범주형자료분석팀

2팀
임지훈
안은선
강세현
심현구
하희나

INDEX

마지막이니까
다들 힘내자구~

1

혼동행렬

혼동행렬

혼동행렬(Confusion Matrix)

모델의 **성능**을 **평가**하기 위한 지표

예측한 값($\hat{Y}$)과 **실제 값**(Y)이 **얼마나 정확히 일치**하는지 보여주는 **행렬**

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

1 혼동행렬

혼동행렬의 4가지 지표

TP(True Positive)

긍정($\hat{Y} = 1$)으로 **예측**하였고 **실제 관측값**도 **긍정**($Y = 1$)인 경우

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

혼동행렬의 4가지 지표

TN(True Negative)

부정($\hat{Y} = 0$)으로 **예측**하였고 **실제 관측값**도 **부정**($Y = 0$)인 경우

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

1 혼동행렬

혼동행렬의 4가지 지표

FP(False Positive)

긍정($\hat{Y} = 1$)으로 **예측**하였으나 **실제 관측값**은 **부정**($Y = 0$)인 경우

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

1종 오류

혼동행렬의 4가지 지표

FN(False Negative)

부정($\hat{Y} = 0$)으로 **예측**하였으나 **실제 관측값**은 **긍정**($Y = 1$)인 경우

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

2종 오류

1 혼동행렬

혼동행렬의 분류평가지표

정확도 (Accuracy/ACC/정분류율)

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

전체 혼동행렬 값에서 **예측값**과 **실제값**이 **일치**하는 **비율**

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

정확도
Accuracy

- ✓ **1에 가까울수록**
  성능이 좋은 모델
- ✓ **불균형 자료**에서는 경향성을
  띄어 **큰 설명력을 갖지 못함**

1 혼동행렬

혼동행렬의 분류평가지표

정밀도(Precision/PPV/Positive)

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

긍정($\hat{Y} = 1$)으로 **예측**한 값들 중 **실제 관측값**도 **긍정**($Y = 1$)인 **비율**

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

정밀도
Precision

- ✓ **1에 가까울수록**
  성능이 좋은 모델
- ✓ **FP**가 더 중요할 때 주로 사용

1 혼동행렬

혼동행렬의 분류평가지표

민감도(Sensitivity/TPR/True Positive Rate)

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

실제로 긍정($Y = 1$)인 관측값을 긍정($\hat{Y} = 1$)으로 올바르게 예측한 비율

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

민감도
Sensitivity

- ✓ 재현율(Recall)
- ✓ 1에 가까울수록 성능이 좋은 모델
- ✓ FN이 더 중요할 때 주로 사용
- ✓ ROC 곡선의 Y축

1 혼동행렬

혼동행렬의 분류평가지표

특이도(Specificity/TNR/True Negative Rate)

$$Specificity = \frac{TN}{TN + FP}$$

실제로 **부정**($Y = 0$)인 관측값을 **부정**($\hat{Y} = 0$)으로 올바르게 **예측**한 **비율**

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

특이도
Specificity

1에 가까울수록

성능이 좋은 모델

혼동행렬의 분류평가지표

FPR/False Positive Rate

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

실제로 **부정**($Y = 0$)인 관측값을 **긍정**($\hat{Y} = 1$)으로 **잘못 예측**한 **비율**

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

FPR
False Positive Rate

- ✓ **0에 가까울수록** 성능이 좋은 모델
- ✓ ROC 곡선의 X축

1 혼동행렬

혼동행렬의 분류평가지표

F1-Score

정밀도(Precision)와 **재현율**(Recall)의 **조화평균**

$$F1\ Score$$

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

1에 가까울수록 **모델의 성능이 좋다**고 판단

1 혼동행렬

혼동행렬의 분류평가지표

F1-Score

정밀도(Precision)와 재현율(Recall)의 조화평균

조화평균을 사용하는 이유

불균형 데이터에서 정확도(Accuracy)의 한계를 보완할 수 있고

정밀도와 재현율 간의 상충관계를 반영하기 위함!

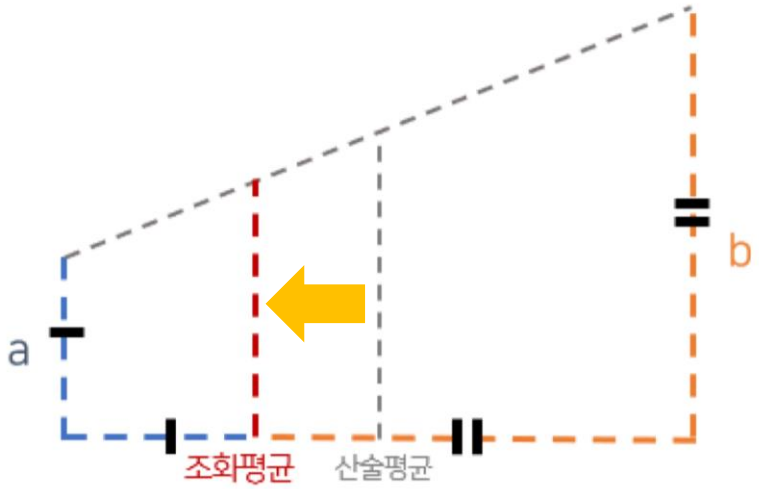
$$F1\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FN + FP}$$

1에 가까울수록 모델의 성능이 좋다고 판단

혼동행렬의 분류평가지표

F1-score에서 조화평균을 사용하는 이유

- ✓ **불균형** 데이터에서 **정확도**(Accuracy) **한계 보완**
- ✓ 정밀도와 재현율 간의 상충관계를 반영하기 위함



더 **큰 값**에 **패널티**를 주어 **작은 값에 가까운 평균**을 구함

1 혼동행렬

혼동행렬의 분류평가지표

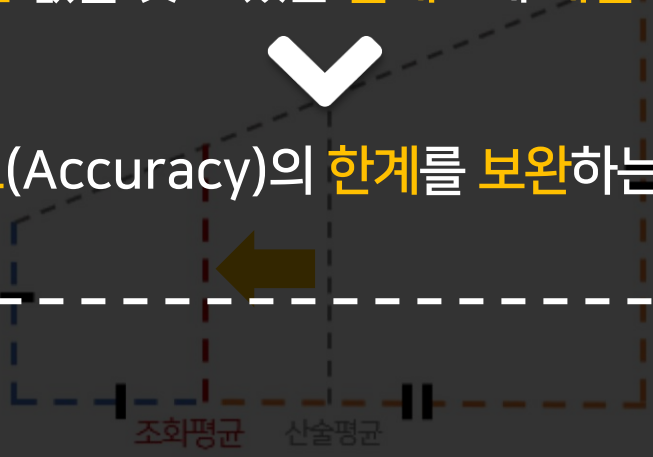
F1-score에서 조화평균을 사용하는 이유

✓ 불균형 데이터에서 정확도(Accuracy) 한계 보완

불균형 데이터(Imbalanced Data)가 주어졌을 때

더 많은 값을 갖고 있는 클래스에 패널티 부과

정확도(Accuracy)의 한계를 보완하는 역할!



더 큰 값에 패널티를 주어 작은 값에 가까운 평균을 구함

혼동행렬의 분류평가지표

F1-score에서 조화평균을 사용하는 이유

- ✓ 불균형 데이터에서 정확도(Accuracy) 한계 보완
- ✓ **정밀도**와 **재현율** 간의 **상충관계**를 반영하기 위함

정밀도나 재현율 중 한 지표만을 이용하여 성능을 평가하지 않고

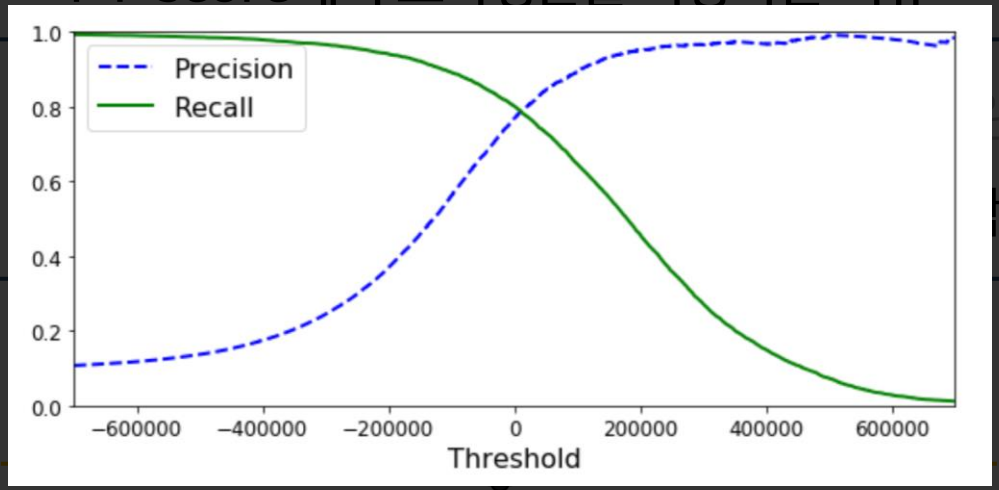
정밀도와 **재현율**을 모두 고려하여 **더 좋은 모델 성능 지표**를 찾을 수 있음

1 혼동행렬

혼동행렬의 분류평가지표

정밀도와 재현율의 상충관계(Trade-off)

F1-score에서 조화평균을 사용하는 이유



정밀도와 재현율은 동시에 높은 값을 가질 수 없음

정밀도나 재현율 중 한 지표만을 이용하여 성능을 평가하지 않고

정밀도와 재현율을 모두 고려하여 더 좋은 모델 성능 지표를 찾을 수 있음

정밀도가 높아지면 재현율이 낮아짐

재현율이 높아지면 정밀도가 낮아짐

혼동행렬의 분류평가지표

F1-score 는 TN(True Negative) 수치를 반영하지 못한다는 한계를 가짐

		관측값(Y)	
		$Y=1$	$Y=0$
예측값 ($\hat{Y}$)	$\hat{Y}=1$	26	27
	$\hat{Y}=0$	24	22

		관측값(Y)	
		$Y=1$	$Y=0$
예측값 ($\hat{Y}$)	$\hat{Y}=1$	26	27
	$\hat{Y}=0$	24	72

$$F1\ Score = \frac{2 * 26}{2 * 26 + 27 + 24} = 0.505$$

같은 값을 가짐

1 혼동행렬

혼동행렬의 분류평가지표

F1-score 는 TN(True Negative) 수치를 반영하지 못한다는 한계를 가짐

TN의 수치를 반영하지 못하는 F1-score의 한계를
MCC 지표를 사용해 보완 가능!

