

# 범주형자료분석팀

2팀  
박지성  
박지민  
서희나  
윤경선  
이지윤

# INDEX

---

1. 혼동행렬
2. ROC 곡선
3. 샘플링
4. 인코딩

## 혼동행렬(Confusion Matrix)이란?



분류 모델의 성능을 평가하는 지표

모델에서 훈련을 통해 **예측한 값( $\hat{Y}$ )**이 **실제값( $Y$ )**을  
얼마나 정확히 예측하였는지 보여주는 행렬

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

T(True)/ F(False) : 실제 값과 예측 값의 일치 여부

P(Positive)/ N(Negative): 모델의 긍정 혹은 부정 예측 여부

## ② 정밀도 (Precision/ PVV/ Positive Predictive Value)

$$\text{Precision} = \frac{TP}{TP + FP}$$

긍정으로 예측한 것 중 실제로도 긍정인 것의 비율

1에 가까울수록 성능이 좋다고 판단



FP가 치명적인 경우에 사용

Ex. 오염된 식수(부정)를 위생적이라고  
판단(긍정)한다는 것이 위생적인 식수를  
오염으로 판단하는 것보다 더 위험

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

## ⑥ F1-Score | 조화평균

큰 값(b)에 **패널티를 주어** 작은 값(a)에 가까운 평균을 산출

→ 조화평균을 이용하는 F1-Score를 불균형 데이터에 적용한다면?



**관측값이 많은 클래스에 패널티를 부여**  
즉, 관측값이 많은 클래스에 대한 의존성 감소



보다 정확하게 모델의 성능 파악 가능

## ⑥ F1-Score

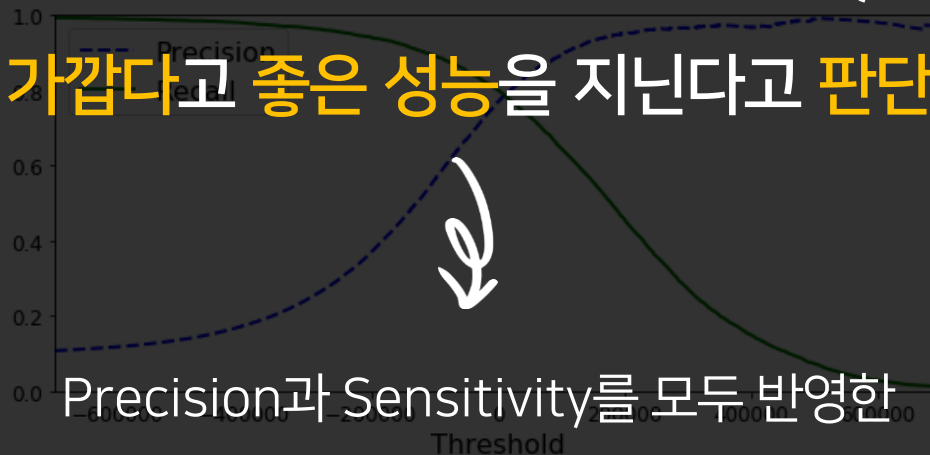


정밀도와 민감도는 상충관계(Trade-off)!



정밀도(Precision)와 민감도(Sensitivity) 중 **하나만**

**1에 가깝다고 좋은 성능을 지닌다고 판단 불가**



Precision과 Sensitivity를 모두 반영한

F1-score를 사용하여 모델의 성능을 파악해야 함

임계값 감소  $\rightarrow$  P로 예측하는 값 증가  $\rightarrow$  FN 감소, FP 증가  
 F1-Score도 1에 가까울수록 해당 모델의 성능이 우수하다고 판단  
 민감도 증가, 정밀도 감소



## Example | F1-Score와 MCC

### F1-Score가 MCC보다

안 좋은 평가지표라는 의미가 아니다!

