<YDMS 2023-1> 4주차 과제

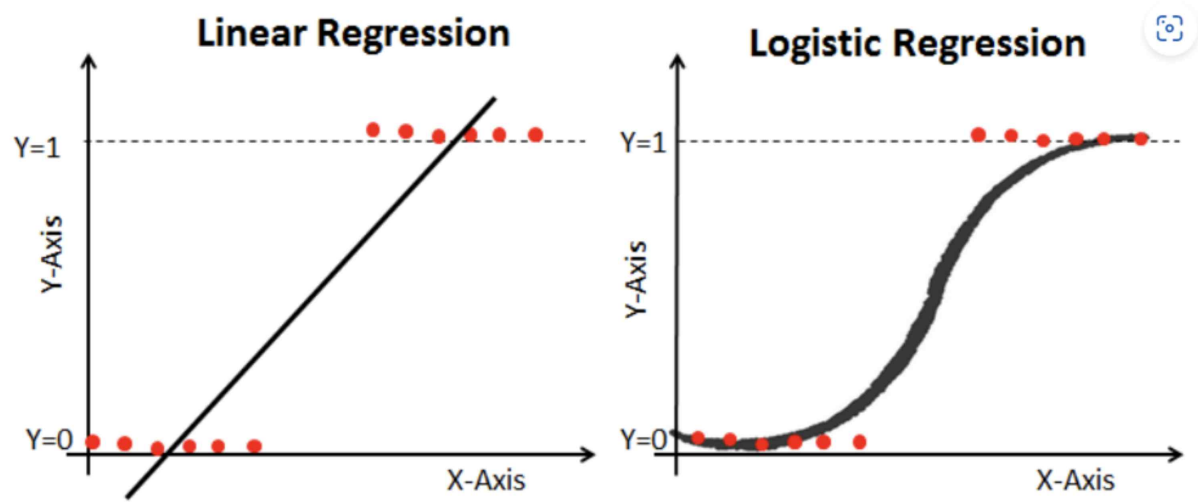
분류

주하연

1. 로지스틱 회귀분석 개념 (+오즈비)

로지스틱 회귀는 이진 분류 작업에 사용되는 방법이다.

선형회귀분석(단순, 다중)은 모두 종속변수가 연속형이었다. 그러나, 로지스틱 회귀분석은 종속변수가 범주형이면서 '0 or 1'의 값을 가지는 경우 사용한다. 독립변수의 값에 따라 종속변수의 확률을 예측한다.



로지스틱 함수는 위의 그림에서 보이는 것과 같이 S자 형태를 가진다.

종속변수가 0 or 1 의 값을 가질 때, 서형회귀로는 fitting을 하기가 힘들다. 그래서 곡선으로 fitting 하기 위해 사용하는 것이 'logistic함수 = 로짓변환' 이다.

a. 로지스틱 회귀의 목적

   i. 독립변수들의 최종적인 outcome인 0,1(binary)이 되는 함수를 찾는 것

   ii. 로짓은 독립변수들에 대한 선형 모형으로 추정 가능

b. logistic 함수

   로지스틱 함수는 로지스틱 회귀분석에서 사용되는 S자 모양의 함수로서, 선형 예측값을 0과 1 사이의 확률값으로 변환하는 데 사용된다. 로지스틱 함수는 종속 변수의 성공 확률을 나타내는데 적합한 함수이다.

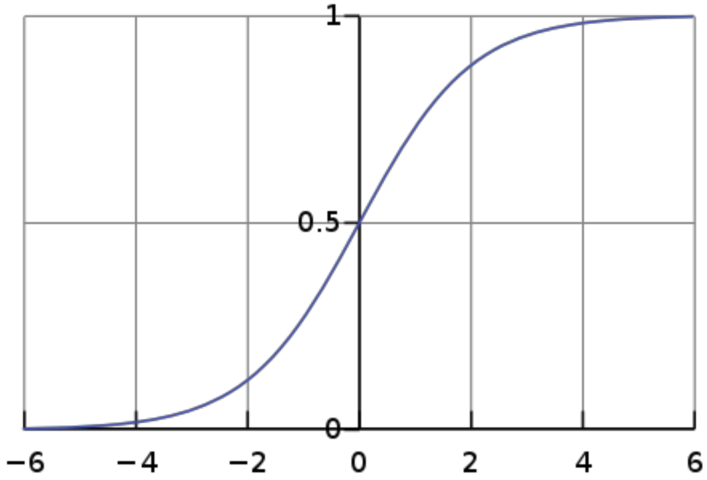
   로지스틱함수는 로짓함수의 역함수이기도 하다.