		관측값(Y)	
		$Y=1$	$Y=0$
예측값 ($\hat{Y}$)	$\hat{Y}=1$	26	27
	$\hat{Y}=0$	24	22

		관측값(Y)	
		$Y=1$	$Y=0$
예측값 ($\hat{Y}$)	$\hat{Y}=1$	26	27
	$\hat{Y}=0$	24	72

$$F1\ Score = \frac{2 * 26}{2 * 26 + 27 + 24} = 0.505$$

같은 값을 가짐

MCC

MCC (Matthews Correlation Coefficient)

혼동행렬의 **모든 구성요소**를 활용하여 계산

상관계수 값이기 때문에 **-1**과 **1** 사이의 값을 가짐

$$\mathrm{MCC}=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

1에 가까울 수록 **완전 예측**

0에 가까울 수록 **랜덤 예측**

-1에 가까울 수록 **역예측**

MCC

MCC (Matthews Correlation Coefficient)

혼동행렬의 **모든 구성요소**를 활용하여 계산

상관계수 값이기 때문에 **-1**과 **1** 사이의 값을 가짐

$$\mathrm{MCC}=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

1에 가까울 수록 **완전 예측**

0에 가까울 수록 **랜덤 예측**

-1에 가까울 수록 **역예측**

MCC

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

TP 관측치가 매우 큰 불균형 데이터

$$\text{F1-score} = \frac{2\times92}{2\times92+4+3} = 0.96$$

$$\text{MCC} = \frac{(92\times1)-(4\times3)}{\sqrt{(92+4)(92+3)(1+4)(1+3)}} = 0.18$$

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

TN 관측치가 매우 큰 불균형 데이터

$$\text{F1-score} = \frac{2\times1}{2\times1+4+3} = 0.22$$

$$\text{MCC} = \frac{(1\times92)-(3\times4)}{\sqrt{(1+3)(1+4)(92+3)(92+4)}} = 0.18$$

MCC

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

TP 관측치가 매우 큰 불균형 데이터

$$\text{F1-score} = \frac{2\times92}{2\times92+4+3} = \mathbf{0.96}$$

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

TN 관측치가 매우 큰 불균형 데이터

$$\text{F1-score} = \frac{2\times1}{2\times1+3+4} = \mathbf{0.22}$$

✓ F1-score는 큰 차이가 나타남

MCC

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

TP 관측치가 매우 큰 불균형 데이터

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

TN 관측치가 매우 큰 불균형 데이터

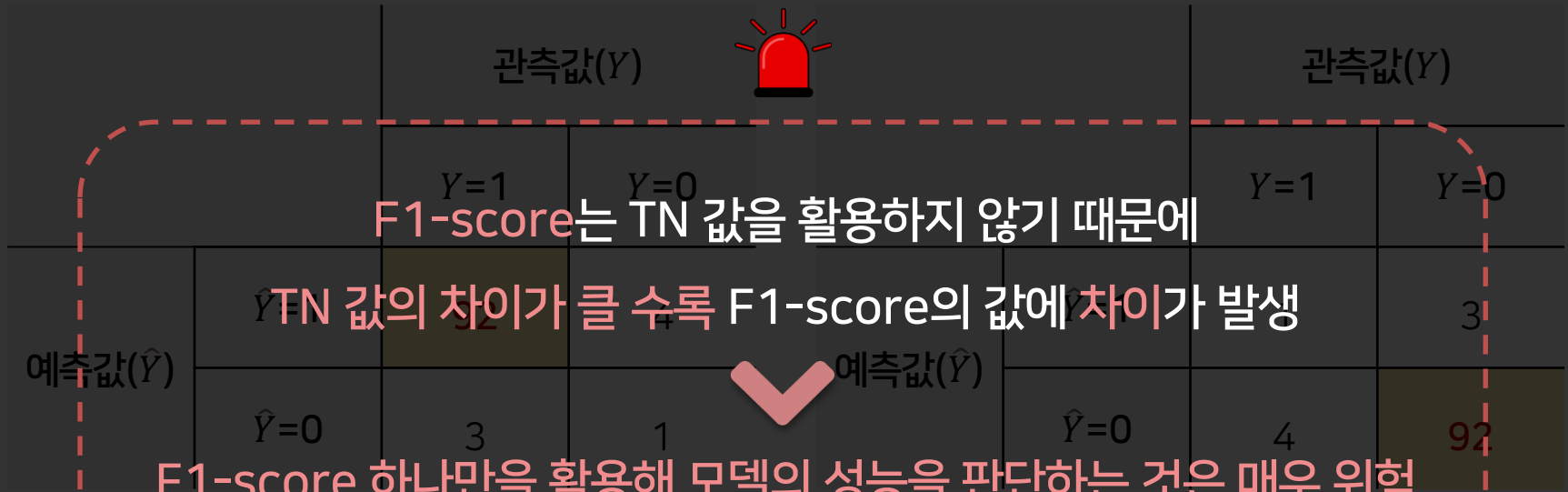
✓ MCC는 0.18로 같음

$$\text{MCC} = \frac{(92\times1)-(4\times3)}{\sqrt{(92+4)(92+3)(1+4)(1+3)}} = \mathbf{0.18}$$

$$\text{MCC} = \frac{(92\times1)-(3\times4)}{\sqrt{(1+3)(1+4)(92+3)(92+4)}} = \mathbf{0.18}$$

1 혼동행렬

MCC



F1-score는 TN 값을 활용하지 않기 때문에

TN 값의 차이가 클 수록 F1-score의 값에 차이가 발생

F1-score 하나만을 활용해 모델의 성능을 판단하는 것은 매우 위험

TP 관측치가 매우 큰 불균형 데이터 — TN 관측치가 매우 큰 불균형 데이터

✓ MCC는 0.18로 같음

$$\text{MCC} = \frac{(92\times1)-(4\times3)}{\sqrt{(92+4)(92+3)(1+4)(1+3)}} = \mathbf{0.18}$$

$$\text{MCC} = \frac{(92\times1)-(3\times4)}{\sqrt{(1+3)(1+4)(92+3)(92+4)}} = \mathbf{0.18}$$

1 혼동행렬

MCC

MCC가 반드시 F1-score보다 좋은 지표일까?

MCC

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

모든 클래스에 대한 균형적인 평가

TP 관측치가 매우 큰 불균형 데이터

$$MCC = \frac{(92\times1)-(4\times3)}{\sqrt{(92+4)(92+3)(1+4)(1+3)}} = 0.18$$

F1-score

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

비대칭 데이터에 대한 평가

관측치가 적은 항목을 Positive로 설정해 TP에 중점

TN 관측치가 매우 큰 불균형 데이터

$$MCC = \frac{(92\times1)-(3\times4)}{\sqrt{(1+3)(1+4)(92+3)(92+4)}} = 0.18$$

✓ MCC는 0.18로 같음

목적에 맞게 지표를 선택하는 것이 중요!

혼동행렬의 단점

혼동행렬의 단점

- ✓ 정보의 손실
- ✓ 임의적인 cut-off point 설정

임의의 cut-off point에 따라
이항변수에 맞게 **범주화**

연속인 값(x)을 **이항**의 값($\hat{Y}$)으로
변환시키는 과정에서
숫자가 갖는 정보를 잃게 됨

어떤 값을 기준으로
분류하는지에 따라
혼동행렬이 크게 달라질 수 있음

분석의 객관성이 떨어짐

혼동행렬의 단점

혼동행렬의 단점

- ✓ 정보의 손실
- ✓ 임의적인 cut-off point 설정

임의의 cut-off point에 따라
이항변수에 맞게 **범주화**

연속인 값(x)을 이항의 값($\hat{Y}$)으로
변환시키는 과정에서
숫자가 갖는 정보를 잃게 됨

어떤 값을 기준으로
분류하는지에 따라
혼동행렬이 크게 달라질 수 있음

분석의 객관성이 떨어짐

1 혼동행렬

혼동행렬의 단점

혼동행렬의 단점

혼동행렬은 특정 cut-off point를 기준으로

관측값과 예측값을 분류하여 나열한 도표이기 때문에

cut-off point가 변화함에 따라 검정력이 어떻게 변화하는지 파악하기 어려움

ROC 곡선을 이용해 한계 보완!

어떤 값을 기준으로

분류하는지에 따라

혼동행렬이 크게 달라질 수 있음

분석의 객관성이 떨어짐

2

ROC 곡선

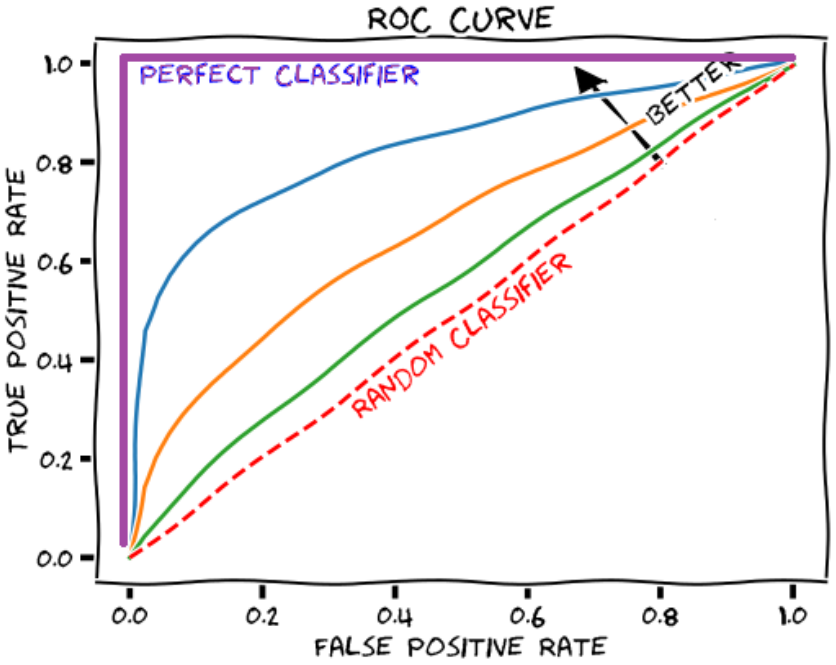
2 ROC 곡선

ROC 곡선

ROC 곡선

0~1 범위의 **모든 cut-off point**에 대해

재현율과 **1-특이도**의 함수로 나타낸 **곡선 그래프**



ROC 곡선

Receiver Operating Characteristic Curve

모든 cut-off point에 대해

혼동행렬을 구하고

각 혼동행렬의 **재현율(TPR)**과

1-특이도(FPR)를

2차원 상의 점으로 찍어 연결

2 ROC 곡선

ROC 곡선

ROC 곡선

0~1 범위의 **모든 cut-off point**에 대해

재현율과 **1-특이도**의 함수로 나타낸 **곡선 그래프**

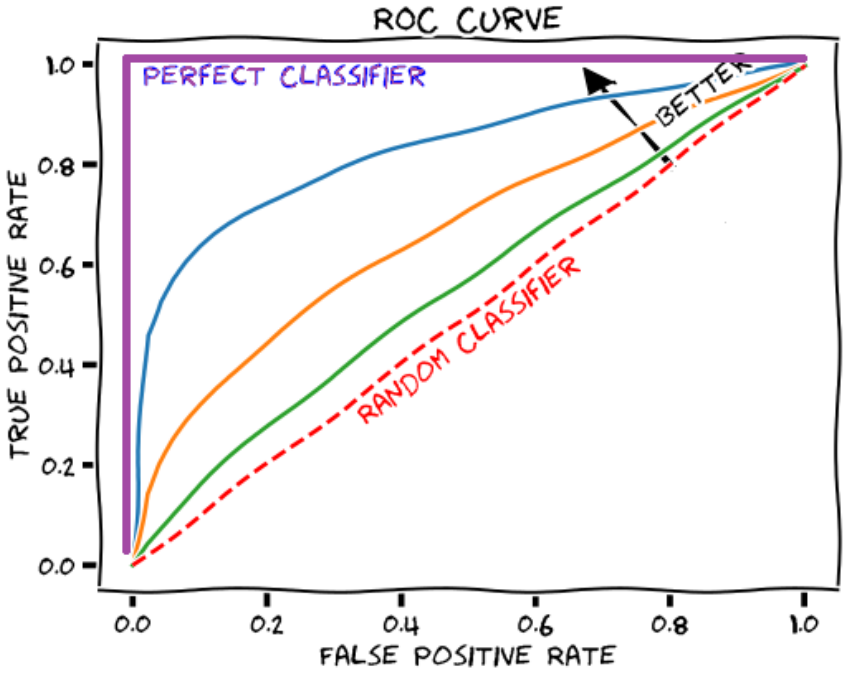
ROC 곡선의 장점

- ✓ 혼동 행렬보다 더 많은 정보를 가짐
- ✓ 주어진 모형에서 가장 적합한 cut-off point를 찾을 수 있음

2 ROC 곡선

ROC 곡선의 형태

재현율 (TPR)

