FLIP ROBO

Car price prediction

Submitted by:
Pallavi Bajpai

ACKNOWLEDGMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and references. I would like to extend my sincere thanks to all of them.

I am highly indebted to my institution i.e. Data Trained and the mentors and chat support team for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Flip and work for their kind co-operation and encouragement which help me in completion of this project.

INTRODUCTION

- Business Problem Framing

  The price of a car depends on a lot of factors like the goodwill of the brand of the car, features of the car, horsepower and the mileage it gives and many more.

  Car price prediction is one of the major research areas in machine learning. One of the main areas of research in machine learning is the prediction of the price of cars. It is based on finance and the marketing domain. It is a major research topic in machine learning because the price of a car depends on many factors.

- Conceptual Background of the Domain Problem

  A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

- Review of Literature

  Considerable about of online research is done in order to understand the problem and requirement of ML in solving this problem.

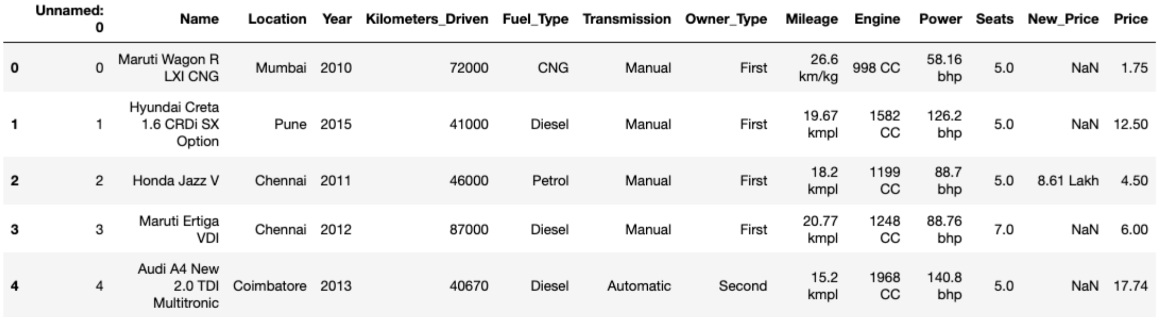
- Motivation for the Problem Undertaken

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. 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 car price valuation model.

Analytical Problem Framing

- ## Data Sources and their formats

- The dataset consists of both numerical and categorical variables. There is a total of 14 explanatory variables describing every aspect of the car price evaluation. Overview of the dataset:



	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

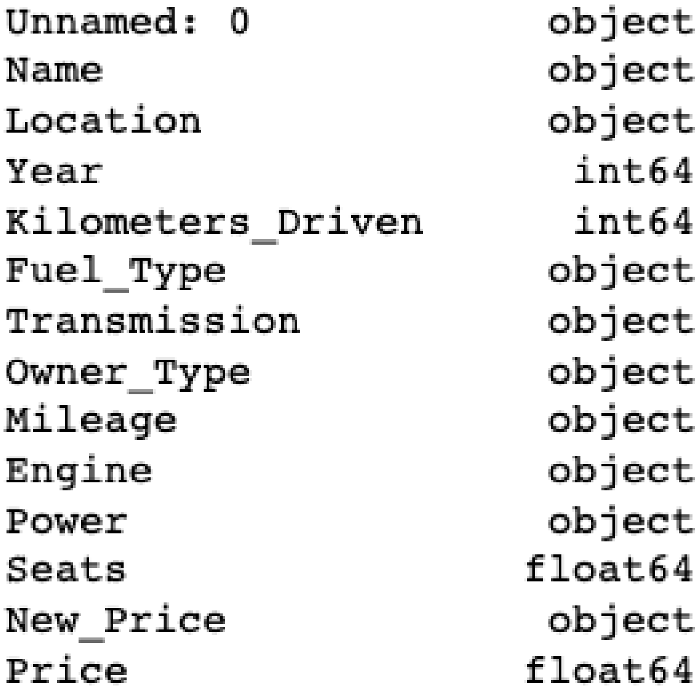
- ## Data Preprocessing Done

- Fair amount of data cleaning process was also involved. This includes checking for space and null values and then dealing with them. Also our available data has 37 columns.

