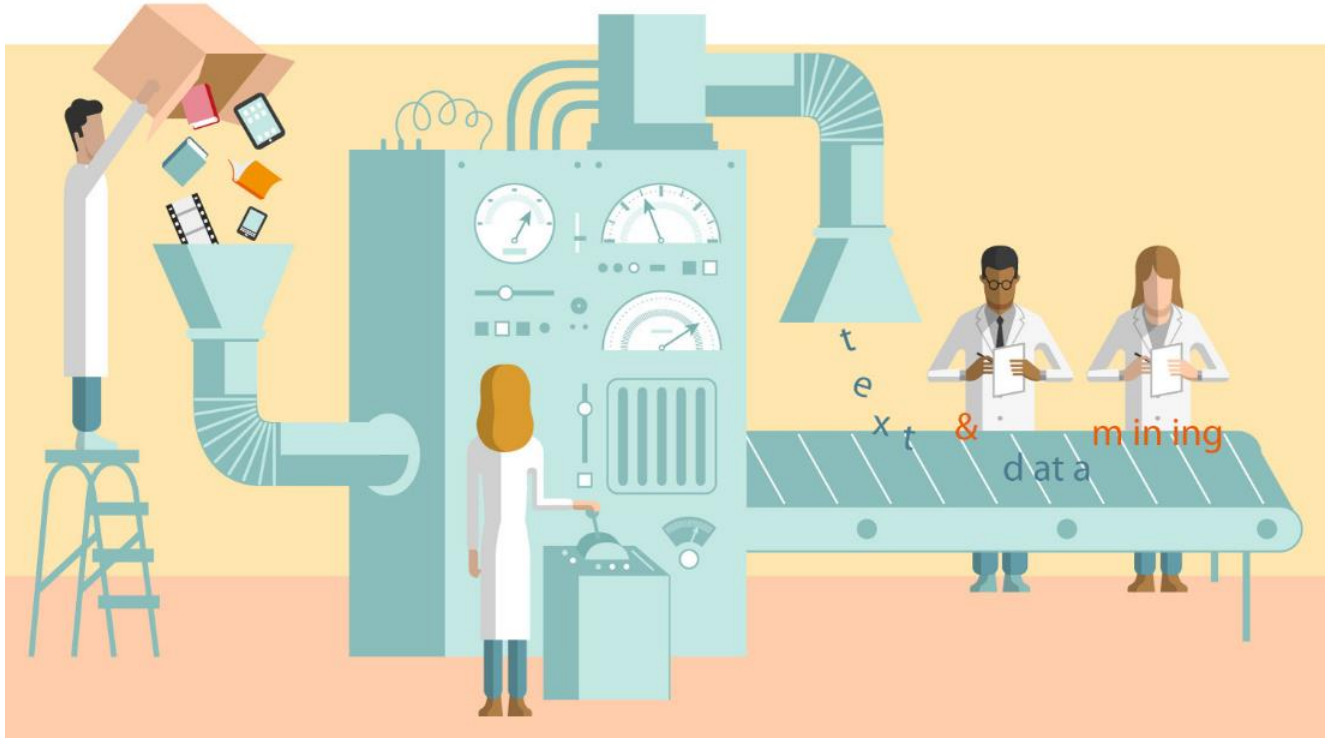
문제해결과 기업성과 달성을 위한 정보시스템 활용

데이터 마이닝 프레임워크

SNU Business School

데이터 마이닝(Data Mining)



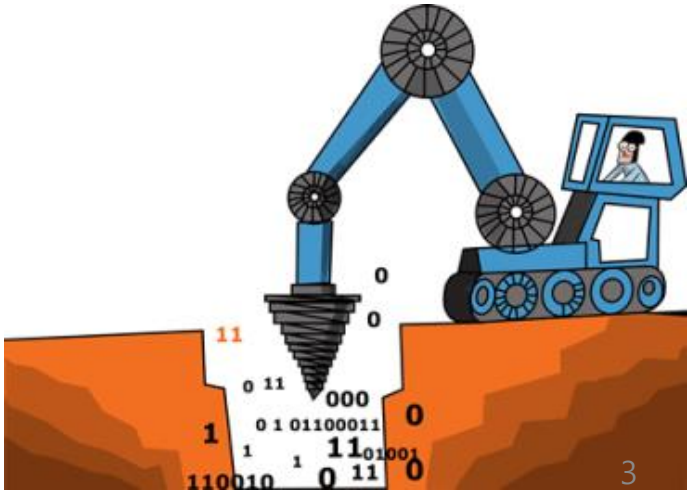
데이터 마이닝이란?

· 데이터 마이닝(Data Mining)[source: https://en.wikipedia.org/wiki/Data_mining]

- 기계학습(machine learning), 통계학, 데이터베이스 시스템 등의 방법론을 활용하여, 대규모의 데이터로부터 패턴을 찾아내는 종합적인 과정

- 데이터 마이닝의 궁극적인 목적인 데이터로부터 현실에서 활용 가능한 “정보(information)”을 찾아내는 것

- 목적을 이루기 위한 데이터 전처리, 모델 수립과 통계적 추론, 평가(evaluation), 시각화 등 전반적인 과정을 포괄하는 개념



데이터 마이닝의 적용 분야

· 데이터 마이닝은 아래와 같은 분야에 적용하여 현실에 유용한 결과를 도출할 수 있음

- **회귀(regression):** 데이터 혹은 데이터 셋 간의 관계를 설명하는 함수를 찾는다(예: 성적과 공부량 간의 관계 규명)

- **분류(classification):** 집단을 정의해 각 관측치를 추론을 통해 구분한다(예: 우수 고객과 이탈 고객)

- **군집화(clustering):** 구체적인 특성을 공유하는 군집을 찾는다. 군집화는 각 군집의 특성을 미리 정의하지 않는다는 점에서 분류와 구분된다(예: 비슷한 구매 패턴을 보이는 고객들의 군집)

- **연관성(association):** 동시에 발생한 사건 간의 관계를 정의한다(예: 고객들이 함께 구매하는 상품들 간의 관계 규명)

