MOMENTUM

CRISP-DM :: Modeling

팀별 활동을 통해 실습과 이론을 동시에~

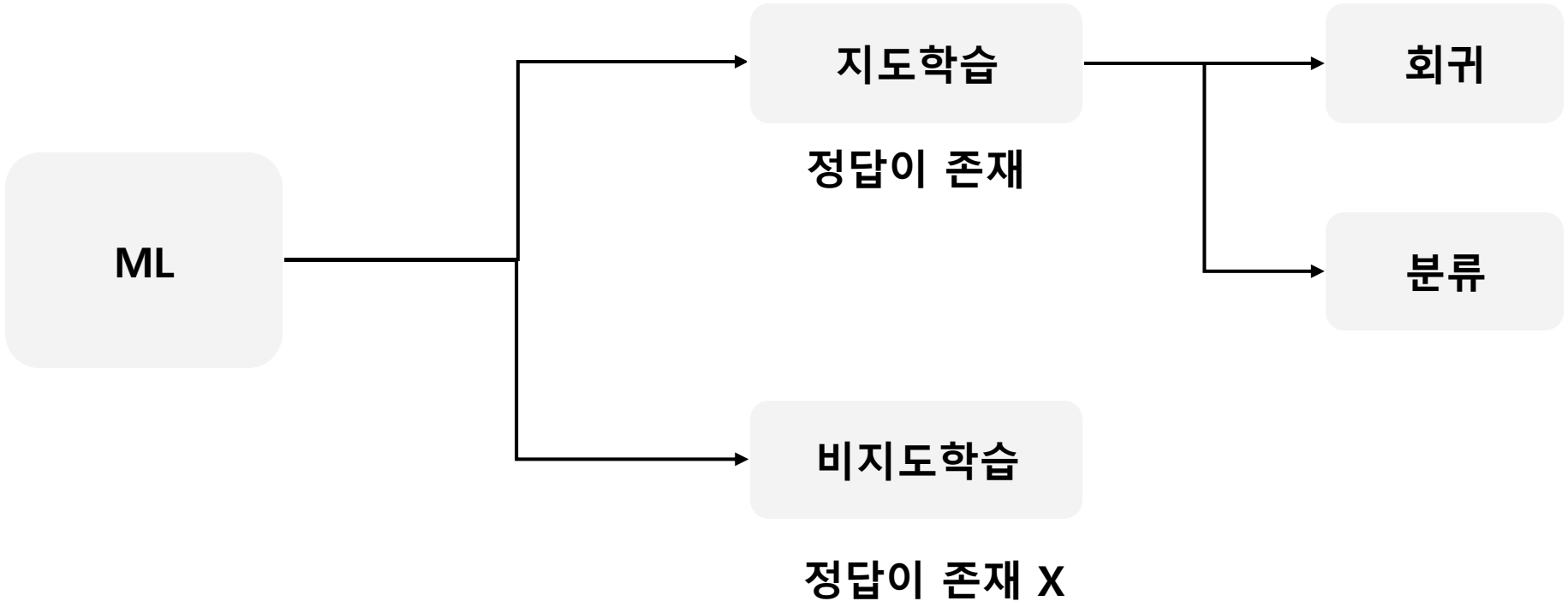
개요

- 지난 시간 복습 브리핑
- 데이터 이해 왜 해야하는가?
- ...

머신러닝이 뭔가요?

머신러닝 모델을 이용하여 데이터의 **패턴을 학습(Fit)**하고,
미래에 대한 **판단이나 예측(Predict)**을 하는 것

머신러닝의 종류



지도학습과 비지도학습

지도학습		비지도학습
정답이 있는 데이터를 학습	**정의**	정답 없이 데이터의 패턴 및 구조 학습
예측 또는 분류	**목표**	군집화, 차원 축소
훈련데이터(X) + 정답데이터(Y)	**입력데이터**	훈련데이터(X)만 있음
고객 이탈예측, 주택 가격 예측	**예시**	고객 군집화, 차원 축소
선형/로지스틱회귀, 랜덤포레스트	**주요 알고리즘**	K-Means, DBSCAN

지도학습 – 회귀와 분류

회귀		분류
숫자(연속값)을 예측	**정의**	카테고리(범주형 값)을 예측해 분류
가격, 시각, 확률 등 수치등을 예측	**목표**	Yes/No, A or B, 고양이 강아지등을 분류
가격 예측, 재구매 확률 예측	**예시**	스펨 메일 분류, 질병 유무 진단, 고객 이탈 예측
선형회귀, 라쏘, 릿지, 랜덤포레스트	**주요 알고리즘**	로지스틱 회귀, 랜덤포레스트

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | 평가하기

무엇을 예측할 것인가?

