왼쪽 상단에 있는 뷰어로 복사 단추를 클릭하여 수정된 규칙 테이블을 출력 뷰어로 다시 복사할 수 있습니다.

혼돈 행렬

분류 모델의 경우 혼돈 행렬은 정확한 예측의 비율을 포함하여 예측 결과 수 대비 실제 관측 결과 수를 보여줍니다.

참고: 혼돈 행렬은 회귀 모델에 사용할 수 없을 뿐 아니라 작성 옵션의 기본 탭에서 **불균형 데이터 처리**를 선택한 경우에도 사용할 수 없습니다.

나. 랜덤 트리 모델 너깃 설정

랜덤 트리 모델 너깃의 설정 탭에서 모델 스코어링 중 신뢰도 및 SQL 생성에 대한 옵션을 지정할 수 있습니다. 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능합니다.

신뢰도 계산 스코어링 작업에 신뢰도를 포함하려면 이 선택란을 선택합니다. 데이터베이스에서 모델 스코어를 계산할 때 신뢰도를 제외하면 보다 효율적인 SQL을 생성할 수 있음을 의미합니다. 회귀 트리에서는 신뢰도를 지정하지 않습니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성 방법을 지정하십시오.

- 기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 폐치하여 SPSS® Modeler에서 스코어를 계산합니다.
- 데이터베이스 외부 스코어 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 폐치하여 SPSS Modeler에서 스코어를 계산합니다.

(8) C&R 트리, CHAID, QUEST, C5.0 의사결정 트리 모형 너깃

의사결정 트리 모형 너깃은 의사결정 트리 모델링 노드(C&R 트리, CHAID, QUEST 또는 C5.0) 중 하나에서 발견한 특정 출력 필드를 예측하기 위한 트리 구조를 나타냅니다. 트리 모델은 트리 작성 노드에서 직접 생성되거나 대화형 트리 작성기를 통해 간접적으로 생성될 수 있습니다. 자세한 정보는 대화형 트리 작성기의 내용을 참조하십시오.

모델 너깃에서는 모델링 노드에 지정된 목표에 따라 서로 다른 옵션이 사용 가능합니다.

- 단일 트리 모델 너깃
- 부스팅, 배깅 및 매우 큰 데이터 세트를 위한 모델 너깃

트리 모델 스코어링

트리 모델 너깃을 포함하는 스트림을 실행하는 경우 트리 유형에 따라 특정 결과가 생성됩니다.

- 분류 트리(범주형 목표)의 경우 각 레코드에 대한 신뢰도와 예측값을 포함하는 2개의 새 필드가 데이터에 추가됩니다. 예측은 레코드가 지정된 터미널 노드의 가장 빈도가 많은 범주에 기반합니다. 지정된 노드의 반응자 대부분이 *여*인 경우 해당 노드에 지정된 모든 레코드의 예측은 *여*입니다.
- 회귀분석 트리에서는 예측값만 생성되고 신뢰도는 지정하지 않습니다.
- 선택적으로 CHAID, QUEST, C&R 트리 모델의 경우 각 레코드가 지정되는 노드의 ID를 표시하도록 추가 필드를 추가할 수 있습니다.

새 필드 이름은 접두문자를 추가하여 모델 이름에서 파생됩니다. C&R 트리, CHAID, QUEST의 경우 접두문자는 예측 필드의 경우 $\$R-$, 신뢰도 필드의 경우 $\$RC-$, 노드 식별자 필드의 경우 $\$RI-$ 입니다. C5.0 트리의 경우 접두문자는 예측 필드의 경우 $\$C-$, 신뢰도 필드의 경우 $\$CC-$ 입니다. 다중 트리 모델 노드가 있는 경우 새 필드 이름은 접두문자에 숫자를 포함하여 필요한 경우 필드를 구별합니다(예: $\$R1-$, $\$RC1-$, $\$R2-$).

트리 모델 너깃에 대한 작업

여러 방법으로 모델과 관련된 정보를 저장하거나 내보낼 수 있습니다.

참고: 이러한 옵션 중 많은 옵션이 트리 작성기 창에서도 사용 가능합니다.

트리 작성기 또는 트리 모델 너깃에서 다음을 수행할 수 있습니다.

- 현재 트리를 기반으로 필터 또는 선택 노드를 생성합니다. 자세한 정보는 필터 및 선택 노드 생성의 내용을 참조하십시오.
- 트리의 터미널 분기를 정의하는 규칙 세트로 트리 구조를 표시하는 규칙 세트 너깃을 생성합니다. 자세한 정보는 의사결정 트리에서 규칙 세트 생성의 내용을 참조하십시오.
- 또한 트리 모델 너깃의 경우에만 모델을 PMML 형식으로 내보낼 수 있습니다. 자세한 정보는 모델 팔레트의 내용을 참조하십시오. 모델에 사용자 정의 분할이 포함된 경우 이 정보는 내보낸 PMML에서 보존되지 않습니다. (분할은 보존되지만 알고리즘을 통해 선택된 것이 아니라 사용자 정의되었다는 사실은 그렇지 않습니다.)
- 현재 트리의 선택된 부분을 기반으로 그래프를 생성합니다. 스트림의 다른 노드에 연결되어 있을 경우에는 너깃에 대해서만 작동합니다. 자세한 정보는 그래프 생성의 내용을 참조하십시오.
- 부스팅 C5.0 모델인 경우에만 **단일 의사결정 트리(캔버스)** 또는 **단일 의사결정 트리(GM 팔레트)**를 선택하여 현재 선택된 규칙에서 파생된 새 단일 규칙 세트를 작성할 수 있습니다. 자세한 정보는 부스팅 C5.0 모델의 내용을 참조하십시오.

참고: 규칙 작성 노드가 C&R 트리 노드로 대체되었어도 처음에 규칙 작성 노드를 사용하여 작성된 기존 스트림의 의사결정 트리 노드는 계속해서 올바르게 작동합니다.

① 단일 트리 모델 너깃

모델링 노드에서 주요 목표로 **단일 트리 작성**을 선택한 경우 결과로 생성되는 모델 너깃은 다음 탭을 포함합니다.

Tab	설명	추가 정보
모델	모델을 정의하는 규칙을 표시합니다.	자세한 정보는 의사결정 트리 모형 규칙 주제를 참조하십시오.
뷰어	모델의 트리 보기를 표시합니다.	자세한 정보는 의사결정 트리 모형 규칙 주제를 참조하십시오.
요약	필드, 작성 설정, 모델 추정 프로세스에 대한 정보를 표시합니다.	자세한 정보는 의사결정 트리 모형 규칙 주제를 참조하십시오.
설정	모델 스코어링 중에 신뢰도 및 SQL 생성에 대한 옵션을 지정할 수 있습니다.	자세한 정보는 의사결정 트리 모형 규칙 주제를 참조하십시오.

Tab	설명	추가 정보
주석	설명 주석을 추가하고 사용자 정의 이름을 지정하고 도구 팁 텍스트를 추가하고 모델에 대한 검색 키워드를 지정할 수 있습니다.	

가. 의사결정 트리 모형 규칙

의사결정 트리 너짓의 모델 탭은 모델을 정의하는 규칙을 표시합니다. 선택적으로 예측자 중요도의 그래프 및 히스토리, 빈도, 대용에 대한 정보가 있는 세 번째 패널이 표시될 수도 있습니다.

참고: CHAID 노드 작성 옵션 탭(목적 패널)에서 **매우 큰 데이터 세트의 모델 작성 옵션**을 선택하면 모델 탭은 트리 규칙 세부사항만 표시합니다.

트리 규칙

왼쪽 분할창에는 알고리즘을 통해 발견한 데이터의 파티셔닝을 정의하는 조건 목록이 표시됩니다. 이는 본질적으로 여러 다른 예측자의 값을 기준으로 하여 개별 레코드를 하위 노드에 지정하는 데 사용할 수 있는 일련의 규칙입니다.

의사결정 트리는 입력 필드 값을 기준으로 하여 데이터를 반복해서 파티셔닝하는 방식으로 작동합니다. 데이터 파티션을 *분기*라 부릅니다. 초기 분기(때로 *루트*라 함)는 모든 데이터 레코드를 포함합니다. 루트는 특정 입력 필드의 값에 따라 서브세트 또는 *하위 분기*로 분할됩니다. 각 하위 분기는 계속해서 다시 다음 *하위 분기*로 차례로 분할되는 식입니다. 트리의 최저 수준에 있는 분기는 더 이상의 분할이 없습니다. 이러한 분기를 *터미널 분기*(또는 *리프*)라 부릅니다.

트리 규칙 세부사항

규칙 브라우저는 분할의 레코드에 대한 출력 필드 값 요약 및 각 파티션이나 분기를 정의하는 입력 값을 표시합니다. 모델 브라우저 사용에 대한 일반 정보는 모델 너짓 찾아보기의 내용을 참조하십시오.

수치 필드에 기반한 분할의 경우 분기가 다음 양식의 행으로 표시됩니다.

```
fieldname relation value [summary]
```

여기서, *relation*은 수치 관계입니다. 예를 들어, 수입 필드에 대해 100보다 큰 값으로 정의된 분기는 다음과 같이 표시됩니다.

```
revenue > 100 [summary]
```

기호 필드에 기반한 분할의 경우 분기가 다음 양식의 행으로 표시됩니다.

```
fieldname = value [summary] or fieldname in [values] [summary]
```

여기서, *values*는 분기를 정의하는 필드 값을 나타냅니다. 예를 들어, *region* 값이 *North*, *West* 또는 *South*일 수 있는 레코드를 포함한 분기는 다음으로 표시됩니다.

```
region in ["North" "West" "South"] [summary]
```

터미널 분기의 경우에는 규칙 조건의 끝에 예측값과 화살표가 추가된 예측도 제공됩니다. 예를 들어, 출력 필드에 대해 *high* 값을 예측하는 *revenue* > 100으로 정의된 리프는 다음으로 표시됩니다.

```
revenue > 100 [Mode: high] → high
```

분기의 요약은 기호 및 수치 출력 필드에 각기 다르게 정의됩니다. 수치 출력 필드가 있는 트리의 경우 요약은 분기의 평균 값이고 분기의 효과는 상위 분기의 평균과 분기 평균 간의 차분입니다. 기호 출력 필드가 있는 트리의 경우에는 요약이 분기의 레코드에 대한 최대 빈도 값 또는 모드입니다.

분기를 완전히 설명하려면 분기를 정의하는 조건 및 트리의 추가 분할을 정의하는 조건을 포함시켜야 합니다. 예를 들어, 트리에서

```
revenue > 100
region = "North"
region in ["South" "East" "West"]
revenue <= 200
```

두 번째 행에 표시된 분기는 *revenue* > 100 및 *region* = "North" 조건으로 정의됩니다.

도구 모음에서 **인스턴스/신뢰도 표시**를 클릭하면 규칙이 적용되는 레코드의 수(*인스턴스*)와 규칙이 참인 레코드의 비율(*신뢰도*)에 대한 정보도 각 규칙이 표시합니다.

예측변수 중요도

선택적으로 모델을 추정할 때 각 예측자의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측자에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하는 것이 좋습니다.

 **참고:** 이 차트는 모델을 생성하기 전에 분석 탭에 **예측자 중요도 계산**이 선택된 경우에만 사용 가능합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

추가 모델 정보

도구 모음에서 **추가 정보 패널 표시**를 클릭하면 선택한 규칙에 대한 자세한 정보를 보여주는 패널이 창의 맨 아래에 표시됩니다. 정보 패널에는 세 개의 탭이 있습니다.

히스토리. 이 탭은 루트 노드에서 아래의 선택한 노드로 분할된 조건을 추적합니다. 선택한 노드에 레코드가 지정된 시기를 판별하는 조건 목록을 제공합니다. 모든 조건이 참인 레코드가 이 노드에 할당됩니다.

빈도. 기호 목표 필드가 있는 모델의 경우 가능한 각 목표 값마다 이 탭은 이 노드에 할당된, 해당 목표 값이 있는 레코드 수를 표시합니다(훈련 데이터에서). 퍼센트로 표시된(최대 3자리수의 소수점이하자리수로 표시됨) 빈도 그림도 표시됩니다. 수치 목표가 있는 모델의 경우에는 이 탭이 비어 있습니다.

대용. 적용 가능한 경우 선택한 노드에 대한 기본 분할 필드의 대용이 표시됩니다. 대용은 주어진 레코드의 기본 예측자 값이 결측된 경우에 사용되는 대체 필드입니다. 주어진 분할의 허용된 최대 대용 수는 트리 작성 노드에 지정되지만 실제 수는 훈련 데이터에 따라 다릅니다. 일반적으로 결측 데이터가 많을수록 더 많은 대용이 사용될 수 있습니다. 기타 의사결정 트리 모형의 경우에는 이 탭이 비어 있습니다.

 **참고:** 모델에 포함하려면 훈련 단계 중에 대용을 식별해야 합니다. 훈련 표본에 결측값이 없으면 대용이 식별되지 않으며, 검정 또는 스코어링 중에 발견된 결측값이 있는 레코드는 자동으로 레코드 수가 가장 많은 하위 노드로 들어갑니다. 검정 또는 스코어링 중에 결측값이 예상되는 경우 반드시 훈련 표본에서도 값이 결측되었는지 확인하십시오. CHAID 트리에는 대용을 사용할 수 없습니다.

효과

노드의 효과는 평균 값의 증가 또는 감소입니다(상위 노드와 비교한 예측 값). 예를 들어 노드의 평균이 0.2이고 상위의 평균이 0.6이면 노드의 효과는 $0.2-0.6=-0.4$ 입니다. 이 통계는 연속형 대상에만 적용됩니다.

나. 의사결정 트리 모형 뷰어

의사결정 트리 모형 너깃의 뷰어 탭은 트리 작성기의 표시와 비슷합니다. 주된 차이는 모델 너깃을 찾아볼 때 트리를 성장시키거나 수정할 수 없다는 점입니다. 표시를 보고 사용자 정의하는 기타 옵션은 두 구성요소에서 서로 비슷합니다. 자세한 정보는 트리 보기 사용자 정의 주제를 참조하십시오.

참고: 뷰어 탭은 작성 옵션 탭 - 목표 패널에서 **매우 큰 데이터 세트에 대한 모델 작성** 옵션을 선택한 경우 작성된 CHAID 모델 너깃에서는 표시되지 않습니다.

뷰어 탭에서 분할 규칙을 보는 경우 꺾쇠 괄호는 인접한 값이 범위에 포함됨을 의미하지만, 소괄호는 인접한 값이 범위에서 제외됨을 의미합니다. 따라서 표현식 (23,37]은 23(제외)에서 37(포함) 사이의 범위(즉, 24부터 37까지)를 의미합니다. 모델 탭에서도 다음과 같이 동일한 조건이 표시됩니다.

```
Age > 23 and Age <= 37
```

다. 의사결정 트리/규칙 세트 모델 너깃 설정

의사결정 트리 또는 규칙 세트 모델 너깃의 설정 탭에서는 모델 스코어링 중 SQL 생성 및 신뢰도에 대한 옵션을 지정할 수 있습니다. 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능합니다.

신뢰도 계산 신뢰도를 스코어링 작업에 포함하려면 선택합니다. 데이터베이스에서 모델 스코어를 계산할 때 신뢰도를 제외하면 보다 효율적으로 SQL을 생성할 수 있습니다. 회귀 트리에서는 신뢰도를 지정하지 않습니다.

참고: CHAID 모델의 작성 옵션 탭 - 모델 패널에서 **매우 큰 데이터 세트에 대한 모델 작성** 옵션을 선택한 경우 이 선택란은 명목 또는 플래그에 해당하는 범주형 목표의 모델 너깃에서만 사용 가능합니다.

원시 성향 스코어 계산 예 또는 아니오 예측을 반환하는 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 이외에도 스코어링 중에 생성할 수 있는 기타 예측 및 신뢰도 값이 있습니다.

참고: CHAID 모델의 작성 옵션 탭 - 모델 패널에서 **매우 큰 데이터 세트에 대한 모델 작성** 옵션을 선택한 경우 이 선택란은 플래그에 해당하는 범주형 목표의 모델 너깃에서만 사용 가능합니다.

수정된 성향 스코어 계산 원시 성향 스코어는 훈련 데이터에만 기반하며, 이 데이터의 과적합을 위한 많은 모델의 경향으로 인해 지나치게 낙관적일 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 성능을 평가하여 보완하려고 합니다. 이 옵션에서는 파티션 필드가 스트림에 정의되어야 하고 모델 생성 전에 모델링 노드에서 수정된 성향 스코어가 사용 가능해야 합니다.

참고: 수정된 성향 스코어는 증폭된 트리 및 규칙 세트 모델에서 사용할 수 없습니다. 자세한 정보는 부스팅 C5.0 모델의 내용을 참조하십시오.

규칙 식별자 CHAID, QUEST, C&R 트리 모델의 경우 이 옵션은 각 레코드가 지정된 터미널 노드의 ID를 표시하는 필드를 스코어링 출력에 추가합니다.

참고: 이 옵션을 선택하면 SQL 생성을 수행할 수 없습니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **결측값 지원 없이 이 모형의 SQL 생성** 이 옵션을 선택하면 결측값 처리를 위한 오버헤드 없이도 데이터베이스에서 모델 스코어를 계산할 원형 SQL을 생성합니다. 이 옵션은 케이스 스코어링 시 결측값이 발생하면 단순히 예측을 널(\$null\$)로 설정합니다.

참고: 이 옵션은 CHAID 모델에서 사용할 수 없습니다. 다른 모델 유형의 경우 의사결정 트리(규칙 세트가 아님)에서만 사용 가능합니다.

- **결측값 지원을 통해 이 모형의 SQL 생성** CHAID, QUEST, C&R 트리 모델의 경우 전체 결측값 지원을 통해 데이터베이스에서 모델 스코어를 계산할 원형 SQL을 생성합니다. 즉, 모델에 지정된 대로 결측값을 처리하도록 SQL이 생성됩니다. 예를 들어, C&R 트리는 대용 규칙 및 가장 큰 하위 폴백을 사용합니다.

참고: C5.0 모델의 경우 이 옵션은 규칙 세트(의사결정 트리가 아님)에서만 사용할 수 있습니다.

- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

라. 부스팅 C5.0 모델

이 기능은 SPSS® Modeler Professional 및 SPSS Modeler Premium에서 사용 가능합니다.

부스팅 C5.0 모델(규칙 세트 또는 의사결정 트리)을 작성하는 경우 실제로는 관련 모델 세트를 작성하는 것입니다. 부스팅 C5.0 모델의 모델 규칙 브라우저는 각 모델의 추정된 정확도 및 부스팅 모델 앙상블의 전체 정확도와 함께 계층의 최고 수준에 있는 모델 목록을 표시합니다. 특정 모델의 규칙이나 분할을 검토하려면 해당 모델을 선택하고 단일 모델에서 규칙 또는 분기에 행한 것처럼 모델을 펼치십시오.

부스팅 모델 세트에서 특정 모델을 추출한 후 해당 모델만 포함한 새 규칙 세트 모델 너깃을 작성할 수도 있습니다. 부스팅 C5.0 모델에서 새 규칙 세트를 작성하려면 관심 있는 트리 또는 규칙 세트를 선택하고 생성 메뉴에서 **단일 의사결정 트리(GM 팔레트)** 또는 **단일 의사결정 트리(캔버스)**를 선택하십시오.

마. 그래프 생성

트리 노드는 많은 정보를 제공하지만 비즈니스 사용자가 쉽게 액세스할 수 있는 형식이 아닐 경우가 있습니다. 비즈니스 보고서, 프레젠테이션 등에 쉽게 통합할 수 있는 방식으로 데이터를 제공하기 위해 선택한 데이터의 그래프를 만들 수 있습니다. 예를 들어, 모델 너깃의 모델 또는 뷰어 탭이나 대화형 트리의 뷰어 탭에서 선택한 트리 파트에 대한 그래프를 생성할 수 있습니다. 그렇게 함으로써 선택한 트리나 분기 노드의 케이스에 대해서만 그래프를 작성합니다.

참고: 스트림의 다른 노드에 연결되어 있을 때에는 너깃에서만 그래프를 생성할 수 있습니다.

그래프 생성

첫 번째 단계는 그래프에 표시할 정보를 선택하는 것입니다.

- 너깃의 모델 탭에서 왼쪽 분할창의 조건 및 규칙 목록을 펼치고 관심 있는 항목을 하나 선택하십시오.
- 너깃의 뷰어 탭에서 분기 목록을 펼치고 관심 있는 노드를 선택하십시오.
- 대화형 트리의 뷰어 탭에서 분기 목록을 펼치고 관심 있는 노드를 선택하십시오.
참고: 둘 중 어느 뷰어 탭에서도 최상위 노드는 선택할 수 없습니다.

선택한 데이터 표시 방식과 무관하게 그래프를 작성하는 방식은 동일합니다.

1. 생성 메뉴에서 **그래프(선택 사항 기준)**를 선택하거나 또는 뷰어 탭에서 맨 아래 왼쪽 구석의 **그래프(선택 사항 기준)** 단추를 클릭하십시오. 그래프보드 기본 탭이 표시됩니다.
참고: 기본 및 세부사항 탭은 그래프보드를 이러한 방식으로 표시할 때에만 사용 가능합니다.
2. 기본 또는 세부사항 탭 설정을 사용하여 그래프에 표시할 세부사항을 지정하십시오.
3. 확인을 눌러 그래프를 생성하십시오.

그래프 머리말은 포함하도록 선택한 노드 또는 규칙을 식별합니다.

② 부스팅, 배깅 및 매우 큰 데이터 세트의 모델 너깃

모형 정확도 개선(boosting), 모형 안정성 개선(bagging) 또는 **매우 큰 데이터 세트용 모델 작성**을 모델링 노드의 기본 목표로 선택하는 경우 IBM® SPSS® Modeler에서는 다중 모델의 앙상블을 작성합니다. 자세한 정보는 앙상블 모델 주제를 참조하십시오.

결과로 생성된 모델 너깃은 다음 탭을 포함합니다. 모델 탭에서는 다양한 모델 보기를 제공합니다.

표 1. 모델 너깃에서 사용 가능한 탭

Tab	보기	설명	추가 정보
모델	모델 요약	양상불 품질 및 (부스팅 모델 및 연속형 목표 제외) 다양성, 서로 다른 모델에서 예측 다양성의 측도에 대한 요약을 표시합니다.	자세한 정보는 모델 요약(양상불 뷰어) 주제를 참조하십시오.
	예측변수 중요도	모델 추정 시 각 예측자의 상대적 중요도(입력 필드)를 나타내는 차트를 표시합니다.	자세한 정보는 예측변수 중요도(양상불 뷰어) 주제를 참조하십시오.
	예측자 빈도	모델 세트에 사용된 각 예측자의 상대적 빈도를 나타내는 차트를 표시합니다.	자세한 정보는 예측자 빈도(양상불 뷰어) 주제를 참조하십시오.
	구성요소 모형 정확도	양상불에서 각 서로 다른 모델의 예측 정밀도의 차트를 구성합니다.	
	구성요소 모델 세부사항	양상불에서 각 서로 다른 각 모델에 대한 정보를 표시합니다.	자세한 정보는 구성요소 모델 세부사항 (양상불 뷰어) 주제를 참조하십시오.
	정보	필드, 작성 설정, 모델 추정 프로세스에 대한 정보를 표시합니다.	자세한 정보는 모델 너깃 요약/정보 주제를 참조하십시오.
설정		스코어링 작업에 신뢰도를 포함할 수 있습니다.	자세한 정보는 의사결정 트리/규칙 세트 모델 너깃 설정 주제를 참조하십시오.
주석		설명 주석을 추가하고 사용자 정의 이름을 지정하고 도구 팁 텍스트를 추가하고 모델에 대한 검색 키워드를 지정할 수 있습니다.	

(9) C&R 트리, CHAID, QUEST, C5.0, Apriori 규칙 세트 모델 너깃

규칙 세트 모델 너깃은 연관 규칙 모델링 노드(Apriori) 또는 트리 작성 노드(C&R 트리, CHAID, QUEST 또는 C5.0) 중 하나를 통해 검색한 특정 출력 필드를 예측하기 위한 규칙을 표시합니다. 연관 규칙의 경우 세분화되지 않은 규칙 너깃에서 규칙 세트가 생성되어야 합니다. 트리의 경우에는 대화형 트리 작성기, C5.0 모델 작성 노드 또는 트리 모델 너깃에서 규칙 세트가 생성될 수 있습니다. 세분화되지 않은 규칙 너깃과 다르게, 규칙 세트 너깃은 예측을 생성하도록 스트림에 둘 수 있습니다.

규칙 세트 너깃을 포함한 스트림을 실행하는 경우 데이터에 대한 각 레코드의 예측 값과 신뢰도를 포함한 두 개의 새 필드가 스트림에 추가됩니다. 새 필드 이름은 접두문자를 추가하여 모델 이름에서 파생됩니다. 연관 규칙 세트의 접두문자는 예측 필드의 경우 \$A-이고 신뢰도 필드는 \$AC-입니다. C5.0 규칙 세트의 접두문자는 예측 필드의 경우 \$C-이고 신뢰도 필드는 \$CC-입니다. C&R 트리 규칙 세트의 접두문자는 예측 필드의 경우 \$R-이고 신뢰도 필드는 \$RC-입니다. 한 계열에 동일한 출력 필드를 예측하는 여러 규칙 세트 너깃이 있는 스트림에서는, 새 필드 이름의 접두문자에 서로를 구별하는 번호가 포함됩니다. 스트림의 첫 번째 연관 규칙 세트 너깃은 일반 이름을 사용하고, 두 번째 노드는 \$A1- 및 \$AC1-으로 시작하는 이름을 사용하며, 세 번째 노드는 \$A2- 및 \$AC2-으로 시작하는 이름을 사용하는 식입니다.

규칙의 적용 방식. 연관 규칙에서 생성된 규칙 세트는 특정 레코드의 경우 둘 이상의 예측이 생성될 수 있고 이 예측이 모두 일치하는 것은 아니므로 다른 모델 너깃과 차이가 있습니다. 규칙 세트에서 예측을 생성하는 두 가지 방법이 있습니다.

참고: 의사결정 트리에서 생성되는 규칙 세트는 의사결정 트리에서 파생된 규칙이 상호 배타적이어서 사용된 방법과 상관 없이 동일한 결과를 리턴합니다.

- **투표.** 이 방법은 레코드에 적용되는 모든 규칙의 예측을 결합하려 시도합니다. 각 레코드마다 모든 규칙을 검토하고 레코드에 적용되는 각 규칙을 사용하여 예측 및 연관된 신뢰도를 생성합니다. 각 출력 값의 신뢰도 수치 합계를 계산하고 신뢰도 합계가 가장 큰 값을 최종 예측으로 선택합니다. 최종 예측의 신뢰도는 해당 레코드에 실행한 규칙 수로 나눈 값의 신뢰도 합계입니다.
- **첫 번째 적용.** 이 방법은 단순히 규칙을 순서대로 검정합니다. 레코드에 적용되는 첫 번째 규칙은 예측을 생성하는 데 사용된 규칙입니다.

스트림 옵션으로 사용되는 방법을 제어할 수 있습니다.

노드 생성. 생성 메뉴로 규칙 세트에 기반하여 새 노드를 작성할 수 있습니다.

- **필터 노드** 규칙 세트의 규칙에 사용되지 않는 필드를 필터링하기 위한 새 필터 노드를 작성합니다.
- **선택 노드** 선택한 규칙이 적용되는 레코드를 선택할 새 선택 노드를 작성합니다. 생성된 노드는 규칙의 모든 전항이 참인 레코드를 선택합니다. 이 옵션의 경우 규칙을 선택해야 합니다.
- **규칙 추적 노드** 각 레코드의 예측을 작성하는 데 사용된 규칙을 표시하는 필드를 계산할 새 SuperNode를 작성합니다. 규칙 세트가 첫 번째 적용 방법을 사용하여 평가되는 경우 이는 단순히 실행할 첫 번째 규칙을 나타내는 기호입니다. 규칙 세트가 투표 방법을 사용하여 평가되는 경우에는 투표 메커니즘에 대한 입력을 표시하는 보다 복잡한 문자열입니다.
- **단일 의사결정 트리(캔버스) / 단일 의사결정 트리(GM 팔레트).** 현재 선택한 규칙에서 파생된 새 단일 규칙 세트 너깃을 작성합니다. **중복된** C5.0 모델에만 사용 가능합니다. 자세한 정보는 부스팅 C5.0 모델의 내용을 참조하십시오.
- **모델을 팔레트로** 모델 팔레트로 모델을 리턴합니다. 이 옵션은 동료가 모델 자체가 아닌 모델을 포함한 스트림을 보냈을 경우에 유용합니다.

참고: 규칙 세트 너깃의 설정 및 요약 탭은 의사결정 트리 모형의 탭과 동일합니다.

① 규칙 세트 모델 탭

규칙 세트 너깃의 모델 탭은 알고리즘을 통해 데이터에서 추출된 규칙 목록을 표시합니다.

규칙은 후항(예측 범주)별로 세분화되며 다음 형식으로 표시됩니다.

```
if antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

여기서, consequent 및 antecedent_1 ~ antecedent_n은 모든 조건입니다. 규칙은 "antecedent_1 ~ antecedent_n이 모두 참이고 consequent도 참일 가능성이 있는 레코드"로 해석됩니다. 도구 모음에서 **인스턴스/신뢰도 표시** 단추를 클릭하면 각 규칙이 규칙이 적용되는-- 즉, 전항이 참인 레코드의 수(**인스턴스**)와 전체 규칙이 참인 레코드의 비율(**신뢰도**)에 대한 정보도 표시합니다.

C5.0 규칙 세트에 대한 신뢰도는 다소 다르게 계산됨에 유의하십시오. C5.0은 규칙의 신뢰도를 계산할 때 다음 공식을 사용합니다.

```
(1 + number of records where rule is correct)
/
(2 + number of records for which the rule's antecedents are true)
```

이 신뢰도 추정값 계산은 의사결정 트리에서 규칙을 생성하는 프로세스(C5.0이 규칙 세트를 작성할 때 수행함)에 대해 조정됩니다.

(10) AnswerTree 3.0에서 프로젝트 가져오기

IBM® SPSS® Modeler는 다음과 같이 표준 파일 > 열기 대화 상자를 사용하여 AnswerTree 3.0 또는 3.1에 저장된 프로젝트를 가져올 수 있습니다.

1. IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

파일 > 스트림 열기

2. 유형 드롭 다운 목록의 파일에서 **AT 프로젝트 파일(*.atp, *.ats)**을 선택하십시오.

가져온 각 프로젝트는 다음 노드가 있는 IBM SPSS Modeler 스트림으로 변환됩니다.

- 사용된 데이터 소스를 정의하는 하나의 소스 노드(예를 들어, IBM SPSS Statistics 데이터 파일 또는 데이터베이스 소스).
- 프로젝트의 각 트리마다(여러 개일 수 있음) 유형, 역할(입력 또는 예측변수 필드 대 출력 또는 예측 필드), 결측값, 기타 옵션을 포함하여 각 필드(변수)에 대한 특성을 정의하는 하나의 유형 노드가 작성됩니다.

- 프로젝트의 각 트리마다 학습 또는 검정 표본의 데이터를 분할하는 파티션 노드가 작성되고 트리를 생성하기 위한 모수를 정의하는 트리 작성 노드(C&R 트리, QUEST 또는 CHAID 노드)가 작성됩니다.

3. 생성된 트리를 보려면 스트림을 실행하십시오.

설명

- IBM SPSS Modeler의 생성된 의사결정 트리를 AnswerTree로 내보낼 수 없습니다. AnswerTree에서 IBM SPSS Modeler로의 가져오기는 단방향 트립입니다.
- AnswerTree에 정의된 이익은 프로젝트를 IBM SPSS Modeler로 가져오면 보존되지 않습니다.

5) 베이지안 신경망 모델

(1) 베이지안 네트워크 노드

베이지안 네트워크 노드로 관측 및 기록한 증거를 "상식적인" 실세계 지식과 결합해서 겹보기에 링크되지 않은 속성을 사용하여 발생 우도를 설정함으로써 확률 모델을 작성할 수 있습니다. 이 노드는 주로 분류에 사용하는 TAN(Tree Augmented Naïve Bayes) 및 Markov Blanket 네트워크에 초점을 맞춥니다.

베이지안 네트워크는 여러 다양한 상황에서 예측을 수행하는 데 사용됩니다. 다음은 몇 가지 예입니다.

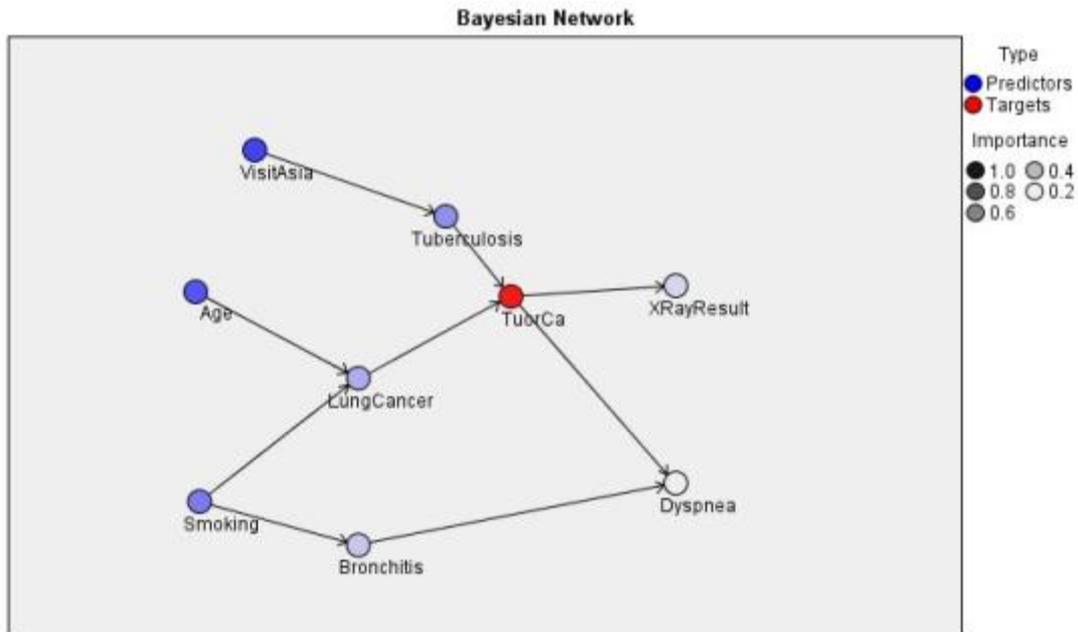
- 채무 불이행 위험이 낮은 대출 기회 선택.
- 센서 입력 및 기존 레코드를 기준으로 하여 설비의 서비스, 부품 또는 교체가 필요한 시기 추정.
- 온라인 문제점 해결 도구를 통한 고객 문제점 해결.
- 실시간으로 휴대 전화 네트워크 문제점 진단 및 해결.
- 최상의 기회에 자원을 집중시키기 위한 연구 개발 프로젝트의 잠재적 위험 및 보상 평가.

베이지안 네트워크는 데이터 세트의 변수(종종 노드라 부름)와 이 변수 사이의 확률적 또는 조건부 독립성을 표시하는 그래픽 모델입니다. 노드 간의 인과 관계를 베이지안 네트워크를 통해 표시할 수 있지만 네트워크의 링크(아크로도 알려짐)가 반드시 직접적인 원인과 결과를 표시하지는 않습니다. 예를 들어, 그래프에 표시된 증상과 질병 간의 확률적 독립성이 참인 경우 특정 증상 및 기타 관련 데이터의 유무가 제공되면 베이지안 네트워크를 사용하여 특정 질병이 있는 환자의 확률을 계산할 수 있습니다. 네트워크는 정보가 누락된 지점에서 매우 강력하며 존재하는 정보를 사용하여 가능한 최상의 예측을 수행합니다.

베이지안 네트워크의 공통 기본 예는 Lauritzen 및 Spiegelhalter에 의해 작성되었습니다(1988). 이 예는 종종 "아시아" 모델이라 불리며 의사의 새 환자를 진단(대략 인과 관계에 해당

하는 링크의 방향)하는 데 사용할 수 있는 네트워크의 단순화된 버전입니다. 각 노드는 환자의 조건에 관련시킬 수 있는 패킷을 나타냅니다. 예를 들어, "Smoking"은 환자가 확실한 흡연자임을 나타내고 "VisitAsia"는 환자가 최근에 아시아를 방문했음을 표시합니다. 확률 관계는 노드 간의 링크로 표시됩니다. 예를 들어, 흡연은 기관지염과 폐암이 모두 진행 중인 환자의 발생을 늘리는 반면 나이는 폐암 발생 가능성에만 연관된 것처럼 보입니다. 이와 마찬가지로, 폐 x-레이 상의 이상은 결핵 또는 폐암으로 인한 것이 수 있는 반면에 환자가 기관지염이나 폐암도 앓는 경우에는 숨가쁨(호흡 곤란)으로 고통받는 환자의 발생이 증가합니다.

그림 1. Lauritzen 및 Spiegelhalter의 아시아 네트워크 예



베이지안 네트워크의 사용을 결심할 수 있는 여러 원인이 있습니다.

- 인과 관계에 대해 훈련하도록 돕습니다. 이를 통해 문제 영역을 이해하고 개입 결과를 예측할 수 있습니다.
- 네트워크는 데이터 과적합을 피할 수 있는 효과적인 접근법을 제공합니다.
- 관련된 관계의 명확한 시각화를 쉽게 관측할 수 있습니다.

요구사항. 대상 필드는 범주형이어야 하며 측정 수준은 *명목*, *순서* 또는 *플래그*가 가능합니다. 입력은 임의의 유형의 필드일 수 있습니다. 연속(수치 범위) 입력 필드는 자동으로 구간화되지만 분포가 왜곡될 경우 베이지안 네트워크 노드 이전에 구간화 노드를 사용하여 수동으로 필드를 구간화해서 더 나은 결과를 얻을 수 있습니다. 예를 들어, **수퍼바이저 필드**가 베이지안 네트워크 노드 **목표** 필드와 동일한 최적 구간화를 사용하십시오.

예. 한 은행의 분석가는 대출 상환을 불이행할 것 같은 잠재적 고객 또는 고객을 예측할 수 있기를 원합니다. 베이지안 신경망 모델을 사용하여 채무를 불이행할 것 같은 고객의 특성을 식별하고 잠재적 채무 불이행자를 예측하는 데 가장 적합한 모델을 설정하기 위해 여러 다른 유형의 모델을 작성할 수 있습니다.

예. 한 통신 사업자는 사업을 그만두려는("이탈"이라 함) 고객 수를 줄이고 전월의 각 데이터를 사용하여 매월 모델을 업데이트하려 합니다. 베이지안 신경망 모델을 사용하여 이탈할 것 같은 고객의 특성을 식별하고 매월 새 데이터로 모델 학습을 계속할 수 있습니다.

① 베이지안 네트워크 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

각 분할의 작성 모델. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성 주제를 참조하십시오.

파티션. 이 필드에서는 모델 작성의 훈련, 검정, 검증 단계를 위한 개별 표본으로 데이터를 분할하는 데 사용되는 필드를 지정할 수 있습니다. 한 표본을 사용하여 모델을 생성하고 다른 표본을 사용하여 이를 검정함으로써 현재 데이터에 유사한 보다 큰 데이터 세트에 모델이 일반화되는 정도를 잘 표시할 수 있습니다. 유형 또는 파티션 노드를 사용하여 다중 파티션 필드를 정의한 경우 파티셔닝을 사용하는 각 모델링 노드에 있는 필드 탭에서 단일 파티션 필드를 선택해야 합니다. (하나의 파티션만 존재하는 경우 파티션이 사용될 때마다 자동으로 사용됩니다.) 또한 분석에서 선택한 파티션을 적용하려면 노드의 모델 옵션 탭에서 파티셔닝을 사용 가능하게 해야 합니다. (이 옵션을 선택 취소하면 필드 설정을 변경하지 않고도, 파티셔닝을 사용하지 않을 수 있습니다.)

분할. 분할 모델의 경우 단일 또는 복수 분할 필드를 선택하십시오. 이는 유형 노드에서 필드 역할을 분할로 설정하는 것과 유사합니다. 측정 수준이 플래그, 명목, 순서 또는 연속인 필드만 분할 필드로 지정할 수 있습니다. 분할 필드로 선택된 필드는 목표, 입력, 파티션, 빈도 또는 가중 필드로 사용할 수 없습니다. 자세한 정보는 분할 모델 작성 주제를 참조하십시오.

기존 모델 훈련 계속. 이 옵션을 선택하면 모델이 실행될 때마다 모델 너깃 모델 탭에 표시된 결과가 재생성되고 업데이트됩니다. 예를 들어, 새 데이터 소스나 업데이트한 데이터 소스를 기존 모델에 추가했을 때 이를 수행합니다.

참고: 이 옵션은 기존 네트워크를 업데이트만 할 수 있으며 노드 또는 연결을 추가하거나 제거할 수는 없습니다. 모델을 재훈련할 때마다 네트워크의 모양이 동일하게 되고 조건부 확률 및 예측자 중요도만 변경됩니다. 새 데이터가 이전 데이터와 대체로 비슷한 경우에는 유의적이라 여기는 사항이 동일하므로 이는 문제가 되지 않지만, 유의적인 항목을 검사 또는 업데이트하려면 (얼마나 유의적인지에 반대됨) 새 모델 즉, 새 네트워크를 작성해야 합니다.

구조 유형. 베이지안 네트워크를 작성할 때 사용할 구조를 선택하십시오.

- **TAN.** TAN(Tree Augmented Naïve Bayes 모델)은 표준 Naïve Bayes 모델보다 개선된 단순 베이지안 신경망 모델을 작성합니다. 이는 각 예측자가 목표변수 외에 또 다른 예측자에 종속되도록 허용해서 분류 정확도가 증가하기 때문입니다.
- **Markov Blanket.** 데이터 세트에서 목표변수의 상위, 하위, 하위의 상위를 포함한 노드 세트를 선택합니다. Markov blanket은 본질적으로 네트워크에서 목표변수를 예측하는 데 필요한 모든 변수를 식별합니다. 이 네트워크 작성 방법이 보다 정확하다고 간주되지만 큰 데이터 세트의 경우 포함된 변수의 수가 많아서 처리 시간이 길어질 수 있습니다. 처리량을 줄이려면 고급 탭의 필드선택 옵션을 사용하여 유의적으로 목표변수에 관련된 변수를 선택할 수 있습니다.

필드선택 전처리 단계 포함. 이 상자를 선택하면 고급 탭에서 필드선택 옵션을 사용할 수 있습니다.

모수 학습 방법. 베이지안 네트워크 모수는 상위 값이 주어진 각 노드의 조건부 확률을 말합니다. 상위 값이 알려진 노드 간에 조건부 확률 표를 추정하는 작업을 제어하는 데 사용할 수 있는 두 가지 가능한 선택이 있습니다.

- **최대우도.** 큰 데이터 세트를 사용할 때 이 상자를 선택하십시오. 기본 선택사항입니다.
- **작은 셀 빈도의 Bayes 조정.** 더 작은 데이터 세트의 경우 많은 0의 수 외에 모델 과적합 위험이 있습니다. 평활을 적용하여 0의 수 효과 및 신뢰할 수 없는 추정 효과를 줄여서 이 문제를 완화하려면 이 옵션을 선택하십시오.

② 베이지안 네트워크 노드 고급 옵션

노드 고급 옵션으로 모델 작성 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

결측값. 기본적으로 IBM® SPSS® Modeler에서는 모델에 사용된 모든 필드의 유효한 값을 포함하는 레코드만 사용합니다. (이 기능을 때때로 결측값의 **목록별 삭제**라고도 합니다.) 결측 데이터가 많은 경우 이 접근 방식을 사용하면 너무 많은 레코드가 제거되므로 데이터가 부족하여 좋은 모델을 생성하지 못할 수도 있습니다. 이러한 경우에는 **완전한 레코드만 사용** 옵션을 선택 취소할 수 있습니다. 그러면 IBM SPSS Modeler에서는 일부 필드에 결측값이 있는 레코드를 포함하여 모델을 추정할 수 있을 만큼 많은 정보를 사용하려고 합니다. (이 기능을 때때로 결측값의 **대응별 삭제**라고도 합니다.) 그러나 일부 상황에서 이러한 방식으로 불완전한 레코드를 사용하면 모델 추정 시 계산상의 문제점이 발생할 수 있습니다.

모든 확률 추가. 출력 필드의 각 범주에 대한 확률을 노드에서 처리하는 각 레코드에 추가하는지 여부를 지정합니다. 이 옵션을 선택하지 않으면 예측 범주의 확률만 추가됩니다.

독립성 검정. 독립성 검정은 두 개의 변수에 대한 쌍을 이룬 관측이 서로 독립적인지 여부를 평가합니다. 사용할 검정 유형을 선택하십시오. 사용 가능한 옵션은 다음과 같습니다.

- **우도비.** 두 가지 다른 가설 하에서 결과의 최대 확률 간 비율을 계산하여 목표 예측자 독립성을 검정합니다.
- **Pearson 카이제곱.** 지정된 빈도 분포 다음에 관측 이벤트 발생의 상대 빈도가 나오는 귀무가설을 사용하여 목표 예측자 독립성을 검정합니다.

베이지안 신경망 모델은 검정 대응에 추가 변수가 사용되는 조건부 독립성 검정을 수행합니다. 또한 모델은 목표와 예측자 간 관계 외에 예측자 자체 사이의 관계도 탐색합니다.

참고: 독립성 검정 옵션은 모델 탭에서 Markov Blanket의 필드선택 전처리 단계 포함 또는 구조 유형을 선택한 경우에만 사용할 수 있습니다.

유의 수준. 독립성 검정 설정과 함께 사용되어 검정을 수행할 때 사용할 분리점 값을 설정할 수 있게 합니다. 이 값이 낮을수록 네트워크에 링크가 더 적게 남습니다. 기본 수준은 0.01입니다.

참고: 이 옵션은 모델 탭에서 Markov Blanket의 필드선택 전처리 단계 포함 또는 구조 유형을 선택한 경우에만 사용할 수 있습니다.

최대 조건 세트 크기. Markov Blanket 구조를 작성하는 알고리즘은 증가량 크기 조건 세트를 사용하여 독립성 검정을 수행하고 네트워크에서 불필요한 링크를 제거합니다. 많은 수의 조건 변수를 포함한 검정은 처리하는 데 많은 시간과 메모리가 필요하므로 포함할 변수의 수를 제한할 수 있습니다. 이는 특히 많은 변수 사이에 종속성이 강한 데이터를 처리할 때 유용합니다. 하지만 결과적인 네트워크에는 일부 불필요한 링크가 포함될 수 있음에 유의하십시오.

독립성 검정에 사용할 조건 변수의 최대 수를 지정하십시오. 기본 설정은 5입니다.

참고: 이 옵션은 모델 탭에서 Markov Blanket의 필드선택 전처리 단계 포함 또는 구조 유형을 선택한 경우에만 사용할 수 있습니다.

필드선택. 이 옵션으로 모델 작성 프로세스의 속도를 올리기 위해 모델을 처리할 때 사용되는 입력 수를 제한할 수 있습니다. 이는 특히 가능한 많은 수의 잠재 입력으로 인해 Markov Blanket 구조를 작성할 때 유용합니다. 이 옵션을 사용하여 목표변수에 유의적으로 관련된 입력을 선택할 수 있습니다.

참고: 필드선택 옵션은 모델 탭에서 필드선택 전처리 단계 포함을 선택한 경우에만 사용할 수 있습니다.

- **항상 선택 내용 입력** - 필드 선택기(텍스트 필드의 오른쪽에 있는 단추)를 사용하여 데이터 세트에서 베이지안 신경망 모델을 작성할 때 항상 사용할 필드를 선택하십시오. 목표 필드는 항상 선택됩니다. 다른 검정에서 유의적으로 간주하지 않을 경우 모델 작성 프로세스 중 베이지안 네트워크가 여전히 이 목록에서 항목을 삭제할 수 있습니다. 따라서 이 옵션은 목록에 있는 항목이 생성되는 베이지안 모델에 반드시 나타나는지 확인하는 것이 아니라, 단순히 목록에 있는 항목이 모델 작성 프로세스 자체에 사용되었는지 확인합니다.

- **최대 입력 수.** 데이터 세트에서 베이지안 신경망 모델을 작성할 때 사용할 총 입력 수를 지정하십시오. 입력 가능한 최고 수는 데이터 세트의 총 입력 수입니다.

참고: 항상 선택 내용 입력에서 선택한 필드 수가 최대 입력 수의 값을 초과할 경우 오류 메시지가 표시됩니다.

(2) 베이지안 신경망 모델 너깃

참고: 모델링 노드 모델 탭에서 기존 모수 훈련 계속을 선택하면 모델을 재생성할 때마다 모델 너깃 모델 탭에 표시되는 정보가 업데이트됩니다.

모델 너깃 모델 탭이 두 개의 분할창으로 분할됩니다.

왼쪽 분할창

기본 이 보기에는 목표와 가장 중요한 예측자 간 관계 및 예측자 사이의 관계를 표시하는 노드의 네트워크 그래프가 있습니다. 각 예측자의 중요도가 색상 농도에 따라(더 진한 색이 중요한 예측자를 나타내고 그 반대일 수도 있음) 표시됩니다.

범위를 나타내는 노드의 구간 값은 노드 위에 마우스 포인터를 두면 도구팁에 표시됩니다.

IBM® SPSS® Modeler에서 그래프 도구를 사용하여 그래프를 상호작용, 편집, 저장할 수 있습니다. 예를 들어, MS Word와 같은 다른 애플리케이션에 사용할 수 있습니다.

팁: 네트워크에 많은 노드가 있는 경우 그래프를 더 쉽게 판독할 수 있도록 노드를 클릭하여 선택한 후 끌 수 있습니다.

분포 이 보기는 네트워크에 있는 각 노드의 조건부 확률을 미니 그래프로 표시합니다. 도구팁에 값을 표시하려면 그래프 위에 마우스 포인터를 두십시오.

오른쪽 분할창

예측자 중요도 모델을 추정할 때 각 예측자의 상대적 중요도를 나타내는 차트를 표시합니다. 추가 정보는 예측변수 중요도의 내용을 참조하십시오.

조건부 확률 왼쪽 분할창에서 노드 또는 미니 분포 그래프를 선택하면 오른쪽 분할창에 연관된 조건부 확률 테이블이 표시됩니다. 이 테이블에는 상위 노드의 개별 값 조합과 각 노드 값의 조건부 확률 값이 있습니다. 또한 상위 노드의 개별 값 조합과 각 레코드 값에 대해 관측한 레코드 수도 있습니다.

① 베이지안 신경망 모델 설정

베이지안 신경망 모델 너깃의 설정 탭은 작성된 모델을 수정하기 위한 옵션을 지정합니다. 예를 들어, 베이지안 네트워크 노드를 사용하여 동일한 데이터와 설정으로 여러 다른 모델을 작성한 후 각 모델에서 이 탭을 사용하여 결과에 미치는 영향을 보기 위해 설정을 약간 수정할 수 있습니다.

 **참고:** 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능합니다.

원시 성향 스코어 계산. (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 이외에도 스코어링 중에 생성할 수 있는 기타 예측 및 신뢰도 값이 있습니다.

수정된 성향 스코어 계산. 원시 성향 스코어는 훈련 데이터에만 기반하며, 이 데이터의 과적합을 위한 많은 모델의 경향으로 인해 지나치게 낙관적일 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 성능을 평가하여 보완하려고 합니다. 이 옵션에서는 파티션 필드가 스트림에 정의되어야 하고 모델 생성 전에 모델링 노드에서 수정된 성향 스코어가 사용 가능해야 합니다.

모든 확률 추가 출력 필드의 각 범주에 대한 확률을 노드에서 처리하는 각 레코드에 추가하는지 여부를 지정합니다. 이 옵션을 선택하지 않으면 예측 범주의 확률만 추가됩니다.

이 선택란의 기본 설정은 모델링 노드의 고급 탭에 있는 해당 선택란을 통해 판별됩니다. 자세한 정보는 베이지안 네트워크 노드 고급 옵션의 내용을 참조하십시오.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

② 베이지안 신경망 모델 요약

모델 너깃의 요약 탭에서는 모델 자체(분석), 모델에 사용된 필드(필드), 모델 작성 시 사용된 설정(작성 설정), 모델 훈련(훈련 요약)에 대한 정보를 표시합니다.

먼저 노드를 찾아볼 때 요약 탭 결과를 접습니다. 관심이 있는 결과를 보려면 항목 왼쪽에 있는 펼치기 제어를 사용하여 펼치거나 **모두 펼치기** 단추를 클릭하여 모든 결과를 표시합니다. 보기를 마친 경우 결과를 숨기려면 펼치기 제어를 사용하여 숨기려는 특정 결과를 접거나 **모두 접기** 단추를 클릭하여 모든 결과를 접으십시오.

분석. 특정 모델에 대한 정보를 표시합니다.

필드. 모델을 작성할 때 목표로 사용되는 필드와 입력을 나열합니다.

작성 설정. 모델을 작성할 때 사용되는 설정에 대한 정보를 포함합니다.

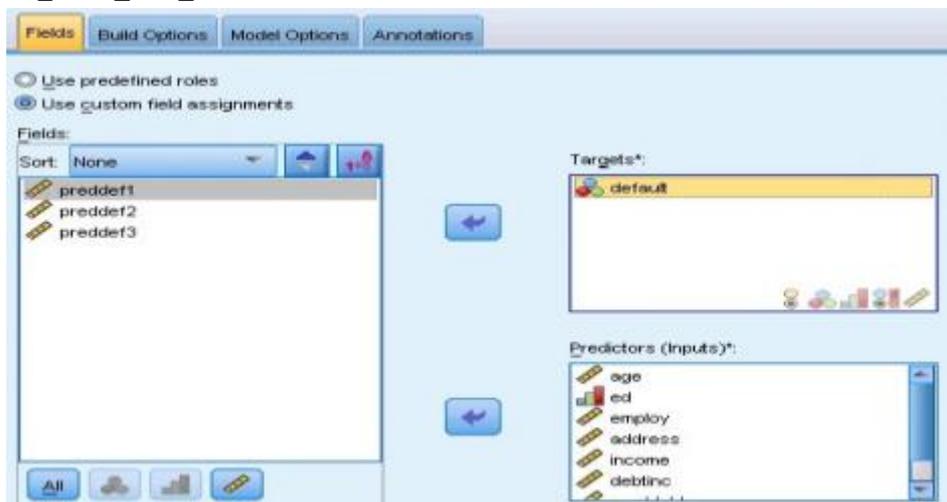
훈련 요약. 모델 유형, 이를 작성하는 데 사용된 스트림, 이를 작성한 사용자, 작성 시점, 모델 작성 시 경과 시간을 표시합니다.

6) 신경망

신경망은 모델 구조 및 가정에서 최소의 요구를 가지고 있는 광범위한 예측 모델과 근사할 수 있습니다. 관계 양식은 학습 프로세스 동안 판별됩니다. 목표와 예측변수 사이의 선형 관계가 적절한 경우, 신경망의 결과는 거의 전형적인 선형 모델의 결과와 근사해야 합니다. 비선형 관계가 더 적절한 경우, 신경망은 자동으로 "올바른" 모델 구조와 근사하게 됩니다.

이 신축성에 대한 절충은 신경망이 쉽게 해석 가능하지 않다는 것입니다. 목표 및 예측변수 사이의 관계를 생성하는 기본적인 프로세스를 설명하는 경우, 한층 전형적인 통계 모델을 사용하는 것이 좋습니다. 그러나 모델 해석가능성이 중요하지 않으면, 신경망을 사용하여 좋은 예측을 확보할 수 있습니다.

그림 1. 필드 탭



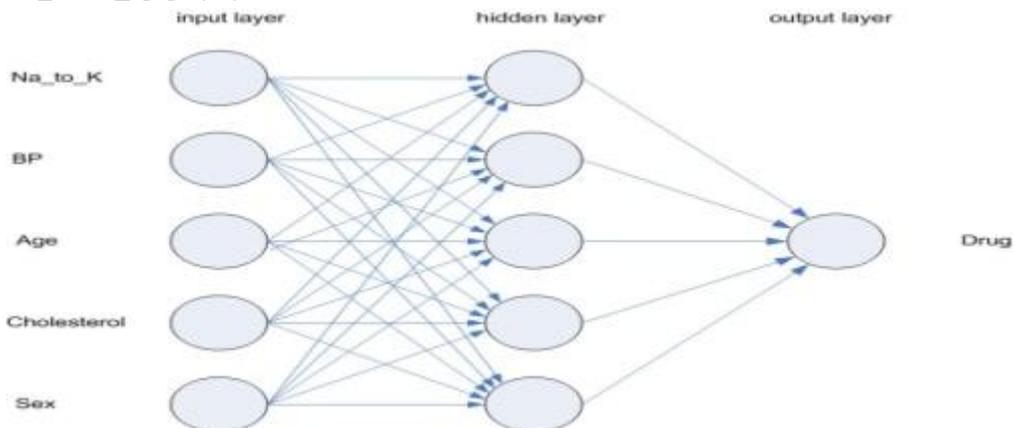
필드 요구 사항. 최소 하나의 목표와 하나의 입력이 있어야 합니다. 둘 다 또는 없음으로 설정된 필드는 무시됩니다. 목표 또는 예측변수(입력)에 대해 어떤 측정 수준 제한도 없습니다. 자세한 정보는 모델링 노트 필드 옵션의 내용을 참조하십시오.

모델 작성 동안 신경망에 초기 가중치가 지정되므로 데이터의 필드 순서에 따라 최종 모델이 생성됩니다. SPSS® Modeler가 학습을 위해 신경망에 데이터를 제시하기 전에 필드 이름을 기준으로 자동으로 정렬합니다. 즉, 데이터 업스트림에서 필드의 순서를 명시적으로 변경해도 모델 작성기에서 난수 시드가 설정될 때 생성된 신경망 모델에는 영향을 미치지 않습니다. 그러나 정렬 순서를 변경하는 방법으로 입력 필드 이름을 변경하면 모델 작성기에서 난수 시드가 설정된 경우에도 다른 신경망 모델이 생성됩니다. 필드 이름에 다른 정렬 순서가 지정되어도 모델 품질에 유의적으로 영향을 미치지 않습니다.

(1) 신경망 모델

신경망은 신경계가 작동하는 방식의 단순 모델입니다. 기본 단위는 뉴런이며, 일반적으로 다음 그림에 표시된 것처럼 레이어로 조직됩니다.

그림 1. 신경망의 구조



신경망은 인간 두뇌가 정보를 처리하는 방식의 단순화된 모델입니다. 뉴런의 추상 버전과 유사한 수많은 상호 연결된 처리 단위를 시뮬레이션하여 작동합니다.

처리 단위는 레이어에 배열됩니다. 신경망에는 일반적으로 세 개의 부분이 있습니다. 입력 필드를 나타내는 단위가 있는 **입력층**과, 하나 이상의 은닉층, 대상 필드를 나타내는 단위가 있는 **출력층**입니다. 단위는 다양한 연결 세기(또는 **가중치**)로 연결됩니다. 입력 데이터는 첫 번째 레이어에 표시되고, 값은 각각의 뉴런에서 다음 레이어의 모든 뉴런으로 전파됩니다. 결국, 결과는 출력층에서 전달됩니다.

네트워크는 개별 레코드를 조사하고, 각각 레코드에 대한 예측을 생성한 후, 부정확한 예측을 할 때마다 가중치를 조정하여 훈련합니다. 이 프로세스는 여러 번 반복되며, 네트워크는 하나 이상의 중지 기준이 충족될 때까지 해당 예측을 계속 향상시킵니다.

처음에, 모든 가중치는 임의적이고, 넷에서 나오는 응답은 아마도 의미가 없을 수 있습니다. 네트워크는 **학습(training)**을 통해 학습(learning)합니다. 출력이 알려지는 예제는 반복해서 네트워크에 제시되고, 네트워크가 제공하는 응답은 알려진 결과와 비교됩니다. 이 비교의 정보는 점차로 가중치를 변경하면서, 네트워크를 통해 뒤로 전달됩니다. 학습이 진행되면서, 네트워크는 알려진 결과를 복제할 때 점차적으로 정확하게 됩니다. 학습되면, 네트워크는 결과를 알 수 없는 나중 케이스에도 적용 가능하게 됩니다.

(2) 레거시 스트림이 있는 신경망 사용

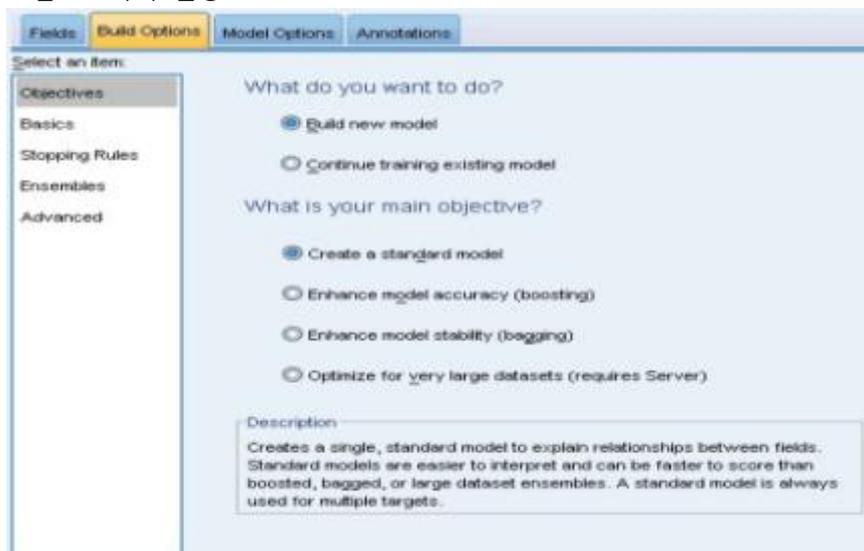
IBM® SPSS® Modeler 버전 14에서는 매우 큰 데이터 세트에 대한 최적화 및 부스팅(boosting) 및 배깅(bagging) 기술을 지원하는 새 신경망 노드를 도입했습니다. 이전 노드가 포함된 기존 스트림은 이후 릴리스에서도 계속 모델을 작성하고 스코어링합니다. 그러나 이 지원은 나중 릴리스에서 제거될 예정이므로, 새 버전을 사용할 것을 권장합니다.

버전 13부터, 알 수 없는 값(즉, 학습 데이터에 없는 값)은 더 이상 결측값으로 자동 처리되지 않고, \$null\$ 값으로 스코어링됩니다. 따라서 버전 13 이상에서 더 오래된(13 이전) 신경망 모델을 사용하여 알 수 없는 값이 있는 필드를 넘어 아닌 것으로 스코어링하려면, 알 수 없는 값을 결측 값으로 표시해야 합니다(예를 들어, 유형 노드를 사용하여).

호환성을 위해, 여전히 이전 노드를 포함하는 레거시 스트림은 **도구 > 스트림 특성 > 옵션에서 세트 크기 제한** 옵션을 사용할 수 있습니다. 이 옵션은 버전 14 이상의 코호넨 넷 및 K-Means 노드에만 적용됩니다.

(3) 목적 (신경망)

그림 1. 목적 설정



원하는 작업

- **새 모델 작성.** 완전히 새 모델을 작성합니다. 노드의 일반적인 작업입니다.
- **기존 모델 학습 계속.** 노드에 의해 성공적으로 작성된 마지막 모델로 계속 학습합니다. 원래 데이터에 액세스할 필요 없이 기존 모델을 업데이트하거나 새로 고칠 수 있으며 새 레코드 또는 업데이트된 레코드만 스트림에 입력되므로 상당히 빠르게 수행할 수 있습니다. 이전 모델에 대한 세부 사항이 모델링 노드와 함께 저장되어 스트림 또는 모델 팔레트에서 이전 모델 너깃을 더 이상 사용할 수 없는 경우에도 이 옵션을 사용할 수 있습니다.

 **참고:** 이 옵션이 활성화되면 필드 및 작성 옵션 탭의 다른 모든 제어가 비활성화됩니다.

원하는 기본 목적 적절한 목적을 선택하십시오.

- **표준 모델 작성.** 이 방법은 예측자를 사용하여 목표를 예측하는 단일 모델을 작성합니다. 일반적으로 표준 모델은 부스팅되었거나 배깁되었거나 큰 데이터 세트 앙상블보다 해석하기 쉽고 스코어링이 빠릅니다.

 **참고:** 분할 모델에 대해 기존 모델 계속 훈련과 함께 이 옵션을 사용하려면 Analytic Server에 연결되어야 합니다.

- **모형 정확도(부스팅) 개선.** 이 방법은 더 정확한 예측을 하기 위해 모델의 시퀀스를 생성하는 부스팅을 사용하여 앙상블 모델을 작성합니다. 앙상블은 표준 모델보다 작성 및 스코어 계산하는 데 오래 걸릴 수 있습니다.

부스팅은 일련의 "구성요소 모델"(각각 전체 데이터 세트에서 작성되는)을 생성합니다. 각각의 연속 구성요소 모델을 작성하기 전에, 레코드는 이전 구성요소 모델의 잔차를 기반으로 가중치가 부여됩니다. 잔차가 큰 케이스에는 다음 구성요소 모델이 해당 레코드 예측에 제대로 초점을 맞추도록 상대적으로 높은 분석 가중치가 부여됩니다. 이 구성요소 모델들은 함께 앙상블 모델을 형성합니다. 앙상블 모델은 결합 규칙을 사용하여 새 레코드를 스코어링합니다. 사용 가능한 규칙은 목표의 측정 수준에 따라 다릅니다.

- **모델 안정성(배깅) 개선.** 이 방법은 더 신뢰할 만한 예측을 하기 위해 여러 모델을 생성하는 배깅(붓스트랩 집계)을 사용하여 앙상블 모델을 작성합니다. 앙상블은 표준 모델보다 작성 및 스코어 계산하는 데 오래 걸릴 수 있습니다.

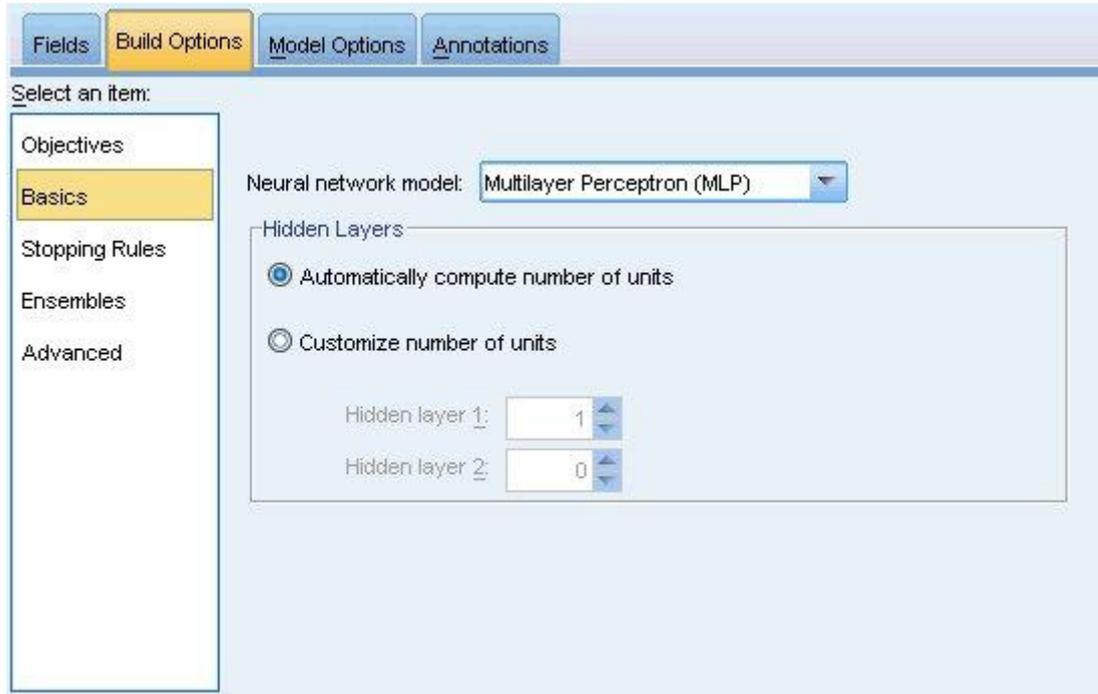
붓스트랩 통합(배깅)은 원래 데이터 세트에서 복원 표본추출하여 훈련 데이터 세트의 복제를 생성합니다. 이는 원래 데이터 세트와 동일한 크기의 붓스트랩 표본을 작성합니다. 그리고 나서 "구성요소 모델"이 각 복제에 작성됩니다. 이 구성요소 모델들은 함께 앙상블 모델을 형성합니다. 앙상블 모델은 결합 규칙을 사용하여 새 레코드를 스코어링합니다. 사용 가능한 규칙은 목표의 측정 수준에 따라 다릅니다.

- **매우 큰 데이터 세트를 위한 모델 작성.** 이 방법은 데이터 세트를 별도의 데이터 블록으로 분할하여 앙상블 모델을 작성합니다. 위의 모델을 작성하기에 데이터 세트가 너무 크거나 증분 모델 작성의 경우 이 옵션을 선택하십시오. 이 옵션은 작성하는 데 시간이 덜 걸릴 수 있지만 표준 모델보다 스코어를 계산하는 데 더 오래 걸릴 수 있습니다.

여러 목표가 있을 때, 이 방법은 선택된 목적에 상관없이, 표준 모델을 작성합니다.

(4) 기본 (신경망)

그림 1. 기본 설정



신경망 모델. 모델 유형은 네트워크가 은닉층을 통해 목표에 예측변수를 연결하는 방법을 판별합니다. **다중 레이어 퍼셉트론(MLP)**은 학습 및 스코어링 시간이 더 소요될 수 있는 한층 복잡한 관계에 대해 허용됩니다. **방사형 기저함수(RBF)**에서는 MLP와 비교하여 예측력이 감소될 수 있고 학습 및 스코어링 시간이 낮아질 수 있습니다.

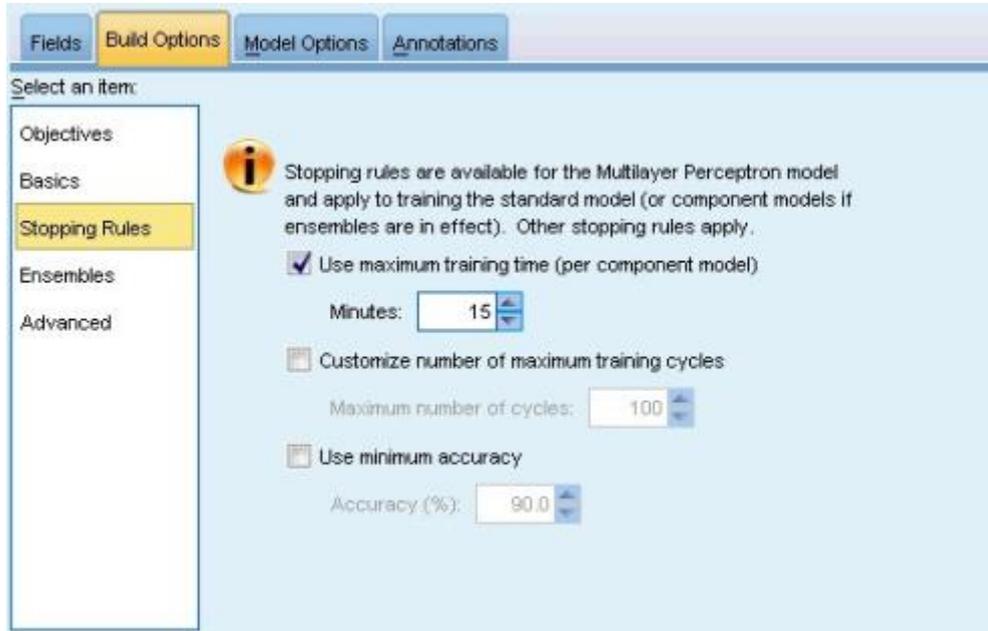
은닉층. 신경망의 은닉층에는 관측할 수 없는 단위가 포함됩니다. 각 은닉 단위의 값은 예측변수의 함수입니다. 함수의 정확한 양식은 부분적으로 네트워크 유형에 따라 다릅니다. 다중 레이어 퍼셉트론에는 하나 또는 두 개의 은닉층이 포함될 수 있습니다. 방사형 기저함수 네트워크에는 하나의 은닉층이 있습니다.

- **단위 수 자동 계산.** 이 옵션은 하나의 은닉층으로 네트워크를 작성하고, 은닉층에서 "최상의" 노드 수를 계산합니다.
- **단위 수 사용자 정의.** 이 옵션을 사용하여 항상 각 은닉층의 단위 수를 지정할 수 있습니다. 첫 번째 은닉층에는 하나 이상의 단위가 있어야 합니다. 두 번째 은닉층에 대해 0개 단위를 지정하면 단일 은닉층으로 다중 레이어 퍼셉트론이 작성됩니다.

참고: 노드 수가 연속형 예측변수 수에 모든 범주(플래그, 명목형 및 순서) 예측변수에서 총 범주 수를 합한 값을 초과하지 않도록 값을 선택해야 합니다.

(5) 중지 규칙(신경망)

그림 1. 중지 규칙 설정



학습 다중 레이어 퍼셉트론 네트워크를 중지할 시기를 판별하는 규칙입니다. 이 설정은 방사형 기저함수 알고리즘이 사용될 때 무시됩니다. 학습은 최소 하나의 순환(데이터 전달)에서 진행되므로, 다음 기준에 따라 중지될 수 있습니다.

최대 학습 시간(구성요소 모델당) 사용. 실행할 알고리즘에 대한 최대 시간(분)을 지정할 것인지 여부를 선택하십시오. 0보다 큰 숫자를 지정하십시오. 앙상블 모델이 작성될 때, 이는 앙상블의 각 구성요소 모델에 대해 허용된 학습 시간입니다. 학습은 현재 순환을 완료하기 위해 지정된 시간 한계를 약간 넘어설 수 있습니다.

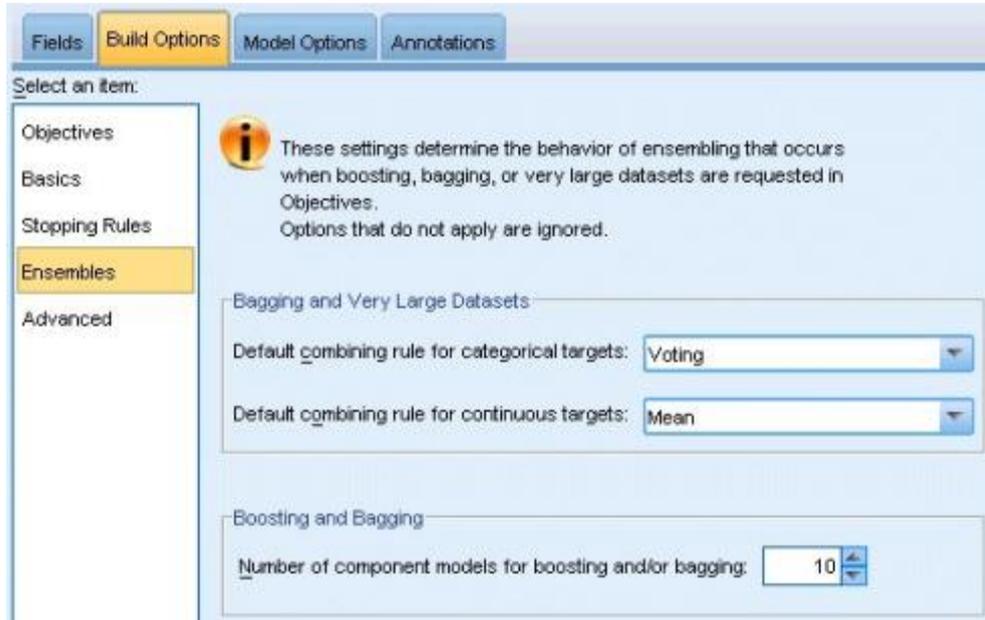
최대 학습 주기 수 사용자 정의. 허용되는 최대 학습 주기 수. 최대 주기 수를 초과하면, 학습이 중지됩니다. 0보다 큰 정수를 지정하십시오.

최소 정확도 사용. 이 옵션을 사용하면, 지정된 정확도가 될 때까지 학습이 계속됩니다. 이러한 상황은 발생하지 않을 수 있지만, 언제든지 학습을 중단하고 지금까지 달성한 최상의 정확도로 넷을 저장할 수 있습니다.

과적합 방지 세트의 오류가 각각의 주기 후에 감소하지 않는 경우, 학습 오류의 상대적 변화가 작은 경우, 또는 현재 학습 오류의 비율이 초기 오류와 비교하여 작은 경우 학습 알고리즘도 중지됩니다.

(6) 앙상블 (신경망)

그림 1. 앙상블 설정



이 설정은 부스팅, 배깅 또는 아주 큰 데이터 세트가 목표에서 요청될 때 발생하는 앙상블 동작을 결정합니다. 선택한 목표에 해당하지 않는 옵션은 무시됩니다.

배깅 및 아주 큰 데이터 세트. 앙상블을 스코어링할 때 앙상블 스코어값을 계산하기 위해 기본 모델에서 예측값을 조합하는 데 사용되는 규칙입니다.

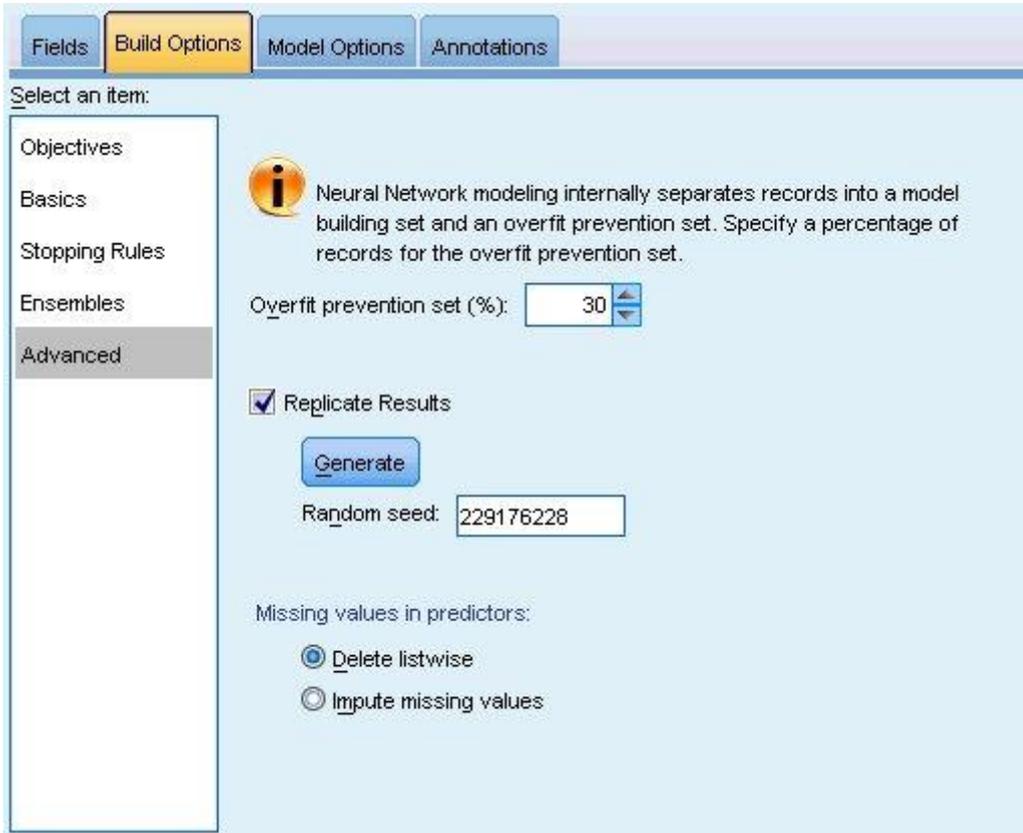
- **범주형 목표의 기본 결합 규칙.** 범주형 목표에 대한 앙상블 예측값은 투표, 최고 확률 또는 최고 평균 확률을 사용하여 조합될 수 있습니다. **투표**는 기본 모델에서 최고 확률을 가지는 범주를 선택합니다. **최고 확률**은 모든 기본 모델에서 단일 최고 확률을 획득하는 범주를 선택합니다. **최고 평균 확률**은 범주 확률이 기본 모델에서 평균이 될 때 최고값이 있는 범주를 선택합니다.
- **연속형 목표의 기본 결합 규칙.** 연속형 목표에 대한 앙상블 예측값은 기본 모델의 예측값 평균 또는 중앙값을 사용하여 조합될 수 있습니다.

모형 정확도를 향상시키는 것이 목표인 경우 결합 규칙 선택이 무시됨에 유의하십시오. 부스팅은 항상 가중 다수 투표를 사용하여 범주형 목표를 스코어링하고 가중 중앙값을 사용하여 연속형 목표를 스코어링합니다.

부스팅 및 배깅. 모형 정확도 또는 안정성을 향상시키는 것이 목표일 때 작성할 기본 모형 수를 지정하십시오. 배깅의 경우, 붓스트랩 표본의 수입니다. 양의 정수여야 합니다.

(7) 고급 (신경망)

그림 1. 고급 설정



고급 설정은 다른 설정 그룹에 적합하지 않은 옵션을 제어합니다.

과적합 방지 세트. 신경망 방법은 내부적으로 레코드를 모델 작성 세트와 과적합 방지 세트로 분리하며, 이는 방법에서 데이터에 우연 변동이 모델링되지 않도록 학습 동안 오류 추적에 사용되는 독립된 데이터 레코드 세트입니다. 레코드 퍼센트를 지정합니다. 기본값은 30입니다.

결과 복제. 난수 시드를 설정하면 분석을 복제할 수 있습니다. 정수를 지정하거나, **생성**을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 의사 난수 정수를 작성합니다. 기본적으로, 분석은 시드 229176228로 복제됩니다.

예측변수의 결측값. 결측값 처리 방법을 지정합니다. **목록별 삭제**는 모델 작성에서 예측변수의 결측값이 있는 레코드를 제거합니다. **결측값 대체**는 예측변수의 결측값을 바꾸고 분석에서 해당 레코드를 사용합니다. 연속형 필드는 최소 및 최대 관측값의 평균을 대치하고, 범주형 필드는 가장 빈번하게 발생하는 범주를 대치합니다. 필드 탭에 지정된 다른 필드에서 결측값이 있는 레코드는 항상 모델 작성 시 제거됨에 유의하십시오.

(8) 모델 옵션(신경망)

그림 1. 모델 옵션 탭

Model Name: Automatic Custom

Make Available for Scoring

i Predicted value and confidence are always available for scoring.

Confidence is based on:

The probability of the predicted value

The increase in probability from the next most likely value

Predicted probability for categorical targets

Maximum categories to save: 25

Propensity scores for flag targets

모델 이름. 목표 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 목표 필드 이름입니다. 목표가 여러 개일 경우, 모델 이름은 필드 이름이며 순서대로 앰퍼샌드로 연결됩니다. 예를 들어, *field1 field2 field3*가 목표가면, 모델 이름은 *field1 & field2 & field3*입니다.

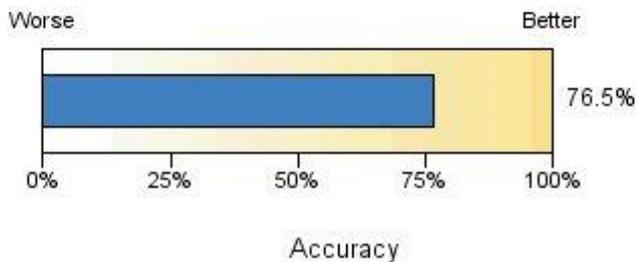
스코어링에 사용 가능. 모델이 스코어링되면 이 그룹에서 선택한 항목이 생성되어야 합니다. (모든 목표에 대한) 예측값과 (범주형 목표에 대한) 신뢰도는 모델이 스코어링될 때 항상 계산됩니다. 계산된 신뢰도는 예측값의 확률(가장 높은 예측 확률) 또는 가장 높은 예측 확률과 두 번째로 가장 높은 예측 확률의 차이를 기준으로 할 수 있습니다.

- **범주형 목표의 예측 확률.** 범주형 목표에 대한 예측 확률이 생성됩니다. 범주마다 하나의 필드가 작성됩니다.
- **플래그 목표를 위한 성향 스코어.** (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 모델은 원시 성향 스코어를 생성합니다. 파티션이 적용 중일 경우 모델은 검정 분할에 근거한 수정된 성향 스코어도 생성합니다.

(9) 모델 요약 (신경망)

그림 1. 신경망 모델 요약 보기

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4



모델 요약 보기는 신경망 예측 또는 분류 정확도를 한 눈에 파악할 수 있도록 요약한 스냅샷입니다.

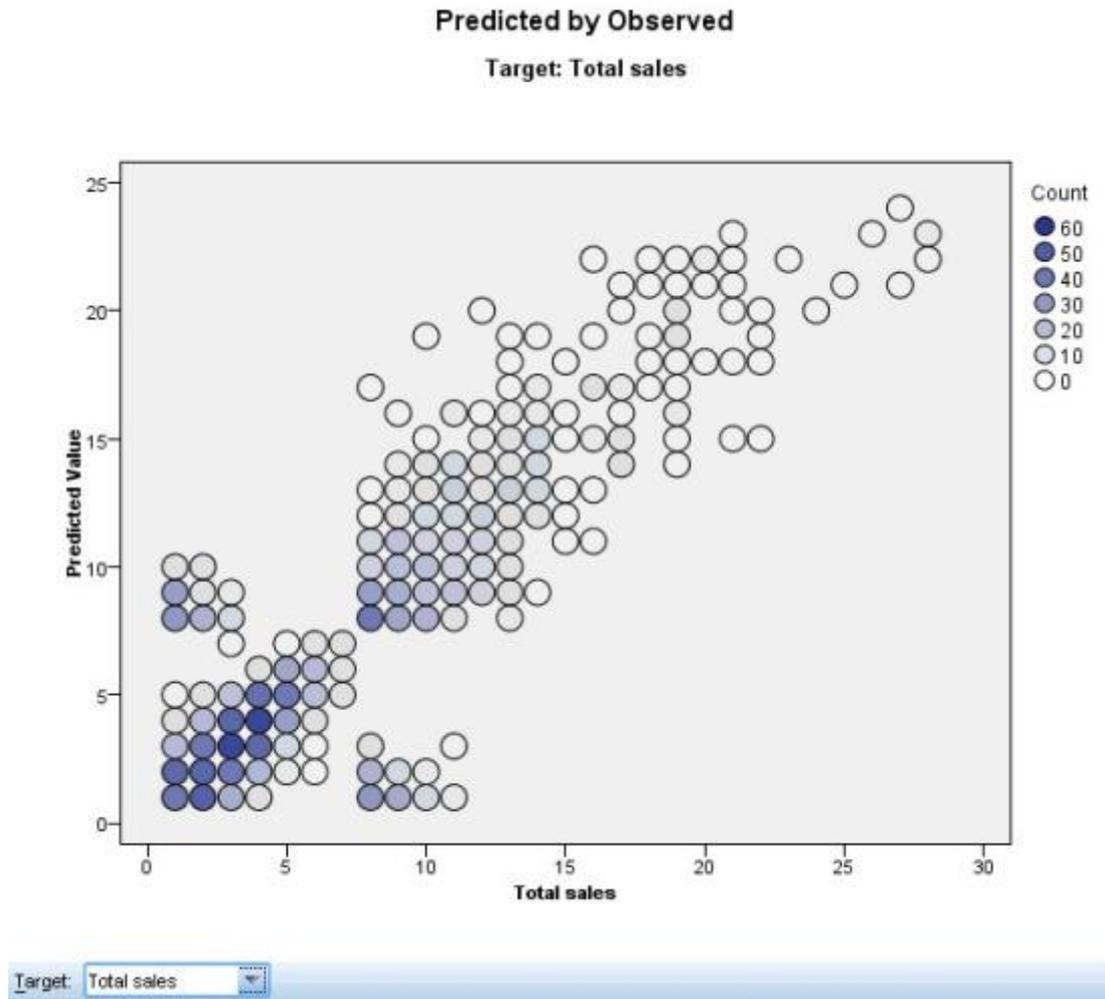
모형 요약. 테이블은 대상, 학습된 신경망의 유형, 학습을 중지한 중지 규칙(다중 레이어 퍼셉트론 네트워크를 학습한 경우 표시됨), 네트워크의 각 은닉층에서 뉴런의 수를 식별합니다.

신경망 품질. 차트는 최종 모델의 정확도를 표시하며 더 크게 표시된 것이 더 나은 형식입니다. 범주형 목표의 경우, 이는 단순히 예측값이 관측값과 일치하는 레코드의 퍼센트입니다. 연속형 대상의 경우 정확도가 R^2 값으로 제공됩니다.

다중 대상. 대상이 여러 개인 경우 각 대상은 테이블의 **대상** 행에 표시됩니다. 차트에 표시되는 정확도는 개별 목표 정확도의 평균입니다.

(11) 관측값 별 예측값 (신경망)

그림 1. 관측값 별 예측값 보기

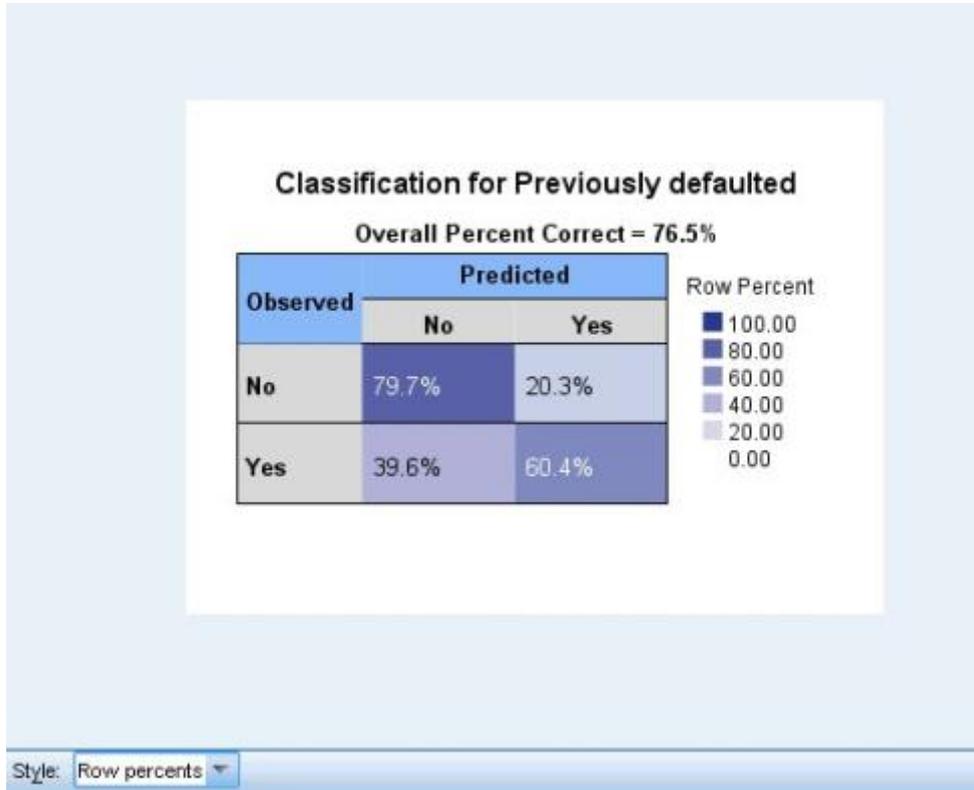


연속형 목표에 대해, 수직축에 예측값의 구간화된 산점도를 표시하고 수평축에 관측값을 표시합니다.

여러 목표. 여러 연속형 목표가 있는 경우, 각 목표가 별도의 차트에 표시되고 표시되는 목표를 제어하는 **목표** 드롭다운 목록이 있습니다.

(12) 분류 (신경망)

그림 1. 분류 보기, 행 퍼센트 유형



범주형 목표의 경우, 정확한 전체 퍼센트와 함께 관측값 대 예측값의 교차 분류를 히트 맵에 표시합니다.

테이블 유형. 다양한 표시 유형이 있으며, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **행 퍼센트.** 셀의 행 백분율(전체 행의 퍼센트로 표시되는 셀 개수)을 표시합니다. 이는 기본값입니다.
- **셀 개수.** 셀의 셀 개수를 표시합니다. 히트 맵의 음영이 행 퍼센트의 기준입니다.
- **히트 맵.** 셀의 값은 표시하지 않고 음영만 표시합니다.
- **압축.** 셀의 행 또는 열 머리말, 값을 표시하지 않습니다. 목표에 범주가 많은 경우에 유용할 수 있습니다.

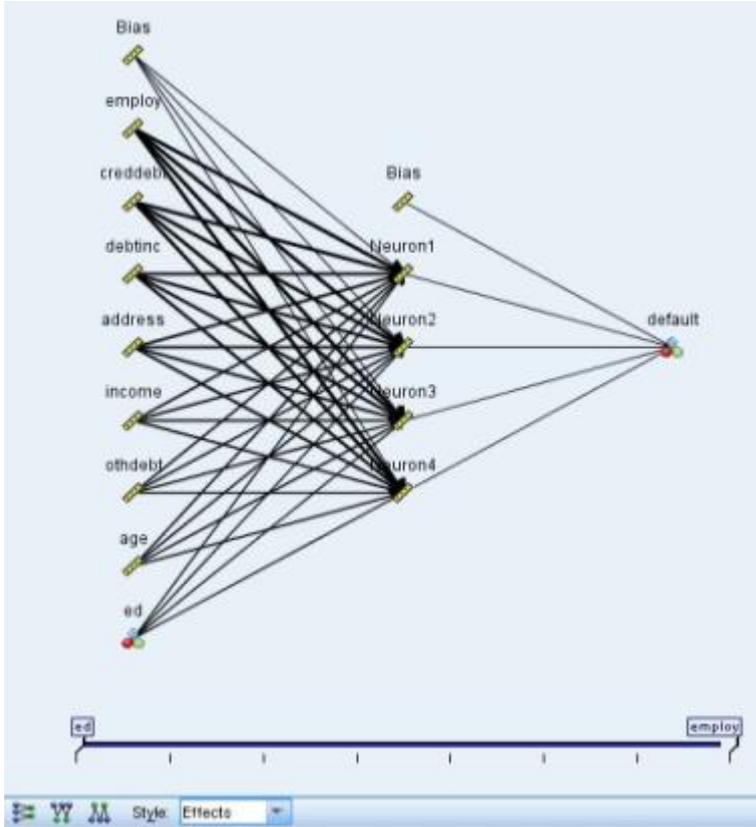
결측. 레코드에 목표의 결측값이 있으면 모든 유효한 행 아래의 (**결측**) 행에 표시됩니다. 결측값이 있는 레코드는 정확한 전체 퍼센트에 기여하지 않습니다.

여러 목표. 여러 범주형 목표가 있는 경우, 각 목표가 별도의 테이블에 표시되고 표시되는 목표를 제어하는 **목표** 드롭다운 목록이 있습니다.

큰 테이블. 표시된 목표에 범주가 100개 이상 있으면 테이블이 표시되지 않습니다.

(13) 네트워크 (신경망)

그림 1. 네트워크 보기, 왼쪽의 입력, 효과 유형



신경망의 그래픽 표현을 표시합니다.

차트 유형. 두 가지 표시 유형이 있으며, 유형 드롭 다운 목록에서 액세스할 수 있습니다.

- **효과.** 측정 척도가 연속형 또는 범주형 여부에 관계없이 다이어그램에서 각각의 예측변수 및 목표를 하나의 노드로 표시합니다. 이는 기본값입니다.
- **계수.** 범주형 예측변수 및 목표에 대해 여러 표시기 노드를 표시합니다. 계수 유형 다이어그램에 있는 연결 선의 색상은 추정되는 시냅스 가중값에 따라 지정됩니다.

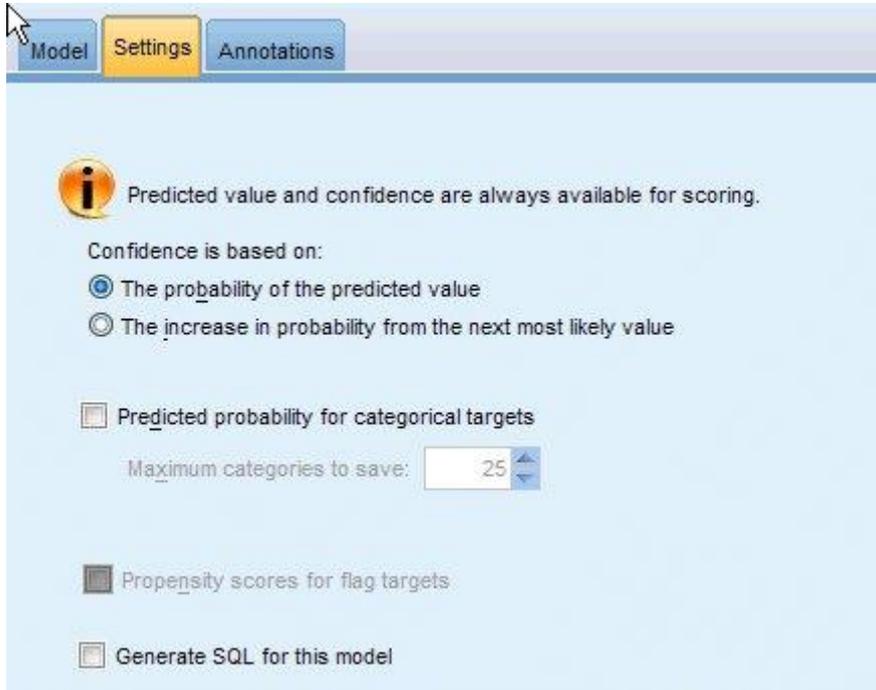
다이어그램 방향. 기본적으로, 네트워크 다이어그램은 왼쪽에서는 입력에, 오른쪽에서는 목표에 맞춰 배열됩니다. 도구 모음 제어를 사용하여, 입력이 위쪽에 있고 목표가 아래쪽에 있거나, 입력이 아래쪽에 있고 목표가 위쪽에 있도록 방향을 변경할 수 있습니다.

예측변수 중요도. 다이어그램의 연결선은 예측변수 중요도를 기준으로 가중되며 선 너비가 클수록 중요도가 큼니다. 네트워크 다이어그램에 표시되는 예측변수를 제어하는 예측변수 중요도 슬라이더가 도구 모음에 있습니다. 모델을 변경하지는 않지만 가장 중요한 예측변수에 집중할 수 있습니다.

여러 목표. 여러 개의 목표가 있을 경우 모든 목표가 차트에 표시됩니다.

(14) 설정 (신경망)

그림 1. 설정 탭



모델이 스코어링되면 이 탭에서 선택한 항목이 생성되어야 합니다. (모든 목표에 대한) 예측값과 (범주형 목표에 대한) 신뢰도는 모델이 스코어링될 때 항상 계산됩니다. 계산된 신뢰도는 예측값의 확률(가장 높은 예측 확률) 또는 가장 높은 예측 확률과 두 번째로 가장 높은 예측 확률의 차이를 기준으로 할 수 있습니다.

- 범주형 목표의 예측 확률. 범주형 목표에 대한 예측 확률이 생성됩니다. 범주마다 하나의 필드가 작성됩니다.
- 플래그 목표를 위한 성향 스코어. (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 모델은 원시 성향 스코어를 생성합니다. 파티션이 적용 중일 경우 모델은 검정 분할에 근거한 수정된 성향 스코어도 생성합니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우, 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.

이 모형의 SQL 생성 이 옵션을 선택하면 데이터베이스에서 모델 스코어를 계산할 원형 SQL을 생성합니다.

 **참고:** 이 옵션은 결과를 빠르게 제공할 수 있지만, 모델의 복잡도가 증가하여 원형 SQL의 크기와 복잡도도 증가할 수 있습니다.

데이터베이스 외부 스코어 선택된 경우, 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

7) 의사결정 목록

의사결정 목록 모델은 전체 표본에 상대적인 이분형(예 또는 아니오) 결과의 더 높거나 낮은 우도를 표시하는 **세그먼트** 또는 하위 그룹을 식별합니다. 예를 들어, 가장 덜 이탈할 것 같거나 특정 오피 또는 캠페인에 가장 우호적일 것 같은 고객을 찾을 수 있습니다. 의사결정 목록 뷰어는 모델을 완전히 제어해서 세그먼트를 편집하고, 비즈니스 규칙을 직접 추가하고, 각 세그먼트가 스코어링되는 방식을 지정하며, 모든 세그먼트에 대한 적중률을 최적화할 여러 다른 방식으로 모델을 사용자 정의할 수 있게 합니다. 이러한 이유로 이는 메일링 목록을 생성하거나 그렇지 않은 경우 특정 캠페인에 대해 목표화할 레코드를 식별할 때 특히 잘 맞습니다. 예를 들어, 여러 **마이닝 작업**을 사용하여 동일한 모델 내에서 고성능 및 저성능 세그먼트를 식별하고 스코어링 단계에서 각 세그먼트를 적합하게 포함 또는 제외시켜서 모델링 접근법을 조합할 수도 있습니다.

세그먼트, 규칙, 조건

모델은 세그먼트 목록으로 구성되며 각 세그먼트는 일치하는 레코드를 선택하는 규칙을 통해 정의됩니다. 주어진 규칙에는 여러 조건이 있을 수 있습니다. 예를 들어, 다음과 같습니다.

