Car Price Prediction Project

Submitted By :

Chamlin Najir Rahman

Batch 1838

Internship 25

2022

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. And I want to express my huge gratitude to Mr. Mohd Kashif (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team “Datatrained” 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.

INTRODUCTION

- ## Business Problem Framing

  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

  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.

- ## Review of Literature

  The second-hand car market has continued to expand even as the reduction in the market of new cars. According to the recent report on India's pre-owned car market by Indian Blue Book, nearly 4 million used cars were purchased and sold in 2018-19. The second-hand car market has created the business for both buyers and sellers. Most of the people prefer

to buy the used cars because of the affordable price and they can resell that again after some years of usage which may get some profit. The price of used cars depends on many factors like fuel type, 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.

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 particular 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 were null values in the dataset. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 50% null values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. Since we have scrapped the data from cardekho 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 plotting like distribution plot, bar plot, line plot, strip plot and count plot. With these plotting I was able to

understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using z-score method and I removed skewness using yeo-johnson method. I have used all the regression algorithms while building model then tuned the best model and saved the best model. At last I have predicted the car-price using saved model.

- # 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. I n this particular datasets I have object type of data which has been changed as per our analysis about the dataset. The information about features is as follows :

- **Car_Name** : Name of the car with year
- **Fuel_type** : Type of fuel used in car engine
- **Running_in_kms** : Total distance covered 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** : Number of seats in the car
- **color** : Color of the car
- **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 dropped Unnamed:0, cargo_volume and Insp_scorecolumn as I found they are useless.
- Next as a part of feature extraction I converted the data types of all the columns and I have extracted useful information from the raw dataset. Thinking that this data will help us more than raw data.

- # 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 line 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
2. Jupyter

Model/s Development and Evaluation

- ## 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 ExtraTreeRegressor 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
- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet
- Support Vector Regression (poly)
- Support Vector Regression (linear)
- Support Vector Regression (rbf)
- K Neighbors Regressor

- ExtraTreesRegressor
- GradientBoostingRegressor
- DecisionTreeRegressor
- Ada Boost Regressor

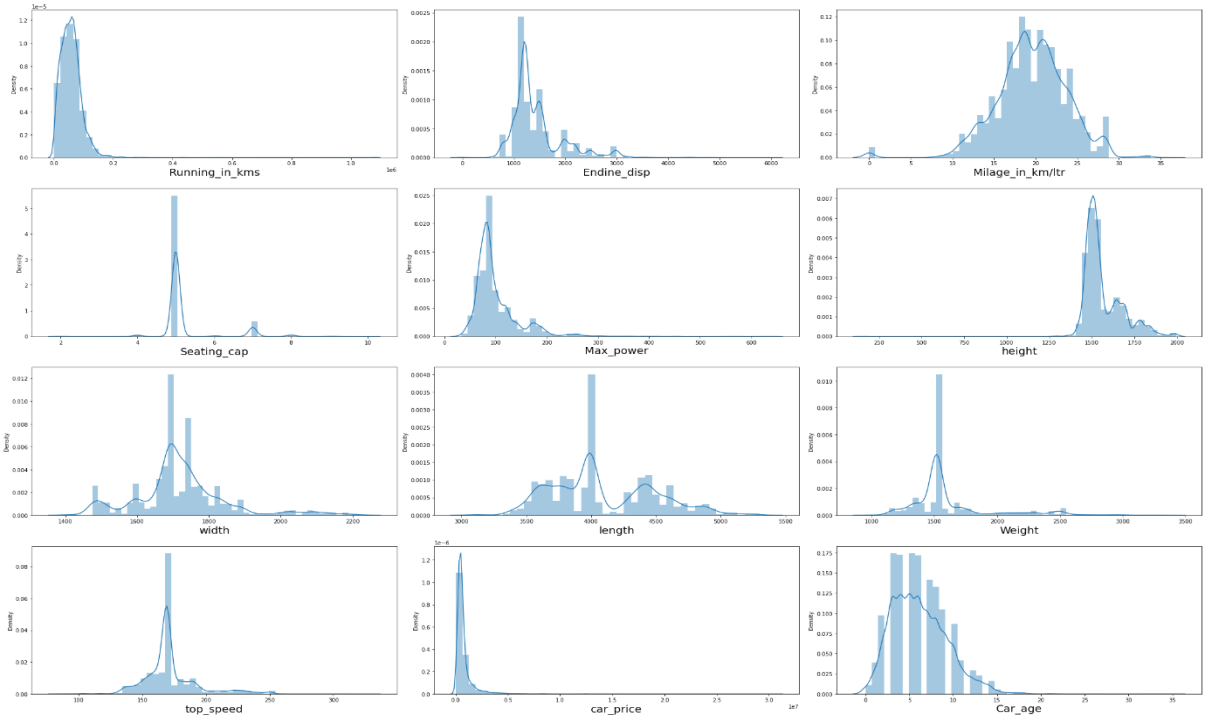
• Key Metrics for success in solving problem under consideration