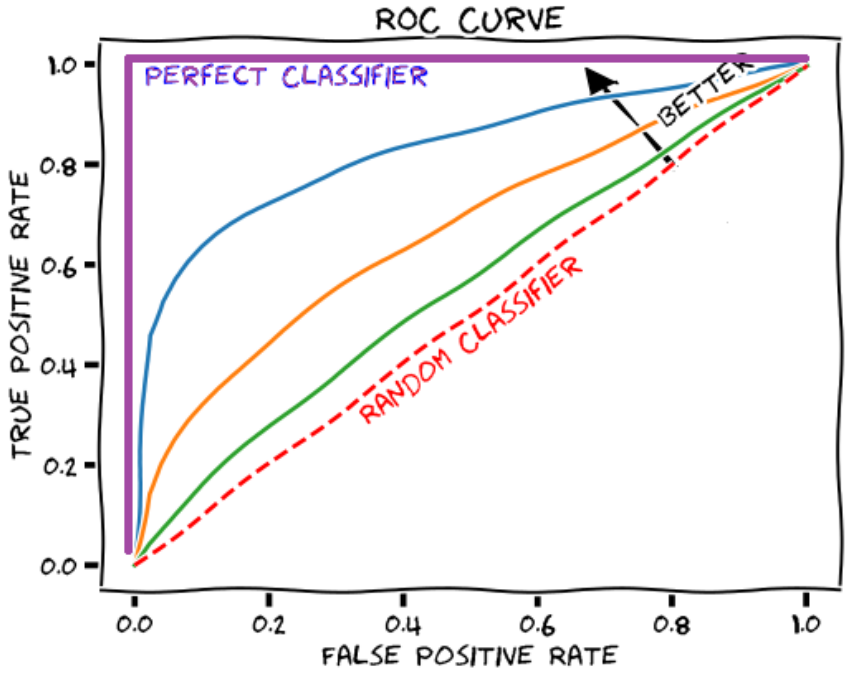


1-특이도 (FPR)

(0,0)과 (1,1)을 잇는 **우상향** 그래프의 형태

2 ROC 곡선

ROC 곡선의 형태



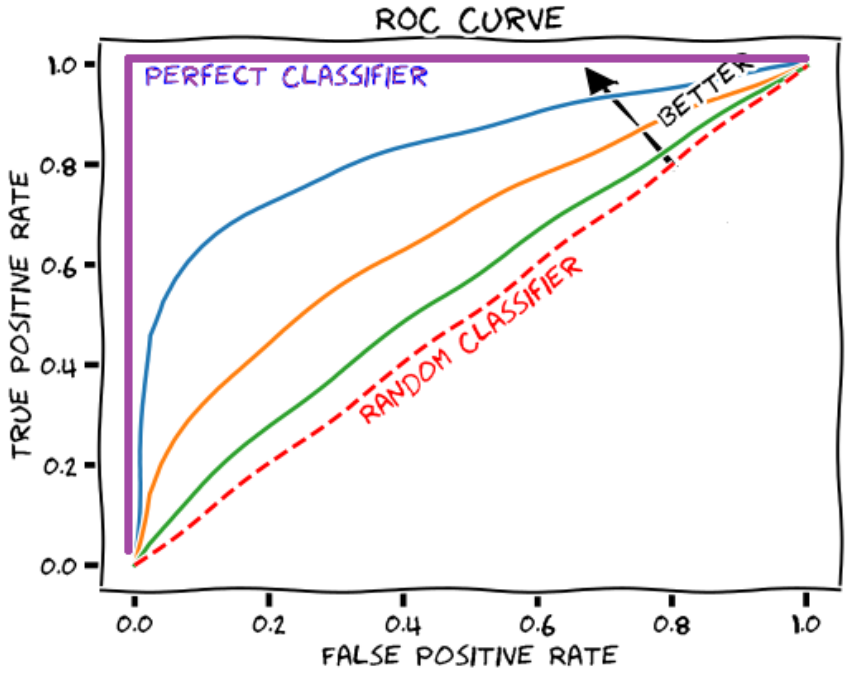
Cut-off point가 0에 가까운 경우

Cut-off point가 0에 가까워짐 → 대부분 $\hat{Y} = 1$ 로 예측 → TP와 FP 증가

→ TN과 FN 감소 → TPR과 FPR 모두 1에 가까워짐

2 ROC 곡선

ROC 곡선의 형태



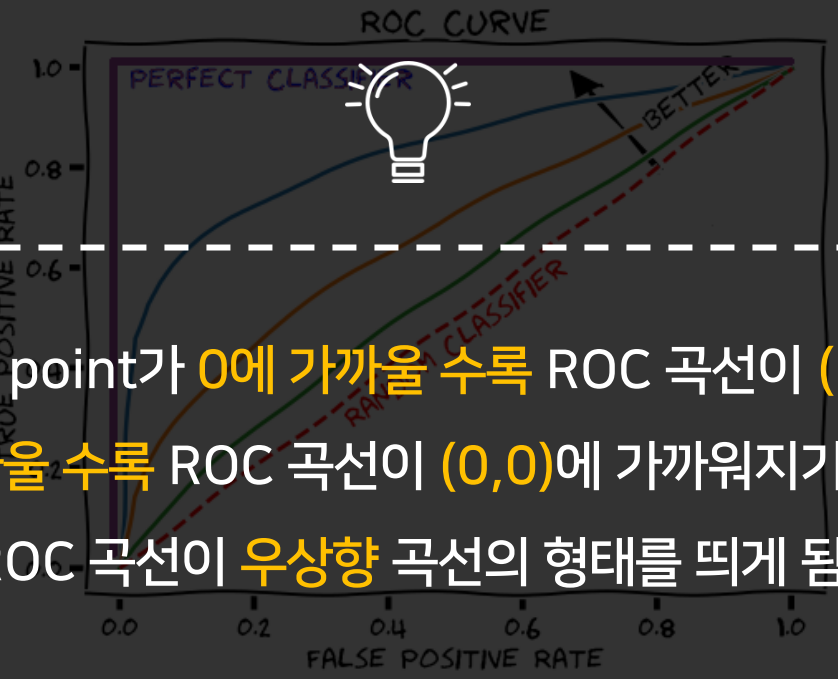
Cut-off point가 1에 가까운 경우

Cut-off point가 1에 가까워짐 → 대부분 $\hat{Y} = 0$ 로 예측 → TP와 FP 감소

→ TN과 FN 증가 → TPR과 FPR 모두 0에 가까워짐

2 ROC 곡선

ROC 곡선의 형태



cut-off point가 0에 가까울 수록 ROC 곡선이 (1,1)에,
1에 가까울 수록 ROC 곡선이 (0,0)에 가까워지기 때문에
ROC 곡선이 우상향 곡선의 형태를 띄게 됨

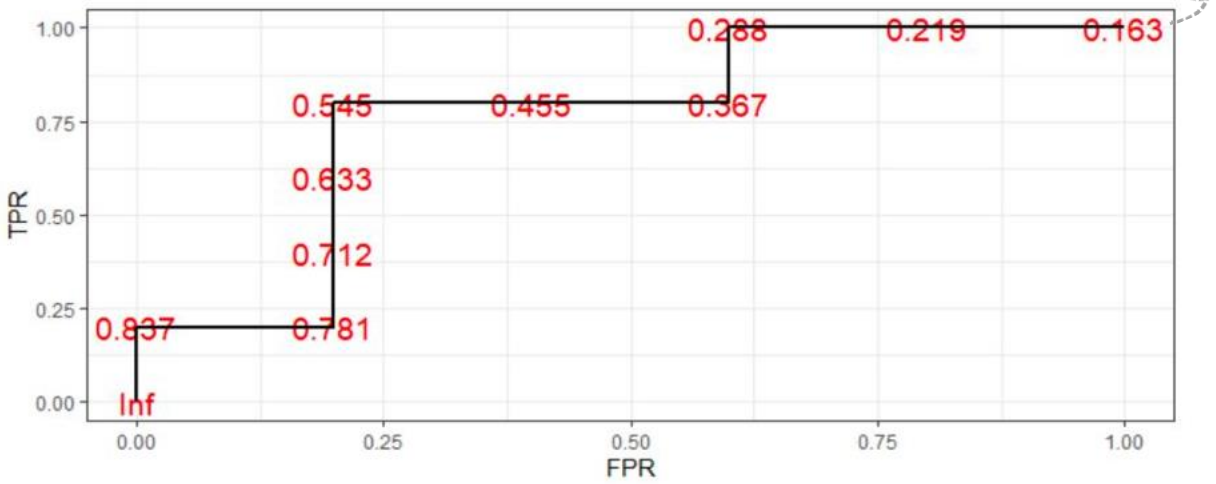
Cut-off point가 1에 가까운 경우

Cut-off point가 1에 가까워짐 → 대부분 $\hat{Y} = 0$ 로 예측 → TP와 FP 감소
→ TN과 FN 증가 → TPR과 FPR 모두 0에 가까워짐

