범주형자료분석팀

2팀
조장희
위재성
김지현
조수미
송지현
김민지

INDEX

혼동행렬 (Confusion Matrix)

분류 모델의 성능을 평가할 때 사용되는 지표

예측값($\hat{Y}$)이 실제 관측값(Y)을 얼마나 정확히 예측했는지 보여주는 행렬

		관측값 (Y)	
		$Y = 1$	$Y = 0$
예측값 ($\hat{Y}$)	$\hat{Y} = 1$	TP	FP
	$\hat{Y} = 0$	FN	TN

T(True)와 F(False) : 실제와 예측이 같은지 혹은 다른지

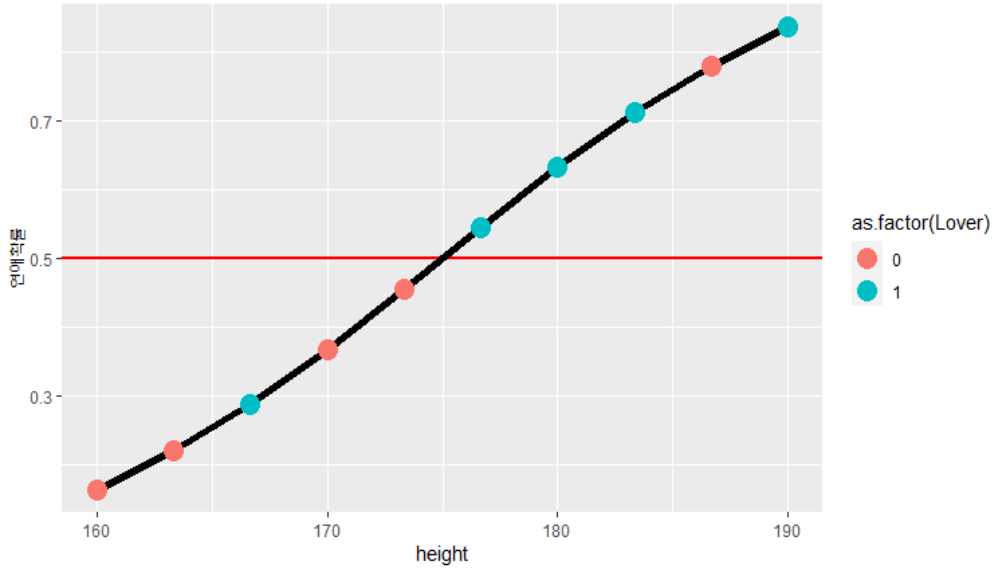
P(Positive)와 N(Negative) : 예측을 긍정 혹은 부정이라 했는지 여부

혼동행렬

- Cut-off point에 의존적

EX) 10명의 키에 따른 연애 여부 예측

<Cut-off point = 0.5>



Cut off = 0.5		실제 연애(Y)	
		$Y=1$	$Y=0$
연애 예측 ($\hat{Y}$)	$\hat{Y}=1$	4	1
	$\hat{Y}=0$	1	4

혼동행렬

분류 평가지표

상황에 따라 사용해야 하는 평가지표가 달라짐!

정확도 (Accuracy)	F1-score	MCC (매튜 상관계수)
민감도 (Sensitivity)	정밀도 (precision)	특이도 (Specificity)

1. 정확도 (Accuracy/ ACC/ 정분류율)

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

전체 경우에서 실제값과 예측값이 같은 경우의 비율

즉, 예측이 실제값과 얼마나 정확히 일치하는지를 나타내는 지표

		관측값 (Y)	
		$Y = 1$	$Y = 0$
예측값 $(\hat{Y})$	$\hat{Y} = 1$	TP	FP
	$\hat{Y} = 0$	FN	TN

