

범주형자료분석팀

3팀

권가민
김수인
김준령
박윤아
이정민

INDEX

1. 혼동행렬
2. ROC 곡선
3. 샘플링
4. 인코딩

1

혼동행렬

혼동행렬

혼동행렬 (Confusion Matrix)

분류 모델의 성능을 평가하기 위한 지표

예측값 (\hat{Y})과 실제값 (Y)을 비교

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

T(True) / F(False) : 실제값과 예측값의 일치 여부

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

혼동행렬

혼동행렬의 네 가지 상황

① TP (True Positive)

긍정 ($\hat{Y} = 1$)으로 예측하였으며 실제 관측값도 긍정 ($Y = 1$)인 경우

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

혼동행렬

혼동행렬의 네 가지 상황

② TN (True Negative)

부정 ($\hat{Y} = 0$)으로 예측하였으며 실제 관측값도 부정 ($Y = 0$)인 경우

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

혼동행렬

혼동행렬의 네 가지 상황

③ FP (False Positive)

긍정 ($\hat{Y} = 1$)으로 예측하였으나 실제 관측값은 부정 ($Y = 0$)인 경우

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

1종 오류

혼동행렬

혼동행렬의 네 가지 상황

④ FN (False Negative)

부정 ($\hat{Y} = 0$)으로 예측하였으나 실제 관측값은 긍정 ($Y = 1$)인 경우

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

2중 오류

혼동행렬

혼동행렬의 한계



정보의 손실이 일어날 수 있음



임의적인 Cut-off Point를 설정



혼동행렬

혼동행렬의 한계



정보의 손실이 일어날 수 있음

이항변수에 맞게
모델의 예측값($\hat{\pi}$) 범주화



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

혼동행렬

혼동행렬의 한계



임의적인 Cut-off Point를 설정

분석자가 임의로
Cut-off Point 설정



분석의 객관성을 떨어트림

ROC 곡선을 이용해 한계를 보완!

분류평가지표

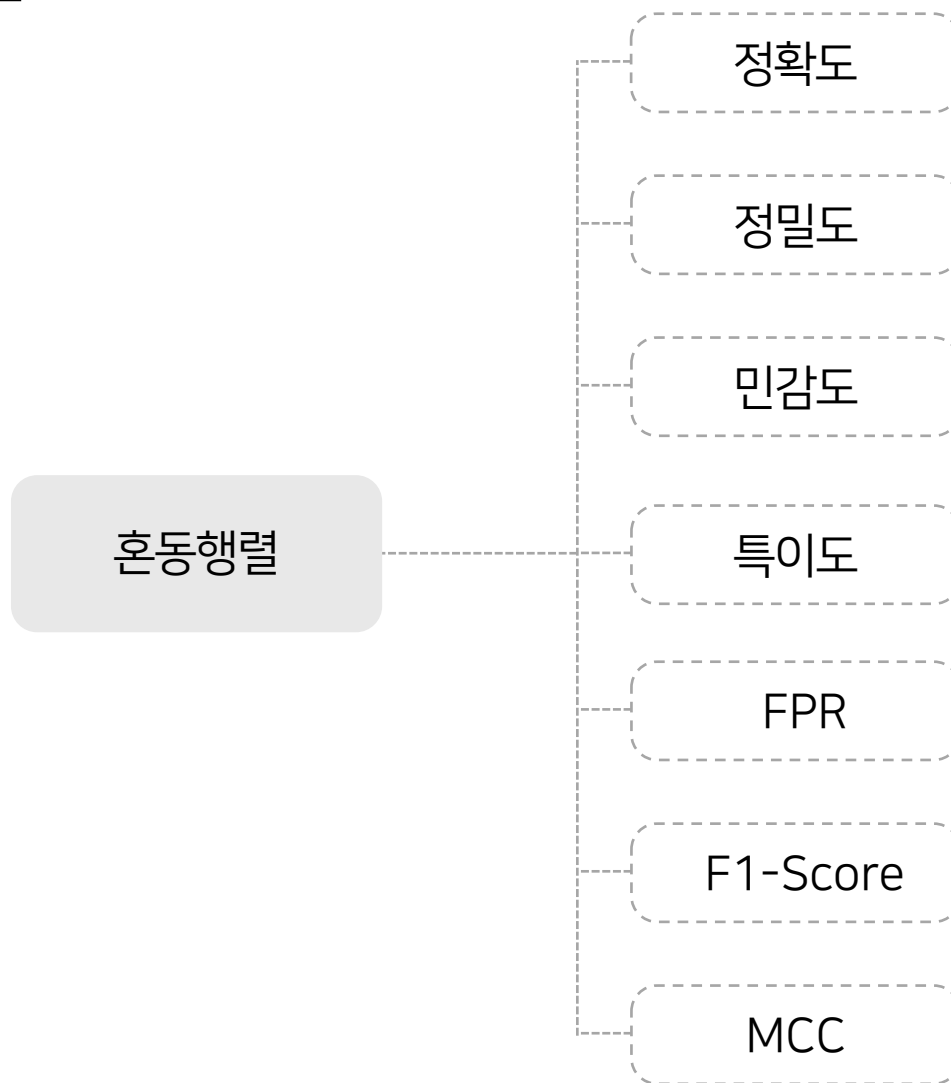
분류평가지표

혼동행렬을 활용하여 분류모델의 성능을 평가하는 지표



각 지표의 특징과 한계를 파악하여
분류평가의 목적에 맞는 적절한 지표를 선택해야 함!

분류평가지표



분류평가지표

① 정확도

정확도 (Accuracy)

전체 경우에서 예측값과 실제값이 일치하는 비율

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

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

분류평가지표

① 정확도

정확도 (Accuracy)

전체 경우에서 예측값과 실제값이 일치하는 비율

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



- 1에 가까울수록 좋은 성능
- 불균형 데이터에서는 관측치가 많은 수준(Class)에 의존하여
정확한 성능 파악 불가

분류평가지표

② 정밀도

정밀도 (Precision)

긍정 ($\hat{Y} = 1$)으로 예측한 값들 중 실제 관측치 역시 긍정 ($Y = 1$)인 비율

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

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

분류평가지표

② 정밀도

정밀도 (Precision)

긍정 ($\hat{Y} = 1$)으로 예측한 값들 중 실제 관측치 역시 긍정 ($Y = 1$)인 비율

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



- 1에 가까울수록 좋은 성능
- FP가 치명적일 때 주로 사용

유죄 ($Y = 1$)를 무죄로 선고 ($\hat{Y} = 0$)하는 것보다
무죄 ($Y = 0$)를 유죄로 선고 ($\hat{Y} = 1$)하는 것이 더 위험

분류평가지표

③ 민감도 / 재현율

민감도 (Sensitivity) / 재현율 (Recall)

실제 긍정 ($Y = 1$)인 관측값 중 긍정 ($\hat{Y} = 1$)으로 예측한 비율

$$Sensitivity(Recall) = \frac{TP}{TP + FN}$$

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

분류평가지표

③ 민감도 / 재현율

민감도 (Sensitivity) / 재현율 (Recall)

실제 긍정 ($Y = 1$)인 관측값 중 긍정 ($\hat{Y} = 1$)으로 예측한 비율

$$Sensitivity(Recall) = \frac{TP}{TP + FN}$$



- 1에 가까울수록 좋은 성능
- ROC 곡선의 Y축
- FN이 치명적일 때 주로 사용

암세포가 없는 사람 ($Y = 0$)에게 암이라고 진단 ($\hat{Y} = 1$)하는 것보다
실제로 암세포가 있는 사람 ($Y = 1$)에게 건강하다 ($\hat{Y} = 0$)고 진단하는 것이 더 위험

분류평가지표

④ 특이도

특이도 (Specificity)

실제 부정 ($Y = 0$)인 관측값 중 부정 ($\hat{Y} = 0$)으로 예측한 비율

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

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

분류평가지표

④ 특이도

특이도 (Specificity)

실제 부정 ($Y = 0$)인 관측값 중 부정 ($\hat{Y} = 0$)으로 예측한 비율

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



1에 가까울수록 좋은 성능

분류평가지표

⑤ FPR

FPR (False Positive Rate)

실제 부정 ($Y = 0$)인 관측값 중 긍정 ($\hat{Y} = 1$)으로 잘못 예측한 비율

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

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

분류평가지표

⑤ FPR

FPR (False Positive Rate)

