Statistical Machine Learning

1주차

담당: 15기 염윤석

0.Orientation

Machine Learning Session

분반장 : 15기 염윤석 & 16기 신인섭

Time : 매주 목요일 저녁 7시~9시 (7시 55분 ~8시 5분 휴식)

Materials : 분반장 자체 제작 PPT & 실습 과제 코드 ipynb

→ 다음 주차 자료는 매주 세션이 끝난 이후, KUBIG Github에 업로드될 예정!

(과제 제출 : KUBIG Github > KUBIG_2023_Fall > 1. 방학분반 > 머신러닝 >1. 강의자료)

→ 과제는 해당 세션 주, 수요일 자정까지 제출!

(과제 제출 : KUBIG Github > KUBIG_2023_Fall > 1. 방학분반 > 머신러닝 >2.과제제출 > n주차 폴더)

Projects

- 4주차 이후, 조별 프로젝트(추후, 공지)
- 8.31 KUBIG Contest (추후, 공지)

1. How does a Machine learn?

2. End to end ML Project

1. How does a machine Learn?

What is Learning Something

- Learning is used when:
  - – Humans are unable to explain their expertise (speech recognition)
  - – Solution changes in time (routing on a computer network)
  - – Solution needs to be adapted to particular cases (user biometrics)

What is Machine Learning

"The field of study that gives computers the ability to learn without being explicitly programmed. "

- Arthur Lee Samuel, 1959

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. "

- Tom Mitchell, 1997

"Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam "

Task (T): Classifying emails as spam or not spam.

Experience (E): Watching you label emails as spam or not spam.

Performance(P): The number (or fraction) of emails correctly classified as spam/not spam.

Why use Machine Learning

"Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam "

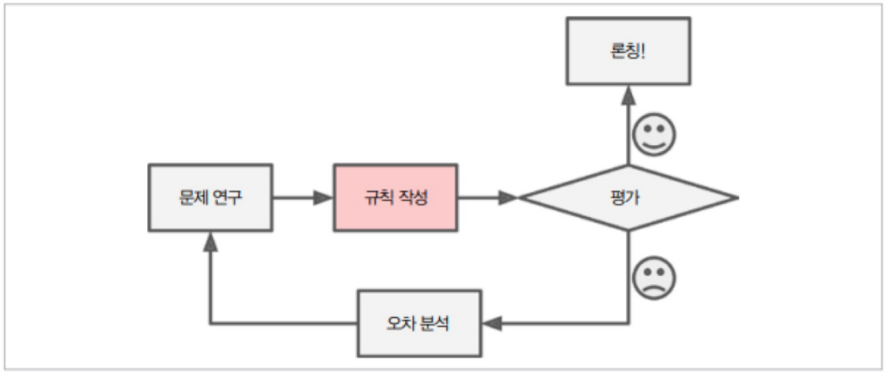


그림 1-1 전통적인 접근 방법

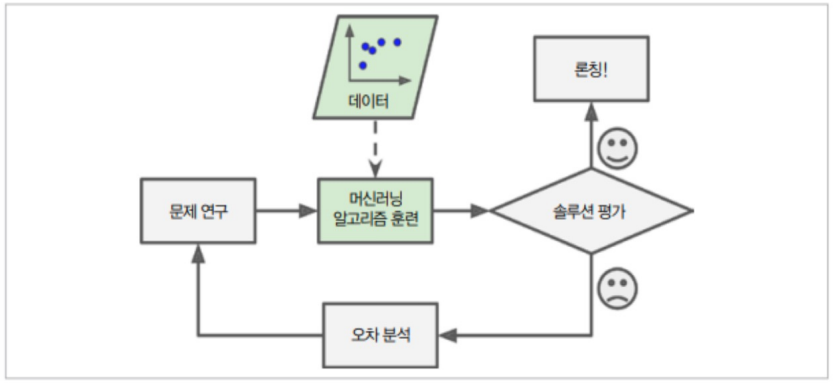


그림 1-2 머신러닝 접근 방법

Why use Machine Learning

"Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam "

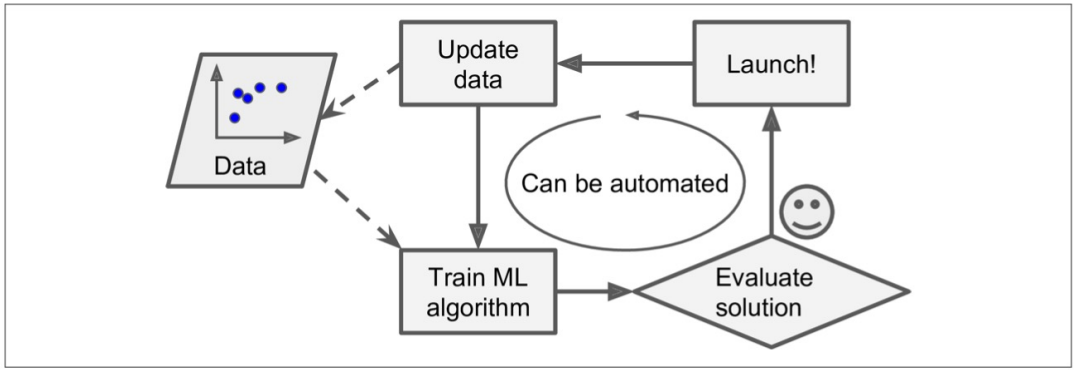


Figure 1-3. Automatically adapting to change

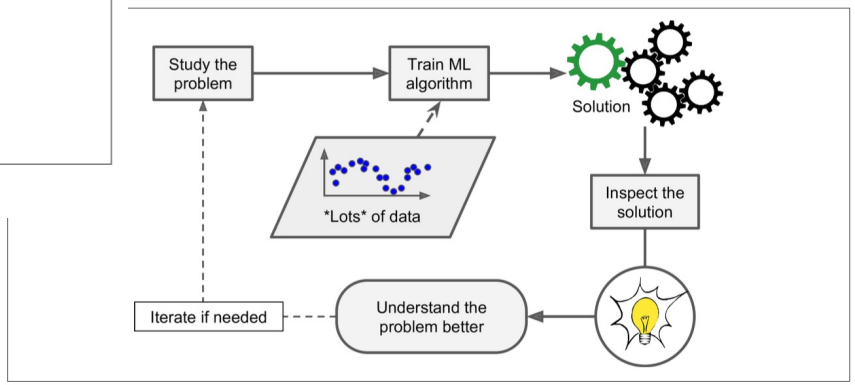
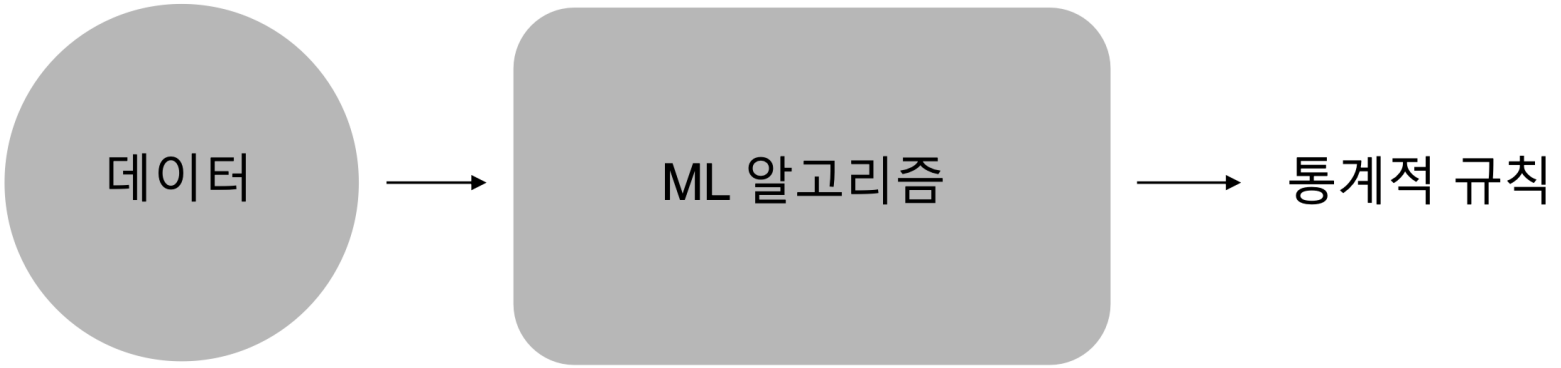


