IBM SPSS Data Preparation 26

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 7 페이지의 『주의사항』에 있는 정보를 확인하십시오.

제품 정보

이 개정판은 새 개정판에서 별도로 명시하지 않는 한, IBM® SPSS Statistics의 버전 26, 릴리스 0. 수정사항 0 및 모든 후속 릴리스와 수정에 적용됩니다.

목차

데이터 준비

다음 데이터 준비 기능은 SPSS® Statistics Professional Edition 또는 데이터 준비 옵션에 포함되어 있습니다.

데이터 준비 소개

컴퓨팅 시스템이 강력해짐에 따라 정보에 대한 요구도 비례하여 늘어났으며 이로 인해 점점 더 많은 데이터 콜렉션(더 많은 케이스, 더 많은 변수, 더 많은 데이터 입력 오류)이 생성되었습니다. 이러한 오류는 데이터 웨어하우징의 최종 목적인 예측 모형 예측값의 문제가 되므로 데이터를 "깨끗한 상태"로 유지해야 합니다. 그러나 데이터 검증에 대한 자동 프로세스를 구현하기 위해 필수적인 케이스 수동 확인 능력을 벗어날 정도로 데이터 웨어하우징 양이 증가되었습니다.

데이터 준비 추가 기능 모듈을 사용하면 활성 데이터 세트에 있는 특이 케이스, 잘못된 케이스, 변수 및 데이터 값을 식별하고 모형화를 위해 데이터를 준비할 수 있습니다.

데이터 준비 프로시저 사용

데이터 준비 프로시저 사용은 특정 요구에 따라 다릅니다. 데이터를 로드한 후의 일반적인 경로는 다음과 같습니다.

메타데이터 준비

사용자 데이터 파일에서 변수를 검토하고 유효한 값, 레이블 및 측정 수준을 판별합니다. 불가능하지만 공통적으로 잘못 코딩되는 변수값 조합을 식별합니다. 이 정보를 기반으로 하여 검증 규칙을 정의합니다. 이는 시간이 걸리는 작업일 수 있지만 보통의 기반에서 유사한 속성이 있는 데이터 파일을 검증해야 하는 경우 노력할 가치가 있는 작업입니다.

데이터 유효성 검증

기본 확인 및 정의된 검증 규칙에 대한 확인을 실행하여 유효하지 않은 케이스, 변수 및 데이터 값을 식별합니다. 유효하지 않은 데이터가 있는 경우 원인을 조사하여 정정하십시오. 이를 수행하는 데 메타데이터 준비를 통한 다른 단계가 필요할 수 있습니다.

모형 준비

자동 데이터 준비를 사용하여 모형설정을 향상시키는 원래 필드의 변환을 확보합니다. 많은 예측 모형에 대한 문제를 일으킬 수 있는 잠재적인 통계 이상값을 식별합니다. 일부 이상값은 식별되지 않은 유효하지 않은 변수 값으로 인한 결과입니다. 이를 수행하는 데 메타데이터 준비를 통한 다른 단계가 필요할 수 있습니다.

데이터 파일이 "깨끗"하면 다른 추가 기능 모듈에서 모형을 작성할 준비가 됩니다.

특이 케이스 식별

비정상 탐지 프로시저는 해당 군집 집단 기준에서의 편차를 기반으로 하는 특이 케이스를 검색합니다. 이 프로시저는 데이터 추정 분석 이전의 탐사 데이터 분석 단계에서 데이터 감사를 위해 특이 케이스를 신속하게 발견하도록 설계되었습니다. 이 알고리즘은 일반적인 비정상 탐지를 위해 설계되었으며, 이는 비정상 케이스 정의(예: 헬스케어 산업에서의 비정상적인 지불 패턴 탐지 또는 재무 산업에서의 돈세탁 탐지)가 비정상 정의가 잘 정의될 수 있는 특정 애플리케이션에 한정되지 않음을 의미합니다.

예 행정 처리 결과에 대한 예측 모형을 작성하기 위해 고용된 데이터 분석가는 해당 모형이 특이한 관측값에 민감할 수 있으므로 데이터 품질에 관심을 기울입니다. 이와 같이 범위를 벗어난 관측값 중 실제로 특별한 케이스를 나타내는 일부 값은 예측 모형을 작성하는 데 적합하지 않은 것으로 거를 수 있지만, 데이터 입력 오류로 인해 발생한 다른 관측값은 기술적으로 볼 때는 "올바른" 값이므로 데이터 검증 프로시저를 통해 발견할 수 없습니다. 특이 케이스 식별 프로시저는 분석가가 이를 처리하는 방법을 결정할 수 있도록 이러한 이상값을 찾아서 보고합니다.

