FLIP ROBO

Car Price Prediction Project

Submitted by:

DEVANSH

PALIWAL

INTRODUCTION

- Business Problem Framing

  With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make carprice valuation model.

- Conceptual Background of the Domain Problem

  With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The data would contains both numeric and text data. The test data will be encoded and used along with numerical data to feed into the model.

- Data Sources and their formats

  The data is scraped from OLX and Cardekho sites. There are 10 columns in the dataset. The description of each of the column is given below:

  - “brand”: Brand of the car.
  - “model”: Model of the car.
  - “variant”: Variant of the car.
  - “manufactured_year”: The manufactured year of the car.
  - “age_of_car”: Derived from the manufactured_year of the car
  - “kilometers_driven”: The distance for which the car has been driven since it was bought.
  - “fuel_type”: The fuel type of the car.
  - “owners” : The number of owners of the car.
  - “location”: The location where the car is being sold.
  - “price”: The selling price of the car.

- Data Preprocessing Done
  - Imputed missing values.
  - Cleaned the data to remove unwanted symbols like currency symbols.
  - Derived the age of the car from manufactured_year.
  - Converted the price to show in the same range.
  - Removed ‘km’/’kms’ from kilometers_driven.
  - Used 1,2,3 and 4 in owners feature. 4 represents 4 or 4+ owners.
  - Corrected data in fuel_type to be grouped under the main categories.

- Data Inputs- Logic- Output Relationships

  The input data consists of int values which are the encoded text values using get_dummies() from Pandas. The input data also contains float values.

  The model approximates the function between the input and the output.

- Hardware and Software Requirements and Tools Used

  1. Google Colab
  2. SKLEARN
  3. MATPLOTLIB
  4. PANDAS
  5. NUMPY
  7. glob

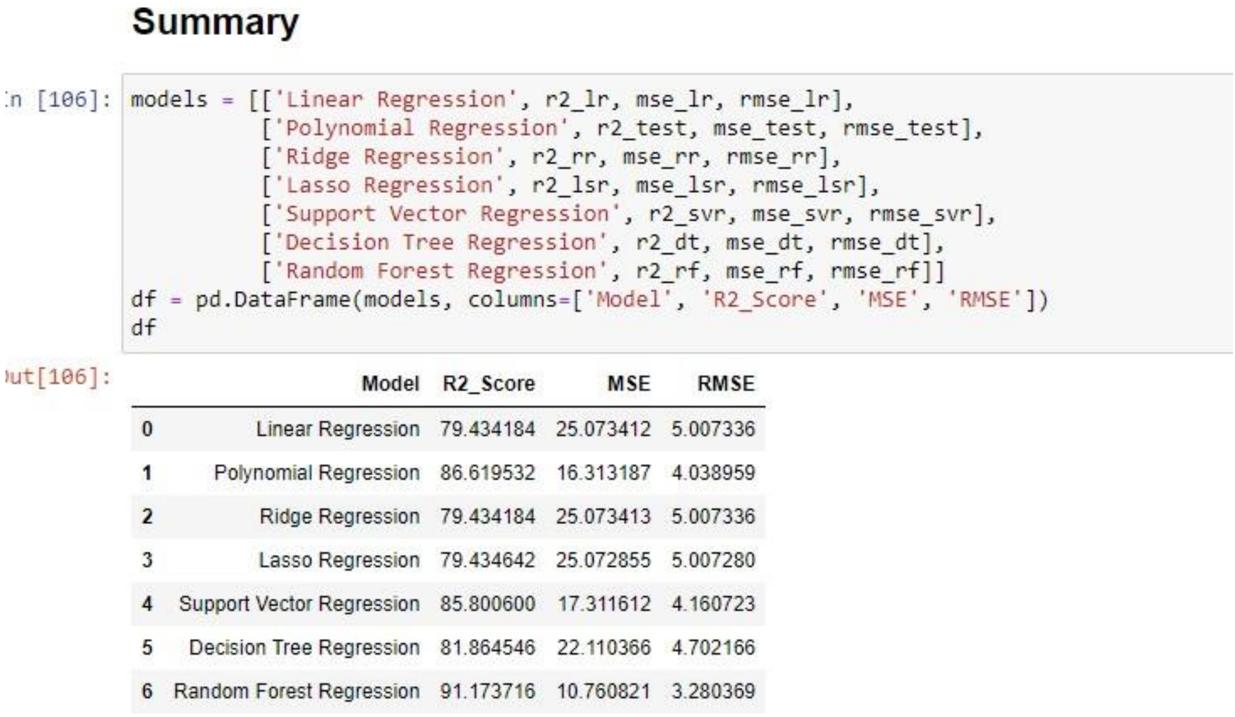
Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  - Clean the dataset.