Figure 1-4. Machine Learning can help humans learn

What is Statistical Machine Learning



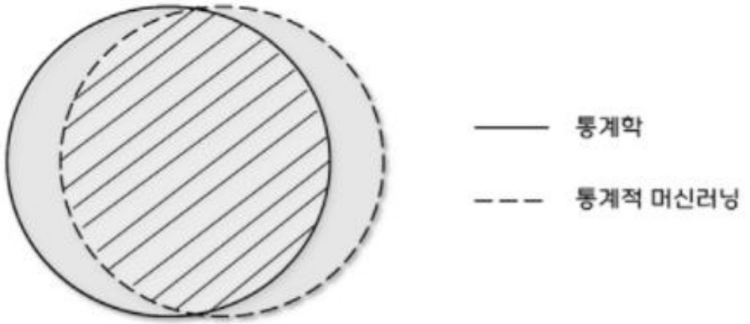
What is Statistical Machine Learning

전통적인 통계학

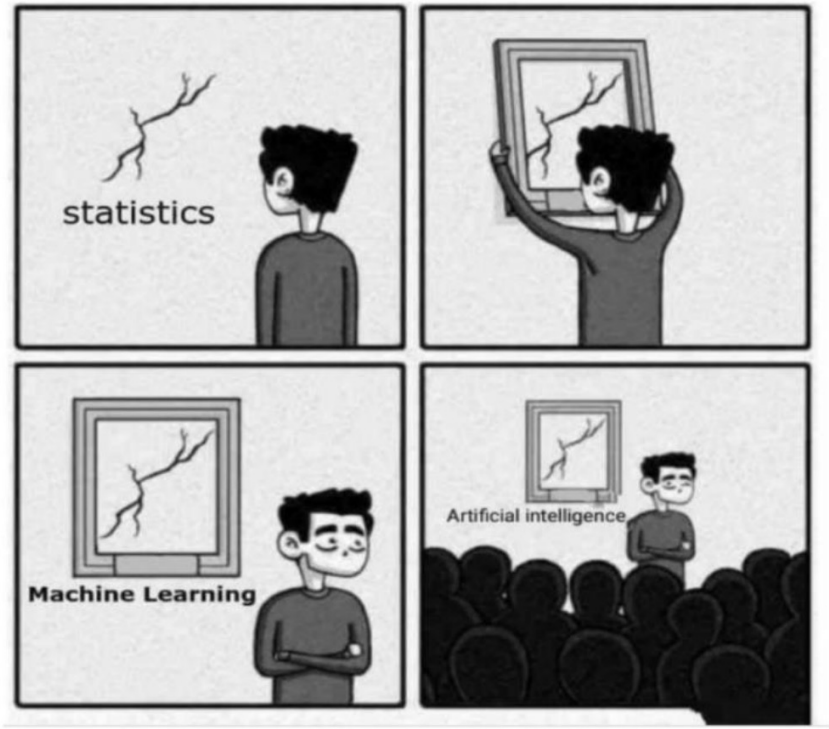
- 규칙의 통계적 추론에 중점
  (전문적인 통계적, 수학적 지식)
- 자료의 특성(다변량, 시계열,
  범주형 등)에 따라 분석.

