# 머신러닝 기반 연구방법로

## 케글 데이터셋 기반 실습 과제

건국대학교 정보통신대학원 인공지능전공

202476713

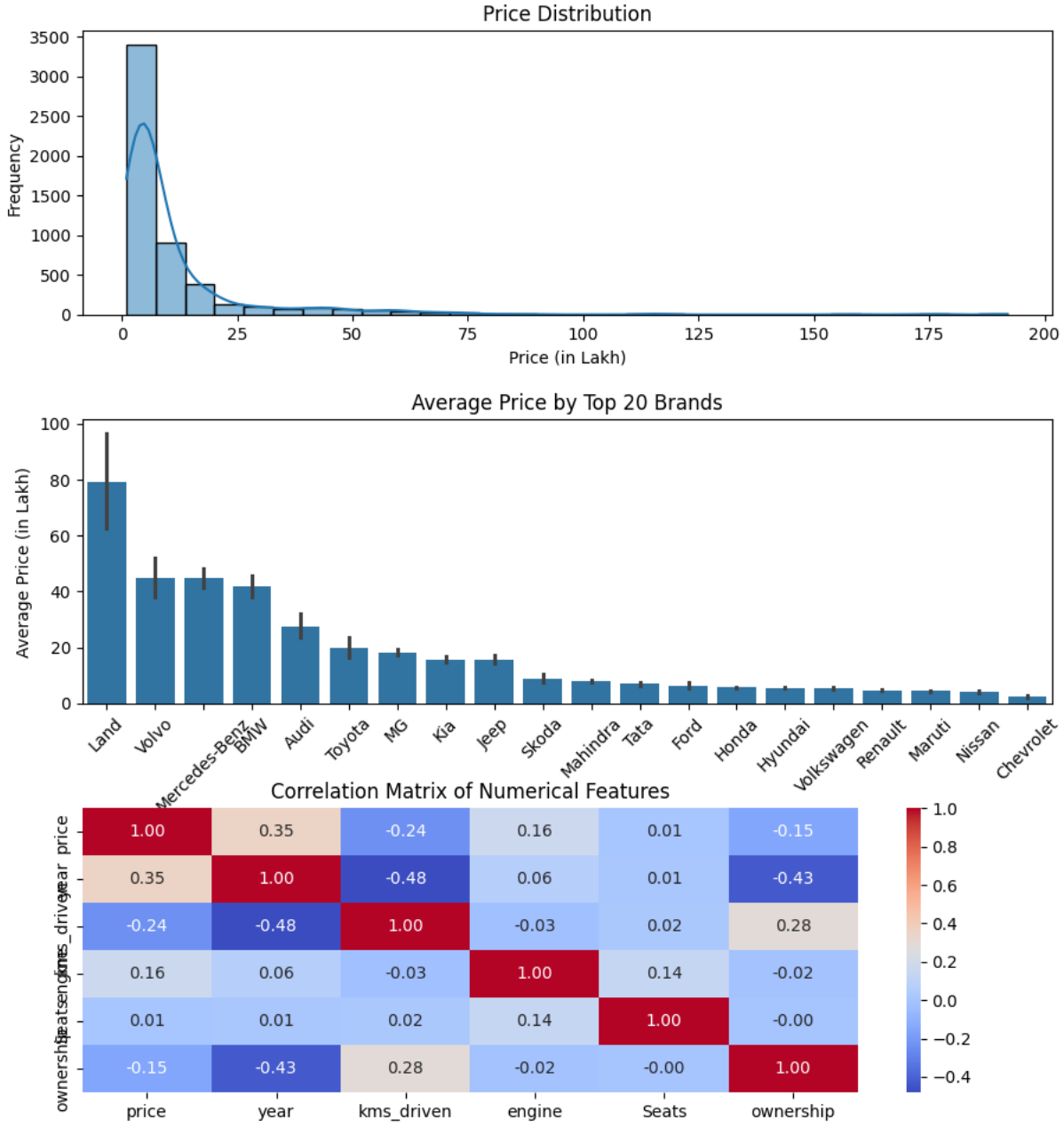
고한준

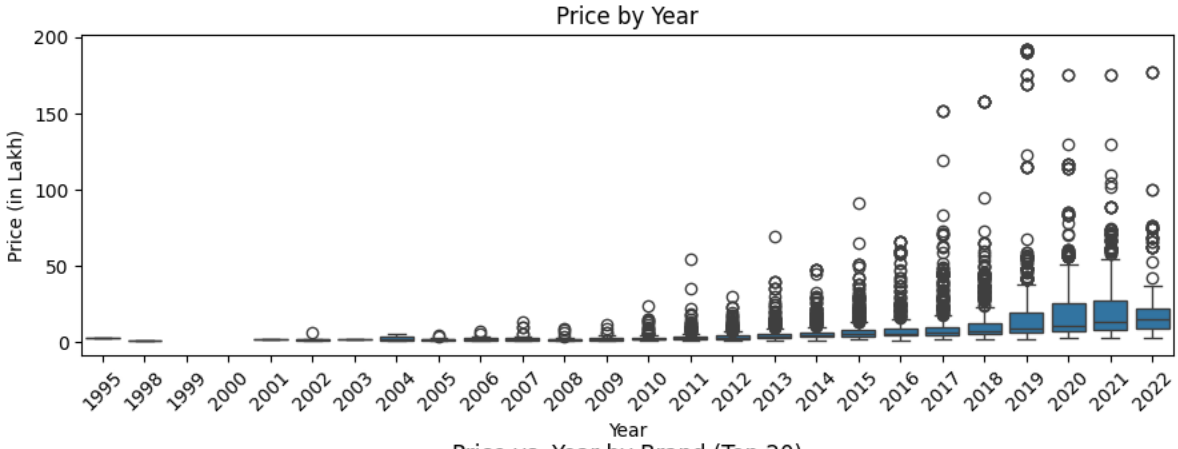
# 데이터 개요

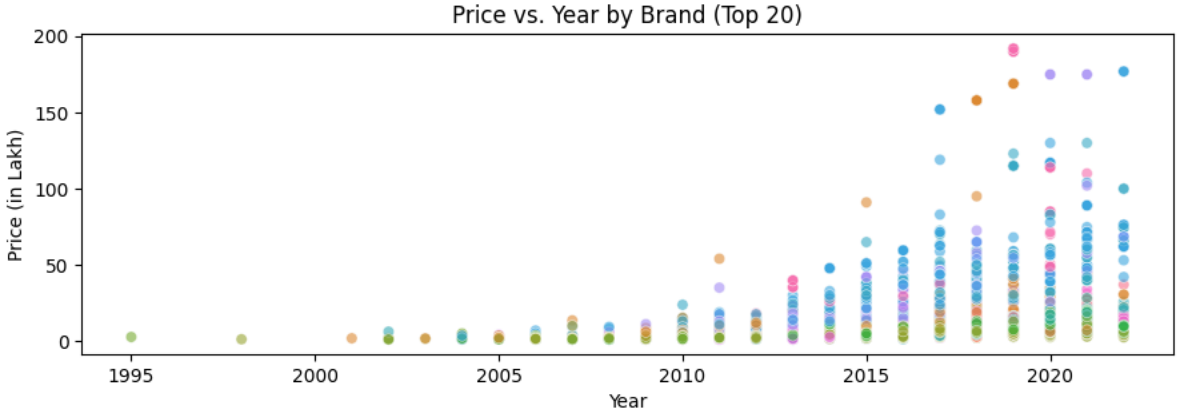
해당 데이터셋은 중고차 가격 예측을 위한 기반 데이터로 데어터는 Index, car_name, car_price, kms_driven, fuel_type, transmission, ownership, manufacture, engine, seats 10개의 칼럼으로 구성되어 있고, 총 5511개의 데이터가 존재한다.

