IBM SPSS Modeler 18.1.1
모델링 노드

IBM

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 423 페이지의 『주의사항』에 있는 정보를 확인하십시오.

제품 정보

이 개정판은 새 개정판에 별도로 명시하지 않는 한, IBM SPSS Modeler의 버전 18, 릴리스 1, 수정 1 및 모든 후속 릴리스와 수정에 적용됩니다.

목차

서론 ix
IBM Business Analytics 소개 ix
기술 지원 ix

제 1 장 IBM SPSS Modeler 정보 1
IBM SPSS Modeler 제품 1
IBM SPSS Modeler 1
IBM SPSS Modeler Server 2
IBM SPSS Modeler Administration Console 2
IBM SPSS Modeler Batch 2
IBM SPSS Modeler Solution Publisher . . . 2
IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터 3
IBM SPSS Modeler 에디션 3
문서 3
SPSS Modeler Professional 문서 4
SPSS Modeler Premium 문서 5
애플리케이션 예제 5
Demos 폴더 5
라이센스 추적 5

제 2 장 모델링 소개 7
스트림 작성 9
모델 찾아보기 14
모델 평가 19
레코드 스코어링 22
요약 23

제 3 장 모델링 개요 25
모델링 노드의 개요 25
분할 모델 작성 30
분할 및 파티셔닝 31
분할 모델을 지원하는 모델링 노드 31
분할 영향을 받는 기능 32
모델링 노드 필드 옵션 33
빈도 및 가중 필드 사용 35
모델링 노드 분석 옵션 37
성향 스코어 38
오분류 비용 40
모델 너깃 40
모델 링크 41
모델 교체 43
모델 팔레트 44
모델 너깃 찾아보기 45
모델 너깃 요약/정보 47
예측변수 중요도 47
앙상블 뷰어 49
분할 모델의 모델 너깃 51
스트림에서 모델 너깃 사용 52
모델링 노드 재생성 53
PMML로서 모델 가져오기 및 내보내기 . . . 54
스코어링 어댑터에 대한 모델 게시 56
세분화되지 않은 모델 56

제 4 장 선별 모델 59
필드 및 레코드 선별 59
필드선택 노드 59
필드선택 모델 설정 60
필드선택 옵션 61
필드선택 모델 너깃 62
필드선택 모델 결과 62
중요도에 따라 필드 선택 63
필드선택 모델에서 필터 생성 63
이상 항목 발견 노드 64
이상 항목 발견 모델 옵션 65
이상 항목 발견 고급 옵션 65
이상 항목 발견 모델 너깃 67
이상 항목 발견 모델 세부사항 67
이상 항목 발견 모델 요약 67
이상 항목 발견 모델 설정 67

제 5 장 자동화된 모델링 노드 69
자동화된 모델링 노드 알고리즘 설정 70
자동화된 모델링 노드 중지 규칙 70
자동 분류자 노드 71
자동 분류자 노드 모델 옵션 72
자동 분류자 노드 고급 옵션 74
오분류 비용 77
자동 분류자 노드 삭제 옵션 77
자동 분류자 노드 설정 옵션 77
자동 숫자 노드 78
자동 숫자 노드 모델 옵션 79
자동 숫자 노드 고급 옵션 80
자동 숫자 노드 설정 옵션 83
자동 군집 노드 83
자동 군집 노드 모델 옵션 84

자동 군집 노드 고급 옵션 85
자동 군집 노드 삭제 옵션 86
자동화된 모델 너깃 87
노드 및 모델 생성 88
평가 차트 생성. 89
평가 그래프 89

제 6 장 의사결정 트리 91
의사결정 트리 모형 91
대화형 트리 작성기 93
트리 성장 및 가지치기 93
사용자 정의 분할 정의 94
분할 세부사항 및 대용 95
트리 보기 사용자 정의 96
이익. 97
위험 101
트리 모델 및 결과 저장 101
필터 및 선택 노드 생성 104
의사결정 트리에서 규칙 세트 생성 105
직접 트리 모델 작성 106
의사결정 트리 노드. 106
C&R 트리 노드 108
CHAID 노드 108
QUEST 노드 109
의사결정 트리 노드 필드 옵션 110
의사결정 트리 노드 작성 옵션 110
의사결정 트리 노드 모델 옵션 117
C5.0 노드 118
C5.0 노드 모델 옵션 119
Tree-AS 노드 121
Tree-AS 노드 필드 옵션 121
Tree-AS 노드 작성 옵션 122
Tree-AS 노드 모델 옵션 124
Tree-AS 모델 너깃 125
랜덤 트리 노드 127
랜덤 트리 노드 필드 옵션 127
랜덤 트리 노드 작성 옵션 128
랜덤 트리 노드 모델 옵션 130
랜덤 트리 모델 너깃 130
C&R 트리, CHAID, QUEST, C5.0 의사결정 트리 모형 너깃 133
단일 트리 모델 너깃 134
부스팅, 배깅 및 매우 큰 데이터 세트의 모델 너깃 139
C&R 트리, CHAID, QUEST, C5.0, Apriori 규칙 세트 모델 너깃 140
규칙 세트 모델 탭 141
AnswerTree 3.0에서 프로젝트 가져오기 . . . 142

제 7 장 베이지안 신경망 모형 143
베이지안 네트워크 노드 143
베이지안 네트워크 노드 모델 옵션 145
베이지안 네트워크 노드 고급 옵션 146
베이지안 신경망 모형 너깃 147
베이지안 신경망 모형 설정 148
베이지안 신경망 모형 요약 149

제 8 장 신경망 151
신경망 모델 151
레거시 스트림이 있는 신경망 사용 152
목적 153
기본 155
중지 규칙 156
앙상블 157
고급 158
모델 옵션 159
모델 요약 160
예측변수 중요도 161
관측값 별 예측값 162
분류 162
네트워크. 163
설정 165

제 9 장 의사결정 목록 167
의사결정 목록 모델 옵션. 168
의사결정 목록 노드 고급 옵션 170
의사결정 목록 모델 너깃. 170
의사결정 목록 모델 너깃 설정 171
의사결정 목록 뷰어. 172
작업 모델 분할창 172
대안 탭 173
스냅샷 탭 174
의사결정 목록 뷰어에 대한 작업 175

제 10 장 통계 모델. 189
선형 노드 190
선형 모델 190
Linear-AS 노드 198
Linear-AS 모델 199
로지스틱 노드. 202
로지스틱 노드 모델 옵션 203
로지스틱 회귀분석 모델에 항 추가 206
로지스틱 노드 고급 옵션 207
로지스틱 회귀분석 수렴 옵션 208
로지스틱 회귀분석 고급 출력 208

로지스틱 회귀분석 단계별 옵션 209
로지스틱 모델 너깃 210
로지스틱 너깃 모델 세부사항 211
로지스틱 모델 너깃 요약 212
로지스틱 모델 너깃 설정 212
로지스틱 모델 너깃 고급 출력 213
PCA/요인 노드 214
PCA/요인 노드 모델 옵션 215
PCA/요인 노드 고급 옵션 215
PCA/요인 노드 회전 옵션 216
PCA/요인 모델 너깃 217
PCA/요인 모델 너깃 방정식 217
PCA/요인 모델 너깃 요약 217
PCA/요인 모델 너깃 고급 출력 217
판별 노드 218
판별 노드 모델 옵션 218
판별 노드 고급 옵션 219
판별 노드 출력 옵션 219
판별 노드 단계 옵션 221
판별 모델 너깃 221
GenLin 노드 223
GenLin 노드 필드 옵션 223
GenLin 노드 모델 옵션 224
GenLin 노드 고급 옵션 225
일반화 선형 모델 반복 227
일반화 선형 모델 고급 출력 228
GenLin 모델 너깃 229
일반화 선형 혼합 모델 230
GLMM 노드 230
GLE 노드 245
목표 246
모델 효과 248
가중치 및 오프셋 250
작성 옵션 250
추정 251
모델 선택 252
모델 옵션 253
GLE 모델 너깃 253
Cox 노드 255
Cox 노드 필드 옵션 255
Cox 노드 모델 옵션 256
Cox 노드 고급 옵션 257
Cox 노드 설정 옵션 259
Cox 모델 너깃 259

제 11 장 군집 모델 261
코호넨 노드 262
코호넨 노드 모델 옵션 263
코호넨 노드 고급 옵션 264
코호넨 모델 너깃 265
코호넨 모델 요약 265
K-평균 노드 266
K-평균 노드 모델 옵션 266
K-평균 노드 고급 옵션 267
K-평균 모델 너깃 267
K-평균 모델 요약 268
이단계 군집 노드 268
이단계 군집 노드 모델 옵션 269
이단계 군집 모델 너깃 270
이단계 모델 요약 270
TwoStep-AS 군집 노드 270
Twostep-AS 군집분석 270
TwoStep-AS 군집 모델 너깃 276
TwoStep-AS 군집 모델 너깃 설정 276
K-평균-AS 노드 277
K-평균-AS 노드 필드 277
K-평균-AS 노드 작성 옵션 277
군집 뷰어 278
군집 뷰어 - 모델 탭 279
군집 뷰어 탐색 283
군집 모델에서 그래프 생성 285

제 12 장 연관 규칙 287
테이블 대 트랜잭션 데이터 288
Apriori 노드 289
Apriori 노드 모델 옵션 289
Apriori 노드 고급 옵션 290
CARMA 노드 292
CARMA 노드 필드 옵션 292
CARMA 노드 모델 옵션 293
CARMA 노드 고급 옵션 294
연관 규칙 모델 너깃 295
연관 규칙 모델 너깃 세부사항 295
연관 규칙 모델 너깃 설정 299
연관 규칙 모델 너깃 요약 300
연관 모델 너깃에서 규칙 세트 생성 301
필터링된 모델 생성 301
연관 규칙 스코어링 302
연관 모델 배포 303
시퀀스 노드 305
시퀀스 노드 필드 옵션 306
시퀀스 노드 모델 옵션 307
시퀀스 노드 고급 옵션 307
시퀀스 모델 너깃 309

연관 규칙 노드 313
연관 규칙 - 필드 옵션. 314
연관 규칙 - 규칙 작성. 315
연관 규칙 - 변환 316
연관 규칙 - 출력 317
연관 규칙 - 모델 옵션. 318
연관성 규칙 모델 너깃 319

제 13 장 시계열 모델 323
왜 예측인가? 323
시계열 데이터. 323
시계열의 공정특성 변수 323
자기상관 및 편자기상관 함수 328
계열 변환 329
예측변수 계열. 329
STP(Spatio-Temporal Prediction) 모델링 노드 330
STP(Spatio-Temporal Prediction) - 필드 옵션 330
STP(Spatio-Temporal Prediction) - 시간 구간 331
STP(Spatio-Temporal Prediction) - 기본 작성 옵션 333
STP(Spatio-Temporal Prediction) - 고급 작성 옵션 333
STP(Spatio-Temporal Prediction) - 출력 334
STP(Spatio-Temporal Prediction) - 모델 옵션 335
STP(Spatio-Temporal Prediction) 모델 너깃 335
TCM 노드 336
시간 인과 모델 336
TCM 모델 너깃 348
시간 인과 모델 시나리오. 349
시계열 노드 355
시계열 노드 - 필드 옵션. 356
시계열 노드 - 데이터 지정 사항 옵션 . . . 356
시계열 노드 - 작성 옵션. 360
시계열 노드 - 모델 옵션. 365
시계열 모델 너깃 367

제 14 장 자체 학습 응답 노드 모델. 371
SLRM 노드 371
SLRM 노드 필드 옵션 371
SLRM 노드 모델 옵션 372
SLRM 노드 설정 옵션 373
SLRM 모델 너깃 374
SLRM 모델 설정 374

제 15 장 지원 벡터 머신 모델. 377
SVM 정보 377
SVM 작동 방법 377
SVM 모델 조정 378
SVM 노드 379
SVM 노드 모델 옵션 380
SVM 노드 고급 옵션 380
SVM 모델 너깃 381
SVM 모델 설정 382
LSVM 노드 382
LSVM 노드 모델 옵션 383
LSVM 작성 옵션 383
LSVM 모델 너깃(대화형 출력). 384
LSVM 모델 설정 385

제 16 장 최근접 이웃 모델 387
KNN 노드. 387
KNN 노드 목표 옵션. 387
KNN 노드 설정. 388
KNN 모델 너깃. 392
최근접 이웃 모델 보기 393
KNN 모델 설정. 395

제 17 장 Python 노드 397
SMOTE 노드. 398
SMOTE 노드 설정 398
XGBoost Linear 노드. 399
XGBoost Linear 노드 필드. 399
XGBoost Linear 노드 작성 옵션. 400
XGBoost Linear 노드 모델 옵션. 401
XGBoost Tree 노드 401
XGBoost Tree 노드 필드. 401
XGBoost Tree 노드 작성 옵션. 402
XGBoost Tree 노드 모델 옵션. 404
t-SNE 노드. 404
t-SNE 노드 고급 옵션. 404
t-SNE 노드 출력 옵션. 406
t-SNE 모델 너깃. 407
랜덤 포리스트 노드. 407
랜덤 포리스트 노드 필드. 408
랜덤 포리스트 노드 작성 옵션. 408
랜덤 포리스트 노드 모델 옵션. 410
랜덤 포리스트 모델 너깃. 410
One-Class SVM 노드. 410
One-Class SVM 노드 필드. 411
One-Class SVM 노드 고급. 411
One-Class SVM 노드 옵션. 412

제 18 장 Spark 노드 415
등위-AS 노드 415
등위-AS 노드 필드 415
등위-AS 노드 작성 옵션 416
등위-AS 모델 너깃 416
XGBoost-AS 노드 416
XGBoost-AS 노드 필드 417
XGBoost-AS 노드 작성 옵션 417
XGBoost-AS 노드 모델 옵션 420
K-평균-AS 노드 420
K-평균-AS 노드 필드 420
K-평균-AS 노드 작성 옵션 420

주의사항 423
상표 424
제품 문서의 이용 약관 425

용어 427
가 427
바 427
사 427
아 428
자 428
차 429
카 429
파 429
하 429
숫자 430
A 430
B 430
F 430
M 430
R 431
W 431

색인 433

서론

IBM® SPSS® Modeler는 IBM Corp. 엔터프라이즈 중심의 데이터 마이닝 워크벤치입니다. SPSS Modeler는 상세한 데이터 이해를 통해 조직이 고객과 시민과의 관계를 향상시킬 수 있도록 도움을 줍니다. 조직은 SPSS Modeler에서 확보한 통찰력을 통해 수익 창출이 가능한 고객을 보유하고, 교차 판매 기회를 식별하고, 새 고객을 모으고, 사기 행위를 적발하고, 위험을 줄이고, 정부 서비스 지원을 향상시킬 수 있습니다.

SPSS Modeler의 시각적 인터페이스를 통해 사용자는 보다 쉽게 비즈니스에 특정한 전문 지식을 적용할 수 있으므로, 더 강력한 예측 모델을 생성하고 솔루션 출시 시점을 단축할 수 있습니다. SPSS Modeler에서는 예측, 분류, 세분화, 연관 발견 알고리즘과 같은 많은 모델링 기법을 제공합니다. 모델을 작성하면 IBM SPSS Modeler Solution Publisher에서 의사결정자 또는 데이터베이스까지 엔터프라이즈 범위로 모델을 전달할 수 있습니다.

IBM Business Analytics 소개

IBM Business Analytics 소프트웨어는 의사 결정자가 비즈니스 성능을 개선하기 위해 신뢰하는 완벽하고 일관되며 정확한 정보를 제공합니다. 비즈니스 지능, 예측 분석, 금융 성과와 전략 관리 및 분석 응용 프로그램의 종합 포트폴리오는 현재 성과와 앞으로의 결과를 예측하는 능력에 분명하고 즉각적이면서 실행 가능한 통찰력을 제공합니다. 풍부한 업계 솔루션, 입증된 사례 및 전문 서비스가 결합되어 어떠한 크기의 조직이라도 생산성을 극대화하고 자신있는 자동 결정을 내릴 수 있으며 더 나은 결과를 가져올 수 있습니다.

이 포트폴리오의 일부인 IBM SPSS Predictive Analytics 소프트웨어를 통해 조직은 미래의 사건을 예측하고 더 나은 비즈니스 결과를 얻기 위한 통찰력에 대해 적극적인 조치를 할 수 있습니다. 전 세계의 기업, 정부 및 학계 고객들은 고객을 매료시키고 유지하며 성장하게 만드는 동시에 불공정 행위를 줄이고 위험을 낮추는 IBM SPSS 기술의 경쟁 이점을 활용합니다. 일상 업무에서 IBM SPSS 소프트웨어를 활용한다면 예측형 기업으로 거듭날 수 있습니다. 즉 비즈니스 목표 달성을 위해 의사 결정의 방향을 정하고 이를 자동화하며 측정 가능한 경쟁 우위를 달성할 수 있습니다. 자세한 내용을 보거나 담당자에게 문의하려면 http://www.ibm.com/spss 사이트를 방문하십시오.

기술 지원

기술 지원은 유지 관리 고객에게 제공됩니다. IBM Corp. 제품 사용 및 지원된 하드웨어 환경 중 하나에 대해 설치하는 데 도움이 필요한 경우 기술 지원부로 문의하십시오. 기술 지원에 문의하려면 http://www.ibm.com/support의 IBM Corp. 웹 사이트를 참고하십시오. 지원을 요청하려면 본인의 신상과 소속 조직(회사) 및 지원 동의서를 제시해야 합니다.

제 1 장 IBM SPSS Modeler 정보

IBM SPSS Modeler는 비즈니스 전문 지식을 사용하여 예측 모형을 신속하게 개발하고 이를 비즈니스 운영에 배포하여 의사결정의 정확성을 향상시켜주는 데이터 마이닝 도구 세트입니다. 산업 표준 CRISP-DM 모델을 중심으로 설계된 IBM SPSS Modeler는 데이터에서 보다 나아진 비즈니스 결과에 이르는 전체 데이터 마이닝 프로세스를 지원합니다.

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.

SPSS Modeler는 독립형 제품으로 구매하거나 SPSS Modeler Server와 통합하여 클라이언트로 사용할 수 있습니다. 다음 절에 요약된 바와 같이 여러가지 추가 옵션도 사용할 수 있습니다. 자세한 정보는 https://www.ibm.com/analytics/us/en/technology/spss/의 내용을 참조하십시오.

IBM SPSS Modeler 제품

IBM SPSS Modeler 제품군 및 연관 소프트웨어는 다음으로 구성됩니다.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console(IBM SPSS Deployment Manager와 함께 포함)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터

IBM SPSS Modeler

SPSS Modeler는 개인용 컴퓨터에 설치하여 실행되는 기능적으로 완벽한 버전의 제품입니다. 로컬 모드에서 독립형 제품으로 SPSS Modeler를 실행하거나 대형 데이터 세트에 대한 성능 향상을 위해 분산 모드에서 IBM SPSS Modeler Server와 함께 사용할 수 있습니다.

SPSS Modeler를 사용하여 프로그래밍하지 않고 신속하게 직관적으로 정확한 예측 모델을 작성할 수 있습니다. 고유한 시각적 인터페이스를 사용하면 데이터 마이닝 프로세스를 쉽게 시각화할 수 있습니다. 제품에 포함된 고급 분석 지원을 통해 데이터에서 이전에 숨겨진 패턴과 추세를 발견할 수 있습니다. 결과를 모델링하고 결과에 영향을 주는 요인을 이해하여 비즈니스 기회를 활용하고 위험을 줄일 수 있습니다.

SPSS Modeler는 두 개의 에디션(SPSS Modeler Professional과 SPSS Modeler Premium)으로 사용할 수 있습니다. 자세한 정보는 3 페이지의 『IBM SPSS Modeler 에디션』의 내용을 참조하십시오.

IBM SPSS Modeler Server

SPSS Modeler는 클라이언트/서버 설계를 사용하여 자원 집약적 작업에 대한 요청을 강력한 서버 소프트웨어로 분배하여 대형 데이터 세트에 대한 성능을 향상시킵니다.

SPSS Modeler Server는 하나 이상의 IBM SPSS Modeler 설치와 함께 서버 호스트의 분산 분석 모드에서 계속해서 실행되는 별도로 라이센스가 부여된 제품입니다. 이런 방법으로 클라이언트 컴퓨터로 데이터를 다운로드하지 않고 서버에서 메모리 집약적 작업을 수행할 수 있기 때문에 SPSS Modeler Server는 대형 데이터 세트에 대한 우수한 성능을 제공합니다. 또한 IBM SPSS Modeler Server는 SQL 최적화 및 In-Database 모델링 기능에 대한 지원을 제공하여 성능 및 자동화의 이점도 추가로 제공합니다.

IBM SPSS Modeler Administration Console

Modeler Administration Console은 옵션 파일을 통해서도 구성 가능한 다수의 SPSS Modeler Server 구성 옵션을 관리하기 위한 그래픽 사용자 인터페이스입니다. 콘솔은 IBM SPSS Deployment Manager에 포함되며, SPSS Modeler Server 설치를 모니터하고 구성하는 데 사용할 수 있고, 현재 SPSS Modeler Server 고객이 무료로 사용할 수 있습니다. 이 애플리케이션은 Windows 컴퓨터에만 설치할 수 있지만 지원되는 플랫폼에 설치된 서버를 관리할 수 있습니다.

IBM SPSS Modeler Batch

데이터 마이닝은 일반적으로 대화식 처리인 반면, 그래픽 사용자 인터페이스가 없어도 명령행에서 SPSS Modeler를 실행할 수 있습니다. 예를 들어, 사용자 개입 없이 수행할 장기 실행 또는 반복 작업이 있습니다. SPSS Modeler Batch는 정규 사용자 인터페이스에 대한 액세스 없이 SPSS Modeler의 전체 분석 기능에 대한 지원을 제공하는 특수 버전의 제품입니다. SPSS Modeler Batch를 사용하려면 SPSS Modeler Server가 필요합니다.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher는 외부 런타임 엔진을 통해 실행하거나 외부 애플리케이션에 포함될 수 있는 SPSS Modeler 스트림의 패키지 버전을 작성할 수 있게 하는 도구입니다. 이런 방법으로 SPSS Modeler가 설치되지 않는 환경에 사용할 수 있도록 전체 SPSS Modeler 스트림을 출판하고 배포할 수 있습니다. SPSS Modeler Solution Publisher는 별도의 라이센스가 필요한 IBM SPSS Collaboration and Deployment Services - Scoring 서비스의 일부로 분배됩니다. 이 라이센스가 있으면 출판된 스트림을 실행할 수 있게 하는 SPSS Modeler Solution Publisher Runtime을 수신합니다.

SPSS Modeler Solution Publisher에 대한 자세한 정보는 IBM SPSS Collaboration and Deployment Services 문서를 참조하십시오. IBM SPSS Collaboration and Deployment Services Knowledge Center에는 "IBM SPSS Modeler Solution Publisher" 및 "IBM SPSS Analytics Toolkit" 섹션이 포함되어 있습니다.

IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터

SPSS Modeler와 SPSS Modeler Server가 IBM SPSS Collaboration and Deployment Services 리포지토리와 상호작용할 수 있게 하는 IBM SPSS Collaboration and Deployment Services용 어댑터를 상당수 사용할 수 있습니다. 이런 방법으로 리포지토리에 배포된 SPSS Modeler 스트림을 여러 사용자가 공유하거나 씬 클라이언트 애플리케이션 IBM SPSS Modeler Advantage에서 액세스할 수 있습니다. 리포지토리를 호스팅하는 시스템에 어댑터를 설치하십시오.

IBM SPSS Modeler 에디션

SPSS Modeler는 다음 에디션으로 사용할 수 있습니다.

SPSS Modeler Professional

SPSS Modeler Professional은 CRM 시스템, 인구 통계, 구매 동작, 판매 데이터에서 추적된 동작 및 상호작용과 같은 대부분의 구조화된 데이터 유형에 대한 작업을 하는 데 필요한 모든 도구를 제공합니다.

SPSS Modeler Premium

SPSS Modeler Premium은 특수 데이터 및 비정형 텍스트 데이터에 대한 작업을 할 수 있도록 SPSS Modeler Professional을 확장하는 별도로 라이센스가 부여된 제품입니다. SPSS Modeler Premium에는 IBM SPSS Modeler Text Analytics가 포함됩니다.

IBM SPSS Modeler Text Analytics는 고급 언어 기술 및 자연어 처리(NLP)를 사용하여 다양한 비정형 텍스트 데이터를 빠르게 처리하고, 주요 개념을 추출 및 구성하고, 이러한 개념을 범주로 분류합니다. 추출된 개념과 범주는 인구 통계와 같은 기존의 구조화된 데이터와 결합할 수 있고 보다 나은 집중적인 의사결정을 내리기 위해 전체 IBM SPSS Modeler 데이터 마이닝 세트를 사용하여 모델링에 적용할 수 있습니다.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription에서는 일반적인 IBM SPSS Modeler 클라이언트와 동일한 모든 예측 분석 공정능력을 제공합니다. 구독 에디션을 사용하면 제품 업데이트를 정기적으로 다운로드할 수 있습니다.

문서

문서는 SPSS Modeler의 도움말 메뉴에서 사용할 수 있습니다. 이 문서는 Knowledge Center를 열고, 제품 외부에서 공개적으로 제공됩니다.

설치 지시사항을 포함하여 각 제품에 대한 전체 문서는 제품 다운로드의 일부로 별도의 압축 폴더에 PDF 형식으로도 제공됩니다. 또한 PDF 문서는 http://www.ibm.com/support/docview.wss?uid=swg27046871 웹 페이지에서 다운로드할 수 있습니다.

SPSS Modeler Professional 문서

SPSS Modeler Professional 문서 스위트(설치 지시사항은 제외)는 다음과 같습니다.

- **IBM SPSS Modeler 사용자 안내서.** SPSS Modeler 사용에 관한 일반적인 소개이며, 데이터 스트림 작성, 결측값 처리, CLEM 표현식 작성 프로젝트 및 보고서에 대한 작업, IBM SPSS Collaboration and Deployment Services 또는 IBM SPSS Modeler Advantage에 배포하기 위한 스트림 패키지 방법이 포함됩니다.
- **IBM SPSS Modeler 소스, 프로세스 및 출력 노드.** 여러 형식의 데이터를 읽고 처리하며, 출력하는 데 사용하는 모든 노드에 대한 설명입니다. 실질적으로 이는 모델링 노드 이외의 모든 노드를 의미합니다.
- **IBM SPSS Modeler 모델링 노드.** 데이터 마이닝 모델을 작성하는 데 사용하는 모든 노드에 대한 설명입니다. IBM SPSS Modeler는 기계 학습, 인공지능 및 통계로부터 취한 다양한 모델링 방법을 제공합니다.
- **IBM SPSS Modeler 애플리케이션 안내서.** 이 안내서의 예제는 특정 모델링 방법과 기법을 중점적으로 간략히 소개합니다. 이 안내서의 온라인 버전은 도움말 메뉴에서도 사용할 수 있습니다. 자세한 정보는 5 페이지의 『애플리케이션 예제』 주제를 참조하십시오.
- **IBM SPSS Modeler Python 스크립팅 및 자동화.** 노드와 스트림을 조작하는 데 사용할 수 있는 특성을 포함하여 Python 스크립팅을 통한 시스템 자동화에 대한 정보입니다.
- **IBM SPSS Modeler 배포 안내서.** IBM SPSS Deployment Manager에서 작업 처리 단계로 IBM SPSS Modeler 스트림 실행에 대한 정보입니다.
- **IBM SPSS Modeler CLEF 개발자 안내서.** CLEF는 데이터 처리 루틴 또는 모델링 알고리즘과 같은 써드파티 프로그램을 IBM SPSS Modeler의 노드로 통합하는 기능을 제공합니다.
- **IBM SPSS Modeler In-Database 마이닝 안내서.** 데이터베이스의 능력을 사용하여 성능을 향상시키고 써드파티 알고리즘을 통해 분석 기능 범위를 확장하는 방법에 대한 정보입니다.
- **IBM SPSS Modeler Server 관리 및 성능 안내서.** IBM SPSS Modeler Server 구성 및 관리 방법에 대한 정보입니다.
- **IBM SPSS Deployment Manager 사용자 안내서.** IBM SPSS Modeler Server 모니터링 및 구성용 Deployment Manager 애플리케이션에 포함된 관리 콘솔 사용자 인터페이스를 사용하는 데 대한 정보
- **IBM SPSS Modeler CRISP-DM 안내서.** SPSS Modeler에서 데이터 마이닝에 CRISP-DM 방법론을 사용하기 위한 단계별 안내서입니다.
- **IBM SPSS Modeler Batch 사용자 안내서.** 일괄처리 모드 실행 및 명령행 인수 세부사항을 포함하여 일괄처리 모드에서 IBM SPSS Modeler 사용을 위한 전체 안내서입니다. 이 안내서는 PDF 형식으로만 사용할 수 있습니다.

SPSS Modeler Premium 문서

SPSS Modeler Premium 문서 스위트(설치 지시사항은 제외)는 다음과 같습니다.

- **SPSS Modeler Text Analytics 사용자 안내서.** SPSS Modeler에서 텍스트 분석 사용에 대한 정보로, 텍스트 마이닝 노드, 대화식 워크벤치, 템플리트 및 기타 자원에 대해 설명합니다.

애플리케이션 예제

SPSS Modeler의 데이터 마이닝 도구가 광범위한 비즈니스 및 조직의 문제점을 해결하는 데 도움을 주는 가운데, 애플리케이션 예제는 특정 모델링 방법 및 기술에 대해 대상화된 간략한 소개를 제공합니다. 여기서 사용된 데이터 세트는 일부 데이터 마이너에서 관리하는 거대한 데이터 스토어보다 훨씬 작지만, 관련된 개념과 방법은 실제 애플리케이션으로 확장 가능합니다.

예제에 액세스하려면 SPSS Modeler의 도움말 메뉴에서 **애플리케이션 예제**를 클릭하십시오.

데이터 파일 및 샘플 스트림은 제품 설치 디렉토리 아래에 있는 Demos 폴더에 설치됩니다. 추가 정보는 『Demos 폴더』의 내용을 참조하십시오.

데이터베이스 모델링 예제. *IBM SPSS Modeler In-Database* 마이닝 안내서의 예제를 참조하십시오.

스크립팅 예제. *IBM SPSS Modeler* 스크립팅 및 자동화 안내서의 예제를 참조하십시오.

Demos 폴더

애플리케이션 예에서 사용하는 데이터 파일 및 샘플 스트림은 제품 설치 디렉토리 아래의 Demos 폴더에 설치됩니다(예: C:\Program Files\IBM\SPSS\Modeler\<version>\Demos). Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서, 또는 **파일** > **스트림 열기** 대화 상자의 최근 디렉토리 목록에서 Demos를 클릭해서도 이 폴더에 액세스할 수 있습니다.

라이센스 추적

SPSS Modeler를 사용할 때 라이센스 사용이 정기적으로 추적되고 로그됩니다. 로그되는 라이센스 메트릭은 *AUTHORIZED_USER* 및 *CONCURRENT_USER*이며 로그되는 메트릭의 유형은 SPSS Modeler에 대해 가진 라이센스의 유형에 의해 결정됩니다.

생성되는 로그 파일은 사용자가 라이센스 사용 보고서를 생성할 수 있는 IBM 라이센스 메트릭 도구에 의해 처리될 수 있습니다.

라이센스 로그 파일은 SPSS Modeler 클라이언트 로그 파일이 기록되는 디렉토리와 동일한 디렉토리(기본적으로 %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log)에 작성됩니다.

제 2 장 모델링 소개

모델은 입력 필드 또는 변수 세트에 기반하여 결과를 예측하는 데 사용할 수 있는 규칙, 수식 또는 방정식의 세트입니다. 예를 들어, 금융 기관은 모델을 사용하여 과거 신청자들에 대해 이미 알고 있는 정보에 기반하여 대출 신청자의 위험이 낮은지(안전), 높은지(위험)를 예측할 수 있습니다.

결과를 예측하는 기능은 예측 분석의 중요한 목표이며, 모델링 프로세스를 이해하는 것이 IBM SPSS Modeler 사용의 핵심입니다.

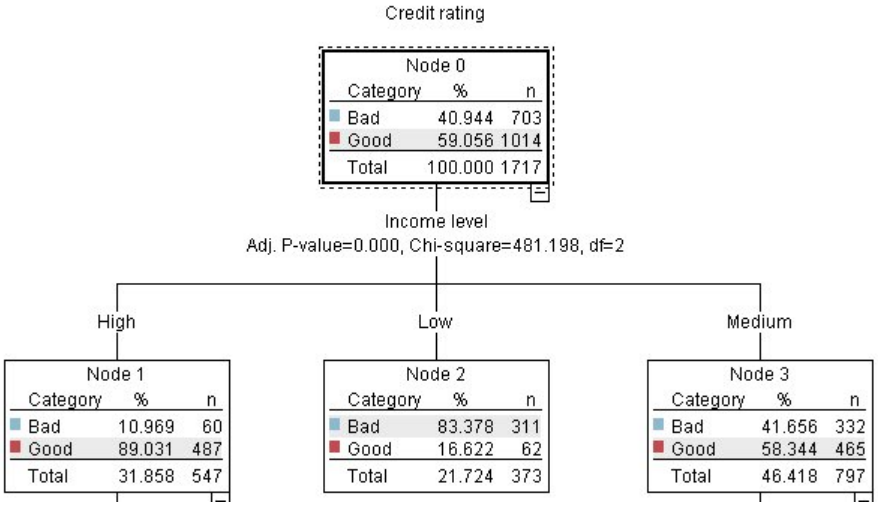


그림 *1.* 단순 의사결정 트리 모델

이 예에서는 일련의 의사결정 규칙을 사용하여 레코드를 분류하고 반응을 예측하는 **의사결정 트리** 모델을 사용합니다. 예를 들어, 다음과 같습니다.

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

이 예에서는 CHAID(Chi-squared Automatic Interaction Detection) 모델을 사용하지만, 일반적인 소개 목적으로 사용하는 것이며, 대부분의 개념은 IBM SPSS Modeler의 다른 모델링 유형에도 포괄적으로 적용됩니다.

모델을 이해하려면 먼저 모델에 입력된 데이터를 이해해야 합니다. 이 예의 데이터는 은행 고객에 대한 정보를 포함합니다. 다음 필드가 사용됩니다.

필드 이름	설명
Credit_rating	신용 등급: 0=나쁨, 1=좋음, 9=결측값
Age	나이
Income	수입 수준: 1=낮음, 2=중간, 3=높음

필드 이름	설명
Credit_cards	보유한 신용카드 수: 1=5개 미만, 2=5개 이상
교육	교육 수준: 1=고등학교, 2=대학교
Car_loans	자동차 구입 대출 건수: 1=없음 또는 1건, 2=2건 이상

은행은 대출을 갚았는지(신용 등급 = 양호), 아니면 체납했는지(신용 등급 = 불량) 여부를 포함하여 은행에서 대출을 받은 고객에 대한 히스토리 정보로 구성된 데이터베이스를 유지보수합니다. 은행은 이 기존 데이터를 사용하여 향후 대출 신청자가 대출을 체납할 확률을 예측할 수 있는 모델을 작성하려고 합니다.

의사결정 트리 모델을 사용하면 두 개 그룹의 고객 특성을 분석하고 대출 기본값의 우도를 예측할 수 있습니다.

이 예에서는 *Demos* 폴더의 *streams* 하위 폴더에서 사용 가능한 스트림, *modelingintro.str*을 사용합니다. 데이터 파일은 *tree_credit.sav*입니다. 자세한 정보는 5 페이지의 『Demos 폴더』의 내용을 참조하십시오.

스트림을 살펴보십시오.

1. 주 메뉴에서 다음을 선택하십시오.

   파일 > **스트림 열기**
2. 열기 대화 상자의 도구 모음에서 금색 너깃 아이콘을 클릭하고 Demos 폴더를 선택하십시오.
3. 스트림 폴더를 두 번 클릭하십시오.
4. *modelingintro.str* 파일을 두 번 클릭하십시오.

스트림 작성

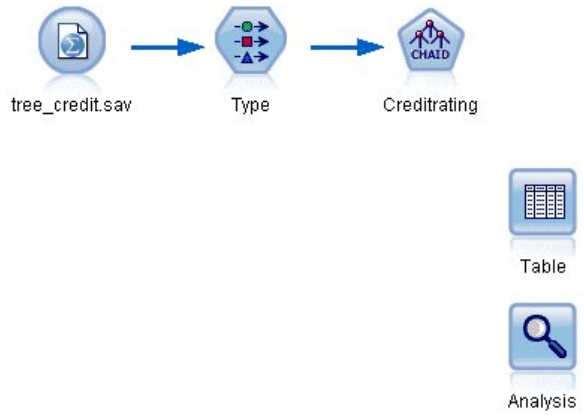


그림 2. 스트림 모델링

모델을 작성하는 스트림을 작성하려면 세 개 이상의 요소가 필요합니다.

- 일부 외부 소스에서 데이터를 읽는 소스 노드(이 경우 IBM SPSS Statistics 데이터 파일).
- 필드 특성(예: 필드가 포함하는 데이터 유형인 측정 수준)을 지정하는 소스 또는 유형 노드와 모델링에서 목표 또는 입력에 해당하는 각 필드의 역할.
- 스트림을 실행할 때 모델 너깃을 생성하는 모델링 노드.

이 예에서는 CHAID 모델링 노드를 사용합니다. CHAID(Chi-squared Automatic Interaction Detection)는 의사결정 트리에서 분할을 수행할 최상의 위치를 파악하기 위해 카이제곱 통계라고 알려진 특정 통계 유형을 사용하여 의사결정 트리를 작성하는 분류 방법입니다.

측정 수준이 소스 노드에 지정된 경우 별도의 유형 노드를 제거할 수 있습니다. 기능상 결과는 동일합니다.

이 스트림에는 모델 너깃을 작성하여 스트림에 추가한 후 스코어링 결과를 보기 위해 사용되는 테이블 및 분석 노드도 있습니다.

통계 파일 소스 노드는 IBM SPSS Statistics 형식으로 *tree_credit.sav* 데이터 파일(*Demos* 폴더에 설치됨)의 데이터를 읽습니다. (이름이 *$CLEO_DEMOS*인 특수 변수는 현재 IBM SPSS Modeler 설치 아래 이 폴더를 참조하는 데 사용됩니다. 그러면 현재 설치 폴더 또는 버전에 상관없이 경로가 유효합니다.)

그림 *3.* 통계 파일 소스 노드를 포함하는 데이터 읽기

유형 노드는 각 필드의 **측정 수준**을 지정합니다. 측정 수준은 필드에서 데이터 유형을 나타내는 범주입니다. 현재 소스 데이터 파일은 서로 다른 세 개의 측정 수준을 사용합니다.

연속형 필드(예: 나이 필드)는 연속된 숫자 값을 포함하지만, **명목** 필드(예: 신용 등급 필드)는 두 개 이상의 서로 다른 값(예: 불량, 양호 또는 신용 내역 없음)을 포함합니다. **순서** 필드(예: 소득 수준 필드)에서는 내재된 순서(이 경우 낮음, 중간, 높음)가 있는 여러 고유 값을 포함하는 데이터를 설명합니다.

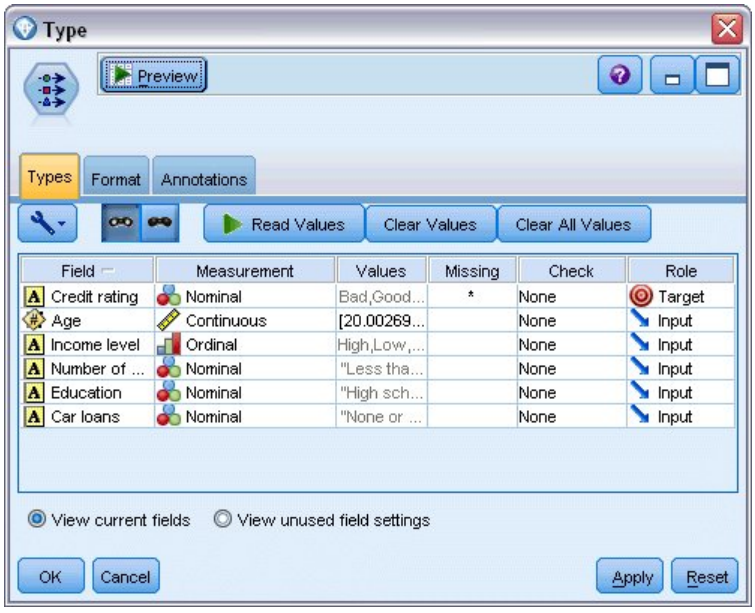


그림 *4.* 유형 노드에서 목표 및 입력 필드 설정

각 필드에서 유형 노드는 **역할**로 지정되어 각 필드가 모델링에서 수행하는 파트를 나타내기도 합니다. 역할이 신용 등급 필드에서 목표로 설정되어 있으며, 이 필드는 해당 고객이 대출을 체납했는지 여부를 표시합니다. 이는 **목표** 또는 값을 예측하려는 필드입니다.

다른 필드에서 역할을 입력으로 설정합니다. 때때로 입력 필드는 **예측변수**라고도 합니다. 또는 목표 필드의 값을 예측하기 위해 모델링 알고리즘에서 해당 값을 사용하는 필드입니다.

CHAID 모델링 노드는 모델을 생성합니다.

모델링 노드의 필드 탭에서 **사전 정의된 역할 사용** 옵션은 선택되어 있습니다. 즉, 유형 노드에 지정된 대로 목표 및 입력이 사용됩니다. 현재 필드 역할을 변경할 수 있지만, 이 예에서는 그대로 사용합니다.

1. 작성 옵션 탭을 클릭하십시오.

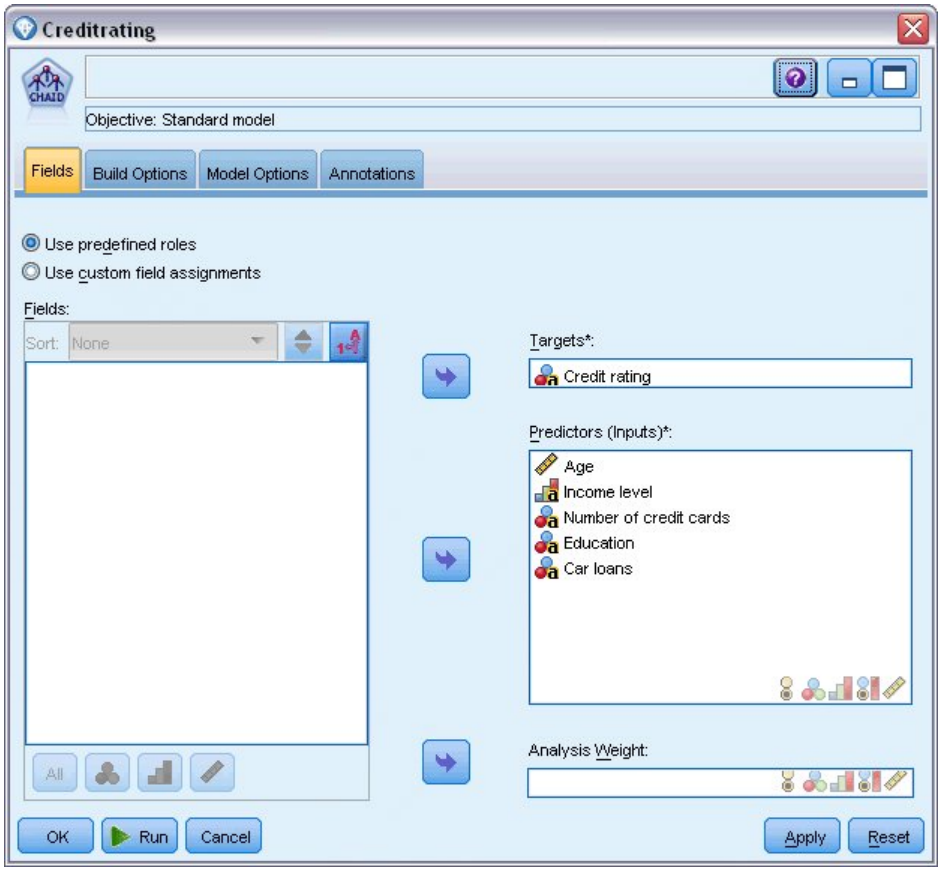


그림 *5. CHAID* 모델링 노드, 필드 탭

다음은 작성할 모델 종류를 지정할 수 있는 여러 옵션입니다.

완전히 새로운 모델을 사용하고자 합니다. 그래서 기본 옵션 **새 모델 작성**을 사용하겠습니다.

또한 개선사항 없이 하나의 표준 의사결정 트리 모델만 사용할 것이므로, 기본 목표 옵션 **단일 트리 작성**은 그대로 둡니다.

선택적으로 모델을 미세 조정할 수 있는 대화형 모델링 세션을 시작할 수 있지만, 이 예에서는 기본 모드 설정 **모델 생성**을 사용해서만 모델을 생성합니다.

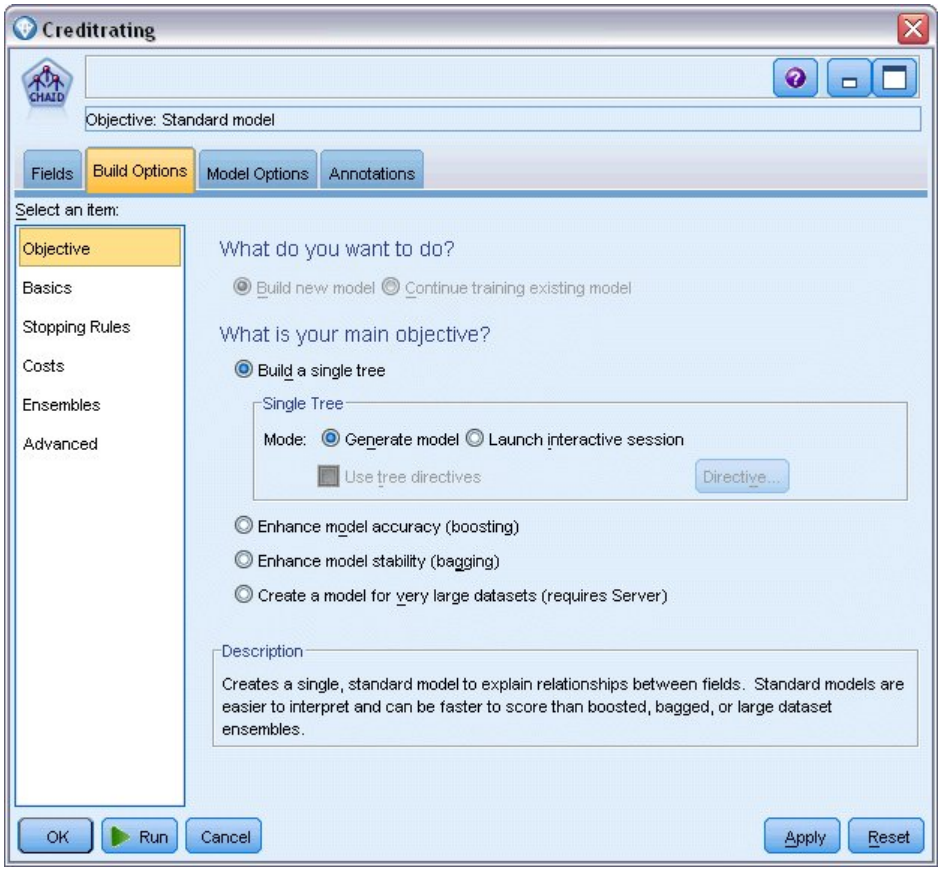


그림 *6. CHAID* 모델링 노드, 작성 옵션 탭

이 예에서는 트리를 단순하게 유지하기 위해 상위 및 하위 노드의 최소 케이스 수를 증가시켜 트리 성장을 제한합니다.

2. 작성 옵션 탭의 왼쪽에 있는 네비게이터 분할창에서 **중지 규칙**을 선택하십시오.
3. **절대값 사용** 옵션을 선택하십시오.
4. **부모 분기에서 최소 레코드 수**를 400으로 설정하십시오.
5. **자식 분기에서 최소 레코드 수**를 200으로 설정하십시오.

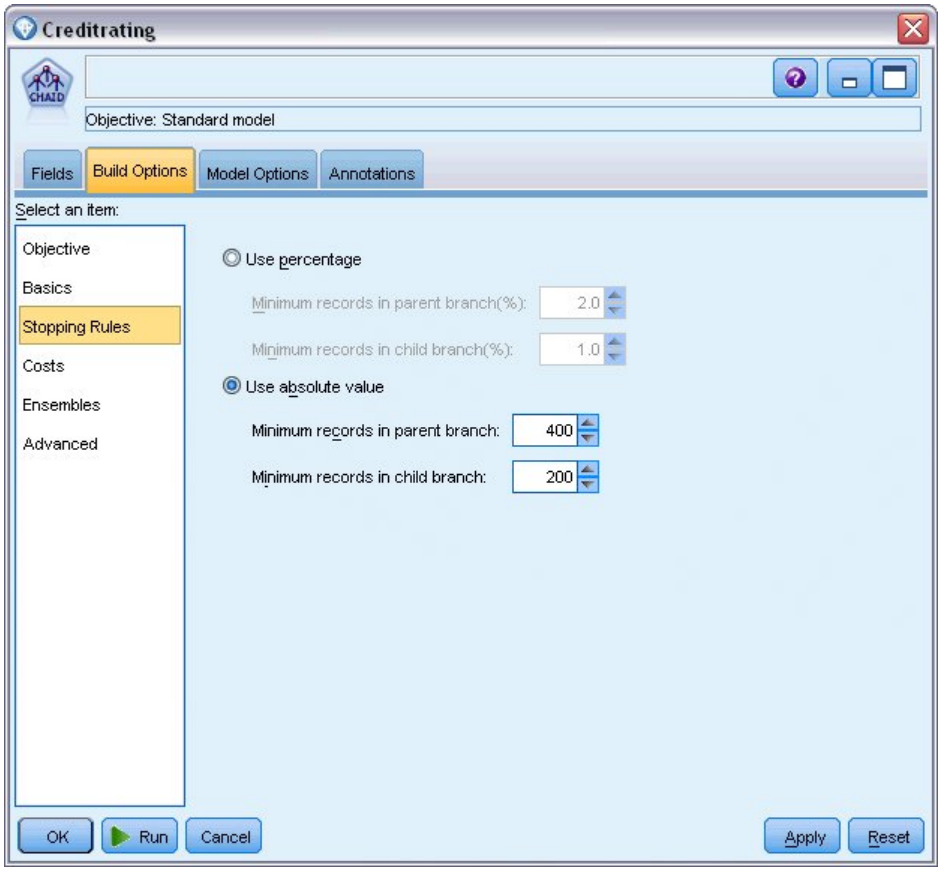


그림 7. 의사결정 트리 작성에 대한 중지 기준 설정

이 예에서 다른 모든 기본 옵션을 사용할 수 있습니다. 따라서 모델을 작성하려면 **실행**을 클릭하십시오. (또는 노드를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **실행**을 선택하거나 노드를 선택하고 도구 메뉴에서 **실행**을 선택하십시오.)

모델 찾아보기

실행을 완료하면 애플리케이션 창의 오른쪽 상단에 있는 모델 팔레트에 모델 너깃이 추가되고, 작성된 모델링 노드에 대한 링크를 포함하는 스트림 캔버스에도 배치됩니다. 모델 세부사항을 보려면 모델 너깃을 마우스 오른쪽 단추로 클릭하고 **찾아보기**(모델 팔레트에서) 또는 **편집**(캔버스에서)을 선택하십시오.

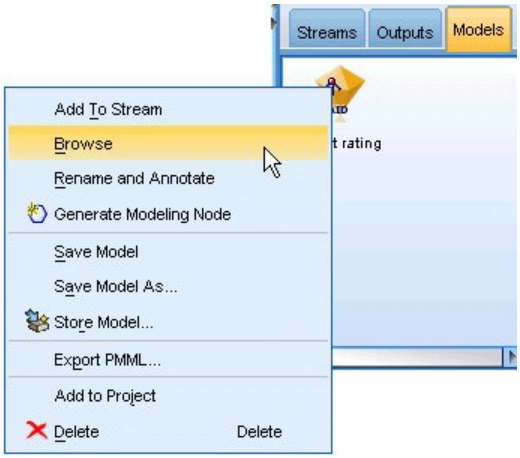


그림 *8.* 모델 팔레트

CHAID 너깃의 경우 모델 탭에서는 규칙 세트 양식으로 세부사항을 표시합니다. 특히 서로 다른 입력 필드 값에 기반하여 하위 노드에 개별 레코드를 지정하는 데 사용할 수 있는 일련의 규칙 세트입니다.

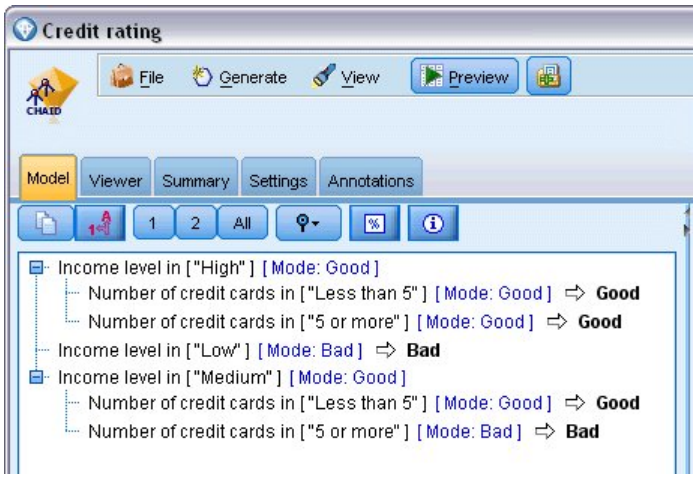


그림 *9. CHAID* 모델 너깃, 규칙 세트

각 의사결정 트리 터미널 노드(즉, 추가로 분할되지 않는 트리 노드)의 경우 안전 또는 위험의 예측이 리턴됩니다. 각각의 케이스에서 예측은 **모드** 또는 해당 노드 범위에 포함된 레코드의 경우 가장 일반적인 반응으로 판별됩니다.

규칙 세트의 오른쪽에서 모델 탭은 모델 추정 시 각 예측변수의 상대적 중요도를 나타내는 예측변수 중요도 차트를 표시합니다. 여기에서 소득 수준이 이 케이스에 가장 중요하며, 또 다른 중요한 요인은 신용카드 수임을 알 수 있습니다.

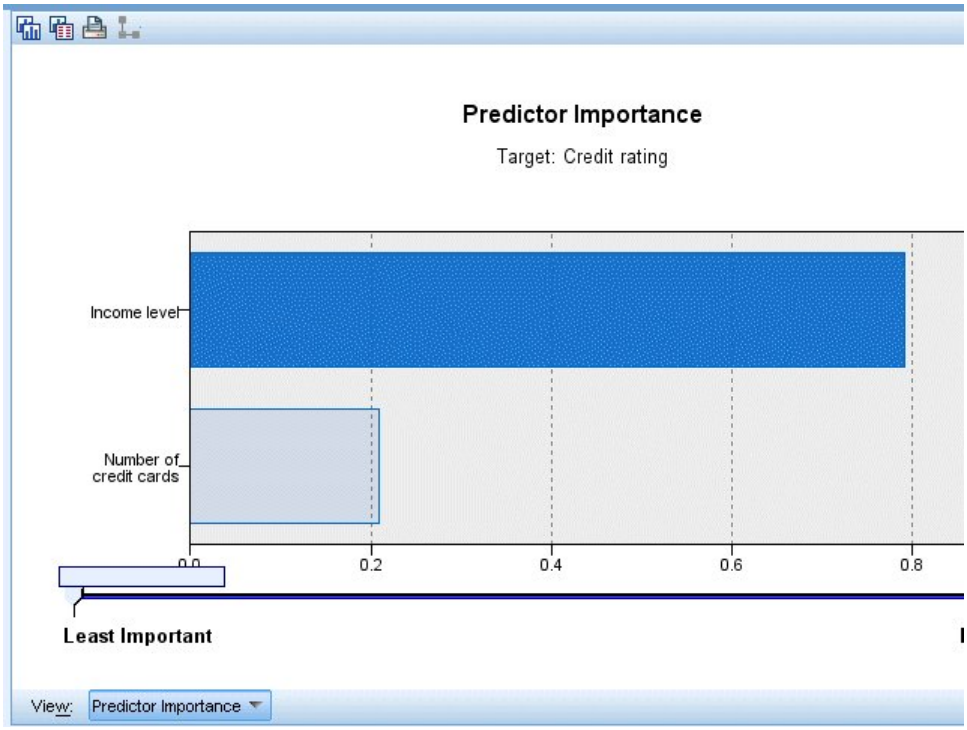


그림 *10.* 예측변수 중요도 차트

모델 너깃의 뷰어 탭에서는 각 의사결정 포인트에서 노드를 포함하는 동일한 모델(트리 양식)을 표시합니다. 도구 모음에서 확대/축소 제어를 사용하여 특정 노드에서 확대하거나 축소하여 추가 트리를 확인합니다.

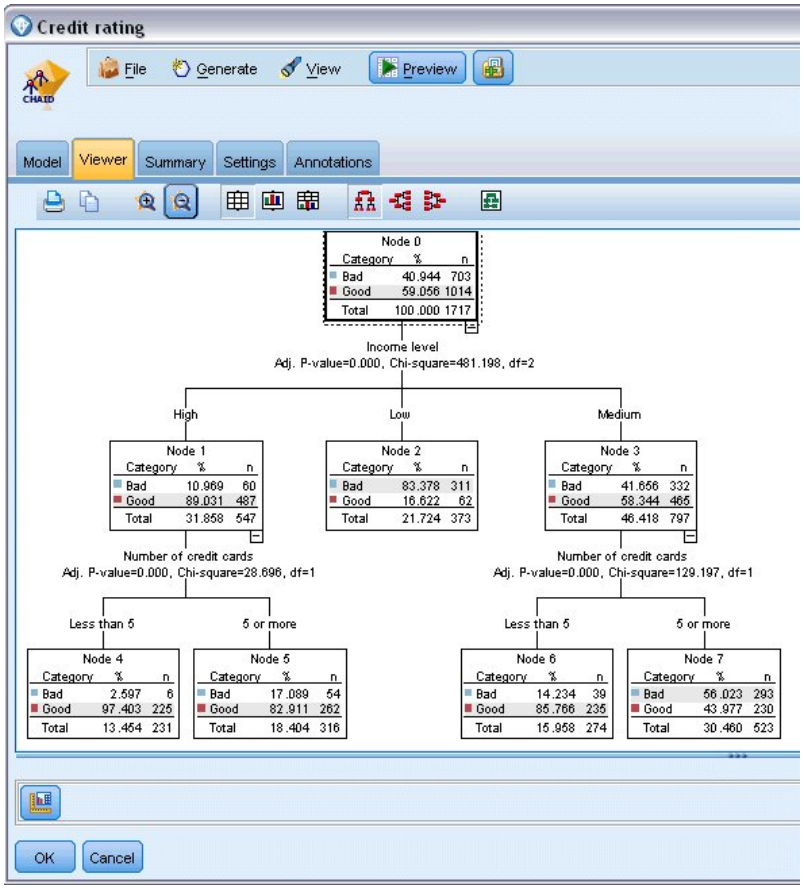


그림 *11.* 축소가 선택된 모델 너깃의 뷰어 탭

트리의 위쪽을 보면 첫 번째 노드(노드 0)는 데이터 세트의 모든 레코드에 대한 요약을 제공합니다. 데이터 세트에서 케이스의 40% 이상만 잘못된 위험으로 분류됩니다. 이는 상당히 높은 비율이므로, 트리가 책임이 있는 요인에 관한 단서를 제공할 수 있는지 살펴보도록 하겠습니다.

첫 번째 분할이 소득 수준을 기준으로 함을 확인할 수 있습니다. 소득 수준이 낮음 범주인 레코드는 노드 2에 지정되고, 이 범주가 대출 체납자 중 가장 높은 퍼센트를 포함한다는 점도 놀랍지 않습니다. 확실히 이 범주의 고객에게 대출할 경우 높은 위험이 수반됩니다.

그러나 이 범주의 고객 중 16%는 실제로 체납하지 않았기 때문에 예측이 항상 올바른 것은 아닙니다. 모델이 모든 반응을 현실적으로 예측할 수는 없지만, 좋은 모델은 사용 가능한 데이터에 기반하여 각 레코드의 가능성 높은 반응을 예측하도록 이끌어야 합니다.

같은 방식으로 소득이 높은 고객을 살펴보면(노드 1) 대다수(89%)가 위험에 안전하다는 사실을 알 수 있습니다. 그러나 이러한 고객에서 10명 중 2명 이상이 체납했습니다. 여기서 위험을 최소화하기 위해 대출 기준을 미세 조정할 수 있을까요?

모델이 보유한 신용카드 번호에 기반하여 이러한 고객을 두 개 하위 범주(노드 4와 5)로 구분하는 방법에 주의하십시오. 소득이 높은 고객의 경우 신용카드가 5개 미만인 고객에게만 대출하는 경우 89%에

서 97%로 성공률을 높일 수 있으며, 보다 만족스러운 결과가 나옵니다.

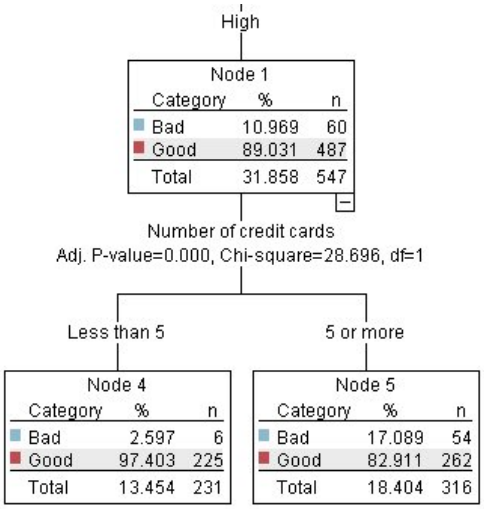


그림 *12.* 소득이 높은 고객의 트리 보기

중간 수입 범주(노드 3)에 속하는 고객에 대해서는 어떻게 생각하십니까? 이들은 안전과 위험 등급 사이에서 훨씬 균등하게 구분됩니다.

다시 하위 범주(이 경우 노드 6과 7)는 도움이 될 수 있습니다. 현재, 신용카드가 5개 미만인 중간 소득 고객에게만 대출하면 안전 등급이 58%에서 85%로 늘어나 크게 개선시켰습니다.

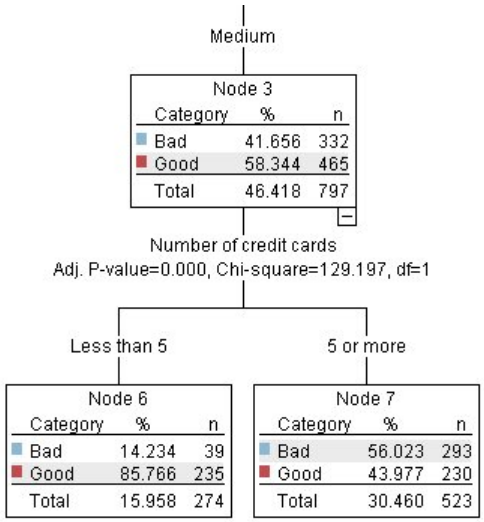


그림 *13.* 중간 소득 고객의 트리 보기

지금까지 모델에 입력된 모든 레코드가 특정 노드에 지정되고, 해당 노드의 가장 일반적인 반응에 기반하여 안전 또는 위험이라는 예측이 지정되는 과정을 훈련했습니다.

개별 레코드에 예측을 지정하는 이러한 프로세스는 **스코어링**이라고 합니다. 모델을 추정하는 데 사용된 동일한 레코드를 스코어링하면 훈련 데이터(결과를 아는 데이터)에서 작업의 정확도를 평가할 수 있습니다. 이제 이 방법에 대해 알아보도록 하겠습니다.

모델 평가

스코어링 작업 방식을 이해하기 위해 모델을 찾아보고자 합니다. 그러나 작동 방식의 정확도를 평가하려면 일부 레코드의 스코어를 계산하고 모델에서 예측한 반응과 실제 결과를 비교해야 합니다. 모델을 추정하는 데 사용했던 동일한 레코드 스코어를 계산하려고 합니다. 그러면 관측 반응과 예측 반응을 비교할 수 있습니다.

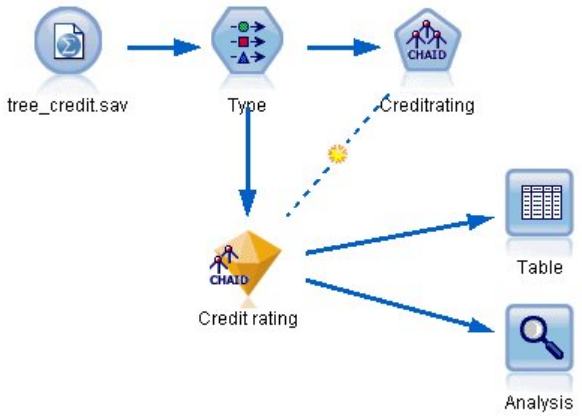
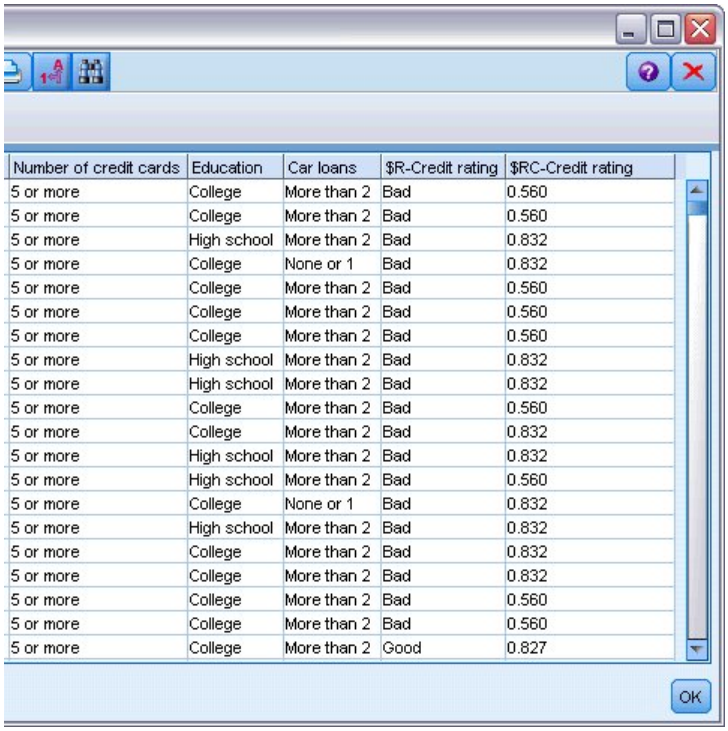


그림 *14.* 모델 평가를 위해 출력 노드에서 모델 너깃 첨부

1. 스코어 또는 예측을 확인하려면 모델 너깃에 테이블 노드를 첨부하고 테이블 노드를 두 번 클릭하고 **실행**을 클릭하십시오.

   테이블은 모델에서 작성된 이름이 *$R-Credit rating*인 필드에서 예측 스코어를 표시합니다. 이러한 값을 실제 반응을 포함하는 원래 신용 등급 필드와 비교할 수 있습니다.

   보통 스코어링 중에 생성된 필드 이름은 표준 접두문자가 아닌 목표 필드에 기반합니다. 접두문자 *$G* 및 *$GE*는 일반화 선형 모델에서 생성되고, *$R*은 이 케이스에서 CHAID 모델에서 생성된 예측에 사용되는 접두문자이며, *$RC*는 신뢰도 값에 사용되고, *$X*는 일반적으로 앙상블을 사용하여 생성되며 *$XR*, *$XS*, *$XF*는 목표 필드가 연속형, 범주형, 세트 또는 플래그 필드인 경우 각각 접두문자로 사용됩니다. 서로 다른 모델 유형은 서로 다른 접두문자 집합을 사용합니다. **신뢰도**는 모델의 추정값으로, 각 예측값의 정확도를 0.0에서 1.0의 척도로 나타낸 값입니다.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

OK

그림 *15.* 생성된 스코어 및 신뢰도 값을 표시하는 테이블

예상대로 예측값은 전부는 아니지만 많은 레코드에서 실제 반응과 매치됩니다. 이유는 각 CHAID 터미널 노드에 혼합된 반응이 있기 때문입니다. 예측은 가장 일반적인 항목과 매치되지만 해당 노드에서 나머지 모든 항목에서는 잘못될 수 있습니다. (체납하지 않은 낮은 소득 고객 중 16%라는 소수를 소환합니다.)

이러한 상황을 방지하기 위해 모든 노드가 혼합된 반응 없이 100% 모두 안전 또는 위험이 될 때까지 트리를 더 작은 분기로 계속 분할할 수 있습니다. 그러나 이러한 모델은 매우 복잡해질 수 있으며, 다른 데이터 세트에 대해 일반화되지 않을 수도 있습니다.

올바른 예측이 얼마나 되는지 정확히 파악하기 위해 테이블을 검토하고 예측 필드 *$R-Credit rating*의 값이 신용 등급 값과 매치하는 레코드 수의 합계를 계산할 수 있습니다. 다행히 분석 노드를 사용할 수 있는 훨씬 쉬운 방법이 있으며 여기서는 이를 자동으로 수행합니다.

2. 모델 너깃을 분석 노드에 연결하십시오.
3. 분석 노드를 두 번 클릭하고 **실행**을 클릭하십시오.

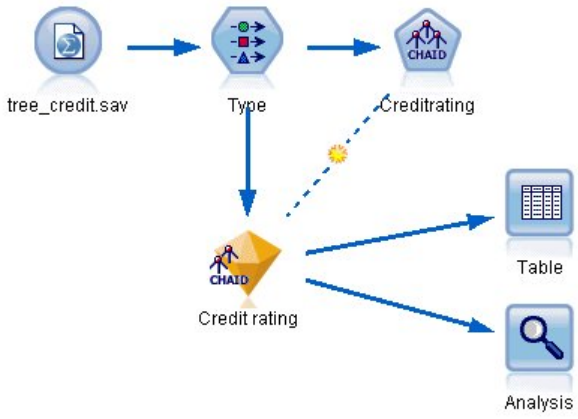


그림 *16.* 분석 노드 첨부

이 분석에서는 2464개 레코드 중 1899개(77% 초과)의 경우 모델에서 예측한 값이 실제 반응과 매치됨을 보여줍니다.

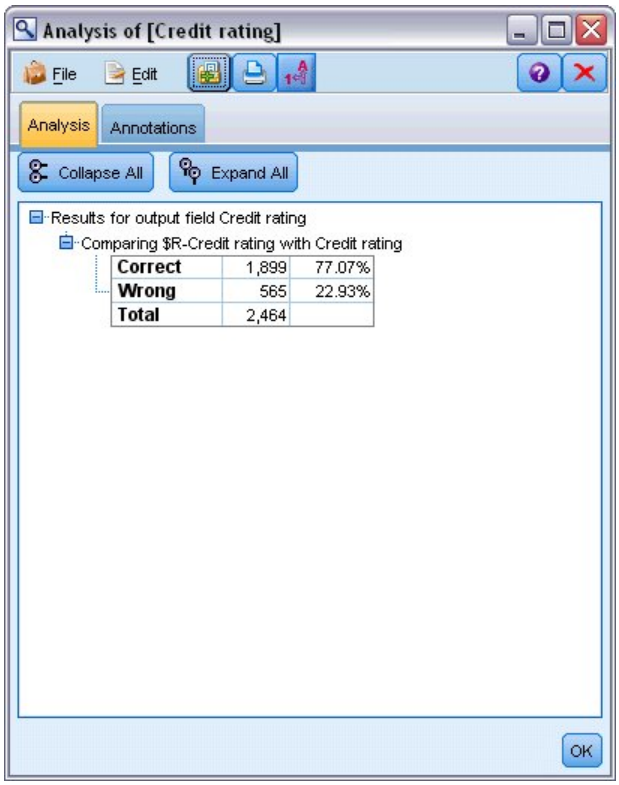


그림 *17.* 관측 및 예측 반응을 비교한 분석 결과

이 결과는 스코어링하는 레코드가 모델을 추정하는 데 사용된 것과 동일하다는 사실로 제한됩니다. 실제 상황에서는 파티션 노드를 사용하여 훈련 및 평가를 위해 데이터를 별도의 샘플로 분할할 수 있습니다.