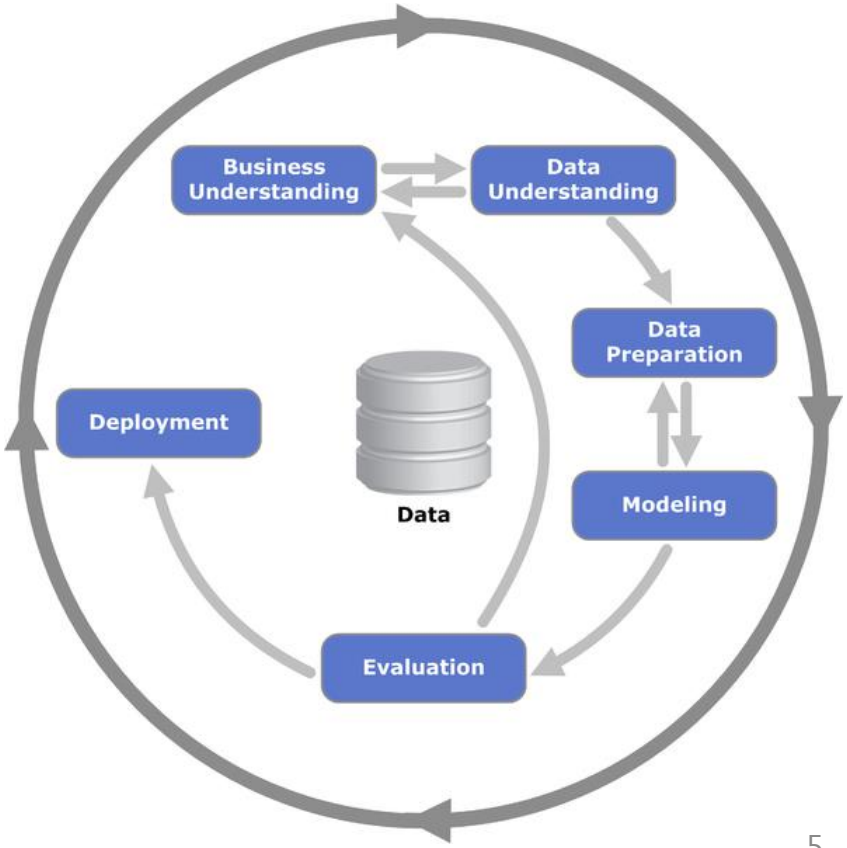
- **이상치 탐색(anomaly detection):** 정상적이지 않은 데이터 레코드나 데이터 에러를 찾는다(예: 아웃라이어 탐색)

- **요약(summarization):** 시각화(visualization)과 리포트(report) 등을 통해 데이터를 보다 간단하게 표현한다(예: 그래프 작성)

데이터 마이닝 프로세스

· 데이터 마이닝 프로세스는 크게 아래 여섯 단계로 나누어질 수 있음
[source: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining]

- 비즈니스 이해(Business understanding)
- 데이터 이해(Data understanding)
- 데이터 준비(Data preparation)
- 모델링(Modeling)
- 평가(Evaluation)
- 적용(Deployment)



데이터 마이닝 프로세스

· 데이터 마이닝의 9가지 법칙(“Nine laws of Data Mining”)
[source: http://khabaza.codimension.net/index_files/9laws.htm, https://www.forbes.com/sites/metabrown/2016/01/27/9-laws-for-data-analytics-profits/#7d1c8b9f16e0]

- 1. 비즈니스 목표의 법칙(“Business Goals Law”)
- 2. 비즈니스 지식의 법칙(“Business Knowledge Law”)
- 3. 데이터 준비의 법칙(“Data Preparation Law”)
- 4. 공짜 점심의 법칙(“NFL-DM”)
- 5. 왓킨스의 법칙(“Watkin’s Law”)
- 6. 인사이트의 법칙(“Insight Law”)
- 7. 예측의 법칙(“Prediction Law”)
- 8. 가치의 법칙(“Value Law”)
- 9. 변화의 법칙(“Law of Change”)

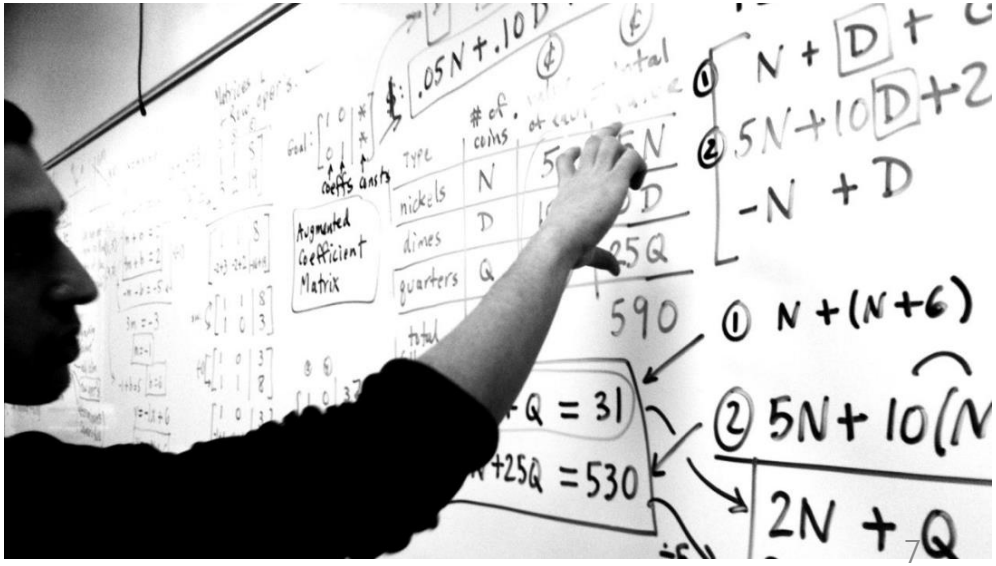
데이터 마이닝 프로세스

· 비즈니스 이해(Business Understanding): 현실 세계의 문제를 데이터 마이닝 문제로 환원

- 프로젝트의 목적과 요구사항을 비즈니스적인 관점에서 이해

- 이러한 지식을 데이터 마이닝 컨텍스트에서 문제 정의(problem definition)

- 데이터 분석을 위한 전반적인 계획을 수립



데이터 마이닝 프로세스

· 데이터 마이닝의 제1법칙: 비즈니스 목표의 법칙

- 비즈니스 목표(Business Objectives)는 모든 데이터 마이닝 솔루션의 시작이다("Business objectives are the origin of every data mining solution")

