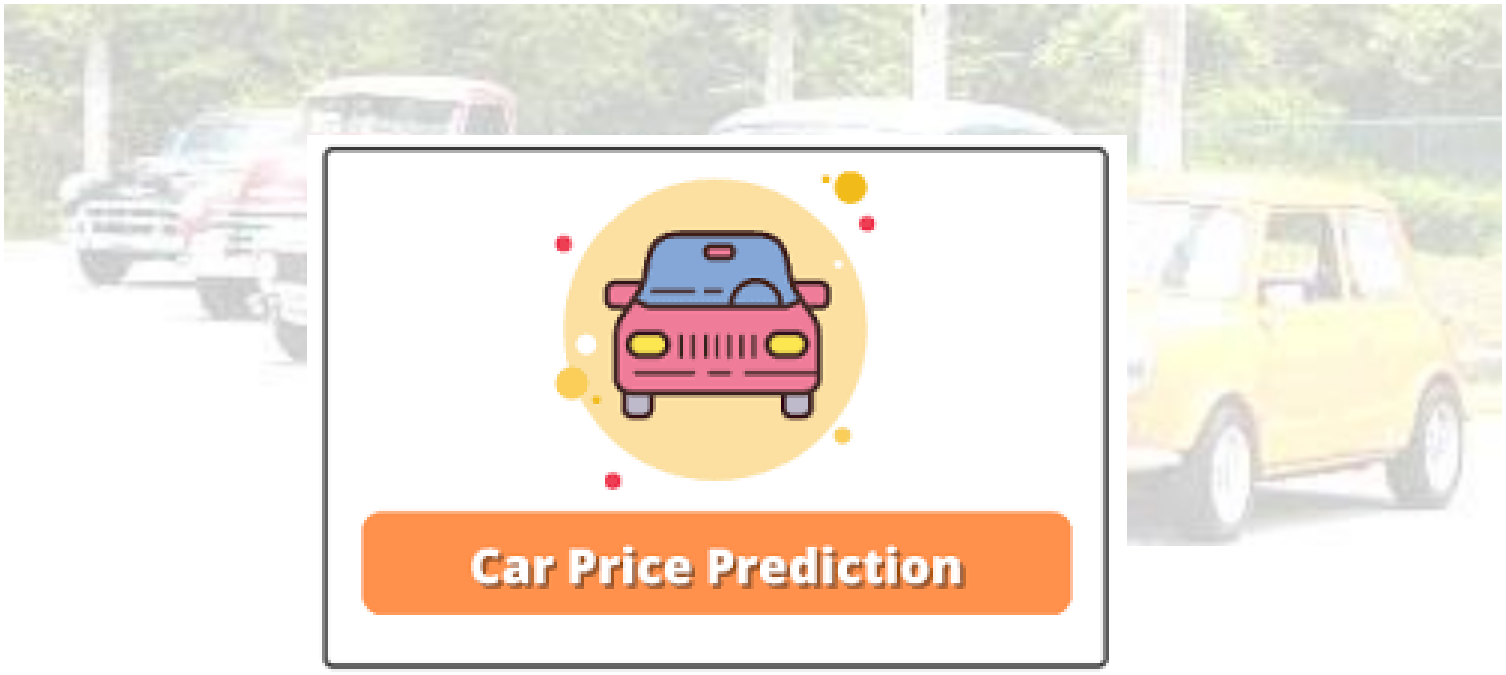
FLIP ROBO

PROJECT REPORT ON:

USED CAR PRICE PREDICTION



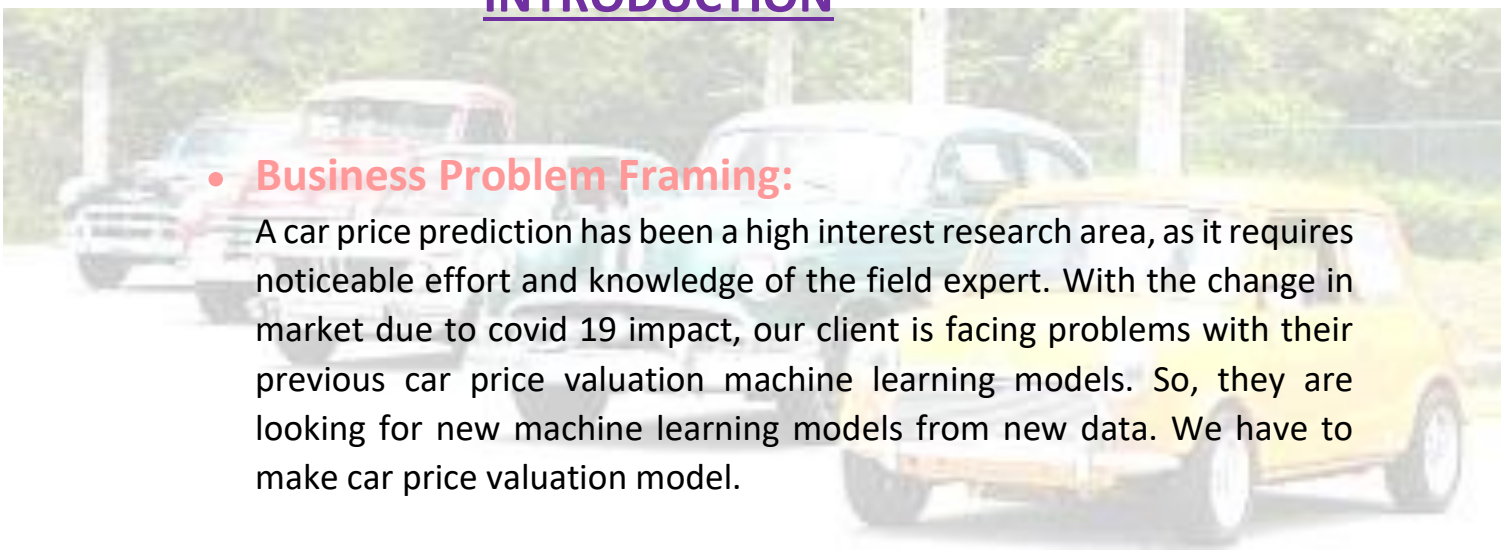
Submitted by:

VISHAL LAKHERA

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 analytical and statistical skills. And I want to express my huge gratitude to Mrs. Sapna Verma (SME Flip Robo), She is the person who has helped me to get out of all the difficulties I faced during the project and also inspired me in so many aspects . I have ended up with a project worth your while. A huge thanks to my academic team “Data trained” who helped me learn and nurtured me through these months.

INTRODUCTION



- ## Business Problem Framing:

  A car price prediction has been a high interest research area, as it requires noticeable effort and knowledge of the field expert. 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:

  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.

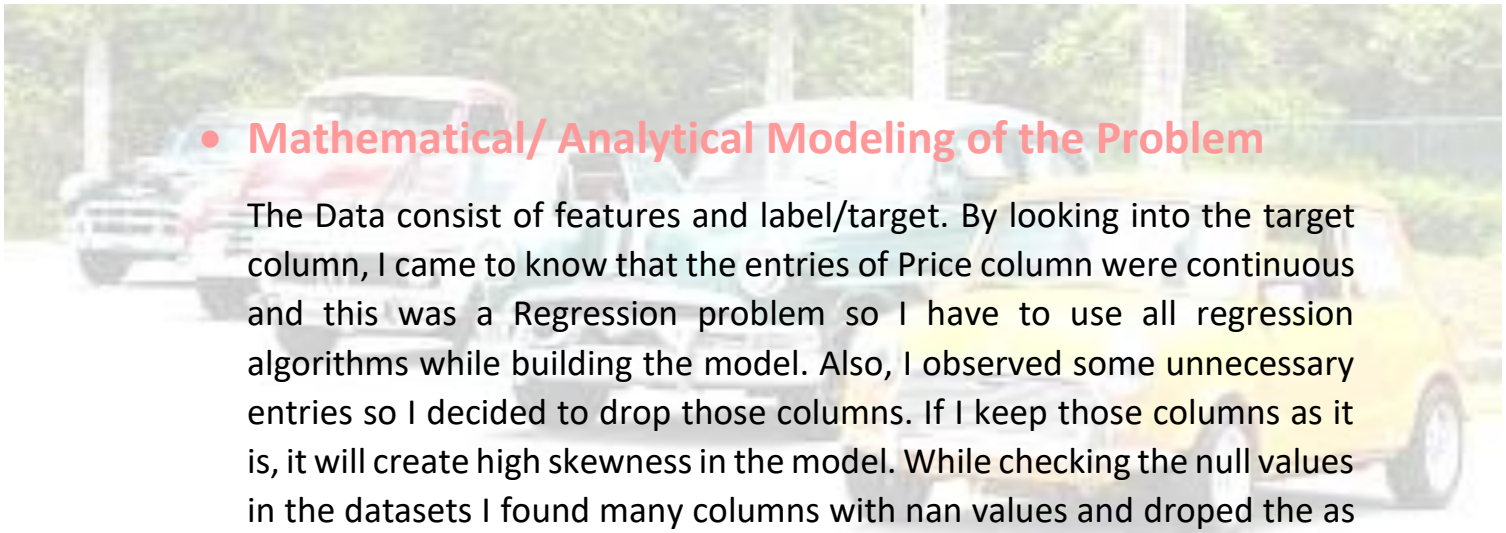
- ## Review of Literature:

  From all the research I come into conclusion that predicting the price of used car is not that easy. There are various factor which may effect the price of car , which don’t seem important to us but do effect the prices, so we need to take every feature into consideration.

- **Motivation for the Problem Undertaken:**

I have to model the price of cars with the available independent variables. This model will then be used by the Dealer to understand how exactly the prices vary with the variables.. Further, the model will be a good way for the dealer to understand the pricing dynamics of a new market. The relationship between car prices and the economy is an important motivating factor for predicting house prices.

Analytical Problem Framing



- **Mathematical/ Analytical Modeling of the Problem**

The Data consist of features and label/target. By looking into the target column, I came to know that the entries of Price column were continuous and this was a Regression problem so I have to use all regression algorithms while building the model. Also, I observed some unnecessary entries so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. While checking the null values in the datasets I found many columns with nan values and droped the as our data is very large and removing some entries won't effect our Model. To get better insight on the features I have used ploting like distribution plot, bar plot, reg plot and strip 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 models while building model. At last I have predicted the sale price for test dataset using the saved model of train dataset.

• Data Sources and their formats

The data is scraped from Cars24 sites. There are 14 columns in the dataset. The description of each of the column is given below:

1 - Location : Location in which the car was manufactured like Delhi, Mumbai, etc
2 - Name : Name of the car, like maruti swift,hyundai i10,etc
3 - Varient : varient of that particular car usually have 3 varient lower, middle and high (diff rent company use diffrent name to classify there model)
4 - PurchaseYr : Year of manufacture
5 - History : History of car wether its accidental or not.
6 - Owners : 'First', 'Second', 'Third', 'Fourth & Above'
7 - DrivenKm : Total Distance traveled by that car (continuous data)
8 - Fule : What kind of fule it uses (petrol, desiel ,CNG ,etc)
9 - Recentservicing : At what milometer reading the car has been recently underwent servicing.
10 - Transmission : 'Manual', 'Automatic'
11 - InsuaranceUpto : insuarnce valid upto which date.
12 - InsuType : What kind of insuarance (1st party, 2nd party, comprehensive, zero depreciation, etc)
13 Emi : Emi per month for that particular car(contineous)
14 PickupDt : Recent Date at which we can pick that car
15 Price -:Price of that car

Most of the column are object type and some re integer type.

• Data Preprocessing Done

I. Removing Unnecessary columns.
II. Removing Nan values.
III. Checking for duplicated and Removing them.
IV. Droping Pickup data as there is no issue of availability , car is available easily.
V. Converting year of purchase to how old the car is ,since that seen to be more helpful for our model building.

VI. Column=['Transmission','RecentServicing','InsuType] has extracted some incorrect data , so replacing them with the mode of that column.

VII. As some continuous data (Emi ,Price , RecentServicing , DrivenKm) are extracted in string format so converting it into Integer for model building.

VIII. Extracting year upto which the car has been insuared from "insuaranceUpto" as that data seem to be more useful and converting it into for how many year it has been insuared.

• Data Inputs- Logic- Output Relationships

The input data consists of int values .The input data also contains object and float values.

Different input effect the output differently ,for instance prices increases with increase in Emi , price decreases with older cars, etc. we'll have proper insight about effect of input on output further on Data Visualization.