우리의 목표는 무엇인가?

예측 목표 : 고객이 서비스에서 이탈하는지 여부 예측

예측 타겟 변수 : 이탈여부 (1=이탈, 0 =유지)

ML 프로세스

요구사항 정의 | **데이터 수집 및 전처리** | 모델 정의하기 | 평가하기

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180일	200,000원	5일 전 전정,?	남자	ON	이탈안함
002	30일	15,000원	20일 전	여자	OFF	이탈함

어떤 데이터를 수집할 것인가?

데이터를 종류별로 어떻게 전처리할 것인가?

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	1	1	0
002	30	15000	20	0	0	1

컴퓨터가 이해할 수 있도록 데이터 처리하기

스케일링, 인코딩,결측치 처리

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | 평가하기

어떤 모델을 사용할 것인가?

지도학습인가?

비지도학습인가?

분류인가?

회귀인가?

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	1	1	0
002	30	15000	20	0	0	1

X : 독립변수 Y : 종속변수

X와 Y를 모두 학습하니 지도 학습이며, 이진 분류 모델을 사용해야함

X에 따른 Y(이탈여부)를 모델에 학습 (훈련데이터) Fit한다

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
102	200	123123	3	1	1	?
150	3	100	201	0	1	?

X : 독립변수 Y : 종속변수

훈련시킨 모델을 통해 다른 고객의 이탈 여부를 예측한다.

같은 독립 변수를 갖고 있는 고객들의 데이터를 X로 넣고 Y를 Predict한다

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | 평가하기

그래서 잘 예측이 되었는가?

예측	실제
102번 : 이탈	102번 : 이탈X
150번 : 이탈X	150번 : 이탈X

실제값과 예측값을 비교해
모델의 성능을 평가한다.

Scikit-learn 라이브러리

파이썬 기반 머신러닝 라이브러리

분류, 회귀, 군집, 차원축소등 다양한 ML 기능 제공

복잡한 수학 없이 간단한 코드로 머신러닝 모델을 구현할 수 있음

데이터 수집 및 전처리

- 데이터 준비
- 데이터 전처리
- 학습/테스트 분리 (Train/Test Split)

모델 정의하기

- 모델 학습(Fit)
- 예측 (Predict)

평가하기

- 예측 결과 평가

데이터 수집 및 전처리

train data의 구성 – 학습용 (Fit)

- **X (독립변수)**
  - 모델이 학습할 수 있도록 주는 데이터
- **Y (Target data, 종속변수)**
  - 모델이 예측 해야하는 정답

Test data의 구성 - 실제 예측용 (Predict)

< Train data>

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	1	1	0
002	30	15000	20	0	0	1

< Test data>

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
102	200	123123	3	1	1	?
150	3	100	201	0	1	?

데이터 수집 및 전처리

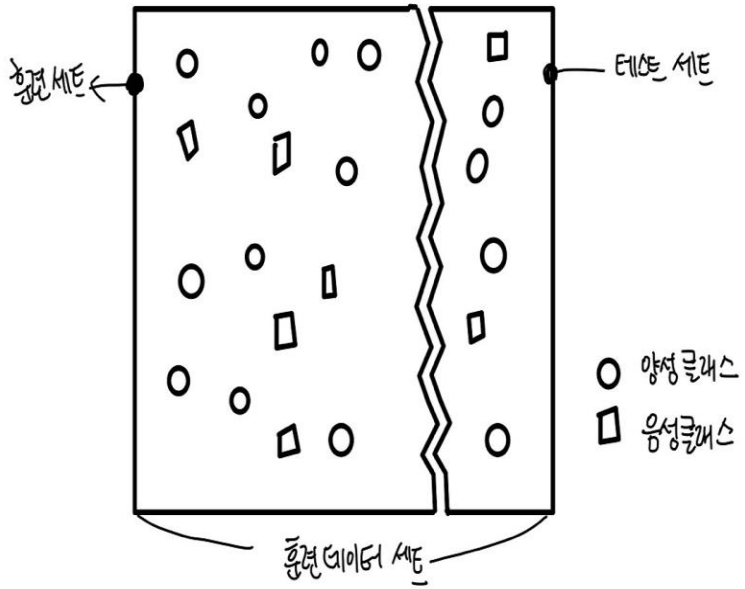
train test split

- 그런데... train한 모델이 성능이 잘 나오는지 알고 싶어요..
- 근데 제가 공부한 데이터로 답을 맞춰보면 당연히 다 맞겠죠!
- 그래서 train 데이터의 일부를 test 데이터로 모의고사를 봅니다!