```
RFM_SCORE > 10 and  
MONTHS_CURRENT <= 9
```

규칙은 나열된 순서대로 적용됩니다. 즉, 첫 번째 일치하는 규칙이 주어진 레코드의 결과를 판별합니다. 독립적으로 작용하는 규칙이나 조건은 겹칠 수 있지만 규칙의 순서는 모호성을 해결합니다. 일치하는 규칙이 없으면 나머지 규칙에 레코드가 할당됩니다.

완벽한 스코어링 제어

의사결정 목록 뷰어를 통해 세그먼트를 보고, 수정 및 인식하고 스코어링 용도로 포함 또는 제외시킬 세그먼트를 선택할 수 있습니다. 예를 들어, 한 고객 그룹을 미래 오피에서 제외시키고 다른 고객 그룹을 포함하도록 선택한 후 이 선택이 전반적인 적중률에 영향을 미치는 방식을 즉시 확인할 수 있습니다. 의사결정 목록 모델은 포함된 세그먼트에 *예* 스코어를 리턴하고 나머지를 포함하여 다른 모든 세그먼트에는 *\$null\$*을 리턴합니다. 이러한 직접적 스코어링 제어를 통해 의사결정 목록 모델은 메일링 목록 생성에 이상적으로 되어 콜센터 또는 마케팅 애플리케이션과 같은 고객 관계 관리에 광범위하게 사용됩니다.

마이닝 작업, 속도, 선택

모델링 프로세스는 **마이닝 작업**으로 구동됩니다. 각 마이닝 작업은 효과적으로 새 모델링 실행을 시작하고 선택할 새 대체 모델 세트를 리턴합니다. 기본 작업은 의사결정 목록 노드의 초기 지정 사항을 기준으로 하지만 임의의 수의 사용자 정의 작업을 정의할 수 있습니다. 작업을 반복해서 적용할 수도 있습니다. 예를 들어, 전체 훈련 세트에 고확률 검색을 실행한 후 나머지에 저확률 검색을 실행해서 저성능 세그먼트를 제거할 수 있습니다.

데이터 선택

모델 작성 및 평가를 위한 데이터 선택과 사용자 정의 모델 속도를 정의할 수 있습니다. 예를 들어, 모델을 특정 지역에 맞게 조정하도록 마이닝 작업에 데이터 선택을 지정한 후 모델이 전체 국가에서 수행하는 정도를 평가하는 사용자 정의 속도를 작성할 수 있습니다. 마이닝 작업과 달리, 속도는 기본 모델을 변경하지 않지만 얼마나 잘 수행하는지 평가할 또 다른 렌즈를 제공합니다.

비즈니스 지식 추가

알고리즘을 통해 식별된 세그먼트를 세부적으로 조정하거나 확장해서 의사결정 목록 뷰어는 비즈니스 지식을 모델에 바로 통합시킬 수 있게 합니다. 모델이 생성한 세그먼트를 편집하거나 직접 지정한 규칙을 기반으로 추가 세그먼트를 추가할 수 있습니다. 그런 다음 변경사항을 적용하고 결과를 미리 볼 수 있습니다.

더 깊이 있는 통찰을 위해 Excel을 포함한 동적 링크를 사용하여 데이터를 Excel로 내보내서, 이를 사용하여 프리젠테이션 차트를 작성하고 복합 이익 및 ROI와 같이 모델을 작성하는 동안 의사결정 목록 뷰어에서 볼 수 있는 사용자 정의 속도를 계산할 수 있습니다.

예. 금융 기관의 마케팅 부서는 현재 각 고객에 올바른 오퍼를 매치하여 미래 캠페인에서 보다 수익성이 좋은 결과를 산출하고자 합니다. 의사결정 목록 모델을 사용하여 이전 홍보를 기반으로 가장 우호적으로 반응할 것 같은 고객 특성을 식별하고 결과에 따라 메일링 목록을 생성할 수 있습니다.

요구사항. 예측하려는 이분형 결과(예/아니오)를 나타내는 **플래그** 또는 **명목** 유형 측정 수준의 단일 범주형 대상 필드 및 최소 하나의 입력 필드. 대상 필드 유형이 명목인 경우 **적중** 또는 **반응**으로 처리할 단일 값을 수동으로 선택해야 합니다. 다른 모든 값은 통틀어 **적중 아님**으로 처리됩니다. 선택적 빈도 필드가 지정될 수도 있습니다. 연속 날짜/시간 필드는 무시됩니다. 연속 수치 범위 입력은 모델링 노드의 고급 탭에 지정된 알고리즘을 통해 자동으로 구간화됩니다. 구간화를 보다 세부적으로 제어하려면 업스트림 구간화 노드를 추가하고 측정 수준 **순서**의 입력으로 구간화된 필드를 사용하십시오.

(1) 의사결정 목록 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

모드. 모델 작성에 사용되는 방법을 지정합니다.

- **모델 생성.** 노드가 실행될 때 모델 팔레트에서 자동으로 모델을 생성합니다. 결과적인 모델은 스코어링 용도로 스트림에 추가될 수 있지만 더 이상의 편집은 불가능합니다.
- **대화형 세션 시작.** 대화형 의사결정 목록 뷰어 모델링(출력) 창을 열어 여러 대안 중에서 선택하고 모델이 점진적으로 성장 또는 수정되도록 알고리즘을 여러 다른 설정으로 반복해서 적용할 수 있게 합니다. 자세한 정보는 의사결정 목록 뷰어의 내용을 참조하십시오.
- **저장된 대화형 세션 정보 사용.** 이전에 저장된 설정을 사용하여 대화형 세션을 시작합니다. 대화형 설정은 메뉴 생성(모델이나 모델링 노드를 작성하기 위해) 또는 파일 메뉴(세션이 시작된 노드를 업데이트하기 위해)를 사용하여 의사결정 목록 뷰어로부터 저장될 수 있습니다.

목표 값. 모델링하려는 결과를 나타내는 대상 필드의 값을 지정합니다. 예를 들어, 대상 필드 이탈이 코딩된 0 = no 및 1 = yes인 경우 이탈할 것 같은 레코드를 표시하는 규칙을 식별하려면 1을 지정하십시오.

세그먼트 찾기. 목표변수 검색이 발생의 **높은 확률** 또는 **낮은 확률**을 찾아야 하는지 여부를 표시합니다. 찾은 후 실행은 모델을 개선하기 위한 유용한 방법이며 특히 나머지의 확률이 낮을 때 유용할 수 있습니다.

최대 세그먼트 수. 리턴할 세그먼트의 최대 수를 지정합니다. 상위 N 개의 세그먼트가 작성되며, 최상의 세그먼트는 확률이 가장 높거나 둘 이상 모델의 확률이 동일한 경우에는 범위가 가장 높은 세그먼트입니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

최소 세그먼트 크기. 아래 두 개의 설정은 최소 세그먼트 크기에 영향을 미칩니다. 두 값 중에서 더 큰 값이 우선합니다. 예를 들어, 퍼센트 값이 절대값보다 큰 수이면 퍼센트 설정이 우선합니다.

- **이전 세그먼트의 퍼센트로(%).** 최소 그룹 크기를 레코드의 퍼센트로 지정합니다. 허용된 최소 설정은 0이고 허용된 최대 설정은 99.9입니다.
- **절대값으로(N).** 최소 그룹 크기를 레코드의 절대 수로 지정합니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

세그먼트 규칙.

최대 속성 수. 세그먼트 규칙별 최대 조건 수를 지정합니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

- **속성 재사용 허용.** 사용할 경우 각 순환이 심지어 이전 순환에 사용한 속성을 포함하여 모든 속성을 고려할 수 있습니다. 세그먼트에 대한 조건은 순환에 작성되고 각 순환은 새 조건을 추가합니다. 순환 수는 **최대 속성 수** 설정을 사용하여 정의됩니다.

새 조건의 신뢰구간(%). 세그먼트 유의성을 검정하기 위한 신뢰수준을 지정합니다. 이 설정은 세그먼트별 조건 수 규칙 외에 리턴되는 세그먼트 수에(있는 경우) 상당한 역할을 합니다. 값이 높을수록 리턴되는 결과 세트는 더 작습니다. 허용된 최소 설정은 50이고 허용된 최대 설정은 99.9입니다.

(2) 의사결정 목록 노드 고급 옵션

고급 옵션으로 모델 작성 프로세스를 미세 조정할 수 있습니다.

구간화 방법. 연속형 필드(동일한 개수나 동일한 너비) 구간화에 사용되는 방법입니다.

구간 수. 연속형 필드에 작성할 구간 수입니다. 허용된 최소 설정은 2이고 최대 설정은 없습니다.

모델 검색 범위. 다음 순환에 사용할 수 있는 순환별 모델 결과의 최대 수입니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

규칙 검색 범위. 다음 순환에 사용할 수 있는 순환별 규칙 결과의 최대 수입니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

구간 병합 요인. 이웃 항목과 병합되었을 때 증가하는 세그먼트의 최소 크기입니다. 허용된 최소 설정은 1.01이고 최대 설정은 없습니다.

- **조건의 결측값 허용.** True는 규칙의 IS MISSING 검정을 허용합니다.
 - **중간 결과 삭제.** True일 경우 검색 프로세스의 최종 결과만 리턴됩니다. 최종 결과는 검색 프로세스에서 더 이상 세분화되지 않는 결과입니다. False이면 중간 결과도 리턴됩니다.
- 최대 대안 수.** 마이닝 작업을 실행할 때 리턴될 수 있는 최대 대안 수를 지정합니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

마이닝 작업은 실제 대안 수만 지정된 최대값까지 리턴함에 유의하십시오. 예를 들어, 최대값이 100으로 설정되어 있고 대안이 3개만 있으면 이 3개만 표시됩니다.

(3) 의사결정 목록 모델 너깃

모델은 세그먼트 목록으로 구성되며 각 세그먼트는 일치하는 레코드를 선택하는 규칙을 통해 정의됩니다. 모델을 생성하기 전에 세그먼트를 쉽게 보거나 수정하고 포함 또는 제외시킬 세그먼트를 선택할 수 있습니다. 스코어링에 사용할 경우 의사결정 목록 모델은 포함된 세그먼트에 *예*를 리턴하고 나머지를 포함한 다른 모든 것에는 *\$null\$*을 리턴합니다. 이러한 직접적인 스코어링 제어를 통해 의사결정 목록 모델은 이상적으로 메일링 목록을 생성하며 콜센터 또는 마케팅 애플리케이션을 포함한 고객 관계 관리에 광범위하게 사용됩니다.

의사결정 목록 모델을 포함한 스트림을 실행할 때에는 노드가 스코어 즉, 포함된 필드의 경우 *1(예를 의미함)* 또는 제외된 필드는 *\$null\$*, 레코드가 있는 세그먼트의 확률(적중률), 세그먼트의 ID 번호를 포함한 세 가지 새 필드를 추가합니다. 새 필드의 이름은 예측 중인 출력 필드의 이름에서 파생되며 스코어의 경우 *\$D-*, 확률은 *\$DP-*, 세그먼트 ID의 경우에는 *\$DI-* 접두문자가 붙습니다.

모델은 작성될 때 지정된 목표 값을 기준으로 하여 스코어링됩니다. *\$null\$*로 스코어링되도록 수동으로 세그먼트를 제외시킬 수 있습니다. 예를 들어, 적중률이 평균 미만인 세그먼트를 찾기 위해 낮은 확률 검색을 실행할 경우 세그먼트를 제외하지 않으면 이 “낮은” 세그먼트가 *예*로 스코어링됩니다. 필요에 따라 파생 또는 채움 노드를 사용하여 널이 *아니오*로 다시 코딩될 수 있습니다.

PMML

의사결정 목록 모델은 “첫 번째 적중” 선택 기준의 PMML RuleSetModel로 저장될 수 있습니다. 하지만 모든 규칙의 스코어가 동일할 것으로 예상됩니다. 목표 필드 또는 목표 값의 변경을 허용하려면 여러 규칙 세트 모델을 한 파일에 저장해서 순서대로 적용할 수 있습니다. 그러면 첫 번째 모델에 일치하지 않는 케이스가 두 번째에 전달되는 식으로 진행됩니다. 알고리즘 이름 *DecisionList*는 이러한 비표준 작동을 표시하는 데 사용되며 이 이름의 규칙 세트 모델만 의사결정 목록 모델로 인식되고 그렇게 스코어링됩니다.

① 의사결정 목록 모델 너깃 설정

의사결정 목록 모델 너깃의 설정 탭으로 성향 스코어를 사용하고 SQL 최적화를 사용 또는 사용하지 않게 설정할 수 있습니다. 이 탭은 모델 너깃을 스트림에 추가한 후에만 사용 가능합니다.

원시 성향 스코어 계산. (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 이외에도 스코어링 중에 생성할 수 있는 기타 예측 및 신뢰도 값이 있습니다.

수정된 성향 스코어 계산. 원시 성향 스코어는 훈련 데이터에만 기반하며, 이 데이터의 과적합을

위한 많은 모델의 경향으로 인해 지나치게 낙관적일 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 성능을 평가하여 보완하려고 합니다. 이 옵션에서는 파티션 필드가 스트림에 정의되어야 하고 모델 생성 전에 모델링 노드에서 수정된 성향 스코어가 사용 가능해야 합니다.

이 유형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링** 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **이 유형의 SQL 생성** 이 옵션을 선택하면 데이터베이스에서 모델 스코어를 계산할 원형 SQL을 생성합니다.

참고: 이 옵션은 결과를 빠르게 제공할 수 있지만, 모델의 복잡도가 증가하여 원형 SQL의 크기와 복잡도도 증가할 수 있습니다.

- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

(4) 의사결정 목록 뷰어

사용이 간편한 작업 기반의 의사결정 목록 뷰어 그래픽 인터페이스는 모델 작성 프로세스의 복잡함을 제거하여 사용자가 상세성이 낮은 수준인 데이터 마이닝 기법에서 벗어나서, 목표 설정, 목표 그룹 선택, 결과 분석, 최적 모델 선택과 같이 사용자 개입이 필요한 분석 파트에 전념할 수 있도록 합니다.

① 작업 모델 분할창

작업 모델 분할창은 마이닝 작업 및 기타 조치를 포함하여, 작업 모델에 적용되는 현재 모델을 표시합니다.

ID. 순차 세그먼트 순서를 식별합니다. 모델 세그먼트는 ID 번호에 따라 순차적으로 계산됩니다.

세그먼트 규칙. 세그먼트 이름 및 정의된 세그먼트 조건을 제공합니다. 기본적으로 세그먼트 이름은 심표를 구분 문자로 지정해서 조건에 사용하는 필드 이름 또는 연결된 필드 이름입니다.

스코어. 예측하려는 필드를 나타내며, 값이 다른 필드(예측변수)의 값에 관련된 것으로 추정됩니다.

참고: 다음 옵션은 모델 속도 구성 대화 상자를 통해 표시 여부가 토글될 수 있습니다.

범위. 원형 차트는 전체 범위와 관련이 있는 각 세그먼트의 범위를 시각적으로 식별합니다.

범위(n). 전체 범위와 관련이 있는 각 세그먼트의 범위를 나열합니다.

빈도. 전체에 관련하여 수신된 적중 수를 나열합니다. 예를 들어, 전체가 79이고 빈도가 50이면 79 중에서 50이 선택한 세그먼트에 반응했음을 의미합니다.

확률. 세그먼트 확률을 표시합니다. 예를 들어, 전체가 79이고 빈도가 50이면 세그먼트의 확률이 63.29%(50을 79로 나눔)라는 의미입니다.

오류. 세그먼트 오류를 표시합니다.

분할창의 맨 아래에 있는 정보는 전체 모델의 범위, 빈도, 확률을 표시합니다.

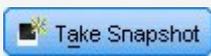
작업 모델 도구 모음

작업 모델 분할창은 도구 모음을 통해 다음 기능을 제공합니다.

참고: 일부 기능은 모델 세그먼트를 마우스 오른쪽 단추로 클릭해서도 사용 가능합니다.

표 1. 작업 모델 도구 모음 단추

도구 모음 단추	설명
	새 모델 너깃을 작성하기 위한 옵션을 제공하는 새 모델 생성 대화 상자를 실행합니다.
	대화형 세션의 현재 상태를 저장합니다. 마이닝 작업, 모델 스냅샷, 데이터 선택, 사용자 정의 속도를 포함한 현재 설정으로 의사결정 목록 모델링 노드가 업데이트됩니다. 세션을 이 상태로 복원하려면 모델링 노드의 모델 탭에서 저장된 세션 정보 사용 상자를 선택하고 실행 을 클릭하십시오.
	모델 속도 구성 대화 상자를 표시합니다. 자세한 정보는 모델 속도 구성의 내용을 참조하십시오.

도구 모음 단추	설명
	데이터 선택 구성 대화 상자를 표시합니다. 자세한 정보는 데이터 선택 구성의 내용을 참조하십시오.
	스냅샷 탭을 표시합니다. 자세한 정보는 스냅샷 탭의 내용을 참조하십시오.
	대안 탭을 표시합니다. 자세한 정보는 대안 탭의 내용을 참조하십시오.
	현재 모델 구조의 스냅샷을 찍습니다. 스냅샷은 스냅샷 탭에 표시되며 일반적으로 모델 비교 용도로 사용됩니다.
	새 모델 세그먼트를 작성하기 위한 옵션을 제공하는 세그먼트 삽입 대화 상자를 실행합니다.
	모델 세그먼트에 조건을 추가하거나 이전에 정의된 모델 세그먼트 조건을 변경하기 위한 옵션을 제공하는 세그먼트 규칙 편집 대화 상자를 실행합니다.
	선택한 세그먼트를 모델 계층에서 위로 이동시킵니다.
	선택한 세그먼트를 모델 계층에서 아래로 이동시킵니다.
	선택한 세그먼트를 삭제합니다.
	선택한 세그먼트를 모델에 포함할지 여부를 토글합니다. 제외할 경우 세그먼트 결과가 나머지에 추가됩니다. 세그먼트를 재활성화할 옵션이 있다는 점에서 이는 세그먼트 삭제와 차이가 있습니다.

② 대안 탭

세그먼트 찾기를 클릭하면 생성되는 대안 탭은 작업 모델 분할창의 선택한 모델 또는 세그먼트에 대한 대체 마이닝 결과를 모두 나열합니다.

대체를 작업 모델로 올리려면 필요한 대체를 강조 표시하고 **로드**를 클릭하십시오. 작업 모델 분할창에 대체 모델이 표시됩니다.

참고: 대안 탭은 의사결정 목록 모델링 노드 고급 탭에서 **최대 대안 수**를 둘 이상의 대체를 작성하도록 설정한 경우에만 표시됩니다.

생성된 각 모델 대체는 특정 모델 정보를 표시합니다.

이름. 각 대체는 순차적으로 번호가 지정됩니다. 일반적으로 첫 번째 대체가 최상의 결과를 포함합니다.

목표. 목표 값을 표시합니다. 예를 들어, 1은 "참"과 같습니다.

세그먼트 수. 대체 모델에 사용되는 세그먼트 규칙의 수입입니다.

범위. 대체 모델의 범위입니다.

빈도. 전체에 관련된 적중 수입입니다.

확률. 대체 모델의 확률 퍼센트를 표시합니다.

참고: 대체 결과는 모델과 함께 저장되지 않으며 결과는 활성 세션 중에만 유효합니다.

③ 스냅샷 탭

스냅샷은 특정 시점의 모델 보기입니다. 예를 들어, 다른 대체 모델을 작업 모델 분할창으로 로드하려 하지만 현재 모델에 대한 작업을 잃고 싶지 않을 때 모델 스냅샷을 찍을 수 있습니다. 스냅샷 탭은 임의의 수의 작업 모델 상태에 대해 수동으로 찍은 모든 모델 스냅샷을 나열합니다.

참고: 스냅샷은 모델과 함께 저장됩니다. 첫 번째 모델을 로드할 때 스냅샷을 찍을 것을 권장합니다. 그러면 이 스냅샷이 원래 모델 구조를 유지하게 되어 사용자가 언제나 원래 모델 상태로 돌아갈 수 있습니다. 생성된 스냅샷 이름은 생성된 시기를 나타내는 시간소인으로 표시됩니다.

모델 스냅샷 작성

1. 작업 모델 분할창에 표시할 적합한 모델/대체를 선택하십시오.
2. 작업 모델에 필요한 변경을 수행하십시오.
3. **스냅샷 생성**을 클릭하십시오. 스냅샷 탭에 새 스냅샷이 표시됩니다.

이름. 스냅샷 이름입니다. 스냅샷 이름을 두 번 클릭해서 스냅샷 이름을 변경할 수 있습니다.

목표. 목표 값을 표시합니다. 예를 들어, 1은 "참"과 같습니다.

세그먼트 수. 모델에 사용하는 세그먼트 규칙의 수입입니다.

범위. 모델의 범위입니다.

빈도. 전체에 관련된 적중 수입입니다.

확률. 모델의 확률 퍼센트를 표시합니다.

4. 스냅샷을 작업 모델로 올리려면 필요한 스냅샷을 강조 표시하고 **로드**를 클릭하십시오. 작업 모델 분할창에 스냅샷 모델이 표시됩니다.
5. **삭제**를 클릭하거나 스냅샷을 마우스 오른쪽 단추로 클릭하고 메뉴에서 **삭제**를 선택하여 스냅샷을 삭제할 수 있습니다.

④ 의사결정 목록 뷰어에 대한 작업

고객 반응과 작동을 가장 잘 예측할 모델은 다양한 단계에서 작성됩니다. 의사결정 목록 뷰어를 실행하면 작업 모델은 마이닝 작업을 시작하고, 필요에 따라 세그먼트/측도를 수정하며, 새 모델 또는 모델링 노드를 생성할 수 있도록 준비되어 있는 정의된 모델 세그먼트 및 측도로 채워져 있습니다.

만족스러운 모델을 개발할 때까지 하나 이상의 세그먼트 규칙을 추가할 수 있습니다. 마이닝 작업을 실행하거나 **세그먼트 규칙 편집** 기능을 사용하여 세그먼트 규칙을 모델에 추가할 수 있습니다.

모델 작성 프로세스에서는, 측도 데이터에 대해 모델을 검증하거나, 모델을 차트로 시각화하거나, 사용자 정의 Excel 측도를 생성해서 모델의 성능을 평가할 수 있습니다.

모델의 품질에 확신이 생기면 새 모델을 생성하여 IBM® SPSS® Modeler 캔버스나 모델 팔레트에 둘 수 있습니다.

가. 마이닝 작업

마이닝 작업은 새 규칙이 생성되는 방식을 판별하는 매개변수 콜렉션입니다. 새로운 상황에 모델을 탄력적으로 적응할 수 있도록 일부 매개변수가 선택 가능합니다. 작업은 작업 템플릿(유형), 목표, 선택 작성(데이터 세트 마이닝)으로 이루어집니다.

다양한 마이닝 작업 조작은 다음 섹션에서 자세히 설명합니다.

ㄱ. 마이닝 작업 실행

의사결정 목록 뷰어를 통해 마이닝 작업을 실행하거나 모델 간에 세그먼트 규칙을 붙여넣어서 세그먼트 규칙을 수동으로 모델에 추가할 수 있습니다. 마이닝 작업은 새 세그먼트 규칙의 생성 방식(검색 처리 방법, 소스 속성, 검색 범위, 신뢰수준 등과 같은 데이터 마이닝 매개변수 설정), 예측할 고객 동작, 조사할 데이터에 대한 정보를 보유합니다. 마이닝 작업의 목표는 가장 가능한 세그먼트 규칙을 검색하는 것입니다.

마이닝 작업을 실행해서 모델 세그먼트 규칙을 생성하려면 다음을 수행하십시오.

1. **나머지** 행을 클릭하십시오. 작업 모델 분할창에 세그먼트가 이미 표시된 경우 세그먼트 중 하나를 선택하고 선택한 세그먼트를 기준으로 하여 추가 규칙을 찾을 수도 있습니다. 나머지 또는 세그먼트를 선택한 후 다음 방법 중 하나를 사용하여 모델이나 대체 모델을 생성하십시오.

- 도구 메뉴에서 **세그먼트 찾기**를 선택하십시오.
- **나머지** 행/세그먼트를 마우스 오른쪽 단추로 클릭하고 **세그먼트 찾기**를 선택하십시오.
- 작업 모델 분할창에서 **세그먼트 찾기** 단추를 클릭하십시오.

작업이 처리 중인 동안에는 작업공간의 맨 아래에 진행률이 표시되어 작업이 완료되는 시기를 알려 줍니다. 엄밀히 작업 완료에 걸리는 시간은 마이닝 작업의 복잡도 및 데이터 세트의 크기에 따라 다릅니다. 결과에 모델이 하나만 있으면 작업이 완료되는 즉시 작업 모델 분할창에 표시되지만, 결과에 둘 이상의 모델이 포함된 경우에는 대안 탭에 표시됩니다.

참고: 작업 결과는 모델과 함께 완료되거나, 모델 없이 완료되거나, 실패합니다.

새 세그먼트 규칙을 찾는 프로세스는 모델에 추가되는 새 규칙이 없을 때까지 반복됩니다. 이는 고객의 모든 유의적 그룹을 찾았음을 의미합니다.

기존 모델 세그먼트에 마이닝 작업을 실행할 수 있습니다. 작업 결과가 찾고 있는 내용이 아닌 경우 동일한 세그먼트에 다른 마이닝 작업을 시작하도록 선택할 수 있습니다. 그러면 선택한 세그먼트를 기반으로 추가 규칙이 발견됩니다. 선택한 세그먼트 "아래에" 있는(즉, 선택한 세그먼트 이후에 모델에 추가된) 세그먼트는 각 세그먼트가 선행자에 종속되기 때문에 새 세그먼트로 대체됩니다.

나. 마이닝 작업 작성 및 편집

마이닝 작업은 데이터 모델을 구성하는 규칙 컬렉션을 검색하는 메커니즘입니다. 선택된 템플릿에 정의된 검색 기준과 함께, 작업은 목표(메일링에 응답할 고객 수와 같은 분석을 유발한 실제 질문)를 정의하고 사용할 데이터 세트를 식별하기도 합니다. 마이닝 작업의 목표는 가장 가능한 모델을 검색하는 것입니다.

마이닝 작업 작성

마이닝 작업을 작성하려면 다음을 수행하십시오.

1. 추가 세그먼트 조건을 마이닝하려는 세그먼트를 선택하십시오.
2. **설정**을 클릭하십시오. 마이닝 작업 작성/편집 대화 상자가 열립니다. 이 대화 상자는 마이닝 작업을 정의하기 위한 옵션을 제공합니다.

- 필요한 변경을 수행하고 **확인**을 클릭하여 작업 모델 분할창으로 돌아가십시오. 의사결정 목록 뷰어는 대체 작업 또는 설정이 선택될 때까지 이 설정을 각 작업에 실행할 기본값으로 사용합니다.
- 세그먼트 찾기**를 클릭하여 선택된 세그먼트에 대한 마이닝 작업을 시작하십시오.

마이닝 작업 편집

마이닝 작업 작성/편집 대화 상자는 새 마이닝 작업을 정의하거나 기존 마이닝 작업을 편집하기 위한 옵션을 제공합니다.

마이닝 작업에 사용 가능한 대부분의 매개변수는 의사결정 목록 노드에 제공된 것과 유사합니다. 예외는 아래에 표시됩니다. 자세한 정보는 의사결정 목록 모델 옵션의 내용을 참조하십시오.

로드 설정: 둘 이상의 마이닝 작업을 작성한 경우 필수 작업을 선택하십시오.

새 파일... 현재 표시된 작업의 설정을 기준으로 하여 새 마이닝 작업을 작성하려면 클릭하십시오.

자세한 정보는 새 설정의 내용을 참조하십시오.

목표

목표 필드: 예측하려는 필드를 나타내며, 값이 다른 필드(예측자)의 값에 관련된 것으로 추정됩니다.

목표 값. 모델링하려는 결과를 나타내는 대상 필드의 값을 지정합니다. 예를 들어, 대상 필드 이 탈이 코딩된 0 = no 및 1 = yes인 경우 이탈할 것 같은 레코드를 표시하는 규칙을 식별하려면 1을 지정하십시오.

단순 설정

최대 대안 수. 마이닝 작업을 실행할 때 표시할 대안 수를 지정합니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

고급 설정

편집... 고급 설정을 정의할 수 있는 **고급 매개변수 편집** 대화 상자를 엽니다. 자세한 정보는 고급 매개변수 편집의 내용을 참조하십시오.

데이터

선택 작성 의사결정 목록 뷰어가 새 규칙을 찾기 위해 분석해야 하는 평가 속도를 지정하는 옵션을 제공합니다. 나열된 평가 속도는 데이터 선택 구성 대화 상자에서 작성/편집됩니다.

사용 가능한 필드. 모든 필드를 표시하거나 표시할 필드를 수동으로 선택하기 위한 옵션을 제공합니다.

편집... 사용자 정의 옵션이 선택되면 마이닝 작업으로 찾은 세그먼트 속성으로 사용 가능한 필드를 선택할 수 있는 사용 가능 필드 사용자 정의 대화 상자가 열립니다. 자세한 정보는 사용 가능 필드 사용자 정의의 내용을 참조하십시오.

- 새 설정

마이닝 작업 작성/편집 대화 상자에서 새로 작성...을 클릭하면 새 설정 대화 상자가 표시됩니다.

적절한 이름을 입력한 후 확인을 클릭하십시오. 마이닝 작업 작성/편집 대화 상자가 열리고 설정을 수정할 수 있습니다.

- 고급 매개변수 편집

고급 매개변수 편집 대화 상자는 다음 구성 옵션을 제공합니다.

구간화 방법. 연속형 필드(동일한 개수나 동일한 너비) 구간화에 사용되는 방법입니다.

구간 수. 연속형 필드에 작성할 구간 수입니다. 허용된 최소 설정은 2이고 최대 설정은 없습니다.

모델 검색 범위. 다음 순환에 사용할 수 있는 순환별 모델 결과의 최대 수입니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

규칙 검색 범위. 다음 순환에 사용할 수 있는 순환별 규칙 결과의 최대 수입니다. 허용된 최소 설정은 1이고 최대 설정은 없습니다.

구간 병합 요인. 이웃 항목과 병합되었을 때 증가하는 세그먼트의 최소 크기입니다. 허용된 최소 설정은 1.01이고 최대 설정은 없습니다.

- **조건의 결측값 허용.** True는 규칙의 IS MISSING 검정을 허용합니다.
- **중간 결과 삭제.** True일 경우 검색 프로세스의 최종 결과만 리턴됩니다. 최종 결과는 검색 프로세스에서 더 이상 세분화되지 않는 결과입니다. False이면 중간 결과도 리턴됩니다.

- 사용 가능 필드 사용자 정의

사용 가능 필드 사용자 정의 대화 상자에서는 마이닝 작업으로 찾은 세그먼트 속성으로 사용 가능한 필드를 선택할 수 있습니다.

사용 가능. 세그먼트 속성으로 현재 사용 가능한 필드를 나열합니다. 목록에서 필드를 제거하려면 해당 필드를 선택하고 **제거 >>**를 클릭하십시오. 선택한 필드가 사용 가능 목록에서 사용 불가능 목록으로 이동합니다.

사용 불가능. 세그먼트 속성으로 사용할 수 없는 필드를 나열합니다. 사용 가능 목록에 필드를 포함시키려면 해당 필드를 선택하고 **<< 추가**를 클릭하십시오. 선택한 필드가 사용 불가능 목록에서 사용 가능 목록으로 이동합니다.

ㄷ. 데이터 선택 구성

데이터 선택(마이닝 데이터 세트)을 구성하여 의사결정 목록 뷰어가 새 규칙을 찾기 위해 분석해야 하는 평가 측도를 지정하고 측도의 기준으로 사용되는 데이터 선택을 선택할 수 있습니다.

데이터 선택을 구성하려면 다음을 수행하십시오.

1. 도구 메뉴에서 **데이터 선택 구성**을 선택하거나 세그먼트를 마우스 오른쪽 단추로 클릭하고 옵션을 선택하십시오. 데이터 선택 구성 대화 상자가 열립니다.
참고: 데이터 선택 구성 대화 상자에서 기존 데이터 선택을 편집하거나 삭제할 수도 있습니다.
2. **새 데이터 선택 추가** 단추를 클릭하십시오. 새 데이터 선택 항목이 기존 테이블에 추가됩니다.
3. **이름**을 클릭하고 적합한 선택 이름을 입력하십시오.
4. **파티션**을 클릭하고 적합한 파티션 유형을 선택하십시오.
5. **조건**을 클릭하고 적합한 조건 옵션을 선택하십시오. **지정**이 선택되면 특정 필드 조건을 정의하기 위한 옵션을 제공하는 선택 조건 지정 대화 상자가 열립니다.
6. 적합한 조건을 정의하고 **확인**을 클릭하십시오.

데이터 선택은 마이닝 작업 작성/편집 대화 상자의 선택 작성 드롭 다운 목록에서 사용 가능합니다. 목록을 통해 특정 마이닝 작업에 사용되는 평가 측도를 선택할 수 있습니다.

• 선택 조건 지정

선택 조건 지정 대화 상자에서는 데이터 선택 파티션에 대한 선택 조건을 정의하는 옵션을 제공합니다. 이 조건은 정의된 파티션에서 레코드를 선택하는 데 사용됩니다. 예를 들어, **훈련** 파티션에 대한 기본 조건은 **모든 레코드**입니다. 신규 또는 기존 데이터 선택에 대한 선택 조건을 지정하면 **모든 레코드**를 필터링하여 더 간결한 결과(예: married=yes)를 얻을 수 있습니다. 데이터 선택이 마이닝 작업에서 사용되는 경우 결과 세트는 결혼한 개인에 대한 적중만 리턴합니다.

속성 - 사용 가능한 입력 필드를 나열합니다.

연산자 - 드롭 다운 목록에서는 선택된 입력 필드에 대해 유효한 모든 연산자를 제공합니다.

값 - 선택된 입력 필드에 대한 값을 쿼리합니다. 예를 들어, 입력 필드 **married**는 YES 및 NO 옵션을 제공합니다.

◆ 값 삽입

값 삽입 대화 상자에는 선택된 입력 필드에 대해 사용 가능한 값이 표시됩니다.

적절한 값을 선택한 후 **삽입**을 클릭하십시오.

나. 실행 취소 및 다시 실행 조치

의사결정 목록 뷰어를 사용하면 수행한 최근 10개 조치를 실행 취소하고 다시 실행할 수 있습니다. 예를 들어, 다음을 수행할 수 있습니다.

- 새 모델 규칙 실행 취소
- 모델에 대해 작성한 변경사항 실행 취소
- 모델 세그먼트의 요소 삭제 취소
- 대체 모델 선택 다시 실행

조치를 실행 취소하거나 다시 실행하려면 편집 메뉴에서 **실행 취소** 또는 **다시 실행**을 선택하십시오.

다. 세그먼트 규칙

작업 템플릿을 기준으로 하여 마이닝 작업을 실행해서 모델 세그먼트 규칙을 찾을 수 있습니다. 세그먼트 삽입 또는 세그먼트 규칙 편집 기능을 사용하여 모델에 세그먼트 규칙을 수동으로 추가할 수 있습니다.

새 세그먼트 규칙을 찾기 위해 마이닝하도록 선택하면 대화형 목록 대화 상자의 뷰어 탭에 결과가 표시됩니다(있는 경우). 모델 앨범 대화 상자에서 대체 결과 중 하나를 선택하고 **로드**를 클릭하여 모델을 빠르게 세분화할 수 있습니다. 이러한 방식으로 최적의 목표 그룹을 정확하게 설명하는 모델을 작성할 준비가 되었을 때 다른 결과를 시험할 수 있습니다.

ㄱ. 세그먼트 삽입

세그먼트 삽입 기능을 사용하여 모델에 세그먼트 규칙을 수동으로 추가할 수 있습니다.

모델에 세그먼트 규칙 조건을 추가하려면 다음을 수행하십시오.

1. 대화형 목록 대화 상자에서 새 세그먼트를 추가하려는 위치를 선택하십시오. 선택한 세그먼트 바로 위에 새 세그먼트가 삽입됩니다.
2. 편집 메뉴에서 **세그먼트 삽입**을 선택하거나 세그먼트를 마우스 오른쪽 단추로 클릭하여 이 선택에 액세스하십시오.
새 세그먼트 규칙 조건을 삽입할 수 있는 세그먼트 삽입 대화 상자가 열립니다.
3. **삽입**을 클릭하십시오. 새 규칙 조건의 속성을 정의할 수 있는 조건 삽입 대화 상자가 열립니다.
4. 드롭다운 목록에서 필드 및 연산자를 선택하십시오.
참고: **Not in** 연산자를 선택할 경우 선택한 조건은 제외 조건으로 작용하며 규칙 삽입 대화 상자에 빨간색으로 표시됩니다. 예를 들어, `region = 'TOWN'` 조건이 빨간색으로 표시되는 경우 이는 TOWN이 결과 세트에서 제외됨을 의미합니다.
5. 하나 이상의 값을 입력하거나 **값 삽입** 아이콘을 클릭하여 값 삽입 대화 상자를 표시하십시오. 대화 상자에서 선택된 필드의 정의된 값을 선택할 수 있습니다. 예를 들어, **married** 필드는 **예** 및 **아니오** 값을 제공합니다.
6. **확인**을 클릭하여 세그먼트 삽입 대화 상자로 돌아가십시오. 모델에 작성한 세그먼트를 추가하려면 **확인**을 한번 더 클릭하십시오.

새 세그먼트가 지정된 모델 위치에 표시됩니다.

ㄴ. 세그먼트 규칙 편집

세그먼트 규칙 편집 기능으로 세그먼트 규칙 조건을 추가, 변경 또는 삭제할 수 있습니다.

세그먼트 규칙 조건을 변경하려면 다음을 수행하십시오.

1. 편집하려는 모델 세그먼트를 선택하십시오.
2. 편집 메뉴에서 **세그먼트 규칙 편집**을 선택하거나 이 선택에 액세스하려면 규칙을 마우스 오른쪽 단추로 클릭하십시오.
세그먼트 규칙 편집 대화 상자가 열립니다.
3. 적합한 조건을 선택하고 편집을 클릭하십시오.
선택된 규칙 조건의 속성을 정의할 수 있는 조건 편집 대화 상자가 열립니다.

4. 드롭다운 목록에서 필드 및 연산자를 선택하십시오.
참고: **Not in** 연산자를 선택할 경우 선택한 조건은 제외 조건으로 작용하며 세그먼트 규칙 편집 대화 상자에 빨간색으로 표시됩니다. 예를 들어, region = 'TOWN' 조건이 빨간색으로 표시되는 경우 이는 TOWN이 결과 세트에서 제외됨을 의미합니다.
5. 하나 이상의 값을 입력하거나 **값 삽입** 단추를 클릭하여 값 삽입 대화 상자를 표시하십시오. 대화 상자에서 선택된 필드의 정의된 값을 선택할 수 있습니다. 예를 들어, **married** 필드는 **예** 및 **아니오** 값을 제공합니다.
6. **확인**을 클릭하여 세그먼트 규칙 편집 대화 상자로 돌아가십시오. 작업 모델로 돌아가려면 확인을 한번 더 클릭하십시오.

업데이트된 규칙 조건과 함께 선택한 세그먼트가 표시됩니다.

- **조건 삽입/편집**

조건 삽입/편집 대화 상자는 세그먼트 규칙 조건을 정의하는 데 필요한 옵션을 제공합니다.

속성 - 사용 가능한 입력 필드를 나열합니다.

연산자 - 드롭 다운에서 선택된 입력 필드에 대해 유효한 모든 연산자를 제공합니다.

값 - 선택된 입력 필드에 대한 값을 쿼리합니다. 예를 들어, 입력 필드 **married**는 값 옵션 **YES** 및 **NO**를 제공합니다.

참고: **Not in** 연산자를 선택하면 선택된 조건은 제외 조건으로 작동하며 **세그먼트 삽입/편집** 대화 상자에서 빨간색으로 표시됩니다. 예를 들어, region = 'TOWN' 조건이 빨간색으로 표시되면 이는 TOWN이 결과 세트에서 제외됨을 의미합니다.

- **세그먼트 규칙 조건 삭제**

세그먼트 규칙 조건을 삭제하려면 다음을 수행하십시오.

1. 삭제하려는 규칙 조건을 포함한 모델 세그먼트를 선택하십시오.
2. 편집 메뉴에서 **세그먼트 규칙 편집**을 선택하거나 이 선택에 액세스하려면 세그먼트를 마우스 오른쪽 단추로 클릭하십시오.
하나 이상의 세그먼트 규칙 조건을 삭제할 수 있는 세그먼트 규칙 편집 대화 상자가 열립니다.
3. 적합한 규칙 조건을 선택하고 **삭제**를 클릭하십시오.
4. **확인**을 클릭하십시오.

하나 이상의 세그먼트 규칙 조건을 삭제하면 작업 모델 분할창이 측도 기준을 새로 고칩니다.

ㄷ. 세그먼트 복사

의사결정 목록 뷰어는 모델 세그먼트를 복사하기 위한 편리한 방법을 제공합니다. 세그먼트를 한 모델에서 다른 모델에 적용하려는 경우 단순히 한 모델에서 세그먼트를 복사(또는 잘라내기)한 후 다른 모델에 이를 붙여넣으십시오. 대체 미리보기 패널에 표시되는 모델에서 세그먼트를 복사한 후 작업 모델 분할창에 표시된 모델에 이를 붙여넣을 수도 있습니다. 이 잘라내기, 복사, 붙여넣기 기능은 시스템 클립보드를 사용하여 임시 데이터를 저장하거나 검색합니다. 이는 클립보드에서 조건 및 목표가 복사됨을 의미합니다. 클립보드 콘텐츠는 단독으로 의사결정 목록 뷰어에 사용하도록 예약되지 않지만 다른 애플리케이션에 붙여넣을 수도 있습니다. 예를 들어, 클립보드 콘텐츠를 텍스트 편집기에서 붙여넣으면 조건 및 목표가 XML 형식으로 붙여넣어집니다.

모델 세그먼트를 복사하거나 잘라내려면 다음을 수행하십시오.

1. 다른 모델에 사용하려는 모델 세그먼트를 선택하십시오.
2. 편집 메뉴에서 **복사**(또는 **잘라내기**)를 선택하거나 모델 세그먼트를 마우스 오른쪽 단추로 클릭하고 **복사** 또는 **잘라내기**를 선택하십시오.
3. 적합한 모델을 여십시오(모델 세그먼트가 붙여넣기됨).
4. 모델 세그먼트 중 하나를 선택하고 **붙여넣기**를 클릭하십시오.

참고: 잘라내기, 복사, 붙여넣기 명령 대신 키 조합 **Ctrl+X**, **Ctrl+C**, **Ctrl+V**를 사용할 수도 있습니다.

복사한(또는 잘라낸) 세그먼트는 이전에 선택한 모델 세그먼트 위에 삽입됩니다. 붙여넣은 세그먼트 및 아래 세그먼트의 축도를 다시 계산합니다.

참고: 이 절차의 두 모델은 모두 동일한 기본 모델 템플릿을 기준으로 하고 동일한 목표를 포함해야 합니다. 그렇지 않으면 오류 메시지가 표시됩니다.

ㄹ. 대체 모델

결과가 둘 이상인 경우 대체 탭에 각 마이닝 작업의 결과가 표시됩니다. 각 결과는 목표와 가장 근접하게 일치하는 선택된 데이터의 조건 및 "적합한" 대안으로 이루어집니다. 표시되는 총 대안 수는 분석 프로세스에 사용된 검색 기준에 따라 다릅니다.

대체 모델을 보려면 다음을 수행하십시오.

1. 대안 탭에서 대체 모델을 클릭하십시오. 대체 미리보기 패널에서 대체 모델 세그먼트가 표시되거나 현재 모델 세그먼트가 대체됩니다.
2. 작업 모델 분할창에서 대체 모델에 대해 작업하려면 모델을 선택하고 대체 미리보기 패널에서 **로드**를 클릭하거나 대안 탭에서 대체 이름을 마우스 오른쪽 단추로 클릭하고 **로드**를 선택하십시오.

참고: 새 모델을 생성할 때 대체 모델은 저장되지 않습니다.

라. 모델 사용자 정의

데이터는 정적 성향이 아닙니다. 고객은 이사하고, 결혼하고, 작업을 변경합니다. 제품은 시장 초점을 잃고 쓸모없게 됩니다.

의사결정 목록 뷰어는 비즈니스 사용자가 새로운 상황에 쉽고 빠르게 모델을 적응할 수 있는 탄력성을 제공합니다. 지정된 모델 세그먼트를 편집, 우선 순위 지정, 삭제 또는 비활성화해서 모델을 변경할 수 있습니다.

ㄱ. 세그먼트 우선 순위 지정

모델 규칙을 선택한 순서대로 순위를 지정할 수 있습니다. 기본적으로 모델 세그먼트는 우선 순위 순으로(첫 번째 세그먼트가 가장 높은 우선 순위) 표시됩니다. 하나 이상의 세그먼트에 다른 우선 순위를 지정하면 모델이 이에 따라 변경됩니다. 필요한 대로 세그먼트를 더 높거나 낮은 우선 순위 위치로 이동시켜서 모델을 변경할 수 있습니다.

모델 세그먼트의 우선 순위를 지정하려면 다음을 수행하십시오.

1. 다른 우선 순위를 지정하려는 모델 세그먼트를 선택하십시오.
2. 작업 모델 분할창 도구 모음에서 두 개의 화살표 단추 중 하나를 클릭하여 선택된 모델 세그먼트를 목록의 위나 아래로 이동하십시오.

우선 순위를 지정하고 나면 이전의 모든 평가 결과가 다시 계산되고 새 값이 표시됩니다.

ㄴ. 세그먼트 삭제

하나 이상의 세그먼트를 삭제하려면 다음을 수행하십시오.

1. 모델 세그먼트를 선택하십시오.
2. 편집 메뉴에서 **세그먼트 삭제**를 선택하거나 작업 모델 분할창의 도구 모음에서 삭제 단추를 클릭하십시오.

수정된 모델에 대해 측도가 다시 계산되고 이에 따라 모델이 변경됩니다.

ㄷ. 세그먼트 제외

특정 그룹을 검색하는 경우 모델 세그먼트 선택에 대한 비즈니스 조치를 기반으로 할 것입니다. 모델을 배포할 때 모델 내 세그먼트를 제외하도록 선택할 수 있습니다. 제외된 세그먼트는 널값

으로 스코어링됩니다. 세그먼트를 제외한다고 해서 세그먼트가 사용되지 않는 것은 아닙니다. 이는 이 규칙에 일치하는 모든 레코드가 메일링 목록에서 제외된다는 의미입니다. 규칙이 여전히 적용되지만 다른 방식으로 적용됩니다.

특정 모델 세그먼트를 제외하려면 다음을 수행하십시오.

1. 작업 모델 분할창에서 세그먼트를 선택하십시오.
2. 작업 모델 분할창의 도구 모음에서 **세그먼트 제외** 토크 단추를 클릭하십시오. 선택한 세그먼트의 선택한 목표 옆에 이제 **제외됨**이 표시됩니다.

참고: 삭제된 세그먼트와 달리, 제외된 세그먼트는 최종 모델에 재사용할 수 있습니다. 제외된 세그먼트는 차트 결과에 영향을 미칩니다.

ㄹ. 목표 값 변경

목표 값 변경 대화 상자에서 현재 목표 필드의 목표 값을 변경할 수 있습니다.

목표 값이 작업 모델과 다른 스냅샷 및 세션 결과는 이 행의 테이블 배경을 노랑으로 변경해서 식별됩니다. 이는 스냅샷/세션 결과가 최신 결과가 아님을 나타내는 것입니다.

마이닝 작업 작성/편집 대화 상자에 현재 작업 모델의 목표 값이 표시됩니다. 목표 값은 마이닝 작업과 함께 저장되지 않습니다. 대신에 작업 모델 값에서 가져옵니다.

현재 작업 모델과 목표 값이 다른(예를 들어, 대체 결과를 편집하거나 스냅샷 사본을 편집해서) 작업 모델로 저장된 모델을 올리는 경우 저장된 모델의 목표 값이 작업 모델과 동일하게 변경됩니다(작업 모델 분할창에 표시된 목표 값은 변경되지 않음). 모델 메트릭이 새 목표로 재평가됩니다.

마. 새 모델 생성

새 모델 생성 대화 상자는 모델의 이름을 지정하고 새 노드가 작성되는 위치를 선택할 수 있는 옵션을 제공합니다.

모델 이름. 사용자 정의를 선택하여 자동 생성된 이름을 조정하거나 스트림 캔버스에 표시되는 노드의 고유 이름을 작성하십시오.

노드 작성 위치. 캔버스를 선택하면 작업 캔버스에 새 모델을 두고 **GM 팔레트**를 선택하면 모델 팔레트에 새 모델을 둡니다. 둘 다를 선택할 경우에는 작업 캔버스와 모델 팔레트에 모두 새 모델을 둡니다.

대화형 세션 상태 포함. 사용할 경우 생성된 모델에 대화형 세션 상태가 유지됩니다. 나중에 모델에서 모델링 노드를 생성할 때 상태가 계속 유지되며 대화형 세션을 초기화하는 데 사용됩니다. 선택된 옵션과 상관 없이 모델 자체는 새 데이터를 동일하게 스코어링합니다. 옵션을 선택하지 않으면 모델이 여전히 작성 노드를 작성할 수 있지만, 이전 세션이 중단한 곳에서 시작하지 않고 새 대화형 세션을 시작하는 보다 일반적인 작성 노드가 됩니다. 노드 설정을 변경하지만 저장된 상태로 실행할 경우에는 저장된 상태의 설정을 위해 변경한 설정이 무시됩니다.

참고: 표준 메트릭은 모델과 함께 잔존하는 유일한 메트릭입니다. 추가 메트릭은 대화형 상태로 유지됩니다. 생성된 모델이 저장된 대화형 마이닝 작업 상태를 나타내지 않습니다. 의사결정 목록 뷰어가 실행되면 뷰어를 통해 원래 작성한 설정이 표시됩니다.

자세한 정보는 모델링 노드 재생성의 내용을 참조하십시오.

바. 모델 평가

성공적 모델링을 위해서는 프로덕션 환경에서 구현이 발생하기 전에 주의 깊게 모델을 평가해야 합니다. 의사결정 목록 뷰어는 실세계에서 모델의 영향을 평가하는 데 사용할 수 있는 많은 통계 및 비즈니스 측도를 제공합니다. 여기에는 Gains 차트 및 Excel과의 완전한 상호 운용성이 포함되므로 배포 영향을 평가하기 위해 비용/이익 시나리오를 시뮬레이션할 수 있습니다.

다음 방식으로 모델을 평가할 수 있습니다.

- 의사결정 목록 뷰어에서 사용 가능한 사전 정의된 통계 및 비즈니스 모델 측도(확률, 빈도)를 사용.
- Microsoft Excel에서 가져온 측도를 평가.
- Gains 차트를 사용하여 모델을 시각화.

ㄱ. 모델 측도 구성

의사결정 목록 뷰어는 열로 계산 및 표시되는 측도를 정의하기 위한 옵션을 제공합니다. 각 세그먼트는 열로 표시되는 기본값 커버, 빈도, 확률, 오류 측도를 포함할 수 있습니다. 열로 표시할 새 측도를 작성할 수도 있습니다.

모델 측도 정의

모델에 측도를 추가하거나 기존 측도를 정의하려면 다음을 수행하십시오.

1. 도구 메뉴에서 **모델 측도 구성**을 선택하거나 모델을 마우스 오른쪽 단추로 클릭하여 이 선택을 수행하십시오. 모델 측도 구성 대화 상자가 열립니다.
2. **새 모델 측도 추가** 단추(표시 열의 오른쪽에)를 클릭하십시오. 새 측도가 테이블에 표시됩니다.

3. 측도 이름을 제공하고 적합한 유형, 표시 옵션, 선택사항을 제공하십시오. 표시 열에 작업 모델에 대한 측도를 표시할지 여부가 나타납니다. 기존 측도를 정의할 때 적합한 메트릭 및 선택사항을 선택하고 작업 모델에 대한 측도를 표시할지 여부를 표시하십시오.
4. **확인**을 클릭하여 의사결정 목록 뷰어 작업공간으로 돌아가십시오. 새 측도에 대한 열 표시를 선택한 경우 작업 모델에 대한 새 측도가 표시됩니다.

Excel의 사용자 정의 메트릭

자세한 정보는 Excel로 평가의 내용을 참조하십시오.

- **측도 새로 고침**

새 고객 세트에 기존 모델을 적용하는 때와 같은 특정 경우에는 모델 측도를 다시 계산해야 할 수 있습니다.

모델 측도를 다시 계산하려면(새로 고치려면) 다음을 수행하십시오.

편집 메뉴에서 **모든 측도 새로 고침**을 선택하십시오.

or

F5를 누르십시오.

모든 측도가 다시 계산되고 작업 모델의 새 값이 표시됩니다.

ㄴ. Excel로 평가

의사결정 목록 뷰어를 Microsoft Excel과 통합하여 모델 작성 프로세스 내에서 직접 자신의 값 계산 및 이익 수식을 사용하여 비용/이익 시나리오를 시뮬레이션할 수 있습니다. Excel을 포함한 링크를 통해 Excel로 데이터를 내보내서 프리젠테이션 도표를 작성하고, ROI 측도 및 복합 이익과 같은 사용자 정의 측도를 계산하며, 모델을 작성하는 동안 의사결정 목록 뷰어에서 이를 볼 수 있습니다.

참고: Excel 스프레드시트에 대해 작업하려면 분석 CRM 전문가가 Microsoft Excel과 의사결정 목록 뷰어의 동기화에 대한 구성 정보를 정의해야 합니다. 구성은 Excel 스프레드시트 파일에 포함되어 있으며 의사결정 목록 뷰어에서 Excel로(그리고 반대로) 전송되는 정보를 표시합니다.

다음 단계는 MS Excel이 설치되어 있을 때에만 유효합니다. Excel이 설치되지 않은 경우 Excel과 모델 동기화에 대한 옵션이 표시되지 않습니다.

모델을 MS Excel과 동기화하려면 다음을 수행하십시오.

1. 모델을 열고 대화형 세션을 실행한 후 도구 메뉴에서 **모델 속도 구성**을 선택하십시오.
2. **Excel로 사용자 정의 속도 계산** 옵션에 **예**를 선택하십시오. 워크북 필드가 활성화되어 사전 구성된 Excel 워크북 템플릿을 선택할 수 있습니다.
3. **Excel에 연결** 단추를 클릭하십시오. 열기 대화 상자가 열려서 로컬 또는 네트워크 파일 시스템에서 사전 구성된 템플릿 위치를 탐색할 수 있습니다.
4. 적합한 Excel 템플릿을 선택하고 **열기**를 클릭하십시오. 선택한 Excel 템플릿이 실행됩니다. Windows 작업 표시줄을 사용하여(또는 Alt-Tab을 눌러서) 사용자 정의 속도의 입력 선택 대화 상자로 다시 이동하십시오.
5. Excel 템플릿에 정의된 매트릭 이름과 모델 매트릭 이름 간의 적합한 매핑을 선택하고 **확인**을 클릭하십시오.

일단 링크가 설정되면 스프레드시트에 모델 규칙을 표시하는 사전 구성된 Excel 템플릿으로 Excel이 시작됩니다. Excel로 계산된 결과는 의사결정 목록 뷰어에 새 열로 표시됩니다.

참고: Excel 매트릭은 모델이 저장될 때 남지 않습니다. 매트릭은 활성 세션 중에만 유효합니다. 하지만 Excel 매트릭을 포함한 스냅샷을 작성할 수 있습니다. 스냅샷 보기에 저장된 Excel 매트릭은 히스토리 비교 용도로만 유효하며 다시 열 때 새로 고쳐지지 않습니다. 자세한 정보는 스냅샷 탭의 내용을 참조하십시오. Excel 템플릿에 대한 연결을 다시 설정할 때까지는 Excel 매트릭이 스냅샷에 표시되지 않습니다.

- **사용자 정의 속도에 대한 입력 선택**

사용자 정의 속도에 대한 입력 선택 대화 상자에서는 사용자 정의 Excel 입력 매트릭을 기존 모델 속도에 매핑하는 데 필요한 옵션을 제공합니다.

입력 - Excel 템플릿의 사용자 정의 Excel 입력 매트릭을 나열합니다.

모델 속도 - 사용 가능한 모델 속도를 나열합니다. 각각의 사용자 정의 Excel 입력 매트릭에 매핑할 적절한 속도를 선택하십시오.

- **MS Excel 통합 설정**

의사결정 목록 뷰어와 Microsoft Excel의 통합은 사전구성된 Excel 스프레드시트 템플릿 사용을 통해 수행됩니다. 템플릿은 다음 세 개의 워크시트로 구성됩니다.

모델 속도. 가져온 의사결정 목록 뷰어 속도, 사용자 정의 Excel 속도, 계산 총계(설정 워크시트에 정의됨)를 표시합니다.

설정. 가져온 의사결정 목록 뷰어 측도 및 사용자 정의 Excel 측도를 기준으로 하여 계산을 생성할 변수를 제공합니다.

구성. 의사결정 목록 뷰어에서 가져올 측도를 지정하고 사용자 정의 Excel 측도를 정의하기 위한 옵션을 제공합니다.

경고: 구성 워크시트의 구조는 엄격히 정의되어 있습니다. 녹색 음영 영역의 셀을 편집하지 **마십시오.**

- **모델로부터의 메트릭.** 계산에 사용되는 의사결정 목록 뷰어 메트릭을 표시합니다.
- **모델로의 메트릭.** 의사결정 목록 뷰어에 리턴할 Excel이 생성한 메트릭을 표시합니다. Excel 생성 메트릭은 의사결정 목록 뷰어에 새 측도 열로 표시됩니다.

참고: Excel 메트릭은 새 모델을 생성할 때 모델과 함께 남지 않습니다. 메트릭은 활성 세션 중에만 유효합니다.

• 모델 측도 변경

다음 예는 여러 방법으로 모델 측도를 변경하는 방법을 설명합니다.

- 기존 측도를 변경합니다.
- 모델로부터 추가 표준 측도를 가져옵니다.
- 모델로 추가 사용자 정의 측도를 내보냅니다.

기존 측도 변경

1. 템플릿을 열고 구성 워크시트를 선택하십시오.
2. **이름** 또는 **설명**을 강조 표시한 후 덮어써서 편집하십시오.
측도를 변경하려는 경우(예를 들어, 사용자에게 빈도 대신 확률을 프롬프트하려면) **모델의 지표**에서 이름과 설명만 변경하면 됩니다. 그러면 변경한 사항이 모델에 표시되고 사용자는 맵핑할 적합한 측도를 선택할 수 있습니다.

모델로부터 추가 표준 측도 가져오기

1. 템플릿을 열고 구성 워크시트를 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.
도구 > 보호 > 시트 비보호
3. 노란색 음영 처리되어 있으며 **End** 단어를 포함한 셀 A5를 선택하십시오.
4. 메뉴에서 다음을 선택하십시오.
삽입 > 행

5. 새 측도의 이름 및 설명을 입력하십시오. 예를 들어, 오류 및 세그먼트와 연관된 오류와 같습니다.
6. 셀 C5에 수식 =COLUMN('Model Measures'!N3)을 입력하십시오.
7. 셀 D5에 수식 =ROW('Model Measures'!N3)+1을 입력하십시오.
이 수식은 현재 비어 있는 모델 측도 워크시트의 열 N에 새 측도를 표시합니다.
8. 메뉴에서 다음을 선택하십시오.
도구 > 보호 > 시트 보호
9. 확인을 클릭하십시오.
10. 모델 측도 워크시트에서 셀 N3의 새 열 제목이 오류인지 확인하십시오.
11. 열 N을 모두 선택하십시오.
12. 메뉴에서 다음을 선택하십시오.
형식 > 셀
13. 기본적으로 모든 셀에는 일반 번호 범주가 있습니다. 그림이 표시되는 방식을 변경하려면 퍼센트를 클릭하십시오. 이렇게 하면 Excel에서 그림을 확인할 수 있고 그래프로 출력하는 것처럼 다른 방식으로 데이터를 이용할 수도 있습니다.
14. 확인을 클릭하십시오.
15. 스프레드시트를 고유 이름 및 파일 확장자 .xlt를 지정해서 Excel 2003 템플릿으로 저장하십시오. 새 템플릿을 쉽게 찾을 수 있도록 로컬 또는 네트워크 파일 시스템의 사전 구성된 템플릿 위치에 저장할 것을 권장합니다.

모델로 추가 사용자 정의 측도 내보내기

1. 이전 예에서 오류 열에 추가한 템플릿을 열고 구성 워크시트를 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.
도구 > 보호 > 시트 비보호
3. 노란색으로 음영 처리되어 있으며 End 단어를 포함한 셀 A14를 선택하십시오.
4. 메뉴에서 다음을 선택하십시오.
삽입 > 행
5. 새 측도의 이름 및 설명을 입력하십시오. 예를 들어, 오류 척도 및 Excel의 오류에 척도 적용과 같습니다.
6. 셀 C14에 수식 =COLUMN('Model Measures'!O3)을 입력하십시오.
7. 셀 D14에 수식 =ROW('Model Measures'!O3)+1을 입력하십시오.
이 수식은 열 O이 모델에 새 측도를 제공함을 지정합니다.
8. 설정 워크시트를 선택하십시오.
9. 셀 A17에 설명 '- 오류 척도'를 입력하십시오.
10. 셀 B17에 척도 요인 10을 입력하십시오.
11. 모델 측도 워크시트에서 셀 O3에 새 열의 제목으로 설명 오류 척도를 입력하십시오.
12. 셀 O4에 수식 =N4*Settings!\$B\$17을 입력하십시오.

13. 셀 O4의 모서리를 선택하여 아래의 셀 O22로 끌어서 수식을 각 셀로 복사하십시오.
14. 메뉴에서 다음을 선택하십시오.
도구 > 보호 > 시트 보호
15. **확인**을 클릭하십시오.
16. 스프레드시트를 고유 이름 및 파일 확장자 .xlt를 지정해서 Excel 2003 템플릿으로 저장하십시오. 새 템플릿을 쉽게 찾을 수 있도록 로컬 또는 네트워크 파일 시스템의 사전 구성된 템플릿 위치에 저장할 것을 권장합니다.

이 템플릿을 사용하여 Excel에 연결하면 새 사용자 정의 측도로 오류 값이 사용 가능합니다.

사. 모델 시각화

모델의 영향을 이해하는 최상의 방법은 모델을 시각화하는 것입니다. Gains 차트를 사용하여 여러 대안의 효과를 실시간으로 연구해서 모델의 비즈니스 및 기술적 이익을 매일 유용하게 통찰할 수 있습니다. Gains 차트 섹션은 무작위 의사결정에 따른 모델의 혜택을 보여주고 대체 모델이 있을 때 여러 차트를 직접 비교합니다.

ㄱ. Gains 차트

Gains 차트는 테이블에서 *이익* % 열의 값을 표시합니다. 이익은 다음 방정식을 사용하여 트리에 있는 총 적중 수에 상대적인 각 증분의 적중 비율로 정의됩니다.

$$(\text{증분의 적중 수} / \text{총 적중 수}) \times 100\%$$

Gains 차트는 트리의 모든 적중을 주어진 퍼센트까지 캡처하기 위해 넷을 캐스트해야 하는 범위를 효과적으로 설명합니다. 모델이 사용되지 않는 경우 대각선이 전체 표본의 기대반응을 표시합니다. 이 경우 한 사람이 다른 항목에 응답하는 것과 같기 때문에 반응률은 일정합니다. 두 배로 산출하려면 두 배 더 많은 사람들에게 질문해야 합니다. 곡선은 이익에 기반하여 더 높은 백분위수에 위치한 사람만 포함하여 반응을 얼마나 개선시킬 수 있는지 표시합니다. 예를 들어, 상위 50%만 포함하면 70% 이상의 긍정적인 반응이 돌아옵니다. 곡선이 가파를수록 이익이 높아집니다.

Gains 차트를 보려면 다음을 수행하십시오.

1. 의사결정 목록 노드를 포함한 스트림을 열고 노드에서 대화형 세션을 실행하십시오.
2. **Gains** 탭을 클릭하십시오. 지정된 파티션에 따라 하나 또는 두 개의 차트(예를 들어, 모델 측도에 훈련 및 검정 파티션이 모두 정의될 때 두 개의 차트가 표시됨)를 볼 수 있습니다.

기본적으로 차트는 세그먼트로 표시됩니다. **분위수**를 선택한 후 드롭 다운 메뉴에서 적합한 분위수 방법을 선택하여 차트를 분위수로 표시하도록 전환할 수 있습니다.

- 차트 옵션

차트 옵션 기능은 차트화할 모델 및 스냅샷, 도표화할 파티션, 세그먼트 레이블의 표시 여부를 선택할 수 있는 옵션을 제공합니다.

도표화할 모델

현재 모델. 차트화할 모델을 선택할 수 있습니다. 작업 모델이나 작성된 스냅샷 모델을 선택할 수 있습니다.

도표화할 파티션

왼쪽 차트의 파티션. 드롭 다운 목록에 정의된 모든 파티션 또는 전체 데이터를 표시할 수 있는 옵션이 제공됩니다.

오른쪽 차트의 파티션. 드롭 다운 목록에 정의된 모든 파티션, 전체 데이터 또는 왼쪽 차트만 표시할 수 있는 옵션이 제공됩니다. **왼쪽만 그래프**를 선택하면 왼쪽 차트만 표시됩니다.

세그먼트 레이블 표시. 사용할 경우 각 세그먼트 레이블이 차트에 표시됩니다.

8) 통계 모델

통계 모델은 산술 방정식을 사용하여 데이터에서 추출한 정보를 인코딩합니다. 일부 경우에,, 통계 모델링 기법을 통해 적당한 모델을 매우 신속하게 제공할 수 있습니다. 신경망과 같은 보다 탄력적인 머신 학습 기법을 통해 궁극적으로 더 나은 결과를 제공할 수 있는 환경의 문제점인 경우에서도 기준선 예측 모델로 일부 통계 모델을 사용하여 고급 기법의 성능을 판별할 수 있습니다.

다음과 같은 통계 모델링 노드를 사용할 수 있습니다.



선형 회귀 모형은 목표와 하나 이상의 예측변수 간의 선형 관계를 기반으로 연속형 목표를 예측합니다.



로지스틱 회귀분석은 입력 필드 값을 기반으로 레코드를 분류하는 통계 기법입니다. 선형 회귀와 유사하지만 숫자 범위 대신 범주형 대상 필드를 사용합니다.



PCA/요인 노드에서는 강력한 데이터 축소 기법을 제공하여 데이터의 복잡도를 줄입니다. 비선형 주성분분석(PCA)은 구성요소가 서로 직각(수직)인 전체 필드 세트에서 변동을 캡처하는 입력 필드의 선형 조합을 찾습니다. 요인 분석은 관측된 필드 세트 내에서 상관관계 패턴을 설명하는 기본 요인을 식별하려고 시도합니다. 두 접근 방식 모두 목표는 원래 필드 세트의 정보를 효과적으로 요약하는 적은 수의 파생 필드를 찾는 것입니다.



판별 분석은 로지스틱 회귀분석보다 엄격한 가정을 하지만 해당 가정이 충족되면 로지스틱 회귀 분석의 귀중한 대안 또는 보조물이 될 수 있습니다.



일반화 선형 모델은 종속변수가 요인과 선형적으로 관련되고 지정된 연결함수를 통해 공변되도록 일반 선형 모델을 확장합니다. 더욱이 모델을 사용하면 종속변수가 비정규 분포를 가질 수 있습니다. 선형 회귀, 로지스틱 회귀분석, 카운트 데이터에 대한 로그선형 모델, 간격 중도절단 생존 모델을 포함하여 상당수 통계 모델의 기능을 포함합니다.



일반화 선형 혼합 모델(GLMM)은 목표가 비정규 분포를 가질 수 있고 지정된 연결함수를 통해 요인 및 공변량과 선형적으로 관련되며 관측값을 상관시킬 수 있도록 선형 모델을 확장합니다. 일반화 선형 혼합 모델은 단순 선형 회귀에서 비정규 장기적인 데이터에 대한 복합 다중 수준 모델에 이르기까지 다양한 모델을 포함합니다.



Cox 회귀 노드를 통해 중도절단된 레코드가 있는 데서 시간 대 이벤트 데이터에 대한 생존 모델을 작성할 수 있습니다. 이 모델은 주어진 입력 변수 값에 대해 주어진 시간(t)에 흥미있는 이벤트가 발생한 확률을 예측하는 생존함수를 생성합니다.

(1) 선형 노드

선형 회귀는 숫자 입력 필드 값에 기반하여 레코드를 분류하는 일반적인 통계 기법입니다. 선형 회귀는 예측 및 실제 출력 값 사이의 차이를 최소화하는 직선 또는 곡선에 적합합니다.

요구사항. 숫자 필드만 선형 회귀 모형에서 사용할 수 있습니다. 정확히 하나의 목표 필드(역할이 **목표**로 설정됨) 및 하나 이상의 예측자(역할이 **입력**으로 설정됨)를 보유해야 합니다. 역할이 모두 또는 **없음**인 필드는 비슷자 필드이므로 무시됩니다. (필요한 경우 비슷자 필드는 파생 노드를 사용하여 기록할 수 있습니다.)

강도. 선형 회귀 모형은 비교적 단순하며 예측 생성을 위해 쉽게 해석되는 수학 공식을 제공합니다. 선형 회귀는 장기적으로 안정된 통계 프로시저이므로 이 모델의 특성도 널리 알려져 있습니다. 또한 보통 선형 모델은 빠르게 훈련할 수 있습니다. 선형 노드에서는 방정식에서 중요하지 않은 입력 필드를 제거하기 위해 자동 필드 선택에 대한 방법을 제공합니다.

참고: 목표 필드가 연속적 범위가 아니라 범주형인 경우(예: 예/아니오 또는 이탈/이탈하지 않음) 로지스틱 회귀분석을 대안으로 사용할 수 있습니다. 또한 로지스틱 회귀분석에서는 비슷자 입력에 대한 지원도 제공하므로 이러한 필드를 기록하지 않아도 됩니다. 자세한 정보는 로지스틱 노드의 내용을 참조하십시오.

① 선형 모델

선형 모델은 목표와 하나 이상의 예측자 사이의 선형 관계를 기반으로 연속형 목표를 예측합니다.

선형 모델은 비교적 단순하며 스코어링에 대해 쉽게 해석되는 수식을 제공합니다. 이러한 모델의 특성은 같은 데이터 세트의 다른 모델 유형(예: 신경망 또는 의사결정 트리)과 비교하여 잘 이해되며 일반적으로 아주 빨리 작성될 수 있습니다.

예제. 주택 소유자의 보험 청구에 대해 조사하기 위한 리소스가 제한된 보험 회사가 보험 청구액을 추정하기 위한 모델을 만들고자 합니다. 이 모델을 서비스 센터에 배포하면 영업 담당자는 고객과 통화하면서 청구 정보를 입력하여 지난 데이터를 토대로 '예상' 청구 비용을 즉시 산출할 수 있습니다.

필드 요구사항. 목표와 하나 이상의 입력이 있어야 합니다. 기본적으로 모두 또는 없음의 사전 정의된 역할이 있는 필드는 사용되지 않습니다. 목표가 연속형(척도)이어야 합니다. 예측자(입력)에 측정 수준 제한 사항이 없습니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되며 연속형 필드는 공변량으로 사용됩니다.

가. 목적 (선형 모델)

원하는 작업

- **새 모델 작성.** 완전히 새 모델을 작성합니다. 노드의 일반적인 작업입니다.
- **기존 모델 학습 계속.** 노드에 의해 성공적으로 작성된 마지막 모델로 계속 학습합니다. 원래 데이터에 액세스할 필요 없이 기존 모델을 업데이트하거나 새로 고칠 수 있으며 새 레코드 또는 업데이트된 레코드만 스트림에 입력되므로 상당히 빠르게 수행할 수 있습니다. 이전 모델에 대한 세부 사항이 모델링 노드와 함께 저장되어 스트림 또는 모델 팔레트에서 이전 모델 너깃을 더 이상 사용할 수 없는 경우에도 이 옵션을 사용할 수 있습니다.

참고: 이 옵션이 활성화되면 필드 및 작성 옵션 탭의 다른 모든 제어가 비활성화됩니다.

원하는 기본 목적 적절한 목적을 선택하십시오.

- **표준 모델 작성.** 이 방법은 예측자를 사용하여 목표를 예측하는 단일 모델을 작성합니다. 일반적으로 표준 모델은 부스팅되었거나 배깅되었거나 큰 데이터 세트 앙상블보다 해석하기 쉽고 스코어링이 빠릅니다.

 **참고:** 분할 모델에 대해 기존 모델 계속 훈련과 함께 이 옵션을 사용하려면 Analytic Server에 연결되어야 합니다.

- **모형 정확도(부스팅) 개선.** 이 방법은 더 정확한 예측을 하기 위해 모델의 시퀀스를 생성하는 부스팅을 사용하여 앙상블 모델을 작성합니다. 앙상블은 표준 모델보다 작성 및 스코어 계산하는 데 오래 걸릴 수 있습니다.

부스팅은 일련의 "구성요소 모델"(각각 전체 데이터 세트에서 작성되는)을 생성합니다. 각각의 연속 구성요소 모델을 작성하기 전에, 레코드는 이전 구성요소 모델의 잔차를 기반으로 가중치가 부여됩니다. 잔차가 큰 케이스에는 다음 구성요소 모델이 해당 레코드 예측에 제대로 초점을 맞추도록 상대적으로 높은 분석 가중치가 부여됩니다. 이 구성요소 모델들은 함께 앙상블 모델을 형성합니다. 앙상블 모델은 결합 규칙을 사용하여 새 레코드를 스코어링합니다. 사용 가능한 규칙은 목표의 측정 수준에 따라 다릅니다.

- **모델 안정성(배깅) 개선.** 이 방법은 더 신뢰할 만한 예측을 하기 위해 여러 모델을 생성하는 배깅(붓스트랩 집계)을 사용하여 앙상블 모델을 작성합니다. 앙상블은 표준 모델보다 작성 및 스코어 계산하는 데 오래 걸릴 수 있습니다.

붓스트랩 통합(배깅)은 원래 데이터 세트에서 복원 표본추출하여 훈련 데이터 세트의 복제를 생성합니다. 이는 원래 데이터 세트와 동일한 크기의 붓스트랩 표본을 작성합니다. 그리고 나서 "구성요소 모델"이 각 복제에 작성됩니다. 이 구성요소 모델들은 함께 앙상블 모델을 형성합니다. 앙상블 모델은 결합 규칙을 사용하여 새 레코드를 스코어링합니다. 사용 가능한 규칙은 목표의 측정 수준에 따라 다릅니다.

- **매우 큰 데이터 세트를 위한 모델 작성.** 이 방법은 데이터 세트를 별도의 데이터 블록으로 분할하여 앙상블 모델을 작성합니다. 위의 모델을 작성하기에 데이터 세트가 너무 크거나 증분 모델 작성의 경우 이 옵션을 선택하십시오. 이 옵션은 작성하는 데 시간이 덜 걸릴 수 있지만 표준 모델보다 스코어를 계산하는 데 더 오래 걸릴 수 있습니다.

부스팅, 배깅 및 매우 큰 데이터 세트와 관련된 설정은 앙상블 (선형 모델)의 내용을 참조하십시오.

나. 기본 (선형 모델)

자동으로 데이터 준비. 이 옵션은 모델의 예측력을 최대화하기 위해 프로시저에서 목표 및 예측자를 내부적으로 변환할 수 있습니다. 모든 변형은 모델과 함께 저장되며 스코어링을 위해 새 데이터에 적용됩니다. 변환된 필드의 원래 버전은 모델에서 제외됩니다. 기본적으로 다음 자동 데이터 준비가 수행됩니다.

- **날짜 및 시간 처리.** 각 날짜 예측자가 참조 날짜(1970-01-01) 이후의 경과 시간이 포함된 새 연속형 예측자로 변환됩니다. 각 시간 예측자가 참조 시간(00:00:00) 이후의 경과 시간이 포함된 새 연속형 예측자로 변환됩니다.
- **측정 수준 조정.** 고유 값이 5개 미만인 연속형 예측자가 순서 예측자로 다시 캐스팅됩니다. 고유 값이 10개보다 많은 순서 예측자가 연속형 예측자로 다시 캐스팅됩니다.
- **이상치 처리.** 절사 값을 넘는 연속형 예측자 값(평균에서 3배 표준 편차)이 절사 값으로 설정됩니다.
- **결측값 처리.** 명목 예측자의 결측값이 훈련 파티션의 최빈값으로 대체됩니다. 순서 예측자의 결측값이 훈련 파티션의 중앙값으로 대체됩니다. 연속형 예측자의 결측값이 훈련 파티션의 평균으로 대체됩니다.
- **지도되는 병합.** 목표와 관련하여 처리되는 필드 수를 줄여 더욱 경제적인 모델을 만들 수 있습니다. 입력과 목표 간의 관계를 기반으로 유사한 범주가 식별됩니다. 유의적으로 다르지 않은 범주(즉 0.1보다 큰 P-값을 가지는 범주)가 병합됩니다. 모든 범주가 하나로 병합되는 경우, 필드의 원래 버전과 파생된 버전이 예측자로서 값이 없기 때문에 모델에서 제외됩니다.

신뢰수준. 계수 보기에서 모형 계수의 구간 추정값을 계산하는 데 사용되는 신뢰수준입니다. 0보다 크고 100보다 작은 값을 지정하십시오. 기본값은 95입니다.

다. 모델 선택(선형 모델)

모델 선택 방법. 모델 선택 방법(자세한 내용은 아래 참조) 중 하나를 선택하거나 주효과 모델 향으로 단순히 사용 가능한 예측변수를 모두 입력하는 **모든 예측변수 포함**을 선택하십시오. 기본적으로 **단계별 전진**이 사용됩니다.

단계별 전진 선택. 모델에 아무 효과 없이 시작하여 단계적 기준에 따라 더는 추가하거나 제거할 수 없을 때까지 한번에 한 단계에서 효과를 추가 및 제거합니다.

- **입력/제거 기준.** 모델에 효과를 추가해야 할지 제거해야 할지 결정하는 데 사용되는 통계입니다. **정보 기준(AICC)**은 모델에 제공된 학습 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **F 통계량**은 모델 오차에서 향상도의 통계 검정을 기준으로 합니다. **수정된 R 제곱**은 학습 세트의 적합도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **과적합 방지 기준(ASE)**은 과적합 방지 세트의 적합도(평균제곱 오차, ASE)를 기준으로 합니다. 과적합 방지 세트는 모델을 학습시키는 데 사용되지 않는 원래 데이터 세트의 약 30%의 무작위 부표본입니다.

F 통계량 외의 다른 기준을 선택한 경우, 각 단계의 기준에서 최대 양의 증가에 해당하는 효과가 모델에 추가됩니다. 기준에서 감소에 해당하는 모델의 효과는 제거됩니다.

F 통계량을 기준으로 선택한 경우, 각 단계에서 지정한 임계값보다 작은 최소 p-값이 있는 효과가(다음보다 작은 p-값이 있는 효과 포함) 모델에 추가됩니다. 기본값은 0.05입니다. 지정한 임계값보다 큰 p-값이 있는 모델의 효과는(다음보다 큰 p-값이 있는 효과 제거) 제거됩니다. 기본값은 0.10입니다.

- **최종 모델에서 최대 효과 수를 사용자 정의하십시오.** 기본적으로 사용 가능한 모든 효과를 모델에 입력할 수 있습니다. 또는 단계적 알고리즘이 지정된 최대 효과 수가 있는 단계로 끝나는 경우, 알고리즘이 현재 효과 세트로 중지됩니다.
- **최대 단계 수를 사용자 정의하십시오.** 단계적 알고리즘이 특정 단계 수 이후 중지됩니다. 기본적으로 사용 가능한 효과 수는 3회입니다. 또는 양의 정수로 최대 단계 수를 지정하십시오.

최적 서브세트 선택. "가능한 모든" 모델을 확인하거나 최소한 단계별 전진보다 큰 가능한 모델 서브세트를 확인하여 최적 서브세트 기준에 따라 최적 서브세트를 선택합니다. **정보 기준(AICC)**은 모델에 제공된 학습 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **수정된 R 제곱**은 학습 세트의 적합도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **과적합 방지 기준(ASE)**은 과적합 방지 세트의 적합도(평균제곱 오차, ASE)를 기준으로 합니다. 과적합 방지 세트는 모델을 학습시키는 데 사용되지 않는 원래 데이터 세트의 약 30%의 무작위 부표본입니다.

기준의 최대값이 있는 모델이 최적 모델로 선택됩니다.

참고: 최적 서브세트 선택이 단계별 전진 선택보다 계산이 더 집중됩니다. 최적 서브세트가 부스팅, 배깅 또는 아주 큰 데이터 세트와 함께 수행되는 경우, 단계별 전진 선택을 사용하여 작성되는 표준 모델보다 작성하는 데 상당히 많은 시간이 걸릴 수 있습니다.

라. 앙상블 (선형 모델)

이 설정은 부스팅, 배깅 또는 아주 큰 데이터 세트가 목표에서 요청될 때 발생하는 앙상블 동작을 결정합니다. 선택한 목표에 해당하지 않는 옵션은 무시됩니다.

배깅 및 아주 큰 데이터 세트. 앙상블을 스코어링할 때 앙상블 스코어값을 계산하기 위해 기본 모델에서 예측값을 조합하는 데 사용되는 규칙입니다.

- **연속형 목표의 기본 결합 규칙.** 연속형 목표에 대한 앙상블 예측값은 기본 모델의 예측값 평균 또는 중앙값을 사용하여 조합될 수 있습니다.

모형 정확도를 향상시키는 것이 목표인 경우 결합 규칙 선택이 무시됨에 유의하십시오. 부스팅은 항상 가중 다수 투표를 사용하여 범주형 목표를 스코어링하고 가중 중앙값을 사용하여 연속형 목표를 스코어링합니다.

부스팅 및 배깅. 모형 정확도 또는 안정성을 향상시키는 것이 목표일 때 작성할 기본 모형 수를 지정하십시오. 배깅의 경우, 붓스트랩 표본의 수입니다. 양의 정수여야 합니다.

마. 고급 (선형 모델)

결과 복제. 난수 시드를 설정하면 분석을 복제할 수 있습니다. 과적합 방지 세트에 있는 레코드를 선택하는 데 난수 생성기가 사용됩니다. 정수를 지정하거나, **생성**을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 의사 난수 정수를 작성합니다. 기본값은 54752075입니다.

바. 모델 옵션(선형 모델)

모델 이름. 대상 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 대상 필드 이름입니다.

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 대상 필드의 이름이며 접두부 $\$L$ -이 붙습니다. 예를 들어, *sales*라는 이름의 대상 필드의 경우, 새 필드 이름은 $\$L$ -*sales*가 됩니다.

사. 모델 요약 (선형 모델)

모델 요약 보기는 모델과 그 적합성을 한 눈에 파악할 수 있도록 요약한 스냅샷입니다.

테이블

다음은 포함한 최고 수준 모델 설정을 식별합니다.

- 필드 탭에 지정된 목표의 이름,
- 자동 데이터 준비가 기본 설정에 지정된 대로 수행되었는지 여부,
- 모델 선택 설정에 지정된 모델 선택 방법 및 선택 기준. 최종 모델에 대한 선택 기준 값도 더 작고 개선된 형식으로 표현됩니다.

차트

차트는 최종 모델의 정확도를 표시하며 더 크게 표시된 것이 더 나은 형식입니다. 값은 $100 \times$ 최종 모델에 대해 수정된 R^2 입니다.

아. 자동 데이터 준비 (선형 모델)

이 보기에는 제외되는 필드 및 변환된 필드가 어떻게 자동 데이터 준비(ADP) 단계에서 유도되었는지에 대한 정보가 표시됩니다. 변환되었거나 제외된 각 필드에 대해 테이블에 필드 이름, 분석에서 역할, ADP 단계에서 실행한 작업이 나열됩니다. 필드는 필드 이름의 알파벳 오름차순으로 정렬됩니다. 각 필드에서 할 수 있는 작업은 다음과 같습니다.

- **기간 유도:** 개월은 날짜를 포함하는 필드의 값에서부터 현재 시스템 날짜까지 경과 시간(개월 수)을 계산합니다.
- **기간 유도:** 시간은 시간을 포함하는 필드의 값에서부터 현재 시스템 시간까지 경과 시간(시)을 계산합니다.
- 측정 수준을 연속형에서 순서로 변경은 고유 값이 5개 미만인 연속형 필드를 순서 필드로 다시 캐스팅합니다.
- 측정 수준을 순서에서 연속형으로 변경은 고유 값이 10개보다 많은 순서 필드를 연속형 필드로 다시 캐스팅합니다.
- **이상치 자름**은 절사 값을 넘는 연속형 예측자 값(평균에서 3배 표준 편차)을 절사 값으로 설정합니다.
- **결측값 대체**는 명목 필드의 결측값을 최빈값으로, 순서 필드를 중앙값으로, 연속형 필드를 평균으로 바꿉니다.
- **범주를 병합하여 목표와의 연관 최대화**는 입력과 목표 사이의 관계를 기반으로 '유사한' 예측자 범주를 식별합니다. 유의적으로 다르지 않은 범주(즉 0.05보다 큰 p -값을 가지는 범주)가 병합됩니다.
- **상수 예측자 제외 / 이상치 처리 후 / 범주 병합 후**는 다른 ADP 작업을 마친 후 단일 값을 가진 예측자를 제거합니다.

자. 예측자 중요도 (선형 모델)

일반적으로, 가장 중요한 예측자 필드에 모델링 노력을 집중하고 가장 쓸모없는 예측자를 삭제하거나 무시하기를 원합니다. 예측자 중요도 차트를 사용하면 모델 추정 시 각 예측자의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측자에 대한 값의 합은 1.0이 됩니다. 예측자 중요도는 모형 정확도와는 관련이 없습니다. 단지 예측 시 각 예측자의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

차. 관측값 별 예측값 (선형 모델)

수직축에 예측값의 구간화된 산점도를 표시하고 수평축에 관측값을 표시합니다. 이상적으로, 점이 45도 줄에 있어야 합니다. 이 보기는 레코드가 모델에 의해 특히 잘못 예측되었는지 여부를 알려줄 수 있습니다.

카. 잔차 (선형 모델)

모델 잔차의 진단 차트를 표시합니다.

차트 유형. 여러 가지 표시 유형이 있으며, **유형** 드롭 다운 목록에서 액세스할 수 있습니다.

- **히스토그램.** 정규 분포의 오버레이가 있는 스튜던트화 잔차의 구간화된 히스토그램입니다. 선형모델은 잔차에 정규 분포가 있다고 가정하므로 히스토그램은 원칙적으로 평활선과 비슷해야 합니다.
- **P-P 도표.** 스튜던트화 잔차를 정규 분포와 비교하는 구간화된 확률대확률 도표입니다. 도표화된 점의 기울기가 정규선보다 덜 가파른 경우 잔차는 정규 분포보다 큰 변동을 표시하며, 기울기가 더 가파른 경우 잔차는 정규 분포보다 적은 변동을 표시합니다. 도표화된 점에 S 형태 곡선이 있으면 잔차 분포가 비대칭됩니다.

타. 이상치 (선형 모델)

이 테이블에는 모델에 대해 불필요한 영향력을 발휘하는 레코드가 나열되고 레코드 ID(필드 탭에 지정된 경우), 목표값 및 Cook의 거리가 표시됩니다. Cook의 거리는 특정 레코드를 모형 계수 계산에서 제외할 때 모든 레코드의 잔차가 얼마나 변경될 수 있는지에 대한 척도입니다. 큰 Cook의 거리는 레코드를 제외하면 계수가 상당히 변경됨을 나타내므로 영향력이 큰 것으로 고려되어야 합니다.

영향력이 큰 레코드는 주의 깊게 관찰하여 모델 추정에서 덜 중요하게 고려할 수 있을지, 허용 가능한 임계값에 대해 이상값을 자를지, 또는 영향력이 큰 레코드를 완전히 제거할지 결정해야 합니다.

파. 효과 (선형 모델)

이 보기는 모델에서 각 효과의 크기를 표시합니다.

유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **다이아그램.** 예측자 중요도를 줄여 효과가 위에서 아래로 정렬되는 차트입니다. 다이아그램의 연결선은 효과 유의수준을 기준으로 가중되며 선이 굵을수록 더 유의한 효과(더 작은 p-값)입니다. 마우스 커서를 연결선 위에 놓으면 p-값 및 효과의 중요도를 알려주는 도구 팁이 표시됩니다. 이는 기본값입니다.
- **테이블.** 전체 모형과 개별 모형 효과에 대한 ANOVA 테이블입니다. 예측자 중요도를 줄여 개별 효과가 위에서 아래로 정렬됩니다. 기본적으로는 테이블이 축소되어 전체 모델의 결과만 보여줌에 유의하십시오. 개별 모형 효과의 결과를 보려면 테이블에서 수정된 모형 셀을 클릭합니다.

예측변수 중요도. 보기에 표시되는 예측자를 제어하는 예측자 중요도 슬라이더가 있습니다. 모델을 변경하지는 않지만 가장 중요한 예측자에 집중할 수 있습니다. 기본적으로 상위 10개 효과가 표시됩니다.

유의수준. 예측자 중요도를 기준으로 표시되는 것 외에 보기에 표시되는 효과를 더욱 제어하는 유의수준 슬라이더가 있습니다. 슬라이더 값보다 큰 유의수준 값이 있는 효과는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 효과에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의수준을 기준으로 필터링된 효과가 없습니다.

하. 계수 (선형 모델)

이 보기는 모델에서 각 계수의 값을 표시합니다. 요인(범주형 예측자)이 모델 내에서 코딩된 지표이므로 요인을 포함하는 **효과**에는 일반적으로 여러 관련 **계수**가 있으며, 중복(참조) 모수에 해당하는 범주를 제외하고 각 범주에 하나씩 있습니다.

유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **다이아그램.** 먼저 절편을 표시한 다음 예측자 중요도를 줄여 위에서 아래로 효과를 정렬하는 차트입니다. 요인을 포함하는 효과 내에서 계수가 데이터 값의 오름차순으로 정렬됩니다. 다이어그램의 연결선은 계수의 부호에 따라 색상이 지정되고(다이아그램 키 참조) 계수 유의수준을 기준으로 가중되며 선이 굵을수록 더 유의한 계수(더 작은 p -값)입니다. 마우스 커서를 연결선 위에 놓으면 계수 값, p -값, 모수와 연결된 효과의 중요도를 보여주는 도구 팁이 표시됩니다. 이것이 기본 유형입니다.
- **테이블.** 개별 모형 계수의 값, 유의수준 검정 및 신뢰구간을 표시합니다. 절편 이후, 예측자 중요도를 줄여 효과가 위에서 아래로 정렬됩니다. 요인을 포함하는 효과 내에서 계수가 데이터 값의 오름차순으로 정렬됩니다. 기본적으로는 테이블이 축소되어 각 모델 모수의 계수, 유의수준 및 중요도만 표시됨에 유의하십시오. 표준 오차, t 통계 및 신뢰구간을 보려면 테이블에서 계수 셀을 클릭합니다. 테이블에서 모델 모수의 이름 위에 마우스 커서를 놓으면 모수의 이름, 모수와 연결된 효과, (범주형 예측자의 경우) 모델 모수와 연결된 값 레이블을 보여주는 도구 팁이 표시됩니다. 이것은 자동 데이터 준비에서 유사한 범주의 범주형 예측자를 병합할 때 만들어지는 새 범주를 보려는 경우에 특히 유용합니다.

예측변수 중요도. 보기에 표시되는 예측자를 제어하는 예측자 중요도 슬라이더가 있습니다. 모델을 변경하지는 않지만 가장 중요한 예측자에 집중할 수 있습니다. 기본적으로 상위 10개 효과가 표시됩니다.

유의수준. 예측자 중요도를 기준으로 표시되는 것 외에 보기에 표시되는 계수를 더욱 제어하는 유의수준 슬라이더가 있습니다. 슬라이더 값보다 큰 유의수준 값이 있는 계수는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 계수에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의수준을 기준으로 필터링된 계수가 없습니다.

거. 추정 평균(선형 모델)

유의한 예측변수에 대해 표시되는 차트입니다. 차트는 다른 모든 예측변수를 상수로 유지하면서 수직축에 목표의 모델 추정값을 표시하고 수평축에 예측변수의 각 값을 표시합니다. 목표에서 각 예측변수 계수의 효과를 시각화하는 데 유용합니다.

 **참고:** 유의한 예측자가 없는 경우, 추정 평균이 생성되지 않습니다.

너. 모델 작성 요약 (선형 모델)

모델 선택 알고리즘으로 모델 선택 설정에서 **없음** 외에 다른 항목을 선택한 경우 모델 작성 프로세스에 대한 세부 정보를 제공합니다.

단계별 전진. 단계별 전진이 선택 알고리즘인 경우, 테이블에 단계적 알고리즘의 마지막 10개 단계가 표시됩니다. 각 단계에 대해 선택 기준값 및 해당 단계에서 모델의 효과가 표시됩니다. 각 단계가 모델에 얼마나 기여하는지 알 수 있습니다. 지정된 단계에서 모델에 어떤 효과가 있는지 쉽게 알 수 있도록 각 열에서 행을 정렬할 수 있습니다.

최적 서브세트. 최적 서브세트가 선택 알고리즘인 경우, 테이블에 상위 10개 모델이 표시됩니다. 각 모델에 대해 선택 기준값 및 모델의 효과가 표시됩니다. 상위 모델의 안정성을 알 수 있으며, 차이가 별로 없는 유사한 효과가 많은 경우 "상위" 모델에서 상당히 신뢰할 수 있습니다. 아주 다른 효과가 있는 경우 몇몇 효과가 매우 비슷하므로 결합하거나 하나를 제거해야 합니다. 지정된 단계에서 모델에 어떤 효과가 있는지 쉽게 알 수 있도록 각 열에서 행을 정렬할 수 있습니다.

더. 설정 (선형 모델)

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 대상 필드의 이름이며 접두부 L -이 붙습니다. 예를 들어, *sales*라는 이름의 대상 필드의 경우, 새 필드 이름은 L -*sales*가 됩니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우, 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.

- 이 모형의 SQL 생성 이 옵션을 선택하면 데이터베이스에서 모델 스코어를 계산할 원형 SQL 을 생성합니다.

참고: 이 옵션은 결과를 빠르게 제공할 수 있지만, 모델의 복잡도가 증가하여 원형 SQL 의 크기와 복잡도도 증가할 수 있습니다.

- 데이터베이스 외부 스코어 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 폐치하고 SPSS Modeler에서 스코어를 계산합니다.

(2) Linear-AS 노드

IBM® SPSS® Modeler에는 두 가지 다른 버전의 선형 노드가 있습니다.

- 선형은 IBM SPSS Modeler Server에서 실행되는 기존 노드입니다.
- Linear-AS는 IBM SPSS Analytic Server에 연결되었을 때 실행할 수 있습니다.

선형 회귀는 숫자 입력 필드 값에 기반하여 레코드를 분류하는 일반적인 통계 기법입니다. 선형 회귀는 예측 및 실제 출력 값 사이의 차이를 최소화하는 직선 또는 곡선에 적합합니다.

요구사항. 숫자 필드 및 범주형 예측자만 선형 회귀 모형에서 사용할 수 있습니다. 정확히 하나의 목표 필드(역할이 목표로 설정됨) 및 하나 이상의 예측자(역할이 입력으로 설정됨)를 보유해야 합니다. 역할이 모두 또는 없음인 필드는 비슷자 필드이므로 무시됩니다. (필요한 경우 비슷자 필드는 파생 노드를 사용하여 기록할 수 있습니다.)

강도. 선형 회귀 모형은 비교적 단순하며 예측 생성을 위해 쉽게 해석되는 수학적 공식을 제공합니다. 선형 회귀는 장기적으로 안정된 통계 프로시저이므로 이 모델의 특성도 널리 알려져 있습니다. 또한 보통 선형 모델은 빠르게 훈련할 수 있습니다. 선형 노드에서는 방정식에서 중요하지 않은 입력 필드를 제거하기 위해 자동 필드 선택에 대한 방법을 제공합니다.

참고: 목표 필드가 연속적 범위가 아니라 범주형인 경우(예: 예/아니오 또는 이탈/이탈하지 않음) 로지스틱 회귀분석을 대안으로 사용할 수 있습니다. 또한 로지스틱 회귀분석에서는 비슷자 입력에 대한 지원도 제공하므로 이러한 필드를 기록하지 않아도 됩니다. 자세한 정보는 로지스틱 노드의 내용을 참조하십시오.

① Linear-AS 모델

선형 모델은 목표와 하나 이상의 예측자 사이의 선형 관계를 기반으로 연속형 목표를 예측합니다.

선형 모델은 비교적 단순하며 스코어링에 대해 쉽게 해석되는 수식을 제공합니다. 이러한 모델의 특성은 같은 데이터 세트의 다른 모델 유형(예: 신경망 또는 의사결정 트리)과 비교하여 잘 이해되며 일반적으로 아주 빨리 작성될 수 있습니다.

예제. 주택 소유자의 보험 청구에 대해 조사하기 위한 리소스가 제한된 보험 회사가 보험 청구액을 추정하기 위한 모델을 만들고자 합니다. 이 모델을 서비스 센터에 배포하면 영업 담당자는 고객과 통화하면서 청구 정보를 입력하여 지난 데이터를 토대로 '예상' 청구 비용을 즉시 산출할 수 있습니다.

필드 요구사항. 목표와 하나 이상의 입력이 있어야 합니다. 기본적으로 모두 또는 없음의 사전 정의된 역할이 있는 필드는 사용되지 않습니다. 목표가 연속형(척도)이어야 합니다. 예측자(입력)에 측정 수준 제한 사항이 없습니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되며 연속형 필드는 공변량으로 사용됩니다.

가. 기본(linear-AS 모델)

절편 포함. 이 옵션은 X축이 0일 때 Y축에 오프셋을 포함합니다. 절편은 보통 모델에 포함됩니다. 데이터가 선형 회귀로 전달된다고 가정할 경우에는 절편을 제외할 수 있습니다.

이원 상호작용 고려. 이 옵션은 모델이 가능한 각각의 입력 쌍을 비교하여 하나의 입력 쌍 추세가 다른 입력 쌍 추세에 영향을 주는지 확인하도록 합니다. 영향을 주는 경우, 해당 입력은 계획 행렬에 포함될 가능성이 더 큼니다.

계수 추정에 대한 신뢰구간(%). 계수 보기에서 모델 계수의 추정값을 계산하는 데 사용되는 신뢰 구간입니다. 0보다 크고 100보다 작은 값을 지정하십시오. 기본값은 95입니다.

범주형 예측변수에 대한 정렬 순서. 이 제어는 "마지막" 범주를 결정하기 위해 요인(범주형 입력)의 범주 순서를 결정합니다. 입력이 범주형이 아니거나 사용자 정의 참조 범주가 지정된 경우 정렬 순서 설정은 무시됩니다.

나. 모델 선택(linear-AS 모델)

모델 선택 방법. 모델 선택 방법(자세한 내용은 아래 참조) 중 하나를 선택하거나 주효과 모델 향으로 단순히 사용 가능한 예측변수를 모두 입력하는 **모든 예측변수 포함**을 선택하십시오. 기본적으로 **단계별 전진**이 사용됩니다.

단계별 전진 선택. 모델에 아무 효과 없이 시작하여 단계적 기준에 따라 더는 추가하거나 제거할 수 없을 때까지 한번에 한 단계에서 효과를 추가 및 제거합니다.

- **입력/제거 기준.** 모델에 효과를 추가해야 할지 제거해야 할지 결정하는 데 사용되는 통계입니다. **정보 기준(AICC)**은 모델에 제공된 학습 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **F 통계량**은 모델 오차에서 향상도의 통계 검정을 기준으로 합니다. **수정된 R 제곱**은 학습 세트의 적합도를 기준으로 하며 너무 복잡한 모델에 페널티를

부여하도록 조정됩니다. **과적합 방지 기준(ASE)**은 과적합 방지 세트의 적합도(평균제곱 오차, ASE)를 기준으로 합니다. 과적합 방지 세트는 모델을 학습시키는 데 사용되지 않는 원래 데이터 세트의 약 30%의 무작위 부표본입니다.

F 통계량 외의 다른 기준을 선택한 경우, 각 단계의 기준에서 최대 양의 증가에 해당하는 효과가 모델에 추가됩니다. 기준에서 감소에 해당하는 모델의 효과는 제거됩니다.

F 통계량을 기준으로 선택한 경우, 각 단계에서 지정한 임계값보다 작은 최소 p-값이 있는 효과가(다음보다 작은 p-값이 있는 효과 포함) 모델에 추가됩니다. 기본값은 0.05입니다. 지정한 임계값보다 큰 p-값이 있는 모델의 효과는(다음보다 큰 p-값이 있는 효과 제거) 제거됩니다. 기본값은 0.10입니다.

- **최종 모델에서 최대 효과 수를 사용자 정의하십시오.** 기본적으로 사용 가능한 모든 효과를 모델에 입력할 수 있습니다. 또는 단계적 알고리즘이 지정된 최대 효과 수가 있는 단계로 끝나는 경우, 알고리즘이 현재 효과 세트로 중지됩니다.
- **최대 단계 수를 사용자 정의하십시오.** 단계적 알고리즘이 특정 단계 수 이후 중지됩니다. 기본적으로 사용 가능한 효과 수는 3회입니다. 또는 양의 정수로 최대 단계 수를 지정하십시오.

최적 서브세트 선택. "가능한 모든" 모델을 확인하거나 최소한 단계별 전진보다 큰 가능한 모델 서브세트를 확인하여 최적 서브세트 기준에 따라 최적 서브세트를 선택합니다. **정보 기준(AICC)**은 모델에 제공된 학습 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **수정된 R 제곱**은 학습 세트의 적합도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **과적합 방지 기준(ASE)**은 과적합 방지 세트의 적합도(평균제곱 오차, ASE)를 기준으로 합니다. 과적합 방지 세트는 모델을 학습시키는 데 사용되지 않는 원래 데이터 세트의 약 30%의 무작위 부표본입니다.

기준의 최대값이 있는 모델이 최적 모델로 선택됩니다.

참고: 최적 서브세트 선택이 단계별 전진 선택보다 계산이 더 집중됩니다. 최적 서브세트가 부스팅, 배깅 또는 아주 큰 데이터 세트와 함께 수행되는 경우, 단계별 전진 선택을 사용하여 작성되는 표준 모델보다 작성하는 데 상당히 많은 시간이 걸릴 수 있습니다.

다. 모델 옵션(Linear-AS 모델)

모델 이름. 대상 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 대상 필드 이름입니다.

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 대상 필드의 이름이며 접두부 L -이 붙습니다. 예를 들어, *sales*라는 이름의 대상 필드의 경우, 새 필드 이름은 L -*sales*가 됩니다.

라. 대화형 출력(linear-AS 모델)

Linear-AS 모델 실행 후, 다음 출력을 사용할 수 있습니다.

모델 정보

모델 정보 보기는 모델에 대한 중요 정보를 제공합니다. 테이블은 다음과 같은 일부 상위 수준 모델 설정을 식별합니다.

- 필드 탭에 지정된 목표의 이름
- 회귀분석 가중값 필드
- 모델 선택 설정에 지정된 모델 빌딩 방법
- 예측변수 수 입력
- 최종 모델에서 예측변수의 개수
- 수정된 Akaike 정보 기준(AICC). AICC는 $-2(\text{제한된})$ 로그 우도를 기반으로 혼합 모델을 선택 및 비교하기 위한 척도입니다. 값이 작을수록 모형이 우수함을 나타냅니다. AICC는 작은 표본 크기의 AIC를 "수정합니다". 표본 크기가 증가함에 따라 AICC는 AIC로 수렴됩니다.
- R 제곱. 이것은 선형 모델의 적합도 척도로, 간혹 결정계수라고도 합니다. 이 항목은 회귀 모형으로 설명한 종속변수의 변동 비율이 됩니다. 값 범위는 0 - 1입니다. 값을 작을수록 모델이 데이터에 적합하지 않음을 의미합니다.
- 수정된 R 제곱

레코드 요약

레코드 요약 보기는 모델에서 포함되고 제외되는 레코드(케이스) 수 및 퍼센트에 대한 정보를 제공합니다.

예측변수 중요도

일반적으로, 가장 중요한 예측자 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하기를 원합니다. 예측자 중요도 차트를 사용하면 모델 추정 시 각 예측자의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측자에 대한 값의 합은 1.0이 됩니다. 예측자 중요도는 모형 정확도와는 관련이 없습니다. 단지 예측 시 각 예측자의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

관측값 별 예측값

수직축에 예측값의 구간화된 산점도를 표시하고 수평축에 관측값을 표시합니다. 이상적으로, 점이 45도 줄에 있어야 합니다. 이 보기는 레코드가 모델에 의해 특히 잘못 예측되었는지 여부를 알려줄 수 있습니다.

마. 설정(linear-AS 모델)

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 대상 필드의 이름이며 접두부 L -이 붙습니다. 예를 들어, *sales*라는 이름의 대상 필드의 경우, 새 필드 이름은 L -*sales*가 됩니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값: 서버 스코어링 어댑터(설치된 경우)를 사용하거나 아니면 프로세스에서 스코어 매기기.** 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우, 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 사용자 모델에 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우, 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어.** 선택한 경우, 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

(3) 로지스틱 노드

로지스틱 회귀분석(명목 회귀라고도 함)은 입력 필드 값에 기반하여 레코드를 분류하는 통계 기법입니다. 이는 선형 회귀와 비슷하지만, 숫자 대신 범주형 목표 필드를 사용합니다. 이항 모델(두 개의 이산형 범주를 포함하는 목표의 경우) 및 다항 모델(셋 이상의 범주를 포함하는 목표의 경우)이 모두 지원됩니다.

로지스틱 회귀분석은 각 출력 필드 범주와 연관된 확률에 입력 필드 값을 상관시키는 방정식 세트를 작성하여 작동합니다. 모델이 생성되면 새 데이터의 확률을 추정하는 데 사용할 수 있습니다. 각 레코드의 경우 가능한 각 출력 범주에 대해 소속 확률이 계산됩니다. 확률이 가장 높은 목표 범주는 해당 레코드의 예측된 출력 값으로 지정됩니다.

이항 예. 통신사업자가 경쟁자에게 빠져나가고 있는 고객 수에 대해 걱정하고 있습니다. 서비스 이용 데이터를 사용하여 이항 모델을 작성하고 이를 통해 다른 제공자로 이전될 가능성이 있는

고객을 예측하고 가능한 한 많은 고객을 보유하도록 제안을 사용자 정의할 수 있습니다. 목표는 서로 다른 2개의 범주(전송될 수도 있고, 전송되지 않을 수도 있음)를 포함하므로 이항 모델이 사용됩니다.

참고: 이항 모델에서만 문자열 필드는 8자로 제한됩니다. 필요한 경우 재분류 노드를 사용하거나 값 익명화 노드를 사용하여 더 긴 문자열을 기록할 수 있습니다.

다항 예. 통신 제공업체가 서비스 사용 패턴을 기준으로 고객층을 세그먼트화하여 고객을 4개의 그룹으로 범주화했습니다. 인구 통계 데이터를 사용하여 소속그룹을 예측하면 다항 모델을 작성하여 잠재 고객을 그룹으로 분류하고 개별 고객에 대한 제안을 사용자 정의할 수 있습니다.

요구사항. 하나 이상의 입력 필드와 둘 이상의 범주를 포함하는 정확히 하나의 범주형 목표 필드. 이항 모델의 경우 목표에서 측정 수준은 *플래그*여야 합니다. 다항 모델의 경우 목표에서 측정 수준은 *플래그*, 또는 두 개 이상의 범주를 포함하는 *명목*일 수 있습니다. *둘 다* 또는 *없음*으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다.

강도. 로지스틱 회귀 모형은 종종 꽤 정확합니다. 이 모델은 기호 및 숫자 입력 필드를 처리할 수 있습니다. 이들은 차선 추측을 쉽게 식별할 수 있도록 모든 목표 범주에 대한 예측 확률을 제공할 수 있습니다. 로지스틱 모델은 소속그룹이 범주형 필드인 경우에 가장 효과적입니다. 소속그룹이 연속 범위 필드의 값에 기반하는 경우(예: 높은 IQ 대 낮은 IQ) 값의 전체 범위에서 제공하는 더 다양한 정보를 활용하도록 선형 회귀를 사용하는 방법을 고려해야 합니다. 또한 필드선택이나 트리 모델과 같은 다른 접근 방식이 대형 데이터 세트에서 더 빠르게 이 작업을 수행할 수 있어도 로지스틱 모델도 자동 필드선택을 수행할 수 있습니다. 마지막으로 로지스틱 모델은 많은 분석가와 데이터 마이너가 자세히 이해하고 있기 때문에 일부는 이를 다른 모델링 기법을 비교할 수 있는 기준선으로 사용할 수 있습니다.

큰 데이터 세트를 처리할 때 고급 출력 옵션인 우도비 검정을 사용하지 않으면 성능을 크게 향상시킬 수 있습니다. 자세한 정보는 로지스틱 회귀분석 고급 출력 주제를 참조하십시오.

중요사항: 임시 디스크 공간이 낮으면 이항 로지스틱 회귀분석이 작성되지 못하며 오류가 표시됩니다. 대형 데이터 세트(10GB 이상)에서 작성하는 경우 동일한 크기의 디스크 여유 공간이 필요합니다. 환경 변수 SPSSTMPDIR을 사용하여 임시 디렉토리의 위치를 설정할 수 있습니다.

① 로지스틱 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

프로시저. 작성할 모델(이항 또는 다항 모델)을 지정합니다. 대화 상자에서 사용 가능한 옵션은 선택한 모델링 프로시저 유형에 따라 달라집니다.

- **이항.** 목표 필드가 두 개의 이산값(이분형)을 포함하는 플래그 또는 명목 필드인 경우(예: 예/아니오, 설정/해제, 남성/여성) 사용합니다.
- **다항.** 목표 필드가 셋 이상의 값을 포함하는 명목 필드일 때 사용합니다. **주효과**, **완전요인 모델** 또는 **사용자 정의**를 지정할 수 있습니다.

방정식에 상수항 포함. 이 옵션에서는 결과로 생성된 방정식이 상수항을 포함하는지 여부를 판별합니다. 대부분의 상황에서 이 옵션은 선택한 상태로 두어야 합니다.

이항 모델

이항 모델의 경우 다음 방법과 옵션이 사용 가능합니다.

방법. 로지스틱 회귀 모형을 작성할 때 사용할 방법을 지정합니다.

- **Enter.** 이는 기본 방법으로 모든 항을 방정식에 직접 입력합니다. 모델 작성 시 필드 선택은 수행되지 않습니다.
- **단계별 전진.** 필드 선택의 단계별 전진 방법은 이름이 함축하는 바와 같이 단계별로 방정식을 작성합니다. 이 초기 모형은 방정식에 모형 항(상수 제외)이 없는 가장 단순한 모형입니다. 각 단계마다 모델에 아직 추가되지 않은 항을 평가하여 최상의 항이 모델의 예측력을 상당히 증가시킬 경우 이 항이 추가됩니다. 또한 현재 모델에 있는 항은 모델을 크게 손상시키지 않고도 제거할 수 있는지 판별하기 위해 재평가됩니다. 이 경우 제거됩니다. 프로세스가 반복되고 다른 항이 추가 및/또는 제거됩니다. 모델을 향상시키기 위해 추가할 수 있는 더 이상의 항이 없으며 모델을 손상시키지 않고 제거할 수 있는 더 이상의 항이 없으면 최종 모델이 생성됩니다.
- **단계별 후진.** 단계별 후진은 단계별 전진 방법과 본질적으로 반대입니다. 이 방법에서 초기 모델은 모든 항을 예측자로 포함합니다. 각 단계에서 모델의 항이 평가되고 모델을 크게 손상시키지 않고 제거할 수 있는 항이 제거됩니다. 또한 이전에 제거된 항은 해당 항 중 최상의 항이 모델의 예측력을 크게 높이는지 여부를 판별하기 위해 재평가됩니다. 이 경우 모델로 다시 추가됩니다. 모델을 많이 손상시키지 않고 제거할 수 있는 더 이상의 항이 없으며 모델을 향상시키기 위해 추가할 수 있는 더 이상의 항이 없으면 최종 모델이 생성됩니다.

범주형 입력. 범주형으로 식별된 필드(즉, 플래그, 명목 또는 순서의 측정 수준으로 설정됨)를 나열합니다. 각 범주형 필드에 대해 대비 및 기본 범주를 지정할 수 있습니다.

- **필드 이름.** 이 열은 범주형 입력의 필드 이름을 포함합니다. 이 열에 연속형 또는 수치형 입력을 추가하려면 목록 오른쪽에 있는 필드 추가 아이콘을 클릭하고 필요한 입력을 선택합니다.
- **대비.** 범주형 필드에 대한 회귀계수의 해석은 사용되는 대비에 따라 달라집니다. 대비는 추정된 평균을 비교하기 위해 가설 검정을 어떻게 설정할지 결정합니다. 예를 들어 범주형 필드에 함축적 순서가 있다는 점을 알면(예: 패턴 또는 그룹화) 해당 순서를 모델링하기 위해 대비를 사용할 수 있습니다. 사용 가능한 대비는 다음과 같습니다.
표시기. 대비는 소속 범주가 있는지 여부를 나타냅니다. 이는 기본 방법입니다.

단순. 참조 범주를 제외한 예측자 필드의 각 범주는 참조 범주와 비교됩니다.

차이. 첫 번째 범주를 제외한 예측자 필드의 각 범주는 이전 범주의 평균 효과와 비교됩니다. 역 Helmert 대비라고도 합니다.

Helmert. 마지막 범주를 제외한 예측자 필드의 각 범주는 후속 범주의 평균 효과와 비교됩니다.

반복. 처음 범주를 제외한 예측자 필드의 각 범주는 선행하는 범주와 비교됩니다.

다항. 직교 다항 대비. 범주는 동일한 간격으로 떨어져 있어야 합니다. 다항 대비는 숫자 필드에서만 사용 가능합니다.

편차. 참조 범주를 제외한 예측자 필드의 각 범주는 전체 효과와 비교됩니다.

- **기본 범주.** 선택한 대비 유형에서 참조 범주가 판별되는 방식을 지정합니다. **첫 번째**를 선택하여 입력 필드(문자순으로 정렬됨)의 첫 번째 범주를 사용하거나 **마지막**을 선택하여 마지막 범주를 사용하십시오. 기본 범주는 **범주형 입력** 영역에 나열된 변수에 적용됩니다.

 **참고:** 이 필드는 대비 설정이 차이, Helmert, 반복 또는 다항인 경우 사용할 수 없습니다.

전체 반응에 대한 각 필드 효과의 추정값은 참조 범주와 관련된 각 기타 범주의 우도에서 증가 또는 감소로 계산됩니다. 이를 통해 특정 반응을 제공할 수 있는 필드 및 값을 식별하는 데 도움이 될 수 있습니다.

기본 범주는 출력에서 0.0으로 표시됩니다. 이는 자체를 비교할 경우 빈 결과가 생성되기 때문입니다. 다른 모든 범주는 기본 범주와 관련하여 방정식으로 표시됩니다. 자세한 정보는 로지스틱 너짓 모델 세부사항의 내용을 참조하십시오.

다항 모델

다항 모델의 경우 다음 방법과 옵션이 사용 가능합니다.

방법. 로지스틱 회귀 모형을 작성할 때 사용할 방법을 지정합니다.

- **Enter.** 이는 기본 방법으로 모든 항을 방정식에 직접 입력합니다. 모델 작성 시 필드 선택은 수행되지 않습니다.
- **단계 선택.** 필드 선택의 단계선택법은 이름이 함축하는 바와 같이 단계별로 방정식을 작성합니다. 이 초기 모형은 방정식에 모형 항(상수 제외)이 없는 가장 단순한 모형입니다. 각 단계마다 모델에 아직 추가되지 않은 항을 평가하여 최상의 항이 모델의 예측력을 상당히 증가시킬 경우 이 항이 추가됩니다. 또한 현재 모델에 있는 항은 모델을 크게 손상시키지 않고도 제거할 수 있는지 판별하기 위해 재평가됩니다. 그러한 경우에는 항이 제거됩니다. 프로세스가 반복되고 다른 항이 추가 및/또는 제거됩니다. 모델을 향상시키기 위해 추가할 수 있는 더 이상의 항이 없으며 모델을 손상시키지 않고 제거할 수 있는 더 이상의 항이 없으면 최종 모델이 생성됩니다.
- **전진.** 필드선택의 전진 방법은 모델이 단계적으로 작성된다는 점에서 단계선택법과 유사합니다. 그러나, 이 방법을 사용하는 경우 초기 모델이 가장 단순한 모델이고, 모델이 상수 및 항만 추가할 수 있습니다. 각 단계에서 아직 모델에 없는 항은 모델을 향상시키는 정도에 기반하여 검정되며, 이러한 항 중 최상의 항이 모델에 추가됩니다. 더 이상 추가할 수 있는 항이 없거나 최상의 후보 항이 모델에서 충분한 개선을 보이지 않으면 최종 모델이 생성됩니다.
- **후진.** 후진 방법은 전진 방법과 본질적으로 반대입니다. 이 방법에서 초기 모델은 모든 항을 예측자로 포함하고, 항은 모델에서 제거만 가능합니다. 모형에 거의 기여하지 않는 모델 항은 모델을 크게 손상시키지 않고 제거할 수 있는 항이 없을 때까지 하나씩 제거되며, 이후에 최종 모델이 생성됩니다.
- **단계별 후진.** 단계별 후진 방법은 본질적으로 단계선택법의 반대 개념입니다. 이 방법에서 초기 모델은 모든 항을 예측자로 포함합니다. 각 단계에서 모델의 항이 평가되고 모델을 크게 손상시키지 않고 제거할 수 있는 항이 제거됩니다. 또한 이전에 제거된 항은 해당 항 중 최상의 항이 모델의 예측력을 크게 높이는지 여부를 판별하기 위해 재평가됩니다. 이 경우 모델로 다시 추가됩니다. 모델을 많이 손상시키지 않고 제거할 수 있는 더 이상의 항이 없으며 모델을 향상시키기 위해 추가할 수 있는 더 이상의 항이 없으면 최종 모델이 생성됩니다.

참고: 단계별 전진 및 후진을 포함한 자동 방법은 적응력이 높은 학습 방법이며 학습 데이터의 과적합 경향이 높습니다. 이 방법을 사용하는 경우 새 데이터 또는 파티션 노드를 사용하여 작성된 검증용 검정 표본을 통해 결과로 생성된 모델의 유효성을 확인하는 것이 특히 중요합니다.

목표의 기본 범주. 참조 범주가 판별되는 방식을 지정합니다. 이는 목표의 다른 모든 범주에 대한 회귀분석 방정식이 평가되는 기준선으로 사용됩니다. **첫 번째**를 선택하여 현재 목표 필드(문자순으로 정렬됨)의 첫 번째 범주를 사용하거나 **마지막**을 선택하여 마지막 범주를 사용하십시오. 또는 **지정**을 선택하여 특정 범주를 선택하고 목록에서 원하는 값을 선택할 수 있습니다. 사용 가능한 값은 유형 노드의 각 필드에서 정의할 수 있습니다.

종종 기본 범주로 거의 고려하지 않는 범주(예: 특가품)를 지정합니다. 그러면 다른 범주는 상대적인 방식으로 이 기본 범주와 관련되어 고유한 범주에 존재할 수 있는 방법을 식별합니다. 이를 통해 특정 반응을 제공할 수 있는 필드 및 값을 식별하는 데 도움이 될 수 있습니다.

기본 범주는 출력에서 0.0으로 표시됩니다. 이는 자체를 비교할 경우 빈 결과가 생성되기 때문입니다. 다른 모든 범주는 기본 범주와 관련하여 방정식으로 표시됩니다. 자세한 정보는 로지스틱 너짓 모델 세부사항의 내용을 참조하십시오.

모델 유형. 모델에서 항을 정의하는 세 가지 옵션이 있습니다. **주효과** 모델은 개별적으로 입력 모델만 포함하고 입력 필드 사이의 상호작용(승법 효과)은 검정하지 않습니다. **완전요인** 모델은 입력 필드 주효과와 함께 모든 상호작용을 포함합니다. 완전요인 모델은 보다 효과적으로 복잡한 관계를 캡처할 수 있지만, 해석이 훨씬 더 어렵고 과적합으로 어려움을 겪을 수 있습니다. 가능한 조합의 수가 잠재적으로 많을 수 있으므로 자동 필드선택 방법(입력 이외의 방법)은 완전요인 모델에서 사용되지 않습니다. **사용자 정의** 모델은 사용자가 지정한 항(주효과 및 상호작용)만 포함합니다. 이 옵션을 선택하면 모형 항 목록을 사용하여 모형에서 항을 추가하거나 제거합니다.

모형 항. 사용자 정의 모델을 작성할 때에는 모델의 항을 명시적으로 지정해야 합니다. 목록에는 모델 항의 현재 세트가 표시됩니다. 모형 항 목록의 오른쪽에 있는 단추를 사용하여 모형 항을 추가 및 제거할 수 있습니다.

- 모형에 항을 추가하려면 **새 모형 항 추가** 단추를 클릭하십시오. 자세한 정보는 로지스틱 회귀 모형에 항 추가 주제를 참조하십시오.
- 항을 삭제하려면 원하는 항을 선택하고 **선택한 모형 항 삭제** 단추를 클릭하십시오.

② 로지스틱 회귀 모형에 항 추가

사용자 정의 로지스틱 회귀 모형을 요청하면 로지스틱 회귀 모형 탭에서 **새 모형 항 추가** 단추를 클릭하여 모형에 항을 추가할 수 있습니다. 항을 지정할 수 있는 새 항 대화 상자가 열립니다.

추가할 항 유형. 사용 가능한 필드 목록에서 입력 필드 선택에 따라 모델에 항을 추가하는 여러 방법이 있습니다.

- **단일 상호작용.** 선택한 모든 필드의 상호작용을 나타내는 항을 삽입합니다.
- **주효과.** 선택된 각 입력 필드마다 주효과 항(필드 자체)을 하나씩 삽입합니다.
- **모든 2원 효과 상호작용.** 선택한 입력 필드의 가능한 각 쌍에서 이원 상호작용 항(입력 필드의 곱)을 삽입합니다. 예를 들어, 사용 가능한 필드 목록에서 입력 필드 A , B , C 를 선택한 경우 이 방법은 $A * B$, $A * C$, $B * C$ 항을 삽입합니다.
- **모든 3원 효과 상호작용.** 선택한 입력 필드의 가능한 각 조합(한 번에 3개 항 사용)에서 삼원 상호작용 항(입력 필드의 곱)을 삽입합니다. 예를 들어, 사용 가능한 필드 목록에서 입력 필드 A , B , C , D 를 선택한 경우, 이 방법은 $A * B * C$, $A * B * D$, $A * C * D$, $B * C * D$ 항을 삽입합니다.
- **모든 4원 효과 상호작용.** 선택한 입력 필드의 가능한 각 조합(한 번에 4개 항 사용)에서 4원 상호작용 항(입력 필드의 곱)을 삽입합니다. 예를 들어 사용 가능한 필드 목록에서 입력 필드 A , B , C , D , E 를 선택한 경우, 이 방법은 $A * B * C * D$, $A * B * C * E$, $A * B * D * E$, $A * C * D * E$, $B * C * D * E$ 항을 삽입합니다.

사용 가능한 필드. 모형 항을 구성할 때 사용할 사용 가능한 입력 필드를 나열합니다.

미리보기. 선택한 필드 및 항 유형에 따라 **삽입**을 클릭한 경우 모델에 추가되는 항을 표시합니다.

삽입. 모델에 항을 삽입하고(필드의 현재 선택 및 항 유형을 기준으로 하여) 대화 상자를 닫습니다.

③ 로지스틱 노드 고급 옵션

로지스틱 회귀분석에 대한 자세한 지식이 있는 경우 고급 옵션을 통해 훈련 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

척도(다항 모델만 해당). 모수 공분산행렬의 추정값을 조정하는 데 사용되는 산포도 배율 값을 지정할 수 있습니다. **Pearson**에서는 Pearson 카이제곱 통계를 사용하여 배율 값을 추정합니다. **편차**에서는 편차 함수(우도비 카이제곱) 통계를 사용하여 배율 값을 추정합니다. 또한 사용자 정의 배율 값을 지정할 수도 있습니다. 이때 양의 숫자 값이어야 합니다.

모든 확률 추가. 이 옵션을 선택하면 출력 필드의 각 범주에 대한 확률을 노드에서 처리하는 각 레코드에 추가됩니다. 이 옵션을 선택하지 않으면 예측 범주의 확률만 추가됩니다.

예를 들어, 세 개 범주가 있는 다항 모델의 결과를 포함하는 테이블은 다섯 개의 새 열을 포함합니다. 한 열은 올바르게 예측된 결과의 확률, 다음 열은 이 예측이 적중했는지 또는 빗나갔는지 확률, 다음에 나오는 세 개 열은 각 범주의 예측이 적중했는지 또는 빗나갔는지 확률을 표시합니다. 자세한 정보는 로지스틱 모델 너깃의 내용을 참조하십시오.

참고: 이 옵션은 이항 모델의 경우 항상 선택됩니다.

비정칙성 공차. 비정칙성을 확인할 때 사용되는 공차를 지정합니다.

수렴. 이 옵션으로 모델 수렴에 대한 모수를 제어할 수 있습니다. 모델을 실행할 때 수렴 설정은 여러 다른 모수가 어느 정도 적합한지 확인하기 위해 모수가 반복적으로 실행되는 횟수를 제어합니다. 모수를 더 자주 시도할 수록 결과에 더 근접합니다(즉, 결과가 수렴됨). 자세한 정보는 로지스틱 회귀분석 수렴 옵션의 내용을 참조하십시오.

출력. 이 옵션을 사용하면 노드에서 작성한 모델 너깃의 고급 출력에 표시되는 추가 통계를 요청할 수 있습니다. 자세한 정보는 로지스틱 회귀분석 고급 출력 주제를 참조하십시오.

단계. 이 옵션을 사용하면 단계 선택, 전진, 후진 또는 단계별 후진 추정 방법을 포함하는 필드를 추가 및 제거하는 기준을 제어할 수 있습니다. (이 단추는 입력 방법을 선택한 경우 사용할 수 없습니다.) 자세한 정보는 로지스틱 회귀분석 단계별 옵션의 내용을 참조하십시오.

④ 로지스틱 회귀분석 수렴 옵션

로지스틱 회귀 모형 추정에 대한 수렴 모수를 설정할 수 있습니다.

최대반복계산. 모델을 추정할 때 최대반복수를 지정합니다.

최대 단계 이분. 단계 이분은 추정 프로세스에서 복잡도를 처리하기 위해 로지스틱 회귀분석에서 사용하는 기법입니다. 일반적인 상황에서는 기본 설정을 사용해야 합니다.

로그-우도 수렴. 로그-우도의 상대값 변화가 이 값 미만이면 반복이 중지됩니다. 값이 0인 경우 이 기준은 적용되지 않습니다.

모수 수렴. 모수 추정값의 절대값 또는 상대값 변화가 이 값 미만이면 반복이 중지됩니다. 값이 0인 경우 이 기준은 적용되지 않습니다.

델타(다항 모델만 해당). 각 빈 셀에 추가할 0에서 1 사이의 값을 지정할 수 있습니다(입력 필드와 출력 필드 값의 조합). 그러면 추정 알고리즘이 데이터에서 레코드 수에 상대적인 필드 값의 가능한 많은 조합이 있는 데이터를 다룰 때 도움이 될 수 있습니다. 기본값은 0입니다.

⑤ 로지스틱 회귀분석 고급 출력

회귀 모델 너깃의 고급 출력에 표시할 선택적 출력을 선택하십시오. 고급 출력을 보려면 모델 너깃을 찾아보고 **고급** 탭을 클릭하십시오. 자세한 정보는 로지스틱 모델 너깃 고급 출력의 내용을 참조하십시오.

이항검정 옵션

모델에서 생성할 출력 유형을 선택합니다. 자세한 정보는 로지스틱 모델 너깃 고급 출력의 내용을 참조하십시오.

표시. 각 단계마다 결과를 표시하거나 모든 단계를 완료할 때까지 기다릴 것인지 여부를 선택합니다.

exp에 대한 신뢰구간(B). 표현식의 각 계수(베타로 표시됨)에 대한 신뢰구간을 선택합니다. 신뢰구간 수준을 지정합니다(기본값은 95%).

잔차 진단. 잔차의 케이스별 진단 테이블을 요청합니다.

- **밖에 나타나는 이상치(표준 편차).** 나열된 변수의 절대 표준화 값이 적어도 지정한 값인 잔차 케이스만 나열합니다. 기본값은 2입니다.

- **모든 케이스.** 잔차의 케이스별 진단 테이블에 있는 모든 케이스를 포함합니다.

참고: 이 옵션에서는 각 입력 레코드를 나열하므로, 이로 인해 모든 레코드마다 한 줄씩 사용하여, 보고서에 너무 큰 테이블이 생성될 수 있습니다.

분류 분리점. 이를 통해 케이스 분류에 대한 절단점을 판별할 수 있습니다. 예측값이 분류 분리점보다 작은 케이스는 음수로 분류되고 예측값이 분류 분리점을 초과하는 케이스는 양수로 분류됩니다. 기본값을 변경하려면 0.01과 0.99 사이의 값을 입력합니다.

다항 옵션

모델에서 생성할 출력 유형을 선택합니다. 자세한 정보는 로지스틱 모델 너짓 고급 출력의 내용을 참조하십시오.

참고: **우도비 검정** 옵션을 선택하면 로지스틱 회귀 모형을 작성하는 데 필요한 처리 시간이 크게 늘어납니다. 모델 작성 시간이 너무 오래 걸리면 이 옵션을 사용하지 않거나 대신 Wald 및 스코어 통계를 활용하는 방법을 고려하십시오. 자세한 정보는 로지스틱 회귀분석 단계별 옵션의 내용을 참조하십시오.

반복계산 히스토리 출력수준. 고급 출력에서 반복 상태를 인쇄하는 단계 구간을 선택합니다.

신뢰구간. 방정식에서 계수의 신뢰구간. 신뢰구간 수준을 지정합니다(기본값은 95%).

⑥ 로지스틱 회귀분석 단계별 옵션

이 옵션을 사용하면 단계 선택, 전진, 후진 또는 단계별 후진 추정 방법을 포함하는 필드를 추가 및 제거하는 기준을 제어할 수 있습니다.

모델에 포함된 항 수(다항 모델만 해당). 후진 및 단계별 후진 모델인 경우 모델에서 최소 항 수를 지정하고, 전진 및 단계선택법 모델인 경우 최대 항 수를 지정할 수 있습니다. 0보다 큰 최소값을 지정하면 통계 기준에 따라 일부 항을 제거했어도 모델이 많은 항을 포함합니다. 전진, 단계 선택, 입력 모델에서 최소값 설정은 무시됩니다. 최대값을 지정하는 경우 일부 항은 통계 기준에 기반하여 선택되었어도 모델에서 생략될 수 있습니다. **최대값 지정** 설정은 후진, 단계별 후진, 입력 모델에서 무시됩니다.

입력 기준(다항 모델만 해당). 처리 속도를 극대화 하려면 **스코어**를 선택하십시오. **우도비** 옵션에서는 다소 더 강력한 추정값을 제공할 수 있지만, 계산 시간이 더 오래 걸립니다. 기본 설정은 스코어 통계를 사용하는 것입니다.

제거 기준. 보다 강력한 모델에서 **우도비**를 선택합니다. 모델 작성에 필요한 시간을 단축하려면 **Wald**를 선택할 수 있습니다. 그러나 데이터에서 분리가 전체 또는 절반만 수행된 경우(모델 너

기의 고급 탭을 사용하여 판별 가능) Wald 통계량은 특히 불안정해지며, 이를 사용해서는 안 됩니다. 기본 설정은 우도비 통계를 사용하는 것입니다. 이항 모델의 경우 추가 옵션 조건부가 있습니다. 이 방법에서는 조건부 모수 추정값에 기반한 우도비 통계의 확률을 토대로 제거 검정을 제공합니다.

기준의 유의수준 임계값. 이 옵션을 사용하면 각 필드와 연관된 통계 확률(p 값)에 기반하여 선택 기준을 지정할 수 있습니다. 연관된 p 값이 입력 값보다 작은 경우에만 모델에 필드가 추가되고 p 값이 제거 값보다 큰 경우에만 필드가 제거됩니다. 입력 값은 제거 값보다 작아야 합니다.

입력 또는 제거에 대한 요구 사항(다항 모델만 해당). 일부 애플리케이션에서는 모델이 상호작용 항과 관련된 필드에서 차수가 낮은 항도 포함하지 않는 한, 모델에 상호작용 항을 추가하는 것은 수학적으로 의미가 없습니다. 예를 들어, A 및 B 도 모델에 포함되지 않는 한, 모델에서 $A * B$ 를 포함하지 않는 것이 좋습니다. 이 옵션을 사용하면 단계선택항 선택 중 이러한 종속성을 처리하는 방법을 판별할 수 있습니다.

- **이산형 효과에 대한 계층 구조.** 관련 필드에서 차수가 더 낮은 모든 효과(주효과 또는 필드가 더 적은 상호작용)가 이미 모델에 있는 경우에만 차수가 더 높은 효과(필드가 더 많은 상호작용)가 모델을 입력하고, 동일한 필드를 포함하는 차수가 더 높은 효과가 모델에 있으면 차수가 더 낮은 효과는 제거되지 않습니다. 이 옵션은 범주형 필드에만 적용됩니다.
- **모든 효과의 계층 구조.** 이 옵션은 모든 입력 필드에 적용된다는 점을 제외하고, 이전 옵션과 동일하게 작동합니다.
- **모든 효과 억제.** 효과는 효과에 포함된 모든 효과가 모델에도 포함되는 경우에만 모델에 포함될 수 있습니다. 이 옵션은 연속형 필드가 다소 다르게 처리된다는 점을 제외하고 모든 효과의 계층 구조 옵션과 유사합니다. 효과에서 다른 효과를 포함하도록 하려면 포함된 효과(차수가 더 낮음)는 포함하는 효과(차수가 더 높음)와 관련된 연속형 필드 모두를 포함해야 하고, 포함된 효과의 범주형 필드는 포함하는 효과에 있는 필드의 서브세트여야 합니다. 예를 들어, A 및 B 가 범주형 필드이고 X 가 연속형 필드이면 항 $A * B * X$ 는 항 $A * X$ 와 $B * X$ 를 포함합니다.
- **없음.** 강제로 적용되는 관계는 없습니다. 항은 모델에서 독립적으로 추가되거나 제거됩니다.

(4) 로지스틱 모델 너깃

로지스틱 모델 너깃은 로지스틱 노드에서 추정하는 방정식을 나타냅니다. 여기에는 로지스틱 회귀 모형에서 캡처한 모든 정보와 모델 구조 및 성능에 대한 정보가 포함됩니다. 이러한 유형의 방정식은 Oracle SVM과 같은 다른 모델에서 생성될 수도 있습니다.

로지스틱 모델 너깃을 포함하는 스트림을 실행하는 경우 노드는 모델의 예측 및 연관된 확률을 포함하는 새 두 개 필드를 추가합니다. 새 필드 이름은 예측하는 출력 필드 이름에서 파생되며, 예측 범주의 경우 접두문자는 $\$L-$, 연관된 확률의 경우 $\$LP$ 입니다. 예를 들어, 이름이 *colorpref*인 출력 필드의 경우 새 필드 이름은 $\$L-colorpref$ 및 $\$LP-colorpref$ 입니다. 또한 로

지스틱 노드에서 **모든 확률 추가** 옵션을 선택한 경우 추가 필드는 각 레코드에 대응하는 범주에 속하는 확률을 포함하여 출력 필드의 각 범주에 추가됩니다. 이러한 추가 필드 이름은 출력 필드 값에 따라 지정되며, 접두문자는 LP 입니다. 예를 들어, *colorpref*의 유효한 값이 *Red*, *Green*, *Blue*인 경우 세 개의 새 필드(LP -Red, LP -Green, LP -Blue)가 추가됩니다.

필터 노드 생성. 생성 메뉴에서는 모델 결과에 기반하여 입력 필드를 전달할 새 필터 노드를 작성할 수 있습니다. 다중공선성으로 인해 모델에서 삭제된 필드는 생성된 노드 및 모델에서 사용되지 않은 필드로 필터링됩니다.

① 로지스틱 너깃 모델 세부사항

다항 모델의 경우 로지스틱 모델 너깃의 모델 탭에서는 왼쪽 분할창에 모델 방정식을 포함하는 분할 표시가, 오른쪽에 예측변수 중요도가 있습니다. 이항 모델의 경우 탭은 예측변수 중요도만 표시합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

모델 방정식

다항 모델의 왼쪽 분할창에서는 로지스틱 회귀분석 모델에서 추정된 실제 방정식을 표시합니다. 대상 필드에는 각 범주에 대한 하나의 방정식이 있습니다(단, 기준선 범주 제외). 방정식은 트리 형식으로 표시됩니다. 이 유형의 방정식은 Oracle SVM과 같은 특정 다른 모델에서도 생성할 수 있습니다.

방정식 기준. 예측변수 값 세트가 주어진 경우 목표 범주 확률을 파생하는 데 사용되는 회귀분석 방정식을 표시합니다. 대상 필드의 마지막 범주는 **기준선 범주**로 간주됩니다. 표시된 방정식은 예측변수 값의 특정 세트에 대한 기준선 범주에 상대적으로 다른 목표 범주의 로그-오즈비를 제공합니다. 주어진 예측변수 패턴의 각 범주에 대한 예측 확률은 이러한 로그-오즈비 값에서 파생됩니다.

확률 계산 방법

각 방정식은 기준선 범주에 상대적으로 특정 목표 범주의 로그-오즈비를 계산합니다. **로그-오즈비(로짓**라고도 함)는 지정된 목표 범주 확률을 결과에 자연로그 함수를 적용하는 기준선 범주의 확률로 나눈 비율입니다. 기준선 범주의 경우 자체에 상대적인 범주의 오즈비는 1.0이므로, 로그-오즈비는 0입니다. 모든 계수가 0인 기준선 범주의 함축적인 방정식으로 간주할 수 있습니다.

특정 목표 범주의 로그-오즈비에서 확률을 파생시키려면 해당 범주의 방정식에서 계산된 로짓 값을 사용하고 다음 수식을 적용합니다.

$$P(\text{group}_i) = \exp(g_i) / \sum_k \exp(g_k)$$

여기서 g 는 계산된 로그-오즈비이고 i 는 범주 지수이며, k 는 1부터 목표 범주의 수까지의 범위에 속합니다.

예측자 중요도(Predictor Importance)

선택적으로 모델을 추정할 때 각 예측자의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측자에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측자 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

참고: 예측변수 중요도를 사용하면 다른 유형의 모델보다 로지스틱 회귀분석 계산 시간이 오래 걸릴 수 있으므로, 기본적으로 분석 탭에서는 선택되어 있지 않습니다. 이 옵션을 선택하면 특히 큰 데이터 세트에서 성능이 느려질 수 있습니다.

② 로지스틱 모델 너깃 요약

로지스틱 회귀 모형의 요약에서는 모델을 생성하는 데 사용된 필드 및 설정을 표시합니다. 또한 이 모델링 노드에 첨부된 분석 노드를 실행하면 해당 분석의 정보도 이 섹션에 표시됩니다. 모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

③ 로지스틱 모델 너깃 설정

로지스틱 모델 너깃의 설정 탭에서는 신뢰도, 확률, 성향 스코어, 모델 스코어링 중 SQL 생성에 대한 옵션을 지정합니다. 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능하고 모델 및 목표 유형에 따라 서로 다른 옵션을 표시합니다.

다항 모델

다항 모델의 경우 다음 옵션이 사용 가능합니다.

신뢰도 계산 스코어링 중 신뢰도 계산 여부를 지정합니다.

원시 성향 스코어 계산(플래그 목표만) 플래그 목표만 포함하는 모델의 경우 목표 필드에 지정된 창의 결과 우도를 나타내는 원시 성향 스코어를 요청할 수 있습니다. 표준 예측 및 신뢰도 값 외에도 제공됩니다. 수정된 성향 스코어는 사용할 수 없습니다. 자세한 정보는 모델링 노드 분석 옵션의 내용을 참조하십시오.

모든 확률 추가 출력 필드의 각 범주에 대한 확률을 노드에서 처리하는 각 레코드에 추가하는지 여부를 지정합니다. 이 옵션을 선택하지 않으면 예측 범주의 확률만 추가됩니다. 예를 들어, 세 개 범주를 포함하는 명목 목표의 경우 스코어링 출력은 각 세 개 범주의 열과 함께 예측되는 모든 범주에 대한 확률을 표시하는 네 번째 열을 포함합니다. 범주 *Red*, *Green*, *Blue*의 확률이 각각 0.6, 0.3, 0.1인 경우 예측 범주는 확률 0.6의 *Red*입니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- 기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- 이 모형의 SQL 생성 이 옵션을 선택하면 데이터베이스에서 모델 스코어를 계산할 원형 SQL을 생성합니다.

참고: 이 옵션은 결과를 빠르게 제공할 수 있지만, 모델의 복잡도가 증가하여 원형 SQL의 크기와 복잡도도 증가할 수 있습니다.

- 데이터베이스 외부 스코어 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

참고: 다항 모델의 경우 모든 확률 추가를 선택하면 SQL 생성은 사용할 수 없습니다. 또는 명목 목표를 포함하는 모델의 경우 신뢰도 계산을 선택하면 사용할 수 없습니다. 신뢰도 계산을 포함하는 SQL 생성은 플래그 목표만 포함하는 다항 모델에서 지원됩니다. SQL 생성은 이항 모델에서 사용할 수 없습니다.

이항 모델

이항 모델의 경우 신뢰도 및 확률은 항상 사용 가능하고 이 옵션을 사용하지 못하도록 하는 설정은 사용 불가능합니다. SQL 생성은 이항 모델에서 사용할 수 없습니다. 이항 모델에서 변경할 수 있는 유일한 설정은 원시 성향 스코어를 계산하는 기능입니다. 다항 모델에서 앞서 언급한 대로, 이는 플래그 목표만 포함하는 모델에 적용됩니다. 자세한 정보는 모델링 노드 분석 옵션의 내용을 참조하십시오.

④ 로지스틱 모델 너짓 고급 출력

로지스틱 회귀분석(명목 회귀라고도 함)의 고급 출력에서는 추정된 모델 및 해당 성능에 대한 자세한 정보를 제공합니다. 고급 출력에 포함된 대부분의 정보는 기술과 관련된 정보로, 이 출력을 적절히 해석하려면 로지스틱 회귀분석에 대한 포괄적인 지식을 요구합니다.

경고. 결과에 대한 경고 또는 잠재적 문제점을 표시합니다.

케이스 처리 요약. 처리된 레코드 수를 나열합니다(모델에서 각 기호 필드로 구분됨).

단계 요약(옵션). 자동 필드 선택을 사용하여 각 모델 작성 단계에서 추가되거나 제거되는 효과를 나열합니다.

참고: 단계 선택, 전진, 후진 또는 단계별 후진 방법에서만 표시됩니다.

반복계산과정(옵션). 초기 추정값부터 시작하여 n 번째 반복마다 모수 추정값의 반복계산과정을 표시합니다. 여기서 n 은 인쇄 간격 값입니다. 기본값은 모든 반복($n=1$)을 인쇄하는 것입니다.

모형 적합 정보(다항 모형) 모든 모수 계수가 0인 모델(절편만)에 대한 모델의 우도비 검정(최종)을 표시합니다.

분류(옵션). 퍼센트로 실제 출력 필드 값과 예측 값의 행렬을 표시합니다.

적합도 카이제곱 통계량(옵션). Pearson 및 우도비 카이제곱 통계량을 표시합니다. 이 통계는 훈련 데이터에 대한 모델의 과적합을 검정합니다.

Hosmer 및 Lemeshow 적합도(옵션). 케이스를 위험도의 십분위수로 그룹화하고 각 십분위수 내의 관측 확률을 기대 확률에 비교하는 결과를 표시합니다. 이 적합도 통계량은 다항 모델, 특히 연속형 공변량을 포함하는 모델과 표본 크기가 작은 연구에 전통적인 적합도 통계량보다 효과적입니다.

유사 R-제곱(옵션). 모형 적합의 Cox 및 Snell, Nagelkerke, McFadden R 제곱 측도를 표시합니다. 이러한 통계는 선형 회귀에서 R -제곱 통계와 유사한 특면이 있습니다.

단조성 측도(옵션). 일치 쌍, 불일치 쌍, 데이터에서 연결된 쌍의 수와 함께 각각이 나타내는 총 쌍의 수에 대한 퍼센트를 표시합니다. Somer의 D , Goodman과 Kruskal의 감마, Kendall의 타우- a , 일치 지수 C 도 이 테이블에 표시됩니다.

정보 기준(옵션). AIC(Akaike's information criterion) 및 Schwarz의 베이저안 정보 기준(BIC)을 표시합니다.

우도비 검정(옵션). 모형 효과의 계수가 통계적으로 0과 다른지 여부를 검정하는 통계를 표시합니다. 유의적 입력 필드는 출력(레이블이 *유의확률임*)에서 유의 수준이 매우 작은 필드입니다.

모수 추정값(옵션). 방정식 계수의 추정값, 해당 계수의 검정, 레이블이 *Exp(B)*인 계수에서 파생된 오즈비, 해당 오즈비의 신뢰구간을 표시합니다.

근사 공분산/상관계수 행렬(옵션). 계수 추정의 상관계수 및/또는 근사 공분산을 표시합니다.

관측빈도와 예측빈도(옵션). 각 공변량 패턴의 경우 각 출력 필드 값의 관측빈도와 예측빈도를 표시합니다. 이 테이블은 숫자 입력 필드를 포함하는 모델에서 특히 클 수 있습니다. 결과로 생성된 테이블이 너무 커서 실용적이지 못하면 생략되고 경고가 표시됩니다.

(5) PCA/요인 노드

PCA/요인 노드에서는 강력한 데이터 축소 기법을 제공하여 데이터의 복잡도를 줄입니다. 이때 비슷하지만 다른 두 가지 접근 방식이 제공됩니다.

- **주성분분석(PRINCALS)**에서는 구성요소가 서로 직교(수직)인 전체 필드 세트에서 분산을 캡처할 때 최상의 작업을 수행하는 입력 필드의 선형 조합을 찾습니다. PCA는 공유 및 고유 분산 모두를 포함하여 모든 분산에 초점을 맞춥니다.
- **요인 분석**은 기본 개념 또는 관측 필드 세트 내 상관관계 패턴을 설명하는 요인을 식별하려고 합니다. 요인 분석은 공유 분산에만 초점을 맞춥니다. 특정 필드에 고유한 분산은 모델 추정 시 고려되지 않습니다. 요인/PCA 노드에서는 여러 요인 분석 방법을 제공합니다.

두 접근 방식 모두 목표는 원래 필드 세트의 정보를 효과적으로 요약하는 소수의 파생된 필드를 찾는 것입니다.

요구사항. 숫자 필드만 PCA-요인 모델에서 사용할 수 있습니다. 요인 분석 또는 PCA를 추정하려면 역할이 *입력* 필드로 설정된 하나 이상의 필드가 필요합니다. 역할이 *목표*, *모두* 또는 *없음*으로 설정된 필드는 비슷자 필드이므로 무시됩니다.

강도. 요인 분석과 PCA는 많은 정보 콘텐츠를 포기하지 않고도 효과적으로 데이터의 복잡도를 줄일 수 있습니다. 이러한 기법을 사용하면 원시 입력 필드보다 빠르게 실행되는 더 강력한 모델을 작성할 수 있습니다.

① PCA/요인 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

추출 방법. 데이터 축소에 사용할 방법을 지정합니다.

- **주성분.** 기본 방법으로, 입력 필드를 요약하는 구성요소를 찾기 위해 PCA를 사용합니다.
- **가중되지 않은 최소제곱법.** 이 요인 분석 방법은 입력 필드 가운데 관계(상관관계)의 패턴을 가장 잘 재현할 수 있는 요인 세트를 검색하는 방식으로 작동합니다.
- **일반화 최소제곱법.** 이 요인 분석 방법은 많은 유일접근(공유되지 않음) 분산을 포함하는 필드의 강조를 해제하기 위해 가중치를 사용한다는 점을 제외하고 가중되지 않은 최소제곱법과 유사합니다.
- **최대우도.** 이 요인 분석 방법에서는 해당 관계 양식에 대한 가정에 기반하여 입력 필드에서 관계(상관관계)의 관측된 패턴을 생성할 수 있는 요인 방정식을 생성합니다.
- **주축 요인 추출.** 이 요인 분석 방법은 공유되는 분산에만 초점을 맞춘다는 점을 제외하고 주성분 방법과 매우 유사합니다.
- **알파 요인 추출.** 이 요인 분석 방법은 분석의 필드를 잠재적 입력 필드 환경에서 표본으로 추출하는 방법을 고려합니다. 이 경우 요인의 통계 신뢰도를 최대화합니다.
- **이미지 요인 추출.** 이 요인 분석 방법은 데이터 추정을 사용하여 공통 분산을 고립시키고 이를 설명하는 요인을 찾습니다.

② PCA/요인 노드 고급 옵션

요인 분석 및 PCA에 대한 자세한 지식이 있는 경우 고급 옵션을 통해 훈련 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

결측값. 기본적으로 IBM® SPSS® Modeler에서는 모델에 사용된 모든 필드의 유효한 값을 포함하는 레코드만 사용합니다. (이 기능을 때때로 결측값의 **목록별 삭제**라고도 합니다.) 결측 데이터가 많은 경우 이 접근 방식을 사용하면 너무 많은 레코드가 제거되므로 데이터가 부족하여 좋은 모델을 생성하지 못할 수도 있습니다. 이 경우 **완전한 레코드만 사용** 옵션을 선택 취소할 수 있습니다. 그러면 IBM SPSS Modeler에서는 일부 필드에 결측값이 있는 레코드를 포함하여 모델을 추정할 수 있을 만큼 많은 정보를 사용하려고 합니다. (이 기능을 때때로 결측값의 **대응별 삭제**라고도 합니다.) 그러나 일부 상황에서 이러한 방식으로 불완전한 레코드를 사용하면 모델 추정 시 계산상의 문제점이 발생할 수 있습니다.

필드. 모델 추정 시 입력 필드의 공분산 행렬이나 상관행렬(기본값) 중 사용할 항목을 지정합니다.

수렴을 위한 최대 반복. 모델을 추정할 때 최대반복수를 지정합니다.

요인 추출. 입력 필드에서 추출한 요인 수를 선택하는 두 가지 방법이 있습니다.

- **고유값 기준.** 이 옵션은 지정된 기준보다 고유값이 큰 모든 요인 또는 구성요소를 보유합니다. 고유값은 입력 필드 세트에서 분산을 요약하기 위해 각 요인 또는 구성요소의 기능을 측정합니다. 모델은 상관행렬 사용 시 고유값이 지정된 값보다 큰 모든 요인 또는 구성요소를 보유합니다. 공분산 행렬을 사용하는 경우 기준은 평균 고유값에 지정된 값을 곱한 값입니다. 해당 배율을 통해 이 옵션은 두 유형의 행렬에서 유사한 의미를 지닙니다.
- **최대 수.** 이 옵션은 고유값의 내림차순으로 지정된 수의 요인 또는 구성요소를 보유합니다. 즉, n 개의 상위 고유값에 대응하는 요인 또는 구성요소가 보유되며, 여기서 n 은 지정된 기준입니다. 기본 추출 기준은 5개의 요인/구성요소입니다.

구성요소/요인 행렬 형식. 이 옵션은 요인 행렬(또는 PCA 모델의 경우 구성요소 행렬) 형식을 제어합니다.

- **값 정렬.** 이 옵션을 선택하면 모델 출력에서 로드되는 요인이 숫자를 기준으로 정렬됩니다.
- **아래 값 숨기기.** 이 옵션을 선택하면 지정된 임계값 아래의 스코어는 행렬에서 숨겨지므로 행렬에서 패턴을 더 쉽게 확인할 수 있습니다.

회전. 이 옵션을 사용하면 모델의 회전 방법을 제어할 수 있습니다. 자세한 정보는 PCA/요인 노드 회전 옵션의 내용을 참조하십시오.

③ PCA/요인 노드 회전 옵션

많은 경우 보유한 요인 세트를 수학적으로 회전하면 유용성과 특히 해석 가능성을 높일 수 있습니다. 회전 방법을 선택하십시오.

- **회전 안 함.** 기본 옵션. 회전이 사용되지 않습니다.
- **베리맥스.** 각 요인의 로딩이 높은 필드의 수를 최소화하는 직교 회전 방법. 이는 요인의 해석을 단순화합니다.
- **직접 오블리민.** 사각(비직교) 회전 방법. 델타가 0(기본값)인 경우 솔루션에 기울기가 나타납니다. 델타가 음수에 가까워질수록 요인의 기울기가 평평해집니다. 기본값 델타 0을 바꾸려면 0.8 이하의 수를 입력합니다.
- **쿼티맥스.** 각 필드를 설명하는 데 필요한 요인 수를 최소화하는 직교 방법. 이는 관측 필드의 해석을 단순화합니다.
- **이쿼맥스.** 요인을 단순화하는 베리맥스 방법과 필드를 단순화하는 쿼티맥스 방법을 조합한 회전 방법. 요인에서 주로 로드한 필드의 수와 필드 설명에 필요한 요인 수는 최소화됩니다.
- **프로맥스.** 요인이 상관되도록 하는 사각 회전. 이 회전은 직접 오블리민 회전보다 빨리 계산될 수 있으므로 큰 데이터 세트에 유용합니다. 카파는 솔루션의 경사(요인이 상관될 수 있는 범위)를 제어합니다.

(6) PCA/요인 모델 너깃

PCA/요인 모델 너깃은 PCA/요인 노드에서 작성된 요인 분석 및 주성분분석(PRINCALS) 모델을 나타냅니다. 여기에서는 훈련 모델에서 캡처한 모든 정보와 모델 성능 및 특성에 대한 정보를 포함합니다.

요인 방정식 모델을 포함하는 스트림을 실행하는 경우 노드는 모델의 각 요인 또는 구성요소에 대한 새 필드를 추가합니다. 새 필드 이름은 모델 이름에서 파생되며 접두문자 F 와 접미문자 $-n$ 이 추가됩니다. 여기서 n 은 요인 또는 구성요소 수입니다. 예를 들어 모델 이름이 *Factor*이고 3개 요인을 포함하는 경우 새 필드 이름은, F -Factor-1, F -Factor-2, F -Factor-3과 같습니다.

요인 모델의 인코딩에 대해 더 잘 이해하려면 일부 추가 다운스트림 분석을 수행할 수 있습니다. 요인 모델 결과를 보는 유용한 방법은, 통계 노드를 사용하여 요인 및 입력 필드 사이의 상관관계를 보는 것입니다. 여기에서는 어떤 요인에 어떤 입력 필드가 과중한 부담을 주는지 표시하고 이를 통해 요인이 기본적인 의미나 해석을 보유하는지 발견하는 데 도움이 될 수 있습니다.

또한 고급 출력에서 사용 가능한 정보를 사용하여 요인 모델을 평가할 수 있습니다. 고급 출력을 보려면 모델 너깃 브라우저의 고급 탭을 클릭하십시오. 고급 출력은 많은 자세한 정보를 포함하며, 요인 분석 또는 PCA의 포괄적인 지식을 가진 사용자를 목표로 합니다. 자세한 정보는 PCA/요인 모델 너깃 고급 출력의 내용을 참조하십시오.

① PCA/요인 모델 너깃 방정식

요인 모델 너깃의 모델 탭은 각 요인의 요인 스코어 방정식을 표시합니다. 요인 또는 구성요소 스코어는 각 입력 필드 값에 해당 계수를 곱하고 결과를 합산하여 계산됩니다.

② PCA/요인 모델 너깃 요약

요인 모델의 요약 탭에서는 모델을 생성하는 데 사용되는 필드 및 설정에 대한 추가 정보와 함께, 요소/PCA 모델에 보유한 요소 수를 표시합니다. 자세한 정보는 모델 너깃 찾아보기 주제를 참조하십시오.

③ PCA/요인 모델 너깃 고급 출력

요인 분석의 고급 출력에서는 추정된 모델 및 해당 성능에 대한 자세한 정보를 제공합니다. 고급 출력에 포함된 대부분의 정보는 기술과 관련된 정보로, 이 출력을 적절히 해석하려면 요인 분석에 대한 포괄적인 지식을 요구합니다.

경고. 결과에 대한 경고 또는 잠재적 문제점을 표시합니다.

공통성. 요인 또는 구성요소에 대해 계산된 각 필드의 분산 비율을 표시합니다. 초기에서는 전체 요인 세트(모델은 입력 필드만큼 많은 요인으로 시작됨)로 초기 공통성을 제공하고 추출은 보유한 요인 세트에 기반한 공통성을 제공합니다.

설명된 총분산. 모델에서 요인별 설명된 총분산을 표시합니다. 초기 **고유값**은 초기 요인의 전체 세트에서 설명된 분산을 표시합니다. **추출 제공합 적재량**에서는 모델에 보유한 요인에서 설명된 분산을 표시합니다. **회전 제공합 적재량**에서는 회전된 요인에서 설명된 분산을 표시합니다. 사각 회전의 경우 **회전 제공합 적재량**은 제공합 적재량만 표시하고 분산 퍼센트는 표시하지 않습니다.

요인 또는 구성요소 행렬. 입력 필드 및 회전되지 않은 요인 사이의 상관관계를 표시합니다.

회전된 요인 또는 구성요소 행렬. 입력 필드 및 직교 회전의 회전된 요인 사이의 상관관계를 표시합니다.

패턴 행렬. 사각 회전의 회전 요인 및 입력 필드 사이의 편상관계수를 표시합니다.

구조행렬. 사각 회전의 회전 요인 및 입력 필드 사이의 단순 상관계수를 표시합니다.

요인 상관행렬. 사각 회전에 대한 요인 가운데 상관관계를 표시합니다.

(7) 판별 노드

판별 분석은 소속그룹에 대한 예측 모델을 작성합니다. 모델은 그룹 간에 최상의 판별을 제공하는 예측자 변수의 선형 조합을 기본으로 하는 판별 함수(그룹이 셋 이상인 경우 판별 함수 세트)로 구성됩니다. 함수는 해당 소속그룹이 알려진 케이스 표본으로부터 생성되며 해당 소속그룹은 알 수 없으나 예측자 변수 측정을 통해 새로운 케이스에 적용될 수는 있습니다.

예제. 한 통신 회사는 판별 분석을 사용하여 사용량 데이터를 기준으로 한 그룹으로 고객을 분류할 수 있습니다. 이를 통해 잠재적 고객을 스코어링하고 가장 가치 있는 그룹에 있을 가능성이 높은 고객을 목표로 할 수 있습니다.

요구사항. 하나 이상의 입력 필드 및 대상 필드가 정확히 하나 필요합니다. 목표는 문자열 또는 정수 저장 공간이 있는 범주형 필드(플래그 또는 명목 측정 수준의)여야 합니다. (필요에 따라 채움 또는 파생 노드를 사용하여 저장 공간을 변환할 수 있습니다.) **둘 다** 또는 **없음**으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다.

강도. 판별 분석과 로지스틱 회귀분석은 둘 모두 적합한 분류 모델입니다. 하지만 판별 분석은 입력 필드에 대한 더 많은 가정을 세웁니다. 예를 들어, 필드가 정상적으로 분포되고 연속형이어야 합니다. 이 요구 사항이 충족되면 특히 표본 크기가 작은 경우에 결과가 더 개선됩니다.

① 판별 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

방법. 다음 옵션을 사용하여 모델에 예측자를 입력할 수 있습니다.

- **Enter.** 이는 기본 방법으로 모든 항목을 방정식에 직접 입력합니다. 모델의 예측력을 상당히 증가시키지 않는 항목은 추가되지 않습니다.
- **단계 선택.** 이 초기 모형은 방정식에 모형 항(상수 제외)이 없는 가장 단순한 모형입니다. 각 단계마다 모델에 아직 추가되지 않은 항목을 평가하여 최상의 항목이 모델의 예측력을 상당히 증가시킬 경우 이 항목이 추가됩니다.

참고: 단계선택법은 훈련 데이터를 과적합할 경향이 높습니다. 이 방법을 사용할 때에는 검증용 검증 표본이나 새 데이터로 결과적인 모델의 유효성을 검증하는 것이 특히 중요합니다.

② 판별 노드 고급 옵션

판별 분석을 자세히 알고 있으면 고급 옵션으로 훈련 프로세스를 세부 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 **모드**를 **고급**으로 설정하십시오.

사전 확률. 이 옵션은 소속그룹의 사전 지식에 대해 분류 계수를 조정할지 여부를 결정합니다.

- **모든 그룹이 동일.** 동일한 사전 확률이 모든 그룹에 가정되며 계수에는 아무런 영향이 없습니다.
- **그룹 크기로 계산.** 표본에서 관측된 그룹 크기는 소속그룹의 사전 확률을 결정합니다. 예를 들어, 관측의 50%가 첫 번째 그룹, 25%가 두 번째 그룹, 25%가 세 번째 그룹에 속하는 분석에 포함된 경우 분류 계수는 다른 두 그룹에 상대적으로 첫 번째 그룹의 소속 가능성을 증가시키도록 조정됩니다.

공분산 행렬 사용. 이 옵션을 선택하여 그룹-내 공분산 행렬이나 개별-그룹 공분산 행렬을 사용하여 케이스를 분류할 수 있습니다.

- **그룹-내.** 그룹 내 풀링 공분산 교차표가 케이스 분류에 사용됩니다.
- **개별-그룹.** 개별-그룹 공분산 교차표가 분류에 사용됩니다. 분류가 판별 함수에 기초하고 원래 변수에 따라 달라지지 않으므로 이 옵션이 2차 판별과 항상 같지는 않습니다.

출력. 이 옵션을 사용하면 노드에서 작성한 모델 너깃의 고급 출력에 표시되는 추가 통계를 요청할 수 있습니다. 자세한 정보는 판별 노드 출력 옵션의 내용을 참조하십시오.

단계. 이 옵션은 단계 선택 추정 방법으로 필드를 추가 및 제거하는 기준을 제어할 수 있습니다. (이 단추는 입력 방법을 선택한 경우 사용할 수 없습니다.) 자세한 정보는 판별 노드 단계 옵션의 내용을 참조하십시오.

③ 판별 노드 출력 옵션

로지스틱 회귀 모델 너깃의 고급 출력에 표시하려는 선택적 출력을 선택하십시오. 고급 출력을 보려면 모델 너깃을 찾아보고 **고급** 탭을 클릭하십시오. 자세한 정보는 판별 모델 너깃 고급 출력의 내용을 참조하십시오.

기술통계. 사용할 수 있는 옵션은 평균(표준 편차 포함), 일변량분산 분석, Box의 M 검정입니다.

- **평균(Means).** 전체 평균 및 그룹 평균, 독립변수에 대한 표준 편차를 표시합니다.
- **일변량분산 분석(Univariate ANOVAs).** 각 독립변수에 대해 그룹 평균의 등식을 검정하는 일원 분산 분석을 수행합니다.
- **Box의 M .** 그룹 공분산 교차표의 등식에 대한 검정을 수행합니다. 표본이 충분히 큰 경우 p 값에 유의수준이 없으면 행렬이 다르고 판단하기 어렵습니다. 이 검정은 다변량 정규성에서 벗어나는 경우 영향을 많이 받습니다.

함수의 계수. 사용할 수 있는 옵션은 Fisher의 분류 계수 및 비표준화 계수입니다.

- **Fisher의 방법(Fisher's).** 분류에 직접 사용할 수 있는 Fisher의 분류 함수 계수를 표시합니다. 각 그룹에 대해 개별적인 일련의 분류 함수 계수가 작성되고 케이스는 판별 스코어(분류 함수 값)가 가장 큰 그룹에 할당됩니다.
- **비표준화(Unstandardized).** 표준화하지 않은 판별 함수 계수를 표시합니다.

행렬. 사용할 수 있는 독립변수에 대한 계수의 행렬은 그룹-내 상관 행렬, 그룹-내 공분산 행렬, 개별-그룹 공분산 행렬, 전체 공분산 행렬입니다.

- **그룹-내 상관행렬.** 상관을 계산하기 전에 모든 그룹에 대한 개별 공분산 교차표의 평균을 구하여 그룹 내 풀링 상관 행렬을 표시합니다.
- **그룹-내 공분산 행렬.** 그룹 내 풀링 공분산 교차표를 표시하는데 이는 전체 공분산 교차표와 다를 수 있습니다. 이 행렬은 모든 그룹에 대해 개별 공분산 교차표를 평균하여 구합니다.
- **개별-그룹 공분산 행렬.** 각 그룹에 대해 개별 공분산 교차표를 표시합니다.
- **전체 공분산.** 단일 표본으로 작성한 것처럼 모든 케이스로부터 공분산 교차표를 표시합니다.

분류. 다음은 분류 결과에 관한 출력입니다.

- **각 케이스에 대한 결과.** 각 케이스마다 실제 그룹, 예측 그룹, 사후 확률, 판별 스코어 등에 대한 코드가 표시됩니다.
- **요약표.** 판별 분석을 기준으로 각 그룹에 정확하게 할당되거나 잘못 할당된 케이스의 수로, "혼동행렬"이라고도 합니다.
- **순차제거복원 분류.** 분석의 각 케이스가 해당 케이스가 아닌 다른 모든 케이스에서 파생된 함수에 따라 분류됩니다. 이 방법을 "U-방법"이라고도 합니다.
- **영역도.** 함수 값에 따라 케이스를 그룹으로 분류하는 데 사용하는 경계의 도표입니다. 숫자는 케이스가 분류된 그룹에 해당합니다. 각 그룹의 평균은 경계 내에서 별표로 표시됩니다. 판별 함수가 하나만 있는 경우에는 맵이 표시되지 않습니다.
- **결합-그룹.** 처음 두 판별 함수 값에 대해 전체 그룹화 산점도를 작성합니다. 함수가 하나만 있는 경우에는 산점도 대신 히스토그램이 표시됩니다.
- **개별-그룹.** 처음 두 판별 함수 값의 개별 그룹 산점도를 작성합니다. 함수가 하나만 있는 경우에는 산점도 대신 히스토그램이 표시됩니다.

단계 선택. 단계 요약은 각 단계 후 모든 변수에 대한 통계를 표시합니다. **대응별 거리에 대한 F**는 각 그룹 쌍의 대응별 F 비율 교차표를 표시합니다. 그룹 간 Mahalanobis의 거리의 유의수준 검정에 F 비율을 사용할 수 있습니다.

④ 판별 노드 단계 옵션

이 옵션은 단계 선택 추정 방법으로 필드 추가에 대한 방법과 기준을 제어할 수 있습니다.

방법. 새 변수 입력 및 제거에 사용할 통계를 선택합니다. 사용 가능한 옵션은 Wilk의 람다, 설명되지 않는 분산, Mahalanobis의 거리, 최소 F -비, Rao의 V 입니다. Rao의 V 를 사용하면 입력할 변수에 대한 V 에 최소값 증가를 지정할 수 있습니다.

- **Wilks의 람다.** 단계별 판별 분석의 변수 선택 방법으로, Wilks의 람다를 낮추는 정도에 따라 방정식에 입력할 변수를 선택합니다. 각 단계에서 전체 Wilks의 람다를 최소화할 변수를 입력합니다.
- **설명되지 않는 분산.** 각 단계에서 그룹 간 설명되지 않은 변동 합계를 최소화하는 변수를 입력합니다.
- **Mahalanobis의 거리.** 독립변수의 케이스 값이 전체 케이스 평균과 얼마나 달라지는지에 대한 척도입니다. Mahalanobis 거리가 크면 케이스가 독립변수 하나 이상에 대해 극단값을 갖는 것으로 식별합니다.
- **최소 F -비.** 그룹 간 Mahalanobis 거리로부터 계산한 F -비를 최대화하는 단계별 분석의 변수 선택 방법입니다.
- **Rao의 V .** 그룹 평균 간 차이에 대한 척도입니다. Lawley-Hotelling 트레이스라고도 하며 각 단계에서 Rao의 V 의 증가를 최대화하는 변수가 입력됩니다. 이 옵션을 선택한 다음 변수가 가져야 하는 최소값을 입력하여 분석에 사용합니다.

기준. 사용 가능한 대안은 **F-값 사용**과 **F-확률 사용**입니다. 변수를 입력하고 제거하기 위한 값을 입력하십시오.

- **F-값 사용.** F 값이 진입값보다 크면 모형에 변수가 입력되고 F 값이 제거값보다 작으면 제거됩니다. 진입값은 제거값보다 커야 하고 두 값 모두 양수여야 합니다. 모형에 더 많은 변수를 입력하려면 진입값을 낮추고 모형에서 변수를 더 많이 제거하려면 제거값을 높입니다.
- **F-확률 사용.** F 값의 유의 수준이 진입값보다 작으면 모형에 변수가 입력되고 유의 수준이 제거값보다 크면 제거됩니다. 진입값은 제거값보다 작아야 하며 두 값 모두 양수여야 합니다. 모형에 변수를 더 많이 입력하려면 진입값을 높이고 모형에서 변수를 더 많이 제거하려면 제거값을 낮춥니다.

⑤ 판별 모델 너깃

판별 모델 너깃은 판별 노드가 추정한 방정식을 나타냅니다. 여기에는 판별 모델에서 캡처한 모든 정보와 모델 구조 및 성능에 대한 정보가 포함되어 있습니다.

판별 모델 너깃을 포함한 스트림을 실행하면 노드는 모델의 예측 및 연관된 확률을 포함한 두 개의 새 필드를 추가합니다. 새 필드의 이름은 예측 중인 출력 필드의 이름에서 파생되며 예측 범주의 경우 *\$D-* 및 연관된 확률의 경우에는 *\$DP-* 접두문자가 붙습니다. 예를 들어, 출력 필드 *colorpref*의 경우 새 필드의 이름은 *\$D-colorpref* 및 *\$DP-colorpref*입니다.

필터 노드 생성. 생성 메뉴로 모델의 결과를 기준으로 하여 입력 필드를 전달할 새 필터 노드를 작성할 수 있습니다.

예측자 중요도(Predictor Importance)

선택적으로 모델을 추정할 때 각 예측자의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측자에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측자 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

가. 판별 모델 너깃 고급 출력

판별 분석의 고급 출력은 추정 모델 및 성능에 관한 자세한 정보를 제공합니다. 고급 출력에 포함된 대부분의 정보는 상당히 기술적 정보이며 이 출력을 제대로 해석하려면 광범위한 판별 분석 지식이 필요합니다. 자세한 정보는 판별 노드 출력 옵션의 내용을 참조하십시오.

나. 판별 모델 너깃 설정

판별 모델 너깃의 설정 탭으로 모델을 스코어링할 때 성향 스코어를 확보할 수 있습니다. 이 탭은 플래그 목표가 있는 모델에, 스트림에 모델 너깃이 추가된 후에만 사용 가능합니다.

원시 성향 스코어 계산. (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 이외에도 스코어링 중에 생성할 수 있는 기타 예측 및 신뢰도 값이 있습니다.

수정된 성향 스코어 계산. 원시 성향 스코어는 훈련 데이터에만 기반하며, 이 데이터의 과적합을 위한 많은 모델의 경향으로 인해 지나치게 낙관적일 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 성능을 평가하여 보완하려고 합니다. 이 옵션에서는 파티션 필드가 스트림에 정의되어야 하고 모델 생성 전에 모델링 노드에서 수정된 성향 스코어가 사용 가능해야 합니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

SQL 생성이 수행되는 방법을 지정하려면 다음 옵션 중 하나를 선택하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

다. 판별 모델 너깃 요약

판별 모델 너깃의 요약 탭은 모델을 생성하는 데 사용하는 필드 및 설정을 표시합니다. 또한 이 모델링 노드에 첨부된 분석 노드를 실행하면 해당 분석의 정보도 이 섹션에 표시됩니다. 모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

(8) GenLin 노드

일반화 선형 모델은 종속변수가 지정된 연결함수를 통해 요인 및 공변량과 선형적으로 관련되도록 일반 선형 모델을 확장합니다. 더욱이 모델을 사용하면 종속변수가 비정규 분포를 가질 수 있습니다. 또한 정상적으로 분포된 반응, 이분형 데이터의 로지스틱 모델, 계수 데이터의 선형로그 모델, 구간 중도절단 생존 데이터에 대한 보 로그-로그 모델은 물론 매우 일반적인 모델 공식을 통해 다른 많은 통계 모델 같이 널리 사용되는 통계 모델을 포함합니다.

예. 운송 회사에서는 일반화 선형 모델을 사용하여 서로 다른 기간에 구성된 선박의 여러 유형에 대한 손상 횟수에 포아송 회귀분석을 맞출 수 있습니다. 그리고 결과로 생성된 모델은 손상될 확률이 높은 선박 유형을 판별하는 데 도움이 될 수 있습니다.

자동차 보험 회사는 일반화 선형 모델을 사용하여 자동차의 손해 배상 청구에 감마회귀를 맞출 수 있습니다. 결과로 생성되는 모델은 청구 규모에 가장 많이 기여하는 요인을 판별하는 데 도움을 줄 수 있습니다.

의료 연구자는 일반화 선형 모델을 사용하여 구간별 검열된 생존 데이터에 보 로그-로그 회귀분석을 맞추어 의료 조건에 대한 재발 시간을 예측할 수 있습니다.

일반화 선형 모델은 입력 필드 값을 출력 필드 값에 연관시키는 방정식을 작성하여 작동됩니다. 모델이 생성되면 새 데이터의 값을 추정하는 데 사용할 수 있습니다. 각 레코드의 경우 가능한 각 출력 범주에 대해 소속 확률이 계산됩니다. 확률이 가장 높은 목표 범주는 해당 레코드의 예측된 출력 값으로 지정됩니다.

요구사항. 하나 이상의 입력 필드와 둘 이상의 범주를 포함하는 정확히 하나의 대상 필드(측정 수준이 연속형 또는 플래그일 수 있음)가 필요합니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다.

강도. 일반화 선형 모델은 매우 탄력적이지만, 모델 구조를 선택하는 프로세스가 자동화되어 있지 않고, "블랙박스" 알고리즘에 필요하지 않은 데이터와 어느 정도 친숙해야 함을 요구합니다.

① GenLin 노드 필드 옵션

일반적으로 모델링 노드 필드 탭에서 제공되는 목표, 입력, 파티션 사용자 정의 옵션 외에도(모델링 노드 필드 옵션 참조) GenLin 노드는 다음과 같은 추가 기능을 제공합니다.

가중 필드 사용. 척도 모수는 반응의 변수와 관련한 추정된 모델 모수입니다. 척도 가중값은 "알려진" 값으로, 관측할 때마다 달라질 수 있습니다. 척도 가중 변수를 지정한 경우 반응의 변수와 관련한 척도 모수는 각 관측에 대해 나눈 것입니다. 0보다 작거나 같고 또는 값이 없는 척도 가중값을 가진 레코드는 분석에 사용되지 않습니다.

시행 세트에서 발생하는 이벤트 수를 나타내는 목표 필드. 반응이 시행 세트에서 발생하는 많은 이벤트면 목표 필드에는 이벤트 수가 포함되며 시행 수가 포함되어 있는 추가 변수를 선택할 수 있습니다. 또는 시행 수가 모든 개체에서 동일한 경우 고정 값을 사용하여 시행을 지정할 수 있습니다. 각 레코드에 대해 시행 수는 이벤트 수보다 크거나 같아야 합니다. 이벤트는 음이 아닌 정수가 되어야 하며 시행 수는 양의 정수가 되어야 합니다.

② GenLin 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

모델 유형. 작성할 모델 유형에 대해 두 가지 옵션이 있습니다. **주효과만**을 사용하면 모델이 입력 필드만 개별적으로 포함하고, 입력 필드 사이의 상호작용(승법 효과)을 검정하지 않습니다. **주효과 및 모든 이원 상호작용**은 모든 이원 상호작용과 입력 필드 주효과를 포함합니다.

오프셋. 오프셋 항은 "구조" 예측자입니다. 계수는 모델로 추정되지 않지만 값 1을 갖는 것으로 가정합니다. 따라서 오프셋 값은 단순히 목표의 선형 예측자에 추가됩니다. 이는 각 케이스가 중요 이벤트마다 다른 노출 수준을 가질 수 있는 포아송 회귀 모형에 특히 유용합니다.

예를 들어, 개별 운전자의 사고 비율을 모델링할 때 3년 경력 중 한 번의 사고를 낸 운전자와 25년 경력 중에 한 번 사고를 낸 운전자 사이에는 중요한 차이가 있습니다! 사고 수는 운전 경력의 자연 로그가 오프셋 항으로 포함되는 경우 로그 링크와 함께 포아송 또는 음이항 반응으로 모델링할 수 있습니다.

다른 분포 및 링크 유형의 조합에는 오프셋 변수의 다른 변환이 필요할 수 있습니다.

참고: 변수 범위 필드를 사용하는 경우 지정된 필드는 입력으로 사용할 수 없습니다. 업스트림 소스 또는 필요한 경우 유형 노드에서 범위 필드의 역할을 **없음**으로 설정하십시오.

플래그 목표의 기본 범주

이분형 반응의 경우 종속변수에 대한 참조범주를 선택할 수 있습니다. 이는 모수 추정값 및 저장된 값과 같은 특정 출력에 영향을 미칠 수 있지만 모형 적합을 변경해서는 안 됩니다. 예를 들어, 이분형 반응이 0과 1의 값을 사용하는 경우 다음과 같습니다.

- 기본적으로 프로시저는 마지막(가장 높은 값) 범주 또는 참조 범주인 1을 만듭니다. 이 상황에서, 모델 저장 확률은 주어진 케이스가 값 0을 사용하는 변화를 추정하고, 모수 추정값은 범주 0의 우도와 관련하여 해석해야 합니다.
- 첫 번째(가장 낮은 값) 범주 또는 참조 범주로 0을 지정하는 경우 모델 저장 확률은 주어진 케이스가 값 1을 사용하는 변화를 추정합니다.
- 사용자 정의 범주를 지정하고 변수에 정의된 레이블이 있는 경우 목록에서 값을 선택하여 참조 범주를 설정할 수 있습니다. 이는 모델을 지정하는 동안 정확히 특정 변수를 어떻게 코딩했는지 기억이 나지 않을 때 편리할 수 있습니다.

모델에 절편 포함. 절편은 보통 모델에 포함됩니다. 데이터가 선형 회귀로 전달된다고 가정할 경우에는 절편을 제외할 수 있습니다.

③ GenLin 노드 고급 옵션

일반화 선형 모델에 대한 자세한 지식이 있는 경우 고급 옵션을 통해 학습 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 **모드**를 **고급**으로 설정하십시오.

대상 필드 분포 및 링크 함수

분포

이 선택은 종속변수의 분포를 지정합니다. 비정규 분포와 항등하지 않은 연결 함수를 지정하는 기능은 일반 선형 모델에서 일반화 선형 모델의 중요한 개선 사항입니다. 많은 분포-연결 함수 조합이 있으며 주어진 데이터 세트에 적합한 함수가 여러 개일 수 있으므로 사전 이론적 고려 사항을 바탕으로 선택하거나 어느 조합이 가장 적합할지를 고려하여 선택하면 됩니다.

- **이항.** 이 분포는 이분형 반응이나 이벤트 수를 나타내는 변수의 경우에만 적합합니다.
- **감마.** 이 분포는 더 큰 양의 값 쪽으로 비대칭되는 양의 척도 값을 가진 변수에 적합합니다. 데이터 값이 0보다 작거나 같고 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다.
- **역가우스.** 이 분포는 더 큰 양의 값 쪽으로 비대칭되는 양의 척도 값을 가진 변수에 적합합니다. 데이터 값이 0보다 작거나 같고 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다.
- **음이항.** 이 분포는 k 성공을 관측하는 데 필요한 시행 횟수로 생각할 수 있으며 양의 정수 값을 갖는 변수에 적합합니다. 데이터 값이 양수이거나, 0보다 작거나 없는 경우 해당 케이스는 분석에 사용되지 않습니다. 음이항 분포 보조 모수의 고정 값은 0 이상의 숫자가 될 수 있습니다. 보조 매개변수가 0으로 설정되면 분포를 사용하는 것은 포아송 분포를 사용하는 것과 동일합니다.
- **정규.** 중앙(평균) 값에 대해 값이 대칭되는 종 형태의 분포를 띠는 척도변수에 적합합니다. 종속변수는 숫자여야 합니다.
- **포아송.** 이 분포는 고정 기간 동안 중요 이벤트의 발생 수로 생각할 수 있으며 양의 정수 값을 갖는 변수에 적합합니다. 데이터 값이 양수이거나, 0보다 작거나 없는 경우 해당 케이스는 분석에 사용되지 않습니다.
- **Tweedie.** 이 분포는 감마 분포의 포아송 혼합으로 표현할 수 있는 변수에 적합합니다. 분포는 연속 특성(음이 아닌 실수 값 사용)과 이산형 분포(단일 값 0에서 양의 확률 매스)의 관점에서 "혼합"된 것입니다. 종속변수는 데이터 값이 0보다 크거나 같은 숫자가 되어야 합니다. 데이터 값이 0보다 작거나 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다. Tweedie 분포에서 모수의 고정 값은 1보다 크고 2보다 작은 숫자가 될 수 있습니다.
- **다항.** 이 분포는 순서 반응을 나타내는 변수에 적합합니다. 종속변수는 숫자나 문자열이 될 수 있으며 최소 두 개의 유효한 개별 데이터 값을 가져야 합니다.

연결함수

연결 함수는 종속변수의 변환으로 모델을 추정할 수 있습니다. 다음 함수를 사용할 수 있습니다.

- **항등.** $f(x)=x$. 종속변수가 변환되지 않습니다. 이 링크는 분포에 사용할 수 있습니다.
- **보 로그-로그.** $f(x)=\log(-\log(1-x))$. 이항 분포에만 적합합니다.
- **누적 Cauchit.** $f(x) = \tan(\pi (x - 0.5))$, 각 응답 범주의 누적 확률에 적용됩니다. 다항 분포에만 적합합니다.
- **누적 보 로그-로그.** $f(x)=\ln(-\ln(1-x))$, 각 응답 범주의 누적 확률에 적용됩니다. 다항 분포에만 적합합니다.
- **누적 로짓.** $f(x)=\ln(x / (1-x))$, 각 응답 범주의 누적 확률에 적용됩니다. 다항 분포에만 적합합니다.
- **누적 음 로그-로그.** $f(x)=-\ln(-\ln(x))$, 각 응답 범주의 누적 확률에 적용됩니다. 다항 분포에만 적합합니다.
- **누적 프로빗.** $f(x)=\Phi^{-1}(x)$, 각 응답 범주의 누적 확률에 적용됩니다. 여기서 Φ^{-1} 은 역 표준 정규 누적 분포 함수입니다. 다항 분포에만 적합합니다.
- **로그.** $f(x)=\log(x)$. 이 링크는 분포에 사용할 수 있습니다.
- **로그 보.** $f(x)=\log(1-x)$. 이항 분포에만 적합합니다.
- **로짓.** $f(x)=\log(x / (1-x))$. 이항 분포에만 적합합니다.
- **음이항.** $f(x)=\log(x / (x+k^{-1}))$, 여기서 k 는 음이항 분포의 보조 모수입니다. 음이항 분포에만 적합합니다.
- **음 로그-로그.** $f(x)=-\log(-\log(x))$. 이항 분포에만 적합합니다.
- **오즈 거듭제곱.** $f(x)=[(x/(1-x))^\alpha - 1]/\alpha$, if $\alpha \neq 0$. $f(x)=\log(x)$ ($\alpha=0$ 의 경우). α 는 필수 숫자 지정 사항이며 실수여야 합니다. 이항 분포에만 적합합니다.
- **프로빗.** $f(x)=\Phi^{-1}(x)$. 여기서 Φ^{-1} 은 역표준 정규 누적 분포 함수입니다. 이항 분포에만 적합합니다.
- **거듭제곱.** $f(x)=x^\alpha$ ($\alpha \neq 0$ 의 경우). $f(x)=\log(x)$ ($\alpha=0$ 의 경우). α 는 필수 숫자 지정 사항이며 실수여야 합니다. 이 링크는 분포에 사용할 수 있습니다.

모수. 이 그룹의 제어를 사용하면 특정 분포 옵션을 선택한 경우 모수 값을 지정할 수 있습니다.

- **음이항에 대한 모수.** 음이항 분포의 경우 값을 지정하거나 시스템에서 추정된 값을 제공하도록 허용합니다.
- **Tweedie에 대한 모수.** Tweedie 분포의 경우 고정값으로 1.0과 2.0 사이의 숫자를 지정합니다. **모수 추정값.** 이 그룹의 제어를 사용하면 추정 방법을 지정하고 모수 추정값에 대한 초기값을 제공할 수 있습니다.
 - **방법.** 모수 추정 방법을 선택할 수 있습니다. Newton-Raphson, Fisher 스코어링 또는 Fisher 스코어링 반복이 Newton-Raphson 방법으로 전환하기 전에 수행되는 하이브리드 방법 중에서 선택합니다. 하이브리드 방법의 Fisher 스코어링 단계 동안 Fisher 반복의 최대 수에 도달하기 전에 수렴이 얻어진 경우 알고리즘은 Newton-Raphson 방법으로 계속 됩니다.

- **척도 모수 방법.** 척도 모수 추정 방법을 선택할 수 있습니다. 최대우도는 모델 효과가 있는 척도 모수를 공동으로 추정합니다. 이 옵션은 응답에 음이항, 포아송 또는 이항 분포인 경우 올바르지 않습니다. 편차 및 Pearson 카이제곱 옵션은 해당 통계값에서 척도 모수를 추정합니다. 또한 척도 모수에 대한 고정 값을 지정할 수 있습니다.
- **공분산 교차표.** 모델 기반 추정량은 Hessian 행렬의 일반화 역의 음수입니다. 동질성 (Huber/White/sandwich라고도 함) 추정량은 변수와 연결함수의 지정이 잘못된 경우에도 공분산의 일관성 있는 추정을 제공하는 "수정된" 모델 기반 추정량입니다.

반복. 이 옵션으로 모델 수렴에 대한 모수를 제어할 수 있습니다. 자세한 정보는 일반화 선형 모델 반복의 내용을 참조하십시오.

출력. 이 옵션을 사용하면 노트에서 작성한 모델 너깃의 고급 출력에 표시되는 추가 통계를 요청할 수 있습니다. 자세한 정보는 일반화 선형 모델 고급 출력의 내용을 참조하십시오.

비정칙성 공차. 비정칙(또는 비가역) 행렬에는 추정 알고리즘에 심각한 문제를 일으킬 수 있는 선형 종속 열이 있습니다. 거의 비정칙인 행렬은 잘못된 결과를 초래할 수 있으므로 프로시저는 행렬식이 공차보다 작은 교차표는 비정칙으로 취급합니다. 양수값을 지정합니다.

④ 일반화 선형 모델 반복

일반화 선형 모델 추정에 대한 수렴 모수를 설정할 수 있습니다.

반복. 다음 옵션을 사용할 수 있습니다.

- **최대반복계산.** 알고리즘에서 실행할 최대 반복 횟수입니다. 음수가 아닌 정수를 지정합니다.
- **최대 단계 이분.** 각 반복에서 단계 크기는 로그 우도 증가 또는 최대 단계 이분에 도달할 때까지 요인이 0.5씩 감소됩니다. 양수를 지정하십시오.
- **데이터 포인트의 분리 확인.** 이 옵션을 선택하면 알고리즘이 모수 추정값이 중복되지 않았는지 확인하기 위한 검정이 수행됩니다. 모든 케이스를 올바르게 분류하는 모델을 프로시저에서 생성할 수 있는 경우에 분리가 발생합니다. 이 옵션은 2진 형식의 2항 응답에 사용 가능합니다.

수렴 기준. 다음 옵션을 사용할 수 있습니다.

- **모수 추정값 변화량.** 이 옵션을 선택하면 모수 추정값의 절대 변화량 또는 상대 변화량이 지정된 값보다 작아지는 반복 후에 알고리즘이 멈춥니다. 지정된 값은 양수여야 합니다.
- **로그-우도 변화.** 이 옵션을 선택하면 로그-우도 함수의 절대 변화량 또는 상대 변화량이 지정된 값보다 작아지는 반복 후에 알고리즘이 멈춥니다. 지정된 값은 양수여야 합니다.
- **Hessian 수렴.** 절대값 지정의 경우 수렴은 Hessian 수렴을 기준으로 하는 통계가 지정된 양의 값보다 작다고 가정합니다. 상대값 지정의 경우 수렴은 통계가 로그 우도의 절대값과 지정된 양의 값의 곱보다 작다고 가정합니다.

⑤ 일반화 선형 모델 고급 출력

일반화 선형 모델 너깃의 고급 출력에 표시할 선택적 출력을 선택하십시오. 고급 출력을 보려면 모델 너깃을 찾아보고 **고급** 탭을 클릭하십시오. 자세한 정보는 GenLin 모델 너깃 고급 출력의 내용을 참조하십시오.

다음 출력을 사용할 수 있습니다.

- **케이스 처리 요약.** 분석에 포함되었거나 제외된 케이스 수와 백분율 및 상관 데이터 요약 테이블을 표시합니다.
- **기술통계.** 종속변수, 공변량 및 요인에 대한 기술 통계와 요약 정보를 표시합니다.
- **모델 정보.** 데이터 세트 이름, 종속변수 또는 이벤트 및 시행 변수, 오프셋 변수, 척도가중 변수, 확률 분포 및 연결 함수를 표시합니다.
- **적합도 통계량.** 편차와 척도화된 편차, Pearson 카이제곱 및 척도화된 Pearson 카이제곱, 로그 우도, Akaike 정보기준(AIC), 유한 표본 수정된 AIC(AICC), 베이저안 정보 기준(BIC) 및 일관된 AIC(CAIC)를 표시합니다.
- **모델 요약 통계.** 각 효과에 대한 제 I 유형 또는 제 III 유형 대비에 대한 통계 및 모형 적합 총괄 검정에 대한 우도비 통계를 포함한 모형 적합 검정을 표시합니다.
- **모수 추정값.** 모수 추정값과 해당 검정 통계량 및 신뢰구간을 표시합니다. 원래 모수 추정값 외에 선택적으로 누승 매개변수 추정을 표시할 수 있습니다.
- **공분산 교차표 기준 모수 추정값.** 추정된 모수 공분산 행렬이 표시됩니다.
- **모수 추정값에 대한 상관행렬.** 추정된 모수 상관 행렬이 표시됩니다.
- **대비 계수(L) 행렬.** EM 평균 탭에서 요청하는 경우 기본 효과 및 주변 평균 추정에 대한 대비계수를 표시합니다.
- **일반 추정가능 함수.** 대비계수(L) 행렬을 생성하는 지표를 표시합니다.
- **반복계산과정.** 모수 추정값과 로그 우도에 대한 반복계산과정을 표시하고 기울기 벡터와 Hessian 행렬의 마지막 평가를 인쇄합니다. 반복계산과정 테이블은 0번째 반복(초기 추정값)부터 시작하여 각 n 번째 반복마다 모수 추정값을 표시합니다. 여기서 n 은 인쇄 구간의 값입니다. 반복계산과정을 요청하는 경우 n 에 관계 없이 마지막 반복은 항상 표시됩니다.
- **LM 검정.** 명목, 감마, 역가우스 분포에서 고정된 숫자로 설정되었거나 편차 또는 Pearson 카이제곱을 사용하여 계산된 척도 모수의 유효성을 평가하기 위해 LM 검정 통계를 표시합니다. 음이항 분포의 경우 고정된 보조 모수를 검정합니다.

모델 효과. 다음 옵션을 사용할 수 있습니다.

- **분석 유형.** 생성할 분석 유형을 지정합니다. 제 I 유형 분석은 일반적으로 모델에서 순서 예측 변수에 대한 사전 이유가 있을 때 적합한 반면 제 III 유형은 보다 일반적으로 적용됩니다. Wald 또는 우도비 통계는 카이제곱 통계량 그룹에서 선택한 것을 기준으로 계산됩니다.
- **신뢰구간 수준(%).** 50보다 크고 100보다 작은 신뢰수준을 지정하십시오. Wald 구간은 매개 변수에 근사 정규 분포가 있다는 가정을 기반으로 합니다. 프로파일 우도 구간은 더 정확하지

만 계산 비용이 들 수 있습니다. 프로파일 우도 구간의 공차 수준은 구간을 계산하는 데 사용되는 반복 알고리즘을 중지하는 데 사용되는 기준입니다.

- **로그-우도 함수.** 이 함수는 로그-우도 함수의 표시 형식을 제어합니다. 전체 함수에는 매개변수 추정에 관해 일관성 있는 추가 항이 포함되어 있습니다. 매개변수 추정에는 효과가 없으며 일부 소프트웨어 제품에서는 출력이 되지 않습니다.

⑥ GenLin 모델 너깃

GenLin 모델 너깃은 GenLin 노드에서 추정된 방정식을 나타냅니다. 여기에는 모델에서 캡처한 모든 정보와 모델 구조 및 성능에 대한 정보가 포함되어 있습니다.

GenLin 모델 너깃을 포함하는 스트림을 실행할 때 노드는 해당 콘텐츠가 목표 필드의 특징에 종속된 새 필드를 추가합니다.

- **플래그 목표.** 예측 범주와 연관된 확률 및 각 범주의 확률을 포함하는 필드를 추가합니다. 처음 2개의 새 필드 이름은 예측할 출력 필드 이름에서 파생되며, 예측 범주의 경우 접두문자는 $\$G-$, 연관된 확률의 경우 $\$GP-$ 입니다. 예를 들어, 이름이 *default*인 출력 필드의 경우 새 필드 이름은 $\$G-default$ 및 $\$GP-default$ 입니다. 마지막 2개 추가 필드 이름은 출력 필드 값에 따라 지정되며, 접두문자는 $\$GP-$ 입니다. 예를 들어 기본값의 적합한 값이 *Yes(예)* 및 *No(아니오)*인 경우 새 필드 이름은 $\$GP-Yes$ 및 $\$GP-No$ 입니다.
- **연속형 목표.** 예측 평균 및 표준 오차를 포함하는 필드를 추가합니다.
- **연속형 목표(일련의 시행에서 이벤트 수 표시).** 예측 평균 및 표준 오차를 포함하는 필드를 추가합니다.
- **순서 목표.** 정렬된 세트의 각 값에 대한 예측 범주 및 연관된 확률을 포함하는 필드를 추가합니다. 필드 이름은 예측하는 정렬된 세트 값에서 파생되며, 예측 범주의 경우 접두문자는 $\$G-$, 연관된 확률의 경우 $\$GP-$ 입니다.

필터 노드 생성. 생성 메뉴로 모델의 결과를 기준으로 하여 입력 필드를 전달할 새 필터 노드를 작성할 수 있습니다.

예측자 중요도(Predictor Importance)

선택적으로 모델을 추정할 때 각 예측자의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측자에 모델링 노력을 집중하고 가장 쓸모 없는 예측자를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측자 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 예측변수 중요도의 내용을 참조하십시오.

가. GenLin 모델 너깃 고급 출력

일반화 선형 모델의 고급 출력은 추정된 모델 및 해당 성능에 대한 자세한 정보를 제공합니다. 고급 출력에 포함된 대부분의 정보는 기술과 관련된 정보로, 이 출력을 적절히 해석하려면 이 분석 유형에 대한 포괄적인 지식을 요구합니다. 자세한 정보는 일반화 선형 모델 고급 출력의 내용을 참조하십시오.

나. GenLin 모델 너깃 설정

GenLin 모델 너깃의 설정 탭에서는 모델 스코어링 시, 그리고 모델 스코어링 중에 SQL 생성을 위해 성향 스코어를 확보할 수 있습니다. 이 탭은 플래그 목표가 있는 모델에, 스트림에 모델 너깃이 추가된 후에만 사용 가능합니다.

원시 성향 스코어 계산. (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 이외에도 스코어링 중에 생성할 수 있는 기타 예측 및 신뢰도 값이 있습니다.

수정된 성향 스코어 계산. 원시 성향 스코어는 훈련 데이터에만 기반하며, 이 데이터의 과적합을 위한 많은 모델의 경향으로 인해 지나치게 낙관적일 수 있습니다. 조정된 성향은 검정 또는 검증 파티션에서 모델 성능을 평가하여 보완하려고 합니다. 이 옵션에서는 파티션 필드가 스트림에 정의되어야 하고 모델 생성 전에 모델링 노드에서 수정된 성향 스코어가 사용 가능해야 합니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조건의 성능을 개선할 수 있습니다.

SQL 생성이 수행되는 방법을 지정하려면 다음 옵션 중 하나를 선택하십시오.

- **기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링** 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어 선택된 경우,** 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

다. GenLin 모델 너깃 요약

GenLin 모델 너깃의 요약 탭에서는 모델을 생성하는 데 사용된 필드 및 설정을 표시합니다. 또한 이 모델링 노드에 첨부된 분석 노드를 실행하면 해당 분석의 정보도 이 섹션에 표시됩니다. 모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

(9) 일반화 선형 혼합 모델

① GLMM 노트

이 노트를 사용하여 일반화 선형 혼합 모델(GLMM)을 작성합니다.

가. 일반화 선형 혼합 모델

일반화 선형 혼합 모델은 선형 모델을 확장하여

- 목표가 지정된 연결 함수를 통해 요인 및 공변량과 선형적으로 관련되도록 합니다.
- 목표는 비정규 분포를 가질 수 있습니다.
- 관측값은 상호 관련될 수 있습니다.

일반화 선형 혼합 모델은 단순 선형 회귀에서 비정규 장기적인 데이터에 대한 복합 다중 수준 모델에 이르기까지 다양한 모델을 포함합니다.

예. 교육청은 일반화 선형 혼합 모델을 사용하여 실험적인 교수법이 수학 스코어 향상에 효과적인지 아닌지를 알 수 있습니다. 동일한 교실의 학생은 동일한 교사가 가르치므로 상호 관련되어야 하고, 동일한 학교 안에 있는 교실 역시 상관될 수 있으므로, 변동의 다양한 소스를 고려하기 위해 학교 및 클래스 수준에서 랜덤 효과를 포함할 수 있습니다.

의료 연구자들은 항경련제가 환자의 발작률을 줄일 수 있는지의 여부를 알아보기 위해 일반화 선형 혼합 모델을 사용할 수 있습니다. 같은 환자에게서 얻은 반복 측정은 대체로 양의 상관을 가지므로 일부 랜덤 효과가 있는 혼합 모델이 적합합니다. 대상 필드(발작 수)는 양수 값을 취하므로, 포아송 분포와 로그 링크가 있는 일반화 선형 혼합 모델이 적절할 수 있습니다.

TV, 전화 및 인터넷 서비스의 케이블 제공업체 대표는 일반화 선형 혼합 모델을 사용하여 잠재 고객에 대해 더 잘 알 수 있습니다. 가능한 응답이 명목 측정 수준이므로, 회사 분석가는 제공된 설문조사 응답자 응답 내에서 서비스 유형(텔레비전, 전화기, 인터넷) 사이의 서비스 사용 질문에 대한 응답 사이의 상관관계를 캡처하기 위해 변량 절편의 일반화 로짓 혼합 모델을 사용합니다.

데이터 구조 탭에서는 관측값들을 연결할 때 데이터 세트를 구성하는 레코드 간의 구조적 관계를 지정할 수 있습니다. 데이터 세트의 레코드가 독립된 관측값인 경우 이 탭에서 아무 것도 지정할 필요가 없습니다.

개체. 지정된 범주형 필드 값의 조합은 데이터 세트 내의 개체를 고유하게 정의해야 합니다. 예를 들어, 단일 환자 ID 필드는 단일 병원의 개체를 정의하기에 충분해야 하지만, 환자 식별 번호가 병원마다 고유하지 않은 경우 병원 ID와 환자 ID의 조합이 필요할 수 있습니다. 반복 측정 설정에서 각 개체마다 여러 관측을 기록하므로 각 개체는 데이터 세트에서 여러 레코드를 차지할 수 있습니다.

개체는 다른 개체에 대해 독립적인 것으로 간주할 수 있는 관측 단위입니다. 예를 들어, 의학 연구에서 한 환자의 혈압 기록은 다른 환자의 기록에 대해 독립적인 것으로 간주할 수 있습니다. 개체마다 반복 측정값이 있고 이러한 관측값 간의 상관을 모델링하려는 경우 개체를 정의하는 것이 매우 중요합니다. 예를 들어, 담당 의사에게 지속적으로 진찰을 받는 한 환자의 혈압 기록은 상호 관련된 것으로 예상할 수 있습니다.

데이터 구조 탭에서 **개체**로 지정된 모든 필드는 잔차 공분산 구조에 대한 개체를 정의하는 데 사용되며, 변량효과 블록에서 변량효과 공분산 구조에 대한 개체를 정의하는 데 가능한 필드 목록을 제공합니다.

반복측도. 여기에 지정된 필드는 반복 관측값을 식별하는 데 사용됩니다. 예를 들어, 단일 변수 주는 의학 연구에서의 10주 동안의 관측을 식별하는 데 사용하거나 월 및 일은 1년 동안의 일별 관측을 식별하는 데 함께 사용할 수 있습니다.

공분산 그룹 정의 기준. 여기에서 지정된 범주형 필드는 반복 효과 공분산 모수의 독립 세트를 정의합니다(그룹 필드의 교차 분류에 의해 정의된 각 범주에 대해 하나씩). 모든 개체의 공분산 유형은 동일하며, 동일한 공분산 그룹 내의 개체는 모수에 대해 같은 값을 가집니다.

공간 공분산 좌표. 이 목록의 변수는 반복된 공분산 유형에 공간 공분산 유형 중 하나가 선택된 경우 반복되는 관찰의 좌표를 지정합니다.

반복 공분산 유형. 잔차에 대한 공분산 구조를 지정합니다. 사용 가능한 구조는 다음과 같습니다.

- 1차 자기회귀(AR1)
- 자기회귀 이동 평균(1,1)(ARMA11)
- 복합 대칭
- 대각선
- 척도화 항등
- 공간: 거듭제곱
- 공간: 지수
- 공간: 가우스
- 공간: 선형
- 공간: 선형-로그
- 공간: 원형
- Toeplitz
- 비구조적
- 분산성분

ㄱ. 목표 (일반화 선형 혼합 모델)

이 설정은 목표, 분포 및 연결함수를 통한 예측변수에 대한 관계를 정의합니다.

목표. 목표는 필수입니다. 어떤 측정 수준도 가질 수 있으며, 목표의 측정 수준은 적합한 분포 및 연결함수를 제한합니다.

- **시행 수를 분모로 사용.** 목표 반응이 시행 세트에서 발생하는 많은 이벤트면 대상 필드에는 이벤트 수가 포함되며 시행 수가 포함되어 있는 추가 필드를 선택할 수 있습니다. 예를 들어, 새 살충제를 실험할 때 개미 표본에 다른 농도의 살충제를 사용하고 죽은 개미 수와 각 표본의 개미 수를 기록할 수 있습니다. 이 경우 죽은 개미 수를 기록한 필드를 목표(이벤트) 필드로 지정하고, 각 표본의 개미 수를 기록한 필드를 시행 필드로 지정해야 합니다. 각 표본의 개미 수가 동일한 경우 시행 수를 고정 값으로 지정할 수 있습니다. 각 레코드에 대해 시행 수는 이벤트 수보다 크거나 같아야 합니다. 이벤트는 음이 아닌 정수가 되어야 하며 시행 수는 양의 정수가 되어야 합니다.

- **참조 범주 사용자 정의.** 범주형 목표의 경우, 참조 범주를 선택할 수 있습니다. 이것은 모수 추정값과 같은 특정 결과에 영향을 미칠 수 있지만 모델 적합을 변경해서는 안 됩니다. 예를 들어 목표가 0, 1, 2 값을 가지면, 기본적으로 프로시저는 마지막(가장 높은 값) 범주 또는 참조 범주인 2를 만듭니다. 이 상황에서, 모수 추정값은 범주 2의 우도에 상대적으로 범주 0 또는 1의 우도와 관련하여 해석해야 합니다. 사용자 정의 범주를 지정하는데 목표에 레이블이 정의된 경우, 목록에서 값을 선택하여 참조 범주를 설정할 수 있습니다. 이는 모델을 지정하는 동안 정확히 특정 필드를 어떻게 코딩했는지 기억이 나지 않을 때 편리할 수 있습니다.

목표 분포 및 선형 모델과의 관계(링크). 예측변수의 값을 고려할 때, 모델은 지정된 형태를 따르기 위해 목표 값의 분포를 예상하고, 목표 값의 경우 지정된 연결함수를 통해 예측변수와 선형적으로 관련됩니다. 여러 공통 모델에 대한 바로 가기가 제공되거나, 최종 목록에 없는 적합하도록 할 특정 분포 및 연결함수 조합이 있는 경우에는 **사용자 정의**를 선택하십시오.

- **선형 모델.** 정규 분포를 항등 링크와 함께 지정합니다. 이는 목표가 선형 회귀 또는 ANOVA 모델을 사용하여 예측될 수 있을 때 유용합니다.
- **감마회귀분석.** 감마 분포를 로그 링크와 함께 지정합니다. 이는 목표에 모든 양수값이 포함되고 더 큰 값 쪽으로 비대칭될 때 사용되어야 합니다.
- **로그선형분석.** 포아송 분포를 로그 링크와 함께 지정합니다. 이는 목표가 고정 기간 동안 발생 개수를 나타낼 때 사용되어야 합니다.
- **음이항회귀분석.** 음수 이항 분포를 로그 링크와 함께 지정합니다. 이는 목표와 분모가 k 성공을 관측하는 데 필요한 시행 수를 나타낼 때 사용되어야 합니다.
- **다항 로지스틱 회귀분석.** 다항 분포를 지정합니다. 이는 목표가 다범주 반응일 때 사용되어야 합니다. 누적 로짓 링크(순서 결과)나 일반화된 로짓 링크(다범주 명목 반응)를 사용합니다.
- **이분형 로지스틱 회귀분석.** 이항 분포를 로짓 링크와 함께 지정합니다. 이는 목표가 로지스틱 회귀분석 모델에 의해 예측된 이분형 반응일 때 사용되어야 합니다.

- **이분형 프로빗**. 이항 분포를 프로빗 링크와 함께 지정합니다. 이는 목표가 기본 정규 분포가 있는 이분형 반응일 때 사용되어야 합니다.
- **구간 중도절단 생존**. 이항 분포를 보 로그-로그 링크와 함께 지정합니다. 이는 몇몇 관측값에 종료 이벤트가 없을 때 생존 분석에서 유용합니다.

분포

이 선택은 목표의 분포를 지정합니다. 비정규 분포와 항등하지 않은 연결함수를 지정하는 기능은 선형 혼합 모델에서 일반화 선형 혼합 모델의 중요한 개선 사항입니다. 많은 분포-연결함수 조합이 있으며 주어진 데이터 세트에 적합한 함수가 여러 개일 수 있으므로 사전 이론적 고려 사항을 바탕으로 선택하거나 어느 조합이 가장 적합할지를 고려하여 선택하면 됩니다.

이항

이 분포는 이분형 반응이나 이벤트 수를 나타내는 목표의 경우에만 적합합니다.

감마

이 분포는 더 큰 양의 값 쪽으로 비대칭되는 양의 척도 값을 가진 목표에 적합합니다. 데이터 값이 0보다 작거나 같고 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다.

역가우스

이 분포는 더 큰 양의 값 쪽으로 비대칭되는 양의 척도 값을 가진 목표에 적합합니다. 데이터 값이 0보다 작거나 같고 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다.

다항

이 분포는 다범주 반응을 나타내는 목표에 적합합니다. 모델 형태는 목표의 측정 수준에 따라 다릅니다.

명목 목표는 목표의 각 범주(참조 범주 제외)에 대해 별도의 모델 모수 세트가 추정되는 명목 다항 모델을 생성합니다. 주어진 예측변수에 대한 모수 추정값은 목표의 각 범주의 우도와 해당 예측변수 사이의 참조 범주와 관련한 관계를 보여줍니다.

순서 목표는 일반적인 절편 항이 목표 범주의 누적 확률과 관련된 **임계값** 모수의 세트로 대체되는 순서 다항 모델을 생성합니다.

음이항

음수 이항 회귀분석은 목표가 높은 분산의 발생 개수를 나타낼 때 사용되는 음수 이항 분포를 로그 링크와 함께 사용합니다.

정규

이것은 중앙(평균) 값에 대해 값이 대칭되는 종 형태의 분포를 띠는 연속형 목표에 적합합니다.

포아송

이 분포는 고정 기간 동안 중요 이벤트의 발생 수로 생각할 수 있으며 양의 정수 값을 갖는 변수에 적합합니다. 데이터 값이 양수이거나, 0보다 작거나 없는 경우 해당 케이스는 분석에 사용되지 않습니다.

연결함수

연결함수는 모델을 추정할 수 있는 목표의 변환입니다. 다음 함수를 사용할 수 있습니다.

항등

$f(x)=x$. 목표는 변환되지 않습니다. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.

보 로그-로그

$f(x)=\log(-\log(1-x))$. 이항 또는 다항 분포에만 적합합니다.

Cauchit

$f(x) = \tan(\pi (x - 0.5))$. 이항 또는 다항 분포에만 적합합니다.

로그

$f(x)=\log(x)$. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.

로그 보

$f(x)=\log(1-x)$. 이항 분포에만 적합합니다.

로짓

$f(x)=\log(x / (1-x))$. 이항 또는 다항 분포에만 적합합니다.

음 로그-로그

$f(x)=-\log(-\log(x))$. 이항 또는 다항 분포에만 적합합니다.

프로빗

$f(x)=\Phi^{-1}(x)$. 여기서 Φ^{-1} 은 역표준 정규 누적 분포 함수입니다. 이항 또는 다항 분포에만 적합합니다.

거듭제곱

$f(x)=x^\alpha$ ($\alpha \neq 0$ 의 경우), $f(x)=\log(x)$ ($\alpha=0$ 의 경우). α 는 필수 숫자 지정 사항이며 실수여야 합니다. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.

ㄴ. 고정 효과 (일반화 선형 혼합 모델)

고정 효과 요인은 일반적으로 관심 있는 해당 값이 모두 데이터 세트에 나타나는 필드로 볼 수 있으며 스코어링에 사용할 수 있습니다. 기본적으로 대화 상자에 지정되지 않은 사전 정의된 입력 역할이 있는 필드가 모델의 고정 효과 부분에 입력됩니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되고 연속형 필드는 공변량으로 사용됩니다.

소스 목록에서 하나 이상의 필드를 선택하고 효과 목록으로 끌어 모델에 효과를 입력합니다. 생성되는 효과의 유형은 선택을 끄는 핫스팟에 따라 다릅니다.

- 주. 끌어 놓은 필드가 효과 목록 하단에 별도의 주효과로 나타납니다.
- **이원.** 끌어 놓은 필드의 모든 가능한 쌍이 효과 목록 하단에 이원 상호작용으로 나타납니다.
- **삼원.** 끌어 놓은 필드의 모든 가능한 세 개로 구성된 세트가 효과 목록 하단에 3원배치 상호작용으로 나타납니다.
- *. 끌어 놓은 모든 필드의 조합은 효과 목록 아래쪽에 단일 상호작용으로 나타납니다.

효과 작성기의 오른쪽에 있는 단추를 사용하여 다양한 동작을 수행할 수 있습니다.

표 1. 효과 작성기 단추 설명

아이콘	설명
	삭제할 항목을 선택하고 삭제 단추를 클릭하여 고정 효과 모델에서 항목을 삭제할 수 있습니다.
	다시 정렬할 항목을 선택하고 위로 또는 아래로 화살표를 클릭하여 고정 효과 모델에서 항목을 다시 정렬할 수 있습니다.
	사용자 정의 항목 추가 (일반화 선형 혼합 모델) 대화 상자에서 사용자 정의 항목 추가 단추를 클릭하여 중첩 항목을 모델에 추가할 수 있습니다.

• 사용자 정의 항목 추가 (일반화 선형 혼합 모델)

이 프로시저에서는 모델에 대해 중첩 항목을 작성할 수 있습니다. 중첩 항목은 요인 또는 공변량의 효과를 모델링하는 데 유용합니다. 이들 값은 다른 요인 수준과 상호작용하지 않습니다. 예를 들어, 식료품 체인점은 여러 점포에서의 고객의 소비 성향을 살펴볼 수 있습니다. 각 고객은 체인점 중의 한 곳만 자주 가기 때문에 *고객* 효과는 *점포* 효과 내에 중첩되었다고 할 수 있습니다.

또한 같은 공변량을 포함하고 있는 다항 항목 같이 상호작용 효과를 포함시키거나 여러 수준의 중첩을 중첩 항목에 추가할 수 있습니다.

❗ 제한: 중첩 항목에는 다음과 같은 제한이 있습니다.

- 상호작용 내의 모든 요인은 고유해야 합니다. 따라서 A가 요인이면 A*A 지정은 유효하지 않습니다.
- 중첩 효과 내의 모든 요인은 고유해야 합니다. 따라서 A가 요인이면 A(A) 지정은 유효하지 않습니다.
- 공변량 내에 효과를 중첩할 수 없습니다. 따라서 A가 요인이고 X가 공변량이면 A(X) 지정은 유효하지 않습니다.

중첩 항목 작성

1. 다른 요인 내에 중첩된 요인 또는 공변량을 선택한 다음 화살표 단추를 클릭합니다.
2. **(포함)**을 클릭합니다.
3. 이전 요인이나 공변량이 중첩된 요인을 선택한 다음 화살표 단추를 클릭합니다.
4. **항 추가**를 클릭합니다.

선택적으로 상호작용 효과를 포함시키거나 여러 수준의 중첩을 중첩 항목에 추가할 수 있습니다.

ㄷ. 랜덤효과 (일반화 선형 혼합 모델)

랜덤 효과 요인은 데이터 파일의 값을 더 큰 모집단 값의 무작위 표본으로 간주할 수 있는 필드입니다. 목표의 과도한 변동을 설명할 때 유용합니다. 기본적으로, 데이터 구조 탭에서 개체를 둘 이상 선택한 경우, 가장 안쪽에 있는 개체 너머에 각 개체에 대해 변량효과 블록이 생성됩니다. 예를 들어, 데이터 구조 탭에서 학교, 클래스 및 학생을 개체로 선택한 경우, 다음 변량효과 블록이 자동으로 생성됩니다.

- 랜덤 효과 1: 개체는 학교입니다(효과 없이, 절편만 있음).
- 랜덤 효과 2: 개체는 학교 * 클래스입니다(효과 없이, 절편만 있음).

다음과 같은 방법으로 랜덤 효과 블록을 사용할 수 있습니다.

1. 새 블록을 추가하려면 **블록 추가...**를 클릭하십시오. 그러면 랜덤 효과 블록 (일반화 선형 혼합 모델) 대화 상자가 열립니다.
2. 기존 블록을 편집하려면, 편집할 블록을 선택하고 **블록 편집...**을 클릭하십시오, 그러면 랜덤 효과 블록 (일반화 선형 혼합 모델) 대화 상자가 열립니다.
3. 하나 이상의 블록을 삭제하려면 삭제하려는 블록을 선택하고 삭제 단추를 클릭하십시오.

• 랜덤 효과 블록 (일반화 선형 혼합 모델)

소스 목록에서 하나 이상의 필드를 선택하고 효과 목록으로 끌어 모델에 효과를 입력합니다. 생성되는 효과의 유형은 선택을 끄는 핫스팟에 따라 다릅니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되고 연속형 필드는 공변량으로 사용됩니다.

- **주.** 끌어 놓은 필드가 효과 목록 하단에 별도의 주효과로 나타납니다.
- **이원.** 끌어 놓은 필드의 모든 가능한 쌍이 효과 목록 하단에 이원 상호작용으로 나타납니다.
- **삼원.** 끌어 놓은 필드의 모든 가능한 세 개로 구성된 세트가 효과 목록 하단에 3원배치 상호작용으로 나타납니다.
- ***** 끌어 놓은 모든 필드의 조합은 효과 목록 아래쪽에 단일 상호작용으로 나타납니다.

효과 작성기의 오른쪽에 있는 단추를 사용하여 다양한 동작을 수행할 수 있습니다.

표 1. 효과 작성기 단추 설명

아이콘	설명
	삭제할 항목을 선택하고 삭제 단추를 클릭하여 모델에서 항목을 삭제할 수 있습니다.

아이콘	설명
	다시 정렬할 항목을 선택하고 위로 또는 아래로 화살표를 클릭하여 모델에서 항목을 다시 정렬할 수 있습니다.
	사용자 정의 항 추가 (일반화 선형 혼합 모델) 대화 상자에서 사용자 정의 항 추가 단추를 클릭하여 중첩 항을 모델에 추가할 수 있습니다.

절편 포함. 절편은 기본적으로 랜덤 효과 모델에 포함되지 않습니다. 데이터가 선형 회귀로 전달 된다고 가정할 경우에는 절편을 제외할 수 있습니다.

이 블록의 매개변수 예측 표시. 랜덤 효과 매개변수 추정값을 표시하려면 지정합니다.

공분산 그룹 정의 기준. 여기에서 지정된 범주형 필드는 랜덤 효과 공분산 모수의 독립 세트를 정의합니다(그룹 필드의 교차 분류에 의해 정의된 각 범주에 대해 하나씩). 각 랜덤 효과 블록에 다른 그룹 필드 세트를 지정할 수 있습니다. 모든 개체의 공분산 유형은 동일하며, 동일한 공분산 그룹 내의 개체는 모수에 대해 같은 값을 가집니다.

개체 조합. 데이터 구조 탭의 사전 설정된 개체 조합에서 랜덤 효과 개체를 지정할 수 있습니다. 예를 들어 데이터 구조 탭에 *학교*, *클래스* 및 *학생*이 순서대로 개체로 정의된 경우, 개체 조합 드롭다운 목록에는 **없음**, **학교**, **학교 * 클래스** 및 **학교 * 클래스 * 학생**이 옵션으로 표시됩니다.

랜덤 효과 공분산 유형. 잔차에 대한 공분산 구조를 지정합니다. 사용 가능한 구조는 다음과 같습니다.

- 1차 자기회귀(AR1)
- 이질적 자기회귀(ARH1)
- 자기회귀 이동 평균(1,1)(ARMA11)
- 복합 대칭
- 이질적 복합 대칭(CSH)
- 대각선
- 척도화 항등
- Toeplitz
- 비구조적
- 분산성분

ㄹ. 가중치 및 범위 (일반화 선형 혼합 모델)

분석 가중치. 척도 모수는 응답의 분산과 관련된 추정 모델 모수입니다. 분석 가중치는 "알려진" 값으로, 관측할 때마다 달라질 수 있습니다. 분석 가중 필드를 지정한 경우 반응의 변수와 관련된 척도 모수는 각 관측에 대해 분석 가중값으로 나눈 것입니다. 0보다 작거나 같고 또는 값이 없는 분석 가중값을 가진 레코드는 분석에 사용되지 않습니다.

오프셋. 오프셋 항은 "구조" 예측자입니다. 계수는 모델로 추정되지 않지만 값 1을 갖는 것으로 가정합니다. 따라서 오프셋 값은 단순히 목표의 선형 예측자에 추가됩니다. 이는 각 케이스가 중요 이벤트마다 다른 노출 수준을 가질 수 있는 포아송 회귀 모형에 특히 유용합니다.

예를 들어, 개별 운전자의 사고 비율을 모델링할 때 3년 경력 중 한 번의 사고를 낸 운전자와 25년 경력 중에 한 번 사고를 낸 운전자 사이에는 중요한 차이가 있습니다! 사고 수는 운전 경력의 자연 로그가 오프셋 항으로 포함되는 경우 로그 링크와 함께 포아송 또는 음이항 반응으로 모델링할 수 있습니다.

다른 분포 및 링크 유형의 조합에는 오프셋 변수의 다른 변환이 필요할 수 있습니다.

ㅁ. 일반 작성 옵션 (일반화 선형 혼합 모델)

이 선택은 모델을 작성하는 데 사용되는 몇몇 고급 기준을 지정합니다.

정렬 순서

이 제어는 "마지막" 범주를 결정하기 위해 목표 및 요인(범주형 입력)의 범주 순서를 결정합니다. 목표가 범주형이 아니거나 사용자 정의 참조 범주가 목표 (일반화 선형 혼합 모델) 설정에 지정된 경우, 목표 정렬 순서 설정이 무시됩니다.

중지 규칙

알고리즘에서 실행할 최대 반복 횟수를 지정할 수 있습니다. 알고리즘은 내부 루프와 외부 루프로 구성되는 이중 반복 프로세스를 사용합니다. 최대 반복 횟수에 지정된 값은 두 루프 모두에 적용됩니다. 음수가 아닌 정수를 지정합니다. 기본값은 100입니다.

사후 추정 설정

이 설정은 몇몇 모델 결과가 보기에 대해 어떻게 계산되는지를 결정합니다.

신뢰수준(%)

모델 계수의 구간 추정값을 계산하는 데 사용되는 신뢰수준입니다. 0보다 크고 100보다 작은 값을 지정하십시오. 기본값은 95입니다.

자유도

이 옵션은 유의수준 검정에 대해 자유도가 어떻게 계산되는지를 지정합니다. 표본 크기가 충분히 크거나, 데이터가 균형을 이루거나, 모델이 간단한 공분산 유형(예: 척도

법 항등 또는 대각선)을 사용하는 경우 **잔차 방법**을 선택합니다. 기본 설정입니다. 표본 크기가 작거나, 데이터가 비균형적이거나, 복잡한 공분산 유형(예: 비구조적)을 사용하는 경우 **검정마다 다름(Satterthwaite approximation)**을 선택합니다. 표본 크기가 작고 제한최대우도(REML) 모델을 사용하는 경우 **Kenward-Roger 근사값**을 선택하십시오.

고정 효과 및 계수의 검정

모수 추정값 공분산 교차표를 계산하는 방법입니다. 모델 가정을 위반할 염려가 있는 경우 강력한 추정을 선택하십시오.

나. 추정 (일반화 선형 혼합 모델)

모델 작성 알고리즘은 내부 루프와 외부 루프로 구성되는 이중 반복 프로세스를 사용합니다. 다음 설정은 내부 루프에 적용됩니다.

정렬 순서

이 제어는 "마지막" 범주를 결정하기 위해 목표 및 요인(범주형 입력)의 범주 순서를 결정합니다. 목표가 범주형이 아니거나 사용자 정의 참조 범주가 목표 (일반화 선형 혼합 모델) 설정에 지정된 경우, 목표 정렬 순서 설정이 무시됩니다.

모수 수렴.

수렴은 모수 추정값의 최대 절대 변화량 또는 최대 상대 변화량이 지정된 값(음수가 아니어야 함)보다 작다고 가정합니다. 지정된 값이 0인 경우 기준은 사용되지 않습니다.

로그-우도 수렴.

수렴은 로그 우도 함수의 절대 변화량 또는 상대 변화량이 지정된 값(음수가 아니어야 함)보다 작다고 가정합니다. 지정된 값이 0인 경우 기준은 사용되지 않습니다.

Hessian 수렴.

절대값 지정의 경우 수렴은 Hessian을 기준으로 하는 통계가 지정된 값보다 작다고 가정합니다. **상대값** 지정의 경우 수렴은 통계가 로그 우도의 절대값과 지정된 값의 곱보다 작다고 가정합니다. 지정된 값이 0인 경우 기준은 사용되지 않습니다.

최대 Fisher 스코어링 단계.

음수가 아닌 정수를 지정합니다. 0의 값은 Newton-Raphson 방법을 지정합니다. 0보다 큰 값은 반복 수 n 까지 Fisher 스코어링 알고리즘을 사용할 것을 지정합니다(여기서 n 은 지정된 정수임). 이후로는 Newton-Raphson을 사용합니다.

비정칙성 공차.

이 값은 비정칙성 확인 시 공차로 사용됩니다. 양수값을 지정하십시오.

중지 규칙

알고리즘에서 실행할 최대 반복 횟수를 지정할 수 있습니다. 알고리즘은 내부 루프와 외부 루프로 구성되는 이중 반복 프로세스를 사용합니다. 최대 반복 횟수에 지정된 값은 두 루프 모두에 적용됩니다. 음수가 아닌 정수를 지정합니다. 기본값은 100입니다.

사후 추정 설정

이 설정은 몇몇 모델 결과가 보기에 대해 어떻게 계산되는지를 결정합니다.

신뢰수준(%)

모델 계수의 구간 추정값을 계산하는 데 사용되는 신뢰수준입니다. 0보다 크고 100보다 작은 값을 지정하십시오. 기본값은 95입니다.

자유도

이 옵션은 유의수준 검정에 대해 자유도가 어떻게 계산되는지를 지정합니다. 표본 크기가 충분히 크거나, 데이터가 균형을 이루거나, 모델이 간단한 공분산 유형(예: 척도법 항등 또는 대각선)을 사용하는 경우 **간차 방법**을 선택합니다. 기본 설정입니다. 표본 크기가 작거나, 데이터가 비균형적이거나, 복잡한 공분산 유형(예: 비구조적)을 사용하는 경우 **검정마다 다름 (Satterthwaite approximation)**을 선택합니다. 표본 크기가 작고 제한최대우도(REML) 모델을 사용하는 경우 **Kenward-Roger 근사값**을 선택하십시오.

고정 효과 및 계수의 검정

모수 추정값 공분산 교차표를 계산하는 방법입니다. 모델 가정을 위반할 염려가 있는 경우 강력한 추정을 선택하십시오.

 **참고:** 기본적으로, 모수 수렴이 사용되며, 1E-6 공차에서 최대 **절대** 변경이 확인됩니다. 이 설정은 버전 22 이전 버전에서 확보되는 결과와 다른 결과를 생성할 수 있습니다. 22 이전 버전으로부터 결과를 생성하려면, 모수 수렴 기준에 대해 **상대값**을 사용하고 1E-6의 기본 공차값을 유지하십시오.

ㄱ. 일반 (일반화 선형 혼합 모델)

모델 이름. 목표 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 목표 필드 이름입니다. 목표가 여러 개일 경우, 모델 이름은 필드 이름이며 순서대로 앰퍼샌드로 연결됩니다. 예를 들어, *field1 field2 field3*가 목표가면, 모델 이름은 *field1 & field2 & field3*입니다.

스코어링에 사용 가능. 모델이 스코어링되면 이 그룹에서 선택한 항목이 생성되어야 합니다. (모든 목표에 대한) 예측값과 (범주형 목표에 대한) 신뢰도는 모델이 스코어링될 때 항상 계산됩니다. 계산된 신뢰도는 예측값의 확률(가장 높은 예측 확률) 또는 가장 높은 예측 확률과 두 번째로 가장 높은 예측 확률의 차이를 기준으로 할 수 있습니다.

- **범주형 목표의 예측 확률.** 범주형 목표에 대한 예측 확률이 생성됩니다. 범주마다 하나의 필드가 작성됩니다.
- **플래그 목표를 위한 성향 스코어.** (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 모델은 원시 성향 스코어를 생성합니다. 파티션이 적용 중일 경우 모델은 검정 분할에 근거한 수정된 성향 스코어도 생성합니다.

o. 평균 추정 (일반화 선형 혼합 모델)

이 탭에서는 요인 및 요인 상호작용의 수준에 대해 추정 주변 평균을 표시할 수 있습니다. 주변 평균 추정은 다항 모델에는 사용할 수 없습니다.

항

전체적으로 범주형 필드로만 구성된 고정 효과의 모델 항은 다음과 같습니다. 모델이 주변 평균 추정을 생성할 각 항을 확인합니다.

대비 유형

대비 필드의 수준에 대해 사용할 대비 유형을 지정합니다.

없음

대비가 생성되지 않습니다.

대응별

지정된 요인의 모든 수준 조합에 대해 쌍대 비교를 생성합니다. 이는 요인 상호작용의 대비에만 사용할 수 있습니다.

편차

요인의 각 수준을 총 평균과 비교합니다.

단순

마지막을 제외한 요인의 각 수준을 마지막 수준과 비교합니다. "마지막" 수준은 작성 옵션에 지정된 요인의 정렬 순서에 따라 결정됩니다. 이러한 대비 유형은 모두 직교하지 않음에 유의하십시오.

대비 필드

선택된 대비 유형을 사용하여 비교되는 수준인 요인은 지정합니다. 없음을 대비 유형으로 선택한 경우 대비 필드를 선택할 수 없거나 선택할 필요가 없습니다.

연속형 필드

나열된 연속형 필드는 연속형 필드를 사용하는 고정 효과의 항에서 추출됩니다. 주변 평균 추정을 계산할 때 공변량이 지정된 값으로 고정됩니다. 평균을 선택하거나 사용자 정의 값을 지정하십시오.

다중비교를 위한 수정

다중 대비를 통해 가설 검정을 수행하는 경우 포함된 대비에 대한 유의 수준에서 전반적인 유의 수준을 조정할 수 있습니다. 조정 방법을 선택할 수 있습니다.

최소유의차

이 방법은 특정 선형 대비가 귀무가설 값과 다르다는 가설을 거부할 전체 확률을 제어하지 않습니다.

순차 Bonferroni(Sequential Bonferroni)

순차 단계별로 낮아지는 거부 Bonferroni 프로시저로, 개별 가설은 거부하지만 동일한 전체 유의 수준을 유지한다는 점에서 훨씬 덜 보수적인 방법입니다.

순차 Sidak(Sequential Sidak)

순차 단계별로 낮아지는 거부 Sidak 프로시저로, 개별 가설은 거부하지만 동일한 전체 유의 수준을 유지한다는 점에서 훨씬 덜 보수적인 방법입니다.

최소유의차 방법은 순차 Sidak 방법보다 덜 보수적이므로 축차 Bonferroni 방법보다도 덜 보수적입니다. 다시 말해, 최소유의차는 적어도 순차 Sidak만큼 개별 가설을 거부하므로 축차 Bonferroni만큼 개별 가설을 거부하게 됩니다.

다음과 관련하여 추정 평균 표시

주변 평균 추정을 목표의 원래 척도를 기준으로 계산할지 연결함수 변환을 기준으로 계산할지 지정합니다.

원 목표 척도

목표의 추정 주변 평균을 계산합니다. 목표가 이벤트/시행 옵션을 사용하여 지정되었을 때 이벤트 수보다는 이벤트/시행 비율에 대한 주변 평균 추정을 제공합니다.

연결함수 변환

선형 예측자의 추정 주변 평균을 계산합니다.

ㄷ. 모델 보기 (일반화 선형 혼합 모델)

기본적으로 모델 요약 보기가 표시됩니다. 다른 모델 보기를 보려면 보기 축소판 그림에서 선택하십시오.

모델 오브젝트에 대한 일반 정보는 모델의 내용을 참조하십시오.

• 모델 요약 (일반화 선형 혼합 모델)

이 보기는 모델과 그 적합성을 한 눈에 파악할 수 있도록 요약한 스냅샷입니다.

테이블. 테이블은 목표 설정에 지정된 목표, 확률 분포 및 연결 함수를 식별합니다. 목표가 이벤트 및 시행에 의해 정의되면, 셀은 이벤트 필드와 시행 필드 또는 고정 시행 수를 표시하기 위해 분할됩니다. 또한 유한 표본 수정된 Akaike 정보 기준(AICC) 및 베이지안 정보 기준(BIC)이 표시됩니다.

- 수정된 Akaike. $-2(\text{제한})$ 로그 우도에 기반한 혼합 모형을 선택 및 비교하기 위한 척도입니다. 값이 작을수록 모형이 우수함을 나타냅니다. AICC는 작은 표본 크기의 AIC를 "수정합니다". 표본 크기가 증가함에 따라 AICC는 AIC로 수렴됩니다.

- **베이지안**. -2 로그 우도에 기반한 모형을 선택 및 비교하기 위한 척도입니다. 값이 작을수록 모형이 우수함을 나타냅니다. BIC도 초과 모수화된 모형(예: 입력이 많은 복잡한 모형)에 "페널티를 부여"하지만 AIC보다 더 엄격하게 부여합니다.

차트. 목표가 범주형인 경우 차트는 최종 모델의 정확도를 표시하며 이는 정확한 분류의 퍼센트입니다.

- **데이터 구조 (일반화 선형 혼합 모델)**

이 보기는 지정한 데이터 구조의 요약을 제공하며 개체와 반복 측도가 올바르게 지정되었는지 확인할 수 있도록 합니다. 첫 번째 개체에 대한 관측 정보가 각 개체 필드와 반복 측도 필드 및 목표에 대해 표시됩니다. 또한 각 개체 필드와 반복 측도 필드에 대한 수준 수가 표시됩니다.

- **관측값 별 예측값 (일반화 선형 혼합 모델)**

이벤트/시행으로 지정된 목표를 포함하여 연속형 목표의 경우, 수직축에 예측값의 구간화된 산점도를 표시하고 수평축에 관측값을 표시합니다. 이상적으로, 점이 45도 줄에 있어야 합니다. 이 보기는 레코드가 모델에 의해 특히 잘못 예측되었는지 여부를 알려줄 수 있습니다.

- **분류 (일반화 선형 혼합 모델)**

범주형 목표의 경우, 정확한 전체 퍼센트와 함께 관측값 대 예측값의 교차 분류를 히트 맵에 표시합니다.

테이블 유형. 다양한 표시 유형이 있으며, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **행 퍼센트**. 셀의 행 백분율(전체 행의 퍼센트로 표시되는 셀 개수)을 표시합니다. 이는 기본값입니다.
- **셀 개수**. 셀의 셀 개수를 표시합니다. 히트 맵의 음영이 행 퍼센트의 기준입니다.
- **히트 맵**. 셀의 값은 표시하지 않고 음영만 표시합니다.
- **압축**. 셀의 행 또는 열 머리말, 값을 표시하지 않습니다. 목표에 범주가 많은 경우에 유용할 수 있습니다.

결측. 레코드에 목표의 결측값이 있으면 모든 유효한 행 아래의 (**결측**) 행에 표시됩니다. 결측값이 있는 레코드는 정확한 전체 퍼센트에 기여하지 않습니다.

여러 목표. 여러 범주형 목표가 있는 경우, 각 목표가 별도의 테이블에 표시되고 표시되는 목표를 제어하는 **목표** 드롭다운 목록이 있습니다.

큰 테이블. 표시된 목표에 범주가 100개 이상 있으면 테이블이 표시되지 않습니다.

- 고정 효과 (일반화 선형 혼합 모델)

이 보기는 모델에서 각 고정 효과의 크기를 표시합니다.

유형. 다양한 표시 유형이 있으면, 유형 드롭다운 목록에서 액세스할 수 있습니다.

- **다이아그램.** 고정 효과 설정에 지정된 순서로 효과가 위에서 아래로 정렬되는 차트입니다. 다이어그램의 연결선은 효과 유의수준을 기준으로 가중되며 선이 굵을수록 더 유의한 효과(더 작은 p -값)입니다. 이는 기본값입니다.
- **테이블.** 전체 모형과 개별 모형 효과에 대한 ANOVA 테이블입니다. 고정 효과 설정에 지정된 순서로 각 효과가 위에서 아래로 정렬됩니다.

유의수준. 보기에 표시되는 효과를 제어하는 유의수준 슬라이더가 있습니다. 슬라이더 값보다 큰 유의수준 값이 있는 효과는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 효과에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의수준을 기준으로 필터링된 효과가 없습니다.

- 고정 계수 (일반화 선형 혼합 모델)

이 보기는 모델에서 각 고정 계수의 값을 표시합니다. 요인(범주형 예측자)이 모델 내에서 코딩된 지표이므로 요인을 포함하는 효과에는 일반적으로 여러 관련 계수가 있으며, 중복 계수에 해당하는 범주를 제외하고 각 범주에 하나씩 있습니다.

유형. 다양한 표시 유형이 있으면, 유형 드롭다운 목록에서 액세스할 수 있습니다.

- **다이아그램.** 절편을 먼저 표시한 다음 고정 효과 설정에 지정된 순서로 효과가 위에서 아래로 정렬되는 차트입니다. 요인을 포함하는 효과 내에서 계수가 데이터 값의 오름차순으로 정렬됩니다. 다이어그램의 연결선은 계수 유의수준을 기준으로 색상이 지정되고 가중되며 선이 굵을수록 더 유의한 계수(더 작은 p -값)입니다. 이것이 기본 유형입니다.
- **테이블.** 개별 모형 계수의 값, 유의수준 검정 및 신뢰구간을 표시합니다. 절편 다음으로, 고정 효과 설정에 지정된 순서로 효과가 위에서 아래로 정렬됩니다. 요인을 포함하는 효과 내에서 계수가 데이터 값의 오름차순으로 정렬됩니다.

다항. 효과에 다항 분포가 있는 경우, 다항 드롭다운 목록이 표시할 목표 범주를 제어합니다. 목록에서 값의 정렬 순서는 작성 옵션 설정의 지정 사항에 의해 결정됩니다.

지수. 이분형 로지스틱 회귀분석(이항 분포 및 로짓 링크), 명목 로지스틱 회귀분석(다항 분포 및 로짓 링크), 음수 이항 회귀분석(음수 이항 분포 및 로그 링크) 및 로그 선형 모델(포아송 분포 및 로그 링크) 등 특정 모델 유형에 대한 지수화된 계수 추정값 및 신뢰구간을 표시합니다.

유의수준. 보기에 표시되는 계수를 제어하는 유의수준 슬라이더가 있습니다. 슬라이더 값보다 큰 유의수준 값이 있는 계수는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 계수에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의수준을 기준으로 필터링된 계수가 없습니다.

- **랜덤효과 공분산 (일반화 선형 혼합 모델)**

이 보기는 랜덤 효과 공분산 행렬(G)을 표시합니다.

유형. 다양한 표시 유형이 있으면, 유형 드롭다운 목록에서 액세스할 수 있습니다.

- **공분산 값.** 고정 효과 설정에 지정된 순서로 효과가 위에서 아래로 정렬되는 공분산 행렬의 히트 맵입니다. 셀 값에 해당하는 상관도표의 색은 키에 표시된 것과 같습니다. 이는 기본값입니다.
- **상관도표.** 공분산 행렬의 히트 맵입니다.
- **압축.** 행 및 열 머리글이 없는 공분산 행렬의 히트 맵입니다.

블록. 여러 랜덤 효과 블록이 있는 경우, 표시할 블록을 선택할 수 있는 블록 드롭다운 목록이 있습니다.

그룹. 랜덤 효과 블록에 그룹 지정이 있는 경우, 표시할 그룹 수준을 선택할 수 있는 그룹 드롭다운 목록이 있습니다.

다항. 효과에 다항 분포가 있는 경우, 다항 드롭다운 목록이 표시할 목표 범주를 제어합니다. 목록에서 값의 정렬 순서는 작성 옵션 설정의 지정 사항에 의해 결정됩니다.

- **공분산 모수 (일반화 선형 혼합 모델)**

이 보기는 잔차 및 랜덤 효과에 대한 공분산 모수 추정값과 관련 통계를 표시합니다. 이것은 공분산 구조가 적합한지에 대한 정보를 제공하는 고급이지만 기본적인 결과입니다.

요약표. 잔차(R) 및 랜덤 효과(G) 공분산 행렬에서 모수의 수, 고정 효과(X) 및 랜덤 효과(Z) 디자인 행렬의 순위(열 수), 데이터 구조를 정의하는 개체 필드에 의해 정의되는 개체 수에 대한 빠른 참조입니다.

공분산 모수 표. 선택된 효과에 대해, 각 공분산 매개변수의 추정값, 표준 오차 및 신뢰구간이 표시됩니다. 표시되는 모수의 수는 효과 및 랜덤 효과 블록에 대한 공분산 구조와 블록의 효과 수에 따라 다릅니다. 비대각선 모수가 유의하지 않은 경우, 더 간단한 공분산 구조를 사용할 수 있습니다.

효과. 랜덤 효과 블록이 있는 경우, 표시할 잔차 또는 랜덤 효과 블록을 선택할 수 있는 효과 드롭다운 목록이 있습니다. 잔차 효과는 항상 사용할 수 있습니다.

그룹. 잔차 또는 랜덤 효과 블록에 그룹 지정이 있는 경우, 표시할 그룹 수준을 선택할 수 있는 그룹 드롭다운 목록이 있습니다.

다항. 효과에 다항 분포가 있는 경우, 다항 드롭다운 목록이 표시할 목표 범주를 제어합니다. 목록에서 값의 정렬 순서는 작성 옵션 설정의 지정 사항에 의해 결정됩니다.

- **평균 추정: 유의한 효과 (일반화 선형 혼합 모델)**

삼원 상호작용으로 시작하여 이원 상호작용, 마지막으로 주효과로 끝나는 10개의 "가장 유의한" 고정 모든 요인 효과를 표시하는 차트입니다. 이 차트는 수직축에 목표의 모델 추정값을 표시하고 수평축에 주효과(또는 상호작용에서 첫 번째 나열된 효과)의 각 값을 표시합니다. 상호작용에서 두 번째 나열된 효과의 각 값에 대해 별도의 선이 만들어집니다. 3원배치 상호작용에서 세 번째 나열된 효과의 각 값에 대한 별도의 차트가 만들어집니다. 기타 모든 예측자는 상수로 유지됩니다. 목표에서 각 예측자 계수의 효과를 시각화하는 데 유용합니다. 유의한 예측자가 없는 경우 평균 추정이 생성되지 않음에 유의하십시오.

신뢰도. 작성 옵션의 일부로 지정된 신뢰수준을 사용하여 주변 평균의 신뢰 한계 상한 및 하한을 표시합니다.

- **평균 추정: 사용자 정의 효과 (일반화 선형 혼합 모델)**

다음은 사용자가 요청한 고정 모든 요인 효과에 대한 테이블과 차트입니다.

유형. 다양한 표시 유형이 있으면, 유형 드롭다운 목록에서 액세스할 수 있습니다.

- **다이아그램.** 이 유형은 수직축에 목표의 모델 추정값 선형 차트를 표시하고 수평축에 주효과(또는 상호작용에서 첫 번째 나열된 효과)의 각 값을 표시합니다. 상호작용에서 두 번째 나열된 효과의 각 값에 대해 별도의 선이 만들어집니다. 삼원 상호작용에서 세 번째 나열된 효과의 각 값에 대한 별도의 차트가 만들어집니다. 기타 모든 예측자는 상수로 유지됩니다. 대비가 요청된 경우, 대비 필드의 수준을 비교하기 위해 또 다른 차트가 표시됩니다. 상호작용의 경우, 대비 필드 외에 효과의 각 수준 조합에 대해 차트가 표시됩니다. **대응별** 대비의 경우, 거리 네트워크 차트입니다. 즉 비교 테이블을 그래픽으로 나타낸 것으로서 네트워크 노드 간의 거리는 표본 간의 차이에 해당합니다. 노란색 선은 통계적으로 유의차에 해당하며, 검정색 선은 비유의차에 해당합니다. 네트워크의 선에 마우스를 올려 놓으면 선으로 연결된 노드 간의 조정된 유의수준차가 도구 팁에 표시됩니다.

편차 대비의 경우, 수직축에 목표의 모델 추정값을 표시하고 수평축에 대비 필드의 값을 표시하는 막대형 차트가 표시됩니다. 상호작용의 경우, 대비 필드 외에 효과의 각 수준 조합에 대해 차트가 표시됩니다. 막대는 대비 필드의 각 수준과 전체 평균 간의 차이를 표시하며, 검은색 가로선으로 나타냅니다.

단순 대비의 경우, 수직축에 목표의 모델 추정값을 표시하고 수평축에 대비 필드의 값을 표시하는 막대형 차트가 표시됩니다. 상호작용의 경우, 대비 필드 외에 효과의 각 수준 조합에 대해 차트가 표시됩니다. 막대는 대비 필드의 각 수준(마지막 제외)과 마지막 수준 간의 차이를 표시하며, 검은색 가로선으로 나타냅니다.

- **테이블.** 이 유형은 목표의 모델 추정값, 표준 오차 및 효과에서 필드의 각 수준 조합에 대한 신뢰구간의 테이블을 표시합니다. 기타 모든 예측자는 상수로 유지됩니다. 대비가 요청된 경우, 추정값, 표준 오차, 유의수준 검정 및 각 대비에 대한 신뢰구간이 있는 또 다른 테이블이 표시됩니다. 상호작용의 경우, 대비 필드 외에 효과의 각 수준 조합에 대한 별도의 행 세트가 있습니다. 또한 전체 검정 결과가 있는 테이블이 표시됩니다. 상호작용의 경우, 대비 필드 외에 효과의 각 수준 조합에 대한 별도의 전체 검정이 있습니다.

신뢰도. 작성 옵션의 일부로 지정된 신뢰수준을 사용하여 주변 평균의 신뢰 한계 상한 및 하한 표시를 토글합니다.

윤곽. 대응별 대비 다이어그램의 윤곽을 토글합니다. 원 윤곽은 망 윤곽보다 대비를 덜 나타내지만 선이 겹쳐지지 않습니다.

- **설정 (일반화 선형 혼합 모델)**

모델이 스코어링되면 이 탭에서 선택한 항목이 생성되어야 합니다. (모든 목표에 대한) 예측값과 (범주형 목표에 대한) 신뢰도는 모델이 스코어링될 때 항상 계산됩니다. 계산된 신뢰도는 예측값의 확률(가장 높은 예측 확률) 또는 가장 높은 예측 확률과 두 번째로 가장 높은 예측 확률의 차이를 기준으로 할 수 있습니다.

- **범주형 목표의 예측 확률.** 범주형 목표에 대한 예측 확률이 생성됩니다. 범주마다 하나의 필드가 작성됩니다.
- **플래그 목표를 위한 성향 스코어.** (yes 또는 no 예측을 반환하는) 플래그 목표가 있는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 나타내는 성향 스코어를 요청할 수 있습니다. 모델은 원시 성향 스코어를 생성합니다. 파티션이 적용 중일 경우 모델은 검정 분할에 근거한 수정된 성향 스코어도 생성합니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

(10) GLE 노드

GLE 모델은 지정된 연결 함수를 통해 요인 및 공변량과 선형적으로 관련된 종속변수를 식별합니다. 더욱이 모델을 사용하면 종속변수가 비정규 분포를 가질 수 있습니다. 이 모델은 정상적으로 분포된 반응에 대한 선형 회귀, 이분형 데이터에 대한 로지스틱 모델, 개수 데이터에 대한 선형로그 모델, 구간 중도절단 생존 데이터에 대한 보 로그-로그 모델 등의 널리 사용되는 통계 모델뿐만 아니라, 매우 일반적인 모델 작성 공식을 통해 작성된 다른 많은 통계 모델을 포함합니다.

예. 운송 회사에서는 일반화 선형 모델을 사용하여 서로 다른 기간에 구성된 선박의 여러 유형에 대한 손상 횟수에 포아송 회귀분석을 맞출 수 있습니다. 그리고 결과로 생성된 모델은 손상될 확률이 높은 선박 유형을 판별하는 데 도움이 될 수 있습니다.

자동차 보험 회사는 일반화 선형 모델을 사용하여 자동차의 손해 배상 청구에 감마회귀를 맞출 수 있습니다. 결과로 생성되는 모델은 청구 규모에 가장 많이 기여하는 요인을 판별하는 데 도움을 줄 수 있습니다.

의료 연구자는 일반화 선형 모델을 사용하여 구간별 검열된 생존 데이터에 보 로그-로그 회귀분석을 맞추어 의료 조건에 대한 재발 시간을 예측할 수 있습니다.

GLE 모델은 입력 필드 값을 출력 필드 값에 관련시키는 방정식을 작성하여 작동됩니다. 모델이 생성되면 새 데이터의 값을 추정하는 데 사용할 수 있습니다.

각 범주형 대상의 각 레코드의 경우 가능한 출력 범주마다 소속 확률이 계산됩니다. 확률이 가장 높은 목표 범주는 해당 레코드의 예측된 출력 값으로 지정됩니다.

요구사항. 하나 이상의 입력 필드와 둘 이상의 범주를 포함하는 정확히 하나의 목표 필드(측정 수준이 연속형, 범주형 또는 플래그일 수 있음)가 필요합니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다.

① 목표(GLE 모델)

이 설정은 목표, 분포 및 연결함수를 통한 예측변수에 대한 관계를 정의합니다.

목표 목표는 필수입니다. 어떤 측정 수준도 가질 수 있으며, 대상의 측정 수준은 적합한 분포 및 연결함수에 영향을 줍니다.

- **사전 정의된 대상 사용** 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 대상 설정을 사용하려면 이 옵션을 선택하십시오.
- **사용자 정의 대상 사용** 대상을 수동으로 지정하려면 이 옵션을 선택하십시오.
- **시행 수를 분모로 사용** 목표 반응이 시행 세트에서 발생하는 다수의 이벤트인 경우, 대상 필드에는 이벤트 수가 포함되며 시행 수를 포함하는 추가 필드를 선택할 수 있습니다. 예를 들어, 새 살충제를 실험할 때 개미 표본에 다른 농도의 살충제를 사용하고 죽은 개미 수와 각 표본의 개미 수를 기록할 수 있습니다. 이 경우 죽은 개미 수를 기록한 필드를 목표(이벤트) 필드로 지정하고, 각 표본의 개미 수를 기록한 필드를 시행 필드로 지정해야 합니다. 각 표본의 개미 수가 동일한 경우 시행 수를 고정 값으로 지정할 수 있습니다. 각 레코드에 대해 시행 수는 이벤트 수보다 크거나 같아야 합니다. 이벤트는 음이 아닌 정수가 되어야 하며 시행 수는 양의 정수가 되어야 합니다.
- **참조 범주 사용자 정의.** 범주형 목표의 경우, 참조 범주를 선택할 수 있습니다. 이것은 모수 추정값과 같은 특정 결과에 영향을 미칠 수 있지만 모델 적합을 변경해서는 안 됩니다. 예를 들어 목표가 0, 1, 2 값을 가지면, 기본적으로 프로시저는 마지막(가장 높은 값) 범주 또는 참조 범주인 2를 만듭니다. 이 상황에서 모수 추정값은 범주 2의 우도에 *비례하여* 범주 0 또는 1의 우도와 관련된 것으로 해석해야 합니다. 사용자 정의 범주를 지정할 때 목표에 정의된 레이블이 있는 경우 목록에서 값을 선택하여 참조 범주를 설정할 수 있습니다. 이는 모델을 지정하는 동안 정확히 특정 필드를 어떻게 코딩했는지 기억이 나지 않을 때 편리할 수 있습니다.

목표 분포 및 선형 모형과의 관계(링크). 예측변수의 값이 주어지면 모델은 목표 값 분포가 지정된 형태를 따르고 목표 값이 지정된 연결함수를 통해 해당 예측변수와 선형적으로 관련될 것으로 예상합니다. 여러 공통 모델에 대한 바로 가기가 제공되거나, 최종 목록에 없는 적합하도록 할 특정 분포 및 연결함수 조합이 있는 경우에는 **사용자 정의**를 선택하십시오.

- **선형 모형** 정규 분포를 항등 링크와 함께 지정합니다. 이는 선형 회귀 또는 ANOVA 모델을 사용하여 목표를 예측할 수 있을 때 유용합니다.
- **감마 회귀분석** 감마 분포를 로그 링크와 함께 지정합니다. 이는 목표에 모든 양수값이 포함되고 더 큰 값 쪽으로 비대칭될 때 사용해야 합니다.
- **로그선형** 포아송 분포를 로그 링크와 함께 지정합니다. 이는 목표가 고정 기간 동안 발생 개수를 나타낼 때 사용해야 합니다.
- **음수 이항 회귀분석** 음수 이항 분포를 로그 링크와 함께 지정합니다. 이는 목표와 분모가 k 성공을 관측하는 데 필요한 시행 수를 나타낼 때 사용해야 합니다.

- **트위디 회귀분석** 항등, 로그 또는 거듭제곱 연결함수를 사용하여 트위디 분포를 지정하고 0과 양의 실수 값의 혼합형인 모델링 반응에 사용할 수 있습니다. 이러한 분포는 **복합 포아송**, **복합 감마**, **포아송-감마** 분포라고도 합니다.
- **다항 로지스틱 회귀분석** 다항 분포를 지정합니다. 이는 목표가 다범주 반응일 때 사용해야 합니다. 누적 로짓 링크(순서 결과)나 일반화된 로짓 링크(다범주 명목 반응)를 사용합니다.
- **이분형 로지스틱 회귀분석** 이항 분포를 로짓 링크와 함께 지정합니다. 이는 목표가 로지스틱 회귀분석 모델에 의해 예측된 이분형 반응일 때 사용해야 합니다.
- **이분형 프로빗** 이항 분포를 프로빗 링크와 함께 지정합니다. 이는 목표가 기본 정규 분포가 있는 이분형 반응일 때 사용해야 합니다.
- **구간 중도절단 생존** 이항 분포를 보 로그-로그 링크와 함께 지정합니다. 이는 몇몇 관측값에 종료 이벤트가 없을 때 생존 분석에서 유용합니다.
- **사용자 정의 분포 및 연결함수의 고유 조합**을 지정합니다.

분포

이 선택사항은 목표의 **분포**를 지정합니다. 비정규 분포와 항등하지 않은 연결함수를 지정하는 기능은 선형 모델 중에서 일반화 선형 모델의 중요한 개선 사항입니다. 많은 분포-연결함수 조합이 있으며 주어진 데이터 세트에 적합한 함수가 여러 개일 수 있으므로 사전 이론적 고려 사항을 바탕으로 선택하거나 어느 조합이 가장 적합할지를 고려하여 선택하면 됩니다.

- **자동** 사용할 분포를 모르는 경우 이 옵션을 선택하십시오. 노드가 데이터를 분석하여 최적의 분포 방법을 추정하여 적용합니다.
- **이항** 이 분포는 이분형 반응이나 이벤트 수를 나타내는 목표에만 적합합니다.
- **감마** 이 분포는 더 큰 양의 값 쪽으로 비대칭되는 양의 척도 값을 가진 목표에 적합합니다. 데이터 값이 0보다 작거나 같고 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다.
- **역 가우스** 이 분포는 더 큰 양의 값 쪽으로 비대칭되는 양의 척도 값을 가진 목표에 적합합니다. 데이터 값이 0보다 작거나 같고 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다.
- **다항** 이 분포는 다범주 반응을 나타내는 목표에 적합합니다. 모델 형태는 목표의 측정 수준에 따라 다릅니다.

명목 목표는 목표의 각 범주(참조 범주 제외)에 대해 별도의 모델 모수 세트가 추정되는 명목 다항 모델을 생성합니다. 주어진 예측변수에 대한 모수 추정값은 목표의 각 범주의 우도와 해당 예측변수 사이의 참조 범주와 관련한 관계를 보여줍니다.

순서 목표는 일반적인 절편 항이 목표 범주의 누적 확률과 관련된 **임계값** 모수의 세트로 대체되는 순서 다항 모델을 생성합니다.

- **음수 이항** 음수 이항 회귀분석은 목표가 높은 분산의 발생 개수를 나타낼 때 사용되는 음수 이항 분포를 로그 링크와 함께 사용합니다.
- **정규 분포** 이것은 중앙(평균) 값에 대해 값이 대칭되는 종 형태의 분포를 띠는 연속형 목표에 적합합니다.

- **포아송** 이 분포는 고정 기간 동안 중요 이벤트의 발생 수로 생각할 수 있으며 양의 정수 값을 갖는 변수에 적합합니다. 데이터 값이 양수이거나, 0보다 작거나 없는 경우 해당 케이스는 분석에 사용되지 않습니다.
- **트위디** 이 분포는 감마 분포의 포아송 혼합으로 표현할 수 있는 변수에 적합합니다. 분포는 연속 특성(음이 아닌 실수 값 사용)과 이산형 분포(단일 값 0에서 양의 확률 매스)를 조합한다는 점에서 "혼합"된 것입니다. 종속변수는 데이터 값이 0보다 크거나 같은 숫자가 되어야 합니다. 데이터 값이 0보다 작거나 또는 없는 경우 해당 케이스는 분석에 사용되지 않습니다. Tweedie 분포에서 모수의 고정 값은 1보다 크고 2보다 작은 숫자가 될 수 있습니다.

연결함수

연결함수는 모델을 추정할 수 있는 목표의 변환입니다. 다음 함수를 사용할 수 있습니다.

- **자동** 사용할 연결을 모르는 경우 이 옵션을 선택하십시오. 노드가 데이터를 분석하여 최적의 연결함수를 추정하여 적용합니다.
- **항등** $f(x)=x$. 목표는 변환되지 않습니다. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.
- **보 로그-로그** $f(x)=\log(-\log(1-x))$. 이항 또는 다항 분포에만 적합합니다.
- **Cauchit** $f(x) = \tan(\pi (x-0.5))$. 이항 또는 다항 분포에만 적합합니다.
- **로그** $f(x)=\log(x)$. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.
- **로그 보** $f(x)=\log(1-x)$. 이항 분포에만 적합합니다.
- **로짓** $f(x)=\log(x / (1-x))$. 이항 또는 다항 분포에만 적합합니다.
- **음수 로그-로그** $f(x)=-\log(-\log(x))$. 이항 또는 다항 분포에만 적합합니다.
- **프로빗** $f(x)=\Phi^{-1}(x)$. 여기서 Φ^{-1} 은 표준 정규 누적 분포의 역함수입니다. 이항 또는 다항 분포에만 적합합니다.
- **거듭제곱** $f(x)=x^\alpha$, if $\alpha \neq 0$. $f(x)=\log(x)$ ($\alpha=0$ 의 경우). α 는 필수 숫자 지정 사항이며 실수여야 합니다. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.

트위디의 모수 트위디 회귀분석 단일 선택 단추 또는 트위디를 분포 방법으로 선택한 경우에만 사용할 수 있습니다. 1과 2 사이의 값을 선택하십시오.

② 모델 효과(GLE 모델)

고정 효과 요인은 일반적으로 관심 있는 해당 값이 모두 데이터 세트에 나타나는 필드로 볼 수 있으며 스코어링에 사용할 수 있습니다. 기본적으로 대화 상자에 지정되지 않은 사전 정의된 입력 역할이 있는 필드가 모델의 고정 효과 부분에 입력됩니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되고 연속형 필드는 공변량으로 사용됩니다.

소스 목록에서 하나 이상의 필드를 선택하고 효과 목록으로 끌어 모델에 효과를 입력합니다. 생성되는 효과의 유형은 선택을 끄는 핫스팟에 따라 다릅니다.

- 주 끌어 놓은 필드가 효과 목록 맨 아래에 별도의 주 효과로 나타납니다.
- 이원 끌어 놓은 필드의 모든 가능한 쌍이 효과 목록 맨 아래에 이원 상호작용으로 나타납니다.
- 삼원 끌어 놓은 필드의 모든 가능한 쌍이 효과 목록 맨 아래에 삼원 상호작용으로 나타납니다.
- * 끌어 놓은 모든 필드의 조합이 효과 목록 맨 아래에 단일 상호작용으로 나타납니다.

효과 작성기의 오른쪽에 있는 단추를 사용하여 다양한 동작을 수행할 수 있습니다.

표 1. 효과 작성기 단추 설명	
아이콘	설명
	삭제할 항을 선택하고 삭제 단추를 클릭하여 고정 효과 모델에서 항목을 삭제할 수 있습니다.
	다시 정렬할 항을 선택하고 위로 또는 아래로 화살표를 클릭하여 고정 효과 모델에서 항목을 다시 정렬할 수 있습니다.
	사용자 정의 항 추가 대화 상자에서 사용자 정의 항 추가 단추를 클릭하여 중첩 항을 모델에 추가하십시오.

절편 포함 절편이 대체로 모델에 포함됩니다. 데이터가 선형 회귀로 전달된다고 가정할 경우에는 절편을 제외할 수 있습니다.

가. 사용자 정의 항 추가(GLE 모델)

이 프로시저에서는 모델에 대해 중첩 항을 작성할 수 있습니다. 중첩 항은 요인 또는 공변량의 효과를 모델링하는 데 유용합니다. 이들 값은 다른 요인 수준과 상호작용하지 않습니다. 예를 들어, 식료품 체인점은 여러 점포에서의 고객의 소비 성향을 살펴볼 수 있습니다. 각 고객은 체인점 중의 한 곳만 자주 가기 때문에 고객 효과는 점포 효과 *내*에 중첩되었다고 할 수 있습니다.

또한 같은 공변량을 포함하고 있는 다항 항 같이 상호작용 효과를 포함시키거나 여러 수준의 중첩을 중첩 항에 추가할 수 있습니다.

제한사항. 중첩 항에는 다음과 같은 제한이 있습니다.

- 상호작용 내의 모든 요인은 고유해야 합니다. 따라서 A 가 요인이면 $A*A$ 지정은 유효하지 않습니다.
- 중첩 효과 내의 모든 요인은 고유해야 합니다. 따라서 A 가 요인이면 $A(A)$ 지정은 유효하지 않습니다.
- 공변량 내에 효과를 중첩할 수 없습니다. 따라서 A 가 요인이고 X 가 공변량이면 $A(X)$ 지정은 유효하지 않습니다.

중첩 항 작성

1. 다른 요인 내에 중첩된 요인 또는 공변량을 선택한 다음 화살표 단추를 클릭합니다.
2. **(포함)**을 클릭합니다.
3. 이전 요인이나 공변량이 중첩된 요인을 선택한 다음 화살표 단추를 클릭합니다.
4. **항 추가**를 클릭합니다.

선택적으로 상호작용 효과를 포함시키거나 여러 수준의 중첩을 중첩 항에 추가할 수 있습니다.

③ 가중치 및 오프셋(GLE 모델)

분석 가중치 척도 모수는 반응의 분산과 관련한 추정된 모델 모수입니다. 분석 가중치는 "알려진" 값으로, 관측할 때마다 달라질 수 있습니다. **분석 가중치** 필드를 지정한 경우, 반응의 분산과 관련한 척도 모수는 각 관측값에 대해 분석 가중치로 나눈 값입니다. 0 이하이거나 값이 없는 분석 가중치를 가진 레코드는 분석에 사용되지 않습니다.

오프셋 오프셋 항은 구조 예측변수입니다. 계수는 모델로 추정되지 않지만 값 1을 갖는 것으로 가정합니다. 따라서 오프셋 값은 단순히 목표의 선형 예측변수에 추가됩니다. 이는 각 케이스가 중요 이벤트마다 다른 노출 수준을 가질 수 있는 포아송 회귀 모형에 특히 유용합니다.

예를 들어, 개별 운전자의 사고 비율을 모델링할 때 3년 경력 중 한 번의 사고를 낸 운전자와 25년 경력 중에 한 번 사고를 낸 운전자 사이에는 중요한 차이가 있습니다. 사고 수는 운전 경력의 자연 로그가 오프셋 항으로 포함되는 경우 로그 링크와 함께 포아송 또는 음이항 반응으로 모델링할 수 있습니다.

다른 분포 및 링크 유형의 조합에는 오프셋 변수의 다른 변환이 필요할 수 있습니다.

④ 작성 옵션(GLE 모델)

이 선택은 모델을 작성하는 데 사용되는 몇몇 고급 기준을 지정합니다.

정렬 순서 이 제어는 "마지막" 범주를 결정하기 위해 목표 및 요인(범주형 입력)의 범주 순서를 결정합니다. 목표가 범주형이 아니거나 사용자 정의 참조 범주가 목표(GLE 모델) 설정에 지정된 경우, 목표 정렬 순서 설정이 무시됩니다.

사후 추정 설정 이 설정은 표시를 위해 일부 모델 결과가 계산되는 방법을 결정합니다.

- **신뢰수준 %** 모델 계수의 구간 추정값을 계산하는 데 사용되는 신뢰수준입니다. 0보다 크고 100보다 작은 값을 지정하십시오. 기본값은 95입니다.

- **자유도** 이 옵션은 유의수준 검정을 위해 자유도가 계산되는 방법을 지정합니다. 표본 크기가 충분히 크거나, 데이터가 균형을 이루거나, 척도법 항등 또는 대각선처럼 모델이 간단한 공분산 유형을 사용하는 경우 **모든 검정에 대해 고정(잔차 방법)**을 선택하십시오. 이는 기본값입니다. 표본 크기가 작거나, 데이터가 비균형적이거나, 비구조적처럼 모델이 복잡한 공분산 유형을 사용하는 경우 **검정마다 다름(Satterthwaite approximation)**을 선택하십시오.
- **고정 효과 및 계수의 검정**. 모수 추정값 공분산 교차표를 계산하는 방법입니다. 모델 가정을 위반할 염려가 있는 경우 강력한 추정을 선택하십시오.

영향력 있는 이상값 발견 다항 분포를 제외한 모든 분포에서 영향력 있는 이상값을 식별하려면 이 옵션을 선택하십시오.

추세 분석 수행 산점도 도표에서 추세 분석을 수행하려면 이 옵션을 선택하십시오.

⑤ 추정(GLE 모델)

방법 사용할 최대우도 추정 방법을 선택하십시오. 사용가능 옵션은 다음과 같습니다.

- Fisher 스코어링
- Newton-Raphson
- 하이브리드

최대 Fisher 반복 수 음수가 아닌 정수를 지정하십시오. 0의 값은 Newton-Raphson 방법을 지정합니다. 0보다 큰 값은 반복 수 n 까지 Fisher 스코어링 알고리즘을 사용할 것을 지정합니다 (여기서 n 은 지정된 정수임). 이후로는 Newton-Raphson을 사용합니다.

척도 모수 방법 척도 모수에 대한 추정 방법을 선택하십시오. 사용가능 옵션은 다음과 같습니다.

- 최대우도 추정값
- 고정값. 사용할 값을 설정할 수도 있습니다.
- 편차
- Pearson 카이제곱

음이항 방법 음이항 보조 모수에 대한 추정 방법을 선택하십시오. 사용가능 옵션은 다음과 같습니다.

- 최대우도 추정값
- 고정값. 사용할 값을 설정할 수도 있습니다.

음이 아닌 최소 제곱(NNLS) 수행 음이 아닌 최소 제곱(NNLS) 추정을 수행하려면 이 옵션을 선택하십시오. NNLS는 계수가 음수가 될 수 없는 제약된 최소 제곱 문제 유형입니다. 모든 데이터 세트가 NNLS에 적합한 것은 아니므로, 예측변수와 목표값 간에 양의 상관 관계가 필요하거나 아예 필요하지 않습니다.

모수 수렴(Parameter Convergence) 수렴은 모수 추정값의 최대 절대 변화량 또는 최대 상대 변화량이 지정된 값(음수가 아니어야 함)보다 작다고 가정합니다. 지정된 값이 0인 경우 기준은 사용되지 않습니다.

로그-우도 수렴 수렴은 로그 우도 함수의 절대 변화량 또는 상대 변화량이 지정된 값(음수가 아니어야 함)보다 작다고 가정합니다. 지정된 값이 0인 경우 기준은 사용되지 않습니다.

Hessian 수렴 절대값 지정의 경우 수렴은 Hessian을 기준으로 하는 통계가 지정된 값보다 작다고 가정합니다. **상대값** 지정의 경우 수렴은 통계가 로그 우도의 절대값과 지정된 값의 곱보다 작다고 가정합니다. 지정된 값이 0인 경우 기준은 사용되지 않습니다.

최대 반복 수 알고리즘에서 실행할 최대 반복 횟수를 지정할 수 있습니다. 알고리즘은 내부 루프와 외부 루프로 구성되는 이중 반복 프로세스를 사용합니다. 최대 반복 횟수에 지정된 값은 두 루프 모두에 적용됩니다. 음수가 아닌 정수를 지정합니다. 기본값은 100입니다.

비정칙성 공차 이 값은 비정칙성 확인 시 공차로 사용됩니다. 양수값을 지정하십시오.

참고: 기본적으로, **모수 수렴**이 사용되며, 1E-6 공차에서 최대 **절대** 변경이 확인됩니다. 이 설정은 버전 17 이전 버전에서 확보되는 결과와 다른 결과를 생성할 수 있습니다. 17 이전 버전에서 결과를 생성하려면 모수 수렴 기준에 대해 **상대값**을 사용하고 1E-6의 기본 공차값을 유지하십시오.

⑥ 모델 선택(GLE 모델)

모델 선택 또는 정규화 사용 이 분할창에서 제어를 활성화하려면 이 선택란을 선택하십시오.

방법 모델 선택 방법을 선택하거나 사용할 정규화(능형을 사용하는 경우)를 선택하십시오. 다음 옵션에서 선택할 수 있습니다.

- **Lasso L1** 정규화라고도 하며 이 방법은 예측자 수가 많은 경우 단계별 전진보다 빠릅니다. 이 방법은 모수를 축소하여(즉, 페널티를 부과하여) 과적합을 방지합니다. 일부 모수를 0으로 축소하여 변수 선택 lasso를 수행할 수 있습니다.
- **능형 L2** 정규화라고도 하며 이 방법은 모수를 축소하여(즉, 페널티를 부과하여) 과적합을 방지합니다. 이 방법은 동일한 비율로 모수를 모두 축소하며, 변수 선택 방법은 아닙니다.
- **Elastic net L1 + L2** 정규화라고도 하며 이 방법은 모수를 축소하여(즉, 페널티를 부과하여) 과적합을 방지합니다. 일부 모수를 0으로 축소하여 변수 선택을 수행할 수 있습니다.
- **단계별 전진** 이 방법은 모델에서 아무런 효과 없이 시작하여 단계 선택 기준에 따라 더는 추가하거나 제거할 수 없을 때까지 한 번에 한 단계에서 효과를 추가하거나 제거합니다.

이원 상호작용 자동 발견 - 이원 상호작용을 자동으로 발견하려면 이 옵션을 선택하십시오.

페널티 모수

Lasso 또는 Elastic Net 방법을 선택하는 경우에만 이 옵션을 사용할 수 있습니다.

페널티 모수 자동 선택 설정할 모수 페널티를 모르는 경우 이 선택란을 선택하면 노드가 페널티를 식별하고 적용합니다.

Lasso 페널티 모수 Lasso 모델 선택 방법에서 사용할 페널티 모수를 입력하십시오.

Elastic net 페널티 모수 1 Elastic net 모델 선택 방법에서 사용할 L1 페널티 모수를 입력하십시오.

Elastic net 페널티 모수 2 Elastic net 모델 선택 방법에서 사용할 L2 페널티 모수를 입력하십시오.

단계별 전진

단계별 전진 방법을 선택하는 경우에만 이 옵션을 사용할 수 있습니다.

P-값이 특정 값 이상인 경우 효과 포함 효과를 계산에 포함할 수 있는 최소 확률 값을 지정하십시오.

P-값이 특정 값을 초과하는 경우 효과 제거 효과를 계산에 포함할 수 있는 최대 확률 값을 지정하십시오.

최종 모델에서 최대 효과 수 사용자 정의 최대 효과 수 옵션을 활성화하려면 이 선택란을 선택하십시오.

최대 효과 수 단계별 전진 작성 방법을 사용하는 경우 최대 효과 수를 지정하십시오.

최대 단계 수 사용자 정의 최대 단계 수 옵션을 활성화하려면 이 선택란을 선택하십시오.

최대 단계 수 단계별 전진 작성 방법을 사용하는 경우 최대 단계 수를 지정하십시오.

⑦ 모델 옵션(GLE 모델)

모델 이름 - 대상 필드를 기반으로 모델 이름을 자동으로 생성하거나 **사용자 정의** 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 대상 필드 이름입니다. 목표가 여러 개일 경우, 모델 이름은 필드 이름이며 순서대로 앰퍼샌드로 연결됩니다. 예를 들어, field1, field2 및 field3가 목표이면 모델 이름은 *field1 & field2 & field3*입니다.

예측변수 중요도 계산 적절한 중요도 측도를 생성하는 모델의 경우, 모델 추정에서 각 예측변수의 상대적 중요도를 나타내는 차트를 표시할 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 예측변수 중요도는 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 몇몇 모델은 기본적으로 해제되어 있음에 유의하십시오.

추가 정보는 예측변수 중요도의 내용을 참조하십시오.

⑧ GLE 모델 너깃

가. GLE 모델 너깃 출력

GLE 모델을 작성하면 출력에서 다음 정보를 사용할 수 있습니다.

모델 정보 테이블

모델 정보 테이블에서는 모델에 대한 주요 정보를 제공합니다. 테이블은 다음과 같은 일부 상위 수준 모델 설정을 식별합니다.

- 유형 노드 또는 GLE 노드 필드 탭에서 선택된 대상 필드 이름.
- 모델링된 참조 대상 범주 퍼센트.
- 확률 분포 및 연관된 연결함수.
- 사용되는 모델 작성 방법.
- 예측변수의 수 입력 및 최종 모델의 수.
- 분류 정확도 퍼센트.
- 모델 유형.
- 대상이 연속형인 경우 모델의 정확도 퍼센트.

레코드 요약

요약표에는 모델에 적합한 레코드 수 및 제외되는 레코드 수가 표시됩니다. 표시된 세부사항에는 포함되고 제외되는 레코드의 수와 퍼센트 뿐만 아니라 가중되지 않은 레코드 수(빈도 가중치를 사용한 경우)가 포함됩니다.

예측변수 중요도

예측변수 중요도 그래프는 막대형 차트로 모델에 있는 상위 10개 입력(예측자)의 중요도를 표시합니다.

차트에 필드가 10개가 넘으면 차트 아래 슬라이더를 사용하여 차트에 포함되는 예측변수의 선택을 변경할 수 있습니다. 슬라이더의 표시기 마크는 고정된 너비이며, 슬라이더의 각 마크는 10개 필드를 나타냅니다. 슬라이더와 함께 표시기 마크를 이동하여 예측변수 중요도로 정렬된 다음 또는 이전 10개 필드를 표시할 수 있습니다.

차트를 두 번 클릭하면 그래프 설정을 편집할 수 있는 별도의 대화 상자를 열 수 있습니다. 예를 들어, 그래프 크기, 사용된 글꼴의 크기와 색상과 같은 항목을 수정할 수 있습니다. 별도의 이 편집 대화 상자를 닫으면 출력 탭에 표시된 차트에 변경이 적용됩니다.

예측되는 도표의 잔차

이 도표를 사용하여 이상값을 식별하거나 비선형성 오차 분산 또는 상수가 아닌 오차 분산을 진단할 수 있습니다. 이상적 도표는 기준선 주변에 무작위로 흩어져 있는 점을 표시합니다.

예측되는 패턴의 경우 선형 예측변수의 예측값에서 표준화 편차 잔차 분포의 평균값은 0이고 범위는 상수입니다. 예측되는 패턴은 0을 통과하는 가로선입니다.

나. GLE 모델 너깃 설정

GLE 모델 너깃의 설정 탭에서 모델 스코어링 중에 원시 성향 및 SQL 생성에 대한 옵션을 지정합니다. 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능합니다.

원시 성향 스코어 계산 플래그 대상만 포함하는 모델의 경우, 목표 필드에 지정된 실제 결과의 우도를 표시하는 원시 성향 스코어를 요청할 수 있습니다. 표준 예측 및 신뢰도 값 외에도 제공됩니다. 수정된 성향 스코어는 사용할 수 없습니다.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성 방법을 지정하십시오.

- **기본값: 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링** 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

(11) Cox 노드

Cox 회귀분석은 시간 대 이벤트 데이터에 대한 예측 모델을 작성합니다. 모델은 예측자 변수의 주어진 값에 대해 특정 시간 t 에 중요 이벤트가 발생했을 확률을 예측하는 생존함수를 생성합니다. 생존 함수의 모양과 예측자의 회귀계수는 관측된 개체에서 추정됩니다. 그런 다음 예측자 변수의 측정값을 가지고 있는 새로운 케이스에 모델을 적용할 수 있습니다. 관측하는 동안 중요 이벤트가 발생하지 않는 중도절단 개체의 정보는 모델 예측에 유용하게 기여함에 유의하십시오.

예. 고객 이탈을 줄이기 위한 일환으로 한 통신 회사는 다른 서비스로 빠르게 전환하는 고객과 연관된 요인을 판별하기 위해 "이탈 시간" 모델링에 관심이 있습니다. 이를 위해 임의의 고객 표본을 선택하고 이들이 고객으로 소모한 시간(여전히 활성 고객이 아닌지 여부) 및 다양한 인구 통계학적 필드를 데이터베이스에서 끌어옵니다.

요구사항. 하나 이상의 입력 필드와 목표 필드가 정확히 하나 필요하며 Cox 노드 내에 생존 시간 필드를 지정해야 합니다. "거짓" 값이 생존을 나타내고 "참" 값은 관심 있는 이벤트가 발생했음을 표시하도록 목표 필드를 코딩해야 합니다. 측정 수준이 *플래그*이고 문자열 또는 정수 저장 공간이 있어야 합니다. (필요에 따라 채움 또는 파생 노드를 사용하여 저장 공간을 변환할 수 있습니다.) *둘 다* 또는 *없음*으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다. 생존 시간은 수치 필드일 수 있습니다.

 **참고:** Cox 회귀 모델을 스코어링할 때 범주 변수의 빈 문자열을 모델 작성의 입력으로 사용하는 경우 오류가 보고됩니다. 빈 문자열을 입력으로 사용하지 마십시오.

날짜 & 시간. 생존 시간을 직접 정의하기 위해 날짜 & 시간 필드를 사용할 수 없습니다. 날짜 & 시간 필드가 있으면 이 필드를 사용해서 연구를 시작한 날짜와 관측 날짜 차이를 기준으로 하여 생존 시간을 포함한 필드를 작성해야 합니다.

Kaplan-Meier 모델 분석. Cox 회귀분석은 입력 필드 없이 수행할 수 있습니다. 이는 Kaplan-Meier 모델 분석에 해당합니다.

① Cox 노드 필드 옵션

생존 시간. 노드를 실행 가능하게 하려면 수치 필드(측정 수준이 *연속인*)을 선택하십시오. 생존 시간은 예상되는 레코드의 수명을 나타냅니다. 예를 들어, 고객 시간을 이탈로 모델링할 경우 이는 고객이 조직과 함께 한 기간을 기록하는 필드입니다. 고객이 가입 또는 이탈한 날짜는 모델에 영향을 미치지 않고 고객이 함께 한 기간만 관련됩니다.

생존 시간은 단위가 없는 기간으로 처리됩니다. 입력 필드가 생존 시간에 일치하는지 확인해야 합니다. 예를 들어, 월별 이탈을 측정하는 조사에서는 연도별 매출이 아닌 월별 매출을 입력으로 사용합니다. 데이터에 기간이 아닌 시작 및 종료 날짜가 있으면 Cox 노드의 기간 업스트림에 이 날짜를 기록해야 합니다.

이 대화 상자에서 나머지 필드는 IBM® SPSS® Modeler 전반에 걸쳐 사용되는 표준 필드입니다. 자세한 정보는 모델링 노드 필드 옵션 주제를 참조하십시오.

② Cox 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

분할 모델 작성. 분할 필드로 지정되는 입력 필드의 각 가능한 값마다 별도의 모델을 작성합니다. 자세한 정보는 분할 모델 작성의 내용을 참조하십시오.

방법. 다음 옵션을 사용하여 모델에 예측자를 입력할 수 있습니다.

- **Enter.** 기본 방법으로, 모델에 직접 모든 항목을 입력합니다. 모델 작성 시 필드 선택은 수행되지 않습니다.
- **단계 선택.** 필드 선택의 단계선택법은 이름에 내포되어 있듯이 단계별로 모델을 작성합니다. 초기 모형은 가능한 가장 단순한 모형으로, 모형의 모형 항목이 없습니다(상수 제외). 각 단계마다 모델에 아직 추가되지 않은 항목을 평가하여 최상의 항목이 모델의 예측력을 상당히 증가시킬 경우 이 항목이 추가됩니다. 또한 현재 모델에 있는 항목은 모델을 크게 손상시키지 않고도 제거할 수 있는지 판별하기 위해 재평가됩니다. 그러한 경우에는 항목이 제거됩니다. 프로세스가 반복되고 다른 항목이 추가 및/또는 제거됩니다. 모델을 향상시키기 위해 추가할 수 있는 더 이상의 항목이 없으며 모델을 손상시키지 않고 제거할 수 있는 더 이상의 항목이 없으면 최종 모델이 생성됩니다.
- **단계별 후진.** 단계별 후진 방법은 본질적으로 단계선택법의 반대 개념입니다. 이 방법에서 초기 모델은 모든 항목을 예측자로 포함합니다. 각 단계에서 모델의 항목이 평가되고 모델을 크게 손상시키지 않고 제거할 수 있는 항목이 제거됩니다. 또한 이전에 제거된 항목은 해당 항목 중 최상의 항목이 모델의 예측력을 크게 높이는지 여부를 판별하기 위해 재평가됩니다. 이 경우 모델로 다시 추가됩니다. 모델을 많이 손상시키지 않고 제거할 수 있는 더 이상의 항목이 없으며 모델을 향상시키기 위해 추가할 수 있는 더 이상의 항목이 없으면 최종 모델이 생성됩니다.

참고: 단계 선택 및 단계별 후진을 포함한 자동 방법은 적응력이 높은 학습 방법이며 학습 데이터의 과적합 경향이 높습니다. 이 방법을 사용하는 경우 새 데이터 또는 파티션 노드를 사용하여 작성된 검증용 검정 표본을 통해 결과로 생성된 모델의 유효성을 확인하는 것이 특히 중요합니다.

그룹. 그룹 필드를 지정하면 노드가 각 필드 범주의 개별 모델을 계산합니다. 문자열이나 정수 저장 공간이 있는 범주형 필드(플래그 또는 명목)일 수 있습니다.

모델 유형. 모델의 항목을 정의하는 두 가지 옵션이 있습니다. **주효과** 모델은 개별적으로 입력 모델만 포함하고 입력 필드 사이의 상호작용(승법 효과)은 검정하지 않습니다. **사용자 정의** 모델은 사용자가 지정한 항목(주효과 및 상호작용)만 포함합니다. 이 옵션을 선택하면 모형 항목 목록을 사용하여 모형에서 항목을 추가하거나 제거합니다.

모형 항목. 사용자 정의 모델을 작성할 때에는 모델의 항목을 명시적으로 지정해야 합니다. 목록에는 모델 항목의 현재 세트가 표시됩니다. 모형 항목 목록의 오른쪽에 있는 단추로 모형 항목을 추가 및 제거할 수 있습니다.

- 모형에 항을 추가하려면 **새 모형 항 추가** 단추를 클릭하십시오. 자세한 정보는 Cox 회귀 모형에 항 추가 주제를 참조하십시오.
- 항을 삭제하려면 원하는 항을 선택하고 **선택한 모형 항 삭제** 단추를 클릭하십시오.

가. Cox 회귀 모형에 항 추가

사용자 정의 모형을 요청할 때 모형 탭에서 새 모형 항 추가 단추를 클릭하여 모형에 항을 추가할 수 있습니다. 항을 지정할 수 있는 새 대화 상자가 열립니다.

추가할 항 유형. 사용 가능한 필드 목록에서 입력 필드 선택에 따라 모델에 항을 추가하는 여러 방법이 있습니다.

- **단일 상호작용.** 선택한 모든 필드의 상호작용을 나타내는 항을 삽입합니다.
- **주효과.** 선택된 각 입력 필드마다 주효과 항(필드 자체)을 하나씩 삽입합니다.
- **모든 2원 효과 상호작용.** 선택한 입력 필드의 가능한 각 쌍에서 이원 상호작용 항(입력 필드의 곱)을 삽입합니다. 예를 들어, 사용 가능한 필드 목록에서 입력 필드 A , B , C 를 선택한 경우 이 방법은 $A * B$, $A * C$, $B * C$ 항을 삽입합니다.
- **모든 3원 효과 상호작용.** 선택한 입력 필드의 가능한 각 조합(한 번에 3개 항 사용)에서 삼원 상호작용 항(입력 필드의 곱)을 삽입합니다. 예를 들어, 사용 가능한 필드 목록에서 입력 필드 A , B , C , D 를 선택한 경우, 이 방법은 $A * B * C$, $A * B * D$, $A * C * D$, $B * C * D$ 항을 삽입합니다.
- **모든 4원 효과 상호작용.** 선택한 입력 필드의 가능한 각 조합(한 번에 4개 항 사용)에서 4원 상호작용 항(입력 필드의 곱)을 삽입합니다. 예를 들어 사용 가능한 필드 목록에서 입력 필드 A , B , C , D , E 를 선택한 경우, 이 방법은 $A * B * C * D$, $A * B * C * E$, $A * B * D * E$, $A * C * D * E$, $B * C * D * E$ 항을 삽입합니다.

사용 가능한 필드. 모형 항을 구성할 때 사용할 사용 가능한 입력 필드를 나열합니다. 목록에 적합하지 않은 입력 필드가 포함될 수 있으므로 모든 모형 항에 입력 필드만 포함되었는지 확인해야 함에 유의하십시오.

미리보기. 선택한 필드 및 위에 선택된 항 유형을 기준으로 하여 **삽입**을 클릭할 때 모델에 추가할 항을 표시합니다.

삽입. 모델에 항을 삽입하고(필드의 현재 선택 및 항 유형을 기준으로 하여) 대화 상자를 닫습니다.

③ Cox 노드 고급 옵션

수렴. 이 옵션으로 모델 수렴에 대한 모수를 제어할 수 있습니다. 모델을 실행할 때 수렴 설정은 여러 다른 모수가 어느 정도 적합한지 확인하기 위해 모수가 반복적으로 실행되는 횟수를 제어

합니다. 모수를 더 자주 시도할 수록 결과에 더 근접합니다(즉, 결과가 수렴됨). 자세한 정보는 Cox 노드 수렴 기준의 내용을 참조하십시오.

출력. 이 옵션을 통해 노드가 작성한 생성된 모델의 고급 출력에 표시될 생존 곡선을 포함하여, 추가 통계량 및 도표를 요청할 수 있습니다. 자세한 정보는 Cox 노드 고급 출력 옵션의 내용을 참조하십시오.

단계. 이 옵션은 단계 선택 추정 방법으로 필드를 추가 및 제거하는 기준을 제어할 수 있습니다. (이 단추는 입력 방법을 선택한 경우 사용할 수 없습니다.) 자세한 정보는 Cox 노드 단계별 기준의 내용을 참조하십시오.

가. Cox 노드 수렴 기준

최대반복계산. 프로시저가 솔루션을 찾는 데 걸리는 시간을 제어하는 모델의 최대반복계산을 지정할 수 있습니다.

로그-우도 수렴. 로그-우도의 상대값 변화가 이 값 미만이면 반복이 중지됩니다. 값이 0인 경우 이 기준은 적용되지 않습니다.

모수 수렴. 모수 추정값의 절대값 또는 상대값 변화가 이 값 미만이면 반복이 중지됩니다. 값이 0인 경우 이 기준은 적용되지 않습니다.

나. Cox 노드 고급 출력 옵션

통계량. 추정값의 상관관계 및 $\exp(B)$ 에 대한 신뢰구간을 포함하여 모델 모수의 통계를 구할 수 있습니다. 이러한 통계는 각 단계마다 또는 마지막 단계에서만 요청할 수 있습니다.

기준선 위험 함수 표시. 공변량의 평균에 기준선 위험 함수 및 누적 생존을 표시할 수 있습니다.

도표

도표는 추정된 모델을 계산하고 결과를 해석하는 데 유용합니다. 생존, 위험 함수, 로그 - 로그, 1 - 생존 함수를 도표로 나타낼 수 있습니다.

- **생존.** 선형 척도로 누적 생존 함수를 표시합니다.
- **위험 함수.** 선형 척도에 누적 위험 함수를 표시합니다.
- **로그 - 로그.** $\ln(-\ln)$ 변환이 추정값에 적용된 후의 누적 생존 추정값을 표시합니다.
- **1 - 생존 함수.** 선형 척도에 1 - 생존 함수를 도표화합니다.

각 값의 선구분 집단변수 도표 표시. 이 옵션은 범주형 필드에만 사용 가능합니다.

도표에 사용할 값. 이 함수는 예측자의 값에 종속되므로 함수 대 시간으로 도표 표시하려면 예측자의 상수 값을 사용해야 합니다. 기본값은 각 예측자의 평균을 상수 값으로 사용하는 것이지만 눈금을 사용하여 도표에 직접 값을 입력할 수 있습니다. 범주형 입력의 경우에는 표시기 코딩이 사용되므로 각 범주마다(마지막 범주 제외) 회귀계수가 있습니다. 따라서 범주형 입력은 표시기 대비에 해당하는 범주의 케이스 비율에 일치하는 각 표시기 대비의 평균 값이 있습니다.

다. Cox 노드 단계별 기준

제거 기준. 보다 강력한 모델에서 **우도비**를 선택합니다. 모델 작성에 필요한 시간을 단축하려면 **Wald**를 선택할 수 있습니다. 조건부 모수 추정값에 기반한 우도비 통계의 확률을 기준으로 하여 제거 검정을 제공하는 추가 옵션 **조건부**가 있습니다.

기준의 유의수준 임계값. 이 옵션으로 각 필드와 연관된 통계 확률(p 값)을 기준으로 하여 선택 기준을 지정할 수 있습니다. 연관된 p 값이 **입력** 값보다 작은 경우에만 모델에 필드가 추가되고 p 값이 **제거** 값보다 큰 경우에만 필드가 제거됩니다. **입력** 값은 **제거** 값보다 작아야 합니다.

④ Cox 노드 설정 옵션

미래의 생존 예측. 하나 이상의 미래 시간을 지정하십시오. 생존 즉, 시간 값별로 예측을 하나씩, 각 시간 값의 레코드마다 터미널 이벤트 발생 없이 각 케이스가 최소 이 기간 동안(지금부터) 생존할지 여부를 예측합니다. 생존은 목표 필드의 "거짓" 값임에 유의하십시오.

- **정규 간격.** 생존 시간 값은 지정된 **시간 간격** 및 **스코어링할 기간** 수에서 생성됩니다. 예를 들어, 각 시간 간격이 2인 3 기간이 요청되면 생존은 미래 시간 2, 4, 6으로 예측됩니다. 모든 레코드가 값과 동시에 평가됩니다.
- **시간 필드.** 선택된 시간 필드의 각 레코드마다 생존 시간이 제공(예측 필드가 하나 생성됨)되므로 각 레코드를 다른 시간에 평가할 수 있습니다.

과거 생존 시간. 이제까지의 레코드 생존 시간을 지정합니다. 예를 들어, 기존 고객의 참여를 필드로 지정합니다. 미래 시간의 생존 우도 스코어링은 과거 생존 시간의 조건부입니다.

참고: 미래와 과거 생존 시간의 값은 모델을 훈련하는 데 사용된 데이터의 생존 시간 범위 내에 있어야 합니다. 시간이 이 범위를 벗어나는 레코드는 널로 스코어링됩니다.

모든 확률 추가. 출력 필드의 각 범주에 대한 확률을 노드에서 처리하는 각 레코드에 추가하는지 여부를 지정합니다. 이 옵션을 선택하지 않으면 예측 범주의 확률만 추가됩니다. 확률은 각 미래 시간마다 계산됩니다.

누적 위험 함수 계산. 누적 위험 값이 각 레코드에 추가되는지 여부를 지정합니다. 누적 위험은 각 미래 시간마다 계산됩니다.

⑤ Cox 모델 너깃

Cox 회귀 모형은 Cox 노드가 추정된 방정식을 나타냅니다. 여기에는 모델에서 캡처한 모든 정보와 모델 구조 및 성능에 대한 정보가 포함되어 있습니다.

생성된 Cox 회귀 모형을 포함한 스트림을 실행하면 노드는 모델의 예측 및 연관된 확률을 포함한 두 개의 새 필드를 추가합니다. 새 필드의 이름은 예측 중인 출력 필드의 이름에서 파생되며 접두문자는 예측 범주의 경우 $\$C$ - 및 연관된 확률의 경우에는 $\$CP$ -이고 접미문자는 미래 시간 간격 수 또는 시간 간격을 정의하는 시간 필드의 이름입니다. 예를 들어, *churn* 이름의 두 개의 미래 시간 구간이 정기적으로 정의된 출력 필드의 경우 새 필드 이름은 $\$C-churn-1$, $\$CP-churn-1$, $\$C-churn-2$, 및 $\$CP-churn-2$ 입니다. 미래 시간이 시간 필드 *tenure*로 정의되는 경우에는 새 필드 이름이 $\$C-churn_tenure$ 및 $\$CP-churn_tenure$ 입니다.

Cox 노드에서 **모든 확률 추가** 설정 옵션을 선택한 경우 각 레코드의 생존 및 실패 확률을 포함하여 두 개의 추가 필드가 각 미래 시간마다 추가됩니다. 이 추가 필드는 출력 필드의 이름을 기반으로 이름이 지정되며, 접두문자는 생존 확률의 경우 $\$CP$ -<거짓 값>- 및 이벤트가 발생한 확률의 경우 $\$CP$ -<참 값>-이고 접미문자는 미래 시간 간격의 수입입니다. 예를 들어, "거짓" 값이 0이고 "참" 값이 1이며 두 개의 미래 시간 구간이 주기적으로 정의된 출력 필드의 경우 새 필드의 이름은 $\$CP-0-1$, $\$CP-1-1$, $\$CP-0-2$, $\$CP-1-2$ 입니다. 미래 시간이 단일 시간 필드 *tenure*로 정의되면 단일 미래 간격이 있으므로 새 필드는 $\$CP-0-1$ 및 $\$CP-1-1$ 입니다.

Cox 노드에서 **누적 위험 함수 계산** 설정 옵션을 선택한 경우에는 각 레코드의 누적 위험 함수를 포함한 추가 필드 하나가 각 미래 시간마다 추가됩니다. 이 추가 필드는 출력 필드의 이름을 기반으로 이름이 지정되며 접두문자는 $\$CH$ -이고 접미문자는 미래 시간 간격의 수 또는 시간 간격을 정의하는 시간 필드의 이름입니다. 예를 들어, 두 개의 미래 시간 구간이 주기적으로 정의된 *churn*이란 출력 필드의 경우 새 필드의 이름은 $\$CH-churn-1$ 및 $\$CH-churn-2$ 입니다. 미래 시간이 시간 필드 *tenure*로 정의되면 새 필드는 $\$CH-churn-1$ 입니다.

가. Cox 회귀분석 출력 설정

SQL 생성을 제외하면, 너깃의 설정 탭은 모델 노드의 설정 탭과 제어가 동일합니다. 너깃 제어의 기본값은 모델 노드에 설정된 값으로 판별됩니다. 자세한 정보는 Cox 노드 설정 옵션의 내용을 참조하십시오.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

나. Cox 회귀분석 고급 출력

Cox 회귀분석의 고급 출력은 생존 곡선을 포함하여 추정된 모델 및 성능에 대한 자세한 정보를 제공합니다. 고급 출력에 포함된 대부분의 정보는 상당히 기술적 정보이며 이 출력을 제대로 해석하려면 광범위한 Cox 회귀분석 지식이 필요합니다.

9) 군집 모델

군집 모델은 유사한 레코드 그룹을 식별하고 레코드에 이들이 속하는 그룹에 따라 레이블을 붙이는 데 초점을 둡니다. 이는 그룹 및 해당 특성에 대한 사전 지식 없이도 수행됩니다. 실제로는 심지어 얼마나 많은 그룹을 찾을지 정확히 알지 못할 수도 있습니다. 이 점이 바로 군집 모델을 다른 머신 학습 기법과 구별합니다. 예측할 모델에 대한 사전정의된 출력 또는 대상 필드가 없습니다. 이 모델은 모델의 분류 성능을 판단할 외부 표준이 없기 때문에 종종 **자율 학습** 모델이라 부르기도 합니다. 이 모델에 대한 **올바른** 또는 **잘못된** 응답이 없습니다. 이들 값은 데이터에서 관심 있는 집단을 캡처하고 이러한 집단에 대한 유용한 설명을 제공하는 기능으로 판별됩니다.

군집방법은 레코드 간 그리고 군집 간의 거리 측정을 기반으로 합니다. 레코드는 동일한 군집에 속한 레코드 사이의 거리를 최소화하려는 방식으로 군집에 할당됩니다.

다음 군집방법이 제공됩니다.



K-평균 노드는 데이터 세트를 고유 그룹(또는 군집)으로 군집화합니다. 이 방법은 고정된 수의 군집을 정의하고 반복적으로 레코드를 군집에 지정하며, 추가 세분화가 더 이상 모델을 향상시킬 수 없을 때까지 군집중심을 조정합니다. 결과를 예상하는 대신 *k*-평균은 자율 학습으로 알려진 프로세스를 사용하여 입력 필드 세트의 패턴을 찾아냅니다.



이단계 노드는 2단계 군집방법을 사용합니다. 첫 번째 단계는 원시 입력 데이터를 관리 가능한 하위 군집 세트로 압축하기 위해 데이터를 통한 단일 전달을 수행합니다. 두 번째 단계는 계층적 군집 방법을 사용하여 하위 군집을 점점 더 큰 군집으로 계속해서 병합하는 것입니다. 이단계는 학습 데이터에 대한 최적 군집 수를 자동으로 평가하는 장점이 있습니다. 혼합 필드 유형과 대형 데이터 세트를 효율적으로 처리할 수 있습니다.



코호넨 노드는 데이터 세트를 고유 그룹으로 군집화하는 데 사용할 수 있는 신경망 유형을 생성합니다. 네트워크가 완전히 숙달되면, 유사 레코드는 출력 맵 가까이 있지만, 다른 레코드는 멀리 떨어져 있을 것입니다. 모델 너지에서 각 단위별로 캡처된 관측값을 살펴 강한 단위를 식별할 수 있습니다. 이것은 적당한 군집 수에 대한 감각을 제공할 것입니다.



Hierarchical Density-Based Spatial Clustering(HDBSCAN)은 자율 학습을 사용하여 데이터 세트의 군집 또는 밀집된 영역을 찾습니다. SPSS® Modeler의 HDBSCAN 노드에는 HDBSCAN 라이브러리의 핵심 기능과 일반적으로 사용되는 매개변수가 표시됩니다. 이 노드는 Python으로 구현되며, 초기에 그룹이 어떤 그룹인지 모를 때 이 노드를 사용하여 데이터 세트를 구별되는 그룹으로 군집화할 수 있습니다.

군집 모델은 종종 후속 분석의 입력으로 사용되는 군집 또는 세그먼트를 작성하는 데 사용됩니다. 일반적인 예로는 마케터가 전체 시장을 동종의 하위 그룹으로 분할하는 데 사용하는 시장 세그먼트가 있습니다. 각 세그먼트에는 목표를 향한 마케팅 노력의 성공에 영향을 미치는 특수 공정특성 변수가 있습니다. 마케팅 전략을 최적화하기 위해 데이터 마이닝을 사용 중이면 일반적으로 적합한 세그먼트를 식별하고 예측 모델에 세그먼트 정보를 사용해서 모델을 상당히 개선할 수 있습니다.

(1) 코호넨 노드

코호넨 네트워크는 **knet** 또는 **자가 조직 맵**이라고도 하며 군집화를 수행하는 신경망의 한 유형입니다. 이 유형의 네트워크는 초기에 그룹이 어떤 그룹인지 모를 때 데이터 세트를 구별되는 그룹으로 군집화하는 데 사용할 수 있습니다. 레코드가 그룹화되므로 그룹 또는 군집 내 레코드는 서로 유사한 경향이 있지만, 다른 그룹의 레코드는 비슷하지 않습니다.

기본 단위는 뉴런이며 두 개의 레이어, **입력층** 및 **출력층**(출력 맵이라고도 함)으로 구성됩니다. 모든 입력 뉴런은 전체 출력 뉴런에 연결되고 연결은 연관된 **강도** 또는 **가중값**이 있습니다. 훈련 중에는 각 단위가 각 레코드에서 "우승"하기 위해 다른 모든 단위와 서로 경쟁합니다.

출력 맵은 단위가 서로 연결되지 않은 뉴런의 2차원 눈금입니다.

입력 데이터는 입력층에 표시되고 값은 출력층으로 전파됩니다. 반응이 가장 강력한 출력 뉴런은 승자라고 하며 해당 입력의 답이 됩니다.

처음에는 모든 가중치가 무작위입니다. 한 단위가 레코드에서 우승하면 가중값(집합적으로 **이웃 항목**이라 부르는 다른 근처의 단위 가중값과 함께)이 해당 레코드의 예측자 값 패턴에 더 일치하도록 조정됩니다. 모든 입력 레코드가 표시되고 이에 따라 가중값이 업데이트됩니다. 이 프로세스는 변경이 매우 적게 될 때까지 여러 번 반복됩니다. 훈련이 진행되면서 눈금 단위의 가중값은 군집의 2차원 "맵"(자가 조직 맵)을 형성하도록 조정됩니다.

네트워크가 완전히 훈련되면 유사한 레코드는 출력 맵에서 서로 가까워야 하는 반면에 아주 상이한 레코드는 멀리 떨어져 있습니다.

IBM® SPSS® Modeler에서 대부분의 학습 방법과는 달리, 코호넨 네트워크는 목표 필드를 사용하지 *않습니다*. 목표 필드가 없는 이 학습 유형은 **자유 학습**이라고 합니다. 결과를 예측하는 대신, 코호넨 넷은 입력 필드 세트에서 패턴을 파악하려고 합니다. 일반적으로 코호넨 넷은 많은 관측(**강력한** 단위)을 요약하는 소수의 단위와 관측에 실제로 대응하지 않는 여러 단위(**취약한** 단위)로 종료됩니다. 강력한 단위(그리고 때때로 눈금에서 이들에 인접한 다른 단위)는 가능한 군집 중심을 나타냅니다.

다른 코호넨 네트워크의 사용은 **차원 축소**에서 찾을 수 있습니다. 2차원 눈금의 공간적 특성을 통해 k 원래 예측자에서, 원래 예측자와 유사 관계를 유지하는 2개의 파생된 기능으로의 맵핑을 제공합니다. 일부 경우에 이는 요인 분석 또는 PCA와 동일한 종류의 혜택을 제공할 수 있습니다.

요구사항. 코호넨 넷을 훈련하려면 역할이 **입력**으로 설정된 하나 이상의 필드가 필요합니다. 역할이 **목표**, **둘 다** 또는 **없음**으로 설정된 필드는 무시됩니다.

강도. 코호넨 네트워크 모델을 작성하기 위해 소속그룹에 데이터는 없어도 됩니다. 찾으려는 여러 그룹을 알지 않아도 됩니다. 코호넨 네트워크는 많은 수의 단위로 시작되고, 훈련이 진행되면 이 단위는 데이터에서 자연 군집 방향으로 이끌립니다. 강력한 단위를 식별하기 위해 모델 너깃에 있는 각 단위에서 캡처한 관측값 수를 살펴볼 수 있습니다. 이를 통해 적절한 군집 수를 파악할 수 있습니다.

① 코호넨 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

기존 모델 훈련 계속. 기본적으로 코호넨 노드를 실행할 때마다 완전히 새로운 네트워크가 작성됩니다. 이 옵션을 선택하면 노드에 의해 성공적으로 생성된 마지막 모델로 계속 훈련합니다.

피드백 그래프 표시. 이 옵션을 선택하면 훈련 중에 2차원 배열의 시각적 표시가 나타납니다. 각 노드의 강도는 색상으로 표시됩니다. 빨간색은 많은 레코드를 얻은 단위(**강력한** 단위)를 뜻하고, 흰색은 얻은 레코드가 거의 없거나 전혀 없는 단위(**취약한** 단위)를 뜻합니다. 모델 작성에 걸린 시간이 비교적 짧으면 피드백이 표시되지 않을 수도 있습니다. 이 기능은 훈련 시간을 늦출 수 있다는 점에 주의하십시오. 훈련 시간을 단축하려면 이 옵션을 선택 취소하십시오.

중지 기준. 기본 중지 기준은 내부 모수에 따라 훈련을 중지합니다. 또한 중지 기준으로 시간을 지정할 수도 있습니다. 훈련할 네트워크에 대한 시간(분 단위)을 입력합니다.

난수 시드 설정. 난수 시드가 설정되지 않은 경우 노드가 실행될 때마다 네트워크 가중값을 초기화하는 데 사용되는 무작위 값 시퀀스가 달라집니다. 이로 인해 노드 설정 및 데이터 값이 정확히 같아도 노드가 다른 실행에 다른 모델을 작성할 수 있습니다. 이 옵션을 선택해서 결과적인 모델을 정확히 재생성할 수 있도록 난수 시드를 특정 값으로 설정할 수 있습니다. 특정 난수 시드는 항상 동일한 시퀀스의 무작위 값을 생성하며 어느 경우든 노드를 실행하면 항상 동일한 모델이 생성됩니다.

참고: 데이터베이스에서 읽은 레코드에서 **난수 시드 설정** 옵션을 사용하는 경우 노드를 실행할 때마다 동일한 결과를 보장하려면 표본 추출 전에 정렬 노드가 필요할 수도 있습니다. 난수 시드는 레코드 순서에 의존하여 관계형 데이터베이스에서는 동일하게 보장되지 않기 때문입니다.

참고: 모델에서 명목(설정된) 필드를 포함하려고 하지만, 모델 작성에 메모리 문제가 있거나 모델 작성 시간이 오래 걸리면 큰 세트 필드를 기록하여 값의 수를 줄이거나 큰 세트에 대한 프록시로 값이 더 적은 다른 필드 사용을 고려하십시오. 예를 들어, 개별 제품에 대한 값을 포함하는 *product_id* 필드에 문제점이 있는 경우 모델에서 이를 제거하고 대신 덜 자세한 *product_category* 필드를 추가하는 방법을 고려할 수 있습니다.

최적화. 특정 요구를 기준으로 한 모델 작성 중에 성능을 늘리려면 이 옵션을 선택하십시오.

- 성능을 개선하기 위해 알고리즘이 디스크 스피어링을 사용하지 않게 하려면 **속도**를 선택하십시오.
- 적합한 시기에 속도가 느려지더라도 알고리즘이 디스크 스피어링을 사용하게 하려면 **메모리**를 선택하십시오. 이 옵션은 기본적으로 선택됩니다.

참고: 분산 모드에서 실행할 때에는 options.cfg 파일에 지정된 관리자 옵션이 이 설정을 대체할 수 있습니다.

군집 레이블 첨부. 기본적으로 새 모델에서 선택되지만, IBM® SPSS® Modeler의 이전 버전에서 로드된 모델에서는 선택 취소되어 있습니다. 이 옵션은 K-평균 및 이단계 노드 모두에서 작성된 동일한 유형의 단일 범주형 스코어 필드를 작성합니다. 이 문자열 필드는 서로 다른 모델 유형에 대한 순위화 측도를 계산할 때 자동 군집 노드에서 사용됩니다. 자세한 정보는 자동 군집 노드의 내용을 참조하십시오.

② 코호넨 노드 고급 옵션

코호넨 네트워크에 대한 자세한 지식을 가진 사용자는 고급 옵션을 통해 학습 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

너비 및 길이 각 차원과 함께 출력 단위의 수로 2차원 출력 맵의 크기(너비와 길이)를 지정합니다.

학습률 감소 선형 또는 지수형 학습률 감소를 선택합니다. **학습률**은 시간에 따라 감소하는 가중치 요인으로, 네트워크는 데이터의 대규모 기능을 인코딩하는 작업부터 시작하여 점차적으로 보다 미세한 수준의 세부사항에 초점을 맞출 수 있습니다.

1단계 및 2단계: 코호넨 넷 학습은 2개 단계로 분할됩니다. 1단계는 대략적인 추정 단계로, 데이터에서 전체 패턴을 캡처하는 데 사용됩니다. 2단계는 조정 단계로, 데이터의 보다 미세한 기능을 모델링하도록 맵을 조정하는 데 사용됩니다. 각 단계에는 세 개의 모수가 있습니다.

- **이웃** 이웃의 시작 크기(반경)를 설정합니다. 이 옵션은 학습 중 획득한 단위와 함께 업데이트되는 "근접" 단위 수를 판별합니다. 1단계 중에 이웃 크기는 *1단계 이웃*으로 시작하고 (*2단계 이웃 + 1*)로 감소합니다. 2단계 중에 이웃 크기는 *2단계 이웃*부터 시작하여 1.0으로 감소합니다. *1단계 이웃*은 *2단계 이웃*보다 커야 합니다.
- **초기 에타** 학습률 **에타**의 시작값을 설정합니다. 1단계 중에 에타는 *1단계 초기 에타*로 시작하고 *2단계 초기 에타*로 감소합니다. 2단계 중에 에타는 *2단계 초기 에타*로 시작하고 0으로 감소합니다. *1단계 초기 에타*는 *2단계 초기 에타*보다 커야 합니다.
- **순환.** 각 학습 단계에서 순환 수를 설정합니다. 각 단계는 전체 데이터에서 지정된 패스 수만큼 계속됩니다.

(2) 코호넨 모델 너깃

코호넨 모델 너깃은 학습된 코호넨 네트워크에서 캡처한 모든 정보와 네트워크 아키텍처에 대한 정보를 포함합니다.

코호넨 모델 너깃을 포함하는 스트림을 실행하는 경우 노드는 해당 레코드에 가장 강력하게 반응하는 코호넨 출력 눈금에서 단위의 X 및 Y 좌표를 포함하는 두 개의 새 필드를 추가합니다. 새 필드 이름은 모델 이름에서 파생됩니다($\$KX$ - 및 $\$KY$ -의 접두문자가 추가됨). 예를 들어, 모델 이름이 *Kohonen*인 경우 새 필드 이름은 $\$KX$ -*Kohonen* 및 $\$KY$ -*Kohonen*으로 지정됩니다.

코호넨 넷의 인코딩에 대해 더 잘 이해하려면 모델 너깃 브라우저에서 모델 탭을 클릭하십시오. 그러면 군집 뷰어를 표시하고, 여기에서 군집, 필드, 중요도 수준의 그래픽 표시를 제공합니다. 자세한 정보는 군집 뷰어 - 모델 탭의 내용을 참조하십시오.

눈금으로 군집을 시각화하려는 경우 plot 노드를 사용해 $\$KX$ - 및 $\$KY$ - 필드를 구성하여 코호넨 넷의 결과를 볼 수 있습니다. (각 단위의 레코드가 서로 상위에서 구성되지 않도록 방지하려면 plot 노드에서 **X-변동** 및 **Y-변동**을 선택해야 합니다.) 도표에서 코호넨 넷이 데이터 군집을 작성하는 방법을 연구하기 위해 기호 필드도 오버레이할 수 있습니다.

코호넨 네트워크에 대한 통찰력을 얻기 위한 또 다른 강력한 기법은 네트워크에서 찾은 군집을 구별하는 특성을 검색하기 위해 규칙 귀납을 사용하는 것입니다.

모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

① 코호넨 모델 요약

코호넨 모델 너깃의 요약 탭에서는 네트워크의 설계 또는 토폴로지에 대한 정보를 표시합니다. 2차원 코호넨 기능 맵(출력 레이어)의 길이 및 너비는 $\$KX$ - *model_name* 및 $\$KY$ - *model_name*으로 표시됩니다. 입력과 출력 레이어의 경우 해당 레이어에 있는 단위 수가 나열됩니다.

(3) K-평균 노드

K-평균 노드는 **군집분석** 방법을 제공합니다. 초기에 그룹이 어떤 그룹인지 모를 때 데이터 세트를 구별되는 그룹으로 군집화하는 데 사용할 수 있습니다. IBM® SPSS® Modeler에서 대부분의 학습 방법과는 달리, K-평균 모델은 목표 필드를 사용하지 않습니다. 목표 필드가 없는 이 학습 유형은 **자율 학습**이라고 합니다. 결과를 예측하는 대신, K-평균은 입력 필드 세트에서 패턴을 파악하려고 합니다. 레코드가 그룹화되므로 그룹 또는 군집 내 레코드는 서로 유사한 경향이 있지만, 다른 그룹의 레코드는 비슷하지 않습니다.

K-평균은 데이터에서 파생된 시작 군집 중심의 세트를 정의하여 작동합니다. 그런 다음, 레코드의 입력 필드 값에 기반하여 가장 유사한 군집에 각 레코드를 지정합니다. 모든 케이스가 지정된 후에 군집 중심은 각 군집에 지정된 새 레코드 세트를 반영하도록 업데이트됩니다. 그러면 다른 군집에 재지정해야 하는지 여부를 확인하기 위해 레코드를 다시 확인하고, 최대 반복 수에 도달할 때까지 레코드 지정/군집 반복 프로세스를 계속합니다. 그렇지 않으면 한 반복과 다음 실패 사이의 변경이 지정된 임계값을 초과하지 않습니다.

참고: 결과로 생성된 모델은 훈련 데이터의 순서에서 특정 범위에 따라 달라집니다. 데이터를 다시 정렬하고 모델을 재작성할 경우 최종 군집 모델이 달라질 수 있습니다.

요구사항. K-평균 모델을 훈련하려면 역할이 입력으로 설정된 하나 이상의 필드가 필요합니다. 역할이 **출력**, **둘 다** 또는 **없음**으로 설정된 필드는 무시됩니다.

강도. K-평균 모델을 작성하기 위해 소속그룹에 데이터는 없어도 됩니다. K-평균 모델은 대형 데이터 세트 군집을 위한 가장 빠른 방법이기도 합니다.

① K-평균 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

지정된 군집 수. 생성할 군집 수를 지정합니다. 기본값은 5입니다.

거리 필드 생성. 이 옵션을 선택하면 모델 너깅은 지정된 군집의 가운데에서 각 레코드까지의 거리를 포함하는 필드를 포함합니다.

군집 레이블. 생성된 소속군집 필드에서 값의 형식을 지정합니다. 소속군집은 지정된 레이블 접두문자(예: "Cluster 1", "Cluster 2" 등)를 포함하는 문자열 또는 숫자로 표시할 수 있습니다.

참고: 모델에서 명목(설정된) 필드를 포함하려고 하지만, 모델 작성에 메모리 문제가 있거나 모델 작성 시간이 오래 걸리면 큰 세트 필드를 기록하여 값의 수를 줄이거나 큰 세트에 대한 프록시로 값이 더 적은 다른 필드 사용을 고려하십시오. 예를 들어, 개별 제품에 대한 값을 포함하는 *product_id* 필드에 문제점이 있는 경우 모델에서 이를 제거하고 대신 덜 자세한 *product_category* 필드를 추가하는 방법을 고려할 수 있습니다.

최적화. 특정 요구를 기준으로 한 모델 작성 중에 성능을 늘리려면 이 옵션을 선택하십시오.

- 성능을 개선하기 위해 알고리즘이 디스크 스페이킹을 사용하지 않게 하려면 **속도**를 선택하십시오.
- 적합한 시기에 속도가 느려지더라도 알고리즘이 디스크 스페이킹을 사용하게 하려면 **메모리**를 선택하십시오. 이 옵션은 기본적으로 선택됩니다.

참고: 분산 모드에서 실행할 때에는 options.cfg 파일에 지정된 관리자 옵션이 이 설정을 대체할 수 있습니다.

② K-평균 노드 고급 옵션

k 평균 군집에 대한 자세한 지식을 가진 사용자는 고급 옵션을 통해 훈련 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

중지 기준. 모델 훈련에 사용할 중지 기준을 지정합니다. 기존 중지 기준은 20회 반복 또는 변경 < 0.000001 중 먼저 나타나는 조건입니다. **사용자 정의**를 선택하여 고유한 중지 기준을 지정합니다.

- **최대반복계산.** 이 옵션을 사용하면 지정된 반복 수 이후에 모델 훈련을 중지할 수 있습니다.
- **공차 변경.** 이 옵션을 사용하면 반복에서 군집중심의 가장 큰 변화량이 지정된 수준 미만인 경우 모델 훈련을 중지할 수 있습니다.

세트의 인코딩 값. 숫자 필드 그룹으로 세트 필드를 기록하는 데 사용할 값(0에서 1.0 사이)을 지정합니다. 기본값은 0.5의 제곱근(약 0.707107)으로, 기록된 플래그 필드의 적절한 가중치를 제공합니다. 값이 1.0에 가까울수록 숫자 필드보다 세트 필드 가중치가 높아집니다.

(4) K-평균 모델 너깃

K-평균 모델 너깃은 훈련 데이터 및 추정 프로세스에 대한 정보와 함께, 군집 모델에서 캡처한 모든 정보를 포함합니다.

K-평균 모델링 노드를 포함하는 스트림을 실행하는 경우 노드는 해당 레코드의 지정된 군집 중심과의 거리 및 소속군집을 포함하는 새 2개 필드를 추가합니다. 새 필드 이름은 모델 이름(접두문자가 소속군집에서는 $\$KM$ -이고 군집 중심과의 거리에서는 $\$KMD$ -임)에서 파생됩니다. 예를 들어, 모델 이름이 *Kmeans*인 경우 새 필드 이름은 $\$KM-Kmeans$ 및 $\$KMD-Kmeans$ 로 지정됩니다.

K-평균 모델에 대한 통찰력을 얻기 위한 강력한 방법은 모델에서 찾은 군집을 구별하는 특성을 검색하기 위해 규칙 귀납을 사용하는 것입니다. 모델 너깃 브라우저에서 모델 탭을 클릭하여 군집, 필드, 중요도 수준에 대한 그래픽 표시를 제공하는 군집 뷰어를 표시할 수도 있습니다. 자세한 정보는 군집 뷰어 - 모델 탭의 내용을 참조하십시오.

모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

① K-평균 모델 요약

K-평균 모델 너깃의 요약 탭은 훈련 데이터, 추정 프로세스, 모델에서 정의하는 군집에 대한 정보를 포함합니다. 군집 수와 반복계산과정도 표시됩니다. 이 모델링 노드에 첨부된 분석 노드를 실행하면 해당 분석의 정보도 이 섹션에 표시됩니다.

(5) 이단계 군집 노드

이단계 군집 노드는 **군집분석** 양식을 제공합니다. 초기에 그룹이 어떤 그룹인지 모를 때 데이터 세트를 구별되는 그룹으로 군집화하는 데 사용할 수 있습니다. 코호넨 노드 및 K-평균 노드의 경우 이단계 군집 모델이 목표 필드를 사용하지 않습니다. 이단계 군집은 결과를 예측하려 시도하지 않고 입력 필드 세트의 패턴을 밝히려 시도합니다. 레코드가 그룹화되므로 그룹 또는 군집 내 레코드는 서로 유사한 경향이 있지만, 다른 그룹의 레코드는 비슷하지 않습니다.

이단계 군집은 이단계 군집방법입니다. 첫 번째 단계는 데이터를 한 번 전달하며 이 과정에서 원시 입력 데이터가 관리 가능한 부군집 세트로 압축됩니다. 두 번째 단계는 계층적 군집방법을 사용하여 데이터를 또 한번 전달할 필요 없이 부군집을 점점 더 큰 군집으로 병합합니다. 계층적 군집은 미리 군집 수를 선택하지 않아도 되는 장점이 있습니다. 많은 계층적 군집 방법은 개별 레코드로 시작해서 군집을 시작하고 더 큰 군집을 생성하기 위해 반복적으로 군집을 병합합니다. 이러한 접근법은 종종 많은 양의 데이터로 실패하지만 이단계의 초기 사전 군집화는 심지어 큰 데이터 세트의 경우에도 계층적 군집을 빠르게 작성합니다.

참고: 결과적인 모델은 어느 정도까지는 훈련 데이터의 순서에 종속됩니다. 데이터를 다시 정렬하고 모델을 재작성할 경우 최종 군집 모델이 달라질 수 있습니다.

요구사항. 이단계 군집 모델을 훈련하려면 역할이 입력으로 설정된 하나 이상의 필드가 필요합니다. 역할이 **목표**, **둘 다** 또는 **없음**으로 설정된 필드는 무시됩니다. 이단계 군집 알고리즘은 결측값을 핸들하지 않습니다. 모델을 작성할 때 입력 필드가 공백인 레코드는 무시됩니다.

강도. 이단계 군집은 혼합 필드 유형을 핸들할 수 있으며 큰 데이터 세트를 효율적으로 핸들할 수 있습니다. 여러 군집 솔루션을 검정하여 최상의 솔루션을 선택하는 기능이 있으므로 처음에 요청할 군집 수를 알고 있지 않아도 됩니다. **이상치** 또는 결과를 오염시킬 수 있는 매우 비정상적인 케이스를 자동으로 제외하도록 이단계 군집을 설정할 수 있습니다.

IBM® SPSS® Modeler에는 두 가지 다른 버전의 TwoStep 군집 노드가 있습니다.

- **TwoStep 군집**은 IBM SPSS Modeler Server에서 실행하는 일반적인 노드입니다.
- **TwoStep-AS 군집**은 IBM SPSS Analytic Server에 연결되었을 때 실행할 수 있습니다.

① TwoStep 군집 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

숫자 필드 표준화. 기본적으로, TwoStep은 평균이 0이고 분산이 1인 동일 척도로 모든 숫자 입력 필드를 표준화합니다. 숫자 필드에 대한 원래 척도를 유지하려면 이 옵션을 선택 취소하십시오. 기호 필드는 영향을 받지 않습니다.

이상값 제외. 이 옵션을 선택하면, 실질적 군집에 적합한 것으로 보이지 않는 레코드는 자동으로 분석에서 제외됩니다. 그러면 이와 같은 케이스로 결과가 왜곡되는 일이 없어집니다.

이상값 발견은 사전군집 단계에서 발생합니다. 이 옵션이 선택될 때, 다른 하위 군집에 상대적으로 적은 레코드를 가지고 있는 하위 군집은 잠재된 이상값으로 간주되어, 하위 군집 트리가 해당 레코드를 제외하고 다시 작성됩니다. 하위 군집이 잠재된 이상값을 포함하고 있는 것으로 고려되는 크기는 퍼센트 옵션으로 제어됩니다. 잠재된 이상값 레코드 중 일부는 새 하위 군집 프로파일에 대해 충분히 유사한 경우 다시 작성된 하위 군집에 추가될 수 있습니다. 병합할 수 없는 잠재된 이상값의 나머지는 이상값으로 간주되어 "잡음" 군집에 추가되고 계층적 군집 단계에서 제외됩니다.

이상값 처리를 사용하는 TwoStep 모델을 사용하여 데이터를 스코어링할 때, 가장 근접한 실질적 군집에서 특정 임계값 거리(로그-우도 기반)보다 긴 새 케이스는 이상값으로 간주되고 이름이 -1인 "잡음" 군집에 지정됩니다.

군집 레이블. 생성된 소속군집 필드에 대한 형식을 지정하십시오. 소속군집은 지정된 레이블 접두문자가 있는 문자열(예: "Cluster 1", "Cluster 2" 등)이나 숫자로 표시할 수 있습니다.

자동으로 군집 수 계산. TwoStep 군집은 학습 데이터에 대한 최적 군집 수를 선택하기 위해 매우 빠르게 큰 군집 수 솔루션을 분석할 수 있습니다. 최대 및 최소 군집 수를 설정하여 시도할 해법범위를 지정하십시오.

군집 수 지정. 모델에 포함할 군집 수를 알고 있는 경우, 이 옵션을 선택하고 군집 수를 입력하십시오.

거리 척도. 이 선택에서는 두 군집 간 유사성이 계산되는 방식이 결정됩니다.

- **로그-우도.** 우도 척도는 변수에 확률 분포를 둡니다. 연속형 변수는 정규 분포로, 범주형 변수는 다항분포로 가정됩니다. 모든 변수를 독립변수로 가정합니다.
- **유클리디안.** 유클리디안 척도는 두 군집 사이의 "직선" 거리입니다. 모든 변수가 연속형 변수인 경우에만 이 옵션을 사용할 수 있습니다.

군집 기준. 이 기능을 선택하면 자동 군집 알고리즘이 군집 수를 결정하는 방식을 결정할 수 있습니다. Bayesian 정보 기준(BIC) 또는 Akaike 정보 기준(AIC)을 지정할 수 있습니다.

(6) TwoStep 군집 모델 너깃

TwoStep 군집 모델 너깃은 군집 모델이 캡처한 모든 정보 외에 학습 데이터 및 추정 프로세스에 대한 정보를 포함합니다.

TwoStep 군집 모델 너깃을 포함한 스트림을 실행하면 노드가 이 레코드의 소속군집이 포함된 새 필드를 추가합니다. 새 필드 이름은 모델 이름에서 파생되고 접두문자 $\$T$ -가 붙습니다. 예를 들어, 모델 이름이 *TwoStep*이면 새 필드 이름은 $\$T$ -*TwoStep*입니다.

TwoStep 모델을 통찰하는 강력한 기술은 규칙 귀납을 사용하여 모델이 찾은 군집을 구별하는 특성을 발견하는 것입니다. 모델 너깃 브라우저에서 모델 탭을 클릭하여 군집, 필드, 중요도 수준에 대한 그래픽 표시를 제공하는 군집 뷰어를 표시할 수도 있습니다. 자세한 정보는 군집 뷰어 - 모델 탭의 내용을 참조하십시오.

모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

① 이단계 모델 요약

이단계 군집 모델 너깃의 요약 탭은 훈련 데이터, 추정 프로세스 및 사용되는 작성 설정에 대한 정보와 함께, 발견된 군집 수를 표시합니다.

자세한 정보는 모델 너깃 찾아보기 주제를 참조하십시오.

(7) TwoStep-AS 군집 노드

IBM® SPSS® Modeler에는 두 가지 다른 버전의 TwoStep 군집 노드가 있습니다.

- **TwoStep 군집**은 IBM SPSS Modeler Server에서 실행하는 일반적인 노드입니다.
- **TwoStep-AS 군집**은 IBM SPSS Analytic Server에 연결되었을 때 실행할 수 있습니다.

① Twostep-AS 군집분석

TwoStep 군집은 명확하지 않은 데이터 세트 안에서 자연적 집단(또는 군집)을 드러내도록 설계된 탐색 도구입니다. 이 프로시저에 사용되는 알고리즘에는 전형적인 군집 기법과 차별화되는 몇 가지의 바람직한 기능이 있습니다.

- **범주형 및 연속형 변수의 처리.** 변수를 독립변수로 가정하여 범주형 변수 및 연속형 변수를 결합 다항 정규 분포로 표시할 수 있습니다.

- **군집 수 자동 선택.** 군집 솔루션별로 모델 선택 기준 값을 비교하여, 프로시저가 최적의 군집 수를 자동으로 결정할 수 있습니다.
- **확장성.** TwoStep 알고리즘은 레코드를 요약하는 군집 기능(CF) 트리를 구성하여 큰 데이터 파일을 분석할 수 있습니다.

예를 들어, 소매 및 소비 제품 회사는 해당 고객의 구매 버릇, 성, 나이, 소득 수준 및 기타 다른 속성을 설명하는 정보에 정기적으로 군집 기법을 적용합니다. 이 회사는 판매를 증가시키고 브랜드 로열티를 구축하기 위해 해당 마케팅 및 제품 개발 전략을 각 고객 그룹에 맞게 조정합니다.

가. 필드 탭 (Twostep-AS 군집)

필드 탭은 분석에 사용되는 필드를 지정합니다.

사전 정의된 역할 사용. 정의된 입력 역할을 가지고 있는 모든 필드가 선택됩니다.

사용자 정의 필드 할당 사용. 해당되는 정의된 역할 할당에 상관없이 필드를 추가하고 제거하십시오. 임의 역할을 가지고 있는 필드를 선택하고 **예측변수(입력)** 목록의 내부 또는 외부로 이동할 수 있습니다.

나. 기본 (Twostep-AS 군집)

군집 수

자동 결정

프로시저는 지정된 범위 내에서 최상의 군집 수를 판별합니다. **최소**는 1보다 커야 합니다. 이는 기본 선택입니다.

고정 수 지정

프로시저는 지정된 군집 수를 생성합니다. 수는 1보다 커야 합니다.

군집 기준

이 선택사항은 자동 군집 알고리즘이 군집 수를 결정하는 방식을 제어합니다.

베이지안 정보 기준(BIC)

-2 로그 우도에 기반한 모형을 선택 및 비교하기 위한 척도입니다. 값이 작을수록 모형이 우수함을 나타냅니다. BIC도 초과 모수화된 모형(예: 입력이 많은 복잡한 모형)에 "페널티를 부여"하지만 AIC보다 더 엄격하게 부여합니다.

Akaike 정보 기준(AIC)

-2 로그 우도에 기반한 모형을 선택 및 비교하기 위한 척도입니다. 값이 작을수록 모형이 우수함을 나타냅니다. AIC는 초과 모수화된 모형(예: 입력이 많은 복잡한 모형)에 "페널티를 부여합니다".

자동 군집방법

자동 결정을 선택하는 경우, 군집 수를 자동으로 판별하기 위해 사용되는 다음 군집방법에서 선택하십시오.

군집 기준 설정 사용

정보 기준 수렴은 두 개의 현재 군집 솔루션과 첫 번째 군집 솔루션에 해당되는 정보 기준의 비율입니다. 사용되는 기준은 그룹 기준 그룹에서 선택되는 기준입니다.

거리 점프

거리 점프는 두 개의 연속 군집 솔루션에 해당하는 거리의 비율입니다.

최대값

두 번째 점프에 해당하는 군집 수를 생성하기 위해 정보 기준 수렴 방법과 거리 점프 방법의 결과를 결합합니다.

최소값

첫 번째 점프에 해당하는 군집 수를 생성하기 위해 정보 기준 수렴 방법과 거리 점프 방법의 결과를 결합합니다.

기능 중요도 방법

기능 중요도 방법은 기능(필드)이 군집 솔루션에서 얼마나 중요한 지를 판별합니다. 출력에는 전체 기능 중요도와 각 군집에서 각 기능 필드의 중요도에 대한 정보가 포함됩니다. 최소 임계값과 일치하지 않는 기능은 제외됩니다.

군집 기준 설정 사용.

군집 기준 그룹에서 선택되는 기준을 기반으로, 기본 방법입니다.

효과 크기

기능 중요도는 유의수준 값 대신에 효과 크기를 기반으로 합니다.

다. 기능 트리 기준(Twostep-AS 군집)

이 설정은 군집 기능 트리가 작성되는 방법을 결정합니다. 군집 기능 트리를 작성하고 레코드를 요약하면, 이단계 알고리즘이 큰 데이터 파일을 분석할 수 있습니다. 다시 말하면, 이단계 군집에서는 군집을 작성하기 위해 군집 기능 트리를 사용하여, 많은 케이스를 처리할 수 있도록 합니다.

거리 측도

이 선택에서는 두 군집 간 유사성이 계산되는 방식이 결정됩니다.

로그 우도

우도 측도는 확률 분포를 필드에 놓습니다. 연속형 필드는 정규 분포로, 범주형 변수는 다항 분포로 간주됩니다. 모든 필드는 독립변수로 간주됩니다.

유클리디안

유클리디안 측도는 두 군집 사이의 "직선" 거리입니다. 유클리드 제곱값 측도 및 Ward 방법은 군집 사이의 유사성을 계산하기 위해 사용됩니다. 모든 필드가 연속형인 경우에만 사용할 수 있습니다.

이상치 군집

이상치 군집 포함

보통 군집에서 이상치인 케이스에 대한 군집을 포함합니다. 이 옵션을 선택하지 않으면 모든 케이스가 보통 군집에 포함됩니다.

기능 트리 리프 내의 케이스 수가 보다 작음.

기능 트리 리프의 케이스 수가 지정된 값보다 적을 경우, 리프는 이상치로 간주됩니다. 값은 1보다 큰 정수여야 합니다. 이 값을 변경하는 경우, 더 높은 값은 결과적으로 기타 이상치 군집이 될 수 있습니다.

이상치의 위쪽 퍼센트.

군집 모델이 작성될 때, 이상치는 이상치 강도에 의해 순위가 매겨집니다. 이상치의 상위 퍼센트에 있어야 하는 이상치 강도는 케이스가 이상치로 분류되는지 여부를 판별하기 위한 임계값으로 사용됩니다. 값이 높을 수록 더 많은 케이스가 이상치로 분류됨을 의미합니다. 값은 1 - 100 사이여야 합니다.

추가 설정

초기 거리 변화 임계값

군집 기능 트리 증가에 사용되는 초기 임계값. 트리 리프에 리프를 삽입하여 이 임계값보다 작은 기밀도(tightness)이 생성되는 경우, 리프는 분할되지 않습니다. 기밀도가 이 임계값을 초과하면 리프는 분할됩니다.

리프 노드 최대 분기

리프 노드가 가질 수 있는 최대 하위 노드 수.

비리프 노드 최대 분기

비리프 노드가 가질 수 있는 최대 하위 노드 수.

최대 트리 깊이

군집 트리가 가질 수 있는 최대 수준 수.

측정 수준에서 가중 설정 조정

연속형 필드에 대한 가중치를 증가시켜서 범주형 필드의 영향력을 줄입니다. 이 값은 범주형 필드에 대한 가중치 감소에 대한 분모를 나타냅니다. 따라서, 예를 들어 6 기본값은 범주형 필드에 1/6의 가중치를 부여합니다.

메모리 할당

군집 알고리즘이 사용하는 최대 메모리 양(MB). 프로시저가 이 최대값을 초과하면, 메모리에 맞지 않는 정보를 저장하기 위해 디스크를 사용합니다.

지연된 분할

군집 가능 트리의 재작성 지연. 군집 알고리즘은 새 케이스를 평가하는 만큼 여러 번 군집 가능 트리를 다시 작성합니다. 이 옵션은 작업을 지연하고 트리가 다시 작성되는 횟수를 줄여서 성능을 개선할 수 있습니다.

라. 표준화

군집 알고리즘은 표준화된 연속형 필드로 작동합니다. 기본적으로 모든 연속형 필드는 표준화됩니다. 시간 및 계산 노력을 절약하기 위해, 이미 표준화된 연속형 필드를 **표준화하지 않음** 목록으로 이동할 수 있습니다.

마. 필드선택

필드선택 화면에서, 필드가 제외되는 시기를 결정하는 규칙을 설정할 수 있습니다. 예를 들어, 다양한 결측값을 가지고 있는 필드를 제외시킬 수 있습니다.

필드 제외 규칙

결측값의 퍼센트가 보다 큼.

지정된 값보다 큰 결측값의 퍼센트를 가지고 있는 필드는 분석에서 제외됩니다. 값은 0보다 크고 100보다 작은 양수여야 합니다.

범주형 필드의 범주의 수가 보다 큼.

지정된 범주 수보다 많은 범주를 가지고 있는 범주형 필드는 분석에서 제외됩니다. 값은 1보다 큰 양수여야 합니다.

단일값에 대한 추세가 있는 필드

연속형 필드에 대한 변동계수가 보다 작음.

지정된 값보다 작은 변동계수의 연속형 필드가 분석에서 제외됩니다. 변동계수는 평균에 대한 표준 편차의 비율입니다. 값이 낮을수록 값에서 변동이 낮을 수 있습니다. 값은 0 - 1 사이여야 합니다.

범주형 필드에 대한 단일 범주 내의 케이스 퍼센트가 보다 큼.

단일 범주에서 지정된 값보다 큰 케이스 퍼센트의 범주형 필드가 분석에서 제외됩니다. 값은 0보다 크고 100보다 작아야 합니다.

적응형 필드선택

이 옵션은 가장 중요하지 않은 필드를 찾아서 제거하기 위해 추가 데이터 전달을 실행합니다.

바. 모델 출력

모델 작성 요약

모델 지정 사항

모델 지정 사항의 요약, 최종 모델에 있는 군집 수 및 최종 모델에 포함된 입력(필드).

레코드 요약

모델에서 포함되고 제외되는 레코드(케이스) 수 및 퍼센트.

제외된 입력

최종 모델에 포함되지 않은 필드의 경우, 필드가 제외된 원인.

평가

모델 품질

각 군집에 대한 중요 및 적합성과 전체 모델 적합도의 테이블.

기능 중요도 막대형 차트

모든 군집에 걸친 기능(필드) 중요도의 막대형 차트. 차트에서 막대가 긴 기능(필드)이 막대가 짧은 필드보다 중요합니다. 또한 중요도 내림차순으로 정렬됩니다(맨 위에 있는 막대가 가장 중요합니다).

기능 중요도 단어 클라우드

모든 군집에 걸친 기능(필드) 중요도의 단어 클라우드. 텍스트가 많은 기능(필드)이 텍스트가 작은 기능(필드)보다 중요합니다.

이상치 군집

이상치를 포함하지 않도록 선택한 경우 이 옵션은 사용 안함으로 설정됩니다.

대화형 테이블 및 차트

이상치 강도와 보통 군집에 대한 이상치 군집의 상대 유사성의 테이블 및 차트. 테이블에서 다른 행을 선택하면 차트에서 다른 이상치 군집에 대한 정보를 표시합니다.

피벗 표

이상치 강도와 보통 군집에 대한 이상치 군집의 상대 유사성의 테이블. 이 테이블에는 대화형 화면과 동일한 정보가 포함됩니다. 이 테이블은 테이블 피벗 및 편집에 대한 모든 표준 기능을 지원합니다.

최대 수

출력에 표시할 최대 이상치 수. 20보다 큰 이상치 군집이 있는 경우, 피벗 표가 대신 표시됩니다.

설명

교차 군집 기능 중요도 프로파일

대화형 테이블 및 차트

군집 솔루션에서 사용되는 각 입력(필드)에 대한 기능 중요도 및 군집 중심의 테이블 및 차트. 테이블에서 다른 행을 선택하면 다른 차트가 표시됩니다. 범주형 필드의 경우 막대형 차트가 표시됩니다. 연속형 필드의 경우, 평균과 표준편차의 차트가 표시됩니다.

피벗 표.

각 입력(필드)에 대한 기능 중요도 및 군집 중심의 테이블. 이 테이블에는 대화형 화면과 동일한 정보가 포함됩니다. 이 테이블은 테이블 피벗 및 편집에 대한 모든 표준 기능을 지원합니다.

군집 내 기능 중요도

각 군집에 대한, 각 입력(필드)에 대한 군집 중심 및 기능 중요도. 각 군집에 대해 별도의 테이블이 있습니다.

군집 거리

군집 사이의 거리를 표시하는 패널 차트. 각 군집에 대해 별도의 패널이 있습니다.

군집 레이블

Text

각 군집에 대한 레이블은 접두문자에 대해 지정된 값이며, 뒤에 순차 번호가 있습니다.

번호

각 군집에 대한 레이블은 순차 번호입니다.

사. 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

(8) TwoStep-AS 군집 모델 너깃

TwoStep-AS 모델 너깃은 출력 뷰어의 모델 탭에서 모델의 세부사항을 표시합니다. 뷰어 사용에 대한 자세한 정보는 출력 작업의 내용을 참조하십시오.

TwoStep-AS 군집 모델 너깃은 군집 모델이 캡처한 모든 정보 외에 학습 데이터 및 추정 프로세스에 대한 정보를 포함합니다.

TwoStep-AS 군집 모델 너깃을 포함한 스트림을 실행하면 노드가 이 레코드의 소속군집이 포함된 새 필드를 추가합니다. 새 필드 이름은 모델 이름에서 파생되고 접두문자 **\$AS**가 붙습니다. 예를 들어, 모델 이름이 TwoStep이면 새 필드 이름은 **\$AS-TwoStep**입니다.

TwoStep-AS 모델을 통찰하는 강력한 기술은 규칙 귀납을 사용하여 모델이 찾은 군집을 구별하는 특성을 발견하는 것입니다.

모델 브라우저 사용에 대한 일반 정보는 모델 너깃 찾아보기의 내용을 참조하십시오.

① TwoStep-AS 군집 모델 너깃 설정

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 폐치하여 SPSS® Modeler에서 스코어를 계산합니다.
- **네이티브 SQL로 변환하여 스코어링** 선택된 경우, 네이티브 SQL을 생성하여 데이터베이스 내에서 모델에 스코어를 계산합니다.

 **참고:** 이 옵션은 빠른 결과를 제공하지만 모델의 복잡도가 증가할 수록 기본 SQL의 복잡도와 크기가 증가합니다.

- 데이터베이스 외부 스코어 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

(9) K-평균-AS 노드

K-평균은 일반적으로 가장 많이 사용되는 군집 알고리즘 중 하나입니다. 여기에서는 데이터 포인트를 사전 정의된 갯수의 군집으로 모읍니다.⁸⁾ SPSS® Modeler에서 K-평균-AS 노드는 Spark로 구현됩니다.

K-평균 알고리즘에 대한 자세한 정보는

<https://spark.apache.org/docs/2.2.0/ml-clustering.html>의 내용을 참조하십시오.

K-평균-AS 노드는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

① K-Means-AS 노드 필드

필드 탭은 분석에 사용되는 필드를 지정합니다.

사전 정의된 역할 사용: 이 옵션은 업스트림 유형 노드의 필드 정보를 사용하도록 노드에 지시합니다. 기본적으로 선택되어 있습니다.

사용자 정의 필드 할당 사용: 수동으로 입력 필드를 지정하려면 이 옵션을 선택한 다음 하나의 입력 필드 또는 다중 필드를 선택하십시오. 이 옵션을 사용하는 것은 유형 노드에서 **입력** 내에서 필드 역할을 설정하는 것과 유사합니다.

② K-평균-AS 노드 작성 옵션

작성 옵션 탭에서는 모델 작성에 대한 일반 옵션, 군집 중심 초기화를 위한 초기화 옵션 및 컴퓨팅 반복 및 난수 시드를 위한 고급 옵션을 포함한 K-평균-AS 노드에 대한 작업 옵션을 지정할 수 있습니다. 자세한 정보는 SparkML에 대한 K-평균의 JavaDoc을 참조하십시오.⁹⁾

일반

모델 이름. 특정 군집에 대한 스코어링 이후 생성되는 필드 이름입니다. **자동(기본값)**을 선택하거나 **사용자 정의를** 선택하고 이름을 입력하십시오.

8) "Clustering." Apache Spark. MLlib: Main Guide. Web. 3 Oct 2017.

9) "Class KMeans." Apache Spark. JavaDoc. Web. 3 Oct 2017.

군집 수. 생성할 군집 수를 지정합니다. 기본값은 5이고 최소값은 2입니다.

초기화

초기화 모드. 군집 중심 초기화를 위한 방법을 지정합니다. **K-평균II**가 기본값입니다. 이러한 두 가지 방법에 대한 세부사항은 확장 가능한 K-평균++의 내용을 참조하십시오.¹⁰⁾

초기화 단계. **K-평균II** 초기화 모드가 선택되면, 초기화 단계 수를 지정하십시오. 2가 기본값입니다.

고급

고급 설정. 다음과 같이 고급 옵션을 설정하려는 경우 이 옵션을 선택하십시오.

최대 반복. 군집 중심을 검색할 때 수행할 최대반복수를 지정하십시오. 20이 기본값입니다.

허용치. 반복 알고리즘의 수렴허용치를 지정하십시오. 1.0E-4가 기본값입니다.

난수 시드 설정. 난수 생성기에서 사용한 시드를 생성하려면 이 옵션을 선택하고 **생성**을 클릭하십시오.

표시

그래프 표시. 출력에 그래프를 포함시키려는 경우 이 옵션을 선택하십시오.

다음 표는 SPSS® Modeler K-평균-AS 노드의 설정과 K-평균 Spark 매개변수 사이의 관계를 표시합니다.

SPSS Modeler 설정	스크립트이름(특성이름)	K-평균SparkML매개변수
입력 필드	features	
군집 수	clustersNum	k
초기화 모드	initMode	initMode
초기화 단계	initSteps	initSteps

10) Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. <http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf>.

SPSS Modeler 설정	스크립트이름(특성이름)	K-평균SparkML매개변수
최대 반복	maxIter	maxIter
허용치	toleration	tol
난수 시드	randomSeed	seed

(10) 군집 뷰어

군집 모델은 일반적으로 검토한 변수를 기준으로 하여 유사한 레코드 그룹(또는 군집)을 찾는 데 사용됩니다. 여기서 동일한 그룹의 멤버 간 유사성은 높고 다른 그룹의 멤버 간 유사성은 낮습니다. 결과를 사용하여 명확하지 않은 연관을 식별할 수 있습니다. 예를 들어, 고객 선호도, 수입 수준, 구매 습관의 군집분석을 통해 특정 마케팅 캠페인에 반응할 가능성이 높은 고객의 유형을 식별할 수 있습니다.

다음 두 가지 방법으로 군집 표시의 결과를 해석할 수 있습니다.

- 군집을 조사하여 해당 군집의 고유한 특성을 판별합니다. *한 군집에 모든 고수의 차용자가 포함되어 있습니까? 다른 군집보다 이 군집에 있는 레코드 수가 많습니까?*
- 군집에서 필드를 조사하여 군집 사이에 값이 분포되는 방식을 판별합니다. *개인의 교육 수준이 군집의 소속을 판별합니까? 높은 신용 스코어가 군집 간의 소속을 구별합니까?*

기본 보기와 군집 뷰어의 링크된 다양한 보기를 사용하여 이러한 질문에 대답할 수 있는 정보를 얻을 수 있습니다.

누가 군집 기법을 사용합니까?

군집 기법은 다음을 비롯하여 광범위한 상황에 유용합니다.

- **시장 세분화.** 고객층에서 고유한 그룹을 식별하여 정확한 영업 활동의 목표를 지정할 수 있습니다.
- **제품 번들화.** 특정 고객 유형에 어필하는 경향이 있는 제품 그룹을 식별합니다.
- **형식적 분류.** 정식 분류학에 속하는 식물 또는 동물 등과 같은 그룹을 분류합니다.
- **의료 진단.** 생물학적 패턴을 사용하여 질병을 식별하거나 진단하는 규칙을 밝혀냅니다.

IBM® SPSS® Modeler에서 다음과 같은 군집 모델 너깃을 생성할 수 있습니다.

- 코호넨 넷 모델 너깃
- K-평균 모델 너깃
- 이단계 군집 모델 너깃

군집 모델 너깃에 대한 정보를 보려면 모델 노드를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **찾아보기**를 선택하거나 스트림의 노드에 대해 **편집**을 선택하십시오. 또는 자동 군집 모델링 노드를 사용하는 경우에는 자동 군집 모델 너깃 내에서 필수 군집 너깃을 두 번 클릭하십시오. 자세한 정보는 자동 군집 노드의 내용을 참조하십시오.

① 군집 뷰어 - 모델 탭

군집 모델의 모델 탭은 군집 간의 필드에 대한 통계 및 분포를 그래픽 표시로 요약하여 보여줍니다. 이를 **군집 뷰어**라 합니다.

참고: 모델 탭은 IBM® SPSS® Modeler 13 이전 버전에 작성된 모델에는 사용할 수 없습니다.

군집 뷰어는 2개 패널 즉, 왼쪽의 기본 보기와 오른쪽의 링크 또는 보조 보기로 구성되어 있습니다. 2개의 기본 보기가 있습니다.

- 모델 요약(기본값). 자세한 정보는 모델 요약 보기의 내용을 참조하십시오.
- 군집. 자세한 정보는 군집 보기의 내용을 참조하십시오.

4개의 링크/보조 보기가 있습니다.

- 예측자 중요도. 자세한 정보는 군집 예측자 중요도 보기의 내용을 참조하십시오.
- 군집 크기(기본값). 자세한 정보는 군집 크기 보기의 내용을 참조하십시오.
- 셀 분포. 자세한 정보는 셀 분포 보기의 내용을 참조하십시오.
- 군집 비교. 자세한 정보는 군집 비교 보기의 내용을 참조하십시오.

가. 모델 요약 보기

모델 요약 보기에는 결과를 불량, 양호 또는 우수로 표시하기 위해 음영 처리된 군집 결합 및 분리의 실루엣 측도를 비롯하여 군집 모델의 스냅샷 또는 요약이 표시됩니다. 이 스냅샷을 사용하여 품질이 불량한지 여부를 신속하게 확인할 수 있습니다. 불량한 경우 모델링 노드로 돌아가서 군집 모델 설정을 수정하여 결과를 개선하도록 결정할 수 있습니다.

불량, 우수, 양호 결과는 군집 구조 해석에 관한 Kaufman 및 Rousseeuw(1990) 작업을 기준으로 합니다. 모델 요약 보기에서 양호 결과는 Kaufman 및 Rousseeuw의 평가를 군집 구조에 대한 합리적 또는 강력한 증거로 반영하는 데이터이고, 우수는 약한 증거 평가를 반영하는 데이터이며, 불량은 충분한 증거가 없다는 평가를 반영하는 데이터에 해당합니다.

실루엣 측도는 전체 레코드에 대한 평균을 구합니다($(B - A) / \max(A, B)$). 여기서, A는 군집 중심까지의 레코드 거리이고 B는 속해 있지 않은 가장 가까운 군집 중심까지의 레코드 거리입니다.

다. 실루엣 계수 1은 모든 케이스가 군집 중심에 직접 위치해 있다는 의미입니다. 값 -1은 모든 케이스가 일부 다른 군집의 군집 중심에 있음을 의미합니다. 0 값은 평균적으로 케이스가 해당 군집 중심과 가장 가까운 다른 군집 사이의 등거리에 있음을 의미합니다.

요약에 다음 정보를 포함한 테이블이 있습니다.

- **알고리즘.** 사용한 군집화 알고리즘입니다(예: "이단계").
- **입력 변수.** 필드 수이며 **입력** 또는 **예측자**라고도 합니다.
- **군집.** 솔루션의 군집 수입니다.

나. 군집 보기

군집 보기에는 각 군집의 군집 이름, 크기, 프로파일이 포함된 변수별 군집 눈금이 있습니다.

눈금의 열은 다음 정보를 포함합니다.

- **군집.** 알고리즘을 통해 작성된 군집 번호입니다.
- **레이블.** 각 군집에 적용되는 레이블입니다(기본적으로 공백임). 셀을 두 번 클릭하여 군집 콘텐츠를 설명하는 레이블을 입력하십시오(예: "고급 승용차 구매자").
- **설명.** 군집 콘텐츠에 대한 설명입니다(기본적으로 비어 있음). 셀을 두 번 클릭하여 군집에 대한 설명을 입력하십시오(예: "\$100,000 이상 수입의 55세 이상 전문직 종사자").
- **크기.** 각 군집의 크기이며, 전체 군집 표본에 대한 퍼센트로 표시됩니다. 눈금에 있는 각 크기 셀에는 군집 내의 크기 퍼센트를 보여주는 세로 막대, 숫자 형식의 크기 퍼센트, 군집 케이스 빈도가 표시됩니다.
- **변수.** 개별 입력 또는 예측자이며 기본적으로 전체 중요도별로 정렬됩니다. 열의 크기가 같으면 군집 번호의 오름차순으로 표시됩니다.
전체 변수 중요도는 셀 배경 음영 색상으로 표시됩니다. 중요도가 가장 높은 변수는 가장 어둡게 표시되고, 중요도가 가장 낮은 변수는 음영 처리되지 않습니다. 테이블 위의 가이드는 각 변수 셀 색상과 연결된 중요도를 나타냅니다.

셀 위에 마우스를 올려 놓으면 변수의 전체 이름/레이블 및 셀의 중요도 값이 표시됩니다. 보기 및 변수 유형에 따라 추가 정보가 표시될 수 있습니다. 군집 중심 보기에는 셀 통계량 및 셀 값이 포함됩니다(예: "평균: 4.32"). 범주형 변수의 경우 최대 빈도(모달) 범주의 이름 및 퍼센트가 셀에 표시됩니다.

군집 보기 내에서 다양한 방법을 선택하여 군집 정보를 표시할 수 있습니다.

- **군집 및 변수 전치.** 자세한 정보는 군집 및 변수 전치의 내용을 참조하십시오.
- **변수 정렬.** 자세한 정보는 변수 정렬의 내용을 참조하십시오.
- **군집 정렬.** 자세한 정보는 군집 정렬의 내용을 참조하십시오.
- **셀 콘텐츠 선택.** 자세한 정보는 셀 콘텐츠의 내용을 참조하십시오.

ㄱ. 군집 및 변수 전치

기본적으로 군집은 열로 표시되고 변수는 행으로 표시됩니다. 이 표시를 반대로 바꾸려면 **변수 정렬 기준** 단추의 왼쪽에 있는 **군집 및 변수 전치** 단추를 클릭하십시오. 예를 들어, 표시되는 군집 수가 많아서 데이터를 보기 위해 가로 스크롤 양을 줄여야 하는 경우에 이 작업을 수행할 수 있습니다.

ㄴ. 변수 정렬

변수 정렬 기준 단추를 사용하여 변수 셀을 표시할 방법을 선택할 수 있습니다.

- **전체 중요도.** 기본 정렬 순서입니다. 전체 중요도의 내림차순으로 변수가 정렬되고 정렬 순서는 군집 전체에 동일합니다. 중요도 값이 같은 변수는 변수 이름의 오름차순으로 나열됩니다.
- **군집 내 중요도.** 각 군집의 중요도에 따라 변수가 정렬됩니다. 중요도 값이 같은 변수는 변수 이름의 오름차순으로 나열됩니다. 이 옵션을 선택하면 일반적으로 정렬 순서가 군집 전체에서 달라집니다.
- **이름.** 이름의 문자순으로 변수가 정렬됩니다.
- **데이터 순서.** 데이터 세트의 순서대로 변수가 정렬됩니다.

ㄷ. 군집 정렬

기본적으로 군집은 크기의 내림차순으로 정렬됩니다. **군집 정렬 기준** 단추를 사용하여 이름의 문자순으로 또는 고유 레이블을 작성한 경우에는 대신에 영숫자 레이블순으로 군집을 정렬할 수 있습니다.

레이블이 동일한 변수는 군집 이름별로 정렬됩니다. 군집이 레이블별로 정렬되어 있을 때 군집의 레이블을 편집하는 경우 정렬 순서가 자동으로 업데이트됩니다.

ㄹ. 셀 콘텐츠

셀 단추를 사용하여 변수 및 평가 필드에 대한 셀 콘텐츠 표시를 변경할 수 있습니다.

- **군집중심.** 기본적으로 셀은 변수 이름/레이블 및 각 군집/변수 조합에 대한 중심 경향을 표시합니다. 연속형 필드에는 평균이 표시되고 범주형 필드에는 범주 퍼센트와 함께 최빈값(가장 자주 발생하는 범주)이 표시됩니다.
- **절대 분포.** 변수 이름/레이블 및 각 군집 내에 있는 변수의 절대 분포를 표시합니다. 범주형 변수의 경우 데이터 값의 오름차순으로 정렬된 범주가 오버레이된 막대형 차트가 표시됩니다. 연속형 변수의 경우에는 각 군집에 동일한 엔드포인트 및 구간을 사용하는 매끄러운 평활 밀도 도표가 표시됩니다.
진한 빨간색 표시는 군집 분포를 나타내는 반면 연한 빨간색 표시는 전체 데이터를 나타냅니다.

- **상대 분포.** 변수 이름/레이블 및 셀의 상대 분포를 표시합니다. 일반적으로 이 표시는 상대 분포가 표시된다는 점을 제외하면 절대 분포의 표시 내용과 유사합니다. 진한 빨간색 표시는 군집 분포를 나타내는 반면 연한 빨간색 표시는 전체 데이터를 나타냅니다.
- **기본 보기.** 군집 수가 많을 경우 스크롤하지 않으면 모든 세부사항을 보기가 어려울 수 있습니다. 스크롤하는 양을 줄이려면 이 보기를 선택하여 보다 간결한 형태의 테이블로 표시하도록 변경하십시오.

다. 군집 예측자 중요도 보기

예측자 중요도 보기는 모델을 추정할 때 각 필드의 상대적 중요도를 표시합니다.

라. 군집 크기 보기

군집 크기 보기는 각 군집을 포함한 원형 차트를 표시합니다. 원형 차트의 각 조각마다 개별 군집의 백분율 크기가 표시됩니다. 각 조각에 마우스를 올려 놓으면 해당 조각 내의 개수가 표시됩니다.

차트 아래의 테이블에 다음 크기 정보가 나열됩니다.

- 가장 작은 군집의 크기(전체 개수 및 퍼센트).
- 가장 큰 군집의 크기(전체 개수 및 퍼센트 모두).
- 가장 큰 군집의 크기 대 가장 작은 군집의 크기 비율.

마. 셀 분포 보기

셀 분포 보기는 군집 기본 패널의 테이블에서 선택하는 변수 셀에 대해 확장되어 보다 자세한 데이터 분포의 도표를 표시합니다.

바. 군집 비교 보기

군집 비교 보기는 눈금 스타일의 레이아웃으로 구성되며 행에 변수가 있고 열에 군집이 선택되어 있습니다. 이 보기를 사용하면 군집을 구성하는 요인을 보다 잘 이해할 수 있습니다. 또한 전체 데이터는 물론 개별 데이터를 서로 비교하여 군집 간의 차이점을 확인할 수 있습니다.

표시할 군집을 선택하려면 군집 기본 패널에서 군집 열 맨 위를 클릭하십시오. Ctrl 키 또는 Shift 키를 클릭한 채로 마우스 단추를 클릭하여 비교할 군집을 둘 이상 선택 또는 선택 취소하십시오.

참고: 최대 다섯 개의 군집을 표시하도록 선택할 수 있습니다.

군집은 선택한 순서대로 표시됩니다. 필드 순서는 **변수 정렬 기준** 옵션으로 판별됩니다. **군집 내 중요도**를 선택하면 필드가 항상 전체 중요도별로 정렬됩니다.

배경 도표에 각 변수의 전체 분포가 표시됩니다.

- 범주형 변수는 점도표로 표시됩니다. 점 크기는 각 군집의 빈도가 가장 높은/전형 범주를 나타냅니다(변수별).
- 연속형 변수는 상자도표로 표시됩니다. 이 도표는 전체 중위수 및 사분위수 범위를 표시합니다.

이러한 배경 보기에 오버레이된 항목은 선택된 군집의 상자도표입니다.

- 연속형 변수의 경우 사격형 표식과 가로 선은 각 군집의 중앙값 및 사분위수 범위를 나타냅니다.
- 각 군집은 보기 맨 위에 다른 색상으로 표시됩니다.

② 군집 뷰어 탐색

군집 뷰어는 대화형 표시장치입니다. 다음을 수행할 수 있습니다.

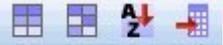
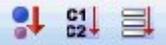
- 필드 또는 군집을 선택하여 세부사항을 봅니다.
- 군집을 비교하여 원하는 항목을 선택합니다.
- 표시를 변경합니다.
- 축을 전치합니다.
- 생성 메뉴를 사용하여 파생, 필터, 선택 노드를 생성합니다.

도구 모음 사용

도구 모음 옵션을 사용하여 왼쪽 및 오른쪽 패널에 표시되는 정보를 제어할 수 있습니다. 도구 모음 제어를 사용하여 표시의 방향(위에서 아래로, 왼쪽에서 오른쪽 또는 오른쪽에서 왼쪽)을 바꿀 수 있습니다. 또한 뷰어를 기본 설정으로 재설정하고 대화 상자를 열어 기본 패널에 군집 보기 내용을 지정할 수도 있습니다.

변수 정렬 기준, 군집 정렬 기준, 셀, 표시 옵션은 기본 패널에서 **군집** 보기를 선택한 경우에만 사용할 수 있습니다. 자세한 정보는 군집 보기의 내용을 참조하십시오.

표 1. 도구 모음 아이콘

아이콘	주제
	군집 및 변수 전치 참조
	변수 정렬 기준 참조
	군집 정렬 기준 참조
	셀 참조

군집 모델에서 노드 생성

생성 메뉴로 군집 모델에 기반하여 새 노드를 작성할 수 있습니다. 이 옵션은 생성된 모델의 모델 탭에서 사용할 수 있으며 현재 표시 또는 선택(즉, 표시된 모든 군집 또는 선택된 모든 군집)을 기준으로 하여 노드를 생성할 수 있습니다. 예를 들어, 단일 피처를 선택한 후 필터 노드를 생성하여 다른 모든(표시되지 않은) 변수를 삭제할 수 있습니다. 생성된 노드는 캔버스에 연결되지 않은 상태로 배치됩니다. 또한 모델 팔레트에 모델 너깃의 사본을 생성할 수 있습니다. 실행하기 전에 노드를 연결하고 원하는 대로 편집할 것을 기억하십시오.

- **모델링 노드 생성.** 스트림 캔버스에 모델링 노드를 작성합니다. 예를 들어, 이 옵션은 이러한 모델 설정을 사용하려는 스트림은 있지만 이 설정을 생성하는 데 사용하는 모델링 노드가 더 이상 없는 경우에 유용합니다.
- **모델을 팔레트로.** 모델 팔레트에 너깃을 작성합니다. 이 옵션은 동료가 모델 자체가 아닌 모델을 포함한 스트림을 보냈을 경우에 유용합니다.
- **필터 노드.** 군집 모델에 사용되지 않고/거나 현재 군집 뷰어 표시에 표시되지 않는 필드를 필터링하기 위한 새 필터 노드를 작성합니다. 이 군집 노드로부터의 유형 노드 업스트림이 있는 경우 생성된 필터 노드는 역할이 **목표**인 모든 필드를 삭제합니다.
- **필터 노드(선택에서).** 군집 뷰어의 선택을 기준으로 하여 필드를 필터링할 수 있는 새 필터 노드를 작성합니다. Ctrl 키를 누른 채로 마우스 단추를 클릭하여 여러 필드를 선택하십시오. 군집 뷰어에 선택된 필드는 삭제된 다운스트림이지만 실행하기 전에 필터 노드를 편집해서 이 동작을 변경할 수 있습니다.
- **선택 노드.** 현재 군집 뷰어 표시에 표시되는 군집의 소속을 기준으로 하여 레코드를 선택할 수 있는 새 선택 노드를 작성합니다. 선택 조건은 자동으로 생성됩니다.
- **선택 노드(선택에서).** 군집 뷰어에 선택된 군집의 소속을 기준으로 하여 레코드를 선택할 수 있는 새 선택 노드를 작성합니다. Ctrl 키를 누른 채로 마우스 단추를 클릭하여 여러 군집을 선택하십시오.
- **파생 노드.** 군집 뷰어에 표시된 모든 군집의 소속을 기준으로 하여 레코드에 참 또는 거짓 값을 할당하는 플래그 필드를 파생시킬 새 파생 노드를 작성합니다. 파생 조건은 자동으로 생성됩니다.

- **파생 노드(선택에서)**. 군집 뷰어에 선택된 군집의 소속을 기준으로 하여 플래그 필드를 파생하는 새 파생 노드를 작성합니다. Ctrl 키를 누른 채로 마우스 단추를 클릭하여 여러 군집을 선택하십시오.

생성 메뉴를 사용하여 노드 생성 외에 그래프를 작성할 수도 있습니다. 자세한 정보는 군집 모델에서 그래프 생성의 내용을 참조하십시오.

군집 보기 표시 제어

기본 패널에서 군집 보기에 표시되는 내용을 제어하려면 **표시** 단추를 클릭하십시오. 그러면 표시 대화 상자가 열립니다.

변수. 기본적으로 선택되어 있습니다. 모든 입력 변수를 숨기려면 선택란을 선택 취소하십시오.

평가 필드. 표시할 평가 필드(군집 모델을 작성하는 데 사용하지 않았지만 군집을 평가하기 위해 모델 뷰어로 보낸 필드)를 선택하십시오. 기본적으로 아무 것도 표시되지 않습니다. 참고 평가 필드는 둘 이상의 값을 포함하는 문자열이어야 합니다. 사용 가능한 평가 필드가 없으면 이 선택란을 사용할 수 없습니다.

군집 설명. 기본적으로 선택되어 있습니다. 모든 군집 설명 셀을 숨기려면 선택란을 선택 취소하십시오.

군집 크기. 기본적으로 선택되어 있습니다. 모든 군집 크기 셀을 숨기려면 선택란을 선택 취소하십시오.

최대 범주 수. 범주형 변수의 차트에 표시할 최대 범주 수를 지정하며 기본값은 20입니다.

③ 군집 모델에서 그래프 생성

군집 모델은 많은 정보를 제공하지만 비즈니스 사용자가 쉽게 액세스할 수 있는 형식이 아닐 경우가 있습니다. 비즈니스 보고서, 프레젠테이션 등에 쉽게 통합할 수 있는 방식으로 데이터를 제공하기 위해 선택한 데이터의 그래프를 만들 수 있습니다. 예를 들어, 군집 뷰어에서 선택한 군집에 대한 그래프를 생성할 수 있어서 해당 군집의 케이스에 대한 그래프만 작성합니다.

참고: 모델 너깃이 스트림의 다른 노드에 연결되어 있을 때에는 군집 뷰어에서만 그래프를 생성할 수 있습니다.

그래프 생성

1. 군집 뷰어를 포함한 모델 너깃을 여십시오.

2. 모델 탭의 보기 드롭다운 목록에서 군집을 선택하십시오.
3. 기본 보기에서 그래프를 생성할 단일 또는 복수 군집을 선택하십시오.
4. 생성 메뉴에서 **그래프(선택 사항 기준)**를 선택하십시오. 그래프보드 기본 탭이 표시됩니다.
참고: 기본 및 세부사항 탭은 그래프보드를 이러한 방식으로 표시할 때에만 사용 가능합니다.
5. 기본 또는 세부사항 탭 설정을 사용하여 그래프에 표시할 세부사항을 지정하십시오.
6. 확인을 눌러 그래프를 생성하십시오.

그래프 머리말은 포함하도록 선택한 단일 또는 복수 군집과 모델 유형을 식별합니다.

10) 연관 규칙

연관 규칙은 특정 결론(예를 들어, 특정 제품의 구매)을 조건 세트(예를 들어, 여러 다른 제품 구매)와 연관시킵니다. 예를 들어, 다음 규칙은

