FLIP ROBO

PROJECT REPORT ON:

“Car Price Prediction”

SUBMITTED BY

Adi Rajit Mahesh

ACKNOWLEDGMENT

I would like to extend my deepest gratitude to “Flip Robo” team who have given me this opportunity to work on this amazing project. I would also like to thank “Srishti Mann” ma’am to extend helped me out during my project.

A huge thanks to “Datatrained” who are the reason for who I am today. Lastly, I would like to thank my parents who have always been my backbone in every step of life.

Contents

1. Introduction
   1.1 Business Problem Framing:
   1.2 Conceptual Background of the Domain Problem
   1.3 Review of Literature
   1.4 Motivation for the Problem Undertaken

2. Analytical Problem Framing
   2.1 Mathematical/ Analytical Modelling of the Problem
   2.2 Data Sources and their formats
   2.3 Data Pre-processing Done
   2.4 Data Inputs-Logic-Output Relationships
   2.5 Hardware and Software Requirements and Tools Used

3. Data Analysis and Visualization
   3.1 Identification of possible problem-solving approaches (methods)
   3.2 Testing of Identified Approaches (Algorithms)
   3.3 Key Metrics for success in solving problem under consideration
   3.4 Visualization
   3.5 Run and evaluate selected models
   3.6 Interpretation of the Results

4. Conclusion
   4.1 Key Findings and Conclusions of the Study
   4.2 Learning Outcomes of the Study in respect of Data Science
   4.3 Limitations of this work and Scope for Future Work

1. INTRODUCTION

1.1 Business Problem Framing:

Car price prediction is somehow interesting and popular problem. As per information that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 from which 84% of them are cars for personal usage. This number is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will increase in future. This adds additional significance to the problem of the car price prediction. Accurate car price prediction involves expert knowledge, because price usually depends on many distinctive features and factors. Typically, most significant ones are brand and model, age, horsepower and mileage. The fuel type used in the car as well as fuel consumption per mile highly affect price of a car due to a frequent change in the price of a fuel. Different features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the car price. In this report, we applied different methods and techniques in order to achieve higher precision of the used car price prediction.

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.

1.2 Conceptual Background of the Domain Problem

The prices of new cars in the industry are fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling.

There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less than its market value.

There are one of the biggest target groups that can be interested in results of this study. If used car sellers better understand what makes a car desirable, what are the important features for a used car, then they may consider this knowledge and offer a better service.

1.3 Review of Literature

The second-hand car market has continued to expand even as the reduction in the market of new cars. According to the recent report on India's pre-owned car market by Indian Blue Book, nearly 4 million used cars were purchased and sold in 2018-19. The second-hand car market has created the business for both buyers and sellers. Most of the people prefer to buy the used cars because of the affordable price and they can resell that again after some years of usage which may get some profit. The price of used cars depends on many factors like fuel type, color, model, mileage, transmission, engine, number of seats etc., the used cars price in the market will keep on changing. Thus, the evaluation model to predict the price of the used cars is required.

1.4 Motivation for the Problem Undertaken

There are websites that offers an estimate value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used car's market value.

2. Analytical Problem Framing

2.1 Mathematical/ Analytical Modeling of the Problem

As a first step I have scrapped the required data from www.cardekho.com, and saved the data in excel format.

In this particular problem car_price is my target column and data was continuous. So clearly, it is a regression problem and I have to use all regression algorithms to build the models. There were null values in the dataset. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 50% null values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. Since we have scrapped the data from cardekho.com the raw data was not in the format, so we have use feature engineering to extract the required feature format. To get better insight on the features I have used plotting like distribution plot, bar plot, reg plot, strip plot and count plot. With these plotting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using z-score method and I removed skewness using yeo-Johnson method. I have used all the regression algorithms while building model then tuned the best model and saved the best model. At last, I have predicted the car-price using saved model.

2.2 Data Sources and their formats

The data was collected from cardekho.com website in excel format. The data was scrapped using selenium. After scrapping required features the dataset is saved as excel file.

Also, my dataset was having 12608 rows and 20 columns including target. In thi s particular datasets I have object type of data which has been changed as per our analysis about the dataset. The information about features is as follows.

Features Information:

- Car_Name: Name of the car with Year
- Fuel_type: Type of fuel used for car engine
- Running_in_kms: Car running in kms till the date

- Endine_disp: Engine displacement/engine CC
- Gear_transmission: Type of gear transmission used in car
- Milage_in_km/ltr: Overall milage of car in Km/ltr
- Seating_cap: Availability of number of seats in the car
- color: Car color
- Max_power: Maximum power of engine used in car in bhp
- front_brake_type: type of brake system used for front-side wheels
- rear_brake_type: type of brake system used for back-side wheels
- cargo_volume: the total cubic feet of space in a car's cargo area.
- height: Total height of car in mm
- width: Width of car in mm
- length: Total length of the car in mm
- Weight : Gross weight of the car in kg
- Insp_score : inspection rating out of 10
- top_speed : Maximum speed limit of the car in km per hours
- City_url : Url of the page of cars from a particular city
- Car_price : Price of the car

2.3 Data Pre-processing Done

- ✓ As a first step I have scrapped the required data using selenium from cardekho website.
- ✓ And I have imported required libraries and I have imported the dataset which was in excel format.
- ✓ Then I did all the statistical analysis like checking shape, nunique, value counts, info etc….
- ✓ While checking for null values I found null values in the dataset and I replaced them using imputation technique.
- ✓ I have also dropped Unnamed: 0, cargo_volume and Insp_score column as I found they are useless.
- ✓ Next as a part of feature extraction I converted the data types of all the columns and I have extracted useful information from the raw dataset. Thinking that this data will help us more than raw data.

2.4 Data Inputs- Logic- Output Relationships

- ✓ Since I had numerical columns, I have plotted dist plot to see the distribution of skewness in each column data.
- ✓ I have used bar plot for each pair of categorical features that shows the relation between label and independent features.
- ✓ I have used reg plot and strip plot to see the relation between numerical columns with target column.
- ✓ I can notice there is a linear relationship between maximum columns and target.

2.5 Hardware and Software Requirements and Tools Used

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

Hardware required: -
1. Processor — core i5 and above
2. RAM — 4 GB and above
3. SSD — 250GB and above

Software/s required: -

1.Anaconda

Libraries required:-

- **import pandas as pd**: **pandas** are a popular Python-based data analysis toolkit which can be imported using `import pandas as pd`. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.
- **import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

- **import seaborn as sns:** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas’ data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.
- **Import matplotlib.pyplot as plt:** matplotlib. pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
    - ✓ from sklearn.preprocessing import LabelEncoder
    - ✓ from sklearn.preprocessing import StandardScaler
    - ✓ from sklearn.linear_model import LinearRegression
    - ✓ from sklearn.ensemble import RandomForestRegressor
    - ✓ from sklearn.tree import DecisionTreeRegressor
    - ✓ from xgboost import XGBRegressor
    - ✓ from sklearn.ensemble import GradientBoostingRegressor
    - ✓ from sklearn.metrics import classification_report
    - ✓ from sklearn.metrics import accuracy_score
    - ✓ from sklearn.model_selection import cross_val_score

3.Data Analysis and Visualization

3.1 Identification of possible problem-solving approaches (methods)

- ✓ Since the data collected was not in the format, we have to clean it and bring it to the proper format for our analysis. To remove outliers, I have used z-score method. And to remove skewness I have used yeo-Johnson method. We have dropped all the unnecessary columns in the dataset according to our understanding. Use of Pearson's correlation coefficient to check the correlation between dependent and independent features. Also, I have used Standardization to scale the data. After scaling we have to remove multicolinearity using VIF. Then followed by model building with all Regression algorithms

3.2 Testing of Identified Approaches (Algorithms)

Since car_price was my target and it was a continuous column with improper format which has to be changed to continuous float datatype column, so this particular problem was Regression problem. And I have used all Regression algorithms to build my model. By looking into the difference of r2 score and cross validation score I found Random Forest Regression as the best model with least difference. Also, to get the best model we have to run through multiple models and to avoid the confusion of overfitting we have go through cross validation. Below is the list of Regression algorithms I have used in my project.