실제 부정 ($Y = 0$)인 관측값 중 긍정 ($\hat{Y} = 1$)으로 잘못 예측한 비율

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

잘못 예측한 비율이므로



- 0에 가까울수록 좋은 성능
- ROC 곡선의 X축

분류평가지표

⑥ F1-Score

F1-Score

정밀도와 재현율의 조화평균

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



1에 가까울수록 모델의 성능이 좋다고 판단

분류평가지표

⑥ F1-Score

F1-Score

정밀도와 재현율의 조화평균

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



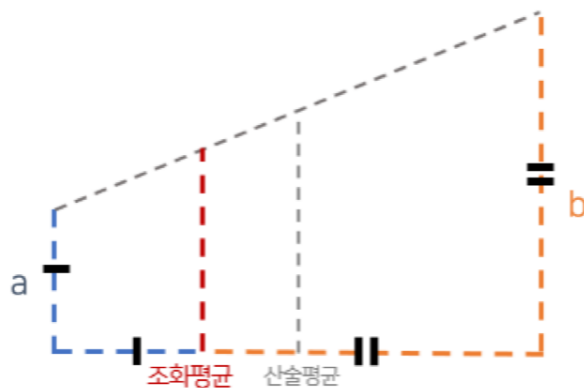
1에 가까울수록 모델의 성능이 좋다고 판단

분류평가지표

⑥ F1-Score

조화평균을 사용하는 이유

불균형 데이터에서 정확도의 한계 보완
상충관계(Trade-off)에 있는 정밀도와 재현율을 모두 균형 있게 반영



더 큰 값에 패널티를 주어 작은 값에 가까운 평균 도출

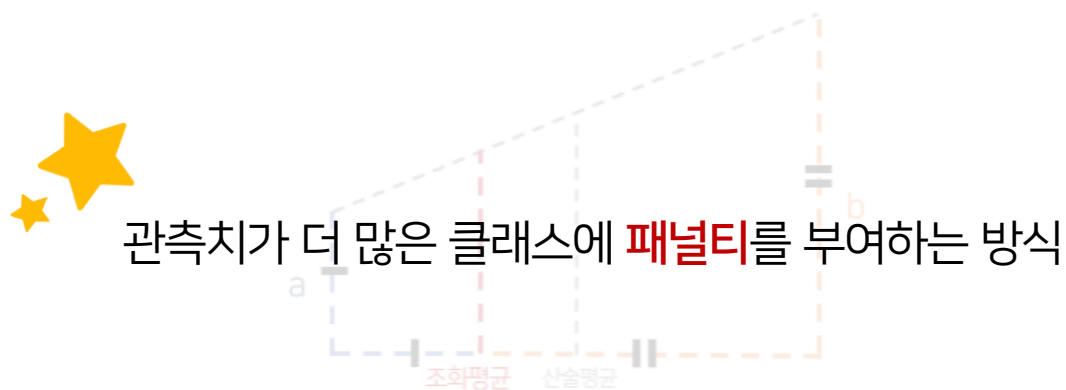
분류평가지표

⑥ F1-Score

조화평균을 사용하는 이유

불균형 데이터에서 정확도의 한계 보완

상충관계(Trade-off)에 있는 정밀도와 재현율을 모두 균형 있게 반영



더 큰 값에 패널티를 주어 작은 값에 가까운 평균 도출

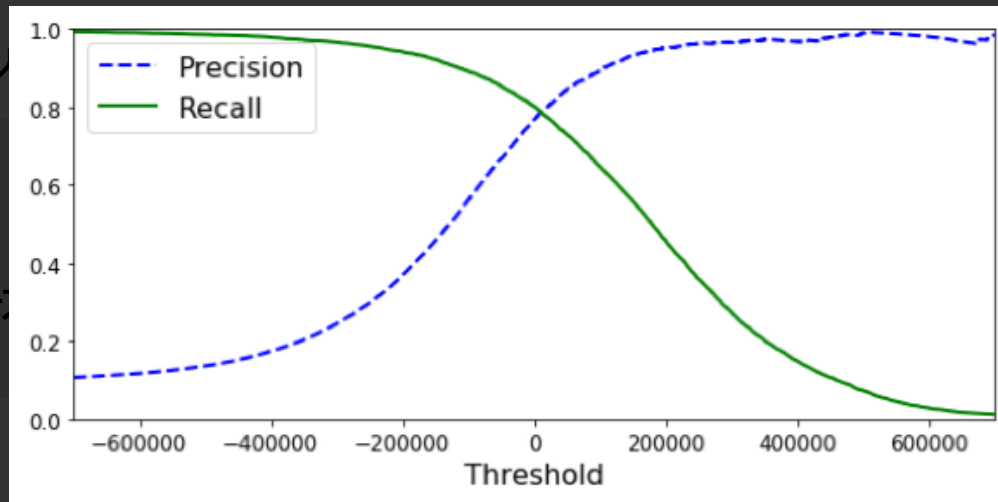
분류평가지표 정밀도와 재현율의 상충관계 (Trade-off)

⑥ F1-Score

조화평균을 사용

상충관계

계 반영



정밀도와 재현율은 **동시에 큰 값을 지닐 수 없음**

- 관측치가 더 많은 클래스에 **패널티**를 줌
 - 해당 클래스에 대한 의존성을 줄여 두 클래스를 균형있게 반영
- 정밀도가 높아지면 재현율이 낮아짐

재현율이 높아지면 정밀도가 낮아짐

분류평가지표

⑥ F1-Score

한계

F1-Score는 TN(True Negative) 수치를 반영하지 않는다는 한계를 가짐

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

		실제값 (Y)	
		Y = 1	Y = 0
예측값 (\hat{Y})	$\hat{Y} = 1$	26	27
	$\hat{Y} = 0$	24	101

$$F1\ score = \frac{2 \times 26}{2 \times 26 + 27 + 24} = 0.505$$



같은 값을 가짐

분류평가지표

⑥ F1-Score

한계

F1-Score는 TN(True Negative) 수치를 반영하지 않는다는 한계를 가짐

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

		실제값 (Y)	
		Y = 1	Y = 0
예측값 (\hat{Y})	$\hat{Y} = 1$	26	27
	$\hat{Y} = 0$	24	101

$$F1\ score = \frac{2 \times 26}{2 \times 26 + 27 + 24} = 0.505$$



같은 값을 가짐

분류평가지표

⑥ F1-Score

한계

F1-Score는 TN(True Negative) 수치를 반영하지 않는다는 한계를 가짐

		실제값 (Y)				실제값 (Y)	
		Y = 1	Y = 0			Y = 1	Y = 0
예측값 (\hat{Y})	$\hat{Y} = 1$	26	27			26	27
	$\hat{Y} = 0$	2	101			24	101

TN (True Negative)의 수치를 반영하지 않는

F1-score의 한계를

MCC 지표를 사용하여 보완 가능!



$$F1\ score = \frac{2 \times 26}{2 \times 26 + 27 + 24} = 0.505$$



같은 값을 가짐

분류평가지표

⑦ MCC

MCC (Matthews Correlation Coefficient)

혼동행렬의 모든 구성요소를 활용하여 계산

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



상관계수 값이기 때문에 -1과 1 사이의 값을 가짐

분류평가지표

⑦ MCC

MCC (Matthews Correlation Coefficient)

혼동행렬의 모든 구성요소를 활용하여 계산

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

-1에 가까울수록

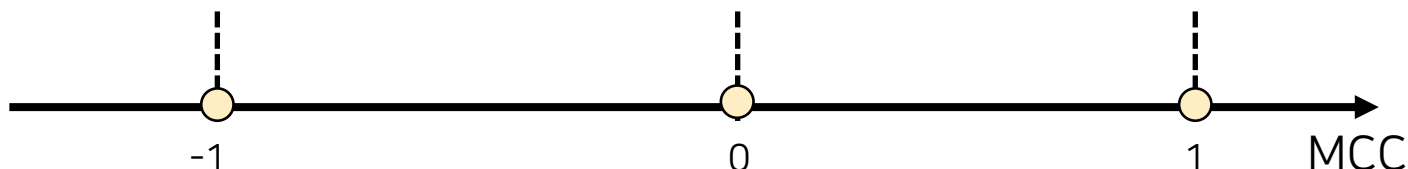
역 예측

0에 가까울수록

랜덤 예측

1에 가까울수록

완전 예측



분류평가지표

⑦ MCC

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

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

TP 관측치가 매우 큰 불균형 데이터

$$F1\ score = \frac{2 \times 92}{2 \times 92 + 4 + 3} = 0.96,$$

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

분류평가지표

⑦ MCC

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

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

TN 관측치가 매우 큰 불균형 데이터

$$F1\ score = \frac{2 \times 1}{2 \times 1 + 3 + 4} = 0.22,$$

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

분류평가지표

⑦ MCC

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

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

$$F1\ score = \frac{2 \times 92}{2 \times 92 + 4 + 3} = 0.96$$

$$F1\ score = \frac{2 \times 1}{2 \times 1 + 3 + 4} = 0.22$$

F1-Score는 큰 차이가 나타남

분류평가지표

⑦ MCC

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

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

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

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

MCC는 0.18로 같음

분류평가지표

⑦ MCC



F1-Score는 TN값을 반영하지 않기 때문에

TN값의 차이가 클수록 F1-Score의 값에 차이가 발생

		실제값 (Y)				실제값 (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



F1-Score 한 가지만 이용하여

모델의 성능을 판단하는 것은 매우 위험

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

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

MCC는 0.18로 같음

분류평가지표



⑦ MCC

MCC가 반드시 F1-Score보다 좋은 지표일까?

분석 목적이 모든 클래스에 대한 **균형적인 평가**라면 **MCC**

희귀질환과 같이 중요하지만 관측치가 적은 경우, 이를 **Positive**로 두고 **F1-Score**

		실제값 (Y)				실제값 (Y)	
		Y = 1	Y = 0			Y = 1	Y = 0
예측값 (Ŷ)	Ŷ = 1	92	2	예측값 (Ŷ)	Ŷ = 1	1	2
	Ŷ = 0	3	1		Ŷ = 0	4	92



목적에 맞는 분류평가지표를 선택하는 것이 **중요!**

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

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

MCC는 0.18로 같음

2

ROC 곡선

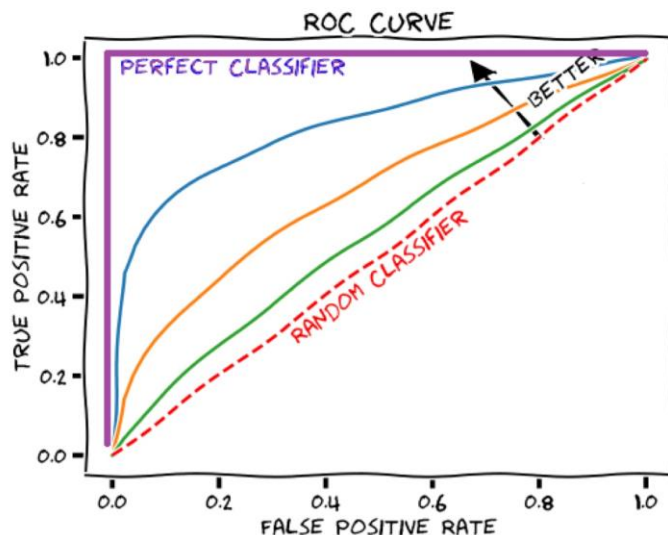
ROC 곡선 (Receiver Operating Characteristic Curve)

ROC 곡선

0~1 사이의 가능한 모든 cut-off point에 대해
재현율을 1-특이도의 함수로 나타낸 곡선 혹은 직선

Y축

FPR, X축



모든 cut-off point에 대한 혼동행렬을 구하고
각 혼동행렬에 해당하는 X, Y값을
그래프에 점으로 표시해 연결한 것!

