평가

분류(Classification) 성능 평가 지표

- 정확도(Accuracy)
- 오차행렬(Confusion Matrix)
- 정밀도(Precision)
- 재현율(Recall)
- F1 Score
- ROC AUC

정확도(Accuracy)

정확도(Accuracy) = 예측 결과가 동일한 데이터 건수 / 전체 예측 데이터 건수

- 정확도는 직관적으로 모델 예측 성능을 나타내는 평가 지표이다. 하지만 이진 분류의 경우 데이터의 구성에 따라 모델의 성능을 왜곡할 수 있기 때문에 정확도 수치 하나만 가지고 성능을 평가하지 않는다.
- 특히 정확도는 불균형한(imbalanced) 레이블 값 분포에서 모델의 성능을 판단할 경우 적합한 평가 지표가 아니다.

정확도의 문제점

y = [0,0,0,0,0,0,0,0,0,1]

- 위의 타겟에서 맞춰야 할 양성 데이터는 1개이다.
- 모델을 만들지 않고 임의로 모두 0으로 예측을 해도 정확도는 90%가 나오게 된다!

오차 행렬(Confusion Matrix)

- 오차 행렬은 이진 분류의 예측 오류가 얼마인지와 더불어 어떤 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표이다.

	예측 Negative(0)	예측 Positive(1)
실제 Negative(0)	TN (True Negative)	FP (False Positive)
실제 Positive(1)	FN (False Negative)	TP (True Positive)

오차 행렬(Confusion Matrix)

- 오차 행렬은 이진 분류의 예측 오류가 얼마인지와 더불어 어떤 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표이다.

	예측 Negative(0)	예측 Positive(1)
실제 Negative(0)	음성을 음성으로 잘 예측	음성을 양성으로 잘못 예측
실제 Positive(1)	양성을 음성으로 잘못 예측	양성을 양성으로 잘 예측

오차 행렬(Confusion Matrix)

- 만약 [0,0,0,0,0,0,0,0,0,1] 을 모두 0으로 예측하면? 정확도는 90% 이지만…

	예측 Negative(0)	예측 Positive(1)
실제 Negative(0)	9	0
실제 Positive(1)	1	0

분명히 양성 데이터가 있음에도 불구하고 Positive로 하나도 예측하지 않은 것을 확인 할 수 있다.

정밀도(Precision)와 재현율(Recall)

- 정밀도 = TP / (FP + TP)
  - Negative를 Positive로 잘못 예측하면 정밀도가 내려간다.

	예측 Negative(0)	예측 Positive(1)
실제 Negative(0)	TN	FP
실제 Positive(1)	FN	TP

정밀도는 예측을 Positive로 한 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율

정밀도(Precision)와 재현율(Recall)

- 재현율 = TP / (FN + TP)
  - Positive를 Negative로 잘못 예측하면 재현율이 내려간다.

	예측 Negative(0)	예측 Positive(1)
실제 Negative(0)	TN	FP
실제 Positive(1)	FN	TP

재현율은 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율

업무에 따른 재현율과 정밀도의 상대적 중요도

- 재현율(Recall)이 상대적으로 더 중요한 지표인 경우
  - 실제 Positive 데이터 예측을 Negative로 잘못 판단하면 업무상 큰 영향이 발생하는 경우.
    암 진단, 금융 사기 판별 등

- 정밀도(Precision)가 상대적으로 더 중요한 지표인 경우
  - 실제 Negative 데이터 예측을 Positive로 잘못 판단하면 업무상 큰 영향이 발생하는 경우.
    스팸 메일 등

업무에 따른 재현율과 정밀도의 상대적 중요도

- 재현율(Recall)이 상대적으로 더 중요한 지표인 경우
  - 실제 Positive 데이터 예측을 Negative로 잘못 판단하면 업무상 큰 영향이 발생하는 경우.
    암 진단, 금융 사기 판별 등

- 정밀도(Precision)가 상대적으로 더 중요한 지표인 경우
  - 실제 Negative 데이터 예측을 Positive로 잘못 판단하면 업무상 큰 영향이 발생하는 경우.
    스팸 메일 등

불균형한 레이블 클래스를 가지는 이진 분류 모델에서는 많은 데이터 중에서 중점적으로 찾아야 하는 매우 적은 수의 결괏값에 Positive(1)를 설정한다.