- ➢ Linear Regression
- ➢ Random Forest Regression
- ➢ Decision Tree Regression
- ➢ XGB Regression
- ➢ Gradient Boosting Regression

3.3 Key Metrics for success in solving problem under consideration

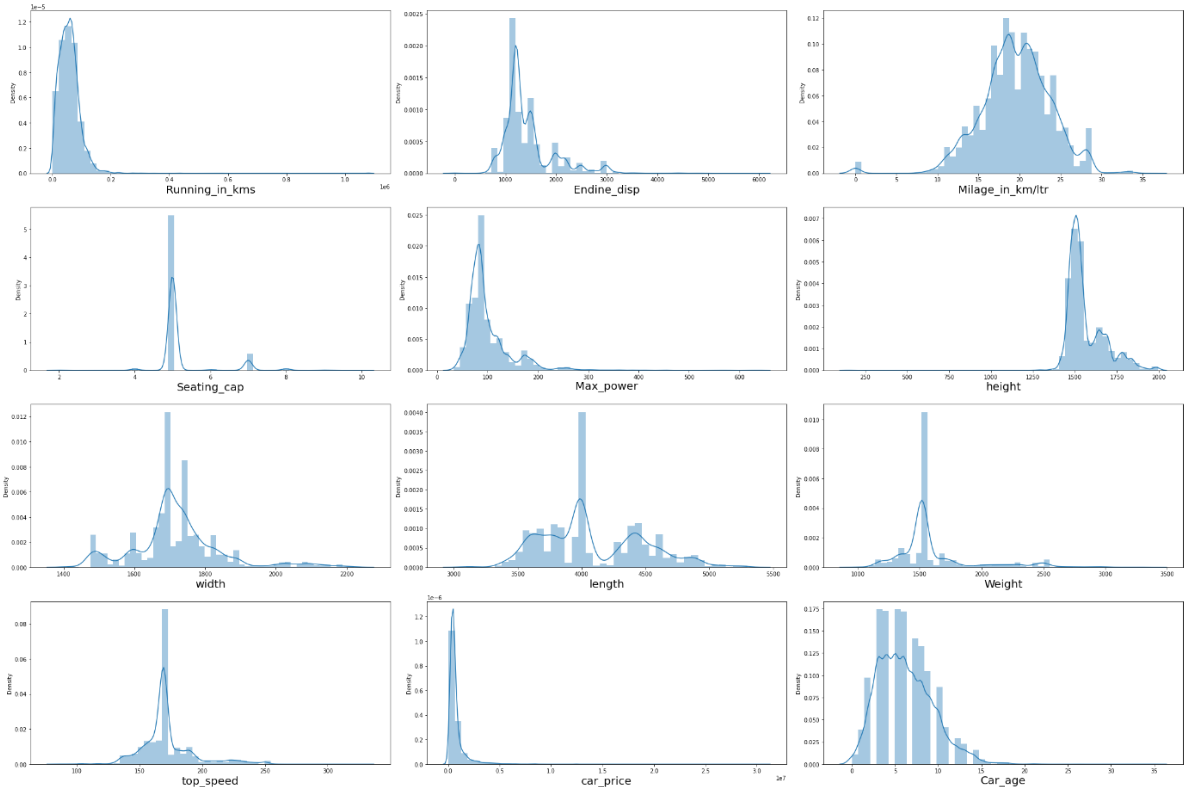
I have used the following metrics for evaluation:

- I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.
- I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.
- I have used r2 score which tells us how accurate our model is.

3.4 Visualizations

I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and reg plot, strip plot for bivariate analysis.

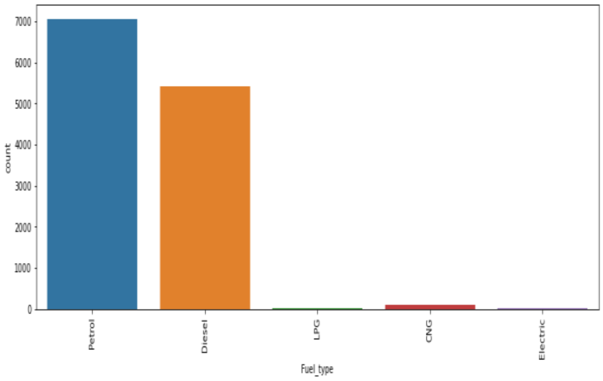
1. **Univariate Analysis for numerical columns:**

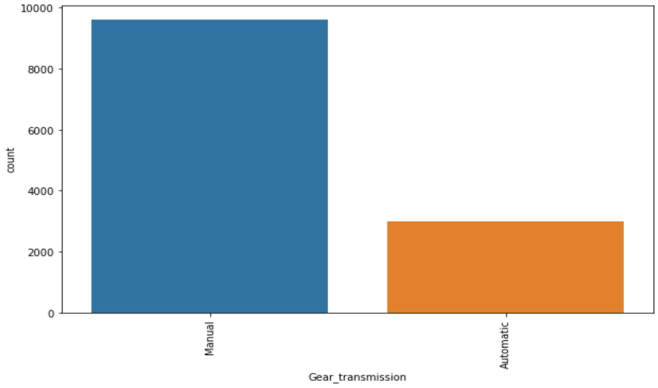


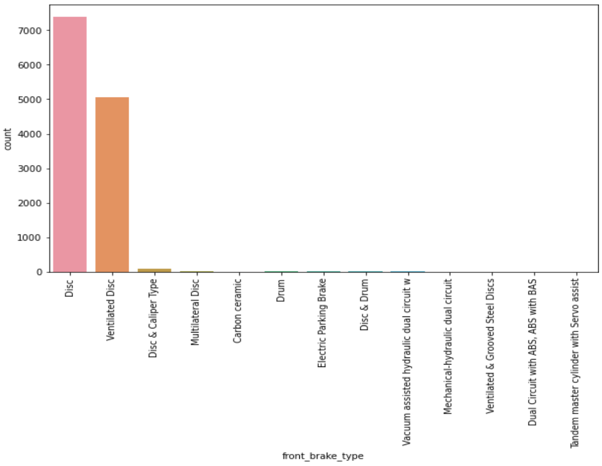
Observations:

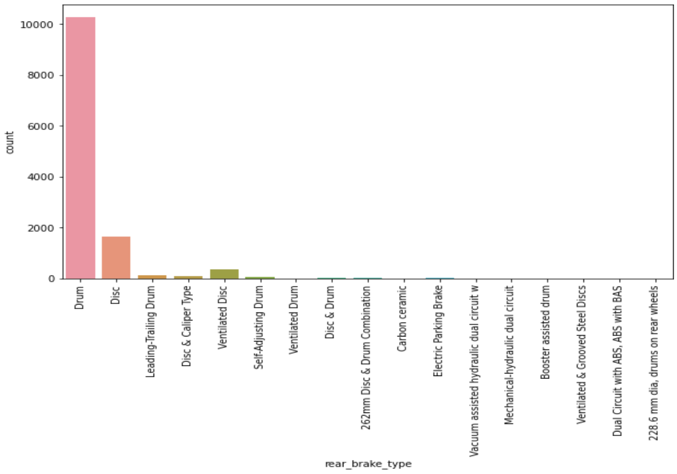
- ✓ We can clearly see that there is skewness in most of the columns so we have to treat those using suitable methods.

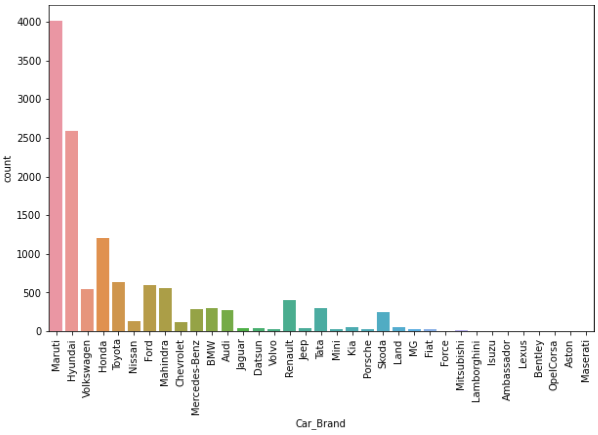
2. Univariate analysis for categorical column:

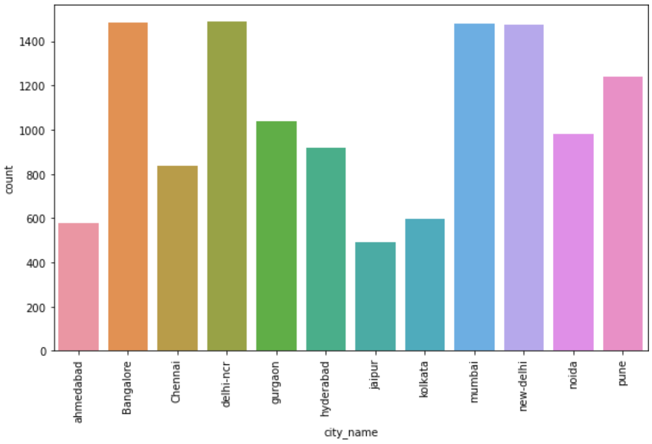








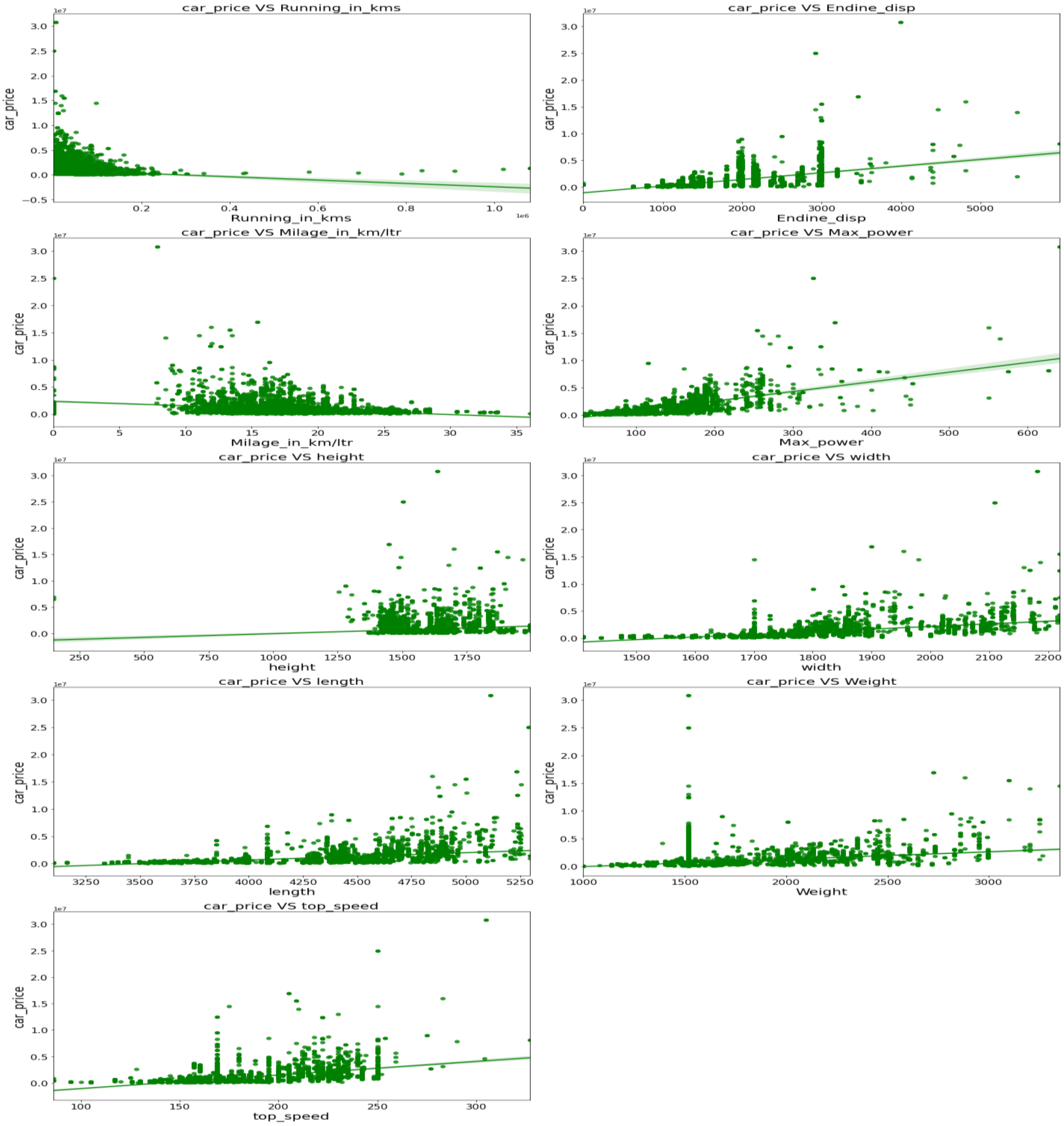


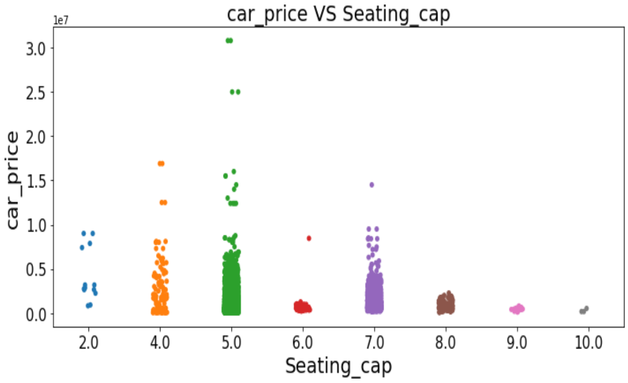


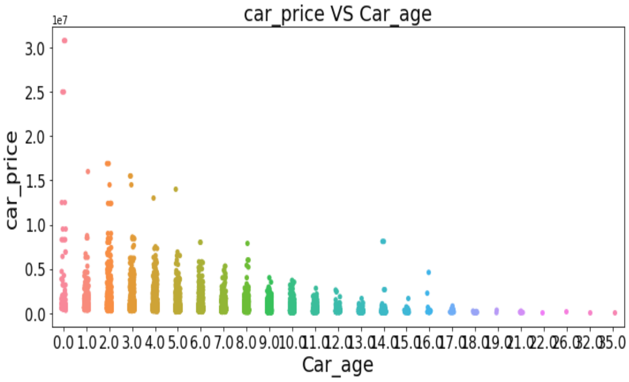
Observations:

- ✓ Maximum cars are petrol driven and also diesel driven.
- ✓ Maximum cars are with Manual gear transmission.
- ✓ Disc front brake cars are more in number followed by Ventilated Disc.
- ✓ Drum rare break cars are more in number.
- ✓ Maximum cars under sale are Maruti followed by Hyundai.
- ✓ In Bangalore, delhi-ncr, Mumbai and new-Delhi we can find maximum cars for sale. Since these are most populated places.

3. Bivariate analysis for numerical columns:



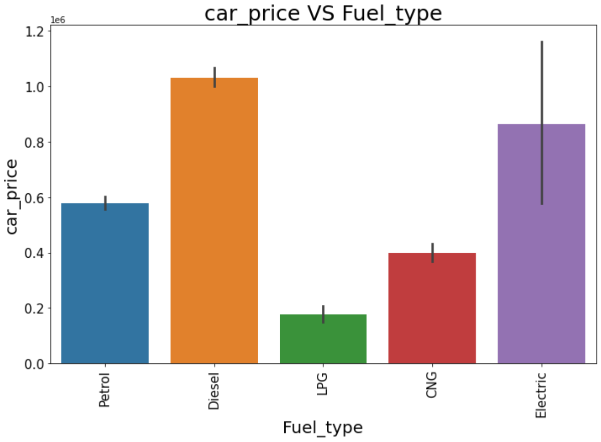


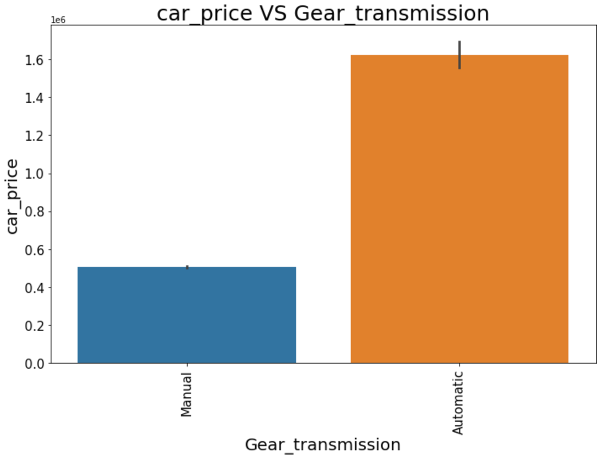


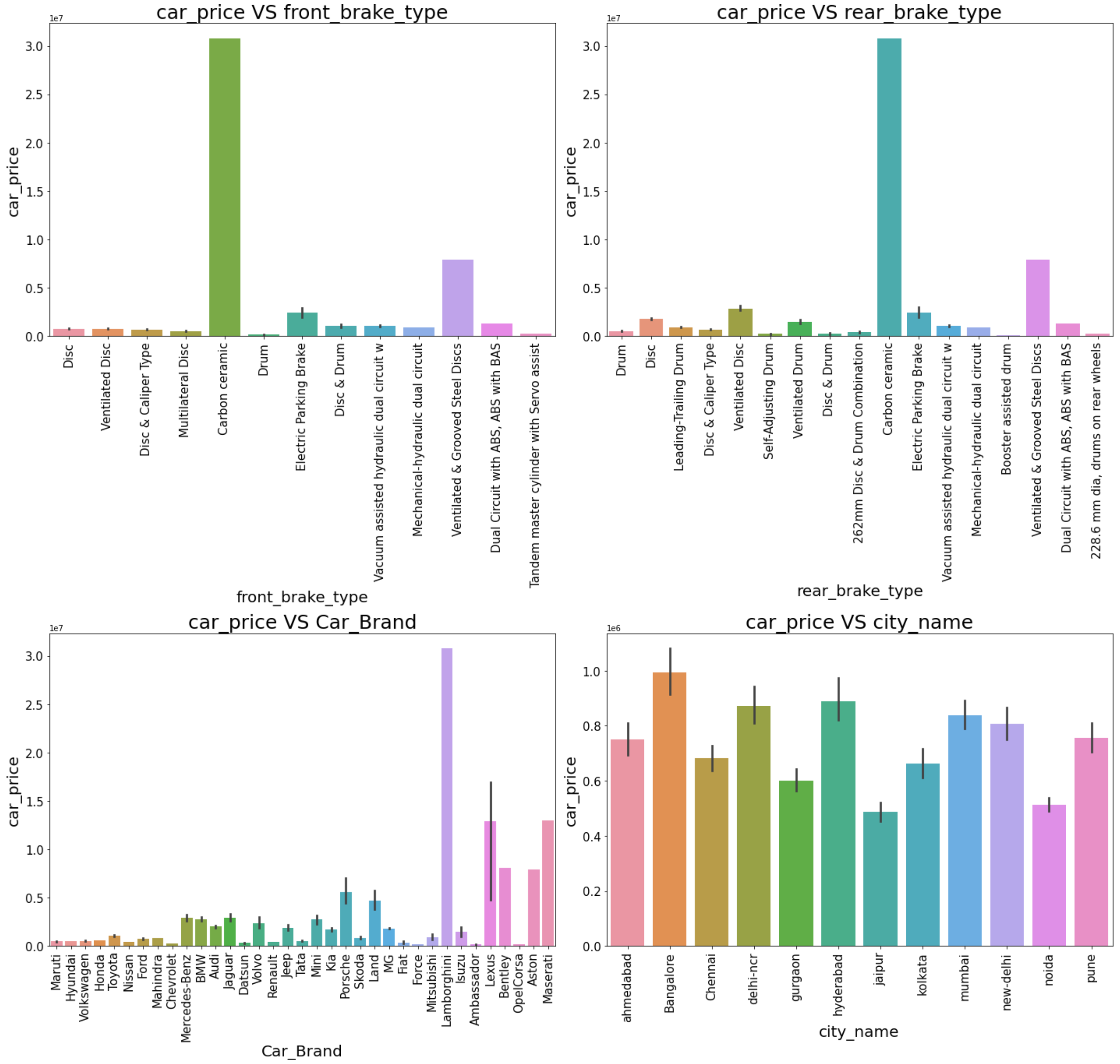
Observations:

- ✓ Maximum cars are having below 20k driven kms. And car price is high for less driven cars.
- ✓ Maximum cars are having 1000-3000 Endine_disp. And car price is high for 3000 Endine_disp.
- ✓ Maximum cars are having milage of 10-25kms. And, milage has no proper relation with car price.
- ✓ As Max_power is increasing car price is also increasing.
- ✓ Car_price has no proper relation with height.
- ✓ As the width is increasing car price is also increasing.
- ✓ As length is increasing car price is also increasing.
- ✓ Weight also has linear relationship with car price.
- ✓ As top_speed is increasing car price is also increasing.
- ✓ Cars with 5 and 4 seats are having highest price.
- ✓ As the age of the car increases the car price decreases.