I have used the following metrics for evaluation:

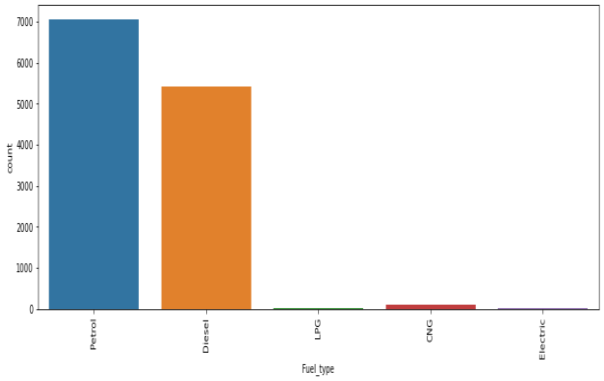
- I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.
- I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.
- I have used r2 score which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

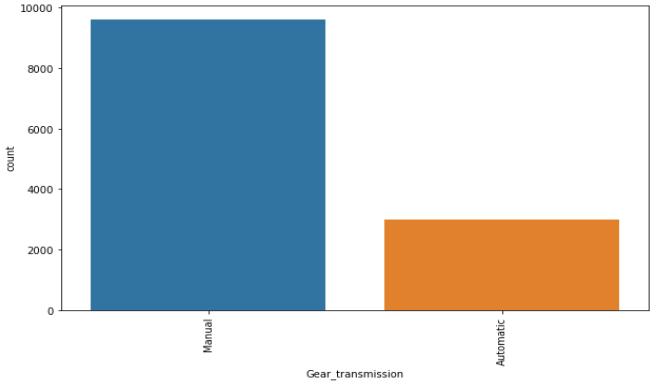
• Visualizations

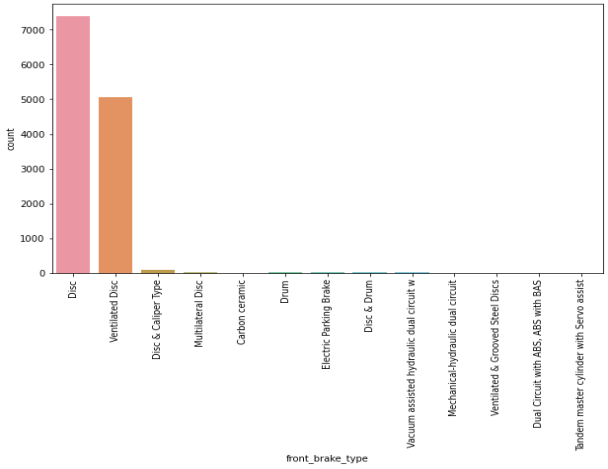
I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and line plot, strip plot for bivariate analysis.