통계 이 프로시저는 동등 집단, 연속형 및 범주형 변수에 대한 동등 집단 기준, 동등 집단 기준의 편차를 기반으로 하는 비정상 지수, 특이한 케이스에 가장 기여하는 변수의 변수 영향 값을 생성합니다.

데이터 고려사항

데이터. 이 프로시저는 연속형 변수와 범주형 변수에 모두 적용됩니다. 각 행은 고유한 관측값을 나타내고, 각 열은 동등 집단이 기반으로 하는 고유한 변수를 나타냅니다. 결과 표시를 위해 데이터 파일에서 케이스 식별 변수를 사용할 수 있지만 분석에서 사용할 수는 없습니다. 결측값은 허용됩니다. 지정한 가중변수는 무시됩니다.

탐지 모형은 새 검정 데이터 파일에 적용될 수 있습니다. 검정 데이터 요소는 학습 데이터 요소와 같아야 합니다. 또한 알고리즘 설정에 따라 모형설정에 사용된 결측값 처리는 스코어링 이전에 검정 데이터 파일에 적용될 수 있습니다.

케이스 순서. 솔루션은 케이스 순서에 따라 다를 수 있습니다. 순서가 미치는 영향을 최소화하려면 케이스 순서를 무작위로 설정해야 합니다. 제공된 솔루션의 안정성을 확인하기 위해 다른 무작위 순서로 정렬된 케이스가 있는 여러 다른 솔루션을 확보하려고 할 수 있습니다. 파일이 너무 큰 솔루션에서는 다른 무작위 순서로 정렬된 샘플 케이스를 사용하여 실행을 여러 번 수행할 수 있습니다.

가정. 이 알고리즘은 모든 변수가 상수가 아니고 독립적이며, 케이스에 입력 변수에 대한 결측값이 없다고 가정합니다. 연속형 변수마다 개별 정규(가우시안)분포를 가지며 범주형 변수마다 다항분포 특성을 가진다고 가정합니다. 실제 내부 검정을 통해 이 프로시저가 독립성과 분산에 대한 가정에 그리 큰 영향을 받지 않는다는 결론을 얻었지만 이러한 가정을 충족하는 것이 좋습니다.

특이 케이스 식별

1. 메뉴에서 다음을 선택합니다.

   데이터 > **특이 케이스 식별...**

2. 하나 이상의 분석 변수를 선택합니다.
3. 선택적으로 결과 레이블 지정에 사용할 케이스 식별자 변수를 선택합니다.
4. **적용**을 클릭합니다.

측정 수준을 알 수 없는 필드

측정 수준 경보는 데이터 세트에서 하나 이상의 변수(필드)에 대한 측정 수준을 알 수 없는 경우 표시됩니다. 측정 수준은 이 프로시저의 계산 결과에 영향을 미치기 때문에 모든 변수에 정의된 측정 수준이 있어야 합니다.

데이터 스캔

활성 데이터 세트의 데이터를 읽고 현재 알 수 없는 측정 수준이 있는 필드에 기본 측정 수준을 할당합니다. 데이터 세트가 큰 경우 시간이 걸릴 수 있습니다.

수동으로 할당

알 수 없는 측정 수준이 있는 필드를 모두 나열합니다. 해당 필드에 측정 수준을 할당할 수 있습니다. 데이터 편집기의 변수 목록 분할창에서도 측정 수준을 할당할 수 있습니다.

측정 수준이 이 프로시저에서 중요한 요소이므로 모든 필드에 대해 측정 수준이 정의될 때까지 이 프로시저를 실행할 수 없습니다.

특이 케이스 식별: 결과

결과 대화 상자에서는 표로 작성된 결과를 생성하는 옵션을 제공합니다.

특이 케이스 및 특이로 처리되는 원인 목록

선택하면 이 옵션으로 세 개의 표가 생성됩니다.

- 비정상 케이스 지수 목록은 특이로 식별되는 케이스를 표시하고 해당 비정상 지수 값을 표시합니다.
- 비정상 케이스 동등 ID 목록은 특이 케이스 및 해당 동등 집단에 관련된 정보를 표시합니다.
- 비정상 원인 목록은 케이스 번호, 원인변수, 변수 영향 값, 변수 값, 각 원인의 변수 노름(norm)을 표시합니다.

