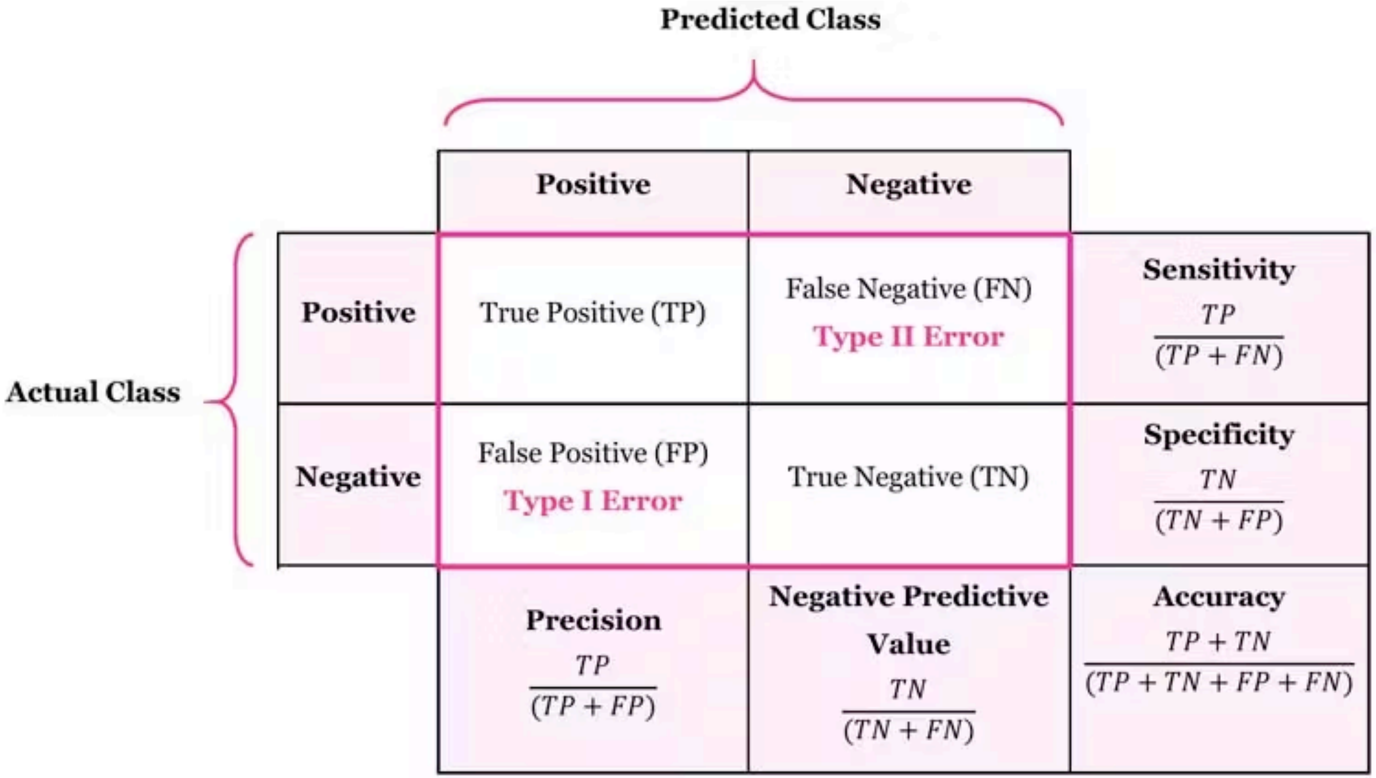
230924_7기_데분중 평가지표

BDA 7기 데분중

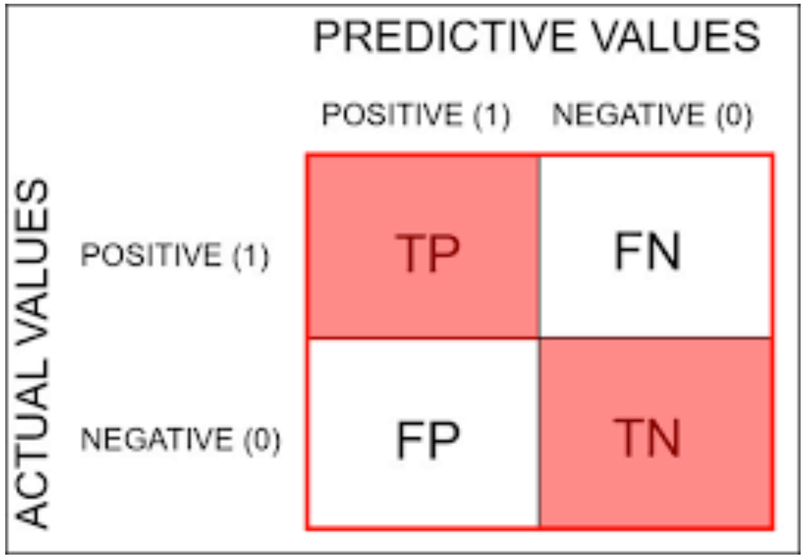
혼동행렬(Confusion Matrix)



