IBM SPSS Modeler CRISP-DM 안내서

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 37 페이지의 『 주의사항 』에 있는 정보를 확인하십시오.

제품 정보

이 개정판은 새 개정판에 별도로 명시하지 않는 한, IBM® SPSS® Modeler 버전 18, 릴리스 2, 수정 2 및 모든 후속 릴리스와 수정에 적용됩니다.

목차

서론

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

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

IBM Business Analytics 소개

IBM Business Analytics 소프트웨어는 의사 결정자가 비즈니스 성과를 향상시키기 위해 신뢰하는 완벽하고 일관되며 정확한 정보를 제공합니다. 비즈니스 지능, 예측 분석, 금융 성과와 전략 관리 및 분석 응용 프로그램의 종합 포트폴리오는 현재 성과와 앞으로의 결과를 예측하는 능력에 분명하고 즉각적이면서 실행 가능한 통찰력을 제공합니다. 다양한 업계 솔루션, 입증된 사례 및 전문 서비스와 결합되어 어떠한 크기의 조직이라도 생산성을 극대화하고 자신있게 의사 결정을 자동화하고 더 나은 결과를 제공할 수 있습니다.

이 포트폴리오의 일부인 IBM SPSS Predictive Analytics 소프트웨어는 조직이 향후 상황을 예측하고 그 통찰을 바탕으로 적극적인 사전 조치를 취해 더 우수한 비즈니스 성과를 거둘 수 있도록 지원합니다. 전 세계의 기업, 정부 및 학계 고객들은 고객을 매료시키고 유지하며 성장하게 만드는 동시에 불공정 행위를 줄이고 위험을 낮추는 IBM SPSS 기술의 경쟁 이점을 활용합니다. 일상 업무에서 IBM SPSS 소프트웨어를 활용한다면 예측형 기업으로 거듭날 수 있습니다. 즉 비즈니스 목표 달성을 위해 의사 결정의 방향을 정하고 이를 자동화하며 측정 가능한 경쟁 우위를 달성할 수 있습니다. 자세한 내용을 보거나 담당자에게 문의하려면 http://www.ibm.com/spss 사이트를 방문하십시오.

기술 지원

유지보수 고객은 기술 지원을 받을 수 있습니다. IBM 제품 사용 및 지원되는 하드웨어 환경 중 하나에 대해 설치하는 데 도움이 필요한 경우 기술 지원에 문의하십시오. 기술 지원에 문의하려면 IBM 웹 사이트 (http://www.ibm.com/support)를 참조하십시오. 지원을 요청하려면 본인의 신상과 소속 조직(회사) 및 지원 동의서를 제시해야 합니다.

제 1 장 CRISP-DM 소개

CRISP-DM 도움말 개요

CRISP-DM(Cross-Industry Standard Process for Data Mining의 약자)은 데이터 마이닝 작업을 안내하기 위해 업계에서 검증된 방법입니다.

- **방법론**으로서, 이 방법은 프로젝트의 일반적 단계에 대한 설명, 각 단계와 관련된 작업 및 이러한 작업 사이의 관계 설명을 포함합니다.
- **프로세스 모델**로서, CRISP-DM은 데이터 마이닝 라이프사이클의 개요를 제공합니다.

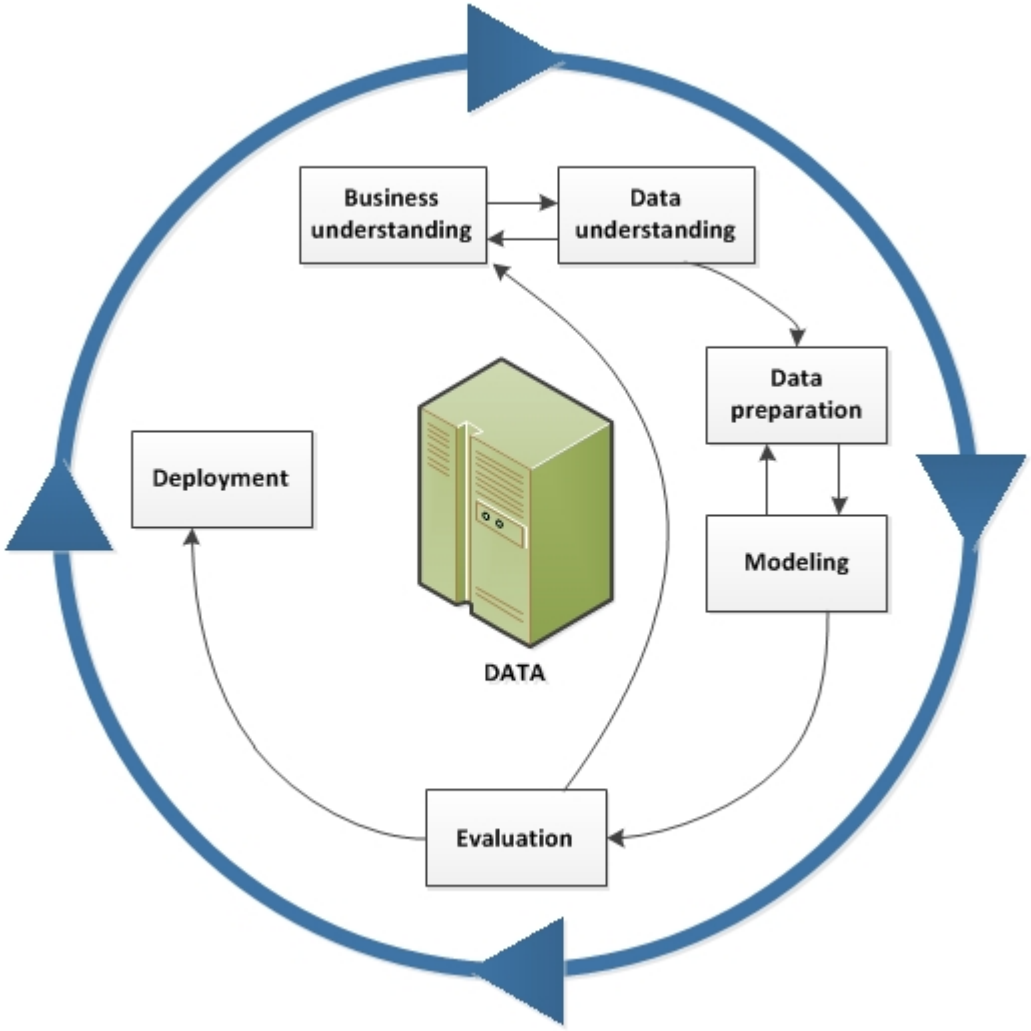

그림 *1.* 데이터 마이닝 라이프사이클

라이프사이클 모델은 6개의 단계로 구성되며 단계 사이에는 가장 중요하고 빈번한 종속 항목을 표시하는 화살표가 있습니다. 단계의 순서는 엄격하지 않습니다. 결국, 대부분의 프로젝트는 필요에 따라 단계 사이를 앞뒤로 이동합니다.

CRISP-DM 모델은 유연하므로 쉽게 사용자 정의할 수 있습니다. 예를 들어, 조직에서 자금 세탁을 감지하는 것이 목표라면 특정 모델링 목적 없이 대량의 데이터를 조사하게 될 것입니다. 모델링 대신에, 사용자의 작업은 재무 데이터에서 의심스런 패턴을 밝히기 위해 데이터 탐색 및 시각화에 초점을 맞출 것입니다. CRISP-DM을 사용하면 특정 요구사항에 맞는 데이터 마이닝 모델을 작성할 수 있습니다.

이러한 상황에서 모델링, 평가 및 배포 단계는 데이터 이해 및 준비 단계보다 관련성이 부족할 수 있습니다. 그러나 장기 계획 및 이후의 데이터 마이닝 목적을 위해 이러한 이후 단계 동안 제기된 일부 질문을 고려하는 것은 여전히 중요합니다.

IBM SPSS Modeler의 CRISP-DM

IBM SPSS Modeler는 두 가지 방법으로 CRISP-DM 방법론을 통합하여 효과적 데이터 마이닝을 위한 특색 있는 지원을 제공합니다.

- CRISP-DM 프로젝트 도구는 일반적 데이터 마이닝 프로젝트의 단계에 따라 프로젝트 스트림, 출력 및 주석을 구성하도록 돕습니다. 스트림 및 CRISP-DM 단계에 대한 설명을 기반으로 프로젝트 동안 언제든지 보고서를 생성할 수 있습니다.
- CRISP-DM의 도움말은 데이터 마이닝 프로젝트를 수행하는 프로세스를 안내합니다. 도움말 시스템에는 각 단계의 작업 목록뿐만 아니라 CRISP-DM이 실제 세계에서 어떻게 동작하는지에 예제도 포함되어 있습니다. 기본 창 도움말 메뉴에서 **CRISP-DM 도움말**을 선택하여 CRISP-DM 도움말에 액세스할 수 있습니다.

CRISP-DM 프로젝트 도구

CRISP-DM 프로젝트 도구는 프로젝트의 성공을 도울 수 있는 데이터 마이닝의 조직화된 접근법을 제공합니다. 이것은 본질적으로 표준 IBM SPSS Modeler 프로젝트 도구의 확장입니다. 결국, CRISP-DM 보기와 표준 클래스 보기 간에 전환하여 CRISP-DM의 유형 또는 단계별로 구성된 스트림 및 출력을 볼 수 있습니다.

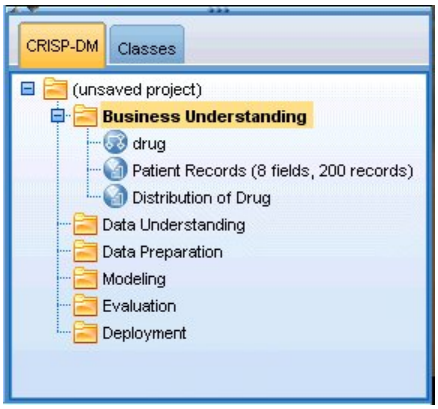

그림 *2. CRISP-DM* 프로젝트 도구

프로젝트 도구의 CRISP-DM 보기를 사용하여 다음을 수행할 수 있습니다.

- 데이터 마이닝 단계에 따라 프로젝트의 스트림 및 출력을 구성합니다.
- 각 단계에 대해 조직의 목적에 관한 설명을 작성합니다.
- 각 단계에 대한 사용자 정의 도구 팁을 작성합니다.
- 특별한 그래프 또는 모델로부터 도출된 결론에 관한 설명을 작성합니다.
- 프로젝트 팀에 배포하기 위한 HTML 보고서 또는 업데이트를 생성합니다.

CRISP-DM에 대한 도움말

IBM SPSS Modeler는 일반 CRISP-DM 프로세스 모델에 대한 온라인 안내서를 제공합니다. 이 안내서는 프로젝트 단계에 따라 구성되며 다음과 같은 지원을 제공합니다.

- CRISP-DM의 각 단계에 대한 개요 및 태스크 목록
- 다양한 이정표에 대한 보고서 생성을 위한 도움말
- 프로젝트 팀이 CRISP-DM을 사용하여 데이터 마이닝을 어떻게 활용할 수 있을지 보여주는 실제 예제
- CRISP-DM에 대한 추가 자원의 링크

기본 창 도움말 메뉴에서 **CRISP-DM 도움말**을 선택하여 CRISP-DM 도움말에 액세스할 수 있습니다.

추가 자원

CRISP-DM에 대한 IBM SPSS Modeler 지원뿐만 아니라, 데이터 마이닝 프로세스의 이해를 넓히기 위한 몇 가지 방법이 있습니다.

- CRISP-DM 컨소시엄에서 작성하고 이 릴리스에서 제공한 CRISP-DM 매뉴얼을 읽으십시오.
- SPSS Inc.에서 제공하는 *Data Mining with Confidence*(ISBN 1-56827-287-1, copyright 2002) 문서를 읽으십시오.

제 2 장 비즈니스 이해

비즈니스 이해 개요

IBM SPSS Modeler에서 작업하기 전에도 조직이 데이터 마이닝에서 무엇을 얻고자 하는지에 대해 탐색할 시간을 가져야 합니다. 이러한 토론에 가능한 한 많은 핵심 인사를 참여시키고 결과를 문서화합니다. 이 CRISP-DM 단계의 최종 단계는 여기서 수집된 정보를 사용하여 프로젝트 계획 생성 방법을 토론하는 것입니다.

이 연구가 불필요해 보일 수도 있지만 그렇지 않습니다. 데이터 마이닝 작업에 대한 사업적 이유를 파악함으로써 귀중한 자원을 소비하기 전에 모든 사람이 동일한 배경 정보를 가질 수 있게 됩니다.

비즈니스 목표 결정

첫 번째 태스크는 데이터 마이닝을 위한 비즈니스 목적에 맞게 가능한 한 많은 통찰력을 얻는 것입니다. 이는 보기만큼 쉽지 않을 수도 있지만 문제점, 목적 및 자원을 명료화하여 이후의 위험을 최소화할 수 있습니다.

CRISP-DM 방법론은 이를 달성하기 위해 조직화된 방식을 제공합니다.

태스크 목록

- 현재 비즈니스 상황에 대한 배경 정보 수집을 시작합니다.
- 핵심 의사결정자에 의해 결정된 특정 비즈니스 목표를 문서화합니다.
- 특정 비즈니스 관점에서 데이터 마이닝 성공을 판단하는 데 사용되는 기준을 절충합니다.