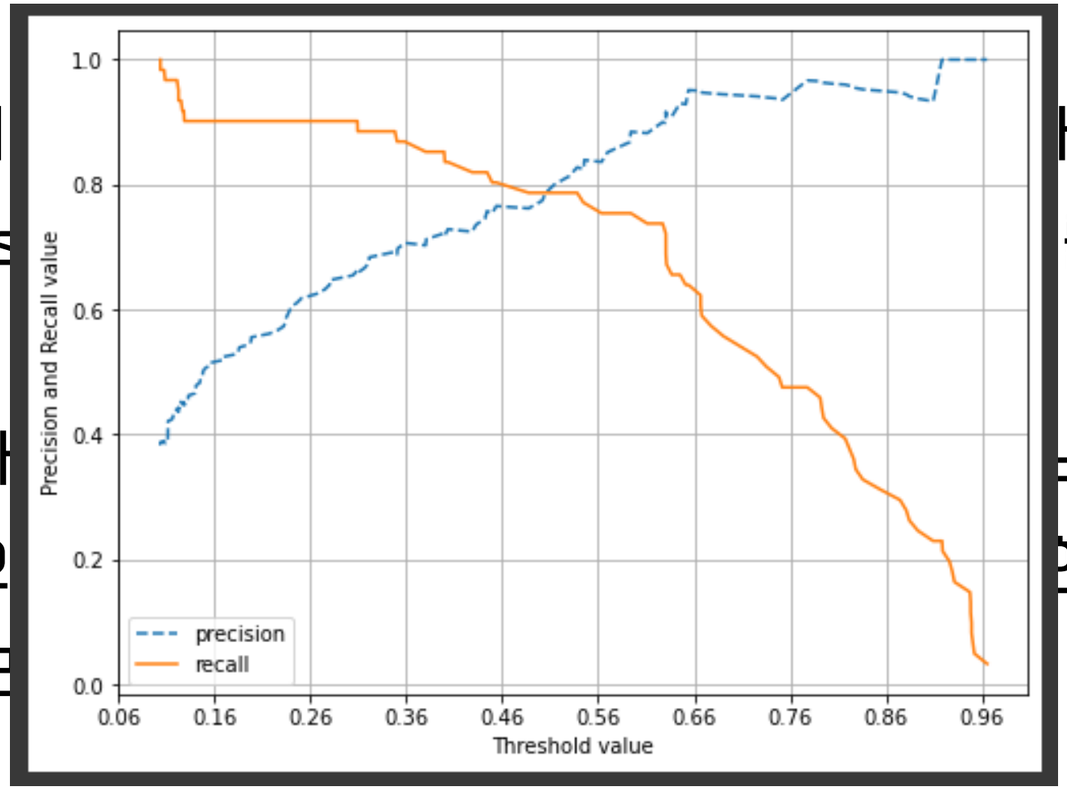
정밀도/재현율 트레이드 오프

- 분류하려는 업무의 특성상 정밀도 또는 재현율이 특별히 강조돼야 할 경우 분류의 결정 임곗값(Threshold)을 조정해 정밀도 또는 재현율의 수치를 높일 수 있다.

- 하지만 정밀도와 재현율은 상호 보완적인 지표이기 때문에 어느 한쪽을 강제로 높이면 다른 하나의 수치는 떨어지기가 쉽다. 이를 정밀도 / 재현율의 트레이드 오프(trade-off)라고 한다.

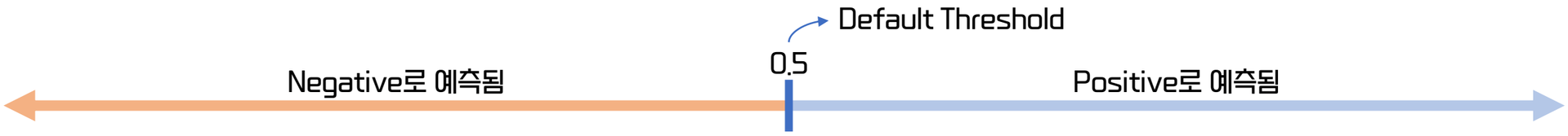
정밀도/재현율 트레이드 오프

- 분류하려는 업무의 …해야 할 경우 분류의 결정 임곗값(Thres… …를 높일 수 있다.
- 하지만 정밀도와 재… …는 한쪽을 강제로 높이면 다른 하나의… …현율의 트레이드 오프(trade-off)라…



