II. 빅데이터 탐색

01. 데이터 전처리

1.1 데이터 정제

-	KeyWord
데이터 정제	데이터 전처리, 데이터 정제, 데이터 세분화
결측값 처리	결측값, 단순 대치법, 다중 대치법
이상값 처리	이상값, 이상값 검출, 통계 기법, ESD, 기하평균, 사분위 수, 표준화 점수, 딕슨의 Q검정, 그럽스 T-검정, 카이제곱 검정, 시각화, 머신러닝 기법, 마할라노비스 거리, LOF, iForest, 이상값 처리, 삭제, 대체법, 변환, 박스 플롯 해석, 분류하여 처리

==========

1) 데이터 정제

(1) 데이터 전처리의 중요성

- 데이터 전처리는 반드시 거쳐야 하는 과정
- 분석 결과에 직접 영향을 주므로, 반복적인 전처리 수행 필요
- 데이터 분석의 단계 중 가장 많은 시간 소요(전체 중 80% 정도)
- 데이터 전처리 순서: 데이터 정제 -> 결측값 처리 -> 이상값 처리 -> 분석 변수 처리

(2) 데이터 정제(Data Cleansing) 개념

- 오류 데이터값을 정확한 데이터로 수정하거나 삭제하는 과정
- 결측값을 채우거나 이상값을 제거하여 데이터 신뢰도를 높이는 작업

(3) 데이터 정제 절차

- 데이터 오류 원인 분석 -> 정제 대상 선정 -> 정제 방법 결정
- 오류 원인 분석: 결측값 / 노이즈 / 이상값
    - 결측값(Missing Value): 누락된 값(입력X)
        - 처리 방법: 평균값, 중앙값, 최빈값 등의 중심 경향값 넣기 / 분포 기반 처리
    - 노이즈(Noise): 잘못 판단된 값(입력되지 않았는데, 입력되었다고 판단됨)
        - 처리 방법: 일정 간격으로 이동하면서 평균값 대체 / 일정 범위 중간값 대체
    - 이상값(Outlier): 범위에서 많이 벗어난 값(지나치게 작은 값 or 큰 값)
        - 처리 방법: 하한보다 낮으면 하한값 / 상한보다 높으면 상한값 대체
- 정제 대상 선정

    - 모든 데이터를 대상으로 정제 활동
    - 품질 저하 위협이 있는 데이터는 더 많은 정제 필요
    - 품질 저하 위협: 내부 < 외부 / 정형 < 비정형&반정형
- 정제 방법 결정: 삭제 / 대체 / 예측값 삽입
    - 정제 여부 결정
        - 정제 규칙을 이용하여 검색
        - 노이즈, 이상값은 비정형 데이터에서 특히 자주 발생함
    - 삭제: 오류 데이터 부분 or 전체 삭제
    - 대체: 평균값, 최빈값, 중앙값 대체
    - 예측값 삽입: 회귀식 등을 이용하여 예측값 생성

(4) 데이터 정제 기술

- 데이터 일관성 유지를 위한 정제 기법: 변환/ 파싱/ 보강
- 다른 시스템에서 들어온 데이터에 일관성 부여
    - 변환(Transform): 다양한 형태 -> 일관된 형태로 변환
        - (ex) YYYYMMDD -> YY/MM/DD
    - 파싱(Parsing): 유의미한 최소 단위로 분할 (정제 규칙을 적용하기 위함)
        - (ex) 주민등록번호 -> 생년월일, 성별
    - 보강(Enhancement): 변환 / 파싱 / 수정 / 표준화 등을 통한 추가 정보를 반영
        - (ex) 주민등록번호 -> 성별 추출 후 반영
- 데이터 정제 기술: ETL / 맵리듀스 / 스파크 / 스톰 / CEP / 피그 / 플럼
    - 분산 처리 시스템을 기반으로 정제
    - 성능 보장을 위해 인메모리 기반 기술을 사용하기도 함
    - 정제된 후, 데이터 변경(분석)에 활용됨
    - ETL: 데이터를 추출 -> 가공 -> 데이터 웨어하우스 / 데이터 마트에 저장
    - 맵리듀스
        - 대용량 데이터셋을 분산, 병렬 컴퓨팅 처리
        - 모든 데이터를 키-값 쌍으로 구성
        - (맵: 데이터 추출) + (리듀스: 중복 없게 처리)
        - 배치 형태: 많은 데이터 처리 시 성능 느림
    - 스파크/스톰
        - 인메모리 기반 데이터 처리 방식
        - 스파크: 맵리듀스 기반으로 성능 개선 / 실시간, 배치 처리 둘 다 가능
    - CEP(Complex Event Processing)
        - 실시간 이벤트 처리에 대한 결괏값 수집, 처리
        - 실시간 데이터: IoT 센싱 데이터/ 로그/ 음성 데이터 등
    - 피그(Pig)
        - 대용량 데이터 집합을 분석하기 위한 플랫폼
        - 피그 라틴이라는 자체 언어 제공
    - 플럼(Flume)
        - 로그 데이터를 수집, 처리하는 기법
        - 실시간에 근접하게 처리함

