

**PROJECT REPORT ON:**

**“Used-Car Price Prediction Project”**

**SUBMITTED BY**

**Sumit Dhandhania**

**ACKNOWLEDGMENT**

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. Mr. Shwetank Mishra (SME Flip Robo) who has inspired me in so many aspects and also encouraged me a lot with his valuable words and with his unconditional support I have ended up with a beautiful Project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who provided me with assistance directly or indirectly to complete the project.

**Contents:**

1. Introduction
   1. Business Problem Framing:
   2. Conceptual Background of the Domain Problem
   3. Review of Literature
   4. Motivation for the Problem Undertaken
2. Analytical Problem Framing
   1. Mathematical/ Analytical Modelling of the Problem
   2. Data Sources and their formats
   3. Data Pre-processing Done
   4. Data Inputs-Logic-Output Relationships
   5. Hardware and Software Requirements and Tools Used
3. Data Analysis and Visualization
   1. Identification of possible problem-solving approaches (methods)
   2. Testing of Identified Approaches (Algorithms)
   3. Key Metrics for success in solving problem under consideration
   4. Visualization
   5. Run and evaluate selected models
   6. Interpretation of the Results
4. Conclusion
5. Key Findings and Conclusions of the Study
6. Learning Outcomes of the Study in respect of Data Science
7. 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 colour, 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. Conceptual Background of the Domain Problem

The prices of new cars in the industry is 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. Review of Literature

1. Customer retention survives when the companies can fulﬁll customer expectations and
2. additionally maintain it in long-term relationships to ensure long-term buying decisions
3. [13–15]. The topic of customer retention is argued in business economics commonly
4. within the perspective of relationship marketing, which considers customer relation-
5. ships as one of the primary concerns with the long-term objective of developing and
6. maintaining them [16–18]. Many previous studies indicated that companies should
7. always manage customer satisfaction to achieve the retention stage. According to [19]
8. “satisfaction is an overall customer attitude towards a service provider”.In[20],
9. authors added that satisfaction is an emotional reaction regarding what customers
10. expect and what they receive, including the fulﬁllment of needs and goals. Customer
11. retention states a desired outcome in the future to satisfaction, so long-term of rela-
12. tionship is demonstrated by satisfaction. Although customer satisfaction does not
13. guarantee repurchase, it still plays a vital role in ensuring customer retention. While
14. many studies on customer retention had long focused on customer satisfaction, addi-
15. tional factors are stated as an inﬂuence in customer retention, such as trust and com-
16. mitment. [21], in “The Commitment-Trust Theory of Relationship Marketing,”which
17. is the most inﬂuential Relationship Marketing, suggests that the center of successful
18. relationship marketing is the relationship of commitment and trust. They urged the
19. importance of commitment and trust that leads to build a positive correlation between
20. company and customers and encourage efﬁciency, productivity, and effectiveness. The
21. degree of trust between service provider and customer is signiﬁcantly inﬂuenced by the
22. quality of the service, which results in an effective commitment to the provider, and
23. Customer retention survives when the companies can fulﬁll customer expectations and
24. additionally maintain it in long-term relationships to ensure long-term buying decisions
25. [13–15]. The topic of customer retention is argued in business economics commonly
26. within the perspective of relationship marketing, which considers customer relation-
27. ships as one of the primary concerns with the long-term objective of developing and
28. maintaining them [16–18]. Many previous studies indicated that companies should
29. always manage customer satisfaction to achieve the retention stage. According to [19]
30. “satisfaction is an overall customer attitude towards a service provider”.In[20],
31. authors added that satisfaction is an emotional reaction regarding what customers
32. expect and what they receive, including the fulﬁllment of needs and goals. Customer
33. retention states a desired outcome in the future to satisfaction, so long-term of rela-
34. tionship is demonstrated by satisfaction. Although customer satisfaction does not
35. guarantee repurchase, it still plays a vital role in ensuring customer retention. While
36. many studies on customer retention had long focused on customer satisfaction, addi-
37. tional factors are stated as an inﬂuence in customer retention, such as trust and com-
38. mitment. [21], in “The Commitment-Trust Theory of Relationship Marketing,”which
39. is the most inﬂuential Relationship Marketing, suggests that the center of successful
40. relationship marketing is the relationship of commitment and trust. They urged the
41. importance of commitment and trust that leads to build a positive correlation between
42. company and customers and encourage efﬁciency, productivity, and effectiveness. The
43. degree of trust between service provider and customer is signiﬁcantly inﬂuenced by the
44. quality of the service, which results in an effective commitment to the provider, and

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here’s where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone’s needs.Real Estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. Real Estate Property is not only the basic need of a man but today it also represents the richness and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic index.

An attempt has been made in this article to review the available literature in the

area of microfinance. Approaches to microfinance, issues related to measuring social

impact versus profitability of MFIs, issue of sustainability, variables impacting

sustainability, affect of regulations of profitability and impact assessment of MFIs

have been summarized in the above article. We hope that the above review of

literature will provide researchers a platform for further research and help the

An attempt has been made in this article to review the available literature in the

area of microfinance. Approaches to microfinance, issues related to measuring social

impact versus profitability of MFIs, issue of sustainability, variables impacting

sustainability, affect of regulations of profitability and impact assessment of MFIs

have been summarized in the above article. We hope that the above review of

literature will provide researchers a platform for further research and help the

An attempt has been made in this article to review the available literature in the

area of microfinance. Approaches to microfinance, issues related to measuring social

impact versus profitability of MFIs, issue of sustainability, variables impacting

sustainability, affect of regulations of profitability and impact assessment of MFIs

have been summarized in the above article. We hope that the above review of

literature will provide researchers a platform for further research and help the

An attempt has been made in this article to review the available literature in the

area of microfinance. Approaches to microfinance, issues related to measuring social

impact versus profitability of MFIs, issue of sustainability, variables impacting

sustainability, affect of regulations of profitability and impact assessment of MFIs

have been summarized in the above article. We hope that the above review of

literature will provide researchers a platform for further research and help the

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, colour, 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. 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**

* 1. Mathematical/ Analytical Modelling of the Problem

As a first step I have scrapped the required data from carwale.com website. I have fetched data for different locations and saved it to csv format.