f(x) = 1 / (1 + e^(-x)) -> 로지스틱 함수

여기서 x 는 선형 예측값이다. 로지스틱 함수는 x 를 입력으로 받아 0과 1사이의 값을 출력한다. x 가 증가하면 f(x) 도 증가하며, x가 감소하면 f(x) 도 감소한다.

이러한 특성으로 로지스틱 함수는 종속변수의 이진 분류를 위해 사용된다.

ex) f(x) = 0.8 이라면, 종속변수가 1이 될 확률 = 0.8



i. 로짓함수

logit(p) = ln(p / (1 - p)) -> 로짓함수

여기서 p는 종속 변수의 성공 확률이다. 그러면 (1-p)는 실패 확률이다.

따라서 (p / (1 - p)) 는 '실패에 비해 성공할 확률의 비' (Odds)를 의미한다.

$$\text{오즈} = \frac{\text{사건이 일어날 확률}(p)}{\text{사건이 일어나지 않을 확률}\ (1-p)}$$

이 Odds에 log를 취한 것이 로짓함수 (logit(p)) 이다. (오즈의 단점때문에 log를 취함 -> 오즈비 설명란에 적어둠)

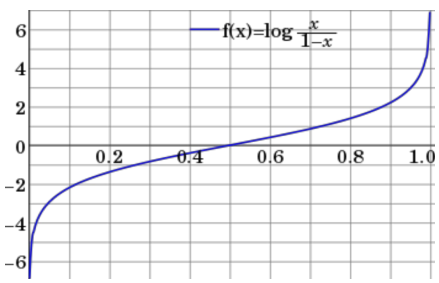
-> 오즈에 로그(log)를 붙이면 더 활용범위가 높아진다.

-무한대 < log(오즈비)=logit(p) < 무한대

ex) 게임에서 이길 확률 = 1/5, 게임에서 질 확률 = 4/5

=> 게임에서 이길 Odds = 1/4

=> 해석 : 5번 중에, 4번 질 동안 1번 이긴다



c. 오즈비 (OR : Odds ratio)

관측치가 발생할 확률과 발생하지 않을 확률 간의 비율 -> 연구에서는 이 값을 이용하여 해석

오즈비는 주로 해석적인 목적으로 사용되며, 통계적 유의성 검정이나 신뢰구간을 통해 해당 독립변수의 영향력을 평가하는 데 활용된다.

OR = (1 / 1-p) / (q / 1-q)

오즈비는 두 그룹간의 오즈(odds)의 비율로 정의된다. 예를 들어, 오즈비가 2라면, 두 그룹간의 발생 비율이 2배 차이가 난다는 의미이다.

- 오즈비가 1보다 큼 : 한 그룹이 다른 그룹보다 더 높은 발생 비율을 가짐

= 해당 독립 변수가 종속 변수에 긍정적인 영향을 미친다고 해석 가능

- 오즈비가 1보다 작음 : 한 그룹이 다른 그룹보다 더 낮은 발생 비율을 가짐

= 해당 독립 변수가 종속 변수에 부정적인 영향을 미친다고 해석 가능

But, 오즈의 문제는 범위가 0 < OR < 무한대 이며 비대칭이다. 따라서 이를 해결하기 위해 로그를 취한다. (=> logit(p))

오즈비는 통계에서 강력한 도구이지만, 이 외에도 몇가지 제한사항이 있다.

i. 오즈비는 인과 관계를 나타내지 않는다. 연관성은 나타내지만, 하나의 사건이 다른 사건을 유발한다는 것은 나타내지 않는다.

