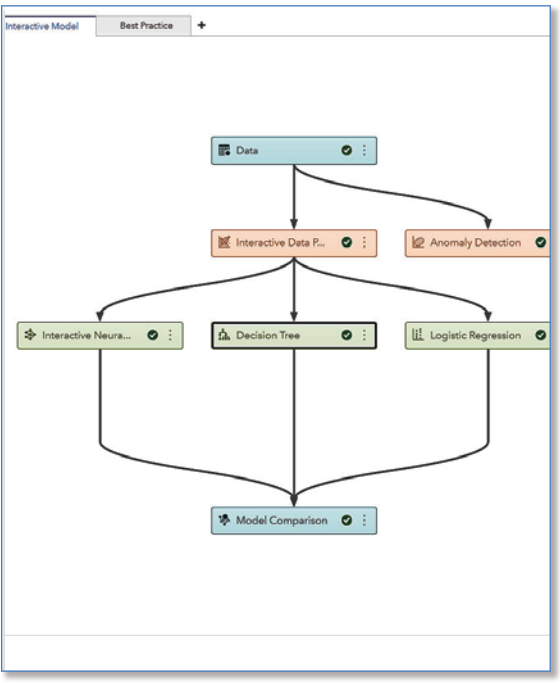
SAS® Visual Data Mining and Machine Learning

복잡한 분석 문제 해결을 위한 하나의 통합 솔루션



SAS® Visual Data Mining and Machine Learning는 어떤 솔루션인가?

SAS® Visual Data Mining and Machine Learning은 분석 라이프 사이클과 관련된 모든 구성 과정을 처리할 수 있는 시각적 인터페이스를 제공합니다. 정형/비정형 데이터를 분석할 수 있는 혁신적인 머신 러닝 기법 외에도 기타 모든 작업들이 분석 프로세스에 통합됩니다. 데이터 준비 및 탐색에서부터 모델 개발 및 배포에 이르기까지 모든 사용자가 동일한 통합 환경에서 작업하게 됩니다. 확장성과 탄력성을 모두 갖춘 프로세싱은 유연성과 속도를 바탕으로 복잡한 질의에 대한 해답을 빠르게 제시합니다.

SAS® Visual Data Mining and Machine Learning이 중요한 이유는?

SAS Visual Data Mining and Machine Learning은 고급 분석, 데이터 준비, 시각화, 모델 평가 및 모델 배포를 단일 환경으로 통합한 최초의 솔루션으로, 자주 사용되는 오픈 소스 언어를 통한 프로그래밍을 지원합니다. 일관된 협업 환경에서 반복적인 과정을 통한 결과 산출로 기업 프로세스를 개선하는 동시에 새로운 성장 기회를 찾아낼 수 있습니다.

SAS® Visual Data Mining and Machine Learning은 누구를 위한 솔루션인가?

복잡한 대용량의 데이터를 분석하여 예측 모델을 개발해야 하는 사용자라면 누구나 사용할 수 있습니다. 여기에는 데이터 사이언티스트, 통계 전문가, 데이터 마이닝 전문가, 비즈니스 분석가, 데이터 엔지니어 및 연구원 등이 포함됩니다.

데이터 수집량은 끊임없이 증가하고 있습니다. 반면 전문적인 데이터 사이언티스트와 분석 전문가는 부족합니다. 기업들은 날로 복잡해지는 문제에 대한 해답을 적시에 얻지 못해 어려움을 겪고 있습니다. 신종 사기 형태를 파악하기 위해 모든 거래를 분석하거나 고객 만족도를 높이기 위해 날로 늘어나는 소셜 미디어 채팅 데이터를 분석하고, 신속하고 정확한 권고 시스템을 개발하려는 기업은 정교한 머신 러닝 소프트웨어를 통해 중요한 문제를 해결할 수 있습니다.

SAS Visual Data Mining and Machine Learning은 통합된 시각적 파이프라인 인터페이스를 통해 가공되지 않은 데이터에서 새로운 인사이트를 도출하기 위한 모든 과정을 처리합니다. 다양한 분야의 분석 전문가들은 데이터 액세스 및 준비, 피처 엔지니어링, 탐색적 분석, 머신 러닝 모델 비교, 예측 모델 구현을 위한 스코어 코드 생성 작업을 빠르게 수행할 수 있습니다.