In this particular problem, I have Price as my target column and it was a continuous column. So clearly it is a regression problem and I have to use all regression algorithms while building the model. There were no null values in the dataset. 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. 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 price using saved model.

2.2 Data Sources and their formats

The data was collected from carwale.com.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 32000 rows and 8 columns including target. In this dataset 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:**

Independent Variables:

- Unnamed: 0 - A unique continuous identifier

- Brand- A categorical variable describing the company brand of the car

- Model - A categorical variable describing the model of the car

- Kilometers Driven - A continuous variable describing the total kilmeters car has ran

- Fuel - A categorical variable describing the fuel type of the car

- City - A categorical variable describing the city car is located in

- Year - A numeric categorical variable describing the year of manufacturing

Dependent Variable :

Price - A continuous variable describing the price of the car

* 1. Data Preprocessing Done
* As a first step I have scrapped the required data using selenium from carwale.com 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 droped 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 usefull 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 distplot 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.
  1. 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 — 8 GB or above

3. SSD — 250GB or above

**Software/s required**: -

1.Anaconda

**Libraries** **required: -**

To run the program and to build the model we need some basic libraries as follows:

**# Importing necessary packages**

**# Importing fundamental packages**

**import warnings**

**warnings.filterwarnings("ignore")**

**import pyforest**

**## For visualization**

**import matplotlib.pyplot as plt**

**%matplotlib inline**

**import seaborn as sns**

**from plotly.subplots import make\_subplots**

**import plotly.graph\_objects as go**

**import plotly.express as px**

**## Data Pre-Processing Packages**

**from scipy.stats import chi2\_contingency**

**from sklearn import metrics**

**from sklearn import preprocessing**

**from sklearn.preprocessing import StandardScaler**

**from sklearn.preprocessing import MinMaxScaler**

**from sklearn.preprocessing import RobustScaler**

**from scipy.stats import zscore**

**from sklearn.decomposition import PCA**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.preprocessing import OrdinalEncoder**

**from sklearn.preprocessing import LabelBinarizer**

**from sklearn.preprocessing import LabelEncoder**

**from sklearn.experimental import enable\_iterative\_imputer**

**from sklearn.impute import IterativeImputer**

**from sklearn.impute import KNNImputer**

**from sklearn.datasets import make\_blobs**

**from sklearn.cluster import DBSCAN**

**from collections import Counter**

**from sklearn.cluster import KMeans**

**from sklearn.cluster import MiniBatchKMeans**

**from sklearn.preprocessing import PowerTransformer**

**from sklearn.model\_selection import KFold**

**from sklearn.feature\_selection import RFE**

**from sklearn.preprocessing import scale**

**from sklearn.datasets import make\_classification**

**## To create copy of data**

**import copy**

**## Pipeline Packages**

**from sklearn.pipeline import Pipeline**

**from sklearn.pipeline import make\_pipeline**

**## Ensemble Learning Algorithms Packages**

**from sklearn import tree**

**import xgboost as xgb**

**from sklearn.model\_selection import RepeatedStratifiedKFold**

**from sklearn.model\_selection import RandomizedSearchCV**

**from scipy.stats import loguniform**

**from sklearn.linear\_model import Ridge**

**from sklearn.linear\_model import LinearRegression**

**from sklearn.linear\_model import Lasso**

**from sklearn.linear\_model import ElasticNet**

**from sklearn.svm import SVR**

**from sklearn.ensemble import RandomForestRegressor**

**from sklearn.tree import DecisionTreeRegressor**

**from sklearn.ensemble import AdaBoostRegressor**

**from sklearn.ensemble import GradientBoostingRegressor**

**from xgboost import XGBRegressor**

**from sklearn.ensemble import BaggingRegressor**

**## Evaluation Metrics Packages**

**import statsmodels.api as sm**

**from scipy.stats import f**

**from sklearn.metrics import r2\_score**

**from statsmodels.formula.api import ols**

**from sklearn.model\_selection import GridSearchCV**

**from sklearn.model\_selection import RandomizedSearchCV**

**from sklearn.utils import resample**

**from sklearn import metrics**

**from statsmodels.stats.outliers\_influence import variance\_inflation\_factor**

**from sklearn import preprocessing**

**from sklearn.metrics import silhouette\_score**

**from sklearn.metrics import accuracy\_score**

**from sklearn.model\_selection import cross\_val\_score**

**from sklearn.metrics import mean\_squared\_error as mse**

**from sklearn.metrics import classification\_report**

**# Saving the model**

**import pickle**

With this sufficient libraries we can go ahead with our model building.

**3.Data Analysis and Visualization**

3.1 Identification of possible problem-solving approaches (methods)

* 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 MinMax scaler the data. After scaling we have to remove multicolinearity using VIF. Then followed by model building with all Regression algorithms
  1. Testing of Identified Approaches (Algorithms)

Since 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 DecisionTreeRegressor as a 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 are the list of Regression algorithms I have used in my project.

