

범주형자료분석팀

2팀
조장희
위재성
김지현
조수미
송지현
김민지

INDEX

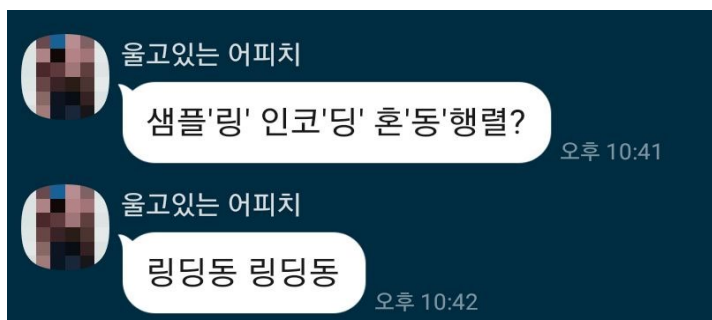
1. 혼동행렬

2. ROC 곡선

3. 샘플링

4. 인코딩

오늘 배울 내용은 바로~~~
링딩동 링딩동
샘플링!! 인코딩!! 혼동!!행렬



아이디어 주신 익명의 어피치님께
감사의 말씀을,,

1

혼동행렬

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 (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 (Confusion Matrix)

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

EX) TP(True Positive):

코로나 양성이라고 예측된 환자가 -> 코로나 양성인 경우 (맞춤)

FN(False Negative):

코로나가 음성이라고 예측된 환자가 -> 코로나 양성인 경우 (틀림)



범주형 자료분석 <혼동행렬의 한계>

1. 정보의 손실 발생

모형은 예측확률 $\hat{\pi}$ 을 연속적인 값으로 반환하지만,

예측은 cut-off point를 기준으로

이항변수(0 또는 1)로 범주화하기 때문

예측값 (\hat{Y})

$\hat{Y} = 0$

FN

TN

2. cut-off point의 선택이 임의적

EX) TP(True Positive):

Cut-off point가 달라지면 혼동행렬도 달라짐 (맞춤)

FN(False Negative):

클래스 불균형이 심한 경우 혼동행렬이 크게 바뀜

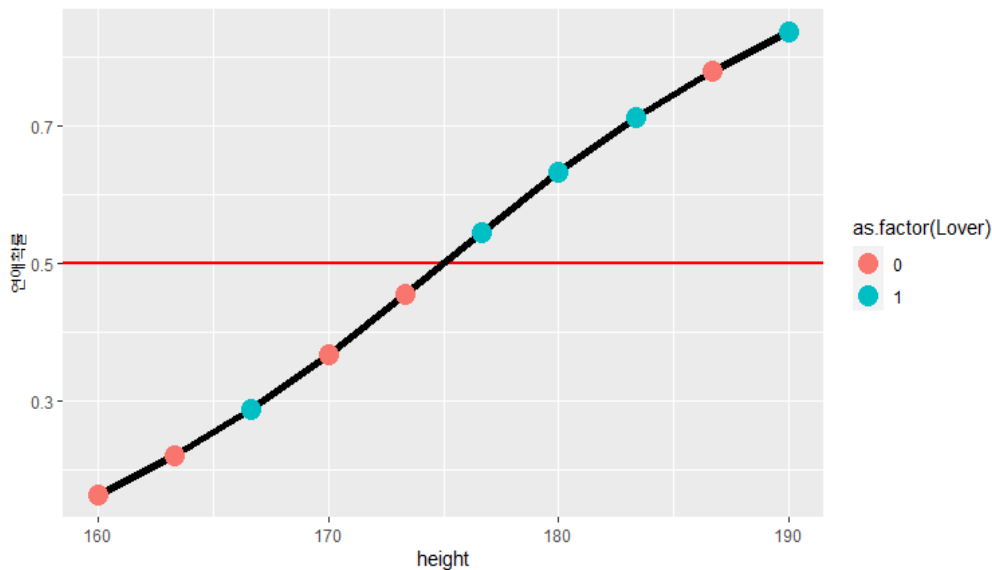
코로나가 음성이라고 예측된 환자가 -> 코로나 양성인 경우 (틀림)

(한계점 보완하는 방법은 ROC에서 배울 예정 ㅎㅎ)

- Cut-off point에 의존적

EX) 10명의 키에 따른 연애 여부 예측

<Cut-off point = 0.5>

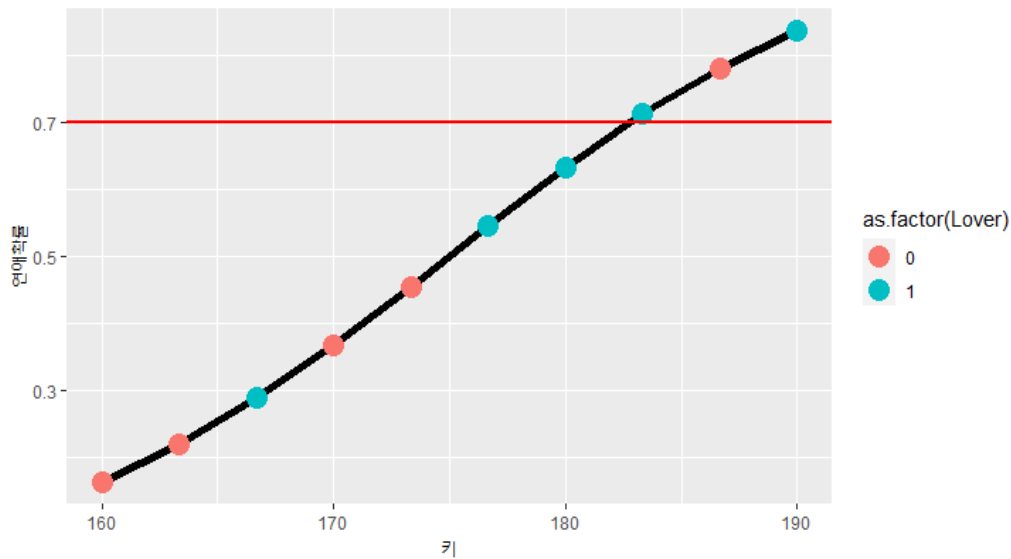


Cut off = 0.5		실제 연애(Y)	
		Y = 1	Y = 0
연애 예측 (Ŷ)	Ŷ = 1	4	1
	Ŷ = 0	1	4

- Cut-off point에 의존적

EX) 10명의 키에 따른 연애 여부 예측

<Cut-off point = 0.7>

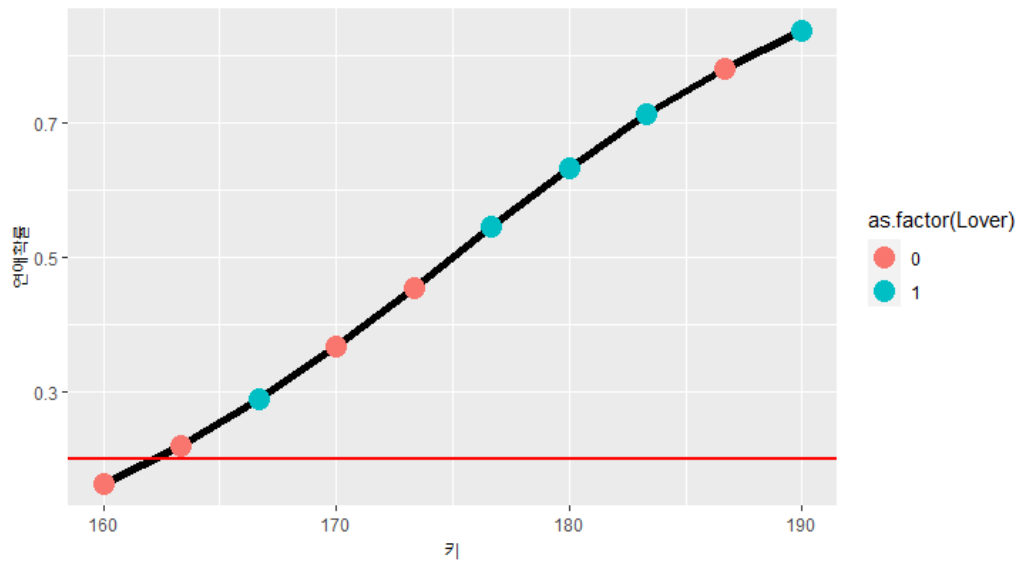


Cut off = 0.7		실제 연애(Y)	
		Y = 1	Y = 0
연애 예측 (\hat{Y})	$\hat{Y} = 1$	2	1
	$\hat{Y} = 0$	3	4