E-소매 예제--비즈니스 목표 찾기

CRISP-DM을 사용하는 웹 마이닝 시나리오

많은 회사에서 인터넷 쇼핑으로의 전환을 시도함에 따라 컴퓨터/전자제품의 기존 e-소매업자는 새로운 사이트에 의한 경쟁 심화에 직면하고 있습니다. 고객이 인터넷으로 이전하는 것만큼 빨리(또는 그보다 빨리) 인터넷 매장이 생기는 현실에 직면해서, 기업은 고객 획득의 비용 상승에도 불구하고 수익성을 유지하는 방법을 모색해야 합니다. 제안되는 솔루션 중 하나는 각 회사가 보유한 현재 고객의 가치를 극대화하기 위해 기존 고객 관계를 돈독하게 하는 것입니다.

따라서 다음과 같은 연구 목표가 수반됩니다.

- 더 나은 추천을 통해 교차 판매를 개선합니다.
- 보다 개인화된 서비스를 통해 고객 충성도를 높입니다.

시험적으로, 연구는 다음과 같은 경우 성공으로 판단합니다.

- 교차 판매가 10% 증가합니다.
- 고객이 사이트 방문 시 더 많은 시간을 보내고 더 많은 페이지를 봅니다.
- 연구를 제 시간에 예산 내에서 마칩니다.

비즈니스 배경 컴파일

조직의 비즈니스 상황을 이해함으로써 다음의 관점에서 현재 처리 중인 사항을 쉽게 파악할 수 있습니다.

- 사용 가능한 자원(인원 및 자재)
- 문제점
- 목적

데이터 마이닝 프로젝트의 결과에 영향을 미칠 수 있는 질문에 대한 실제적 해답을 찾기 위해 현재 비즈니스 상황에 대해 약간의 연구가 필요할 것입니다.

태스크 1--조직 구조 결정

- 기업 부문, 부서 및 프로젝트 그룹을 보여주는 조직 차트를 개발합니다. 관리자의 이름 및 책무를 포함해야 합니다.
- 조직 내 핵심 인사들을 식별합니다.
- 재정 지원 및/또는 도메인 전문지식을 제공할 내부 스폰서를 식별합니다.
- 운영 위원회가 있는지 판별하고, 멤버 목록을 획득합니다.
- 데이터 마이닝 프로젝트에 의해 영향을 받을 비즈니스 단위를 식별합니다.

태스크 2--문제점 영역 설명

- 마케팅, 고객 관리, 비즈니스 개발 등의 문제점 영역을 식별합니다.
- 일반적인 관점에서 문제점을 설명합니다.
- 프로젝트의 필수조건을 명시합니다. 프로젝트 이면의 동기는 무엇입니까? 비즈니스에서 데이터 마이닝을 이미 사용하고 있습니까?
- 비즈니스 그룹에서 데이터 마이닝 프로젝트의 상태를 확인합니다. 관련 업무가 승인되었습니까? 아니면 데이터 마이닝이 비즈니스 그룹에 대한 핵심 기술로 "광고"되어야 합니까?
- 필요한 경우, 조직에 발표할 데이터 마이닝 관련 정보 프리젠테이션을 준비합니다.

태스크 3--현재 솔루션 설명

- 비즈니스 문제점을 처리하기 위해 현재 사용되는 솔루션을 설명합니다.
- 현재 솔루션의 장점과 단점을 설명합니다. 또한 조직 내에서 이 솔루션이 받아들여지는 수준을 설명합니다.

비즈니스 목표 정의

여기서 관련 항목을 구체화합니다. 연구 및 회의의 결과로서, 프로젝트 스폰서와 해당 결과에 영향을 받는 기타 비즈니스 단위가 합의한 구체적인 기본 목표를 구축해야 합니다. 이 목적은 결국 "고객 이탈 감소" 같은 막연한 것에서 분석의 가이드가 될 구체적인 데이터 마이닝 목표로 바뀌게 됩니다.

태스크 목록

나중에 프로젝트 계획에 통합할 수 있도록 다음 사항을 기록해 두십시오. 목적을 현실에 맞게 유지하십시오.

- 데이터 마이닝을 사용하여 해결하려는 문제점을 설명합니다.
- 모든 비즈니스 질문을 가능한 한 정확히 지정합니다.
- 다른 비즈니스 요구사항(예: 교차 판매 기회를 증가시키면서 기존 고객을 잃지 않음)을 판별합니다.
- 예상되는 혜택을 비즈니스 용어로 지정합니다(예: 우수 고객의 이탈률을 10% 줄임).

비즈니스 성공 기준

앞에 있는 목적은 분명할 수 있지만 목적을 달성하게 되면 아시겠습니까? 추가적으로 나아가기 전에 데이터 마이닝 프로젝트에 대한 비즈니스 성공의 성질을 정의하는 것은 중요합니다. 성공 기준은 다음 두 개의 범주로 분류됩니다.

- **객관적.** 이러한 기준은 감사 정확도 또는 합의된 이탈 감소율의 구체적인 증가처럼 단순할 수 있습니다.
- **주관적.** "유효 처리의 군집 찾기"와 같은 주관적 기준은 특정하기가 더 어렵지만 누가 최종 결정을 하느냐는 합의할 수 있습니다.

태스크 목록

- 가능한 한 정확하게, 이 프로젝트에 대한 성공 기준을 문서화하십시오.
- 각 비즈니스 목표가 성공을 위한 상관성 있는 기준을 가지는지 확인하십시오.
- 주관적 성공 측정의 결정자를 배정하십시오. 가능한 경우, 해당 기대치에 대한 설명을 작성하십시오.

상황 평가

이제 분명히 정의된 목적이 있으므로 현재 자신의 위치에 대해 평가할 때입니다. 이 단계에서는 다음과 같은 질문을 제기하는 것이 포함됩니다.

- 어떤 종류의 데이터를 분석에 사용할 수 있습니까?
- 프로젝트를 완료하는 데 필요한 인원이 있습니까?
- 관련된 가장 큰 위험 요인은 무엇입니까?
- 각 위험에 대한 비상 계획을 가지고 있습니까?

E-소매 예제--상황 평가

CRISP-DM을 사용하는 웹 마이닝 시나리오

이것은 전자제품 e-소매업자의 첫 번째 웹 마이닝 시도이며 회사는 데이터 마이닝 시작을 돕기 위해 데이터 마이닝 전문가를 초빙하기로 결정했습니다. 컨설턴트가 마주하는 첫 번째 태스크 중 하나는 데이터 마이닝을 위한 회사의 자원을 평가하는 것입니다.

직원. 서버 로그와 제품 및 구매 데이터베이스를 관리하는 사내 전문가는 있지만 분석을 위한 데이터 웨어하우징 및 데이터 정리에 대한 경험은 거의 없다는 것은 분명합니다. 그러므로 데이터베이스 전문가도 초빙할 수 있습니다. 회사는 연구 결과가 지속적인 웹 마이닝 프로세스의 일부가 되기를 희망하므로 경영진은 현재 업무 중에 생성된 직위가 영구적 직위가 될 것인지 여부도 고려해야 합니다.

데이터. 이 회사는 저명한 회사이므로 끌어다 쓸 웹 로그 및 구매 데이터가 많이 있습니다. 사실 이 초기 연구의 경우 회사는 분석을 사이트에 등록한 고객으로 제한할 것입니다. 연구가 성공적이라면 프로그램을 확장할 수 있습니다.

위험. 컨설턴트에 대한 금전적 경비 및 직원이 연구에 소요하는 시간을 제외하면 이 모험에서 즉각적인 위험은 그리 많지 않습니다. 그러나 시간은 항상 중요하므로 이 초기 프로젝트는 하나의 회계 분기에 대해서만 스케줄링됩니다.

또한 현재는 추가 현금 흐름이 많지 않으므로 예산에 맞게 연구를 진행해야 합니다. 이러한 목적 중에 달성하기 어려운 목적이 있으면 비즈니스 관리자는 프로젝트의 범위를 줄여야 한다고 제안했습니다.

자원 명세

사용자의 자원에 대한 정확한 자원 명세를 기록하는 것이 꼭 필요합니다. 하드웨어, 데이터 소스 및 직원 관련 사항을 꼼꼼히 확인함으로써 많은 시간을 절약하고 골치아픈 문제를 피할 수 있습니다.

태스크 1--하드웨어 자원 조사

- 어떤 하드웨어를 지원해야 합니까?

태스크 2--데이터 소스 및 지식 저장소 식별

- 어느 데이터 소스가 데이터 마이닝에 사용 가능합니까? 데이터 유형 및 형식에 대한 설명을 작성하십시오.
- 데이터는 어떻게 저장됩니까? 데이터 웨어하우스 또는 운영 데이터베이스에 액세스할 수 있습니까?
- 외부 데이터(예: 인구 통계 정보)를 구매할 계획입니까?
- 필요한 데이터의 액세스를 막는 보안 문제가 있습니까?

태스크 3--인적 자원 식별

- 비즈니스 및 데이터 전문가에게 액세스할 수 있습니까?
- 필요할지도 모르는 데이터베이스 관리자 및 기타 지원 담당자를 식별했습니까?

이와 같은 질문을 했으면 단계 보고서에 대해 담당자 및 자원 목록을 포함시키십시오.

요구사항, 가정 및 제약조건

프로젝트에 대한 책임을 정직하게 평가하면 노력의 성과를 거둘 가능성이 더 높아집니다. 이러한 관심사를 가능한 한 명확하게 해두면 이후의 문제를 피하는 데 도움이 됩니다.

태스크 1--요구사항 결정

기본적인 요구사항은 앞에서 논의한 비즈니스 목적이지만 다음을 고려하십시오.

- 데이터 또는 프로젝트 결과에 대한 보안 및 법적 제한사항이 있습니까?
- 모든 사람이 프로젝트 스케줄링 요구사항에 맞춰져 있습니까?
- 결과 배포에 대한 요구사항(예: 웹에 게시 또는 스코어를 데이터베이스에 읽어들이기)이 있습니까?

태스크 2--가정 명시

- 프로젝트에 영향을 미칠 수 있는 경제적 요인(예: 자문료 또는 경쟁사 제품)이 있습니까?
- 데이터 품질 가정이 있습니까?
- 프로젝트 스폰서/관리 팀은 어떤 결과를 볼 것으로 기대합니까? 즉, 모델 자체를 이해하고 싶습니까, 아니면 단순히 결과를 보고 싶습니까?

태스크 3--제약조건 확인

- 데이터 액세스에 필요한 모든 비밀번호가 있습니까?
- 데이터 사용에 대한 모든 법적 제약조건을 확인했습니까?
- 모든 재정적 제약조건을 프로젝트 예산에서 다루었습니까?

위험 및 비상사태

프로젝트 과정에서 가능한 위험을 고려하는 것도 현명합니다. 위험의 유형은 다음과 같습니다.

- 스케줄링(프로젝트가 예상보다 오래 걸리면 어떨까?)
- 재정(프로젝트 스폰서가 예산 문제에 부딪치면 어떨까?)
- 데이터(데이터의 품질 또는 범위가 올바르지 않다면 어떨까?)
- 결과(초기 결과가 기대한 것보다 인상적이지 않다면 어떨까?)

다양한 위험을 고려한 후, 재해를 피하는 데 도움이 될 비상 계획을 수립하십시오.

태스크 목록

- 가능한 각 위험을 문서화합니다.
- 각 위험에 대한 비상 계획을 문서화합니다.

용어

비즈니스 및 데이터 마이닝 팀이 "동일한 언어로 말하게" 하려면 설명이 필요한 기술 용어 및 전문어의 용어집 컴파일을 고려해야 합니다. 예를 들어, 비즈니스에 대한 "이탈"이 특별하고 고유한 의미를 가지고 있다면 전체 팀을 위해 이에 대해 명시적으로 설명할 가치가 있습니다. 마찬가지로, 팀에서는 Gains 차트의 사용법에 대한 명확한 설명이 필요할 수 있습니다.

태스크 목록

- 팀 멤버에게 혼동되는 용어 또는 전문어 목록을 작성해 둡니다. 비즈니스 용어와 데이터 마이닝 용어를 모두 포함시키십시오.
- 인트라넷 또는 기타 프로젝트 문서에 목록을 공개하는 것을 고려하십시오.

비용/혜택 분석

이 단계에서는 **최종 결과가 어떻습니까?**라는 질문에 응답합니다. 최종 평가의 일부로서, 프로젝트의 비용을 성공의 잠재적 혜택과 비교하는 것이 중요합니다.

태스크 목록

다음에 대한 예상 비용을 분석에 포함시키십시오.

- 데이터 수집 및 사용된 외부 데이터
- 결과 배포
- 운영 비용

그리고 나서 다음과 같은 혜택을 고려하십시오.

- 기본 목표 충족
- 데이터 탐색에서 얻은 추가 통찰력
- 데이터 이해 증진을 통한 잠재적 혜택

데이터 마이닝 목적 결정

이제 비즈니스 목적이 분명하므로 이를 데이터 마이닝으로 실현할 때입니다. 예를 들어, "이탈을 줄이는" 비즈니스 목표는 다음을 포함하는 데이터 마이닝 목적으로 변환될 수 있습니다.

- 최근의 구매 데이터를 기반으로 우수 고객 식별
- 각 고객의 이탈 가능성을 예측하기 위해 가용 고객 데이터를 사용하여 모델 작성
- 이탈 성향 및 고객 가치를 기반으로 각 고객에게 순위 지정

이러한 데이터 마이닝 목적은 충족될 경우 업체에서 우수 고객 중에 이탈을 줄이는 데 사용될 수 있습니다.

이와 같이, 업체와 기술은 효과적인 데이터 마이닝을 위해 함께 협력해야 합니다. 데이터 마이닝 목적을 결정하는 방법에 대한 구체적인 팁을 계속 읽어보십시오.

