3 주차 강의

1. 분석프로세스

2022.03.22

Institutor: JS LEE

Wrap Up

데이터 마이닝은

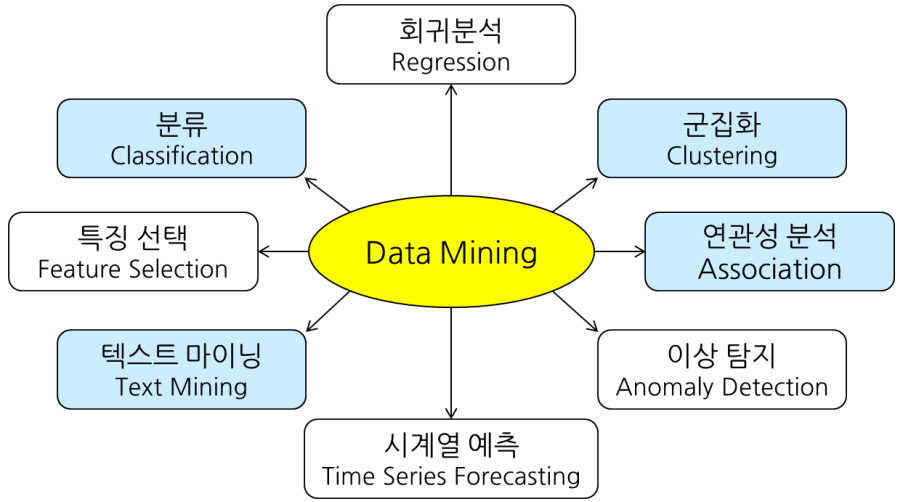
- ✓ 대규모 데이터에 대한 귀납적 추론
- ✓ **데이터 속의 유용한(의미 있는) 패턴(규칙, 관계)을 찾고 이를 일반화하는 프로세스**
- ✓ 관측 데이터에 적합한 모델을 구축하는 과정

데이터 마이닝 유형

	지도학습 (supervised learning)	자율학습 (unsupervised learning)
의미	• 학습용 데이터를 기준으로 모델을 만들고 이를 새로운 데이터에 적용하여 예측분석에 이용 • 입력변수들을 기준으로 타겟(출력, 결과)변수 예측	• 데이터 포인트들 간의 관계를 기반으로 데이터에서 패턴을 찾아내는 작업
특징	• 타겟(출력, 결과)변수 존재함	• 타겟(출력, 결과)변수 존재하지 않음
분석 기법	• 신경망, 회귀분석, 의사결정나무, 판별분석, 로지스틱회귀분석 …	• 군집분석, 연관규칙, …

Wrap Up

분야	설명	알고리즘	사례
분류	•데이터 포인트가 미리 정의된 클래스 중 어디에 속하는지에 대해 예측 •예측은 학습용 데이터셋을 기반으로 함	•**의사결정나무,Random Forest, Xgboost,** 신경망 베이지안 모델, 규칙 유도, k-최근접 이웃	•유권자들을 정당에 따라서 알려진 버킷으로 할당 •새 고객을 정의된(이미 알려진) 고객 그룹 중 하나의 그룹에 할당
회귀 분석	•수치형 타겟변수를 예측 •예측은 학습용 데이터셋을 기반으로 함	•선형회귀	•내년도 실업률 예측 •보험료 추정
이상 탐지	•특정 데이터 포인트가 데이터셋의 다른 데이터 포인트와 비교하여 특이값인지 예측	•거리 기반, 밀도 기반, 지역 특이값 요소(LOF)	•신용카드의 사기거래 탐지, 네트워크 침입 탐지

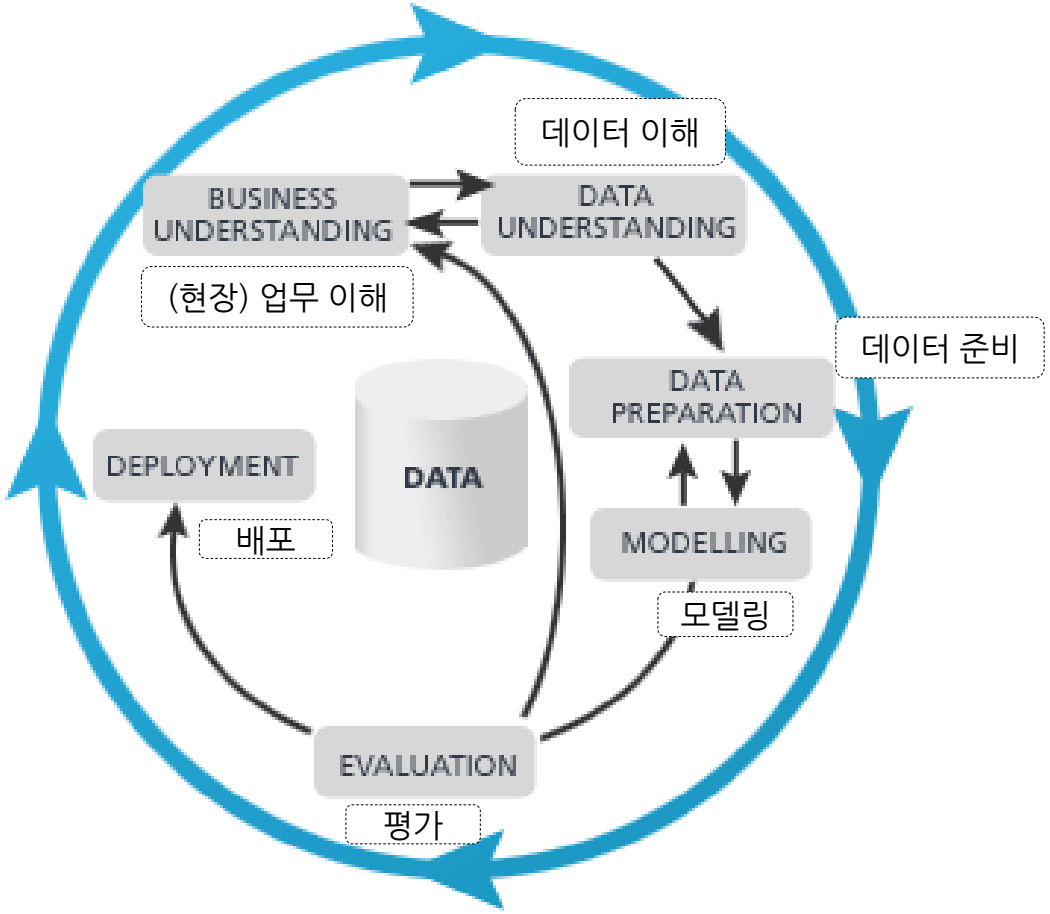


분야	설명	알고리즘	사례
시계열 분석	•과거 값에 기반하여 미래의 타겟변수 값을 예측	•지수평활, 자기회귀 누적 이동평균(ARIMA), 회귀분석	•매출액 예측, 생산 예측, 추정할 필요가 있는 성장 현상
군집화	•데이터셋 내의 속성들을 기준으로 하여 데이터셋의 데이터 포인트들을 군집으로 구별	•**k-평균**, 밀도 기반 군집화 (예: DBSCAN)	•거래, 웹 및 고객 통화 데이터를 기반으로 한 고객 세분화
연관성 분석	•거래 데이터를 기반으로 항목집합 내의 관계를 식별	•**빈발패턴-성장 알고리즘 (FP-Growth),** 선험적 (Apriori) 알고리즘	•소매업에서 구매 이력 데이터를 기반으로 한 교차 판매 기회 발견