4. Bivariate Analysis for categorical columns:






car_price VS front_brake_type

car_price VS rear_brake_type

car_price VS Car_Brand

car_price VS city_name

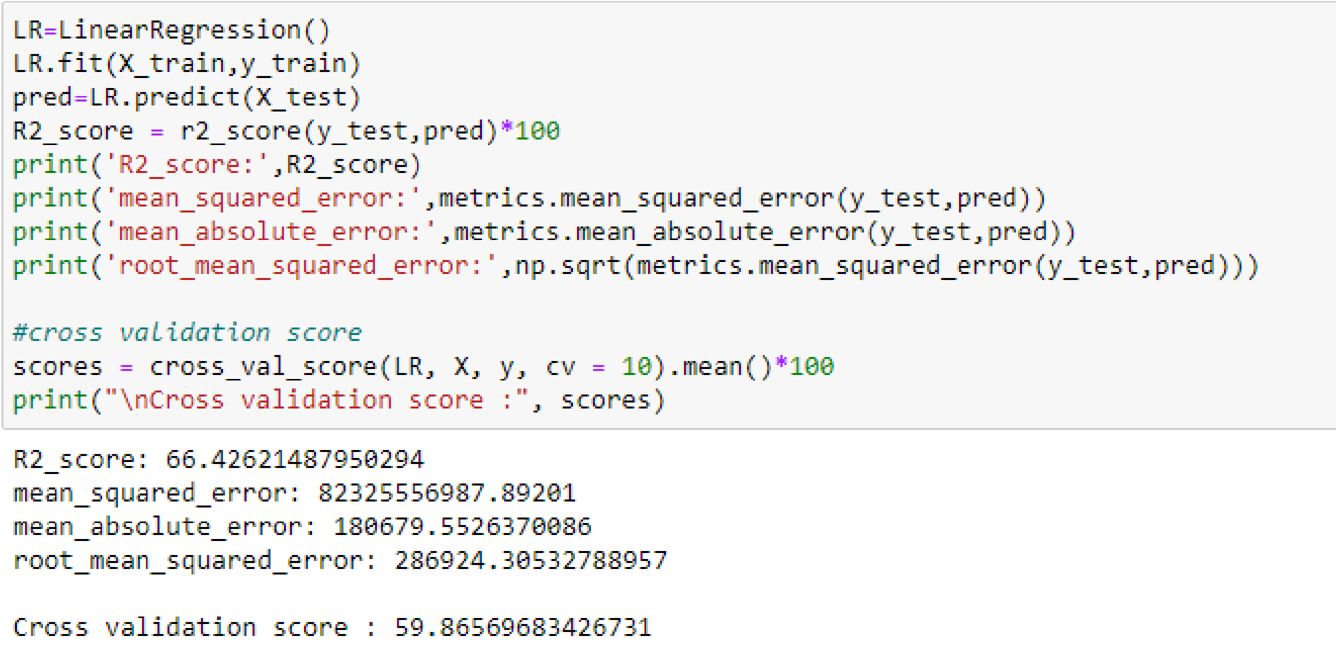
Observations:

- ✓ For Diesel and Electric cars, the price is high compared to Petrol, LPG and CNG.
- ✓ Cars with automatic gear are costlier than manual gear cars.
- ✓ Cars with Carbon Ceramic front brake are costlier compared to other cars.
- ✓ Cars with carbon Ceramic rear brake are costlier compared to other cars.
- ✓ Lamborghini brand cars are having highest sale price.
- ✓ In Bangalore, Hyderabad and delhi-ncr the car prices are high as they are highly populated cities.

.5 Run and Evaluate selected models

1. Model Building

1) Linear Regression:



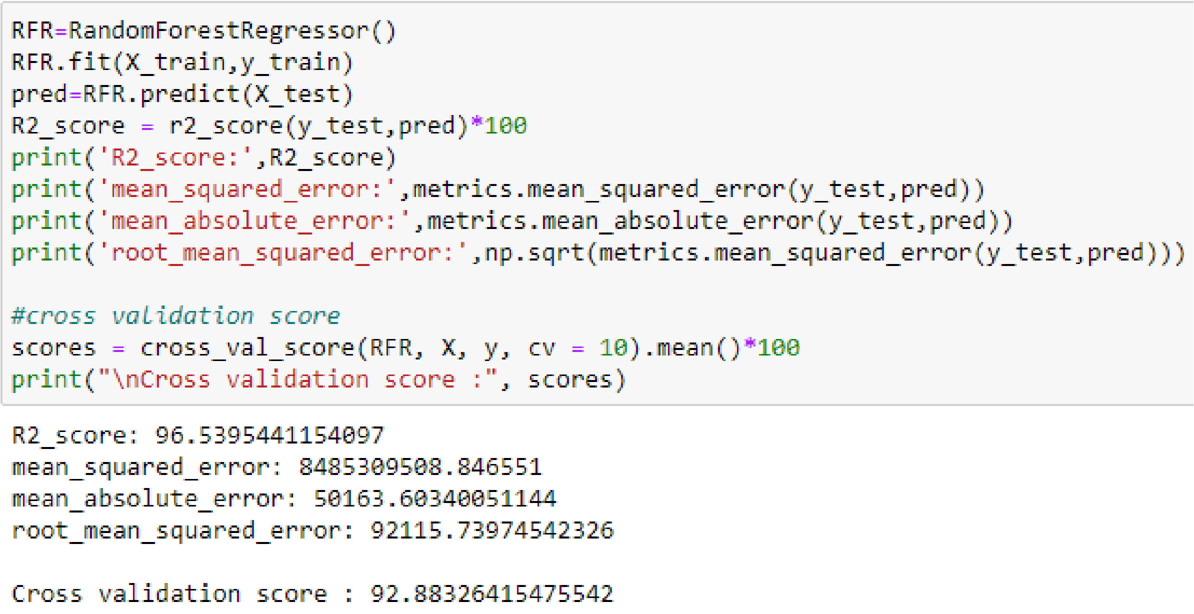
```
LR=LinearRegression()
LR.fit(X_train,y_train)
pred=LR.predict(X_test)
R2_score = r2_score(y_test,pred)*100
print('R2_score:',R2_score)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))

#cross validation score
scores = cross_val_score(LR, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)
```

```
R2_score: 66.42621487950294
mean_squared_error: 82325556987.89201
mean_absolute_error: 180679.5526370086
root_mean_squared_error: 286924.30532788957

Cross validation score : 59.86569683426731
```