- 문제를 이해하지 못하면, 문제를 풀 수 없다

- 데이터 마이닝은 기술(technology)보다는 비즈니스 목표를 그 핵심에 두고 있는 프로세스이다

- "Without a business objective (whether or not this is articulated), there is no data mining"

데이터 마이닝 프로세스

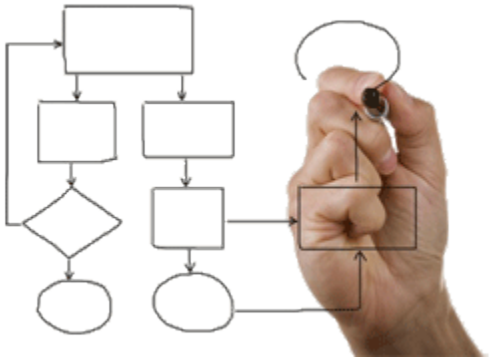
· 데이터 마이닝의 제2법칙: 비즈니스 지식의 법칙

- 데이터 마이닝 프로세스의 모든 단계에서 비즈니스 지식(business knowledge)은 핵심적이다

- 달리 말해, 비즈니스 지식이 없다면 데이터 마이닝의 모든 단계는 효과가 있을 수 없다

- 비즈니스 지식은 전체적인 프로세스를 이끌어 나가고, 데이터 마이닝의 결과 중에 유의미한 것을 알아볼 수 있는 혜안을 제공한다

- “If you don’t have someone who knows the business on the team, you won’t get good results”



데이터 마이닝 프로세스

· 데이터 이해(Data Understanding): 데이터에 대한 기초적인 이해 습득

- 데이터 수집(data collection): 웹, 소셜 미디어, …

- 데이터 품질(data quality) 검증: 결측치(missing values), 이상치(anomalies) 등

- 데이터 종류(data type): 정형 데이터(수치형), 비정형 데이터(텍스트, 음성, 이미지) 등

- 데이터 탐색(data exploration): 데이터 시각화, 기초 통계량 계산 등

데이터 마이닝 프로세스

· 데이터 준비(Data Preparation): 데이터 분석을 위한 기초 작업

- 데이터를 불러오는 것부터 모델에 입력하기 직전까지의 모든 작업을 포함

- 데이터 전처리(data preprocessing): 데이터를 모델에 맞는 형식으로 변환

- 데이터 클리닝(data cleaning): 데이터의 품질을 높이기 위해 불량한 데이터를 삭제

데이터 마이닝 프로세스

· 데이터 마이닝의 제3법칙: 데이터 준비의 법칙

- 데이터 마이닝 프로세스에서 데이터 준비는 절반 이상의 비중을 차지한다

- 비공식적으로, 데이터 수집 및 전처리 단계에 50~80%의 노력이 투자된다

- 프로세스의 자동화가 시간을 단축시켜 주지만, 컴퓨터의 한계점으로 인한 오류 발생의 위험도 존재한다

- “Most of the time and effort goes into the dirty work of cleaning data and getting it in shape for analysis.”

데이터 마이닝 프로세스

· 데이터 마이닝의 제4법칙: 공짜 점심의 법칙

- 특정 응용 분야에 적합한 모델은 지속적인 실험을 통해서만 발굴될 수 있다

- Wolpert의 “No Free Lunch” (NFL) 이론: 모든 데이터 마이닝 문제에 최적인 기계학습 알고리즘은 존재하지 않는다
=> 각 문제에 맞는 최적화된 알고리즘이 존재한다는 것

- 그러한 알고리즘을 찾기 위해 반복적인 시험(back-and-forth iterations)을 하고 이를 통해 현실의 문제에 최적화된 솔루션을 찾을 수 있다

- “There is no free lunch for the data miner.”

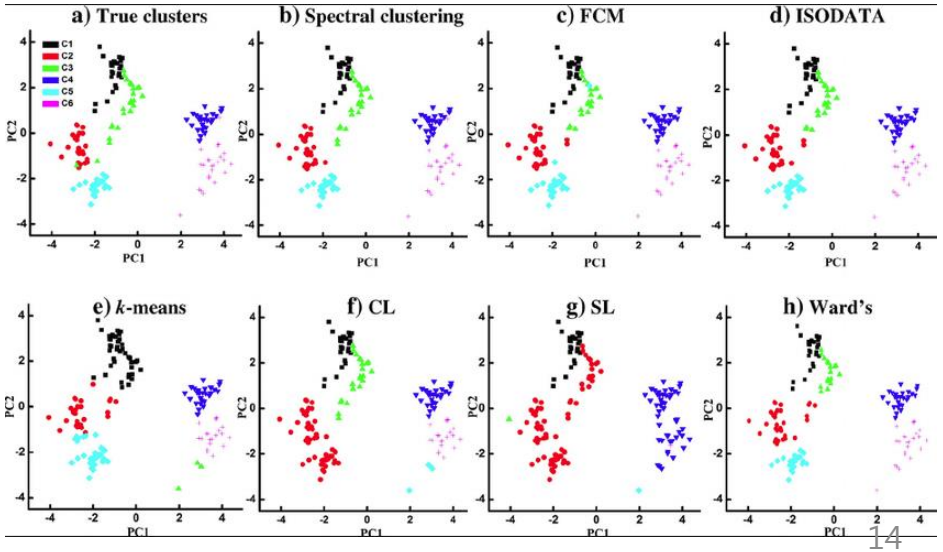
데이터 마이닝 프로세스

· 모델링(Modeling): 데이터 마이닝 테크닉 적용

- 데이터로부터 의미 있는 정보를 추출하기 위해 다양한 데이터 마이닝 테크닉을 적용

- 다양한 테크닉이 적용되고 입력 파라미터가 최적화됨

- 회귀/분류/군집화/연관성 탐색



데이터 마이닝 프로세스

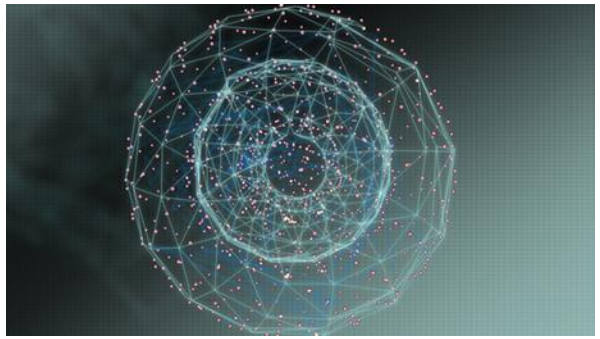
· 데이터 마이닝의 제5법칙: 왓킨스의 법칙

- 모든 비즈니스 관련 데이터에는 의미 있는 정보(patterns)가 존재한다

- 데이터의 생성은 비즈니스의 과정과 함께하므로, 이와 함께 의미 있는 정보가 축적될 수 밖에 없음

- 정보를 찾기 위해서는 비즈니스 지식(business knowledge)로부터 시작해야 한다

- “There will always be patterns for every data mining problem in every domain unless there is no relevant data.”



