IBM SPSS Modeler 18.1.1
애플리케이션 안내서

IBM

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 383 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

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 제품 개요 7
시작하기 7
IBM SPSS Modeler 시작 7
명령행에서 시작 8
IBM SPSS Modeler Server에 연결 8
Analytic Server에 연결 11
팀 디렉토리 변경 12
다중 IBM SPSS Modeler 세션 시작 13
IBM SPSS Modeler 인터페이스 살펴보기 . . . 13
IBM SPSS Modeler 스트림 캔버스 14
노드 팔레트 15
IBM SPSS Modeler 관리자 16
IBM SPSS Modeler 프로젝트 17
IBM SPSS Modeler 도구 모음 18
도구 모음 사용자 정의 20
IBM SPSS Modeler 창 사용자 정의 20
스트림 아이콘 크기 변경 21
IBM SPSS Modeler에서 마우스 사용 22
단축키 사용 22
인쇄 23
IBM SPSS Modeler 자동화 24

제 3 장 모델링 소개 25
스트림 작성 27
모델 찾아보기 32
모델 평가 37
레코드 스코어링 40
요약 41

제 4 장 플래그 대상에 대한 자동화된 모델링 . . 43
모델링 고객 반응(자동 분류자) 43
히스토리 데이터 43
스트림 작성 44
모델 생성 및 비교 49
요약 54

제 5 장 연속형 대상에 대한 자동화된 모델링 . . 55
특성 값(자동 숫자) 55
학습 데이터 55
스트림 작성 56
모델 비교 59
요약 61

제 6 장 자동 데이터 준비(ADP) 63
스트림 작성 63
모델 정확도 비교 68

제 7 장 분석용 데이터 준비(데이터 검토) . . . 73
스트림 작성 73
통계 및 도표 찾아보기 76
이상치 및 결측값 처리 79

제 8 장 약물 치료(예비 그래프/C5.0) 83
텍스트 데이터에서 읽기 83
테이블 추가 86
분포 그래프 작성 87
산점도 작성 89
웹 그래프 작성 90
새 필드 파생 91
모델 작성 94
모델 찾아보기 96
분석 노드 사용 97

제 9 장 예측변수 선별(필드선택) 99
스트림 작성 100
모델 작성 102
결과 비교 104
요약 105

제 10 장 입력 데이터 문자열 길이 감소(재분류 노드) 107
입력 데이터 문자열 길이 감소(재분류) 107
데이터 재분류. 107

제 11 장 고객 반응 모델링(의사결정 목록) . . 113
히스토리 데이터 113
스트림 작성 114
모델 작성 116
Excel을 사용하여 사용자 정의 측도 계산 . . . 129
Excel 템플리트 수정 135
결과 저장 137

제 12 장 통신 고객 분류(다항 로지스틱 회귀분석). 139
스트림 작성 139
모델 찾아보기. 143

제 13 장 통신 서비스 제공자를 바꾸는 고객(이항 로지스틱 회귀분석). 149
스트림 작성 149
모델 찾아보기. 156

제 14 장 대역폭 활용 시계열 분석(시계열) . . 163
시계열 노드를 사용하여 예측 163
스트림 작성 164
데이터 탐색 165
날짜 정의 169
대상 정의 171
시간 구간 설정 172
모델 작성 173
모델 탐색적 데이터분석 175
요약 181
시계열 모델 다시 적용 181
스트림 검색 182
저장된 모델 검색 183
모델링 노드 생성 184
새 모델 생성 184
새 모델 탐색 185
요약 188

제 15 장 카탈로그 판매 예측(시계열) 189
스트림 작성 190
데이터 탐색 193
지수평활. 193
ARIMA 198
요약 202

제 16 장 고객에 대한 오퍼 작성 (Self-Learning) 203
스트림 작성 204
모델 찾아보기. 209

제 17 장 대출 체납자 예측(베이지안 네트워크) 215
스트림 작성 215
모델 찾아보기. 219

제 18 장 월 단위로 모델 재교육(베이지안 네트워크) 223
스트림 작성 223
모델 평가 228

제 19 장 소매 판매 프로모션(신경망/C&RT) 235
데이터 탐색 235
학습 및 검정 238

제 20 장 상태 모니터링(신경망/C5.0) 241
데이터 탐색 242
데이터 준비 243
학습 245
검정 245

제 21 장 통신 고객 분류(판별 분석). 247
스트림 작성 247
모델 탐색 252
통신회사 고객을 분류하기 위해 판별 분석을 사용하여 결과 분석. 253
요약 257

제 22 장 구간 중도절단 생존 데이터 분석(일반화 선형 모델). 259
스트림 작성 259
모델 효과 검정 263
치료법 전용 모델 적합화. 264
모수 추정값 265
예측된 재발 및 생존 확률 266
주기 기준 반복 확률 모델링. 270
모델 효과 검정 275
축소된 모델 적합화. 275
모수 추정값 277
예측된 재발 및 생존 확률 278
요약 282
관련 프로시저. 283
권장 참고 자료 283

제 23 장 선박 손상 비율을 분석하기 위해 포아송 회귀분석 사용(일반화 선형 모델) 285
"과분산된" 포아송 회귀분석 적합화 285
적합도 통계 289
총괄 검정 289
모델 효과 검정 290
모수 추정값 290
대체 모델 적합화 291
적합도 통계 293
요약 294
관련 프로시저 294
권장 참고 자료 294

제 24 장 자동차 보험 청구에 감마회귀 적합화 (일반화 선형 모델) 295
스트림 작성 295
모수 추정값 299
요약 299
관련 프로시저 299
권장 참고 자료 300

제 25 장 세포 표본 분류(SVM) 301
스트림 작성 302
데이터 탐색 307
다른 함수 시도 309
결과 비교 311
요약 312

제 26 장 고객 이탈 시간을 모델링하기 위해 Cox 회귀분석 사용 313
적합한 모델 작성 313
중도절단 케이스 317
범주형 변수 코딩 318
변수 선택 319
공변량 평균값 321
생존함수 곡선 322
위험함수 곡선 323
평가 323
보유 고객 예측 수 추적 329
스코어링 340
요약 345

제 27 장 장바구니 분석(규칙 귀납/C5.0) . . . 347
데이터 액세스 347
바스켓 내용에서 유사성 검색 349
고객 집단 프로파일링 352
요약 353

제 28 장 새 차량 오퍼링(KNN) 평가 355
스트림 작성 356
출력 탐색 360
예측변수 공간 361
피어 차트 362
이웃 및 거리 테이블 364
요약 364

제 29 장 비즈니스 메트릭(TCM)에서 인과 관계 찾기 365
스트림 작성 365
분석 실행 366
전체 모델 품질 차트 368
전체 모델 시스템 369
영향 다이어그램 371
이상치의 근본 원인 판별 373
시나리오 실행 377

주의사항 383
상표 384
제품 문서의 이용 약관 385

색인 387

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 디자인된 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 포함됨)
- 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 모형을 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 사용될 수 있으며 현재 SPSS Modeler Server 고객에게 무료로 제공됩니다. 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 비구조적 텍스트 데이터에 대한 작업을 하도록 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 예측 분석 공정능력을 제공합니다. 구독 에디션을 사용하면 제품 업데이트를 정기적으로 다운로드할 수 있습니다.

문서

문서는 SPSS Modeler의 도움말 메뉴에서 사용할 수 있습니다. 제품 외부에서 공개적으로 사용할 수 있는 Knowledge Center에서 열 수 있습니다.

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

SPSS Modeler Professional 문서

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

- **IBM SPSS Modeler 사용자 안내서.** General introduction to using SPSS Modeler, 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 방법을 포함하여 SPSS Modeler 사용에 대한 일반 소개입니다.
- **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 버전을 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.** Deployment Manager 애플리케이션에 포함된 관리 콘솔 사용자 인터페이스를 사용하여 IBM SPSS Modeler Server를 모니터링하고 구성하는 방법에 대한 정보입니다.
- **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 마이너에 의해 관리되는 거대한 데이터 스토어보다 훨씬 작지만, 관련된 개념과 방법은 실제 애플리케이션에 대해 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. 자세한 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 제품 개요

시작하기

데이터 마이닝 애플리케이션으로서 IBM SPSS Modeler는 대규모 데이터 세트에서 유용한 관계를 찾기 위한 전략적 접근 방식을 제공합니다. 보다 전통적인 통계 방법과는 대조적으로 시작할 때 찾아야 하는 것을 꼭 알고 있을 필요가 없습니다. 유용한 정보를 찾을 때 까지 서로 다른 모델을 맞춰보고 서로 다른 관계를 조사하면서 데이터를 탐색할 수 있습니다.

IBM SPSS Modeler 시작

애플리케이션을 시작하려면 다음을 클릭하십시오.

시작 > **[모든] 프로그램** > **IBM SPSS Modeler <버전>** > **IBM SPSS Modeler <버전>**

몇 초 후에 기본 창이 표시됩니다.

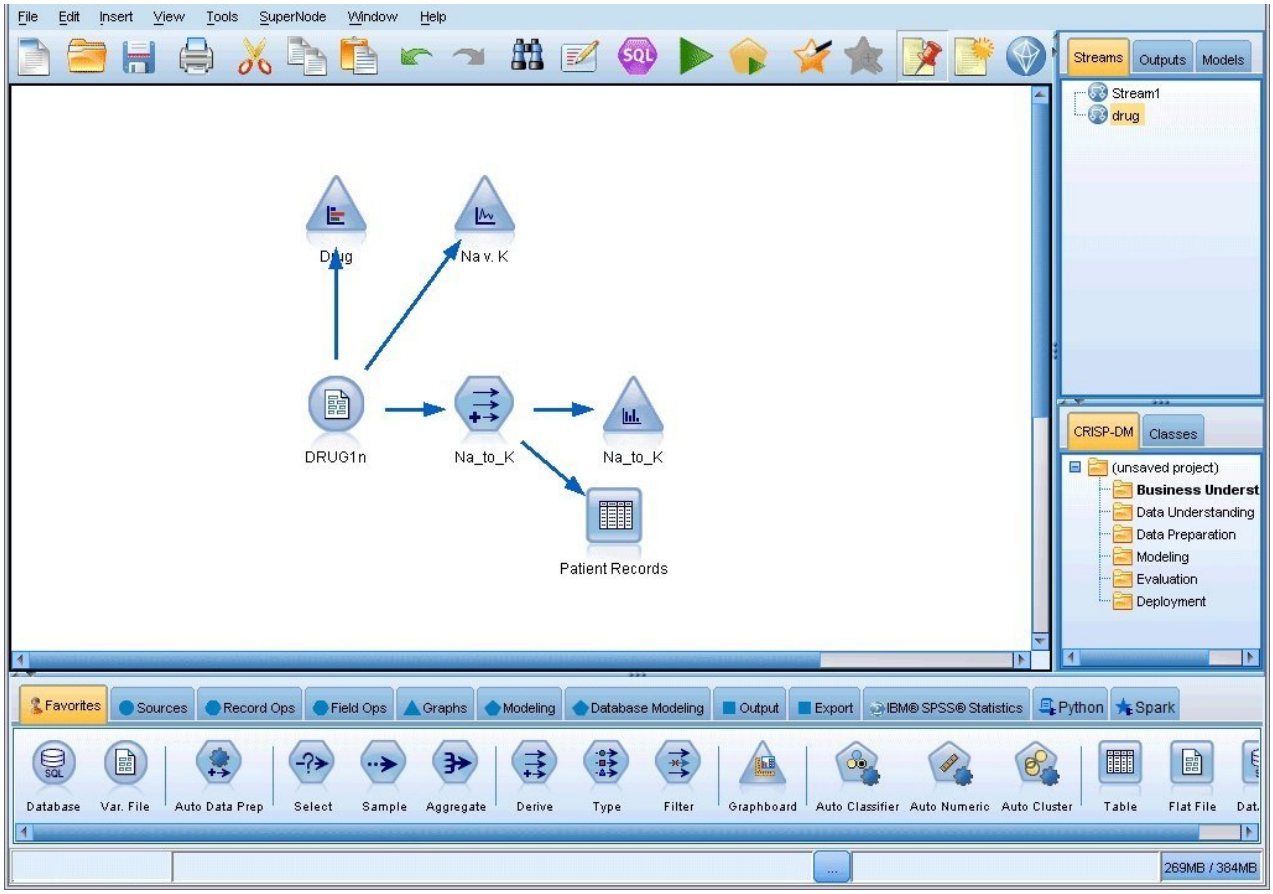


그림 *1. IBM SPSS Modeler* 기본 애플리케이션 창

명령행에서 시작

운영 체제의 명령행을 사용하여 다음과 같이 IBM SPSS Modeler를 시작할 수 있습니다.

1. IBM SPSS Modeler가 설치된 컴퓨터에서 DOS 또는 명령 프롬프트 창을 여십시오.
2. 대화식 모드로 IBM SPSS Modeler 인터페이스를 시작하려면 `modelerclient` 명령 뒤에 필수 인수를 입력하십시오. 예를 들어, 다음과 같습니다.

```
modelerclient -stream report.str -execute
```

사용 가능한 인수(플래그)를 사용하면 필요에 따라 서버에 연결하고 스트림을 로드하며, 스크립트를 실행하거나 다른 모수를 지정할 수 있습니다.

IBM SPSS Modeler Server에 연결

IBM SPSS Modeler는 독립형 애플리케이션으로서나 IBM SPSS Modeler Server에 직접 또는 IBM SPSS Modeler Server 또는 IBM SPSS Collaboration and Deployment Services로부터 프로세스 조정자 플러그인을 통해 서버 클러스터에 연결된 클라이언트로서 실행할 수 있습니다. 현재 연결 상태는 IBM SPSS Modeler 창의 왼쪽 아래에 표시됩니다.

서버에 연결하려고 할 때마다 연결하려는 서버 이름을 수동으로 입력하거나 이전에 정의한 이름을 선택할 수 있습니다. 그러나 IBM SPSS Collaboration and Deployment Services가 있는 경우 서버 로그인 대화 상자에서 서버 목록이나 서버 클러스터를 통해 검색할 수 있습니다. 네트워크에서 실행 중인 Statistics 서비스를 통해 찾아보는 기능은 프로세스 조정자를 통해 사용 가능하게 됩니다.

서버에 연결하기

1. 도구 메뉴에서 **서버 로그인**을 클릭하십시오. 서버 로그인 대화 상자가 열립니다. 또는 IBM SPSS Modeler 창의 연결 상태 영역을 두 번 클릭하십시오.
2. 대화 상자를 사용하여 로컬 서버 컴퓨터에 연결하기 위한 옵션을 지정하거나 테이블에서 연결을 선택하십시오.
   - 연결을 추가하거나 편집하려면 **추가** 또는 **편집**을 클릭하십시오. 자세한 정보는 10 페이지의 『IBM SPSS Modeler Server 연결 추가 및 편집』의 내용을 참조하십시오.
   - 프로세스 조정자에서 서버 또는 서버 클러스터에 액세스하려면 **검색**을 클릭하십시오. 자세한 정보는 10 페이지의 『IBM SPSS Collaboration and Deployment Services에서 서버 검색』의 내용을 참조하십시오.

   서버 테이블. 이 테이블에는 정의된 서버 연결 세트가 포함됩니다. 테이블은 기본 연결, 서버 이름, 설명 및 포트 번호를 표시합니다. 새 연결을 수동으로 추가하고 기존 연결을 선택하거나 검색할 수도 있습니다. 특정 서버를 기본 연결로 설정하려면 연결의 테이블에서 기본 열의 선택란을 선택하십시오.

   기본 데이터 경로. 서버 컴퓨터에서 데이터에 사용된 경로를 지정하십시오. 필요한 위치로 이동하려면 생략 기호 단추**(...)**를 클릭하십시오.

   자격 설정. 로컬 컴퓨터 사용자 이름과 비밀번호 세부사항을 사용하여 서버에 로그인하려고 시도하는 **싱글 사인온** 기능을 사용하려면 이 상자를 선택 해제 상태로 두십시오. 싱글 사인온이 가능하지 않거나 싱글 사인온을 사용 안함으로 설정하기 위해 이 상자를 선택한 경우에는(예: 관리자 계정에 로그인), 신임 정보를 입력하라는 다음 필드가 사용 가능하게 됩니다.

   사용자 ID. 서버에 로그온하는 데 사용할 사용자 이름을 입력합니다.

   비밀번호. 지정된 사용자 이름과 연관된 비밀번호를 입력합니다.

   도메인. 서버에 로그온하는 데 사용된 도메인을 지정합니다. 도메인 이름은 서버 컴퓨터가 클라이언트 컴퓨터와 다른 Windows 도메인에 있을 때에만 필요합니다.
3. 연결을 완료하려면 **확인**을 클릭하십시오.

서버에서 연결 해제하기

1. 도구 메뉴에서 **서버 로그인**을 클릭하십시오. 서버 로그인 대화 상자가 열립니다. 또는 IBM SPSS Modeler 창의 연결 상태 영역을 두 번 클릭하십시오.
2. 대화 상자에서 로컬 서버를 선택하고 **확인**을 클릭하십시오.

IBM SPSS Modeler Server 연결 추가 및 편집

서버 로그인 대화 상자에서 수동으로 서버 연결을 편집하거나 추가할 수 있습니다. 추가를 클릭하여 서버 연결 세부사항을 입력할 수 있는 서버 추가/편집 대화 상자에 액세스할 수 있습니다. 기존 연결을 선택하고 서버 로그인 대화 상자에서 편집을 클릭하면 사용자가 변경을 수행할 수 있도록 해당 연결에 대한 세부사항과 함께 서버 추가/편집 대화 상자가 열립니다.

참고: IBM SPSS Collaboration and Deployment Services에서 추가된 서버 연결은 편집할 수 없습니다. 이름, 포트 및 기타 세부사항이 IBM SPSS Collaboration and Deployment Services에 정의되어 있기 때문입니다. 우수 사례에서는 IBM SPSS Collaboration and Deployment Services 및 SPSS Modeler 클라이언트 둘 모두와 통신하는 데에는 동일 포트를 사용해야 한다고 지시합니다. 이들은 options.cfg 파일에 `max_server_port` 및 `min_server_port`로 설정할 수 있습니다.

서버 연결 추가

1. 도구 메뉴에서 **서버 로그인**을 클릭하십시오. 서버 로그인 대화 상자가 열립니다.
2. 이 대화 상자에서 **추가**를 클릭하십시오. 서버 로그인 서버 추가/편집 대화 상자가 열립니다.
3. 서버 연결 세부사항을 입력하고 **확인**을 클릭하여 연결을 저장하고 서버 로그인 대화 상자로 돌아가십시오.

- **서버.** 사용 가능한 서버를 지정하거나 목록에서 하나를 선택하십시오. 서버 컴퓨터는 영숫자 이름(예: *myserver*)이나 서버 컴퓨터에 지정된 IP 주소(예: 202.123.456.78)로 식별할 수 있습니다.
- **포트.** 서버가 청취 중인 포트 번호를 제공하십시오. 기본값이 작동하지 않으면 시스템 관리자에게 올바른 포트 번호를 문의하십시오.
- **설명.** 이 서버 연결에 대한 선택적 설명을 입력하십시오.
- **보안 연결을 확인하십시오(SSL 사용).** SSL(**Secure Sockets Layer**) 연결을 사용해야 하는지 여부를 지정합니다. SSL은 네트워크를 통해 전송된 데이터를 보안하기 위해 주로 사용하는 프로토콜입니다. 이 기능을 사용하려면 IBM SPSS Modeler Server를 호스팅하는 서버에서 SSL이 사용 가능으로 설정되어야 합니다. 필요한 경우, 세부사항은 로컬 관리자에게 문의하십시오.

서버 연결 편집하기

1. 도구 메뉴에서 **서버 로그인**을 클릭하십시오. 서버 로그인 대화 상자가 열립니다.
2. 이 대화 상자에서 편집할 연결을 선택한 다음 **편집**을 클릭하십시오. 서버 로그인 서버 추가/편집 대화 상자가 열립니다.
3. 서버 연결 세부사항을 변경하고 **확인**을 클릭하여 변경사항을 저장하고 서버 로그인 대화 상자로 돌아가십시오.

IBM SPSS Collaboration and Deployment Services에서 서버 검색

서버 연결에 수동으로 들어가는 대신에 IBM SPSS Collaboration and Deployment Services에서 사용 가능한 프로세스 조정자를 통해 네트워크에서 사용 가능한 서버나 서버 클러스터를 선택할 수 있습니다. 서버 클러스터는 프로세스 조정자가 처리 요청에 대한 응답에 가장 적합한 서버를 판별하는 데 사용하는 서버 그룹입니다.

서버 로그인 대화 상자에서 서버를 수동으로 추가할 수 있지만 사용 가능한 서버를 검색하면 올바른 서버 이름과 포트 번호를 몰라도 서버에 연결할 수 있습니다.이 정보는 자동으로 제공됩니다.그러나 사용자 이름, 도메인 및 비밀번호 등과 같은 올바른 로그온 정보는 여전히 필요합니다.

참고: 프로세스 조정자 기능에 대한 액세스가 없는 경우에는 연결하려는 서버 이름을 수동으로 입력하거나 이전에 정의된 이름을 선택할 수도 있습니다. 자세한 정보는 10 페이지의 『IBM SPSS Modeler Server 연결 추가 및 편집』의 내용을 참조하십시오.

서버와 군집 검색하기

1. 도구 메뉴에서 **서버 로그인**을 클릭하십시오. 서버 로그인 대화 상자가 열립니다.
2. 이 대화 상자에서 **검색**을 클릭하여 서버 검색 대화 상자를 엽니다. 프로세스 조정자를 찾으려고 시도할 때 IBM SPSS Collaboration and Deployment Services에 로그온되어 있지 않은 경우에는 로그온하라는 메시지가 표시됩니다.
3. 목록에서 서버나 서버 클러스터를 선택하십시오.
4. **확인**을 클릭하여 대화 상자를 닫고 서버 로그인 대화 상자에서 테이블에 이 연결을 추가하십시오.

Analytic Server에 연결

여러 Analytic Server를 사용할 수 있으면 분석 서버 연결 대화 상자를 사용하여 IBM SPSS Modeler에서 사용할 서버를 두 개 이상 정의할 수 있습니다. 관리자가 이미 `<Modeler_install_path>/config/options.cfg` 파일에 기본 Analytic Server를 설정했을 가능성이 있습니다. 단, 서버를 정의한 후에도 사용 가능한 다른 서버를 사용할 수 있습니다. 예를 들어, Analytic Server 소스 및 내보내기 노드를 사용할 때, 각 분기가 실행될 때 고유 Analytic Server를 사용하고 데이터가 IBM SPSS Modeler Server에 가져오지 않도록 스트림의 여러 다른 분기에서 다른 Analytic Server 연결을 사용할 수 있습니다. 분기에 둘 이상의 Analytic Server 연결이 포함된 경우, 데이터가 Analytic Server에서 IBM SPSS Modeler Server로 풀링됩니다. 제한사항을 포함하여 자세한 정보는 Analytic Server 스트림 특성의 내용을 참조하십시오.

새 Analytic Server 연결을 작성하려면 **도구** > **Analytic Server 연결**로 이동하여 대화 상자의 다음 섹션에 필수 정보를 제공하십시오.

연결

URL `https://hostname:port/contextroot` 형식으로 Analytic Server의 URL을 입력하십시오. 여기서, `hostname`은 Analytic Server의 IP 주소 또는 호스트 이름이며 `port`는 포트 번호이며 `contextroot`는 Analytic Server의 컨텍스트 루트입니다.

테넌트 IBM SPSS Modeler Server가 멤버인 테넌트의 이름을 입력하십시오. 테넌트를 모르는 경우 관리자에게 문의하십시오.

인증

모드 다음 인증 모드에서 선택하십시오.

- **사용자 이름 및 비밀번호**의 경우, 사용자가 사용자 이름 및 비밀번호를 입력해야 합니다.
- **저장된 신임 정보**의 경우, 사용자가 IBM SPSS Collaboration and Deployment Services Repository에서 신임 정보를 선택해야 합니다.
- **Kerberos**의 경우, 사용자가 서비스 프린시펄 이름 및 구성 파일 경로를 입력해야 합니다. 이 정보를 모르는 경우 관리자에게 문의하십시오.

사용자 이름 Analytic Server 사용자 이름을 입력하십시오.

비밀번호 Analytic Server 비밀번호를 입력하십시오.

연결 연결을 클릭하여 새 연결을 테스트하십시오.

연결

위 정보를 지정한 후에 **연결**을 클릭하면 연결이 이 연결 테이블에 추가됩니다. 연결을 제거하려면 연결을 선택하고 **제거**를 클릭하십시오.

관리자가 `options.cfg` 파일에서 기본 Analytic Server 연결을 정의한 경우, **기본 연결 추가**를 클릭하여 이를 사용 가능한 연결에 추가할 수도 있습니다. 사용자 이름 및 비밀번호를 입력하도록 프롬프트됩니다.

팀 디렉토리 변경

IBM SPSS Modeler Server가 수행하는 일부 작업에는 임시 파일을 작성해야 할 수도 있습니다. 기본적으로 IBM SPSS Modeler은 시스템 임시 디렉토리를 사용하여 임시 파일을 작성합니다. 다음 단계를 사용하여 임시 디렉토리의 위치를 변경할 수 있습니다.

1. `spss`라는 새 디렉토리와 *servertemp*라는 서브디렉토리를 작성하십시오.
2. IBM SPSS Modeler 설치 디렉토리의 */config* 디렉토리에 있는 *options.cfg*를 편집하십시오. 이 파일에서 `temp_directory` 모수를 편집하여 `temp_directory, "C:/spss/servertemp"`가 되도록 하십시오.
3. 이를 수행한 후에는 IBM SPSS Modeler Server 서비스를 다시 시작해야 합니다. Windows 제어판에서 **서비스** 탭을 클릭하여 이를 수행할 수 있습니다. 서비스를 중지한 다음 다시 시작하여 변경사항을 활성화하십시오. 시스템을 다시 시작하면 서비스도 다시 시작됩니다.

모든 임시 파일이 이 새 디렉토리에 작성됩니다.

참고:

- 슬래시를 사용해야 합니다.
- IBM SPSS Collaboration and Deployment Services 작업을 통해 평가 스트림을 실행하는 경우에는 `temp_directory` 설정이 적용되지 않습니다. 해당 작업을 실행하는 경우, 임시 파일이 작성됩

니다. 기본적으로 파일이 IBM SPSS Modeler Server 설치 디렉토리에 저장됩니다. IBM SPSS Modeler에서 IBM SPSS Modeler Server 연결을 작성할 때 임시 파일이 저장되는 기본 데이터 폴더를 변경할 수 있습니다.

다중 IBM SPSS Modeler 세션 시작

한 번에 둘 이상의 IBM SPSS Modeler 세션을 실행해야 하는 경우에는 IBM SPSS Modeler 및 Windows 설정을 약간 변경해야 합니다. 예를 들어, 두 개의 별도의 서버 사용권이 있고 동일한 클라이언트 시스템에서 두 개의 서로 다른 서버에 대해 두 개의 스트림을 실행하려는 경우 이를 수행할 수도 있습니다.

다중 IBM SPSS Modeler 세션 사용

1. 다음을 클릭하십시오.

   시작 > **[모든] 프로그램** > **IBM SPSS Modeler**
2. IBM SPSS Modeler 단축키(아이콘이 있음)에서 마우스 오른쪽 단추를 클릭하고 **특성**을 선택하십시오.
3. **목표** 텍스트 상자에서 `-noshare`를 문자열 끝에 추가하십시오.
4. Windows 탐색기에서 다음을 선택하십시오.

   도구 > **폴더 옵션...**
5. 파일 유형 탭에서 IBM SPSS Modeler 스트림 옵션을 선택하고 **고급**을 클릭하십시오.
6. 파일 유형 편집 대화 상자에서 IBM SPSS Modeler으로 열기를 선택하고 **편집**을 클릭하십시오.
7. **조치를 수행하는 데 사용하는 애플리케이션** 텍스트 상자에서 **-stream** 인수 앞에 `-noshare`를 추가하십시오.

IBM SPSS Modeler 인터페이스 살펴보기

데이터 마이닝 프로세스의 각 지점에서 사용하기 용이한 IBM SPSS Modeler 인터페이스는 사용자의 특정 비즈니스 전문 지식을 활용합니다. 예측, 분류, 세분화, 연관 발견과 같은 모델링 알고리즘은 강력하고 정확한 모델을 보장합니다. 모델 결과를 쉽게 배포하고 데이터베이스, IBM SPSS Statistics, 다양한 기타 애플리케이션으로 읽어들일 수 있습니다.

IBM SPSS Modeler에 대한 작업은 데이터에 대해 작업하는 3단계 프로세스입니다.

- 먼저, IBM SPSS Modeler로 데이터를 읽어들입니다.
- 그 다음에는 일련의 조작을 통해 데이터를 실행합니다.
- 마지막으로, 데이터를 대상으로 보냅니다.

각 조작을 통해 소스에서 마침내 대상(모델 또는 데이터 출력 유형)으로 레코드별로 데이터가 플로우되기 때문에 이러한 작업 시퀀스를 **데이터 스트림**이라고 합니다.

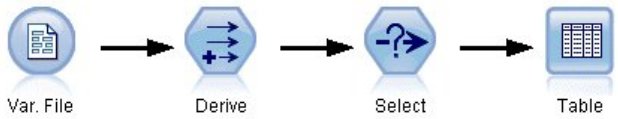


그림 2. 단순 스트림

IBM SPSS Modeler 스트림 캔버스

스트림 캔버스는 IBM SPSS Modeler 창의 가장 큰 영역이며 데이터 스트림을 작성하고 조작하는 위치입니다.

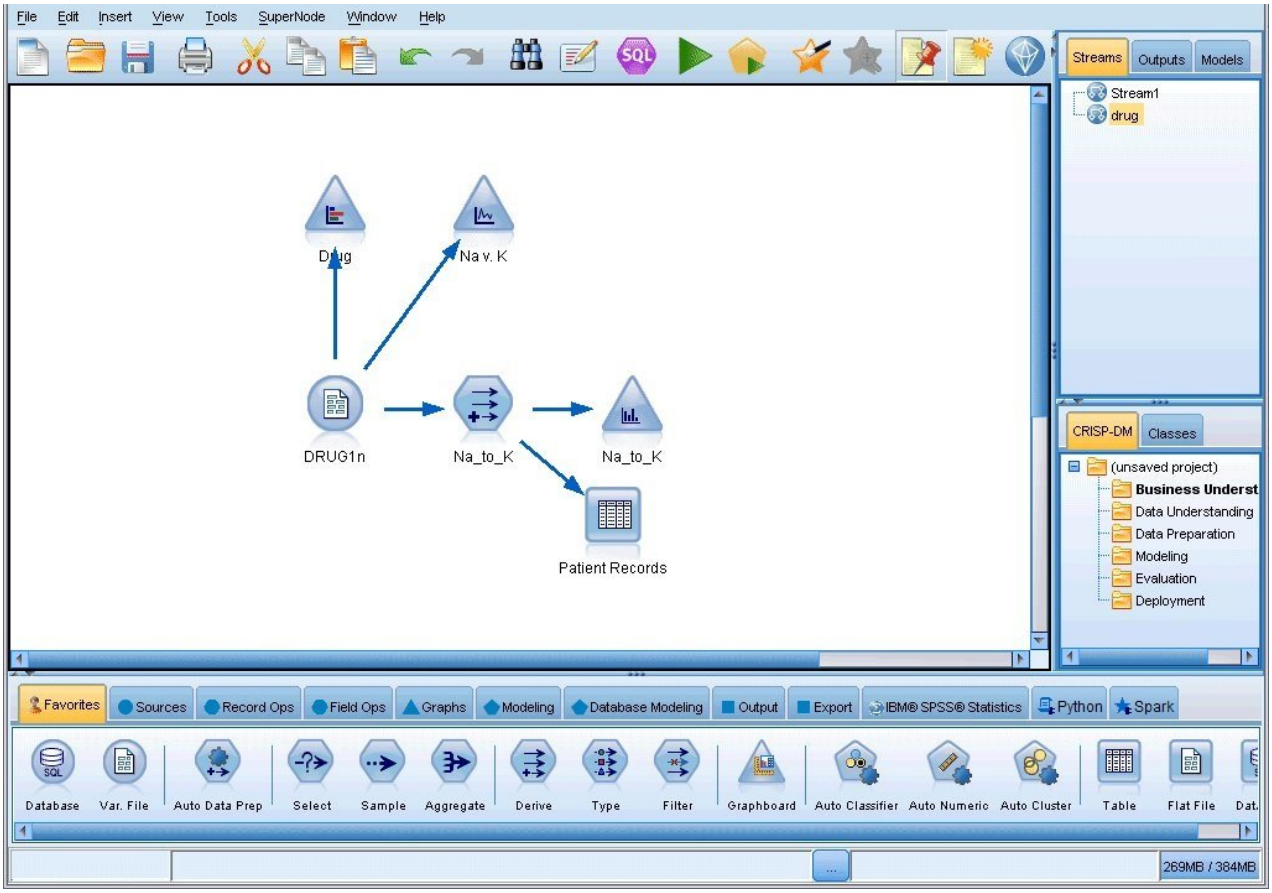


그림 3. *IBM SPSS Modeler* 작업공간(기본 보기)

스트림은 인터페이스의 기본 캔버스에서 비즈니스 관련 데이터 연산 다이어그램을 그려서 작성됩니다. 각 연산은 아이콘 또는 **노드**로 표시되며, 노드는 각 연산을 통한 데이터 플로우를 표시하는 **스트림**에서 함께 연결됩니다.

동일한 스트림 캔버스에서 또는 새 스트림 캔버스를 열어 IBM SPSS Modeler에서 동시에 다중 스트림에 대한 작업을 할 수 있습니다. 세션 동안 스트림은 IBM SPSS Modeler 창의 오른쪽 상단에 있는 스트림 관리자에 저장됩니다.

참고: 기본 제공되는 트랙 패드의 **포스클릭 및 햅틱 피드백** 설정이 사용 가능한 상태에서 MacBook을 사용하는 경우, 노드 팔레트에서 스트림 캔버스로 노드를 끌어다 놓으면 중복 노드가 캔버스에 추가됩니다. 이 문제를 방지하려면 **포스클릭 및 햅틱 피드백** 트랙 패드 시스템 환경 설정을 사용할 수 없음으로 지정하도록 권장합니다.

노드 팔레트

SPSS Modeler에서 대다수의 데이터와 모델링 도구는 스트림 캔버스 아래의 창 아래에 걸쳐서 노드 팔레트에서 사용할 수 있습니다.

예를 들어, **레코드 작업** 팔레트 탭에는 선택, 병합 및 붙여쓰기 등과 같이 데이터 레코드에서 작업을 수행하는 데 사용할 수 있는 노드가 포함됩니다.

노드를 캔버스에 추가하려면 노드 팔레트에서 아이콘을 두 번 클릭하거나 이를 캔버스로 끄십시오. 그런 다음 이들을 연결하여 데이터 플로우를 나타내는 스트림을 작성할 수 있습니다.

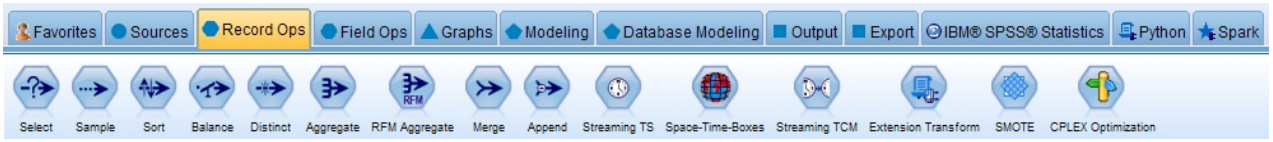


그림 *4.* 노드 팔레트에서 레코드 작업 탭

각 팔레트 탭에는 다음과 같이 스트림 작업의 서로 다른 단계에 사용하는 관련 노드의 콜렉션이 포함됩니다.

- **소스** 노드는 데이터를 SPSS Modeler로 가져옵니다.
- **레코드 작업** 노드는 선택, 병합 및 붙여쓰기 등의 데이터 레코드 작업을 수행합니다.
- **필드 작업** 노드는 필터링, 새 필드 파생 및 주어진 필드의 측정 수준 판별 등의 데이터 필드 작업을 수행합니다.
- **그래프** 노드는 모델링 전후 데이터를 그래프로 표시합니다. 그래프에는 도표, 히스토그램, 웹 노드 및 평가 차트를 포함합니다.
- **모델링** 노드는 신경망, 의사결정 트리, 군집 알고리즘 및 데이터 순차규칙 등과 같이 SPSS Modeler에서 사용 가능한 모델링 알고리즘을 사용합니다.
- **데이터베이스 모델링** 노드는 Microsoft SQL Server, IBM Db2 및 Oracle과 Netezza 데이터베이스에서 사용 가능한 모델링 알고리즘을 사용합니다.
- **출력** 노드는 SPSS Modeler에서 볼 수 있는 데이터, 차트 및 모델 결과의 다양한 출력을 생성합니다.

- **내보내기** 노드는 IBM SPSS Data Collection 또는 Excel과 같은 외부 애플리케이션에서 볼 수 있는 다양한 출력을 생성합니다.
- **IBM SPSS Statistics** 노드는 IBM SPSS Statistics에서 데이터를 가져오거나 내보낼 뿐만 아니라 IBM SPSS Statistics 프로시저를 실행합니다.
- **Python** 노드는 Python 알고리즘을 실행하는 데 사용할 수 있습니다.
- **Spark** 노드는 Spark 알고리즘을 실행하는 데 사용할 수 있습니다.

SPSS Modeler과 친숙해짐에 따라 원하는 대로 팔레트 내용을 사용자 정의할 수 있습니다.

노드 팔레트의 왼쪽에서는 감독, 연관 또는 세분화를 선택하여 표시되는 노드를 필터링할 수 있습니다.

노드 팔레트 아래에 있는 보고서 분할창은 데이터를 데이터 스트림으로 읽어 올 때 등과 같이 다양한 작업의 진행률에 대한 피드백을 제공합니다. 마찬가지로 노드 팔레트 아래에 있는 상태 분할창은 애플리케이션이 현재 하는 작업뿐만 아니라 사용자 피드백이 필요한 시기 표시에 대한 정보를 제공합니다.

참고: 기본 제공되는 트랙 패드의 **포스클릭 및 햅틱 피드백** 설정이 사용 가능한 상태에서 MacBook을 사용하는 경우, 노드 팔레트에서 스트림 캔버스로 노드를 끌어다 놓으면 중복 노드가 캔버스에 추가됩니다. 이 문제를 방지하려면 **포스클릭 및 햅틱 피드백** 트랙 패드 시스템 환경 설정을 사용할 수 없음으로 지정하도록 권장합니다.

IBM SPSS Modeler 관리자

창의 오른쪽 상단은 관리자 분할창입니다. 여기에는 스트림, 출력, 모델을 관리하는 데 사용하는 세 개의 탭이 있습니다.

스트림 탭을 사용하여 세션에 작성된 스트림을 열고 이름을 변경하며, 저장 및 삭제할 수 있습니다.

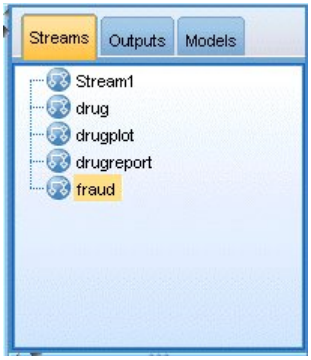


그림 *5.* 스트림 탭

그림 *6.* 출력 탭

출력 탭은 IBM SPSS Modeler에서 스트림 작업을 통해 생성된 그래프, 테이블과 같은 다양한 파일을 포함합니다. 이 탭에 나열된 테이블, 그래프, 보고서를 표시하고 저장하며, 이름을 변경하고 닫을 수 있습니다.

그림 *7.* 모델 너깃이 포함된 모델 탭

모델 탭은 관리자 탭 중 가장 강력합니다. 이 탭은 현재 세션에 대해 IBM SPSS Modeler에 생성된 모델을 포함하는 모든 모델 **너깃**을 포함합니다. 이러한 모델을 모델 탭에서 직접 찾아보거나 캔버스의 스트림에 추가할 수 있습니다.

IBM SPSS Modeler 프로젝트

창의 오른쪽 하단은 데이터 마이닝 **프로젝트**(데이터 마이닝 작업과 관련된 파일 그룹)를 작성하고 관리하는 데 사용하는 프로젝트 분할창입니다. IBM SPSS Modeler에서 사용자가 작성하는 프로젝트를 볼 수 있는 두 가지 방법이 있는데, 클래스 보기와 CRISP-DM 보기입니다.

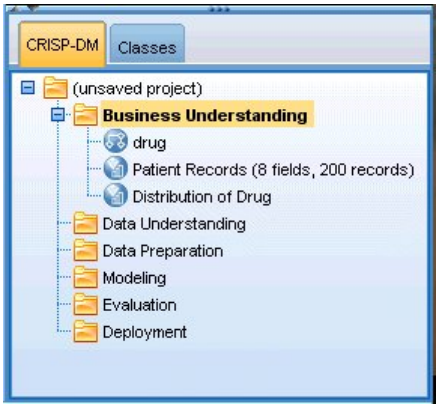


그림 *8. CRISP-DM* 보기

CRISP-DM 탭은 업계에서 입증된 비독점 방법론인 CRISP-DM(Cross-Industry Standard Process for Data Mining)에 따라 프로젝트를 구성하는 방법을 제공합니다. 숙련된 데이터 마이너와 초보 데이터 마이너 모두를 위해 CRISP-DM 도구를 사용하면 데이터 마이닝을 위한 노력을 쉽게 체계화하고 전달할 수 있습니다.

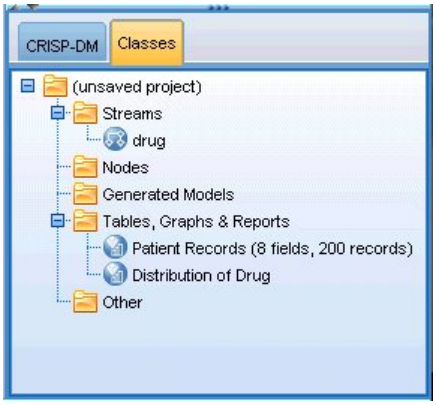


그림 *9.* 클래스 보기

클래스 탭은 IBM SPSS Modeler에서 작업을 범주적으로(작성하는 오브젝트 유형별로) 구성하는 방법을 제공합니다. 이 보기는 데이터, 스트림, 모델 인벤토리를 만들 때 유용합니다.

IBM SPSS Modeler 도구 모음

IBM SPSS Modeler 창의 맨 위에서 다양한 유용한 기능을 제공하는 아이콘 도구 모음을 볼 수 있습니다. 다음은 도구 모음 단추 및 해당 함수입니다.

새 스트림 작성

스트림 열기

스트림 저장

현재 스트림 인쇄

잘라내기 & 클립보드로 이동

클립보드에 복사

선택 붙여넣기

마지막 작업 실행 취소

다시 실행

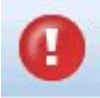
노드 검색

스트림 특성 편집

SQL 생성 미리보기

현재 스트림 실행

스트림 선택 실행



스트림 중지(스트림이 실행 중인 동안에만 활성)

수퍼노드 추가

확대(수퍼노드 전용)

축소(수퍼노드 전용)

스트림에 마크업 없음

주석 삽입

스트림 마크업 숨김(있는 경우)

숨겨진 스트림 마크업 표시

스트림을 IBM SPSS Modeler Advantage에서 열기

스트림 마크업은 스트림 주석, 모델 링크 및 스코어링 분기 표시로 구성됩니다.

모델 링크는 *IBM SPSS* 모델링 노드 안내서에 설명되어 있습니다.

도구 모음 사용자 정의

다음과 같이 도구 모음의 다양한 측면을 변경할 수 있습니다.

- 표시되는지 여부
- 아이콘에 사용 가능한 도구 팁이 있는지 여부
- 큰 아이콘 또는 작은 아이콘을 사용하는지 여부

도구 모음 표시를 켜고 끄기

1. 기본 메뉴에서 다음을 클릭하십시오.

   보기 > **도구 모음** > **표시**

도구 팁이나 아이콘 크기 설정 변경

1. 기본 메뉴에서 다음을 클릭하십시오.

   보기 > **도구 모음** > **사용자 정의**

필요에 따라 **도구 팁 표시** 또는 **단추 크게**를 클릭하십시오.

IBM SPSS Modeler 창 사용자 정의

SPSS Modeler 인터페이스의 다양한 부분 사이에 디바이더를 사용할 때 기본 설정을 충족하도록 도구의 크기를 조정하거나 닫을 수 있습니다. 예를 들어, 대용량 스트림에 대해 작업 중인 경우 각 디바이더에 있는 작은 화살표를 사용하여 노드 팔레트, 관리자 분할창 및 프로젝트 분할창을 닫을 수 있습니다. 이는 스트림 캔버스를 최소화하여 대용량 또는 다량의 스트림을 위한 충분한 작업 공간을 제공합니다.

또는 보기 메뉴에서 **노드 팔레트**, **관리자** 또는 **프로젝트**를 클릭하여 이러한 항목을 켜거나 끌 수도 있습니다.

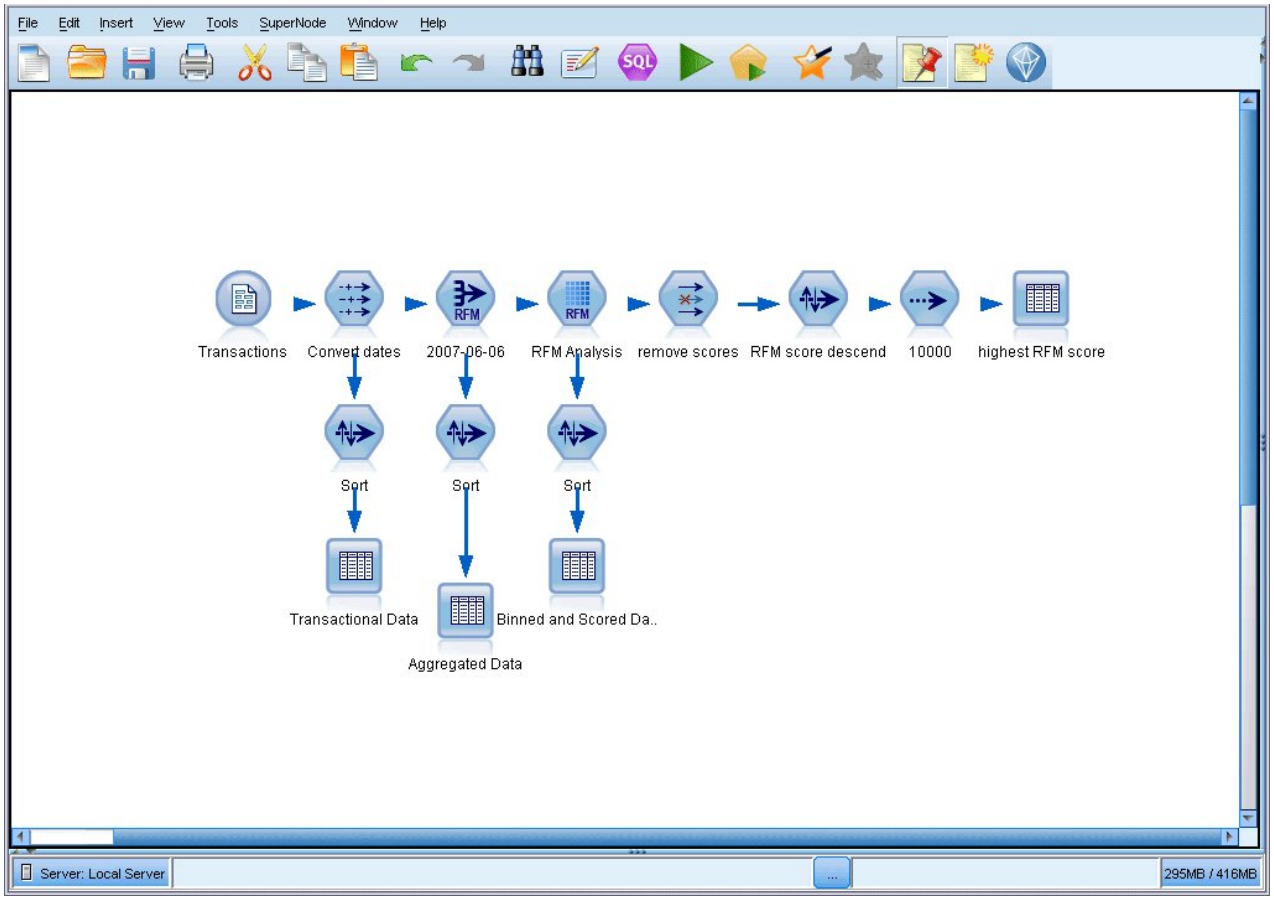


그림 *10.* 최대 스트림 캔버스

노드 팔레트, 관리자 및 프로젝트 분할창을 닫는 대신에 SPSS Modeler 창의 측면과 아래쪽에 있는 스크롤바를 가로 세로로 이동하여 스트림 캔버스를 스크롤 가능한 페이지로 사용할 수 있습니다.

또한 스트림 설명, 모델 링크 및 스코어링 분기 표시로 구성된 화면 마크업 표시를 제어할 수도 있습니다. 이 표시를 켜거나 끄려면 다음을 클릭하십시오.

보기 > **스트림 마크업**

스트림 아이콘 크기 변경

다음 방법으로 스트림 아이콘 크기를 변경할 수 있습니다.

- 스트림 특성 설정을 통해
- 스트림의 팝업 메뉴를 통해
- 키보드 사용

전체 스트림 보기를 표준 아이콘 크기의 8% - 200% 사이의 많은 크기 중 하나로 스케일링할 수 있습니다.

전체 스트림 스케일링 방법(스트림 특성 방법)

1. 메뉴에서 다음을 선택하십시오.

   도구 > **스트림 특성** > **옵션** > **레이아웃**
2. 아이콘 크기 메뉴에서 원하는 크기를 선택하십시오.
3. **적용**을 클릭하여 결과를 확인하십시오.
4. **확인**을 클릭하여 변경사항을 저장하십시오.

전체 스트림 스케일링 방법(메뉴 방법)

1. 캔버스에서 스트림 배경을 마우스 오른쪽 단추로 클릭하십시오.
2. **아이콘 크기**를 선택하고 원하는 크기를 선택하십시오.

전체 스트림 스케일링 방법(키보드 방법)

1. 기본 키보드에서 Ctrl +[-]를 눌러 크기를 축소하십시오.
2. 기본 키보드에서 Ctrl +Shift + [-]를 눌러 크기를 확대하십시오.

이 기능은 복합 스트림에 대한 전체 보기를 확보할 경우 특히 유용합니다. 또한 이를 사용하여 스트림을 인쇄하는 데 필요한 페이지 수를 최소화할 수 있습니다.

IBM SPSS Modeler에서 마우스 사용

IBM SPSS Modeler에서 마우스의 가장 흔한 용도는 다음과 같습니다.

- **단일 클릭.** 마우스 오른쪽이나 왼쪽 단추를 사용하여 메뉴에서 옵션을 선택하고, 팝업 메뉴를 열고, 다양한 기타 표준 제어와 옵션에 액세스할 수 있습니다. 노드를 이동하거나 끌려면 단추를 클릭한 후 계속 누르십시오.
- **두 번 클릭.** 마우스 왼쪽 단추를 두 번 클릭하여 노드를 스트림 캔버스에 놓고 기존 노드를 편집하십시오.
- **가운데 클릭.** 마우스 가운데 단추를 클릭하여 커서를 끌어서 스트림 캔버스에 있는 노드를 연결하십시오. 노드를 연결 해제하려면 마우스 가운데 단추를 두 번 클릭하십시오. 버튼이 3개인 마우스가 아닌 경우에는 마우스를 클릭하고 끄는 중에 Alt 키를 눌러서 이 기능을 시뮬레이션할 수 있습니다.

단축키 사용

IBM SPSS Modeler의 많은 비주얼 프로그래밍 작업에는 연관된 단축키가 있습니다. 예를 들어, 노드를 클릭하고 키보드에서 삭제 키를 눌러 노드를 삭제할 수 있습니다. 마찬가지로, Ctrl 키를 누른 상태에서 S 키를 눌러서 스트림을 빠르게 저장할 수 있습니다. 이와 같은 제어 명령은 Ctrl 및 다른 키 조합(예: Ctrl+S)으로 표시됩니다.

표준 Windows 작업에 사용하는 많은 단축키가 있습니다(예: 잘라내기는 Ctrl+X). 이러한 단축키는 다음 애플리케이션 특정 단축키와 함께 IBM SPSS Modeler에서 지원됩니다.

참고: 어떤 경우에는 IBM SPSS Modeler에서 사용한 이전 단축키가 표준 Windows 단축키와 충돌할 수 있습니다. 이러한 이전 단축키는 Alt 키의 추가와 함께 지원됩니다. 예를 들어, Ctrl+Alt+C는 캐시를 켜기/끄기를 토글하는 데 사용할 수 있습니다.

표 *1.* 지원되는 단축키

단축키	기능
Ctrl+A	모두 선택
Ctrl+X	잘라내기
Ctrl+N	새 스트림
Ctrl+O	스트림 열기
Ctrl+P	인쇄
Ctrl+C	복사
Ctrl+V	붙여넣기
Ctrl+Z	실행 취소
Ctrl+Q	선택한 노드의 모든 노드 다운스트림 선택
Ctrl+W	모든 다운스트림 노드 선택 취소(Ctrl+Q로 토글)
Ctrl+E	선택한 노드에서 실행
Ctrl+S	현재 스트림 저장
Alt+화살표 키	선택한 노드를 캔버스 스트림에서 사용된 화살표의 방향으로 이동
Shift+F10	선택한 노드의 팝업 메뉴 열기

표 *2.* 이전 단축키의 지원되는 단축키

단축키	기능
Ctrl+Alt+D	노드 중복
Ctrl+Alt+L	노드 로드
Ctrl+Alt+R	노드 이름 변경
Ctrl+Alt+U	사용자 입력 노드 작성
Ctrl+Alt+C	캐시 켜기/끄기 토글
Ctrl+Alt+F	캐시 비우기
Ctrl+Alt+X	수퍼노드 확장
Ctrl+Alt+Z	확대/축소
Delete	노드나 연결 삭제

인쇄

다음 오브젝트를 IBM SPSS Modeler에서 인쇄할 수 있습니다.

- 스트림 다이어그램
- 그래프
- 테이블
- 보고서(보고서 노드 및 프로젝트 보고서에서)
- 스크립트(스트림 특성, 독립형 스크립트 또는 수퍼노드 스크립트 대화 상자에서)

- 모델(모델 브라우저, 현재 초점이 있는 대화 상자 탭, 트리 뷰어)
- 주석(출력을 위해 주석 탭 사용)

오브젝트 인쇄

- 미리보지 않고 인쇄하려면 도구 모음에서 인쇄 단추를 클릭하십시오.
- 인쇄하기 전에 페이지를 설정하려면 파일 메뉴에서 **페이지 설정**을 선택하십시오.
- 인쇄 전에 미리보려면 파일 메뉴에서 **인쇄 미리보기**를 선택하십시오.
- 프린터 선택을 위한 옵션이 있는 표준 인쇄 대화 상자를 보려면 파일 메뉴에서 **인쇄**를 선택하십시오.

IBM SPSS Modeler 자동화

고급 데이터 마이닝은 복잡하고 때로는 시간이 긴 프로세스일 수 있으므로 IBM SPSS Modeler에는 몇몇 코딩 유형 및 자동화 지원이 포함됩니다.

- **CLEM**(Control Language for Expression Manipulation)은 IBM SPSS Modeler 스트림과 함께 플로우하는 데이터를 분석하고 조작하기 위한 언어입니다. 데이터 마이너는 비용과 수입 데이터에서 수익을 파생하는 것과 같은 단순한 작업 또는 웹 로그 데이터를 유용한 정보가 있는 필드와 레코드 세트로 변환하는 것과 같은 복잡한 작업을 수행하기 위해 스트림 작업에서 CLEM을 집중적으로 사용합니다.
- **스크립팅**은 사용자 인터페이스에서 프로세스를 자동화하기 위한 강력한 도구입니다. 스크립트는 사용자가 마우스나 키보드를 사용하여 수행하는 것과 동일한 유형의 조치를 수행합니다. 또는 출력을 지정하고 생성된 모델을 조작할 수도 있습니다.

제 3 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.

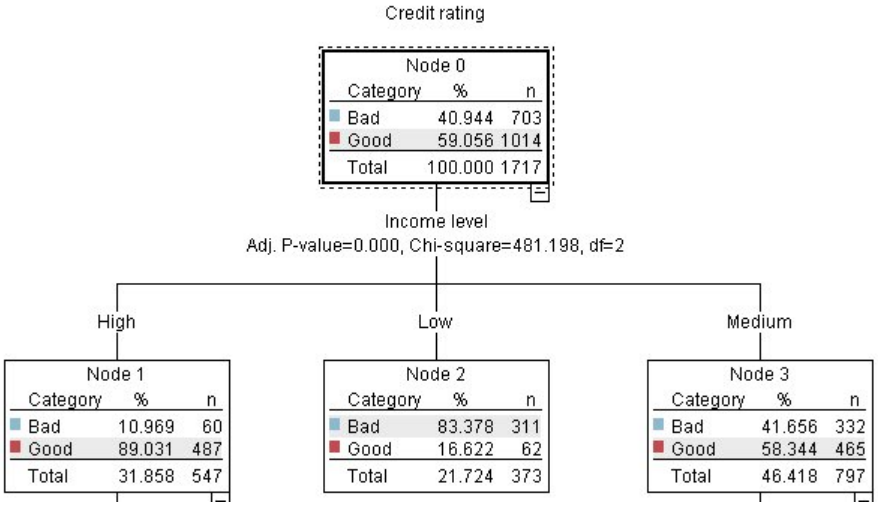


그림 *11.* 단순 의사결정 트리 모델

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

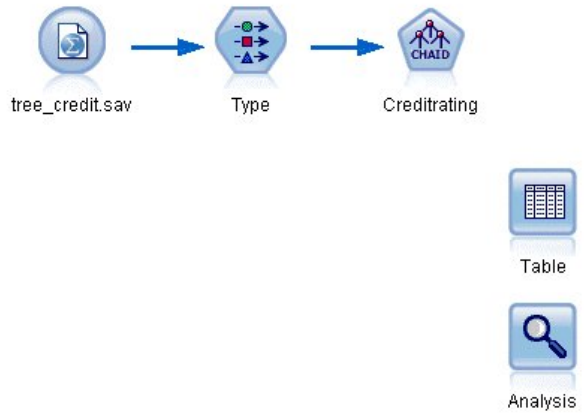


그림 *12.* 스트림 모델링

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 *13.* 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.

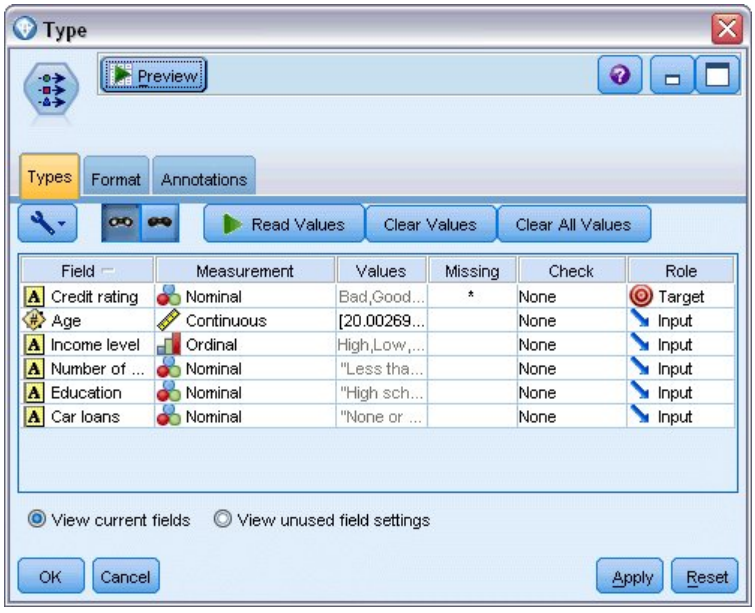


그림 *14.* 유형 노드에서 목표 및 입력 필드 설정

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 대상 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.

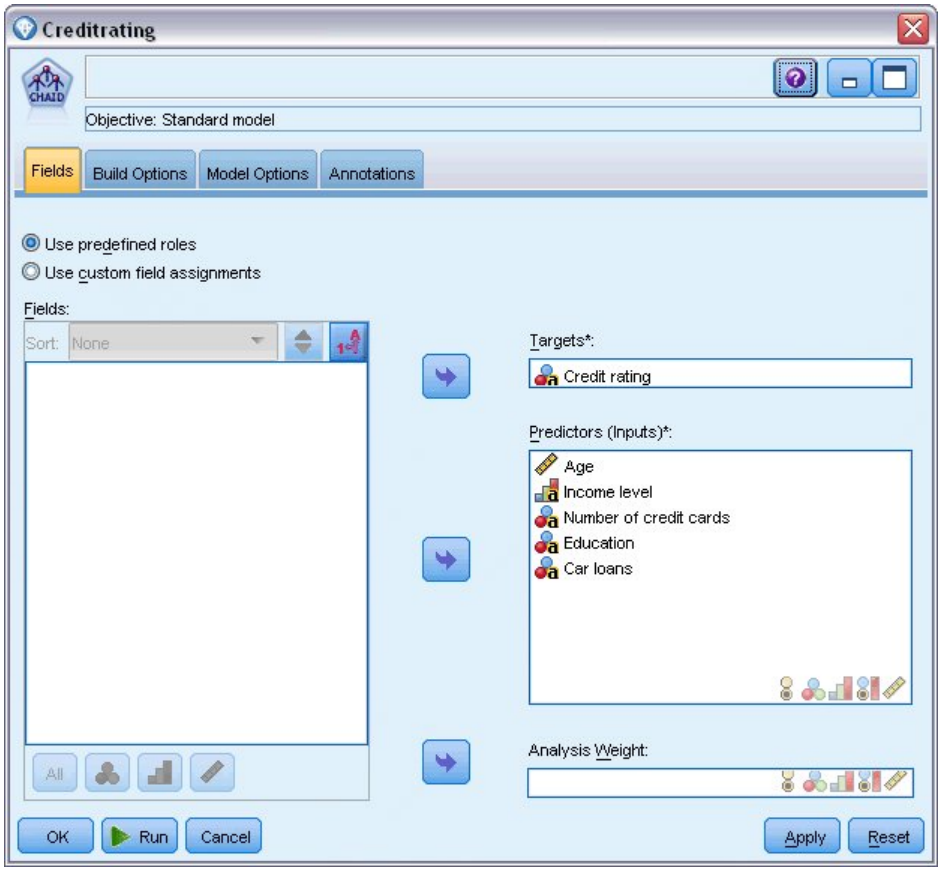


그림 *15. CHAID* 모델링 노드, 필드 탭

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.

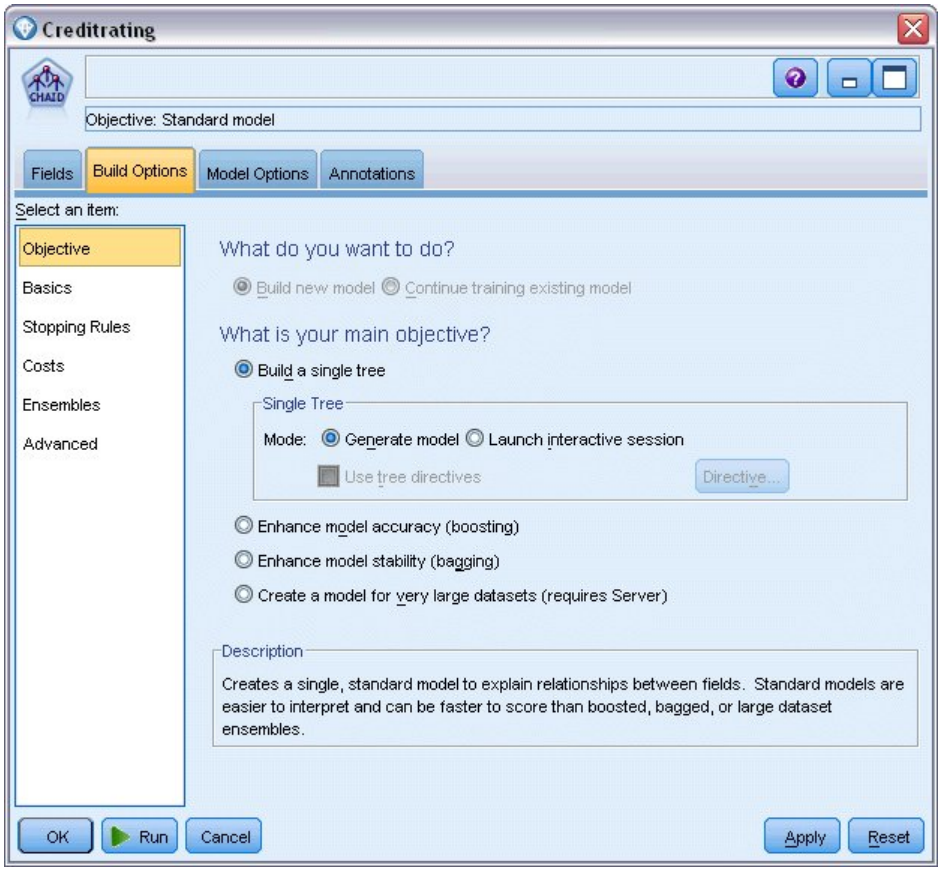


그림 *16. CHAID* 모델링 노드, 작성 옵션 탭

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. **상위 분기 최소 레코드 수**를 400으로 설정하십시오.
5. **하위 분기 최소 레코드 수**를 200으로 설정하십시오.

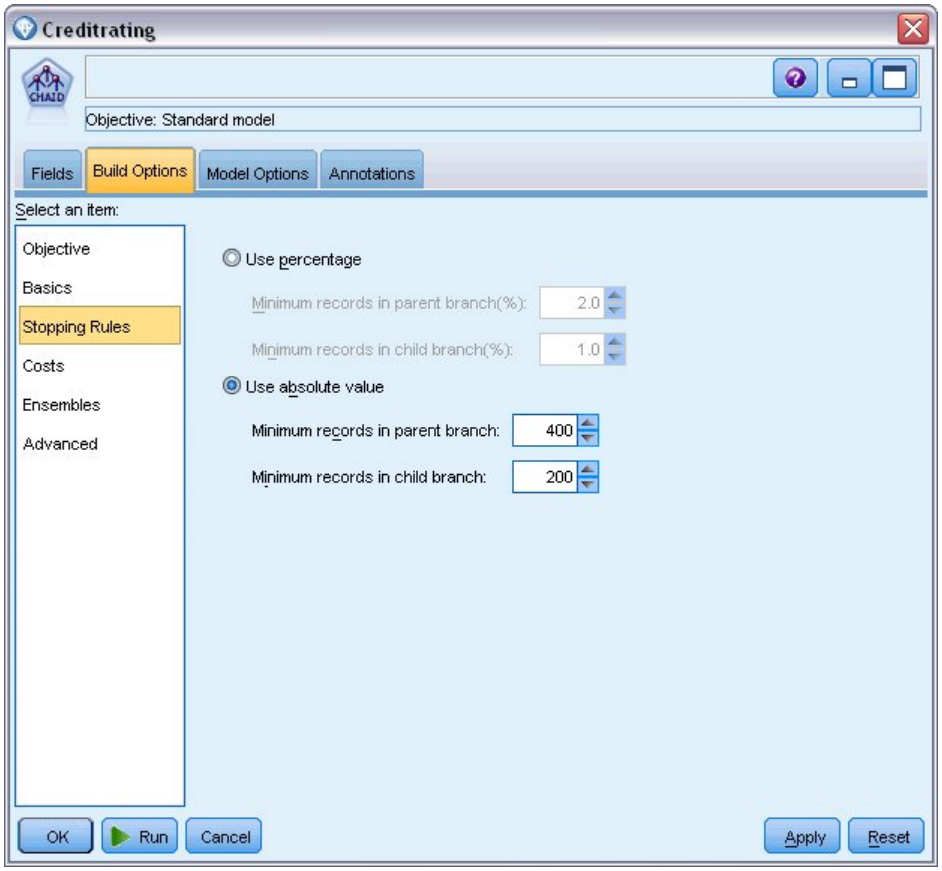


그림 *17.* 의사결정 트리 작성에 대한 중지 기준 설정

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.

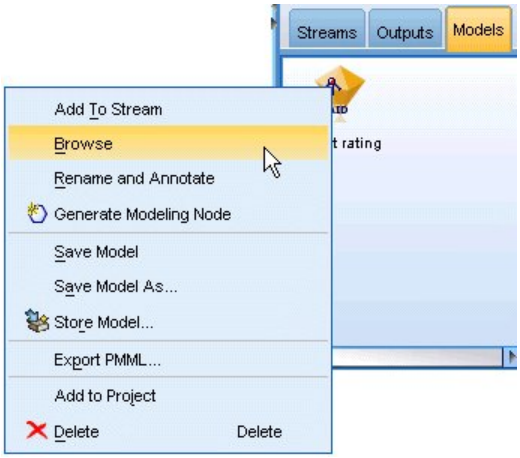


그림 *18.* 모델 팔레트

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.

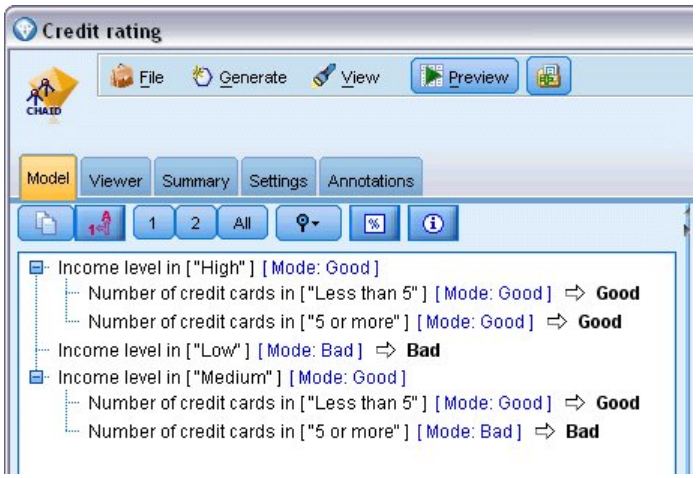


그림 *19. CHAID* 모델 너깃, 규칙 세트

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.

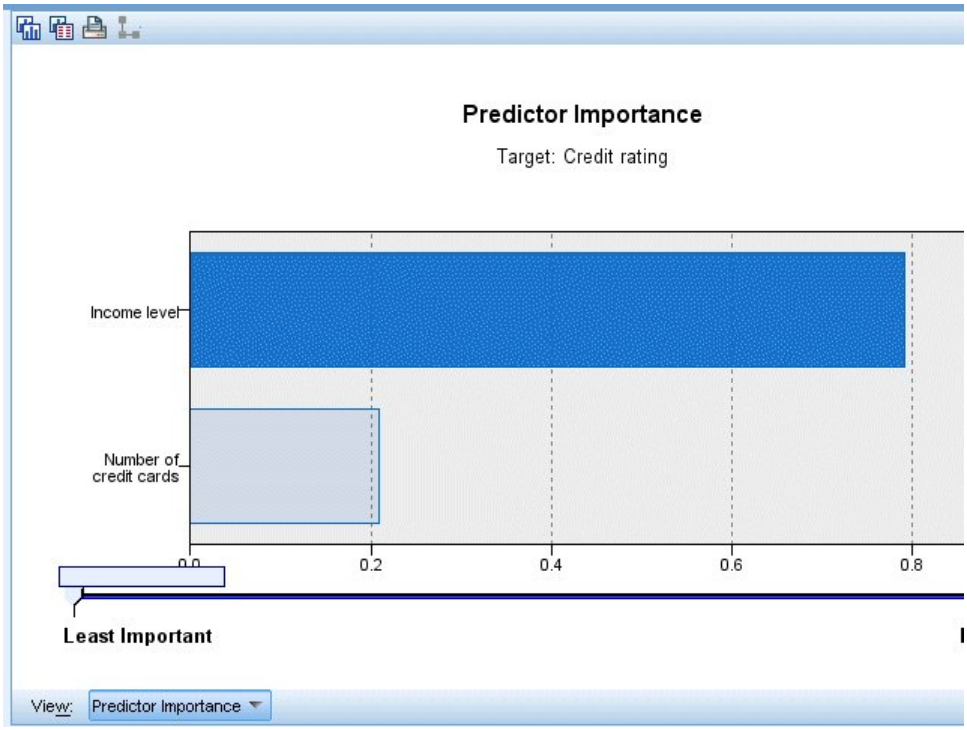


그림 *20.* 예측변수 중요도 차트

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.

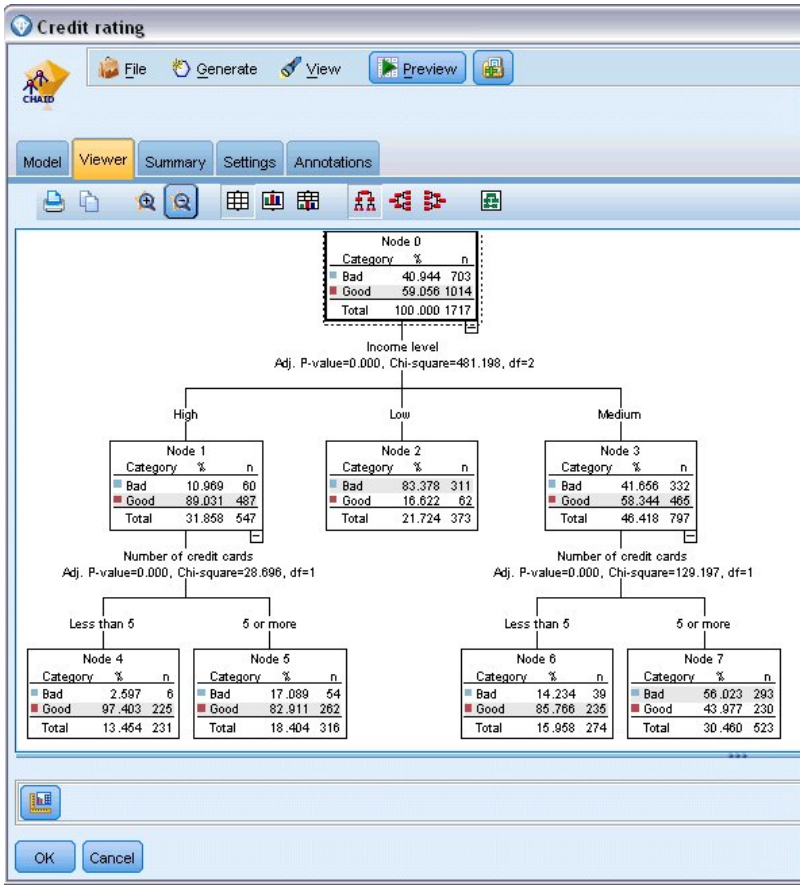


그림 *21.* 축소가 선택된 모델 너깃의 뷰어 탭

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.

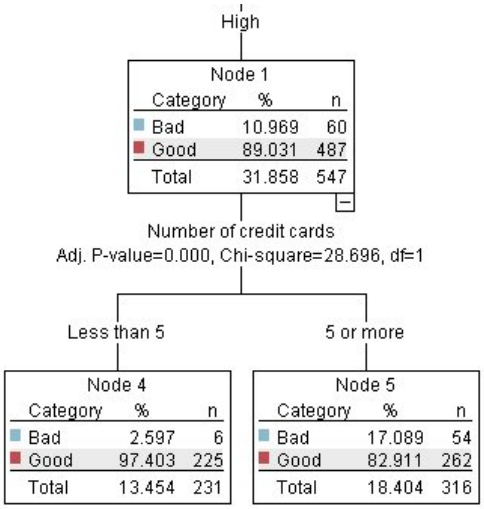


그림 22. 소득이 높은 고객의 트리 보기

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.

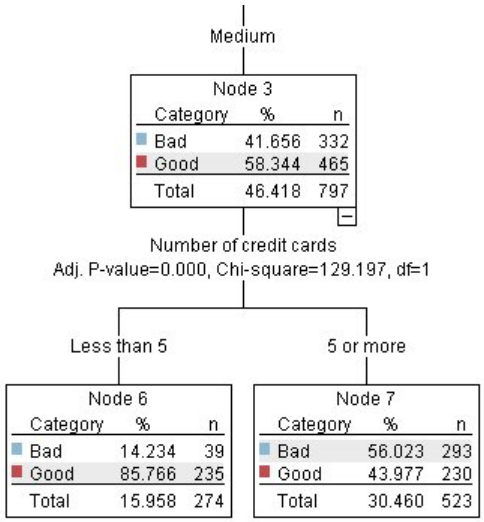


그림 23. 중간 소득 고객의 트리 보기

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.

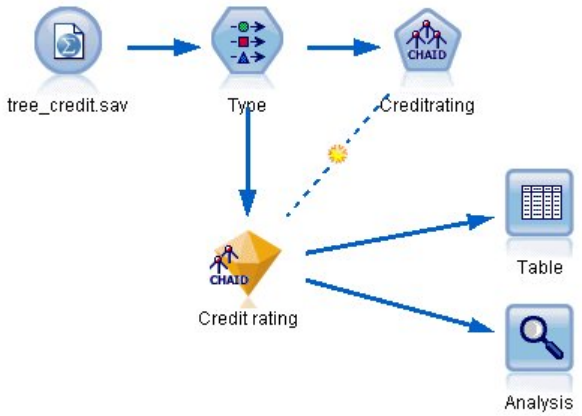
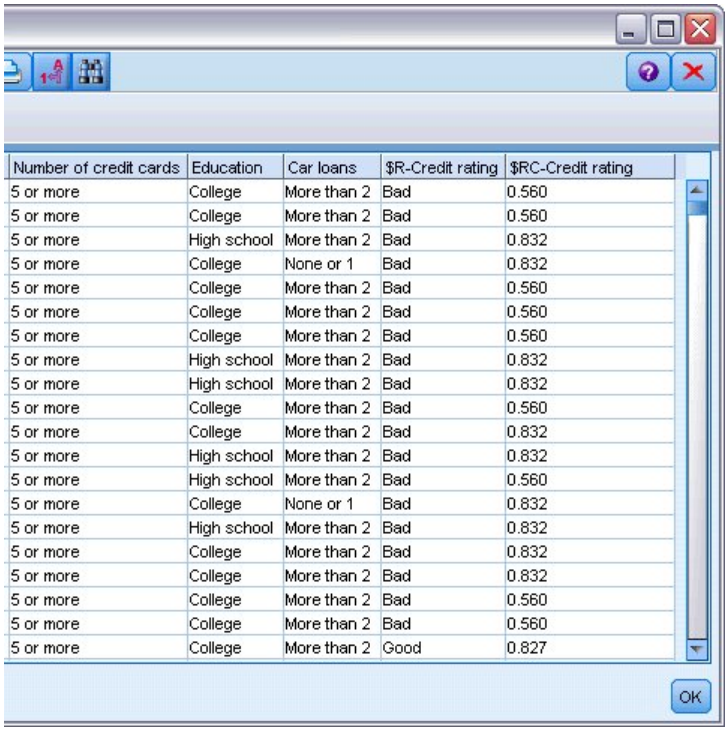


그림 *24.* 모델 평가를 위해 출력 노드에서 모델 너깃 첨부

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 대상 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 대상 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 *25.* 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.

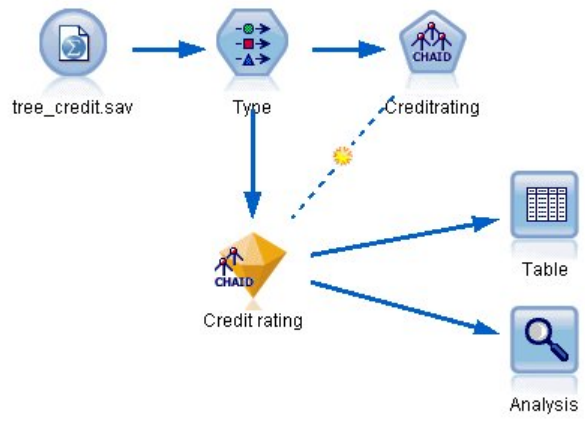


그림 *26.* 분석 노드 첨부

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

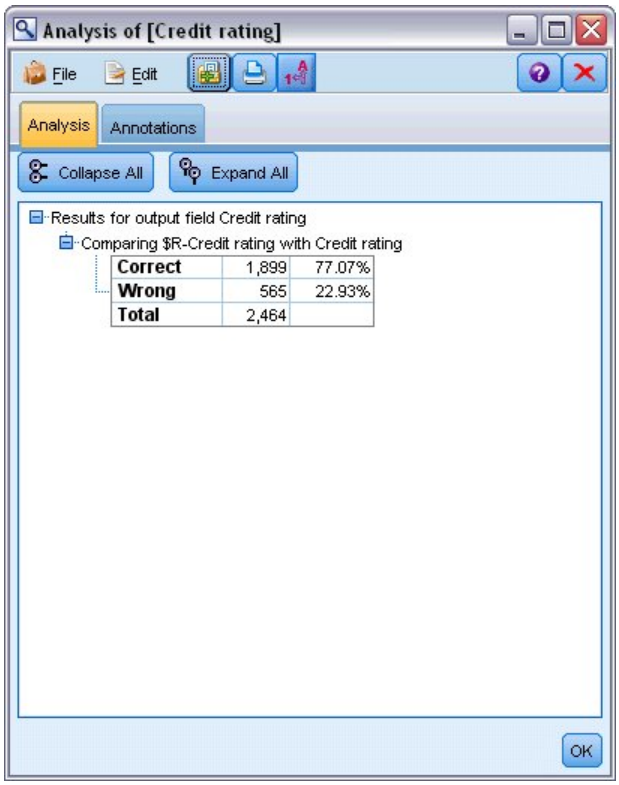


그림 *27.* 관측 및 예측 반응을 비교한 분석 결과

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 대상 필드(결과를 아는 스터디 레코드)를 포함하는 모델링의 목표입니다.

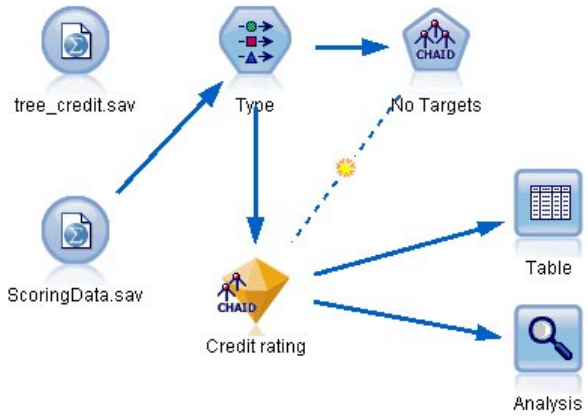


그림 *28.* 스코어링에 대한 새 데이터 첨부

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(대상 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 4 장 플래그 대상에 대한 자동화된 모델링

모델링 고객 반응(자동 분류자)

자동 분류자 노드를 사용하면 플래그(지정된 고객에게 대출이 자동 설정되는지 여부 또는 특정 오퍼에 대한 반응 여부 등) 또는 명목형(변수군) 대상에 대해 수많은 다른 모델을 자동으로 작성하고 비교할 수 있습니다. 이 예에서는 플래그(예 또는 아니오) 결과를 검색할 것입니다. 상대적으로 단순한 스트림 내에서 노드가 후보 모델 세트를 생성하고 순위를 매기며 가장 잘 수행하는 모델들을 선택하고 이러한 모델들을 단일 통합(앙상블) 모델로 결합합니다. 이 방법은 자동화의 간편함과 어느 하나의 모델에서 얻을 수 있는 것보다 더 정확한 예측을 얻을 수 있는 다중 모델 결합의 혜택을 조합합니다.

이 예는 현재 각 고객에 올바른 오퍼를 매치하여 보다 수익성이 좋은 결과를 산출하고자 하는 금융 회사를 기반으로 합니다.

이 방법은 자동화의 혜택을 강조합니다. 연속형(숫자 범위) 대상을 사용하는 유사한 예의 경우, 특성 값(자동 숫자)을 참조하십시오.

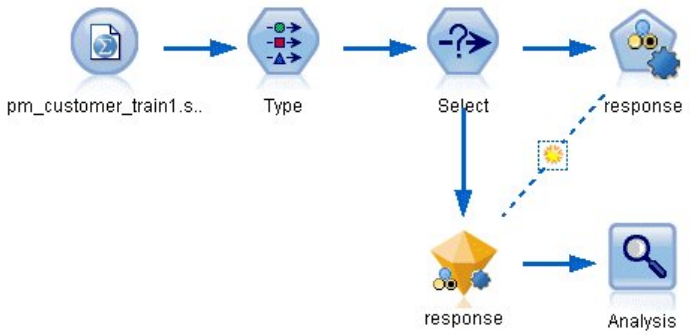


그림 *29.* 자동 분류자 샘플 스트림

이 예에서는 *streams* 아래의 Demo 폴더에 설치된 *pm_binaryclassifier.str* 스트림을 사용합니다. 사용된 데이터 파일은 *pm_customer_train1.sav*입니다. 자세한 정보는 『히스토리 데이터』의 내용을 참조하십시오.

히스토리 데이터

pm_customer_train1.sav 파일에는 *campaign* 필드의 값에 의해 표시되는 대로 과거 캠페인에서 특정 고객에 대해 수행된 오퍼를 추적하는 히스토리 데이터가 있습니다. 가장 많은 수의 레코드가 프리미엄 계정 캠페인에 해당됩니다.

캠페인 필드의 값은 실제로 데이터에서 정수로 코딩됩니다. 예를 들어, *2* = 프리미엄 계정입니다. 나중에 이러한 값에 더 유의미한 출력을 제공할 수 있는 레이블을 정의할 수 있습니다.

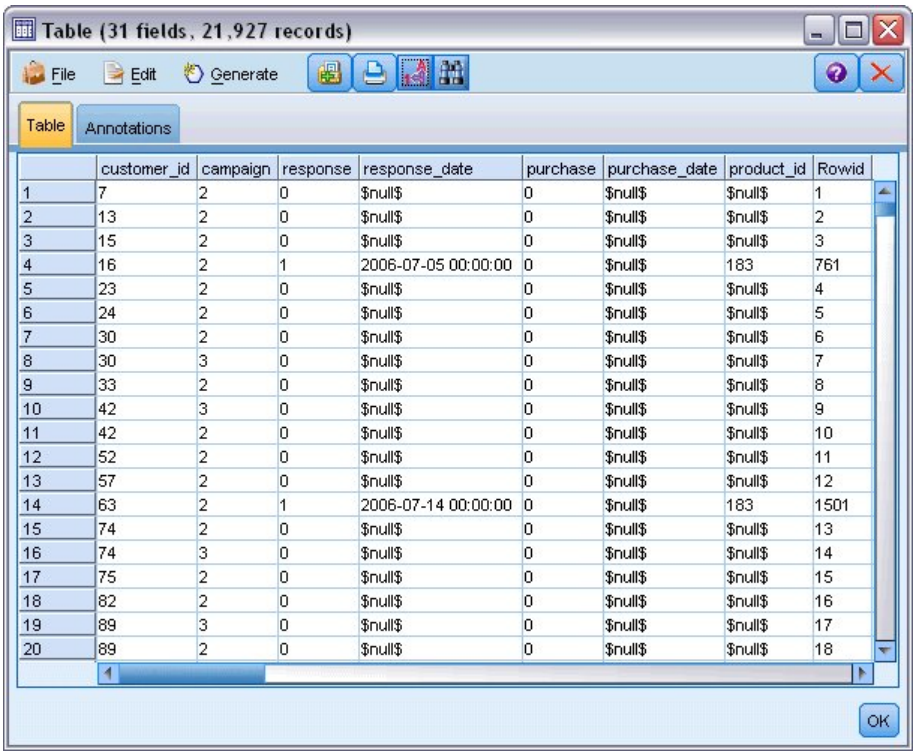


Table (31 fields, 21,927 records)

File Edit Generate

Table Annotations

	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

OK

그림 *30.* 이전 프로모션에 관한 데이터

파일에는 오퍼가 수락되었는지 여부(0 = 아니오, 1 = 예)를 표시하는 반응 필드도 포함됩니다. 이는 예측할 **대상 필드** 또는 값이 됩니다. 각 고객에 대한 인구 통계학 및 금융 정보를 포함하는 수많은 필드도 포함됩니다. 이러한 필드는 수입, 연령 또는 월별 트랜잭션 수와 같은 공정특성 변수를 기준으로 하여 개인 또는 그룹에 대한 반응률을 예측하는 모델을 작성하거나 "교육"하는 데 사용될 수 있습니다.

스트림 작성

1. 사용자의 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *pm_customer_train1.sav*를 가리키는 Statistics 파일 소스 노드를 추가하십시오. 이 폴더를 참조하기 위한 단축키로 파일 경로에서 `$CLEO_DEMOS/`를 지정할 수 있습니다. 표시된 대로 경로에 백슬래시가 아니라 슬래시가 사용되어야 합니다.

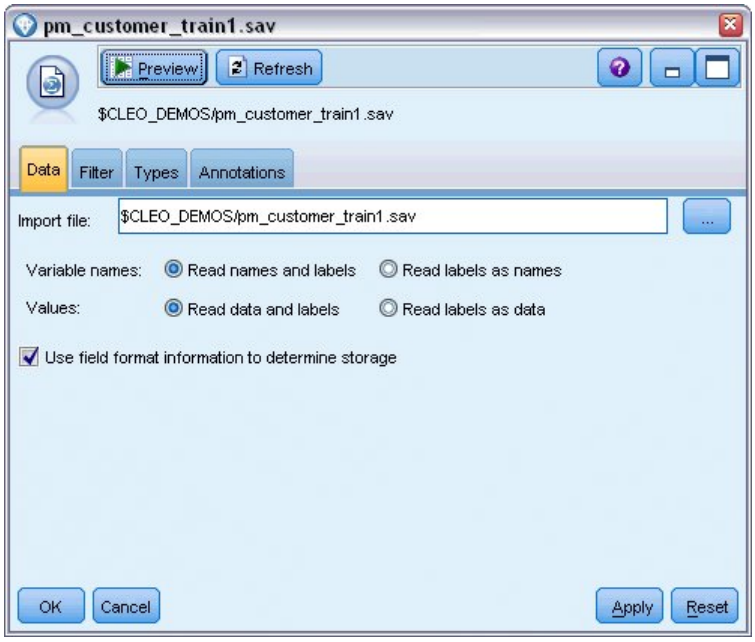


그림 *31.* 데이터에서 읽기

2. 유형 노드를 추가하고 대상 필드로 *response*를 선택하십시오(역할 = **대상**). 이 필드에 대한 측정을 **플래그**로 설정하십시오.

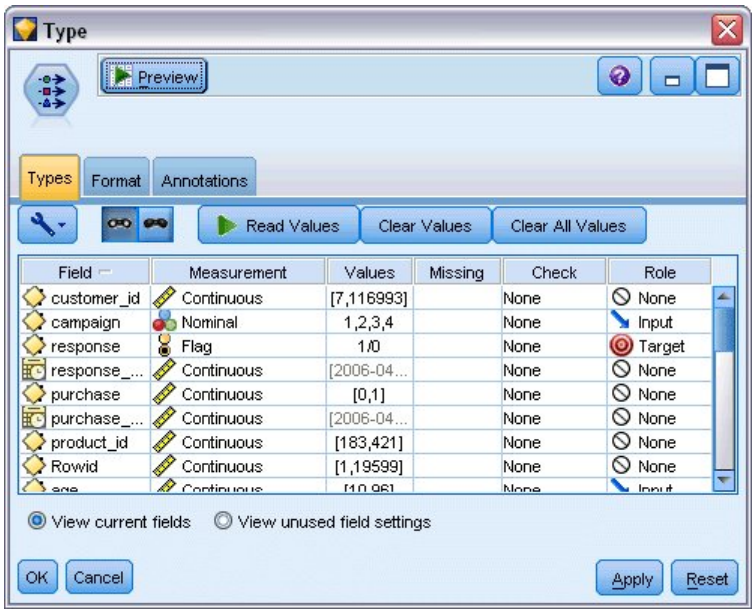


그림 *32.* 측정 수준 및 역할 설정

3. *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* 및 *X_random* 필드에 대한 역할을 **없음**으로 설정하십시오. 모델을 작성하고 있을 때 이러한 필드는 무시됩니다.
4. 유형 노드에서 **값 읽기** 단추를 클릭하여 값이 인스턴스화되도록 하십시오.

앞에서 본대로 소스 데이터에 네 가지 다른 캠페인에 대한 정보가 포함되며 각각은 다른 유형의 고객 계정을 대상으로 합니다. 이러한 캠페인은 각 정수가 어떤 계정 유형을 나타내는지 쉽게 기억할 수 있으므로 데이터에서 정수로 인코딩됩니다. 이제 각각에 대한 레이블을 정의할 것입니다.

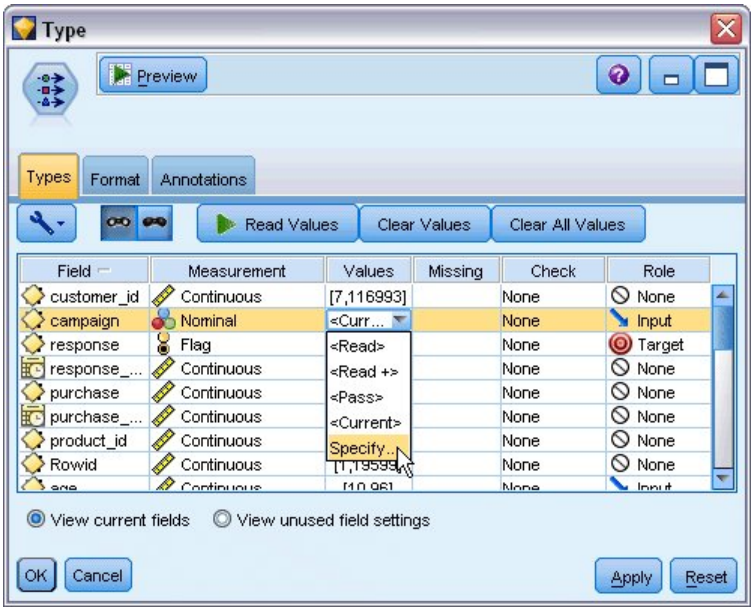


그림 *33.* 필드에 대한 값을 지정하도록 선택

5. **캠페인** 필드에 대한 행에서 **값** 열의 항목을 클릭하십시오.
6. 드롭 다운 목록에서 **지정**을 선택하십시오.

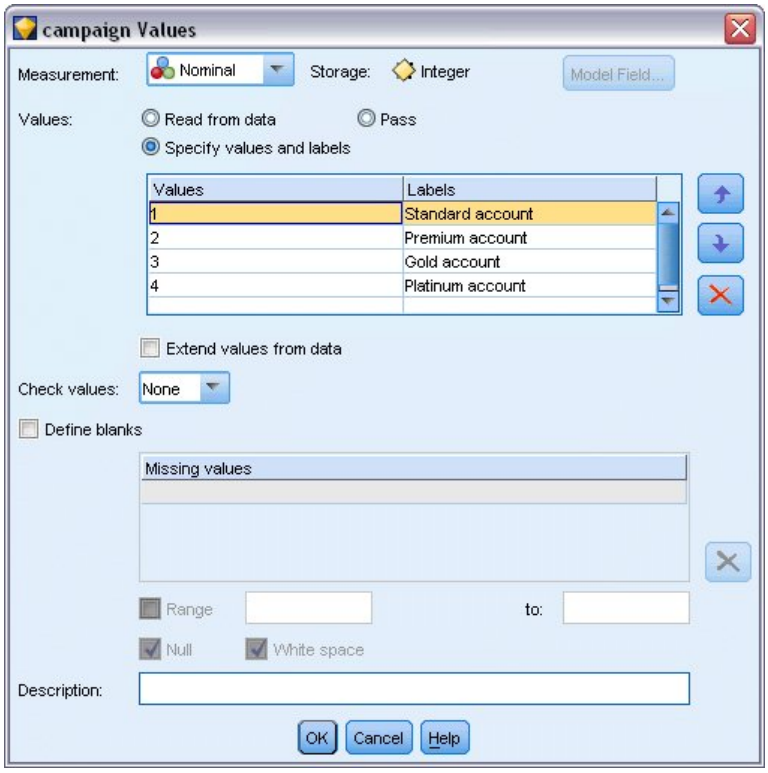
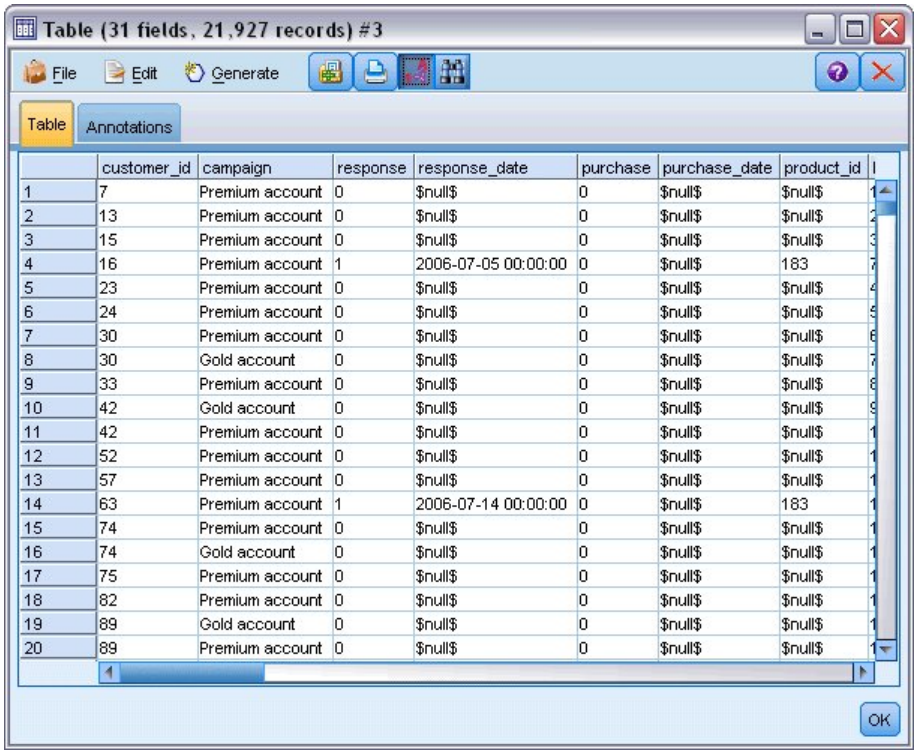


그림 *34.* 필드 값에 대한 레이블 정의

7. **레이블** 열에서 **캠페인** 필드의 각 네 값에 대해 표시될 레이블을 입력하십시오.
8. **확인**을 클릭하십시오.

이제 출력 창에 정수 대신 레이블을 표시할 수 있습니다.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold account	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold account	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold account	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold account	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

그림 *35.* 필드 값 레이블 표시

9. 테이블 노드를 유형 노드에 연결하십시오.
10. 테이블 노드를 열고 **실행**을 클릭하십시오.
11. 출력 창에서 **필드 및 값 레이블 표시** 도구 모음 단추를 클릭하여 레이블을 표시하십시오.
12. **확인**을 클릭하여 출력 창을 닫으십시오.

데이터에 네 가지 다른 캠페인에 대한 정보가 포함되나 한 번에 한 캠페인에 집중하여 분석을 수행할 것입니다. 가장 많은 레코드 수가 프리미엄 계정 캠페인(데이터에서는 *campaign = 2*로 코딩됨)에 해당되므로 선택 노드를 사용하여 이러한 레코드만 스트림에 포함시킬 수 있습니다.

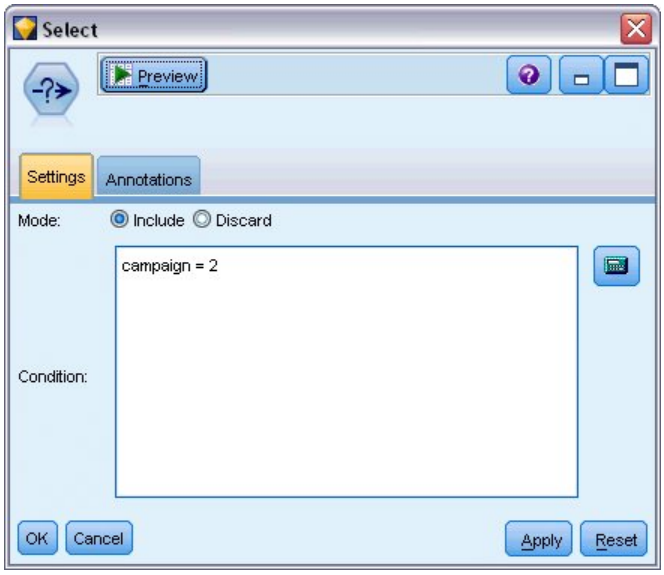


그림 *36.* 단일 캠페인용 레코드 선택

모델 생성 및 비교

1. 자동 분류자 노드를 연결하고 모델 순위 매기기에 사용할 메트릭으로 **전체 정확도**를 선택하십시오.
2. **사용할 모델 수**를 3으로 설정하십시오. 이는 모드를 실행할 때 세 개의 최적 모델이 작성됨을 의미합니다.

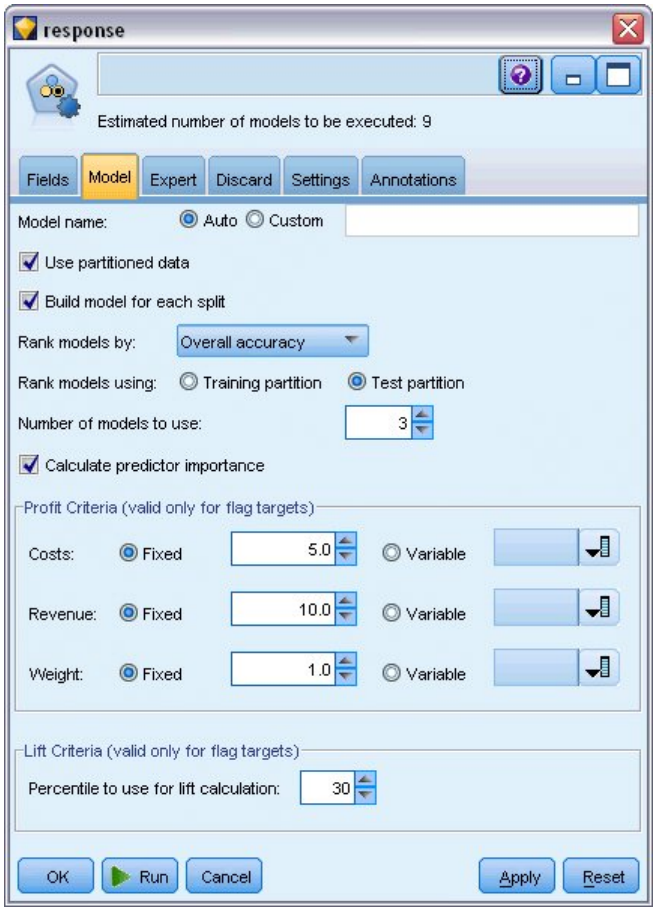


그림 *37.* 자동 분류자 노드 모델 탭

전문가 탭에서 최대 11개까지 다른 모델 알고리즘을 선택할 수 있습니다.

3. **판별분석** 및 **SVM** 모델 유형을 선택 취소하십시오. (이러한 모델은 해당 데이터를 학습하는 데 시간이 많이 걸리므로 선택 취소하여 예의 속도를 높일 수 있습니다. 기다릴 수 있으면 선택한 상태로 두어도 좋습니다.)

   모델 탭에서 **사용할 모델 수**를 3으로 설정했으므로 노드가 나머지 아홉 가지 알고리즘의 정확도를 계산하여 세 개의 최고의 정확도를 포함하는 단일 모델 너깃을 작성합니다.

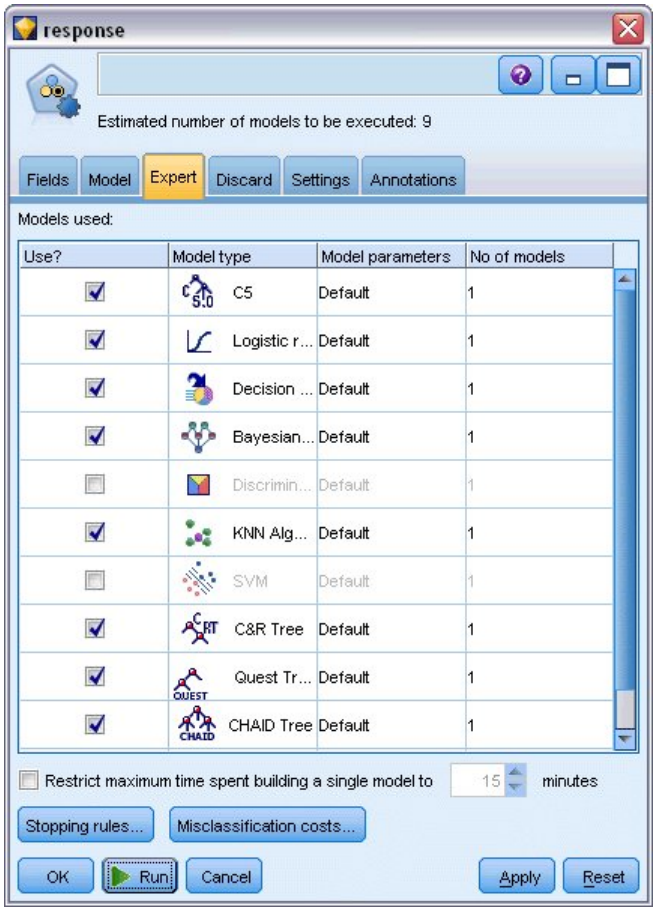


그림 *38.* 자동 분류자 노드 전문가 탭

4. 앙상블 방법의 경우, 설정 탭에서 **신뢰 가중 투표**를 선택하십시오. 그러면 각 레코드에 대해 하나로 통합된 스코어가 생성되는 방법이 결정됩니다.

   단순 투표를 사용하는 경우, 세 모델 중 두 모델이 예를 예측하면 예가 2 대 1로 이깁니다. 신뢰 가중 투표의 경우, 각 예측의 신뢰도를 기반으로 하여 투표에 가중치가 적용됩니다. 따라서 한 모델이 두 예 예측이 결합된 신뢰도보다 더 높은 아니오를 예측하면 아니오가 이깁니다.

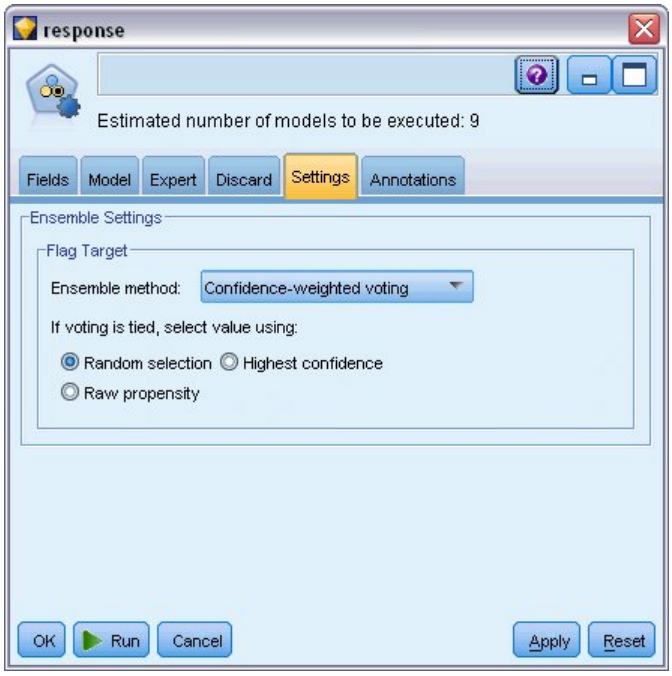


그림 *39.* 자동 분류자 노드*:* 설정 탭

5. **실행**을 클릭하십시오.

몇 분 후에 생성된 모델 너깃이 작성되고 캔버스 및 창의 오른쪽 상단 코너의 모델 팔레트에 배치됩니다. 모델 너깃을 찾아보거나 이를 저장하거나 수많은 방법으로 배치할 수 있습니다.

모델 너깃을 여십시오. 여기에 실행 동안 작성된 각 모델에 대한 세부사항이 나열됩니다. (실제 상황에서는 대형 데이터 세트에 대해 수백 개의 모델이 작성될 수 있으므로 몇 시간이 걸릴 수도 있습니다.) 43 페이지의 그림 29의 내용을 참조하십시오.

임의의 개별 모델을 추가 탐색하려면 **모델** 열에서 모델 너깃 아이콘을 두 번 클릭하여 드릴다운하여 개별 모델 결과를 찾아볼 수 있습니다. 거기서부터 모델링 노드, 모델 너깃 또는 평가 도표를 생성할 수 있습니다. **그래프** 열에서 썸네일을 두 번 클릭하여 전체 크기의 그래프를 생성하십시오.

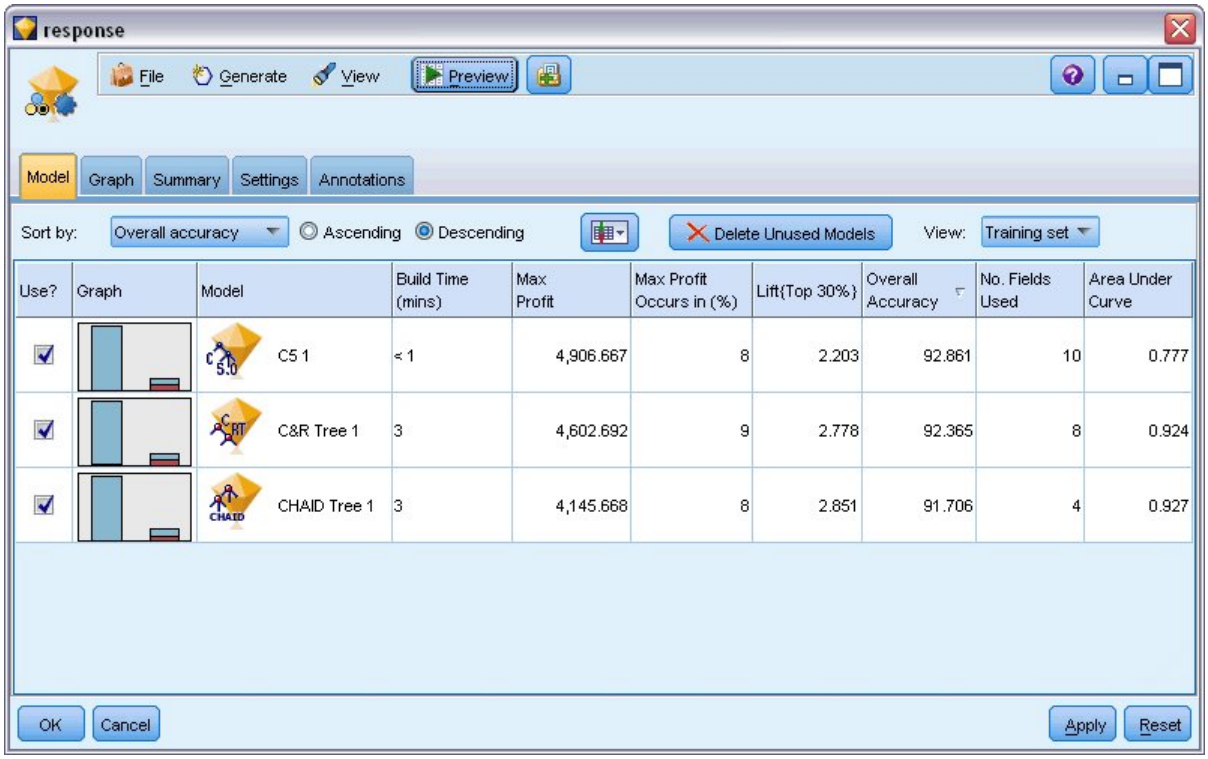


그림 *40.* 자동 분류자 결과

기본적으로 모델은 전체 정확도를 기준으로 하여 정렬됩니다. 사용자가 자동 분류자 노드 모델 탭에서 선택한 측도가 전체 정확도이기 때문입니다. C51 모델이 이 측도에서 최선인 것으로 순위가 매겨졌으나 C&R 트리 및 CHAID 모델은 정확도에서 거의 유사합니다.

해당 열의 헤더를 클릭하여 다른 열 기준으로 정렬하거나 도구 모음의 **정렬기준** 드롭 다운 목록에서 원하는 측도를 선택할 수 있습니다.

이러한 결과를 기준으로 하여 모두 세 개의 가장 정확한 모델을 사용하기로 결정했습니다. 여러 모델의 예측을 결합하면 개별 모델의 제한을 피할 수 있어 전반적인 정확도가 향상됩니다.

사용? 열에서 C51, C&R 트리 및 CHAID 모델을 선택하십시오.

분석 노드(출력 팔레트)를 모델 너깃 뒤에 연결하십시오. 마우스 오른쪽 단추로 분석 노드를 클릭하고 **실행**을 선택하여 스트림을 실행하십시오.

앙상블 모델에 의해 생성된 통합 스코어가 *$XF-response*라는 필드에 표시됩니다. 학습 데이터에 대해 측정될 때 예측값은 (원래 반응 필드에 기록된 대로) 실제 반응과 전체 정확도 92.82%로 일치합니다.

이 케이스에서 세 개별 모델 중 최적 모델(C51의 경우, 92.86%)처럼 정확하지는 않은 반면 유의미하다고 보기에는 차이가 너무 작습니다. 일반적으로 학습 데이터가 아닌 데이터 세트에 적용되는 경우 앙상블 모델이 가장 잘 수행할 확률이 높습니다.

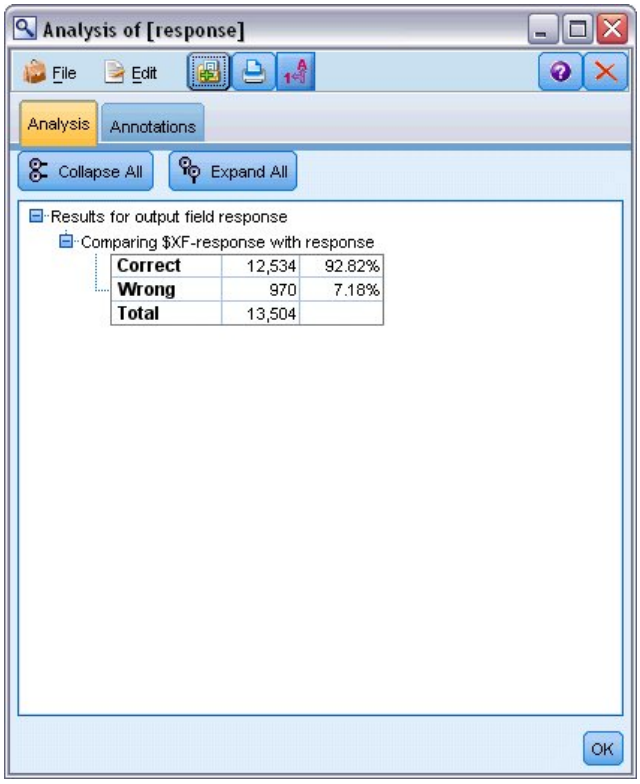


그림 *41.* 세 가지 앙상블 모델의 분석

요약

합계를 계산하기 위해 자동 분류자 노드를 사용하여 다른 모델의 수를 비교하고 세 개의 가장 정확한 모델을 사용하고 앙상블 자동 분류자 모델 너깃 내의 스트림에 이를 추가했습니다.

- 전체 정확도를 기준으로 할 때, 학습 데이터에서 C51, C&R 트리 및 CHAID 모델의 수행이 가장 뛰어났습니다.
- 앙상블 모델은 개별 모델과 거의 동일하게 수행했으며 기타 데이터 세트에 적용될 때는 더 잘 수행할 수도 있습니다. 사용자의 목표가 가능한 한 프로세스를 자동화하는 것인 경우, 이 접근법을 사용하면 특정한 어느 한 모델을 깊이 조사할 필요 없이 대부분의 상황에서 강력한 모델을 얻을 수 있습니다.

제 5 장 연속형 대상에 대한 자동화된 모델링

특성 값(자동 숫자)

자동 숫자 노드를 사용하면 다양한 모델의 연속형(숫자 범위) 결과(자산의 과세 가격 예측 등)를 자동으로 작성하고 비교할 수 있습니다. 단일 노드를 사용하여 일련의 후보 모델을 추정하고 비교할 수 있으며 나중에 분석할 수 있도록 모델의 서브세트를 생성할 수 있습니다. 이 노드는 자동 분류자 노드에서와 같은 방식으로 작동하나 플래그 또는 명목 대상이 아니라 연속형 대상에 대해 작동합니다.

노드는 최선의 후보 모델들을 단일 통합(앙상블) 모델 너깃으로 결합합니다. 이 방법은 자동화의 간편함과 어느 하나의 모델에서 얻을 수 있는 것보다 더 정확한 예측을 얻을 수 있는 다중 모델 결합의 혜택을 조합합니다.

이 예에서는 부동산 세금 조정 및 평가를 담당하는 가상의 지방 자치체에 초점을 맞춥니다. 이를 좀 더 정확하게 수행하기 위해 물 유형, 이웃, 크기, 기타 알려진 요인에 기반하여 자산 가치를 예측하는 여러 모델을 작성합니다.

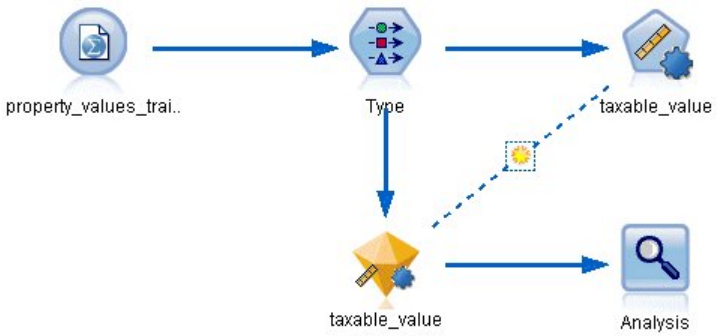


그림 *42.* 자동 숫자 샘플 스트림

이 예에서는 *streams* 아래의 Demos 폴더에 설치된 *property_values_numericpredictor.str* 스트림을 사용합니다. 사용된 데이터 파일은 *property_values_train.sav*입니다. 자세한 정보는 5 페이지의 『Demos 폴더』의 내용을 참조하십시오.

학습 데이터

데이터 파일은 예측할 **대상 필드** 또는 값이 되는 *taxable_value*라는 이름의 필드를 포함합니다. 기타 필드는 이웃, 건물 유형 및 내부 볼륨 등의 정보를 포함할 수 있으며 예측 변수로 사용될 수 있습니다.

필드 이름	레이블
property_id	특성 ID
neighborhood	구/군/시 내의 영역

필드 이름	레이블
building_type	건물 유형
year_built	건설 연도
volume_interior	내부 볼륨
volume_other	차고 및 추가 건물의 볼륨
lot_size	건축 용지 크기
taxable_value	과세 가격

이름이 *property_values_score.sav*인 스코어링 데이터 파일도 Demos 폴더에 포함됩니다. 여기에는 *taxable_value* 필드가 없는 것 외에 동일한 필드가 포함됩니다. 과세 가격이 알려진 데이터 세트를 사용하여 모델을 학습한 후에 이 가격이 아직 알려지지 않은 레코드를 스코어링할 수 있습니다.

스트림 작성

1. 사용자의 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *property_values_train.sav*를 가리키는 Statistics 파일 소스 노드를 추가하십시오. 이 폴더를 참조하기 위한 단축키로 파일 경로에서 `$CLEO_DEMOS/`를 지정할 수 있습니다. 표시된 대로 경로에 백슬래시가 아니라 슬래시가 사용되어야 합니다.)

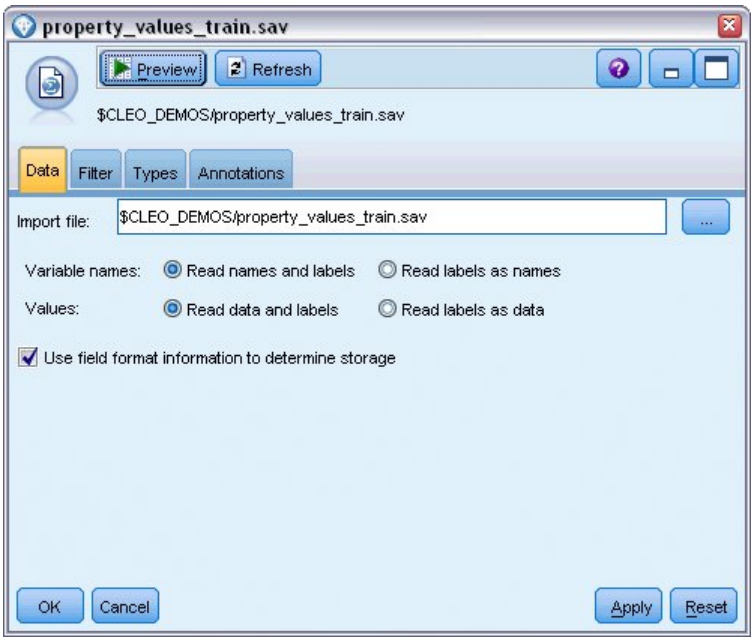


그림 *43.* 데이터에서 읽기

2. 유형 노드를 추가하고 대상 필드로 *taxable_value*를 선택하십시오(역할 = **대상**). 다른 모든 필드에 대해서는 역할이 예측변수로 사용됨을 나타내는 **입력**으로 설정되어야 합니다.

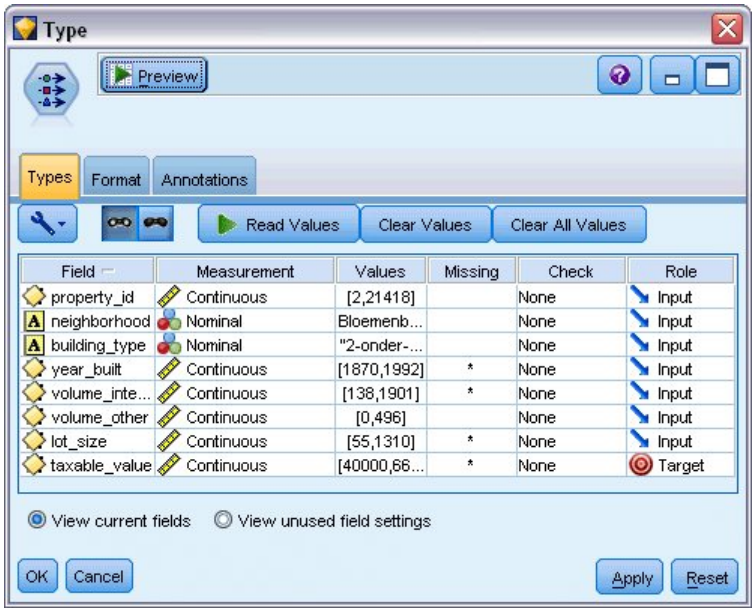


그림 *44.* 대상 필드 설정

3. 자동 숫자 노드를 연결하고 모델 순위 매기기에 사용할 메트릭으로 **상관관계**를 선택하십시오.
4. **사용할 모델 수**를 3으로 설정하십시오. 이는 모드를 실행할 때 세 개의 최적 모델이 작성됨을 의미합니다.

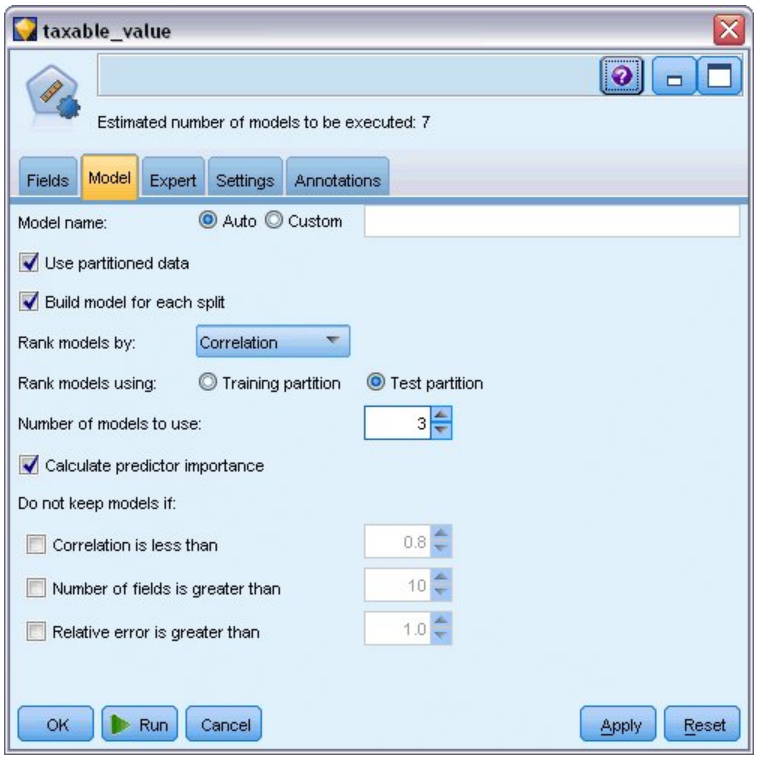


그림 *45.* 자동 숫자 노드 모델 탭

5. 전문가 탭에서 기본 설정을 그대로 두십시오. 노드가 총 일곱 개의 모델에 대해 각 알고리즘에 대해 단일 모델을 추정합니다. (또는 각 모델 유형에 대한 다중 변량을 비교하도록 해당 설정을 수정할 수 있습니다.)

   모델 탭에서 **사용할 모델 수**를 3으로 설정했으므로 노드가 일곱 가지 알고리즘의 정확도를 계산하여 세 개의 최고의 정확도를 포함하는 단일 모델 너깃을 작성합니다.

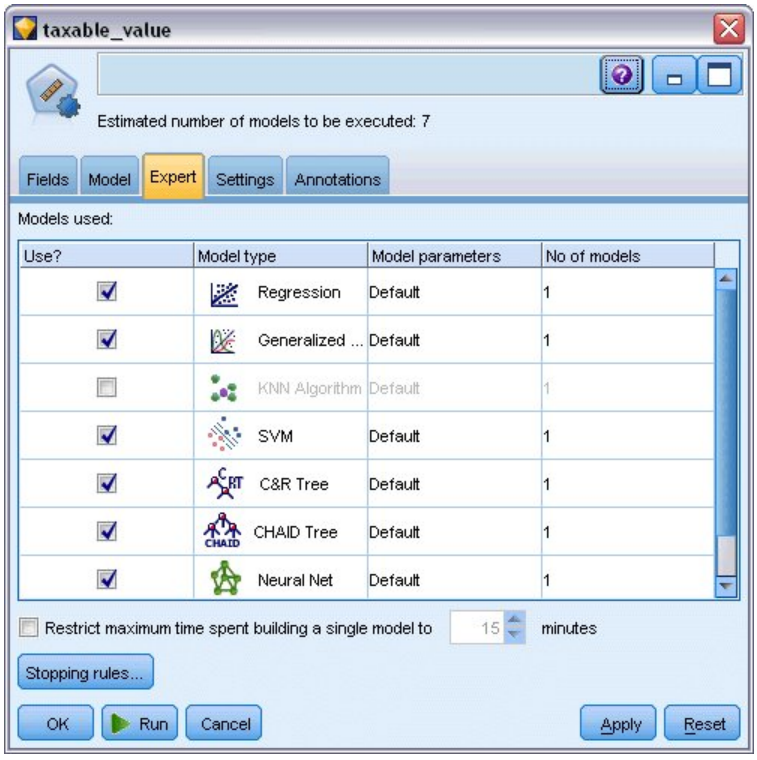


그림 *46.* 자동 숫자 노드 전문가 탭

6. 설정 탭에서 기본 설정을 그대로 두십시오. 이는 연속형 대상이므로 개별 모델에 대한 점수의 평균을 내어 앙상블 점수가 생성됩니다.