2) Random Forest Regression:



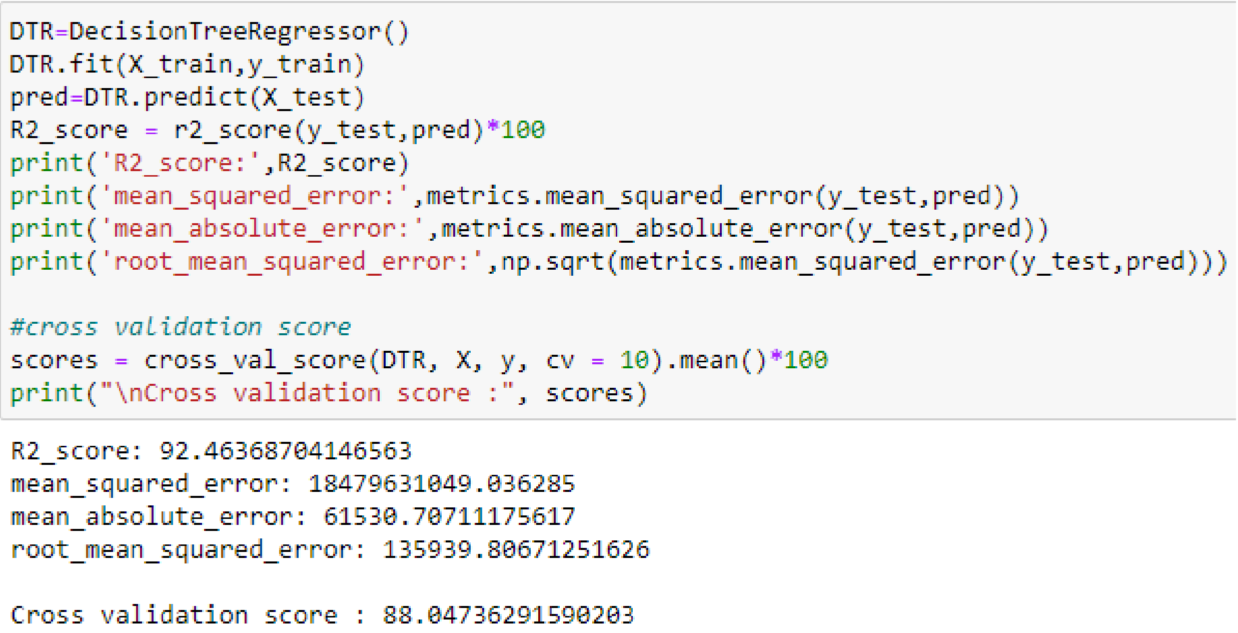
```
RFR=RandomForestRegressor()
RFR.fit(X_train,y_train)
pred=RFR.predict(X_test)
R2_score = r2_score(y_test,pred)*100
print('R2_score:',R2_score)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))

#cross validation score
scores = cross_val_score(RFR, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)
```

```
R2_score: 96.5395441154097
mean_squared_error: 8485309508.846551
mean_absolute_error: 50163.60340051144
root_mean_squared_error: 92115.73974542326

Cross validation score : 92.88326415475542
```

3) Decision Tree Regression:



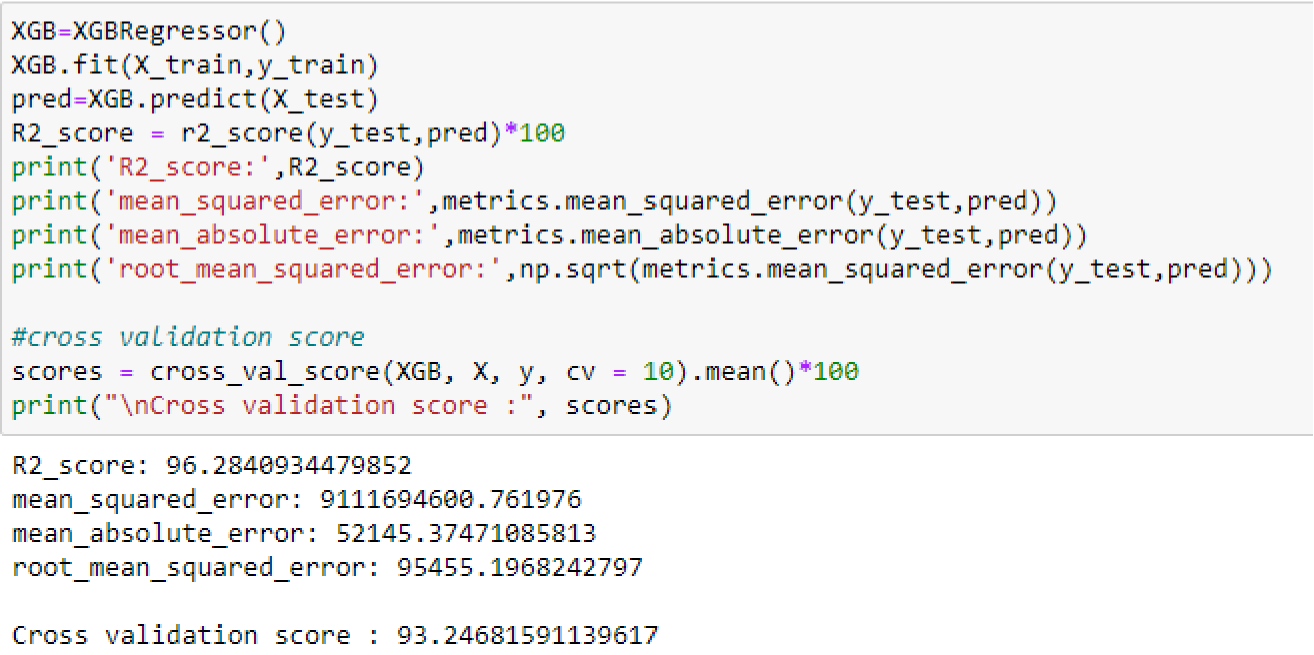
```
DTR=DecisionTreeRegressor()
DTR.fit(X_train,y_train)
pred=DTR.predict(X_test)
R2_score = r2_score(y_test,pred)*100
print('R2_score:',R2_score)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))

#cross validation score
scores = cross_val_score(DTR, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)
```

```
R2_score: 92.46368704146563
mean_squared_error: 18479631049.036285
mean_absolute_error: 61530.70711175617
root_mean_squared_error: 135939.80671251626

Cross validation score : 88.04736291590203
```

4) XGB Regression:



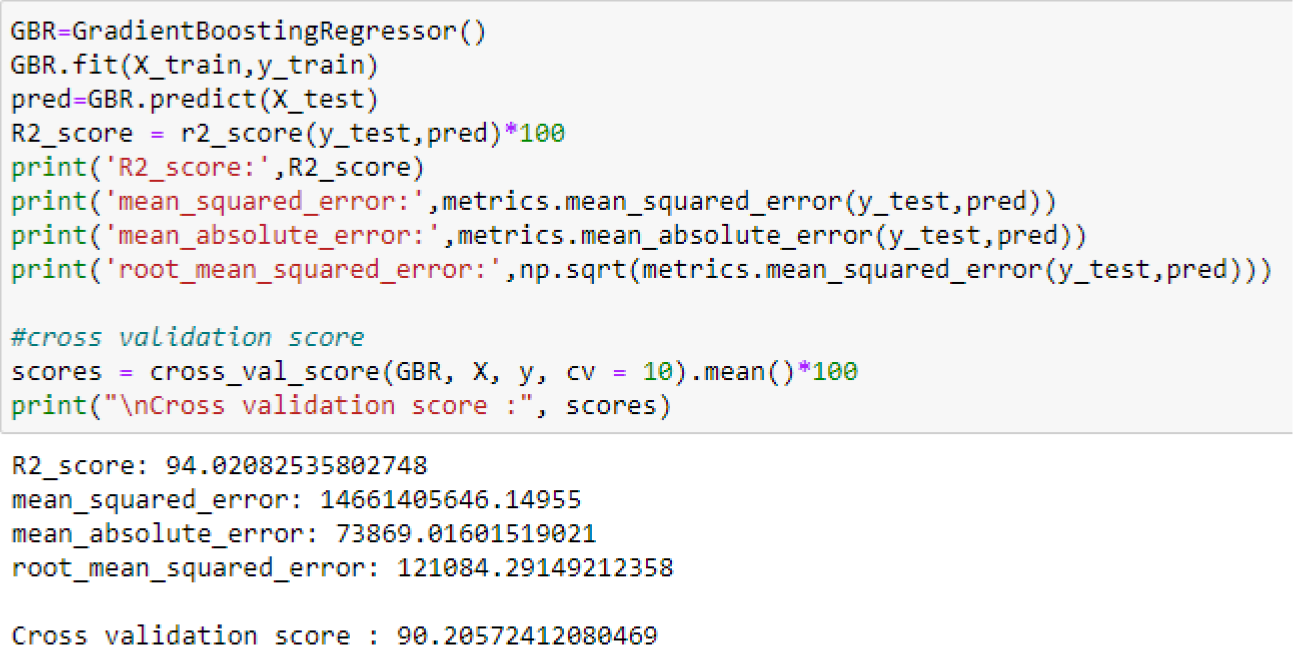
```
XGB=XGBRegressor()
XGB.fit(X_train,y_train)
pred=XGB.predict(X_test)
R2_score = r2_score(y_test,pred)*100
print('R2_score:',R2_score)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))

#cross validation score
scores = cross_val_score(XGB, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)
```

```
R2_score: 96.2840934479852
mean_squared_error: 9111694600.761976
mean_absolute_error: 52145.37471085813
root_mean_squared_error: 95455.1968242797

Cross validation score : 93.24681591139617
```

5) Gradient Boosting Regression:



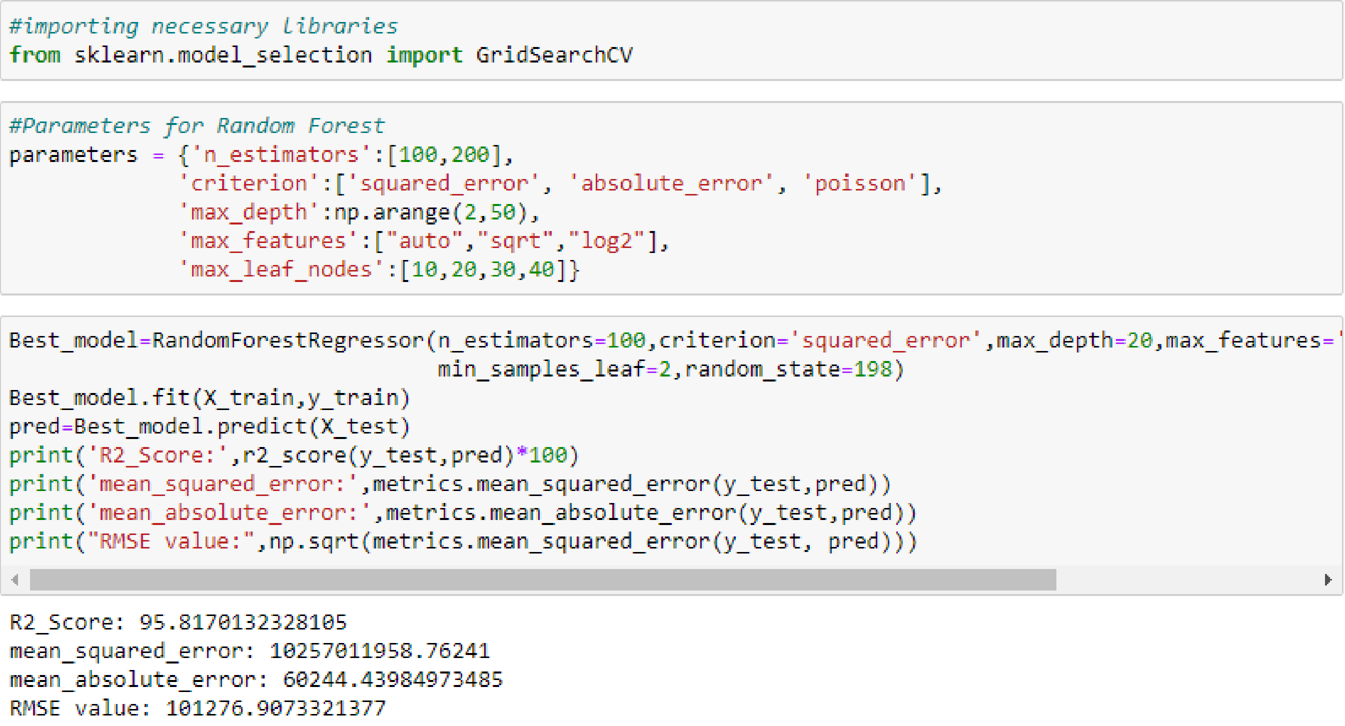
```
GBR=GradientBoostingRegressor()
GBR.fit(X_train,y_train)
pred=GBR.predict(X_test)
R2_score = r2_score(y_test,pred)*100
print('R2_score:',R2_score)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print('root_mean_squared_error:',np.sqrt(metrics.mean_squared_error(y_test,pred)))

#cross validation score
scores = cross_val_score(GBR, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)
```

```
R2_score: 94.02082535802748
mean_squared_error: 14661405646.14955
mean_absolute_error: 73869.01601519021
root_mean_squared_error: 121084.29149212358

Cross validation score : 90.20572412080469
```

