범주형자료분석팀

2팀

김정훈
김상태
김민정
박정현
이윤희

INDEX

0. 1주차 리뷰

1. Confusion matrix

2. 평가 지표

3. ROC & AUC

4. 인코딩

Confusion Matrix(오류 행렬) 란?

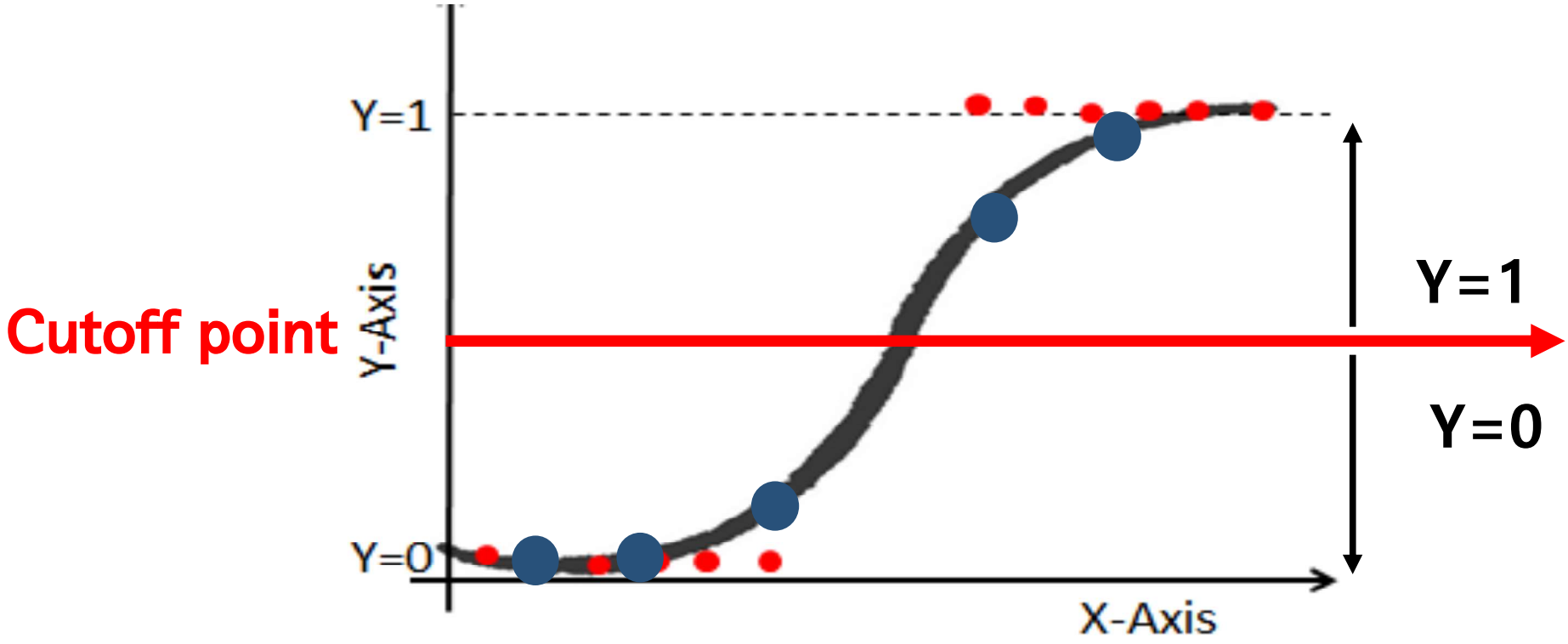
Training 을 통해 도출한 예측값 ($\hat{y}$)과 실제 관측값(y)을 비교한 표

True(y)	Predicted($\hat{y}$)	
	Y=1	N=0
Y=1	TP	FN
N=0	FP	TN

Cutoff Point

Cutoff point!