데이터 마이닝 목적

비즈니스 문제점에 대한 기술적 솔루션을 정의하기 위해 비즈니스 및 데이터 분석가와 작업을 진행할 때처럼, 항목들을 구체적으로 정의해 두십시오.

태스크 목록

- 데이터 마이닝 문제점의 **유형**(군집, 예측, 분류 등)을 설명합니다.
- 특정 시간 단위를 사용하여 기술적 목적을 문서화합니다(예: 유효 기간이 3개월인 예측).
- 가능한 경우, 바람직한 결과의 실제 수치(예: 기존 고객의 80%에 대한 이탈 점수 생성)를 제공하십시오.

E-소매 예제--데이터 마이닝 목적

CRISP-DM을 사용하는 웹 마이닝 시나리오

해당 데이터 마이닝 컨설턴트의 도움으로, e-소매업자는 회사의 비즈니스 목표를 데이터 마이닝 용어로 변환할 수 있었습니다. 이 분기에 완료할 초기 연구의 목적은 다음과 같습니다.

- 이전 구매에 대한 히스토리 정보를 사용하여 "관련" 항목들을 링크하는 모델을 생성합니다. 사용자가 항목 설명을 볼 때, 관련 그룹(**장바구니 분석**)의 다른 항목에 대한 링크를 제공합니다.
- 웹 로그를 사용하여 서로 다른 고객들이 무엇을 찾고 있는지 판별하고 이러한 항목을 강조표시하도록 사이트를 재설계합니다. 서로 다른 각 고객 "유형"마다 사이트의 메인 페이지가 다르게 표시됩니다(**프로파일링**).
- 웹 로그를 사용하여 사용자가 어디에서 왔고 사이트의 어디에 있었는지를 감안해서 사용자가 다음에 어디로 이동할지를 예측합니다(**시퀀스 분석**).

데이터 마이닝 성공 기준

데이터 마이닝 작업이 순조롭게 진행되려면 성공을 기술적인 용어로도 정의해야 합니다. 이전에 결정된 데이터 마이닝 목적을 사용하여 성공에 대한 벤치마크를 공식화합니다. IBM SPSS Modeler는 평가 노드 및 분석 노드 등의 도구를 제공하여 결과의 정확도와 타당성을 분석할 수 있게 합니다.

태스크 목록

- 모델 평가 방법을 설명합니다(예: 정확도, 성과 등).
- 성공을 평가하기 위한 벤치마크를 정의합니다. 특정 수치를 제공합니다.
- 최선의 주관적 측정을 정의하고 성공의 결정자를 결정합니다.
- 모델 결과의 성공적인 배포가 데이터 마이닝 성공의 일부인지 여부를 고려합니다. 배포 계획을 지금 시작합니다.

프로젝트 계획 생성

이 시점에서 데이터 마이닝 프로젝트의 계획을 생성할 준비가 되었습니다. 지금까지 제기한 질문과, 공식화한 비즈니스 및 데이터 마이닝 목적은 이 로드맵의 기반이 될 것입니다.

프로젝트 계획 작성

프로젝트 계획은 모든 데이터 마이닝 작업에 대한 마스터 문서입니다. 제대로 작성되면 프로젝트와 연관된 모든 사람들에게 모든 데이터 마이닝 단계에 대한 목적, 자원, 위험 및 스케줄을 알려줄 수 있습니다. 이 계획뿐 아니라 이 단계에서 수집된 문서를 회사의 인트라넷에 게시할 수도 있습니다.

태스크 목록

계획을 작성할 때는 다음과 같은 질문에 응답해야 합니다.

- 프로젝트 태스크 및 제안된 계획을 관련된 모든 사람과 논의했습니까?
- 모든 단계 또는 태스크에 대한 시간 추정값이 포함되었습니까?
- 결과 또는 비즈니스 솔루션을 배포하는 데 필요한 작업 및 자원을 포함시켰습니까?
- 의사결정 사항과 검토 요청이 계획에 강조표시되었습니까?
- 다중 반복이 일반적으로 발생하는 단계(예: 모델링)를 식별 표시했습니까?

샘플 프로젝트 계획

연구에 대한 개요 계획은 아래 표와 같습니다.

표 *1.* 샘플 프로젝트 계획 개요

단계	시간	자원	위험
비즈니스 이해	1주	모든 분석가	경제 변화
데이터 이해	3주	모든 분석가	데이터 문제점, 기술 문제점
데이터 준비	5주	데이터 마이닝 컨설턴트, 일부 데이터베이스 분석가 시간	데이터 문제점, 기술 문제점
모델링	2주	데이터 마이닝 컨설턴트, 일부 데이터베이스 분석가 시간	기술 문제점, 적절한 모델을 찾을 수 없음
평가	1주	모든 분석가	경제 변화, 결과를 구현할 수 없음
배포	1주	데이터 마이닝 컨설턴트, 일부 데이터베이스 분석가 시간	경제 변화, 결과를 구현할 수 없음

도구 및 기법 평가

IBM SPSS Modeler를 이미 데이터 마이닝 성공을 위한 도구로 사용하도록 선택했으므로 이 단계를 사용하여 어느 데이터 마이닝 기법이 비즈니스 요구사항에 가장 적절한지 조사할 수 있습니다. IBM SPSS Modeler는 데이터 마이닝의 각 단계에 대한 전체적인 도구를 제공합니다. 다양한 기법을 언제 사용할지에 대해서는 온라인 도움말의 모델링 섹션을 참고하십시오.

다음 단계에 대한 준비 여부

IBM SPSS Modeler에서 데이터 탐색과 작업 시작 전에 다음 질문에 대한 응답을 해야 합니다.

비즈니스 관점에서:

- 해당 사업체가 이 프로젝트로부터 무엇을 얻고자 합니까?
- 관련 업무의 성공적인 완료를 어떻게 정의하시겠습니까?
- 목적을 달성하는 데 필요한 예산과 자원이 있습니까?
- 이 프로젝트에 필요한 모든 데이터에 액세스할 수 있습니까?
- 이 프로젝트와 연관된 위험 및 비상 계획을 팀과 논의했습니까?
- 비용과 혜택 분석의 결과가 이 프로젝트의 가치를 증명했습니까?

위의 질문에 응답한 후, 그 응답을 데이터 마이닝 목적으로 변환했습니까?

데이터 마이닝 관점에서:

- 얼마나 구체적으로 데이터 마이닝이 비즈니스 목적의 달성에 도움을 줄 수 있습니까?
- 어떤 데이터 마이닝 기법이 최상의 결과를 산출할지에 대한 아이디어가 있습니까?
- 언제 결과가 정확하거나 충분히 효과적인지 어떻게 알 수 있습니까? (데이터 마이닝 성공에 대한 측정 기준을 설정했습니까?)
- 어떻게 모델링 결과가 배포될 것입니까? 사용자의 프로젝트 계획에서 배포를 고려했습니까?
- 프로젝트 계획에 CRISP-DM의 모든 단계가 포함됩니까?
- 위험 및 종속 항목을 계획에서 다룹니까?

위의 질문에 "예"라고 응답할 수 있다면 데이터를 더 자세히 살펴볼 준비가 된 것입니다.

제 3 장 데이터 이해

데이터 이해 개요

CRISP-DM의 데이터 이해 단계에서는 마이닝에 사용 가능한 데이터를 자세히 살펴봐야 합니다. 이 단계는 다음 단계(데이터 준비) 동안 예기치 못한 문제를 피하는 데 있어서 중요하며, 일반적으로 프로젝트의 가장 긴 부분입니다.

데이터 이해를 위해서는 데이터에 액세스하고 테이블 및 그래픽(IBM SPSS Modeler에서 CRISP-DM 프로젝트 도구를 사용하여 구성 가능)을 사용하여 데이터를 탐색해야 합니다. 이렇게 하면 데이터의 품질을 판별하고 프로젝트 문서에서 이러한 단계의 결과를 설명할 수 있습니다.

초기 데이터 수집

CRISP-DM의 이 시점에서는, 데이터에 액세스하여 데이터를 IBM SPSS Modeler로 가져올 준비가 되었습니다. 데이터는 다음과 같은 다양한 소스에서 가져옵니다.

- **기존 데이터.** 이 데이터는 트랜잭션 데이터, 설문조사 데이터, 웹 로그 등의 매우 다양한 데이터를 포함합니다. 기존 데이터가 해당 요구사항을 충족하기에 충분한지 여부를 고려하십시오.
- **구매한 데이터.** 조직에서 인구 통계 등의 보충 데이터를 사용합니까? 그렇지 않다면 해당 데이터가 필요할지 여부를 고려하십시오.
- **추가 데이터.** 위의 소스가 해당 요구사항을 충족시키지 못하면 설문조사를 수행하거나 추가 추적을 시작하여 기존 데이터 저장소를 보충할 수 있습니다.

태스크 목록

IBM SPSS Modeler에서 데이터를 살펴보고 다음과 같은 질문을 고려하십시오. 결과물에 대한 설명을 작성하십시오. 자세한 정보는 14 페이지의 『데이터 수집 보고서 작성』의 내용을 참조하십시오.

- 데이터베이스의 어느 속성(열)이 가장 유망해 보입니까?
- 무관해 보여서 제외할 수 있는 속성은 무엇입니까?
- 일반화 가능한 결론을 도출하거나 정확한 예측을 하기에 충분한 데이터가 있습니까?
- 선택할 모델링 방법의 속성이 너무 많이 있습니까?
- 다양한 데이터 소스를 병합할 것입니까? 그렇다면, 병합할 때 문제를 일으킬지도 모르는 영역이 있습니까?
- 각 데이터 소스에서 결측값을 어떻게 처리할지에 대해 고려했습니까?

E-소매 예제--초기 데이터 수집

CRISP-DM을 사용하는 웹 마이닝 시나리오

이 예제에서 e-소매업자는 다음을 비롯하여 몇 가지 중요한 데이터 소스를 사용합니다.

웹 로그. 원시 액세스 로그에는 고객이 웹 사이트를 탐색하는 방식에 대한 모든 정보가 있습니다. 웹 로그에서 이미지 파일에 대한 참조와 기타 비정보 엔트리는 데이터 준비 과정에서 제거되어야 합니다.

구매 데이터. 고객이 주문을 제출하면 해당 주문에 속한 모든 정보가 저장됩니다. 구매 데이터베이스에 있는 주문은 웹 로그의 해당 세션에 맵핑되어야 합니다.

제품 데이터베이스. 제품 속성은 "관련" 제품을 결정할 때 유용할 수 있습니다. 제품 정보는 해당 주문에 맵핑되어야 합니다.

고객 데이터베이스. 이 데이터베이스는 등록된 고객에게서 수집된 추가 정보를 포함합니다. 많은 고객이 질문지를 채우지 않기 때문에, 전체 레코드는 완성되지 않습니다. 고객 정보는 웹 로그의 해당 구매 및 세션에 맵핑되어야 합니다.

현재, 회사에서는 해당 분석가가 현재 가진 데이터를 관리하느라 바쁘기 때문에 외부 데이터베이스를 구매하거나 설문조사 수행에 자금을 지출할 계획이 없습니다. 그러나 언젠가 데이터 마이닝 결과의 확장 배포를 고려할 필요가 있는 경우 미등록 고객의 추가 인구 통계 데이터를 구매하는 것이 상당히 유용할 수 있습니다. 또한 e-소매업자의 고객 데이터베이스가 평균 인터넷 쇼핑객과 어떻게 다른지를 보기 위해 인구 통계 정보를 확보하는 것이 유용할 수 있습니다.

데이터 수집 보고서 작성

이전 단계에서 수집된 자료를 사용하여 데이터 수집 보고서 작성을 시작할 수 있습니다. 작성이 완료되면 이 보고서는 프로젝트 웹 사이트에 추가되거나 팀에게 베포될 수 있습니다. 이 보고서는 다음 단계(데이터 설명, 탐색 및 품질 확인)에서 준비된 보고서와 결합될 수도 있습니다. 이러한 보고서는 데이터 준비 단계에서 사용자의 작업을 안내할 것입니다.

데이터 설명

데이터를 설명하는 방법은 여러 가지가 있지만 대부분의 설명은 데이터의 수량 및 품질(얼마나 많은 데이터가 사용 가능하며 데이터의 상태는 어떠한가)에 초점을 맞춥니다. 아래에는 데이터를 설명할 때 다루어야 하는 몇 가지 핵심 특성이 나열되어 있습니다.

- **데이터의 양.** 대부분의 모델링 기법에는 데이터 크기와 연관된 장단점이 있습니다. 큰 데이터 세트는 더 정확한 모델을 생성할 수 있지만 처리 시간이 늘어날 수 있습니다. 데이터의 서브세트를 사용하는 것이 적절한지 여부를 고려하십시오. 최종 보고서에 대한 설명을 작성할 때, 모든 데이터 세트의 크기 통계를 포함해야 하고 데이터를 설명할 때 레코드 수 뿐만 아니라 필드(속성)도 고려해야 하는 것을 명심하십시오.
- **값 유형.** 데이터는 **숫자** 또는 **범주형**(문자열) 또는 **부울**(true/false)와 같은 다양한 형식을 가질 수 있습니다. 값 유형에 주의하면 이후 모델링 중에 문제를 피할 수 있습니다.
- **코딩 체계.** 데이터베이스의 값은 성별 또는 제품 유형과 같은 특성의 표현인 경우가 빈번합니다. 예를 들어, 한 데이터 세트에서는 *M* 및 *F*를 사용하여 남성 및 여성을 나타내고, 다른 데이터 세트에서는 숫자 값 *1* 및 *2*를 사용할 수 있습니다. 데이터 보고서에서 충돌하는 체계를 메모해 두십시오.