ii. 데이터의 품질에 민감하다. 데이터에 편향이 있으면 오즈비가 왜곡될 수 있다.

iii. 표본 크기에 민감하다. 표본 크기가 작으면 오즈비가 불안정할 수 있다.

d. 다항 로지스틱 회귀분석 (Multinomial Logistic Regression)

독립변수	종속변수	분석방법
명목척도 서열척도 등간척도 비율척도	명목척도(2개)	이분형 로지스틱 회귀분석
	명목척도(3개 이상)	다항 로지스틱 회귀분석
	서열척도(3개 이상)	다항 로지스틱 회귀분석

로지스틱 회귀분석은 종속변수에 따라서 이분형 로지스틱과 다항 로지스틱 분석으로 구분된다.

다항 로지스틱 회귀분석은 종속 변수가 세 개 이상의 범주일 때 사용할 수 있다.

ex) 고객이 제품 A,B 또는 C를 구매할 확률을 예측, 환자가 질병 A,B 또는 C에 걸릴 확률을 예측

e. 로지스틱 회귀분석이 사용되는 예

i. 의학 : 질병의 발생 여부를 예측

ex) 어떤 환자가 암에 걸릴 확률 예측, 특정 증상이 있는 환자가 특정 질병을 가

지고 있는지를 판단

ii. 마케팅 : 고객이 제품을 구매할 확률을 예측

ex) 고객의 특성이나 이전 구매이력등을 독립 변수로 활용하여 고객의 구매 여부를 예측하고, 이를 기반으로 타겟 마케팅이나 마케팅 전략을 개발

iii. 신용평가 : 개인의 신용 등급을 예측

ex) 개인의 소득, 신용카드 사용 이력, 대출 이력 등을 독립 변수로 활용하여 해당 개인의 신용 등급을 예측하는 데에 활용

iv. 사기탐지 : 금융 거래에서 사기 행위를 탐지

ex) 거래 패턴, 이상 거래 패턴, 고객의 행동 패턴 등을 독립 변수로 활용하여 사기 행위의 가능성을 예측, 사기 탐지 모델 개발

v. 인사관리 : 직원의 이직 여부를 예측

ex) 직원의 근무 조건, 만족도 조사 결과, 보상 등을 독립 변수로 활용하여 직원의 이직 가능성을 예측하고, 이를 기반으로 인사 정책을 조정하거나 직원 ㅇ지를 위한 전략을 수립

f. 선형 회귀분석과 로지스틱 회귀분석 비교

- 선형 회귀분석 모형 : 연속 종속변수와 하나 이상의 독립 변수 간의 관계를 파악하는 데 사용된다. 독립 변수와 종속 변수가 각각 1개인 경우를 단순 선형 회귀라고 하고, 독립 변수의 수가 늘어나면 다중 선형 회귀라고 한다. 각 선형 회귀 유형은 일반적으로 최소 제곱법을 사용하여 계산되는 일련의 데이터 포인트를 통해 최적합선을 그리는 것을 목표로 한다.

선형 회귀분석과 유사하게 로지스틱 회귀분석도 종속 변수와 하나 이상의 독립 변수 간의 관계를 추정하는 데 사용되지만, 범주형 변수와 연속형 변수에 대한 예측을 수행하는 데 사용된다. 범주형 변수는 참/거짓, 예/아니오, 1/0 등이 될 수 있다. 측정 단위도 확률을 계산한다는 점에서 선형 회귀분석과 다르지만 로짓함수는 S곡선을 직선으로 변환한다.

두 모형 모두 회귀 분석에서 미래의 결과를 예측하는 데 사용되지만, 일반적으로 선형 회귀번석이 더 이해하기 쉽다. 또한, 선형 회귀분석의 경우, 로지스틱 회귀분석에서 모든 응답 범주에 걸쳐 값을 표현하는 데 필요한 표본의 크기만큼 큰 표본이 필요하지 않다.

2. 의사결정나무(Decision tree) 개념 (+ 관련 모델 조사)

데이터를 분류하거나 예측하기 위한 분석 방법 중 하나이다.