		관측값(Y)						관측값(Y)	
		Y = 1	Y = 0					Y = 1	Y = 0
예측값 ( $\hat{Y}$ )	$\hat{Y} = 1$	92	4		예측값 ( $\hat{Y}$ )	$\hat{Y} = 1$	1	3	
	$\hat{Y} = 0$	3	1			$\hat{Y} = 0$	4	92	

IF. 분석의 목적이 클래스에 대한 균형적인 평가

혼동행렬의 모든 요소를 반영하는 MCC를 사용



F1 - Score : 왼쪽 0.96 / 오른쪽 0.22 로 상이한 결과

IF. 희귀질환처럼 연구 대상이지만 관측치가 적다면

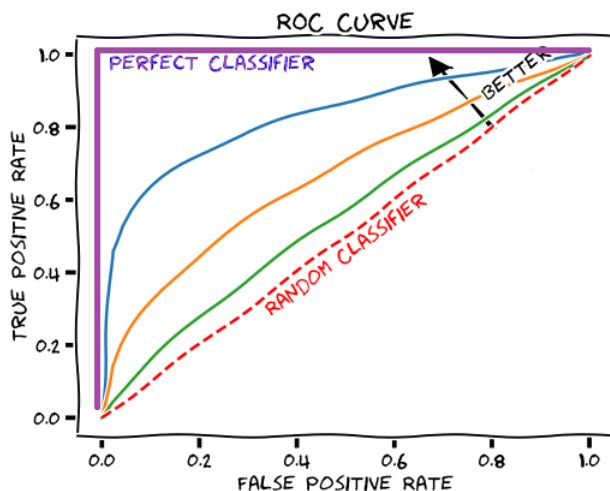
MCC : 모두 0.18로 동일한 결과!

해당 대상이 존재하는 경우를 Positive로 두고 F1-Score를 사용

## ROC 곡선 (Receiver Operating Characteristic Curve)이란?

모든 cut-off point에 대하여 **재현율(x)**를 **1-특이도(y)**의 함수로 나타낸 곡선

모든 cut-off point에 대하여 confusion matrix를 구하고 이를 통해 구현한  
'재현율'과 '1-특이도' 값을 2차원 상의 점으로 찍어 연결한 상태