분류 기준값

Threshold 0.5 일반적

정확도 Accuracy



$$\text{정확도(Accuracy)} = \frac{\text{예측 결과가 동일한 데이터 건수}}{\text{전체 예측 데이터 건수}}$$

Accuracy = (TP+TN) / (TP+TN+FP+FN)

1은 1로, 0은 0으로 모델이 분류한다는 비중

Error rate = (FP+FN) / (TP+TN+FP+FN)

1- 정확도는 = 오분류율 (Error rate)

정확도 이슈

불균형 데이터

사기탐지 사기 1 정상 0

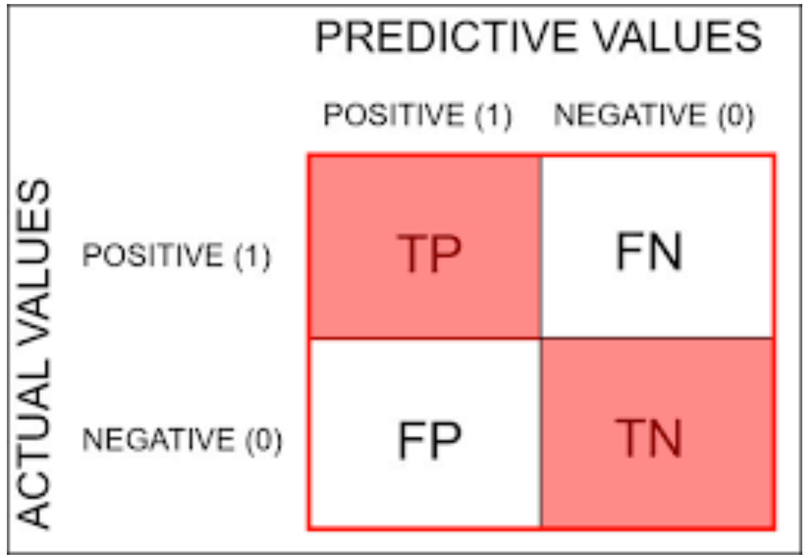
양성예측 양성 1 음성 0

무슨 문제가 발생할까?

정밀도 (Precision)

‘1’으로 예측하여 분류한 관측치 중 실제값도 ‘1’인 비중

Precision = TP / (TP+FP)

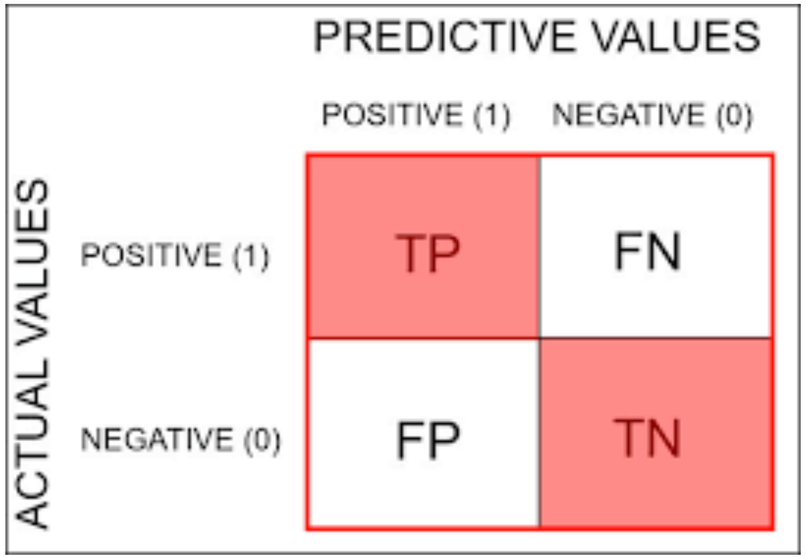


Precision = 실제값1 예측값1

(실제값1 예측값 1+실제값 0 예측값 1)

모델이 실제 1을 제대로 분류하는 성능이 얼마나 우수한지?