통계적 머신러닝

- 규칙의 일반화에 중점
- 목적변수의 관측여부에 따라
  지도학습, 비지도학습으로 분석



What is Statistical Machine Learning



Types of Machine learning

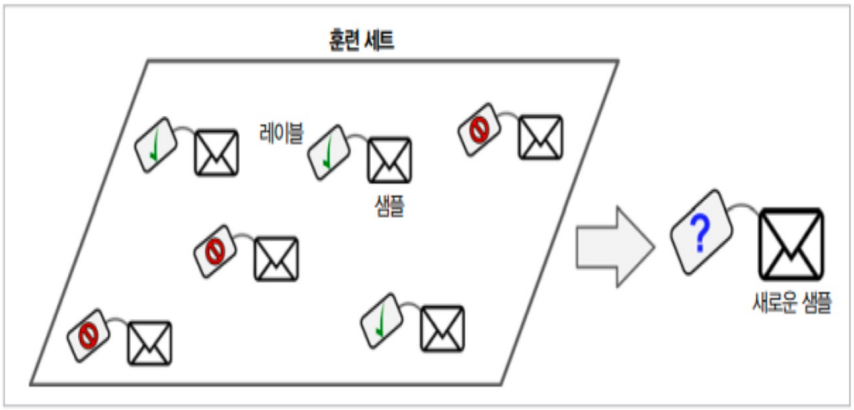
•Whether or not they are trained with human supervision
→ supervised, unsupervised, semi-supervised, and Reinforcement Learning

•Whether or not they can learn incrementally on the fly
→ online vs batch learning

•Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do
→ instance-based vs model-based learning

