CAR PRICE PREDICTION

Submitted by:

AVINASH PATEL

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my teacher as well as my intern company FLIPROBO who gave me this golden opportunity to do this wonderful project on the topic CAR PRICE PREDICTION, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I want to mention some sites which helped me when I got stuck somewhere while completing my projects.

Those sites are :

http://scikit-learn.org

https://www.w3schools.com

https://www.youtube.com

https://www.kaggle.com

INTRODUCTION

- Business Problem Framing

  Our client does not provide us with data. So we have to scrape data from different websites and we have to scrape more than 5000 rows of data.

  So, In order to improve the prediction of price, the client wants some predictions that could help them in further investment and improvement in selection of cars according to the price.

- Conceptual Background of the Domain Problem

  I think we have to get some knowledge on how the price of the car gets affected according to other columns and then we have to scrape those data also, to which extent the price gets increased or decreased according to the increase and decrease of other attributes and what are they.

- Review of Literature

  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. Customers through a strategy of disruptive innovation that focuses on the subscriber.

- Motivation for the Problem Undertaken

  As I am an intern I want to work on as many project as possible for me. So ,I am highly motivated to do this project and to learn new thing and also learn from my mistakes.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  First of all I checked that how many of the column are categorical columns and how many of them are numerical columns. By doing so I find out that there are Four columns with categorical data i.e. Car Name, Transmission, Fuel Type, Body Type.

  Now I started to visualise the column individually.

  On creating a Catplot for the columns that I have treated I find out that many of the columns are right skewed i.e. the bell curve is elongated parallel to the x-axis on the right side.

  After that I created a boxplot again to see how much outliers are still left and some of the columns still have many outliers.

  Then I proceed to create a heatmap for each column and find out that the maximum correlation with our target variable is -53% which is of Transmission column.

- Data Sources and their formats

  As no Data is provided by our client I scraped 6506 rows of data with 7 columns of each. I scraped data from cardekho.com , carwale.com,olacars.com.

- Data Pre-processing Done

  For Kilometres Run: As it is the number of kilometres rum while with the previous owner it has km written with it which makes it a

categorical column so I treated this column by deleting all the subscripts with km and keep only the integer values..

For Fuel Type: As for this column I deleted the +1 from some fuel types and then after doing that I left with sone fuel types which are the same but the capitalisation is different so I merged the rows with the same name as one name.

Treating Price column: Treating this column is a little tricky because not only it contains commas but also it has three different values one is the full amount the other is subscript Lakh and the other subscript crore. So I have to first delete all the comas and the segregate all the values with lakh and multiplied it with 100000 and for the rows which have subscript crore I multiplied them with 10000000.

Treating Transmission column: This column has three different types of automatic written which the program is treating three different and the same case is with manual also. So I merged all the automatic rows to single name AUTOMATIC and I have done the same for the MANUAL rows ,after replacing all the automatics and the manuals we have left with only two row type automatic and manual.

- Data Inputs- Logic- Output Relationships

  The Remaining Data after all the pre-processing acts as an input data for the Model(in our case XGBoost Regressor) the already trained model gives us an output of an estimated price.

- Hardware and Software Requirements and Tools Used

  Ram:8GB
  ROM:200MB
  Processor : Ryzen 5 3600X(6 cores 12 threads)
  Tool : Jupiter Notebook
  Language Used: Python

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  First I imported all the basic libraries of python used in the problem solving. Then I read carefully the data and the description given to me. After that I imported the provided csv file into the Jupiter notebook. Now first I checked for any nan values in the dataset after that I checked how many number of rows and columns I am dealing with and I also done some more analysis of the raw data to get a general idea of there max and mins and there inter quartile ranges etc.

  After analysing the data I started cleaning the data as explained above.

