FLIP ROBO

PROJECT REPORT ON:

" Car Price Prediction "

SUBMITTED BY

RAHUL RANJAN

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

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 has helped me directly or indirectly to complete the project.

Contents:

- Introduction
  - Business Problem Framing:
  - Conceptual Background of the Domain Problem
  - Review of Literature
  - Motivation for the Problem Undertaken

- Analytical Problem Framing
  - Mathematical/ Analytical Modeling of the Problem
  - Data Sources and their formats
  - Data Preprocessing Done
  - Data Inputs-Logic-Output Relationships
  - Hardware and Software Requirements and Tools Used

- Data Analysis and Visualization
  - Identification of possible problem-solving approaches (methods)
  - Testing of Identified Approaches (Algorithms)
  - Key Metrics for success in solving problem under consideration
  - Visualization
  - Run and Evaluate selected models
  - Interpretation of the Results

- Conclusion

- Key Findings and Conclusions of the Study
- Learning Outcomes of the Study in respect of Data Science
- 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 changes 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.

• 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 it's market value.

There are one of the biggest target group 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.

• Review of Literature

- Customer retention survives when the companies can fulfill customer expectations and
- additionally maintain it in long-term relationships to ensure long-term buying decisions
- [13–15]. The topic of customer retention is argued in business economics commonly
- within the perspective of relationship marketing, which considers customer relation-

- ships as one of the primary concerns with the long-term objective of developing and
- maintaining them [16–18]. Many previous studies indicated that companies should
- always manage customer satisfaction to achieve the retention stage. According to [19]
- “satisfaction is an overall customer attitude towards a service provider”. In [20],
- authors added that satisfaction is an emotional reaction regarding what customers
- expect and what they receive, including the fulfillment of needs and goals. Customer
- retention states a desired outcome in the future to satisfaction, so long-term of rela-
- tionship is demonstrated by satisfaction. Although customer satisfaction does not
- guarantee repurchase, it still plays a vital role in ensuring customer retention. While
- many studies on customer retention had long focused on customer satisfaction, addi-
- tional factors are stated as an influence in customer retention, such as trust and com-
- mitment. [21], in “The Commitment-Trust Theory of Relationship Marketing,” which
- is the most influential Relationship Marketing, suggests that the center of successful
- relationship marketing is the relationship of commitment and trust. They urged the
- importance of commitment and trust that leads to build a positive correlation between
- company and customers and encourage efficiency, productivity, and effectiveness. The
- degree of trust between service provider and customer is significantly influenced by the
- quality of the service, which results in an effective commitment to the provider, and
- Customer retention survives when the companies can fulfill customer expectations and
- additionally maintain it in long-term relationships to ensure long-term buying decisions

- [13–15]. The topic of customer retention is argued in business economics commonly
- within the perspective of relationship marketing, which considers customer relation-
- ships as one of the primary concerns with the long-term objective of developing and
- maintaining them [16–18]. Many previous studies indicated that companies should
- always manage customer satisfaction to achieve the retention stage. According to [19]
- "satisfaction is an overall customer attitude towards a service provider". In [20],
- authors added that satisfaction is an emotional reaction regarding what customers
- expect and what they receive, including the fulfillment of needs and goals. Customer
- retention states a desired outcome in the future to satisfaction, so long-term of rela-
- tionship is demonstrated by satisfaction. Although customer satisfaction does not
- guarantee repurchase, it still plays a vital role in ensuring customer retention. While
- many studies on customer retention had long focused on customer satisfaction, addi-
- tional factors are stated as an influence in customer retention, such as trust and com-
- mitment. [21], in "The Commitment-Trust Theory of Relationship Marketing," which
- is the most influential Relationship Marketing, suggests that the center of successful
- relationship marketing is the relationship of commitment and trust. They urged the
- importance of commitment and trust that leads to build a positive correlation between
- company and customers and encourage efficiency, productivity, and effectiveness. The
- degree of trust between service provider and customer is significantly influenced by the
- 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.

- 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

- ## Mathematical/ Analytical Modeling of the Problem

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

In this perticular problem I have car_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 was 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 website 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 ploting like distribution plot, bar plot, reg plot, strip plot and count plot. With these ploting 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 tunned 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 this perticular 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

Data Preprocessing 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 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 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.