하나의 표본 파티션을 사용하여 모델을 생성하고 다른 표본으로 이를 검정하면 다른 데이터 세트에서 일반화할 때 효율성을 효과적으로 표시할 수 있습니다.

분석 노드에서는 이미 실제 결과를 알고 있는 레코드에 대해 모델을 검정할 수 있습니다. 다음 단계에서는 모델을 사용하여 결과를 모르는 레코드의 스코어를 계산하는 방법을 보여줍니다. 예를 들어, 여기에는 현재 은행 고객은 아니지만, 판촉 메일링의 잠재적 목표인 사람이 포함될 수 있습니다.

레코드 스코어링

이전에는 모델의 정확도를 평가하기 위해 모델을 추정하는 데 사용된 동일한 레코드 스코어를 계산했습니다. 지금은 모델 작성 시 사용한 항목으로부터 서로 다른 레코드 세트 스코어를 계산하는 방법을 확인하고자 합니다. 이는 아직 모르는 결과를 예측할 수 있는 패턴을 식별하기 위해 목표 필드(결과를 아는 스터디 레코드)를 포함하는 모델링의 목표입니다.

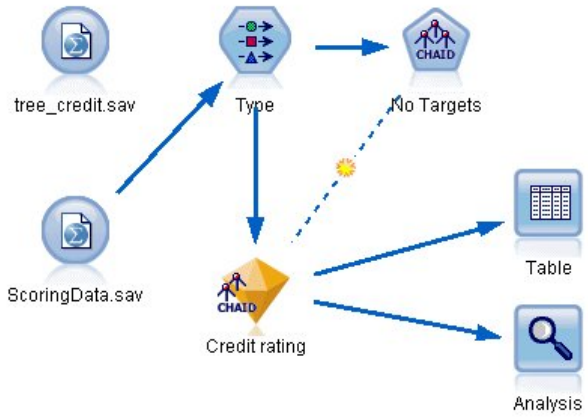


그림 *18.* 스코어링에 대한 새 데이터 첨부

다른 데이터 파일을 가리키도록 통계량 파일 소스 노드를 업데이트하거나 스코어를 계산하려는 데이터를 읽을 새 소스 노드를 추가할 수 있습니다. 어느 방법이든, 새 데이터 세트는 모델(나이, 소득 수준, 교육 등)에서 사용하는 동일한 입력 필드를 포함해야 합니다(목표 필드 신용 등급은 포함하지 않음).

또는 예상 입력 필드를 포함하는 스트림에 모델 너깃을 추가할 수 있습니다. 파일이나 데이터베이스 등 읽는 위치에 상관없이 소스 유형은 필드 이름과 유형이 모델에서 사용하는 항목과 일치하는 한, 중요하지 않습니다.

또한 모델 너깃을 별도의 파일로 저장하거나 이 형식을 지원하는 다른 애플리케이션에 대한 PMML 형식으로 모델을 내보내거나 엔터프라이즈 범위의 배치, 스코어링 및 모델 관리를 제공하는 IBM SPSS Collaboration and Deployment Services 리포지토리에 모델을 저장할 수 있습니다.

사용된 인프라에 상관없이 모델은 동일한 방식으로 작동합니다.

요약

이 예에서는 모델 작성, 평가, 스코어링에 대한 기본 단계를 설명합니다.

- 모델링 노드는 결과가 알려진 레코드를 연구하여 모델을 추정하고 모델 너깃을 작성합니다. 때때로 이 작업을 모델 훈련이라고도 합니다.
- 모델 너깃은 레코드 스코어링을 위해 예상 필드를 포함하는 스트림에 추가할 수 있습니다. 결과를 이미 아는 사용자(예: 기존 고객)의 레코드 스코어를 계산하면 수행 성과를 평가할 수 있습니다.
- 모델의 수행 성과에 만족하면 반응 수준을 예측하도록 새 데이터(예: 잠재 고객)의 스코어를 계산할 수 있습니다.
- 모델을 훈련하거나 추정하는 데 사용되는 데이터는 분석 또는 히스토리 데이터라고도 합니다. 스코어링 데이터는 운영 데이터라고도 합니다.

제 3 장 모델링 개요

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 안내서에서는 모델링 프로세스에 대한 일반적인 소개와 함께 이러한 여러 방법의 예제를 제공합니다. 이 안내서는 온라인 자습서 및 PDF 형식으로 사용 가능합니다. 자세한 정보는 5 페이지의 『애플리케이션 예제』 주제를 참조하십시오.

모델링 방법은 다음 범주로 나뉩니다.

- 감독
- 연관
- 세분화

감독 모델

감독 모델은 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, 상대오차 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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 찾은 경향을 줄여 더 많은 분할을 허용하는 입력을 선호하도록 설계되었습니다. 입력 필드는 숫자 범위(연속)일 수 있지만 목표 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. 목표 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 목표 필드를 사용합니다.

일반화 선형 모델은 종속변수가 요인과 선형적으로 관련되고 지정된 연결 함수를 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 연결 함수를 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 재교육할 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 가까이에 있는 *k* 오브젝트의 범주 또는 값과 연관시킵니다. 유사한 케이스는 서로 가까이에 있고 유사하지 않은 케이스는 서로 멀리 떨어져 있습니다.

STP(Spatio-Temporal Prediction) 노드는 위치 데이터, 예측(예측자)을 위한 입력 필드, 시간 필드 및 목표 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 모델은 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 연관이 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 훈련 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.

CARMA 모델은 입력 또는 목표 필드를 지정하지 않아도 데이터에서 규칙 세트를 추출합니다. Apriori와 대조적으로 CARMA 노드는 단지 전항 지원이 아니라 규칙 지원(전항 및 후항 둘 다에 대한 지원)을 위한 작성 설정을 제공합니다. 이는 생성된 규칙을 보다 다양한 애플리케이션에 사용하여, 예를 들어 후항이 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 Server과 함께 사용할 수 있습니다.

세분화 모델

세분화 모델은 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 모델은 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 구별해줍니다. 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 사용합니다(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 측도를 제공하려고 시도하는 기본 측도를 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 것 입니다. 모델 너깃에서 각 단위별로 캡처된 관측값을 살펴 강한 단위를 식별할 수 있습니다. 이것은 적당한 군집 수에 대한 감각을 제공할 것 입니다.

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 훈련 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 안내서를 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 스코어할 수 있습니다.

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 너깃에서 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 확인 상자 설정을 통해 단일 모델을 작성할 것인지, 아니면 분할 모델을 작성할 것인지 선택할 수 있습니다.

분할 필드가 정의되었지만, 확인 상자가 선택되지 않은 경우 단일 모델만 생성됩니다. 마찬가지로 확인 상자를 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 52 페이지의 『분할 모델 뷰어』의 내용을 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.

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.

분할 모델을 지원하는 모델링 노드

많은 모델링 노드에서 분할 모델을 작성할 수 있습니다. 예외로는, 자동 군집, 시계열, PCA/요인, 필드선택, SLRM, 임의 트리, Tree-AS, Linear-AS, LSVM, 연관 모델(Apriori, Carma, 시퀀스), 군집 모델(K-평균, 코호넨, 2단계, 이상 항목), Statistics 모델, In-Database 모델링에 사용된 노드가 있습니다.

분할 모델링을 지원하는 모델링 노드는 다음과 같습니다.


C&R 트리


Bayes Net


선형


QUEST


GenLin


GLMM


CHAID


KNN


시계열


C5.0


Cox


STP


신경망


자동 분류자


One-Class SVM


의사결정 목록


자동 숫자


XGBoost Tree


회귀분석


로지스틱


XGBoost Linear


판별


SVM

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 서브세트가 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 예측변수 중요도의 계산(모델 추정 시 예측변수 입력 필드의 상대적 중요도)을 지원하지 않습니다. 예측변수 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 387 페이지의 『KNN 노드』의 내용을 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.

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 확인 상자를 선택합니다. 이 형식의 레코드는 2개 필드(ID와 내용에 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 288 페이지의 『테이블 대 트랜잭션 데이터』의 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 입력으로 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 검정함으로써 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 분할로 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 30 페이지의 『분할 모델 작성』의 내용을 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』의 내용을 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』의 내용을 참조하십시오.
- **후항.** 규칙 귀납 노드(Apriori)의 경우 결과로 생성된 규칙 세트에서 후항으로 사용할 필드를 선택합니다. (이는 유형 노드에서 역할이 목표 또는 둘 다인 필드에 대응합니다.)
- **전항.** 규칙 귀납 노드(Apriori)의 경우 결과로 생성된 규칙 세트에서 전항으로 사용할 필드를 선택합니다. (이는 유형 노드에서 역할이 입력 또는 둘 다인 필드에 대응합니다.)

일부 모델은 이 절에서 설명한 것과 다른 필드 탭이 있기도 합니다.

- 자세한 정보는 306 페이지의 『시퀀스 노드 필드 옵션』의 내용을 참조하십시오.
- 자세한 정보는 292 페이지의 『CARMA 노드 필드 옵션』의 내용을 참조하십시오.

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.

모델링 노드 분석 옵션

많은 모델링 노드가 원래 및 수정된 성향 스코어와 함께 예측변수 중요도 정보를 확보할 수 있는 분석 탭을 포함합니다.

모델 평가

예측변수 중요도 계산. 적절한 중요도 측도를 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 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 몇몇 모델은 기본적으로 해제되어 있음에 유의하십시오. 예측변수 중요도는 의사결정 목록 모델에 사용할 수 없습니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 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 38 페이지의 『성향 스코어』의 내용을 참조하십시오.

원시 성향 스코어 계산. 원시 성향 스코어는 훈련 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 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 훈련 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 138 페이지의 『부스팅 C5.0 모델』의 내용을 참조하십시오.

다음을 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 훈련하는 데 사용되는 파티션만큼 많은 레코드를 최소한으로 포함해야 합니다.

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 *4.* 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 아니오인 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 *0.85*인 아니오 예측은 *0.15*의 원시 성향(또는 *1 - 0.85*)으로 변환됩니다.

표 *5.* 신뢰도 스코어

고객	예측	신뢰도
Joe Smith	응답함	.35
Jane Smith	응답하지 않음	.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 *37* 페이지의 『모델링 노드 분석 옵션』의 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 맵핑되는 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.

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 내용을 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.

참고: 의사결정 트리 모형에서만 작성 시 비용을 지정할 수 있습니다.

모델 너깃

그림 *19.* 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 52 페이지의 『스트림에서 모델 너깃 사용』의 내용을 참조하십시오.

그림 *20.* 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 안내서를 참조하십시오.

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 43 페이지의 『모델 교체』의 내용을 참조하십시오.

모델 링크 정의 및 제거

캔버스에서 수동으로 링크를 정의 및 제거할 수 있습니다. 새 링크를 정의하는 경우 커서가 링크 커서로 변경됩니다.

그림 *21.* 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 제거**를 선택하십시오.

기존 링크 제거(주 메뉴)

1. 링크를 제거하려는 모델링 노드 또는 너깃을 클릭하십시오.
2. 주 메뉴에서 다음을 선택하십시오.

   편집 > **노드** > **모델 링크 제거**

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(43 페이지의 『모델 교체』 참조)를 보유합니다.

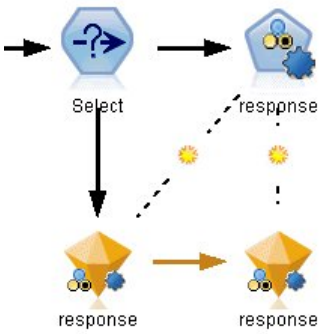


그림 22. 링크된 너깃 복사 및 붙여넣기

링크된 모델링 노드와 함께 너깃을 복사하고 붙여넣는 경우 오브젝트를 붙여넣는 위치(동일한 스트림 또는 새 스트림)에 상관없이 링크가 유지됩니다.

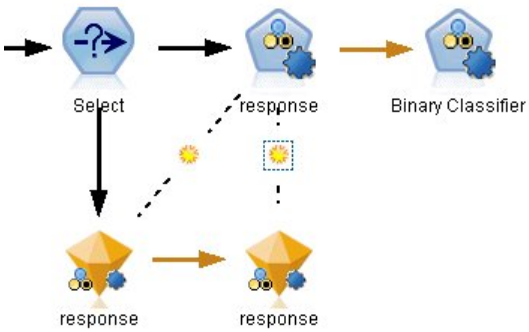


그림 *23.* 링크된 너깃 복사 및 붙여넣기

참고:: 모델링 노드 없이 링크된 너깃을 복사하고 너깃을 새 스트림 또는 모델링 노드를 포함하지 않는 수퍼 노드에 붙여넣는 경우 링크가 끊어지고 너깃만 붙여넣습니다.

모델 링크 및 수퍼 노드

모델링 노드 또는 링크된 모델(둘 중 하나만)의 모델 너깃을 포함하도록 수퍼 노드를 정의하는 경우 링크가 끊어집니다. 수퍼 노드를 확장해도 링크는 복원되지 않습니다. 수퍼 노드 작성을 실행 취소해야만 복원할 수 있습니다.

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 *24.* 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 *25.* 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.

모델 팔레트

모델 팔레트(관리자 창의 모델 탭에 있음)에서는 다양한 방식으로 모델 너깃을 사용, 검사, 수정할 수 있습니다.

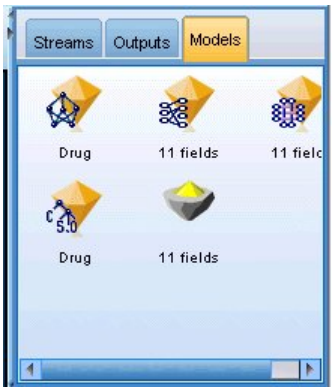


그림 *26.* 모델 팔레트

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 54 페이지의 『PMML로서 모델 가져오기 및 내보내기』의 내용을 참조하십시오.

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 내용을 인쇄합니다. 하위 메뉴에서 인쇄하려는 정보를 선택합니다.
- **인쇄 보기.** 현재 보기 또는 모든 보기를 인쇄합니다.
- **텍스트 내보내기.** 너깃의 내용을 텍스트 파일로 내보냅니다. 하위 메뉴에서 내보내려는 정보를 선택합니다.
- **HTML 내보내기.** 너깃의 내용을 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 54 페이지의 『PMML로서 모델 가져오기 및 내보내기』의 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 56 페이지의 『스코어링 어댑터에 대한 모델 게시』의 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 행의 수는 10개입니다. 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.

모델 너깃 요약/정보

모델 너깃의 요약 탭 또는 정보 보기에서는 필드, 작성 설정, 모들 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 배깅 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 사용디는 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.

예측변수 중요도

일반적으로, 가장 중요한 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

예측변수 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 예측변수 중요도는 모델링 노드의 분석 탭에서 사용할 수 있습니다. 자세한 정보는 37 페이지의 『모델링 노드 분석 옵션』의 내용을 참조하십시오. KNN 모델의 경우 389 페이지의 『이웃』의 내용을 참조하십시오.

참고: 예측변수 중요도는 분할 모델에서 지원되지 않습니다. 예측변수 중요도 설정은 분할 모델을 작성할 때 무시됩니다. 자세한 정보는 30 페이지의 『분할 모델 작성』의 내용을 참조하십시오.

예측변수 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, 기본적으로 이 모델에서는 사용되지 않습니다. 많은 예측변수를 포함하는 데이터 세트를 사용하는 경우 필드선택 노드를 사용하는 초기 선별은 더 빠른 결과를 제공할 수 있습니다(아래 참조).

- 예측변수 중요도는 사용 가능한 경우 검정 파티션에서 계산됩니다. 그렇지 않으면 훈련 데이터가 사용됩니다.
- SLRM 모델의 경우 예측변수 중요도가 사용 가능하지만, SLRM 알고리즘에 의해 계산됩니다. 자세한 정보는 374 페이지의 『SLRM 모델 너깃』의 내용을 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 49 페이지의 『중요도에 기반하여 변수 필터링』의 내용을 참조하십시오.

예측변수 중요도 및 필드선택

모델 너깃에 표시된 예측변수 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, 예측변수 중요도 차트는 이 특정 모델에 대한 각 입력의 상대적 중요도를 표시합니다. 따라서 필드선택은 선별 입력보다 보수적입니다. 예를 들어, 직위 및 작업 범주가 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, 예측변수 중요도는 모델을 미세 조정할 때 가장 유용합니다.

단일 모델과 자동화된 모델링 노드 사이의 예측변수 중요도 차이

개별 노드에서 단일 모델을 작성하는지, 아니면 자동화된 모델링 노드를 사용하여 결과를 생성하는지에 따라 예측변수 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 예측변수 중요도 값을 얻기 위해 자동화된 모델링 노드 대신 단일 노드를 사용하도록 제안합니다.

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(예측변수 중요도)**

또는

> **필드 선택(예측변수 중요도)**

최상위 변수. 가장 중요한 예측변수를 지정된 수까지 포함하거나 제외합니다.

다음보다 큰 중요도. 상대적 중요도가 지정된 값보다 큰 모든 예측변수를 포함하거나 제외합니다.

앙상블 뷰어

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.

모델 요약: 모델 요약 보기는 앙상블 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 앙상블의 경우, 참조 모델은 전체 훈련 파티션에 작성된 표준 모델입니다. 부스팅된 앙상블의 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 앙상블을 작성하는 데 사용된 구성요소 모델 중에서 "의견의 다양성"이 표시됩니다. 크게 표시되는 것이 더 다양한 형식입니다. 이것은 기본 모델에서 예측이 얼마나 다양한지에 대한 측도입니다. 다양성은 부스팅된 앙상블 모델에 대해 사용할 수 없으며 연속형 목표에 대해 표시되지 않습니다.

예측변수 중요도: 일반적으로, 가장 중요한 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

예측변수 중요도는 모든 앙상블 모델에 사용할 수 있는 것은 아닙니다. 예측변수 세트는 구성요소 모델에서 다양할 수 있지만, 적어도 하나의 구성요소 모델에서 사용된 예측변수에 대해 중요도가 계산될 수 있습니다.

예측변수 빈도: 예측변수 세트는 모델링 방법의 선택 또는 예측변수 선택으로 인해 구성요소 모델에서 다양할 수 있습니다. 예측변수 빈도 그림은 앙상블의 구성요소 모델에서 예측변수의 분포를 표시하는 점 도표입니다. 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.

구성요소 모델 정확도: 차트는 구성요소 모델의 예측 정확도에 대한 점 도표입니다. 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 앙상블에 대한 색상 코드화된 선을 표시합니다. 스코어링에 사용될 모델에 해당하는 선 옆에 체크 표시가 나타납니다.

상호작용성. 결합 규칙을 변경하면 차트가 업데이트됩니다.

부스팅된 앙상블. 부스팅된 앙상블에 대해 선형 차트가 표시됩니다.

구성요소 모델 세부사항: 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. 훈련 표본의 가중치를 부여한 입력 레코드 수입니다.

자동 데이터 준비:

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.

이상값 자르기 작업은 표시되는 경우, 절사 값을 넘는 연속형 예측변수 값(평균에서 3배 표준 편차)이 절사 값으로 설정되었음을 나타냅니다.

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.

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.

스트림에서 모델 너깃 사용

모델 너깃은 새 데이터 스코어를 계산하도록 스트림에 배치되고 새 노드를 생성합니다. 데이터 **스코어링**을 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 301 페이지의 『연관 모델 너깃에서 규칙 세트 생성』의 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.

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 168 페이지의 『의사결정 목록 모델 옵션』의 내용을 참조하십시오.
- 시계열 모델의 경우 **기존 모델을 사용하여 추정 계속** 옵션이 사용 가능하므로, 현재 데이터로 이전 모델을 재생성할 수 있습니다. 자세한 정보는 시계열 모델 옵션의 내용을 참조하십시오.

PMML로서 모델 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 55 페이지의 『PMML을 지원하는 모델 유형』 주제를 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 디렉토리와 모델의 고유 이름을 지정하십시오.

참고: 사용자 옵션 대화 상자에서 PMML 내보내기의 옵션을 변경할 수 있습니다. 기본 메뉴에서 다음을 클릭하십시오.

도구 > **옵션** > **사용자 옵션**

PMML 탭을 클릭하십시오.

저장된 모델을 PMML로서 가져오기

IBM SPSS Modeler 또는 또 다른 애플리케이션에서 PMML로서 내보낸 모델은 모델 팔레트로 가져올 수 있습니다. 자세한 정보는 55 페이지의 『PMML을 지원하는 모델 유형』의 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. **열기**를 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.

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
- 선형 회귀
- 신경망
- C5.0
- 로지스틱 회귀분석
- Genlin
- SVM
- Apriori
- Carma
- K-평균
- 코호넨
- 이단계
- GLMM(고정 효과 전용 GLMM 모델에만 지원됨)
- 의사결정 목록
- Cox
- 순차규칙(순차규칙 PMML 모델에 대한 스코어링은 지원되지 않음)
- Statistics 모델

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 Modeler은 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 Modeler은 인지되고 스코어된 유효한 PMML을 스코어링하려고 시도합니다. 모든 PMML이 스코어할지 또는 이를 생성한 애플리케이션과 같은 방식으로 스코어할지는 보장이 되지 않습니다.

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.

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 *27.* 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 목표 필드의 값을 예측할 규칙 세트 노드를 생성합니다. 자세한 정보는 301 페이지의 『연관 모델 너깃에서 규칙 세트 생성』의 내용을 참조하십시오.

제 4 장 선별 모델

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, 필드선택을 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.)

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 결측값이 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 필드 선택이 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.

예. 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 목표 필드는 세 가지 각 제안에 고객이 반응하는지 여부를 보여줍니다. 회사는 이 데이터를 사용하여 고객이 향후에 유사한 제안에 반응할 가능성을 예측할 수 있습니다.

요구사항. 단일 목표 필드(해당 역할이 목표으로 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.

필드선택 모델 설정

모델 탭의 설정에는 입력 필드 선별 기준을 미세 조정할 수 있는 설정과 함께 표준 모델 옵션이 포함되어 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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.

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 62 페이지의 『필드선택 모델 결과』의 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 측도에 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 측도. 값은 0(연관 없음)부터 1(완벽한 연관)까지 범위입니다.
- **람다.** 목표 값을 예측하기 위해 변수를 사용하는 경우 오차 내 비례 축소를 반영하는 연관성 측도. 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.

필드선택 모델 너깃

필드선택 모델 너깃에서는 필드선택 노드에서 순위화한 대로, 선택한 목표와 관련된 각 입력의 중요도를 표시합니다. 순위화 전에 선별된 필드도 나열됩니다. 자세한 정보는 59 페이지의 『필드선택 노드』의 내용을 참조하십시오.

필드선택 모델 너깃을 포함하는 스트림을 실행할 때 모델은 모델 탭에서 현재 선택이 표시한 대로, 선택한 입력만 유지하는 필터 역할을 합니다. 예를 들어, 중요로 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.

필드선택 모델 결과

필드선택 모델 너깃의 모델 탭에서는 상위 분할창에 있는 모든 입력의 순위 및 중요도를 표시하고, 왼쪽에 있는 열의 확인 상자를 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 확인 상자를 사용하여 이러한 필드를 포함하거나 삭제할 수 있습니다. 자세한 정보는 60 페이지의 『필드선택 모델 설정』의 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 선택』의 내용을 참조하십시오.
- 중요, 주변 또는 중요하지 않음으로 입력을 순위화할 때 임계값은 테이블 아래 범례에 표시됩니다. 이러한 값은 모델링 노드에 지정됩니다. 자세한 정보는 61 페이지의 『필드선택 옵션』의 내용을 참조하십시오.

중요도에 따라 필드 선택

필드선택 모델 너깃을 사용하여 데이터를 스코어링하는 경우 순위화 또는 선별된 필드(왼쪽 열의 확인 상자로 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.

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 너깃도 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.

이상 항목 발견 노드

이상 항목 발견 모델은 데이터에서 이상값 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 이상값이 알려진 패턴을 따르지 않을 경우에도 이상값을 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.

이상값을 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 필드 수를 조사할 수 있습니다. 각 레코드는 해당 피어 그룹에 다른 레코드와 비교되어 가능한 이상 항목을 식별할 수 있습니다. 케이스가 보통의 중심에서 멀어질 수록 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(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, 필드선택을 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.)

예. 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 **입력**으로 설정된 필드만 입력으로 사용할 수 있음에 유의하십시오. 목표 필드(**목표** 또는 **둘 다**에 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.

이상 항목 발견 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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($0-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.

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 시스템이 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 이상값의 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 이상값 레이블이 붙습니다. 이상값 군집에는 식별 번호 –1이 지정됩니다. 이상값 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 이상값 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 이상값이 배치됩니다. 그렇지 않은 경우에는 군집화의 두 번째 단계 중에 이상값이 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.

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 사항에 따라 판별된 많은 새 필드가 스트림에 추가됩니다. 자세한 정보는 『이상 항목 발견 모델 설정』의 내용을 참조하십시오. 새 필드 이름은 다음 테이블에 요약된 것처럼 모델 이름을 기준으로 지정되고 *$O* 접두문자가 붙습니다.

표 *6.* 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 *n*번째인 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 설정』의 내용을 참조하십시오.

이상 항목 발견 모델 세부사항

생성된 이상 항목 발견 모델의 모델 탭은 모델의 피어 그룹에 대한 정보를 표시합니다.

피어 그룹 크기 및 통계는 훈련 데이터를 기준으로 한 추정값으로, 동일한 데이터로 실행하더라도 실제 스코어링 결과와 약간 다를 수 있음에 유의하십시오.

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.

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 65 페이지의 『이상 항목 발견 모델 옵션』의 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 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.

제 5 장 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 측도에 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, 상대오차 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 측도를 제공하려고 시도하는 기본 측도를 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 샘플 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 59 페이지의 『필드선택 노드』의 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.

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 훈련 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.

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 Net, 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 77 페이지의 『자동 분류자 노드 삭제 옵션』의 내용을 참조하십시오. 자세한 정보는 86 페이지의 『자동 군집 노드 삭제 옵션』의 내용을 참조하십시오.

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 실험할 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 69 페이지의 제 5 장 『자동화된 모델링 노드』의 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 플래그인 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 범주형이며, 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 상관 없이 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 35 페이지의 『빈도 및 가중 필드 사용』의 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 74 페이지의 『자동 분류자 노드 고급 옵션』의 내용을 참조하십시오.

자동 분류자 노드 모델 옵션

자동 분류자 노드의 모델 탭으로 모델을 비교하는 데 사용되는 기준과 함께 작성할 모델 수를 지정할 수 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 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** 모델이 제대로 예측한, 총 레코드 수에 상대적인 레코드 퍼센트입니다.
- **ROC 곡선 아래 영역** ROC 곡선은 모델 성능에 대한 지수를 제공합니다. 곡선이 참조선보다 위에 있을수록 검정이 더 정확합니다.
- **이익(누적)** 지정된 비용, 수입, 가중치 기준에 따라 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 적중 비율입니다(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 필드의 수를 기준으로 하여 모델의 순위를 정합니다.

모델 순위화 사용. 파티션이 사용 중인 경우 순위가 훈련 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.

예측변수 중요도 계산. 적절한 중요도 측도를 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 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

이익 기준. 참고. 플래그 목표만 해당됩니다. 이익은 각 레코드의 수입에서 레코드의 비용을 뺀 값입니다. 분위수의 이익은 단순히 분위수의 전체 레코드 이익 합계입니다. 이익은 적중에만 적용된다고 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. 참고. 플래그 목표만 해당됩니다. 리프트 계산에 사용할 백분위수를 지정합니다. 결과를 비교할 때 이 값도 변경될 수 있음에 유의하십시오. 자세한 정보는 87 페이지의 『자동화된 모델 너깃』의 내용을 참조하십시오.

자동 분류자 노드 고급 옵션

자동 분류자 노드의 고급 탭으로 파티션을 적용하고(가능한 경우), 사용할 알고리즘을 선택하고, 중지 규칙을 지정할 수 있습니다.

모델 선택. 기본적으로 전체 모델이 작성하도록 선택되지만 Analytic Server가 있으면, 모델을 Analytic Server에서 실행할 수 있는 모델로 제한하도록 선택하고 분할 모델을 작성하거나 매우 큰 데이터 세트를 처리할 준비가 되도록 사전 설정할 수 있습니다.

참고: 자동 분류자 노드에 로컬 Analytic Server 모델 작성은 지원되지 않습니다..

사용한 모델. 왼쪽 열의 확인 상자를 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 모수. 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 훈련 방법 중에서 하나를 선택하기 보다는 모두를 선택하여 단일 전달에서 여섯 개의 모델을 훈련할 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, 이단계, SVM, KNN, Bayes Net, 의사결정 목록 모델만) 한 모델의 최대 시간 제한을 설정합니다. 예를 들어, 일부 복잡한 상호작용으로 인해 특정 모델의 훈련에 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 가까이에 있는 *k* 오브젝트의 범주 또는 값과 연관시킵니다. 유사한 케이스는 서로 가까이에 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 목표 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 찾은 경향을 줄여 더 많은 분할을 허용하는 입력을 선호하도록 설계되었습니다. 입력 필드는 숫자 범위(연속)일 수 있지만 목표 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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. 목표 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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.

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 내용을 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.

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 72 페이지의 『자동 분류자 노드 모델 옵션』의 내용을 참조하십시오.

선택적으로 지정된 모든 기준을 만족하는 모델을 처음 생성할 때 실행을 중지하도록 노드를 구성할 수 있습니다. 자세한 정보는 70 페이지의 『자동화된 모델링 노드 중지 규칙』의 내용을 참조하십시오.

자동 분류자 노드 설정 옵션

자동 분류자 노드의 설정 탭에서 너깃에 사용 가능한 스코어 시간 옵션을 미리 구성할 수 있습니다.

앙상블 모델이 생성한 필드 필터링. 앙상블 노드에 반영되는 개별 모델이 생성한 모든 추가 필드의 출력에서 제거합니다. 모든 입력 모델에서 결합된 스코어에만 관심이 있는 경우 이 확인 상자를 선택합

니다. 예를 들어, 분석 노드 또는 평가 노드를 사용하여 각 개별 입력 모델의 정확도와 결합된 스코어의 정확도를 비교하려는 경우에는 이 옵션을 선택 취소해야 합니다.

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 하우징 값을 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 69 페이지의 제 5 장 『자동화된 모델링 노드』의 내용을 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 노드를 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 35 페이지의 『빈도 및 가중 필드 사용』의 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-AS가 있습니다. 추가 정보는 80 페이지의 『자동 숫자 노드 고급 옵션』의 내용을 참조하십시오.

자동 숫자 노드 모델 옵션

자동 숫자 노드의 모델 탭에서는 모델 비교에 사용되는 기준과 함께 저장할 모델 수를 지정할 수 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 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 측도로, 값이 1에 가까울수록 더 강한 관계임을 나타냅니다. (상관관계 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 목표 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 정확

하므로 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 측도를 제공합니다.

모델 순위화 사용. 파티션이 사용 중인 경우 순위의 기준(훈련 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.

예측변수 중요도 계산. 적절한 중요도 측도를 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 47 페이지의 『예측변수 중요도』의 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 70 페이지의 『자동화된 모델링 노드 중지 규칙』의 내용을 참조하십시오.

자동 숫자 노드 고급 옵션

자동 숫자 노드의 고급 탭에서는 중지 규칙을 사용 및 지정할 알고리즘과 옵션을 선택할 수 있습니다.

모델 선택. 기본적으로 전체 모델이 작성하도록 선택되지만 Analytic Server가 있으면, 모델을 Analytic Server에서 실행할 수 있는 모델로 제한하도록 선택하고 분할 모델을 작성하거나 매우 큰 데이터 세트를 처리할 준비가 되도록 사전 설정할 수 있습니다.

참고: 자동 숫자 노드에 로컬 Analytic Server 모델 작성은 지원되지 않습니다..

사용한 모델. 왼쪽 열의 확인 상자를 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 모수. 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 훈련 방법 중에서 하나를 선택하기 보다는 모두를 선택하여 단일 전달에서 여섯 개의 모델을 훈련할 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, 이단계, SVM, KNN, Bayes Net, 의사결정 목록 모델만) 한 모델의 최대 시간 제한을 설정합니다. 예를 들어, 일부 복잡한 상호작용으로 인해 특정 모델의 훈련에 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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 연결 함수를 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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 연결 함수를 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.

자동 숫자 노드 설정 옵션

자동 숫자 노드의 설정 탭에서는 너깃에서 사용 가능한 스코어-시간 옵션을 사전 구성할 수 있습니다.

앙상블 모델이 생성한 필드 필터링. 앙상블 노드에 반영되는 개별 모델이 생성한 모든 추가 필드의 출력에서 제거합니다. 모든 입력 모델에서 결합된 스코어에만 관심이 있는 경우 이 확인 상자를 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.

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 집단을 캡처하고 이 집단에 대한 유용한 설명을 제공하는 기능으로 판별됩니다. 자세한 정보는 261 페이지의 제 11 장 『군집 모델』의 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 유형에는 이단계, K-평균, 코호넨, One-Class SVM 및 K-평균-AS가 있습니다.

자동 군집 노드 모델 옵션

자동 군집 노드의 모델 탭으로 모델을 비교하는 데 사용되는 기준과 함께 저장할 모델 수를 지정할 수 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 측도를 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 훈련 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 측도

기본 순위화 측도, 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, A는 케이스로부터 케이스가 속한 군집의 중심값까지의 거리이고 B는 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.

자동 군집 노드 고급 옵션

자동 군집 노드의 고급 탭으로 파티션을 적용하고(가능한 경우), 사용할 알고리즘을 선택하고, 중지 규칙을 지정할 수 있습니다.

모델 선택. 기본적으로 전체 모델이 작성하도록 선택되지만 Analytic Server가 있으면, 모델을 Analytic Server에서 실행할 수 있는 모델로 제한하도록 선택하고 분할 모델을 작성하거나 매우 큰 데이터 세트를 처리할 준비가 되도록 사전 설정할 수 있습니다.

참고: 자동 군집 노드에서의 Analytic Server 모델 로컬 작성은 지원되지 않습니다..

사용한 모델. 왼쪽 열의 확인 상자를 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 모수. 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 훈련 방법 중에서 하나를 선택하기 보다는 모두를 선택하여 단일 전달에서 여섯 개의 모델을 훈련할 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, 이단계, SVM, KNN, Bayes Net, 의사결정 목록 모델만) 한 모델의 최대 시간 제한을 설정합니다. 예를 들어, 일부 복잡한 상호작용으로 인해 특정 모델의 훈련에 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 것 입니다. 모델 너깃에서 각 단위별로 캡처된 관측값을 살펴 강한 단위를 식별할 수 있습니다. 이것은 적당한 군집 수에 대한 감각을 제공할 것 입니다.

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 훈련 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. 여기에서는 데이터 포인트를 사전 정의된 군집 수로 모읍니다. SPSS Modeler에서 K-평균-AS는 Spark로 구현됩니다. K-평균 알고리즘에 대한 자세한 정보는 https://spark.apache.org/docs/2.2.0/ml-clustering.html의 내용을 참조하십시오. K-평균-AS 노드는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

자동 군집 노드 삭제 옵션

자동 군집 노드의 삭제 탭을 사용하여 일정한 기준에 일치하지 않는 모델을 자동으로 삭제할 수 있습니다. 이 모델은 모델 너깃에 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 84 페이지의 『자동 군집 노드 모델 옵션』의 내용을 참조하십시오.

선택적으로 지정된 모든 기준을 만족하는 모델을 처음 생성할 때 실행을 중지하도록 노드를 구성할 수 있습니다. 자세한 정보는 70 페이지의 『자동화된 모델링 노드 중지 규칙』의 내용을 참조하십시오.

자동화된 모델 너깃

자동화된 모델링 노드를 실행하면 노드는 가능한 모든 옵션 조합에 대해 후보 모델을 추정하고, 사용자가 지정한 측도에 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 측도가 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 설정**이 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 72 페이지의 『자동 분류자 노드 모델 옵션』의 내용을 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 72 페이지의 『자동 분류자 노드 모델 옵션』의 내용을 참조하십시오.

연속형 목표

- 연속(숫자 범위) 모델의 경우 그래프는 각 모델의 관측 값에 대한 예측을 구성하여, 둘 사이의 상관 관계에 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 79 페이지의 『자동 숫자 노드 모델 옵션』의 내용을 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 84 페이지의 『자동 군집 노드 모델 옵션』의 내용을 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.

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.

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.

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 예측변수 중요도를 모두 표시합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

자동 분류자의 경우 분포 그래프가 표시되지만, 자동 숫자의 경우 다중 도표(산점도라고도 함)가 표시됩니다.

제 6 장 의사결정 트리

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 귀납이라 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 훈련하기 전에 관련 필드로 데이터를 축소하는 데 사용할 수 있습니다.

의사결정 트리 모형 너깃은 if-then 규칙 콜렉션(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 파티셔닝하는 방식을 확인하려는 경우에 유용합니다. 트리-AS 노드 출력은 규칙 세트를 작성할 필요 없이 너깃에 직접 규칙 목록을 포함시키므로 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 프로파일을 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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 목표 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. 목표 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 훈련 레코드를 세그먼트로 분할합니다. 여기서 트리의 노드는 노드의 케이스의 100%가 목표 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 선택합니다.

상호작용 식별 특정 하위 그룹에만 관련된 관계를 식별하고 정규 모수 모델에 이를 지정합니다.

범주 병합 및 연속형 변수 밴딩 최소의 정보 손실로 그룹 예측변수 범주 및 연속형 변수를 다시 코딩합니다.

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 너깃을 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 예측변수 필드를 선택하고 필요한 경우 추가 모델 옵션을 지정하십시오. 특정 지시사항의 경우 각 트리 작성 노드에 대한 문서를 참조하십시오.
3. 작성 옵션 탭의 목적 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.

주석

- 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 106 페이지의 『직접 트리 모델 작성』의 내용을 참조하십시오.

트리 성장 및 가지치기

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 정의』의 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 108 페이지의 『C&R 트리 노드』의 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 *28.* 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.

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 95 페이지의 『예측변수 세부사항 보기』의 내용을 참조하십시오.

3. 각 분할에 대한 기본 조건을 수락하거나 **사용자 정의**를 선택하여 분할에 대한 조건을 적절히 지정할 수 있습니다.
   - 연속형(숫자 범위) 예측변수의 경우 **범위 값 편집** 필드를 사용하여 각 새 노드에 포함되는 값의 범위를 지정할 수 있습니다.
   - 범주형 예측변수의 경우 **세트 값 편집** 또는 **순서 값 편집** 필드를 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 대용』의 내용을 참조하십시오.)

예측변수 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 예측변수에서 나열됩니다. 예측변수가 숫자 범위인 경우 *F* 통계량이 표시됩니다. 카이제곱 통계량은 분할 필드에서 목표 필드가 얼마나 독립되어 있는지 정도의 측도입니다. 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 예측변수의 개선도가 표시됩니다. 개선도가 클수록 예측변수가 사용된 경우 상위와 하위 노드 사이의 불순도가 더 많이 감소합니다. (순수한 노드는 모든 케이스가 단일 목표 범주에 속하는 노드입니다. 트리에서 불순도가 낮을수록 모델이 데이터에 더 적합합니다.) 즉, 일반적으로 개선도가 높은 그림은 이 트리 유형에서 유용한 분할을 표시합니다. 사용되는 불순도 측도는 트리 작성 노드에서 지정됩니다.

분할 세부사항 및 대용

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 대용 및 1차 분할 필드 사이의 대응에 대한 측도이며, 일반적으로 분할 필드와 가장 비슷한 항목이 "최상"의 대용입니다. C&R 트리 및 QUEST의 경우 1차 예측변수 대신 사용되는 대용도 함께 나열됩니다.

선택한 노드의 분할을 편집하려면 대용 패널 왼쪽에 있는 아이콘을 클릭하여 분할 정의 대화 상자를 열면 됩니다. (단축 아이콘으로, 아이콘을 클릭하여 1차 분할 필드로 선택하기 전에 목록에서 대용을 선택할 수 있습니다.)

대용. 적용 가능한 경우 선택한 노드에 대한 기본 분할 필드의 대용이 표시됩니다. 대용은 주어진 레코드의 기본 예측변수 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.

대용은 CHAID 트리에서 사용되지 않지만, 사용자 정의 분할을 정의할 때 특정 하위 노드에 이를 지정하는 옵션이 제공됩니다. 자세한 정보는 94 페이지의 『사용자 정의 분할 정의』의 내용을 참조하십시오.

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 세부사항을 보십시오. 자세한 정보는 95 페이지의 『분할 세부사항 및 대용』의 내용을 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.

이익

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.

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 2배임을 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 행당 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 구성하는 데 둘 이상의 노드

가 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.

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 *yes*는 $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. **ROI**는 노드에서 총 이익을 총 지출로 나눈 값입니다.

주석

- 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.

회귀분석 이익

회귀분석 트리의 경우 노드별, 누적 노드별, 분위수 보기 사이에서 선택할 수 있습니다. 테이블에는 평균값이 표시됩니다. 차트는 수량에서만 사용할 수 있습니다.

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 측도를 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.

(증분의 적중 수 / 총 적중 수) x 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, 훈련 데이터 세트에 있는 전체 적중 퍼센트와 적중에 해당하는 각 증분에 있는 레코드 퍼센트를 비교합니다.

(증분의 적중 수 / 증분의 레코드 수) / (총 적중 수 / 총 레코드 수)

반응 차트

반응 차트는 테이블의 반응*(%)* 열에 있는 값을 구성합니다. 반응은 다음 방정식을 사용하여 계산된, 적중에 해당하는 증분에 있는 레코드의 퍼센트입니다.

(증분의 반응 수 / 증분의 레코드 수) x 100%

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 *n*개 노드.
   - 지정된 레코드 수까지 상위 노드.
   - 훈련 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 104 페이지의 『필터 및 선택 노드 생성』의 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.

위험

위험은 모든 수준에서 오분류의 확률을 알려줍니다. 위험 탭에서는 포인트 위험 추정값과 오분류 테이블(범주형 출력의 경우)을 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.

트리 모델 및 결과 저장

다음을 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 104 페이지의 『모델, 이익, 위험 정보 내보내기』의 내용을 참조하십시오.

트리 작성기 또는 트리 모델 너깃에서 다음을 수행할 수 있습니다.

- 현재 트리를 기반으로 필터 또는 선택 노드를 생성합니다. 자세한 정보는 104 페이지의 『필터 및 선택 노드 생성』의 내용을 참조하십시오.
- 트리의 터미널 분기를 정의하는 규칙 세트로 트리 구조를 표시하는 규칙 세트 너깃을 생성합니다. 자세한 정보는 105 페이지의 『의사결정 트리에서 규칙 세트 생성』의 내용을 참조하십시오.
- 또한 트리 모델 너깃의 경우에만 모델을 PMML 형식으로 내보낼 수 있습니다. 자세한 정보는 44 페이지의 『모델 팔레트』 주제를 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. 참고: 스트림의 다른 노드에 연결되어 있을 때에는 너깃에 대해서만 작동합니다. 자세한 정보는 138 페이지의 『그래프 생성』의 내용을 참조하십시오.

참고: 대화형 트리 자체는 저장할 수 없습니다. 작업을 유실하지 않으려면 트리 작성기 창을 닫기 전에 모델을 생성하고/하거나 트리 지시문을 업데이트하십시오.

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 **모두**에서 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 지시문』의 내용을 참조하십시오.

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 104 페이지의 『트리 지시문 업데이트』의 내용을 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 시작**을 선택하여 제어를 사용 가능하게 하고 **트리 지시문 사용**을 선택하고 **지시문**을 클릭하십시오.

지시문 명령문

지시문은 루트 노드부터 시작하여 트리를 성장시키는 조건을 지정합니다. 예를 들어, 트리를 한 수준 성장시키려면:

```
Grow Node Index 0 Children 1 2
```

예측변수를 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:

```
Grow Tree
```

트리 성장 및 가지치기를 수행하려면(C&R 트리만 해당):

```
Grow_And_Prune Tree
```

연속형 예측변수에 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 예측변수에서 분할하려면:

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

값이 여러 개인 명목 예측변수에서 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 예측변수에서 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.

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 102 페이지의 『트리 성장 지시문』의 내용을 참조하십시오.

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 **그래프**를 적절히 선택하고 하위 메뉴에서 내보낼 특정 항목을 선택하십시오.

해당되는 경우 내보내기는 현재 선택에 기반합니다.

텍스트 또는 HTML 형식 내보내기. 훈련 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 훈련 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.

필터 및 선택 노드 생성

트리 작성기 창에서 또는 의사결정 트리 모형 너깃을 찾아볼 때 메뉴에서 다음을 선택하십시오.

생성 > **필터 노드**

또는

> **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.

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 대부분의 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 띌 만큼 큰 영역은 데이터에 결측값이 있는 경우입니다.

규칙 세트는 범주형 목표 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.

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 작성**을 선택한 경우 **모드**를 **모델 생성**으로 설정해야 합니다.

   C5.0의 경우 모델 탭에서 **출력 유형**을 **의사결정 트리**로 설정하십시오.

   Tree-AS의 경우 작성 옵션 탭의 기본 패널에서 **트리 성장 알고리즘** 유형을 선택하십시오.
3. 목표 및 예측변수 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.

의사결정 트리 노드

IBM SPSS Modeler의 의사결정 트리 노드는 다음 트리 작성 알고리즘에 대한 액세스를 제공합니다.

- C&R 트리
- QUEST
- CHAID
- C5.0
- Tree-AS
- 랜덤 트리

자세한 정보는 91 페이지의 『의사결정 트리 모형』의 내용을 참조하십시오.

알고리즘은 데이터를 작은 하위 그룹으로 분할하여 반복적으로 의사결정 트리를 구성할 수 있다는 점에서 유사합니다. 그러나 일부 중요한 차이가 있습니다.

입력 필드. 입력 필드(예측변수)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 제곱 방법이 사용됩니다. CHAID 및 Tree-AS는 카이제곱 검정을 사용합니다. QUEST는 범주형 예측변수에 대해 카이제곱 검정을 사용하고 연속형 입력에 대해 공차 분석을 사용합니다. C5.0의 경우 정보 이론 측도가 사용됩니다(정보 이익 비율).

결측값 처리. 모든 알고리즘은 예측변수 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, 예측변수 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 너깃에서 설정된 규칙을 생성할 수 있도록 합니다. 자세한 정보는 105 페이지의 『의사결정 트리에서 규칙 세트 생성』의 내용을 참조하십시오.

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 큽니다. 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 예측변수 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 훈련 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.

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 유의성을 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 예측변수에 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.

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 계속적인(수적 범위) 입력 필드 또는 많은 범주의 입력 필드를 위해 분류 트리 방법에서 찾은 경향을 줄이는 것입니다.

- QUEST는 노드에서 입력 필드를 평가하기 위해 유의성 검정에 기반하여 일련의 규칙을 사용합니다. 선택 목적으로 노드의 각 입력에서 최소 단일 검정을 수행해야 합 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.

의사결정 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, 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 수동으로 대상, 예측변수 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 선택하십시오.

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 35 페이지의 『빈도 및 가중 필드 사용』의 내용을 참조하십시오.

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.

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 104 페이지의 『트리 지시문 업데이트』의 내용을 참조하십시오.
- **모델 정확도(부스팅) 개선** 모델 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 블

록으로 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.

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 예측변수에 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 정의**를 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 트리와 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 대용. 대용은 결측값을 처리하기 위한 방법입니다. 트리의 각 분할에서 알고리즘은 선택한 분할 필드와 가장 유사한 입력 필드를 식별합니다. 이러한 필드를 해당 분할의 대용이라고 합니다. 레코드를 분류해야 하지만 분할 필드에 결측값이 있으면 대용 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.

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.

의사결정 트리 노드 - 앙상블

이 설정은 부스팅, 배깅 또는 아주 큰 데이터 세트가 목적에서 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.

모델 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. 모델 정확도 또는 안정성을 향상시키는 것이 목표일 때 작성할 기본 모델 수를 지정하십시오. 배깅의 경우, 붓스트랩 표본의 수입니다. 양의 정수여야 합니다.

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 내용을 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.

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, 예측변수 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, k는 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 내용을 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 **평준화** 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 투잉 불순도 측도를 사용하는 트리의 트리 성장 프로세스로 비용 정보를 직접 통합할 수 있습니다. (이 옵션을 선택하지 않으면 비용 정보는 투잉 측도에 기반하여 레코드를 분류하고 트리의 위험 추정값을 계산하는 데만 사용됩니다.)

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 내용을 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.

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. **불순도**는 각 그룹 내 광범위한 출력 필드 값 범위를 보유하는 트리에서 정의된 부집단의 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 부집단을 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.

- **Gini**는 분기에 대한 범주 소속 확률에 기반한 일반적인 불순도 측도입니다.
- **투잉**은 이분형 분할을 강조하는 불순도 측도로, 분할에서 대략적으로 균등한 크기의 분기를 생성할 수 있습니다.
- **정렬**은 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, **생성**을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 유사 난수 정수를 작성합니다.

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, **생성**을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 유사 난수 정수를 작성합니다.

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. 예측변수의 다양한 범주 조합을 검정할 때 유의성 값을 조정합니다. 값은 범주 수와 예측변수의 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.

- **피어슨.** 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 변

화가 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. 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, **생성**을 클릭하여 1과 2147483647 사이(1과 2147483647 포함)의 유사 난수 정수를 작성합니다.

의사결정 트리 노드 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 예측변수 중요도 정보 및 플래그 목표의 원래 및 수정된 성향 스코어를 확보할 수도 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

모델 평가

예측변수 중요도 계산. 적절한 중요도 측도를 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 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 몇몇 모델은 기본적으로 해제되어 있음에 유의하십시오. 예측변수 중요도는 의사결정 목록 모델에 사용할 수 없습니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 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 38 페이지의 『성향 스코어』의 내용을 참조하십시오.

원시 성향 스코어 계산. 원시 성향 스코어는 훈련 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 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 훈련 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 138 페이지의 『부스팅 C5.0 모델』의 내용을 참조하십시오.

다음을 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 훈련하는 데 사용되는 파티션만큼 많은 레코드를 최소한으로 포함해야 합니다.

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 데이터를 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 훈련 데이터의 특정 서브세트를 설명하고 훈련 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 특성이 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 훈련하려면 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 훈련 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.

C5.0 노드 모델 옵션

모델 이름. 생성할 모델의 이름을 지정합니다.

- **자동.** 이 옵션이 선택되면 목표 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 30 페이지의 『분할 모델 작성』의 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 = RED*인 레코드가 *COLOR = BLUE*인 레코드와 매우 유사한 경우에는 두 개의 분할 즉, 한 그룹에 *GREEN*을 작성하고 다른 그룹에는 *BLUE* 및 *RED*를 함께 작성합니다.

부스팅 사용. C5.0 알고리즘에는 정확도 비율을 개선하기 위한 **부스팅**이라는 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 훈련해야 합니다. **시행 수** 옵션으로 부스팅 모델에 사용되는 모델 수를 제어할 수 있습니다. 이 기능은 불량 데이터를 더 잘 핸들하도록 몇 가지를 독점적으로 개선한 Freund & Schapire의 연구에 기반을 두고 있습니다.

교차 검증. 이 옵션이 선택되면 C5.0은 훈련 데이터의 서브세트에 작성된 모델 세트를 사용하여 전체 데이터 세트에 작성된 모델의 정확도를 추정합니다. 이 옵션은 데이터 세트가 너무 작아서 일반 훈련 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 훈련의 경우 대부분의 C5.0 모수가 자동으로 설정됩니다. **고급** 훈련에서는 훈련 모수를 보다 직접적으로 제어할 수 있습니다.

단순 모드 옵션

선호. 기본적으로 C5.0은 가능한 가장 정확한 트리를 생성하려 시도합니다. 일부 인스턴스에서 이는 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(%). 훈련 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 훈련 세트에서 최소 이 수만큼의 레코드가 포함된 경우에만 분할됩니다. 기본값은 2입니다. 불량 데이터의 **초과 훈련**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.

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 유의성을 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 예측변수에 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.

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, 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 수동으로 대상, 예측변수 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 선택하십시오.

예측변수 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 35 페이지의 『빈도 및 가중 필드 사용』의 내용을 참조하십시오.

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.

Tree-AS 노드 - 기본

의사결정 트리 작성 방법에 대한 기본 옵션을 지정하십시오.

트리 성장 알고리즘 사용하려는 **CHAID** 알고리즘 유형을 선택합니다. **Exhaustive CHAID**는 각 예측변수에 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.

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 설정**으로 전환하는 레코드 수를 지정합니다. 기본값은 1,000,000입니다.

분할 유의수준 노드 분할에 대한 유의 수준(알파)을 지정합니다. 값은 0.01과 0.99 사이에 있어야 합니다. 값이 낮을수록 노드 수가 적은 트리를 생성합니다.

병합 유의수준 범주 병합에 대한 유의 수준(알파)을 지정합니다. 값은 0.01과 0.99 사이에 있어야 합니다. 이 옵션은 Exhaustive CHAID에 사용할 수 없습니다.

Bonferroni 방법을 사용하여 유의성 값 조정 예측변수의 다양한 범주 조합을 검정할 때 유의성 값을 조정합니다. 값은 범주 수와 예측변수의 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.

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 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.

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. 엡

실론은 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.

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 내용을 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.

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.

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 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

신뢰도 계산 모델 스코어링 중에 신뢰도 필드를 추가하려면 이 확인 상자를 선택합니다.

규칙 식별자 레코드가 지정된 리프 노드의 ID를 포함하는 모델의 스코어링 중에 필드를 추가하려면 이 확인 상자를 선택합니다.

Tree-AS 모델 너깃

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 예측변수로 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

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.

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 **평균**값을 표시합니다.

다음 **테이블 내용** 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 **페이지**의 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 범주**에 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.

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 확인 상자를 선택합니다. 데이터베이스에서 모델 스코어를 계산할 때 신뢰도를 제외하면 보다 효율적인 SQL을 생성할 수 있음을 의미합니다. 회귀 분석 트리에서는 신뢰도를 지정하지 않습니다.

규칙 식별자 각 레코드가 지정된 터미널 노드의 ID를 표시하는 스코어링 출력에서 필드를 추가하려면 이 확인 상자를 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.

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 트리는 C&R 트리와 비교했을 때 두 개의 기능이 추가되었습니다.

- 첫 번째 기능은 원래 데이터 세트에서 복원 표본추출하여 훈련 데이터 세트의 복제본을 작성하는 배깅입니다. 이 동작을 수행하면 원래 데이터 세트와 동일한 크기의 붓스트랩 표본이 작성된 다음 구성요소 모델이 각 복제본에 작성됩니다. 이 구성요소 모델들은 함께 앙상블 모델을 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 없음으로 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 배깅 및 필드 표본추출 사용으로 인해 과적합이 발생할 가능성이 훨씬 줄어들어 새 데이터를 사용할 때 검정에 표시되는 결과가 반복될 가능성이 더 높아집니다.

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, 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 수동으로 대상, 예측변수 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 선택하십시오.

예측변수 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 35 페이지의 『빈도 및 가중 필드 사용』의 내용을 참조하십시오.

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.

랜덤 트리 노드 - 기본

의사결정 트리 작성 방법에 대한 기본 옵션을 지정하십시오.

작성할 모델 수 노드가 작성할 수 있는 최대 모델 수를 지정하십시오.

표본 크기 기본적으로 붓스트랩 표본의 크기는 원본 학습 데이터와 동일합니다. 대형 데이터 세트를 처리하는 경우 표본 크기를 줄이면 성능이 높아질 수 있습니다.

불균형한 데이터 처리 모델의 대상이 플래그 결과이고(예: 구매 또는 구매 안 함) 원하는 결과 대 원하지 않는 결과의 비율이 매우 작으면 데이터의 균형이 맞지 않고 모델이 수행하는 붓스트랩 표본추출이 모델 정확도에 영향을 줄 수 있습니다. 정확도를 향상시키려면 이 확인 상자를 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 확인 상자를 선택하십시오.

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.

참고: 분할에 대한 예측자 수는 데이터의 총 예측자 수보다 클 수 없습니다.

더 이상 정확도를 개선할 수 없는 경우에 작성 중단 모델 작성 시간을 줄이려면 이 옵션을 선택하여 결과의 정확도를 개선할 수 없는 경우에 모델 작성을 중지하십시오.

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 내용을 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.

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.
- **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.

랜덤 트리 노드 - 고급

의사결정 트리 작성 방법에 대한 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 작은 경우 이 필드가 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).

랜덤 트리 노드 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 모델 스코어링 중에 예측변수의 중요도를 계산하도록 선택할 수도 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

랜덤 트리 모델 너깃

랜덤 트리 모델 너깃 출력

랜덤 트리 모델을 작성하면 출력 뷰어에 다음 정보가 제공됩니다.

모델 정보 테이블

모델 정보 테이블에서는 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 예측변수 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)
  - 모델 정확도
  - 오분류 규칙
- 분류 모델(**불균형 데이터 처리** 선택)
  - Gmean

  - 참 긍정 비율(클래스로 세분화됨)
- 회귀 모델
  - 제곱근 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

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 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 내용** 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.

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 확인 상자를 선택합니다. 데이터베이스에서 모델 스코어를 계산할 때 신뢰도를 제외하면 보다 효율적인 SQL을 생성할 수 있음을 의미합니다. 회귀 분석 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 페치하여 SPSS Modeler에서 스코어를 계산합니다.

- **데이터베이스 외부 스코어** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

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 93 페이지의 『대화형 트리 작성기』 주제를 참조하십시오.

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 예인 경우 해당 노드에 지정된 모든 레코드의 예측은 예입니다.
- 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 104 페이지의 『필터 및 선택 노드 생성』의 내용을 참조하십시오.
- 트리의 터미널 분기를 정의하는 규칙 세트로 트리 구조를 표시하는 규칙 세트 너깃을 생성합니다. 자세한 정보는 105 페이지의 『의사결정 트리에서 규칙 세트 생성』의 내용을 참조하십시오.
- 또한 트리 모델 너깃의 경우에만 모델을 PMML 형식으로 내보낼 수 있습니다. 자세한 정보는 44 페이지의 『모델 팔레트』 주제를 참조하십시오. 모델에 사용자 정의 분할이 포함된 경우 이 정보

는 내보낸 PMML에서 보존되지 않습니다. (분할은 보존되지만 알고리즘을 통해 선택된 것이 아니라 사용자 정의되었다는 사실은 그렇지 않습니다.)

- 현재 트리의 선택된 부분을 기반으로 그래프를 생성합니다. 참고: 스트림의 다른 노드에 연결되어 있을 때에는 너깃에 대해서만 작동합니다. 자세한 정보는 138 페이지의 『그래프 생성』의 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 138 페이지의 『부스팅 C5.0 모델』의 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.

단일 트리 모델 너깃

모델링 노드에서 주요 목표로 **단일 트리 작성**을 선택한 경우 결과로 생성되는 모델 너깃은 다음 탭을 포함합니다.

표 7. 단일 트리 너깃의 탭

탭	설명	추가 정보
모델	모델을 정의하는 규칙을 표시합니다.	자세한 정보는 『의사결정 트리 모형 규칙』의 내용을 참조하십시오.
뷰어	모델의 트리 보기를 표시합니다.	자세한 정보는 136 페이지의 『의사결정 트리 모형 뷰어』의 내용을 참조하십시오.
요약	필드, 작성 설정, 모델 추정 프로세스에 대한 정보를 표시합니다.	자세한 정보는 47 페이지의 『모델 너깃 요약/정보』의 내용을 참조하십시오.
설정	모델 스코어링 중에 신뢰도 및 SQL 생성에 대한 옵션을 지정할 수 있습니다.	자세한 정보는 137 페이지의 『의사결정 트리/규칙 세트 모델 너깃 설정』의 내용을 참조하십시오.
주석	설명 주석을 추가하고 사용자 정의 이름을 지정하고 도구 팁 텍스트를 추가하고 모델에 대한 검색 키워드를 지정할 수 있습니다.	

의사결정 트리 모형 규칙

의사결정 트리 너깃의 모델 탭은 모델을 정의하는 규칙을 표시합니다. 선택적으로 예측변수 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** 옵션을 선택하면 모델 탭은 트리 규칙 세부사항만 표시합니다.

트리 규칙

왼쪽 분할창에는 알고리즘을 통해 발견한 데이터의 파티셔닝을 정의하는 조건 목록이 표시됩니다. 이는 본질적으로 여러 다른 예측변수의 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 분기라 부릅니다. 초기 분기(때로 루트라 함)는 모든 데이터 레코드를 포함합니다. 루트는 특정 입력 필드의 값에 따라 서브세트 또는 하위 분기로 분할됩니다. 각 하위 분기는 계속해서 다시 다음 하위 분기로 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 45 페이지의 『모델 너깃 찾아보기』의 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 예측변수의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다.

참고: 이 차트는 모델을 생성하기 전에 분석 탭에 **예측변수 중요도 계산**이 선택된 경우에만 사용 가능합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 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 소수점이하 자리수로 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 예측변수 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.