ROC 곡선 (Receiver Operating Characteristic Curve)

ROC 곡선의 장점



혼동 행렬보다 더 많은 정보를 가짐



주어진 모형에서 가장 적합한 cut-off point를 찾을 수 있음



ROC 곡선 (Receiver Operating Characteristic Curve)

ROC 곡선의 형태

(0,0)과 (1,1)을 잇는 위로 볼록한 우상향하는 형태



(0,0)에서 (1,1)을 잇는 형태인 이유

Cut-off point가 0에 가까운 값을 지님 → 대부분 $\hat{Y} = 1$ 로 예측 →

TP&FP 증가 → TN&FN 감소 → TPR&FPR (1,1)에 가까워짐

(0,0)에 가까워지는 과정은 완전히 반대의 과정을 따름

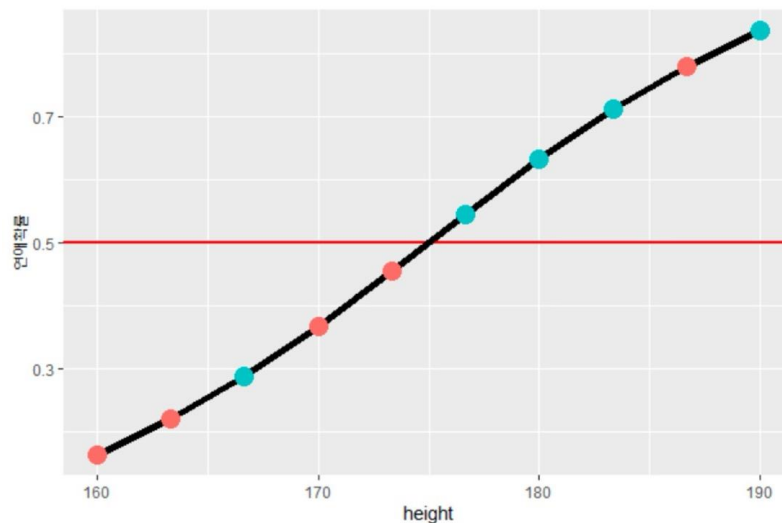
ROC 곡선 (Receiver Operating Characteristic Curve)

최적의 Cut-off point 찾기 예시

$$\log \left[\frac{\pi(x)}{1 - \pi(x)} \right] = -19 + 0.1x$$

● : 실제로 연애중인 관측치 ($Y = 1$)

● : 실제로 연애를 하고 있지 않은 관측치 ($Y = 0$)



키에 따른 연애 여부

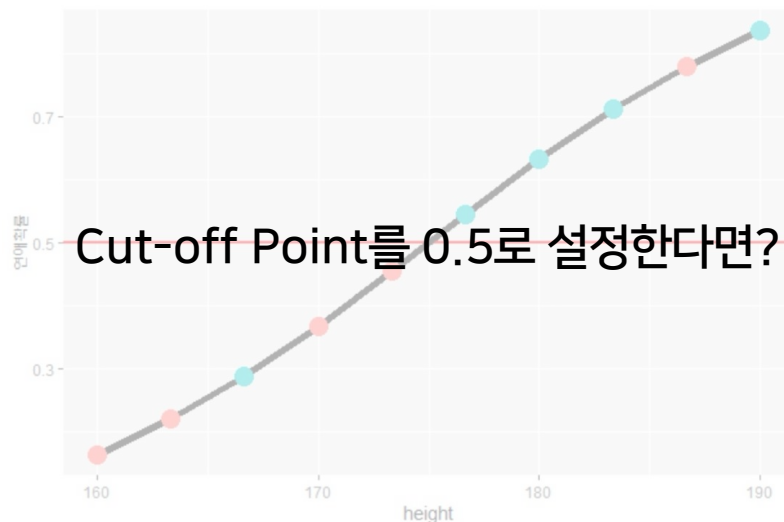
ROC 곡선 (Receiver Operating Characteristic Curve)

최적의 Cut-off point 찾기 예시

$$\log \left[\frac{\pi(x)}{1 - \pi(x)} \right] = -19 + 0.1x$$

● : 실제로 연애중인 관측치 ($Y = 1$)

● : 실제로 연애를 하고 있지 않은 관측치 ($Y = 0$)



ROC 곡선 (Receiver Operating Characteristic Curve)

최적의 Cut-off point 찾기 예시

$$\log \left[\frac{\pi(x)}{1 - \pi(x)} \right] = -19 + 0.1x$$

● : 실제로 연애중인 관측치 ($Y = 1$)

● : 실제로 연애를 하고 있지 않은 관측치 ($Y = 0$)

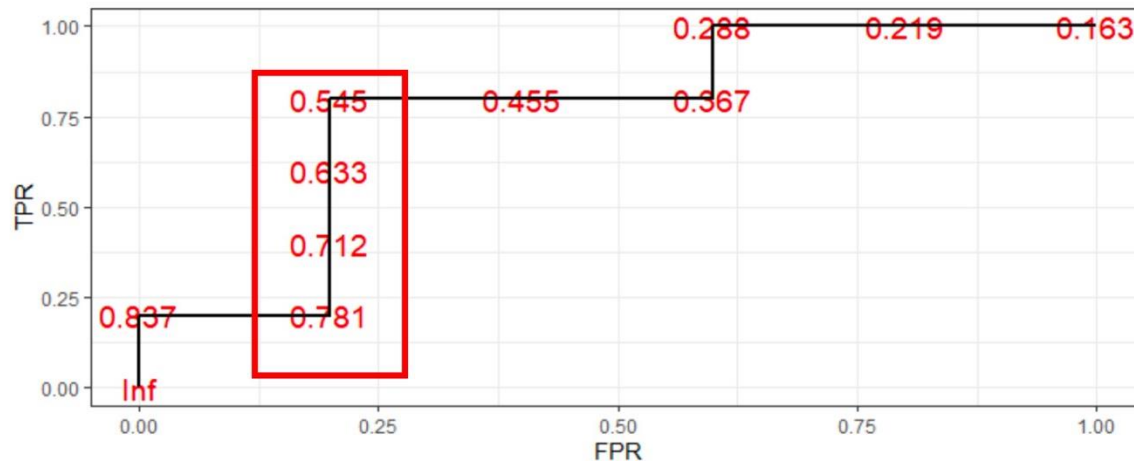
		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	4	1
	$\hat{Y} = 0$	1	4

$$TPR = \frac{TP}{TP+FN} = \frac{4}{4+1} = 0.8$$

$$FPR = \frac{FP}{FP+TN} = \frac{1}{1+4} = 0.2$$

ROC 곡선 (Receiver Operating Characteristic Curve)

최적의 Cut-off point 찾기 예시

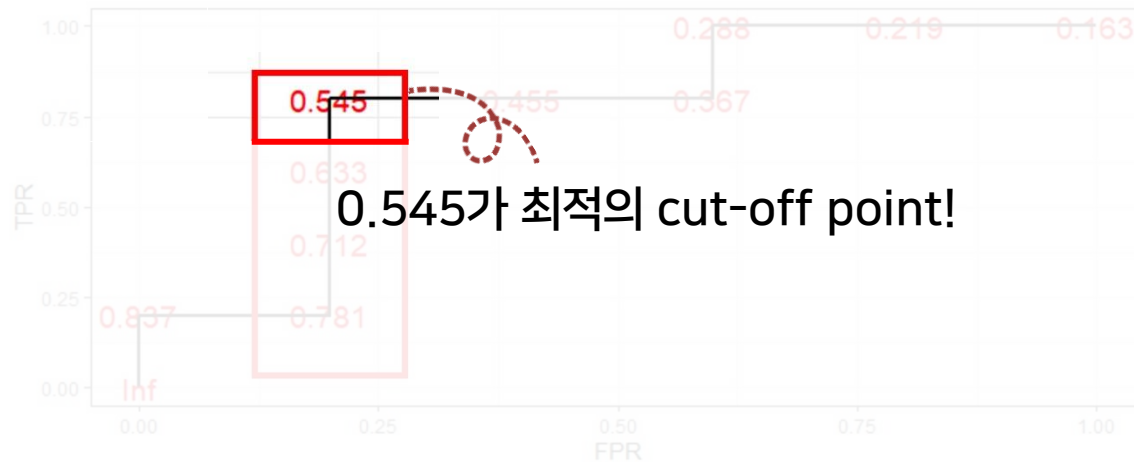


모든 cut-off point에 대한 TPR, FPR을 구하고 곡선을 그린 경우

Y값(TPR)이 같을 때, X값(FPR)이 더 작을수록
X값(FPR)이 같을 때, Y값(TPR)이 더 클수록
좋은 cut-off point임

ROC 곡선 (Receiver Operating Characteristic Curve)

최적의 Cut-off point 찾기 예시



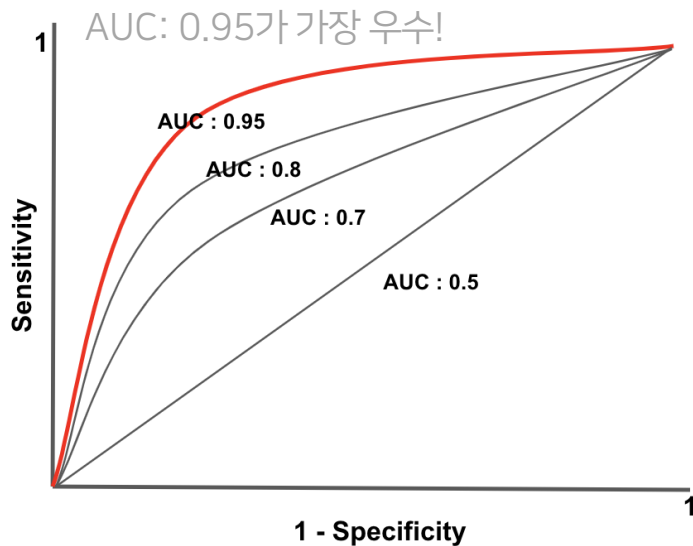
모든 cut-off point에 대한 TPR, FPR을 구하고 곡선을 그린 경우