이 지식을 갖췄으므로 이제 데이터 설명 보고서를 작성하고 결과물을 더 많은 대상과 공유할 준비가 되었습니다.

E-소매 예제--데이터 설명

CRISP-DM을 사용하는 웹 마이닝 시나리오

웹 마이닝 애플리케이션에는 처리할 레코드 및 속성이 많이 있습니다. 이 데이터 마이닝 프로젝트를 수행하는 e-소매업자가 초기 연구를 사이트에 등록한 30,000여 명의 고객으로 제한했을지라도 웹 로그에는 여전히 수백 만 개의 레코드가 있습니다.

이러한 데이터 소스에 있는 대부분의 값 유형은 (날짜 및 시간이든, 액세스한 웹 페이지 수이든, 등록 질문지의 다중 선택 질문에 대한 응답이든 간에) 기호입니다. 이러한 변수의 일부는 수치인 새 변수(예: 방문한 웹 페이지 수, 웹 사이트에서 보낸 시간)를 작성하는데 사용됩니다. 데이터 소스 내의 몇 가지 기존 숫자 변수로는 주문한 각 제품 수, 구매 중에 지출한 금액, 제품 데이터베이스의 제품 중량과 치수 내역 등이 있습니다.

데이터 소스는 매우 다른 속성들을 포함하기 때문에 다양한 데이터 소스에 대한 코딩 체계에는 중복되는 것이 거의 없습니다. 중복되는 유일한 변수는 고객 ID, 제품 코드 등의 "키"입니다. 이러한 변수는 데이터 소스 간에 동일한 코딩 체계를 가져야 합니다. 그렇지 않으면 데이터 소스를 병합할 수 없습니다. 병합을 위해 이러한 키 필드를 다시 코딩하려면 약간의 추가 데이터 준비가 필요할 것입니다.

데이터 설명 보고서 작성

데이터 마이닝 프로젝트를 효과적으로 진행하려면 다음과 같은 메트릭을 사용하여 정확한 데이터 설명 보고서 생성의 의미를 고려하십시오.

데이터 양

- 데이터의 형식은 무엇입니까?
- 데이터를 캡처하는 데 사용되는 방법(예: ODBC)을 식별합니다.
- (행 및 열 수 면에서) 데이터베이스가 얼마나 큽니까?

데이터 품질

- 데이터는 비즈니스 질문과 관련된 특성을 포함합니까?
- 어떤 데이터 유형(기호, 숫자 등)이 존재합니까?
- 핵심 속성의 기본 통계를 계산했습니까? 이것은 비즈니스 질문에 대해 어떤 통찰력을 제공했습니까?
- 관련 속성의 우선순위를 지정할 수 있습니까? 그럴 수 없다면 추가 통찰력을 제공할 비즈니스 분석가가 있습니까?

데이터 탐색

IBM SPSS Modeler에서 사용 가능한 테이블, 차트 및 기타 시각화 도구를 사용하여 데이터를 탐색하려면 CRISP-DM의 이 단계를 사용하십시오. 이러한 분석은 비즈니스 이해 단계 동안 구축된 데이터 마이닝 목적을 달성하는 데 도움이 될 수 있습니다. 또한 이를 통해 가설을 공식화하고 데이터 준비 동안 발생하는 데이터 변환 작업을 구체화할 수 있습니다.

E-소매 예제--데이터 탐색

CRISP-DM을 사용하는 웹 마이닝 시나리오

CRISP-DM은 이 시점에서 초기 탐색의 수행을 제안하지만 e-소매업자가 발견한 바와 같이 원시 웹 로그에서 데이터 탐색은 불가능하지는 않지만 어렵습니다. 일반적으로 웹 로그 데이터는 의미 있게 탐색 가능한 데이터를 생성하도록 데이터 준비 단계에서 먼저 처리되어야 합니다. CRISP-DM에서의 이 출발은 프로세스가 사용자의 특별한 데이터 마이닝 요구사항에 맞게 사용자 정의될 수 있으며 사용자 정의되어야 한다는 사실을 강조합니다. CRISP-DM은 주기적이며, 일반적으로 데이터 마이너가 단계 사이에 앞뒤로 이동합니다.

웹 로그는 탐색 전에 처리되어야 하지만 e-소매업자가 사용 가능한 다른 데이터 소스는 탐색에 더 용이합니다. 구매 데이터베이스를 탐색에 사용할 경우 고객에 대한 흥미로운 요약값(예: 고객의 지출 금액, 구매당 구입하는 물품 수, 고객 출신지)을 찾을 수 있습니다. 고객 데이터베이스의 요약값은 등록 질문지의 문항에 대한 응답의 분포를 보여줍니다.

탐색은 데이터의 오류를 찾는 데에도 유용합니다. 대부분의 데이터 소스는 자동으로 생성되는 반면, 제품 데이터베이스의 정보는 수동으로 입력되었습니다. 나열된 제품 치수의 몇 가지 빠른 요약값을 통해 "119인치"("19인치"의 오타)와 같은 오타를 발견할 수 있습니다.

데이터 탐색 보고서 작성

그래프를 만들고 사용 가능 데이터에 대한 통계를 실행하면서, 데이터가 어떻게 기술적 및 비즈니스 목적에 해답을 제공할 수 있는지에 대한 가설 형성을 시작하십시오.

태스크 목록

데이터 탐색 보고서에 포함할 결과물에 대한 설명을 작성하십시오. 다음 질문에 응답하십시오.

- 데이터에 대해 어떤 종류의 가설을 형성했습니까?
- 어느 속성이 추가 분석에 유망해 보입니까?
- 탐색을 통해 데이터에 대한 새 특성이 밝혀졌습니까?
- 이러한 탐색이 초기 가설을 어떻게 변경했습니까?
- 나중에 사용하기 위해 데이터의 특정 서브세트를 식별할 수 있습니까?
- 데이터 마이닝 목적을 다시 살펴보십시오. 이 탐색으로 인해 목적이 변경되었습니까?

데이터 품질 확인

데이터는 완전한 경우가 거의 없습니다. 결국, 대부분의 데이터는 때때로 분석을 까다롭게 하는 코딩 오류, 결측값 또는 기타 유형의 불일치를 포함합니다. 잠재적 위험을 피하기 위한 한 가지 방법은 모델링 전에 사용 가능 데이터의 철저한 품질 분석을 수행하는 것입니다.

IBM SPSS Modeler의 보고서 작성 도구(예: 데이터 검토, 테이블 및 기타 출력 노드)는 다음과 같은 유형의 문제점을 찾는 데 도움이 될 수 있습니다.

- **누락된 데이터**. 무응답(예: *$null$*, *?* 또는 *999*)으로 코딩되었거나 비어 있는 값이 포함됩니다.
- **데이터 오류**. 일반적으로 데이터를 입력할 때 생기는 오타 오류입니다.
- **측정 오류**. 올바르게 입력되었지만 부정확한 측정 체계를 기반으로 하는 데이터가 포함됩니다.
- **코딩 불일치**. 일반적으로 비표준 측정 단위 또는 값 불일치(예: 성별에 대해 *M*과 남성을 모두 사용)가 포함됩니다.
- **잘못된 메타데이터**. 필드의 분명한 의미와, 필드 이름 또는 정의에 명시된 의미 사이의 불일치가 포함됩니다.

이러한 품질 문제에 대한 설명을 작성하십시오. 자세한 정보는 16 페이지의 『데이터 품질 보고서 작성』의 내용을 참조하십시오.

E-소매 예제--데이터 품질 확인

CRISP-DM을 사용하는 웹 마이닝 시나리오

데이터 품질의 확인은 설명 및 탐색 프로세스 과정에서 이루어집니다. e-소매업자가 마주치게 되는 몇 가지 문제는 다음과 같습니다.

누락된 데이터. 알려진 누락 데이터는 등록된 일부 사용자가 질문지에 응답하지 않은 경우를 포함합니다. 질문지에 의해 제공된 추가 정보가 없으면, 이러한 고객은 일부 후속 모델에서 배제되어야 할 것입니다.

데이터 오류. 대부분의 데이터 소스는 자동으로 생성되므로 이것은 큰 문제가 아닙니다. 제품 데이터베이스의 오타 오류는 탐색 프로세스에서 발견될 수 있습니다.

측정 오류. 측정 오류의 가능성이 가장 큰 소스는 질문지입니다. 문항이 신중하지 못하거나 적절하게 표현되지 않으면 e-소매업자가 얻고자 하는 정보를 제공하지 못할 수 있습니다. 역시, 탐색 프로세스에서 특별한 응답 분포를 가지는 문항에 특별한 주의를 기울이는 것이 중요합니다.

데이터 품질 보고서 작성

데이터 품질에 대한 탐색 및 확인을 기반으로, 이제 CRISP-DM의 다음 단계를 안내할 보고서를 준비할 준비가 되었습니다. 자세한 정보는 15 페이지의 『데이터 품질 확인』의 내용을 참조하십시오.

태스크 목록

앞에서 논의했듯이, 몇 가지 유형의 데이터 품질 문제점이 있습니다. 다음 단계로 이동하기 전에, 다음과 같은 품질 문제를 고려하고 솔루션을 계획하십시오. 모든 응답을 데이터 품질 보고서에서 문서화하십시오.

- 누락된 속성 및 비어 있는 필드를 식별했습니까? 그렇다면, 이러한 결측값 속에 숨은 의미가 있습니까?
- 이후의 병합 또는 변환에서 문제를 일으킬 수 있는 맞춤법 불일치가 있습니까?
- 편차를 탐색하여, 편차가 "잡음"인지 또는 추가적으로 분석할 가치가 있는 현상인지 판별했습니까?
- 값에 대해 타당성 검사를 수행했습니까? 분명하게 모순적인 사항(예: 소득 수준이 높은 청소년)에 대해 설명을 작성하십시오.
- 가설에 아무 영향을 미치지 않는 데이터를 제외하는 것을 고려했습니까?
- 데이터가 플랫 파일에 저장됩니까? 그렇다면, 구분자가 파일 간에 일치합니까? 각 레코드는 동일한 수의 필드를 포함합니까?

다음 단계에 대한 준비 여부

IBM SPSS Modeler에서 모델링 데이터를 준비하기 전에 다음 사항을 고려하십시오.

데이터를 얼마나 잘 이해하고 있습니까?

- 모든 데이터 소스가 분명히 식별되고 액세스됩니까? 문제점 또는 제한사항을 알고 있습니까?
- 사용 가능 데이터로부터 핵심 속성을 식별했습니까?
- 이러한 속성은 가설을 공식화하도록 도왔습니까?
- 모든 데이터 소스의 크기를 기재했습니까?

- 적당한 곳에서 데이터의 서브세트를 사용할 수 있습니까?
- 각 관심 속성의 기본 통계를 계산했습니까? 의미 있는 정보가 나타났습니까?
- 핵심 속성에 대한 추가 통찰력을 얻기 위해 탐색 그래픽을 사용했습니까? 이 통찰력을 통해 가설이 변동되었습니까?
- 이 프로젝트에 대한 데이터 품질 문제는 무엇입니까? 이 문제를 해결할 계획이 있습니까?
- 데이터 준비 단계가 분명합니까? 예를 들면, 어느 데이터 소스를 병합시키고 어느 속성을 필터링하거나 선택할지 압니까?

이제 비즈니스와 데이터 이해를 모두 갖췄으므로 IBM SPSS Modeler를 사용하여 모델링에 대한 데이터를 준비할 때입니다.

제 4 장 데이터 준비

데이터 준비 개요

데이터 준비는 데이터 마이닝에서 가장 중요하면서 시간이 많이 걸리는 측면 중 하나입니다. 결국, 데이터 준비는 일반적으로 프로젝트에 대한 시간 및 작업량의 50-70%를 차지하는 것으로 추정됩니다. 앞선 비즈니스 이해 및 데이터 이해 단계에 적당한 에너지를 쏟으면 이 오버헤드를 최소화할 수 있지만, 여전히 마이닝 데이터를 준비하고 패키징하는 데 상당한 노력을 기울여야 합니다.

조직과 해당 목적에 따라, 데이터 준비는 일반적으로 다음과 같은 작업을 포함합니다.

- 데이터 세트 및/또는 레코드 병합
- 데이터의 표본 서브세트 선택
- 레코드 통합
- 새 속성 파생
- 모델링 데이터 정렬
- 공백 또는 결측값을 제거하거나 대체
- 학습 및 검정 데이터 세트로 분할

데이터 선택

이전 CRISP-DM 단계에서 수행된 초기 데이터 수집을 기반으로, 데이터 마이닝 목적과 관련된 데이터 선택을 시작할 준비가 되었습니다. 일반적으로 데이터를 선택하는 방법은 두 가지가 있습니다.

- **항목(행) 선택**. 어느 계정, 제품 또는 고객을 포함시킬까 등에 관한 의사결정이 포함됩니다.
- **속성 또는 특성(열) 선택**. 트랜잭션 금액 또는 가계 소득 등의 특성 사용에 관한 의사결정이 포함됩니다.

E-소매 예제--데이터 선택

CRISP-DM을 사용하는 웹 마이닝 시나리오

어느 데이터를 선택할지에 대한 e-소매업자의 의사결정 다수는 이미 데이터 마이닝 프로세스의 초기 단계에서 수행되었습니다.

항목 선택. 초기 연구는 사이트에 등록한 30,000여 명의 고객으로 제한될 것이므로 비등록 고객의 구매 및 웹 로그를 제외하기 위해 필터를 설정해야 합니다. 웹 로그의 이미지 파일 및 기타 비정보 엔트리를 제거하기 위해 기타 필터가 설정되어야 합니다.