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 96 페이지의 『트리 보기 사용자 정의』의 내용을 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 꺽쇠 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
```

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. 회귀분석 트리에서는 신뢰도를 지정하지 않습니다.

참고: CHAID 모델의 작성 옵션 탭 - 모델 패널에서 **매우 큰 데이터 세트에 대한 모델 작성** 옵션을 선택한 경우 이 확인 상자는 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 확인 상자는 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 138 페이지의 『부스팅 C5.0 모델』의 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 사용 가능합니다.
- **데이터베이스 외부 스코어** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

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.

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. 그래프 보드 기본 탭이 표시됩니다.

   참고: 기본 및 세부사항 탭은 그래프 보드를 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.

부스팅, 배깅 및 매우 큰 데이터 세트의 모델 너깃

모델 정확도 개선(boosting), **모델 안정성 개선(bagging)** 또는 **매우 큰 데이터 세트용 모델 작성**를 모델링 노드의 기본 목표로 선택하는 경우 IBM SPSS Modeler에서는 다중 모델의 앙상블을 작성합니다. 자세한 정보는 49 페이지의 『앙상블 모델』의 내용을 참조하십시오.

결과로 생성된 모델 너깃은 다음 탭을 포함합니다. 모델 탭에서는 다양한 모델 보기를 제공합니다.

표 *8.* 모델 너깃에서 사용 가능한 탭

탭	보기	설명	추가 정보
모델	모델 요약	앙상블 품질 및 (부스팅 모델 및 연속형 목표 제외) 다양성, 서로 다른 모델에서 예측 다양성의 측도에 대한 요약을 표시합니다.	자세한 정보는 49 페이지의 『모델 요약』의 내용을 참조하십시오.
	예측변수 중요도	모델 추정 시 각 예측변수의 상대적 중요도(입력 필드)를 나타내는 차트를 표시합니다.	자세한 정보는 50 페이지의 『예측변수 중요도』의 내용을 참조하십시오.
	예측변수 빈도	모델 세트에 사용된 각 예측변수의 상대적 빈도를 나타내는 차트를 표시합니다.	자세한 정보는 50 페이지의 『예측변수 빈도』의 내용을 참조하십시오.
	구성요소 모델 정확도	앙상블에서 각 서로 다른 모델의 예측 정밀도의 차트를 구성합니다.	
	구성요소 모델 세부사항	앙상블에서 각 서로 다른 각 모델에 대한 정보를 표시합니다.	자세한 정보는 51 페이지의 『구성요소 모델 세부사항』의 내용을 참조하십시오.

표 *8.* 모델 너깃에서 사용 가능한 탭 *(*계속*)*

탭	보기	설명	추가 정보
	정보	필드, 작성 설정, 모델 추정 프로세스에 대한 정보를 표시합니다.	자세한 정보는 47 페이지의 『모델 너깃 요약/정보』의 내용을 참조하십시오.
설정		스코어링 작업에 신뢰도를 포함할 수 있습니다.	자세한 정보는 137 페이지의 『의사결정 트리/규칙 세트 모델 너깃 설정』의 내용을 참조하십시오.
주석		설명 주석을 추가하고 사용자 정의 이름을 지정하고 도구 팁 텍스트를 추가하고 모델에 대한 검색 키워드를 지정할 수 있습니다.	

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 적중.** 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 적중 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. **증폭된** C5.0 모델에만 사용 가능합니다. 자세한 정보는 138 페이지의 『부스팅 C5.0 모델』의 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 모델의 탭과 동일합니다.

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.

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 훈련 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.

주석

- 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.

제 7 장 베이지안 신경망 모형

베이지안 네트워크 노드

베이지안 네트워크 노드로 관측 및 기록한 증거를 "상식적인" 실세계 지식과 결합해서 겉보기에 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 **노드**라 부름)와 이 변수 사이의 확률적 또는 조건부 독립성을 표시하는 그래픽 모델입니다. 노드 간의 인과 관계를 베이지안 네트워크를 통해 표시할 수 있지만 네트워크의 링크(**아크**로도 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 해당하는 링크의 방향)하는 데 사용할 수 있는 네트워크의 단순화된 버전입니다. 각 노드는 환자의 조건에 관련시킬 수 있는 패싯을 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.

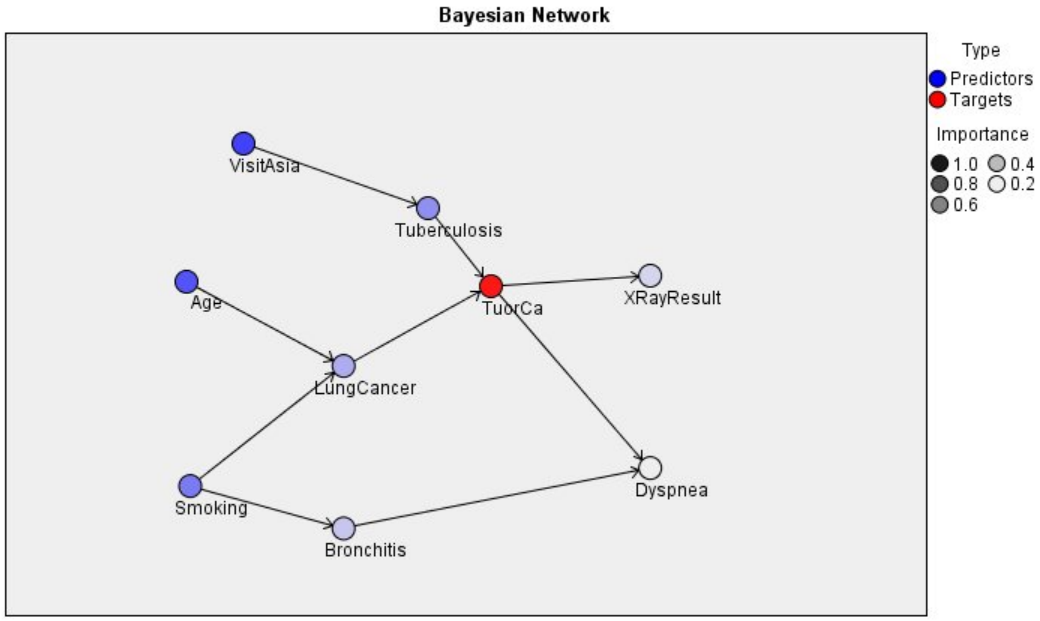


그림 *29. Lauritzen* 및 *Spegelhalter*의 아시아 네트워크 예

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. 목표 필드는 범주형이어야 하며 측정 수준은 명목, 순서 또는 플래그가 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 모형을 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 모형을 사용하여 이탈할 것 같은 고객의 특성을 식별하고 매월 새 데이터로 모델 훈련을 계속할 수 있습니다.

베이지안 네트워크 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 내용을 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 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 분할로 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 30 페이지의 『분할 모델 작성』의 내용을 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 예측변수 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 모형을 작성합니다. 이는 각 예측변수가 목표변수 외에 또 다른 예측변수에 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 포함

된 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.

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 삭제라고도 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 예측변수 독립성을 검정합니다.
- **Pearson 카이제곱.** 지정된 빈도 분포 다음에 관측 이벤트 발생의 상대 빈도가 나오는 귀무가설을 사용하여 목표 예측변수 독립성을 검정합니다.

베이지안 신경망 모형은 검정 대응에 추가 변수가 사용되는 조건부 독립성 검정을 수행합니다. 또한 모델은 목표와 예측변수 간 관계 외에 예측변수 자체 사이의 관계도 탐색합니다.

참고: 독립성 검정 옵션은 모델 탭에서 Markov Blanket의 **필드선택 전처리 단계 포함** 또는 **구조 유형**을 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 유형**을 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 유형**을 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 포함**을 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 모형을 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 수**의 값을 초과할 경우 오류 메시지가 표시됩니다.

베이지안 신경망 모형 너깃

참고: 모델링 노드 모델 탭에서 **기존 모수 훈련 계속**을 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 예측변수 간 관계 및 예측변수 사이의 관계를 표시하는 노드의 네트워크 그래프가 있습니다. 각 예측변수의 중요도가 색상 농도에 따라(더 진한 색이 중요한 예측변수를 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

예측변수 중요도 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타내는 차트를 표시합니다. 추가 정보는 47 페이지의 『예측변수 중요도』의 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.

베이지안 신경망 모형 설정

베이지안 신경망 모형 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 확인 상자의 기본 설정은 모델링 노드의 고급 탭에 있는 해당 확인 상자를 통해 판별됩니다. 자세한 정보는 146 페이지의 『베이지안 네트워크 노드 고급 옵션』의 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** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

베이지안 신경망 모형 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.

제 8 장 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.

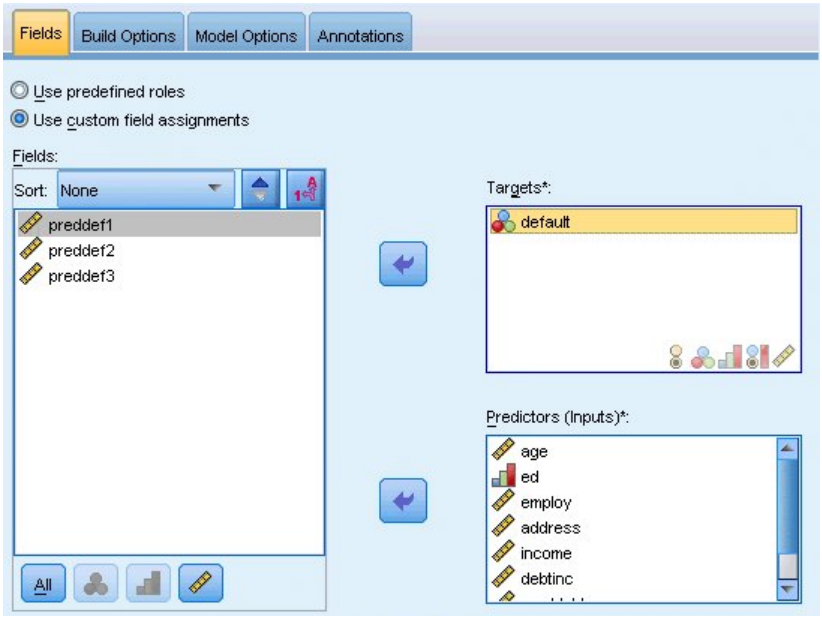


그림 *30.* 필드 탭

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 33 페이지의 『모델링 노드 필드 옵션』의 내용을 참조하십시오.

신경망 모델

신경망은 신경계가 작동하는 방식의 단순 모델입니다. 기본 단위는 **뉴런**이며, 일반적으로 다음 그림에 표시된 것처럼 **레이어**로 조직됩니다.

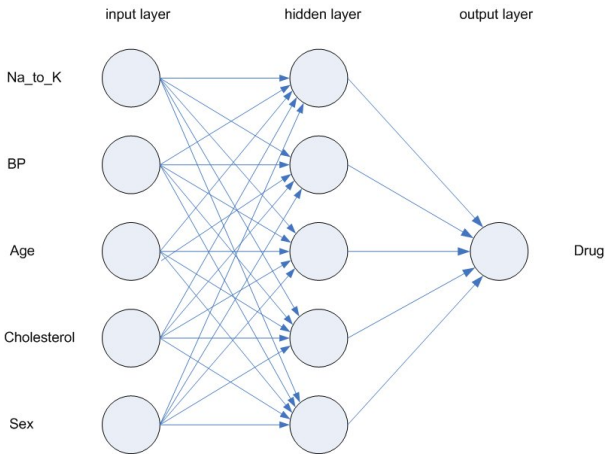


그림 *31.* 신경망의 구조

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**, 목표 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 **훈련**을 통해 훈련합니다. 출력이 알려지는 예제는 반복해서 네트워크에 제시되고, 네트워크가 제공하는 응답은 알려진 결과와 비교됩니다. 이 비교의 정보는 점차로 가중치를 변경하면서, 네트워크를 통해 뒤로 전달됩니다. 훈련이 진행되면서, 네트워크는 알려진 결과를 복제할 때 점차적으로 정확하게 됩니다. 훈련되면, 네트워크는 결과를 알 수 없는 나중 케이스에도 적용 가능하게 됩니다.

레거시 스트림이 있는 신경망 사용

IBM SPSS Modeler 버전 14에서는 매우 큰 데이터 세트에 대한 최적화 및 boosting 및 bagging 기술을 지원하는 새 신경망 노드를 도입했습니다. 이전 노드를 포함하는 기존 스트림은 이 릴리스에서 계속 모델을 작성하고 스코어링할 것입니다. 그러나 이 지원은 나중 릴리스에서 제거될 예정이므로, 지금부터는 새 버전을 사용할 것을 권장합니다.

버전 13부터, 알 수 없는 값(즉, 훈련 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 널이 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 > 옵션**에서 세트 크기 제한 옵션을 사용할 수 있습니다. 이 옵션은 버전 14 이상의 코호넨 넷 및 *K*-Means 노드에만 적용됩니다.

목적

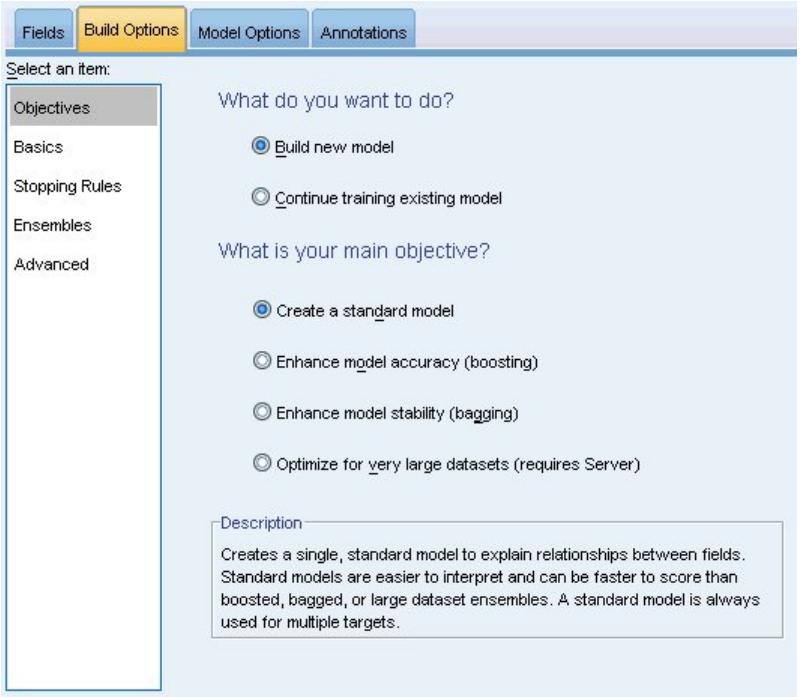


그림 *32.* 목적 설정

원하는 작업

- **새 모델 작성.** 완전히 새 모델을 작성합니다. 노드의 일반적인 작업입니다.
- **기존 모델 계속 훈련.** 노드에 의해 성공적으로 작성된 마지막 모델로 계속 훈련합니다. 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 예측변수를 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 훈련**과 함께 이 옵션을 사용하려면 Analytic Server에 연결되어야 합니다.

- **모델 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.

기본

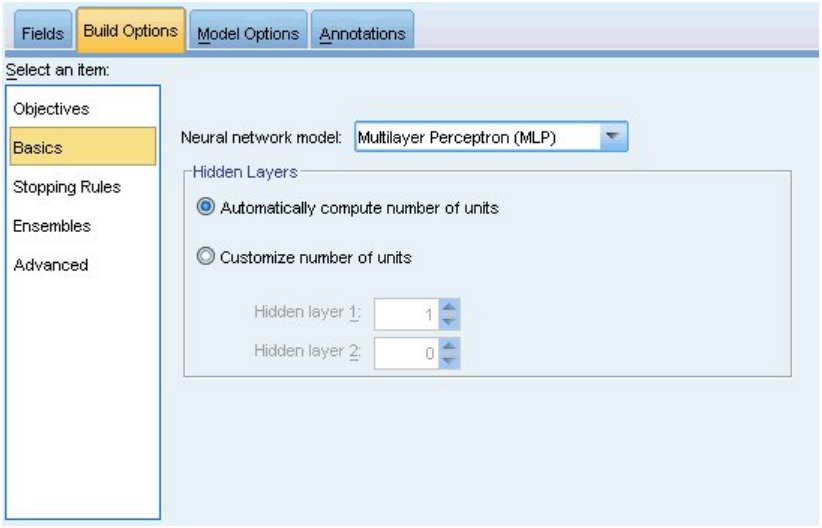


그림 *33.* 기본 설정

신경망 모델. 모델 유형은 네트워크가 은닉층을 통해 목표에 예측변수를 연결하는 방법을 판별합니다. **다중 레이어 퍼셉트론(MLP)**은 훈련 및 스코어링 시간이 더 소요될 수 있는 한층 복잡한 관계에 대해 허용됩니다. **방사형 기저함수(RBF)**에서는 MLP와 비교하여 예측력이 감소될 수 있고 훈련 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.

중지 규칙

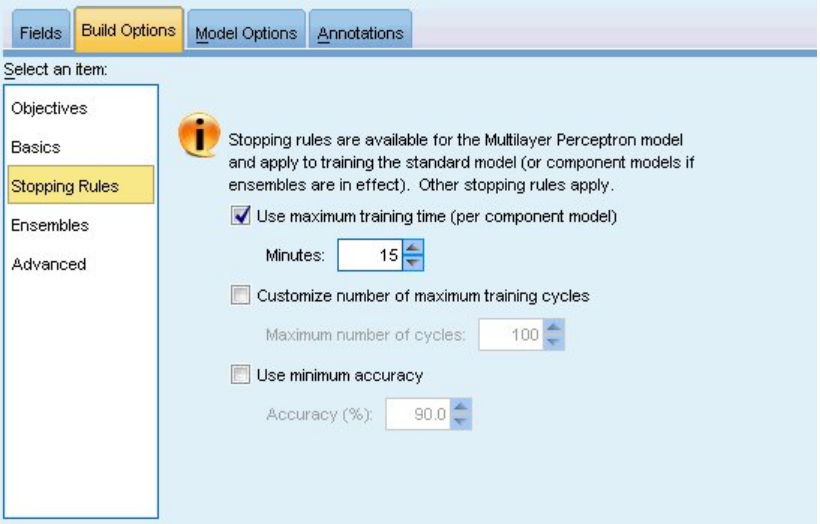


그림 *34.* 중지 규칙 설정

훈련 다중 레이어 퍼셉트론 네트워크를 중지할 시기를 판별하는 규칙입니다. 이 설정은 방사형 기저함수 알고리즘이 사용될 때 무시됩니다. 훈련은 최소 하나의 순환(데이터 전달)에서 진행되므로, 다음 기준에 따라 중지될 수 있습니다.

최대 훈련 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 훈련 시간입니다. 훈련은 현재 순환을 완료하기 위해 지정된 시간 한계를 약간 넘어설 수 있습니다.

최대 훈련 주기 수 사용자 정의. 허용되는 최대 훈련 주기 수. 최대 주기 수를 초과하면, 훈련이 중지됩니다. 0보다 큰 정수를 지정하십시오.

최소 정확도 사용. 이 옵션을 사용하면, 지정된 정확도가 될 때까지 훈련이 계속됩니다. 이러한 상황은 발생하지 않을 수 있지만, 언제든지 훈련을 중단하고 지금까지 달성한 최상의 정확도로 넷을 저장할 수 있습니다.

과적합 방지 세트의 오류가 각각의 주기 후에 감소하지 않는 경우, 훈련 오류의 상대적 변화가 작은 경우, 또는 현재 훈련 오류의 비율이 초기 오류와 비교하여 작은 경우 훈련 알고리즘도 중지됩니다.

앙상블

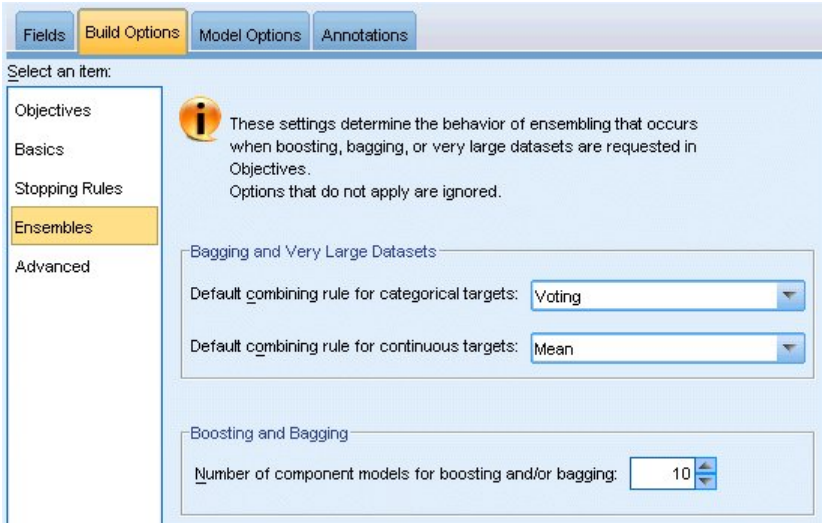


그림 *35.* 앙상블 설정

이 설정은 부스팅, 배깅 또는 아주 큰 데이터 세트가 목적에서 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.

모델 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. 모델 정확도 또는 안정성을 향상시키는 것이 목표일 때 작성할 기본 모델 수를 지정하십시오. 배깅의 경우, 붓스트랩 표본의 수입니다. 양의 정수여야 합니다.

고급

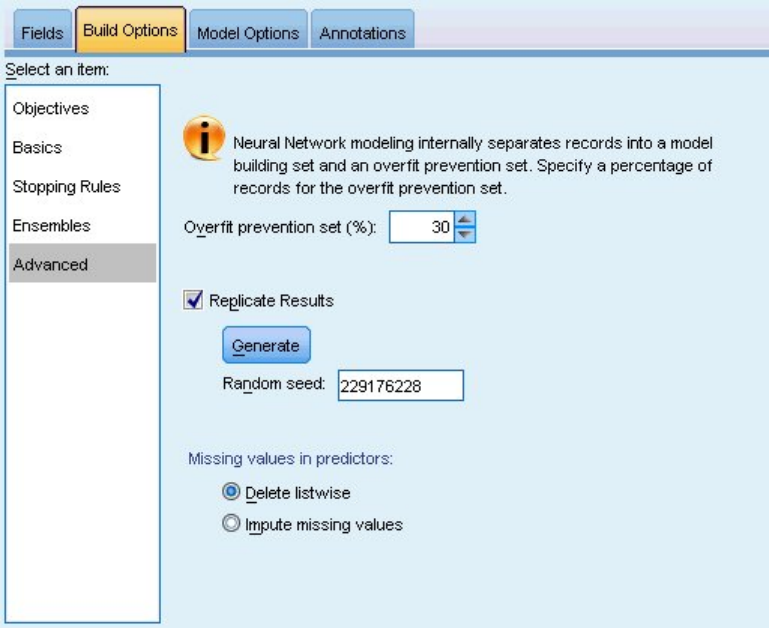


그림 *36.* 고급 설정

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 훈련 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 유사 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.

모델 옵션

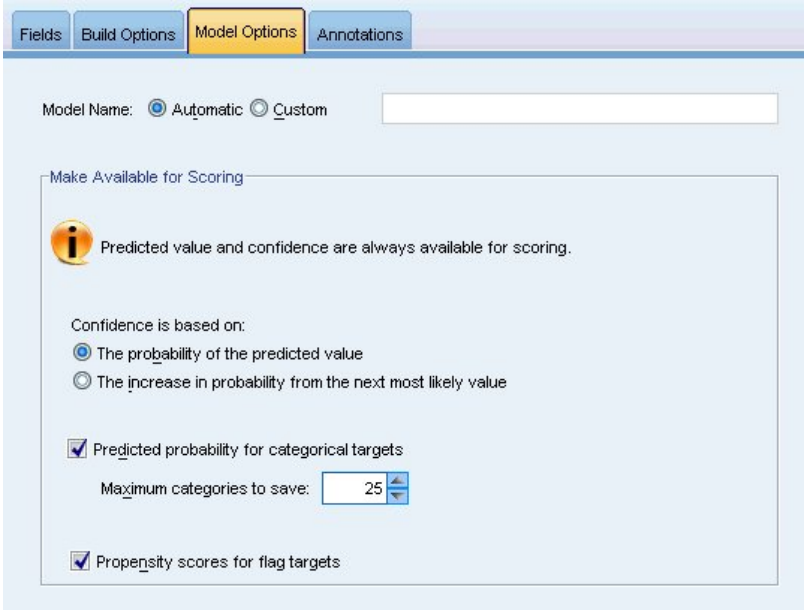


그림 *37.* 모델 옵션 탭

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.

모델 요약

Model Summary

Target	Previously defaulted
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

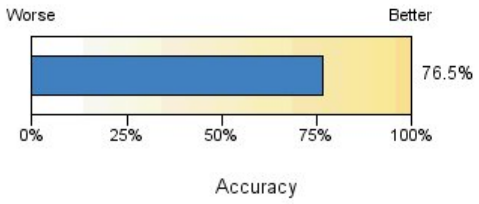


그림 *38.* 신경망 모델 요약 보기

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.

예측변수 중요도

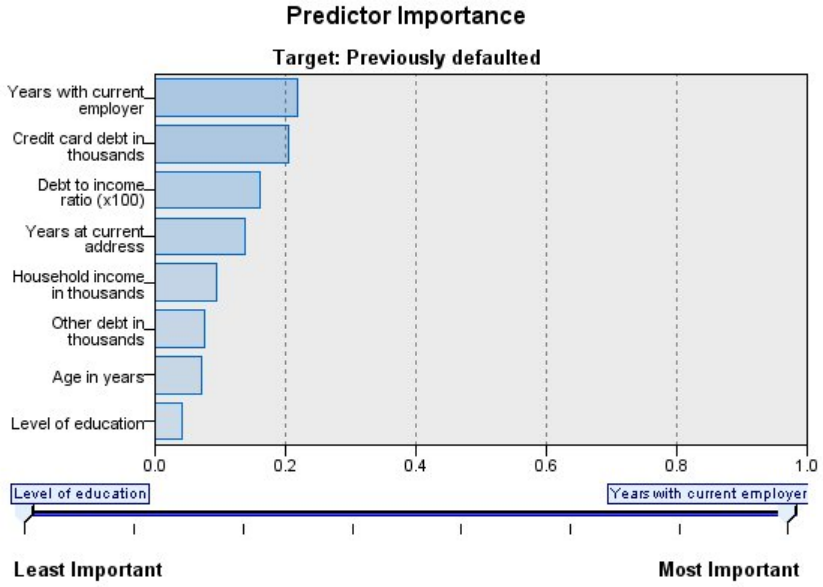


그림 *39.* 예측변수 중요도 보기

일반적으로, 가장 중요한 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

여러 목표. 여러 목표가 있는 경우, 각 목표가 별도의 도표에 표시되고 표시할 목표를 제어하는 **목표** 드롭다운 목록이 있습니다.

관측값 별 예측값

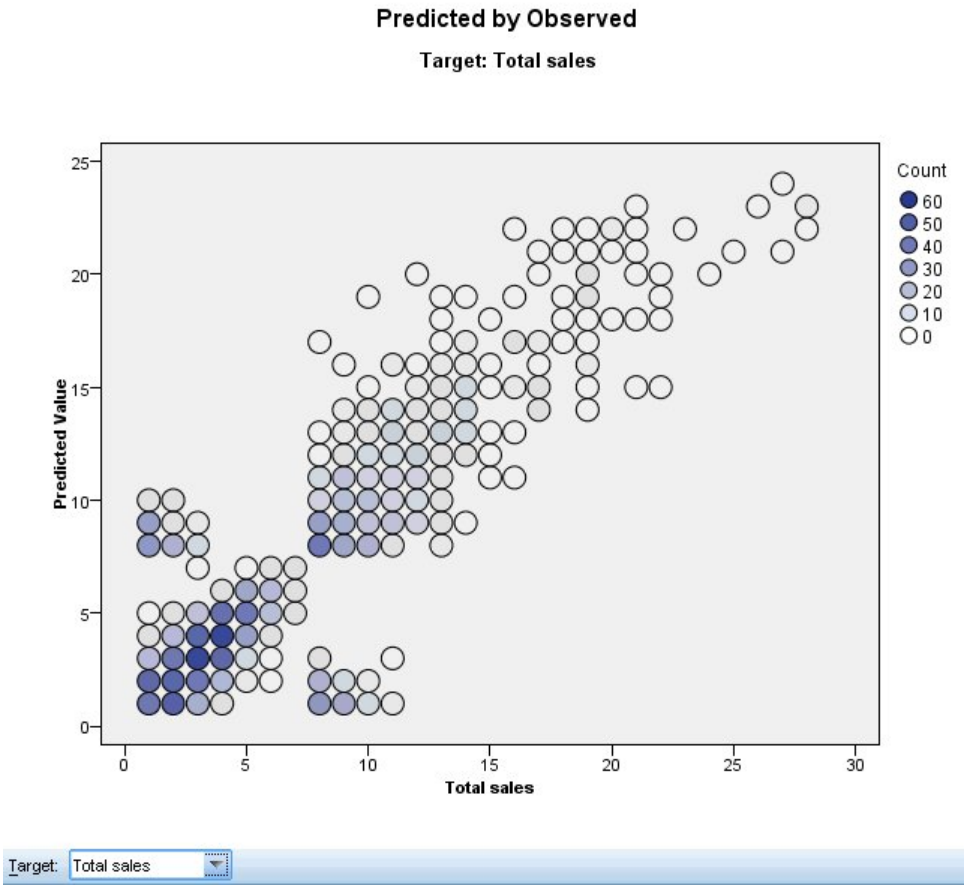


그림 *40.* 관측값 별 예측값 보기

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.

분류

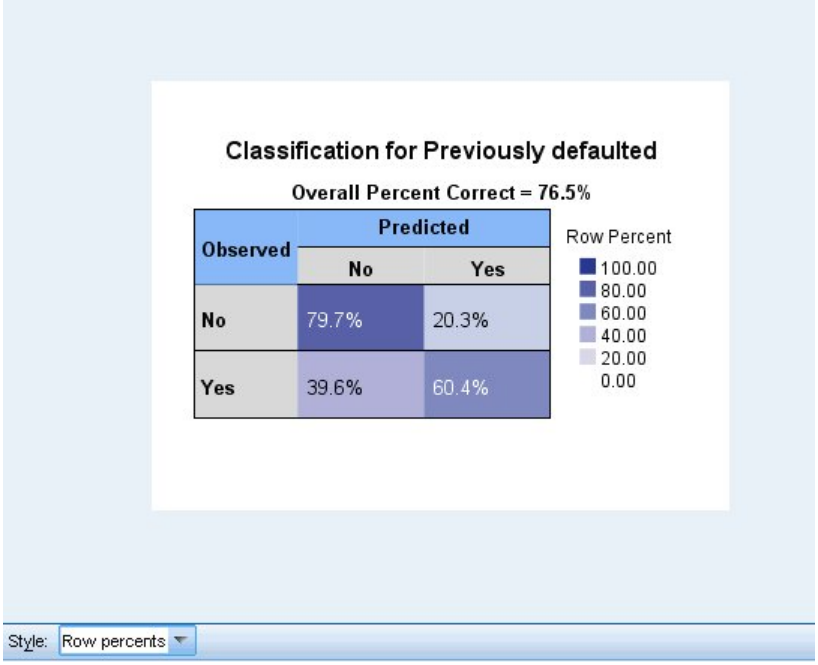


그림 *41.* 분류 보기, 행 퍼센트 유형

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.

네트워크

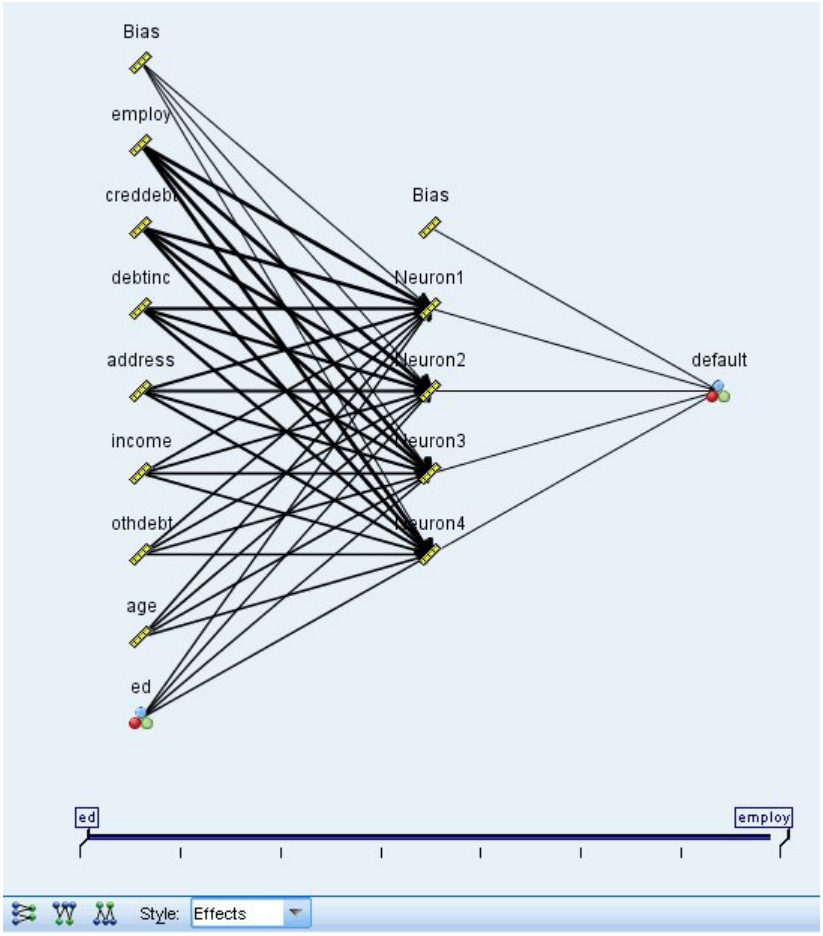


그림 *42.* 네트워크 보기, 왼쪽의 입력, 효과 유형

신경망의 그래픽 표현을 표시합니다.

차트 유형. 두 가지의 다른 표시 유형이 있으며, **유형** 드롭다운 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 큽니다. 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.

설정

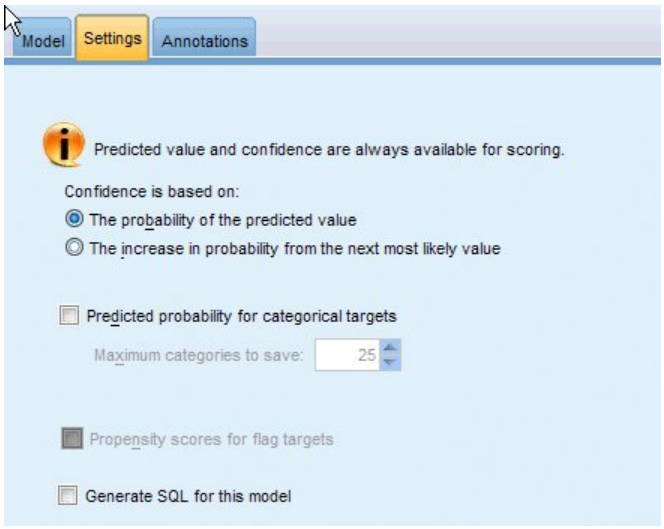


그림 *43.* 설정 탭

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 다시 데이터를 페치하고 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.

제 9 장 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 오퍼 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 오퍼에서 제외시키고 다른 고객 그룹을 포함하도록 선택한 후 이 선택이 전반적인 적중률에 영향을 미치는 방식을 즉시 확인할 수 있습니다. 의사결정 목록 모델은 포함한 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, 측도, 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 측도를 정의할 수 있습니다. 예를 들어, 모델을 특정 지역에 맞게 조정하도록 마이닝 작업에 데이터 선택을 지정한 후 모델이 전체 국가에서 수행하는 정도를 평가하는 사용자 정의 측도를 작성할 수 있습니다. 마이닝 작업과 달리, 측도는 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 측도를 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 목표 필드 및 최소 하나의 입력 필드. 목표 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 순서의 입력으로 구간화된 필드를 사용하십시오.

의사결정 목록 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 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 열어서 여러 대안 중에서 선택하고 모델이 점진적으로 성장 또는 수정되도록 알고리즘을 여러 다른 설정으로 반복해서 적용할 수 있게 합니다. 자세한 정보는 172 페이지의 『의사결정 목록 뷰어』의 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 목표 필드의 값을 지정합니다. 예를 들어, 목표 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.

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.

의사결정 목록 모델 너깃

모델은 **세그먼트** 목록으로 구성되며 각 세그먼트는 일치하는 레코드를 선택하는 **규칙**을 통해 정의됩니다. 모델을 생성하기 전에 세그먼트를 쉽게 보거나 수정하고 포함 또는 제외시킬 세그먼트를 선택할 수 있습니다. 스코어링에 사용할 경우 의사결정 목록 모델은 포함된 세그먼트에 예를 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 예로 스코어링됩니다. 필요에 따라 파생 또는 채움 노드를 사용하여 널이 아니오로 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.

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 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** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

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.

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 쉼표를 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 183 페이지의 『모델 측도 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

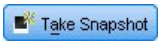
작업 모델 분할창은 도구 모음을 통해 다음 기능을 제공합니다.

참고: 일부 기능은 모델 세그먼트를 마우스 오른쪽 단추로 클릭해서도 사용 가능합니다.

표 9. 작업 모델 도구 모음 단추.

도구 모음 단추	설명
	새 모델 너깃을 작성하기 위한 옵션을 제공하는 새 모델 생성 대화 상자를 실행합니다.

표 9. 작업 모델 도구 모음 단추 (계속).

	대화형 세션의 현재 상태를 저장합니다. 마이닝 작업, 모델 스냅샷, 데이터 선택, 사용자 정의 측도를 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 측도 구성 대화 상자를 표시합니다. 자세한 정보는 183 페이지의 『모델 측도 구성』의 내용을 참조하십시오.
	데이터 선택 구성 대화 상자를 표시합니다. 자세한 정보는 178 페이지의 『데이터 선택 구성』의 내용을 참조하십시오.
	스냅샷 탭을 표시합니다. 자세한 정보는 174 페이지의 『스냅샷 탭』의 내용을 참조하십시오.
	대안 탭을 표시합니다. 자세한 정보는 『대안 탭』의 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.

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 수입니다.

범위. 대체 모델의 범위입니다.

빈도. 전체에 관련한 적중 수입니다.

확률. 대체 모델의 확률 퍼센트를 표시합니다.

참고: 대체 결과는 모델과 함께 저장되지 않으며 결과는 활성 세션 중에만 유효합니다.

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 수입니다.

   범위. 모델의 범위입니다.

   빈도. 전체에 관련한 적중 수입니다.

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.

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.

마이닝 작업

마이닝 작업은 새 규칙이 생성되는 방식을 판별하는 매개변수 콜렉션입니다. 새로운 상황에 모델을 탄력적으로 적응할 수 있도록 일부 매개변수가 선택 가능합니다. 작업은 작업 템플리트(유형), 목표, 선택 작성(데이터 세트 마이닝)으로 이루어집니다.

다양한 마이닝 작업 조작은 다음 섹션에서 자세히 설명합니다.

- 『마이닝 작업 실행』
- 176 페이지의 『마이닝 작업 작성 및 편집』
- 178 페이지의 『데이터 선택 구성』

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.

마이닝 작업 작성 및 편집: 마이닝 작업은 데이터 모델을 구성하는 규칙 콜렉션을 검색하는 메커니즘입니다. 선택된 템플리트에 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.
3. 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.
4. **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 168 페이지의 『의사결정 목록 모델 옵션』의 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.

목표

목표 필드: 예측하려는 필드를 나타내며, 값이 다른 필드(예측변수)의 값에 관련된 것으로 추정됩니다.

목표 값. 모델링하려는 결과를 나타내는 목표 필드의 값을 지정합니다. 예를 들어, 목표 필드 이탈이 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 편집』의 내용을 참조하십시오.

데이터

선택 작성 의사결정 목록 뷰어가 새 규칙을 찾기 위해 분석해야 하는 평가 측도를 지정하는 옵션을 제공합니다. 나열된 평가 측도는 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 178 페이지의 『사용 가능 필드 사용자 정의』의 내용을 참조하십시오.

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.

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.

세그먼트 규칙

작업 템플리트를 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.

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 `TOWN`이 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.

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 **편집**을 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 `TOWN`이 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 **확인**을 한번 더 클릭하십시오.

업데이트된 규칙 조건과 함께 선택한 세그먼트가 표시됩니다.

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.

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 내용은 단독으로 의사결정 목록 뷰어에 사용하도록 예약되지 않지만 다른 애플리케이션에 붙여넣을 수도 있습니다. 예를 들어, 클립보드 내용을 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 측도를 다시 계산합니다.

참고: 이 절차의 두 모델은 모두 동일한 기본 모델 템플리트를 기준으로 하고 동일한 목표를 포함해야 합니다. 그렇지 않으면 오류 메시지가 표시됩니다.

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.

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.

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.

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.

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

널값으로 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 토글** 단추를 클릭하십시오. 선택한 세그먼트의 선택한 목표 열에 이제 **제외됨**이 표시됩니다.

참고: 삭제된 세그먼트와 달리, 제외된 세그먼트는 최종 모델에 재사용할 수 있습니다. 제외된 세그먼트는 차트 결과에 영향을 미칩니다.

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.

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 다**를 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 53 페이지의 『모델링 노드 재생성』의 내용을 참조하십시오.

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.

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 184 페이지의 『Excel로 평가』의 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.

또는

F5를 누르십시오.

모든 측도가 다시 계산되고 작업 모델의 새 값이 표시됩니다.

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 측도 구성**을 선택하십시오.
2. **Excel로 사용자 정의 측도 계산** 옵션에 **예**를 선택하십시오. **워크북** 필드가 활성화되어 사전 구성된 Excel 워크북 템플리트를 선택할 수 있습니다.
3. **Excel에 연결** 단추를 클릭하십시오. 열기 대화 상자가 열려서 로컬 또는 네트워크 파일 시스템에서 사전 구성된 템플리트 위치를 탐색할 수 있습니다.
4. 적합한 Excel 템플리트를 선택하고 **열기**를 클릭하십시오. 선택한 Excel 템플리트가 실행됩니다. Windows 작업 표시줄을 사용하여(또는 Alt-Tab을 눌러서) 사용자 정의 측도의 입력 선택 대화 상자로 다시 이동하십시오.
5. Excel 템플리트에 정의된 매트릭 이름과 모델 메트릭 이름 간의 적합한 맵핑을 선택하고 **확인**을 클릭하십시오.

일단 링크가 설정되면 스프레드시트에 모델 규칙을 표시하는 사전 구성된 Excel 템플리트로 Excel이 시작됩니다. Excel로 계산된 결과는 의사결정 목록 뷰어에 새 열로 표시됩니다.

참고: Excel 메트릭은 모델이 저장될 때 남지 않습니다. 메트릭은 활성 세션 중에만 유효합니다. 하지만 Excel 메트릭을 포함한 스냅샷을 작성할 수 있습니다. 스냅샷 보기에 저장된 Excel 메트릭은 히스토리 비교 용도로만 유효하며 다시 열 때 새로 고쳐지지 않습니다. 자세한 정보는 174 페이지의 『스냅샷 탭』의 내용을 참조하십시오. Excel 템플리트에 대한 연결을 다시 설정할 때까지는 Excel 메트릭이 스냅샷에 표시되지 않습니다.

***MS Excel* 통합 설정:** 의사결정 목록 뷰어와 Microsoft Excel의 통합은 사전구성된 Excel 스프레드시트 템플리트 사용을 통해 수행됩니다. 템플리트는 다음 세 개의 워크시트로 구성됩니다.

모델 측도. 가져온 의사결정 목록 뷰어 측도, 사용자 정의 Excel 측도, 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. 템플리트를 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. 템플리트를 열고 구성 워크시트를 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.

   도구 > **보호** > **시트 비보호**
3. 노랑으로 음영 처리되어 있으며 **End** 단어를 포함한 셀 A5를 선택하십시오.
4. 메뉴에서 다음을 선택하십시오.

   삽입 > **행**
5. 새 측도의 **이름** 및 **설명**을 입력하십시오. 예를 들어, **오류** 및 **세그먼트와 연관된 오류**와 같습니다.

6. 셀 C5에 수식 **=COLUMN('Model Measures'!N3)**을 입력하십시오.
7. 셀 D5에 수식 **=ROW('Model Measures'!N3)+1**을 입력하십시오.

   이 수식은 현재 비어 있는 모델 측도 워크시트의 열 N에 새 측도를 표시합니다.
8. 메뉴에서 다음을 선택하십시오.

   도구 > **보호** > **시트 보호**
9. **확인**을 클릭하십시오.
10. 모델 측도 워크시트에서 셀 N3의 새 열 제목이 **오류**인지 확인하십시오.
11. 열 N을 모두 선택하십시오.
12. 메뉴에서 다음을 선택하십시오.

    형식 > **셀**
13. 기본적으로 모든 셀에는 **일반** 번호 범주가 있습니다. 그림이 표시되는 방식을 변경하려면 **퍼센트**를 클릭하십시오. 이렇게 하면 Excel에서 그림을 확인할 수 있고 그래프로 출력하는 것처럼 다른 방식으로 데이터를 이용할 수도 있습니다.
14. **확인**을 클릭하십시오.
15. 스프레드시트를 고유 이름 및 파일 확장자 *.xlt*를 지정해서 Excel 2003 템플리트로 저장하십시오. 새 템플리트를 쉽게 찾을 수 있도록 로컬 또는 네트워크 파일 시스템의 사전 구성된 템플리트 위치에 저장할 것을 권장합니다.

모델로 추가 사용자 정의 측도 내보내기

1. 이전 예에서 오류 열에 추가한 템플리트를 열고 구성 워크시트를 선택하십시오.
2. 메뉴에서 다음을 선택하십시오.

   도구 > **보호** > **시트 비보호**
3. 노랑으로 음영 처리되어 있으며 **End** 단어를 포함한 셀 A14를 선택하십시오.
4. 메뉴에서 다음을 선택하십시오.

   삽입 > **행**
5. 새 측도의 **이름** 및 **설명**을 입력하십시오. 예를 들어, **오류 척도** 및 **Excel의 오류에 척도 적용**과 같습니다.
6. 셀 C14에 수식 **=COLUMN('Model Measures'!O3)**을 입력하십시오.
7. 셀 D14에 수식 **=ROW('Model Measures'!O3)+1**을 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 **10**을 입력하십시오.

11. 모델 측도 워크시트에서 셀 03에 새 열의 제목으로 설명 **오류 척도**를 입력하십시오.
12. 셀 O4에 수식 **=N4*Settings!B17**을 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 *.xlt*를 지정해서 Excel 2003 템플리트로 저장하십시오. 새 템플리트를 쉽게 찾을 수 있도록 로컬 또는 네트워크 파일 시스템의 사전 구성된 템플리트 위치에 저장할 것을 권장합니다.

이 템플리트를 사용하여 Excel에 연결하면 새 사용자 정의 측도로 오류 값이 사용 가능합니다.

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.

Gains 차트: Gains 차트는 테이블에서 이익 % 열의 값을 표시합니다. 이익은 다음 방정식을 사용하여 트리에 있는 총 적중 수에 상대적인 각 증분의 적중 비율로 정의됩니다.

(증분의 적중 수 / 총 적중 수) x 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.

제 10 장 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 목표 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 연결 함수를 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 연결 함수를 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.

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 예측변수(역할이 **입력**으로 설정됨)를 보유해야 합니다. 역할이 **모두** 또는 **없음**인 필드는 비숫자 필드이므로 무시됩니다. (필요한 경우 비숫자 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 비숫자 입력에 대한 지원도 제공하므로 이러한 필드를 기록하지 않아도 됩니다. 자세한 정보는 202 페이지의 『로지스틱 노드』의 내용을 참조하십시오.

선형 모델

선형 모델은 목표와 하나 이상의 예측변수 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. 자택 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. 예측변수(입력)에 측정 수준 제한 사항이 없습니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되며 연속형 필드는 공변량으로 사용됩니다.

목적

원하는 작업

- **새 모델 작성.** 완전히 새 모델을 작성합니다. 노드의 일반적인 작업입니다.
- **기존 모델 계속 훈련.** 노드에 의해 성공적으로 작성된 마지막 모델로 계속 훈련합니다. 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 예측변수를 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 훈련**과 함께 이 옵션을 사용하려면 Analytic Server에 연결되어야 합니다.
- **모델 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 193 페이지의 『앙상블』의 내용을 참조하십시오.

기본

자동으로 데이터 준비. 이 옵션은 모델의 예측력을 최대화하기 위해 프로시저에서 목표 및 예측변수를 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 예측변수가 참조 날짜(1970-01-01) 이후의 경과 시간이 포함된 새 연속형 예측변수로 변환됩니다. 각 시간 예측변수가 참조 시간(00:00:00) 이후의 경과 시간이 포함된 새 연속형 예측변수로 변환됩니다.
- **측정 수준 조정.** 고유 값이 5개 미만인 연속형 예측변수가 순서 예측변수로 다시 캐스팅됩니다. 고유 값이 10개보다 많은 순서 예측변수가 연속형 예측변수로 다시 캐스팅됩니다.
- **이상값 처리.** 절사 값을 넘는 연속형 예측변수 값(평균에서 3배 표준 편차)이 절사 값으로 설정됩니다.
- **결측값 처리.** 명목 예측변수의 결측값이 훈련 파티션의 최빈값으로 대체됩니다. 순서 예측변수의 결측값이 훈련 파티션의 중앙값으로 대체됩니다. 연속형 예측변수의 결측값이 훈련 파티션의 평균으로 대체됩니다.
- **지도되는 병합.** 목표와 연관하여 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 예측변수로서 값이 없기 때문에 모델에서 제외됩니다.

신뢰수준. 계수 보기에서 모델 계수의 구간 추정값을 계산하는 데 사용되는 신뢰수준입니다. 0보다 크고 100보다 작은 값을 지정하십시오. 기본값은 95입니다.

모델 선택

모델 선택 방법. 모델 선택 방법(자세한 내용은 아래 참조) 중 하나를 선택하거나 주효과 모델 항으로 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 훈련 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를

부여하도록 조정됩니다. **F 통계량**은 모델 오차에서 향상도의 통계 검정을 기준으로 합니다. **수정된 R-제곱**은 훈련 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 훈련하는 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 훈련 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **수정된 R-제곱**은 훈련 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 훈련하는 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.

앙상블

이 설정은 부스팅, 배깅 또는 아주 큰 데이터 세트가 목적에서 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.

모델 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. 모델 정확도 또는 안정성을 향상시키는 것이 목표일 때 작성할 기본 모델 수를 지정하십시오. 배깅의 경우, 붓스트랩 표본의 수입니다. 양의 정수여야 합니다.

고급

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 유사 난수 정수를 작성합니다. 기본값은 54752075입니다.

모델 옵션

모델 이름. 목표 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 목표 필드 이름입니다.

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 목표 필드의 이름이며 접두부 *$L-*이 붙습니다. 예를 들어, *sales*라는 이름의 목표 필드의 경우, 새 필드 이름은 *$L-sales*가 됩니다.

모델 요약

모델 요약 보기는 모델과 그 적합성을 한 눈에 파악할 수 있도록 요약한 스냅샷입니다.

테이블. 테이블은 다음을 포함한 일부 상위 레벨 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 × 최종 모델에 대해 수정된 R^2입니다.

자동 데이터 준비

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: 개월**은 날짜를 포함하는 필드의 값에서부터 현재 시스템 날짜까지 경과 시간(개월 수)을 계산합니다.
- **기간 유도: 시간**은 시간을 포함하는 필드의 값에서부터 현재 시스템 시간까지 경과 시간(시)을 계산합니다.

- **측정 수준을 연속형에서 순서로 변경**은 고유 값이 5개 미만인 연속형 필드를 순서 필드로 다시 캐스팅합니다.
- **측정 수준을 순서에서 연속형으로 변경**은 고유 값이 10개보다 많은 순서 필드를 연속형 필드로 다시 캐스팅합니다.
- **이상값 자름**은 절사 값을 넘는 연속형 예측변수 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' 예측변수 범주를 식별합니다. 유의적으로 다르지 않은 범주(즉 0.05보다 큰 *p*-값을 가지는 범주)가 병합됩니다.
- **상수 예측변수 제외 / 이상값 처리 후 / 범주 병합 후** 는 다른 ADP 작업을 마친 후 단일 값을 가진 예측변수를 제거합니다.

예측변수 중요도

일반적으로, 가장 중요한 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 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.

잔차

모델 잔차의 진단 차트를 표시합니다.

차트 유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 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.

이상값

이 테이블에는 모델에 대해 불필요한 영향력을 발휘하는 레코드가 나열되고 레코드 ID(필드 탭에 지정된 경우), 목표값 및 Cook의 거리가 표시됩니다. Cook의 거리는 특정 레코드를 모델 계수 계산에서

제외할 때 모든 레코드의 잔차가 얼마나 변경될 수 있는지에 대한 측도입니다. 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.

효과

이 보기는 모델에서 각 효과의 크기를 표시합니다.

유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **다이어그램.** 예측변수 중요도를 줄여 효과가 위에서 아래로 정렬되는 차트입니다. 다이어그램의 연결선은 효과 유의성을 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 모델과 개별 모델 효과에 대한 ANOVA 테이블입니다. 예측변수 중요도를 줄여 개별 효과가 위에서 아래로 정렬됩니다. 기본적으로는 테이블이 축소되어 전체 모델의 결과만 보여줌에 유의하십시오. 개별 모델 효과의 결과를 보려면 테이블에서 **수정된 모델** 셀을 클릭합니다.

예측변수 중요도. 보기에 표시되는 예측변수를 제어하는 예측변수 중요도 슬라이더가 있습니다. 모델을 변경하지는 않지만 가장 중요한 예측변수에 집중할 수 있습니다. 기본적으로 상위 10개 효과가 표시됩니다.

유의성. 예측변수 중요도를 기준으로 표시되는 것 외에 보기에 표시되는 효과를 더욱 제어하는 유의성 슬라이더가 있습니다. 슬라이더 값보다 큰 유의성 값이 있는 효과는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 효과에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의성을 기준으로 필터링된 효과가 없습니다.

계수

이 보기는 모델에서 각 계수의 값을 표시합니다. 요인(범주형 예측변수)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.

- **다이어그램.** 먼저 절편을 표시한 다음 예측변수 중요도를 줄여 위에서 아래로 효과를 정렬하는 차트입니다. 요인을 포함하는 효과 내에서 계수가 데이터 값의 오름차순으로 정렬됩니다. 다이어그램의 연결선은 계수의 부호에 따라 색상이 지정되고(다이어그램 키 참조) 계수 유의성을 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 모델 계수의 값, 유의성 검정 및 신뢰구간을 표시합니다. 절편 이후, 예측변수 중요도를 줄여 효과가 위에서 아래로 정렬됩니다. 요인을 포함하는 효과 내에서 계수가 데이터 값의 오름

차순으로 정렬됩니다. 기본적으로는 테이블이 축소되어 각 모델 모수의 계수, 유의성 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 예측변수의 경우) 모델 모수와 연결된 값 레이블을 보여주는 도구 팁이 표시됩니다. 이것은 자동 데이터 준비에서 유사한 범주의 범주형 예측변수를 병합할 때 만들어지는 새 범주를 보려는 경우에 특히 유용합니다.

예측변수 중요도. 보기에 표시되는 예측변수를 제어하는 예측변수 중요도 슬라이더가 있습니다. 모델을 변경하지는 않지만 가장 중요한 예측변수에 집중할 수 있습니다. 기본적으로 상위 10개 효과가 표시됩니다.

유의성. 예측변수 중요도를 기준으로 표시되는 것 외에 보기에 표시되는 계수를 더욱 제어하는 유의성 슬라이더가 있습니다. 슬라이더 값보다 큰 유의성 값이 있는 계수는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 계수에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의성을 기준으로 필터링된 계수가 없습니다.

평균 추정

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 예측변수가 없는 경우 평균 추정이 생성되지 않습니다.

모델 작성 요약

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.

설정

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 목표 필드의 이름이며 접두부 *$L-*이 붙습니다. 예를 들어, *sales*라는 이름의 목표 필드의 경우, 새 필드 이름은 *$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 다시 데이터를 페치하고 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.

Linear-AS 노드

IBM SPSS Modeler에는 두 가지 다른 버전의 선형 노드가 있습니다.

- **선형**은 IBM SPSS Modeler Server에서 실행되는 기존 노드입니다.
- **Linear-AS**는 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 예측변수만 선형 회귀 모형에서 사용할 수 있습니다. 정확히 하나의 목표 필드(역할이 **목표**로 설정됨) 및 하나 이상의 예측변수(역할이 **입력**으로 설정됨)를 보유해야 합니다. 역할이 **모두** 또는 **없음**인 필드는 비숫자 필드이므로 무시됩니다. (필요한 경우 비숫자 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 비숫자 입력에 대한 지원도 제공하므로 이러한 필드를 기록하지 않아도 됩니다. 자세한 정보는 202 페이지의 『로지스틱 노드』의 내용을 참조하십시오.

Linear-AS 모델

선형 모델은 목표와 하나 이상의 예측변수 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. 자택 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. 예측변수(입력)에 측정 수준 제한 사항이 없습니다. 범주형(플래그, 명목 및 순서) 필드는 모델에서 요인으로 사용되며 연속형 필드는 공변량으로 사용됩니다.

기본

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 큽니다.

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.

모델 선택

모델 선택 방법. 모델 선택 방법(자세한 내용은 아래 참조) 중 하나를 선택하거나 주효과 모델 항으로 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 훈련 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **F 통계량**은 모델 오차에서 향상도의 통계 검정을 기준으로 합니다. **수정된 R-제곱**은 훈련 세트의 적합도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니

다.**과적합 방지 기준(ASE)**은 과적합 방지 세트의 적합도(평균제곱 오차, ASE)를 기준으로 합니다. 과적합 방지 세트는 모델을 훈련하는 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 훈련 세트의 우도를 기준으로 하며 너무 복잡한 모델에 페널티를 부여하도록 조정됩니다. **수정된 R-제곱**은 훈련 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 훈련하는 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.

모델 옵션

모델 이름. 목표 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 목표 필드 이름입니다.

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 목표 필드의 이름이며 접두부 *$L-*이 붙습니다. 예를 들어, *sales*라는 이름의 목표 필드의 경우, 새 필드 이름은 *$L-sales*가 됩니다.

대화형 출력

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(제한된) 로그 우도를 기반으로 혼합 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모형이 우수함을 나타냅니다. AICC는 작은 표본 결과의 AIC를 "수정합니다". 표본 결과가 증가함에 따라 AICC는 AIC로 수렴됩니다.
- R 제곱. 이것은 선형 모델의 적합도 측도로, 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 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 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.

설정

예측값은 모델이 스코어링될 때 항상 계산됨에 유의하십시오. 새 필드의 이름은 목표 필드의 이름이며 접두부 *$L-*이 붙습니다. 예를 들어, *sales*라는 이름의 목표 필드의 경우, 새 필드 이름은 *$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 다시 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어.** 선택한 경우, 이 옵션은 데이터베이스에서 다시 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

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 소속집단을 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 플래그여야 합니다. 다항 모델의 경우 목표에서 측정 수준은 플래그, 또는 두 개 이상의 범주를 포함하는 명목일 수 있습니다. 둘 다 또는 없음으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다.

강도. 로지스틱 회귀분석 모델은 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 소속집단이 범주형 필드인 경우에 가장 효과적입니다. 소속집단이 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 208 페이지의 『로지스틱 회귀분석 고급 출력』의 내용을 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.

로지스틱 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 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**, **완전요인모델** 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 회귀분석 모델을 작성할 때 사용할 방법을 지정합니다.

- **입력.** 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 모델은 방정식에 모델 항(상수 제외)이 없는 가장 단순한 모델입니다. 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 예측변수로 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 예측변수 필드의 각 범주는 참조 범주와 비교됩니다.

  차이. 첫 번째 범주를 제외한 예측변수 필드의 각 범주는 이전 범주의 평균 효과와 비교됩니다. 역 Helmert 대비라고도 합니다.

  Helmert. 마지막 범주를 제외한 예측변수 필드의 각 범주는 후속 범주의 평균 효과와 비교됩니다.

  반복. 처음 범주를 제외한 예측변수 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 예측변수 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 211 페이지의 『로지스틱 너깃 모델 세부사항』의 내용을 참조하십시오.

다항 모델

다항 모델의 경우 다음 방법과 옵션이 사용 가능합니다.

방법. 로지스틱 회귀분석 모델을 작성할 때 사용할 방법을 지정합니다.

- **입력.** 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 모델은 방정식에 모델 항(상수 제외)이 없는 가장 단순한 모델입니다. 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.
- **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 예측변수로 포함하고, 항은 모델에서 제거만 가능합니다. 모델에 거의 기여하지 않는 모델 항은 모델을 크기 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 예측변수로 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 211 페이지의 『로지스틱 너깃 모델 세부사항』의 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 모델 항 목록을 사용하여 모델에서 항을 추가하거나 제거합니다.

모델 항. 사용자 정의 모델을 작성할 때에는 모델의 항을 명시적으로 지정해야 합니다. 목록에는 모델 항의 현재 세트가 표시됩니다. 모델 항 목록의 오른쪽에 있는 단추를 사용하여 모델 항을 추가 및 제거할 수 있습니다.

- 모델에 항을 추가하려면 새 모델 항 추가 단추를 클릭하십시오.
- 항을 삭제하려면 원하는 항을 선택하고 선택한 모델 항 삭제 단추를 클릭하십시오.

로지스틱 회귀분석 모델에 항 추가

사용자 정의 로지스틱 회귀분석 모델을 요청하면 로지스틱 회귀분석 모델 탭에서 새 모델 항 추가 단추를 클릭하여 모델에 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. 모델 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.

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 정정하는 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 210 페이지의 『로지스틱 모델 너깃』의 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 208 페이지의 『로지스틱 회귀분석 수렴 옵션』의 내용을 참조하십시오.

출력. 이 옵션을 사용하면 노드에서 작성한 모델 너깃의 고급 출력에 표시되는 추가 통계를 요청할 수 있습니다. 자세한 정보는 『로지스틱 회귀분석 고급 출력』의 내용을 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 209 페이지의 『로지스틱 회귀분석 단계별 옵션』의 내용을 참조하십시오.

로지스틱 회귀분석 수렴 옵션

로지스틱 회귀분석 모델 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.

로지스틱 회귀분석 고급 출력

회귀 모형 너깃의 고급 출력에 표시할 선택적 출력을 선택하십시오. 고급 출력을 보려면 모델 너깃을 찾아보고 **고급** 탭을 클릭하십시오. 자세한 정보는 213 페이지의 『로지스틱 모델 너깃 고급 출력』의 내용을 참조하십시오.

이항검정 옵션

모델에서 생성할 출력 유형을 선택합니다. 자세한 정보는 213 페이지의 『로지스틱 모델 너깃 고급 출력』의 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 이상값(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 213 페이지의 『로지스틱 모델 너깃 고급 출력』의 내용을 참조하십시오.

참고: **우도비 검정** 옵션을 선택하면 로지스틱 회귀분석 모델을 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 옵션』의 내용을 참조하십시오.

반복계산 히스토리 출력수준. 고급 출력에서 반복 상태를 인쇄하는 단계 구간을 선택합니다.

신뢰구간. 방정식에서 계수의 신뢰구간. 신뢰구간 수준을 지정합니다(기본값은 95%).

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 너깃의 고급 탭을 사용하여 판별 가능) Wald 통계량은 특히 불안정해지며, 이를 사용해서는 안 됩니다. 기본 설정은 우도비 통계를 사용하는 것입니다. 이항 모델의 경우 추가 옵션 **조건부**가 있습니다. 이 방법에서는 조건부 모수 추정값에 기반한 우도비 통계의 확률을 토대로 제거 검정을 제공합니다.

기준의 유의성 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 필드가 모델에 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.

로지스틱 모델 너깃

로지스틱 모델 너깃은 로지스틱 노드에서 추정하는 방정식을 나타냅니다. 여기에는 로지스틱 회귀분석 모델에서 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 로지스틱 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.

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 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

모델 방정식

다항 모델의 왼쪽 분할창에서는 로지스틱 회귀분석 모델에서 추정된 실제 방정식을 표시합니다. 목표 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. 목표 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.

$$P(group_i) = \exp(g_i) \; / \; \Sigma_k \exp(g_k)$$

여기서 g는 계산된 로그-오즈비이고 i는 범주 지수이며, k는 1부터 목표 범주의 수까지의 범위에 속합니다.

예측변수 중요도

선택적으로 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측변수 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 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.

로지스틱 모델 너깃 요약

로지스틱 회귀분석 모델의 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 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

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 참의 결과 우도를 나타내는 원시 성향 스코어를 요청할 수 있습니다. 표준 예측 및 신뢰도 값 외에도 제공됩니다. 수정된 성향 스코어는 사용할 수 없습니다. 자세한 정보는 37 페이지의 『모델링 노드 분석 옵션』의 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** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

참고: 다항 모델의 경우 **모든 확률 추가**를 선택하면 SQL 생성은 사용할 수 없습니다. 또는 명목 목표를 포함하는 모델의 경우 **신뢰도 계산**을 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 37 페이지의 『모델링 노드 분석 옵션』의 내용을 참조하십시오.

로지스틱 모델 너깃 고급 출력

로지스틱 회귀분석(**명목 회귀**라고도 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.

모델 적합 정보(다항 모델) 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 결과가 작은 연구에 전통적인 적합도 통계량보다 효과적입니다.

유사 R-제곱(옵션). 모델 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 C도 이 테이블에 표시됩니다.

정보 기준(옵션). AIC(Akaike's information criterion) 및 Schwarz의 베이지안 정보 기준(BIC)를 표시합니다.

우도비 검정(옵션). 모델 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.

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 **요인**을 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 없음으로 설정된 필드는 비숫자 필드이므로 무시됩니다.

강도. 요인 분석과 PCA는 많은 정보 내용을 포기하지 않고도 효과적으로 데이터의 복잡도를 줄일 수 있습니다. 이러한 기법을 사용하면 원시 입력 필드보다 빠르게 실행되는 더 강력한 모델을 작성할 수 있습니다.

PCA/요인 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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. 특히, 이 방법은 훈련 데이터가 다변량 정규 분포를 따른다고 가정합니다.
- **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.

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. **고유값**은 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 옵션』의 내용을 참조하십시오.

PCA/요인 노드 회전 옵션

많은 경우 보유된 요인 세트를 수학적으로 회전하면 유용성과 특히 해석 가능성을 높일 수 있습니다. 회전 방법을 선택하십시오.

- **회전 안 함.** 기본 옵션. 회전이 사용되지 않습니다.
- **베리멕스.** 각 요인의 로딩이 높은 필드의 수를 최소화하는 직교 회전 방법. 이는 요인의 해석을 단순화합니다.
- **직접 오블리민.** 사각(비직교) 회전 방법. **델타**가 0(기본값)인 경우 솔루션에 기울기가 나타납니다. 델타가 음수에 가까워질수록 요인의 기울기가 평평해집니다. 기본값 델타 0을 바꾸려면 0.8 이하의 수를 입력합니다.
- **쿼티멕스.** 각 필드를 설명하는 데 필요한 요인 수를 최소화하는 직교 방법. 이는 관측 필드의 해석을 단순화합니다.
- **이쿼멕스.** 요인을 단순화하는 베리멕스 방법과 필드를 단순화하는 쿼티멕스 방법을 조합한 회전 방법. 요인에서 주로 로드한 필드의 수와 필드 설명에 필요한 요인 수는 최소화됩니다.
- **프로멕스.** 요인이 상관되도록 하는 사각 회전. 이 회전은 직접 오블리민 회전보다 빨리 계산될 수 있으므로 큰 데이터 세트에 유용합니다. **카파**는 솔루션의 경사(요인이 상관될 수 있는 범위)를 제어합니다.

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-3*과 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 출력』의 내용을 참조하십시오.

PCA/요인 모델 너깃 방정식

요인 모델의 모델 탭은 각 요인의 요인 스코어 방정식을 표시합니다. 요인 또는 구성요소 스코어는 각 입력 필드 값에 해당 계수를 곱하고 결과를 합산하여 계산됩니다.

PCA/요인 모델 너깃 요약

요인 모델의 요약 탭에서는 모델을 생성하는 데 사용되는 필드 및 설정에 대한 추가 정보와 함께, 요소/PCA 모델에 보유된 요소 수를 표시합니다. 자세한 정보는 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

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 보유된 요인 세트에 기반한 공통성을 제공합니다.

설명된 총분산. 모델에서 요인별 설명된 총분산을 표시합니다. 초기 고유값은 초기 요인의 전체 세트에서 설명된 분산을 표시합니다. 추출 제곱합 적재량에서는 모델에 보유된 요인에서 설명된 분산을 표시합니다. 회전 제곱합 적재량에서는 회전된 요인에서 설명된 분산을 표시합니다. 사각 회전의 경우 회전 제곱합 적재량은 제곱합 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.

판별 노드

판별 분석은 소속집단에 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 소속집단이 알려진 케이스 표본으로부터 생성되며 해당 소속집단은 알 수 없으나 예측자 변수 측정을 통해 새로운 케이스에 적용될 수는 있습니다.

예. 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 목표 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 없음으로 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.

판별 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 내용을 참조하십시오.

방법. 다음 옵션을 사용하여 모델에 예측변수를 입력할 수 있습니다.

- **입력.** 이는 기본 방법으로 모든 항을 방정식에 직접 입력합니다. 모델의 예측력을 상당히 증가시키지 않는 항은 추가되지 않습니다.
- **단계 선택.** 이 초기 모델은 방정식에 모델 항(상수 제외)이 없는 가장 단순한 모델입니다. 각 단계마다 모델에 아직 추가되지 않은 항을 평가하여 최상의 항이 모델의 예측력을 상당히 증가시킬 경우 이 항이 추가됩니다.

참고: 단계선택법은 훈련 데이터를 과적합할 경향이 높습니다. 이 방법을 사용할 때에는 검증용 검정 표본이나 새 데이터로 결과적인 모델의 유효성을 검증하는 것이 특히 중요합니다.

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 소속집단의 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 소속집단의 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 행렬이 케이스 분류에 사용됩니다.
- 개별-그룹. 개별-그룹 공분산 행렬이 분류에 사용됩니다. 분류가 판별 함수에 기초하고 원래 변수에 따라 달라지지는 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 옵션』의 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 221 페이지의 『판별 노드 단계 옵션』의 내용을 참조하십시오.

판별 노드 출력 옵션

로지스틱 회귀분석 모델 너깃의 고급 출력에 표시하려는 선택적 출력을 선택하십시오. 고급 출력을 보려면 모델 너깃을 찾아보고 **고급** 탭을 클릭하십시오. 자세한 정보는 222 페이지의 『판별 모델 너깃 고급 출력』의 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.
- *Box*의 *M*. 그룹 공분산 행렬의 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.

- *Fisher*의 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 행렬의 평균을 구하여 그룹 내 풀링 상관 행렬을 표시합니다.
- 그룹-내 공분산 행렬. 그룹 내 풀링 공분산 행렬을 표시하는데 이는 전체 공분산 행렬과 다를 수 있습니다. 이 행렬은 모든 그룹에 대해 개별 공분산 행렬을 평균하여 구합니다.
- 개별-그룹 공분산 행렬. 각 그룹에 대해 개별 공분산 행렬을 표시합니다.
- 전체 공분산. 단일 표본으로 작성한 것처럼 모든 케이스로부터 공분산 행렬을 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 유의성 검정에 *F* 비율을 사용할 수 있습니다.

판별 노드 단계 옵션

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.

- *Wilks*의 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.
- *Mahalanobis*의 거리. 독립변수의 케이스 값이 전체 케이스 평균과 얼마나 달라지는지에 대한 측도입니다. 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.
- *Rao*의 *V*. 그룹 평균 간 차이에 대한 측도입니다. Lawley-Hotelling 트레이스라고도 하며 각 단계에서 Rao의 V의 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 모델에 변수가 입력되고 F 값이 제거값보다 작으면 제거됩니다. 진입값은 제거값보다 커야 하고 두 값 모두 양수이어야 합니다. 모델에 더 많은 변수를 입력하려면 진입값을 낮추고 모델에서 변수를 더 많이 제거하려면 제거값을 높입니다.
- *F*-확률 사용. F 값의 유의 수준이 진입값보다 작으면 모델에 변수가 입력되고 유의 수준이 제거값보다 크면 제거됩니다. 진입값은 제거값보다 작아야 하며 두 값 모두 양수이어야 합니다. 모델에 변수를 더 많이 입력하려면 진입값을 높이고 모델에서 변수를 더 많이 제거하려면 제거값을 낮춥니다.

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 메뉴에서는 모델 결과에 기반하여 입력 필드를 전달할 새 필터 노드를 작성할 수 있습니다.

예측변수 중요도

선택적으로 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측변수 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

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 219 페이지의 『판별 노드 출력 옵션』의 내용을 참조하십시오.

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.

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.

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 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

GenLin 노드

일반화 선형 모델은 종속변수가 지정된 연결 함수를 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 중도 절단 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 목표 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.

GenLin 노드 필드 옵션

일반적으로 모델링 노드 필드 탭에서 제공되는 목표, 입력, 파티션 사용자 정의 옵션 외에도(33 페이지의 『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.

GenLin 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 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" 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.

참고*:* 변수 변위 필드를 사용하는 경우 지정된 필드는 입력으로 사용할 수 없습니다. 업스트림 소스 또는 필요한 경우 유형 노드에서 변위 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 모델 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.

GenLin 노드 고급 옵션

일반화 선형 모델에 대한 자세한 지식이 있는 경우 고급 옵션을 통해 훈련 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 **모드**를 **고급**으로 설정하십시오.

목표 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.

링크 함수.

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 k는 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 $\alpha \neq 0$. $f(x)=\log(x)$(α=0의 경우). α는 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). $f(x)=\log(x)$(α=0의 경우). α는 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 계속됩니다.
- **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 분포 인 경우 올바르지 않습니다. 편차 및 Pearson 카이제곱 옵션은 해당 통계값에서 척도 모수를 추정합니다. 또한 척도 모수에 대한 고정 값을 지정할 수 있습니다.

- **공분산 행렬.** 모델 기반 추정량은 Hessian 행렬의 일반화 역의 음수입니다. 동질성(Huber/White/sandwich라고도 함) 추정량은 변수와 연결 함수의 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 반복』의 내용을 참조하십시오.

출력. 이 옵션을 사용하면 노드에서 작성한 모델 너깃의 고급 출력에 표시되는 추가 통계를 요청할 수 있습니다. 자세한 정보는 228 페이지의 『일반화 선형 모델 고급 출력』의 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.

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 응답 에 사용 가능합니다.

수렴 기준. 다음 옵션을 사용할 수 있습니다.

- **모수 수렴.** 이 옵션을 선택하면 모수 추정값의 절대 변화량 또는 상대 변화량이 지정된 값보다 작아지는 반복 후에 알고리즘이 멈춥니다. 지정된 값은 양수여야 합니다.
- **로그-우도 수렴.** 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.

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 230 페이지의 『GenLin 모델 너깃 고급 출력』의 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), 베이지안 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 모델 적합 총괄 검정에 대한 우도비 통계를 포함한 모델 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 $0^{번째}$ 반복(초기 추정값)부터 시작하여 각 $n^{번째}$ 반복마다 모수 추정값을 표시합니다. 여기서 n은 인쇄 구간의 값입니다. 반복계산과정을 요청하는 경우 n에 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 예측변수에 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.** 50보다 크고 100보다 작은 신뢰수준을 지정하십시오. Wald 구간은 매개변수에 근사 정규 분포가 있다는 가정을 기반으로 합니다. 프로파일 우도 구간은 더 정확하지만 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.

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 내용이 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 메뉴에서는 모델 결과에 기반하여 입력 필드를 전달할 새 필터 노드를 작성할 수 있습니다.

예측변수 중요도

선택적으로 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측변수 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

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 228 페이지의 『일반화 선형 모델 고급 출력』의 내용을 참조하십시오.

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 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.

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 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

일반화 선형 혼합 모델

GLMM 노드

이 노드를 사용하여 일반화 선형 혼합 모델(GLMM)을 작성합니다.

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. 목표 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 *ID*와 환자 *ID*의 조합이 필요할 수 있습니다. 반복 측도 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 개체로 지정된 모든 필드는 잔차 공분산 구조에 대한 개체를 정의하는 데 사용되며 랜덤 효과 블록에서 랜덤 효과 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)
- 자귀회귀 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

목표: 이 설정은 목표, 분포 및 연결 함수를 통한 예측변수에 대한 관계를 정의합니다.

목표. 목표는 필수입니다. 어떤 측정 수준도 가질 수 있으며, 목표의 측정 수준은 적합한 분포 및 연결 함수를 제한합니다.

- **시행 수를 분모로 사용.** 목표 반응이 시행 세트에서 발생하는 많은 이벤트면 목표 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 연결 함수를 통해 예측변수와 선형적으로 관련됩니다. 여러 공통 모델에 대한 바로 가기가 제공되거나, 최종 목록에 없는 적합하도록 할 특정 분포 및 연결 함수 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 연결 함수를 지정하는 기능은 선형 혼합 모델에서 일반화 선형 혼합 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 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.

연결 함수

연결 함수는 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). α는 필수 숫자 지정 사항이며 실수여야 합니다. 이 링크는 다항을 제외한 모든 분포에 사용할 수 있습니다.

고정 효과: 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 입

력 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 핫스폿에 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 *10.* 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 추가할 수 있습니다.

절편 포함. 이 모델에는 대개 절편이 포함됩니다. 데이터가 선형 회귀로 전달된다고 가정할 경우에는 절편을 제외할 수 있습니다.

사용자 정의 항 추가: 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 **내에 중첩**되었다고 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.

랜덤효과: 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 랜덤 효과 블록이 생성됩니다. 예를 들어 데이터 구조 탭에서 학교, 클래스 및 학생을 개체로 선택한 경우, 다음 랜덤 효과 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』 대화 상자가 열립니다.
2. 기존 블록을 편집하려면, 편집할 블록을 선택하고 **블록 편집...**을 클릭하십시오, 그러면 『랜덤 효과 블록』 대화 상자가 열립니다.
3. 하나 이상의 블록을 삭제하려면 삭제하려는 블록을 선택하고 삭제 단추를 클릭하십시오.

랜덤 효과 블록*:* 소스 목록에서 하나 이상의 필드를 선택하고 효과 목록으로 끌어 모델에 효과를 입력합니다. 생성되는 효과의 유형은 선택을 끄는 핫스폿에 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 *11.* 효과 작성기 단추 설명.

아이콘	설명
	삭제할 항을 선택하고 삭제 단추를 클릭하여 모델에서 항목을 삭제할 수 있습니다.
	다시 정렬할 항을 선택하고 위로 또는 아래로 화살표를 클릭하여 모델에서 항목을 다시 정렬할 수 있습니다.
	235 페이지의 『사용자 정의 항 추가』 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 전달된다고 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 학생이 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)
- 자귀회귀 이동 평균(1,1)(ARMA11)
- 복합 대칭
- 대각선
- 척도화 항등
- Toeplitz
- 비구조적
- 분산성분

가중치 및 변위: 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

관련한 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" 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.

일반 작성 옵션: 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 232 페이지의 『목표』 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, 척도법 항등 또는 대각선처럼 모델이 간단한 공분산 유형을 사용하는 경우 **모든 검정에 대해 고정(잔차 방법)**을 선택합니다. 이는 기본값입니다. 표본 크기가 작거나, 데이터가 비균형적이거나, 비구조적처럼 모델이 복잡한 공분산 유형을 사용하는 경우 **검정마다 다름(Satterthwaite approximation)**을 선택합니다.
- **고정 효과 및 계수의 검정.** 모수 추정값 공분산 행렬을 계산하는 방법입니다. 모델 가정을 위반할 염려가 있는 경우 강력한 추정을 선택하십시오.

추정: 모델 작성 알고리즘은 내부 루프와 외부 루프로 구성되는 이중 반복 프로세스를 사용합니다. 다음 설정은 내부 루프에 적용됩니다.

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.

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.

일반: 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.

평균 추정: 이 탭을 사용하면 요인 대 요인 상호작용의 수준에 대해 주변 평균 추정을 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**이 선택되면 대비가 생성되지 않습니다. **대응별**은 지정된 요인의 모든 수준 조합에 대해 쌍대 비교를 생성합니다. 이는 요인 상호작용의 대비에만 사용할 수 있습니다. **편차** 대비는 요인의 각 수준을 총 평균과 비교합니다. **단순** 대비는 마지막을 제외한 요인의 각 수준을 마지막 수준과 비교합니다. "마지막" 수준은 작성 옵션에 지정된 요인의 정렬 순서에 의해 결정됩니다. 이러한 대비 유형은 모두 직교하지 않음에 유의하십시오.
- **대비 필드.** 선택된 대비 유형을 사용하여 비교되는 수준인 요인은 지정합니다. **없음**을 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.

다음과 관련하여 추정 평균 표시. 주변 평균 추정을 목표의 원래 척도를 기준으로 계산할지 연결 함수 변환을 기준으로 계산할지 지정합니다.**원래 목표 척도**는 목표의 주변 평균 추정을 계산합니다. 목표가 이벤트/시행 옵션을 사용하여 지정되었을 때 이벤트 수보다는 이벤트/시행 비율에 대한 주변 평균 추정을 제공합니다. **연결 함수 변환**은 선형 예측변수의 주변 평균 추정을 계산합니다.

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.

모델 보기: 기본적으로 모델 요약 보기가 표시됩니다. 다른 모델 보기를 보려면 보기 축소판 그림에서 선택하십시오.

모델 요약*:* 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(제한) 로그 우도에 기반한 혼합 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. AICC는 작은 표본 크기의 AIC를 "수정합니다". 표본 크기가 증가함에 따라 AICC는 AIC로 수렴됩니다.

- 베이지안. -2 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. BIC도 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여" 하지만 AIC보다 더 엄격하게 부여합니다.

차트. 목표가 범주형인 경우 차트는 최종 모델의 정확도를 표시하며 이는 정확한 분류의 퍼센트입니다.

데이터 구조: 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: 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: 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: 이 보기는 모델에서 각 고정 효과의 크기를 표시합니다.

유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **다이어그램.** 고정 효과 설정에 지정된 순서로 효과가 위에서 아래로 정렬되는 차트입니다. 다이어그램의 연결선은 효과 유의성을 기준으로 가중되며 선이 굵을수록 더 유의한 효과(더 작은 p-값)입니다. 이는 기본값입니다.
- **테이블.** 전체 모델과 개별 모델 효과에 대한 ANOVA 테이블입니다. 고정 효과 설정에 지정된 순서로 각 효과가 위에서 아래로 정렬됩니다.

유의성. 보기에 표시되는 효과를 제어하는 유의성 슬라이더가 있습니다. 슬라이더 값보다 큰 유의성 값이 있는 효과는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 효과에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의성을 기준으로 필터링된 효과가 없습니다.

고정 계수: 이 보기는 모델에서 각 고정 계수의 값을 표시합니다. 요인(범주형 예측변수)이 모델 내에서 코딩된 지표이므로 요인을 포함하는 **효과**에는 일반적으로 여러 관련 **계수**가 있으며, 중복 계수에 해당하는 범주를 제외하고 각 범주에 하나씩 있습니다.

유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **다이어그램.** 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 모델 계수의 값, 유의성 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 지수화한 계수 추정값 및 신뢰구간을 표시합니다.

유의성. 보기에 표시되는 계수를 제어하는 유의수준 슬라이더가 있습니다. 슬라이더 값보다 큰 유의성 값이 있는 계수는 숨겨집니다. 모델을 바꾸지는 않지만 가장 중요한 계수에 집중할 수 있습니다. 기본적으로 이 값은 1.00이므로, 유의성을 기준으로 필터링된 계수가 없습니다.

랜덤효과 공분산: 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*:* 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*: 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 예측변수는 상수로 유지됩니다. 목표에서 각 예측변수 계수의 효과를 시각화하는 데 유용합니다. 유의한 예측변수가 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*: 다음은 사용자가 요청한 고정 모든 요인 효과에 대한 테이블과 차트입니다.

유형. 다양한 표시 유형이 있으면, **유형** 드롭다운 목록에서 액세스할 수 있습니다.

- **다이어그램.** 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 예측변수는 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 나타납니다.

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 나타납니다.

- **테이블.** 이 유형은 목표의 모델 추정값, 표준 오차 및 효과에서 필드의 각 수준 조합에 대한 신뢰구간의 테이블을 표시합니다. 기타 모든 예측변수는 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: 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 다시 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

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.

목표

이 설정은 목표, 분포 및 연결 함수를 통한 예측변수에 대한 관계를 정의합니다.

목표 목표는 필수입니다. 어떤 측정 수준도 가질 수 있으며, 대상의 측정 수준은 적합한 분포 및 연결 함수에 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, 목표 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 연결 함수를 통해 해당 예측변수와 선형적으로 관련될 것으로 예상합니다. 여러 공통 모델에 대한 바로 가기가 제공되거나, 최종 목록에 없는 적합하도록 할 특정 분포 및 연결 함수 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.
- **감가 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 연결 함수를 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 연결 함수의 고유 조합을 지정합니다.

분포

이 선택사항은 목표의 **분포**를 지정합니다. 비정규 분포와 항등하지 않은 연결 함수를 지정하는 기능은 선형 모델 중에서 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.

- **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.

연결 함수

연결 함수는 모델을 추정할 수 있는 목표의 변환입니다. 다음 함수를 사용할 수 있습니다.

- **자동** 사용할 연결을 모르는 경우 이 옵션을 선택하십시오. 노드가 데이터를 분석하여 최적의 연결 함수를 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 $\alpha \neq 0$. $f(x)=\log(x)$(α=0의 경우). α는 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.

모델 효과

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 핫스폿에 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 *12.* 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 시용자 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.

사용자 정의 항 추가

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 내에 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.

가중치 및 오프셋

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.

작성 옵션

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 246 페이지의 『목표』 설정에 지정된 경우, 목표 정렬 순서 설정이 무시됩니다.

추정 후 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 선택합니다. 이는 기본값입니다. 표본 크기가 작거나, 데이터가 비균형적이거나, 비구조적처럼 모델이 복잡한 공분산 유형을 사용하는 경우 **검정마다 다름(Satterthwaite approximation)**을 선택합니다.
- **고정 효과 및 계수의 검정**. 모수 추정값 공분산 행렬을 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.

추정

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 **값**을 설정할 수도 있습니다.
- 편차
- Pearson 카이제곱

음이항 방법 음이항 보조 모수에 대한 추정 방법을 선택하십시오. 사용가능 옵션은 다음과 같습니다.

- 최대우도 추정값
- 고정값. 사용할 **값**을 설정할 수도 있습니다.

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.

모델 선택

모델 선택 또는 정규화 사용 이 분할창에서 제어를 활성화하려면 이 확인 상자를 선택하십시오.

방법 모델 선택 방법을 선택하거나 사용할 정규화(**능형**을 사용하는 경우)를 선택하십시오. 다음 옵션에서 선택할 수 있습니다.

- **Lasso** L1 정규화라고도 하며 이 방법은 예측변수 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 **방법**을 선택하는 경우에만 이 옵션을 사용할 수 있습니다.

페널티 모수 자동 선택 설정할 모수 페널티를 모르는 경우 이 확인 상자를 선택하면 노드가 페널티를 식별하고 적용합니다.

