통계적 데이터 분석 절차

1) 기술적 데이터 분석(DDA)

- 데이터를 수집하고, 수집한 데이터의 구조와 타입을 확인하는 단계
- 비즈니스 목표에 부합하는 주요 데이터를 선택
  (목표변수 Y, Target, Label, Output / 비즈니스 목표에 있어 가장 핵심적인 변수 <-> (설명변수 X, Feature, Input / 목표변수를 제외한 나머지 변수)
- 기술 통계량 확인 (대푯값 / 산포 / 평균 / ...)
- 데이터 전처리

2) 탐색적 데이터 분석(EDA)

- 목표변수와 설명변수 간 관계(상관성, 연관성, 유사성)들을 파악하는 관계
- 데이터 시각화

3) 확증적 데이터 분석(CDA)

- 앞서 확인한 데이터 간의 관계를 가설로 수립하여 객관적인 값(p-value)을 계산해 검증하는 작업
- 통계적 가설 검정
- 비즈니스 Insight 도출

4) 예측적 데이터 분석(PDA)

- 기존의 데이터 간 관계를 새로 수식화하여 새로운 데이터가 들어올 때, 다음 상황을 예측 / 대응
- 기계학습(머신러닝)

1. 기술적 데이터 분석 (DDA)

- 데이터의 구조와 타입을 확인

- 정형데이터 :
  1) Index (데이터의 순서) -> 데이터의 개수 확인 (데이터 개수에 따라 분석 방법이 달라질 수 있음)
  2) Column (데이터 항목) -> 데이터의 항목의 데이터 타입을 확인
     - 연속형 (숫자) : int (정수) / float (실수)

- 범주형 (문자) : object (범주형)
- 순서형 (날짜) : datetime (날짜)

3) Value (데이터 값) -> 데이터의 형식을 확인 (Format)

- 데이터의 기술 통계량

1) 연속형 (숫자) :

- 대푯값 : 해당 숫자 데이터를 대표하는 값을 확인 (ex. 평균 Mean / 중앙값 Median)

  A : 2, 3, 1, 4, 5 -> 평균 : 3, 중앙값 : 3

  B : 2, 3, 1, 4, 1000 -> 평균 : 약 200 (이상치 Outlier), 중앙값 : 3
- 산포 : 중심(대푯값)으로부터 데이터가 얼마나 떨어져 있는가 (분산 / 표준편차)

  Ex) 자동차 생산 -> 신형 전기차 G 차량 (보증기간 5년)

  -> 타이어 휠 생산 업체 A / B

  -> A : 7년 / B : 10년
- 분포의 모양 : 데이터의 분포의 모양 (가설 검정)

2) 범주형 (문자) : 항목과 빈도수

2. 탐색적 데이터 분석 (EDA)

- 데이터 시각화

1. 상대방을 설득시키기 위한 자료를 만들 때

2. 데이터를 빠르게 이해하고 파악할 때

- 단일변수 시각화

1) 연속형 : 해당 숫자데이터의 분포를 확인하는 목적

- Histogram : 숫자데이터의 분포를 막대그래프로 표현 (X-연속형 / Y-빈도수)
- Kernel Density Estimator (KDE, 확률밀도) : 모수(모집단의 통계량)의 추정치 분포를 확률로 표현 (X-연속형 / Y-확률)
- Box Plot (상자그림) : 숫자 데이터의 분포의 위치를 Box로 표현 (IQR : 전체 데이터의 50%가 분포해 있는 범위

2) 범주형 : 해당 데이터의 빈도수를 확인하는 목적

- 막대 그래프 / 파이차트

- 다변수 시각화

1) X 범주형 / Y 연속형 : 집단 간 통계량을 비교하는 경우

-> 막대 그래프 (Bar Chart)

2) X 연속형 / Y 연속형 : 두 숫자 데이터의 상관성을 확인하는 경우 (비례 / 반비례)

-> 산점도 (Scatter Plot)

3) X 순서형 / Y 연속형 : 시간/날짜 데이터에 따라 연속형 자료의 추세를 확인하는 경우

-> 선그래프 (Line Chart)

4) 기타 : 관리도 (Control Chart) / 파레토도 (Pareto Plot) / ...