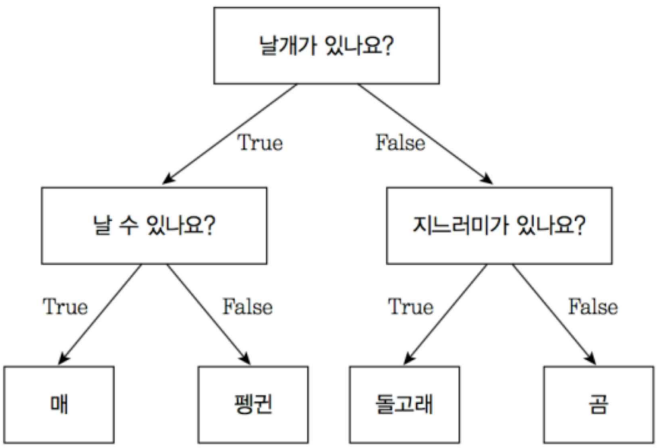
의사결정나무는 트리 구조로 표현되며, 데이터의 속성을 기반으로 분할 규칙을 생성하고

이를 통해 데이터를 분류하거나 예측한다.

비모수 기반의 지도학습 알고리즘으로, 특정 규칙에 의해 분할을 하며 X와 Y를 연결시켜 분류 또는 예측 (회귀와 분류 모두 가능)

각 내부 노드는 특정 변수의 값을 검사하며, 분기점으로서 다른 내부 노드나 단말 노트(리프)로 이어진다. 리프노드에서는 최종 예측 결과가 출력된다.

- 스무고개 방식의 기계학습 모형, 각 네모칸을 트리의 노드(Node) 라고 함
- 들어오는 선은 최대 1개, 나가는 선은 적어도 2개
- 가장 위쪽에 위치하는 것 = 뿌리 마디 (Root Node), 마지막 것 = 리프 노트(leaf node)
- 분석결과 : '조건 A이고 조건 B이면 결과집단 C' 라는 형태의 규칙으로 표현됨



a. 의사결정나무 장점
   i. Simple : 직관적이고 이해하기 쉬움 -> 결과를 해석하고 설명하기 쉽다
   ii. Little data preparation : 데이터 준비가 수월함
   iii. Robust : 별다른 가정을 하지 않고 단순히 쪼개기만 함
   iv. Performs well with large dataset : 데이터가 클 때 성능이 좋음
   v. fast : 빠르게 훈련할 수 있음
   vi. 변수 중요도 : 변수들의 상대적인 중요도를 평가할 수 있으며, 변수 선택 기준에 따라 분기점을 결정함으로써 가장 중요한 변수를 식별할 수 있다.

b. 의사결정나무 단점
   i. 최적점을 찾기 힘들다

ii. 불안정성, 계층적 구조로 변동성이 큼(= 조금만 바뀌어도 결과가 확 바뀜)

iii. 범주를 2개로만 나누어서 처리(yes / no)

iv. Overfitting(과적합) - 트리의 깊이가 깊어지면 훈련데이터에 대한 예측은 높아지지마느 새로운 데이터에 대한 일반화 성능은 저하될 수 있다.

-> 이를 완화하기 위해 가지치기(Pruning) 기법이 사용된다.

v. Selection bias

vi. 특정 컨셉을 표현하기 어려움

c. 의사결정나무 분석과정

i. 목표변수(Target)와 관계가 있는 독립변수들의 선택

ii. 분석목적과 자료의 구조에 따라 적절한 분리기준과 정지규칙을 정하여 의사결정나무의 구조 작성

- 정지규칙 : 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 여러가지 규칙을 의미한다. 이러한 규칙에는 최대 나무의 깊이, 자식마디의 최소 관측치 수, 또는 카이제곱 검정통계량, 지니지수, 엔트로피 지수 등이 될 수 있다.

iii. 부적절한 나뭇가지는 제거 (가지치기)