분석 프로세스

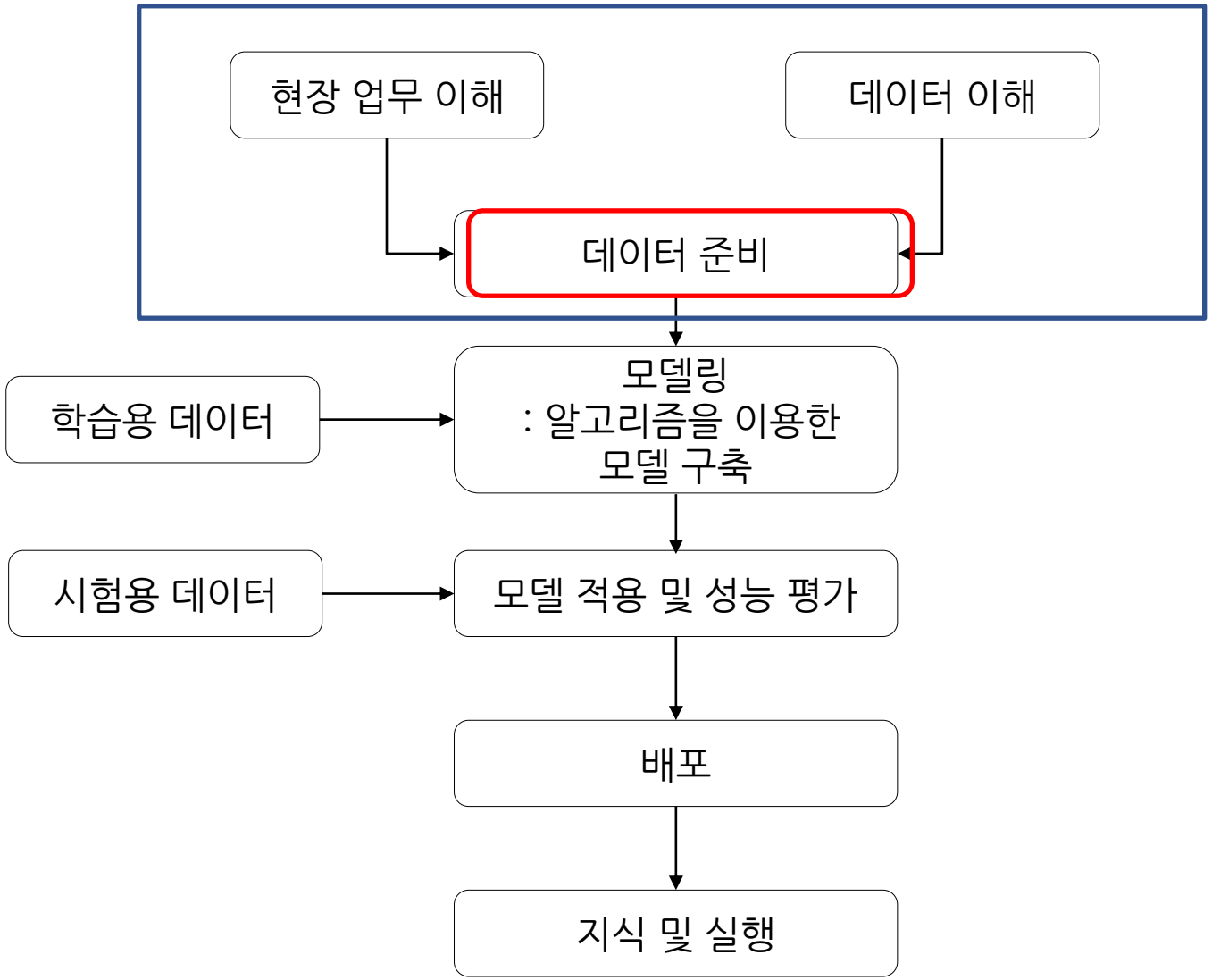
대표적인 데이터 마이닝 프레임워크

CRISP-DM(Cross Industry Standard Process for Data Mining)



다른 프레임워크
- SEMMA(Sample, Explore, Modify, Mo
- DMAIC(Define, Measure, Analyze, Improve, Control)
- KDD(Knowledge Discovery in Databases, Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation framework)

분석 프로세스



* 전체 프로세스에서 시간이 가장 많이 소요되는 부분은?

분석 프로세스

Play the role of a Data Scientist.

If you've made the decision to transition into a Data Scientist, you must have done a lot of extra reading to fully understand what it entails to become a Data Scientist. You will go from describing trends in your data to uncover new data using your existing data and build machine learning models to support your hypothesis.

Data Scientists:

- Spend a lot of their time cleaning data using languages like Python or R. → 데이터 준비
- Build predictive models using machine learning algorithms such as gradient boosting, linear regression, logistic regression, decision trees, Random Forest, and more. → 모델링
- Evaluate the models they create to get a high percentage accuracy in order to validate the analysis → 평가
- Test and improve the accuracy of already built ML models.
- Build visualizations to narrate the advanced analysis result.

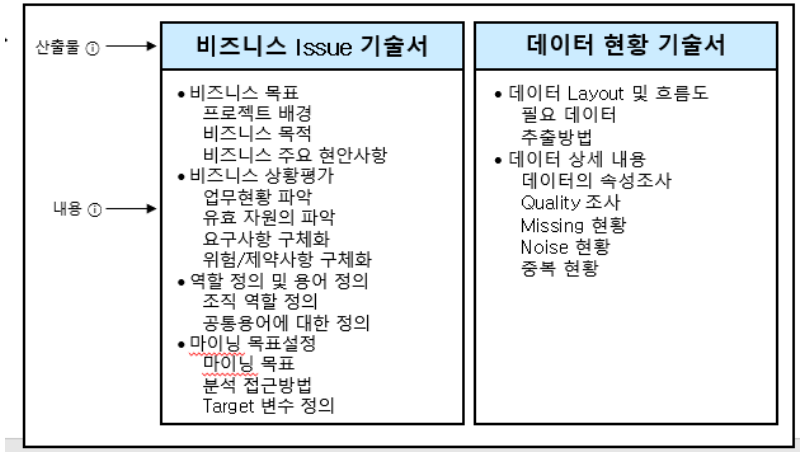
분석 프로세스

1. Business Understanding : 업무이해

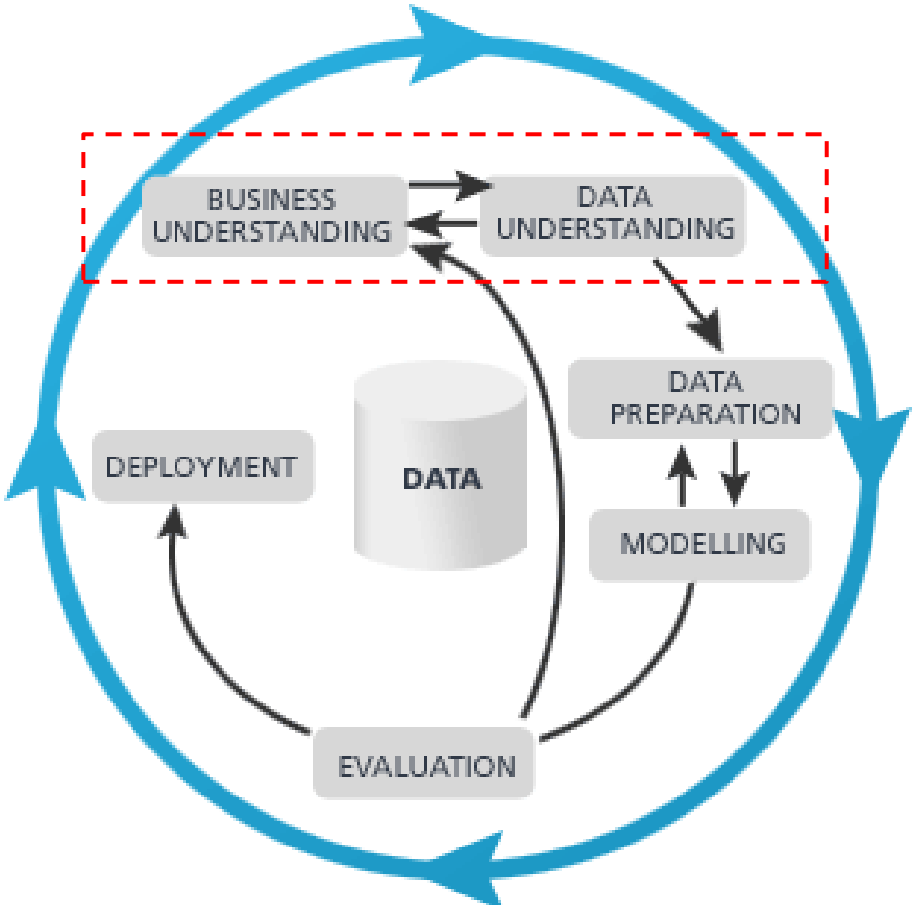
- 분석을 통해 무엇을 얻고자 하는지에 결정.
- 비즈니스 배경 지식 이해
- 분석 주제 선정
- 분석 방향성 선정 : 분석 대상 고객, 분석 기간, 분석에 사용할 데이터의 범위 등
- 분석의 성공의 기준 정의 등
- 분석결과 활용은 어떻게 할 것인지?