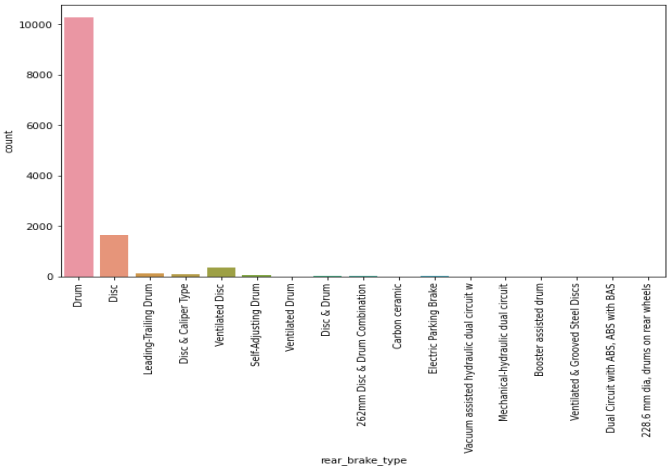


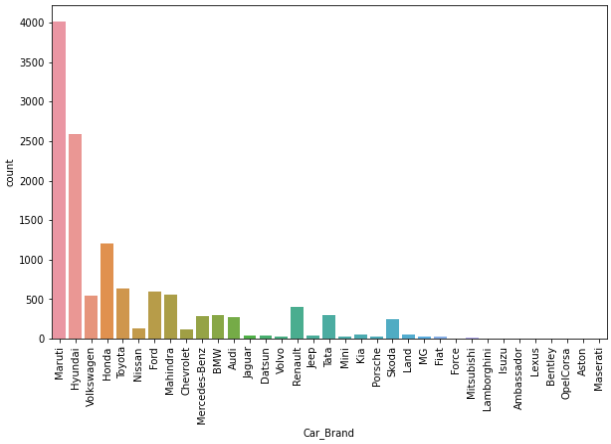
Univariate analysis for categorical column :