데이터 수집 및 전처리

train test split

- 100개의 데이터 중, 80개는 학습하고
- 나머지 20개는 모의고사를 보면서 실제로 답을 잘 맞추는지 확인해요.
- **그 후 실제 test 데이터로 예측을 단 한 번만 수행!**



훈련데이터세트 나누기

데이터 수집 및 전처리

인코딩(Encoding)

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	남자	1	0
002	30	15000	20	여자	0	1
003	12	31	2	외계인	0	0

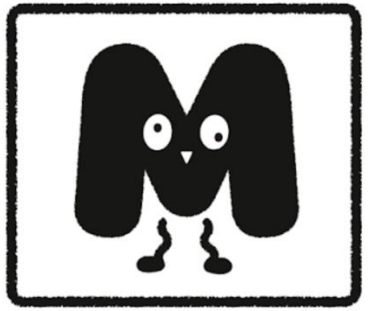
고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	0	1	0
002	30	15000	20	1	0	1
003	12	31	2	3	0	0

데이터 수집 및 전처리

인코딩(Encoding) – 라벨 인코딩

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	남자	1	0
002	30	15000	20	여자	0	1
003	12	31	2	외계인	0	0

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	0	1	0
002	30	15000	20	1	0	1
003	12	31	2	3	0	0



- 그럼 여자보다 외계인의 영향력이 큰건가?
- 성별에 순서가 있는건가?

데이터 수집 및 전처리

인코딩(Encoding) – 원핫 인코딩

고객ID	가입일수	총 결제금액	최근 접속일	성별	알림 설정	이탈여부
001	180	200000	5	남자	1	0
002	30	15000	20	여자	0	1
003	12	31	2	외계인	0	0

- 성별을 개별 컬럼으로 만들어주고, 해당 항목에만 1을 부여
- 컬럼의 수(차원)이 너무 많아질 가능성 존재

고객ID	가입일수	총 결제금액	최근 접속일	성별_남자	성별_여자	성별_외계인	알림 설정	이탈여부
001	180	200000	5	1	0	0	1	0
002	30	15000	20	0	1	0	0	1
003	12	31	2	0	0	1	0	0

데이터 수집 및 전처리

정규화(Normalization) - MinMaxScaler

고객ID	가입일수	총 결제금액	최근 접속일	성별_남자	성별_여자	성별_외계인	알림 설정	이탈여부
001	180	200000	5	1	0	0	1	0
002	30	15000	20	0	1	0	0	1
003	12	31	2	0	0	1	0	0