2. Data Understanding : 데이터 이해 - 계속

- 어떠한 데이터가 존재하는지?
- 데이터 양은 어느 정도인지?
- 데이터의 품질은 어떠한지?
- 분석에 필요한 데이터가 가용한지?



분석 프로세스



현장을 모르는 전략가의 위험

" MilkShake Mistake"

분석 프로세스

2. Data Understanding : 데이터 이해 - 계속

“새로운 자산으로 떠오르는 다크 데이터 (Dark Data)”

- **다크 데이터**(dark data): 다양한 컴퓨터 네트워크 운영을 통해 얻는 데이터이지만 의사 결정이나 이해를 위한 수단으로 사용되지 않음.

- ‘다크 데이터(dark data)’는 사진, 동영상, 음성 등 분석이 어려운 비정형 데이터로 역시 빅데이터의 한 종류라고 할 수 있으나, 저장은 되어있으나 구조화되지 않고 다른 데이터와 상호 작용이 없는 데이터 혹은 사용할 수 없는 데이터를 의미함

〈 Dark Data 〉

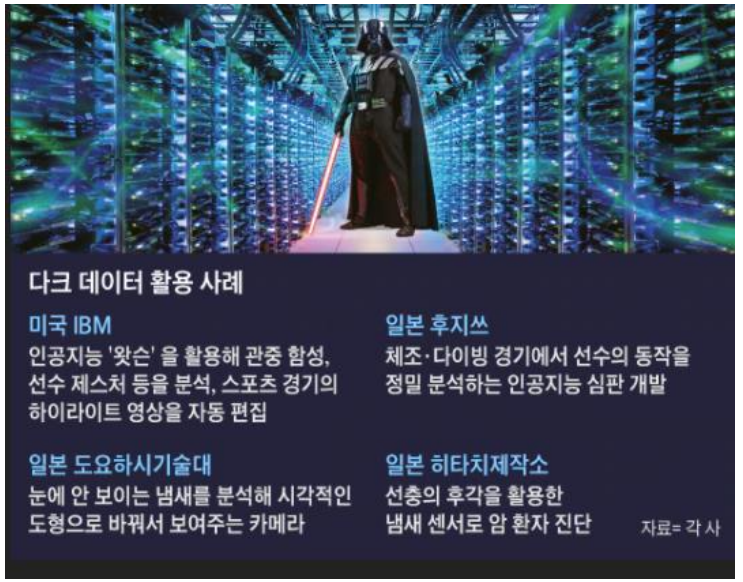
분석 프로세스

2. Data Understanding : 데이터 이해 - 계속

- 데이터는 도처에 존재한다
- 전 세계적으로 매일 250경 바이트의 데이터가 생성된다 (IBM)
- 그러나 활용되는 데이터는 극소수
- 존재하는 데이터의 90%는 암흑데이터

기업 내부의 데이터 외에 공공데이터 및 수 많은 데이터들이 저장됨에 따라 창의적 데이터 활용이 중요함

〈 Dark Data 활용사례〉



분석 프로세스

2. Data Understanding : 데이터 이해 - 계속

1) 조직 내의 데이터 : 조직에 의해 관리되고 생산되는 데이터

예. 삼육대

- 학사정보, 교내 활동 정보, 취업 정보, 도서관 이용정보 등
  출석정보, E-class 이용정보 등
  여러분들이 모든 발자취가 데이터가 됨

재학생 충원율
(정원내 재학생수/(학생정원-학생모집정지인원)×100%)

전체 재학생수	5,464 명
정원내 재학생수	4,832 명
학생정원	4,644 명
학생모집정지인원	0 명
충원율(정원내)	104.0 %

ⓘ 자료기준일: 2021.4.1.

취업률
(취업자수/(졸업자수-제외자수)×100%)

졸업자수	1,276 명
취업자수	772 명
제외자수	110 명
취업률	66.2 %

ⓘ 자료기준일: 2020.12.31

신입생 충원율
(정원내 입학자수/정원내 모집인원×100%)

전체 입학자수	1,207 명
정원내 입학자수	1,121 명
정원내 모집인원	1,121 명
충원율(정원내)	100.0 %

ⓘ 자료기준일: 2021년3월학기 신입생

학생 1인당 교육비
(총교육비/재학생수)

총교육비	76,338,160,967 원
재학생수	5,975 명
학생 1인당 교육비	12,776,261.2 원

ⓘ 자료기준일: 2020년 결산, 재학생 : 2019.4.1. 학부+대학원

전임교원 확보율
(전임교원수/교원 법정정원×100%)

전임교원수(재학생기준)	221 명
재학생수	6,216 명
교원 법정정원(재학생기준)	291 명
전임교원 확보율(재학생기준)	75.95 %

ⓘ 자료기준일: 2021.4.1. 학부+대학원

학생 1인당 장학금
(장학금총액/재학생수)

재학생수	5,368 명
장학금 총액	21,224,506,104 원
교내 + 사설 및 기타	8,434,974,123 원
학생 1인당 장학금	3,953,894.6 원

ⓘ 자료기준일: 2020년기준, 재학생 : 2020.4.1., 10.1. 합계/2

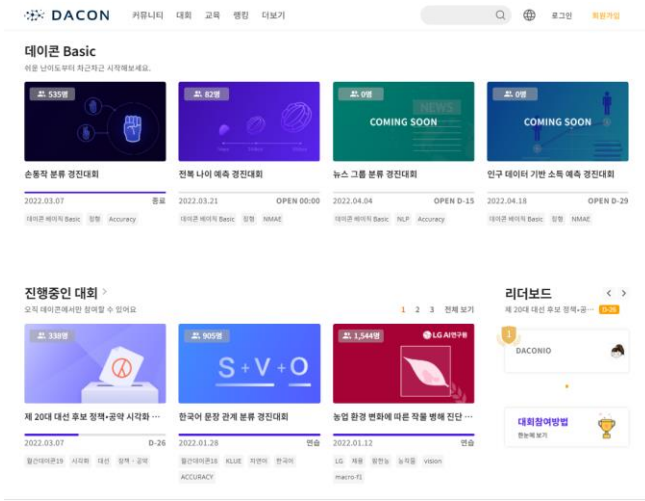
2) 그 외 데이터들을 얻을 수 있는 곳

① 공공데이터 포털에 굉장히 많은 유용한 데이터들 존재
- https://www.data.go.kr/index.do

② 지자체나 공공기관 등에서 각종 '(빅)데이터 포털 운영
(예: 서울 열린 데이터광장, 경기데이터드림, 보건의료 빅데이터 개방시스템 등등)

③ 경진대회에서 제공하는 데이터

- Kaggle: 기업이나 단체에서 데이터와 해결과제 등록하면 참가자들이 데이터분석을 위한 모델을 개발하고 경쟁
- Dacon: Kaggle과 유사한 국내 경진대회 플랫폼