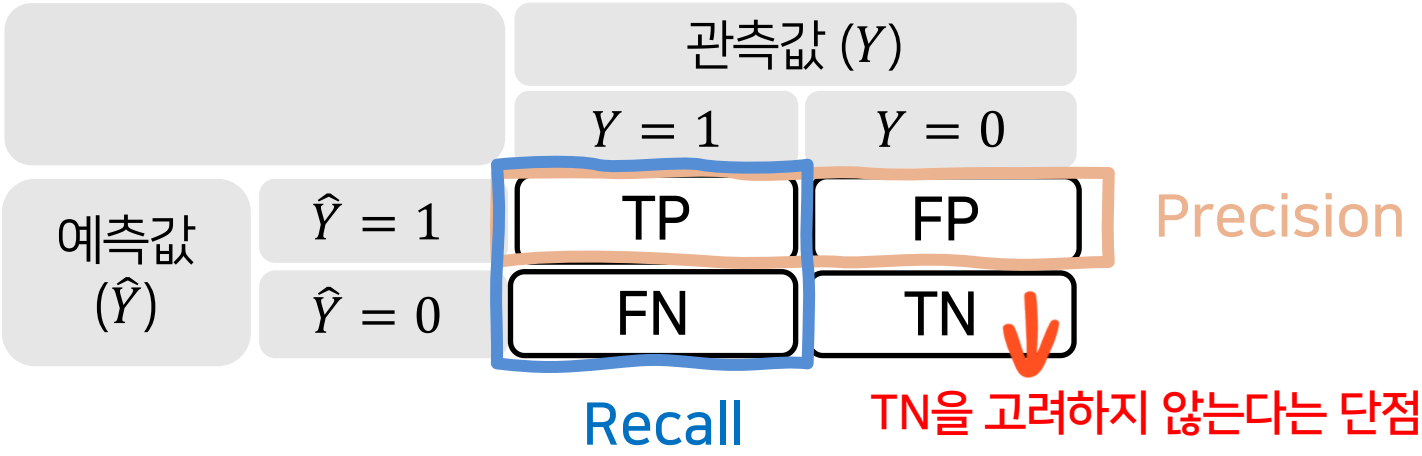
- 1에 가까울 수록 좋은 모형
- imbalanced data일 때 해당 범주에 지나치게 의존하여 문제발생