```
beer <= cannedveg & frozenmeal (173, 17.0%, 0.84)
```

cannedveg 및 *frozenmeal*이 함께 발생할 때 *beer*도 종종 발생함을 보여줍니다. 이 규칙은 신뢰도가 84%로, 데이터의 17% 또는 173개 레코드에 적용됩니다. 연관 규칙 알고리즘은 웹 드와 같은 시각화 기법을 사용하여 수동으로 찾을 수 있는 연관을 자동으로 찾습니다.

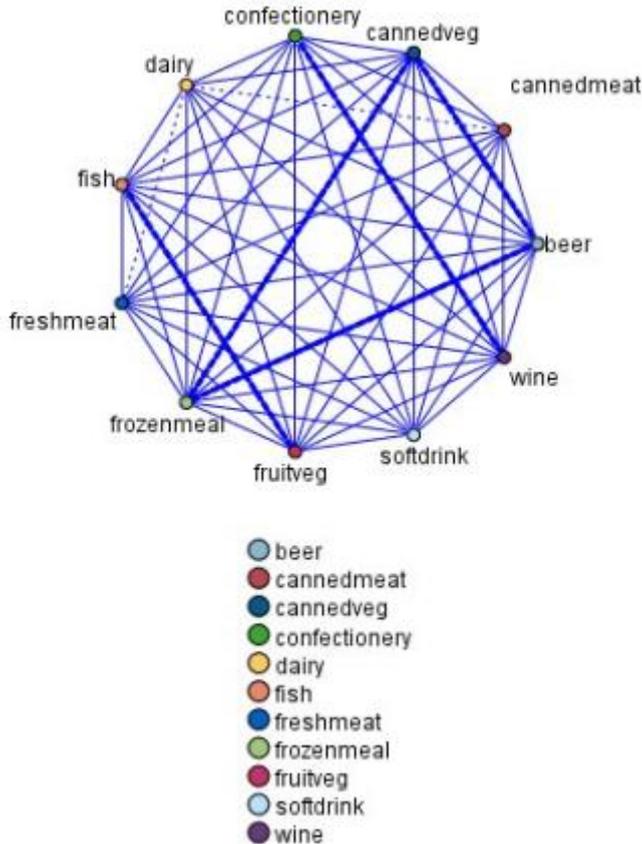
보다 표준적인 의사결정 트리 알고리즘(C5.0 및 C&R 트리)에 비해 연관 규칙 알고리즘을 사용했을 때의 이점은 모든 속성 간에 연관이 존재할 수 있다는 점입니다. 의사결정 트리 알고리즘은 단일 결론만 포함하는 규칙을 작성하지만, 연관 알고리즘은 각각 다른 결론을 보유할 수 있는 많은 규칙을 찾으려고 합니다.

연관 알고리즘의 단점은 잠재적으로 매우 큰 검색 공간에서 패턴을 찾으려 시도하기 때문에 의사결정 트리 알고리즘에 비해 실행하는 데 훨씬 더 많은 시간이 필요할 수 있다는 점입니다. 이 알고리즘은 **생성 및 검정** 방법을 사용하여 규칙을 찾고(단순 규칙은 초기에 생성됨) 이 규칙을 데이터 세트와 대조하여 검증합니다. 우수한 규칙을 저장한 후 다양한 제약조건이 있는 모든 규칙을 특수화합니다. **특수화**는 규칙에 조건을 추가하는 프로세스입니다. 그런 다음 새 규칙을 데이터와 대조하여 검증하고 프로세스는 발견한 최상의 또는 가장 관심 있는 규칙을 반복해서 저장합니다. 사용자는 대개 규칙에 허용할 가능한 전항 수에 몇 가지 한계를 설정하고, 정보 이론에 기반한 다양한 기법 또는 효율적인 색인화 스킴을 사용하여 잠재적으로 큰 검색 공간을 줄여 나갑니다.

처리가 끝나면 최상의 결과 테이블이 제시됩니다. 의사결정 트리와 달리, 이 연관 규칙 세트는 표준 모델(예를 들어, 의사결정 트리 또는 신경망)을 통해 가능한 방식으로 직접 예측을 수행할 수는 없습니다. 규칙의 여러 다른 가능한 결론이 존재하기 때문입니다. 연관 규칙을 분류 규칙

세트로 변환하려면 또 다른 변환 수준이 필요합니다. 이러한 이유로 연관 알고리즘을 통해 생성된 연관 규칙을 **세분화되지 않은 모델**이라 부릅니다. 사용자가 세분화되지 않은 모델을 찾아볼 수는 있지만 세분화되지 않은 모델에서 분류 모델을 생성하도록 시스템에 알리지 않으면 이 모델을 명시적으로 분류 모델로서 사용할 수 없습니다. 이 작업은 브라우저에서 메뉴 생성 옵션을 통해 수행합니다.

그림 1. 장바구니 항목 간의 연관을 보여주는 웹 노드



두 가지 연관 규칙 알고리즘이 지원됩니다.



Apriori 노드는 데이터에서 규칙 세트를 추출하고 정보 내용이 가장 많은 규칙을 꺼냅니다. Apriori는 규칙을 선택하는 5개의 서로 다른 방법을 제공하며 정교한 색인화 스킴을 사용하여 대형 데이터 세트를 효율적으로 처리합니다. 큰 문제점의 경우, Apriori는 일반적으로 학습 속도가 빠릅니다. 보유할 수 있는 규칙 수에 임의 제한이 없으며 최대 32개의 전제조건을 가진 규칙을 처리할 수 있습니다. Apriori에서는 입력 및 출력 필드가 모두 범주형이어야 하지만 이런 유형의 데이터에 최적화되어 있기 때문에 우수한 성능을 제공합니다.



순차규칙 노드는 순차 또는 시간 지향 데이터에서 연관 규칙을 발견합니다. 순차규칙은 예측 가능한 순서로 발생하는 경향이 있는 항목 세트 목록입니다. 예를 들어, 면도기와 애프터셰이브 로션을 구매하는 고객은 다음 번 구매 시에 면도용 크림을 구매할 수 있습니다. 순차규칙 노드는 순차규칙을 찾는 데 효율적인 2패스 방법을 사용하는 CARMA 연관 규칙 알고리즘을 기반으로 합니다.

(1) 테이블 대 트랜잭션 데이터

연관 규칙 모델에 사용되는 데이터는 아래에 설명된 대로 트랜잭션 또는 표 형식일 수 있습니다. 이는 일반적인 설명이며 특정 요구사항은 각 모델 유형마다 문서에 설명된 것처럼 다양할 수 있습니다. 모델을 스코어링할 때에는 스코어링할 데이터가 모델을 작성하는 데 사용하는 데이터의 형식을 미리링해야 함에 유의하십시오. 표 형식 데이터를 사용하여 작성한 모델은 표 형식 데이터만 스코어링하는 데 사용할 수 있고, 트랜잭션 데이터를 사용하여 작성한 모델은 트랜잭션 데이터만 스코어링할 수 있습니다.

트랜잭션 형식

트랜잭션 데이터는 각 트랜잭션이나 항목마다 별도의 레코드가 있습니다. 예를 들어, 고객이 여러 항목을 구매하는 경우 각 항목은 고객 ID로 링크된 연관된 항목이 있는 별도의 레코드가 됩니다. 이를 종종 **till-roll** 형식이라 합니다.

고객	구매
1	jam
2	milk
3	jam
3	bread
4	jam
4	bread
4	milk

Apriori, CARMA, 시퀀스 노드는 모두 트랜잭션 데이터를 사용할 수 있습니다.

표 형식 데이터

표 형식 데이터(장바구니 또는 참 표 데이터라고도 함)는 각 플래그 필드가 특정 항목의 유무를 나타내는 개별 플래그로 표시된 항목이 있습니다. 각 레코드는 연관된 항목의 전체 세트를 나타냅니다. 일정한 모델에 더 많은 특정 요구사항이 있어도 플래그 필드는 범주형 또는 수치일 수 있습니다.

고객	Jam	Bread	Milk
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Apriori, CARMA, GSAR, 시퀀스 노드는 모두 표 형식 데이터를 사용할 수 있습니다.

(2) Apriori 노드

Apriori 노드는 데이터의 연관 규칙을 검색합니다. 연관 규칙은 양식의 명령문입니다.