2. Hyper Parameter Tuning:



```
#importing necessary libraries
from sklearn.model_selection import GridSearchCV
```

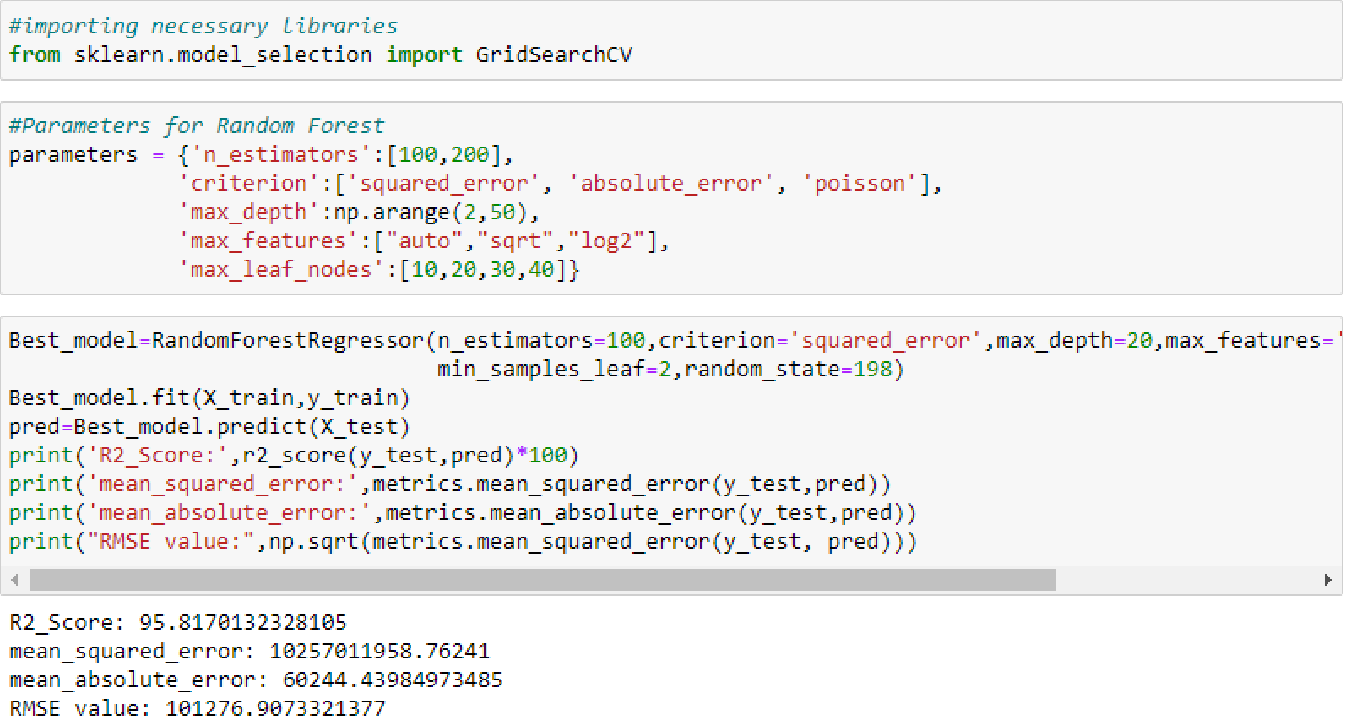
```
#Parameters for Random Forest
parameters = {'n_estimators':[100,200],
             'criterion':['squared_error', 'absolute_error', 'poisson'],
             'max_depth':np.arange(2,50),
             'max_features':["auto","sqrt","log2"],
             'max_leaf_nodes':[10,20,30,40]}
```

```
Best_model=RandomForestRegressor(n_estimators=100,criterion='squared_error',max_depth=20,max_features='
                                 min_samples_leaf=2,random_state=198)
Best_model.fit(X_train,y_train)
pred=Best_model.predict(X_test)
print('R2_Score:',r2_score(y_test,pred)*100)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print("RMSE value:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

```
R2_Score: 95.8170132328105
mean_squared_error: 10257011958.76241
mean_absolute_error: 60244.43984973485
RMSE value: 101276.9073321377
```

5. Saving the model and Prediction:

- I have saved my best model using pickle as follows.



```
#importing necessary libraries
from sklearn.model_selection import GridSearchCV
```

```
#Parameters for Random Forest
parameters = {'n_estimators':[100,200],
             'criterion':['squared_error', 'absolute_error', 'poisson'],
             'max_depth':np.arange(2,50),
             'max_features':["auto","sqrt","log2"],
             'max_leaf_nodes':[10,20,30,40]}
```

```
Best_model=RandomForestRegressor(n_estimators=100,criterion='squared_error',max_depth=20,max_features='
                                 min_samples_leaf=2,random_state=198)
Best_model.fit(X_train,y_train)
pred=Best_model.predict(X_test)
print('R2_Score:',r2_score(y_test,pred)*100)
print('mean_squared_error:',metrics.mean_squared_error(y_test,pred))
print('mean_absolute_error:',metrics.mean_absolute_error(y_test,pred))
print("RMSE value:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

```
R2_Score: 95.8170132328105
mean_squared_error: 10257011958.76241
mean_absolute_error: 60244.43984973485
RMSE value: 101276.9073321377
```

3.6 Interpretation of the Results

- ✓ The dataset was scrapped from cardekho website.
- ✓ The dataset was very challenging to handle it had 20 features with 12608 samples.
- ✓ Firstly, the datasets were having any null values, so I have used imputation method to replace the NaN values.
- ✓ I then plotted a few graphs to analyze and visualize the data in a better manner, I used bar graphs and reg plots to see the relationship between the features and the target
- ✓ As there were a large number of outliers in most of the features, I used Z-score method to remove the outliers and reduced the skewness in the data by using PowerTransformer method
- ✓ Next I used StandardScaler to scale the data, this helps the model in not getting biased
- ✓ I then used the data and built various regression models to get the best model
- ✓ After building 5 different models, I found that Random Forest Regressor gave us the best r2 score of 96.53%, and tried to improve the score by hyper parameter tuning
- ✓ I then saved the model and checked the prediction results againt the actual results

4. CONCLUSION

4.1 Key Findings and Conclusions of the Study

In this project report, I have used machine learning algorithms to predict the used car prices. We have mentioned the step-by-step procedure to analyse the dataset and finding the correlation between the features. Thus, we can select the features which are correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a hyper parameter tuning was done to the best model and the accuracy has been improved. Hence, we calculated the performance of each model using different performance metrics and compared them based on those metrics. Then we have also saved the best model and predicted the used car price. It was good the predicted and actual values were almost same.

4.2 Learning Outcomes of the Study in respect of Data Science

I found that the dataset was quite interesting to handle as it contains all types of data in it and it was self-scrapped from www.cardekho.com website using selenium.
Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analyzed. New analytical techniques of machine learning can be used in used car price research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove unrealistic values and null values. This study is an exploratory attempt to use five machine learning algorithms in estimating used car price prediction, and then compare their results.

To conclude, the application of machine learning in predicting used car price is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to crediting online platforms, and presenting an alternative approach to the valuation of used car price. Future direction of research may consider incorporating additional used car data from a larger economical background with more features.

4.3 Limitations of this work and Scope for Future Work

- ✓ First drawback is scrapping the data as it is fluctuating process.
- ✓ Followed by a greater number of outliers and skewness these two will reduce our model accuracy.
- ✓ Also, we have tried best to deal with outliers, skewness and null values. So, it looks quite good that we have achieved an accuracy of 92.29% even after dealing all these drawbacks.
- ✓ Also, this study will not cover all Regression algorithms instead, it is focused on the chosen algorithm, starting from the basic Ensembling techniques to the advanced ones