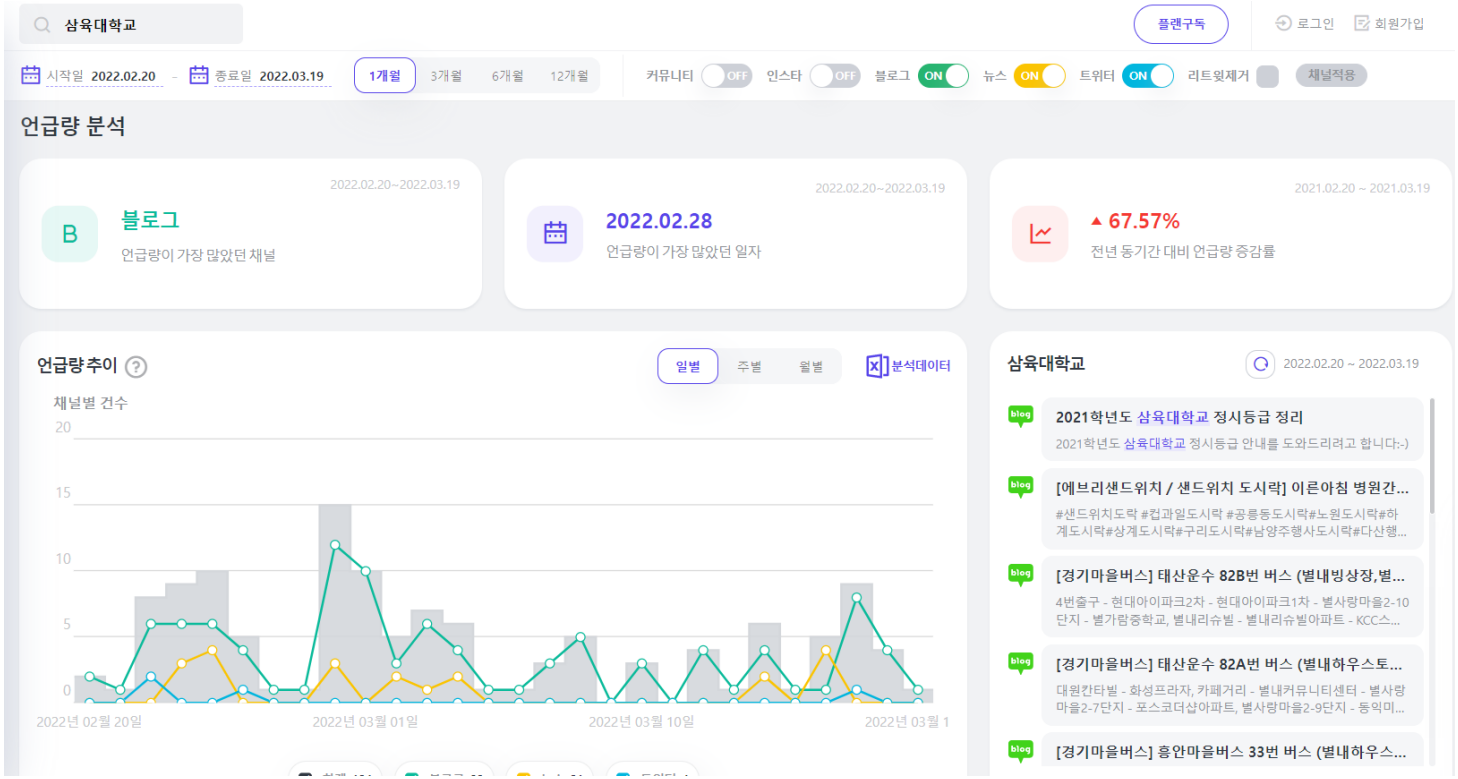


분석 프로세스

2. Data Understanding : 데이터 이해 - 계속 2) 그 외 데이터들을 얻을 수 있는 곳

④ SNS에서 데이터를 획득, 가공, 판매하는 업체

- 썸트렌드 (https://insight.some.co.kr), 네이버데이터랩 (https://datalab.naver.com), 구글트렌드(https://trends.google.co.kr) 등
- 다만 무료로 데이터를 제공하는 범위나 방식은 제한적



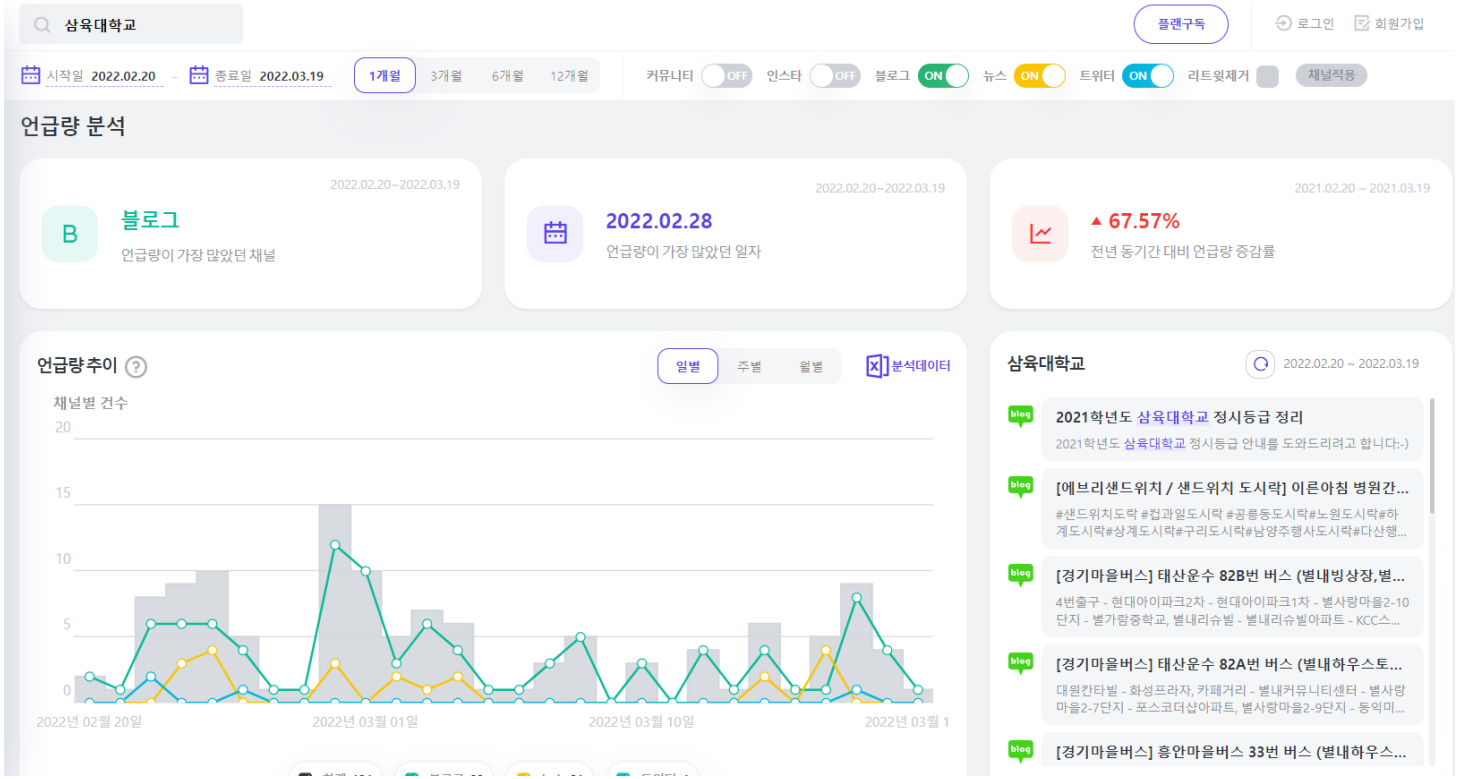
썸트렌드로 삼육대학교 분석한 결과

분석 프로세스

2. Data Understanding : 데이터 이해 - 계속 2) 그 외 데이터들을 얻을 수 있는 곳

④ SNS에서 데이터를 획득, 가공, 판매하는 업체

- 썸트렌드 (https://insight.some.co.kr), 네이버데이터랩 (https://datalab.naver.com), 구글트렌드(https://trends.google.co.kr) 등
- 다만 무료로 데이터를 제공하는 범위나 방식은 제한적



썸트렌드로 삼육대학교 분석한 결과

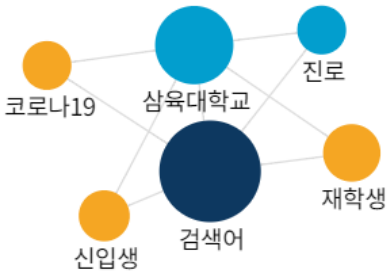
분석 프로세스

2. Data Understanding : 데이터 이해 - 계속 2) 그 외 데이터들을 얻을 수 있는 곳

⑤ 신문기사 데이터 얻을 수 있는 곳

- 빅카인즈 (https://www.bigkinds.or.kr/)