```
if antecedent(s) then consequent(s)
```

예를 들어, "면도기와 애프터쉐이브 로션을 구매하는 고객은 다음 번 구매 시에 면도용 크림을 구매할 수 있습니다." Apriori는 최고 정보 콘텐츠의 규칙을 빼내어 데이터에서 규칙 세트를 추출합니다. Apriori는 규칙을 선택하는 다섯 가지 다른 방법을 제공하고 정교한 색인화 스킴을 사용하여 큰 데이터 세트를 효과적으로 처리합니다.

요구사항. Apriori 규칙 세트를 작성하려면 하나 이상의 입력 필드와 하나 이상의 목표 필드가 필요합니다. 입력 및 출력 필드(입력, 목표 또는 둘 다 역할의 필드)는 기호여야 합니다. 없음 역할의 필드는 무시됩니다. 노드를 실행하기 전에 필드 유형이 완전히 인스턴스화되어야 합니다. 데이터는 표 또는 트랜잭션 형식이 가능합니다. 자세한 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

강도. 큰 문제가 있을 경우 Apriori는 일반적으로 더 빠르게 학습합니다. .. 보유 가능한 규칙 수에 대한 임의의 한계도 없고 최대 32개의 전제조건이 있는 규칙을 핸들할 수 있습니다. Apriori는 다섯 가지 다른 학습 방법을 제공해서 보다 탄력적으로 당면한 문제에 데이터 마이닝 방법을 부합시킵니다.

① Apriori 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

최소 전향 지원. 규칙 세트의 규칙을 유지하기 위한 지원 기준을 지정할 수 있습니다. **지원**은 전향(규칙의 "if" 파트)이 참인 훈련 데이터의 레코드 퍼센트를 말합니다. (이 지원 정의는 CARMA 및 시퀀스 노드에 사용되는 것과 차이가 있습니다. 자세한 정보는 시퀀스 노드 모델 옵션의 내용을 참조하십시오.) 매우 작은 데이터 서브세트에 적용되는 규칙을 사용하는 경우 이 설정을 늘려보십시오.

참고: Apriori에 대한 지원 정의는 전향이 있는 레코드 수를 기준으로 합니다. 그리고 지원 정의가 규칙의 모든 항목(즉, 전향과 후향 모두)이 있는 레코드 수를 기준으로 하는 CARMA 및 시퀀스 알고리즘과 상반됩니다. 연관 모델의 결과는 (전향) 지원 및 규칙 지원 측도를 모두 표시합니다.

최소 규칙 신뢰도. 신뢰도 기준을 지정할 수도 있습니다. **신뢰도**는 규칙의 전향이 참인 레코드를 기준으로 하며 후향도 참인 레코드의 퍼센트입니다. 즉, 올바른 규칙에 기반한 예측 퍼센트입니다. 신뢰도가 지정된 기준보다 낮은 규칙은 삭제됩니다. 너무 많은 규칙을 사용하는 경우 이 설정을 늘려 보십시오. 너무 적은 규칙을 사용하는(또는 규칙을 전혀 사용하지 않는) 경우에는 이 설정을 줄여 보십시오.

참고: 필요한 경우 값을 강조 표시하고 고유한 값을 입력할 수 있습니다. 신뢰도 값을 1.0 미만으로 줄이면 프로세스의 사용 가능한 메모리가 많이 필요한 점 외에 규칙을 작성하는 데 극단적으로 오랜 시간이 걸릴 수 있음에 유의하십시오.

최대 전향 수. 규칙의 최대 전제조건 수를 지정할 수 있습니다. 이는 규칙의 복잡도를 제한하기 위한 방법입니다. 규칙이 너무 복잡하거나 너무 세부적인 경우 이 설정을 줄여 보십시오. 이 설정은 훈련 시간에도 큰 영향을 미칩니다. 규칙 세트의 훈련 시간이 너무 오래 걸리면 이 설정을 줄여 보십시오.

플래그의 참 값만 이용. 표(참 표) 형식의 데이터에 이 옵션을 선택하면 결과적인 규칙에 참 값만 포함됩니다. 이는 규칙을 보다 쉽게 이해할 수 있도록 합니다. 트랜잭션 형식의 데이터에는 옵션이 적용되지 않습니다. 자세한 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

참고: CARMA 모델 작성 노드는 필드 유형이 플래그인 경우 모델을 작성할 때 비어 있는 레코드를 무시하는 반면 Apriori 모델 작성 노드는 비어 있는 레코드를 포함합니다. 비어 있는 레코드는 모델 작성에 사용된 모든 필드의 값이 거짓인 레코드입니다.

최적화. 특정 요구를 기준으로 한 모델 작성 중에 성능을 늘리려면 이 옵션을 선택하십시오.

- 성능을 개선하기 위해 알고리즘이 디스크 스펀링을 사용하지 않게 하려면 **속도**를 선택하십시오.

- 적합한 시기에 속도가 느려지더라도 알고리즘이 디스크 스피어링을 사용하게 하려면 **메모리**를 선택하십시오. 이 옵션은 기본적으로 선택됩니다.

 **참고:** 분산 모드에서 실행할 때에는 *options.cfg* 파일에 지정된 관리자 옵션이 이 설정을 대체할 수 있습니다. 자세한 정보는 *IBM® SPSS® Modeler Server 관리자 안내서*를 참조하십시오.

② Apriori 노드 고급 옵션

Apriori 작업에 대한 세부 지식이 있는 사용자는 다음 고급 옵션으로 귀납 프로세스를 세부 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

평가 척도. Apriori는 잠재적 규칙을 평가하는 다섯 가지 방법을 지원합니다.

- **규칙 신뢰도.** 기본 방법은 규칙의 신뢰도(또는 정확도)를 사용하여 규칙을 평가합니다. 이 척도의 경우 **평가 척도 하한**이 모델 탭의 **최소 규칙 신뢰도** 옵션과 중복되므로 사용되지 않습니다. 자세한 정보는 Apriori 노드 모델 옵션의 내용을 참조하십시오.
- **신뢰도 차이.** (사전 신뢰도에 대한 절대 신뢰도 차이라고도 합니다.) 이 평가 척도는 규칙의 신뢰도와 사전 신뢰도 간 절대차입니다. 이 옵션은 결과가 고르게 분포되지 않는 편향을 방지합니다. 이를 통해 "명백한" 규칙이 유지되지 않게 합니다. 규칙을 유지하려는 신뢰도의 최소 차이로 평가 척도 하한을 설정하십시오.
- **신뢰도 비율.** (1에 대한 신뢰 지수 차이라고도 합니다.) 이 평가 척도는 1에서 뺀 사전 신뢰도에 대한 규칙 신뢰도의 비율입니다(또는 비율이 1보다 큰 경우에는 역수). 신뢰도 차이와 마찬가지로 이 방법은 고르지 않은 분포를 고려합니다. 이는 특히 희박한 이벤트를 예측하는 규칙을 찾을 때 좋습니다. 규칙을 유지하려는 차이로 평가 척도 하한을 설정하십시오.
- **정보 차이.** (사전 확률에 대한 정보 차이라고도 합니다.) 이 척도는 **정보 이득** 척도를 기준으로 합니다. 특정 후향의 확률이 논리 값(비트)으로 간주되면 정보 이득은 전향을 기준으로 하여 판별할 수 있는 이 비트의 비율입니다. 정보 차이는 전향이 주어진 정보 이득과 후향의 사전 신뢰도만 주어진 정보 이득 간 차이입니다. 이 방법의 중요한 기능은 더 많은 레코드를 처리하는 규칙이 주어진 신뢰도 수준에 선호되도록 지원을 고려하는 것입니다. 규칙을 유지하려는 정보 차이로 평가 척도 하한을 설정하십시오.
- **정규화 카이제곱.** (정규화 카이제곱 척도라고도 합니다.) 이 척도는 전향과 후향 간 연관의 통계 지수입니다. 척도는 0과 1 사이의 값을 사용하도록 정규화되어 있습니다. 이 척도는 정보 차이 척도보다 훨씬 더 지원에 종속적입니다. 규칙을 유지하려는 정보 차이로 평가 척도 하한을 설정하십시오.

전향 없는 규칙 허용. 후향(항목 또는 항목 세트)만 포함한 규칙을 허용하려면 이 옵션을 선택하십시오. 이 옵션은 공통 항목 또는 항목 세트 판별에 관심이 있는 경우에 유용합니다. 예를 들어, *cannedveg*는 *cannedveg* 구매가 데이터의 공통 발생임을 표시하는 전향이 없는 단일 항목

규칙입니다. 일부 경우 가장 확실한 예측에만 관심이 있다면 이러한 규칙을 포함시키려 할 수 있습니다. 이 옵션은 기본적으로 해제 상태입니다. 관례상, 전항이 없는 규칙에 대한 전항 지원은 100%로 표현되며 규칙 지원은 신뢰도와 동일합니다.

(3) CARMA 노드

CARMA 노드는 연관 규칙 발견 알고리즘을 사용하여 데이터의 연관 규칙을 발견합니다. 연관 규칙은 다음 양식의 명령문입니다.