주요 특징

- **분석 팀의 생산성 향상** 전체 머신 러닝 파이프라인을 지원하기 때문에 다양한 사용자들이 단일 협업 환경에서 정교한 모델의 개발 및 확장을 통해 정확한 결과를 얻을 수 있습니다.
- **데이터와 의사결정 간 지연 시간 최소화** 대화형 방식의 시각적 프로그래밍 인터페이스를 통해 데이터 준비부터 모델 구축까지 소요되는 시간을 크게 단축할 수 있습니다. 또한 고속 프로세싱으로 빠른 결과를 제공합니다.
- **복잡한 분석 문제의 빠른 해결** 최신 분산형 In-Memory 플랫폼인 SAS® Viya기반에서 획기적인 속도의 예측 모델링 및 머신러닝 기능을 지원합니다. 메모리에 데이터가 저장되어 반복적인 분석 작업 시 데이터 로딩 시간을 단축할 수 있습니다. 따라서 수 초/수 분 만에 분석 모델링 처리를 완료하여 아무리 까다로운 문제라도 빠르게 해답을 찾을 수 있습니다.
- **최적의 해답을 찾기 위한 다양한 접근법 탐색** 분산 분석 엔진의 뛰어난 성능과 머신러닝 파이프 라인을 위한 풍부한 피쳐 빌딩 블록을 사용하면 다양한 접근법을 보다 쉽고 빠르게 탐색 및 비교할 수 있습니다. 또한 자동 튜닝(Auto-Tunining) 기능을 통해 다양한 시나리오를 테스트하여 가장 적합한 모델을 찾고 높은 수준의 신뢰도를 갖는 해답을 얻을 수 있습니다.
- **다양한 프로그래밍 언어 지원** Python, Java, Lua 프로그래머는 SAS 프로그램 작성법을 익히지 않고도 검증된 SAS 머신러닝 알고리즘을 사용할 수 있습니다.
- **자동 생성된 SAS 스코어 코드로 신속한 예측 모델 배포** 모든 머신러닝 모델을 구현할 수 있는 스코어 코드가 다양한 프로그래밍 언어로 자동 생성되므로 시간이 훨씬 더 절약됩니다.

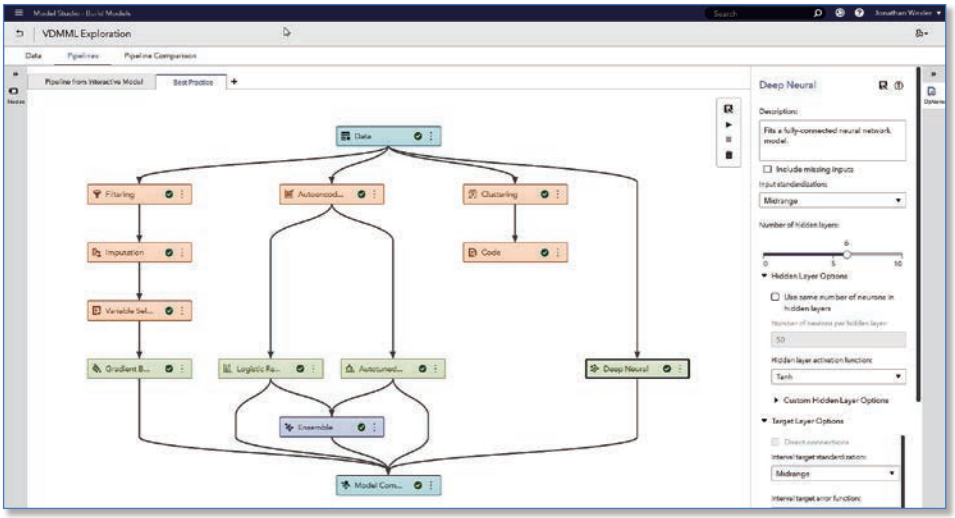
주요 기능

SAS Visual Data Mining and Machine Learning은 데이터 액세스 및 전처리에서부터 정교한 모델 개발 및 배포에 이르기까지 머신 러닝과 딥 러닝의 모든 요소를 포함하는 종합적인 시각 환경을 제공합니다. In-Memory 분산 프로세싱을 통해 대량의 데이터와 복잡한 모델링을 처리하여 해답을 신속하게 도출하고 리소스를 보다 효율적으로 사용할 수 있습니다.

