



CHAPTER 07.

데이터마이닝





AGENDA

- 01 데이터마이닝의 가치
- 02 데이터마이닝 프로세스



01

데이터마이닝의 가치

기업의 성공요소

• 기업의 성공요소

- 시간이 지날수록 고객의 **기대 상승**, **고객 만족 실현**이 어려워 짐.
- **대체제품**이 많아 짐에 따라 고객 충성심을 얻기 힘들어 짐.
- 시장의 **경쟁**이 심화되며 수익성 좋은 고객을 유지하는 전략이 필요하게 됨.
 ∴ 기업은 성공하기 위해 **고객의 니즈를 미리 예측하는 것이 필요하게 됨!**

• 고객의 니즈 예측

- 방대한 양의 고객, 시장, 제품에 관한 매우 상세한 정보를 수집하고 저장하여, 이를 각기 다른 프로그램을 통해 처리함 → **데이터마이닝**을 통해 인사이트 도출

- **미래의 트렌드와 행동**, 그리고 새로운 **기회**를 만들기 위한 예측
- **전략적 비즈니스**를 수행할 수 있는 능력을 제공
- 올바른 **목표고객**을 선정하고, 비슷한 행동과 니즈에 따라 **고객층**을 분류

데이터마이닝의 필요성

- 데이터마이닝의 필요성

- 1) 성공적인 데이터마이닝 인프라는 **기술, 인간 기술** 및 **기업 경영**과의 긴밀한 통합으로 구성되어 새로운 지식을 **비즈니스 활동 및 가치**로 전환할 수 있음.
- 2) 데이터마이닝 절차를 **표준화**하는 것은 필요한 결과의 품질을 보장하므로 반복적인 절차를 구축해야 함.
- 3) **기업 내부의 지식**을 보다 잘 유지하고 보관할 수 있게 함.
- 4) 신입 사원을 더 빨리 **교육**하는 데에도 중요하게 쓰일 수 있음.

데이터마이닝의 비즈니스 가치

• 데이터마이닝의 비즈니스 가치

- 1) 고객관리의 맥락에서, 데이터마이닝은 고객을 **이해**하고 고객의 **니즈**를 보다 잘 파악 할 수 있도록 도와 줌.
- 2) 과학적으로 **향상된 타겟팅**을 제공하여 마케팅의 창조적인 측면에만 집중하는 것보다 **비용 절감 및 매출 증가**를 얻을 수 있게 함.
- 3) 올바른 **목표고객**을 선정하거나 유사한 행동 및 니즈를 가진 **세분 고객**(이전에는 알려지지 않았던)을 식별하는 데 도움을 줄 수 있음.
- 4) 데이터마이닝 기술을 사용한 훌륭한 타겟 목록은 **구매율**을 높이며 **매출**에도 긍정적인 영향을 미침.

• 데이터마이닝의 적용

- 기업과 비즈니스를 중단할 것 같은 고객을 조기에 식별할 수 있는 예측 모델을 통해 **이탈을 감소**시킴.
- 높은 **성장 잠재력**을 가진 고객을 식별하여 **고객수익성**을 높일 수 있음.
- 보다 선별적인 **타겟팅**을 통해 마케팅 **비용을 감소**시킴.



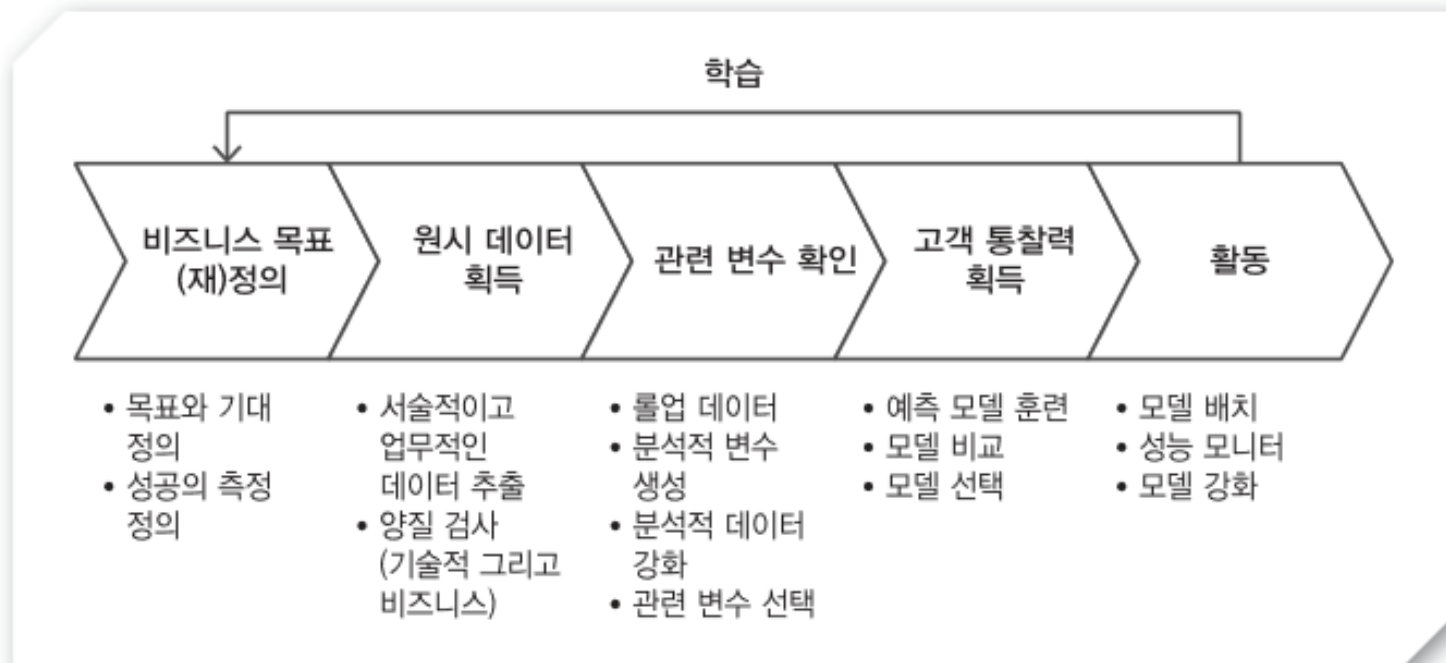
02

데이터마이닝 프로세스

데이터마이닝 프로세스 (1)