분류 결정 임곗값(Threshold)에 따른 Positive 예측 확률 변화

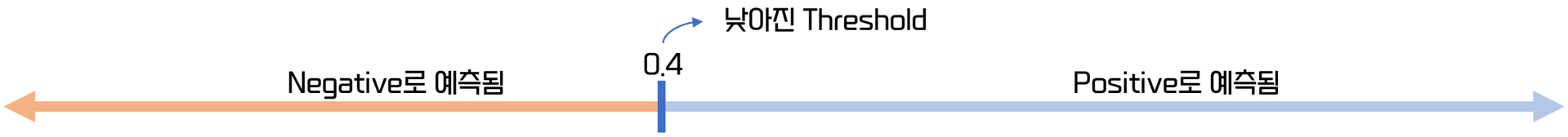
- 정밀도 = TP / (FP + TP)
  - Positive로 잘 예측 / Positive로 잘못 예측 + Positive로 잘 예측
  - 실제는 Negative 인데, Positive로 잘못 예측하면 정밀도는 떨어진다.
- 재현율 = TP / (FN + TP)
  - Positive로 잘 예측 / Negative로 잘못 예측 + Positive로 잘 예측
  - 실제는 Positive인데, Negative로 잘못 예측하면 재현율은 떨어진다.



Positive로 예측되는 기준. Threshold! 이 수치를 넘어가면 Positive가 된다. 기본은 0.5
사이킷런 Estimator의 predict_proba() 메소드는 분류 결정 예측 확률을 반환하고, 이 값이 0.5가 넘으면 1로 예측한다.

분류 결정 임곗값(Threshold)에 따른 Positive 예측 확률 변화

- 정밀도 = TP / (FP + TP)
  - Positive로 잘 예측 / Positive로 잘못 예측 + Positive로 잘 예측
  - 실제는 Negative 인데, Positive로 잘못 예측하면 정밀도는 떨어진다.
- 재현율 = TP / (FN + TP)
  - Positive로 잘 예측 / Negative로 잘못 예측 + Positive로 잘 예측
  - 실제는 Positive인데, Negative로 잘못 예측하면 재현율은 떨어진다.



Threshold를 낮추면 Positive로 예측될 확률이 커지게 된다. 즉 Negative로 잘못 예측할 확률이 낮아지므로 재현율이 상승하게 된다.

분류 결정 임곗값(Threshold)에 따른 Positive 예측 확률 변화

- 정밀도 = TP / (FP + TP)
  - Positive로 잘 예측 / Positive로 잘못 예측 + Positive로 잘 예측
  - 실제는 Negative 인데, Positive로 잘못 예측하면 정밀도는 떨어진다.
- 재현율 = TP / (FN + TP)
  - Positive로 잘 예측 / Negative로 잘못 예측 + Positive로 잘 예측
  - 실제는 Positive인데, Negative로 잘못 예측하면 재현율은 떨어진다.

Threshold를 높이면 Negative로 예측될 확률이 커지게 된다. 즉 Positive로 잘못 예측할 확률이 낮아지므로 정밀도가 상승하게 된다.

정밀도와 재현율의 맹점

- 정밀도를 100%로 만드는 법. 정밀도 = TP / (TP + FP), FP = 0
  - 확실한 기준이 되는 경우만 Positive로 예측하고 나머지는 모두 Negative로 예측.
  - 전체 1000명의 환자 중 1명만 확실한 환자면 이 1명만 Positive로 예측하고 나머지를 모두 Negative로 예측하더라도 FP=0, TP=1이 되므로 1 / (1 + 0)으로 100%가 된다.

- 재현율을 100%로 만드는 법. 재현율 = TP / (TP + FN), FN = 0
  - 모든 데이터를 Positive로 예측
  - 전체 1000명의 환자를 모두 Positive로 예측하면 실제 Positive인 사람이 30명 정도라도 TN이 수치에 포함되지 않고 FN은 아예 0이므로 30 / (30 + 0)으로 100%가 된다.