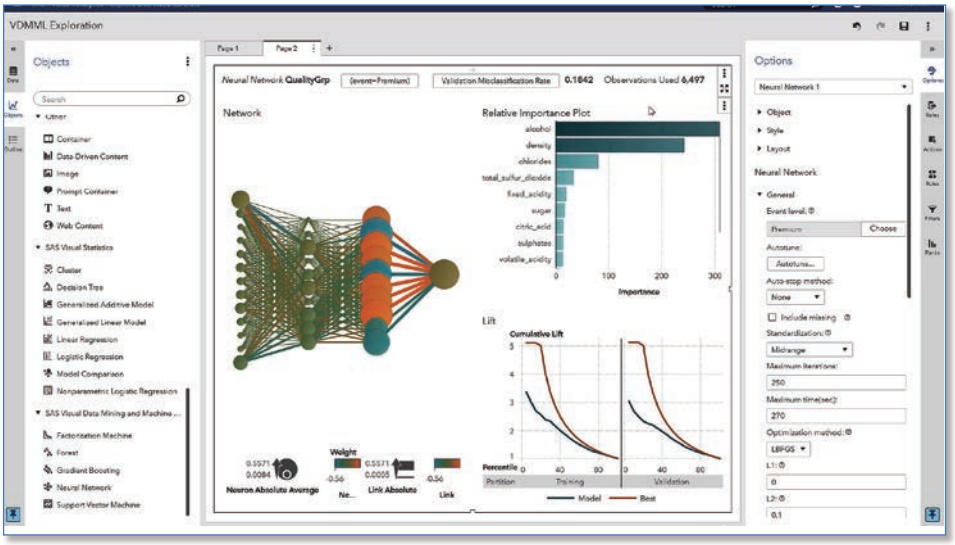
유연하고 접근이 쉬운 시각적 분석 환경

다수의 사용자들이 간편한 인터페이스를 통해 정형/비정형 데이터를 크기에 상관없이 동시에 분석할 수 있습니다. 시각적 파이프라인이 분석 라이프 사이클을 논리적 시퀀스로 분할하여 각 프로젝트와 이의 목표를 정의합니다. 데이터 분리는 산발적으로 수행될 수 있습니다. 시각적 인터페이스(Model Studio)는 데이터 준비부터 전처리 및 새로운 변수 생성(Feature Engineering), 탐색, 모델 개발 및 배포에 이르는 가장 일반적인 머신 러닝 단계에 대한 통합된 환경을 제공합니다. 대화형 작업 방식을 통해 정교한 알고리즘을 복잡한 대용량 데이터에 간편하게 적용할 수 있습니다. 이는 향후 자동화 작업에 사용할 수 있는 SAS 코드를 생성합니다. 또한 코드와 모범 사례 템플릿도 쉽게 공유할 수 있습니다. Model Studio는 모델 개발, 확장 및 공유를 위한 뛰어난 협업 환경을 제공합니다.

- 전체 분석 라이프 사이클 프로세스를 위한 시각적 인터페이스
- 코딩이 불필요한 드래그-앤-드롭 방식의 인터랙티브 인터페이스
- 파이프라인의 각 노드마다 자동 코드 생성 지원
- 사용자가 머신 러닝 작업을 빠르게 시작할 수 있는 모범 사례 템플릿(초급, 중급 또는 고급)
- 협업 환경을 통해 다른 사용자 사이에서도 데이터, 코드 조각 및 모범 사례를 손쉽게 공유



시각적 파이프라인 접근은 복잡한 머신 러닝 및 딥 러닝 모드를 생성 및 배포하기 위한 효율적인 협업 환경을 제공합니다.



SAS Visual Data Mining and Machine Learning은 사용자에게 고급 머신 러닝 알고리즘의 빠른 배포와 간편한 해석 기능을 제공합니다.

뛰어난 확장성이 보장되는 In-Memory 분석 프로세싱

이 솔루션은 다수의 사용자가 In-Memory 데이터에 안전하게 동시 액세스할 수 있는 환경을 제공합니다. 데이터와 분석 워크로드 작업이 다수의 노드로 동시에 분산되면 각 노드에서 매우 빠른 속도의 멀티스레드 방식으로 처리됩니다. 모든 데이터, 테이블 및 객체가 필요한 시간만큼 메모리에 저장되므로 데이터 프로세싱이 효율적으로 이뤄집니다. 마지막으로 자체 내장된 고장 방지 능력 및 메모리 관리 기능으로 고급 워크플로우를 데이터에 적용하여 프로세스를 차질 없이 완료할 수 있습니다.