- 값의 범위를 0~1로 사이로 바꾸어 주는 것
- 200000과 2의 차이가 크므로, 총 결제 금액의 영향력이 비대해 지는 것을 방지

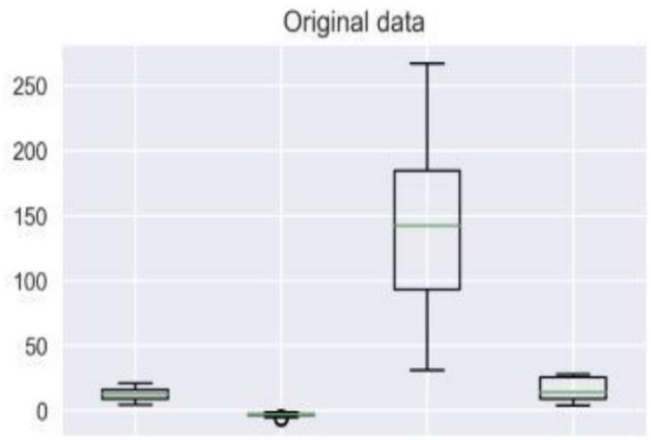
데이터 수집 및 전처리

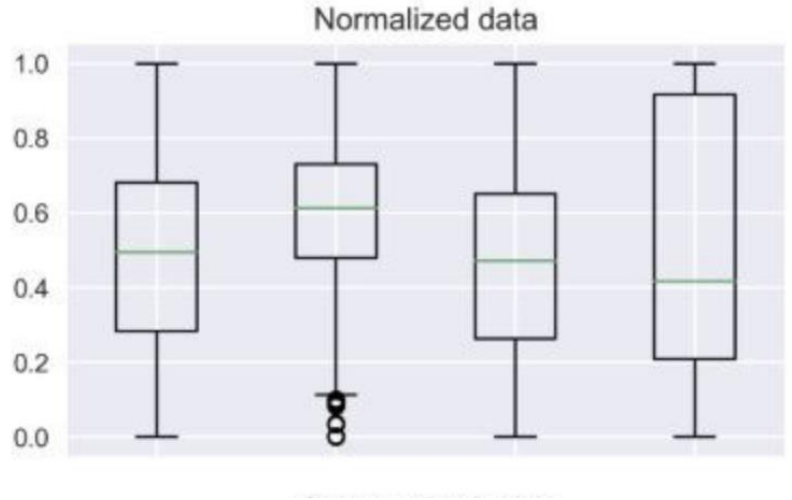
표준화(Standardization) - StandardScaler

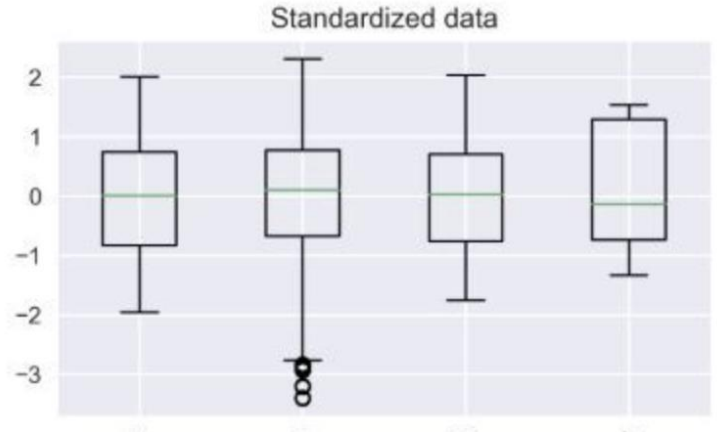
고객ID	가입일수	총 결제금액	최근 접속일	성별_남자	성별_여자	성별_외계인	알림 설정	이탈여부
001	180	200000	5	1	0	0	1	0
002	30	15000	20	0	1	0	0	1
003	12	31	2	0	0	1	0	0

- 값의 범위를 평균이 0 분산 1로 바꾸어 주는 것
- 200000과 2의 차이가 크므로, 총 결제 금액의 영향력이 비대해 지는 것을 방지
- 표준정규분포화 시켜준다!