Lasso 페널티 모수 Lasso 모델 선택 **방법**에서 사용할 페널티 모수를 입력하십시오.

Elastic net 페널티 모수 1 Elastic net 모델 선택 **방법**에서 사용할 L1 페널티 모수를 입력하십시오.

Elastic net 페널티 모수 2 Elastic net 모델 선택 **방법**에서 사용할 L2 페널티 모수를 입력하십시오.

단계별 전진

단계별 전진 **방법**을 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 확인 상자를 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 확인 상자를 선택하십시오.

최대 단계 수 단계별 전진 작성 방법을 사용하는 경우 최대 단계 수를 지정하십시오.

모델 옵션

모델 이름 - 목표 필드를 기반으로 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 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 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

GLE 모델 너깃

GLE 모델 너깃 출력

GLE 모델을 작성하면 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.

- 유형 노드 또는 GLE 노드 필드 탭에서 선택된 목표 필드 이름.
- 모델링된 참조 대상 범주 퍼센트.
- 확률 분포 및 연관된 연결 함수.
- 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.

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 다시 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택된 경우, 이 옵션은 데이터베이스에서 다시 사용자 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

Cox 노드

Cox 회귀는 시간-이벤트 데이터의 예측 모델을 작성합니다. 모델은 예측자 변수의 주어진 값에 대해 특정 시간 *t*에 중요 이벤트가 발생했을 확률을 예측하는 생존함수를 생성합니다. 생존 함수의 모양과 예측변수의 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 플래그이고 문자열 또는 정수 저장 공간이 있어야 합니다. (필요에 따라 채움 또는 파생 노드를 사용하여 저장 공간을 변환할 수 있습니다.) 둘 다 또는 없음으로 설정된 필드는 무시됩니다. 모델에 사용된 필드는 유형이 완전히 인스턴스화되어 있어야 합니다. 생존 시간은 수치 필드일 수 있습니다.

참고: Cox 회귀 모형을 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.

Cox 노드 필드 옵션

생존 시간. 노드를 실행 가능하게 하려면 수치 필드(측정 수준이 연속인)를 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 33 페이지의 『모델링 노드 필드 옵션』의 내용을 참조하십시오.

Cox 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 내용을 참조하십시오.

방법. 다음 옵션을 사용하여 모델에 예측변수를 입력할 수 있습니다.

- **입력.** 기본 방법으로, 모델에 직접 모든 항을 입력합니다. 모델 작성 시 필드 선택은 수행되지 않습니다.
- **단계 선택.** 필드 선택의 단계선택법은 이름에 내포되어 있듯이 단계별로 모델을 작성합니다. 초기 모델은 가능한 가장 단순한 모델로, 모델의 모델 항이 없습니다(상수 제외). 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 후진 방법은 본질적으로 단계선택법의 반대 개념입니다. 이 방법에서 초기 모델은 모든 항을 예측변수로 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: 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 항을 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 항(주효과 및 상호작용)만 포함합니다. 이 옵션을 선택하면 모델 항 목록을 사용하여 모델에서 항을 추가하거나 제거합니다.

모델 항. 사용자 정의 모델을 작성할 때에는 모델의 항을 명시적으로 지정해야 합니다. 목록에는 모델 항의 현재 세트가 표시됩니다. 모델 항 목록의 오른쪽에 있는 단추로 모델 항을 추가 및 제거할 수 있습니다.

- 모델에 항을 추가하려면 새 모델 항 추가 단추를 클릭하십시오.
- 항을 삭제하려면 원하는 항을 선택하고 선택한 모델 항 삭제 단추를 클릭하십시오.

Cox 회귀 모형에 항 추가

사용자 정의 모델을 요청할 때 모델 탭에서 새 모델 항 추가 단추를 클릭하여 모델에 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. 모델 항을 구성할 때 사용할 사용 가능한 입력 필드를 나열합니다. 목록에 적합하지 않은 입력 필드가 포함될 수 있으므로 모든 모델 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.

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 제어합니다. 모수를 더 자주 시도할 수록 결과에 더 근접합니다(즉, 결과가 수렴됨). 자세한 정보는 258 페이지의 『Cox 노드 수렴 기준』의 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 258 페이지의 『Cox 노드 고급 출력 옵션』의 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 기준』의 내용을 참조하십시오.

Cox 노드 수렴 기준

최대반복계산. 프로시저가 솔루션을 찾는 데 걸리는 기간을 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.

Cox 노드 고급 출력 옵션

통계. 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 생존함수를 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 ? 생존함수를 도표화합니다.

각 값의 선구분 집단변수 도표 표시. 이 옵션은 범주형 필드에만 사용 가능합니다.

도표에 사용할 값. 이 함수는 예측변수의 값에 종속되므로 함수 대 시간으로 도표 표시하려면 예측변수의 상수 값을 사용해야 합니다. 기본값은 각 예측변수의 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.

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 유의성 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 필드가 모델에 추가되고 *p* 값이 **제거** 값보다 큰 경우에만 필드가 제거됩니다. **입력** 값은 **제거** 값보다 작아야 합니다.

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 수**에서 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.

Cox 모델 너깃

Cox 회귀 모형은 Cox 노드가 추정한 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 간격이 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 수입니다. 예를 들어, "거짓" 값이 0이고 "참" 값이 1이며 두 개의 미래 시간 간격이 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 *$CH-*이고 접미문자는 미래 시간 간격의 수 또는 시간 간격을 정의하는 시간 필드의 이름입니다. 예를 들어, 두 개의 미래 시간 간격이 주기적으로 정의된 *churn*이란 출력 필드의 경우 새 필드의 이름은 *$CH-churn-1* 및 *$CH-churn-2*입니다. 미래 시간이 시간 필드 *tenure*로 정의되면 새 필드는 *$CH-churn-1*입니다.

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 259 페이지의 『Cox 노드 설정 옵션』의 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** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

Cox 회귀분석 고급 출력

Cox 회귀분석의 고급 출력은 생존 곡선을 포함하여 추정한 모델 및 성능에 대한 자세한 정보를 제공합니다. 고급 출력에 포함된 대부분의 정보는 상당히 기술적 정보이며 이 출력을 제대로 해석하려면 광범위한 Cox 회귀분석 지식이 필요합니다.

제 11 장 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 군집화 모델을 다른 머신 학습 기법과 구별합니다. 예측할 모델에 대한 사전정의된 출력 또는 목표 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. 값은 데이터에서 관심 있는 집단을 캡처하고 이 집단에 대한 유용한 설명을 제공하는 기능으로 판별됩니다.

군집화 방법은 레코드 간 그리고 군집 간의 거리 측정을 기반으로 합니다. 레코드는 동일한 군집에 속한 레코드 사이의 거리를 최소화하려는 방식으로 군집에 할당됩니다.

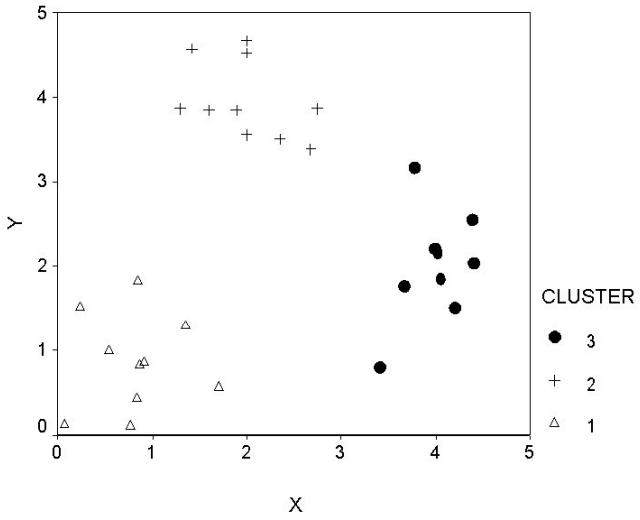


그림 *44.* 단순 군집화 모델

세 가지 군집화 방법이 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 훈련 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 것 입니다. 모델 너깃에서 각 단위별로 캡처된 관측값을 살펴 강한 단위를 식별할 수 있습니다. 이것은 적당한 군집 수에 대한 감각을 제공할 것 입니다.

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.

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 **뉴런**이며 두 개의 레이어, **입력층** 및 **출력층**(**출력 맵**이라고도 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 **승자**라고 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 예측변수 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 **자율 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 예측변수에서, 원래 예측변수와 유사 관계를 유지하는 2개의 파생된 기능으로의 맵핑을 제공합니다. 일부 경우에 이는 요인 분석 또는 PCA와 동일한 종류의 혜택을 제공할 수 있습니다.

출력 눈금의 기본 크기를 계산하는 방법은 IBM SPSS Modeler의 이전 버전과 달라짐에 유의하십시오. 일반적으로 새로운 방법은 더 빠르게 훈련하고 더 효율적으로 일반화하는 더 작은 출력층을 생성합니다. 기본 크기의 열악한 결과를 얻은 경우 고급 탭에서 출력 눈금의 크기를 늘리십시오. 자세한 정보는 264 페이지의 『코호넨 노드 고급 옵션』의 내용을 참조하십시오.

요구사항. 코호넨 넷을 훈련하려면 역할이 입력으로 설정된 하나 이상의 필드가 필요합니다. 역할이 목표, 둘 다 또는 없음으로 설정된 필드는 무시됩니다.

강도. 코호넨 네트워크 모델을 작성하기 위해 소속집단에 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.

코호넨 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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: **난수 시드 설정** 옵션을 데이터베이스에서 읽은 레코드와 함께 사용할 경우에는 노드를 실행할 때마다 동일한 결과가 보장되도록 표본추출 이전에 정렬 노드가 필요할 수 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 스필링을 사용하지 않게 하려면 **속도**를 선택하십시오.
- 적합한 시기에 속도가 느려지더라도 알고리즘이 디스크 스필링을 사용하게 하려면 **메모리**를 선택하십시오. 이 옵션은 기본적으로 선택됩니다.

  참고: 분산 모드에서 실행할 때에는 *options.cfg*에 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 83 페이지의 『자동 군집 노드』의 내용을 참조하십시오.

코호넨 노드 고급 옵션

코호넨 네트워크에 대한 자세한 지식을 가진 사용자는 고급 옵션을 통해 훈련 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 훈련은 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 훈련 중 획득한 단위와 함께 업데이트되는 " 근접" 단위 수를 판별합니다. 1단계 중에 이웃 크기는 *1*단계 이웃으로 시작하고 (*2*단계 이웃 + 1)로 감소합니다. 2단계 중에 이웃 크기는 *2*단계 이웃부터 시작하여 1.0으로 감소합니다. *1*단계 이웃은 *2*단계 이웃보다 커야 합니다.

- **초기 에타** 학습률 **에타**의 시작값을 설정합니다. 1단계 중에 에타는 *1*단계 초기 에타로 시작하고 *2* 단계 초기 에타로 감소합니다. 2단계 중에 에타는 *2*단계 초기 에타로 시작하고 0으로 감소합니다. *1*단계 초기 에타는 *2*단계 초기 에타보다 커야 합니다.
- **순환** 각 훈련 단계에서 순환 수를 설정합니다. 각 단계는 전체 데이터에서 지정된 패스 수만큼 계속됩니다.

코호넨 모델 너깃

코호넨 모델 너깃은 훈련된 코호넨 네트워크에서 캡처한 모든 정보와 네트워크 설계에 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 279 페이지의 『군집 뷰어 - 모델 탭』의 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. 자세한 정보는 118 페이지의 『C5.0 노드』 주제를 참조하십시오.

모델 브라우저 사용에 대한 일반 정보는 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

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.

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 없음으로 설정된 필드는 무시됩니다.

강도. K-평균 모델을 작성하기 위해 소속집단에 데이터는 없어도 됩니다. K-평균 모델은 대형 데이터 세트 군집을 위한 가장 빠른 방법이기도 합니다.

K-평균 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 너깃은 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 **숫자**로 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 스필링을 사용하지 않게 하려면 **속도**를 선택하십시오.
- 적합한 시기에 속도가 느려지더라도 알고리즘이 디스크 스필링을 사용하게 하려면 **메모리**를 선택하십시오. 이 옵션은 기본적으로 선택됩니다.

  참고: 분산 모드에서 실행할 때에는 *options.cfg*에 지정된 관리자 옵션이 이 설정을 대체할 수 있습니다.

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.

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. 자세한 정보는 118 페이지의 『C5.0 노드』 주제를 참조하십시오. 모델 너깃 브라우저에서 모델 탭을 클릭하여 군집, 필드, 중요도 수준에 대한 그래픽 표시를 제공하는 군집 뷰어를 표시할 수도 있습니다. 자세한 정보는 279 페이지의 『군집 뷰어 - 모델 탭』의 내용을 참조하십시오.

모델 브라우저 사용에 대한 일반 정보는 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

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.

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 없음으로 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. **이상값** 또는 결과를 오염시킬 수 있는 매우 비정상적인 케이스를 자동으로 제외하도록 이단계 군집을 설정할 수 있습니다.

중요사항:

IBM SPSS Modeler에는 두 가지 다른 버전의 이단계 군집 노드가 있습니다.

- **이단계 군집**은 IBM SPSS Modeler Server에서 실행하는 일반적인 노드입니다.
- **TwoStep-AS 군집**은 IBM SPSS Analytic Server에 연결되었을 때 실행할 수 있습니다.

이단계 군집 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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, 이단계는 평균이 0이고 분산이 1인 동일 척도로 모든 숫자 입력 필드를 표준화합니다. 숫자 필드에 대한 원래 배율을 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 충분하게 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 이단계 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 접두문자**가 있는 **문자열**(예: "Cluster 1", "Cluster 2" 등)이나 **숫자**로 표시할 수 있습니다.

자동으로 군집 수 계산. 이단계 군집은 훈련 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. 이단계는 최적 군집 수를 결정하기 위해 2 단계 프로세스를 사용합니다. 첫 번째 단계에서, 모델의 군집 수에 대한 상한은 추가 군집이 추가되는 대로 베이즈 정보 기준(BIC)의 변경을 기반으로 선택됩니다. 두 번째 단계에서, 군집 사이 최소 거리의 변경은 최소 BIC 솔루션 보다 군집이 적은 모든 모델에 대해 발견됩니다. 거리에서 가장 큰 변경은 최종 군집 모델을 식별하기 위해 사용됩니다.

군집 수 지정. 모델에 포함할 군집 수를 알고 있는 경우, 이 옵션을 선택하고 군집 수를 입력하십시오.

거리 측도. 이 선택에서는 두 군집 간 유사성이 계산되는 방식이 결정됩니다.

- **로그-우도.** 우도 측도는 변수에 확률 분포를 둡니다. 연속형 변수는 정규 분포로, 범주형 변수는 다항분포로 가정됩니다. 모든 변수를 독립변수로 가정합니다.
- **유클리디안.** 유클리디안 측도는 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.

이단계 군집 모델 너깃

이단계 군집 모델 너깃은 군집 모델이 캡처한 모든 정보 외에 훈련 데이터 및 추정 프로세스에 대한 정보를 포함합니다.

이단계 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.

이단계 모델을 통찰하는 강력한 기술은 규칙 귀납을 사용하여 모델이 찾은 군집을 구별하는 특성을 발견하는 것입니다. 자세한 정보는 118 페이지의 『C5.0 노드』 주제를 참조하십시오. 모델 너깃 브라우저에서 모델 탭을 클릭하여 군집, 필드, 중요도 수준에 대한 그래픽 표시를 제공하는 군집 뷰어를 표시할 수도 있습니다. 자세한 정보는 279 페이지의 『군집 뷰어 - 모델 탭』의 내용을 참조하십시오.

모델 브라우저 사용에 대한 일반 정보는 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

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 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

TwoStep-AS 군집 노드

IBM SPSS Modeler에는 두 가지 다른 버전의 이단계 군집 노드가 있습니다.

- **이단계 군집**은 IBM SPSS Modeler Server에서 실행하는 일반적인 노드입니다.
- **TwoStep-AS 군집**은 IBM SPSS Analytic Server에 연결되었을 때 실행할 수 있습니다.

Twostep-AS 군집분석

이단계 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.** 이단계 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.

필드 탭

필드 탭은 분석에서 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.

기본

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. **수**는 1보다 커야 합니다.

군집 기준

이 선택사항은 자동 군집 알고리즘이 군집 수를 결정하는 방식을 제어합니다.

베이지안 정보 기준(BIC)

-2 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. BIC도 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여"하지만 AIC보다 더 엄격하게 부여합니다.

Akaike 정보 기준(AIC)

-2 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. AIC는 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)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 유의성 값 대신에 효과 크기를 기반으로 합니다.

기능 트리 기준

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 다항분포로 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.

이상값 군집

이상값 군집 포함

보통 군집에서 이상값인 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 이상값으로 간주됩니다. 값은 1보다 큰 정수여야 합니다. 이 값을 변경하는 경우, 더 높은 값은 결과적으로 기타 이상값 군집이 될 수 있습니다.

이상값의 위쪽 퍼센트.

군집 모델이 작성될 때, 이상값은 이상값 강도에 의해 순위가 매겨집니다. 이상값의 상위 퍼센트에 있어야 하는 이상값 강도는 케이스가 이상값으로 분류되는지 여부를 판별하기 위한 임계값으로 사용됩니다. 값이 높을 수록 더 많은 케이스가 이상값으로 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 기능 트리의 재작성 지연. 군집 알고리즘은 새 케이스를 평가하는 만큼 여러 번 군집 기능 트리를 다시 작성합니다. 이 옵션은 작업을 지연하고 트리가 다시 작성되는 횟수를 줄여서 성능을 개선할 수 있습니다.

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.

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.

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.

이상값 군집

이상값을 포함하지 않도록 선택한 경우 이 옵션은 사용 안함으로 설정됩니다.

대화형 테이블 및 차트

이상값 강도와 보통 군집에 대한 이상값 군집의 상대 유사성의 테이블 및 차트. 테이블에서 다른 행을 선택하면 차트에서 다른 이상값 군집에 대한 정보를 표시합니다.

피벗 테이블

이상값 강도와 보통 군집에 대한 이상값 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 이상값 수. 20보다 큰 이상값 군집이 있는 경우, 피벗 테이블이 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 테이블.

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

텍스트

각 군집에 대한 레이블은 **접두문자**에 대해 지정된 값이며, 뒤에 순차 번호가 있습니다.

번호 각 군집에 대한 레이블은 순차 번호입니다.

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.

TwoStep-AS 군집 모델 너깃

TwoStep-AS 모델 너깃은 출력 뷰어의 모델 탭에서 모델의 세부사항을 표시합니다. 뷰어 사용에 대한 자세한 정보는 모델러 사용자 안내서(ModelerUsersGuide.pdf)에서 "출력에 대한 작업" 절을 참조하십시오.

TwoStep-AS 군집 모델 너깃은 군집 모델이 캡처한 모든 정보 외에 훈련 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. 자세한 정보는 118 페이지의 『C5.0 노드』 주제를 참조하십시오.

모델 브라우저 사용에 대한 일반 정보는 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

TwoStep-AS 군집 모델 너깃 설정

설정 탭에서는 TwoStep-AS 모델 너깃에 대한 추가 옵션을 제공합니다.

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** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

K-평균-AS 노드

K-평균은 일반적으로 가장 많이 사용되는 군집 알고리즘 중 하나입니다. 여기에서는 데이터 포인트를 사전 정의된 갯수의 군집으로 모읍니다.[1] SPSS Modeler에서 K-평균-AS 노드는 Spark로 구현됩니다.

K-평균 알고리즘에 대한 자세한 정보는 https://spark.apache.org/docs/2.2.0/ml-clustering.html의 내용을 참조하십시오.

K-평균-AS 노드는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

K-평균-AS 노드 필드

필드 탭은 분석에서 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 직접 입력 필드를 지정하려는 경우 이 옵션을 선택한 후 입력 필드를 선택하십시오. 이 옵션을 사용하는 것은 **입력**에서 필드 역할을 유형 노드로 설정하는 것과 유사합니다.

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.[1]

일반

모델 이름. 특정 군집에 대한 스코어링 이후 생성되는 필드 이름입니다. **자동**(기본값)을 선택하거나 **사용자 정의**를 선택하고 이름을 입력하십시오.

군집 수. 생성할 군집 수를 지정합니다. 기본값은 **5**이고 최소값은 **2**입니다.

초기화

초기화 모드. 군집 중심 초기화를 위한 방법을 지정합니다. **K-평균||**가 기본값입니다. 이러한 두 가지 방법에 대한 세부사항은 확장 가능한 K-평균++를 참조하십시오.[2]

초기화 단계. K-평균|| 초기화 모드가 선택되면, 초기화 단계 수를 지정하십시오. **2**가 기본값입니다.

고급

고급 설정. 다음과 같이 고급 옵션을 설정하려는 경우 이 옵션을 선택하십시오.

최대 반복. 군집 중심을 검색할 때 수행할 최대반복수를 지정하십시오. **20**이 기본값입니다.

허용치. 반복 알고리즘의 수렴허용치를 지정하십시오. **1.0E-4**가 기본값입니다.

난수 시작값 설정. 난수 생성기에 사용될 시드를 생성하려면 이 옵션을 선택한 후 **생성**을 클릭하십시오.

표시

그래프 표시. 출력에 그래프를 포함시키려는 경우 이 옵션을 선택하십시오.

다음 테이블은 SPSS Modeler K-평균-AS 노드의 설정과 K-평균 Spark 매개변수 사이의 관계를 표시합니다.

표 *13. Spark* 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	K-평균 SparkML 매개변수
입력 필드	features	
군집 수	clustersNum	k
초기화 모드	initMode	initMode
초기화 단계	initSteps	initSteps
최대 반복	maxIter	maxIter
허용치	toleration	tol
난수 시드	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

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 고수익 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.

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 83 페이지의 『자동 군집 노드』의 내용을 참조하십시오.

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 280 페이지의 『모델 요약 보기』의 내용을 참조하십시오.
- 군집. 자세한 정보는 280 페이지의 『군집 보기』의 내용을 참조하십시오.

4개의 링크/보조 보기가 있습니다.

- 예측변수 중요도. 자세한 정보는 282 페이지의 『군집 예측변수 중요도 보기』의 내용을 참조하십시오.
- 군집 크기(기본값). 자세한 정보는 282 페이지의 『군집 크기 보기』의 내용을 참조하십시오.
- 셀 분포. 자세한 정보는 282 페이지의 『셀 분포 보기』의 내용을 참조하십시오.
- 군집 비교. 자세한 정보는 282 페이지의 『군집 비교 보기』의 내용을 참조하십시오.

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. 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 군집 알고리즘입니다(예: "이단계").
- **입력 변수.** 필드 수이며 **입력** 또는 **예측변수**라고도 합니다.
- **군집.** 솔루션의 군집 수입니다.

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 내용을 설명하는 레이블을 입력하십시오(예: "고급 승용차 구매자").
- **설명.** 군집 내용에 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 예측변수이며 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 전치』의 내용을 참조하십시오.
- 변수 정렬. 자세한 정보는 『변수 정렬』의 내용을 참조하십시오.
- 군집 정렬. 자세한 정보는 『군집 정렬』의 내용을 참조하십시오.
- 셀 내용 선택. 자세한 정보는 『셀 내용』의 내용을 참조하십시오.

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.

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.

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.

셀 내용: **셀** 단추를 사용하여 변수 및 평가 필드에 대한 셀 내용 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.

군집 예측변수 중요도 보기

예측변수 중요도 보기는 모델을 추정할 때 각 필드의 상대적 중요도를 표시합니다.

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.

셀 분포 보기

셀 분포 보기는 군집 기본 패널의 테이블에서 선택하는 변수 셀에 대해 확장되어 보다 자세한 데이터 분포의 도표를 표시합니다.

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.

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 열어서 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 280 페이지의 『군집 보기』의 내용을 참조하십시오.

표 *14.* 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 목표인 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 285 페이지의 『군집 모델에서 그래프 생성』의 내용을 참조하십시오.

군집 보기 표시 제어

기본 패널에서 군집 보기에 표시되는 내용을 제어하려면 **표시** 단추를 클릭하십시오. 그러면 표시 대화상자가 열립니다.

변수. 기본적으로 선택되어 있습니다. 모든 입력 변수를 숨기려면 확인 상자를 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 확인 상자를 사용할 수 없습니다.

군집 설명. 기본적으로 선택되어 있습니다. 모든 군집 설명 셀을 숨기려면 확인 상자를 선택 취소하십시오.

군집 크기. 기본적으로 선택되어 있습니다. 모든 군집 크기 셀을 숨기려면 확인 상자를 선택 취소하십시오.

최대 범주 수. 범주형 변수의 차트에 표시할 최대 범주 수를 지정하며 기본값은 20입니다.

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. 그래프 보드 기본 탭이 표시됩니다.

   참고: 기본 및 세부사항 탭은 그래프 보드를 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.

제 12 장 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 IBM SPSS Modeler의 웹 노드와 같은 시각화 기법을 사용하여 수동으로 찾을 수 있는 연관을 자동으로 찾습니다.

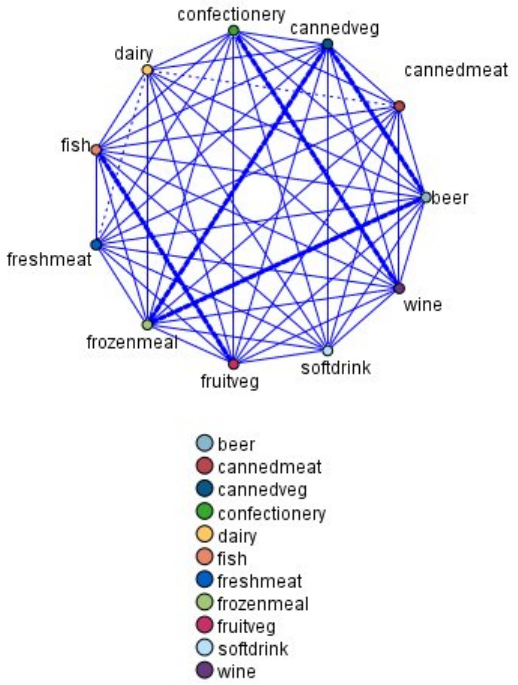


그림 *45.* 장바구니 항목 간의 연관을 보여주는 웹 노드

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 프로

세스는 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 체계를 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.

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 훈련 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 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.

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 미러링해야 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

고객	구매
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.

Apriori 노드

Apriori 노드는 데이터의 연관 규칙도 검색합니다. Apriori는 규칙을 선택하는 다섯 가지 다른 방법을 제공하고 정교한 색인화 기법을 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 288 페이지의 『테이블 대 트랜잭션 데이터』의 내용을 참조하십시오.

강도. 큰 문제가 있을 경우 Apriori는 일반적으로 더 빠르게 훈련합니다. .. 보유 가능한 규칙 수에 대한 임의의 한계도 없고 최대 32개의 전제조건이 있는 규칙을 핸들할 수 있습니다. Apriori는 다섯 가지 다른 훈련 방법을 제공해서 보다 탄력적으로 당면한 문제에 데이터 마이닝 방법을 부합시킵니다.

Apriori 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

최소 전항 지원. 규칙 세트의 규칙을 유지하기 위한 지원 기준을 지정할 수 있습니다. **지원**은 전항(규칙의 "if" 파트)이 참인 훈련 데이터의 레코드 퍼센트를 말합니다. (이 지원 정의는 CARMA 및 시퀀스 노드에 사용되는 것과 차이가 있습니다. 자세한 정보는 307 페이지의 『시퀀스 노드 모델 옵션』 주제를 참조하십시오.) 매우 작은 데이터 서브세트에 적용되는 규칙을 사용하는 경우 이 설정을 늘려 보십시오.

참고: Apriori에 대한 지원 정의는 전항이 있는 레코드 수를 기준으로 합니다. 그리고 지원 정의가 규칙의 모든 항목(즉, 전항과 후항 모두)이 있는 레코드 수를 기준으로 하는 CARMA 및 시퀀스 알고리즘과 상반됩니다. 연관 모델의 결과는 (전항) 지원 및 규칙 지원 측도를 모두 표시합니다.

최소 규칙 신뢰도. 신뢰도 기준을 지정할 수도 있습니다. **신뢰도**는 규칙의 전항이 참인 레코드를 기준으로 하며 후항도 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 전항 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 288 페이지의 『테이블 대 트랜잭션 데이터』의 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 스필링을 사용하지 않게 하려면 **속도**를 선택하십시오.
- 적합한 시기에 속도가 느려지더라도 알고리즘이 디스크 스필링을 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 안내서를 참조하십시오.

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 측도. Apriori는 잠재적 규칙을 평가하는 다섯 가지 방법을 지원합니다.

- **규칙 신뢰도.** 기본 방법은 규칙의 신뢰도(또는 정확도)를 사용하여 규칙을 평가합니다. 이 측도의 경우 **평가 측도 하한**이 모델 탭의 **최소 규칙 신뢰도** 옵션과 중복되므로 사용되지 않습니다. 자세한 정보는 289 페이지의 『Apriori 노드 모델 옵션』의 내용을 참조하십시오.
- **신뢰도 차이.** (**사전 신뢰도에 대한 절대 신뢰도 차이**라고도 합니다.) 이 평가 측도는 규칙의 신뢰도와 사전 신뢰도 간 절대차입니다. 이 옵션은 결과가 고르게 분포되지 않는 편향을 방지합니다. 이를 통해 "명백한" 규칙이 유지되지 않게 합니다. 예를 들어, 고객의 80%가 가장 인기 있는 제품을 구매하려는 경우가 있습니다. 정확도 85%의 인기 있는 제품 구매를 예측하는 규칙은 85% 정확도가 절대 척도에서 상당히 좋아 보이지만 사용자에게 많은 정보를 제공하지 않습니다. 규칙을 유지하려는 신뢰도의 최소 차이로 평가 측도 하한을 설정하십시오.
- **신뢰도 비율.** (**1에 대한 신뢰 지수 차이**라고도 합니다.) 이 평가 측도는 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. 예를 들어, 1% 환자에게만 발생하는 드문 질병이 있다고 가정하십시오. 절대 척도에서 10% 정확도가 매우 인상적일 것 같지는 않아도 이 조건을 10% 정도 예측할 수 있는 규칙은 어림짐작에 비하면 큰 향상입니다. 규칙을 유지하려는 차이로 평가 측도 하한을 설정하십시오.
- **정보 차이.** (**사전 확률에 대한 정보 차이**라고도 합니다.) 이 측도는 **정보 이득** 측도를 기준으로 합니다. 특정 후항의 확률이 논리 값(**비트**)으로 간주되면 정보 이득은 전항을 기준으로 하여 판별할 수 있는 이 비트의 비율입니다. 정보 차이는 전항이 주어진 정보 이득과 후항의 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 측도 하한을 설정하십시오.

  참고: 이 측도의 척도는 다른 척도에 비해 다소 덜 직관적이므로 만족스러운 규칙 세트를 얻기 위해 다른 하한으로 실험해야 합 수 있습니다.
- **정규화 카이제곱.** (**정규화 카이제곱 측도**라고도 합니다.) 이 측도는 전항과 후항 간 연관의 통계 지수입니다. 측도는 0과 1 사이의 값을 사용하도록 정규화되어 있습니다. 이 측도는 정보 차이 측도보다 훨씬 더 지원에 종속적입니다. 규칙을 유지하려는 정보 차이로 평가 측도 하한을 설정하십시오.

  참고: 정보 차이 측도의 경우 이 측도의 척도는 다른 척도에 비해 다소 덜 직관적이므로 만족스러운 결과 세트를 얻기 위해 다른 하한으로 실험해야 할 수 있습니다.

전항 없는 규칙 허용. 후항(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, cannedveg는 *cannedveg* 구매가 데이터의 공통 발생임을 표시하는 전항이 없는 단일 항목 규칙입니다. 일부 경우 가장 확실한 예측에만 관심이 있으면 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.

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 목표 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 다로 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 288 페이지의 『테이블 대 트랜잭션 데이터』의 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 40 페이지의 『모델 너깃』의 내용을 참조하십시오.

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 288 페이지의 『테이블 대 트랜잭션 데이터』의 내용을 참조하십시오.

표 형식 데이터

트랜잭션 형식 사용을 선택하지 않을 경우 다음 필드가 표시됩니다.

- **입력.** 하나 이상의 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 입력으로 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 검정함으로써 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 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.

CARMA 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 나눠서 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.

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 후항 값이 있는 규칙 제외. "two-headed" 후항 즉, 두 개의 항목을 포함한 후항을 제외하도록 선택합니다. 예를 들어, bread & cheese & fish -> wine&fruit 규칙은 two-headed 후항, 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 수**의 값을 입력하여 지원 수준을 얼마나 빨리 감소시켜야 하는지 지정하십시오.

전항 없는 규칙 허용. 후항(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 전항이 없는 단일 항목 규칙입니다. 일부 경우 가장 확실한 예측에만 관심이 있으면 이러한 규칙을 포함시키려 할 수 있습니다. 이 옵션은 기본적으로 선택되지 않습니다.

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.

모델 스코어링

세분화된 모델 너깃(Apriori, CARMA, 시퀀스)이 스트림에 추가되고 스코어링에 사용될 수 있습니다. 자세한 정보는 52 페이지의 『스트림에서 모델 너깃 사용』의 내용을 참조하십시오. 스코어링에 사용되는 모델 너깃은 각 대화 상자마다 추가 설정 탭을 포함합니다. 자세한 정보는 299 페이지의 『연관 규칙 모델 너깃 설정』의 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 301 페이지의 『연관 모델 너깃에서 규칙 세트 생성』의 내용을 참조하십시오.

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 *15.* 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
- 후항
- 리프트
- 배포성

표시/숨기기 메뉴. 표시/숨기기 메뉴(기준 도구 모음 단추)는 규칙 표시 옵션을 제어합니다.

그림 *46.* 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 전항과 같은 추가 규칙 정보를 나중에 병합할 수도 있습니다.
- **인스턴스**는 규칙이 적용되는 즉, 전항이 참인 고유 ID 수에 대한 정보를 표시합니다. 예를 들어, bread -> cheese 규칙이 주어진 경우 전항 *bread*를 포함한 훈련 데이터의 레코드 수를 **인스턴스**라 부릅니다.
- **지원**은 전항 지원 즉, 훈련 데이터를 기준으로 하여 전항이 참인 ID의 비율을 표시합니다. 예를 들어, 훈련 데이터의 50%가 bread 구매를 포함하면, 규칙 bread -> cheese의 전항 지원은 50%입니다. 참고: 여기에 정의된 지원은 인스턴스와 동일하지만 퍼센트로 표현됩니다.
- **신뢰도**는 전항 지원에 대한 규칙 지원 비율을 표시합니다. 후항도 참인 전항이 지정된 ID의 비율을 표시합니다. 예를 들어, 훈련 데이터의 50%가 bread를 포함(전항 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, 전항, 후항이 참인 ID의 비율을 표시합니다. 예를 들어, 훈련 데이터의 20%가 bread와 cheese를 모두 포함하면 규칙 bread -> cheese의 규칙 지원은 20%입니다.
- **리프트**는 후항이 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 = 2입니다. 다른 규칙에 사람들이 11% 신뢰도로 빵을 살 것이라 지정되면 규칙의 리프트는 1에 근접합니다. 이는 전항이 있다고 해서 후항이 있을 확률이 많이 차이가 나지 않음을 의미합니다. 일반적으로 리프트가 1이 아닌 규칙이 리프트가 1에 가까운 규칙보다 흥미롭습니다.
- **배포성**은 훈련 데이터가 전항 조건을 충족시키지만 후항 조건을 충족시키지 않는 퍼센트 측도입니다. 제품 구매 조건에서, 이는 기본적으로 총 고객 기반이 전항을 소유하지만(또는 구매했지만) 후항을 아직 구매하지 않은 퍼센트를 의미합니다. 배포성 통계는 ((Antecedent Support in # of Records - Rule Support in # of Records) / Number of Records) * 100으로 정의되며 여기서, *Antecedent Support*는 전항이 참인 레코드 수를 의미하고 *Rule Support*는 전항과 후항이 모두 참인 레코드 수를 의미합니다.

필터 단추. 메뉴의 필터 단추(깔때기 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 *47.* 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 서브세트를 포함한 모델을 생성함에 유의하십시오. 자세한 정보는 아래 298 페이지의 『규칙의 필터 지정』의 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 *48.* 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 매기기**를 선택하고 시작 번호를 지정하십시오.

자세한 정보는 301 페이지의 『필터링된 모델 생성』의 내용을 참조하십시오.

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 전항과 후항이 보다 분명히 표시되도록 규칙을 필터링할 것을 고려해야 합니다. 규칙 브라우저의 모델 탭에서 필터링 옵션을 사용하여 필터 조건을 지정할 대화 상자를 열 수 있습니다.

후항. 지정된 후항의 포함 또는 제외를 기준으로 한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. 규칙에 최소 하나의 지정된 후항이 포함되는 필터를 작성하려면 **하나 이상 포함**을 선택하십시오. 또는 **제외**를 선택하여 지정된 후항을 제외하는 필터를 작성하십시오. 목록 상자 오른쪽에 있는 선택도구 아이콘을 사용하여 후항을 선택할 수 있습니다. 그러면 생성된 규칙에 있는 모든 후항이 나열된 대화 상자가 열립니다.

참고: 후항은 둘 이상의 항목을 포함할 수 있습니다. 필터는 지정된 항목 중 하나가 후항에 포함되었는지만 검사합니다.

전항. 지정된 전항의 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 전항이 나열된 대화 상자가 열립니다.

- 지정된 모든 전항을 규칙에 포함해야 포함 필터로 필터를 설정하려면 **모두 포함**을 선택하십시오.
- 규칙에 최소 하나의 지정된 전항이 포함되는 필터를 작성하려면 **하나 이상 포함**을 선택하십시오.
- 지정된 전항이 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.

전항 지원. 규칙의 전항 지원의 수준을 기준으로 한 규칙 필터링의 옵션을 활성화하려면 **필터 사용**을 선택하십시오. 전항 지원은 인기 지수와 유사하도록, 현재 규칙과 동일한 전항을 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.

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. 그래프 보드 기본 탭이 표시됩니다.

   참고: 기본 및 세부사항 탭은 그래프 보드를 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.

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 301 페이지의 『연관 모델 너깃에서 규칙 세트 생성』의 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)
- 리프트
- 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 후항(bread & cheese & fruit -> wine & pate)이 있는 규칙은 모든 후항(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 전항이 장바구니의 컨텐츠와 정확히 일치하도록 **Ignore unmatched basket items**를 선택 취소해야 합니다. 기본적으로 일치하지 않는 항목은 무시됩니다.

예측이 장바구니에 없는지 검사. 장바구니에 후항도 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 후항이 이미 존재하는 규칙이 유용할 수 있습니다. 후자의 경우 가장 유용한 옵션은 아래의 **장바구니에서 예측을 검사하지 않음**일 수 있습니다.

장바구니에 예측이 있는지 검사 장바구니에 후항도 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 후항이 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 페치하여 SPSS Modeler에서 스코어를 계산합니다.
- **데이터베이스 외부 스코어** 선택할 경우 이 옵션은 데이터베이스에서 데이터를 다시 페치하여 SPSS Modeler에서 스코어를 계산합니다.

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.

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 56 페이지의 『세분화되지 않은 모델』의 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.

목표 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 목표 필드의 기본값을 지정할 수 있습니다.

필터링된 모델 생성

연관 모델 너깃(예: Apriori, CARMA, 또는 시퀀스 규칙 세트 노드)에서 필터링된 모델을 생성하려면 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.

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 288 페이지의 『테이블 대 트랜잭션 데이터』의 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 *16.* 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 *17.* 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 *18.* 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 *19.* 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.

연관 모델 배포

연관 모델을 스코어링할 때 예측 및 신뢰도는 별도의 열에 출력됩니다(여기서, *P*는 예측을 나타내고, *C*는 신뢰도, *I*는 규칙 ID를 나타냄). 이는 입력 데이터가 표 또는 트랜잭션 형식인 경우입니다. 자세한 정보는 302 페이지의 『연관 규칙 스코어링』의 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.

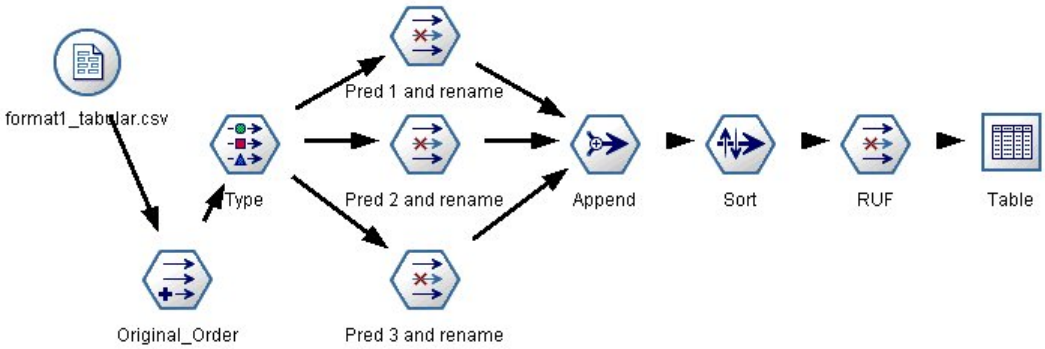


그림 *49.* 표 형식 데이터를 *till-roll* 형식으로 전치하는 데 사용되는 예 스트림

1. 파생 노드의 `@INDEX` 함수를 사용하여 현재 예측 순서를 확인하고 *Original_order*와 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.

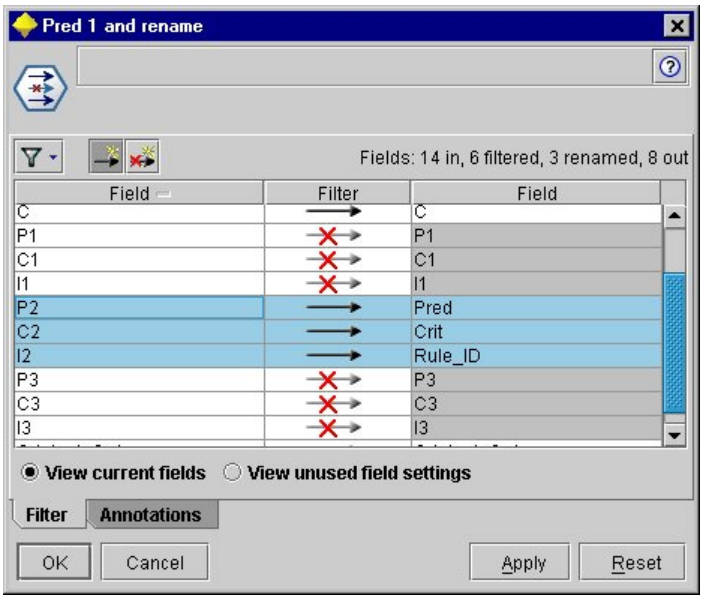


그림 *50.* 예측 *2*의 필드 이름 변경 중 예측 *1* 및 *3*의 필드 필터링.

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.

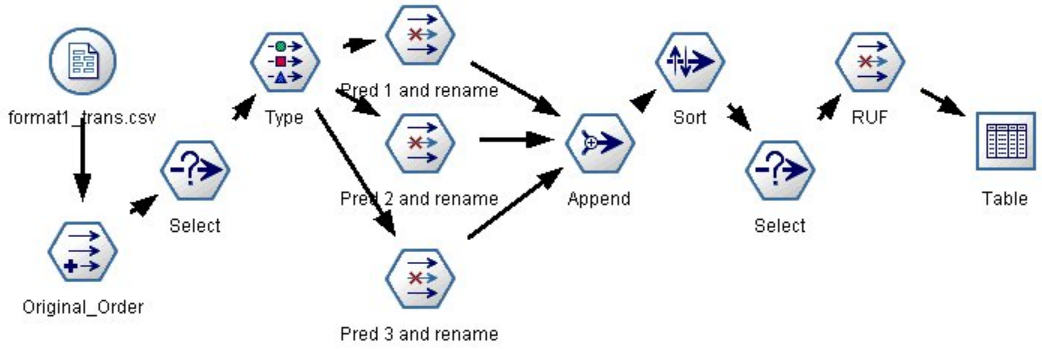


그림 *51.* 트랜잭션 데이터를 *till-roll* 형식으로 전치하는 데 사용되는 예 스트림

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 CLEM 표현식을 사용하여 레코드를 선택합니다. `ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`.
- 두 번째 선택 노드는 `Rule_ID`의 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.

시퀀스 노드

시퀀스 노드는 `bread -> cheese` 형식으로 순차 또는 시간 중심의 데이터에서 패턴을 검색합니다. 시퀀스의 요소는 단일 트랜잭션을 구성하는 **항목 세트**입니다. 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. **시퀀스**는 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 내용 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 지정하

지 않은 경우 시퀀스 노드는 시간 값으로 행 번호를 사용하여 함축된 시간소인을 사용합니다. 내용 필드는 임의의 측정 수준 및 역할을 보유할 수 있지만, 모든 내용 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 수퍼 노드를 생성할 수 있습니다.

시퀀스 노드 필드 옵션

시퀀스 노드를 실행하기 전에 시퀀스 노드의 필드 탭에서 ID 및 내용 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).

내용 필드. 모델의 내용 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 288 페이지의 『테이블 대 트랜잭션 데이터』의 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 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.)

시퀀스 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 지원은 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. 신뢰도는 규칙에서 예측을 수행하는 모든 ID 중 올바른 예측에 성공한 ID의 퍼센트를 나타냅니다. 이는 전체 시퀀스를 찾은 ID 수를 훈련 데이터에 기반하여 전항을 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 309 페이지의 『시퀀스 모델 너깃』의 내용을 참조하십시오.

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 *20.* 트랜잭션 목록 예.

ID	시간	내용
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 -> bread`와 같은 시퀀스는 확인할 수 없습니다. `apples` 및 `bread` 사이의 차이가 최소 차이보다 작기 때문입니다. 마찬가지로, 다음과 같은 대체 데이터를 고려하십시오.

표 *21.* 트랜잭션 목록 예.

ID	시간	내용
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.

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 312 페이지의 『시퀀스 모델 너깃 설정』의 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 예측 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 312 페이지의 『시퀀스 모델 너깃 설정』의 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 302 페이지의 『연관 규칙 스코어링』의 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 합니

다. 그러나 예측은 추가된 레코드의 시간소인에 연결되지 않습니다. 단순히 현재 레코드까지 현재 ID의 트랜잭션 히스토리가 주어진 경우 미래의 특정 포인트에 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.

표 *22.* 예 레코드.

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 -> Bread`는 계속 적용됩니다.

새 노드 생성

생성 메뉴에서는 시퀀스 모델에 기반하여 새 수퍼 노드를 작성할 수 있습니다.

- **규칙 수퍼 노드.** 스코어가 계산된 데이터에서 시퀀스의 발생을 발견하고 계산할 수 있는 수퍼 노드를 작성합니다. 규칙이 선택되지 않으면 이 옵션은 사용할 수 없습니다. 자세한 정보는 312 페이지의 『시퀀스 모델 너깃에서 규칙 수퍼 노드 생성』의 내용을 참조하십시오.
- **모델을 팔레트로.** 모델을 모델 팔레트로 리턴합니다. 이 옵션은 동료가 모델 자체가 아닌 모델을 포함한 스트림을 보냈을 경우에 유용합니다.

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.

표 *23.* 규칙 형식

전항	후항
beer and cannedveg	beer
fish fish	fish

첫 번째 규칙 예는 동일한 트랜잭션에서 *"beer"* 및 *"cannedveg"*를 포함하는 *ID*의 경우 *"beer"*가 뒤에 나올 수 있습니다. 두 번째 규칙 예는 한 트랜잭션에서 *"fish"*를 포함했고, 다른 트랜잭션에서도 *"fish"*를 포함하는 *ID*의 경우 다음에도 *"fish"*가 나올 수 있음으로 해석됩니다. 첫 번째 규칙에서 *beer* 및 *cannedveg*는 동시에 구매했고, 두 번째 규칙에서 *fish*는 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 %
- 후항
- 첫 번째 전항
- 마지막 전항
- 항목 수(전항)

예를 들어 다음 표에서는 항목 수를 내림차순으로 정렬합니다. 전항 세트에 여러 항목을 포함하는 규칙은 항목이 더 적은 규칙보다 앞에 옵니다.

표 *24.* 항목 수로 정렬된 규칙

전항	후항
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(전항 및 후항 모두)가 나타나는 고유 ID의 번호에 대한 정보를 표시합니다. (이는 인스턴스 번호가 전항만 적용하는 ID의 번호를 참조하는 연관 모델과는 다름에 유의하십시오.) 예를 들어, 규칙이 bread -> cheese인 경우 *bread* 및 *cheese*를 모두 포함하는 훈련 데이터의 ID 번호는 **인스턴스**로 참조됩니다.
- **지원**은 전항이 참인 훈련 데이터에서 ID의 비율을 표시합니다. 예를 들어, 훈련 데이터 중 50%가 전항 *bread* 를 포함하면 bread -> cheese 규칙에 대한 지원은 50%입니다. (전항 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 훈련 데이터에 기반하여 전항을 찾은 ID 수로 나누어 계산합니

다. 예를 들어, 훈련 데이터의 50%가 cannedveg를 포함하지만(전항 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 훈련 데이터의 비율을 표시합니다. 예를 들어, 훈련 데이터 중 20%가 *bread* 및 *cheese*를 모두 포함하는 경우 규칙 bread -> cheese의 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(깔때기 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 *52.* 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 서브세트를 포함한 모델을 생성함에 유의하십시오. 자세한 정보는 아래 298 페이지의 『규칙의 필터 지정』의 내용을 참조하십시오.

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.

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 45 페이지의 『모델 너깃 찾아보기』의 내용을 참조하십시오.

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.

연관 규칙 노드

연관 규칙은 다음 양식의 명령문입니다.

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 풀아웃합니다. 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.

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 예측변수)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, 예측변수 및 기타 역할을 지정하려면 이 옵션을 선택하십시오.

필드 **사용자 정의 필드 할당 사용**을 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("후항"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("전항"이라고도 함)으로만 표시될 수 있습니다. 이 목록에 필드가 존재하는 것은 필드가 임의 규칙에서 사용됨을 의미하지는 않습니다. 단지 사용되는 경우 조건만 가능합니다.

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 288 페이지의 『테이블 대 트랜잭션 데이터』의 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.

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 나눠집니다. 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 길이

를 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.

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 후항 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 후항 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 전항 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 대화형이며 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.

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 후항 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 측도.

반복 예상값 허용 스코어링 동안 동일한 예상값을 갖는 여러 규칙을 포함하려면 이 확인 상자를 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.

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 52 페이지의 『스트림에서 모델 너깃 사용』의 내용을 참조하십시오. 스코어링에 사용되는 모델 너깃은 각 대화 상자마다 추가 설정 탭을 포함합니다. 자세한 정보는 320 페이지의 『연관 규칙 모델 너깃 설정』의 내용을 참조하십시오.

연관 규칙 모델 너깃 세부사항

연관 규칙 모델 너깃은 출력 뷰어의 모델 탭에 모델 세부사항을 표시합니다. 뷰어 사용에 대한 자세한 정보는 모델러 사용자 안내서(ModelerUsersGuide.pdf)에서 "출력에 대한 작업" 절을 참조하십시오.

GSAR 모델링 작업은 다음 표에 표시된 것처럼 접두문자 $A로 많은 새 필드를 작성합니다.

표 *25.* 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).

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 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. 다음 기준에서 선택할 수 있습니다.

- **신뢰도**
- **규칙 지원**
- **리프트**
- **조건 지원**
- **배포성**

반복 예상값 허용 스코어링 시 후항이 같은 여러 규칙을 포함하려면 이 확인 상자를 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 확인 상자를 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 스코어링**이 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.

제 13 장 시계열 모델

왜 예측인가?

예측하는 것은 시간 경과에 따라 하나 이상의 계열의 값을 예측하는 것을 의미합니다. 예를 들어, 제조 또는 분포에 대해 자원을 할당하기 위해 제품 또는 서비스 라인에 대한 예상 요구를 예측할 수 있습니다. 계획 의사결정에 구현에는 시간이 소요되므로, 예측은 많은 계획 프로세스에서 중요한 도구입니다.

모델링 시계열의 메소드에서는 히스토리가 자체를 반복한다고 가정합니다. 정확하지 않으면, 과거 훈련으로 나중에 더 나은 결정을 할 수 있습니다. 예를 들어 내년 판매를 예측하기 위해, 올해 판매를 보는 것에서 시작하여 최근 몇 년 동안 어떤 추세 또는 패턴이 개발되었는지 보기 위해 거꾸로 작업합니다. 그러나 패턴은 측정이 어려울 수 있습니다. 예를 들어 한 행에서 몇 주에 걸쳐 증가할 경우, 이것이 계절 순환의 일부인지 또는 장기 추세의 시작인지 하는 것이 쉽지 않습니다.

통계 모델링 기법을 사용하여, 과거 데이터의 패턴을 분석하고 해당 패턴을 투영하여 계열의 미래 값이 속하게 될 범위를 판별할 수 있습니다. 결과는 사용자 의사결정의 기초가 되는 한층 정확한 예측값입니다.

시계열 데이터

시계열은 보통의 구간으로 측정되는 순서화된 측정 콜렉션입니다(예: 매일 주가 또는 매주 판매 데이터). 측정은 관심이 있는 어떤 것도 가능하며, 각 계열은 보통 다음 중 하나로 분류됩니다.

- **종속변수.** 예측하려는 계열.
- **예측변수.** 목표 설명에 도움이 될 수 있는 계열(예: 광고 예산을 사용한 판매 예측). 예측변수는 ARIMA 모델에만 사용할 수 있습니다.
- **이벤트.** 예측 가능한 반복 발생 인시던트(예: 판매 프로모션) 설명에 사용되는 특수 예측변수 계열.
- **개입.** 일회성 인시던트(예: 정전 또는 사원 파업) 설명에 사용되는 특수 예측변수 계열.

구간은 시간 단위를 나타낼 수 있지만, 모든 측정에 동일해야 합니다. 또한 어떤 측정도 없는 구간은 결측값으로 설정되어야 합니다. 따라서, 측정이 있는 구간 수(결측값이 있는 구간을 포함하여)는 데이터의 기록 범위 시간 길이를 정의합니다.

시계열의 공정특성 변수

계열의 과거 작동을 훈련하면 패턴을 식별하고 더 좋은 예측을 하는데 도움이 됩니다. 도표화될 때, 많은 시계열은 다음 기능 중 하나 이상을 보여줍니다.

- 추세
- 계절 및 비계절 순환
- 펄스 및 단계

- 이상값

추세

추세는 시간이 경과하면서 증가하거나 감소할 계열 값의 추세 또는 계열 수준에서 점증적인 상향 또는 하향 이동입니다.

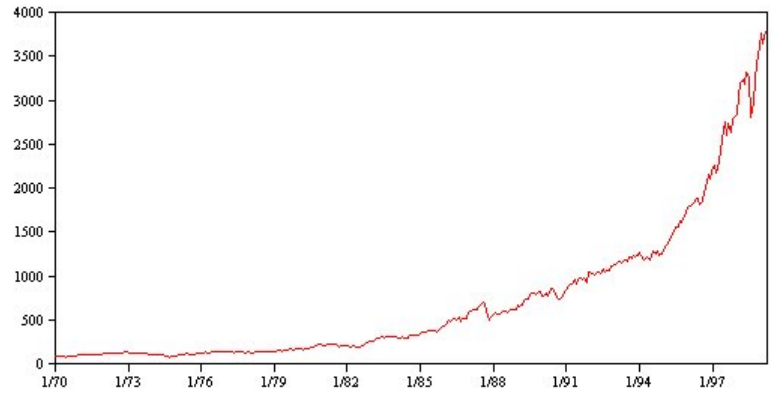


그림 *53.* 추세

추세는 **로컬** 또는 **글로벌**이지만, 단일 계열이 두 유형 모두를 나타낼 수 있습니다. 히스토리에서, 증권 거래소 지수의 계열 도표는 상향 글로벌 추세를 보여줍니다. 로컬 하향 추세는 불경기에 나타내고, 로컬 상향 추세는 경기 좋을 때 나타났습니다.

추세는 **선형** 또는 **비선형**이 될 수도 있습니다. 선형 추세는 주성분에 대한 단리의 효과에 비교할 만한, 계열 수준에 대한 양(+) 또는 음(-) 가법 증분입니다. 비선형 추세는 종종 승법 추세로, 증분은 이전 계열 값에 비례합니다.

글로벌 선형 추세는 적합하며 지수평활 및 ARIMA 모델 둘 다에 의해 잘 예측됩니다. ARIMA 모델 작성 시, 추세를 보여주는 계열은 일반적으로 추세 효과를 제거하기 위해 차별화됩니다.

계절 순환

계절 순환은 계열 값에 반복적 예측가능 패턴입니다.

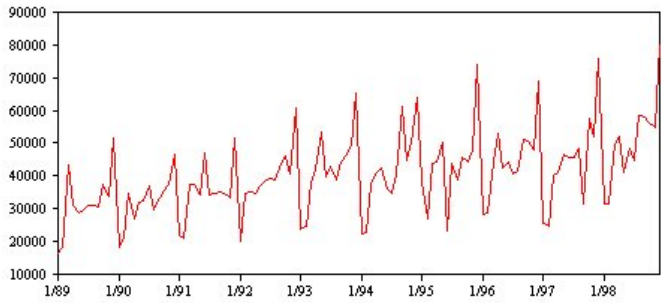


그림 *54.* 계절 순환

계절 순환은 사용자 계열의 구간과 연결됩니다. 예를 들어, 매월 데이터는 일반적으로 분기 및 연도에 대해 순환됩니다. 매월 계열은 첫 번째 분기가 낮은 유의적 분기별 순환이나 모든 12월이 최대인 매년 순환을 보여줄 수 있습니다. 계절 순환을 보여주는 계열은 **계절성**을 드러낸다고 합니다.

계절 패턴은 좋은 적합 및 예측을 확보할 때 유용하며, 계절성을 캡처하는 지수평활 및 ARIMA 모델이 있습니다.

비계절 순환

비계절 순환은 계열 값에서 반복되는 예측할 수 없는 패턴입니다.

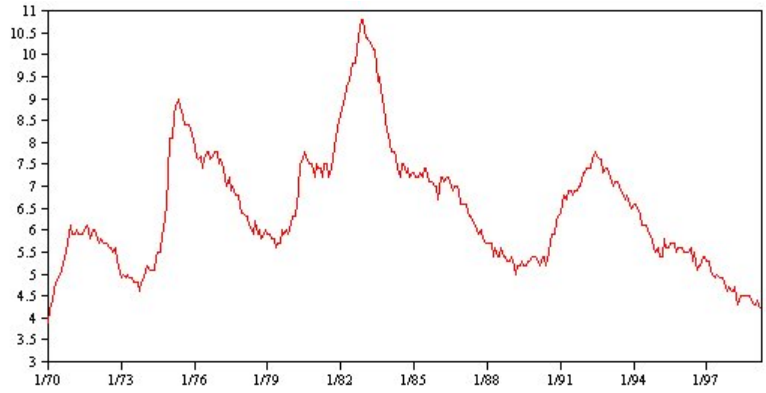


그림 *55.* 비계절 순환

일부 계열(예: 실업률)은 주기적 작동을 명확하게 표시합니다. 그러나 시간이 경과하면서 순환의 주기성은 변하여, 높거나 낮을 때를 예측하는 것이 어려워집니다. 다른 계열에 예측할 수 있는 순환이 있을 수 있지만, 그레고리 달력에 확실하게 맞지 않거나 1년보다 오랫동안의 순환이 있을 수 있습니다. 예를 들어, 조석은 음력에 따르고, 올림픽에 관련된 국제 여행 및 거래는 4년마다 불어나며, 그레고리오력 날짜가 연도마다 변경되는 종교 휴일이 많이 있습니다.

비계절 순환 패턴은 모델링하기에 어려우므로, 일반적으로 예측 시 불확실성이 증가합니다. 예를 들어, 증권 시장은 예측변수의 노력을 무시한 다양한 계열 인스턴스를 제공합니다. 비계절 패턴이 존재할 경우 동일한 모든 비계절 패턴을 고려해야 합니다. 많은 경우에, 합리적으로 잘 히스토리 데이터에 적합한 모델을 계속 식별할 수 있어서, 예측 시 불확실성을 최소화할 최상의 기회가 제공됩니다.

펄스 및 단계

계열이 많으면 수준에서 비약적인 변경사항이 발생합니다. 변경사항은 일반적으로 두 가지 유형으로 발생합니다.

- 계열 수준에서, 갑작스러운, 임시 이동 또는 **펄스**
- 계열 수준에서, 갑작스러운, 영구 이동 또는 **단계**

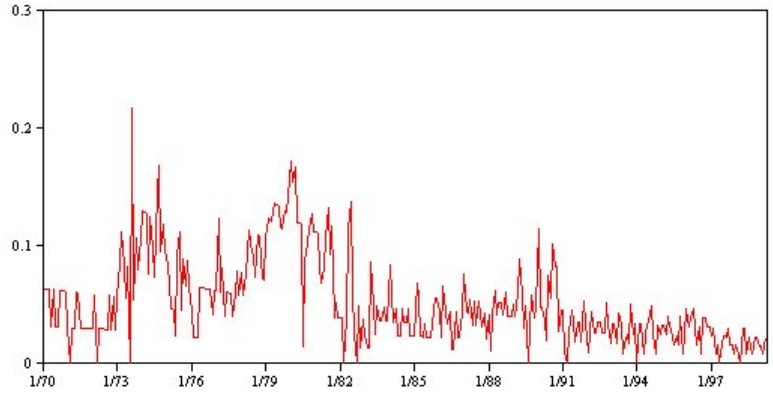


그림 *56.* 펄스가 있는 계열

단계 또는 펄스가 관찰되는 때, 타당한 설명을 찾는 것이 중요합니다. 시계열 모델은 점진적인(갑작스럽지 않은) 변경을 고려하도록 계획됩니다. 결과적으로, 이 모델은 펄스를 과소평가하고 단계에서 파멸되어, 나쁜 모델 적합과 불확실한 예측을 유도할 수 있습니다. (일부 계절성 인스턴스는 수준에서 갑작스런 변경을 나타낸 것처럼 보일 수 있지만, 수준이 한 계절 주기에서부터 다음 주기까지 일정합니다.)

교란을 설명할 수 있으면, **개입** 또는 **이벤트**를 사용하여 모델링할 수 있습니다. 예를 들어, 1973년 8월 동안, OPEC(Organization of Petroleum Exporting Countries)에 의해 부과된 석유 금수 조치는 물가 상승률에서 강한 변화를 야기시켰으며, 그 후 수개월에서 정상 수준으로 돌아갔습니다. 금수조치 달에 대해 **포인트 개입**을 지정하면, 모델의 적합도가 개선되어 예측이 간접적으로 향상될 수 있습니다. 예를 들어, 소매점에서 모든 항목이 50% 세일로 표시된 날에 평소보다 판매가 훨씬 높았다는 것을 발견할 수 있습니다. 되풀이 되는 **이벤트**로서 50% 할인 프로모션을 지정하여, 모델 적합을 개선하고 나중 날짜에 프로모션을 반복하는 효과를 예측할 수 있습니다.

이상값

설명할 수 없는 시계열 수준에서의 이동(Shift)을 **이상값**이라고 합니다. 이러한 관측값은 계열의 나머지와 일치하지 않고, 상당히 분석에 영향을 미칠 수 있어서, 결국 시계열 모델의 예측 기능에 영향을 미칠 수 있습니다.

다음 그림은 시계열에서 일반적으로 발생하는 이상값의 몇 가지 유형을 표시합니다. 파랑 선은 이상값 없이 계열을 나타냅니다. 빨강 선은 계열에 이상값을 포함한 경우가 있을 수 있는 패턴을 제안합니다. 이 이상값은 모두 **결정적**으로 분류됩니다. 계열의 평균 수준에만 영향을 주기 때문입니다.

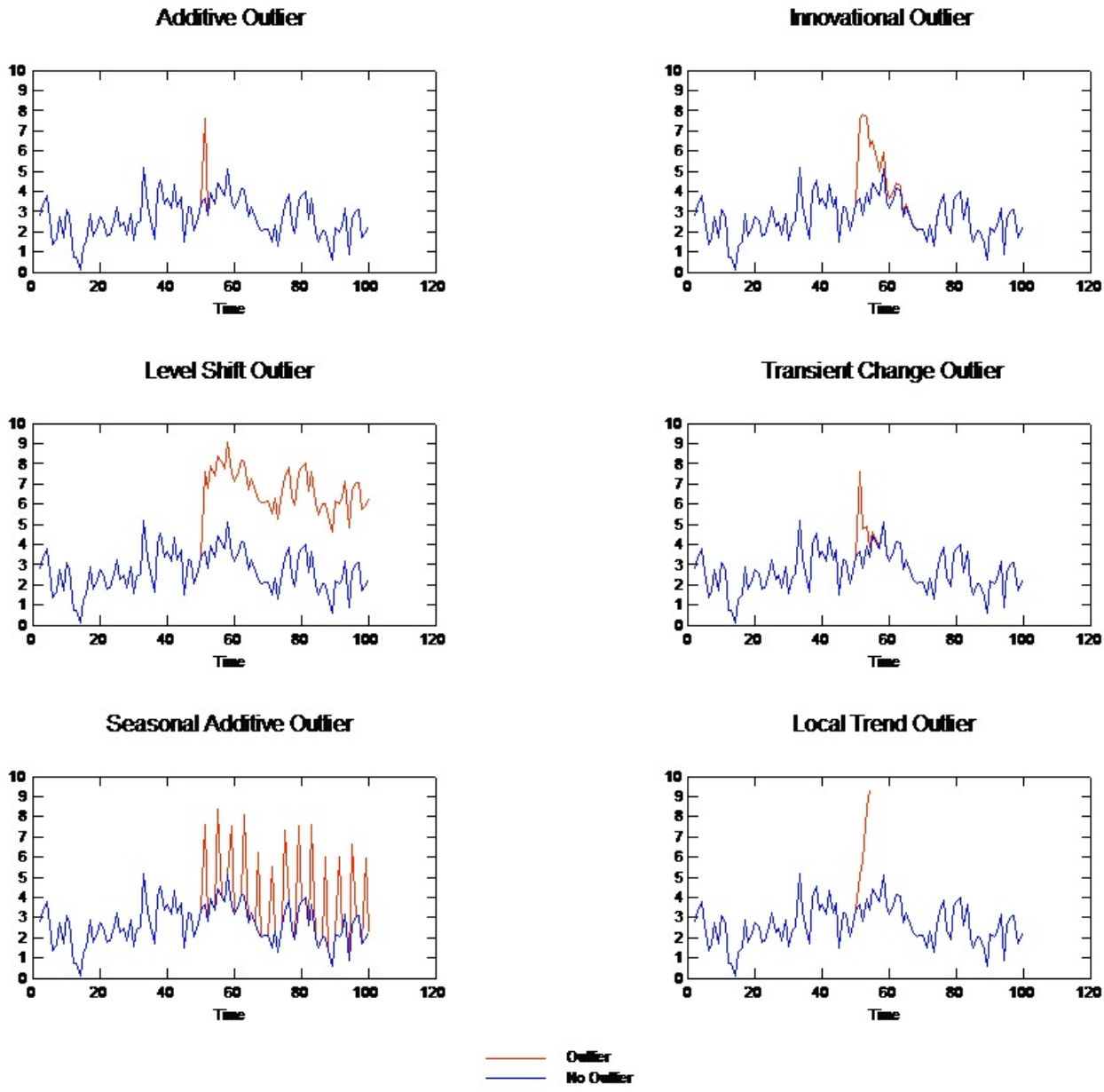


그림 *57.* 이상값 유형

- **가법 이상값.** 가법 이상값은 단일 관측값에 대해 발생하는 너무 크거나 너무 작은 값으로 나타납니다. 연속 관측값은 가법 이상값에 의해 영향을 받지 않습니다. 연속 가법 이상값은 일반적으로 **가법 이상값 패치**라고 합니다.
- **혁신적 이상값.** 혁신적 이상값은 효과가 후속 관측값에 머물고 있는 초기 영향에 의해 특징지어집니다. 이상값의 영향력은 시간이 진행되면서 증가할 수 있습니다.
- **수준 이동 이상값.** 수준 이동에 대해, 이상값 이후에 나타나는 모든 관측값은 새 수준으로 이동합니다. 가법 이상값과 반대로, 수준 이동 이상값은 많은 관측값에 영향을 주고 영구적인 효과를 갖습니다.

- **일시적 변경 이상값.** 일시적 변경 이상값은 수준 이동 이상값과 유사하지만 이상값 효과가 후속 관측값에서 기하급수적으로 감소합니다. 결국, 계열은 해당되는 정규 수준으로 돌아갑니다.
- **계정 가법 이상값.** 계절 가법 이상값은 정규적으로 반복해서 발생하는 너무 크거나 너무 작은 값으로 나타납니다.
- **국소적 추세 이상값.** 국소적 추세 이상값은 초기 이상값의 시작 후에 이상값에서 패턴에 의해 야기되는 계열의 일반적 이동을 발생시킵니다.

시계열에서의 이상값 발견에는 존재하는 이상값의 위치, 유형 및 크기 판별이 포함됩니다. Tsay(1988년)는 결정적인 이상값을 식별하기 위해 평균 수준 변경을 발견하기 위한 반복 프로시저를 제안했습니다. 이 프로세스에는 이상값을 포함하는 다른 모델에 대해, 이상값이 없는 것으로 가정하는 시계열 모델을 비교하는 과정이 포함됩니다. 모델 사이의 차이로 인해 지정된 포인트를 이상값으로 처리하는 효과의 추정값이 생성됩니다.

자기상관 및 편자기상관 함수

자기상관 및 편자기상관은 현재 및 지난 계열 값 사이의 연관성 측도이며 나중 값 예측에 가장 유용한 지난 계열 값을 표시합니다. 이러한 지식을 사용하여, ARIMA 모델에서 프로세스 순서를 판별할 수 있습니다. 좀더 구체적으로 말하면 다음과 같습니다.

- **자기상관 함수(ACF).** *k* 시차로, 이는 *k* 구간인 계열 값 사이의 상관관계입니다.
- **편자기상관 함수(PACF).** *k* 시차로, 이는 *k* 구간(사이의 구간 값 고려)인 계열 값 사이의 상관관계입니다.

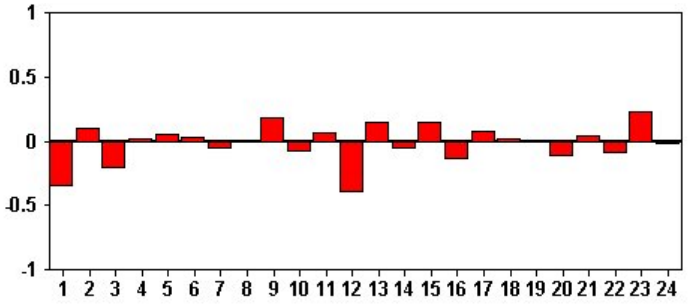


그림 *58.* 계열에 대한 *ACF* 도표

ACF 도표의 *x* 축은 자기상관이 계산되는 시차를 표시합니다. *y* 축은 상관관계의 값(–1과 1 사이)을 표시합니다. 예를 들어, ACF 도표의 시차 1 위치에 있는 말뚝표시는 각 계열 값과 이전 값 사이의 강한 상관관계를 표시하고, 시차 2에 있는 말뚝표시는 각 값과 이전에 두 포인트를 발생하는 값 사이의 강한 상관관계를 표시합니다.

- 양수 상관관계는 큰 현재 값이 지정된 시차에서 큰 값과 일치함을 표시하고, 음수 상관관계는 큰 현재 값이 지정된 시차에서 작은 값과 일치함을 표시합니다.
- 상관관계의 절대값은 연관의 강도 측정으로, 절대값이 클수록 관계가 더 강함을 표시합니다.

계열 변환

변환은 종종 모델을 추정하기 전에 계열을 안정시키는데 유용합니다. 이는 특히 모델 추정 전에 계열이 **정상**이어야 하는 ARIMA 모델에 중요합니다. 글로벌 수준(평균)과 수준(분산)으로부터의 평균 편차가 계열 전체적으로 일정한 경우 계열은 정상입니다.

가장 관심이 가는 계열이 정상이 아니어도, 자연로그, 차분 또는 계절 차분과 같은 변환을 적용하여 계열을 정상으로 만들 수 있으면 ARIMA은 효과적입니다.

분산 안정화 변환. 시간이 경과하면서 분산이 변경되는 계열은 종종 자연로그 또는 제곱근 변환을 사용하여 안정화될 수 있습니다. 이를 함수 변환이라고도 합니다.

- **자연로그.** 자연로그는 계열 값에 적용됩니다.
- **제곱근.** 제곱근 함수는 계열 값에 적용됩니다.

자연로그와 제곱근 변환은 음수값이 있는 계열에 사용할 수 없습니다.

수준 안정화 변환. ACF에서 값이 완만하게 감소하면 각 계열 값은 이전 값과 강하게 상관되어 있음을 나타냅니다. 계열 값에서 변경을 분석하여, 안정 수준을 확보합니다.

- **단순 차분.** 계열에서 각 값과 이전 값 사이의 차분이 계산됩니다(계열에서 가장 오래된 값 제외). 이는 차분화된 계열에는 원래 계열보다 1 작은 값이 있음을 의미합니다.
- **계절 차분.** 계열에서 각 값과 이전 계절 값 사이의 차분이 계산되는 것을 제외하고 단순 차분과 같습니다.

로그 또는 제곱근 변환을 사용하여 단순 또는 계절 차분이 동시에 사용중인 경우, 분산 안정화 변환이 항상 첫 번째로 적용됩니다. 단순 및 계절 차분 모두 사용 중인 경우, 결과 계열 값은 단순 차분 또는 계절 차분이 첫 번째로 적용되는지 여부에 관계없이 동일합니다.

예측변수 계열

예측변수 계열에는 예측할 계열의 작동을 설명하는 데 도움이 될 수 있는 관련 데이터가 포함됩니다. 예를 들어, 웹 기반 또는 카탈로그 기반 소매상은 메일을 보낸 카탈로그 수, 개통된 전화 회선 수 또는 회사 웹 페이지를 클릭한 횟수를 기반으로 판매를 예측할 수 있습니다.

계열이 예측하고자 하는 장래만큼 연장되고 결측값 없이 전체 데이터를 가지고 있는 경우 어떤 계열도 예측변수로 사용할 수 있습니다.

모델링할 예측변수를 추가할 때 주의하여 사용하십시오. 많은 수의 예측변수를 추가하면 모델 예측에 소요되는 시간이 증가합니다. 예측변수를 추가하면 히스토리 데이터가 적합하도록 하는 모델의 기능이 개선될 수 있지만, 모델이 더 나은 예측 작업을 수행함을 의미하는 것은 아니므로, 추가된 복잡도에 가치가 없을 수도 있습니다. 이상적으로, 목적은 좋은 예측 작업을 수행하는 가장 단순한 모델을 식별하는 것이어야 합니다.

일반 규칙과 같이, 예측변수 수는 표본 크기를 15로 나눈 값보다 적어야 합니다(최대, 15개 케이스마다 하나의 예측변수).

결측 데이터가 있는 예측변수. 불완전 또는 결측 데이터가 있는 예측 변수는 예측에 사용할 수 없습니다. 이는 히스토리 데이터와 나중 값 둘 다에 적용됩니다. 일부 경우에는, 모델 예측 시 가장 오래된 데이터를 제외하기 위해 모델 예측 범위를 설정하여 이러한 제한사항을 피할 수 있습니다.

STP(Spatio-Temporal Prediction) 모델링 노드

STP(Spatio-Temporal prediction)에는 건물 또는 시설에 대한 에너지 관리, 머신 설비 엔지니어를 위한 성능 분석 및 예측, 또는 대중 교통수단 계획과 같은 많은 잠재된 애플리케이션이 있습니다. 이러한 애플리케이션에서 에너지 사용량과 같은 측정에는 종종 공간과 시간이 소요됩니다. 이러한 측정 기록에 관련될 수 있는 질문으로는, 나중 관측값에 영향을 주게 될 요인은 무엇인가?, 원하는 변경에 영향을 주기 위해 수행할 수 있는 것은 무엇인가? 또는 시스템을 더 좋게 관리하기 위해 수행할 수 있는 것은 무엇인가? 등이 있습니다. 이러한 질문을 처리하기 위해, 다른 위치에서 나중 값을 예측할 수 있는 통계적인 기법을 사용할 수 있고, 가정(what-if) 분석을 수행하기 위해 조정 가능한 요인을 명시적으로 모델링할 수 있습니다.

STP 분석에는 위치 데이터, 예측(예측변수)에 대한 입력 필드, 시간 필드 및 목표 필드가 포함되는 데이터가 사용됩니다. 각 위치에서는 데이터에 측정 시 각 예측변수의 값을 나타내는 여러 행이 있습니다. 데이터 분석 후, 분석에서 사용되는 모양 데이터 내의 위치에서 목표값을 예측하기 위해 사용할 수 있습니다. 또한 나중 시점에 대한 입력 데이터를 알 수 있는 시기를 예측할 수도 있습니다.

참고: STP 노드는 IBM SPSS Collaboration and Deployment Services에서 모델 평가 또는 챔피언 챌린저 단계를 지원하지 않습니다.

STP 사용 작업 예제를 보여주고(`server_demo.str`) `room_data.csv` 및 `score_data.csv` 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. `server_demo.str` 파일은 `streams` 디렉토리에 있습니다.

STP(Spatio-Temporal Prediction) - 필드 옵션

필드 탭에서, 이미 업스트림 노드에 정의된 필드 역할 설정을 사용할 것인지 여부를 선택하거나 필드에 수동으로 할당합니다.

사전 정의된 역할 사용

이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표 및 예측변수만)을 사용합니다.

사용자 정의 필드 할당 사용

이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 지정하려면 이 옵션을 선택합니다.

필드 선택할 수 있는 데이터의 모든 필드를 표시합니다. 이 목록에서 화면의 오른쪽에 있는 다양한 상자로 수동으로 항목을 지정하려면 화살표 단추를 사용하십시오. 아이콘은 각 필드에 대한 유효한 측정 수준을 나타냅니다.

참고: STP에서는 올바르게 작동하기 위해 위치당, 시간 구간당 하나의 레코드가 필요하므로, 필수 필드가 됩니다.

필드 분할창의 아래쪽에서, 측정 수준에 상관없이 모든 필드를 선택하려면 **모두** 단추를 클릭하고, 해당 측정 수준을 가지고 있는 모든 필드를 선택하려면 개별적인 측정 수준 단추를 클릭하십시오.

목표 예측에 대한 목표로 하나의 필드를 선택합니다.

참고: 연속형의 측정 수준을 가지고 있는 필드만 선택할 수 있습니다.

위치 모델에서 사용할 위치 유형을 선택합니다.

참고: 특정 지역과 관련된 측정 수준을 가지고 있는 필드만 선택할 수만 있습니다.

위치 레이블

형태 데이터에는 종종 레이어에서 기능의 이름(예: 주 또는 국가의 이름)을 표시하는 필드가 포함됩니다. 출력에서 선택한 **위치** 필드에 레이블을 붙이기 위해 범주형 필드를 선택하여 이름 또는 레이블을 위치와 연관시키려면 이 필드를 사용하십시오.

시간 필드

예측에 사용할 시간 필드를 선택합니다.

참고: 연속형의 측정 수준과, 시간, 날짜, 시간소인 또는 정수의 저장 유형을 가지고 있는 필드만 선택할 수 있습니다.

예측변수(입력)

예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

참고: 연속형의 측정 수준을 가지고 있는 필드만 선택할 수 있습니다.

STP(Spatio-Temporal Prediction) - 시간 구간

시간 구간 분할창에서, 시간이 지나면서 시간 구간과 필수 통합을 설정하기 위한 옵션을 선택할 수 있습니다.

STP 모델을 작성하기 전에 시간 필드를 지수로 변환하기 위해 데이터 준비가 필요합니다. 시간 필드를 변환하려면 레코드 사이에 일정한 구간이 있어야 합니다. 데이터에 아직 이 정보가 포함되지 않으면, 모델링 노드를 사용하기 전에 이 구간을 설정하기 위해 이 분할창에서 옵션을 사용하십시오.

시간 구간 데이터 세트를 변환하려고 하는 구간을 선택하십시오. 사용 가능한 옵션은 필드 탭에서 모델에 대한 **시간 필드**로 선택한 필드의 저장 유형에 따라 다릅니다.

- **주기** 정수 시간 필드의 경우에만 사용할 수 있으며, 사용 가능한 다른 구간과 일치하지 않는, 각 측정 사이의 구간이 일정한 일련의 구간입니다.

- **년** 날짜 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.
- **분기** 날짜 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다. 이 옵션을 선택하면, 첫 번째 분기의 **시작 월**을 선택하도록 요청하는 프롬프트가 표시됩니다.
- **월** 날짜 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.
- **주** 날짜 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.
- **일** 날짜 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.
- **시** 시간 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.
- **분** 시간 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.
- **초** 시간 또는 시간소인 시간 필드에 대해서만 사용할 수 있습니다.

시간 구간을 선택할 때, 추가 필드를 완료하도록 요청하는 프롬프트가 표시됩니다. 사용 가능한 필드는 시간 구간과 저장 유형에 따라 다릅니다. 표시될 수 있는 필드는 다음 목록에 있습니다.

- **주당 일 수**
- **하루 중 시간**
- **주 시작 요일** 주의 첫 번째 요일
- **하루 시작 시간** 새 날이 시작되는 것으로 간주되는 시간.
- **구간 값** 1, 2, 3, 4, 5, 6, 10, 12, 15, 20 또는 30 옵션 중에서 하나를 선택할 수 있습니다.
- **시작 월** 회계연도가 시작되는 월.
- **시작 주기 주기**를 사용 중인 경우, 시작 주기를 선택하십시오.

데이터가 지정된 시간 구간 설정과 일치함 데이터에 이미 정확한 시간 구간 정보가 포함되고 변환이 필요하지 않은 경우, 이 확인 상자를 선택하십시오. 이 상자를 선택할 때, **통합** 영역의 필드는 사용할 수 없습니다.

통합

데이터가 지정된 시간 구간 설정과 일치함 확인 상자에서 선택을 지우는 경우에만 사용할 수 있습니다. 지정된 구간과 일치하도록 필드를 통합하는 경우에 옵션을 지정하십시오. 예를 들어, 매주 및 매월 데이터의 혼합이 있는 경우, 일정한 매월 구간이 되도록 매주 값을 통합 또는 "롤업"할 수 있습니다. 여러 필드 유형의 통합에 사용될 기본 설정을 선택하고 특정 필드에 대해 원하는 사용자 정의 설정을 작성하십시오.

- **연속형** 개별적으로 지정되지 않은 모든 연속형 필드에 적용될 기본 통합 방법을 설정하십시오. 다음의 여러 방법에서 선택할 수 있습니다.
  - 합계
  - 평균
  - 최소값
  - 최대값
  - 중앙값

– 첫 번째 사분위수
– 세 번째 사분위수

지정된 필드에 대한 사용자 정의 설정 개별 필드에 특정 통합 함수를 적용하려면, 이 테이블에서 함수를 선택하고 통합 방법을 선택하십시오.

- **필드** 필드 선택 대화 상자를 표시하고 필수 필드를 선택하려면 **필드 추가** 단추를 사용하십시오. 선택된 필드는 이 열에 표시됩니다.
- **통합 함수** 드롭 다운 목록에서, 지정된 시간 구간으로 필드를 변환하기 위한 통합 함수를 선택하십시오.

STP(Spatio-Temporal Prediction) - 기본 작성 옵션

기본 모델 작성 옵션을 설정하려면 이 대화 상자에서 설정을 사용하십시오.

모델 설정

절편 포함

절편(모델에서 상수항)을 포함하면 솔루션의 전체 정확도가 증가할 수 있습니다. 데이터가 원점을 통과하여 전달된다고 가정할 수 있는 경우 절편을 제외할 수 있습니다.

최대 자기회귀적 순서

자기회귀 순서는 현재 값 예측에 사용될 이전 값을 지정합니다. 새 값을 계산하기 위해 사용되는 이전 레코드 수를 지정하려면 이 옵션을 사용하십시오. 1과 5 사이의 정수를 선택할 수 있습니다.

공간 공분산

추정 방법

사용할 추정 방법을 선택하십시오. **모수** 또는 **비모수**를 선택할 수 있습니다. **모수** 방법의 경우 세 가지 **모델** 유형 중 하나에서 선택할 수 있습니다.

- **가우스**
- **지수**
- **거듭제곱 지수** 이 옵션을 선택하는 경우, 사용할 **거듭제곱** 수준도 지정해야 합니다. 이 수준은 0 - 1 사이의 값(0.1씩 증분 변경됨)이 될 수 있습니다.

STP(Spatio-Temporal Prediction) - 고급 작성 옵션

자세한 STP 지식을 가지고 있는 사용자는 다음 옵션을 사용하여 모델 작성 프로세스를 자세히 조정할 수 있습니다.

결측값의 최대 백분율

모델에 포함될 수 있는 결측값을 포함하는 레코드의 최대 퍼센트를 지정하십시오.

모델 작성에서 가설 검정 유의 수준(N)

두 가지의 적합도 검정, 효과 F 검정 및 계수 T 검정을 포함하여, STP 모델 예측의 모든 검정에 사용될 유의 수준 값을 지정하십시오. 이 수준은 0 - 1의 값(0.01 증분에서 변경)이 될 수 있습니다.

STP(Spatio-Temporal Prediction) - 출력

모델을 작성하기 전에, 출력 뷰어에 포함할 출력을 선택하려면 이 분할창의 옵션을 사용하십시오.

모델 정보

모델 지정 사항

모델 출력에 모델 지정 사항을 정보를 포함하려면 이 옵션을 선택하십시오.

임시 정보 요약

모델 출력에 임시 정보 요약을 포함하려면 이 옵션을 선택하십시오.

평가

모델 품질

모델 품질을 모델 출력에 포함시키려면 이 옵션을 선택하십시오.

평균 구조 모델에서 효과 검정

모델 출력에 효과 검정 정보를 포함하려면 이 옵션을 선택하십시오.

설명

모델 계수의 평균 구조

모델 출력에 평균 구조 모델 계수 정보를 포함하려면 이 옵션을 선택하십시오.

자기회귀 계수

모델 출력에 자기회귀 계수 정보를 포함하려면 이 옵션을 선택하십시오.

공백 감소 검정

모델 출력에 공간 공분산 또는 공백 감소 검정 정보를 포함하려면 이 옵션을 선택하십시오.

매개변수식 공간 공분산 모델 모수 도표

모델 출력에 매개변수식 공간 공분산 모델 모수 도표 정보를 포함하려면 이 옵션을 선택하십시오.

참고: 기본 탭에서 **매개변수식** 추정 방법을 선택한 경우 이 옵션만 사용할 수 있습니다.

상관계수 히트 맵

모델 출력에 목표 값의 맵을 포함하려면 이 옵션을 선택하십시오.

참고: 사용자 모델에 500개보다 많은 위치가 있는 경우 맵 출력이 작성되지 않습니다.

상관계수 맵

모델 출력에 상관관계 맵을 포함하려면 이 옵션을 선택하십시오.

참고: 사용자 모델에 500개보다 많은 위치가 있는 경우 맵 출력이 작성되지 않습니다.

위치 군집

모델 출력에 위치 군집 출력을 포함하려면 이 옵션을 선택하십시오. 맵 데이터에 대한 액세스가 필요하지 않은 출력만 군집 출력의 일부로 포함됩니다.

참고: 이 출력은 비모수 공간 공분산 모델에 대해서만 작성될 수 있습니다.

이 옵션을 선택하는 경우 다음을 설정할 수 있습니다.

- **유사성 임계값** 출력 군집이 단일 군집으로 병합되기에 충분히 유사한 것으로 간주되는 임계값을 선택하십시오.
- **표시할 최대 군집 수** 모델 출력에 포함될 수 있는 군집 수의 상한을 설정하십시오.

STP(Spatio-Temporal Prediction) - 모델 옵션

모델 이름 목표 필드를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다. 자동으로 생성된 이름은 목표 필드 이름입니다.

불확실성 요인(%) 불확실성 요인은 미래를 예측할 때 불확실성 증가를 나타내는 퍼센트 값입니다. 예측 불확실성의 상한 및 하한은 미래의 각 단계에 대해 이 퍼센트 기준으로 증가합니다. 모델 출력에 적용될 불확실성 요인을 설정하십시오. 그러면 예측값의 상한 및 하한이 설정됩니다.

STP(Spatio-Temporal Prediction) 모델 너깃

STP(Spatio-Temporal Prediction) 모델 너깃은 출력 뷰어의 모델 탭에서 모델의 세부사항을 표시합니다. 뷰어 사용에 대한 자세한 정보는 모델러 사용자 안내서(ModelerUsersGuide.pdf)에서 "출력에 대한 작업" 절을 참조하십시오.

STP(spatio-temporal prediction) 모델링 조작은 다음 표에 표시된 대로 접두문자 $STP-를 사용하여 여러 새 필드를 작성합니다.

표 *26. STP* 모델링 작업으로 작성된 새 필드

필드 이름	설명
$STP-<Time>	모델 작성의 일부로 작성된 시간 필드. 작성 옵션 탭의 시간 구간 분할창에 있는 설정에 따라 이 필드의 작성 방법이 결정됩니다. <시간>은 필드 탭에서 **시간 필드**로 선택된 필드의 원래 이름입니다. **참고:** 이 필드는 원래 **시간 필드**를 모델 작성의 일부로 변환한 경우에만 작성됩니다.
$STP-<Target>	이 필드에는 목표 값에 대한 예측이 포함됩니다. <Target>은 모델에 대한 원래 **목표** 필드의 이름입니다.
$STPVAR-<Target>	이 필드에는 VarianceOfPointPrediction 값이 포함됩니다. <Target>은 모델에 대한 원래 **목표** 필드의 이름입니다.
$STPLCI-<Target>	이 필드는 LowerOfPredictionInterval 값(즉, 신뢰도의 하한)을 포함합니다. <Target>은 모델에 대한 원래 **목표** 필드의 이름입니다.

표 *26. STP* 모델링 작업으로 작성된 새 필드 *(계속)*

$STPUCI-<Target>	이 필드는 UpperOfPredictionInterval 값(신뢰도의 상한)을 포함합니다. <Target>은 모델에 대한 원래 **목표** 필드의 이름입니다.

STP(Spatio-Temporal Prediction) 모델 설정

모델링 작업에서 허용 가능한 것으로 간주되는 불확실성의 수준을 제어하려면 설정 탭을 사용하십시오.

불확실성 요인(%) 불확실성 요인은 미래를 예측할 때 불확실성 증가를 나타내는 퍼센트 값입니다. 예측 불확실성의 상한 및 하한은 미래의 각 단계에 대해 이 퍼센트 기준으로 증가합니다. 모델 출력에 적용될 불확실성 요인을 설정하십시오. 그러면 예측값의 상한 및 하한이 설정됩니다.

TCM 노드

이 노드를 사용하여 시간 인과 모델(TCM)을 작성합니다.

시간 인과 모델

시간 인과 모델링은 시계열 데이터에서 핵심 인과 관계를 검색하려고 시도합니다. 시간 인과 모델링에서, 목표 계열 세트와 해당 목표에 대한 후보 입력 세트를 지정합니다. 프로시저는 목표마다 자기회귀 시계열 모델을 작성하고 목표와 인과 관계를 갖는 입력만 포함합니다. 이 접근 방법은 목표 계열에 대해 명시적으로 예측변수를 지정해야 하는 전형적인 시계열 모델링과 다릅니다. 시간 인과 모델링에는 일반적으로 여러 관련 시계열에 대한 모델 작성이 포함되므로, 결과를 모델 시스템이라고 합니다.

시간 인과 모델링의 컨텍스트에서, 인과 용어는 그랜저 인과성을 가리킵니다. X 및 Y 둘 다의 지난 값 관점에서 Y에 대한 회귀추정이 Y의 지난 값에 대해서만 회귀추정하는 것보다 Y에 대해 더 나은 모델이 생성되는 경우 시계열 X는 다른 시계열 Y의 "그렌저 인과" 관계라고 합니다.

참고: 시간 인과 모델링 노드는 IBM SPSS Collaboration and Deployment Services에서 모델 평가 또는 챔피언 챌린저 단계를 지원하지 않습니다.

예제

비즈니스 의사결정자는 비즈니스를 설명하는 시간 기반 메트릭의 큰 세트 내에서 인과 관계를 알아내기 위해 시간 인과 모델링을 사용할 수 있습니다. 분석은 몇 개의 제어 가능한 입력을 드러낼 수 있습니다. 이 입력은 핵심성과지표(KPI)에 대한 가장 큰 영향을 가지고 있습니다.

대규모 IT 시스템의 관리자는 시간 인과 모델링을 사용하여 상호 관련된 작동 메트릭의 큰 세트에서 이상 항목을 발견할 수 있습니다. 인과 모델은 이상 항목 발견을 넘어서 이상 항목의 가장 근본적인 원인을 발견할 수 있도록 합니다.

필드 요구 사항

최소 하나의 목표가 있어야 합니다. 기본적으로, 사전 정의된 역할이 없음인 필드는 사용되지 않습니다.

데이터 구조

시간 인과 모델링은 두 가지 유형의 데이터 구조를 지원합니다.

열 기반 데이터

열 기반 데이터의 경우, 각 시계열 필드에는 단일 시계열에 대한 데이터가 포함됩니다. 이 구조는 시계열 모델 생성기에서 사용되는 전형적인 시계열 데이터 구조입니다.

다차원 데이터

다차원 데이터의 경우, 각 시계열 필드에는 여러 시계열에 대한 데이터가 포함됩니다. 특정 필드 내에서 별도의 시계열은 차원 필드라고 하는 범주형 필드의 값 세트로 식별됩니다.예를 들어, 두 개의 다른 판매 채널(소매 및 웹)에 대한 판매 데이터가 단일 *sales* 필드에 저장될 수 있습니다. 이름이 *channel*이고 값이 'retail' 및 'web'인 차원 필드는 두 판매 채널 각각과 연관되는 레코드를 식별합니다.

참고: 시간 인과 모델을 작성하려면 충분한 데이터 점이 필요합니다. 곱은 제한조건을 사용합니다.