• 데이터마이닝 프로세스 개요

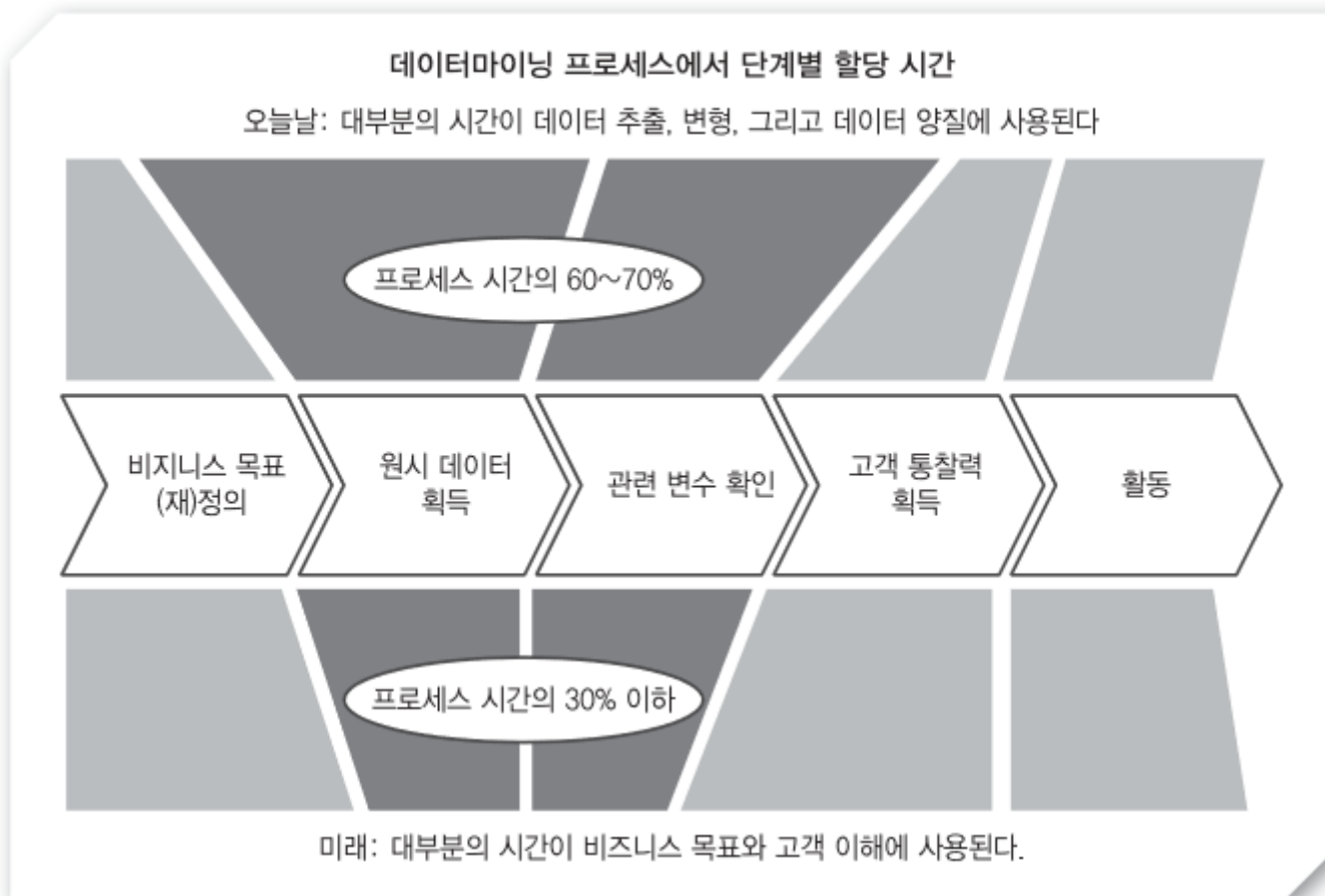
〈그림 7-1〉 데이터마이닝 프로세스 개요



데이터마이닝 프로세스 (2)

• 데이터마이닝 프로세스 단계별 할당 시간

〈그림 7-2〉 데이터마이닝 프로세스에서 단계별 할당하는 시간



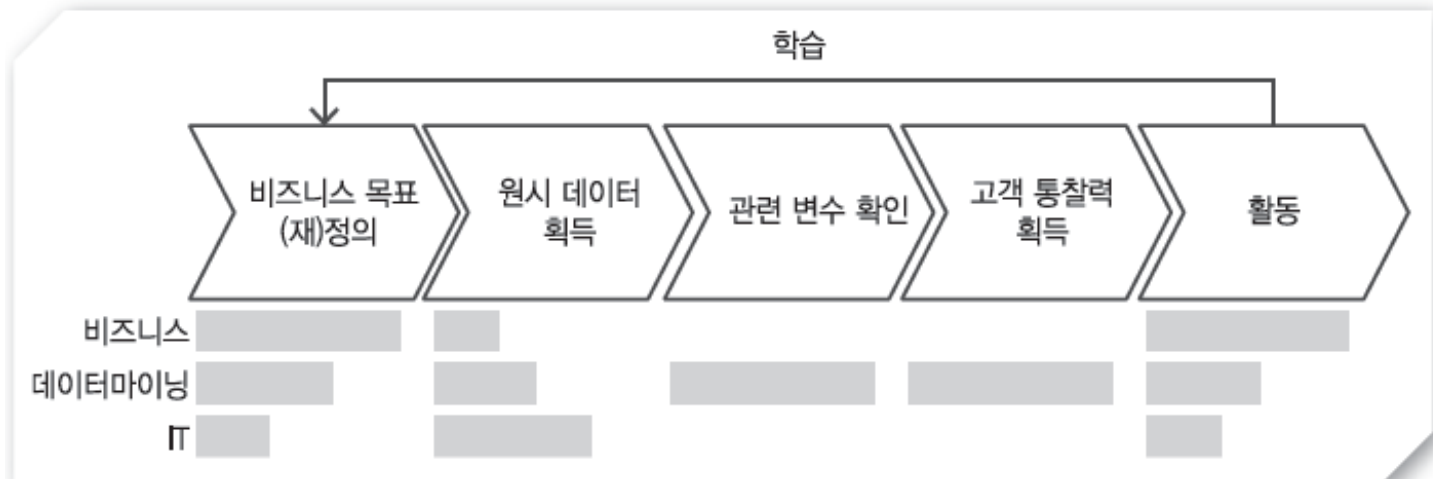
데이터마이닝 프로세스 (3)

- 데이터마이닝 프로세스 단계별 할당 시간

- 데이터마이닝 프로젝트의 많은 사례에서 **데이터 준비 단계**가 전체 프로젝트 기간의 60%에서 70%를 사용함.
 - 이는 **고객 행동을 묘사하는 관련 변수의 비가동률과 관련된 이슈** 때문
 - ex) 고객 중심의 관점이 없는 다른 부서에 의해 관리되는 기존 데이터 자료에는 접근이 어렵기 때문
- 많은 시간이 요구되는 데이터 추출 및 조작, 그리고 데이터 품질 모니터링 및 향상 단계를 **자동화**하는 것이 중요함.
- 데이터 지식을 배치 모드와 같은 실행할 수 있는 프로그램에 **순차적**이고 **체계적**인 코딩을 적용하여 비즈니스 목표의 정확한 정의, 고객 통찰력 추출 및 지식을 기반으로 한 효과적인 행동과 같은 가치 창출 작업에 집중할 수 있어야 함.

자원의 개입

- 데이터마이닝에 관여된 주요 그룹은 비즈니스 그룹(예: 마케팅, 제품 관리), 데이터마이닝 및 IT자원임.
 - **비즈니스 그룹** : 비즈니스 목표를 정의하는 데 관여하며, 기업 활동에 새로운 통찰력을 배치할 때 주도적인 역할을 함.
 - **데이터마이닝 그룹** : 비즈니스 목표를 이해하고 데이터 정제 과정에서 비즈니스 그룹을 지원함. 프로젝트의 범위를 수정하고 사용 가능한 데이터에 의한 제한 사항에 맞추어 기대치를 재조정.
 - **IT 자원** : 모델링에 필요한 필수 데이터의 소싱 및 추출에 필요함.



〈그림 7-3〉 전형적인 데이터마이닝 프로젝트에서 비즈니스, 데이터마이닝, IT 자원의 관여 수준

데이터 조작 (1)

- 사용된 데이터의 **차원 수**는 극적으로 변할 수 있음.
- 단순한 2차원 데이터 테이블에서 **열**을 설명 변수로 **행**을 단일 관찰이라고 가정하면 각각은 동일한 기본 객체(예: 고객 식별 번호, 거래 식별 번호)에 대한 변수 모음에 속함.

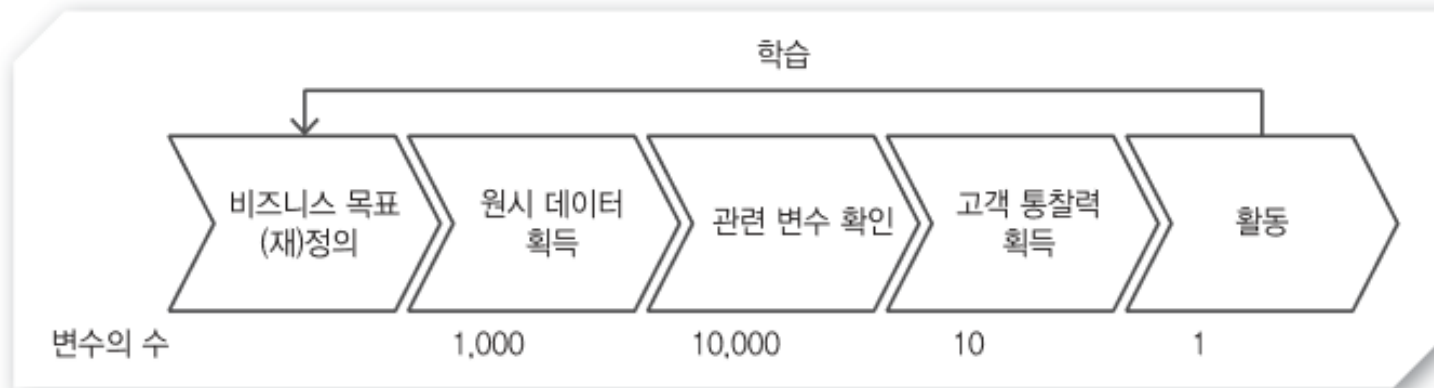
열 조작의 유형

- **변환:** 생년월일을 나이로 변환
- **도출:** 기존에 있던 변수를 바탕으로 새로운 변수 생성
(예: 판매와 비용 정보로부터 월 간 수익 산출)
- **제거:** 가능한 다양한 이유로 전체 변수가 추가 처리에서 제외
(예: 예측에 도움이 되지 않는 변수 또는 이미 모델에 있는 하나 이상의 변수와 상관관계가 있는 변수는 제거할 수 있음)

데이터 조작 (2)

- 사용되는 **변수의 수**는 데이터마이닝 프로세스 중에 크게 바뀔 수 있음.