데이터 수집 및 전처리


Original data
250
200
150
100
50
0

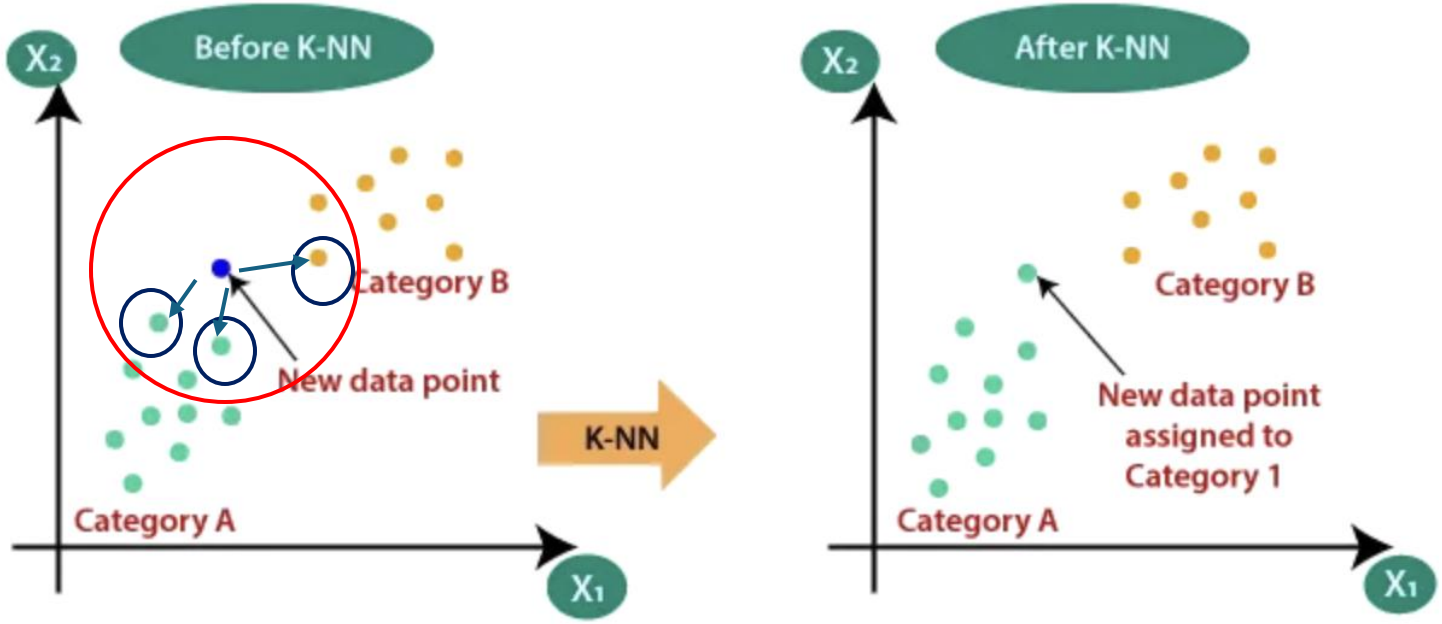

Normalized data
1.0
0.8
0.6
0.4
0.2
0.0


Standardized data
2
1
0
−1
−2
−3

모델 정의하기

KNN (K-Nearest Neighbors) 알고리즘

- K개의 가장 가까운 주변 데이터의 종류를 보고 예측 하겠다.



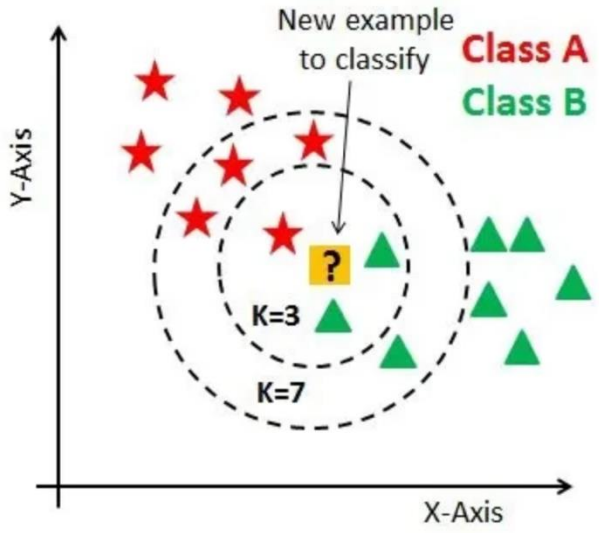
출처: MEDIUM.COM

K= 3일 경우
초록 2 노랑 1이므로
예측 데이터는 초록색이 된다.

모델 정의하기

KNN (K-Nearest Neighbors) 알고리즘

- K개의 가장 가까운 주변 데이터의 종류를 보고 예측 하겠다.