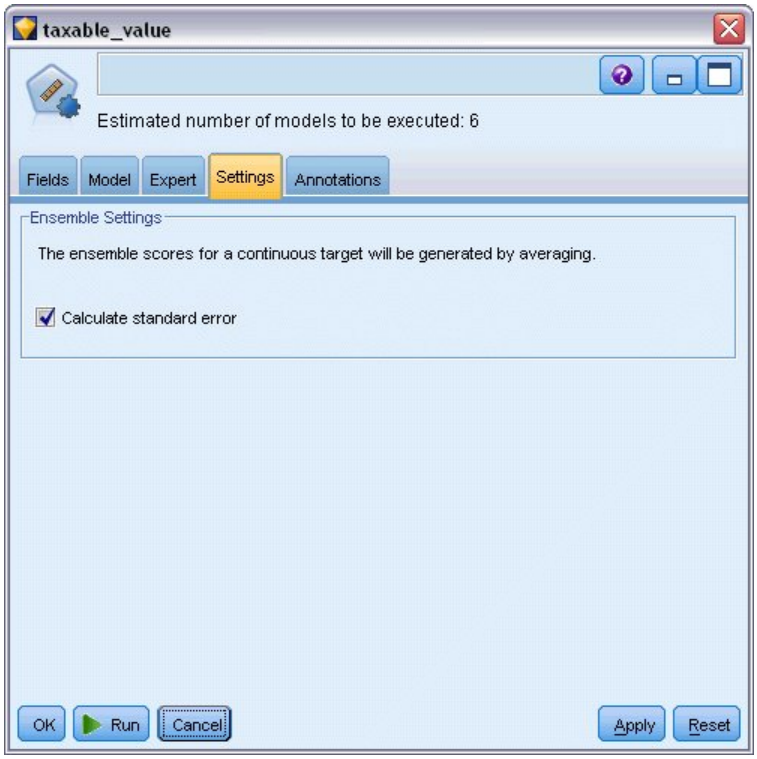


그림 *47.* 자동 숫자 노드 설정 탭

모델 비교

1. 실행 단추를 클릭하십시오.

모델 너깃이 작성되고 캔버스 및 창의 오른쪽 상단 코너의 모델 팔레트에도 배치됩니다. 너깃을 찾아보거나 이를 저장하거나 수많은 방법으로 배치할 수 있습니다.

모델 너깃을 여십시오. 여기에 실행 동안 작성된 각 모델에 대한 세부사항이 나열됩니다. (실제 상황에서는 대형 데이터 세트에 대해 수백 개의 모델이 추정될 수 있으므로 몇 시간이 걸릴 수도 있습니다.) 55 페이지의 그림 42의 내용을 참조하십시오.

임의의 개별 모델을 추가 탐색하려면 **모델** 열에서 모델 너깃 아이콘을 두 번 클릭하여 드릴다운하여 개별 모델 결과를 찾아볼 수 있습니다. 거기서부터 모델링 노드, 모델 너깃 또는 평가 도표를 생성할 수 있습니다.

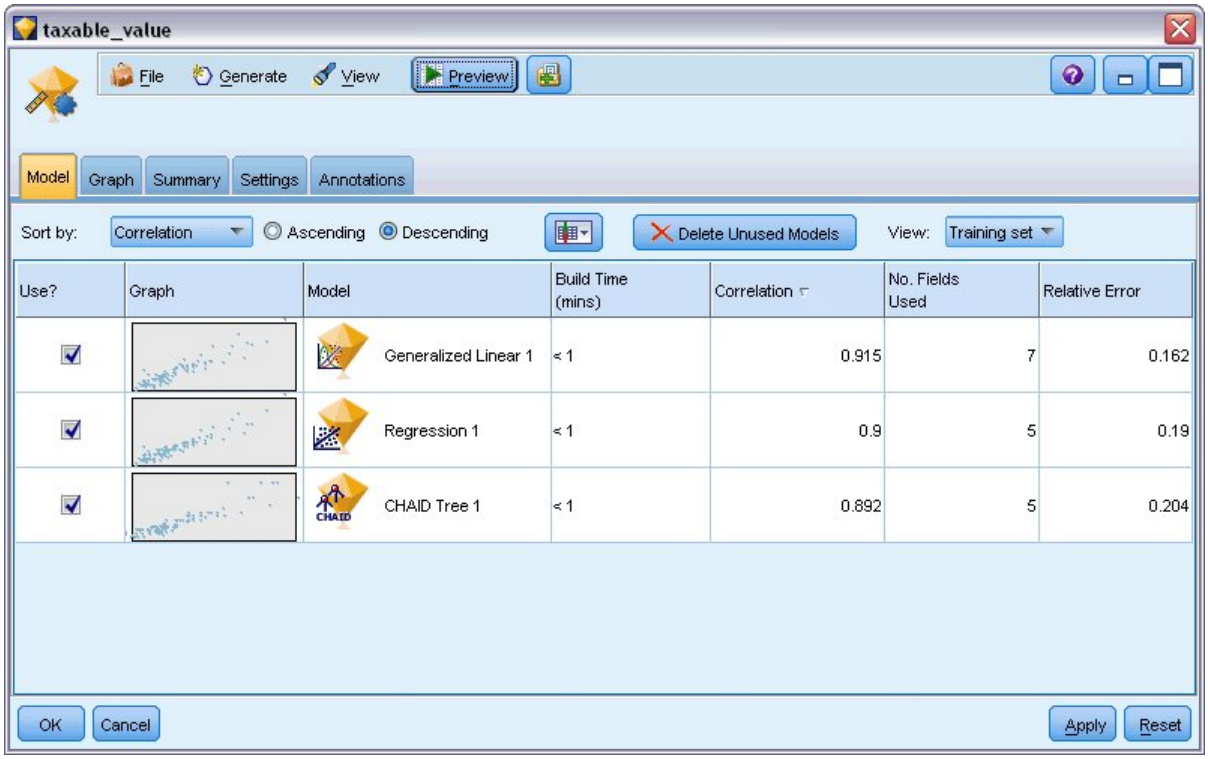


그림 *48.* 자동 숫자 결과

기본적으로 모델은 상관관계를 기준으로 하여 정렬됩니다. 사용자가 자동 숫자 노드에서 선택한 측도가 상관관계이기 때문입니다. 순위화의 목적으로 상관관계의 절대값이 사용됩니다. 값이 1에 가까울수록 더 강한 관계임을 나타냅니다. 일반화된 선형 모델이 이 측도에서 가장 상위 순위이나 여러 다른 모델이 거의 근접한 정확도를 나타냅니다. 또한 일반화된 선형 모델은 상대 오차도 가장 낮습니다.

해당 열의 헤더를 클릭하여 다른 열 기준으로 정렬하거나 도구 모음의 **정렬기준** 목록에서 원하는 측도를 선택할 수 있습니다.

각 그래프는 모델의 관측값 대 예측값 도표를 표시하며 둘 사이의 상관관계에 대한 빠른 시각적 표시를 제공합니다. 좋은 모델인 경우, 이 예의 모든 모델에서와 같이 점이 대각선을 따라 모여 있어야 합니다.

그래프 열에서 썸네일을 두 번 클릭하여 전체 크기의 그래프를 생성하십시오.

이러한 결과를 기준으로 하여 모두 세 개의 가장 정확한 모델을 사용하기로 결정했습니다. 여러 모델의 예측을 결합하면 개별 모델의 제한을 피할 수 있어 전반적인 정확도가 향상됩니다.

사용? 열에서 세 모델이 모두 선택되어 있는지 확인하십시오.

분석 노드(출력 팔레트)를 모델 너깃 뒤에 연결하십시오. 마우스 오른쪽 단추로 분석 노드를 클릭하고 **실행**을 선택하여 스트림을 실행하십시오.

앙상블 모델에 의해 생성된 평균 스코어가 0.922의 상관관계로 *$XR-taxable_value*라는 필드에 추가됩니다. 이는 세 가지 개별 모델의 상관관계보다 높습니다. 또한 앙상블 점수는 낮은 평균 절대 오차를 표시하며 기타 데이터 세트에 적용될 때 어떠한 개별 모델보다 더 잘 수행할 수 있습니다.

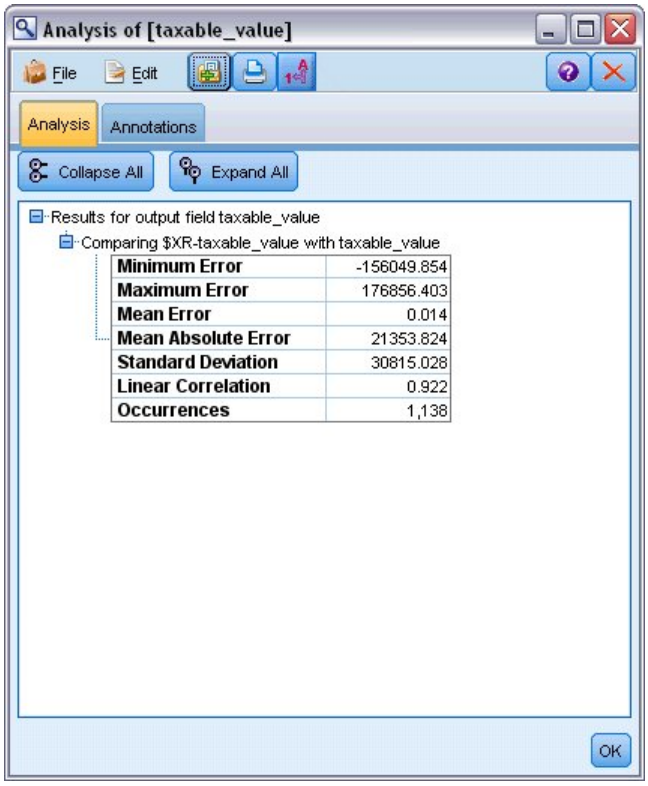


그림 *49.* 자동 숫자 샘플 스트림

요약

합계를 계산하기 위해 자동 숫자 노드를 사용하여 다른 모델의 수를 비교하고 세 개의 가장 정확한 모델을 선택하고 앙상블 자동 숫자 모델 너깃 내의 스트림에 이를 추가했습니다.

- 전체 정확도를 기준으로 할 때, 학습 데이터에서 일반화된 선형, 회귀분석 및 CHAID 모델의 수행이 가장 뛰어났습니다.
- 앙상블 모델은 두 개별 모델보다 우수한 수행을 나타냈으며 기타 데이터 세트에 적용될 때는 더 잘 수행할 수도 있습니다. 사용자의 목표가 가능한 한 프로세스를 자동화하는 것인 경우, 이 접근법을 사용하면 특정한 어느 한 모델을 깊이 조사할 필요 없이 대부분의 상황에서 강력한 모델을 얻을 수 있습니다.

제 6 장 자동 데이터 준비(ADP)

분석을 위한 데이터 준비는 모든 데이터 마이닝 프로젝트에서 가장 중요한 단계 중 하나이며 일반적으로 가장 많은 시간이 소요되는 단계 중 하나입니다. 자동 데이터 준비(ADP) 노드는 데이터 분석, 수정사항 식별, 문제가 있거나 유용할 것 같지 않은 필드 필터링, 적절한 경우 새 속성 파생 및 지능형 선별 기술을 통한 성능 개선 등의 작업을 자동으로 처리합니다. 완전 자동화된 방식으로 노드를 사용하여 노드가 수정사항을 선택하고 적용할 수 있게 하거나, 변경사항이 작성 및 승인되기 전에 원하는 대로 변경을 미리 보거나 거부할 수 있습니다.

ADP 노드를 사용하면 관련된 통계 개념에 대한 사전 지식 없이 데이터 마이닝을 위한 데이터를 쉽고 빠르게 준비할 수 있습니다. 기본 설정을 사용하여 노드를 실행하면 모델을 더 빠르게 작성하고 스코어링할 수 있습니다.

이 예에서는 *ADP_basic_demo.str*이라는 스트림을 사용하며 이는 *telco.sav*라는 데이터 파일을 참조하여 모델을 작성할 때의 기본 ADP 노드를 사용하여 발견할 수 있는 증가하는 정확도를 증명합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *ADP_basic_demo.str* 파일은 *streams* 디렉토리에 있습니다.

스트림 작성

1. 스트림을 작성하려면 IBM SPSS Modeler 설치의 *Demos* 디렉토리에 있는 *telco.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.

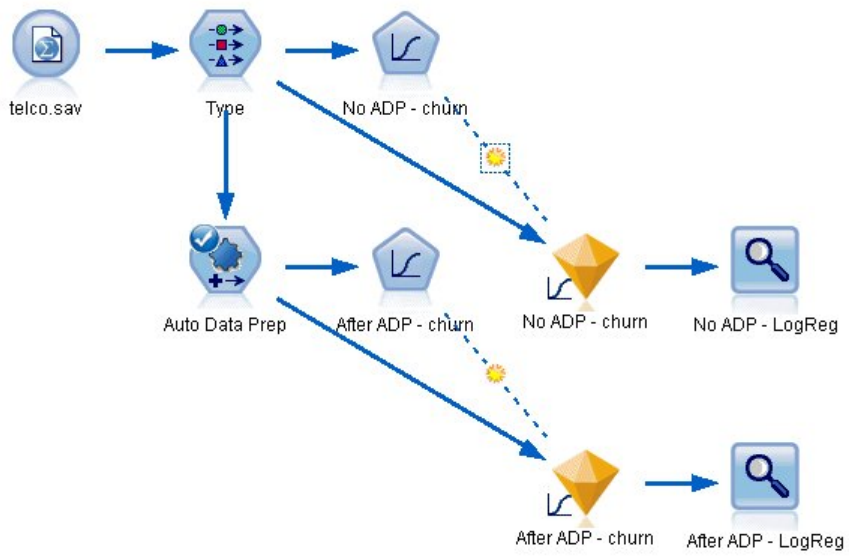


그림 *50.* 스트림 작성

2. 유형 노드를 소스 노드에 첨부하고 서비스 제공자를 바꾸는 고객 필드에 대한 측정 수준을 **플래그**로 설정하고 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.

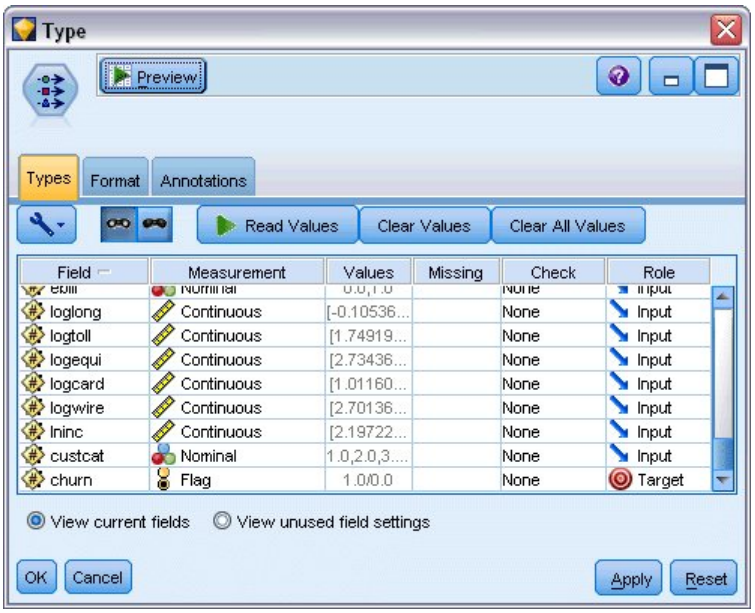


그림 *51.* 대상 선택

3. 로지스틱 노드를 유형 노드에 연결하십시오.
4. 로지스틱 노드에서 모델 탭을 클릭하여 **이항** 프로시저를 선택하십시오. 모델 이름 필드에서 **사용자 정의** 및 ADP 없음 - 서비스 제공자를 바꾸는 고객을 선택하십시오.

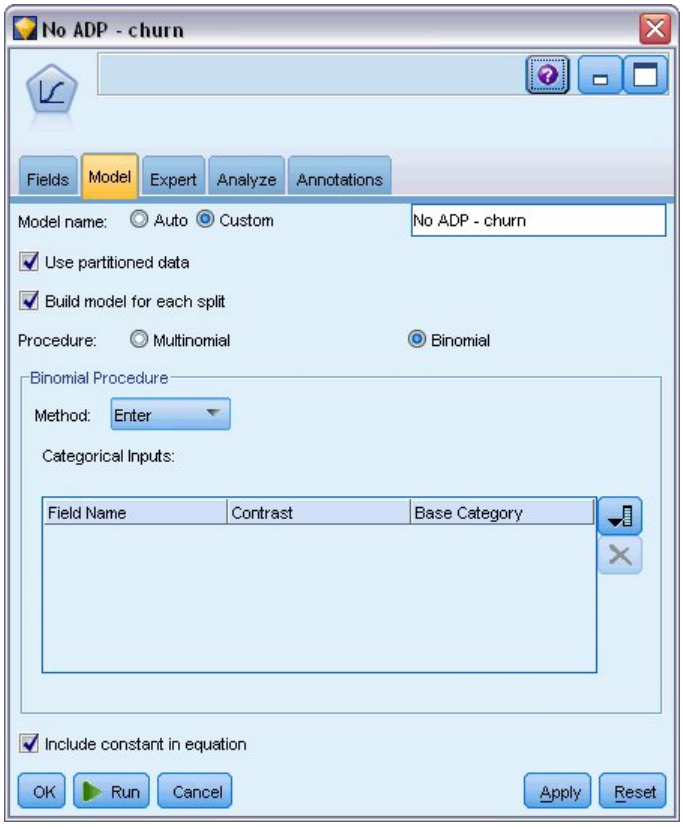


그림 *52.* 모델 옵션 선택

5. ADP 노드를 유형 노드에 연결하십시오. 목적 탭에서 기본 설정을 그대로 두고 속도 및 정확도의 균형을 모두 유지하면서 데이터를 분석하고 준비하십시오.
6. 목적 탭의 맨 위에 있는 **데이터 분석**을 클릭하여 데이터를 분석하고 처리하십시오.

   ADP 노드의 기타 옵션을 사용하면 정확도에 더 집중할 것인지 처리 속도에 더 집중할 것인지 결정하거나 데이터 준비 처리 단계의 많은 부분을 정교하게 조정할 수 있습니다.

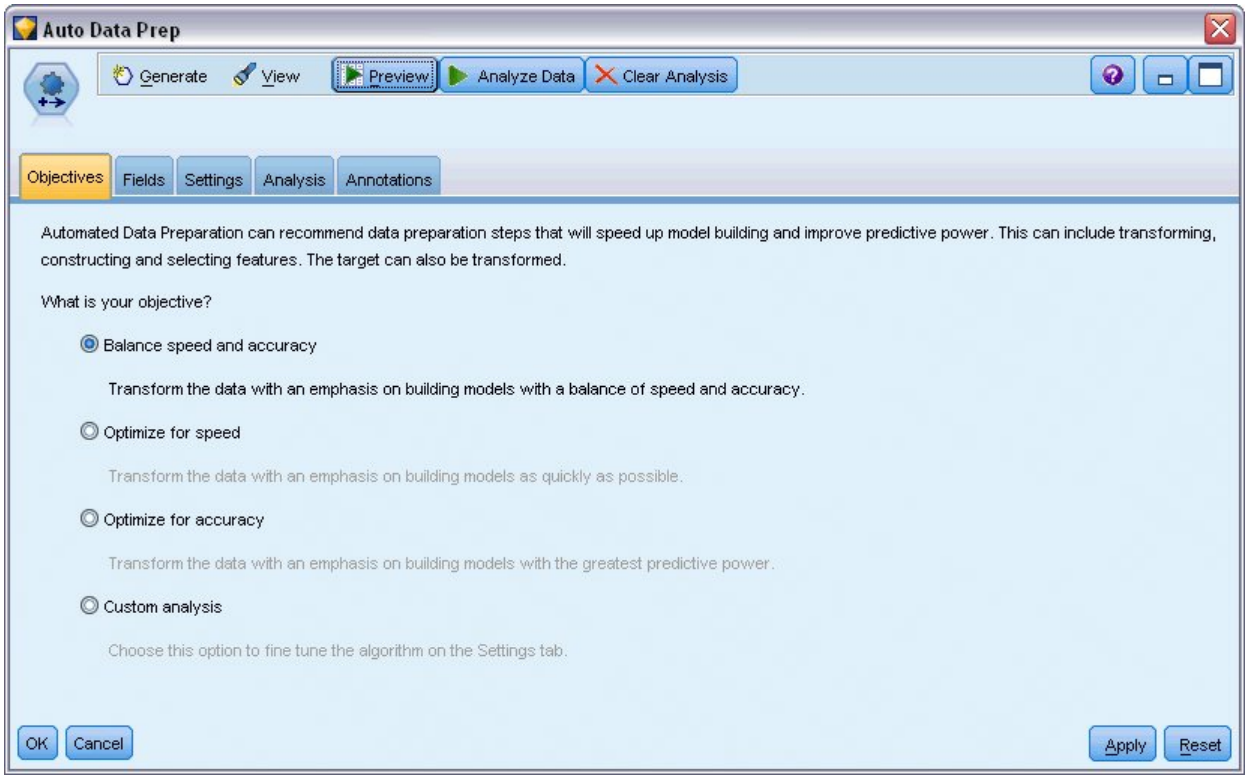


그림 *53. ADP* 기본 목적

데이터 처리의 결과는 분석 탭에 표시됩니다. **필드 처리 요약**은 41개의 데이터 필드를 ADP 노드로 가져오고 19개가 처리를 돕기 위해 변환되었으며 3개가 사용되지 않고 삭제되었음을 표시합니다.

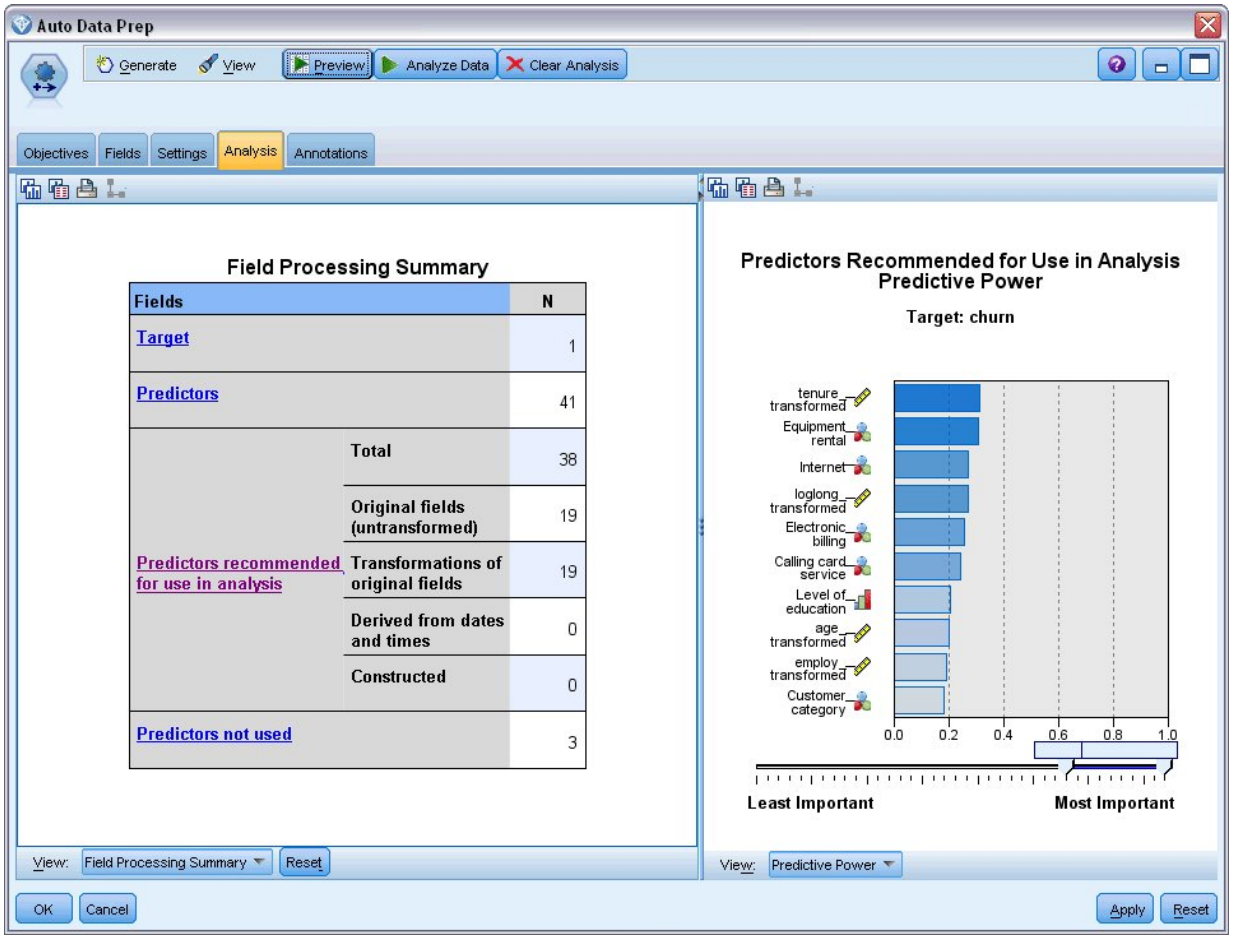


그림 *54.* 데이터 처리 요약

7. 로지스틱 노드를 ADP 노드에 연결하십시오.
8. 로지스틱 노드에서 모델 탭을 클릭하여 **이항** 프로시저를 선택하십시오. 모델링 이름 필드에서 **사용자 정의** 및 ADP 이후 - 서비스 제공자를 바꾸는 고객을 선택하십시오.

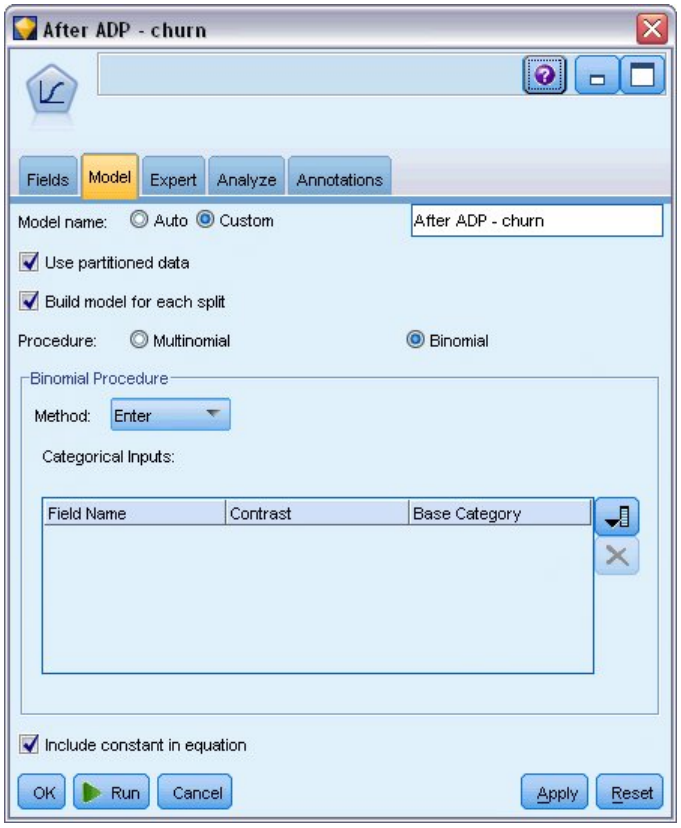


그림 *55.* 모델 옵션 선택

모델 정확도 비교

1. 두 로지스틱 모델을 실행하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에 추가됩니다.

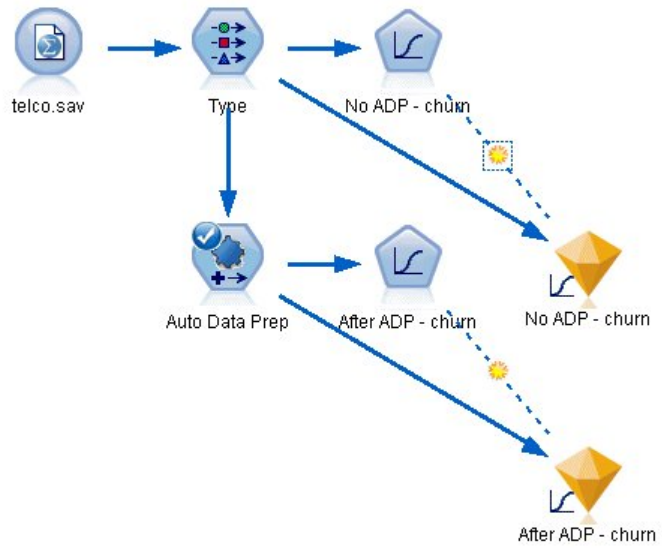


그림 *56.* 모델 너깃 첨부

2. 분석 노드를 모델 너깃에 첨부하고 기본 설정을 사용하여 분석 노드를 실행하십시오.

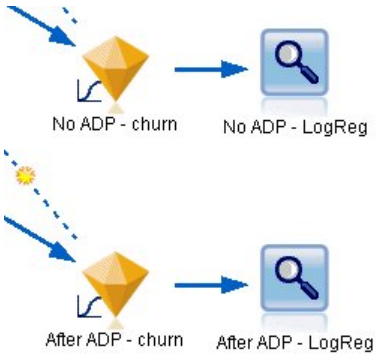


그림 *57.* 분석 노드 첨부

ADP에서 파생되지 않은 모델을 분석하면 기본 설정을 사용하여 로지스틱 회귀분석 노드를 통해 데이터를 실행할 때 모델의 정확도가 낮음(단 10.6%)을 보여줍니다.

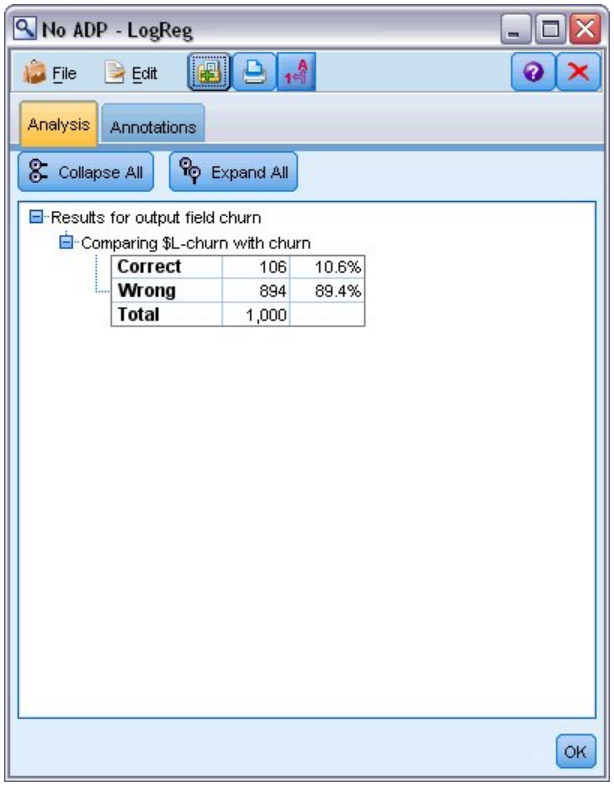


그림 *58. ADP*에서 파생되지 않은 모델 결과

기본 ADP 설정을 통해 데이터를 실행한 ADP에서 파생된 모델의 분석은 78.8% 적합한 훨씬 더 정확한 모델을 작성했음을 표시합니다.

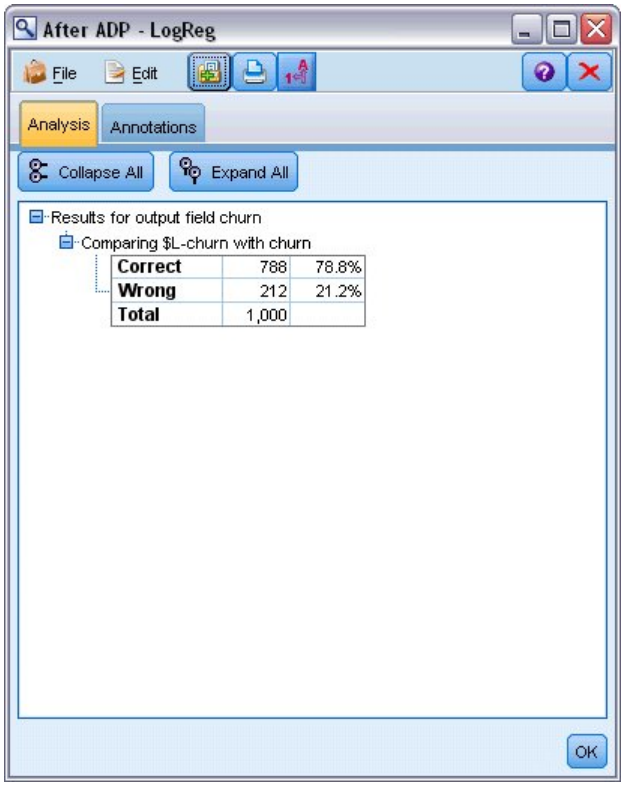


그림 *59. ADP*에서 파생된 모델 결과

요약하면 ADP 노드 실행만으로 데이터 처리를 정교하게 조정할 수 있으며 직접적인 데이터 조작을 거의 수행하지 않고도 훨씬 정확한 모델을 작성할 수 있었습니다.

특정 이론이 맞거나 틀렸음을 입증하는 데 관심이 있거나 특정 모델을 작성하고자 하는 경우에는 모델 설정에 직접 작업하는 것이 유용하다는 것이 확실하나 많은 양의 데이터를 준비해야 하는 경우에는 시간이 줄어든다는 점에서 ADP 노드가 장점이 있습니다.

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 설치 디스크의 *\Documentation* 디렉토리에서 사용 가능한 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

이 예의 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 실세계에서 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

제 7 장 분석용 데이터 준비(데이터 검토)

데이터 검토 노드는 IBM SPSS Modeler로 가져오는 데이터에 대한 포괄적인 정보를 간략하게 제공합니다. 데이터 검토 보고서가 초기 데이터 검토 시에 사용되는 경우 각 데이터 필드에 대한 히스토그램 및 분포 그래프와 함께 요약을 표시하여 사용자가 결측값, 이상치 및 극단값에 대한 처리를 지정할 수 있는 경우가 종종 있습니다.

이 예에서는 *telco.sav*라는 데이터 파일을 참조하는 *telco_dataaudit.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *telco_dataaudit.str* 파일은 *streams* 디렉토리에 있습니다.

스트림 작성

1. 스트림을 작성하려면 IBM SPSS Modeler 설치의 *Demos* 디렉토리에 있는 *telco.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.

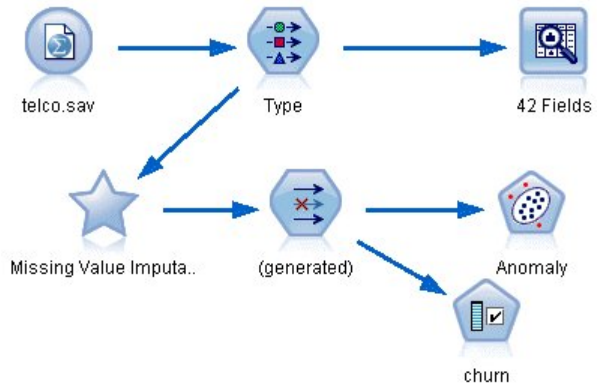


그림 *60.* 스트림 작성

2. 유형 노드를 추가하여 필드를 정의하고 *churn*을 대상 필드로 지정하십시오(역할 = **대상**). 이 필드가 유일한 대상이 되도록 다른 모든 필드에 대해서는 역할이 **입력**으로 설정되어야 합니다.

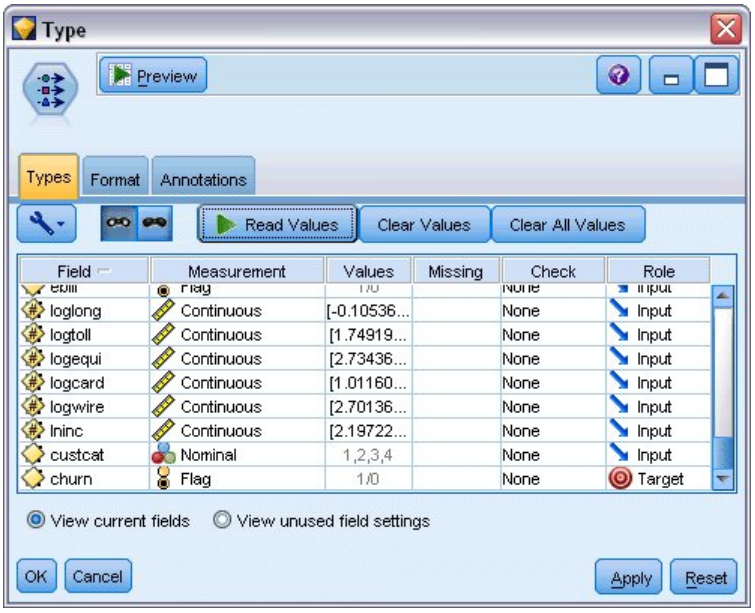


그림 *61.* 대상 설정

3. 필드 측정 수준이 올바르게 정의되었는지 확인하십시오. 예를 들어, 값이 0 및 1인 대부분의 필드는 플래그로 간주할 수 있으나 성별과 같은 특정 필드는 두 개의 값이 있는 명목 필드로 간주하는 것이 더 정확합니다.

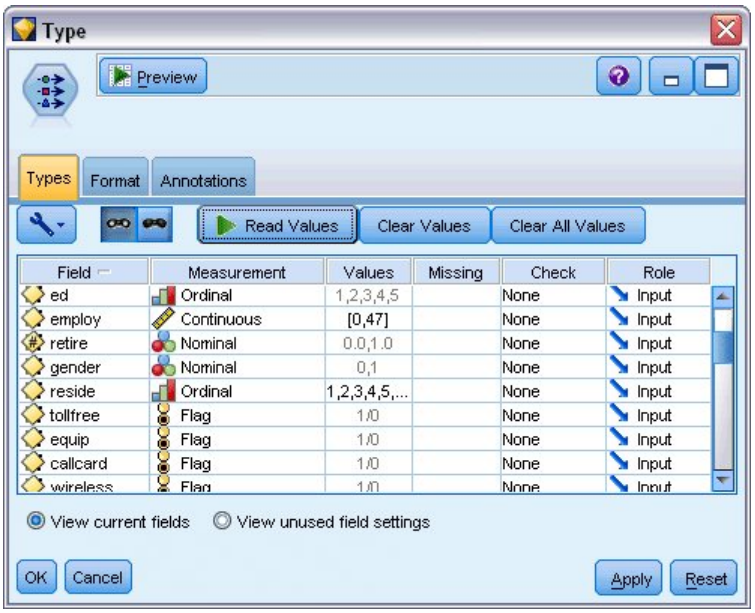


그림 *62.* 측정 수준 설정

팁: 유사한 값(0/1 등)을 가진 다중 필드에 대한 특성을 변경하려면 값 열 헤더를 클릭하여 필드를 해당 열 기준으로 정렬한 다음 Shift 키를 사용하여 변경할 키를 모두 선택하십시오. 그런 다음 마우스 오른쪽 단추로 선택영역을 클릭하여 선택된 필드의 측정 수준 또는 기타 속성을 변경하십시오.

4. 데이터 검토 노드를 스트림에 연결하십시오. 모든 필드를 보고서에 포함하도록 설정 탭에서 기본 설정을 그대로 두십시오. *churn*이 유형 노드 내에서 유일하게 정의된 대상 필드이므로 자동으로 오버레이로 사용됩니다.

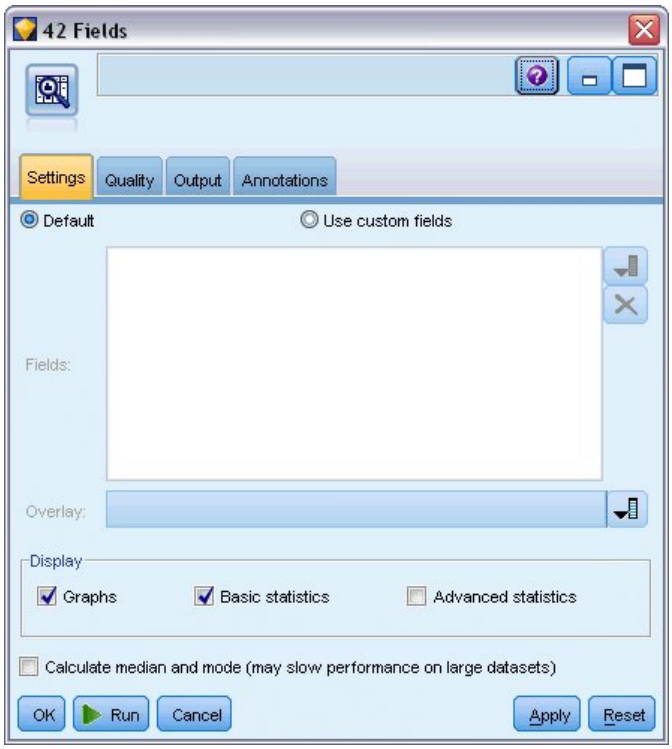


그림 *63.* 데이터 검토 노드, 설정 탭

품질 탭에서 결측값, 이상치 및 극단값을 발견하기 위한 기본 설정을 그대로 두고 **실행**을 클릭하십시오.

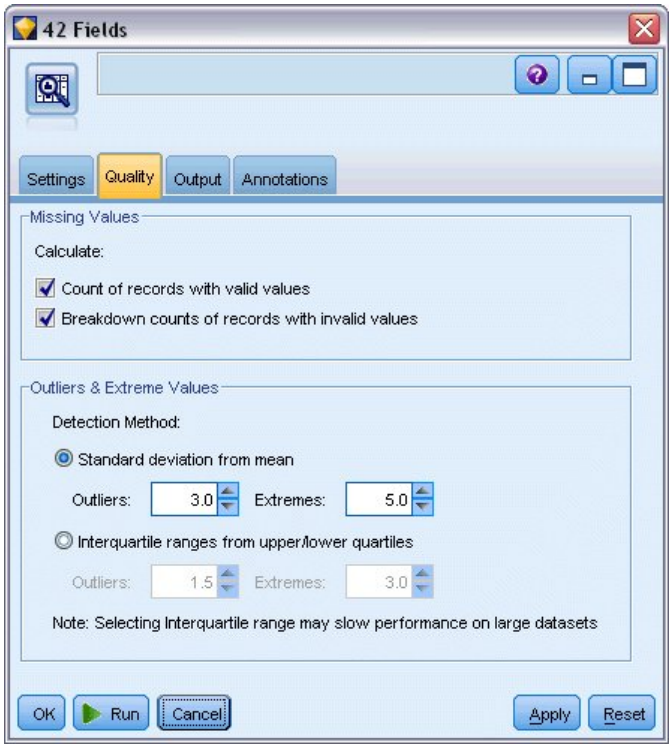


그림 *64.* 데이터 검토 노드, 품질 탭

통계 및 도표 찾아보기

썸네일 그래프 및 각 필드에 대한 기술통계와 함께 데이터 검토 브라우저가 표시됩니다.

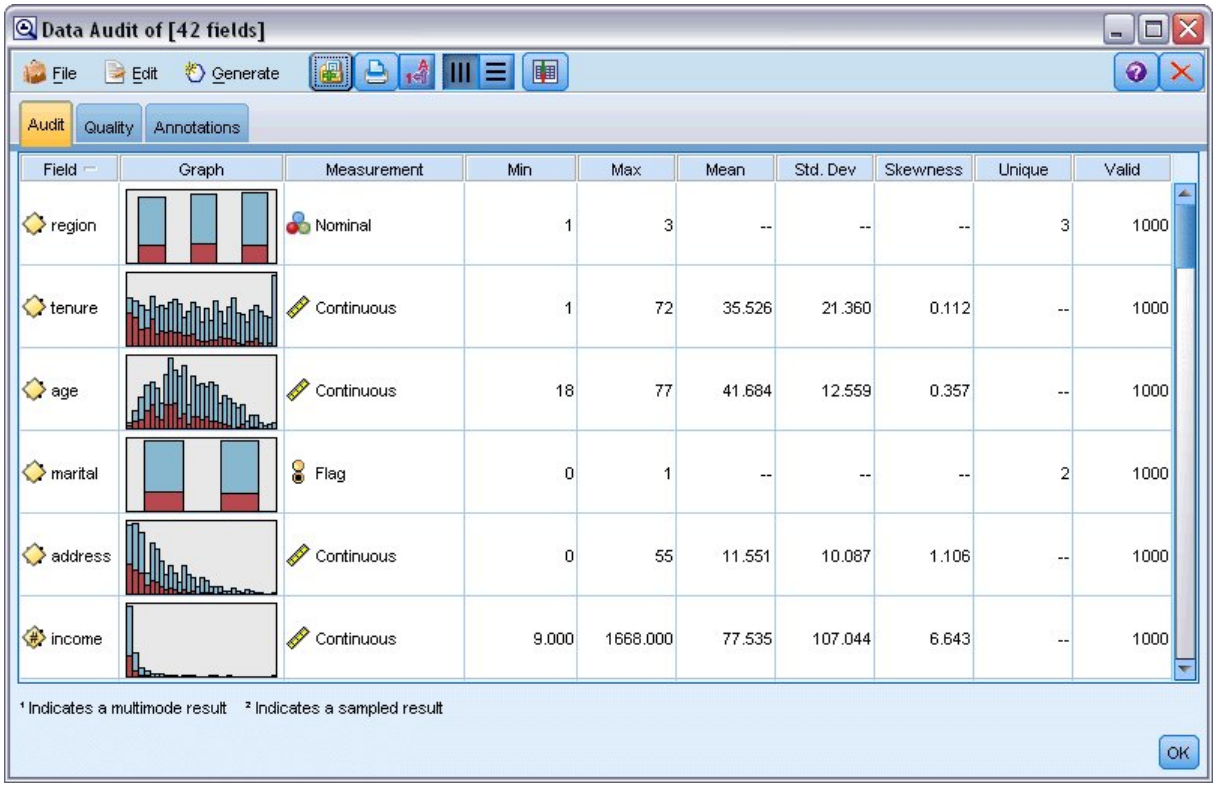


그림 *65.* 데이터 검토 브라우저

필드 및 값 레이블을 표시하고 차트 맞춤을 수평에서 수직으로 토글하려면(범주형 필드에만 해당됨) 도구 모음을 사용하십시오.

1. 또한 도구 모음 또는 편집 메뉴를 사용하여 표시할 통계를 선택할 수 있습니다.

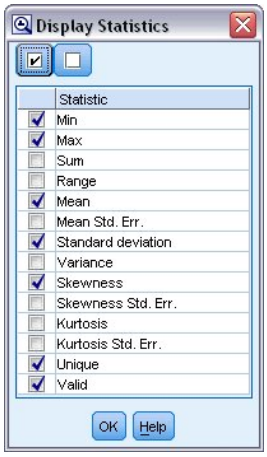


그림 *66.* 통계 표시

감사 보고서에서 임의의 썸네일 그래프를 두 번 클릭하여 해당 차트의 전체 크기 버전을 볼 수 있습니다. *churn*이 스트림 내의 유일한 대상 필드이므로 자동으로 오버레이로 사용됩니다. 그래프 창 도구 모음을 사용하여 필드 및 값의 표시를 토글하거나 편집 모드 단추를 클릭하여 차트를 추가적으로 사용자

정의할 수 있습니다.

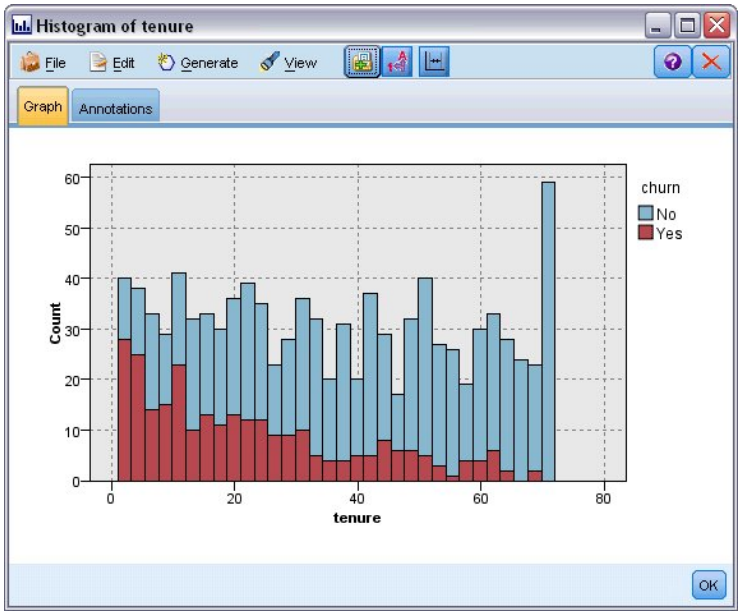


그림 *67.* 재직 히스토그램

또는 하나 이상의 썸네일을 선택하여 각각에 대한 그래프 노드를 생성할 수 있습니다. 생성된 노드는 스트림 캔버스에 배치되며 특정 그래프를 다시 작성하기 위해 스트림에 추가될 수 있습니다.

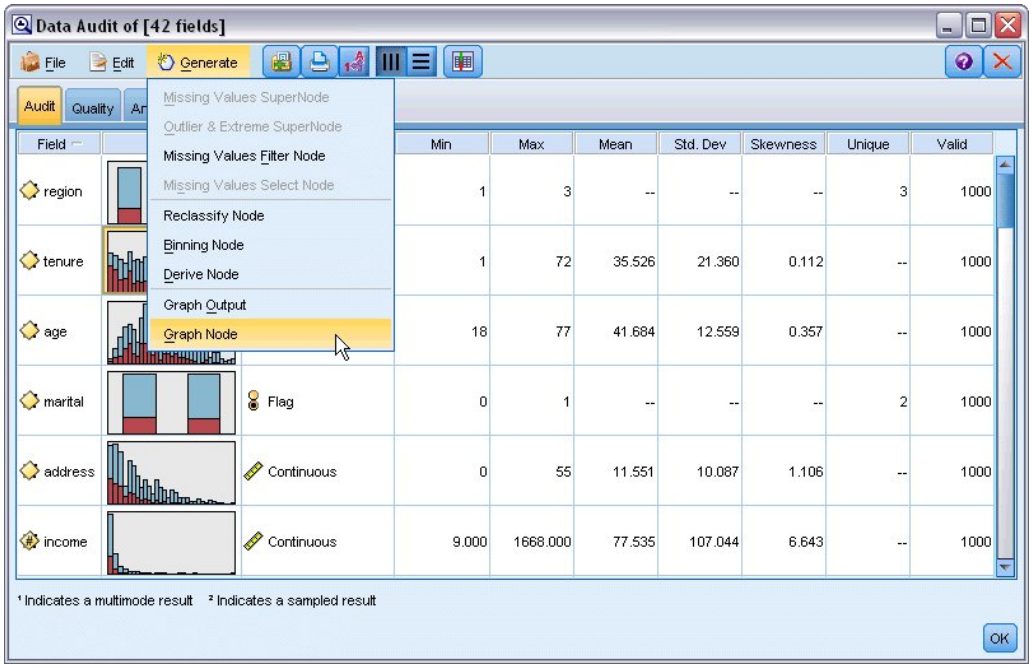


그림 *68.* 그래프 노드 생성

이상치 및 결측값 처리

감사 보고서의 품질 탭은 이상치, 극단값 및 결측값에 대한 정보를 표시합니다.

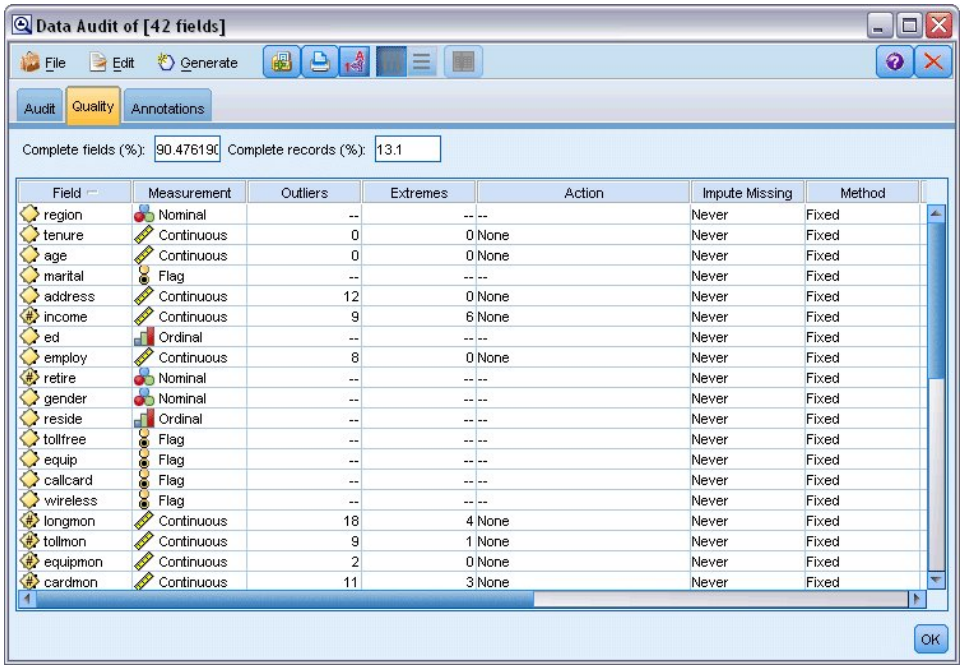


그림 *69.* 데이터 검토 브라우저, 품질 탭

또한 이러한 값을 처리하기 위한 방법을 지정하고 수퍼 노드를 생성하여 자동으로 변환을 적용할 수 있습니다. 예를 들어, 하나 이상의 필드를 선택하여 이러한 필드에 대해 C&RT 알고리즘을 포함하여 수많은 방법을 사용하여 결측값을 대치하거나 대체하도록 선택할 수 있습니다.

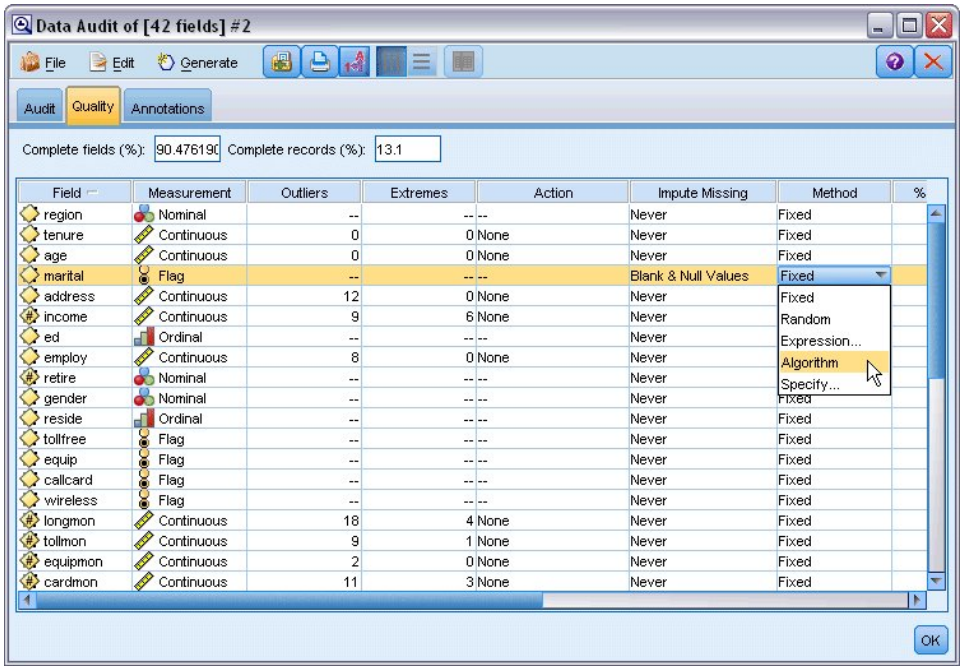


그림 *70.* 대치 방법 선택

하나 이상의 필드에 대해 대치 방법을 지정한 다음 결측값 수퍼 노드를 생성하려면 메뉴에서 다음을 선택하십시오.

생성 > 결측값 수퍼 노드

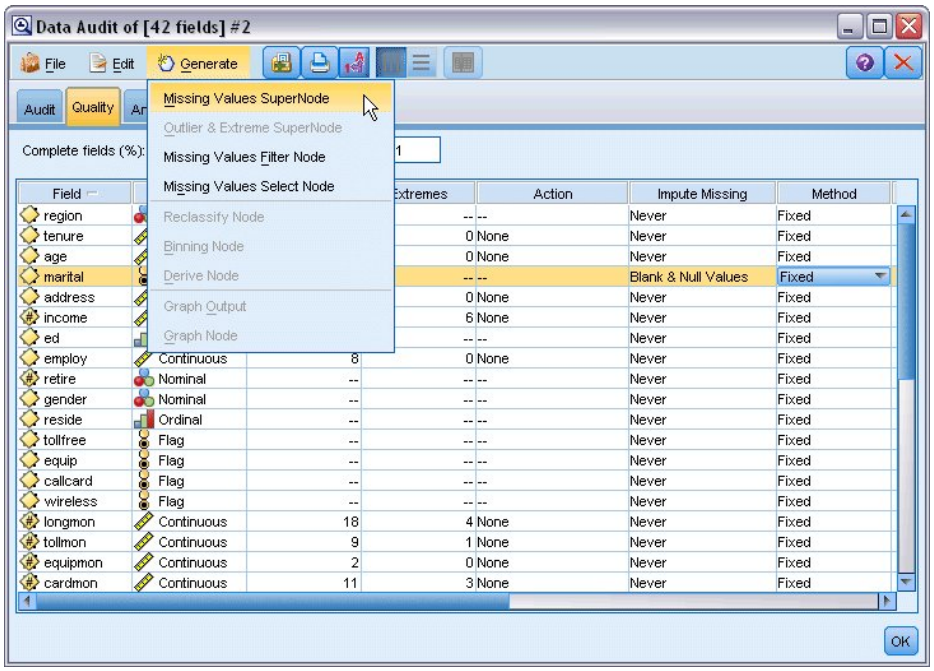


그림 *71.* 수퍼 노드 생성

생성된 수퍼 노드가 스트림 캔버스에 추가되며 여기서 이를 스트림에 추가하여 변환을 적용할 수 있습니다.

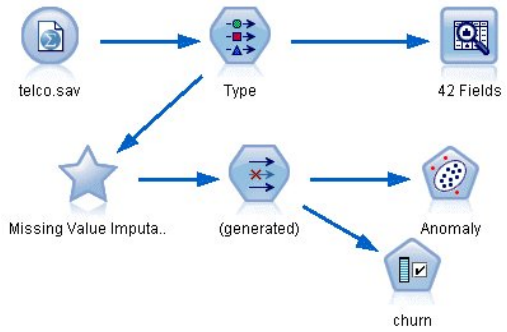


그림 72. 결측값 수퍼 노드가 있는 스트림

수퍼 노드는 실제로 요청된 변환을 수행하는 일련의 노드를 포함합니다. 수퍼 노드를 편집하고 **확대**를 클릭하면 작동 방법을 이해할 수 있습니다.

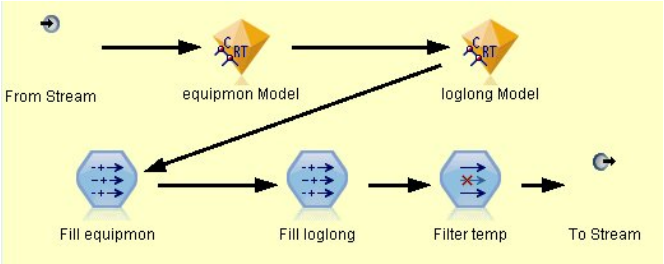


그림 73. 수퍼 노드 확대

예를 들어, 알고리즘 방법을 사용하여 대치된 각 필드의 경우, 공백과 널을 모델에 의해 예측된 값으로 대체하는 채움 노드와 함께 별도의 C&RT 모델이 있습니다. 작동을 추가적으로 사용자 정의하기 위해 수퍼 노드 내의 특정 노드를 추가, 편집 또는 제거할 수 있습니다.

또는 선택 또는 필터 노드를 생성하여 결측값이 있는 필드 또는 레코드를 제거할 수 있습니다. 예를 들어, 품질 퍼센트가 지정된 임계값보다 낮은 모든 필드를 필터링할 수 있습니다.

그림 74. 필터 노드 생성

이상치 및 극단값은 유사한 방법으로 처리할 수 있습니다. 각 필드에 대해 취할 조치(강제 실행, 삭제 또는 무효화 등)를 지정하고 수퍼 노드를 생성하여 변환을 적용할 수 있습니다.

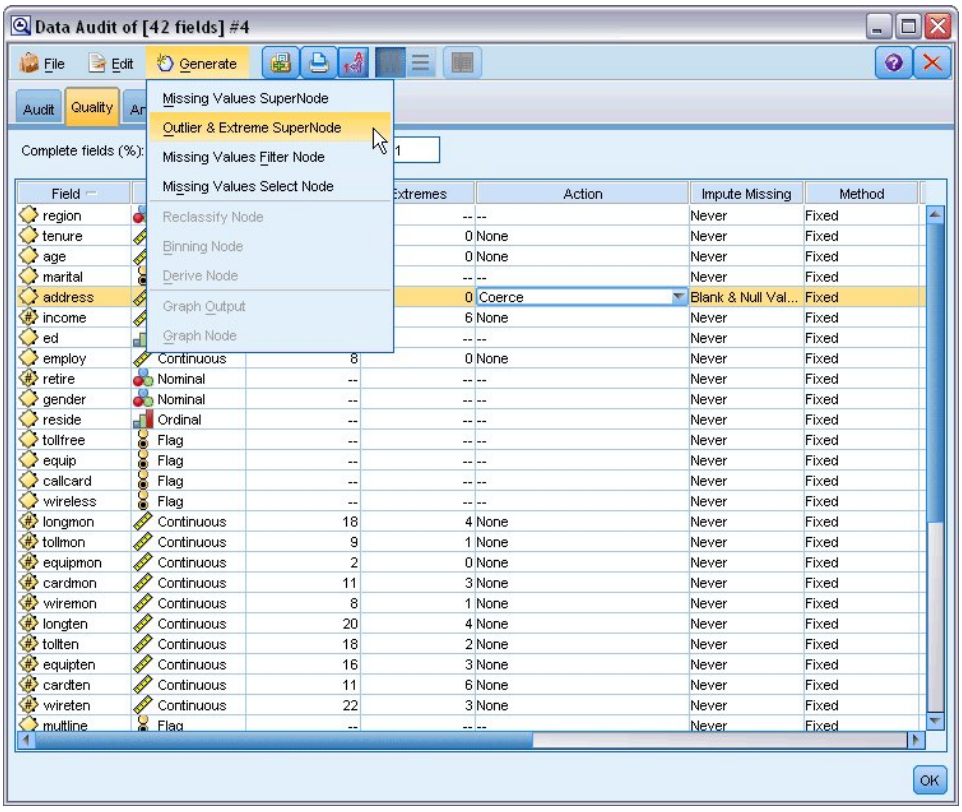


그림 *75.* 필터 노드 생성

감사를 완료하고 생성된 노드를 스트림에 추가한 후에 분석을 계속 진행할 수 있습니다. 필요에 따라 이상 항목 발견, 필드선택 또는 수많은 기타 방법을 사용하여 데이터를 추가적으로 선별할 수 있습니다.

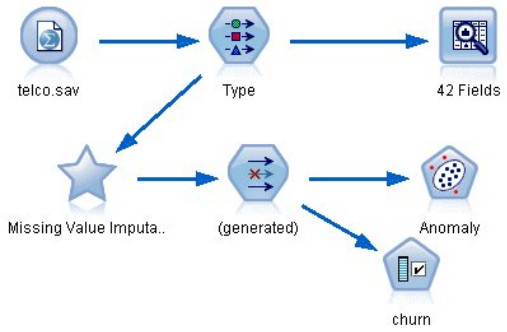


그림 *76.* 결측값 수퍼 노드가 있는 스트림

제 8 장 약물 치료(예비 그래프/C5.0)

이 절에서는 사용자가 연구에 대한 데이터를 컴파일하는 의학 연구자라고 가정합니다. 사용자는 모두 동일한 질병을 앓고 있는 일련의 환자에 대한 데이터를 수집해왔습니다. 치료 과정 중에 각 환자는 다섯 가지 약물 치료 중 하나에 반응했습니다. 작업 중 일부는 데이터 마이닝을 사용하여 동일한 질병을 앓는 미래의 환자에게 어느 약물이 적합한지 찾는 것입니다.

이 예에서는 *DRUG1n*이라는 데이터 파일을 참조하는 *druglearn.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *druglearn.str* 파일은 *streams* 디렉토리에 있습니다.

데모에서 사용되는 데이터 필드는 다음과 같습니다.

데이터 필드	설명
나이	(숫자)
성별	남 또는 여
BP	혈압: 높음, 정상 또는 낮음
콜레스테롤	혈액 콜레스테롤: 정상 또는 높음
Na	혈액 내 나트륨 농도
K	혈액 내 칼륨 농도
약제	환자가 반응한 처방 약물

텍스트 데이터에서 읽기

그림 77. 가변파일 노드 추가

가변파일 노드를 사용하여 구분자에 의한 배열 텍스트 데이터를 읽을 수 있습니다. **소스** 탭을 클릭하여 노드를 찾거나 기본적으로 이 노드를 포함하는 **즐겨찾기** 탭을 사용하여 팔레트에서 가변파일 노드를 추가할 수 있습니다. 그런 다음, 새로 배치된 노드를 두 번 클릭하여 대화 상자를 여십시오.

생략 기호(...)로 표시된 파일 선택란 바로 오른쪽의 단추를 클릭하여 IBM SPSS Modeler가 사용자의 시스템에 설치된 디렉토리를 찾아보십시오. *Demos* 디렉토리를 열고 *DRUG1n*이라는 파일을 선택하십시오.

파일에서 필드 이름 읽기가 선택되었는지 확인하고 방금 대화 상자에 로드된 필드 및 값이 있는지 확인하십시오.

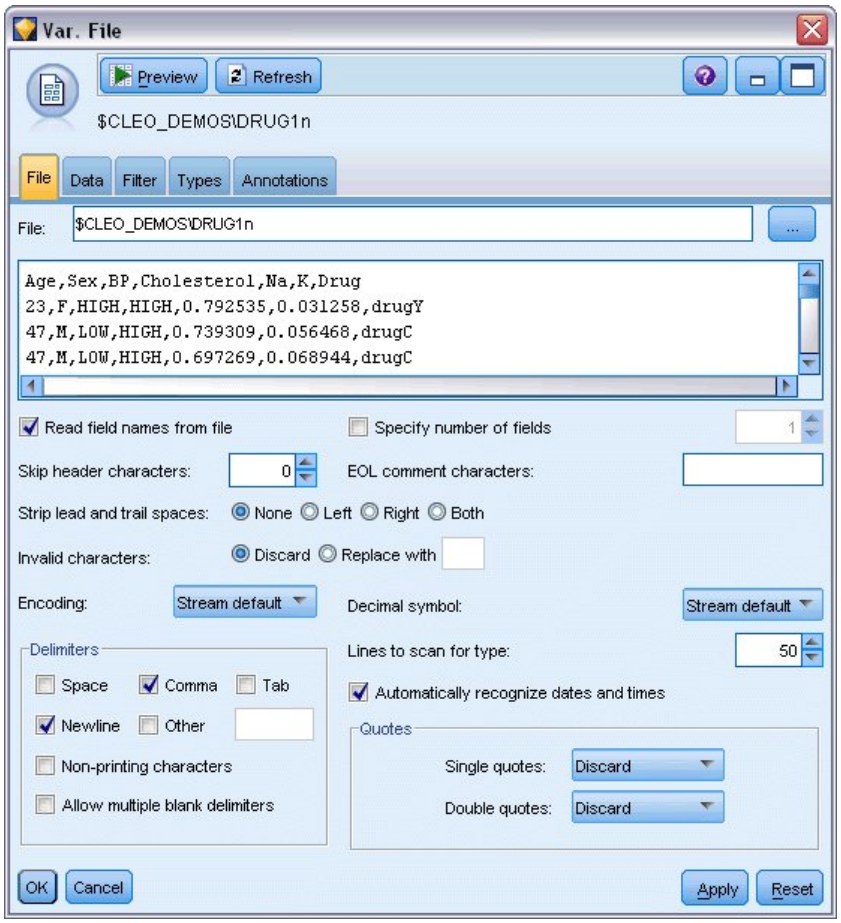


그림 *78.* 가변파일 대화 상자

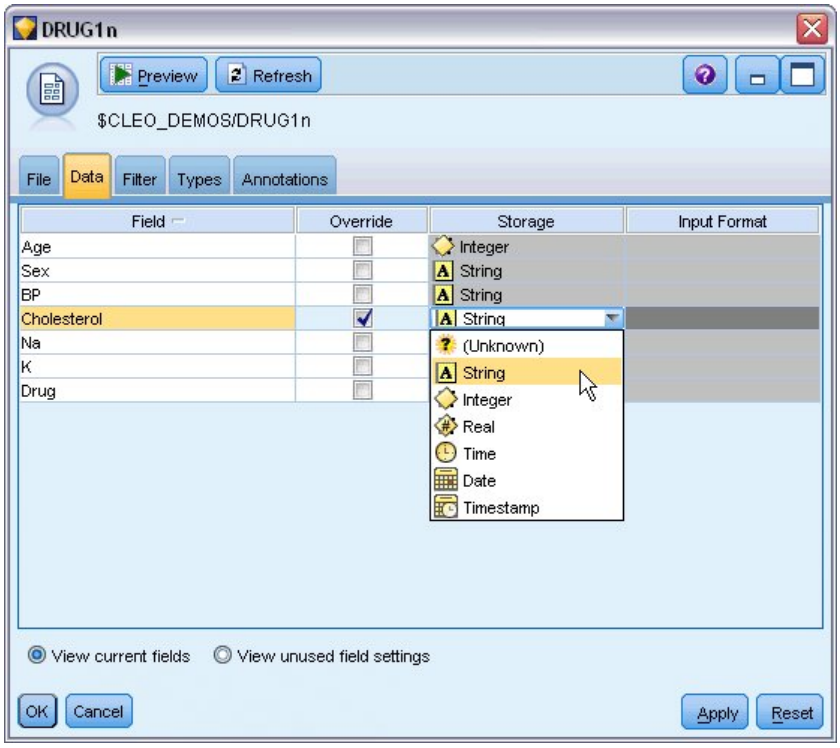


그림 *79.* 필드에 대한 저장 유형 변경

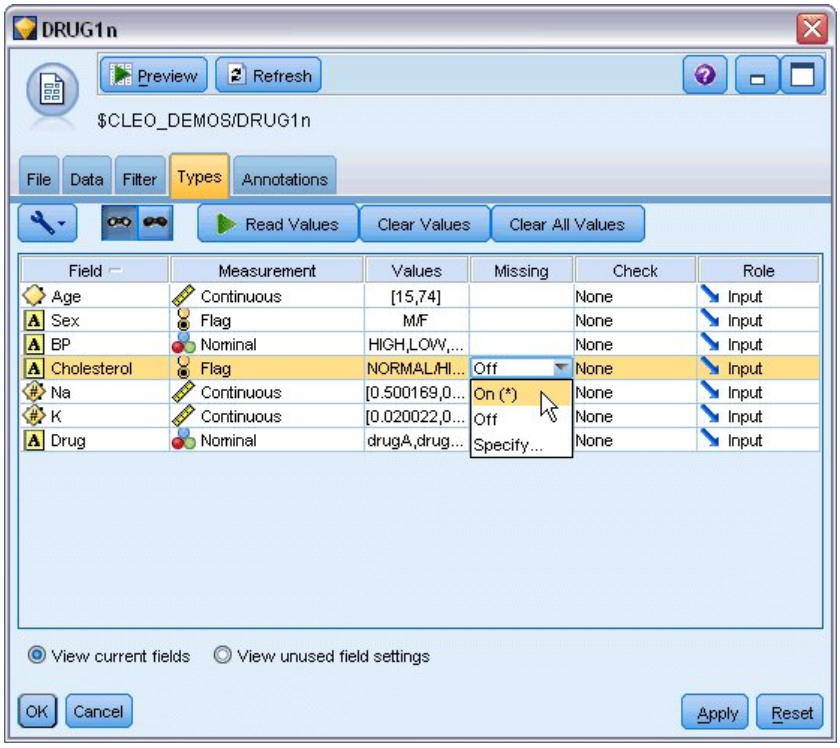


그림 *80.* 유형 탭에서 값 옵션 선택

데이터 탭을 클릭하여 필드에 대한 **저장 공간**을 대체하고 변경하십시오. 저장 공간은 데이터 필드의 측정 수준 또는 사용 유형인 **측정**과 다릅니다. **유형** 탭을 사용하면 데이터의 필드 유형에 대한 자세한 정보를 알 수 있습니다. 또한 **값 읽기**를 선택하면 사용자가 값 열에서 선택한 사항을 기반으로 하여 각 필드에 대한 실제 값을 볼 수 있습니다. 이 프로세스를 **인스턴스화**라고 합니다.

테이블 추가

이제 데이터 파일을 로드했으며 일부 레코드 값을 한 눈에 보려고 합니다. 한 가지 방법은 테이블 노드를 포함하는 스트림을 작성하는 것입니다. 스트림에 테이블 노드를 배치하려면 팔레트에서 아이콘을 두 번 클릭하거나 이를 끌어서 캔버스에 놓으십시오.

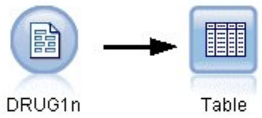
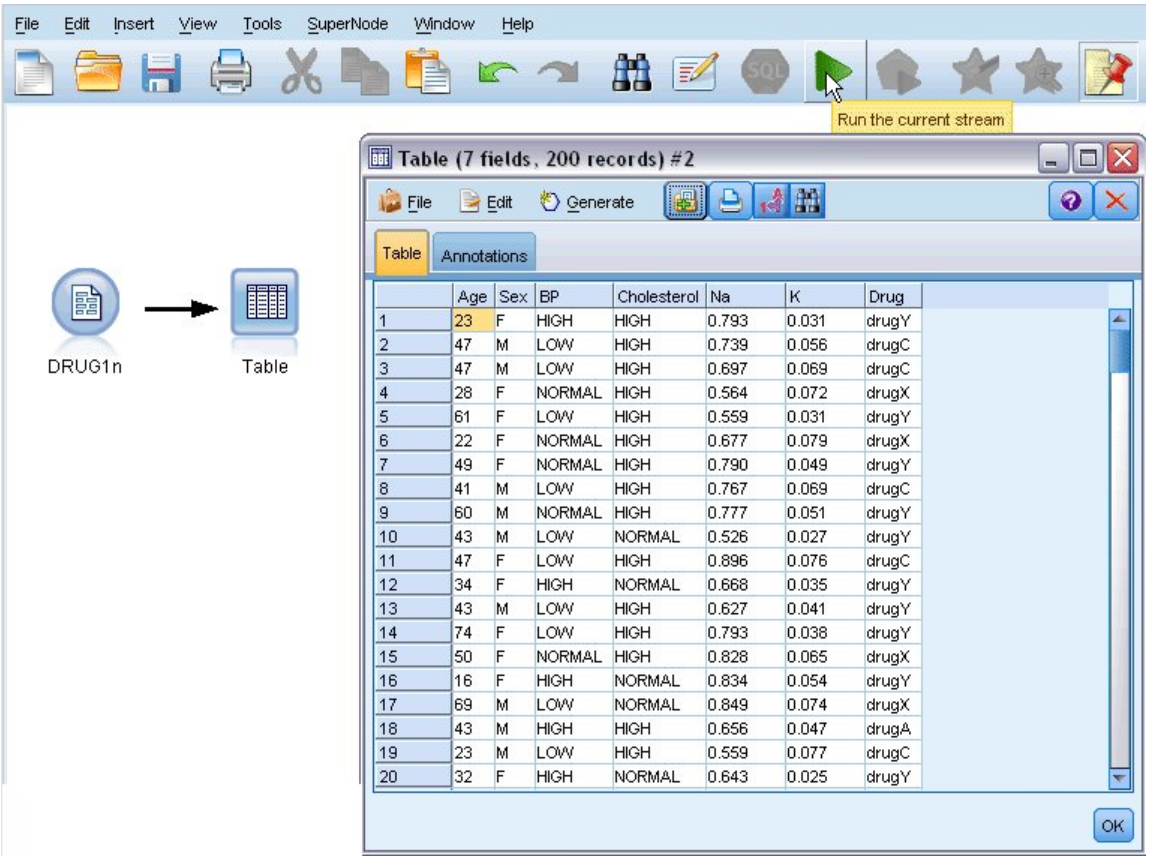


그림 *81.* 데이터 소스에 연결된 테이블 노드



	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0.793	0.031	drugY
2	47	M	LOW	HIGH	0.739	0.056	drugC
3	47	M	LOW	HIGH	0.697	0.069	drugC
4	28	F	NORMAL	HIGH	0.564	0.072	drugX
5	61	F	LOW	HIGH	0.559	0.031	drugY
6	22	F	NORMAL	HIGH	0.677	0.079	drugX
7	49	F	NORMAL	HIGH	0.790	0.049	drugY
8	41	M	LOW	HIGH	0.767	0.069	drugC
9	60	M	NORMAL	HIGH	0.777	0.051	drugY
10	43	M	LOW	NORMAL	0.526	0.027	drugY
11	47	F	LOW	HIGH	0.896	0.076	drugC
12	34	F	HIGH	NORMAL	0.668	0.035	drugY
13	43	M	LOW	HIGH	0.627	0.041	drugY
14	74	F	LOW	HIGH	0.793	0.038	drugY
15	50	F	NORMAL	HIGH	0.828	0.065	drugX
16	16	F	HIGH	NORMAL	0.834	0.054	drugY
17	69	M	LOW	NORMAL	0.849	0.074	drugX
18	43	M	HIGH	HIGH	0.656	0.047	drugA
19	23	M	LOW	HIGH	0.559	0.077	drugC
20	32	F	HIGH	NORMAL	0.643	0.025	drugY

그림 *82.* 도구 모음에서 스트림 실행

팔레트에서 노드를 두 번 클릭하면 해당 노드가 스트림 캔버스 내에서 선택된 노드에 자동으로 연결됩니다. 또는 노드가 아직 연결되지 않은 경우, 마우스 가운데 단추를 사용하여 소스 노드를 테이블 노드에 연결할 수 있습니다. 마우스 가운데 단추를 시뮬레이션하려면 마우스를 사용하는 동안 Alt 키를 아래로 누르십시오. 테이블을 보려면 도구 모음에서 녹색 화살표 단추를 클릭하여 스트림을 실행하거나 마우스 오른쪽 단추로 테이블 노드를 클릭하여 **실행**을 선택하십시오.

분포 그래프 작성

데이터 마이닝 동안, 시각적인 요약값을 작성하여 데이터를 탐색하는 것이 유용한 경우가 종종 있습니다. IBM SPSS Modeler에서는 사용자가 요약할 데이터의 유형에 따라 선택할 수 있는 다양한 유형의 여러 가지 그래프를 제공합니다. 예를 들어, 각 약물에 반응하는 환자의 비율을 찾으려면 분포 노드를 사용하십시오.

스트림에 분포 노드를 추가하고 이를 소스 노드에 연결한 다음 노드를 두 번 클릭하여 표시 옵션을 편집하십시오.

분포를 표시할 대상 필드로 *Drug*를 선택하십시오. 그런 다음 대화 상자에서 **실행**을 클릭하십시오.

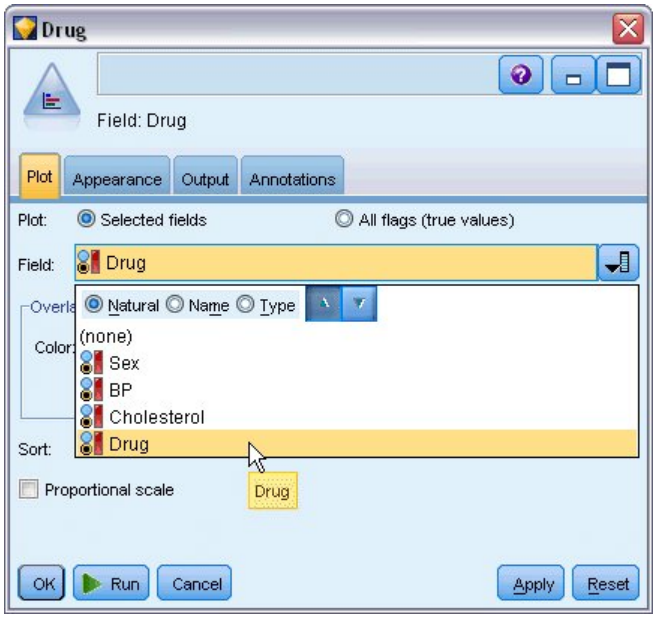


그림 *83.* 대상 필드로 *drug* 선택

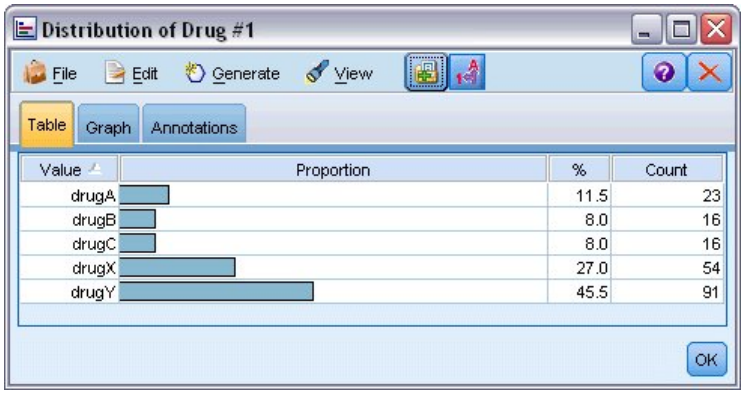


그림 *84.* 약물 유형에 대한 반응 분포

결과 그래프는 데이터의 "모양"을 파악하는 데 도움이 됩니다. 이 그래프는 환자들이 *Y* 약물에 가장 많이 반응하고 *B* 및 *C* 약물에 가장 덜 반응했음을 표시합니다.

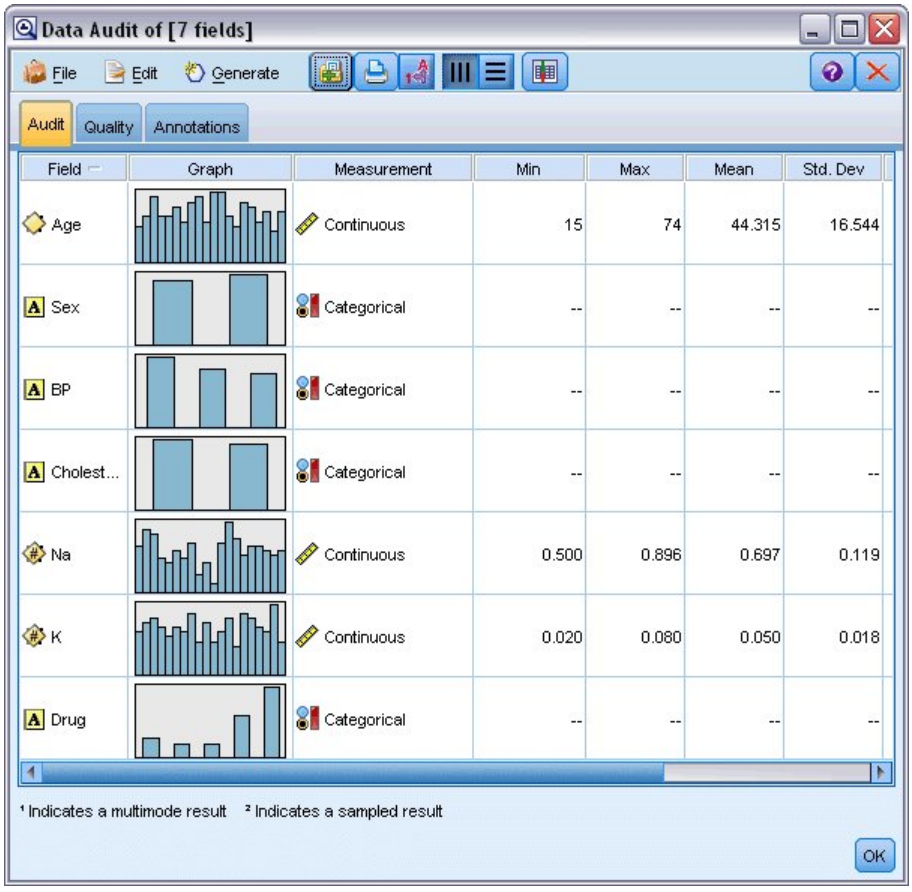


그림 *85.* 데이터 검토 결과

또는 모든 필드의 분포 및 히스토그램을 한 눈에 신속하게 파악하기 위해 데이터 검토 노드를 연결하고 실행할 수 있습니다. 데이터 검토 노드는 출력 탭에서 사용할 수 있습니다.

산점도 작성

이제 목표변수인 *Drug*에 영향을 미치는 요인에 대해 살펴보겠습니다. 연구원으로서 혈액 내의 나트륨 및 칼륨 농도가 중요 요인인 것을 알고 있습니다. 두 요인 모두 숫자 값이므로 약물 범주를 색상 오버레이로 사용하여 나트륨 대 칼륨의 산점도를 작성할 수 있습니다.

plot 노드를 작업공간에 배치하고 이를 소스 노드에 연결한 다음 두 번 클릭하여 노드를 편집하십시오.

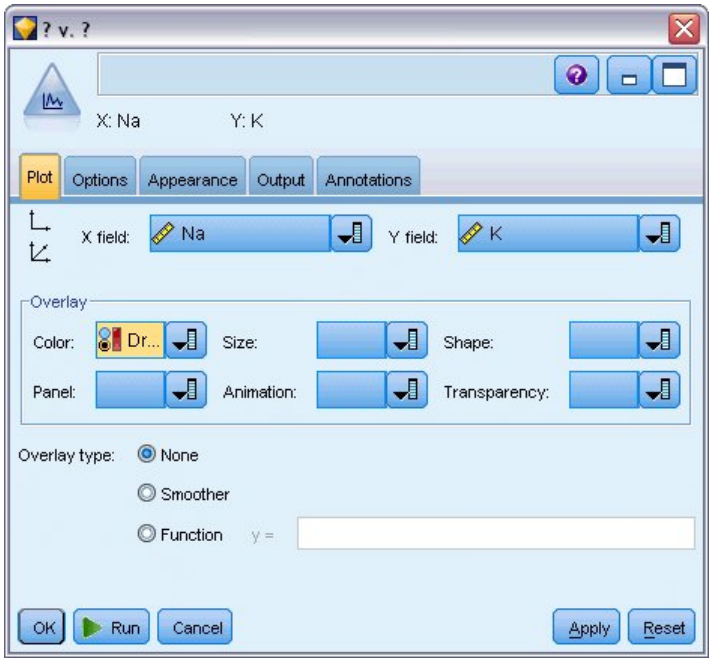


그림 *86.* 산점도 작성

도표 탭에서 *Na*를 X 필드로 선택하고 *K*를 Y 필드로 선택하고 *Drug*를 오버레이 필드로 선택하십시오. 그런 다음 **실행**을 클릭하십시오.

도표는 항상 해당 임계값 위에 있는 약물이 *Y* 약물이며 결코 해당 임계값 아래에 있지 않은 약물이 *Y* 약물임을 명확히 표시합니다. 이 임계값은 비율, 즉, 나트륨(*Na*) 대 칼륨(*K*)의 비율입니다.

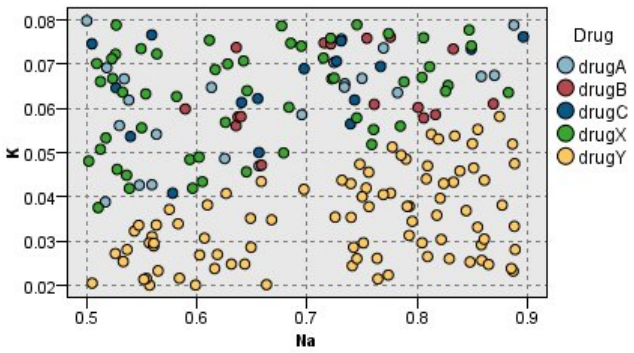


그림 *87.* 약물 분포 산점도

웹 그래프 작성

많은 데이터 필드가 범주형이므로 다른 범주 사이의 연관을 맵핑하는 웹 그래프 도표 작성을 시도할 수 있습니다. 웹 노드를 작업공간 내의 소스 노드에 연결하는 것부터 시작하십시오. 웹 노드 대화 상자에서 *BP*(혈압) 및 *Drug*를 선택하십시오. 그런 다음 **실행**을 클릭하십시오.

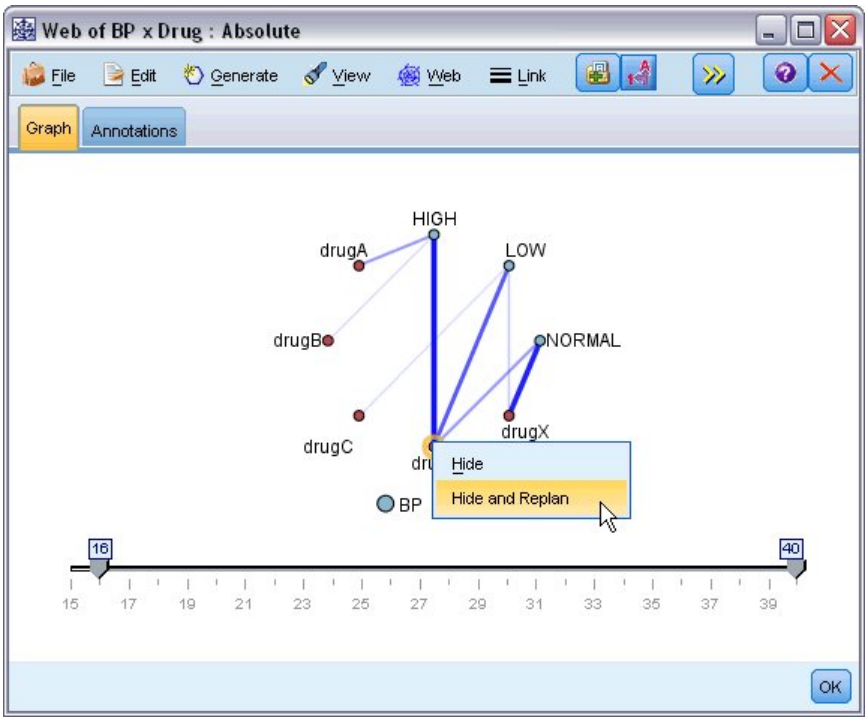


그림 *88.* 약물 대 혈압 웹 그래프

도표에서 *Y* 약물이 세 수준의 혈압 수준과 모두 연관되어 있음이 표시됩니다. 이는 놀라운 사실이 아닙니다. 이미 *Y* 약물이 최선임을 파악했기 때문입니다. 기타 약물에 초점을 맞추기 위해 *Y* 약물을 숨깁니다. **보기** 메뉴에서 **편집 모드**를 선택한 다음 마우스 오른쪽 단추로 *Y* 포인트를 클릭하고 **숨기기 및 재계획**을 선택할 수 있습니다.

단순화된 도표에서, *Y* 약물과 해당 링크가 모두 숨겨집니다. 이제 *A* 및 *B* 약물만 고혈압과 연관되었음을 명확히 알 수 있습니다. *C* 및 *X* 약물만 저혈압과 연관됩니다. 정상 혈압은 *X* 약물과만 연관됩니다. 이 포인트에서는 지정된 환자에 대해 *A* 및 *B* 약물 사이 또는 *C* 및 *X* 약물 사이에서 선택하는 방법을 알지 못합니다. 여기서 모델링이 도움이 될 수 있습니다.

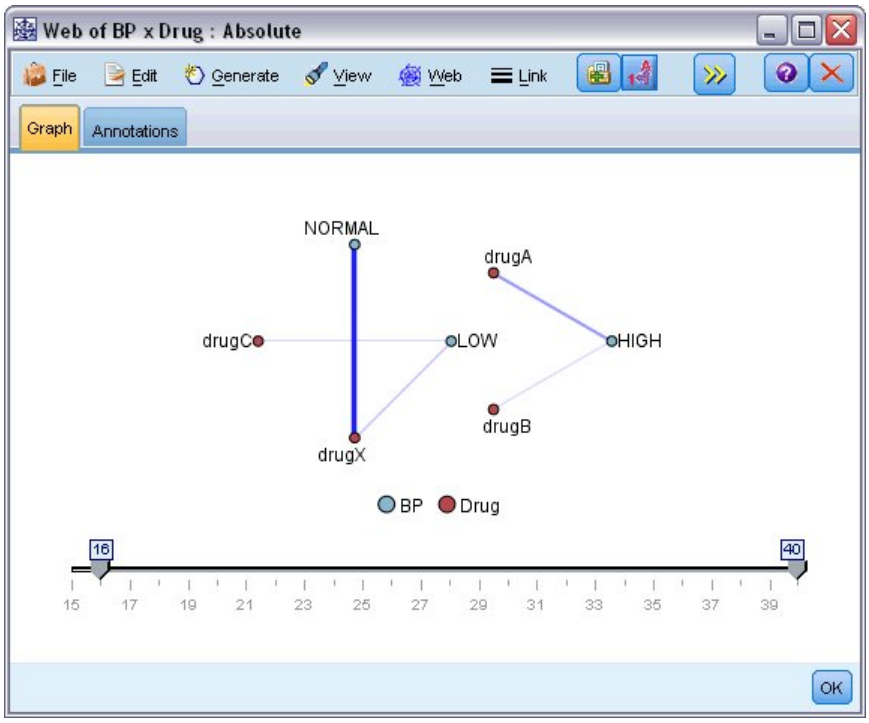


그림 *89. Y* 약물 및 링크가 숨겨진 웹 그래프

새 필드 파생

나트륨 대 칼륨의 비율이 *Y* 약물을 사용할 시기를 예측할 수 있는 것으로 보이므로 각 레코드에 대해 이 비율의 값을 포함하는 필드를 파생할 수 있습니다. 나중에 다섯 가지의 약물 중 각 약물을 사용할 시기를 예측하는 모델을 작성할 때 이 필드가 유용할 수 있습니다. 스트림 레이아웃을 단순화하기 위해 DRUG1n 소스 노드를 제외한 모든 노드를 삭제하는 것부터 시작합니다. 파생 노드(필드 작업 탭)를 DRUG1n에 연결한 다음 파생 노드를 두 번 클릭하여 이를 편집하십시오.

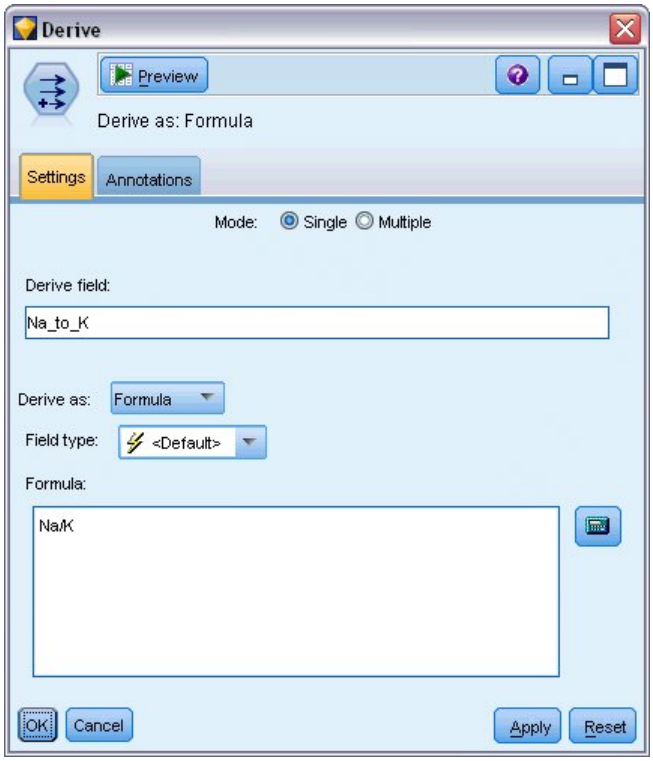


그림 *90.* 파생 노드 편집

새 필드의 이름을 *Na_to_K*로 지정하십시오. 나트륨 값을 칼륨 값으로 나누어서 새 필드를 얻으므로 수식에 대해 `Na/K`를 입력하십시오. 또한 필드의 바로 오른쪽에 있는 아이콘을 클릭하여 수식을 작성할 수도 있습니다. 그러면 표현식 작성기가 열리고 함수, 피연산자, 필드 및 해당 값의 내장된 목록을 사용하여 대화형으로 표현식을 작성할 수 있습니다.

히스토그램 노드를 파생 노드에 연결하여 사용자의 새 필드의 분포를 확인할 수 있습니다. 히스토그램 노드 대화 상자에서 도표로 표시될 필드로 *Na_to_K*를 지정하고 오버레이 필드로 *Drug*를 지정하십시오.

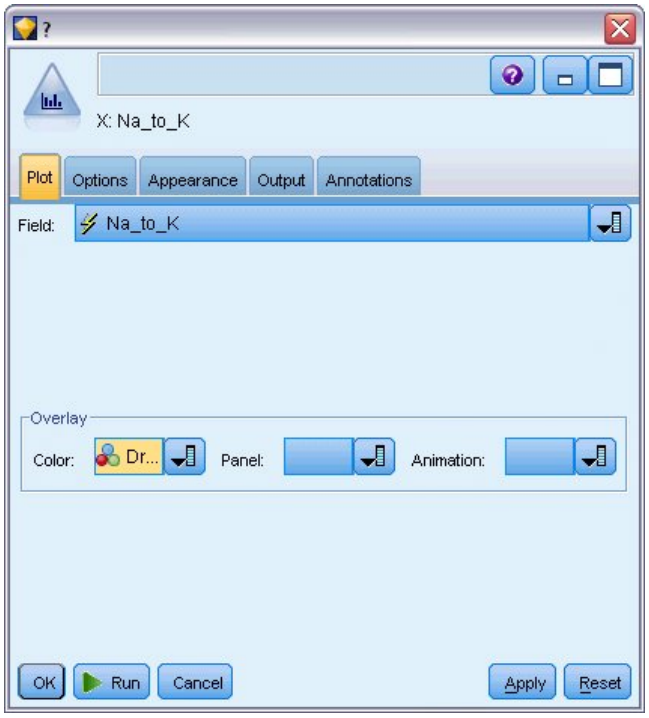


그림 *91.* 히스토그램 노드 편집

스트림을 실행하면 여기에 표시된 그래프를 얻을 수 있습니다. 표시에 따라 *Na_to_K* 값이 약 15 이상일 때 *Y* 약물이 선택 약물이 되는 것으로 결론 내릴 수 있습니다.

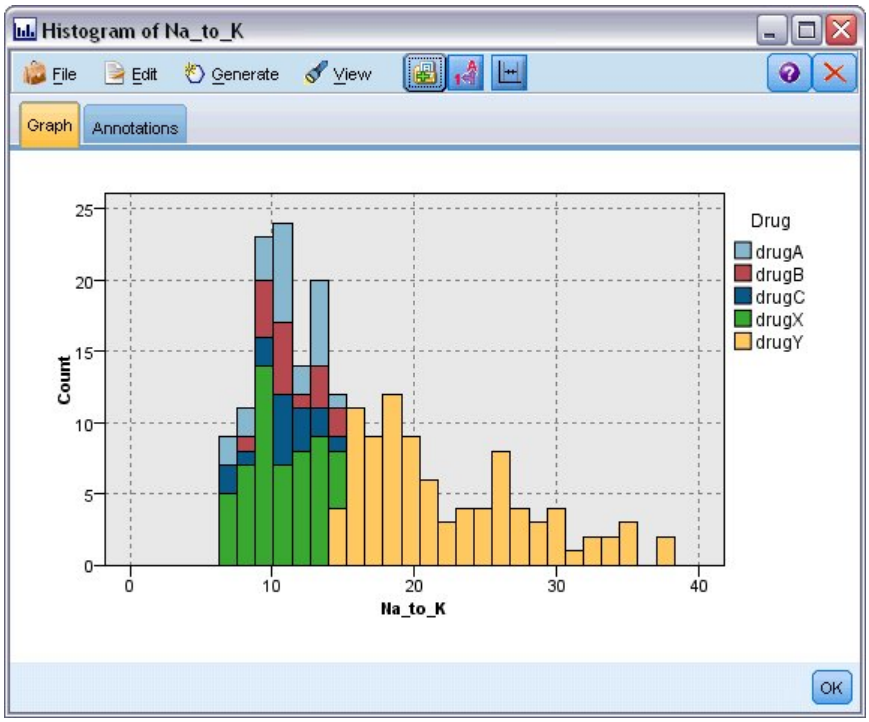


그림 *92.* 히스토그램 표시

모델 작성

데이터를 탐색하고 조작하여 일부 가설을 형성할 수 있었습니다. 혈압과 마찬가지로 혈액 내 나트륨 대 칼륨의 비율이 약물 선택에 영향을 미치는 것처럼 보입니다. 그러나 아직은 모든 관계를 완전히 설명할 수 없습니다. 여기서 모델링이 몇 가지 대답을 제공할 수 있습니다. 이 경우, 규칙 작성 모델, C5.0을 사용하여 데이터를 맞추려고 시도할 것입니다.

파생된 필드인 *Na_to_K*를 사용하고 있으므로 원본 필드인 *Na* 및 *K*가 모델링 알고리즘에서 두 번 사용되지 않도록 필터링할 수 있습니다. 필터 노드를 사용하여 이 작업을 수행할 수 있습니다.

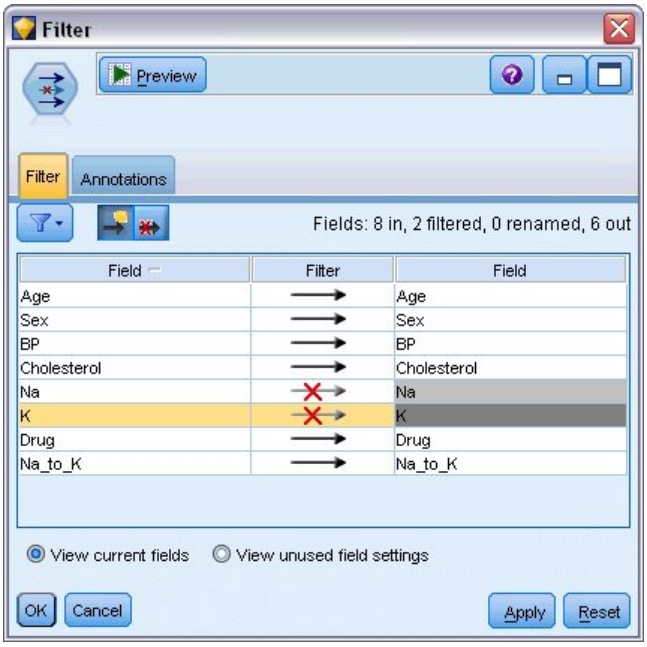


그림 *93.* 필터 노드 편집

필터 탭에서 *Na* 및 *K* 옆의 화살표를 클릭하십시오. 필드가 이제 필터링되었음을 표시하기 위해 빨강 X가 화살표 위에 표시됩니다.

그런 다음 필터 노드에 연결된 유형 노드를 첨부하십시오. 유형 노드를 사용하면 사용 중인 필드 유형 및 결과 예측에 사용된 방법을 표시할 수 있습니다.

유형 탭에서 *Drug* 필드에 대한 역할을 **대상**으로 설정하십시오. 즉, *Drug*가 예측할 필드임을 표시합니다. 기타 필드에 대한 역할은 예측변수로 사용되도록 **입력**으로 설정된 상태로 두십시오.

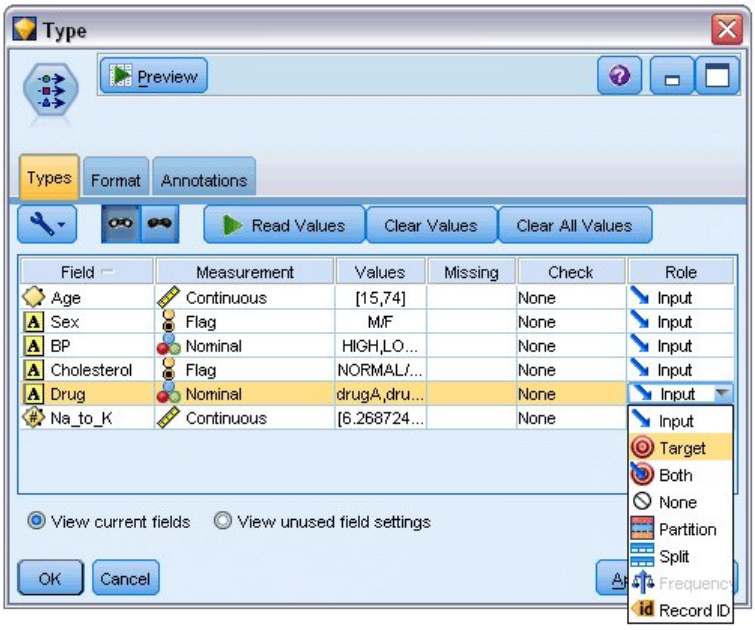


그림 *94.* 유형 노드 편집

모델을 추정하려면 표시된 대로 C5.0 노드를 작업공간에 배치하고 이를 스트림의 끝에 첨부하십시오. 그런 다음 녹색 **실행** 도구 모음 단추를 클릭하여 스트림을 실행하십시오.

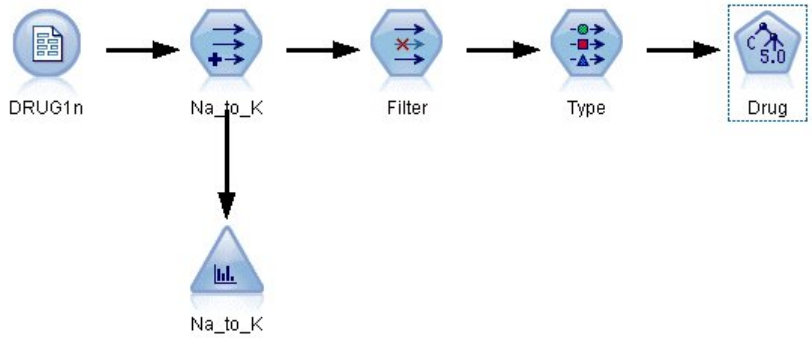


그림 *95. C5.0* 노드 추가

모델 찾아보기

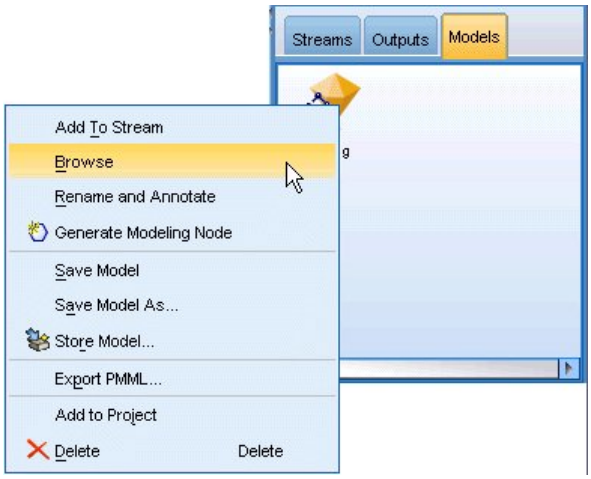


그림 *96.* 모델 찾아보기

C5.0 노드가 실행될 때 모델 너깃이 스트림에 추가되고 창의 오른쪽 상단 코너에 있는 모델 팔레트에도 추가됩니다. 모델을 찾아보려면 오른쪽 마우스 단추로 아무 아이콘이나 클릭하고 컨텍스트 메뉴에서 **편집** 또는 **찾아보기**를 선택하십시오.

규칙 브라우저는 C5.0 노드에 의해 생성된 규칙 세트를 의사결정 트리 형식으로 표시합니다. 처음에는 트리가 접혀 있습니다. 트리를 펼치려면 **모두** 단추를 클릭하여 모든 수준을 표시하십시오.

그림 *97.* 규칙 브라우저

이제 퍼즐의 누락된 부분을 볼 수 있습니다. *Na*-대-*K* 비율이 14.64 미만이며 고혈압인 사람의 경우, 나이가 약물 선택을 결정합니다. 저혈압인 사람의 경우, 콜레스테롤 수준이 최선의 예측변수인 것처럼 보입니다.

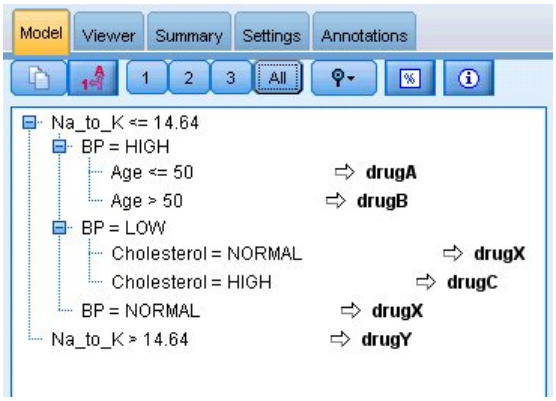


그림 *98.* 완전히 펼쳐진 규칙 브라우저

뷰어 탭을 클릭하면 동일한 의사결정 트리를 좀 더 자세한 그래프로 볼 수 있습니다. 여기서는 각 혈압 범주에 대한 케이스 수 및 케이스 퍼센트를 더 쉽게 볼 수 있습니다.

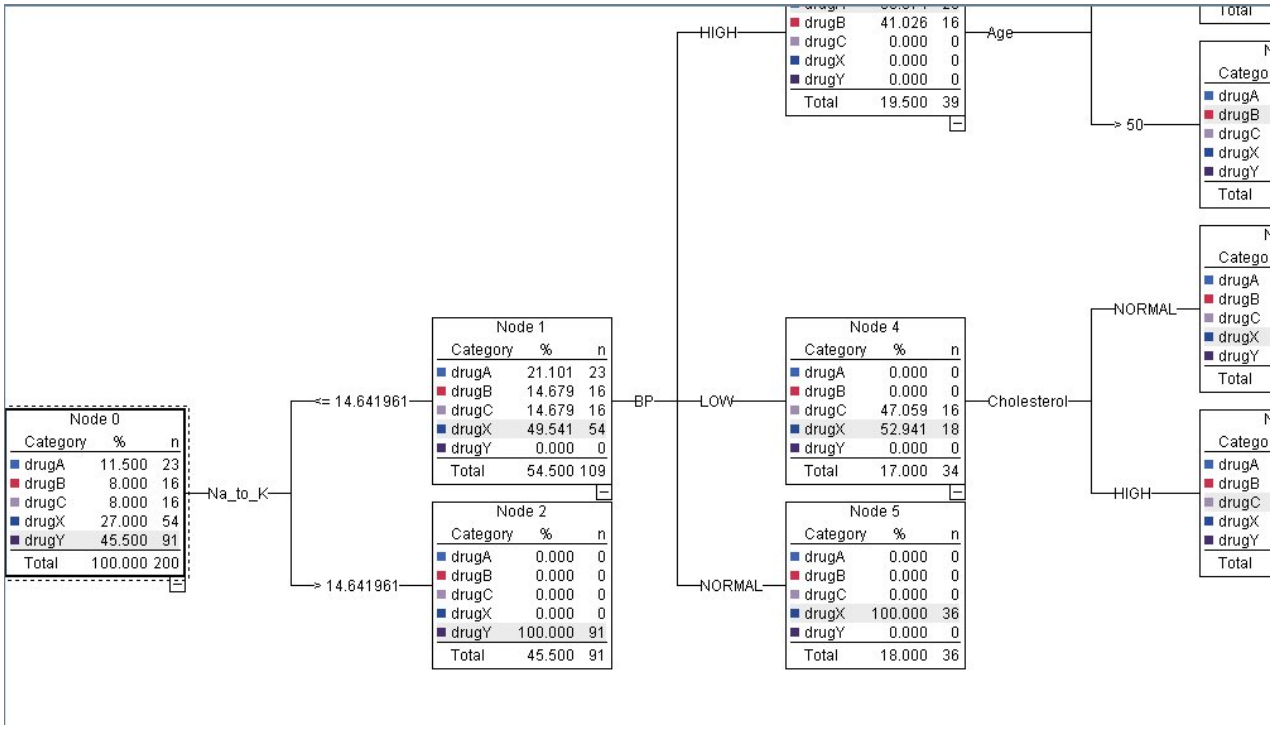


그림 *99.* 그래프 형식의 의사결정 트리

분석 노드 사용

분석 노드를 사용하여 모델의 정확도를 평가할 수 있습니다. (출력 노드 팔레트의) 분석 노드를 모델 너깃에 첨부하고 분석 노드를 연 다음 **실행**을 클릭하십시오.

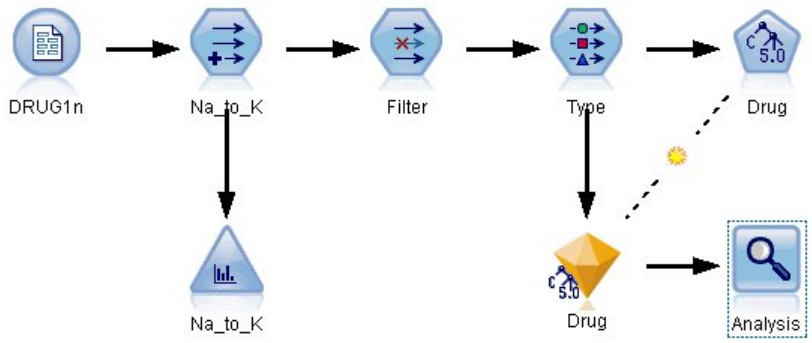


그림 *100.* 분석 노드 추가

분석 노드 출력은 이 인공 데이터 세트를 사용하여 모델이 데이터 세트 내의 모든 레코드에 대해 약물 선택을 정확히 예측했음을 표시합니다. 실제 데이터 세트를 사용하여 100% 정확도를 얻는 것은 거의 불가능하나 분석 노드를 사용하면 모델이 사용자의 특정 애플리케이션에 대해 허용 가능한 정확도인지 판별하는 데 도움을 얻을 수 있습니다.

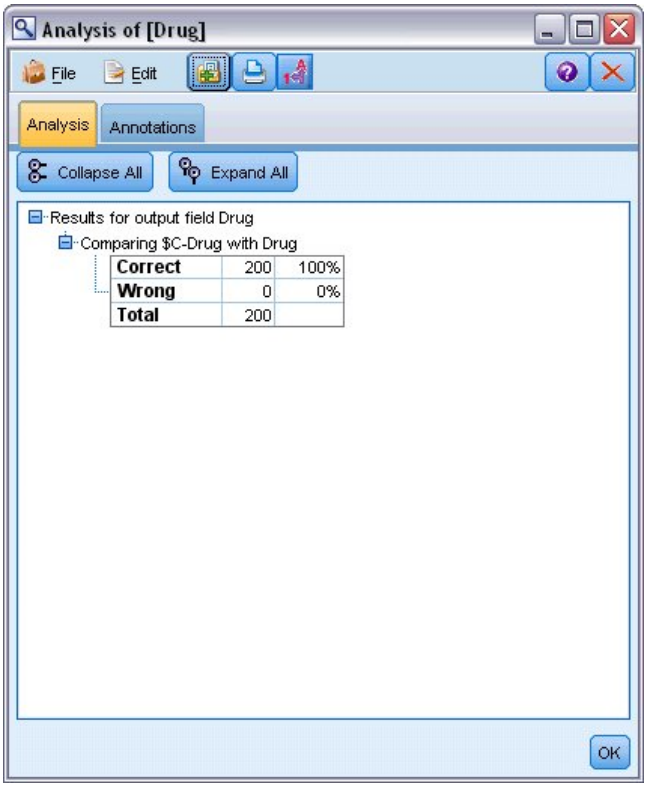


그림 *101.* 분석 노드 출력

제 9 장 예측변수 선별(필드선택)

필드선택 노드는 사용자가 특정 결과를 예측할 때 가장 중요한 필드를 식별하는 데 도움을 줍니다. 필드선택 노드는 수백 또는 수천 개의 예측변수군에서 가장 중요한 예측변수를 선별하고 순위를 매기고 선택합니다. 궁극적으로 더 적은 수의 예측변수를 사용하며 더 빠르게 실행되며 더 쉽게 이해할 수 있는 더 빠르고 더 효율적인 모델을 갖게 될 것입니다.

이 예에 사용된 데이터는 가상의 통신회사에 대한 데이터 웨어하우스를 나타내며 회사 고객 5,000명이 특별 프로모션에 보인 반응에 대한 정보를 포함합니다. 이때 데이터에는 고객의 나이, 고용, 수입, 통신 사용 통계량을 포함하는 여러 필드가 있습니다. 세 개의 "대상" 필드는 세 가지 각 제안에 고객이 반응하는지 여부를 보여줍니다. 회사는 이 데이터를 사용하여 고객이 향후에 유사한 제안에 반응할 가능성을 예측할 수 있습니다.

이 예에서는 *customer_dbase.sav*라는 데이터 파일을 참조하는 *featureselection.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *featureselection.str* 파일은 *streams* 디렉토리에 있습니다.

이 예에서는 대상으로 한 오퍼에만 초점을 맞춥니다. CHAID 트리 작성 노드를 사용하여 어떤 고객이 프로모션에 반응할 가능성이 가장 높은지 설명하는 모델을 개발합니다. 여기에는 두 가지 접근법이 있습니다.

- 필드선택을 사용하지 않는 방법입니다. 데이터 세트 내의 모든 예측변수 필드가 CHAID 트리에 대한 입력으로 사용됩니다.
- 필드선택을 사용하는 방법입니다. 상위 10개의 예측변수를 사용하기 위해 필드선택 노드가 사용됩니다. 그런 다음 CHAID 트리에 입력됩니다.

두 결과 트리 모델을 비교하여 필드선택이 효율적인 결과를 생성함을 알 수 있습니다.

스트림 작성

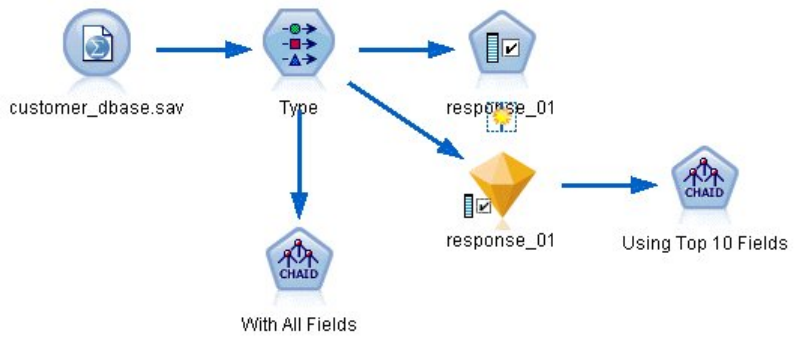


그림 *102.* 필드선택 예 스트림

1. 통계 파일 소스 노드를 공백 스트림 캔버스에 배치하십시오. 이 노드를 IBM SPSS Modeler 설치의 *Demos* 디렉토리에서 사용 가능한 *customer_dbase.sav* 예 데이터 파일로 지정하십시오. (또는 *streams* 디렉토리에서 예 스트림 파일 *featureselection.str*을 여십시오.)
2. 유형 노드를 추가하십시오. 유형 탭에서 아래쪽으로 스크롤하여 *response_01*에 대한 역할을 대상으로 변경하십시오. 또한 목록 위쪽의 고객 ID(*custid*) 및 기타 반응 필드(*response_02* 및 *response_03*)에 대한 역할을 없음으로 변경하십시오. 모든 기타 필드에 대해서는 역할을 입력으로 설정된 상태로 두고 **값 읽기** 단추를 클릭한 다음 **확인**을 클릭하십시오.

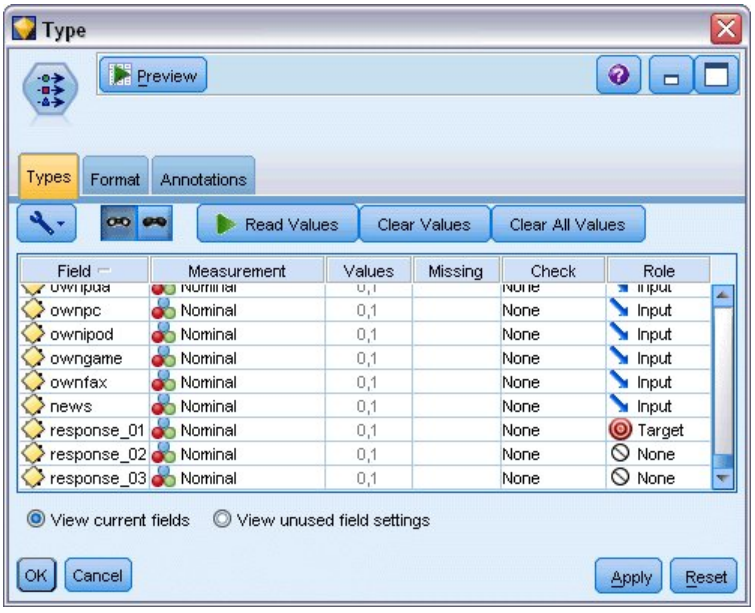


그림 *103.* 유형 노드 추가

3. 필드선택 모델링 노드를 스트림에 연결하십시오. 이 노드에서 필드를 선별하거나 자격이 없는 것으로 식별하기 위한 규칙 및 기준을 지정할 수 있습니다.
4. 스트림을 실행하여 필드선택 모델 너깃을 작성하십시오.

5. 스트림 또는 모델 팔레트에서 마우스 오른쪽 단추로 모델 너깃을 클릭하고 **편집** 또는 **찾아보기**를 선택하여 결과를 보십시오.

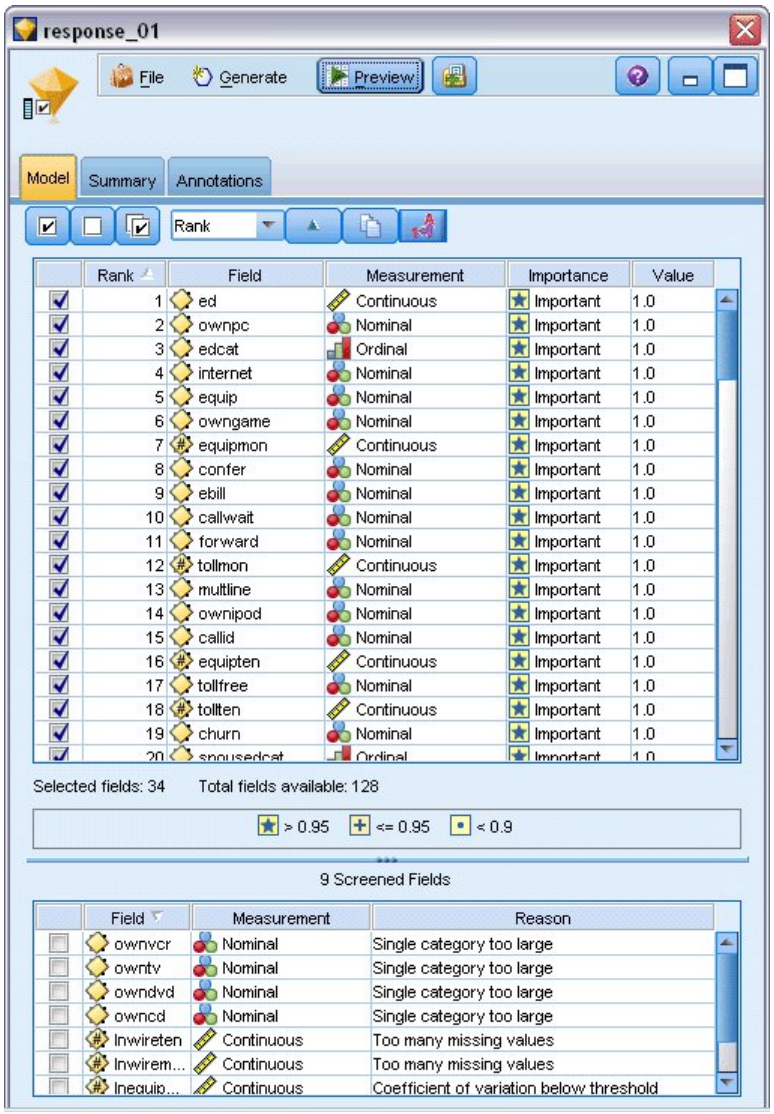


그림 *104.* 필드선택 모델 너깃의 모델 탭

위쪽 패널에는 예측에 유용한 것으로 판별된 필드가 표시됩니다. 이러한 필드는 중요도를 기준으로 순위가 매겨집니다. 아래쪽 패널에는 분석에 사용되지 않도록 선별되는 필드 및 이유가 표시됩니다. 위쪽 필드를 탐색하여 후속 모델링 세션에 사용할 필드를 결정할 수 있습니다.

6. 이제 다운스트림을 사용할 필드를 선택할 수 있습니다. 원래 34개의 필드가 중요한 것으로 식별되었으나 예측변수군을 더 줄이고자 합니다.
7. 첫 번째 열의 확인 표시를 사용하여 위쪽에서 10개의 예측변수만 선택하여 원하는 않는 예측변수를 선택 취소하십시오. (11행에서 확인 표시를 클릭하고 Shift 키를 아래로 누른 상태에서 34행의 확인 표시를 클릭하십시오.) 모델 너깃을 닫으십시오.

8. 필드선택 없이 결과를 비교하려면 하나는 필드선택을 사용하고 다른 하나는 사용하지 않는 두 개의 CHAID 모델링 노드를 추가해야 합니다.
9. 한 CHAID 노드를 유형 노드에 연결하고 다른 하나는 필드선택 모델 너깃에 연결하십시오.
10. 각 CHAID 노드를 열고 옵션 작성 탭을 선택하고 목적 분할창에서 **새 모델 작성**, **단일 트리 작성** 및 **대화형 세션 시작**이 선택되어 있는지 확인하십시오.

기본 분할창에서 **최대 트리 깊이**가 5로 설정되어 있는지 확인하십시오.

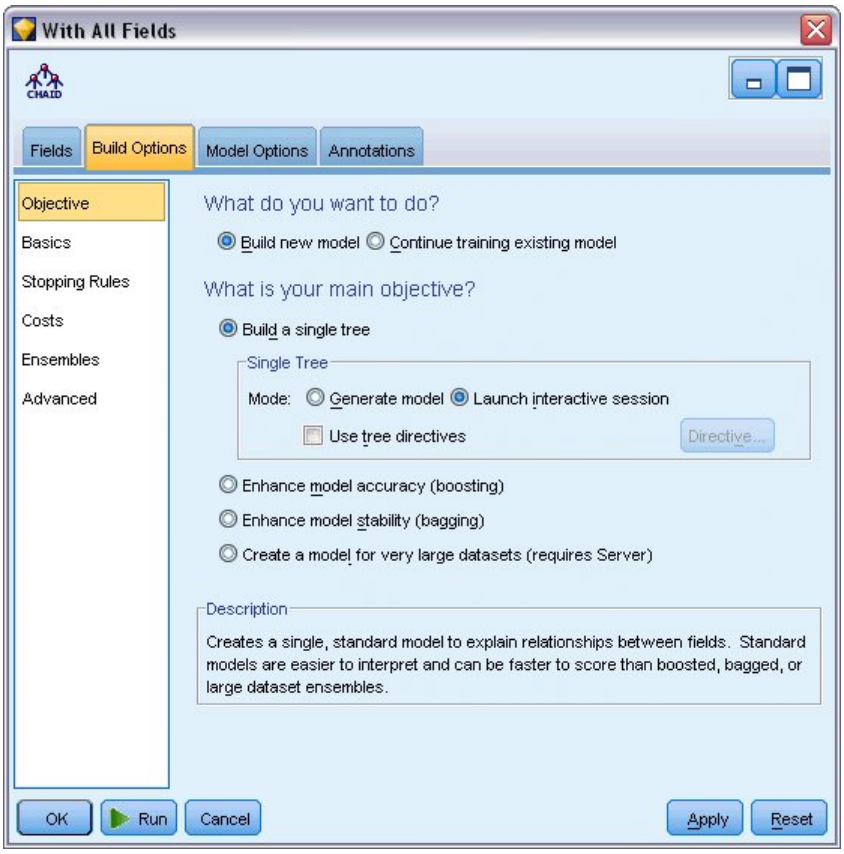


그림 *105. CHAID* 모델링 노드의 모든 예측변수 필드에 대한 목적 설정

모델 작성

1. 유형 노드에 연결된 데이터 세트 내의 모든 예측변수를 사용하는 CHAID 노드를 실행하십시오. 노드가 실행됨에 따라 실행에 시간이 얼마나 많이 걸리는 지 깨닫게 될 것입니다. 결과 창에 테이블이 표시됩니다.
2. 메뉴에서 **트리 > 트리 증가**를 선택하여 트리를 증가시키고 펼쳐진 트리를 표시하십시오.