Y값(TPR)이 같을 때, X값(FPR)이 더 작을수록
X값(FPR)이 같을 때, Y값(TPR)이 더 클수록
좋은 cut-off point임

AUC 곡선

AUC (Area Under Curve)

ROC 곡선 아래의 면적으로 0~1 사이의 값을 가짐



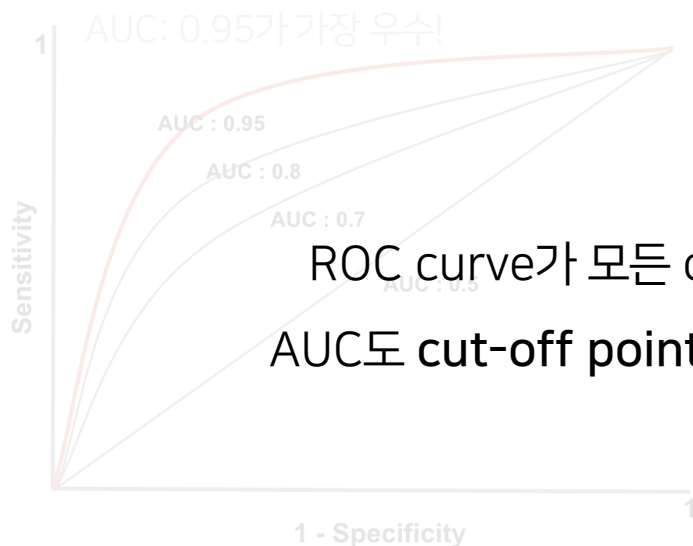
ROC 곡선이 볼록할수록 좋은 모델이므로
AUC 값이 클수록 모델의 성능이 우수!



AUC 곡선

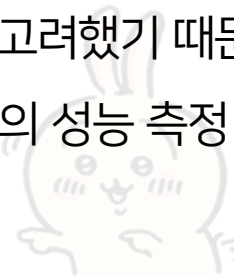
AUC (Area Under Curve)

ROC 곡선 아래의 면적으로 0~1 사이의 값을 가짐



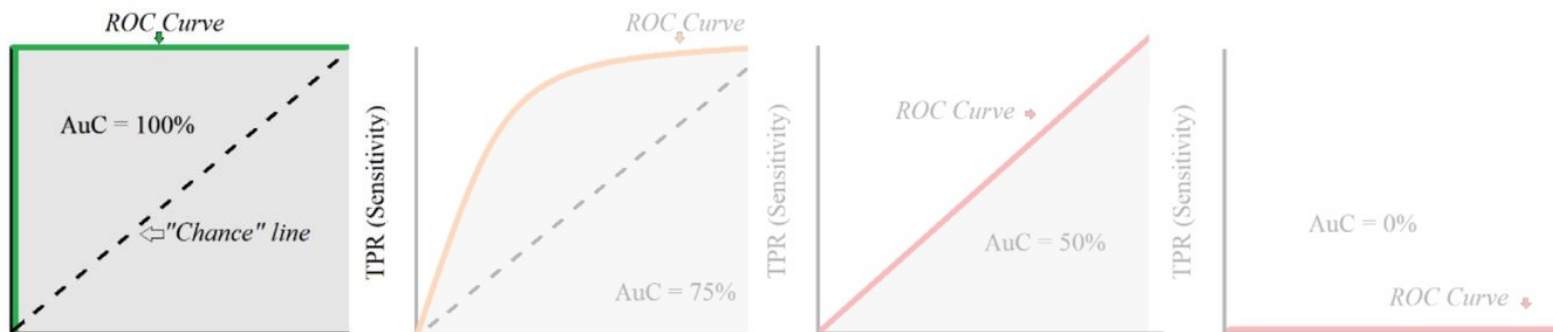
ROC 곡선이 볼록할수록 좋은 모델이므로
AUC 값이 클수록 모델의 성능이 우수!

ROC curve가 모든 cut-off point를 고려했기 때문에
AUC도 cut-off point와 상관없이 모델의 성능 측정 가능!



AUC 곡선

AUC의 해석

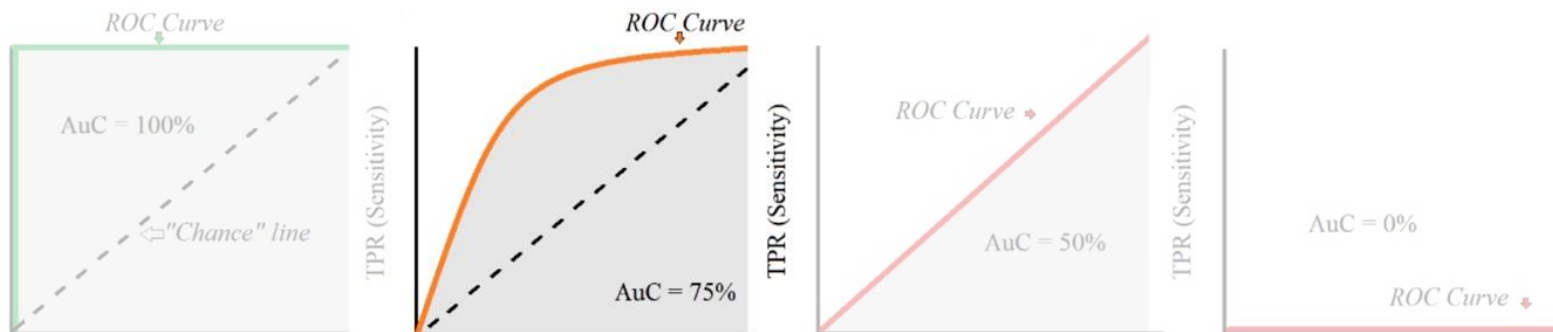


$$AUC = 1$$

모델을 100% 정확히 예측했다는 의미로
모델이 과적합(Overfitting)된 것은 아닌지 확인해야 함

AUC 곡선

AUC의 해석

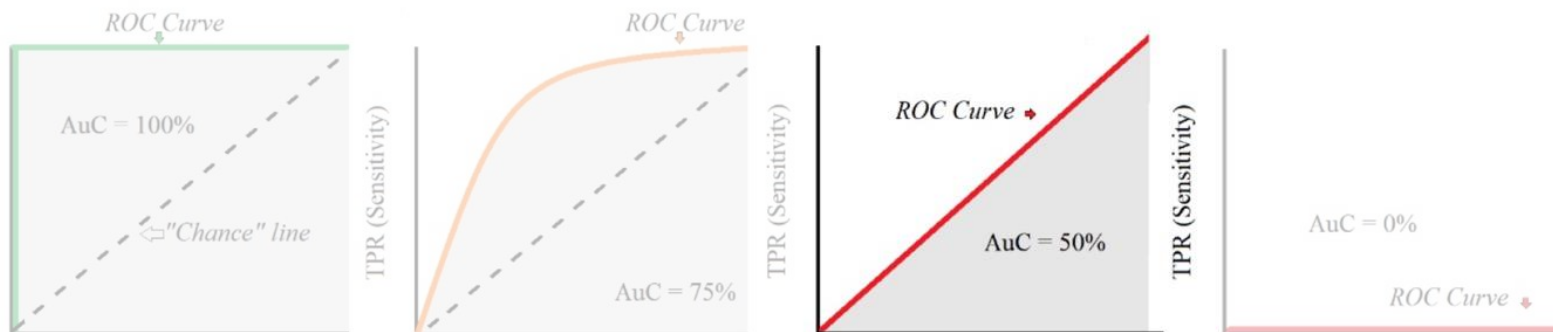


$$\text{AUC} = 0.75$$

모델이 실제 값을 75% 수준으로 맞췄다는 의미로
일반적으로 AUC가 0.8 이상이면 성능이 우수하다고 함

AUC 곡선

AUC의 해석

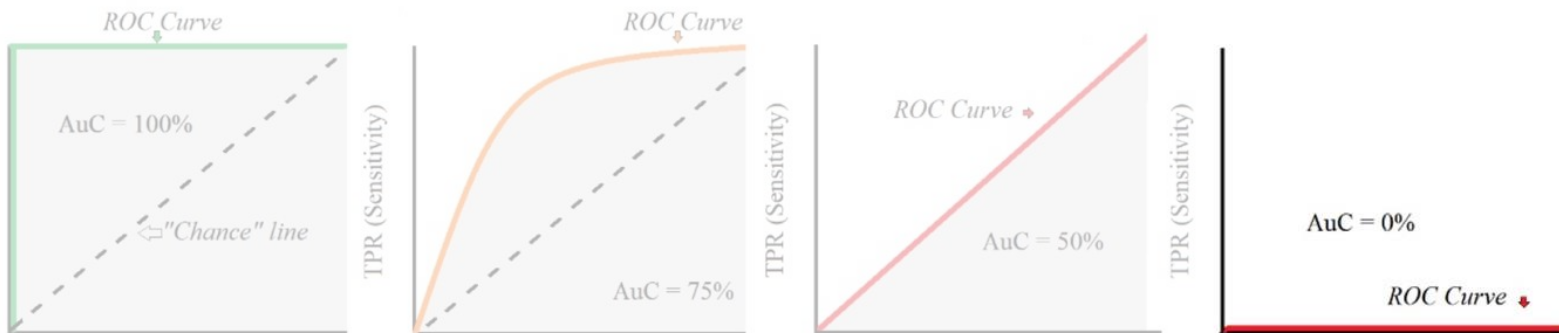


$$AUC = 0.5$$

모델이 실제 값을 50% 맞췄다는 의미로 무작위 예측과 다름이 없으며
보통 AUC는 0.5 이상의 값을 보여야 정상임

AUC 곡선

AUC의 해석



AUC = 0

모델이 100% 반대로, 즉 $Y=1$ 과 $Y=0$ 을 반대로 예측했다는 의미



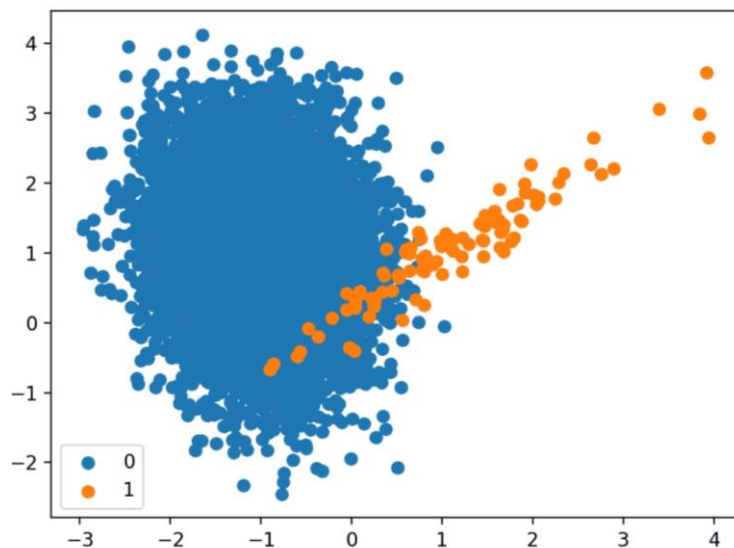
3

샘플링

샘플링 (Sampling)