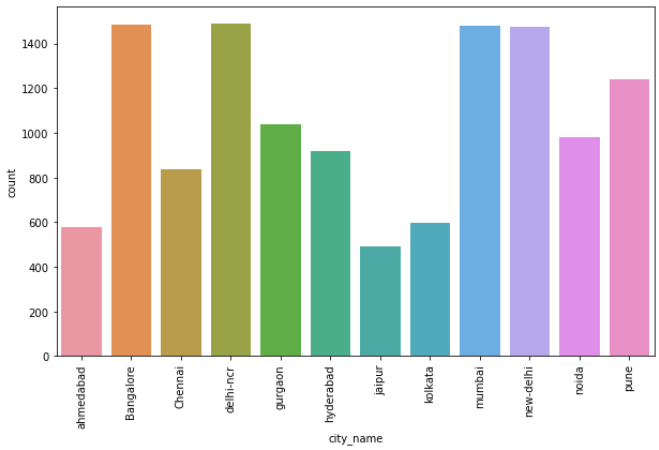




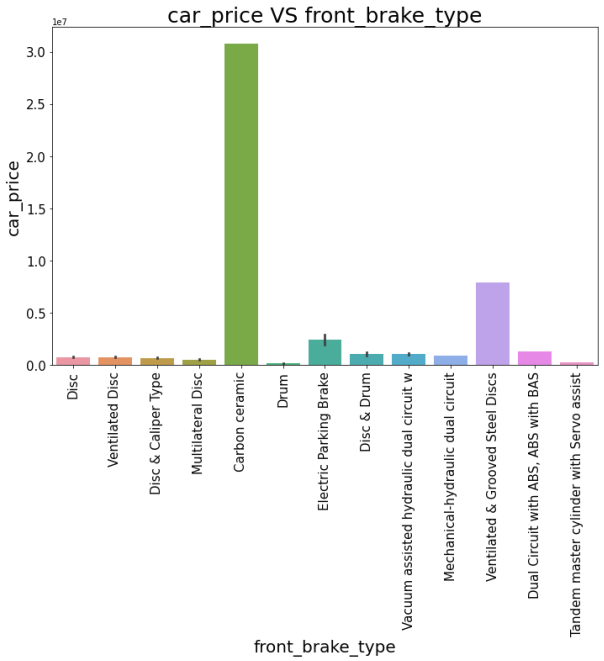


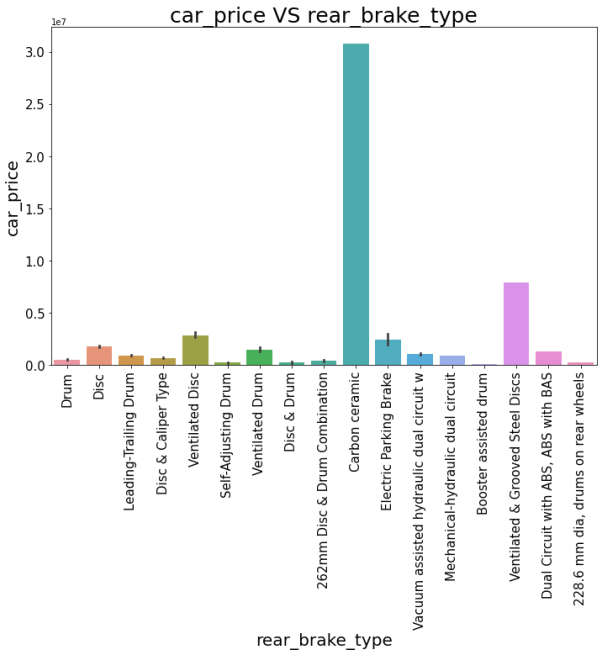


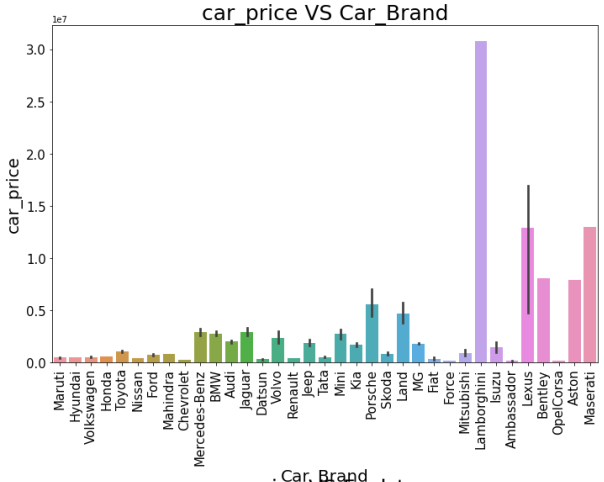


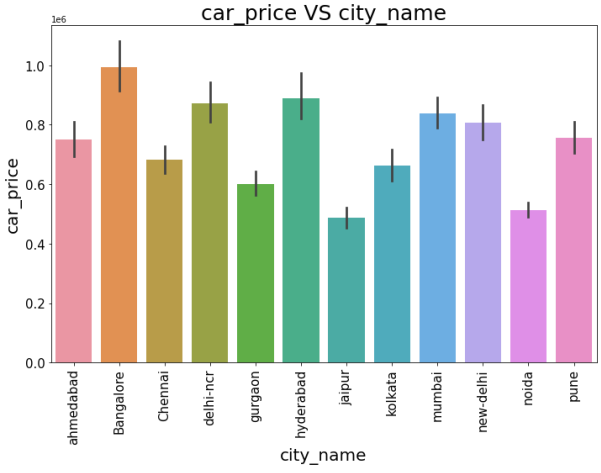


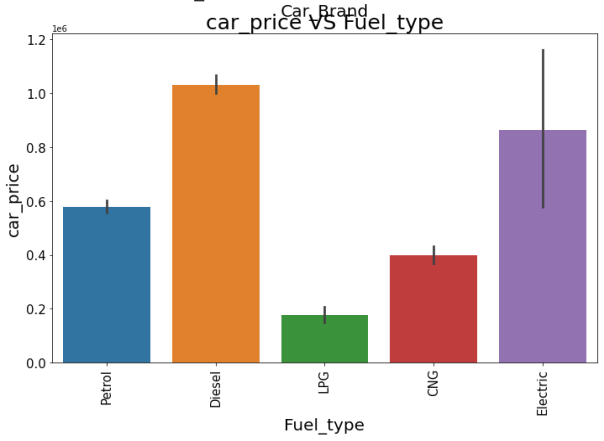
Bivariate Analysis for categorical columns:

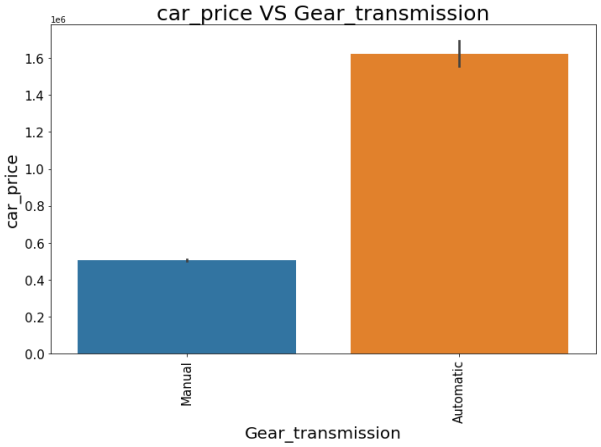












- Interpretation of the Results

```
# Extra Trees Regressor

model=ExtraTreesRegressor(n_estimators=300)
reg(model,x,y)
```

```
Mean Square Error Score is: 80223.9491520162
r2 Score is: 96.61260163301473
Cross Validation Score: 94.1813671074118
Difference between r2 Score and Cross Validation Score is 2.4312345256029317
```

After finding all the scores of various models, we found that ExtraTreesRegressor model gives the highest r2 score. Hence we choose this model and proceed further with the process.

This is the best model I achieved

Hyper Parameter Tunning :

```
grid=GridSearchCV(ExtraTreesRegressor(), params, cv=5)
```

```
grid.fit(x_train,y_train)
```

```
GridSearchCV(cv=5, estimator=ExtraTreesRegressor(),
             param_grid={'criterion': ['squared_error', 'absolute_error'],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'n_estimators': [300]})
```

```
grid.best_params_
```

```
{'criterion': 'absolute_error', 'max_features': 'auto', 'n_estimators': 300}
```

So here we have found the best parameters for our model, and now we can finally train our model.

```
ETR=ExtraTreesRegressor(criterion='absolute_error', max_features='auto', n_estimators=300)
```

```
ETR.fit(x_train,y_train)
pred=ETR.predict(x_test)
r2=r2_score(y_test, pred)*100
print('r2 score for the final model : ',r2)
```

```
r2 score for the final model :  96.48965467469864
```

Saving the best model

Saving the best model :

```
import pickle
filename='Car_Price.pkl'
pickle.dump(ETR,open(filename,'wb'))
```

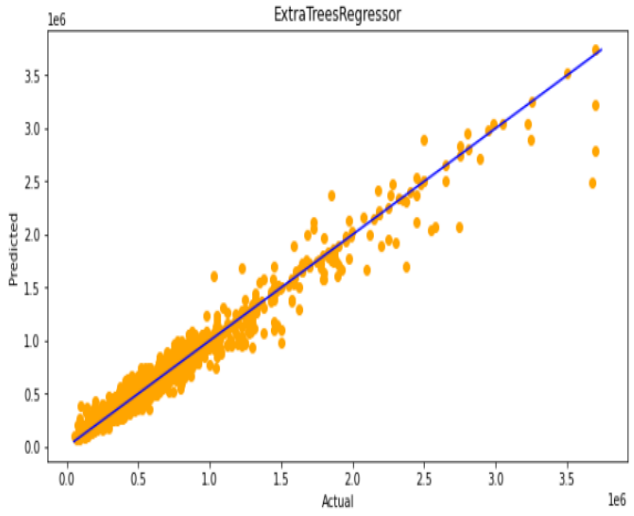
Predictions

```
prediction = ETR.predict(x_test)
prediction
```

```
array([ 605000.        ,  387800.        ,  601763.33333333, ...,
        627750.        ,  443560.        , 1717170.        ])
```

```
pd.DataFrame([prediction[:],y_test[:]],index=["Predicted","Original"]).T
```

	Predicted	Original
0	6.050000e+05	605000.0
1	3.878000e+05	388000.0
2	6.017633e+05	637000.0
3	4.990000e+05	499000.0
4	9.028800e+05	918000.0
...	...	...
2260	4.395567e+05	570000.0
2261	2.874000e+05	295000.0
2262	6.277500e+05	705000.0
2263	4.435600e+05	410000.0



Blue line denotes the actual values and the orange dots are the predicted values.

CONCLUSION

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 tuning was done to the best model and the accuracy has been improved. Hence we calculated the performance of each model using different performance

metrics and compared them based on those metrics. Then we have also saved the best modeland predicted the used car price. It was good the the predicted and actual values were almost same.