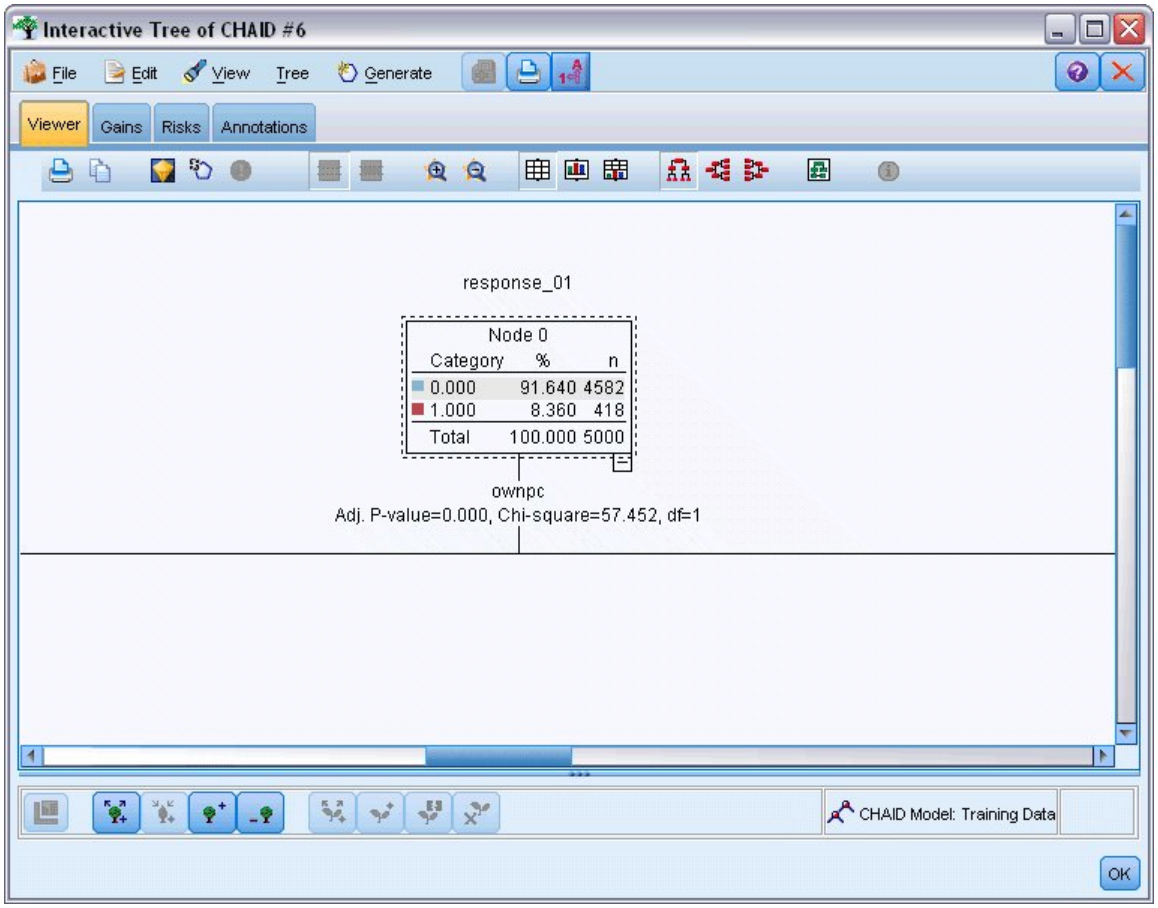


그림 *106.* 트리 작성기에서 트리 증가

3. 이제 10개의 예측변수만 사용하는 다른 CHAID 노드에 대해 동일한 작업을 수행하십시오. 트리 작성기를 열 때 다시 트리를 증가하십시오.

두 번째 모델은 첫 번째 모델보다 더 빨리 실행되어야 합니다. 이 데이터 세트가 매우 작기 때문에 실행 시간의 차이는 몇 초 정도일 수 있습니다. 그러나 훨씬 더 큰 실세계 데이터 세트의 경우, 차이가 훨씬 커서 몇 분에서 몇 시간이 될 수도 있습니다. 필드선택을 사용하면 처리 속도가 상당히 빨라집니다.

또한 두 번째 트리는 첫 번째 트리보다 더 적은 수의 트리 노드를 포함합니다. 이해하기에 더 쉽습니다. 그러나 이 트리를 사용하도록 결정하기 전에 해당 트리가 효율적인지 여부 및 모든 예측변수를 사용하는 모델과 비교할 방법을 찾아야 합니다.

결과 비교

두 결과를 비교하려면 유효성 측도가 필요합니다. 이를 위해 트리 작성기의 이익 탭을 사용할 것입니다. 데이터 세트 내의 모든 레코드와 비교할 때 노드 내의 레코드가 목표 범주에 속할 확률을 측정하는 **리프트**를 살펴 볼 것입니다. 예를 들어, 리프트 값이 148%인 경우 노드의 레코드가 데이터 세트 내의 모든 레코드에 비해 목표 범주에 포함될 가능성이 1.48배임을 의미합니다. 리프트는 이익 탭의 지수 열에서 표시됩니다.

1. 전체 예측 변수군에 대한 트리 작성기에서 이익 탭을 클릭하십시오. 목표 범주를 1.0으로 변경하십시오. 먼저 사분위수 도구 모음 단추를 클릭하여 표시를 사분위수로 변경하십시오. 그런 다음 드롭다운 목록에서 이 단추의 오른쪽에 있는 **사분위수**를 선택하십시오.
2. 다음 그림에서 보듯이 유사한 두 개의 비교 가능한 이익 테이블을 갖게 되도록 트리 작성기에서 10 개의 예측변수군에 대해 이 프로시저를 반복하십시오.

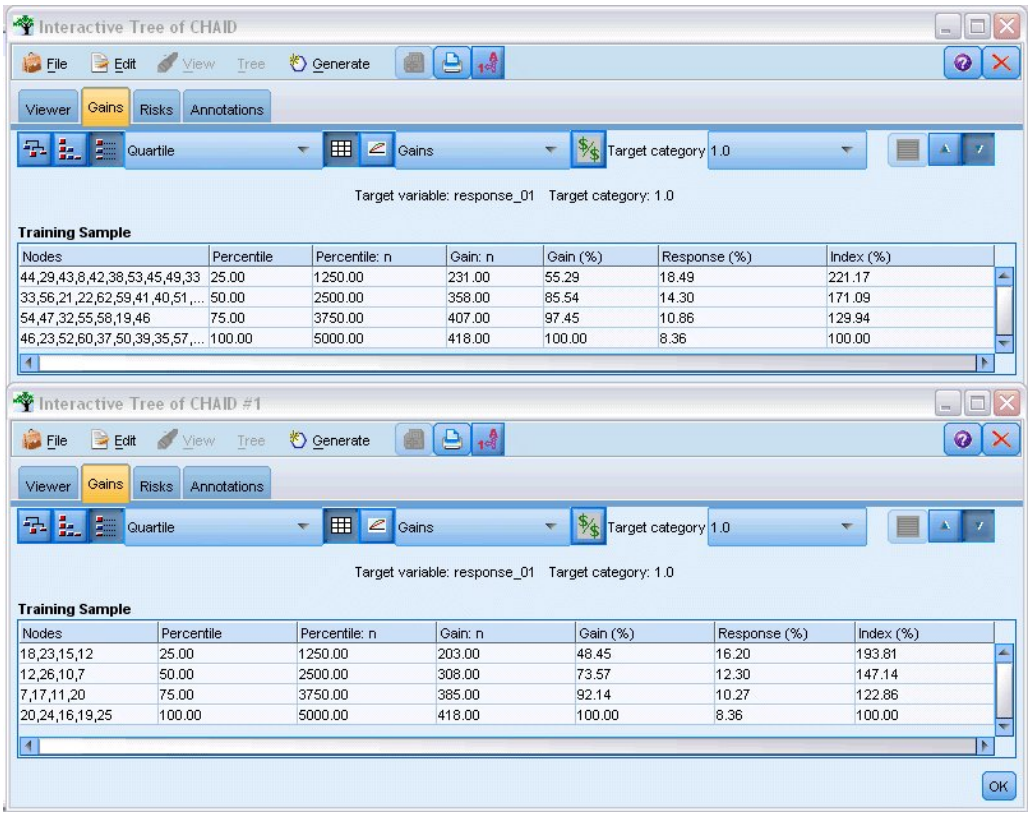


그림 *107.* 두 개의 *CHAID* 모델에 대한 이익 차트

각 이익 테이블은 트리에 대한 터미널 노드를 사분위수로 그룹화합니다. 두 모델의 유효성을 비교하려면 각 테이블의 위쪽 사분위수에서 리프트(지수 값)를 보십시오.

모든 예측변수가 포함된 경우, 모델이 221%의 리프트를 표시합니다. 즉, 이러한 노드에서 공정특성 변수가 있는 케이스가 목표 프로모션에 응답할 가능성이 2.2배임을 나타냅니다. 이러한 공정특성 변수를 보려면 위쪽 행에서 클릭하여 선택하십시오. 그런 다음 뷰어 탭으로 전환하면 해당 노드가 검은색으로

윤관석이 표시됩니다. 트리를 따라 강조표시된 각 터미널 노드로 내려가서 예측변수가 어떻게 분할되었는지 보십시오. 위쪽 사분위수 단독으로 10개의 노드를 포함합니다. 실세계 스코어링 모델로 변환될 때 10 개의 서로 다른 고객 프로파일을 관리하는 것이 힘들 수 있습니다.

(필드선택에 의해 식별된 대로) 상위 10개의 예측변수만 포함되면 리프트가 거의 194%입니다. 이 모델이 모든 예측변수를 사용하는 모델만큼 좋지는 않으나 유용한 것은 확실합니다. 여기서, 위쪽 사분위수에는 네 개의 노드만 포함하므로 훨씬 단순합니다. 따라서 필드선택 모델이 모든 예측변수를 사용하는 모델보다 선호된다고 판별할 수 있습니다.

요약

필드선택의 장점을 검토하려고 합니다. 더 적은 수의 예측변수를 사용하면 더 적은 비용이 듭니다. 즉, 데이터를 더 적게 수집하고 처리하고 모델에 피드할 수 있습니다. 계산 시간이 줄어듭니다. 이 예에서는 추가 필드선택 단계를 통해 적은 수의 예측변수군으로도 모델 작성이 현저히 빨라집니다. 더 큰 실세계 데이터 세트를 사용하는 경우, 시간 절약이 더 증가할 수 있습니다.

더 적은 수의 예측변수를 사용하면 스코어링이 단순해집니다. 예에서 보듯이 프로모션에 반응할 가능성이 높은 고객에 대해 네 가지의 프로파일만 식별합니다. 더 많은 수의 예측변수를 사용하면 모델이 과적합될 위험이 있습니다. 더 단순한 모델이 기타 데이터 세트에도 더 잘 일반화될 수 있습니다. (단, 이를 확인하기 위해서는 검증을 수행해야 합니다.)

필드선택 작업을 수행하기 위해 트리가 사용자를 대신하여 가장 중요한 예측변수를 식별할 수 있는 트리 작성 알고리즘을 사용했을 수 있습니다. 실제로 이 목적으로 CHAID 알고리즘이 자주 사용되며 트리를 수준별로 증가시켜 깊이 및 복잡도를 제어할 수도 있습니다. 그러나 필드선택 노드가 더 빠르며 사용하기 쉽습니다. 여기서는 한 단계 내에서 신속하게 모든 예측변수의 순위를 매기므로 사용자가 신속하게 가장 중요한 필드를 식별할 수 있습니다. 또한 포함할 예측변수의 수를 사용자가 변경할 수 있습니다. 10개 대신 15개 또는 20개의 예측변수를 사용하여 이 예를 쉽게 다시 실행하고 결과를 비교하여 최적의 모델을 판별할 수 있습니다.

제 10 장 입력 데이터 문자열 길이 감소(재분류 노드)

입력 데이터 문자열 길이 감소(재분류)

이항 로지스틱 회귀분석 및 이항 로지스틱 회귀분석 모델을 포함하는 자동 분류자 모델의 경우, 문자열 필드가 최대 여덟 자로 제한됩니다. 문자열이 여덟 자를 넘는 경우, 재분류 노드를 사용하여 다시 코딩될 수 있습니다.

이 예에서는 *drug_long_name*이라는 데이터 파일을 참조하는 *reclassify_strings.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *reclassify_strings.str* 파일은 *streams* 디렉토리에 있습니다.

이 예에서는 너무 긴 문자열에서 생성될 수 있는 오류의 종류를 표시하고 재분류 노드를 사용하여 문자열 세부사항을 허용되는 길이로 변경하는 방법을 설명하기 위해 스트림의 작은 부분에 초점에 맞춥니다. 이 예에서는 이항 로지스틱 회귀분석 노드를 사용하지만 이항 로지스틱 회귀분석 모델을 생성하기 위해 자동 분류자 노드를 사용하는 경우에도 동일하게 적용 가능합니다.

데이터 재분류

1. 가변파일 소스 노드를 사용하여 *Demos* 폴더의 *drug_long_name* 데이터 세트에 연결하십시오.

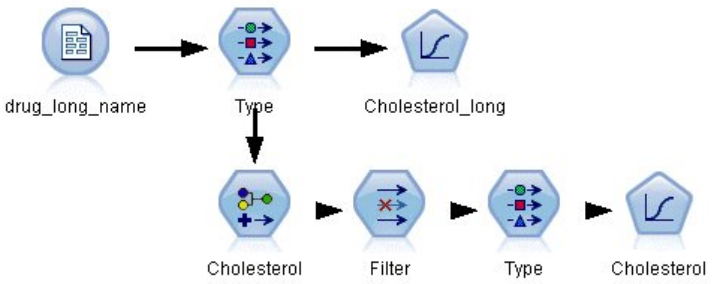


그림 *108.* 이항 로지스틱 회귀분석에 대한 문자열 재분류를 표시하는 샘플 스트림

2. 유형 노드를 소스 노드에 연결하고 **Cholesterol_long**을 대상으로 선택하십시오.
3. 로지스틱 회귀분석 노드를 유형 노드에 추가하십시오.
4. 로지스틱 회귀분석 노드에서 모델 탭을 클릭하여 **이항** 프로시저를 선택하십시오.

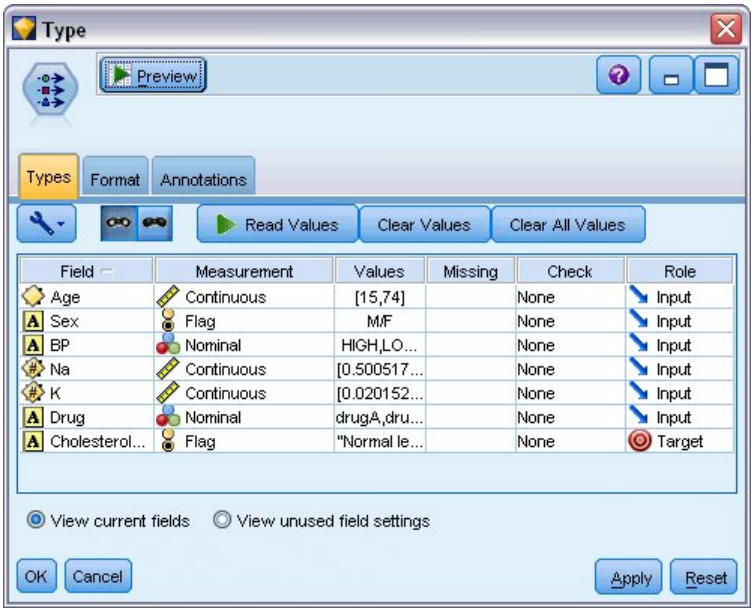


그림 *109. "Cholesterol_long"* 필드의 긴 문자열 세부사항

5. *reclassify_strings.str*에서 로지스틱 회귀분석 노드를 실행할 때 **Cholesterol_long** 문자열 값이 너무 길다는 점을 경고하는 오류 메시지가 표시됩니다.

   이 유형의 오류 메시지가 발생하면 이 예의 나머지 부분에서 설명하는 프로시저에 따라 데이터를 수정하십시오.

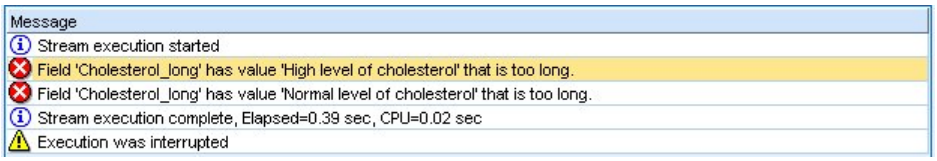


그림 *110.* 이항 로지스틱 회귀분석 노드를 실행할 때 표시되는 오류 메시지

6. 재분류 노드를 유형 노드에 연결하십시오.
7. 재분류 필드에서 **Cholesterol_long**을 선택하십시오.
8. 새 필드 이름으로 **Cholesterol**을 입력하십시오.
9. **가져오기** 단추를 클릭하여 **Cholesterol_long** 값을 원래 값 열에 추가하십시오.
10. 새 값 열에서, **High level of cholesterol**이라는 원래 값 옆에 **High**를 입력하고 **High level of cholesterol** 원래 값 옆에 **Normal**을 입력하십시오.

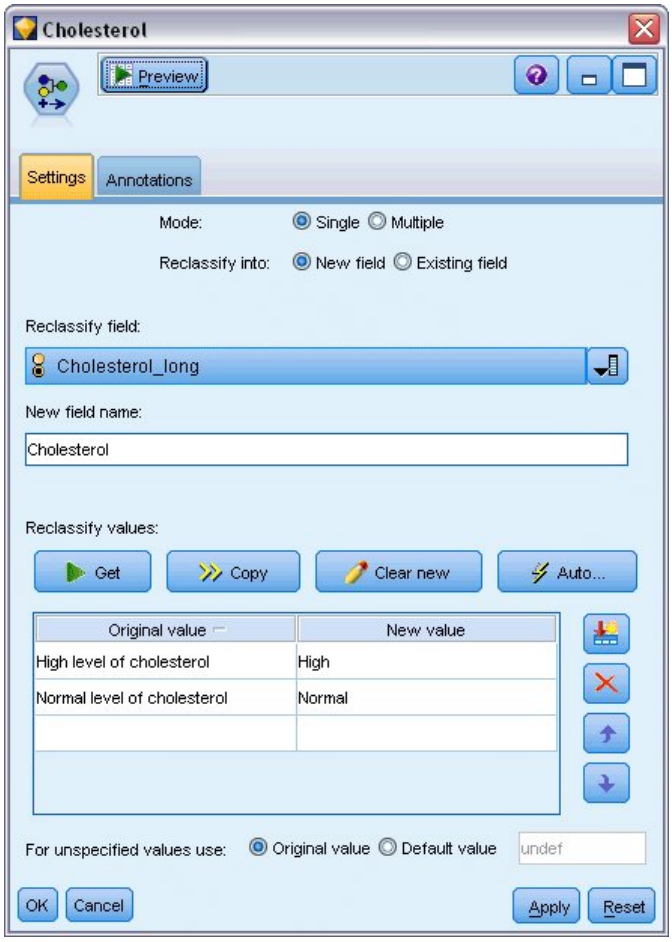


그림 *111.* 긴 문자열 재분류

11. 필터 노드를 재분류 노드에 연결하십시오.
12. 필터 열에서 **Cholesterol_long**을 클릭하여 제거하십시오.

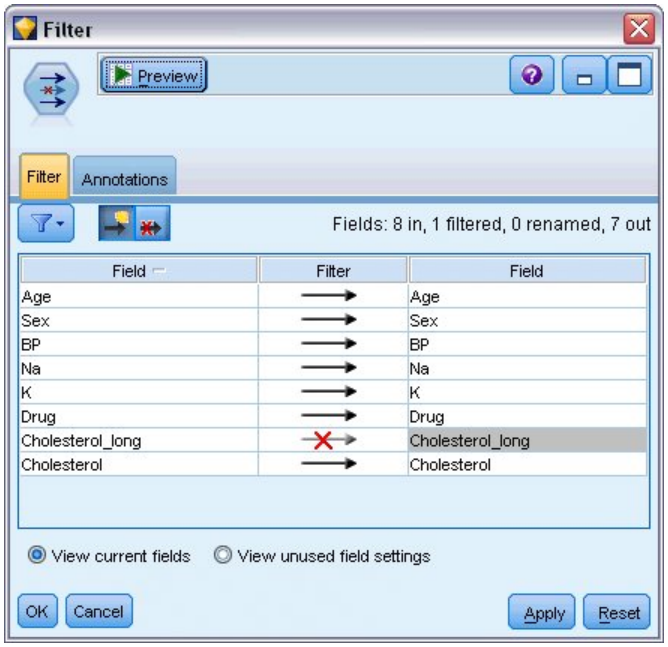


그림 *112.* 데이터에서 *"Cholesterol_long"* 필드 필터링

13. 유형 노드를 필터 노드에 연결하고 **Cholesterol**을 대상으로 선택하십시오.

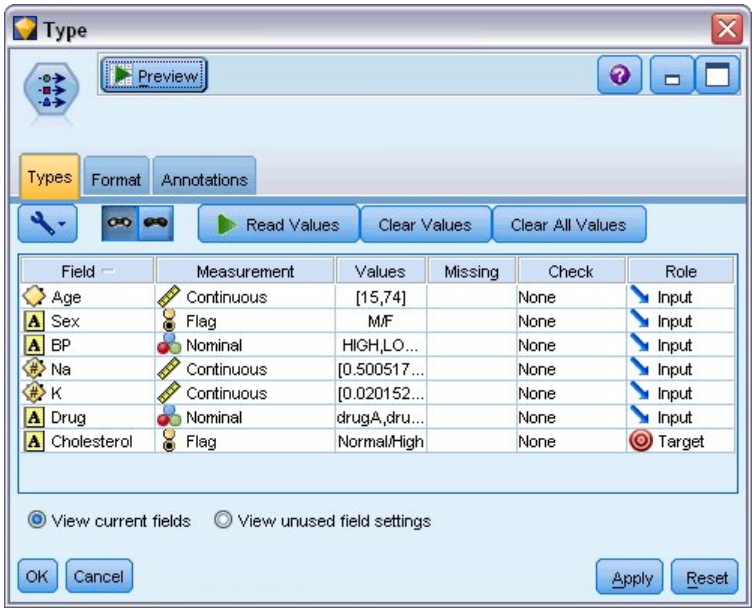


그림 *113. "Cholesterol"* 필드의 짧은 문자열 세부사항

14. 로지스틱 노드를 유형 노드에 추가하십시오.
15. 로지스틱 노드에서 모델 탭을 클릭하여 **이항** 프로시저를 선택하십시오.
16. 이제 이항 로지스틱 노드를 실행하여 오류 메시지를 표시하지 않고 모델을 생성할 수 있습니다.

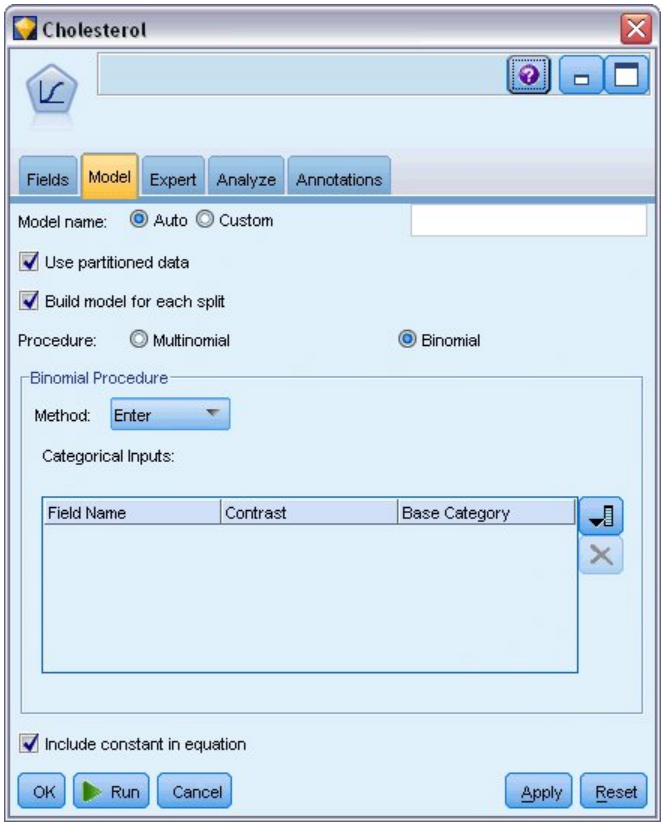


그림 *114.* 프로시저로 이항 선택

이 예는 스트림의 일부만 표시합니다. 긴 문자열을 재분류해야 하는 경우에 스트림의 유형에 대한 추가 정보가 필요하면 다음 예를 사용할 수 있습니다.

- 자동 분류자 노드입니다. 자세한 정보는 43 페이지의 『모델링 고객 반응(자동 분류자)』의 내용을 참조하십시오.
- 이항 로지스틱 회귀분석 노드입니다. 자세한 정보는 149 페이지의 제 13 장 『통신 서비스 제공자를 바꾸는 고객(이항 로지스틱 회귀분석)』의 내용을 참조하십시오.

IBM SPSS Modeler 사용 방법에 대한 자세한 정보(사용자 안내서, 노드 참조서 및 알고리즘 안내서 등)는 설치 디스크의 *Documentation* 디렉토리에서 사용 가능합니다.

제 11 장 고객 반응 모델링(의사결정 목록)

의사결정 목록 알고리즘은 지정된 이분형(예 또는 아니오) 결과의 더 높거나 낮은 우도를 표시하는 규칙을 생성합니다. 의사결정 목록 모델은 콜센터 또는 마케팅 애플리케이션 등의 고객 관계 관리에 광범위하게 사용됩니다.

이 예는 현재 각 고객에 올바른 오퍼를 매치하여 미래 마케팅 캠페인에서 보다 수익성이 좋은 결과를 산출하고자 하는 금융 회사를 기반으로 합니다. 특히, 예에서 의사결정 목록 모델을 사용하여 이전 홍보를 기반으로 가장 우호적으로 반응할 것 같은 고객 특성을 식별하고 결과에 따라 메일링 목록을 생성할 수 있습니다.

의사결정 목록 모델은 모델에서 모수를 조정하고 즉시 결과를 볼 수 있으므로 특히 대화형 모델링에 적합합니다. 자동으로 수많은 다른 모델을 작성하고 결과의 순위를 매길 수 있는 다른 접근법의 경우, 자동 분류자 노드를 대신 사용할 수 있습니다.

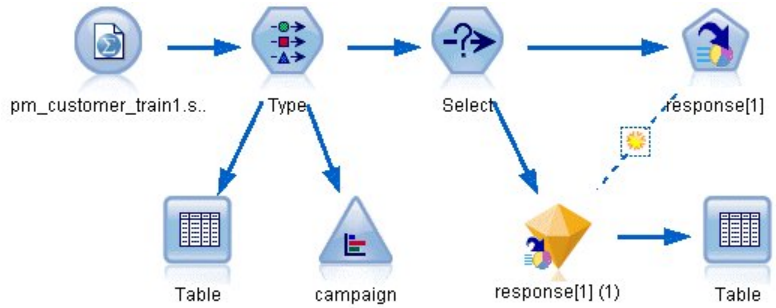
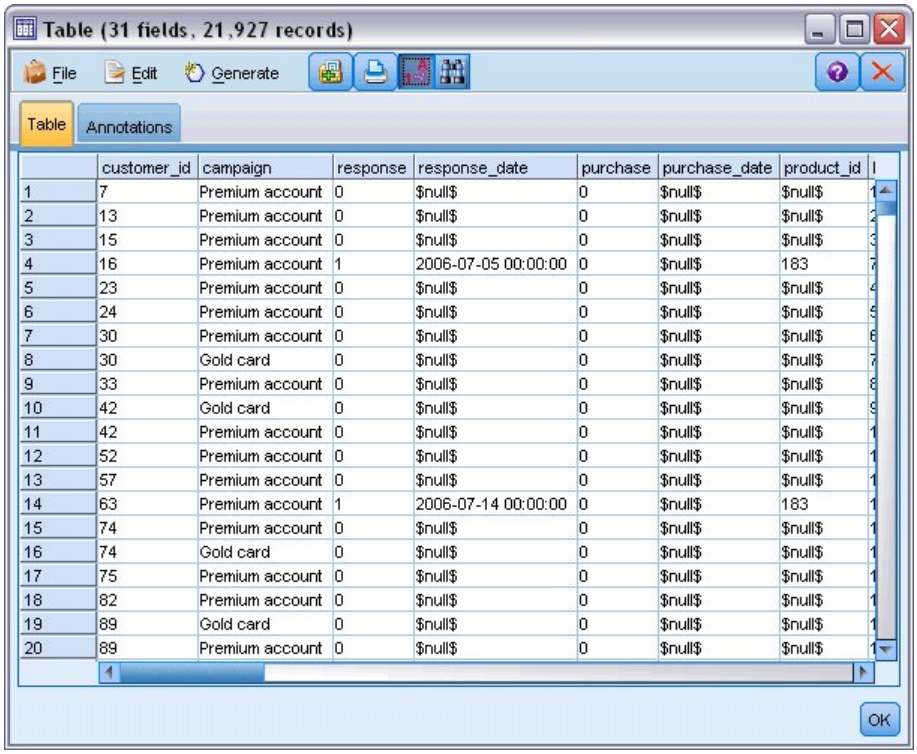


그림 *115.* 의사결정 목록 샘플 스트림

이 예에서는 *pm_customer_train1.sav* 데이터 파일을 참조하는 *pm_decisionlist.str* 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *pm_decisionlist.str* 파일은 *streams* 디렉토리에 있습니다.

히스토리 데이터

pm_customer_train1.sav 파일에는 *campaign* 필드의 값에 의해 표시되는 대로 과거 캠페인에서 특정 고객에 대해 수행된 오퍼를 추적하는 히스토리 데이터가 있습니다. 가장 많은 수의 레코드가 프리미엄 계정 캠페인에 해당됩니다.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold card	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold card	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold card	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold card	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

그림 *116.* 이전 프로모션에 관한 데이터

캠페인 필드의 값은 유형 노드에서 정의된 레이블을 사용하여 실제로 데이터에서 정수로 코딩됩니다. 예를 들어, *2* = 프리미엄 계정입니다. 도구 모음을 사용하여 테이블에서 값 레이블의 표시를 토글할 수 있습니다.

또한 파일에는 특정 공정특성 변수를 기반으로 하여 다른 그룹에 대한 반응률을 예측할 수 있는 모델을 작성하거나 "학습"하는 데 사용될 수 있는 각 고객에 대한 인구 통계학적 및 재정적 정보를 포함하는 수많은 필드가 포함될 수 있습니다.

스트림 작성

1. 사용자의 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *pm_customer_train1.sav*를 가리키는 통계량 파일 노드를 추가하십시오. 이 폴더를 참조하기 위한 단축키로 파일 경로에서 `$CLEO_DEMOS/`를 지정할 수 있습니다.

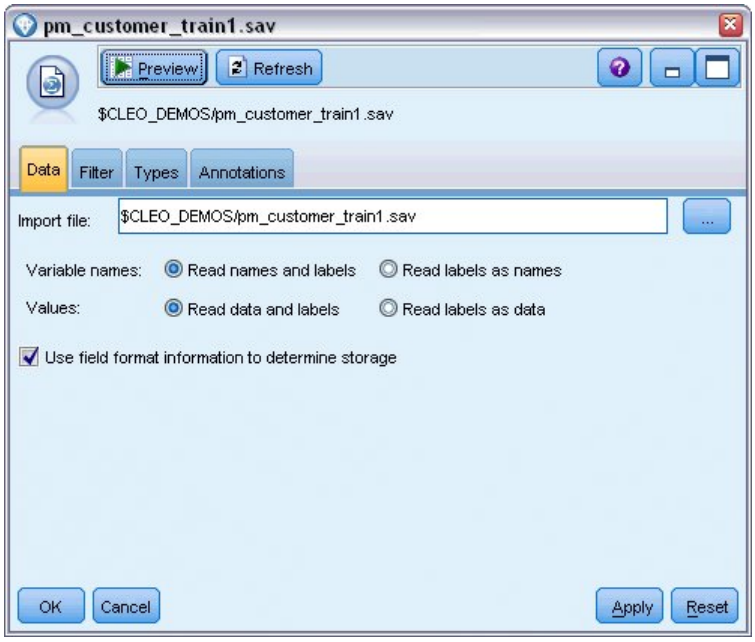


그림 *117.* 데이터에서 읽기

2. 유형 노드를 추가하고 대상 필드로 *response*를 선택하십시오(역할 = **대상**). 이 필드에 대한 측정 수준을 **플래그**로 설정하십시오.

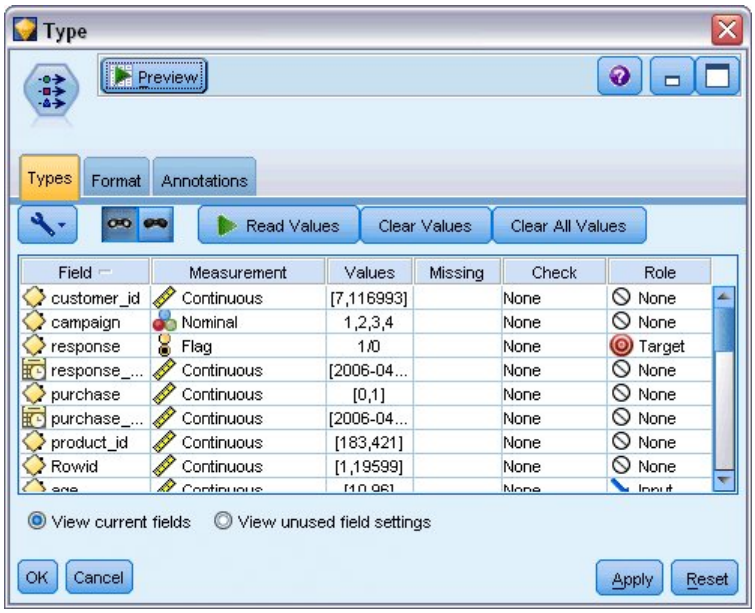


그림 *118.* 측정 수준 및 역할 설정

3. *customer_id*, *campaign*, *response_date*, *purchase*, *purchase_date*, *product_id*, *Rowid* 및 *X_random* 필드에 대한 역할을 **없음**으로 설정하십시오. 이러한 필드 모두 데이터에서는 사용되나 실제 모델 작성에는 사용되지 않습니다.
4. 유형 노드에서 **값 읽기** 단추를 클릭하여 값이 인스턴스화되도록 하십시오.

데이터에 네 가지 다른 캠페인에 대한 정보가 포함되나 한 번에 한 캠페인에 집중하여 분석을 수행할 것입니다. 가장 많은 레코드 수가 프리미엄 캠페인(데이터에서는 *campaign* = 2로 코딩됨)에 해당되므로 선택 노드를 사용하여 이러한 레코드만 스트림에 포함시킬 수 있습니다.

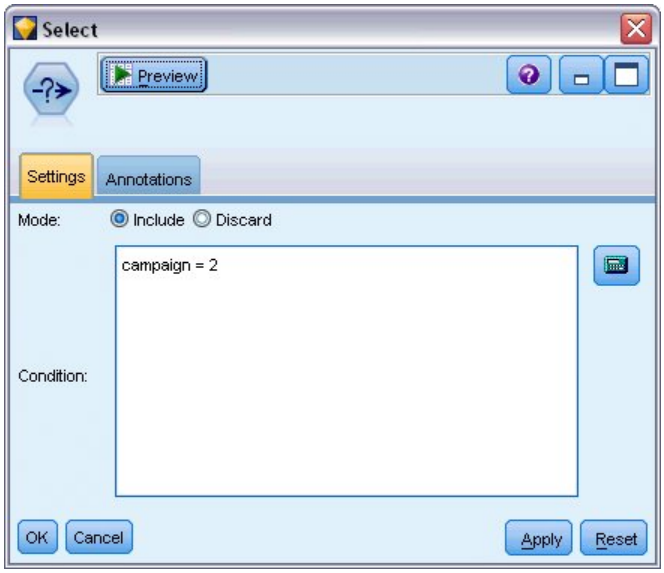


그림 *119.* 단일 캠페인용 레코드 선택

모델 작성

1. 스트림에 의사결정 목록 노드를 연결하십시오. 모델 탭에서 **목표 값**을 1로 설정하여 검색할 결과를 표시하십시오. 이 케이스에서는 이전 오퍼에 예라고 응답한 고객을 검색합니다.

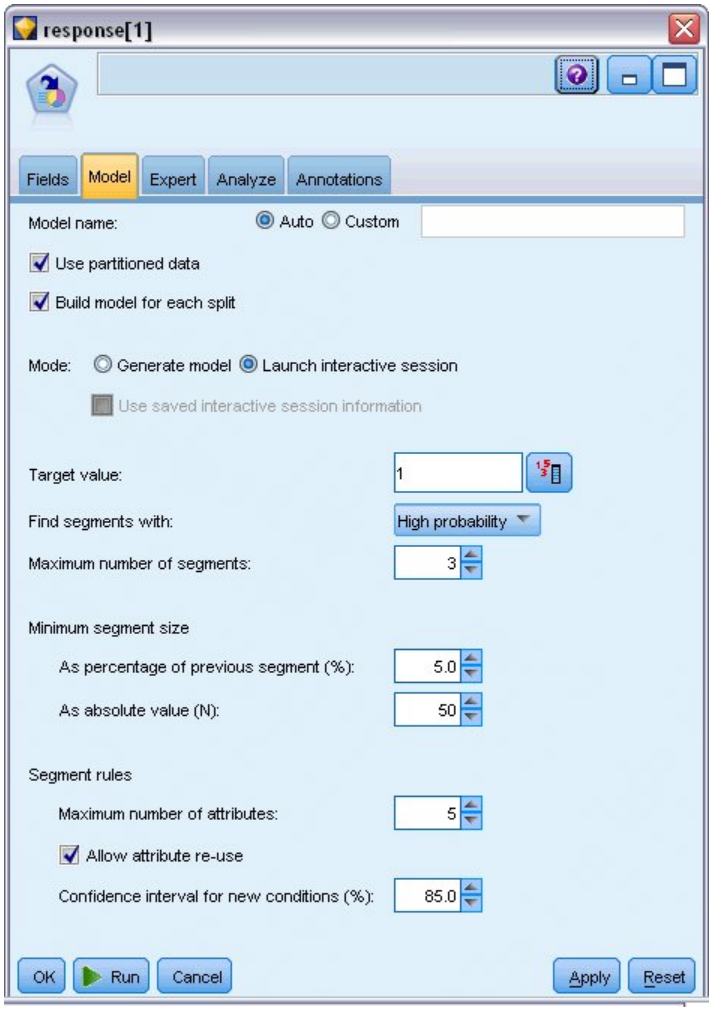


그림 *120.* 의사결정 목록 노드, 모델 탭

2. **대화형 세션 시작**을 선택하십시오.
3. 이 예를 위해서 모델을 단순한 상태로 유지하려면 최대 세그먼트 수를 3으로 설정하십시오.
4. 새 조건에 대한 신뢰구간을 85%로 변경하십시오.
5. 전문가 탭에서 **모드**를 **전문가**로 설정하십시오.

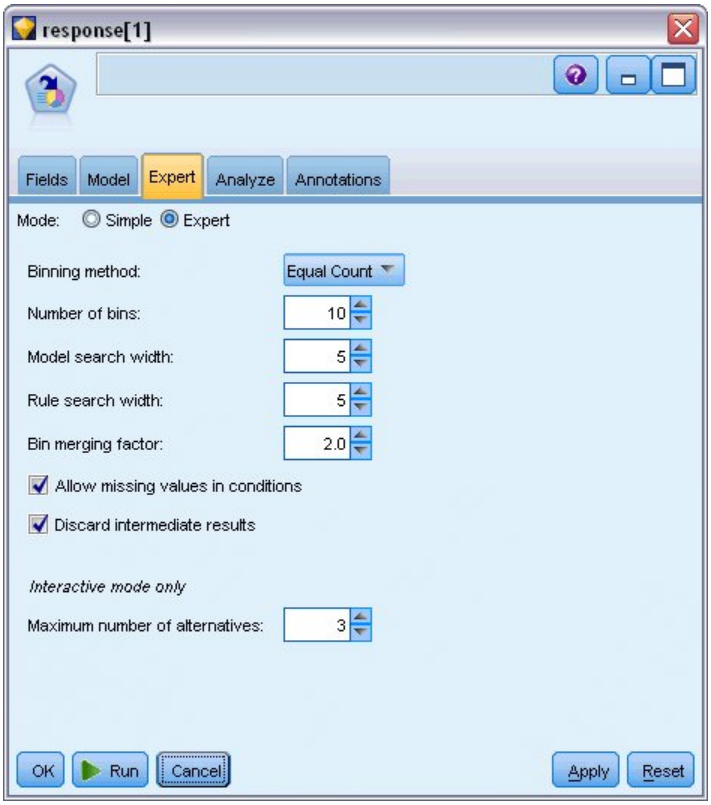


그림 *121.* 의사결정 목록 노드, 전문가 탭

6. **최대 대안 수**를 3으로 늘리십시오. 이 옵션은 사용자가 모델 탭에서 선택한 **대화형 세션 시작** 설정과 함께 작동합니다.
7. **실행**을 클릭하여 대화형 목록 뷰어를 표시하십시오.

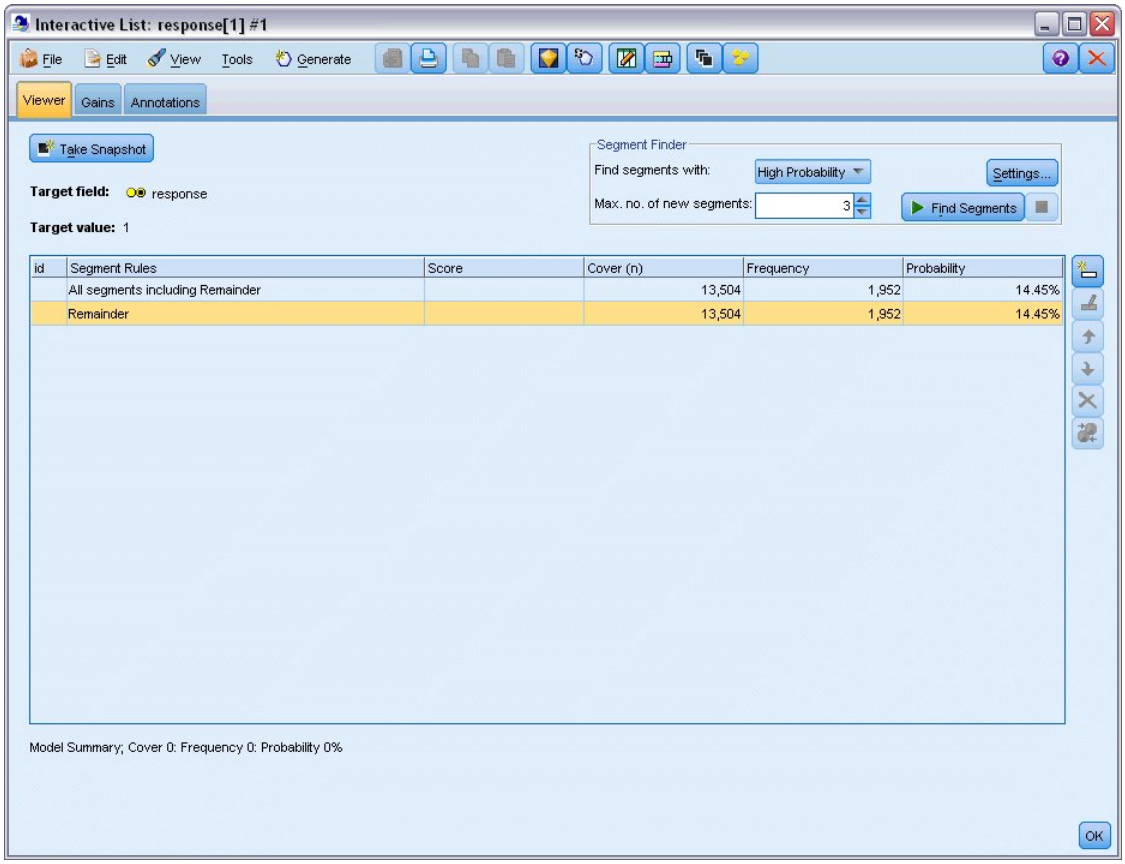


그림 *122.* 대화형 목록 뷰어

아직 세그먼트가 정의되지 않았으므로 모든 레코드가 나머지에 해당됩니다. 표본의 13,504 레코드 중 1,952 레코드가 예라고 응답하여 전체 적중 비율은 14.45%입니다. 고객이 선호하는 응답을 한 경향이 높거나 낮은 세그먼트를 식별하여 이 비율을 높일 수 있습니다.

8. 대화형 목록 뷰어의 메뉴에서 다음을 선택하십시오.

도구 > **세그먼트 찾기**

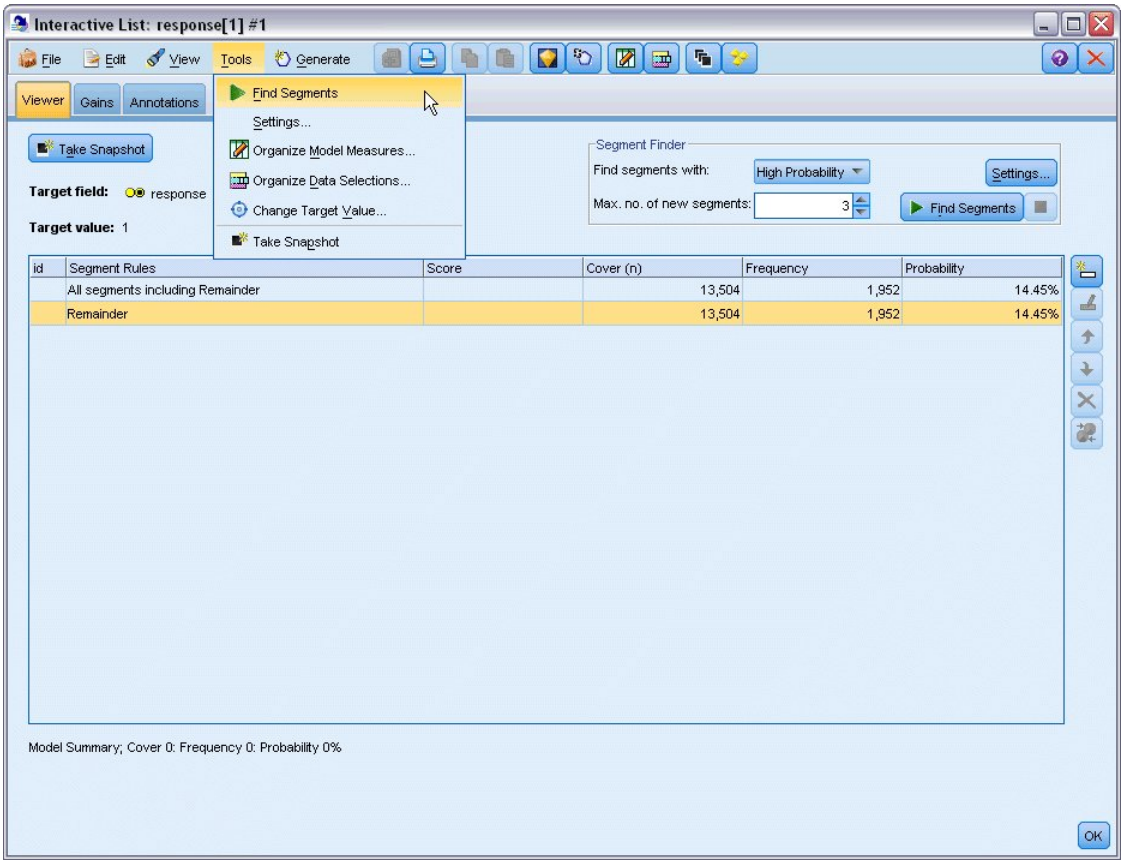


그림 *123.* 대화형 목록 뷰어

이는 사용자가 의사결정 목록 노드에서 지정된 설정을 기준으로 하여 기본 마이닝 작업을 실행합니다. 완료된 작업은 모델 앨범 대화 상자의 대안 탭에 나열된 세 가지 대안 모델을 리턴합니다.

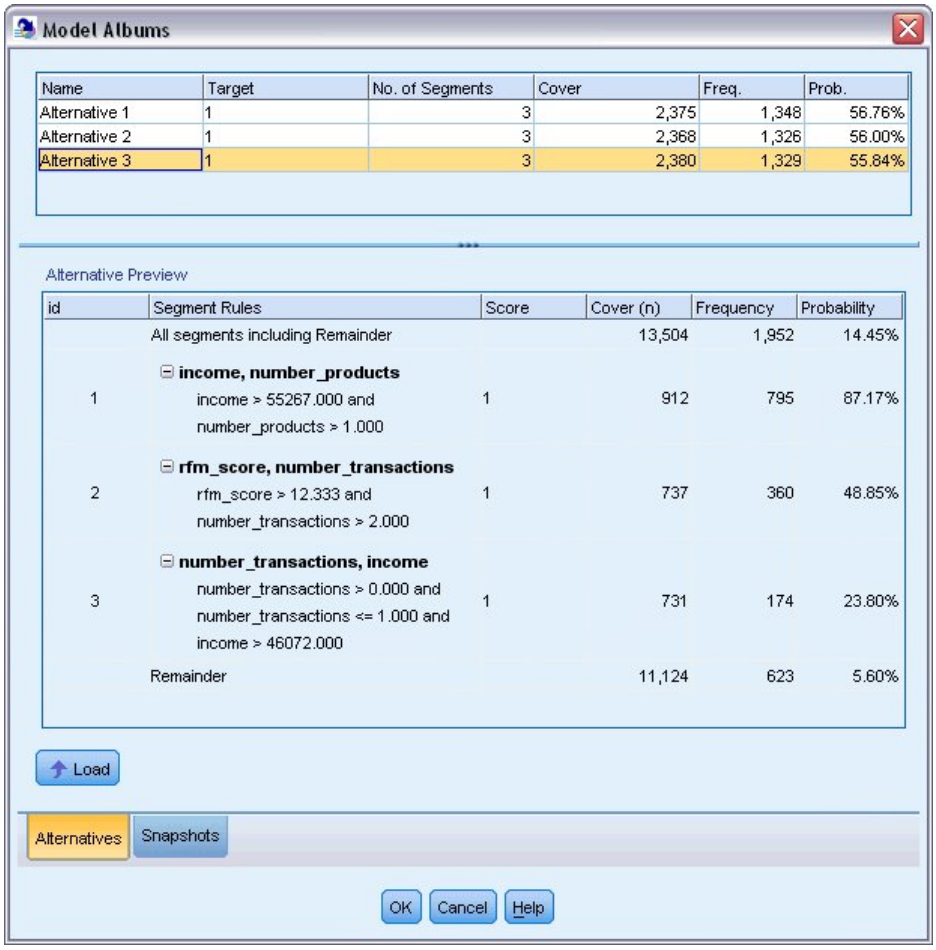


그림 *124.* 사용 가능한 대안 모델

9. 목록에서 첫 번째 대안 모델을 선택하십시오. 세부사항은 대안 미리보기 패널에 표시됩니다.

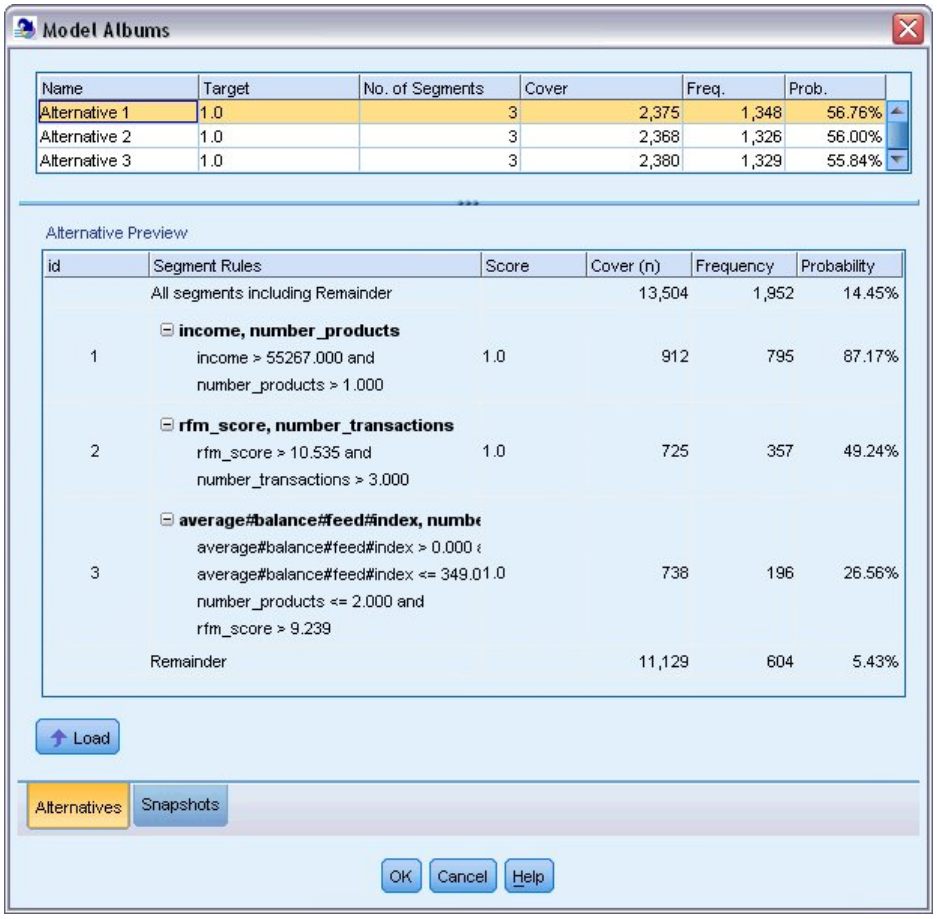


그림 *125.* 선택된 대안 모델

대안 미리보기 패널을 사용하면 작업 모델을 변경할 필요 없이 원하는 수의 대안을 신속하게 찾아볼 수 있으며 쉽게 다른 접근법을 시험해 볼 수 있습니다.

참고*:* 여기에 표시된 것처럼, 모델을 더 잘 보려면 대화 상자 내의 대안 미리보기 패널을 최대화하십시오. 패널 경계를 끌어서 최대화할 수 있습니다.

수입, 월별 트랜잭션 수, RFM 점수와 같은 예측변수를 기준으로 하는 규칙을 사용하면 모델이 표본 전체에 대해 이보다 높은 응답 비율의 세그먼트를 식별합니다. 세그먼트가 결합된 경우, 이 모델에서 적중 비율을 56.76%로 개선할 수 있음을 제안합니다. 단, 모델은 나머지에 해당하는 수백 개의 적중이 있는 11,000개 이상의 레코드를 남겨두고 전체 표본의 작은 부분만을 포함합니다. 낮은 성능의 세그먼트를 제외하면서 이러한 적중을 더 많이 캡처하는 모델이 필요할 것입니다.

10. 다른 모델링 방법을 시도하려면 메뉴에서 다음을 선택하십시오.

도구 > **설정**

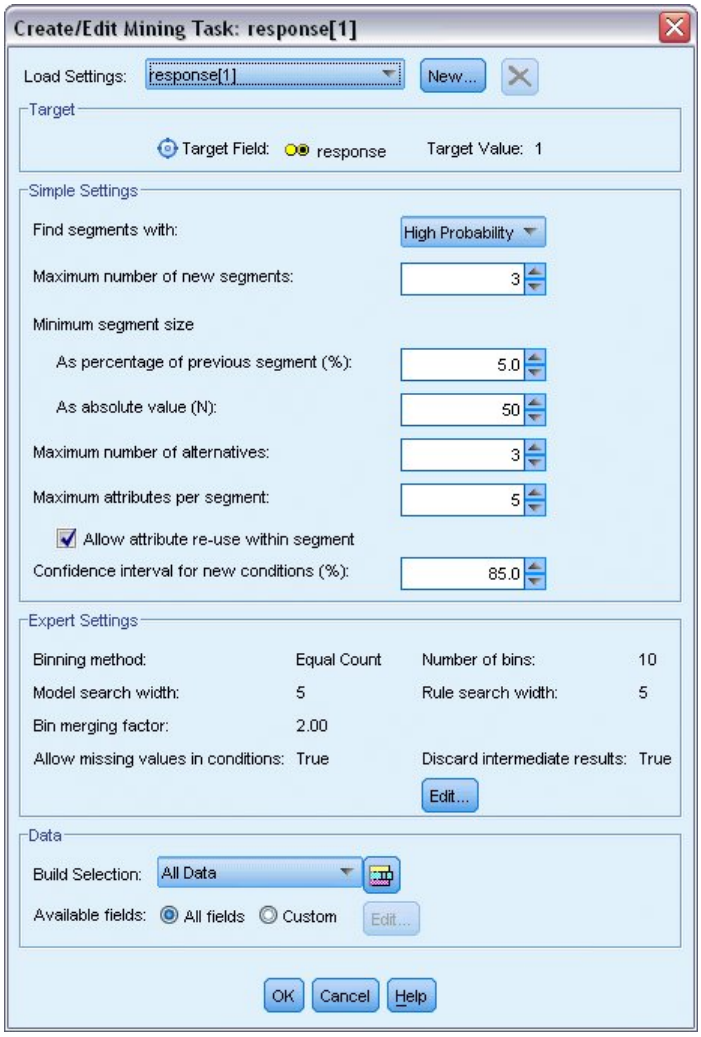


그림 *126.* 마이닝 작업 작성/편집 대화 상자

11. **새로 만들기** 단추(오른쪽 상단 코너)를 클릭하여 두 번째 마이닝 작업을 작성하고 새 설정 대화 상자의 작업 이름으로 아래로 검색을 선택하십시오.

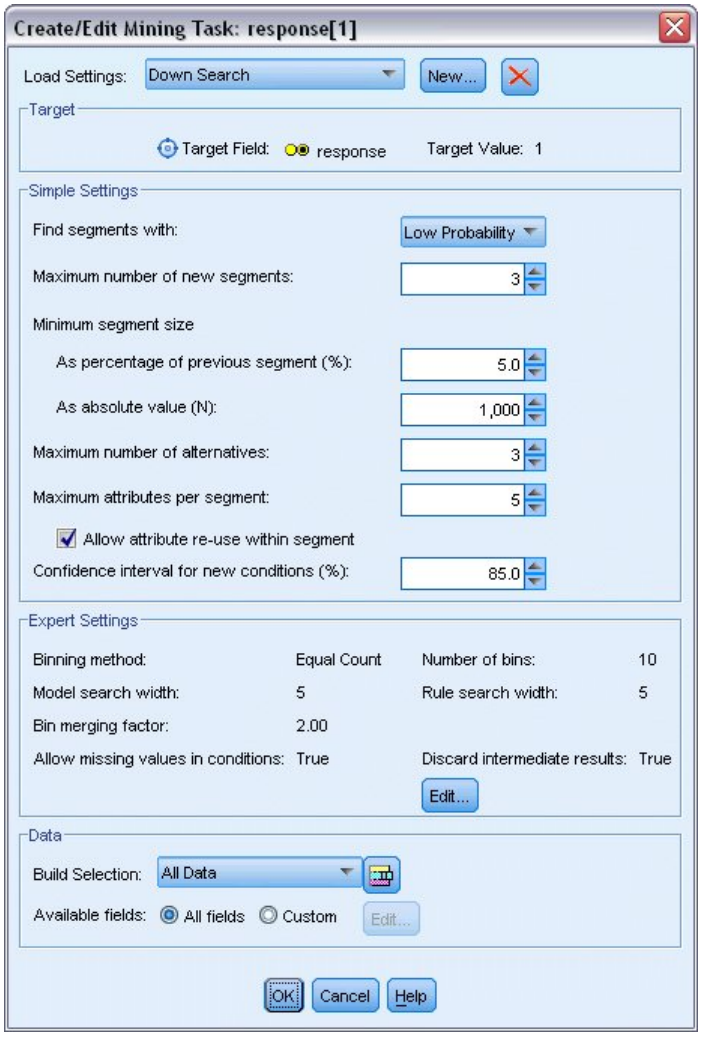


그림 *127.* 마이닝 작업 작성/편집 대화 상자

12. 작업에 대한 검색 방향을 **낮은 확률**로 변경하십시오. 그러면 알고리즘이 가장 높은 반응률이 아니라 가장 낮은 반응률을 가진 세그먼트를 검색하게 됩니다.
13. 최소 세그먼트 크기를 1,000으로 늘리십시오. **확인**을 클릭하여 대화형 목록 뷰어로 돌아가십시오.
14. 대화형 목록 뷰어에서 세그먼트 파인더 패널이 새 작업 세부사항을 표시하는지 확인하고 **세그먼트 찾기**를 클릭하십시오.

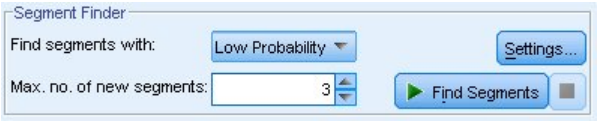


그림 *128.* 새 마이닝 작업에서 세그먼트 찾기

작업 결과 모델 앨범 대화 상자의 대안 탭에 표시되고 이전 결과와 동일한 방법으로 미리 볼 수 있는 새 대안 집합이 리턴됩니다.

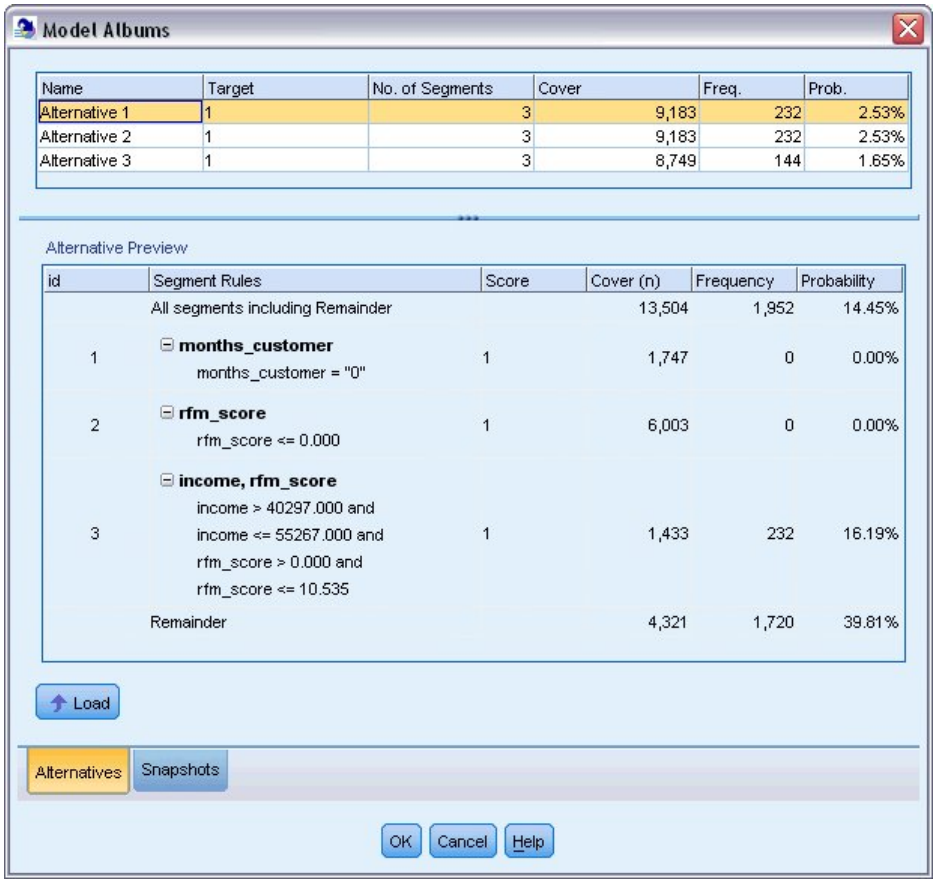


그림 *129.* 아래로 검색 모델 결과

이번에는 각 모델이 높은 반응 확률이 아니라 낮은 반응 확률의 세그먼트를 식별합니다. 첫 번째 대안을 볼 때 단순히 이러한 세그먼트를 제외하기만 하면 나머지에 대한 적중률이 39.81%로 증가합니다. 이는 앞에서 본 모델보다는 낮으나 적용 범위는 더 높아집니다. 즉, 총 적중이 더 많습니다.

두 방법을 결합하여, 즉, 낮은 확률 검색을 사용하여 관심이 없는 레코드를 제거하고 뒤 이어 높은 확률 검색을 사용하여 이 결과를 개선할 수 있습니다.

15. **로드**를 클릭하여 이 모델(첫 번째 아래로 검색 대안)을 작업 모델로 만들고 **확인**을 클릭하여 모델 앨범 대화 상자를 닫으십시오.

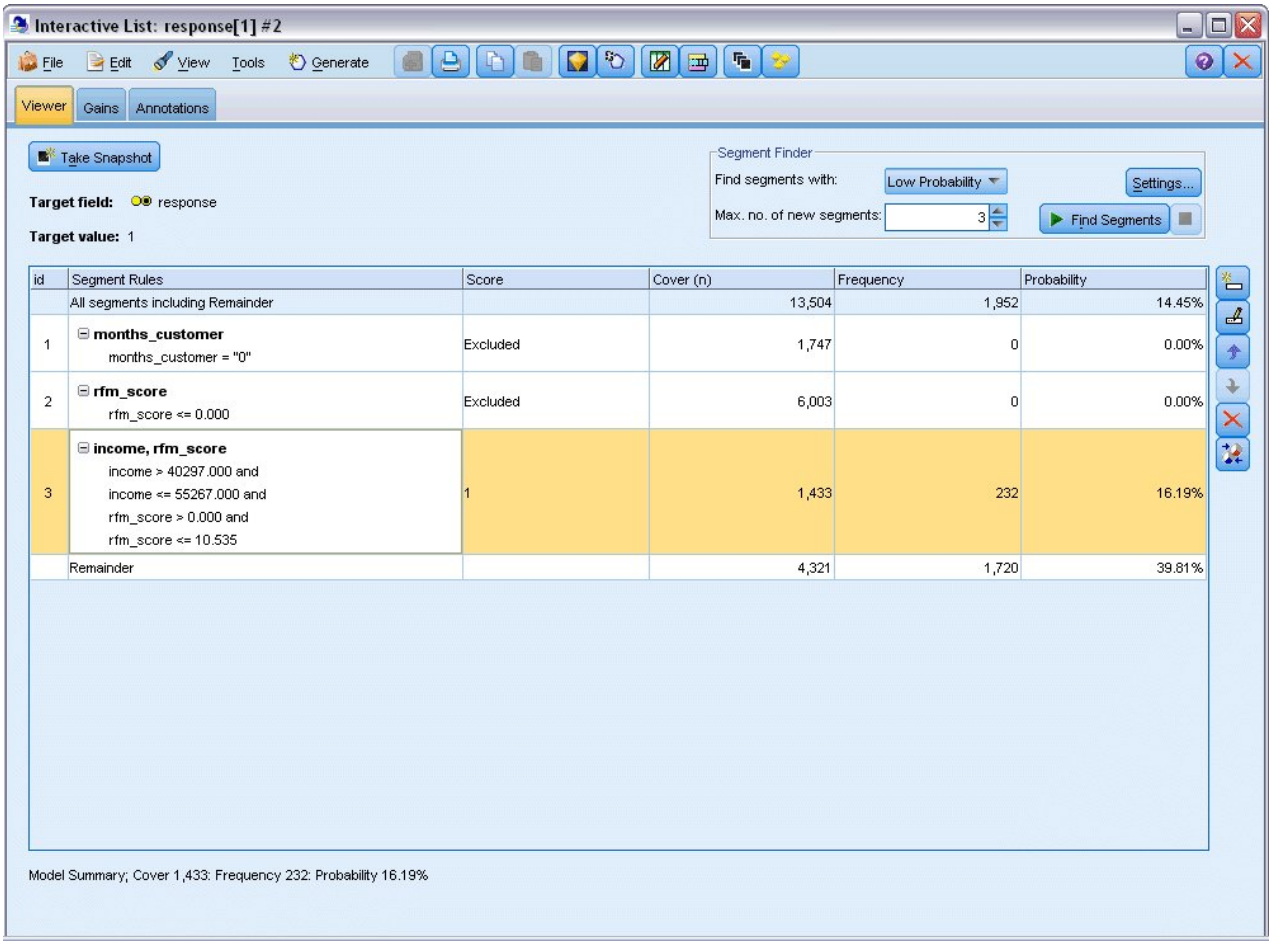


그림 *130.* 세그먼트 제외

16. 첫 번째 두 개의 세그먼트를 각각 마우스 오른쪽 단추로 클릭하고 **세그먼트 제외**를 선택하십시오. 이러한 세그먼트는 합쳐서 거의 8,000 개의 적중이 0인 레코드를 캡처하므로 이후 오퍼를 위해 제외해야 합니다. (제외된 세그먼트는 이를 나타내기 위해 널로 스코어링됩니다.)
17. 세 번째 세그먼트를 마우스 오른쪽 단추로 클릭하고 **세그먼트 삭제**를 선택하십시오. 이 세그먼트에 대한 적중 비율은 16.19%로서 기준선 비율인 14.45%와 그다지 다르지 않으므로 이를 그대로 유지하는 것을 정당화하는 정보를 충분히 추가하지 못합니다.

    참고: 세그먼트를 삭제하는 것은 세그먼트를 제외하는 것과 같지 않습니다. 세그먼트를 제외하면 단순히 스코어링 방법이 변경되는 반면 삭제하면 모델에서 완전히 제거됩니다.

    가장 낮은 성과의 세그먼트를 제외하였으므로 이제 나머지에서 높은 성과의 세그먼트를 검색할 수 있습니다.
18. 다음 마이닝 작업이 나머지에만 적용되도록 테이블에서 나머지 행을 클릭하여 이를 선택하십시오.

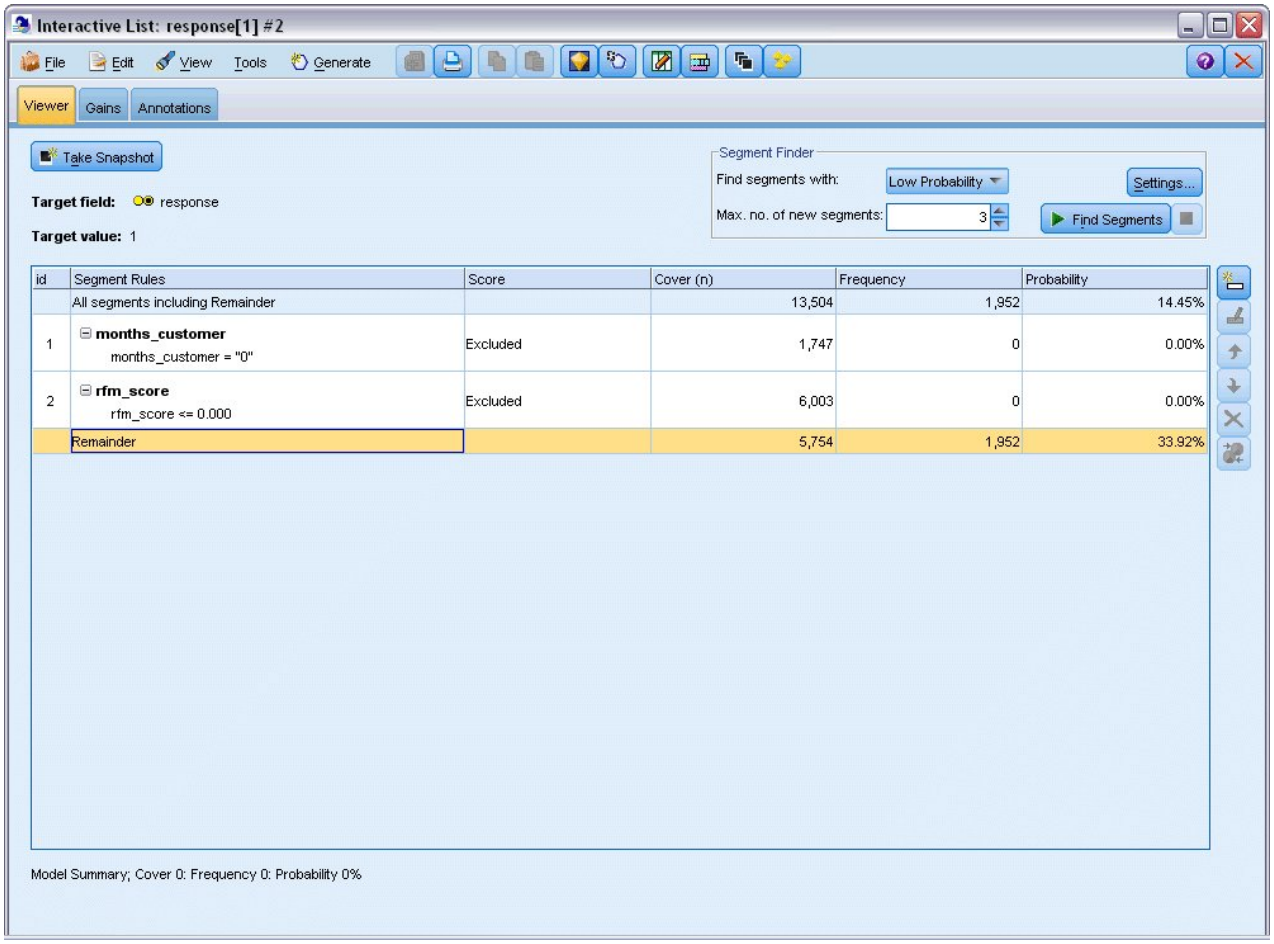


그림 *131.* 세그먼트 선택

19. 선택된 나머지에 대해 **설정**을 클릭하여 마이닝 작업 작성/편집 대화 상자를 다시 여십시오.
20. **설정 로드**의 위쪽에서 기본 마이닝 작업: **반응[1]**을 선택하십시오.
21. 새 세그먼트의 수를 5로, 최소 세그먼트 크기를 500으로 늘리도록 **단순 설정**을 편집하십시오.
22. **확인**을 클릭하여 대화형 목록 뷰어로 돌아가십시오.

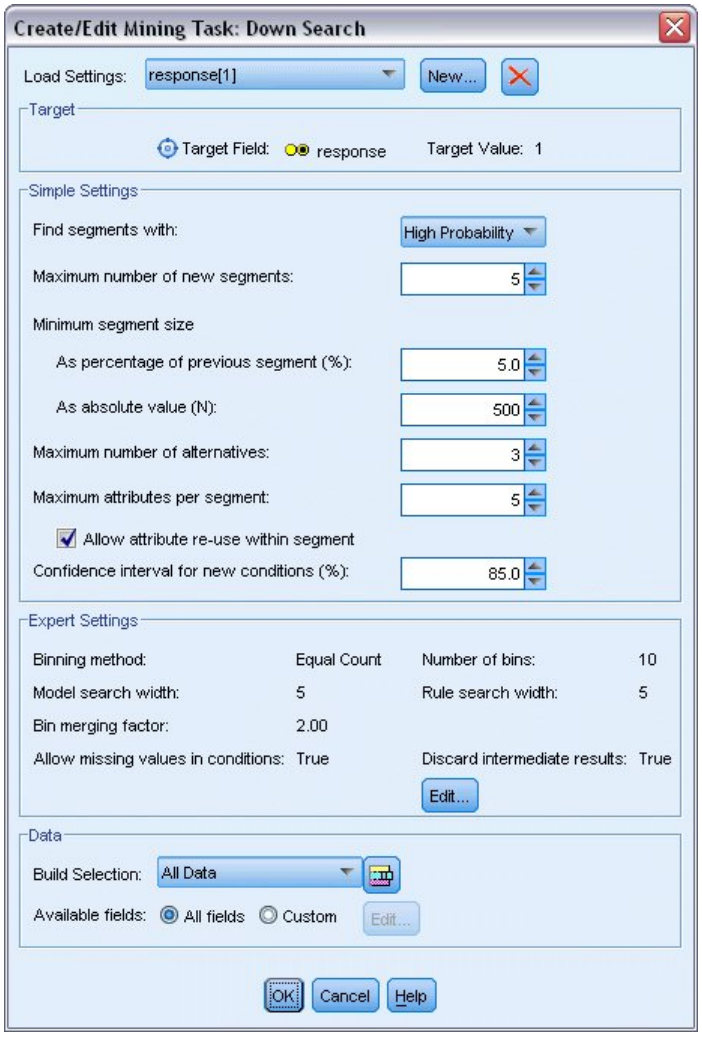


그림 *132.* 기본 마이닝 작업 선택

23. **세그먼트 찾기**를 클릭하십시오.

그러면 대안 모델의 또 다른 집합이 표시됩니다. 한 마이닝 작업의 결과를 다른 마이닝 작업에 피드함으로써 이러한 최신 모델에 높은 성과 및 낮은 성과 세그먼트가 혼합되어 포함될 수 있습니다. 낮은 반응률의 세그먼트는 제외되고, 즉, 널로 스코어링되고 포함된 세그먼트는 1로 스코어링됩니다. 전체 통계량에서 첫 번째 대안 모델은 45.63%의 적중 비율을 나타내고 이전의 어떤 모델보다 높은 적용 범위(3,456 레코드 중 1,577 레코드 적중)를 가짐으로써 이러한 제외를 반영합니다.

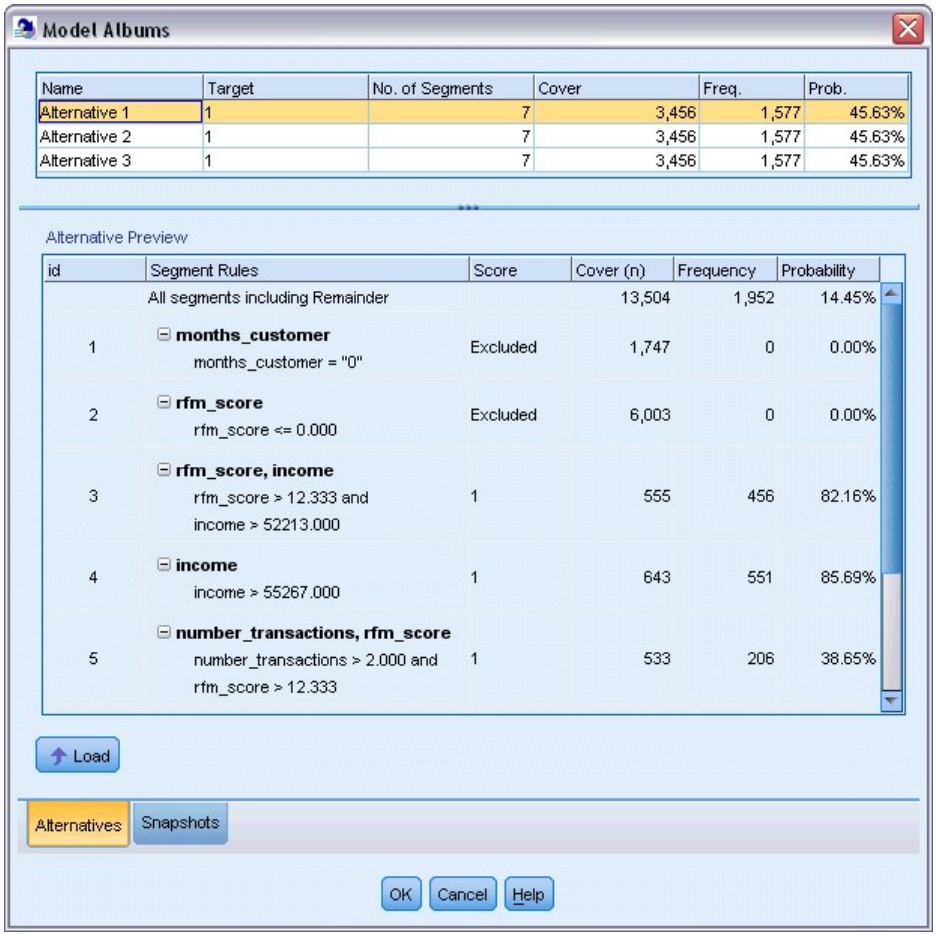


그림 *133.* 결합된 모델에 대한 대안

24. 첫 번째 대안을 미리 본 다음 **로드**를 클릭하여 이를 작업 모델로 만드십시오.

Excel을 사용하여 사용자 정의 측도 계산

1. 모델이 실제로 수행하는 방법에 대해 더 깊은 통찰을 얻으려면 도구 메뉴에서 **모델 측도 구성**을 선택하십시오.

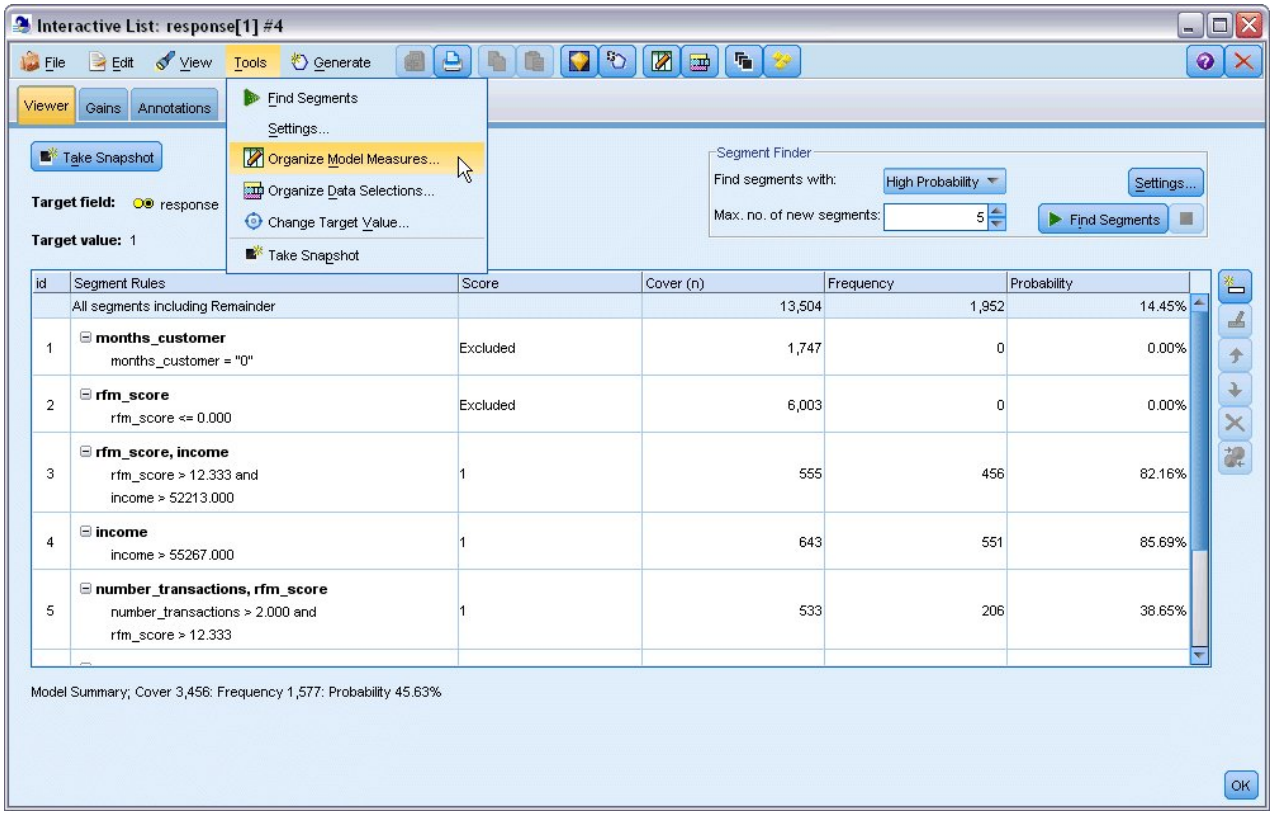


그림 *134.* 모델 측도 구성

모델 측도 구성 대화 상자를 사용하면 대화형 목록 뷰어에서 표시할 측도(열)를 선택할 수 있습니다. 또한 측도가 모든 레코드 또는 선택된 서브세트에 대해 계산되는지 지정할 수 있으며 적용되는 경우에 한해 숫자가 아니라 원형 차트로 표시할 수 있습니다.

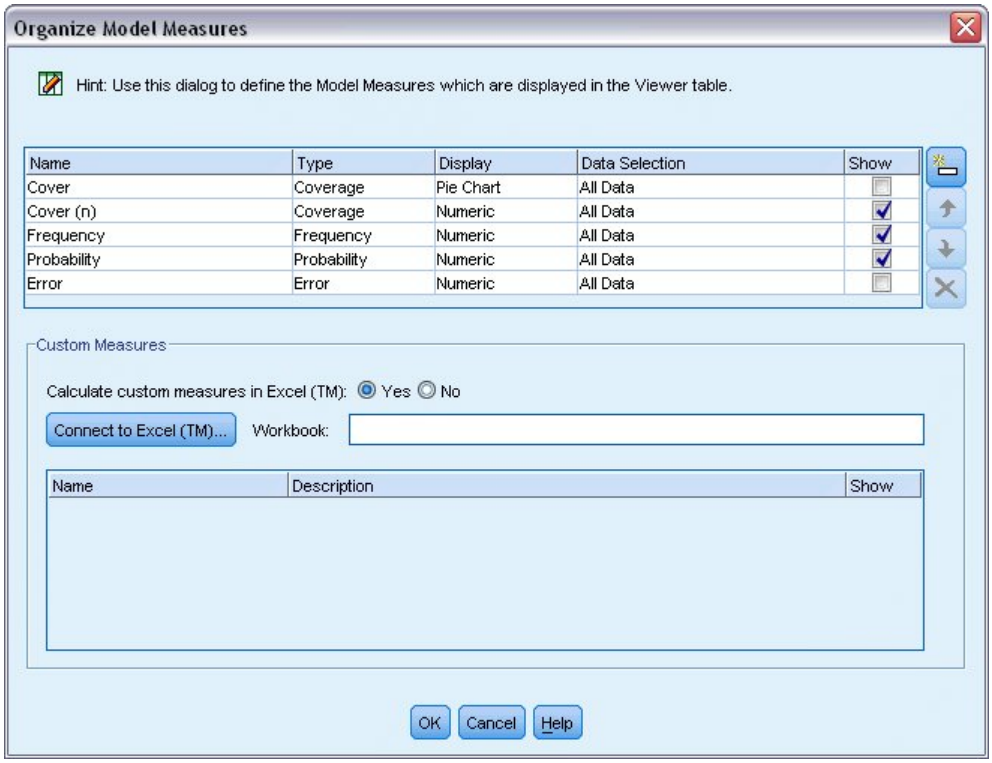


그림 *135.* 모델 측도 구성 대화 상자

또한 Microsoft Excel이 설치되어 있으면 사용자 정의 측도를 계산하여 이를 대화형 표시에 추가하는 Excel 템플리트에 링크할 수 있습니다.

2. 모델 측도 구성 대화 상자에서 **Excel(TM)에서 사용자 정의 측도 계산**을 **예**로 설정하십시오.
3. **Excel(TM)에 연결**을 클릭하십시오.
4. IBM SPSS Modeler 설치의 *Demos* 폴더 내의 *streams* 아래에 있는 *template_profit.xlt* 워크북을 선택하고 **열기**를 클릭하여 스프레드시트를 시작하십시오.

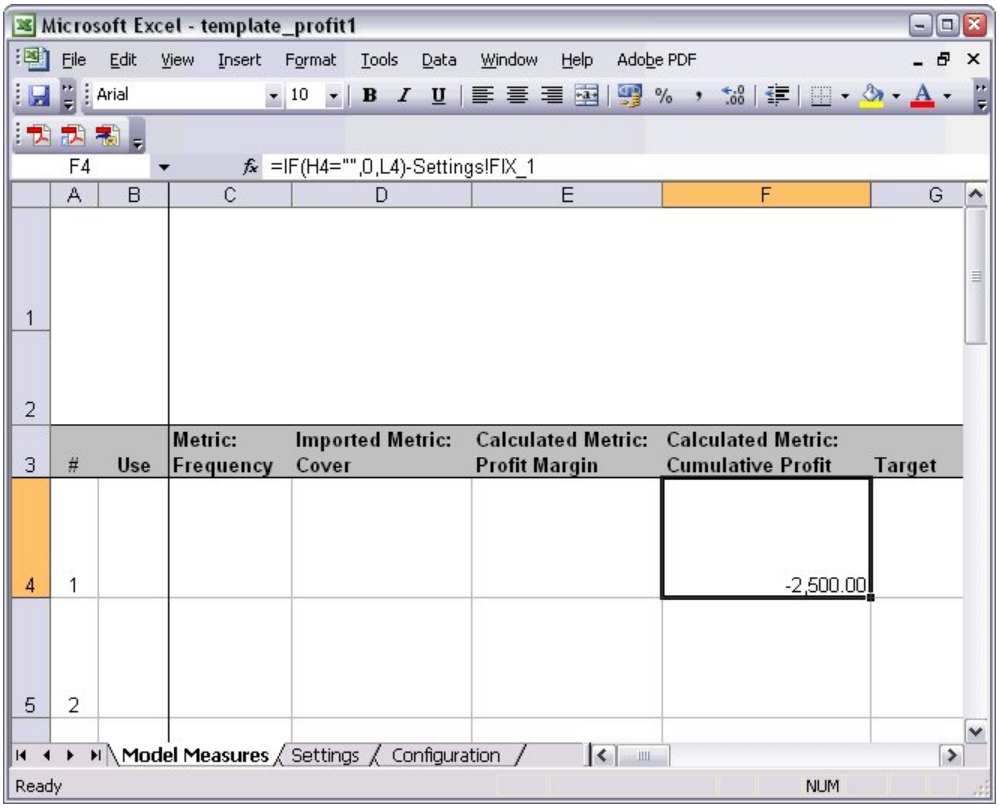


그림 *136. Excel* 모델 측도 워크시트

Excel 템플리트에는 세 개의 워크시트가 포함되어 있습니다.

- **모델 측도**는 모델에서 가져온 모델 측도를 표시하고 모델로 다시 내보내기 위한 사용자 정의 측도를 계산합니다.
- **설정**에는 사용자 정의 측도 계산에 사용할 모수가 포함됩니다.
- **설정**은 모델로 가져오거나 모델에서 내보낼 측도를 정의합니다.

모델로 다시 내보낼 메트릭은 다음과 같습니다.

- **이익.** 세그먼트의 순 이익
- **누적 이익.** 캠페인의 총 이익

다음 수식에 의해 정의된 바와 같습니다.

```
Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost

Cumulative Profit = Total Profit Margin - Fixed cost
```

빈도 및 범위는 모델에서 가져옵니다.

비용 및 수입 모수는 설정 워크시트에서 사용자에 의해 지정됩니다.

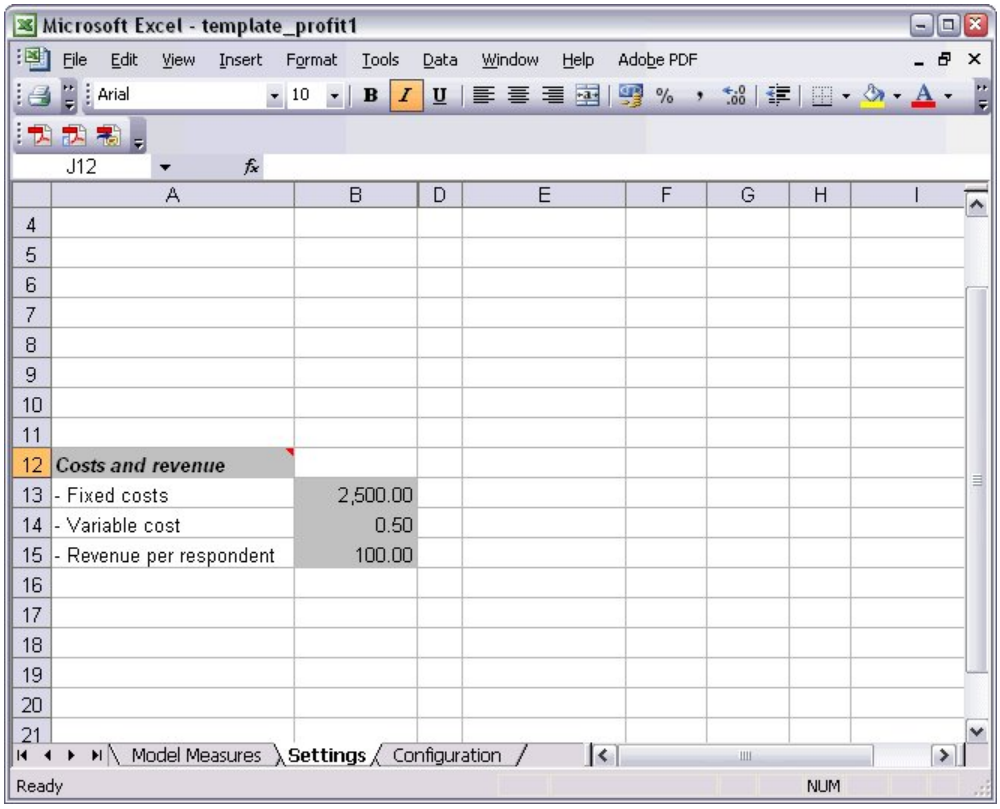


그림 *137. Excel* 설정 워크시트

고정 비용은 디자인 및 계획과 같이 캠페인의 설정 비용입니다.

가변 비용은 엔벨로프 및 스탬프와 같이 오퍼를 각 고객에게 확장하는 비용입니다.

반응자당 수입은 오퍼에 반응한 고객의 순 수입입니다.

5. 다시 모델로 링크하려면 Windows 작업 표시줄을 사용하거나 Alt+Tab을 눌러 대화형 목록 뷰어로 다시 이동하십시오,

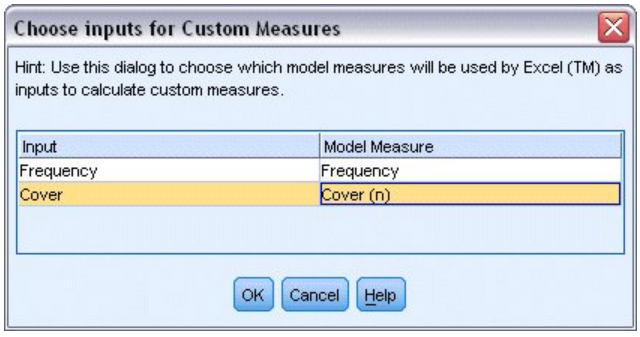


그림 *138.* 사용자 정의 측도에 대한 입력 선택

모델에서 템플리트에서 정의된 특정 모수로 입력을 맵핑할 수 있는 사용자 정의 측도에 대한 입력 선택 대화 상자가 표시됩니다. 왼쪽 열에는 사용 가능한 측도가 나열되고 오른쪽 열에서는 설정 워크시트에서 정의된 대로 이를 스프레드시트 모수에 맵핑합니다.

6. **모델 측도** 열에서 각 입력에 대해 **빈도** 및 **범위(n)**를 선택하고 **확인**을 클릭하십시오.

   이 케이스에서는 템플리트(빈도 및 범위(n))의 모수 이름이 입력과 일치하나 다른 이름이 사용될 수 있습니다.

7. 모델 측도 구성 대화 상자에서 **확인**을 클릭하여 대화형 목록 뷰어를 업데이트하십시오.

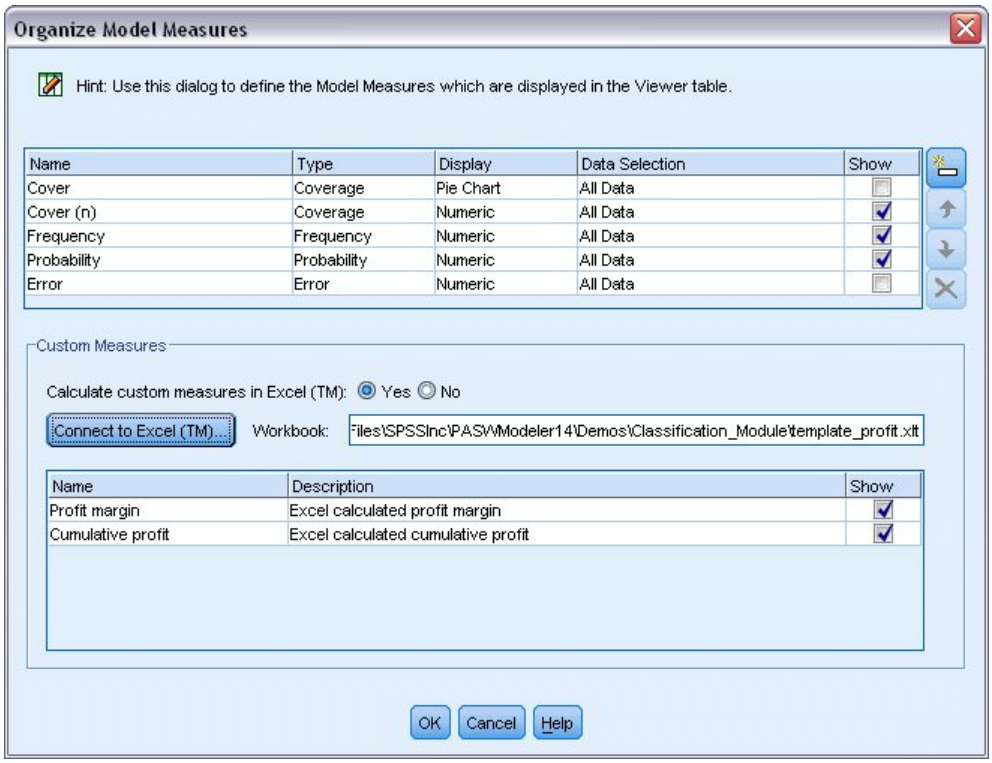


그림 *139. Excel*의 사용자 정의 측도를 표시하는 모델 측도 구성 대화 상자

이제 창에 새 측도가 새 열로 추가되고 모델이 업데이트될 때마다 다시 계산됩니다.

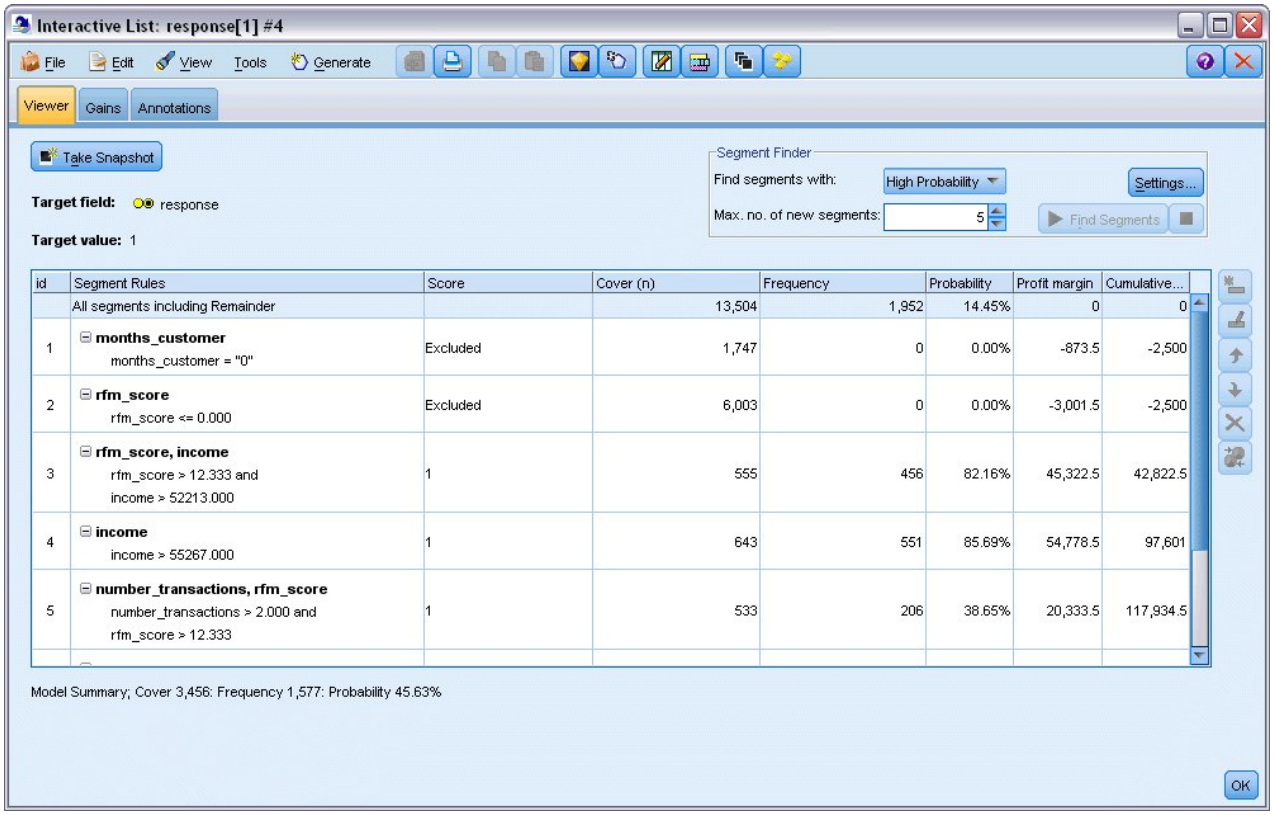


그림 *140.* 대화형 목록 뷰어에 표시되는 *Excel*의 사용자 정의 측도

Excel 템플리트를 편집하여 원하는 수의 사용자 정의 측도를 작성할 수 있습니다.

Excel 템플리트 수정

IBM SPSS Modeler가 대화형 목록 뷰어와 함께 사용하기 위한 기본 Excel 템플리트와 함께 제공되나 사용자가 설정을 변경하거나 자신의 템플리트를 직접 작성하고자 할 수 있습니다. 예를 들어, 템플리트의 비용이 사용자의 조직에 적합하지 않아 수정이 필요할 수 있습니다.

참고: 기존 템플리트를 수정하거나 직접 작성하려면 파일을 Excel 2003 *.xlt* 접미문자를 사용하여 저장해야 합니다.

새 비용 및 수입 세부사항을 사용하여 기본 템플리트를 수정하고 새 그림으로 대화형 목록 뷰어를 업데이트하려면 다음을 수행하십시오.

1. 대화형 목록 뷰어의 도구 메뉴에서 **모델 측도 구성**을 선택하십시오.
2. 모델 측도 구성 대화 상자에서 **Excel™에 연결**을 클릭하십시오.
3. *template_profit.xlt* 워크북을 선택하고 **열기**를 클릭하여 스프레드시트를 시작하십시오.
4. 설정 워크시트를 선택하십시오.
5. **고정 비용**을 3,250.00으로 편집하고 **반응자당 수입**을 150.00으로 편집하십시오.

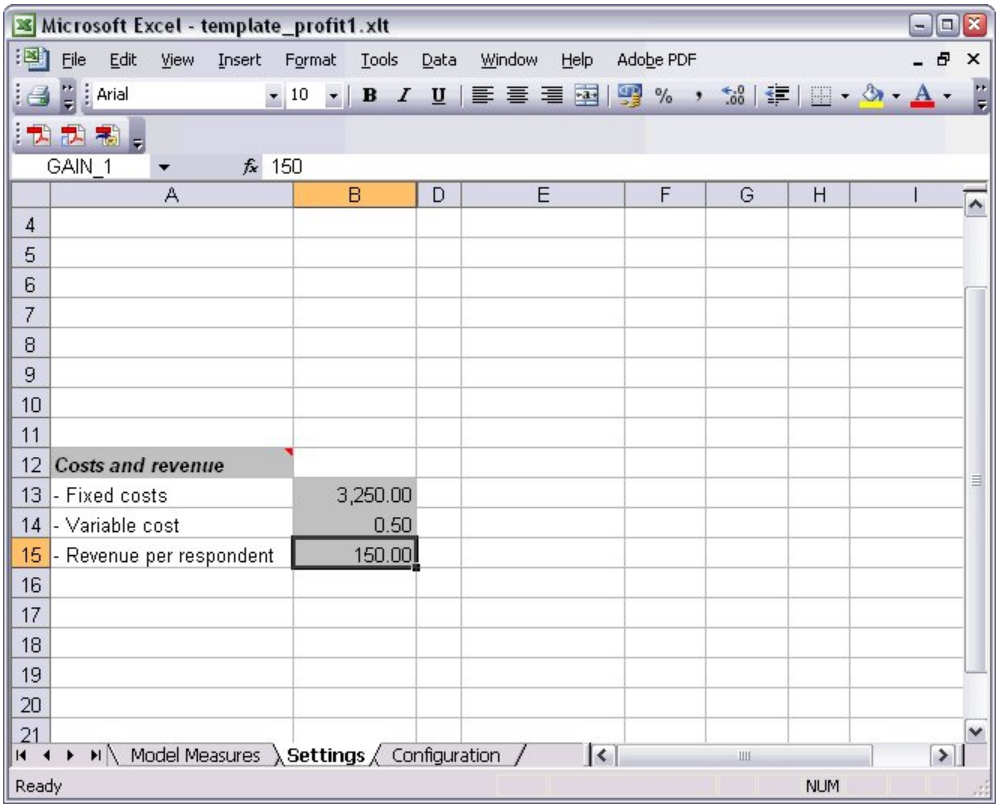


그림 *141. Excel* 설정 워크시트의 수정된 값

6. 수정된 템플리트를 고유한 관련 파일 이름을 사용하여 저장하십시오. Excel 2003 *.xlt* 확장자를 가지고 있는지 확인하십시오.

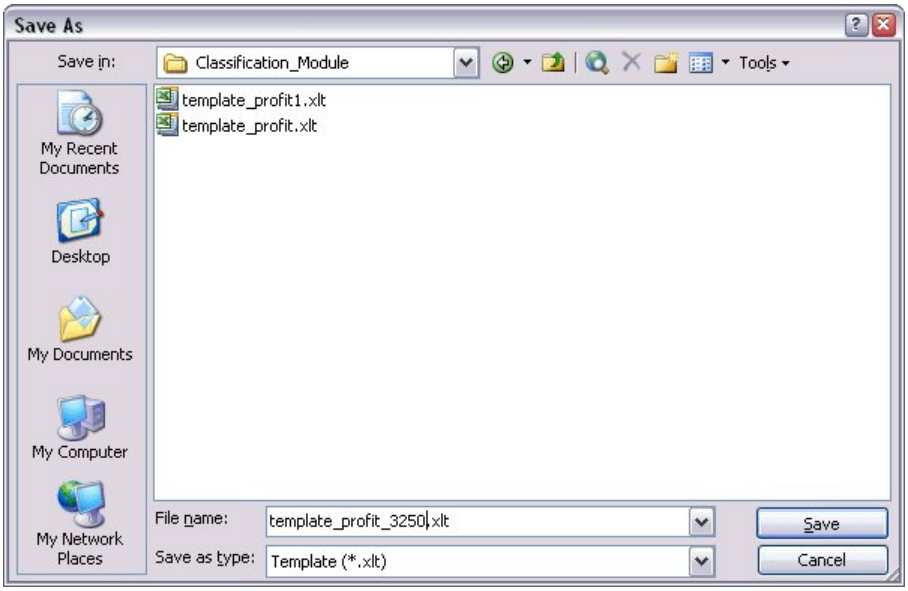


그림 *142.* 수정된 *Excel* 템플리트 저장

7. Windows 작업 표시줄을 사용하거나 Alt+Tab을 눌러 대화형 목록 뷰어로 다시 이동하십시오.

사용자 정의 측도에 대한 입력 선택 대화 상자에서 표시할 측도를 선택하고 **확인**을 클릭하십시오.

8. 모델 측도 구성 대화 상자에서 **확인**을 클릭하여 대화형 목록 뷰어를 업데이트하십시오.

이 예는 Excel 템플리트를 수정하는 한 가지 간단한 방법만 표시하지만 전체 데이터를 추가적으로 변경하고 데이터를 대화형 목록 뷰어에 전달하거나 Excel로 작업하여 그래프와 같은 다른 출력을 생성할 수 있습니다.

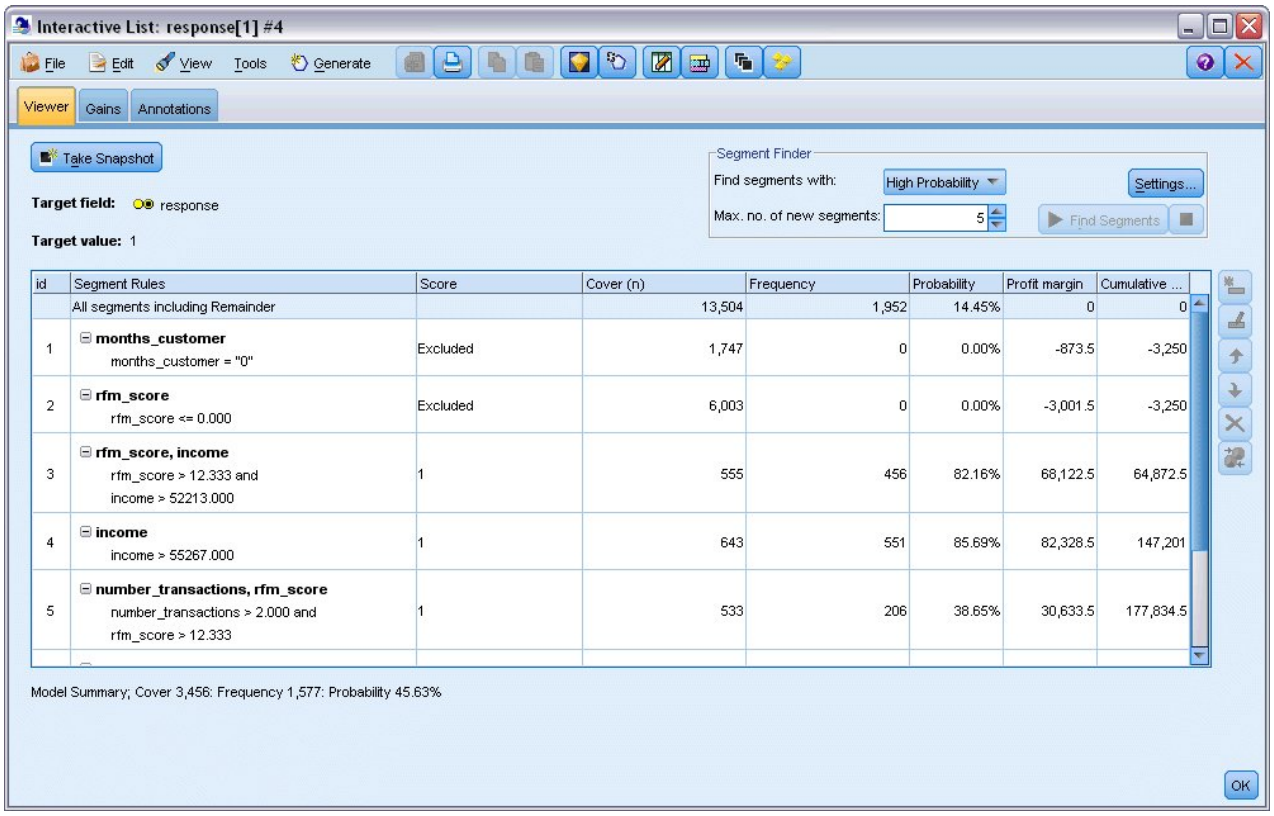


그림 *143.* 대화형 목록 뷰어에 표시되는 *Excel*에서 수정된 사용자 정의 측도

결과 저장

대화형 세션 동안 나중에 사용하기 위해 모델을 저장하려면 스냅샷 탭에 나열될 모델의 스냅샷을 작성하십시오. 대화형 세션 동안 언제든지 저장된 임의의 스냅샷으로 돌아갈 수 있습니다.

이 방법을 계속 사용하여 추가 마이닝 작업을 시험하여 추가 세그먼트를 검색할 수 있습니다. 또한 기존 세그먼트를 편집하고 사용자 자신의 비즈니스 규칙을 기준으로 하여 사용자 정의 세그먼트를 삽입하고 데이터 선택영역을 작성하여 특정 그룹에 대한 모델을 최적화하고 수많은 다른 방법으로 모델을 사용자 정의할 수 있습니다. 최종적으로 각 세그먼트가 스코어링되는 방법을 지정하기 위해 각 세그먼트를 명시적으로 포함하거나 제외할 수 있습니다.

결과에 만족하면 생성 메뉴를 사용하여 스트림에 추가하거나 스코어링 목적으로 배포할 수 있는 모델을 생성할 수 있습니다.

또는 다른 날을 위해 대화형 세션의 현재 상태를 저장하려면 파일 메뉴에서 **모델링 노드 업데이트**를 선택하십시오. 그러면 마이닝 작업, 모델 스냅샷, 데이터 선택, 사용자 정의 측도를 포함한 현재 설정으로 의사결정 목록 모델링 노드가 업데이트됩니다. 다음에 스트림을 실행할 때 세션을 현재 상태로 복원하려면 의사결정 목록 모델링 노드에서 **저장된 세션 정보 사용**이 선택되어 있는지만 확인하면 됩니다.

제 12 장 통신 고객 분류(다항 로지스틱 회귀분석)

로지스틱 회귀분석은 입력 필드 값을 기반으로 레코드를 분류하는 통계 기법입니다. 이는 선형 회귀와 비슷하지만, 숫자 대신 범주형 대상 필드를 사용합니다.

예를 들어, 통신 제공업체가 서비스 사용 패턴을 기준으로 고객층을 세그먼트화하여 고객을 4개의 그룹으로 범주화한다고 가정합니다. 소속그룹을 예측하기 위해 인구 통계학적 데이터를 사용하면 개별 잠재 고객에 대한 제공을 사용자 정의할 수 있습니다.

이 예에서는 *telco.sav*라는 데이터 파일을 참조하는 *telco_custcat.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *telco_custcat.str* 파일은 *streams* 디렉토리에 있습니다.

이 예에서는 사용 패턴을 예측하기 위해 인구 통계학적 데이터를 사용하는 데 초점을 맞춥니다. 대상 필드 *custcat*에는 다음과 같이 네 개의 고객 그룹에 해당하는 네 개의 가능한 값이 있습니다.

값	레이블
1	기본 서비스
2	E-서비스
3	플러스 서비스
4	전체 서비스

대상에 다중 범주가 있으므로 다항 모델이 사용됩니다. 예/아니오, 참/거짓 또는 이탈/이탈하지 않음과 같이 두 개의 고유 범주가 있는 대상의 경우, 이항 모델이 대신 작성될 수 있습니다. 자세한 정보는 149 페이지의 제 13 장 『통신 서비스 제공자를 바꾸는 고객(이항 로지스틱 회귀분석)』의 내용을 참조하십시오.

스트림 작성

1. *Demos* 폴더에서 *telco.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

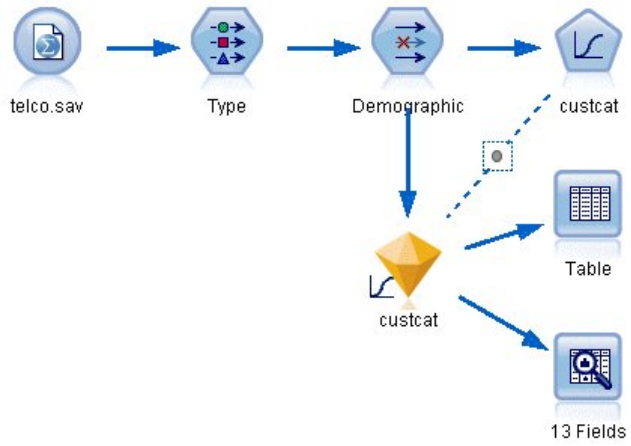


그림 *144.* 다항 로지스틱 회귀분석을 사용하여 고객을 분류하기 위한 샘플 스트림

a. 유형 노드를 추가하고 모든 측정 수준이 올바르게 설정되었는지 확인하고 **값 읽기**를 클릭하십시오. 예를 들어, 값이 0 및 1인 대부분의 필드는 플래그로 간주할 수 있습니다.

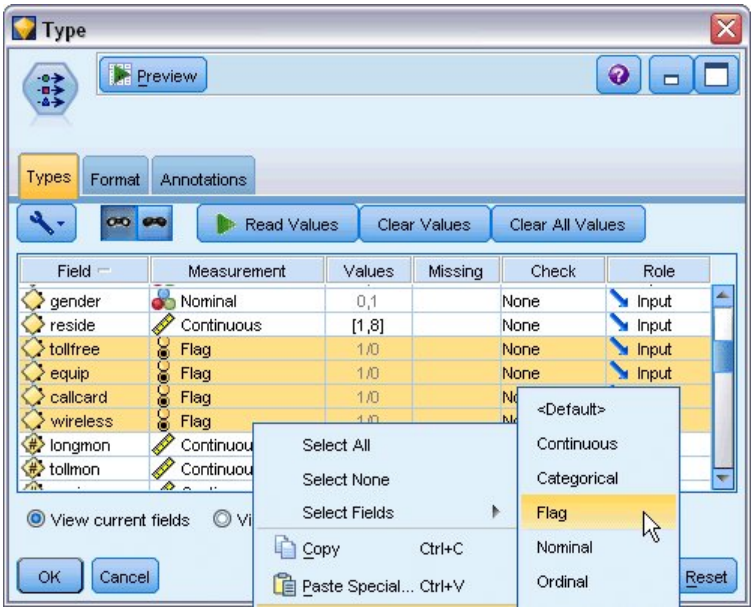


그림 *145.* 다중 필드에 대한 측정 수준 설정

팁: 유사한 값(0/1 등)을 가진 다중 필드에 대한 특성을 변경하려면 값 열 헤더를 클릭하여 필드를 값 기준으로 정렬한 다음 Shift 키를 누른 상태에서 마우스 또는 화살표를 사용하여 변경할 키를 모두 선택하십시오. 그런 다음 마우스 오른쪽 단추로 선택영역을 클릭하여 선택된 필드의 측정 수준 또는 기타 속성을 변경하십시오.

성별은 플래그 대신 두 개의 값 변수군이 있는 필드로 간주하는 것이 더 정확하므로 해당 측정 값을 **명목형**으로 두십시오.

b. 통신사용등급 필드에 대한 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.

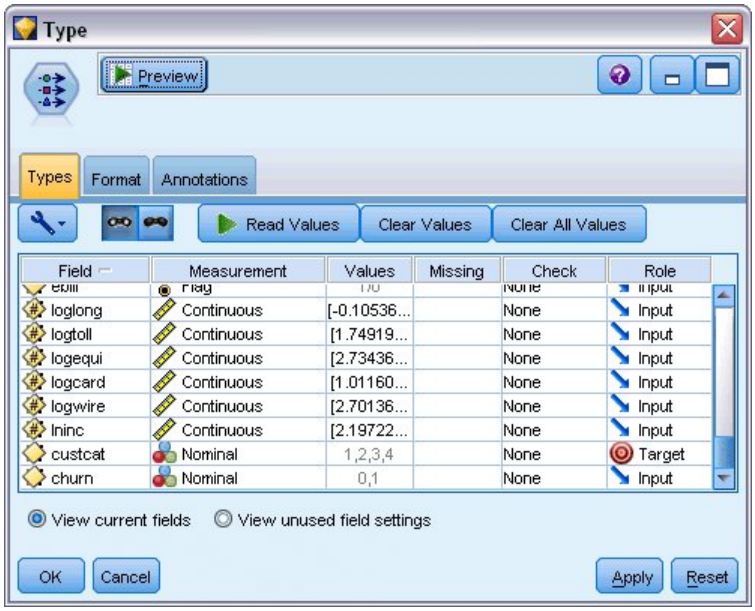


그림 *146.* 필드 역할 설정

이 예에서는 인구 통계에 초점을 맞추므로 관련 필드(지역, 연령, 혼인 여부, 주소, 수입, 교육수준, 고용, 은퇴, 성별, 거주 및 통신사용등급)만 포함하도록 필터 노드를 사용하십시오. 기타 필드는 이 분석 목적으로는 제외될 수 있습니다.

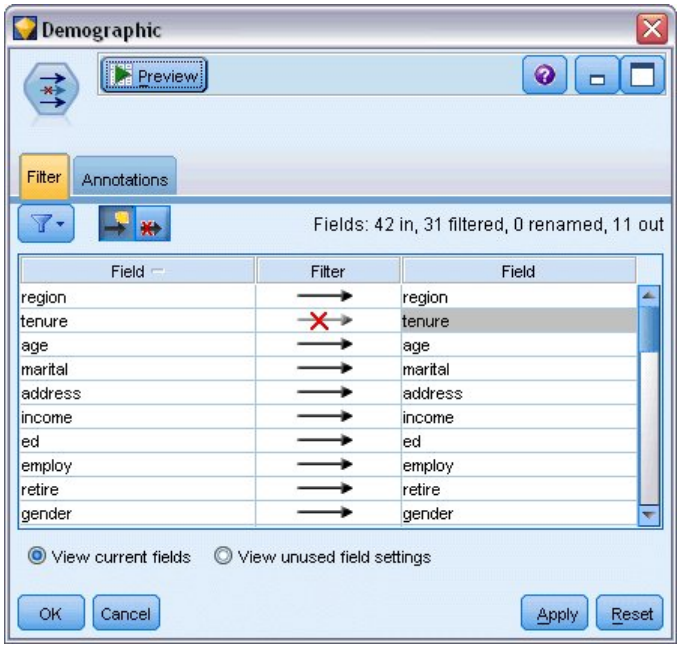


그림 *147.* 인구 통계학 필드 필터링

(또는 이러한 필드를 제외하지 않고 해당 필드에 대한 역할을 **없음**으로 지정하거나 모델링 노드에 사용할 필드를 선택할 수도 있습니다.)

2. 로지스틱 노드에서 **모델** 탭을 클릭하고 **단계선택법**을 선택하십시오. **다항**, **주효과** 및 **방정식에 상수항 포함**도 선택하십시오.

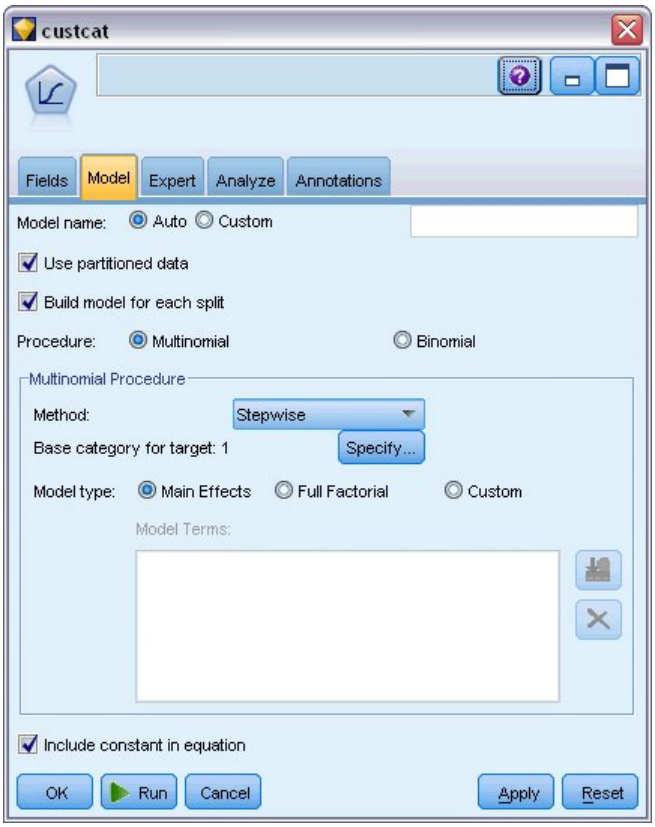


그림 *148.* 모델 옵션 선택

대상에 대한 기본 범주를 1로 두십시오. 모델이 기타 고객을 기본 서비스에 가입한 고객과 비교합니다.

3. 전문가 탭에서 **전문가** 모드를 선택하고 **출력**을 선택한 다음 고급 출력 대화 상자에서 **분류표**을 선택하십시오.

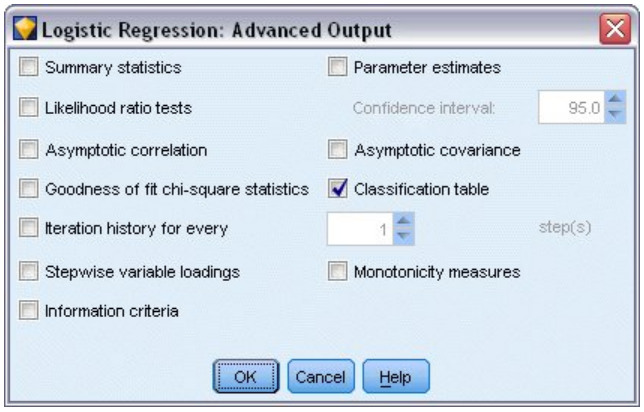


그림 *149.* 출력 옵션 선택

모델 찾아보기

1. 노드를 실행하여 모델을 생성하십시오. 모델은 오른쪽 상단 코너의 모델 팔레트에 추가됩니다. 세부사항을 보려면 마우스 오른쪽 단추로 생성된 모델 노드를 클릭하고 **찾아보기**를 선택하십시오.

모델 탭은 대상 필드의 각 범주에 레코드를 지정하는 데 사용된 방정식을 표시합니다. 네 개의 가능한 범주가 있으며 그 중 하나는 방정식 세부사항이 표시되지 않는 기본 범주입니다. 나머지 세 개의 방정식에 대해서는 세부사항이 표시됩니다. 범주 3은 플러스 서비스를 나타내는 것 등입니다.

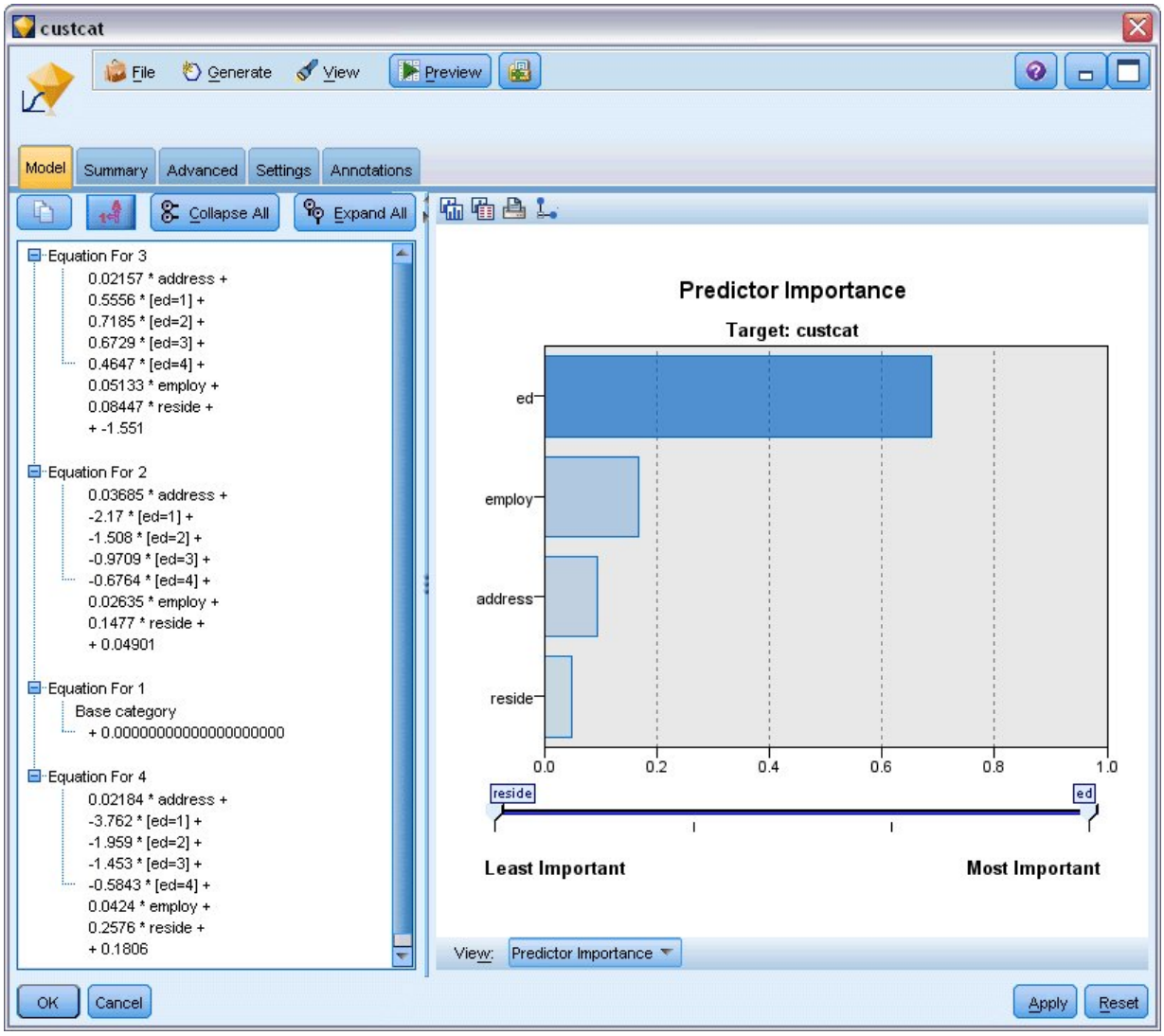


그림 *150.* 모델 결과 찾아보기

요약 탭은 (다른 사항 사이에서) 대상 및 모델에 의해 사용되는 입력(예측변수 필드)을 표시합니다. 이러한 필드는 실제로 단계 선택법을 기준으로 하여 선택된 필드이며 고려하도록 제출된 전체 목록이 아닙니다.