클래스 불균형

각 수준(클래스)에 따른 관측치 개수의 차이가 큰 경우



Ex) 올림픽 메달이 있는 사람과 없는 사람



샘플링 (Sampling)

샘플링의 필요성

클래스 불균형

Y=0
5개

Y=1
95개

각 클래스에 따른 관측한 레수의 차이가 클 경우 모두 $\hat{Y} = 1$ 으로 예측할 경우

Y = 0 은 전혀 예측하지 못했지만

정확도는 95%인 모델이 됨

모델의 성능을 정확히 파악하기 어려움



Ex) 올림픽 메달이 있는 사람과 없는 사람



샘플링 (Sampling)

샘플링의 필요성

클래스 불균형



각 클래스에 따른 관측한 레수의 차이가 클 경우 모두 $\hat{Y} = 1$ 으로 예측할 경우

$Y = 0$ 은 전혀 예측하지 못했지만

정확도는 95%인 모델이 됨

모델의 성능을 정확히 파악하기 어려움



Ex) 올림픽 메달이 있는 사람과 없는 사람

샘플링 (Sampling)

샘플링의 필요성 예시

		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	6	5
	$\hat{Y} = 0$	4	5

$Y=1$ 수준에서의 정확도 0.6

$Y=0$ 수준에서의 정확도 0.5

혼동행렬의 전체 정확도: 0.55

		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	6	50
	$\hat{Y} = 0$	4	50

$Y=1$ 수준에서의 정확도 0.6

$Y=0$ 수준에서의 정확도 0.5

혼동행렬의 전체 정확도: 0.509

샘플링 (Sampling)

샘플링의 필요성 예시

		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	6	5
	$\hat{Y} = 0$	4	5

$Y=1$ 수준에서의 정확도 0.6

$Y=0$ 수준에서의 정확도 0.5

혼동행렬의 전체 정확도: 0.55

		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	6	50
	$\hat{Y} = 0$	4	50

$Y=1$ 수준에서의 정확도 0.6

$Y=0$ 수준에서의 정확도 0.5

혼동행렬의 전체 정확도: 0.509

샘플링 (Sampling)

샘플링의 필요성 예시

		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	6	5
	$\hat{Y} = 0$	4	5

$Y=1$ 수준에서의 정확도 0.6

$Y=0$ 수준에서의 정확도 0.5

혼동행렬의 전체 정확도: 0.55

		관측값(Y)	
		$Y = 1$	$Y = 0$
예측값 (\hat{Y})	$\hat{Y} = 1$	6	50
	$\hat{Y} = 0$	4	50

$Y=1$ 수준에서의 정확도 0.6

$Y=0$ 수준에서의 정확도 0.5

혼동행렬의 전체 정확도: 0.509

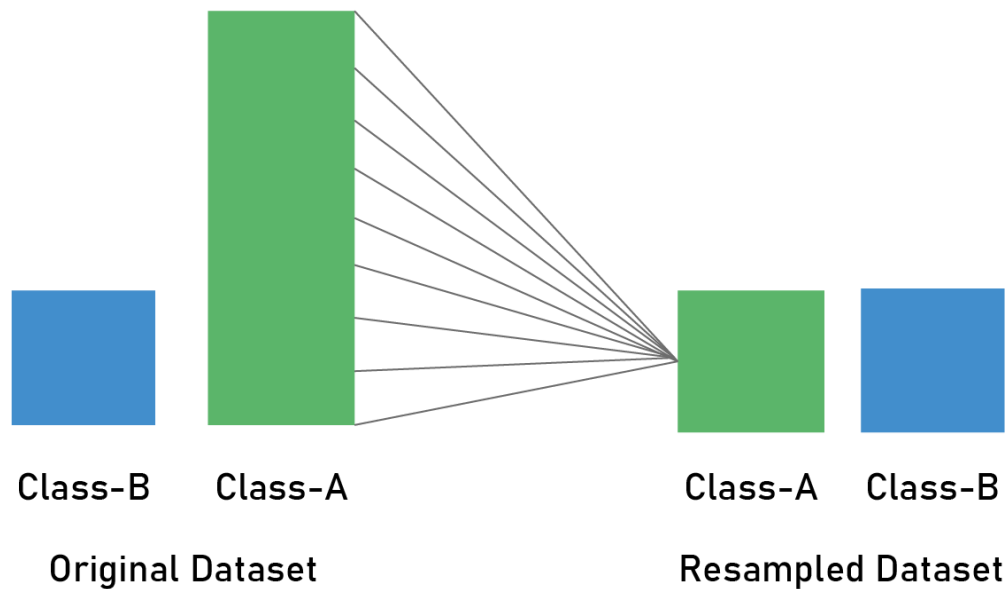


정확도는 **관측치가 많은 수준**의 영향을 받음!

샘플링 (Sampling)

언더 샘플링 (Under Sampling)

소수의 클래스는 변형하지 않고,
다수의 클래스를 소수의 클래스에 맞추어 **관측치를 감소**시키는 방법



샘플링 (Sampling)

언더 샘플링 (Under Sampling)

소수의 클래스는 변형하지 않고,
다수의 클래스를 소수의 클래스에 맞추어 **관측치를 감소**시키는 방법



데이터 사이즈가 줄어들어 메모리 사용이나 처리 속도에 있어 유리하지만
관측치 손실로 인해 정보가 누락되는 문제 발생

샘플링 (Sampling)

언더 샘플링의 종류

Tomek Links Method

두 클래스 간 경계에 있는 데이터를 제거하는 방법

Tomek Link로 묶이는 값이 한정적이므로 효과가 크지 않음

Random Under Sampling

랜덤으로 다수의 클래스에 해당하는 데이터를 제거하는 방법

추출된 샘플들이 기존 데이터에 대해 대표성을 띄지 못할 수도 있음

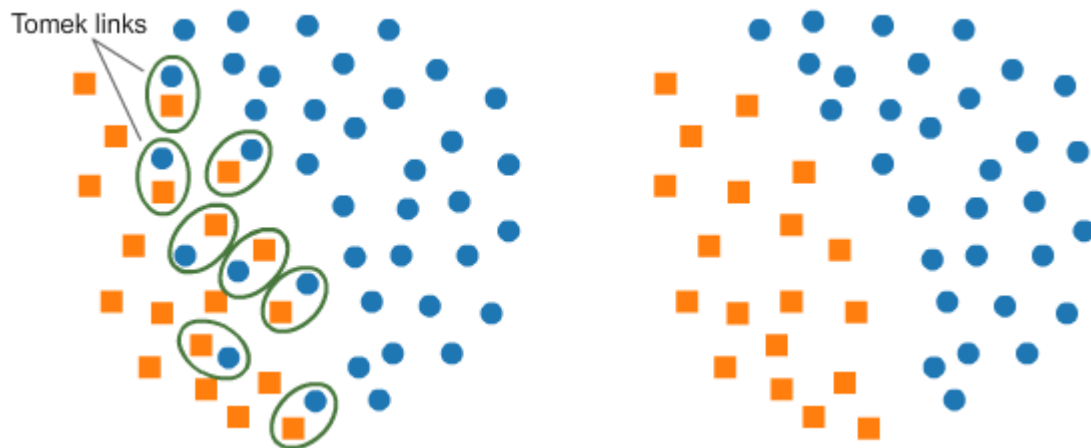
샘플링 (Sampling)

언더 샘플링의 종류

Tomek Links Method

두 클래스 간 경계에 있는 데이터를 제거하는 방법

Tomek Link로 묶이는 값이 한정적이므로 효과가 크지 않음



샘플링 (Sampling)

언더 샘플링의 종류

Tomek Links Method

두 클래스 간 경계에 있는 데이터를 제거하는 방법

Tomek Link로 묶이는 값이 한정적이므로 효과가 크지 않음

Tomek links

분포가 높은 클래스의 중심분포는 어느정도 유지되지만,
여전히 데이터가 삭제되어 정보의 누락 발생!