```
if antecedent(s) then consequent(s)
```

예를 들어, 무선 카드와 최고급 무선 라우터를 구매한 웹 고객은 무선 음악 서버도 구매할 가능성이 있습니다(제공된 경우). CARMA 모델은 입력 또는 대상 필드를 지정하지 않아도 데이터에서 규칙 세트를 추출합니다. 이는 생성된 규칙을 보다 광범위한 애플리케이션에 사용할 수 있음을 의미합니다. 예를 들어, 이 노드가 생성한 규칙을 사용하여 후항이 이번 연휴 기간에 홍보하려는 항목인 제품 또는 서비스(전항) 목록을 찾을 수 있습니다. IBM® SPSS® Modeler를 사용하여 전항 제품을 구매한 클라이언트를 판별하고 후항 제품을 홍보하도록 설계된 마케팅 캠페인을 수행할 수 있습니다.

요구사항. Apriori와 다르게 CARMA 노드는 *입력* 또는 *목표* 필드가 필요하지 않습니다. 이는 알고리즘의 작동 방식에 필수적이며 모든 필드를 둘 *다*로 설정해서 Apriori 모델을 작성하는 것과 동일합니다. 작성된 모델을 필터링하여 전항 또는 후항으로만 나열되는 항목을 제한할 수 있습니다. 예를 들어, 모델 브라우저를 사용하여 후항이 이번 연휴 기간에 홍보하려는 항목인 제품 또는 서비스(전항) 목록을 찾을 수 있습니다.

CARMA 규칙 세트를 작성하려면 ID 필드 및 하나 이상의 내용 필드를 지정해야 합니다. ID 필드에는 역할 또는 측정 수준이 있을 수 있습니다. *없음* 역할의 필드는 무시됩니다. 노드를 실행하기 전에 필드 유형이 완전히 인스턴스화되어 있어야 합니다. Apriori와 마찬가지로 데이터는 표 형식 또는 트랜잭션 형식이 가능합니다. 자세한 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

강도. CARMA 노드는 CARMA 연관 규칙 알고리즘을 기준으로 합니다. Apriori와 다르게, CARMA 노드는 전항 지원이 아닌 규칙 지원(전항과 후항 모두에 대한 지원)에 대한 작성 설정을 제공합니다. CARMA는 여러 후항이 있는 규칙도 허용합니다. Apriori처럼, 예측을 작성하기 위해 CARMA 노드가 생성한 모델이 데이터 스트림에 삽입될 수 있습니다. 자세한 정보는 모델 너깅의 내용을 참조하십시오.

① CARMA 노드 필드 옵션

CARMA 노드를 실행하기 전에 CARMA 노드의 필드 탭에 입력 필드를 지정해야 합니다. 대부분의 모델링 노드가 동일한 필드 탭 옵션을 공유하는 반면에 CARMA 노드는 여러 고유 옵션을 포함합니다. 모든 옵션은 아래에 설명되어 있습니다.

유형 노드 설정 사용. 이 옵션은 업스트림 유형 노드의 필드 정보를 사용하도록 노드에 지시합니다. 이는 기본값입니다.

사용자 정의 설정 사용. 이 옵션에서는 업스트림 유형 노드에 지정된 항목 대신, 여기에 지정된 필드 정보를 사용하도록 노드에 지시합니다. 이 옵션을 선택한 후 트랜잭션 또는 표 형식의 데이터를 읽는지 여부에 따라 아래 필드를 지정하십시오.

트랜잭션 형식 사용. 이 옵션은 데이터가 트랜잭션 또는 테이블 형식인지에 따라 이 대화 상자의 나머지 부분에서 필드 제어를 변경합니다. 트랜잭션 데이터를 포함하는 다중 필드를 사용하는 경우 특정 레코드에서 이 필드에 지정된 항목은 단일 시간소인을 포함하는 단일 트랜잭션에서 찾은 항목을 나타낸다고 가정합니다. 자세한 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

표 형식 데이터

트랜잭션 형식 사용을 선택하지 않을 경우 다음 필드가 표시됩니다.

- **입력** 하나 이상의 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 *입력*으로 설정하는 것과 유사합니다.
- **파티션.** 이 필드에서는 모델 작성에 대한 훈련, 검정, 검증 단계에 대한 별도의 샘플로 데이터를 파티션하는 데 사용되는 필드를 지정할 수 있습니다. 한 표본을 사용하여 모델을 생성하고 다른 표본을 사용하여 이를 검증함으로써 현재 데이터에 유사한 보다 큰 데이터 세트로 모델이 일반화되는 정도를 잘 표시할 수 있습니다. 유형 또는 파티션 노드를 사용하여 다중 파티션 필드를 정의한 경우 파티셔닝을 사용하는 각 모델링 노드에 있는 필드 탭에서 단일 파티션 필드를 선택해야 합니다. (하나의 파티션만 존재하는 경우 파티션이 사용될 때마다 자동으로 사용됩니다.) 또한 분석에서 선택한 파티션을 적용하려면 노드의 모델 옵션 탭에서 파티셔닝을 사용 가능하게 해야 합니다. (이 옵션을 선택 취소하면 필드 설정을 변경하지 않고도, 파티셔닝을 사용하지 않을 수 있습니다.)

트랜잭션 데이터

트랜잭션 형식 사용을 선택하면 다음 필드가 표시됩니다.

- **ID.** 트랜잭션 데이터의 경우 목록에서 ID 필드를 선택합니다. 숫자 또는 기호 필드를 ID 필드

로 사용할 수 있습니다. 이 필드의 각 고유한 값은 분석의 특정 단위를 표시해야 합니다. 예를 들어, 장바구니 애플리케이션에서 각 ID는 단일 고객을 나타낼 수 있습니다. 웹 로그 분석 애플리케이션의 경우 각 ID는 컴퓨터(IP 주소별) 또는 사용자(로그인 데이터별)를 나타낼 수 있습니다.

- **연속적 ID.** (Apriori 및 CARMA 노드만) ID가 동일한 모든 레코드를 데이터 스트림에서 함께 그룹화하도록 데이터를 사전 정렬한 경우 이 옵션을 선택하여 처리 속도를 높입니다. 데이터가 사전 정렬되지 않았거나 정렬 여부가 확실하지 않은 경우 이 옵션을 선택하지 않은 상태로 두면 노드가 데이터를 자동으로 정렬합니다.

참고: 데이터가 정렬되지 않은 상태에서 이 옵션을 선택하면 모델에서 유효하지 않은 결과가 발생할 수 있습니다.

- **내용.** 모델의 내용 필드를 지정합니다. 이 필드는 연관 모델링에서 관심이 있는 항목을 포함합니다. 다중 플래그 필드(데이터가 표 형식인 경우) 또는 단일 명목 필드(데이터가 트랜잭션 형식인 경우)를 지정할 수 있습니다.

② CARMA 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

최소 규칙 지원(%). 지원 기준을 지정할 수 있습니다. **규칙 지원**은 훈련 데이터에서 전체 규칙을 포함한 ID 비율을 나타냅니다. (이 지원 정의는 Apriori 노드에 사용된 전항 지원과 차이가 있음에 유의하십시오.) 더 많은 공통 규칙에 초점을 맞추려면 이 설정을 늘리십시오.

최소 규칙 신뢰도(%). 규칙 세트의 규칙을 유지하기 위한 신뢰도 기준을 지정할 수 있습니다. **신뢰도**는 올바른 예측이 작성된 ID(규칙이 예측을 작성하는 모든 ID 중에서) 퍼센트를 나타냅니다. 이 퍼센트는 훈련 데이터를 기준으로 하여 전체 규칙이 있는 ID 수를 전항이 있는 ID 수로 나뉘서 계산합니다. 신뢰도가 지정된 기준보다 낮은 규칙은 삭제됩니다. 너무 많은 규칙을 사용하거나 관심이 없는 경우 이 설정을 늘려 보십시오. 너무 적은 규칙을 사용하는 경우에는 이 설정을 줄여 보십시오.

참고: 필요한 경우 값을 강조 표시하고 고유한 값을 입력할 수 있습니다. 신뢰도 값을 1.0 미만으로 줄이면 프로세스의 사용 가능한 메모리가 많이 필요한 점 외에 규칙을 작성하는 데 극단적으로 오랜 시간이 걸릴 수 있음에 유의하십시오.

최대 규칙 크기. 구별되는 item sets(items에 반대로)의 최대 수를 규칙에 설정할 수 있습니다. 관심 있는 규칙이 상대적으로 짧은 경우 이 설정을 줄여서 규칙 세트 작성 속도를 올릴 수 있습니다.

참고: CARMA 모델 작성 노드는 필드 유형이 플래그인 경우 모델을 작성할 때 비어 있는 레코드를 무시하는 반면 Apriori 모델 작성 노드는 비어 있는 레코드를 포함합니다. 비어 있는 레코드는 모델 작성에 사용된 모든 필드의 값이 거짓인 레코드입니다.

③ CARMA 노드 고급 옵션

CARMA 노드 작업에 대한 세부 지식이 있는 사용자는 다음 고급 옵션으로 모델 작성 프로세스를 세부 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

복수 후향 값이 있는 규칙 제외. “two-headed” 후향 즉, 두 개의 항목을 포함한 후향을 제외하도록 선택합니다. 예를 들어, bread & cheese & fish -> wine&fruit 규칙은 two-headed 후향, wine&fruit을 포함합니다. 기본적으로 이러한 규칙이 포함됩니다.

가지치기 값 설정. 사용된 CARMA 알고리즘은 메모리를 보존하기 위해 처리 중 잠재적 항목 세트 목록에서 희소 항목 세트를 주기적으로 제거(**가지치기**)합니다. 가지치기 빈도를 조정하려면 이 옵션을 선택하십시오. 지정하는 숫자로 가지치기 빈도가 판별됩니다. 더 작은 값을 입력하여 알고리즘의 메모리 요구 사항을 줄이거나(그러나 잠재적으로 필요한 훈련 시간이 늘어남) 더 큰 값을 입력하여 훈련 속도를 높이십시오(그러나 잠재적으로 메모리 요구 사항이 늘어남). 기본값은 500입니다.

지원 변경. 고르지 않게 포함될 경우 빈번할 것으로 보이는 희소 항목 세트를 제외해서 효율성을 높이려면 이 옵션을 선택하십시오. 지원 레벨을 높여서 시작한 후 모델 탭에 지정된 수준까지 감소시키면 됩니다. **예상 트랜잭션** 수의 값을 입력하여 지원 수준을 얼마나 빨리 감소시켜야 하는지 지정하십시오.

전향 없는 규칙 허용. 후향(항목 또는 항목 세트)만 포함한 규칙을 허용하려면 이 옵션을 선택하십시오. 이 옵션은 공통 항목 또는 항목 세트 판별에 관심이 있는 경우에 유용합니다. 예를 들어, *cannedveg*는 *cannedveg* 구매가 데이터의 공통 발생임을 표시하는 전향이 없는 단일 항목 규칙입니다. 일부 경우 가장 확실한 예측에만 관심이 있으면 이러한 규칙을 포함시키려 할 수 있습니다. 이 옵션은 기본적으로 선택되지 않습니다.

(4) 연관 규칙 모델 너깃

연관 규칙 모델 너깃은 다음 연관 규칙 모델링 노드 중 하나를 통해 검색된 규칙을 나타냅니다.

- Apriori
- CARMA

모델 너깃은 모델 작성 중 데이터에서 추출된 규칙에 대한 정보를 포함합니다.

 **참고:** 트랜잭션 데이터를 ID별로 정렬하지 않는 경우 연관 규칙 너깃 스코어링이 올바르지 않을 수 있습니다.

결과 보기

대화 상자에서 모델 탭을 사용하여 연관 모델(Apriori 및 CARMA)과 시퀀스 모델이 생성한 규칙을 찾아볼 수 있습니다. 모델 너깃을 찾아보면 규칙에 대한 정보가 표시되고 새 노드 생성 또는 모델 스코어링 이전에 결과를 필터링하고 정렬하는 옵션이 제공됩니다. 자세한 정보는 연관 규칙 모델 너깃 세부사항 주제를 참조하십시오.

모델 스코어링

세분화된 모델 너깃(Apriori, CARMA, 시퀀스)이 스트림에 추가되고 스코어링에 사용될 수 있습니다. 자세한 정보는 스트림에서 모델 너깃 사용의 내용을 참조하십시오. 스코어링에 사용되는 모델 너깃은 각 대화 상자마다 추가 설정 탭을 포함합니다. 자세한 정보는 연관 규칙 모델 너깃 설정의 내용을 참조하십시오.

세분화되지 않은 모델 너깃은 원시 형식의 스코어링에 사용할 수 없습니다. 대신에, 규칙 세트를 생성해서 이 규칙 세트를 스코어링에 사용할 수 있습니다. 자세한 정보는 연관 모델 너깃에서 규칙 세트 생성의 내용을 참조하십시오.

① 연관 규칙 모델 너깃 세부사항

연관 규칙 모델 너깃의 모델 탭에서 알고리즘을 통해 추출된 규칙을 포함한 테이블을 볼 수 있습니다. 테이블의 각 행은 규칙을 표시합니다. 첫 번째 열이 후항(규칙의 "then" 파트)을 표시하는 반면 다음 열은 전항(규칙의 "if" 파트)을 표시합니다. 후속 열은 신뢰도, 지원, 리프트와 같은 규칙 정보를 포함합니다.

연관 규칙은 종종 다음 테이블의 형식으로 표시됩니다.

표 1. 연관 규칙 예	
후항	전항
Drug = drugY	Sex=F BP = HIGH

예 규칙은 *성별 = "F"* 및 *BP = "HIGH"*이면 약품은 *drugY*로 해석하거나 다른 방식, *성별 = "F"* 및 *BP = "HIGH"*인 레코드의 경우 약품은 *drugY*로 표현됩니다. 대화 상자 도구 모음을 사용하여 신뢰도, 지원, 인스턴스와 같은 추가 정보를 표시할 수 있습니다.

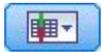
정렬 메뉴. 도구 모음의 정렬 메뉴 단추는 규칙 정렬을 제어합니다. 정렬 방향 단추(위로 또는 아래로 화살표)를 사용하여 정렬 방향(오름차순 또는 내림차순)을 변경할 수 있습니다.

다음 기준에 따라 규칙을 정렬할 수 있습니다.

- 지원
- 신뢰도
- 규칙 지원
- 후향
- 평가
- Lift
- 배포성

메뉴 표시/숨기기. 표시/숨기기 메뉴(기준 도구 모음 단추)는 규칙 표시 옵션을 제어합니다.

그림 1. 표시/숨기기 단추



다음 표시 옵션을 사용할 수 있습니다.

- **규칙 ID**는 모델 작성 중 지정된 규칙 ID를 표시합니다. 규칙 ID로 주어진 예측에 대해 적용하고 있는 규칙을 식별할 수 있습니다. 규칙 ID로 배포성, 제품 정보 또는 전향과 같은 추가 규칙 정보를 나중에 병합할 수도 있습니다.
- **인스턴스**는 규칙이 적용되는 즉, 전향이 참인 고유 ID 수에 대한 정보를 표시합니다. 예를 들어, bread -> cheese 규칙이 주어진 경우 전향 bread를 포함한 학습 데이터의 레코드 수를 **인스턴스**라 부릅니다.
- **지원**은 전향 지원 즉, 학습 데이터를 기준으로 하여 전향이 참인 ID의 비율을 표시합니다. 예를 들어, 학습 데이터의 50%가 bread 구매를 포함하면, 규칙 bread -> cheese의 전향 지원은 50%입니다. **참고:** 여기에 정의된 지원은 인스턴스와 동일하지만 퍼센트로 표현됩니다.
- **신뢰도**는 전향 지원에 대한 규칙 지원 비율을 표시합니다. 후향도 참인 전향이 지정된 ID의 비율을 표시합니다. 예를 들어, 학습 데이터의 50%가 bread를 포함(전향 지원을 나타냄)하지만 20%만 bread와 cheese를 모두 포함(규칙 지원을 나타냄)하는 경우에는, 규칙 bread -> cheese의 신뢰도가 Rule Support / Antecedent Support 또는 이 예의 경우 40%입니다.
- **규칙 지원**은 전체 규칙, 전향, 후향이 참인 ID의 비율을 표시합니다. 예를 들어, 학습 데이터의 20%가 bread와 cheese를 모두 포함하면 규칙 bread -> cheese의 규칙 지원은 20%입니다.
- **평가**는 고급 연관 규칙 기준(신뢰도 차이, 신뢰도 비율, 정보 차이, 정규화 카이제곱) 중 하나를 선택할 경우에 포함됩니다. 이러한 고급 기준 측도는 사용자가 설정한 **평가 속도 하한** 값과 비교되며, 고급 기준 규칙을 선택한 경우에만 적용됩니다. 평가 통계에서 고급 연관 규칙 기준 각각에 대한 의미는 다음과 같습니다.
 - **신뢰도 차이:** 사후 신뢰 - 사전 신뢰
 - **신뢰도 비율:** (사후 신뢰 - 사전 신뢰)/사후 신뢰
 - **정보 차이:** 정보 이득 속도
 - **정규화 카이제곱:** 정규화 카이제곱 통계

이러한 통계는 각각 사용자가 설정한 평가 척도 하한 값과 비교되며, 통계가 이 값을 초과할 경우 규칙이 선택됩니다.

- **리프트**는 후향이 있을 사전 확률에 대한 규칙의 신뢰도 비율을 표시합니다. 예를 들어, 전체 모집단의 10%가 빵을 살 경우 사람들이 20% 신뢰도로 빵을 살 것이라 예측하는 규칙의 리프트는 $20/10 = 2$ 입니다. 다른 규칙에 사람들이 11% 신뢰도로 빵을 살 것이라 지정되면 규칙의 리프트는 1에 근접합니다. 이는 전향이 있다고 해서 후향이 있을 확률이 많이 차이가 나지 않음을 의미합니다. 일반적으로 리프트가 1이 아닌 규칙이 리프트가 1에 가까운 규칙보다 흥미롭습니다.
- **배포성**은 학습 데이터가 전향 조건을 충족시키지만 후향 조건을 충족시키지 않는 퍼센트 척도입니다. 제품 구매 조건에서, 이는 기본적으로 총 고객 기반이 전향을 소유하지만(또는 구매했지만) 후향을 아직 구매하지 않은 퍼센트를 의미합니다. 배포성 통계는 $((Antecedent\ Support\ in\ \#\ of\ Records - Rule\ Support\ in\ \#\ of\ Records) / Number\ of\ Records) * 100$ 으로 정의되며 여기서, *Antecedent Support*는 전향이 참인 레코드 수를 의미하고 *Rule Support*는 전향과 후향이 모두 참인 레코드 수를 의미합니다.

필터 단추. 메뉴의 필터 단추(갈때기 아이콘)는 활성 규칙 필터가 표시되는 패널을 보여주기 위해 대화 상자의 맨 아래까지 확장됩니다. 필터는 모델 탭에 표시되는 규칙 번호의 범위를 좁히는 데 사용됩니다.

그림 2. 필터 단추



필터를 작성하려면 펼쳐진 패널의 오른쪽에 있는 필터 아이콘을 클릭하십시오. 그러면 규칙 표시에 대한 제약조건을 지정할 수 있는 별도의 대화 상자가 열립니다. 필터 단추는 종종 생성 메뉴와 함께 사용되어 먼저 규칙을 필터링한 후 규칙의 서브셋을 포함한 모델을 생성함에 유의하십시오. 자세한 정보는 아래 규칙의 필터 지정의 내용을 참조하십시오.

규칙 찾기 단추. 규칙 찾기 단추(쌍안경 아이콘)로 지정된 규칙 ID에 대해 표시되는 규칙을 검색할 수 있습니다. 인접한 대화 상자에는 사용 가능한 수 중에서 현재 표시된 규칙 수가 표시됩니다. 규칙 ID는 모델에 따라 당시에 발견된 순서대로 지정되며 스코어링 중 데이터에 추가됩니다.

그림 3. 규칙 찾기 단추



규칙 ID를 다시 정렬하려면 다음을 수행하십시오.

1. 먼저 신뢰도 또는 리프트와 같은 원하는 측정에 따라 규칙 표시 테이블을 정렬해서 IBM® SPSS® Modeler에 규칙 ID를 재배열할 수 있습니다.
2. 그런 다음 생성 메뉴의 옵션을 사용하여 필터링된 모델을 작성하십시오.
3. 필터링된 모델 대화 상자에서 **다음으로 시작하여 연속적으로 규칙 번호 다시 매기기를 선택** 하고 시작 번호를 지정하십시오.

자세한 정보는 필터링된 모델 생성의 내용을 참조하십시오.

가. 규칙의 필터 지정

기본적으로 Apriori, CARMA, 시퀀스와 같은 규칙 알고리즘은 많은 수의 규칙을 생성할 수 있습니다. 규칙 스코어링을 찾아보거나 능률화할 때 명확성을 개선하려면 관심 있는 전향과 후향이 보다 분명히 표시되도록 규칙을 필터링할 것을 고려해야 합니다. 규칙 브라우저의 모델 탭에서 필터링 옵션을 사용하여 필터 조건을 지정할 대화 상자를 열 수 있습니다.

필터를 작성하려면 펼쳐진 패널의 오른쪽에 있는 필터 편집 단추(갈때기 아이콘)를 클릭하십시오. 그러면 규칙 표시에 대한 제약조건을 지정할 수 있는 별개의 필터 편집 대화 상자가 열립니다.

후향. 지정된 후향의 포함 또는 제외를 기준으로 한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. 규칙에 최소 하나의 지정된 후향이 포함되는 필터를 작성하려면 **하나 이상 포함**을 선택하십시오. 또는 **제외**를 선택하여 지정된 후향을 제외하는 필터를 작성하십시오. 목록 상자 오른쪽에 있는 선택도구 아이콘을 사용하여 후향을 선택할 수 있습니다. 그러면 생성된 규칙에 있는 모든 후향이 나열된 대화 상자가 열립니다.

참고: 후향은 둘 이상의 항목을 포함할 수 있습니다. 필터는 지정된 항목 중 하나가 후향에 포함되었는지만 검사합니다.

전향. 지정된 전향의 포함 또는 제외를 기준으로 한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. 목록 상자 오른쪽에 있는 선택도구 아이콘을 사용하여 항목을 선택할 수 있습니다. 그러면 생성된 규칙에 있는 모든 전향이 나열된 대화 상자가 열립니다.

- 지정된 모든 전향을 규칙에 포함해야 포함 필터로 필터를 설정하려면 **모두 포함**을 선택하십시오.
- 규칙에 최소 하나의 지정된 전향이 포함되는 필터를 작성하려면 **하나 이상 포함**을 선택하십시오.
- 지정된 전향이 포함된 규칙을 제외하는 필터를 작성하려면 **제외**를 선택하십시오.

신뢰도. 규칙의 신뢰도 수준에 기반한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. **최소** 및 **최대** 제어를 사용하여 신뢰도 범위를 지정할 수 있습니다. 생성된 모델을 찾아볼 때 신뢰도가 퍼센트로 나열됩니다. 출력을 스코어링할 때에는 신뢰도가 0 - 1의 숫자로 표현됩니다.

전향 지원. 규칙의 전향 지원의 수준을 기준으로 한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. 전향 지원은 인기 지수와 유사하도록, 현재 규칙과 동일한 전향을 포함한 훈련 데이터의 비율을 표시합니다. **최소** 및 **최대** 제어를 사용하여 지원 수준을 기준으로 규칙을 필터링하는 데 사용되는 범위를 지정할 수 있습니다.

리프트. 규칙의 리프트 측정을 기준으로 한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. **참고:** 리프트 필터링은 릴리스 8.5 이후에 작성된 연관 모델이나 리프트 측정을 포함한 이전 모델에만 사용 가능합니다. 시퀀스 모델은 이 옵션을 포함하지 않습니다.

이 대화 상자에서 사용한 모든 필터를 적용하려면 **확인**을 클릭하십시오.

나. 규칙의 그래프 생성

연관 노드는 많은 정보를 제공하지만 비즈니스 사용자가 쉽게 액세스할 수 있는 형식이 아닐 경우가 있습니다. 비즈니스 보고서, 프레젠테이션 등에 쉽게 통합할 수 있는 방식으로 데이터를 제공하기 위해 선택한 데이터의 그래프를 만들 수 있습니다. 모델 탭에서 선택한 규칙에 대한 그래프를 생성할 수 있으므로 해당 규칙의 케이스에 대한 그래프만 작성할 수 있습니다.

1. 모델 탭에서 관심이 있는 규칙을 선택하십시오.
2. 생성 메뉴에서 **그래프(선택 사항 기준)**를 선택하십시오. 그래프보드 기본 탭이 표시됩니다.
참고: 기본 및 세부사항 탭은 그래프보드를 이러한 방식으로 표시할 때에만 사용 가능합니다.
3. 기본 또는 세부사항 탭 설정을 사용하여 그래프에 표시할 세부사항을 지정하십시오.
4. 확인을 눌러 그래프를 생성하십시오.

그래프 머리말은 포함하도록 선택한 규칙 및 전항 세부사항을 식별합니다.

② 연관 규칙 모델 너깃 설정

이 설정 탭은 연관 모델(Apriori 및 CARMA)의 스코어링 옵션을 지정하는 데 사용됩니다. 이 탭은 모델 너깃이 스코어링 용도로 스트림에 추가된 후에만 사용 가능합니다.

참고: 세분화되지 않은 모델을 찾아보기 위한 대화 상자에는 설정 탭이 없습니다(스코어링이 불가능하므로). "세분화되지 않은" 모델을 스코어링하려면 먼저 규칙 세트를 생성해야 합니다. 자세한 정보는 연관 모델 너깃에서 규칙 세트 생성의 내용을 참조하십시오.

최대 예측 수 장바구니 항목의 각 세트마다 포함된 최대 예측 수를 지정합니다. 이 옵션은 아래의 규칙 기준과 함께 사용되어 "top" 예측을 생성합니다. 여기서, *top*은 아래에 지정된 최상위 레벨의 신뢰도, 지원, 리프트 등을 나타냅니다.

규칙 기준 규칙의 강도를 판별하는 데 사용된 측도를 선택합니다. 규칙은 항목 세트의 최상의 예측을 리턴하기 위해 여기에 선택된 기준의 강도별로 정렬됩니다. 사용 가능한 기준이 다음 목록에 표시됩니다.

- 신뢰도
- 지원
- 규칙 지원(지원 * 신뢰도)
- Lift
- 배포성

반복 예측 허용 스코어링 시 후항이 동일한 여러 규칙을 포함하려면 선택하십시오. 예를 들어, 이 옵션을 선택하면 다음 규칙이 스코어링됩니다.

bread & cheese -> wine
cheese & fruit -> wine

스코어링 시 반복 예측을 제외하려면 이 옵션을 해제하십시오.

참고: 여러 후향(bread & cheese & fruit -> wine & pate)이 있는 규칙은 모든 후향(wine & pate)이 이전에 예측된 경우에만 반복 예상값을 고려합니다.

일치하지 않는 장바구니 항목 무시 항목 세트의 추가 항목 존재를 무시하려면 선택하십시오. 예를 들어, [tent & sleeping bag & kettle]을 포함한 장바구니에 이 옵션이 선택되면 장바구니에 추가 항목(kettle)이 있어도 tent & sleeping bag -> gas_stove 규칙이 적용됩니다.

추가 항목을 제외시켜야 하는 몇 가지 상황이 있을 수 있습니다. 예를 들어, 텐트, 침낭, 주전자를 구매하는 누군가에게 이미 주전자의 존재를 통해 표시되는 가스 스토브가 있을 수도 있습니다. 즉, 가스 스토브가 최상의 예측이 아닐 수도 있습니다. 이러한 경우 규칙 전향이 장바구니의 콘텐츠와 정확히 일치하도록 **Ignore unmatched basket items**를 선택 취소해야 합니다. 기본적으로 일치하지 않는 항목은 무시됩니다.

예측이 장바구니에 없는지 검사. 장바구니에 후향도 없는지 확인하려면 선택하십시오. 예를 들어, 스코어링 목적이 가구 제품을 추천하는 것이면 이미 식탁이 들어있는 장바구니가 다른 식탁을 구매할 가능성은 없습니다. 이러한 경우에 이 옵션을 선택해야 합니다. 반면에, 제품이 신선식품 또는 일회용품(예를 들어, 치즈, 분유 또는 티슈)이면 장바구니에 후향이 이미 존재하는 규칙이 유용할 수 있습니다. 후자의 경우 가장 유용한 옵션은 아래의 **장바구니에서 예측을 검사하지 않음**일 수 있습니다.

장바구니에 예측이 있는지 검사 장바구니에 후향도 있는지 확인하려면 이 옵션을 선택하십시오. 이 접근법은 기존 고객 또는 트랜잭션에 대해 통찰력을 얻고자 할 때 유용합니다. 예를 들어, 최고 리프트의 규칙을 식별한 후 이 규칙에 적합한 고객을 탐색할 수 있습니다.

장바구니에서 예측을 검사하지 않음 스코어링 시 장바구니에 후향이 있는지 여부와 상관 없이 모든 규칙을 포함하려면 선택하십시오.

이 모형의 SQL 생성 데이터베이스에서 데이터를 사용할 때 실행할 SQL 코드를 데이터베이스로 다시 밀어넣어 많은 조작의 성능을 개선할 수 있습니다.

다음 옵션 중 하나를 선택하여 SQL 생성을 수행하는 방법을 지정하십시오.

- **기본값:** 서버 스코어링 어댑터를 사용하거나(설치된 경우) 아니면 프로세스에서 스코어링 스코어링 어댑터가 설치된 데이터베이스에 연결된 경우 스코어링 어댑터 및 연관된 사용자 정의 함수(UDF)를 사용하여 SQL을 생성하고 데이터베이스 내에서 모델의 스코어를 계산합니다. 스코어링 어댑터를 사용할 수 없는 경우 이 옵션은 데이터베이스에서 데이터를 다시 폐치하여 SPSS® Modeler에서 스코어를 계산합니다.

- 데이터베이스 외부 스코어 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

③ 연관 규칙 모델 너깃 요약

연관 규칙 모델 너깃의 요약 탭은 규칙 세트에서 규칙의 지원, 리프트, 신뢰도, 배포성의 최소값 및 최대값과 검색한 규칙 수를 표시합니다.

④ 연관 모델 너깃에서 규칙 세트 생성

연관 모델 너깃(예: Apriori 및 CARMA)은 데이터 스코어를 직접 계산하는 데 사용되거나 먼저 **규칙 세트**라고 하는 규칙의 서브세트를 생성할 수 있습니다. 세분화되지 않은 모델(스코어링을 위해 직접 사용할 수 없음)에 대한 작업을 수행할 때 특히 규칙 세트가 유용합니다. 자세한 정보는 세분화되지 않은 모델의 내용을 참조하십시오.

규칙 세트를 생성하려면 모델 너깃 브라우저의 생성 메뉴에서 **규칙 세트**를 선택하십시오. 규칙을 규칙 세트로 변환하는 경우 다음 옵션을 지정할 수 있습니다.

규칙 세트 이름. 새로 생성된 규칙 세트 노드의 이름을 지정할 수 있습니다.

노드 작성 위치. 새로 생성된 규칙 세트 노드의 위치를 제어합니다. **캔버스**, **GM 팔레트** 또는 **모두**를 선택하십시오.

대상 필드. 생성된 규칙 세트 노드에서 사용할 출력 필드를 판별합니다. 목록에서 단일 출력 필드를 선택합니다.

최소 지원. 생성된 규칙 세트에서 유지할 규칙의 최소 지원을 지정합니다. 지원이 지정된 값보다 적은 규칙은 새 규칙 세트에 포함되지 않습니다.

최소 신뢰도. 생성된 규칙 세트에서 유지할 규칙의 최소 신뢰도를 지정합니다. 신뢰도가 지정된 값보다 적은 규칙은 새 규칙 세트에 포함되지 않습니다.

기본값. 규칙이 실행되지 않는 스코어 계산된 레코드에 지정된 대상 필드의 기본값을 지정할 수 있습니다.

⑤ 필터링된 모델 생성

연관 모델 너깃(예: Apriori, CARMA, 또는 시퀀스 규칙 세트 노드)에서 필터링된 모델을 생성

하려면 모델 너깃 브라우저의 생성 메뉴에서 **필터링된 모델**을 선택하십시오. 그러면 현재 브라우저에 표시된 규칙만 포함하는 서브세트 모델을 작성합니다. **참고:** 세분화되지 않은 모델로 필터링된 모델은 생성할 수 없습니다.

필터링 규칙에 대해 다음 옵션을 지정할 수 있습니다.

새 모델 이름. 새 필터링된 모델 노드 이름을 지정할 수 있습니다.

노드 작성 위치. 새 필터링된 모델 노드 위치를 제어합니다. **캔버스**, **GM 팔레트** 또는 **모두**를 선택하십시오.

규칙 번호 지정. 필터링된 모델에 포함된 규칙 서브세트에서 규칙 ID의 번호를 지정하는 방법을 지정합니다.

- **원래 규칙 ID 번호 보존.** 원래 규칙 번호 지정을 유지보수하려면 선택합니다. 기본적으로 규칙에는 알고리즘에서 발견 순서에 대응하는 ID가 주어집니다. 이 순서는 사용되는 알고리즘에 따라 달라집니다.
- **다음으로 시작하여 연속적으로 규칙 번호 다시 매기기.** 필터링된 규칙에 대해 새 규칙 ID를 지정하려면 선택합니다. 새 ID는 모델 탭의 규칙 브라우저 테이블에 표시되는 정렬 순서(여기에 지정된 번호로 시작)에 기반하여 지정됩니다. 오른쪽을 향하는 화살표를 사용하여 ID의 시작 번호를 지정할 수 있습니다.

⑥ 연관 규칙 스코어링

연관 규칙 모델 너깃을 통해 새 데이터를 실행해서 생성된 스코어는 별도의 필드에 리턴됩니다. 각 예측별로 세 가지 새 필드가 추가됩니다. 여기서, *P*는 예측을 나타내고, *C*는 신뢰도, *I*는 규칙 ID를 나타냅니다. 이 출력 필드 구성은 입력 데이터가 트랜잭션 또는 표 형식인지에 따라 다릅니다. 이 형식의 개요는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

예를 들어, 다음 세 가지 규칙을 기준으로 하여 예측을 생성하는 모델을 사용해서 장바구니 데이터를 스코어링 중이라고 가정하십시오.