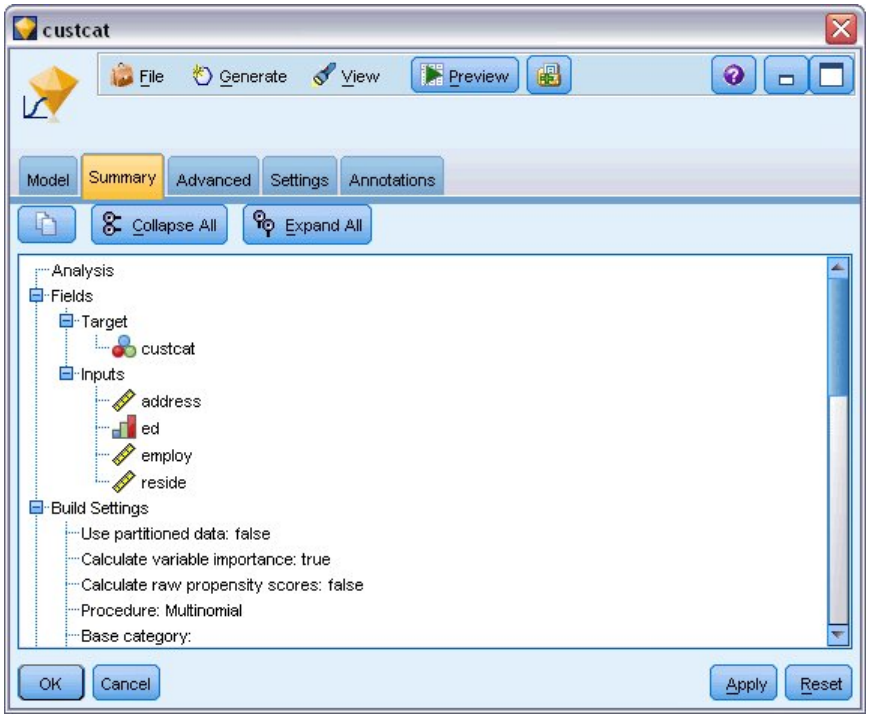


그림 *151.* 대상 및 입력 필드를 표시하는 모델 요약

고급 탭에 표시되는 항목은 모델링 노드의 고급 출력 대화 상자에서 선택된 옵션에 따라 다릅니다.

항상 표시되는 한 항목은 대상 필드의 각 범주에 해당하는 레코드의 퍼센트를 표시하는 케이스 처리 요약입니다. 이는 비교의 기준으로 사용할 널 모델을 제공합니다.

예측변수를 사용한 모델을 작성하지 않고 가장 잘 추측하는 방법은 모든 고객을 가장 일반적인 그룹에 지정하는 것이며 이는 플러스 서비스에 해당하는 방법입니다.

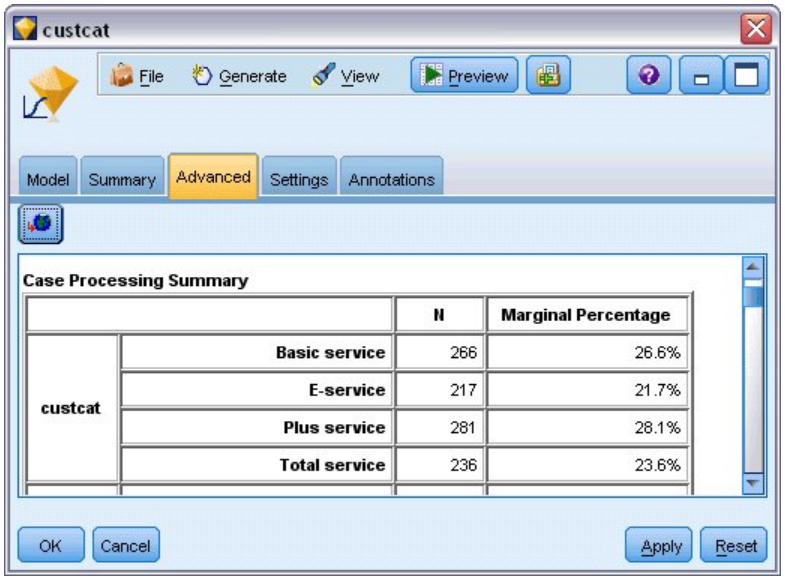


그림 *152.* 케이스 처리 요약

학습 데이터를 기준으로 할 때, 모든 고객을 널 모델에 지정하면 당시에 281/1000 = 28.1% 비율로 올바릅니다. 고급 탭은 모델 예측을 탐색할 수 있는 추가 정보를 포함합니다. 그런 다음 널 모델의 결과를 사용하는 예측과 비교하여 모델이 사용자의 데이터를 사용하여 얼마나 잘 작업하는지 알아볼 수 있습니다.

고급 탭의 아래쪽에 있는 분류표은 사용자의 모델에 대해 당시에 39.9% 올바른 결과를 표시합니다.

특히 이 모델은 전체 서비스 고객(범주 4)을 식별하는 데는 탁월하나 E-서비스 고객(범주 2)을 식별하는 데는 매우 부족합니다. 범주 2의 고객에 대해 더 나은 정확도를 원하는 경우, 해당 고객을 식별할 수 있는 다른 예측변수를 찾아야 합니다.

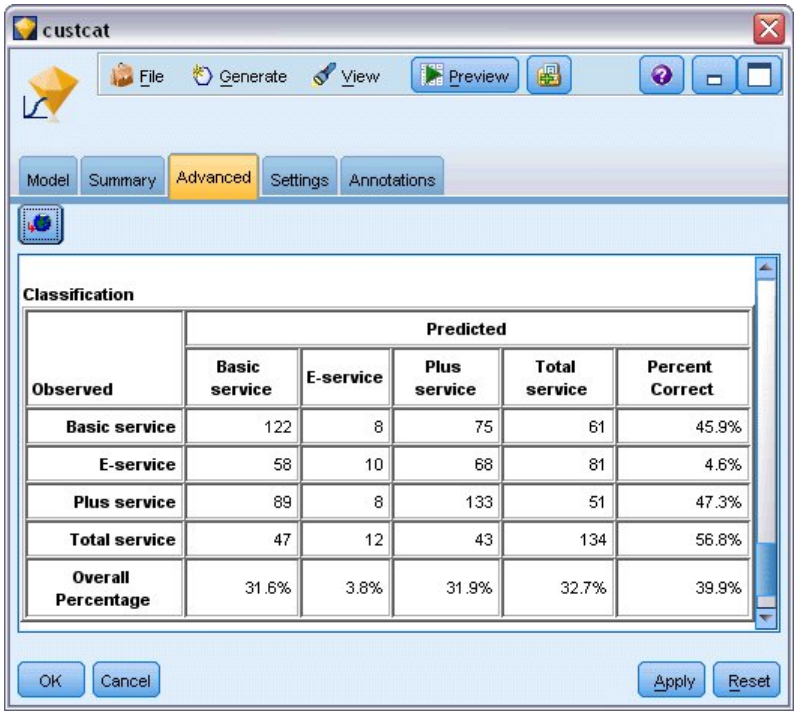


그림 *153.* 분류표

예측할 대상에 따라 모델이 사용자의 요구에 완벽하게 적합할 수 있습니다. 예를 들어, 범주 2의 고객을 식별하는 데 관심이 없는 경우, 해당 모델로 충분할 것입니다. 이는 E-서비스가 수익이 거의 없는 특가품인 경우일 수 있습니다.

예를 들어, 투자수익률(ROI)이 가장 높은 고객이 범주 3 또는 4인 경우, 이 모델은 사용자가 원하는 정보를 제공할 수 있습니다.

모델이 실제로 데이터에 얼마나 적합한지 평가하려면 모델을 작성할 때 고급 출력 대화 상자에서 수많은 진단을 사용할 수 있어야 합니다. IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 설치 디스크의 *\Documentation* 디렉토리에서 사용 가능한 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

이러한 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 실세계에서 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

제 13 장 통신 서비스 제공자를 바꾸는 고객(이항 로지스틱 회귀분석)

로지스틱 회귀분석은 입력 필드 값을 기반으로 레코드를 분류하는 통계 기법입니다. 이는 선형 회귀와 비슷하지만, 숫자 대신 범주형 대상 필드를 사용합니다.

이 예에서는 *telco.sav*라는 데이터 파일을 참조하는 *telco_churn.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *telco_churn.str* 파일은 *streams* 디렉토리에 있습니다.

예를 들어, 통신사업자가 경쟁자에게 빠져나가고 있는 고객 수에 대해 걱정하고 있습니다. 서비스 이용 데이터를 사용하여 다른 제공자로 바꿀 가능성이 있는 고객을 예측할 수 있으면 가능한 한 많은 고객을 보유하도록 제안을 사용자 정의할 수 있습니다.

이 예에서는 고객 손실(서비스 제공자를 바꾸는 고객)을 예측하기 위한 사용 데이터에 초점을 맞춥니다. 대상에 두 개의 고유 범주가 있으므로 이항 모델이 사용됩니다. 다중 범주가 있는 대상의 경우, 다항 모델이 대신 작성될 수 있습니다. 자세한 정보는 139 페이지의 제 12 장 『통신 고객 분류(다항 로지스틱 회귀분석)』의 내용을 참조하십시오.

스트림 작성

1. *Demos* 폴더에서 *telco.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

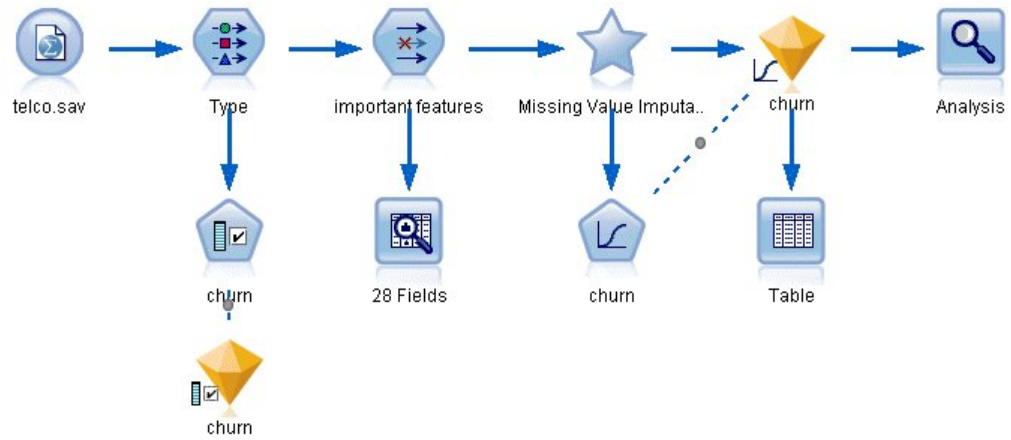


그림 *154.* 이항 로지스틱 회귀분석을 사용하여 고객을 분류하기 위한 샘플 스트림

2. 유형 노드를 추가하고 필드를 정의하고 모든 측정 수준이 올바르게 설정되었는지 확인하십시오. 예를 들어, 값이 0 및 1인 대부분의 필드는 플래그로 간주할 수 있으나 성별과 같은 특정 필드는

두 개의 값이 있는 명목 필드로 간주하는 것이 더 정확합니다.

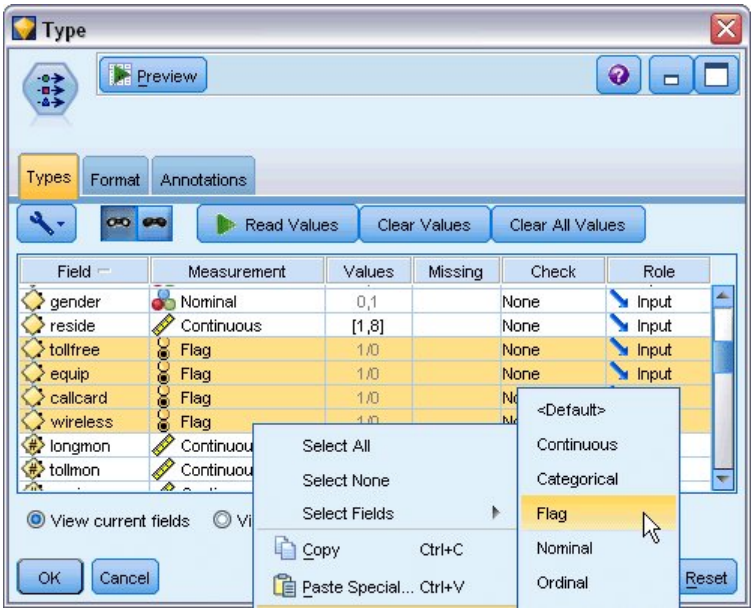


그림 *155.* 다중 필드에 대한 측정 수준 설정

팁: 유사한 값(0/1 등)을 가진 다중 필드에 대한 특성을 변경하려면 값 열 헤더를 클릭하여 필드를 값 기준으로 정렬한 다음 Shift 키를 누른 상태에서 마우스 또는 화살표를 사용하여 변경할 키를 모두 선택하십시오. 그런 다음 마우스 오른쪽 단추로 선택영역을 클릭하여 선택된 필드의 측정 수준 또는 기타 속성을 변경하십시오.

3. 서비스 제공자를 바꾸는 고객 필드에 대한 측정 수준을 **플래그**로 설정하고 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.

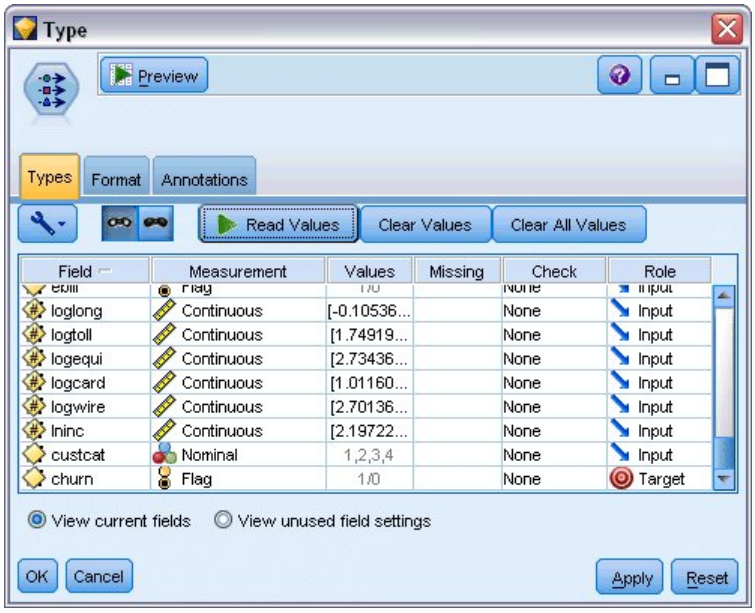


그림 *156.* 서비스 제공자를 바꾸는 고객 필드에 대한 측정 수준 및 역할 설정

4. 필드선택 모델링 노드를 유형 노드에 연결하십시오.

   필드선택 노드를 사용하면 예측변수/대상 관계와 관련하여 유용한 정보를 추가하지 않는 예측변수 또는 데이터를 제거할 수 있습니다.

5. 스트림을 실행하십시오.
6. 결과 모델 너깃을 열고 **생성** 메뉴에서 **필터**를 선택하여 필터 노드를 생성하십시오.

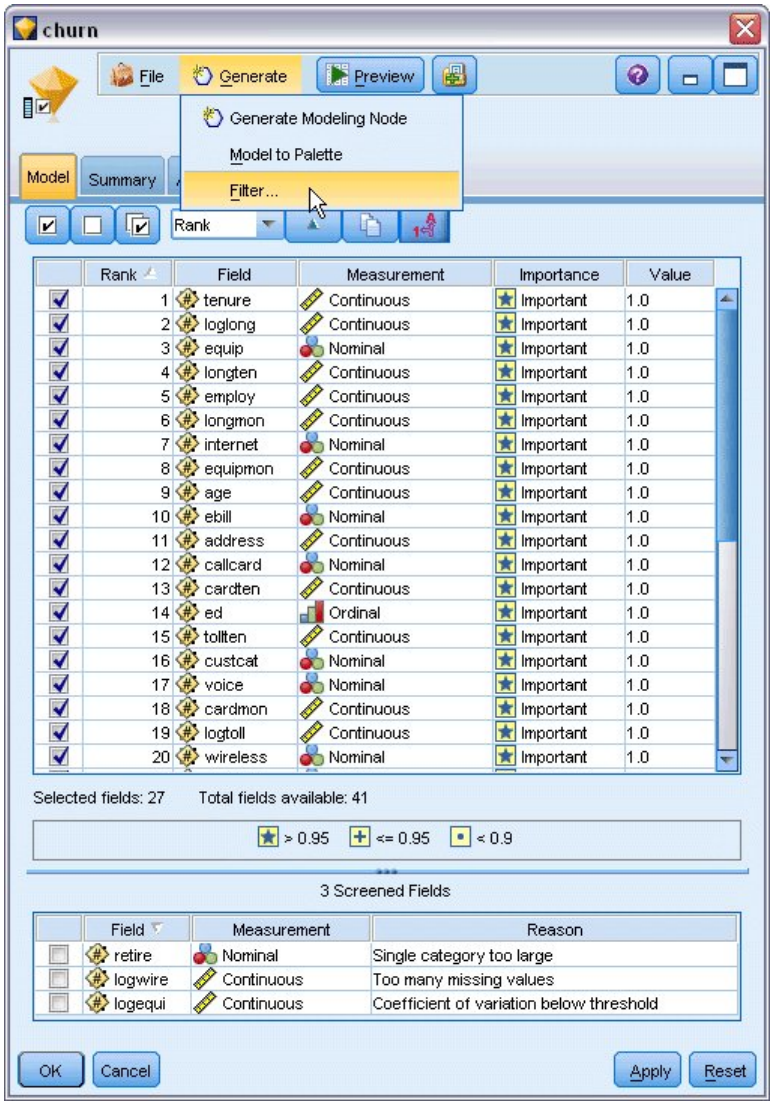


그림 *157.* 필드선택 노드에서 필터 노드 생성

telco.sav 파일의 모든 데이터가 서비스 제공자를 바꾸는 고객 예측에 유용한 것은 아닙니다. 필터를 사용하여 예측변수로 사용하기에 중요한 것으로 간주되는 데이터만 선택할 수 있습니다.

7. 필터 생성 대화 상자에서 **표시된 모든 필드: 중요**를 선택하고 **확인**을 클릭하십시오.
8. 생성된 필터 노드를 유형 노드에 연결하십시오.

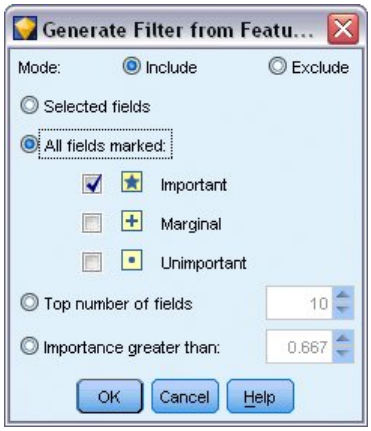


그림 *158.* 중요 필드 선택

9. 데이터 검토 노드를 생성된 필터 노드에 연결하십시오.

   데이터 검토 노드를 열고 **실행**을 클릭하십시오.

10. 데이터 검토 브라우저의 품질 탭에서 % 완료 열을 클릭하여 오름차순 숫자 순서로 열을 정렬하십시오. 그러면 누락 데이터가 많은 모든 필드를 식별할 수 있습니다. 이 경우, 수정해야 하는 유일한 필드는 완료 비율이 50% 미만인 *logtoll*입니다.

11. *logtoll*에 대한 결측값 대치에서 **지정**을 클릭하십시오.

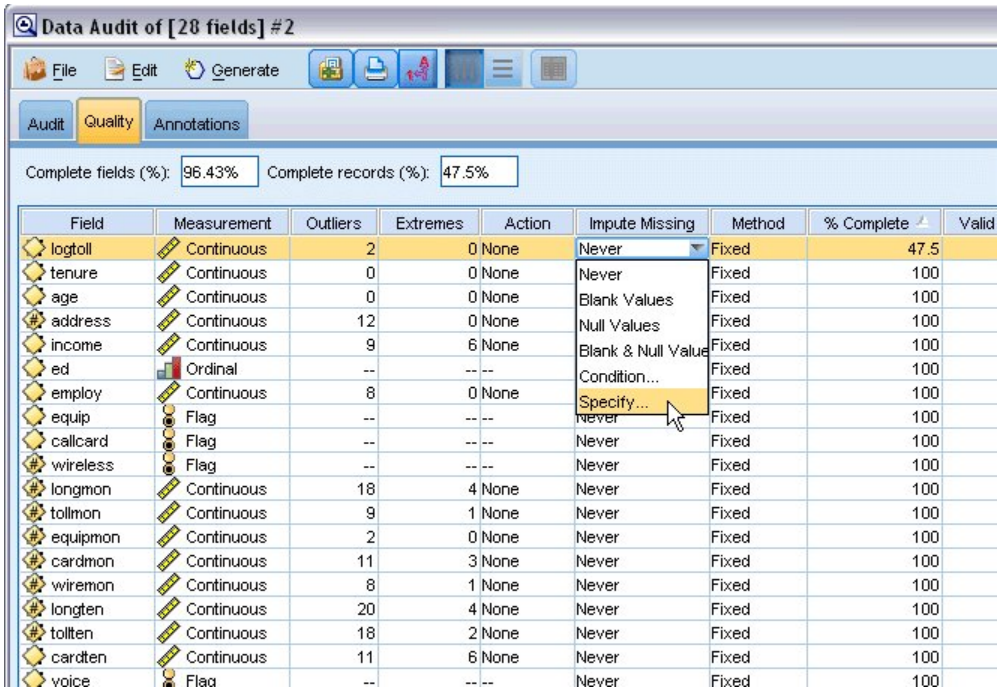


그림 *159. logtoll*에 대한 결측값 대치

12. **다음 경우에 대치**에 대해 **공백값 및 널값**을 선택하십시오. **다음으로 고정**에 대해 **평균**을 선택하고 **확인**을 클릭하십시오.

평균을 선택하면 대치된 값이 전체 데이터의 전체 값의 평균에 악영향을 미치지 않습니다.

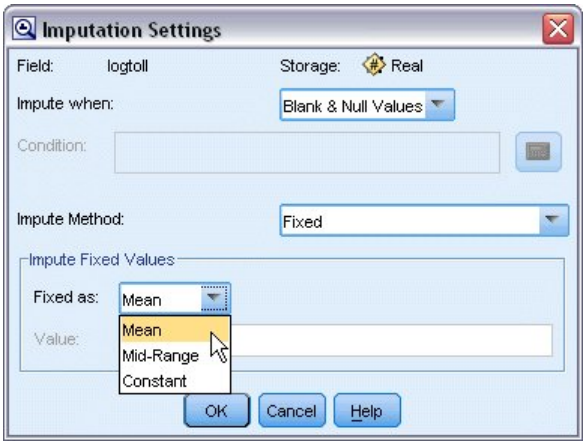


그림 *160.* 대치 설정 선택

13. 데이터 검토 브라우저 품질 탭에서 결측값 수퍼 노드를 생성하십시오. 이를 위해서는 메뉴에서 다음을 선택하십시오.

생성 > **결측값 수퍼 노드**

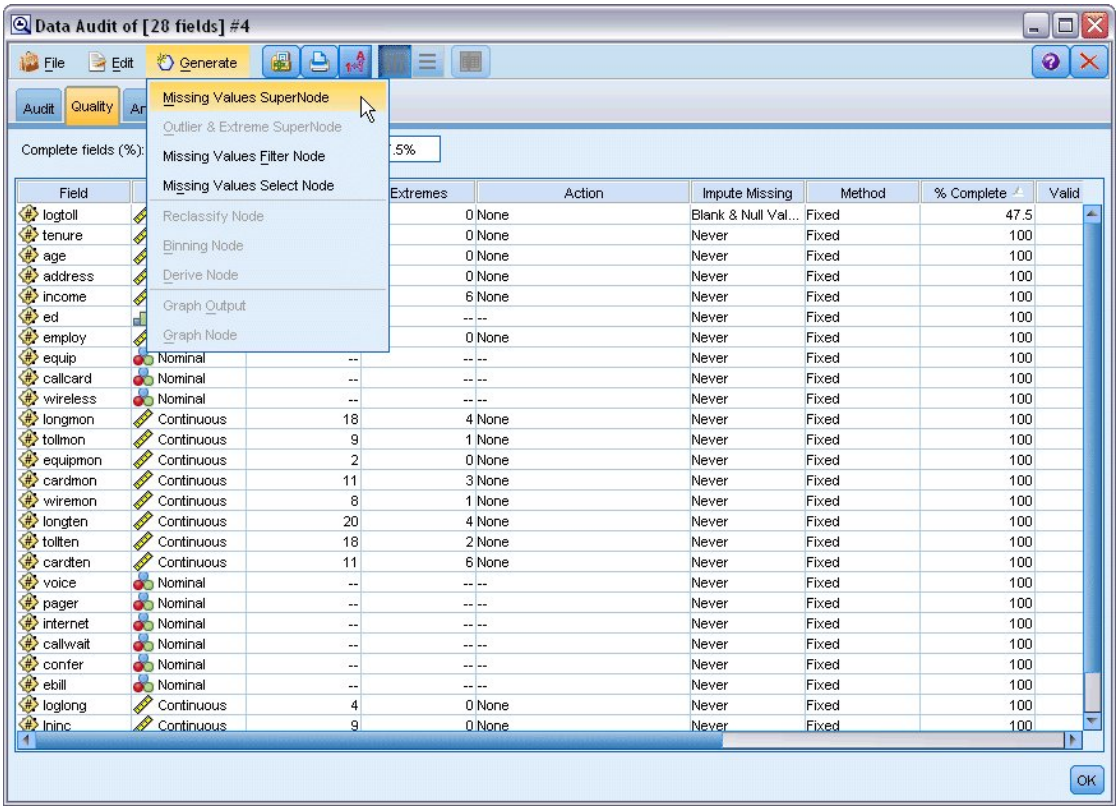


그림 *161.* 결측값 수퍼 노드 생성

결측값 수퍼 노드 대화 상자에서 **표본 크기**를 50%로 늘리고 **확인**을 클릭하십시오.

수퍼 노드가 결측값 대치라는 제목과 함께 스트림 캔버스에 표시됩니다.

14. 수퍼 노드를 필터 노드에 연결하십시오.

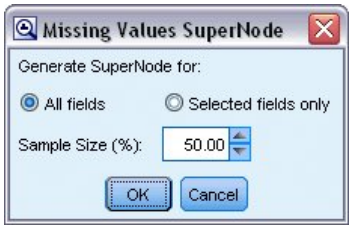


그림 *162.* 표본 크기 지정

15. 로지스틱 노드를 수퍼 노드에 추가하십시오.
16. 로지스틱 노드에서 모델 탭을 클릭하여 **이항** 프로시저를 선택하십시오. 이항 프로시저 영역에서 **전진** 방법을 선택하십시오.

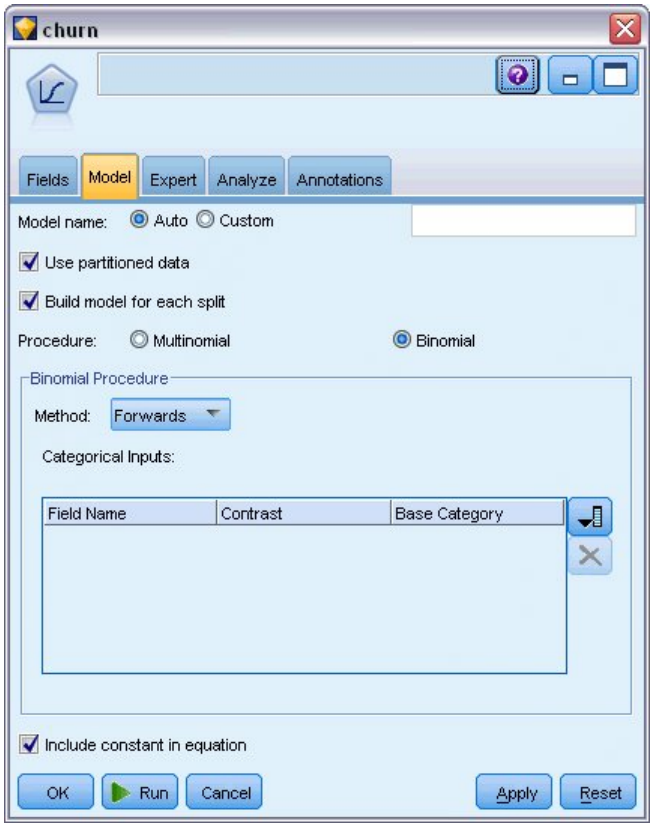


그림 *163.* 모델 옵션 선택

17. 전문가 탭에서 모델을 **전문가** 모드를 선택한 다음 **출력**을 클릭하십시오. 고급 출력 대화 상자가 표시됩니다.
18. 고급 출력 대화 상자에서 **표시** 유형으로 각 단계마다를 선택하십시오. **반복계산 히스토리** 및 **모수 추정값**을 선택하고 **확인**을 클릭하십시오.

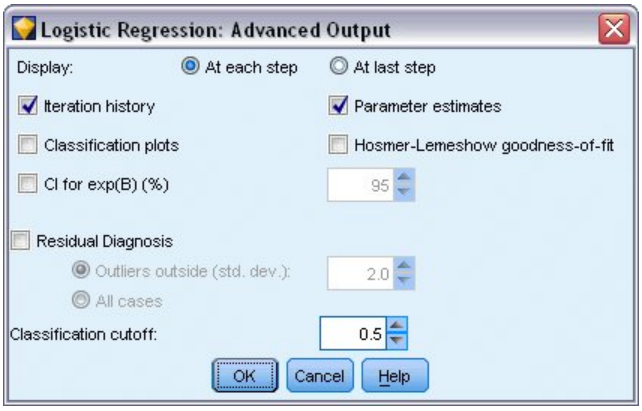


그림 *164.* 출력 옵션 선택

모델 찾아보기

1. 로지스틱 노드에서 **실행**을 클릭하여 모델을 작성하십시오.

   모델 너깃이 스트림 캔버스에 추가되고 오른쪽 상단 코너에 있는 모델 팔레트에도 추가됩니다. 세부사항을 보려면 마우스 오른쪽 단추로 모델 너깃을 클릭하고 **편집** 또는 **찾아보기**를 선택하십시오.

   요약 탭은 (다른 사항 사이에서) 대상 및 모델에 의해 사용되는 입력(예측변수 필드)을 표시합니다. 이러한 필드는 실제로 전진 기법을 기준으로 하여 선택된 필드이며 고려하도록 제출된 전체 목록이 아닙니다.

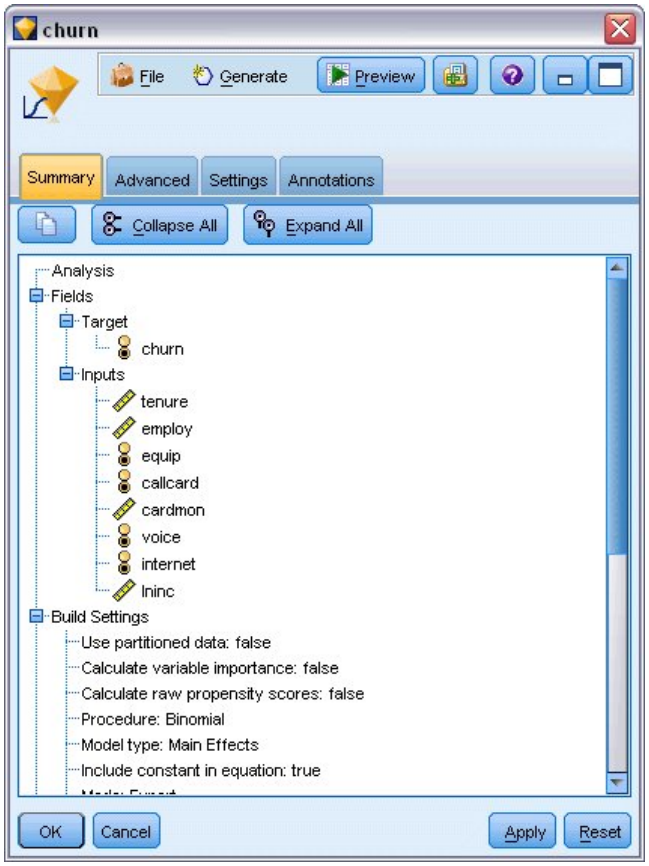


그림 *165.* 대상 및 입력 필드를 표시하는 모델 요약

고급 탭에 표시되는 항목은 로지스틱 노드의 고급 출력 대화 상자에서 선택된 옵션에 따라 다릅니다. 항상 표시되는 한 항목은 분석에 포함된 레코드의 수 및 퍼센트를 표시하는 케이스 처리 요약입니다. 또한 하나 이상의 입력 필드가 사용 불가능하고 어떠한 케이스도 선택되지 않은 누락 케이스(있는 경우에 한함)의 수를 나열합니다.

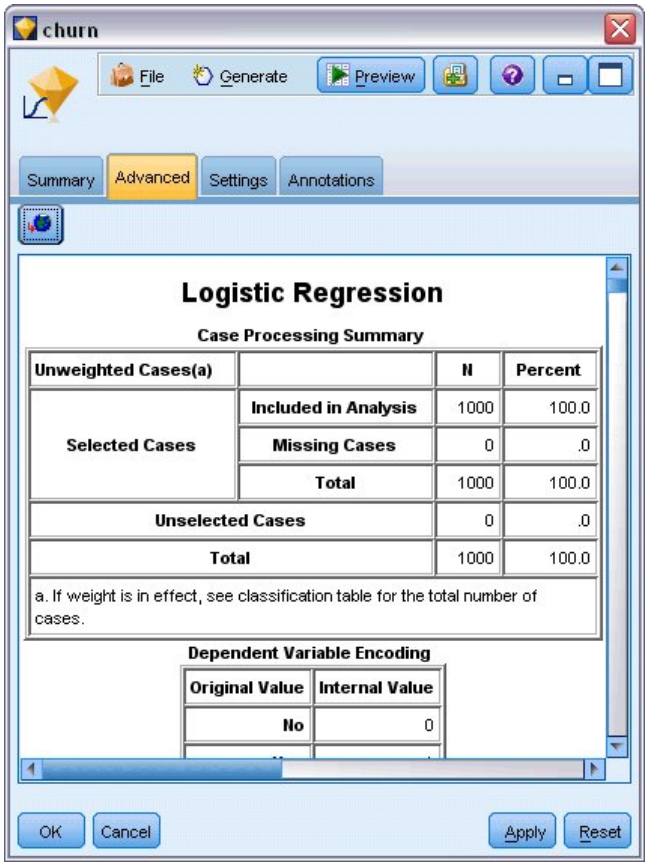


그림 *166.* 케이스 처리 요약

2. 케이스 처리 요약에서 아래로 스크롤하여 블록 0: 시작 블록 아래의 분류표을 표시할 수 있습니다.

   단계별 전진 방법은 최종 작성 모델과 비교할 때 기초가 될 수 있는 널 모델, 즉, 예측변수가 없는 모델부터 시작합니다. 널 모델은 편의상 모든 것을 0으로 예측합니다. 따라서 서비스 제공자를 바꾸지 않은 726명의 고객이 올바르게 예측되었다는 이유만으로 널 모델의 정확도가 72.6%가 됩니다. 단, 서비스 제공자를 바꾼 고객은 전혀 올바르게 예측되지 않았습니다.

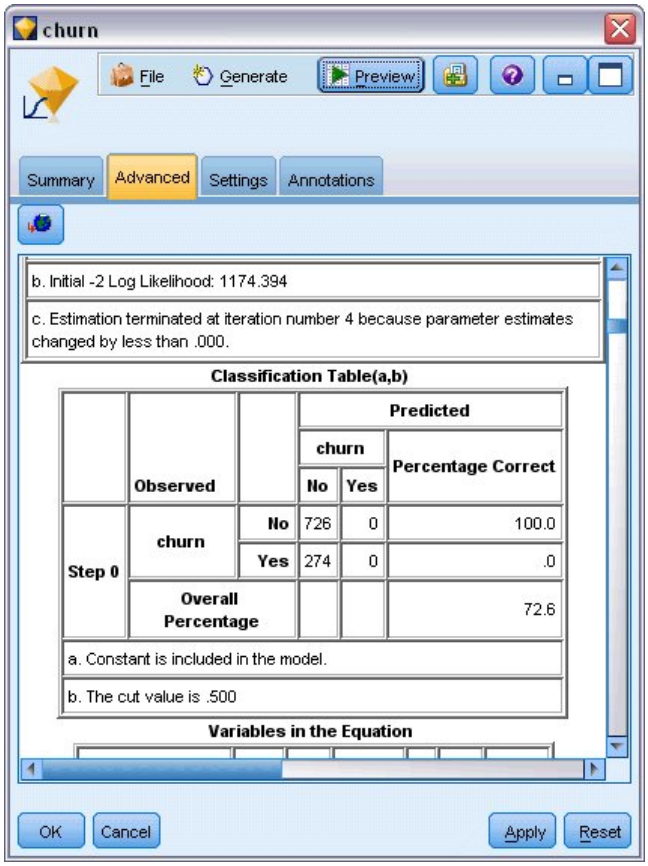


그림 *167.* 분류표 - 블록 *0* 시작

3. 이제 아래로 스크롤하여 블록 1: 방법 = 단계별 전진 아래에 분류표을 표시하십시오.

   이 분류표은 각 단계에서 예측변수가 추가될 때 모델에 대한 결과를 표시합니다. 단 하나의 예측변수만 사용된 후 첫 번째 단계에서 이미 서비스 제공자를 바꾸는 고객 예측에서 정확도가 0.0%에서 29.9%로 증가했습니다.

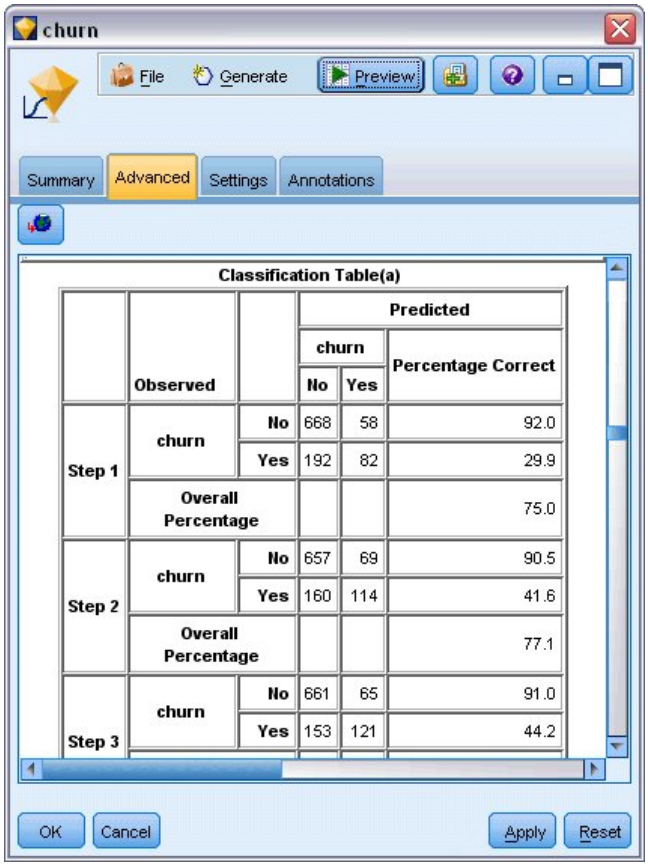


그림 *168.* 분류표 - 블록 *1*

4. 이 분류표의 맨 아래쪽으로 스크롤하십시오.

분류표은 마지막 단계가 8단계라는 것을 표시합니다. 이 단계에서 알고리즘이 더 이상 모델에 추가 예측변수를 추가할 필요가 없다고 판단했습니다. 서비스 제공자를 바꾸지 않는 고객의 정확도가 91.2%로 약간 낮아졌지만 서비스 제공자를 바꾼 고객에 대한 정확도는 원래 0%에서 47.1%로 높아졌습니다. 이는 예측변수를 사용하지 않았던 원래 널 모델에 비해 유의적인 개선도입니다.

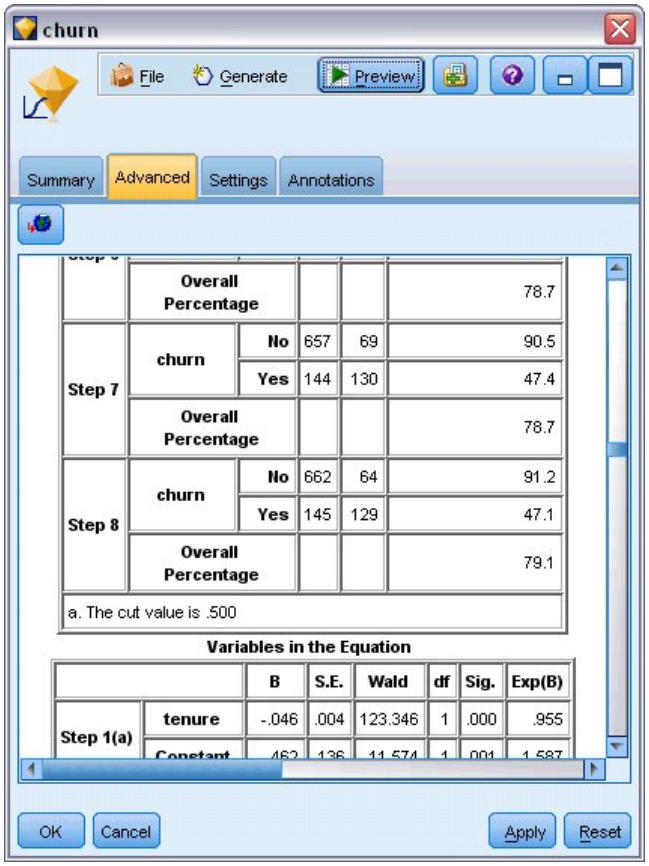


그림 *169.* 분류표 - 블록 *1*

서비스 제공자를 바꾸는 사용자를 감소시키고자 하는 고객에 대해 이를 거의 절반으로 감소시킬 수 있으면 수입 스트림을 보호하기 위한 주요한 단계가 될 것입니다.

참고: 또한 이 예는 전체 퍼센트를 지침으로 사용하는 경우에 어떤 경우에 어떻게 모델의 정확도가 오도될 수 있는지 표시합니다. 원래 널 모델의 전체 정확도가 72.6%인 반면 최종적으로 예측된 모델의 전체 정확도는 79.1%입니다 그러나 앞에서 보았듯이 실제 개별 범주 예측의 정확도는 매우 달랐습니다.

모델이 실제로 데이터에 얼마나 적합한지 평가하려면 모델을 작성할 때 고급 출력 대화 상자에서 수많은 진단을 사용할 수 있어야 합니다. IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 설치 디스크의 *\Documentation* 디렉토리에서 사용 가능한 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

이러한 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 실세계에서 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

제 14 장 대역폭 활용 시계열 분석(시계열)

시계열 노드를 사용하여 예측

국내 광대역 제공자의 분석가가 대역폭 사용을 예측하기 위해 사용자 가입을 예측해야 합니다. 국내 가입자 기반을 구성하는 각 현지 시장에 대한 예측값이 필요합니다. 시계열 모델링을 사용하여 향후 3개월 동안 여러 현지 시장에 대한 예측을 생성할 것입니다. 두 번째 예는 소스 파일이 시계열 노드에 입력하기에 적합한 형식이 아닌 경우에 이를 변환하는 방법을 표시합니다.

이러한 예에서는 *broadband_1.sav*라는 데이터 파일을 참조하는 *broadband_create_models.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 폴더에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다.
broadband_create_models.str 파일은 스트림 폴더에 있습니다.

마지막 예는 또 다른 3개월로 예측을 확장하기 위해 업데이트된 데이터 세트에 저장된 모델을 적용하는 방법을 보여줍니다.

IBM SPSS Modeler에서 단일 작업으로 다중 시계열 모델을 생성할 수 있습니다. 사용하는 소스 파일에 85개의 서로 다른 시장에 대한 시계열 데이터가 있으나 단순하게 작업할 수 있도록 이러한 시장 중 다섯 개 및 모든 시장에 대한 총계만 모델링할 것입니다.

broadband_1.sav 데이터 파일에는 85개의 각 현지 시장에 대한 사용 데이터가 있습니다. 이 예의 용도로는 처음 다섯 개의 계열만 사용됩니다. 이러한 다섯 개의 계열 및 총계에 대한 별도의 모델이 작성됩니다.

또한 파일에는 각 레코드의 월 및 연도를 표시하는 날짜 필드가 포함됩니다. 이 필드는 레코드에 레이블을 지정하기 위해 사용됩니다. 날짜 필드를 문자열로 IBM SPSS Modeler로 읽어 오지만 필드를 IBM SPSS Modeler에서 사용하기 위해서는 채움 노드를 사용하여 저장 유형을 숫자 날짜 형식으로 변환해야 합니다.

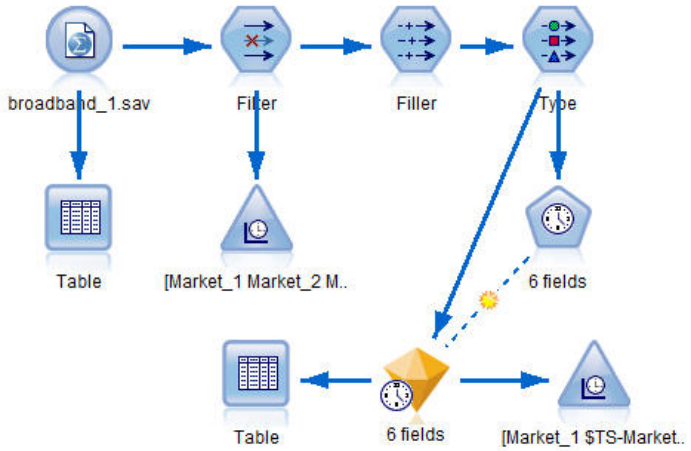
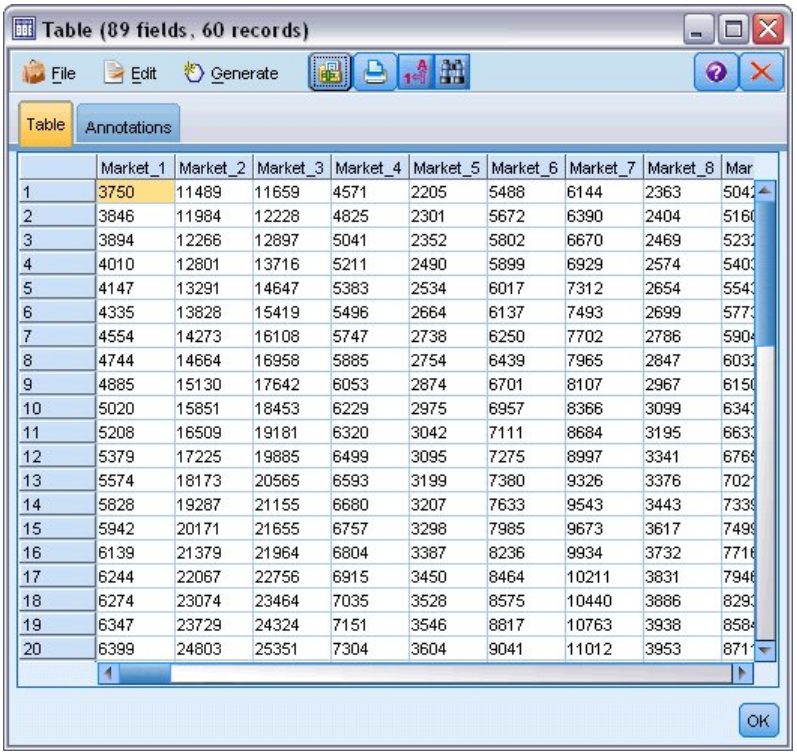


그림 *170.* 시계열 모델링을 표시하기 위한 샘플 스트림

시계열 노드의 경우, 각 계열이 각 구간에 대한 행이 있는 별도의 열에 있어야 합니다. IBM SPSS Modeler에서는 필요한 경우 데이터가 이 형식과 일치하도록 데이터를 변환하는 방법을 제공합니다.



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	504
2	3846	11984	12228	4825	2301	5672	6390	2404	516
3	3894	12266	12897	5041	2352	5802	6670	2469	523
4	4010	12801	13716	5211	2490	5899	6929	2574	540
5	4147	13291	14647	5383	2534	6017	7312	2654	554
6	4335	13828	15419	5496	2664	6137	7493	2699	577
7	4554	14273	16108	5747	2738	6250	7702	2786	590
8	4744	14664	16958	5885	2754	6439	7965	2847	603
9	4885	15130	17642	6053	2874	6701	8107	2967	615
10	5020	15851	18453	6229	2975	6957	8366	3099	634
11	5208	16509	19181	6320	3042	7111	8684	3195	663
12	5379	17225	19885	6499	3095	7275	8997	3341	676
13	5574	18173	20565	6593	3199	7380	9326	3376	702
14	5828	19287	21155	6680	3207	7633	9543	3443	733
15	5942	20171	21655	6757	3298	7985	9673	3617	749
16	6139	21379	21964	6804	3387	8236	9934	3732	771
17	6244	22067	22756	6915	3450	8464	10211	3831	794
18	6274	23074	23464	7035	3528	8575	10440	3886	829
19	6347	23729	24324	7151	3546	8817	10763	3938	858
20	6399	24803	25351	7304	3604	9041	11012	3953	871

그림 *171.* 광대역 현지 시장에 대한 월별 가입 데이터

스트림 작성

1. 새 스트림을 작성하고 *broadband_1.sav*를 가리키는 통계 파일 소스 노드를 추가합니다.

2. 필터 노드를 사용하여 *Market_6*에서 *Market_85*까지의 필드 및 *MONTH_* 및 *YEAR_* 필드를 필터링하여 모델을 단순화합니다.

팁: 단일 작업에서 인접한 다중 필드를 선택하려면 *Market_6* 필드를 클릭하고 마우스 왼쪽 단추를 누른 상태로 마우스를 *Market_85* 필드까지 아래로 끄십시오. 선택된 필드가 파란색으로 강조표시됩니다. 다른 필드를 추가하려면 Ctrl 키를 아래로 누른 상태에서 *MONTH_* 및 *YEAR_* 필드를 클릭하십시오.

그림 *172.* 모델 단순화

데이터 탐색

모델을 설정하기 전에 항상 데이터의 특성을 알아보는 것이 좋습니다. 데이터가 계절 변동을 나타냅니까? 자동 모델 생성기가 각 계열에 대해 자동으로 최선의 계절성 또는 비계절성 모델을 찾을 수 있지만 데이터에 계절성이 없는 경우에는 검색을 비계절성 모델로 제한함으로써 더 빠르게 결과를 얻을 수 있는 경우가 있습니다. 각 현지 시장에 대해 데이터를 탐색하지 않고도 전체 다섯 개의 시장에 대해 가입자의 총 수를 도표로 작성하여 계절성의 존재 또는 부재에 대해 대략적으로 파악할 수 있습니다.

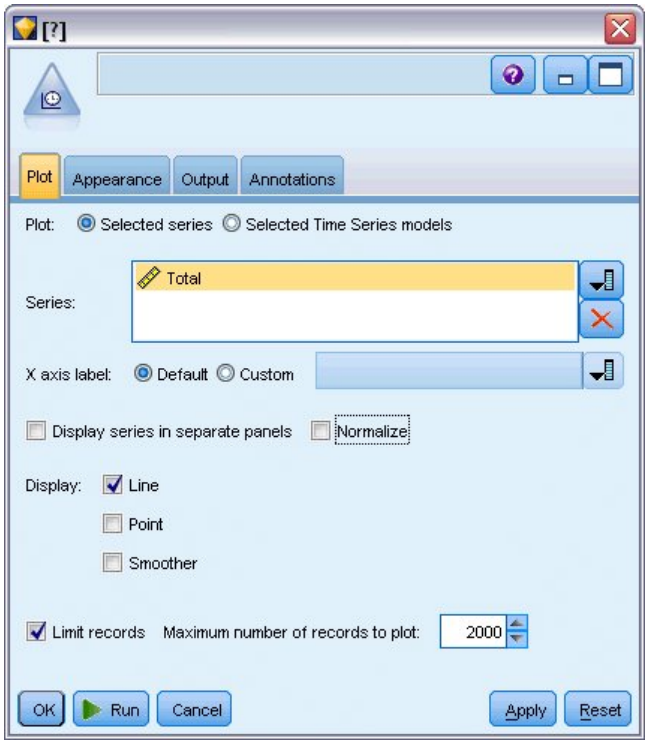


그림 *173.* 가입자의 총 수 도표 작성

1. 그래프 팔레트로부터 시간 구성 노드를 필터 노드에 첨부하십시오.
2. 총계 필드를 계열 목록에 추가하십시오.
3. **별도 패널에 계열 표시** 및 **표준화** 선택란을 선택 취소하십시오.
4. **실행**을 클릭하십시오.

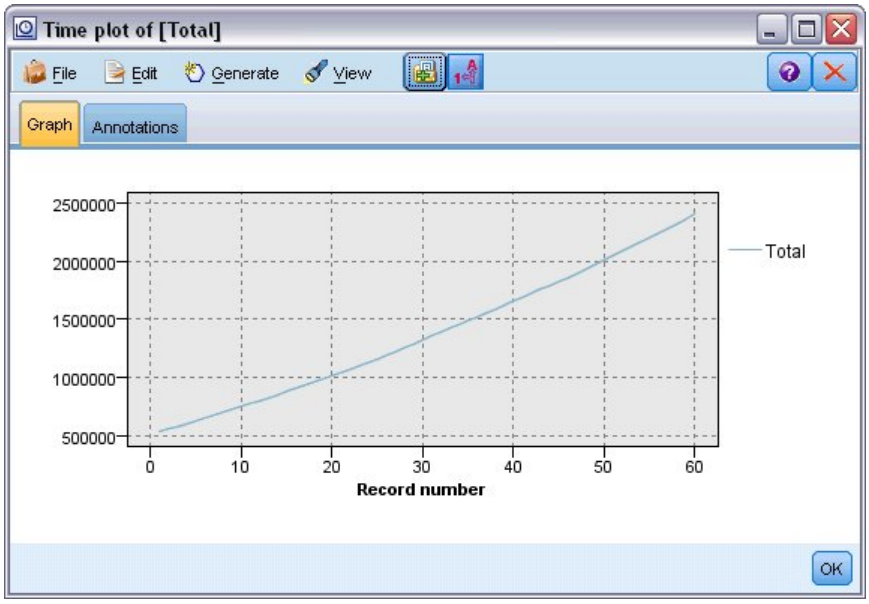


그림 *174.* 총계 필드의 시간 도표

계열이 계절 변동을 나타내지 않는 평활 상향 추세를 나타냅니다. 계절이 있는 개별 계절성이 있을 수 있지만 일반적으로 계절성이 데이터의 두드러진 특징이 아닌 것으로 나타납니다.

계절 모델을 제외하기 전에 각 계열을 조사하는 것이 좋습니다. 그런 다음 계절이 나타나는 계절성을 구별하여 별도로 모델링할 수 있습니다.

IBM SPSS Modeler를 사용하면 쉽게 다중 계열을 함께 도표로 작성할 수 있습니다.

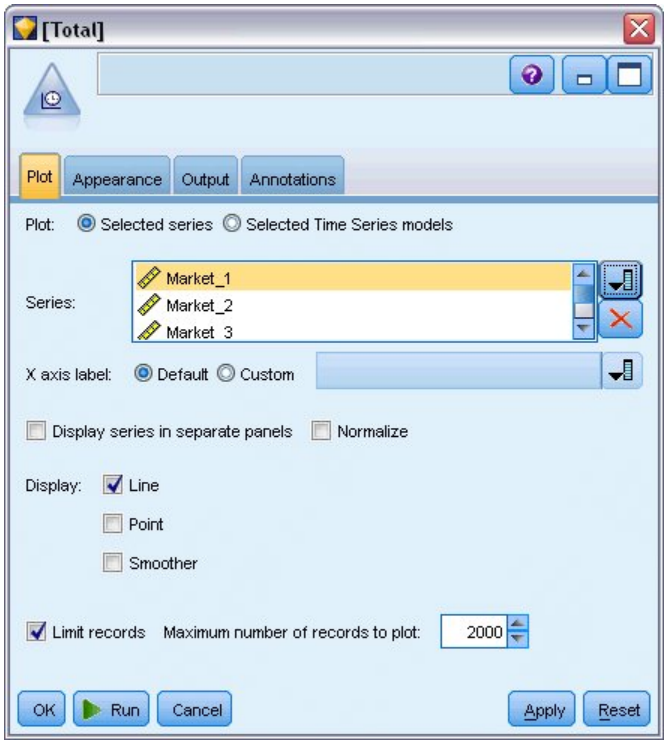


그림 *175.* 다중 시계열 도표 작성

5. 시간 구성 노드를 다시 여십시오.
6. 총계 필드를 계열 목록에서 제거하십시오(해당 필드를 선택한 다음 빨간색 X 단추를 클릭하십시오).
7. *Market_1*에서 *Market_5*까지의 필드를 목록에 추가하십시오.
8. **실행**을 클릭하십시오.

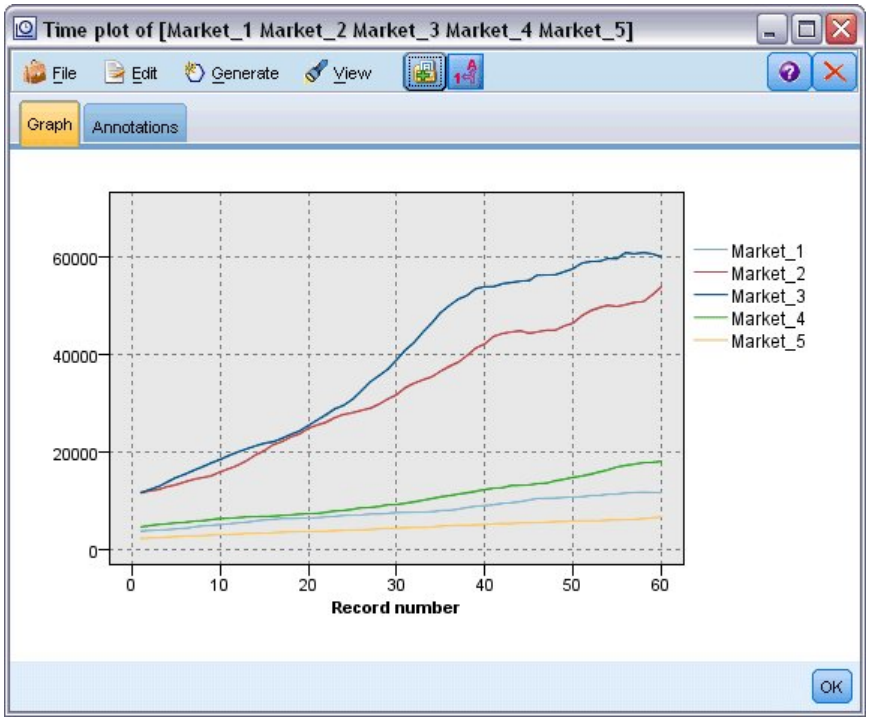


그림 *176.* 다중 필드의 시간 도표

각 시장을 조사한 결과 각 케이스에서 안정된 상승세가 발견되었습니다. 일부 시장이 다른 시장보다 조금 더 탄력적이라 하더라도 계절성이 보인다는 증거는 없습니다.

날짜 정의

이제 *DATE_* 필드의 저장 유형을 날짜 형식으로 변경해야 합니다.

1. 채움 노드를 필터 노드에 첨부하십시오.
2. 채움 노드를 열고 필드 선택기 단추를 클릭하십시오.
3. **DATE_**를 선택하여 이를 **필드 채우기**에 추가하십시오.
4. **바꾸기** 조건을 **항상**으로 설정하십시오.
5. **바꿀 문자열**의 값을 **to_date(DATE_)**로 설정하십시오.

그림 *177.* 날짜 저장 유형 설정

날짜 필드의 형식과 일치하도록 기본 날짜 형식을 변경하십시오. 예상대로 작업하기 위해서는 날짜 변환이 필수입니다.

6. 메뉴에서 **도구 > 스트림 특성 > 옵션**을 선택하여 스트림 옵션 대화 상자를 표시하십시오.
7. **날짜/시간** 분할창을 선택하고 기본 **날짜 형식**을 **MON YYYY**로 설정하십시오.

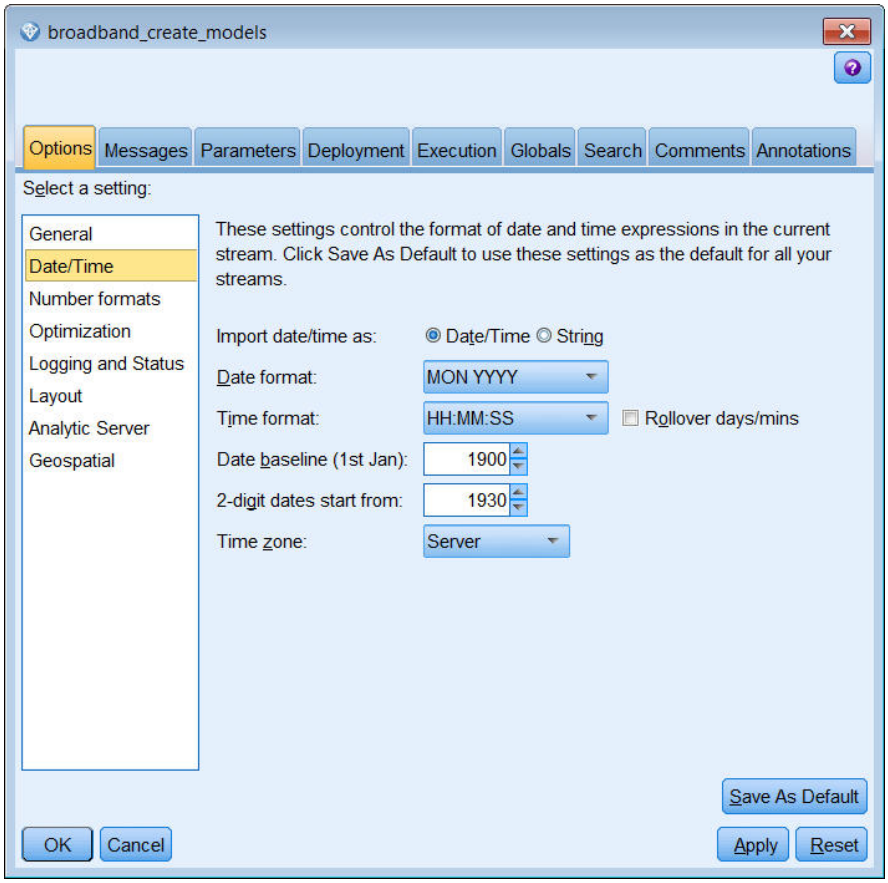


그림 *178.* 날짜 형식 설정

대상 정의

1. 유형 노드를 추가하고 *DATE_* 필드에 대한 역할을 **없음**으로 설정하십시오. 모든 기타(*Market_n* 필드 및 총계 필드)에 대한 역할을 **대상**으로 설정하십시오.
2. **실제 값** 단추를 클릭하여 값 열을 채우십시오.

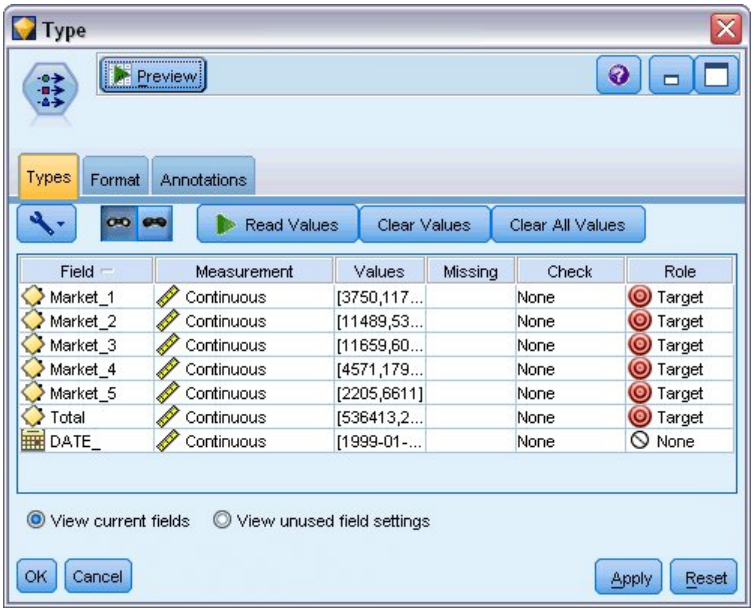


그림 *179.* 다중 필드에 대한 역할 설정

시간 구간 설정

1. 모델링 팔레트에서 시계열 노드를 스트림에 추가하고 이를 유형 노드에 첨부하십시오.
2. 데이터 지정 사항 탭의 관측값 분할창에서 **날짜/시간 필드**로 DATE_를 선택하십시오.
3. **시간 구간**으로 Months를 선택하십시오.

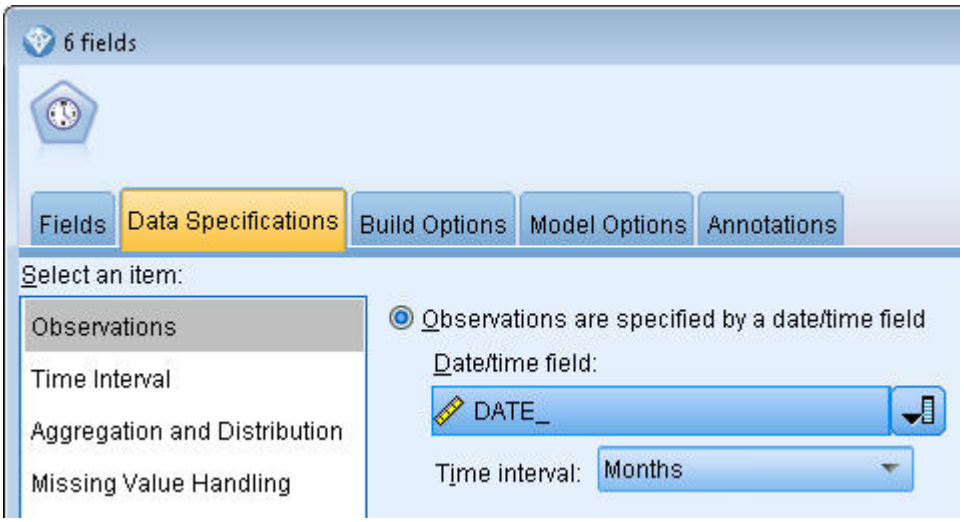


그림 *180.* 시간 구간 설정

4. 모델 옵션 탭에서 **레코드를 미래로 확장** 선택란을 선택하십시오.
5. 값을 **3**으로 설정하십시오.

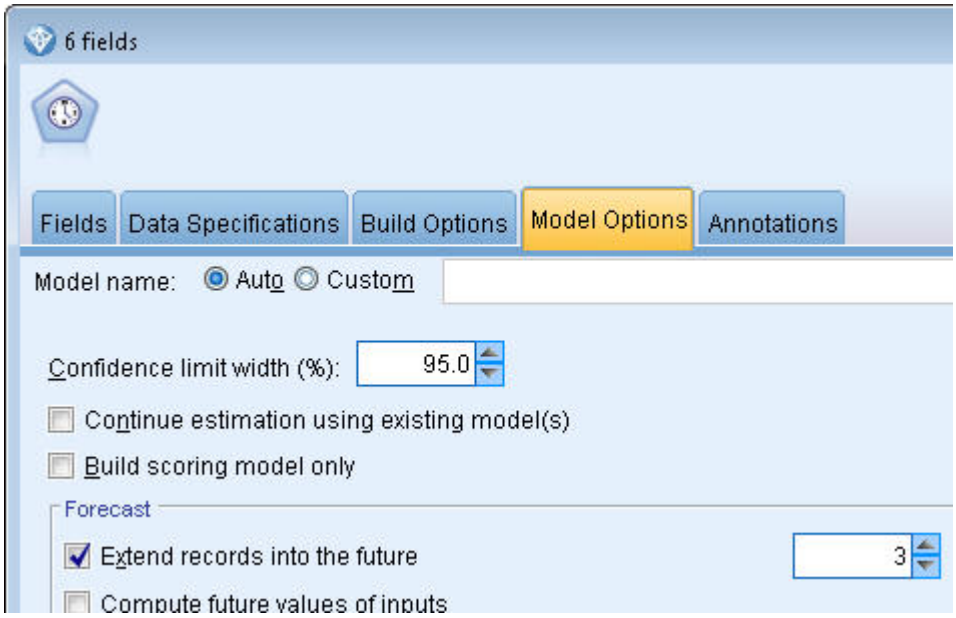


그림 *181.* 예측 기간 설정

모델 작성

1. 시계열 노드에서 필드 탭을 선택하십시오. **필드** 목록에서 다섯 개의 시장을 모두 선택하여 **대상** 및 **후보 입력** 목록 둘 다에 복사하십시오. 또한 총계 필드를 선택하여 **대상** 목록에 복사하십시오.
2. 일반 분할창에서 작성 옵션 탭을 선택하고 자동 모델 생성기 **방법**이 모든 기본 설정을 사용하여 선택되었는지 확인하십시오. 그러면 자동 모델 생성기가 각 시계열에 대해 사용할 가장 적합한 모델을 결정할 수 있습니다.**실행**을 클릭하십시오.

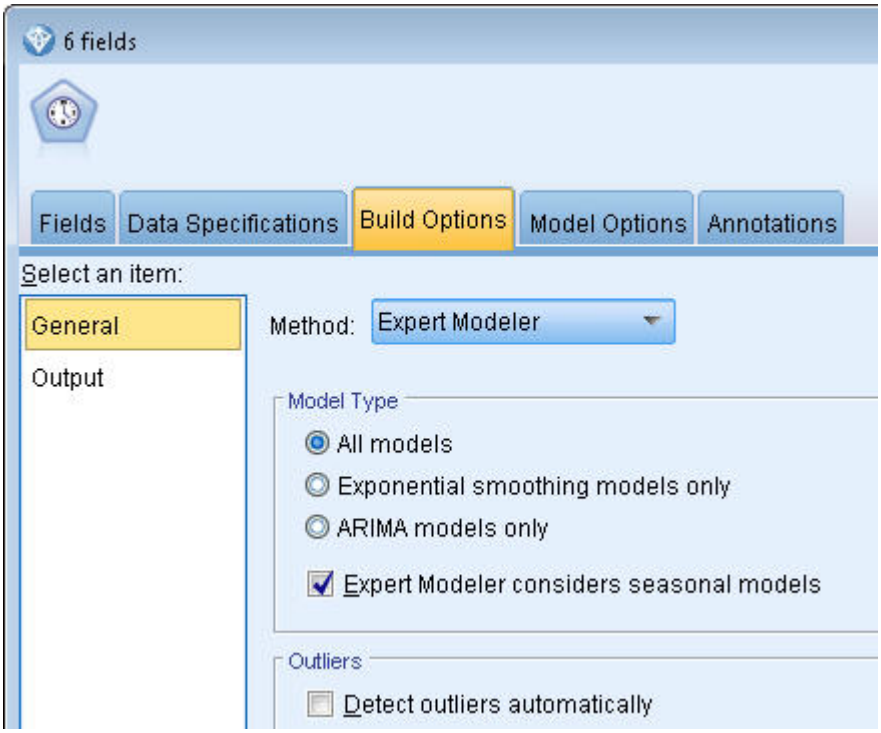


그림 *182.* 시계열에 대한 자동 모델 생성기 선택

3. 시계열 노드에 시계열 모델 너깃을 첨부하십시오.
4. 테이블 노드를 시계열 모델 너깃에 첨부하고 **실행**을 클릭하십시오.

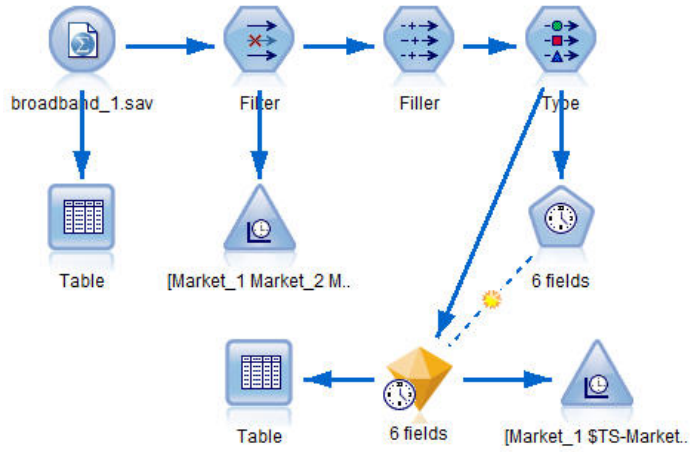


그림 *183.* 시계열 모델링을 표시하기 위한 샘플 스트림

이제 원 데이터에 세 개의 새 행(61에서 63까지)이 추가되었습니다. 이러한 행은 예측 기간에 대한 행이며 이 케이스의 경우, 2004년 1월에서 3월까지입니다.

몇 가지 새 열도 지금 표시됩니다. 즉, 시계열 노드별 *$TS-* 열이 추가됩니다. 열은 각 행(예: 시계열 데이터의 각 행)에 대해 다음을 표시합니다.

Column	설명
$TS-*colname*	원 데이터의 각 열에 대해 생성된 모델 데이터입니다.
$TSLCI-*colname*	생성된 모델 데이터의 각 열에 대한 하한 신뢰구간 값입니다.
$TSUCI-*colname*	생성된 모델 데이터의 각 열에 대한 상한 신뢰구간 값입니다.
$TS-Total	이 행에서 $TS-*colname* 값의 총계.
$TSLCI-Total	이 행에서 $TSLCI-*colname* 값의 총계입니다.
$TSUCI-Total	이 행에 대한 $TSUCI-*colname* 값의 총계입니다.

예측 작업에 대한 가장 유의적 열은 *$TS-Market_n*, *$TSLCI-Market_n* 및 *$TSUCI-Market_n* 열입니다. 특히 61에서 63까지의 이러한 열은 사용자 가입 데이터 및 각 현지 시장의 신뢰구간을 포함합니다.

모델 탐색적 데이터분석

1. 시계열 모델 너깃을 두 번 클릭하고 출력 탭을 선택하여 각 시장에 대해 생성된 모델에 관한 데이터를 표시하십시오.

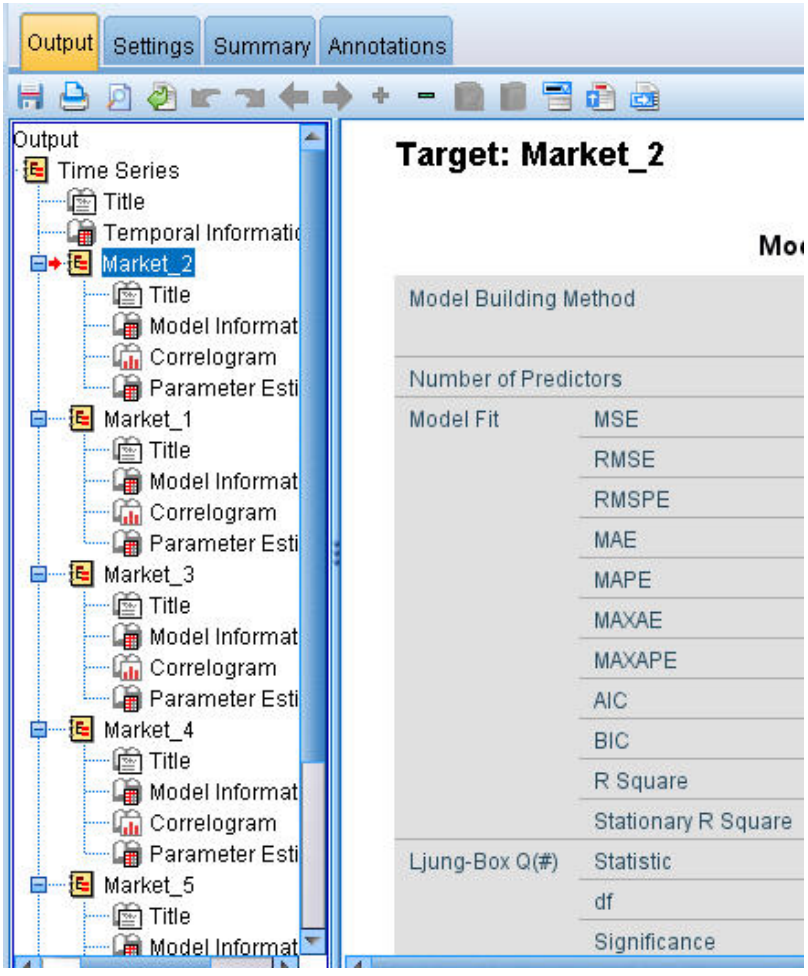


그림 *184.* 시장에 대해 생성된 시계열 모델

왼쪽 출력 열에서 임의의 시장에 대한 **모델 정보**를 선택하십시오. **예측변수 수** 줄은 각 대상에 대해 예측변수로 사용된 필드의 수를 표시합니다. 이 케이스의 경우, 없습니다.

모델 정보의 나머지 줄은 각 모델에 대한 다양한 적합도 측도를 표시합니다. **정상 R 제곱** 값은 모델이 설명하는 계열의 총 변동 비율에 대한 추정값을 제공합니다. 값이 높을수록(최대값은 1.0임) 모델의 적합도가 높습니다.

Q(#) 통계, **자유도** 및 **유의성** 줄은 Ljung-Box 통계, 모델 내의 잔차 오류의 무작위성 검증과 연관됩니다. 오류가 무작위적일수록 더 나은 모델일 확률이 높습니다. **Q(#)**는 Ljung-Box 통계 자체인 반면 **자유도**는 특정 목표 추정 시 달라질 수 있는 모델 모수의 수를 표시합니다.

유의성 줄은 모델이 올바르게 지정되었는지 여부에 대한 또 다른 지표는 제공하는 Ljung-Box 통계의 유의수준을 제공합니다. 유의수준이 0.05 미만이면 잔차 오차가 임의가 아님을 의미합니다. 즉, 모델에 의해 고려되지 않은 관측된 계열 내의 구조가 있음을 의미합니다.

정상 R 제곱 및 **유의성** 값 둘 모두를 고려할 때 자동 모델 생성기가 *Market_3* 및 *Market_4*에 대해 선택한 모델이 매우 적합합니다. *Market_1*, *Market_2* 및 *Market_5*의 **유의성** 값이 둘 다 0.05 미만이며 이러한 시장에 대해 더 적합한 모델을 사용한 몇 가지 실험이 필요함을 의미합니다.

추가 적합도 측도의 수가 표시됩니다. **R 제곱** 값은 모델이 설명하는 시계열의 총 변동에 대한 추정을 제공합니다. 이 통계에 대한 최대값은 1.0이며 이 면에서 해당 모델은 양호합니다.

RMSE는 제곱근 평균제곱오차이며 계열의 실제 값이 모델에 의해 예측된 값과 얼마나 다른지에 대한 측도이며 계열 자체에 대해 사용된 단위와 동일한 단위로 표시됩니다. 이는 오차의 측정이므로 이 값이 가능한 한 낮기를 원합니다. 첫 눈에 보기에 *Market_2* 및 *Market_3*에 대한 모델이 지금까지 봐 온 통계에 따르면 여전히 허용 가능한 것으로 보이나 다른 세 시장에 비해서는 덜 성공적인 것으로 보입니다.

이러한 추가 적합도 측도에는 평균 절대 퍼센트 오차(**MAPE**) 및 이의 최대값(**최대 절대 퍼센트 오차(MAXAPE)**)이 포함됩니다. 절대 퍼센트 오차는 대상 계열이 모델 예측 수준에서 얼마나 달라지는지에 대한 측도이며 퍼센트 값으로 표시됩니다. 모든 모델에 대한 평균 및 최대를 검토하여 예측의 불확실성을 확인할 수 있습니다.

평균 절대 퍼센트 오차(MAPE) 값은 모든 모델이 약 1%의 평균 불확실성을 나타내고 있음을 표시하며 이는 매우 낮은 수치입니다. 최대 절대 퍼센트 오차(MAXAPE) 값은 최대 절대 퍼센트 오차를 표시하며 예측에 대한 최악의 케이스 시나리오를 예상하는 데 유용합니다. 이는 대부분의 모델에 대한 가장 큰 퍼센트 오차가 대략 1.8%에서 3.7% 사이에 해당함을 보여주며 이 또한 *Market_4*만 거의 7% 이상으로 그림에서 매우 낮은 변수군입니다.

MAE(평균 절대 오차) 값은 예측 오차의 절대값의 평균을 표시합니다. 제곱근 평균제곱오차(RMSE) 값과 같이 계열 자체와 동일한 단위로 표현됩니다. **최대 절대 오차(MAXAE)**는 동일한 단위의 가장 큰 예측 오차를 표시하며 예측값에 대한 최악의 시나리오를 표시합니다.

이러한 절대값이 흥미롭기는 하지만 대상 계열이 가변적인 크기의 마켓에 대한 가입자 수를 나타내므로 이 케이스에서 더 유용한 값은 퍼센트 오차(MAPE 및 MAXAPE)의 값입니다.

MAPE 및 MAXAPE 값이 모델에 대해 허용 가능한 불확실성 양을 나타냅니까? 확실히 매우 낮습니다. 허용 가능한 위험은 문제점에 따라 변경되므로 이는 비즈니스 감각이 작용해야 하는 상황입니다. 이제 적합도 통계가 허용 가능한 범위에 있다고 가정하고 잔차 오차를 살펴볼 것입니다.

모델 잔차에 대한 자기상관 함수(ACF) 및 편자기상관 함수(PACF)의 값을 탐색하면 단순히 적합도 통계만 보는 것보다 모델에 대해 더 양적인 통찰을 가질 수 있습니다.

잘 지정된 시계열 모델은 계절성, 추세, 순환 및 기타 중요한 요인을 포함하여 무작위성이 없는 모든 변동을 캡처합니다. 이 케이스에서는 시간 경과에 따라 어떠한 오류도 자신과 상관관계가 있어서는 안됩니다. 즉, 자기상관이 없어야 합니다. 어느 한 자기상관 함수의 유의적 구조는 기본 모델이 불완전함을 암시합니다.

2. 네 번째 시장의 경우, 왼쪽 열에서 **상관 곡선**을 클릭하여 모델에서 잔차 오차에 대한 자기상관 함수(ACF) 및 편자기상관 함수(PACF)의 값을 표시하십시오.

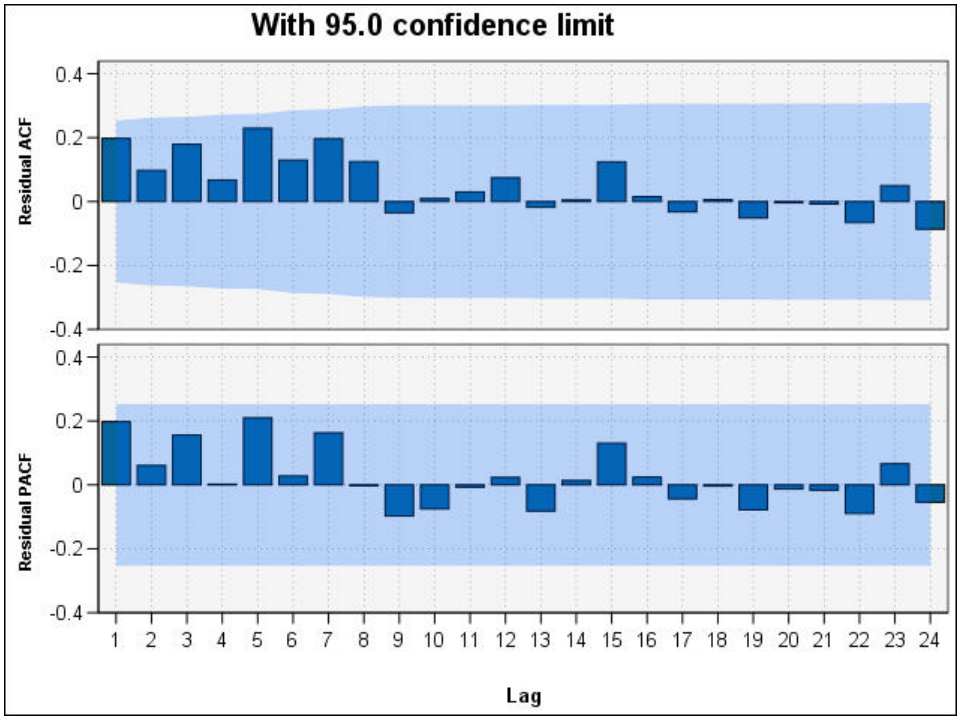


그림 *185.* 네 번째 시장에 대한 *ACF* 및 *PACF* 값

이러한 도표에서 오차 변수의 원래 값은 24시간까지 시차가 발생하고 원래 값과 비교되어 시간 경과에 따른 상관이 있는지 판단합니다. 모델이 허용 가능하려면 위쪽 (ACF) 도표의 어느 막대도 양(+)(위쪽) 또는 음(-)(아래쪽) 방향으로 음영 처리된 영역 밖으로 확장되어서는 안됩니다.

이런 현상이 발생하면 아래쪽의 (PACF) 도표를 검사하여 구조가 확정적인지 확인해야 합니다. PACF 도표는 개입 시점에 계열 값에 대한 제어 후에 상관관계를 찾습니다.

*Market_4*에 대한 값은 모두 음영 처리된 영역 내에 있으므로 기타 시장에 대한 값을 계속 확인할 수 있습니다.

3. 나머지 각 시장 및 총계에 대한 **상관 곡선**을 클릭하십시오.

   기타 시장에 대한 값은 모두 음영 표시된 영역 외부의 값을 표시하며 **유의성** 값에서 이전에 예상한 내용을 확인할 수 있습니다. 이제 몇 가지 포인트에서 해당 시장에 대해 다양한 모델을 실험하여 더 나은 적합도를 얻을 수 있는지 확인할 것입니다. 단, 이 예제의 나머지 부분에서는 *Market_4* 모델에서 배울 수 있는 또 다른 점에 대해 집중할 것입니다.
4. 그래프 팔레트로부터 시간 구성 노드를 시계열 모델 너깃에 첨부하십시오.
5. 도표 탭에서 **별도 패널에 계열 표시** 선택란을 선택 취소하십시오.
6. **계열** 목록에서 필드 선택기 단추를 클릭하고 *Market_4* 및 *$TS-Market_4* 필드를 선택한 다음 **확인**을 클릭하여 이를 목록에 추가하십시오.
7. **실행**을 클릭하여 첫 번째 현지 시장에 대한 실제 및 예측 데이터의 선 그래프를 표시하십시오.

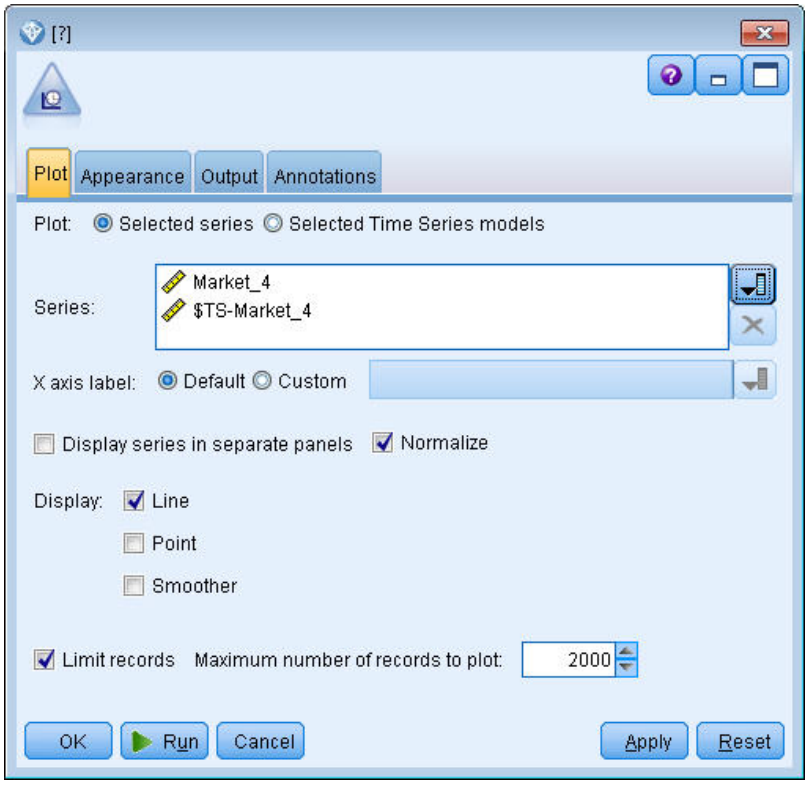


그림 *186.* 도표 작성할 필드 선택

예측(*$TS-Market_4*) 선이 실제 데이터의 끝을 지나서 얼마나 확장되는지 보십시오. 이제 이 시장에서 다음 석 달 동안의 예상 수요에 대한 예측을 갖게 되었습니다.

전체 시계열에 대한 실제 및 예측 데이터의 선이 그래프에서 매우 근접하며 이는 이 특정 시계열에 대해 신뢰할 수 있는 모델임을 나타냅니다.

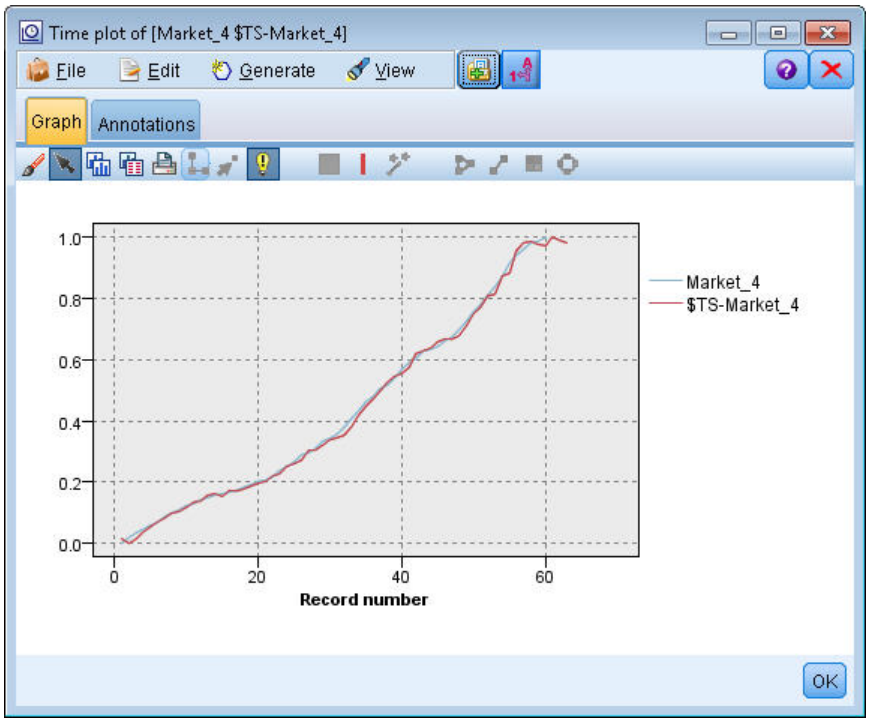


그림 *187. Market_4*에 대한 실제 및 예측 데이터의 시간 도표

이후 예제에서 사용할 수 있도록 모델을 파일에 저장합니다.

8. **확인**을 클릭하여 현재 그래프를 닫으십시오.
9. 시계열 모델 너깃을 여십시오.
10. **파일 > 노드 저장**을 선택하고 파일 위치를 지정하십시오.
11. **저장**을 클릭합니다.

이 특정 시장에 대한 신뢰할 수 있는 모델을 갖게 되었으나 예측에 어떠한 오차 범위가 있습니까? 신뢰구간을 탐색하여 이 지표를 얻을 수 있습니다.

12. 스트림에서 마지막 시간 구성 노드(**Market_4 $TS-Market_4**로 레이블된 노드)를 두 번 클릭하여 대화 상자를 다시 여십시오.
13. 필드 선택기 단추를 클릭하여 *$TSLCI-Market_4* 및 *$TSUCI-Market_4* 필드를 **계열** 목록에 추가하십시오.
14. **실행**을 클릭하십시오.

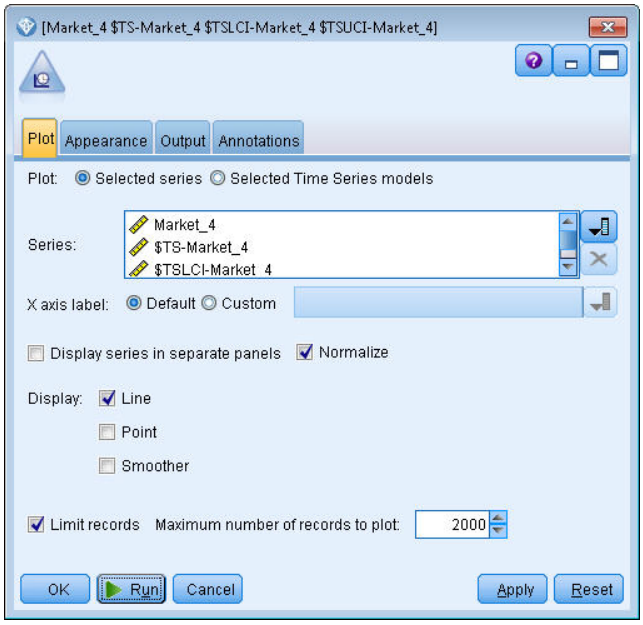


그림 *188.* 도표에 추가 필드 추가

이제 이전과 동일하나 신뢰구간에 대한 상한(*$TSUCI*) 및 하한(*$TSLCI*) 한계가 추가된 그래프를 갖게 되었습니다.

더 먼 미래를 예측함에 따라 불확실성이 증가함이 표시되면서 신뢰구간의 경계가 예측 기간에 대해 어떻게 분기되는지 알 수 있을 것입니다.

그러나 각 시간 주기가 경과함에 따라 예측의 기반이 되는 실제 사용 데이터의 또 다른 개월(이 케이스의 경우) 가치를 갖게 됩니다. 스트림에 새 데이터를 읽어 오고 신뢰할 수 있음을 알고 있는 모델을 다시 적용하십시오. 자세한 정보는 181 페이지의 『시계열 모델 다시 적용』의 내용을 참조하십시오.

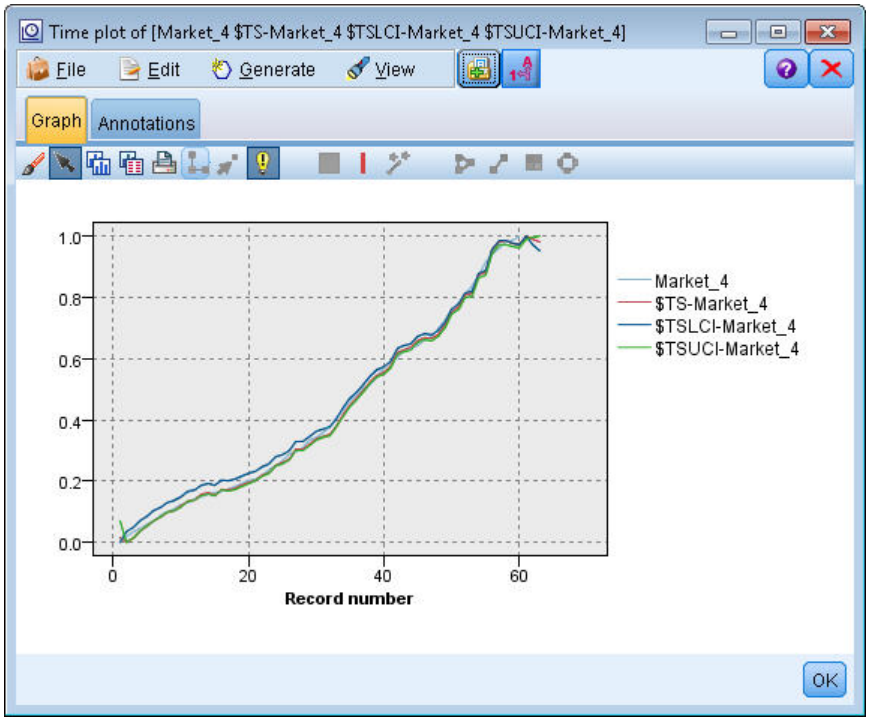


그림 *189.* 신뢰구간이 추가된 시간 도표

요약

자동 모델 생성기를 사용하여 다중 시계열에 대한 예측값을 생성하는 방법을 학습하고 결과 모델을 외부 파일에 저장했습니다.

다음 예에서는 비표준 시계열 데이터를 시계열 노드에 입력하기에 적합한 형식으로 변환하는 방법에 대해 학습할 것입니다.

시계열 모델 다시 적용

이 예는 처음 시계열 예부터 시계열 모델을 적용하나 독립적으로 사용될 수도 있습니다. 자세한 정보는 163 페이지의 『시계열 노드를 사용하여 예측』의 내용을 참조하십시오.

원래 시나리오에서와 같이, 국내 광대역 제공자의 분석가가 대역폭 요구사항을 예측하기 위해 수많은 현지 시장 각각에 대해 사용자 가입에 대한 월별 예측을 생성해야 합니다. 사용자는 이미 자동 모델 생성기를 사용하여 모델을 작성하고 향후 3개월에 대한 예측을 수행했습니다.

데이터 웨어하우스가 원래 예측 기간에 대한 실제 데이터로 업데이트되었으므로 해당 데이터를 사용하여 또 다른 3개월 동안의 예측 범위를 확장하고자 할 수 있습니다.

이 예에서는 *broadband_2.sav*라는 데이터 파일을 참조하는 *broadband_apply_models.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 폴더에 있습니다. Windows 시작 메뉴

의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다.
broadband_apply_models.str 파일은 스트림 폴더에 있습니다.

스트림 검색

이 예에서는 처음 예에서 저장된 시계열 모델에서 시계열 노드를 다시 작성합니다. 모델을 저장하지 않았더라도 걱정할 필요 없습니다. *Demos* 폴더에서 모델이 제공됩니다.

1. *Demos* 아래의 *streams* 폴더에서 *broadband_apply_models.str* 스트림을 여십시오.

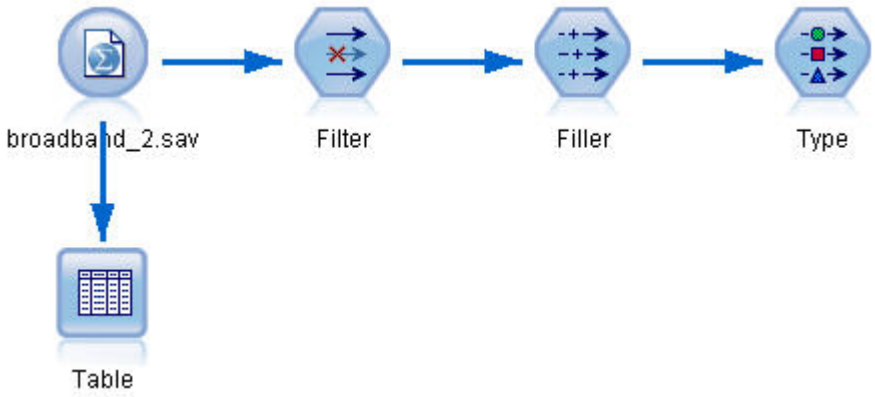
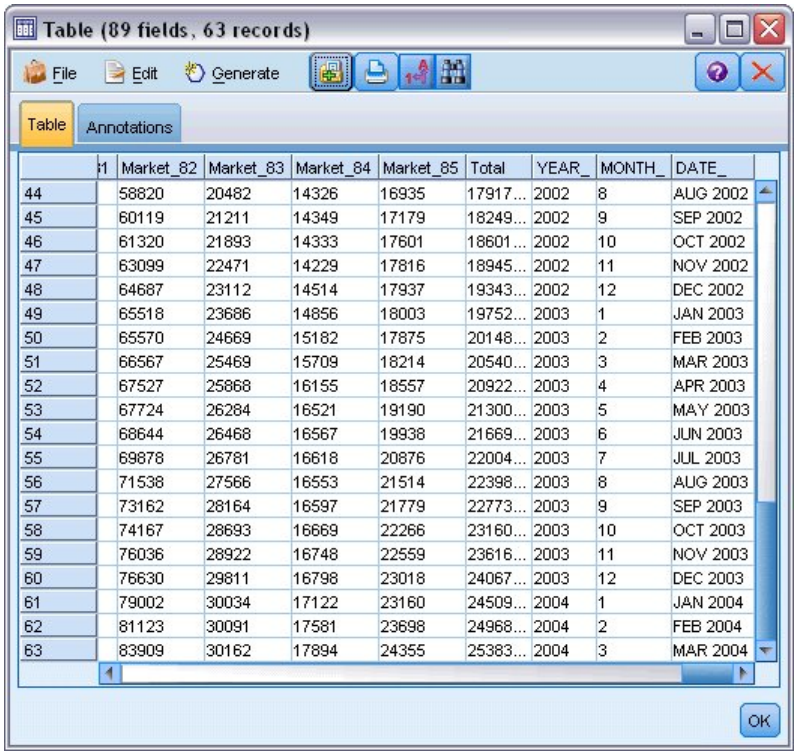


그림 *190.* 스트림 열기

   업데이트된 월별 데이터는 *broadband_2.sav*에 수집됩니다.

2. 테이블 노드를 IBM SPSS Statistics 파일 소스 노드에 첨부하고 테이블 노드를 열고 **실행**을 클릭하십시오.

   참고: 2004년 1월에서 3월까지에 대한 실제 판매 데이터를 사용하여 데이터 파일의 행 61에서 63까지가 업데이트되었습니다.



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

그림 *191.* 업데이트된 판매 데이터

저장된 모델 검색

1. IBM SPSS Modeler 메뉴에서 **삽입** > **파일에서 노드**를 선택하고 *Demos* 폴더에서 *TSmodel.nod* 파일을 선택하십시오. 또는 처음 시계열 예에서 저장된 시계열 모델을 사용하십시오.

이 파일에는 이전 예의 시계열 모델이 포함됩니다. 삽입 작업은 해당되는 시계열 모델 너깃을 캔버스에 배치합니다.

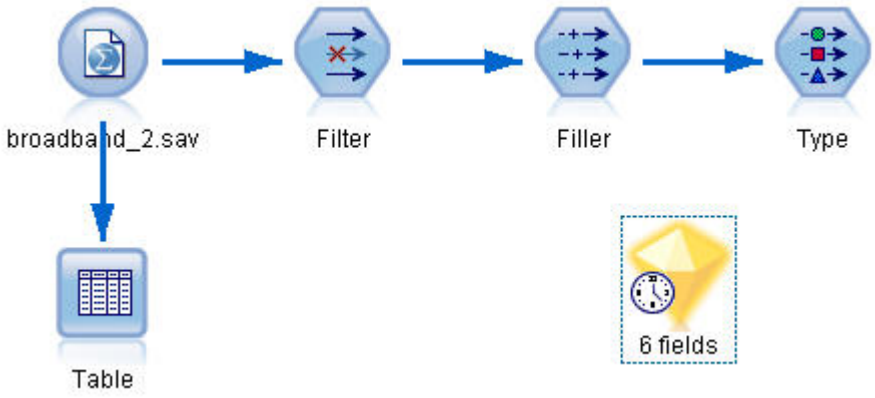


그림 *192.* 모델 너깃 추가

모델링 노드 생성

1. 시계열 모델 너깃을 열고 **생성** > **모델링 노드 생성**을 선택하십시오.

그러면 시계열 모델링 노드가 캔버스에 배치됩니다.

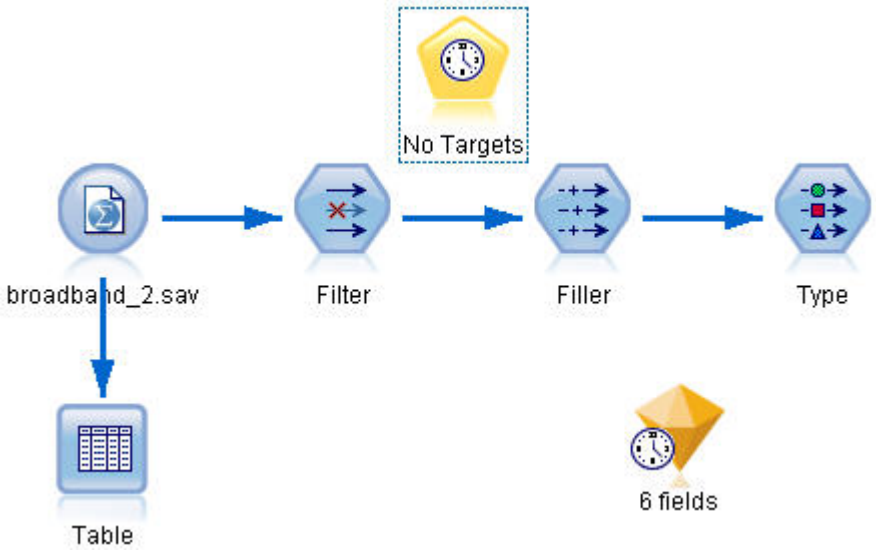


그림 *193.* 모델 너깃에서 모델링 노드 생성

새 모델 생성

1. 시계열 모델 너깃을 닫고 이를 캔버스에서 삭제하십시오.

   이전 모델은 60행의 데이터를 바탕으로 작성되었습니다. 업데이트된 판매 데이터(63행)를 기반으로 하여 새 모델을 생성해야 합니다.
2. 새로 생성된 시계열 작성 노드를 스트림에 첨부하십시오.

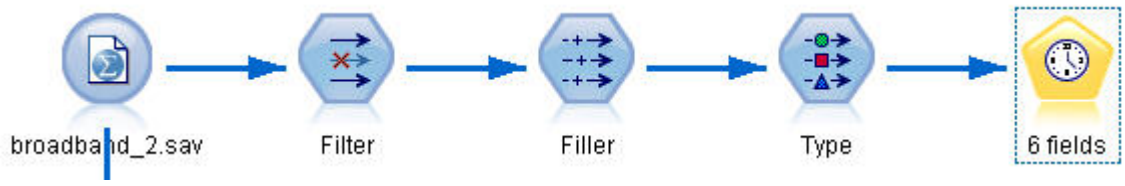


그림 *194.* 스트림에 모델링 노드 첨부

3. 시계열 노드를 여십시오.
4. **모델 옵션** 탭에서 **기존 모델을 사용하여 추정 계속**이 선택되어 있는지 확인하십시오.

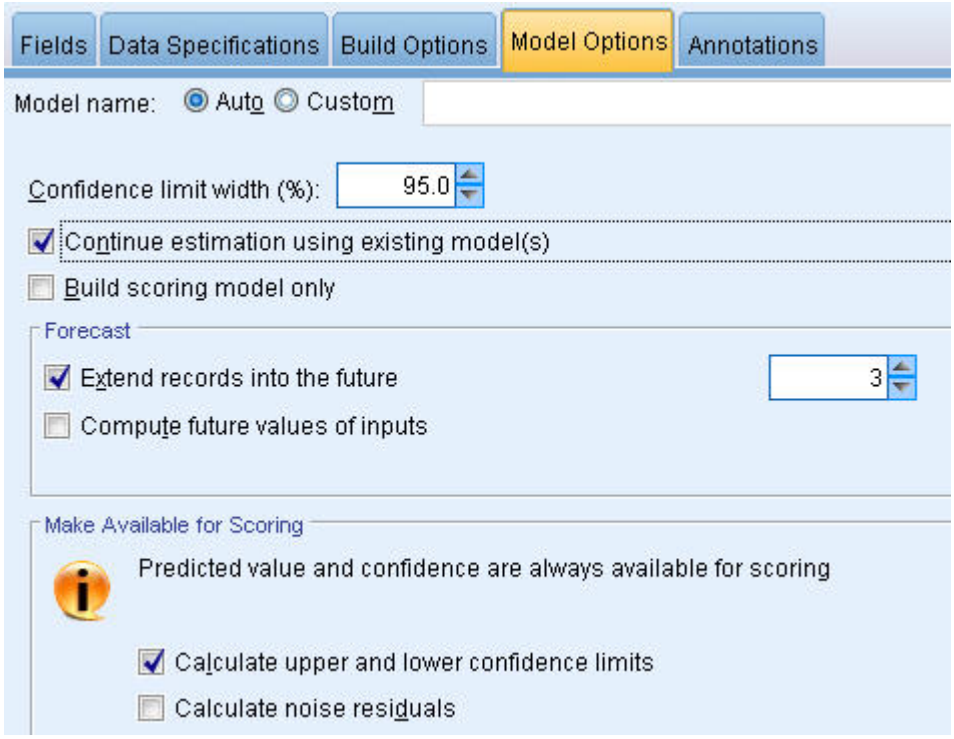


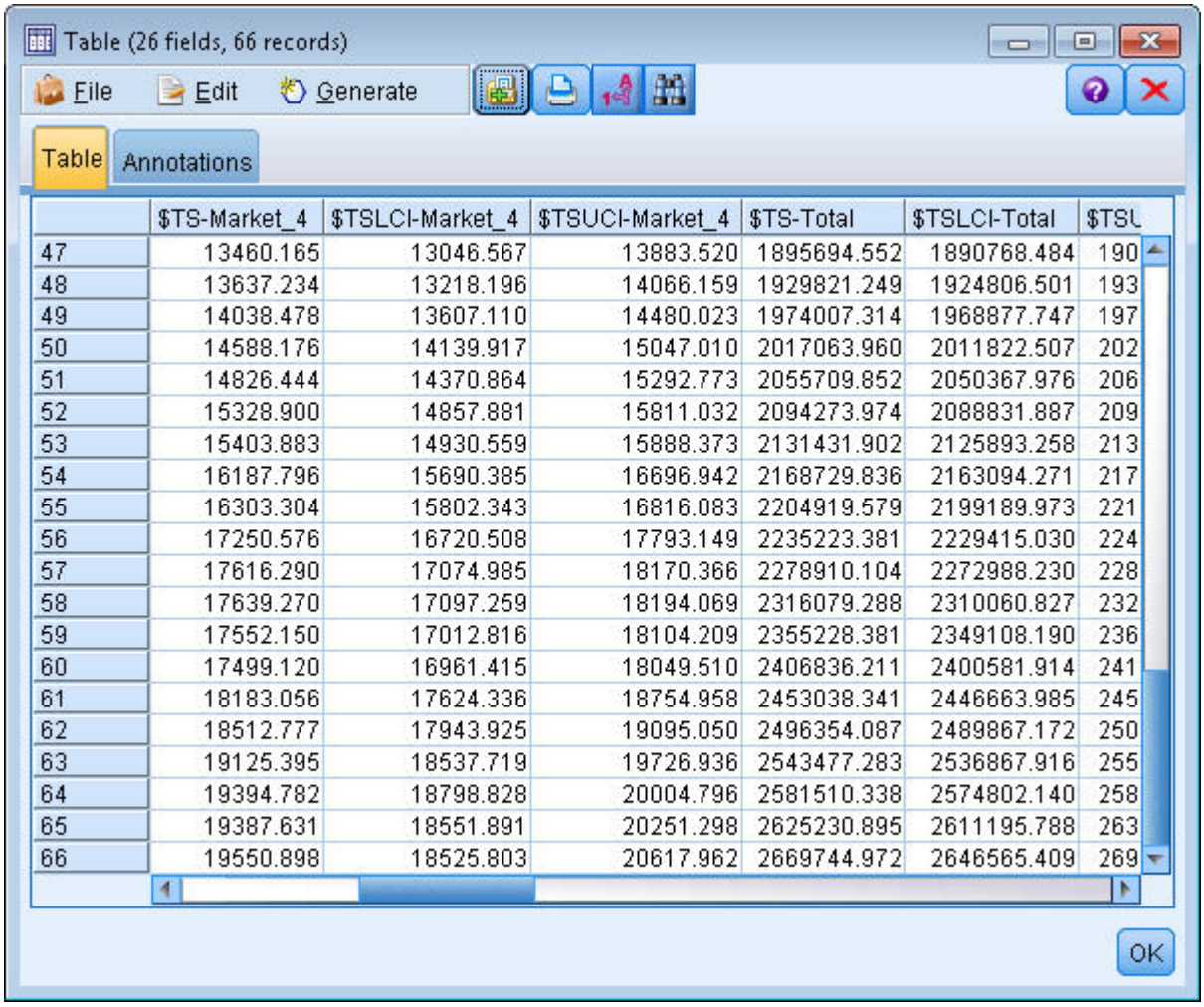
그림 *195.* 시계열 모델에 저장된 설정 재사용

5. **레코드를 미래로 확장**이 **3**으로 설정됩니다.
6. **실행**을 클릭하여 새 모델 너깃을 캔버스 및 모델 팔레트에 배치하십시오.

새 모델 탐색

1. 테이블 노드를 캔버스의 새 시계열 모델 너깃에 첨부하십시오.
2. 테이블 노드를 열고 **실행**을 클릭하십시오.

사용자가 저장된 설정을 재사용하고 있으므로 새 모델은 여전히 삼 개월 앞을 예측합니다. 단, 추정 기간이 1월 대신 3월에 끝나므로 이번에는 4월에서 6월까지(64줄에서 66줄까지) 예측합니다.



	$TS-Market_4	$TSLCI-Market_4	$TSUCI-Market_4	$TS-Total	$TSLCI-Total	$TSL
47	13460.165	13046.567	13883.520	1895694.552	1890768.484	190
48	13637.234	13218.196	14066.159	1929821.249	1924806.501	193
49	14038.478	13607.110	14480.023	1974007.314	1968877.747	197
50	14588.176	14139.917	15047.010	2017063.960	2011822.507	202
51	14826.444	14370.864	15292.773	2055709.852	2050367.976	206
52	15328.900	14857.881	15811.032	2094273.974	2088831.887	209
53	15403.883	14930.559	15888.373	2131431.902	2125893.258	213
54	16187.796	15690.385	16696.942	2168729.836	2163094.271	217
55	16303.304	15802.343	16816.083	2204919.579	2199189.973	221
56	17250.576	16720.508	17793.149	2235223.381	2229415.030	224
57	17616.290	17074.985	18170.366	2278910.104	2272988.230	228
58	17639.270	17097.259	18194.069	2316079.288	2310060.827	232
59	17552.150	17012.816	18104.209	2355228.381	2349108.190	236
60	17499.120	16961.415	18049.510	2406836.211	2400581.914	241
61	18183.056	17624.336	18754.958	2453038.341	2446663.985	245
62	18512.777	17943.925	19095.050	2496354.087	2489867.172	250
63	19125.395	18537.719	19726.936	2543477.283	2536867.916	255
64	19394.782	18798.828	20004.796	2581510.338	2574802.140	258
65	19387.631	18551.891	20251.298	2625230.895	2611195.788	263
66	19550.898	18525.803	20617.962	2669744.972	2646565.409	269

그림 *196.* 새 예측을 표시하는 테이블

3. 시간 도표 그래프 노드를 시계열 모델 너깃에 추가하십시오.

   이번에는 시계열 모델을 위해 특별히 계획된 시간 도표 계획을 사용할 것입니다.

4. 도표 탭에서 **X 축 레이블**을 **사용자 정의**로 설정하고 Date_를 선택하십시오.
5. **도표** 탭에서 **선택된 시계열 모델** 옵션을 선택하십시오.
6. **계열** 목록에서 필드 선택기 단추를 클릭하고 $TS-Market_4 필드를 선택하고 **확인**을 클릭하여 이를 목록에 추가하십시오.

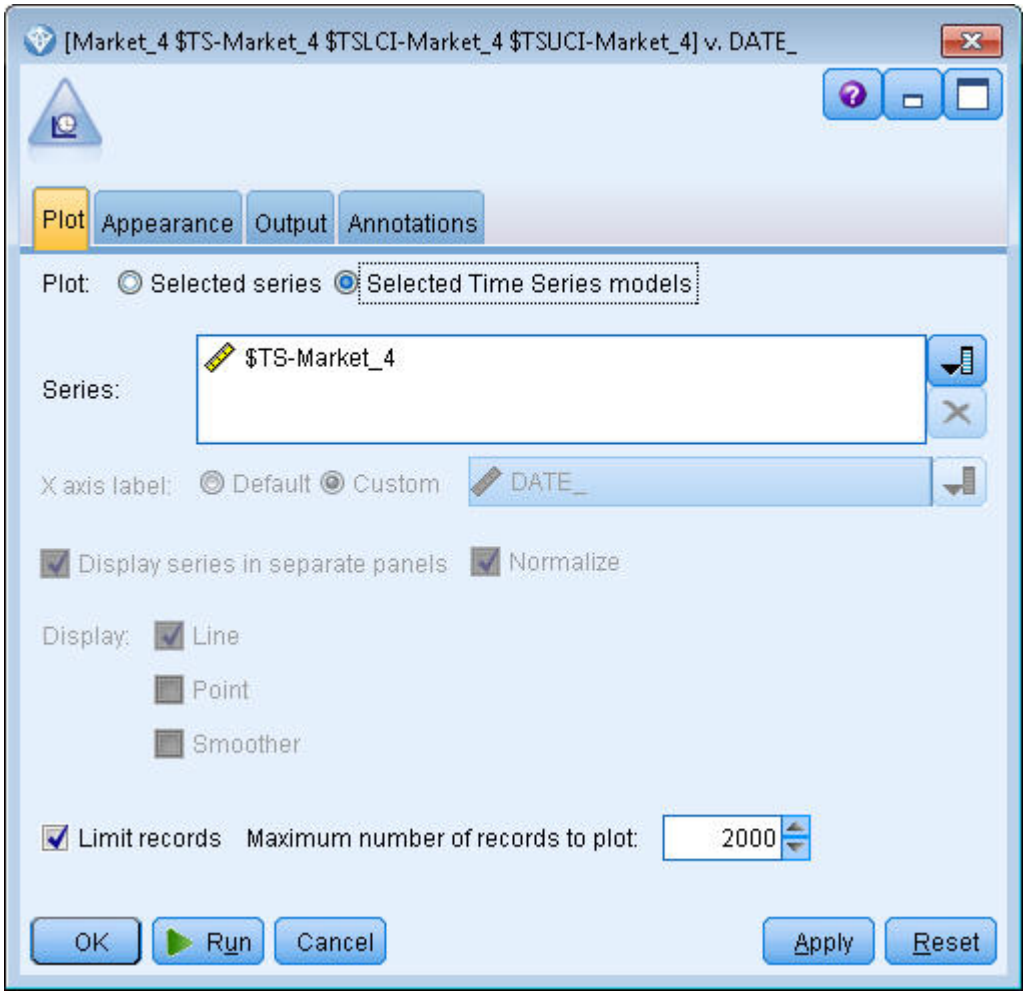


그림 *197.* 도표 작성할 필드 지정

7. **실행**을 클릭하십시오.

이제 2004년 6월까지의 예측(예측값) 판매 및 신뢰구간(파란색 음영 처리된 영역)과 함께 2004년 3월까지의 Market_4에 대한 실제 판매를 표시하는 그래프를 갖게 됩니다.

처음 예에서와 같이, 전체 기간을 통해 예측 값이 실제 데이터에 근접하게 따라가므로 다시 한 번 좋은 모델임을 알 수 있습니다.

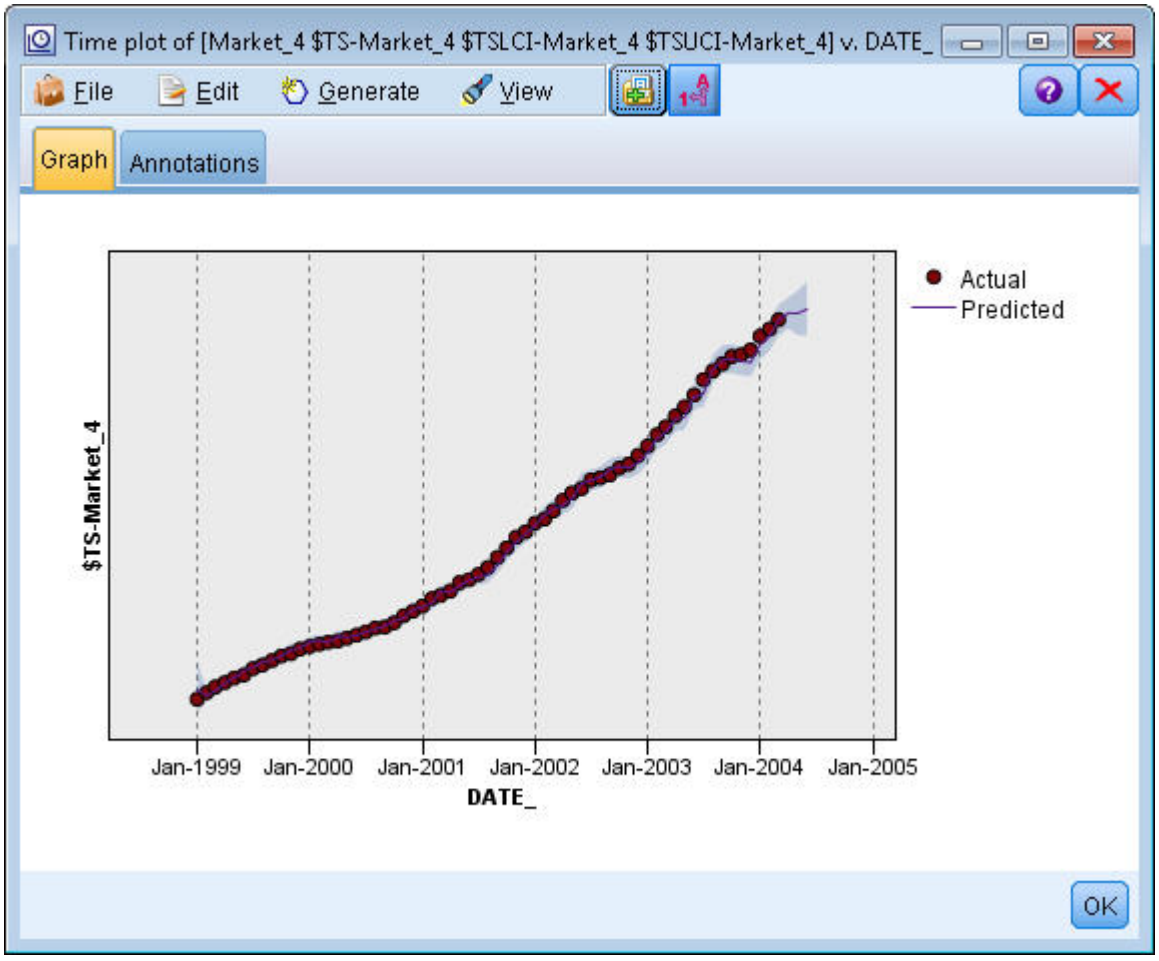


그림 *198. 6*월로 확장된 예측

요약

더 많은 현재 데이터를 사용할 수 있게 될 때 이전 예측을 확장하기 위해 저장된 모델을 적용하는 방법을 학습하고 모델을 다시 작성하지 않고 이 작업을 수행했습니다. 물론 모델이 변경되었다고 생각할 이유가 있으면 모델을 다시 작성해야 합니다.

제 15 장 카탈로그 판매 예측(시계열)

한 카탈로그 회사가 지난 10년 간의 판매 데이터를 기반으로 하여 남성 의류 라인의 월별 판매를 예측하는 데 관심이 있습니다.

이 예에서는 *catalog_seasfac.sav*라는 데이터 파일을 참조하는 *catalog_forecast.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *catalog_forecast.str* 파일은 *streams* 디렉토리에 있습니다.

이전 예에서 자동 모델 생성기가 사용자의 시계열에 가장 적합한 모델을 판단하는 방법을 보았습니다. 이제 모델을 스스로 선택할 때 사용 가능한 두 가지 방법(지수평활 및 ARIMA)을 더 자세히 살펴볼 때입니다.

적절한 모델을 결정할 때 도움을 받기 위해 먼저 시계열을 작성하는 것이 좋습니다. 시계열에 대한 시각적 조사는 종종 선택에 도움이 되는 강력한 안내서가 될 수 있습니다. 특히, 다음 사항을 자문해야 합니다.

- 계열에 전체 추세가 있습니까? 있으면 추세가 상수로 표시됩니까? 또는 시간 경과에 따라 사라지는 것으로 표시됩니까?
- 계열에 계절성이 표시됩니까? 표시되는 경우, 계절적 변동이 시간 경과와 함께 증가합니까? 또는 연속된 기간 도안 상수로 표시됩니까?

스트림 작성

1. 새 스트림을 작성하고 *catalog_seasfac.sav*를 가리키는 통계 파일 소스 노드를 추가합니다.

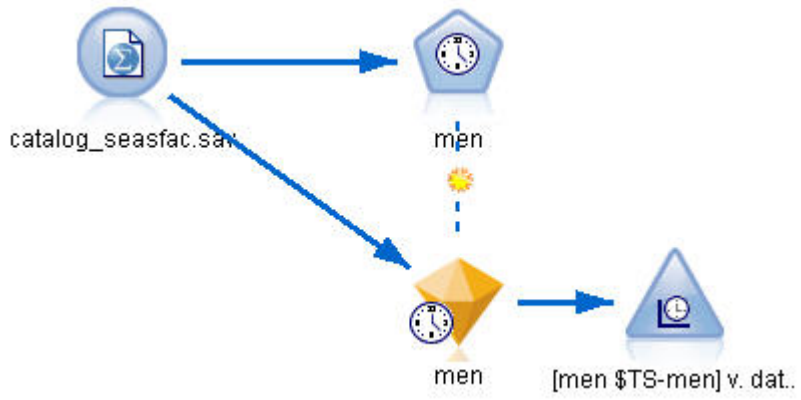


그림 *199.* 카탈로그 판매 시계열 분석

2. IBM SPSS Statistics 파일 소스 노드를 열고 유형 탭을 선택하십시오.
3. **값 읽기**를 클릭한 다음 **확인**을 클릭하십시오.
4. **men** 필드에 대한 역할 열을 클릭하고 역할을 **대상**으로 설정하십시오.

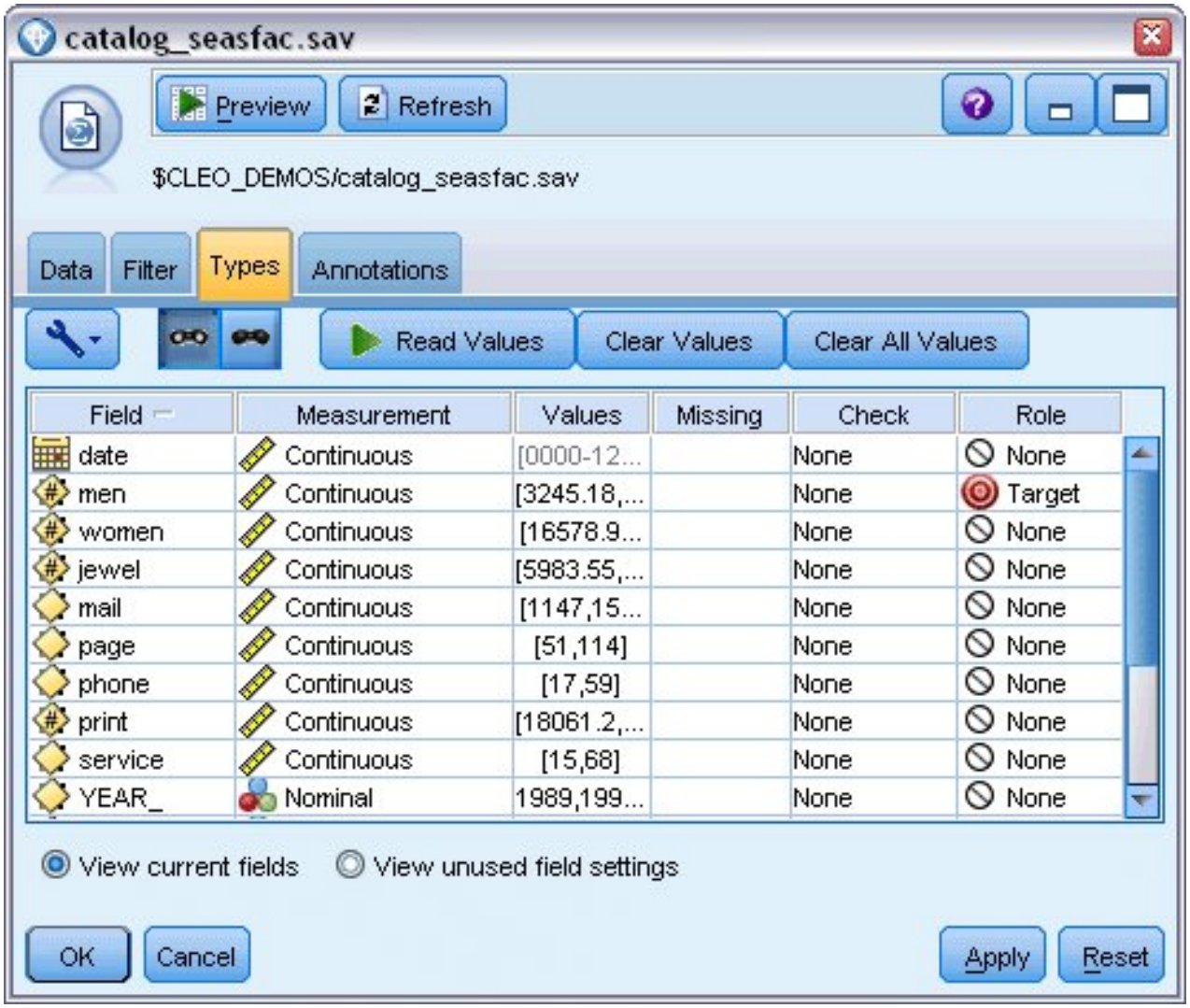


그림 *200.* 대상 필드 지정

5. 모든 기타 필드에 대한 역할을 **없음**으로 설정하고 **확인**을 클릭하십시오.
6. 시간 도표 그래프 노드를 IBM SPSS Statistics 파일 소스 노드에 추가하십시오.
7. 시간 구성 노드를 열고 도표 탭에서 men을 **계열** 목록에 추가하십시오.
8. **X 축 레이블**을 **사용자 정의**로 설정하고 date를 선택하십시오.
9. **표준화** 선택란을 선택 취소하십시오.