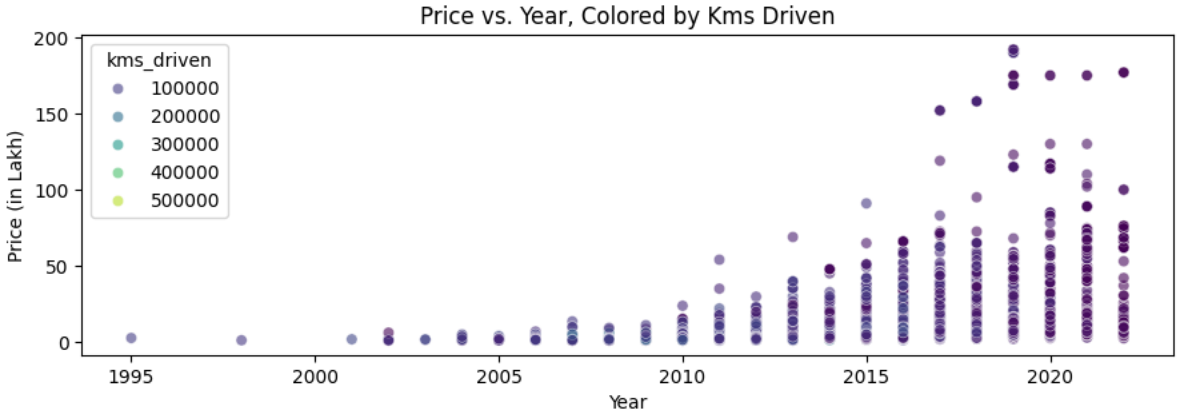
https://www.kaggle.com/datasets/milanvaddoriya/old-car-price-prediction/data

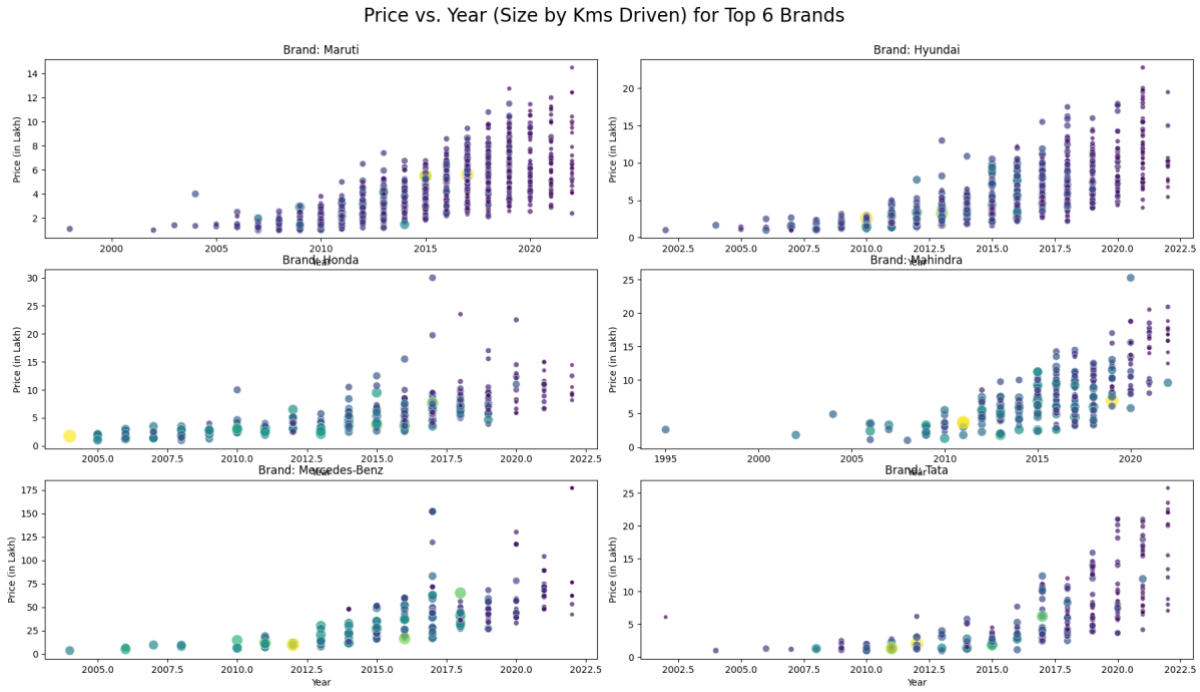
# 시각화




Price by Year
Price (in Lakh)
Year
0
50
100
150
200
1995
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022