3. 확증적 데이터 분석 (CDA)

- 통계적 가설 검정 : 규명하고자 하는 가설을 수립해, 객관적인 수치로 가설의 참/거짓을 규명

- 가설 :

- 귀무가설 (보통가설) : 기각할 목적으로 수립하는 가설 (대립가설의 반대)
  통계학에서 기본적인 가정 : 항목 간 독립된 상태로 데이터가 수집
  -> 집단 간 평균이 같다 / 정규분포 / 상관성이 없다 / 연관성이 없다 ...
- 대립가설 : 채택을 목적으로 수립하는 가설
  (규명하고자 하는 바 / 의심하는 바 / 이상상태)
  -> 정규분포 X / 집단 간 평균이 다르다 / 상관성, 연관성이 있다

- P-Value (확률값) : 귀무가설이 참일 확률 (0% ~ 100%)

- 유의수준 (5%) : 수립된 가설을 채택/기각하기 위한 기준 값

  - P-value > 5% (0.05) : 귀무가설 참

  - P-value < 5% (0.05) : 대립가설 참

- 단일변수 :

  1) 연속형 (숫자) : 해당 숫자데이터의 분포가 정규분포를 따르는가? (정규성 검정 Normal Test)

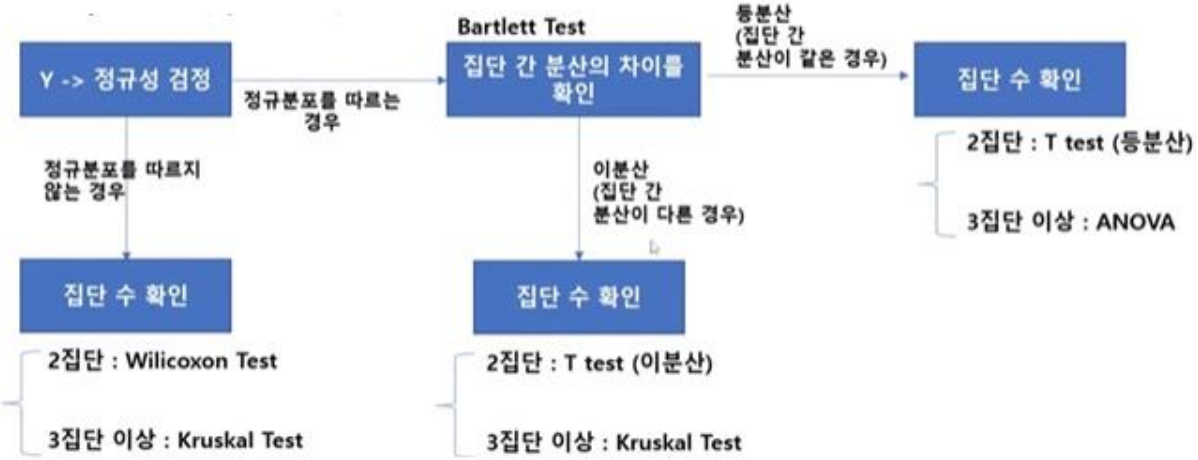
-> 다른 통계적인 분석을 수행할 때, 숫자데이터의 분포의 모양에 따라 분석기법이 달라짐

-> 귀무가설 : 해당 데이터의 분포는 정규분포를 따른다.

-> 대립가설 : 해당 데이터의 분포는 정규분포를 따르지 않는다.

- 다변수 :

1) X 범주형 / Y 연속형 : 집단 간 통계량(평균/분산)을 비교



2) X 연속형 / Y 연속형 : 두 연속형 숫자데이터의 상관성을 확인

-> 두 연속형 데이터가 정규분포 : Pearson R Test

-> 정규분포 X : Spearman R Test

3) X 범주형 / Y 범주형 : 두 범주형 데이터가 서로 연관/독립인지 확인

-> Chi^2 Test (카이제곱검정)

4. 예측적 데이터 분석 (PDA)

- 기계학습 (Machine Learning) : 컴퓨터가 데이터 간 (Y - 목표변수 / X - 설명변수) 규칙/수식을 학습을 통해 도출하는 작업

- 학습 핵심 Point (기계학습의 실무 핵심 Point) :

1) 학습 능력 : 컴퓨터가 데이터로부터 적절한 수식을 도출하는 능력
2) 일반화 능력 : 학습을 통해 도출된 수식(Model)이 새로운 데이터에 대해 잘 예측/대응하는 능력

- 기계학습의 핵심 3요소

1. 데이터(교과서) : 학습 목적에 맞는 깔끔한 데이터셋 구축
   -> 특성공학 (Feature Engineering)
2. 알고리즘(선생님) : 학습에 있어 데이터로부터 적절한 수식/규칙을 도출할 수 있는 적절한 알고리즘을 선택하는 것이 중요
   -> 선형회귀 / 의사결정나무 / KNN / SVM / Ensemble ...
3. 하드웨어(학생) : CPU / GPU
   -> 비용 (Cost)