- Cut-off point에 의존적

EX) 10명의 키에 따른 연애 여부 예측

<Cut-off point = 0.2>



Cut off = 0.2		실제 연애(Y)	
		Y = 1	Y = 0
연애 예측 (Ŷ)	Ŷ = 1	5	4
	Ŷ = 0	0	1

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

• unbalanced data일 때 해당 범주에 지나치게 의존하여 문제발생

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
- unbalanced data일 때 해당 범주에 지나치게 의존하여 문제발생

2. 정밀도(Precision/PPV/Positive Predictive Value)

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

맞다고 예측한 것 중 실제로 맞는 것이 얼마인지를 비율로 나타내는 지표

즉, $\hat{Y} = 1$ 이라고 말했던 것 중에서 실제로 $Y = 1$ 인 경우의 비율

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

- FP가 더 critical한 경우에 사용
- unbalanced data일 때 특정 범주에 대한 의존성을 줄여주는 장점

2. 정밀도(Precision/PPV/Positive Predictive Value)

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

맞다고 예측한 것 중 실제로 맞는 것이 얼마인지를 비율로 나타내는 지표

즉, $\hat{Y} = 1$ 이라고 말했던 것 중에서 실제로 $Y = 1$ 인 경우의 비율

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

- FP가 더 critical한 경우에 사용
- unbalanced data일 때 특정 범주에 대한 의존성을 줄여주는 장점

3. 민감도(Sensitivity/TPR/True Positive Rate)

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

실제로 맞는 것 중에서 맞다고 예측한 것이 얼마인지를 비율로 나타내는 지표

즉, 실제로 $Y = 1$ 인 것 중에서 $\hat{Y} = 1$ 이라고 예측한 것의 비율

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

- FN이 더 critical한 경우에 사용

- unbalanced data일 때 특정 범주에 대한 의존성을 줄여주는 장점

3. 민감도(Sensitivity/TPR/True Positive Rate)

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

실제로 맞는 것 중에서 맞다고 예측한 것이 얼마인지를 비율로 나타내는 지표

즉, 실제로 $Y = 1$ 인 것 중에서 $\hat{Y} = 1$ 이라고 예측한 것의 비율

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

- FN이 더 critical한 경우에 사용
- unbalanced data일 때 특정 범주에 대한 의존성을 줄여주는 장점

정밀도와 민감도



정밀도

Ex) 재판
죄 있는 사람을
무죄라고 하는 것(FN)은
흔히 발생하지만,

결백한 사람을 유죄로
판결하는 경우(FP)는
심각한 문제를 초래

FP가 더 치명적일 때
정밀도를 사용



민감도

Ex) 코로나
코로나에 안 걸린 사람을
걸렸다고 하는 것(FP)는
다시 검사를 하면 되지만,

코로나에 걸린 사람을
안 걸렸다고 하는 것(FN)은
방역에 큰 악영향을 미침

FN이 더 치명적일 때
민감도를 사용

4. 특이도(Specificity/TNR/True Negative Rate)

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

실제로 아닌 것 중 아니라고 예측한 것이 얼마인지 비율로 나타낸 지표

즉, $Y = 0$ 인 것 중에서 $\hat{Y} = 0$ 이라고 예측한 것의 비율

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

• 1에 가까울수록 좋음

4. 특이도(Specificity/TNR/True Negative Rate)

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

실제로 아닌 것 중 아니라고 예측한 것이 얼마인지 비율로 나타낸 지표

즉, $Y = 0$ 인 것 중에서 $\hat{Y} = 0$ 이라고 예측한 것의 비율

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

- 1에 가까울수록 좋음

FPR (False Positive Rate)

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

특이도의 반대 경우!
실제로 아닌 것 중에 맞다고 예측한 것의 비율

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

- 0에 가까울수록 좋음

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

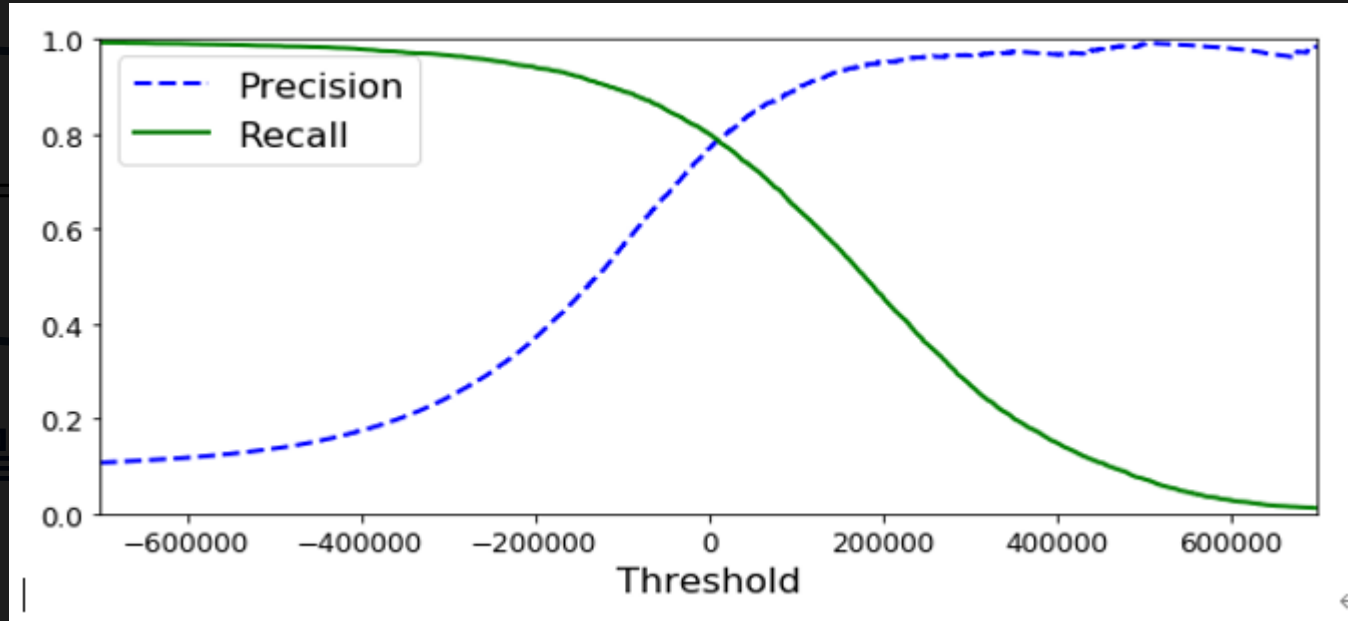
		관측값 (Y)		
		Y = 1	Y = 0	
예측값 (\hat{Y})	$\hat{Y} = 1$	TP	FP	Precision
	$\hat{Y} = 0$	FN	TN	

Recall

TN을 고려하지 않는다는 단점

Precision과 Recall의 관계

5. F1-Score

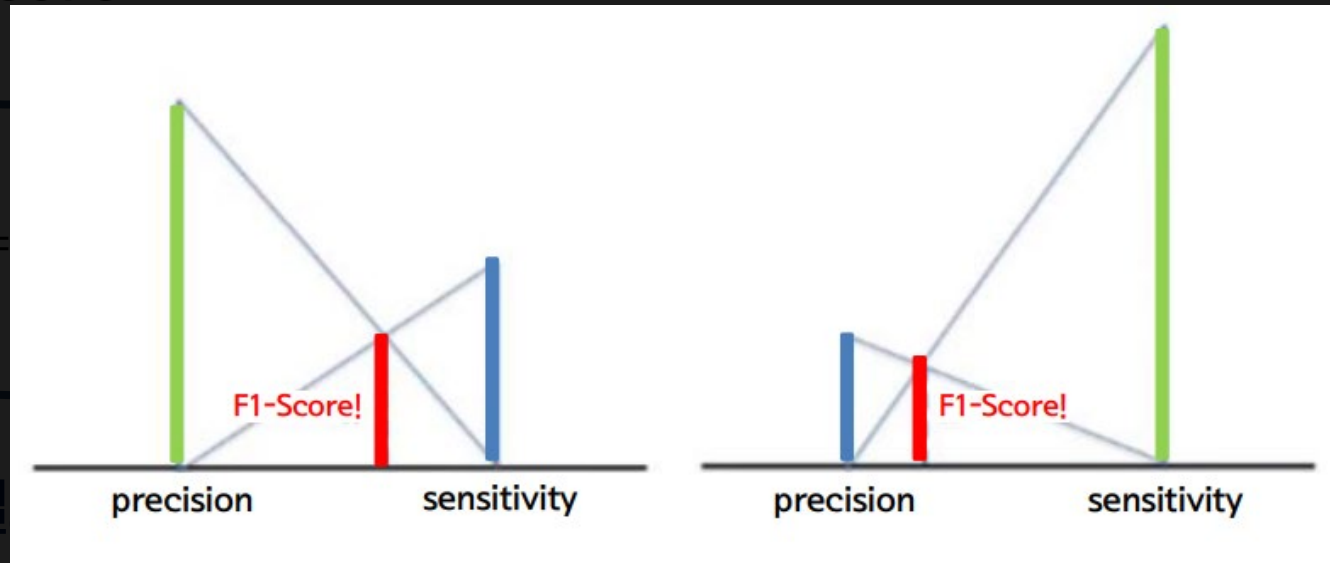


정밀도와 민감도는 trade-off (상충관계)이다.
 즉, 정밀도가 커지면 민감도는 작아진다.