속성 선택. 구매 데이터베이스는 e-소매업자의 고객과 관련된 민감한 정보를 포함할 것이므로 고객 이름, 주소, 전화번호, 신용카드 번호 등의 속성을 필터링하는 것이 중요합니다.

데이터 포함 또는 제외

포함하거나 제외할 데이터의 서브세트를 결정할 때는 의사결정 이면의 근본적 이유를 문서화하십시오.

고려할 질문

- 지정된 속성이 데이터 마이닝 목적과 관련되어 있습니까?
- 특정 데이터 세트 또는 속성의 품질이 결과의 타당성에 방해가 됩니까?
- 이러한 데이터를 구제할 수 있습니까?
- 성별 또는 인종과 같은 특수 필드를 사용하는 것에 대한 제약조건이 있습니까?

여기서의 의사결정이 데이터 이해 단계에서 공식화된 가설과 다릅니까? 그렇다면 프로젝트 보고서에서 해당 추론을 문서화하십시오.

데이터 정리

데이터를 정리하려면 분석을 위해 포함하도록 선택한 데이터에서 문제점을 자세히 살펴봐야 합니다. IBM SPSS Modeler에서 레코드 및 필드 작업 노드를 사용하여 데이터를 정리하는 방법이 몇 가지 있습니다.

표 *2.* 데이터 정리

데이터 문제점	**가능한 솔루션**
누락된 데이터	행 또는 특성을 제외합니다. 또는 공백을 예상 값으로 채웁니다.
데이터 오류	로직을 사용하여 오류를 수동으로 찾아서 대체합니다. 또는 특성을 제외합니다.
코딩 불일치	하나의 코딩 체계를 결정한 후 값을 변환해서 대체합니다.
누락되었거나 잘못된 메타데이터	의심스런 필드를 수동으로 탐색하고 올바른 의미를 추적합니다.

데이터 이해 단계 동안 준비된 데이터 품질 보고서에는 해당 데이터의 특정 문제점 유형에 대한 세부사항이 포함되어 있습니다. IBM SPSS Modeler에서 이 보고서를 데이터 조작에 대한 시작점으로 사용할 수 있습니다.

E-소매 예제--데이터 정리

CRISP-DM을 사용하는 웹 마이닝 시나리오

e-소매업자는 데이터 품질 보고서에서 지적된 문제점을 해결하기 위해 데이터 정리 프로세스를 사용합니다.

누락된 데이터. 온라인 질문지를 완성하지 않은 고객은 이후의 일부 모델에서 배제되어야 할 것입니다. 이러한 고객에게 질문지를 작성하도록 다시 요청할 수도 있지만 e-소매업자가 지출하기에 부담스런 시간과 비용이 소요될 것입니다. e-소매업자가 할 수 있는 것은 질문지에 응답한 고객과 응답하지 않은 고객 사이의 구매 차이를 모델링하는 것입니다. 이러한 두 고객 집합이 유사한 구매 습관을 가지는 경우, 누락된 질문지는 큰 문제가 되지 않습니다.

데이터 오류. 탐색 프로세스 중에 발견된 오류는 여기에서 수정될 수 있습니다. 그러나 대부분은 고객이 페이지를 백엔드 데이터베이스에 제출하기 전에 적절한 데이터 입력이 웹 사이트에서 시행됩니다.

측정 오류. 질문지의 문항이 적절하게 표현되지 않으면 데이터의 품질에 크게 영향을 미칠 수 있습니다. 누락된 질문지와 마찬가지로, 새 대체 질문에 대한 응답을 수집할 시간이나 비용이 없을 수 있으므로 이것은 어려운 문제입니다. 문제가 되는 문항에 대해 최적의 솔루션은 선택 프로세스로 돌아가서 해당 문항을 추가 분석에서 배제하는 것일 수 있습니다.

데이터 정리 보고서 작성

데이터 정리 작업에 대해 보고서를 작성하는 것은 데이터 변경사항 추적에 필수적입니다. 작업 세부사항이 준비되어 있으면 이후의 데이터 마이닝 프로젝트에 도움이 될 것입니다.

태스크 목록

보고서를 작성할 때 다음과 같은 질문을 고려하는 것이 좋습니다.

- 데이터에서 어떤 유형의 잡음이 발생했습니까?
- 잡음을 제거하기 위해 어떤 접근법을 사용했습니까? 어떤 기법이 성공적이었습니까?
- 구제할 수 없는 케이스 또는 속성이 있습니까? 잡음으로 인해 제외된 데이터를 기록하십시오.

새 데이터 구축

새 데이터를 구축해야 하는 경우가 빈번합니다. 예를 들어, 각 트랜잭션에서 보증기간 연장 구매에 플래그를 지정하는 새 열을 작성하는 것이 유용할 수 있습니다. 이 새 필드(*purchased_warranty*)는 IBM SPSS Modeler에서 '플래그로 설정' 노드를 사용하여 쉽게 생성할 수 있습니다.

새 데이터를 구축하는 방법은 두 가지가 있습니다.

- 속성(열 또는 특성) 파생
- 레코드(행) 생성

IBM SPSS Modeler는 해당 '레코드 및 필드 작업' 노드를 사용하여 여러 가지 데이터 구축 방법을 제공합니다.

E-소매 예제--데이터 구축

CRISP-DM을 사용하는 웹 마이닝 시나리오

웹 로그의 처리는 많은 새 속성을 만들어낼 수 있습니다. 로그에 기록된 이벤트의 경우, e-소매업자는 시간소인을 작성하고 방문자 및 세션을 식별하며 액세스된 페이지와 이벤트가 나타내는 활동의 유형을 기재할 수 있습니다. 이러한 변수의 일부는 추가 속성(예: 세션 내에서 이벤트 사이의 시간)을 작성하는데 사용됩니다.

추가 속성은 병합 또는 기타 데이터 구조변환의 결과로서 작성될 수 있습니다. 예를 들어, 각 행이 세션이 되도록 행별 이벤트 웹 로그가 "롤업"되면 총 동작 수, 총 소요 시간 및 세션 중에 수행된 총 구매 수를 기록하는 새 속성이 작성됩니다. 각 행이 고객이 되도록 웹 로그가 고객 데이터베이스와 병합되면 세션 수, 총 동작 수, 총 소요 시간 및 각 고객의 총 구매 수를 기록하는 새 속성이 작성됩니다.

새 데이터를 구축한 후에, e-소매업자는 탐색 과정을 거쳐 데이터 작성이 올바로 수행되었는지 확인합니다.

속성 파생

IBM SPSS Modeler에서 다음과 같은 필드 작업 노드를 사용하여 새 속성을 파생시킬 수 있습니다.

- **파생 노드**를 사용하여 기존 필드에서 파생된 새 필드를 작성합니다.
- **플래그로 설정 노드**를 사용하여 플래그 필드를 작성합니다.

태스크 목록

- 속성을 파생시킬 때 모델링에 대한 데이터 요구사항을 고려하십시오. 모델링 알고리즘에서 특정 데이터 유형(예: 숫자)를 기대합니까? 그렇다면 필요한 변환을 수행하십시오.
- 모델링 전에 데이터를 표준화해야 합니까?
- 통합, 평균화 또는 귀납을 사용하여 누락된 속성을 구축할 수 있습니까?
- 사용자의 배경 지식을 기반으로, 기존 필드에서 파생될 수 있는 중요한 사실(예: 웹 사이트에서 보낸 시간의 길이)이 있습니까?

데이터 통합

동일한 비즈니스 질문 세트에 대해 여러 데이터 소스를 가지는 것은 드물지 않습니다. 예를 들어, 동일한 클라이언트 세트에 대해 주택 담보 대출 데이터뿐 아니라 구매한 인구 통계 데이터에도 액세스할 수 있습니다. 이러한 데이터 세트가 동일한 고유 식별자(예: 주민등록번호)를 포함하는 경우, 이 키 필드를 사용하여 IBM SPSS Modeler에서 데이터 세트를 병합할 수 있습니다.

기본적인 데이터 통합 방법이 두 가지 있습니다.

- 데이터 **병합**의 경우, 유사한 레코드를 가지지만 속성이 다른 두 데이터 세트를 병합합니다. 데이터는 각 레코드의 동일한 키 식별자(예: 고객 ID)를 사용하여 병합됩니다. 열 또는 특성에서 결과 데이터가 증가합니다.
- 데이터 **추가**의 경우, 유사한 속성을 가지지만 레코드가 다른 두 개 이상의 데이터 세트를 통합합니다. 데이터는 유사한 필드(예: 제품 이름 또는 계약 기간)를 기반으로 통합됩니다.

E-소매 예제--데이터 통합

CRISP-DM을 사용하는 웹 마이닝 시나리오

다중 데이터 소스를 사용하는 경우, e-소매업자가 데이터를 통합할 수 있는 여러 가지 방법이 있습니다.

- **이벤트 데이터에 고객 및 제품 속성 추가.** 다른 데이터베이스의 속성을 사용하여 웹 로그 이벤트를 모델링하려면 각 이벤트와 연관된 고객 ID, 제품 번호 및 구매 주문 번호가 올바르게 식별되어야 하고 해당 속성이 처리된 웹 로그에 병합되어야 합니다. 고객 또는 제품이 이벤트와 연관될 때마다 병합 파일이 고객 및 제품 정보를 복제한다는 점을 참고하십시오.

- **고객 데이터에 구매 및 웹 로그 정보 추가.** 고객의 가치를 모델링하려면 고객의 구매 및 세션 정보를 적절한 데이터베이스에서 골라서 합산하여 고객 데이터베이스와 병합해야 합니다. 여기에는 데이터 구축 프로세스에서 논의된 새 속성의 작성이 포함됩니다.

데이터베이스를 통합한 후에, e-소매업자는 탐색 과정을 거쳐 데이터 병합이 올바로 수행되었는지 확인합니다.

통합 태스크

데이터의 이해를 높이기 위해 적절한 노력을 기울이지 않으면 데이터 통합이 복잡해질 수 있습니다. 데이터 마이닝 목적에 가장 관련성 있다고 보이는 항목 및 속성에 대해 숙고한 후 데이터의 통합을 시작하십시오.

태스크 목록

- IBM SPSS Modeler에서 병합 또는 붙여쓰기 노드를 사용하여, 모델링에 유용하다고 생각되는 데이터 세트를 통합합니다.
- 모델링을 진행하기 전에 출력 결과의 저장을 고려하십시오.
- 병합 후에, 데이터는 **통합된** 값에 의해 단순화될 수 있습니다. 통합이란 다중 레코드 및/또는 테이블로부터 정보를 요약하여 새로운 값이 계산된다는 의미입니다.
- 새 레코드(예: 수년간의 결합된 납세 신고의 평균 공제)를 생성해야 할 수도 있습니다.

데이터 형식화

모델 작성 전의 최종 단계로서, 특정 기법에서 데이터에 특정 형식 또는 순서가 필요한지 여부를 확인하는 것이 유용합니다. 예를 들어, 시퀀스 알고리즘에서 모델 실행 전에 데이터를 예비 정렬해야 하는 경우가 드물지 않습니다. 모델이 정렬을 수행할 수 있더라도, 모델링 전에 정렬 노드를 사용하면 처리 시간을 절약할 수 있습니다.

태스크 목록

데이터를 형식화할 때 다음과 같은 질문을 고려하십시오.

- 어느 모델을 사용할 계획입니까?
- 이러한 모델에서는 특정 데이터 형식 또는 순서가 필요합니까?

변경이 권장되는 경우, IBM SPSS Modeler의 처리 도구가 필요한 데이터 조작을 적용하도록 도와줄 수 있습니다.

모델링 준비 여부

IBM SPSS Modeler에서 모델 빌드 전에 다음 질문에 대한 응답을 해야 합니다.

- IBM SPSS Modeler 내에서 모든 데이터가 액세스 가능합니까?
- 초기 탐색 및 이해를 기반으로 데이터의 관련 서브세트를 선택할 수 있었습니까?
- 데이터를 효과적으로 정리했거나 구제할 수 없는 항목을 제거했습니까? 의사결정 사항을 최종 보고서에서 문서화하십시오.
- 다중 데이터 세트가 적절하게 통합되었습니까? 문서화해야 하는 병합 문제점이 있었습니까?
- 사용할 계획인 모델링 도구의 요구사항을 조사했습니까?
- 모델링 전에 해결 가능한 형식화 문제가 있습니까? 여기에는 요구되는 형식화 문제와 함께 모델링 시간을 줄일 수 있는 태스크가 포함됩니다.

위의 질문에 응답할 수 있다면 데이터 마이닝의 핵심인 모델링의 준비가 된 것입니다.

제 5 장 모델링

모델링 개요

이것은 사용자의 힘든 작업이 성과를 거두기 시작하는 지점입니다. 준비하느라 노력한 데이터가 IBM SPSS Modeler의 분석 도구에 입력되고, 해당 결과가 비즈니스 이해 단계 동안 제기된 비즈니스 문제점에 대한 해결의 실마리를 제공하기 시작합니다.

일반적으로 모델링은 여러 번 반복해서 수행됩니다. 일반적으로 데이터 마이너는 기본 매개변수를 사용하여 몇 가지 모델을 실행한 후 모수를 미세 조정하거나, 선택한 모델에서 요구되는 조작을 위해 데이터 준비 단계로 되돌아갑니다. 단일 모델과 단일 실행으로 조직의 데이터 마이닝 질문이 만족스럽게 응답되는 경우는 드뭅니다. 이로 인해 데이터 마이닝이 매우 흥미로워집니다. 지정된 문제점을 살펴보는 방법은 여러 가지가 있으며 IBM SPSS Modeler는 이 작업을 도울 수 있는 매우 다양한 도구를 제공합니다.