* RandomForestRegressor
* XGBRegressor
* ExtraTreesRegressor
* GradientBoostingRegressor
* DecisionTreeRegressor
* BaggingRegressor
  1. Key Metrics for success in solving problem under consideration

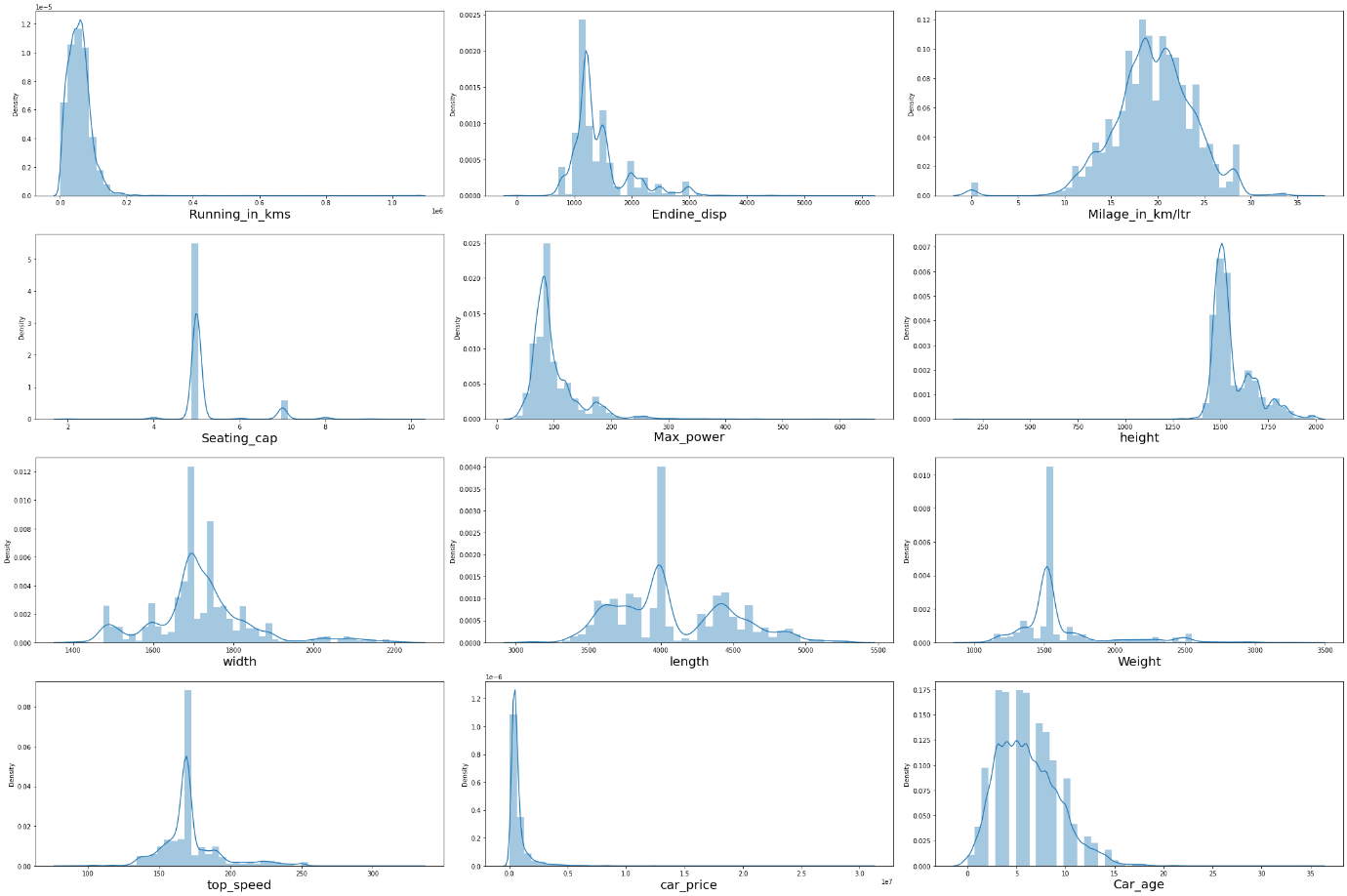
I have used the following metrics for evaluation:

* 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

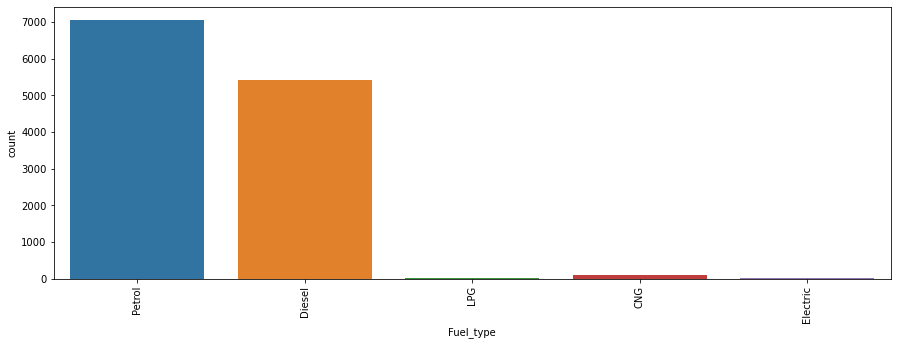
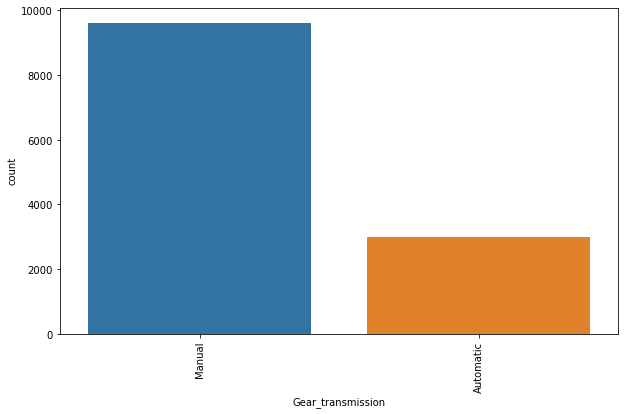
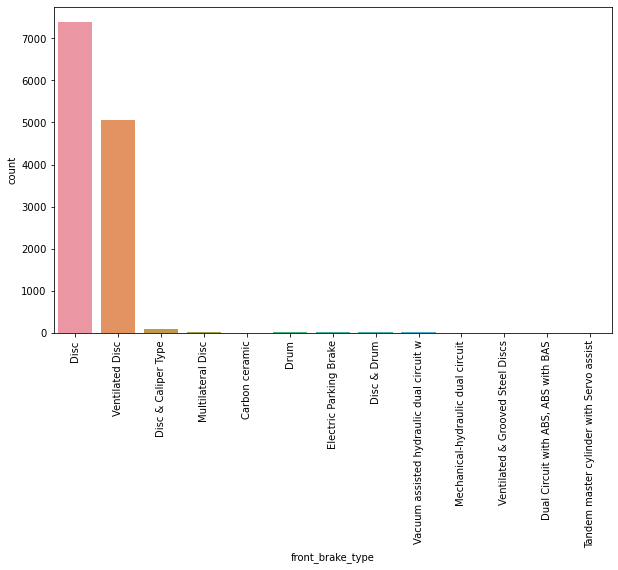
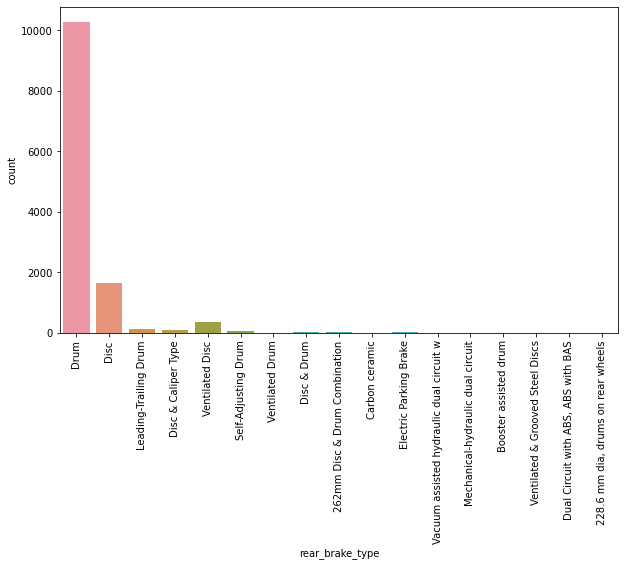
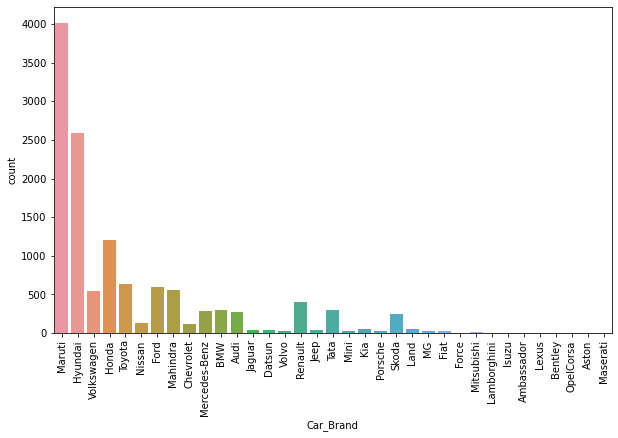
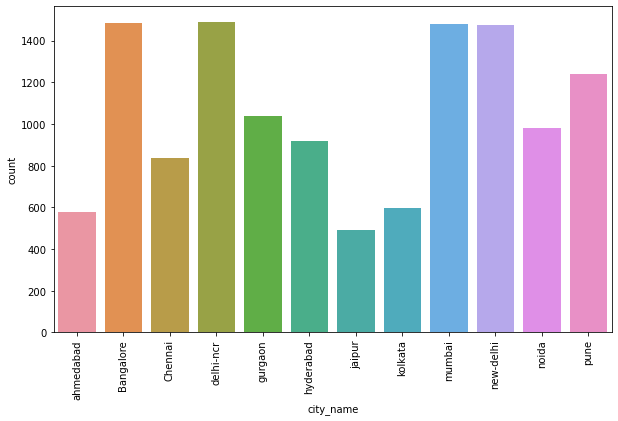
I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and reg plot, strip plot for bivariate analysis.