- Encoding the data to get numerical input data.
- Compare different models and identify the suitable model.
- R2 score is used as the primary evaluation metric.
- MSE and RMSE are used as secondary metrics.

- Testing of Identified Approaches (Algorithms)
    - DecisionTreeRegressor
    - LinearRegression
    - Ridge
    - SVR
    - RandomForestRegressor
    - AdaBoostRegressor
    - GradientBoostingRegressor

- Run and Evaluate selected models



Summary

```
models = [['Linear Regression', r2_lr, mse_lr, rmse_lr],
          ['Polynomial Regression', r2_test, mse_test, rmse_test],
          ['Ridge Regression', r2_rr, mse_rr, rmse_rr],
          ['Lasso Regression', r2_lsr, mse_lsr, rmse_lsr],
          ['Support Vector Regression', r2_svr, mse_svr, rmse_svr],
          ['Decision Tree Regression', r2_dt, mse_dt, rmse_dt],
          ['Random Forest Regression', r2_rf, mse_rf, rmse_rf]]
df = pd.DataFrame(models, columns=['Model', 'R2_Score', 'MSE', 'RMSE'])
df
```

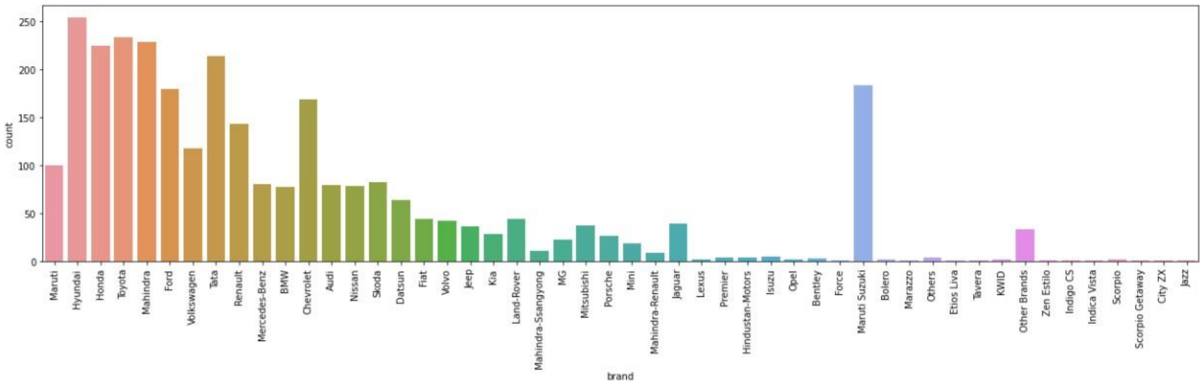
	Model	R2_Score	MSE	RMSE
0	Linear Regression	79.434184	25.073412	5.007336
1	Polynomial Regression	86.619532	16.313187	4.038959
2	Ridge Regression	79.434184	25.073413	5.007336
3	Lasso Regression	79.434642	25.072855	5.007280
4	Support Vector Regression	85.800600	17.311612	4.160723
5	Decision Tree Regression	81.864546	22.110366	4.702166
6	Random Forest Regression	91.173716	10.760821	3.280369

Observations:

- Random Forest is giving a very good score.