Price vs. Year by Brand (Top 20)
Price (in Lakh)
Year
0
50
100
150
200
1995
2000
2005
2010
2015
2020


Price vs. Year, Colored by Kms Driven
kms_driven
100000
200000
300000
400000
500000
Price (in Lakh)
Year
0
50
100
150
200
1995
2000
2005
2010
2015
2020




# 모델 선정, 학습, 결과

가격 예측 모델이기 때문에 회귀 모델 중에 선택하였다.

선정한 모델은 Linear Regression, Decision Tree, Random Forest 세 모델로 선정하여 각각 모델로 학습을 진행하였다.

학습과 테스트 데이터는 8:2로 나눠서 진행하였다.

| 모델 | MAE | RMSE | R2 Score |
|---|---|---|---|
| Linear Regression | 5.617188 | 13.612262 | 0.526988 |
| Decision Tree | 3.359249 | 9.883336 | 0.750645 |
| Random Forest | 2.927580 | 7.489171 | 0.856821 |

** MAE/RMSE: 낮을수록 좋습니다.

** R2 Score: 높을수록 좋습니다. (1에 가까울수록 모델이 데이터를 잘 설명함을 의미)

# 최적화 및 결과

최적화는 하이퍼파라미터를 사용하여 최적화 하였다.

선형회귀는 하이퍼파라미터 없이 진행하였고, 의사결정트리와 랜덤포레스트는 아래와 같은 항목들을 하이퍼파라미터로 적용하여 최적의 파라미터를 생성하였다.

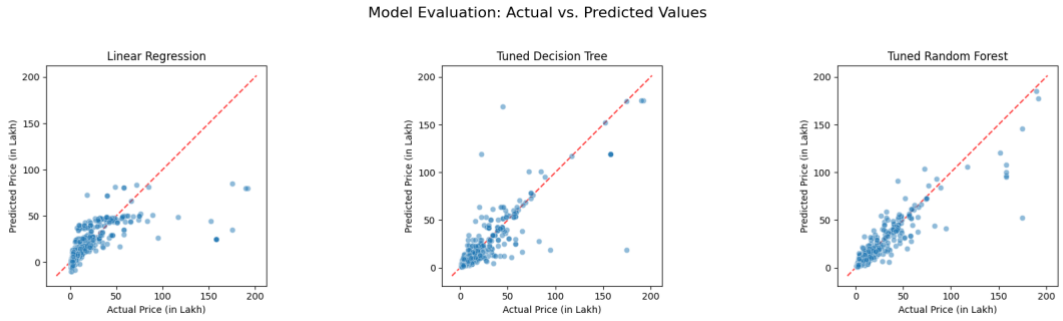
```
'n_estimators': [100, 200],
'max_depth': [10, 20, None],
'min_samples_split': [2, 5],
'min_samples_leaf': [1, 2]
```