데이터 마이닝 프로세스

· 평가(Evaluation): 분석 결과를 평가

- 결과 평가는 모델링 만큼이나 중요한 단계임

- 일반적으로 정확도(accuracy)와 오차(error)를 바탕으로 평가되나, 그 외의 다양한 평가 방법이 활용될 수 있음

- 평가 결과 데이터 마이닝 결과의 활용 여부가 결정됨

데이터 마이닝 프로세스

· 적용(Deployment): 분석 결과를 적용함

- 분석을 통해 얻은 정보와 지식을 현실 세계의 문제를 해결하기 위해 적용함

- 정보와 지식은 최종 소비자가 쉽게 이해할 수 있는 형식으로 전달되어야 함(일반적으로 결과의 적용은 데이터 분석가가 아닌 다른 사람이 하는 경우가 많음)

- 지나치게 기술적인 부분보다는 결과와 인사이트 중심으로 전달하는 것이 중요하다!

데이터 마이닝 프로세스

· 데이터 마이닝의 제6법칙: 인사이트의 법칙

- 데이터 마이닝은 비즈니스의 인지 과정(perceptual process)의 일부이다

- 즉, 데이터 마이닝은 인간이 쉽게 발견하지 못하는 정보(pattern)를 데이터로부터 찾아 줌으로써 비즈니스 인지 과정을 돕는다

- 즉, 인공지능은 지능을 증폭시키는 역할(“intelligence amplifier”)을 한다고 할 수 있다

- “Data mining amplifies perception in the business domain.”

데이터 마이닝 프로세스

· 데이터 마이닝의 제7법칙: 예측의 법칙

- 데이터 마이닝의 핵심은 바로 그럴듯한 결과를 “예측(predict)”하는 데에 있다

- 모델을 통해 제한적인 상황 속에서 “예측”을 하고 이를 보다 일반적인 경우로 확장(generalize)한다

- 회귀, 분류, 군집, 연관 분석 등 대부분의 데이터 마이닝 모델은 “예측”과 결부되어 있다

- “Prediction increases information locally by generalisation.”

데이터 마이닝 프로세스

· 데이터 마이닝의 제8법칙: 가치의 법칙

- 데이터 마이닝의 결과는 정확도(accuracy)나 안정성(stability)에 의해서 결정되지 않는다

- 예측 모델의 가치는 개선된 행동(action)을 촉발하고, 새로운 인사이트를 통해 개선된 전략을 발생하는 데에 있다

- “The value of a predictive model is not determined by any technical measure.”

데이터 마이닝 프로세스

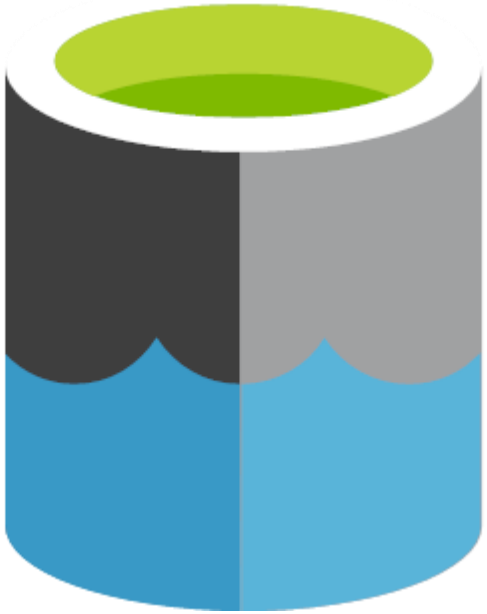
· 데이터 마이닝의 제9법칙: 변화의 법칙

- 데이터 마이닝에 의해 발견된 정보는 영원히 지속되지 않는다

- 데이터가 갖더라도 컨텍스트나 목적에 따라 데이터의 가치나 이를 통해 도출되는 인사이트 등이 얼마든지 달라질 수 있다

- “All patterns are subject to change because they reflect not only a changing world but also our changing understanding.”

데이터 레이크(Data Lake)

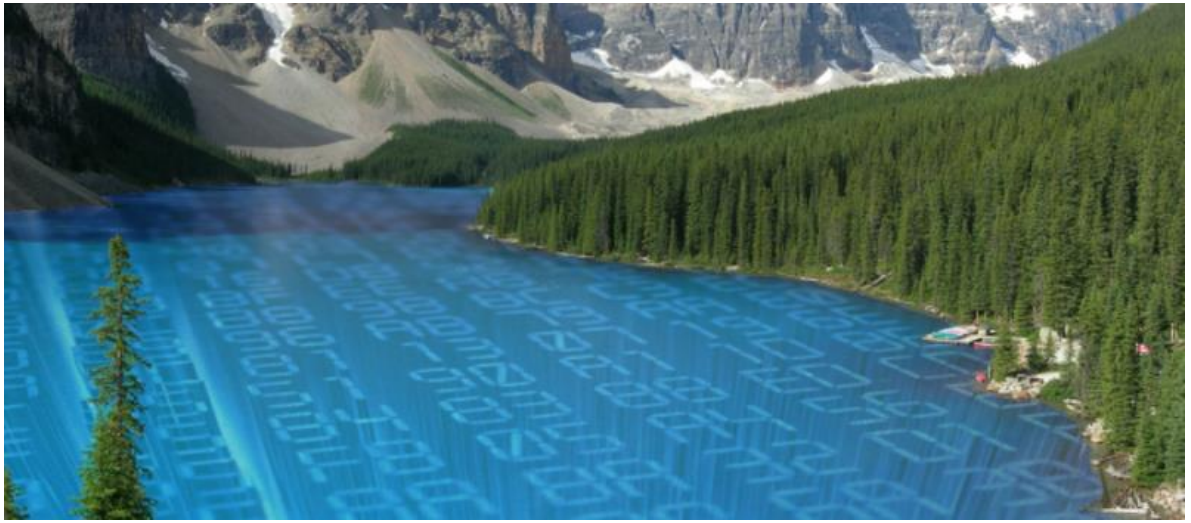


데이터 마이닝의 최근 트렌드

· 데이터 웨어하우스(Data Warehouse)에서 데이터 레이크(Data Lake)로
[source: https://en.wikipedia.org/wiki/Data_lake]