```
m>(L + KL + 1)
```

여기서, m은 데이터 점의 수이고 L은 시차의 수이며 K는 예측자 수입니다. 데이터 점의 수(m)가 조건을 충족할 수 있도록 데이터 세트가 충분히 큰지 확인하십시오.

모델링할 시계열

필드 탭에서, 모델 시스템에서 포함할 계열을 지정하려면 **시계열** 설정을 사용하십시오.

데이터에 적용되는 데이터 구조에 맞는 옵션을 선택하십시오. 다차원 데이터의 경우, 차원 필드를 지정하기 위해 **차원 선택**을 클릭하십시오. 차원 필드의 지정된 순서는 모두 연속 대화 상자 및 출력에 차원이 나타나는 순서를 정의합니다. 차원 필드를 다시 정렬하려면 차원 선택 하위 대화 상자에서 위로 및 아래로 화살표 단추를 사용하십시오.

열 기반 데이터에 대해 계열 용어의 의미는 필드 용어 의미와 같습니다. 다차원 데이터의 경우, 시계열을 포함하는 필드는 메트릭 필드로 언급됩니다. 다차원 데이터에 대해 시계열은 차원 필드 각각에 대한 메트릭 필드 및 값으로 정의됩니다. 다음 고려사항은 열 기반 및 다차원 데이터에 적용됩니다.

- 후보 입력으로, 또는 목표 및 입력 둘 다로 지정되는 계열은 각 목표에 대해 모델에서 포함을 위해 고려됩니다. 각 목표에 대한 모델에는 항상 목표 자체의 시차 값이 포함됩니다.
- 강제 입력으로 지정되는 계열은 항상 각 목표에 대해 모델에 포함됩니다.
- 최소 하나의 계열을 목표로 또는 목표 및 입력 둘 다로 지정해야 합니다.
- **사전 정의된 역할 사용**이 선택되면, 입력 역할을 가지고 있는 필드가 후보 입력으로 설정됩니다. 사전 정의된 어떤 규칙도 강제 입력에 맵핑하지 않습니다.

다차원 데이터

다차원 데이터의 경우, 눈금에서 메트릭 필드 및 연관된 역할을 지정합니다. 눈금의 각 행은 단일 메트릭 및 역할을 지정합니다. 기본적으로, 모델 시스템에는 눈금의 각 행에 대한 차원 필드의 모든 조합에 대한 계열이 포함됩니다. 예를 들어, *region* 및 *brand*에 대한 차원이 있는 경우, 기본적으로 목표로 메트릭 *sales*를 지정하면, 이는 각각의 *region* 및 *brand* 조합에 대해 별도의 sales 목표 계열이 있음을 의미합니다.

눈금의 각 행에 대해, 차원의 생략 기호 단추를 클릭하여 차원 필드에 대한 값 세트를 사용자 정의할 수 있습니다. 이 동작은 차원 값 선택 하위 대화 상자를 엽니다. 또한 눈금 행을 추가, 삭제 또는 복사할 수도 있습니다.

계열 개수 열은 현재 연관 메트릭에 대해 지정된 차원 값 세트 수를 표시합니다. 표시된 값은 실제 계열 수(세트당 하나의 계열)보다 클 수 있습니다. 이 조건은 지정된 차원 값 조합 중 일부가 연관 메트릭에 의해 포함된 계열에 해당되지 않는 경우에 발생합니다.

차원 값 선택을 선택하십시오: 다차원 데이터의 경우, 특정 역할이 있는 특정 메트릭 필드에 적용되는 차원 값을 지정하여 분석을 사용자 정의할 수 있습니다. 예를 들어, *sales*가 메트릭 필드이고 *channel*이 'retail' 및 'web' 값을 가지고 있는 차원인 경우, 'web' 판매가 입력이고 'retail' 판매가 목표임을 지정할 수 있습니다. 또한 분석에 사용되는 모든 메트릭 필드에 적용되는 차원 서브세트를 지정할 수도 있습니다. 예를 들어, *region*이 지리적 지역을 나타내는 차원 필드인 경우 분석을 특정 지역으로 제한할 수 있습니다.

모든 값

현재 차원 필드의 모든 값이 포함됨을 지정합니다. 이 옵션은 기본값입니다.

포함하거나 제외할 값 선택

현재 차원 필드의 값 세트를 지정하려면 이 옵션을 사용하십시오. **모드**에 대해 **포함**이 선택되는 경우, **선택된 값** 목록에 지정되는 값만 포함됩니다. **모드**에 대해 **제외**가 선택되는 경우, **선택된 값** 목록에 지정된 값이 아닌 다른 모든 값이 포함됩니다.

선택할 값 세트를 필터링할 수 있습니다. 필터 조건에 충족하는 값은 **매치됨** 탭에 나타나고, 필터 조건과 일치하지 않는 값은 **선택되지 않은 값** 목록의 **매치하지 않음** 탭에 나타납니다. **모두** 탭은 필터 조건과 관계없이 선택되지 않은 모든 값을 나열합니다.

- 필터를 지정할 때 와일드카드 문자를 표시하기 위해 별표(*)를 사용할 수 있습니다.
- 현재 필터를 지우려면, 표시된 값 필터링 대화 상자에 검색어로 비어 있는 값을 지정하십시오.

관측값

필드 탭에서, 관측값을 정의하는 필드를 지정하려면 **관측값** 설정을 사용하십시오.

날짜/시간에 의해 정의되는 관측값

관측값이 날짜, 시간 또는 시간소인 필드에 의해 정의됨을 지정할 수 있습니다. 관측값을 정의하는 필드 외에, 관측값을 설명하는 적절한 시간 구간을 선택하십시오. 지정된 시간 구간에 따라, 관측값(증분) 사이의 구간이나 주당 일 수와 같은 다른 설정을 지정할 수도 있습니다. 다음 고려사항은 시간 간격에 적용됩니다.

- 관측값이 시간에서 비정규적으로 간격이 있는 경우(판매 순서가 처리되는 시간과 같이), **비정규** 값을 사용하십시오. **비정규**가 선택될 때, 데이터 지정 사항 탭의 **시간 간격** 설정에서 분석에 사용되는 시간 구간을 지정해야 합니다.
- 관측값이 날짜와 시간을 나타내고 시간 구간이 시, 분 또는 초인 경우 **하루 중 시간(시)**, **하루 중 시간(분)** 또는 **하루 중 시간(초)**을 사용하십시오. 관측값이 날짜에 대한 참조 없이 시간(기간)을 나타내고 시간 구간이 시, 분 또는 초일 경우, **시(비주기적)**, **분(비주기적)** 또는 **초(비주기적)**를 사용하십시오.
- 선택된 시간 간격을 기초로, 프로시저는 결측 관측값을 발견할 수 있습니다. 프로시저에서는 모든 관측값이 시간에서 동일하게 간격을 두고 결측 관측값이 없다고 가정하므로, 결측 관측값을 발견해야 합니다. 예를 들어, 시간 구간이 일(Days)이고 날짜 2014-10-27 뒤에 2014-10-29가 있는 경우, 2014-10-28에 대해 결측 관측값이 있습니다. 값은 결측 관측값에 대해 대체됩니다. 결측값 처리에 대한 설정은 데이터 지정 사항 탭으로부터 지정할 수 있습니다.
- 지정된 시간 구간은 프로시저가 함께 통합해야 하는 동일한 시간 구간의 여러 관측값을 발견하고 관측값에 동일하게 간격이 있도록 월의 첫 번째와 같은 구간 경계에 관측값을 맞출 수 있도록 합니다. 예를 들어, 시간 구간이 월일 경우, 동일 월에 있는 여러 날짜가 함께 통합됩니다. 이 유형의 통합을 그룹화라고 합니다. 기본적으로, 관측값은 그룹화될 때 합산됩니다. 데이터 지정 사항 탭의 **통합 및 분포** 설정에서, 그룹화에 다른 방법(예: 관측값의 평균)을 지정할 수 있습니다.
- 일부 시간 구간의 경우, 추가 설정은 동일하게 간격이 있는 정규 구간에서 중단을 설정할 수 있습니다. 예를 들어, 시간 구간이 일(Days)이지만 평일만 유효한 경우, 주에 5일이 있고 주는 월요일에 시작함을 지정할 수 있습니다.

관측값이 주기 또는 순환 주기로 정의됨

관측값은 임의의 순환 수준 수까지, 주기 또는 반복 주기 순환을 나타내는 하나 이상의 정수 필드로 정의할 수 있습니다. 이 구조에서, 표준 시간 구간 중 하나에 맞지 않은 관측값 계열을 설정할 수 있습니다. 예를 들어, 10개월만 있는 회계연도는 연도를 나타내는 순환 필드와, 월을 나타내는 주기 필드로 설명할 수 있습니다. 여기서 하나의 주기 길이는 10입니다.

순환 주기를 지정하는 필드는 주기적 수준의 계층 구조를 정의합니다. 가장 낮은 수준은 **주기** 필드에 의해 정의됩니다. 다음 최상위 수준은 수준이 1인 순환 필드에 의해 지정되고, 그 다음은 수준 2의 순환 필드로 지정되며 뒤로도 마찬가지입니다. 가장 높은 수준을 제외하고, 각 수준의 필드 값은 다음 최상위 수준에 관하여 주기적이어야 합니다. 최상위 수준의 값은 주기적이 될 수 없습니다. 예를 들어, 10달 회계연도의 경우 월은 연도 내에서 주기적이며 연도는 주기적이 아닙니다.

- 특정 수준에 있는 순환의 길이는 다음으로 가장 낮은 수준의 주기성입니다. 회계연도 예의 경우, 단 하나 순환 수준이 있고 순환 길이는 10입니다. 다음으로 가장 낮은 수준이 월을 나타내고 지정된 회계 연도에 10달이 있기 때문입니다.
- 1에서 시작하지 않은 주기적 필드의 시작 값을 지정하십시오. 이 설정은 결측값을 발견하는 데 필요합니다. 예를 들어, 주기적 필드는 2에서 시작하지만 시작 값은 1로 지정되는 경우, 프로시저는 해당 필드의 각 순환에 있는 첫 번째 주기에 대해 결측값이 있다고 가정합니다.

분석에 대한 시간 구간

분석에 사용되는 시간 구간은 관측값의 시간 구간과 다를 수 있습니다. 예를 들어, 관측값의 시간 구간이 일(Days)일 경우, 분석의 시간 구간으로는 월을 선택할 수 있습니다. 데이터는 모델이 작성되기 전에 매일 데이터에서 매월 데이터까지 통합됩니다. 또한 데이터를 장기 시간 구간에서 단기 시간 구간으로 분포할 것을 선택할 수도 있습니다. 예를 들어, 관측값이 분기별인 경우, 데이터를 분기별에서 월별 데이터로 분포할 수 있습니다.

분석이 행해지는 시간 구간에 대해 사용 가능한 선택은 해당 관측값 정의 방법과 관측값의 시간 구간에 따라 다릅니다. 특히, 관측값이 순환 주기로 정의될 경우, 통합만 지원됩니다. 그러한 경우, 분석의 시간 구간은 관측값의 시간 구간보다 크거나 같아야 합니다.

분석 시간 구간은 데이터 지정 사항 탭의 **시간 구간** 설정에서 지정됩니다. 데이터가 통합되거나 분포되는 방법은 데이터 지정 사항 탭의 **통합 및 분포** 설정에서 지정됩니다.

통합 및 분포

통합 함수

분석에 사용되는 시간 구간이 관측에 사용되는 시간 구간보다 길 경우, 입력 데이터는 통합됩니다. 예를 들어, 관측값의 시간 구간이 일(Days)이고 분석의 시간 구간이 월일 경우 통합이 수행됩니다. mean, sum, mode, min 또는 max 통합 함수를 사용할 수 있습니다.

분포 함수

분석에 사용되는 시간 구간이 관측의 시간 구간보다 짧을 경우, 입력 데이터는 분포됩니다. 예를 들어, 관측값의 시간 구간이 분기이고 분석의 시간 구간이 월일 경우 분포가 수행됩니다. mean 또는 sum 분포 함수를 사용할 수 있습니다.

그룹화 함수

그룹화는 관측값이 날짜/시간에 의해 정의되고 여러 관측값이 동일 시간 구간에 발생하는 경우에 적용됩니다. 예를 들어, 관측값의 시간 구간이 월일 경우, 동일 월에 있는 여러 날짜가 그룹화되어 날짜가 발생하는 월과 연관됩니다. mean, sum, mode, min 또는 max와 같은 그룹화 함수를 사용할 수 있습니다. 다음 그룹화는 항상 관측값이 날짜/시간에 의해 정의되고 관측값의 시간 간격이 비정규로 지정된 경우에 수행됩니다

참고: 그룹화가 통합 양식이어도, 그룹화는 결측값 처리 이전에 수행됩니다(정상 통합은 결측값 처리 이후에 수행됩니다). 관측값의 시간 구간이 비정규로 지정되는 경우, 통합은 그룹화 함수로만 수행됩니다.

교차-일 관측값을 이전 일로 통합

1일 경계를 교차하는 시간을 사용하는 관측값이 전날의 값에 통합되는지 여부를 지정합니다. 예를 들어, 20:00시에 시작하는 8시간 노동의 시간별 관측값의 경우, 이 설정은 00:00 및 04:00 사이의 관측값이 전날 통합 결과에 포함되는지 여부를 지정합니다. 이 설정은 관측값의 시간 구간이 하루 중 시간(시), 하루 중 시간(분) 또는 하루 중 시간(초)이고 분석의 시간 구간이 일(Days)인 경우에만 적용됩니다.

지정된 필드에 대한 사용자 정의 설정

필드 기준으로 필드에 통합, 분포 및 그룹화 함수를 지정할 수 있습니다. 이 설정은 통합, 분포 및 그룹화 함수에 대한 기본 설정을 대체합니다.

결측값

입력 데이터의 결측값은 대치된 값으로 바뀝니다. 다음 방법으로 바꿀 수 있습니다.

선형 보간법

선형 보간법을 사용하여 결측값을 바꿉니다. 결측값 이전의 마지막 유효한 값과 결측값 이후의 첫 번째 유효한 값이 보간법에 사용됩니다. 계열에서 첫 번째 또는 마지막 관측값에 결측값이 있는 경우, 계열의 시작 또는 종료에서 두 개의 가장 근접한 비결측 값이 사용됩니다.

계열 평균

결측값을 전체 계열에 대한 평균으로 바꿉니다.

근접한 값들의 평균

결측값을 유효한 근접 값의 평균으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

근접한 값들의 중앙값

결측값을 근접한 유효한 값의 중앙값으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

선형 추세

이 옵션은 단순 선형 회귀 모형을 적합시키기 위해 계열에서 모든 비결측 관측값을 사용합니다. 이 모델은 결측값을 대치하기 위해 사용됩니다.

기타 설정:

결측값의 최대 퍼센트(%)

어떤 계열에 대해서도 허용되는 최대 결측값 퍼센트를 지정합니다. 지정된 최대값보다 많은 결측값이 있는 계열은 분석에서 제외됩니다.

일반 데이터 옵션

차원 필드당 최대 고유 값 수

이 설정은 다차원 데이터에 적용되며 하나의 차원 필드에 대해 허용되는 최대 고유 값 수를 지정합니다. 기본적으로, 이 한계는 10000으로 설정되지만, 임의로 큰 숫자로 증가될 수 있습니다.

일반 작성 옵션

신뢰구간 너비(%)

이 설정은 예측 및 모델 모수 둘 다의 신뢰구간을 제어합니다. 100보다 작은 양수 값을 지정할 수 있습니다. 기본적으로, 95% 신뢰구간이 사용됩니다.

각 목표에 대한 최대 입력 수

이 설정은 각 목표에 대한 모델에서 허용되는 최대 입력 수를 지정합니다. 1 - 20 범위의 정수를 지정할 수 있습니다. 각 목표에 대한 모델에는 항상 자체의 시차 값이 포함되므로, 이 값을 1로 설정하면 입력만 목표 자체가 됩니다.

모델 허용 한도

이 설정은 각 목표에 대한 최상의 입력 세트를 판별하기 위해 사용되는 반복 프로세스를 제어합니다. 0보다 큰 값을 지정할 수 있습니다. 기본값은 0.001입니다. 모델 공차는 예측자 선택에 대한 중단 기준입니다. 이는 최종 모델에 포함되는 예측자 수에 영향을 미칠 수 있습니다. 단, 목표가 스스로 매우 잘 예측할 수 있는 경우, 기타 예측자가 최종 모델에 포함되지 않을 수 있습니다. 일부 시행 착오가 필요할 수 있습니다. 예를 들어, 이 값을 높게 설정한 경우, 이를 더 작은 값으로 설정하여 기타 예측자가 포함될 수 있는지 여부를 알 수 있습니다.

이상값 임계값(%)

모델에서 계산된 확률(이상값)이 이 임계값을 초과하는 경우 관측값은 이상값으로 플래그가 붙습니다. 50 - 100 범위의 값을 지정할 수 있습니다.

각 입력의 시차 수

이 설정은 각 목표에 대한 모델에서 각 입력의 시차 항 수를 지정합니다. 기본적으로, 시차 항 수는 분석에서 사용되는 시간 구간에서 자동으로 결정됩니다. 예를 들어, 시간 구간이 월(증분: 한 달)인 경우 시차 수는 12입니다. 선택적으로, 시차 수를 명시적으로 지정할 수 있습니다. 지정된 값은 1 - 20 범위의 정수여야 합니다.

기존 모델을 사용하여 추정 계속

이미 시간 인과 모델을 생성한 경우, 새 모델을 작성하기 보다는 해당 모델에 대해 지정된 기준 설정을 재사용하려면 이 옵션을 선택하십시오. 이 방식에서는 이전(그러나, 최근의 데이터)과 동일한 모델 설정을 기반으로 하는 새 예측을 다시 추정하고 생성하여 시간을 절약할 수 있습니다.

표시할 계열

이 옵션은 출력이 표시되는 계열(목표 또는 입력)을 지정합니다. 지정된 계열에 대한 출력의 내용은 **출력 옵션** 설정으로 판별됩니다.

최적 적합 모델과 연관된 목표 표시

기본적으로, R 제곱 값으로 판별되는 10개의 최적 적합 모델과 연관되는 목표에 대해 출력이 표시됩니다. 최적 적합 모델의 다른 고정 숫자를 지정하거나 최적 적합 모델의 백분율을 지정할 수 있습니다. 다음 적합도 측도에서 선택할 수도 있습니다.

R 제곱

선형모델의 적합도 측도로서 결정계수라고도 합니다. 이 항목은 모델로 설명한 목표변수의 변동 비율이 됩니다. 값 범위는 0 - 1입니다. 값을 작을수록 모델이 데이터에 적합하지 않음을 의미합니다.

제곱근 평균 제곱 퍼센트 오차

모델 예측값이 계열의 관측값과 얼마나 다른지의 측도입니다. 사용된 단위와 상관이 없으므로 다른 단위의 계열을 비교하는 데 사용할 수 있습니다.

제곱근 평균제곱오차

평균 제곱 오차의 제곱근입니다. 종속 계열이 모델 예측 수준과 얼마나 다른지에 대한 측도로서, 종속 계열과 같은 단위로 표시됩니다.

BIC 베이지안 정보 기준. -2 축소 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. BIC도 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여"하지만 AIC보다 더 엄격하게 부여합니다.

AIC Akaike 정보 기준. -2 축소 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. AIC는 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여합니다".

개별 계열 지정

출력할 개별 계열을 지정할 수 있습니다.

- 열 기반 데이터의 경우, 원하는 계열을 포함하는 필드를 지정합니다. 지정된 필드의 순서는 출력에 필드가 나타나는 순서를 정의합니다.
- 다차원 데이터의 경우, 계열을 포함하는 메트릭 필드에 대한 눈금에 진입을 추가하여 특수 계열을 지정합니다. 그런 다음 계열을 정의하는 차원 필드의 값을 지정합니다.
  - 각 차원 필드의 값을 눈금에 직접 입력하거나 사용 가능한 차원 값 목록에서 선택할 수 있습니다. 사용 가능한 차원 값 목록에서 선택하려면 원하는 차원의 셀에서 생략 기호 단추를 클릭하십시오. 이 동작을 수행하면 차원 값 선택 하위 대화 상자가 열립니다.
  - 쌍안경 아이콘을 클릭하고 검색어를 지정하여, 차원 값 선택 하위 대화 상자에서 차원 값 목록을 검색할 수 있습니다. 공백은 검색어의 일부로 처리됩니다. 검색어의 별표(*)는 와일드카드 문자를 표시하지 않습니다.
  - 눈금에서 계열의 순서는 출력에 나타나는 순서를 정의합니다.

열 기반 데이터 및 다차원 데이터 둘 다에 대해, 출력은 30 계열로 제한됩니다. 이 한계에는 사용자가 지정하는 개별 계열(입력 또는 목표)과 최적 적합 모델과 연관되는 목표가 포함됩니다. 개별적으로 지정된 계열은 최적 적합 모델과 연관된 목표보다 우선순위가 높습니다.

출력 옵션

이러한 옵션은 출력의 내용을 지정합니다. **목표에 대한 출력결과** 그룹의 옵션은 **표시할 계열** 설정의 최적 적합 모델과 연관되는 목표에 대한 출력을 생성합니다. **계열에 대한 출력결과** 그룹의 옵션은 **표시할 계열** 설정에 지정된 개별 계열에 대한 출력을 생성합니다.

전체 모델 시스템

모델 시스템에서 계열 사이의 인과 관계에 대한 그래픽 표현을 표시합니다. 표시된 목표에 대한 모델 적합 통계 및 이상값 둘 다의 테이블이 출력 항목의 일부로 포함됩니다. **계열에 대한 출력결과** 그룹에서 이 옵션이 선택되면, **표시할 계열** 설정에 지정된 개별 계열마다 별도의 출력 항목이 작성됩니다.

계열 사이의 인과 관계에는 연관된 유의 수준이 있으며, 유의 수준이 적으면 한층 유의적 연결을 표시합니다. 지정된 값보다 큰 유의 수준의 관계는 숨길 것을 선택할 수 있습니다.

모델 적합 통계량 및 이상값

표시를 위해 선택한 목표 계열에 대한 이상값 및 모델 적합 통계의 테이블. 이 테이블은 전체 모델 시스템 시각화의 테이블과 같은 정보를 포함합니다. 이 테이블은 테이블 피벗 및 편집에 대한 모든 표준 기능을 지원합니다.

모델 효과 및 모델 모수

모델의 테이블은 표시를 위해 선택한 목표 계열에 대한 검정 및 모델 모수에 영향을 줍니다. 모델 효과 검정에는 모델에 포함된 각 입력에 대한 F 통계량 및 연관된 유의성 값이 포함됩니다.

영향 다이어그램

관심이 있는 계열과 이 계열이 영향을 주거나 이 계열에 영향을 주는 다른 계열 사이의 인과 관계의 그래픽 표현을 표시합니다. 관심 계열에 영향을 주는 계열을 원인이라고 합니다. **효과**를 선택하면 효과를 표시하기 위해 초기화된 영향 다이어그램이 생성됩니다. **원인**을 선택하면 원인을 표시하기 위해 초기화된 영향 다이어그램이 생성됩니다. **원인 및 효과 모두**을 선택하면 두 개의 개별 영향(으)로 생성됩니다(하나는 원인으로 초기화되고, 하나는 효과로 초기화됩니다. 영향 다이어그램을 표시하는 출력 항목에서 원인 및 효과 사이에 대화형으로 토글할 수 있습니다.

표시할 원인 또는 효과의 수준 수를 지정할 수 있습니다. 첫 번째 수준은 관심 계열입니다. 각각의 추가 수준은 관심 계열의 한층 간접적인 원인 또는 효과를 보여줍니다. 예를 들어, 효과의 표시에서 세 번째 수준은 직접 입력으로 두 번째 수준에 계열을 포함하는 계열로 구성됩니다. 세 번째 수준에 있는 계열은 관심 계열에 의해 간접적으로 영향을 받습니다. 관심 계열은 두 번째 수준에 있는 계열의 직접 입력입니다.

계열 도표

표시를 위해 선택되는 목표 계열의 관측 및 예측 값 도표. 예측이 요청될 때, 도표는 또한 예측에 대한 예측된 값과 신뢰구간을 표시합니다.

잔차도표

표시를 위해 선택되는 목표 계열에 대한 모델 잔차의 도표.

상위 입력

목표에 대한 상위 세 개의 입력과 함께, 시간이 경과하면서 표시되는 각 목표의 도표. 상위 입력은 유의성 값이 가장 낮은 입력입니다. 입력 및 목표에 대해 다른 척도를 수용하기 위해, y 축이 각 계열에 대한 z 스코어를 나타냅니다.

예측표

표시를 위해 선택되는 목표 계열에 대한 예측된 값 및 해당 예측의 신뢰구간 테이블.

이상값 근본 원인 분석

관심 계열에서 각 이상값의 원인될 가능성이 가장 큰 계열을 판별합니다. 이상값 근본 원인 분석은 **표시할 계열** 설정에 대해 개별 계열의 목록에 포함되는 각 목표 계열에 대해 수행됩니다.

출력

대화형 이상값 테이블 및 차트

각 관심 계열에 대한 이상값 및 해당 이상값의 근본 원인에 대한 테이블 및 차트. 테이블에는 이상값마다 하나의 행이 포함됩니다. 차트는 영향 다이어그램입니다. 테이블에서 행을 선택하면 영향 다이어그램에서, 관심 계열에서 연관된 가중값의 가장 큰 원인이 되는 계열까지의 경로가 강조 표시됩니다.

이상값의 피벗 테이블

각 관심 계열에 대한 이상값 및 해당 이상값의 근본 원인 테이블. 이 테이블에는 대화형 화면에 있는 테이블과 동일한 정보가 포함됩니다. 이 테이블은 테이블 피벗 및 편집에 대한 모든 표준 기능을 지원합니다.

인과 수준

근본 원인에 대한 검색에 포함할 수준 수를 지정할 수 있습니다. 여기에서 사용되는 수준의 개념은 영향 다이어그램에 대해 설명된 것과 동일합니다.

모든 모델에 대해 모델 적합

모든 모델 및 선택된 적합 통계량에 대한 모델 적합의 히스토그램. 다음 적합 통계량을 사용할 수 있습니다.

R 제곱

선형모델의 적합도 측도로서 결정계수라고도 합니다. 이 항목은 모델로 설명한 목표변수의 변동 비율이 됩니다. 값 범위는 0 - 1입니다. 값을 작을수록 모델이 데이터에 적합하지 않음을 의미합니다.

제곱근 평균 제곱 퍼센트 오차

모델 예측값이 계열의 관측값과 얼마나 다른지의 측도입니다. 사용된 단위와 상관이 없으므로 다른 단위의 계열을 비교하는 데 사용할 수 있습니다.

제곱근 평균제곱오차

평균 제곱 오차의 제곱근입니다. 종속 계열이 모델 예측 수준과 얼마나 다른지에 대한 측도로서, 종속 계열과 같은 단위로 표시됩니다.

BIC 베이지안 정보 기준. -2 축소 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. BIC도 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여"하지만 AIC보다 더 엄격하게 부여합니다.

AIC Akaike 정보 기준. -2 축소 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. AIC는 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여합니다".

시간 경과에 따른 이상값

추정 기간에서 각 시간 구간에 대한 모든 목표에서의 이상값 수 막대형 차트.

계열 변환

모델 시스템에서 계열에 적용된 변환의 테이블. 가능한 변환은 결측값 대체, 통합 및 분포입니다.

추정 기간

기본적으로 추정 기간은 모든 계열에 걸쳐 최초 관측값 시간에 시작되고 최근 관측값 시간에 종료됩니다.

시작 및 종료 시간 기준

추정 기간의 시작 및 종료 둘 다를 지정하거나 시작 또는 종료만 지정할 수 있습니다. 추정 기간의 시작 또는 종료를 생략하는 경우, 기본값이 사용됩니다.

- 날짜/시간 필드에 의해 관측값이 정의된 경우, 날짜/시간 필드에 사용되는 동일한 형식으로 시작 및 종료 값을 입력하십시오.
- 순환 주기에 의해 정의된 관측값의 경우, 순환 주기 필드마다 값을 지정하십시오. 각 필드는 별도의 열에 표시됩니다.

최근이거나 최초의 시간 간격(L)

선택적 오프셋으로, 데이터의 최초 시간 구간에 시작하거나 최근 시간 구간에 종료하는, 지정된 시간 구간 수로 추정 기간을 정의합니다. 이 컨텍스트에서, 시간 구간은 분석의 시간 구간을 가리킵니다. 예를 들어, 관측값이 매월 단위이지만 분석의 시간 구간은 분기일 수 있습니다. **최근**과 **시간 구간 수**로 24 값을 지정하면 최근 24개 분기를 의미합니다.

선택적으로, 지정된 시간 구간 수를 제외할 수 있습니다. 예를 들어, 최근 24 시간 구간을 지정하고 제외할 수로 1를 지정하면, 추정 기간은 마지막 구간 앞에 있는 24개 구간으로 구성됩니다.

모델 옵션

모델 이름

모델에 대한 사용자 정의 이름을 지정하거나 자동으로 생성되는 이름(*TCM*)을 승인할 수 있습니다.

예측 **레코드를 미래로 확장** 옵션은 추정 기간이 끝난 이후에 예측할 시간 구간 수를 설정합니다. 이 경우의 시간 구간은 데이터 지정 사항 탭에 지정된 분석의 시간 구간입니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다. 이 설정에 최대 한계는 없습니다.

대화형 출력

시간 인과 모델링의 출력에 대화형 출력 오브젝트 수가 포함됩니다. 출력 뷰어에서 오브젝트를 활성화(두 번 클릭)하여 대화형 기능을 사용할 수 있습니다.

전체 모델 시스템

모델 시스템에서 계열 사이의 인과 관계를 표시합니다. 해당 입력에 특수 목표를 연결하는 모든 선은 색상이 동일합니다. 이 선의 두께는 인과 연결의 유의성을 나타냅니다. 선이 두꺼울 수록 연결 유의성이 큽니다. 목표도 아닌 입력은 검은색 사각형으로 표시됩니다.

- 상위 모델, 지정된 계열, 모든 계열 또는 입력이 없는 모델에 대한 관계를 표시할 수 있습니다. 상위 모델은 **표시할 계열** 설정에 최적 적합 모델에 대해 지정된 기준에 맞는 모델입니다.
- 차트에서 계열 이름을 선택하고, 마우스 오른쪽 단추로 클릭한 후 컨텍스트 메뉴에서 **영향 다이어그램 생성**을 선택하여 하나 이상의 계열에 대한 영향 다이어그램을 생성할 수 있습니다.
- 지정된 값보다 큰 유의 수준을 가지고 있는 인과 관계는 숨길 것을 선택할 수 있습니다. 유의 수준이 작을수록 인과 관계는 한층 유의함을 나타냅니다.
- 차트에서 계열 이름을 선택하고, 마우스 오른쪽 단추로 클릭한 후 컨텍스트 메뉴에서 **계열에 대한 관계 강조 표시**를 선택하여 특정 계열에 대한 관계를 표시할 수 있습니다.

영향 다이어그램

관심이 있는 계열과 이 계열이 영향을 주거나 이 계열에 영향을 주는 다른 계열 사이의 인과 관계의 그래픽 표현을 표시합니다. 관심 계열에 영향을 주는 계열을 원인이라고 합니다.

- 원하는 계열의 이름을 지정하여 관심 계열을 변경할 수 있습니다. 영향 다이어그램에서 노드는 두 번 클릭하면 관심 계열이 해당 노드와 연관되는 계열로 변경됩니다.
- 원인 및 효과 사이의 화면을 토글하고 표시할 원인 또는 효과의 수준 수를 변경할 수 있습니다.
- 노드를 한 번 클릭하면 노드와 연관된 계열의 세부 순서도가 열립니다.

이상값 근본 원인 분석

관심 계열에서 각 이상값의 원인될 가능성이 가장 큰 계열을 판별합니다.

- 이상값 테이블에서 이상값에 대한 행을 선택하여 이상값에 대한 근본 원인을 표시할 수 있습니다. 순차도표에 이상값의 아이콘을 클릭하여 근본 원인을 표시할 수도 있습니다.
- 노드를 한 번 클릭하면 노드와 연관된 계열의 세부 순서도가 열립니다.

전체 모델 품질

특정 적합 통계량에 대한, 모든 모델의 모델 적합 히스토그램. 막대형 차트에서 막대를 클릭하면 선택된 막대와 연관되는 모델만 표시하도록 점도표를 필터링합니다. 계열 이름을 지정하여 점도표에서 특정 목표 계열에 대한 모델을 찾을 수 있습니다.

이상값 분포

추정 기간에서 각 시간 구간에 대한 모든 목표에서의 이상값 수 막대형 차트. 막대형 차트에서 막대를 클릭하면 선택된 막대와 연관되는 이상값만 표시하도록 점도표를 필터링합니다.

TCM 모델 너깃

TCM 모델링 작업은 다음 표에 표시된 대로, 접두문자 $TCM-를 사용하는 여러 새 필드를 작성합니다.

표 *27. TCM* 모델링 작업에서 작성한 새 필드

필드 이름	설명
$TCM-*colname*	각 목표 계열에 대한 모델에서 예측한 값.
$TCMLCI-*colname*	각 예측된 계열에 대한 하한 신뢰구간.
$TSUCI-*colname*	각 예측된 계열에 대한 상한 신뢰구간.
$TCMResidual-*colname*	생성된 모델 데이터의 각 열에 대한 잡음 잔차 값.

TCM 모델 너깃 설정

설정 탭에서는 TCM 모델 너깃에 대한 추가 옵션을 제공합니다.

예측

레코드를 미래로 확장 옵션은 추정 기간이 끝난 이후에 예측할 시간 구간 수를 설정합니다. 이 케이스에서 시간 구간은 TCM 노드의 데이터 지정 탭에 지정된 분석의 시간 구간입니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다.

스코어링에 사용 가능

각 모델의 스코어를 계산할 새 필드 만들기 각 모델의 스코어를 계산하기 위해 작성할 새 필드를 지정할 수 있습니다.

- **잡음 잔차** 이 옵션을 선택한 경우 각 목표 필드에서 모델 잔차와 이러한 값의 총계에 대한 새 필드(기본 접두문자가 $TCM-임)를 작성합니다.
- **신뢰 상한 및 하한** 이 옵션을 선택한 경우 각 목표 필드에서 각각 하한 및 하한 신뢰구간과 이러한 값의 총계에 대한 새 필드(기본 접두문자가 $TCM-임)를 작성합니다.

스코어링에 포함된 목표 모델 스코어에 포함할 수 있는 목표를 선택합니다.

시간 인과 모델 시나리오

시간 인과 모델 시나리오 프로시저는 활성 데이터 세트의 데이터를 사용하여 시간 인과 모델 시스템에 대해 사용자 정의 시나리오를 실행합니다. 시나리오는 지정된 시간 범위에 걸쳐 해당 계열에 대한 사용자 정의 값 세트 및 시계열(루트 계열이라고 하는)에 의해 정의됩니다. 지정된 값은 루트 계열에 의해 영향을 받는 시계열에 대해 예측을 생성하기 위해 사용됩니다. 프로시저에는 시간 인과 모델링 프로시저에 의해 작성된 모델 시스템 파일이 필요합니다. 활성 데이터 세트는 모델 시스템 파일을 작성하기 위해 사용된 것과 동일한 데이터인 것으로 가정합니다.

예제

시간 인과 모델링 프로시저를 사용하여, 비즈니스 의사결정자가 많은 중요한 성능 표시기에 영향을 주는 핵심 메트릭을 발견했습니다. 메트릭은 제어 가능하므로, 의사 결정자는 다음 사분기에서 메트릭에 대한 다양한 값 세트의 효과를 조사하려고 합니다. 조사는 시간 인과 모델 시나리오 프로시저에 모델 시스템 파일을 로드하고 핵심 메트릭에 대한 값 세트를 지정하여 쉽게 수행할 수 있습니다.

시나리오 주기 정의

시나리오 주기는 시나리오 실행에 사용되는 값을 지정하는 주기입니다. 추정 기간 종료 이전 또는 이후에 시작할 수 있습니다. 선택적으로, 시나리오 주기의 끝을 넘어서 예측할 경우 지정할 수 있습니다. 기본적으로, 시나리오 주기의 끝을 통과하여 예측이 생성됩니다. 모든 시나리오는 동일한 시나리오 주기 및 예측 거리 지정 사항을 사용합니다.

참고: 예측은 시나리오 주기의 시작 이후 첫 번째 시간 주기에서 시작합니다. 예를 들어, 시나리오 주기가 2014-11-01에 시작하고 시간 구간이 월일 경우, 첫 번째 예측은 2014-12-01에 수행됩니다.

시작, 종료 및 예측 시간 범위에 의해 지정됨

- 날짜/시간 필드에 의해 관측값이 정의된 경우, 날짜/시간 필드에 사용되는 동일한 형식으로 시작, 종료 및 예측 값을 입력하십시오. 날짜/시간 필드의 값은 연관된 시간 구간 맨 앞에 맞춰집니다. 예를 들어, 분석의 시간 구간이 월일 경우, 값 10/10/2014는 월의 맨 앞인 10/01/2014에 맞춰집니다.
- 순환 주기에 의해 정의된 관측값의 경우, 순환 주기 필드마다 값을 지정하십시오. 각 필드는 별도의 열에 표시됩니다.

추정 기간의 종료에 대해 상대적인 시간 간격에 의해 지정

추정 기간의 종료에 상대적인 시간 구간 수 측면에서 시작 및 종료를 정의합니다. 여기서 시간 구간은 분석의 시간 구간입니다. 추정 기간의 끝은 시간 구간 0으로 정의됩니다. 추정 기간 끝 이전의 시간 구간은 음수 값을 가지며, 추정 기간 끝 이후의 구간은 양수 값을 갖습니다. 시나리오 주기의 끝을 넘어서 예측하기 위한 구간 수를 지정할 수도 있습니다. 기본값은 0입니다.

예를 들어, 분석의 시간 구간이 월이고 시작 구간으로 1, 종료 구간으로 3, 끝을 지나 예측할 거리로 1을 지정한다고 가정합시다. 시나리오 주기는 추정 기간 끝 이후의 3개월입니다. 예측은 시나리오 주기의 두 번째 및 세 번째 달에 대해, 그리고 시나리오 주기 끝을 지나 추가 한 달동안 생성됩니다.

시나리오 및 시나리오 집단 추가

시나리오 탭은 실행할 시나리오를 지정합니다. 시나리오를 정의하려면, 먼저 **시나리오 주기 정의**를 클릭하여 시나리오 주기를 정의해야 합니다. 시나리오 및 시나리오 그룹(다차원 데이터에만 적용됨)은 연관된 **시나리오 추가** 또는 **시나리오 그룹 추가** 단추를 클릭하여 작성됩니다. 연관된 눈금에서 특수 시나리오 또는 시나리오 그룹을 선택하여, 편집, 복사 또는 삭제할 수 있습니다.

열 기반 데이터

눈금의 **루트 필드** 열은 값이 시나리오 값으로 바뀌는 시계열 필드를 지정합니다. **시나리오 값** 열은 가장 빠른 것으로 최근 순서로 지정된 시나리오 값을 표시합니다. 시나리오 값이 표현식으로 정의된 경우, 열은 표현식을 표시합니다.

다차원 데이터

개별 시나리오

개별 시나리오 눈금의 각 행은 값이 지정된 시나리오 값으로 바뀌는 시계열을 지정합니다. 계열은 각각의 차원 필드에 대해 지정된 값과 **루트 메트릭** 열에 지정된 필드의 조합에 의해 정의됩니다. **시나리오 값** 열의 내용은 열 기반 데이터와 같습니다.

시나리오 그룹

시나리오 그룹은 단일 루트 메트릭 필드를 기반으로 하는 하나의 시나리오 세트와 여러 차원 값 세트를 정의합니다. 지정된 메트릭 필드에 대한 각 차원 값 세트(차원 필드당 하나의 값)는 시계열을 정의합니다. 해당 값이 시나리오 값으로 바뀌는 이와 같은 시계열 각각에 대해 개별 시나리오가 생성됩니다. 시나리오 그룹에 대한 시나리오 값은 표현식으로 지정되며, 표현식은 그룹의 각 시계열에 적용됩니다.

계열 개수 열은 시나리오 그룹과 연관되는 차원 값 세트 수를 표시합니다. 표시된 값은 시나리오 그룹과 연관되는 실제 시계열 수(세트당 하나의 계열)보다 클 수 있습니다. 이 조건은 지정된 차원 값 조합 중 일부가 그룹에 대한 루트 메트릭에 의해 포함된 계열에 해당되지 않는 경우에 발생합니다.

시나리오 그룹의 예로서, 메트릭 필드 *advertising*과 두 개의 차원 필드 *region* 및 *brand*를 고려해 보십시오. *region* 및 *brand*의 모든 조합을 포함하고 루트 메트릭으로 *advertising*을 기반으로 하는 시나리오 그룹을 정의할 수 있습니다. *advertising* 필드와 연관되는 시계열 각각에 대해 20퍼센트씩 *advertising*을 증가시키는 효과를 조사하기 위한 표현식으로 `advertising*1.2`를 지정할 수도 있습니다. 네 개의 *region* 값과 두 개의 *brand* 값이 있는 경우, 8개의 시계열이 있으므로 그룹에 의해 8개의 시나리오가 정의됩니다.

시나리오 정의: 시나리오 정의를 위한 설정은 데이터가 열 기반 또는 다차원 데이터 여부에 따라 다릅니다.

루트 계열

시나리오에 대한 루트 계열을 지정합니다. 각 시나리오는 단일 루트 계열을 기반으로 합니다. 열 기반 데이터의 경우, 루트 계열을 정의하는 필드를 선택합니다. 다차원 데이터의 경우, 계열을 포함하는 메트릭 필드에 대한 눈금에 진입을 추가하여 루트 계열을 지정합니다. 그런 다음 루트 계열을 정의하는 차원 필드의 값을 지정합니다. 차원 값 지정에 다음 사항이 적용됩니다.

- 각 차원 필드의 값을 눈금에 직접 입력하거나 사용 가능한 차원 값 목록에서 선택할 수 있습니다. 사용 가능한 차원 값 목록에서 선택하려면 원하는 차원의 셀에서 생략 기호 단추를 클릭하십시오. 이 동작을 수행하면 차원 값 선택 하위 대화 상자가 열립니다.
- 쌍안경 아이콘을 클릭하고 검색어를 지정하여, 차원 값 선택 하위 대화 상자에서 차원 값 목록을 검색할 수 있습니다. 공백은 검색어의 일부로 처리됩니다. 검색어의 별표(*)는 와일드카드 문자를 표시하지 않습니다.

영향 목표 지정

루트 계열에 의해 영향을 받는 특정 목표를 알고 해당 목표에 대한 효과를 탐색하려면 이 옵션을 사용하십시오. 기본적으로, 루트 계열에 의해 영향을 받는 목표는 자동으로 결정됩니다. 옵션 탭의 설정으로 시나리오에 의해 영향을 받는 계열의 폭을 지정할 수 있습니다.

열 기반 데이터에 대해 원하는 목표를 선택하십시오. 다차원 데이터의 경우, 계열을 포함하는 목표 메트릭 필드에 대한 눈금에 진입을 추가하여 목표 계열을 지정합니다. 기본적으로, 지정된 메트릭 필드에 포함되는 모든 계열이 포함됩니다. 하나 이상의 차원 필드에 대해 포함된 값을 사용자 정의하여 포함된 계열 세트를 사용자 정의할 수 있습니다. 포함되는 차원 값을 사용자 정의하려면, 원하는 차원에 대한 생략 기호 단추를 클릭하십시오. 이 동작으로 차원 값 선택 대화 상자가 열립니다.

계열 개수 열(다차원 데이터에 대한)은 현재 연관된 목표 메트릭에 대해 지정된 차원 값 세트 수를 표시합니다. 표시된 값은 실제 영향을 받는 목표 계열 수(세트당 하나의 계열)보다 클 수 있습니다. 이 조건은 지정된 차원 값 조합 중 일부가 연관 목표 메트릭에 의해 포함된 계열에 해당되지 않는 경우에 발생합니다.

시나리오 ID

각 시나리오에는 고유한 식별자가 있어야 합니다. ID는 시나리오와 연관되는 출력에 표시됩니다. 식별자 값에 대해 유일성이 아닌 다른 제한사항은 없습니다.

루트 계열에 대한 시나리오 값 지정

시나리오 주기에서 루트 계열의 명시적 값을 지정하려면 이 옵션을 사용하십시오. 눈금에 나열되는 각 시간 구간에 대한 숫자 값을 지정해야 합니다. **읽기**, **예측** 또는 **Read\Forecast**를 클릭하여 시나리오 주기에서 각 구간에 대한 루트 계열(실제 또는 예측된)의 값을 확보할 수 있습니다.

루트 계열에 대한 시나리오 값의 표현식 지정

시나리오 주기에서 루트 계열의 값을 계산하기 위한 표현식을 정의할 수 있습니다. 직접적으로 표현식을 입력하거나 계산기 단추를 클릭하고 시나리오 값 표현식 작성기로부터 표현식을 작성할 수 있습니다.

- 표현식은 모델 시스템에 목표 또는 입력을 포함할 수 있습니다.
- 시나리오 주기가 기존 데이터를 넘어서 연장될 때, 표현식은 표현식에 있는 필드의 예측된 값에 적용됩니다.
- 다차원 데이터의 경우, 표현식의 각 필드는 루트 메트릭에 대해 지정된 필드 및 차원 값에 의해 정의되는 시계열을 지정합니다. 이는 표현식을 평가하기 위해 사용된 시계열입니다.

예로서, 루트 필드가 *advertising*이고 표현식이 `advertising*1.2`인 것으로 가정해 보십시오. 시나리오에서 사용되는 *advertising*의 값은 기존 값에서 20 퍼센트 증가를 나타냅니다.

참고: 시나리오는 시나리오 탭에서 **시나리오 추가**를 클릭하여 작성됩니다.

차원 값 선택: 다차원 데이터의 경우, 시나리오 또는 시나리오 그룹에 의해 영향을 받는 목표를 정의하는 차원 값을 사용자 정의할 수 있습니다. 또한 시나리오 그룹의 루트 계열 세트를 정의하는 차원 값을 사용자 정의할 수도 있습니다.

모든 값

현재 차원 필드의 모든 값이 포함됨을 지정합니다. 이 옵션은 기본값입니다.

값 선택(L)

현재 차원 필드의 값 세트를 지정하려면 이 옵션을 사용하십시오. 선택할 값 세트를 필터링할 수 있습니다. 필터 조건에 충족하는 값은 **매치됨** 탭에 나타나고, 필터 조건과 일치하지 않는 값은 **선택되지 않은 값** 목록의 **매치하지 않음** 탭에 나타납니다. **모두** 탭은 필터 조건과 관계없이 선택되지 않은 모든 값을 나열합니다.

- 필터를 지정할 때 와일드카드 문자를 표시하기 위해 별표(*)를 사용할 수 있습니다.
- 현재 필터를 지우려면, 표시된 값 필터링 대화 상자에 검색어로 비어 있는 값을 지정하십시오.

영향을 받는 목표의 차원 값을 사용자 정의하려면 다음을 수행하십시오.

1. 시나리오 정의 또는 시나리오 그룹 정의 대화 상자에서, 차원 값을 사용자 정의할 목표 메트릭을 선택하십시오.
2. 사용자 정의하려는 차원의 열에서 줄생략 기호 단추를 클릭하십시오.

시나리오 그룹의 루트 계열에 대한 차원 값을 사용자 정의하려면 다음을 수행하십시오.

1. 시나리오 그룹 정의 대화 상자에서, 사용자 정의할 차원의 생략 기호 단추(루트 계열 눈금에서)를 클릭하십시오.

시나리오 그룹 정의:

루트 계열

시나리오 그룹을 위한 루트 계열의 세트를 지정합니다. 개별 시나리오는 세트의 시계열마다 생성됩니다. 원하는 계열을 포함하는 메트릭 필드에 대한 눈금에 진입을 추가하여 루트 계열을 지정합니다. 그런 다음 세트를 정의하는 차원 필드의 값을 지정합니다. 기본적으로, 지정된 루트 메트릭 필드에 포함되는 모든 계열이 포함됩니다. 하나 이상의 차원 필드에 대해 포함된 값을 사용자 정의하여 포함된 계열 세트를 사용자 정의할 수 있습니다. 포함되는 차원 값을 사용자 정의하려면, 차원에 대한 생략 기호 단추를 클릭하십시오. 이 동작으로 차원 값 선택 대화상자가 열립니다.

계열 개수 열은 현재 연관 루트 메트릭에 대해 포함되는 차원 값 세트 수를 표시합니다. 표시된 값은 시나리오 그룹의 실제 루트 계열 수(세트당 하나의 계열)보다 클 수 있습니다. 이 조건은 지정된 차원 값 조합 중 일부가 루트 메트릭에 의해 포함된 계열에 해당되지 않은 경우에 발생합니다.

영향 목표 계열 지정

루트 계열 세트에 의해 영향을 받는 특정 목표를 알고 해당 목표에 대한 효과를 탐색하려면 이 옵션을 사용하십시오. 기본적으로, 각 루트 계열에 의해 영향을 받는 목표는 자동으로 결정됩니다. 옵션 탭의 설정으로 각각의 개별 시나리오에 의해 영향을 받는 계열의 폭을 지정할 수 있습니다.

계열을 포함하는 메트릭 필드에 대한 눈금에 진입을 추가하여 목표 계열을 지정합니다. 기본적으로, 지정된 메트릭 필드에 포함되는 모든 계열이 포함됩니다. 하나 이상의 차원 필드에 대해 포함된 값을 사용자 정의하여 포함된 계열 세트를 사용자 정의할 수 있습니다. 포함되는 차원 값을 사용자 정의하려면, 원하는 차원에 대한 생략 기호 단추를 클릭하십시오. 이 동작으로 차원 값 선택 대화 상자가 열립니다.

계열 개수 열은 현재 연관된 목표 메트릭에 대해 지정된 차원 값 세트 수를 표시합니다. 표시된 값은 실제 영향을 받는 목표 계열 수(세트당 하나의 계열)보다 클 수 있습니다. 이 조건은 지정된 차원 값 조합 중 일부가 연관 목표 메트릭에 의해 포함된 계열에 해당되지 않는 경우에 발생합니다.

시나리오 ID 접두문자

각 시나리오 그룹에는 고유한 집단이 있어야 합니다. 접두문자는 시나리오 그룹에 각각의 개별 시나리오와 연관된 출력에 표시되는 ID를 구성하기 위해 사용됩니다. 개별 시나리오에 대한 ID는 접두문자이며, 루트 계열을 식별하는 각 차원 필드의 값이 뒤에 밑줄로 붙습니다. 차원 값은 밑줄로 구분됩니다. 접두문자의 값에 대해 유일성이 아닌 다른 제한사항은 없습니다.

루트 계열에 대한 시나리오 값의 표현식

시나리오 그룹에 대한 시나리오 값은 표현식으로 지정되며, 표현식은 그룹의 각 루트 계열의 값을 계산하기 위해 사용됩니다. 직접적으로 표현식을 입력하거나 계산기 단추를 클릭하고 시나리오 값 표현식 작성기로부터 표현식을 작성할 수 있습니다.

- 표현식은 모델 시스템에 목표 또는 입력을 포함할 수 있습니다.

- 시나리오 주기가 기존 데이터를 넘어서 연장될 때, 표현식은 표현식에 있는 필드의 예측된 값에 적용됩니다.
- 그룹에서 각 루트 계열에 대해, 표현식의 필드는 루트 계열을 정의하는 해당 필드 및 차원 값에 의해 정의되는 시계열을 지정합니다. 이는 표현식을 평가하기 위해 사용된 시계열입니다. 예를 들어, 루트 계열이 `region='north'` 및 `brand='X'`에 의해 정의되는 경우, 표현식에 사용되는 시계열은 동일한 차원 값에 의해 정의됩니다.

예제로서, 루트 메트릭 필드가 *advertising*이고 두 개의 차원 필드 *region* 및 *brand*가 있다고 가정해 보십시오. 또한, 시나리오 그룹에 차원 필드 값의 모든 조합이 포함된다고 가정합니다. *advertising* 필드와 연관되는 시계열 각각에 대해 20 퍼센트씩 *advertising*을 증가시키는 효과를 조사하기 위한 표현식으로 `advertising*1.2`를 지정할 수도 있습니다.

참고: 시나리오 그룹은 다차원 데이터에만 적용되고, 시나리오 탭에서 **시나리오 그룹 추가**를 클릭하여 작성됩니다.

옵션

영향 목표의 최대 수준

영향을 받는 목표의 최대 수준 수를 지정합니다. 각각의 연속 수준(최대 5까지)에는 루트 계열에 의해 한층 간접적으로 영향을 받는 목표가 포함됩니다. 특히, 첫 번째 수준에는 직접 입력으로 루트 계열을 가지고 있는 목표가 포함됩니다. 두 번째 수준의 목표에는 직접 입력으로 첫 번째 수준에 있는 목표가 포함됩니다. 그 뒤로도 마찬가지입니다. 이 설정의 값을 증가시키면 계산의 복잡도가 증가하여 성능에 영향을 줄 수 있습니다.

자동 검색된 최대 목표

자동으로 각 루트 계열에 대해 발견되는, 영향을 받는 최대 목표 수를 지정합니다. 이 설정의 값을 증가시키면 계산의 복잡도가 증가하여 성능에 영향을 줄 수 있습니다.

영향 다이어그램

각 시나리오의 루트 계열과 영향을 미치는 목표 계열 사이의 인과 관계에 대한 그래픽 표현을 표시합니다. 시나리오 값과 영향을 받은 목표의 예측값 둘 다에 대한 테이블이 출력 항의 일부로 포함됩니다. 그래프에는 영향을 받는 목표의 예측값 도표가 포함됩니다. 영향 다이어그램에서 노드를 한 번 클릭하면 노드와 연관된 계열의 세부 순서도가 열립니다. 시나리오마다 별도의 영향 다이어그램이 생성됩니다.

계열 도표

각 시나리오에서 영향을 받는 목표 각각에 대한 예측값 계열 도표가 생성됩니다.

예측 및 시나리오 테이블

각 시나리오에 대한 시나리오 값 및 예측값의 테이블. 이 테이블에는 영향 다이어그램에 있는 테이블과 같은 정보가 포함됩니다. 이 테이블은 테이블 피벗 및 편집에 대한 모든 표준 기능을 지원합니다.

도표 및 테이블에 신뢰구간 포함

시나리오 예측에 대한 신뢰구간이 차트 및 테이블 출력 둘 다에 포함되는지 여부를 지정합니다.

신뢰구간 너비(%)

이 설정은 시나리오 예측에 대한 신뢰구간을 제어합니다. 100보다 작은 양수 값을 지정할 수 있습니다. 기본적으로, 95% 신뢰구간이 사용됩니다.

시계열 노드

시계열 노드는 로컬 또는 분산 환경의 데이터와 함께 사용할 수 있으며 분산 환경에서는 IBM SPSS Analytic Server의 기능을 이용할 수 있습니다. 이 노드를 사용하여 시계열에 대한 지수평활, 일변량 ARIMA(Autoregressive Integrated Moving Average) 또는 다변량 ARIMA(또는 전이 함수) 모델을 추정하고 시계열 데이터에 기반하여 예측을 생성합니다.

지수평활은 향후 값을 예측하기 위해 이전 계열 관측의 가중된 값을 사용하는 시계열 분석 방법입니다. 이와 같이 지수평활은 데이터의 이론적 이해에 기반하지 않습니다. 새 데이터가 들어오면 해당 예측을 조정하여 한 번에 하나의 포인트를 예측합니다. 이 방법은 추세, 계절성 또는 모두를 나타내는 계열의 시계열 분석에 유용합니다. 추세 및 계절성을 다르게 처리하는 다양한 지수평활 모델 중에서 선택할 수 있습니다.

ARIMA 모델은 지수평활 모델을 수행하는 모델링 추세 및 계절 구성요소에 대해 보다 정교한 방법을 제공하고, 특히 모델의 독립 변수(예측변수)의 추가된 혜택을 누릴 수 있습니다. 여기에는 차이 정도와 함께 자기회귀 및 이동 평균 순서를 명시적으로 지정하는 작업이 포함됩니다. 예측변수를 포함하고 이 모든 항목에 대해 전이 함수를 정의하고, 이상값 또는 명시적 이상값 세트의 자동 발견을 지정할 수 있습니다.

참고: 실제로 메일로 보내는 카탈로그 수 또는 회사 웹 페이지의 적중 수와 같은 예측할 계열의 동작을 설명하는 데 도움이 될 수 있는 예측변수를 포함시키려는 경우 ARIMA 모델이 가장 유용합니다. 지수평활 모델에서는 왜 원래대로 작동하는지 이유를 이해하지 않고도 시계열 동작을 설명합니다. 예를 들어, 12개월마다 히스토리에서 최고치를 기록하는 계열은 이유는 알지 못해도 계속해서 이러한 수치를 보일 수 있습니다.

자동 모델 생성기 옵션도 사용할 수 있습니다. 이 경우 하나 이상의 목표 변수에 대해 최적합 ARIMA 또는 지수평활 모델을 자동으로 식별 및 추정하려고 하므로, 시행착오를 통해 적절한 모델을 식별하지 않아도 됩니다. 의심되는 경우 자동 모델 생성기 옵션을 사용하십시오.

예측자 변수가 지정된 경우 자동 모델 생성기는 ARIMA 모델에 포함되도록 종속 계열과 통계적으로 중요한 관계에 있는 해당 변수를 선택합니다. 모델 변수는 차이 및/또는 제곱근이나 자연 로그 변환을 사용하여 적절한 위치에서 변환됩니다. 기본적으로 자동 모델 생성기는 모든 지수평활 모델과 모든 ARIMA 모델을 고려하며, 각 목표 필드에서 이들 중 최상의 모델을 선택합니다. 그러나 지수평활 모델 중 최상의 항목을 선택하거나 ARIMA 모델 중 최상의 항목만을 선택하도록 자동 모델 생성기를 제한할 수 있습니다. 또한 이상값 자동 발견을 지정할 수 있습니다.

시계열 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, 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 수동으로 대상, 예측변수 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. 예측의 대상으로 하나 이상의 필드를 선택하십시오.

후보 입력. 예측의 입력으로 하나 이상의 필드를 선택하십시오.

이벤트 및 개입 이 영역을 사용하여 특정 입력 필드를 이벤트 또는 개입 필드로 지정하십시오. 이와 같이 지정하면 이벤트(판매 프로모션과 같은 예측 가능한 반복 상황) 또는 개입(정전 또는 직원 파업과 같은 일회성 사건)의 영향을 받을 수 있는 시계열 데이터를 포함하는 항목으로 필드를 식별합니다.

시계열 노드 - 데이터 지정 사항 옵션

데이터 지정 사항 탭에서는 모델에 포함될 데이터에 대한 모든 옵션을 설정할 수 있습니다. **날짜/시간 필드**와 **시간 간격**을 모두 지정하는 경우에만 **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.

시계열 노드 - 관측값

이 분할창의 설정을 사용하여 관측값을 정의하는 필드를 지정할 수 있습니다.

날짜/시간 필드로 지정된 관측값

관측값이 날짜, 시간 또는 시간소인 필드를 통해 정의되도록 지정할 수 있습니다. 관측값을 정의하는 필드 외에, 관측값을 설명하는 적절한 시간 구간을 선택하십시오. 지정된 시간 구간에 따라, 관측값(증분) 사이의 구간이나 주당 일 수와 같은 다른 설정을 지정할 수도 있습니다. 다음 고려사항은 시간 간격에 적용됩니다.

- 관측값이 시간에서 비정규적으로 간격이 있는 경우(판매 순서가 처리되는 시간과 같이), **비정규** 값을 사용하십시오. **비정규**가 선택될 때, 데이터 지정 사항 탭의 **시간 간격** 설정에서 분석에 사용되는 시간 구간을 지정해야 합니다.

- 관측값이 날짜와 시간을 나타내고 시간 구간이 시, 분 또는 초인 경우 **하루 중 시간(시)**, **하루 중 시간(분)** 또는 **하루 중 시간(초)**을 사용하십시오. 관측값이 날짜에 대한 참조 없이 시간(기간)을 나타내고 시간 구간이 시, 분 또는 초일 경우, **시(비주기적)**, **분(비주기적)** 또는 **초(비주기적)**를 사용하십시오.
- 선택된 시간 간격을 기초로, 프로시저는 결측 관측값을 발견할 수 있습니다. 프로시저에서는 모든 관측값이 시간에서 동일하게 간격을 두고 결측 관측값이 없다고 가정하므로, 결측 관측값을 발견해야 합니다. 예를 들어, 시간 간격이 일(Days)이고 날짜 2015-10-27 뒤에 2015-10-29가 있는 경우, 2015-10-28에 대해 결측 관측값이 있습니다. 결측 관측값에 대해 값이 대체됩니다. 데이터 지정 사항 탭의 **결측값 처리** 영역에서 결측값 처리 설정을 지정하십시오.
- 지정된 시간 구간은 프로시저가 함께 통합해야 하는 동일한 시간 구간의 여러 관측값을 발견하고 관측값에 동일하게 간격이 있도록 월의 첫 번째와 같은 구간 경계에 관측값을 맞출 수 있도록 합니다. 예를 들어, 시간 구간이 월일 경우, 동일 월에 있는 여러 날짜가 함께 통합됩니다. 이 유형의 통합을 그룹화라고 합니다. 기본적으로, 관측값은 그룹화될 때 합산됩니다. 데이터 지정 사항 탭의 **통합 및 분포** 설정에서, 그룹화에 다른 방법(예: 관측값의 평균)을 지정할 수 있습니다.
- 일부 시간 구간의 경우, 추가 설정은 동일하게 간격이 있는 정규 구간에서 중단을 설정할 수 있습니다. 예를 들어, 시간 구간이 일(Days)이지만 평일만 유효한 경우, 주에 5일이 있고 주는 월요일에 시작함을 지정할 수 있습니다.

관측값이 주기 또는 순환 주기로 정의됨

관측값은 임의의 순환 수준 수까지, 주기 또는 반복 주기 순환을 나타내는 하나 이상의 정수 필드로 정의할 수 있습니다. 이 구조를 사용할 경우 표준 시간 구간 중 하나에 맞지 않은 관측값 계열을 기술할 수 있습니다. 예를 들어, 10개월만 있는 회계연도는 연도를 나타내는 순환 필드와, 월을 나타내는 주기 필드로 설명할 수 있습니다. 여기서 하나의 주기 길이는 10입니다.

순환 주기를 지정하는 필드는 주기적 수준의 계층 구조를 정의합니다. 가장 낮은 수준은 **주기** 필드에 의해 정의됩니다. 다음 최상위 수준은 수준이 1인 순환 필드에 의해 지정되고, 그 다음은 수준 2의 순환 필드로 지정되며 뒤로도 마찬가지입니다. 가장 높은 수준을 제외하고, 각 수준의 필드 값은 다음 최상위 수준에 관하여 주기적이어야 합니다. 최상위 수준의 값은 주기적이 될 수 없습니다. 예를 들어, 10달 회계연도의 경우 월은 연도 내에서 주기적이며 연도는 주기적이 아닙니다.

- 특정 수준에 있는 순환의 길이는 다음으로 가장 낮은 수준의 주기성입니다. 회계연도 예의 경우, 단 하나 순환 수준이 있고 순환 길이는 10입니다. 다음으로 가장 낮은 수준이 월을 나타내고 지정된 회계 연도에 10달이 있기 때문입니다.
- 주기적 필드의 시작 값(1부터 시작하지 않음)을 지정하십시오. 이 설정은 결측값을 발견하는 데 필요합니다. 예를 들어, 주기적 필드는 2에서 시작하지만 시작 값은 1로 지정되는 경우, 프로시저는 해당 필드의 각 순환에 있는 첫 번째 주기에 대해 결측값이 있다고 가정합니다.

시계열 노드 - 분석 시간 구간

분석에 사용할 시간 구간은 관측값의 시간 구간과 다를 수 있습니다. 예를 들어, 관측값의 시간 구간이 일(Days)일 경우, 분석의 시간 구간으로는 월을 선택할 수 있습니다. 그런 다음 모델이 작성되기 전에

매일 데이터에서 매월 데이터까지 데이터가 통합됩니다. 또한 데이터를 장기 시간 구간에서 단기 시간 구간으로 분포할 것을 선택할 수도 있습니다. 예를 들어, 관측값이 분기별인 경우, 데이터를 분기별에서 월별 데이터로 분포할 수 있습니다.

이 분할창의 설정을 사용하여 분석 시간 구간을 지정할 수 있습니다. 데이터가 통합되거나 분포되는 방법은 데이터 지정 사항 탭의 **통합 및 분포** 설정에서 지정됩니다.

분석이 행해지는 시간 구간에 대해 사용 가능한 선택은 해당 관측값 정의 방법과 관측값의 시간 구간에 따라 다릅니다. 특히, 관측값이 순환 주기로 정의될 경우 통합만 지원됩니다. 그러한 경우, 분석의 시간 구간은 관측값의 시간 구간보다 크거나 같아야 합니다.

시계열 노드 - 통합 및 분포 옵션

이 분할창의 설정을 사용하여 관측값의 시간 구간과 관련한 입력 데이터 통합 또는 분포 설정을 지정할 수 있습니다.

통합 함수

분석에 사용되는 시간 구간이 관측에 사용되는 시간 구간보다 길 경우, 입력 데이터는 통합됩니다. 예를 들어, 관측값의 시간 구간이 일(Days)이고 분석의 시간 구간이 월일 경우 통합이 수행됩니다. mean, sum, mode, min 또는 max 통합 함수를 사용할 수 있습니다.

분포 함수

분석에 사용되는 시간 구간이 관측의 시간 구간보다 짧을 경우, 입력 데이터는 분포됩니다. 예를 들어, 관측값의 시간 구간이 분기이고 분석의 시간 구간이 월일 경우 분포가 수행됩니다. mean 또는 sum 분포 함수를 사용할 수 있습니다.

그룹화 함수

그룹화는 관측값이 날짜/시간에 의해 정의되고 여러 관측값이 동일 시간 구간에 발생하는 경우에 적용됩니다. 예를 들어, 관측값의 시간 구간이 월일 경우, 동일 월에 있는 여러 날짜가 그룹화되어 날짜가 발생하는 월과 연관됩니다. mean, sum, mode, min 또는 max와 같은 그룹화 함수를 사용할 수 있습니다. 다음 그룹화는 항상 관측값이 날짜/시간에 의해 정의되고 관측값의 시간 간격이 비정규로 지정된 경우에 수행됩니다

참고: 그룹화가 통합 양식이어도, 그룹화는 결측값 처리 이전에 수행됩니다(정상 통합은 결측값 처리 이후에 수행됩니다). 관측값의 시간 구간이 비정규로 지정되는 경우, 통합은 그룹화 함수로만 수행됩니다.

교차-일 관측값을 이전 일로 통합

1일 경계를 교차하는 시간을 사용하는 관측값이 전날의 값에 통합되는지 여부를 지정합니다. 예를 들어, 20:00시에 시작하는 8시간 노동의 시간별 관측값의 경우, 이 설정은 00:00 및 04:00 사이의 관측값이 전날 통합 결과에 포함되는지 여부를 지정합니다. 이 설정은 관측값의 시간 구간이 하루 중 시간(시), 하루 중 시간(분) 또는 하루 중 시간(초)이고 분석의 시간 구간이 일(Days)인 경우에만 적용됩니다.

지정된 필드에 대한 사용자 정의 설정

필드 기준으로 필드에 통합, 분포 및 그룹화 함수를 지정할 수 있습니다. 이 설정은 통합, 분포 및 그룹화 함수에 대한 기본 설정을 대체합니다.

시계열 노드 - 결측값 옵션

이 분할창의 설정을 사용하여 입력 데이터의 결측값을 대체값으로 바꾸는 방법을 지정할 수 있습니다. 다음 방법으로 바꿀 수 있습니다.

선형 보간법

선형 보간법을 사용하여 결측값을 바꿉니다. 결측값 이전의 마지막 유효한 값과 결측값 이후의 첫 번째 유효한 값이 보간법에 사용됩니다. 계열에서 첫 번째 또는 마지막 관측값에 결측값이 있는 경우, 계열의 시작 또는 종료에서 두 개의 가장 근접한 비결측 값이 사용됩니다.

계열 평균

결측값을 전체 계열에 대한 평균으로 바꿉니다.

근접한 값들의 평균

결측값을 유효한 근접 값의 평균으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

근접한 값들의 중앙값

결측값을 근접한 유효한 값의 중앙값으로 바꿉니다. 근접한 값들의 계산너비는 평균을 계산하는데 사용되는 결측값 전후의 유효값 수입니다.

선형 추세

이 옵션은 단순 선형 회귀 모델 적합을 위해 계열의 모든 비결측 관측값을 사용합니다. 그런 다음 결측값을 대치하는 데 사용됩니다 .

기타 설정:

최저 데이터 품질 스코어(%)

각 시계열에 대한 시간 변수 및 입력 데이터에 대한 데이터 품질 측도를 계산합니다. 데이터 품질 스코어가 이 임계값보다 낮은 경우 해당 시계열을 버립니다.

시계열 노드 - 추정 기간

추정 기간 분할창에서 모형 추정에 사용될 레코드의 범위를 지정할 수 있습니다. 기본적으로 추정 기간은 모든 계열에 걸쳐 최초 관측값 시간에 시작되고 최근 관측값 시간에 종료됩니다.

시작 및 종료 시간 기준

추정 기간의 시작 및 종료 둘 다를 지정하거나 시작 또는 종료만 지정할 수 있습니다. 추정 기간의 시작 또는 종료를 생략하는 경우, 기본값이 사용됩니다.

- 날짜/시간 필드에 의해 관측값이 정의된 경우, 날짜/시간 필드에 사용되는 동일한 형식으로 시작 및 종료 값을 입력하십시오.
- 순환 주기에 의해 정의된 관측값의 경우, 순환 주기 필드마다 값을 지정하십시오. 각 필드는 별도의 열에 표시됩니다.

최근이거나 최초의 시간 간격(L)

선택적 오프셋으로, 데이터의 최초 시간 구간에 시작하거나 최근 시간 구간에 종료하는, 지정된 시간 구간 수로 추정 기간을 정의합니다. 이 컨텍스트에서, 시간 구간은 분석의 시간 구간을 가리킵니다. 예를 들어, 관측값이 매월 단위이지만 분석의 시간 구간은 분기일 수 있습니다. **최근**과 **시간 구간 수**로 24 값을 지정하면 최근 24개 분기를 의미합니다.