```
Rule_15 bread&wine -> meat (confidence 54%)
Rule_22 cheese -> fruit (confidence 43%)
Rule_5 bread&cheese -> frozveg (confidence 24%)
```

표 형식 데이터. 표 형식 데이터의 경우 세 개의 예측(기본값이 3)이 단일 레코드에 리턴됩니다.

표 1. 표 형식의 스코어

ID	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat	0.54	15	fruit	0.43	22	frozveg	.24	5

트랜잭션 데이터. 트랜잭션 데이터의 경우 각 예측마다 별도의 레코드가 생성됩니다. 예측이 여전히 개별 열에 추가되지만 스코어는 계산할 때 리턴됩니다. 이로 인해 아래의 샘플 출력에 표시된 대로 레코드의 예측이 불완전하게 됩니다. 두 번째 및 세 번째 예측(P2 및 P3)이 연관된 신뢰도 및 규칙 ID와 함께 첫 번째 레코드에서 공백입니다. 하지만 스코어가 리턴될 때 마지막 레코드에 세 가지 모든 예측이 포함됩니다.

표 2. 트랜잭션 형식의 스코어

ID	항목	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread	meat	0.54	14	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$
Fred	cheese	meat	0.54	14	fruit	0.43	22	\$null\$	\$null\$	\$null\$
Fred	wine	meat	0.54	14	fruit	0.43	22	frozveg	0.24	5

보고 또는 배포 용도로 완전한 예측만 포함하려면 선택 노드를 사용하여 완전한 레코드를 선택하십시오.

참고: 이 예에 사용된 필드 이름은 명확한 표현을 위해 축약되었습니다. 실제 사용 중에는 연관 모델의 결과 필드가 다음 표에 표시된 대로 이름 지정됩니다.

표 3. 연관 모델의 결과 필드 이름

새 필드	예 필드 이름
예측	\$A-TRANSACTION_NUMBER-1
신뢰도(또는 기타 기준)	\$AC-TRANSACTION_NUMBER-1
규칙 ID	\$A-Rule_ID-1

여러 후항이 있는 규칙

CARMA 알고리즘은 여러 후항이 있는 규칙을 허용합니다. 예를 들어, 다음과 같습니다.