- 데이터가 조직과 사회에서 갖는 중요성이 증가하면서 데이터 마트(Data Mart), 데이터 사일로(Data Silo), 데이터 웨어하우스(Data Warehouse) 등 데이터를 저장하기 위한 다양한 개념들이 등장함

- 앞서 살펴본 4V로 대표되는 “빅 데이터”의 시대가 도래하면서 ‘데이터 레이크’의 개념이 대두되고 있음



데이터 마이닝의 최근 트렌드

· 데이터 웨어하우스(Data Warehouse)에서 데이터 레이크(Data Lake)로

- 데이터 웨어하우스(Data Warehouse): 사용자의 의사 결정에 도움을 주기 위하여, 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스
[source:https://ko.wikipedia.org/wiki/%EB%8D%B0%EC%9D%B4%ED%84%B0_%EC%9B%A8%EC%96%B4%ED%95%98%EC%9A%B0%EC%8A%A4]

- 데이터 레이크: 데이터를 자연적인 형태(natural format)으로 시스템이나 저장소에 저장하는 방법으로, 다양한 데이터의 개요나 구조 간에 호환(collocation)을 가능케 함[source: https://en.wikipedia.org/wiki/Data_lake]

■ 조직과 관련된 로 데이터(raw data) 뿐만 아니라 리포팅, 시각화, 애널리틱스 등의 작업을 위해 활용되는 변형된 형태의 데이터(transformed data)도 함께 포함

■ 관계형 데이터베이스에 적합한 정형 데이터(structured data), CSV, 로그(logs), XML, JSON 등 반정형 데이터(semi-structured data) 및 전자메일, 각종 문서,이미지, 오디오 등 비정형 데이터(unstructured data)를 포함

■ 즉, 모든 형태의 데이터를 포함하는 중앙화된 데이터 저장소이다

■ 앞서 살펴본 아파치 하둡의 분산 파일 시스템(distributed file system)은 데이터 레이크의 대표적인 사례이다

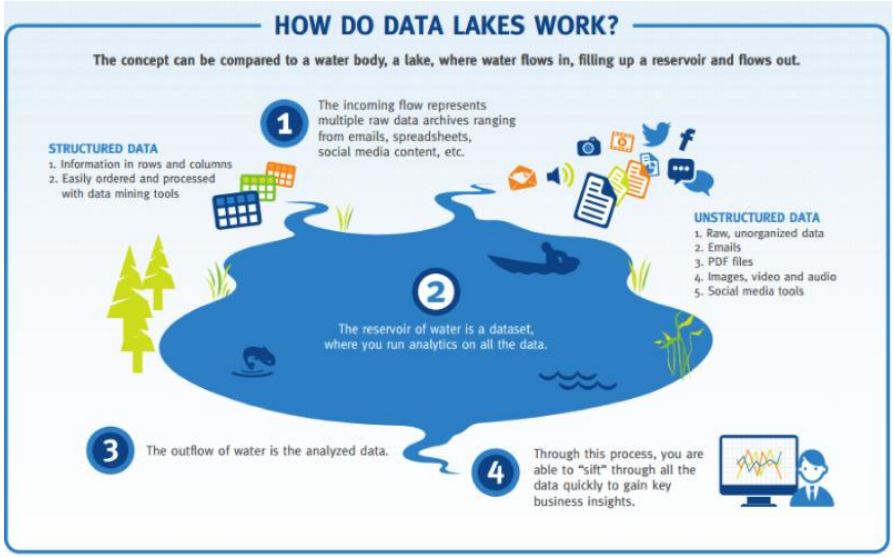
데이터 마이닝의 최근 트렌드

· 데이터 웨어하우스(Data Warehouse)에서 데이터 레이크(Data Lake)로

- 구글, 마이크로소프트, 아마존 등 대부분의 선도 IT 기업들은 이미 클라우드(cloud) 서비스, 분산 처리 등 데이터 레이크와 관련된 산업에 뛰어들고 있는 추세임[source: https://blog.equinix.com/blog/2016/11/10/why-companies-are-jumping-into-data-lakes/]

- 데이터 레이크가 각광을 받고 있는 이유는 바로 “빅 데이터” 시대를 맞아 데이터 애널리틱스(data analytics)에 대한 수요가 급격히 증가했기 때문임[source: https://infocus.emc.com/william_schmarzo/why-do-i-need-a-data-lake-for-big-data/]