혼동행렬

5. F1-Score

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FN + FP}$$

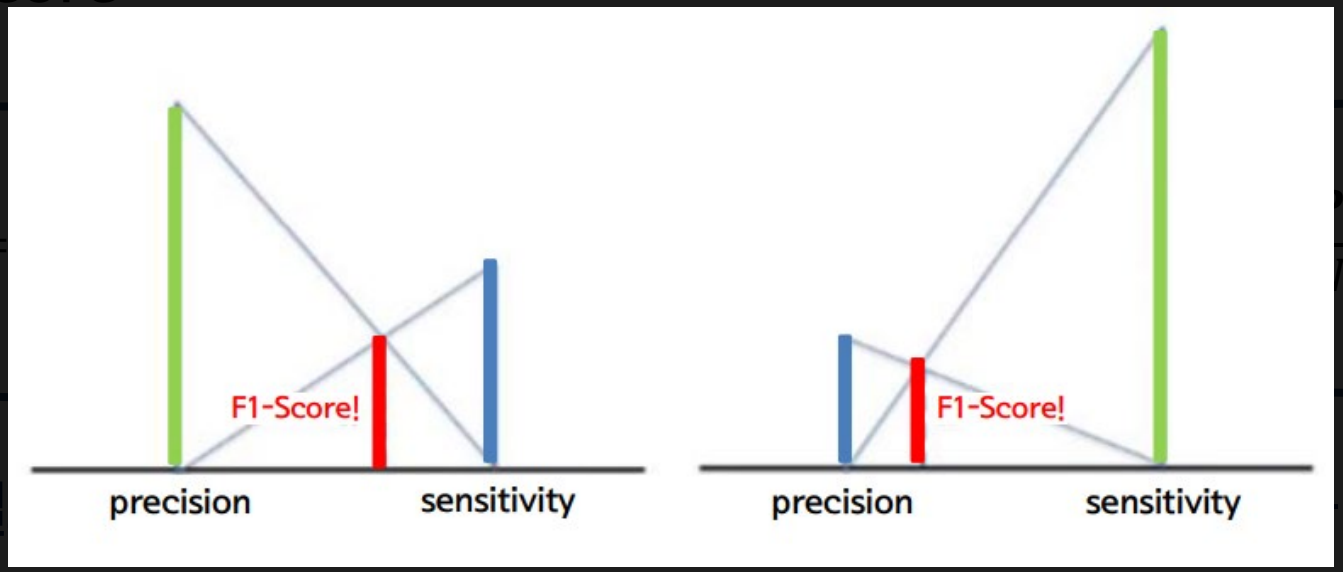
정밀도(Precision)와 민감도(Recall) 두 지표의 조화평균



혼동행렬

Q. 왜 산술평균이 아닌 조화평균을 구하는가?

5. F1-Score



정밀도와 민감도의 trade-off를 고려해서 두 지표 모두 균형 있게 반영하기 위함

조화평균은 큰 값을 갖는 쪽에 페널티를 주어 작은 값에 가까운 평균을 구함

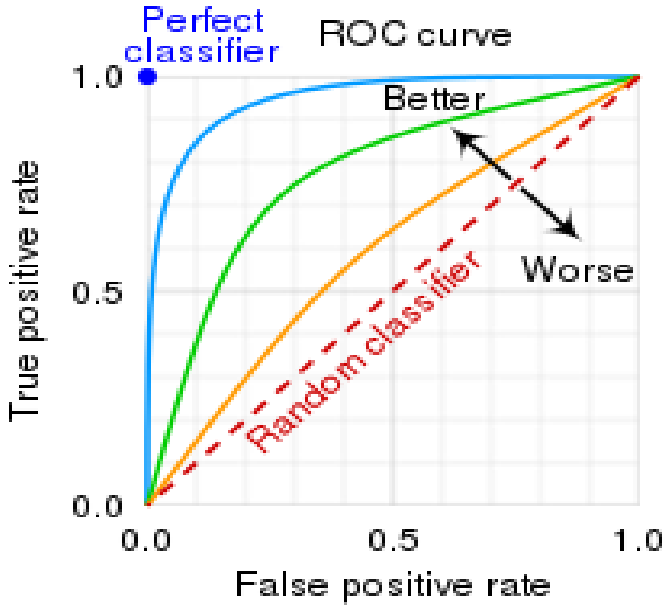
Imbalanced data에서 큰 값을 가지는 클래스에 대해 페널티를 줄 수 있음!

ROC & AUC

ROC Curve

ROC 곡선이란?

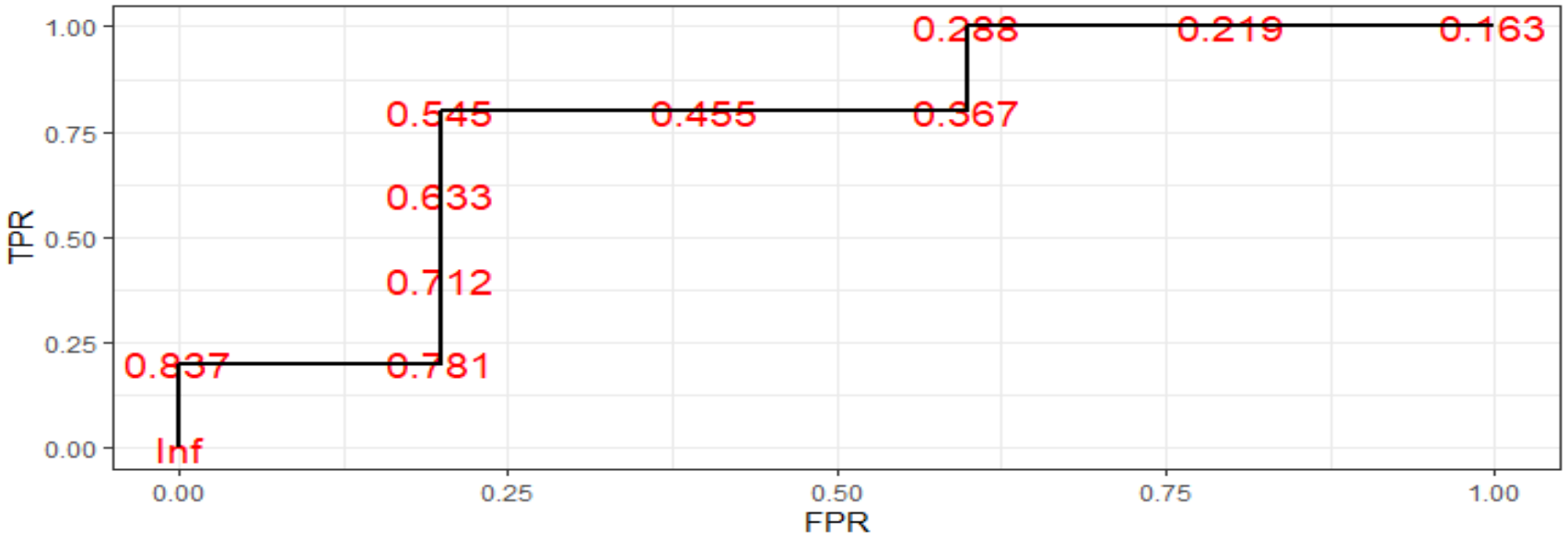
모든 cut-off point에 대해
TPR(민감도)와 FPR(1-특이도)를 나타낸 곡선



