FLIP ROBO

Car Price Prediction Project

Submitted by:

Chaganti Sai Mahathi

ACKNOWLEDGMENT

References:

[1] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”;(IJICT 2014)

[2] Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, ”Car Price Prediction Using Machine Learning”; (TEM Journal 2019)

[3] Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, “Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory”; (Hohai University Changzhou, China)

[4] Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, “Prediction of Prices for Used Car by using Regression Models” (ICBIR 2018)

[5] Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, “Prediction car prices using qualify qualitative data and knowledge-based system” (Hanoi National University)

[6] https://towardsdatascience.com/used-car-price-prediction-using-machine-learning-e3be02d977b2

[7] https://medium.com/geekculture/used-carprice-prediction-complete-machine-learning-project-d25559cf2d2a

INTRODUCTION

From a long time since being, a continuous paradigm of transactions of commodities has been into existence. Earlier these transactions were in the form of barter system which later was translated into a monetary system. And with consideration into these, all changes that were brought about the pattern of re-selling items was affected as well. There are two ways in which the re-selling of the item is carried out. One is offline and the other being online. In offline transactions, there is a mediator present in between who is very vulnerable to being corrupt and make overly profitable transactions. The second option is online wherein there is a certain platform which lets the user find the price he might get if he goes for selling

• Kilometres travelled – We know that the number of kilometres travelled by a vehicle has a huge role to play while putting the vehicle up for sale. The more the vehicle has travelled, the older it is.

• Fiscal power – It is the power output of the vehicle. More output yields better value out of a vehicle.

• Year of registration – It is the year when the vehicle was registered with the Road Transport Authority. The newer the vehicle is; the better value it will yield. By every passing year, the value will depreciate.

• Fuel Type – There were two types of fuel types present in the dataset that we had. Petrol and Diesel. It was relatively less dominant. It's due to the above factors that we need a system that can develop a self-learning machine learning-based system. This was the basis on which a set of objectives was supposed to be formulated. One thing that was pre-determined was that this is going to be a real-time project.

OBJECTIVE

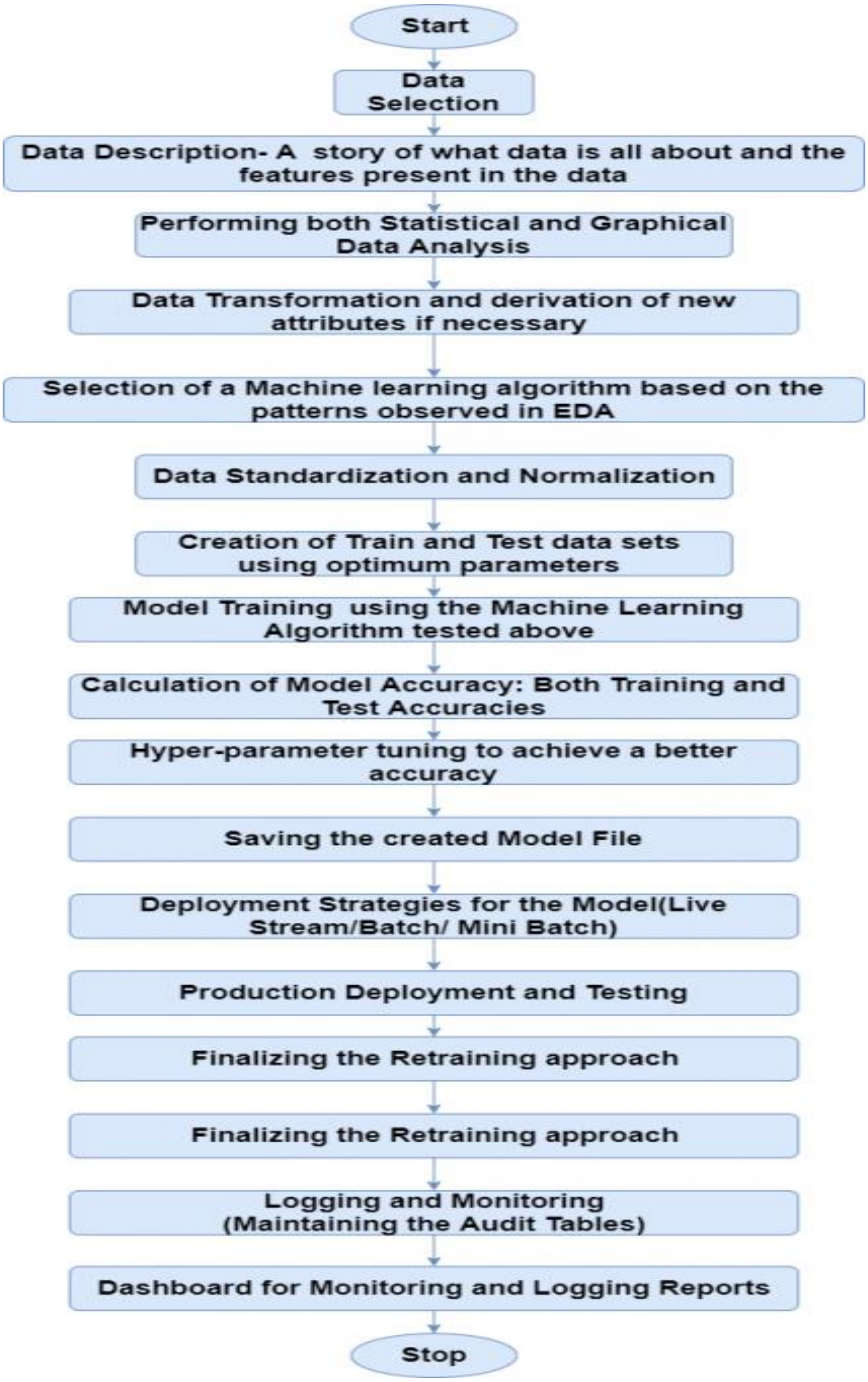
• To build a supervised machine learning model for forecasting value of a vehicle based on multiple attributes