〈그림 7-4〉 각 프로세스별 변수의 수



데이터 조작 (3)

- 수천만 개의 **행**을 처리한다고 가정한다면, 확장성 및 좋은 샘플링 방법이 모든 데이터마이닝 환경에 필수적임.

행 조작의 유형

- **집합:** 특정 고객, 제품 유형 및 기타 항목에 대해 주어진 기간 동안 특정 유형의 거래 수, 평균 및 표준편차가 포함.
- **변화 탐지:** 특정 변수가 고객 거주지의 우편번호 또는 신용 등급과 같은 값을 변경할 때 탐지하기 위해 사용됨 .
- **결측치 탐지:** 결측치가 감지되면 추가 처리에서 전체 행 제거, 결측치를 상수값으로 바꾸기 또는 결측값의 분포나 다른 데이터 필드와의 상관관계를 기반으로 임의의 값으로 대체하는 방법이 있음.
- **이상치 탐지:** 경우에 따라 관측치에는 극한값을 포함한 변수가 있을 수 있음. 따라서 보다 정교한 형태로 동시에 여러 변수를 살펴보고, 다변수의 이상치는 주의해야 함. 이상치는 다른 값 또는 대응하는 행에 매핑 될 수 있으며 이후 처리에서 제외될 수 있음.

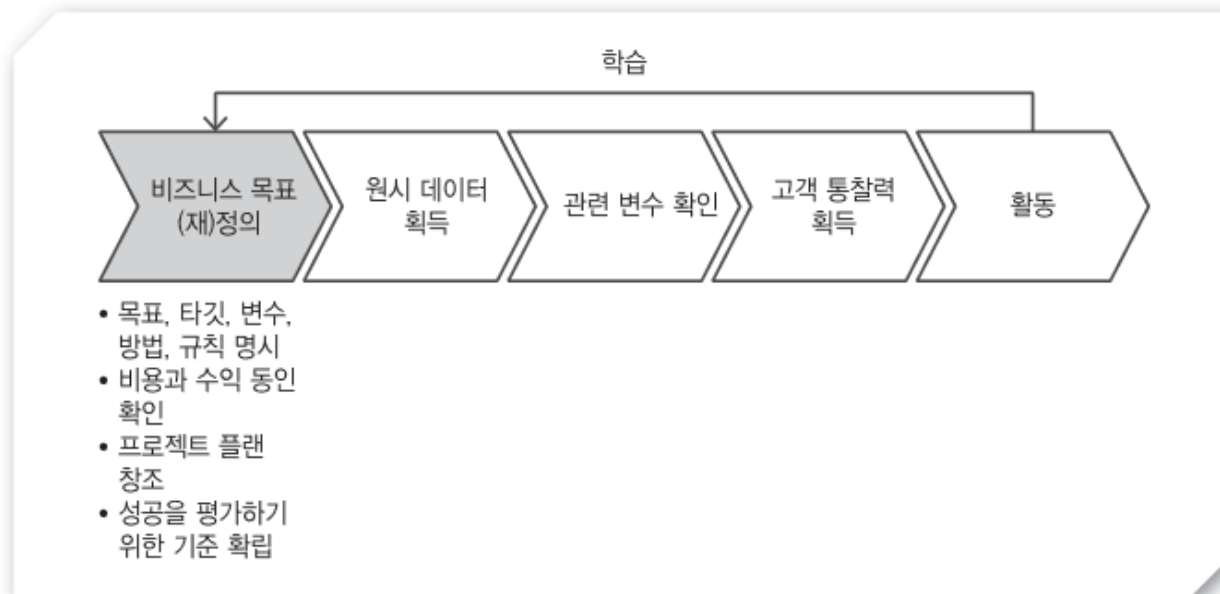
데이터 조작 (4)

- 새로운 데이터를 **샘플링** 하고 각기 다른 목적을 위해 다양한 스트림으로 **분할**하는 것이 일반적임.
 - **학습용**: 모델을 구축하기 위해 사용
 - **검증용**: 모델 구축에 사용된 표본 외의 테스트 및 최종 모델 후보 선택에 사용
 - **스코어링 데이터**: 모델 기반 예측에 사용됨. 일반적으로 이 데이터 셋은 이전 데이터 셋에 비해 큼.
- 데이터 셋은 신중하게 검사해야 하며 얻어진 결과의 **통계적 중요성을 보장**하기 위해 설계되어야 함.

비즈니스 목표 정의 (1)

• 수익성 있는 고객의 획득

〈그림 7-5〉 데이터마이닝 프로세스: 비즈니스 목표 정의



- **교차판매 또는 상향 판매 모델** - 고객의 선호도를 구매 가능성으로 변환한 제품 또는 서비스와 함께 모델링
- **고객 이탈 방지 관리** - 과거의 행동을 기반으로 고객의 이탈 가능성을 정확하게 모델링
- **일부 응용 프로그램** - 누가 어떤 제품이나 서비스를 구매할 것인지를 예측하는 것뿐만 아니라, 거래에서 얼마나 쓸 것인지 또한 예측

비즈니스 목표 정의 (2)

• 목표 변수의 값 설정

이미 스탠더드 신용카드를 소지하고 있는 고객에게 플래티넘 신용카드를 상향 판매하는 중에, **여러 개의 스탠더드 및 플래티넘 카드의 유형**이 존재한다.

비즈니스 목표는 아직 P2와 P3 유형의 플래티넘 카드를 소유하고 있지 않거나 S1, S3, S4 유형의 스탠더드 카드를 이미 소유하고 있는 고객만을 위한 **상향 판매**이다.

- 1) S1, S3 또는 S4 유형의 스탠더드 카드를 먼저 구매한 다음, P2 또는 P3 유형의 플래티넘 카드를 구입하거나 P2 또는 P3 유형의 플래티넘 카드를 바로 구입한 모든 고객의 **목표 변수는 1**로 설정
- 2) 다른 모든 고객의 **목표 변수는 0**으로 설정
- 3) 이러한 제한 사항과 고려사항이 데이터에 적용되면, 모델은 목표 변수가 0인 고객과 목표 변수가 1인 고객을 **구별하도록 학습**
- 4) 학습을 마친 후, 모델은 특정 고객이 플래티넘 카드를 살 **가능성이 있는지 예측함**. 이 때 비즈니스 그룹은 가망고객이 마케팅 캠페인에 포함되어야 한다고 생각하는 가능성을 임계값으로 설정함

비즈니스 목표 정의 (3)

• 비즈니스 혹은 캠페인을 위한 선정 규칙

- 규칙은 목표 그룹에 **배제**되거나 **포함**될 고객을 정의함.

- **신용 상품**은 고객의 신용 등급에 따라 판매를 **제한**함.
기업은 나쁜 신용등급 보유 혹은 마케팅 목적으로 연락을 하지 않을 것을 명시적으로 요구한 고객에게 새로운 서비스를 제공받을 수 없도록 '**블랙리스트**'를 보유함.
- 일부 국가에서는 승인되지 않은 다이렉트 메일이나 전화를 받고 싶어하지 않는 고객에 대한 리스트를 **중앙에서 관리**함.
- 특정한 지역 혹은 기타 다른 그룹에서 시장점유율을 얻기 위한 전략적 문제로 인해 캠페인에 **포함**시켜야 할 고객 그룹이 있을 수 있음.
이러한 경우에는 캠페인에 어차피 포함되기 때문에 해당 그룹의 구성원이 높은 모델 점수를 얻는 것과 관련이 없음.