대용량 데이터 및 분석 프로세싱을 위한 런타임이 크게 줄어들 뿐만 아니라 네트워크 트래픽이 감소하기 때문에 최신 멀티코어 아키텍처를 최대한 이용하여 원하는 해결책을 빠르게 찾아낼 수 있습니다.

- 대용량 데이터 세트를 대상으로 한 복잡한 분석에 분산형 In-Memory 프로세싱을 적용하여 신속하게 해답 도출
- 데이터를 다시 로드하거나 중간 결과를 디스크에 기록할 필요 없이 분석 작업을 단일 In-Memory 작업으로 연계해서 처리
- 다수의 사용자가 동일한 In-Memory 데이터에 동시에 액세스할 수 있어 효율성 향상
- 데이터 및 중간 결과가 필요한 시간만큼 메모리에 상주하므로 지연 시간 감소
- 워크로드 관리 기능이 내장되어 있어 컴퓨팅 리소스 활용에 최적화
- 자체 내장된 결함 허용 관리 기능으로 맡은 작업을 차질 없이 완료

혁신적 통계, 데이터 마이닝 및 머신 러닝 기법

SAS Visual Data Mining and Machine Learning은 단일 환경에서 최신 통계, 머신 러닝, 딥 러닝, 텍스트 분석 등 매우 광범위한 알고리즘을 구현합니다.

분석 기능으로는 군집화, 여러 가지 회귀분석, 랜덤 포레스트, 그래디언트 부스팅 모델, 서포트 벡터 머신, 자연어 처리(NLP; Natural Language Processing), 주제 탐지 등이 있습니다. 이러한 강력한 기법들은 정형/비정형 데이터로부터 새로운 패턴, 트렌드, 관계 등을 도출합니다. 솔루션 역시 맞춤형 추천 시스템을 개발할 수 있도록 고급 추정(Matrix Factorization)기능을 제공합니다.

SAS Visual Data Mining and Machine Learning은 높은 속도로 대량의 데이터 세트를 처리할 수 있는 성능 때문에 딥 러닝 기법에 특히 이상적입니다. 딥 러닝 알고리즘에는 심층 신경망과 이미지 분류를 위한 컨볼루션 신경망, 그리고 더욱 정확한 텍스트 분석을 위한 순환 신경망 등이 있습니다.

신경망, 그래디언트 부스팅, 랜덤 포레스트 등 복잡한 학습 알고리즘은 최적의 성능에 맞게 자동으로 조정되어 시간과 리소스를 절감하는 데 매우 유용합니다.

- 랜덤 포레스트:
  - 단일 타깃을 예측하는 의사결정트리의 자동 앙상블
  - 독립된 트레이닝 실행 자동 배포
  - 모델 모수에 대한 지능형 자동 조정 지원
  - 운영 환경 스코어링을 위한 SAS 코드 자동 생성
- 그래디언트 부스팅:
  - 선택된 레이블 변수에 대한 최적의 데이터 파티션을 자동으로 반복 검색
  - 잔차에 따른 조정 가중치를 적용하여 입력 데이터의 표본을 여러 차례 자동으로 다시 추출
  - 최종 지도 학습 모델의 가중 평균 자동 생성
  - 이진, 명목 및 구간 레이블 지원
  - 트리 수, 적용할 분할 기준, 서브트리의 깊이, 컴퓨팅 리소스 등 다양한 옵션으로 트리 트레이닝을 사용자 지정
  - 검증된 데이터 스코어링에 따라 중지 기준을 자동으로 적용하여 과대 적합 방지
  - 운영 환경 스코어링을 위한 SAS 코드 자동 생성
- 신경망:
  - 매개 변수를 지능형 방식으로 자동 조정하여 최적의 모델을 찾아내도록 설정
  - 카운트 데이터의 모델링 지원
  - 대부분 신경망 변수에 사용할 수 있는 지능형 기본값
  - 사용자 지정 가능한 신경망 아키텍처 및 가중치
  - 딥 러닝 지원 목적으로 임의로 여러개의 히든 레이어(Hidden Layer) 사용 가능
  - 입력 및 목표 변수의 자동 표준화
  - 검증 데이터 하위 세트의 자동 선택 및 사용
  - 자동 OOB(Out-Of-Bag) 검증을 통한 조기 중지로 과대 적합 방지
  - 모델 모수에 대한 지능형 자동 조정 지원
  - 운영 환경 스코어링을 위한 SAS 코드 자동 생성
- 서포트 벡터 머신:
  - 이진 목표 레이블 모델링
  - 모델 트레이닝을 위한 선형 및 다항 커널 지원
  - 연속형 및 범주형 입/출력 변수 추가 가능
  - 입력 변수의 자동 조정
  - IP(Interior-Point) 기법 및 AS(Active-Set) 기법 적용 가능
  - 모델 검증 목적의 데이터 파티션 지원
  - 패널티 선택을 위한 교차 검증 지원
  - 운영 환경 스코어링을 위한 SAS 코드 자동 생성
- 베이지안 네트워크:
  - Naive, TAN(Tree-Augmented Naive), BAN(Bayesian Network-Augmented Naive), 상위-하위 베이지안 네트워크, 마르코프(Markov) 블랭킷 등 다양한 베이지안 네트워크 구조 학습
  - 독립성 검정을 통한 효율적인 변수 선택
  - 특정 모수에서 최상의 모델을 자동 선택
  - 데이터 스코어링을 위한 SAS 코드 또는 분석 스토어 생성
  - 다수의 노드에서 데이터를 로드한 후 동시에 계산 실행