• The system that is being built must be feature based i.e. feature wise prediction must be possible.

METHODOLGY:

There are two primary phases in the system: 1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly. 2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

For this project, we have implemented the machine learning life cycle to create a basic web application which will predict the flight prices by applying machine learning algorithm to historical flight data using python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn. Figure.1 shows the steps that we followed from the life cycle:


Start
Data Selection
Data Description- A story of what data is all about and the features present in the data
Performing both Statistical and Graphical Data Analysis
Data Transformation and derivation of new attributes if necessary
Selection of a Machine learning algorithm based on the patterns observed in EDA
Data Standardization and Normalization
Creation of Train and Test data sets using optimum parameters
Model Training using the Machine Learning Algorithm tested above
Calculation of Model Accuracy: Both Training and Test Accuracies
Hyper-parameter tuning to achieve a better accuracy
Saving the created Model File
Deployment Strategies for the Model(Live Stream/Batch/ Mini Batch)
Production Deployment and Testing
Finalizing the Retraining approach
Finalizing the Retraining approach
Logging and Monitoring (Maintaining the Audit Tables)
Dashboard for Monitoring and Logging Reports
Stop

Data selection is the first step where historical data of flight is gathered for the model to predict prices. Our dataset consists of more than 10,000 records of data related. Some of the features of the dataset are source, destination, departure date, departure time, number of stops, arrival time, prices and few more. In the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If these values are not removed it would affect the accuracy of the model. We gained further information such as distribution of data. Next step is data pre-processing where we observed that most of the data was present in string format.. For this One hotencoding and label encoding techniques are used to convert categorical values to model identifiable values.

Feature selection step is involved in selecting important features that are more correlated to the price. There are some features such as extra information and route which are unnecessary features which may affect the accuracy of the model and therefore, they need to be removed before getting our model ready for prediction. After selecting the features which are more correlated to price the next step involves applying machine algorithm and creating a model. As our dataset consist of labelled data, we will be using supervised machine learning algorithms also in supervised we will be using regression algorithms as our dataset contains continuous values in the features. Regression models are used to describe relationship between dependent and independent variables. The machine learning algorithms that we will be using in our project are

Linear Regression: Linear regression is the most common form of regression. Multi linear regression is used to explain the relationship between one continuous dependant variable and two or more explanatory variables. Linear regression has several assumptions such as; Regression residuals must be normally distributed, residuals are homoscedastic and approximately rectangular-shaped, Absence of multicollinearity is assumed in the model.

K Nearest Neighbour Regression: K-Neighbhor's Regressor is a non-parametric lazy learning algorithm. KNN uses a similarity measure such as the Euclidian distance to find a heuristically optimal number of K nearest neighbours based on Root mean square error to label new observations. The output is the property value for the object. This value is the average of the values of its k nearest neighbours. KNN is sensitive to high dimensional data and outliers.

Random Forest Regression: Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression.

