2025.02.01

16기 신입교육세션

3주차 - 회귀 모델링

B.a.f

CONTENTS

01 모델링 과제 피드백

1. RandomForest
2. Gradient Boosting
3. Logistic Regression

B.a.f

스케일링

RandomForest, Gradient Boosting 에서는 스케일링이 필요하지 않다!

RandomForest

여러 개의 결정트리 학습시켜서
예측하는 앙상블 모델

결정트리 : 데이터를 분할하는 방식

모델링에서 특성(x)들의 스케일은
영향을 미치지 않는다
(크기, 단위 중요하지 않음)

Gradient Boosting

여러 개의 약한 학습기 순차적으로
학습시켜 오차 줄여가는 모델

결정트리 : 데이터를 분할하는 방식

모델링에서 특성(x)들의 스케일은
영향을 미치지 않는다
(크기, 단위 중요하지 않음)

변수 스케일링 (Feature Scaling)

- 변수들의 범위(scale)를 통일시켜주는 작업
- 대부분의 모델(Tree 기반 모델 제외)에서 필요한 단계
  - (ex. 특정 회귀 모형, 거리 기반 모델 등)
  - 특정 변수의 값의 범위가 클수록 영향력이 큰 변수라고 잘못 인식

(주의 1) 스케일링된 데이터로 모델 학습 시, 예측값은 원래 스케일로 변환하기

(주의 2) train, valid, test 셋 분리한 후, 각각 스케일링 진행하기

(주의 3) train 셋에 적용한 scaler로 valid, test셋에도 적용하기

하이퍼파라미터 튜닝

하이퍼파라미터 : 모델의 동작 및 학습 과정을 제어하는 매개변수

쉽게 말해 우리가 직접 조정할 수 있는 값들

왜 해야할까?