모든 표는 내림차순의 비정상 지수별로 정렬됩니다. 또한 변수 대화 상자에 케이스 식별자 변수가 지정된 경우 케이스의 ID가 표시됩니다.

요약값

이 그룹의 제어는 분포 요약을 생성합니다.

동등 집단 기준

이 옵션은 연속형 변수 노름 표(분석에 연속형 변수가 사용된 경우)와 범주형 변수 노름 표(분석에 범주형 변수가 사용된 경우)를 표시합니다. 연속형 변수 노름 표는 각 동등 집단의 각 연속형 변수에 대한 평균과 표준편차를 표시합니다. 범주형 변수 노름 표는 각 동등 집단의 각 범주형 변수에 대한 최빈값(가장 널리 사용되는 범주), 빈도 및 빈도 퍼센트를 표시합니다. 연속형 변수의 평균 및 범주형 변수의 최빈값은 분석에서 노름 값으로 사용됩니다.

비정상 지수

비정상 지수 요약은 가장 특이한 항목으로 식별되는 케이스의 비정상 지수에 대한 기술통계량을 표시합니다.

분석 변수에 의한 원인 발생

이 표는 각 원인에 대해 각 변수의 원인 발생 빈도 및 빈도 퍼센트를 표시합니다. 또한 이 표는 각 변수의 영향에 대한 기술통계량을 보고합니다. 옵션 탭에서 최대 원인 수가 0으로 설정된 경우 이 옵션을 사용할 수 없습니다.

처리된 케이스

케이스 처리 요약은 활성 데이터 세트에 있는 모든 케이스, 분석에서 포함되거나 제외된 케이스, 각 동등 집단의 케이스에 대한 수 및 수 퍼센트를 표시합니다.

특이 케이스 식별: 저장

저장 대화 상자에서는 변수 및 모형 저장 옵션을 제공합니다.

변수 저장

이 그룹의 제어를 사용하여 모형 변수를 활성 데이터 세트에 저장할 수 있습니다. 또한 저장할 변수와 이름이 충돌하는 기존 변수를 대체하도록 선택할 수도 있습니다.

비정상 지수

각 케이스에 대한 비정상 지수 값을 지정된 이름의 변수에 저장합니다.

동등 집단

각 케이스의 동등 집단 ID, 케이스 빈도 및 크기(퍼센트)를 지정된 루트 이름의 변수에 저장합니다. 예를 들어 루트 이름이 *Peer*로 지정되면 *Peerid*, *PeerSize* 및 *PeerPctSize* 변수가 생성됩니다. *Peerid*는 동등 집단 ID 케이스이고, *PeerSize*는 집단의 크기이며, *PeerPctSize*는 집단의 크기 퍼센트입니다.

원인 지정된 루트 이름을 사용하여 원인 변수 세트를 저장합니다. 원인 변수 세트는 원인 변수 이름, 변수 영향 측도와 해당 값, 노름(norm) 값으로 구성됩니다. 세트 수는 옵션 탭에서 요청된 원인 수에 따라 다릅니다. 예를 들어, 루트 이름이 *Reason*으로 지정되면 *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* 및 *ReasonNorm_k* 변수가 생성됩니다. 여기서 *k*는 *k*번째 원인입니다. 원인 수가 0으로 설정되면 이 옵션을 사용할 수 없습니다.

동일한 이름 또는 루트 이름을 가진 기존변수 바꾸기
선택한 경우 저장할 변수와 이름이 충돌하는 기존 변수를 바꿀 수 있습니다.

모형 파일 내보내기
모형을 외부 XML 파일로 저장할 수 있습니다.

특이 케이스 식별: 결측값

결측값 대화 상자에서 사용자 결측값과 시스템 결측값 처리를 제어할 수 있습니다.

분석에서 결측값 제외
결측값이 있는 케이스를 분석에서 제외합니다.

분석에 결측값 포함
연속형 변수의 결측값은 해당 총평균을 대체하고 범주형 변수의 결측 범주는 유효한 범주로 집단화되어 처리됩니다. 처리된 변수는 분석에 사용됩니다. 선택적으로 각 케이스에서 결측 변수의 비율을 나타내는 추가 변수를 작성하도록 요청하여 분석에서 해당 변수를 사용할 수 있습니다.

특이 케이스 식별: 옵션

결과 대화 상자에는 특이 케이스 기준 설정 및 동등 집단 수의 범위 정의가 포함됩니다.