```
bread -> wine&cheese
```

“two-headed” 규칙을 스코어링할 때에는 예측이 다음 표에 표시된 형식으로 리턴됩니다.

표 4. 복수 후항이 있는 예측을 포함한 결과 스코어링

ID	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat& veg	0.54	16	fruit	0.43	22	frozveg	.24	5

일부 경우 배포 전에 이러한 스코어를 분할해야 할 수 있습니다. 여러 후항이 있는 예측을 분할하려면 CLEM 문자열 함수를 사용하여 필드를 구문 분석해야 합니다.

⑦ 연관 모델 배포

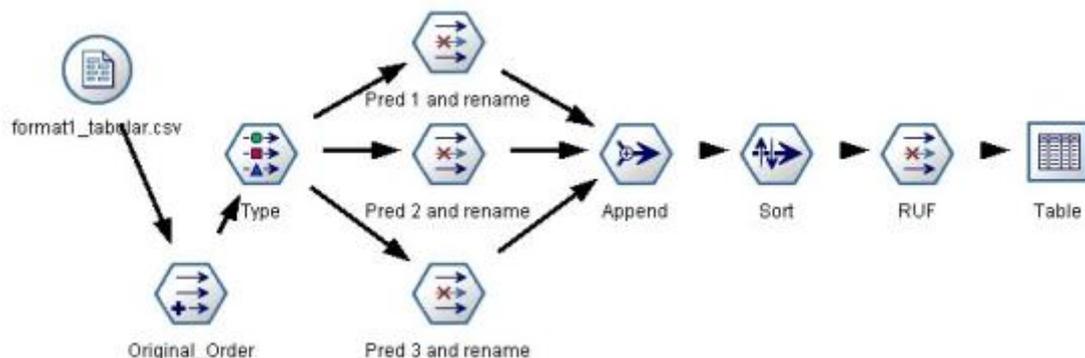
연관 모델을 스코어링할 때 예측 및 신뢰도는 별도의 열에 출력됩니다(여기서, P는 예측을 나타내고, C는 신뢰도, I는 규칙 ID를 나타냄). 이는 입력 데이터가 표 또는 트랜잭션 형식인 경우입니다. 자세한 정보는 연관 규칙 스코어링의 내용을 참조하십시오.

배포를 위한 스코어를 준비하는 경우 애플리케이션이 출력 데이터를 열이 아닌 행에 예측이 있는 형식(각 행별로 예측이 하나씩, 이를 때로 "till-roll" 형식이라 함)으로 전치하도록 요구함을 알 수 있습니다.

표 스코어 전치

다음 단계에 설명된 대로 IBM® SPSS® Modeler의 단계를 조합하여 표 스코어를 열에서 행으로 전치할 수 있습니다.

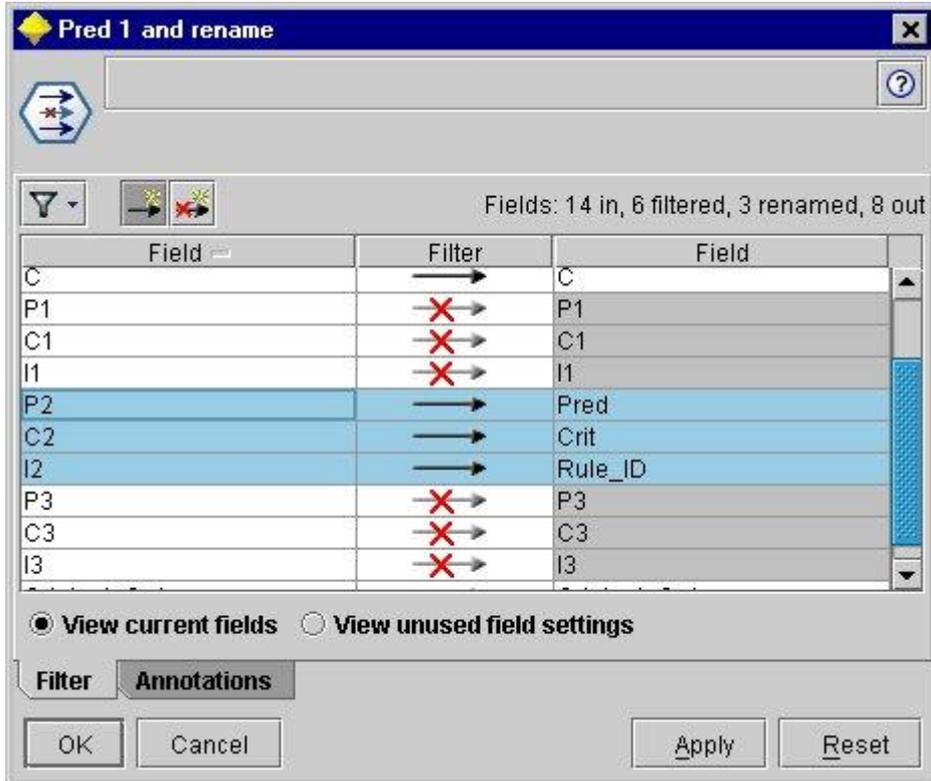
그림 1. 표 형식 데이터를 till-roll 형식으로 전치하는 데 사용되는 예 스트림



1. 파생 노드의 @INDEX 함수를 사용하여 현재 예측 순서를 확인하고 Original_order와 같은 새 필드에 이 표시기를 저장하십시오.

2. 모든 필드가 인스턴스화되도록 유형 노드를 추가하십시오.
3. 필터 노드를 사용하여 기본 예측, 신뢰도, ID 필드(P1, C1, I1)의 이름을 나중에 레코드를 붙여쓰는 데 사용할 *Pred*, *Crit*, *Rule_ID*와 같은 공통 필드로 변경하십시오. 생성된 각 예측마다 필터 노드가 하나씩 필요합니다.

그림 2. 예측 2의 필드 이름 변경 중 예측 1 및 3의 필드 필터링.



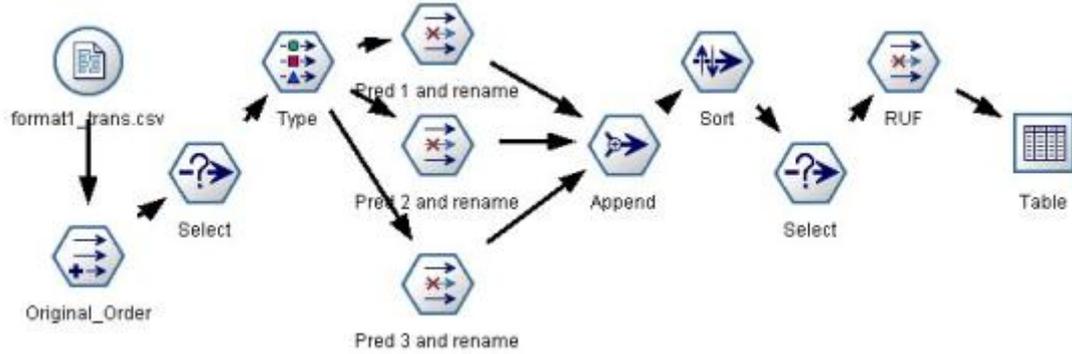
4. 붙여쓰기 노드를 사용하여 공유 *Pred*, *Crit*, *Rule_ID*의 값을 붙여쓰십시오.
5. 정렬 노드를 첨부하여 필드 *Original_order*의 레코드를 오름차순으로 정렬하고 신뢰도, 리프트, 지원과 같은 기준별로 예측을 정렬하는 데 사용되는 필드인 *Crit*의 레코드를 내림차순으로 정렬하십시오.
6. 또 다른 필터 노드를 사용하여 출력에서 필드 *Original_order*를 필터링하십시오.

이 때 데이터는 배포할 준비가 됩니다.

트랜잭션 스코어 전치

트랜잭션 스코어 전치의 프로세스는 유사합니다. 예를 들어, 아래에 표시된 스트림은 배포에 필요한 각 행마다 단일 예측이 있는 형식으로 스코어를 전치합니다.

그림 3. 트랜잭션 데이터를 till-roll 형식으로 전치하는 데 사용되는 예 스트림



두 개의 선택 노드가 추가되었고 프로세스는 이전에 표 형식 데이터에 설명한 것과 동일합니다.

- 첫 번째 선택 노드는 규칙 ID를 인접한 레코드에 비교하는 데 사용되며 고유 또는 정의되지 않은 레코드만 포함합니다. 이 선택 노드는 다음과 같이 CLEM 표현식을 사용하여 레코드를 선택합니다. $ID \neq @OFFSET(ID, -1)$ or $@OFFSET(ID, -1) = undef$
- 두 번째 선택 노드는 Rule_ID의 값이 널값인 규칙이나 관련이 없는 규칙을 삭제하는 데 사용됩니다. 이 선택 노드는 다음 CLEM 표현식을 사용하여 레코드를 삭제합니다. $not(@NULL(Rule_ID))$.

배포를 위한 스코어 전치에 대한 자세한 정보는 기술 지원에 문의하십시오.

(5) 시퀀스 노드

시퀀스 노드는 bread -> cheese 형식으로 순차 또는 시간 중심의 데이터에서 패턴을 검색합니다. 시퀀스의 요소는 단일 트랜잭션을 구성하는 항목 세트입니다. 예를 들어, 상점에 가서 빵과 우유를 구입하고 며칠 후 다시 상점에서 치즈를 구입한 경우 이 사람의 구매 활동은 두 개 항목 세트로 표시됩니다. 첫 번째 항목 세트는 빵과 우유를 포함하고 두 번째 항목 세트는 치즈를 포함합니다. 시퀀스는 예측 가능한 순서로 발생하는 경향이 있는 항목 세트의 목록입니다. 시퀀스 노드는 빈번한 시퀀스를 발견하고 예측을 수행하는 데 사용할 수 있는 생성된 모델 노드를 작성합니다.

요구사항. 시퀀스 규칙 세트를 작성하려면 ID 필드, 선택적 시간 필드, 하나 이상의 콘텐츠 필드를 지정해야 합니다. 이러한 설정은 모델링 노드의 필드 탭에서 수행해야 합니다. 업스트림 유형 노드에서는 읽을 수 없습니다. ID 필드에는 역할 또는 측정 수준이 있을 수 있습니다. 시간 필드를 지정하면 역할을 보유할 수 있지만, 저장 공간은 숫자, 날짜, 시간 또는 시간소인이어야 합니다. 시간 필드를 지정하지 않은 경우 시퀀스 노드는 시간 값으로 행 번호를 사용하여 함축된 시간소인을 사용합니다. 콘텐츠 필드는 임의의 측정 수준 및 역할을 보유할 수 있지만, 모든 콘텐츠 필드는 유형이 동일해야 합니다. 숫자인 경우 정수 범위(실수가 아님)여야 합니다.

강도. 시퀀스 노드는 시퀀스를 찾는 효율적인 2단계 방법을 사용하는 CARMA 연관 규칙 알고리즘에 기반합니다. 또한 시퀀스 노드에서 작성하여 생성된 모델 노드를 데이터 스트림에 삽입하여 예측을 작성할 수 있습니다. 생성된 모델 노드는 특정 시퀀스의 발견과 계산, 그리고 특정 시퀀스에 기반한 예측을 수행하기 위해 슈퍼 노드를 생성할 수 있습니다.

① 시퀀스 노드 필드 옵션

시퀀스 노드를 실행하기 전에 시퀀스 노드의 필드 탭에서 ID 및 콘텐츠 필드를 지정해야 합니다. 시간 필드를 사용하려면 여기에서 해당 항목도 지정해야 합니다.

ID 필드. 목록에서 ID 필드를 선택합니다. 숫자 또는 기호 필드를 ID 필드로 사용할 수 있습니다. 이 필드의 각 고유한 값은 분석의 특정 단위를 표시해야 합니다. 예를 들어, 장바구니 애플리케이션에서 각 ID는 단일 고객을 나타낼 수 있습니다. 웹 로그 분석 애플리케이션의 경우 각 ID는 컴퓨터(IP 주소별) 또는 사용자(로그인 데이터별)를 나타낼 수 있습니다.

- **연속적 ID.** ID가 동일한 모든 레코드를 데이터 스트림에서 함께 그룹화하도록 데이터를 사전 정렬한 경우 이 옵션을 선택하여 처리 속도를 높입니다. 데이터가 사전 정렬되지 않았거나 정렬 여부가 확실하지 않은 경우 이 옵션을 선택하지 않은 상태로 두면 시퀀스 노드가 데이터를 자동으로 정렬합니다.

참고: 데이터가 정렬되지 않은 상태에서 이 옵션을 선택하면 시퀀스 모델에서 유효하지 않은 결과가 발생할 수 있습니다.

시간 필드. 이벤트 시간을 표시하기 위해 데이터에서 필드를 사용하려면 **시간 필드 사용**을 선택하고 사용할 필드를 지정하십시오. 시간 필드는 숫자, 날짜, 시간 또는 시간소인이어야 합니다. 시간 필드를 지정하지 않은 경우 레코드는 데이터 소스에서 순차적으로 도달한다고 가정하며, 레코드 번호는 시간 값으로 사용됩니다(첫 번째 레코드는 "1" 시간에, 두 번째는 "2" 시간에 나타나는 방식).

콘텐츠 필드. 모델의 콘텐츠 필드를 지정합니다. 이 필드는 시퀀스 모델링에서 관심이 있는 항목을 포함합니다.

시퀀스 노드는 표 형식 또는 트랜잭션 형식의 데이터를 처리할 수 있습니다. 트랜잭션 데이터를 포함하는 다중 필드를 사용하는 경우 특정 레코드에서 이 필드에 지정된 항목은 단일 시간소인을 포함하는 단일 트랜잭션에서 찾은 항목을 나타낸다고 가정합니다. 자세한 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

파티션. 이 필드에서는 모델 작성에 대한 훈련, 검정, 검증 단계에 대한 별도의 샘플로 데이터를 파티션하는 데 사용되는 필드를 지정할 수 있습니다. 한 표본을 사용하여 모델을 생성하고 다른 표본을 사용하여 이를 검정함으로써 현재 데이터에 유사한 보다 큰 데이터 세트로 모델이 일반

화되는 정도를 잘 표시할 수 있습니다. 유형 또는 파티션 노드를 사용하여 다중 파티션 필드를 정의한 경우 파티셔닝을 사용하는 각 모델링 노드에 있는 필드 탭에서 단일 파티션 필드를 선택해야 합니다. (하나의 파티션만 존재하는 경우 파티션이 사용될 때마다 자동으로 사용됩니다.) 또한 분석에서 선택한 파티션을 적용하려면 노드의 모델 옵션 탭에서 파티셔닝을 사용 가능하게 해야 합니다. (이 옵션을 선택 취소하면 필드 설정을 변경하지 않고도, 파티셔닝을 사용하지 않을 수 있습니다.)

② 시퀀스 노드 모델 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

최소 규칙 지원(%) 지원 기준을 지정할 수 있습니다. **규칙 지원**은 전체 시퀀스를 포함하는 훈련 데이터에서 ID 비율을 나타냅니다. 보다 일반적인 시퀀스에 초점을 맞추려면 이 설정을 늘립니다.

최소 규칙 신뢰도(%) 시퀀스 세트에서 시퀀스를 유지하기 위해 신뢰도 기준을 지정할 수 있습니다. **신뢰도**는 규칙에서 예측을 수행하는 모든 ID 중 올바른 예측에 성공한 ID의 퍼센트를 나타냅니다. 이는 전체 시퀀스를 찾은 ID 수를 훈련 데이터에 기반하여 전향을 찾은 ID 수로 나누어 계산합니다. 신뢰도가 지정된 기준보다 낮은 시퀀스는 삭제됩니다. 시퀀스 또는 관련도가 낮은 시퀀스가 너무 많으면 이 설정을 늘리십시오. 확보한 시퀀스가 너무 적은 경우 이 설정을 줄이십시오.

참고: 필요한 경우 값을 강조 표시하고 고유한 값을 입력할 수 있습니다. 신뢰도 값을 1.0 미만으로 줄이면 프로세스의 사용 가능한 메모리가 많이 필요한 점 외에 규칙을 작성하는 데 극단적으로 오랜 시간이 걸릴 수 있음에 유의하십시오.

최대 시퀀스 크기 시퀀스에서 개별 항목의 최대 수를 설정할 수 있습니다. 관심이 있는 시퀀스가 비교적 짧으면 이 설정을 줄여 시퀀스 세트 작성 속도를 높일 수 있습니다.

스트림에 추가할 예측 결과로 생성된 모델 노드에서 스트림에 추가할 예측 수를 지정합니다. 추가 정보는 시퀀스 모델 너깃의 내용을 참조하십시오.

③ 시퀀스 노드 고급 옵션

시퀀스 노드의 작업에 대한 자세한 지식을 가진 사용자는 다음 고급 옵션을 통해 모델 작성 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

최대 기간 설정. 이 옵션을 선택하면 시퀀스는 지정된 값 이하인 기간(첫 번째 항목 세트와 마지막 항목 세트 사이의 시간)으로 제한됩니다. 시간 필드를 지정하지 않으면 기간은 원시 데이터의 행(레코드) 관점으로 표현됩니다. 사용되는 시간 필드가 시간, 날짜 또는 시간소인 필드인 경우 기간은 초 단위로 표현됩니다. 숫자 필드의 경우 기간은 필드 자체와 동일한 단위로 표현됩니다.

가지치기 값 설정. 시퀀스 노드에 사용된 CARMA 알고리즘은 주기적으로 메모리를 보존하는 프로세스 중에 잠재적 항목 세트의 목록에서 자주 사용하지 않는 항목 세트를 제거(가지치기)합니다. 이 옵션을 선택하여 가지치기 빈도를 조정하십시오. 지정된 번호는 가지치기 빈도를 판별합니다. 더 작은 값을 입력하여 알고리즘의 메모리 요구 사항을 줄이거나(그러나 잠재적으로 필요한 훈련 시간이 늘어남) 더 큰 값을 입력하여 훈련 속도를 높이십시오(그러나 잠재적으로 메모리 요구 사항이 늘어남).

메모리에서 최대 시퀀스 설정. 이 옵션을 선택하면 CARMA 알고리즘은 지정된 시퀀스 번호로 모델을 작성하는 동안 후보 시퀀스의 메모리 보관을 제한합니다. IBM® SPSS® Modeler에서 시퀀스 모델 작성 중에 너무 많은 메모리를 사용하는 경우 이 옵션을 선택합니다. 여기에 지정한 최대 시퀀스 값은 모델을 작성할 때 내부적으로 추적되는 후보 시퀀스의 번호입니다. 이 번호는 최종 모델에서 예상되는 시퀀스 번호보다 훨씬 더 커야 합니다.

항목 세트 사이에서 차이 제한. 이 옵션을 사용하면 항목 세트를 구분하는 시간 구간에 제약조건을 지정할 수 있습니다. 이 옵션을 선택하면 시간 차이가 여기에서 지정한 **최소 차이**보다 작거나 **최대 차이**보다 큰 항목 세트는 시퀀스의 일부를 구성한다고 간주되지 않습니다. 이 옵션을 사용하여 매우 짧은 기간에 발생하거나 긴 시간 구간을 포함하는 시퀀스를 계산하지 않도록 합니다.

참고: 사용되는 시간 필드가 시간, 날짜 또는 시간소인 필드인 경우 시간 차이는 초 단위로 표현됩니다. 숫자 필드인 경우 시간 차이는 시간 필드와 동일한 단위로 표현됩니다.

예를 들어, 다음 트랜잭션 목록을 고려하십시오.

표 1. 트랜잭션 목록 예

ID	시간	컨텐츠
1001	1	apples
1001	2	bread
1001	5	cheese
1001	6	dressing

최소 차이가 2로 설정된 해당 데이터에서 모델을 작성하는 경우 다음 시퀀스를 가져옵니다.