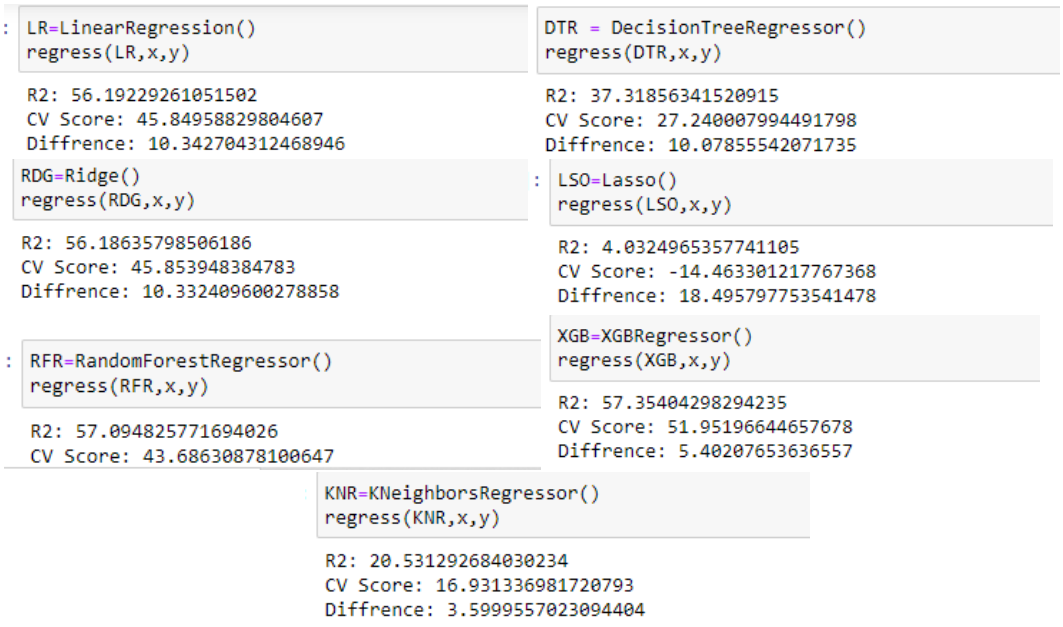
  After cleaning I did some visualisation of the data to get some insights from our cleaned data.

  I removed skewness and some excess outliers. On the completion of all these steps I moved on to the model selection part. Here I tested many models and choose one of the model to proceed.

  After I selected one of the models I did hyperparameter tuning and finally I saved my model.

- Testing of Identified Approaches (Algorithms)
  - LinearRegression
  - DecisionTreeRegressor
  - Ridge
  - Lasso
  - RandomForestRegressor
  - KNeighborsRegressor
  - XGBRegressor
- Run and Evaluate selected models

After running all the models the lowest difference between the accuracy score and the cross validation is for XGBOOST Regressor.



```
LR=LinearRegression()
regress(LR,x,y)
```

```
R2: 56.19229261051502
CV Score: 45.84958829804607
Diffrence: 10.342704312468946
```

```
DTR = DecisionTreeRegressor()
regress(DTR,x,y)
```

```
R2: 37.31856341520915
CV Score: 27.240007994491798
Diffrence: 10.07855542071735
```

```
RDG=Ridge()
regress(RDG,x,y)
```

```
R2: 56.18635798506186
CV Score: 45.853948384783
Diffrence: 10.332409600278858
```

```
LSO=Lasso()
regress(LSO,x,y)
```

```
R2: 4.0324965357741105
CV Score: -14.463301217767368
Diffrence: 18.495797753541478
```

```
RFR=RandomForestRegressor()
regress(RFR,x,y)
```

```
R2: 57.094825771694026
CV Score: 43.68630878100647
```

```
XGB=XGBRegressor()
regress(XGB,x,y)
```

```
R2: 57.35404298294235
CV Score: 51.95196644657678
Diffrence: 5.40207653636557
```

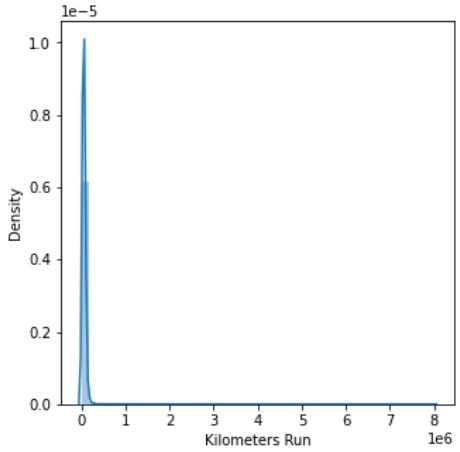
```
KNR=KNeighborsRegressor()
regress(KNR,x,y)
```