- 기계학습의 종류 :

1. 지도 학습 (Supervised Learning) : 목표변수(Y)와 설명변수(X)의 관계를 수식화하여, 새로운 X가 들어올 때, Y를 예측하거나 분류하는 기법
   - 회귀 분석 (Regression, Y – 연속형) : 정확한 목표변수를 예측
   - 분류 분석 (Classification, Y – 범주형) : 특정 항목을 정확하게 구분
   Ex) 주가예측프로그램 -> 주가(Y) - 회귀 / 스팸 메시지 분류기 -> 스팸여부(Y) - 분류

   - 지도 학습 절차

   1. 데이터 핸들링 (데이터 불러오기, 파생변수 생성, 이상치 제거, ...)
   2. 회귀/분류 기법을 적용시킬 목표변수(Y)와 설명변수(X)를 설정
   3. 학습 데이터(Train Set)와 검증 데이터(Test Set)를 분할
   4. 학습 데이터를 이용해 수식을 생성 (Modeling)
      - 특성 공학 (Feature Engineering)
      - 학습 알고리즘
   5. 평가 (Evaluation)
      - 학습 능력 -> 학습 성능 평가 (Train Set)

      Y_train (실제값) – Y_train_pred (예측값)

      - 일반화 능력 -> 일반화 성능 평가 (Test Set)

      Y_test (실제값) – Y_test_pred (예측값)

2. 비지도 학습 (Unsupervised Learning) : 설명변수(X) 간 유사성, 연관성 등을 계산

하여 유사한 데이터끼리 묶거나, 비슷한 데이터를 군집화하는 기법

- 군집분석 (Clustering) : 서로 유사한(특성이 비슷한) 데이터끼리 묶어주는 기법
- 연관분석 (Association Analysis) : 데이터 간 유사도를 계산하여 서로 연관성 높은 데이터를 찾는 기법

Ex) 장바구니 분석 / 추천 시스템 / ...

3. 강화 학습 (Reinforcement Learning) : 컴퓨터가 시뮬레이션을 통해 주어진 상황 (사용자 정의한 상황)에 대해 보상이 좋은 방향으로 학습 (데이터가 없어도 학습이 가능!)
   - 게임 AI / 시뮬레이션

- 모델 평가 기법 (Evaluation)

1) 분류에서의 평가

정확도 Accuracy = $\frac{\textit{정확하게 분류한 데이터 수}}{\textit{전체 데이터 수}}$

오차 행렬 (Confusion Matrix)

	Real Negative	Real Positive
Predict Negative	True Negative (TN)	False Negative (FN)
Predict Positive	False Positive (FP)	True Positive (TP)

정밀도 Precision = $\frac{TP}{FP+TP\ (Predict\ Positive)}$

재현율 Recall = $\frac{TP}{FN+TP\ (Real\ Positive)}$

F1 Score = $\frac{2\times(\textit{정밀도}\times\textit{재현율})}{\textit{정밀도}+\textit{재현율}}$

- 회귀 모델 평가 지표 (Model Evaluation)

- R Square (결정계수 R^2) : 회귀선이 데이터를 얼마나 잘 대변하는가 (0 ~ 1)

$$R^2 = 1 - \frac{SSE}{SST}$$

총 변동 : $\sum\left(\textit{실제값} - \textit{평균}\right)^2$, SST (Total Sum of Squares)

회귀 변동 : $\sum\left(\textit{예측값} - \textit{평균}\right)^2$, SSR (Regression Sum of Squares)

오차 변동 : $\sum\left(\textit{실제값} - \textit{예측값}\right)^2$, SSE (Error Sum of Squares)

- Mean Squared Error (MSE) : $\frac{\Sigma(\text{실제값} - \text{예측값})^2}{\text{데이터 수}}$
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE) : $\frac{\Sigma|\text{실제값} - \text{예측값}|}{\text{데이터 수}}$

- Overfitting (과적합) : 학습 데이터에 대해서는 성능이 높게 나오지만, 검증 데이터에 대해서는 성능이 매우 저조하게 나오는 현상

해결책

- 특성공학 기법을 이용해 데이터를 깔끔하게 다듬는다.