Sensitivity

Q. 왜 산술평균이 아닌 조화평균을 구하는가?

5. F1-Score



관측값 (Y)

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

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

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

Sensitivity

6. MCC (Matthews correlation coefficient, 매튜 상관계수/파이계수)

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

상관계수이므로 -1과 1 사이의 값을 갖음.

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

- 1에 가까울수록 **완전 예측**
- -1에 가까울수록 **완전 역예측**
- 0에 가까울수록 **랜덤 예측**

혼동행렬의 모든 부분을 사용 => F1-Score의 단점을 보완

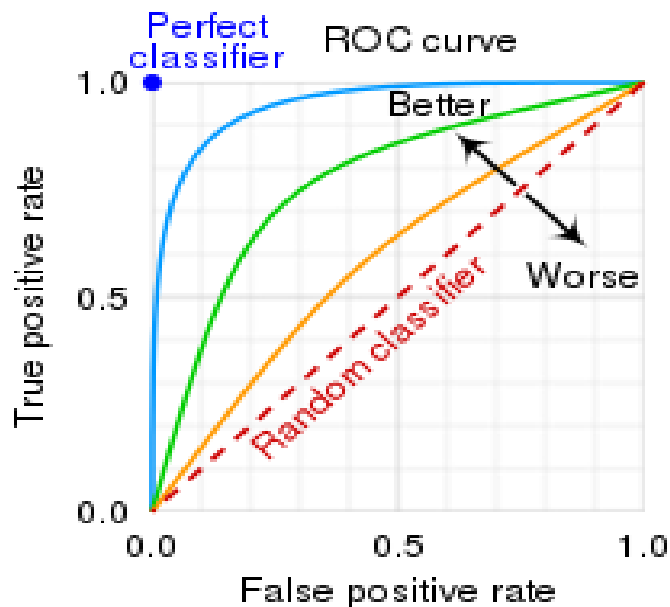
2

ROC & AUC

ROC Curve

ROC 곡선이란?

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



2

ROC & AUC



Confusion
Matrix의 한계



ROC curve로 해결!



Cut-off point에
의존적



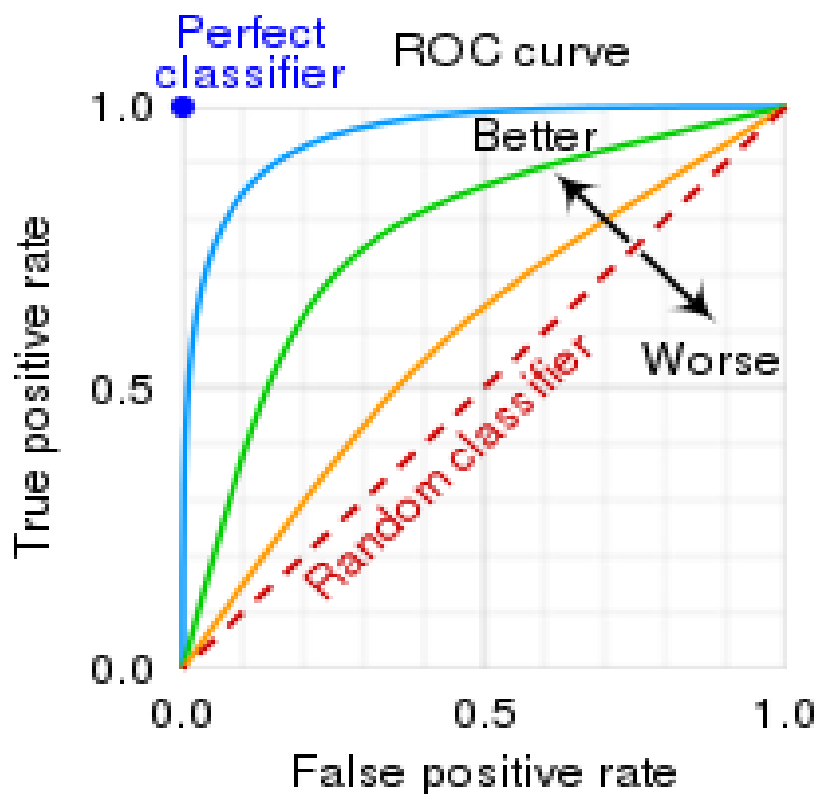
모든 cut-off point를
고려할 수 있다

정보의 손실



모든 예측 검정력을 구하기에
정보의 손실이 적다

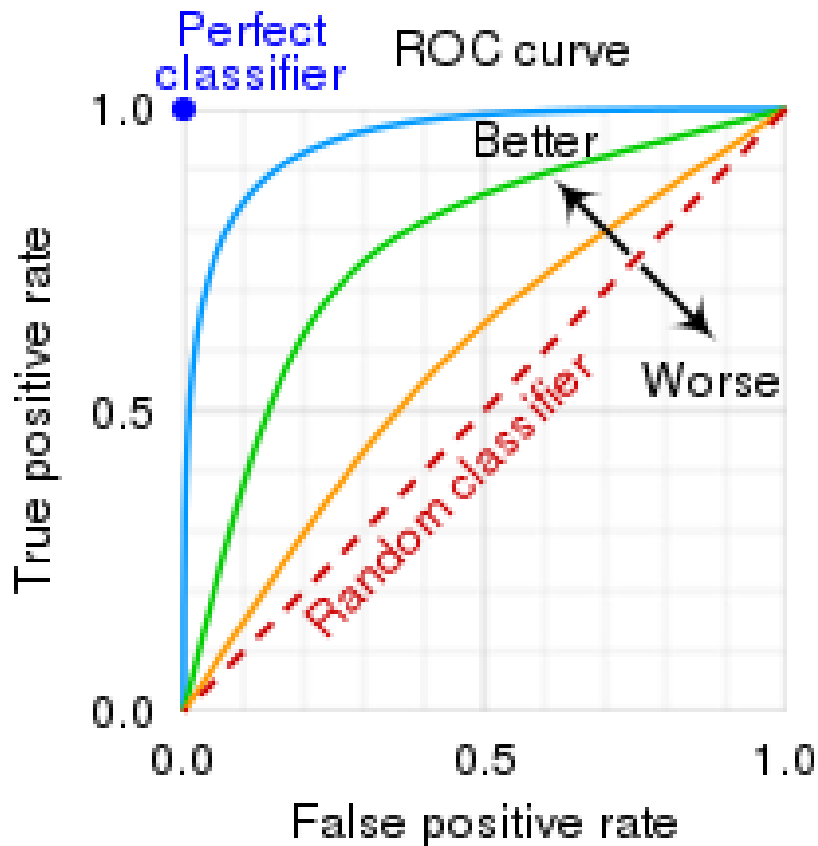
ROC Curve의 형태



우상향하는 위로 볼록한 곡선

- X축: FPR(1-특이도)
- Y축 : TPR(민감도)
- X는 작을수록, Y는 클수록 좋음
- $0 \leq X, Y \leq 1$

ROC Curve의 형태



Cuf-off point가

0에 가까워 질수록 $\rightarrow (1,1)$

1에 가까워 질수록 $\rightarrow (0,0)$

...why?