같은 모델을 사용해도 하이퍼파라미터 값들에 따라 모델 성능 달라짐!
따라서 모델을 최적화시키기 위해 튜닝은 필수적

튜닝 방법

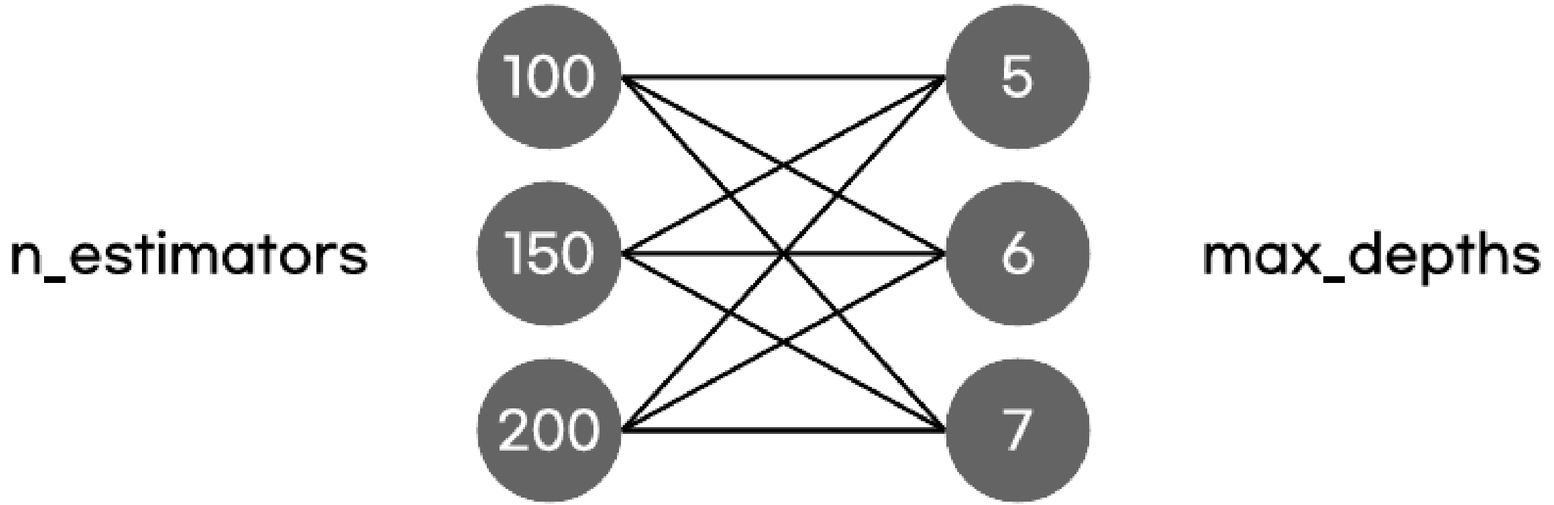
- Grid Search
- Manual Search

하이퍼파라미터 튜닝

Grid Search

우리가 지정한 하이퍼파라미터들의 후보군들의 조합 중 Best 조합을 선별

(예시)



라이브러리가 존재하여 사용이 간편하지만

조합이 늘어갈 때마다 시간 소요가 크다는 단점

하이퍼파라미터 튜닝

Manual Search

사용자의 직관이나 경험으로 하이퍼파라미터를 조정하여 사용

1. 먼저 임의의 값을 대입해 결과를 살핀다.
2. 그 결과에 따라 값을 조정해가며 변화를 관찰한다.
3. 값을 하나씩 대입해보고 조정하는 과정을 반복하면서 최적의 값을 찾는다.

매우 단순하고 쉬운 방법이지만
그만큼 최적의 파라미터 값들과 조합을 찾는 것이 힘들다

최종 변수 선택

변수를 선택해야하는 이유 ?

1. 종속 변수 예측에 영향을 주지 않는 경우
2. 독립변수들끼리 다중공선성이 발생한 경우

» 정확한 예측을 위해 적절한 변수 선택 또는 PCA 같은 차원 축소가 필요

(1) 변수 선택 방법

- 전진선택법 (forward selection)
- 후진선택법 (backward selection)
- 단계선택법 (stepwise method)
- 변수중요도를 보고 판단
- 변수의 정의를 보고 판단

