데이터 전처리

언더 샘플링 & 오버 샘플링

- 레이블이 불균형한 분포를 가진 데이터를 학습 시킬 때 예측 성능의 문제가 발생할 수 있음

> 금융 사기 데이터에서 사기를 나타내는 데이터는 적을 수 밖에 없음
>
> 사기 검출 (Fraud Detection) & 이상 검출 (Anomaly Detection)

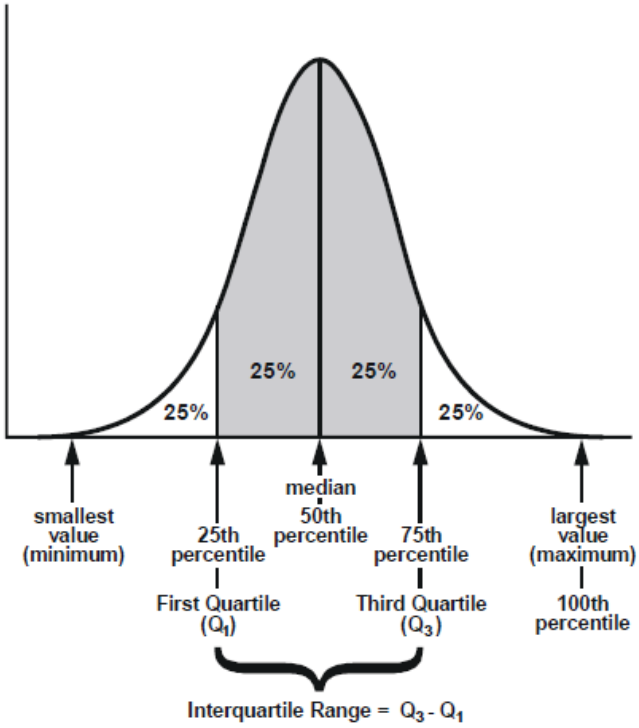
  - 이런 극도로 불균형한 레이블 분포로 인한 문제를 해결하기 위해 적절한 학습 데이터 확보 방안은 대표적으로 `오버 샘플링 (Oversampling)`과 `언더 샘플링 (Undersampling)` 방법이 존재 (오버 샘플링이 주로 사용됨)

    1. 오버 샘플링 (Oversampling)
       - 적은 비율의 라벨 데이터 세트를 증식하여 학습을 위한 충분한 데이터를 확보하는 방법
       - 동일한 데이터를 단순히 증식시키면 과적합이 되기 때문에 사용하지 않고, 원본 데이터의 피처 값들을 아주 약간만 변경하여 데이터 세트를 증식
       - 대표 적인 방법으로 SMOTE (Synthetic Minority Over-sampling Technique)이 있음
         - 적은 데이터 세트의 있는 개별 데이터들의 K 최근접 이웃(K Nearest Neighbor)을 찾아서 이 데이터와 K개 이웃들의 차이를 일정 값으로 만들어 기존 데이터와 약간만 차이가 가는 새로 운 데이터들을 생성하는 방식
    2. 언더 샘플링 (Undersampling)
       - 많은 비율의 데이터 세스트를 적은 데이터 세트 수준으로 감소 시키는 방식
         - 보통 레이블 10,000건, 이상 레이블 100건이라고 할 때, 보통 레이블을 100건으로 줄이는 방식
       - 오히려 정상 레이블의 학습 수행을 방해하는 단점이 있어 잘 쓰이지 않는 방식임

- 대부분의 선형 모델은 중요 피처들의 값이 정규 분포 형태를 유지하는 것을 선호
  - 사이킷런의 StandardScaler 클래스를 이용해 피처를 중규 분호 형태로 변환 가능
  - 로그변환 - 데이터 분포도가 심각하게 왜곡되어 있을 경우, 원래 값을 log 값으로 변환해 원래 큰 값을 상대적으로 작은 값을 변환

이상치 데이터 (Outlier)

- Outlier (이상치)는 전체 데이터의 패턴에서 벗어난 이상 값을 가진 데이터
- 이상치를 찾는 방식
  - IQR
    - 사분위(Quantile) 값의 편차를 이용하는 기법 (Box Plot 으로 시각화 가능)
    - IQR 은 Q3 - Q1



  - 최솟값 : 분위수 Q1에 IQR*1.5 뺀 수*
  - 최댓값 : 분위수 Q3에 IQR*1.5 더한 수
- 이상치는 결정값과 가장 상관성(Correlation)이 높은 데이터 피처의 이상치를 검출하는 것이 효율적임