(5) 데이터 세분화(Data Segmentation)

- 데이터 세분화 개념
    - 데이터를 기준에 따라 나누고 선택한 매개변수를 기반으로 유사한 데이터를 그룹화함
    - 데이터 세분화 방법: 응집분석법 / 분할분석법 / 인공신경망 모델 / K-평균 군집화
        - 군집화
            - 이질적인 집단을 몇개의 동질적인 소집단으로 세분화
            - 군집화 방법: 계층적 / 비 계층적 방법으로 구분
        - 계층적 방법: 응집분석법 / 분할분석법
            - 응집분석법: 각 객체를 하나의 소집단으로 간주 -> 유사한 소집단을 합침
            - 분할분석법: 전체 집단에서 시작 -> 유사성 떨어지는 객체를 분리
        - 비 계층적 방법: 인공신경망 모델 / K-평균 군집화
            - 인공신경망 모델: 생물학의 신경망으로부터 영감을 얻은 통계학적 학습모델
            - K-평균 군집화: K개 소집단의 중심좌표와 각 객체 간의 거리를 계산 -> 중심좌표 업데이트

==========

2) 데이터 결측값 처리

(1) 데이터 결측값 개념

- 데이터 결측값(Missing Value): 입력이 누락된 값(NA / 999999 / Null)

(2) 데이터 결측값 종류

- 완전 무작위 결측 / 무작위 결측 / 비 무작위 결측
- 완전 무작위 결측 (MCAR)
    - 결측값이 다른 변수들과 아무 상관 없음
    - (ex) Y가 누락될 확률은 X 또는 Y와 관련이 없음
- 무작위 결측 (MAR)
    - 특정 변수와 관련되어 일어남 / 변수의 결과는 관계 없음
    - (ex) Y가 누락될 확률은 X의 값에만 의존
- 비 무작위 결측 (MNAR)
    - 누락된 값(변수의 결과)이 다른 변수와 관계 있음
    - (ex) Y가 누락될 확률은 Y 자체의 관찰되지 않는 값에 달려 있음

(3) 데이터 결측값 처리 절차

- 결측값 식별 -> 부호화 -> 대체
- 결측값 식별(Identify): 데이터 형태와 현황 파악
- 결측값 부호화(Encode)
    - 컴퓨터 처리 가능한 형태로 부호화 -> NA / NaN / inf / NULL
    - NA: 기록되지 않은 값(Not Available)
    - NaN: 수학적으로 정의되지 않은 값(Not a Number)
    - inf: 무한대(infinite)
    - NULL: 값이 없음
- 결측값 대체(Impute): 대체 알고리즘을 통해 결측값 처리