- 가지치기란 끝마디가 너무 많으면 모형이 과대 적합된 상태로 현실문제에 적용할 수 있는 적절한 규칙이 나오지 않게 된다. 따라서 분류된 관측치의 비율 또는 MSE 등을 고려하여 적절한 수준의 가지치기 규칙을 제공하여야 한다.

iv. 이익(Gain), 위험(Risk), 비용(Cost) 등을 고려하여 모형평가

v. 분류(Classification) 및 예측(Prediction)

d. 관련모델(의사결정나무 알고리즘)

i. CHAID

널리 사용되는 알고리즘이다. 이는 명목형, 순서형, 연속형 등 모든 종류의 목표변수와 분류변수에 적용이 가능하다.

ii. CART (Classification and Regression Trees)

분류와 회귀분석을 모두 지원하는 알고리즘이다.

하나의 부모마디 밑에 2개의 자식마디만이 생기는 이지(Binary) 분리 알고리즘이다. CHAID와 마찬가지로 목표변수나 분류변수의 척도에 관계없이 적용할 수 있다는 장점이 있다.

https://tyami.github.io/machine%20learning/decision-tree-4-CART/

https://heeya-stupidbutstudying.tistory.com/entry/ML-%EA%B2%B0%EC%A0%95%ED%8A%B8%EB%A6%ACDecision-Tree-%ED%8C%8C%ED%97%A4%EC%B9%98%EA%B8%B0

iii. ID3 (Iterative Dichotomiser 3)

분류를 위해 설계된 의사결정나무 알고리즘이다. 이는 엔트로피를 사용하여 데이터를 분할 기준을 선택한다. 엔트로피는 주어진 데이터 집합의 불확실성을 측정하는 척도로 사용된다. ID3 알고리즘은 분할 이후의 엔트로피 감소량이 최대가 되는 변수를 선택하여 의사결정나무를 구축한다.

하지만, ID3는 범주형 변수에만 적용할 수 있으며, 연속형 변수에는 적용하기 어렵다.

iv. C4.5

ID3의 개선된 버전이다. ID3와 마찬가지로 엔트로피를 사용하지만, 정보획득량 대신 정보이득비율을 사용하여 변수를 선택한다. 정보이득비율은 변수의 분할 후 엔트로피 감소량을 변수 분할 전 에느로피로 정규화한 값이다.

이는 범주형 변수와 연속형 변수 모두에 적용할 수 있다.

가지치기라는 과적합 방지 기법을 사용하여 의사결정나뮤의 일반화 성능을 향상시킨다.

v. C5.0

ID3와 C4.5의 개선된 버전이다. C5.0은 로스함수를 최소화 하는 방식으로 분할 기준을 선택하며, 가중치를 사용하여 변수의 중요도를 평가하는 기능을 제공한다.

vi. MARS

vii. QUEST

3. 분류성능 평가지표

분류 모델을 평가할 때 주로 Confusion Matrix를 기반으로 Accuracy, Precision, Recall, F1-score를 측정한다.

머신러닝 모델은 일반적으로 모델 회귀냐 분류냐에 따라 서로 다른 평가 지표를 사용해야 한다. 회귀모델 평가지표는 MAE, MSE, RMSE, R2 등이 있다.

https://velog.io/@73syjs/%EB%B6%84%EB%A5%98-%EB%AA%A8%EB%8D%B8%EC%9D%98-%ED%8F%89%EA%B0%80-%EC%A7%80%ED%91%9C

a. 오차 행렬(Confusion Matrix)

분류 모델(classifier)의 성능을 측정하는 데 자주 사용되는 표로 모델이 두 개의 클래스를 얼마나 헷갈려하는지 알 수 있다.