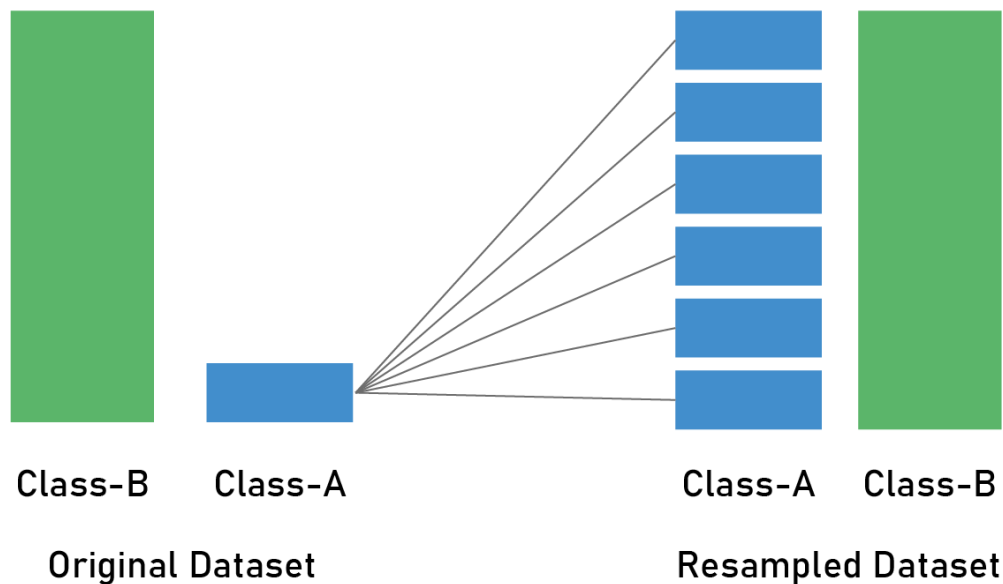


오버 샘플링 이용

샘플링 (Sampling)

오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞추어 **관측치를 증가**시키는 방법



샘플링 (Sampling)

오버 샘플링 (Over Sampling)

소수의 클래스를 다수의 클래스에 맞추어 **관측치를 증가**시키는 방법



데이터 사이즈가 커져 메모리 사용이나 처리 속도에 있어 불리하지만
정보의 손실이 발생하지 않아 일반적으로 언더 샘플링보다 성능이 좋음!

샘플링 (Sampling)

오버 샘플링의 종류

SMOTE (Synthetic Minority Over-sampling Method)

소수 범주의 데이터를 가상으로 만들어내는 방법

Random Over Sampling

랜덤으로 소수 클래스의 데이터를 복제하는 방법

동일한 데이터의 수가 늘어나 **과적합**될 가능성이 큼



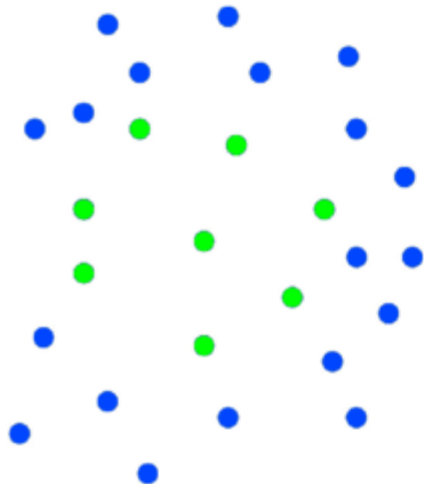
샘플링 (Sampling)

오버 샘플링의 종류

SMOTE (Synthetic Minority Over-sampling Method)

소수 범주의 데이터를 가상으로 만들어내는 방법

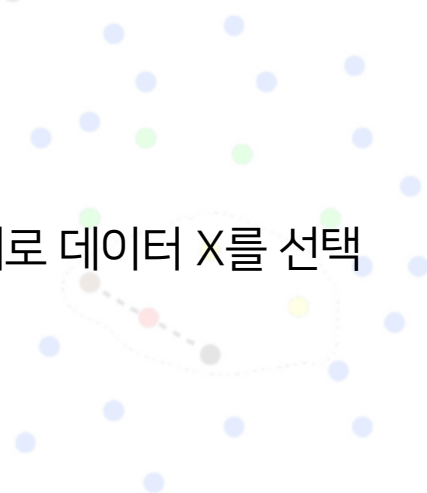
a



b



c



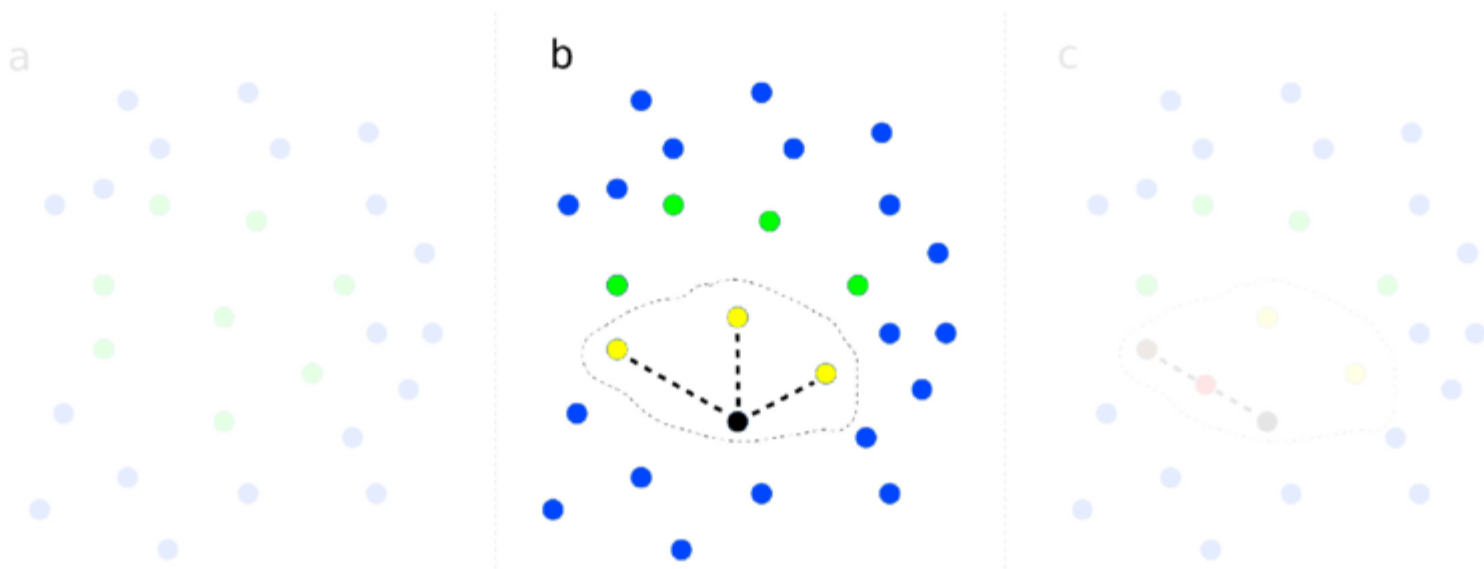
소수 클래스에서 무작위로 데이터 X를 선택

샘플링 (Sampling)

오버 샘플링의 종류

SMOTE (Synthetic Minority Over-sampling Method)

소수 범주의 데이터를 가상으로 만들어내는 방법



X 기준으로 소수 클래스에서 k 개의 가장 가까운 데이터를 탐색 후 선택 (KNN활용)

샘플링 (Sampling)

오버 샘플링의 종류

SMOTE (Synthetic Minority Over-sampling Method)

소수 범주의 데이터를 가상으로 만들어내는 방법



샘플링 (Sampling)

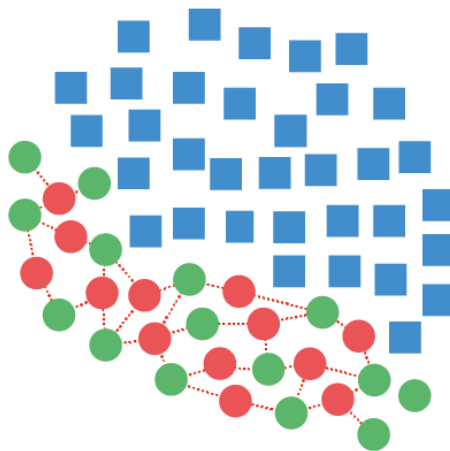
오버 샘플링의 종류

SMOTE (Synthetic Minority Over-sampling Method)

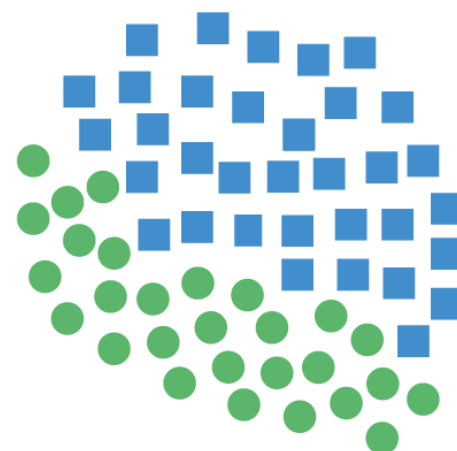
소수 범주의 데이터를 가상으로 만들어내는 방법



Original Dataset



Generating Samples



Resampled Dataset

샘플링 (Sampling)

오버 샘플링의 종류

SMOTE (Synthetic Minority Over-sampling Method)

소수 범주의 데이터를 가상으로 만들어내는 방법



소수 클래스의 데이터 간 거리만 고려하기 때문에
새롭게 생성된 데이터가 다른 클래스의 데이터와 겹치거나
노이즈가 발생할 위험이 있어 **고차원 데이터**에서는 효율적이지 않을 수 있음

4

인코딩

인코딩 (Encoding)

인코딩

범주형 자료를 **수치화**하는 과정

사용자가 입력한 문자나 기호들을 **컴퓨터가 이용할 수 있는 신호**로 변환



범주형 데이터는 주로 문자열, 기호로 표현되어 있으므로
문자열 그대로 모델을 학습시키기 어렵기 때문에
수치적인 값으로 변환하는 인코딩 과정 필요

인코딩 (Encoding)

인코딩의 종류

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	

인코딩의 종류

Classicm	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	Jones Stein	
Hashing	Backward Difference	Estimator	
	Orthogonal Polynomial	Ordered Target	

모든 상황에서 BEST인 인코딩 방식은 없으며,
데이터에 따라 적절한 인코딩 방식을 선택해야 함



One-Hot Encoding

One-Hot Encoding (Dummy Encoding)

데이터에 **가변수**(Dummy Variable)를 추가함으로써
인코딩을 진행하는 기법



가변수 : 설명변수를 0과 1로 변환한 변수

One-Hot Encoding

예시

범주팀 팀원들		가민	수인	윤아	정민	준령
가민		1	0	0	0	0
수인		0	1	0	0	0
윤아		0	0	1	0	0
정민		0	0	0	1	0
준령		0	0	0	0	1

범주 안에 속한 각 수준들의 이름이 가변수의 명칭이 됨

클래스의 개수만큼 열을 추가하여 해당 범주에는 1, 그 외 값에는 0 부여

One-Hot Encoding

예시

범주형 팀원들	가민	수인	윤아	정민	준령
가민	1	0	0	0	0
수인	0	1	0	0	0
윤아	0	0	1	0	0
정민	0	0	0	1	0
준령	0	0	0	0	1



기준이 되는 열 삭제

J개의 수준을 갖는 범주형 변수를 표현하기 위해서는

J-1개의 가변수만으로 충분하기 때문



One-Hot Encoding

트리 기반 모델 인코딩

예시

범주형 변수	범주	수인	수인	문자	문자	준령
가민	0	0	0	0	0	0
수인	0	1	0	0	0	0
윤아	0	0	1	0	0	0
정민	0	0	0	1	0	0
준령	0	0	0	0	1	1

트리 기반 모델은 사용 가능한 모든 부분을 활용해 트리를 생성

삭제된 기준 범주가 트리를 생성하는 데 중요한 요소라면



트리가 잘못 학습될 수 있음



트리 기반 모델의 경우,

삭제하는 가변수 없이 J개의 가변수를 생성하는 것이 좋음!