선택적으로, 지정된 시간 구간 수를 제외할 수 있습니다. 예를 들어, 최근 24 시간 구간을 지정하고 제외할 수로 1를 지정하면, 추정 기간은 마지막 구간 앞에 있는 24개 구간으로 구성됩니다.

시계열 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.

이 탭은 해당 모델에만 적용되는 사용자 정의를 설정하는 여러 개의 분할창으로 구성됩니다.

시계열 노드 - 일반 작성 옵션

이 분할창에서 사용 가능한 옵션은 **방법** 목록에서 선택한 다음 세 가지 설정에 따라 다릅니다.

- **자동 모델 생성기** 자동 모델 생성기를 사용하려면 이 옵션을 선택합니다. 그러면 각 종속 계열에 대한 최적 적합 모델을 자동으로 찾습니다.
- **지수평활** 이 옵션을 사용하여 사용자 정의 지수평활 모델을 지정합니다.
- **ARIMA** 이 옵션을 사용하여 사용자 정의 ARIMA 모델을 지정합니다.

자동 모델 생성기

모델 유형에서 작성하려는 모델의 유형을 선택하십시오.

- **모든 모델** 자동 모델 생성기에서 ARIMA 및 지수평활 모델을 모두 고려합니다.
- **지수평활 모델만.** 자동 모델 생성기에서 지수평활 모델만 고려합니다.
- **ARIMA 모델만** 자동 모델 생성기에서 ARIMA 모델만 고려합니다.

자동 모델 생성기에서 계절 모델 고려 이 옵션은 활성 데이터 세트에 대해 주기성이 정의된 경우에만 사용할 수 있습니다.이 옵션을 선택하면 자동 모델 생성기는 계절 및 비계절 모델을 모두 고려합니다. 이 옵션을 선택하지 않은 경우 자동 모델 생성기에서 비계절 모델만 고려합니다.

자동 모델 생성기에서 정교한 지수평활 모델 고려 이 옵션을 선택하면 자동 모델 생성기가 총 13개의 지수평활 모델(7개는 원래 시계열 노드에 있고 6개는 버전 18.1에서 추가됨)을 검색합니다. 이 옵션을 선택하지 않으면 자동 모델 생성기가 원래 7개의 지수평활 모델만 검색합니다.

이상치에서 다음 옵션 중 하나를 선택하십시오.

자동으로 이상값 발견 기본적으로 이상값 자동 발견을 수행하지 않습니다. 이상값 자동 발견을 수행하려면 이 옵션을 선택하고 원하는 이상값 유형을 선택합니다.

입력 필드는 이 목록에 포함되기 전에 플래그, 명목 또는 순서와 같은 측정 수준이 있어야 하며 숫자여야 합니다(예: 플래그 필드의 경우 참/거짓이 아닌 1/0).

자동 모델 생성기는 **필드** 탭의 이벤트 또는 개입 필드로 식별된 입력에 대해 임의 전이 함수가 아닌 단순 회귀분석만 고려합니다.

지수평활

모델 유형 지수평활 모델은 계절 또는 비계절로 분류됩니다.[1] 계절 모델은 데이터 지정 사항 탭의 시간 간격 분할창을 사용하여 정의된 주기성이 계절인 경우에만 사용할 수 있습니다. 계절 주기성은 다음과 같습니다. 주기적 기간, 연도, 분기, 월, 한 주의 요일, 하루의 시간, 하루의 분, 하루의 초. 다음 모델 유형을 사용할 수 있습니다.

- **단순** 이 모델은 추세나 계절성이 없는 계열에 적합합니다. 유일하게 관련된 평활 모수는 수준입니다. 단순 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 1, 이동 평균 차수가 1, 및 상수 없음인 ARIMA와 가장 비슷합니다.
- **Holt의 선형 추세** 이 모델은 선형 추세가 있고 계절성이 없는 계열에 적합합니다. 관련 평활 모수는 수준 및 추세이며, 이 모델에서는 서로의 값으로 제한되지 않습니다. Holt 모델은 Brown 모델보다 일반적이지만 대형 계열의 추정값 계산에는 더 오래 걸립니다. Holt 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 2, 이동 평균 차수가 2인 ARIMA와 가장 비슷합니다.
- **진폭감소 추세** 이 모델은 점점 소멸되는 선형 추세가 있고 계절성이 없는 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 진폭감소 추세입니다. 진폭감소 지수평활은 자동 선형회귀 차수가 1, 차이 차수가 1, 이동 평균 차수가 2인 ARIMA와 가장 비슷합니다.
- **승법 추세** 이 모델은 계열의 규모가 변경되는 추세가 있고 계절성이 없는 계열에 적합합니다. 관련된 평활 모수는 수준과 추세입니다. 승법 추세 지수평활은 ARIMA 모델과 다릅니다.
- **Brown의 선형 추세** 이 모델은 선형 추세가 있고 계절성이 없는 계열에 적합합니다. 관련 평활 모수는 수준 및 추세지만, 이 모델에서는 동일하다고 가정합니다. 따라서 Brown의 모형은 Holt 모형의 특별한 케이스입니다. Brown의 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 2, 이동 평균 차수가 2이며, 이동 평균의 두 번째 차수에 대한 계수가 첫 번째 차수에 대한 계수의 절반과 같은 ARIMA와 가장 비슷합니다.
- **단순 계절** 이 모델은 추세와 계절 효과가 없고 시간에 따라 일정한 계열에 적합합니다. 관련된 평활 모수는 수준과 계절입니다. 계절 지수평활은 자동 선형회귀 차수가 0, 차이 차수가 1, 계절 차이 차수가 1, 이동 평균의 경우 차수 1, p, p+1인 ARIMA와 가장 유사합니다. 여기서 p는 계절 간격에서 기간 수입니다. 월별 데이터의 경우 p = 12입니다.
- **Winters의 가법** 이 모델은 선형 추세와 계절 효과가 있고 시간에 따라 일정한 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 계절입니다. Winters의 가법 지수평활은 자동 선형회귀 차수가 0,

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

차이 차수가 1, 계절 차이 차수가 1, 이동 평균의 경우 차수가 p+1인 ARIMA와 가장 유사합니다. 여기서 p는 계절 간격에서 기간 수입니다. 월별 데이터의 경우 p = 12입니다.

- **가법 계절이 있는 진폭감소 추세** 이 모델은 감소되는 선형 추세와 계절 효과가 있고 시간에 따라 일정한 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 진폭감소 추세 및 계절입니다. 진폭감소 추세 및 가법 계절 지수평활은 ARIMA 모델과 다릅니다.
- **가법 계절이 있는 승법 추세** 이 모델은 계열의 규모가 변경되는 추세와 계절 효과가 있고 시간에 따라 일정한 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 계절입니다. 승법 추세 및 가법 계절 지수평활은 ARIMA 모델과 다릅니다.
- **승법 계절성** 이 모델은 선형 추세와 계절 효과가 없고 계절 규모가 변경되는 계열에 적합합니다. 관련된 평활 모수는 수준과 계절입니다. 승법 계절 지수평활은 ARIMA 모델과 다릅니다.
- **Winters의 승법** 이 모델은 선형 추세와 계절 효과가 있고 계절 규모가 변경되는 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 계절입니다. Winters의 승법 지수평활은 ARIMA 모델과 다릅니다.
- **승법 계절이 포함된 진폭감소 추세.** 이 모델은 감소되는 선형 추세와 계절 효과가 있고 계절 규모가 변경되는 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 진폭감소 추세 및 계절입니다. 진폭감소 추세 및 승법 계절 지수평활은 ARIMA 모델과 다릅니다.
- **승법 계절이 있는 승법 추세** 이 모델은 추세와 계절 효과가 있고 둘 다 계절 규모와 함께 변경되는 계열에 적합합니다. 관련된 평활 모수는 수준, 추세, 계절입니다. 승법 추세 및 승법 계절 지수평활은 ARIMA 모델과 다릅니다.

목표 변환 모델링하기 전에 각 종속 변수에서 수행할 변환을 지정할 수 있습니다.

- **없음** 변환을 수행하지 않습니다.
- **제곱근** 제곱근 변환을 수행합니다.
- **자연로그** 자연로그 변환을 수행합니다.

ARIMA

사용자 정의 ARIMA 모델의 구조를 지정하십시오.

ARIMA 차수 눈금의 해당 셀에 모델의 다양한 ARIMA 구성요소 값을 입력합니다. 모든 값은 0또는 음이 아닌 정수여야 합니다. 자기회귀 및 이동 평균 구성요소의 경우 값은 최대 차수를 나타냅니다. 더 낮은 양의 차수가 모두 모델에 포함됩니다. 예를 들어, 2를 지정하면 모델에 차수 2와 1이 포함됩니다. 계절 열의 셀은 활성 데이터 세트에 대해 주기성이 정의된 경우에만 사용할 수 있습니다.

- **자기회귀(p)** 모델의 자기회귀 차수 수입니다. 자기회귀 차수는 계열의 이전 값 중 현재 값 예측에 사용될 값을 지정합니다. 예를 들어, 자기회귀 차수 2는 과거 2개 시간 주기의 계열 값을 현재 값 예측에 사용하도록 지정합니다.

- **차이(d)** 모델을 추정하기 전 계열에 적용할 차이 차수를 지정합니다. 추세가 존재하며(추세가 있는 계열은 일반적으로 비정상이며 ARIMA 모델링은 정상성을 가정) 해당 효과 제거를 위해 사용되는 경우 차이가 필요합니다. 차이 차수는 계열 추세 수준에 해당합니다. 1차 차이는 선형 추세, 2차 차이는 2차 추세 등을 나타냅니다.
- **이동 평균(q)** 모델의 이동 평균 차수 수입니다. 이동 평균 차수는 이전 값에 대한 계열 평균 편차를 사용하여 현재 값을 예측하는 방법을 지정합니다. 예를 들어, 이동 평균 차수 1과 2는 계열의 현재 값을 예측하는 경우 지난 2개 시간 주기 각각의 계열 평균값 편차를 고려하도록 지정합니다.

계절 계절 자기회귀, 이동 평균, 차이 구성요소는 해당 비계절 구성요소와 동일한 역할을 합니다. 그러나 계절 차수의 경우 현재 계열 값이 한 개 이상의 계절 주기에 의해 구분된 이전 계열 값의 영향을 받습니다. 예를 들어, 월별 데이터(계절 주기 12)의 경우 계절 차수 1은 현재 계열 값이 현재 계열 이전의 계열 값 12 주기의 영향을 받는다는 것을 의미합니다. 월별 데이터의 경우 계절 차수 1은 비계절 차수 12를 지정하는 것과 동일합니다.

자동으로 이상값 발견 이상값 자동 발견을 수행하려면 이 옵션을 선택하고 사용 가능한 하나 이상의 이상값 유형을 선택합니다.

발견할 이상값 유형 발견할 이상값 유형을 선택합니다. 지원되는 유형은 다음과 같습니다.

- 가법(기본값)
- 수준 이동(기본값)
- 혁신적
- 일시적
- 계절 가법
- 국소적 추세
- 가법 수정

전이 함수 차수 및 변환 변환을 지정하고 ARIMA 모델의 일부 또는 전체 입력 모델에 대해 전이 함수를 정의하려면 **설정**을 클릭하십시오. 전이 및 변환 세부사항을 입력할 수 있는 별도의 대화 상자가 표시됩니다.

모델에 상수 포함 전체 평균 계열 값이 0인지 잘 모르는 경우 상수를 포함하는 것이 표준입니다. 차이를 적용하는 경우에는 상수를 제외하는 것이 좋습니다.

추가 세부사항

- 이상치 유형에 대한 자세한 정보는 326 페이지의 『이상값』의 내용을 참조하십시오.
- 전송 및 변환 기능에 대한 자세한 정보는 『전이 및 변환 함수』의 내용을 참조하십시오.

전이 및 변환 함수: 전이 함수 차수 및 변환 대화 상자에서는 변환을 지정하고 ARIMA 모델의 일부 또는 전체 입력 모델에 대해 전이 함수를 정의할 수 있습니다.

목표 변환 이 분할창에서는 모델링하기 전에 각 목표변수에 대해 수행할 변환을 지정할 수 있습니다.

- **없음** 변환을 수행하지 않습니다.
- **제곱근** 제곱근 변환을 수행합니다.
- **자연로그** 자연로그 변환을 수행합니다.

후보 입력 전이 함수 및 변환 전이 함수를 사용하여 목표 계열의 미래 값을 예측하는 데 입력 필드의 과거 값을 사용하는 방식을 지정할 수 있습니다. 분할창 왼쪽에 있는 목록에는 모든 입력 필드가 표시됩니다. 이 분할창의 나머지 정보는 선택한 입력 필드에 따라 다릅니다.

전이 함수 차수 구조 눈금의 해당 셀에 전이 함수의 다양한 구성요소 값을 입력합니다. 모든 값은 0또는 음이 아닌 정수여야 합니다. 분자 및 분모 구성요소의 경우 값은 최대 차수를 나타냅니다. 더 낮은 양의 차수가 모두 모델에 포함됩니다. 또한 분자 구성요소에 대해 차수 0은 항상 포함됩니다. 예를 들어, 분자로 2를 지정하면 모델에 차수 2, 1, 0이 포함됩니다. 분모로 3을 지정하면 모델에 차수 3, 2, 1이 포함됩니다. 계절 열의 셀은 활성 데이터 세트에 대해 주기성이 정의된 경우에만 사용할 수 있습니다.

분자 전이 함수의 분자 차수는 종속 계열의 현재 값을 예측하기 위해 사용되는 선택된 독립(예측변수) 계열의 이전 값을 지정합니다. 예를 들어, 분자 차수 1은 각 종속 계열의 현재 값 예측에 과거 1개 시간 주기의 독립 계열 값과 독립 계열의 현재 값을 사용하도록 지정합니다.

분모 전이 함수의 분모 차수는 종속 계열의 현재 값을 예측하기 위해 사용되는 선택된 독립(예측변수) 계열의 이전 값에 대한 계열 평균의 편차를 지정합니다. 예를 들어 분모 차수 1은 각 종속 계열의 현재 값을 예측하는 경우 과거 1개 시간 주기의 독립 계열 평균 값 편차를 고려하도록 지정합니다.

차이 모델을 추정하기 전 선택된 독립(예측변수) 계열에 적용할 차이 차수를 지정합니다. 추세가 있으며 해당 효과 제거를 위해 사용되는 경우 차이가 필요합니다.

계절 계절 분자, 분모 및 차이 구성요소는 해당 비계절 구성요소와 동일한 역할을 합니다. 그러나 계절 차수의 경우 현재 계열 값이 한 개 이상의 계절 주기에 의해 구분된 이전 계열 값의 영향을 받습니다. 예를 들어, 월별 데이터(계절 주기 12)의 경우 계절 차수 1은 현재 계열 값이 현재 계열 이전의 계열 값 12 주기의 영향을 받는다는 것을 의미합니다. 월별 데이터의 경우 계절 차수 1은 비계절 차수 12를 지정하는 것과 동일합니다.

지연 지연을 설정하면 지정된 구간 수만큼 입력 필드의 영향력이 지연됩니다. 예를 들어, 지연이 5로 설정된 경우 시간 t의 입력 필드 값은 다섯 구간이 경과할 때까지 예측에 영향을 주지 않습니다(t + 5).

변환 독립 변수 세트에서 전이 함수를 지정하면 해당 변수에서 수행할 선택적 변환도 포함됩니다.

- **없음** 변환을 수행하지 않습니다.
- **제곱근** 제곱근 변환을 수행합니다.
- **자연로그** 자연로그 변환을 수행합니다.

시계열 노드 - 작성 출력 옵션

ACF 및 PACF 출력의 최대 시차수. 자기상관(ACF) 및 편자기상관(PACF)은 현재 및 지난 계열 값 사이의 연관성 측도이며 나중 값 예측에 가장 유용한 지난 계열 값을 표시합니다. 자기상관 및 편자기상관 테이블과 도표에 표시되는 최대 시차 수를 설정할 수 있습니다.

예측변수 중요도 계산. 적절한 중요도 측도를 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 예측변수는 삭제 또는 무시하려 합니다. 예측변수 중요도는 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 일부 모델에 대해 기본적으로 해제되어 있습니다.

시계열 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.

신뢰구간 너비(%). 모델 예측 및 잔차 자기상관에 대해 신뢰구간을 계산합니다. 100보다 작은 양수 값을 지정할 수 있습니다. 기본적으로, 95% 신뢰구간이 사용됩니다.

기존 모델을 사용하여 추정 계속. 이미 시계열 모델을 생성한 경우 해당 모델에 대해 지정된 기준 설정을 재사용하고 처음부터 새 모델을 작성하는 대신, 모델 팔레트에서 새 모델 노드를 생성하려면 이 옵션을 선택하십시오. 이 방식에서는 이전(그러나, 최근의 데이터)과 동일한 모델 설정을 기반으로 하는 새 예측을 다시 추정하고 생성하여 시간을 절약할 수 있습니다. 따라서 예를 들어, 특정 시계열의 원래 모델이 Holt의 선형 추세인 경우 해당 데이터의 재추정 및 시계열 분석에 동일한 모델 유형이 사용됩니다. 시스템은 새 데이터에 대한 최상의 모델 유형을 찾으려고 다시 시도하지 않습니다.

스코어링 모델만 작성. 모델에 저장되는 데이터의 양을 줄이려면 이 상자를 선택하십시오. 이 옵션을 사용하면 매우 많은 시계열(수만 개)이 포함된 모델을 작성할 때 성능을 향상시킬 수 있습니다. 여전히 일반적인 방식으로 데이터를 스코어링할 수 있습니다.

레코드를 미래로 확장. 추정 기간이 끝난 이후에 예측할 시간 구간 수를 설정할 수 있는 다음 **예측에 사용할 미래 값** 섹션을 사용합니다. 이 경우의 시간 구간은 데이터 지정 사항 탭에 지정한 분석의 시간 구간입니다. 이 설정에 최대 한계는 없습니다. 다음 옵션을 사용하면 입력의 미래 값을 자동으로 계산하거나 하나 이상의 예측 변수에 대한 예측 값을 수동으로 지정할 수 있습니다.

예측에 사용할 미래 값

- **입력의 미래 값 계산.** 이 옵션을 선택하면 예측자의 예측 값, 잡음 예측, 분산 추정 및 미래 값이 자동으로 계산됩니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다.
- **데이터에 추가할 값이 있는 필드 선택.** 예측변수 필드를 사용하는 경우(역할이 입력으로 설정됨), 예측하려는 레코드(검증용 제외)마다 각 예측변수에 대한 예측 기간 동안의 추정값을 지정할 수 있습니다. 값을 수동으로 지정하거나 목록에서 선택할 수 있습니다.

– **필드.** 필드 선택기 단추를 클릭하고 예측변수로 사용할 수 있는 필드를 선택하십시오. 여기서 선택한 필드가 모델링에 사용되거나 사용되지 않을 수 있습니다. 실제로 필드를 예측변수로 사용하려면 다운스트림 모델링 노드에서 해당 필드를 선택해야 합니다. 이 대화 상자에서는 단지 각 노드에서 개별적으로 미래 값을 지정하지 않고도 여러 다운스트림 모델링 노드에서 값을 공유할 수 있도록 미래 값을 지정하는 편리한 장소를 제공합니다. 작성 옵션 탭의 선택사항을 통해 사용 가능한 필드의 목록을 제한할 수도 있습니다.

  삭제되었거나 작성 옵션 탭에서 선택한 사항을 업데이트했기 때문에 더 이상 스트림에서 사용할 수 없는 필드에 미래 값이 지정된 경우 필드는 빨간색으로 표시됩니다.

– **값.** 필드마다, 함수 목록에서 선택하거나 **지정**을 클릭하여 수동으로 값을 입력하거나 또는 사전 정의된 값 목록에서 선택할 수 있습니다. 예측변수 필드가 제어 가능한 항목 또는 미리 알 수 있는 항목과 관련이 있으면 값을 수동으로 입력해야 합니다. 예를 들어, 룸 예약 수를 기반으로 호텔의 다음 달 수입을 예측하는 경우, 해당 기간 동안의 실제 예약 수를 지정할 수 있습니다. 반면, 예측변수 필드가 제어 불가능한 항목(예: 주식 가격)과 관련이 있으면 most recent value 또는 mean of recent points와 같은 함수를 사용할 수 있습니다.

  사용 가능한 함수는 필드의 측정 수준에 따라 다릅니다.

표 *28.* 측정 수준에 사용 가능한 함수

측정 수준	함수
연속형 또는 명목 필드	공백 최신 점의 평균 최신 값 지정
플래그 필드	공백 최신 값 True False 지정

최신 점의 평균은 마지막 세 개의 데이터 점 평균에서 미래 값을 계산합니다.

최신 값은 미래 값을 최신 데이터 점의 값으로 설정합니다.

True/False는 플래그 필드의 미래 값을 지정된 대로 True나 False로 설정합니다.

지정은 수동으로 미래 값을 지정하거나 사전 정의된 목록에서 선택하여 대화 상자를 엽니다.

스코어링에 사용 가능

모델 너깃에 대한 대화 상자에 표시할 스코어링 옵션에 대한 기본값을 여기서 설정할 수 있습니다.

- **신뢰구간 상한 및 하한 계산** 선택할 경우 이 옵션은 각 목표 필드에서 상한 및 하한 신뢰구간에 대한 새 필드(기본 접두문자: $TSLCI- 및 $TSUCI-)를 작성합니다.

- **잡음 잔차 계산** 이 옵션을 선택한 경우 각 목표 필드에서 모델 잔차와 이러한 값의 총계에 대한 새 필드(기본 접두문자가 $TSResidual-임)를 작성합니다.

모델 설정

출력에 표시할 최대 모델 수. 출력에 포함될 최대 모델 수를 지정하십시오. 작성된 모델 수가 이 임계값을 초과할 경우 모델이 출력에 표시되지는 않지만 스코어링에 계속 사용할 수 있습니다. 기본값은 10입니다. 더 큰 모델 수를 표시하면 성능이나 안정성이 저하될 수 있습니다.

시계열 모델 너깃

시계열 모델 너깃 출력

시계열 모델을 작성하면 출력 뷰어에 다음 정보가 제공됩니다. 시계열 모델의 출력 뷰어에 표시될 수 있는 모델은 10개로 제한됩니다.

임시 정보 요약

요약에는 다음 정보가 표시됩니다.

- 시간 필드
- 증분
- 시작점 및 끝점
- 고유한 포인트 수

요약은 모든 목표에 적용됩니다.

모델 정보 테이블

모델 정보 테이블에서 모델에 대한 키 정보를 제공하며 각 대상마다 반복됩니다. 이 테이블에는 항상 다음과 같은 고급 모델 설정이 포함되어 있습니다.

- 유형 노드 또는 시계열 노드 필드 탭에서 선택된 목표 필드의 이름
- 모델 작성 방법(에: 지수평활 또는 ARIMA)
- 모델에 입력된 예측변수 수
- 모델 유형 적합에 사용된 레코드 수. 다양한 모형 유형의 예로는 RMSE, MAE, AIC, BIC, R 제곱 등이 있습니다.

또한 데이터가 필요한 조건을 충족할 경우 Ljunq-Box Q 통계도 표시할 수 있습니다. 다음 조건에서는 이 통계를 사용할 수 **없습니다**.

- 비결측 데이터 포인트 수가 원하는 합계 항 수(18에 고정됨)보다 작거나 같을 경우
- 모수의 수가 원하는 합계 항 수보다 크거나 같을 경우
- 계산된 합계 항 수가 허용되는 가장 작은 k 값(7에서 고정됨)보다 작을 경우
- 테이블이 각 목표에 대해 반복하는 경우

예측변수 중요도

예측변수 중요도 그래프는 막대형 차트로 모델에 있는 상위 10개 입력(예측자)의 중요도를 표시하며, 각 대상마다 반복됩니다.

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.

상관도표

상관도표 또는 자기상관 도표는 각 목표에 대해 표시되며, 시차 대 잔차(예상 값과 실제 값 사이의 차이)의 자기상관 함수(ACF) 또는 편자기상관 함수(PACF)를 표시합니다. 신뢰구간은 차트 전체에서 강조표시됩니다.

모수 추정값

모수 추정값 테이블은 각 목표에 대해 반복되며, 다음과 같은 세부사항(해당하는 경우)을 표시합니다.

- 목표 이름
- 적용된 변환
- 모델의 이 모수에 사용된 시차(ARIMA)
- 계수 값
- 모수 추정값의 표준 오차
- 모수 추정값을 표준 오차로 나눈 값
- 모수 추정값의 유의 수준.

시계열 모델 너깃 설정

설정 탭에서는 시계열 모델 너깃에 대한 추가 옵션을 제공합니다.

예측

레코드를 미래로 확장하는 옵션은 추정 기간이 끝난 이후에 예측할 시간 구간 수를 설정합니다. 이 경우의 시간 구간은 시계열 노드의 데이터 지정 사항 탭에 지정된 분석의 시간 구간입니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다.

입력의 미래 값 계산. 이 옵션을 선택하면 예측변수의 예측 값, 잡음 예측, 분산 추정 및 미래 시간 값이 계산됩니다.

예측에 사용할 미래 값

- **입력의 미래 값 계산.** 이 옵션을 선택하면 예측자의 예측 값, 잡음 예측, 분산 추정 및 미래 값이 자동으로 계산됩니다. 예측이 요청되면 물론 목표가 아닌 입력 계열에서 자기회귀분석 모델이 자동으로 작성됩니다. 그런 다음, 이 모델을 사용하여 예측 기간에 해당 입력 계열의 값을 생성합니다.
- **데이터에 추가할 값이 있는 필드 선택.** 예측변수 필드를 사용하는 경우(역할이 입력으로 설정됨), 예측하려는 레코드(검증용 제외)마다 각 예측변수에 대한 예측 기간 동안의 추정값을 지정할 수 있습니다. 값을 수동으로 지정하거나 목록에서 선택할 수 있습니다.
  - **필드.** 필드 선택기 단추를 클릭하고 예측변수로 사용할 수 있는 필드를 선택하십시오. 여기서 선택한 필드가 모델링에 사용되거나 사용되지 않을 수 있습니다. 실제로 필드를 예측변수로 사용하려면 다운스트림 모델링 노드에서 해당 필드를 선택해야 합니다. 이 대화 상자에서는 단지 각 노드에서 개별적으로 미래 값을 지정하지 않고도 여러 다운스트림 모델링 노드에서 값을 공유할 수 있도록 미래 값을 지정하는 편리한 장소를 제공합니다. 작성 옵션 탭의 선택사항을 통해 사용 가능한 필드의 목록을 제한할 수도 있습니다.

    삭제되었거나 작성 옵션 탭에서 선택한 사항을 업데이트했기 때문에 더 이상 스트림에서 사용할 수 없는 필드에 미래 값이 지정된 경우 필드는 빨간색으로 표시됩니다.
  - **값.** 필드마다, 함수 목록에서 선택하거나 **지정**을 클릭하여 수동으로 값을 입력하거나 또는 사전정의된 값 목록에서 선택할 수 있습니다. 예측변수 필드가 제어 가능한 항목 또는 미리 알 수 있는 항목과 관련이 있으면 값을 수동으로 입력해야 합니다. 예를 들어, 룸 예약 수를 기반으로 호텔의 다음 달 수입을 예측하는 경우, 해당 기간 동안의 실제 예약 수를 지정할 수 있습니다. 반면, 예측변수 필드가 제어 불가능한 항목(예: 주식 가격)과 관련이 있으면 most recent value 또는 mean of recent points와 같은 함수를 사용할 수 있습니다.

    사용 가능한 함수는 필드의 측정 수준에 따라 다릅니다.

표 *29.* 측정 수준에 사용 가능한 함수

측정 수준	함수
연속형 또는 명목 필드	공백 최신 점의 평균 최신 값 지정
플래그 필드	공백 최신 값 True False 지정

최신 점의 평균은 마지막 세 개의 데이터 점 평균에서 미래 값을 계산합니다.

최신 값은 미래 값을 최신 데이터 점의 값으로 설정합니다.

True/False는 플래그 필드의 미래 값을 지정된 대로 True나 False로 설정합니다.

지정은 수동으로 미래 값을 지정하거나 사전 정의된 목록에서 선택하여 대화 상자를 엽니다.

스코어링에 사용 가능

각 모델의 스코어를 계산할 새 필드 만들기. 각 모델의 스코어를 계산하기 위해 작성할 새 필드를 지정할 수 있습니다.

- **잡음 잔차.** 이 옵션을 선택한 경우 각 목표 필드에서 모델 잔차와 이러한 값의 총계에 대한 새 필드(기본 접두문자가 $TSResidual-임)를 작성합니다.
- **신뢰 상한 및 하한.** 이 옵션을 선택한 경우 각 목표 필드에서 각각 하한 및 하한 신뢰구간과 이러한 값의 총계에 대한 새 필드(기본 접두문자가 $TSLCI- 및 $TSUCI-임)를 작성합니다.

스코어링에 포함된 목표 모델 스코어에 포함할 수 있는 목표를 선택합니다.

제 14 장 자체 학습 응답 노드 모델

SLRM 노드

SLRM(**Self-Learning Response Model**) 노드에서는 전체 데이터 세트를 사용할 때마다 모델을 다시 작성하지 않고도 데이터 세트가 성장하면 지속적으로 업데이트 또는 재추정할 수 있는 모델을 작성할 수 있습니다. 예를 들어, 이는 여러 제품이 있고 제품에 대한 제안을 제공할 경우 고객이 구매할 가능성이 높은 제품을 식별하려는 경우에 유용합니다. 이 모델에서는 고객에게 가장 적합한 제안과 수락할 제안의 확률을 예측할 수 있습니다.

모델은 처음에 임의로 작성된 제안과 이 제안에 대한 반응을 포함하는 작은 데이터 세트를 사용하여 작성할 수 있습니다. 데이터 세트가 성장하면 모델을 업데이트할 수 있으므로 고객에게 더 적합한 제안과 나이, 성별, 직업, 소득과 같은 기타 입력 필드에 기반하여 제안 수락 확률을 더 효과적으로 예측할 수 있습니다. 사용 가능한 제안은 데이터 세트의 목표 필드를 변경하지 않고도 노드 대화 상자에서 필드를 추가하거나 제거하여 변경할 수 있습니다.

IBM SPSS Collaboration and Deployment Services와 함께 사용하면 모델에 대한 정기적인 자동 업데이트를 설정할 수 있습니다. 사용자의 통찰력이나 작업 없이도 이 프로세스를 통해 데이터 마이너에 의한 사용자 정의 개입이 불가능하거나 필요하지 않은 조직과 애플리케이션에서 저렴하고 탄력적인 솔루션을 제공합니다.

예제. 금융 기관에서 각 고객이 수락할 확률이 높은 제안을 매치시켜 보다 수익성 높은 결과를 달성하고자 합니다. 자체 학습 모델을 사용하여 이전 프로모션에 기반하여 가장 우호적인 반응을 끌어낼 것 같은 고객 특성을 식별하고 최근 고객 반응에 기반하여 실시간으로 모델을 업데이트할 수 있습니다.

SLRM 노드 필드 옵션

SLRM 노드를 실행하기 전에 노드의 필드 탭에서 목표 및 목표 반응 필드 모두를 지정해야 합니다.

목표 필드. 목록에서 목표 필드를 선택합니다. 예를 들어, 고객에게 제공하려는 서로 다른 제품을 포함하는 명목 집합 필드입니다.

참고: 목표 필드는 숫자가 아닌 문자열 저장 공간이어야 합니다.

목표 반응 필드. 목록에서 목표 반응 필드를 선택합니다. 예를 들어, 수락됨 또는 기각됨이어야 합니다.

참고: 이 필드는 플래그여야 합니다. 플래그의 참 값은 제안 수락을 나타내고 거짓 값은 제안 거절을 나타냅니다.

이 대화 상자에서 나머지 필드는 IBM SPSS Modeler 전반에 걸쳐 사용되는 표준 필드입니다. 자세한 정보는 33 페이지의 『모델링 노드 필드 옵션』의 내용을 참조하십시오.

참고: 소스 데이터가 연속(숫자 범위) 입력 필드로 사용할 범위를 포함하는 경우 메타데이터는 각 범위의 최소 및 최대 세부사항 모두를 포함해야 합니다.

SLRM 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 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.

목표 필드 값 기본적으로 **모두 사용**으로 설정되어 있습니다. 즉, 선택한 목표 필드 값과 연관된 모든 제안을 포함하는 모델이 작성됨을 의미합니다. 목표 필드 제안 중 일부만 포함하는 모델을 생성하려면 **지정**을 클릭하고 **추가**, **편집**, **삭제** 단추를 사용하여 모델을 작성하려는 제안 이름을 입력하거나 추가합니다. 예를 들어, 제공하는 모든 제품을 나열하는 목표를 선택하는 경우 이 필드를 사용하여 여기에 입력한 몇 개로만 제안된 제품을 제한할 수 있습니다.

모델 평가. 이 패널의 필드는 스코어링에 영향을 주지 않는다는 점에서 모델과는 독립되어 있습니다. 대신, 모델이 결과를 예측하는 정도에 대한 시각적 표현을 작성할 수 있습니다.

참고: 모델 너깃에 모델 평가 결과를 표시하려면 **모델 평가 표시** 상자도 선택해야 합니다.

- **모델 평가 포함.** 선택한 각 제안에 대한 모델의 예측 정확도를 표시하는 그래프를 작성하려면 이 상자를 선택합니다.
- **난수 시드 설정.** 무작위 퍼센트에 기반한 모델의 정확도를 추정할 때 이 옵션을 사용하면 다른 세션에서 동일한 결과를 복제할 수 있습니다. 난수 생성기에서 사용하는 시작값을 지정하면 노드를 실행할 때마다 동일한 레코드를 지정하도록 보장할 수 있습니다. 원하는 시드 값을 입력하십시오. 이 옵션을 선택하지 않으면 노드를 실행할 때마다 다른 표본이 생성됩니다.
- **시뮬레이션된 표본 크기.** 모델을 평가할 때 표본에 사용할 레코드 수를 지정합니다. 기본값은 100입니다.
- **반복계산 수.** 이를 통해 지정된 반복 이후에 모델 평가 작성을 중지할 수 있습니다. 최대 반복 수를 지정합니다. 기본값은 20입니다.

  참고: 표본 크기가 크고 반복 수가 많으면 모델 작성에 걸리는 시간이 길어집니다.

모델 평가 표시. 모델 너깃에 결과의 그래픽 표현을 표시하려면 이 옵션을 선택합니다.

SLRM 노드 설정 옵션

노드 설정 옵션을 사용하면 모델 작성 프로세스를 미세하게 조정할 수 있습니다.

레코드당 최대 예측 수 이 옵션을 사용하면 데이터 세트에 있는 각 레코드에서 수행된 예측 수를 제한할 수 있습니다. 기본값은 3입니다.

예를 들어, 6개의 제안이 있지만(예: 저축, 저당, 자동차 구입 대출, 펜션, 신용카드, 보험), 추천할 수 있는 최상의 제안 2개만 알고 싶습니다. 이 경우 이 필드를 2로 설정합니다. 모델을 작성하고 테이블에 첨부하면 레코드당 두 개의 예측 열(그리고 수락할 제안의 확률에서 연관된 신뢰도)을 확인할 수 있습니다. 예측은 6개의 가능한 제안으로 구성될 수 있습니다.

임의화 수준 편향(예를 들어, 작거나 불완전한 데이터 세트)을 방지하고 모든 가능한 제안을 동등하게 처리하기 위해 제안 선택의 임의화 수준과 추천 제안으로 포함할 제안의 확률을 추가할 수 있습니다. 임의화는 퍼센트로 표현되고, 0.0(임의화하지 않음)과 1.0(완전히 임의화) 사이의 소수점 값으로 표시됩니다. 기본값은 0.0입니다.

난수 시드 설정 제안 선택에 임의화 수준을 추가할 때 이 옵션을 사용하면 다른 세션에서 동일한 결과를 복제할 수 있습니다. 난수 생성기에서 사용하는 시작값을 지정하면 노드를 실행할 때마다 동일한 레코드를 지정하도록 보장할 수 있습니다. 원하는 시드 값을 입력하십시오. 이 옵션을 선택하지 않으면 노드를 실행할 때마다 다른 표본이 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.

정렬 순서 작성된 모델에서 제안을 표시하는 순서를 선택합니다.

- **내림차순** 모델은 스코어가 가장 높은 제안부터 먼저 표시합니다. 이는 수락할 확률이 가장 높은 제안을 의미합니다.
- **오름차순** 모델은 스코어가 가장 낮은 제안부터 먼저 표시합니다. 이는 기각할 확률이 가장 높은 제안을 의미합니다. 예를 들어, 특정 제안에 대한 마케팅 캠페인에서 제거할 고객을 결정할 때 유용할 수 있습니다.

목표 필드의 기본 설정 모델 작성 시 적극적으로 올리거나 제거할 데이터의 특정 측면이 존재할 수 있습니다. 예를 들어, 고객에게 올릴 최고의 재무 제안을 선택하는 모델 작성 시 각 고객에 대해 스코어 계산의 효율성에 상관없이 하나의 특정 제안을 항상 포함하고자 할 수 있습니다.

이 패널에서 제안을 포함하고 해당 기본 설정을 편집하려면 **추가**를 클릭하고 제안의 이름(예: 저축 또는 담보)을 입력하고 **확인**을 클릭하십시오.

- **값** 이는 추가한 제안의 이름을 표시합니다.
- **환경 설정** 제안에 적용할 환경 설정의 수준을 지정합니다. 환경 설정은 퍼센트로 표현되고, 0.0(선호되지 않음)과 1.0(가장 선호됨) 사이의 소수점 값으로 표시됩니다. 기본값은 0.0입니다.
- **항상 포함** 특정 제안이 항상 예측에 포함되도록 하려면 이 상자를 선택합니다.

참고: 환경 설정이 0.0으로 설정된 경우 **항상 포함** 설정은 무시됩니다.

모델 신뢰도 고려 잘 구성되고, 데이터가 풍부한 모델(여러 재생성을 통해 미세 조정됨)은 항상 데이터가 적은 완전히 새로운 모델과 비교했을 때 더 정확한 결과를 생성해야 합니다. 보다 성숙한 모델의 증가된 신뢰도를 활용하려면 이 상자를 선택합니다.

SLRM 모델 너깃

참고: 결과는 모델 옵션 탭에서 **모델 평가 포함** 및 **모델 평가 표시**를 모두 선택한 경우 이 탭에만 표시됩니다.

SLRM 모델을 포함하는 스트림을 실행할 때 노드는 각 목표 필드 값(제안)의 예측 정확도와 사용된 각 예측의 중요도를 추정합니다.

참고: 모델링 노드 모델 탭에서 **기존 모델 훈련 계속**을 선택한 경우 모델을 재생성할 때마다 모델 너깃에 표시된 정보가 업데이트됩니다.

IBM SPSS Modeler 12.0 이상을 사용하여 작성된 모델에서는 모델 너깃 모델 탭이 다음 두 개 열로 구분됩니다.

왼쪽 열.

- **보기.** 둘 이상의 제안이 있는 경우 결과를 표시할 하나를 선택하십시오.
- **모델 성능.** 여기에는 각 제안에 대한 추정된 모델 정확도를 표시합니다. 검정 세트는 시뮬레이션을 통해 생성됩니다.

오른쪽 열.

- **보기. 반응을 포함하는 연관** 또는 **변수 중요도** 세부사항을 표시할 것인지 여부를 선택합니다.
- **반응을 포함하는 연관.** 목표 변수를 포함하는 각 예측변수의 연관(상관관계)을 표시합니다.
- **예측변수 중요도.** 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타냅니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 SLRM에서 그래프는 SLRM 알고리즘에 의해 시뮬레이션으로 생성되어도 예측변수 중요도를 표시하는 다른 모델과 동일한 방식으로 해석할 수 있습니다. 이는 모델에서 차례로 각 예측변수를 제거하고 이로 인해 모델의 정확도에 미치는 영향을 확인하여 수행합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

SLRM 모델 설정

SLRM 모델 너깃의 설정 탭에서는 작성된 모델을 수정하는 옵션을 지정합니다. 예를 들어, SLRM 노드를 사용하여 동일한 데이터와 설정을 통해 여러 다른 모델을 작성하고 각 모델에서 이 탭을 사용하여 결과에 영향을 미치는 방식을 확인하고자 설정을 약간 수정할 수 있습니다.

참고: 이 탭은 스트림에 모델 너깃을 추가한 후에만 사용 가능합니다.

레코드당 최대 예측 수 이 옵션을 사용하면 데이터 세트에 있는 각 레코드에서 수행된 예측 수를 제한할 수 있습니다. 기본값은 3입니다.

예를 들어, 6개의 제안이 있지만(예: 저축, 저당, 자동차 구입 대출, 펜션, 신용카드, 보험), 추천할 수 있는 최상의 제안 2개만 알고 싶습니다. 이 경우 이 필드를 2로 설정합니다. 모델을 작성하고 테이블에 첨부하면 레코드당 두 개의 예측 열(그리고 수락할 제안의 확률에서 연관된 신뢰도)을 확인할 수 있습니다. 예측은 6개의 가능한 제안으로 구성될 수 있습니다.

임의화 수준 편향(예를 들어, 작거나 불완전한 데이터 세트)을 방지하고 모든 가능한 제안을 동등하게 처리하기 위해 제안 선택의 임의화 수준과 추천 제안으로 포함할 제안의 확률을 추가할 수 있습니다. 임의화는 퍼센트로 표현되고, 0.0(임의화하지 않음)과 1.0(완전히 임의화) 사이의 소수점 값으로 표시됩니다. 기본값은 0.0입니다.

난수 시드 설정 제안 선택에 임의화 수준을 추가할 때 이 옵션을 사용하면 다른 세션에서 동일한 결과를 복제할 수 있습니다. 난수 생성기에서 사용하는 시작값을 지정하면 노드를 실행할 때마다 동일한 레코드를 지정하도록 보장할 수 있습니다. 원하는 시드 값을 입력하십시오. 이 옵션을 선택하지 않으면 노드를 실행할 때마다 다른 표본이 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.

정렬 순서 작성된 모델에서 제안을 표시하는 순서를 선택합니다.

- **내림차순** 모델은 스코어가 가장 높은 제안부터 먼저 표시합니다. 이는 수락할 확률이 가장 높은 제안을 의미합니다.
- **오름차순** 모델은 스코어가 가장 낮은 제안부터 먼저 표시합니다. 이는 기각할 확률이 가장 높은 제안을 의미합니다. 예를 들어, 특정 제안에 대한 마케팅 캠페인에서 제거할 고객을 결정할 때 유용할 수 있습니다.

목표 필드의 기본 설정 모델 작성 시 적극적으로 올리거나 제거할 데이터의 특정 측면이 존재할 수 있습니다. 예를 들어, 고객에게 올릴 최고의 재무 제안을 선택하는 모델 작성 시 각 고객에 대해 스코어 계산의 효율성에 상관없이 하나의 특정 제안을 항상 포함하고자 할 수 있습니다.

이 패널에서 제안을 포함하고 해당 기본 설정을 편집하려면 **추가**를 클릭하고 제안의 이름(예: 저축 또는 담보)을 입력하고 **확인**을 클릭하십시오.

- **값** 이는 추가한 제안의 이름을 표시합니다.
- **환경 설정** 제안에 적용할 환경 설정의 수준을 지정합니다. 환경 설정은 퍼센트로 표현되고, 0.0(선호되지 않음)과 1.0(가장 선호됨) 사이의 소수점 값으로 표시됩니다. 기본값은 0.0입니다.
- **항상 포함** 특정 제안이 항상 예측에 포함되도록 하려면 이 상자를 선택합니다.

  참고: 환경 설정이 0.0으로 설정된 경우 **항상 포함** 설정은 무시됩니다.

모델 신뢰도 고려 잘 구성되고, 데이터가 풍부한 모델(여러 재생성을 통해 미세 조정됨)은 항상 데이터가 적은 완전히 새로운 모델과 비교했을 때 더 정확한 결과를 생성해야 합니다. 보다 성숙한 모델의 증가된 신뢰도를 활용하려면 이 상자를 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 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.

제 15 장 지원 벡터 머신 모델

SVM 정보

지원 벡터 머신(SVM)은 훈련 데이터 과적합 없이도 모델의 예측 정확도를 최대화하는 강력한 분류 및 회귀분석 기법입니다. 특히 SVM은 많은 예측변수 필드(예: 수천 개의 필드)에서 데이터를 분석하는 데 적합합니다.

SVM은 고객 관계 관리(CRM), 얼굴 및 기타 이미지 인식, 생물정보학, 텍스트 마이닝 개념 추출, 침입 감지, 단백질 구조 예측, 음성 및 대화 인식을 비롯해 다양한 분야에서 응용할 수 있습니다.

SVM 작동 방법

SVM은 고차원 기능 공간으로 데이터를 맵핑하여 작동하므로, 데이터가 이외의 경우 선형으로 분리 불가능한 경우에도 데이터 포인트는 분류 가능합니다. 범주 사이의 구분 문자가 존재하며, 데이터는 구분 문자를 초평면으로 그릴 수 있는 방식으로 변형됩니다. 그러면 새 데이터 특성을 사용하여 새 레코드를 포함해야 하는 그룹을 예측할 수 있습니다.

예를 들어 데이터 포인트가 두 개의 서로 다른 범주에 포함되는 다음 그림을 고려하십시오.

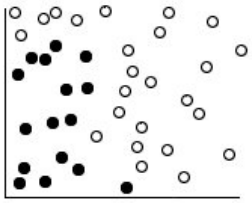


그림 *59.* 원래 데이터 세트

다음 그림과 같이 두 개 범주를 곡선으로 분리할 수 있습니다.

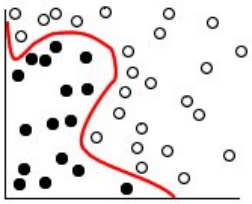


그림 *60.* 구분 문자가 추가된 데이터

변환 후 두 범주 사이의 경계는 다음 그림과 같이 하이픈으로 정의할 수 있습니다.

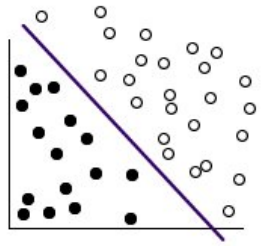


그림 *61.* 변환 데이터

변환에 사용된 산술 함수는 **커널** 함수라고도 합니다. IBM SPSS Modeler에서 SVM은 다음 커널 유형을 지원합니다.

- 선형
- 다항
- 방사형 기저함수(RBF)
- 시그모이드

선형 커널 함수는 데이터의 선형 분리가 복잡하지 않은 경우 권장됩니다. 즉, 다른 기능 중 하나를 사용해야 합니다. 각각 서로 다른 알고리즘과 모수를 사용하므로 서로 다른 함수로 실험하여 각 경우에 최상의 모델을 얻도록 해야 합니다.

SVM 모델 조정

범주 사이의 선구분 변수 외에도 분류 SVM 모델은 두 범주 사이의 공백을 정의하는 한계선도 찾습니다.

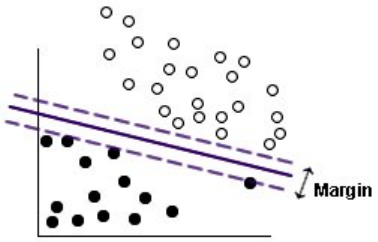


그림 *62.* 예비 모델을 포함하는 데이터

여백에 놓인 데이터 포인트는 **지원 벡터**라고도 합니다.

두 범주 사이의 여백이 넓을수록 새 레코드의 범주를 예측할 때 모델 효율성이 더 높습니다. 이전 예에서 여백은 별로 넓지 않았으며, 모델은 **과적합**으로 간주됩니다. 여백을 확장시키기 위해 적은 수의 오분류는 허용 가능합니다. 이러한 예가 다음 그림에 표시됩니다.

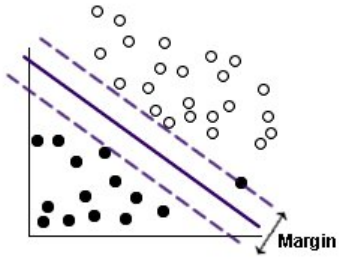


그림 *63.* 향상된 모델을 포함하는 데이터

일부 경우에 선형 분할이 더 어렵습니다. 다음 그림에서 해당 예를 보여줍니다.

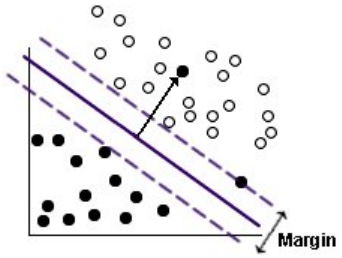


그림 *64.* 선형 분할의 문제점

이와 같은 케이스에서는 넓은 여백과 오분류된 적은 데이터 포인트 사이에서 최적의 균형을 찾는 것이 목표입니다. 커널 함수에는 이러한 두 값 사이의 균형을 제어하는 **정규화 모수**(C라고도 함)가 있습니다. 최상의 모델을 찾기 위해 이 모수와 다른 커널 모수의 서로 다른 값으로 실험해야 할 수도 있습니다.

SVM 노드

SVM 노드를 사용하면 지원 벡터 머신을 사용하여 데이터를 분류할 수 있습니다. SVM은 특히 포괄적인 데이터 세트(즉, 예측변수 필드가 많음)에서 사용하는 데 적합합니다. 노드에서 기본 설정을 사용하여 기본 모델을 비교적 신속하게 생성하거나 고급 설정을 사용하여 SVM 모델의 서로 다른 유형을 실험할 수 있습니다.

모델을 작성하면 다음을 수행할 수 있습니다.

- 모델 너깃을 찾아서 모델을 작성할 때 입력 필드의 상대적 중요도를 표시합니다.
- 모델 너깃에 테이블 노드를 첨부하여 모델 출력을 봅니다.

예. 의료 분야의 연구자가 암 진행 위험이 있다고 판단된 환자로부터 추출한 여러 조직 표본의 특성을 포함하는 데이터 세트를 확보했습니다. 원래 데이터 분석에서는 많은 특성이 양성과 악성 표본 사이에서 크게 다르다고 나왔습니다. 연구자는 다른 환자의 표본에서 유사한 셀 특성 값을 사용할 수 있는 SVM 모델을 개발하여 표본이 양성인지, 악성인지 여부를 조기에 표시하고자 합니다.

SVM 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 내용을 참조하십시오.

SVM 노드 고급 옵션

지원 벡터 머신에 대한 자세한 지식이 있는 경우 고급 옵션을 통해 훈련 프로세스를 미세 조정할 수 있습니다. 고급 옵션에 액세스하려면 고급 탭에서 모드를 **고급**으로 설정하십시오.

모든 확률 추가(범주형 목표에만 유효함). 선택한 경우 명목 또는 플래그 목표 필드의 가능한 각 값에 대한 확률이 노드에서 처리하는 각 레코드에 대해 표시되도록 지정합니다. 이 옵션을 선택하지 않으면 예측된 값의 확률만 명목 또는 플래그 목표 필드에 표시됩니다. 이 확인 상자의 설정은 모델 너깃 표시에서 대응하는 확인 상자의 기본 상태를 판별합니다.

중지 기준. 최적화 알고리즘의 중지 시점을 판별합니다. 값의 범위는 1.0E–1에서 1.0E–6까지이며, 기본값은 1.0E–3입니다. 값을 줄이면 더 정확한 모델이 생성되지만, 모델 훈련 시간이 더 오래 걸립니다.

정규화 모수(C). 여백의 극대화와 훈련 오차항의 최소화 사이에서 균형을 제어합니다. 값은 일반적으로 1과 10 사이(경계값 포함)여야 합니다. 기본값은 10입니다. 값을 늘리면 훈련 데이터에서 분류 정확도가 향상되거나 회귀분석 오차가 감소하지만 과적합으로 이어질 수도 있습니다.

회귀분석 정밀도(엡실론). 목표 필드의 측정 수준이 연속형인 경우에만 사용됩니다. 오류가 여기에 지정된 값보다 작으면 오류를 허용할 수 있습니다. 값을 늘리면 모델링 속도가 단축되지만, 정확도가 떨어질 수 있습니다.

커널 유형. 변환에 사용된 커널 함수의 유형을 판별합니다. 서로 다른 커널 유형으로 인해 구분 문자를 서로 다른 방식으로 계산하므로, 다양한 옵션으로 실험하는 것이 좋습니다. 기본값은 방사형 기저함수(**RBF**)입니다.

RBF 감마. 커널 유형이 **RBF**로 설정된 경우에만 사용 가능합니다. 일반적으로 값은 $3/k$와 $6/k$ 사이여야 합니다. 여기서, k는 입력 필드 수입니다. 예를 들어, 12개의 입력 필드가 있는 경우 0.25와 0.5 사이의 값을 시도해볼 수 있습니다. 값을 늘리면 훈련 데이터에서 분류 정확도가 향상되거나 회귀분석 오차가 감소하지만 과적합으로 이어질 수도 있습니다.

감마. 커널 유형이 **다항** 또는 **시그모이드**로 설정된 경우에만 사용 가능합니다. 값을 늘리면 훈련 데이터에서 분류 정확도가 향상되거나 회귀분석 오차가 감소하지만 과적합으로 이어질 수도 있습니다.

편향. 커널 유형이 **다항** 또는 **시그모이드**로 설정된 경우에만 사용 가능합니다. 커널 함수에서 coef0 값을 설정합니다. 기본값 0은 대부분의 경우에 적합합니다.

차수. 커널 유형이 **다항**으로 설정된 경우에만 사용 가능합니다. 맵핑 공백의 복잡도(차원)를 제어합니다. 보통 10보다 큰 값은 사용하지 않습니다.

SVM 모델 너깃

SVM 모델은 여러 새 필드를 작성합니다. 이러한 필드 중 가장 중요한 필드는 **$S-fieldname** 필드로, 모델에서 예측하는 목표 필드 값을 표시합니다.

모델에서 작성되는 새 필드와 수와 이름은 목표 필드(이 필드는 다음 표에서 *fieldname*으로 표시됨)의 측정 수준에 따라 달라집니다.

이러한 필드와 해당 값을 보려면 SVM 모델 너깃에서 테이블 노드를 추가하고 테이블 노드를 실행하십시오.

표 *30.* 목표 필드의 측정 수준은 '명목' 또는 '플래그'임

새 필드 이름	설명
$S-*fieldname*	목표 필드의 예측값.
$SP-*fieldname*	예측값의 확률.
$SP-*value*	명목 또는 플래그인 각 가능한 값의 확률로, 모델 너깃의 설정 탭에서 **모든 확률 추가**가 선택된 경우에만 표시됩니다.
$SRP-*value*	(플래그 목표만 해당) 원형(SRP) 및 수정된(SAP) 성향 스코어로 목표 필드의 "참" 결과가 나올 우도를 표시합니다. 이러한 스코어는 모델을 생성하기 전에 SVM 모델링 노드의 분석 탭에서 대응하는 확인 상자를 선택한 경우에만 표시됩니다. 자세한 정보는 37 페이지의 『모델링 노드 분석 옵션』의 내용을 참조하십시오.
$SAP-*value*	

표 *31.* 목표 필드의 측정 수준은 '연속'임

새 필드 이름	설명
$S-*fieldname*	목표 필드의 예측값.

예측변수 중요도

선택적으로 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타내는 차트도 모델 탭에 표시될 수 있습니다. 일반적으로 가장 중요한 예측변수에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하는 것이 좋습니다. 이 차트는 모델을 생성하기 전에 분석 탭에서 **예측변수 중요도 계산**을 선택한 경우에만 사용 가능합니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

참고*:* 예측변수 중요도를 사용하면 다른 유형의 모델보다 SVM 계산 시간이 오래 걸릴 수 있으므로, 기본적으로 분석 탭에서는 선택되어 있지 않습니다. 이 옵션을 선택하면 특히 큰 데이터 세트에서 성능이 느려질 수 있습니다.

SVM 모델 설정

설정 탭을 사용하면 결과를 볼 때 표시할 추가 필드를 지정할 수 있습니다(예: 너깃에 첨부된 테이블 노드 실행). 이러한 각 옵션의 효과는 옵션을 선택하고 미리보기 단추를 클릭하여 확인할 수 있습니다. 미리보기 출력의 오른쪽으로 스크롤하면 추가 필드가 나옵니다.

모든 확률 추가(범주형 목표에만 유효함). 이 옵션을 선택하면 명목 또는 플래그 목표 필드의 가능한 모든 값에 대한 확률이 노드에서 처리하는 각 레코드에서 표시됩니다. 이 옵션을 선택 해제하면 예측 값 및 해당 확률만 명목 또는 플래그 목표 필드에 표시됩니다.

이 확인 상자의 기본 설정은 모델링 노드에서 대응하는 확인 상자로 판별됩니다.

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.

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.

LSVM 노드

LSVM 노드를 사용하면 선형 지원 벡터 머신을 사용하여 데이터를 분류할 수 있습니다. LSVM은 특히 포괄적인 데이터 세트(즉, 예측변수 필드가 많음)에서 사용하는 데 적합합니다. 노드에서 기본 설정을 사용하여 기본 모델을 비교적 신속하게 생성하거나 작성 옵션을 사용하여 서로 다른 설정을 실험할 수 있습니다.

LSVM 노드는 SVM 노드와 유사하지만, 선형이고 많은 수의 레코드를 처리하기에 더 적합합니다.

모델을 작성하면 다음을 수행할 수 있습니다.

- 모델 너깃을 찾아서 모델을 작성할 때 입력 필드의 상대적 중요도를 표시합니다.
- 모델 너깃에 테이블 노드를 첨부하여 모델 출력을 봅니다.

예. 의료 분야의 연구자가 암 진행 위험이 있다고 판단된 환자로부터 추출한 여러 조직 표본의 특성을 포함하는 데이터 세트를 확보했습니다. 원래 데이터 분석에서는 많은 특성이 양성과 악성 표본 사이에서 크게 다르다고 나왔습니다. 연구자는 다른 환자의 표본에서 유사한 셀 특성 값을 사용할 수 있는 LSVM 모델을 개발하여 표본이 양성인지 또는 악성인지 여부를 조기에 표시하고자 합니다.

LSVM 노드 모델 옵션

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

예측변수 중요도 계산. 적절한 중요도 측도를 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 특히 큰 데이터 세트에 대해 작업할 때 일부 모델의 경우 계산 시간이 오래 걸릴 수 있어서 몇몇 모델은 기본적으로 해제되어 있음에 유의하십시오. 예측변수 중요도는 의사결정 목록 모델에 사용할 수 없습니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

LSVM 작성 옵션

모델 설정

절편 포함. 절편(모델에서 상수항)을 포함하면 솔루션의 전체 정확도가 증가할 수 있습니다. 데이터가 원점을 통과하여 전달된다고 가정할 수 있는 경우 절편을 제외할 수 있습니다.

범주형 대상에 대한 정렬 순서. 범주형 대상에 대한 정렬 순서를 지정합니다. 연속형 대상에서는 이 설정이 무시됩니다.

회귀분석 정밀도(엡실론). 목표 필드의 측정 수준이 연속형인 경우에만 사용됩니다. 오류가 여기에 지정된 값보다 작으면 오류를 허용할 수 있습니다. 값을 늘리면 모델링 속도가 단축되지만, 정확도가 떨어질 수 있습니다.

결측값을 가진 레코드 제외. **참**으로 설정하면 단일 값이 결측값인 경우 레코드가 제외됩니다.

페널티 설정

페널티 함수. 과적합 가능성을 줄이는 데 사용하는 페널티 함수 유형을 지정합니다. 옵션은 **L1** 또는 **L2**입니다.

L1과 **L2**는 계수의 페널티를 추가하여 과적합의 가능성을 줄입니다. 이 둘 사이의 차이점은 다수의 기능이 있을 때 **L1**에서 모델 작성 중에 계수를 0으로 설정하여 기능 선택사항을 사용한다는 것입니다. **L2**에는 이 기능이 없으므로 다수의 기능이 있을 때 사용하지 않아야 합니다.

페널티 모수(람다). 페널티(정규화) 모수를 지정합니다. 이 설정은 **페널티 함수**가 설정된 경우에만 사용됩니다.

LSVM 모델 너깃(대화형 출력)

LSVM 모델을 실행한 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 대상 이름
- 모델 선택 설정에 지정된 모델 빌딩 방법
- 예측변수 수 입력
- 최종 모델에서 예측변수의 개수
- 정규화 유형(L1 또는 L2)
- 페널티 모수(람다). 이는 정규화 모수입니다.
- 회귀분석 정밀도(엡실론). 이 값보다 작은 경우 오차가 허용됩니다. 값이 커지면 모델링 속도는 빨라지지만 정확도가 떨어질 수 있습니다. 목표 필드의 측정 수준이 연속형인 경우에만 사용됩니다.
- 분류 정확도 퍼센트. 이는 분류에만 적용됩니다.
- 평균 제곱 오차. 이는 회귀분석에만 적용됩니다.

레코드 요약

레코드 요약 보기는 모델에서 포함되고 제외되는 레코드(케이스) 수 및 퍼센트에 대한 정보를 제공합니다.

예측변수 중요도

일반적으로, 가장 중요한 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 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.

참고: 예측변수 중요도를 사용하면 LSVM 및 SVM 계산 시간이 다른 유형의 모델보다 길어질 수 있습니다. 이 옵션을 선택하면 특히 큰 데이터 세트에서 성능이 느려질 수 있습니다.

혼돈 행렬

요약표라고도 하는 혼돈 행렬에는 LSVM 분석을 기준으로 각 그룹에 올바르거나 올바르지 않게 지정된 케이스 수가 표시됩니다.

LSVM 모델 설정

SVLM 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 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 다시 데이터를 페치하고 SPSS Modeler에서 스코어를 계산합니다.

제 16 장 최근접 이웃 모델

KNN 노드

최근접 이웃 분석은 다른 케이스와의 유사성을 기준으로 케이스를 분류하는 방법입니다. 머신 학습에서 이 분석 방법은 저장된 모든 패턴이나 케이스와 정확히 일치할 필요가 없는 데이터 패턴을 인식하는 방법으로 개발되었습니다. 유사한 케이스는 서로 가까이에 있고 유사하지 않은 케이스는 서로 멀리 떨어져 있습니다. 따라서 두 케이스 사이의 거리는 두 케이스의 상이성 측도가 됩니다.

서로 인접한 케이스를 "이웃"이라고 합니다. 새 케이스(검증용)가 있는 경우, 해당 모델에서 각 케이스와의 거리가 계산됩니다. 가장 유사한 케이스(최근접 이웃)의 분류가 기록되고 새 케이스가 최근접 이웃의 수가 가장 많은 범주에 배치됩니다.

탐색할 최근접 이웃 수를 지정할 수 있으며, 이 값을 *k*라고 합니다. 그림은 새 케이스가 두 개의 다른 *k* 값을 사용하여 분류되는 방법을 보여줍니다. *k* = 5일 경우, 대부분의 최근접 이웃이 범주 *1*에 속하기 때문에 새 케이스는 범주 *1*에 위치합니다. 그러나 *k* = 9일 경우, 대부분의 최근접 이웃이 범주 *0*에 속하기 때문에 새 케이스는 범주 *0*에 위치합니다.

또한 최근접 이웃 분석은 연속적인 목표 값을 계산하는 데 사용할 수 있습니다. 이 경우, 가장 가까운 이웃의 평균 또는 중앙값 목표 값이 사용되어 새 케이스의 예측값을 가져옵니다.

KNN 노드 목표 옵션

목적 탭은 최근접 이웃 값에 기반하여 입력 데이터에서 목표 필드 값을 예측하는 모델을 작성하거나, 관심이 있는 특정 케이스에 대한 최근접 이웃인 항목을 찾을 수 있는 위치입니다.

어떤 종류의 분석을 수행하시겠습니까?

목표 필드 예측. 최근접 이웃의 값에 기반하여 목표 필드 값을 예측하려는 경우 이 옵션을 선택합니다.

최근접 이웃만 식별. 특정 입력 필드에 대한 최근접 이웃인 항목만 보려는 경우 이 옵션을 선택합니다.

최근접 이웃만 식별하려는 경우 정확도 및 속도와 관련하여 이 탭의 나머지 옵션은 목표 예측에만 관련되어 있으므로 사용되지 않습니다.

귀하의 목표는 무엇입니까?

목표 필드 예측 시 이 옵션 그룹을 사용하면 목표 필드를 예측할 때 가장 중요한 요인이 무엇인지(속도, 정확도 또는 둘 다) 결정할 수 있습니다. 또는 직접 설정을 사용자 정의할 수도 있습니다.

균형, 속도 또는 정확도 옵션을 선택하면 알고리즘은 해당 옵션에 대한 설정을 가장 적절히 혼합한 조합을 미리 선택합니다. 고급 사용자는 이러한 선택을 대체하고자 할 수도 있습니다. 이는 설정 탭의 다양한 패널에서 수행할 수 있습니다.

속도와 정확도의 균형. 작은 범위 내에서 최상의 이웃 수를 선택합니다.

속도. 고정된 이웃 항목 수를 찾습니다.

정확도. 더 광범위한 범위 내 최상의 이웃 수를 선택하고, 거리를 계산할 때 예측변수 중요도를 사용합니다.

사용자 정의 분석. 설정 탭에서 알고리즘을 미세 조정하려면 이 옵션을 선택합니다.

참고: 결과로 생성된 KNN 모델의 크기는 다른 모델과 달리 훈련 데이터의 수량에서 선형으로 증가합니다. KNN 모델을 작성하려는 경우 "메모리 부족" 오류를 보고하는 오류가 나타납니다. 이 경우 IBM SPSS Modeler에서 사용하는 최대 시스템 메모리를 늘리십시오. 이를 수행하려면 다음을 선택하십시오.

도구 > **옵션** > **시스템 옵션**

그리고 **최대 메모리** 필드에 새 크기를 입력하십시오. IBM SPSS Modeler를 다시 시작할 때까지 시스템 옵션 대화 상자에서 변경된 내용은 적용되지 않습니다.

KNN 노드 설정

설정 탭은 최근접 이웃 분석에 특정한 옵션을 지정하는 위치입니다. 화면 왼쪽의 세로 막대에는 옵션을 지정하는 데 사용하는 패널이 나열됩니다.

모델

모델 패널에서는 모델 작성 방법, 사용할 모델(파티셔닝 또는 분할 모델), 필드가 모두 동일한 범위에 포함되도록 숫자 입력 필드의 변환 여부, 관심이 있는 케이스를 관리하는 방법을 제어하는 옵션을 제공합니다. 또한 모델에 대한 사용자 정의 이름을 선택할 수도 있습니다.

참고: 분할된 데이터 사용 및 **케이스 레이블 사용**은 동일한 필드를 사용할 수 없습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 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 30 페이지의 『분할 모델 작성』의 내용을 참조하십시오.

수동으로 필드를 선택하려면... 기본적으로 노드는 유형 노드에서 파티션 및 분할 필드 설정(있는 경우)을 사용하지만 여기서 이 설정을 대체할 수 있습니다. **파티션** 및 **분할** 필드를 활성화하려면 **필드** 탭을 선택하고 **사용자 정의 설정 사용**을 선택하고 여기로 리턴하십시오.

- **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 표본

을 사용하여 이를 검정함으로써 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 분할로 설정하는 것과 유사합니다. **플래그**, **명목** 또는 **순서** 유형의 필드만 분할 필드로 지정할 수 있습니다. 분할 필드로 선택된 필드는 목표, 입력, 파티션, 빈도 또는 가중 필드로 사용할 수 없습니다. 자세한 정보는 30 페이지의 『분할 모델 작성』의 내용을 참조하십시오.

범위 입력 요소 정규화. 연속형 입력 필드에 대한 값을 표준화하려면 이 상자를 선택합니다. 정규화된 기능에는 추정 알고리즘 성능을 개선할 수 있는 동일한 범위의 값이 있습니다. 조정된 정규화, [2*(x–min)/(max–min)]–1이 사용됩니다. 조정된 정규화 값은 -1과 1 사이의 범위에 있습니다.