재현율 (Recall) or 민감도(Sensitivity)



$$Recall = \frac{TP}{TP + FN}$$

모델이 정밀도가 높아도 실제 '1'인 관측치를 너무 적게 찾아내면 좋은모델이라 할 수 있을까? (최소 50% 이상 유지)

Recall = 실제값1 예측값1

(실제값1 예측값 1+실제값 1 예측값 0)

정밀도와 재현율 관계

Precision = TP / (TP+FP)

정밀도가 중요한 경우?

실제 Negative 음성 데이터를 Positive 로 잘못판단한 경우

Recall = TP / (TP+FN)

재현율이 중요한 경우?

실제 Positive 양성 데이터를 Negative로 잘못판단

결국 FP와 FN을 낮추는 것에 초점과 TP를 올리는 것 초점

둘은 보완적인 지표 둘 다 높은 수치면 가장 좋지만

하나만 높은 수치는 좋은 수치가 아니다.

Trade off

F1 스코어

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

정밀도와 재현율 조화평균

정밀도 0.5 재현율 0.5	0.5
정밀도 0.7 재현율 0.4	0.51

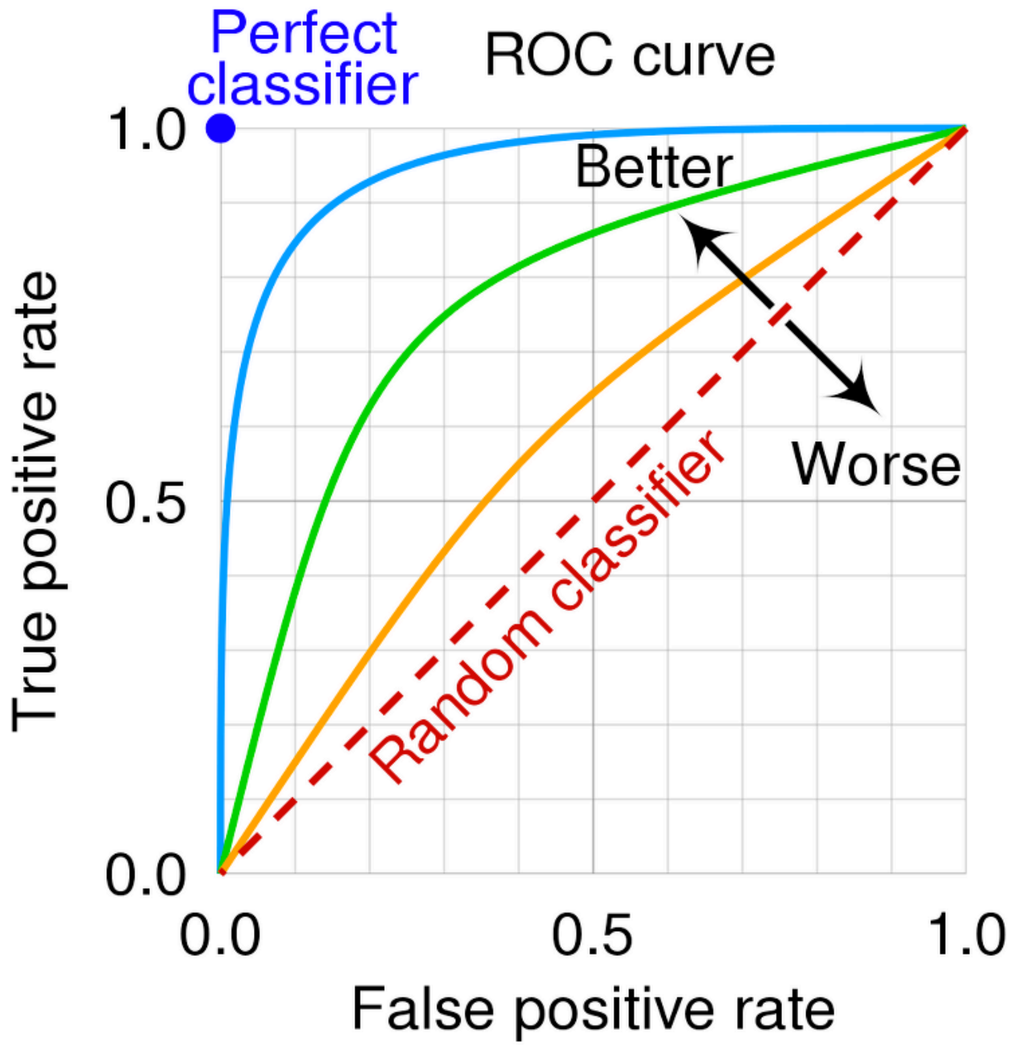
둘은 Trade off 관계

정밀도를 높인다? 실제 ‘1’일 것이 거의 확실한 관측치만 ‘1’로 예측

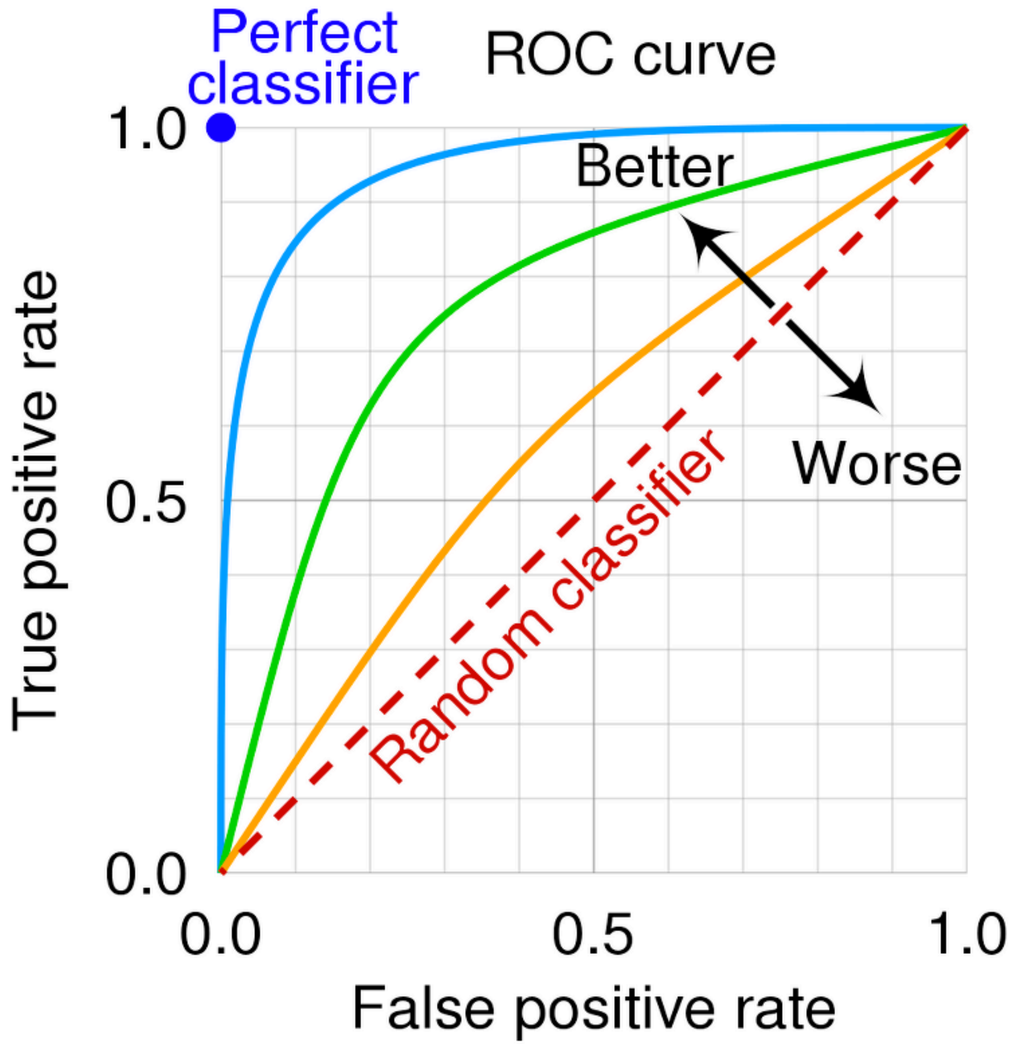
따라서 전체 실제’1’중에 ‘1’로 분류되는 비중인 민감도는 감소할 수 밖에 없음

ROC, AUC

- ROC(Receiver operator Characteristic)
- AUC(Area Under Curve) 최소 0.5 이상
- X축 FPR (False Positive Rate)
- Y축 TPR (True Positive Rate)
- 임계값을 1부터 0까지 변화시키면서 FPR을 구하고 FPR 변화에 따른 TPR값 구하는 것이 ROC곡선



ROC, AUC



FPR = FP / (FP+TN) = 1 - TNR = 1-특이성

FPR을 0부터 1까지 변경하면서 TPR의 변화값

임계값의 변경

필수과제 1

특이도 Specificity

G-mean (기하평균)