■ 데이터를 분석하고 활용함에 있어 다양한 데이터 구조(data structures)가 지원되면 상당한 이점을 가짐



데이터 마이닝의 최근 트렌드

· 데이터 웨어하우스(Data Warehouse)에서 데이터 레이크(Data Lake)로

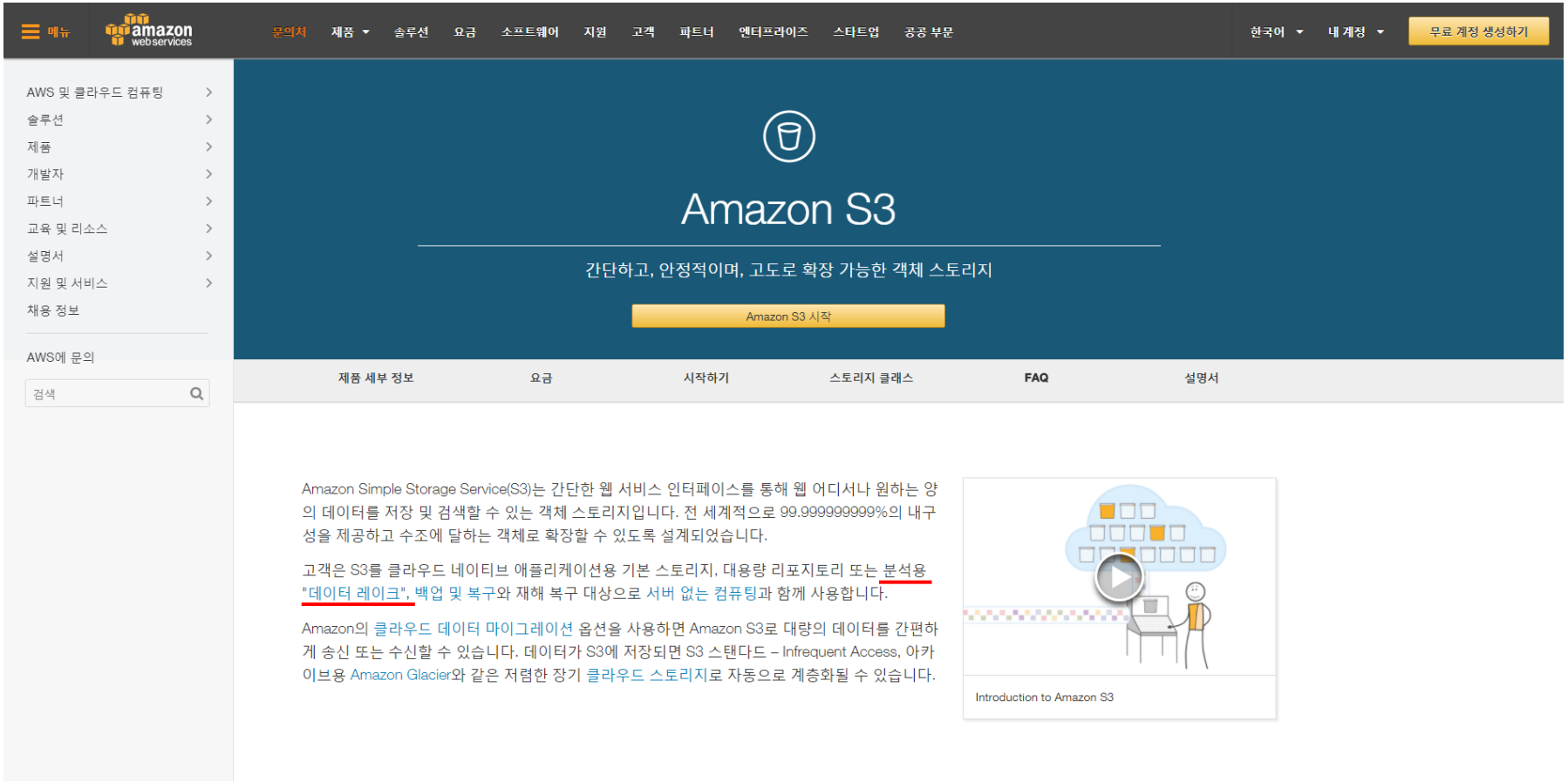
- 데이터 웨어하우스 vs. 데이터 레이크[source: http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html]

데이터 웨어하우스	vs.	데이터 레이크
정형 데이터 처리된(processed) 데이터	**데이터 형태**	정형/비정형/반정형 데이터 로 데이터/처리된 데이터
데이터를 읽어오기 전에 스키마를 미리 정의(schema-on-write)	**데이터 프로세싱**	데이터를 읽어오면서 스키마를 정의 (schema-on-read)
다량의 데이터를 저장하기엔 비효율적	**데이터 저장**	다량의 데이터를 저렴하게 저장하기에 최적화
기민성이 떨어짐 초기 설정을 바꾸기 힘듬	**유연성**	매우 기민함 초기 설정을 바꾸고 새로운 설정을 추가하기가 용이
안정	**보안**	불안정
비즈니스 전문가	**주 사용자**	데이터 사이언티스트

데이터 마이닝의 최근 트렌드

· 데이터 레이크 시스템 예시 - 아마존 S3 (Amazon S3)[source: https://en.wikipedia.org/wiki/Amazon_S3]

- 간단한 웹 인터페이스를 통해 웹 어디서나 데이터를 저장 및 검색할 수 있는 객체 스토리지(object storage)



데이터 마이닝의 최근 트렌드

· 데이터 레이크 시스템 예시 - 아마존 S3 (Amazon S3)[source: https://en.wikipedia.org/wiki/Amazon_S3]

- 아마존 S3에 저장된 객체의 숫자는 지난 2013년 2조를 넘어서며 기하급수적으로 증가하고 있다

- 웹 호스팅, 이미지 호스팅, 백업 시스템 저장소 등 다양한 기능을 제공

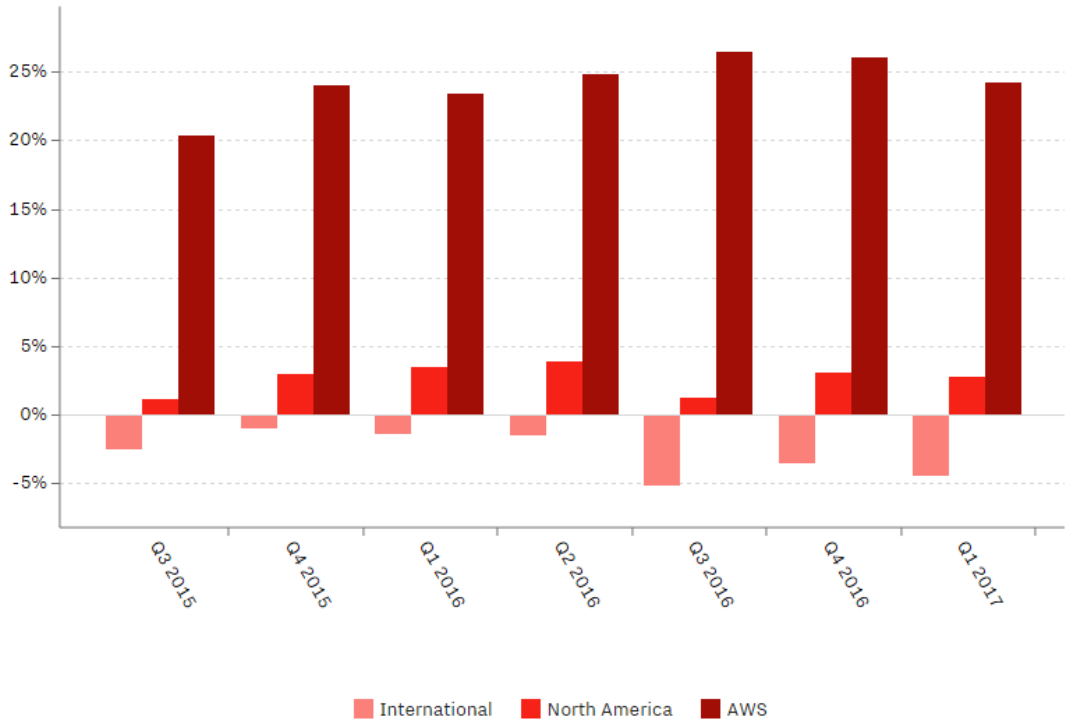
- 에어비엔비(AirBnb), 나스닥(Nasdaq), 넷플릭스(Netflix) 등 다량의 데이터를 활용하는 조직들은 이미 아마존 S3의 서비스를 활용하고 있음