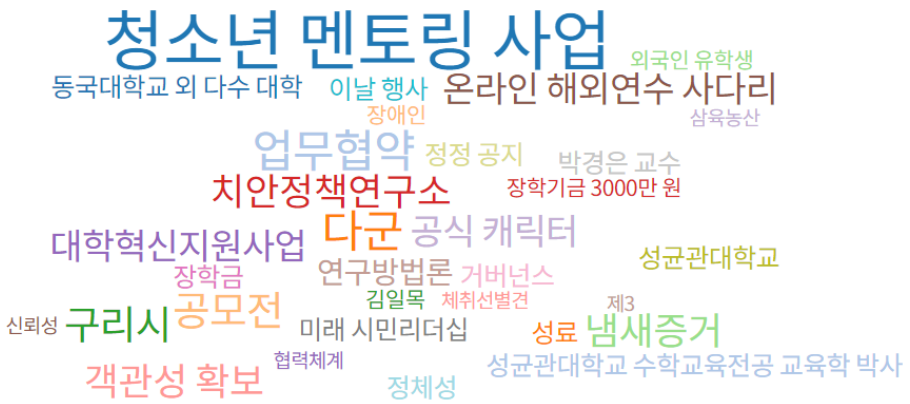
최근 3개월간 삼육대학교 뉴스 검색 결과



데이터 다운로드 | 관계도 분석 | 키워드 트렌드 | 연관어 분석 | 정보 추출

검색한 뉴스의 메타데이터(언론사, 기고자, 제목 등)와 개체명(인물, 기관, 장소 등) 분석 데이터를 엑셀파일로 제공하는 서비스입니다.
데이터 다운로드는 최대 20,000건의 데이터가 다운로드 됩니다. 미리보기는 최대 20개까지 보여집니다.
'키워드' 항목은 본문 내에서 추출된 키워드 중 단순 숫자(1, 2, 2018, 2019 등), 이메일 주소, 시간을 뜻하는 단어(밤, 낮, 새벽 등)를 제외한 결과가 표시됩니다.

	뉴스 식별자	일자	언론사	기고자	제목	통
1	01100301.20220316082943003	20220316	내일신문	장세풍 기자 spjang@naeil.com	삼육대, 공식 캐릭터 디자인 공모전 개최	문화
2	02100201.20220316163356001	20220316	머니투데이	권현수|기자|	삼육대, 공식 캐릭터 디자인 공모전 열어	지
3	01100901.20220316162413001	20220316	중앙일보		삼육대, 공식 캐릭터 디자인 공모전 개최	문화
4	02100201.20220314172226001	20220314	머니투데이	권현수|기자|	삼육대, SU인재상 제도 마련...'봉사왕' 12명 선발	사회
5	02100201.20220304151727001	20220304	머니투데이	권현수|기자|	삼육대, 대학일자리플러스센터 거점형 운영대학 선정	경제
6	02100201.20220302160820001	20220302	머니투데이	권현수|기자|	삼육대, 특허청 '지식재산 전략적 활용 지원사업' 선정	지
7	02100201.20220302140635001	20220302	머니투데이	권태혁|기자|	삼육대, 2022학년도 입학식 진행...1220명 입학	사회
8	02100101.20220228164453001	20220228	매일경제	박나은	"후배들에게 도움이 되고 싶어요" 모교에 장학금 1천만원 기부한 '사회초년생'	사회
9	02100201.20220228111120001	20220228	머니투데이	권태혁|기자|	삼육대 정소영 동문, 대학 발전기금 1000만 원 기탁	사회
10	02100201.20220224174055001	20220224	머니투데이	권현수|기자|	삼육대, 'SW 산학협력 네트워크 데이' 열어	IT_
11	02100201.20220224170655001	20220224	머니투데이	권현수|기자|	삼육대-삼육식품, 연구협업센터 공동 운영 협약 체결	지



분석 프로세스

2. Data Understanding : 데이터 이해 - 계속

C A D

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

B

➢ 용어

A: 속성(attribute), 변수(variable), 요인(factor), 필드(field), 특징/특성(feature), 열(column)

B: 사례(case, example, instance), 표본(sample), 관찰치(observation), 레코드(record), 행(row)

C: 입력(input), 예측변수(predictor), 독립변수(independent variable): 출력에 영향을 주는 변수, 보통 X로 표기

D: 출력(output), 반응(response), 종속변수(dependent), 성과/결과변수(outcome), 목표변수(target variable), 레이블(label): 지도 학습으로 예측되는 변수, 보통 Y로 표기

분석 프로세스

2. Data Understanding : 데이터 이해 - 계속

- 우리가 사용하는 데이터의 80~90%는 비정형 데이터다.” - Merrill Lynch
- 어떠한 데이터 유형이든 분석을 위해서는 수치화 해야 함.

데이터 유형	설명
정형데이터 (Structured Data)	• 정해진 규칙(Rule)에 맞게 구성된 데이터 • 관계형 데이터베이스 시스템의 테이블과 같이 고정된 컬럼에 저장되는 데이터와 파일, 또는 지정된 행과 열에 의해 데이터의 속성이 구별되는 스프레드시트 형태의 데이터
비정형데이터 (Unstructured Data)	• 정해진 규칙이 없어서 값의 의미를 쉽게 파악하기 힘든 경우 (예> 텍스트, 음성, 이미지 등)

<정형데이터 유형>

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

Employed (취업여부) 와 Write off (대손 상각)은 정형데이터 유형일까 비정형 데이터 유형일까?

분석 프로세스

2. Data Understanding : 데이터 이해

각 문장의 단어를 구분하여, 그 구분된 데이터를 하나의 변수화 함

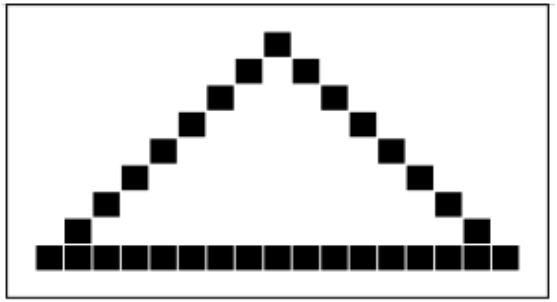
문서1	This is a book on data mining.
문서2	This book describes data mining and text mining using RapidMiner.

토큰 => 칼럼
토큰이 나온 횟수 => 셀
문서 =>레코드

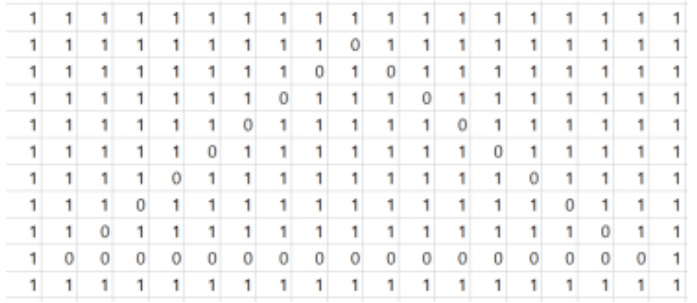
[표9.1] 비-구조화형식의 텍스트의 단어행렬

	this	is	a	book	on	data	mining	describes	text	rapidminer	and	using
문서1	1	1	1	1	1	1	1	0	0	0	0	0
문서2	1	0	0	1	0	1	2	1	1	1	1	1

이미지는 수많은 픽셀로 이루어지고, 픽셀은 RGB 색상 값을 가짐.
이미지는 픽셀 * 색상 값의 분포로 표현



11 x 19 픽셀

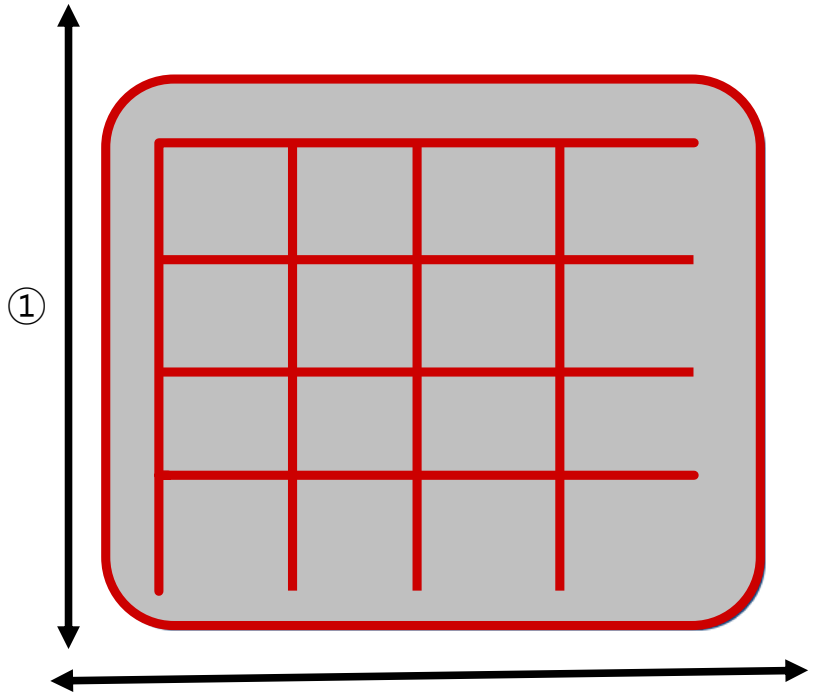


11 x 19 행렬

분석 프로세스

3. Data Preparation : 데이터 준비 : 계속

- 데이터 탐색(data exploration)
- 데이터 품질(data quality)
- 결측치(missing value)
- 특이값(Outlier)
- 데이터 유형 및 전환(data type and conversion)
- 변환(transformation)
- 특징 선택(feature selection)
- 데이터 샘플링(data sampling)
- 새로운 변수생성