• 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 :-

```
#importing required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
import warnings
warnings.filterwarnings('ignore')
```

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

- **import pandas as pd**: **pandas** is 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.ensemble import RandomForestRegressor
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.neighbors import KNeighborsRegressor
- from sklearn.ensemble import GradientBoostingRegressor
- from sklearn.ensemble import ExtraTreesRegressor
- from sklearn.metrics import classification_report
- from sklearn.metrics import accuracy_score
- from sklearn.model_selection import cross_val_score

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)

- 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 Standardisation to scale the data. After scaling we have to remove multicolinearity using VIF. Then followed by model building with all Regression algorithms

- ## 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 perticular 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
- K Neighbours Regressor
- ExtraTreesRegressor
- GradientBoostingRegressor
- DecisionTreeRegressor
- BaggingRegressor

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

I have used the following metrics for evaluation:

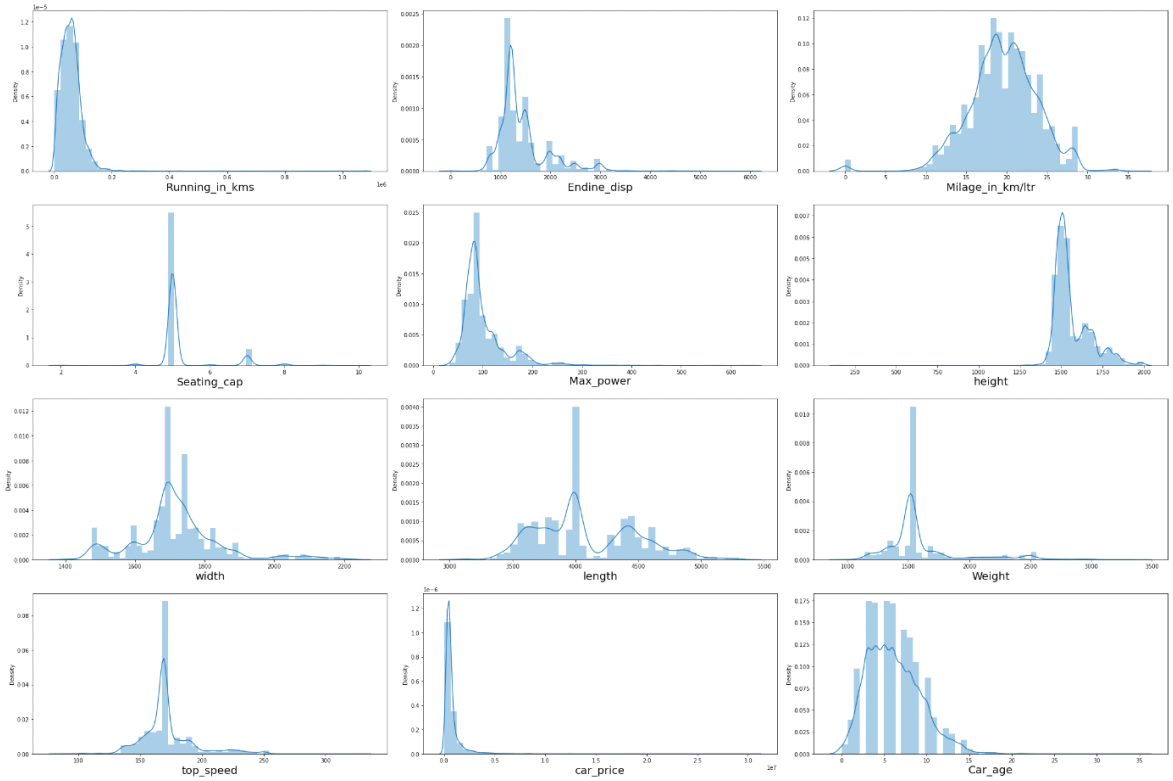
- I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.

- 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

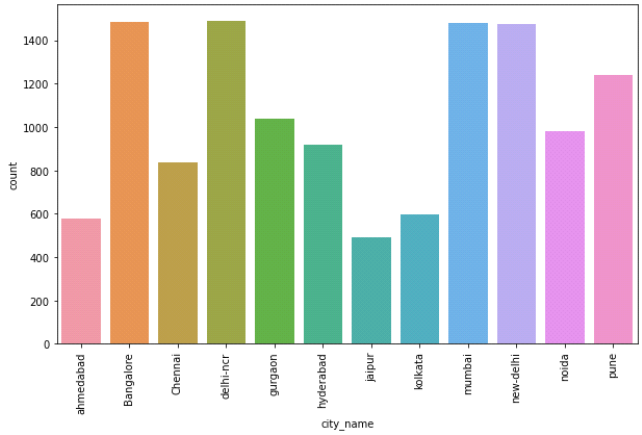
I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and reg plot, strip plot for bivariate analysis.