케이스 레이블 사용. 드롭 다운 목록을 사용하려면 이 상자를 선택합니다. 이 드롭 다운 목록을 통해 모델 뷰어에 있는 예측변수 공간 차트, 동위 차트, 사분면 맵에서 관심이 있는 케이스를 식별하기 위해 해당 값을 레이블로 사용하는 필드를 선택할 수 있습니다. 측정 수준이 명목, 순서 또는 플래그인 필드를 선택하여 레이블 필드로 사용할 수 있습니다. 여기서 필드를 선택하지 않으면 소스 데이터에서 행 번호로 식별되는 최근접 이웃이 있는 레코드가 모델 뷰어 차트에 표시됩니다. 모델 작성 후 데이터를 조작하려는 경우 표시에서 케이스를 식별하려고 할 때마다 소스 데이터를 다시 참조하지 않도록 케이스 레이블을 사용합니다.

초점 레코드 식별. 드롭 다운 목록을 사용하려면 이 상자를 선택합니다. 이 드롭 다운 목록을 통해 관심이 있는 특정 입력 필드(플래그 필드만 해당)를 표시할 수 있습니다. 여기서 필드를 지정하면 모델을 작성할 때 해당 필드를 나타내는 점이 모델 뷰어에서 처음에 선택됩니다. 여기서 초점 레코드 선택은 선택사항입니다. 모델 뷰어에서 수동으로 선택할 경우 포인트는 임시로 초점 레코드가 될 수 있습니다.

이웃

이웃 패널에는 최근접 이웃 수를 계산하는 방법을 제어하는 옵션 세트가 있습니다.

최근접 이웃 수(k). 특정 케이스에 대한 최근접 이웃 수를 지정합니다. 많은 수의 이웃을 사용한다고 해서 반드시 더 정확한 모델을 얻을 수 있는 것은 아님에 유의하십시오.

목표가 목표 예측인 경우 두 가지 선택 사항이 제공됩니다.

- **고정된 k 지정.** 찾으려는 최근접 이웃의 고정된 수를 지정하려는 경우 이 옵션을 사용합니다.
- **k 자동 선택.** 대신 **최소** 및 **최대** 필드를 사용하여 값의 범위를 지정하고 프로시저를 통해 해당 범위 내에서 "최상"의 이웃 수를 선택할 수 있습니다. 최근접 이웃 수를 판별하는 방법은 필드선택 패널에서 필드선택을 요청하는지 여부에 따라 달라집니다.

  필드선택이 유효하면 요청된 범위에서 *k*의 모든 값에 대해 필드선택이 수행되고 *k* 및 동반되는 기능 세트(오차율 또는 목표가 연속형인 경우 오차제곱합이 가장 낮음)가 선택됩니다.

필드선택이 유효하지 않으면 *V*-검증(교차 검증)을 사용하여 "최상"의 이웃 수를 선택합니다. 중첩 지정에 대한 제어에 관해서는 교차 검증 패널을 참조하십시오.

거리 계산. 이는 케이스의 유사성 측도에 사용되는 거리 메트릭을 지정하는 데 사용되는 메트릭입니다.

- **유클리디안 거리.** 두 케이스(x와 y) 간 거리는 케이스 값 간 차이 제곱값 합(모든 차원에 대한)의 제곱근입니다.
- **시티 블록 거리.** 두 케이스 간 거리는 케이스 값 간 절대 차의 합(모든 차원에 대한)입니다. 이를 Manhattan 거리라고도 합니다.

선택적으로 목표가 목표 예측인 경우 거리를 계산할 때 정규화 중요도로 기능의 가중치를 부여하도록 할 수 있습니다. 예측변수의 기능 중요도는 전체 모델의 오차율 또는 오차제곱합에 대한 모델에서 예측변수가 제거된 모델의 오차율 또는 오차제곱합으로 계산됩니다. 정규화 중요도는 합이 1이 되도록 기능 중요도 값을 다시 부여하여 계산합니다.

거리 계산 시 중요도별로 변수 가중치. (목표가 목표 예측인 경우에만 표시됩니다.) 이웃 사이의 거리를 계산할 때 예측변수 중요도를 사용하려는 경우 이 상자를 선택합니다. 그러면 예측변수 중요도는 모델 너깃에 표시되고 예측에 사용됩니다. 결과적으로 스코어링에도 영향을 줍니다. 자세한 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

범위 목표 예측. (목표가 목표 예측인 경우에만 표시됩니다.) 연속형 목표(숫자 범위)를 지정하면 최근접 이웃의 평균 또는 중앙값 중 예측값을 계산할 때 사용할 항목을 정의합니다.

필드선택

이 패널은 목표가 목표 예측인 경우에만 활성화됩니다. 이를 통해 필드선택에 대한 옵션을 요청하고 지정할 수 있습니다. 기본적으로 모든 기능은 필드선택을 위해 준비되어 있지만 선택적으로 기능의 서브세트를 선택하여 모델에 적용할 수 있습니다.

변수 선택 수행. 필드선택 옵션을 사용하려면 이 상자를 선택합니다.

- **강제 입력.** 이 상자 옆의 필드 선택기 단추를 클릭하고 모델에 적용할 하나 이상의 기능을 선택하십시오.

중지 기준. 각 단계에서, 모델에 추가할 경우 가장 작은 오차를 생성하는 기능(범주형 목표의 경우 오차율, 연속형 목표의 경우 오차제곱합으로 계산됨)은 모델 세트에 포함하도록 고려됩니다. 전진 선택법은 지정 조건이 충족될 때까지 지속됩니다.

- **지정된 기능 수를 선택하면 중지.** 알고리즘은 고정된 기능 수는 물론 모델에 적용된 기능 수를 추가합니다. 양수를 지정하십시오. 선택할 기능 수에 대한 값을 줄이면 경제적인 모델을 만들 수 있지만 중요한 기능이 빠질 위험이 있습니다. 선택할 기능 수에 대한 값을 증가시키면 중요한 기능을 모두 얻을 수 있지만 실제로 모델 오차를 증가시키는 기능을 추가하게 될 위험이 있습니다.
- **절대 오차 비율 변경이 최소값 이하인 경우 중지.** 절대 오차 비율 변경이 기능을 더 추가해도 모델을 개선할 수 없음을 의미할 경우 알고리즘이 중지됩니다. 양수를 지정하십시오. 최소 변화량을 줄이면 더 많은 기능을 포함할 수 있지만 모델에 많은 값을 추가하지 않는 기능이 포함될 위험도 있

습니다. 최소 변경량을 증가시키면 많은 기능을 채택하지 않을 수 있지만 모델에 중요한 기능을 잃을 위험이 있습니다. 최소 변화량에 대한 "최적"의 값은 데이터 및 애플리케이션에 따라 달라집니다. 어떤 기능이 가장 중요한 기능인지 판단하려면 결과의 필드선택 오차 로그를 참조하십시오. 자세한 정보는 395 페이지의 『예측변수 선택 오차 로그』의 내용을 참조하십시오.

교차 검증

이 패널은 목표가 목표 예측인 경우에만 활성화됩니다. 이 패널의 옵션은 최근접 이웃을 계산할 때 교차 검증을 사용할지 여부를 제어합니다.

교차 검증은 표본을 다수의 부표본 또는 **중첩**으로 나눕니다. 그런 다음 최근접 이웃 모델을 생성하고 각 부표본에서 차례대로 데이터를 제외합니다. 첫 번째 모델은 첫 번째 표본 중첩의 케이스를 제외한 모든 케이스를 기준으로 하며, 두 번째 모델은 두 번째 표본 중첩의 케이스를 제외한 모든 케이스를 기준으로 하는 방식입니다. 각 모델의 경우 모델 생성에서 제외된 부표본에 적용하여 해당 모델의 오차를 추정합니다. "최적"의 가장 가까운 모델 수는 중첩 전체에서 가장 낮은 오차를 생성하는 수입니다.

교차 검증 중첩. *V*-중첩 교차 검증은 "최적"의 이웃 수를 결정하는 데 사용됩니다. 교차 검증 중첩은 성능 문제로 인해 필드선택과 함께 사용할 수 없습니다.

- **중첩에 케이스 무작위 할당.** 교차 검증에 사용할 중첩 수를 지정합니다. 프로시저에서는 케이스를 무작위로 중첩에 할당합니다. 중첩에는 1부터 *V*까지 번호가 매겨집니다.
- **난수 시드 설정.** 무작위 퍼센트에 기반한 모델의 정확도를 추정할 때 이 옵션을 사용하면 다른 세션에서 동일한 결과를 복제할 수 있습니다. 난수 생성기에서 사용하는 시작값을 지정하면 노드를 실행할 때마다 동일한 레코드를 지정하도록 보장할 수 있습니다. 원하는 시드 값을 입력하십시오. 이 옵션을 선택하지 않으면 노드를 실행할 때마다 다른 표본이 생성됩니다.
- **필드를 사용하여 케이스 할당.** 활성 데이터 세트의 각 케이스를 폴더에 지정하는 숫자 필드를 지정합니다. 필드는 숫자여야 하며, 1부터 *V*까지의 값을 사용합니다. 이 범위의 값이 결측되고, 분할 모델이 유효한 경우 분할 필드에서 이로 인해 오류가 발생합니다.

분석

분석 패널은 목표가 목표 예측인 경우에만 활성화됩니다. 이를 사용하여 모델이 다음을 포함하도록 추가 변수를 포함할 것인지 여부를 지정할 수 있습니다.

- 가능한 각 목표 필드 값의 확률
- 케이스와 최근접 이웃 사이의 거리
- 원시 및 수정된 성향 스코어(플래그 목표만 해당)

모든 확률 추가. 이 옵션을 선택하면 명목 또는 플래그 목표 필드의 가능한 모든 값에 대한 확률이 노드에서 처리하는 각 레코드에서 표시됩니다. 이 옵션을 선택 해제하면 예측값 및 해당 확률만 명목 또는 플래그 목표 필드에 표시됩니다.

케이스 및 k 최근접 이웃 사이의 거리 저장. 각 초점 레코드의 경우 각 초점 레코드의 *k* 최근접 이웃(훈련 표본에서) 및 대응하는 *k* 최근접 거리에 대해 별도의 변수가 작성됩니다.

성향 스코어

성향 스코어는 모델링 노드 및 모델 너깃의 설정 탭에서 사용 가능합니다. 이 기능은 선택한 목표가 플래그 필드인 경우에만 사용 가능합니다. 자세한 정보는 38 페이지의 『성향 스코어』의 내용을 참조하십시오.

원시 성향 스코어 계산. 원시 성향 스코어는 훈련 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 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 훈련 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 138 페이지의 『부스팅 C5.0 모델』의 내용을 참조하십시오.

KNN 모델 너깃

KNN 모델은 다음 표에 표시된 대로, 여러 새 필드를 작성합니다. 이러한 필드와 해당 값을 보려면 KNN 모델 너깃에서 테이블 노드를 추가하고 테이블 노드를 실행하거나 너깃에서 미리보기 단추를 클릭하십시오.

표 *32. KNN* 모델 필드

새 필드 이름	설명
$KNN-*fieldname*	목표 필드의 예측값.
$KNNP-*fieldname*	예측값의 확률.

표 *32. KNN* 모델 필드 *(*계속*)*

새 필드 이름	설명
$KNNP-*value*	명목 또는 플래그 필드에서 각 가능한 값의 확률. 모델 너깃의 설정 탭에서 **모든 확률 추가**가 선택된 경우에만 포함됩니다.
$KNN-neighbor-*n*	초점 레코드에서 *n*번째로 인접한 이웃의 이름. 모델 너깃의 설정 탭에서 **최근접**이 0이 아닌 값으로 설정된 경우에만 포함됩니다.
$KNN-distance-*n*	초점 레코드에서 *n*번째로 인접한 이웃 초점 레코드와의 상대적 거리. 모델 너깃의 설정 탭에서 **최근접**이 0이 아닌 값으로 설정된 경우에만 포함됩니다.

최근접 이웃 모델 보기

모델 보기

모델 보기에는 2-패널 창이 있습니다.

- 첫 번째 패널에서는 기본 보기라고 불리는 모델 개요가 표시됩니다.
- 두 번째 패널에서는 두 가지 보기 유형 중 하나가 표시됩니다.

  보조 모델 보기는 모델에 대한 자세한 정보를 보여줍니다. 단 모델 자체에 초점을 맞추지는 않습니다.

  연결된 보기는 사용자가 기본 보기 부분에서 드릴다운할 때 해당 모델의 특정 기능에 대한 자세한 내용을 보여줍니다.

기본적으로, 첫 번째 패널에는 예측변수 공간이 표시되고 두 번째 패널에는 예측변수 중요도 차트가 표시됩니다. 예측변수 중요도 차트를 사용할 수 없는 경우(즉, **중요도별 기중치 기능**이 설정 탭의 이웃 패널에서 선택되지 않은 경우, 보기 드롭다운에서 첫 번째 사용 가능한 보기가 표시됩니다.

보기에 어떤 사용 가능한 정보도 없는 경우 보기 드롭다운에서 생략됩니다.

예측변수 공간: 예측변수 공간 차트는 예측변수 공간(또는 네 개 이상의 예측 변수가 있는 하위 공간)의 대화형 그래프입니다. 각각의 축은 모델에서 예측변수를 나타내고, 차트에서 포인트의 위치는 훈련 및 검증 분할에서 케이스에 대한 해당 예측변수의 값을 표시합니다.

키. 예측변수 값 외에도, 도표의 포인트는 다른 정보를 전달합니다.

- 형태는 포인트가 속해 있는 파티션(훈련 또는 검증)을 나타냅니다.
- 포인트의 색상/음영은 해당 케이스에 대한 목표 값을 나타냅니다. 이 경우 개별 색상값은 범주형 목표의 범주를 나타내며 음영은 연속적인 목표의 범위값을 나타냅니다. 훈련 파티션에 대해 표시된 값은 관측값입니다. 검증용 파티션의 경우에는 예측값이 됩니다. 목표가 지정되지 않은 경우 이 키가 표시되지 않습니다.
- 윤곽이 많다는 것은 케이스가 초점이라는 것을 의미합니다. 초점 레코드는 해당되는 *k*개의 최근접 이웃에 링크되어 표시됩니다.

제어 및 상호작용성. 차트의 많은 제어는 예측변수 공간을 탐색하도록 허용합니다.

- 차트에 표시할 예측변수 서브세트를 선택하고 차원에 표시되는 예측변수를 변경할 수 있습니다.
- "초점 레코드"는 단지 예측변수 공간 차트에서 선택되는 포인트입니다. 초점 레코드 변수를 지정한 경우, 포컬 레코드를 표시하는 포인트는 초기에 선택됩니다. 그러나 포인트는 사용자가 선택하는 경우 임시로 초점 레코드가 될 수 있습니다. 포인트 선택에 대한 "보통의" 제어가 적용됩니다. 포인트를 클릭하면 해당 포인트가 선택되고 다른 모든 포인트는 선택 취소됩니다. Ctrl 키를 누른 상태에서 포인트를 클릭하면 선택된 포인트 세트에 포인트가 추가됩니다. 동위 차트와 같은 링크된 보기는 예측변수 공간에서 선택된 케이스에 따라 자동으로 업데이트됩니다.
- 초점 레코드에 대해 표시할 최근접 이웃 수(*k*)를 변경할 수 있습니다.
- 차트에서 마우스를 포인트 위에 두면 케이스 레이블값 또는 케이스 레이블이 정의되지 않은 경우 케이스 번호, 그리고 관측 및 예측 목표 값을 포함한 도구 팁이 표시됩니다.
- "재설정" 단추를 사용하여 예측변수 공간을 해당되는 원래 상태로 되돌릴 수 있습니다.

예측변수 공간 차트에서 축 변경: 예측변수 공간 차트의 축에 표시되는 기능을 제어할 수 있습니다.

축 설정을 변경하려면 다음을 수행하십시오.

1. 예측변수 공간에 대해 편집 모드를 선택하려면 왼쪽 패널에서 편집 모드 단추(페인트 붓 아이콘)를 클릭하십시오.
2. 오른쪽 패널에서 (어떤 것에 대한) 보기를 변경하십시오. **구역 표시** 패널은 두 개의 주요 패널 사이에 나타납니다.
3. **구역 표시** 확인 상자를 클릭하십시오.
4. 예측변수 공간에서 데이터 점을 클릭하십시오.
5. 동일한 데이터 유형의 예측변수로 축을 바꾸려면 다음을 수행하십시오.
   - 바꾸려고 하는 예측변수의 구역 레이블(작은 X 단추가 있는) 위로 새 예측변수를 끌어오십시오.
6. 축을 다른 데이터 유형의 예측변수로 바꾸려면 다음을 수행하십시오.
   - 바꾸려고 하는 예측변수의 구역 레이블에서 작은 X 단추를 클릭하십시오. 예측변수 공간은 2차원 보기로 변경됩니다.
   - **차원 추가** 구역 레이블 위로 새 예측변수를 끌어오십시오.
7. 편집 모드에서 나가려면 왼쪽 패널에서 탐색 모드 단추(화살촉 아이콘)를 클릭하십시오.

예측변수 중요도: 일반적으로, 가장 중요한 예측변수 필드에 모델링 노력을 집중하고 가장 쓸모 없는 예측변수를 삭제하거나 무시하기를 원합니다. 예측변수 중요도 차트를 사용하면 모델 추정 시 각 예측변수의 상대적인 중요도를 표시하여 원하는 작업을 수행할 수 있습니다. 값이 상대적이므로 표시된 모든 예측변수에 대한 값의 합은 1.0이 됩니다. 예측변수 중요도는 모델 정확도와는 관련이 없습니다. 단지 예측 시 각 예측변수의 중요도와 관련이 있으며 예측이 정확한지, 정확하지 않은지 여부와는 관련이 없습니다.

최근접 이웃 거리: 이 테이블은 초점 레코드에 대해서만 *k* 최근접 이웃 및 거리를 표시합니다. 초점 레코드 ID가 모델링 노드에 지정된 경우에만 사용 가능하며, 이 변수에 의해 식별된 초점 레코드만 표시합니다.

각 행:

- **초점 레코드** 열에는 초점 레코드에 대한 케이스 레이블 변수의 값이 포함됩니다. 케이스 레이블이 정의되지 않으면, 이 열에는 초점 레코드의 케이스 번호가 포함됩니다.
- **최근접 이웃** 그룹 아래에서 *i*번째 열에는 초점 레코드의 *i*번째 최근접 이웃에 대한 케이스 레이블링 변수의 값이 포함됩니다. 케이스 레이블이 정의되지 않으면, 이 열에는 초점 레코드의 *i*번째 최근접 이웃의 케이스 번호가 포함됩니다.
- **최근접 거리** 그룹 아래에서 *i*번째 열에는 초점 레코드까지의 *i*번째 최근접 이웃 거리가 포함됩니다.

동위: 이 차트는 각 예측변수 및 목표에 대해 초점 케이스와 해당되는 *k*개의 최근접 이웃을 표시합니다. 예측변수 공간에서 초점 케이스를 선택하는 경우에 사용할 수 있습니다.

동위 도표는 두 가지 방식으로 예측변수 공간에 연결됩니다.

- 예측변수 공간에서 선택된 케이스(초점)는 해당되는 *k*개의 최근접 이웃과 함께 동위 도표에 표시됩니다.
- 예측변수 공간에서 선택된 *k*의 값이 동위 도표에서 사용됩니다.

예측변수 선택. 동위 도표에서 표시할 예측변수를 선택할 수 있습니다.

4분위 맵: 이 차트는 *y*축에 목표가 있고 *x*축에 척도 예측변수가 있는(패널 기준: 예측변수) 산점도(또는 목표의 측정 수준에 따라 점도표)에 초점 케이스와 해당되는 *k* 최근접 이웃을 표시합니다. 목표가 있고 초점 케이스가 예측변수 공간에서 선택된 경우에 사용할 수 있습니다.

- 훈련 파티션의 변수 평균에서 연속적인 변수에 대해 참조선이 그려집니다.

예측변수 선택. 4분위 맵에서 표시할 예측변수를 선택할 수 있습니다.

예측변수 선택 오차 로그: 차트 위의 포인트는 예측변수가 *x*축 상에 나열된(또한 *x*축 왼쪽에 모든 기능이 나열된) 모델의 *y*축에 오류(목표의 측정 수준에 따라 오차율 또는 오차제곱합)를 표시합니다. 목표와 필드선택이 유효하면 이 차트를 사용할 수 있습니다.

분류표: 이 테이블에는 파티션 기준으로 목표의 관측값 대 예측값의 교차 분류가 표시됩니다. 목표가 있고 범주형(플래그, 명목 또는 순서)인 경우에 사용할 수 있습니다.

- 검증용 파티션의 **(결측)** 행에는 목표에 대한 결측값을 가진 검증 케이스가 포함됩니다. 이러한 케이스는 검증용 표본에 기여하지만, 전체 퍼센트 값은 정확도 퍼센트 값에 기여하지 않습니다.

오류 요약: 목표변수가 있는 경우 이 테이블을 사용할 수 있습니다. 이 테이블은 모델과 연관되는 오류를 표시합니다(연속형 목표의 제곱합과 범주형 목표의 오차율(100% – 전체 정확도 퍼센트)).

KNN 모델 설정

설정 탭을 사용하면 결과를 볼 때 표시할 추가 필드를 지정할 수 있습니다(예: 너깃에 첨부된 테이블 노드 실행). 이러한 각 옵션의 효과는 옵션을 선택하고 미리보기 단추를 클릭하여 확인할 수 있습니다. 미리보기 출력의 오른쪽으로 스크롤하면 추가 필드가 나옵니다.

모든 확률 추가(범주형 목표에만 유효함). 이 옵션을 선택하면 명목 또는 플래그 목표 필드의 가능한 모든 값에 대한 확률이 노드에서 처리하는 각 레코드에서 표시됩니다. 이 옵션을 선택 해제하면 예측 값 및 해당 확률만 명목 또는 플래그 목표 필드에 표시됩니다.

이 확인 상자의 기본 설정은 모델링 노드에서 대응하는 확인 상자로 판별됩니다.

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.

최근접 표시. 이 값을 *n*으로 설정한 경우(여기서, *n*은 0이 아닌 양의 정수임) 초점 레코드에 대한 *n* 최근접 이웃은 초점 레코드와의 상대적 거리와 함께 모델에 포함됩니다.

제 17 장 Python 노드

SPSS Modeler는 Python 원시 알고리즘을 사용하는 노드를 제공합니다. 노드 팔레트의 **Python** 탭은 Python 알고리즘을 실행하는 데 사용할 수 있는 다음 노드를 포함합니다. 이러한 노드는 Windows 64, Linux64 및 Mac에서 지원됩니다.

SMOTE(Synthetic Minority Over-sampling Technique) 노드는 불균형 데이터 세트를 처리하기 위해 초과 표본추출 알고리즘을 제공합니다. 또한 데이터 균형을 조정하기 위한 고급 방법을 제공합니다. SPSS Modeler의 SMOTE 프로세스 노드는 Python으로 구현되며, `imbalanced-learn`© Python 라이브러리가 필요합니다.

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.

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.

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.

랜덤 포리스트 노드는 트리 모델을 기본 모델로 사용하는 배깅 알고리즘의 고급 구현을 사용합니다. SPSS Modeler의 이 랜덤 포리스트 모델링 노드는 Python으로 구현되며, `scikit-learn`© Python 라이브러리가 필요합니다.

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.

SMOTE 노드

SMOTE(Synthetic Minority Over-sampling Technique) 노드는 불균형 데이터 세트를 처리하기 위해 초과 표본추출 알고리즘을 제공합니다. 또한 데이터 균형을 조정하기 위한 고급 방법을 제공합니다. SMOTE 프로세스 노드는 Python으로 구현되며, `imbalanced-learn`© Python 라이브러리가 필요합니다. imbalanced-learn 라이브러리에 대한 상세 정보는 http://contrib.scikit-learn.org/imbalanced-learn/about.html[1]을 참조하십시오.

노드 팔레트의 Python 탭은 SMOTE 노드와 다른 Python 노드로 구성됩니다.

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5(http://jmlr.org/papers/v18/16-365.html)

SMOTE 노드 설정

SMOTE 노드의 **설정** 탭에서 다음 설정을 정의하십시오.

대상 설정

목표 필드 목표 필드를 선택하십시오. 플래그, 명목, 순서, 및 이산형 측정 유형이 모두 지원됩니다. 파티션 섹션에서 **파티션된 데이터 사용** 옵션을 선택한 경우 훈련 데이터만 초과 표본추출됩니다.

초과 표본 비율통계량

초과 표본 비율을 자동으로 선택하려면 **자동**을 선택하고, 사용자 정의 비율 값을 설정하려면 **비율통계량 설정(소수/다수)**을 선택하십시오. 비율은 다수 클래스의 표본 수 대비 소수 클래스의 표본 수입니다. 값은 **0**보다 크고 **1**보다 작거나 같아야 합니다.

난수 시드

난수 시드 설정 난수 생성기에 사용될 시드를 생성하려면 이 옵션을 선택한 후 **생성**을 클릭하십시오.

방법

알고리즘 종류 사용할 SMOTE 알고리즘의 유형을 선택하십시오.

표본 규칙

K 이웃 합성 표본을 생성하는 데 사용할 최근접 이웃의 수를 지정하십시오.

M 이웃 소수 표본이 위험한지 판별하는 데 사용할 최근접 이웃의 수를 지정하십시오. 이 옵션은 **Borderline1** 또는 **Borderline1** SMOTE 알고리즘 유형을 선택한 경우에만 사용됩니다.

파티션

파티션된 데이터 사용. 훈련 데이터를 초과 표본추출하려는 경우에만 이 옵션을 선택하십시오.

SMOTE 노드에는 imbalanced-learn© Python 라이브러리가 필요합니다. 다음 표는 SPSS Modeler SMOTE 노드 대화 상자의 설정과 Python 알고리즘 간의 관계를 보여줍니다.

표 *33. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	Python API 매개변수 이름
초과 표본 비율통계량(숫자 입력 컨트롤)	sample_ratio_value	ratio
난수 시드	random_seed	random_state
K_Neighbours	k_neighbours	k
M_Neighbours	m_neighbours	m
알고리즘 종류	algorithm_kind	kind

XGBoost Linear 노드

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.

부스팅 알고리즘에 대한 자세한 정보는 http://xgboost.readthedocs.io/en/latest/tutorials/index.html에서 제공하는 XGBoost 자습서를 참조하십시오. [1]

XGBoost 교차 검증 기능은 SPSS Modeler에서 지원되지 않습니다. 이 기능에는 SPSS Modeler 파티션 노드를 사용할 수 있습니다. 또한 SPSS Modeler의 XGBoost는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost Linear 노드 필드

필드 탭은 분석에서 사용되는 필드를 지정합니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표 및 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 목표 및 예측변수를 수동으로 할당하려면 이 옵션을 선택하십시오.

필드 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 목표 및 예측변수 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 사용할 필드를 선택하십시오.

예측변수 예측의 입력으로 하나 이상의 필드를 선택하십시오.

XGBoost Linear 노드 작성 옵션

작성 옵션 탭에서는 선형 부스트 매개변수 및 모델 작성 등의 **기본 옵션** 및 목적에 대한 **학습 작업 옵션**을 포함한 XGBoost Linear 노드에 대한 작성 옵션을 지정할 수 있습니다. 이러한 옵션에 대한 추가 정보는 다음 온라인 자원을 참조하십시오.

- XGBoost 매개변수 참조[1]
- XGBoost Python API[2]
- XGBoost 홈 페이지[3]

기본

하이퍼-모수 최적화(Rbfopt 기준). 모델이 표본에 대해 기대빈도 또는 하한 오차율을 달성할 수 있도록 모수의 최적 조합을 자동으로 검색하는 Rbfopt 기준 하이퍼-모수 최적화를 사용하려면 이 옵션을 선택하십시오. Rbfopt에 대한 세부사항은 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html의 내용을 참조하십시오.

알파 가중치에 대한 L1 정규화 항입니다. 이 값을 늘리면 더 보수적인 모델이 생성됩니다.

람다 가중치에 대한 L2 정규화 항입니다. 이 값을 늘리면 더 보수적인 모델이 생성됩니다.

람다 편향 편향에 대한 L2 정규화 항입니다. (편향에 대한 L1 정규화 항은 중요하지 않으므로 없습니다.)

숫자 부스트 반올림. 부스팅 반복 횟수입니다.

학습 작업

목적 **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic**, **multi** 중 하나를 학습 작업 목적 유형으로 선택하십시오.

난수 시드 난수 생성기에 사용될 시드를 생성하려면 **생성**을 클릭하십시오.

다음 표는 SPSS Modeler XGBoost Linear 노드 대화 상자의 설정과 Python XGBoost 라이브러리 매개변수 간의 관계를 보여줍니다.

표 *34. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	XGBoost 매개변수
목표	TargetField	
예측변수	InputFields	
람다	lambda	lambda
알파	alpha	alpha
람다 편향	lambdaBias	lambda_bias
숫자 부스트 반올림	numBoostRound	num_boost_round
목적	objectiveType	objective
난수 시드	random_seed	seed

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

XGBoost Linear 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.

XGBoost Tree 노드

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.

부스팅 알고리즘에 대한 자세한 정보는 http://xgboost.readthedocs.io/en/latest/tutorials/index.html에서 제공하는 XGBoost 자습서를 참조하십시오. [1]

XGBoost 교차 검증 기능은 SPSS Modeler에서 지원되지 않습니다. 이 기능에는 SPSS Modeler 파티션 노드를 사용할 수 있습니다. 또한 SPSS Modeler의 XGBoost는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost Tree 노드 필드

필드 탭은 분석에서 사용되는 필드를 지정합니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표 및 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 목표 및 예측변수를 수동으로 할당하려면 이 옵션을 선택하십시오.

필드 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 목표 및 예측변수 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 사용할 필드를 선택하십시오.

예측변수 예측의 입력으로 하나 이상의 필드를 선택하십시오.

XGBoost Tree 노드 작성 옵션

작성 옵션 탭에서는 모델 작성 및 트리 성장을 위한 **기본 옵션**, 목적에 대한 **학습 작업 옵션** 및 과적합 제어 및 불균형 데이터 세트 처리를 위한 **고급 옵션**을 포함한 XGBoost Tree 노드에 대한 작성 옵션을 지정할 수 있습니다. 이러한 옵션에 대한 추가 정보는 다음 온라인 자원을 참조하십시오.

- XGBoost 매개변수 참조[1]
- XGBoost Python API[2]
- XGBoost 홈 페이지[3]

기본

하이퍼-모수 최적화(Rbfopt 기준). 모델이 표본에 대해 기대빈도 또는 하한 오차율을 달성할 수 있도록 모수의 최적 조합을 자동으로 검색하는 Rbfopt 기준 하이퍼-모수 최적화를 사용하려면 이 옵션을 선택하십시오. Rbfopt에 대한 세부사항은 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html의 내용을 참조하십시오.

트리 방법 사용할 XGBoost Tree 생성 알고리즘을 선택하십시오.

숫자 부스트 반올림 부스팅 반복 횟수를 지정하십시오.

최대 깊이 최대 트리 깊이를 지정하십시오. 이 값을 늘리면 모델이 더 복잡해져서 과적합이 발생할 수 있습니다.

최소 하위 가중치 하위에 필요한 인스턴스 가중치(hessian)의 최소 합계를 지정하십시오. 트리 파티션 단계로 인해 인스턴스 가중치의 합계가 이 **최소 하위 가중치**보다 적은 리프 노드가 생성된 경우 작성 프로세스가 추가 파티셔닝을 중단합니다. 선형 회귀 모드에서 이 가중치는 단순하게 각 노드에 필요한 최소 인스턴스 수에 해당합니다. 가중치가 클수록 더 보수적인 알고리즘이 생성됩니다.

최대 델타 단계 각 트리의 가중치 추정에 허용할 최대 델타 단계를 지정하십시오. **0**으로 설정할 경우 제한조건이 없습니다. 양수 값으로 설정할 경우 업데이트 단계를 더 보수적으로 설정할 수 있습니다. 이 매개변수는 보통 필요하지 않지만, 클래스의 균형이 극도로 맞지 않는 경우 로지스틱 회귀분석에 유용할 수 있습니다.

학습 작업

목적 **reg:linear**, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic**, **multi** 중 하나를 학습 작업 목적 유형으로 선택하십시오.

난수 시드 난수 생성기에 사용될 시드를 생성하려면 **생성**을 클릭하십시오.

고급

하위 표본 하위 표본은 훈련 인스턴스의 비율입니다. 예를 들어 이 값을 **0.5**로 설정할 경우, XGBoost는 트리 성장을 위해 데이터 인스턴스의 절반을 무작위로 수집하며 이로 인해 과적합이 방지됩니다.

에타 과적합 방지를 위해 업데이트 단계 중에 사용되는 단계 크기 축소입니다. 부스팅 단계 후마다 새로운 기능의 가중치를 직접 가져올 수 있습니다. 에타는 또한 부스팅 프로세스를 더 보수적으로 만들기 위해 기능 가중치도 축소합니다.

감마 트리의 리프 노드에 추가 파티션을 만드는 데 필요한 최소 손실 감소입니다. 감마 설정이 클수록 더 보수적인 알고리즘이 생성됩니다.

트리별 Colsample 각 트리를 생성할 때 열의 하위 표본 비율입니다.

수준별 Colsample 각 수준에서 각 분할에 대한 열의 하위 표본 비율입니다.

람다 가중치에 대한 L2 정규화 항입니다. 이 값을 늘리면 더 보수적인 모델이 생성됩니다.

알파 가중치에 대한 L1 정규화 항입니다. 이 값을 늘리면 더 보수적인 모델이 생성됩니다.

척도 양수 가중치 양수 및 음수 가중치의 균형을 제어합니다. 불균형 클래스에 유용합니다.

다음 표는 SPSS Modeler XGBoost Tree 노드 대화 상자의 설정과 Python XGBoost 라이브러리 매개변수 간의 관계를 보여줍니다.

표 *35. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	XGBoost 매개변수
목표	TargetField	
예측변수	InputFields	
트리 방법	treeMethod	tree_method
숫자 부스트 반올림	numBoostRound	num_boost_round
최대 깊이	maxDepth	max_depth
최소 하위 가중치	minChildWeight	min_child_weight
최대 델타 단계	maxDeltaStep	max_delta_step
목적	objectiveType	objective
난수 시드	random_seed	seed
하위 표본	sampleSize	subsample
에타	eta	eta
감마	gamma	gamma
트리별 Colsample	colsSampleRatio	colsample_bytree
수준별 Colsample	colsSampleLevel	colsample_bylevel
람다	lambda	lambda
알파	alpha	alpha
척도 양수 가중치	scalePosWeight	scale_pos_weight

[1] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[2] "Plotting API" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

XGBoost Tree 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.

---

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 맵에서 많은 척도로 구조 표시
- 다중, 이형, 매니폴드 또는 군집에 있는 데이터 표시
- 중심에 포인트를 모으는 경향을 줄임

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.

t-SNE 노드 고급 옵션

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

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

방법. Barnes Hut 또는 **정확**을 선택하십시오. 기본적으로 기울기 계산 알고리즘은 정확한 값을 찾는(정확) 방법보다 훨씬 빠르게 실행되는 Barnes-Hut 근사값을 사용합니다. Barnes-Hut 근사값을 사용하면 t-SNE 기술이 대형 실제 데이터 세트에 적용될 수 있습니다. 정확 알고리즘은 최근접 이웃 오류를 방지하여 더 정확하게 작업을 수행합니다.

초기화. 임베드 초기화에 대해 **임의** 또는 **PCA**를 선택하십시오.

목표 필드 출력 그래프에 대한 컬러 맵으로 표시할 목표 필드를 선택하십시오. 여기서 목표 필드가 지정되지 않으면 그래프가 단색으로 표시됩니다.

최적화

당혹도. 당혹도(perlexity)는 기타 매니폴드 학습 알고리즘에서 사용되는 최근접 이웃 수와 연관됩니다. 일반적으로 데이터 세트가 클수록 더 큰 당혹도가 필요합니다. **5**와 **50** 사이의 값을 선택하는 것을 고려해 보십시오. 기본값은 **30**이고, 범위는 **2 - 9999999**입니다.

조기 과장. 이 설정은 원래 공간의 자연 군집이 임베드된 공간에서 얼마나 타이트할지 및 그 사이에 얼마나 많은 공간이 있을지를 제어합니다. 기본값은 **12**이고, 범위는 **2 - 9999999**입니다.

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

최대 반복. 최적화에 대한 최대 반복 수입니다. 기본값은 **1000**이고, 범위는 **250 - 9999999**입니다.

각도 크기. 한 포인트에서 측정된 원거리 노드의 각도 크기입니다. **0**과 **1** 사이의 값을 입력하십시오. 기본값은 **0.5**입니다.

난수 시드

난수 시드 설정 난수 생성기에 사용될 시드를 생성하려면 이 옵션을 선택한 후 **생성**을 클릭하십시오.

최적화 중단 조건

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

최소 기울기 노름. 기울기 노름이 이 최소 임계값 미만이면 최적화가 중단됩니다. 기본값은 **1.0E-7**입니다.

메트릭. 기능 배열에서 인스턴스 사이의 거리를 계산할 때 사용할 메트릭입니다. 메트릭이 문자열이면 메트릭 매개변수 또는 `pairwise.PAIRWISE_DISTANCE_FUNCTIONS`에 나열된 메트릭의 경우 `scipy.spatial.distance.pdist`에 허용된 옵션 중 하나여야 합니다. 사용 가능한 메트릭 유형 중 하나를 선택하십시오. 기본값은 **유클리디안**입니다.

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 향상됩니다. 또는 **모든 데이터 사용**을 선택하여 모든 데이터 점을 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 사용으로 설정하려면 선택하십시오. 구간화는 실제로 도표화하기 전에 그래프를 세분화된 눈금으로 나누고 각각의 눈금 셀에 표시되는 연결의 수를 계수합니다. 최종 그래프에서는 구간 중심값(구간에 있는 모든 연결 점의 평균)에서 셀당 하나의 연결이 사용됩니다.
- **표본.** 지정된 수의 레코드로 데이터에서 무작위로 표본을 추출하려면 선택하십시오.

다음 표는 SPSS Modeler t-SNE 노드 대화 상자의 고급 탭의 설정과 Python t-SNE 라이브러리 매개변수 간의 관계를 표시합니다.

표 *36. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	Python t-SNE 매개변수
최빈값	mode_type	
시각화 유형	n_components	n_components
방법	method	method
임베드 초기화	init	init
목표	target_field	target_field
당혹도	perplexity	perplexity
조기 과장	early_exaggeration	early_exaggeration
학습 비율	learning_rate	learning_rate
최대 반복수	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 당혹도를 사용하여 t-SNE 수행	isGridSearch	

t-SNE 노드 출력 옵션

출력 탭에서 t-SNE 노드 출력에 대한 옵션을 지정하십시오.

출력결과 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. 비교를 위해 다른 필드를 사용하여 도표를 작성하거나 분류 또는 회귀 모델에서 예측자로서 출력을 사용하려는 경우 t-SNE 노드에는 이 출력 파일에 대한 액세스가 필요합니다.

t-SNE 모델은 고정 파일 소스 노드를 사용하여 가장 쉽게 액세스하는 x, y(및 z) 좌표 필드의 결과 파일을 작성합니다. 자세한 정보는 @@@@@@의 내용을 참조하십시오.

다음 표는 SPSS Modeler t-SNE 노드 대화 상자의 출력 탭의 설정과 Python t-SNE 라이브러리 매개변수 간의 관계를 표시합니다.

표 *37. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	Python t-SNE 매개변수
출력결과 이름	output_Rename	output_Rename
출력 모드	output_to	output_to
파일 이름	full_filename	full_filename
파일 유형	output_file_type	output_file_type
목표	target_field	target_field

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 유형을 선택한 경우, 여기서 결과는 2차원의 포인트 값입니다. **3D**를 선택한 경우, 결과는 3차원의 포인트 값입니다.

랜덤 포리스트 노드

랜덤 포리스트©는 트리 모델을 기본 모델로 사용하는 배깅 알고리즘의 고급 구현을 사용합니다. 랜덤 포리스트에서 앙상블 내의 각 트리는 훈련 세트의 대체(예: 붓스트랩 표본)로 그려진 표본에서 작성됩니다. 트리 생성 동안 노드를 분할하는 경우, 선택된 분할이 더 이상 모든 변수 간의 최상의 분할이 아닙니다. 대신, 선택된 분할이 변수의 임의의 세브세트 간의 최상의 분할이 됩니다. 이 임의성으로 인해 일반적으로 포리스트의 편향이 (단일 비임의 트리의 편향에 비해) 약간 상승하나 평균화로 인해 일반적으로 편향의 증가에 대한 보상 이상으로 분산도 감소하므로 전체적으로 더 나은 모델을 생성합니다.[1]

SPSS Modeler의 랜덤 포리스트 노드는 Python으로 구현됩니다. 노드 팔레트의 Python 탭은 이 노드와 다른 Python 노드로 구성됩니다.

랜덤 포리스트 알고리즘에 대한 자세한 정보는 https://scikit-learn.org/stable/modules/ensemble.html#forest의 내용을 참조하십시오.

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

랜덤 포리스트 노드 필드

필드 탭은 분석에서 사용되는 필드를 지정합니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표 및 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 목표 및 예측변수를 수동으로 할당하려면 이 옵션을 선택하십시오.

필드 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 목표 및 예측변수 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 사용할 필드를 선택하십시오.

예측변수 예측의 입력으로 하나 이상의 필드를 선택하십시오.

랜덤 포리스트 노드 작성 옵션

작성 옵션 탭에서는 **기본 옵션** 및 **고급 옵션**을 포함한 랜덤 포리스트 노드에 대한 작성 옵션을 지정할 수 있습니다. 이러한 옵션에 대한 자세한 정보는 https://scikit-learn.org/stable/modules/ensemble.html#forest의 내용을 참조하십시오.

기본

작성할 트리 수. 포리스트 내의 트리 수를 선택하십시오.

최대 깊이 지정. 선택하지 않으면 모든 리프가 순수하게 될 때까지 또는 모든 리프가 min_samples_split 개 미만의 표본을 포함할 때까지 노드가 펼쳐집니다.

최대 깊이 트리의 최대 깊이입니다.

최소 리프 노드 크기. 리프 노드에 필요한 최소 표본 수입니다.

분할에 사용할 변수 수. 최상의 분할을 검색할 때 고려할 변수의 수입니다.

- auto인 경우, 분류자에 대해서는 max_features=sqrt(n_features), 회귀분석에 대해서는 max_features=sqrt(n_features)입니다.
- sqrt인 경우, max_features=sqrt(n_features)입니다.
- log2인 경우, max_features=log2(n_features)입니다.

고급

트리 작성 시 붓스트랩 표본 사용. 선택하면 트리를 작성할 때 붓스트랩 표본이 사용됩니다.

일반화 정확도를 추정하기 위해 준비된 표본 사용. 선택하면 일반화 정확도를 추정하기 위해 준비된 표본이 사용됩니다.

극단적으로 임의화된 트리 사용. 선택하면 일반 랜덤 포리스트 대신 극단적으로 임의화된 트리가 사용됩니다. 극단적으로 임의화된 트리에서 임의성은 분할이 계산되는 방법보다 한 단계 더 나아갑니다. 랜덤 포리스트에서와 같이 후보 변수의 임의의 세브세트가 사용되나 최상의 판별 임계값을 찾는 대신 각 후보 변수에 대해 임계값이 임의로 추출되고 이러한 임의적으로 생성된 임계값 중 최상이 분할 규칙으로 선택됩니다. 일반적으로 이로 인해 모델의 분산이 약간 감소하고 편향이 약간 증가합니다. [1]

결과 복제. 선택하면 모델 작성 프로세스가 복제되어 동일한 스코어링 결과를 얻을 수 있습니다.

난수 시드. 난수 생성기에 사용될 시드를 생성하려면 **생성**을 클릭하십시오.

하이퍼-모수 최적화(Rbfopt 기준). 모델이 표본에 대해 기대빈도 또는 하한 오차율을 달성할 수 있도록 모수의 최적 조합을 자동으로 검색하는 Rbfopt 기준 하이퍼-모수 최적화를 사용하려면 이 옵션을 선택하십시오. Rbfopt에 대한 세부사항은 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html의 내용을 참조하십시오.

목표 도달하고자 하는 목표 함수 값(표본에 대한 모델의 오차율)이며 예를 들어, 알 수 없는 최적 값이 있습니다. `0.01` 등의 허용 가능한 값을 설정하십시오.

최대 반복. 모델을 시도하는 최대 반복 수입니다. 기본값은 `1000`입니다.

최대 평가. 모델을 시도하기 위한 정확한 모드에서 함수 평가의 최대 수입니다. 기본값은 `300`입니다.

다음 표는 SPSS Modeler 랜덤 포리스트 노드 대화 상자의 설정과 Python 랜덤 포리스트 라이브러리 매개변수 간의 관계를 보여줍니다.

표 *38. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	랜덤 포리스트 매개변수
목표	target	
예측변수	inputs	
작성할 트리의 수	n_estimators	n_estimators
최대 깊이 지정	specify_max_depth	specify_max_depth
최대 깊이	max_depth	max_depth
최소 리프 노드 크기	min_samples_leaf	min_samples_leaf
분할에 사용할 변수의 수	max_features	max_features
트리 작성 시 붓스트랩 표본 사용	bootstrap	bootstrap
일반화 정확도를 추정하기 위해 준비된 표본 사용	oob_score	oob_score
극단적으로 임의화된 트리 사용	extreme	

표 *38. Python* 라이브러리 매개변수에 맵핑되는 노드 특성 (계속)

SPSS Modeler 설정	스크립트 이름(특성 이름)	랜덤 포리스트 매개변수
결과 복제	use_random_seed	
난수 시드	random_seed	random_seed
하이퍼-모수 최적화(Rbfopt 기준)	enable_hpo	
목표(HPO용)	target_objval	
최대 반복(HPO용)	max_iterations	
최대 평가(HPO용)	max_evaluations	

[1]L. Breiman, "Random Forests," Machine Learning, 45(1), 5-32, 2001.

랜덤 포리스트 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.

랜덤 포리스트 모델 너깃

랜덤 포리스트 모델 너깃은 랜덤 포리스트 모델이 캡처한 모든 정보를 포함합니다. 다음 섹션을 사용할 수 있습니다.

모델 정보

이 보기는 입력 필드, one-hot 인코딩 값 및 모델 매개변수를 포함하여 모델에 대한 중요 정보를 제공합니다.

예측변수 중요도

이 보기는 모델을 추정할 때 각 예측변수의 상대적 중요도를 나타내는 차트를 표시합니다. 추가 정보는 47 페이지의 『예측변수 중요도』의 내용을 참조하십시오.

One-Class SVM 노드

One-Class SVM© 노드에서는 자율 학습 알고리즘을 사용합니다. 이 노드는 이상 탐지에 사용할 수 있습니다. 주어진 표본 세트의 소프트 경계를 탐지하여 새 포인트를 해당 세트에 속하거나 속하지 않는 것으로 분류합니다. 이 One-Class SVM 모델링 노드는 Python으로 구현되며, scikit-learn© Python 라이브러리가 필요합니다. scikit-learn 라이브러리에 대한 상세 정보는 http://contrib.scikit-learn.org/imbalanced-learn/about.html[1]을 참조하십시오.

노드 팔레트의 Python 탭은 One-Class SVM 노드와 다른 Python 노드로 구성됩니다.

참고: One-Class SVM은 비감독 이상치와 이상 탐지에 사용됩니다. 대부분의 경우, 알고리즘이 주어진 표본에 대해 올바른 경계를 설정할 수 있도록 알려진 "정상" 데이터 세트를 사용하여 모델을 작성하는 것이 좋습니다. 모델에 대한 매개변수(예: nu, gamma, kernel)는 결과에 상당히 영향을 미칩니다. 그러므로 상황에 맞는 최적의 설정을 찾을 때까지 이러한 옵션을 실험해야 합니다.

[1]Smola, Schölkopf. "A Tutorial on Support Vector Regression". *Statistics and Computing Archive*, vol. 14, no. 3, August 2004, pp. 199-222(http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288)

One-Class SVM 노드 필드

필드 탭은 분석에서 사용되는 필드를 지정합니다.

사전 정의된 역할 사용 입력 역할이 정의되어 있는 모든 필드를 선택하려면 이 옵션을 선택하십시오.

사용자 정의 필드 할당 사용 필드를 수동으로 선택하려면 이 옵션을 선택하고 입력 필드 및 분할 필드를 선택하십시오.

입력 분석에 사용할 입력 필드를 선택하십시오. 유형 없음 또는 알 수 없음을 제외한 모든 저장 유형과 측정 유형이 지원됩니다. 필드의 저장 유형이 문자열인 경우, 이 필드의 값은 one-hot 인코딩 알고리즘을 통해 one-vs-all 방식으로 바이너리화됩니다.

분할 분할 필드로 사용할 필드를 선택하십시오. 플래그, 명목, 순서, 및 이산형 측정 유형이 모두 지원됩니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

One-Class SVM 노드 고급

One-Class SVM 노드의 고급 탭에서는 **단순** 모드 또는 **고급** 모드를 선택할 수 있습니다. **단순**을 선택할 경우 모든 매개변수가 아래에 표시된 기본값으로 설정됩니다. **고급**을 선택할 경우 이러한 매개변수에 대해 사용자 정의 값을 지정할 수 있습니다. 이러한 옵션에 대한 세부사항은 http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM을 참조하십시오.

중지 기준 중지 기준에 대한 허용 오차를 지정하십시오. 기본값은 **1.0E-3**(0.001)입니다.

회귀분석 정밀도(nu) 훈련 오차 및 지원 벡터의 분수 부분에 대한 한도입니다. 기본값은 **0.1**입니다.

커널 유형 알고리즘에 사용할 커널 유형입니다. 옵션은 **RBF**, **다항**, **시그모이드**, **선형** 또는 **사전 계산됨**입니다. 기본값은 **RBF**입니다.

감마 지정 감마를 지정하려면 이 옵션을 선택하십시오. 그렇지 않으면 자동 감마가 적용됩니다.

감마 감마 설정은 RBF, 다항 및 시그모이드 커널 유형에만 사용할 수 있습니다.

Coef0. Coef0은 다항 및 시그모이드 커널 유형에만 사용할 수 있습니다.

차수. 차수는 다항 커널 유형에만 사용할 수 있습니다.

축소 휴리스틱 사용 축소 휴리스틱을 사용하려면 이 옵션을 선택하십시오. 이 옵션은 기본적으로 선택 취소되어 있습니다.

난수 시드 설정 확률 추정 데이터를 셔플링할 때 사용할 난수 시드를 설정하려면 이 옵션을 선택하십시오. 이 옵션은 기본적으로 선택 취소되어 있습니다.

커널 캐시 크기 지정(MB). 커널 캐시의 크기를 지정하려면 이 옵션을 선택하십시오. 이 옵션은 기본적으로 선택 취소되어 있습니다. 선택할 경우 기본값은 **200**MB입니다.

하이퍼-모수 최적화(Rbfopt 기준). 모델이 표본에 대해 기대빈도 또는 하한 오차율을 달성할 수 있도록 모수의 최적 조합을 자동으로 검색하는 Rbfopt 기준 하이퍼-모수 최적화를 사용하려면 이 옵션을 선택하십시오. Rbfopt에 대한 세부사항은 http://rbfopt.readthedocs.io/en/latest/rbfopt_settings.html의 내용을 참조하십시오.

목표 도달하고자 하는 목표 함수 값(표본에 대한 모델의 오차율)이며 예를 들어, 알 수 없는 최적 값이 있습니다. **0.01** 등의 허용 가능한 값을 설정하십시오.

최대 반복. 모델을 시도하는 최대 반복 수입니다. 기본값은 **1000**입니다.

최대 평가. 모델을 시도하는 함수 평가의 최대 수입니다. 여기서, 초점은 속도에 대한 정확도입니다. 기본값은 **300**입니다.

One-Class SVM 노드에는 scikit-learn© Python 라이브러리가 필요합니다. 다음 표는 SPSS Modeler SMOTE 노드 대화 상자의 설정과 Python 알고리즘 간의 관계를 보여줍니다.

표 *39. Python* 라이브러리 매개변수에 맵핑되는 노드 특성

매개변수 이름	스크립트 이름(특성 이름)	Python API 매개변수 이름
중지 기준	stopping_criteria	tol
회귀분석 정밀도	precision	nu
커널 유형	kernel	kernel
감마	gamma	gamma
Coef0	coef0	coef0
차수	degree	degree
축소 휴리스틱 사용	shrinking	shrinking
커널 캐시 크기 지정(숫자 입력 상자)	cache_size	cache_size
난수 시드	random_seed	random_state

One-Class SVM 노드 옵션

One-Class SVM 노드의 옵션 탭에서는 다음 옵션을 설정할 수 있습니다.

평행 좌표 그래픽의 유형. SPSS Modeler는 작성된 모델을 나타내기 위해 평행 좌표 그래픽을 그립니다. 경우에 따라 일부 데이터 열/기능의 값이 다른 것보다 훨씬 크게 표시되어 그래프의 일부 다른 부분을 보기 힘들 수 있습니다. 이 경우, 모든 수직축에 독립형 축 척도를 제공하려면 **독립 수직축** 옵션을 선택하고, 모든 수직축이 동일한 축 척도를 공유하도록 하려면 **일반 수직축**을 선택하십시오.

그래픽의 최대 행 수. 그래프 출력에 표시할 최대 데이터 행 수를 지정하십시오. 기본값은 100입니다. 성능상의 이유로 최대 20개의 필드가 표시됩니다.

그래픽에 모든 입력 필드 그리기. 그래프 출력에 모든 입력 필드를 표시하려면 이 옵션을 선택하십시오. 기본적으로 각 데이터 필드는 수직축으로 그려집니다. 성능상의 이유로 최대 30개의 필드가 표시됩니다.

그래픽에 그릴 사용자 정의 필드. 그래프 출력에 모든 입력 필드를 표시하는 대신, 이 옵션을 선택한 후 표시할 필드의 서브세트를 선택할 수 있습니다. 이로 인해 성능이 향상될 수 있습니다. 성능상의 이유로 최대 20개의 필드가 표시됩니다.

제 18 장 Spark 노드

SPSS Modeler는 Spark 원시 알고리즘을 사용하는 노드를 제공합니다. 노드 팔레트의 **Spark** 탭은 Spark 알고리즘을 실행하는 데 사용할 수 있는 다음 노드를 포함합니다. 이러한 노드는 Windows 64 및 Mac에서 지원됩니다.

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.

등위 회귀분석은 회귀분석 알고리즘 계열에 속합니다. SPSS Modeler에서 등위-AS 노드는 Spark로 구현됩니다. 등위 회귀분석 알고리즘에 대한 자세한 정보는 https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html의 내용을 참조하십시오.

K-평균은 일반적으로 가장 많이 사용되는 군집 알고리즘 중 하나입니다. 여기에서는 데이터 포인트를 사전 정의된 군집 수로 모읍니다. SPSS Modeler에서 K-평균-AS는 Spark로 구현됩니다. K-평균 알고리즘에 대한 자세한 정보는 https://spark.apache.org/docs/2.2.0/ml-clustering.html의 내용을 참조하십시오. K-평균-AS 노드는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

등위-AS 노드

등위 회귀분석은 회귀분석 알고리즘 계열에 속합니다. SPSS Modeler에서 등위-AS 노드는 Spark로 구현됩니다.

등위 회귀분석 알고리즘에 대한 자세한 정보는 https://spark.apache.org/docs/2.2.0/mllib-isotonic-regression.html의 내용을 참조하십시오. [1]

[1] "Regression - RDD-based API." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

등위-AS 노드 필드

필드 탭은 분석에서 사용되는 필드를 지정합니다.

필드 데이터 소스에 있는 모든 필드를 나열합니다. 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 목표, 입력 및 가중 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 목표로 사용할 필드를 선택하십시오.

입력. 하나 이상의 입력 필드를 선택합니다.

가중치. 지수 가중치를 위한 가중 필드를 선택하십시오. 설정되지 않으면, 기본 가중값인 **1**이 사용됩니다.

등위-AS 노드 작성 옵션

작성 옵션 탭에서는 기능 지수 및 등위 유형을 포함하여 등위-AS 노드의 작성 옵션을 지정할 수 있습니다. 자세한 정보는 http://spark.apache.org/docs/latest/api/java/org/apache/spark/ml/regression/IsotonicRegression.html의 내용을 참조하십시오.[1]

입력 필드 지수. 입력 필드의 지수를 지정하십시오. 기본값은 **0**입니다.

등위 유형. 이 설정은 출력 순차규칙이 등위/증가 또는 순서역전/감소여야 하는지 여부를 판별합니다. 기본값은 **등위**입니다.

[1] "Class IsotonicRegression." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

등위-AS 모델 너깃

등위-AS 모델 너깃에는 등위 회귀 모형에서 캡처한 모든 정보가 포함됩니다. 다음 섹션을 사용할 수 있습니다.

모델 요약

이 보기는 입력 필드, 목표 필드 및 모델 작성 옵션을 포함하여 모델에 대한 중요 정보를 제공합니다.

모델 차트

이 보기는 산점도 다이어그램을 표시합니다.

XGBoost-AS 노드

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.

부스팅 알고리즘에 대한 자세한 정보는 http://xgboost.readthedocs.io/en/latest/tutorials/index.html에서 제공하는 XGBoost 자습서를 참조하십시오. [1]

XGBoost 교차 검증 기능은 SPSS Modeler에서 지원되지 않습니다. 이 기능에는 SPSS Modeler 파티션 노드를 사용할 수 있습니다. 또한 SPSS Modeler의 XGBoost는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

[1] "XGBoost Tutorials." *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

XGBoost-AS 노드 필드

필드 탭은 분석에서 사용되는 필드를 지정합니다.

사전 정의된 역할 사용 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표 및 예측변수 등)을 사용합니다.

사용자 정의 필드 할당 사용 목표 및 예측변수를 수동으로 할당하려면 이 옵션을 선택하십시오.

필드 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 목표 및 예측변수 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 사용할 필드를 선택하십시오.

예측변수 예측의 입력으로 하나 이상의 필드를 선택하십시오.

XGBoost-AS 노드 작성 옵션

작성 옵션 탭에서는 모델 작성 및 불균형 데이터 세트 처리를 위한 **일반 옵션**, 목적 및 평가 메트릭에 대한 **학습 작업 옵션** 및 특정 부스터에 대한 **부스터 매개변수**를 포함한 XGBoost-AS 노드에 대한 작성 옵션을 지정할 수 있습니다. 이러한 옵션에 대한 자세한 정보는 다음 온라인 자원을 참조하십시오.

- XGBoost 홈 페이지[1]
- XGBoost 매개변수 참조[2]
- XGBoost Spark API[3]

일반사항

작업자 수. XGBoost 모델을 교육하는 데 사용되는 작업자 수입니다.

스레드 수. 작업자당 사용되는 스레드 수입니다.

외부 메모리 사용. 외부 메모리를 캐시로 사용하는지 여부입니다.

부스터 유형. 사용할 부스터(**gbtree**, **gblinear** 또는 **dart**)입니다.

부스터 라운드 수. 부스팅 라운드 수입니다.

척도 양수 가중치 이 설정은 양수 및 음수 가중치의 균형을 제어하며 불균형 클래스에 유용합니다.

난수 시드 난수 생성기에 사용될 시드를 생성하려면 **생성**을 클릭하십시오.

학습 작업

목적 reg:linear, **reg:logistic**, **reg:gamma**, **reg:tweedie**, **count:poisson**, **rank:pairwise**, **binary:logistic**, **multi** 중 하나를 학습 작업 목적 유형으로 선택하십시오.