J개의 수준을 갖는 범주형 변수를 표현하기 위해서는

J-1개의 가변수만으로 충분하기 때문

One-Hot Encoding

장점



해석이 용이함

기준 범주에 대한 정보가 intercept로 존재하기 때문에 기준 범주를 기준으로 모델을 해석할 수 있음



명목형 변수 값을 가장 잘 반영하는 방법



해당 수준에 속하는 경우만 1, 나머지는 0으로 표현되기 때문에
다중공선성 문제를 해결할 수 있음

One-Hot Encoding



장점



고차원 범주형 자료의 경우,
해석이 용이함
One-Hot Encoding은 **너무 많은 가변수를 생성하기 때문에**



기준 범주에 대한 정보가 **데이터의 차원이 늘어나는 문제가 발생** 모델을 해석할 수 있음



명목형 변수 값을 **↓** 잘 반영하는 방법

이는 모델의 학습 속도를 느리게 하며
해당 수준에서 1과 0으로 표현하기 때문에
상당한 computing power를 요구함
다중공선성 문제를 해결할 수 있음

Label Encoding

Label Encoding

명목형 **자료**가 주어졌을 때
범주형 변수의 각 수준에 **점수**를 할당하는 방법



단순히 점수를 할당하므로
각 수준에 부여한 숫자들 사이에는
어떠한 의미나 연관성이 존재하지 않음

Label Encoding

예시

혈액형	점수
A형	1
B형	2
O형	3
AB형	4

위 표는 혈액형이라는 범주형 변수의 각 수준에
1~4 점수를 할당한 것

Label Encoding

예시

혈액형	점수
A형	1
B형	2
O형	3
AB형	4

할당되는 점수는 각 수준을 구분하는 역할일 뿐,

크기나 순서 등의 다른 의미를 가지지 않음

또한 시작점과 간격을 분석자가 설정할 수 있음

Label Encoding

장점

Label Encoding은 One-Hot Encoding과 달리
가변수를 생성하지 않아 차원이 늘어나지 않기 때문에 빨리 학습할 수 있음

단점

할당된 점수에 순서나 연관성이 있다고 잘못 판단해
정보 왜곡이 발생할 수 있음

Ordinal Encoding

Ordinal Encoding

순서형 자료가 주어졌을 때
각 순서에 대응하는 점수를 **차등적으로** 할당하는 방식



Label Encoding과는 달리,
할당된 점수들 간 순서나 연관성 존재

Ordinal Encoding

예시

만족도	점수
매우 별로	1
별로	2
보통	3
좋음	4
매우 좋음	5

만족도의 각 수준에 1~5 점수를 할당

일반적으로 1부터 시작해 **차등적으로** 점수를 부여,
각 점수는 **순서를 반영함**

Ordinal Encoding

장점

Label Encoding처럼 차원이 늘어나지 않으므로
모델이 데이터를 빠르게 처리할 수 있음

단점

범주 내 수준 간 차이를 정확히 반영하기 어려움
이런 차이를 명확히 반영하기 위해서는
해당 데이터에 대한 도메인을 명확히 파악해야 함

"별로"와 "보통"의 차이는 그 이상이어서 3 정도임에도 1로 나타낼 수 있음

Ordinal Encoding



장점

Label Encoding과 마찬가지로
범주형 변수의 **각 수준을 구별하는 것**에 초점을 맞추고 있음

단점

다음으로 소개할 3가지 **Target Encoding**은,
범주형 변수의 각 수준을 구별할 뿐만 아니라
해당 변수와 반응 변수 간의 **수치적인 관계**를 반영해 인코딩을 진행
해당 데이터에 대한 도메인을 명확히 파악해야 함

위 예시에서 각 수준별 차이는 모두 1로 동일하지만, 실제로는 그 차이가 클 수 있음
예를 들어, "별로 " 와 "보통 " 의 차이는 그 이상이어서 3 정도임에도 1로 나타낼 수 있음

Mean Encoding

Mean Encoding

Target Encoding의 일종으로,
범주형 변수의 각 수준에서 도출된 **반응변수의 평균**으로
수준별 점수를 할당하는 인코딩 방식



Mean Encoding

예시

[Y] 키(cm)	[X] 학과	[X] Mean Encoding
168	경영	172
180	경영	172
168	경영	172
174	통계	166
156	통계	166
163	통계	166
171	통계	166
165	경제	171.66
171	경제	171.66
170	경제	171.66

반응변수가 키(cm),
범주형 설명변수가 학과일 때
Mean Encoding을 적용한 예시



학과 변수에는 3개의 수준이 있으며,
각 클래스에 해당하는 평균을 구하면
각각 172, 166, 171.66이 나옴

Mean Encoding

예시

[Y] 키(cm)	[X] 학과	[X] Mean Encoding
168	경영	172
180	경영	172
168	경영	172
174	통계	166
156	통계	166
163	통계	166
171	통계	166
165	경제	171.66
171	경제	171.66
170	경제	171.66

반응변수가 키(cm),
범주형 설명변수가 학과일 때
Mean Encoding을 적용한 예시



학과 변수에는 3개의 수준이 있으며,
각 클래스에 해당하는 평균을 구하면
각각 172, 166, 171.66이 나옴

Mean Encoding

예시

[Y] 키(cm)	[X] 학과	[X] Mean Encoding
168	경영	172
180	경영	172
168	경영	172
174	통계	166
156	통계	166
163	통계	166
171	통계	166
165	경제	171.66
171	경제	171.66
170	경제	171.66

반응변수가 키(cm),
범주형 설명변수가 학과일 때
Mean Encoding을 적용한 예시



이를 통해 학과라는 범주형 변수 대신에
클래스별 반응변수에 대한
평균으로 인코딩을 진행

Mean Encoding

장점



앞선 Encoding 기법들과 달리 설명변수와 반응변수 간의 관계를 고려하여 점수를 할당하였다는 점에서 당위성을 가짐



One-Hot Encoding과 달리 차원이 증가하지 않아 학습 속도가 빠름



Mean Encoding

한계점



Train set에 없던 새로운 수준이 Test set에 등장하면 활용할 수 없음



이상치에 취약함



Data Leakage가 발생하여 모델을 학습시킬 때 과적합이 발생할 위험이 있음



관측치 값이 적은 범주의 경우 모델링에 부정확한 결과가 도출될 가능성이 있음

Mean Encoding

한계점



Train set에 없던 새로운 수준이 Test set에 등장하면 활용할 수 없음



예를 들어 Test set의 학과라는 변수에
'데이터사이언스'라는 새로운 수준이 있다면 점수 할당이 어려움

Mean Encoding

한계점



이상치에 취약함



반응변수가 지나치게 크거나 작을 경우도 포함해 평균을 내기 때문에,
이상치의 영향을 많이 받음

Mean Encoding

한계점



Data Leakage가 발생하여 모델을 학습시킬 때 과적합이 발생할 위험이 있음



설명변수를 인코딩하기 위해 반응변수 값을 활용하기 때문에
Data Leakage가 발생할 수 있음

Data Leakage: 데이터가 누출된 것

즉, 반응변수 Y에 대한 정보가 모델 학습 시 train feature에 들어가는 것

Mean Encoding

한계점



관측치 값이 적은 범주의 경우 모델링에 부정확한 결과가 도출될 가능성이 있음



예를 들어, Train set에서 5개 만으로 진행한 Encoding 값이
Test set에서 관측된 50개의 데이터를 대표한다고 하기 어려움

Leave-One-Out Encoding

Leave-One-Out Encoding (LOO Encoding)

인코딩하고자 하는 현재 행을 제외한 **나머지 행들의 평균**을 구해,
해당 값을 할당하는 방식



현재 행이 이상치라면, 이 값을 제외하고 평균을 구하기 때문에
이상치의 영향을 덜 받게 됨

Leave-One-Out Encoding

예시

[Y] 키(cm)	[X] 학과	[X] LOO Encoding
168	경영	174
180	경영	168
168	경영	174
174	통계	163.33
156	통계	169.33
163	통계	167
171	통계	164.33
165	경제	175
171	경제	167.5
170	경제	172.5

Mean Encoding과 같이
수준별 평균을 구해 점수를 할당하지만,
현재 행을 제외한 평균을 구하므로
같은 수준에 대해
서로 다른 값이 할당될 수 있음

Leave-One-Out Encoding

예시

[Y] 키(cm)	[X] 학과	[X] LOO Encoding
168	경영	174
180	경영	168
168	경영	174
174	통계	163.33
156	통계	169.33
163	통계	167
171	통계	164.33
165	경제	175
171	경제	167.5
170	경제	172.5

경영학과에서 키가 168인 학우의 경우,
자신을 제외한 나머지 2개 행의
반응변수 값의 평균인 174를 할당

Leave-One-Out Encoding

예시

[Y] 키(cm)	[X] 학과	[X] LOO Encoding
168	경영	174
180	경영	168
168	경영	174
174	통계	163.33
156	통계	169.33
163	통계	167
171	통계	164.33
165	경제	175
171	경제	167.5
170	경제	172.5