The model approximates the function between the input and the output.

• 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:

• 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.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

a) Cleaning data: Removing unnecessary data which will create problem in our further analysis.
b) Converting continuous data from object type to integer type.
c) Encoding categorical columns using Label Encoder.
d) Removing outlier using z-score method.
e) Removing skewness of data using yeo- johnson Method.
f) Compare different models and identify the suitable model.
g) R2 score is used as the primary evaluation metric.
h) MSE and RMSE are used as secondary metrics.

- ## Testing of Identified Approaches (Algorithms)

Since Price is my target and it was a continuous 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 GradientBoostingRegressor 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.

a) RandomForestRegressor
b) XGBRegressor
c) ExtraTreesRegressor()
d) GradientBoostingRegressor
e) DecisionTreeRegressor

- Run and Evaluate selected models

1)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 = 5).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: 99.99939529178053
mean_squared_error: 300712.8588567396
mean_absolute_error: 227.37685250529287
root_mean_squared_error: 548.3729195143936

Cross validation score : 99.99847621728804

R2_Score - Cross Validation Score : 0.0009190744924865157
```

2) XGBRegressor

```
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 = 5).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: 99.99636776598855
mean_squared_error: 1806258.6855069029
mean_absolute_error: 746.6975344036697
root_mean_squared_error: 1343.9712368599646

Cross validation score : 99.99507167985598

R2_Score - Cross Validation Score : 0.001296086132569485
```

3) ExtraTreesRegressor

```
ETR=ExtraTreesRegressor()
ETR.fit(X_train,y_train)
pred=ETR.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(ETR, X, y, cv = 5).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: 99.99893587679534
mean_squared_error: 529173.4438955539
mean_absolute_error: 244.87226534932958
root_mean_squared_error: 727.4430863617813

Cross validation score : 99.99339816635819

R2_Score - Cross Validation Score : 0.0055377104371530095
```

4) GradientBoosting"

```
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 = 5).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: 99.99132260865663
mean_squared_error: 4315144.187320909
mean_absolute_error: 1520.7671973002384
root_mean_squared_error: 2077.2925136631357

Cross validation score : 99.99086535013998

R2_Score - Cross Validation Score : 0.0004572585166471299
```

5) 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 = 5).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: 99.99787335372329
mean_squared_error: 1057551.1644318984
mean_absolute_error: 317.2194777699365
root_mean_squared_error: 1028.373066757341

Cross validation score : 99.99553846522936

R2_Score - Cross Validation Score : 0.0023348884939338177
```

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

I have used the following metrics for evaluation:

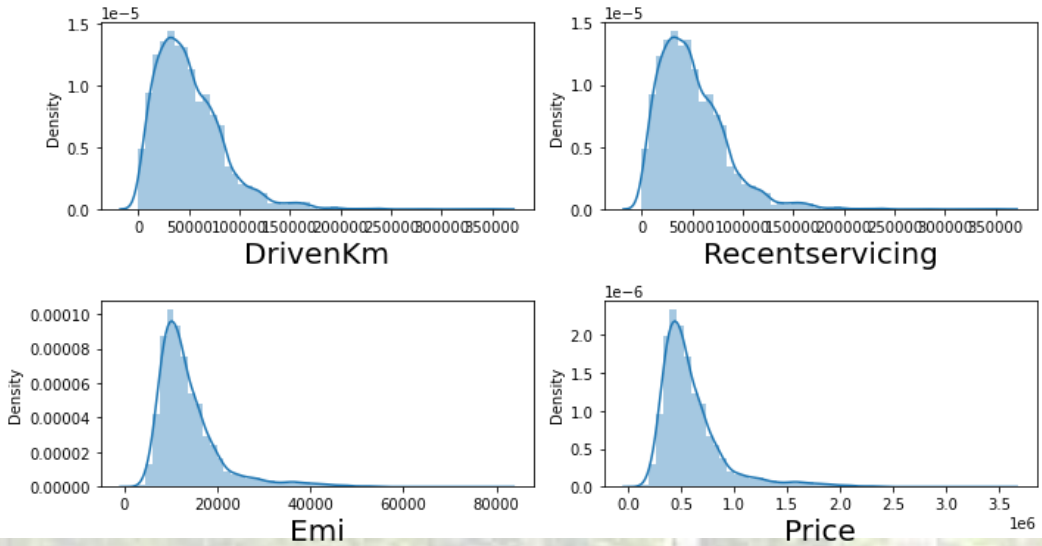
▪ I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.

