# 기계학습 실습

## 결과 보고서

2024년 12월

건국대학교 정보통신대학원

정보보안학과

박성철(202376629)

## ■ 개 요

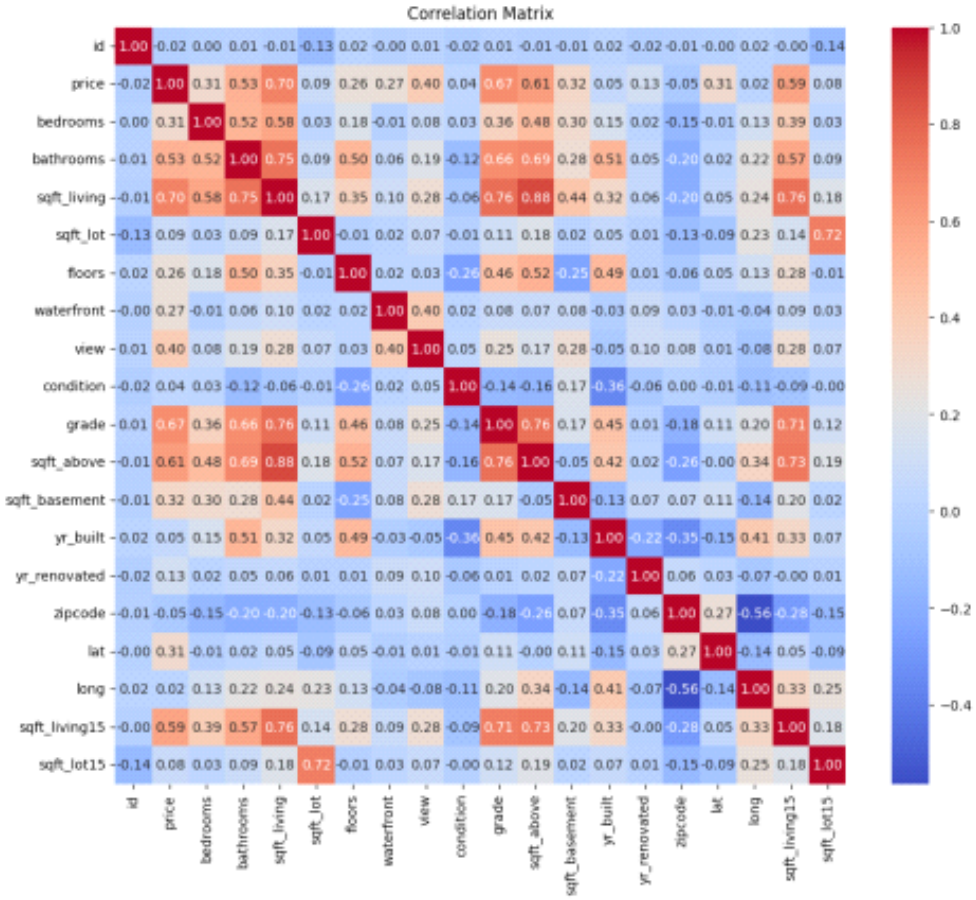
- 데이터(kc_house_data.csv)에서 다양한 Feature(예: sqft_living, grade, bathrooms 등)을 사용하여 주택가격(price) 예측에 대한 기계학습 모델 성능 비교.