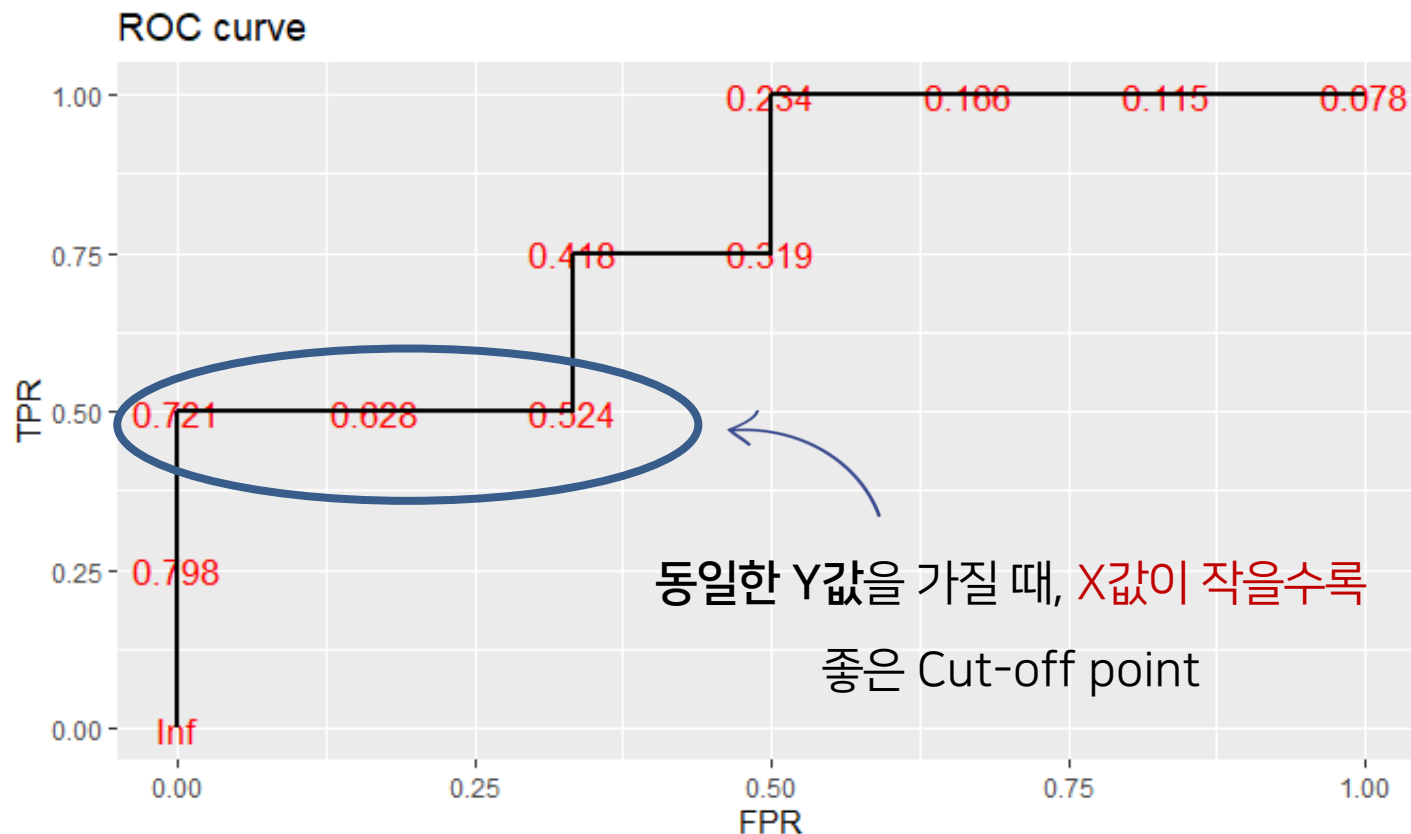




## ROC 곡선 그리기



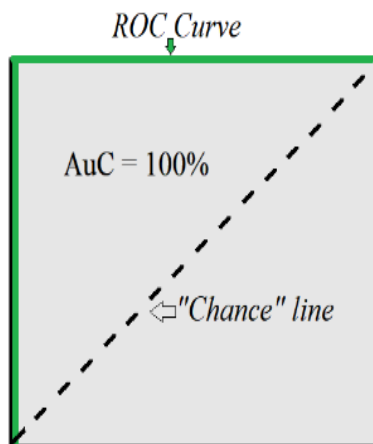
X축은 FPR Y축은 TPR 값으로 ROC 곡선 그리기



## 2

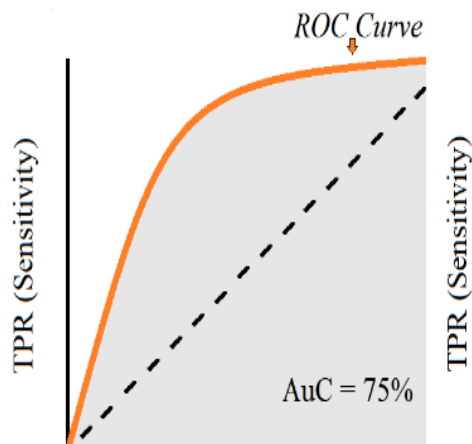
## ROC 곡선

## AUC (Area Under the Curve)



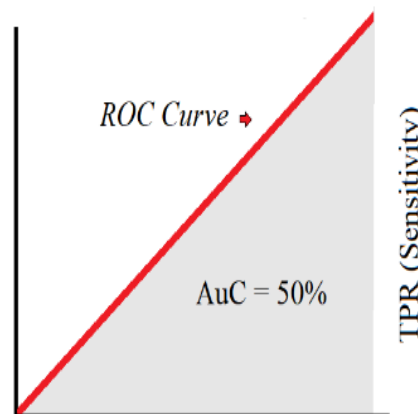
✓ AUC = 1

완벽하게 예측  
(과적합 의심)