키가 180인 학우 역시
자신을 제외한 평균(168)을 할당

Leave-One-Out Encoding

예시

[Y] 키(cm)	[X] 학과	[X] LOO Encoding
168	경영	174
180	경영	168
168	경영	174
174	통계	163.33
156	통계	169.33
163	통계	167
171	통계	164.33
165	경제	175
171	경제	167.5
170	경제	172.5

키가 180인 학우 역시
자신을 제외한 평균(168)을 할당



Mean Encoding과 달리
같은 범주이더라도
점수가 제각각 다름

Leave-One-Out Encoding

장점



이상치의 영향을 덜 받음

앞선 Mean Encoding과 같은 장점을 가지면서, 한계점을 일부 극복함



모든 반응변수의 정보가 다 반영되지 않기 때문에
과적합의 가능성이 Mean Encoding보다 낮음

Leave-One-Out Encoding

한계점



Train set에 없던 새로운 수준이 Test set에 등장하면 활용할 수 없음



이상치에 취약함



Data Leakage가 발생하여 모델을 학습시킬 때 과적합이 발생할 위험이 있음



관측치 값이 적은 범주의 경우 모델링에 부정확한 결과가 도출될 가능성이 있음

Leave-One-Out Encoding

한계점



Train set에 없던 새로운 수준이 Test set에 등장하면 활용할 수 없음



이상치에 취약함

Mean Encoding과 동일한 한계점을 지님



Data Leakage가 발생하여 모델을 학습시킬 때 과적합이 발생할 위험이 있음



관측치 값이 적은 범주의 경우 모델링에 부정확한 결과가 도출될 가능성이 있음

Ordered Target Encoding

Ordered Target Encoding (CatBoost Encoding)

같은 수준에 속한 행들 중
현재 행 **이전 행 값들의 평균**을 점수로 할당하는 방식



범주형 변수가 많은 데이터 처리에 유용한 부스팅 모델인
CatBoost에서 사용하는 방식으로 CatBoost Encoding이라고도 불림

Ordered Target Encoding

예시

[Y] 키(cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	169.5
174	통계	166	169.5
165	경제	171.66	169.5
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	171.66	165
170	경제	171.66	172.5
168	경영	172	174
171	통계	166	164.33

각 수준의 첫번째 행은
같은 수준에 해당하는
이전 값이 없으므로
전체 평균을 구해 할당

Ordered Target Encoding

예시

[Y] 키(cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	169.5
174	통계	166	169.5
165	경제	171.66	169.5
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	171.66	165
170	경제	171.66	172.5
168	경영	172	174
171	통계	166	164.33

키 174인 통계학과 학우는
모든 수준을 통합해 구한
전체 평균 169.5를 할당

각 수준의 첫번째 행은
같은 수준에 해당하는
이전 값이 없으므로
전체 평균을 구해 할당

Ordered Target Encoding

예시

[Y] 키(cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	169.5
174	통계	166	169.5
165	경제	171.66	169.5
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	171.66	165
170	경제	171.66	172.5
168	경영	172	174
171	통계	166	164.33

두번째 통계학과 행에는
유일한 이전 행인 첫 번째 통계학과
반응변수 값 174를 할당

Ordered Target Encoding

예시

[Y] 키(cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	169.5
174	통계	166	169.5
165	경제	171.66	169.5
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	171.66	165
170	경제	171.66	172.5
168	경영	172	174
171	통계	166	164.33

세번째 통계학과 행은
 앞선 두 통계학과 학우들의
 키 평균을 구해 $\frac{174+156}{2} = 165$ 할당

Ordered Target Encoding

예시

[Y] 키(cm)	[X] 학과	Mean Encoding	Ordered Target Encoding
168	경영	172	169.5
174	통계	166	169.5
165	경제	171.66	169.5
156	통계	166	174
180	경영	172	168
163	통계	166	165
180	경제	171.66	165
170	경제	171.66	172.5
168	경영	172	174
171	통계	166	164.33

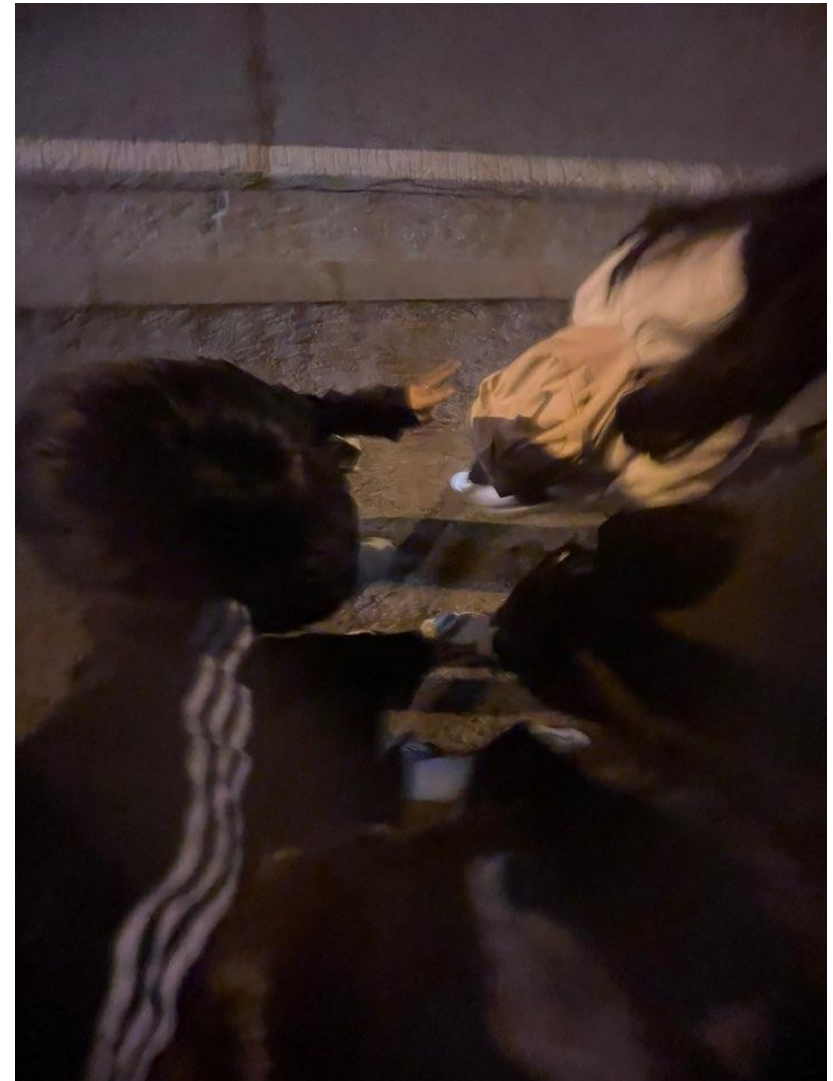
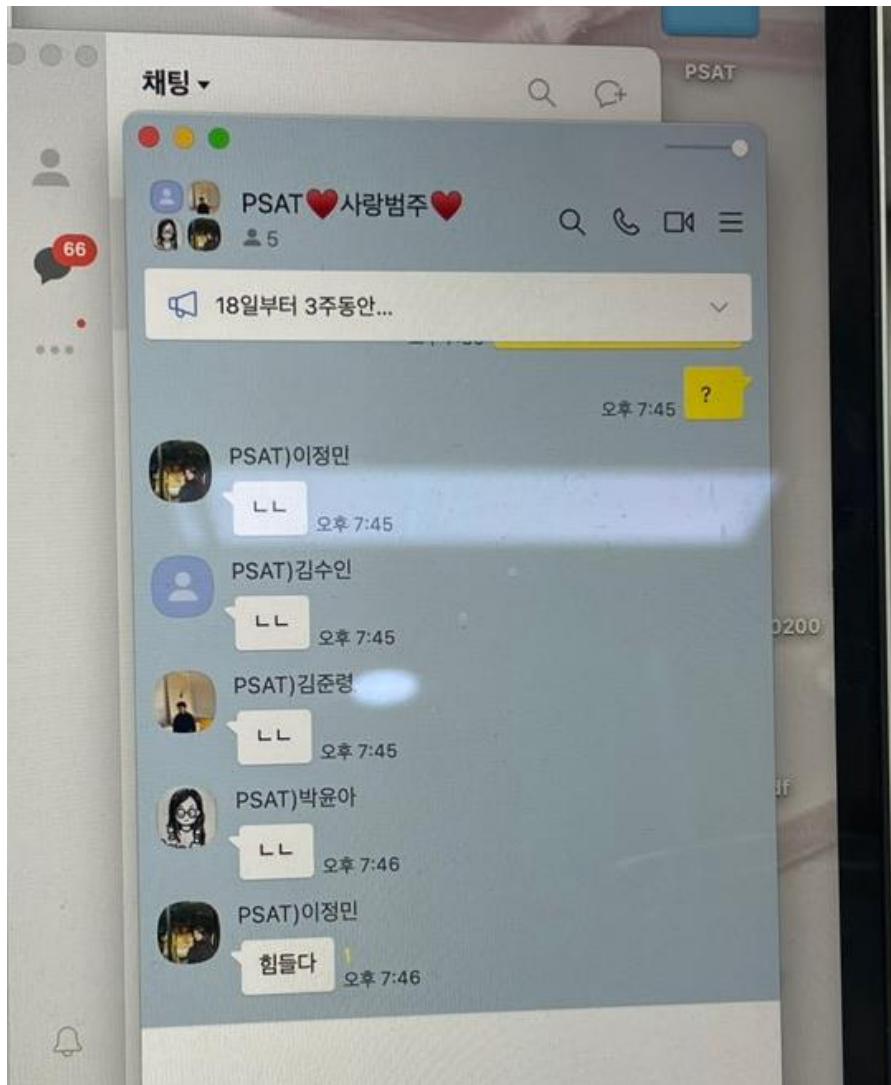


Mean Encoding과 달리
Ordered Target Encoding은
같은 수준에 해당하는 행이라도
다른 값이 할당될 수 있음
즉, 각 행에 더 다양한 값이 할당됨



퀸가민님의 은총 (제사범주 아님)







THANK YOU