2 ROC 곡선

ROC 곡선의 해석

빨간색 숫자들은
cut-off point를 의미



모든 cut-off point에 대한 혼동행렬을 구해 TPR, FPR을 구함

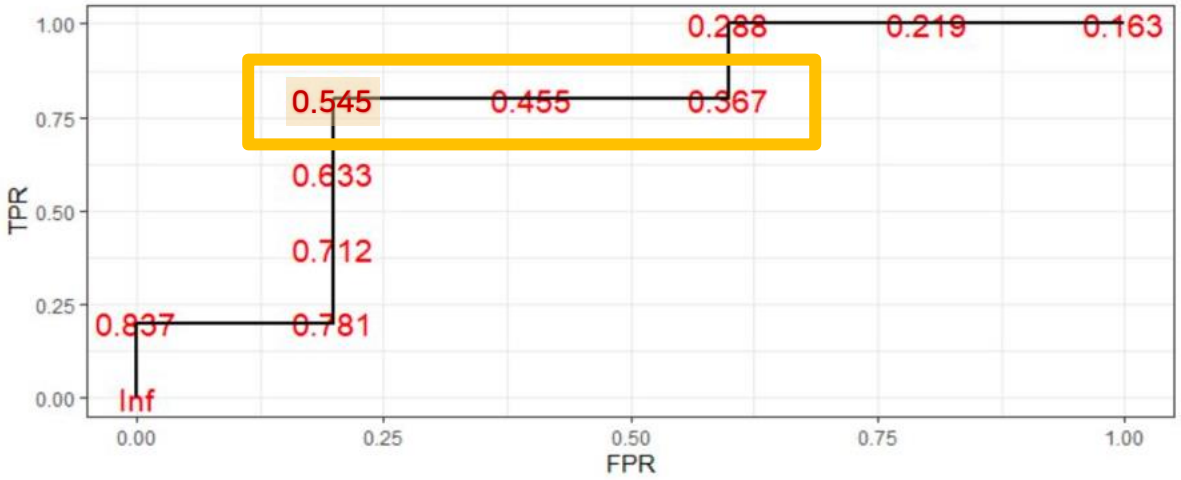
ROC 곡선의 해석

✓ TPR(Y값)이 같을 때 FPR(X값)이 더 작을수록 좋은 cut-off point

✓ FPR(X값)이 같을 때 TPR(Y값)이 더 클 수록 좋은 cut-off point

2 ROC 곡선

ROC 곡선의 해석



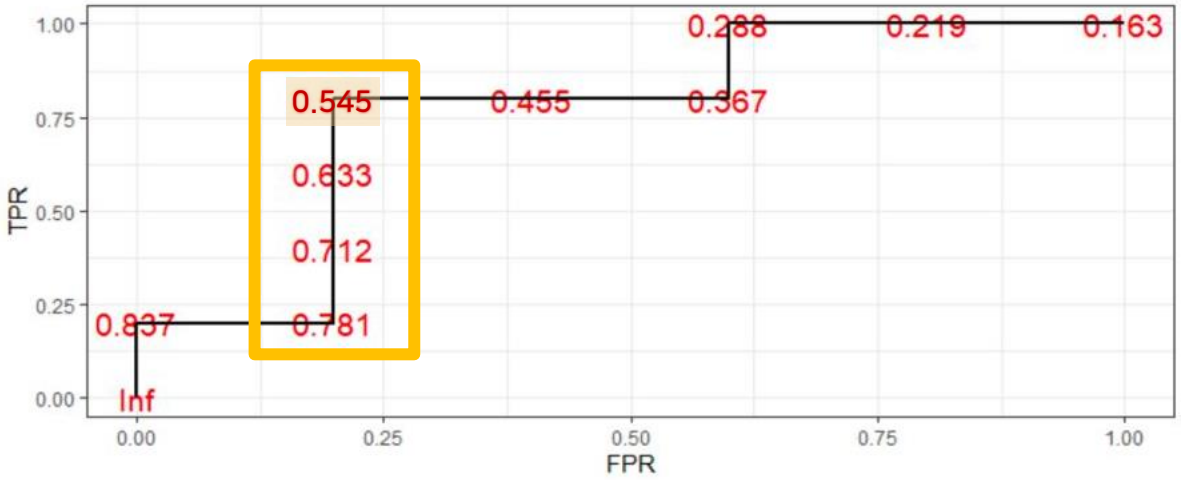
모든 cut-off point에 대한 혼동행렬을 구해 TPR, FPR을 구함

ROC 곡선의 해석

- ✓ TPR(Y값)이 같을 때 FPR(X값)이 더 작을수록 좋은 cut-off point
- ✓ FPR(X값)이 같을 때 TPR(Y값)이 더 클 수록 좋은 cut-off point

2 ROC 곡선

ROC 곡선의 해석



모든 cut-off point에 대한 혼동행렬을 구해 TPR, FPR을 구함

ROC 곡선의 해석

✓ TPR(Y값)이 같을 때 FPR(X값)이 더 작을수록 좋은 cut-off point

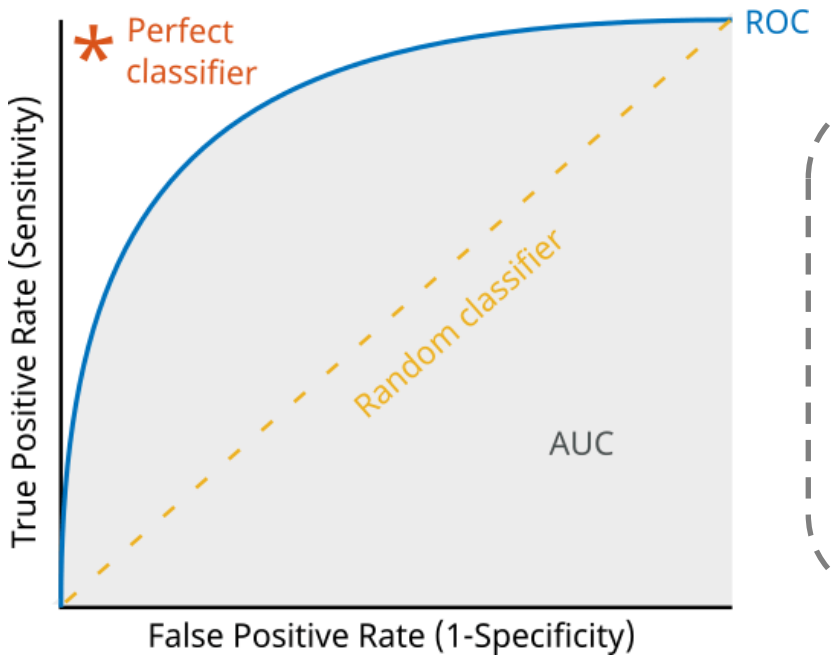
✓ FPR(X값)이 같을 때 TPR(Y값)이 더 클 수록 좋은 cut-off point

2 ROC 곡선

AUC

AUC

ROC 곡선 아래의 면적을 의미



AUC

Area Under Curve

- ✓ 0과 1사이의 값을 가짐
- ✓ 모든 cut-off point를 고려하기 때문에 **특정 cut-off point와 상관없이** **모델 성능 측정 가능**