Cutoff point가 0에 가까워 질 때 ROC Curve의 형태 (기준점이 낮아짐)

Cutoff point
 ≈ 0

대부분 $\hat{Y}=1$
로 예측

TP, FP 증가
TN, FN 감소

FPR ≈ 1
TPR ≈ 1



Cutoff point가 1에 가까워 질 때

ROC Curve의 형태 (기준점이 높아짐)

Cutoff point
 ≈ 1 대부분 $\hat{Y} = 0$
으로 예측TP, FP 감소
TN, FN 증가FPR ≈ 0
TPR ≈ 0 

0에 가까운 질수록

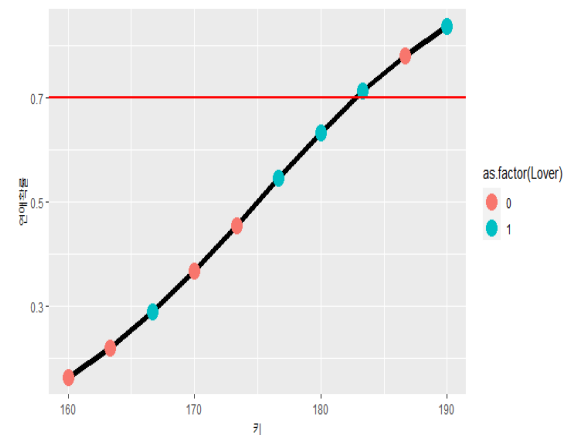
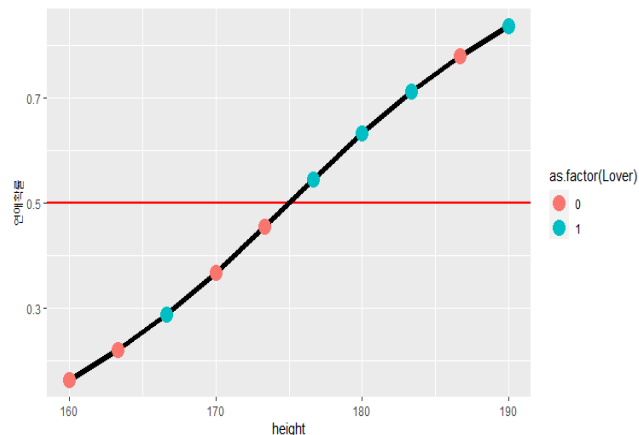
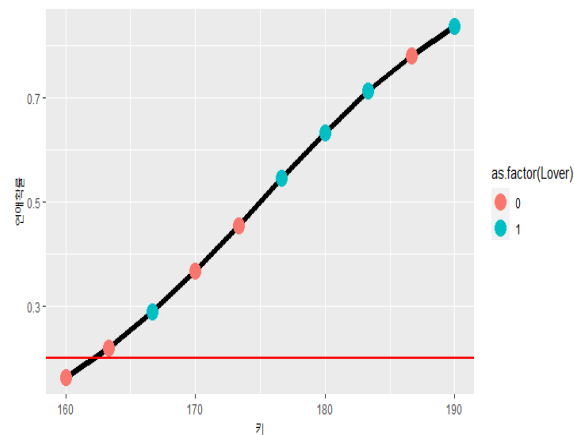
질수록

ROC 곡선으로 적합한 Cutoff point 찾기

STEP 1

모든 Cut off point 에 대한 Confusion Matrix 생성

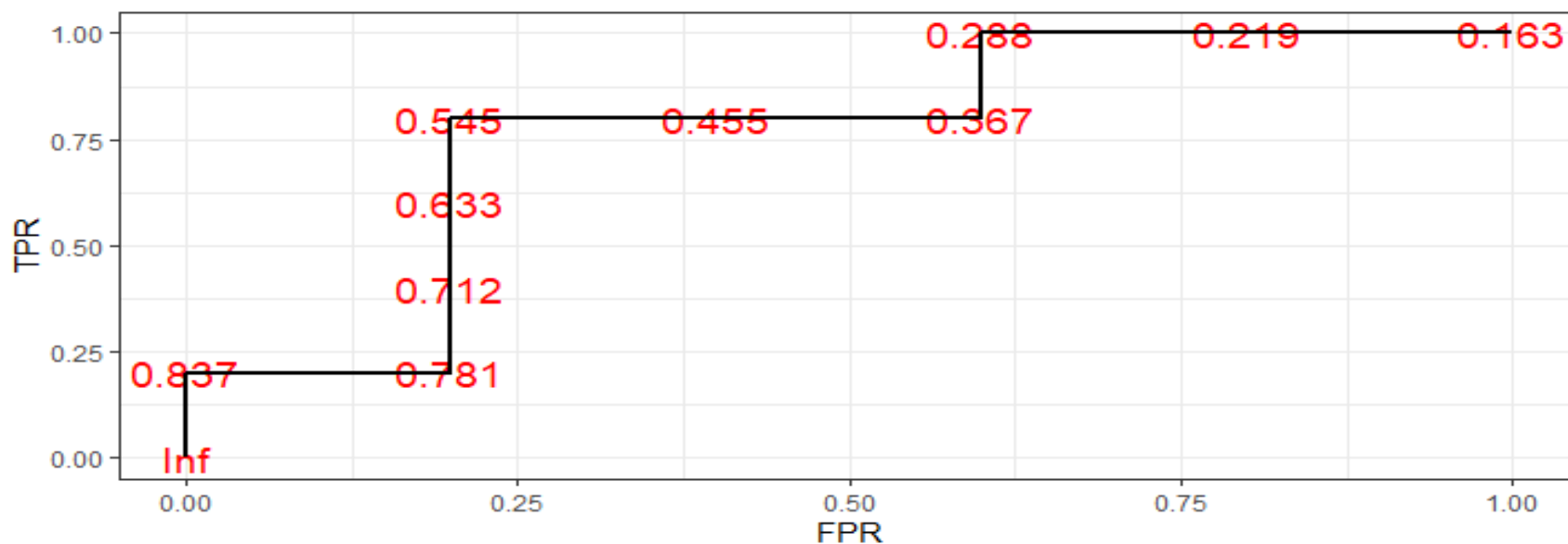
Cutoff가 달라질 때마다 예측이 다르게 됨 > 각각 다른 TPR & FPR 값