① 분석에 누구를 포함시킬 것인가?

② 어떠한 정보를 기준으로 분석할 것인가?

More generally, data scientists may spend considerable time early in the process defining the variables used later in the process. This is one of the main points at which human creativity, common sense, and business knowledge come into play. Often the quality of the data mining solution rests on how well the analysts structure the problems and craft the variables (and sometimes it can be surprisingly hard for them to admit it).

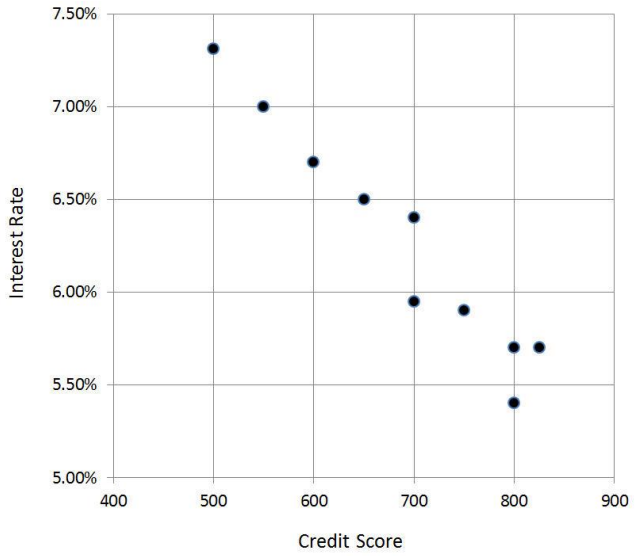
Refer :Data Science for Business)

분석 프로세스

3. Data Preparation : 데이터 준비 : 계속

◆ 데이터 탐색(Data Exploration) : 데이터의 전반적인 이해

- 탐색적 데이터 분석(EDA, Exploratory Data Analysis)
- 기술 통계량 계산, 데이터 시각화
- 데이터의 구조, 값들의 분포, 극단 값의 존재, 데이터셋 안의 상관관계
- 데이터 요약 제공
  - 평균, 중간값, 최빈값, 표준 편차, 각 속성의 범위 등
- 산점도 등으로 시각화 하여 데이터의 이해가 용이하도록 지원

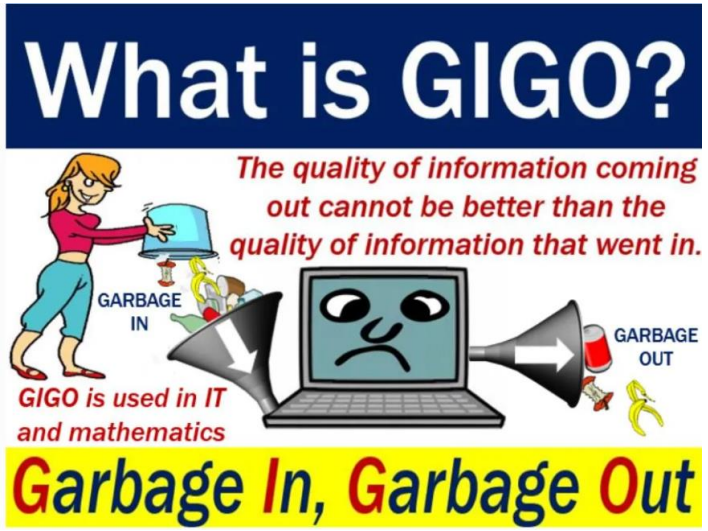


분석 프로세스

3. Data Preparation : 데이터 준비 : 계속

◆ 데이터 품질(Data Quality) : Garbage In Garbage Out

- 데이터가 어디서부터 수집되었고, 처리, 저장되었는지 판단
- 데이터 정제(data cleansing)
  - ✓ 중복 레코드 삭제
  - ✓ 이상 레코드 격리 및 처리 (예> 한 학기 수강신청학점이 45인 경우)
  - ✓ 속성값들의 표준화
  - ✓ 누락된 값 치환
- 데이터 웨어하우스: 데이터 정제 및 변환, 데이터 품질 관리, 기록 저장
- 잘 관리된 데이터 웨어하우스 → 높은 품질의 데이터 제공



분석 프로세스

3. Data Preparation : 데이터 준비 : 계속

◆ 결측치(Missing Values)

- 레코드의 속성값이 누락된 것
- 결측치 관리
  - ✓ 데이터 수집의 과정이나 계통을 추적
  - ✓ 누락된 값의 근원지를 파악
- 결측치가 무작위로 나타나고 빈도가 적을 경우
  - ✓ 결측치를 인위적으로 데이터 범위 내에서 치환하여 처리
- 알고리즘에 따라 결측치 문제를 해결하는 방법이 필요

◆ 데이터 유형 및 전환(data types and conversion)

- **데이터 유형**
  - ✓ 연속적인 수치형(continuous numeric)
  - ✓ 정수 수치형(integer numeric)
  - ✓ 범주형(categorical)
  - ✓ 서수형(ordinal)
- **유형 전환(type conversion)**
  - ✓ 알고리즘마다 입력값으로 허용하는 유형이 다름
  - ✓ 필요에 따라 전환(conversion)하여 입력값으로 사용
  - ✓ 인코딩(encoding): 범주형 → 수치형
  - ✓ 양자화(또는 binning): 수치형 →범주형

분석 프로세스

3. Data Preparation : 데이터 준비 : 계속

◆ 변환(transformation)

- 하나의 특정 속성이 결과값을 지배하지 않도록 하는 것이 중요
- 표준화 또는 정규화 방법을 통해 일관성 있는 범위의 값으로 변환
  - ✓ (예) 모든 속성값을 0~1사이로 변환
  - ✓ 주의: 이상치(outlier)가 있을 경우 정규화 결과가 왜곡될 수 있음

- 일부 데이터 마이닝에서는 속성의 개수를 줄일 필요
  - ✓ Feature Extraction

    : 속성을 합쳐서 새로운 하나의 속성으로 생성

    주성분 분석

  - ✓ Feature selection

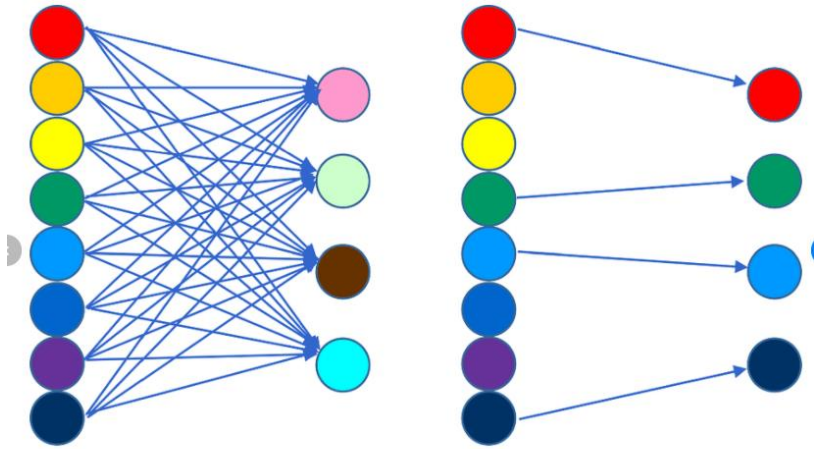
    : 의미 있는 속성만 선택

    :모든 속성들이 목적하는 결과값을 예측하는데 똑같이 중요하거나 유용하지는 않음

※ **특이값(outlier, 이상치)**

- 데이터 셋에서 정상적이지 않은 데이터 포인트
- 모델을 왜곡할 수 있기 때문에 처리가 필요

https://tink.com/blog/product/outlier-detection-categorisation/



Difference between feature extraction and feature selection

Ding, Ye & Zhou, Kui & Bi, Weihong. (2020). Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer. Soft Computing. 24. 10.1007/s00500-019-04628-6.

