<불균형 데이터란?>

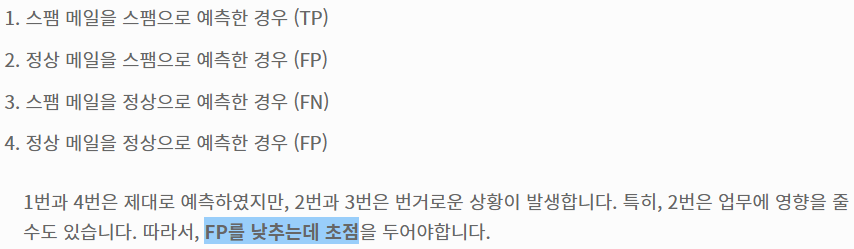
* 남/여, 구매여부 등 클래스 분포를 예측해야 하는 분류문제에서 라벨 값의 분포가 100:1, 200:1 등으로 불균형하게 나타나는 상태

<불균형 데이터로 인한 문제>

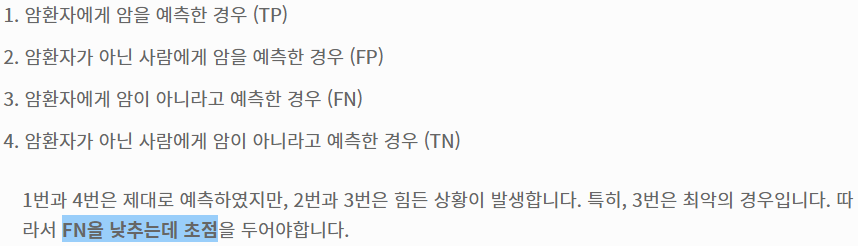
* 과적합 문제 발생
  + 과적합은 변수가 많아서 생기는 모델 복잡성, 데이터 불균형 등 원인이 다양
  + Train set에서는 높은 성능을 보이지만 새로운 데이터에서는 예측 성능이 더 낮을 수 있다.
  + 모델을 만들 때 일반화(어떠한 데이터에서도 비슷한 성능을 보여주는 것) 정도가 높은 모델을 만드는 것이 중요 과적합 문제는 이것을 방해함.
* 데이터가 불균형하다면 분포도가 높은 클래스에 모델이 가중치를 많이 두기 때문에 모델 자체에서는 ‘분포가 높은 것으로 예측하게 된다면 어느정도는 맞힐 수 있겠지?’라고 판단하기 때문에 Accuracy는 높아질 수 있지만 분포가 작은 값에 대한 Precision은 낮을 수 있고, 분포가 작은 클래스의 재현율이 낮아지는 문제가 발생할 수 있음

<분류 문제의 평가 방법>

1. Confusion matrix
2. Accuracy
   1. (정상기업을 제대로 정상기업으로 분류+부도기업을 제대로 부도기업으로 분류) / 전체
   2. 가장 직관적인 모델 예측 성능을 나타내는 평가 지표이지만, 데이터의 불균형이 심하다면 모델 성능을 판단하기에 좋은 지표는 아님
3. Precision 정밀도
   1. 부도 기업을 제대로 부도로 예측 / 실제 부도



1. Recall 재현율
   1. 부도 기업을 제대로 부도로 예측 / 예측 부도



불균형 데이터에서는 Accuracy대신 Precision과 Recall를 모델 평가 지표로 주로 사용함.

하지만 정밀도와 재현율은 상호 보완적인 평가 지표이기 때문에 강제로 하나를 올리면, 다른 하나는 떨어지긱 쉽다. (정밀도/ 재현율 trade-off)

임계치에 따라 정밀도와 재현율이 양 극으로 치우치는 경우를 방지하기 위해 이를 조합한 지표를 사용하기도 하는데 이것이 F1 score이다.

1. F1score (Recall과 Precision의 조화평균)
   1. Recall과 Precision 둘 중 한쪽으로 치우치지 않는 경우 높은 값을 가진다.

위의 지표들은 극단적인 불균형 데이터 셋에서는 의미가 없다. 이를 보완하기 위한 것이 ROC곡선과 AUC

1. ROC curve :ROC 곡선은 FPR, TPR 관계를 그린 곡선
   1. FPR: False Positive Rate
   2. TPR: True Positive Rate (=재현율)
2. AUC ROC 곡선 아래 면적

불균형 - > F1 (Recall & Precision) -> 균형잡혀야 된다. 하지만 우리는 Recall에 더 치중해야 한다. 따라서 F1을 고려하지 않고 대안으로 ROC, AUC를 체택한다.

Recall, Precision, F1-Score으로 분류를 판단하기에는 TP 분류를 알기 어렵다. 또한, 은행 여신 리스크 관리 분석 목적에 맞게 Recall에 더 집중해야 한다.

따라서 목적에 맞는 부도기업 정분류를 model 1차 평가지표로 사용하고 Data 정분류율인 Accurcay를 2차 평가지표로 사용하여 가장 좋은 모델을 찾아낼 것이다.

또한 Precision, F1-Score 보완책으로 ROC, AUC를 활용한다.