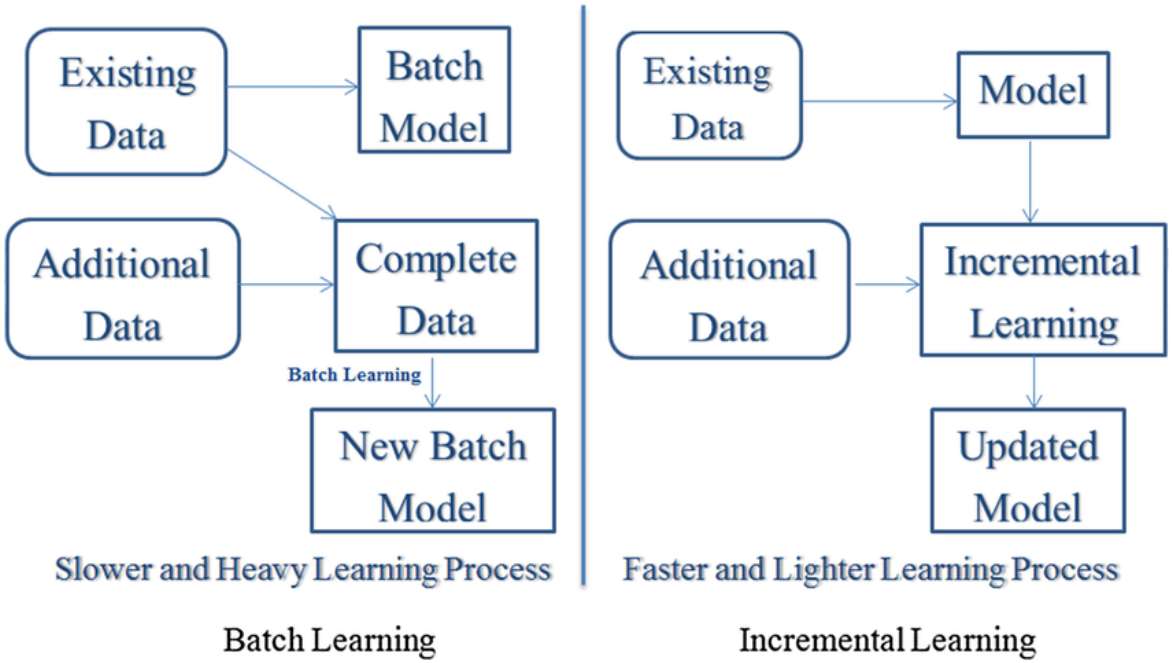
Types of Machine learning

Supervised Learning vs Unsupervised Learning



Types of Machine learning

On-line learning vs Batch Learning



Batch Learning Incremental Learning

Types of Machine learning

Instance-based vs Model-based

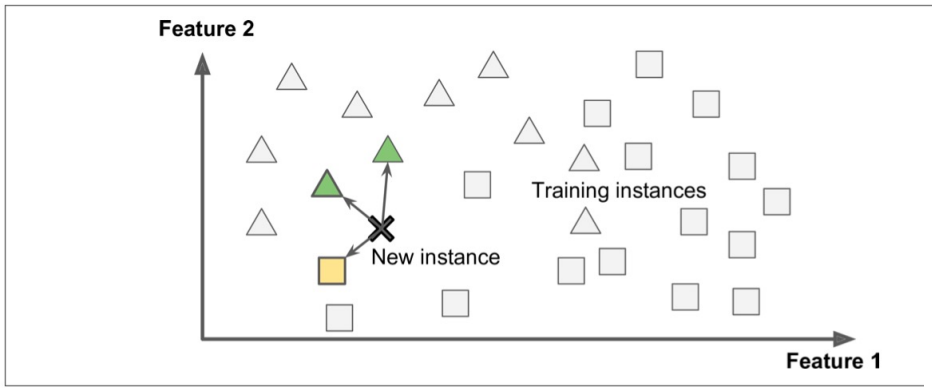


Figure 1-15. Instance-based learning

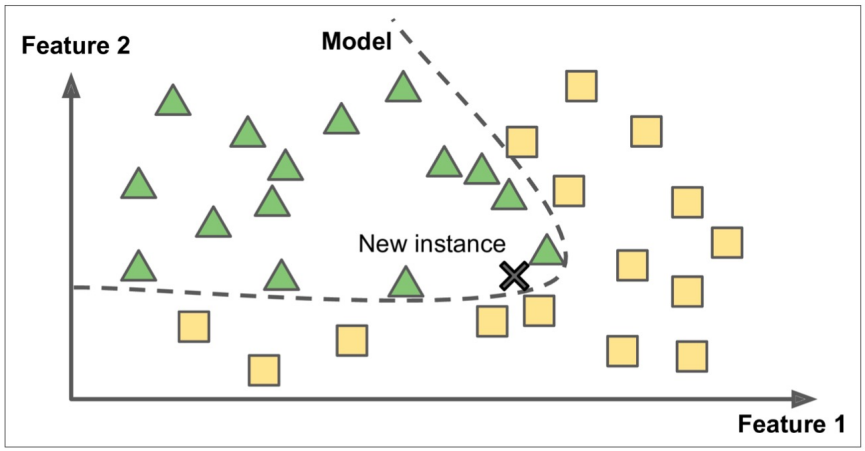
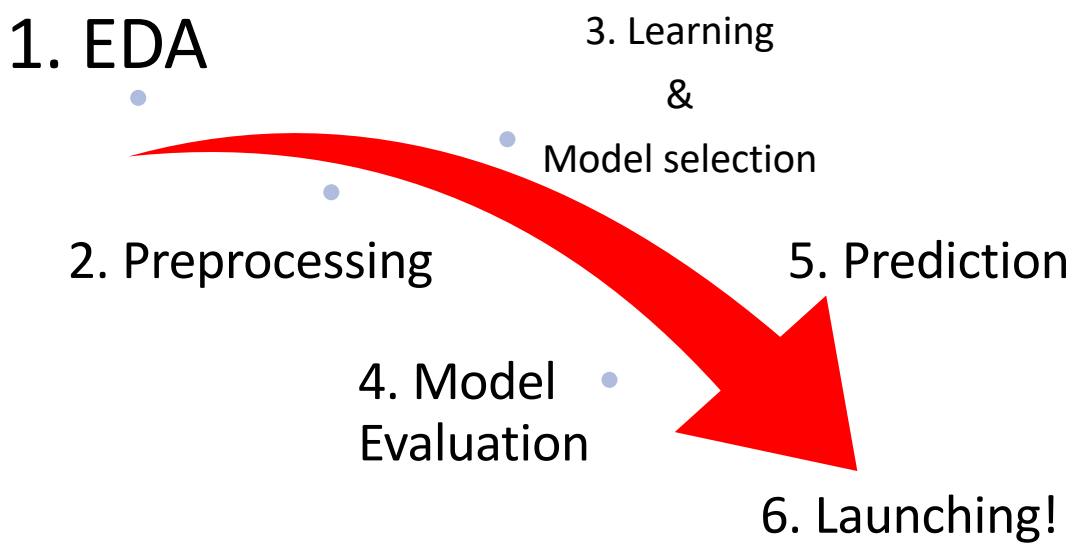


Figure 1-16. Model-based learning

2. End to end ML Project

End to End ML pipeline



EDA (Exploratory Data Analysis)

- EDA 란?

