4) 모델 평가

모델 비교

- 하나 이상의 모델에 대한 Lift 차트, ROC 차트, C-통계 및 오분류표 등의 모델 비교 요약 생성.
- 인터랙티브 방식의 컷오프(Cut-off) 설정으로 평가 지표 및 오분류표 결과의 자동 업데이트.
- 인터랙티브 방식으로 설정된 백순위수의 Lift 값 제공.
- 모델 진단(model diagnostics)과 모델 적합성(Model fitting) 평가를 동시에 진행하여 모델 성능의 영향도 확인.

모델 스코어링 (Model scoring)

모델을 Base SAS DATA step 코드 형식으로 추출해서 새로운 데이터에 적용할 수 있습니다.

시스템 사양

서버환경 :

운영 체제 Operating Systems

- Red Hat Enterprise Linux 6
- SuSE Linux Enterprise Server 11
- Oracle Linux 6.1
- Windows (only for nondistributed deployments): Windows Server 2008 R2 Enterprise SP1, Windows Server 2008 R2 Datacenter SP1, Windows Server 2012 Standard, Windows Server 2012 Datacenter

하드웨어

- DB어플라이언스 : Teradata, Oracle 및 Pivotal (previously Greenplum)
- Hadoop 배포판: Cloudera 및 Hortonworks 지원 (기타 별도 문의)
- 기본 : HP and Dell (with preconfigured hardware and software packaging options)
- 기타 IBM, Cisco 등 타 벤더 하드웨어 지원 (별도 문의)

미드-티어

- SAS Web Application Server (included)

클라이언트 환경

브라우저

- Internet Explorer 9 and above (native mode)
- Firefox 6 and up
- Chrome 15 and up

플래쉬 플레이어

- Adobe Flash Player 11.1 이상

필수 소프트웨어

- SAS Visual Analytics 6.4
- 추가 소프트웨어 설치 필요할 수 있음 : 데이터 접속을 위한 다양한 SAS/ACCESS® 엔진 등. (SAS Visual Analytics 사용 시 1개 선택 SAS/ACCESS interface 포함)

For More Information

SAS Visual Statistics에 관한 더 자세한 내용이나 백서 다운로드, 스크린샷, 기타 관련 자료는 sas.com/visual-statistics 페이지를 참조하시기 바랍니다.

한국쌔스소프트웨어(유) 서울 강남구 테헤란로 408 (대치동, 대치빌딩 8~10층) (우 135-839)

www.sas.com/korea

THE POWER TO KNOW.

SAS® Visual Statistics

› fact sheet

인터랙티브한 시각화 기술로 고성능 분석 모델을 생성하고, 실행, 평가하여 빠른 결과 획득

SAS® Visual Statistics는 어떤 솔루션인가?

SAS® Visual Statistics는 인터랙티브하고 직관적인 드래그-앤-드롭(drag-and-drop) 웹 브라우저 인터페이스를 제공하여 아무리 큰 데이터라도 신속하게 탐색 모델(Descriptive Model) 및 예측 모델(Predictive Model) 모델을 생성할 수 있게 해줍니다. 또한 SAS® LASR™ Analytic Server의 분산 형 인-메모리 프로세싱 기술을 통해 분석 모델 개발 시간을 획기적으로 단축하고 복잡한 분석 계산 작업을 수분 내에 처리할 수 있도록 도와줍니다.

SAS® Visual Statistics가 왜 중요한가?

데이터 사이언티스트와 통계 전문가는 세그먼트 별로 최적의 예측 모델(Predictive Model)을 생성하고, 새로운 아이디어를 테스트하고, 즉석에서 해당 모델을 정밀하게 조정할 수 있습니다. 사용자는 아무리 까다로운 문제도 효과적으로 해결할 수 있으며 새로운 기회를 신속히 포착하여 보다 정확한 정보를 기반으로 의사 결정을 내릴 수 있습니다.

SAS® Visual Statistics는 누구를 위한 솔루션인가?

SAS® Visual Statistics는 복잡하고 다양한 데이터를 즉각적으로 시각화하여 분석하고 예측 모델(Predictive Model)을 인터랙티브한 방식으로 생성·평가하여 정확한 인사이트를 신속히 도출할 필요가 있는 통계 전문가와 데이터 사이언티스트, 비즈니스 분석가 등을 위해 설계된 솔루션입니다.

데이터 시각화 기술에 강력한 예측 모델링 기능을 더하다!