- 학습 알고리즘 통제

- 특성공학 기법 : 컴퓨터가 학습을 수행할 때, 일반화된 수식이 적절하게 도출될 수 있도록 데이터를 다듬는 기법

1. Imputation (결측값 처리)
   - 결측값을 다른 값으로 대치하여 데이터의 공백 없이 학습을 수행
   - 제거 : 결측값이 있는 모든 행을 다 제거 (df1.dropna( ))
   - 대치 :
     - 연속형 -> 평균 (중앙값) / 범주형 -> 최빈값
     - 알고리즘을 이용한 대치 (KNN)
     - 보간법 (df1.interpolate( ))
2. Scaling & Encoding

   1) Scaling : 연속형 데이터의 Scale 맞춰주는 작업

   Ex) 계약기간 12 ~ 60 / 비용 10,000,000 ~ 100,000,000 / 연령 20 ~ 80

   -> Standard Scaler : 모든 숫자 데이터를 평균 0 / 표준편차 1

   -> Min-Max Scaler : 모든 숫자 데이터를 최솟값이 0 / 최댓값 1

   -> Robust Scaler : 모든 숫자 데이터를 중앙값 0 / IQR 1 (Inter Quantile Range : 사분범위, 중앙값을 기준으로 전체데이터의 50%가 포함된 구간)

   2) Encoding : 범주형 데이터를 연속형 숫자 데이터로 변환

   -> Label Encoding : 범주형 데이터의 각 항목을 정수로 변환

   -> One Hot Encoding
3. Cross Validation : 데이터셋이 모두 학습에 참여할 수 있도록 학습 데이터셋을 교차로 바꿔가며 학습을 진행하는 방법

4. Hyper Parameter Tuning : 데이터를 학습할 때 발생하는 알고리즘의 수학적 구조를 사용자가 통제
    - Hyper Parameter : 알고리즘 내 수학 구조 / 학습을 수행하면서 발생하는 구조
    - Hyper Parameter Tuning : 데이터를 학습할 때 알고리즘에 의해 발생하는 구조들을 사용자가 통제
    - Random Search : Hyper Parameter를 랜덤으로 부여하여 가장 적절한 Hyper Parameter를 찾는 방법
    - Grid Search : Hyper Parameter를 사용자가 사전에 지정한 방법대로 (경우의 수에 따라) 부여하여 가장 적절한 Hyper Parameter를 찾는 방법

    Ex) Decision Tree

    - Depth / Leaf / Split / …
    - Random Search :
      Depth – 10 / Leaf – 7 / Split – 15,
      Depth – 20 / Leaf – 3 / Split – 5, … 랜덤으로 부여된 것들 중에서 적절한 것을 찾는 방법
    - Grid Search :
      Depth – 5~10 / Leaf – 5~10 / Split – 1~20 해당 범위 내에서 가장 적절한 것을 찾는 방법 -> 모든 경우의 수에 대한 모델을 생성
5. Imbalanced Data Sampling (분류) : 서로 다른 비율의 데이터를 맞춰주는 작업
    - Under Sampling : 데이터의 비율이 적은 쪽으로 한 쪽의 데이터를 줄여주는 기법
      (Random Under Sampling / Tomek's Link / KNN / …)
    - Over Sampling : 데이터의 비율이 큰 쪽으로 한 쪽의 데이터를 늘려주는 기법
      (Random Over Sampling / SMOTE / ADASYN / …)

    Ex) 불량여부(Y) <-> X1, X2, X3, …

    -> 정상 950 / 불량 50

    Imbalanced Sampling : 정상 50 / 불량 50 or 정상 950 / 불량 950

- 산포의 개념

    - 편차 (Deviation) : 개별 값 – 평균
    - 편차 합 (Sum of Deviation) : $\sum\left(개별\ 값 - 평균\right)$
    - 편차 제곱 (Square Deviation) : $\left(개별\ 값 - 평균\right)^2$
    - 편차 제곱 합 (Sum of Square Deviation) : $\sum\left(개별\ 값 - 평균\right)^2$
    - 분산 (Variance) : $\sum\left(개별\ 값 - 평균\right)^2 \ /\ 데이터\ 수$
    - 표준편차 (Standard Deviation) : $\sqrt{\sum\left(개별\ 값 - 평균\right)^2 \ /\ 데이터\ 수}$
    - 범위 (Range) : 최댓값(Max) – 최솟값(Min)
    - 사분 범위 (Inter Quantile Range, IQR) : 데이터를 오름차순 정렬한 뒤, 중앙값 (50%)을 기준으로 앞, 뒤 25% 구간 계산 (전체데이터 25% ~ 75%에 해당하는 구간)