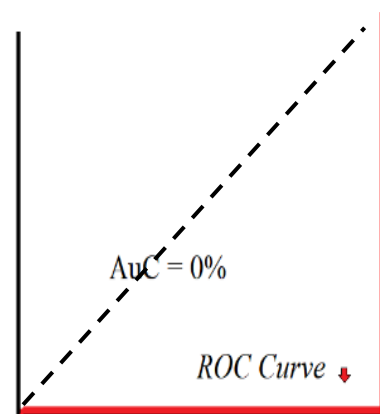
✓ AUC = 0.75

80%이상  
성능 좋음



✓ AUC = 0.5

50% 예측  
무작위 예측



✓ AUC = 0

100% 반대로  
예측

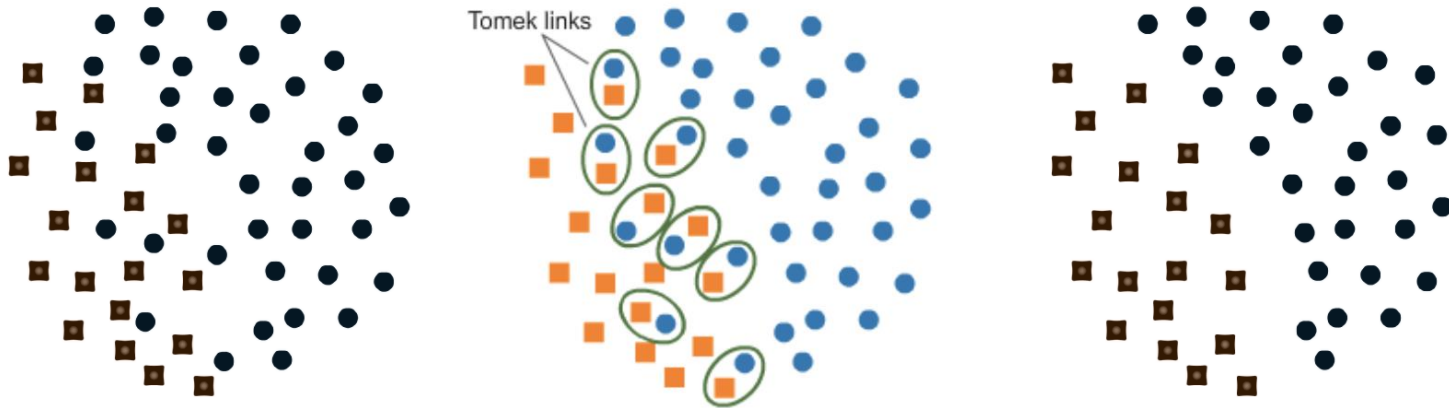
## 언더 샘플링의 종류

### ② Tomek Links Method



앞서 선택한 두 점간 거리가 주위에 있는 다른 클래스의 데이터들과 연결한 거리보다 짧다면, 두 점 간 **Tomek Link**가 있다고 말함

(초록색 동그라미로 강조)



## 오버 샘플링의 종류

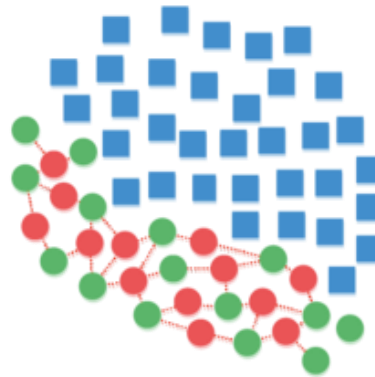
### ② SMOTE (Synthetic Minority Over-sampling Method)



① 에서 선택한 하나의 데이터에서 ② 에서 선택한 K개의 데이터 사이에 직선을 그리고, 그 직선 상 가상의 소수 클래스 데이터를 생성



Original Dataset



Generating Samples




Resampled Dataset

## Ordinal Encoding

- ✓ 순서형 자료가 주어졌을 때 사용
- ✓ 순서가 있는 각 수준에 대응하는 점수를 할당하는 방식
- ✓ 각 수준에 할당된 점수들 간 순서와 연관성이 존재

만족도	점수
매우 나쁨	1
나쁨	2
보통	3
좋음	4
매우 좋음	5



1부터 시작하여  
순서가 있도록  
수치를 할당

## Target Encoding

One-Hot  
Encoding

Label  
Encoding

Ordinal  
Encoding

각 수준을 구분할 뿐, 값 자체에 특별한 의미 X



Target Encoding



- ✓ 각 수준을 구분
- ✓ 설명변수 X와 반응변수 Y 간 **수치적 관계 반영**하여 인코딩

## Ordered Target Encoding (CatBoost Encoding)

현재 행 이전의 값들 중 같은 수준에 속한 행들의 평균을 구해 이를 점수로 할당

Mean Encoding과 비교해 보았을 때

각 수준에 더 다양한 점수가

할당되었음을 알 수 있음

		Mean Encoding	Ordered Target Encoding
	경영	172	169.5
	통계	166	169.5
	경제	171.66	169.5
	통계	166	174
	경영	172	168
	통계	166	165
	경제	171.66	165
	경제	171.66	172.5
	경영	172	174
	통계	166	164.33

과적합의 정도나  
가능성이 낮아짐