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

****

**Observations:**

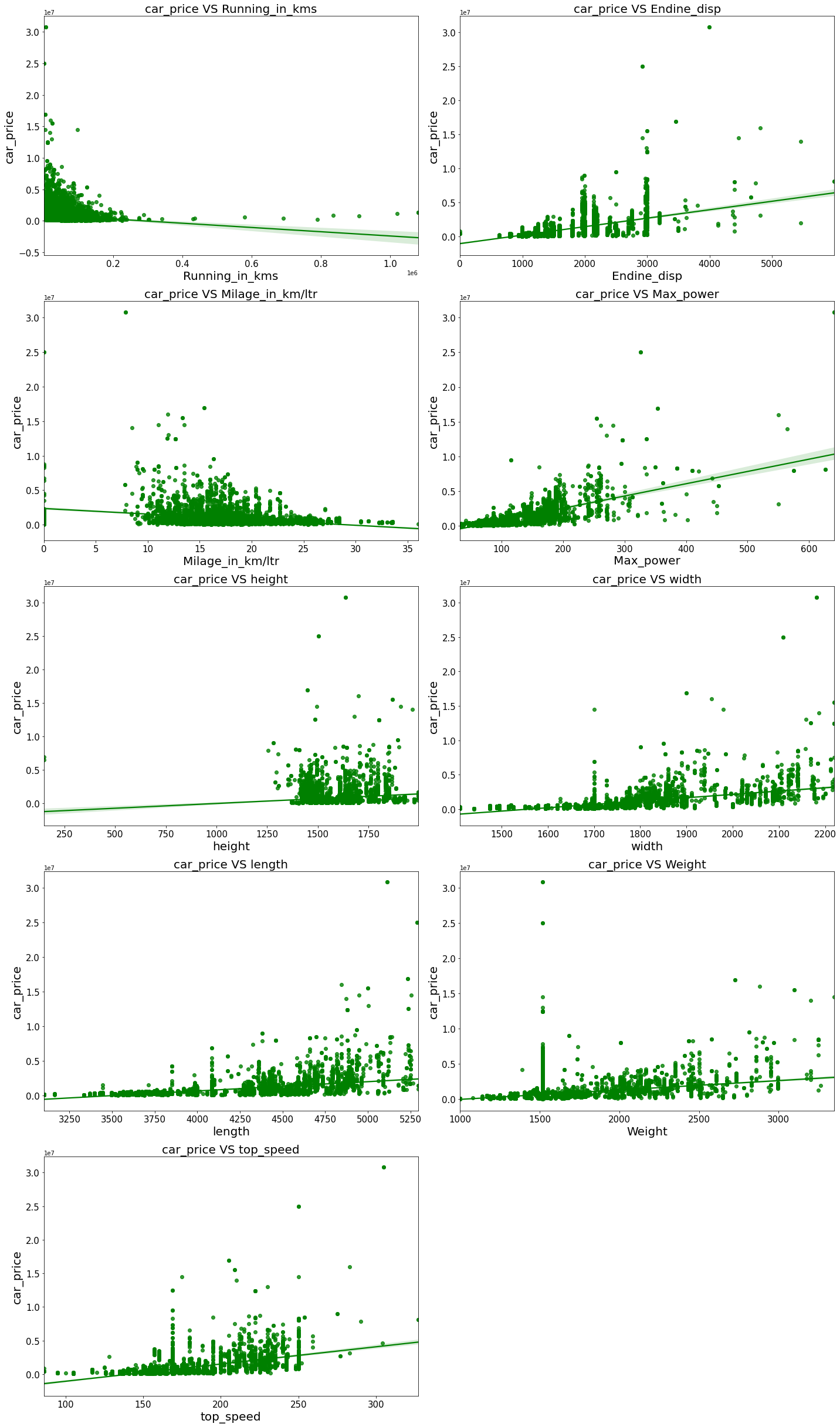
* We can clearly see that there is skewness in most of the columns so we have to treat them using suitable methods.

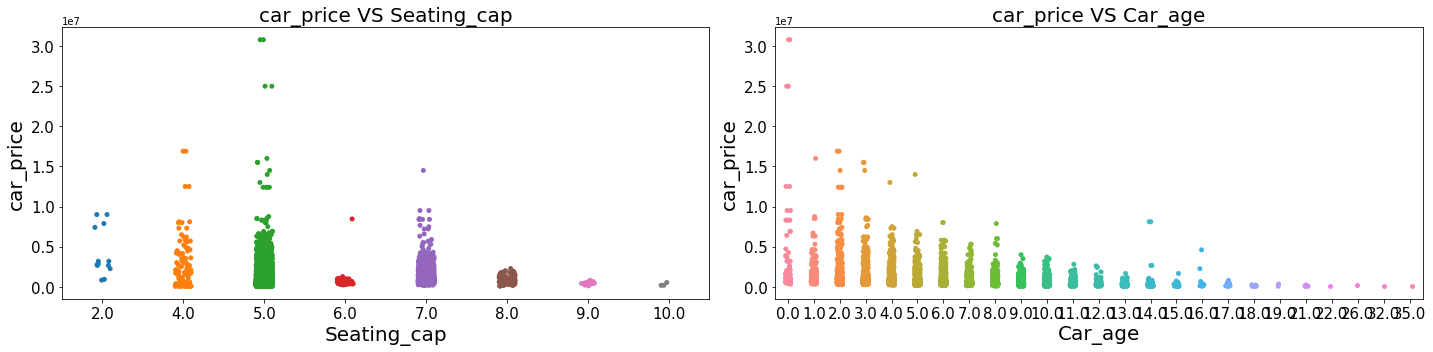
1. **Univariate analysis for categorical column:**

**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.

1. **Bivariate analysis for numerical columns:**

****

****

**Observations:**

- Maximum cars use "Petrol" as fuel type. Total= 21000 (65.63%)

- Least number of cars use "CNG" as Fuel type. Total= 1000 (3.13%)

- Maximum cars are located in Mumbai (Total=10,000=31.25%), followed by Hyderabad (Total=8000=25%), followed by Delhi (Total=4000=12.50%)

- Least number of cars are located in Chennai, Kanpur and Lucknow (Total= 1000 = 3.13% each)

- Major sales are happening for the cars manufactured in the year 2018 (Total= 7000, 21.86%) followed by 2021 & 2017 (Total= 5000, 15.63% each)

- Least sales are happening for the cars manufactured in the year 2020, 2011 & 2008 (Total= 1000 , 3.13% each)