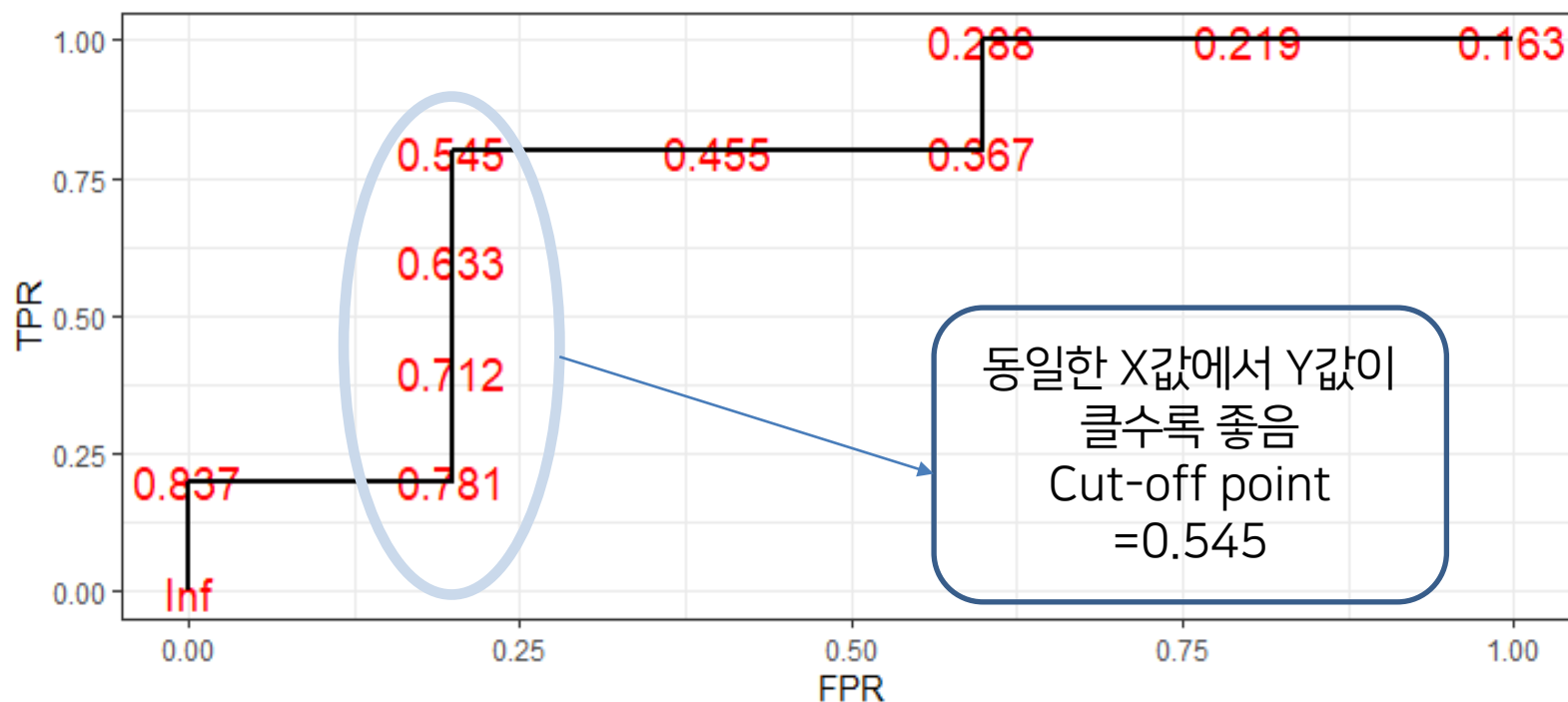
ROC 곡선으로 적합한 Cutoff point 찾기

STEP 2

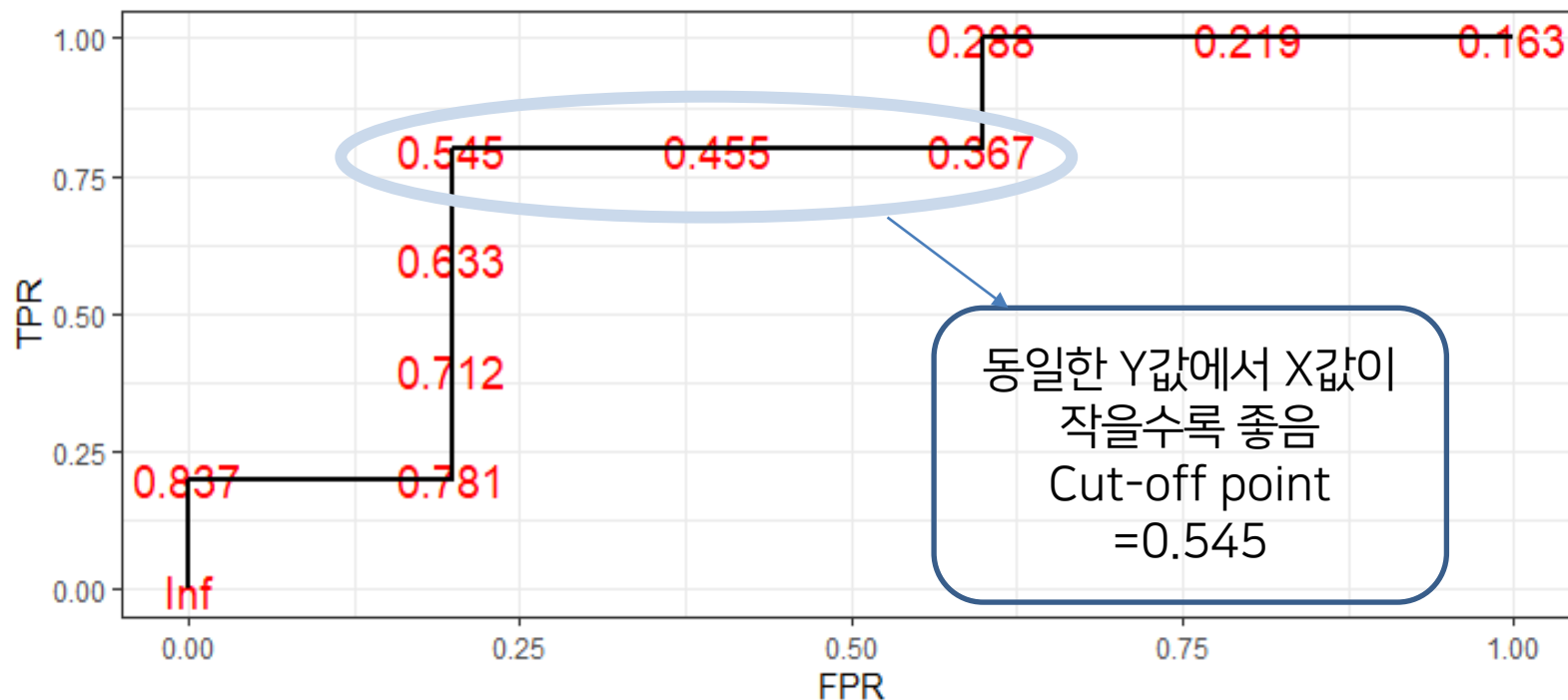
각각 다른 TPR & FPR 값으로 ROC 곡선 그리기



ROC 곡선으로 적합한 Cutoff point 찾기



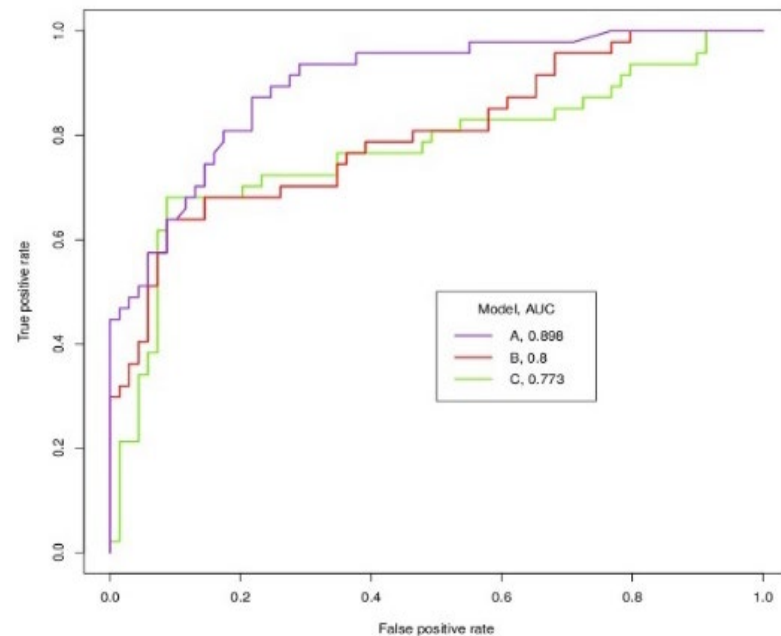
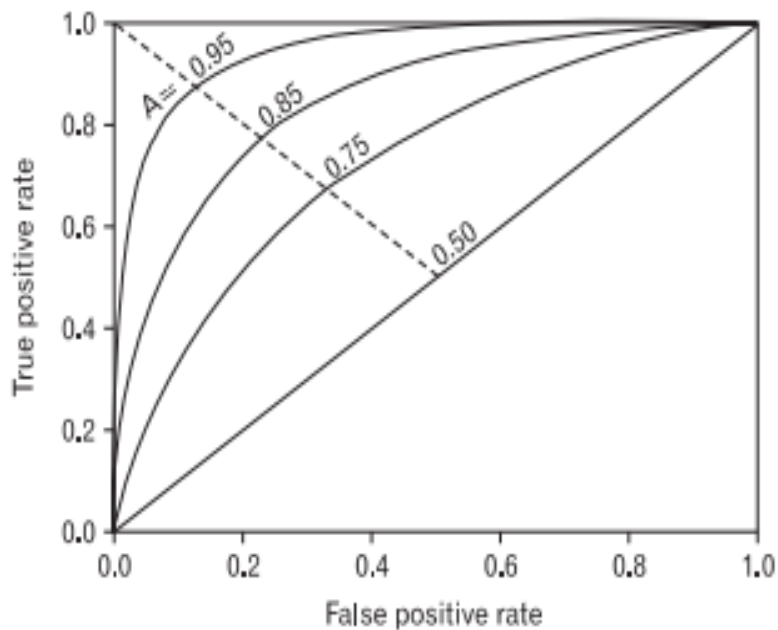
ROC 곡선으로 적합한 Cutoff point 찾기



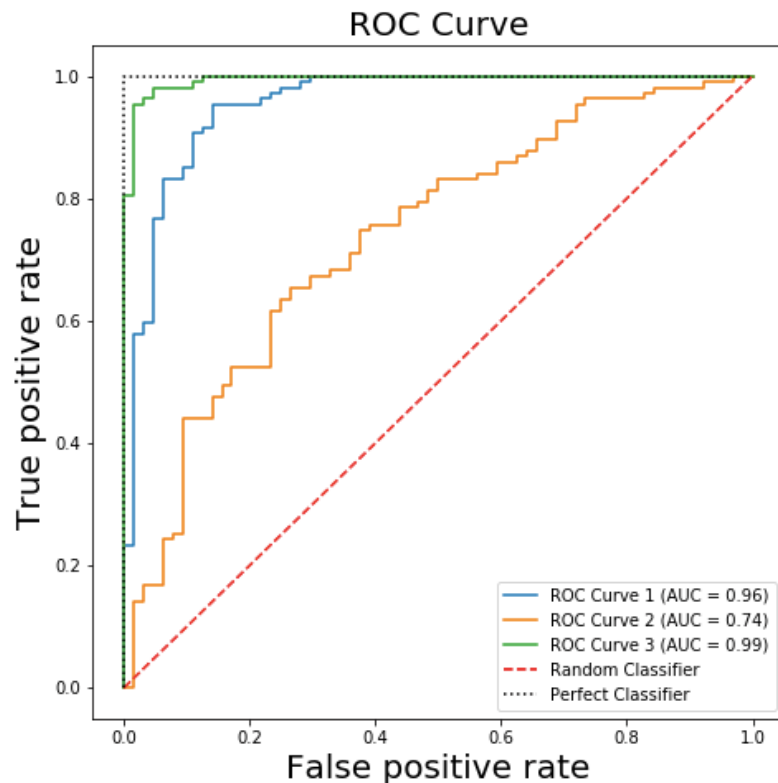
AUC란?