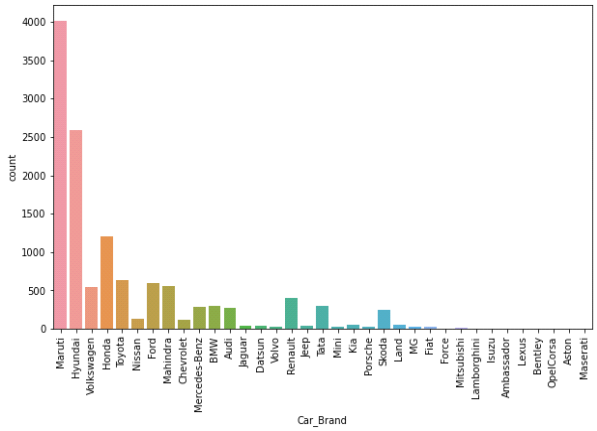
- **Univariate Analysis for numerical columns:**

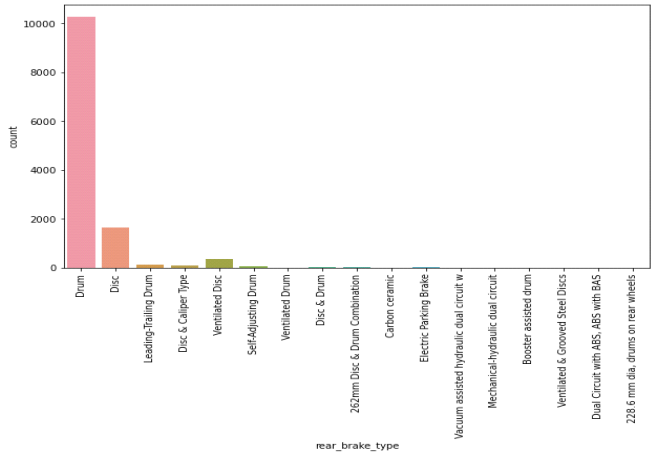


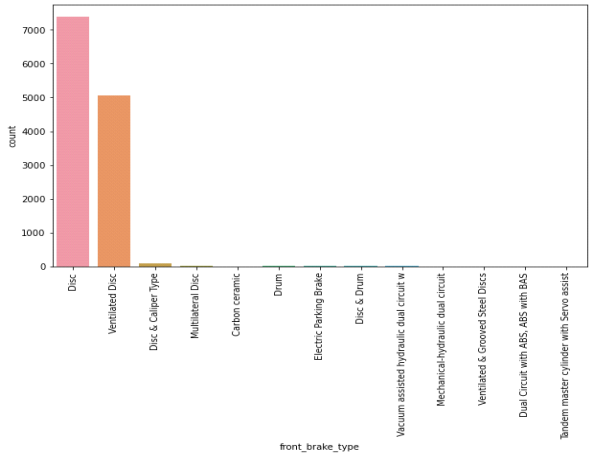
Observations:

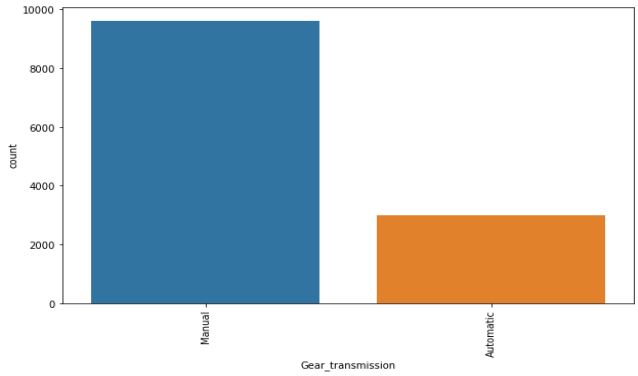
- We can clearly see that there is skewness in most of the columns so we have to treat them using suitable methods.