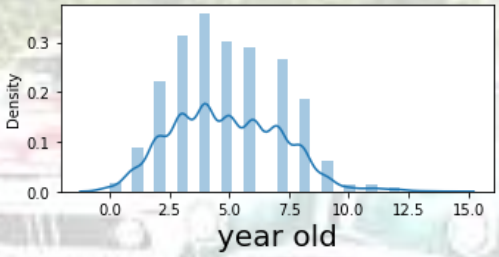
▪ I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.

▪ I have used r2 score which tells us how accurate our model is.

- **Visualizations**

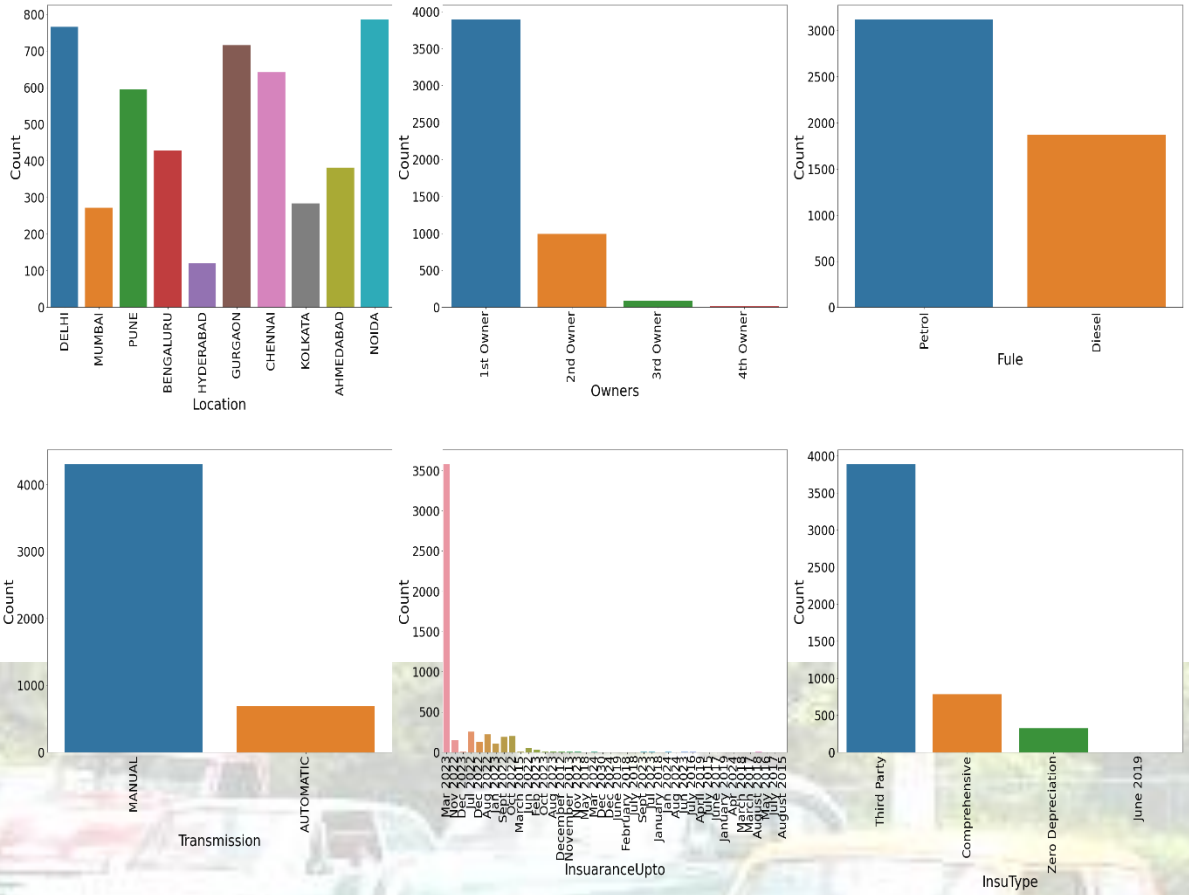
Distplot





All the columns seemes to be skewed except year old.