Area Under the Curve

: ROC curve 아래의 면적

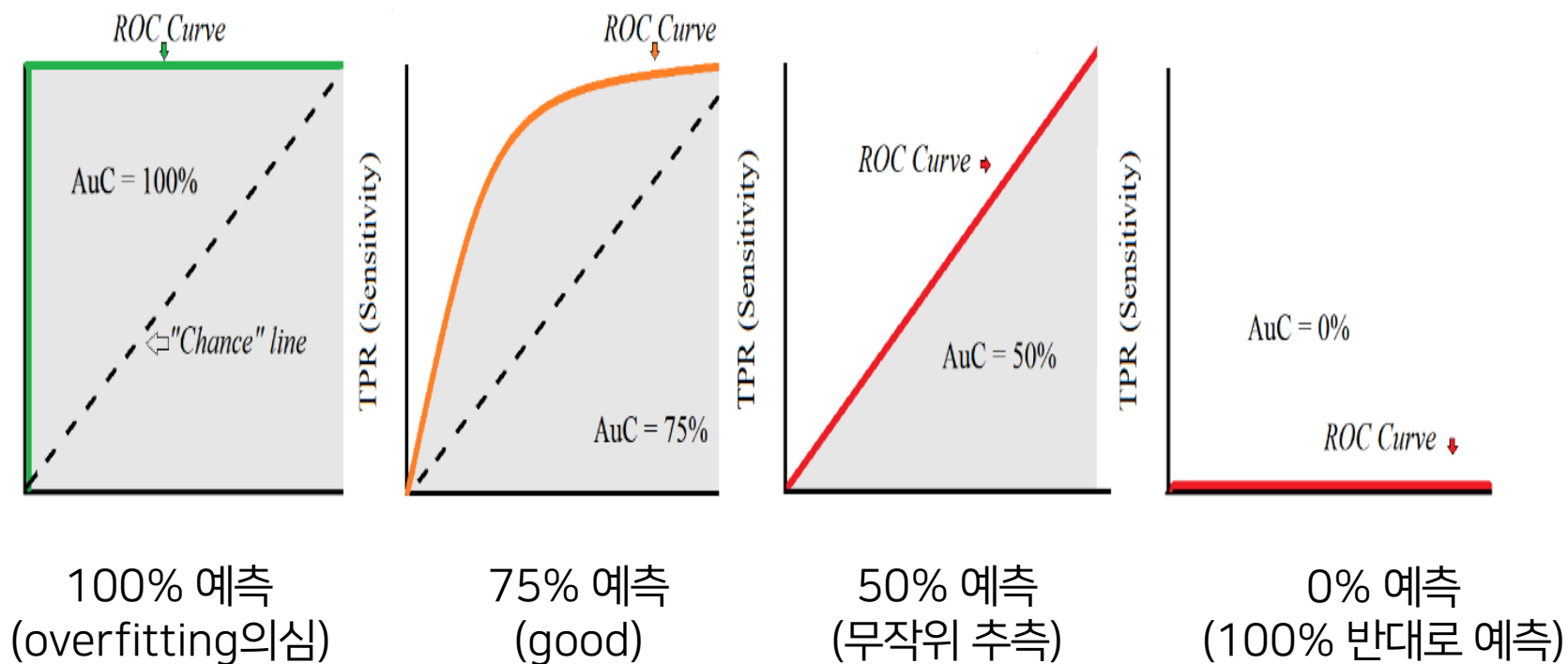


AUC의 특징



- 모델의 성능을 비교하는 지표
- Cut-off point와 상관없이 모델의 성능 측정 가능
- $0 \leq \text{AUC} \leq 1$
- AUC가 1에 가까워질 수록 모델의 성능이 좋음

AUC의 특징



3

샘플링

클래스 불균형

각 클래스(=수준)가 갖고 있는 데이터의 양에
큰 차이가 있는 경우

※Unbalanced data와 같은 의미

		관측값(Y)	
		Y = 1	Y = 0
예측값(\hat{Y})	$\hat{Y} = 1$	93	4
	$\hat{Y} = 0$	2	1

		관측값(Y)	
		Y = 1	Y = 0
예측값(\hat{Y})	$\hat{Y} = 1$	60	2
	$\hat{Y} = 0$	35	3

클래스 불균형인 경우

정확도가 양이 더 많은 클래스에 의존적이게 되는 문제 발생

클래스 불균형

		관측값(Y)	
		Y = 1	Y = 0
예측값(\hat{Y})	$\hat{Y} = 1$	93	4
	$\hat{Y} = 0$	2	1

		관측값(Y)	
		Y = 1	Y = 0
예측값(\hat{Y})	$\hat{Y} = 1$	60	2
	$\hat{Y} = 0$	35	3

Y = 1이 95개, Y = 0이 5개로 클래스 불균형을 갖는 반응변수 Y가 있다.

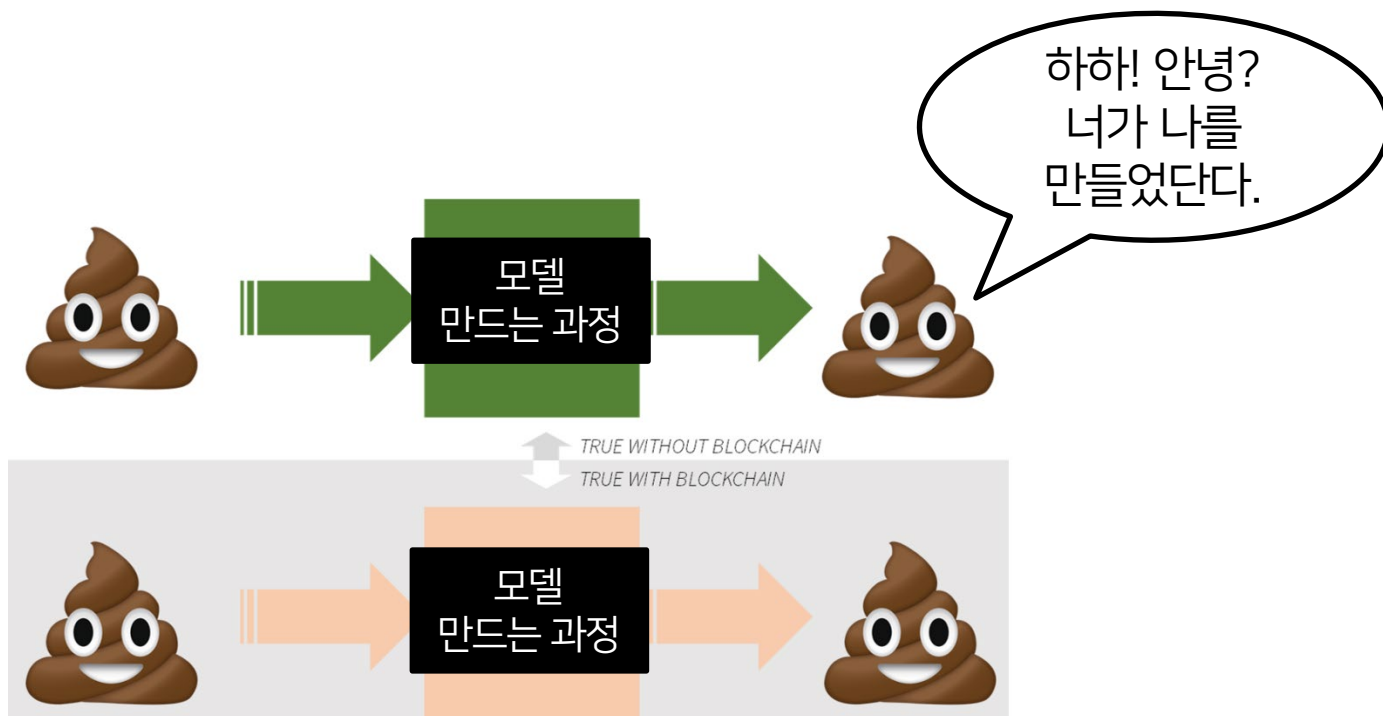
$$\text{왼쪽 Accuracy} = \frac{93+2}{93+4+2+1} = 0.94,$$

$$\text{오른쪽 Accuracy} = \frac{60+3}{60+2+35+3} = 0.63$$

왼쪽의 정확도가 훨씬 높게 나와

모델의 성능을 판별하기 어렵게 됨!

샘플링



GIGO (Garbage In, Garbage Out)이라는 말이 있듯이

좋은 모델에는 좋은 train set이 필요한 법!