출처: MEDIUM.COM

장점

- 간단하고 쉬움
- 적은 데이터에서도 분류 가능

단점

- 차원이 늘어나면, 정확도 저하
- 이상치에 민감함

모델 정의하기

로지스틱 회귀 알고리즘

- 선형 회귀처럼 독립변수들의 선형 조합을 사용하되, 결과를 0과 1사이의 확률로 예측하는 모델

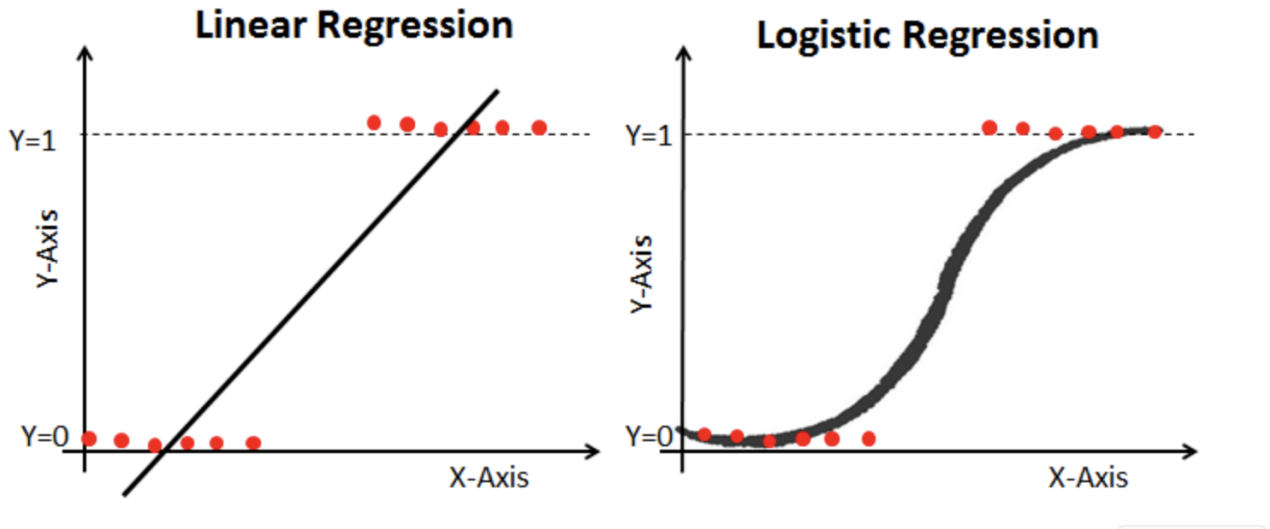
~~- 오즈를 계산해서 로짓 변환을 해서 그걸 시그모이드함수에 넣어서 다시 확률롤 출력한다...~~

- 그래서 얘가 1일 확률이 어느정도고 0일 확률은 어느정도야??

- **1일 확률이 0.5 이상 -> 1로 분류**
- **0일 확률이 0.5 이상 -> 0으로 분류**

모델 정의하기

로지스틱 회귀 알고리즘



장점

- 해석이 수월하다

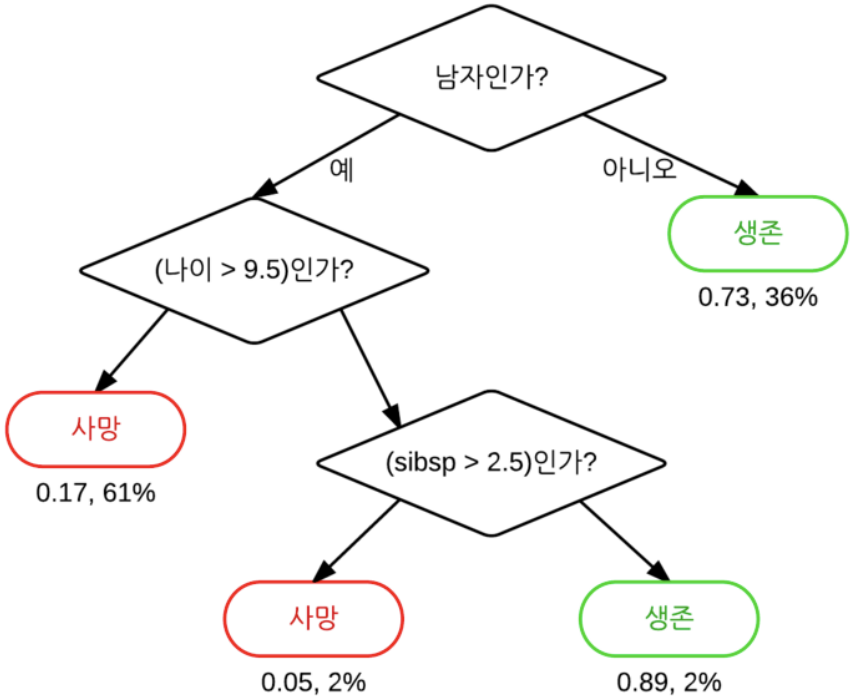
단점

- 회귀분석이므로, 분포에 대한 가정이 필요하다.