Table Join : Key Column을 기준으로 서로 다른 Table 병합

(Key Column : Table을 병합할 때, 기준이 되는 항목)

- Inner Join : Key Column을 기준으로 공통된 값을 병합
- Outer Join : Key Column을 기준으로 두 Table 모든 값을 병합
- Right / Left Join : Key Column 값에서 왼쪽 혹은 오른쪽의 값을 기준으로 병합

최적화 (Optimization)

- 문제 상황에서 여러 해결방안 중 가장 최적의 해결방안을 찾아가는 작업

수학적 접근

- 특정 함수의 값을 최소화(또는 최대화)시키는 최적의 수식값(최적의 파라미터)의 조합을 찾는 문제
- 최적화 문제는 최대화(Maximization)와 최소화(Minimization)으로 나누어 볼 수 있다.

    - 최대화(Maximization) : 함수의 목표변수(Y, Label, Output)를 최대가 되게끔 파라미터(계수와 절편, Weight)의 조합을 찾는 문제

-> 이윤 / 점수 / ...

- 최소화(Minimization) : 함수의 목표변수(Y, Label, Output)를 최소가 되게끔 파라미터(계수와 절편, Weight)의 조합을 찾는 문제

-> 비용 / 손실 / 오차 / ...

- 기본적으로 머신러닝(기계학습) 데이터의 관계를 파악해 새로운 데이터가 들어올 때, 목표값을 예측/분류하는 수식을 만드는 과정(Modeling)에서 최적화를 이용해 Model(수식) 구축
- 함수식이 예측한 예측값과 실제 데이터 상에서 수집된 실제값을 비교하여, 실제값과 예측값의 차이가 최소가 되는 방향으로 최적화 수행

데이터 분석에서 최적화

- 머신러닝 Model을 구축하여 예측/분류를 수행할 때, 정확한 Model을 도출하기 위해 여러 파라미터를 비교하여 찾는 작업
- 여러 알고리즘을 비교하여 가장 적절한 알고리즘을 찾는 작업
- 회귀분석에서 최소제곱법(Least Squared Method)과 경사하강법(Gradient Descent)과 같은 방법으로 Model 구축

최적화의 기본 원리 (Minimization 최소화 관점에서)

- 현재의 위치에서 함수의 값(Y)이 감소하는 방향으로 파라미터 값을 변경해 나가는 원리
- 파라미터 값을 변경해도 더 이상 감소하는 부분이 없을 때까지(Local Minima) 파라미터를 변경해 나감

최적화의 핵심 Point

- 어떤 방향(양, 음, 고차원의 경우 특정 벡터 방향)으로 감소/증가할 것인가?
- 얼만큼 파라미터를 변경해 나갈 것인가?

-> 위의 방향과 변경할 값의 크기에 따라 최적화하는 기법들이 달라짐

최적화의 다양한 기법

- Least Square Method (최소제곱법)
- Gradient Descent (경사하강법)
- Newton's Method
- Gauss Newton Method

- Levenberg-Marquardt Method
- Bayesian Method

최소제곱법 (Ordinary Least Square OLS / Least Square Method)

- 데이터를 대표하는 회귀선을 찾을 때, 회귀선이 예측한 예측값과 실제값의 차이(잔차, Residual)의 제곱의 합 또는 평균이 최소화(Minimization)되는 방향으로 파라미터를 결정하는 방법
- 실제값 : 실제 X와 매칭되는 Y값
- 예측값 : 생성된 회귀선이 계산한 Y값
- 잔차 (Residual) : 실제값 – 예측값
- RSS (Residual Sum of Squares)

경사하강법 (Gradient Descent)

- 점진적인 반복 계산을 통해 함수가 최소가 되는 파라미터를 찾아주는 기법
- 점진적 반복 계산을 수학에서 그래디언트(Gradient) 개념을 이용하여 적용
- 그래디언트 (Gradient) : 함수가 증감하는 방향과 크기를 표현
- 초기값(m, b값에 대한 초기값)을 설정한 상태에서, Gradient가 감소하는 방향으로 점진적 계산을 통해 최적의 m값과 b값을 찾는 기법
- Learning Rate (학습율) : 얼마나 점진적으로 파라미터를 변화해 나갈지 곱해주는 값
- 적절한 Learning Rate를 찾아, RSS가 최소가 되는 지점을 찾아야 함
  - Learning Rate 값이 너무 크면, RSS가 최소가 되는 지점을 놓칠 수 있음
  - Learning Rate 값이 너무 작으면, 파라미터를 찾는데 시간이 오래 걸림