(2) 차원축소

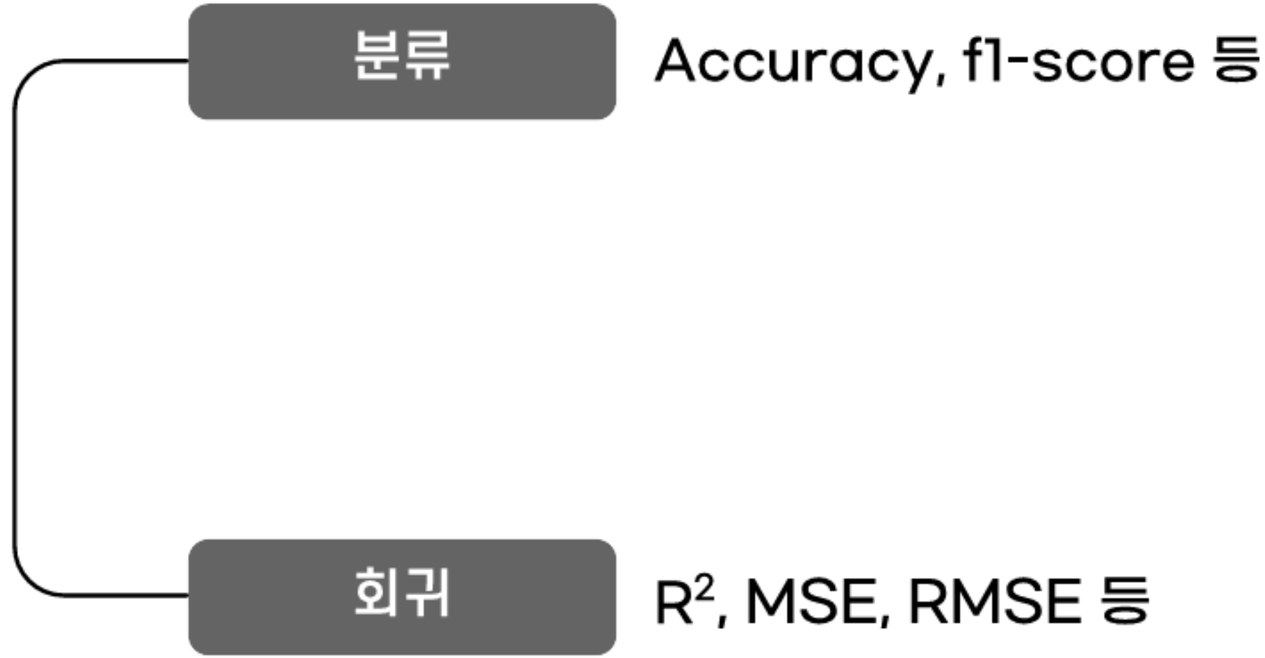
- PCA
- FA
- MDS

평가지표

우리는 Test셋의 종속 변수 값을 알 수 없는 경우가 대부분
따라서 valid셋을 활용하여 성능 개선

(모델링 순서)

1. 모델 선언
2. 모델 학습
3. valid 셋을 활용하여 성능 확인
4. 3번 과정을 반복하여 모델 최적화
5. 최종 모델로 test셋을 예측하며 마무리

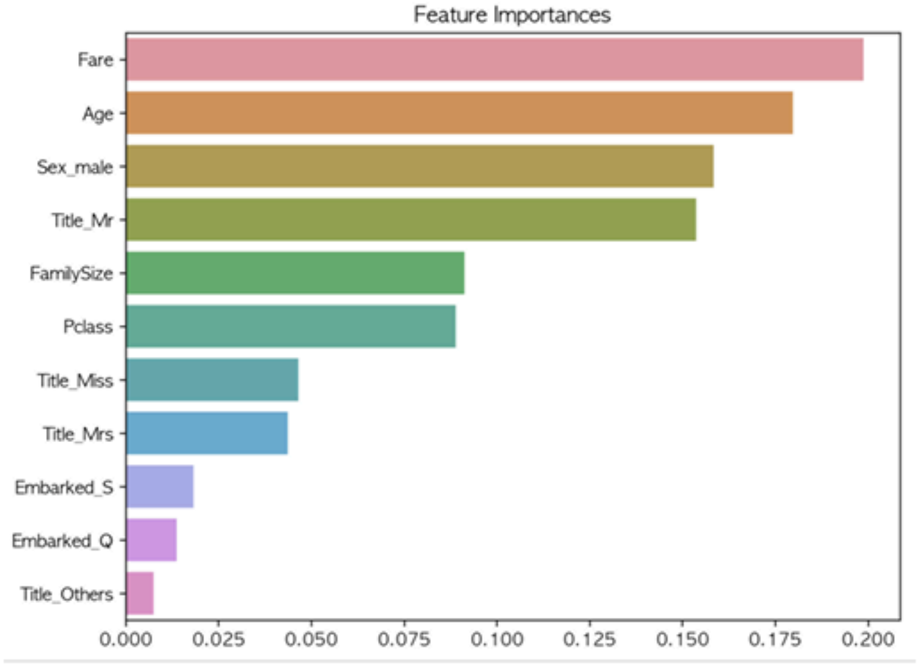


**종속변수에 스케일링 시 값 복원 후 평가지표 확인

변수 중요도

머신러닝 모델

: 머신러닝 모델들은 변수 중요도를 알 수 있음 ** Linear Regression - 회귀 계수와 변수가 유의한지 확인
ex) RandomForest, XGBoost, Decision Tree, Lgbm 등



변수중요도를 보고 해석 가능
분석 목표에 맞는 기대효과와 해결방안 생각 가능
데이터 분석가에게 중요한 역할

02 회귀 모델

B.a.f

회귀 문제

regression

- 다음 달 전력사용량은 얼마일까?
- 설날 선물 수요량은 얼마일까 ?
- 영화를 보러 몇 명이 올까 ?