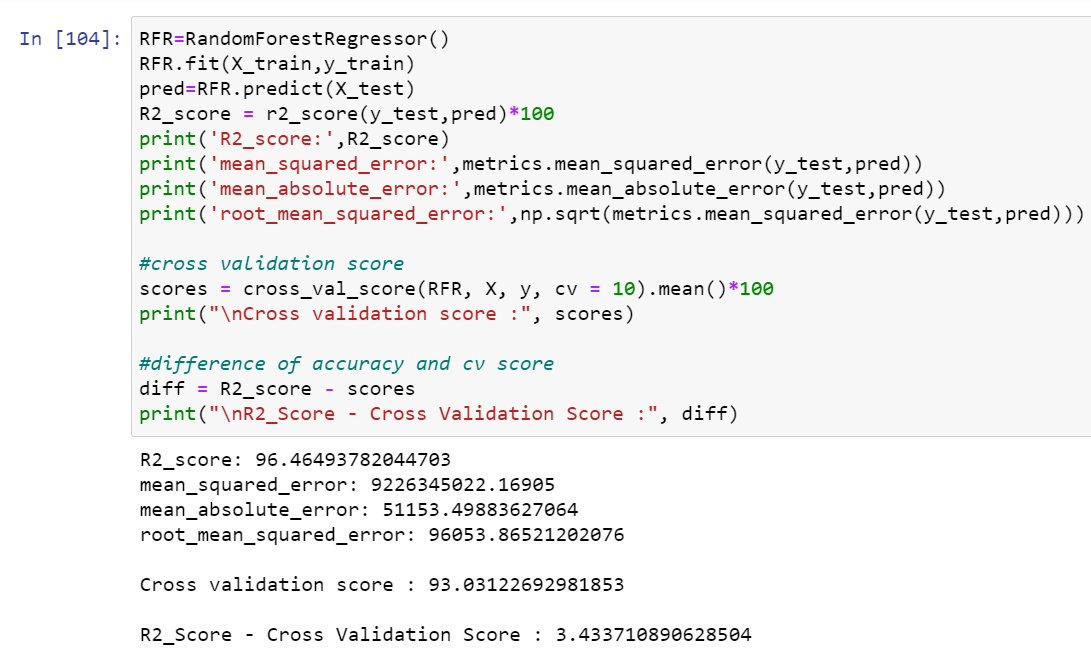
**Observations:**

* 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 break are costlier compared to other cars.
* Cars with carbon Ceramic rear break 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

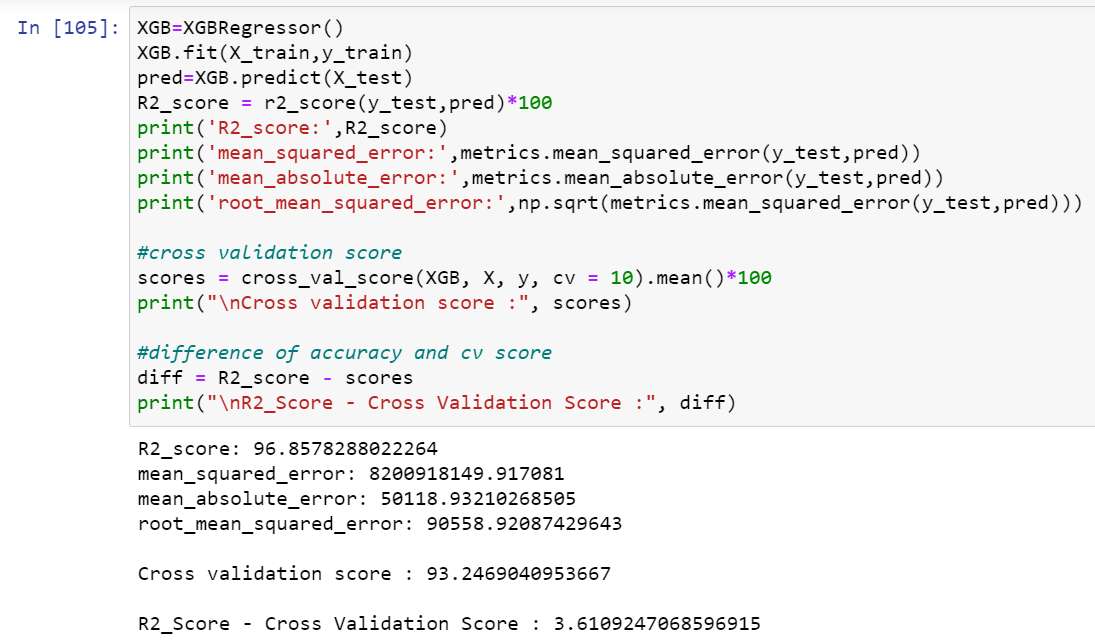
**1. Model Building:**

1. **RandomForestRegressor:**



* RandomForestRegressor has given me 96.46% r2\_score and the difference between r2\_score and cross validation score is 3.43%, but still we have to look into multiple models.

1. **XGBRegressor:**

****

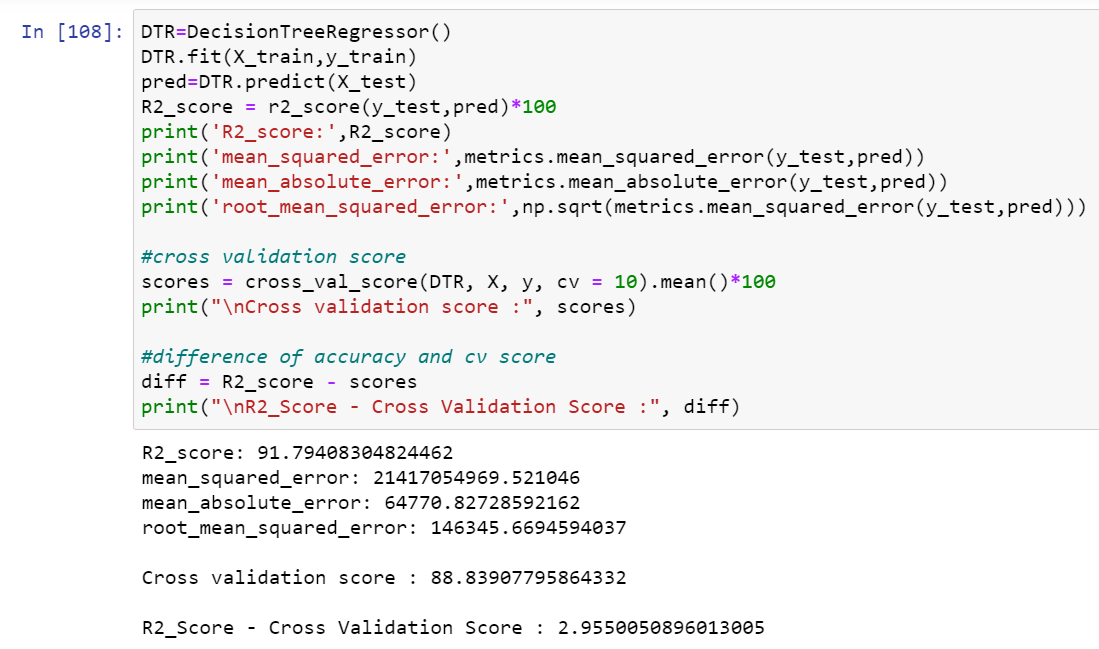
* XGBRegressor is giving me 96.86% r2\_score and the difference between r2\_score and cross validation score is 3.61%.