평가 메트릭. 검증 데이터에 대한 평가 메트릭입니다. 기본 메트릭은 목적에 따라 지정됩니다(회귀분석의 경우 **rmse**, 분류의 경우 **오차**, 순위화의 경우 **평균 정밀도**). 사용 가능한 옵션은 **rmse**, **mae**, **logloss**, **error**, **merror**, **mlogloss**, **uac**, **ndcg**, **map** 또는 **gamma-deviance**이며 기본값은 **rmse**입니다.

부스터 모수

람다 가중치에 대한 L2 정규화 항입니다. 이 값을 늘리면 더 보수적인 모델이 생성됩니다.

알파 가중치에 대한 L1 정규화 항입니다. 이 값을 늘리면 더 보수적인 모델이 생성됩니다.

람다 편향 편향에 대한 L2 정규화 항입니다. (편향에 대한 L1 정규화 항은 중요하지 않으므로 없습니다.)

트리 방법 사용할 XGBoost Tree 생성 알고리즘을 선택하십시오.

최대 깊이 최대 트리 깊이를 지정하십시오. 이 값을 늘리면 모델이 더 복잡해져서 과적합이 발생할 수 있습니다.

최소 하위 가중치 하위에 필요한 인스턴스 가중치(hessian)의 최소 합계를 지정하십시오. 트리 파티션 단계로 인해 인스턴스 가중치의 합계가 이 **최소 하위 가중치**보다 적은 리프 노드가 생성된 경우 작성 프로세스가 추가 파티셔닝을 중단합니다. 선형 회귀 모드에서 이 가중치는 단순하게 각 노드에 필요한 최소 인스턴스 수에 해당합니다. 가중치가 클수록 더 보수적인 알고리즘이 생성됩니다.

최대 델타 단계 각 트리의 가중치 추정에 허용할 최대 델타 단계를 지정하십시오. **0**으로 설정할 경우 제한조건이 없습니다. 양수 값으로 설정할 경우 업데이트 단계를 더 보수적으로 설정할 수 있습니다. 이 매개변수는 보통 필요하지 않지만, 클래스의 균형이 극도로 맞지 않는 경우 로지스틱 회귀분석에 유용할 수 있습니다.

하위 표본 하위 표본은 훈련 인스턴스의 비율입니다. 예를 들어 이 값을 **0.5**로 설정할 경우, XGBoost는 트리 성장을 위해 데이터 인스턴스의 절반을 무작위로 수집하며 이로 인해 과적합이 방지됩니다.

에타 과적합 방지를 위해 업데이트 단계 중에 사용되는 단계 크기 축소입니다. 부스팅 단계 후마다 새로운 기능의 가중치를 직접 가져올 수 있습니다. 에타는 또한 부스팅 프로세스를 더 보수적으로 만들기 위해 기능 가중치도 축소합니다.

감마 트리의 리프 노드에 추가 파티션을 만드는 데 필요한 최소 손실 감소입니다. 감마 설정이 클수록 더 보수적인 알고리즘이 생성됩니다.

트리별 Colsample 각 트리를 생성할 때 열의 하위 표본 비율입니다.

수준별 Colsample 각 수준에서 각 분할에 대한 열의 하위 표본 비율입니다.

정규화 알고리즘. 정규화 알고리즘은 일반 옵션 아래에서 다트 부스터 유형이 선택된 경우에 사용됩니다. 사용 가능한 옵션은 **트리** 또는 **포리스트**이며 기본값은 **트리**입니다.

표본추출 알고리즘. 표본추출 알고리즘은 일반 옵션 아래에서 다트 부스터 유형이 선택된 경우에 사용됩니다. **균일** 알고리즘은 놓인 트리를 균일하게 선택합니다. **가중된** 알고리즘은 가중치 비율에 따라 놓인 트리를 선택합니다. 기본값은 **균일**입니다.

드롭아웃 비율. 드롭아웃 비율은 일반 옵션 아래에서 다트 부스터 유형이 선택된 경우에 사용됩니다.

건너뛰기 드롭아웃 확률. 건너뛰기 드롭아웃 확률은 일반 옵션 아래에서 다트 부스터 유형이 선택된 경우에 사용됩니다. 드롭아웃을 건너뛰면 새 트리가 **gbtree**와 동일한 방법으로 추가됩니다.

다음 표는 SPSS Modeler XGBoost-AS 노드 대화 상자의 설정과 XGBoost Spark 매개변수 간의 관계를 보여줍니다.

표 *40. Spark* 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	XGBoost Spark 매개변수
목표	target_fields	
예측변수	input_fields	
람다	lambda	lambda
작업자 수	nWorkers	nWorkers
스레드 수	numThreadPerTask	numThreadPerTask
외부 메모리 사용	useExternalMemory	useExternalMemory
부스터 유형	boosterType	boosterType
부스팅 라운드 수	numBoostRound	round
척도 양수 가중치	scalePosWeight	scalePosWeight
목적	objectiveType	objective
평가 메트릭	evalMetric	evalMetric
람다	lambda	lambda
알파	alpha	alpha
람다 편향	lambdaBias	lambdaBias
트리 방법	treeMethod	treeMethod
최대 깊이	maxDepth	maxDepth
최소 하위 가중치	minChildWeight	minChildWeight
최대 델타 단계	maxDeltaStep	maxDeltaStep
하위 표본	sampleSize	sampleSize
에타	eta	eta
감마	gamma	gamma
트리별 Colsample	colsSampleRation	colSampleByTree
수준별 Colsample	colsSampleLevel	colsSampleLevel
정규화 알고리즘	normalizeType	normalizeType
표본추출 알고리즘	sampleType	sampleType
드롭아웃 비율	rateDrop	rateDrop

표 *40. Spark* 매개변수에 맵핑되는 노드 특성 *(계속)*

SPSS Modeler 설정	스크립트 이름(특성 이름)	XGBoost Spark 매개변수
건너뛰기 드롭아웃 확률	skipDrop	skipDrop

[1] "Scalable and Flexible Gradient Boosting." Web. © 2015-2016 DMLC.

[2] "XGBoost Parameters" *Scalable and Flexible Gradient Boosting*. Web. © 2015-2016 DMLC.

[3] "ml.dmlc.xgboost4j.scala.spark 모수." *DMLC for Scalable and Reliable Machine Learning*. Web. 3 Oct 2017.

XGBoost-AS 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.

K-평균-AS 노드

K-평균은 일반적으로 가장 많이 사용되는 군집 알고리즘 중 하나입니다. 여기에서는 데이터 포인트를 사전 정의된 갯수의 군집으로 모읍니다.[1] SPSS Modeler에서 K-평균-AS 노드는 Spark로 구현됩니다.

K-평균 알고리즘에 대한 자세한 정보는 https://spark.apache.org/docs/2.2.0/ml-clustering.html의 내용을 참조하십시오.

K-평균-AS 노드는 범주형 변수에 대해 one-hot 인코딩을 자동으로 수행합니다.

[1] "Clustering." *Apache Spark*. MLlib: Main Guide. Web. 3 Oct 2017.

K-평균-AS 노드 필드

필드 탭은 분석에서 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 직접 입력 필드를 지정하려는 경우 이 옵션을 선택한 후 입력 필드를 선택하십시오. 이 옵션을 사용하는 것은 **입력**에서 필드 역할을 유형 노드로 설정하는 것과 유사합니다.

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.[1]

일반

모델 이름. 특정 군집에 대한 스코어링 이후 생성되는 필드 이름입니다. **자동**(기본값)을 선택하거나 **사용자 정의**를 선택하고 이름을 입력하십시오.

군집 수. 생성할 군집 수를 지정합니다. 기본값은 **5**이고 최소값은 **2**입니다.

초기화

초기화 모드. 군집 중심 초기화를 위한 방법을 지정합니다. **K-평균||**가 기본값입니다. 이러한 두 가지 방법에 대한 세부사항은 확장 가능한 K-평균++를 참조하십시오.[2]

초기화 단계. K-평균|| 초기화 모드가 선택되면, 초기화 단계 수를 지정하십시오. **2**가 기본값입니다.

고급

고급 설정. 다음과 같이 고급 옵션을 설정하려는 경우 이 옵션을 선택하십시오.

최대 반복. 군집 중심을 검색할 때 수행할 최대반복수를 지정하십시오. **20**이 기본값입니다.

허용치. 반복 알고리즘의 수렴허용치를 지정하십시오. **1.0E-4**가 기본값입니다.

난수 시작값 설정. 난수 생성기에 사용될 시드를 생성하려면 이 옵션을 선택한 후 **생성**을 클릭하십시오.

표시

그래프 표시. 출력에 그래프를 포함시키려는 경우 이 옵션을 선택하십시오.

다음 테이블은 SPSS Modeler K-평균-AS 노드의 설정과 K-평균 Spark 매개변수 사이의 관계를 표시합니다.

표 *41. Spark* 매개변수에 맵핑되는 노드 특성

SPSS Modeler 설정	스크립트 이름(특성 이름)	K-평균 SparkML 매개변수
입력 필드	features	
군집 수	clustersNum	k
초기화 모드	initMode	initMode
초기화 단계	initSteps	initSteps
최대 반복	maxIter	maxIter
허용치	toleration	tol
난수 시드	randomSeed	seed

[1] "Class KMeans." *Apache Spark*. JavaDoc. Web. 3 Oct 2017.

[2] Bahmani, Moseley, et al. "Scalable K-Means++." Feb 28, 2012. http://theory.stanford.edu/%7Esergei/papers/vldb12-kmpar.pdf.

주의사항

이 정보는 미국에서 제공되는 제품 및 서비스용으로 작성된 것입니다. 이 자료는 IBM에서 다른 언어로 사용 가능합니다. 그러나 자료에 접근하기 위해서는 해당 언어로 된 제품 또는 제품 버전의 사본이 필요할 수 있습니다.

IBM은 다른 국가에서 이 책에 기술된 제품, 서비스 또는 기능을 제공하지 않을 수도 있습니다. 현재 사용할 수 있는 제품 및 서비스에 대한 정보는 한국 IBM 담당자에게 문의하십시오. 이 책에서 IBM 제품, 프로그램 또는 서비스를 언급했다고 해서 해당 IBM 제품, 프로그램 또는 서비스만을 사용할 수 있다는 것을 의미하지는 않습니다. IBM의 지적 재산권을 침해하지 않는 한, 기능상으로 동등한 제품, 프로그램 또는 서비스를 대신 사용할 수도 있습니다. 그러나 비IBM 제품, 프로그램 또는 서비스의 운영에 대한 평가 및 검증은 사용자의 책임입니다.

IBM은 이 책에서 다루고 있는 특정 내용에 대해 특허를 보유하고 있거나 현재 특허 출원 중일 수 있습니다. 이 책을 제공한다고 해서 특허에 대한 라이센스까지 부여하는 것은 아닙니다. 라이센스에 대한 의문사항은 다음으로 문의하십시오.

07326
서울특별시 영등포구
국제금융로 *10, 3IFC*
한국 아이.비.엠 주식회사
대표전화서비스*: 02-3781-7114*

2바이트(DBCS) 정보에 관한 라이센스 문의는 한국 IBM에 문의하거나 다음 주소로 서면 문의하시기 바랍니다.

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM은 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여(단, 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 이 책을 "현상태대로" 제공합니다. 일부 국가에서는 특정 거래에서 명시적 또는 묵시적 보증의 면책사항을 허용하지 않으므로, 이 사항이 적용되지 않을 수도 있습니다.

이 정보에는 기술적으로 부정확한 내용이나 인쇄상의 오류가 있을 수 있습니다. 이 정보는 주기적으로 변경되며, 변경된 사항은 최신판에 통합됩니다. IBM은 이 책에서 설명한 제품 및/또는 프로그램을 사전 통지 없이 언제든지 개선 및/또는 변경할 수 있습니다.

이 정보에서 언급되는 비IBM 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 해당 웹 사이트의 자료는 본 IBM 제품 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

IBM은 귀하의 권리를 침해하지 않는 범위 내에서 적절하다고 생각하는 방식으로 귀하가 제공한 정보를 사용하거나 배포할 수 있습니다.

(i) 독립적으로 작성된 프로그램과 기타 프로그램(본 프로그램 포함) 간의 정보 교환 및 (ii) 교환된 정보의 상호 이용을 목적으로 본 프로그램에 관한 정보를 얻고자 하는 라이센스 사용자는 다음 주소로 문의하십시오.

07326
서울특별시 영등포구
국제금융로 *10, 3IFC*
한국 아이.비.엠 주식회사
대표전화서비스*: 02-3781-7114*

이러한 정보는 해당 조건(예를 들면, 사용료 지불 등)하에서 사용될 수 있습니다.

이 정보에 기술된 라이센스가 부여된 프로그램 및 프로그램에 대해 사용 가능한 모든 라이센스가 부여된 자료는 IBM이 IBM 기본 계약, IBM 프로그램 라이센스 계약(IPLA) 또는 이와 동등한 계약에 따라 제공한 것입니다.

인용된 성능 데이터와 고객 예제는 예시 용도로만 제공됩니다. 실제 성능 결과는 특정 구성과 운영 조건에 따라 다를 수 있습니다.

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.

IBM의 향후 방향 또는 의도에 관한 언급은 별도의 통지없이 변경될 수 있습니다.

이 정보에는 일상의 비즈니스 운영에서 사용되는 자료 및 보고서에 대한 예제가 들어 있습니다. 이들 예제에는 개념을 가능한 완벽하게 설명하기 위하여 개인, 회사, 상표 및 제품의 이름이 사용될 수 있습니다. 이들 이름은 모두 가공의 것이며 실제 인물 또는 기업의 이름과 유사하더라도 이는 전적으로 우연입니다.

상표

IBM, IBM 로고 및 ibm.com은 전세계 여러 국가에 등록된 International Business Machines Corp.의 상표 또는 등록상표입니다. 기타 제품 및 서비스 이름은 IBM 또는 타사의 상표입니다. IBM 상표의 최신 목록은 웹 사이트(www.ibm.com/legal/copytrade.shtml)에서 "Copyright and trademark information"을 참조하십시오.

Adobe, Adobe 로고, PostScript 및 PostScript 로고는 미국 및/또는 기타 국가에서 사용되는 Adobe Systems Incorporated의 등록상표 또는 상표입니다.

Intel, Intel 로고, Intel Inside, Intel Inside 로고, Intel Centrino, Intel Centrino 로고, Celeron, Intel Xeon, Intel SpeedStep, Itanium 및 Pentium은 미국 또는 기타 국가에서 사용되는 Intel Corporation 또는 그 계열사의 상표 또는 등록상표입니다.

Linux는 미국 또는 기타 국가에서 사용되는 Linus Torvalds의 등록상표입니다.

Microsoft, Windows, Windows NT 및 Windows 로고는 미국 또는 기타 국가에서 사용되는 Microsoft Corporation의 상표입니다.

UNIX는 미국 및 기타 국가에서 사용되는 The Open Group의 등록상표입니다.

Java 및 모든 Java 기반 상표와 로고는 Oracle 및/또는 그 계열사의 상표 또는 등록상표입니다.

제품 문서의 이용 약관

다음 이용 약관에 따라 이 책을 사용할 수 있습니다.

적용성

본 이용 약관은 IBM 웹 사이트의 모든 이용 약관에 추가됩니다.

개인적 사용

모든 소유권 사항을 표시하는 경우에 한하여 귀하는 이 책을 개인적, 비상업적 용도로 복제할 수 있습니다. 귀하는 IBM의 명시적 동의 없이 본 발행물 또는 그 일부를 배포 또는 전시하거나 2차적 저작물을 만들 수 없습니다.

상업적 사용

모든 소유권 사항을 표시하는 경우에 한하여 귀하는 이 책을 귀하 기업집단 내에서만 복제, 배포 및 전시할 수 있습니다. 귀하는 귀하의 기업집단 외에서는 IBM의 명시적 동의 없이 이 책의 2차적 저작물을 만들거나 이 책 또는 그 일부를 복제, 배포 또는 전시할 수 없습니다.

권한

본 허가에서 명시적으로 부여된 경우를 제외하고, 이 책이나 이 책에 포함된 정보, 데이터, 소프트웨어 또는 기타 지적 재산권에 대한 어떠한 허가나 라이센스 또는 권한도 명시적 또는 묵시적으로 부여되지 않습니다.

IBM은 이 책의 사용이 IBM의 이익을 해친다고 판단되거나 위에서 언급된 지시사항이 준수되지 않는다고 판단하는 경우 언제든지 부여한 허가를 철회할 수 있습니다.

귀하는 미국 수출법 및 관련 규정을 포함하여 모든 적용 가능한 법률 및 규정을 철저히 준수하는 경우에만 본 정보를 다운로드, 송신 또는 재송신할 수 있습니다.

IBM은 이 책의 내용과 관련하여 아무런 보장을 하지 않습니다. 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여 (단 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 현 상태대로 제공합니다.

용어

가

개별-그룹 공분산*(Separate-Groups Covariance)*. 각 그룹에 대해 개별 공분산 행렬을 표시합니다.

개별-그룹 도표*(Separate-Groups Plots)*. 처음 두 판별 함수 값의 개별 그룹 산점도를 작성합니다. 함수가 하나만 있는 경우에는 산점도 대신 히스토그램이 표시됩니다.

개별-그룹*(Separate-Groups)*. 개별-그룹 공분산 행렬이 분류에 사용됩니다. 분류가 판별 함수에 기초하고 원래 변수에 따라 달라지지는 않으므로 이 옵션이 2차 판별과 항상 같지는 않습니다.

결합-그룹 도표*(Combined-Groups Plots)*. 처음 두 판별 함수 값에 대해 전체 그룹화 산점도를 작성합니다. 함수가 하나만 있는 경우에는 산점도 대신 히스토그램이 표시됩니다.

고유*(Unique)*. 모든 효과를 동시에 평가하고 유형에 관계없이 다른 모든 효과에 대해 각 효과를 조정합니다.

공분산*(Covariance)*. 두 변수 간 연관을 표준화하지 않은 측도로서, N-1로 나눈 교차곱 편차와 같습니다.

그룹 내 공분산*(Within-Groups Covariance)*. 그룹 내 풀링 공분산 행렬을 표시하는데 이는 전체 공분산 행렬과 다를 수 있습니다. 이 행렬은 모든 그룹에 대해 개별 공분산 행렬을 평균하여 구합니다.

그룹 내 상관*(Within-Groups Correlation)*. 상관을 계산하기 전에 모든 그룹에 대한 개별 공분산 행렬의 평균을 구하여 그룹 내 풀링 상관 행렬을 표시합니다.

그룹 내*(Within-Groups)*. 그룹 내 풀링 공분산 행렬이 케이스 분류에 사용됩니다.

바

범위*(Range)*. 숫자변수의 가장 큰 값과 가장 작은 값의 차이로 최대값에서 최소값을 뺀 값을 의미합니다.

베이지안 정보 기준*(BIC)(Bayesian Information Criterion (BIC))*. -2 로그 우도에 기반한 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. BIC도 초과 모수화된 모델(예: 입력이 많은 복잡한 모델)에 "페널티를 부여"하지만 AIC보다 더 엄격하게 부여합니다.

분류 결과*(Classification Results)*. 판별 분석을 기준으로 각 그룹에 정확하게 할당되거나 잘못 할당된 케이스의 수로, "혼동 행렬"이라고도 합니다.

분산*(Variance)*. 평균에 대한 산포 측도로, 평균으로부터의 제곱합 편차를 케이스 수에서 1을 뺀 값으로 나눈 값과 같습니다. 분산은 변수 자체의 제곱 단위로 측정됩니다.

비표준화*(Unstandardized)*. 표준화하지 않은 판별 함수 계수를 표시합니다.

사

생존 도표*(Survival Plot)*. 선형 척도로 누적 생존함수를 표시합니다.

설명되지 않는 분산*(Unexplained Variance)* . 각 단계에서 그룹 간 설명되지 않은 변동 합계를 최소화하는 변수를 입력합니다.

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.

순차제거복원 분류*(Leave-one-out Classification)* . 분석의 각 케이스가 해당 케이스가 아닌 다른 모든 케이스에서 파생된 함수에 따라 분류됩니다. 이 방법을 "U-방법"이라고도 합니다.

아

영역 맵*(Territorial Map)* . 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.

왜도*(Skewness)* . 분포의 비대칭성에 대한 측도입니다. 정규 분포는 대칭이므로 왜도 값이 0입니다. 양의 왜도가 많은 분포는 오른쪽이 깁니다. 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 -2보다 작거나 +2보다 큰 경우 정규성을 거부할 수 있습니다. 왜도가 큰 양의 값인 경우 오른쪽이 길어지고 큰 음의 값인 경우 왼쪽이 길어집니다.

위험 함수 도표*(Hazard Plot)* . 선형 척도에 누적 위험 함수를 표시합니다.

유효함*(Valid)* . 시스템 결측값 또는 사용자 결측값이 지정되어 있지 않은 케이스가 유효 케이스입니다.

일변량분산 분석*(Univariate ANOVAs)* . 각 독립변수에 대해 그룹 평균의 등식을 검정하는 일원 분산 분석을 수행합니다.

자

전체 공분산*(Total Covariance)* . 단일 표본으로 작성한 것처럼 모든 케이스로부터 공분산 행렬을 표시합니다.

정규화된 *BIC(Normalized BIC)* . 정규화된 Bayesian 정보 기준입니다. 모델의 전반적인 적합도에 대한 일반적인 측도로서 모델 복합성을 설명해 줍니다. 평균 제곱 오차를 기반으로 하는 스코어이며, 모델 내 모수의 수와 계열의 길이에 대한 페널티를 포함합니다. 페널티는 모수가 더 많은 모델의 이점을 제거하게 되므로 동일 계열의 경우 다른 모델끼리 통계를 쉽게 비교할 수 있습니다.

정상 *R*-제곱*(Stationary R-squared)* . 모델의 정상 부분과 단순 평균 모델을 비교하는 측도입니다. 추세나 계절 패턴이 있는 경우 보통 R-제곱보다 이 측도를 사용하는 것이 좋습니다. 정상 R-제곱의 범위는 음의 무한대에서 1까지입니다. 음수 값은 고려 중인 모델이 기준선 모델보다 나쁨을 의미하며 양수 값은 고려 중인 모델이 기준선 모델보다 좋음을 의미합니다.

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.

차

첨도*(Kurtosis)* . 이상치가 있는 정도에 대한 측도입니다. 정규 분포의 경우 첨도 통계 값은 0입니다. 양(+) 첨도는 데이터가 정규 분포보다 더 극단적인 이상치를 나타냄을 표시합니다. 음(-) 첨도는 데이터가 정규 분포보다 덜 극단적인 이상치를 나타냄을 표시합니다.

첨도의 표준 오차*(Standard Error of Kurtosis)* . 표준 오차에 대한 첨도의 비율을 정규성 검정에 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).

최대값*(Maximum)* . 숫자변수의 가장 큰 값입니다.

최빈값*(Mode)* . 가장 자주 발생하는 값입니다. 여러 값에서 최대 발생 빈도를 공유하는 경우 각각을 최빈값이라고 합니다.

최소 *F*-비 진입방법 최대화*(Maximizing the Smallest F Ratio Methodof Entry)* . 그룹 간 Mahalanobis 거리로부터 계산한 F-비를 최대화하는 단계별 분석의 변수 선택 방법입니다.

최소값*(Minimum)* . 숫자변수의 가장 작은 값입니다.

카

케이스*(Cases)* . 각 케이스마다 실제 그룹, 예측 그룹, 사후 확률, 판별 스코어 등에 대한 코드가 표시됩니다.

파

평균*(Mean)* . 중심 경향에 대한 측도입니다. 합계를 케이스 수로 나눈 산술 평균 값입니다.

평균*(Means)* . 전체 평균 및 그룹 평균, 독립변수에 대한 표준 편차를 표시합니다.

평균의 표준 오차*(Standard Error of Mean)* . 동일 분포로부터 선택한 표본 간에 발생할 수 있는 평균값의 차이에 대한 측도입니다. 이 값을 사용하여 관측 평균과 가설 값을 간략하게 비교할 수 있습니다. 즉, 표준 오차에 대한 차이 비율이 ?2보다 작거나 +2보다 큰 경우 두 값이 다르다고 판단할 수 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.

표준 편차*(standard deviation)* . 평균 주위의 산포 측도이며 분산의 제곱근과 같습니다. 표준 편차는 원래 변수와 같은 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.

하

합계*(Sum)* . 비결측값을 갖는 전체 케이스 값의 총계입니다.

숫자

1 - 생존함수(One Minus Survival) . 선형 척도에 1 ? 생존함수를 도표화합니다.

A

AICC . -2(제한) 로그 우도에 기반한 혼합 모델을 선택 및 비교하기 위한 측도입니다. 값이 작을수록 모델이 우수함을 나타냅니다. AICC는 작은 표본 크기의 AIC를 "수정합니다". 표본 크기가 증가함에 따라 AICC는 AIC로 수렴됩니다.

B

Box의 M 검정(Box's M test) . 그룹 공분산 행렬의 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.

F

F 값 사용(Use F Value) . F 값이 진입값보다 크면 모델에 변수가 입력되고 F 값이 제거값보다 작으면 제거됩니다. 진입값은 제거값보다 커야 하고 두 값 모두 양수이어야 합니다. 모델에 더 많은 변수를 입력하려면 진입값을 낮추고 모델에서 변수를 더 많이 제거하려면 제거값을 높입니다.

F 확률 사용(Use Probability of F) . F 값의 유의 수준이 진입값보다 작으면 모델에 변수가 입력되고 유의 수준이 제거값보다 크면 제거됩니다. 진입값은 제거값보다 작아야 하며 두 값 모두 양수이어야 합니다. 모델에 변수를 더 많이 입력하려면 진입값을 높이고 모델에서 변수를 더 많이 제거하려면 제거값을 낮춥니다.

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.

M

MAE . 평균 절대 오차입니다. 계열이 모델 예측 수준에서 얼마나 달라지는지에 대한 측도입니다. MAE는 원래의 계열 단위로 보고됩니다.

Mahalanobis 거리(Mahalanobis Distance) . 독립변수의 케이스 값이 전체 케이스 평균과 얼마나 달라지는지에 대한 측도입니다. Mahalanobis 거리가 크면 케이스가 독립변수 하나 이상에 대해 극단값을 갖는 것으로 식별합니다.

MAPE . 평균 절대 백분율 오차로서, 종속 계열이 모델 예측 수준에서 얼마나 달라지는지에 대한 측도입니다. 사용된 단위와 상관이 없으므로 다른 단위의 계열을 비교하는 데 사용할 수 있습니다.

MaxAE . 절대 오차의 최대값으로서, 종속 계열과 같은 단위로 표현되는 최대 예측 오차입니다. MaxAPE와 마찬가지로 예측에 대한 최악의 케이스 시나리오를 예상하는 데 유용합니다. 절대 오차의 최대값과 절대 백분율 오차의 최대값은 다른 계열 지점에서 발생할 수 있습니다. 예를 들어 대형 계열 값의 절대 오차는 소형 계열 값의 절대 오차보다 약간 큽니다. 그런 경우, 절대 오차의 최대값은 더 큰 계열 값에서 발생하며 최대 절대 백분율 오차는 더 작은 계열 값에서 발생합니다.

MaxAPE . 최대 절대 백분율 오차로서, 백분율로 표현되는 최대 예측 오차입니다. 이 측도는 예측에 대한 최악의 시나리오를 예상하는 데 유용합니다.

R

*Rao*의 *V(*판별 분석*)(Rao's V (Discriminant Analysis))* . 그룹 평균 간 차이에 대한 측도입니다. Lawley-Hotelling 트레이스라고도 하며 각 단계에서 Rao의 V의 증가를 최대화하는 변수가 입력됩니다. 이 옵션을 선택한 다음 변수가 가져야 하는 최소값을 입력하여 분석에 사용합니다.

RMSE . RMSE(Root Mean Square Error)는 평균 제곱 오차의 제곱근입니다. 종속 계열이 모델 예측 수준과 얼마나 다른지에 대한 측도로서, 종속 계열과 같은 단위로 표시됩니다.

*R-*제곱*(R-Squared)* . 선형모델의 적합도 측도로서 결정계수라고도 합니다. 이 항목은 회귀 모델로 설명한 종속변수의 변동 비율이 됩니다. 값 범위는 0 - 1입니다. 값을 작을수록 모델이 데이터에 적합하지 않음을 의미합니다.

W

*Wilks*의 람다 최소화*(Minimize Wilks' Lambda)* . 단계별 판별 분석의 변수 선택 방법으로, Wilks의 람다를 낮추는 정도에 따라 방정식에 입력할 변수를 선택합니다. 각 단계에서 전체 Wilks의 람다를 최소화할 변수를 입력합니다.

색인

[가]

가법 이상값 326
  패치 326
가져오기
  PMML 44, 54, 55
가중 최소제곱법 33
가중 필드 33, 35
개입
  식별 325
결측 데이터
  예측변수 계열 329
결측값
  필드 선별 60
  CHAID 트리 94
  SQL에서 제외 126, 132, 137, 254
결합 규칙
  선형 모델에서 193
  신경망에서 157
계열
  변환 329
계열 변환 329
계절 가법 이상값 326
계절 차분 변환 329
계절성 325
  식별 324
계층적 모델
  일반화 선형 혼합 모델 231
고급 매개변수 177
고급 옵션
  베이지안 네트워크 노드 146
  시퀀스 노드 307
  코호넨 모델 264
  Apriori 노드 290
  CARMA 노드 294
  Cox 회귀 모형 257
  K-평균 모델 267
고급 출력
  요인/PCA 노드 217
  Cox 회귀 모형 258
고유값
  PCA/요인 모델 215
공분산 교차표
  일반화 선형 모델 228
과적합 방지
  신경망에서 158
과적합 방지 기준
  선형 모델에서 192
  linear-AS 모델에서 199
관리자
  모델 탭 44
관측값 별 예측값
  Linear-AS 모델 200
  LSVM 모델 384
국소적 추세 이상값 326
군집 노드 277, 420
군집 뷰어
  개요 279
  군집 모델 정보 278
  군집 및 변수 이동 281
  군집 및 변수 전치 281
  군집 보기 280
  군집 비교 보기 282
  군집 예측변수 중요도 보기 282
  군집 정렬 281
  군집 중심 보기 280
  군집 크기 보기 282
  군집 표시 정렬 281
  군집의 비교 282
  군집의 크기 282
  그래프 생성 285
  기본 보기 281
  모델 요약 280
  변수 정렬 281
  변수 표시 정렬 281
  사용 283
  셀 내용 정렬 281
  셀 내용 표시 281
  셀 분포 보기 282
  셀의 분포 282
  예측변수 중요도 282
  요약 보기 280
군집분석
  군집 수 269
  이단계 군집 270, 271, 272, 274, 276
  이상 항목 발견 65
군집화 262, 266, 267, 268, 270, 278
  군집 보기 279
군집화 (계속)
  전체 표시 279
규칙
  규칙 지원 295, 310
  연관 규칙 289, 292
규칙 귀납 108, 109, 118, 121, 127, 289
규칙 노드 작성 133
규칙 세트 105, 137, 140, 141, 299, 301
  의사결정 트리에서 생성 105
규칙 수퍼 노드
  시퀀스 규칙에서 생성 312
규칙 ID 295
그래프 생성
  연관 규칙 299
근사 공분산
  로지스틱 회귀분석 모델 208
근사 상관
  로지스틱 회귀분석 모델 208, 213
기본 범주
  로지스틱 노드 203
기술통계량
  일반화 선형 모델 228

[나]

내보내기
  모델 너깃 44
  PMML 54, 55
  SQL 45
내용 필드
  시퀀스 노드 306
  CARMA 노드 292

[다]

다중 레이어 퍼셉트론(MLP)
  신경망에서 155
다중 수준 모델
  일반화 선형 혼합 모델 231
다항 로지스틱 회귀분석
  일반화 선형 혼합 모델 231
다항 로지스틱 회귀분석 모델 202, 203
단계 개입
  식별 325

단계 선택 필드 선택
판별 노드 221
단계별 옵션
로지스틱 회귀분석 모델 209
Cox 회귀 모형 258
단계별 전진
선형 모델에서 192
linear-AS 모델에서 199
대비 계수 행렬
일반화 선형 모델 228
대안 탭 173
대용
의사결정 트리 95, 112, 122
대체 규칙 분할창 178
대체 모델 181
대화형 트리 93, 95, 96
결과 내보내기 104
그래프 생성 138
대용 95
모델 생성 101, 102
사용자 정의 분할 94
이익 97, 99, 100
ROI 99
데이터 선택 구성 178
데이터 스코어링 52
데이터 축소
PCA/요인 모델 214
동위
최근접 이웃 분석에서 395
등위-AS 노드 415, 416
등위-AS 모델 너깃 416

[라]

람다
필드선택 61
랜덤 트리 모델
고급 설정 129
구간화 129
모델 정보 130
모델링 노드 127, 132
예측변수 중요도 130
오분류 비용 129
작성 옵션 128
출력 130
트리 깊이 128
표본 크기 128
필드 옵션 127
랜덤 포리스트 노드 407, 408, 410
랜덤 포리스트 모델 너깃 410
레이블
값 54
변수 54
레코드 요약
Linear-AS 모델 200
LSVM 모델 384
로그 변환 329
시계열 모델 생성기 363
로그선형분석
일반화 선형 혼합 모델 231
로그-오즈비
로지스틱 회귀분석 모델 211
로드
모델 너깃 44
로지스틱 회귀분석
일반화 선형 혼합 모델 231
로지스틱 회귀분석 모델 189
고급 옵션 207
고급 출력 208, 213
다항 옵션 203
단계별 옵션 209
모델 너깃 210, 211, 212
모델 방정식 211
모델링 노드 202
상호작용 206
수렴 옵션 208
예측변수 중요도 211
이항검정 옵션 203
주효과 206
항 추가 206
리프트 295
연관 규칙 298
의사결정 트리 이익 97
리프트 도표
의사결정 트리 이익 99
링크
모델 41

[마]

마이닝 작업 175
시작 176
작성 176
편집 176
마이닝 작업 실행 175
명목 회귀 202
모델
가져오기 44
교체 43
분할 30, 31, 32
요약 탭 47
모델 교체 43
모델 규칙 추가 178
모델 너깃 40, 56, 126, 132, 133, 137, 138, 140, 141, 230, 254
내보내기 44, 45
데이터 스코어링 52
메뉴 45
분할 모델 51, 52
스트림에서 사용 52
앙상블 모델 49
요약 탭 47
인쇄 45
저장 45
저장 및 로드 44
처리 노드 생성 52
모델 링크 41
및 수퍼 노드 43
복사 및 붙여넣기 42
정의 및 제거 41
모델 링크 복사 42
모델 링크 제거 41
모델 보기
일반화 선형 혼합 모델 240
최근접 이웃 분석에서 393
모델 사용자 정의 181
모델 새로 고침
자체 학습 반응 모델 372
모델 시각화 187
모델 옵션
베이지안 네트워크 노드 145
Cox 회귀 모형 256
SLRM 노드 372
모델 적합
로지스틱 회귀분석 모델 213
모델 정보
랜덤 트리 모델 130
시계열 모델 367
일반화 선형 모델 228
GLE 모델 253
Linear-AS 모델 200
LSVM 모델 384
Tree-AS 모델 125

모델 측도
새로 고침 183
정의 183
모델 팔레트 40, 44
모델 평가 183
모델링 노드 64, 118, 143, 262, 266, 268, 270, 277, 289, 305, 371, 415, 416, 417, 420
모수 추정 333
모수 추정값
로지스틱 회귀분석 모델 213
일반화 선형 모델 228
목표 값 변경 182
문서 3
미리보기
모델 내용 45

[바]

반복계산과정
로지스틱 회귀분석 모델 208
일반화 선형 모델 228
반응 차트
의사결정 트리 이익 97, 99
방사형 기저함수(RBF)
신경망에서 155
배깅 110
선형 모델에서 191
신경망에서 153
배포성 측도 295
베리멕스 회전
PCA/요인 모델 216
베이지안 신경망 모형
고급 옵션 146
모델 너깃 147
모델 너깃 설정 148
모델 너깃 요약 149
모델 옵션 145
모델링 노드 143
변수 중요도
자체 학습 반응 모델 374
변환 연관 규칙 316
부스팅 110, 119, 138
선형 모델에서 191
신경망에서 153
분류 이익
의사결정 트리 97, 99
분류 트리 108, 109, 118, 121, 127
분류표
로지스틱 회귀분석 모델 208
최근접 이웃 분석에서 395
분산 계수
필드 선별 60
분산 분석
일반화 선형 혼합 모델 231
분산 안정화 변환 329
분할
의사결정 트리 94, 95
분할 모델
대 파티셔닝 31
모델링 노드 31
영향을 받는 기능 32
작성 30
분할 모델 너깃 51
뷰어 52
요약 탭 47
불순도 측도
의사결정 트리 115
C&R 트리 노드 115
뷰어 탭
그래프 생성 138
의사결정 트리 모형 136
비계절 순환 325
비모수 추정 333
비선형 추세
식별 324
비용
오분류 40
의사결정 트리 113, 114, 124, 129
빈도 필드 35

[사]

사용 가능 필드 178
사용자 정의 분할
의사결정 트리 94, 95
사전 신뢰도에 대한 절대 신뢰도 차이
apriori 평가 측도 290
사전 확률
의사결정 트리 113
삭제
모델 링크 41
상관행렬
일반화 선형 모델 228
상호작용
로지스틱 회귀분석 모델 206
새 모델 생성 182
생성된 시퀀스 규칙 세트 301
선택 노드
의사결정 트리에서 생성 104
선택 작성
정의 176
선형 모델 190
결과 복제 194
결합 규칙 193
계수 196
관측값 별 예측값 195
너깃 설정 197
모델 선택 192
모델 옵션 194
모델 요약 194
모델 작성 요약 197
목적 191
분산분석표 196
신뢰수준 192
앙상블 193
예측변수 중요도 195
이상값 195
자동 데이터 준비 192, 194
잔차 195
정보 기준 194
평균 추정 197
R 제곱 통계 194
선형 지원 벡터 머신 모델
모델 너깃 384
모델 옵션 383
모델링 노드 382
설정 385
작성 옵션 383
선형 추세
식별 324
선형 커널
지원 벡터 머신 모델 377
선형 회귀 모형 189
가중 최소제곱법 33
모델링 노드 190, 198
설정 옵션
Cox 회귀 모형 259
SLRM 노드 373
성능 개선 209, 289
성능 최적화 289
성향 스코어
데이터 균형 38
의사결정 목록 모델 171

성향 스코어 (계속)
일반화 선형 모델 230
판별 모델 222
세그먼트
규칙 조건 삭제 180
복사 180
삭제 181
삽입 179
우선 순위 지정 181
제외 181
편집 179
세그먼트 규칙 생성 175
세분화되지 않은 규칙 모델 295, 300, 301
세분화되지 않은 모델 56, 62, 63
수렴 옵션
로지스틱 회귀분석 모델 208
일반화 선형 모델 227
CHAID 노드 116
Cox 회귀 모형 258
Tree-AS 노드 123
수렴에 대한 엡실론
CHAID 노드 116
Tree-AS 노드 123
수정 Bonferroni
CHAID 노드 116
Tree-AS 노드 122
수정된 성향 스코어
데이터 균형 38
의사결정 목록 모델 171
일반화 선형 모델 230
판별 모델 222
수정된 R-제곱
선형 모델에서 192
linear-AS 모델에서 199
수준 안정화 변환 329
수준 이동 이상값 326
수퍼노드
및 모델 링크 43
순서화 투잉 불순도 측도 115
스냅샷
작성 174
스냅샷 탭 174
스코어 통계 208, 209
시각화
군집 모델 279
그래프 생성 138, 285, 299
의사결정 트리 136
시간 인과 모델 336, 337, 338, 340, 341, 342, 344, 346, 347
모델링 노드 336
시간 인과 모델 시나리오 349, 350, 351, 353, 354
시간 인과 모델링
모델 너깃 348
모델 너깃 설정 348
시간 필드
시퀀스 노드 306
CARMA 노드 292
시계열 모델
결측값 옵션 359
관측값 옵션 356
데이터 지정 사항 옵션 356
모델 너깃 설정 368
모델 옵션 365
모델 정보 367
모델링 노드 355
변환 363
시간 구간 옵션 357
예측변수 중요도 367
일반 작성 옵션 360
작성 옵션 360
작성 출력 옵션 365
전이 함수 차수 363
지수평활 355, 360
추정 기간 359
출력 367
통합 및 분포 옵션 358
필드 옵션 356
ARIMA 360, 363
ARIMA 모델 355
시작하기 172
시차
ACF 및 PACF 328
시퀀스 모델
고급 옵션 307
규칙 수퍼 노드 생성 312
내용 필드 306
데이터 형식 306
모델 너깃 309, 310, 312
모델 너깃 설정 312
모델 너깃 세부사항 310
모델 너깃 요약 312
모델링 노드 305
시간 필드 306
시퀀스 브라우저 312
시퀀스 모델 (계속)
예측 309
옵션 307
정렬 312
테이블 대 트랜잭션 데이터 307
필드 옵션 306
ID 필드 306
시퀀스 발견 305
시퀀스 브라우저 312
신경망 151
결과 복제 158
결측값 158
결합 규칙 157
과적합 방지 158
관측값 별 예측값 162
너깃 설정 165
네트워크 163
다중 레이어 퍼셉트론(MLP) 155
모델 옵션 159
모델 요약 160
목적 153
방사형 기저함수(RBF) 155
분류 162
앙상블 157
예측변수 중요도 161
은닉층 155
중지 규칙 156
신경망 노드 151
신경망 모델
필드 옵션 33
신뢰구간
로지스틱 회귀분석 모델 208
신뢰도
규칙 세트 137
로지스틱 회귀분석 모델 212
시퀀스 노드 307
시퀀스에 대해 310
연관 규칙 295, 298, 310
의사결정 트리 모형 126, 132, 137
Apriori 노드 289
CARMA 노드 293
GLE 모델 254
신뢰도 비율
apriori 평가 측도 290
신뢰도 스코어 38
신뢰도 차이
apriori 평가 측도 290

[아]

알고리즘 40
앙상블
  선형 모델에서 193
  신경망에서 157
앙상블 뷰어 49
  구성요소 모델 세부사항 51
  구성요소 모델 정확도 50
  모델 요약 49
  예측변수 빈도 50
  예측변수 중요도 50
  자동 데이터 준비 51
애플리케이션 예제 3
연결 함수
  일반화 선형 혼합 모델 232
  GLE 모델 246
연관 규칙 313
연관 규칙 노드 313
연관 규칙 모델 33, 126, 132, 137, 140, 141, 309, 310, 312
  규칙 세트 생성 301
  규칙 스코어링 302
  그래프 생성 299
  모델 너깃 295
  모델 너깃 세부사항 295
  모델 너깃 요약 300
  배포 303
  설정 299
  스코어 전치 303
  시퀀스에 대해 305
  필터 지정 298
  필터링된 모델 생성 301
  apriori 289
  CARMA 292
연관 규칙 모델 옵션 318
연관 규칙 변환 316
연관 규칙 작성 315
연관 규칙 출력 317
연관 규칙의 출력 317
연관성 규칙 모델
  모델 너깃 319
  모델 너깃 설정 320
  모델 너깃 세부사항 320
  필드 옵션 314
영역도
  판별 노드 219
예제
  개요 5
  애플리케이션 안내서 3
예측
  개요 323
  예측변수 계열 329
예측변수
  대용 95
  분석을 위한 선택 61, 62, 63
  선별 62, 63
  의사결정 트리 95
  중요도 순위화 61, 62, 63
예측변수 계열 329
  결측 데이터 329
예측변수 공간 차트
  최근접 이웃 분석에서 393
예측변수 선별 62, 63
예측변수 선택
  최근접 이웃 분석에서 395
예측변수 순위화 61, 62, 63
예측변수 중요도
  랜덤 트리 모델 130
  로지스틱 회귀분석 모델 211
  모델 결과 37, 47, 49
  선형 모델 195
  시계열 모델 367
  신경망 161
  일반화 선형 모델 229
  최근접 이웃 분석에서 394
  판별 모델 221
  필드 필터링 49
  GLE 모델 253
  Linear-AS 모델 200
  LSVM 모델 384
  Tree-AS 모델 125
오류 요약
  최근접 이웃 분석에서 395
오분류 비용 40
  C5.0 노드 119
요인 모델
  결측값 처리 215
  고급 옵션 215
  고급 출력 217
  고유값 215
  모델 너깃 217
  모델 옵션 215
  모델링 노드 214
  반복 215
요인 모델 (계속)
  방정식 217
  요인 수 215
  요인 스코어 215
  회전 216
우도비 검정
  로지스틱 회귀분석 모델 208, 213
우도비 카이제곱
  필드선택 61
  CHAID 노드 116
  Tree-AS 노드 122
원시 성향 스코어 38
위험
  내보내기 104
위험 추정값
  의사결정 트리 이익 101
유사 R-제곱
  로지스틱 회귀분석 모델 213
유의 수준
  병합 116, 122
의사결정 목록 모델
  검색 방향 168
  검색 범위 170
  고급 옵션 170
  구간화 방법 170
  대안 탭 173
  모델 옵션 168
  모델링 노드 167
  목표 값 168
  뷰어 작업공간 172
  뷰어에 대한 작업 175
  설정 171
  세그먼트 170
  스냅샷 탭 174
  스코어링 170
  요구 사항 167
  작업 모델 분할창 172
  PMML 170
  SQL 생성 171
의사결정 트리 가지치기 108, 112
의사결정 트리 모형 93, 96, 106, 108, 109, 110, 118, 121, 127, 133, 136, 138
  결과 내보내기 104
  그래프 생성 138
  대용 95
  모델링 노드 105
  뷰어 136
  사용자 정의 분할 94

의사결정 트리 모형 (계속)
생성 101, 102
예측변수 95
오분류 비용 113, 114, 124, 129
이익 97, 99, 100
ROI 99
이단계 군집 270, 271, 272, 274, 276
이단계 군집 모델 269, 270
군집 수 269
군집화 270
모델 너깃 270
모델 너깃에서 그래프 생성 285
모델링 노드 268
옵션 269
이상값 처리 269
필드의 표준화 269
이벤트
식별 325
이상 항목 발견 모델 67
결측값 65
스코어링 67
이상 항목 지수 65
이상 항목 필드 65, 67
잡음 수준 65
절사 값 65, 67
조정 계수 65
피어 그룹 65, 67
이상값 326
가법 수정 326
결정적 326
계열에서 325
계절 가법 326
국소적 추세 326
수준 이동 326
일시적 변경 326
혁신적 326
이익
내보내기 104
의사결정 트리 97, 99
의사결정 트리 이익 99
차트 187
이익 기반 선택 100
이쿼멕스 회전
PCA/요인 모델 216
이항 로지스틱 회귀분석 모델 202, 203
인스턴스 295, 310
일반 선형 모델
일반화 선형 혼합 모델 231
일반 추정가능 함수
일반화 선형 모델 228
일반화 선형 모델
고급 옵션 225
고급 출력 228, 230
모델 너깃 229, 230
모델 양식 224
모델링 노드 223, 245
성향 스코어 230
수렴 옵션 227
일반화 선형 혼합 모델 231
필드 223
일반화 선형 혼합 모델 231
고정 계수 242
고정 효과 234, 241
공분산 모수 243
관측값 별 예측값 241
데이터 구조 241
랜덤 효과 236
랜덤 효과 공분산 242
랜덤 효과 블록 236
모델 보기 240
모델 요약 240
목표 분포 232
분류표 241
분석 가중치 237
사용자 정의 항 235
설정 244
스코어링 옵션 239
연결 함수 232
오프셋 237
주변 평균 추정 239
평균 추정 243
일시적 변경 이상값 326
입력 필드
분석을 위한 선택 60
선별 60
입력 필드 선별 60

[자]

자기상관 함수
계열 328
자동 군집 모델 69
결과 브라우저 창 87
모델 너깃 87
모델 삭제 86
모델 순위화 84
자동 군집 모델 (계속)
모델 유형 85
모델링 노드 83, 84
모델링 노드 및 너깃 생성 88
알고리즘 설정 70
중지 규칙 70
파티션 85
평가 차트 89
자동 데이터 준비
선형 모델에서 194
자동 분류자 모델 69
결과 브라우저 창 87
모델 너깃 87
모델 삭제 77
모델 순위화 72
모델 유형 74
모델링 노드 71, 72
모델링 노드 및 너깃 생성 88
설정 77
소개 71
알고리즘 설정 70
중지 규칙 70
파티션 74
평가 그래프 89
평가 차트 89
자동 수치 모델 69
결과 브라우저 창 87
모델 너깃 87
모델 유형 80
모델링 노드 78, 79
모델링 노드 및 너깃 생성 88
모델링 옵션 79
설정 83
알고리즘 설정 70
중지 규칙 70, 80
평가 그래프 89
평가 차트 89
자동화된 모델링 노드
자동 군집 모델 69
자동 분류자 모델 69
자동 수치 모델 69
자연 로그 변환 329
시계열 모델 생성기 363
자율 학련 262
자체 구성 맵 262
자체 학습 반응 모델
모델 너깃 374
모델 새로 고침 372

자체 학습 반응 모델 (계속)
모델링 노드 371
변수 중요도 374
설정 374
필드 옵션 371
작업 모델 분할창 172
장기적인 모델
일반화 선형 혼합 모델 231
장바구니 데이터 302, 303
적중
의사결정 트리 이익 97
적합도 통계량
로지스틱 회귀분석 모델 213
일반화 선형 모델 228
전이 함수 363
계절 차수 363
분모 차수 363
분자 차수 363
지연 363
차이 차수 363
전항
없는 규칙 294
정규화 카이제곱
apriori 평가 측도 290
정보 기준
선형 모델에서 192
linear-AS 모델에서 199
정보 차이
apriori 평가 측도 290
제곱근 변환 329
시계열 모델 생성기 363
주기성
시계열 모델 생성기 363
주성분분석(PRINCALS). PCA 모델 참조 214, 217
주효과
로지스틱 회귀분석 모델 206
중요도
모델에서 예측변수 37, 47, 49
예측변수 순위화 61, 62, 63
필드 필터링 49
중첩, 교차 검증 391
지수
의사결정 트리 이익 97
지수평활 355
지시문
의사결정 트리 104
지원
규칙 지원 295, 310
시퀀스 노드 307
시퀀스에 대해 310
연관 규칙 298
전항 지원 295, 310
Apriori 노드 289
CARMA 노드 293, 294
지원 벡터 머신 모델
고급 옵션 380
과적합 378
모델 너깃 381, 392
모델 옵션 380
모델링 노드 379
설정 382
정보 377
조정 378
커널 기능 377
직접 오블리민 회전
PCA/요인 모델 216

[차]

차분 변환 329
차원 축소 262
차트 옵션 187
참 표 데이터 302, 303
참조 범주
로지스틱 노드 203
첫 번째 적중 규칙 세트 140
초점 레코드 388
최근접 이웃 거리
최근접 이웃 분석에서 394
최근접 이웃 모델
교차 검증 옵션 391
모델 옵션 388
모델링 노드 387
목표 옵션 387
분석 옵션 391
설정 옵션 388
이웃 옵션 389
정보 387
필드선택 옵션 390
최근접 이웃 분석
모델 보기 393
최적 서브세트
선형 모델에서 192
linear-AS 모델에서 199
추세
식별 324
측도 새로 고침 183

[카]

카이제곱
필드선택 61
CHAID 노드 116
Tree-AS 노드 122
커널 기능
지원 벡터 머신 모델 377
코호넨 모델 262, 263, 264
고급 옵션 264
모델 너깃 265
모델 너깃에서 그래프 생성 285
모델링 노드 262
신경망 262, 265
이분형 세트 인코딩 옵션(제거됨) 263
이웃 262, 264
중지 기준 263
피드백 그래프 263
학습률 264
쿼티멕스 회전
PCA/요인 모델 216

[타]

통계 모델 189
투잉 불순도 측도 115
투표 규칙 세트 140
트랜잭션 데이터 302, 303
시퀀스 노드 306
Apriori 노드 33
CARMA 노드 292
MS 연관 규칙 노드 33
트리 깊이 112, 122, 128
트리 맵
그래프 생성 138
의사결정 트리 모형 136
트리 작성기 93, 96
결과 내보내기 104
그래프 생성 138
대용 95
모델 생성 101, 102
사용자 정의 분할 94
예측변수 95
이익 97, 99, 100

트리 작성기 (계속)
ROI 99
트리 지시문 110
의사결정 트리 104
CHAID 노드 102
C&R 트리 노드 102
QUEST 노드 102

[파]

파티션 306
선택 306
판별 모델
고급 옵션 219
고급 출력 219, 222
단계별 기준(필드 선택) 221
모델 너깃 221, 222, 223
모델 양식 218
모델링 노드 218
성향 스코어 222
수렴 기준 219
스코어링 221
펄스
계열에서 325
편자기상관 함수
계열 328
편집
고급 매개변수 177
평가 그래프
자동 분류자 모델 89
자동 수치 모델 89
평가 차트
자동 군집 모델 89
자동 분류자 모델 89
자동 수치 모델 89
평가 측도
Apriori 노드 290
포아송 회귀분석
일반화 선형 혼합 모델 231
포인트 개입
식별 325
표 출력 전치 303
표 형식 데이터 302
시퀀스 노드 306
전치 303
Apriori 노드 33
CARMA 노드 292
프로멕스 회전
PCA/요인 모델 216
프로빗 분석
일반화 선형 혼합 모델 231
피어 그룹
이상 항목 발견 65
필드 옵션
모델링 노드 33
Cox 노드 255
SLRM 노드 371
필드 중요도
모델 결과 37, 47, 49
필드 순위화 61, 62, 63
필드 필터링 49
필드선택 모델 62, 63
예측변수 선별 60, 62
예측변수 순위화 60, 62
중요도 60, 62
필터 노드 생성 63
필터 노드
의사결정 트리에서 생성 104
필터링 규칙 295, 310
연관 규칙 298

[하]

함수 변환 329
혁신적 이상값 326
혼돈 행렬
LSVM 모델 384
혼합 모델
일반화 선형 혼합 모델 231
확률
로지스틱 회귀분석 모델 211
회귀 모형
모델링 노드 190, 198
회귀분석 이익
의사결정 트리 99, 100
회귀분석 트리 108, 109, 121, 127
회전
PCA/요인 모델 216
후항
복수 후항 294

[숫자]

1에 대한 신뢰 지수 차이
apriori 평가 측도 290
4분위 맵
최근접 이웃 분석에서 395

A

Akaike 정보 기준
선형 모델에서 192
linear-AS 모델에서 199
ANOVA
선형 모델에서 196
apriori 모델
고급 옵션 290
모델링 노드 289
모델링 노드 옵션 289
테이블 대 트랜잭션 데이터 33
평가 측도 290
ARIMA 모델 355
전이 함수 363

B

Box의 M 검정
판별 노드 219

C

C5.0 모델
가지치기 119
모델 너깃 133, 140, 141
모델 너깃에서 그래프 생성 138
모델링 노드 118, 119, 136, 137, 138
부스팅 119, 138
오분류 비용 119
옵션 119
CARMA 모델
고급 옵션 294
내용 필드 292
데이터 형식 292
모델링 노드 292
모델링 노드 옵션 293
복수 후항 302
시간 필드 292
테이블 대 트랜잭션 데이터 294
필드 옵션 292
ID 필드 292
CHAID 모델
모델 너깃 133
모델 너깃에서 그래프 생성 138

CHAID 모델 (계속)
모델링 노드 93, 106, 108, 136, 137
목적 110
앙상블 113
오분류 비용 114
작성 옵션 110
중지 옵션 112, 123
트리 깊이 112, 122
필드 옵션 110
exhaustive CHAID 112, 122
Cox 회귀 모형 260
고급 옵션 257
고급 출력 258, 260
단계별 기준 258
모델 너깃 259
모델 옵션 256
모델링 노드 255
설정 옵션 259
수렴 기준 258
필드 옵션 255
Cramér의 V
필드선택 61
C&R 트리 모델
가지치기 112
대용 112
모델 너깃 133
모델 너깃에서 그래프 생성 138
모델링 노드 93, 106, 108, 136, 137
목적 110
불순도 측도 115
빈도 가중치 33
사전 확률 113
앙상블 113
오분류 비용 113
작성 옵션 110
중지 옵션 112
케이스 가중치 33
트리 깊이 112
필드 옵션 110

D

DTD 54

E

Excel로 평가 184
exhaustive CHAID 93, 112, 122

F

F 통계량
선형 모델에서 192
필드선택 61
linear-AS 모델에서 199

G

Gini 불순도 측도 115
GLE 모델
모델 선택 옵션 252
모델 정보 253
모델 효과 248
모델링 노드 254
목표 분포 246
분석 가중치 250
사용자 정의 항 249
스코어링 옵션 253
연결 함수 246
예측변수 중요도 253
오프셋 250
작성 옵션 250
출력 253

H

Hosmer 및 Lemeshow 적합도
로지스틱 회귀분석 모델 213

I

IBM SPSS Modeler 1
문서 3
IBM SPSS Modeler Server 2
ID 필드
시퀀스 노드 306
CARMA 노드 292

K

KNN. 최근접 이웃 모델을 참조하십시오. 387
K-평균 모델 266, 267
거리 필드 266
고급 옵션 267
군집화 266, 268
모델 너깃 267, 268
K-평균 모델 (계속)
모델 너깃에서 그래프 생성 285
세트의 인코딩 값 267
중지 기준 267
K-평균-AS 노드 277, 420

L

L 행렬
일반화 선형 모델 228
linearnode 노드 190
Linear-AS 노드 199
Linear-AS 모델 199
관측값 별 예측값 200
너깃 설정 201
레코드 요약 200
모델 선택 199
모델 옵션 200
모델 정보 200
범주형 예측변수에 대한 정렬 순서 199
신뢰구간 199
신뢰수준 199
예측변수 중요도 200
이원 상호작용 고려 199
절편 포함 199
정보 기준 200
출력 200
R 제곱 통계 200
LM 검정
일반화 선형 모델 228
LSVM 모델
관측값 별 예측값 384
레코드 요약 384
모델 정보 384
예측변수 중요도 384
출력 384
혼돈 행렬 384

M

MLP(다중 레이어 퍼셉트론)
신경망에서 155
MS Excel 설정 통합 형식 184

N

nodeName 노드 231

O

One-Class SVM 노드 410, 411, 412

P

p 값 61
PCA 모델
  결측값 처리 215
  고급 옵션 215
  고급 출력 217
  고유값 215
  모델 너깃 217
  모델 옵션 215
  모델링 노드 214
  반복 215
  방정식 217
  요인 수 215
  요인 스코어 215
  회전 216
Pearson 카이제곱
  필드선택 61
  CHAID 노드 116
  Tree-AS 노드 122
PMML
  모델 가져오기 44, 54, 55
  모델 내보내기 44, 54, 55
python 노드 398, 399, 400, 401, 402, 404, 406, 407, 408, 410, 411, 412

Q

QUEST 모델
  가지치기 112
  대용 112
  모델 너깃 133
  모델 너깃에서 그래프 생성 138
  모델링 노드 93, 106, 109, 136, 137
  목적 110
  사전 확률 113
  앙상블 113
  오분류 비용 113
  작성 옵션 110
  중지 옵션 112
  트리 깊이 112
  필드 옵션 110

R

R 제곱
  선형 모델에서 194, 200
RBF(방사형 기저함수)
  신경망에서 155
ROI
  의사결정 트리 이익 99

S

SLRM. 자체 학습 반응 모델 참조 371
SMOTE 노드 398
spark 노드 277, 415, 416, 417, 420
SQL
  규칙 세트 137
  내보내기 45
  랜덤 트리 모델 132
  로지스틱 회귀분석 모델 212
  GLE 모델 254
  Tree-AS CHAID 모델 126
STP 노드 330
STP 모델
  모델 너깃 335
  시간 구간 옵션 331
  필드 옵션 330
STP(Spatio-Temporal Prediction) 330
STP(spatio-temporaral prediction) 고급 작성 옵션 333
STP(spatio-temporaral prediction) 모델 옵션 335
STP(spatio-temporaral prediction) 출력 334
STP(spatio-temporaral prediction)에 대한 모델 옵션 335
STP(spatio-temporaral prediction)에 대한 작성 옵션 333
STP(spatio-temporaral prediction)의 출력 334
SVM 모델 과적합 378
SVM. 지원 벡터 머신 모델을 참조하십시오. 377

T

T 통계량
  필드선택 61
TCM 노드 336
TCM 모델
  모델 너깃 348
  모델 너깃 설정 348
  모델링 노드 336
till-roll 데이터 302, 303
Tree-AS 모델
  구간화 122
  모델 정보 125
  모델링 노드 121, 126
  예측변수 중요도 125
  오분류 비용 124
  작성 옵션 110, 122
  중지 옵션 123
  출력 125
  트리 깊이 122
  필드 옵션 121
TwoStep-AS 군집 모델
  모델링 노드 270
TwoStep-AS 모델
  모델 너깃 276
  모델 너깃 설정 276
two-headed 규칙 294
t-SNE 노드 404, 406
t-SNE 모델 너깃 407

W

Wald 통계량 208, 209

X

XGBoost Linear 노드 399, 400, 401
XGBoost Tree 노드 401, 402, 404
XGBoost-AS 노드 416, 417, 420