SAS® Visual Statistics 특징

- **빅 데이터의 실시간 분석을 통한 보다 정교한 인사이트 확보**
  실시간으로 빅데이터로부터 의미 있는 정보를 얻고, 이를 분석 · 평가하여 새로운 수익 신장의 활로를 찾아낼 수 있습니다. 또한 인터랙티브한 데이터 시각화 기반의 예측 분석 기능으로 비즈니스 분석가와 통계 전문가는 데이터 기반의 더욱 정확한 의사결정을 내릴 수 있습니다.
- **인터랙티브 모델링 환경에서 효율적인 분석 작업**
  웹 브라우저 인터페이스에서 단순한 드래그-앤-드롭(drag-and-drop) 방식으로 탐색 모델(Descriptive Model) 및 예측 모델(Predictive Model) 생성할 수 있게 해줍니다. 멀티 유저 환경에서 변수를 추가/변경하고 이상점(outlier)을 제거하는 등 간단한 방법으로 모델을 변경할 수 있고, 변경사항이 모델 성능에 어떠한 영향을 미치는지 즉시 확인할 수 있습니다.
- **더욱 빠르고 정확한 모델 생성 및 실행**
  SAS의 멀티코어 프로세싱 환경에서는 몇 시간씩 소요되던 모델링 시간을 단 몇 분으로 단축해줍니다. 특정 그룹이나 세그먼트 별 모델을 생성하고 멀티 시나리오를 동시에 실행할 수 있으며, 분석 전문가는 더 많은 What-If 질문을 던지고 신속하게 답을 얻을 수 있습니다. 그리고, 모델을 빠르게 변경하여 적용함으로써 양질의 작업 결과물을 보장합니다.
- **혁신적인 인-메모리 컴퓨팅 기술**
  인-메모리 엔진을 탑재한 SAS® Visual Statistics는 복잡한 계산 및 분석에서 진가를 발휘합니다. 모델러는 지금까지 분석하기 어려웠던 빅데이터를 활용해 새로운 아이디어를 빠르게 테스트하고, 다양한 모델링 기법을 적용하여 모델의 성능을 향상시킬 수 있습니다.

THE POWER TO KNOW.

SAS® Visual Statistics 주요 기능

1) 데이터 준비

다이내믹 Group-By 프로세싱

매번 데이터를 정렬하거나 인덱싱할 필요 없이 각 그룹 또는 세그먼트에 대해 동시에 복수의 모델을 생성하고 결과를 처리할 수 있습니다.

- 그룹 변수 별 개별 모델링이 아닌 그룹 변수 별 동시 모델링 가능, 데이터 재배열 없이 그룹 프로세싱하여 각 그룹에 대해 즉각적으로 더 많은 결과물 산출 가능.
- 의사결정트리 또는 클러스터링 분석 결과에서 생성된 세그먼트 별 모델링 가능.

인-메모리 분석 프로세싱

보다 신속하게 모델을 생성할 수 있게 해주는 기술로, 사용자는 디스크에 데이터를 쓰거나 데이터 셔플링(data shuffling)을 진행할 필요가 없습니다.

- 전수 데이터를 메모리로 로드하면 이후부터는 새로운 작업을 수행할 때마다 다시 로드할 필요 없이 데이터 작업이 가능하며, 변경 사항(새 변수 추가, 이상점 제거 등)이 모델에 미치는 영향 즉시 확인.
- 동시 프로세싱이 가능하도록 설계되어 복수의 사용자가 복잡한 모델을 동시에 생성 및 실행 가능

이 외에도 데이터 및 분석 작업을 복수의 서버 노드에 걸쳐 분산 처리하고 각 노드에서 다중 작업이 가능하므로 사용자는 놀라운 속도 향상을 경험할 수 있습니다.

2) 데이터 탐색

인터랙티브 데이터 시각화 및 탐색 부가 기능

SAS® Visual Statistics은 SAS® Visual Analytics에 Add-on 할 수 있으며, 사용이 매우 간편한 데이터 조작 및 시각적 데이터 탐색 기능을 제공합니다.

- 수많은 변수들 중에서 모델 적합 결과에 영향을 미치는 주요 변수를 쉽고 빠르게 파악 가능.
- 추가 분석 작업을 위해 이상점(outlier) 혹은 영향력 포인트 파악 및 측정·제거.
- 막대 그래프, 히스토그램, 상자 그림(box plots), 히트 맵, 버블 차트, 네트워크 다이어그램 등 데이터 탐색(SAS® Visual Analytics 이용).
- 상관 행렬(correlation matrices), 산점도(scatter plots), 상자 그림(box plots) 에서 변수를 탐색하고 선택된 특정 변수를 이용하여 모델링 가능.

로지스틱 회귀(Logistic regression) 를 이용해 이진 결과를 예측합니다.

경쟁 모델 간의 비교를 통해 어떤 모델이 최상의 결과를 낼지 결정합니다.

3) 모델링

탐색 모델링 (Descriptive Modeling)

클러스터링은 이질적 모집단을 보다 동질적인 여러 개의 소집단, 즉 클러스터로 세분화하는 작업으로, 다른 유형의 데이터 마이닝을 위한 예비 작업으로 활용되는 경우가 많습니다. 일례로, 우리는 시장 세분화를 통해서 유사한 구매 습성을 가진 고객들을 클러스터링하여 어떤 판촉 전략이 가장 효과적인지 알아낼 수 있습니다.