분석 프로세스

3. Data Preparation : 데이터 준비 : 계속

Marriage	Gender	Income	Child	Propose	SMS
Y	1	468.11	1	Y	1
Y	1	68.75	0	Y	1
Y	1	212.04	0		1
N	.	.	0	Y	1
y	2	585.05	0	Y	1
Y	1	-47.69	2		1
Y	1	487.7	0		1
N	.	.	1		2
Y	.	.	.	Y	1
n	1	0.00	0		1
Y	3	89982.12	0	Y	1
Y	2	585.05	0	Y	1

Marriage:결혼여부
Y:Yes, N:No

Gender : 성별
1: 남, 2: 여
단, 2000년 이전 출생자
국내거주자

Income : 월급

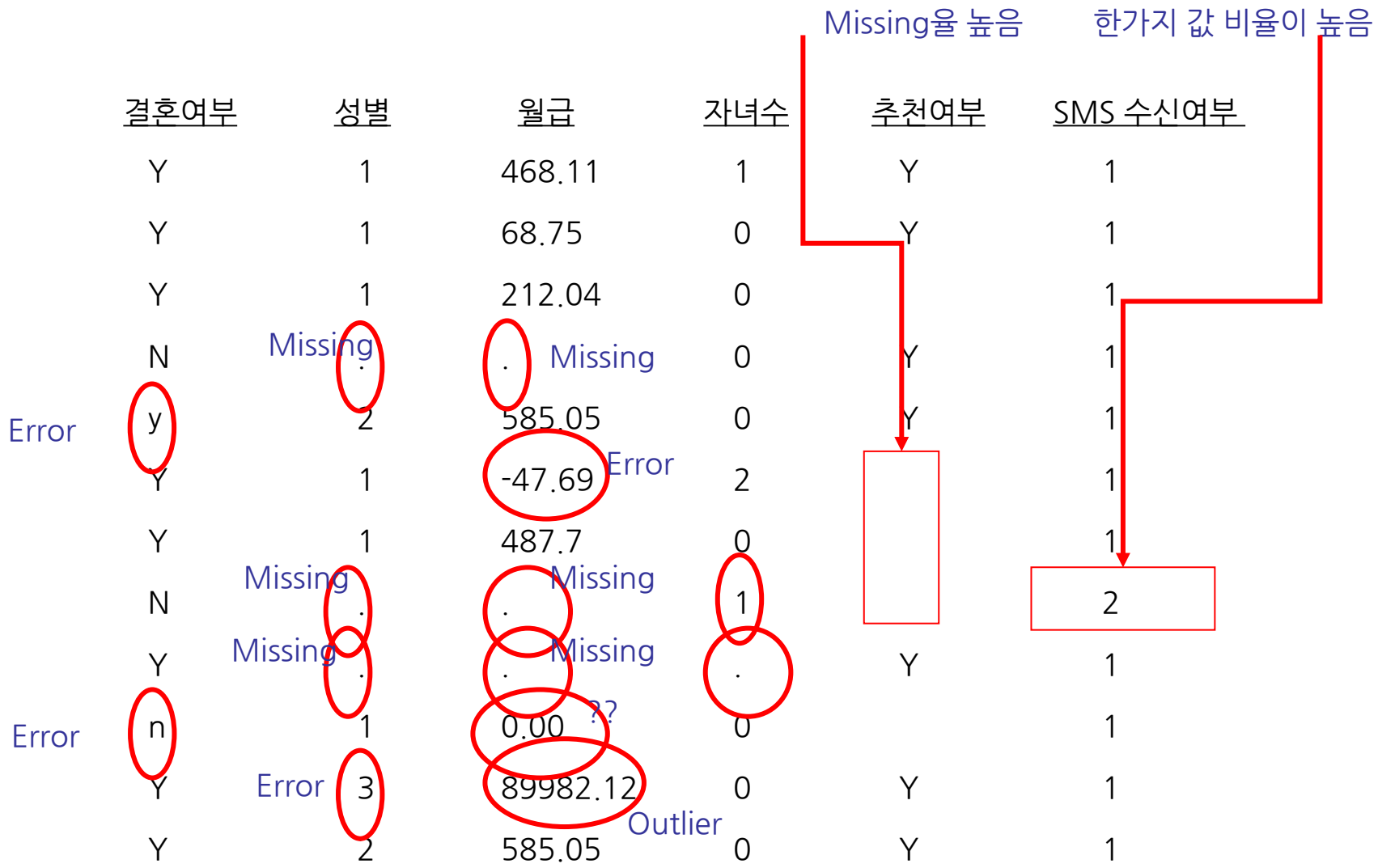
N of Child : 자녀수

Proposer : 추천여부
Y:Yes, N: No

SMS : SMS 수신여부
1: YES , 2: No

분석 프로세스

3. Data Preparation : 데이터 준비 : 계속



분석 프로세스

3. Data Preparation : 데이터 준비

◆ 새로운 변수 생성

1) 졸업까지 걸리는 기간 : 졸업일자 - 입학일자

2) PF 이수 학점 비중 : PF 이수 학점 수 / 전체 이수학점 수

3) 이수율: 이수학점 / 신청학점

4) 휴학 횟수

우리는 유용한 분석을 위해 가용데이터와 필요데이터 사이의 Gap을 최대한 줄여야 합니다.

이 GAP을 줄이기 위해, 보유데이터를 활용하여 파생변수를 생성하는 방법과 데이터를 획득하는 방법이 있습니다.

또한 마이닝을 통해 이 지식을 더할 수 있음.

분석 프로세스

3. Data Preparation : 데이터 준비

데이터를 이해하는 과정을 통해 수집한 데이터를 분석할 수 있은 상태로 만들어주는 과정

- 쓸 수 없는 데이터는 버리고, 수정이 필요한 부분은 수정해주고, 추가로 필요한 자료는 결합하는 작업
- 데이터 정제: 이상치 (예를 들어, 1학기 수강학점이 45학점인 학생이 있다면?)…

  변수 간 불일치 (예를 들어, 나이가 25세인데, 거주기간이 30년)…

  결측치 (예를 들어, 나이 정보가 없음)… 이런 경우 확 인 후 최대한 수정
- 데이터 변환: 필요한 경우 변수의 기록 방식을 변환

  (예를 들어, 소득의 경우 로그를 취하거나, 연령을 연령대로

분석 프로세스

4. Modeling ; 모델링

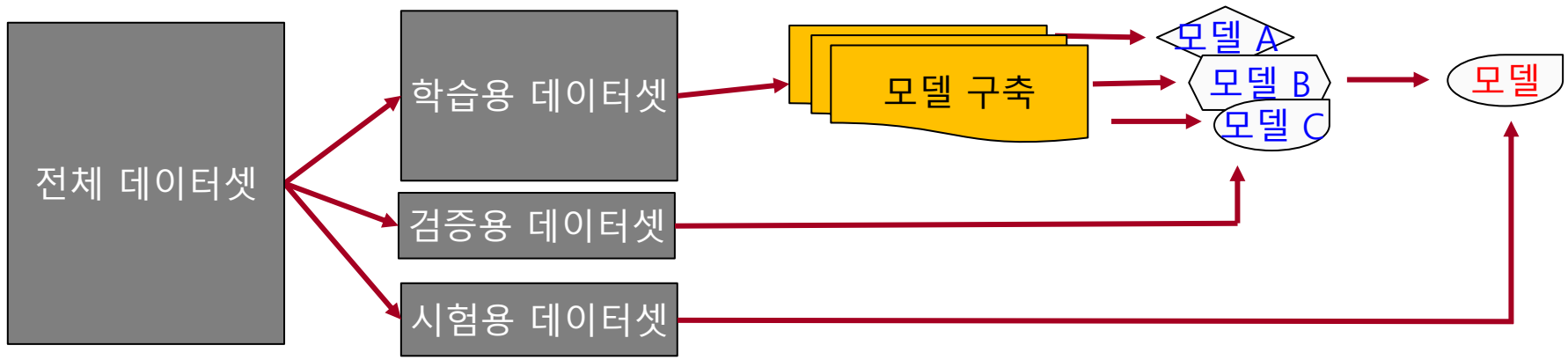
◆ 데이터셋 구분

- 일반적으로 데이터셋을 다음과 같이 세 가지로 구분함

  ✓ 학습용 데이터셋(training dataset): 모델을 구축하기 위해 사용되는 데이터

  ✓ 검증용 데이터셋(validation dataset): 모델이 얼마나 잘 구축되었는지 평가하고, 일부 모델들을 조정하며, 구축된 모델들 중에서 가장 좋은 것을 선택하기 위해 사용되는 데이터