모델링 기법 선택

어느 유형의 모델링이 조직의 요구사항에 가장 적절한지에 대한 아이디어가 이미 있을 수 있지만 이제 어느 모델링을 사용할 것인지에 대해 확고한 결정을 할 때입니다. 가장 적절한 모델을 결정하는 것은 일반적으로 다음 고려사항을 기반으로 할 것입니다.

- **마이닝에 사용 가능한 데이터 유형.** 예를 들어, 관심 필드가 범주형(기호)입니까?
- **사용자의 데이터 마이닝 목적.** 단순히 트랜잭션 데이터 저장소에 대한 통찰력을 얻고 흥미로운 구매 패턴을 알아내고 싶습니까? 아니면 학자금 대출의 연체 성향 등을 나타내는 스코어를 생성해야 합니까?
- **특정 모델링 요구사항.** 모델이 특정 데이터 크기 또는 유형을 요구합니까? 쉽게 발표 가능한 결과가 있는 모델이 필요합니까?

IBM SPSS Modeler의 모델 유형 및 해당 요구사항에 대한 자세한 정보는 IBM SPSS Modeler 문서 또는 온라인 도움말을 참조하십시오.

E-소매 예제--모델링 기법

e-소매업자가 채택하는 모델링 기법은 회사의 데이터 마이닝 목적에 따라 좌우됩니다.

권장사항 개선. 가장 간단한 형태로, 함께 구입하는 횟수가 가장 빈번한 제품을 판별하기 위해 구매 주문을 군집시키는 작업이 포함됩니다. 고객 데이터를 비롯하여 방문 레코드까지도 풍부한 결과를 위해 추가될 수 있습니다. 이단계 또는 코호넨 네트워크 군집 기법이 이 모델링 유형에 적합합니다. 이 후에는, 고객 방문 중 어느 시점에서 가장 적절한 권장사항을 판별하기 위해 C5.0 규칙 세트를 사용하여 군집을 프로파일링할 수 있습니다.

사이트 탐색 개선. 우선, e-소매업자는 자주 사용되지만 사용자가 페이지를 찾기 위해 여러 번 클릭해야 하는 페이지를 식별하는 데 초점을 맞출 것입니다. 여기에는 고객이 웹 사이트에서 선택하는 "고유 경로"를 생성하기 위해 순서 지정 알고리즘을 웹 로그에 적용한 후 수행한 동작 없이(또는 전에) 많은 페이지 방문을 가지는 세션을 구체적으로 찾는 과정이 포함됩니다. 나중에, 더 심층적인 분석에서 군집 기법을 사용하여 다른 "유형"의 방문 및 방문자를 식별할 수 있고 유형에 따라 사이트 컨텐츠를 조직하고 표시할 수 있습니다.

올바른 모델링 기법 선택

IBM SPSS Modeler에서 여러 모델링 기법을 사용할 수 있습니다. 데이터 마이너는 둘 이상의 기법을 사용하여 여러 방향에서 문제점에 접근할 수 있는 경우가 빈번합니다.

태스크 목록

어떤 모델을 사용할지 결정할 때는 다음과 같은 문제가 선택에 영향을 미치는지 여부를 고려하십시오.

- 모델에서 데이터를 검정 및 학습 세트로 분할해야 합니까?
- 지정된 모델에 대해 신뢰성 있는 결과를 생성하기에 충분한 데이터가 있습니까?

- 모델이 특정 수준의 데이터 품질을 요구합니까? 현재 데이터로 이 수준을 충족할 수 있습니까?
- 데이터가 특정 모델에 적절한 유형입니까? 그렇지 않은 경우, 데이터 조작 노드를 사용하여 필요한 변환을 수행할 수 있습니까?

IBM SPSS Modeler의 모델 유형 및 해당 요구사항에 대한 자세한 정보는 IBM SPSS Modeler 문서 또는 온라인 도움말을 참조하십시오.

모델링 가정

선택할 모델링 도구를 좁혀가기 시작하면서 의사결정 프로세스에 대한 설명을 작성하십시오. 데이터 가정뿐 아니라 모델의 요구사항 충족을 위해 이루어진 데이터 조작에 대해 문서화하십시오.

예를 들어, 로지스틱 회귀분석 노드와 신경망 노드는 실행 전에 데이터 유형을 완전히 **인스턴스화**(데이터 유형이 알려짐)해야 합니다. 즉, 모델 빌드 및 실행 전에 유형 노드를 스트림에 추가하고 이를 실행하여 데이터를 시연해야 합니다. 마찬가지로, 예측 모형(예: C5.0)은 희귀 이벤트에 대한 규칙을 예측할 때 데이터를 재조정하는 것이 좋습니다. 이런 유형의 예측을 할 때, 균형 노드를 스트림에 삽입하고 더 균형 잡힌 서브세트를 모델에 공급하면 종종 더 좋은 결과를 얻을 수 있습니다.

이런 유형의 의사결정을 문서화하십시오.

테스트 설계 생성

모델을 실제로 작성하기 전의 최종 단계로서, 모델의 결과를 어떻게 테스트할지에 대해 다시 고려해야 합니다. 두 가지 부분에서 포괄적 테스트 설계를 생성합니다.

- 모델의 "우수성"에 대한 위한 기준 설명
- 이러한 기준을 테스트하기 위한 데이터 정의

모델의 **우수성**은 여러 방법으로 측정될 수 있습니다. 감독된 모델(예: C5.0 및 C&R 트리)의 경우, 우수성의 측정은 일반적으로 특정 모델의 오차율을 추정합니다. 감독되지 않은 모델(예: 코호넨 군집 넷)의 경우, 측정은 해석 또는 배포의 용이성이나 필요한 처리 시간 등의 기준을 포함할 수 있습니다.

모델 작성은 반복적 프로세스라는 것을 기억하십시오. 즉, 사용 및 배포할 모델을 결정하기 전에 일반적으로 여러 모델의 결과를 테스트하게 됩니다.

테스트 설계 작성

테스트 설계는 생성된 모델을 테스트하기 위해 수행할 단계에 대한 설명입니다. 모델링은 반복적 프로세스이기 때문에, 언제 모수 조정을 중지하고 다른 방법 또는 모델을 시도할지를 아는 것이 중요합니다.

태스크 목록

테스트 설계를 작성할 때는 다음과 같은 질문을 고려하십시오.

- 모델을 테스트하기 위해 어떤 데이터가 사용됩니까? 데이터를 학습/검정 세트로 파티션 분할했습니까? (이는 일반적으로 사용되는 모델링 접근법입니다.)
- 감독된 모델(예: C5.0)의 성공을 어떻게 측정할 수 있습니까?
- 감독되지 않은 모델(예: 코호넨 군집 넷)의 성공을 어떻게 측정할 수 있습니까?
- 다른 모델 유형을 시도하기 전에, 몇 회까지 설정을 조정하여 모델을 재실행하시겠습니까?

E-소매 예제--테스트 설계

CRISP-DM을 사용하는 웹 마이닝 시나리오

모델이 평가되는 기준은 고려 중인 모델 및 데이터 마이닝 목적에 따라 달라집니다.

권장사항 개선. 개선된 권장사항이 현재 고객에게 제시될 때까지는 이를 평가할 객관적 방법이 없습니다. 그러나 e-소매업자에게는 비즈니스 관점에서 타당하도록 충분히 단순한 권장사항을 생성하는 규칙이 필요할 수 있습니다. 마찬가지로, 규칙은 서로 다른 고객 및 세션에 대해 서로 다른 권장사항을 생성할 만큼 충분히 복잡해야 합니다.

사이트 탐색 개선. 고객이 웹 사이트의 어느 페이지에 액세스하는지에 대한 증거가 주어지면 e-소매업자는 중요한 페이지에 대한 액세스 용이성의 관점에서 업데이트된 사이트 디자인을 객관적으로 평가할 수 있습니다. 그러나 권장사항과 마찬가지로, 재구성된 사이트에 고객이 얼마나 잘 적응할 것인지를 미리 평가하기는 어렵습니다. 시간 및 재정이 허락할 경우, 일부 유용성 테스트는 적절할 수 있습니다.

모델 작성

이 시점에서는 오랫동안 고려해 온 모델의 작성 준비가 잘 되어 있어야 합니다. 최종 결론을 내리기 전에 여러 가지 모델을 사용하여 실험할 시간 및 공간을 확보하십시오. 일반적으로 대부분의 데이터 마이너는 여러 모델을 작성해서, 모델을 배포하거나 통합하기 전에 결과를 비교합니다.

다양한 모델을 사용한 진행상황을 추적하기 위해, 각 모델에 사용된 설정 및 데이터에 대한 설명을 기록해 두십시오. 이렇게 하면 결과를 다른 사람들과 논의하고 필요한 경우 해당 단계를 재추적할 수 있습니다. 모델 작성 프로세스가 완료되면 데이터 마이닝 의사결정에 사용할 세 가지 정보가 준비될 것입니다.

- **모수 설정**(최상의 결과를 생성할 모수에 대해 사용자가 작성하는 설명이 포함되어 있음)
- 생성된 실제 **모델**
- **모델 결과에 대한 설명**(모델의 실행 및 해당 결과의 탐색 동안 발생한 성능 및 데이터 문제 포함)

E-소매 예제--모델 작성

CRISP-DM을 사용하는 웹 마이닝 시나리오

권장사항 개선. 단순한 구매 데이터베이스부터 시작해서 관련 고객 및 세션 정보를 포함하는 데이터 통합의 다양한 수준에 대해 군집이 생성됩니다. 통합의 각 수준마다, 이단계 및 코호넨 네트워크 알고리즘에 대한 다양한 모수 설정에 따라 군집이 생성됩니다. 이러한 각각의 군집마다, 몇 개의 C5.0 규칙 세트가 서로 다른 모수 설정으로 생성됩니다.

사이트 탐색 개선. 시퀀스 모델링 노드는 고객 경로를 생성하는 데 사용됩니다. 알고리즘을 통해 최소 지원 기준을 지정할 수 있으며, 이는 가장 일반적인 고객 경로에 초점을 맞추는 데 유용합니다. 모수에 대한 다양한 설정이 시도됩니다.

모수 설정

대부분의 모델링 기법은 모델링 프로세스를 제어하기 위해 조정할 수 있는 다양한 모수 또는 설정을 가지고 있습니다. 예를 들어, 의사결정 트리는 트리 깊이, 분할 및 기타 여러 설정을 조정하여 제어할 수 있습니다. 일반적으로 대부분의 사람들은 기본 옵션을 우선 사용하여 모델을 작성한 후 후속 세션에서 모수를 세분화합니다.

가장 정확한 결과를 생성하는 모수를 판별했으면 스트림 및 생성된 모델 노드를 저장하십시오. 또한 최적 설정에 대한 설명을 작성해 두면 새 데이터로 모델을 자동화하거나 다시 작성할 때 도움이 될 수 있습니다.

모델 실행

IBM SPSS Modeler에서 모델 실행은 간단한 작업입니다. 모델 노드를 스트림에 삽입하고 모수를 편집했으면 간단히 모델을 실행하여 조회 가능 결과를 생성합니다. 결과는 작업공간의 오른쪽에 있는 '생성된 모델' 네비게이터에 나타납니다. 모델을 마우스 오른쪽 단추로 클릭하여 결과를 찾아볼 수 있습니다. 대부분의 모델에서, 생성된 모델을 스트림에 삽입하여 결과를 추가적으로 평가하고 배포할 수 있습니다. 또한 모델은 쉽게 재사용할 수 있도록 IBM SPSS Modeler에 저장할 수 있습니다.

모델 설명

모델의 결과를 검사할 때 모델링 경험에 대한 설명을 작성해야 합니다. 노드 주석 대화 상자 또는 프로젝트 도구를 사용하여 모델 자체에 설명을 저장할 수 있습니다.

태스크 목록

각 모델에 대해 다음과 같은 정보를 기록하십시오.

- 이 모델로부터 의미 있는 결론을 끌어낼 수 있습니까?
- 이 모델에 의해 새로운 통찰력 또는 특별한 패턴이 밝혀졌습니까?

- 이 모델에 대한 실행 문제점이 있었습니까? 처리 시간은 얼마나 합리적이었습니까?
- 이 모델은 데이터 품질 문제(예: 결측값 수가 많음)로 어려움이 있었습니까?
- 주목해야 하는 계산 불일치도 있었습니까?

모델 평가

이제 일련의 초기 모델을 가지고 있으므로, 이들을 자세히 살펴보고 어느 모델이 최종이 될 만큼 정확하거나 효과적인지 판별하십시오. 최종이란 "배포될 준비가 된" 또는 "흥미로운 패턴을 보여주는"과 같은 여러 의미를 가질 수 있습니다. 이전에 작성한 테스트 계획을 컨설팅하면 조직의 관점에서 이 평가를 수행하는 데 도움이 될 수 있습니다.

포괄적 모델 평가

고려 중인 각 모델에 대해, 테스트 계획에서 생성된 기준을 기반으로 조직적 평가를 수행하는 것이 좋습니다. 여기서는 생성된 모델을 스트림에 추가하고 평가 차트 또는 분석 노드를 사용하여 결과의 유효성을 분석할 수 있습니다. 또한 결과가 논리적으로 합당한지 또는 사용자의 비즈니스 목적과 관련하여 너무 단순한지(예: 와인 > 와인 > 와인과 같은 구매 시퀀스) 고려해야 합니다.