데이터 마이닝의 최근 트렌드

· 데이터 레이크 시스템 예시 - 아마존 S3 (Amazon S3)

- 아마존 웹 서비스 부문은 최근 10년 간 조직과 사회에서 운용되는 데이터의 양이 증가하면서 급격히 성장해 2017년 1분기 현재 영업이익률 24.3%를 기록하며 아마존이 지속적인 흑자를 기록하는 데 큰 기여를 하고 있음
[source: https://www.recode.net/2017/4/27/15451726/amazon-q1-2017-earnings-profits-net-income-cash-flow-chart]

- 아마존의 부문별 영업이익률



데이터 마이닝 적용 예시 – 야구장 관객 수 예측하기

비즈니스 이해

· 야구장 운영은 구단 유지비, 구장 사용료 등 고정 비용(fixed cost)가 큰 사업으로, 야구장에 관객이 많이 찾지 않는 날은 적자를 감수할 수 밖에 없는 구조이다

- 야구 경기의 경우 악천후나 국가 비상 상태 등 특수한 상황이 아니면 반드시 스케줄에 맞추어 경기를 진행해야 한다

· 관객이 적게 온다고 경기를 취소하거나 연기할 수 없으므로, 특정 조건을 갖춘 날 관객이 얼마나 찾는지 예측할 수 있다는 유의미한 정보가 될 수 있다

· 그러한 정보를 바탕으로 수익 극대화(Profit maximization)을 위한 경영 전략(Corporate strategy)를 마련할 수 있다

- 관객이 적게 찾아올 것 같은 날 프로모션을 진행하거나 연예인 시구를 하는 등의 이벤트를 통해 수익을 늘리면서 동시에 관객이 많이 찾아올 것 같은 날은 별도의 비용을 들이지 않아도 된다

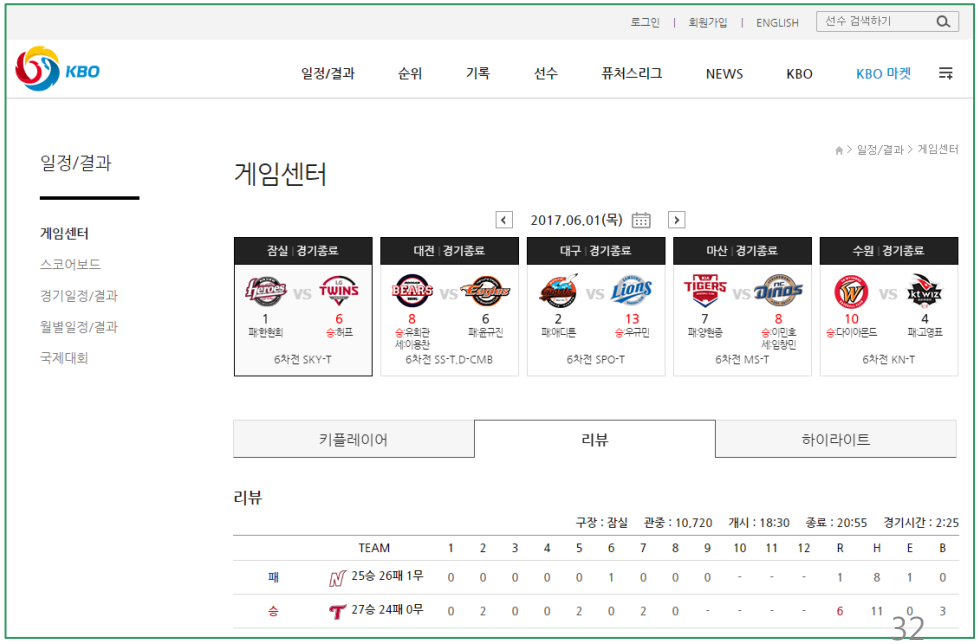
데이터 이해

· 데이터 수집

- 2010년부터 2014년까지 두산 베어스와 LG 트윈스의 잠실야구장 홈경기 데이터를 수집(총 616경기)

- 한국 야구위원회(KB) 홈페이지에서 데이터 수집(Web Crawling)