데이터 준비, 탐색 및 피처 엔지니어링의 통합

많은 시간이 소요되는 분석 데이터 준비 작업을 해결할 수 있도록 드래그-앤-드롭 인터페이스가 지원되어 데이터 엔지니어들은 통합된 작업 파이프라인에서 빠르게 전처리 과정을 수행하거나, 데이터를 추가 및 결합할 수 있습니다. 모든 작업들은 메모리 내에서 실행되어 일관된 데이터

구조를 유지합니다. 또한 고급 분석 기법으로 데이터 내부의 문제를 찾아 해결합니다. 이 외에도 잠재적 예측 변수를 빠르게 찾아내고 대용량 데이터 세트의 차원을 축소하며 원본 데이터에서 새로운 변수를 손쉽게 생성할 수 있습니다.

분석 데이터 준비

- 시각적으로 분산 데이터 관리
- 대용량 데이터 탐색 및 요약
- 카디널리티(Cardinality) 프로파일링:
  - 입력 데이터 소스의 대용량 데이터 프로파일링
  - 변수 측정 및 역할을 위한 지능형 추천 기능
- 표본 추출: 무작위 및 층화 표본 추출과 희소 이벤트를 위한 과대 표본 추출, 추출된 레코드를 위한 지시 변수 지원
- 이항 레이블을 위한 서포트 벡터 머신(SVM):
  - 선형 및 다항 커널을 사용한 모델 학습
  - Interior-Point 기법 및 Active-Set 기법 적용 가능
  - 모델 평가 목적의 데이터 분할 지원
  - 패널티 선택 목적의 교차 평가 지원
  - 파라미터를 지능적으로 자동 조정하여 최적의 모델을 찾아내도록 설정
- Factorization Machine:
  - 사용자ID 및 아이템 평가 조사 희소행렬 기반의 추천 시스템 개발
  - 운영 환경 스코어링을 위한 SAS 스코어 코드 자동 생성
  - 타임 스탬프, 인구 통계, 컨텍스트 정보가 포함된 모델
  - 처음부터 재학습할 필요없이 새로운 트랜잭션으로 모델을 업데이트할 수 있는 웜 리스타트(Warm Restart) 지원
  - 세 개 이상의 범주형 변수가 포함된 추천을 위한 Tensor Factorization 자동 호출
- 네트워크 분석 및 커뮤니티 감지:
  - 그래프 이론 및 네트워크 분석 알고리즘으로 데이터 마이닝 및 머신러닝 기법 강화
  - 관심있는 객체간 Pairwise-Interaction 적용
  - 가성없는 접근 방식으로 네트워크가 발생할 수 있는 여러가지 방법의 탐지 향상
  - 상호작용 빈도 강도를 토대로 네트워크 연결 가중치 모델링

접근성과 클라우드 환경

Python, R, Java 또는 Lua 등 프로그램에 상관없이 모델 구축자와 데이터 사이언티스트라면 누구나 자신이 원하는 코드 환경에서 SAS 기능에 액세스할 수 있습니다. 또한 SAS Viya REST API를 사용하면 SAS의 기능을 다른 애플리케이션에 추가할 수도 있습니다.