Count plot



Observation

1 Noida have maximum no of used car on sale.

2 Most of the car falls into category '1st owner' i.e single user before selling.

3 Car with Petrol as Fule Type are more in numbers , which indicates that maybe customer usually prefer petrol car.

4 Most of the car which are on sale are Mannualy operated one ,automatic car are usually expensive.

5 Most of the car have been insuared upto march 2023

6 Most No. of car have 3rd party insuarance.

Bar Plot

Observation

1 . Cars manufactured in mumbai have higher price then other, whereas in kolkata price is cheapest.

2 . Car who have 1st or 2nd owner have higher price as they are well maintained.

3 . Car who have 1st or 2nd owner have higher price as they are well maintained.

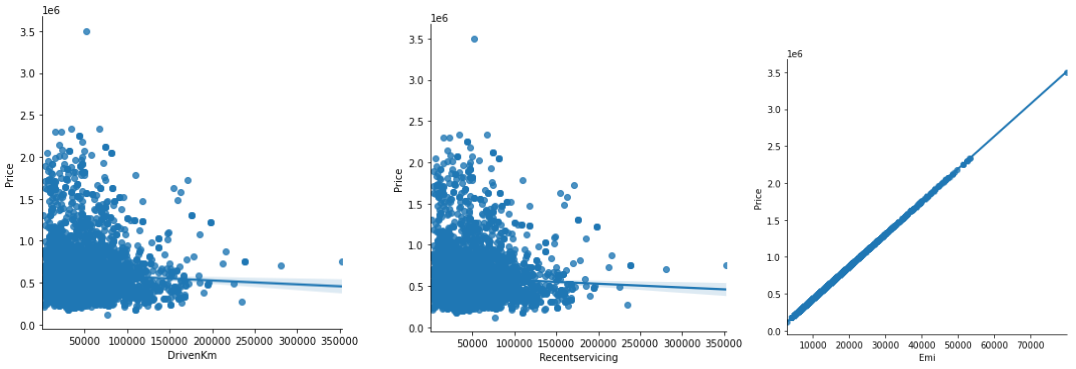
4 . Car who have 1st or 2nd owner have higher price as they are well maintained.

5 . Cars which have been insuared upto May 2018 have higher price.

6 . Cars having insuarance type as "Zero Depriciation" have higher price.

7 . New cars have higher Price.

Lmplot



Observation

1 Kilometers Driven is affecting the price as it has a slightly Negative slope, which indicate more driven car have lower price.
2 Car which have underwent servicing at lower milometer readings have higher price
3 Higher the Emi ,higher the price of a car.

- Interpretation of the Results

I. The Dataset was huge ,handling it was a great task.
II. Since I have extracted all the data using web scraping, hence some of the unnecessary/wrong data was been extracted , so that was handeled according to the situation.
III. Some of the Contineous data was in object type which was later converted to integer type for ease in model building
IV. Proper ploting for proper type of features help us to get better insight on the data
V. 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.

VI. Then scaling both dataset has a good impact like it will help the model not to get baised.

VII. We have to use multiple models while building model in our dataset so as to get the best model out of it.

VIII. At last I have predicted the Price for test dataset using saved model of train dataset. It was good!! that I was able to get the predictions near to actual values.

CONCLUSION

- ## 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 not correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a csv file was generated consisting of predicted car prices. Hence we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the dataframe of predicted prices of test dataset.

I find the best model for this problem was GradientBoostingRegressor.

- ## 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. 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 economical car searches and understanding the market of used cars .

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 missing value and to replace null value and zero values with there respective mean, median or mode accordingly to the situation. This study is an exploratory attempt to use five machine learning algorithms in estimating car prices, and then compare their results. To conclude, the application of machine learning in Car price prediction is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to this topic an alternative approach to the valuation of car prices.

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

a) Retraining of the model is important at frequent intervals so that the predicted prices stay relevant to the economic situation.
b) Market fluctuations are very random to predict.
c) There are lot more features which need to be taken into consideration for price prediction.
d) There may exist some hidden fact which are unknown to online selling sites [like olx,cars24,etc] but effects the pricing of the cars.