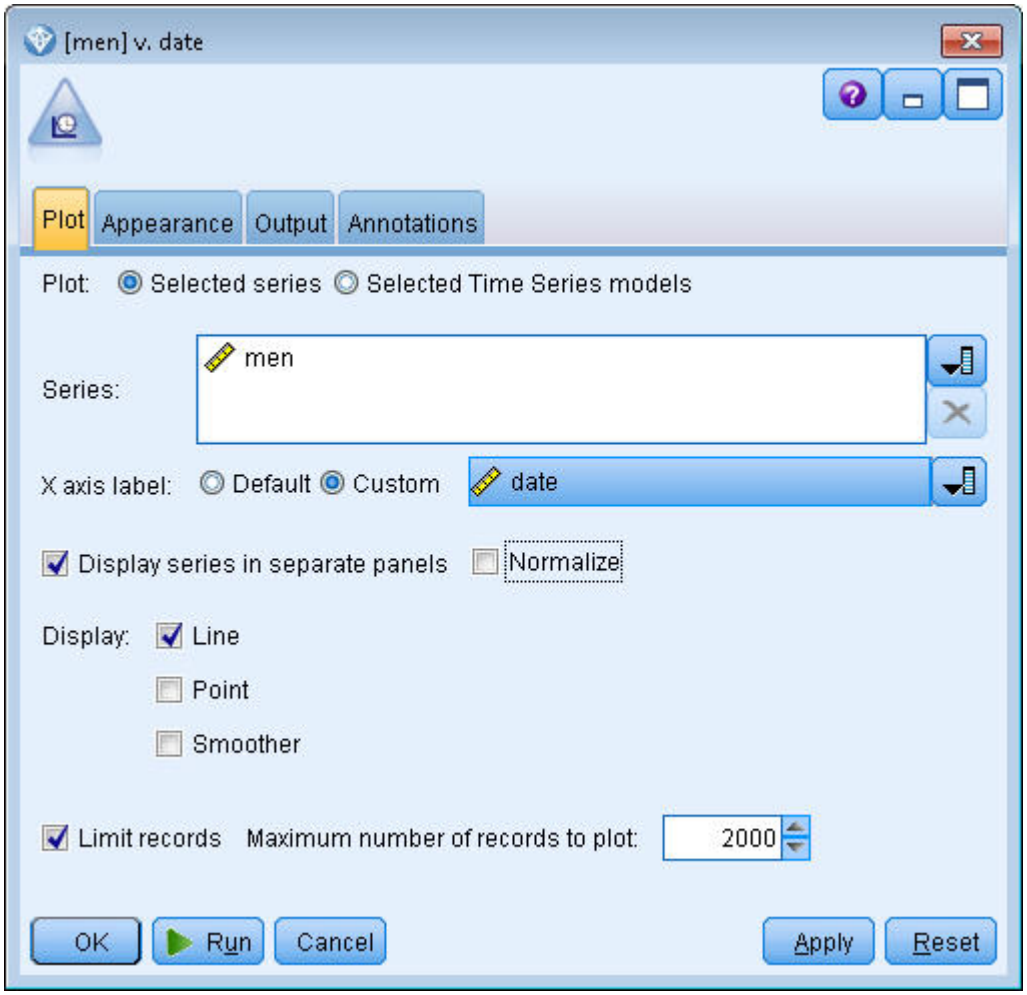


그림 *201.* 시계열 도표 작성

10. **실행**을 클릭하십시오.

데이터 탐색

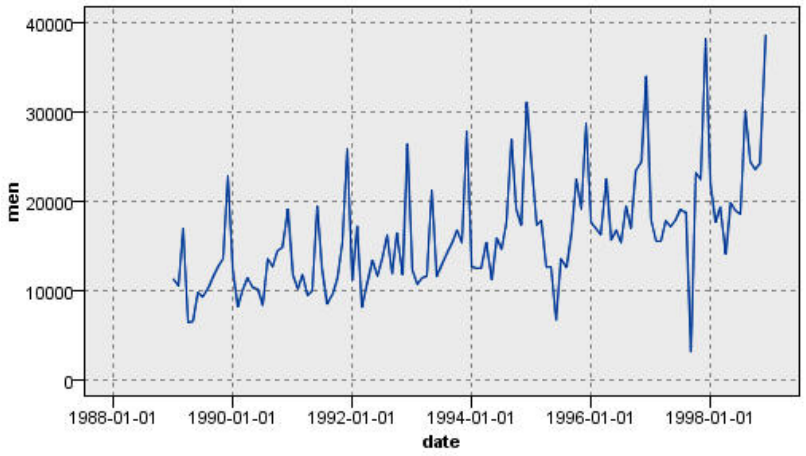


그림 *202.* 남성 의류 실제 판매

계열은 일반 상향 추세를 표시합니다. 즉, 계열 값은 시간이 지나면서 증가하는 추세가 있습니다. 상향 추세는 일정한 것으로 보이며 선형 추세를 나타냅니다.

계열 또한 그래프의 세로선에 의해 표시된 대로 매년 12월에 높아지는 뚜렷한 계절 패턴을 보입니다. 계절 변동은 상향 계열 추세와 함께 증가하는 것으로 보이며 이는 가법 계절성보다 승법 계절성을 제안합니다.

1. **확인**을 클릭하여 도표를 닫으십시오.

이제 계열의 공정특성 변수를 식별했으므로 모델링을 시도할 준비가 되었습니다. 지수평활 방법은 추세, 계절성 또는 모두를 나타내는 계열의 시계열 분석에 유용합니다. 앞에서 본 대로 데이터가 모두 공정특성 변수를 나타냅니다.

지수평활

최적 맞춤 지수평활 모델 작성은 모델 유형(모델에 추세, 계절성 또는 둘 다를 포함시켜야 하는지 여부)을 판별하고 선택된 모델에 대한 최적 맞춤 모수를 얻는 것과 연관됩니다.

시간 경과에 따른 남성 의류 판매 도표에서는 선형 추세 성분 및 승법 계절성 성분이 둘 다 있는 모델을 제안합니다. 이는 Winters 모델을 암시합니다. 그러나 먼저 추세 및 계절성이 없는 단순 모델을 탐색한 다음 Holt 모델(선형 추세는 통합하나 계절성은 통합하지 않음)을 탐색할 것입니다. 이는 모델이 데이터에 적합하지 않음을 식별하는 방법 및 성공적인 모델 작성의 필수 기술에 대해 연습할 수 있게 해줍니다.

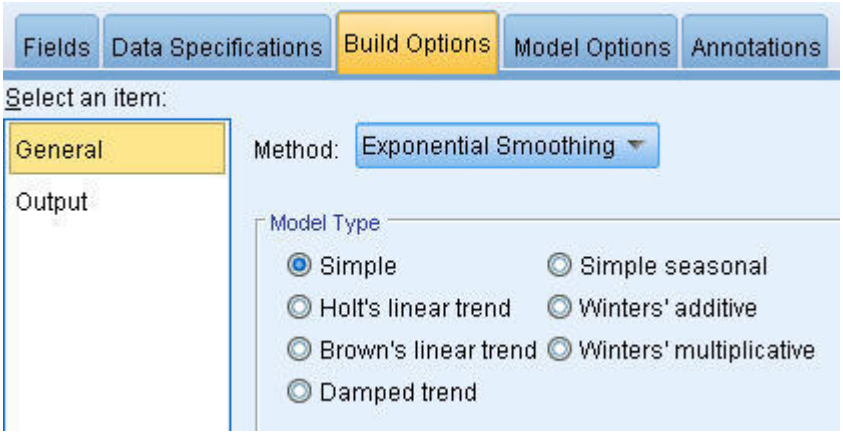


그림 *203.* 지수평활 지정

단순 지수평활 모델을 시작할 것입니다.

1. 시계열 노드를 스트림에 추가하고 이를 소스 노드에 첨부하십시오.
2. 데이터 지정 사항 탭의 관측값 분할창에서 **날짜/시간 필드**로 날짜를 선택하십시오.
3. **시간 구간**으로 월을 선택하십시오.

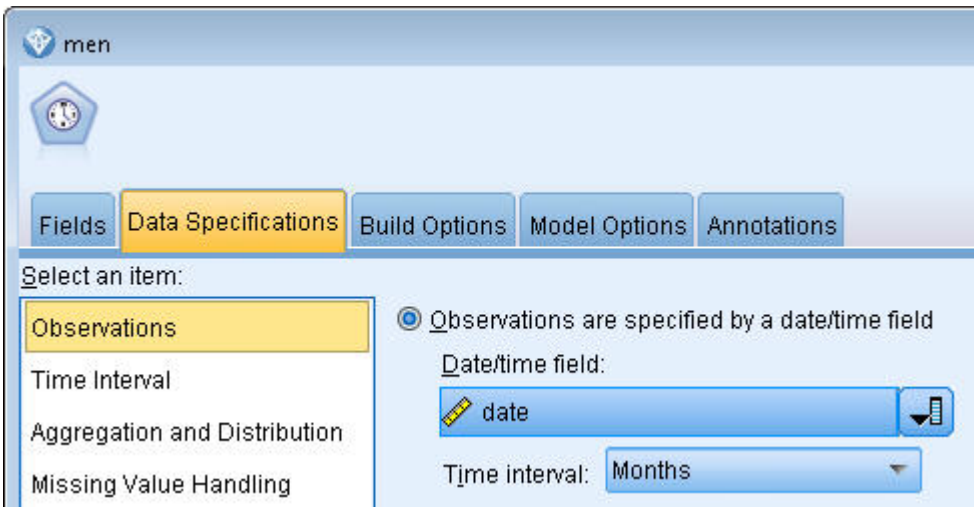


그림 *204.* 시간 구간 설정

4. 작성 옵션 탭의 일반 분할창에서 **방법**을 **지수평활**로 설정하십시오.
5. **모델 유형**을 **단순**으로 설정하십시오.

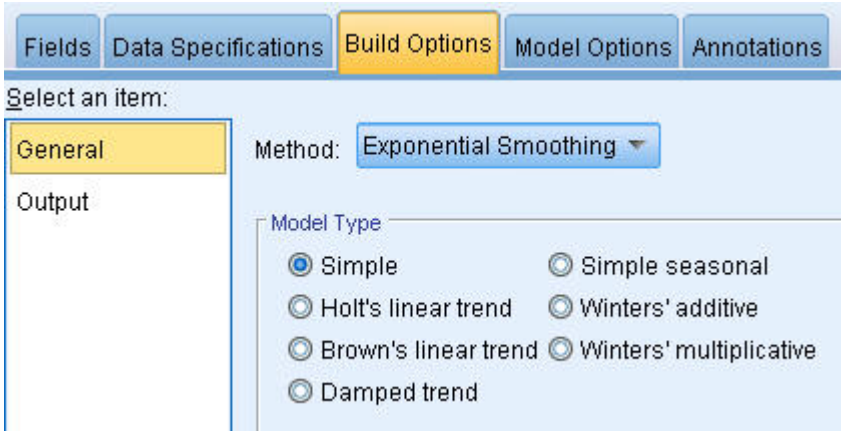


그림 *205.* 모델 작성 방법 설정

6. **실행**을 클릭하여 모델 너깃을 작성하십시오.

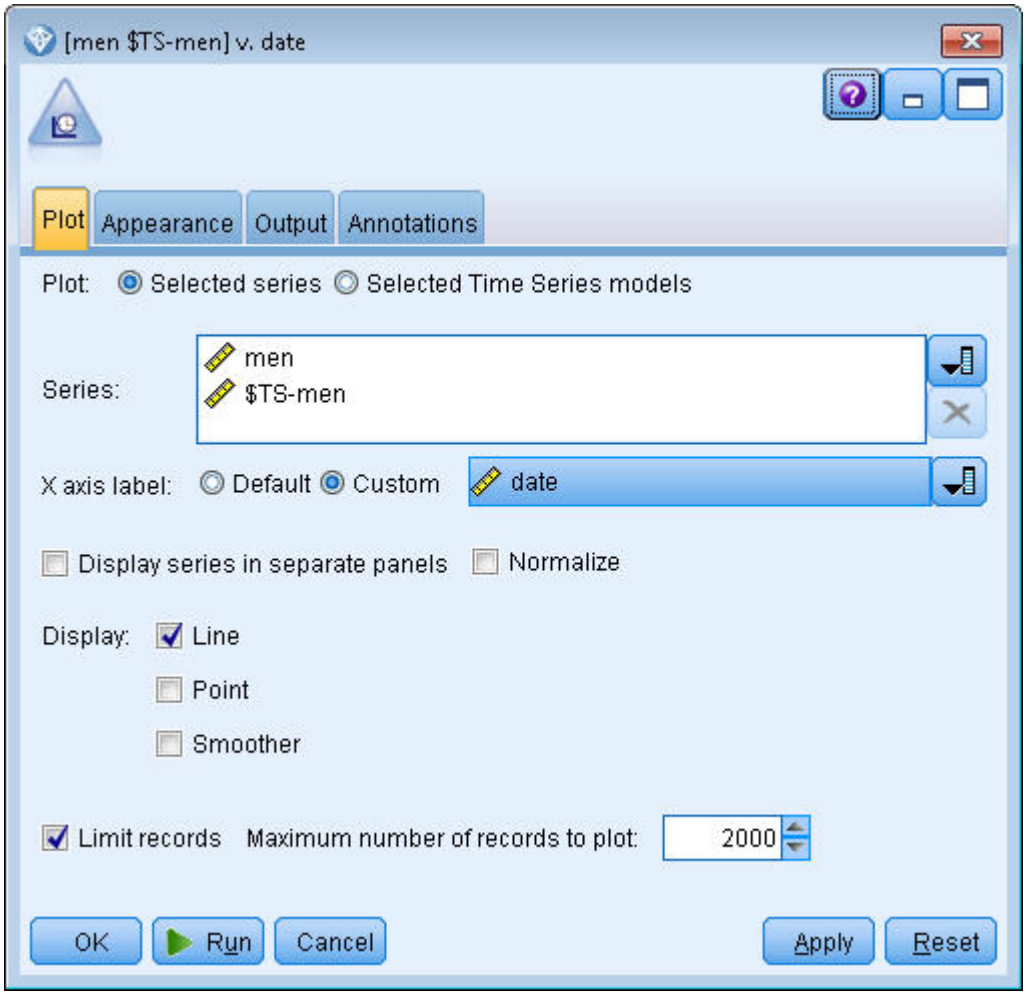


그림 *206.* 시계열 모델 도표 작성

7. 시간 구성 노드를 모델 너깃에 연결하십시오.

8. **도표** 탭에서 men 및 $TS-men을 **계열** 목록에 추가하십시오.
9. **X 축 레이블**을 **사용자 정의**로 설정하고 date를 선택하십시오.
10. **별도 패널에 계열 표시** 및 **표준화** 선택란을 선택 취소하십시오.
11. **실행**을 클릭하십시오.

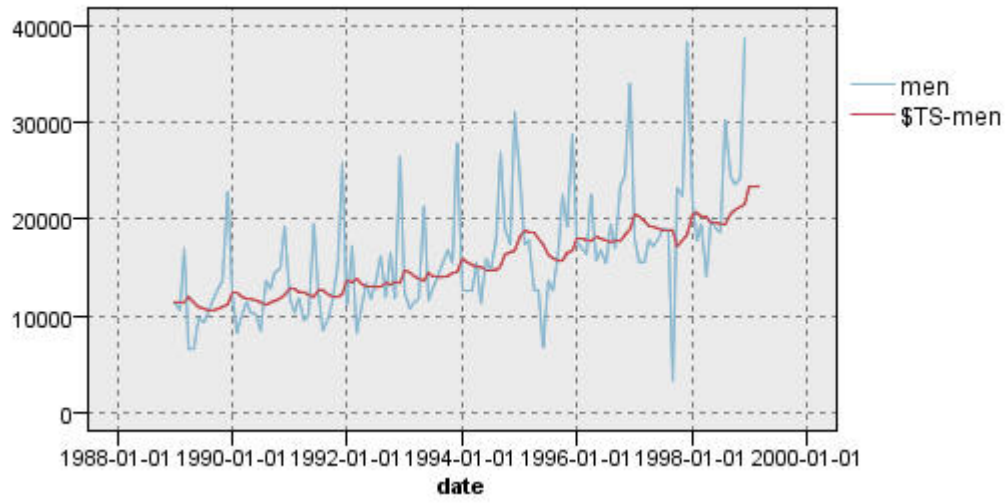


그림 *207.* 단순 지수평활 모델

men 도표는 실제 데이터를 표시하는 반면 **$TS-men**은 시계열 모델을 나타냅니다.

단순 모델은 실제로 무겁다고도 할 수 있는 점진적인 상향 추세를 나타내나 계절성은 고려하지 않습니다. 이 모델을 거부하는 것이 안전합니다.

12. **확인**을 클릭하여 시간 도표 창을 닫으십시오.

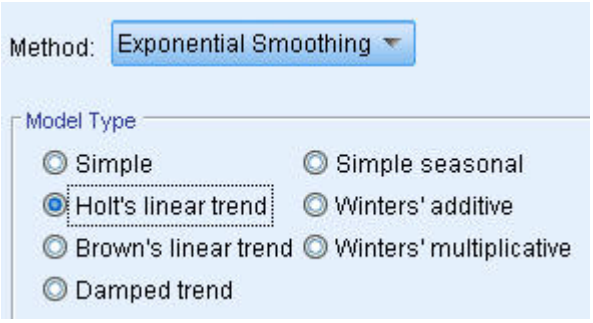


그림 *208. Holt* 모델 선택

Holt의 선형 모델을 시도합니다. 이는 계절성을 캡처할 가능성이 없더라도 단순 모델보다는 추세가 더 적합한 모델이어야 합니다.

13. 시계열 노드를 다시 여십시오.
14. **방법**으로 여전히 **지수평활**이 선택된 작성 옵션 탭의 일반 분할창에서 **모델 유형**으로 **Holt 선형 추세**를 선택하십시오.

15. **실행**을 클릭하여 모델 너깃을 다시 작성하십시오.
16. 시간 구성 노드를 다시 열고 **실행**을 클릭하십시오.

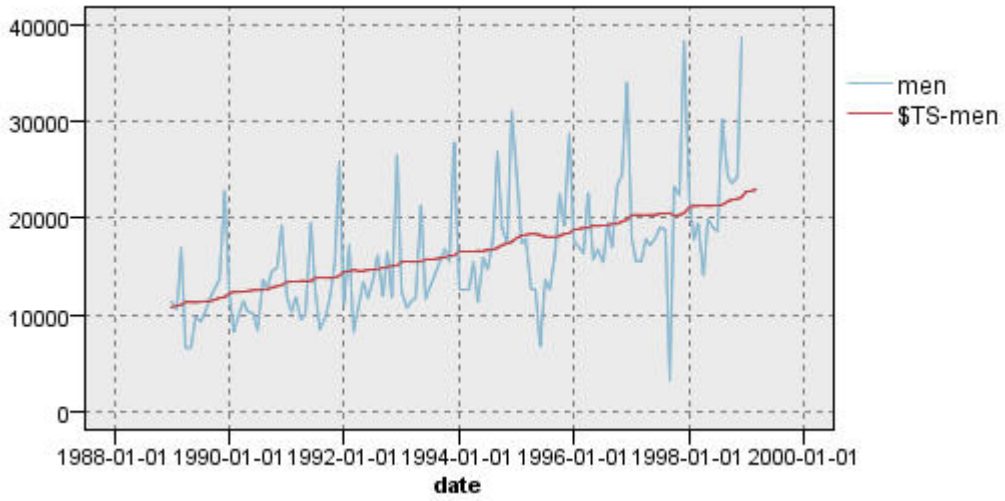


그림 *209. Holt* 선형 추세 모델

Holt 모델은 단순 모델보다 평활한 상향 추세를 표시하나 여전히 계절성을 고려하지 않으므로 이 모델도 삭제할 수 있습니다.

17. 시간 도표 창을 닫으십시오.

초기의 시간 경과에 따른 남성 의류 판매 도표에서 선형 추세 및 승법 계절성을 통합한 모델을 제안했음을 기억할 것입니다. 따라서 더 적합한 후보는 윈터스 모델일 수 있습니다.

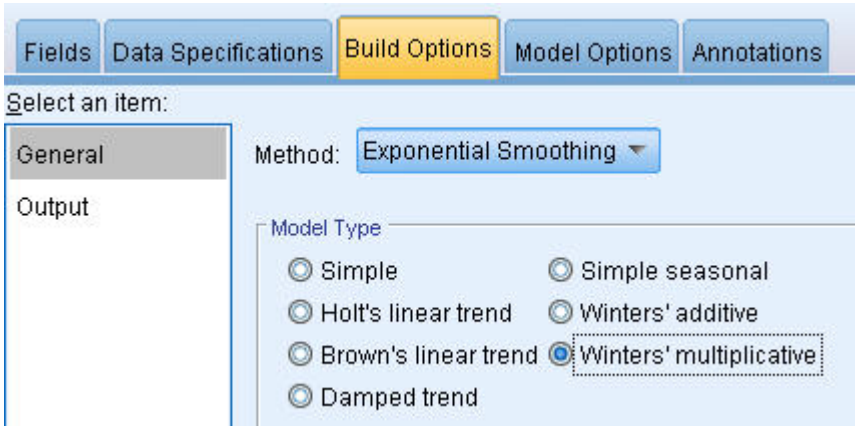


그림 *210.* 윈터스 모델 선택

18. 시계열 노드를 다시 여십시오.
19. **방법**으로 여전히 **지수평활**이 선택된 작성 옵션 탭의 일반 분할창에서 **모델 유형**으로 **Winters의 승법**을 선택하십시오.
20. **실행**을 클릭하여 모델 너깃을 다시 작성하십시오.

21. 시간 구성 노드를 열고 **실행**을 클릭하십시오.

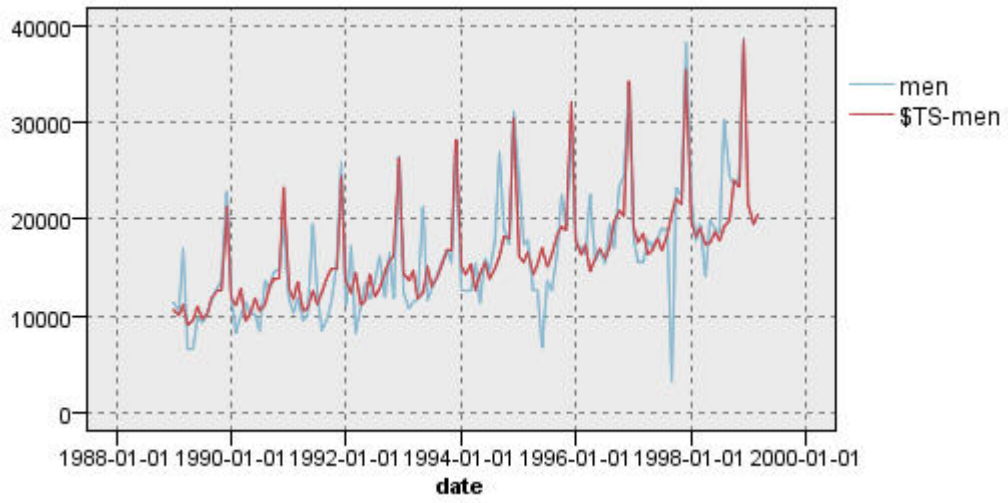


그림 *211. Winters*의 승법 모델

이 모델이 더 나아 보입니다. 즉, 모델이 데이터의 추세 및 계절성을 둘 다 반영합니다.

데이터 세트는 10년 간의 기간을 포함하고 각 연도의 12월에 발생하는 계절 최대를 포함합니다. 예측 결과에 있는 10개의 최대는 실제 데이터의 10개의 연간 최대와 잘 일치됩니다.

그러나 이 결과는 지수평활 프로시저의 제한사항을 강조하기도 합니다. 상향 및 하향 말뚝표시를 둘 다 보면 고려하지 않은 유의적 구조가 있습니다.

기본적으로 계절 변동이 있는 장기 추세를 모델링하는 데 관심이 있으면 지수평활이 좋은 선택이 될 수 있습니다. 이와 같은 더 복잡한 구조를 모델링하려면 ARIMA 프로시저를 사용하는 것을 고려해야 합니다.

ARIMA

ARIMA 프로시저를 사용하면 시계열의 정교하게 조정된 모델링에 적합한 자기회귀 통합 이동 평균(ARIMA) 모델을 작성할 수 있습니다. ARIMA 모델은 지수평활 모델을 수행하는 모델링 추세 및 계절 성분에 대해 보다 정교한 방법을 제공하며 모델의 예측변수를 포함할 수 있는 혜택을 추가했습니다.

예측 모델을 개발하고자 하는 카탈로그 회사의 예를 계속 사용하여 몇 가지 판매 변동을 설명하는 데 사용될 수 있는 여러 계열과 함께 회사에서 남성 의류의 월별 할인에 대한 데이터를 수집하는 방법에 대해 알아보았습니다. 가능한 예측변수에는 우편으로 발송된 카탈로그의 수, 카탈로그의 페이지 수, 주문할 수 있는 개통된 전화 회선 수, 인쇄 광고에 사용된 금액 및 고객 서비스 담당자의 수 등이 있습니다.

이러한 예측변수 중 어느 것이 예측에 유용합니까? 예측변수가 있는 모델이 예측변수가 없는 모델보다 실제로 더 낫습니까? ARIMA 프로시저를 사용하면 예측변수가 있는 시계열 분석 모델을 작성할 수 있으며 예측변수가 없는 지수평활 모델에 대해 예측 기능에 유의적 차이가 있는지 알 수 있습니다.

ARIMA 방법을 사용하면 자기회귀의 순서, 차분 및 이동 평균 및 이러한 성분에 대한 계절 성분을 지정하여 모델을 정교하게 조정할 수 있습니다. 이러한 성분에 대한 최선의 값을 수동으로 결정하는 것은 수많은 시행착오를 겪어야 하는 시간 소모적인 프로세스가 될 수 있으므로 이 예에서는 자동 모델 생성기가 대신 ARIMA 모델을 선택합니다.

데이터 세트에서 기타 변수 중 일부를 예측변수 변수로 처리하여 더 나은 모델을 작성하려고 시도할 수 있습니다. 예측변수로 포함하기에 가장 유용한 변수는 우편으로 발송된 카탈로그의 수(mail), 카탈로그의 페이지(page), 주문할 수 있는 개통된 전화 회선 수(phone), 인쇄 광고에 사용된 금액(print) 및 고객 서비스 담당자의 수(service)입니다.

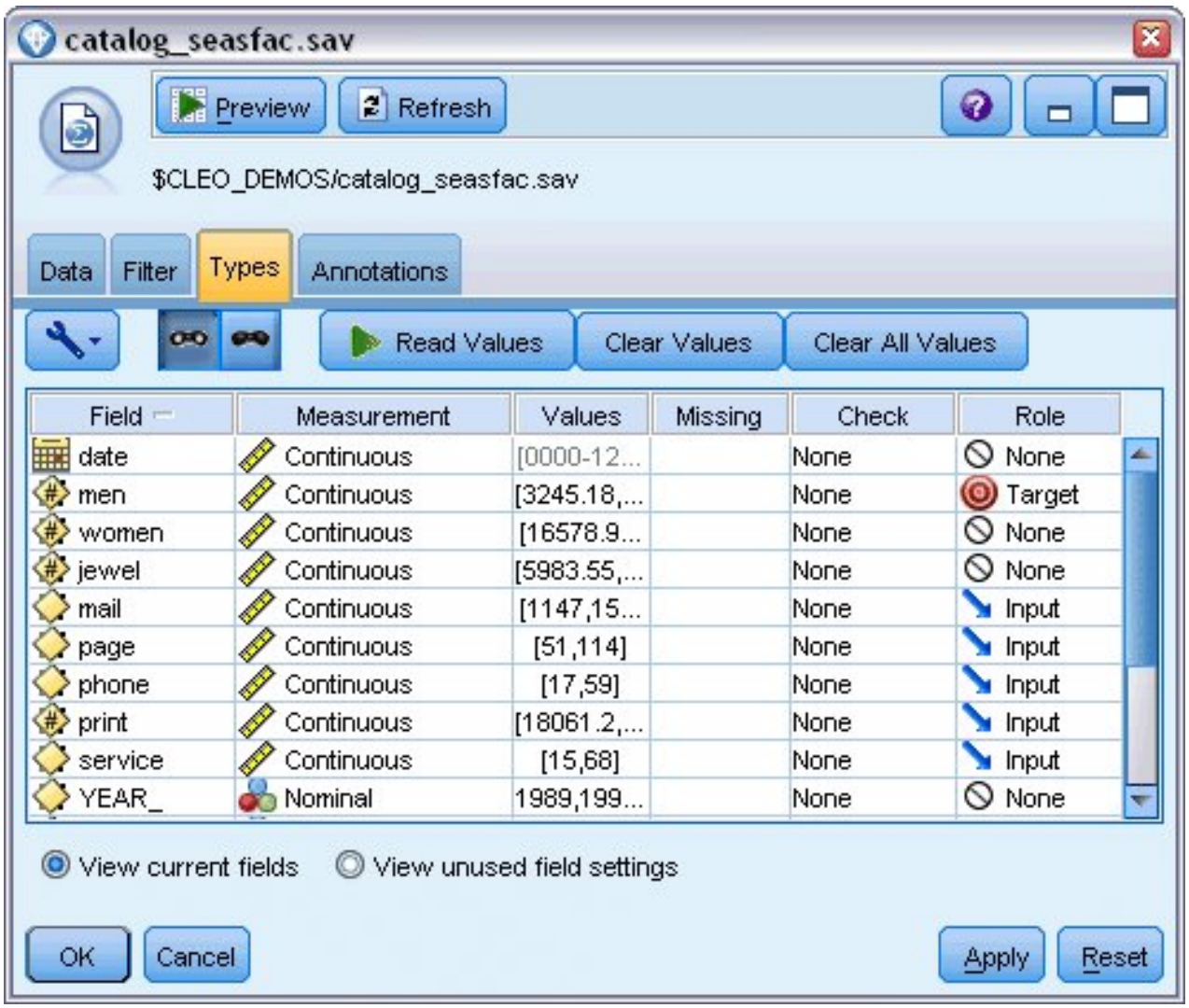


그림 *212.* 예측변수 필드 설정

1. IBM SPSS Statistics 파일 소스 노드를 여십시오.
2. 탭에서 **mail**, page, phone, print 및 service에 대한 역할을 **입력**으로 설정하십시오.
3. men에 대한 역할이 **대상**으로 설정되고 모든 나머지 필드가 **없음**으로 설정되었는지 확인하십시오.
4. **확인**을 클릭하십시오.
5. 시계열 노드를 여십시오.
6. 작성 옵션 탭의 일반 분할창에서 **방법**을 **자동 모델 생성기**로 설정하십시오.
7. **ARIMA 모델만** 옵션을 선택하고 **자동 모델 생성기에서 계절 모델 고려**가 선택되어 있는지 확인하십시오.

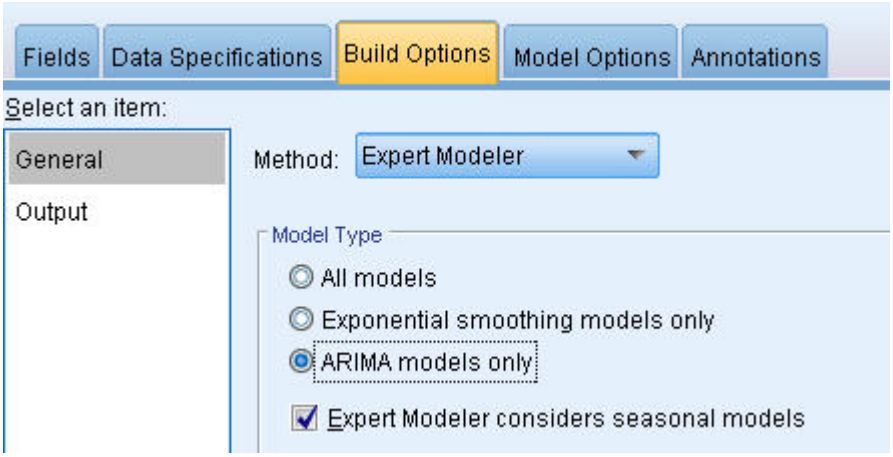


그림 *213. ARIMA* 모델만 선택

8. **실행**을 클릭하여 모델 너깃을 다시 작성하십시오.
9. 모델 너깃을 여십시오.

   출력 탭의 왼쪽 열에서 **모델 정보**를 선택하십시오. 자동 모델 생성기가 다섯 개의 지정된 예측변수 중 두 개만 모델에 유의적인 것으로 선택한 방법에 주목하십시오.

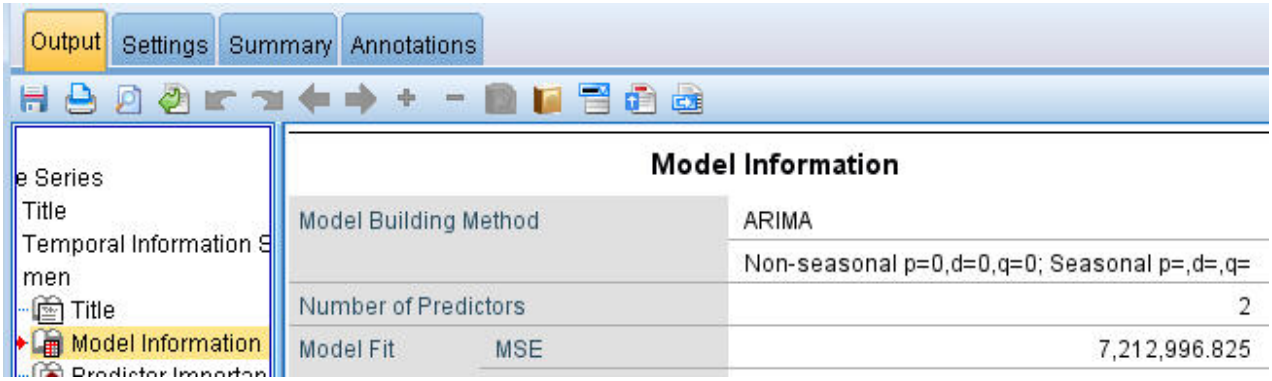


그림 *214.* 자동 모델 생성기에서 두 개의 예측변수 선택

10. **확인**을 클릭하여 모델 너깃을 닫으십시오.

11. 시간 구성 노드를 열고 **실행**을 클릭하십시오.

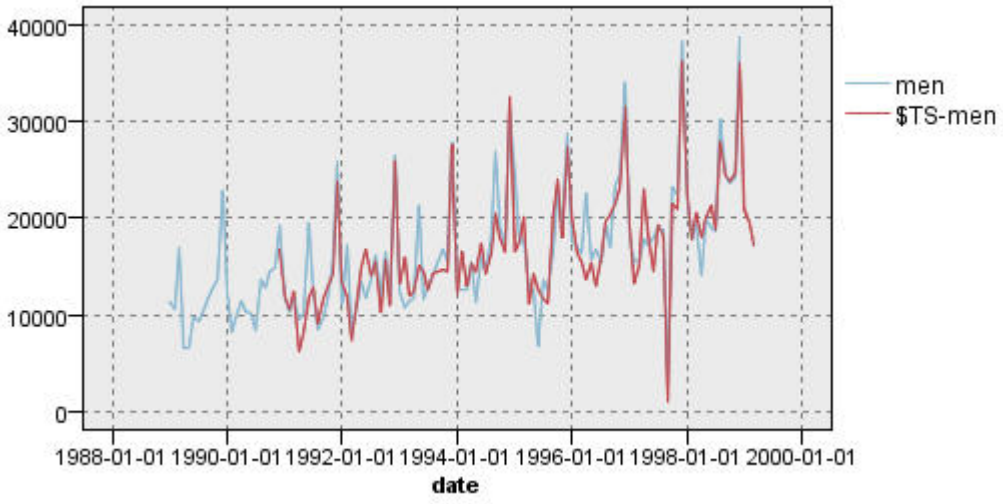


그림 *215.* 예측변수가 지정된 *ARIMA* 모델

이 모델은 큰 아래쪽 말뚝표시를 캡처하고 이를 지금까지의 최적 맞춤으로 작성하여 이전 모델을 개선했습니다.

모델을 더 세분화하려고 시도할 수는 있으나 지금부터의 개선은 최소화될 수 있습니다. 예측변수가 포함된 ARIMA 모델이 더 나은 것이 입증되었으므로 방금 작성한 모델을 사용하겠습니다. 이 예에서는 내년의 판매를 예측할 것입니다.

12. **확인**을 클릭하여 시간 도표 창을 닫으십시오.
13. 시계열 노드를 열고 모델 옵션 탭을 선택하십시오.
14. **레코드를 미래로 확장** 선택란을 선택하고 값을 12로 설정하십시오.
15. **입력의 미래 값 계산** 선택란을 선택하십시오.
16. **실행**을 클릭하여 모델 너깃을 다시 작성하십시오.
17. 시간 구성 노드를 열고 **실행**을 클릭하십시오.

1999에 대한 예측은 좋아 보입니다. 예측한 대로 12월 최대 판매 뒤에 보통 판매 수준으로 돌아가서 일년의 후반 절반 동안 안정적으로 추세가 상승하며 일반적으로 판매가 이전 연도 보다 상승하는 것으로 보입니다.

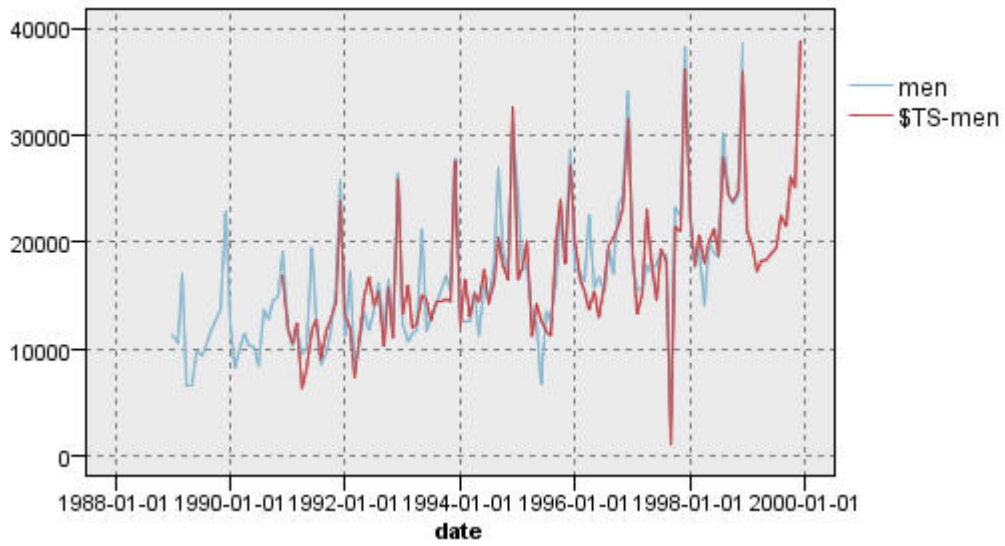


그림 *216. 12*개월로 확장된 판매 예측

요약

상향 추세뿐만 아니라 계절 및 기타 변동도 통합하여 복합 시계열을 성공적으로 모델링했습니다. 또한 시행착오를 통해 방법을 배웠으며 미래 판매를 예측하는 데 사용되는 점점 더 정확한 모델을 작성했습니다.

실제로 예를 들어, 매월 또는 매분기별로 실제 판매 데이터가 업데이트됨에 따라 모델을 다시 적용하여 업데이트된 예측값을 생성해야 합니다. 자세한 정보는 181 페이지의 『시계열 모델 다시 적용』의 내용을 참조하십시오.

제 16 장 고객에 대한 오퍼 작성(Self-Learning)

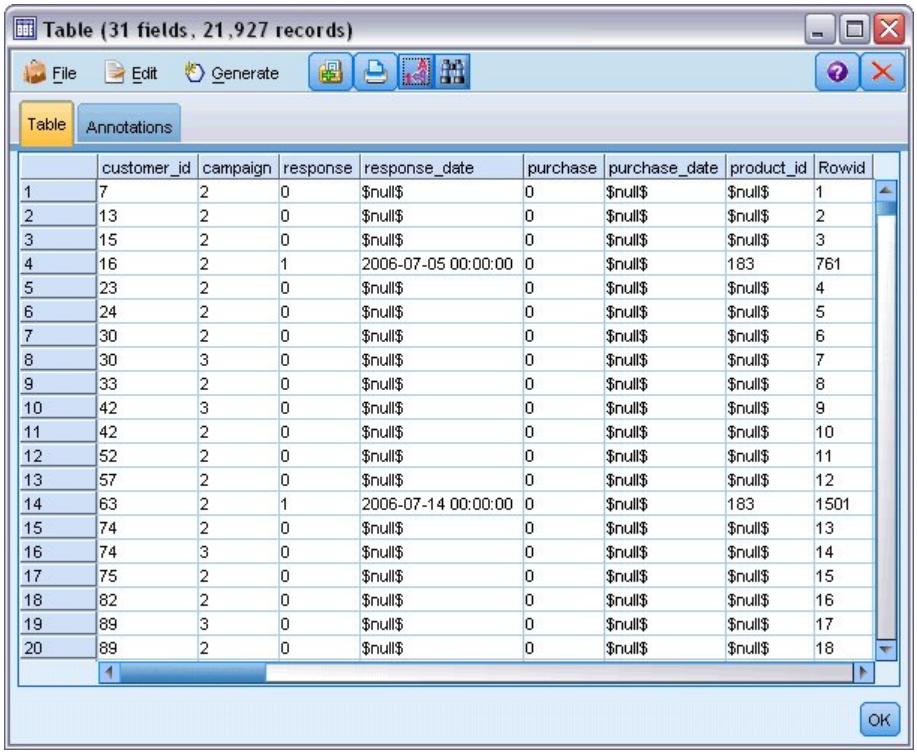
SLRM(Self-Learning Response Model) 노드는 고객에게 가장 적합한 제안과 제안을 수락할 확률을 예측하는 모델을 작성하고 업데이트할 수 있습니다. 이러한 종류의 모델은 마케팅 애플리케이션 또는 콜센터와 같은 고객 관계 관리에서 가장 장점이 많습니다.

이 예에서는 가상의 은행을 기반으로 합니다. 마케팅 부서는 현재 각 고객에 금융 서비스의 올바른 오퍼를 매치하여 미래 캠페인에서 보다 수익성이 좋은 결과를 산출하고자 합니다. 특히, 예에서 Self-Learning Response Model을 사용하여 이전 오퍼 및 반응을 기반으로 가장 우호적으로 반응할 것 같은 고객 특성을 식별하고 결과에 따라 최선의 현재 오퍼를 판촉할 수 있습니다.

이 예는 *pm_customer_train1.sav*, *pm_customer_train2.sav* 및 *pm_customer_train3.sav* 데이터 파일을 참조하는 *pm_selflearn.str* 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 폴더에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *pm_selflearn.str* 파일은 스트림 폴더에 있습니다.

기존 데이터

한 회사에는 지난 캠페인의 고객에 대한 오퍼를 오퍼에 대한 반응과 함께 추적하는 히스토리 데이터가 있습니다. 이러한 데이터는 다른 고객에 대한 반응률을 예측하는 데 사용할 수 있는 인구 통계 및 재정적 정보를 포함합니다.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

그림 *217.* 이전 오퍼에 대한 반응

스트림 작성

1. 사용자의 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *pm_customer_train1.sav*를 가리키는 통계량 파일 소스 노드를 추가하십시오.

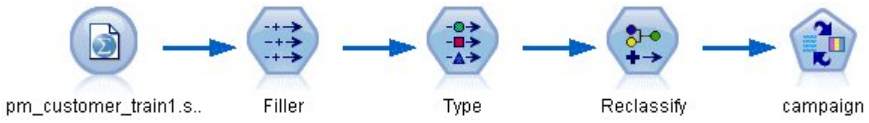


그림 *218. SLRM* 샘플 스트림

2. 채움 노드를 추가하고 필드에서 채우기로 `campaign`을 선택하십시오.
3. 바꾸기 유형으로 **항상**을 선택하십시오.
4. 바꿀 문자열 텍스트 상자에 `to_string(campaign)`을 입력하고 **확인**을 클릭하십시오.

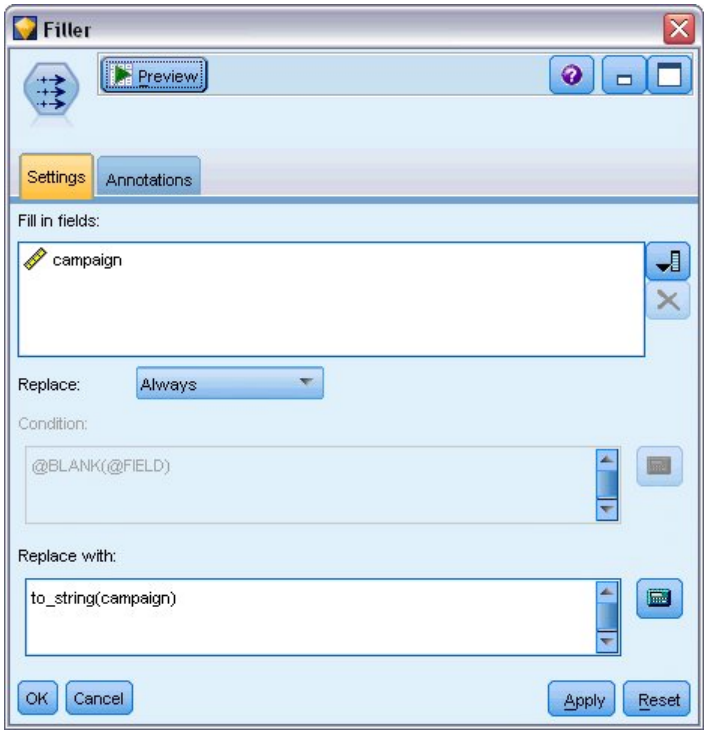


그림 *219.* 캠페인 필드 파생

5. 유형 노드를 추가하고 *customer_id*, *response_date*, *purchase_date*, *product_id*, *Rowid* 및 *X_random* 필드에 대한 역할을 **없음**으로 설정하십시오.

그림 *220.* 유형 노드 설정 변경

6. *campaign* 및 *response* 필드에 대한 역할을 **대상**으로 설정하십시오. 이러한 필드는 예측의 기준이 되는 필드입니다.

   response 필드에 대한 **측정**을 **플래그**로 설정하십시오.
7. **값 읽기**를 클릭한 다음 **확인**을 클릭하십시오.

   캠페인 필드 데이터가 숫자 목록(1, 2, 3 및 4)으로 표시되므로 더 의미 있는 제목을 갖도록 필드를 재분류할 수 있습니다.
8. 재분류 노드를 유형 노드에 연결하십시오.
9. **재분류** 필드에서 **기존 필드**를 선택하십시오.
10. **재분류 필드** 목록에서 **campaign**을 선택하십시오.
11. **가져오기** 단추를 클릭하십시오. 캠페인 값이 원래 값 열에 추가됩니다.
12. 새로운 값 열에서 처음 네 행에 다음과 같은 캠페인 이름을 입력하십시오.
    - **저당**
    - **자동차 대출**
    - **저축**
    - **연금**
13. **확인**을 클릭하십시오.

그림 *221.* 캠페인 이름 재분류

14. SLRM 모델링 노드를 재분류 노드에 연결하십시오. 필드 탭에서 대상 필드에 대해 **캠페인**을 선택하고 대상 반응 필드에 대해 **반응**을 선택하십시오.

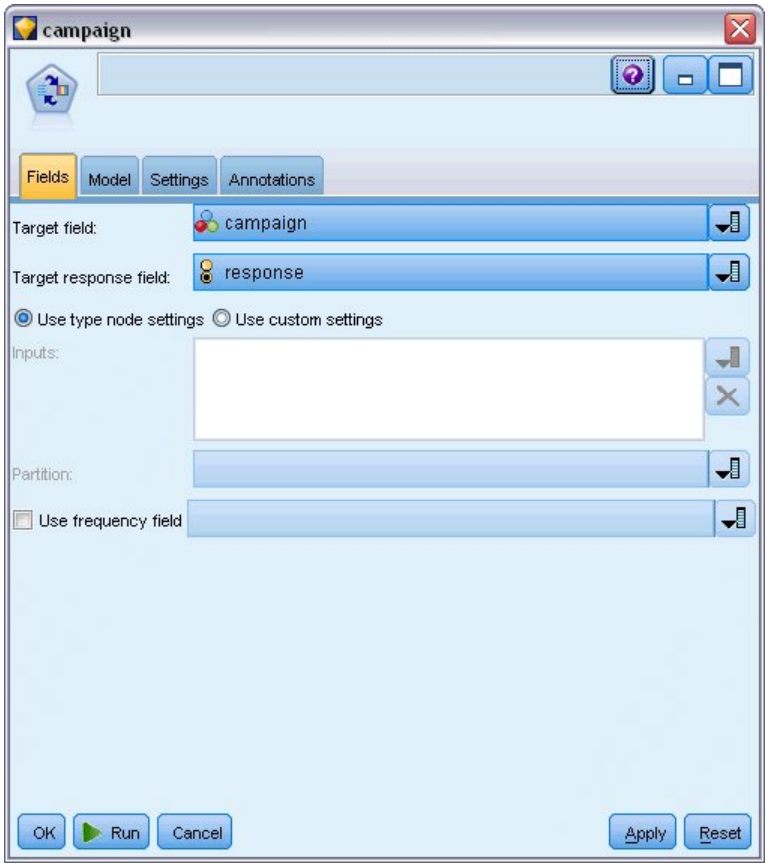


그림 222. 대상 및 대상 반응 선택

15. 설정 탭의 레코드당 예측의 최대 수 필드에서 수를 2로 줄이십시오.

    이는 각 고객에 대해 허용될 확률이 가장 높은 것으로 식별된 두 개의 오퍼가 제공됨을 의미합니다.

16. **모델 신뢰도 고려**가 선택되었는지 확인하고 **실행**을 클릭하십시오.

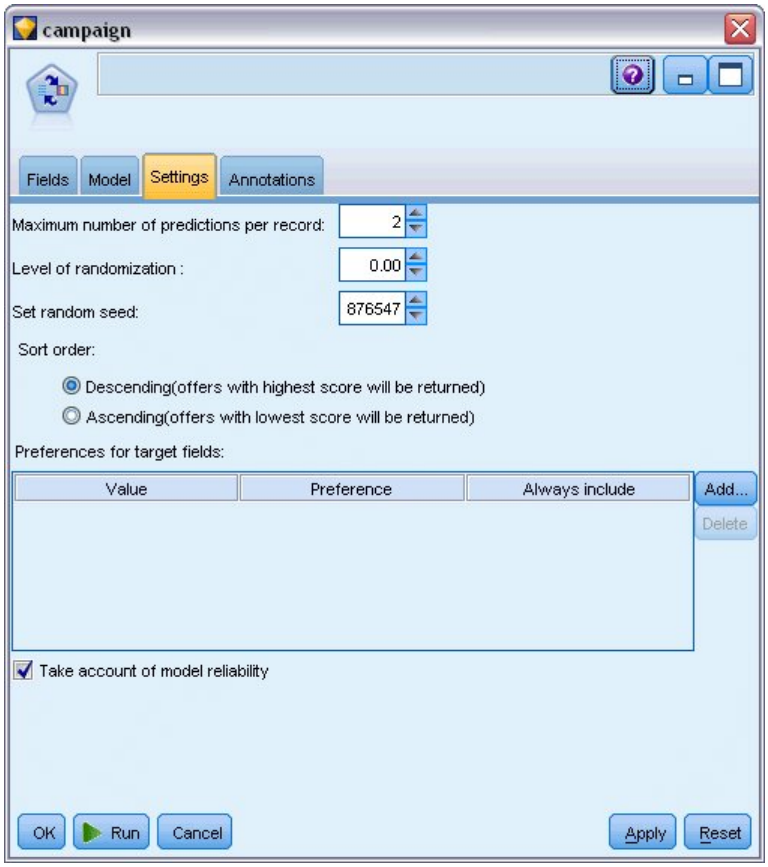


그림 *223. SLRM* 노드 설정

모델 찾아보기

1. 모델 너깃을 여십시오. 모델 탭은 초기에 각 오퍼에 대한 예측의 정확도 추정 및 모델 추정 시 각 예측변수의 상대적인 중요도를 표시합니다.

   대상 변수와 각 예측변수의 상관관계를 표시하려면 오른쪽 분할창의 **보기** 목록에서 **반응에 연관**을 선택하십시오.

2. 예측이 있는 네 개의 각 오퍼 사이에서 전환하려면 왼쪽 분할창의 **보기** 목록에서 필수 오퍼를 선택하십시오.

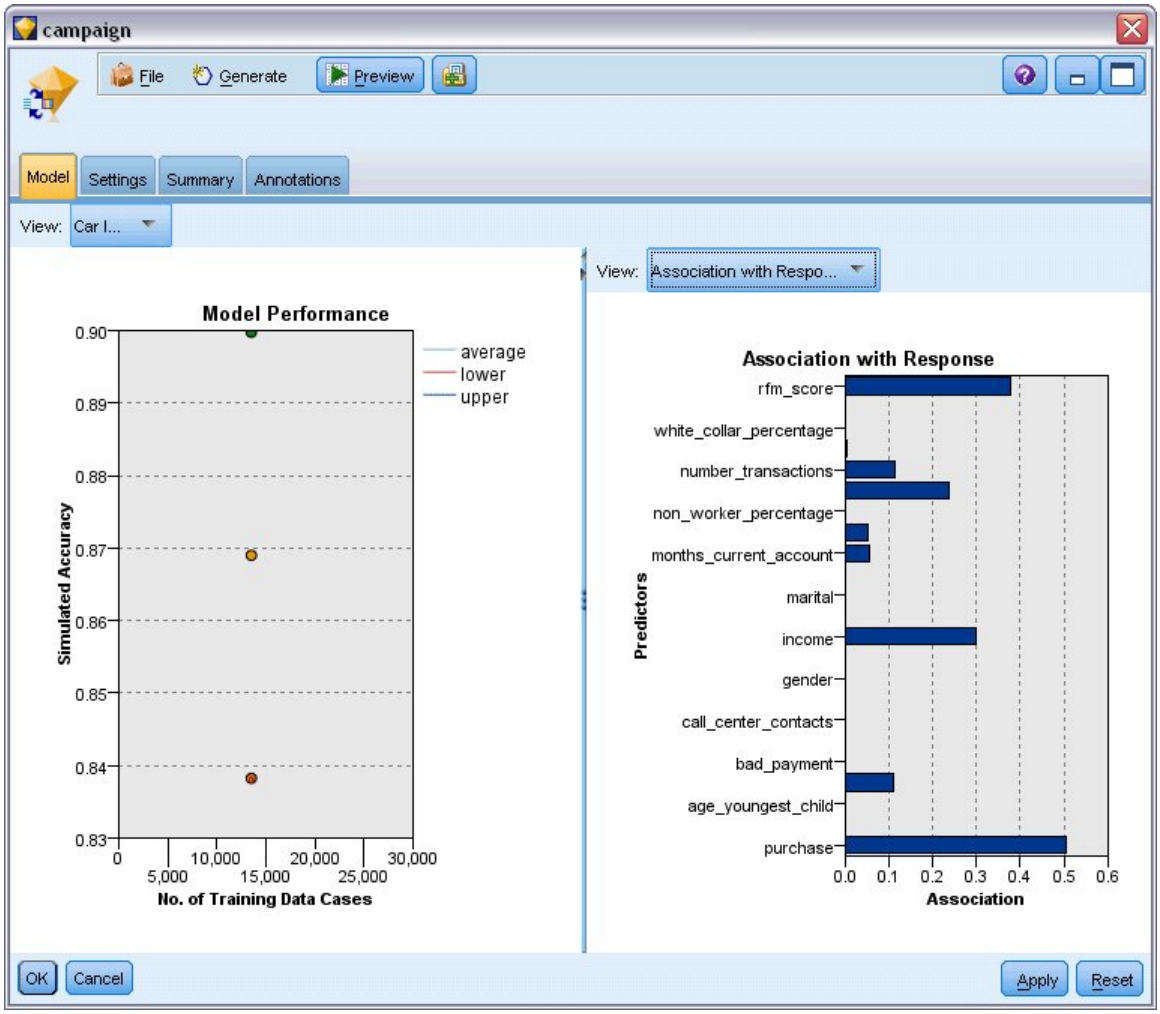


그림 *224. SLRM* 모델 너깃

3. 모델 너깃 창을 닫으십시오.
4. 스트림 캔버스에서 *pm_customer_train1.sav*을 가리키는 IBM SPSS Statistics 파일 소스 노드의 연결을 해제하십시오.
5. 사용자의 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *pm_customer_train2.sav*를 가리키는 통계량 파일 소스 노드를 추가하고 이를 채움 노드에 연결하십시오.

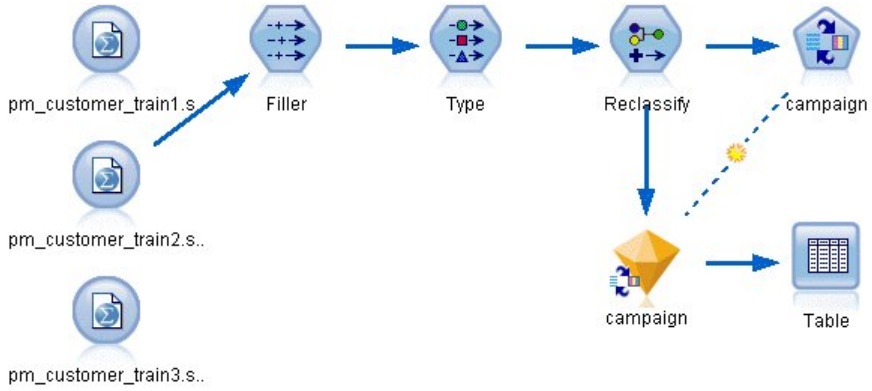


그림 *225.* 두 번째 데이터 소스를 *SLRM* 스트림에 연결

6. SLRM 노드의 모델 탭에서 **기존 모델 학습 계속**을 선택하십시오.

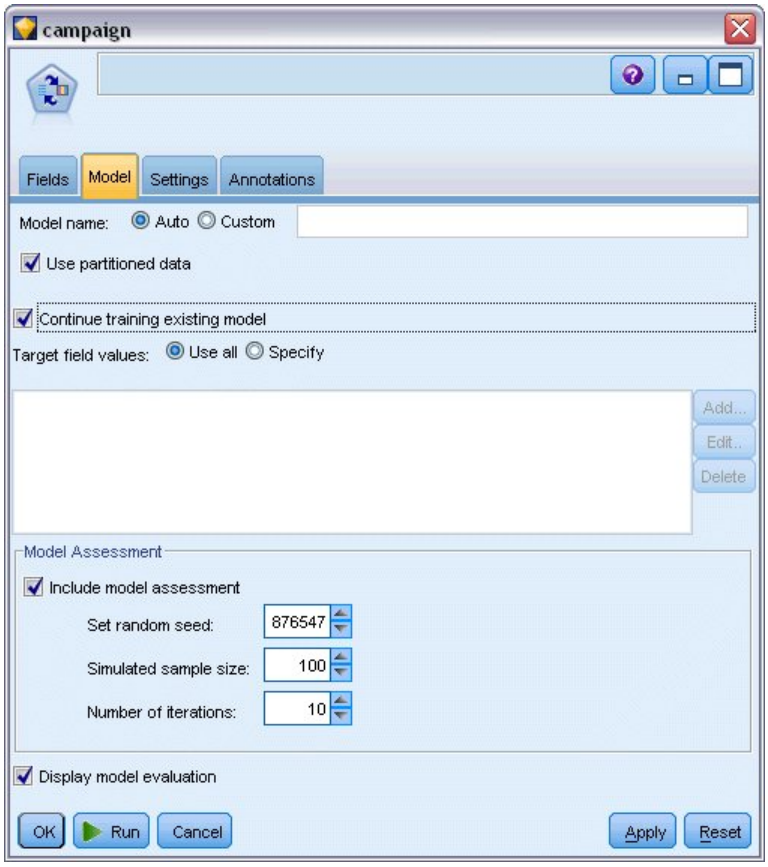


그림 *226.* 모델 학습 계속

7. **실행**을 클릭하여 모델 너깃을 다시 작성하십시오. 세부사항을 보려면 캔버스에서 너깃을 두 번 클릭하십시오.

   이제 모델 탭에 개정된 각 오퍼에 대한 예측의 정확도 추정값이 표시됩니다.

8. 사용자의 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *pm_customer_train3.sav*를 가리키는 통계량 파일 소스 노드를 추가하고 이를 채움 노드에 연결하십시오.

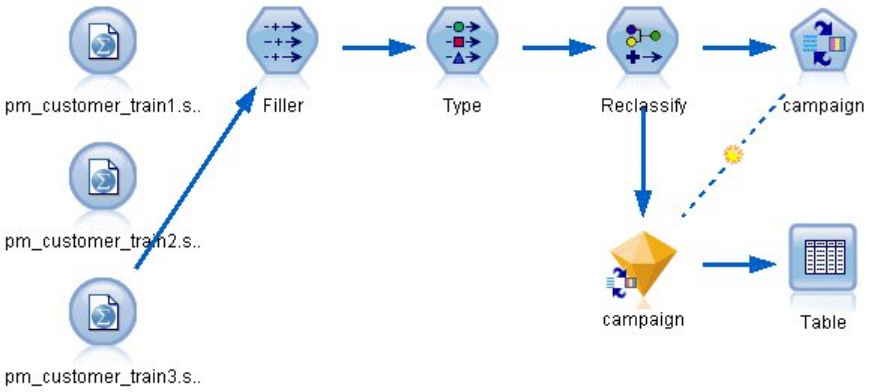


그림 *227.* 세 번째 데이터 소스를 *SLRM* 스트림에 연결

9. **실행**을 클릭하여 모델 너깃을 한 번 더 다시 작성하십시오. 세부사항을 보려면 캔버스에서 너깃을 두 번 클릭하십시오.
10. 이제 모델 탭에 최종적인 각 오퍼에 대한 예측의 정확도 추정값이 표시됩니다.

    추가 데이터 소스를 추가함에 따라 평균 정확도가 (86.9%에서 85.4%로) 약간 떨어진 것을 볼 수 있습니다. 단, 이 변동은 최소 양이며 사용 가능 데이터 내의 약간의 이상 항목에 원인이 있을 수 있습니다.

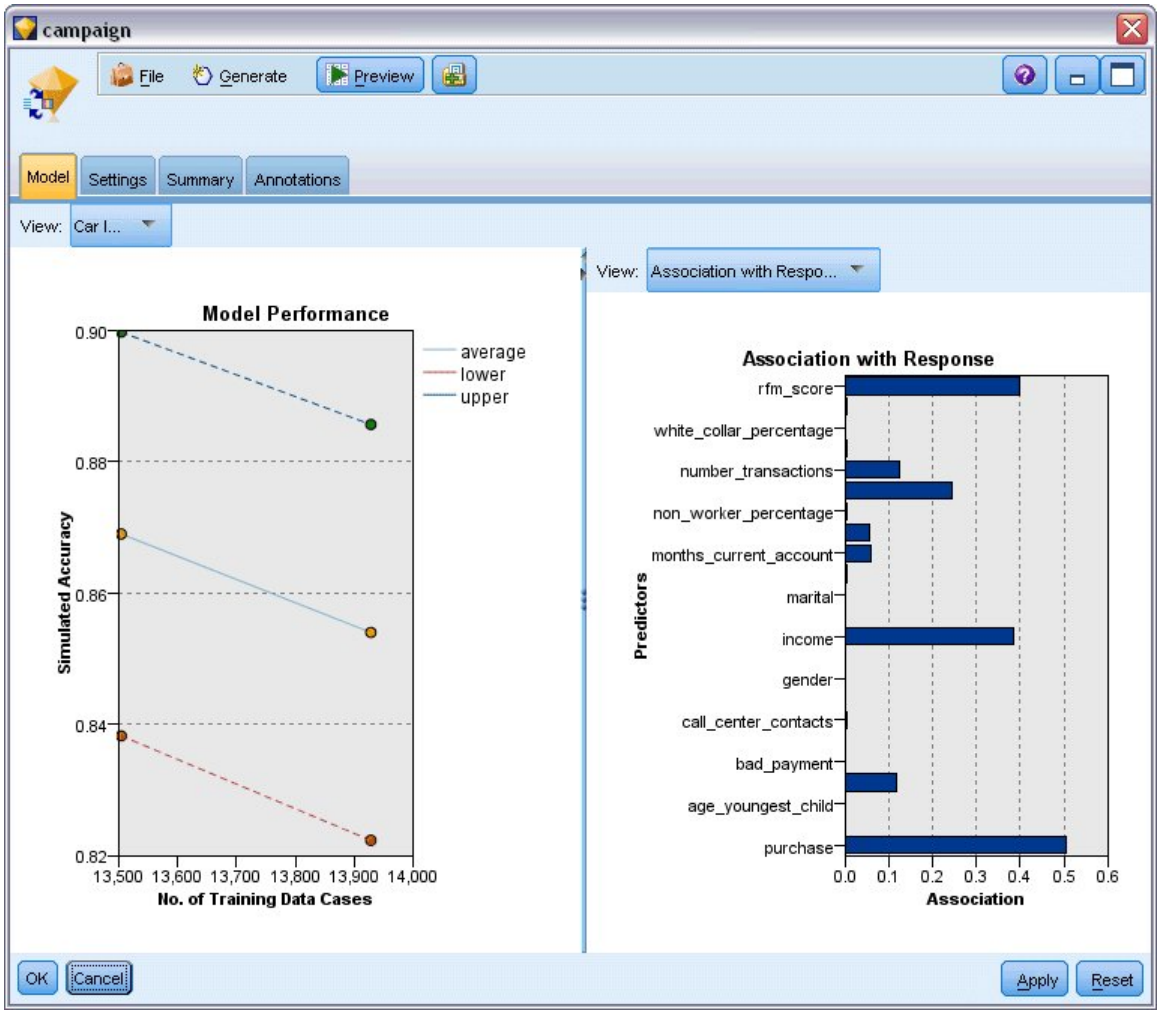
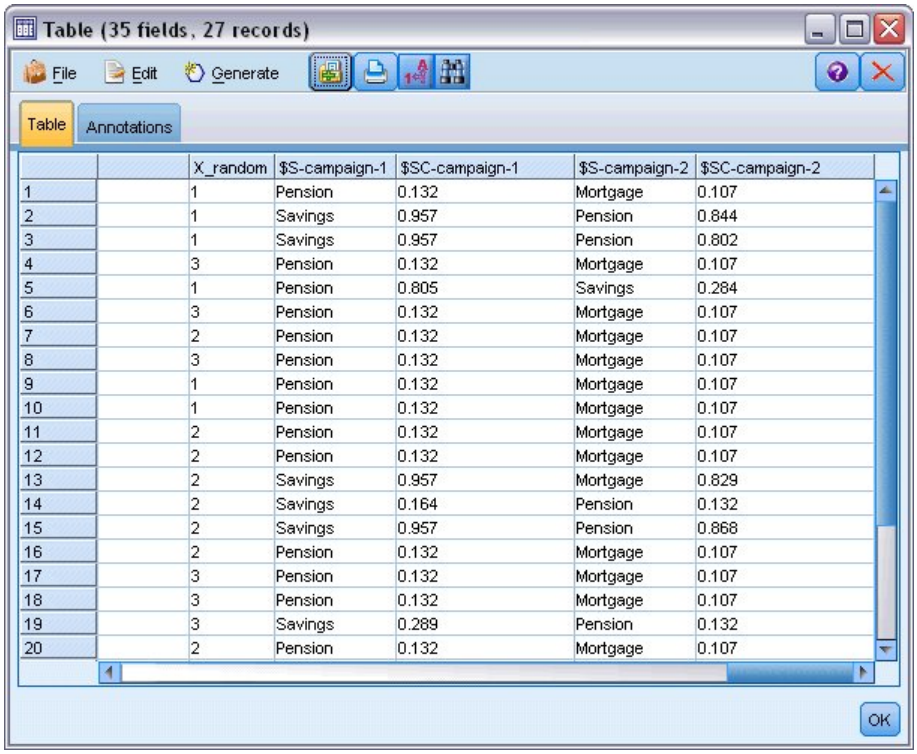


그림 *228.* 업데이트된 *SLRM* 모델 너깃

11. 테이블 노드를 마지막(세 번째)으로 생성된 모델에 연결하고 테이블 노드를 실행하십시오.
12. 테이블의 오른쪽 전체에 걸쳐 스크롤하십시오. 예측은 고객이 수락할 가능성이 가장 높은 오퍼 및 각 고객의 세부사항에 따른 수락 신뢰도를 표시합니다.

예를 들어, 표시된 테이블의 첫 번째 줄에는 이전에 자동차 대출을 사용한 고객이 연금이 제공될 때 이를 수락할 비율이 단지 13.2%(*$SC-campaign-1* 열에서 0.132 값으로 표시됨)의 신뢰도 등급임을 표시합니다. 그러나 두 번째 및 세 번째 열에서는 자동차 대출을 사용한 두 명의 고객이 더 있음을 표시합니다. 이 경우, 해당 고객 및 유사한 히스토리를 가진 다른 고객이 예금계좌가 제공될 때 이를 개설할 신뢰도가 95.7%이며 연금을 수락할 신뢰도가 80% 이상입니다.



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

그림 *229.* 모델 출력 - 예측된 오퍼 및 신뢰도

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 제품 다운로드에서 PDF 파일로 제공되는 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

이러한 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 실세계에서 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

제 17 장 대출 체납자 예측(베이지안 네트워크)

베이지안 네트워크로 관측 및 기록한 증거를 "상식적인" 실제 지식과 결합해서 겉보기에 링크되지 않은 속성을 사용하여 발생 우도를 설정함으로써 확률 모델을 작성할 수 있습니다.

이 예에서는 *bankloan.sav*이라는 데이터 파일을 참조하는 *bayes_bankloan.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있으며 Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 액세스할 수 있습니다. *bayes_bankloan.str* 파일은 *streams* 디렉토리에 있습니다.

예를 들어, 은행이 상환되지 않은 대출 잠재성에 대해 우려하고 있다고 가정합니다. 이전 대출 체납 데이터가 어떤 잠재적 고객이 대출 상환에 문제점이 있을 수 있는지 예측하는 데 사용될 수 있으면 이러한 "나쁜 위험" 고객에게 대출을 거부하거나 대안 상품을 제공할 수 있을 것입니다.

이 예에서는 기존 대출 체납 데이터를 사용하여 잠재적인 미래의 대출 체납자를 예측하고 세 가지 서로 다른 베이지안 신경망 모델 유형을 관찰하여 이 상황에서 예측에 가장 적합한 모델을 설정하는 것에 초점을 둡니다.

스트림 작성

1. *Demos* 폴더에서 *bankloan.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

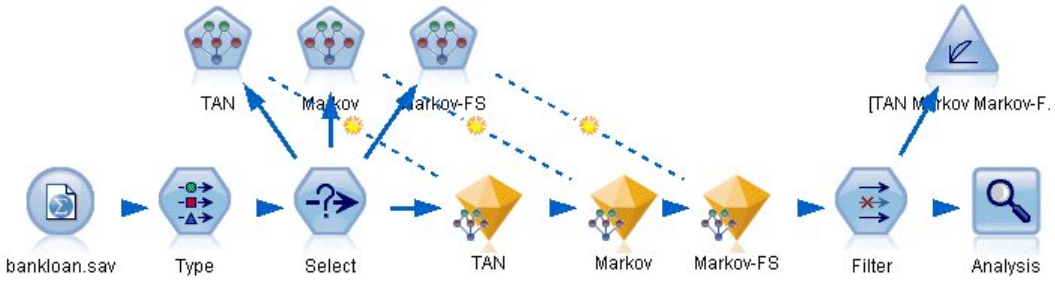


그림 *230.* 베이지안 네트워크 샘플 스트림

2. 유형 노드를 소스 노드에 추가하고 **기본값** 필드의 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.
3. **실제 값** 단추를 클릭하여 값 열을 채우십시오.

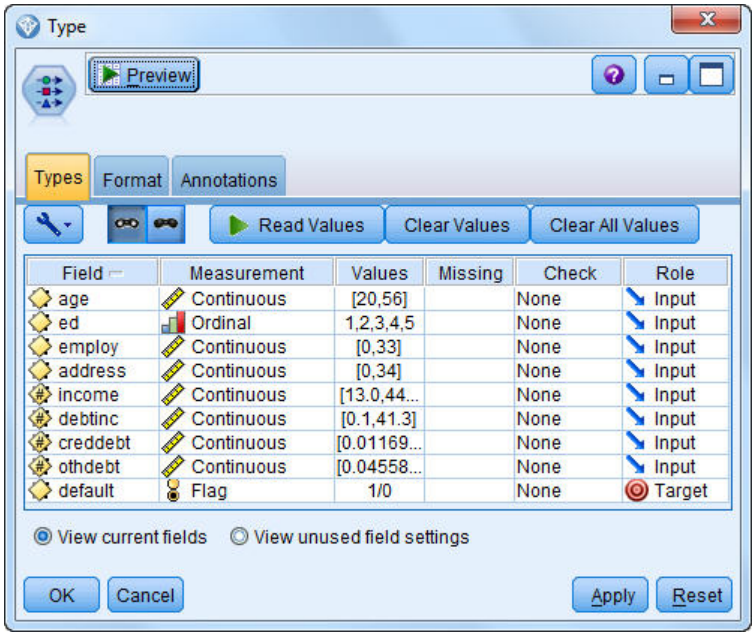


그림 *231.* 대상 필드 선택

대상이 널 값을 가진 케이스는 모델을 작성할 때 소용이 없습니다. 이러한 케이스를 제외하여 모델 평가에서 사용되는 것을 방지할 수 있습니다.

4. 선택 노드를 유형 노드에 연결하십시오.
5. 모드의 경우, **삭제**를 선택하십시오.
6. 조건 선택란에 **default = '$null$'**을 입력하십시오.

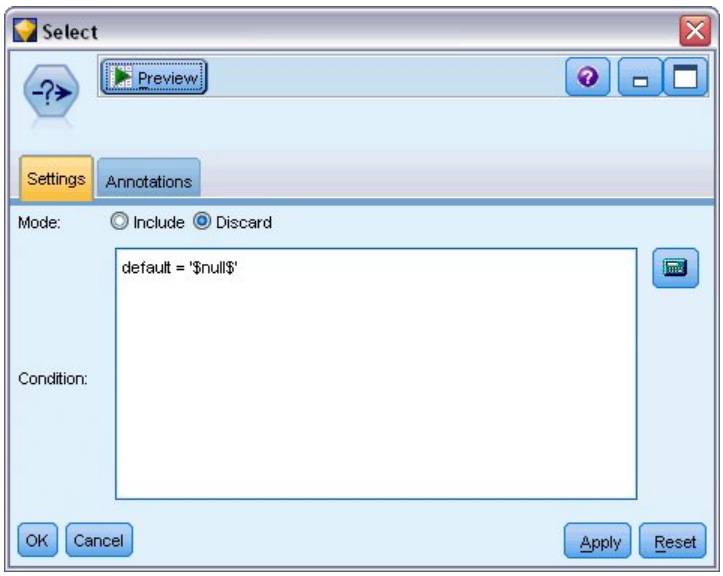


그림 *232.* 널 대상 삭제

여러 다른 유형의 베이지안 네트워크를 작성할 수 있으므로 여러 개를 비교하여 어떤 모델이 가장 나은 예측을 제공하는지 비교할 가치가 있습니다. 처음 작성할 모델은 TAN(Tree Augmented Naïve Bayes) 모델입니다.

7. 베이지안 네트워크 노드를 선택 노드에 연결하십시오.
8. 모델 탭에서 모델 이름으로 **사용자 정의**를 선택하고 텍스트 상자에 TAN을 입력하십시오.
9. 구조 유형으로 **TAN**을 선택하고 **확인**을 클릭하십시오.

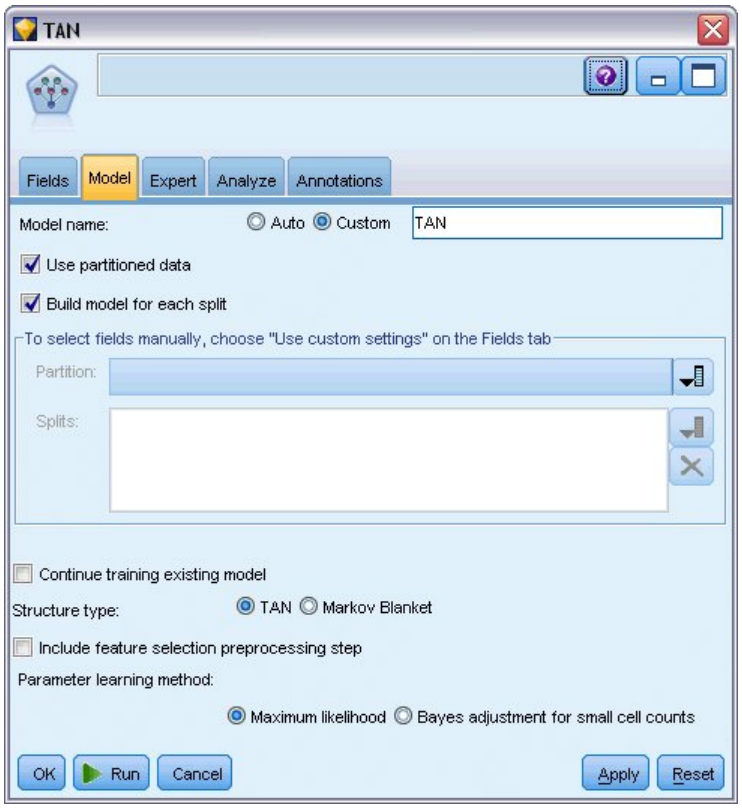


그림 *233. TAN(Tree Augmented Naïve Bayes)* 모델 작성

두 번째로 작성할 모델 유형은 Markov Blanket 구조를 가집니다.

10. 두 번째 베이지안 네트워크 노드를 선택 노드에 연결하십시오.
11. 모델 탭에서 모델 이름으로 **사용자 정의**를 선택하고 텍스트 상자에 Markov를 입력하십시오.
12. 구조 유형으로 **Markov Blanket**을 선택하고 **확인**을 클릭하십시오.

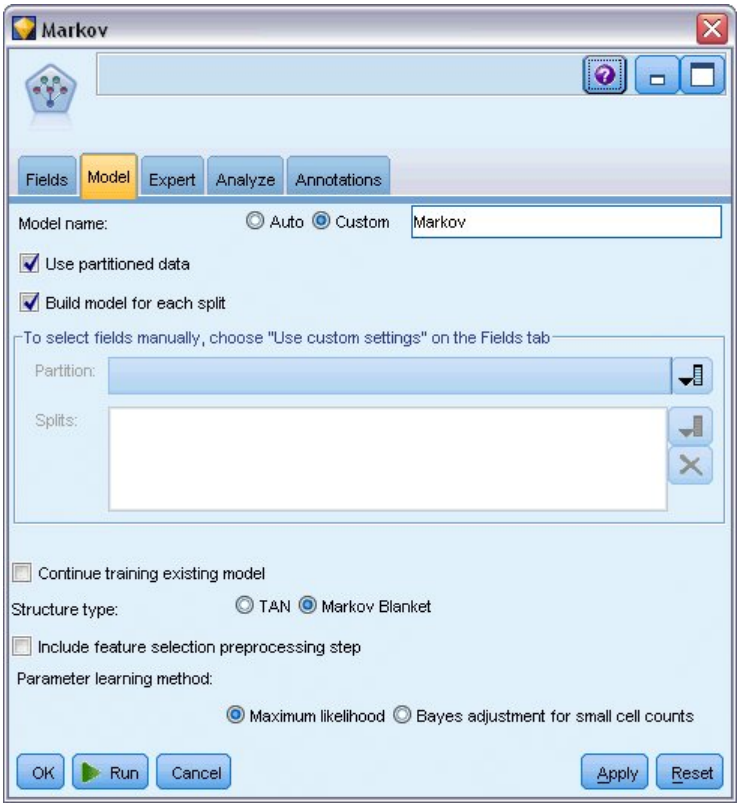


그림 *234. Markov Blanket* 모델 작성

세 번째로 작성할 모델 유형은 Markov Blanket 구조를 가지며 또한 목표변수에 유의적으로 관련된 입력을 선택하기 위해 필드선택 전처리를 사용합니다.

13. 세 번째 베이지안 네트워크 노드를 선택 노드에 연결하십시오.
14. 모델 탭에서 모델 이름으로 **사용자 정의**를 선택하고 텍스트 상자에 `Markov-FS`를 입력하십시오.
15. 구조 유형으로 **Markov Blanket**을 선택하십시오.
16. **필드선택 전처리 단계 포함**을 선택하고 **확인**을 클릭하십시오.

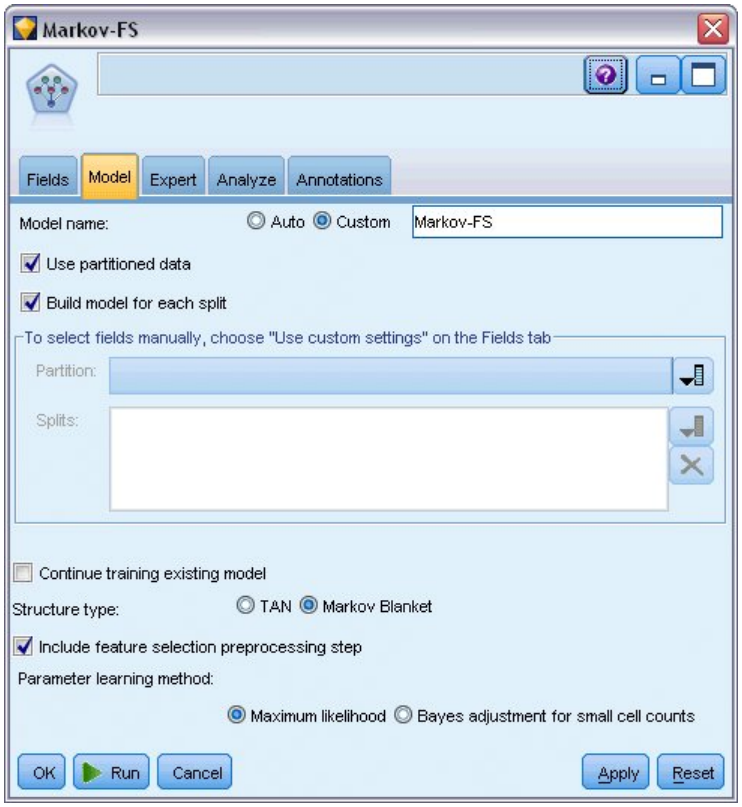


그림 *235.* 필드선택 전처리가 있는 *Markov Blanket* 모델 작성

모델 찾아보기

1. 스트림을 실행하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에 추가됩니다. 세부사항을 보려면 스트림에서 임의의 모델 너깃을 두 번 클릭하십시오.

   모델 너깃 모델 탭이 두 개의 분할창으로 분할됩니다. 왼쪽 분할창에는 목표와 가장 중요한 예측변수 간 관계 및 예측변수 사이의 관계를 표시하는 노드의 네트워크 그래프가 있습니다.

   오른쪽 분할창은 모델 추정에서 각 예측변수의 상대적 중요도를 표시하는 예측변수 중요도 또는 상위 노드의 개별 값 조합과 각 노드 값의 조건부 확률 값을 포함하는 조건부 확률을 표시합니다.

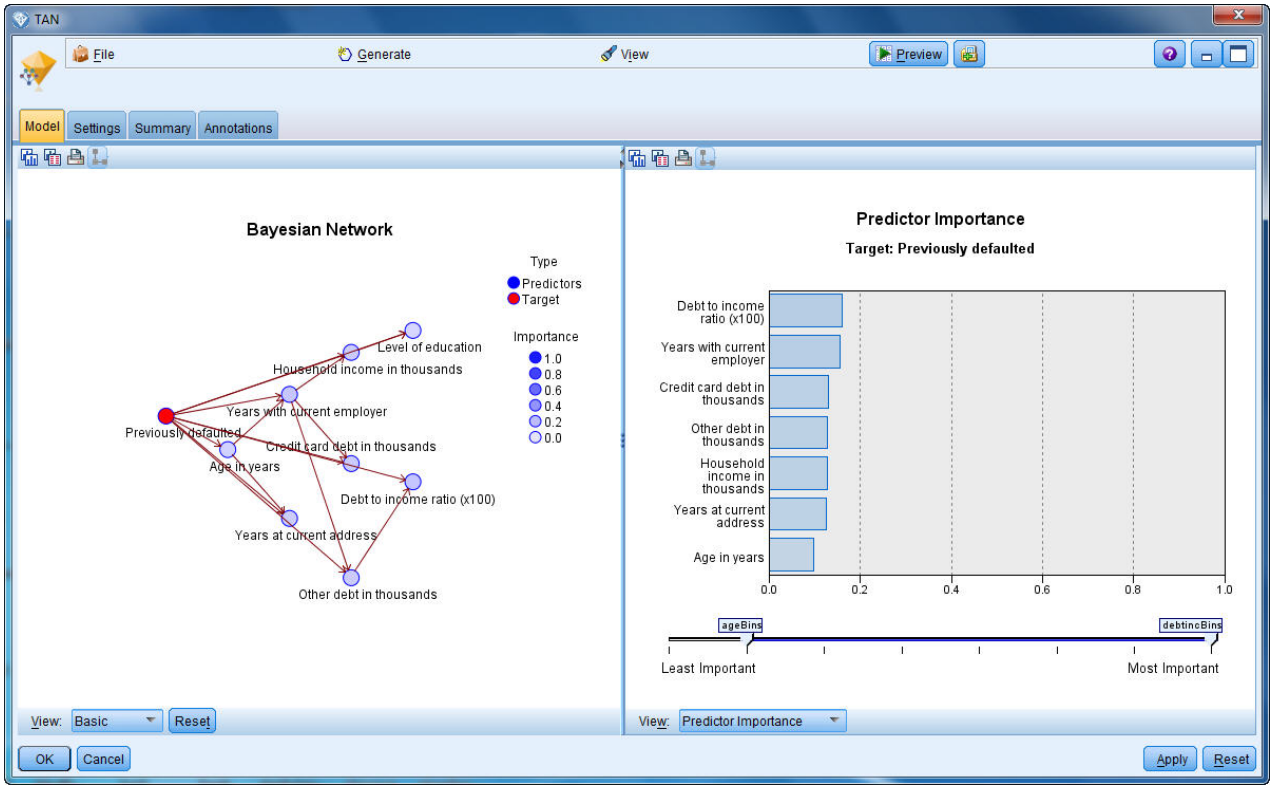


그림 *236. TAN(Tree Augmented Naïve Bayes)* 모델 보기

2. TAN 모델 너깃을 Markov 너깃에 연결하십시오(경고 대화 상자에서 **바꾸기**를 선택하십시오).
3. Markov 너깃을 Markov-FS 너깃에 연결하십시오(경고 대화 상자에서 **바꾸기**를 선택하십시오).
4. 보기 쉽도록 세 개의 너깃을 선택 노드와 맞추십시오.

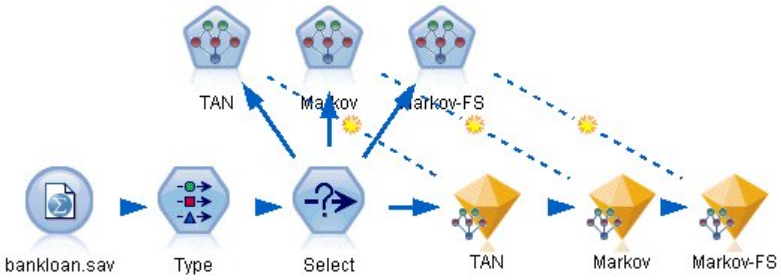


그림 *237.* 스트림에서 너깃 맞춤

5. 작성할 평가 그래프에서 명확하게 표시하기 위해 모델 출력의 이름을 변경하려면 필터 노드를 Markov-FS 모델 너깃에 연결하십시오.
6. 오른쪽 필드 열에서 $B-default를 TAN으로, $B1-default를 Markov로, $B2-default를 Markov-FS로 이름을 변경하십시오.

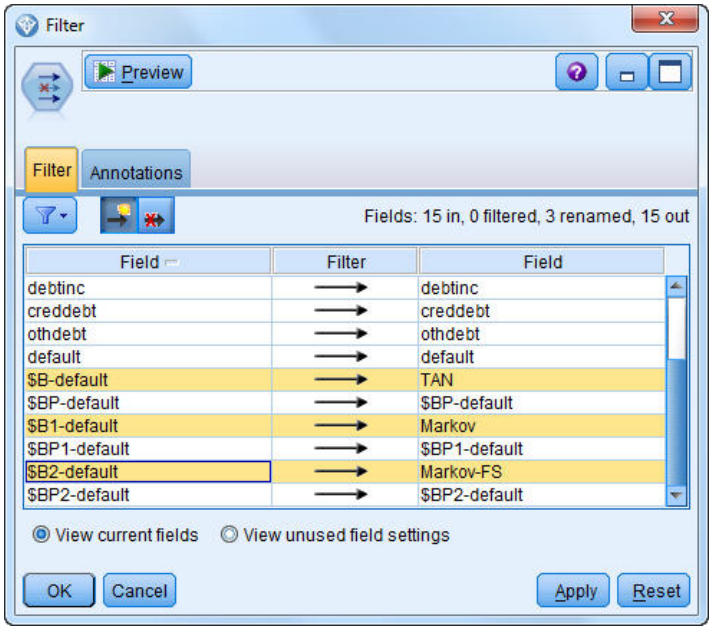


그림 *238.* 모델 필드 이름 변경

모델의 예측 정확도를 비교하기 위해 이익 차트를 작성할 수 있습니다.

7. 평가 그래프 노드를 필터 노드에 연결하고 기본 설정을 사용하여 그래프 노드를 실행하십시오.

그래프는 각 모델 유형이 유사한 결과를 생산함을 표시하나 Markov 모델이 약간 더 낫습니다.

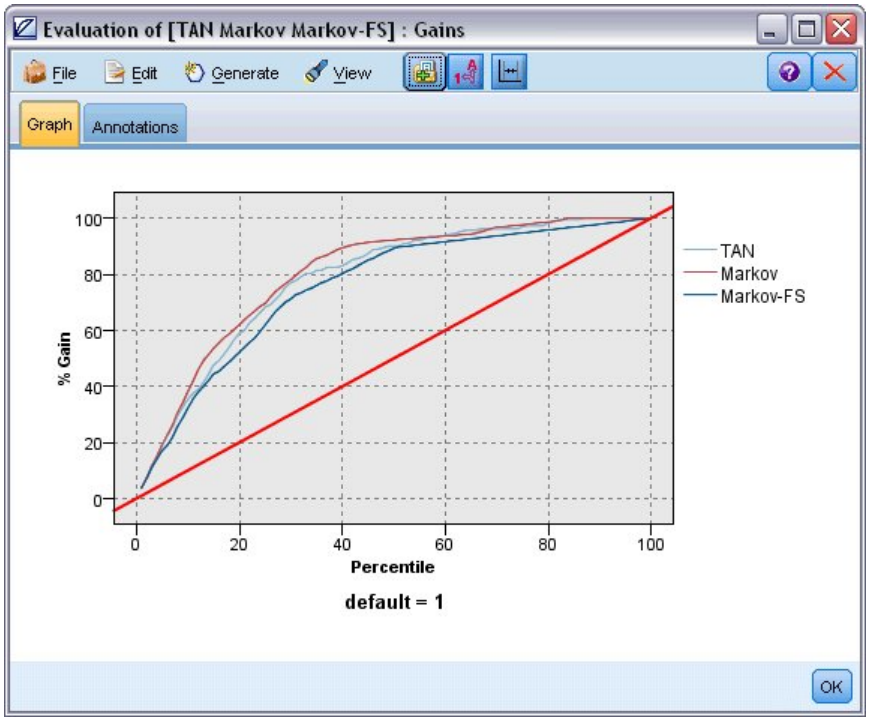


그림 *239.* 평가 모델 정확도

각 모델이 얼마나 잘 예측하는지 확인하기 위해 평가 그래프 대신 분석 노드를 사용할 수 있습니다. 이는 정확성 및 부정확성 예측 둘 다에 대해 퍼센트 단위로 정확도를 표시합니다.

8. 분석 노드를 필터 노드에 연결하고 기본 설정을 사용하여 분석 노드를 실행하십시오.

평가 그래프를 사용함으로써 정확하게 예측하는 데는 Markov 모델이 약간 낫지만, Markov-FS 모델의 경우 Markov 모델보다 아주 약간의 퍼센트로 뒤처짐을 알 수 있습니다. 이는 Markov-FS 모델이 결과를 계산하는 데 필요한 입력이 더 적으므로 데이터 수집, 입력 시간 및 처리 시간이 절약됨을 의미합니다.

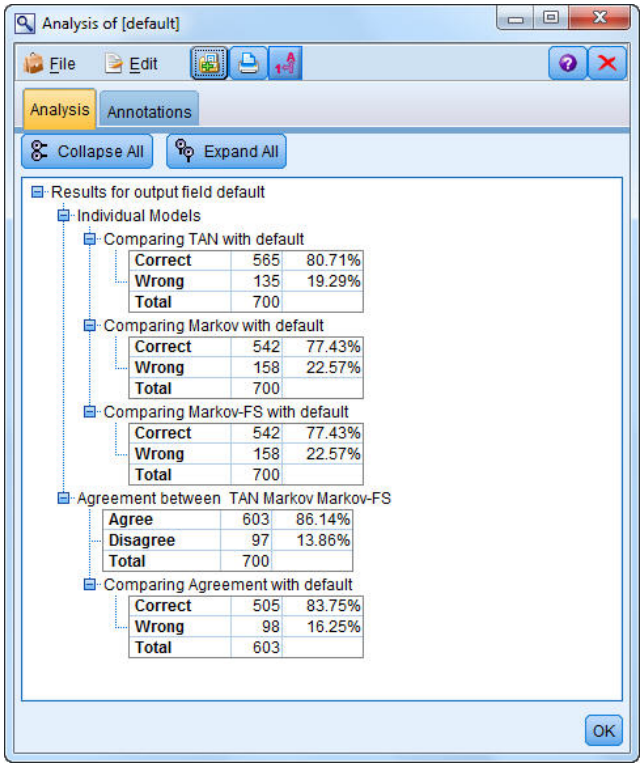


그림 *240.* 모델 정확도 분석

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 설치 디스크의 *Documentation* 디렉토리에서 사용 가능한 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

이러한 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 실세계에서 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

제 18 장 월 단위로 모델 재교육(베이지안 네트워크)

베이지안 네트워크로 관측 및 기록한 증거를 "상식적인" 실제 지식과 결합해서 겉보기에 링크되지 않은 속성을 사용하여 발생 우도를 설정함으로써 확률 모델을 작성할 수 있습니다.

이 예에서는 *telco_Jan.sav* 및 *telco_Feb.sav*라는 데이터 파일을 참조하는 *bayes_churn_retrain.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있으며 Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 액세스할 수 있습니다. *bayes_churn_retrain.str* 파일은 *streams* 디렉토리에 있습니다.

예를 들어, 통신사업자가 경쟁자에게 빠져나가고 있는 고객(서비스 제공자를 바꾸는 고객) 수에 대해 걱정하고 있습니다. 히스토리 고객 데이터가 어떤 고객이 미래에 서비스 제공자를 바꿀 가능성이 높은지 예측하는 데 사용될 수 있으면 이러한 고객을 인센티브 또는 기타 오퍼로 대상화하여 다른 서비스 제공자로 변환하는 것을 억제할 수 있습니다.

이 예에서는 기존 개월의 서비스 제공자를 바꾸는 고객 데이터를 사용하여 어떤 고객이 미래에 서비스 제공자를 바꿀 가능성이 높은지 예측하고 다음 개월의 데이터를 추가하여 정교화하고 모델을 재교육하는 데 초점을 맞춥니다.

스트림 작성

1. *Demos* 폴더에서 *telco_Jan.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

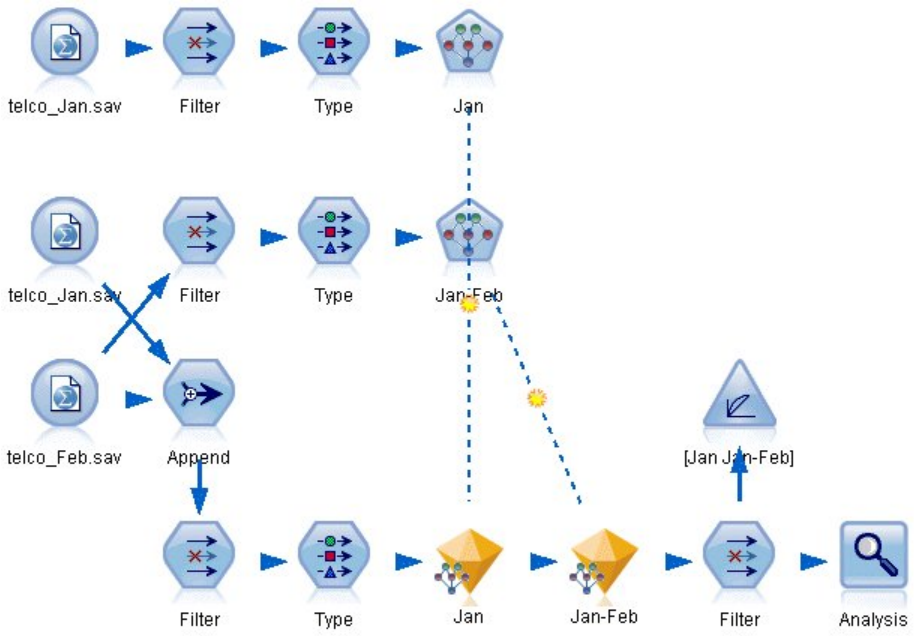


그림 *241.* 베이지안 네트워크 샘플 스트림

이전 분석은 여러 데이터 필드가 서비스 제공자를 바꾸는 고객을 예측할 때 중요도가 거의 없음을 표시합니다. 이러한 필드는 모델을 작성하고 스코어링할 때 처리 속도를 높이기 위해 데이터 세트에서 필터링될 수 있습니다.

2. 필터 노드를 소스 노드에 연결하십시오.
3. 주소, 연령, 서비스 제공자를 바꾸는 고객, 통신사용등급 , 교육수준, 고용, 성별, 결혼여부, 거주, 은퇴 및 재직을 제외한 모든 필드를 제외하십시오.
4. **확인**을 클릭하십시오.

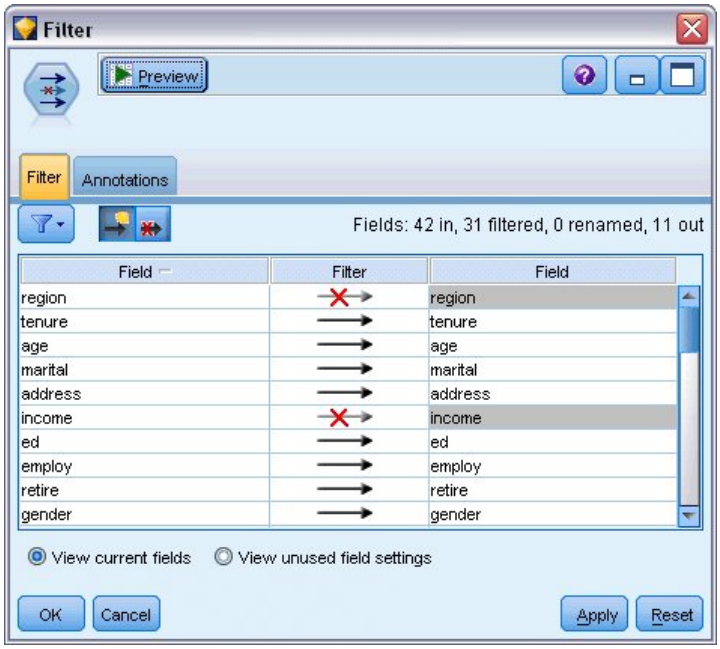


그림 *242.* 불필요한 필드 필터링

5. 유형 노드를 필터 노드에 연결하십시오.
6. 유형 노드를 열고 **실제 값** 단추를 클릭하여 값 열을 채우십시오.
7. 평가 노드가 어떤 값이 참이고 어떤 값이 거짓인지 평가할 수 있도록 서비스 제공자를 바꾸는 고객 필드에 대한 측정 수준을 **플래그**로 설정하고 역할을 **대상**으로 설정하십시오. **확인**을 클릭하십시오.

그림 *243.* 대상 필드 선택

여러 다른 유형의 베이지안 네트워크를 작성할 수 있으나 이 예에서는 TAN(Tree Augmented Naïve Bayes) 모델을 작성할 것입니다. 이는 대형 네트워크를 생성하므로 데이터 변수 사이에 모든 가능한 링크를 포함하여 강력한 초기 모델을 작성했는지 확인하십시오.

8. 베이지안 네트워크 노드를 유형 노드에 연결하십시오.
9. 모델 탭에서 모델 이름으로 **사용자 정의**를 선택하고 텍스트 상자에 Jan을 입력하십시오.
10. 모수 학습 방법으로 **작은 셀 빈도에 대한 Bayes 조정**을 선택하십시오.
11. **실행**을 클릭하십시오. 모델 너깃이 스트림에 추가되고 오른쪽 상단 코너에 있는 모델 팔레트에도 추가됩니다.

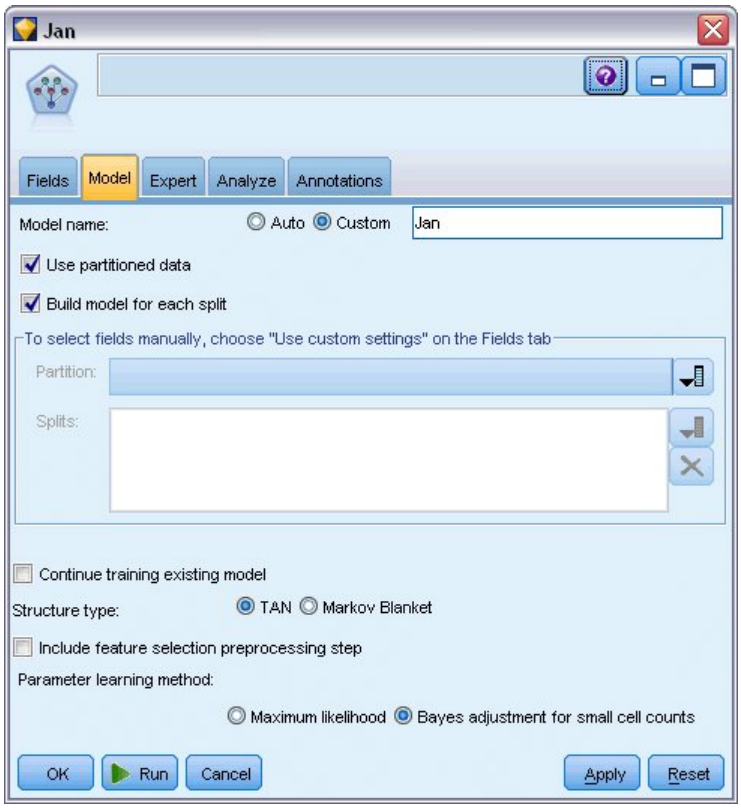


그림 *244. TAN(Tree Augmented Naïve Bayes)* 모델 작성

12. *Demos* 폴더에서 *telco_Feb.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.
13. 이 새 소스 노드를 필터 노드에 연결하십시오(경고 대화 상자에서 **바꾸기**를 선택하여 이전 소스 노드에 대한 연결을 바꾸십시오).

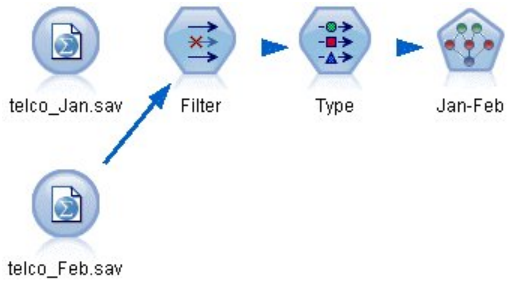


그림 *245.* 두 번째 월의 데이터 추가

14. 베이지안 네트워크 노드의 모델 탭에서 모델 이름으로 **사용자 정의**를 선택하고 텍스트 상자에 Jan-Feb을 입력하십시오.
15. **기존 모델 학습 계속**을 선택하십시오.
16. **실행**을 클릭하십시오. 모델 너깃이 스트림의 기존 모델 너깃을 덮어쓰나 이는 오른쪽 상단 코너에 있는 모델 팔레트에도 추가됩니다.

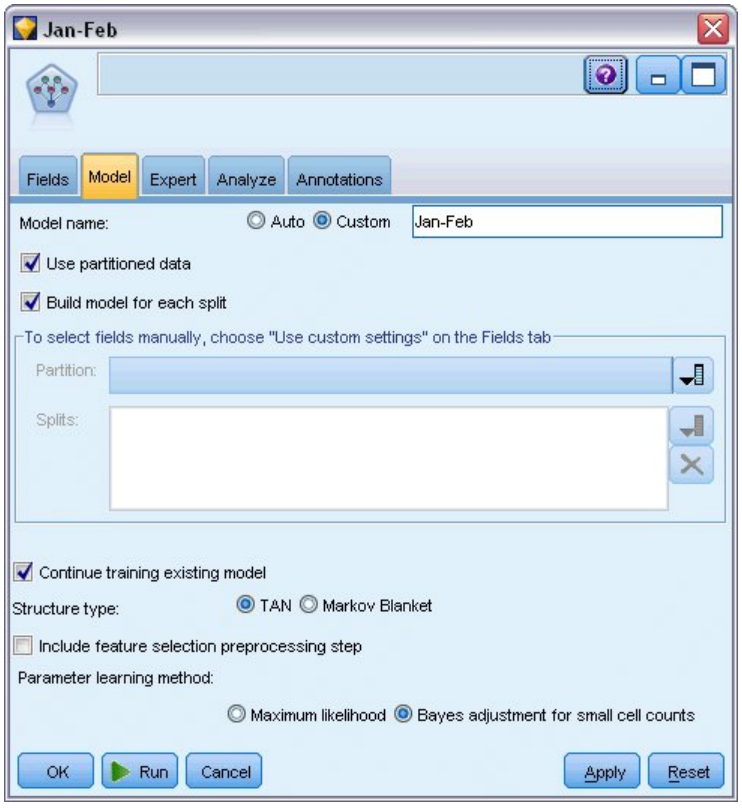


그림 *246.* 모델 재교육

모델 평가

모델을 비교하려면 두 개의 데이터 세트를 조합해야 합니다.

1. 추가 노드를 추가하고 *telco_Jan.sav* 및 *telco_Feb.sav* 두 소스 노드를 모두 연결하십시오.

그림 *247.* 두 개의 데이터 소스 추가

2. 이전 스트림에서 필터 및 유형 노드를 복사하여 스트림 캔버스에 붙여넣으십시오.
3. 추가 노드를 새로 복사한 필터 노드에 연결하십시오.

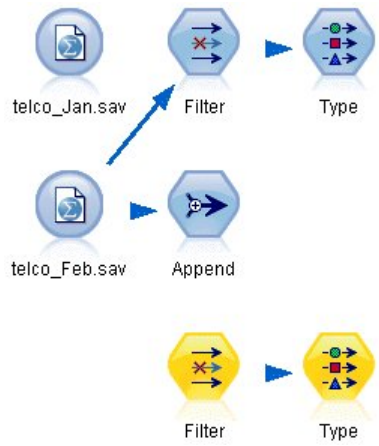


그림 *248.* 복사한 노드를 스트림에 붙여넣기

두 개의 베이지안 신경망 모델에 대한 너깃은 오른쪽 상단 코너의 모델 팔레트에 있습니다.

4. 1월 모델 너깃을 두 번 클릭하여 이를 스트림으로 가져오고 새로 복사한 유형 노드에 연결하십시오.
5. 이미 스트림에 있는 1월-2월 모델 너깃을 1월 모델 너깃에 연결하십시오.
6. 1월 모델 너깃을 여십시오.

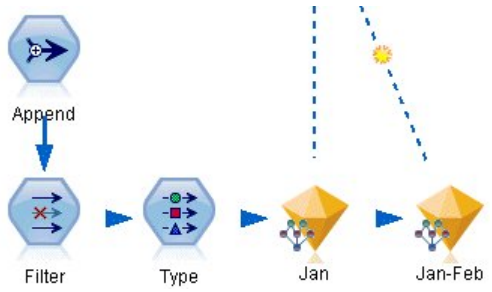


그림 *249.* 스트림에 너깃 추가

베이지안 네트워크 모델 너깃 모델 탭이 두 개의 열로 분할됩니다. 왼쪽 열에는 목표와 가장 중요한 예측변수 간 관계 및 예측변수 사이의 관계를 표시하는 노드의 네트워크 그래프가 있습니다.

오른쪽 열은 모델 추정에서 각 예측변수의 상대적 중요도를 표시하는 예측변수 중요도 또는 상위 노드의 개별 값 조합과 각 노드 값의 조건부 확률 값을 포함하는 조건부 확률을 표시합니다.

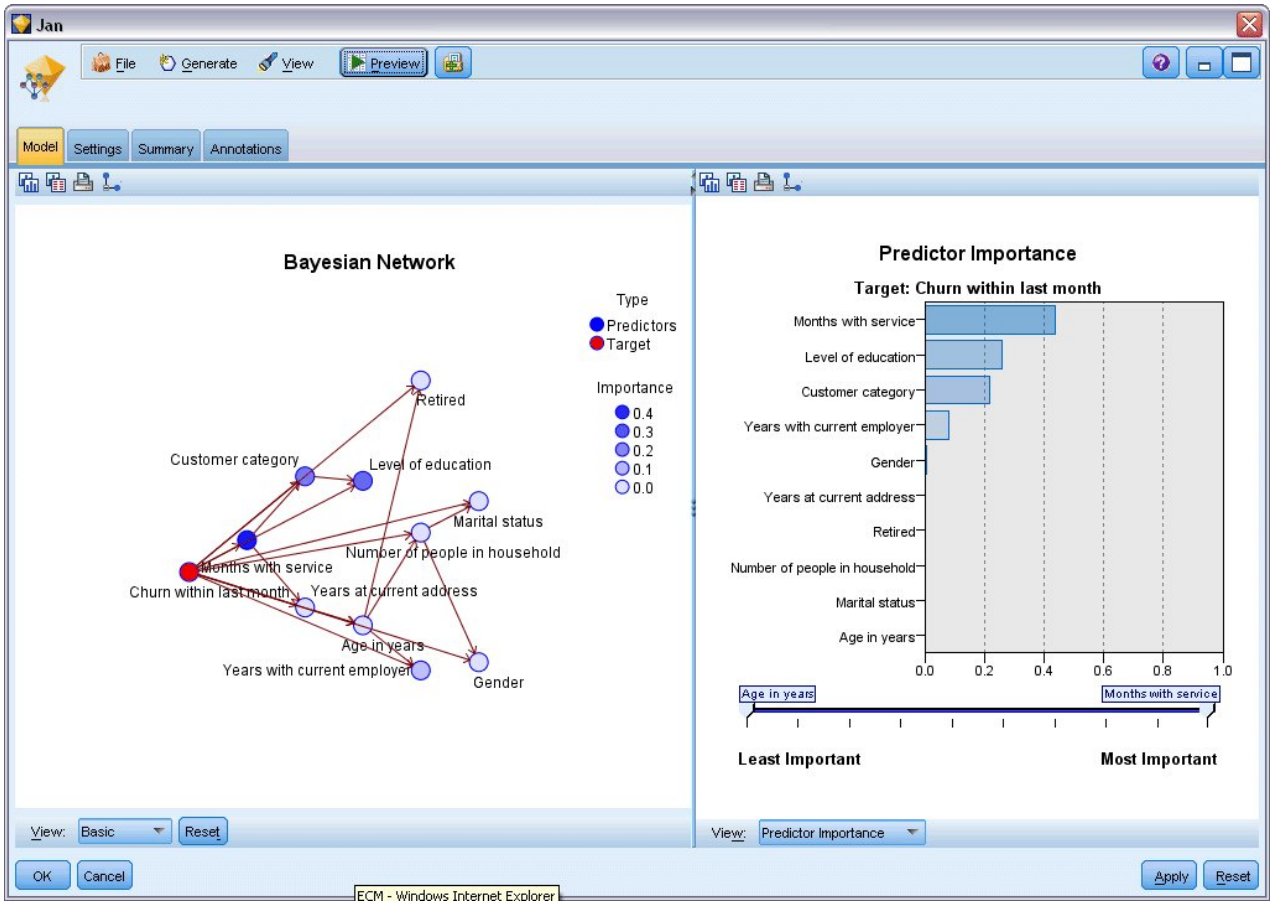


그림 *250.* 예측변수 중요도를 표시하는 베이지안 네트워크 모델

임의의 노드에 대한 조건부 확률을 표시하려면 왼쪽 열의 노드를 클릭하십시오. 필수 세부사항을 표시하기 위해 오른쪽 열이 업데이트됩니다.

각 구간에 대해 노드의 상위 및 동위 노드에 관해 나뉘어진 데이터 값인 조건부 확률이 표시됩니다.

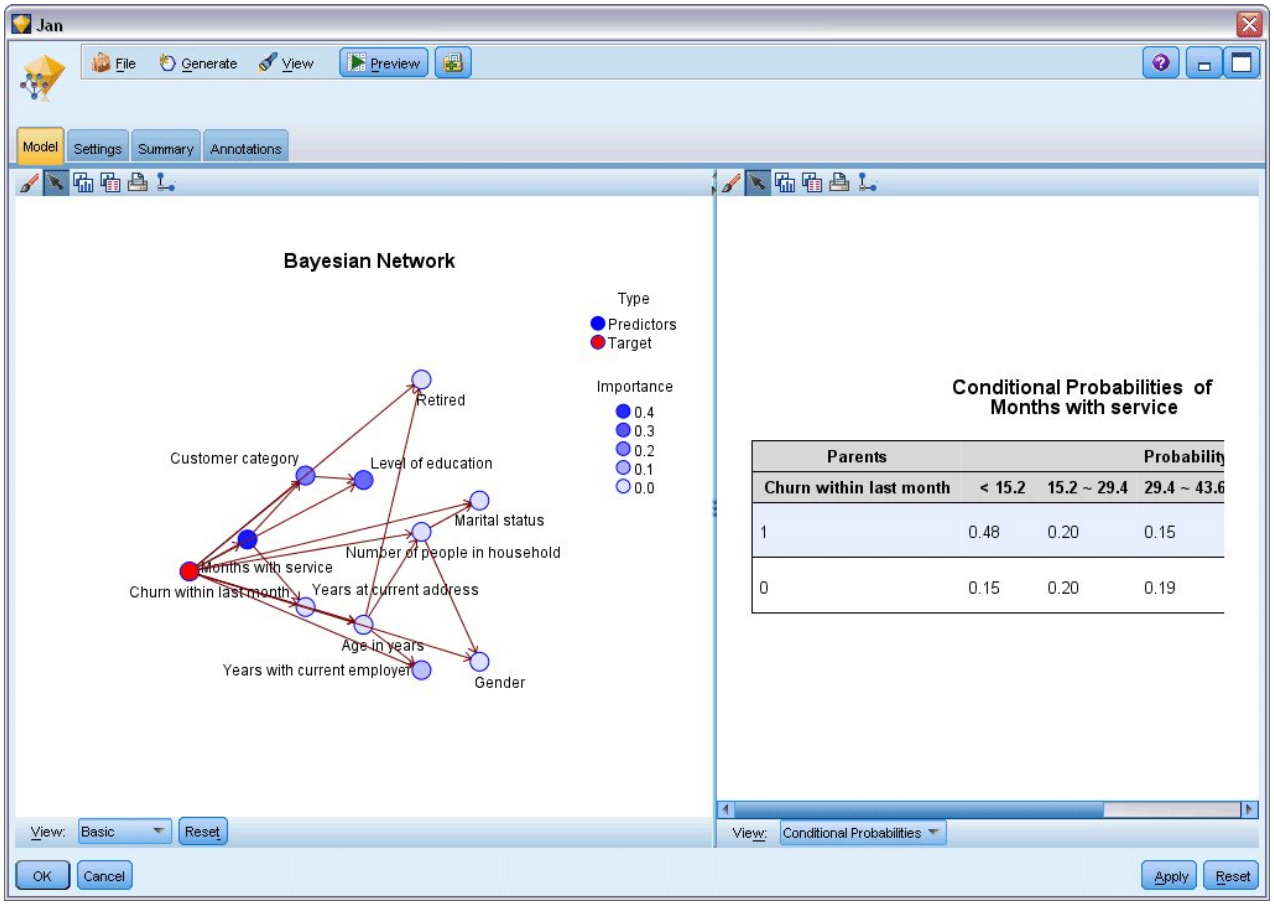


그림 *251.* 조건부 확률을 표시하는 베이지안 네트워크 모델

7. 명확하게 표시하기 위해 모델 출력의 이름을 변경하려면 필터 노드를 1월-2월 모델 너깃에 연결하십시오.
8. 오른쪽 필드 열에서 $B-churn를 1월로, $B1-churn를 1월-2월로 변경하십시오.

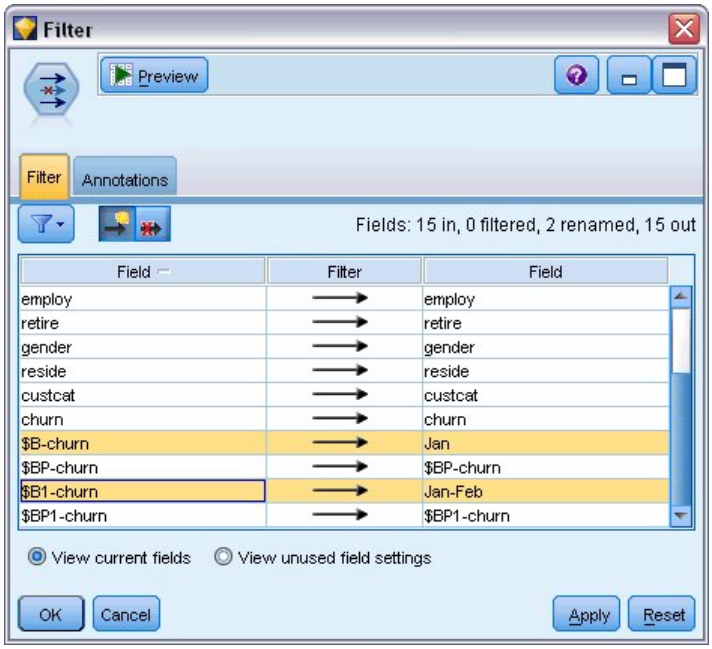


그림 252. 모델 필드 이름 변경

각 모델이 서비스 제공자를 바꾸는 고객을 얼마나 잘 예측하는지 검사하기 위해 분석 노드를 사용하십시오. 분석 노드는 정확성 및 부정확성 예측 둘 다에 대해 퍼센트 단위로 정확도를 표시합니다.

9. 분석 노드를 필터 노드에 연결하십시오.
10. 분석 노드를 열고 **실행**을 클릭하십시오.

그러면 서비스 제공자를 바꾸는 고객을 예언할 때 두 모델 모두 유사한 정도의 정확도를 가진 것을 표시합니다.

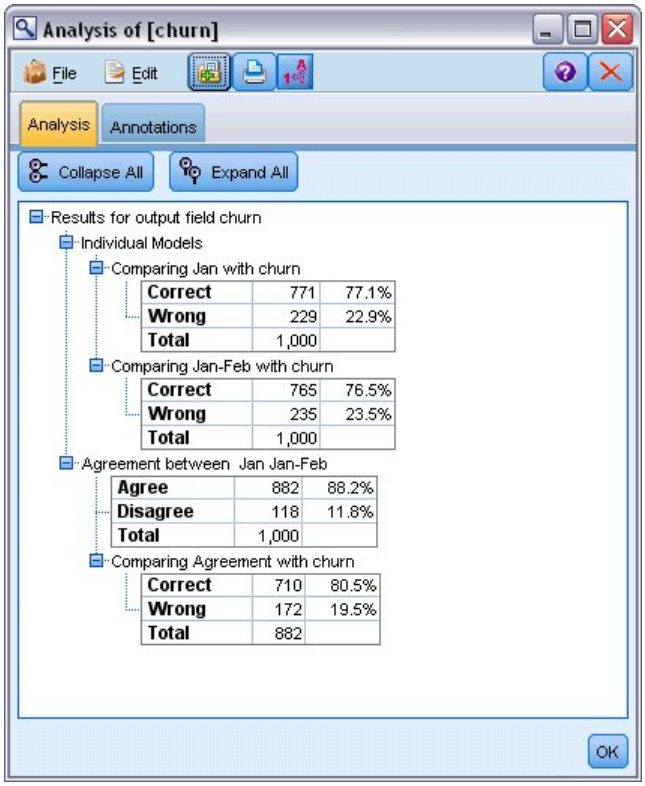


그림 *253.* 모델 정확도 분석

분석 노드에 대한 대안으로 평가 그래프를 사용하여 이익 차트를 작성함으로써 모델의 예측 정확도를 비교할 수 있습니다.

11. 평가 그래프 노드를 필터 노드에 연결하십시오.

기본 설정을 사용하여 그래프 노드를 실행하십시오.

분석 노드를 사용하면 그래프가 각 모델 유형이 유사한 결과를 생산함을 표시합니다. 그러나 두 월의 데이터를 모두 사용하는 재교육된 모델이 예측에서 더 높은 수준의 신뢰도를 가지므로 약간 더 나음을 알 수 있습니다.

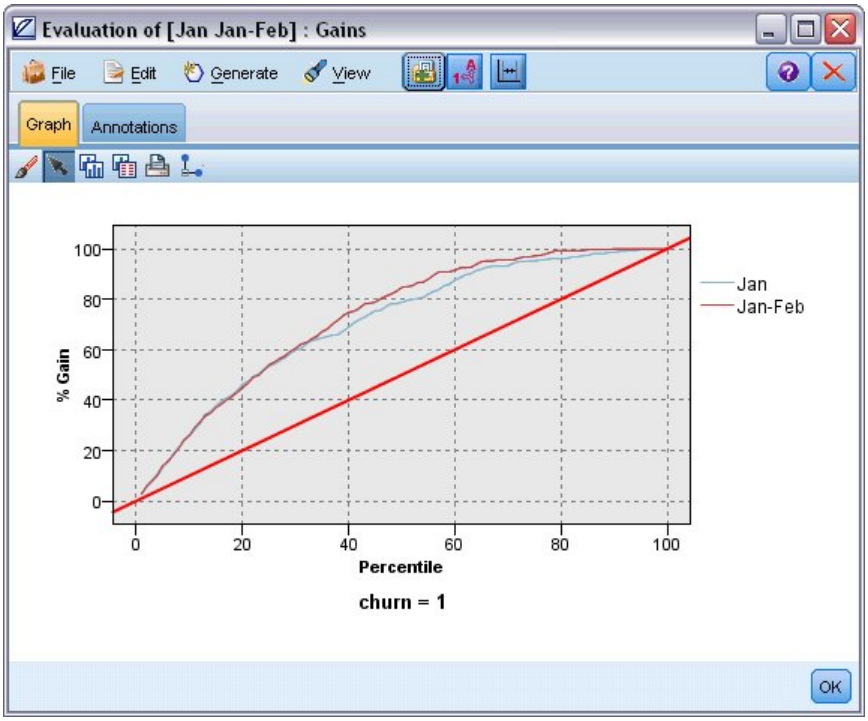


그림 *254.* 평가 모델 정확도

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 설치 디스크의 *Documentation* 디렉토리에서 사용 가능한 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

이러한 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 실세계에서 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

제 19 장 소매 판매 프로모션(신경망/C&RT)

이 예에서는 소매 제품 라인 및 판매에 대한 프로모션의 효과를 설명하는 데이터를 다룹니다. (이 데이터는 가상의 데이터입니다.) 이 예의 목적은 미래의 판매 프로모션의 효과를 예측하는 것입니다. 데이터 마이닝 프로세스는 상태 모니터링 예와 같이 탐색, 데이터 준비, 학습 및 검정 단계로 구성됩니다.

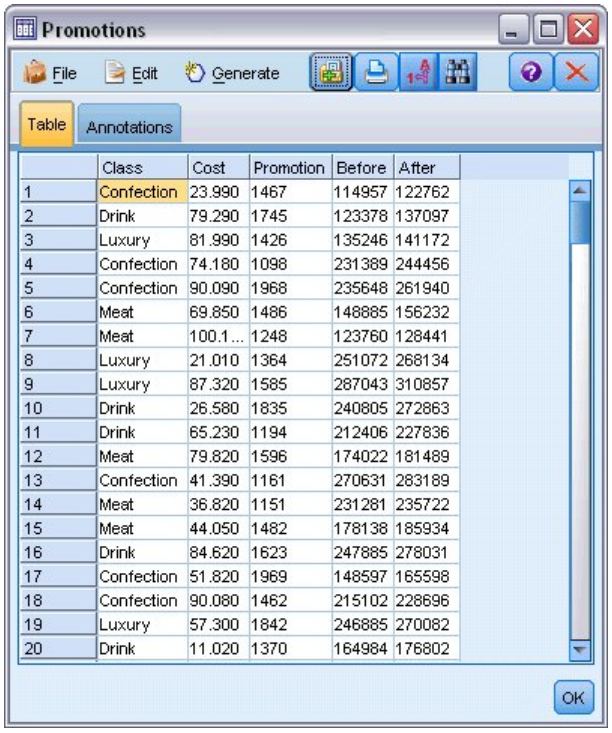
이 예에서는 *GOODS1n* 및 *GOODS2n*이라는 데이터 파일을 참조하는 *goodsplot.str* 및 *goodslearn.str* 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *goodsplot.str* 스트림은 *streams* 폴더에 있고 *goodslearn.str* 파일은 *streams* 디렉토리에 있습니다.

데이터 탐색

각 레코드에는 다음이 포함됩니다.

- *Class*. 제품 유형입니다.
- *Cost*. 단위 가격입니다.
- *Promotion*. 특정 프로모션에 소모된 금액 지수입니다.
- *Before*. 프로모션 전의 수입입니다.
- *After*. 프로모션 후의 수입입니다.

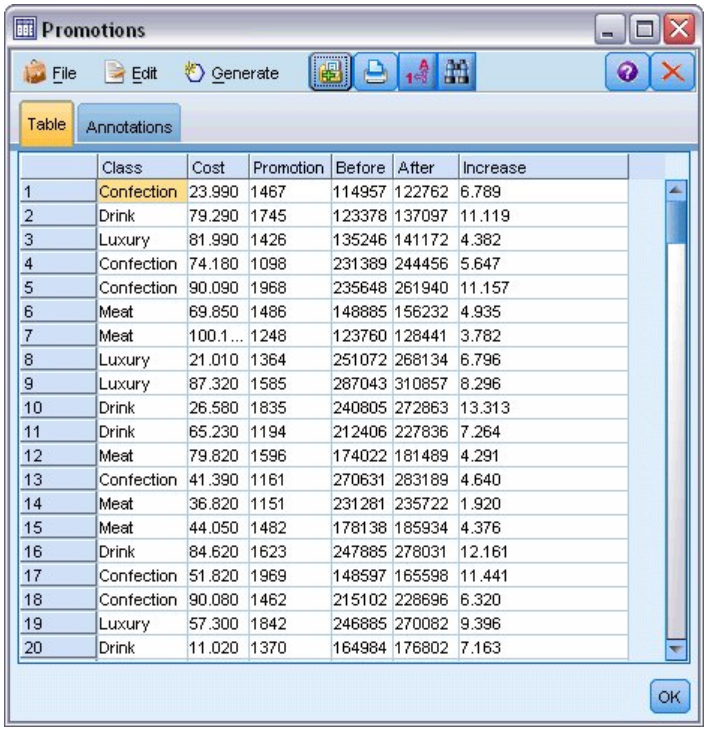
goodsplot.str 스트림에는 데이터를 테이블로 표시할 수 있는 단순 스트림이 포함됩니다. 두 개의 수입 필드(*Before* 및 *After*)가 절대항으로 표현됩니다. 단, 프로모션 후 및 아마도 프로모션의 결과로 인한 수입의 증가가 더 유용한 그림일 수 있습니다.



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

그림 *255.* 제품 판매에 대한 프로모션의 효과

*goodsplot.str*에는 *Increase* 필드에서 프로모션 이전 수입의 퍼센트로 표현되는 이 값을 파생시키고 이 필드를 나타내는 테이블을 표시하기 위한 노드가 포함됩니다.