- Y=0 또는 1(실패/성공)의 값을 갖게 하려면 어떠한 기준값이 필요



평가 지표

평가 지표의 종류

정분류율(Accuracy)

전체에서 실제 관측값과 예측값이 맞은 경우의 비율

True(y)	Predicted($\hat{y}$)	
	Y=1	N=0
Y=1	TP	FN
N=0	FP	TN

$$\text{Accuracy} = TP + TN$$

평가 지표

F1-Score

Precision과 Sensitivity는 일종의 Trade-off 관계

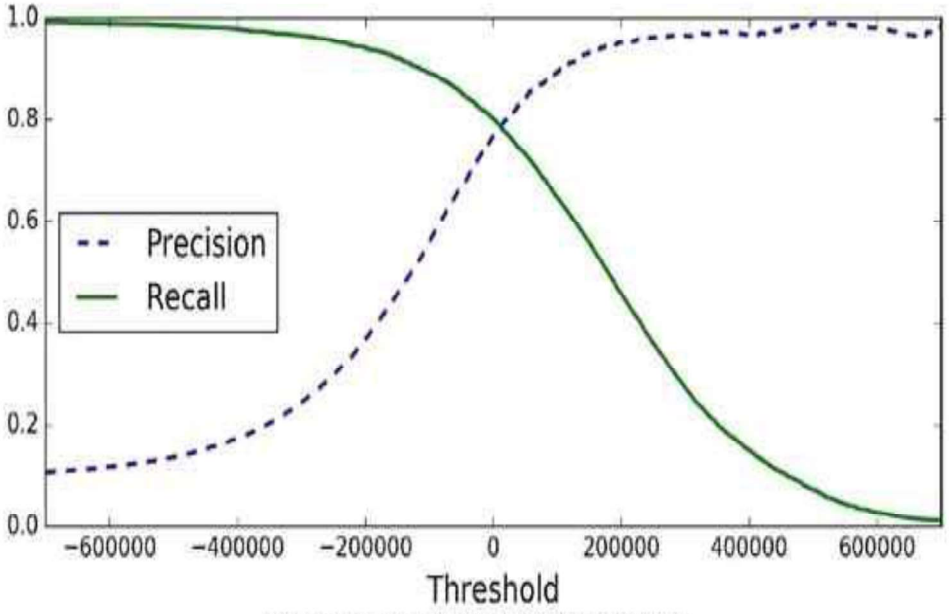


Figure 3-4. Precision and recall versus the decision threshold

설명

- Precision과 Sensitivity의 조화평균

두 지표의 밸런스를 고려하여 정확도 측정

- 두 개의 False 상황을 고려하는 지표

: imbalanced data 평가하는 지표로 좋음

한계

- TN을 사용하지 않는다

평가 지표

MCC (Matthew correlation coefficient)