		예측	
		Positive	Negative
정답	Positive	TP	FN
	Negative	FP	TN

- **T**(True): 예측한 것이 **정답**
- **F**(False): 예측한 것이 **오답**
- **P**(Positive): 모델이 **positive**라고 **예측**
- **N**(Negative): 모델이 **negative**라고 **예측**

- **TP**(True Positive): 모델이 **positive**라고 **예측**했는데 실제로 **정답**이 **positive** (정답)
- **TN**(True Negative): 모델이 **negative**라고 **예측**했는데 실제로 **정답**이 **negative** (정답)
- **FP**(False Positive): 모델이 **positive**라고 **예측**했는데 실제로 **정답**이 **negative** (오답)
- **FN**(False Negative): 모델이 **negative**라고 **예측**했는데 실제로 **정답**이 **positive** (오답)

b. 정확도(Accuracy)

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

전체 예측 결과 중 올바르게 예측한 비율을 나타낸다. 즉, 정확히 맞춘 샘플의 수를 전체 샘플의 수로 나눈 값이다.

정확도는 데이터셋의 클래스 균형이 좋을 때 유용한 지표이다. 그러나 클래스 불균형이 심한 경우에는 정확도만으로는 모델의 성능을 평가하기에는 충분하지 않을 수 있기 때문에 Recall과 Precision을 사용한다.

0 ~ 1 사이의 값을 가지며, 1에 가까울수록 좋다.

c. 정밀도(Precision)

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

실제로 정답이 positive인 것들 중에서 실제로 정답이 Positive인 비율이다.

이는 잘못된 Positive를 최소화하고, 모델이 Positive라고 예측한 것이 실제로 Positive인지에 대한 민감도를 나타낸다.

Precision을 높이기 위해선 FP(모델이 positive라고 예측했는데 정답은 negative인 경우)를 낮추는 것이 중요하다.

0 ~ 1 사이의 값을 가지며, 1에 가까울수록 좋다.

d. 재현율(Recall) = TPR(True Positive Rate)

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

실제로 정답이 positive인 것들 중에서 모델이 positive라고 예측한 비율이다.

이는 잘못된 Positive를 최소화하고, 모델이 실제로 Positive인 것을 놓치지 않는 민감도를 나타낸다.

Recall를 높이기 위해선 FN(모델이 negative라고 예측했는데 정답은 positive인 경우)를 낮추는 것이 중요하다.

0 ~ 1 사이의 값을 가지며, 1에 가까울수록 좋다.

- Precision과 Recall 중 어느 한 평가 지표만 매우 높고, 다른 수치는 매우 낮으면 바람직 하지 않은 결과이다.

e. F1 score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Precision과 Recall의 조화평균이다.

Precision과 Recall은 상호 보완적인 평가지표이기 때문에 F1-score를 사용한다.

Precision과 Recall이 한쪽으로 치우쳐지지 않고 모두 클 때 큰 값을 가진다.

불균형한 클래스 분포에서 모델의 성능을 평가할 때 유용하다.

0 ~ 1사이의 값을 가지며, 1에 가까울수록 좋다.

f. 기타

i. 오분류율(Error Rate)

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

모델이 전체 데이터에서 잘못 맞춘 비율이다.

ii. 특이도(Specificity) - TNR(True Negative Rate)

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

실제 정답이 negative인 것들 중에서 모델이 negative라고 예측한 비율이다.

iii. 위양성률(Fall Out)

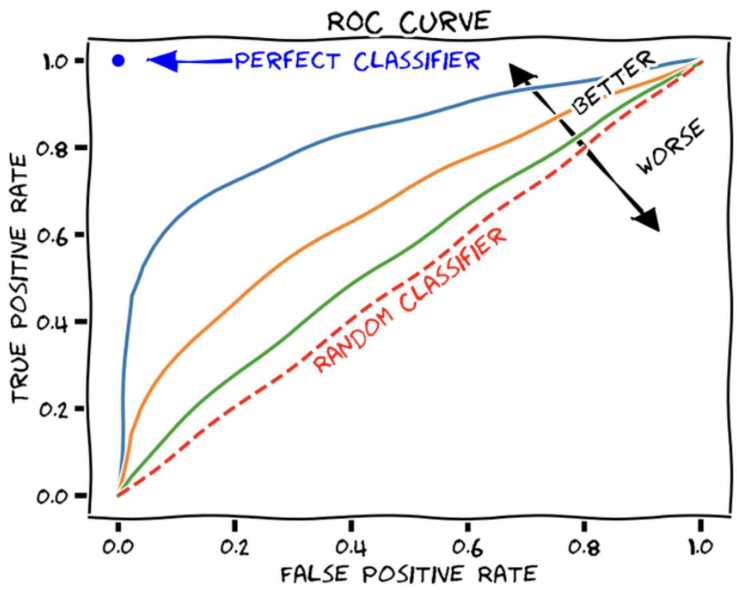
$$FallOut = 1 - Specificity = 1 - \frac{TN}{TN+FP} = \frac{FP}{FP+TN}$$