- Duplicate data was removed.
- Outliers were removed but to an extent to maintain 7-8% of data loss.
- Data was imbalanced, so resampling was done.

- Data Inputs- Logic- Output Relationships

  Dataset have 14 columns out of which 10 are of object type, 2 are of int type and rest 2 are of float type.



```
Unnamed: 0               object
Name                     object
Location                 object
Year                      int64
Kilometers_Driven         int64
Fuel_Type                object
Transmission             object
Owner_Type               object
Mileage                  object
Engine                   object
Power                    object
Seats                   float64
New_Price                object
Price                   float64
```

- State the set of assumptions (if any) related to the problem under consideration

  None
- Hardware and Software Requirements and Tools Used
- Python code was written in Jupiter notebook. Below are the libraries needed in the process.
- Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model_selection.cross_val_score, sklearn.linear_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion_matrix,accuracy_score, sklearn.metrics.f1_score, scipy.stats, seaborn.

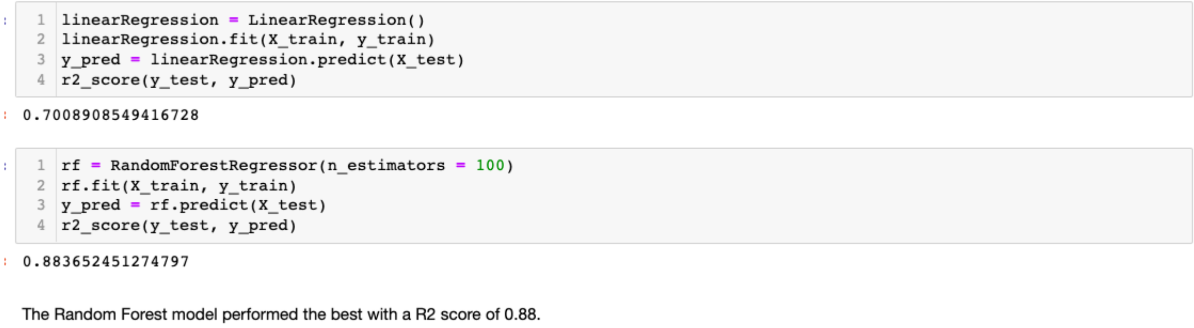
Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)
- Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
- Outliers were detected with the help of scatterplot and boxplot and were removed.
- Skewness was checked using histogram.

- ## Testing of Identified Approaches (Algorithms)

  Below algorithms were used for training and testing of data
- Random Forest Classifier
- Leniar Regression

- ## Run and Evaluate selected models



```
linearRegression = LinearRegression()
linearRegression.fit(X_train, y_train)
y_pred = linearRegression.predict(X_test)
r2_score(y_test, y_pred)
```

```
0.7008908549416728
```

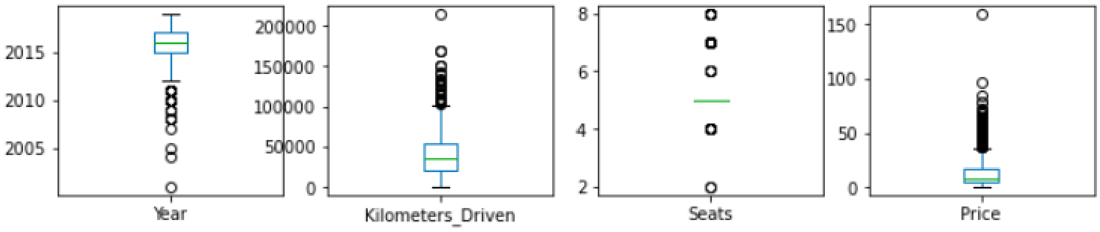
```
rf = RandomForestRegressor(n_estimators = 100)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
r2_score(y_test, y_pred)
```