설명

- 모든 부분을 사용하여 만들어 졌다
- 범위 : -1 ≤ MCC ≤ 1

  1 : 완벽하게 예측

  0 : 랜덤 예측과 같음

  -1 : 완벽하게 예측 실패

- Confusion Matrix 설명에 가장 좋은 지표

다른 지표

- Accuracy

  : imbalanced data X

- F1-Score

  : TN 부분 사용X ,

  : Y와 N이 바뀌게 될 때 성능 지표 바뀜

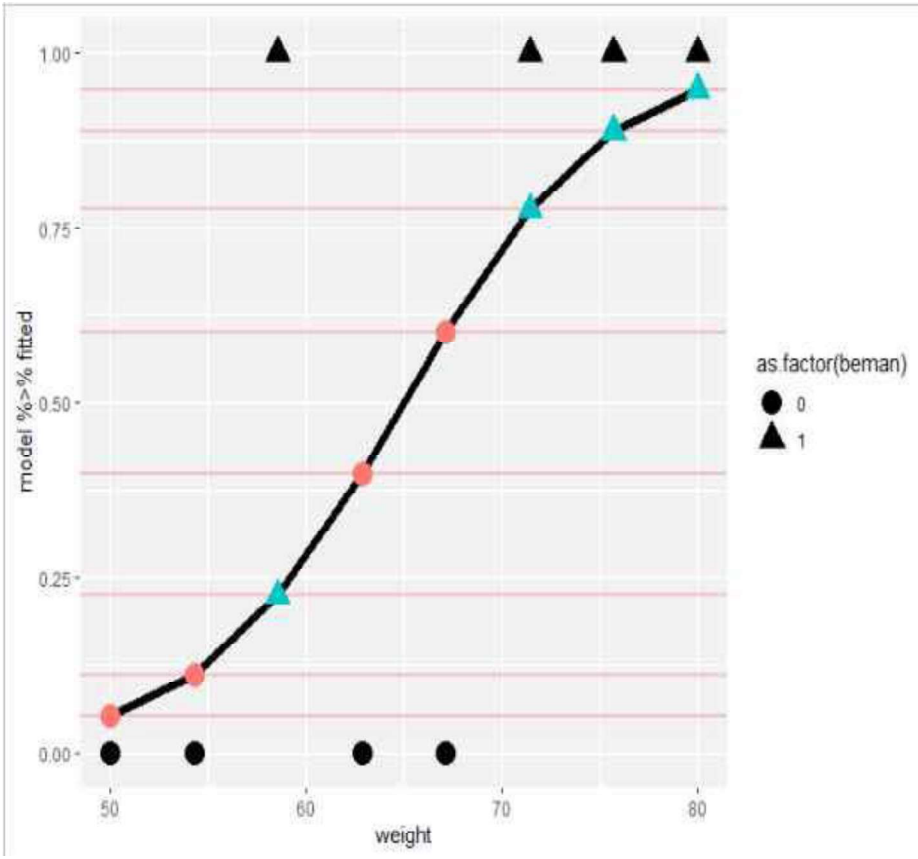
ROC (Receiver Operating Characteristic)

그래프의 모양

- X축 : FPR = 1 – 특이도
- Y축 : TPR = 민감도
- 일반적으로 우상향하는 위로 볼록한 곡선 혹은 직선

Why?

- cutoff가 0에 가까워지면
  (FPR, TPR) → (1, 1)이 됨!
- cutoff가 1에 가까워지면 반대로
  (FPR, TPR) → (0, 0)에 가까워 짐!



AUC

특징

- 분류 임계값(cutoff) 불변의 측도이다!