비즈니스 목표 정의 (4)

- 프로젝트 계획 수립

- 데이터마이닝 프로세스의 시작일과 배송일 및 각 업무에 대한 **책임 자원**을 명시해야 함.
- **비즈니스 그룹**은 데이터마이닝 결과를 검토하고, 일관성 검사 수행, 그리고 배포를 위해 선택된 모델에 대해 최종 결정을 내릴 수 있어야 함.
- 지원되는 캠페인의 **시작과 종료일**과 함께 최종 모델 혹은 점수에 대한 게재일도 정의되어야 함.

- 목표그룹 선정

1. 통제 그룹에는 **무작위**로 선정된 고객만 포함됨.
이 그룹은 **기준 효과**(즉 캠페인의 영향을 받지 않은 보통 고객의 활동)를 측정하는 데 필요함.
2. 다른 셀은 사용된 모델에서 **가장 좋은 고객**만 포함됨.
이 간단한 설정을 통해 평균적인 고객의 행동에 대해 **모델 기반 선정**이 어떻게 수행되는지 측정할 수 있음.

비즈니스 목표 정의 (5)

• 비즈니스 성격과 비용, 수익 동인의 정의

- 이러한 지식은 **최종 모델**과 **목표 그룹의 선정**에 영향을 미침.
- 비즈니스 기술의 적용과 데이터마이닝 프로세스에 미치는 영향을 보여 주는 **비용/수익 매트릭스**를 정의하는 데에도 유용함.

ex) 핸드폰 계약 판매를 위한 콜센터 캠페인의 비용/수익 매트릭스

〈표 7-1〉 비용/수익 매트릭스

비용/수익 매트릭스	구매하지 않은 가능성	구매할 가능성
모델은 가망고객이 구매하지 않을 것으로 예측(연락하지 않음)	비용: \$0	비즈니스 기회 + \$895를 잃음
	첫 해 수익: \$0	
	합계: \$0	
모델은 가망고객이 구매할 것으로 예측 (연락 함)	비용: -\$5	비용: -\$105
	첫 해 수익: \$0	첫 해 수익: +\$1,000
	합계: -\$5	합계: +\$895

비즈니스 목표 정의 (6)

- 캠페인 평가 기준 수립

- 전체 프로젝트의 성공 혹은 실패를 결정하는 핵심임.
- 기대에 대해 **명확하게 정의**하는 것 이 도움이 됨.
- 비즈니스 유형, 고객세분화, 지역, 제품 및 기타 부문으로 구성되는 방식에 따라, 지역별, 판매 채널별, 제품 유형별 등 구매 시간을 **시간의 함수로 측정**하는 데 관심을 가져야 함.
- 기존 타겟팅 방법과 예측 모형을 사용하여 동일하거나 유사한 캠페인 설정으로 과거에 얻은 결과와 **비교**해 보는 것이 유용할 수 있음.

비즈니스 목표 정의 (7)

• Credite Est에서 비즈니스 목표 정의

Credite Est(가명)은 약 60만 명의 고객을 보유한 프랑스 지역의 중형 은행이다.

1965년 창사 이후 유기적으로 성장해 온 이 기업은 운영에 있어 계량적 접근 방식을 취하고 있다. 따라서 마케팅에서 **데이터마이닝**을 통한 **계량적 방법의 사용**은 기업의 본성이다.

은행은 금융 서비스 운영을 제공하며, 고객수익성 측면에서 다양한 고객을 보유하고 있다. 또한, 고객의 행동 특성(예: 제품 소유권)을 기반으로 한 **세분화 체계**를 사용하는 것 외에도 개개인의 고객 수준 공헌이익을 식별할 수 있는 **활동 기반 원가 계산 시스템**을 갖추고 있다. 문제의 프로젝트는 **프로파일링** 기법을 사용하여 새로운 가망고객을 확보하려는 비즈니스 목표를 가지고 있다.

특히, Credite Est은 대중 시장 부문에서 **수익성 높은 고객의 특성**을 파악하는 것이 목표이다. 특성을 좀 더 면밀히 파악하면 가망고객층에서 유사한 프로파일을 타겟팅 할 수 있다.

이 프로젝트의 본질은 은행이 **기업 수준의 데이터를 사용하는 것 이상**을 요구한다.

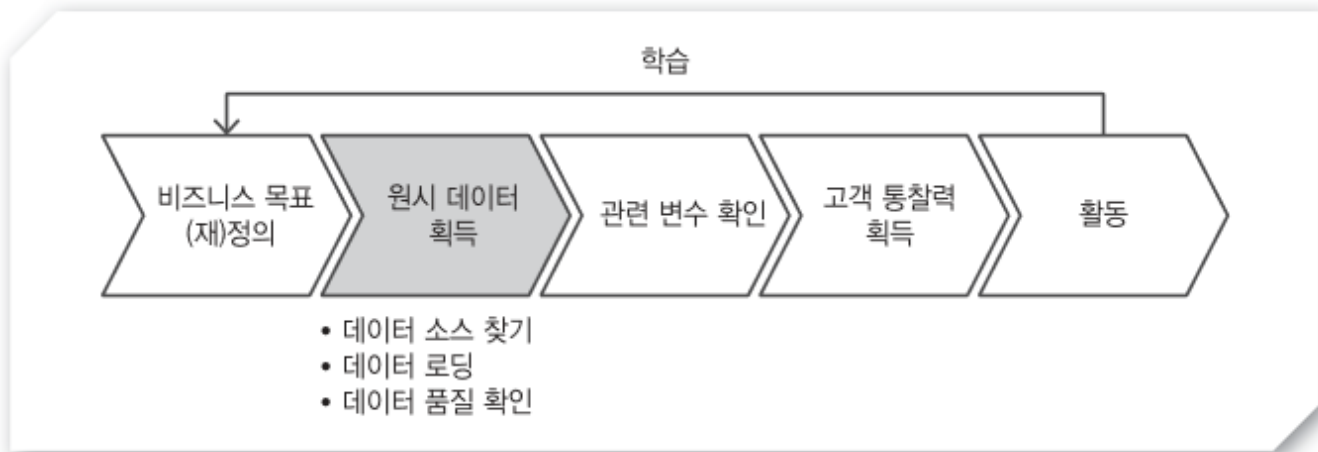
왜냐하면 가망고객에 대한 행동(거래) 데이터는 구할 수 없기 때문이다.

기업은 모든 데이터마이닝 프로젝트를 자체적으로 수행하기 때문에 이러한 프로젝트의 프로세스 관리에 상당한 경험을 갖고 있다.

원시 데이터 획득 (1)

• 원시 데이터 획득

〈그림 7-6〉 데이터마이닝 프로세스: 원시 데이터 획득



- 어떤 데이터를 사용하여 문제를 적절하고 정확하게 설명하며 목표한 활동을 모델링할 수 있는가에 대한 것을 식별함.
- 필요한 데이터가 식별되면 **데이터 마트**라고 불리는 데이터베이스에서 추출 및 통합되므로 차후의 데이터 조작 및 데이터마이닝 단계에서 사용 가능함.
- 분석할 원시 데이터의 **품질을 확인**하고, **기술적 점검**뿐만 아니라 주어진 비즈니스 내용에 데이터가 합당하고 정확한 추정이 이루어질 수 있는지 확인함.