평가를 수행했으면 객관적(모델 정확도) 및 주관적(사용의 용이성 또는 결과의 해석) 기준 모두를 기반으로 모델 순위를 순서대로 지정합니다.

태스크 목록

- IBM SPSS Modeler의 데이터 마이닝 도구(예: 평가 차트, 분석 노드 또는 교차 검증 차트)를 사용하여 모델의 결과를 평가합니다.
- 비즈니스 문제점의 이해를 기반으로 결과의 검토를 수행합니다. 특정 결과의 관련성을 간파할 수 있는 데이터 분석가 또는 기타 전문가와 상의합니다.
- 모델의 결과를 쉽게 배포할 수 있는지 여부를 고려합니다. 결과를 인터넷에서 배포할지 또는 데이터 웨어하우스로 돌려보낼지에 대해 해당 조직과 상의합니다.
- 결과가 성공 기준에 미치는 영향을 분석합니다. 비즈니스 이해 단계 중에 설정된 목적을 충족합니까?

위의 문제를 성공적으로 해결할 수 있었으며 현재 모델이 관련 목적을 충족한다고 판단되면 이제 모델의 보다 철저한 평가와 최종 배포를 진행할 때입니다. 그렇지 않다면 지금까지 배운 것을 바탕으로 모수 설정을 조정하여 모델을 재실행합니다.

E-소매 예제--모델 평가

CRISP-DM을 사용하는 웹 마이닝 시나리오

권장사항 개선. 코호넨 네트워크 중 하나와 이단계 군집은 각각 적절한 결과를 생성하므로 e-소매업자는 이들 중에 선택하기가 어렵다는 것을 발견합니다. 조만간 회사에서는 두 기법을 모두 사용하여 두 기법이 공통적으로 제시하는 권장사항을 수용하고 두 기법에서 차이가 나는 상황을 보다 면밀하게 연구하기를 희망합니다. 적은 노력과 적용된 비즈니스 지식을 통해, e-소매업자는 두 가지 기법 사이의 차이를 해소하기 위한 추가 규칙을 개발할 수 있습니다.

또한 e-소매업자는 세션 정보를 포함하는 결과가 놀랍게도 좋다는 것을 발견합니다. 권장사항을 사이트 탐색에 연결할 수 있다는 것을 암시하는 증거가 있습니다. 고객이 다음에 어디로 갈 수 있을지를 정의하는 규칙 집합을 실시간으로 사용하여 고객이 브라우징 중일 때 사이트 컨텐츠에 직접 영향을 미칠 수 있습니다.

사이트 탐색 개선. 시퀀스 모델은 특정 고객 경로를 예측할 수 있다는 높은 수준의 자신감을 e-소매업자에게 심어주고 사이트 계획에 대해 관리 가능한 수의 변경사항을 제시하는 결과를 생성합니다.

수정된 모수 추적

모델 평가 중에 배웠던 내용을 기반으로, 모델을 다른 방식으로 살펴볼 때입니다. 여기서는 두 가지 옵션이 있습니다.

- 기존 모델의 모수를 조정합니다.
- 다른 모델을 선택하여 데이터 마이닝 문제점을 해결합니다.

두 경우 모두, 모델 작성 태스크로 돌아가서 결과가 성공적이 될 때까지 반복됩니다. 이 단계의 반복에 대해 염려하지 마십시오. 데이터 마이너가 해당 요구사항을 충족하는 모델을 찾기 전에 모델을 여러 번 평가하고 재실행하는 것은 매우 흔한 일입니다. 이것은 동시에 여러 모델을 작성하고 결과를 비교한 후 각각의 모수를 조정하기 위한 좋은 논거입니다.

다음 단계에 대한 준비 여부

모델의 최종 평가를 진행하기 전에 초기 평가가 충분히 철저했는지 여부를 고려하십시오.

태스크 목록

- 모델의 결과를 이해할 수 있습니까?
- 순수하게 논리적 관점에서 모델 결과가 합당합니까? 추가 탐색을 필요로 하는 분명한 불일치가 있습니까?
- 언뜻 보기에, 해당 결과가 조직의 비즈니스 문제를 해결할 수 있겠습니까?
- 모델 정확도를 비교하고 평가하기 위해 분석 노드와 리프트 또는 Gains 차트를 사용했습니까?
- 둘 이상의 모델 유형을 탐색하고 결과를 비교했습니까?
- 모델의 결과가 배포 가능합니까?

데이터 모델링의 결과가 정확하고 관련성이 있다고 판단되면 최종 배포 전에 더 철저한 평가를 수행할 때입니다.

제 6 장 평가

평가 개요

이 시점에서는 대부분의 데이터 마이닝 프로젝트를 완료했습니다. 또한 작성된 모델이 이전에 정의한 **데이터 마이닝 성공 기준**에 따라 기술적으로 올바르고 유효한지를 모델링 단계에서 판별했습니다.

그러나 계속하기 전에 프로젝트 초기에 설정된 **비즈니스 성공 기준**을 사용하여 작업의 결과를 평가해야 합니다. 이는 획득한 결과를 조직에서 충분히 이용할 수 있게 하기 위한 열쇠입니다. 두 가지 결과 유형이 데이터 마이닝에 의해 생성됩니다.

- CRISP-DM의 이전 단계에서 선택된 최종 **모델**
- 모델 자체에서뿐만 아니라 데이터 마이닝 프로세스에서 도출한 결론 또는 추론. 이들을 **결과물**이라고 합니다.

결과 평가

이 단계에서는 프로젝트 결과가 비즈니스 성공 기준을 충족하는지 여부의 평가 내용을 정식화합니다. 이 단계에서는 명시된 비즈니스 목적의 분명한 이해가 필요하므로 핵심 의사결정자를 프로젝트 평가에 포함시키십시오.

태스크 목록

먼저, 데이터 마이닝 결과가 비즈니스 성공 기준을 충족하는지 여부의 평가 내용을 문서화해야 합니다. 보고서에서 다음과 같은 질문을 고려하십시오.

- 결과가 분명하고 쉽게 발표할 수 있는 양식으로 명시되어 있습니까?
- 강조표시해야 하는 특별히 기발하거나 독특한 결과물이 있습니까?
- 비즈니스 목적에 적용 가능한 순서대로 모델 및 결과물의 순위를 지정할 수 있습니까?
- 일반적으로, 이러한 결과가 조직의 비즈니스 목적에 얼마나 잘 부응합니까?
- 결과로 인해 어떤 추가 질문이 발생했습니까? 이러한 질문을 비즈니스 용어로 어떻게 표현할 수 있습니까?

결과를 평가한 후, 최종 보고서에 포함할 승인된 모델 목록을 컴파일하십시오. 이 목록은 데이터 마이닝과 조직의 비즈니스 목적 모두를 충족하는 모델을 포함해야 합니다.

E-소매 예제--결과 평가

CRISP-DM을 사용하는 웹 마이닝 시나리오

e-소매업자가 수행하는 첫 번째 데이터 마이닝의 전체 결과는 비즈니스 관점에서 의사 전달이 다소 용이합니다. 연구에서는 더 나은 제품이 되기 위해 바라는 권장사항과 사이트 디자인 개선사항을 산출했습니다. 사이트 디자인 개선사항은 고객의 브라우징 순서를 기반으로 하며, 이는 고객이 원하지만 도달하기 위해 몇 가지 단계가 필요한 사이트 기능을 보여줍니다. 제품 권장사항이 더 나은지에 대한 증거는 의사결정 규칙이 복잡해질 수 있기 때문에 전달하기가 더 어렵습니다. 최종 보고서를 생성하기 위해, 분석가는 더 쉽게 설명할 수 있는 규칙 세트에서 몇 가지 일반적 추세를 식별하려고 할 것입니다.

모델 순위화. 몇 가지 초기 모델은 비즈니스 타당성이 있다고 보였으므로 해당 그룹 내의 순위는 통계 기준, 해석의 용이성 및 다양성을 기반으로 했습니다. 그러므로 모델은 서로 다른 상황에 대해 서로 다른 권장사항을 제공했습니다.

새로운 질문. 연구에서 도출할 가장 중요한 질문은 "어떻게 e-소매업자가 해당 고객에 대해 더 많이 알 수 있을까"에 대한 질문입니다. 고객 데이터베이스의 정보는 권장사항에 대한 군집을 형성함에 있어 중요한 역할을 합니다. 특별 규칙은 해당 정보가 손실된 고객에게 권장사항을 제공하기 위해 사용 가능한 반면, 권장사항은 등록된 고객에게 제공될 수 있는 특별 규칙보다 속성상 더 일반적입니다.

프로세스 검토

일반적으로 효과적 방법론은 방금 완료된 프로세스의 성공 및 약점에 대해 숙고하기 위한 시간을 포함합니다. 데이터 마이닝은 다르지 않습니다. CRISP-DM의 일부는 이후의 데이터 마이닝 프로젝트가 보다 효과적이 되도록 사용자의 경험으로부터 배우고 있습니다.

태스크 목록

먼저 데이터 준비 단계, 모델 작성 등의 각 단계에 대한 활동 및 의사결정을 요약해야 합니다. 그런 다음 각 단계에 대해 다음과 같은 질문을 고려하고 개선을 위한 제안을 해야 합니다.

- 이 단계는 최종 결과의 가치에 기여했습니까?
- 이 특정 단계 또는 작업을 합리화하거나 개선할 방법이 있습니까?
- 이 단계의 실패 또는 실수가 무엇이었습니까? 다음 번에는 어떻게 회피할 수 있습니까?
- 결실이 없는 특정 모델과 같은 교착 상황이 있었습니까? 작업을 보다 생산적으로 진행할 수 있도록 이러한 교착 상황을 예측할 방법이 있습니까?
- 이 단계에서 의외의 (좋은 또는 나쁜) 상황이 발생했습니까? 되돌아 봤을 때, 이러한 상황을 예측할 수 있는 분명한 방법이 있습니까?
- 해당 단계에서 사용되었을 수 있는 대체 의사결정 또는 처리 방법이 있습니까? 이후의 데이터 마이닝 프로젝트를 위해 해당 대안을 기록하십시오.

E-소매 예제--검토 보고서

CRISP-DM을 사용하는 웹 마이닝 시나리오

초기 데이터 마이닝 프로젝트의 프로세스를 검토함으로써, e-소매업자는 프로세스 단계 사이의 상호관계를 잘 이해하게 되었습니다. 처음에는 CRISP-DM 프로세스를 역추적하는 것이 내키지 않았던 e-소매업자는 이제 프로세스의 순환성이 그 힘을 증강시킨다는 사실을 이해합니다. 또한 프로세스 검토는 e-소매업자가 다음과 같은 사실을 이해하도록 도왔습니다.

- 특별한 상황이 CRISP-DM 프로세스의 다른 단계에서 나타나는 경우 탐색 프로세스로의 회귀가 항상 보장됩니다.
- 데이터 준비(특히 웹 로그의 경우)는 매우 오랜 시간이 걸릴 수 있으므로 인내심이 필요합니다.
- 일단 데이터의 분석 준비가 되면 더 큰 그림과 상관없이 모델 구축을 시작하기가 너무 쉽기 때문에 당면한 비즈니스 문제점에 계속 초점을 맞추는 것이 중요합니다.
- 모델링 단계가 끝나면, 결과 구현 방식을 결정하고 어떠한 추가 연구가 보장되는지 판별함에 있어서 비즈니스 이해가 훨씬 더 중요합니다.

다음 단계 결정

지금까지, 결과를 생성했고 데이터 마이닝 경험을 평가했는데 **다음은 어디로?**라고 궁금해 할지도 모릅니다. 이 단계는 데이터 마이닝을 위한 비즈니스 목적에 비추어 해당 질문에 응답하도록 돕습니다. 본질적으로, 이 시점에서는 두 가지 선택이 있습니다.

- **배포 단계를 진행합니다.** 다음 단계에서는 모델 결과를 비즈니스 프로세스에 통합하고 최종 보고서를 생성하도록 도울 것입니다. 데이터 마이닝 작업이 성공하지 못했을지라도, CRISP-DM의 배포 단계를 사용하여 프로젝트 스폰서에게 배포할 최종 보고서를 작성해야 합니다.
- **뒤로 돌아가서 모델을 세분화하거나 대체합니다.** 해당 결과가 거의 최적이지만 아주 최적은 아니라면 다른 모델링 라운드를 고려하십시오. 이 단계에서 배운 내용을 사용하여 모델을 세분화하고 더 나은 결과를 생성할 수 있습니다.

이 시점에서의 의사결정은 모델링 결과의 정확도 및 관련성을 고려합니다. 결과가 데이터 마이닝 및 비즈니스 목적에 합당하면 배포 단계 준비가 된 것입니다. 어떤 의사결정을 하더라도 평가 프로세스를 철저하게 문서화하십시오.

E-소매 예제--다음 단계

CRISP-DM을 사용하는 웹 마이닝 시나리오

e-소매업자는 프로젝트 결과의 정확도와 연관성을 상당히 신뢰하므로 배포 단계를 진행합니다.

동시에, 프로젝트 팀은 뒤로 돌아가서 예측 기술을 포함하도록 일부 모델을 보완할 준비가 되어 있습니다. 이 시점에서 그들은 최종 보고서의 전달과 의사결정자의 승인을 기다리고 있습니다.