샘플링

언더 샘플링
(Under Sampling)

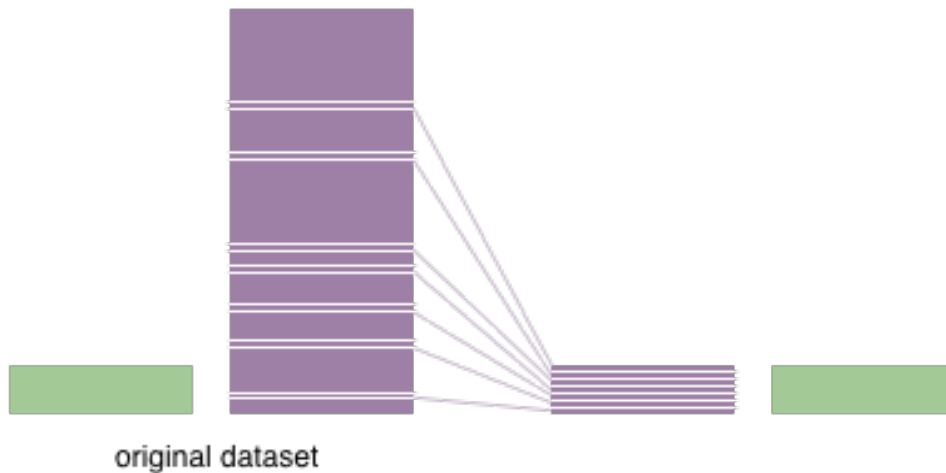
오버 샘플링
(Over Sampling)

샘플링

언더 샘플링 (Under Sampling)

소수 클래스는 그대로 두고,
다수 클래스의 데이터를 소수 클래스에 맞추어 감소시키는 방법

under-sampling



언더 샘플링의 장단점

샘플링

언더 샘플링(under-sampling) **장점** : 데이터 사이즈 감소로 메모리 사용이나 처리 속도 측면에서 유리

소수 클래스는 그대로 두고, 다수 클래스의 데이터를 소수 클래스에 맞추어 감소시키는 방법

단점 : 관측치 손실로 인한 정보 누락

under-sampling



original dataset

샘플링

언더 샘플링 (Under Sampling)

종류	특징
Random Under Sampling	랜덤으로 다수 클래스 데이터 제거
Condensed Nearest Neighbors (CNN)	<ul style="list-style-type: none">- KNN 알고리즘 적용- 다수 클래스 데이터가 밀집된 곳을 위주로 제거- 어떤 데이터를 지울지는 아직 랜덤
Tomek Links	<ul style="list-style-type: none">- 클래스 경계의 노이즈 데이터만 제거- 주로 다른 기법과 결합해서 사용
Edited Nearest Neighbors (ENN)	
One-sided selection (OSS)	Tomek Links + CNN
Neighborhood Cleaning Rule (NCR)	ENN + CNN

샘플링

오버 샘플링 (Over Sampling)

소수클래스의 데이터를 다수 클래스에 맞추어 증가시키는 방법



오버 샘플링의 장단점

샘플링

장점: 정보손실이 없어 언더 샘플링에 비해 성능이 좋음

소수클래스의 데이터를 다수 클래스에 맞추어 증가시키는 방법

단점: 관측치 수 증가로 메모리 사용이나 처리속도 측면에서 상대적으로 불리함

over-sa



original dataset

샘플링

오버 샘플링 (Over Sampling)

1. <Random Over Sampling>

가장 기본적인 방법

무작위로 소수 클래스의 데이터를 복사해 늘린다.

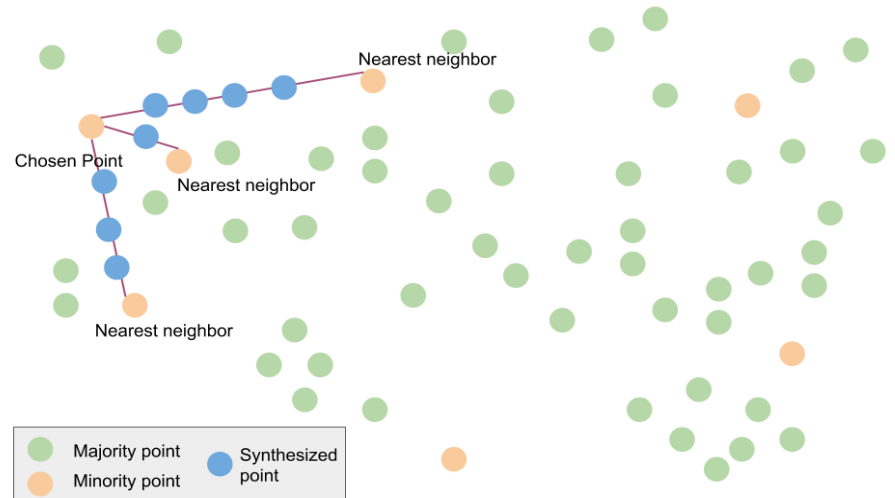
장점: 정보의 손실 X

단점: 임의로 데이터를 복제하기 때문에 과적합될 가능성 O

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.

SMOTE의 장단점

샘플링

오버 샘플링(Over Sampling)

장점 : 단순 랜덤이 아닌 KNN 알고리즘으로 가상의 데이터 생성

SMOTE

=> Random Over Sampling보다 과적합이 발생할 가능성이 낮음.

① 소수 클래스의 데이터 중 랜덤으로 하나를 선택한다.

② 선택한 데이터와 가장 가까운 k 개의 소수 클래스 데이터를 선택한다.

단점 : 새로운 소수의 가상 데이터 생성과정에서 인접한 "다수 데이터의 위치" 고려X

③ 처음에 선택한 데이터와 무작위로 선택한 데이터 사이에 직선을 그리고,

=> 서로 다른 클래스의 데이터가 겹치거나 노이즈 생성될 가능성 높음.

그 직선 상에 가상의 소수 클래스 데이터를 생성한다.

=> 고차원 데이터에서는 효율적이지 않다!

샘플링

오버 샘플링 (Over Sampling)

※ SMOTE를 확장 및 수정한 방법

종류	특징
Borderline-SMOTE	<ul style="list-style-type: none">- 클래스 경계 근처에만 소수클래스 데이터 생성<ul style="list-style-type: none">- 경계 설정에서 KNN 알고리즘 사용
Borderline-SMOTE SVM (SVMSMOTE)	<ul style="list-style-type: none">- 클래스 경계 근처에만 소수클래스 데이터 생성<ul style="list-style-type: none">- 경계 설정에서 SVM 알고리즘 사용
ADASYN	<ul style="list-style-type: none">- 소수 클래스의 밀도 분포를 고려해서 데이터 생성- 밀도에 가중치를 두어 밀도가 낮으면 많이, 높으면 조금 생성

4

인코딩

인코딩

범주형 자료를 수치화

모델 학습을 위해 필요한 과정

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	

One-Hot-Encoding

가변수(Dummy variable)를 만들어주는 인코딩 방법

범주	제니 (기준범주)	로제	리사	지수
제니	1	0	0	0
로제	0	1	0	0
리사	0	0	1	0
지수	0	0	0	1



범주	로제	리사	지수
제니	0	0	0
로제	1	0	0
리사	0	1	0
지수	0	0	1

<기존의 변수>

자기 자신의 범주 -> 1
그 외 범주 -> 0
기준이 되는 범주의 열을 삭제

<One-Hot Encoding>

One-Hot-Encoding

범주	로제	리사	지수
제니	0	0	0
로제	1	0	0
리사	0	1	0
지수	0	0	1

<장점>

1. 해석이 용이
2. 명목형 변수 값 잘 반영
3. 다중공선성 해결

One-Hot-Encoding

But, 범주형 변수가 너무 많을 경우 수많은 가변수 생성 -> 차원 ↑ 문제점

차농남 팀장의 사심,,, (덕밍아웃)

범주	정연	모모	사나	지효	미나	다현	채영	쯔위
나연	0	0	0	0	0	0	0	0
정연	1	0	0	0	0	0	0	0
모모	0	1	0	0	0	0	0	0
사나	0	0	1	0	0	0	0	0
지효	0	0	0	1	0	0	0	0
미나	0	0	0	0	1	0	0	0
다현	0	0	0	0	0	1	0	0
채영	0	0	0	0	0	0	1	0
쯔위	0	0	0	0	0	0	0	1

One-Hot-Encoding

범주	로제	리사	지수
제니	0	0	0
로제	1	0	0
리사	0	1	0
지수	0	0	1

회귀: $J-1$ 의 가변수 사용

(기준 범주 열 삭제)

분류: J 개의 가변수 그대로 사용

Label Encoding

각 범주에 점수를 할당하는 인코딩 방법

명목형 자료에 사용

에스파	점수
카리나	1
윈터	2
지젤	3
닝닝	4

할당된 점수는 임의의 숫자로 서로 순서나 연관성 X

Label Encoding

에스파	점수
카리나	1
윈터	2
지젤	3
닝닝	4

<장점 & 단점>

장점: 차원이 늘어나지 않음

단점: 정보의 왜곡 발생

-> 할당된 점수가 서로 순서나
연관성 있다고 판단할 가능성

Ordinal Encoding

순서형 변수에 대응하는 점수를 할당하는 인코딩 방법

순서형 자료에 사용

빡침 정도	점수
소	1
중	2
대	3
극대	4

1부터 순서대로 점수 할당

할당된 점수들은 순서나 연관성을 가짐!