원시 데이터 획득 (2)

• 1단계: 데이터 소스 찾기

- 원시 데이터 수집을 시작하기 위해 **비즈니스 요구사항**(상향)과 **기술적 제한 사항**(하향)에 따라 진행되는 **하향식**(top-down) 및 **상향식**(bottom-up) 프로세스에 맞춰 데이터 **출처**를 검토함.
- 시스템에서 동일하거나 유사한 정보 필드가 **모순적**인 경우, 인구통계학 정보로 인한 **품질 저하**는 매우 일반적임.
이 때 고급 데이터 정제 프로세스를 갖춘 **데이터웨어하우스 기반 시설**은 고품질 데이터로 작업할 수 있도록 도와 줌.
- 비즈니스 및 데이터 관리 분야 사람들과의 **대화**는 특정 상황에서 일반적으로 사용되는 데이터 소스를 이해할 수 있을 뿐 아니라 중요한 정보가 포함될 수 있는 새로운 소스도 발견할 수 있음.
- 로드해야 할 데이터 소스에 대해 이해했다면 소스 데이터가 매핑되는 단순한 **관계형 데이터 모델**을 설계해야 함.

원시 데이터 획득 (3)

- 2단계: 데이터 로딩

- **어떻게 필요한 데이터를 추출할 것인지**를 명시한 후에 전체 데이터의 하위 집합만 모델링할 수 있는 쿼리를 추가로 정의해야 함(예: 특정 고객 세그먼트, 지역, 기간 등).
- 데이터 관리(IT)에 명시된 **데이터 요구사항**을 제공하도록 요청하고, IT팀은 일괄 모드에서 사전 정의된 일괄 처리 모드로 미리 정의된 시간에 실행되는 **필수적인 데이터 쿼리**를 준비함.
- **데이터 마이너**는 데이터마이닝 환경으로 데이터를 가져 오는 방법을 정의하는데, ftp 프로토콜을 통한 전송이 일반적이며, 네트워크를 통해 직접 액세스 할 수 있도록 공통 파일 서버에 배치할 수도 있음.
- 데이터가 정의된 랜딩 영역에 전달된 후에는 데이터마이닝 환경에서 이전에 정의된 **데이터 모델의 채우기**가 사용됨.

원시 데이터 획득 (4)

- 3단계: 데이터 품질 확인

- 고객 태도에 대한 잘못된 추론, 열악한 서비스를 통한 고객 유실, 또는 의사결정자에게 데이터 전달 지연과 같은 나쁜 데이터 품질로 인한 비용은 영업 이익의 약 **15~25%**.
- 데이터 품질은 **사용 용도**와 **데이터 자체**에 의해 결정됨.

데이터 품질과 관련된 측면

- | | |
|------------------|-------|
| • 정확성(일관성 및 유효성) | • 완전성 |
| • 연관성 | • 신뢰성 |

- 데이터 품질을 확인할 때는 데이터의 기술적 측면(주 키, 중복 레코드, 결측값 등) 뿐만 아니라 비즈니스 내용과 관련된 **품질 문제**도 살펴봐야 함.

원시 데이터 획득 (5)

- 3단계: 데이터 품질 확인
 - 예비 데이터 품질 평가 질문

- 1) 데이터가 원래의 소싱 요구사항과 일치하는가?
- 2) 품질이 충분한가?
- 3) 데이터를 이해하는가?

- 데이터 마이너
 - 데이터 마이너는 **비즈니스와 IT 요구 간의 연결**을 나타냄.
 - 데이터 마이너는 데이터를 **이해**하고 있음을 입증해야 함.
 - 데이터 품질
 - 데이터를 올바르게 해석할 수 있는 능력
 - 기본적인 데이터 해석 및 집계 능력

원시 데이터 획득 (6)

- Credite Est에서 원시 데이터 모으기

현재 고객을 위한 반응 변수는 **고객 공헌이익**이다.

기업은 고객을 운영적 공헌으로 분류하고 그 중에서 **상위 20%**을 프로파일링 하기 위해 고객을 선정한다. 가망고객의 거래 정보는 사용할 수 없다. 이것이 은행이 기존 고객과 가망고객 **모두에게 유용한 정보**에 의존하는 이유이다.

정보 유형 중 하나는 지역의 사회 경제적 지위, 평균 연령, 주택 유형 등 과 같은 **지리적 인구통계 데이터**이다. 이것은 다이렉트 마케팅 에이전시로부터 구입한 다음 기존 고객 개개인의 레코드에 첨부된다. 즉 우편번호에 따라 지리적 인구통계 정보가 기존 고객 레코드에 추가된다.

모델은 지리적 인구통계 정보와 함께 독립 변수인 **고객 영업 이익률의 예측**을 시도한다. 이 프로세스의 근거는 높은 가치의 고객을 특징 짓는 프로파일을 찾아 향후 가망고객의 정보에 적용하고자 하는 것이다.

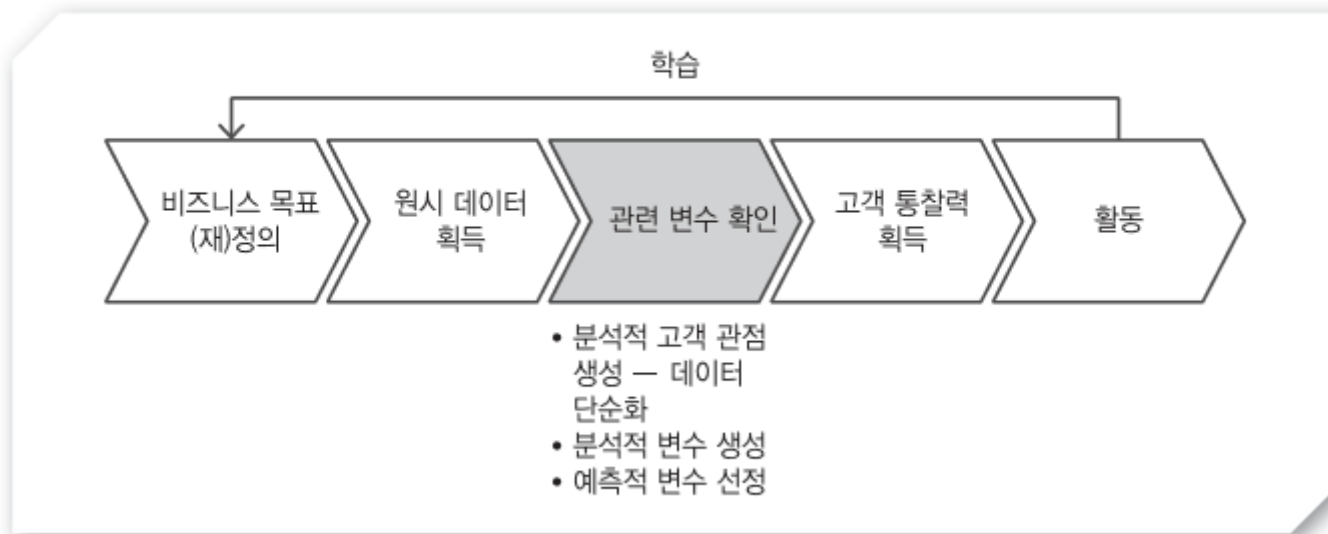
Credite Est는 기존 고객 레코드에 총 65개의 변수를 추가했다.

이는 Claritas뿐만 아니라 프랑스 리스트 매니저인 CIFEA로부터 입수되었다.

관련 변수 확인 (1)

• 관련 변수 확인

〈그림 7-7〉 관련 변수 확인



- 1) 단일값(레코드 혹은 행)에서 시간 경과에 따른 고객 행동에 대한 모든 정보를 모은 원시 데이터에 **평균 뷰**를 생성함.
- 2) 예측력을 가질 수 있는 **새로운 변수를 생성**하여 선행적 비즈니스 지식을 포함시킴.
- 3) 모델링 된 타겟 행동을 충분히 설명해 주는 **소수의 변수만 식별**하고 선정함.

관련 변수 확인 (2)

- 1단계: 분석적 고객 관점 생성 — 데이터 단순화

- 각 고객이 데이터 분석 및 예측을 위한 단위를 구성하기 때문에 개별 고객이 사용할 수 있는 **모든 데이터를 수집하고 통합**해야 함.
- 고객의 과거 행동은 시계열 기반의 **관계형 거래 데이터 베이스**에서 **데이터 쿼리**를 통해 얻을 수 있음.
- 합계, 평균, 중앙값 및 표준편차와 같은 **기술 통계**는 관련된 시계열의 특징을 포착하기 위해 사용됨.
- 최초의 관계형 데이터 구조를 **비정규화(flattening)** 하는 작업은 데이터 마이너가 포함되어 플래트닝 프로세스의 세부 정보를 정의하고 IT 자원을 사용하여 대상 데이터 형식을 얻음.
- 나중에 나울 예측 모델링 단계에서 원시 데이터 변수 외에도 **새로 생성된 변수**를 예측 변수로 사용할 수 있음.
- 목표나 종속 변수에 대한 만족스러운 정의를 찾으면 모든 고객에 대해 값을 생성하고 기존 데이터 테이블에 추가함.