1. **GradientBoostingRegressor:**

****

* GradientBoostingRegressor is giving me 94.93% r2\_score and the difference between r2\_score and cross validation score is 4.74%.

1. **DecisionTreeRegressor:**

****

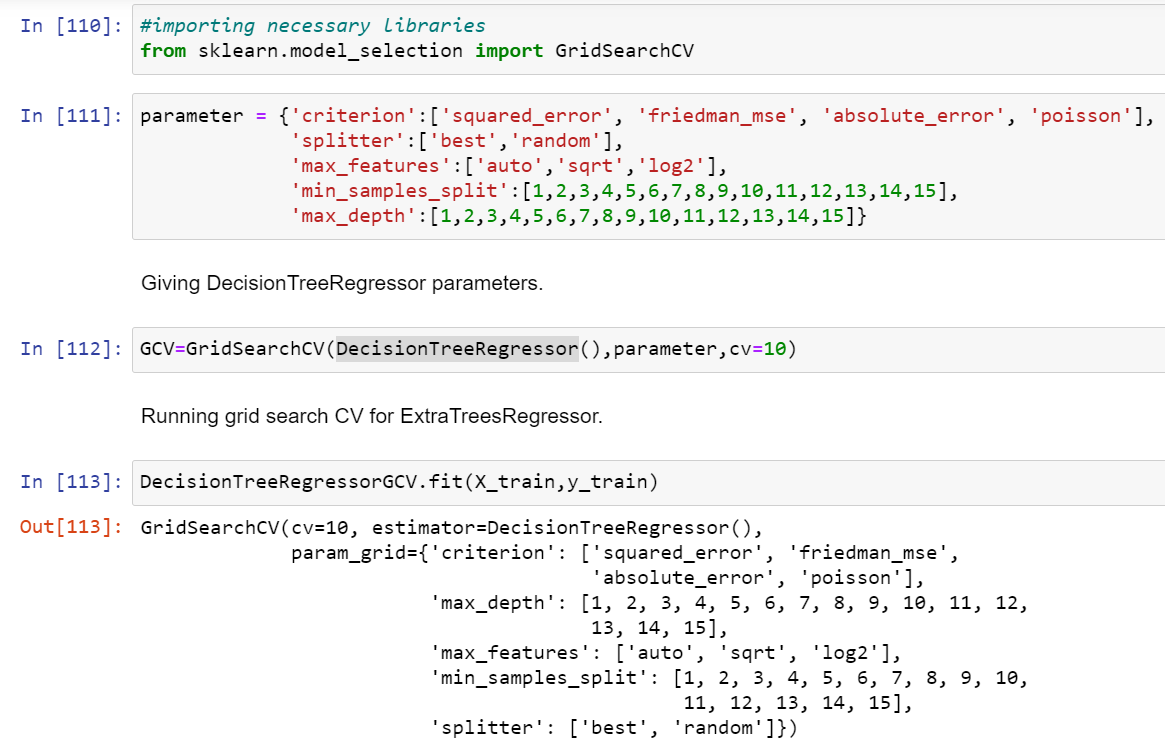
* DecisionTreeRegressor is giving me 91.79% r2\_score and the difference between r2\_score and cross validation score is 2.96%.

1. **BaggingRegressor:**

****

* BaggingRegressor is giving me 95.68% r2\_score and the difference between r2\_score and cross validation score is 3.07%.
* **By looking into the difference of r2\_score and cross validation score i found DecisionTreeRegressor as the best model with 91.79% r2\_score and the difference between r2\_score and cross validation score is 2.96%.**