특이 케이스 식별 기준
다음 설정은 이상 항목 목록에 포함된 케이스 수를 결정합니다.

비정상 지수 값이 가장 높은 케이스 퍼센트
100보다 작거나 같은 양의 정수를 지정하십시오.

비정상 지수 값이 가장 높은 케이스의 고정 수
분석에 사용되는 활성 데이터 세트에 있는 총 케이스 수보다 작거나 같은 양의 정수를 지정하십시오.

비정상 지수 값이 최소값을 충족하거나 초과하는 케이스만 식별
음수가 아닌 수를 지정하십시오. 비정상 지수 값이 지정된 분리점보다 크거나 같은 경우 이 케이스는 이상 항목으로 간주됩니다. 이 옵션은 **케이스 퍼센트** 및 **케이스의 고정 숫자** 옵션과 함께 사용됩니다. 예를 들어 케이스의 고정 숫자를 50으로 지정하고 분리점 값을 2로 지정한 경우 비정상 목록은 비정상 지수 값이 각각 2보다 크거나 같은 최대 50개의 케이스로 구성됩니다.

동등 집단 수
이 프로시저는 지정된 최소값과 최대값 사이의 동등 집단의 최적 수를 검색합니다. 이 값은 양의 정수여야 하고 최소값은 최대값을 넘지 않아야 합니다. 지정된 값이 같은 경우 프로시저는 동등 집단의 고정 숫자를 가정합니다.

참고: 데이터의 변동 크기에 따라 데이터에서 지원할 수 있는 동등 집단 수가 최소값으로 지정된 수보다 작은 상황이 발생할 수 있습니다. 이러한 경우 프로시저에서는 더 작은 동등 집단 수를 생성할 수 있습니다.

최대 원인 수

원인은 변수 영향 측도, 이 원인에 대한 변수 이름, 변수 값 및 해당 동등 집단 값으로 구성됩니다. 음수가 아닌 정수를 지정하십시오. 이 값이 분석에 사용된 처리된 변수의 수와 같거나 큰 경우 모든 변수가 표시됩니다.

DETECTANOMALY 명령 추가 기능

명령 구문을 사용하여 수행할 수 있는 추가 기능은 다음과 같습니다.

- 모든 분석 변수를 명시적으로 지정하지 않고 분석에서 활성 데이터 세트에 있는 몇 가지 변수를 생략합니다(EXCEPT 부명령문 사용).
- 연속형과 범주형 변수 영향의 균형을 맞추도록 조정합니다(CRITERIA 부명령문에서 MLWEIGHT 키워드 사용).

명령 구문에 대한 자세한 내용은 *Command Syntax Reference*를 참조하십시오.

주의사항

이 정보는 미국에서 제공되는 제품 및 서비스용으로 작성된 것입니다. 본 자료는 다른 언어로도 제공될 수 있습니다. 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.

IBM이 제시하는 방향 또는 의도에 관한 모든 언급은 특별한 통지 없이 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.

저작권 라이센스:

이 정보에는 여러 운영 플랫폼에서의 프로그래밍 기법을 보여주는 원어로 된 샘플 응용프로그램이 들어 있습니다. 귀하는 이러한 샘플 프로그램의 작성 기준이 된 운영 플랫폼의 애플리케이션 프로그래밍 인터페이스(API)에 부합하는 애플리케이션을 개발, 사용, 판매 또는 배포할 목적으로 IBM에 추가 비용을 지불하지 않고 이들 샘플 프로그램을 어떠한 형태로든 복사, 수정 및 배포할 수 있습니다. 이러한

샘플 프로그램은 모든 조건하에서 완전히 테스트된 것은 아닙니다. 따라서 IBM은 이러한 프로그램의 신뢰성, 서비스 가능성 또는 기능을 보증하거나 진술하지 않습니다. 본 샘플 프로그램은 일체의 보증 없이 "현상태대로" 제공됩니다. IBM은 귀하의 샘플 프로그램 사용과 관련되는 손해에 대해 책임을 지지 않습니다.

이러한 샘플 프로그램 또는 파생 제품의 각 사본이나 그 일부에는 반드시 다음과 같은 저작권 표시가 포함되어야 합니다.

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. 현재 IBM 상표 목록은 웹 "저작권 및 상표 정보"(www.ibm.com/legal/copytrade.shtml)에 있습니다.

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.

색인

[가]

[다]

[바]

[아]

[타]