2 ROC 곡선

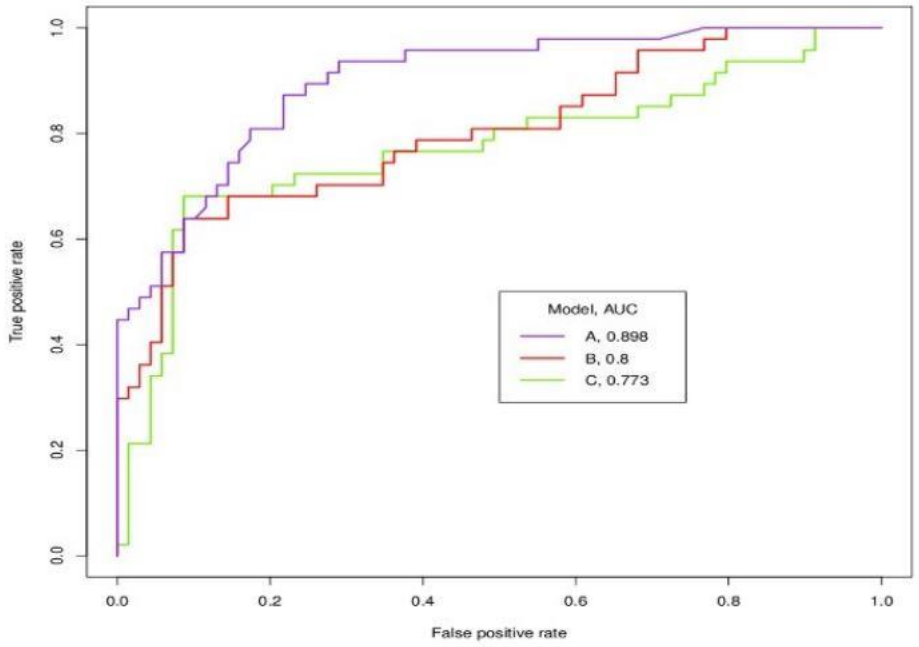
AUC

AUC

AUC

ROC 곡선 아래의 면적을 의미

3개 모델의 ROC 곡선



AUC를 계산했을 때

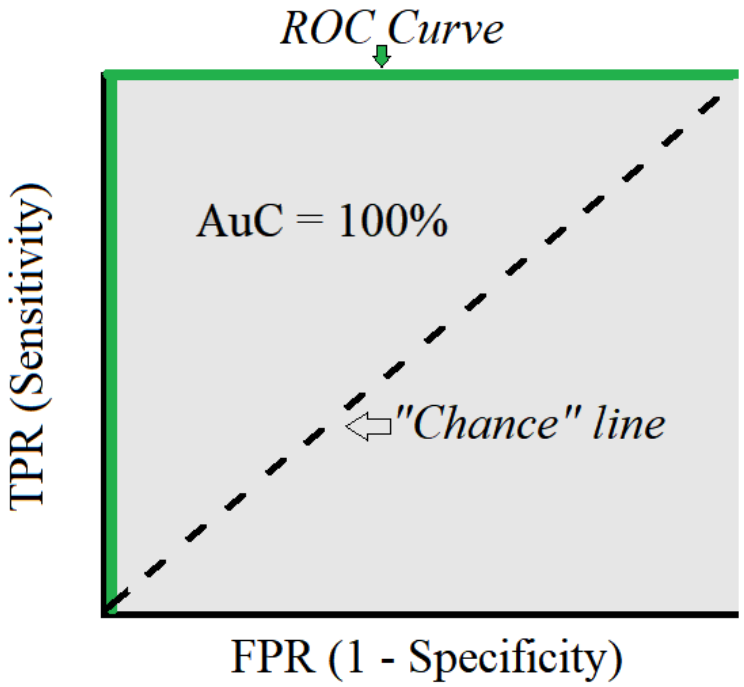
A 모델의 AUC가 0.898로 **가장 높음**

A 모델의 성능이

가장 좋음

2 ROC 곡선

AUC의 해석



AUC = 1인 경우

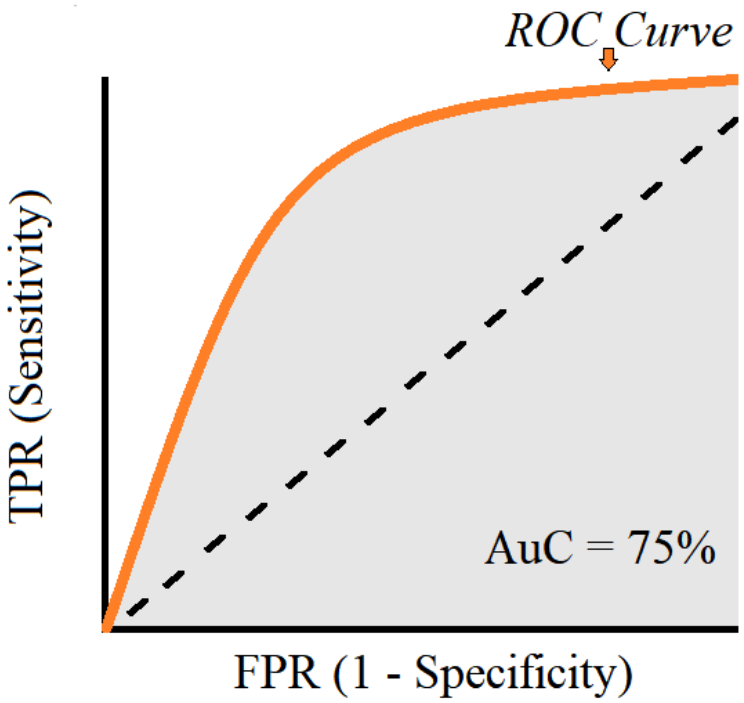
모델이 100% 수준으로

완벽하게 **관측치**를 **예측**

과적합이 발생한 것은 아닌지 **확인 필요**

2 ROC 곡선

AUC의 해석



AUC = 0.75인 경우

모델이 75% 수준으로

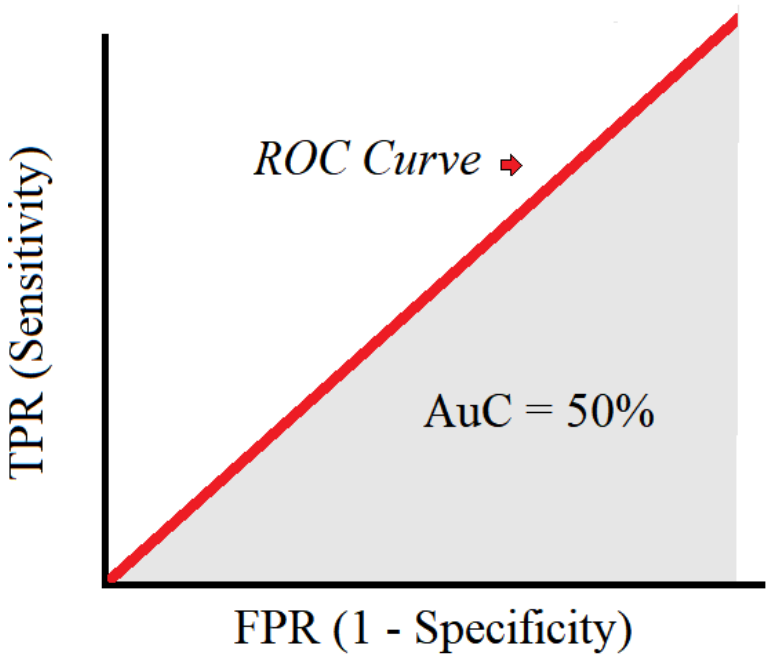
관측치를 예측

일반적으로 모델의 AUC가 **0.8 이상**일 때

성능이 좋다고 판단

2 ROC 곡선

AUC의 해석



AUC = 0.5인 경우

모델이 절반의

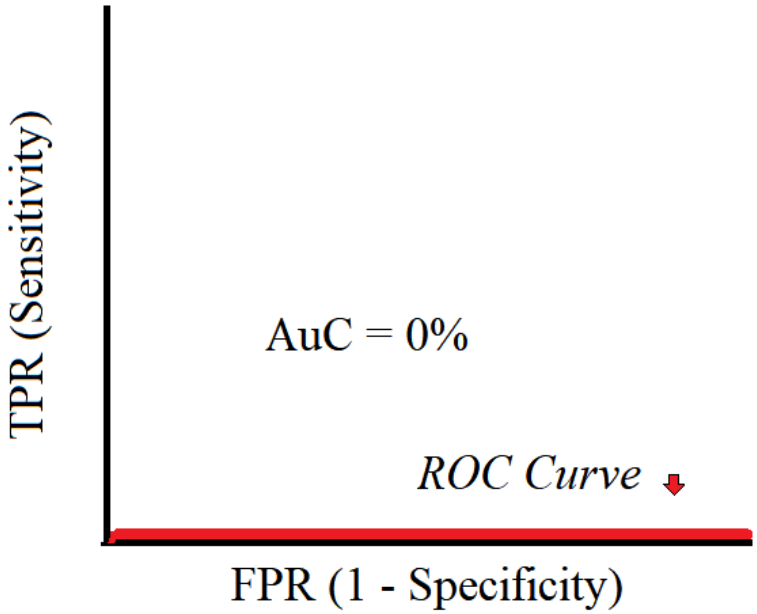
관측치를 예측

무작위로 **예측**했다는 의미

보통 AUC는 **0.5 이상**의 값일 때 **정상**

2 ROC 곡선

AUC의 해석



AUC = 0인 경우

모델이 관측치를

100% 반대로 예측

Y=1과 Y=0을

거꾸로 예측

3

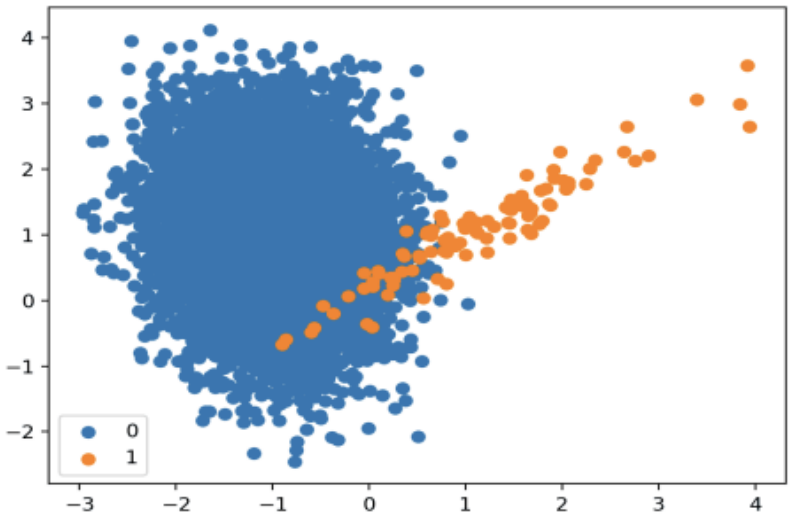
샘플링

3 샘플링

클래스 불균형

클래스 불균형

관측치의 개수가 크게 **차이** 나는 경우



0인 클래스가 1인 클래스에 비해 **많이 관측**됨

ex) RH-형과 RH+형

3 샘플링

클래스 불균형 해결의 필요성

		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	60	5
	$\hat{Y}$=0	40	5
	클래스 별 정확도	0.6	0.5

전체 정확도 : 0.59

		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	50	4
	$\hat{Y}$=0	50	6
	클래스 별 정확도	0.5	0.6

전체 정확도 : 0.509

Y=1의 **관측치 개수가 많아** 전체 정확도는 왼쪽 도표가 높음

클래스 불균형 해결의 필요성

		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	60	5
	$\hat{Y}$=0	40	5
	클래스 별 정확도	0.6	0.5

		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	50	4
	$\hat{Y}$=0	50	6
	클래스 별 정확도	0.5	0.6

정확도 지표는 **클래스 불균형**에 **민감**

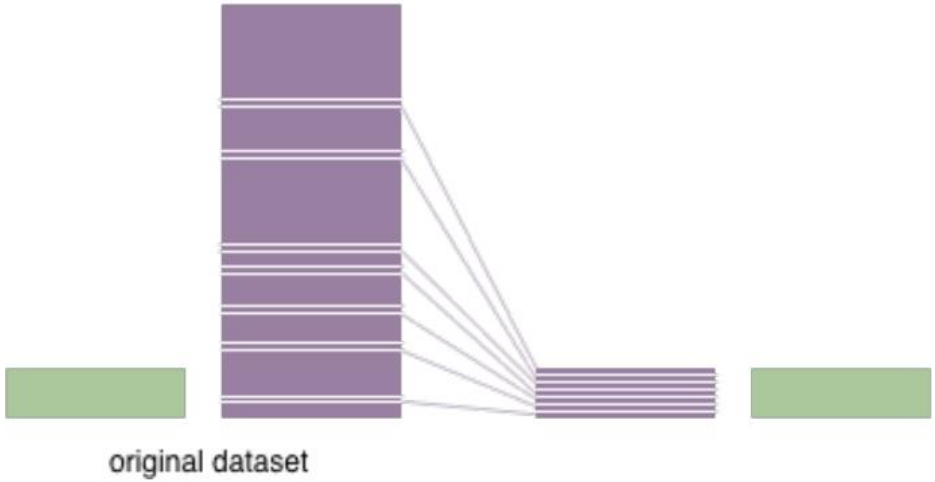
정확한 해석을 위해 **클래스 불균형 해결이 필요**

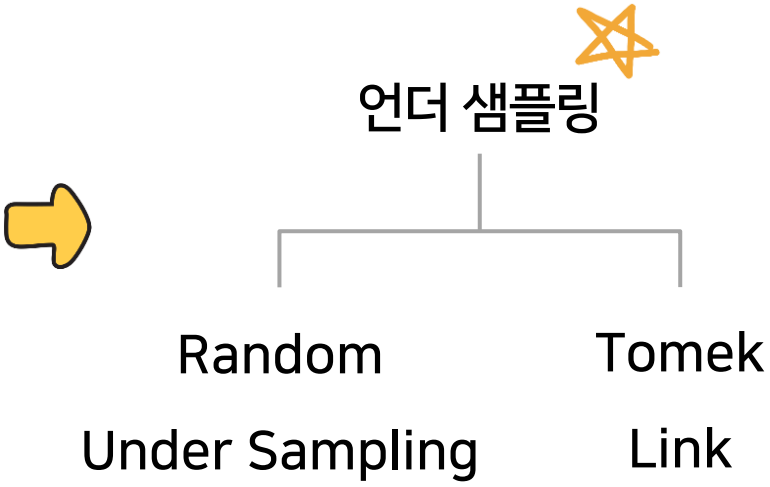
3 샘플링

언더 샘플링

언더 샘플링 (Under Sampling)