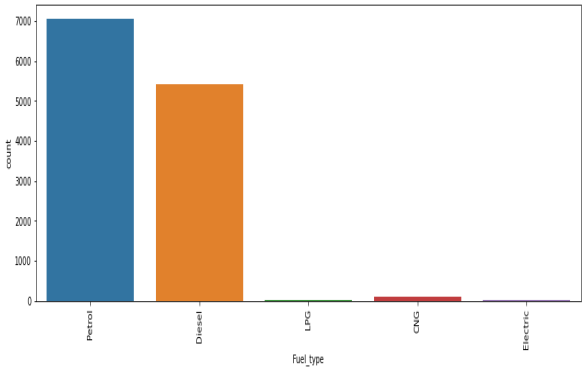

count
1400
1200
1000
800
600
400
200
0
ahmedabad
Bangalore
Chennai
delhi-ncr
gurgaon
hyderabad
jaipur
kolkata
mumbai
new-delhi
noida
pune
city_name

- 
count
4000
3500
3000
2500
2000
1500
1000
500
0
Maruti
Hyundai
Volkswagen
Honda
Toyota
Nissan
Ford
Mahindra
Chevrolet
Mercedes-Benz
BMW
Audi
Jaguar
Datsun
Volvo
Renault
Jeep
Tata
Mini
Kia
Porsche
Skoda
Land
MG
Fiat
Force
Mitsubishi
Lamborghini
Isuzu
Ambassador
Lexus
Bentley
OpelCorsa
Aston
Maserati
Car_Brand


count
10000
8000
6000
4000
2000
0
Drum
Disc
Leading-Trailing Drum
Disc & Caliper Type
Ventilated Disc
Self-Adjusting Drum
Ventilated Drum
Disc & Drum
262mm Disc & Drum Combination
Carbon ceramic
Electric Parking Brake
Vacuum assisted hydraulic dual circuit w
Mechanical-hydraulic dual circuit
Booster assisted drum
Ventilated & Grooved Steel Discs
Dual Circuit with ABS, ABS with BAS
228.6 mm dia. drums on rear wheels
rear_brake_type







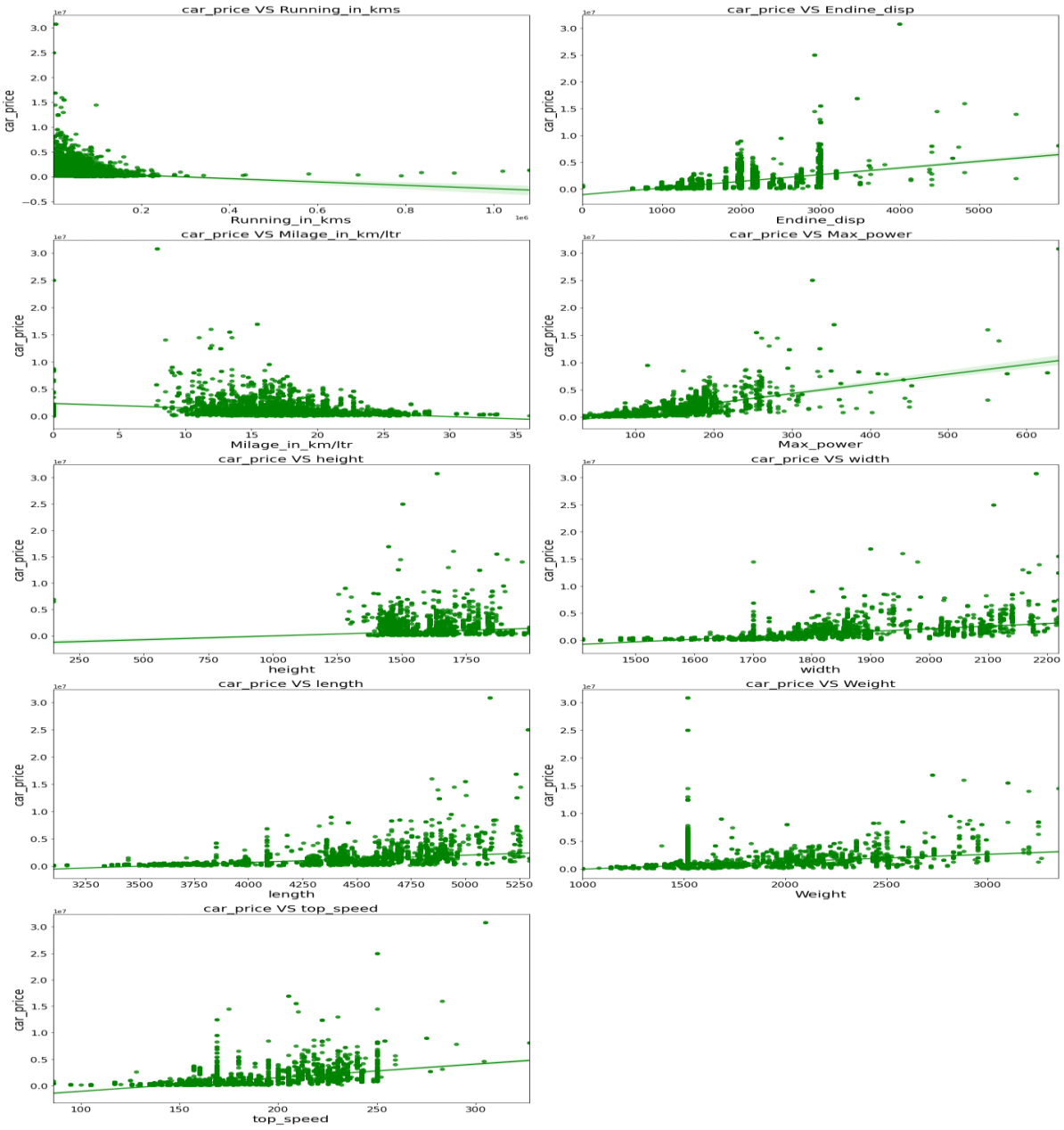
Univariate analysis for categorical column:

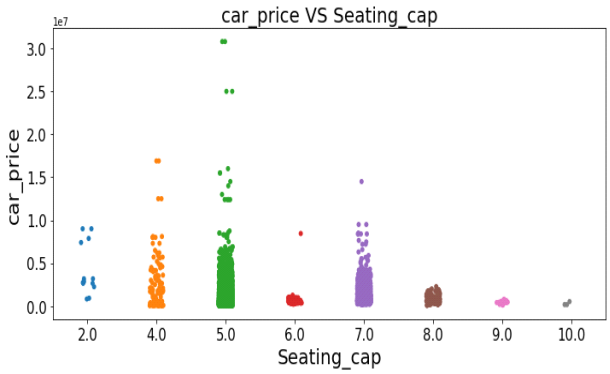
Observations:

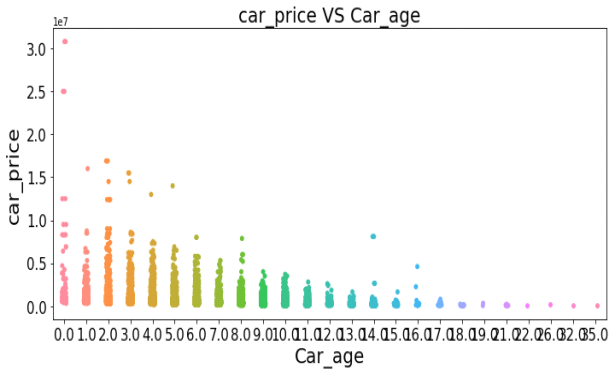
- Maximum cars are petrol driven and also diesel driven.
- Maximum cars are with Manual gear transmission.
- Disc front brake cars are more in number followed by Ventilated Disc.
- Drum rare break cars are more in number.

- 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.