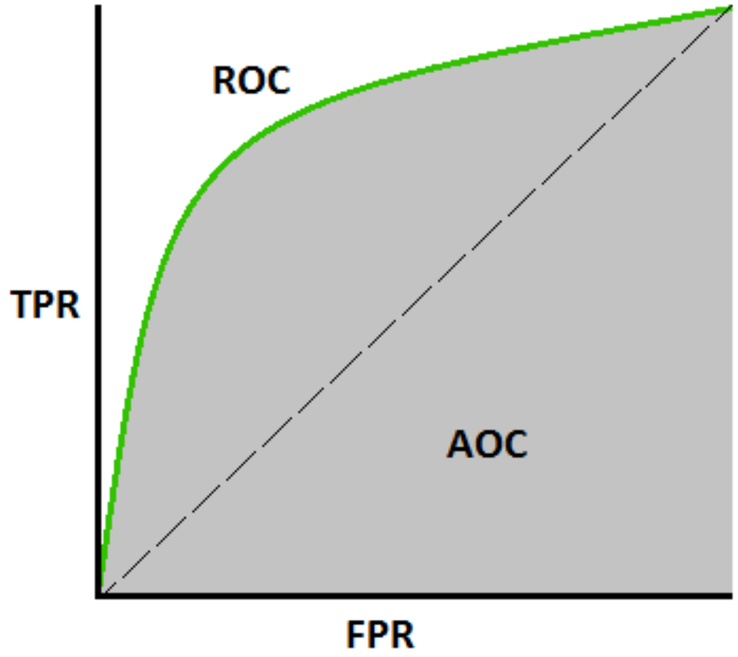
F1 Score

- F1 스코어는 정밀도와 재현율을 결합한 지표로서, 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가지게 된다.

$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall}$$

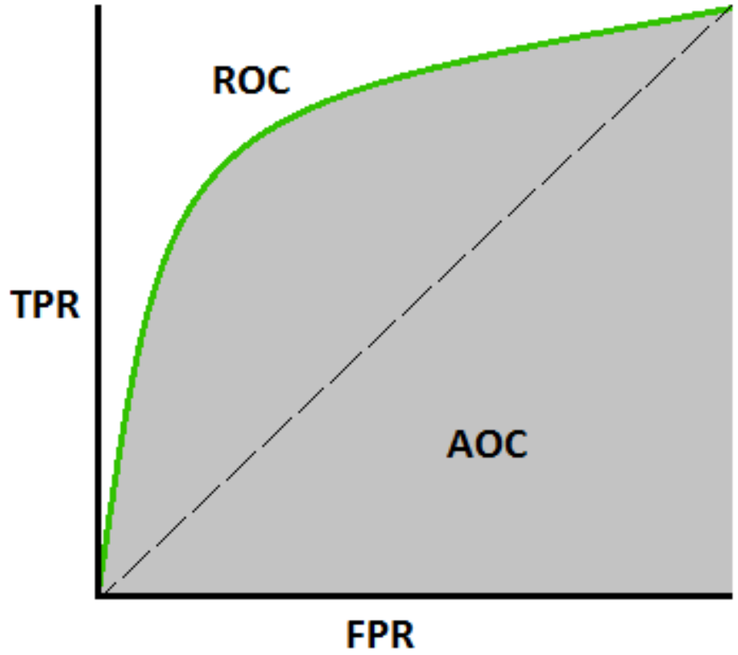
- 만일 모델 A의 정밀도가 0.9, 재현율이 0.10이면 모델 A의 F1 Score는 0.180이다.
- 만일 모델 B의 정밀도가 0.5, 재현율이 0.50이면 모델 B의 F1 Score는 0.5로 B 모델이 A 모델에 비해 우수한 F1 Score를 가지게 된다.

ROC 곡선과 AUC



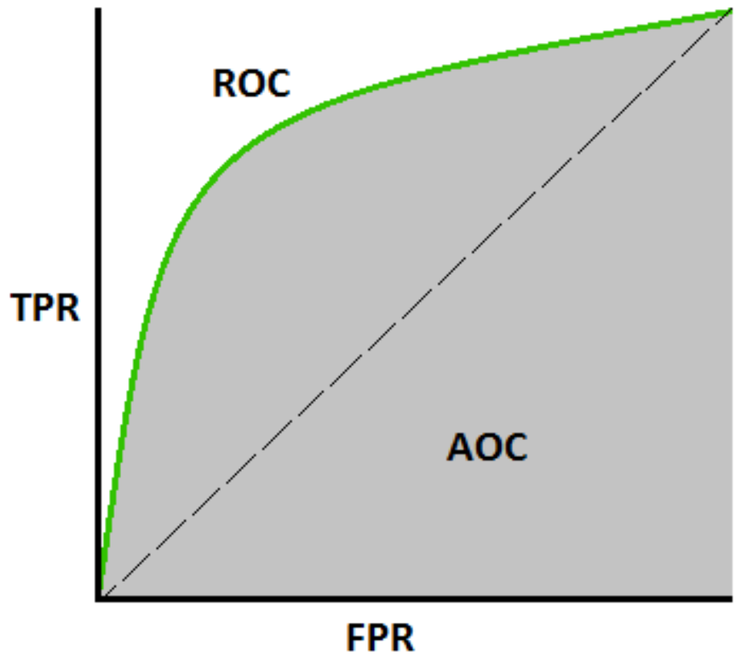
- ROC 곡선(Receiver Operation Characteristic Curve)과 이에 기반한 AUC(Area Under Curve) 스코어는 이진 분류의 예측 성능 측정에서 중요하게 사용되는 지표이다.
- 일반적으로 의학 분야에서 많이 사용되지만, 머신러닝 이진 분류 모델의 예측 성능을 판단하는 중요한 평가 지표이기도 하다.