모델 정의하기

랜덤 포레스트 알고리즘 (앙상블 중 배깅)

- 의사결정 나무(Decision Tree)를 여러개 만들어, 각 나무의 결과를 합치는 방식



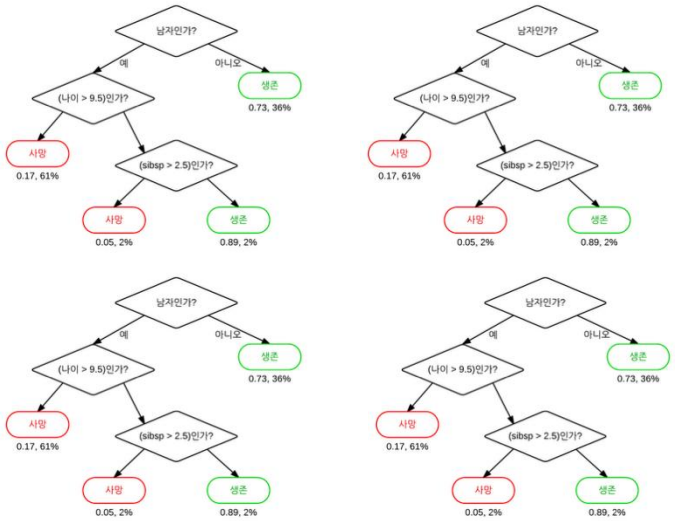
- 사망(0)과 생존(1)의 분류가 잘 되게 기준이 설정되고, 트리가 만들어짐

정보 이득이 높다
분류가 분명하게 된다

모델 정의하기

랜덤 포레스트 알고리즘 (앙상블 중 배깅)

- 의사결정 나무(Decision Tree)를 여러개 만들어, 각 나무의 결과를 합치는 방식



-표본 추출을 통해 의사결정 나무를 여러개 만들어,
해당 결과를 다수결로 판단하거나, 평균으로 합치는 방식

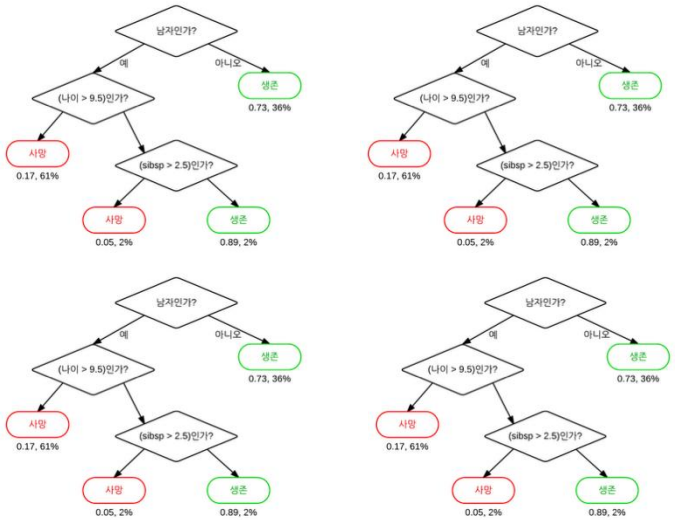
- 1번을 0이라고 분류한 의사결정 트리가 3개
- 1번을 1이라고 분류한 의사결정 트리가 1개

-> 1번은 0으로 분류함

모델 정의하기

랜덤 포레스트 알고리즘 (앙상블 중 배깅)