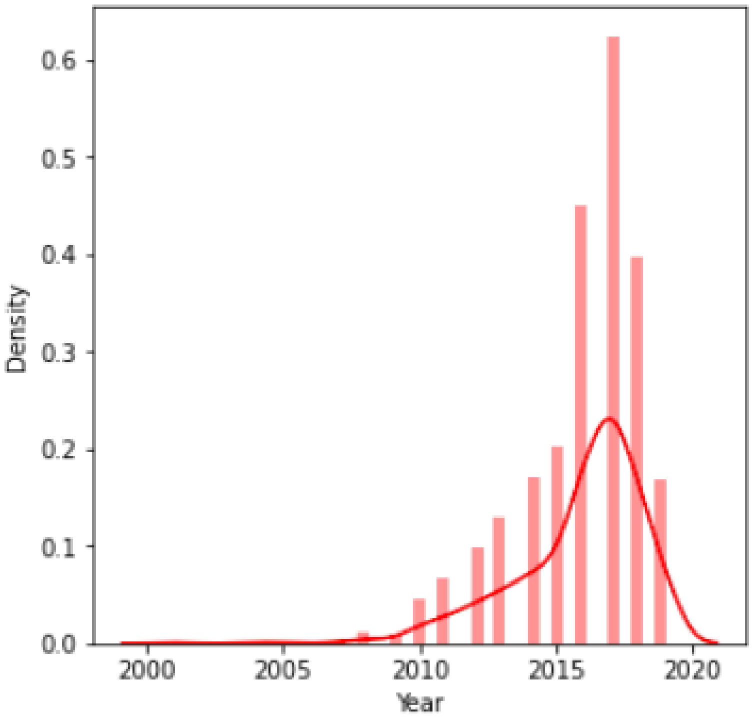
```
0.883652451274797
```

The Random Forest model performed the best with a R2 score of 0.88.

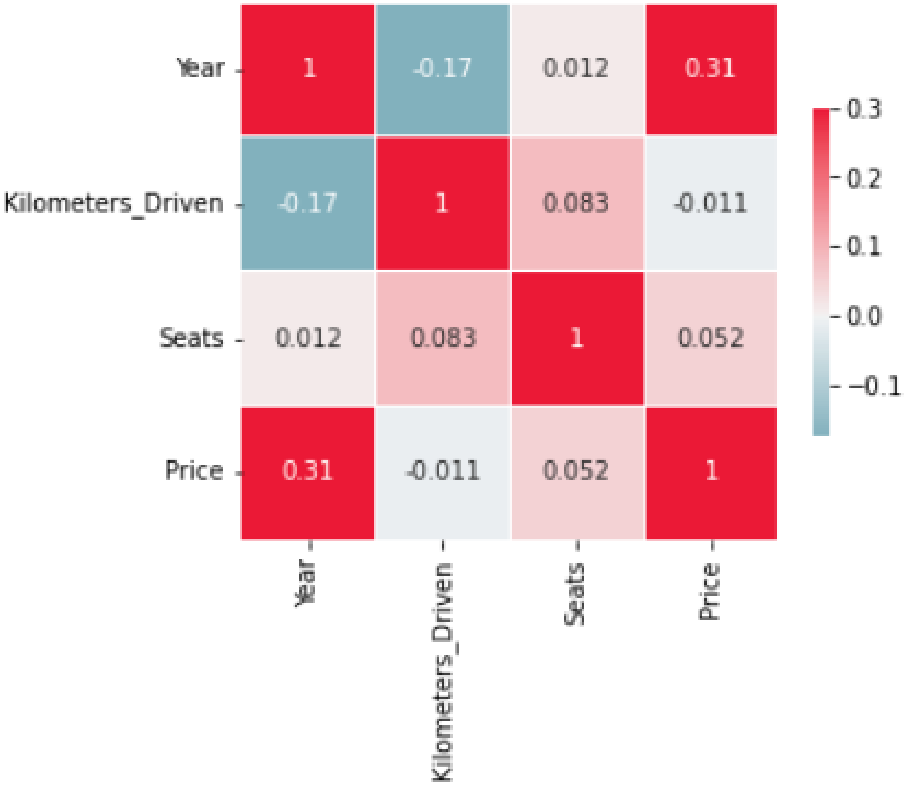
- ## Key Metrics for success in solving problem under consideration