(4) 데이터 결측값 처리 방법

- 단순 대치법(Single Imputation)
  - 결측값을 그럴 듯한 값으로 대체하는 통계적 기법
  - 통계량의 효율성, 일치성 등을 부분적으로 보완해줌
  - 대체된 자료는 결측값 없이 완전한 형태
- 단순 대치법의 종류: 완전 분석법/ 평균 대치법/ 단순 확률 대치법
  - 완전 분석법
    - 불완전 자료는 완전 무시/ 완전하게 관측된 자료만 사용
      -> 효율성 상실 / 통계적 추론의 타당성 문제 발생
  - 평균 대치법
    - 얻어진 자료의 평균값으로 결측값을 대치
    - 비 조건부 평균 대치법: 평균값으로 대치
    - 조건부 평균 대치법: 회귀분석을 활용하여 결측값 대치
  - 단순 확률 대치법
    - 확률값을 부여한 후 대치
    - 핫덱 대체(Hot-Deck): 현재 진행 중인 연구에서 비슷한 성향을 가진 응답자의 자료로 무응답을 대체
    - 콜드덱 대체(Cold-Deck): 외부 출처 / 이전의 비슷한 연구에서 가져온 자료로 대체
    - 혼합 방법: 몇 가지 다른 방법을 혼합
  - 다중 대치법(Multiple Imputation)
    - 다중 대치법 개념
      - 단순 대치법을 1번 이상, m번 대치
      - m개의 대치된 표본을 구하는 방법
      - 3단계: 대치 → 분석 → 결합
    - 여러 번의 대체 표본으로 대체 내 분산, 대체 간 분산을 구하여 추정치의 총 분산을 추정하는 방법

==========

3) 데이터 이상값 처리

(1) 데이터 이상값 개념

- 데이터 이상값(Data Outlier)
  - 관측된 데이터의 범위에서 많이 벗어난
  - 아주 작은 값 or 아주 큰 값
  - 입력 오류, 데이터 처리 오류 등의 이유로 특정 범위에서 벗어난 데이터값

(2) 데이터 이상값 발생 원인

- 입력 오류 / 측정 오류 / 실험 오류 / 고의적인 이상값 / 표본추출 에러
- 입력 오류: 전체 데이터 분포에서 쉽게 발견 가능
- 실험 오류: 실험 조건이 동일하지 않은 경우
- 고의적인 이상값: 자기 보고식 측정(Self Reported Measures)에서 나타나는 에러
- 표본추출 에러: 샘플링을 잘못한 경우

(3) 데이터 이상값 검출 방법

- 개별 데이터 관찰 / 통계 기법 / 시각화 / 머신러닝 기법 / 마할라노비스 거리 / LOF / iForest
- 개별 데이터 관찰: 전체 데이터 추이/ 특이사항/ 무작위 추출하여 관찰
- 통계 기법
    - ESD / 기하평균 / 사분위 수 / 표준화 점수 / 딕슨의 Q검정 / 그럽스 T-검정 / 카이제곱 검정
        - ESD(Extreme Studentized Deviation): 평균에서 3표준편차 떨어진 값은 이상값 -> $\mu - 3\sigma < data < \mu + 3\sigma$
        - 기하평균 활용: 기하평균에서 2.5표준편차 떨어진 값은 이상값 -> 기하평균 - 2.5σ < data < 기하평균 + 2.5σ
        - 사분위 수 활용: $Q_1 - 1.5(Q_3 - Q_1) < data < Q_3 + 1.5(Q_3 - Q_1)$
        - 표준화 점수(Z-score) 활용: 정규분포를 따르는 관측치들이 평균에서 얼마나 떨어져 있는지 나타내어 이상값 검출
        - 딕슨의 Q 검정: 오름차순 정렬 -> 범위에 대한 관측치 간의 차이의 비율을 활용 (데이터 30개 미만인 경우 적절함)
        - 그럽스 T-검정: 정규분포를 만족하는 단변량 자료에서 이상값 검정
        - 카이제곱 검정: 정규분포를 만족하나, 자료 수가 적은 경우
- 시각화: 확률밀도함수/ 히스토그램/ 시계열 차트
- 머신러닝 기법: K-평균 군집화
    - 주어진 데이터를 K개의 클러스터로 묶는 알고리즘
    - 각 클러스터와 거리 차이의 분산을 최소화하는 방식
- 마할라노비스 거리(Mahalanobis Distance)
    - 데이터의 분포를 고려한 거리 측도
    - 관측치가 평균으로부터 벗어난 정도를 측정
    - 모든 변수 간 선형관계 만족 & 각 변수들이 정규분포를 따르는 경우에 적용할 수 있음
- LOF(Local Outlier Factor)
    - 관측치 주변 밀도와 근접한 관측치 주변 밀도의 상대적인 비교를 통해 이상값 탐색
    - 각 관측치에서 k번째 근접이웃까지의 거리 산출 -> 해당 거리 안에 포함되는 관측치 개수로 나눈 역수 값으로 산출
- iForest(Isolation Forest)
    - 관측치 사이의 거리나 밀도가 아닌, 의사결정나무를 이용하여 이상값 탐색
    - 분류모형을 생성하여 모든 관측치를 고립시켜나가면서 분할 횟수로 이상값 탐색
    - 적은 횟수로 잎 노드(Leaf Node)에 도달하는 관측치일수록 이상값일 가능성이 큼