Ordinal Encoding

<장점 & 단점>

장점: 차원이 늘어나지 않음

단점: 범주 간의 정확한 간격 반영 불가

빡침 정도	점수
소	1
중	2
대	3
극대	4



소



극대

Mean Encoding (Target Encoding)

각 수준에 대하여 Y 의 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 Y 와 설명변수 X 간의 수치적 관계 반영

Mean Encoding (Target Encoding)

<장점 & 단점>

장점: 차원이 늘어나지 않음

반응변수와의 관계를 고려한 점수이므로 당위성 존재

단점: 1) 이상치에 영향 크게 받음 -> 반응변수와의 관계 제대로 반영 X

2) 반응변수에 대한 정보 설명변수에 들어감 -> 과적합 가능성 ↑

Leave One Out Encoding (LOO Encoding)

현재 행을 제외하고(leave one out) 평균을 구해 할당하는 인코딩 방법

[Y] 양궁 획득 점수	[X] 국가	[X] 국가 (LOO Encoding)
30	대한민국	30
30	대한민국	30
30	대한민국	30
28	러시아	28.5
29	러시아	28
28	러시아	28.5
28	이탈리아	27
27	이탈리아	27.5
27	이탈리아	27.5

Leave One Out Encoding (LOO Encoding)

[Y] 양궁 획득 점수	[X] 국가	[X] 국가 (LOO Encoding)
28	러시아	28.5
29	러시아	28
28	러시아	28.5

러시아의 Y값인 28,29,28 중에서 -> 첫번째 관측치의 값인 28을 빼고
 -> 남은 29,28의 평균을 내서 -> 28.5라는 LOO Encoding 결과

<장점 & 단점>

- 장점:** 1) Mean Encoding에 비해 이상치의 영향 덜 받음
 2) 과적합 발생 가능성 비교적 ↓
- 단점:** Mean Encoding과 동일

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

Ordered Target Encoding (CatBoost Encoding)

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

=> 그 관측치보다 앞에 있는 "러시아"의 Y값 평균을 구한다

Ordered Target Encoding (CatBoost Encoding)

[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	28.2
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

앞에 같은 범주의 관측치가 없다면 전체 평균 할당

Ordered Target Encoding (CatBoost Encoding)

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

<장점>

같은 범주라고 해도 서로 다른 인코딩 값 가지게 됨

=> 과적합될 가능성 감소

차농남의 교학상장(敎學相長)론 (주분 드가자~)


여러분 안녕하세요ㅎ 범주형 자료분석팀 팀장입니다. 클린업을 마치면서 간단한 소감과 느낀점을 작성해보려고 하는데,,ㅎ 긴 글이지만 읽어주시면 감사하겠습니다(구백)

무엇보다도 3주동안 잘 따라와준 우리 팀원들 너무 고맙습니다ㅠㅠ 범주가 다소 생소하고 쉽지만은 않은 내용이었는데 개떡같이 설명해도 찰떡같이 이해해주는 팀원들을 보면서 너무 고마웠어요ㅠㅠ 저도 지난 학기 범주 팀이어서 스터디를 들었지만 전 범주팀장 현 부회장 지연이의 명강의에도 불구하고 저의 딸리는 이해력으로 인해 고생을 좀 했었는데, 여러분들은 그렇지 않은 것 같아서 참 대단한 사람들이라고 느꼈습니다ㅎㅎ 이해도 빠르네 열정은 어찌나 넘치던지, 모르는 부분 있으면 질문도 열심히 하고 실습도 늦거나 빠리는 사람 없이 열심히 하는 모습을 보면서 감동 받았습니니다ㅠㅠ 실습한 것 보면 이해를 했는지 못했는지 보이는데 다들 완벽하게 이해한 것이 보여서 너무 좋았습니다. 이런 여러분들의 명석한 두뇌와 뜨거운 열정이라면 무엇이든지 쉽게 해낼 수 있을 거예요,,,ㅎㅎ 클린업 발표도 어찌나 잘하던지..!(아직 못해본 팀원도 있긴 하지만 분명 잘할 듯ㅋㅋ) 완벽한 내용 숙지를 바탕으로 간결하면서도 핵심 내용은 다 전달하고, 약간의 긴장감이 있으면서도 여유롭고, 차분하면서도 속사포 랩을 쏟아내는 여러분들을 보면서 감탄을 금치 못했었습니다,,, 이렇게 완벽한 팀원들을 만난 저는 참 행운아라고 생각하고, 이렇게 열정적인 여러분들을 보면서 저 역시 나태해지지 말고 더 열심히 해야겠다는 동기부여도 되었습니다!

3주동안의 클린업을 돌아보면 1주차에는 분할표, 연관성 측도 등 범주의 기본적인 이론들에 대해서 배웠고, 2주차 때는 앞서 배운 이론들을 적용해볼 만한 GLM 모형을, 3주차 때는 실제 분석에 써먹을 수 있는 실용적인 방법들을 배웠어요. 바이오통계입문이나 범주형자료분석 수업을 들으신 분들은 수월하거나 익숙했겠고, 듣지 않았다면 헛갈리고 낯설었을 거예요! 둘 중 어느 경우가 됐든 잘 따라와준 여러분 정말 고맙습니다 ㅎㅎ

팀장이 돼서 팀을 구상할 때 두가지 목표가 있었는데 첫번째는 서로서로 도우면서 으쌔으쌔해서 다같이 성장하는 그런 팀을 만들고 싶었어요. 누구 한 명 낙오되거나 소외되는 사람 없이 모두가 푹푹 뭉치는 그런 팀이요! 교학상장(敎學相長)이라는 사자성어가 있어요. 가르치고 배우면서 성장한다는 뜻인데, 팀장·팀원, 기존·신입 할 것 없이 모두가 서로에게 가르칠 수 있고 배울 수 있다고 생각해요. 그래서 우리가 서로 도우면서 성장하면 결국에는 모두가 훌륭한 엘리트로 발돋움하지 않을까?라는 망상이 있는데, 지금까지 여러분의 모습을 보면 저의 목표는 이미 따 놓은 당상이 아닌가라는 생각이 듭니다ㅎㅎㅎ 우리 서로서로 도와가면서 최고의 팀을 만들어보자구요!

팀을 구상할 때 가졌던 두번째 목표는 귀여움입니다ㅋㅋㅋㅋ 귀여운 사람들끼리 모여서 영차영차 서로서로 도우면 얼마나 귀여울까라는 생각이 들었었어요ㅋㅋㅋㅋ 척직한 팀이 되면 어찌나 걱정했었는데 다행히 우리 팀원들은 한 명도 빠짐없이 귀여워서 너무 기분이 좋습니다 ㅎㅎㅎ 매번 세미나때마다 우리의 귀여움을 한껏 자랑하고 싶었어요. 우리 팀이 드러나는 유일한 시간이 발표시간이다 보니 여러분에게 귀여운 PPT를 강요하지 않았나 싶습니다,,ㅎㅎ 이런 억지 강요에 반발하거나 기분 나쁠 수도 있는데 이 부분은 미안하다고 말하고 싶네요,, ㅎㅎ 그래도 너무 잘 받아줘서 정말 고맙습니다ㅠㅠ 앞으로도 우리의 귀여움 잔뜩 어필해보자구요! ㅎㅎ

주제분석부터는 우리의 귀여움 외에 능력에 초점을 좀더 두어서 보여줄 시간이에요! 그동안은 워밍업이었고 본격적인 프로젝트가 진행될 텐데, 주제 선정부터 분석까지 하나도 쉽지 않을거예요,, 중간 끝나고 과제가 쏟아지는 기간이라 더 바빠져구요ㅠ 주제는 미미리미 생각해 놓는게 좋을거고 본인이 재밌다고 느끼는 분야를 주제로 생각해보는게 좋아요! (물론 선정이 안 될 수도 있지만ㅠㅠ) 힘든 과정이기 때문에 그나마 재미를 느껴야 할만해지거든요,,ㅎㅎ 앞으로도 서로서로 도와가면서 주제분석도 잘해봅시다! 저도 열심히 공부하고 배워서 여러분들 도와 주제분석 재밌게 해보려구요 ㅎㅎ 그럼 중간고사 잘 보시고 주제분석 때 보자구요 우리~~ 사랑합니다 범주팀 



THANK YOU