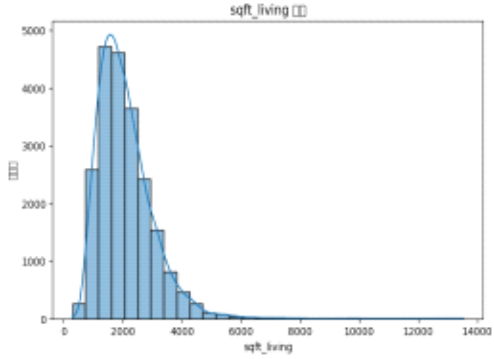
## 1. Kaggle 데이터셋 로드

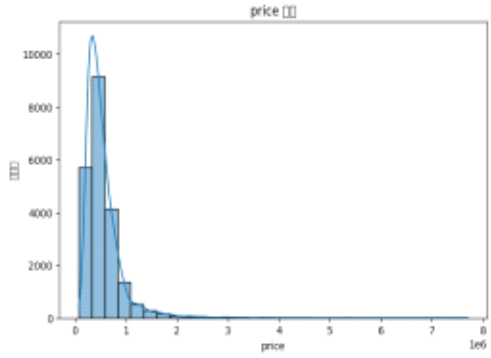
- Kaggle에서 kc_house_data.csv 데이터셋을 Colab에 로드하고 기본 정보를 확인(실습코드 참조)

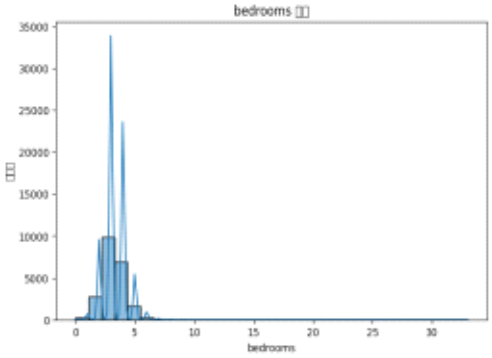
## 2. 데이터 시각화

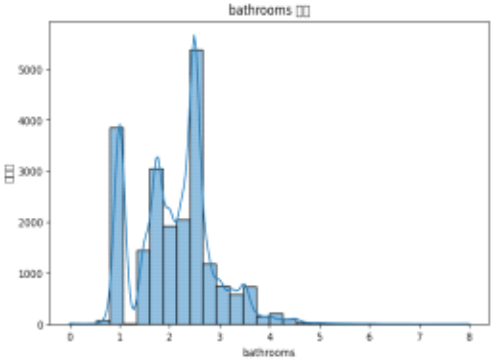
- 주요 특성(sqft_living, price 등)의 분포 및 상관관계를 확인하기 위해 히스토그램, 상관관계 히트맵, 산점도를 시각화.