ROC 곡선과 AUC



- TPR은 True Positive Rate이며, 이는 재현율(Recall)을 의미한다.
  - TPR = Recall = TP / (FN + TP)
- FPR은 False Positive Rate이며, 이는 Negative를 Positive로 잘못 예측한 비율이다.
  - FPR = FP / (FP + TN)

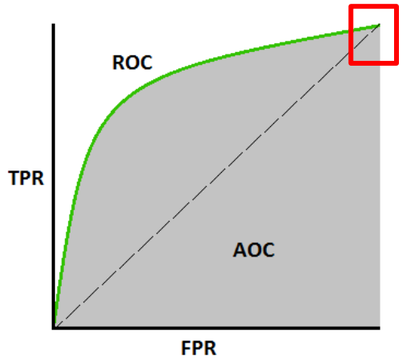
ROC 곡선과 AUC



- ROC 곡선은 FPR이 변할 때 TPR이 어떻게 변하는지를 나타내는 곡선이다.
- FPR를 X 축으로, TPR을 Y축으로 잡으면 FPR의 변화에 따른 TPR의 변화가 곡선 형태로 나타난다.
- 분류의 성능 지표로 사용되는 것은 ROC 곡선 면적에 기반한 AUC 값으로 결정된다. AUC 값은 ROC 곡선 밑의 면적을 구한 것으로서 일반적으로 1에 가까울 수록(커브의 꼭짓점이 좌상단에 위치) 좋은 수치이다.

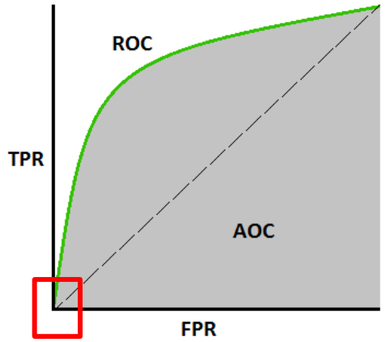
Threshold와 TPR, FPR

- Threshold를 0으로 설정하면 모델의 모든 예측이 Positive가 된다.
  - FN이 0이 되기 때문에 TPR은 1이 된다.
  - TN 또한 0이 되기 때문에 FPR도 1이 된다.

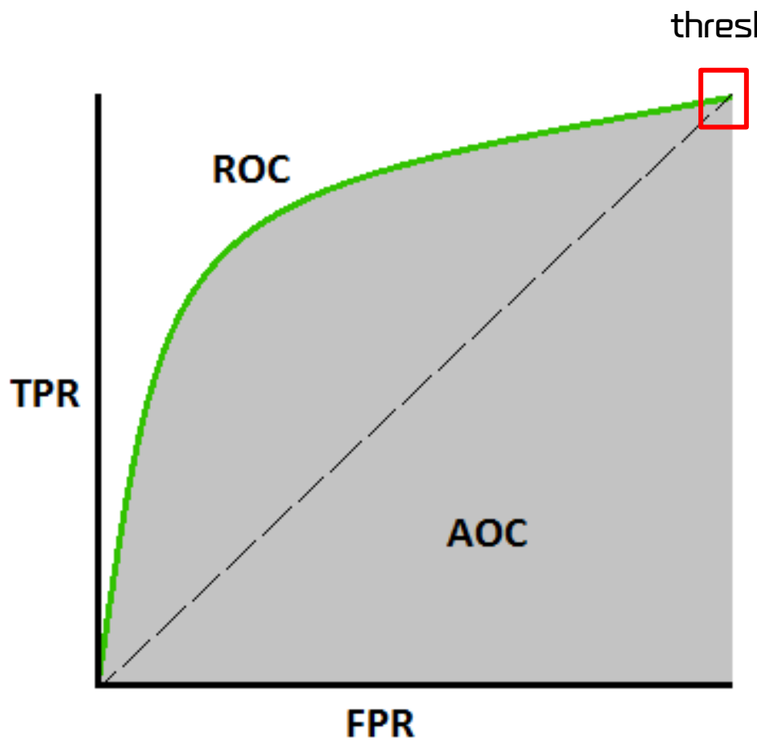


Threshold와 TPR, FPR

- Threshold를 1으로 설정하면 모델의 모든 예측이 Negative가 된다.
  - TP가 0이 되기 때문에 TPR은 0이 된다.
  - FP가 0이 되기 때문에 FPR도 0이 된다.