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

그림 *256.* 프로모션 이후의 수입 증가량

또한 이 스트림은 증가량 히스토그램 및 관련된 제품 범주로 오버레이된 프로모션 지출 비용에 대한 증가량의 산점도를 표시합니다.

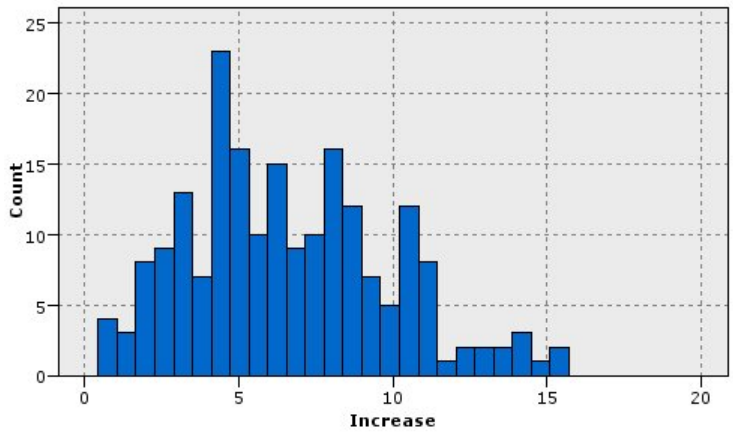


그림 *257.* 수입 증가량의 히스토그램

산점도는 제품의 각 클래스에 대해 수입 증가량 및 프로모션 비용 간에 거의 선형 관계가 있음을 표시합니다. 따라서 의사결정 트리 또는 신경망이 상당한 정확도로 사용 가능한 기타 필드에서 수입이 증가함을 예측할 수 있습니다.

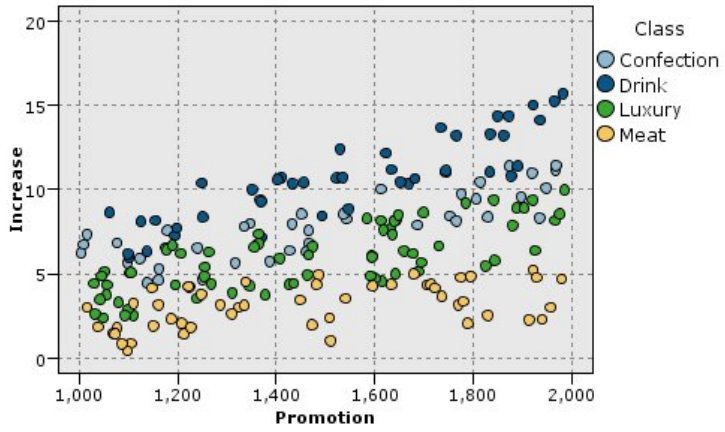


그림 *258.* 수입 증가량 대 프로모션 지출

학습 및 검정

goodslearn.str은 신경망 및 의사결정 트리를 학습하여 수입 증가량을 예측합니다.

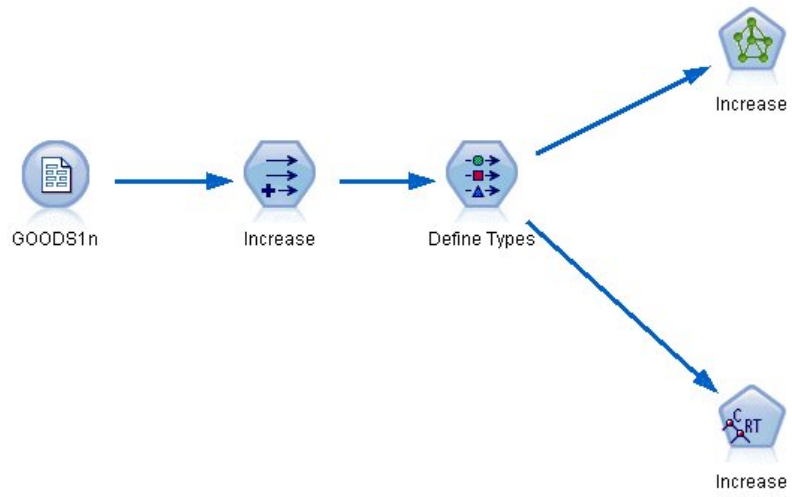


그림 *259.* 모델링 스트림 *goodslearn.str*

일단 모델 노드를 실행하여 실제 모델을 생성한 후에는 학습 프로세스의 결과를 검정할 수 있습니다. 유형 노드 및 새 분석 노드 사이의 의사결정 트리 및 네트워크에 연속적으로 연결하고 입력(데이터) 파일을 GOODS2n으로 변경하고 분석 노드를 실행하여 이를 수행할 수 있습니다. 이 노드의 출력으로부터, 특히 예측된 증가량과 올바른 답 사이의 선형 상관관계로부터 학습된 시스템이 높은 수준의 정확도로 수입의 증가량을 예측함을 알 수 있습니다.

추가 탐색에서는 학습된 시스템이 상대적으로 많은 오류를 만드는 경우에 초점을 맞출 것입니다. 이는 실제 증가에 대해 수입의 예측 증가량을 도표로 작성하여 식별할 수 있습니다. 이 그래프의 이상치는

SPSS Modeler 내의 대화형 그래픽을 사용하여 그 특성에서 선택할 수 있으며 정확도를 개선하기 위해 데이터 설명 또는 학습 프로세스를 조정하는 것도 가능합니다.

제 20 장 상태 모니터링(신경망/C5.0)

이 예에서는 시스템의 상태 정보 및 인식 및 예측 결함 상태를 모니터링하는 데 중점을 둡니다. 데이터는 가상의 시뮬레이션에서 작성되고 시간 경과에 따라 측정된 수많은 연결 계열로 구성됩니다. 각 레코드는 다음 측면에서의 시스템에 대한 스냅샷 보고서입니다.

- *Time*. 정수입니다.
- *Power*. 정수입니다.
- *Temperature*. 정수입니다.
- *Pressure*. 정상이면 0이고 순간 압력 경고이면 1입니다.
- *Uptime*. 마지막으로 이용된 이후의 시간입니다.
- *Status*. 정상적으로 0이며 오류에 대한 오류 코드(101, 202 또는 303)로 변경됩니다.
- *Outcome*. 이 시계열에 표시되는 오류 코드입니다. 오류가 발생하지 않으면 0입니다. (이러한 코드는 발생 이후에 판단에 도움을 줄 뿐입니다.)

이 예에서는 *COND1n* 및 *COND2n*이라는 데이터 파일을 참조하는 *condplot.str* 및 *condlearn.str* 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *condplot.str* 및 *condlearn.str* 파일은 *streams* 디렉토리에 있습니다.

다음 표에서 보듯이 각 시계열에 대해 정상 작동 기간의 레코드 계열 및 결함이 발생하는 기간의 레코드가 있습니다.

시간	전력	온도	압력	가동 시간	상태	결과
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
90	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0

시간	전력	온도	압력	가동 시간	상태	결과
53	938	251	0	209	0	101
54	936	251	0	209	0	101
			...			
208	644	251	0	209	0	101
209	640	251	0	209	101	101

다음 프로세스는 대부분의 데이터 마이닝 프로젝트에 공통적입니다.

- 관심 있는 상태의 예측 또는 인식과 연관되었을 가능성이 있는 속성을 판별하기 위해 데이터를 탐색합니다.
- 해당 속성을 보유하거나(이미 있는 경우) 이를 파생시켜 데이터에 추가합니다(필요한 경우에 한함).
- 결과 데이터를 사용하여 규칙 및 신경망을 학습합니다.
- 독립 검정 데이터를 사용하여 학습한 시스템을 검정합니다.

데이터 탐색

condplot.str 파일은 프로세스의 처음 부분을 설명합니다. 이는 수많은 그래프 도표를 작성하는 스트림을 포함합니다. 온도 또는 전력의 시계열이 시작적 패턴을 포함하는 경우, 임박한 오류 조건을 차별화하고 발생을 예측할 수도 있습니다. 온도 및 전원 둘 다에 대해 아래 스트림이 세 개의 서로 다른 오류 코드와 연관된 시계열을 별도의 그래프에 도표로 작성하여 여섯 개의 그래프가 생성됩니다. 선택 노드는 서로 다른 오류 코드와 연관된 데이터를 분리합니다.

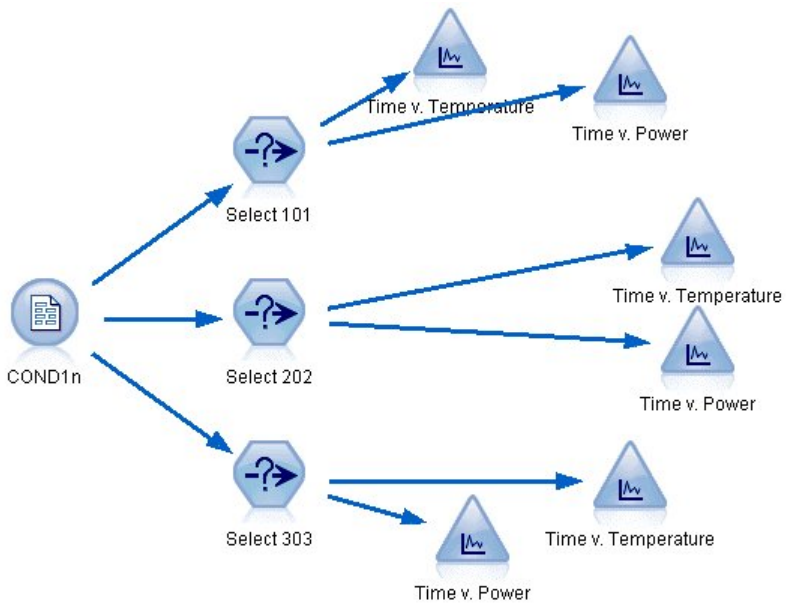


그림 *260. Condplot* 스트림

이 스트림의 결과는 아래 그림에서 표시됩니다.

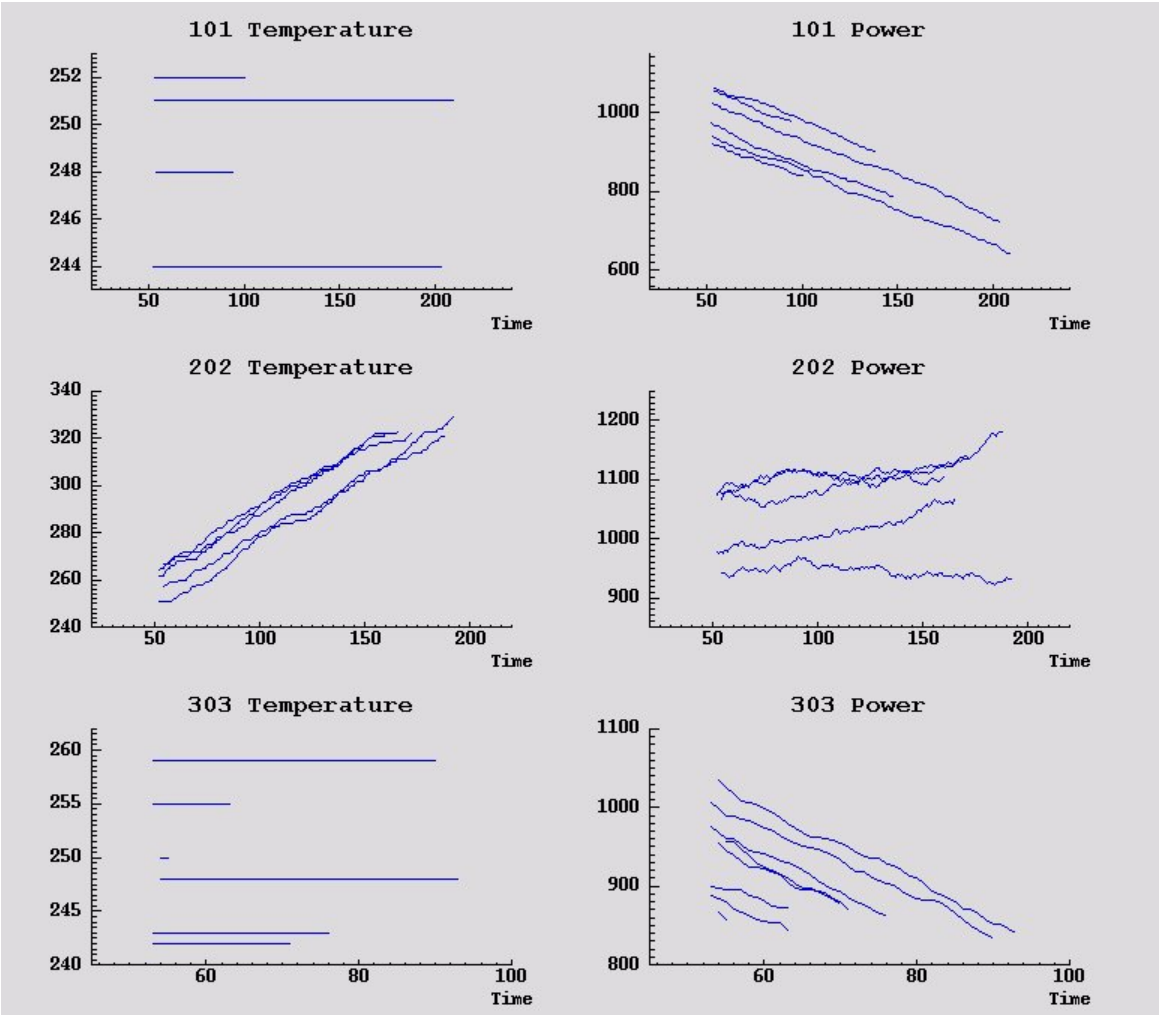


그림 *261.* 시간 경과에 따른 온도 및 전력

이 그래프에서 202 오류는 101 및 303 오류와 명확히 구별됩니다. 202 오류는 시간 경과에 따라 높아지는 온도 및 전력 변동을 표시합니다. 그러나 101과 303 오류를 구별하는 패턴은 덜 명확합니다. 두 오류 모두 균일한 온도 및 전력 하락을 표시하나 전력 하락이 303 오류에서 좀 더 가파른 것으로 보입니다.

이러한 그래프를 기준으로 할 때 온도 및 전력 둘 다에 대해 변화의 존재 여부 및 비율이 표시되며 변동의 존재 및 정도가 결함 예측 및 구별과 관련된 것으로 보입니다. 따라서 이러한 속성은 학습 시스템을 적용하기 전에 데이터에 추가되어야 합니다.

데이터 준비

데이터 탐색 결과를 기준으로 하여 *condlearn.str* 스트림이 관련 데이터를 파생시키고 결함을 예측하는 방법을 학습합니다.

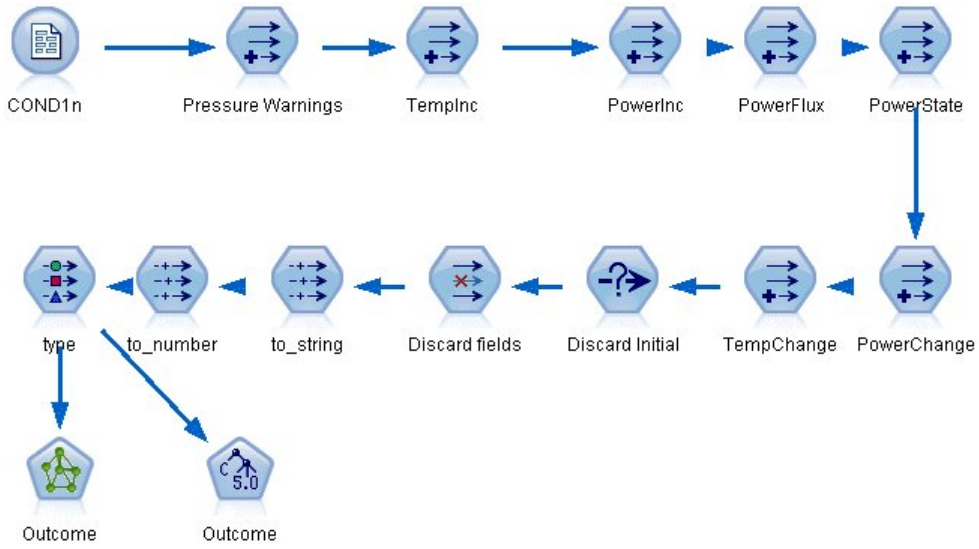


그림 *262. Condlearn* 스트림

이 스트림은 수많은 파생 노드를 사용하여 모델링에 사용할 데이터를 준비합니다.

- **가변파일 노드**. 데이터 파일 *COND1n*을 읽습니다.
- **Derive Pressure Warnings**. 순간적인 압력 경고의 수를 계수합니다. 횟수가 0으로 돌아가는 시기를 재설정합니다.
- **Derive TempInc**. `@DIFF1`을 사용하여 순간적인 온도 변화의 비율을 계산합니다.
- **Derive PowerInc**. `@DIFF1`을 사용하여 순간적인 전력 변화의 비율을 계산합니다.
- **Derive PowerFlux**. 플래그이며 전원이 이 레코드(최대 전원 또는 마지막 전원)에서 마지막 레코드에서와 반대 방향으로 변경된 경우, 참입니다.
- **Derive PowerState**. 안정으로 시작하여 두 개의 연속적인 전력속이 발견될 때 변동으로 변환되는 상태입니다. 5회 간격 동안 전력속이 없거나 횟수가 재설정될 때만 다시 안정으로 전환됩니다.
- **PowerChange**. 마지막 5회 간격에 걸친 *PowerInc*의 평균입니다.
- **TempChange**. 마지막 5회 간격에 걸친 *TempInc*의 평균입니다.
- **Discard Initial(select)**. 경계에서 *Power* 및 *Temperature*의 지나친(올바르지 않은) 점프를 피하기 위해 각 시계열이 처음 레코드를 삭제합니다.
- **Discard fields**. *Uptime*, *Status*, *Outcome*, *Pressure Warnings*, *PowerState*, *PowerChange* 및 *TempChange*에 대한 레코드를 잘라냅니다.
- **유형**. *Outcome*의 역할을 **대상**(예측할 필드)으로 정의합니다. 또한 *Outcome*의 측정 수준을 **명목형**으로, *Pressure Warnings*의 측정 수준을 **연속형**으로 *PowerState*의 측정 수준을 **플래그**로 정의합니다.

학습

*condlearn.str*에서 스트림을 실행하면 C5.0 규칙 및 신경망(넷)을 학습할 수 있습니다. 네트워크를 학습하는 데는 시간이 좀 걸릴 수 있으나 타당한 결과를 생성하는 넷을 저장하기 위해 학습이 일찍 중단될 수 있습니다. 일단 학습이 완료되면 관리자 창의 오른쪽 상단에 있는 모델 탭이 반짝거려서 두 개의 새 너깃이 작성되었음을 알립니다. 하나는 신경망을 나타내고 다른 하나는 규칙을 나타냅니다.

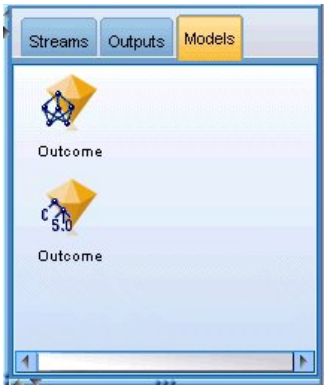


그림 *263.* 모델 너깃이 포함된 모델 관리자

시스템을 검증하거나 모델의 결과를 내보낼 수 있도록 모델 너깃도 기존 스트림에 추가됩니다. 이 예에서는 모델의 결과를 검증할 것입니다.

검정

모델 너깃이 스트림에 추가되고 둘 다 유형 노드에 연결됩니다.

1. 유형 노드가 C5.0 너깃에 연결되는 신경망 너깃에 연결되도록 표시된 대로 너깃의 위치를 다시 지정하십시오.
2. 분석 노드를 C5.0 너깃에 연결하십시오.
3. *COND2n*은 표시되지 않은 검정 데이터를 포함하므로 *COND1n* 대신 *COND2n* 파일을 읽도록 원래 소스 노드를 편집하십시오.

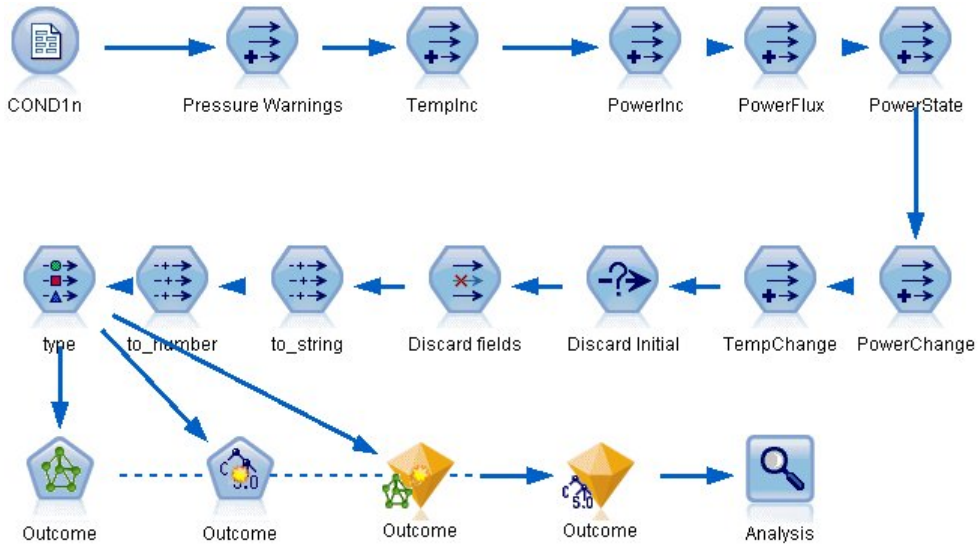


그림 *264.* 학습된 네트워크 검정

4. 분석 노드를 열고 실행을 클릭하십시오.

그러면 학습된 네트워크 및 규칙의 정확도를 반영하는 그림이 생성됩니다.

제 21 장 통신 고객 분류(판별 분석)

판별 분석은 입력 필드 값을 기반으로 레코드를 분류하는 통계 기법입니다. 이는 선형 회귀와 비슷하지만, 숫자 대신 범주형 대상 필드를 사용합니다.

예를 들어, 통신 제공업체가 서비스 사용 패턴을 기준으로 고객층을 세그먼트화하여 고객을 4개의 그룹으로 범주화한다고 가정합니다. 소속그룹을 예측하기 위해 인구 통계학적 데이터를 사용하면 개별 잠재 고객에 대한 제공을 사용자 정의할 수 있습니다.

이 예에서는 *telco.sav*라는 데이터 파일을 참조하는 *telco_custcat_discriminant.str*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다.
telco_custcat_discriminant.str 파일은 *streams* 디렉토리에 있습니다.

이 예에서는 사용 패턴을 예측하기 위해 인구 통계학적 데이터를 사용하는 데 초점을 맞춥니다. 대상 필드 *custcat*에는 다음과 같이 네 개의 고객 그룹에 해당하는 네 개의 가능한 값이 있습니다.

값	레이블
1	기본 서비스
2	E-서비스
3	플러스 서비스
4	전체 서비스

스트림 작성

1. 먼저, 출력에 변수 및 값 레이블을 표시하도록 스트림 특성을 설정하십시오. 메뉴에서 다음을 선택하십시오.

   파일 > **스트림 특성...** > **옵션** > **일반**
2. **출력에 필드와 값 레이블 표시**가 선택되어 있는지 확인하고 **확인**을 클릭하십시오.

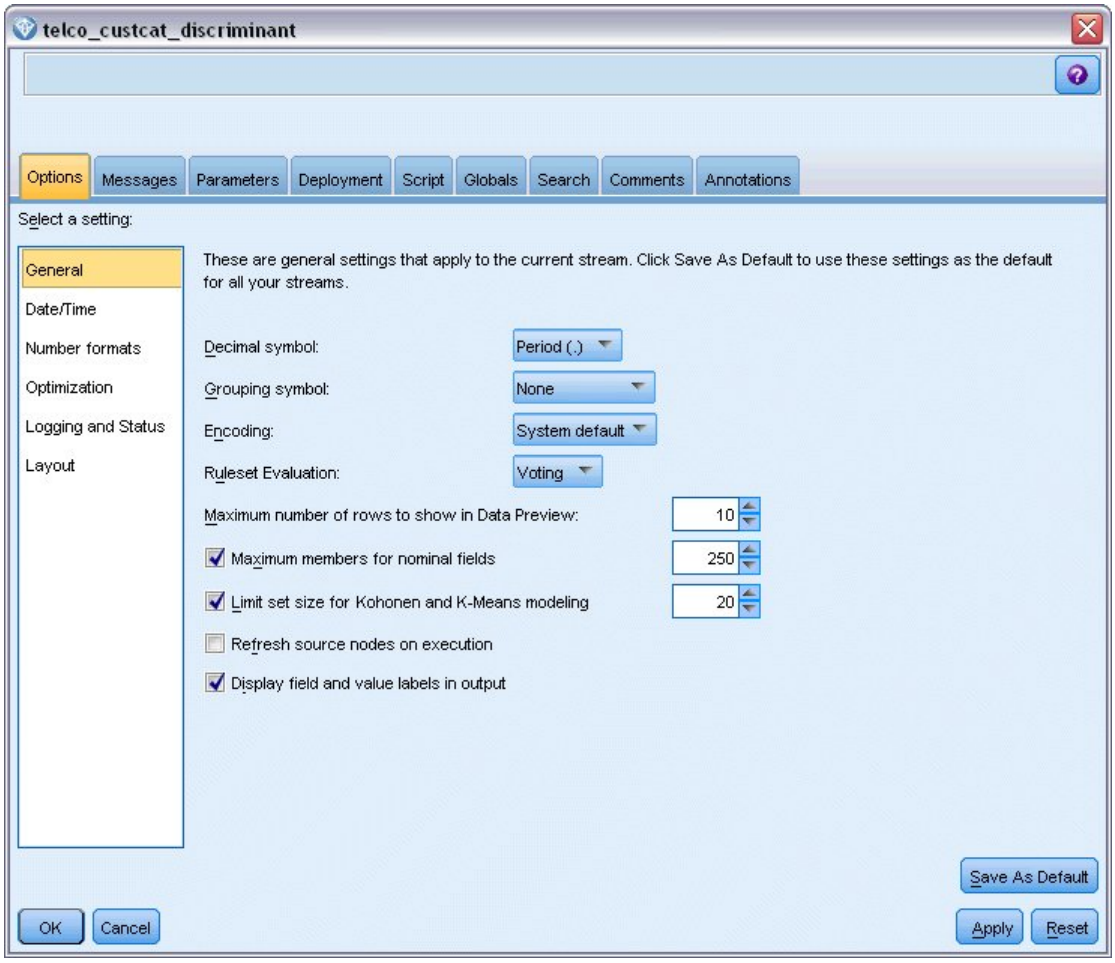


그림 *265.* 스트림 특성

3. *Demos* 폴더에서 *telco.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

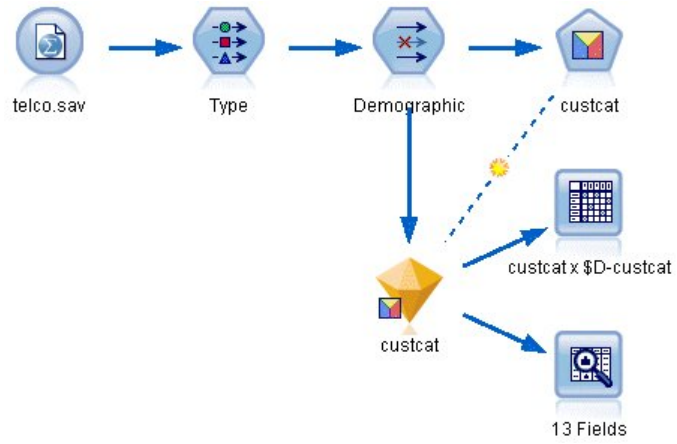


그림 *266.* 판별 분석을 사용하여 고객을 분류하기 위한 샘플 스트림

a. 유형 노드를 추가하고 모든 측정 수준이 올바르게 설정되었는지 확인하고 **값 읽기**를 클릭하십시오. 예를 들어, 값이 0 및 1인 대부분의 필드는 플래그로 간주할 수 있습니다.

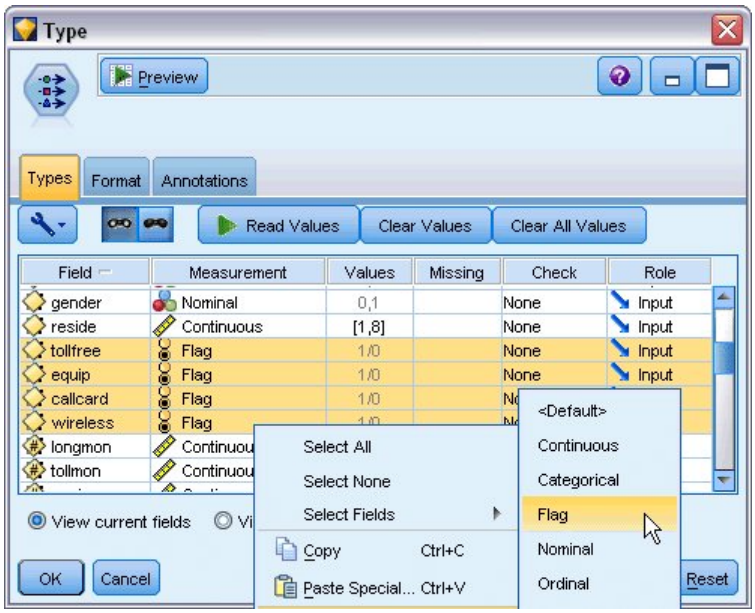


그림 *267.* 다중 필드에 대한 측정 수준 설정

팁: 유사한 값(0/1 등)을 가진 다중 필드에 대한 특성을 변경하려면 값 열 헤더를 클릭하여 필드를 값 기준으로 정렬한 다음 Shift 키를 누른 상태에서 마우스 또는 화살표를 사용하여 변경할 키를 모두 선택하십시오. 그런 다음 마우스 오른쪽 단추로 선택영역을 클릭하여 선택된 필드의 측정 수준 또는 기타 속성을 변경하십시오.

성별은 플래그 대신 두 개의 값 변수군이 있는 필드로 간주하는 것이 더 정확하므로 해당 측정 값을 **명목형**으로 두십시오.

b. 통신사용등급 필드에 대한 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.

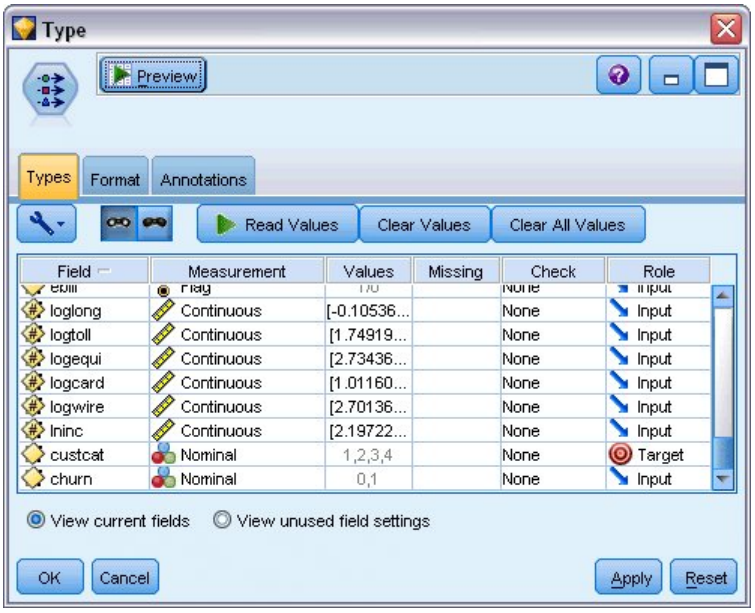


그림 *268.* 필드 역할 설정

이 예에서는 인구 통계에 초점을 맞추므로 관련 필드(지역, 연령, 혼인 여부, 주소, 수입, 교육수준, 고용, 은퇴, 성별, 거주 및 통신사용등급)만 포함하도록 필터 노드를 사용하십시오. 기타 필드는 이 분석 목적으로는 제외될 수 있습니다.

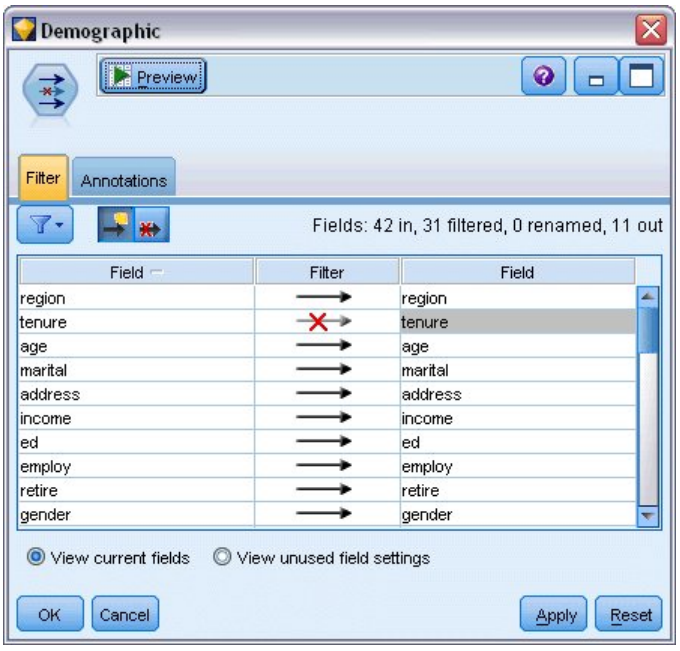


그림 *269.* 인구 통계학 필드 필터링

(또는 이러한 필드를 제외하지 않고 해당 필드에 대한 역할을 **없음**으로 지정하거나 모델링 노드에 사용할 필드를 선택할 수도 있습니다.)

4. 판별 노드에서 모델 탭을 클릭하여 **단계선택법**을 선택하십시오.

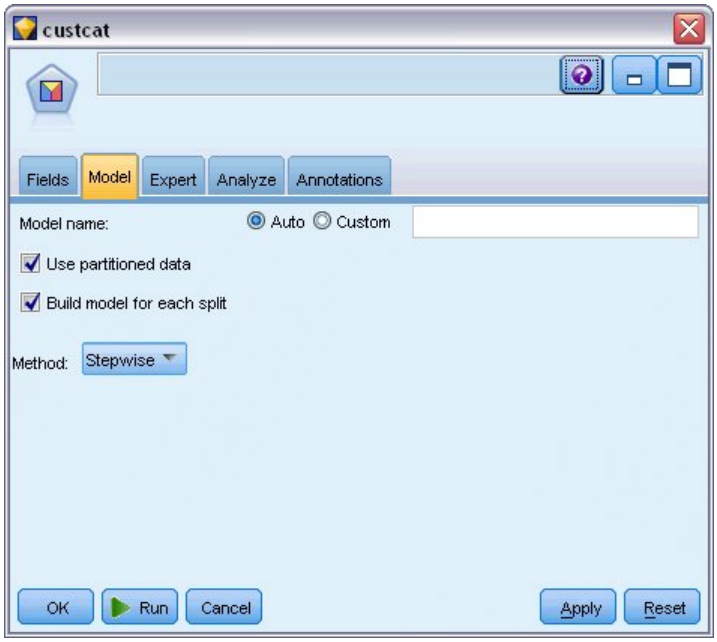


그림 *270.* 모델 옵션 선택

5. 전문가 탭에서 모델을 **전문가**로 설정하고 **출력**을 클릭하십시오.
6. 고급 출력 대화 상자에서 **요약표**, **영역도** 및 **단계 요약**을 선택한 다음 **확인**을 클릭하십시오.

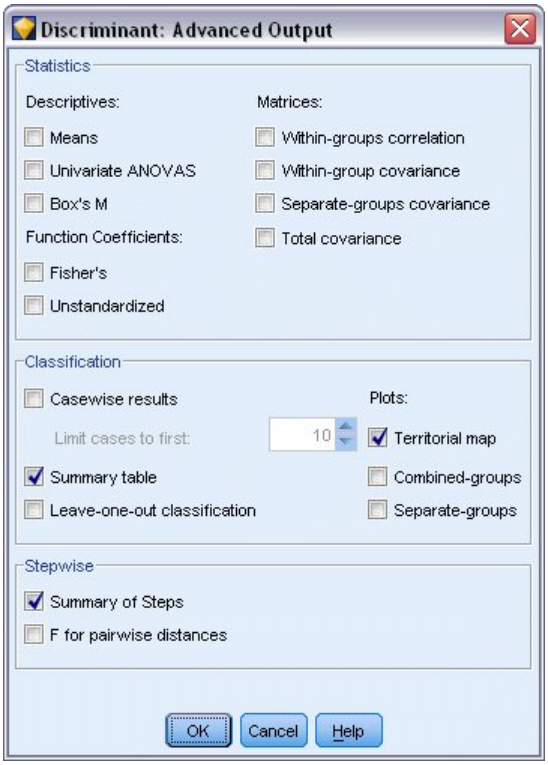


그림 *271.* 출력 옵션 선택

모델 탐색

1. **실행**을 클릭하여 모델을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에 추가됩니다. 세부사항을 보려면 스트림에서 모델 너깃을 두 번 클릭하십시오.

   요약 탭은 (다른 사항 사이에서) 대상을 표시하고 고려하도록 제출된 입력(예측변수 필드)의 전체 목록을 표시합니다.

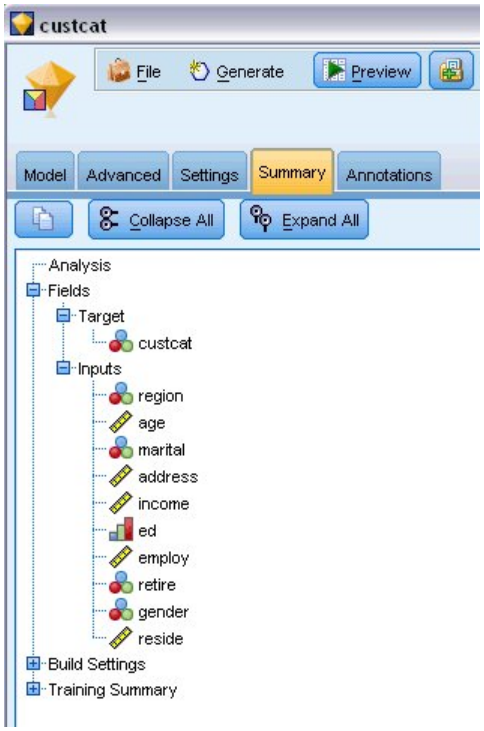


그림 *272.* 대상 및 입력 필드를 표시하는 모델 요약

판별 분석의 결과에 대한 세부사항:

2. 고급 탭을 클릭하십시오.
3. "외부 브라우저에서 실행" 단추(모델 탭 바로 아래)를 클릭하여 웹 브라우저에서 결과를 보십시오.

통신회사 고객을 분류하기 위해 판별 분석을 사용하여 결과 분석

단계 선택 판별 분석

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	Age in years	1.000	1.000	7.521	.978
	Marital status	1.000	1.000	3.500	.990
	Years at current address	1.000	1.000	8.433	.975
	Household income in thousands	1.000	1.000	6.689	.980
	Level of education	1.000	1.000	61.454	.844
	Retired	1.000	1.000	3.005	.991
	Years with current employer	1.000	1.000	16.976	.951
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988

그림 *273.* 분석에 사용되지 않는 변수, *0*단계

예측변수가 많은 경우에 모델에 사용할 "최선"의 변수를 자동으로 선택하는 단계적 방법이 유용할 수 있습니다. 이 단계적 방법은 어떠한 예측변수도 포함하지 않는 모델부터 시작합니다. 각 단계에서 진입

기준(기본적으로 3.84)를 초과하는 가장 큰 입력에 대한 *F* 값을 가진 예측변수가 모델에 추가됩니다.

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
3	Age in years	.535	.535	.252	.795
	Marital status	.605	.593	1.507	.792
	Years at current address	.776	.771	3.514	.787
	Household income in thousands	.688	.657	.687	.794
	Retired	.917	.880	.353	.795
	Gender	.997	.931	.395	.795

그림 *274.* 분석에 사용되지 않는 변수, *3*단계

마지막 단계에서 분석에서 제외된 모든 변수가 3.84보다 작은 입력에 대한 *F* 값을 갖게 되면 더 이상 추가되지 않습니다.

Step		Tolerance	F to Remove	Wilks' Lambda
1	Level of education	1.000	61.454	
2	Level of education	.953	59.108	.951
	Years with current employer	.953	14.933	.844
3	Level of education	.951	60.046	.940
	Years with current employer	.934	15.824	.834
	Number of people in household	.979	4.841	.807

그림 *275.* 분석에 사용되는 변수

이 테이블에는 각 단계마다 분석에 사용되는 변수에 대한 통계량을 표시합니다. 공차는 방정식에서 다른 독립변수에 의해 설명되지 않는 변수 분산의 비율입니다. 허용 오차가 매우 낮은 변수는 모델에 별 정보를 제공하지 못하고 계산에 문제를 일으킬 수 있습니다.

제거에 대한 *F* 값은 변수가 현재 모델에서 제거될 경우(다른 변수가 있는 것으로 가정) 발생되는 사항을 설명하는 데 유용합니다. 변수 입력에 대한 제거에 대한 *F*는 이전 단계에서의 (분석에 사용되지 않는 변수 테이블에서 표시된) 입력에 대한 *F*와 동일합니다.

단계적 방법에 대한 참고 사항

단계적 방법은 편리하지만 제한 사항이 있습니다. 단계적 방법에서는 통계 장점만을 기준으로 모델을 선택하므로 실제 유의성이 없는 예측변수가 선택될 수도 있습니다. 해당 데이터에 어느 정도 경험이 있고 어느 예측변수가 중요한지 예상할 수 있는 경우에는 이 지식을 사용하고 단계적 방법을 피해야 합니다. 그러나 예측변수가 많고 어디서 시작해야 하는지 모르는 경우에는 단계 선택 분석을 실행하고 선택된 모델을 조정하는 것이 전혀 아무 모델이 없는 것보다 나을 수 있습니다.

모델 적합 확인

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198	80.2	80.2	.407
2	.048	19.4	99.6	.214
3	.001	.4	100.0	.031

그림 *276.* 고유값

모델에서 설명되는 거의 모든 분산이 처음 두 개의 함수에 의해 설명됩니다. 세 가지 함수는 자동으로 맞춰지나 극소의 고유값 때문에 세 번째는 무시할 수 있습니다.

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	.000
2 through 3	.953	47.486	4	.000
3	.999	.929	1	.335

그림 *277. Wilks*의 람다

Wilks의 람다는 처음 두 함수만 유의미하다는 데 동의합니다. 이는 각 함수 집합에 대해 나열된 함수의 평균이 전체 그룹에 걸쳐 동일하다는 가설을 검정합니다. 함수 3의 검정은 유의수준이 0.10보다 크므로 이 함수는 모델에 거의 기여하지 않습니다.

구조행렬

	Function		
	1	2	3
Level of education	.966*	-.090	-.244
Years with current employer	-.182	.964*	-.193
Age in years[a]	-.162	.598*	-.285
Household income in thousands[a]	.109	.514*	-.190
Years at current address[a]	-.151	.394*	-.214
Retired[a]	-.108	.230*	-.137
Gender[a]	.008	.054*	.009
Number of people in household	.232	.097	.968*
Marital status[a]	.132	.134	.600*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

그림 *278.* 구조행렬

둘 이상의 판별 함수가 있는 경우, 별표(*)는 각 변수의 정준 함수와의 가장 큰 절대 상관관계를 표시합니다. 각 함수 내에서 표시된 이러한 변수는 상관관계 크기에 의해 정렬됩니다.

- *Level of education*은 첫 번째 함수와 가장 강한 상관관계가 있으며 이 함수와 가장 강한 상관관계가 있는 유일한 변수입니다.
- *Years with current employer*, *Age in years*, *Household income in thousands*, *Years at current address*, *Retired* 및 *Gender*는 두 번째 함수와 가장 강한 상관관계가 있습니다. 단, *Gender* 및 *Retired*는 다른 변수에 비해 약한 상관관계가 있습니다. 다른 변수는 이 함수를 "안정성"으로 표시합니다.
- *Number of people in household* 및 *Marital status*는 세 번째 판별 함수와 가장 강한 상관관계가 있으나 해당 함수가 쓸모가 없는 함수이므로 이러한 변수도 거의 쓸모가 없는 예측변수입니다.

영역도

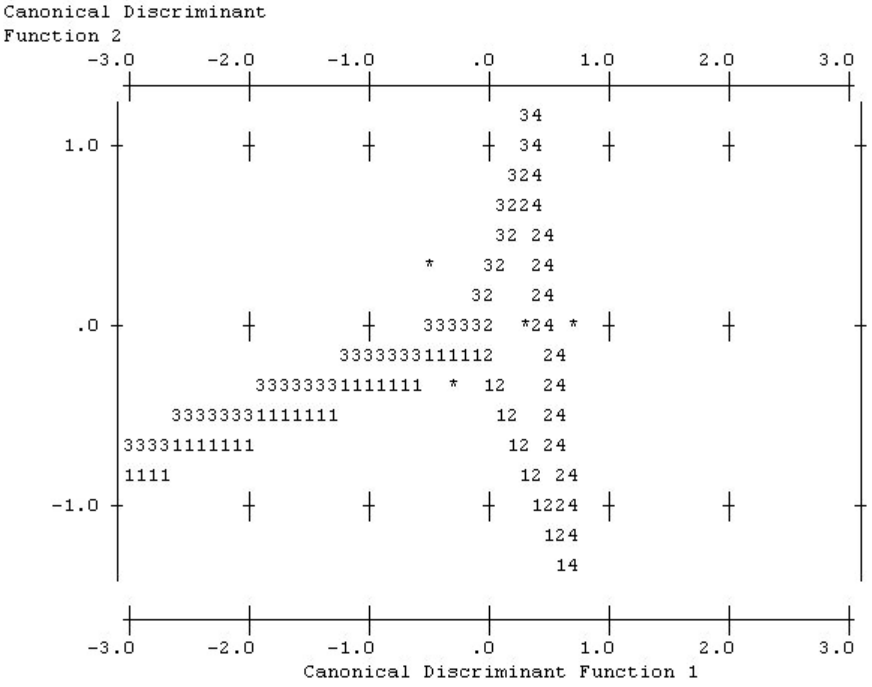


그림 *279.* 영역도

영역도를 사용하면 집단 및 판별 함수 사이의 관계를 파악하는 데 도움이 됩니다. 이러한 영역도는 구조행렬 결과와 결합하여 예측변수 및 집단 사이의 관계에 대한 그래픽 방식의 해석을 제공합니다. 수평축에 표시되는 첫 번째 함수는 집단 4(*Total service* 고객)를 다른 집단과 구분합니다. *Level of education*이 첫 번째 함수와 양의 방향으로 강한 상관관계가 있으므로 이는 *Total service* 고객이 일반적으로 교육 수준이 가장 높은 고객임을 제안합니다. 두 번째 함수는 집단 1 및 3(기본 서비스 및 플러스 서비스 고객)을 구분합니다. 플러스 서비스 고객은 더 오래 일하는 경향이 있고 기본 서비스 고객보다 나이가 많은 경향이 있습니다. 맵에서는 *E*-서비스 고객이 중간 정도의 직업 경험이 있는 잘 교육받은 고객인 경향이 있음을 나타내지만 다른 고객과 잘 구분되지 않습니다.

일반적으로 별표(*)로 표시된 집단 중심값의 영역 선에 대한 근접성은 모든 그룹 간의 구분이 아주 강하지 않음을 나타냅니다.

처음 두 판별 함수만 도표로 작성되고 세 번째 함수는 다소 덜 유의미한 것으로 발견되나 영역도에서는 판별 모델에 대한 포괄적인 보기를 제안합니다.

분류 결과

			Predicted Group Membership				
		Customer category	Basic service	E-service	Plus service	Total service	Total
Original	Count	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47.0	4.1	22.9	25.9	100.0
		E-service	22.6	6.9	26.7	43.8	100.0
		Plus service	36.3	5.0	39.9	18.9	100.0
		Total service	16.9	6.8	15.7	60.6	100.0

a. 39.5% of original grouped cases correctly classified.

그림 *280.* 분류 결과

Wilks의 람다에서 모델이 추측보다는 잘 수행하고 있음을 알 수 있습니다. 그러나 얼마나 잘 수행하는지 판별하기 위해서는 분류 결과의 도움을 받아야 합니다. 주어진 관측 데이터에 대해 "널" 모델(즉, 예측변수가 없는 모델)은 모든 고객을 전형 그룹인 플러스 서비스로 분류합니다. 따라서 널 모델은 전체의 281/1000 = 28.1%가 됩니다. 사용자의 모델은 11.4% 추가 또는 고객의 39.5%를 갖게 됩니다. 특히 이 모델은 전체 서비스 고객을 식별할 때 탁월합니다. 그러나 *E*-서비스 고객을 분류하는 작업은 유난히 잘 수행하지 못합니다. 이러한 고객을 구분하기 위해 다른 예측변수를 찾아야 할 수 있습니다.

요약

각 고객의 인구 통계학적 정보를 기준으로 하여 고객을 사전정의된 네 가지 "서비스 사용법" 집단 중 하나에 분류하는 판별 모델을 작성했습니다. 구조행렬 및 영역도를 사용하여 고객 기준을 세그먼트화할 때 가장 유용한 변수를 식별했습니다. 마지막으로 분류 결과는 모델이 *E*-서비스 고객 분류에서 수행이 가장 저조함을 표시합니다. 해당 고객을 더 잘 분류하는 또 다른 예측변수를 판별하기 위해서는 더 많은 연구가 필요하나 예측할 대상에 따라 모델이 사용자의 요구에 완벽하게 적합할 수 있습니다. 예를 들어, *E*-서비스 고객을 식별하는 데 관심이 없는 경우, 해당 모델로 충분할 것입니다. 이는 E-서비스가 수익이 거의 없는 특가품인 경우일 수 있습니다. 예를 들어, 투자수익률(ROI)이 가장 높은 고객이 플러스 서비스 또는 전체 서비스 고객인 경우, 이 모델은 사용자가 원하는 정보를 제공할 수 있습니다.

이러한 결과가 학습 데이터에만 기반한다는 점을 참고하십시오. 모델이 기타 데이터에 얼마나 잘 일반화되었는지 평가하려면 파티션 노드를 사용하여 검정 및 검증 목적으로 레코드 서브세트를 유지할 수 있습니다.

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 IBM SPSS Modeler 알고리즘 안내서에 나와 있습니다. 이는 설치 디스크의 *\Documentation* 디렉토리에서 사용 가능합니다.

제 22 장 구간 중도절단 생존 데이터 분석(일반화 선형 모델)

구간 중도절단이 있는(관심 있는 이벤트의 정확한 시간이 알려지지 않았으나 지정된 구간 내에 발생했음만 알 수 있는) 경우, 생존 데이터를 분석하고 Cox 모델을 구간 내의 이벤트의 위험함수에 적용하면 결과적으로 보 로그-로그 회귀 모델이 됩니다.

궤양 재발을 방지하기 위한 두 가지 치료법의 효과를 비교하기 위해 계획된 연구의 편상관 정보는 *ulcer_recurrence.sav*에 수집됩니다. 이 데이터 세트는 다른 곳[1]에서 제시되고 분석됩니다. 일반화 선형 모델을 사용하면 보 로그-로그 회귀 모델에 대한 결과를 복제할 수 있습니다.

이 예에서는 *ulcer_recurrence.sav* 데이터 파일을 참조하는 *ulcer_genlin.str*이라는 스트림을 사용합니다. 데이터 파일은 *Demos* 폴더에 있으며 스트림 파일은 *streams* 서브폴더에 있습니다.

스트림 작성

1. *Demos* 폴더에서 *ulcer_recurrence.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

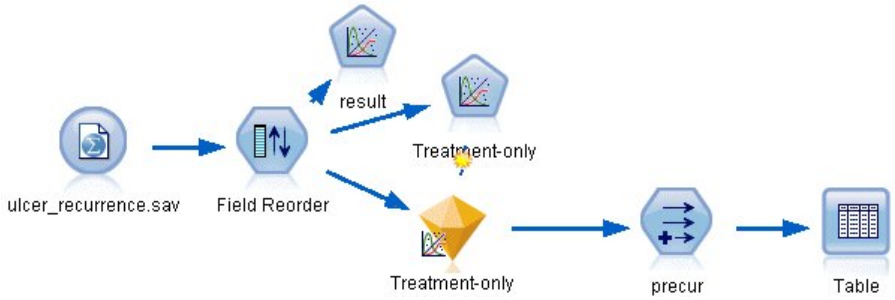


그림 *281.* 궤양 재발을 예측하기 위한 샘플 스트림

2. 소스 노드의 필터 탭에서 *id* 및 *time*을 제외하십시오.

1. Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

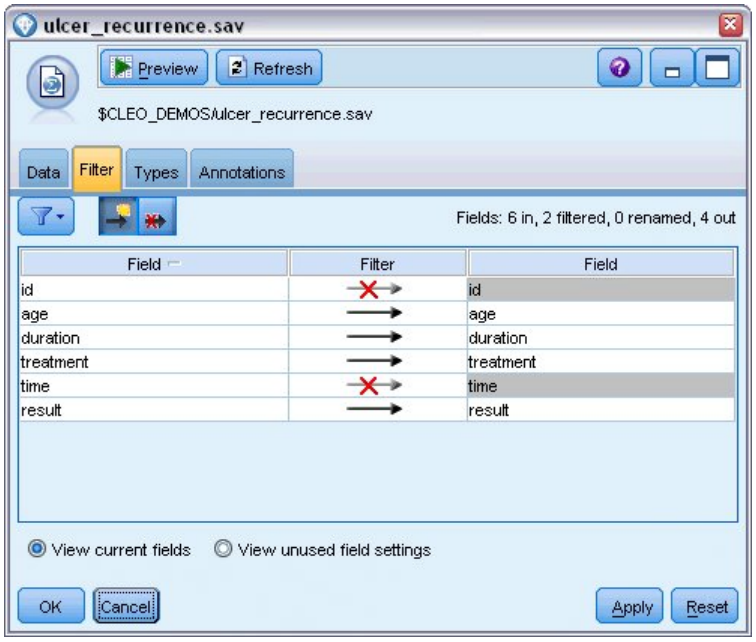


그림 *282.* 원하지 않는 필드 필터링

3. 소스 노드의 유형 탭에서 *result* 필드에 대한 역할을 **대상**으로 설정하고 측정 수준을 **플래그**로 설정하십시오. 1 결과는 궤양이 재발한 것을 나타냅니다. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.
4. **값 읽기**를 클릭하여 데이터를 인스턴스화하십시오.

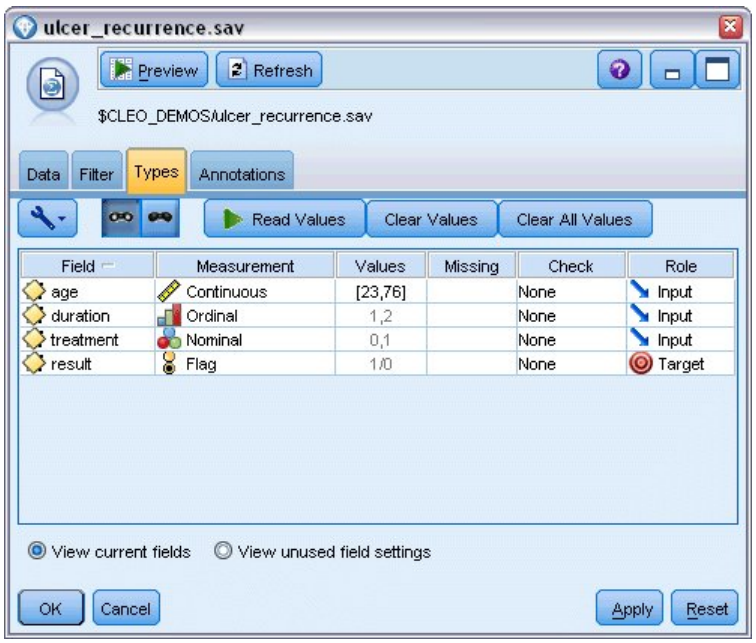


그림 *283.* 필드 역할 설정

5. 필드 다시 정렬 노드를 추가하고 입력 순서로 *duration*, *treatment* 및 *age*를 지정하십시오. 이는 필드가 모델에 입력되는 순서를 결정하고 Collett 결과를 복제하는 데 도움이 됩니다.

그림 *284.* 필드가 원하는 대로 모델에 입력되도록 필드 다시 정렬

6. GenLin 노드를 소스 노드에 연결하고 GenLin 노드에서 **모델** 탭을 클릭하십시오.
7. 대상에 대한 참조범주로 **처음(가장 낮은 값)**을 선택하십시오. 이는 두 번째 범주가 관심 있는 이벤트이며 모델에 대한 해당 효과가 모수 추정값의 해석에 사용됨을 표시합니다. 양수 계수를 가진 연속형 예측변수는 예측변수의 값이 증가하면 반복 확률이 증가함을 나타냅니다. 계수가 큰 명목 예측변수의 범주는 변수군의 기타 범주와 관련하여 반복 확률이 증가함을 표시합니다.

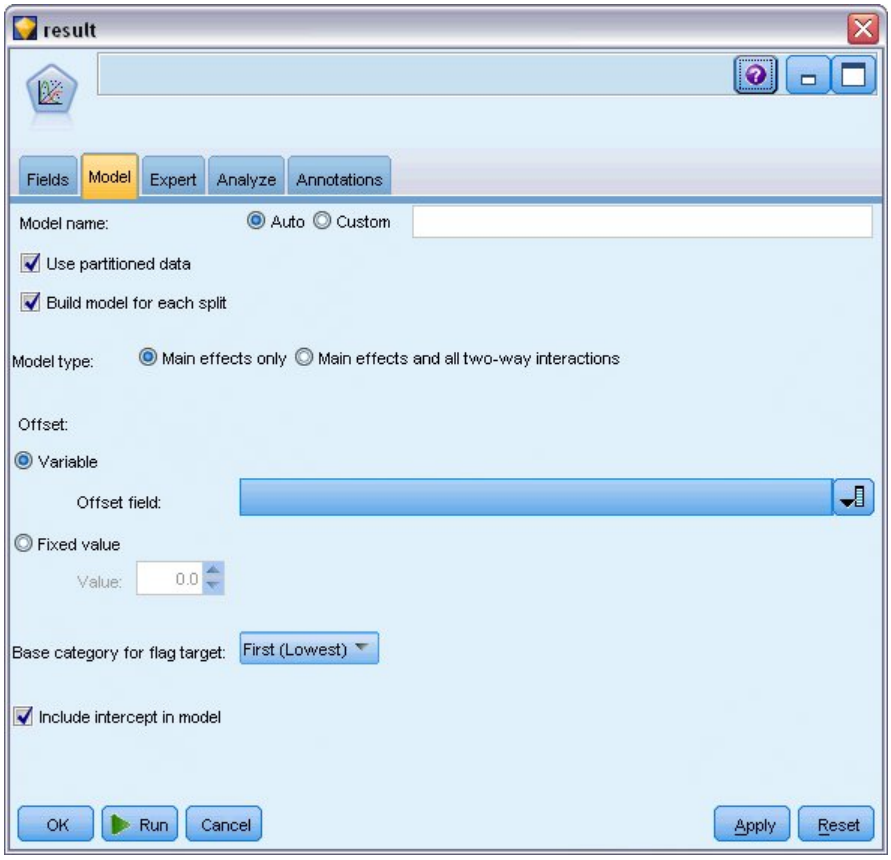


그림 *285.* 모델 옵션 선택

8. **전문가** 탭을 클릭하고 **전문가**를 선택하여 전문가 모델링 옵션을 활성화하십시오.
9. 분포로 **이항**을 선택하고 연결함수로 **보 로그-로그**를 선택하십시오.
10. 척도 모수를 추정하기 위한 방법으로 **고정값**을 선택하고 기본값인 1.0을 그대로 두십시오.
11. 요인에 대한 범주 순서로 **내림차순**을 선택하십시오. 이는 각 요인의 첫 번째 범주가 참조 범주가 되고 모델에 대한 이 선택의 효과가 모수 추정값 해석에 사용됨을 표시합니다.

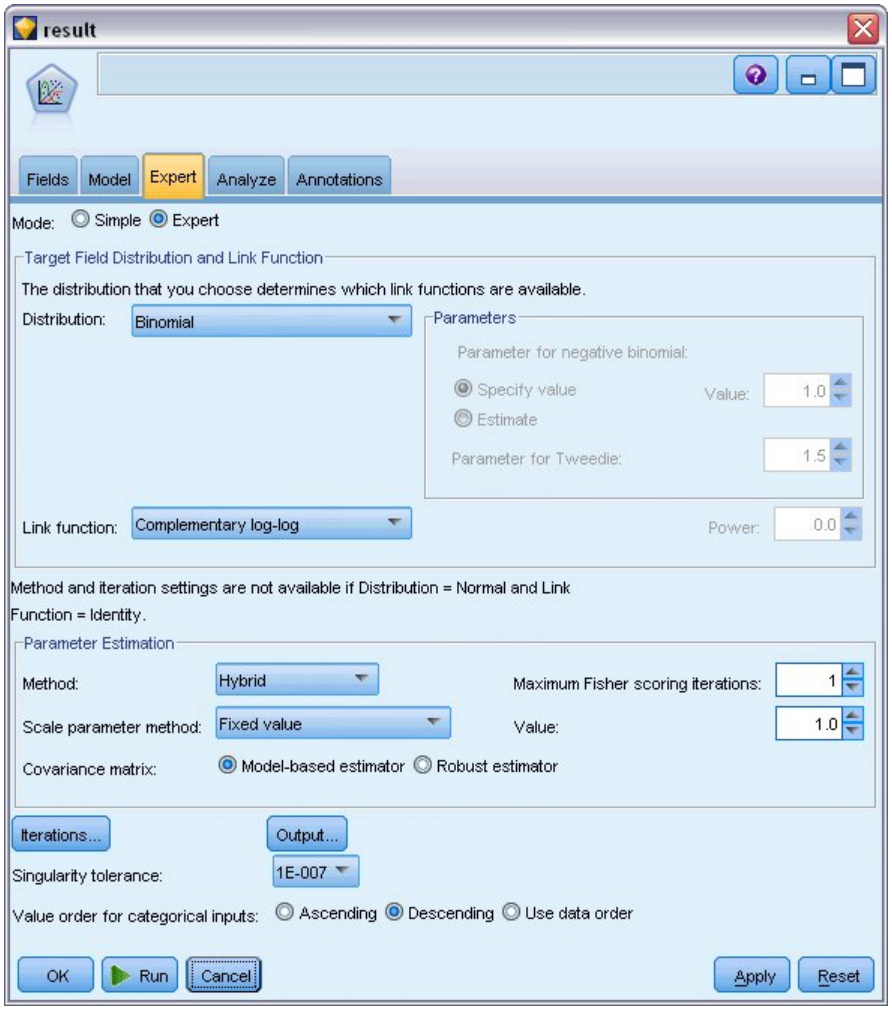


그림 *286.* 고급 옵션 선택

12. 스트림을 실행하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에도 추가됩니다. 모델 세부사항을 보려면 너깃을 마우스 오른쪽 단추로 클릭하고 **편집** 또는 **찾아보기**를 선택하십시오.

모델 효과 검정

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
duration	.003	1	.958
treatment	.382	1	.537
age	.358	1	.550

Dependent Variable: Result
Model: (Intercept), duration, treatment, age

그림 *287.* 주효과 모델에 대한 모델 효과 검정

어떠한 모델 효과도 통계적으로 유의적이지 않으나 치료법 효과에서 관측된 차이는 의학적으로 흥미가 있으므로 모델 항이 치료법만 있는 축소한 모델을 모델 항으로 적합화할 수 있습니다.

치료법 전용 모델 적합화

1. GenLin 노드의 필드 탭에서 **사용자 정의 설정**을 클릭하십시오.
2. 대상으로 *result*를 선택하십시오.
3. 단일 입력으로 *treatment*를 선택하십시오.

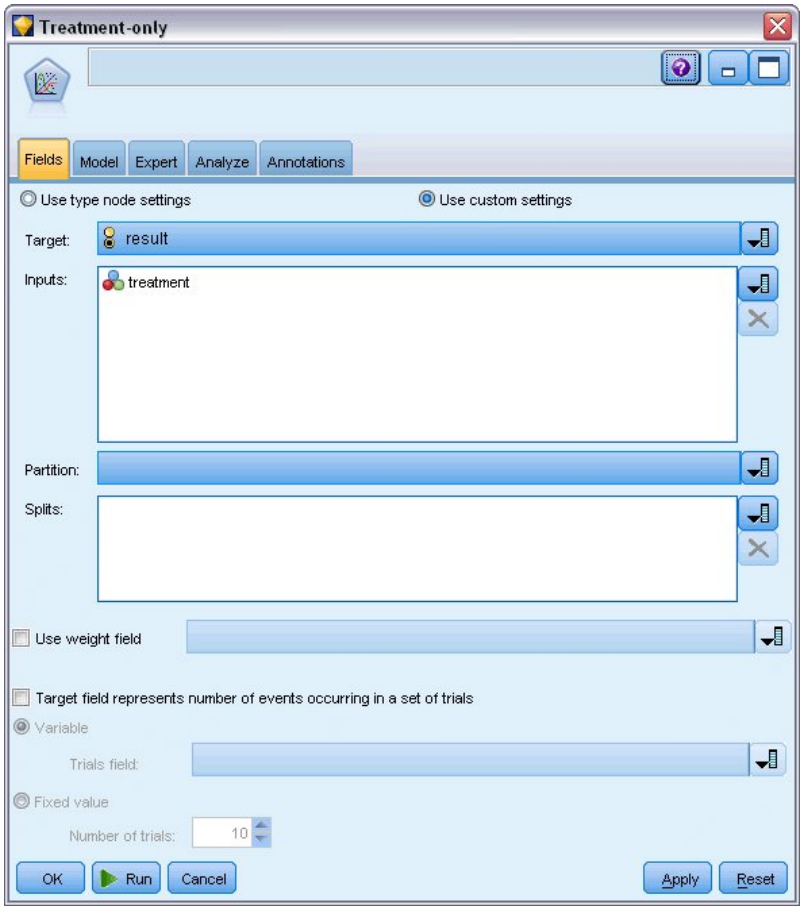


그림 *288.* 필드 옵션 선택

4. 스트림을 실행하여 결과 모델 너깃을 여십시오.

모델 너깃에서 **고급** 탭을 선택하고 아래쪽으로 스크롤하십시오.

모수 추정값

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.442	.5012	-2.425	-.460	8.282	1	.004
[treatment=1]	.378	.6288	-.855	1.610	.361	1	.548
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result
Model: (Intercept), treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

그림 *289.* 치료법 전용 모델에 대한 모수 추정값

치료법 효과(두 치료법 수준 사이의 선형 예측변수의 차이점, 즉, *[treatment=1]*에 대한 계수)가 여전히 통계적으로 유의적이지 않으며 단지 *B* 치료법에 대한 모수 추정값이 *A*에 대한 것보다 크고 처음 12개월 안의 재발 확률이 증가하는 것과 연관되어 *A [treatment=0]* 치료법이 *B [treatment=1]* 치료법보다 나은 것처럼 보입니다. 선형 예측변수(절편 + 치료법 효과)는 로그($-\log(1-P(recur_{12,t}))$의 추정값이며 여기서, $P(recur_{12,t})$는 치료법 t(=*A* 또는 *B*)에 대한 12개월의 재발 확률입니다. 이러한 예측 확률은 데이터 세트 내의 각 관측값에 대해 생성됩니다.

예측된 재발 및 생존 확률

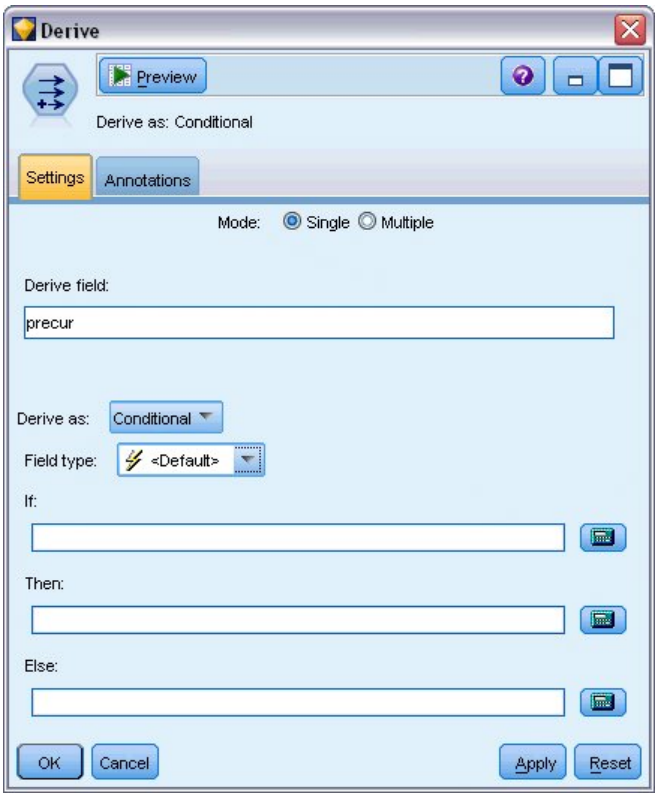


그림 *290.* 파생 노드 설정 옵션

1. 모델은 각 환자에 대해 예측된 결과 및 해당 예측 결과의 확률을 스코어링합니다. 예측된 재발 확률을 보려면 생성된 모델을 팔레트에 복사하고 파생 노드를 연결하십시오.
2. 설정 탭에서 파생 필드로 `precur`을 입력하십시오.
3. 이를 **조건부**로 파생하도록 선택하십시오.
4. 계산기 단추를 클릭하여 **If** 조건에 대한 표현식 작성기를 여십시오.

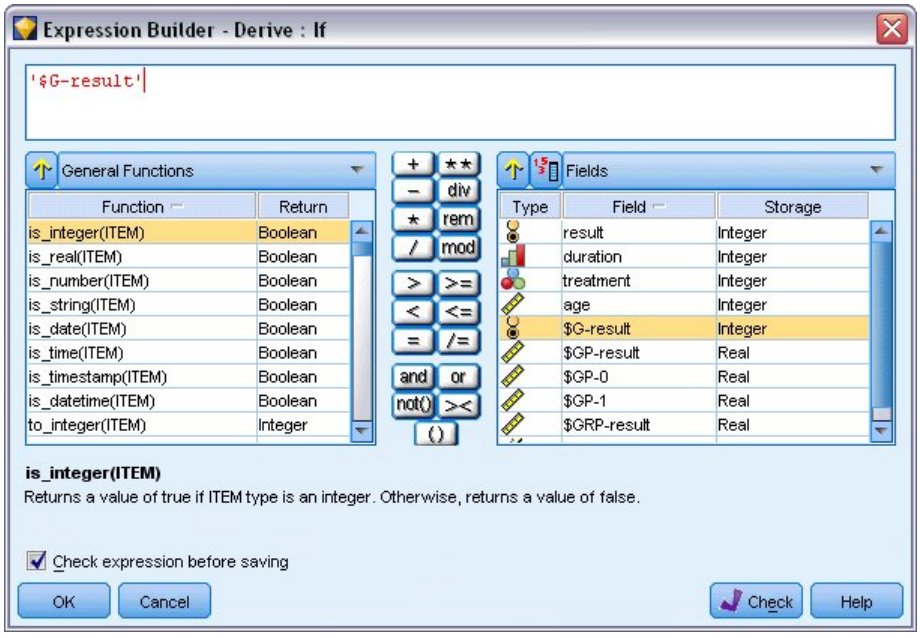


그림 *291.* 파생 노드*: If* 조건에 대한 표현식 작성기

5. *$G-result* 필드를 표현식에 삽입하십시오.
6. **확인**을 클릭하십시오.

   precur 파생 필드는 *$G-result*가 1과 동일하면 **Then** 표현식의 값을 사용하고 0이면 **Else** 표현식의 값을 사용합니다.

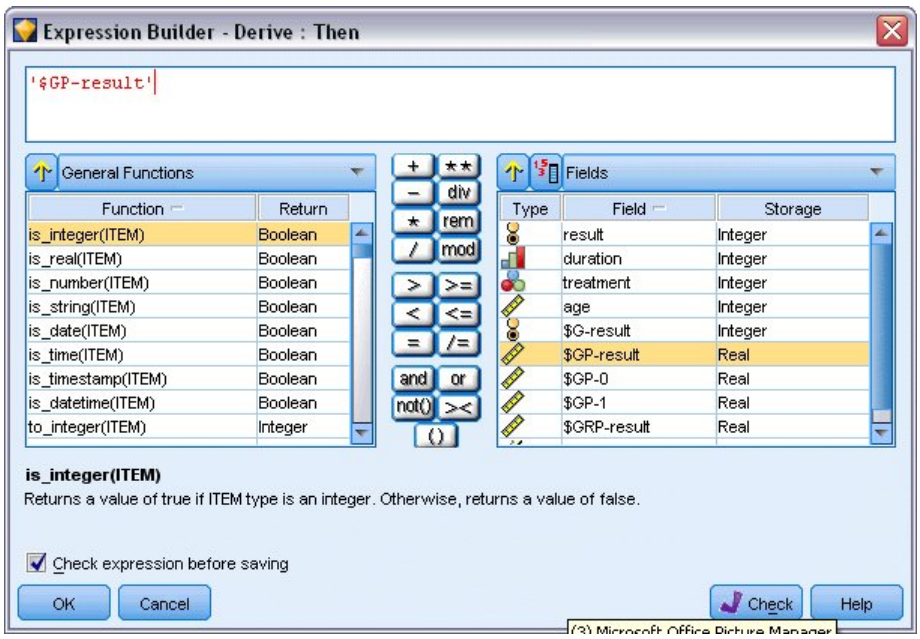


그림 *292.* 파생 노드*: Then* 표현식에 대한 표현식 작성기

7. 계산기 단추를 클릭하여 **Then** 표현식에 대한 표현식 작성기를 여십시오.

8. *$GP-result* 필드를 표현식에 삽입하십시오.
9. **확인**을 클릭하십시오.

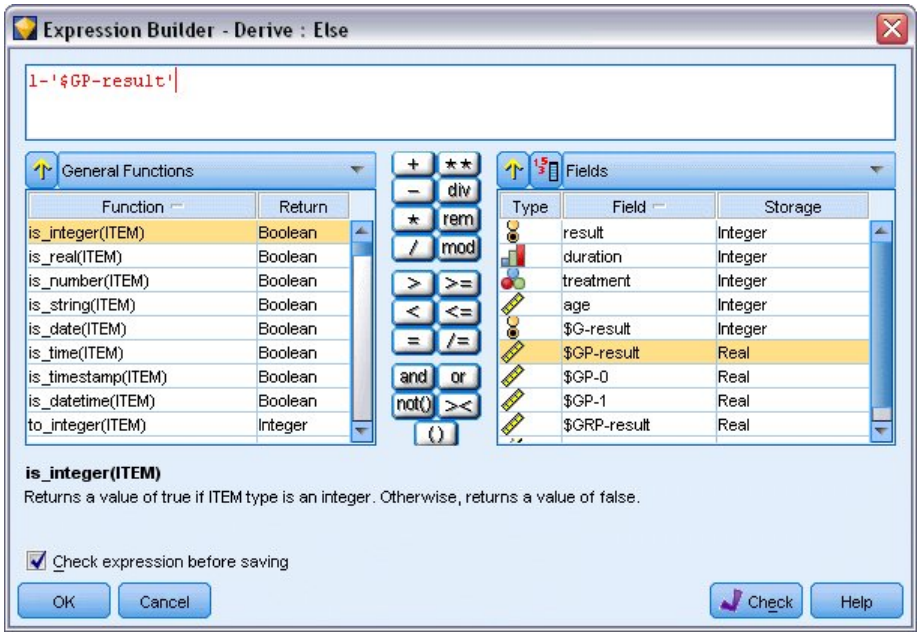
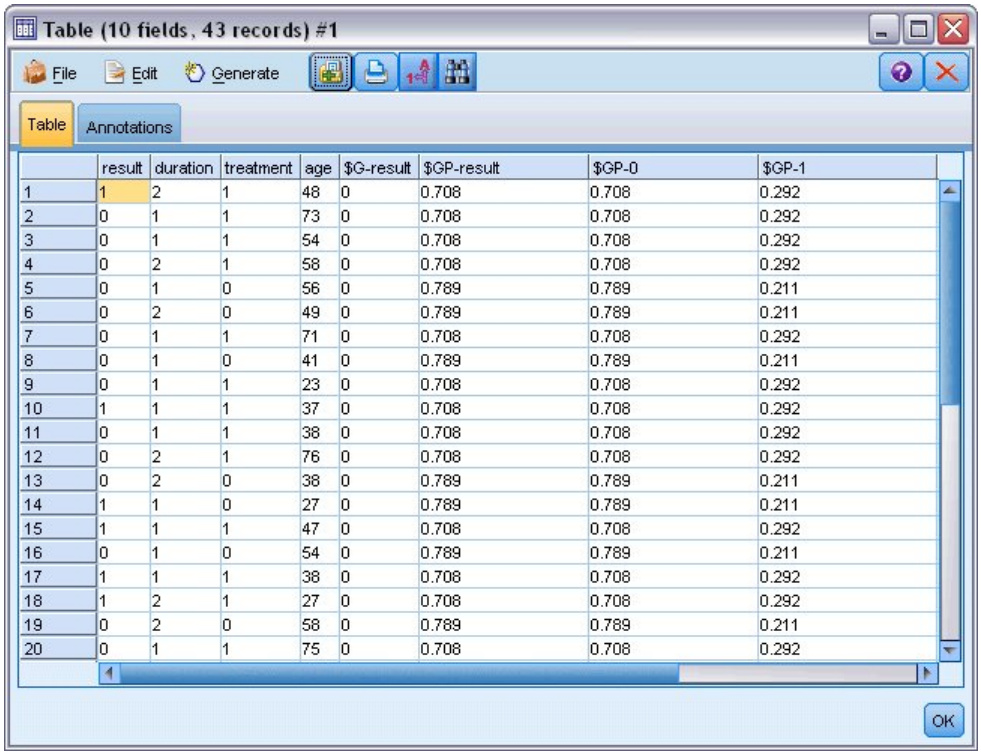


그림 *293.* 파생 노드*: Else* 표현식에 대한 표현식 작성기

10. 계산기 단추를 클릭하여 **Else** 표현식에 대한 표현식 작성기를 여십시오.
11. 표현식에 1-을 입력한 다음 *$GP-result* 필드를 표현식에 삽입하십시오.
12. **확인**을 클릭하십시오.

그림 *294.* 파생 노드 설정 옵션

13. 테이블 노드를 필드 파생 노드에 연결하고 이를 실행하십시오.



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

그림 *295.* 예측 확률

A 치료법에 지정된 환자가 처음 12개월 동안 재발을 경험할 확률이 0.211로 추정되고 *B* 치료법에 대해서는 0.292로 추정됩니다. 1–P(recur$_{12,\ t}$)가 12개월의 생존 확률이며 이는 생존 분석에서 더 관심이 있을 수 있습니다.

주기 기준 반복 확률 모델링

이 모델의 문제점은 첫 번째 검사에서 수집한 정보를 무시한다는 점입니다. 즉, 많은 환자가 처음 6개월 동안은 반복을 경험하지 않았습니다. "더 나은" 모델은 각 구간 동안 이벤트가 발생했는지 여부를 기록하는 이분형 반응을 모델링하는 것입니다. 이 모델을 적합화하려면 원본 데이터 세트를 재구성해야 하며 이에 대해서는 *ulcer_recurrence_recoded.sav*에서 찾을 수 있습니다. 이 파일에는 두 개의 추가 변수가 포함됩니다.

- 케이스가 첫 번째 또는 두 번째 검사 기간에 해당하는지 여부를 기록하는 *Period*
- 지정된 주기 동안 지정된 환자에 대해 재발이 있었는지 여부를 기록하는 *Result by period*

각각의 원본 케이스(환자)는 위험 변수군에 남아 있는 구간당 한 케이스에 기여합니다. 따라서 환자 1이 두 케이스에 기여하게 됩니다. 즉, 재발이 발생하지 않은 첫 번째 검사 주기에 하나, 재발이 기록된 두 번째 검사 주기에 하나입니다. 반면에 환자 10은 첫 번째 주기에서 재발이 기록되었으므로 단일 케이스에만 기여합니다. 환자 16, 28 및 34는 6개월 후에 연구를 중단했으므로 새 데이터 세트에 대해 단일 케이스에만 기여합니다.

1. *Demos* 폴더에서 *ulcer_recurrence_recoded.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

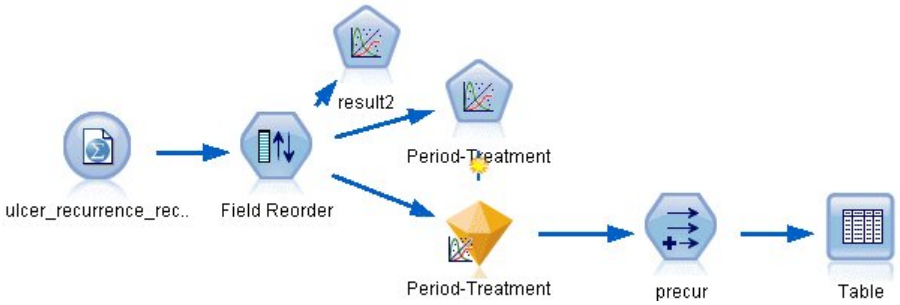


그림 *296.* 궤양 재발을 예측하기 위한 샘플 스트림

2. 소스 노드의 필터 탭에서 *id*, *time* 및 *result*를 제외하십시오.

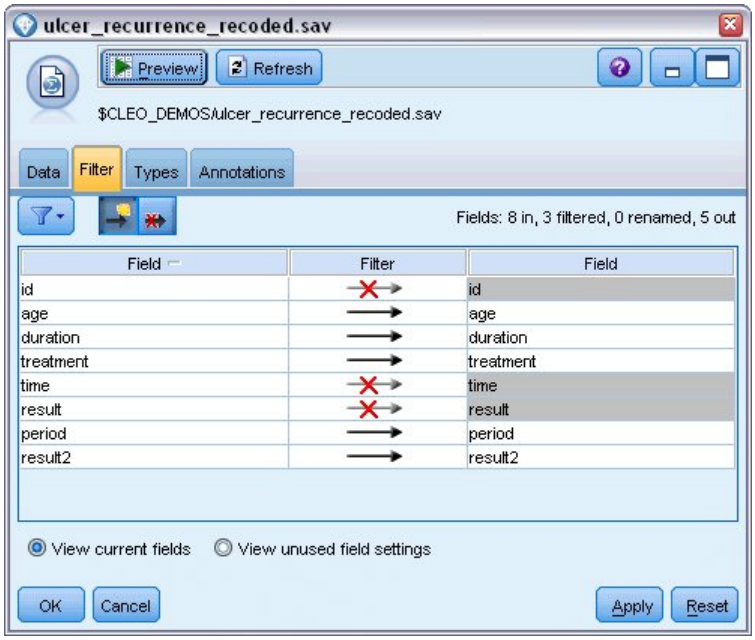


그림 *297.* 원하지 않는 필드 필터링

3. 소스 노드의 유형 탭에서 *result2* 필드에 대한 역할을 **대상**으로 설정하고 측정 수준을 **플래그**로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.

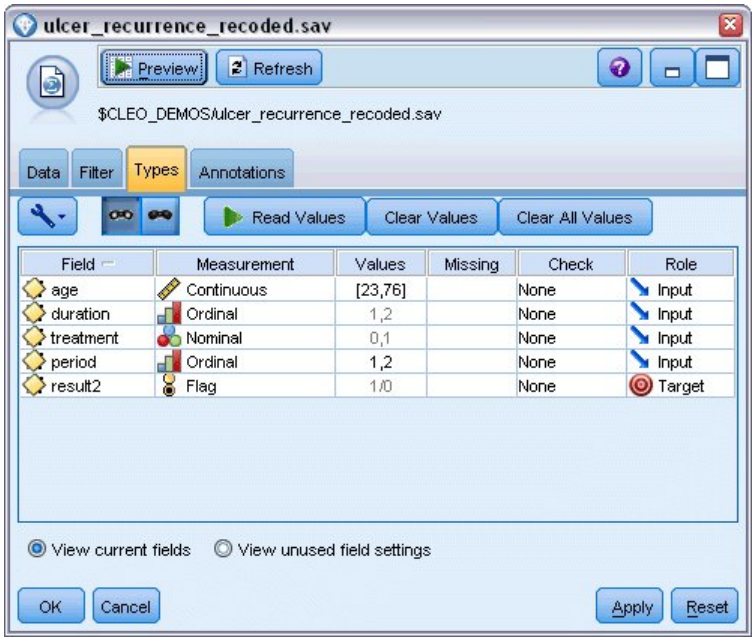


그림 *298.* 필드 역할 설정

4. 필드 다시 정렬 노드를 추가하고 입력 순서로 *period*, *duration*, *treatment* 및 *age*를 지정하십시오. *period*를 첫 번째 입력으로 설정하고 절편 항을 모델에 포함시키지 않으면 더미 변수의 전체 변수군을 적합화하여 주기 효과를 보유할 수 있습니다.

그림 *299.* 필드가 원하는 대로 모델에 입력되도록 필드 다시 정렬

5. GenLin 노드에서 **모델** 탭을 클릭하십시오.

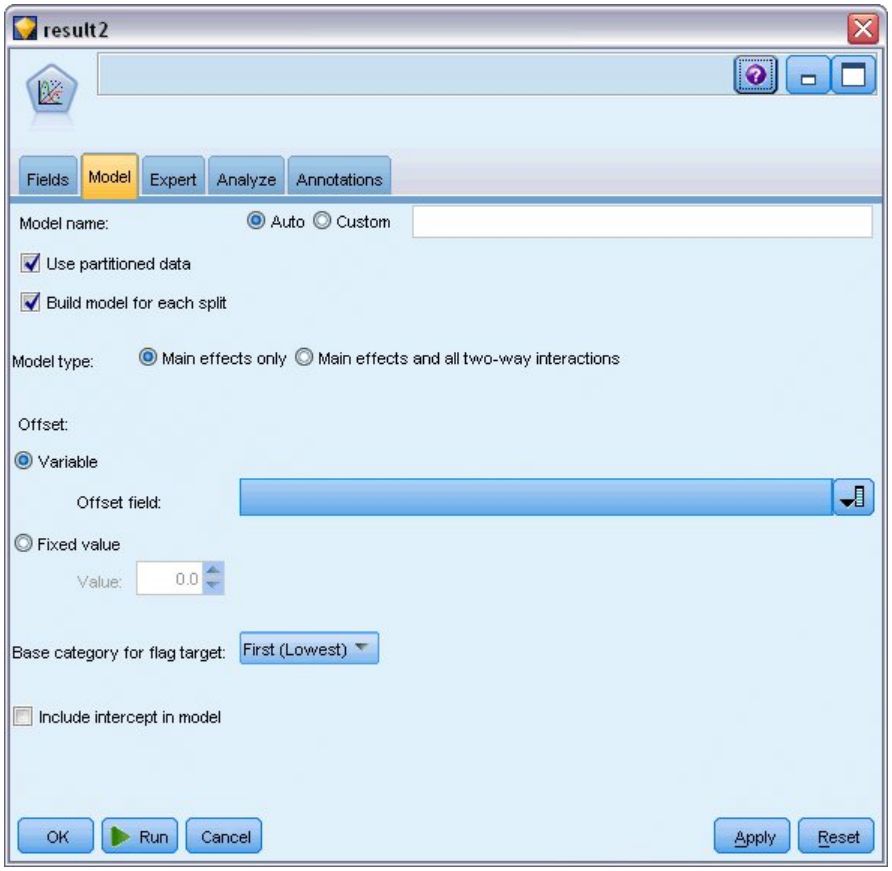


그림 *300.* 모델 옵션 선택

6. 대상에 대한 참조범주로 **처음(가장 낮은 값)**을 선택하십시오. 이는 두 번째 범주가 관심 있는 이벤트이며 모델에 대한 해당 효과가 모수 추정값의 해석에 사용됨을 표시합니다.
7. **모델에 절편 포함**을 선택 취소하십시오.
8. **전문가** 탭을 클릭하고 **전문가**를 선택하여 전문가 모델링 옵션을 활성화하십시오.

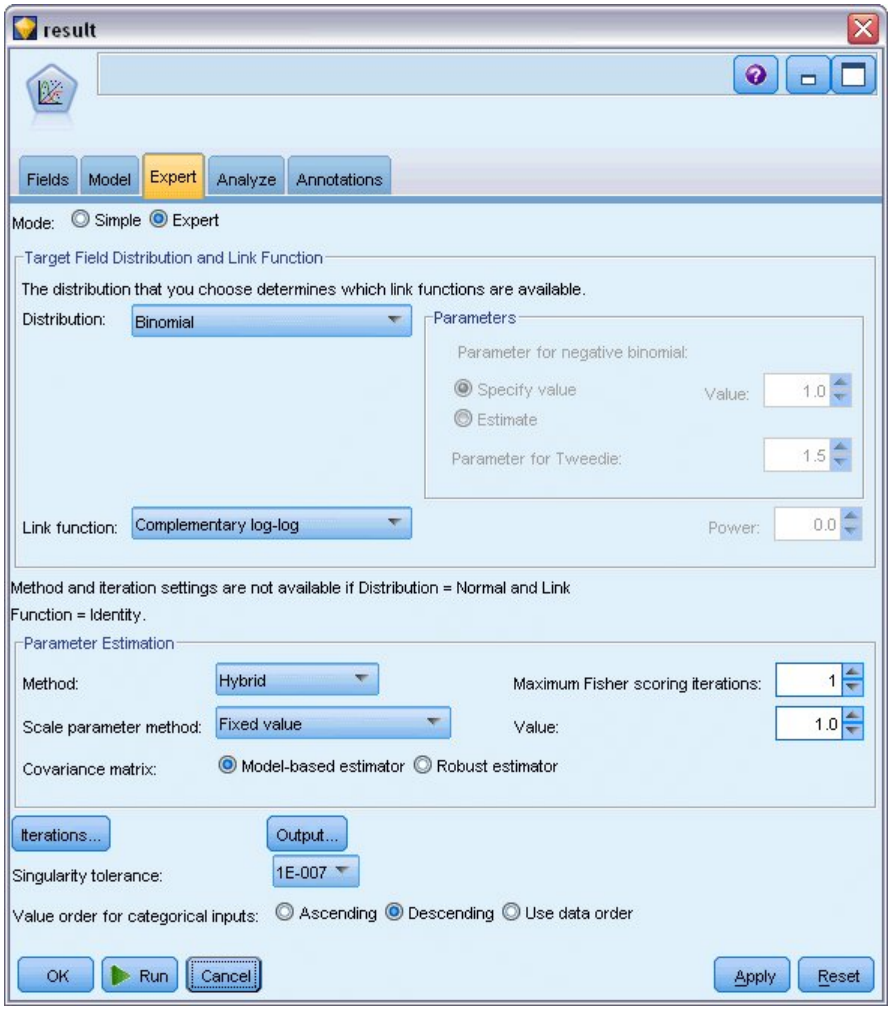


그림 *301.* 고급 옵션 선택

9. 분포로 **이항**을 선택하고 연결함수로 **보 로그-로그**를 선택하십시오.
10. 척도 모수를 추정하기 위한 방법으로 **고정값**을 선택하고 기본값인 1.0을 그대로 두십시오.
11. 요인에 대한 범주 순서로 **내림차순**을 선택하십시오. 이는 각 요인의 첫 번째 범주가 참조 범주가 되고 모델에 대한 이 선택의 효과가 모수 추정값 해석에 사용됨을 표시합니다.
12. 스트림을 실행하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에도 추가됩니다. 모델 세부사항을 보려면 너깃을 마우스 오른쪽 단추로 클릭하고 **편집** 또는 **찾아보기**를 선택하십시오.

모델 효과 검정

Source	Type III		
	Wald Chi-Square	df	Sig.
period	.464	1	.496
duration	.000	1	.988
treatment	.117	1	.732
age	.314	1	.575

Dependent Variable: Result by period
Model: period, duration, treatment, age

그림 *302.* 주효과 모델에 대한 모델 효과 검정

어떠한 모델 효과도 통계적으로 유의적이지 않으나 주기 및 치료법 효과에서 관측된 차이는 의학적으로 흥미가 있으므로 이러한 모델 항만 있는 축소한 모델을 적합화할 수 있습니다.

축소된 모델 적합화

1. GenLin 노드의 필드 탭에서 **사용자 정의 설정**을 클릭하십시오.
2. 대상으로 *result2*를 선택하십시오.
3. 입력으로 *period* 및 *treatment*를 선택하십시오.

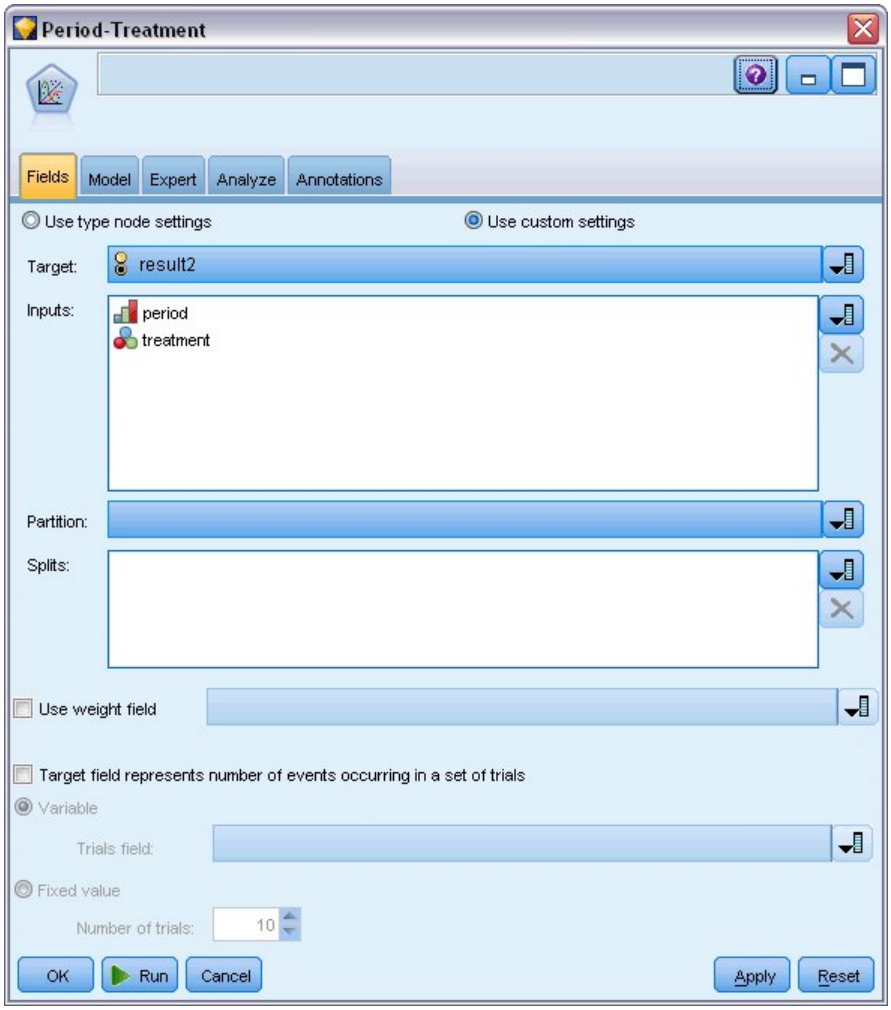


그림 *303.* 필드 옵션 선택

4. 노드를 실행하고 생성된 모델을 찾은 다음 생성된 모델을 팔레트에 복사하고 테이블 노드를 연결한 다음 이를 실행하십시오.

모수 추정값

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
[period=2]	-1.794	.5792	-2.929	-.659	9.597	1	.002
[period=1]	-2.206	.5912	-3.365	-1.047	13.926	1	.000
[treatment=1]	.195	.6279	-1.035	1.426	.097	1	.756
[treatment=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period
Model: period, treatment

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

그림 *304.* 치료법 전용 모델에 대한 모수 추정값

치료법 효과가 여전히 통계적으로 유의적이지 않으며 단지 *B* 치료법에 대한 모수 추정값이 처음 12개월 동안의 재발 확률 증가와 연관되기 때문에 *A* 치료법이 *B* 치료법보다 나은 것처럼 보입니다. 주기 값은 0과 통계적으로 유의미하게 다르나 이는 절편 항이 적합하지 않기 때문입니다. 모델 효과 검정에서 보듯이 주기 효과(*[period=1]* 및 *[period=2]*에 대한 선형 예측변수의 값 사이의 차이)가 통계적으로 유의미하지 않습니다. 선형 예측변수(주기 효과 + 치료법 효과)는 로그$(-\log(1-P(\text{recur}_{p,t}))$의 추정값이며 여기서, $P(\text{recur}_{p,\ t})$는 치료법 t(=*A* 또는 *B*)에 대한 주기 p(=1 또는 2, 6개월 또는 12개월을 나타냄)에서의 재발 확률입니다. 이러한 예측 확률은 데이터 세트 내의 각 관측값에 대해 생성됩니다.

예측된 재발 및 생존 확률

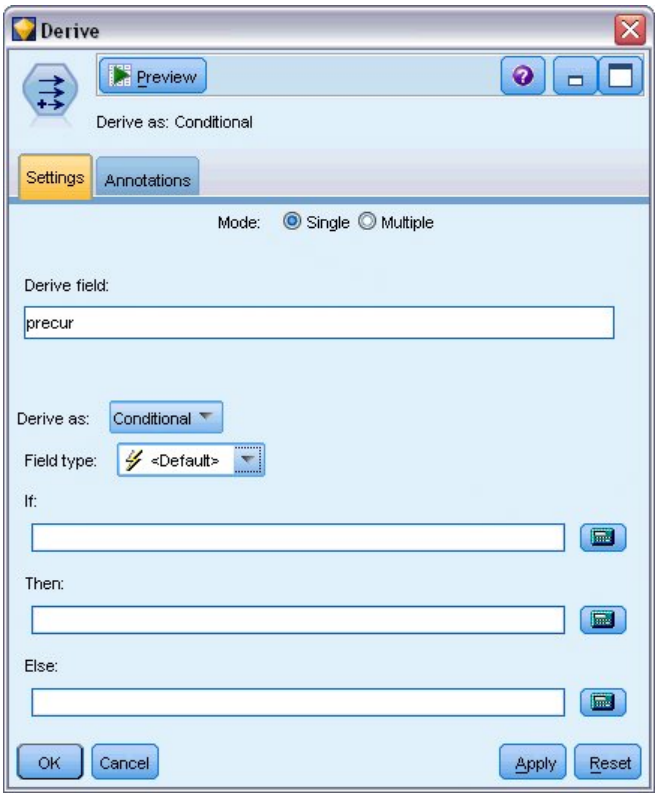


그림 *305.* 파생 노드 설정 옵션

1. 모델은 각 환자에 대해 예측된 결과 및 해당 예측 결과의 확률을 스코어링합니다. 예측된 재발 확률을 보려면 생성된 모델을 팔레트에 복사하고 파생 노드를 연결하십시오.
2. 설정 탭에서 파생 필드로 `precur`을 입력하십시오.
3. 이를 **조건부**로 파생하도록 선택하십시오.
4. 계산기 단추를 클릭하여 **If** 조건에 대한 표현식 작성기를 여십시오.

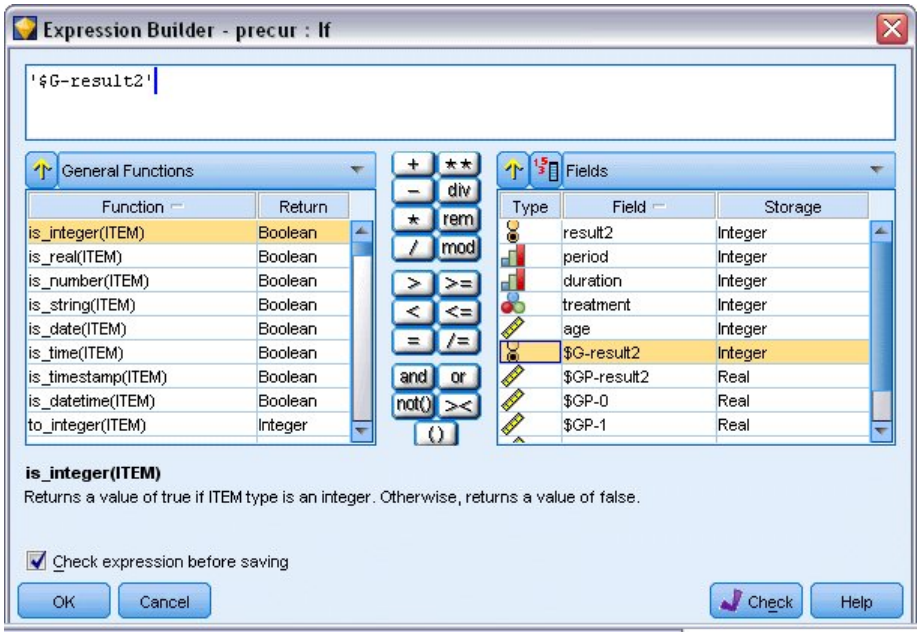


그림 *306.* 파생 노드*: If* 조건에 대한 표현식 작성기

5. *$G-result2* 필드를 표현식에 삽입하십시오.
6. **확인**을 클릭하십시오.

   precur 파생 필드는 *$G-result2*가 1과 동일하면 **Then** 표현식의 값을 사용하고 0이면 **Else** 표현식의 값을 사용합니다.

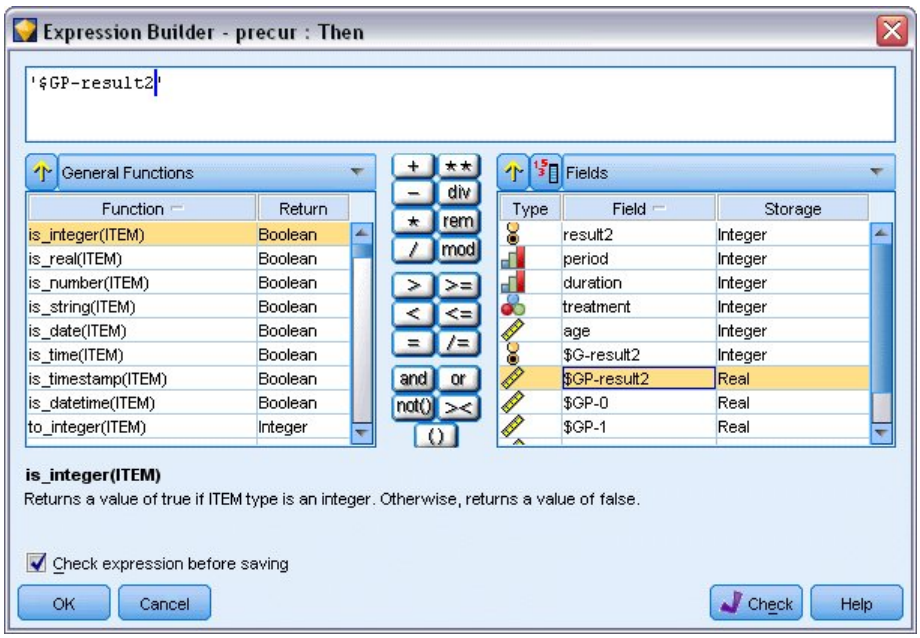


그림 *307.* 파생 노드*: Then* 표현식에 대한 표현식 작성기

7. 계산기 단추를 클릭하여 **Then** 표현식에 대한 표현식 작성기를 여십시오.

8. *$GP-result2* 필드를 표현식에 삽입하십시오.
9. **확인**을 클릭하십시오.

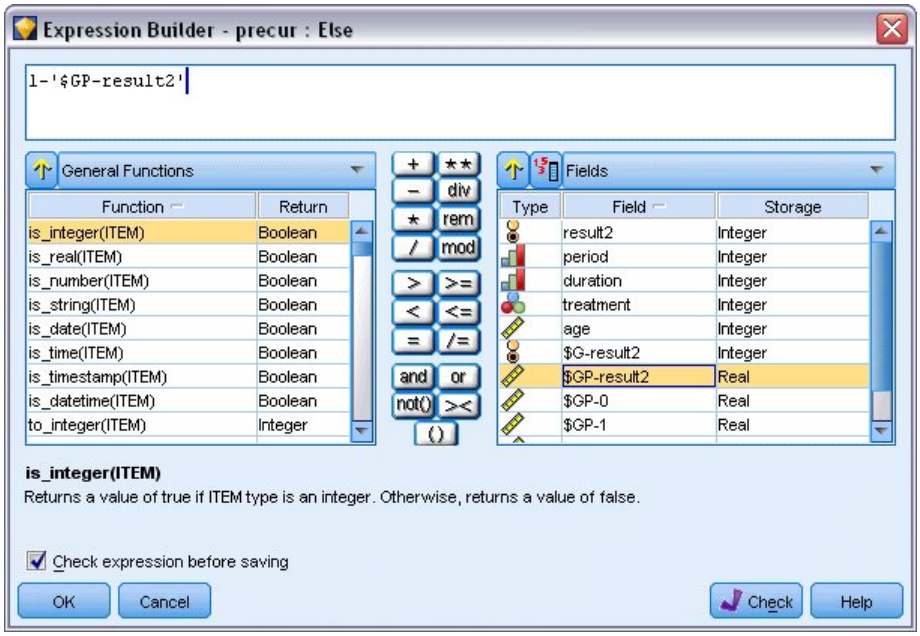


그림 *308.* 파생 노드*: Else* 표현식에 대한 표현식 작성기

10. 계산기 단추를 클릭하여 **Else** 표현식에 대한 표현식 작성기를 여십시오.
11. 표현식에 1-을 입력한 다음 *$GP-result2* 필드를 표현식에 삽입하십시오.
12. **확인**을 클릭하십시오.

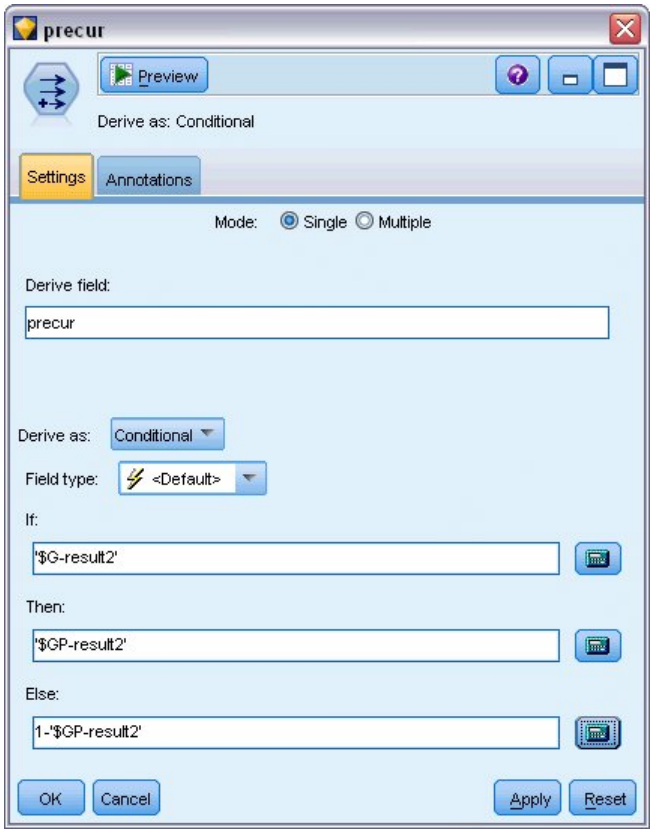
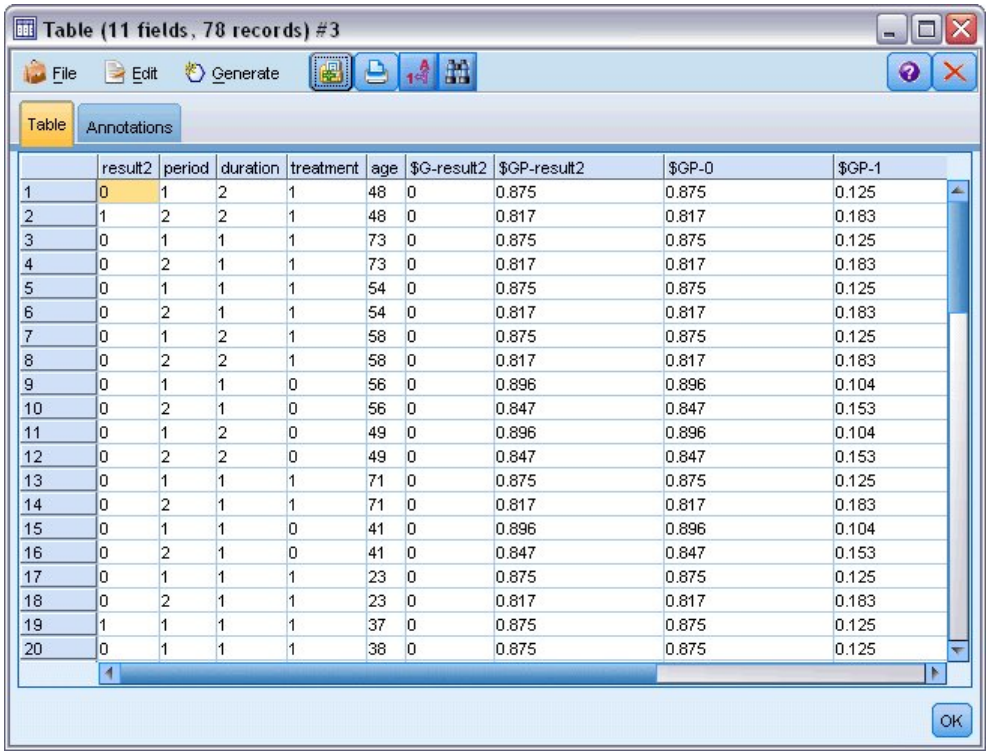


그림 *309.* 파생 노드 설정 옵션

13. 테이블 노드를 필드 파생 노드에 연결하고 이를 실행하십시오.



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

그림 *310.* 예측 확률

표 *3.* 추정된 재발 확률

치료법	**6개월**	**12개월**
A	0.104	0.153
B	0.125	0.183

추정된 재발 확률에서 각 치료법에 대해 12개월 동안의 생존 확률을 $1-(P(recur_{1,\ t}) + P(recur_{2,\ t})\times(1-P(recur_{1,\ t})))$로 추정할 수 있습니다.

A: $1 - (0.104 + 0.153*0.896) = 0.759$

B: $1 - (0.125 + 0.183*0.875) = 0.715$

여기서도 *A*가 더 나은 치료법임을 나타내는 비통계적으로 유의적인 지원이 표시됩니다.

요약

일반화 선형 모델을 사용하면 일련의 보 로그-로그 회귀 모델을 구간 중도절단 생존 데이터에 적합화할 수 있습니다. *A* 치료법을 선택하기 위한 일부 지원이 있는 상태에서 통계적으로 유의미한 결과를 얻으려면 더 많은 연구가 필요합니다. 단, 기존 데이터를 추가적으로 탐색하는 방법도 있습니다.

- 상호작용 효과가 있는 모델을 다시 적합화하는 방법도 효과가 있을 수 있습니다. 특히 *Period* 및 *Treatment group* 사이에 해당됩니다.

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

관련 프로시저

일반화 선형 모델 프로시저는 다양한 모델을 적합화하기 위한 강력한 도구입니다.

- 일반화된 추정 방정식 프로시저는 반복되는 측정을 허용하도록 일반화 선형 모델을 확장합니다.
- 선형 혼합 모델 프로시저를 사용하면 임의 성분 및/또는 반복 측정이 있는 척도 종속변수에 대해 모델을 적합화할 수 있습니다.

권장 참고 자료

일반화 선형 모델에 대한 자세한 정보는 다음 텍스트를 참조하십시오.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

제 23 장 선박 손상 비율을 분석하기 위해 포아송 회귀분석 사용(일반화 선형 모델)

일반화 선형 모델은 개수 데이터의 분석에 대해 포아송 회귀분석을 적합화하는 데 사용될 수 있습니다. 예를 들어, 파도에 의한 화물선 손상에 관한 데이터 세트는 다른 곳[2]에서 제시되고 분석됩니다. 사고 수가 예측변수의 값이 지정된 포아송 비율로 발생하는 것으로 모델링될 수 있으며 결과로 생성된 모델은 손상될 확률이 높은 선박 유형을 판별하는 데 도움이 될 수 있습니다.

이 예에서는 *ships.sav*라는 데이터 파일을 참조하는 *ships_genlin.str* 스트림을 사용합니다. 데이터 파일은 *Demos* 폴더에 있으며 스트림 파일은 *streams* 서브폴더에 있습니다.

이 상황에서는 *Aggregate months of service*가 선박 유형별로 다르므로 원래 값 셀 빈도를 모델링하면 잘못된 결과가 도출될 수 있습니다. 위험에 대한 "노출"의 양을 측정하는 이와 같은 변수는 변위 변수로 일반화 선형 모델에서 처리됩니다. 또한 포아송 회귀분석은 종속변수의 로그가 예측변수에서 선형이라고 가정합니다. 따라서 일반화 선형 모델을 사용하여 포아송 회귀분석을 사고 비율에 적합화하려면 *Logarithm of aggregate months of service*를 사용해야 합니다.

"과분산된" 포아송 회귀분석 적합화

1. *Demos* 폴더에서 *ships.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

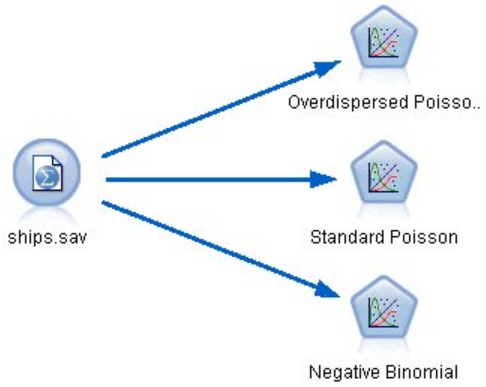


그림 *311.* 손상 비율을 분석하기 위한 샘플 스트림

2. 소스 노드의 필터 탭에서 *months_service* 필드를 제외하십시오. 이 변수의 로그 변환 값은 분석에 사용될 *log_months_service*에 포함됩니다.

2. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

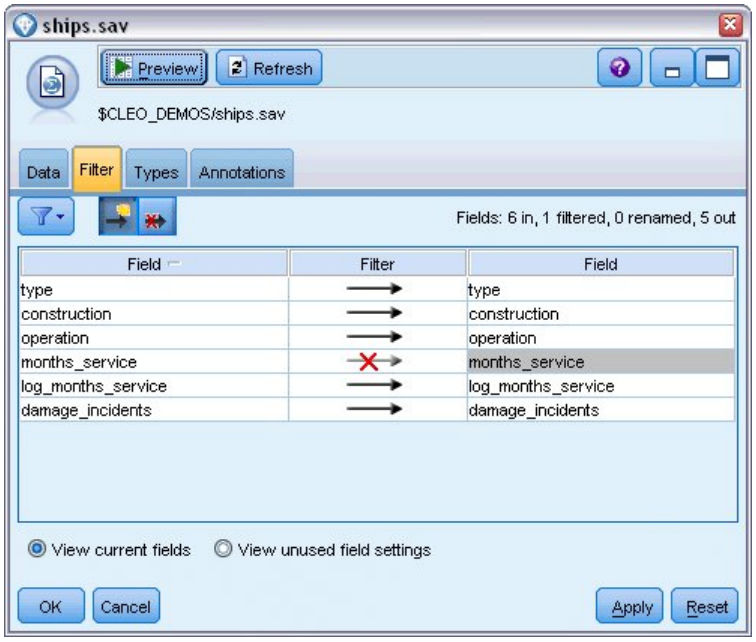


그림 *312.* 불필요한 필드 필터링

(또는 불필요한 필드를 제외하지 않고 유형 탭에서 해당 필드에 대한 역할을 **없음**으로 변경하거나 모델링 노드에서 사용할 필드를 선택할 수 있습니다.)

3. 소스 노드의 유형 탭에서 *damage_incidents* 필드에 대한 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.
4. **값 읽기**를 클릭하여 데이터를 인스턴스화하십시오.

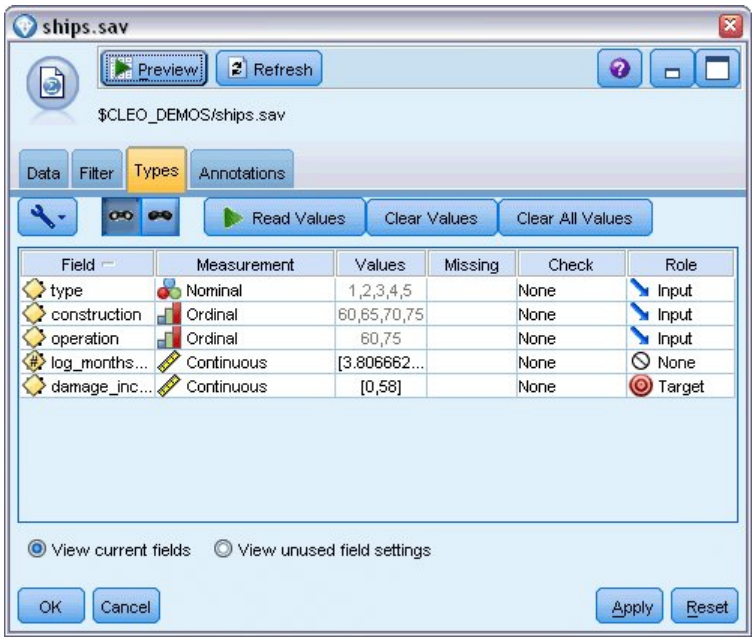


그림 *313.* 필드 역할 설정

5. Genlin 노드를 소스 노드에 연결하고 Genlin 노드에서 **모델** 탭을 클릭하십시오.
6. 변위 변수로 *log_months_service*를 선택하십시오.

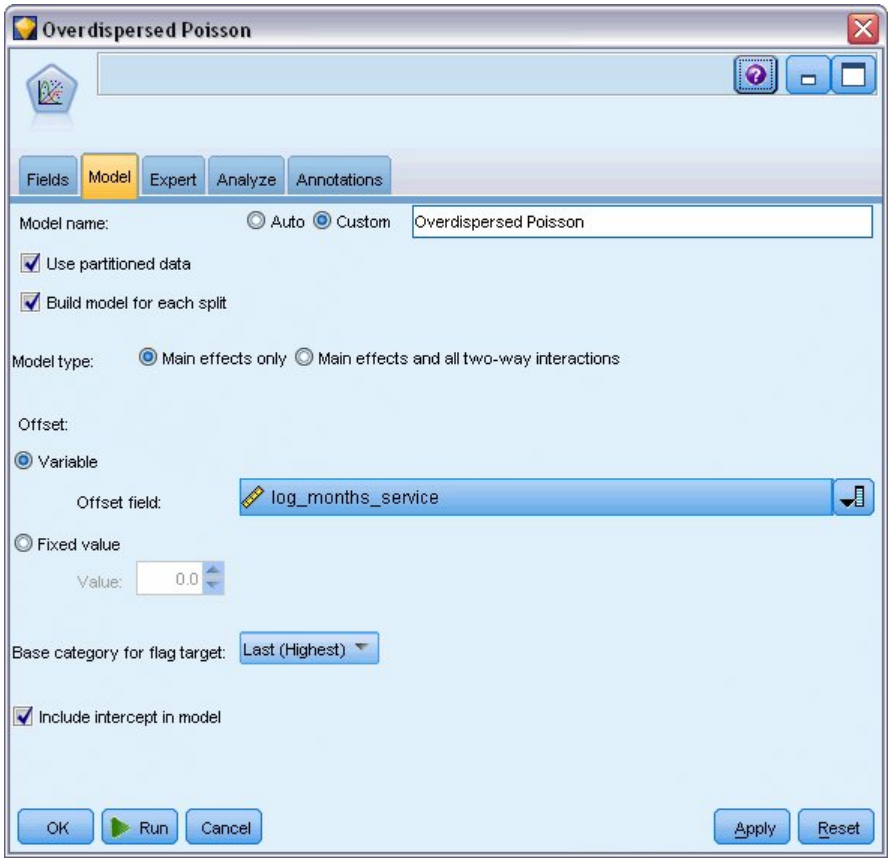


그림 *314.* 모델 옵션 선택

7. **전문가** 탭을 클릭하고 **전문가**를 선택하여 전문가 모델링 옵션을 활성화하십시오.

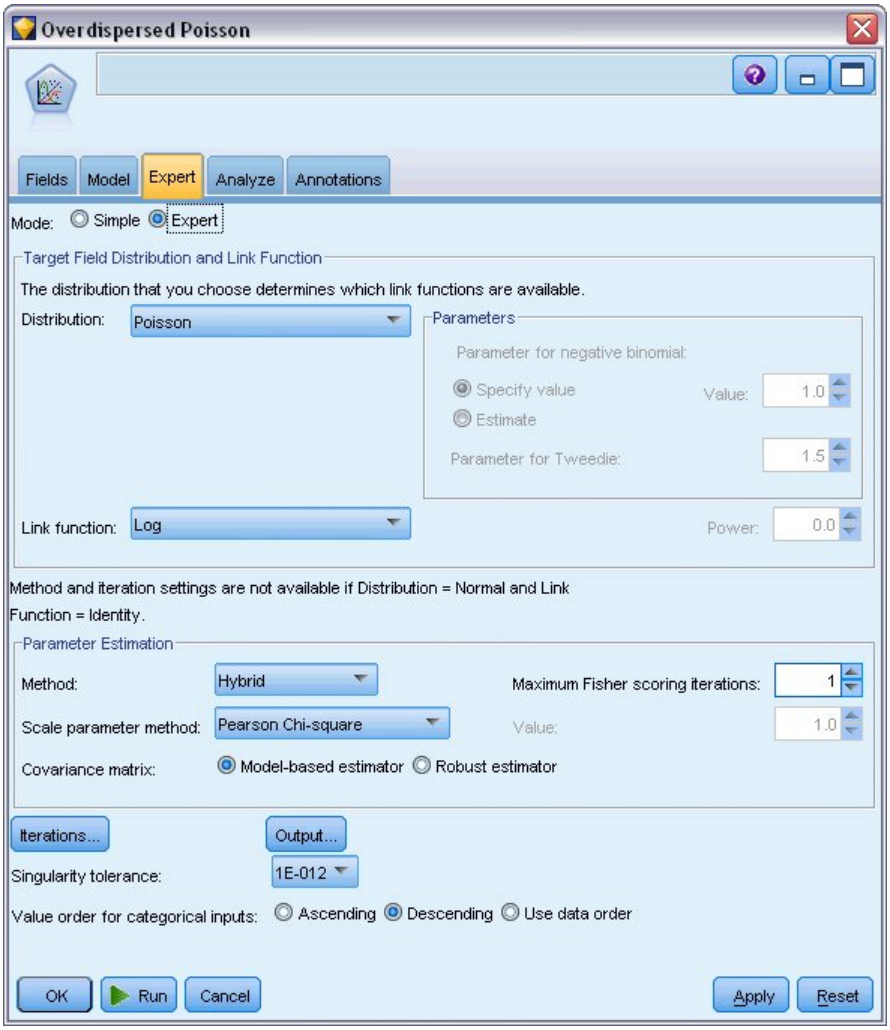


그림 *315.* 고급 옵션 선택

8. 반응에 대한 분포로 **포아송**을 선택하고 연결함수로 **로그**를 선택하십시오.
9. 척도 모수를 추정하기 위한 방법으로 **Pearson 카이제곱**을 선택하십시오. 척도 모수는 일반적으로 포아송 회귀분석에서 1로 가정되나 McCullagh 및 Nelder는 Pearson 카이제곱 추정값을 사용하여 보다 보수적인 분산추정값 및 유의 수준을 얻습니다.
10. 요인에 대한 범주 순서로 **내림차순**을 선택하십시오. 이는 각 요인의 첫 번째 범주가 참조 범주가 되고 모델에 대한 이 선택의 효과가 모수 추정값 해석에 사용됨을 표시합니다.
11. **실행**을 클릭하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에도 추가됩니다. 모델 세부사항을 보려면 너깃을 마우스 오른쪽 단추로 클릭하고 **편집** 또는 **찾아보기**를 선택한 다음 **고급** 탭을 클릭하십시오.

적합도 통계

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

그림 *316.* 적합도 통계

적합도 통계 테이블은 경쟁 모델을 비교하기에 유용한 측도를 제공합니다. 편차에 대한 값/자유도 및 Pearson 카이제곱 통계 또한 척도 모수에 해당되는 추정값을 제공합니다. 이러한 값은 포아송 회귀분석에 대해 1.0 근처여야 합니다. 1.0 보다 크면 과분산된 모델 적합화가 합리적일 수 있음을 나타냅니다.

총괄 검정

Likelihood Ratio Chi-Square	df	Sig.
107.633	8	.000

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Compares the fitted model against the intercept-only model.

그림 *317.* 총괄 검정

총괄 검정은 현재 모델 대 널(이 케이스에서는 절편) 모델의 우도비 카이제곱 검정입니다. 유의수준이 0.05 미만이면 현재 모델이 널 모델보다 낫다는 것을 의미합니다.

모델 효과 검정

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
type	15.415	4	.004
construction	17.242	3	.001
operation	6.249	1	.012

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

그림 *318.* 모델 효과 검정

모델의 각 항에 대해 효과가 있는지 여부를 검정합니다. 유의수준이 0.05 미만인 항은 인식할 수 있는 효과가 있는 것입니다. 각 주효과 항은 모델에 기여합니다.

모수 추정값

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.406	.2828	-6.960	-5.852	513.238	1	.000
[type=5]	.326	.3067	-.276	.927	1.127	1	.288
[type=4]	-.076	.3779	-.817	.665	.040	1	.841
[type=3]	-.687	.4279	-1.526	.151	2.581	1	.108
[type=2]	-.543	.2309	-.996	-.091	5.536	1	.019
[type=1]	0[a]	.	.	.	.	.	.
[construction=75]	.453	.3032	-.141	1.048	2.236	1	.135
[construction=70]	.818	.2208	.386	1.251	13.743	1	.000
[construction=65]	.697	.1946	.316	1.079	12.835	1	.000
[construction=60]	0[a]	.	.	.	.	.	.
[operation=75]	.384	.1538	.083	.686	6.249	1	.012
[operation=60]	0[a]	.	.	.	.	.	.
(Scale)	1.691[b]						

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

그림 *319.* 모수 추정값

모수 추정값 테이블에는 각 예측변수의 효과에 대한 요약 정보가 표시됩니다. 연결함수의 특성으로 인해 이 모델에서 계수 해석이 어려운 반면 공변량에 대한 계수의 부호 및 요인 수준에 대한 공변량의 상대값은 모델에서 예측변수의 효과에 대한 중요한 통찰을 제공합니다.

- 공변량의 경우, 양수(음수) 계수는 예측변수 및 결과 사이의 정관계(역관계)를 표시합니다. 양의 계수를 가진 공변량의 값이 증가하는 것은 손상 사고의 비율이 증가하는 것에 해당됩니다.
- 요인의 경우, 요인 수준의 계수가 클수록 손상 사고가 많음을 나타냅니다. 요인 수준에 대한 계수의 부호는 참조범주에 상대적인 요인 수준의 효과에 따라 결정됩니다.

모수 추정값에 따라 다음과 같이 해석할 수 있습니다.

- 선박 유형 *B [type=2]*는 참조범주인 유형 *A [type=1]*에 비해 통계적으로 유의미하게(*p* 값 0.019) 낮은 손상 비율(추정된 계수 –0.543)을 가집니다. 유형 *C [type=3]*는 실제적으로 추정 모수가 *B*보다 낮지만 *C*의 추정값에서의 변동이 효과를 불명확하게 만듭니다. 요인 수준 간의 모든 관계에 대한 추정 주변 평균을 참조하십시오.
- 1965–69 사이(*[construction=65]*) 및 1970–74 사이(*[construction=70]*)에 건조된 선박은 참조범주인 1960–64 *[construction=60]* 사이에 건조된 선박에 비해 통계적으로 유의미하게(*p* 값 <0.001) 높은 손상 비율(추정된 계수가 각각 0.697 및 0.818임)을 가집니다. 요인 수준 간의 모든 관계에 대한 추정 주변 평균을 참조하십시오.
- 1975–79 사이에 운항 중인 선박(*[operation=75]*)은 1960–1974 사이에 운항 중인 선박(*[operation=60]*)에 비해 통계적으로 유의미하게(*p* 값 0.012) 높은 손상 비율(추정된 계수 0.384)을 가집니다.