SAS Visual Data Mining and Machine Learning은 현장, Cloud Foundry 같은 기술을 통한 사설 클라우드, 혹은 Amazon Web Services 및 Microsoft Azure 같은 공용 클라우드 등 기업에 가장 적합한 환경에 배포할 수 있습니다. 그 밖에도 사전 배포 및 구성되는 SAS의 관리형 SaaS 솔루션을 통해 이 소프트웨어에 액세스할 수 있습니다.

자동 모델 튜닝

지능적인 자동 하이퍼파라미터 튜닝 기능으로 최적의 모델 구성을 찾아낼 수 있습니다. 통합 자동 튜닝 프로세스는 사용자가 탐색하려고 선택한 하이퍼파라미터와 범위에 따라 SAS Viya 분산 플랫폼에서 다양한 탐색 전략을 동원하여 모델을 순차적으로 학습하고 평가합니다. 18가지 모델 평가 지표를 튜닝 목표로 삼고 검증용 데이터나 자체 내장된 교차 검증 메커니즘으로 모델을 평가할 수 있습니다.

통합 텍스트 분석

빅데이터를 염두에 두고 설계되었으므로 대량의 텍스트 문서를 검토하여 강력한 텍스트 전처리, 자연어 처리, 토픽 추출 등을 통해 알려지지 않은 주제와 연관성에 대한 새로운 인사이트를 확보할 수 있습니다. 또한 통합 텍스트 분석 기능은 데이터 사이언티스트가 비정형 데이터에 내재된 인사이트를 사용하여 지도 학습을 개선하는 데 효과적입니다.

통합 텍스트 분석

- 즉시 사용 가능한 30가지 언어 지원: 영어, 아랍어, 중국어, 크로아티아어, 체코어, 덴마크어, 네덜란드어, 페르시아어, 핀란드어, 프랑스어, 독일어, 그리스어, 히브리어, 힌디어, 인도네시아어, 이탈리아어, 일본어, 한국어, 노르웨이어, 폴란드어, 포르투갈어, 러시아어, 슬로바키아어, 슬로베니아어, 스페인어, 스웨덴어, 타갈로그어, 터키어, 태국어, 베트남어
- 음성에서 용어 부분 자동 식별(시스템에 15개 이상 정의 가능)
- 사전 정의된 옵션에서 위치, 시간, 날짜, 주소 등의 표준 객체 추출
- 명사 그룹/복합 명사 목록을 감지할 경우 머신 러닝 프로세싱에서 단일 용어로 처리
- 자동으로 형태가 다른 단어의 동일 어간을 추출
- 동의어 감지 기능으로 단어 변형 자동 탐색
- 빈도-용어 가중치를 선택하여 문서 내에서 Term-Occurrence Effect를 최소화
- 단어 가중치를 사용하여 문서에서 중요한 단어 구분
- 기본 시작 및 중지 목록을 사용하여 구문 분석 및 다운스트림 처리에 필요한 용어 관리
- 복합어를 포함해 용어를 추가, 삭제 및 편집하여 시작과 중지 목록 수정
- 머신 러닝 주제에 따라 단어-문서 행렬로 구성된 텍스트 처리 함수를 문서 집합의 정형 숫자 형식으로 변환
- 결과적으로 의미와 관련하여 생성된 주제를 머신 러닝 모델의 입력 데이터로 사용 가능
- 텍스트 프로세싱, 구문 분석 등 정확한 결과를 위해 SAS의 운영 환경 문서 스코어링을 위한 코드 자동 생성

모델 평가 및 스코어링

단 한 번의 실행으로 여러 가지 모델링 방식을 테스트하고 표준화된 테스트를 통해 다수의 지도 학습 알고리즘의 분석 결과를 비교하여 빠른 시간 내에 챔피언 모델을 확인할 수 있습니다. 챔피언 모델이 확인되면 자동 생성된 SAS 스코어 코드를 사용하여 분산 환경과 기존 환경에서 분석 작업을 수행할 수 있습니다.

모델 평가

- 지도 학습 모델 성과 통계를 자동으로 계산
- 구간 및 범주형 목표에 따른 출력 통계 산출
- 구간 및 범주형 목표에 따른 리프트 테이블 생성
- 범주형 목표에 따른 ROC 테이블 생성

모델 스코어링

- 모델 스코어링을 위한 SAS DATA 스텝 코드 자동 생성
- 트레이닝, 홀드아웃 데이터 및 새로운 데이터에 스코어링 로직 적용

더 자세한 내용은 sas.com/korea/vdmml에서 확인하실 수 있습니다.