관련 변수 확인 (3)

- 2단계: 분석적 변수 생성

- 플래트닝으로 인해 생성된 변수의 기본 세트는 예측 모델을 위한 데이터의 잠재력을 완전히 파악하지 못할 수 있기에 원래 변수에서 **파생된 추가 변수**를 도입하고자 할 수 있음.
- **변수 비닝**(또는 범주화)도 자주 발생하는데, 급여와 같이 매우 비대칭적인 변수를 정의된 값에 따라 낮음, 중간, 높음과 같이 몇 개의 **클래스로 나누는 것**이 포함됨.
- **결측값**은 데이터 셋의 품질을 향상시키는 데 중요함.

- 하나 이상의 결측값이 있는 각 행을 삭제하는 것(가장 선호하지 않는 방법)
- 결측치를 상수값으로 대체하는 것
- 변수의 분포를 기반으로 임의의 값을 생성하는 것
- 기댓값 최대화 알고리즘(expectation-maximization algorithm)과 같이 다른 변수와의 상관관계를 기반으로 임의의 값을 생성하는 것

관련 변수 확인 (4)

• 3단계: 예측 변수 선정

- 모든 예측 변수를 신경망에 입력하면 모델링에 많은 시간이 소비되며 때로는 모델의 **과적합화**(overfitting)를 야기함.
- **변수의 제외**는 일반적으로 많은 변수가 예측력을 갖지 않기 때문에 얻어진 모델의 예측력을 저하시키지 않고 가능함.
- 사용 가능한 모든 변수와 관련된 모든 단일 변량 분포를 확인하고, **상수**와 같이 하나의 값만 가지는 변수는 예측력을 갖고 있지 않기 때문에 즉시 제외함.
- 어떤 경우에는 **공선성**(collinear)의 예측 변수가 로지스틱 회귀와 같은 특정 유형의 모델에서 예측 성과에 부정적인 영향을 미칠 수 있으므로 공선성을 확인하고 해당 변수는 프로세스를 진행하기 전에 제거해야 함.
- 대응 카이제곱 검정, 선형 상관 분석, 대응 표본 선형 회귀를 수행하여 목표 변수와 **상관관계가 거의 없는 변수**도 제거해야 함.

관련 변수 확인 (5)

• Credite Est에서 관련 변수 확인

추가된 모든 정보를 포함하는 하나의 데이터 파일을 만들 때에는 탐색적 분석 단계로 시작한다. 추가된 데이터의 주요 관심사는 **누락된 정보의 양**이다.

모든 추가 변수의 **50% 정도**는 누락된 데이터이다. 다음 단계는 누락된 데이터를 의미 있게 대체할 수 있는지 평가한다. 이러한 작업은 **결측값**의 전체 비율을 42%에서 21%로 개선했다.

다음은 모든 변수의 단일 통계량(평균, 표준편차, 빈도, 이상치)을 통해 변수가 충분한 **무결성**을 지니고 있는지 확인한다. 이 단계에서 변수가 65개에서 54개로 감소했다.

다음 단계는 종속 변수(고객가치)에 대한 독립 변수의 **상관관계**(범주형 변수의 경우 평균 분석)를 계산한다. 이것은 **독립 변수가 변형되거나 새로운 변수를 생성**하는 반복적인 과정이다.

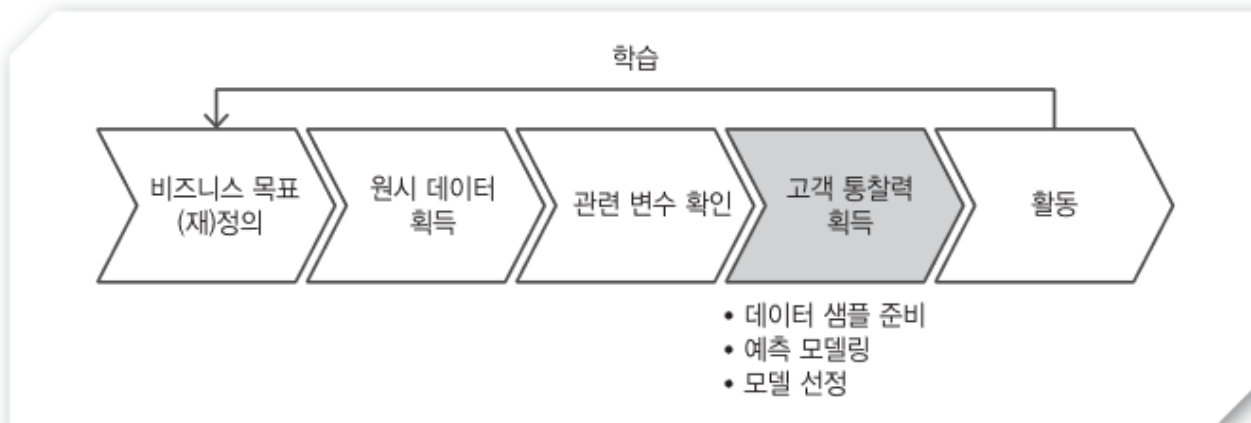
예를 들어, 0~4세, 5~11세, 12~18세의 자녀를 둔 가구를 나타내는 변수가 있다. 여기에서, 자녀 유무에 대한 단순한 더미변수를 만든다.

결국, 데이터 평가 프로세스를 통해 총 17개의 변수가 종속 변수와 적절한 상관관계를 가지는 것으로 나타난다. 이는 다음 단계인 반응 모델에서 사용된다.

고객 통찰력 획득 (1)

• 고객 통찰력 획득

〈그림 7-8〉 데이터마이닝 프로세스: 고객 통찰력 획득



- **예측 모델링**을 통해 계획된 캠페인을 수행하기 위해 필요한 고객 행동 및 특성에 대한 지식을 추출해야 함.
- 서로 다른 모델링 패러다임을 통해 얻은 다양한 유형의 예측 모델은 **지도 학습**(supervised) 및 **비지도 학습**(unsupervised)으로 구분함.
- 모델을 구축하는 것은 구매자 또는 비구매자로 나타나는 고객의 가능성을 예측하기 위해 고객을 설명하는 변수들 사이의 **올바른 관계**를 찾는 것을 의미.
- 이 프로젝트 단계의 결과는 온라인 상품 환경에 적용되는 **예측 모델**이나 **고객 점수**로 나타남.

고객 통찰력 획득 (2)

• 1단계: 데이터 샘플 준비

- 모델을 구축(또는 학습)하기 전에, 통계적으로 **유의한 결과**를 얻을 수 있는 충분한 데이터가 있는지 분석해야 함.
- 데이터가 충분하다면, 모델을 구축하는 **학습용**과 모델 구축에 사용하지 않았던 데이터로써 모델의 성능을 검토하는 **검증용** 데이터 두 개로 나눔.
- 이 과정은 제품 또는 캠페인을 론칭하기 전에 모델의 일반화에 대한 **객관적인 평가**를 제공함.

• 2단계: 예측 모델링

1. 규칙(또는 선형/비선형 분석 모델)은 학습용을 기반으로 구축됨.
2. 이러한 규칙은 캠페인에 필요한 답을 얻기 위해 새로운 데이터 셋에 적용함.

- 학습용을 기반으로 **예측 오차**를 최소화하는 예측 모델을 개발함.
- 이 과정에서 모델에 최적화된 **매개 변수**를 얻을 수 있음.
- 다양한 **통계 기법**을 적용하여 여러 가지 모델을 함께 학습하는 경우가 일반적임.

고객 통찰력 획득 (3)

• 3단계: 모델 선택

- 모든 대안 모델에 대해 학습되고 나면 각각의 **오분류율**을 비교하거나 **리프트 분석**을 수행하여 상대적인 예측 성과를 **비교**함.
- 다른 모델보다 우수한 예측 성과를 보이며 학습용에서 검증용 데이터까지 가장 일반적이라고 생각하는 **모델을 선택**함.
- 모델은 이전에 정의된 **비용/수익 매트릭스**를 적용하여 모델의 **경제적 영향**을 포함해야 함.