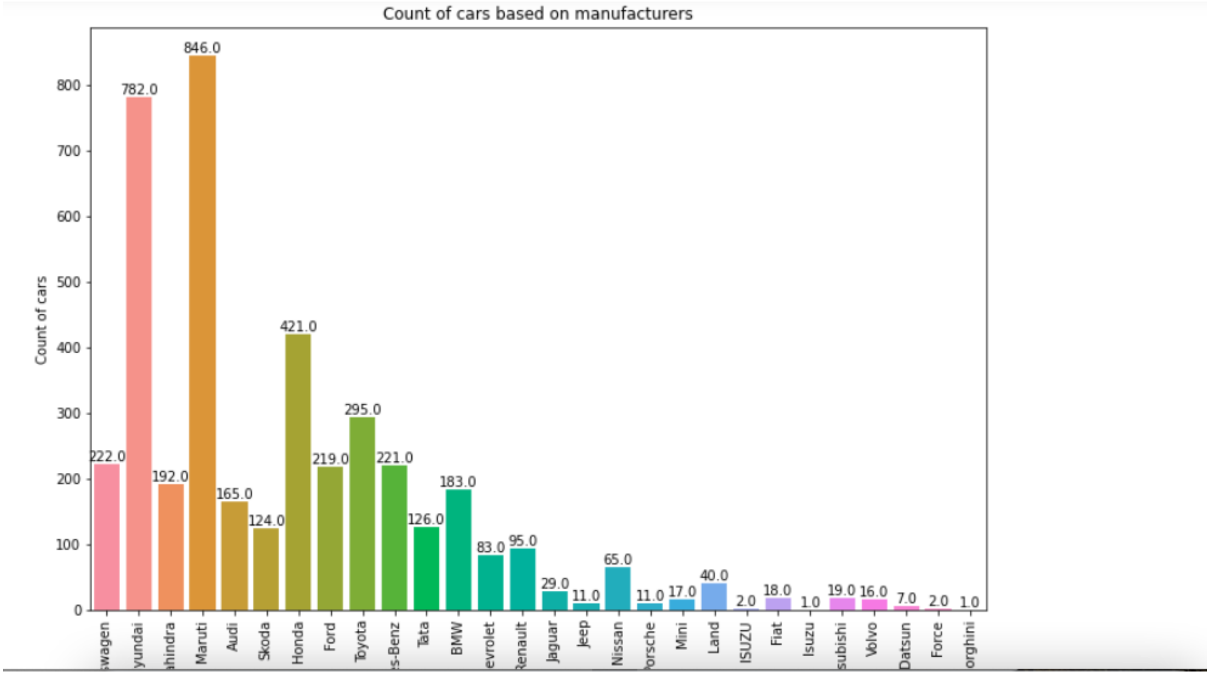
  Accuracy score was the key metric used to finalize the model.
- ## Visualizations




0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
2000
2005
2010
2015
2020
Year

```
<AxesSubplot:>
```


Year
Kilometers_Driven
Seats
Price
1
-0.17
0.012
0.31
-0.17
1
0.083
-0.011
0.012
0.083
1
0.052
0.31
-0.011
0.052
1
Year
Kilometers_Driven
Seats
Price
0.3
0.2
0.1
0.0
−0.1



- ## Interpretation of the Results
- Dataset have 6019 rows × 14 columns
- There are several cars in the dataset, some of them with a count higher than 1. Sometimes the resale value of a car also depends on manufacturer of car and hence, I'll extract the manufacturer from this column and add it to the dataset.
- There are no null values and identify all unique values.
- Maximum cars in the dataset are by the manufacturer Maruti.
- Fuel_Type, Transmission, and Owner_Type All these columns are categorical columns which should be converted to dummy variables before being used.
- The data range is really varied and the high values might affect prediction, thus, it is really important that scaling be applied to this column for sure.
- The Random Forest model performed the best with a R2 score of 0.88.

CONCLUSION

- ## Key Findings and Conclusions of the Study
- Accuracy score was highest in Random forest regressor hence we trained the model with the same.
- ## Learning Outcomes of the Study in respect of Data Science
- Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
- Outliers were detected with the help of scatterplot and boxplot and were removed.
- Skewness was checked using histogram.

- Limitations of this work and Scope for Future Work

  Outliers of all features were not removed as it was giving huge data loss.