- **<u>Bivariate analysis for numerical columns:</u>**



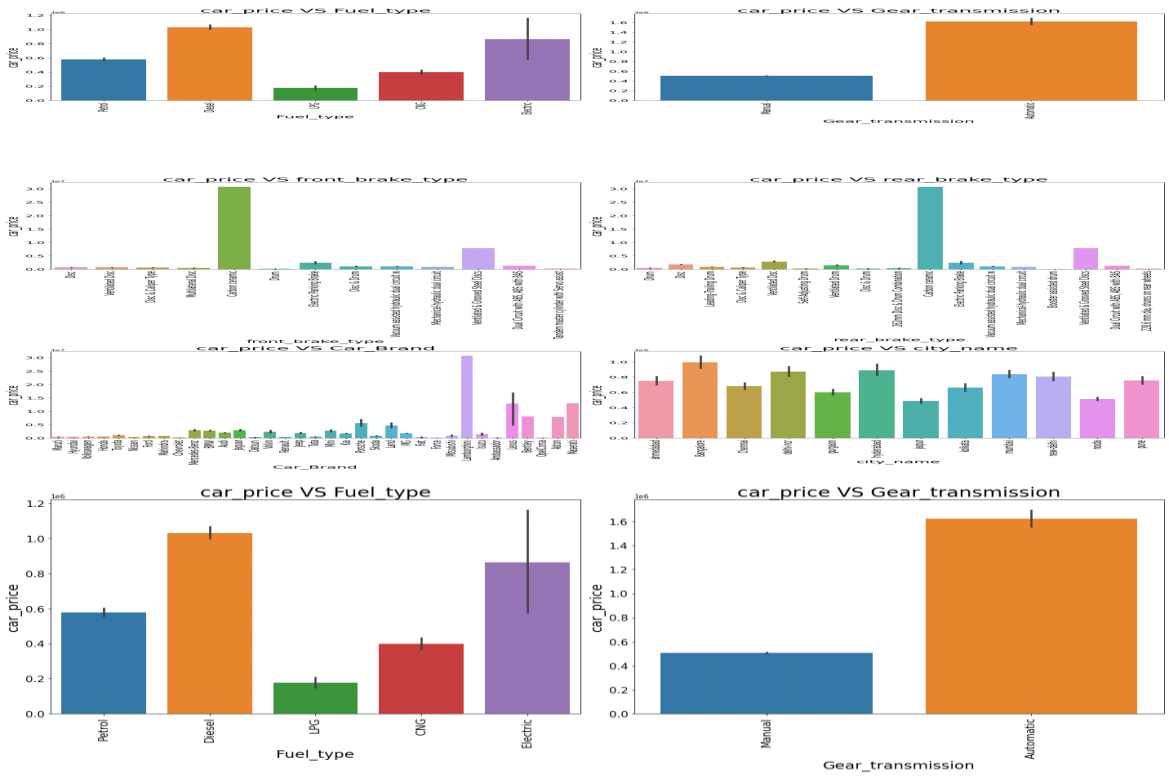


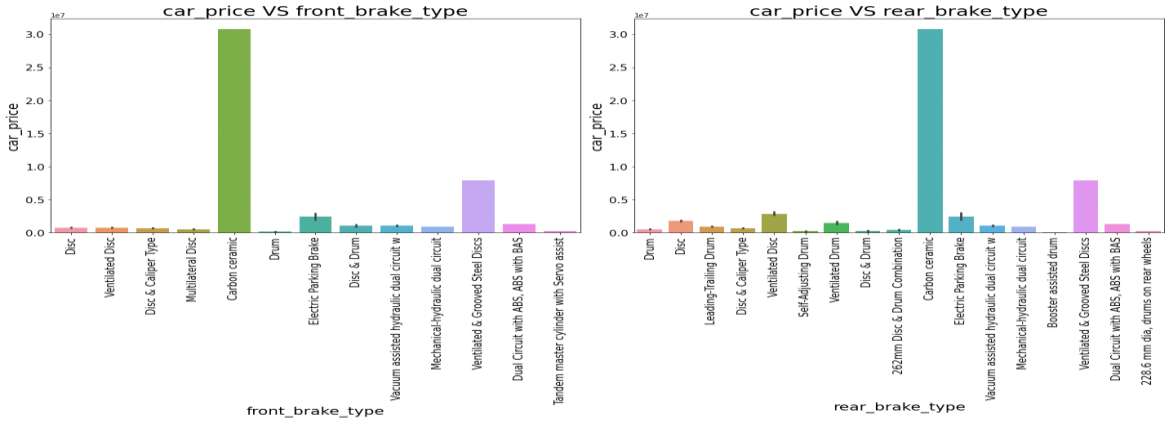


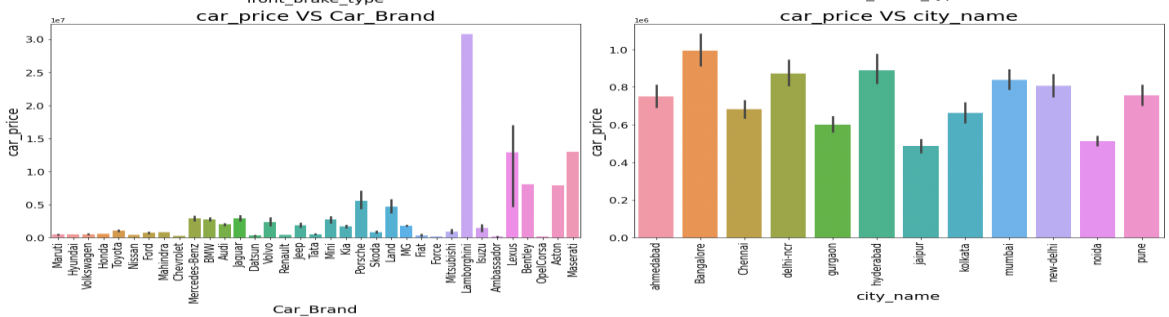
Observations:

- 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.