• Credite Est에서의 고객 통찰력 획득

모델 개발자가 선택한 방법론은 **로지스틱 회귀 분석**이었다.

목표는 가망고객 그룹에서 특정 고객을 타깃 할 것인지 타깃 하지 않을 것인지이기 때문에 종속 변수를 **0** 과 **1**로 분류하였다. 이전 단계에서 최소의 상관관계인 변수들만 유지되었다.

다중 공선성의 문제가 발생되면 높은 공선성을 보이는 변수를 이론에 근거하여 제거하였다.

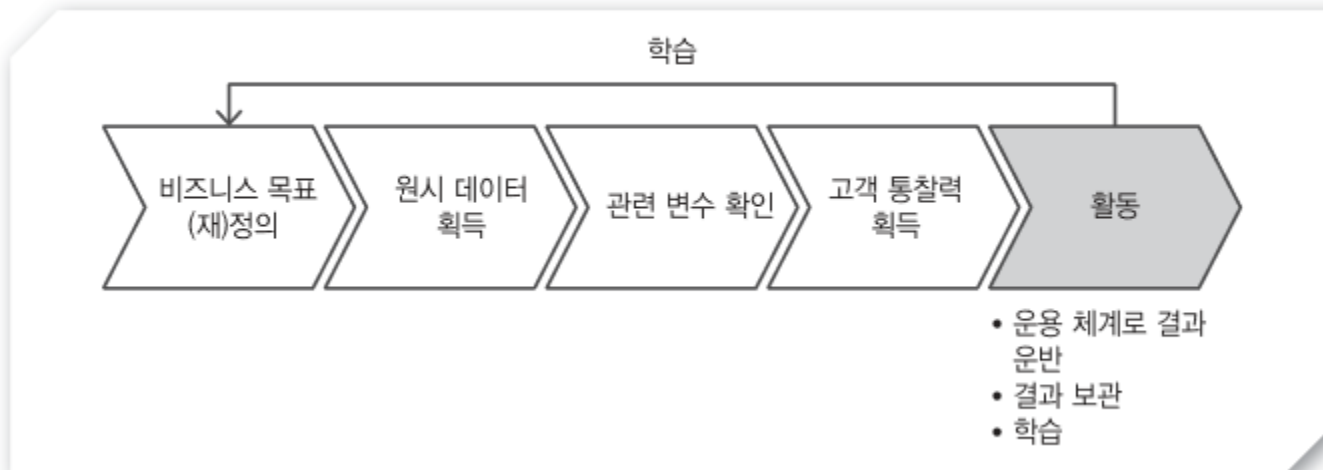
최종 모델은 결측값의 수가 적으면서도 예측 성과를 고려하여 선택되었는데, **중산층 그룹, 기술 그룹, 아동 지표, 주택가치 지표, 직업 군** 등 다섯 가지 예측 변수가 포함되었다.

모델의 분류 성과는 학습용에서 **75.5%**, 검증용에서 **69.8%**로 나타났으며, 예측 모델을 사용하지 않았을 때보다 대략 **20% 정도 향상**되었다.

활동 (1)

• 활동

〈그림 7-9〉 데이터마이닝 프로세스: 활동



- 데이터마이닝 프로젝트의 최종 목표는 **그 결과에 따라 활동하는 것**. 결과의 **전개**라고도 함.
- 프로젝트 계획은 데이터마이닝 결과를 IT 시스템(데이터베이스, 웹사이트, 콜센터 등)을 지원하는 프로세스로 다시 공급하는 데 필요한 **IT 리소스의 참여**와 **가용성**을 파악하고 있어야 함.
- 활동은 결과를 운영 시스템에 **전달**하고, 결과를 **보관**하고, **학습**하는 것으로 세분화 됨.

활동 (2)

- 1단계: 운영 시스템에 결과 전달

- 어떤 **고객 그룹**이 캠페인에 대한 반응률이 더 높게 나타날 것인지 식별하기 위해, 선택한 모델을 전체 고객 데이터에 적용함.
- 각 고객이 획득한 점수와 정의된 임계값은 **해당 고객이 캠페인에 참여할 자격이 있는지를** 결정함.
- 고객의 점수를 매기기 전 모델에서 필요한 변수와 함께 각 고객에 대한 최신 정보가 포함된 **점수 데이터 셋**을 준비해야 함.
이 때, 모델 구축에 사용된 학습용 및 검증용 데이터 셋과 마찬가지로 점수 데이터 셋에서도 동일한 **변수 변환, 파생 및 선정 절차**를 거쳐야 함.
- 결과를 운영 시스템에 전달할 때 모델 점수 정보를 올바른 고객에게 정확하게 연결하기 위해 시스템에 필요한 필수적인 **고객 식별자**를 제공해야 함.

활동 (3)

- 2단계: 결과 보관

- 모델에 대한 정보를 **보관**하지 않는 기업은 보관된 정보로부터 과거 경험을 배우는 기업들만큼 빠른 학습을 기대할 수 없음.

- | | |
|------------------------|--------------|
| • 사용된 원시 데이터 | • 목표 변수 계산 |
| • 각 변수의 변환 | • 모델과 매개 변수화 |
| • 파생 변수를 만들기 위한 공식 | • 점수 임계값 수준 |
| • 학습, 검증, 그리고 점수 데이터 셋 | • 최종 목표고객 선정 |

- 이러한 정보를 아는 것과 이것을 사용할 수 있도록 준비하는 것은 모델 성능에서 **변칙**을 이해할 수 있도록 도와 줌.

활동 (4)

• 3단계: 학습

- 데이터마이닝 프로젝트를 통해 **학습**하는 것은 프로세스의 필수적인 부분이며 이를 **폐쇄형 고리**(closing the loop)라고도 함.
- 데이터마이닝 프로젝트로부터 학습하기 위해, 먼저 **성능 및 비즈니스 영향**에 대한 사실을 파악해야 함.
- 중간적 성능 피드백 없이 캠페인이 끝날 때까지 **파일럿**(blind piloting)을 피할 수 있기 때문에 **캠페인 모니터링**은 데이터마이닝 그룹이 반드시 제공해야 하는 중요한 기능임. 일반적으로 모니터링은 지역, 세분 고객, 제품 등의 응답 및 구매율과 같은 **주요 성과 지표**를 제공함.

캠페인 평가에 의한 표본 학습 예

- 커뮤니케이션 채널의 선택에 따라 구매율 공개
- 화려하고 상세한 상품 브로셔가 포함된 다이렉트 메일이 한 페이지의 흑백 광고지보다 덜 판매되는 것을 발견

활동 (5)

- Credite Est에서의 정보에 따른 활동

최종 모델은 가망고객을 타겟으로 순차적으로 진행되었다. 목표는 향후 라운드에서 **모델을 반복적으로 수정**하는 것이다.

첫 번째 단계로, Credite Est는 최종 모델의 최소 5개 변수 중 3개 이상에 대해 잘못된 값을 가진 주소를 중개인으로부터 구입했다. **가망 고객**은 모델에 의해 점수를 산출하고 가치가 높은 고객이 될 가능성에 따라 평가된다. 10,000명의 가망고객 중 절반은 단기 금융시장 상품을, 절반은 대출 상품을 대상으로 했다.

목적은 각 상품에 대한 **두 샘플의 수용성을 평가하는** 것이다. 또한 기준 시나리오는 **무작위 표본의 가정으로 방문 캠페인을 수행하는** 것이다. 두 표본 모두 타겟 메일링이 기준 시나리오보다 훨씬 성공적이었지만, 이것은 이후 모델 개선을 위한 첫 단계이다.

특히, 응답률을 평가하는 것 외에도, **획득된 고객의 가치**, 즉 프로젝트의 원래 목표를 **추적**하고 **문서화**하는 것이 중요하다.