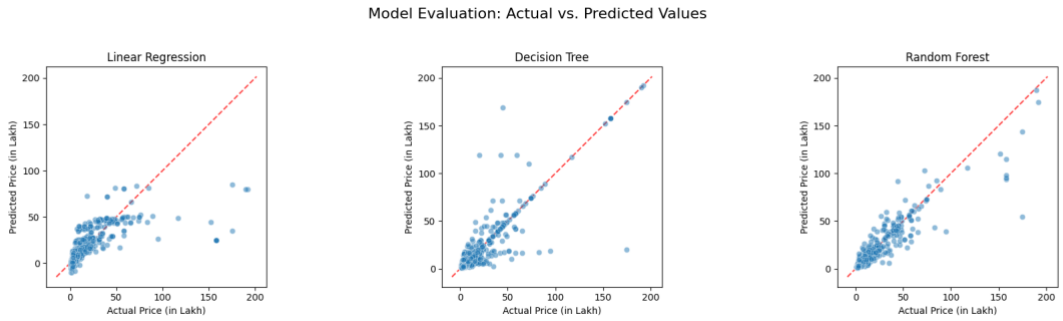
최적화 결과 다음과 같은 파라미터를 도출하였다.

의사결정트리 : {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 10}

램덤포레스트 : {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

| 모델 | MAE | RMSE | R2 Score |
|---|---|---|---|
| Linear Regression | 5.617188 | 13.612262 | 0.526988 |
| Decision Tree | 3.482972 | 9.688931 | 0.760358 |
| Random Forest | 2.903093 | 7.475235 | 0.857353 |




# 결론

25이하의 가격에 대부분의 차량들이 분포하는 경향과 최신 차량의 경우, 기준치를 벗어나는 이상치 가격들이 많이 분포하고 있어서 모델의 학습 결과를 보면 높은 가격에 위치할 경우, 낮은 가격보다 더 낮은 정확도를 보여주는 걸 알 수 있었다.

이를 하이퍼파라미터 최적화를 통해서 개선하려고 하였으나, 의사결정트리의 경우, 높은

가격의 일부 정확도는 높아졌으나 중간 가격의 정확도가 조금 낮아지는 경향을 보였으며, 그래도 여전히 램덤포레스트의 성능이 높게 나왔다.

1차

| 모델 | MAE | RMSE | R2 Score |
|---|---|---|---|
| Linear Regression | 5.617188 | 13.612262 | 0.526988 |
| Decision Tree | 3.359249 | 9.883336 | 0.750645 |
| Random Forest | 2.927580 | 7.489171 | 0.856821 |

최적화

| 모델 | MAE | RMSE | R2 Score |
|---|---|---|---|
| Linear Regression | 5.617188 | 13.612262 | 0.526988 |
| Decision Tree | 3.482972 | 9.688931 | 0.760358 |
| Random Forest | 2.903093 | 7.475235 | 0.857353 |

# 모델 추가

결론에서 높은 가격에서의 정확도가 낮게 나온 것을 바탕으로 모델 성능을 증가시킬 새로운 방안을 제시하여 시도해 보았다.

이전에는 단순히 전체 데이터셋을 회귀모델로 한번에 학습하였지만, 이번에는 가격 100을 기준으로 높은 가격대와 낮은 가격대로 분리하여 두개의 회귀 모델을 통한 방법을 생각하였습니다.

이를 위해 가격 100을 기준으로 높은 가격과 낮은 가격을 분류하는 분류 모델을 먼저 학습시키고 이 모델을 통해서 예측된 가격 분류를 기반으로 다음 높은 가격 회귀 모델과 낮은 가격 회귀 모델을 통해서 최종 가격을 예측하는 모델을 설계하였습니다.

모델은 Two-stage-model 이라고 하였습니다.(분류모델 회귀모델 모두 랜덤포레스트 사용)

| 모델 | MAE | RMSE | R2 Score |
|---|---|---|---|
| Two-stage-model | 3.024058 | 10.913434 | 0.728858 |
| 높은 가격 모델 | 5.824545 | 8.091278 | 0.894624 |
| 낮은 가격 모델 | 2.312256 | 4.675371 | 0.873445 |

과적으로 좋은 성능을 가지지 못하였으나, 낮은 가격 모델과 높은 가격 모델 각각의 성능은 더 좋아진 것을 보아서 결국 앞단에 분류 모델에서의 오차가 더해져서 더 낮은 결과로 나온 것을 예측할 수 있었습니다.