  ✓ 시험용 데이터셋(test dataset): 최종 선택된 모델이 새로운 데이터에 대하여 얼마나 좋은 성과를 갖는지를 평가하기 위해 사용되는 데이터. 모델 구축 및 모델 선택 과정이 끝난 후에만 사용



분석 프로세스

4. Modeling ; 모델링

◆ 알고리즘 또는 모델링 기법(algorithm or modeling technique) - 계속

- 문제와 데이터 가용성에 따라 데이터 마이닝 분야를 결정
  - ✓ 연관성, 분류, 회귀 등
- 선택된 분야에서 적당한 알고리즘을 선택

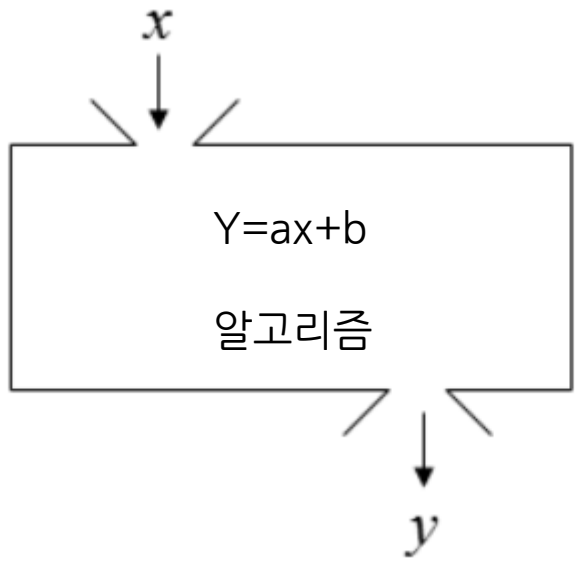
(예) 분류 분야 내에서 다음 알고리즘 중 하나를 선택

- ✓ 의사결정나무(decision tree)
- ✓ 규칙 유도(rule induction)
- ✓ 신경망(neural networks)
- ✓ 베이지안 모델(Bayesian model) 등

  (예) 의사결정나무를 선택했다면 다음 실행 방법 중 하나를 선택

  CART, CHAID, C4.5 등

- 하나의 문제를 해결하는데 다수의 분야와 알고리즘을 사용



분석 프로세스

4. Modeling ; 모델링

알고리즘 또는 모델링 기법(algorithm or modeling technique) - 계속

- 앙상블 모델링(ensemble modeling)
- 결과값을 잘 예측하기 위하여 다수의 모델링 알고리즘을 사용하거나 다수의 학습용 데이터셋을 사용
- 처음 보는 데이터에 대해, 기본 모델들의 예측값들을 종합하여 하나의 최종 예측값을 계산하는 방법

예측의 일반화 오류 감소가 목적
: 기본 모델들이 다양하고 독립적이기 때문에 앙상블 접근법을 사용할 때 모델의 예측 오류가 감소

분석 프로세스

5. Evaluation

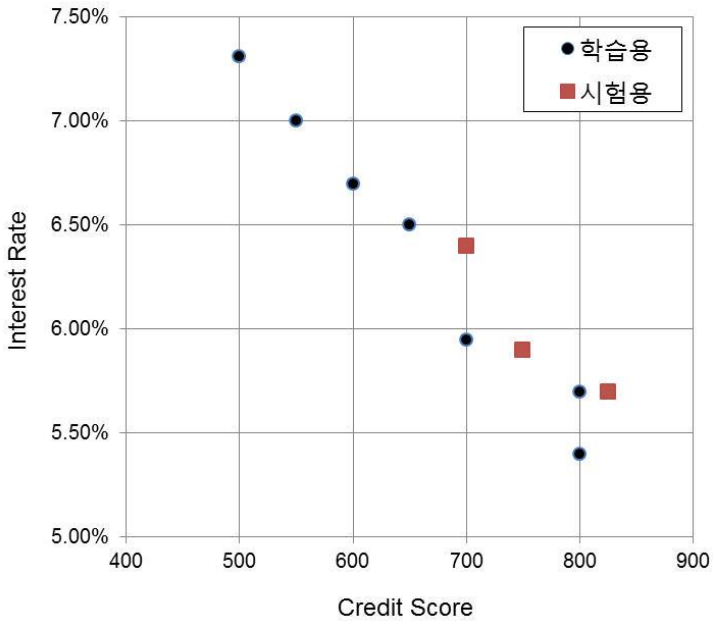
❖ 학습용 및 시험용 데이터셋

[표 2.3] 학습용 데이터셋

Borrower	Credit Score (X)	Interest Rate (Y)
01	500	7.31%
02	600	6.70%
03	700	5.95%
05	800	5.40%
06	800	5.70%
08	550	7.00%
09	650	6.50%

[표 2.4] 시험용 데이터셋

Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40%
07	750	5.90%
10	825	5.70%



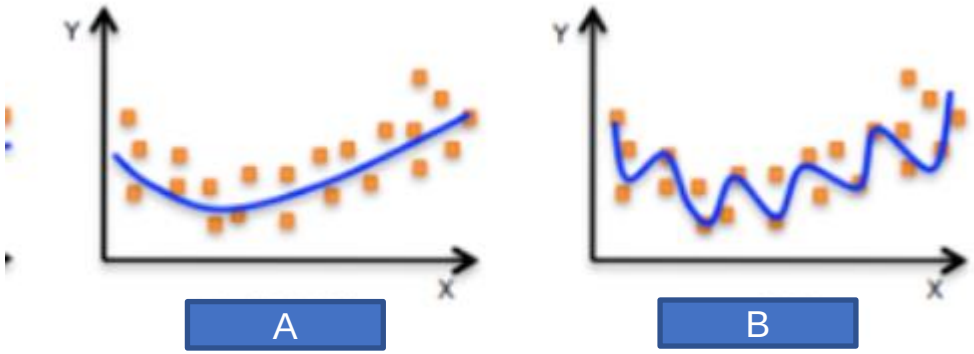
분석 프로세스

5. Evaluation : 평가

- 과적합(overfitting) 회피
  - ✓ 과적합: 학습용 데이터를 그대로 기억만 하는 현상
  - ✓ 과적합된 모델은 실제 현장의 데이터에 대해서는 성능이 좋지 않음

- 모델이 속성 간의 관계를 일반화하거나 학습하는가?
  - ✓ 시험용 데이터셋으로 모델을 검증하고 평가

세상에 단 하나밖에 없는 맞춤복



B 모형이 좋은 모형일까?

- 알고리즘이 학습 데이터에 너무 적합하면 새로운 사례에 대한 예측력이 떨어질 수 있음

설명력이 높으면서도 너무 복잡하지 않은 알고리즘이 가장 좋다
⇒ The Principle of Parsimon

분석 프로세스

6.Deployment : 배포

- 선정된 주제에 따라, 완성된 마이닝 프로세스를 자동화/시스템화 하는 과정
- 거래 시 발생할 수 있는 사기 탐지 같은 경우는 운영시스템에 배치 하는 경우가 증가함.
- 모델링 까지는 데이터과학팀, 시스템화 부터 운영 및 유지 등은 개발팀이 책임을 짐.