다수의 클래스를 소수의 클래스에 맞추어 관측치를 감소시키는 방법



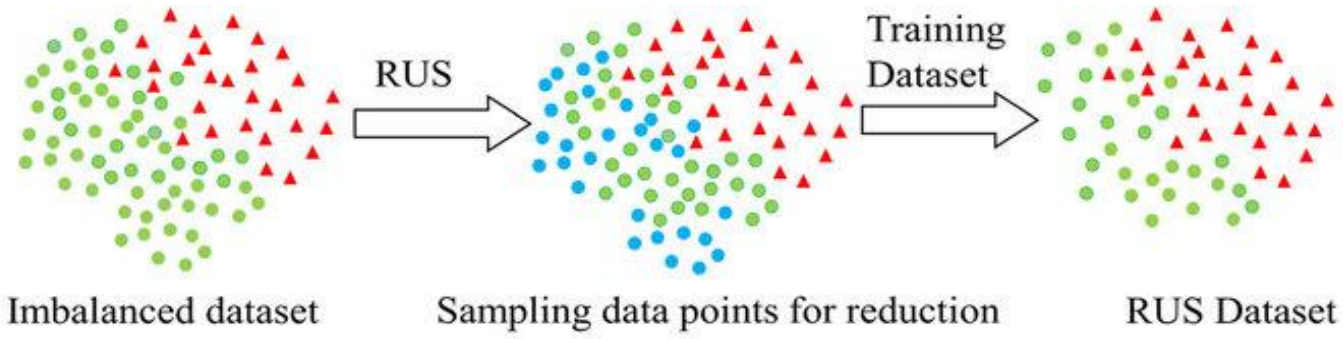


3 샘플링

랜덤 언더 샘플링

랜덤 언더 샘플링 (Random Under Sampling)

임의적으로 **다수의 클래스**의 데이터를 **제거**하여 관측치의 수를 줄이는 방법

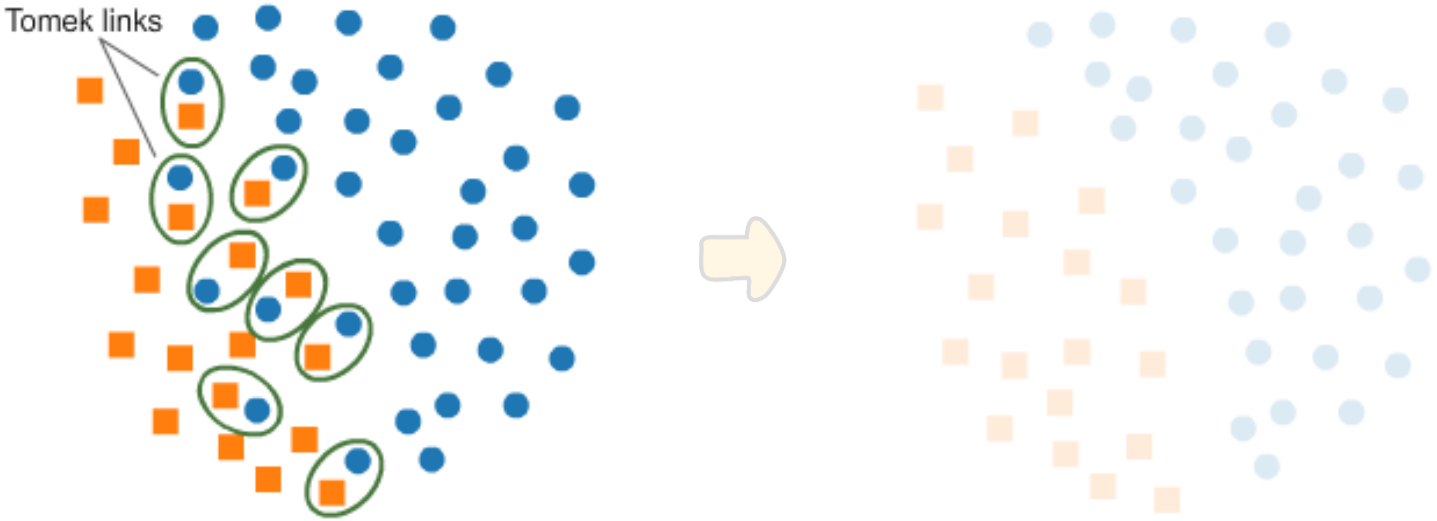


임의적으로 제거한 데이터의 정보가 누락되고

기존 데이터에 대한 **대표성**을 **띄지 못하면 부정확한 결과**를 야기할 수 있음

3 샘플링

Tomek Links



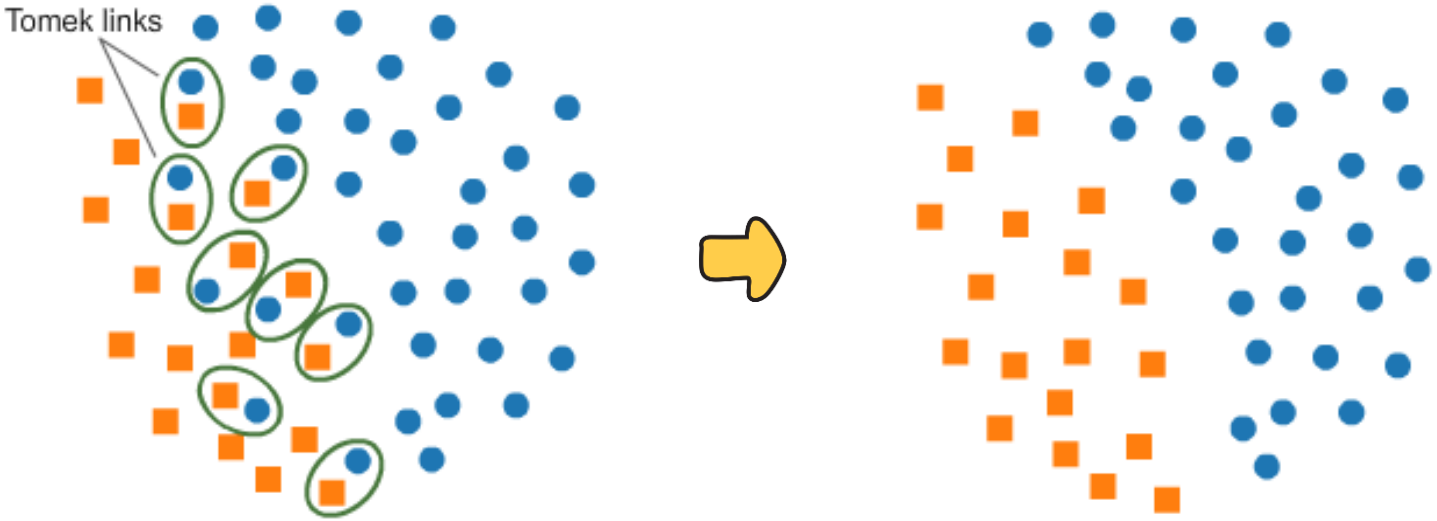
Tomek Links 과정

임의로 서로 다른 클래스의 데이터를 **두 점**을 **선택**하여 이를 **연결**

묶인 데이터 쌍에서 다수의 클래스에 속한 데이터들을 삭제

3 샘플링

Tomek Links

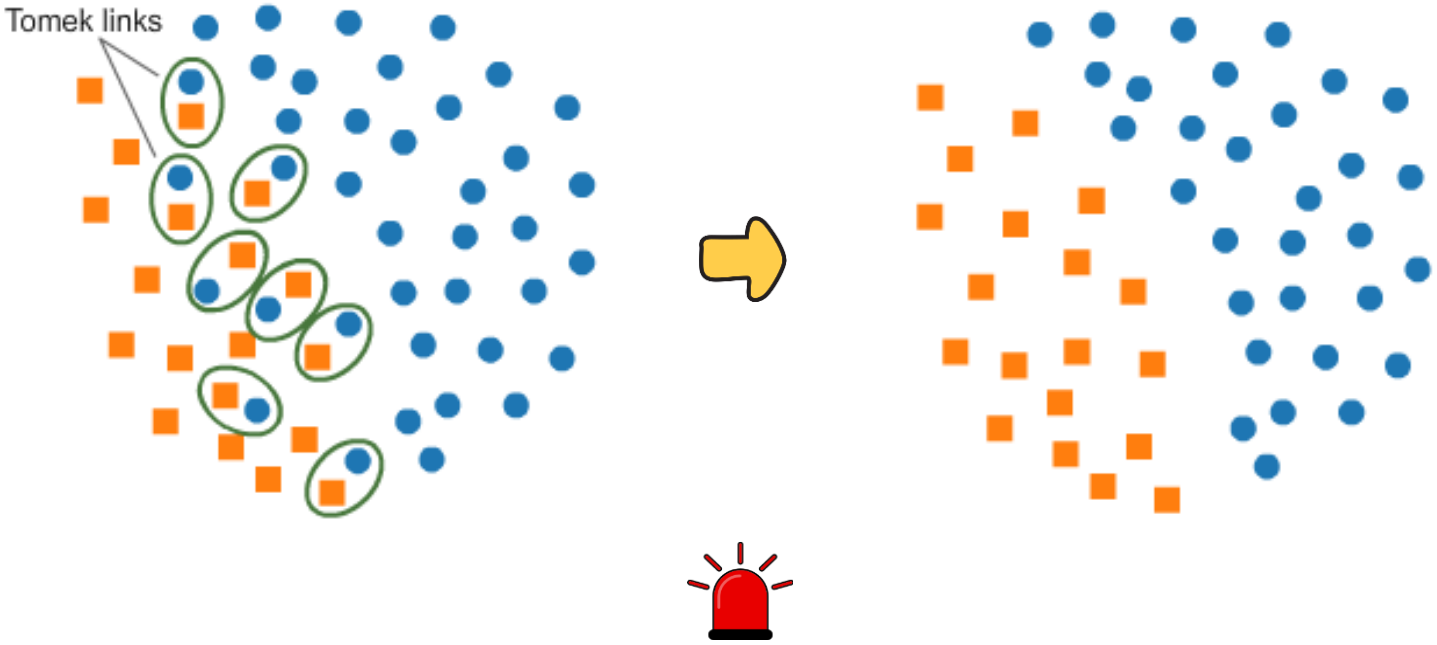


Tomek Links 과정

임의로 서로 다른 클래스의 데이터를 **두 점**을 **선택**하여 이를 **연결**

묶인 데이터 쌍에서 **다수의 클래스**에 속한 데이터들을 **삭제**

Tomek Links



상대적으로 정보의 유실은 크게 방지할 수 있지만

묶이는 값이 한정적이기 때문에 **언더 샘플링의 큰 효과를 얻을 수 없음**

3 샘플링

언더 샘플링

언더 샘플링 (Under Sampling)

다수의 클래스를 소수의 클래스에 맞추어 관측치를 감소시키는 방법

언더 샘플링의 장점

- ✓ 사이즈가 줄어들어 메모리 사용이나 처리 속도 측면에서 유리

언더 샘플링의 단점

- ✓ 관측치의 손실이 일어나기 때문에 정보가 누락

3 샘플링

언더 샘플링

언더 샘플링 (Under Sampling)

다수의 클래스를 소수의 클래스에 맞추어 관측치를 감소시키는 방법

언더 샘플링의 장점

✓ 사이즈가 줄어들어 메모리 사용이나 처리 속도 측면에서 유리

언더 샘플링의 단점

✓ 관측치의 손실이 일어나기 때문에 정보가 누락

3 샘플링

언더 샘플링

언더 샘플링 (Under Sampling)

다수의 클래스를 소수의 클래스에 맞추어 관측치를 감소시키는 방법

언더 샘플링은 관측치를 삭제함으로써 정보를 누락시키는 방법

오버 샘플링을 사용!

