머신러닝 모델의 예측 성능 평가 방법

* 회귀의 성증 평가 지표: 오차 평균값에 기반
* 분류의 성능 평가 지표

결정 클래스 값 종류의 유형에 따라 긍정/부정 같은 2개의 결과 값만 갖는 이진 분류 + 여러 개의 결정 클래스 값 가지는 멀티 분류

1. **정확도(Accuracy)**

정확도 = 예측 결과 동일한 데이터 수 / 전체 예측 데이터 수

직관적으로 성능 나타내지만 이진 분류의 경우 ML 모델의 성능 왜곡 가능

🡪 정확도만 가지고 성능 평가 X

불균형한 레이블 값 분포에서 정확도는 적합한 평가 지표 X

한계점을 극복하기위해 여러 분류 지표와 함께 적용

1. **오차행렬(Confusion Matrix)**

이진 분류의 예측 오류가 얼마인지 + 어떤 유형의 예측 오류인지 보여줌

4분면 행렬에서 실제 레이블 값과 예측 레이블 값 유형에 따라 TN, FP, FN, TP

앞 T/F는 예측 값과 실제 값이 같은 지 다른 지, 뒤 N/P는 예측 결과 값이 부정(0)인지 긍정인지(1) -> array[0,0]: TN, array[0,1]: FP, array[1,0]: FN, array[1,1]: TP

정확도 = (TN+TP)/(TN+FP+FN+TP)

보통 불균형한 이진분류 데이터 세트는 적은 수의 결과값에 Positive 1 부여, 아니면 Negative 0 부여 --> Negative로 예측할 때 TN 커지고 TP, FN, FP 작아짐. 따라서, Negative에 대한 예측 정확도만으로 분류의 정확도가 높아지는 오류 발생

1. **정밀도(Precision)와 재현율(Recall)**

Positive 데이터 세트의 예측 성능에 초점을 맞춘 평가 지표

정밀도 = TP / (FP + TP) : 예측을 Positive로 한 대상 중 예측과 실제 값이 Positive로 일치한 데이터의 비율(양성 예측도)

재현율 = TP / (FN + TP) : 실제 값이 Positive인 대상 중 예측과 실제 값이 Positive로 일치한 데이터의 비율(민감도, True Positive Rate)

재현율이 중요 지표인 경우: 실제 Positive 데이터를 Negative로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우 🡪 예) 암 판단 모델: 실제 Positive 양성 환자를 Negative 음성으로 잘못 판단했을 때는 심각, 금융 사기 적발 모델

정밀도가 중요 지표인 경우: 실제 Negative 데이터를 Positive로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우 🡪 예) 스팸 메일 여부: Negative인 일반 메일을 Positive인 스팸 메일로 분류하면 안됨

둘 다 TP를 높이는데 초점을 맞추지만 재현율은 FN 낮추는데, 정밀도는 FP 낮추는데 초점

**정밀도와 재현율의 트레이드오프**

분류의 결정 임곗값(Threshold)을 조정해 정밀도 또는 재현율의 수치를 높일 수 있음, 둘은 상호 보완적이므로 한 쪽 높이면 다른 쪽 낮아짐

개별 레이블 별로 결정확률 구하고 예측 확률이 큰 레이블 값으로 예측하게 됨. 일반적으로 임계값은 0.5

predict\_proba(): 개별 데이터 별로 예측 확률 반환, 학습 완료된 Classifier에서 호출하여 테스트 데이터를 파라미터로 입력하면 개별 클래스 예측 확률 반환

-> ndarray m\*n (m: 입력 값의 레코드 수, n: 클래스 값 유형) 형태로 반환, 각 열은 개별 클래스의 예측 확률

predict()는 predict\_proba()의 결과 배열에서 분류 결정 임계값보다 큰 값이 들어있는 칼럼의 위치를 받아서 최종적으로 예측 클래스 결정

임계값 낮췄더니 정밀도 감소, 재현율 증가 - True 값이 많아지게 되므로 재현율 값 높아짐, 양성 예측을 많이 하여서 실제 양성을 음성으로 예측하는 횟수 상대적으로 감소

정밀도 또는 재현율 중 하나에 상대적인 중요도를 부여해 각 상황에 맞게 알고리즘 튜닝할 수 있지만, 하나를 극단적으로 높이면 안됨

-> 정밀도 100% 경우: 확실한 기준이 되는 경우만 Positive 나머지는 Negative로 예측

확실한 Positive가 1명이면 이 한 명만 Positive로 예측하고 나머지는 Negative

정밀도 = TP/(TP+FP) = 1/(1+0) = 100%

-> 재현율 100%인 경우: 모두 Positive예측

전체 환자 중 실제 양성인 사람 30명인 경우

재현율 = TP/(TP+FN) = 30/(30+0) = 100%

정밀도와 재현율의 수치가 적절하게 조합되어야 분류의 종합적인 성능 평가 가능

1. **F1 스코어**

정밀도와 재현율을 결합한 지표, 둘 중 어느 쪽으로 치우치지 않을 때 F1 스코어 값이 상대적으로 높음

F1 = 2 / (1/recall + 1/precision) = 2 \* (precision \* recall) / (precision + recall)

f1\_score()이라는 API 제공

1. **ROC AUC**

정리

이진분류의 레이블 값이 불균형하게 분포될 경우 정확도만으로 머신러닝 모델의 예측 성능을 평가할 수 X

오차행렬: Negative와 Positive 값을 가지는 실제 클래스 값과 예측 클래스 값이 True, False에 따라 TN, FP, FN, TP 로 매핑되는 4분면 행렬

정밀도와 재현율: Positive 데이터 세트의 예측 성능에 좀 더 초점을 맞춤. 정밀도 또는 재현율이 특별히 강조되어야 할 경우 임계값 조정하여 정밀도 또는 재현율의 수치를 높일 수 있음.

F1 스코어: 정밀도와 재현율을 결합한 지표, 정밀도와 재현율이 어느 한쪽으로 치우치지 않을 때 높은 값을 가짐

ROC-AUC: 일반적으로 이진분류의 성능 평가를 위해 가장 많이 사용되는 지표