  수집한 데이터를 다양한 시각에서 이해하는 과정

  → 모델링 전, 그래프, 통계적 방법으로 데이터 자료를 살펴보는 과정

- 필요성

  모델에 필요한 정규성 검증, 혹은 데이터 분포 확인 등
  → Preprocessing(전처리) 과정과 같은 feature engineering의 방향성을 결정!

EDA (Exploratory Data Analysis)

- **EDA 란?**

  수집한 데이터를 다양한 시각에서 이해하는 과정

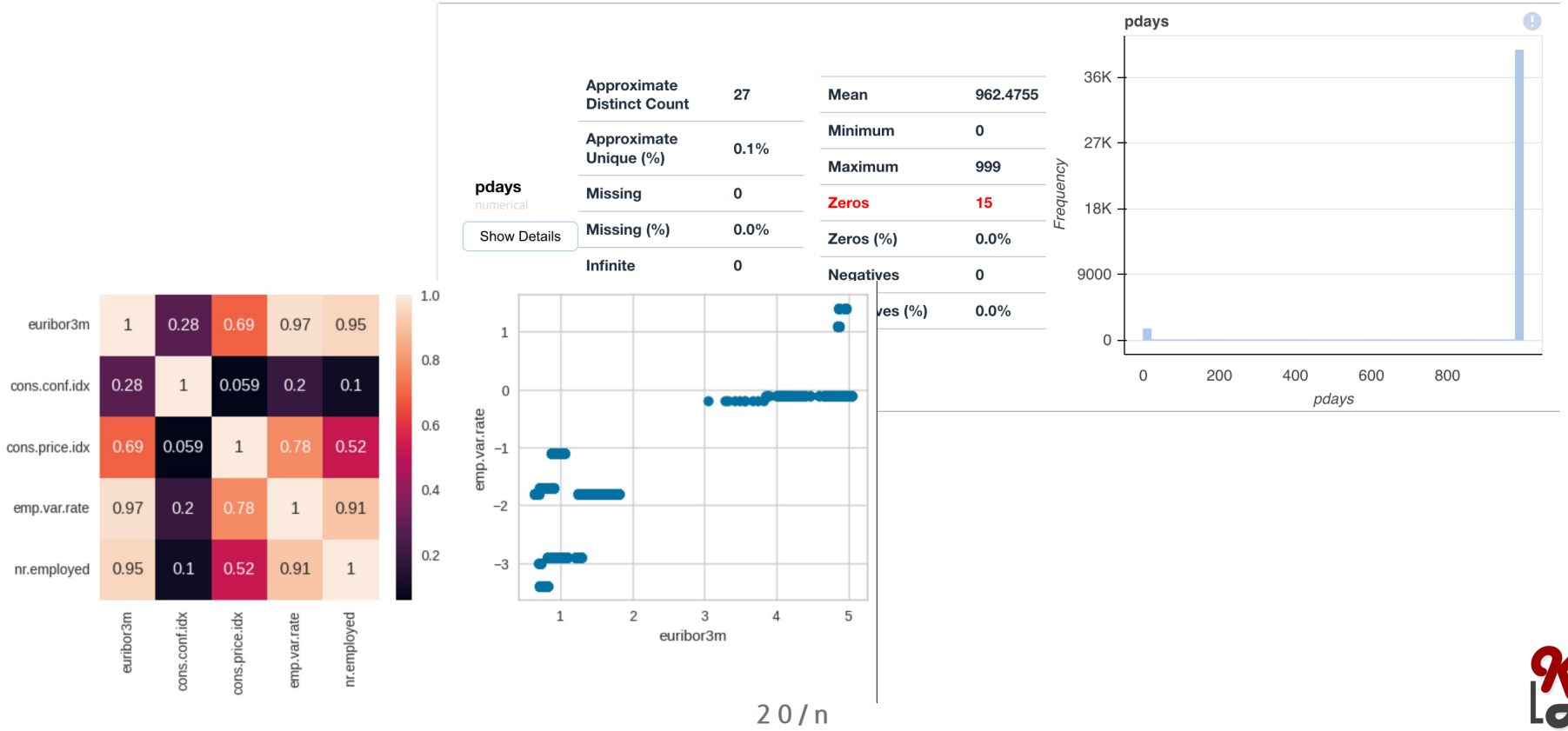
  → 모델링 전, 그래프, 통계적 방법으로 데이터 자료를 살펴보는 과정