대체 모델 적합화

"과분산" 포아송 회귀분석의 한 가지 문제점은 "표준" 포아송 회귀분석에 대해 이를 검정하는 공식적인 방법이 없다는 점입니다. 그러나 과분산이 있는지 판별하기 위해 사용하도록 제안하는 공식적인 검정은 모든 기타 설정은 동일한 "표준" 포아송 회귀분석 및 음이항회귀 사이의 우도비 검정을 수행하는 것입니다. 포아송 회귀분석에 과분산이 없으면 통계 –2×(포아송 모델에 대한 로그-우도 – 음이항 모델에 대한 로그-우도)가 절반은 확률질량이 0이고 나머지는 자유도 1인 카이제곱 분포인 혼합 분포여야 합니다.

1. 척도 모수를 추정하기 위한 방법으로 **고정값**을 선택하십시오. 기본적으로 이 값은 1입니다.

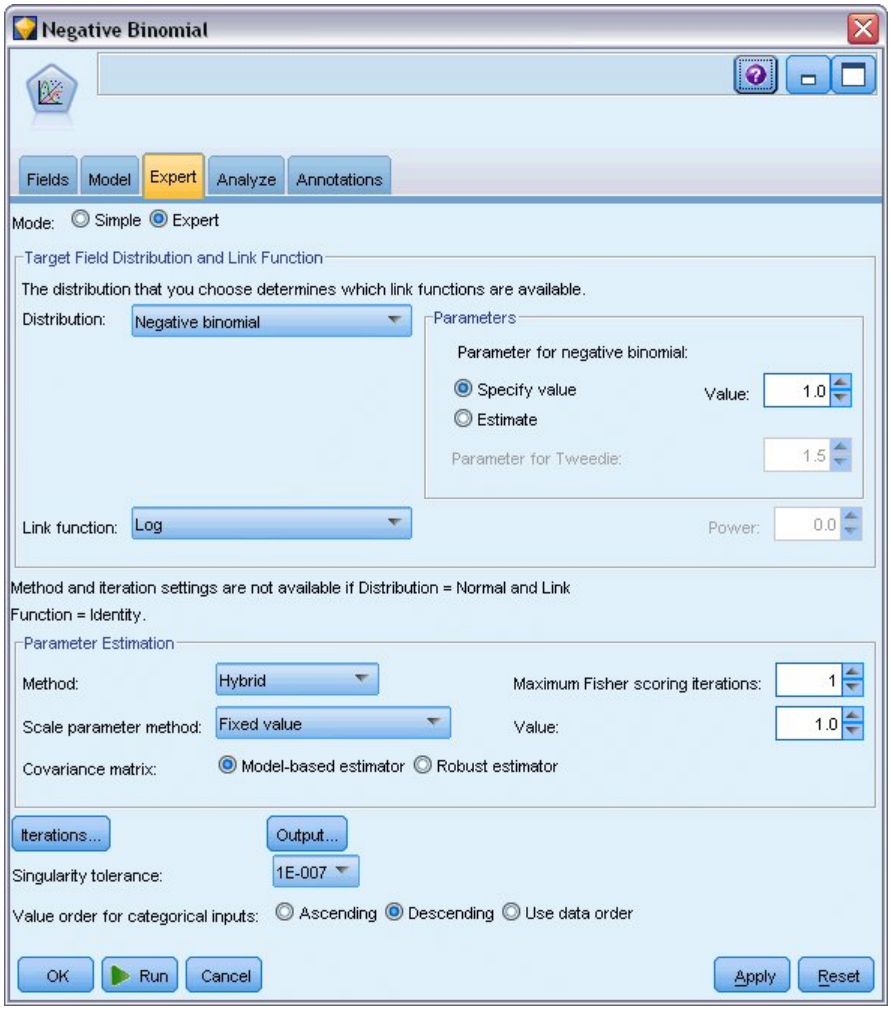


그림 *320.* 전문가 탭

2. 음이항회귀를 적합화하려면 Genlin 노드를 복사하여 이를 소스 노드에 붙여넣고 새 노드를 연 다음 **전문가** 탭을 클릭하십시오.
3. 분포로 **음이항**을 선택하십시오. 보조 모수에 대한 기본값 1을 그대로 두십시오.
4. 스트림을 실행하고 새로 작성된 모델 너깃에서 고급 탭을 찾아보십시오.

적합도 통계

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

그림 *321.* 표준 포아송 회귀분석에 대한 적합도 통계

표준 포아송 회귀분석에 대해 보고된 로그-우도가 –68.281입니다. 이를 음이항 모델과 비교하십시오.

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood[a]	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

그림 *322.* 음이항회귀에 대한 적합도 통계

음이항회귀에 대해 보고된 로그-우도가 –83.725입니다. 이는 실제로 포아송 회귀분석에 대한 로그-우도보다 작습니다. 즉, 우도비 검정을 수행하지 않고도 이 음이항회귀가 포아송 회귀에 비해 개선되지 않았음을 나타냅니다.

그러나 음이항 분포의 보조 모수에 대해 1 값을 선택한 것이 이 데이터 세트에 대해 최적이 아닐 수 있습니다. 과분산을 검정할 수 있는 또 다른 방법은 0과 동일한 보조 모수를 사용하여 음이항 모델을 적합화하고 전문가 탭의 출력 대화 상자에서 LM 검정을 요청하는 것입니다. 검정이 유의적이지 않으면 과분산이 이 데이터 세트의 문제점이 아닌 것입니다.

요약

일반화 선형 모델을 사용하면 세 가지 다른 모델을 개수 데이터에 적합화할 수 있습니다. 음이항회귀는 포아송 회귀분석에 대해 개선되지 않은 것으로 표시되었습니다. 과분산 포아송 회귀분석은 표준 포아송 모델에 대한 적합한 대안을 제공하는 것으로 보이나 선택에 사용할 공식적인 검정이 없습니다.

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

관련 프로시저

일반화 선형 모델 프로시저는 다양한 모델을 적합화하기 위한 강력한 도구입니다.

- 일반화된 추정 방정식 프로시저는 반복되는 측정을 허용하도록 일반화 선형 모델을 확장합니다.
- 선형 혼합 모델 프로시저를 사용하면 임의 성분 및/또는 반복 측정이 있는 척도 종속변수에 대해 모델을 적합화할 수 있습니다.

권장 참고 자료

일반화 선형 모델에 대한 자세한 정보는 다음 텍스트를 참조하십시오.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

제 24 장 자동차 보험 청구에 감마회귀 적합화(일반화 선형 모델)

일반화 선형 모델은 양수 범위 데이터의 분석에 대해 감마회귀를 적합화하는 데 사용될 수 있습니다. 예를 들어, 자동차에 대한 손상 청구에 관한 데이터 세트는 다른 곳[3]에서 제시되고 분석됩니다. 평균 청구 금액은 감마 분포를 사용하고 역 연결함수를 사용하여 종속변수의 평균을 예측변수의 선형 조합에 연관시켜서 모델링할 수 있습니다. 평균 청구 금액을 계산하는 데 사용되는 다양한 청구 수를 고려하기 위해 척도 가중치로 *Number of claims*를 지정하십시오.

이 예에서는 *car_insurance_claims.sav*라는 데이터 파일을 참조하는 *car-insurance_genlin.str*이라는 스트림을 사용합니다. 데이터 파일은 *Demos* 폴더에 있으며 스트림 파일은 *streams* 서브폴더에 있습니다.

스트림 작성

1. *Demos* 폴더에서 *car_insurance_claims.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

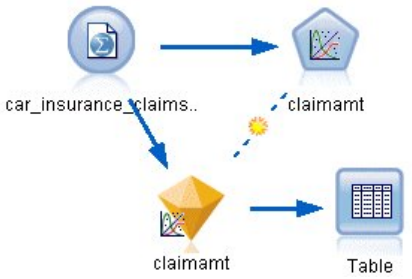


그림 *323.* 자동차 보험 청구를 예측하기 위한 샘플 스트림

2. 소스 노드의 유형 탭에서 *claimamt* 필드에 대한 역할을 **대상**으로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.
3. **값 읽기**를 클릭하여 데이터를 인스턴스화하십시오.

3. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

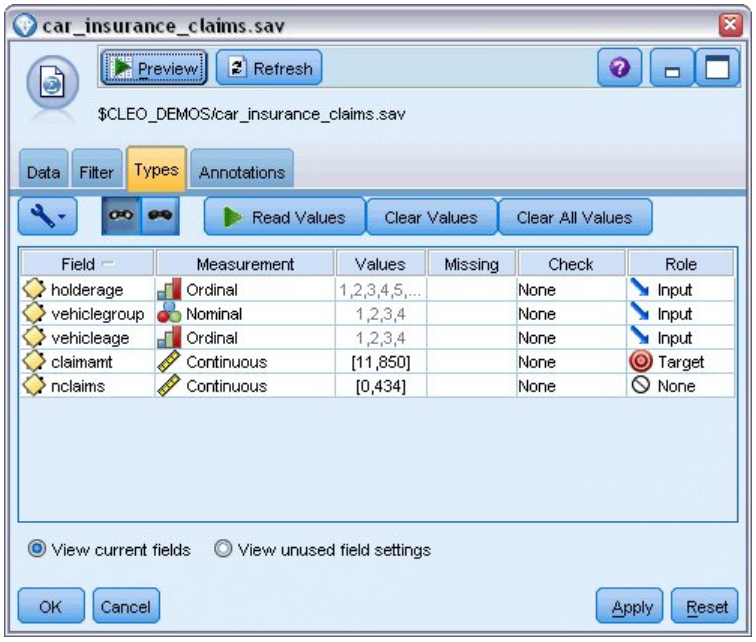


그림 *324.* 필드 역할 설정

4. Genlin 노드를 소스 노드에 연결하고 Genlin 노드에서 필드 탭을 클릭하십시오.
5. 척도 가중값 필드로 *nclaims*를 선택하십시오.

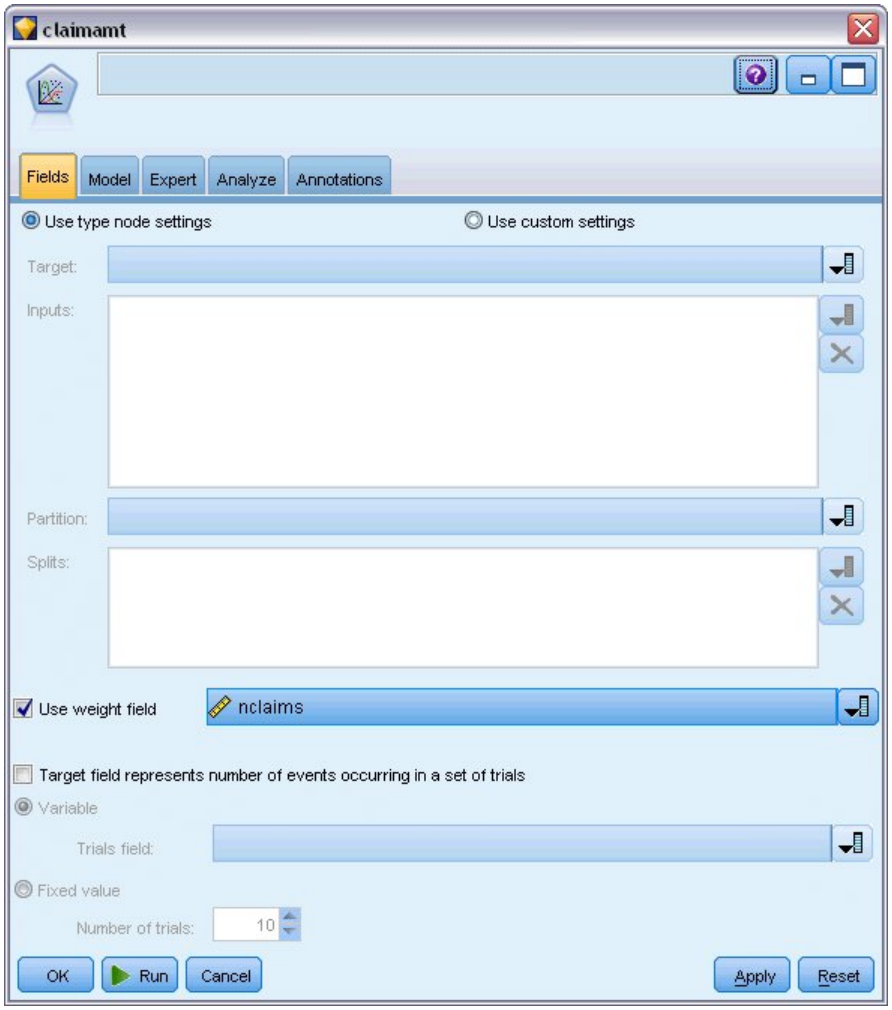


그림 *325.* 필드 옵션 선택

6. 전문가 탭을 클릭하고 **전문가**를 선택하여 전문가 모델링 옵션을 활성화하십시오.

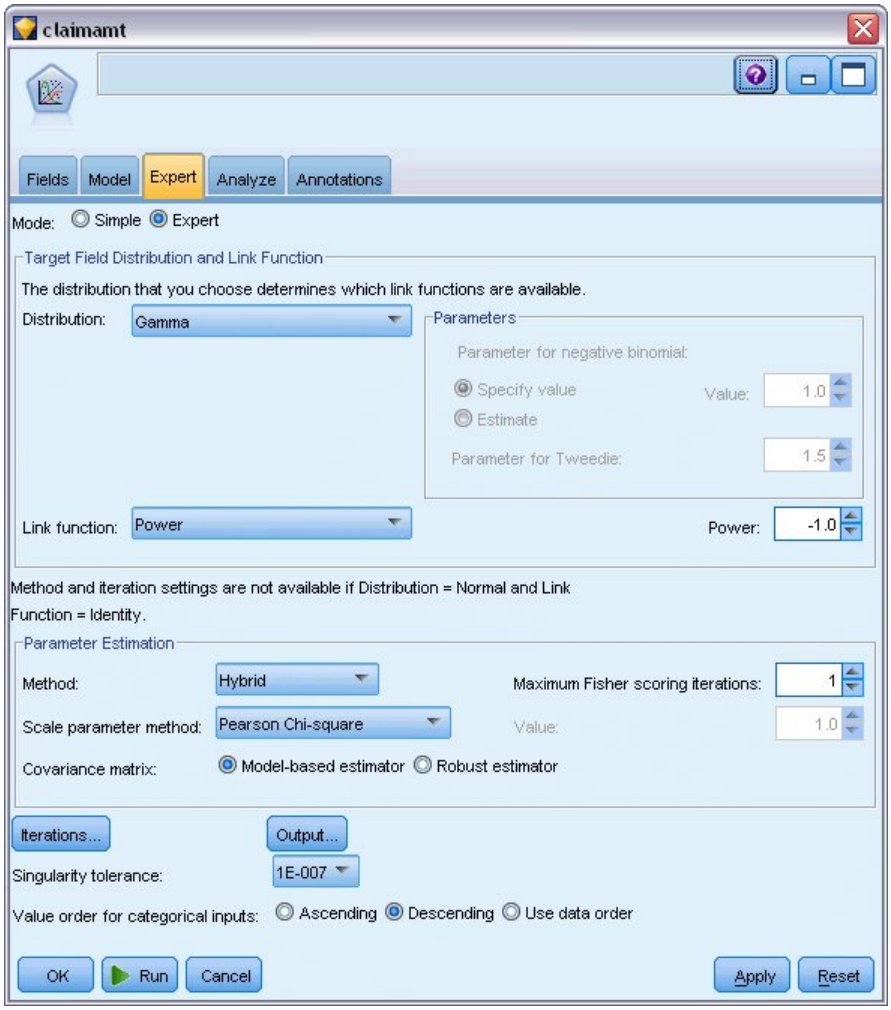


그림 *326.* 고급 옵션 선택

7. 반응 분포로 **감마**를 선택하십시오.
8. 연결함수로 **거듭제곱**을 선택하고 거듭제곱 함수의 지수로 `-1.0`을 입력하십시오. 이는 역 링크입니다.
9. 척도 모수를 추정하기 위한 방법으로 **Pearson 카이제곱**을 선택하십시오. 이는 McCullagh 및 Nelder에 의해 사용되는 방법이므로 여기서는 결과를 복제하도록 이에 따릅니다.
10. 요인에 대한 범주 순서로 **내림차순**을 선택하십시오. 이는 각 요인의 첫 번째 범주가 참조 범주가 되고 모델에 대한 이 선택의 효과가 모수 추정값 해석에 사용됨을 표시합니다.
11. **실행**을 클릭하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 캔버스 및 모델 팔레트에도 추가됩니다. 모델 세부사항을 보려면 모델 너깃을 마우스 오른쪽 단추로 클릭하고 **편집** 또는 **찾아보기**를 선택한 다음 고급 탭을 선택하십시오.

모수 추정값

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003411	.000418	.002591	.004230	66.593	1	.000
[holderage=8]	.000920	.000416	.000105	.001735	4.898	1	.027
[holderage=7]	.000916	.000408	.000117	.001716	5.046	1	.025
[holderage=6]	.000969	.000405	.000176	.001763	5.740	1	.017
[holderage=5]	.001370	.000419	.000548	.002192	10.682	1	.001
[holderage=4]	.000462	.000411	-.000342	.001267	1.268	1	.260
[holderage=3]	.000350	.000412	-.000458	.001158	.720	1	.396
[holderage=2]	.000101	.000436	-.000754	.000956	.054	1	.816
[holderage=1]	.000000[a]	.	.	.	.	.	.
[vehiclegroup=4]	-.001421	.000181	-.001775	-.001067	61.883	1	.000
[vehiclegroup=3]	-.000614	.000170	-.000947	-.000281	13.039	1	.000
[vehiclegroup=2]	.000038	.000169	-.000293	.000368	.050	1	.823
[vehiclegroup=1]	.000000[a]	.	.	.	.	.	.
[vehicleage=4]	.004154	.000442	.003287	.005021	88.175	1	.000
[vehicleage=3]	.001651	.000227	.001207	.002096	53.013	1	.000
[vehicleage=2]	.000366	.000101	.000169	.000564	13.191	1	.000
[vehicleage=1]	.000000[a]	.	.	.	.	.	.
(Scale)	1.209[b]			.001	.0004	.000	.002

Dependent Variable: Average cost of claims
Model: (Intercept), holderage, vehiclegroup, vehicleage

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

그림 *327.* 모수 추정값

총괄 검정 및 모델 효과 검정(표시되지 않음)은 모델이 널 모델보다 더 잘 수행하며 각 주효과 항이 모델에 기여함을 나타냅니다. 모수 추정값 테이블은 요인 수준 및 척도 모수에 대해 McCullagh 및 Nelder에 의해 획득된 값과 동일한 값을 표시합니다.

요약

일반화 선형 모델을 사용하면 감마회귀를 청구 데이터에 적합화할 수 있습니다. 이 모델에서는 감마 분포에 대해 정준 연결함수가 사용되지만 로그 링크도 합리적인 결과를 제공합니다. 일반적으로 다른 연결함수를 가진 모델을 직접 비교하기는 어렵지만 로그 링크는 지수가 0인 특수한 케이스의 제곱 링크이므로 로그 링크가 있는 모델과 제곱 링크가 있는 모델의 편차값을 비교하여 어느 모델이 더 적합한지 판별할 수 있습니다. 예를 들어, McCullagh 및 Nelder의 11.3 절을 참조하십시오.

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

관련 프로시저

일반화 선형 모델 프로시저는 다양한 모델을 적합화하기 위한 강력한 도구입니다.

- 일반화된 추정 방정식 프로시저는 반복되는 측정을 허용하도록 일반화 선형 모델을 확장합니다.

- 선형 혼합 모델 프로시저를 사용하면 임의 성분 및/또는 반복 측정이 있는 척도 종속변수에 대해 모델을 적합화할 수 있습니다.

권장 참고 자료

일반화 선형 모델에 대한 자세한 정보는 다음 텍스트를 참조하십시오.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2 ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., and J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

제 25 장 세포 표본 분류(SVM)

지원 벡터 머신(SVM)은 특히 넓은 범위의 데이터 세트에 적합한 분류 및 회귀분석 기법입니다. 광범위한 데이터 세트는 생물 정보학(생화학 및 생물학 데이터에 정보 기술 적용) 필드에서 발생하는 데이터 세트와 같이 예측변수의 수가 많은 데이터 세트를 의미합니다.

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 이러한 세포 특성 값을 사용할 수 있는 SVM 모델을 개발하여 표본이 양성인지 또는 악성인지 여부를 조기에 표시하고자 합니다.

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

예는 UCI Machine Learning Repository에서 공개적으로 사용 가능한 데이터 세트를 기준으로 합니다. 데이터 세트는 수백 개의 인간 세포 표본 레코드로 구성되며 각 레코드는 세포의 공정특성 변수에 대한 값 세트를 포함합니다. 각 레코드의 필드는 다음과 같습니다.

필드 이름	설명
ID	환자 식별자
Clump	군집 두께
UnifSize	세포 크기의 균일성
UnifShape	세포 모양의 균일성
MargAdh	변연 유착
SingEpiSize	단일 상피 세포 크기
BareNuc	노출 핵
BlandChrom	블랜드 염색질
NormNucl	정상 핵소체
Mit	유사 분열
Class	양성 또는 악성

이 예에서는 각 레코드에서 상대적으로 적은 수의 예측변수가 있는 데이터 세트를 사용할 것입니다.

스트림 작성

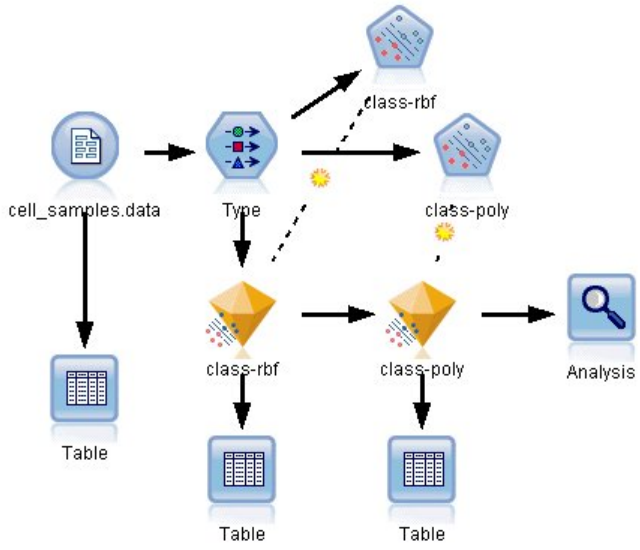
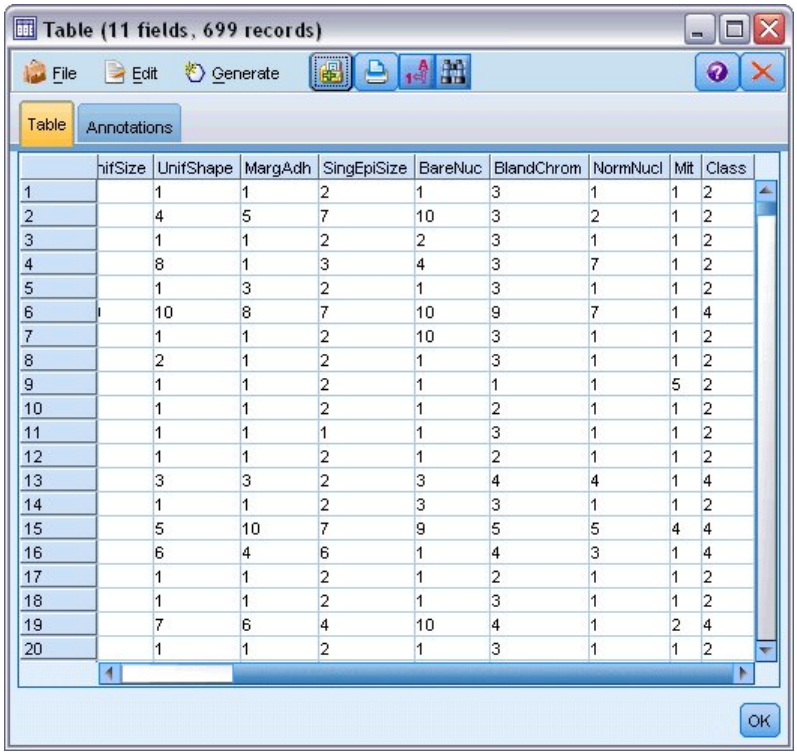


그림 *328. SVM* 모델링을 표시하기 위한 샘플 스트림

1. 새 스트림을 작성하고 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *cell_samples.dat*를 가리키는 변수 파일 소스 노드를 추가하십시오.

   소스 파일에서 데이터를 보십시오.
2. 테이블 노드를 스트림에 추가하십시오.
3. 테이블 노드를 변수 파일 노드에 연결하고 스트림을 실행하십시오.



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

그림 *329. SVM*에 대한 소스 데이터

ID 필드에는 환자 식별자가 포함됩니다. 각 환자의 세포 표본의 공정특성 변수는 *Clump*에서 *Mit* 필드에 포함됩니다. 값은 1에서 10까지 등급이 매겨지며 1이 양성에 가장 가깝습니다.

Class 필드에는 별도의 의학적 프로시저에 의해 확인된 대로 표본이 양상(값 = 2) 또는 악성(값 = 4)인지에 대한 진단이 포함됩니다.

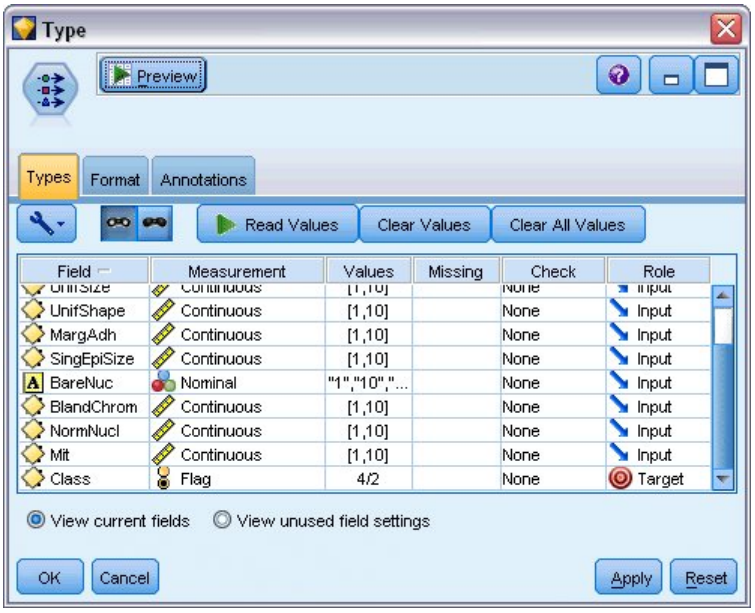


그림 *330.* 유형 노드 설정

4. 유형 노드를 추가하고 이를 변수 파일 노드에 연결하십시오.
5. 유형 노드를 여십시오.

   모델이 *Class*의 값이 양성(=2) 또는 악성(=4)인지 예측하기를 원합니다. 이 필드가 두 가지 가능한 값 중 하나만 가질 수 있으므로 측정 수준이 이를 반영하도록 변경해야 합니다.
6. *Class* 필드(목록의 마지막 필드)에 대한 **측정** 열에서 **연속형** 값을 클릭하여 이를 **플래그**로 변경하십시오.
7. **값 읽기**를 클릭하십시오.
8. **역할** 열에서 *ID*(환자 식별자)에 대한 역할을 **없음**으로 설정하십시오. 예측변수 또는 모델에 대한 대상으로 사용되지 않기 때문입니다.
9. 대상인 *Class*에 대한 역할을 **대상**으로 설정하고 모든 기타 필드(예측변수)의 역할을 **입력**으로 남겨 두십시오.
10. **확인**을 클릭하십시오.

    SVM 노드는 처리를 수행하기 위한 커널 함수 선택사항을 제공합니다. 지정된 데이터 세트에 대해 어떤 함수가 가장 잘 수행할 것인지 쉽게 알 수 있는 방법이 없으므로 다양한 함수를 차례로 선택하여 결과를 비교할 것입니다. 기본값인 RBF(방사형 기저함수)부터 시작합니다.

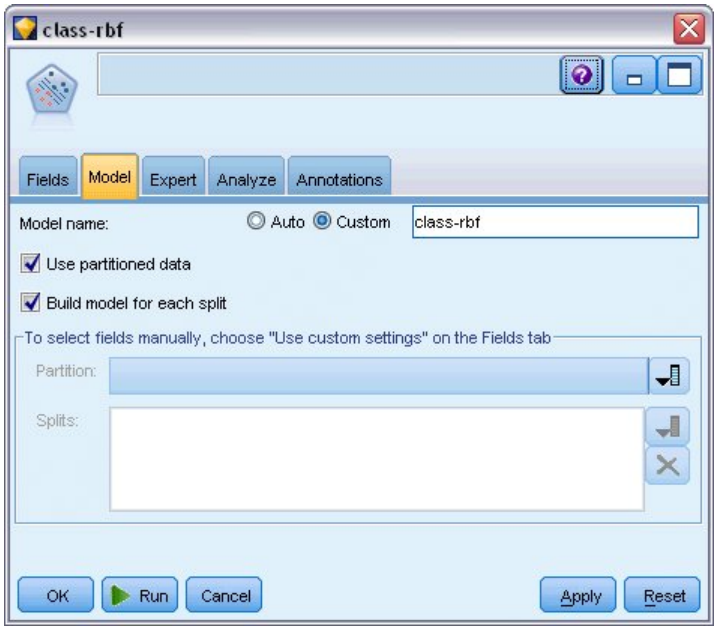


그림 *331.* 모델 탭 설정

11. 모델링 팔레트에서 SVM 노드를 유형 노드에 연결하십시오.
12. SVM 노드를 여십시오. **모델** 탭에서 **모델 이름**에 대해 **사용자 정의** 옵션을 클릭하고 인접 텍스트 필드에 *class-rbf*를 입력하십시오.

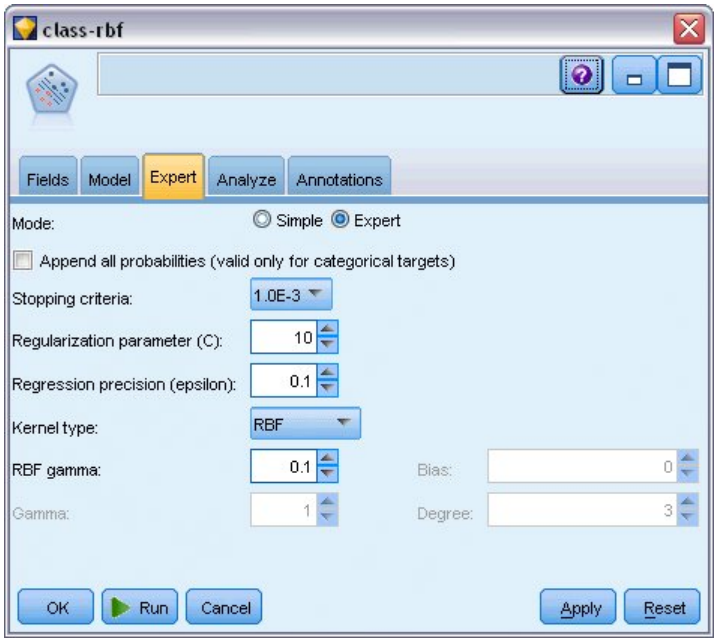


그림 *332.* 기본 전문가 탭 설정

13. **전문가** 탭에서 **모드**를 **전문가**로 설정하십시오. 단, 모든 기본 옵션은 그대로 두십시오. **커널 유형**은 기본적으로 **RBF**로 설정됩니다. 모든 옵션은 단순 모드에서 비활성 상태로 표시됩니다.

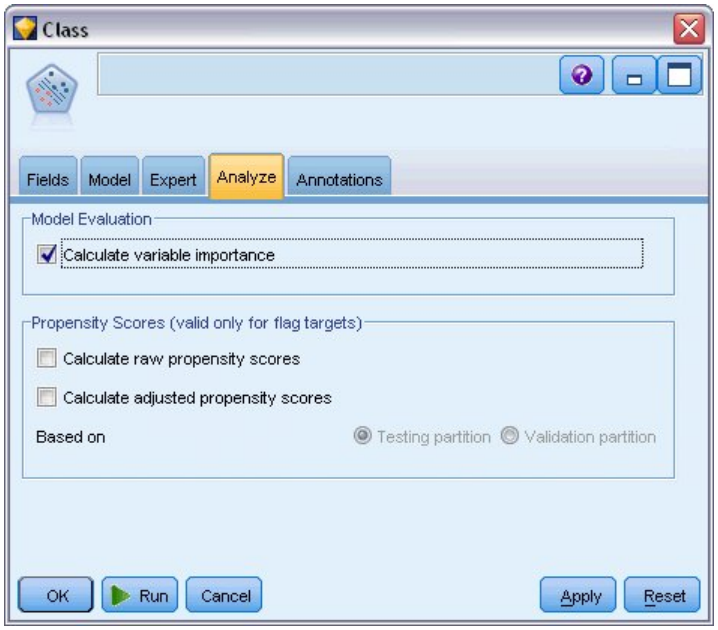


그림 *333.* 분석 탭 설정

14. **분석** 탭에서 **변수 중요도 계산** 선택란을 선택하십시오.
15. **실행**을 클릭하십시오. 모델 너깃이 스트림 및 화면 오른쪽 맨 위의 모델 팔레트에 배치됩니다.
16. 스트림에서 모델 너깃을 두 번 클릭하십시오.

데이터 탐색

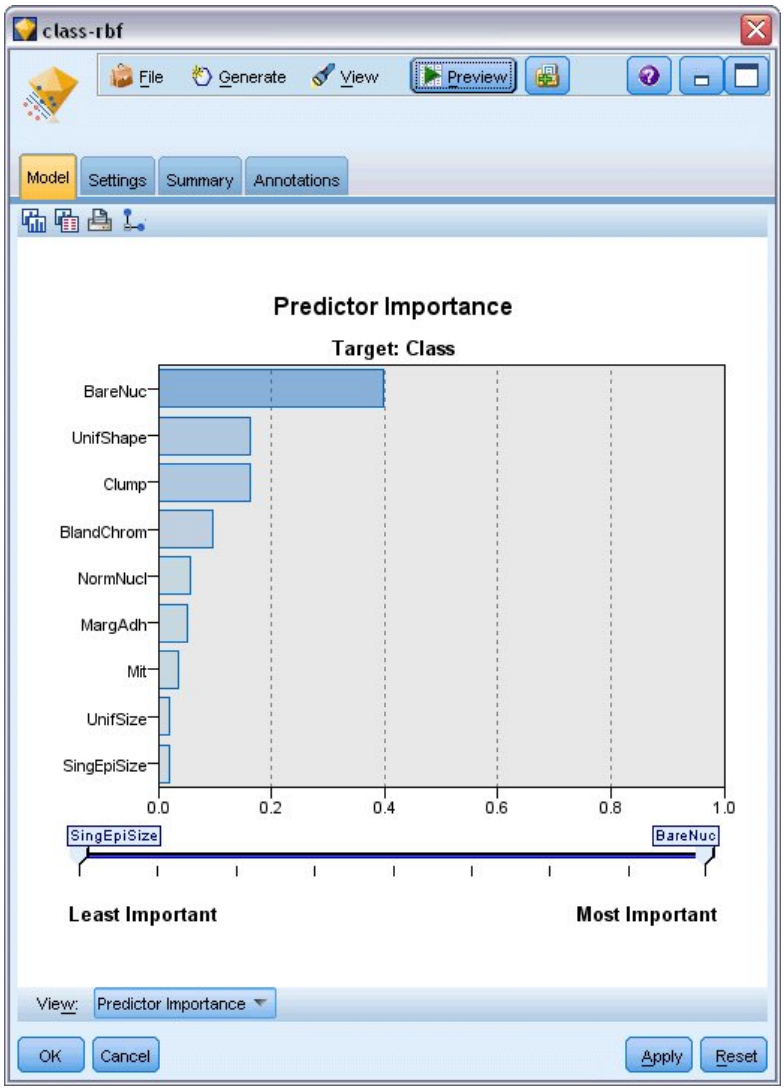
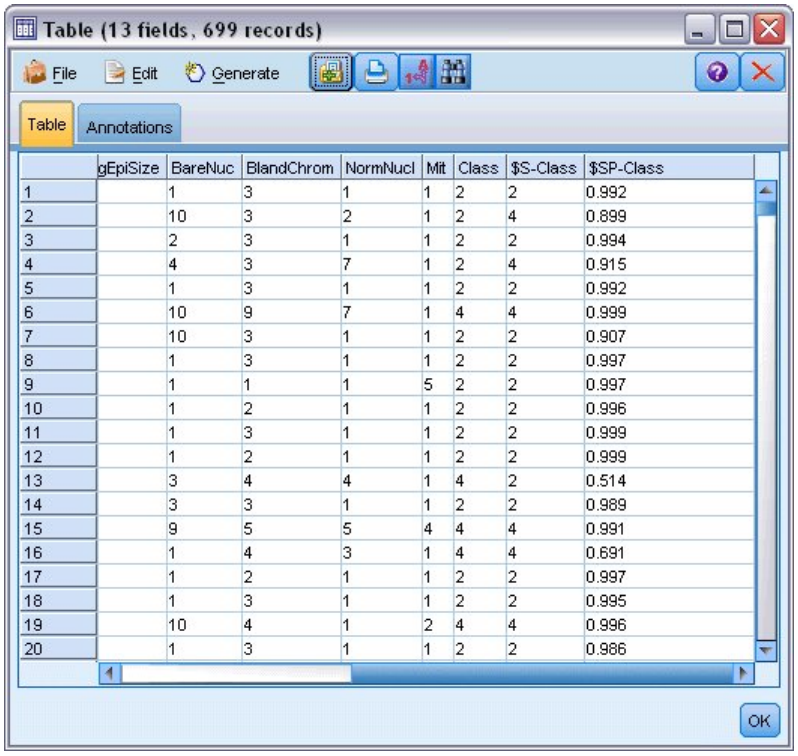


그림 *334.* 예측변수 중요도 그래프

모델 탭에서 예측변수 중요도 그래프는 예측에 대한 다양한 필드의 상대적인 효과를 표시합니다. 그래프에서 *BareNuc*이 가장 큰 효과를 갖고 있으며 동시에 *UnifShape* 및 *Clump*도 상당히 유의적임을 알 수 있습니다.

1. **확인**을 클릭하십시오.
2. 테이블 노드를 *class-rbf* 모델 너깃에 연결하십시오.
3. 테이블 노드를 열고 **실행**을 클릭하십시오.



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

그림 *335.* 예측 및 신뢰도 값에 대해 추가된 필드

4. 모델이 두 개의 추가 필드를 작성했습니다. 테이블 출력 오른쪽으로 스크롤하여 볼 수 있습니다.

새 필드 이름	설명
$S-Class	모델에 의해 예측된 *Class*의 값입니다.
$SP-Class	이 예측에 대한 성향 스코어입니다. (이 예측이 참이 될 우도이며 값은 0.0에서 1.0까지입니다.)

테이블을 보기만 하면 대부분의 레코드에 대해 성향 스코어(*$SP-Class* 열 내)가 상당히 높음을 알 수 있습니다.

그러나 몇 가지 유의적인 예외가 있습니다. 예를 들어, 13줄의 환자 1041801에 대한 레코드에서 0.514 값은 허용할 수 없을 정도로 낮습니다. 또한 *Class*를 *$S-Class*와 비교하면 이 모델이 2줄 및 4줄에서 보듯이 성향 스코어가 상대적으로 높더라도 수많은 잘못된 예측을 수행했음이 명확합니다.

다른 함수 유형을 선택하여 더 잘 수행할 수 있는지 알아 보려고 합니다.

다른 함수 시도

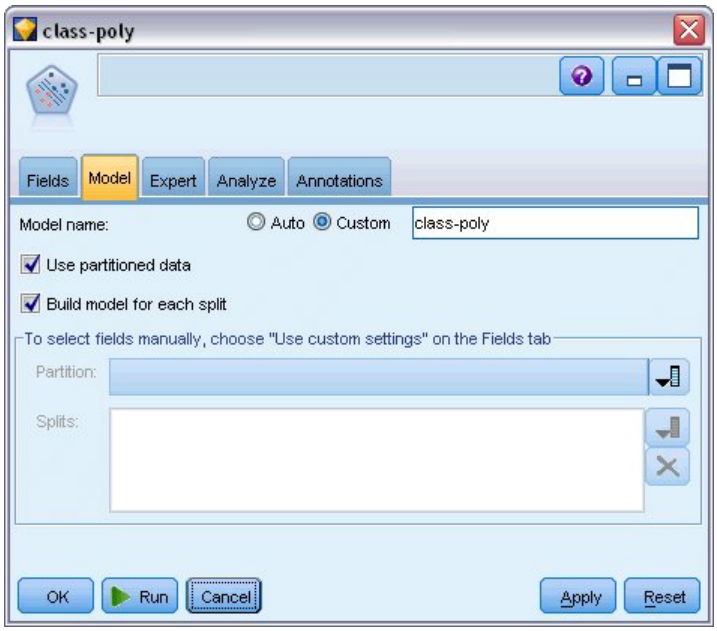


그림 *336.* 모델에 대한 새 이름 설정

1. 테이블 출력 창을 닫으십시오.
2. 두 번째 SVM 모델링 노드를 유형 노드에 연결하십시오.
3. 새 SVM 노드를 여십시오.
4. **모델** 탭에서 사용자 정의 및 *class-poly* 유형을 모델 이름으로 선택하십시오.

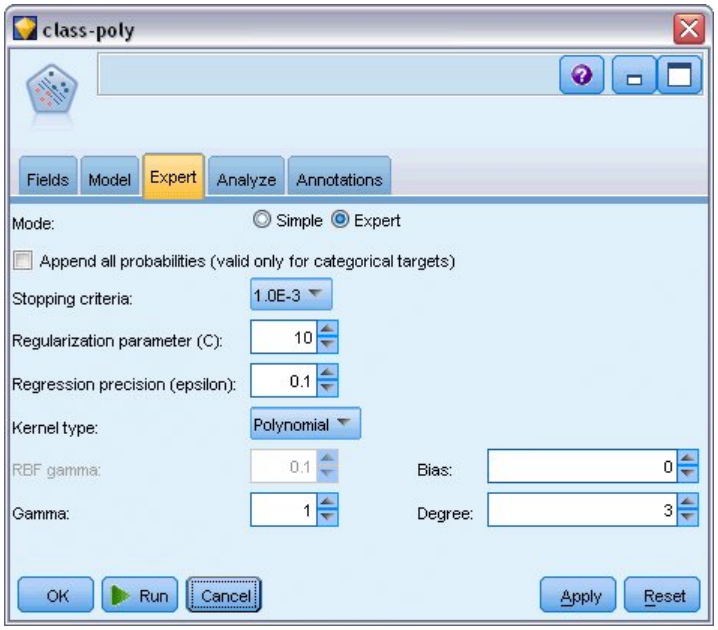


그림 *337.* 다항에 대한 전문가 탭 설정

5. **전문가** 탭에서 **모드**를 **전문가**로 설정하십시오.
6. **커널 유형**을 **다항**으로 설정하고 **실행**을 클릭하십시오. *class-poly* 모델 너깃이 스트림에 추가되고 오른쪽 상단 코너에 있는 모델 팔레트에도 추가됩니다.
7. *class-rbf* 모델 너깃을 *class-poly* 모델 너깃에 연결하십시오(경고 대화 상자에서 **바꾸기**를 선택하십시오).
8. 테이블 노드를 *class-poly* 너깃에 연결하십시오.
9. 테이블 노드를 열고 **실행**을 클릭하십시오.

결과 비교

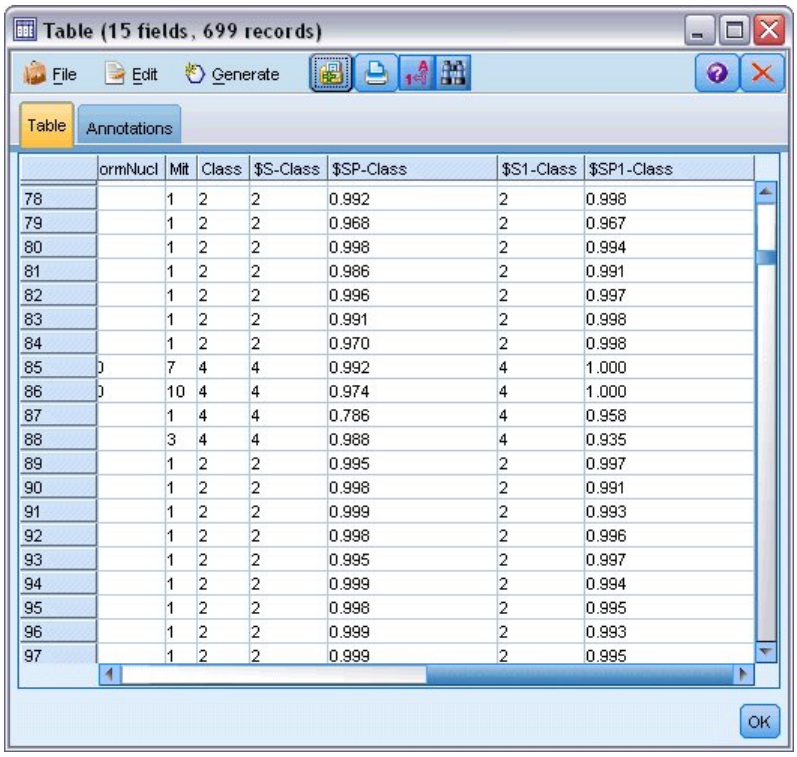


그림 *338.* 다항 함수에 대해 추가된 필드

1. 테이블 출력 오른쪽으로 스크롤하여 새로 추가된 필드를 보십시오.

   다항 함수 유형에 대해 생성된 필드는 *$S1-Class* 및 *$SP1-Class*로 이름이 지정됩니다.

   다항 결과는 훨씬 나아 보입니다. 많은 성향 스코어가 0.995 이상이며 이는 매우 고무적입니다.
2. 모델에서 개선도를 확인하려면 분석 노드를 *class-poly* 모델 너깃에 연결하십시오.

분석 노드를 열고 **실행**을 클릭하십시오.

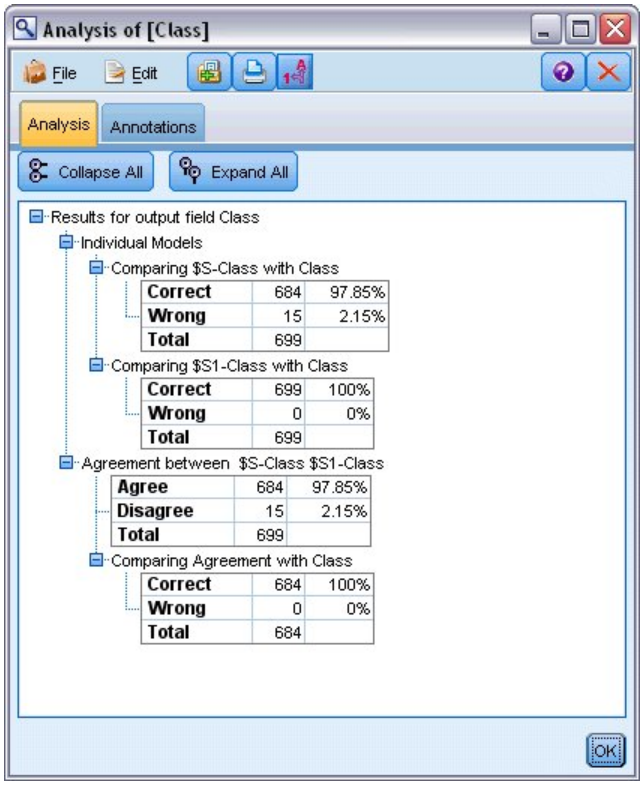


그림 *339.* 분석 노드

분석 노드를 사용하는 이 기술을 통해 동일한 유형의 두 개 이상의 모델 너깃을 비교할 수 있습니다. 분석 노드의 출력은 RBF 함수가 케이스의 97.85%를 올바르게 예측했음을 표시하며 이는 여전히 상당히 좋습니다. 그러나 출력에서는 다항 함수가 모든 단일 케이스에서 진단을 올바르게 예측했음을 표시합니다. 실제로 100% 정확도를 얻는 것은 거의 불가능하나 분석 노드를 사용하면 모델이 사용자의 특정 애플리케이션에 대해 허용 가능한 정확도인지 판별하는 데 도움을 얻을 수 있습니다.

사실상 이 특정 데이터 세트에 대해 다항 함수만큼 잘 수행하는 기타 함수 유형(시그모이드 및 선형)은 없습니다. 그러나 다른 데이터 세트의 경우, 결과가 다를 가능성이 높으므로 항상 전체 옵션 범위를 시도할 가치가 있습니다.

요약

수많은 속성으로부터 분류를 예측하기 위해 다양한 유형의 SVM 커널 함수를 사용했습니다. 서로 다른 커널이 동일한 데이터 세트에 대해 어떻게 다른 결과를 제공하는지, 어떻게 한 모델이 다른 모델에 대해 개선된 정도를 측정하는지 살펴보았습니다.

제 26 장 고객 이탈 시간을 모델링하기 위해 Cox 회귀분석 사용

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 기타 필드를 데이터베이스에서 끌어옵니다.

이 예에서는 *telco.sav*라는 데이터 파일을 참조하는 *telco_coxreg.str* 스트림을 사용합니다. 데이터 파일은 *Demos* 폴더에 있으며 스트림 파일은 *streams* 서브폴더에 있습니다. 자세한 정보는 5 페이지의 『Demos 폴더』의 내용을 참조하십시오.

적합한 모델 작성

1. *Demos* 폴더에서 *telco.sav* 파일을 가리키는 통계 파일 소스 노드를 추가하십시오.

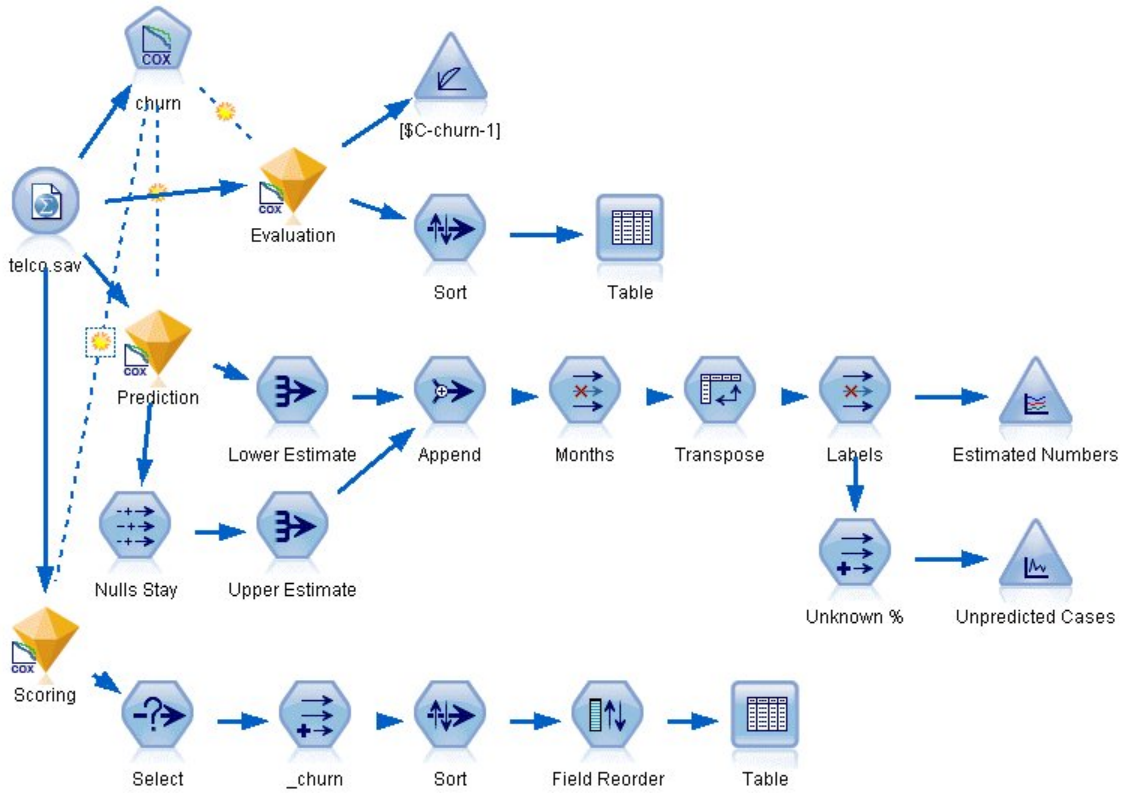


그림 *340.* 이탈 시간을 분석하기 위한 샘플 스트림

2. 소스 노드의 필터 탭에서 *region*, *income*, *longten*에서 *wireten*까지, *loglong*에서 *logwire*까지의 필드를 제외하십시오.

그림 *341.* 불필요한 필드 필터링

(또는 불필요한 필드를 제외하지 않고 유형 탭에서 해당 필드에 대한 역할을 **없음**으로 변경하거나 모델링 노드에서 사용할 필드를 선택할 수 있습니다.)

3. 소스 노드의 유형 탭에서 *churn* 필드에 대한 역할을 **대상**으로 설정하고 측정 수준을 **플래그**로 설정하십시오. 모든 기타 필드의 역할은 **입력**으로 설정해야 합니다.
4. **값 읽기**를 클릭하여 데이터를 인스턴스화하십시오.

그림 *342.* 필드 역할 설정

5. Cox 노드를 소스 노드에 연결하십시오. 즉, **필드** 탭에서 생존 시간 변수로 *tenure*를 선택하십시오.

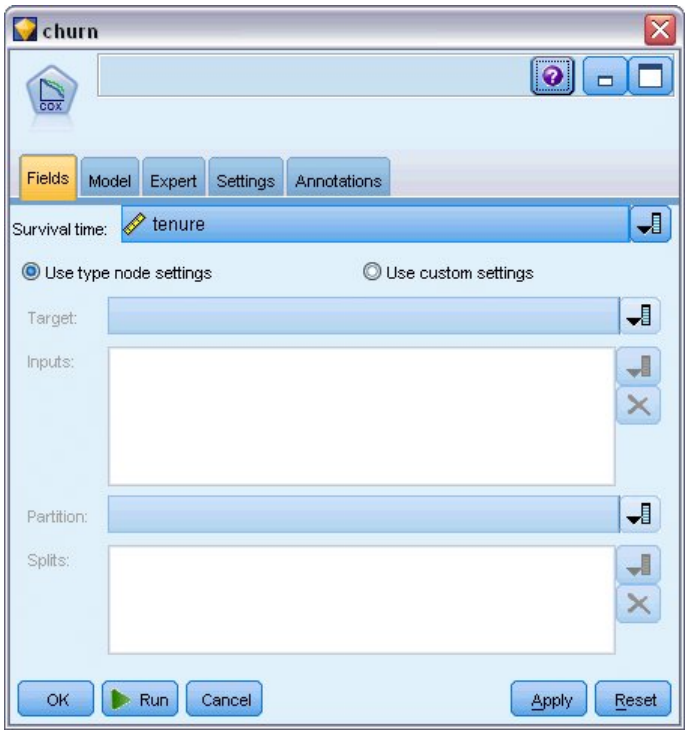


그림 *343.* 필드 옵션 선택

6. **모델** 탭을 클릭하십시오.
7. 변수 선택 방법으로 **단계 선택**을 선택하십시오.

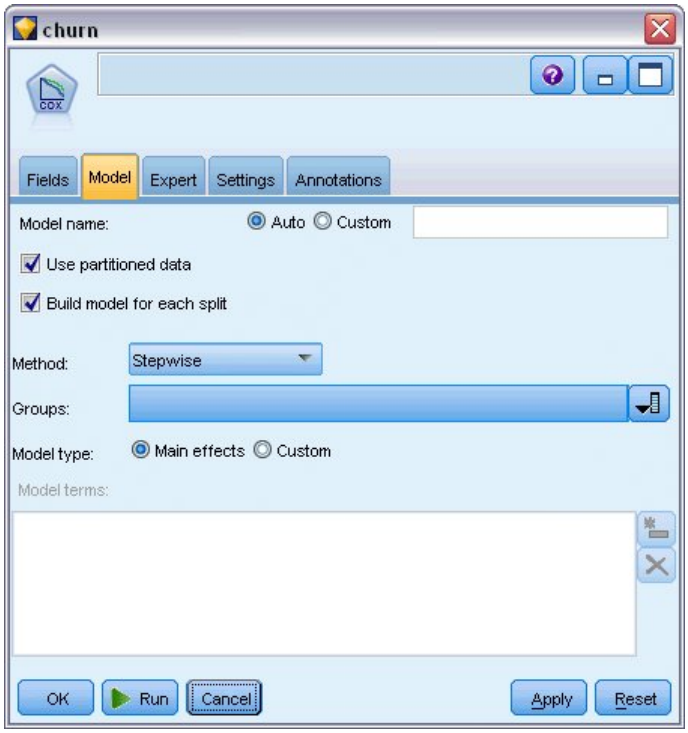


그림 *344.* 모델 옵션 선택

8. **전문가** 탭을 클릭하고 **전문가**를 선택하여 전문가 모델링 옵션을 활성화하십시오.
9. **출력**을 클릭하십시오.

그림 *345.* 고급 출력 옵션 선택

10. 생성할 도표로 **생존함수** 및 **위험함수**를 선택한 다음 **확인**을 클릭하십시오.
11. **실행**을 클릭하여 모델 너깃을 작성하십시오. 이는 오른쪽 상단 코너의 스트림 및 모델 팔레트에 추가됩니다. 세부사항을 보려면 스트림에서 너깃을 두 번 클릭하십시오. 먼저 고급 출력 탭을 보십시오.

중도절단 케이스

		N	Percent
Cases available in analysis	Event[a]	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

그림 *346.* 케이스 처리 요약

상태변수는 지정된 케이스에서 이벤트가 발생했는지 여부를 식별합니다. 이벤트가 발생하지 않은 경우, 케이스가 중도절단되었다고 합니다. 중도절단 케이스는 회귀계수 계산에는 사용되지 않으나 기준선 위험함수 계산에는 사용됩니다. 케이스 처리 요약은 726개의 케이스가 중도절단되었음을 표시합니다. 이러한 케이스는 이탈하지 않은 고객입니다.

범주형 변수 코딩

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			
tollfree[a]	0=No	526	1			
	1=Yes	474	0			
equip[a]	0=No	614	1			
	1=Yes	386	0			
callcard[a]	0=No	322	1			
	1=Yes	678	0			
wireless[a]	0=No	704	1			
	1=Yes	296	0			
multline[a]	0=No	525	1			
	1=Yes	475	0			
voice[a]	0=No	696	1			
	1=Yes	304	0			
pager[a]	0=No	739	1			
	1=Yes	261	0			
internet[a]	0=No	632	1			
	1=Yes	368	0			
callid[a]	0=No	519	1			
	1=Yes	481	0			
callwait[a]	0=No	515	1			
	1=Yes	485	0			
forward[a]	0=No	507	1			
	1=Yes	493	0			
confer[a]	0=No	498	1			
	1=Yes	502	0			
ebill[a]	0=No	629	1			
	1=Yes	371	0			
custcat[a]	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

그림 *347.* 범주형 변수 코딩

범주형 변수 코딩은 범주형 공변량, 특히 이분형 변수에 대한 회귀계수를 해석하는 데 유용한 참조입니다. 기본적으로 참조범주는 "마지막" 범주입니다. 따라서, 예를 들어, 기혼 고객이 데이터 파일에서 변수값 1을 갖더라도 회귀분석 목적으로는 0으로 코딩됩니다.

변수 선택

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2[b]	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3[c]	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4[d]	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5[e]	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6[f]	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7[g]	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8[h]	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9[i]	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10[j]	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11[k]	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12[l]	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

그림 *348.* 총괄 검정

모델 작성 프로세스는 단계별 전진 알고리즘을 사용합니다. 총괄 검정은 모델이 얼마나 잘 수행하는지에 대한 측도입니다. 카이제곱 이전 단계와의 상대적 변화는 이전 단계 및 현재 단계에서의 모델의 -2 로그-우도 간의 차분입니다. 단계가 변수에 추가되는 경우, 변경 유의성이 0.05 미만이면 포함이 의미가 있습니다. 단계가 변수에서 제거되는 경우, 변경 유의성이 0.10을 초과하면 제외가 의미가 있습니다. 12단계에서 12번째 변수가 모델에 추가됩니다.

		B	SE	Wald	df	Sig.	Exp(B)
Step 12	address	-.035	.009	14.543	1	.000	.966
	employ	-.051	.010	25.767	1	.000	.950
	reside	-.103	.046	5.037	1	.025	.902
	equip	-1.948	.381	26.180	1	.000	.143
	callcard	.777	.151	26.451	1	.000	2.175
	longmon	-.233	.022	115.619	1	.000	.792
	equipmon	-.042	.011	15.377	1	.000	.959
	multline	.612	.145	17.854	1	.000	1.844
	voice	-.501	.157	10.197	1	.001	.606
	internet	-.362	.160	5.114	1	.024	.697
	callid	-.464	.148	9.790	1	.002	.629
	ebill	-.399	.156	6.557	1	.010	.671

그림 *349.* 방정식의 변수*(12*단계에만 해당됨*)*

최종 모델에는 *address*, *employ*, *reside*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid* 및 *ebill*이 포함됩니다. 개별 예측변수의 효과를 이해하려면 예측변수에서 단위 증가량에 대한 위험함수 내에서의 예측된 변경으로 해석할 수 있는 Exp(B)를 보십시오.

- *address*에 대한 Exp(B)의 값은 이탈 위험이 고객이 동일한 주소에서 거주하는 각 해마다 100%−(100%×0.966)=3.4%씩 감소함을 의미합니다. 5년 동안 동일한 주소에서 거주하는 고객에 대한 이탈 위험은 100%−(100%×0.966^5)=15.88%만큼 감소합니다.
- *callcard*에 대한 Exp(B)의 값은 전화 카드 서비스에 가입하지 않는 고객의 이탈 위험이 서비스에 가입한 고객의 2.175배임을 의미합니다. 회귀분석에 대해 *No* = 1인 범주형 변수 코딩에서 상기하십시오.
- *internet*에 대한 Exp(B)의 값은 인터넷 서비스에 가입하지 않는 고객의 이탈 위험이 서비스에 가입한 고객의 0.697배임을 의미합니다. 이는 서비스에 가입한 고객이 서비스에 가입하지 않은 고객보다 더 빠르게 회사를 떠나고 있음을 의미하므로 다소 우려되는 일입니다.

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	ed	6.328	4	.176
	ed(1)	.007	1	.934
	ed(2)	.203	1	.652
	ed(3)	.835	1	.361
	ed(4)	5.773	1	.016
	retire	.013	1	.908
	gender	.214	1	.644
	tollfree	3.243	1	.072
	wireless	.668	1	.414
	tollmon	.000	1	.987
	cardmon	3.163	1	.075
	wiremon	1.084	1	.298
	pager	1.808	1	.179
	callwait	.266	1	.606
	forward	2.201	1	.138
	confer	2.568	1	.109
	custcat	.864	3	.834
	custcat(1)	.466	1	.495
	custcat(2)	.450	1	.502
	custcat(3)	.019	1	.889

그림 *350.* 모델에 없는 변수*(12*단계에만 해당됨*)*

모델 밖의 변수는 모드 유의수준이 0.05보다 큰 스코어 통계량을 갖습니다. 단, *tollfree* 및 *cardmon*에 대한 유의수준은 0.05 미만이나 상당히 근접합니다. 추가 연구에서 이를 추적하면 흥미로운 결과가 나올 것 같습니다.

공변량 평균값

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
ed(1)	.204
ed(2)	.287
ed(3)	.209
ed(4)	.234
employ	10.987
retire	.953
gender	.483
reside	2.331
tollfree	.526
equip	.614
callcard	.322
wireless	.704
longmon	11.723
tollmon	13.274
equipmon	14.220
cardmon	13.781
wiremon	11.584
multline	.525
voice	.696
pager	.739
internet	.632
callid	.519
callwait	.515
forward	.507
confer	.498
ebill	.629
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

그림 *351.* 공변량 평균값

이 테이블은 각 예측변수의 평균값을 표시합니다. 이 테이블은 평균값에 대해 구성되며 생존 도표를 볼 때 유용한 참조입니다. 그러나 범주형 예측변수에 대한 표시자 변수의 평균값을 볼 때 실제로 "평균" 고객이 존재하지는 않습니다. 모든 척도 예측변수를 사용하는 경우에도 공변량 값이 모두 평균에 가까운 고객을 찾지 못할 수도 있습니다. 특정 케이스에 대한 생존 곡선을 보려면 도표 대화 상자에서 생존 곡선이 도표로 작성되는 공변량 값을 변경할 수 있습니다. 특정 케이스에 대한 생존 곡선을 보려면 고급 출력 대화 상자의 도표 그룹에서 생존 곡선이 도표로 작성되는 공변량 값을 변경할 수 있습니다.

생존함수 곡선

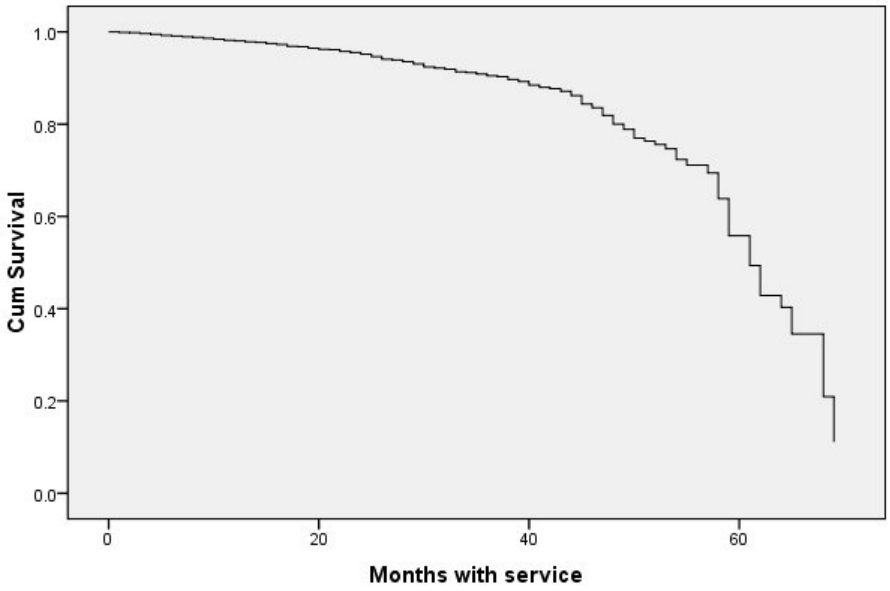


그림 *352.* ″평균″ 고객에 대한 생존함수 곡선

기본 생존함수 곡선은 모델의 시각적 표시이며 "평균" 고객에 대한 이탈 시간을 예측합니다. 수평축 변수는 이벤트 발생 시간을 표시합니다. 수직축 변수는 생존 확률을 표시합니다. 따라서 생존함수 곡선의 임의의 점은 "평균" 고객이 해당 시간이 시간 후에 고객으로 남을 확률을 표시합니다. 지난 55개월 동안 생존함수 곡선이 덜 평활하게 됩니다. 그렇게 긴 시간 동안 회사에 남아 있는 고객이 거의 없으므로 가능한 정보가 적고 이에 따라 곡선이 장방형이 됩니다.

위험함수 곡선

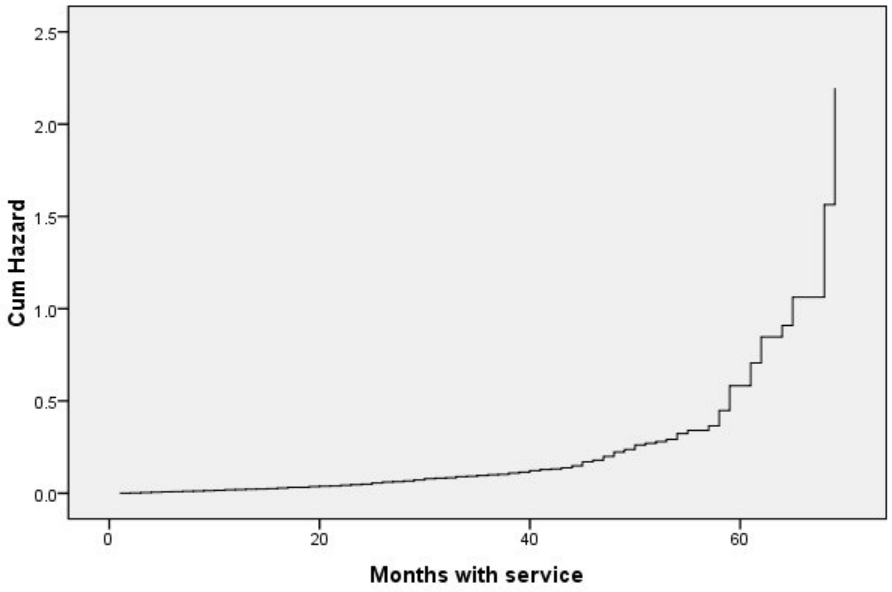


그림 *353.* ″평균″ 고객에 대한 위험함수 곡선

기본 위험함수 곡선은 누적 모델의 시각적 표시이며 "평균" 고객에 대한 잠재적인 이탈을 예측합니다. 수평축 변수는 이벤트 발생 시간을 표시합니다. 수직 축은 생존 확률의 음수 로그와 동일한 누적 위험을 표시합니다. 지난 55개월 동안 위험함수 곡선은 생존함수 곡선과 같이 동일한 원인으로 덜 평활해졌습니다.

평가

단계 선택법을 사용하면 모델이 "통계적으로 유의적인" 예측변수만 갖게 되나 모델이 실제로 목표 예측에 좋은지는 확신할 수 없습니다. 확신하려면 스코어링된 레코드를 분석해야 합니다.

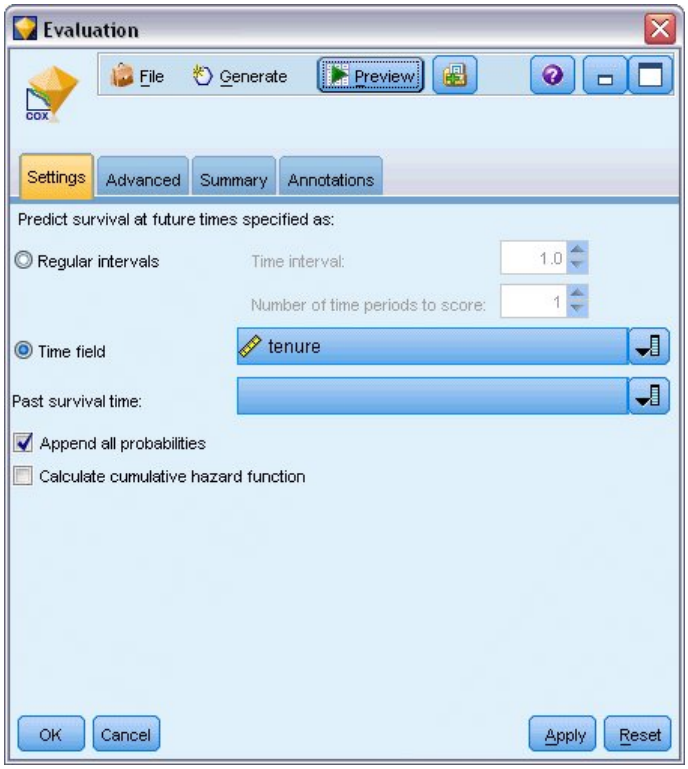


그림 *354. Cox* 너깃*:* 설정 탭

1. 모델 너깃을 캔버스에 두고 이를 소스 노드에 연결한 다음 너깃을 열고 설정 탭을 클릭하십시오.
2. **시간 필드**를 선택하고 *tenure*를 지정하십시오. 각 레코드가 재직 기간에 따라 스코어링됩니다.
3. **모든 확률 추가**를 선택하십시오.

   그러면 고객이 이탈하는지 여부에 따라 분리점으로 0.5를 사용하여 점수를 생성합니다. 이탈 성향이 0.5가 넘으면 이탈자로 스코어링됩니다. 이 숫자에 어떠한 특이한 점이 있는 것은 아니며 다른 분리점이 더 바람직한 결과를 생성할 수도 있습니다. 분리점을 선택하는 데 대한 한 방법으로 평가 노드를 사용하십시오.

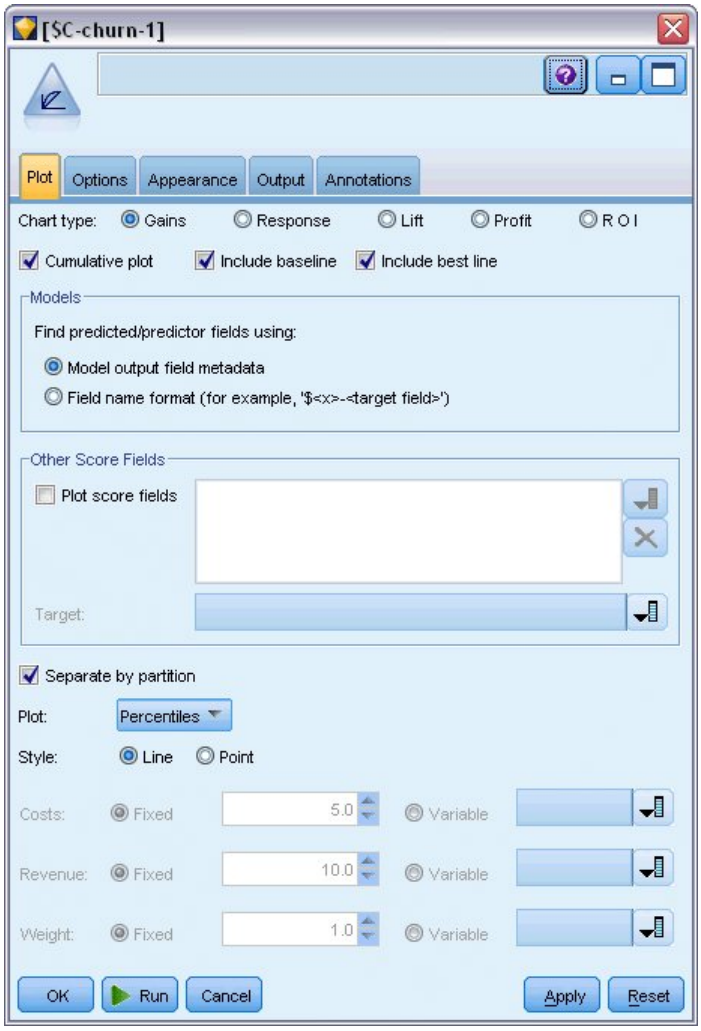


그림 *355.* 평가 노드*:* 도표 탭

4. 평가 노드를 모델 너깃에 연결하십시오. 도표 탭에서 **최적 예측선 포함**을 선택하십시오.
5. **옵션** 탭을 클릭하십시오.

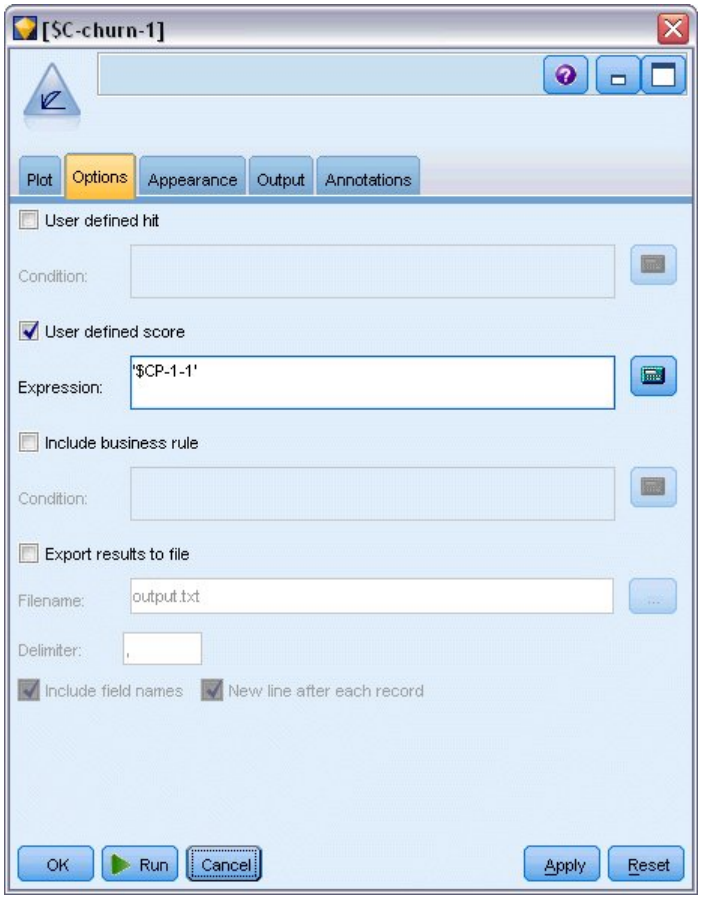


그림 *356.* 평가 노드*:* 옵션 탭

6. **사용자 정의 스코어** 및 `'$CP-1-1'` 유형을 표현식으로 선택하십시오. 이는 이탈 성향에 대해 모델에서 생성한 필드입니다.
7. **실행**을 클릭하십시오.

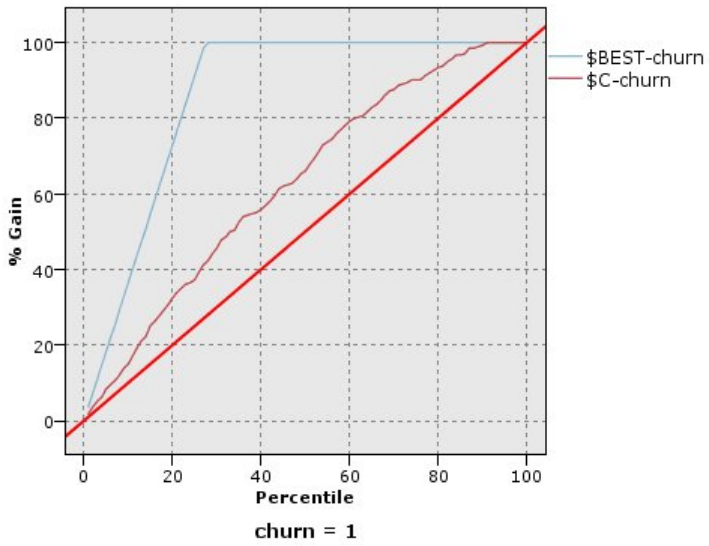


그림 *357.* 이익 차트

누적 이익 차트는 케이스의 총 수의 퍼센트를 대상화하여 "얻은" 지정된 범주 내의 전체 케이스 수의 퍼센트를 표시합니다. 예를 들어, 곡선 위의 한 점은 (10%, 15%)에 있으며 사용자가 모델을 사용하여 데이터 세트를 스코어링하고 예측된 이탈 성향으로 모든 케이스를 정렬한 경우, 상위 10%가 실제로 범주 *1*(이탈자)에 해당하는 전체 케이스의 약 15%를 포함함을 예상할 수 있습니다. 이와 마찬가지로 상위 60%가 약 79.2%의 이탈자를 포함합니다. 스코어링된 데이터 세트의 100%를 선택하면 모든 이탈자가 데이터 세트 내에 포함됩니다.

대각선은 "기준선" 곡선입니다. 스코어링된 데이터 세트에서 무선적으로 20%의 레코드를 선택하면 실제로 범주 *1*에 해당하는 모든 레코드 중 약 20%를 "얻을" 것으로 예상할 수 있습니다. 곡선이 기준선이 위쪽에 있을 수록 더 많은 레코드를 얻습니다. "최고"의 선은 모든 이탈자에 대한 이탈 성향 스코어가 모든 비이탈자보다 높게 지정되는 "완벽한" 모델에 대한 곡선을 보여줍니다. 누적 이익 차트를 사용하면 원하는 이익에 해당하는 퍼센트를 선택한 다음 해당 퍼센트를 적절한 절사 값에 맵핑하여 분류 분리점을 선택하는 데 도움을 받을 수 있습니다.

"바람직한" 이익을 구성하는 요소는 유형 I 및 유형 II 오류의 비용에 따라 다릅니다. 즉, 이탈자를 비이탈자로 분류하는(제 I 유형) 비용은 무엇입니까? 비이탈자를 이탈자로 분류하는(제 II 유형) 비용은 무엇입니까? 고객 유지가 주 관심사인 경우, 제 I 유형 오류를 낮추고자 할 것입니다. 누적 이익 도표에서 이는 예측 성향이 *1*인 상위 60%의 고객에 대해 고객 관리를 늘리는 것에 해당될 것입니다. 이는 가능한 이탈자의 79.2%의 이탈을 억제할 수 있으나 새 고객을 얻기 위해 사용할 수 있는 시간 및 자원이 비용이 됩니다. 현재 고객 기반을 유지하는 비용을 낮추는 것이 높은 우선 순위인 경우, 제 II 유형 오류를 낮추고자 할 것입니다. 차트에서 상위 20%의 고객 관리를 증가시키면 이탈자의 32.5%에서 이탈을 억제하는 것에 해당됩니다. 일반적으로는 둘 다 중요한 관심사이므로 민감도 및 특이도를 적절히 혼합한 고객 분류 의사결정 규칙을 선택해야 합니다.

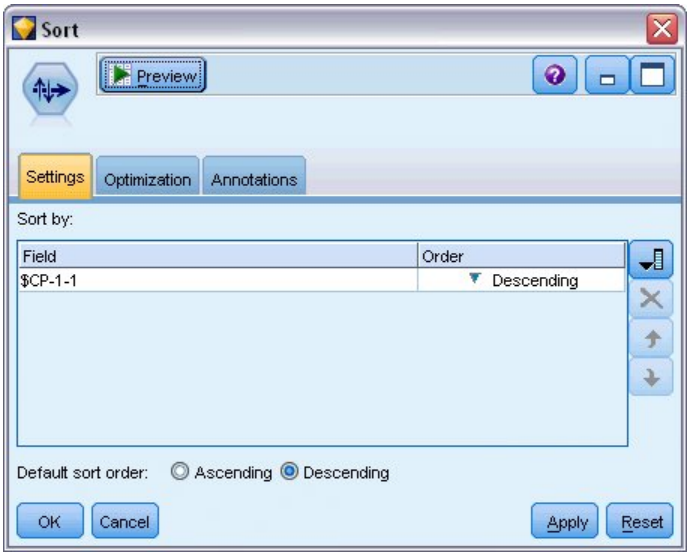


그림 *358.* 정렬 노드*:* 설정 탭

8. 45.6%가 원하는 이익이라고 할 때 이는 상위 30%의 레코드를 사용하는 것에 해당됩니다. 절절한 분류 분리점을 찾으려면 정렬 노드를 모델 너깃에 연결하십시오.
9. 설정 탭에서 내림차순으로 *$CP-1-1* 기준으로 정렬하고 **확인**을 클릭하십시오.

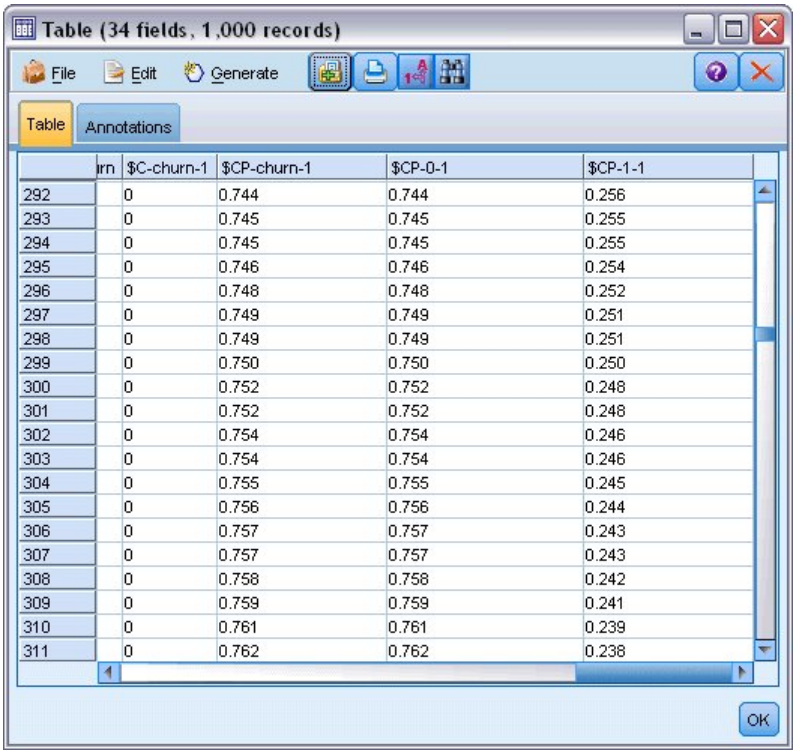


그림 *359.* 테이블

10. 테이블 노드를 정렬 노드에 연결하십시오.
11. 테이블 노드를 열고 **실행**을 클릭하십시오.

출력 아래로 스크롤하면 300번째 레코드에서 *$CP-1-1*의 값이 0.248인 것을 볼 수 있을 것입니다. 0.248을 분류 분리점으로 사용하면 고객 중 약 30%가 이탈자로 스코어링되고 실제 총 이탈자 중 약 45%의 이탈을 억제하는 결과가 발생해야 합니다.

보유 고객 예측 수 추적

일단 모델에 만족하면 데이터 세트에서 예측한 다음 2년에 걸쳐 보유되는 고객의 수를 추적하고자 할 것입니다. 총 가입 기간(이후 시간 + *tenure*)이 모델을 학습하는 데 사용된 데이터의 생존 시간 범위에 속하지 않는 고객인 널값은 흥미 있는 도전 과제를 제시합니다. 이를 처리하는 한 가지 방법은 두 예측변수군을 작성하여 한 예측변수군에서는 널값을 이탈한 것으로 간주하고 또 다른 예측변수군에서는 보유된 것으로 간주하는 것입니다. 이 방법으로 보유되는 고객의 예측 수를 기반으로 하여 상한 및 하한 경계를 설정할 수 있습니다.

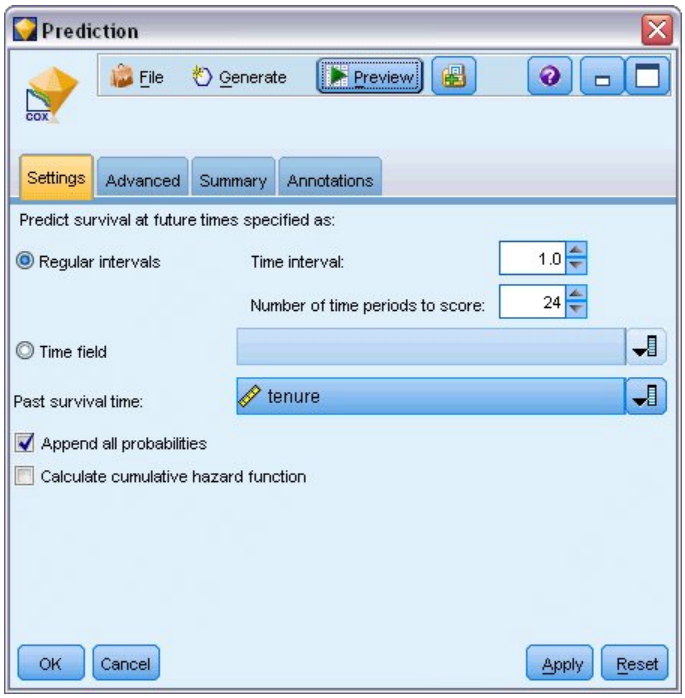


그림 *360. Cox* 너깃*:* 설정 탭

1. 모델 팔레트에서 모델 너깃을 두 번 클릭하거나 너깃을 복사하여 스트림 캔버스에 붙여넣고 새 너깃을 소스 노드에 연결하십시오.
2. 설정 탭에서 너깃을 여십시오.
3. **정상 구간**이 선택되어 있는지 확인하고 시간 간격으로 1.0을 지정하고 스코어링할 기간으로 24를 지정하십시오. 그러면 각 레코드가 다음 24개월 각각에 대해 스코어링됩니다.
4. 지난 생존 시간을 지정할 필드로 *tenure*를 선택하십시오. 스코어링 알고리즘이 각 고객의 해당 회사의 고객으로서의 시간 길이를 계산할 것입니다.
5. **모든 확률 추가**를 선택하십시오.

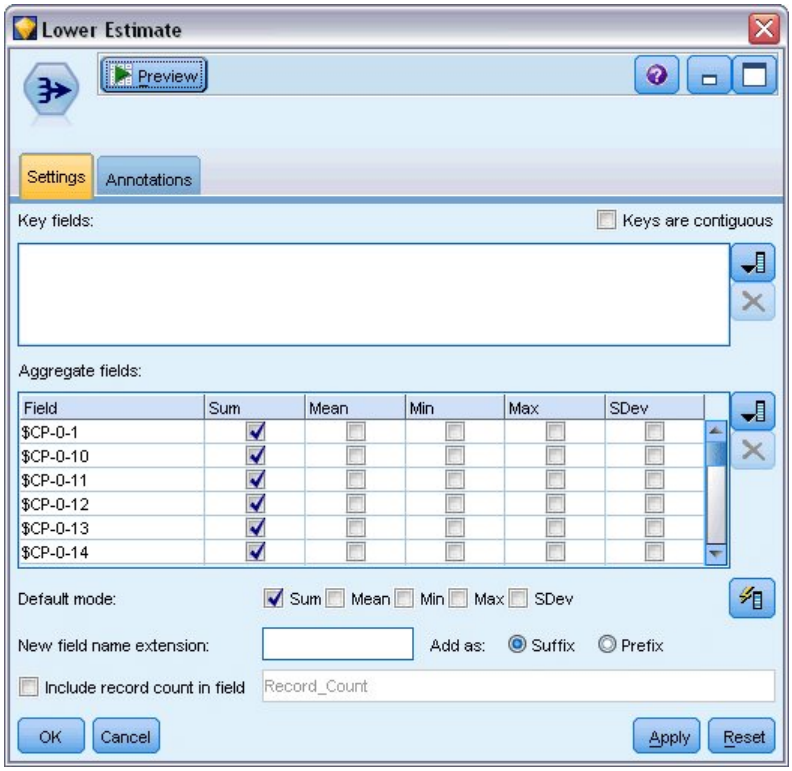


그림 *361.* 통합 노드*:* 설정 탭

6. 통합 노드를 모델 너깃에 연결하고 설정 탭에서 기본 모드로 **평균**을 선택 취소하십시오.
7. *$CP-0-n* 양식의 *$CP-0-1*에서 *$CP-0-24*까지의 필드를 통합할 필드로 선택하십시오. 필드 선택 대화 상자에 있으면 이름순(문자순)으로 필드를 정렬하는 것이 가장 쉬운 방법입니다.
8. **필드에 레코드 수 포함**을 선택 취소하십시오.
9. **확인**을 클릭하십시오. 이 노드는 "하한 경계" 예측을 작성합니다.

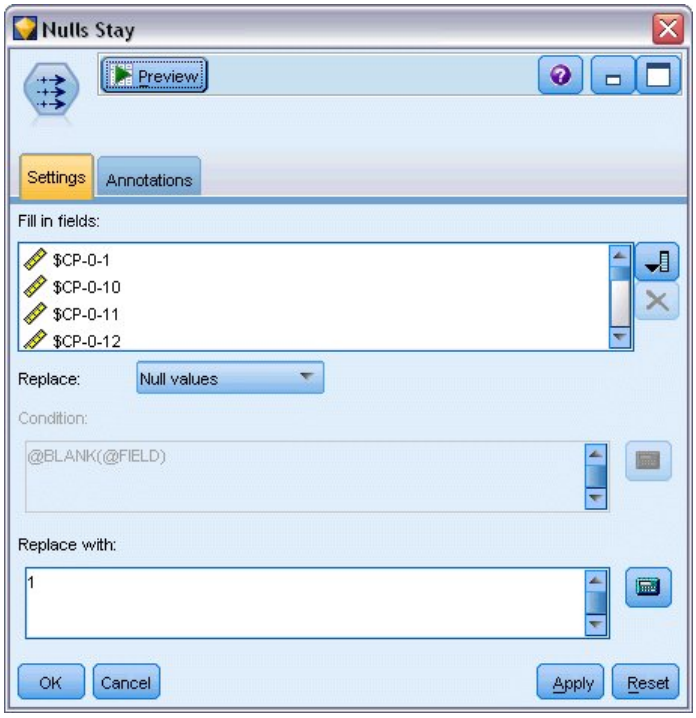


그림 *362.* 채움 노드*:* 설정 탭

10. 방금 통합 노드에 연결한 Coxreg 너깃에 채움 노드를 연결하십시오. 설정 탭에서 *$CP-0-n* 양식의 *$CP-0-1*에서 *$CP-0-24*까지의 필드를 채울 필드로 선택하십시오. 필드 선택 대화 상자에 있으면 이름순(문자순)으로 필드를 정렬하는 것이 가장 쉬운 방법입니다.
11. **널값**을 1값으로 바꾸도록 선택하십시오.
12. **확인**을 클릭하십시오.

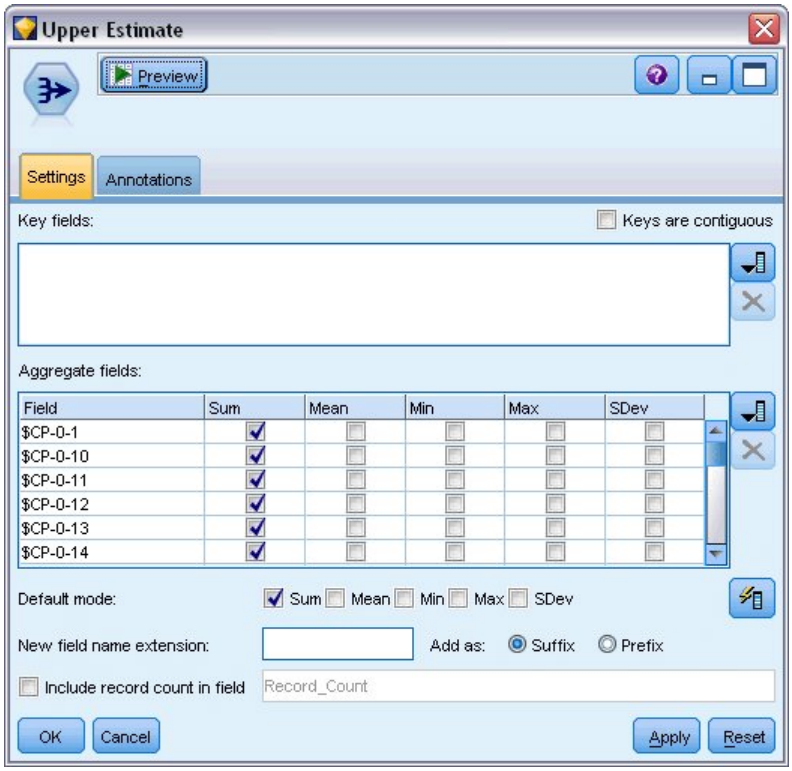


그림 *363.* 통합 노드*:* 설정 탭

13. 통합 노드를 채움 노드에 연결하고 설정 탭에서 기본 모드로 **평균**을 선택 취소하십시오.
14. *$CP-0-n* 양식의 *$CP-0-1*에서 *$CP-0-24*까지의 필드를 통합할 필드로 선택하십시오. 필드 선택 대화 상자에 있으면 이름순(문자순)으로 필드를 정렬하는 것이 가장 쉬운 방법입니다.
15. **필드에 레코드 수 포함**을 선택 취소하십시오.
16. **확인**을 클릭하십시오. 이 노드는 "상한 경계" 예측을 작성합니다.

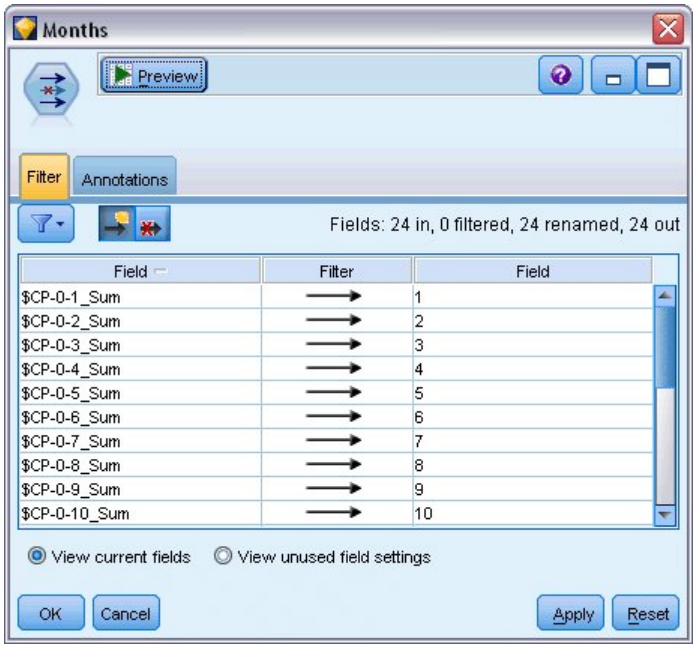


그림 *364.* 필터 노드*:* 설정 탭

17. 추가 노드를 두 개의 통합 노드에 연결한 다음 필터 노드를 추간 노드에 연결하십시오.
18. 필터 노드의 설정 탭에서 필드의 이름을 *1*에서 *24*까지로 변경하십시오. 전치 노드를 사용하여 이러한 필드 이름이 도표 다운스트림에서 *x*축에 대한 값이 됩니다.

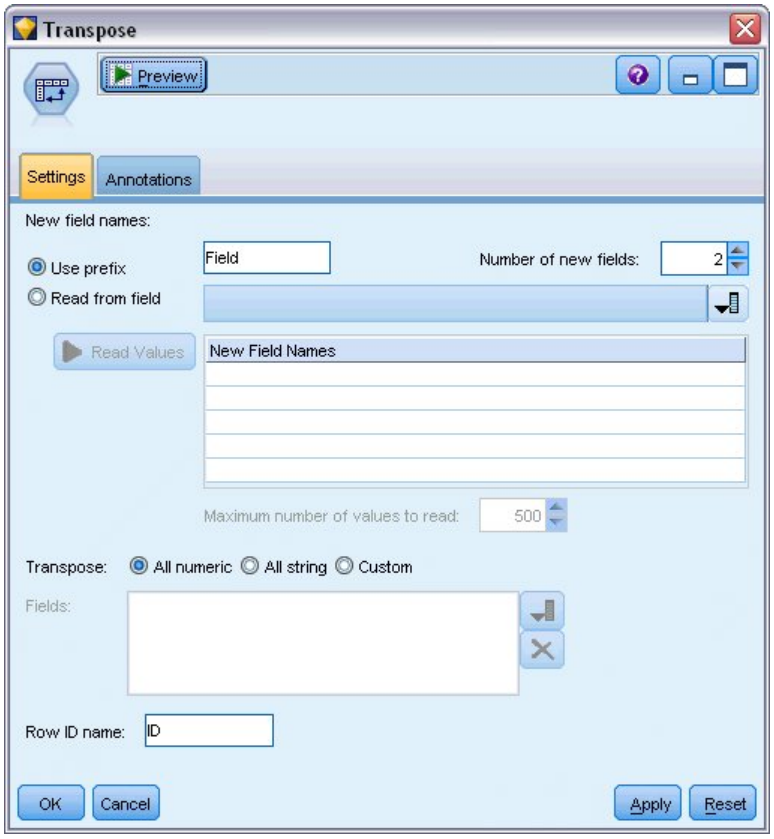


그림 *365.* 전치 노드*:* 설정 탭

19. 전치 노드를 필터 노드에 첨부하십시오.
20. 새 필드의 번호로 2를 입력하십시오.

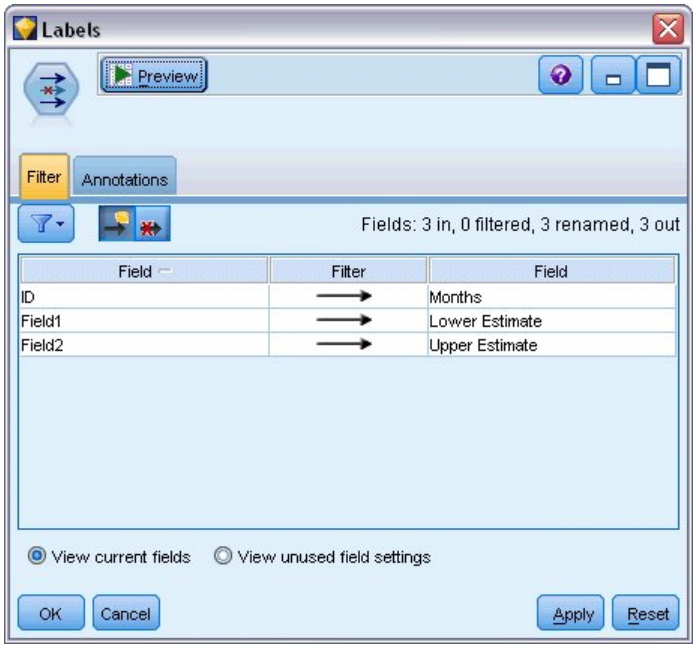


그림 *366.* 필터 노드*:* 필터 탭

21. 필터 노드를 전치 노드에 연결하십시오.
22. 필터 노드의 설정 탭에서 *ID*의 이름을 *Months*로, *Field1*의 이름을 *Lower Estimate*로, *Field2*의 이름을 *Upper Estimate*로 변경하십시오.

그림 *367.* 다중 도표 노드*:* 도표 탭

23. 다중 도표 노드를 필터 노드에 첨부하십시오.
24. 도표 탭에서 *Months*를 X 필드로, *Lower Estimate* 및 *Upper Estimate*를 Y 필드로 지정하십시오.

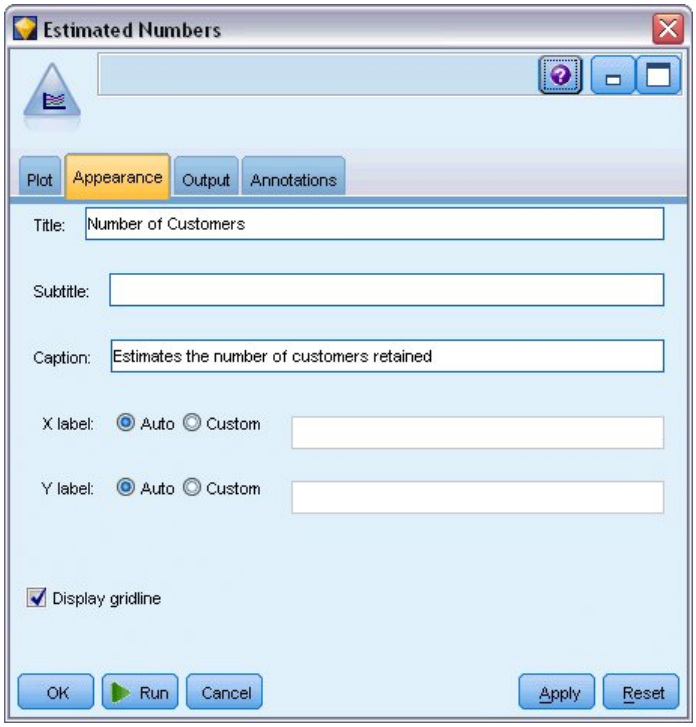


그림 *368.* 다중 도표 노드*:* 외형 탭

25. 외형 탭을 클릭하십시오.
26. 제목으로 고객 수를 입력하십시오.
27. 캡션으로 보유된 고객 수 추정을 입력하십시오.
28. **실행**을 클릭하십시오.

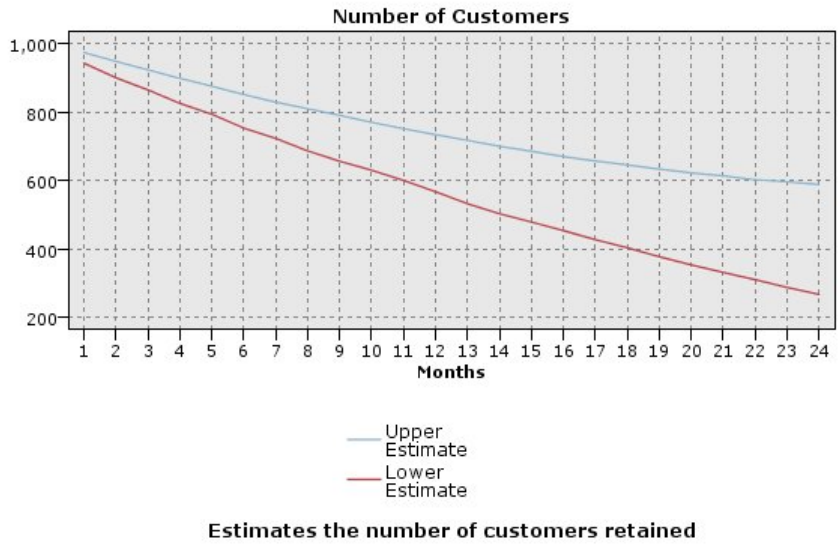


그림 *369.* 보유된 고객 수를 추정하는 다중 도표

보유된 고객의 추정 수에 대한 상한 및 하한 경계가 도표로 작성됩니다. 두 선 사이의 차이가 널로 스코어링되어 상태가 매우 불명확한 고객의 수입니다. 시간이 경과함에 따라 이러한 고객의 수가 증가합니다. 데이터 세트에서 12개월 이후에는 원래 고객의 601에서 735 사이가 보유될 것으로 예상됩니다. 개월 이후에는 288에서 597 사이입니다.

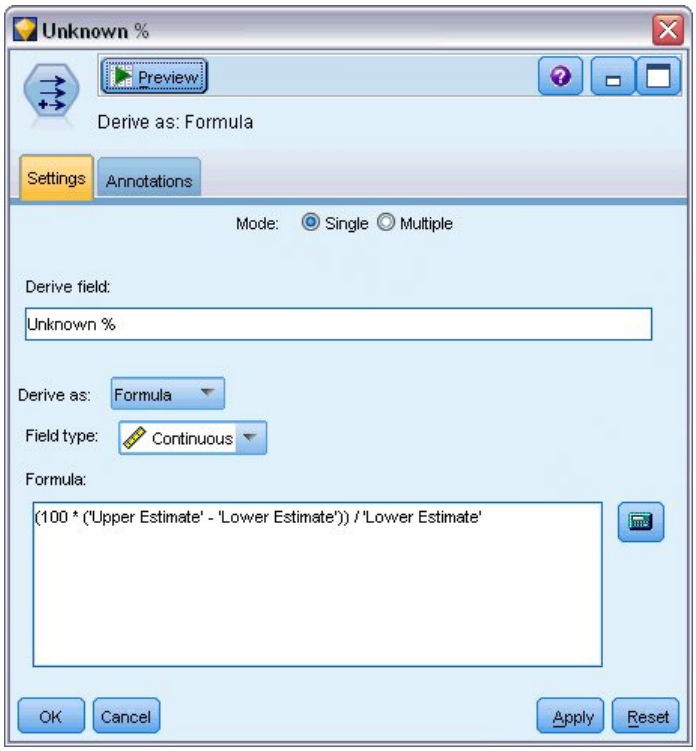


그림 *370.* 파생 노드: 설정 탭

29. 보유되는 고객의 수에 대한 추정값이 어느 정도로 불확실한지 보려면 파생 노드를 필터 노드에 연결하십시오.
30. 파생 노드의 설정 탭에서 파생 노드로 알 수 없는 %를 입력하십시오.
31. 필드 유형으로 **연속형**을 선택하십시오.
32. `(100 * ('상한 추정값' - '하한 추정값')) / '하한 추정값'`을 수식으로 입력하십시오. *Unknown* %는 "의심되는" 고객의 수를 하한 추정값의 퍼센트로 표시한 것입니다.
33. **확인**을 클릭하십시오.

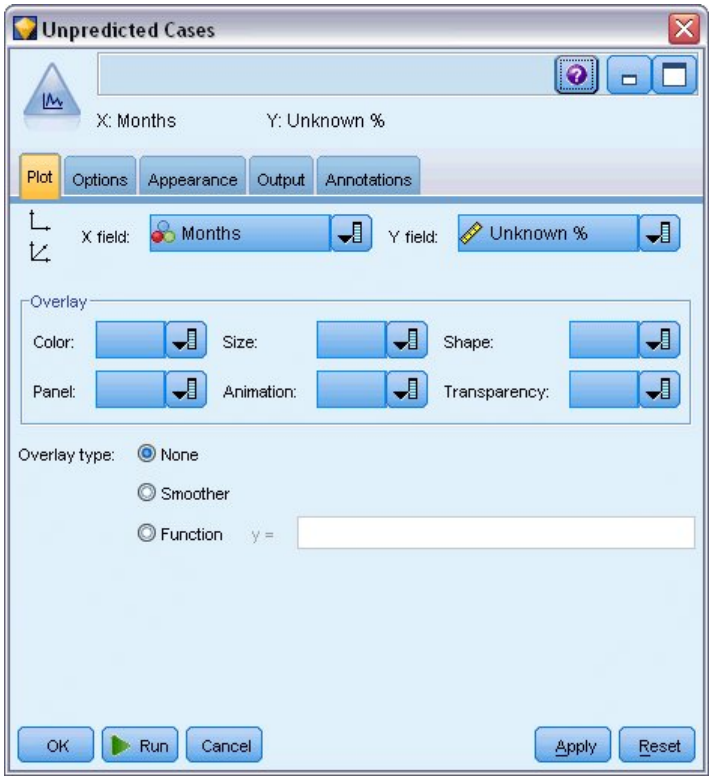


그림 *371. plot* 노드: 도표 탭

34. plot 노드를 파생 노드에 연결하십시오.
35. plot 노드의 도표 탭에서 *Months*를 X 필드로, *Unknown* %를 Y 필드로 선택하십시오.
36. **외형** 탭을 클릭하십시오.

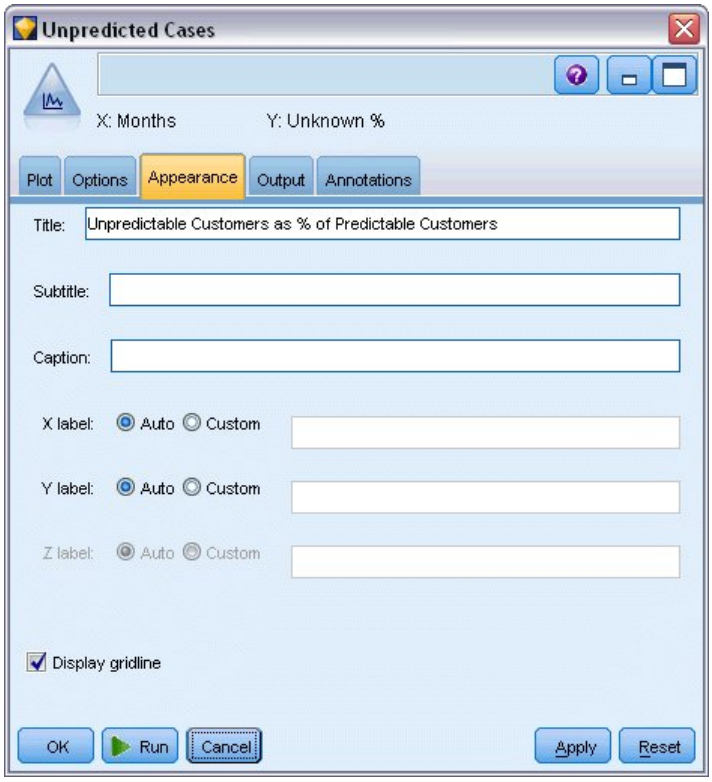


그림 *372. plot* 노드*:* 외형 탭

37. 제목으로 예측 가능한 고객에 대한 %로 표시한 예측 불가능한 고객을 입력하십시오.
38. 노드를 실행하십시오.

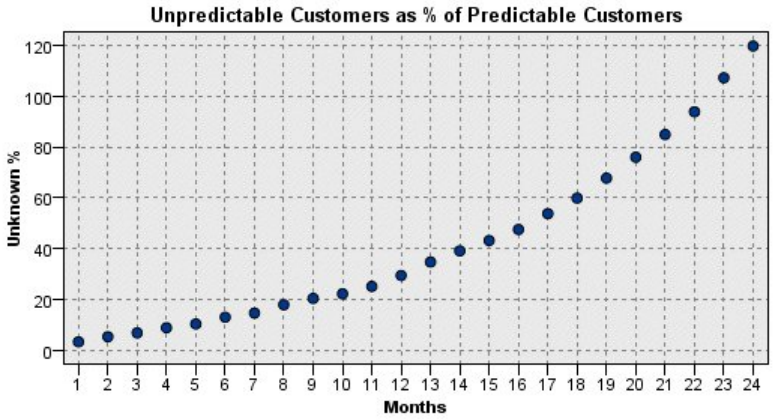


그림 *373.* 예측 불가능한 고객의 도표

첫 해 동안 예측 불가능한 고객의 퍼센트가 상당히 선형 비율로 증가하나 두 번째 해에서는 23개월까지 증가 비율이 급증하고 23개월에는 널 값을 가진 고객 수가 보유 고객의 예상 수를 능가합니다.

스코어링

일단 모델에 만족하면 내년에 이탈할 가능성이 가장 높은 개인을 분기별로 식별하기 위해 고객을 스코어링하고자 할 것입니다.

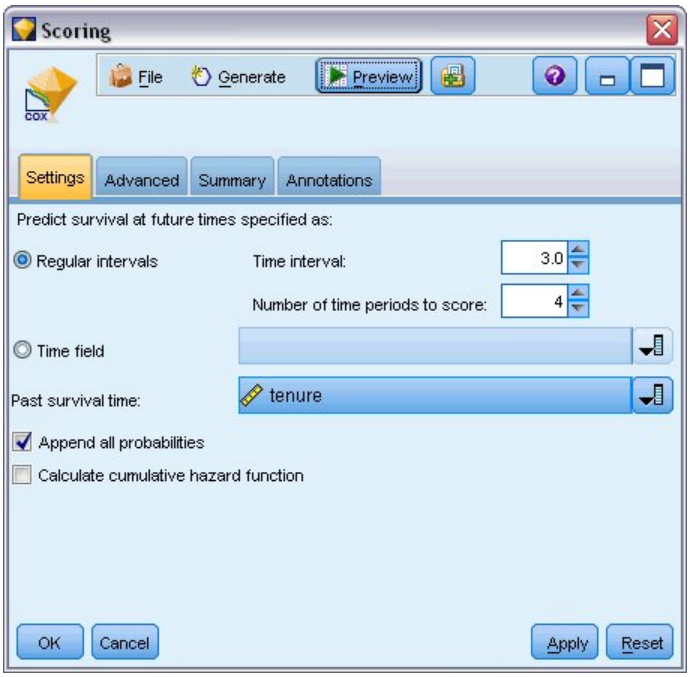


그림 *374. Coxreg* 너깃*:* 설정 탭

1. 세 번째 모델 너깃을 소스 노드에 연결하고 모델 너깃을 여십시오.
2. **정상 구간**이 선택되어 있는지 확인하고 시간 간격으로 3.0을 지정하고 스코어링할 기간으로 4를 지정하십시오. 그러면 각 레코드가 다음 4분기에 대해 스코어링됩니다.
3. 지난 생존 시간을 지정할 필드로 *tenure*를 선택하십시오. 스코어링 알고리즘이 각 고객의 해당 회사의 고객으로서의 시간 길이를 계산할 것입니다.
4. **모든 확률 추가**를 선택하십시오. 이러한 추가 필드로 인해 테이블에서 볼 레코드를 더 쉽게 정렬할 수 있습니다.

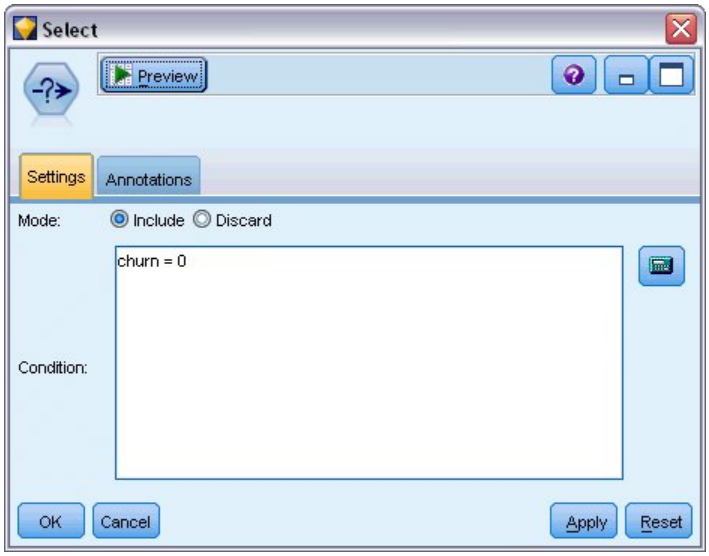


그림 *375.* 선택 노드*:* 설정 탭

5. 선택 노드를 모델 너깃에 연결하십시오. 설정 탭에서 조건으로 `churn=0`을 입력하십시오. 그러면 이미 이탈한 고객이 결과 테이블에서 제거됩니다.

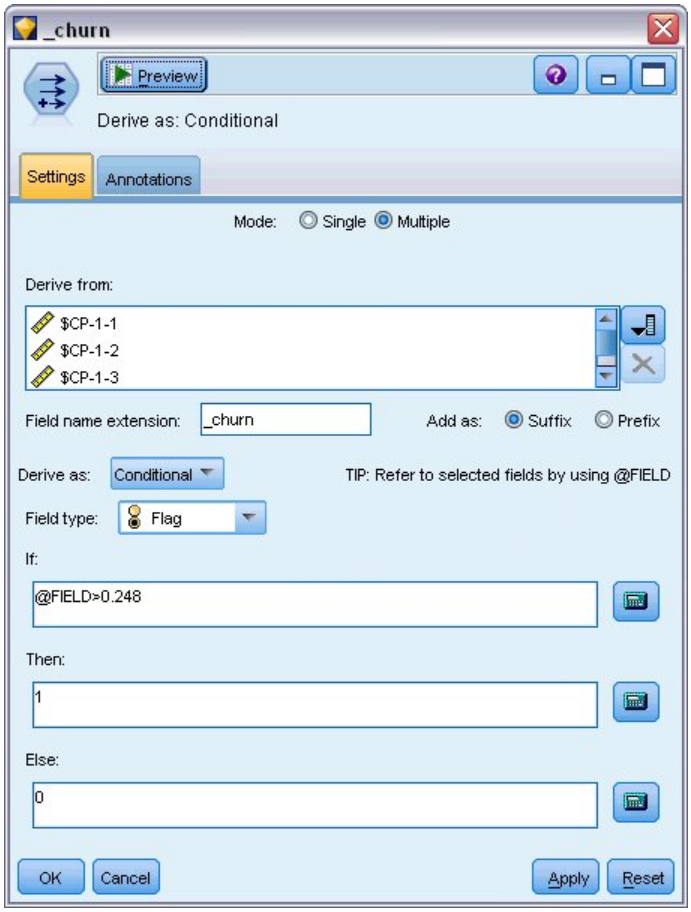


그림 *376.* 파생 노드*:* 설정 탭

6. 파생 노드를 선택 노드에 연결하십시오. 설정 탭에서 모델로 **다중**을 선택하십시오.
7. *$CP-1-1*에서 *$CP-1-4*까지, *$CP-1-n* 양식의 필드에서 파생하도록 선택하고 추가할 접미문자로 _churn을 입력하십시오. 필드 선택 대화 상자에 있으면 이름순(문자순)으로 필드를 정렬하는 것이 가장 쉬운 방법입니다.
8. **조건부**로 필드를 파생시키도록 선택하십시오.
9. 측정 수준으로 **플래그**를 선택하십시오.
10. **If** 조건으로 `@FIELD>0.248`을 입력하십시오. 이는 평가 동안 식별된 분류 분리점입니다.
11. **Then** 표현식으로 `1`을 입력하십시오.
12. **Else** 표현식으로 `0`을 입력하십시오.
13. **확인**을 클릭하십시오.

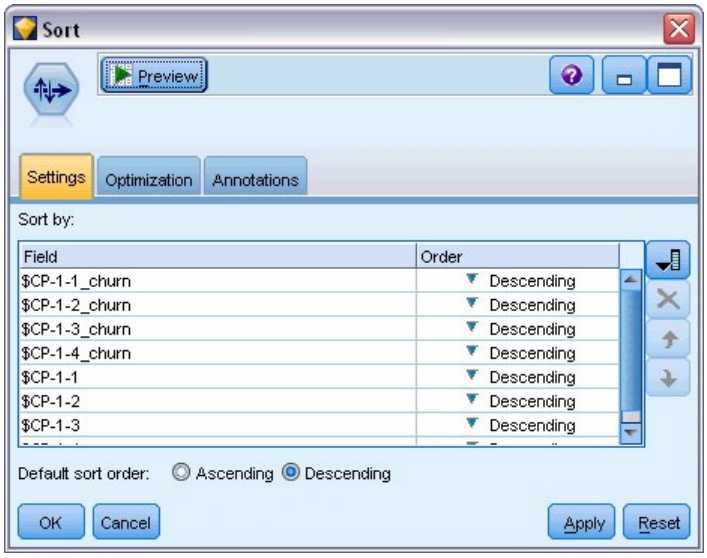


그림 *377.* 정렬 노드*:* 설정 탭

14. 정렬 노드를 파생 노드에 연결하십시오. 설정 탭에서 *$CP-1-1_churn*에서 *$CP-1-4-churn*까지, *$CP-1-1*에서 *$CP-1-4*까지 모두 내림차순으로 정렬하도록 선택하십시오. 이탈할 것으로 예측된 고객이 위쪽에 표시됩니다.

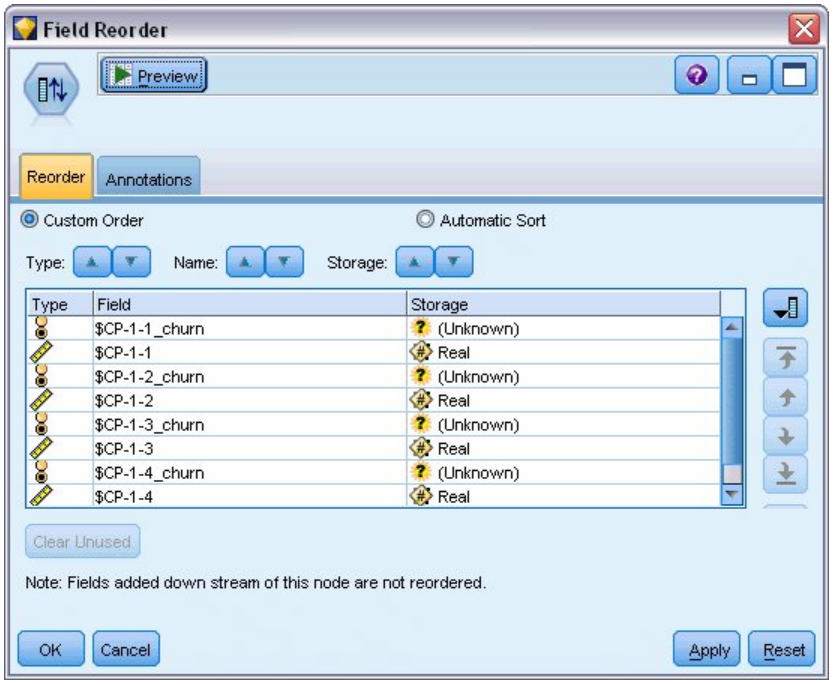
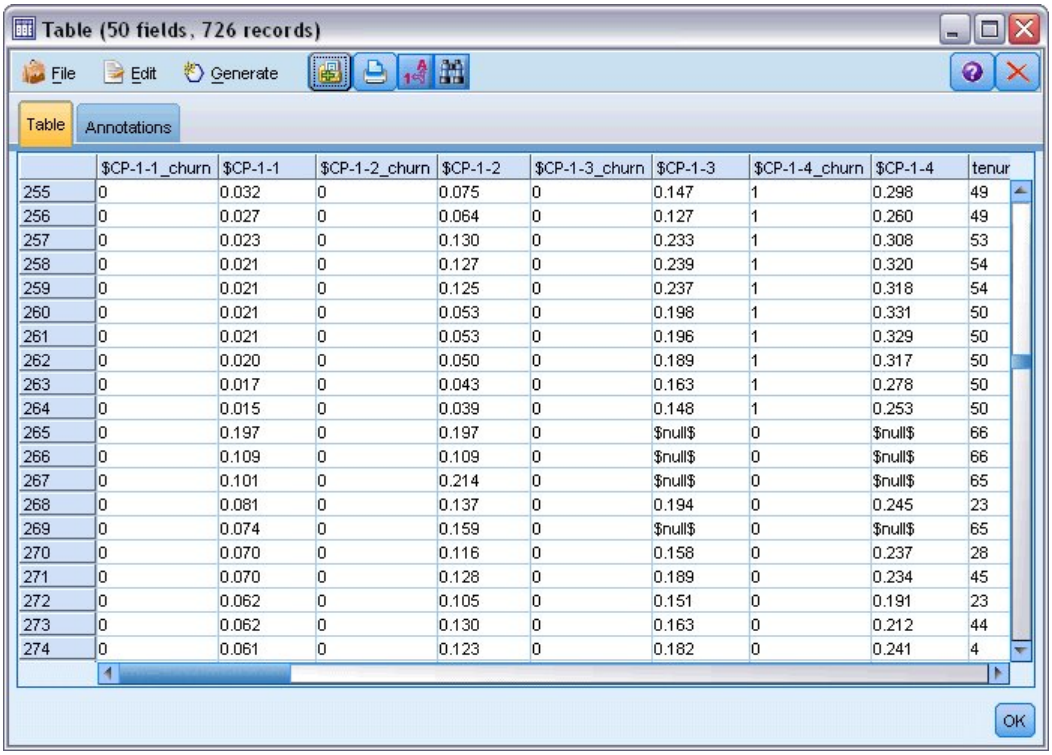


그림 *378.* 필드 다시 정렬 노드*:* 다시 정렬 탭

15. 필드 다시 정렬 노드를 정렬 노드에 연결하십시오. 다시 정렬 탭에서 *$CP-1-1_churn*에서 *$CP-1-4*까지를 다른 필드 앞에 배치하도록 선택하십시오. 그러면 테이블을 더 쉽게 읽을 수 있게 되며

이 조치는 선택적입니다. 필드를 그림에 표시된 위치로 이동하려면 단추를 사용해야 합니다.

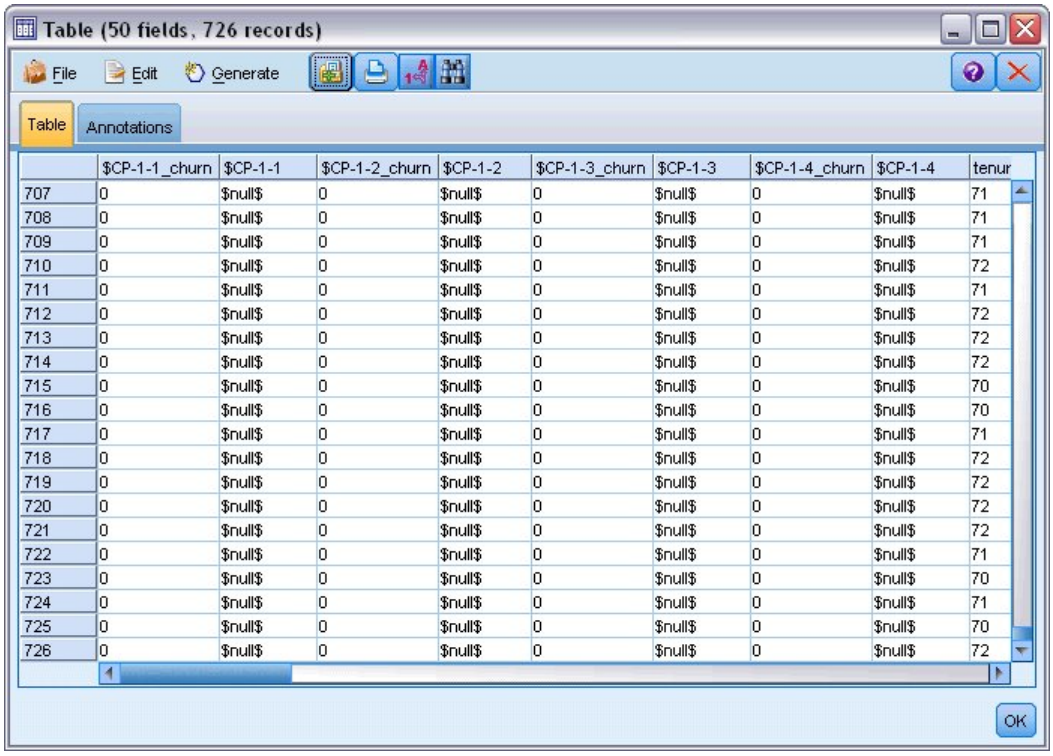


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

그림 *379.* 고객 점수를 표시하는 테이블

16. 테이블 노드를 필드 다시 정렬 노드에 연결하고 이를 실행하십시오.