☑ 과제로 나가는 따릉이는 회귀 문제

모델 종류

- Linear Regression, SVR, 랜덤포레스트, XGBoost 등

03 회귀 모델 학습

1 . 선 형 회 귀 모 델

2 . 다 항 회 귀

3 . 규 제 선 형 회 귀

B . a . f

선형회귀모델

- 독립변수 (X) 와 종속변수(Y)의 관계를 가장 잘 나타낼 수 있는 선형식을 모델링/학습하는 알고리즘

선형(Linear)이란?

$$f(u+v) = f(u) + f(v)$$

$$f(c \times u) = c \times f(u)$$

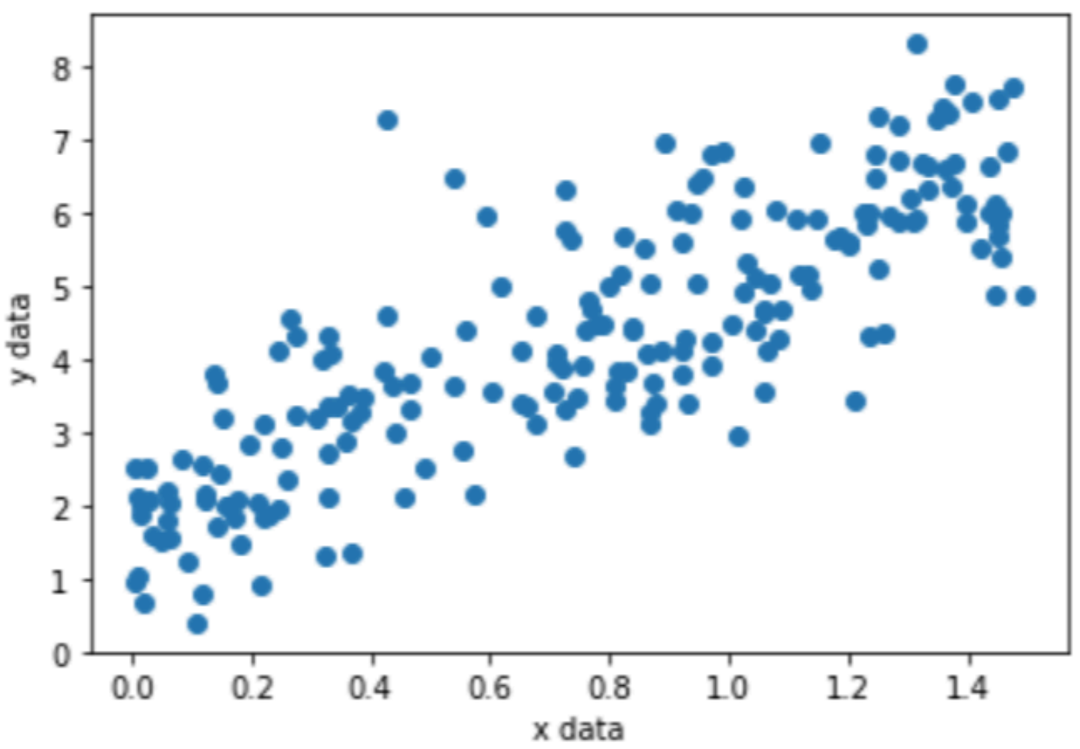
$(u, v : variables, c : constant)$

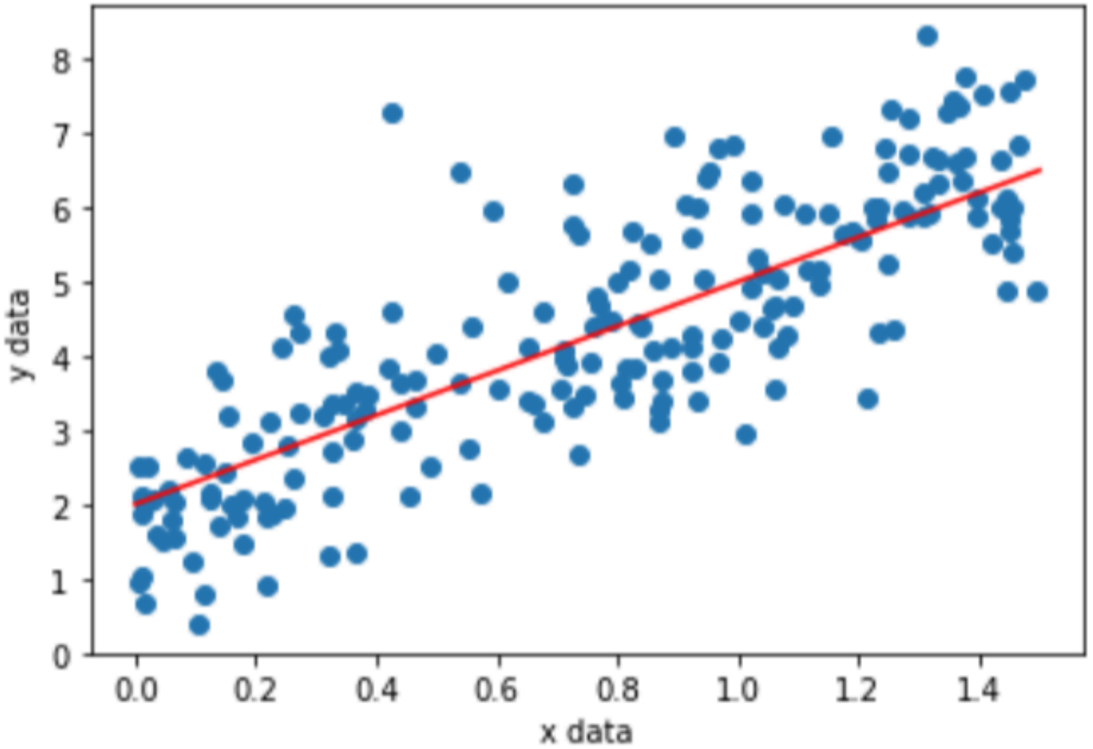
선형회귀모델

$$y = wx + b$$

$$y = w1x1 + w2x2 + b$$

$$y = w1x1 + w2x2 + \ldots + wnxn + b$$



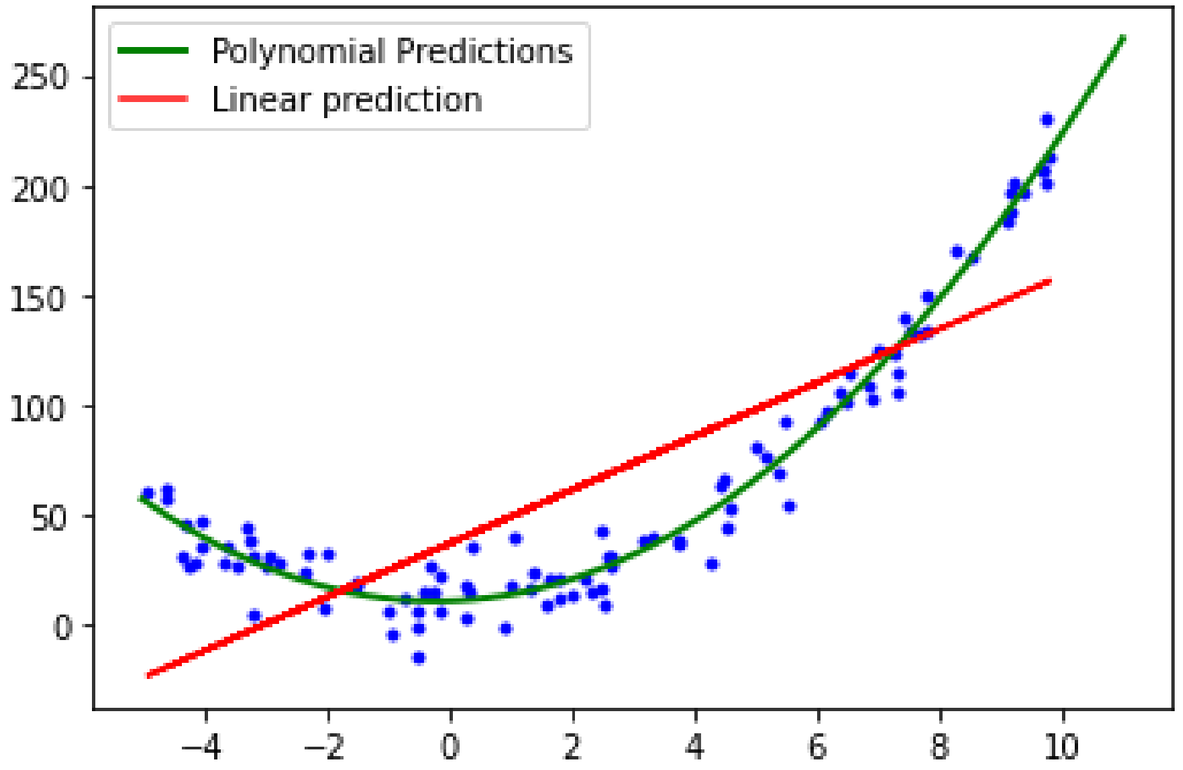


컴퓨터는 우리처럼 직관적으로 선을 찾아낼 수 없으므로

선형 회귀 모델을 학습시켜 데이터를 가장 잘 표현하는 식을 찾는 것 !