(4) 데이터 이상값 처리

- 삭제 / 대체법 / 변환 / 박스플롯 해석 / 분류하여 처리
- 삭제(Deleting Observations)
    - 이상값은 제외하고 분석
    - 추정치의 분산이 작아짐
    - 과소/과대 추정되어 편의 발생할 수 있음
    - 양극단의 값을 절단(Trimming)하기도 함: 기하평균 / 상하단 %를 이용한 제거
    - 절단보다, 극단값 조정 방법을 활용 -> 데이터 손실률↓ 설명력↑
- 대체법(Imputation)
    - 하한값보다 작으면 하한값 대체
    - 상한값보다 크면 상한값 대체
- 변환(Transformation)
    - 자연로그를 취해 데이터 값 감소 -> 실젯값 변형
    - 오른쪽으로 길게 기울어진 분포 -> 평균 중심 대칭 형태로 변환

- 박스플롯 해석(Box-Plot)
  - 수염 밖에 있는 값을 이상값으로 판단
  - 수염(Whiskers): Q_1, Q_3로부터 IQR의 1.5배 내에 있는 가장 멀리 떨어진 데이터까지 이어진 선
  - $IQR = Q_3 - Q_1$
- 분류하여 처리
  - 이상값이 많을 경우 사용하는 방법
  - 서로 다른 그룹으로 묶음 -> 각 그룹에 대해 통계적인 모형 생성 -> 결과 결합

1.2 분석 변수 처리

-	**KeyWord**
변수 선택	변수, 종속변수, 독립변수, 변수선택, 필터기법, 정보 소득, 카이제곱 검정, 피셔 스코어, 상관계수, 래퍼기법, 전진선택법, 후진제거법, 단계적방법, RFE, SFS, 유전 알고리즘, 단변량 선택, mRMR, 임베디드기법, 라쏘, 릿지, 엘라스틱넷, SelectFromModel
차원축소	차원축소, 주성분분석(PCA), 특이값분해(SVD), 요인분석, 독립성분분석(ICA), 다차원척도법(MDS)
파생변수 생성	파생변수
변수 변환	변수변환, 단순 기능 변환, 비닝, 정규화, 표준화
불균형 데이터 처리	불균형 데이터 처리, 임계값이동, 앙상블기법, 언더샘플링, ENN, 토멕 링크 방법, CNN, OSS, 오버샘플링, SMOTE, Borderline-SMOTE, ADASYN

==========

1) 변수 선택

(1) 변수 개념

- 변수(Feature): 데이터 모델에서 예측에 사용되는 입력변수
- RDBMS에서 속성/열 = 머신러닝에서 변수
- 변수 유형: 알려진 값 & 예측값
  - 알려진 값: 변수 / 속성 / 예측변수 / 차원 / 관측치 / 독립변수
  - 예측 값: 라벨 / 클래스 / 목푯값 / 반응 / 종속변수

(2) 변수 유형

- 인과관계에 따라 - 독립변수, 종속변수
  - 독립변수
    - 종속변수에 영향을 주는 변수
    - 종속변수가 특정한 값을 가지게 되는 원인이 된다고 가정함
    - 연구자가 의도적으로 변화시키는 변수
    - 독립변수 = 예측변수/회귀자/통제변수/조작변수/노출변수/리스크 팩터/설명변수/입력변수

    - 종속변수
        - 독립변수로부터 영향을 받는 변수
        - 독립변수의 영향을 받아 그 값이 변할 것이라고 가정함
        - 어떻게 변화하는지 연구하는 변수
- 속성에 따라 - 범주형(명목형, 순서형), 수치형(이산형, 연속형)
    - 명목형: 이름만 의미 부여 / 크기와 순서는 상관 없음 / 명사형
    - 순서형: 순서에 의미 부여 가능
    - 이산형: 하나하나 셀 수 있음
    - 연속형: 구간 안의 모든 값을 가질 수 있음
- 변수 간 관계
    - 독립변수, 종속변수 둘 다 연속형, 범주형 자료로 분석 가능
    - 연속형 자료에서 원인은 공변량(Covariate) 이라고 부름
    - 범주형 자료에서 원인은 요인(Factor) 이라고 부름