첫 번째 분기의 끝에 31명의 고객, 두 번째 분기의 끝에 103명의 고객, 세 번째 분기의 끝에 184명의 고객, 이 해의 끝에 264명의 고객이 이탈할 것으로 예측됩니다. 지정된 두 고객, 즉, 첫 번째 분기에서 이탈 성향이 가장 높은 고객이 나머지 분기에서도 반드시 이탈 성향이 가장 높지는 않다는 점에 주목하십시오. 예를 들어, 256 및 260 레코드를 참조하십시오. 이는 고객의 현재 가입 기간을 뒤따르는 개월에 대한 위험함수의 모양 때문일 가능성이 높습니다. 예를 들어, 프로모션 때문에 가입한 고객은 개인적인 추천 때문에 가입한 고객보다 빨리 전환할 가능성이 높지만 전환하지 않는 경우에는 실제로는 나머지 가입 기간 동안 더 충성할 수 있습니다. 이탈할 가능성이 가장 높은 고객에 대한 다른 관점을 얻기 위해 고객을 다시 정렬하고자 할 수 있습니다.



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

그림 *380.* 널값을 가진 고객을 표시하는 테이블

테이블 아래쪽에 널값이 예측되는 고객이 있습니다. 이러한 고객은 총 가입 기간(이후 시간 + *tenure*)이 모델을 학습하는 데 사용된 데이터의 생존 시간 범위에 속하지 않는 고객입니다.

요약

Cox 회귀분석을 사용하여 이탈 시간에 대한 적합한 모델을 발견하고 향후 2년 동안 유지되는 예상 고객 수를 도표화하고 다음 연도에 이탈할 가능성이 가장 높은 개별 고객을 식별했습니다. 이 모델이 적합한 모델이나 최선의 모델은 아닐 수 있습니다. 이상적으로는 최소한 이 모델을 단계별 전진법을 사용하여 작성한 모델 및 단계별 후진법을 사용하여 작성한 모델과 비교해야 합니다.

IBM SPSS Modeler에서 사용된 모델링 방법의 수학적 토대에 대한 설명은 *IBM SPSS Modeler* 알고리즘 안내서에 나와 있습니다.

제 27 장 장바구니 분석(규칙 귀납/C5.0)

이 예에서는 슈퍼마켓 장바구니의 내용(즉, 함께 구매한 항목의 콜렉션) 및 고객 카드 체계를 통해 얻은 구매자의 연관된 개인 데이터를 설명하는 가상의 데이터를 다룹니다. 목적은 유사한 제품을 구매하며 연령, 수입 등의 인구 통계학적으로 분류될 수 있는 고객 그룹을 발견하는 것입니다.

이 예에서는 데이터 마이닝의 두 단계를 설명합니다.

- 구매한 항목 간 링크를 발견하는 연관 규칙 모델링 및 웹 표시
- 식별된 제품 집단의 구매자를 프로파일링하는 C5.0 규칙 귀납

참고: 이 애플리케이션은 예측 모델링을 직접 사용하지는 않으므로 결과 모델에 대한 정확도 측정이 없으며 데이터 마이닝 프로세스에 연관된 학습/검증 구분이 없습니다.

이 예에서는 *BASKETS1n*이라는 데이터 파일을 참조하는 *baskrule*이라는 스트림을 사용합니다. 이러한 파일은 IBM SPSS Modeler 설치의 데모 디렉토리에 있습니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 데모 폴더에 액세스할 수 있습니다. *baskrule* 파일은 *streams* 디렉토리에 있습니다.

데이터 액세스

파일로부터 필드 이름 읽기를 선택하고 가변파일 노드를 사용하여 *BASKETS1n* 데이터 세트에 연결하십시오. 유형 노드를 데이터 소스에 연결한 다음 이 노드를 테이블 노드에 연결하십시오. *cardid* 필드의 측정 수준을 유형 없음으로 설정하십시오. 각 고객 카드 ID가 데이터 세트 내에서 한 번만 발생하고 모델링에서 쓸모가 없기 때문입니다. *sex* 필드에 대한 측정 수준을 명목형으로 선택하십시오. 그러면 Apriori 모델링 알고리즘이 *sex*를 플래그로 처리하지 않습니다.

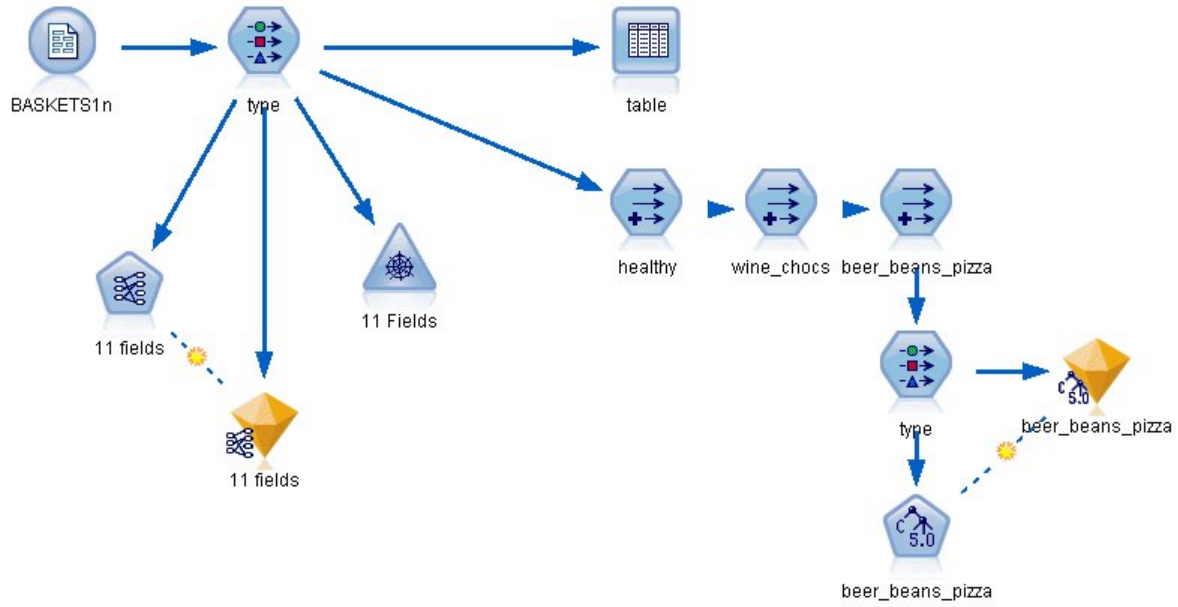


그림 *381. baskrule* 스트림

이제 스트림을 실행하여 유형 노드를 인스턴스화하고 테이블을 표시하십시오. 데이터 세트는 18개의 필드를 포함하며 각 레코드는 바스켓을 나타냅니다.

18개의 필드는 다음과 같은 머리말에서 표시됩니다.

바스켓 요약:

- *cardid*. 이 바스켓을 구매한 고객에 대한 고객 카드 ID입니다.
- *value*. 바스켓의 총 구매 가격입니다.
- *pmethod*. 바스켓에 대한 지불 방법입니다.

카드 소유자의 개인적 세부사항:

- *sex*
- *homeown*. 카드 소유자가 주택을 소유하는지 여부입니다.
- *income*
- *age*

바스켓 내용-제품 범주의 존재 여부에 대한 플래그:

- *fruitveg*
- *freshmeat*
- *dairy*
- *cannedveg*
- *cannedmeat*
- *frozenmeal*

- *beer*
- *wine*
- *softdrink*
- *fish*
- *confectionery*

바스켓 내용에서 유사성 검색

먼저 Apriori를 사용하여 바스켓 내용의 전체 유사성(연관) 그림을 얻어서 연관 규칙을 생성해야 합니다. 유형 노드를 편집하고 전체 제품 범주의 역할을 모두로 설정하고 모든 기타 역할을 없음으로 설정하여 이 모델링 프로세스에서 사용할 필드를 선택하십시오. (모두는 해당 필드가 결과 모델의 입력 또는 출력으로 사용될 수 있음을 의미합니다.)

참고: Shift - 클릭을 사용하여 다중 필드에 대한 옵션을 설정함으로써 열에서 옵션을 지정하기 전에 필드를 선택할 수 있습니다.

그림 *382.* 모델링에 사용할 필드 선택

일단 모델링에 사용할 필드를 지정한 후에는 Apriori 노드를 유형 노드에 연결하여 이를 편집하고 **플래그에 대한 참 값만 이용**을 선택한 다음 Apriori 노드에서 실행을 클릭하십시오. 그 결과, 관리자 창의 오른쪽 상단에 있는 모델 탭에 있는 모델에 컨텍스트 메뉴를 사용하여 **찾아보기**를 선택하면 볼 수 있는 연관 규칙이 포함됩니다.

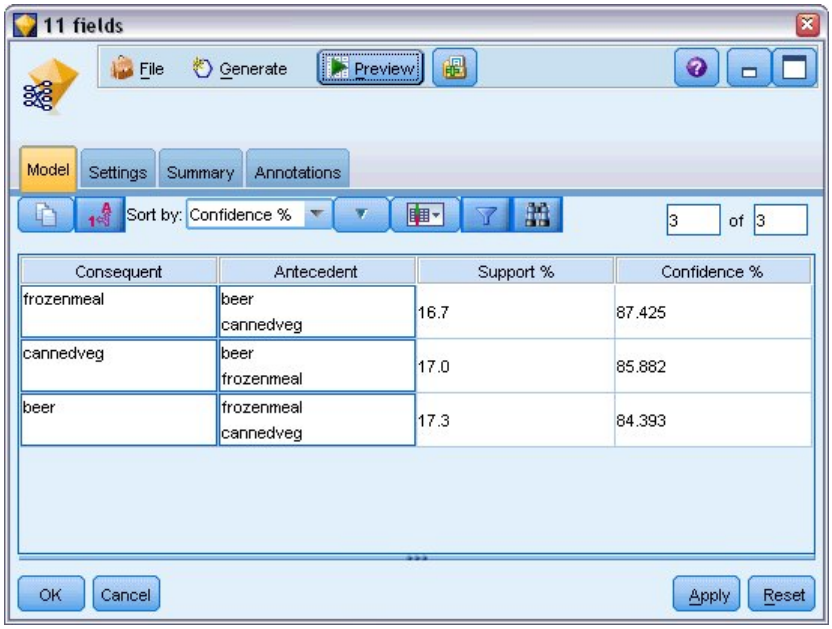


그림 *383.* 연관 규칙

이러한 규칙은 냉동 식품, 통조림 야채 및 맥주 사이의 다양한 연관을 표시합니다. 다음 두 방향 연관 규칙의 존재를 보면,

```
frozenmeal -> beer
beer -> frozenmeal
```

(두 방향 연관만 표시하는) 웹 표시는 이 데이터의 패턴 중 일부만 강조표시함을 알 수 있습니다.

웹 노드를 유형 노드에 연결하고 웹 노드를 편집하고 모든 바스켓 내용 필드를 선택하고 **참 플래그만 표시**를 선택하고 웹 노드에서 실행을 클릭하십시오.

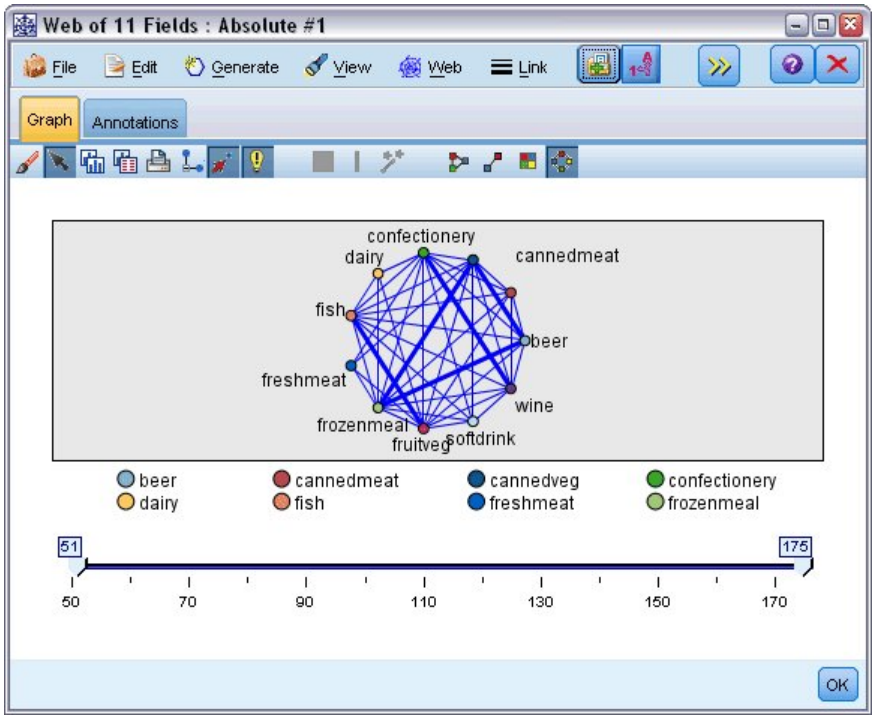


그림 *384.* 제품 연관의 웹 표시

제품 범주의 대부분의 조합이 여러 바스켓에서 발생하므로 모델이 제안하는 고객 그룹을 표시하기에는 이 웹에서 강한 링크가 너무 많습니다.

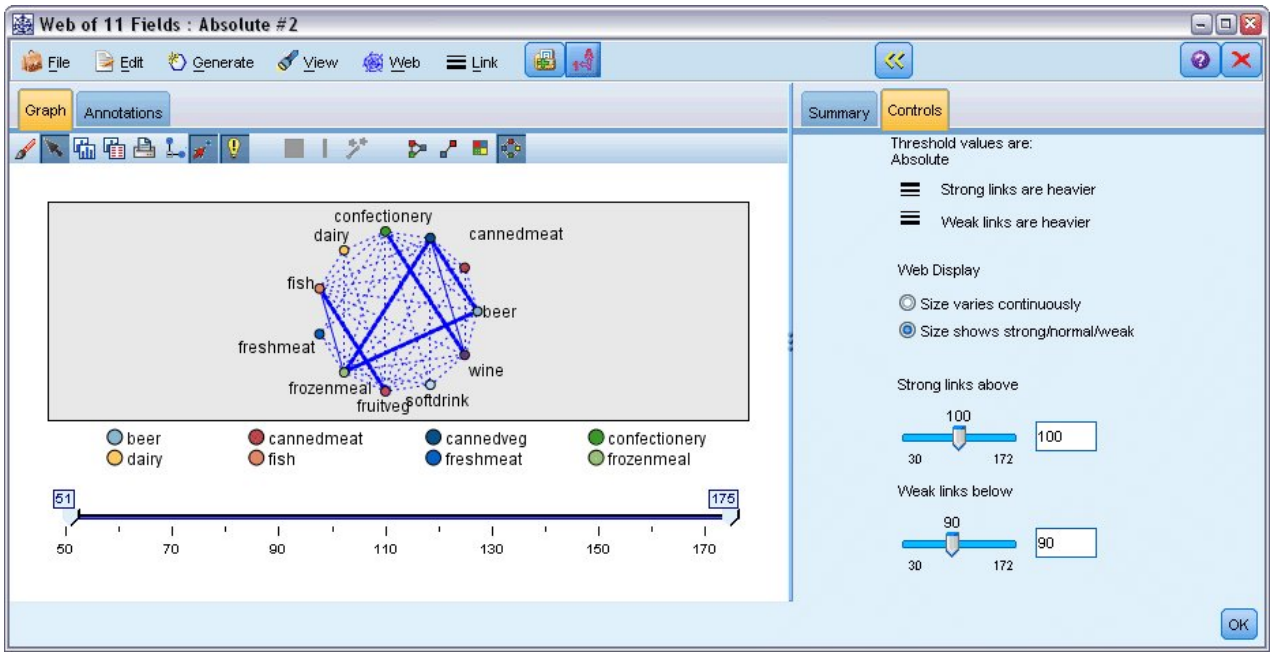


그림 *385.* 제한된 웹 표시

1. 약한 연결 및 강한 연결을 지정하려면 도구 모음에서 노란색 이중 화살표 단추를 클릭하십시오. 그러면 웹 출력 요약 및 제어를 표시하는 대화 상자가 펼쳐집니다.

2. **크기가 강함/보통/약함으로 나타납니다.**를 선택하십시오.
3. 약한 링크를 90 아래로 설정하십시오.
4. 강한 링크를 100 위로 설정하십시오.

결과 표시에서 세 고객 그룹이 두드러집니다.

- "건강한 음식을 먹는 사람"이라고 할 수 있는 생선, 과일 및 채소를 구매한 고객
- 포도주 및 과자류를 구매한 고객
- 맥주, 냉동 식품 및 통조림 채소("맥주, 콩류 및 피자")를 구매한 고객

고객 집단 프로파일링

구매한 제품의 유형을 기반으로 하여 세 가지 유형의 고객을 식별했으나 아직 이러한 고객이 누구인지, 즉, 인구 통계학적 프로파일에 대해 알고 싶어 합니다. 이러한 각 집단에 대해 플래그를 사용하여 각 고객에게 태그를 지정하고 해당 플래그의 규칙 기반 프로파일을 작성하기 위한 규칙 귀납(C5.0)을 사용하여 이를 수행할 수 있습니다.

먼저, 각 집단에 대한 플래그를 파생시켜야 합니다. 이는 방금 작성한 웹 표시를 사용하여 자동으로 생성될 수 있습니다. 마우스 오른쪽 단추를 사용하여 *fruitveg* 및 *fish* 사이의 링크를 클릭한 다음 마우스 오른쪽 단추를 클릭하여 **링크에 대한 파생 노드 생성**을 선택하십시오.

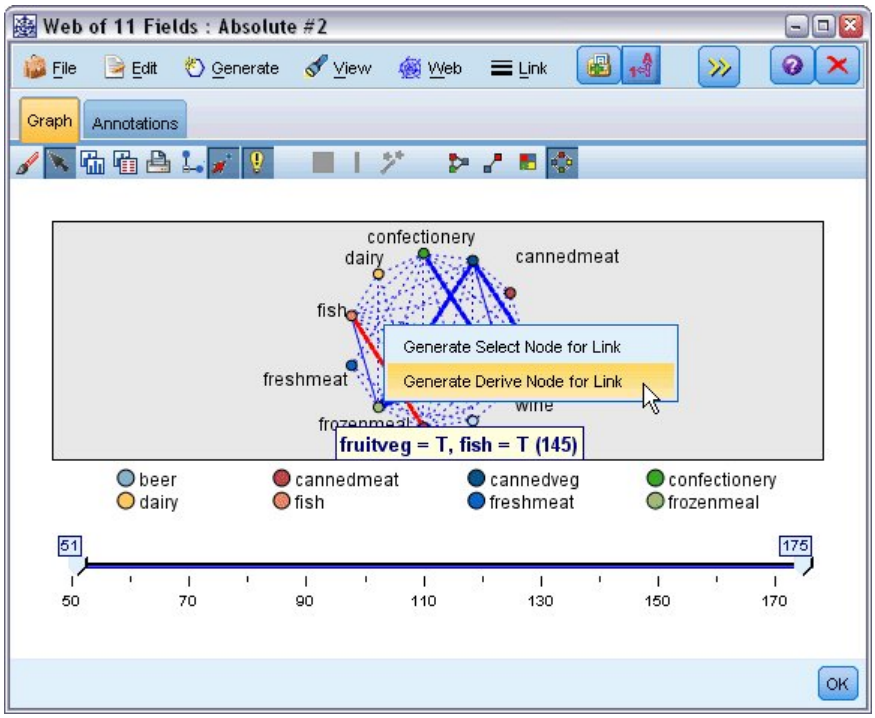


그림 *386.* 각 고객 집단에 대한 플래그 파생

결과 파생 노드를 편집하여 파생 노드 이름을 *healthy*로 변경하십시오. *wine*에서 *confectionery*로의 링크에 대해 연습을 반복하여 결과 파생 필드 이름을 *wine_chocs*로 변경하십시오.

(세 링크와 연관된) 세 번째 집단의 경우, 먼저 링크가 선택되지 않았는지 확인하십시오. 그런 다음 Shift 키를 누른 상태에서 마우스 왼쪽 단추를 클릭하여 *cannedveg*, *beer* 및 *frozenmeal* 삼각형에서 세 링크를 모두 선택하십시오. (편집 모드가 아니라 대화식 모드여야 합니다.) 그런 다음 웹 표시 메뉴에서 다음을 선택하십시오.

생성 > **파생 노드("및")**

결과 파생 필드의 이름을 *beer_beans_pizza*로 변경하십시오.

이러한 고객 집단을 프로파일링하려면 기존 유형 노드를 이러한 세 파생 노드에 연속하여 연결한 다음 또 다른 유형 노드를 연결하십시오. 새 유형 노드에서 모든 필드의 역할을 없음으로 설정하십시오. 단, *value*, *pmethod*, *sex*, *homeown*, *income* 및 *age*는 입력으로 설정해야 하며 관련 고객 집단(예: *beer_beans_pizza*)은 대상으로 설정해야 합니다. C5.0 노드를 연결하고 출력 유형을 **규칙 세트**로 설정하고 노드에서 실행을 클릭하십시오. *beer_beans_pizza*에 대한 결과 모델은 이 고객 집단에 대한 명확한 인구 통계학적 프로파일을 포함합니다.

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

두 번째 유형 노드에서 다른 고객 집단을 출력으로 선택하여 해당 고객 그룹 플래그에 동일한 방법을 적용할 수 있습니다. 이 컨텍스트에서 C5.0 대신 Apriori를 사용하여 더 넓은 범위의 대안 프로파일을 생성할 수 있습니다. 또한 Apriori는 단일 출력 필드로 제한되지 않으므로 모든 고객 집단의 플래그를 동시에 프로파일링하는 데 사용될 수 있습니다.

요약

이 예에서는 모델링(Apriori 사용) 및 시각화(웹 표시 사용) 둘 다를 사용하여 데이터베이스에서 유사성 또는 링크를 발견하는 데 IBM SPSS Modeler를 사용할 수 있는 방법에 대해 설명합니다. 이러한 링크는 데이터의 케이스 집단에 해당되며 이러한 집단을 자세히 조사하여 모델링(C5.0 규칙 세트 사용)을 사용하여 프로파일링할 수 있습니다.

예를 들어, 소매업체 도메인에서 다이렉트 메일에 대한 반응률을 개선하거나 지점별로 재고 제품의 범위를 사용자 정의하여 인구 통계학적 기반의 요구를 충족할 수 있도록 특별 오퍼를 대상화하는 데 고객 집단이 사용될 수 있습니다.

제 28 장 새 차량 오퍼링(KNN) 평가

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.

자동차 제조업체는 승용차 및 트럭이라는 두 가지 새 차량에 대한 프로토타입을 개발해 왔습니다. 제조업체는 새 모델을 범위 안에 포함시키기 전에 시장의 어떤 기존 차량이 프로토타입과 가장 유사한지 판별하기를 원합니다. 즉, 어떤 차량이 "가장 가까운 이웃"이라서 경쟁하게 될 것인지 알고자 합니다.

제조업체는 수많은 범주 아래에 기존 모델에 대한 데이터를 수집해 왔으며 프로토타입의 세부사항을 추가해 왔습니다. 모델을 비교할 범주에는 천 단위 가격(*price*), 엔진 크기(*engine_s*), 마력(*horsepow*), 축간 거리(*wheelbas*), 너비(*width*), 길이(*length*), 공차 중량(*curb_wgt*), 연료 용량(*fuel_cap*) 및 연비(*mpg*) 등이 있습니다.

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

스트림 작성

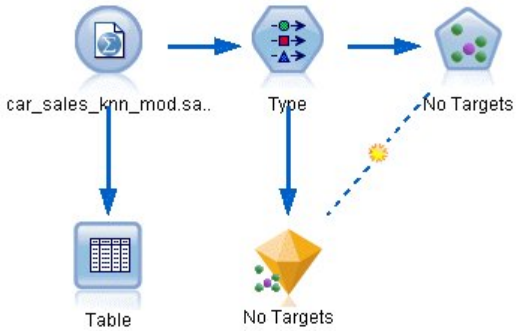
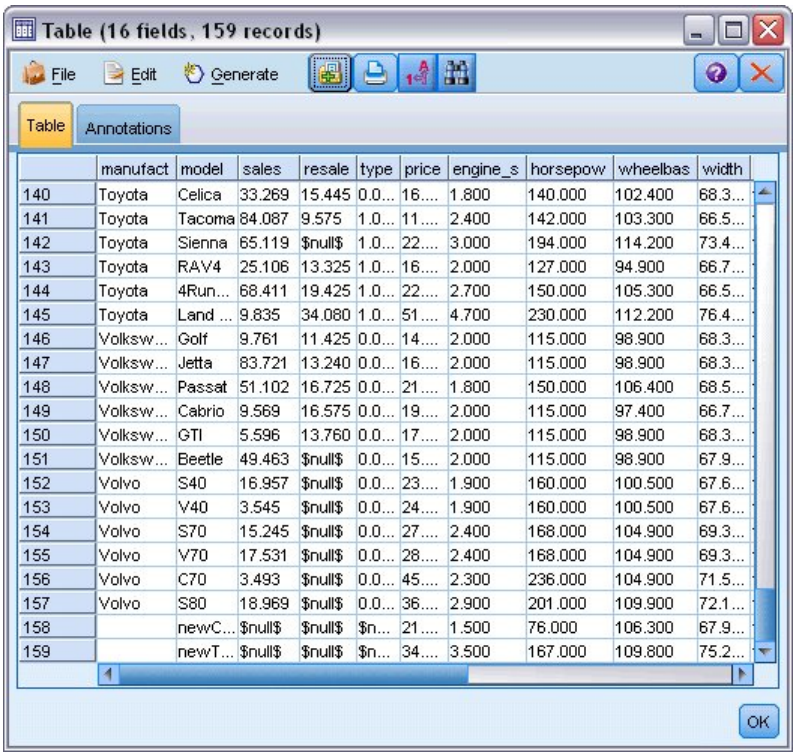


그림 *387. KNN* 모델링에 대한 샘플 스트림

새 스트림을 작성하고 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *car_sales_knn_mod.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.

먼저 제조업체에서 수집한 데이터를 보십시오.

1. 테이블 노드를 통계 파일 노드에 연결하십시오.
2. 테이블 노드를 열고 **실행**을 클릭하십시오.



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

그림 *388.* 승용차 및 트럭에 대한 소스 데이터

newCar 및 *newTruck*이라는 두 가지 프로토유형에 대한 세부사항이 파일의 끝에 추가되었습니다.

소스 데이터로부터 제조업체가 승용차 유형이 아닌 차량을 다소 광범위하게 의미하는 데 "트럭"이라는 분류(*type* 열의 1 값)를 사용함을 알 수 있습니다.

가장 가까운 이웃을 식별할 때 두 프로토타입이 검증용으로 지정될 수 있도록 마지막 열인 *partition*이 필요합니다. 이 방법으로 데이터가 계산에 영향을 미치지 않을 수 있습니다. 해당 데이터가 이 케이스에서 고려할 시장의 범위 밖이기 때문입니다. 두 검증용 레코드의 *partition* 값을 1로 설정하고 기타 모든 레코드를 이 필드에서 0으로 설정하면 나중에 초점 레코드(가장 가까운 이웃을 계산할 레코드)를 설정하게 될 때 이 필드를 사용할 수 있습니다.

나중에 테이블 출력 창을 참조할 것이므로 지금은 열린 상태로 두십시오.

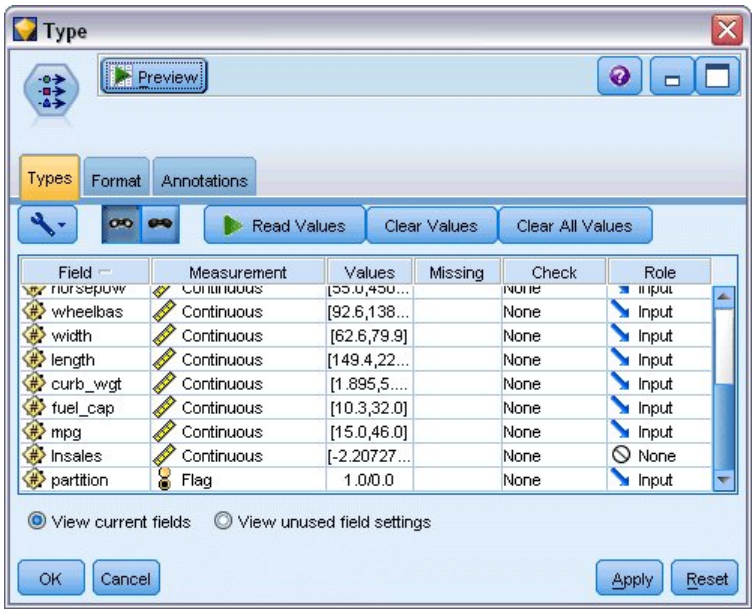


그림 *389.* 유형 노드 설정

3. 스트림에 유형 노드를 추가하십시오.
4. 유형 노드를 통계 파일 노드에 연결하십시오.
5. 유형 노드를 여십시오.

   *price*에서 *mpg*까지의 필드만 비교할 것이므로 이러한 모든 필드에 대한 역할을 **입력**으로 설정된 상태로 둡니다.
6. 모든 기타 필드(*manufact*에서 *type*까지 및 *lnsales*)에 대한 역할을 **없음**으로 설정하십시오.
7. 마지막 필드인 *partition*에 대한 측정 수준을 **플래그**로 설정하십시오. 역할이 **입력**으로 설정되었는지 확인하십시오.
8. **값 읽기**를 클릭하여 데이터 값을 스트림으로 읽으십시오.
9. **확인**을 클릭하십시오.

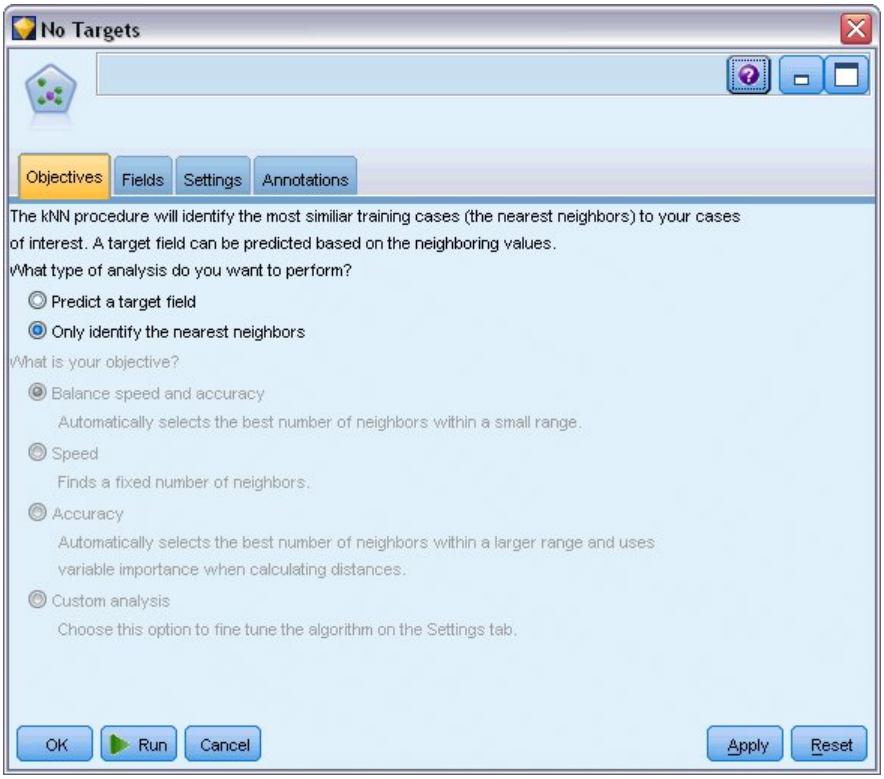


그림 *390.* 가장 가까운 이웃 식별 선택

10. KNN 노드를 유형 노드에 연결하십시오.
11. KNN 노드를 여십시오.

    두 프로토유형에 대해서 가장 가까운 이웃을 찾는 것만 필요하므로 이번에는 대상 필드를 예측하지 않을 것입니다.
12. **목적** 탭에서 **가장 가까운 이웃 항목만 식별**을 선택하십시오.
13. **설정** 탭을 클릭하십시오.

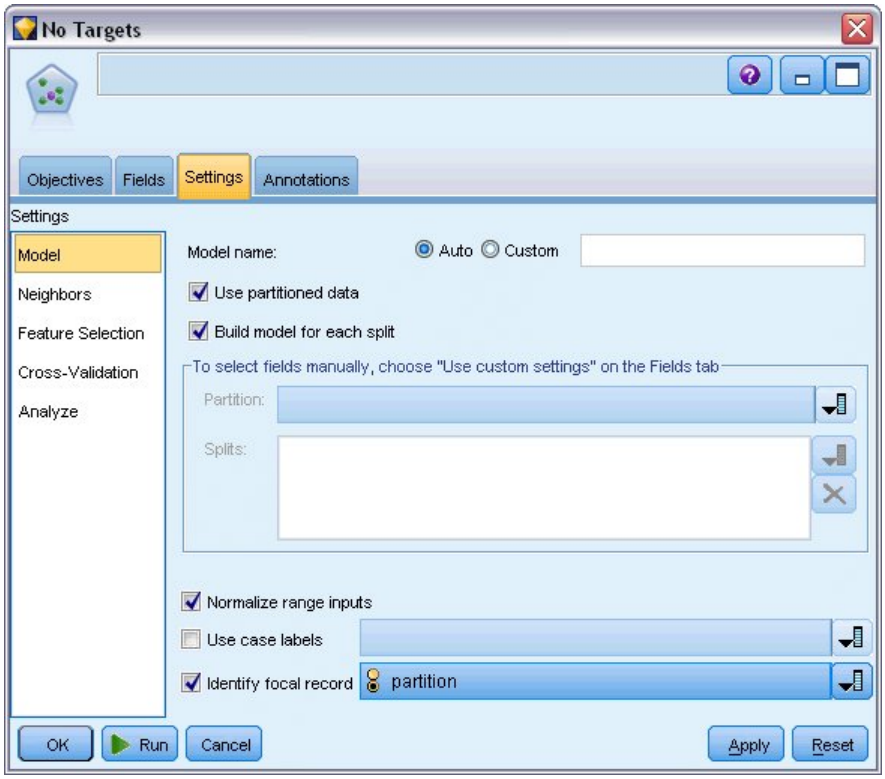


그림 *391.* 초점 레코드를 식별하기 위해 파티션 필드 사용

이제 파티션 필드를 사용하여 초점 레코드(가장 가까운 이웃을 식별하고자 하는 레코드)를 식별할 수 있습니다. 플래그 필드를 사용하면 이 필드의 값이 1로 설정된 레코드가 초점 레코드가 되도록 할 수 있습니다.

이 필드의 값이 1인 유일한 레코드는 *newCar* 및 *newTruck*이므로 이러한 레코드가 초점 레코드가 됩니다.

14. **설정** 탭의 **모델** 패널에서 **초점 레코드 식별** 선택란을 선택하십시오.
15. 이 필드에 대한 드롭 다운 목록에서 **파티션**을 선택하십시오.
16. **실행** 단추를 클릭하십시오.

출력 탐색

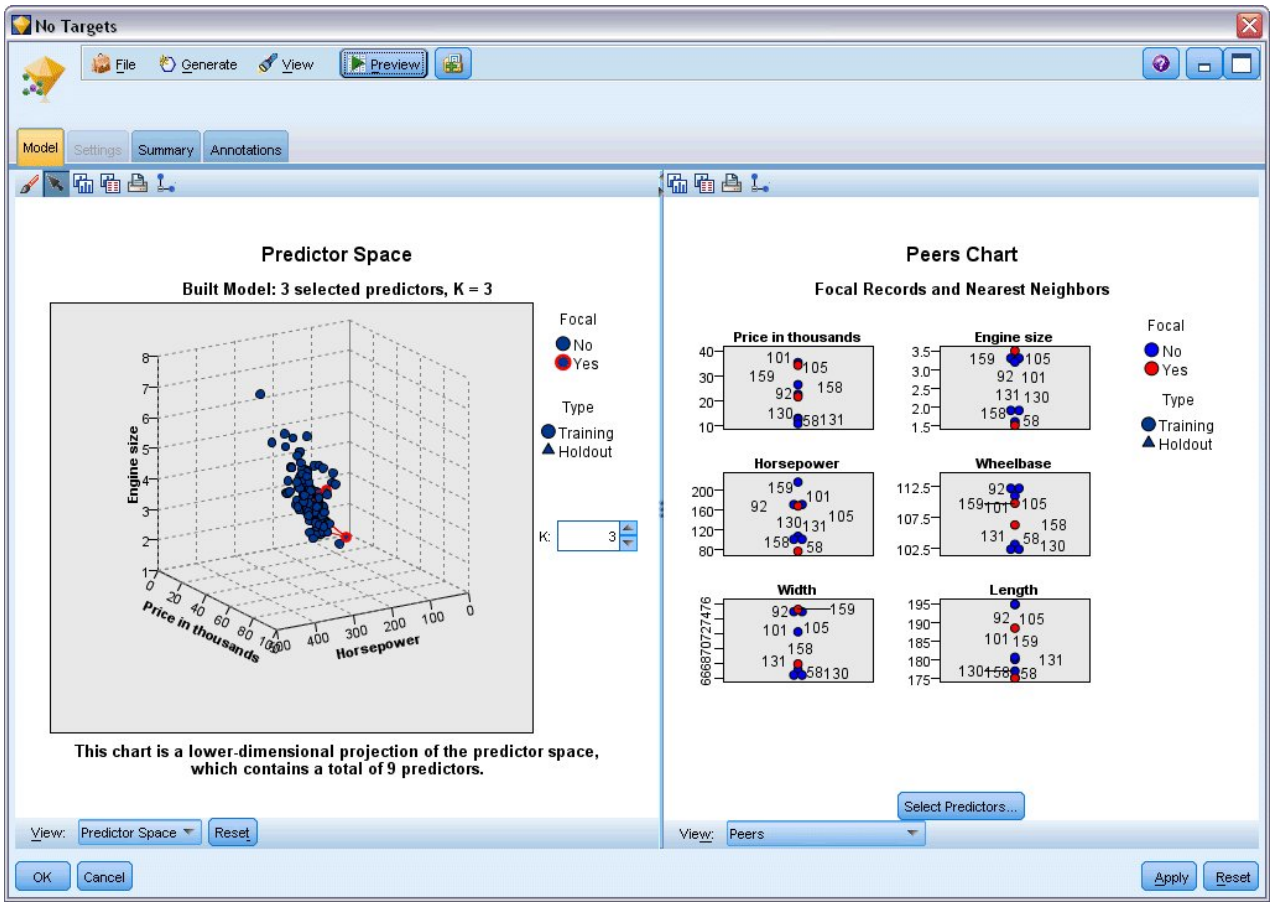


그림 *392.* 모델 뷰어 창

모델 너깃은 스트림 캔버스 및 모델 팔레트에서 생성되었습니다. 너깃 중 하나를 열어서 두 패널 창이 있는 모델 뷰어 표시를 보십시오.

- 첫 번째 패널에서는 기본 보기라고 불리는 모델 개요가 표시됩니다.가장 가까운 이웃 모델에 대한 기본 보기를 **예측변수 공간**이라고 합니다.
- 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.

예측변수 공간

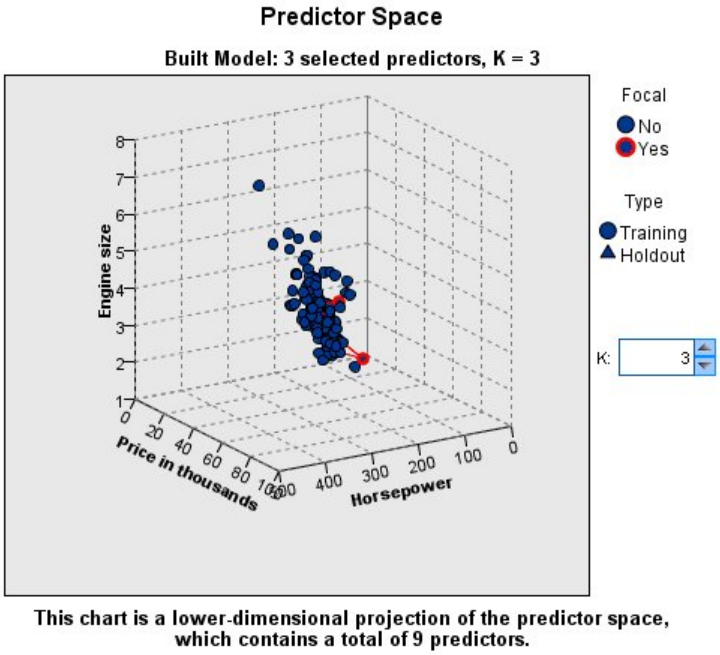


그림 *393.* 예측변수 공간 차트

예측변수 공간 차트는 가격, 엔진 크기 및 마력을 나타내는 세 가지 변수(실제로 소스 데이터의 처음 세 입력 필드)에 대한 데이터 점을 도표로 작성하는 대화형 3차원 그래프입니다.

두 개의 초점 레코드는 빨간색으로 강조표시되며 선을 이용하여 *k* 가장 가까운 이웃에 연결됩니다.

차트를 클릭하여 끌어오는 방법으로 예측변수 공간에서 분포 점을 더 잘 볼 수 있도록 회전시킬 수 있습니다. 기본 보기로 되돌리려면 **재설정** 단추를 클릭하십시오.

피어 차트

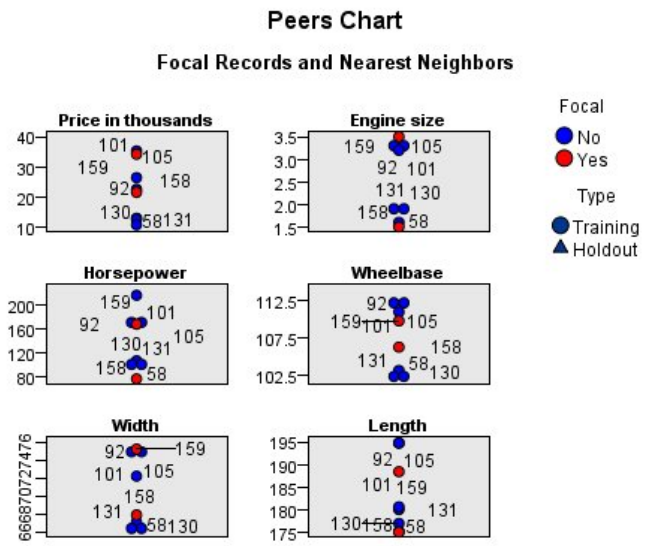


그림 *394.* 피어 차트

기본 보조 보기는 예측변수 공간에서 선택된 두 초점 레코드와 여섯 개의 각 변수에 대한 *k* 가장 가까운 이웃(소스 데이터의 처음 여섯 개의 입력 필드)을 강조표시하는 피어 차트입니다.

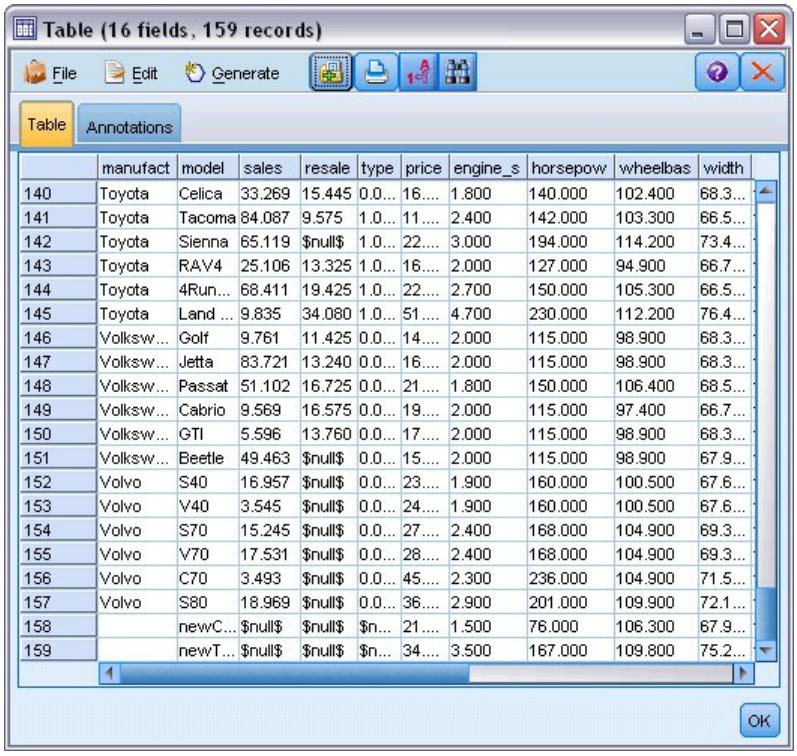
차량은 소스 데이터에서 레코드 번호로 표시됩니다. 이는 식별하는 데 도움을 받기 위해 테이블 노드의 출력이 필요한 곳입니다.

테이블 노드 출력이 여전히 사용 가능한 경우 다음을 수행하십시오.

1. 기본 IBM SPSS Modeler 창의 오른쪽 상단에 있는 관리자 분할창의 **출력** 탭을 클릭하십시오.
2. 항목 **테이블(16필드, 159 레코드)**을 두 번 클릭하십시오.

테이블 출력이 더 이상 사용 가능하지 않은 경우 다음을 수행하십시오.

3. 기본 IBM SPSS Modeler 창에서 테이블 노드를 여십시오.
4. **실행**을 클릭하십시오.



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

그림 *395.* 레코드 번호로 레코드 식별

테이블의 아래쪽으로 스크롤하면 *newCar* 및 *newTruck*이 데이터의 마지막 두 개의 레코드이며 각각 158 및 159로 번호가 매겨진 것을 볼 수 있습니다.

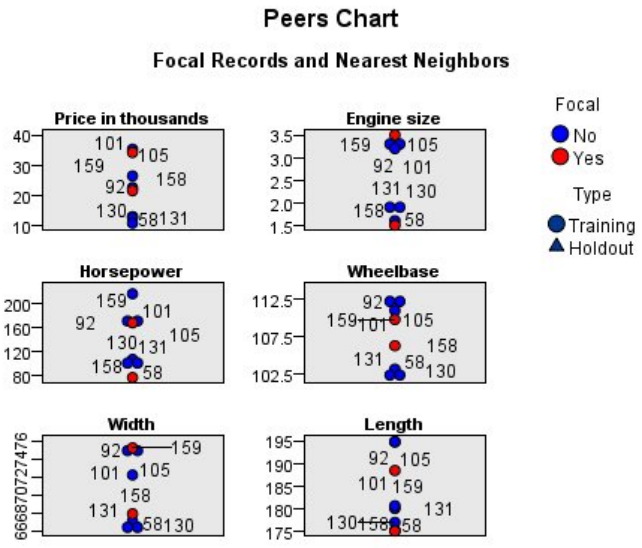


그림 *396.* 피어 차트에서 변수 비교

여기서 피어 차트를 볼 수 있습니다. 예를 들어, *newTruck*(159)은 가장 가까운 이웃 중 가장 큰 엔진을 가진 반면 *newCar*(158)는 해당되는 가장 가까운 이웃보다 작은 엔진을 갖고 있습니다.

여섯 개의 각 변수에 대해 마우스를 개별 점 위로 이동하면 특정 케이스에 대한 각 변수의 실제 값을 볼 수 있습니다.

그러나 *newCar* 및 *newTruck*에 대한 가장 가까운 이웃은 어느 차량입니까?

피어 차트는 약간 복잡할 수 있으므로 단순한 보기로 변경하도록 합니다.

5. 피어 차트의 맨 아래에 있는 **보기** 드롭 다운 목록(현재 **피어**로 표시되는 항목)을 클릭하십시오.
6. **이웃 및 거리 테이블**을 선택하십시오.

이웃 및 거리 테이블

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distar	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

그림 *397.* 이웃 및 거리 테이블

잘 수행되었습니다. 이제 시장에서 두 프로토타입이 서로 가장 가까운 세 모델을 볼 수 있습니다.

newCar(초점 레코드 158)의 경우, Saturn SC(131), Saturn SL(130) 및 Honda Civic(58)입니다.

세 모델 모두 중형 크기의 세단형 승용차이므로 당연한 결과일 수 있습니다. 따라서 *newCar*는 매우 적합하며 특히 연비가 우수합니다.

newTruck(초점 레코드 159)의 경우, 가장 가까운 이웃은 Nissan Quest(105), Mercury Villager(92) 및 Mercedes M-Class(101)입니다.

앞에서 본 것과 같이 반드시 전통적인 관점에서의 트럭일 필요는 없으며 승용차로 분류되지 않는 차량이기만 하면 됩니다. 가장 가까운 이웃에 대한 테이블 노드 출력을 보자면 *newTruck*이 상대적으로 비싸고 가장 무거운 유형 중 하나임을 알 수 있습니다. 단, 연비는 가장 가까운 라이벌보다 더 우수하므로 이를 장점으로 간주해야 합니다.

요약

특정 데이터 세트의 케이스 내의 광범위한 변수군을 비교하기 위해 가장 가까운 이웃 분석을 사용하는 방법에 대해 학습했습니다. 또한 두 개의 매우 다른 검증용 레코드에 대해 이러한 검증용 레코드를 가장 가까이 닮은 케이스를 계산했습니다.

제 29 장 비즈니스 메트릭(TCM)에서 인과 관계 찾기

비즈니스에서는 시간 경과에 따른 비즈니스의 재정 상태를 설명하는 핵심성과지표(KPI)를 추적하며 제어 가능한 수많은 메트릭을 추적합니다. 비즈니스에서는 시간 인과 모델링을 사용하여 제어 가능한 메트릭과 KPI 사이의 인과 관계를 찾는 데 관심이 있습니다. 또한 KPI 사이의 인과 관계를 밝히는 데도 관심이 있습니다.

tcm_kpi.sav 데이터 파일에는 KPI 및 제어 가능한 메트릭에 대한 주간 데이터가 포함됩니다. KPI에 대한 데이터는 접두문자가 *KPI*인 필드에 저장되고 제어 가능한 메트릭에 대한 데이터는 접두문자가 *Lever*인 필드에 저장됩니다.

스트림 작성

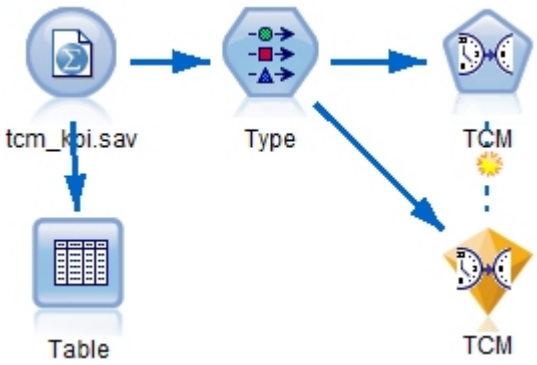
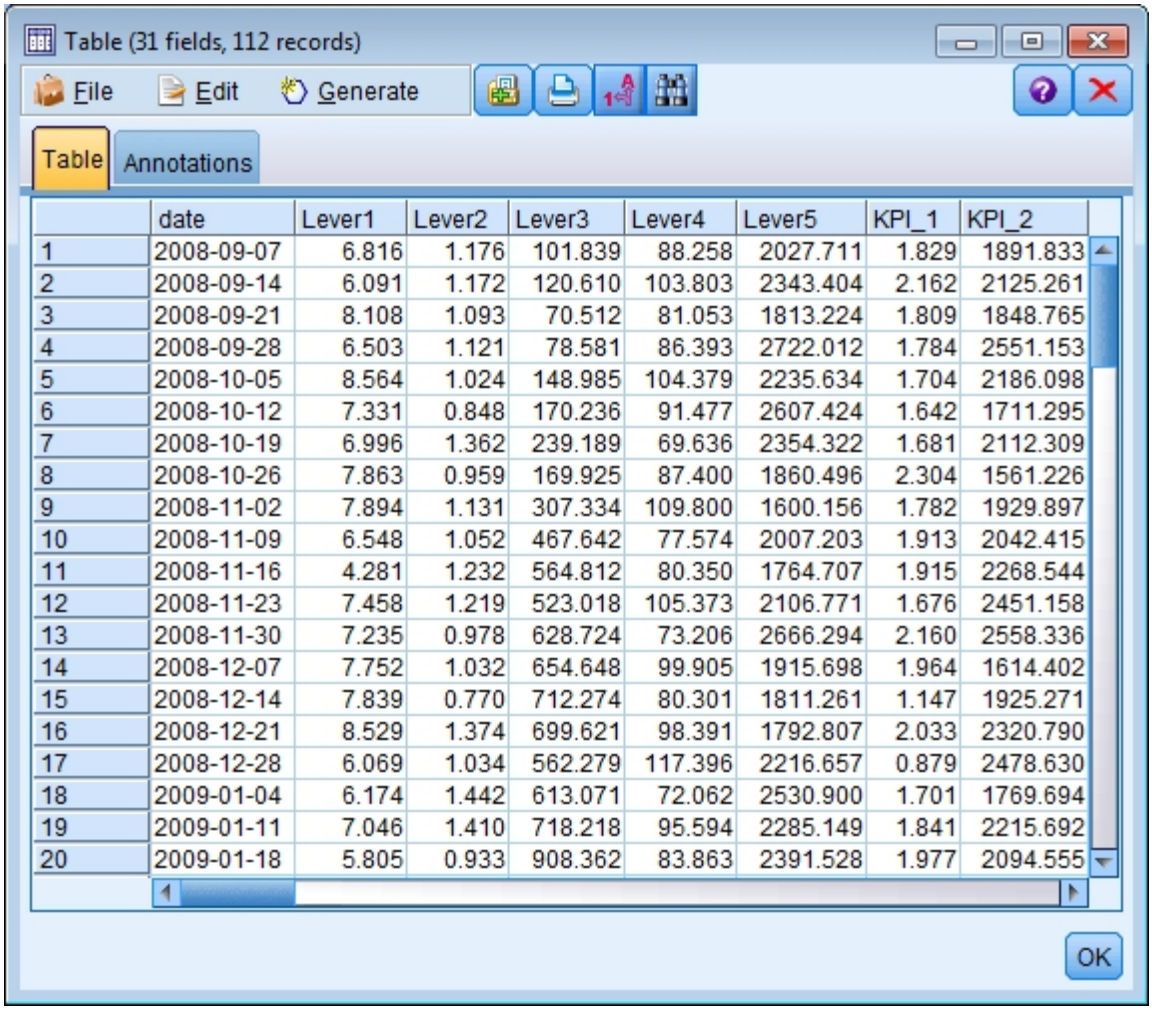


그림 *398. TCM* 모델링에 대한 샘플 스트림

1. 새 스트림을 작성하고 IBM SPSS Modeler 설치의 *Demos* 폴더에 있는 *tcm_kpi.sav*를 가리키는 통계 파일 소스 노드를 추가하십시오.
2. 테이블 노드를 통계 파일 노드에 연결하십시오.
3. 테이블 노드를 열고 **실행**을 클릭하여 데이터를 보십시오. 여기에는 핵심성과지표(KPI) 및 제어 가능한 메트릭에 대한 주간 데이터가 포함됩니다. 핵심성과지표(KPI)에 대한 데이터는 접두문자가 *KPI*인 필드에 저장되고 제어 가능한 메트릭에 대한 데이터는 접두문자가 *Lever*인 필드에 저장됩니다.



	date	Lever1	Lever2	Lever3	Lever4	Lever5	KPI_1	KPI_2
1	2008-09-07	6.816	1.176	101.839	88.258	2027.711	1.829	1891.833
2	2008-09-14	6.091	1.172	120.610	103.803	2343.404	2.162	2125.261
3	2008-09-21	8.108	1.093	70.512	81.053	1813.224	1.809	1848.765
4	2008-09-28	6.503	1.121	78.581	86.393	2722.012	1.784	2551.153
5	2008-10-05	8.564	1.024	148.985	104.379	2235.634	1.704	2186.098
6	2008-10-12	7.331	0.848	170.236	91.477	2607.424	1.642	1711.295
7	2008-10-19	6.996	1.362	239.189	69.636	2354.322	1.681	2112.309
8	2008-10-26	7.863	0.959	169.925	87.400	1860.496	2.304	1561.226
9	2008-11-02	7.894	1.131	307.334	109.800	1600.156	1.782	1929.897
10	2008-11-09	6.548	1.052	467.642	77.574	2007.203	1.913	2042.415
11	2008-11-16	4.281	1.232	564.812	80.350	1764.707	1.915	2268.544
12	2008-11-23	7.458	1.219	523.018	105.373	2106.771	1.676	2451.158
13	2008-11-30	7.235	0.978	628.724	73.206	2666.294	2.160	2558.336
14	2008-12-07	7.752	1.032	654.648	99.905	1915.698	1.964	1614.402
15	2008-12-14	7.839	0.770	712.274	80.301	1811.261	1.147	1925.271
16	2008-12-21	8.529	1.374	699.621	98.391	1792.807	2.033	2320.790
17	2008-12-28	6.069	1.034	562.279	117.396	2216.657	0.879	2478.630
18	2009-01-04	6.174	1.442	613.071	72.062	2530.900	1.701	1769.694
19	2009-01-11	7.046	1.410	718.218	95.594	2285.149	1.841	2215.692
20	2009-01-18	5.805	0.933	908.362	83.863	2391.528	1.977	2094.555

그림 *399.* 핵심성과지표*(KPI)* 및 제어 가능한 메트릭에 대한 소스 데이터

4. 스트림에 유형 노드를 추가하십시오.
5. 유형 노드를 통계 파일 노드에 연결하십시오.

분석 실행

1. TCM 노드를 유형 노드에 연결을 선택한 다음 TCM 노드를 열고 **필드** 탭의 **관측값** 섹션으로 이동하십시오.

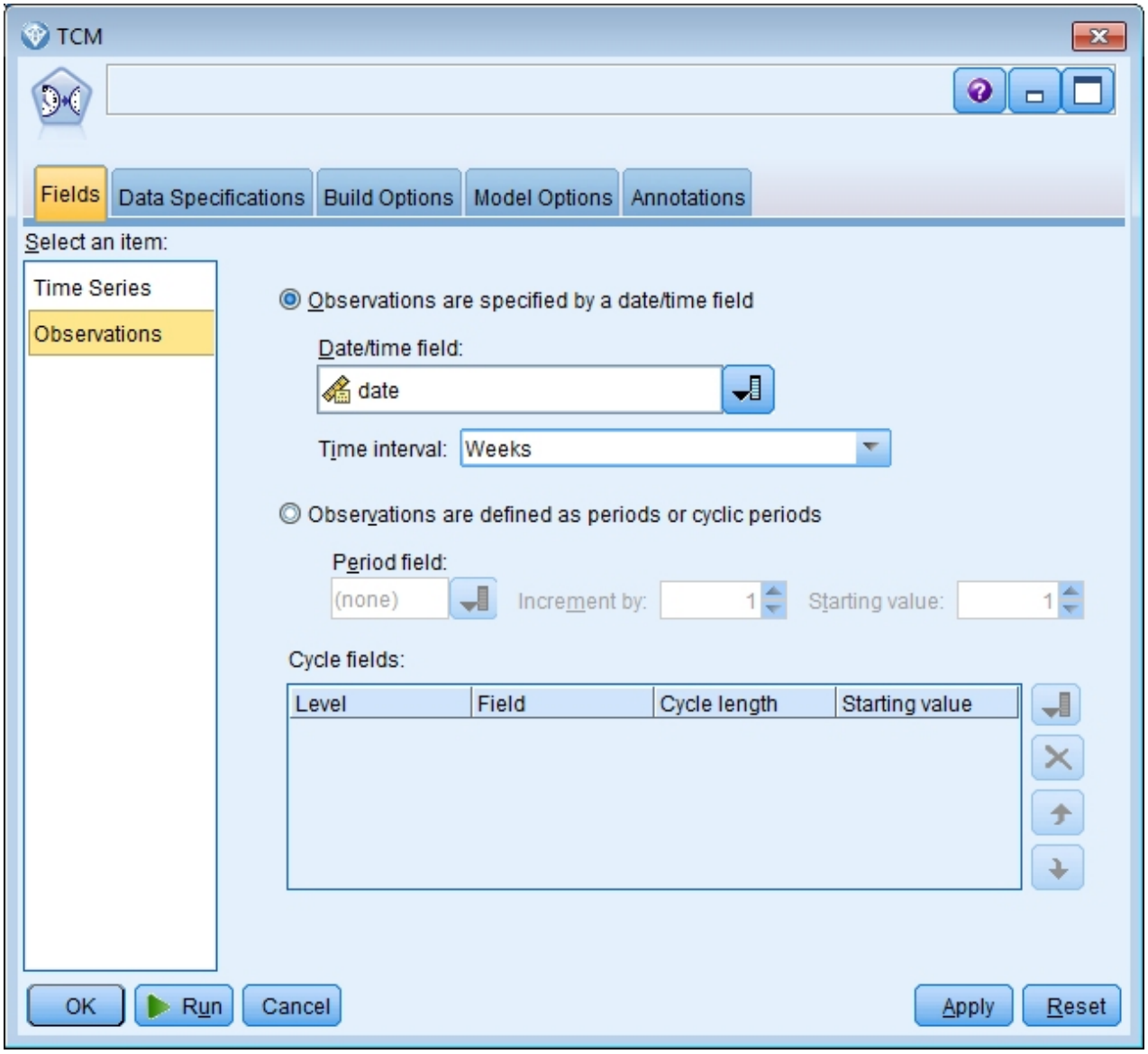


그림 *400.* 시간 인과 모델링, 관측값

2. 날짜/시간 필드에서 *date*를 선택하고 시간 구간 구간에서 *Weeks*를 선택하십시오.
3. **시계열**을 클릭하고 **사전 정의된 역할 사용**을 선택하십시오.

   표본 데이터 세트 `tcm_kpi.sav`에서 *Lever1*에서 *Lever5* 까지의 필드는 입력 역할을 하며 *KPI_1*에서 *KPI_25* 까지의 필드는 모든 역할을 합니다. **사전 정의된 역할 사용**이 선택된 경우, 입력 역할의 필드는 후보 입력으로 처리되고 두 역할을 모두 가진 필드는 시간 인과 모델링에 대한 후보 입력 및 대상 둘 다로 처리됩니다.

   시간 인과 모델링 프로시저는 후보 입력 변수군에서 각 대상에 대해 최선의 입력을 판별합니다. 이 예에서 후보 입력은 *Lever1*에서 *Lever5*까지의 필드 및 *KPI_1*에서 *KPI_25*까지의 필드입니다.
4. **실행**을 클릭하십시오.

전체 모델 품질 차트

기본적으로 생성되는 전체 모델 품질 출력 항목은 막대형 차트로 표시되고 이와 연관된 모든 모델에 대해 모델 적합의 점도표가 표시됩니다. 각 대상 계열에 대해 별도의 모델이 있습니다. 모델 적합은 선택된 적합 통계량에 의해 측정됩니다. 이 예에서는 R 제곱인 기본 적합 통계량을 사용합니다.

전체 모델 품질 항목에는 대화형 변수가 포함됩니다. 변수를 사용하려면 뷰어에서 전체 모델 품질 차트를 두 번 클릭하여 항목을 활성화하십시오.

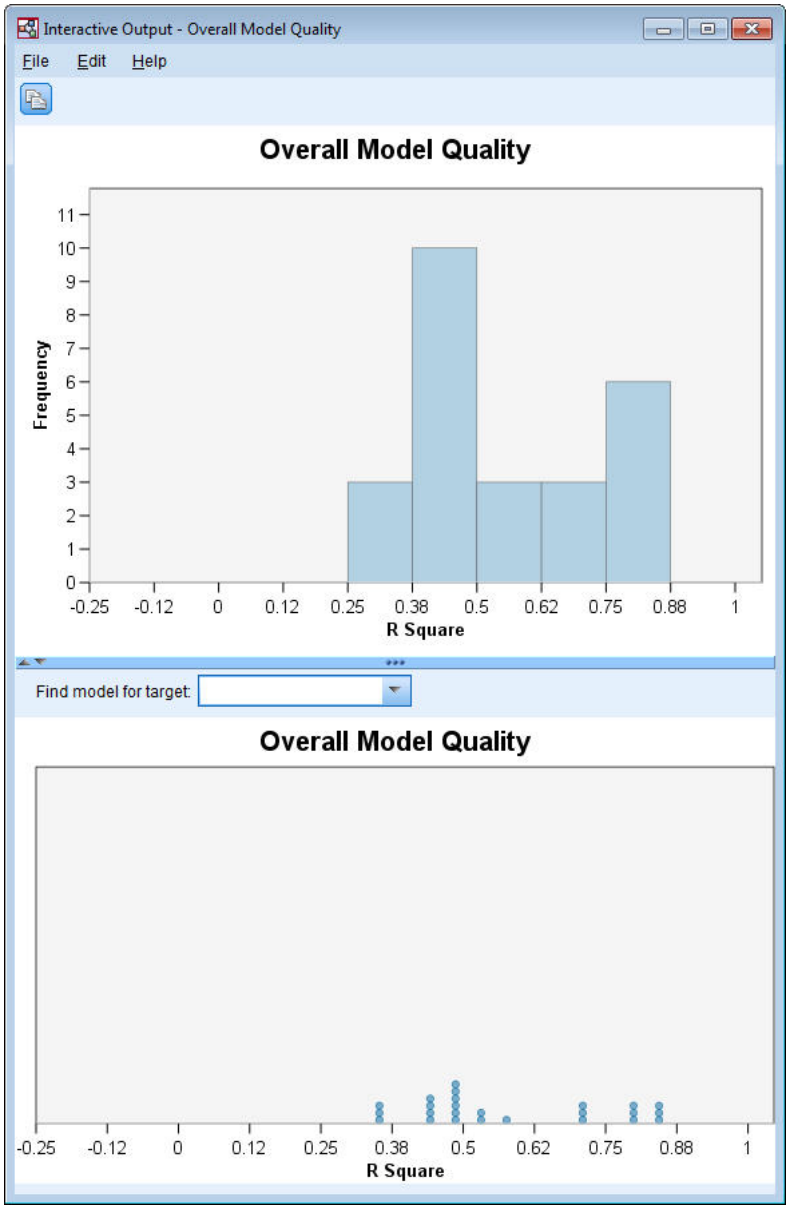


그림 *401.* 전체 모델 품질

막대형 차트에서 막대를 클릭하면 선택된 막대와 연관되는 모델만 표시하도록 점도표를 필터링합니다. 점도표에서 점 위로 마우스를 이동하면 연관된 계열의 이름 및 적합 통계량의 값을 포함하는 도구 팁이 표시됩니다. **대상에 대한 모델 찾기** 선택란에서 계열 이름을 지정하여 점도표에서 특정 목표 계열에 대한 모델을 찾을 수 있습니다.

전체 모델 시스템

기본적으로 생성되는 전체 모델 품질 출력 항목은 모델 시스템에서 계열 사이의 인과 관계에 대한 그래픽 표현을 표시합니다. 기본적으로 R 제곱 적합 통계량의 값에 의해 판별된 상위 10개의 모델에 대한 관계가 표시됩니다. 상위 모델(최적 적합 모델이라고도 함)의 수 및 적합 통계량은 시간 인과 모델링 대화 상자의 작성 옵션 탭의 표시할 계열 설정에서 지정됩니다.

전체 모델 시스템 항목에는 대화형 변수가 포함됩니다. 변수를 사용하려면 뷰어에서 전체 모델 시스템 차트를 두 번 클릭하여 항목을 활성화하십시오. 이 예에서는 시스템 내의 모든 계열 사이의 관계를 보는 것이 가장 중요합니다. 대화형 출력의 **관계 강조표시 대상** 드롭 다운 목록에서 **모든 계열**을 선택하십시오.

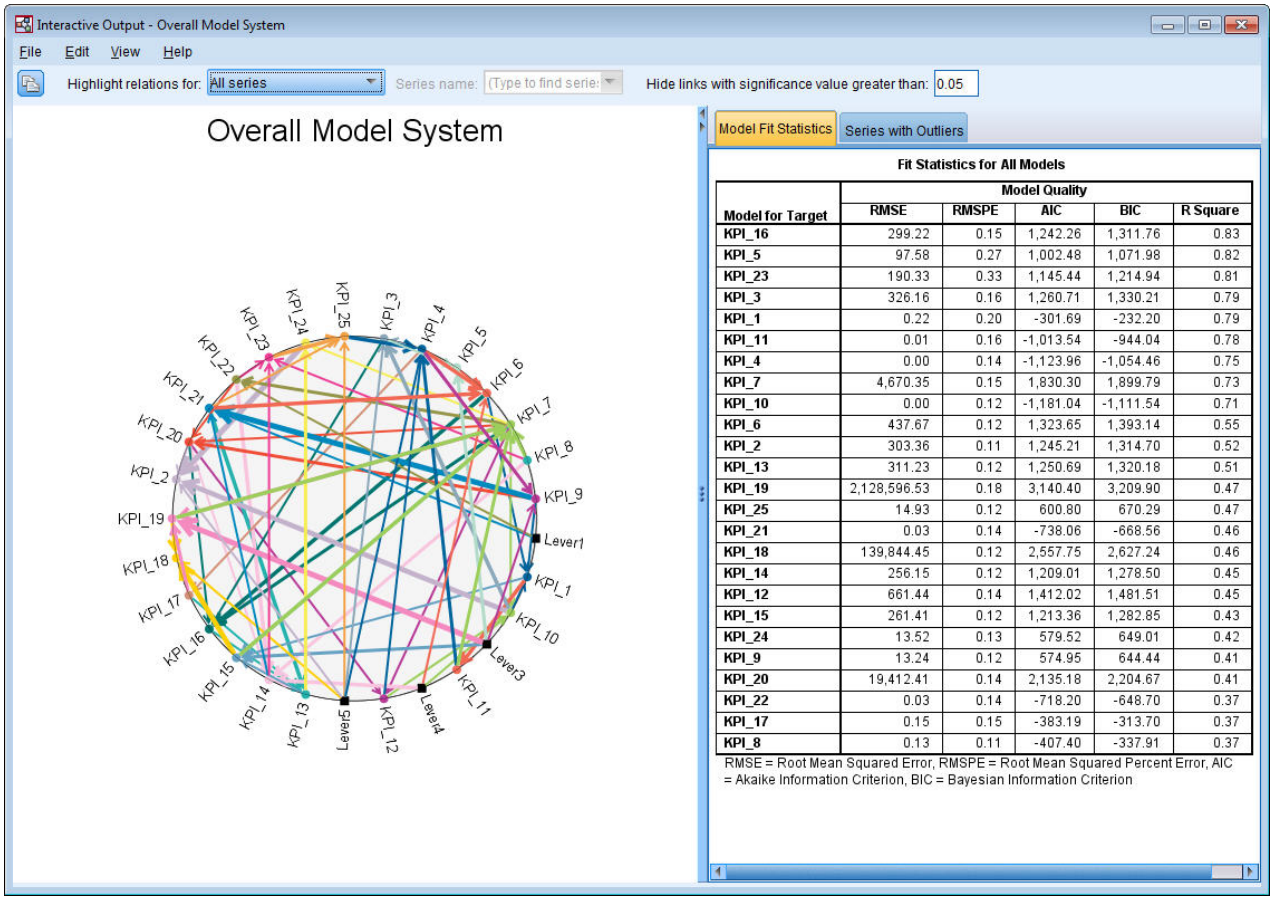


그림 *402.* 전체 모델 시스템, 모든 계열에 대한 보기

해당 입력에 특수 대상을 연결하는 모든 선은 색상이 동일하며 각 선의 화살표는 입력에서 해당 입력의 대상을 지정합니다. 예를 들어, *Lever3*은 *KPI_19*에 대한 입력입니다.

각 선의 두께는 인과 관계의 유의성을 나타냅니다. 선이 두꺼울 수록 관계 유의성이 큽니다. 기본적으로 유의수준이 0.05를 초과하는 인과 관계를 숨깁니다. 0.05 수준에서는 *Lever1*, *Lever3*, *Lever4* 및 *Lever5*만이 KPI 필드와 유의적인 인과 관계를 갖습니다. **다음보다 유의수준이 큰 링크 숨기기**로 레이블된 필드에 값을 입력하여 임계값 유의 수준을 변경할 수 있습니다.

분석에서는 *Lever* 필드 및 KPI 필드 간의 인과 관계를 찾는 것 외에 KPI 필드 사이의 관계도 찾습니다. 예를 들어, *KPI_10*이 *KPI_2*에 대한 모델에 입력으로 선택되었습니다.

보기를 필터링하여 단일 계열에 대한 관계만 표시할 수 있습니다. 예를 들어, *KPI_19*에 대한 관계만 보려면 *KPI_19*에 대한 레이블을 클릭한 다음 마우스 오른쪽 단추를 클릭하고 **계열에 대한 관계 강조 표시**를 선택하십시오.

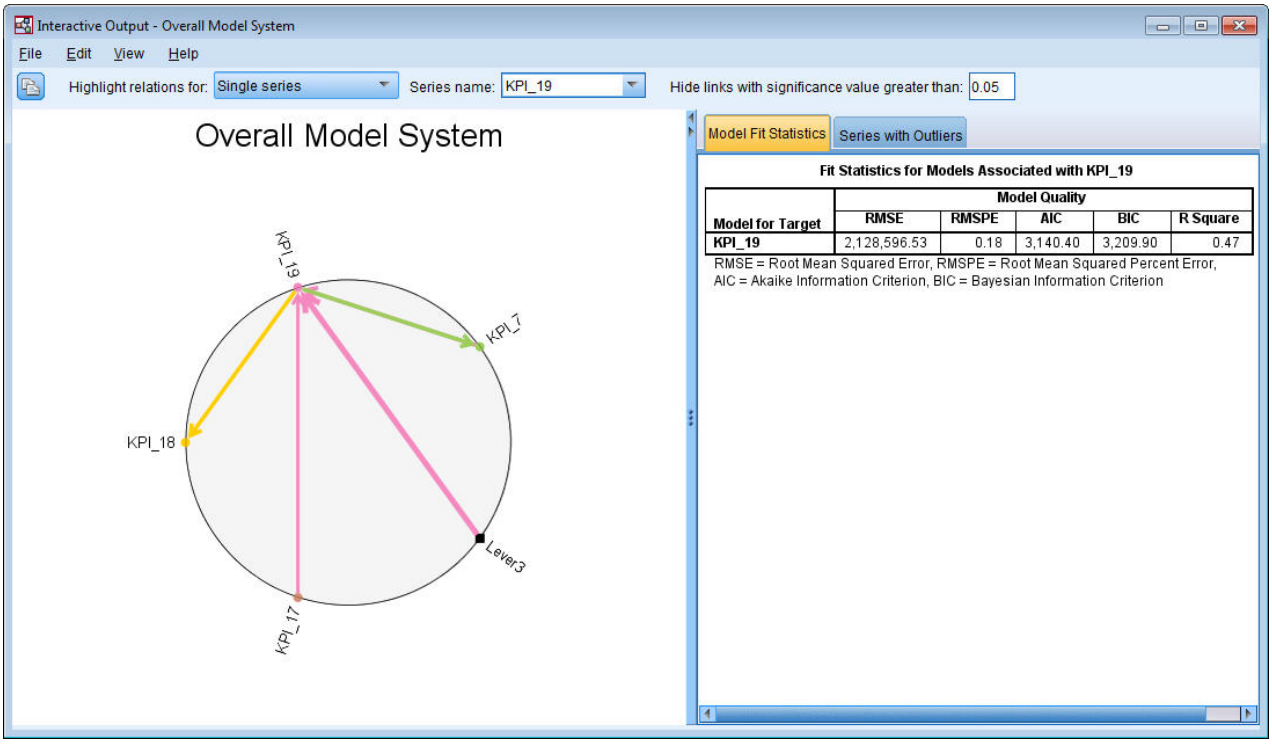


그림 *403.* 전체 모델 시스템, 단일 계열에 대한 보기

이 보기는 유의수준이 0.05 이하인 *KPI_19*에 대한 입력만 표시합니다. 또한 0.05 유의 수준에서 *KPI_19*가 *KPI_18* 및 *KPI_7* 둘 다에 대해 입력으로 선택되었음을 표시합니다.

선택된 계열에 대한 관계가 표시되는 것 외에 출력 항목에는 계열에 대해 발견된 모든 이상치에 대한 정보가 포함됩니다. **이상치가 있는 계열** 탭을 클릭하십시오.

Series with Outliers for KPI_19

Series	Time	Observed Value
KPI_19	2008-10-12	7,358,201.68
	2009-04-05	2.10E+007
	2010-09-19	6,492,157.97

그림 *404. KPI_19*에 대한 이상치

*KPI_19*에 대해 세 개의 이상치가 발견되었습니다. 발견된 연결을 모두 포함하는 모델 시스템이 있으면 이상치 발견을 넘어 특정 이상치의 원인이 될 가능성이 가장 높은 계열을 판별할 수 있습니다. 이 유형의 분석을 이상치 근본 원인 분석이라고 하며 이 케이스 연구의 후반부 주제에서 설명합니다.

영향 다이어그램

영향 다이어그램을 생성하여 특정 계열과 연관된 모든 관계를 전체적으로 볼 수 있습니다. 전체 모델 시스템 차트에서 *KPI_19*의 레이블을 클릭하고 마우스 오른쪽 단추를 클릭한 다음 **영향 다이어그램 작성**을 선택하십시오.

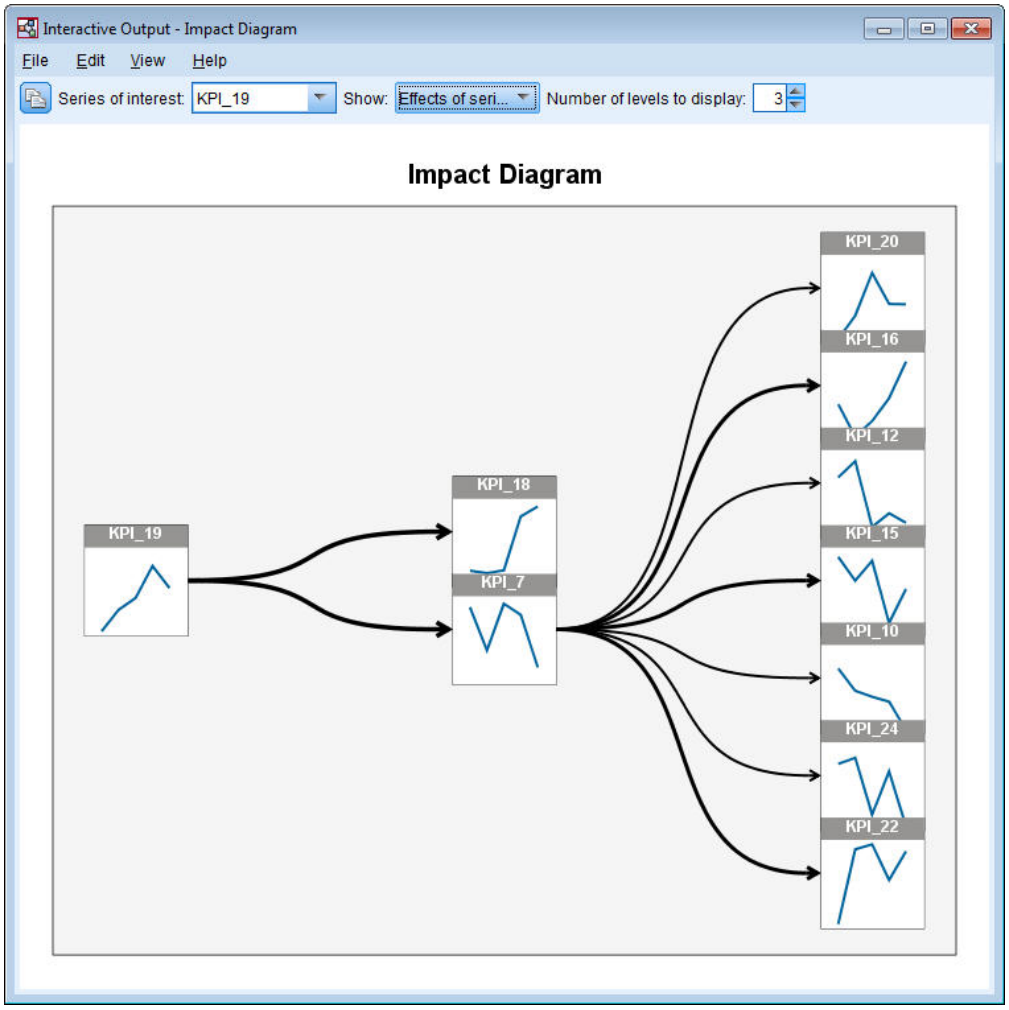


그림 *405.* 효과의 영향 다이어그램

이 예에서 보듯이 전체 모델 시스템에서 영향 다이어그램이 작성된 경우, 초기에는 선택된 계열에 의해 영향을 받는 계열을 표시합니다. 기본적으로 영향 다이어그램은 세 가지 수준의 효과를 표시하며 첫 번째 수준은 관심 계열입니다. 각각의 추가 수준은 관심 계열의 한층 간접적인 효과를 보여줍니다. 더 많거나 적은 수의 효과를 표시하도록 **표시할 수준 수**의 값을 변경할 수 있습니다. 이 예의 영향 다이어그램은 *KPI_19*가 *KPI_18* 및 *KPI_7* 둘 다에 대한 직접 입력임을 표시하나 이는 *KPI_7* 계열에 대한 효과를 통해 수많은 계열에 간접적으로 영향을 미칩니다. 전체 모델 시스템에서 선의 두께는 인과 관계의 유의성을 표시합니다.

영향 다이어그램의 각 노드에 표시되는 차트는 추정 기간의 끝에 있는 연관된 계열의 마지막 L+1 값이 및 모든 예측 값을 표시합니다. 여기서, L은 각 모델에 포함되는 시차 항의 수입니다. 연관된 노드를 한 번 클릭하기만 하면 이러한 값의 세부사항 순차도표를 얻을 수 있습니다.

노드를 두 번 클릭하면 연관된 계열이 관심 계열로 설정되고 해당 계열을 기반으로 하는 영향 다이어그램이 다시 생성됩니다. 또한 **관심 계열** 상자에서 계열 이름을 지정하여 다른 관심 계열을 선택할 수도 있습니다.

영향 다이어그램은 관심 계열에 영향을 미치는 계열을 표시할 수도 있습니다. 이러한 계열을 원인이라고 합니다. *KPI_19*에 영향을 미치는 계열을 보려면 **표시** 드롭 다운에서 **계열의 원인**을 선택하십시오.

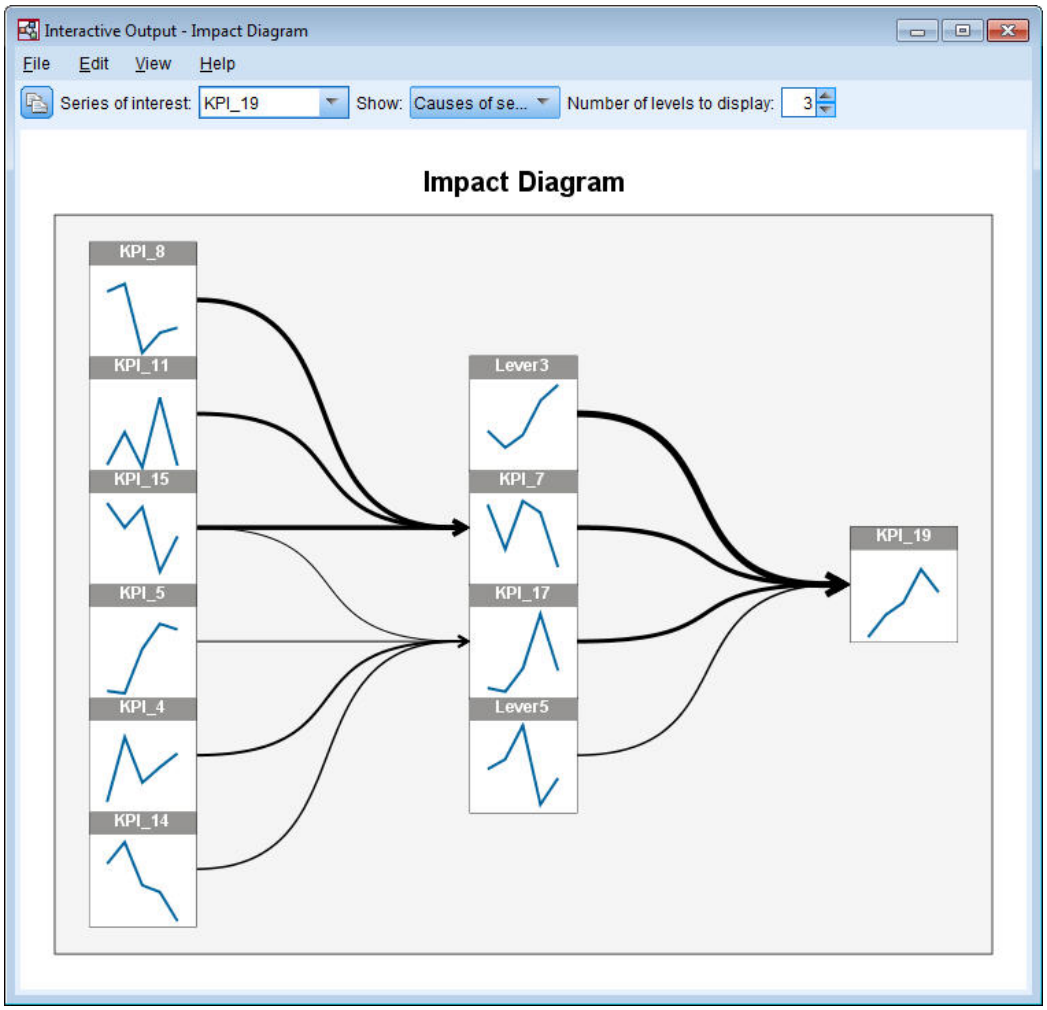


그림 *406.* 원인의 영향 다이어그램

이 보기는 *KPI_19*에 대한 모델에 네 개의 입력이 있으며 *Lever3*이 *KPI_19*와 가장 유의적인 인과 연결이 있음을 표시합니다. 또한 계열이 *KPI_7* 및 *KPI_17*에 대한 영향을 통해 *KPI_19*에 간접적으로 영향을 미침을 표시합니다. 영향에 대해 설명할 때 언급한 동일한 개념의 수준 또한 원인에 적용됩니다. 이와 유사하게 더 많거나 적은 수의 원인을 표시하도록 **표시할 수준 수**의 값을 변경할 수 있습니다.

이상치의 근본 원인 판별

시간 인과 모델 시스템이 있으면 이상치 발견을 넘어 특정 이상치의 원인이 될 가능성이 가장 높은 계열을 판별할 수 있습니다. 이 프로세스를 이상치 근본 원인 분석이라고 하며 계열별 기초에서 요청되어야 합니다. 분석에는 시간 인과 모델 시스템 및 시스템을 작성하는 데 사용된 데이터가 필요합니다. 이 예에서는 활성 데이터 세트가 모델 시스템을 작성하기 위해 사용된 데이터입니다.

이상치 근본 원인 분석을 실행하려면 다음을 수행하십시오.

1. TCM 대화 상자에서 **작성 옵션** 탭으로 이동한 다음 \ **항목 선택** 목록에서 **표시할 계열**을 클릭하십시오.

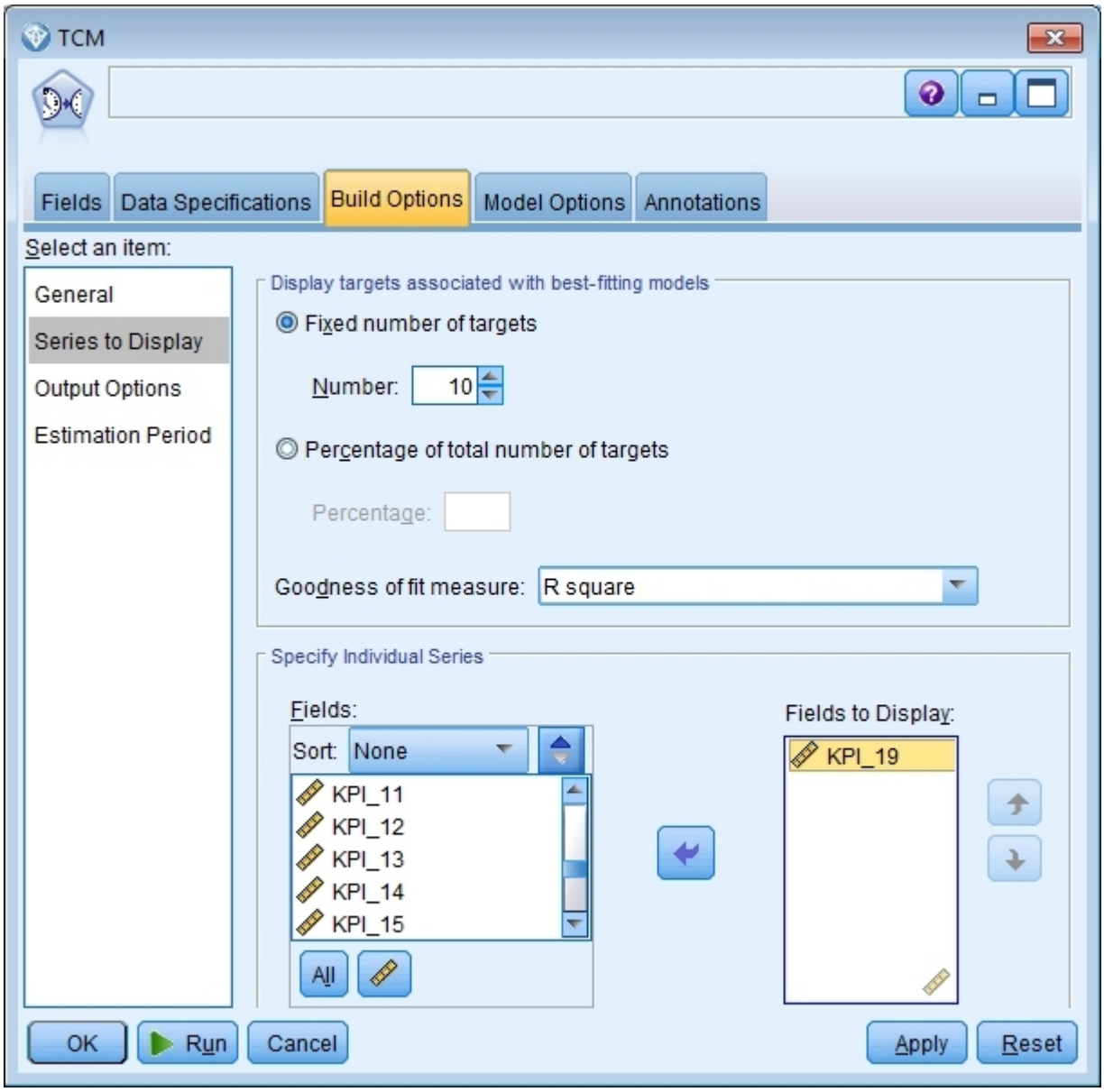


그림 *407.* 표시할 시간 인과 모델 계열

2. *KPI_19*를 **표시할 필드** 목록으로 이동하십시오.
3. 옵션 탭의 **항목 선택** 목록에서 **출력 옵션**을 클릭하십시오.

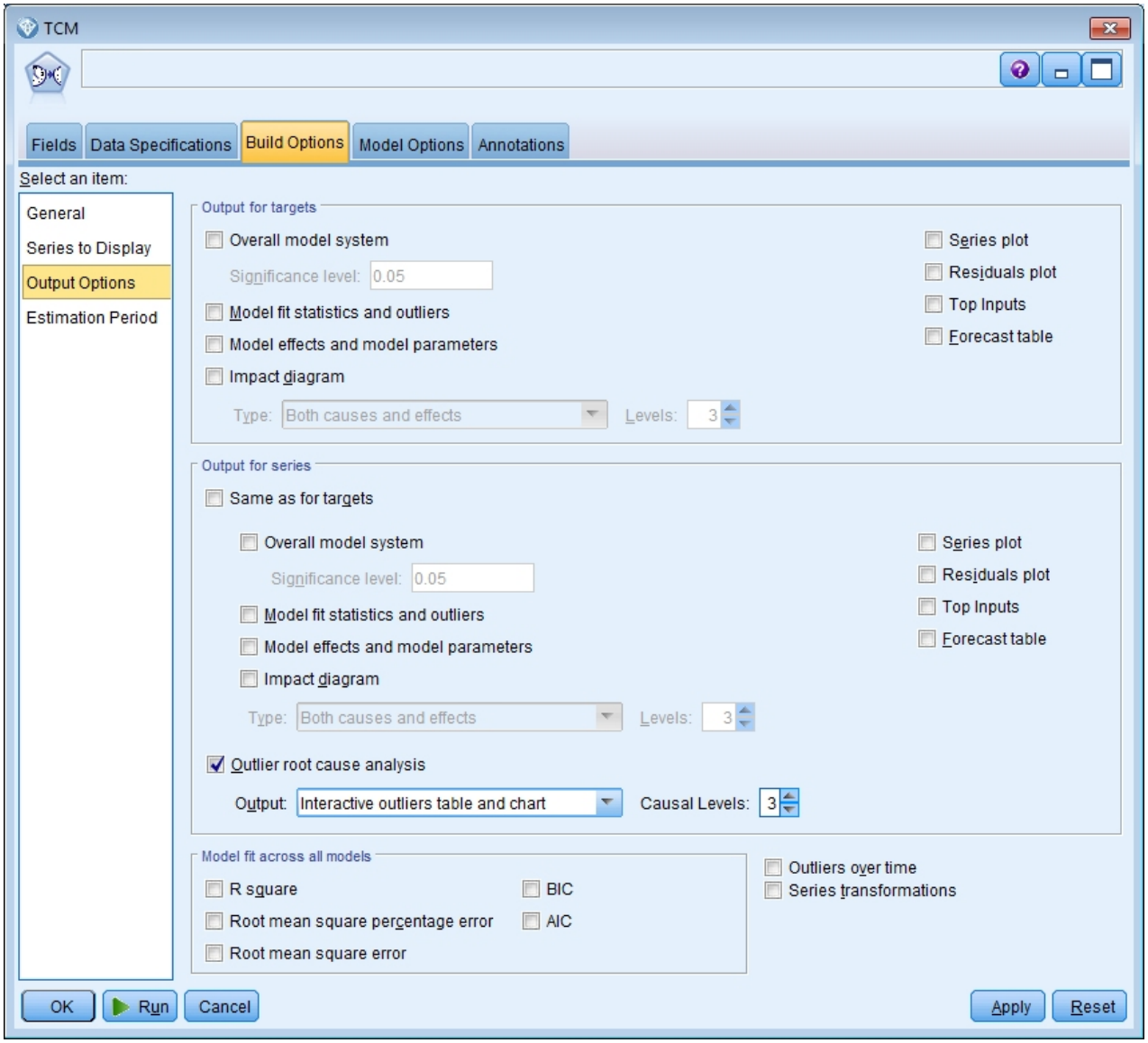


그림 *408.* 시간 인과 모델 출력 옵션

4. **전체 모델 시스템, 목표와 동일, R 제곱** 및 **계열 변환**을 선택 취소하십시오.
5. **이상치 근본 원인 분석**을 선택하고 **출력** 및 **인과 수준**에 대한 기존 설정을 유지하십시오.
6. **실행**을 클릭하십시오.
7. 뷰어에서 *KPI_19*에 대한 이상치 근본 원인 분석 차트를 두 번 클릭하여 이를 활성화하십시오.

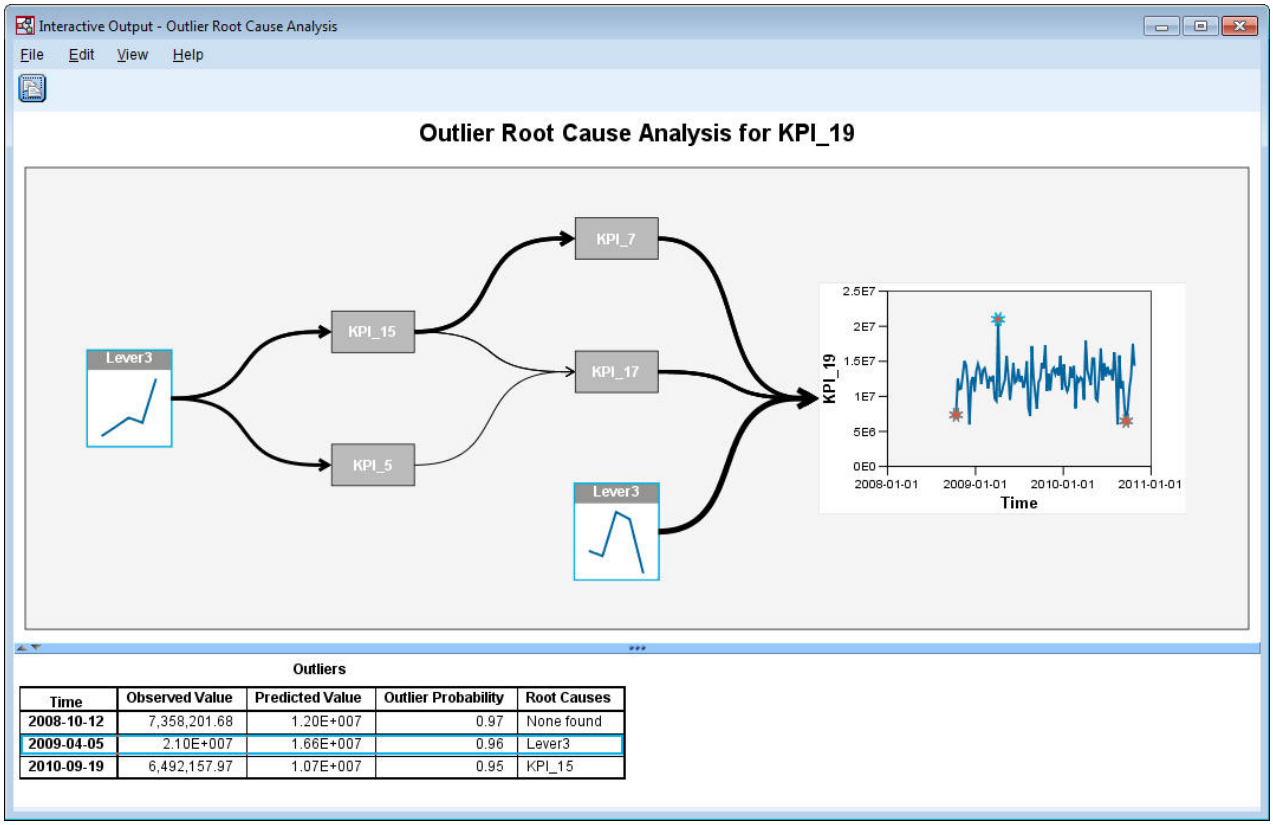


그림 *409. KPI_19*에 대한 이상치 근본 원인 분석

분석 결과가 이상치 테이블에 요약됩니다. 테이블에서는 2009-04-05 및 2010-09-19에는 이상치에 대한 근본 원인이 발견되었으나 2008-10-12에는 이상치에 대한 근본 원인이 발견되지 않았음이 표시됩니다. 이상치 테이블의 행을 클릭하면 2009-04-05에 대한 이상치에 표시된 대로 근본 원인 계열에 대한 경로가 강조표시됩니다. 또한 이 동작은 순차도표에서 선택된 이상치를 강조표시합니다. 또한 순차도표에서 이상치에 대한 아이콘을 직접 클릭하여 해당 이상치에 대한 근본 원인의 경로를 강조표시할 수 있습니다.

2009-04-05에서의 이상치의 경우, 근본 원인이 *Lever3*입니다. 다이어그램에서는 *Lever3*이 *KPI_19*에 대한 직접 입력인 것으로 표시되나 *KPI_19*에 영향을 미치는 기타 계열에 대한 영향을 통하는 *KPI_19*에 대한 간접적인 영향도 있습니다. 이상치 근본 원인 분석에 대한 구성 가능한 매개변수 중 하나는 근본 원인을 검색할 인과 수준의 수입니다. 기본적으로 세 수준을 검색합니다. 근본 원인 계열의 발생이 지정된 수의 인과 숫자까지 표시됩니다. 이 예에서는 *Lever3*이 첫 번째 인과 수준 및 세 번째 인과 수준 둘 다에서 발생합니다.

이상치에 대한 강조표시된 경로의 각 노드에는 노드가 발생하는 수준에 시간 범위가 종속되는 차트가 포함됩니다. 첫 번째 인과 수준 내의 노드의 경우, 범위가 T-1에서 T-L까지이며 여기서, T는 이상치가 발생한 시간이며 L은 각 모델에 포함된 시차 항의 수입니다. 두 번째 인과 수준 내의 노드의 경우, 범위가 T-2에서 T-L-1까지 이며 세 번째 수준의 경우, 범위가 T-3에서 T-L-2까지입니다. 연관된 노드를 한 번 클릭하기만 하면 이러한 값의 세부사항 순차도표를 얻을 수 있습니다.

시나리오 실행

시간 인과 모델 시스템이 있으면 사용자 정의 시나리오를 실행할 수 있습니다. 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. 분석에는 시간 인과 모델 시스템 및 시스템을 작성하는 데 사용된 데이터가 필요합니다. 이 예에서는 활성 데이터 세트가 모델 시스템을 작성하기 위해 사용된 데이터입니다.

시나리오를 실행하려면 다음을 수행하십시오.

1. TCM 출력 대화 상자에서 **시나리오 분석** 단추를 클릭하십시오.
2. 시간 인과 모델 시나리오 대화 상자에서 **시나리오 주기 정의**를 클릭하십시오.

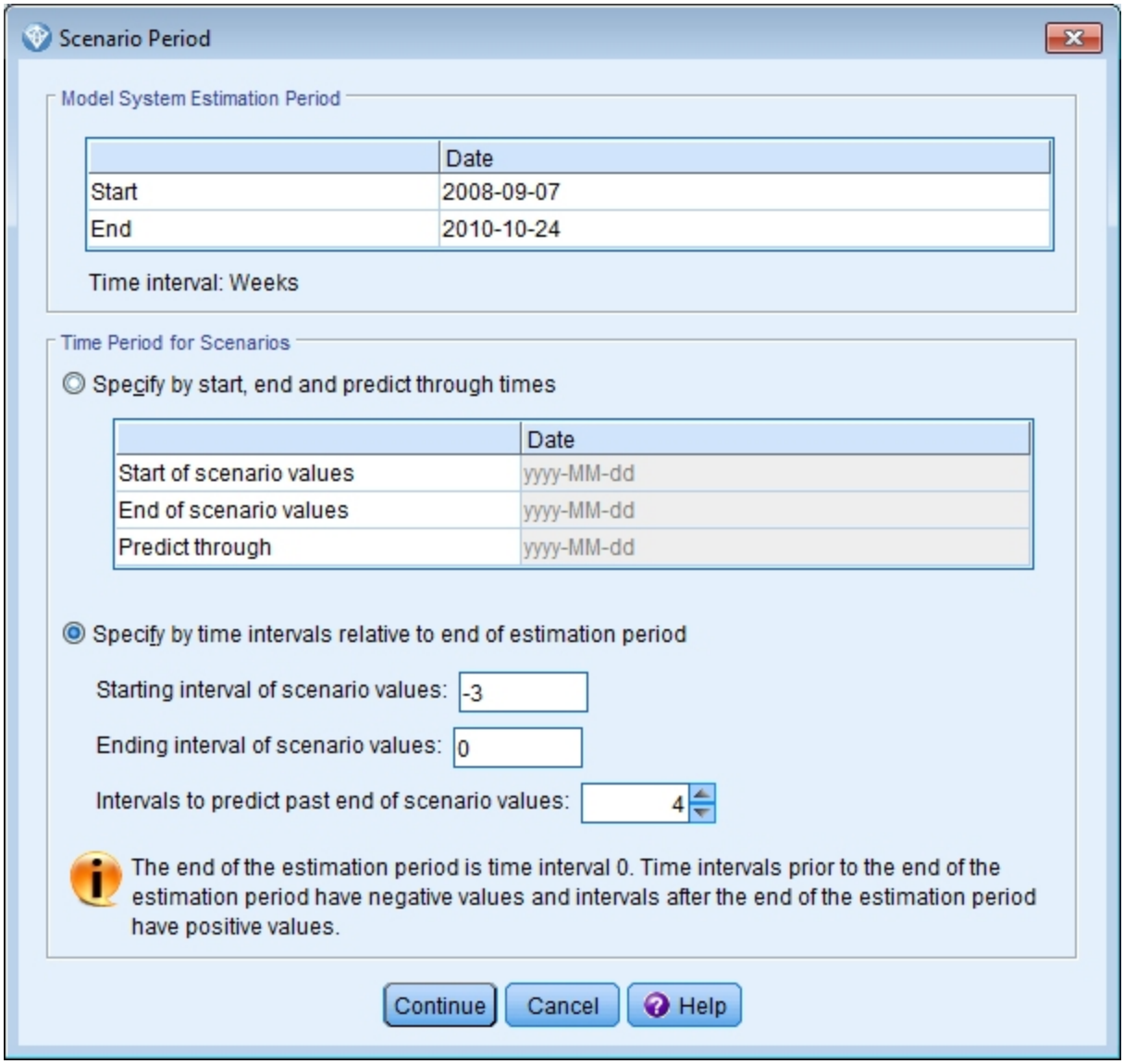


그림 *410.* 시나리오 주기

3. **추정 기간의 종료에 대해 상대적인 시간 간격에 의해 지정**을 선택하십시오.
4. 시작 구간에 -3을 입력하고 종료 구간에 0을 입력하십시오.

   이러한 설정은 각 시나리오가 추정 기간의 마지막 네 개의 시간 구간에 대해 지정된 값을 기반으로 하도록 지정합니다. 이 예의 경우, 마지막 네 개의 시간 구간은 마지막 네 주를 의미합니다. 시나리오 값이 지정된 시간 범위를 시나리오 주기라고 합니다.
5. 시나리오 값의 끝을 지나서 예측하려면 4를 입력하십시오.

   이 설정은 시나리오 주기의 끝을 지나 네 개의 시간 구간에 대해 예측을 생성하도록 지정합니다.
6. **계속**을 클릭하십시오.
7. 시나리오 탭에서 **시나리오 추가**를 클릭하십시오.

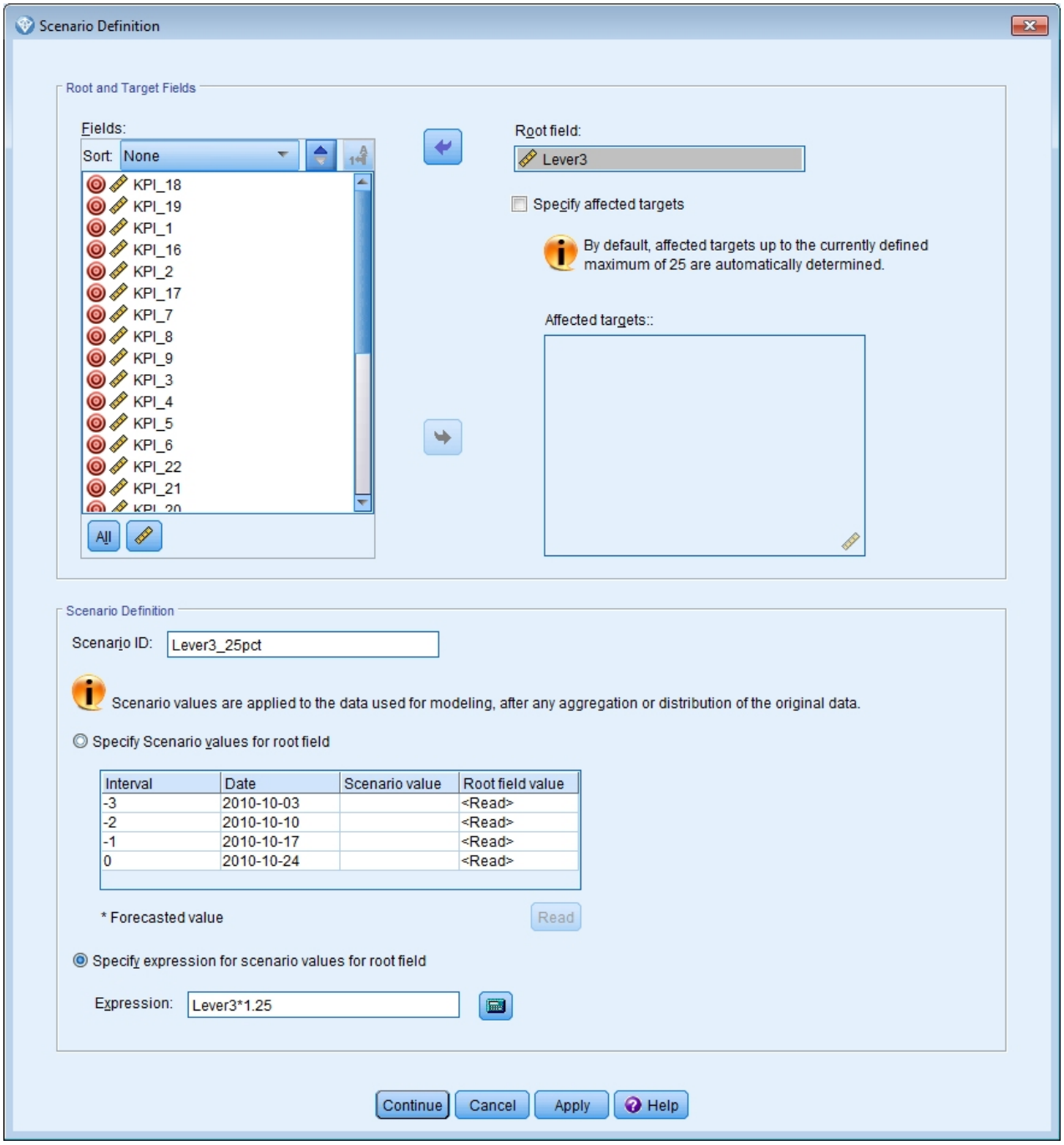


그림 *411.* 시나리오 정의

8. *Lever3*을 **루트 필드** 선택란으로 이동하여 시나리오 주기의 *Lever3*의 지정된 값이 *Lever3*에 의해 인과적으로 영향을 받는 기타 계열의 예측에 어떻게 영향을 미치는지 검사하십시오.
9. 시나리오 ID로 `Lever3_25pct`를 입력하십시오.
10. **루트 필드에 대한 시나리오 값의 표현식 지정**을 선택하고 표현식으로 `Lever3*1.25`를 입력하십시오.

이 설정은 시나리오 주기에서 *Lever3*의 값을 관측값보다 25% 크게 지정합니다. 더 복잡한 표현식의 경우, 계산기 아이콘을 클릭하여 표현식 작성기를 사용할 수 있습니다.

11. **계속**을 클릭하십시오.
12. 10 - 14 단계를 반복하여 루트 필드에 *Lever3*, 시나리오 ID에 `Lever3_50pct`, 표현식에 `Lever3*1.5`가 지정되도록 시나리오를 정의하십시오.

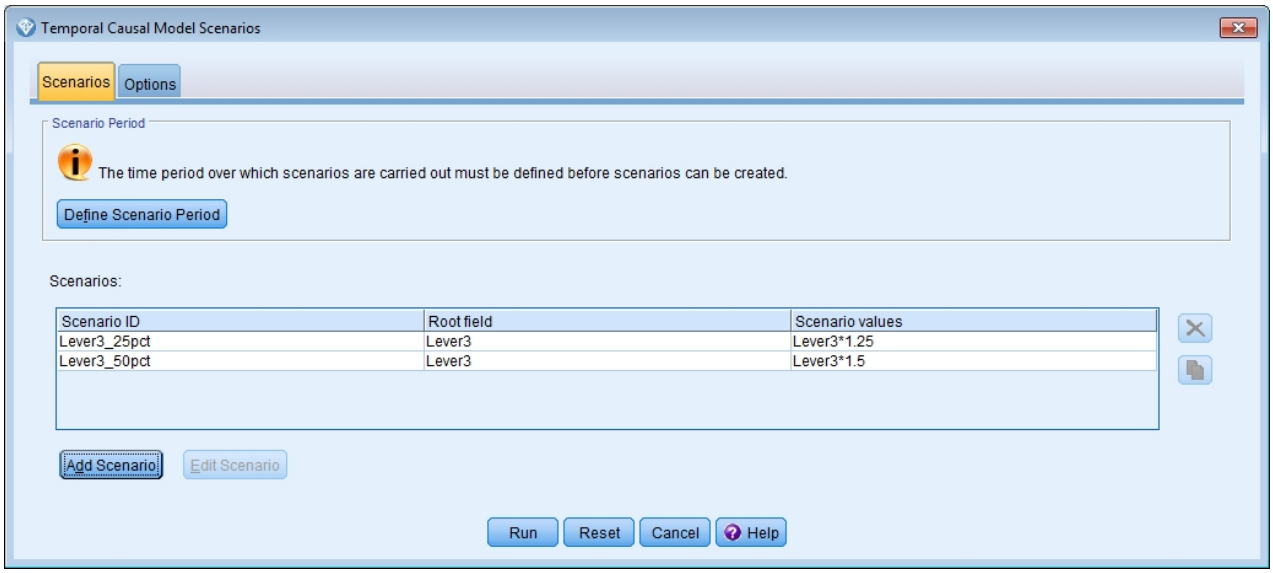
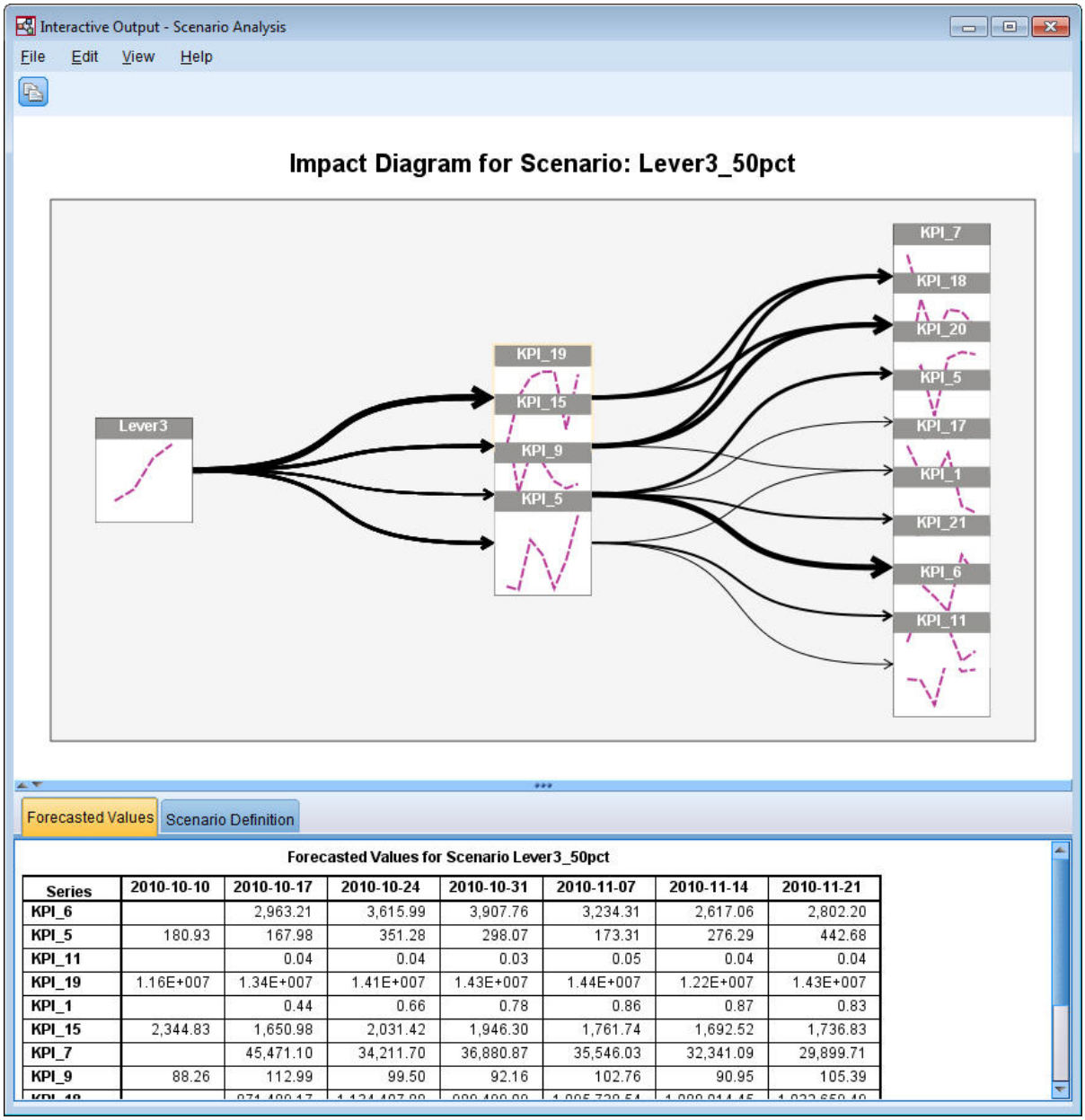


그림 *412.* 시나리오

13. **옵션** 탭을 클릭하고 영향을 받는 목표에 대한 최대 수준으로 *2*를 입력하십시오.
14. **실행**을 클릭하십시오.
15. 뷰어에서 *Lever3_50pct*에 대한 영향 다이어그램 차트를 두 번 클릭하여 이를 활성화하십시오.



Series	2010-10-10	2010-10-17	2010-10-24	2010-10-31	2010-11-07	2010-11-14	2010-11-21
KPI_6		2,963.21	3,615.99	3,907.76	3,234.31	2,617.06	2,802.20
KPI_5	180.93	167.98	351.28	298.07	173.31	276.29	442.68
KPI_11		0.04	0.04	0.03	0.05	0.04	0.04
KPI_19	1.16E+007	1.34E+007	1.41E+007	1.43E+007	1.44E+007	1.22E+007	1.43E+007
KPI_1		0.44	0.66	0.78	0.86	0.87	0.83
KPI_15	2,344.83	1,650.98	2,031.42	1,946.30	1,761.74	1,692.52	1,736.83
KPI_7		45,471.10	34,211.70	36,880.87	35,546.03	32,341.09	29,899.71
KPI_9	88.26	112.99	99.50	92.16	102.76	90.95	105.39

그림 *413.* 시나리오에 대한 영향 다이어그램*: Lever3_50pct*

영향 다이어그램은 원 계열 *Lever3*에 의해 영향을 받는 계열을 표시합니다. 사용자가 영향을 받는 목표에 대한 최대 수준으로 *2*를 지정했으므로 두 수준의 효과가 표시됩니다.

예측값 테이블에는 효과의 두 번째 수준까지, *Lever3*에 의해 영향을 받는 모든 계열에 대한 예측이 포함됩니다. 효과의 첫 번째 수준 내의 목표 계열에 대한 예측은 시나리오 주기의 시작 이후 첫 번째 시간 주기에서 시작합니다. 이 예에서는 첫 번째 수준의 목표 계열에 대한 예측이 2010-10-10에 시작됩니다. 효과의 두 번째 수준 내의 목표 계열에 대한 예측은 시나리오 주기의 시작 이후 두 번째 시간 주기에서 시작합니다. 이 예에서는 두 번째 수준의 목표 계열에 대한 예측이 2010-10-17에 시작됩니다. 예측의 지그재그형 특성은 시계열 모델이 입력의 시차 값을 기반으로 한다는 사실을 반영합니다.

16. *KPI_5*에 대한 노드를 클릭하여 세부사항 순차 다이어그램을 생성하십시오.

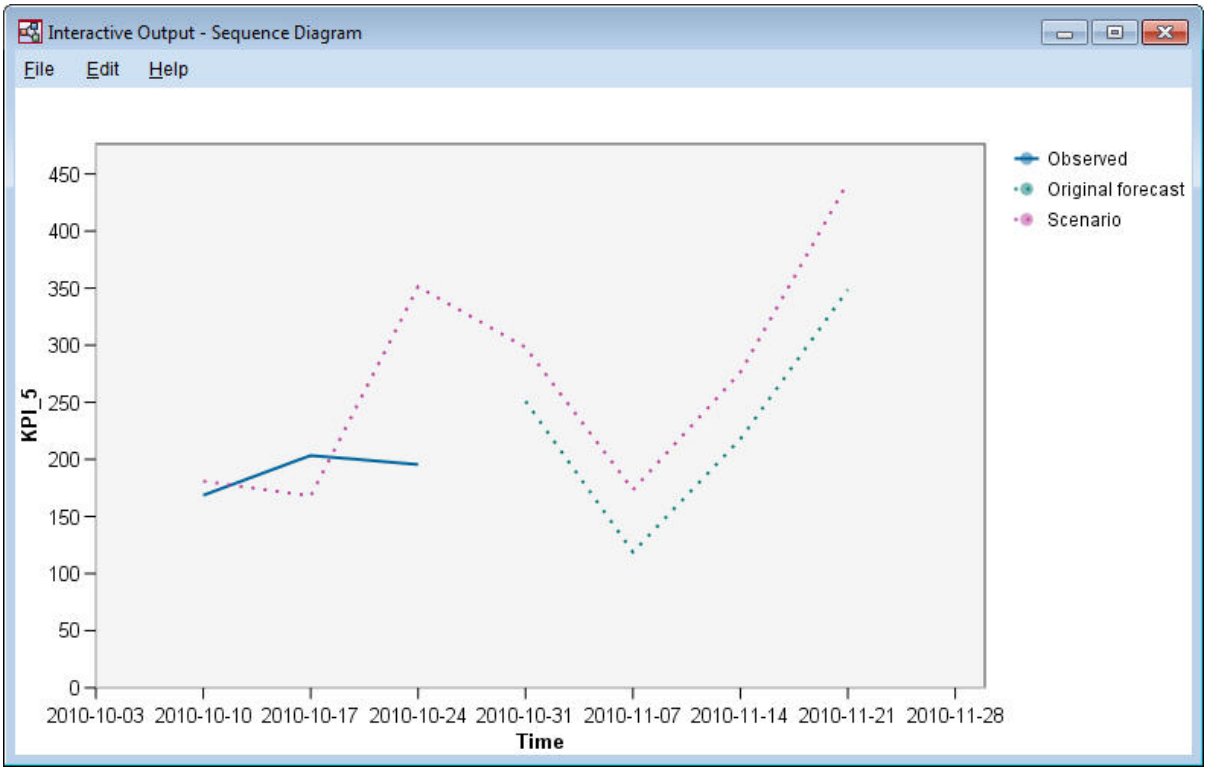


그림 *414. KPI_5*에 대한 순차 다이어그램

순차도표는 시나리오에서 예측된 값을 표시하며 시나리오가 없는 경우의 계열 값도 표시합니다. 시나리오 주기에 추정 주기 내의 시간이 포함되면 계열의 관측값이 표시됩니다. 추정 기간의 종료 이후의 시간인 경우, 원본 예측값이 표시됩니다.

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.

색인

[가]

감마회귀
  일반화 선형 모델에서 295
고유값
  판별 분석 255
공변량 평균값
  Cox 회귀분석 내 321
관리자 16
구간 중도절단 생존 데이터
  일반화 선형 모델에서 259
구조행렬
  판별 분석 255
그래프 노드 90
그룹화된 생존 데이터
  일반화 선형 모델에서 259
기본 창 14

[나]

나머지
  의사결정 목록 모델 116
낮은 확률 검색
  의사결정 목록 모델 116
너깃
  정의된 16
노드 7

[다]

단계적 방법
  판별 분석 253
  Cox 회귀분석 내 319
단축키 22
  키보드 22
대화형 목록 뷰어
  미리보기 분할창 116
  사용 116
  애플리케이션 예 116
데이터
  모델링 94, 96, 97
  보기 86
  읽기 83
  조작 91
도구 모음 18
도메인 이름(Windows)
  IBM SPSS Modeler Server 8

[마]

마우스
  IBM SPSS Modeler에서 사용 22
마우스 가운데 단추
  시뮬레이션 22
마이닝 작업
  의사결정 목록 모델 116
명령행
  IBM SPSS Modeler 시작 8
모델 효과 검정
  일반화 선형 모델에서 263, 275, 290
모델링 94, 96, 97
모수 추정값
  일반화 선형 모델에서 265, 277, 290, 299
문서 3

[바]

범주형 변수 코딩
  Cox 회귀분석 내 318
보기에 맞춰 스트림 스케일링 21
복사 18
분류표
  판별 분석 257
분석 노드 97
붙여넣기 18
비밀번호
  IBM SPSS Analytic Server 11
  IBM SPSS Modeler Server 8
비주얼 프로그래밍 13

[사]

사용자 ID
  IBM SPSS Modeler Server 8
상태 모니터링 241
생성된 모델 팔레트 16
생존함수 곡선
  Cox 회귀분석 내 322
서버
  로그인 8
  연결 추가 10
  COP에서 서버 검색 10
세그먼트
  스코어링에서 제외 116
  의사결정 목록 모델 116
소개
  IBM SPSS Modeler 7
소매 분석 235
소스 노드 83
스크립팅 24
스트림 7, 14
  보기 스케일링 21
  작성 83
시간 인과 모델
  자습서 365
  케이스 연구 365
실행 중단 18
실행 취소 18
싱글 사인온 8

[아]

아래로 검색
  의사결정 목록 모델 116
아이콘
  옵션 설정 21
애플리케이션 예제 3
여러 IBM SPSS Modeler 세션 13
연결
  서버 클러스터 10
  IBM SPSS Analytic Server로 11
  IBM SPSS Modeler Server로 8, 10
영역도
  판별 분석 256
예
  다항 로지스틱 회귀분석 139, 149
  문자열 길이 감소 107
  베이지안 네트워크 215, 223
  새 차량 오퍼링 평가 355
  세포 표본 분류 301

예 (계속)
입력 문자열 길이 감소 107
재분류 노드 107
카탈로그 통신판매 189
통신 139, 149, 163, 181, 247
판별 분석 247
KNN 355
SVM 301
예제
개요 5
상태 모니터링 241
소매 분석 235
애플리케이션 안내서 3
장바구니 분석 347
예측변수
분석을 위한 선택 99
선별 99
중요도 순위화 99
예측변수 선별 99
예측변수 순위화 99
웹 노드 90
위험함수 곡선
Cox 회귀분석 내 323
음이항회귀
일반화 선형 모델에서 291
의사결정 목록 노드
애플리케이션 예 113
의사결정 목록 모델
생성 137
세션 정보 저장 137
애플리케이션 예 113
Excel 템플리트 수정 135
Excel과 연결 129
Excel을 사용하는 사용자 정의 측도 129
의사결정 목록 뷰어 116
인쇄 23
스트림 21
일반화 선형 모델
관련 프로시저 283, 294, 299
모델 효과 검정 263, 275, 290
모수 추정값 265, 277, 290, 299
적합도 289, 293
총괄 검정 289
포아송 회귀분석 285
임시 디렉토리 12

[자]

잘라내기 18
장바구니 분석 347
적합도
일반화 선형 모델에서 289, 293
준비 91
중도절단 케이스
Cox 회귀분석 내 317
중요도
예측변수 순위화 99

[차]

총괄 검정
일반화 선형 모델에서 289
Cox 회귀분석 내 319
최소화 20
출력 16

[카]

캔버스 14
크기 조정 20
클래스 17

[타]

테넌트
IBM SPSS Analytic Server 11
테이블 노드 86

[파]

파생 노드 91
판별 분석
고유값 255
구조행렬 255
단계적 방법 253
분류표 257
영역도 256
Wilks의 람다 255
팔레트 14
포아송 회귀분석
일반화 선형 모델에서 285
포트 번호
IBM SPSS Modeler Server 8, 10
표현식 작성기 91
프로세스 조정자 10
프로젝트 17
필드
분석을 위한 선택 99
선별 99
중요도 순위화 99
필드선택 노드
예측변수 선별 99
예측변수 순위화 99
중요도 99
필드선택 모델 99
필터링 94

[하]

호스트 이름
IBM SPSS Modeler Server 8, 10
확대/축소 18

C

CLEM
소개 24
COP 10
COP에서 연결 검색 10
Cox 회귀분석
범주형 변수 코딩 318
변수 선택 319
생존함수 곡선 322
위험함수 곡선 323
중도절단 케이스 317
CRISP-DM 17

E

Excel
의사결정 목록 모델과 연결 129
의사결정 목록 템플리트 수정 135

I

IBM SPSS Analytic Server
다중 연결 11
연결 11
IBM SPSS Modeler 1, 13
개요 7
명령행에서 실행 8
문서 3

IBM SPSS Modeler (계속)
시작하기 7
IBM SPSS Modeler Server 2
도메인 이름(Windows) 8
비밀번호 8
사용자 ID 8
포트 번호 8, 10
호스트 이름 8, 10
IBM SPSS Modeler Server 연결 추가 10
IBM SPSS Modeler Server에 로그인 8

M

Microsoft Excel
의사결정 목록 모델과 연결 129
의사결정 목록 템플리트 수정 135

S

Self-Learning 반응 모델 노드
모델 찾아보기 209
스트림 작성 204
스트림 작성 예 204
애플리케이션 예 203
SLRM 노드
모델 찾아보기 209
스트림 작성 204
스트림 작성 예 204
애플리케이션 예 203

U

URL
IBM SPSS Analytic Server 11

V

var. 파일 노드 83

W

Wilks의 람다
판별 분석 255