- **Bivariate Analysis for categorical columns:**







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:

- **RandomForestRegressor:**

```
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)

#difference of accuracy and cv score
diff = R2_score - scores
print("\nR2_Score - Cross Validation Score :", diff)
```

```
R2_score: 96.52527915830896
mean_squared_error: 8520288303.59705
mean_absolute_error: 50097.38864782178
root_mean_squared_error: 92305.40777005999

Cross validation score : 93.12997734156582

R2_Score - Cross Validation Score : 3.3953018167431424
```

RFR is giving me 96.52% r2_score.

- RandomForestRegressor has given me 96.52% r2_score and the difference between r2_score and cross validation score is 3.39%, but still we have to look into multiple models.

- **K Neighbours Regressor:**

```
KNN=KNeighborsRegressor()
KNN.fit(X_train,y_train)
pred=KNN.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(KNN, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)

#difference of accuracy and cv score
diff = R2_score - scores
print("\nR2_Score - Cross Validation Score :", diff)
```

```
R2_score: 88.12681183216546
mean_squared_error: 29113989549.61074
mean_absolute_error: 88875.74368650217
root_mean_squared_error: 170628.22026151107

Cross validation score : 84.64325187065668

R2_Score - Cross Validation Score : 3.4835599615087744
```

KNeighborsRegressor is giving me 88.12% r2_score.

- KNeighborsRegressor is giving me 88.12% r2_score and the difference between r2_score and cross validation score is 3.48%.

- **GradientBoostingRegressor:**

```
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)

#difference of accuracy and cv score
diff = R2_score - scores
print("\nR2_Score - Cross Validation Score :", diff)
```

```
R2_score: 94.02883824049444
mean_squared_error: 14641757429.250465
mean_absolute_error: 73839.51146443721
root_mean_squared_error: 121003.12983245708

Cross validation score : 90.20303019250167

R2_Score - Cross Validation Score : 3.8258080479927656
```

GradientBoostingRegressor is giving me 94.02% r2_score.

- GradientBoostingRegressor is giving me 94.02% r2_score and the difference between r2_score and cross validation score is 3.82%.

- **DecisionTreeRegressor:**

```
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)

#difference of accuracy and cv score
diff = R2_score - scores
print("\nR2_Score - Cross Validation Score :", diff)
```

```
R2_score: 92.68356026411895
mean_squared_error: 17940484644.879246
mean_absolute_error: 61529.549201741655
root_mean_squared_error: 133942.09437245352

Cross validation score : 88.05320030399506

R2_Score - Cross Validation Score : 4.630359960123897
```

DecisionTreeRegressor is giving me 92.68% r2_score.

- DecisionTreeRegressor is giving me 92.68% r2_score and the difference between r2_score and cross validation score is 4.63%.

- **BaggingRegressor:**

```
BR=BaggingRegressor()
BR.fit(X_train,y_train)
pred=BR.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(BR, X, y, cv = 10).mean()*100
print("\nCross validation score :", scores)

#difference of accuracy and cv score
diff = R2_score - scores
print("\nR2_Score - Cross Validation Score :", diff)
```

```
R2_score: 95.48437830282303
mean_squared_error: 11072658922.206182
mean_absolute_error: 55081.30680489322
root_mean_squared_error: 105226.70251512295

Cross validation score : 92.15473386175148

R2_Score - Cross Validation Score : 3.329644441071551
```

Bagging Regressor is giving me 95.48% r2_score.

- BaggingRegressor is giving me 95.48% r2_score and the difference between r2_score and cross validation score is 3.32%.

- **By looking into the difference of r2_score and cross validation score i found DecisionTreeRegressor as the best model with 92.68% r2_score and the difference between r2_score and cross validation score is 4.63%.**

3. Hyper Parameter Tunning:

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

```
parameter = {'criterion':['squared_error', 'friedman_mse', 'absolute_error', 'poisson'],
             'splitter':['best','random'],
             'max_features':['auto','sqrt','log2'],
             'min_samples_split':[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15],
             'max_depth':[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]}
```

Giving DecisionTreeRegressor parameters.

```
GCV=GridSearchCV(DecisionTreeRegressor(),parameter,cv=10)
```

Running grid search CV for ExtraTreesRegressor.

```
DecisionTreeRegressorGCV.fit(X_train,y_train)
```

```
GridSearchCV(cv=10, estimator=DecisionTreeRegressor(),
             param_grid={'criterion': ['squared_error', 'friedman_mse',
                                       'absolute_error', 'poisson'],
                         'max_depth': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,
                                       13, 14, 15],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'min_samples_split': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
                                               11, 12, 13, 14, 15],
                         'splitter': ['best', 'random']})
```