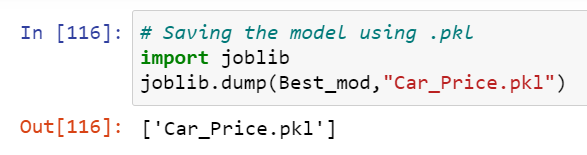
**3. Hyper Parameter Tunning:**

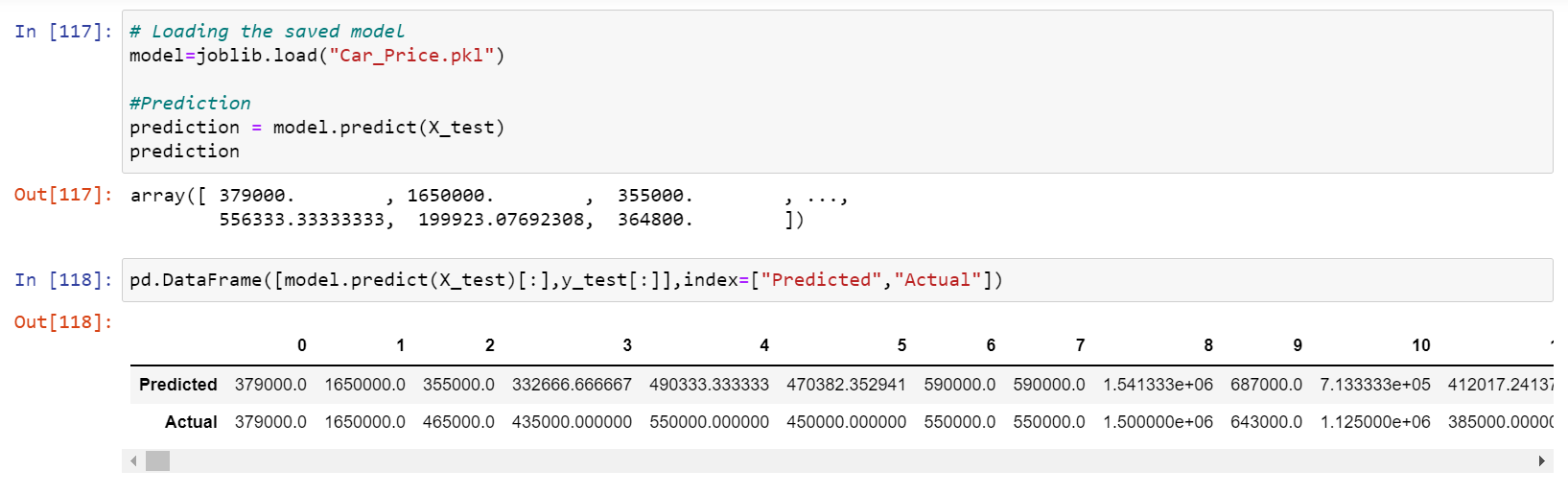
****

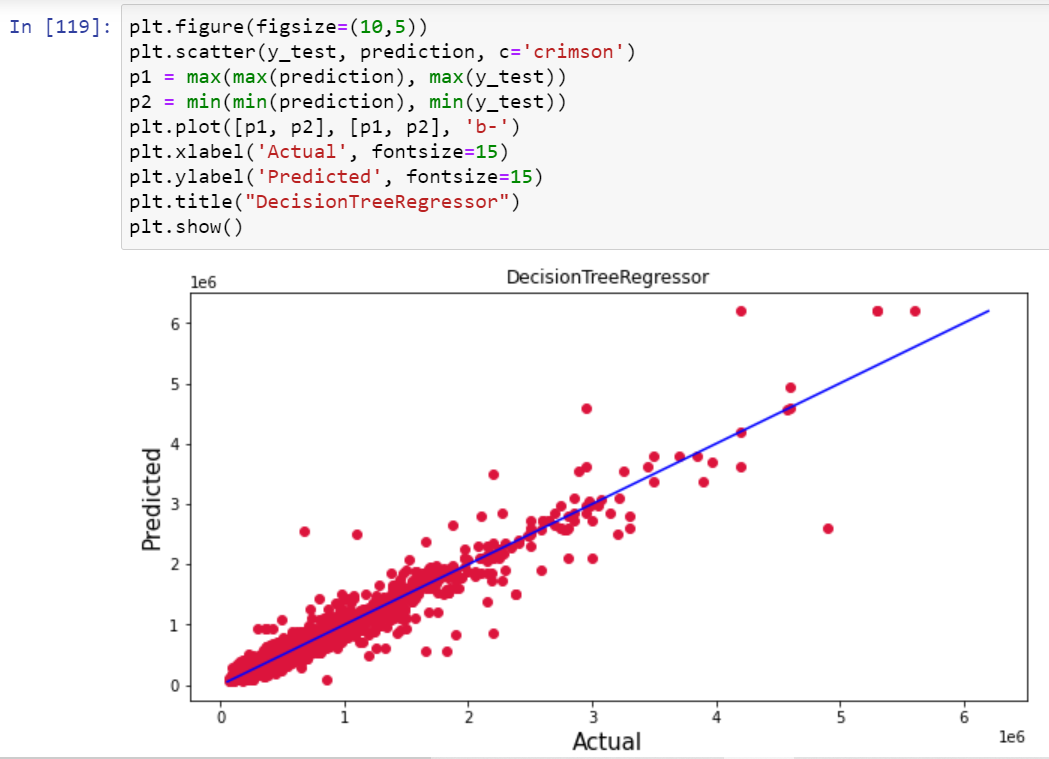
****

* **I have choosed all parameters of DecisionTreeRegressor, after tunning the model with best parameters I have incresed my model accuracy from 91.79% to 92.29%.**

1. **Saving the model and Predictions:**

* ****I have saved my best model using .pkl as follows**.**
* Now loading my saved model and predicting the price values.





* Plotting Actual vs Predicted, To get better insight. Bule line is the actual line and red dots are the predicted values.
  1. Interpretation of the Results
* The dataset was scrapped from carwale.com 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.
* And there was huge number of unnecessary entries in all the features so I have used feature extraction to get the required format of variables.
* And proper ploting for proper type of features will help us to get better insight on the data. I found both numerical columns and categorical columns in the dataset so I have choosen reg plot, strip plot and bar plot to see the relation between target and features.
* I notice a huge amount of outliers and skewness in the data so we have choose proper methods to deal with the outliers and skewness. If we ignore this outliers and skewness we may end up with a bad model which has less accuracy.
* Then scaling dataset has a good impact like it will help the model not to get baised. Since we have removed outliers and skewness from the dataset so we have to choose Standardisation.
* We have to use multiple models while building model using dataset as to get the best model out of it.
* And we have to use multiple metrics like mse, mae, rmse and r2\_score which will help us to decide the best model.
* I found Bagging Regressor as the best model with 100% r2\_score. Also I have improved the accuracy of the best model by running hyper parameter tunning.
* At last I have predicted the used car price using saved model. It was good!! that I was able to get the predictions near to actual values.

**4.CONCLUSION**

4.1 Key Findings and Conclusions of the Study

In this project report, we 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 tunning 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 carwale.com website using selenium. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. 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.

* 1. 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 a 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 ensemble techniques to the advanced ones.