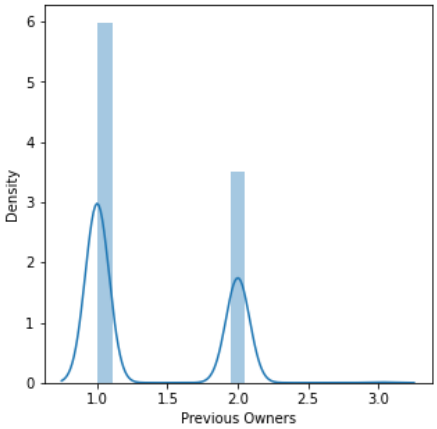
```
R2: 20.531292684030234
CV Score: 16.931336981720793
Diffrence: 3.5999557023094404
```

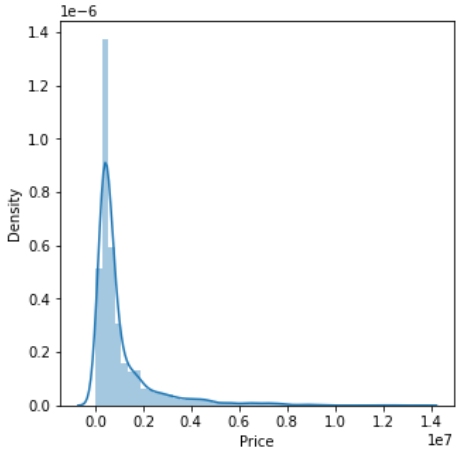
That means XGBoost Regressor is not overfitted nor underfitted.

- Visualizations

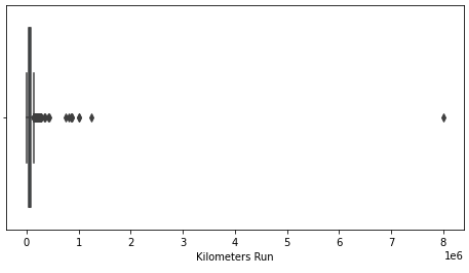
  Distplot:

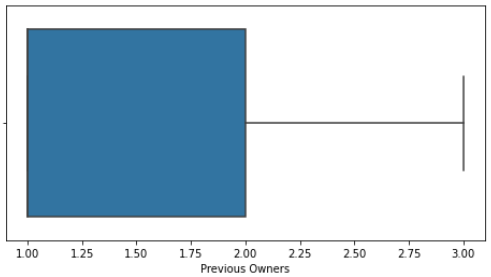

1e−5
Density
Kilometers Run
1e6

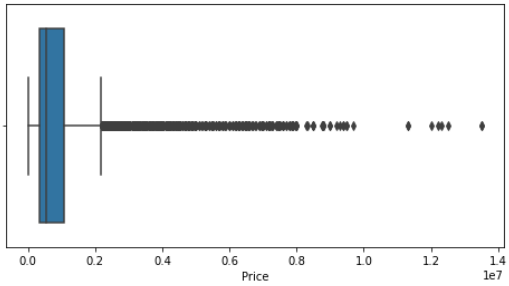

Density
Previous Owners


1e−6
Density
Price
1e7

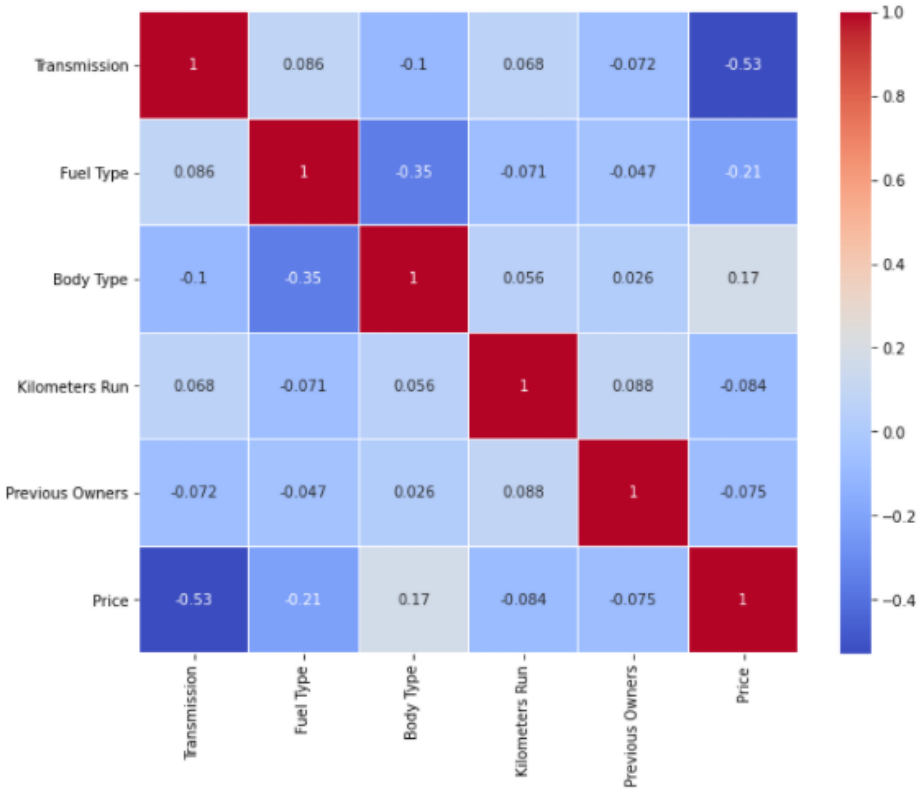
Boxplot:


Kilometers Run
1e6

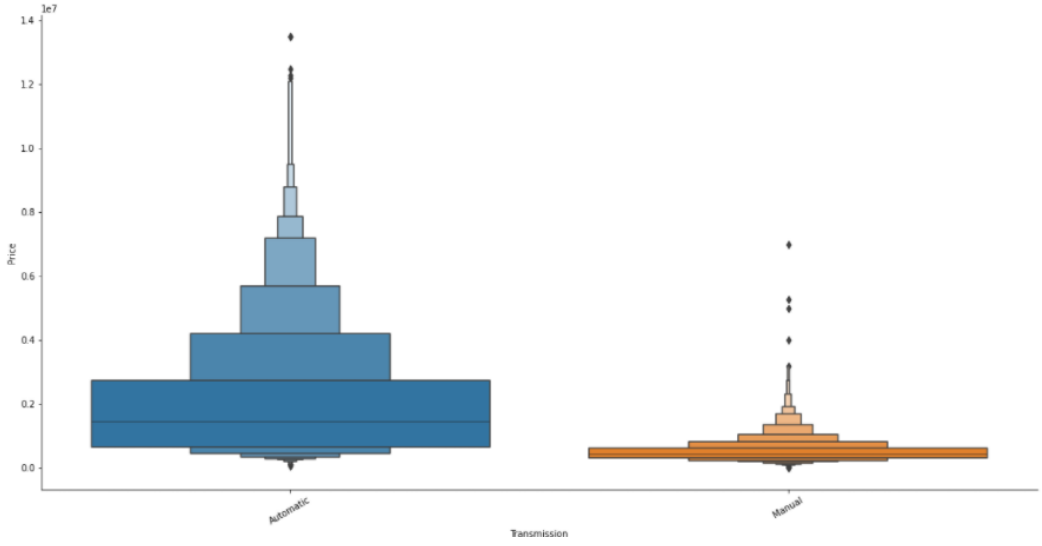

Previous Owners

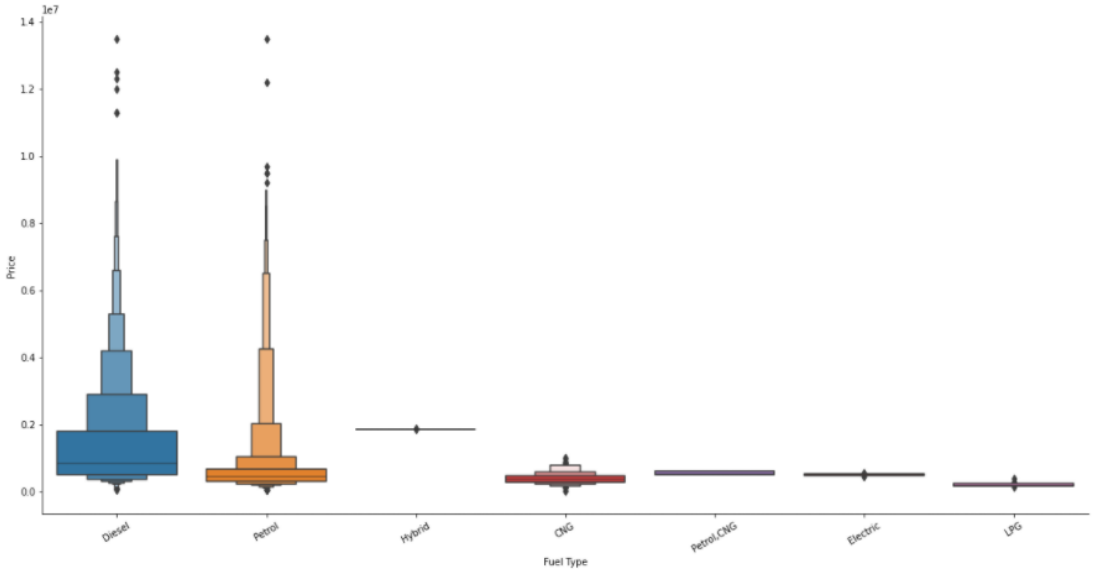

Price
1e7

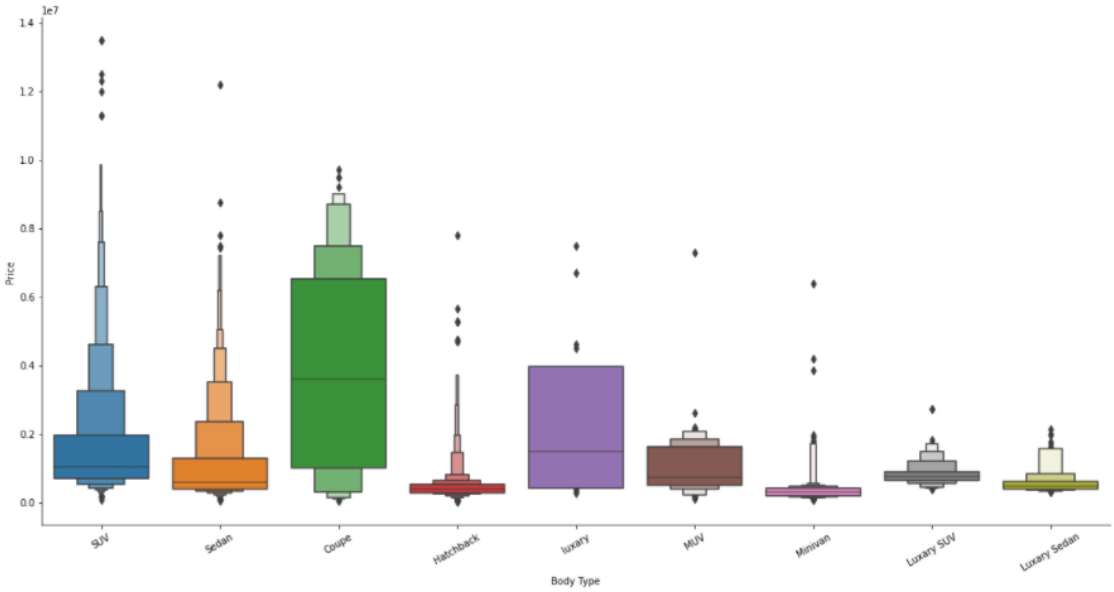
Heatmap:



Catplot:







- Interpretation of the Results

  After all the pre processing and then visualising the data I find out that although there are many columns to the dataset ,but name column is useless and the remaining columns do not have a significant correlation with our Target column except one that is transmission.

CONCLUSION

- ## Key Findings and Conclusions of the Study

  I find out that few columns are correlated to our target variable that means predicting our target variable is difficult for any model because there is no direct correlation between them. Also, many of the columns have values which are unrealistic. Many of the columns have highly skewed values and many columns also have high amounts of outlies in them.

- ## Learning Outcomes of the Study in respect of Data Science

  As I have already told that this dataset has a lot of outliers and I have to look into each and every column to get some of the data cleaned. As per the visualisation part I plotted boxplots to see the number of outliers and in which column they exist and in how much amount.

  I have also plotted Distplot to get a general idea of the columns which have skewness to further treat them.

  I plotted heatmap to see which columns are in a correlation with our target variable.

  I have plotted catplot to see that our Categorical columns are how much varying with price.

- ## Limitations of this work and Scope for Future Work

  As I am not able to fully understand the meaning of all the columns and if I could get someone who is already working in the car company then I could get a better idea of how all the column are working and which outer columns needs to be in the dataset. As per the future scope, yes there is definitely more room to grow as I could get the model efficiency to 65% it could be more on further research.