제 7 장 배포

배포 개요

배포는 새로운 통찰력을 사용하여 조직 내에서 개선을 수행하는 프로세스입니다. 이는 정식 통합(예: 데이터 웨어하우스에 읽혀질 이탈 스코어를 생성하는 IBM SPSS Modeler 모델의 구현)을 의미할 수 있습니다. 또는 배포가 데이터 마이닝에서 얻은 통찰력을 사용하여 조직 내에서 변화를 이끌어내는 것을 의미할 수도 있습니다. 예를 들어, 데이터에서 30살이 넘은 고객의 행동 변화를 나타내는 놀라운 패턴을 발견했을 수 있습니다. 이러한 결과는 정보 시스템에 정식으로 통합되지 않을지도 모르지만 확실히 계획 및 마케팅 의사결정에 유용할 것입니다.

일반적으로, CRISP-DM의 배포 단계는 두 가지 유형의 활동을 포함합니다.

- 결과 배포의 계획 및 모니터링
- 결론 태스크(예: 최종 보고서 생성 및 프로젝트 검토 수행) 완료

조직의 요구사항에 따라, 이러한 단계 중 하나 또는 모두를 완료해야 할 수도 있습니다.

배포 계획

데이터 마이닝 작업의 과실을 공유하고자 조급할지라도, 원활하고 포괄적인 결과 배포를 계획할 시간을 가지십시오.

태스크 목록

- 첫 번째 단계는 결과(모델과 결과물)을 요약하는 것입니다. 이렇게 하면 어느 모델이 데이터베이스 시스템 내에 통합될 수 있으며 어느 결과물을 동료에게 발표할지 판별할 수 있습니다.
- 배포 가능한 각 모델에 대해, 사용자 시스템에서의 배포 및 통합을 위한 단계별 계획을 작성하십시오. 모델 출력에 대한 데이터베이스 요구사항 등의 기술적 세부사항을 참고하십시오. 예를 들어, 사용자 시스템에서는 모델링 출력을 탭 구분 데이터 형식으로 배포해야 할 수도 있습니다.
- 결론적인 각 결과물에 대해, 이 정보를 전략결정자에게 전달하기 위한 계획을 작성하십시오.
- 언급할 가치가 있는 두 결과 유형 모두에 대한 대체 배포 계획이 있습니까?
- 배포가 모니터링되는 방식을 고려하십시오. 예를 들어, IBM SPSS Modeler Solution Publisher를 사용하여 배포된 모델을 어떻게 업데이트합니까? 모델을 더 이상 적용하지 못하는 시기를 어떻게 결정하시겠습니까?
- 배포 문제점을 식별하고 비상 계획을 마련하십시오. 예를 들어, 의사결정자는 모델링 결과에 대한 자세한 정보를 원할 수 있고 더 많은 기술적 세부사항을 제공하도록 요구할 수 있습니다.

E-소매 예제--배포 계획

CRISP-DM을 사용하는 웹 마이닝 시나리오

e-소매업자의 데이터 마이닝 결과가 성공적으로 배포되려면 올바른 정보가 올바른 사람에게 전달되어야 합니다.

의사결정자. 의사결정자는 사이트에 대한 권장사항 및 변경 제안을 통지받아야 하고 이러한 변경이 어떻게 도움이 될지에 대한 간단한 설명을 제공받아야 합니다. 의사결정자가 해당 연구 결과를 승인할 경우, 변경을 구현할 사람들에게 통지되어야 합니다.

웹 개발자. 웹 사이트를 유지보수하는 사람들은 새 권장사항과 사이트 컨텐츠의 조직을 통합해야 할 것입니다. 미래 연구로 인해 어떤 변화가 발생할 수 있는지 알려서 지금 기초를 놓을 수 있게 하십시오. 팀에게 실시간 시퀀스 분석을 기반으로 한 동적 사이트 구축을 준비시키면 나중에 도움이 될 수 있습니다.

데이터베이스 전문가. 고객, 구매 및 제품 데이터베이스를 유지보수하는 사람들은 데이터베이스의 정보가 어떻게 사용되고 있으며 어떤 속성이 이후 프로젝트에 추가될 수 있는지에 대해 지속적인 통지를 받아야 합니다.

무엇보다도 프로젝트 팀은 결과의 배포 및 이후 프로젝트의 계획을 조정하기 위해 각각의 해당 그룹과 지속적으로 의견을 나누어야 합니다.

모니터링 및 유지보수 계획

모델링 결과의 완전한 배포 및 통합에서, 데이터 마이닝 작업은 진행 중일 수 있습니다. 예를 들어, 장바구니 구매의 순서를 예측하기 위해 모델이 배포된 경우 이 모델은 해당 유효성을 보장하고 지속적인 개선을 수행하기 위해 정기적으로 평가할 필요가 있습니다. 마찬가지로, 우수 고객 중에 고객 유지를 증가시키기 위해 배포된 모델은 특정 유지 수준에 도달하게 되면 개조할 필요가 있습니다. 그런 다음 이 모델을 수정 후 재사용하여 더 낮은 수준이지만 가치 피라미드에서 여전히 수익성 있는 수준에서 고객을 유지할 수 있습니다.

태스크 목록

다음 문제에 대한 설명을 작성하고 최종 보고서에 이를 포함시키십시오.

- 각 모델 또는 결과물에 대해, 어느 요인 또는 영향력(예: 시장 가치 또는 계절적 변화)을 추적해야 합니까?
- 각 모델의 타당성 및 정확도를 어떻게 측정하고 모니터링할 수 있습니까?
- 모델이 "만료"되는 시기를 어떻게 판별할 수 있습니까? 정확도 임계값 또는 예상되는 데이터 변화 등에 대한 세부적인 사항을 제공합니다.
- 모델이 만료되면 어떤 상황이 발생할 것입니까? 단순히 최신 데이터로 모델을 다시 빌드하거나 약간의 조정을 할 수 있습니까? 아니면 새로운 데이터 마이닝 프로젝트가 필요할 만큼 변경사항이 광범위하겠습니까?
- 이 모델이 만료된 후 유사한 비즈니스 문제에 이 모델을 사용할 수 있습니까? 각 데이터 마이닝 프로젝트에 대한 비즈니스 목적을 평가하기 위해 충실한 문서화가 중요한 이유가 바로 이것입니다.

E-소매 예제--모니터링 및 유지보수

CRISP-DM을 사용하는 웹 마이닝 시나리오

모니터링의 당면 과제는 새 사이트 조직과 개선된 권장사항이 실제로 효과가 있는지를 판별하는 것입니다. 즉, 사용자가 그들이 찾는 페이지에 더 직접적으로 이동할 수 있습니까? 권장 품목의 교차 판매가 증가했습니까? 모니터링의 몇 주 후에, e-소매업자는 연구의 성공을 판별할 수 있습니다.

자동으로 처리될 수 있는 것은 새로 등록된 사용자의 포함입니다. 고객이 사이트에 등록하면 현재 규칙 세트가 고객의 정보에 적용되어 어떤 권장사항을 제공해야 하는지 결정할 수 있습니다.

권장사항을 결정하는 규칙 세트를 언제 업데이트할지를 결정하는 것은 보다 까다로운 작업입니다. 군집 작성을 위해서는 해당 군집 솔루션의 적절성에 관한 사용자 입력이 필요하기 때문에 규칙 세트의 업데이트는 자동 프로세스가 아닙니다.

이후 프로젝트는 더 복잡한 모델을 생성할 것이므로 모니터링의 필요성과 그 양은 거의 확실히 증가할 것입니다. 가능하면 대부분의 모니터링은 자동이어야 하며 정기적으로 스케줄링된 보고서로 검토 가능해야 합니다. 또는 동적으로 예측을 제공하는 모델의 작성이 회사가 추구하는 방향일 수 있습니다. 이를 위해서는 처음 데이터 마이닝 프로젝트보다 팀의 정교한 작업이 더 필요합니다.

최종 보고서 생성

최종 보고서 작성은 선행 문서의 부족한 면을 보완할 뿐만 아니라 결과를 서로 전달하는 용도로도 사용할 수 있습니다. 당연해 보일 수도 있지만, 결과에 지분을 가진 다양한 사람들에게 결과를 발표하는 것은 중요합니다. 모델링 결과의 구현을 담당하는 기술적 관리자뿐만 아니라 결과를 기반으로 의사결정을 수행할 마케팅 및 관리 스폰서가 관련 대상이 될 수 있습니다.

태스크 목록

먼저, 보고서의 청중을 고려하십시오. 이들은 기술적 개발자 또는 시장 집중 관리자입니까? 해당 요구사항이 서로 다르면 각 청중에 대해 별도의 보고서를 작성해야 할 수도 있습니다. 어느 경우이든, 보고서는 다음 사항을 대부분 포함시켜야 합니다.

- 원래 비즈니스 문제점에 대한 철저한 설명
- 데이터 마이닝을 수행하는 데 사용되는 프로세스
- 프로젝트의 비용
- 원래 프로젝트 계획에서 벗어나는 것에 대한 설명

- 데이터 마이닝 결과(모델 및 결과물)의 요약
- 배포에 대해 제안된 계획의 개요
- 탐색 및 모델링 동안 발견된 흥미로운 리드를 비롯하여 추가 데이터 마이닝 작업에 대한 권장사항

최종 프리젠테이션 준비

프로젝트 보고서 외에, 프로젝트 결과물을 스폰서 또는 관련 부서의 팀에게 발표해야 할 수도 있습니다. 이 경우, 보고서의 정보와 거의 동일한 정보를 사용하지만 더 넓은 관점에서 발표할 수 있습니다. IBM SPSS Modeler의 차트 및 그래프는 이 프리젠테이션 유형을 위해 쉽게 내보낼 수 있습니다.

E-소매 예제--최종 보고서

CRISP-DM을 사용하는 웹 마이닝 시나리오

원래 프로젝트 계획에서 편차가 많이 나는 것은 추가 데이터 마이닝 작업을 위한 흥미로운 리드이기도 합니다. 원래 계획의 목적은 고객이 사이트 방문 시 더 많은 시간을 보내고 더 많은 페이지를 보도록 할 방법을 찾아내는 것이었습니다.

결과적으로, 고객 만족은 단지 고객이 온라인에서 오래 머무르게 하는 문제가 아닙니다. (세션이 구매로 이어졌는지 여부에 따라 분할된) 세션당 소요된 시간의 빈도 분포는 구매로 이어진 대부분의 세션에 대한 세션 시간이 비구매 세션의 두 군집에 대한 세션 시간 사이에 있다는 것을 알아내었습니다.

이것을 알아냈으므로 이제 문제는 구매 없이 사이트에서 오래 머무르는 고객이 단지 아이쇼핑 중인지 또는 찾고 있는 상품이 없는 것인지 알아내는 것입니다. 그 다음 단계는 구매를 촉진하도록 그들이 찾고 있는 상품을 조달할 방법을 알아내는 것입니다.

최종 프로젝트 검토 수행

이것은 CRISP-DM 방법론의 최종 단계이며, 사용자의 최종 인상을 공식화하고 데이터 마이닝 프로세스 동안에 배운 교훈을 대조할 기회를 제공합니다.

태스크 목록

데이터 마이닝 프로세스에 상당히 관여한 사람들과 간단한 인터뷰를 수행해야 합니다. 이러한 인터뷰 동안 고려할 질문은 다음과 같습니다.

- 프로젝트에 대한 전체적인 인상은 어떻습니까?
- 일반적인 데이터 마이닝 및 사용 가능한 데이터와 관련해서 이 프로세스 동안 무엇을 배웠습니까?
- 프로젝트의 어떤 부분이 잘 진행되었습니까? 어디서 어려움이 발생했습니까? 혼란을 완화하는 데 도움이 된 정보가 있었습니까?

데이터 마이닝 결과가 배포된 후, 해당 결과에 의해 영향을 받은 사람들(예: 고객 또는 비즈니스 파트너)과도 인터뷰할 수 있습니다. 여기서의 목적은 프로젝트를 수행할 가치가 있었고 프로젝트를 기획할 때 설정한 혜택이 제공되었는지 판별하는 것이어야 합니다.

이러한 인터뷰의 결과는 데이터 저장소 마이닝 경험에서 배운 교훈에 초점을 맞춰야 하는 최종 보고서에서 프로젝트에 대한 자신의 인상과 함께 요약될 수 있습니다.

E-소매 예제--최종 검토

CRISP-DM을 사용하는 웹 마이닝 시나리오

프로젝트 멤버 인터뷰. 가장 밀접하게 시종일관 연구와 연관된 프로젝트 멤버가 대부분 결과에 열중하고 이후 프로젝트를 고대한다는 것을 e-소매업자는 발견합니다. 데이터베이스 그룹은 조심스럽게 낙관적인 것 같습니다. 이들은 연구의 유용성을 인정하지만 데이터베이스 자원에 대한 추가 부담을 지적합니다. 연구 중에는 컨설턴트가 사용 가능했지만, 상황이 진행되면서 프로젝트 범위가 확장됨에 따라 데이터베이스 유지보수 전담 직원이 필요할 것입니다.

고객 인터뷰. 고객 피드백은 지금까지 주로 긍정적이었습니다. 심사숙고하지 않았던 한 가지 문제는 기존 고객이 사이트 디자인 변경에 대해 어떻게 생각할 것이냐는 것이었습니다. 몇 년 후에, 등록된 고객은 사이트가 조직되는 방식에 대한 특정 기대치를 발전시켰습니다. 등록된 사용자의 피드백은 등록되지 않은 고객의 피드백만큼 그리

긍정적이지 않으며 소수 사람들은 변경을 매우 싫어합니다. e-소매업자는 이 문제를 숙지하고 있어야 하며 변경으로 인해 기존 고객을 잃을 위험을 감수하고 새 고객을 충분히 불러올 수 있을지에 대해 신중히 고려해야 합니다.

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.

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.

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.

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

C

H

가

다

라

마

바

사

아

자

카

타

파

하