- **필요성**

  모델에 필요한 정규성 검증, 혹은 데이터 분포 확인 등
  → Preprocessing(전처리) 과정과 같은 feature engineering의 방향성을 결정!

변수 분포 | 변수간 상관관계 | 이상치 | 결측치

EDA (Exploratory Data Analysis)



Preprocessing

보다 높은 정확성을 갖는 분석을 위해, 원자료에 대해 전환 및 가공을 거치는 단계 (AutoML)의 등장으로 그 중요도 및 비중이 높아지고 있음

정규화 & 표준화

특성변수의 단위에서 나타나는 차이를 조정해주는 역할

One-hot Encoding

범주형 변수를 수치형으로 변환

Bag of Words

텍스트를 수치형 변수로 변환(NLP)

PCA / t-SNE

차원축소

Feature 가 너무 많아 발생하는 Overfitting, curse of dimension을 해결

이상치 및 결측치

머신러닝 모델은 직접 결측치를 처리할 수 없음

SMOTE/ADASYN

불균형 자료 처리

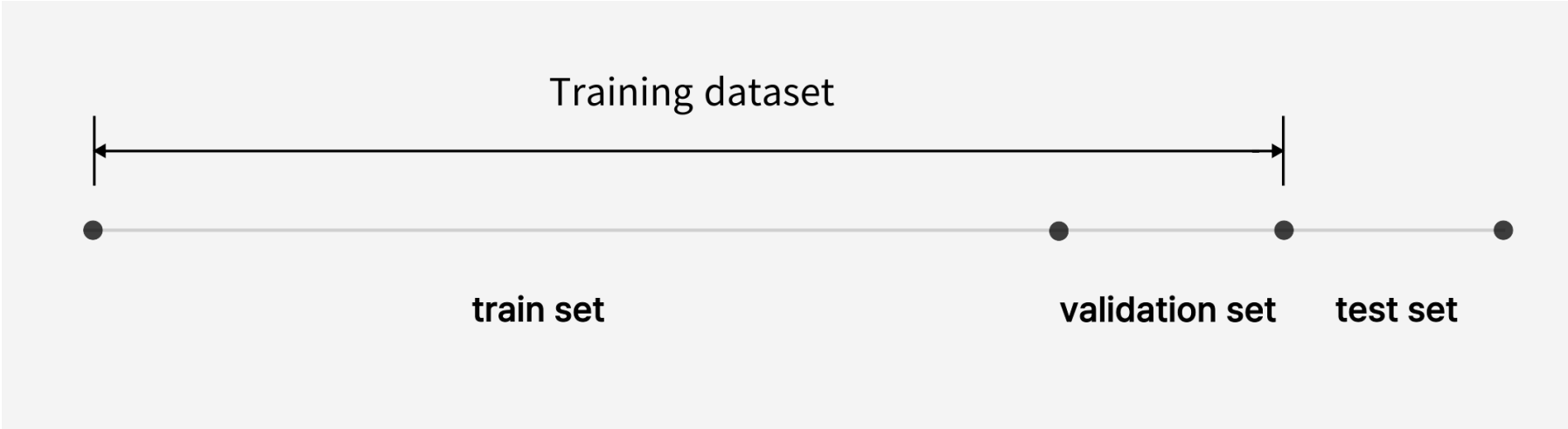
불균형 자료 처리를 해소하기 위한 과대표집 방법

Learning Process

Train data set → 학습 용도
Validation set → 모델의 일반화 및 모델 평가 비교 및 선택 → overfitting 을 방지

학습 데이터를 통해 학습된 모델을 cross validation 을 통한 Hyper-parameter Tuning (초모수 조정)

*초모수란, 모델이 학습하는 가중치가 아닌, 사용자 설정에 의해 달라지는 값들이다.