모든 것에 대한 것을 고려하여 예측 품질을 측정하고 있기 때문

범위 : 0 ~ 1

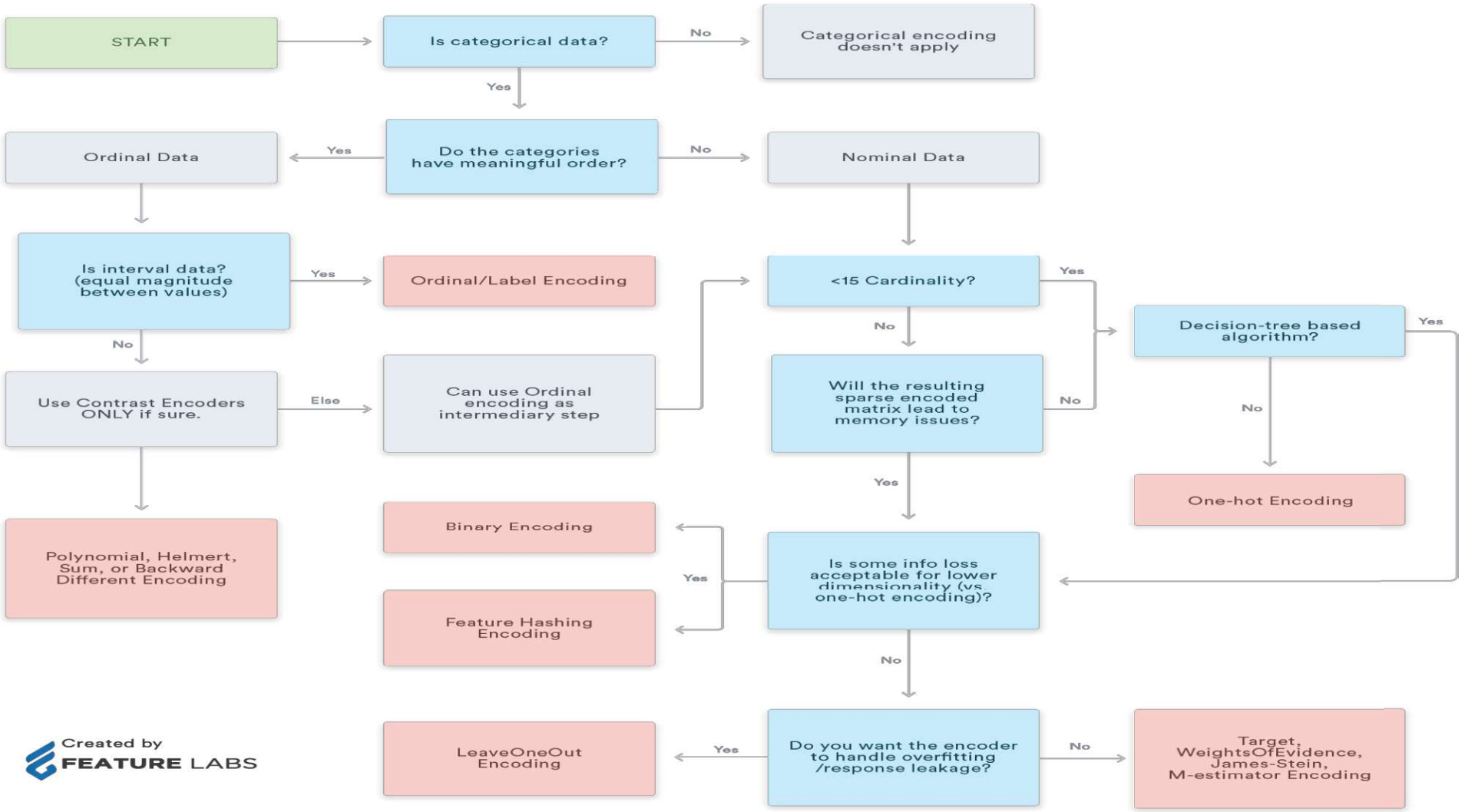
0 : 예측이 100% 잘못된 모델

0.5 : 거의 무작위

1 : 예측이 100% 잘 된 모델

인코딩(Encoding)

인코딩(Encoding)의 종류



4 인코딩(Encoding)①

One-Hot Encoding

방법

- 특정 범주형 변수에 해당하는 각각의 범주들을 1 또는 0의 값을 갖는 vector로 Mapping
- 어원 : 1개만 Hot(True) 나머지는 Cold(False)

특징

- 범주 개수를 N개라 할 때, 최소 N-1개의 가변수를 만드는 Encoding 방법
- Regression의 경우 N-1개의 가변수만 만들어 줘도 무방
- Tree 기반 모델일 경우 N개의 가변수를 만드는 것을 권장

4 인코딩(Encoding)③

Ordinal Encoding

장점

· 변수의 개수가 늘어나지 않는다

· 순서형 범주의 경우 순서 고려 가능

단점

· 범주간의 순서를 표현하는데

어느 정도의 차이 값을 줄지 난해

예시

카카오프렌즈	머리크기
제이지	매우 크다
라이언	매우 크다
프로도	크다
네오	크다
무지	보통이다
튜브	보통이다
어피치	작다
콘	매우 작다

카카오프렌즈	머리크기
제이지	5
라이언	5
프로도	4
네오	4
무지	3
튜브	3
어피치	2
콘	1

콘과 어피치 크기 차이 = 1

어피치와 무지의 크기 차이 = 1 ??