다항회귀분석 (Polynominal Regression)

- 독립변수 (X) 와 종속변수(Y)의 관계가 곡선 형태로 되 있는 경우



다항회귀분석 (Polynominal Regression)

1. **Lasso (L1) 규제**
   a. 가중치의 절대값 합에 비례하는 패널티를 부과
   b. 불필요한 가중치를 0으로 만듬 (필요없는 변수 제거 가능)
   c. 특성(X)이 많은 데이터셋에 유용
2. **Ridge (L2) 규제**
   a. 가중치의 제곱합에 비례하는 패널티를 부과
   b. 가중치를 0에 가깝게 만들지만 완전히 0으로 만들지는 않음
   c. 모든 특성이 적절하게 고려되어 특성 간의 상관관계가 높은 경우에 유용 (다중공선성 문제 완화)
3. **Elastic Net : L1, L2를 함께 사용하는 방법**

04. 회귀 모델 평가

B.a.f

회귀 문제

R^2: 결정계수

- 실제 관측값의 분산대비 예측값의 분산을 계산
- 0~1까지 나타낼 수 있고, 1에 가까울수록 설명력을 높게 가지는 모델

MSE (Mean Squared Error)

- 종속 변수와 단위가 다름
- 에러를 제곱하기 때문에 이상치에 민감

MAE (Mean Absolute Error)

- Error에 절대값을 취해 Error의 크기를 그대로 반영
- 예측변수와 단위가 같고 직관적임
- MSE보다 이상치에 robust함

RMSE

- MSE에 루트를 취한 값
- 종속 변수와 단위가 같음

수식

$$R^2 = \frac{\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=0}^{n}(Y_i - \bar{Y}_i)^2} = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|Y_i - \hat{Y}_i|$$

05. 실습 & 과제

B . a . f

모델링2 과제

(1) 타이타닉 생존자 예측하기

- 세션 때 배운 내용을 최대한 활용하여 “타이타닉 생존자“ 예측 모델링 진행
- Titanic-dataset.csv

(2) 따릉이 대여량 예측하기

- 세션 때 배운 내용을 최대한 활용하여 따릉이 대여량 예측하기
- SeoulBikeData_NaN.csv

*데이터는 노션에서 다운받으실 수 있습니다.

*타이타닉 생존자 예측.ipynb , *따릉이 대여량 예측.ipynb 각각 깃허브 3주차 과제로 올려주세요.

감사합니다

B.a.f