실제 정답이 negative인 것들 중에서 모델이 positive라고 예측한 비율이다.

iv. ROC

ROC Curve는 임계값에 대한 TPR, FPR의 변화를 곡선으로 나타낸 것이다.

X축에 FPR, Y축에 TPR을 두어 최적의 임계값을 찾는 것이다.

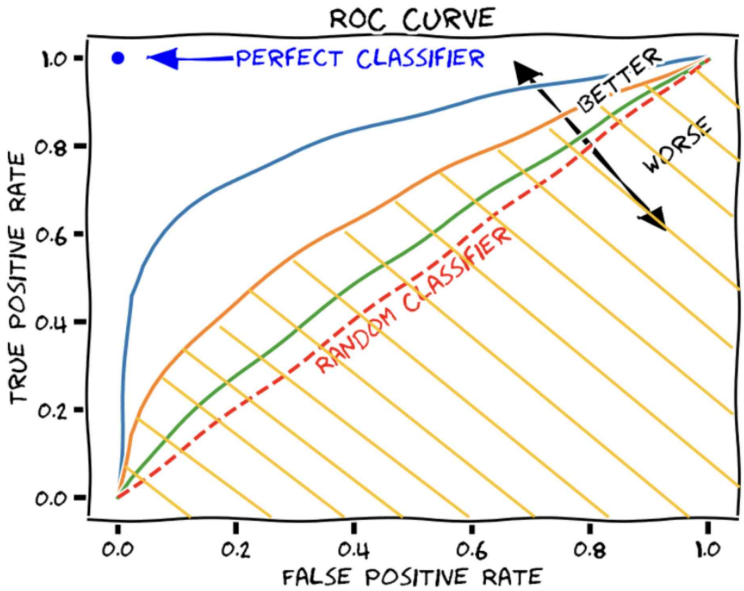


- TPR = recall , FPR = FP / FP+TN

TPR이 클수록, FPR이 작을수록 성능이 좋으며, ROC Curve가 좌측 상단측을 향할 때 모델의 성능이 좋다고 판단할 수 있다. 직선일 때 성능이 가장 안좋다.

v. AUC

AUC score는 말 그대로 커브 아래 공간을 말하는 것인데, 이 공간의 넓이가 넓을수록 성능이 좋은 모델이라는 뜻이다.



4. 본 데이터에는 10년후 관상동맥 질환 발병 여부에 영향을 미칠 것이라 판단되는 독립변수들이 있습니다. 10년 후 관상독맥 질환 발병 여부에 대하여 로지스틱 회귀분석과의사결정나무 모형을 적합시켜 어떠한 변수들이 10년 후 관상동맥 질환 발병 여부에 어떻게 영향을 미치는지 알아보시오 (설명 모형)

- 로지스틱 회귀

https://colab.research.google.com/drive/1Dek3ogTVCMtSl1Q3gLjzOfGGIBa15enC?usp=sharing

- 의사결정나무

https://colab.research.google.com/drive/1jTsUIpPHFXfM5dFqT3UVwQ2AodE4gWMl?usp=sharing

-framingham dataset 시각화

https://colab.research.google.com/drive/1k-lY0I9y-Dsdmh5t7Qmu5nz3wvW6UIfr?usp=sharing

디스커션 : 성능평가를 할 때 성능 지표 중 accuracy는 0.75로 아주 나쁘지는 않게 나온 반면, recall, precision, f1-score는 수치가 0.20~0.27 로 아주 낮게 나왔다. 이럴 때는 학습을 다시 진행해야하나요? 아니면 하이퍼파라미터튜닝을 해줘야하나요?