언더 샘플링의 단점

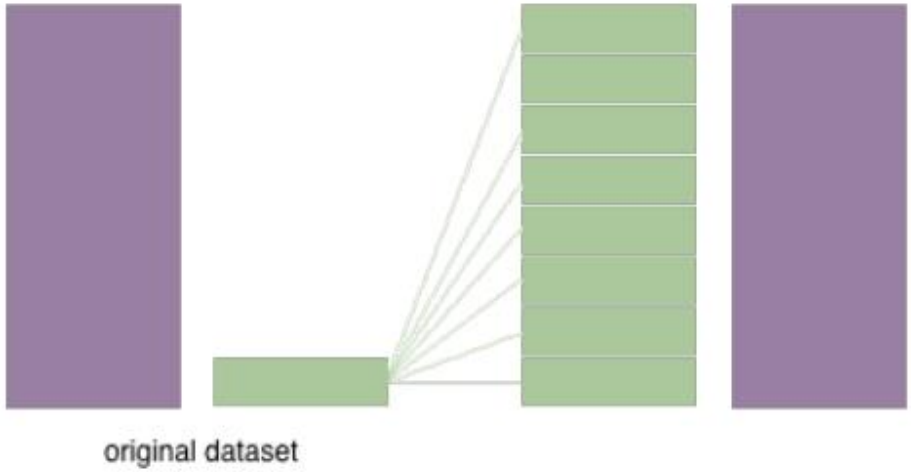
✓ 관측치의 손실이 일어나기 때문에 정보가 누락

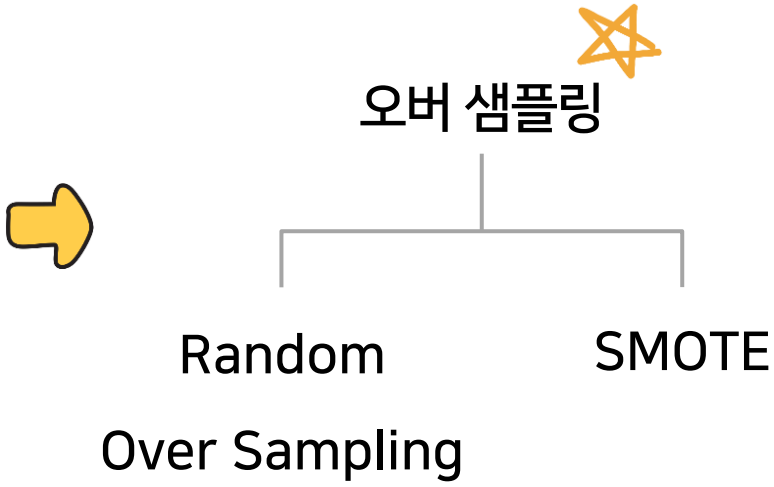
3 샘플링

오버 샘플링

오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞추어 관측치를 증가시키는 방법

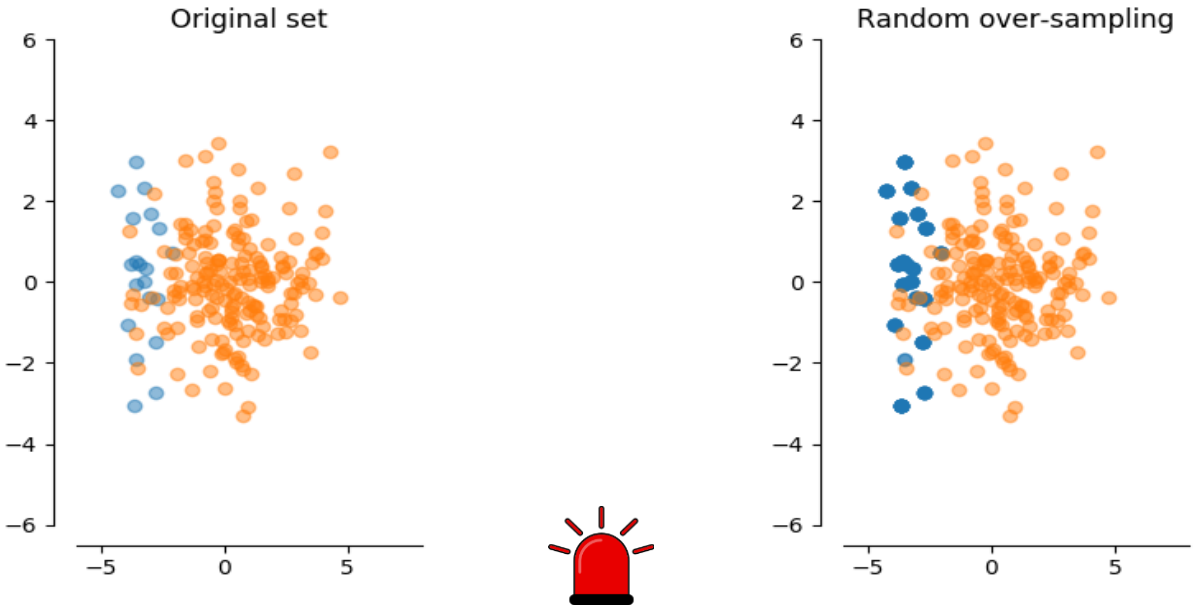




랜덤 오버 샘플링

랜덤 오버 샘플링 (Random Over Sampling)

임의적으로 **소수의 클래스**의 데이터를 **복제**하여 관측치의 수를 늘리는 방법



동일한 데이터의 수가 늘어나 **과적합**될 가능성이 **큼**

3 샘플링

SMOTE

SMOTE 과정

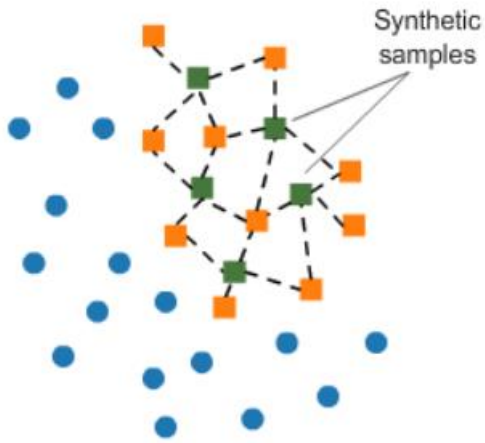
소수 범주의 데이터 중 **무작위**로 하나를 **선택**

선택된 데이터를 기준으로 KNN 알고리즘을 활용해 k개의 가까운 데이터를 선택

선정된 K개의 관측치 중 랜덤으로 일부를 선택 후 앞서 선택된 데이터 사이에 직선을 그려 직선 상에 가상의 소수 클래스 데이터 생성

3 샘플링

SMOTE



SMOTE 과정

소수 범주의 데이터 중 **무작위**로 하나를 **선택**

선택된 데이터를 기준으로 KNN 알고리즘을 활용해 k개의 가까운 데이터를 선택

선정된 K개의 관측치 중 랜덤으로 일부를 선택 후 앞서 선택된 데이터 사이에 직선을 그려 직선 상에 가상의 소수 클래스 데이터 생성

3 샘플링

SMOTE

SMOTE 과정

소수 범주의 데이터 중 **무작위**로 하나를 **선택**

선택된 데이터를 기준으로 **KNN 알고리즘**을 활용해 **k개**의 **가까운 데이터**를 선택

선정된 k개의 관측치 중 랜덤으로 일부를 선택한 후 앞서 선택된 데이터 사이에 직선을 그려 직선 상에 **가상의 소수 클래스 데이터** 생성

3 샘플링

SMOTE

SMOTE는 랜덤 오버 샘플링에 비해 과적합 위험이 낮지만

소수의 클래스의 데이터 간의 거리만을 고려해 데이터를 생성하기 때문에

기존 데이터와 겹치거나 노이즈가 발생할 수 있음

SMOTE 과정

소수 범주의 데이터 중 무작위로 하나를 선택

선택된 데이터를 기준으로 KNN 알고리즘을 활용해 k개의 가까운 데이터를 선택

고차원 데이터에서는 효율적이지 못할 수 있음!

선정된 k개의 관측치 중 랜덤으로 일부를 선택한 후 앞서 선택된 데이터 사이에

직선을 그려 직선 상의 가상의 소수 클래스 데이터 생성

3 샘플링

오버 샘플링

오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞추어 관측치를 증가시키는 방법

오버 샘플링의 장점

✓ 정보의 손실이 없기 때문에 언더 샘플링에 비해 성능이 좋음

오버 샘플링의 단점

✓ 메모리 사용이나 처리속도 측면에서 상대적으로 불리

3 샘플링

오버 샘플링

오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞추어 관측치를 증가시키는 방법

오버 샘플링의 장점

✓ 정보의 손실이 없기 때문에 언더 샘플링에 비해 성능이 좋음

오버 샘플링의 단점

✓ 메모리 사용이나 처리속도 측면에서 상대적으로 불리

3 샘플링

오버 샘플링

오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞추어 관측치를 증가시키는 방법

데이터셋의 구조를 파악해

구조에 맞는 샘플링 기법을 사용!

오버 샘플링의 단점

✓ 메모리 사용이나 처리속도 측면에서 상대

4

인코딩

4 인코딩

인코딩

인코딩 (Encoding)

사용자가 입력한 **문자**나 **기호**를 **컴퓨터 신호**로 **변환**하는 과정

범주형 변수의 **인코딩**을 통해

수치형 변수를 설명변수로 갖는

다양한 분석 기법을 **적용 가능**

인코딩(Encoding)

Classic	Contrast	Bayesian	기타
Ordinal	Simple	Mean Target	Frequency
One-hot	Sum	Leave one out	
Label	Helmert	Weight of Evidence	
Binary	Reverse Helmert	Probability Ratio	
BaseN	Forward Difference	James Stein	
Hashing	Backward Difference	M-estimator	
	Orthogonal Polynomial	Ordered Target	

이걸 다요?