- Key Metrics for success in solving problem under consideration
  - R2 Score is used as the primary key metric for evaluation.
  - MSE and RMSE are used as secondary metrics.
- Visualizations

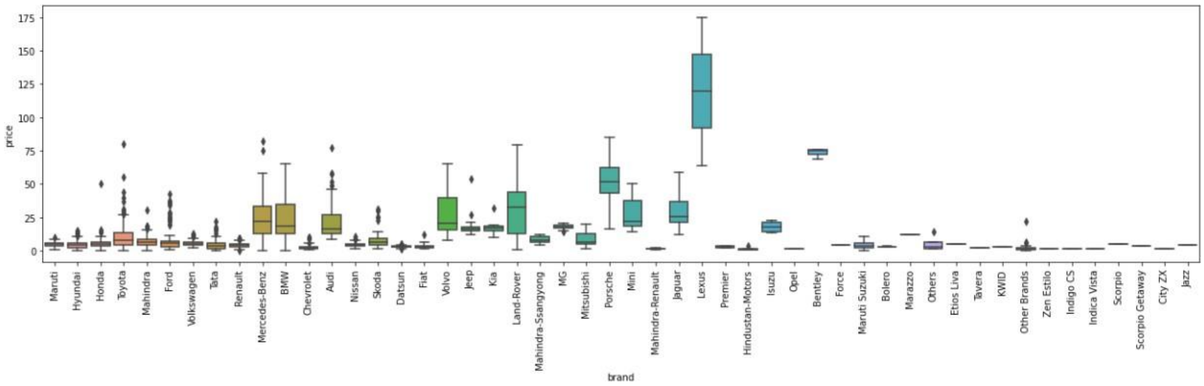
1. Count of brands



Observations:

- Most of the data are from Hyundai, honda, Toyota, Mahindra cars.

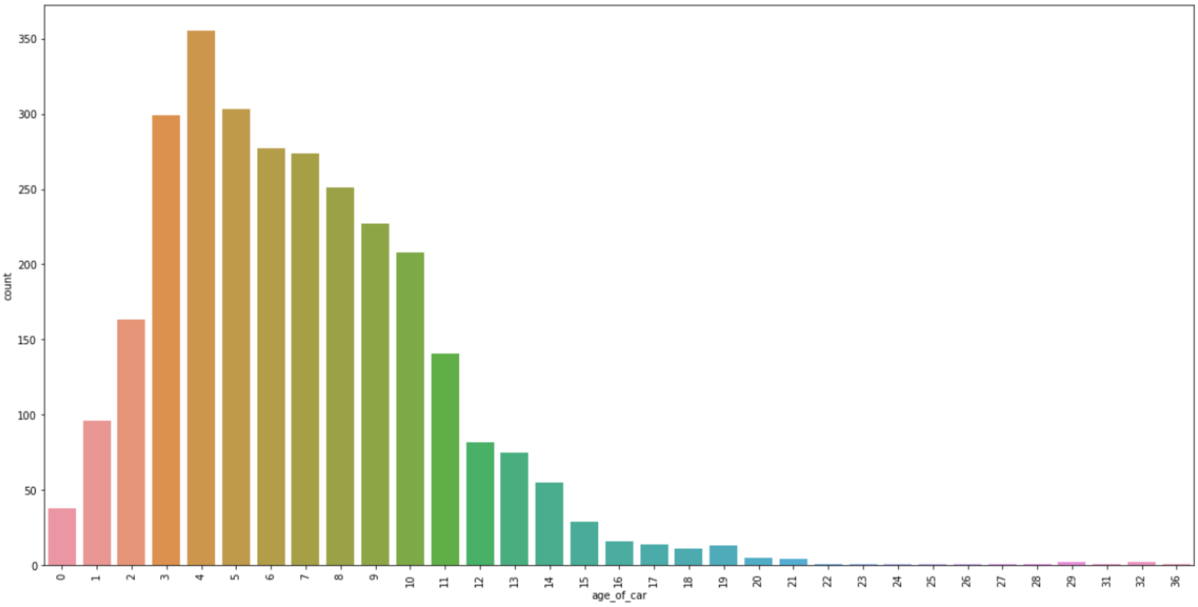
2. Car prices based on brands:



Observation:

- Lexus brand is the most expensive.