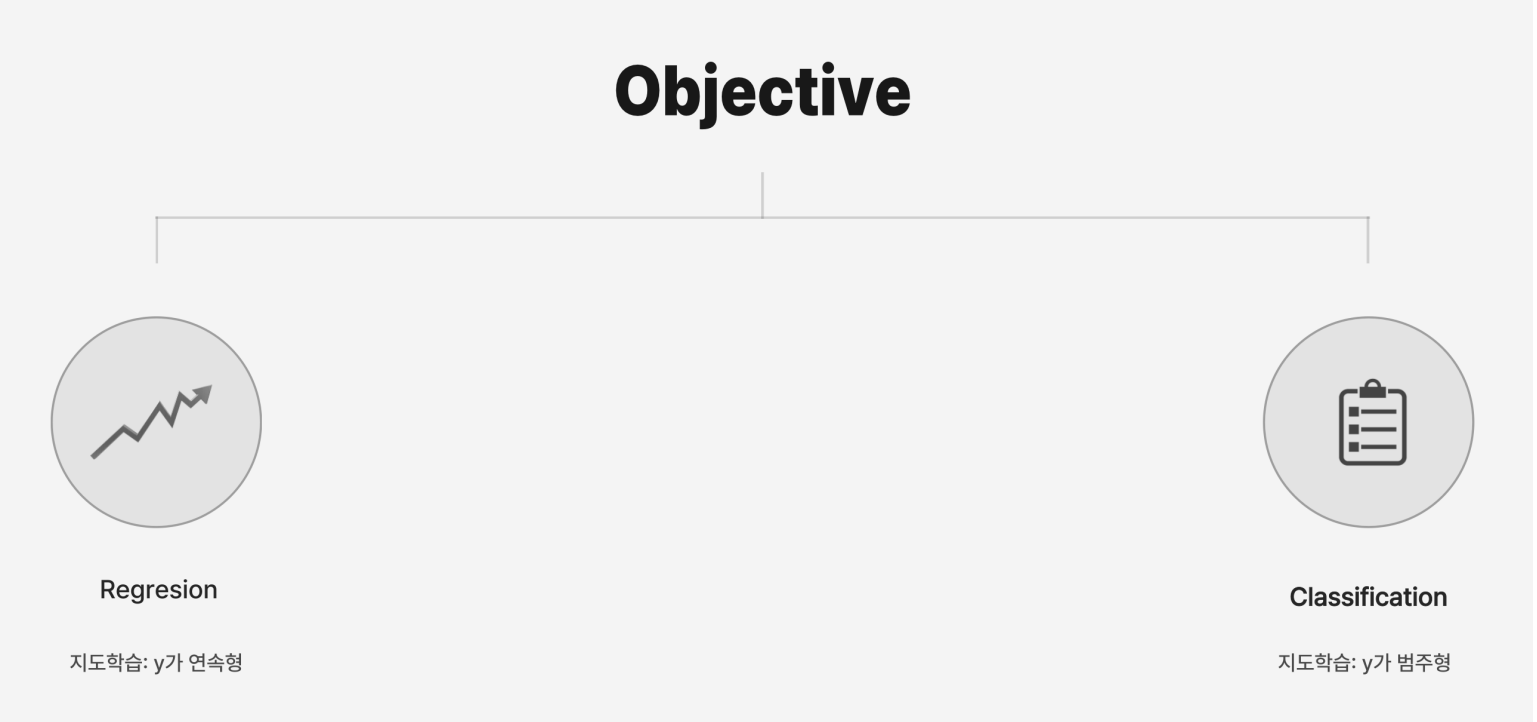


Model Evaluation

Tests set → 최종 모델 성능 검증

이후, 데이터 수 증대 및 규제화, 앙상블과 같은 기법을 통해 최종 모델의 성능을 향상시킬 수 있다.

Prediction & Launching



수고하셨습니다!

해당 세션자료는 KUBIG Github에서 보실 수 있습니다!