- k 평균 군집화(k-means clustering).
- 평행 좌표 그래프를 통해 클러스터 멤버십 평가(인터랙티브 방식).
- 소규모 데이터 세트의 클러스터 프로파일이 적용된 산점도 및 대규모 데이터 세트의 클러스터 프로파일이 적용된 히트 맵.
- 추가 분석을 위한 클러스터 세그먼트 변수 생성.
- 세부 요약 통계(각 클러스터의 평균 및 관측수 등).

예측 모델링 (Predictive Modeling)

예측 모델은 미래 행동을 분류하거나 미래 가치를 추정하는데 활용되며 SAS® Visual Statistics는 선형 회귀(linear regression), 일반화 선형 모델(generalized linear model), 로지스틱 회귀(logistic regression), 의사결정 트리(Decision Tree) 등의 기법을 이용해 손쉽게 예측 모델을 생성할 수 있도록 도와줍니다. 사용자는 분류 기능을 이용해서 사기여부나 고객 이탈 여부와 같은 이진(Binary) 분류나 고객 등급 상향, 하향, 유지 등과 같은 멀티 레벨의 분류를 예측할 수 있습니다.

의사결정 트리 (Decision Tree)

- C4.5 알고리즘 기반(information gain 또는 information ratio).
- 인터랙티브 방식으로 트리 가지치기의 조건 설정.
- 트리 깊이, 최대 가지수, 리프 크기, 트리 가지치기의 조건 설정 기능.
- 사용자 지정 그룹핑(Bin 수 지정).
- Tree Map 및 Tree Overview Display - 트리 구조를 인터랙티브 방식으로 탐색.

일반화 선형 모델 (Generalized linear model)

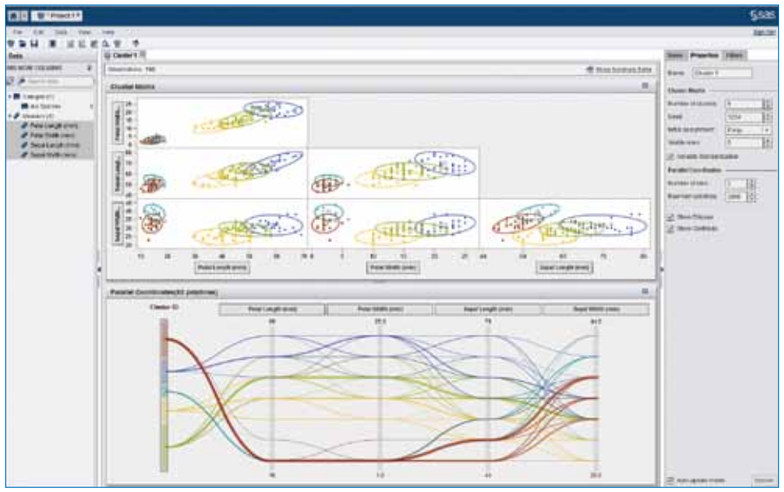
- 지원 배포판 : 베타, 정규, 이항, 지수, 감마, 기하, 포아송, 역가우스, 음이항 분포 지원 등.
- 수렴 및 반복 기준 설정.
- 보정(offset) 변수 지원.
- FREQ/WEIGHT 변수.
- 잔차 진단(Residual diagnostics).
- 모델 요약, 반복 계산 과정, 적합통계량, Type III 검증테이블, 모수 추정 결과.
- 예측 변수의 결측값 처리에 유용한 결측 옵션.

로지스틱 회귀 (Logistic regression)

- Logit 및 Probit 링크 함수를 적용한 이진(Binary) 분류 모델.
- 영향력 통계량(Influence statistics).
- 변수 선택(Variable selection).
- 보정(Offset) 변수 지원.
- FREQ/WEIGHT 변수.
- 잔차 진단(Residual diagnostics).
- 모델 차원(Model dimensions), 반복 계산 과정, 적합통계량, Type III 검증테이블, 모수 추정 결과.
- 예측 변수의 결측값 처리 옵션.

선형 회귀 (Linear regression)

- 영향력 통계(Influence statistics).
- 변수 선택(Variable selection).
- FREQ/WEIGHT 변수.
- 잔차 진단(Residual diagnostics).
- Overall ANOVA, 모델 차원(Model dimension), 적합통계량, Model ANOVA, Type III 테스트, 모수 추정 등의 요약 테이블.
- 예측 변수의 결측값을 처리하는 데 유용한 결측 옵션.



k 평균 군집화(k-means clustering)을 이용해 데이터를 세분화합니다.