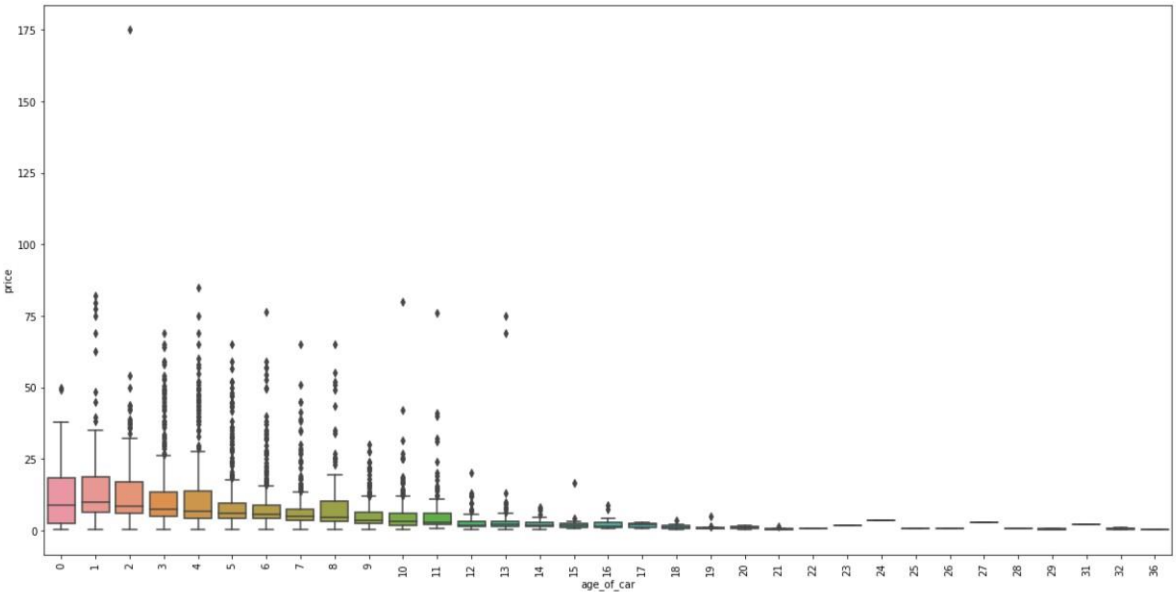
3. Car count based on age of car:



Observations:

- Most of the cars are 4 years old.
- Many cars fall in the range of 3 to 10 years of age.

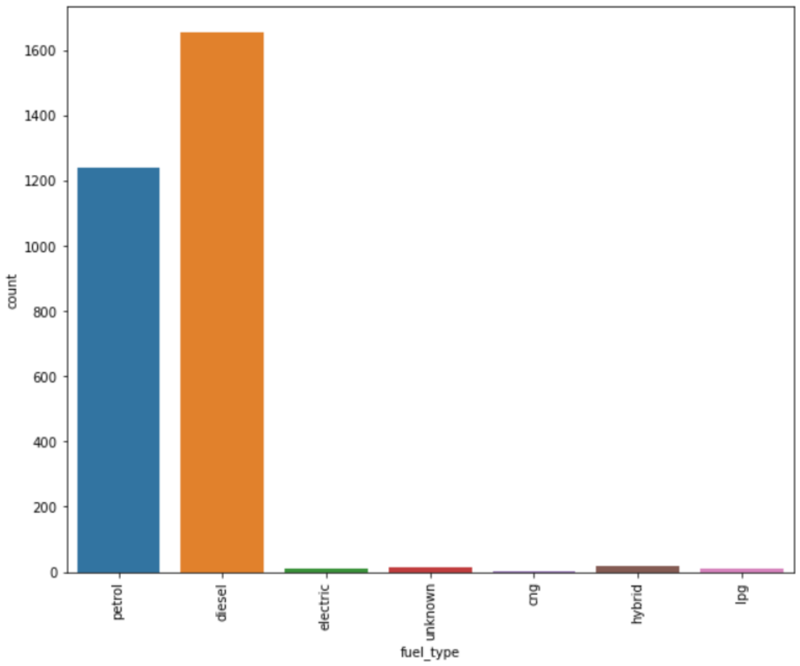
4. Price based on age of the cars:



Observations:

- There are expensive cars in the 1 to 13 years range.

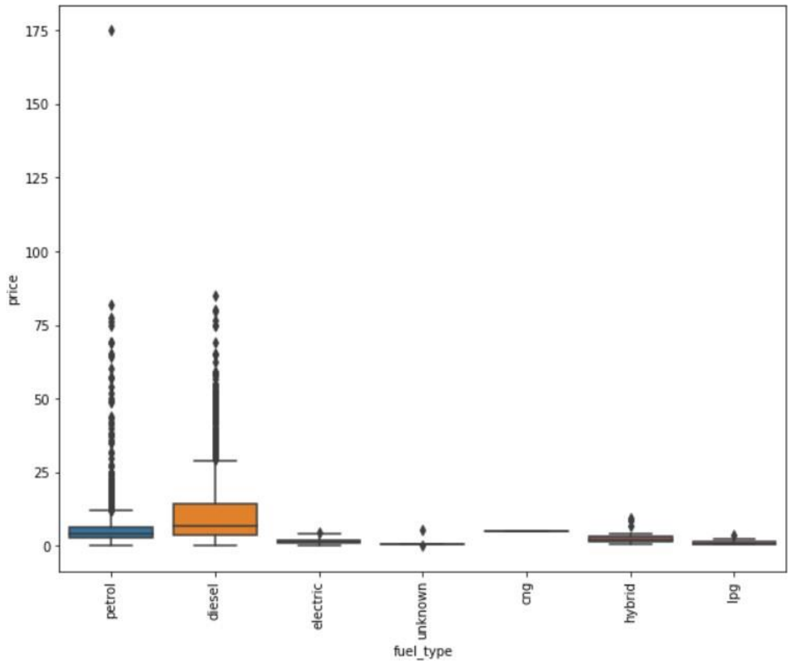
5. Number of cars based on fuel type:



Observations:

- A large number of cars are petrol or diesel cars in the dataset.

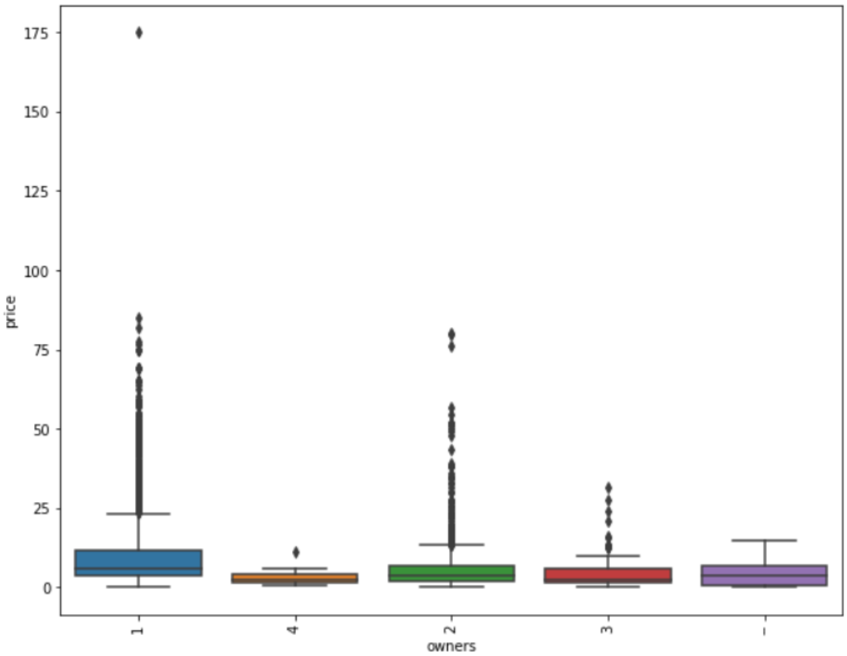
6. Price based on fuel type:



Observations:

- The most expensive cars are Diesel type in general although petrol type has one natural outlier datapoint.

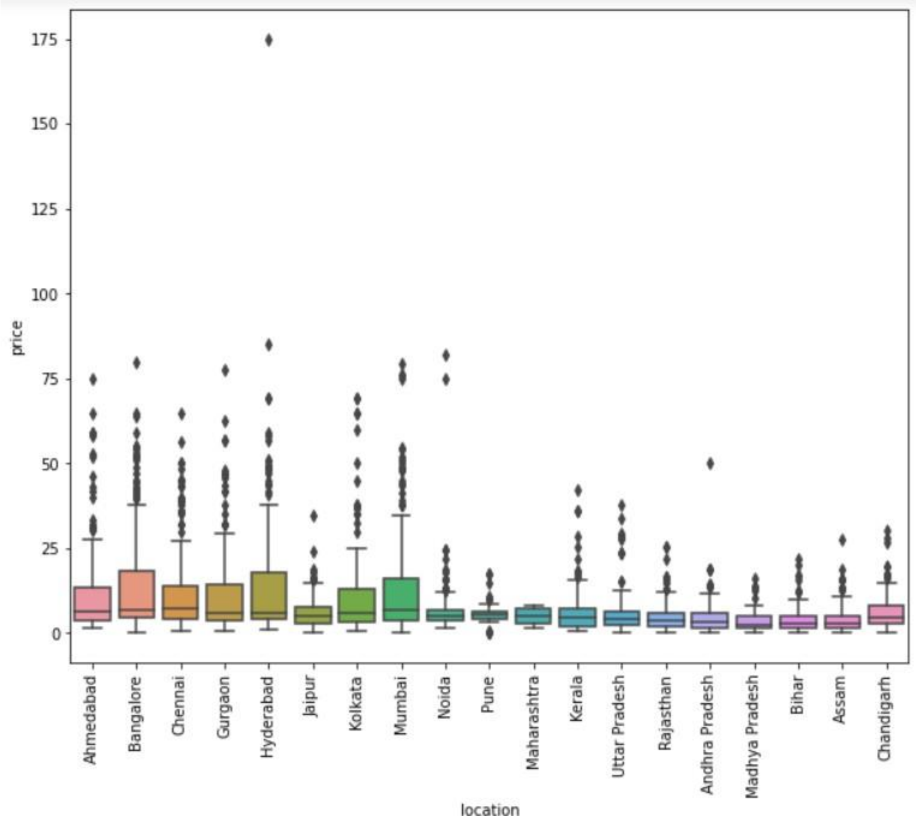
7. Price based on number of owners:



Observations:

- The price is more if the number of previous owners is less.

8. Price based on locations:



Observations:

- Bangalore and Hyderabad see to have more expensive cars.

- Interpretation of the Results
  - We can see from the visuals that the features are impacting the price.
  - There are categorical data that needs to be encoded.

CONCLUSION

- Key Findings and Conclusions of the Study
  - The Brand of the car, Age of the car and the number of previous owners of the car have more influence on the price of the car.
- Limitations of this work and Scope for Future Work

Retraining of the model is important at frequent intervals so that the predicted prices stay relevant to the economic situation.