Yapi Kredi 사례 (1)

• 야피 크레디 — 예측적 모델 기반의 교차판매 캠페인

- 1944년에 설립된 **터키 최초의 민간 은행**인 야피 크레디(Yapi Kredi)는 터키 금융 분야의 개척자임.
- 임대, 인수, 투자 은행 보험, 중개 및 신경제 기업에서 활용하는 계열사 뿐 아니라 860개 이상의 국내 지사 및 다양한 자회사를 보유함.
- 신용카드, 운용 자산, 채권 매입, 사적 연금 기금 및 생명 보험과 비생명 보험에서 선도적인 지위를 가지고 터키에서 자산 규모로 **네 번째로 큰 개인 소유 상업 은행**으로 자리매김 함.

• 도전 과제

- 500만 이상의 고객층에 대한 잠재력을 충분히 탐색하여 **최고의 고객층과 긴밀한 금융 관계의 유지**를 목표로, 고객 당 계약 비율을 5개로 늘리고자 함.

목표 달성에 필요한 기능

- 고급 분석을 통한 **고객세분화**
- 제품 번들의 **세분화별 제공 사항**
- **예측 모델링**과 같은 고급 CRM 도구를 사용하여 **타겟 캠페인**을 통해 수익성 높은 그룹으로 고객 전환

Yapi Kredi 사례 (2)

• 해결책

- 야피 크레디는 고객당 제품 비율을 높이고, 고객으로부터 새로운 투자를 하도록 하며, 새로운 분석적 CRM 방법의 효과를 입증하기 위해, 소비자 금융 상품의 **교차판매**를 위한 일련의 **파일럿 프로젝트**를 수행함.

1) 비즈니스 목표 정의

- **고객과 은행의 관점에서 교차판매에 가장 적합한 제품을 찾는 것**
- 예측 모델은 교차판매 캠페인을 기반으로 제공될 잠재 제품에 대한 심층 분석을 거친 후, 고정 수익을 보장하고 낮은 위험 투자 상품으로 야피 크레디의 B형 뮤추얼 펀드를 선택하기로 결정

선정한 고객그룹

- 자산 증가를 위해 이미 **B형 뮤추얼 펀드에 투자한 고객**
- 제품 비율을 높이고 새로운 자금을 얻는 데 도움을 주는 고객으로 **아직 B형 펀드를 소유하지 않은 고객**

Yapi Kredi 사례 (3)

- 해결책

1) 비즈니스 목표 정의

- 시범 사업에 캠페인과 이용 가능한 자원의 특성을 제공함으로써 고객 지점 방문 동안 발신 전화와 활성의 마케팅을 기초로 **3,000명의 고객과 연락을 하는 것으로** 결정하였고, **1,200명의 목표고객**이 콜센터에 의해 연락이 됨.
- **연락 채널의 응답과 구매율**(지점 혹은 콜센터)은 캠페인의 성공을 평가하는 척도로 선정됨.

2) 원시 데이터 획득

- **데이터 마트의 50개 이상 소스 시스템 테이블**에서 데이터가 추출됨.
- 데이터 마트에는 **우선 순위**가 높은 비즈니스 활동(예: 파일럿 캠페인)에 필요한 데이터가 포함되어 있으며, 이후 데이터 조작 및 데이터마이닝 단계를 위해 짧은 시간 내에 데이터를 즉시 사용할 수 있게 함.

Yapi Kredi 사례 (4)

• 해결책

3) 관련 변수 확인

- 분석 및 예측 모델링에 필요한 올바른 고객 중심의 데이터 형식을 얻으려면 다양한 집계 및 변환이 필요하므로 고객 행동 및 선호도를 파악하는 **고객 속성을 선정**

인구통계	나이, 성별, 결혼 상태, 그룹 구성원, 주소, 직업 및 기타 식별 특성 포함
제품 소유권	각 고객이 보유한 제품 포트폴리오, 개장일, 소유한 제품의 최대 보유 기간과 같은 고객 보유 기간과 관련된 변수
제품 사용	은행 거래의 평균적 숫자와 같은 고객의 사용 빈도와 관련된 변수
채널 사용	고객의 자동 납부 행동, 자동 납부의 평균 금액, 각기 다른 채널의 사용 비율 등과 관련된 변수
자산	유가 증권, 정기 예금, 요구불 예금 등에 투자한 평균 잔액 같은 저축 및 투자 상품과 연관된 변수
부채	평균 대출 잔액, 신용카드에 있는 평균 잔액 등과 같은 대출과 관련된 변수
수익성	파일럿 프로젝트는 모든 고객이 수익성을 이용할 수 없기에 수익성 지표가 만들어짐 지표는 절대 가치를 부여하지 않고 수익성에 따라 고객 순위를 매기는 데 사용됨

Yapi Kredi 사례 (5)

• 해결책

4) 고객 통찰력 획득

- 과거 6개월의 고객 데이터를 바탕으로, 3개월 동안 B형 뮤추얼 펀드에 투자하려는 고객의 성향을 평가하기 위해 **다섯 가지 예측 모델을 개발**.
- 가장 좋은 모델은 상위 10분위의 고객에서 **리프트 값 2.9**를 산출하는 결과가 **로지스틱 회귀 분석**으로 나타남.

5) 활동

- 콜센터와 지점을 통해 캠페인을 시작하려면 각 채널은 어떤 고객에게 연락해야 하는지 **목표고객**을 명확하게 지정해야 함.
- 3,000명의 고객 집합은 16개 지점에, 나머지 1,200명의 고객은 콜센터로 배정함.

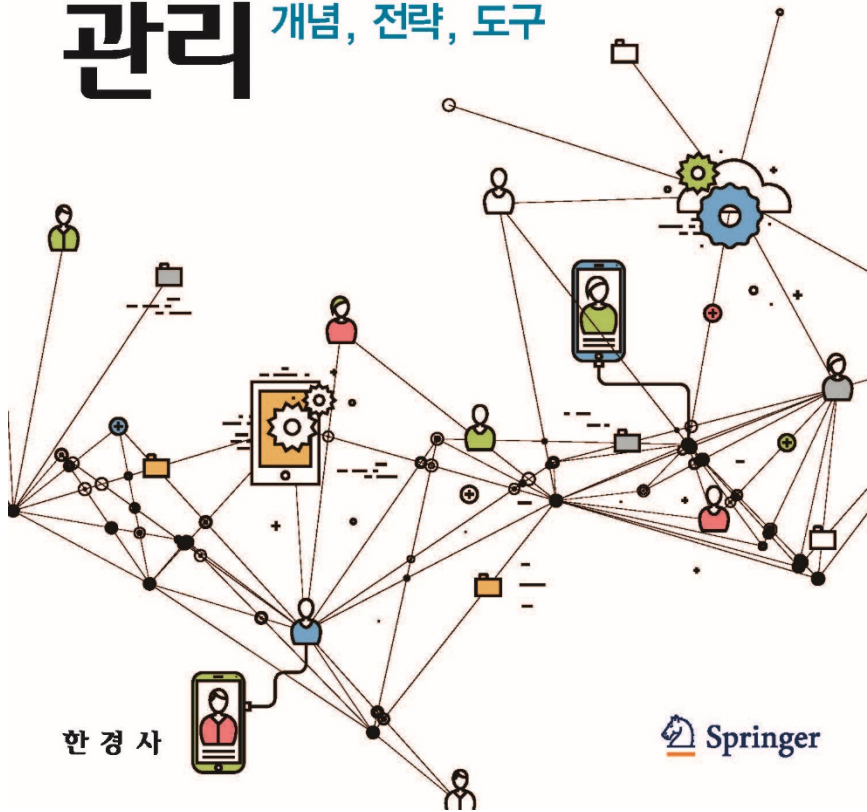
	응답률(%)	펀드 판매액(유로)
지점	6.5	582,000
콜센터	12.2	452,000
합계	18.2	1,034,000



Q & A

고객관계 관리

개념, 전략, 도구



고객관계관리 개념, 전략, 도구 (제2판)

V. Kumar, W. Reinartz 공저

홍태호, 신태수, 안현철, 김은미 공역

한경사, 2018

본 강의보조자료는 고객관계관리 개념, 전략, 도구(제2판)의 한국어판 서적을 기초로 제작되었으며, 해당 서적의 저작권은 '도서출판 한경사'에 있습니다. 저작권법에 의하여 한국 내에서 보호를 받는 저작물이므로 무단전재와 복제를 금합니다.