```

apples -> cheese
apples -> dressing
bread -> cheese
bread -> dressing

```

apples -> bread와 같은 시퀀스는 확인할 수 없습니다. apples 및 bread 사이의 차이가 최소 차이보다 작기 때문입니다. 마찬가지로, 다음과 같은 대체 데이터를 고려하십시오.

표 2. 트랜잭션 목록 예

ID	시간	컨텐츠
1001	1	apples
1001	2	bread
1001	5	cheese
1001	20	dressing

최대 차이를 10으로 설정한 경우 dressing을 포함하는 시퀀스를 확인할 수 없습니다. cheese 및 dressing 사이의 차이가 너무 커서 동일한 시퀀스의 파트로 간주할 수 없기 때문입니다.

④ 시퀀스 모델 너깃

시퀀스 모델 너깃은 시퀀스 노드에서 검색된 특정 출력 필드에서 찾은 시퀀스를 나타내고, 이를 스트림에 추가하여 예측을 생성할 수 있습니다.

시퀀스 노드를 포함하는 스트림을 실행할 때 노드는 시퀀스 모델에서 데이터로 예측 및 각 예측의 연관된 신뢰도 값을 포함하는 한 쌍의 필드를 추가합니다. 기본적으로 상위 3개의 예측(그리고 연관된 신뢰도 값)을 포함하는 3개의 필드 쌍이 추가됩니다. 스트림에 모델 너깃 추가 후 설정 탭에서, 그리고 작성 시 시퀀스 노드 모델 옵션을 설정하여 모델을 작성할 때 생성되는 예측 수를 변경할 수 있습니다. 자세한 정보는 시퀀스 모델 너깃 설정의 내용을 참조하십시오.

새 필드 이름은 모델 이름에서 파생됩니다. 필드 이름은 예측 필드의 경우 *\$S-sequence-n*(여기서 *n*은 *n*번째 예측을 나타냄)이고 신뢰도 필드의 경우 *\$SC-sequence-n*입니다. 일련의 여러 시퀀스 규칙 노드를 포함하는 스트림에서 새 필드 이름은 서로를 구별하도록 접두문자에 숫자를 포함합니다. 스트림에서 첫 번째 시퀀스 세트 노드는 일상적인 이름을 사용하고, 두 번째 노드는 *\$S1-* 및 *\$SC1-*로 시작하는 이름을 사용하고, 세 번째 노드는 *\$S2-*, *\$SC2-* 등으로 시작하는 이름을 사용합니다. 예측은 신뢰도 순서로 표시되므로, *\$S-sequence-1*은 신뢰도가 가장 높은 예측을, *\$S-sequence-2*는 그 다음으로 신뢰도가 높은 예측을 포함하는 식입니다. 사용 가능한 예

측 수가 요청된 예측 수보다 작은 레코드의 경우 나머지 예측은 $\$null\$\$ 의 값을 포함합니다. 예를 들어, 유일한 2개 예측이 특정 레코드에서 수행될 수 있는 경우 $\$S-sequence-3\$ 및 $\$SC-sequence-3\$ 의 값은 $\$null\$\$ 이 됩니다.

각 레코드에서 모델의 규칙은 현재 레코드 및 ID가 동일하고 이전 시간소인의 이전 레코드를 포함하여 지금까지 현재 ID에서 처리된 트랜잭션 세트와 비교됩니다. 이 트랜잭션 세트에 적용되는 신뢰도 값이 가장 높은 k 규칙은 레코드에 대한 k 예측을 생성하는 데 사용됩니다. 여기서, k 는 스트림에 모델을 추가한 후에 설정 탭에 지정된 예측 수입니다. (다중 규칙이 트랜잭션 세트에 대해 동일한 결과를 예측하는 경우 신뢰도가 가장 높은 규칙만 사용합니다.) 자세한 정보는 시퀀스 모델 너짓 설정의 내용을 참조하십시오.

연관 규칙 모델의 다른 유형과 마찬가지로 데이터 형식은 시퀀스 모델을 작성할 때 사용되는 형식과 매치해야 합니다. 예를 들어, 표 형식 데이터를 사용하여 작성된 모델은 표 형식 데이터 스코어를 계산하는 데만 사용할 수 있습니다. 자세한 정보는 연관 규칙 스코어링의 내용을 참조하십시오.

참고: 스트림에서 생성된 시퀀스 세트 노드를 사용하여 데이터를 스코어링하는 경우 모델 작성 시 선택한 공차 또는 차이 설정은 스코어링 목적에서 무시됩니다.

시퀀스 규칙에서 예측

노드는 모델을 작성하는 데 사용된 시간소인 필드가 없는 경우 시간에 종속된 방식(또는 순서에 종속된 방식)으로 레코드를 처리합니다. 레코드는 ID 필드 및 시간소인 필드(있는 경우)로 정렬해야 합니다. 그러나 예측은 추가된 레코드의 시간소인에 연결되지 않습니다. 단순히 현재 레코드까지 현재 ID의 트랜잭션 히스토리가 주어진 경우 *미래의 특정 포인트*에 나타날 가능성이 높은 항목을 참조합니다.

각 레코드의 예측은 해당 레코드의 트랜잭션에 반드시 의존하지 않아도 됩니다. 현재 레코드의 트랜잭션이 특정 규칙을 트리거하지 않으면 현재 ID의 이전 트랜잭션에 기반하여 규칙이 선택됩니다. 즉, 현재 레코드가 시퀀스에 유용한 예측 정보를 추가하지 않는 경우 이 ID의 마지막 유용한 트랜잭션에서 예측이 현재 레코드로 이월됩니다.

예를 들어, 단일 규칙을 포함하는 시퀀스 모델이 있다고 가정합니다.

```
Jam -> Bread (0.66)
```

그리고 다음 레코드를 전달합니다.

ID	구매	예측
001	jam	bread
001	milk	bread

첫 번째 레코드는 예상대로 *bread*의 예측을 생성합니다. 두 번째 레코드는 *bread*의 예측도 포함합니다. *jam*과 뒤에 나오는 *milk*에 대한 규칙이 없기 때문에 *milk* 트랜잭션은 유용한 정보를 추가하지 않고, 규칙 Jam -> Bread는 계속 적용됩니다.

새 노드 생성

생성 메뉴에서는 시퀀스 모델에 기반하여 새 슈퍼 노드를 작성할 수 있습니다.

- **규칙 슈퍼 노드.** 스코어가 계산된 데이터에서 시퀀스의 발생을 발견하고 계산할 수 있는 슈퍼 노드를 작성합니다. 규칙이 선택되지 않으면 이 옵션은 사용할 수 없습니다. 자세한 정보는 시퀀스 모델 너깃에서 규칙 슈퍼 노드 생성 주제를 참조하십시오.
- **모델을 팔레트로.** 모델을 모델 팔레트로 리턴합니다. 이 옵션은 동료가 모델 자체가 아닌 모델을 포함한 스트림을 보냈을 경우에 유용합니다.

가. 시퀀스 모델 너깃 세부사항

시퀀스 모델 너깃의 모델 탭에서는 알고리즘에서 추출된 규칙을 표시합니다. 표의 각 행은 규칙을 나타내고 첫 번째 열에는 전항(규칙의 "if" 부분)이 나오고, 두 번째 열에는 후항(규칙의 "then" 부분)이 나옵니다.

각 규칙은 다음 형식으로 표시됩니다.

전항	후항
beer and cannedveg	beer
fish fish	fish

첫 번째 규칙 예는 동일한 트랜잭션에서 "beer" 및 "cannedveg"를 포함하는 ID의 경우 "beer"가 뒤에 나올 수 있습니다. 두 번째 규칙 예는 한 트랜잭션에서 "fish"를 포함했고, 다른 트랜잭션에서도 "fish"를 포함하는 ID의 경우 다음에도 "fish"가 나올 수 있으므로 해석됩니다. 첫 번째 규칙에서 beer 및 cannedveg는 동시에 구매했고, 두 번째 규칙에서 fish는 별도의 2개 트랜잭션으로 구매했다는 점을 참고하십시오.

정렬 메뉴. 도구 모음의 정렬 메뉴 단추는 규칙 정렬을 제어합니다. 정렬 방향 단추(위로 또는 아래로 화살표)를 사용하여 정렬 방향(오름차순 또는 내림차순)을 변경할 수 있습니다.

다음 기준에 따라 규칙을 정렬할 수 있습니다.

- 지원 %
- 신뢰도
- 규칙 지원 %
- 후향
- 첫 번째 전향
- 마지막 전향
- 항목 수(전향)

예를 들어 다음 표에서는 항목 수를 내림차순으로 정렬합니다. 전향 세트에 여러 항목을 포함하는 규칙은 항목이 더 적은 규칙보다 앞에 옵니다.

표 2. 항목 수로 정렬된 규칙	
전향	후향
beer and cannedveg and frozenmeal	frozenmeal
beer and cannedveg	beer
fish fish	fish
softdrink	softdrink

기준 메뉴 표시/숨기기. 기준 메뉴 표시/숨기기 단추(눈금 아이콘)는 규칙 표시에 대한 옵션을 제어합니다. 다음 표시 옵션을 사용할 수 있습니다.

- **인스턴스**는 전체 시퀀스(전향 및 후향 모두)가 나타나는 고유 ID의 번호에 대한 정보를 표시합니다. (이는 인스턴스 번호가 전향만 적용하는 ID의 번호를 참조하는 연관 모델과는 다르게 유의하십시오.) 예를 들어, 규칙이 bread -> cheese인 경우 *bread* 및 *cheese*를 모두 포함하는 학습 데이터의 ID 번호는 **인스턴스**로 참조됩니다.
- **지원**은 전향이 참인 학습 데이터에서 ID의 비율을 표시합니다. 예를 들어, 학습 데이터 중 50%가 전향 *bread*를 포함하면 bread -> cheese 규칙에 대한 지원은 50%입니다. (전향 모델과 달리, 지원은 앞서 언급한 대로 인스턴스 수에 기반하지 *않습니다*.)
- **신뢰도**는 규칙에서 예측을 수행하는 모든 ID 중 올바른 예측에 성공한 ID의 퍼센트를 표시합니다. 이는 전체 시퀀스를 찾은 ID 수를 학습 데이터에 기반하여 전향을 찾은 ID 수로 나누어 계산합니다. 예를 들어, 학습 데이터의 50%가 cannedveg를 포함하지만(전향 지원을 표시함) 20%만 cannedveg 및 frozenmeal 둘 다를 포함하는 경우 cannedveg -> frozenmeal 규칙의 신뢰도는 Rule Support / Antecedent Support이거나 이 경우 40%입니다.

- 시퀀스 모델에서 **규칙 지원**은 인스턴스에 기반하며, 전체 규칙, 전항, 무항이 참인 학습 데이터의 비율을 표시합니다. 예를 들어, 학습 데이터 중 20%가 *bread* 및 *cheese*를 모두 포함하는 경우 규칙 *bread* -> *cheese*의 규칙 지원은 20%입니다.

비율은 총 트랜잭션이 아닌 유효한 트랜잭션(하나 이상의 관측 항목 또는 참 값을 포함하는 트랜잭션)에 기반합니다. 유효하지 않은 트랜잭션(항목이나 참 값이 없는 트랜잭션)은 이 계산에서 삭제됩니다.

필터 단추. 메뉴의 필터 단추(갈때기 아이콘)는 활성 규칙 필터가 표시되는 패널을 보여주기 위해 대화 상자의 맨 아래까지 확장됩니다. 필터는 모델 탭에 표시되는 규칙 번호의 범위를 좁히는 데 사용됩니다.

그림 1. 필터 단추



필터를 작성하려면 펼쳐진 패널의 오른쪽에 있는 필터 아이콘을 클릭하십시오. 그러면 규칙 표시에 대한 제약조건을 지정할 수 있는 별도의 대화 상자가 열립니다. 필터 단추는 종종 생성 메뉴와 함께 사용되어 먼저 규칙을 필터링한 후 규칙의 서브셋을 포함한 모델을 생성함에 유의하십시오. 자세한 정보는 아래 규칙의 필터 지정의 내용을 참조하십시오.

나. 시퀀스 모델 너깃 설정

시퀀스 모델 너깃의 설정 탭에서는 모델의 스코어링 옵션을 표시합니다. 이 탭은 스코어링을 위해 스트림 캔버스에 모델을 추가한 후에만 사용 가능합니다.

최대 예측 수. 각 바구니 항목 집합에 포함되는 최대 예측 수를 지정합니다. 이 트랜잭션 세트에 적용되는 최상위 신뢰도 값을 가지고 있는 규칙이 지정된 한계까지 레코드에 대한 예상값을 생성하기 위해 사용됩니다.

다. 시퀀스 모델 너깃 요약

시퀀스 규칙 모델 너깃의 요약 탭에서는 규칙에서 지원 및 신뢰도의 최소값 및 최대값과 검색된 규칙 수를 표시합니다. 이 모델링 노드에 첨부된 분석 노드를 실행하면 해당 분석의 정보도 이 섹션에 표시됩니다.

자세한 정보는 모델 너깃 찾아보기 주제를 참조하십시오.

라. 시퀀스 모델 너깃에서 규칙 수퍼 노드 생성

시퀀스 규칙에 기반하여 규칙 수퍼 노드를 생성하려면 다음을 수행하십시오.

1. 시퀀스 규칙 모델 너깃의 모델 탭에 있는 테이블에서 행을 클릭하여 원하는 규칙을 선택하십시오.
2. 규칙 브라우저 메뉴에서 다음을 선택하십시오.

생성 > 규칙 수퍼 노드

중요: 생성된 수퍼 노드를 사용하려면 수퍼 노드로 전달하기 전에 ID 필드와 시간 필드(있는 경우)로 데이터를 정렬해야 합니다. 수퍼 노드는 정렬되지 않은 데이터에서 시퀀스를 적절히 발견하지 못합니다.

규칙 수퍼 노드 생성을 위해 다음 옵션을 지정할 수 있습니다.

발견. 수퍼 노드로 전달된 데이터에서 매치를 정의하는 방법을 지정합니다.

- **전항만.** 수퍼 노드는 후항도 찾았는지 여부에 상관없이 ID가 동일한 레코드 세트 내에서 올바른 순서로 정렬된 선택한 규칙의 전항을 찾을 때마다 매치를 식별합니다. 이 경우 원래 시퀀스 모델링 노드에서 시간소인 공차 또는 항목 차이 제한조건 설정을 고려하지 않습니다. 마지막 전항 항목 세트가 스트림에서 발견되고(그리고 적절한 순서로 다른 모든 전항이 발견되면) 현재 ID의 모든 후속 레코드는 아래 선택된 요약을 포함합니다.
- **전체 시퀀스.** 수퍼 노드는 ID가 동일한 레코드 세트 내에서 올바른 순서로 선택한 규칙의 전항 및 후항을 찾을 때마다 매치를 식별합니다. 이 경우 원래 시퀀스 모델링 노드에서 시간소인 공차 또는 항목 차이 제한조건 설정을 고려하지 않습니다. 스트림에서 후항이 발견되면(그리고 올바른 순서로 모든 전항도 발견됨) 현재 레코드와 현재 ID의 모든 후속 레코드는 아래 선택된 요약을 포함합니다.

표시. 규칙 수퍼 노드 출력에서 데이터에 매치에 대한 요약을 추가하는 방법을 제어합니다.

- **첫 번째 발생의 후항 값.** 데이터에 추가된 값은 매치의 첫 번째 발생에 기반하여 예측된 후항 값입니다. 값은 이름이 *rule_n_consequent*인 새 필드로 추가됩니다. 여기서 *n*은 규칙 번호입니다(스트림에서 규칙 수퍼 노드의 작성 순서에 기반함).
- **첫 번째 발생의 참 값.** 데이터에 추가된 값은 ID에 대해 하나 이상의 매치가 있으면 참, 매치가 없으면 거짓입니다. 값은 이름이 *rule_n_flag*인 새 필드로 추가됩니다.
- **발생 수.** 데이터에 추가된 값은 ID와 매치하는 수입니다. 값은 이름이 *rule_n_count*인 새 필드로 추가됩니다.
- **규칙 번호.** 추가된 값은 선택된 규칙의 규칙 번호입니다. **규칙 번호**는 스트림에 수퍼 노드를 추가하는 순서에 기반하여 지정됩니다. 예를 들어, 첫 번째 규칙 수퍼 노드는 *규칙 1*로 고려되고, 두 번째 규칙 수퍼 노드는 *규칙 2*등으로 고려됩니다. 이 옵션은 스트림에서 다중 규칙 수퍼 노드를 포함하는 경우 가장 유용합니다. 값은 이름이 *rule_n_number*인 새 필드로 추가됩니다.

- 신뢰도 그림 포함. 이 옵션을 선택하면 선택한 요약과 함께 데이터 스트림에 규칙 신뢰도를 추가합니다. 값은 이름이 *rule_n_confidence*인 새 필드로 추가됩니다.

(6) 연관 규칙 노드

연관 규칙은 다음 양식의 명령문입니다.

```
if condition(s) then prediction(s)
```

예를 들어, "면도기와 애프터쉐이브 로션을 구매하는 고객은 다음 번 구매 시에 면도용 크림을 구매할 수 있습니다." 연관 규칙 노드는 데이터에서 규칙 세트를 추출하여, 최상의 정보 내용을 가지고 있는 규칙을 골라냅니다. 연관 규칙 노드는 Apriori 노드와 아주 유사하지만, 일부 주목할 만한 차이가 있습니다.

- 연관 규칙 노드는 트랜잭션 데이터를 처리할 수 없습니다.
- 연관 규칙 노드는 목록 저장 유형과 요약도표 측정 수준에 있는 데이터를 처리할 수 있습니다.
- 연관 규칙 노드는 IBM® SPSS® Analytic Server와 함께 사용할 수 있습니다. 이는 확장성을 제공하고 빅 데이터를 처리하고 빠른 병렬 처리를 이용할 수 있는 수단을 제공합니다.
- 연관 규칙 노드는 생성되는 규칙 수를 제한하는 기능과 같은 추가 설정을 제공하여, 처리 속도를 증가시킵니다.
- 모델 너깃의 출력은 출력 뷰어에 표시됩니다.

참고: 연관 규칙 노드는 IBM SPSS Collaboration and Deployment Services에서 모델 평가 또는 챔피언 챌린저 단계를 지원하지 않습니다.

참고: 필드 유형이 플래그이면 모델을 작성할 때 연관 규칙 노드에서 빈 레코드가 무시됩니다. 비어 있는 레코드는 모델 작성에 사용된 모든 필드의 값이 거짓인 레코드입니다.

연관 규칙 사용 작업 예제를 보여주고(*geospatial_association.str*) *InsuranceData.sav*, *CountyData.sav* 및 *ChicagoAreaCounties.shp* 데이터 파일을 참조하는 스트림은 IBM SPSS Modeler 설치의 Demos 디렉토리에서 사용 가능합니다. Windows 시작 메뉴에 있는 IBM SPSS Modeler 프로그램 그룹에서 Demos 디렉토리에 액세스할 수 있습니다. *geospatial_association.str* 파일은 *streams* 디렉토리에 있습니다.

① 연관 규칙 - 필드 옵션

필드 탭에서, 이미 이전 유형 노드와 같은, 업스트림 노드에서 정의된 필드 역할 설정을 사용할 것인지 여부를 선택하거나, 수동으로 필드 할당을 수행합니다.

사전 정의된 역할 사용

이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(예: 목표 또는 예측자)을 사용합니다. 입력 역할이 있는 필드는 조건인 것으로 간주되고, 목표 역할이 있는 필드가 예측값인 것으로 간주되며, 입력 및 목표로 사용되는 해당 필드는 두 역할 모두를 가지고 있는 것으로 간주됩니다.

사용자 정의 필드 할당 사용

이 화면에서 수동으로 목표, 예측자 및 기타 역할을 지정하려면 이 옵션을 선택하십시오.

필드

사용자 정의 필드 할당 사용을 선택한 경우, 이 목록의 항목을 수동으로 화면 오른쪽의 상자에 지정하려면 화살표 단추를 사용하십시오. 아이콘은 각 필드에 대한 유효한 측정 수준을 나타냅니다.

모두(조건 또는 예측)

이 목록에 추가되는 필드에 모델에서 생성되는 규칙의 예측 역할 또는 조건이 사용될 수 있습니다. 이는 규칙 기준에 의한 규칙이므로, 필드는 한 규칙의 조건이면서 다른 규칙의 예측이 될 수 있습니다.

예측만

이 목록에 추가되는 필드는 규칙의 예측("후향"이라고도 함)으로만 표시될 수 있습니다. 이 목록에 필드가 존재하는 것은 필드가 임의 규칙에서 사용됨을 의미하지는 않습니다. 단지 사용되는 경우 예측만 가능합니다.

조건만

이 목록에 추가되는 필드는 규칙의 조건("전향"이라고도 함)으로만 표시될 수 있습니다. 이 목록에 필드가 존재하는 것은 필드가 임의 규칙에서 사용됨을 의미하지는 않습니다. 단지 사용되는 경우 조건만 가능합니다.

② 연관 규칙 - 규칙 작성

규칙당 항목 수

각 규칙에서 사용할 수 있는 항목 또는 값 수를 지정하려면 이 옵션을 사용하십시오.

 참고: 이 두 필드의 결합된 총계는 10을 초과할 수 없습니다.

최대 조건 수

단일 규칙에 포함될 수 있는 최대 조건 수를 선택하십시오.

최대 예측변수 수

단일 규칙에 포함될 수 있는 최대 예측변수 수를 선택하십시오.

규칙 작성

작성할 규칙의 유형 및 개수를 지정하려면 다음 옵션을 사용하십시오.

최대 규칙 수

모델의 규칙 작성 시 사용하기 위해 고려할 수 있는 최대 규칙 수를 지정하십시오.

상위 N에 대한 규칙 기준

상위 N개 규칙을 설정하기 위해 사용되는 기준을 선택하십시오. N은 **최대 규칙 수** 필드에 입력되는 값입니다. 다음 기준에서 선택할 수 있습니다.

- 신뢰도
- 규칙 지원
- 조건 지원
- 리프트
- 배포성

플래그에 대한 참 값만 이용

데이터가 표 형식인 경우, 결과 규칙에 플래그 필드에 대한 참 값만 포함하려면 이 옵션에 선택하십시오. 참 값을 선택하면 규칙을 더 쉽게 이해할 수 있습니다. 트랜잭션 형식의 데이터에는 옵션이 적용되지 않습니다. 추가 정보는 테이블 대 트랜잭션 데이터의 내용을 참조하십시오.

규칙 기준

규칙 기준 사용을 선택하면, 규칙이 모델에서 사용되기 위해 충족해야 하는 최소 강도를 선택하기 위해 이 옵션을 사용할 수 있습니다.

- **신뢰도** 모델에 의해 생성되는 규칙에 대한 신뢰수준에 대한 최소 퍼센트 값을 지정하십시오. 모델이 이 정도보다 낮은 수준의 규칙을 생성하면 규칙이 삭제됩니다.
- **규칙 지원** 모델에 의해 생성되는 규칙에 대한 규칙 지원 수준의 최소 퍼센트 값을 지정하십시오. 모델이 이 정도보다 낮은 수준의 규칙을 생성하면 규칙이 삭제됩니다.
- **조건 지원** 모델에 의해 생성되는 규칙에 대한 조건 지원 수준에 대한 최소 퍼센트 값을 지정하십시오. 모델이 지정된 정도보다 낮은 수준의 규칙을 생성하면 규칙이 삭제됩니다.
- **리프트** 모델에 의해 생성되는 규칙에 대해 허용되는 최소 리프트 값을 지정하십시오. 모델이 지정된 정도보다 낮은 값의 규칙을 생성하면 규칙이 삭제됩니다.

규칙 제외

일부 경우에, 두 개 이상의 필드 사이의 연관이 알려져 있거나 따로 설명할 필요가 없습니다. 이러한 경우, 필드가 서로 예측하는 규칙을 제외할 수 있습니다. 두 값 모두를 포함하는 규칙을 제외하면, 관련이 없는 입력이 감소하고 유용한 결과를 발견하는 기회가 증가됩니다.

필드

규칙 작성 시 함께 사용하지 않을 연관 필드를 선택하십시오. 예를 들어, 연관 필드는 자동차 제조사 및 차종이나, 학생의 학년 및 나이가 될 수 있습니다. 모델이 규칙을 작성할 때, 규칙의 어느 한 쪽에서(조건 또는 예측) 선택된 필드 중 하나 이상에 규칙이 포함되는 경우, 규칙은 삭제됩니다.

③ 연관 규칙 - 변환

구간화

연속(수치 범위) 필드가 구간화되는 방법을 지정하려면 다음 옵션을 사용하십시오.

구간 수

자동으로 구간화되도록 설정된 연속형 필드는 사용자가 지정하는 동일 간격 구간 수로 나뉩니다. 2 - 10 범위의 숫자를 선택할 수 있습니다.

목록 필드

최대 목록 길이

목록 필드의 길이를 알 수 없는 경우에 모델에 포함할 항목 수를 제한하려면 목록의 최대 길이를 입력하십시오. 1 - 100 범위의 숫자를 선택할 수 있습니다. 목록이 입력하는 수보다 길 경우, 모델은 계속 필드를 사용하게 되지만 이 개수까지만 값을 포함하고, 필드의 추가 값은 무시됩니다.

④ 연관 규칙 - 출력

모델이 작성될 때 생성되는 출력을 제어하려면 이 분할창에서 옵션을 사용하십시오.

규칙 테이블

선택된 각 기준에 대해 최상의 규칙 수를 표시하는(사용자가 지정하는 수를 기반으로) 하나 이상의 테이블 유형을 작성하려면 다음 옵션을 사용하십시오.

신뢰도

신뢰도는 조건 지원에 대한 규칙 지원의 비율입니다. 나열된 조건 값을 가지는 항목의, 예측된 후향 값을 가지고 있는 퍼센트. 출력에 포함될 신뢰도를 기반으로 하는 최상의 N개 연관 규칙을 포함하는 테이블을 작성합니다(N은 **표시할 규칙** 값입니다).

규칙 지원

전체 규칙, 조건 및 예측이 일치하는 항목의 비율. 데이터 세트의 모든 항목에 대해, 규칙에 대해 올바르게 설명되거나 규칙으로 예측된 퍼센트. 이 측도는 규칙의 전체 중요도를 제공합니다. 출력에 포함될 규칙 지원을 기반으로 하는 최상의 N개 연관 규칙을 포함하는 테이블을 작성합니다(N은 **표시할 규칙** 값입니다).

리프트

규칙 신뢰도의 비율 및 예상값을 갖는 사전 확률. 규칙에 대한 신뢰도 값 비율 대 전체 모집단에서 후향 값이 발생하는 퍼센트. 이 비율은 규칙이 변화에서 제대로 개선되는 정도의 측도를 제공합니다. 출력에 포함될 리프트를 기반으로 하는 최상의 N개 연관 규칙을 포함하는 테이블을 작성합니다(N은 **표시할 규칙** 값입니다).

조건 지원

조건이 일치하는 항목의 비율. 출력에 포함될 전향 지원을 기반으로 하는 최상의 N개 연관 규칙을 포함하는 테이블을 작성합니다(N은 **표시할 규칙** 값입니다).

배포성

조건을 충족하지만 예상값을 충족하지 않는 훈련 데이터의 퍼센트 측도. 이 측도는 규칙이 빗나가는 빈도를 보여줍니다. 효과적으로, 신뢰도의 반대입니다. 출력에 포함될 배포성을 기반으로 하는 최상의 N개 연관 규칙을 포함하는 테이블을 작성합니다(N은 **표시할 규칙** 값입니다).

표시할 규칙

테이블에 표시할 최대 규칙 수를 설정하십시오.

모델 정보 테이블.

출력에 포함할 모델 테이블을 선택하려면 다음 옵션 중 하나 이상을 사용하십시오.

- 필드 변환
- 레코드 요약
- 규칙 통계량
- 최대 빈도 값
- 최대 빈도 필드

규칙의 정렬 가능 단어 클라우드.

규칙 출력을 표시하는 단어 클라우드를 작성하려면 다음 옵션을 사용하십시오. 단어는 해당 중요도를 표시하기 위해 증가하는 텍스트 크기로 표시됩니다.

정렬 가능 단어 클라우드 생성.

출력에서 정렬 가능 클라우드 단어를 작성하려면 이 상자를 선택하십시오.

기본 정렬

처음에 단어 클라우드를 작성할 때 사용할 정렬 유형을 선택하십시오. 단어 클라우드는 대화 형이며 다른 규칙 및 정렬을 보기 위해 모델 뷰어에서 기준을 변경할 수 있습니다. 다음 정렬 옵션에서 선택할 수 있습니다.

- 신뢰도.
- 규칙 지원
- 리프트
- 조건 지원.
- 배포성

표시할 최대 규칙

단어 클라우드에 표시될 규칙 수를 설정하십시오. 선택할 수 있는 최대값은 20입니다.

⑤ 연관 규칙 - 모델 옵션

연관 규칙 모델에 대한 스코어링 옵션을 지정하려면 이 탭에 설정을 사용하십시오.

모델 이름 자동으로 목표 필드(또는 이러한 필드가 지정되지 않은 경우 모델 유형)를 기반으로 모델 이름을 생성하거나, 사용자 정의 이름을 지정할 수 있습니다.

최대 예상값 수 스코어 결과에 포함되는 최대 예상값 수를 지정합니다. 이 옵션은 **규칙 기준** 항목과 함께 사용하여 “상위” 예상값을 생성합니다. 여기서 상위는 신뢰도, 지원, 리프트 등의 최상위 수준을 표시합니다.

규칙 기준 규칙의 강도를 결정하기 위해 사용되는 측도를 선택합니다. 규칙은 항목 세트에 대한 상위 예상값을 리턴하기 위해 여기에서 선택하는 기준의 강도별로 정렬됩니다. 5개의 다양한 기준에서 선택할 수 있습니다.

- **신뢰도** 신뢰도는 조건 지원에 대한 규칙 지원의 비율입니다. 나열된 조건 값을 가지는 항목의, 예측된 후향 값을 가지고 있는 퍼센트.
- **조건 지원** 조건이 일치하는 항목의 비율.
- **규칙 지원** 전체 규칙, 조건 및 예상값이 일치하는 항목의 비율. **조건 지원** 값에 **신뢰도** 값을 곱하여 계산합니다.
- **리프트** 규칙 신뢰도의 비율 및 예상값을 갖는 사전 확률.
- **배포성** 조건을 충족하지만 예상값을 충족하지 않는 훈련 데이터의 퍼센트 속도.

반복 예상값 허용 스코어링 동안 동일한 예상값을 갖는 여러 규칙을 포함하려면 이 선택란을 선택하십시오. 예를 들어, 이를 선택하면 다음 규칙이 스코어링될 수 있습니다.

bread & cheese -> wine
cheese & fruit -> wine

참고: 여러 예상값(bread & cheese & fruit -> wine & pate)이 있는 규칙은 모든 예상값(wine & pate)이 이전에 예측된 경우에만 반복 예상값을 고려합니다.

예상값이 입력에 존재하지 않는 경우에만 규칙 스코어링 예상값이 입력에도 존재하지 않는지 확인하려면 이 옵션을 선택하십시오. 예를 들어 스코어링 목적이 가정용 가구 제품을 추천하기 위한 것일 경우 이미 식탁이 들어있는 입력은 다른 것을 구매할 가능성이 적습니다. 이와 같은 경우, 이 옵션을 선택하십시오. 그러나, 제품이 상하기 쉽거나 일회용인 경우(예: 치즈, 아기 유동식 또는 티슈), 후항이 이미 입력에 존재하는 규칙이 가치있을 수도 있습니다. 후자의 경우, 가장 유용한 옵션은 **모든 규칙 스코어링**이 될 수 있습니다.

예상값이 입력에 존재하는 경우에만 규칙 스코어링 예상값이 입력에도 존재하지 않는지 확인하려면 이 옵션을 선택하십시오. 이 접근법은 기존 고객 또는 트랜잭션에 대해 통찰력을 얻고자 할 때 유용합니다. 예를 들어, 최고의 리프트를 가진 규칙을 식별한 다음 어떤 고객이 이러한 규칙에 적합한지 탐색하고 싶을 수 있습니다.

모든 규칙 스코어링 예상값의 존재 여부에 상관없이 스코어링 시 모든 규칙을 포함하려면 이 옵션을 선택하십시오.

⑥ 연관성 규칙 모델 너깃

모델 너깃에는 모델 작성 동안 사용자 데이터로부터 추출된 규칙에 대한 정보가 포함됩니다.

결과 보기

대화 상자의 모델 탭을 사용하여 연관성 규칙 모델에 의해 생성된 규칙을 찾아볼 수 있습니다. 모델 너깃을 찾아보면 새 노드를 생성하거나 모델을 스코어링하기 전에 규칙에 대한 정보를 볼 수 있습니다.

모델 스코어링

세분화된 모델 너깃은 스트림에 추가되어 스코어링에 사용될 수 있습니다. 자세한 정보는 스트림에서 모델 너깃 사용의 내용을 참조하십시오. 스코어링에 사용되는 모델 너깃은 각 대화 상자마다 추가 설정 탭을 포함합니다. 자세한 정보는 연관 규칙 모델 너깃 설정의 내용을 참조하십시오.

가. 연관 규칙 모델 너깃 세부사항

연관 규칙 모델 너깃은 출력 뷰어의 모델 탭에 모델 세부사항을 표시합니다. 뷰어 사용에 대한 자세한 정보는 출력 작업의 내용을 참조하십시오.

GSAR 모델링 작업은 다음 표에 표시된 것처럼 접두문자 \$A로 많은 새 필드를 작성합니다.

표 1. 연관 규칙 모델링 작업에 의해 작성된 새 필드	
필드 이름	설명
\$A-<prediction>#	이 필드에는 스코어링된 레코드에 대한 모델을 통한 예측이 포함됩니다. <prediction>은 모델에서 예측 역할에 포함된 필드의 이름이고, #은 출력 규칙에 대한 일련의 번호입니다(예를 들어, 스코어가 세 개의 규칙을 포함하도록 설정된 경우 일련의 번호는 1-3입니다).
\$AC-<prediction>#	이 필드에는 예측의 신뢰도가 포함됩니다. <prediction>은 모델에서 예측 역할에 포함된 필드의 이름이고, #은 출력 규칙에 대한 일련의 번호입니다(예를 들어, 스코어가 세 개의 규칙을 포함하도록 설정된 경우 일련의 번호는 1-3입니다).
\$A-Rule_ID#	이 열에는 스코어링된 데이터 세트에 있는 각 레코드에 대해 예측된 규칙의 ID가 포함됩니다. #은 출력 규칙에 대한 일련의 번호입니다(예를 들어, 스코어가 세 개의 규칙을 포함하도록 설정되면 일련의 번호는 1-3입니다).

나. 연관 규칙 모델 너깃 설정

연관 규칙 모델 너깃의 설정 탭은 모델에 대한 스코어링 옵션을 표시합니다. 이 탭은 스코어링에 대한 스트림 캔버스에 모델이 추가된 후에만 사용할 수 있습니다.

최대 예상값 수 각 항목 세트에 대해 포함되는 최대 예상값 수를 지정하십시오. 이 트랜잭션 세트에 적용되는 최상위 신뢰도 값을 가지고 있는 규칙이 지정된 한계까지 레코드에 대한 예상값을 생성하기 위해 사용됩니다. **규칙 기준** 옵션과 함께 이 옵션을 사용하여 “상위” 예상값을 생성하십시오. 여기서 **상위**는 신뢰도, 지원, 리프트 등의 최상위 수준을 표시합니다.

규칙 기준 규칙의 강도를 결정하기 위해 사용되는 측도를 선택합니다. 규칙은 항목 세트에 대한 상위 예상값을 리턴하기 위해 여기에서 선택하는 기준의 강도별로 정렬됩니다. 다음 기준에서 선택할 수 있습니다.

- 신뢰도
- 규칙 지원
- 리프트
- 조건 지원
- 배포성

반복 예상값 허용 스코어링 시 후항이 같은 여러 규칙을 포함하려면 이 선택란을 선택하십시오. 예를 들어, 이 옵션을 선택하는 것은 다음 규칙의 스코어를 매길 수 있음을 의미합니다.

```
bread & cheese -> wine  
cheese & fruit -> wine
```

스코어링 시 반복 예상값을 제외하려면 선택란을 지우십시오.

참고: 여러 후항(bread & cheese & fruit -> wine & pate)이 있는 규칙은 모든 후항(wine & pate)이 이전에 예측된 경우에만 반복 예상값을 고려합니다.

예상값이 입력에 존재하지 않는 경우에만 규칙 스코어링 후항이 입력에도 존재하지 않는지 확인하려면 선택하십시오. 예를 들어 스코어링 목적이 가정용 가구 제품을 추천하기 위한 것일 경우 이미 식탁이 들어있는 입력은 다른 것을 구매할 가능성이 적습니다. 이와 같은 경우, 이 옵션을 선택하십시오. 다른 한편으로, 제품이 상하기 쉽거나 일회용인 경우(예: 치즈, 아기 유동식 또는 티슈), 후항이 이미 입력에 존재하는 규칙이 가치있을 수도 있습니다. 후자의 경우, 가장 유용한 옵션은 모든 규칙 스코어링이 될 수 있습니다.

예상값이 입력에 존재하는 경우에만 규칙 스코어링 후항이 입력에도 존재하는지 확인하려면 이 옵션을 선택하십시오. 이 접근법은 기존 고객 또는 트랜잭션에 대해 통찰력을 얻고자 할 때 유용합니다. 예를 들어, 최고의 리프트를 가진 규칙을 식별한 다음 어떤 고객이 이러한 규칙에 적합한지 탐색하고 싶을 수 있습니다.

모든 규칙 스코어링 입력에서 후항의 존재 여부에 상관없이 스코어링 시 모든 규칙을 포함하려면 이 옵션을 선택하십시오.