Tunning the model using GCV.

```
GCV.best_params_
```

```
{'criterion': 'friedman_mse',
 'max_depth': 13,
 'max_features': 'auto',
 'min_samples_split': 4,
 'splitter': 'random'}
```

Got the best parameters for DecisionTreeRegressor.

```
Best_mod=DecisionTreeRegressor(criterion='friedman_mse',max_depth=15,max_features='auto',min_samples_split=4,splitter='random')
Best_mod.fit(X_train,y_train)
pred=Best_mod.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: 93.04049659044364
mean_squared_error: 17065248749.717915
mean_absolute_error: 69732.05841261185
RMSE value: 130634.0260028677
```

This is my model r2_score after tuning.I got 93.04% as r2_score which is gooddd!!!.Before model accuracy was 92.68% now after tunning it is 93.04%.

- **I have choosed all parameters of DecisionTreeRegressor, after tunning the model with best parameters I have incresed my model accuracy from 92.68% to 93.04%.**

- **<u>Saving the model and Predictions:</u>**

```
# Loading the saved model
model=joblib.load("Car_Price.pkl")

#Prediction
prediction = model.predict(X_test)
prediction
```

```
array([1400000.        , 1575000.        ,  660388.88888889, ...,
        620142.85714286,  282500.        ,  291000.        ])
```

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

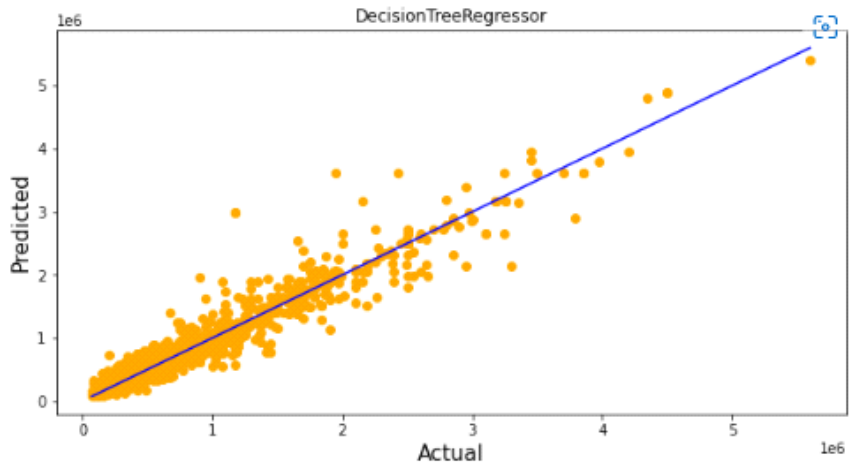
```
# Loading the saved model
model=joblib.load("Car_Price.pkl")

#Prediction
prediction = model.predict(X_test)
prediction
```

```
array([1400000.        , 1575000.        ,  660388.88888889, ...,
        620142.85714286,  282500.        ,  291000.        ])
```

```
pd.DataFrame([model.predict(X_test)[:],y_test[:]],index=["Predicted","Actual"])
```

	0	1	2	3	4	5	6	7	8	9	10
Predicted	1400000.0	1575000.0	660388.888889	634500.0	1.658333e+06	994000.0	1528000.0	587023.809524	709117.647059	643384.615385	587023.809524	62
Actual	1651000.0	1575000.0	651000.000000	715000.0	1.675000e+06	1099000.0	1500000.0	625000.000000	660000.000000	735000.000000	485000.000000	54

```
plt.figure(figsize=(10,5))
plt.scatter(y_test, prediction, c='orange')
p1 = max(max(prediction), max(y_test))
p2 = min(min(prediction), min(y_test))
plt.plot([p1, p2], [p1, p2], 'b-')
plt.xlabel('Actual', fontsize=15)
plt.ylabel('Predicted', fontsize=15)
plt.title("DecisionTreeRegressor")
plt.show()
```



Plotting Actual vs Predicted,To get better insight.Bule line is the actual line and orange dots are the predicted values.

- Plotting Actual vs Predicted, To get better insight. Bule line is the actual line and orange dots are the predicted values.

- ## 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.
  - 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 DecisionTreeRegressor as the best model with 92.68% 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 analyze 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 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 cardekho 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.

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

- First draw back is scrapping the data as it is fluctuating process.
- Followed by more 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 93.04% 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.