- 의사결정 나무(Decision Tree)를 여러개 만들어, 각 나무의 결과를 합치는 방식



장점

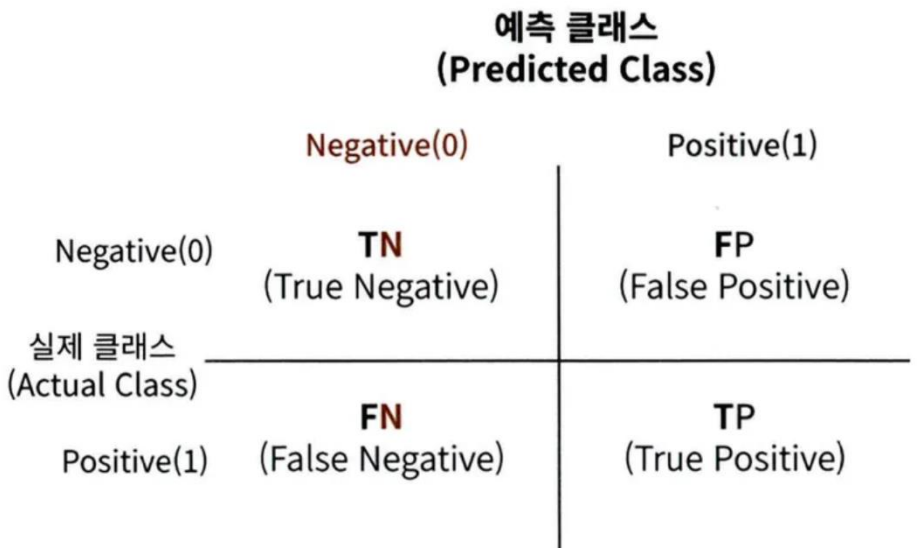
- 쉬움!
- 대용량 데이터 처리가 쉬움
- 스케일링이 필요없음

단점

- 느림!
- 앙상블이므로, 해석이 어려움!

평가하기

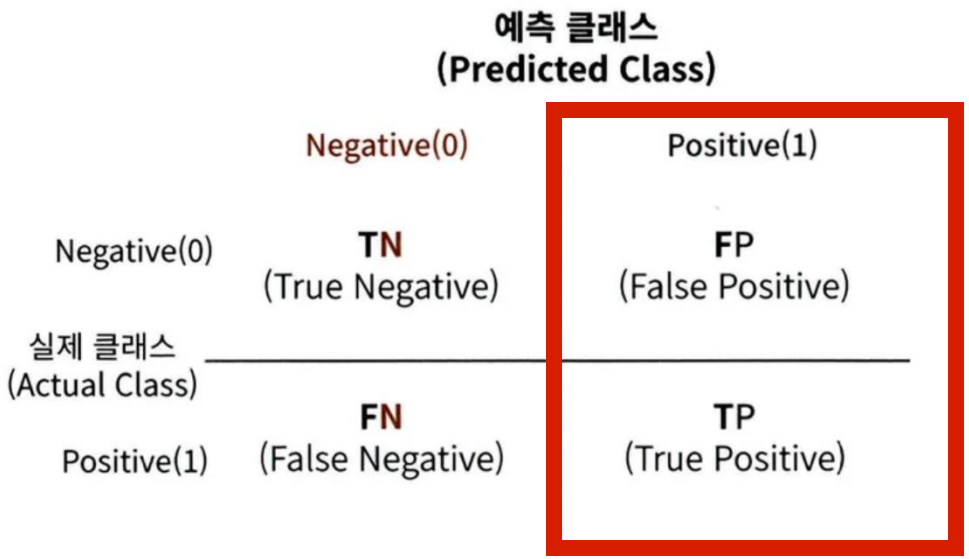
혼동행렬 (Confusion-Matrix)



- TN : N(0)으로 예측했는데 진짜 N임
- FP : P(1)라고 예측했는데 사실 N임
- FN : N(0)으로 예측했는데 사실 P임
- TP : P(1)라고 예측했는데 진짜 P임

평가하기(추가예정)

정밀도



- TP / FP + TP
- P라고 예측한 애들 중 진짜 P인경우

오늘의 실습 데이터 Titanic

컬럼명	설명	자료형	예시
PassengerId	탑승객 고유 ID	int	892
Survived	생존 여부 (타깃 변수)0 = 사망, 1 = 생존	int (0 or 1)	1
Pclass	객실 등급 (1=1등석, 2=2등석, 3=3등석)	int	3
Name	이름	string	“Braund, Mr. Owen Harris”
Sex	성별	string	“male” / “female”
Age	나이 (결측값 존재)	float	22.0
SibSp	함께 탑승한 형제/배우자 수	int	1
Parch	함께 탑승한 부모/자녀 수	int	0
Ticket	티켓 번호	string	“A/5 21171”
Fare	운임 요금	float	7.25
Cabin	객실 번호 (결측값 다수 존재)	string	“C85”
Embarked	탑승 항구 (C = Cherbourg, Q = Queenstown, S = Southampton)	string	“S”

Task #04 :: EDA 및 시각화

다음의 항목들을 Point Plot 을 이용하여 시각화 하시오.

(1) 연월 별 총 매출

(2) 연월 별 총 판매량

*정답과 관련없는 플롯 예시입니다.