인코딩(Encoding)

Classic	Contrast	Bayesian	기타
Ordinal	Simple	Mean Target	Frequency
One-hot	Sum	Leave one out	
Label	Helmert	Weight of Evidence	
Binary	Reverse Helmert	Probability Ratio	
BaseN	Forward Difference	James Stein	
Hashing	Backward Difference	M-estimator	
	Orthogonal Polynomial	Ordered Target	

4 인코딩

Ordinal Encoding

Ordinal Encoding

순서형 변수를 **인코딩**하는 기법

만족도	점수
매우 불만족	1
불만족	2
보통	3
만족	4
매우 만족	5

- ✓ 1을 기준으로 **순서**에 따라 **차등적**으로 **점수**를 부여
- ✓ 각 수준에 부여된 **점수**들 간에 **순서**와 **연관성**이 존재

4 인코딩

Ordinal Encoding

Ordinal Encoding

순서형 변수를 **인코딩**하는 기법

만족도	점수
매우 불만족	1
불만족	2
보통	3
만족	4
매우 만족	5

추가적인 차원 증가가 발생하지 않음

모델이 빠르게 학습 가능

4 인코딩

Ordinal Encoding

Ordinal Encoding

순서형 변수를 인코딩하는 기법

할당한 점수가 각 수준 간의 정확한 간격의 차이를 반영하기 어려움

만족도	점수
매우 불만족	1
불만족	2
보통	3
만족	4
매우 만족	5

해당 과제의 도메인 지식을 이용해
수준 간 차이를 정확히 반영해야 함

차원적인 차원 증가가
생하지 않음

모델이 빠르게 학습 가능

4 인코딩

One-Hot Encoding

One-Hot Encoding

명목형 변수를 **인코딩**하는 기법

가변수(Dummy Variable)를 생성

분류 모델

J개의 가변수 모두 사용 가능

회귀 모델

J-1개의 가변수를 사용해서 J개 수준 표현

4 인코딩

One-Hot Encoding

One-Hot Encoding

명목형 변수를 **인코딩**하는 기법

가변수(Dummy Variable)를 생성

분류 모델

J개의 가변수 모두 사용 가능

다중공선성 문제

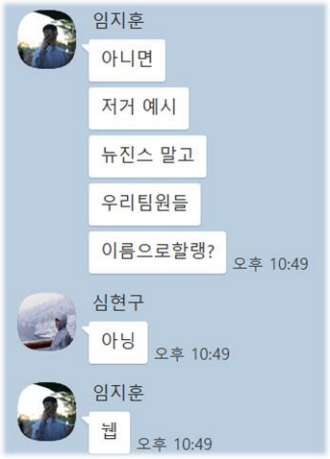
해결 가능!

회귀 모델

J-1개의 가변수를 사용해서 J개 수준 표현

4 인코딩

One-Hot Encoding



또..진스?

뉴진스
하니
민지
다니엘
해린
혜인

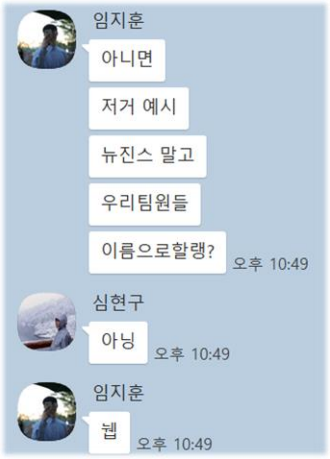
하니	민지	다니엘	해린	혜인
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

각 멤버들이 **새로운 가변수**로 생성되어

열과 **일치**하는 값에는 1, **아닌 값**에는 0이 부여

4 인코딩

One-Hot Encoding



또..진스?

뉴진스
하니
민지
다니엘
해린
혜인

하니	민지	다니엘	해린	혜인
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

회귀분석 모델에서는 **J-1(=4)개 변수**만 생성해도

데이터가 가진 정보는 그대로 유지

4 인코딩

One-Hot Encoding

One-Hot Encoding의 장점

- ✓ **기준 범주**가 intercept 형태로 존재하기 때문에 **해석 용이**
- ✓ **다중공선성** 문제 **해결** 가능

범주형 변수나 그 수준에 따라

너무 많은 가변수가 생성되어 **차원**이 **과다**해짐

많은 컴퓨팅 파워가 요구되고

모델의 학습속도가 느려짐

4 인코딩

One-Hot Encoding

One-Hot Encoding의 장점

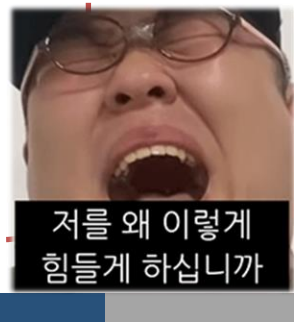
- ✓ **기준 범주**가 intercept 형태로 존재하기 때문에 **해석 용이**
- ✓ **다중공선성** 문제 **해결** 가능

범주형 변수나 그 수준에 따라

너무 많은 가변수가 생성되어 **차원**이 **과다**해짐

많은 컴퓨팅 파워가 요구되고

모델의 학습속도가 느려짐



4 인코딩

Label Encoding

Label Encoding

명목형 변수를 **인코딩**하는 기법

Ordinal Encoding과 같이 각 **수준**에 **점수** 할당

할당한 점수는
수치적인 의미 없음

뉴진스	점수
하니	0
민지	2
다니엘	3
해린	4
혜인	7

1부터 시작할 필요 없음

간격 일정할 필요 없음

명목형 변수를 Encoding 할 때
수준끼리 구분만 되면 충분!

4 인코딩

Label Encoding

Label Encoding

명목형 변수를 **인코딩**하는 기법

Ordinal Encoding과 같이 각 **수준**에 **점수** 할당

뉴진스	점수
하니	0
민지	2
다니엘	3
해린	4
혜인	7

1부터 시작할 필요 없음

간격 일정할 필요 없음

명목형 변수를 Encoding 할 때

수준끼리 구분만 되면 충분!

Label Encoding

Label Encoding의 장점

✓ 가변수의 부재에 따라 **차원**이 **증가하지 않아** 처리속도 빠름

모델 학습과정에서 Ordinal Encoding으로 인식

할당한 **점수**들 간에 **연관성**이 있다고

잘못 판단할 위험이 존재

4 인코딩

Label Encoding

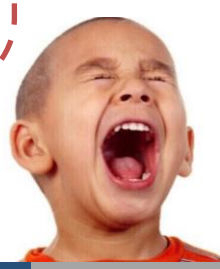
Label Encoding의 장점

✓ 가변수의 부재에 따라 **차원**이 **증가하지 않아** 처리속도 빠름

모델 학습과정에서 **Ordinal Encoding**으로 **인식**

할당한 **점수**들 간에 **연관성**이 있다고

잘못 판단할 위험이 존재



4 인코딩

Mean Encoding

Mean Encoding (Target Encoding)

범주형 변수의 각 **수준**에서 도출된 **반응변수**의 **평균**을

해당 수준에 **동일**하게 할당하는 인코딩 방식

[Y] 키 (cm)	[X] 학과
168	경영
180	경영
168	경영
174	통계
156	통계
163	통계
171	통계
169	경제
180	경제
170	경제

경영 평균 172

통계 평균 166

경제 평균 173

각 학과별 반응변수의

평균을 계산한 값으로

Encoding 진행

4 인코딩

Mean Encoding

Mean Encoding의 장점

- ✓ One-Hot Encoding과 달리 **차원**이 **증가하지 않아** 학습 속도가 빠름
- ✓ 해당 변수와 반응변수의 **관계**를 고려하여 점수를 할당해 **당위성을** 가짐

평균을 활용하기 때문에
이상치에 **취약**해 정보의 왜곡이 발생할 수 있으며
반응변수를 활용해 설명변수를 처리하기 때문에
모델 학습 시 **과적합**이 발생할 수 있음

4 인코딩

Mean Encoding

Mean Encoding의 장점

- ✓ One-Hot Encoding과 달리 **차원**이 **증가하지 않아** 학습 속도가 빠름
- ✓ 해당 변수와 반응변수의 **관계**를 고려하여 점수를 할당해 **당위성을** 가짐

평균을 활용하기 때문에

이상치에 **취약**해 정보의 왜곡이 발생할 수 있으며

반응변수를 활용해 설명변수를 처리하기 때문에

모델 학습 시 **과적합**이 발생할 수 있음

4 인코딩

Mean Encoding

Mean Encoding의 장점

- ✓ One-Hot Encoding의 결과로 변수가 생기지 않아 학습 속도가 빠름
- ✓ 해당 변수와 반응변수의 관계를 모델이 아닌 사람이 설정했다는 점이 **당위적**

Mean Encoding은 Train set에 없던 수준이 Test set에 등장하면 점수를 할당하지 못하고 관측치 값이 적은 범주는 모델링에 부정확한 결과를 도출할 수 있음

적은 데이터로 인코딩한 값은 다량의 Test set에 대한 대표성 감소

평균을 활용하기 때문에
이상치에 **취약**해 정보의 **에러**가 발생할 수 있으며
반응변수를 활용해 설명변수를 치리하기 때문에
모델 학습 시 **과적합** 발생할 수 있음

Smoothing, CV loop, Expanding Mean과 같은 기법으로 보완 가능

Leave-One-Out Encoding

Leave-One-Out Encoding (LOO Encoding)

이상치에 **취약**한 Mean Encoding을 **개선**한 인코딩 기법

[Y] 키 (cm)	[X] 학과	[X] LOO Encoding
168	경영	174
180	경영	168
168	경영	174

수준 별 반응변수의 평균을 계산할 때 **현재 행을 제외**한 나머지 행들의 **평균** 활용

이상치의 영향력 감소!

Leave-One-Out Encoding

Leave-One-Out Encoding (LOO Encoding)

이상치에 취약한 Mean Encoding을 개선한 인코딩 기법

[Y] 키 (cm)	[X] 학과	[X] LOO Encoding
168	경영	174 ---------- $\frac{180+168}{2}$
180	경영	168 ---------- $\frac{168+168}{2}$
168	경영	174 ---------- $\frac{168+180}{2}$

수준이 경영으로 **같지만**
서로 다른 값 할당해
Encoding 진행

Leave-One-Out Encoding

Leave-One-Out Encoding (LOO Encoding)

이상치에 취약한 Mean Encoding을 개선한 인코딩 기법

Leave-One-Out Encoding의 장점

- ✓ **이상치**의 영향을 **덜 받음**
- ✓ 스스로의 반응변수 값은 제외하기 때문에
  과적합 위험성이 Mean Encoding보다 **상대적으로 낮음**

Mean Encoding과 동일한 한계를 가짐

Leave-One-Out Encoding

Leave-One-Out Encoding (LOO Encoding)

이상치에 취약한 Mean Encoding을 개선한 인코딩 기법

Leave-One-Out Encoding의 장점

- ✓ **이상치**의 영향을 **덜 받음**
- ✓ 스스로의 반응변수 값은 제외하기 때문에
  과적합 위험성이 Mean Encoding보다 **상대적으로 낮음**

Mean Encoding과 동일한 한계를 가짐

Ordered Target Encoding

Ordered Target Encoding (CatBoosting Encoding)

같은 수준에 속하는 행들 중 **이전 행들 값의 평균**을 할당하는 인코딩 기법

[Y] 키 (cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	170
174	통계	166	170
169	경제	173	170
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	173	165
170	경제	173	172.5

각 수준의 **첫 번째 행**은
이전 값이 없으므로
전체 데이터의
반응변수의 **평균** 사용

Ordered Target Encoding

Ordered Target Encoding (CatBoosting Encoding)

같은 수준에 속하는 행들 중 **이전 행들 값의 평균**을 할당하는 인코딩 기법

[Y] 키 (cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	170
174	통계	166	170
169	경제	173	170
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	173	165
170	경제	173	172.5

앞선 **두 통계학과 행**의 **반응변수**의 **평균**

$$\frac{174 + 156}{2} = 165$$

안녕히 계세요 여러분 ~
범주팀은 주제 분석 주제를 찾아 떠납니다 ~

감사합니다