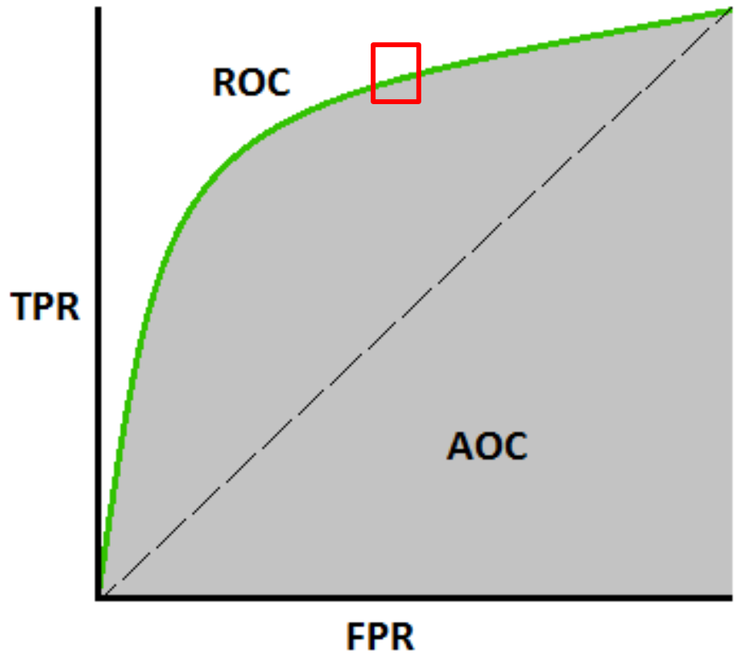


Threshold와 TPR, FPR



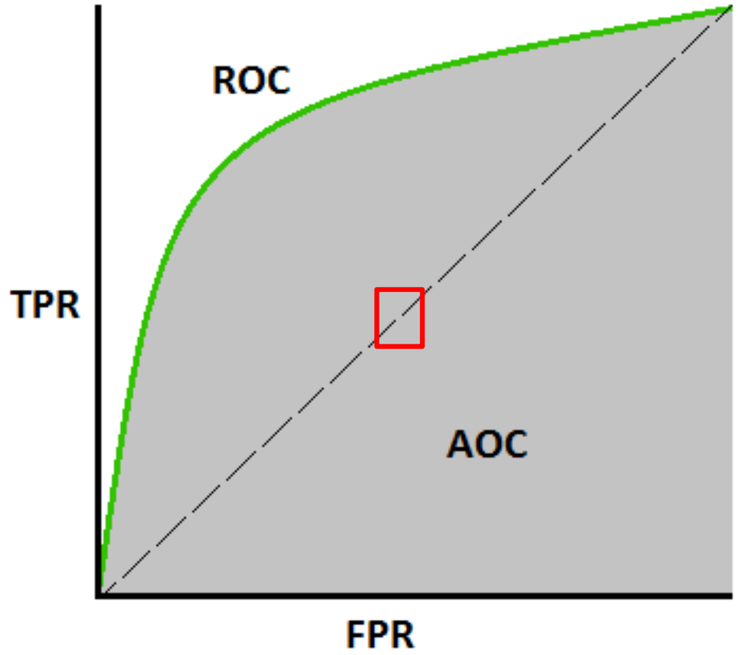
- 임계값을 0에서 부터 조금씩 높여보면

Threshold와 TPR, FPR



- 임계값을 0에서 부터 조금씩 높여보면 FPR은 비교적 빨리 떨어지고, TPR(Recall)은 버티면서 떨어져야 좋은 AOC가 구해진다.

Threshold와 TPR, FPR



- 임계값을 0에서 부터 조금씩 높여보면 FPR은 비교적 빨리 떨어지고, TPR(Recall)은 버티면서 떨어져야 좋은 AOC가 구해진다.
- 중간의 사선은 FPR과 TPR이 같은 비율로 떨어진다는 이야기이다. 즉 랜덤하게 예측된 경우를 뜻한다.