(3) 변수 선택(Feature Selection)

- 독립변수(x)들 중 종속변수(y)에 가장 관련성이 높은 변수만 선정하는 방법
- 변수 선택 특징
    - 해석하기 쉽도록 모델 단순화
    - 훈련 시간 축소
    - 차원의 저주 방지(차원이 증가할수록, 필요한 샘플 데이터가 기하급수적으로 증가하는 현상)
    - 과적합 줄이고 일반화
    - 모델 정확도, 성능 향상 기대
- 변수 선택 방식 분류
    - 비지도 방식: 분류를 참고하지 않고 변수들만으로 선택 수행
    - 지도 방식: 분류를 참고하여 변수 선택

<변수 선택 기법: 필터 / 래퍼 / 임베디드 기법 >

필터 기법	래퍼 기법	임베디드 기법
정보 소득 카이제곱 검정 피셔 스코어 상관계수	RFE SFS 유전 알고리즘 단변량 선택 mRMR	라쏘(LASSO) 릿지(Lidge) 엘라스틱넷(Elastic Net) SelectFromModel

- 필터 기법(Filter Method)
    - 데이터의 통계적 특성으로부터 변수를 선택
    - 절차: 변수 전체집합 -> 베스트 하위집합 선택 -> 알고리즘 학습 -> 성능 평가
    - 특징
        - 통계적 측정 방법으로 변수들의 상관관계를 알아냄
        - 계산 속도 빠름 -> 래퍼 기법 사용 전, 전처리에 사용함
    - 사례: 정보 소득/ 카이제곱 검정/ 피셔 스코어/ 상관계수
        - 정보 소득(Information Gain): 가장 정보 소득이 높은 속성 선택
        - 카이제곱 검정(Chi-Square Test): 관찰 빈도와 기대 빈도의 차이가 유의한가 검정
        - 피셔 스코어(Fisher Score): 최대 가능성 방정식을 풀기 위한 뉴턴의 방법

        - 상관계수(Correlation Coefficient): 두 변수간 상관관계 정도를 나타낸 계수

- 래퍼 기법(Wrapper Method)

    - 변수의 일부만으로 모델링 반복
    - 절차: 변수 전체집합 -> (하위 집합 -> 알고리즘 학습)을 반복 -> 성능 평가
    - 특징
        - 예측 정확도 성능이 가장 좋은 하위 집합을 선택하는 기법
        - 그리디 알고리즘(Greedy Algorithm): 하위 집합을 반복 선택
            - 그리디 알고리즘: 문제를 해결하는 과정에서 그 순간마다 최적이라고 생각되는 결정을 하는 방식으로 진행하여 최종 해답에 도달하는 문제해결방식
        - 일반적으로 필터 기법보다 예측 정확도 높음
        - 시간 오래 걸림 / 과적합 위험 있음
    - 알고리즘 유형: 전진선택법 / 후진제거법 / 단계적방법(전진+후진)
        - 전진선택법: 빈 모델 -> 변수 하나씩 추가(모델을 가장 많이 향상 시키는 변수)
        - 후진제거법: 풀 모델 -> 변수 하나씩 제거(모델에 가장 적은 영향을 주는 변수)
    - 기법 상세: RFE / SFS / 유전 알고리즘 / 단변량 선택 / mRMR
        - RFE(Recursive Feature Elimination): SVM 사용 -> 재귀적으로 제거
        - SFS(Sequential Feature Selection): 그리디 알고리즘 -> 빈 모델에 하나씩 추가
        - 유전 알고리즘(Genetic Algorithm): 자연세계 진화과정에 기초한 전역 최적화 기법(존 홀랜드, 1975)
        - 단변량 선택(Univariate Selection): 각 변수를 개별 검사 -> 변수와 반응변수간 관계 강도 결정
        - mRMR(Minimum Redundancy Maximum Relevance): 특성변수의 중복성 최소화하는 기법

- 임베디드 기법(Embedded Method)