ROC 곡선으로 적합한 Cutoff point 찾기

STEP 2

각각 다른 TPR & FPR 값으로 ROC 곡선 그리기

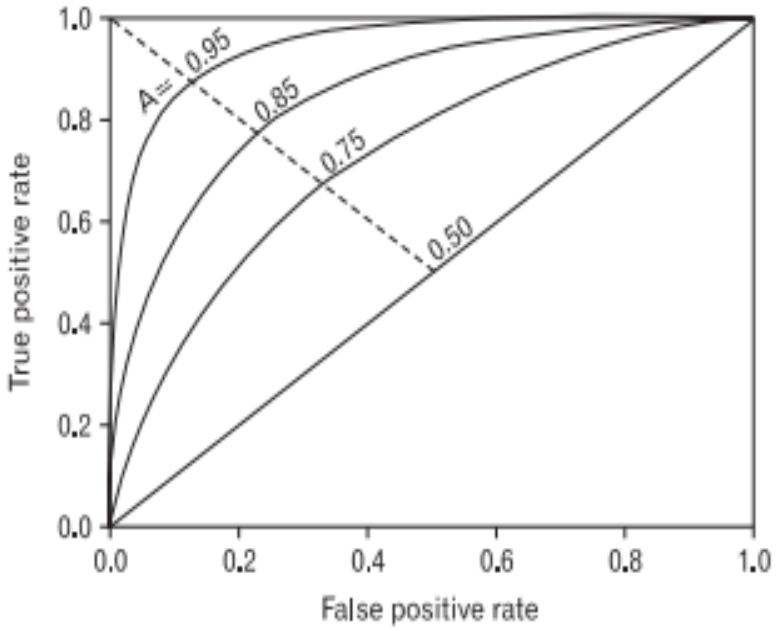


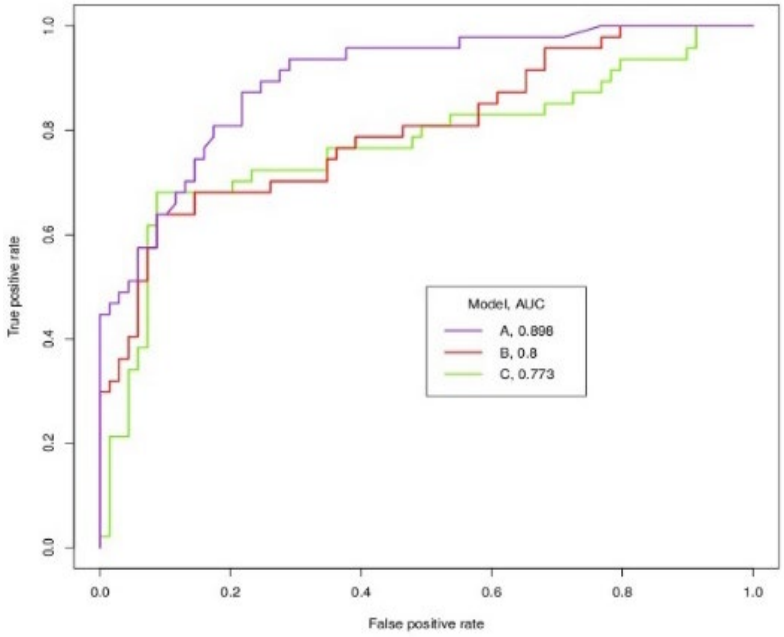
ROC & AUC

AUC란?

Area Under the Curve

: ROC curve 아래의 면적





샘플링

샘플링

오버 샘플링 (Over Sampling)

SMOTE

① 소수 클래스의 데이터 중 랜덤으로 하나를 선택한다.

② 선택한 데이터와 가장 가까운 k개의 소수 클래스 데이터를 선택한다.

③ 처음에 선택한 데이터와 무작위로 선택한 데이터 사이에 직선을 그리고, 그 직선 상에 가상의 소수 클래스 데이터를 생성한다.

Mean Encoding (Target Encoding)

각 수준에 대하여 Y의 평균을 점수로 할당하는 인코딩 방법

[Y] 양궁 획득 점수	[X] 국가	[X] 국가 (Mean Encoding)
30	대한민국	30
30	대한민국	30
30	대한민국	30
28	러시아	28.333
29	러시아	28.333
28	러시아	28.333
28	이탈리아	27.333
27	이탈리아	27.333
27	이탈리아	27.333

반응변수 Y와 설명변수 X간의 수치적 관계 반영

Ordered Target Encoding (CatBoost Encoding)

현재 행 이전 값들로 평균을 구해 할당하는 인코딩 방법

[Y] 양궁 획득 점수	[X] 국가	[X] 국가 (Mean Encoding)	[X] 국가 (CatBoost Encoding)
30	대한민국	30	28.2
28	러시아	27.5	28.2
28	이탈리아	27.333	28.2
30	대한민국	30	30
27	이탈리아	27.333	28
30	대한민국	30	30
29	러시아	27.5	28
28	러시아	27.5	28.5
27	이탈리아	27.333	27.5
25	러시아	27.5	28.333