- 경기 별 관객, 당시 성적, 당시 선발투수 기록, 휴일 여부 관련 데이터 수집

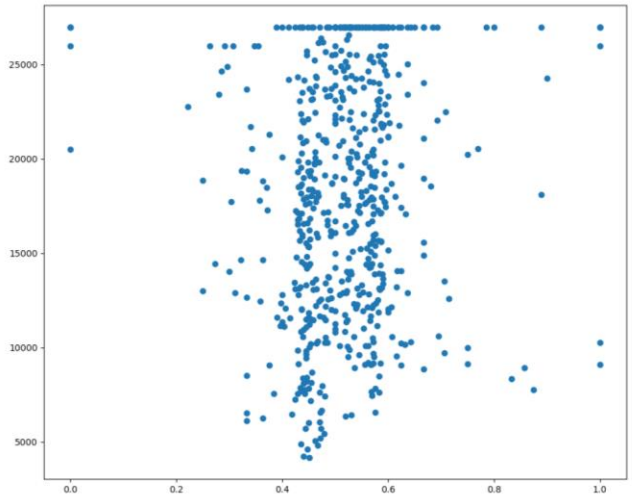


데이터 이해

· 데이터 탐색

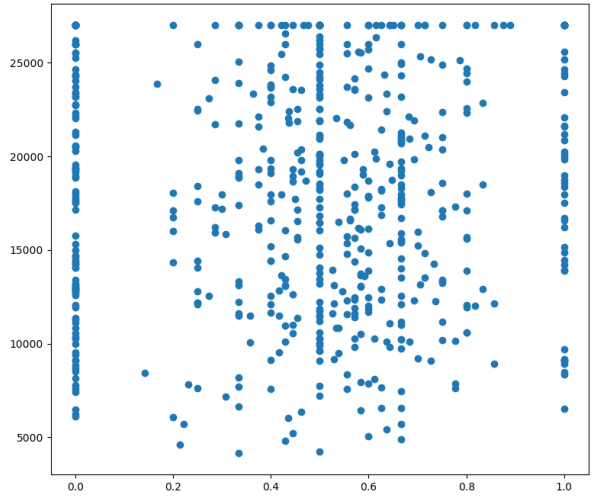
- 데이터 시각화

● 홈 팀 승률과 관객 수 간의 관계



상관계수 = 0.06923221

● 홈 선발투수 승률과 관객 수 간의 관계



상관계수 = -0.01201143

데이터 준비

· 원본 형식의 데이터를 엑셀 데이터프레임에 맞게 가공

- 분석 툴에 넣어 바로 분석을 수행할 수 있도록 가공

	K	L	M	N	O	P	Q	R
1	ome_p_	away_w	away_dr	away_lo	away_p_v	away_p_l	holida	audienc
2	0	0	0	0	0	0	1	27,000
3	0	0	0	1	0	0	1	27,000
4	0	2	0	0	0	0	0	10,830
5	0	2	0	1	1	0	0	6,106
6	1	3	0	1	1	0	1	13,023
7	1	3	0	2	0	1	1	12,808
8	0	2	0	4	0	0	0	8,345
9	0	2	0	5	0	0	0	8,928
10	0	2	0	6	0	2	0	7,774
11	0	3	0	6	0	1	0	18,126
12	0	3	1	6	0	0	1	27,000
13	0	3	1	7	0	1	1	24,281
14	2	9	0	4	1	0	0	6,244
15	0	10	0	4	1	0	0	6,520
16	2	10	0	5	1	1	0	7,571
17	0	6	0	9	0	1	0	20,543
18	1	6	0	10	1	1	1	27,000
19	0	6	0	11	2	0	1	27,000
20	1	13	0	5	1	1	0	10,006
21	1	14	0	5	0	1	0	9,722
22	1	8	0	13	1	0	0	9,116
23	0	9	0	13	0	2	1	22,099
24	3	9	0	14	0	4	1	19,226
25	2	13	0	12	1	1	0	10,287
26	1	8	0	18	1	3	0	10,607
27	1	8	0	19	0	0	1	22,511
28	0	9	0	19	0	3	1	18,562
29	3	17	1	9	1	1	0	23,679

모델링

· 변수

- 출력변수

● 경기별 관객수(y): min = 0; max = 27,000

- 입력변수

● 홈 팀 승률($x1$)

● 어웨이 팀 승률($x2$)

● 홈 팀 선발투수 승률($x3$)

● 어웨이 팀 선발투수 승률($x4$)

● 휴일 여부($x5$): 휴일이면 1, 휴일이 아니면 0

모델링

· 가설: 선형회귀모형

$$y = a_1x_1 + a_2x_2 + \ldots + a_5x_5 + b$$

- 목적: 비용을 최소화하는 a와 b를 찾는다

· 결과

- 추정값: 데이터를 통해 회귀 모형을 돌린 결과 아래와 같은 값이 추정됨

- $b = 10615.2675$
- $a_1 = 6042.6753$
- $a_2 = 2349.4648$
- $a_3 = 281.82658$
- $a_4 = 570.8495$
- $a_5 = 7090.2335$

모델링

· 결과 해석

- 모든 승률이 0이고 휴일이 아닌 날의 관객 추정치는 약 10615 명이다

- 홈 팀의 승률이 1% 올라갈 때 마다 관객이 약 60명 증가한다

- 홈 팀의 선발투수의 승률이 1% 올라갈 때 마다 관객이 약 23명 증가한다

- 어웨이 팀의 승률이 1% 올라갈 때 마다 관객이 약 3명 증가한다

- 어웨이 팀의 선발투수의 승률이 1% 올라갈 때 마다 관객이 약 6명 증가한다

- 휴일이 아닐 때보다 휴일일 때 관객이 약 7000 명 정도 많이 온다

평가

· 학습 데이터 셋(training dataset)과 시험 데이터 셋(test dataset)의 비율을 7:3으로 나누어 오차를 계산함

- 평균 제곱근 오차(RMSE, Root mean squared error)로 계산

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y)^2}$$

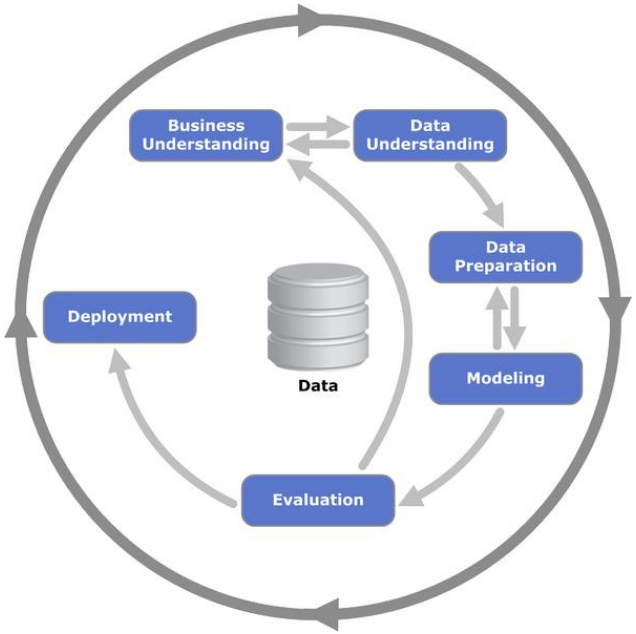
- RMSE = 5,164. 즉, 경기당 약 5,000명 정도의 오차가 발생

적용

· 모델링 결과와 평가 결과를 바탕으로 데이터 마이닝 모델을 적용할 것인지를 결정

· 모델을 바로 적용할 수도 있고, 좀 더 보완하여 적용할 수도 있음

- 데이터, 변수, 테크닉 등을 보강하여 새로운 결과 도출 가능



References

· Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

· 9 laws of data mining. http://khabaza.codimension.net/index_files/9laws.htm