    - 모델 자체에 변수 선택이 포함된 기법
    - 절차: 변수 전체집합 -> (하위 집합 -> 학습 + 평가)를 반복
    - 특징
        - 모델 정확도에 기여하는 변수를 학습
        - 제약조건: 더 적은 계수를 가지는 회귀식을 찾는 방향으로 제어
    - 사례: 라쏘 / 릿지 / 엘라스틱넷 / SelectFromModel
        - 라쏘(LASSO): 가중치 절댓값 합을 최소화 -> L1-norm
        - 릿지(Lidge): 가중치 제곱 합을 최소화 -> L2-norm
            - Norm: 벡터의 크기(길이)를 측정하는 방법. L1-norm은 벡터 p, q 각 원소간 차이의 절댓값의 합, L2-norm은 유클리디안 거리(직선 거리)
        - 엘라스틱넷 (Elastic Net): 가중치 절댓값 합, 제곱 합을 동시에 제약 -> 라쏘와 릿지를 선형결합
        - SelectFromModel: 의사결정나무 기반 알고리즘으로 변수 선택

==========

2) 차원축소

(1) 차원축소(Dimensionality Reduction) 개념

- 차원축소

    - 분석대상인 여러 변수의 정보를 최대한 유지하면서 변수 개수를 최대한 줄이는 탐색적 분석기법
- 특성변수(설명변수)만 사용함 -> 비지도 학습 머신러닝 기법

(2) 차원축소 특징

- 축약된 변수세트: 원래 전체 변수들의 정보를 최대한 유지해야 함
- 결합변수: 변수간 내재된 특성, 관계를 분석 -> 선형/비선형 결합변수 -> 결합변수만으로 전체변수 설명 가능하도록
- 차원축소의 목적: 다른 분석 전 단계 / 분석 후 개선 / 효과적인 시각화 등
- 장점: 고차원보다 저차원으로 학습하면 머신러닝 알고리즘이 더 잘 작동함 / 시각화 쉬움

(3) 차원축소 기법

- 주성분분석(PCA) / 특이값분해(SVD) / 요인분석 / 독립성분분석(ICA) / 다차원척도법(MDS)

- 주성분분석(PCA; Principal Component Analysis)

    - 변수들의 공분산행렬 or 상관행렬 이용
    - 정방행렬만 사용(행 개수 = 열 개수)
    - 고차원 공간의 표본들 -> 선형 연관성 없는 저차원 공간으로 변환

- 특이값분해(SVD; Singular Value Decomposition)

    - M x N 차원 행렬
    - 특이값 추출 -> 데이터셋을 효과적으로 축약

- 요인분석(Factor Analysis)

    - 관찰할 수 없는 잠재적인 변수(Latent Variable)가 존재한다고 가정함
    - 잠재요인을 도출 -> 데이터 안의 구조 해석

- 독립성분분석(ICA; Independent Component Analysis)

    - 다변량 신호 -> 통계적으로 독립적인 하부성분으로 분리
    - 독립성분이 비정규 분포를 따르게 됨

- 다차원척도법(MDS; Multi-Dimensional Scaling)

    - 개체들 간 유사성, 비유사성 측정 -> 2차원 or 3차원 공간상에 점으로 표현(시각화)

- 차원축소 기법 주요 활용분야

    - 탐색적 데이터 분석
    - 주요 특징(변수)을 추출하여 타 분석기법의 설명변수로 활용
    - 텍스트에서 주제, 개념 추출
    - 비정형 데이터(이미지, 사운드 등)에서 특징 패턴 추출
    - 상품 추천시스템 알고리즘 구현 및 개선
    - 다차원 공간 정보를 저차원으로 시각화
    - 공통 요인을 추출하여 잠재된 데이터 규칙 발견

==========

3) 파생변수 생성

(1) 파생 변수(Derived Variable)

- 기존 변수에 특정 조건 or 함수 등을 사용하여 새롭게 재정의한 변수
- 변수 조합 or 함수 적용 -> 새 변수를 만들어 분석
- 논리적 타당성, 기준을 가지고 생성해야 함
- 파생 변수 생성 방법: 단위 변환 / 표현방식 변환 / 요약 통계량 변환 / 변수 결합(함수 등의 수학적 결합)