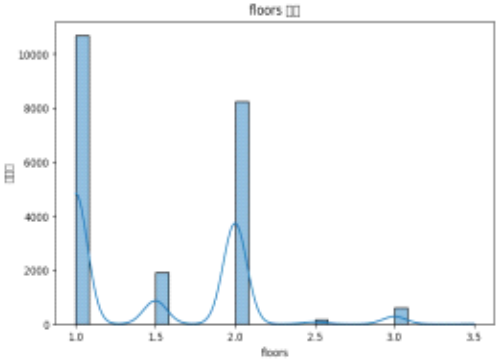



sqft_living 분포
sqft_living


price 분포
price
1e6

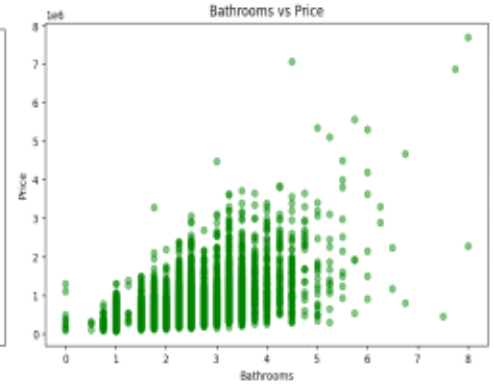

bedrooms 분포
bedrooms

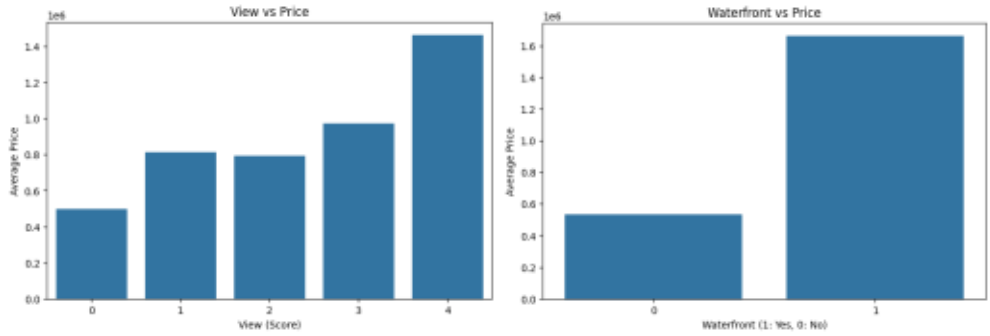

bathrooms 분포
bathrooms


floors 분포
floors


Sqft Living vs Price
1e6
Price
Sqft Living (분포 확인)

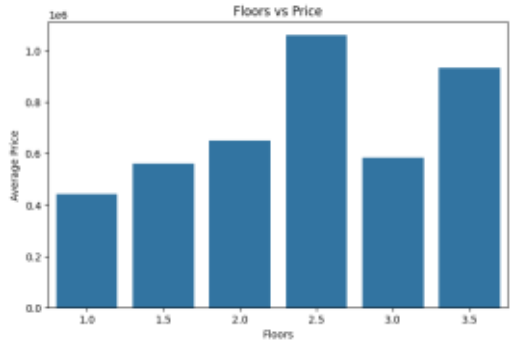

Bedrooms vs Price
1e6
Price
Bedrooms

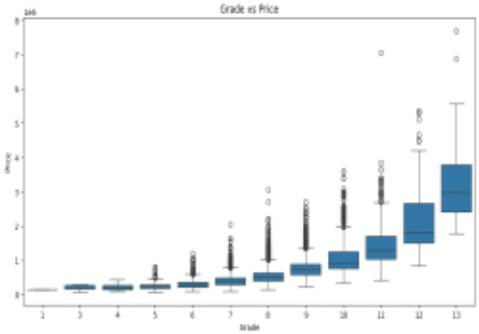

Bathrooms vs Price
1e6
Price
Bathrooms











## 3. 데이터 전처리 및 데이터 분할

- 모델 학습을 위해 데이터를 전처리 후 훈련 데이터와 테스트 데이터를 8:2 비율로 분할.(실습코드 참조)

## 4. 모델 학습

- 랜덤 포레스트, 선형 회귀, XGBoost모델로 학습.(실습코드 참조)

## 5. 모델 평가

- 테스트 데이터를 사용하여 모델을 $R^2$ Score를 이용하여 평가
- 모델 평가 결과

| 구 분 | 랜덤 포레스트 | 선형 회귀 | XGBoost |
|---|---|---|---|
| $R^2$ Score | 0.8540 | 0.7012 | 0.8735 |

## 6. 모델 최적화 및 최적화된 모델로 재학습

- GridSearchCV를 사용하여 최적의 하이퍼파라미터를 적용 후 모델을 재학습(실습코드 참조)

## 7. 최적화된 모델 평가

- 테스트 데이터를 사용하여 최적화된 모델의 성능을 평가.
- 최적화 모델의 최종 평가 결과

| 구 분 | 랜덤 포레스트 | 선형 회귀 | XGBoost |
|---|---|---|---|
| 초기모델 $R^2$ Score | 0.8540 | 0.7012 | 0.8735 |
| 최적화모델 $R^2$ Score | 0.8570 | 0.7012 | 0.8685 |

## 8. 최종 분석 결과

- 최적화된 랜덤 포레스트 모델의 $R^2$ Score는 0.857으로 기존 모델(0.854) 대비 성능이 약간 향상되었으나 XGBoost 모델이 주택가격 예측에 가장 적합하다 판단됨.

## 9. 첨부

- 실습코드 : 기계학습실습