==========

4) 변수 변환

(1) 변수 변환(Variable Transformation)

- 불필요한 변수를 제거, 변수를 반환, 새로운 변수를 생성시킴
- 변수들이 로그 / 제곱 / 지수 등의 모습 -> 변수 변환 -> 선형관계로 만들면 분석하기 쉬움

(2) 변수 변환 방법

- 단순 기능 변환 / 비닝 / 정규화 / 표준화
- 단순 기능 변환(Simple Functions Transformation)
    - 한쪽으로 치우친 변수를 단순한 함수로 변환
    - 로그: 오른쪽으로 기울어진 분포를 변경
    - 제곱 / 세제곱 / 루트: 로그에 비해선 덜 사용됨
- 비닝(Binning)
    - 기존 데이터를 범주화 / 몇개의 Bin or Bucket으로 분할
    - 기존 비즈니스에 대한 도메인 지식 필요
- 정규화(Normalization)
    - 데이터를 특정 구간으로 바꾸는 척도법
    - 최소-최대 정규화 / Z-스코어 정규화
- 표준화(Standardization)
    - 0 중심 양쪽으로 데이터를 분포시킴
    - Z = (x - x_bar)/s
- 변수 변환 사례(일반적인 마케팅의 사례)
    - 매출 / 판매수량 / 가격 / 가구소득: 로그
    - 지리적 거리: 역수 / 로그
    - 효용에 근거한 시장점유율 / 선호점유율: e^x / (1+e^x)
    - 우측으로 꼬리가 긴 분포 -> 루트 / 로그
    - 좌측으로 꼬리가 긴 분포 -> 제곱

==========

5) 불균형 데이터 처리

(1) 불균형 데이터 처리

- 타겟 데이터 수가 매우 극소수인 경우, 불균형 데이터 처리 수행

- 불균형 데이터 처리를 수행 -> 정밀도(Precision) 향상
- 처리 기법: 언더샘플링/ 오버샘플링/ 임곗값이동/ 앙상블기법

(2) 언더 샘플링 (Under Sampling)

- 다수 클래스 중 일부만 선택 -> 데이터 비율을 맞춤
- 일부만 선택 -> 데이터 소실 매우 큼 / 중요한 정상 데이터 소실 가능성
- 기법: 랜덤 언더 샘플링 / ENN / 토멕 링크 방법 / CNN / OSS
    - 랜덤 언더 샘플링: 무작위 선택
    - ENN(Edited Nearest Neighbours): 소수클래스 주위에 인접한 다수클래스 데이터 제거
    - 토멕 링크 방법(Tomek Link Method): 다수클래스의 토멕 링크 제거
        - 토멕 링크: 클래스 구분 경계선 가까이 있는 데이터
    - CNN(Condensed Nearest Neighbor): 다수클래스에 밀집된 데이터가 없을 때까지 제거
    - OSS (One Sided Selection): 토멕 링크 + CNN

(3) 오버 샘플링(Over Sampling)

- 소수 클래스 데이터를 복제 or 생성 -> 데이터 비율을 맞춤
- 장점: 정보 손실 없음 / 알고리즘 성능 높음
- 단점: 과적합 가능성 / 검증 성능 나빠질 수 있음
- 기법: 랜덤 오버 샘플링 / SMOTE / Borderline-SMOTE / ADASYN
    - 랜덤 오버 샘플링: 무작위 복제
    - SMOTE(Synthetic Minority Over-sampling TEchnique)
        - 가상의 직선 위에 데이터 추가
        - 소수클래스의 중심 데이터와 주변 데이터 사이에 가상의 직선을 만듦
    - Borderline-SMOTE: 다수클래스와 소수클래스의 경계선에서 SMOTE
    - ADASYN(ADAptive SYNthetic): 모든 소수클래스에서 다수클래스의 관측비율 계산 -> SMOTE 적용

(4) 임계값 이동(Threshold-Moving)

- 임계값을 데이터가 많은 쪽으로 이동시킴
- 임계값(Threshold): 귀무가설 기각 여부를 구분하는 값
- 변화 없이 학습 -> 테스트 단계에서 임계값 이동

(5) 앙상블 기법(Ensemble Technique)

- 서로 같거나 다른 여러 모형들의 예측/분류 결과를 종합