Decision Tree Regression: Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]

2. Result [End Nodes]

Here we use different techniques such as Data Analysis, Data mining and Data Visualisation, Data Pre-processing and model building is done for the prediction.

Steps For Data Collection, Processing, EDA and Modelling the data:

Data collection and Analysis:

Data is collected by using various web scraping techniques using selenium from different websites and creating the data frame for all the data collected. The collected data has many features which gives different information about theused car price.

Feautures:

brand:Brand name of each car.

Model:Designed model of each car.

variant:variant number given for each car.

manufacturing year:year of manufacture.

EMI:Monthly installments for car buying.

Driven_kilometers:how many kilometers it can drive.

Fuel:What kind of fuel like petrol or diesel used.

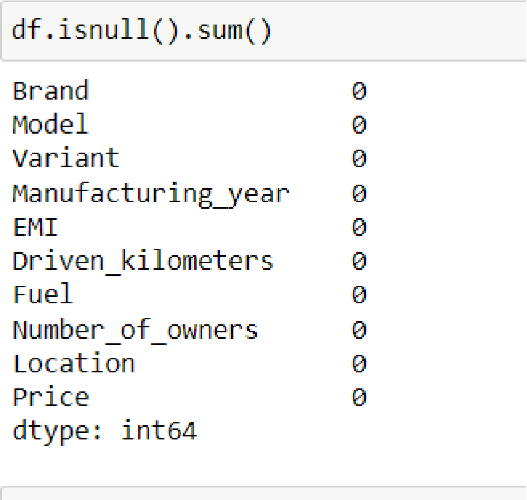
Number of Owners:no of owners existed before selling a car.

Location :location where the car present.

Price: Final Price of the car.

A) Removing null values:

1. After looking in to the data we got the ticket price label with some null values which are removed by replace the comma with space and changing them to float and using np.nan method.

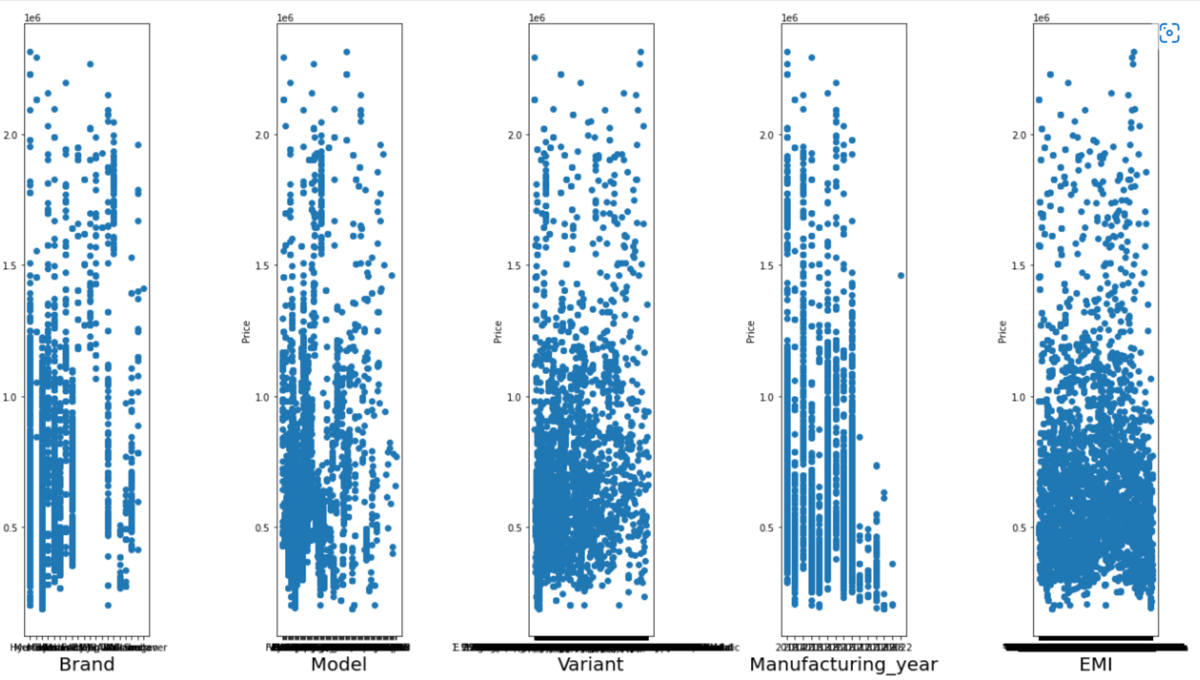


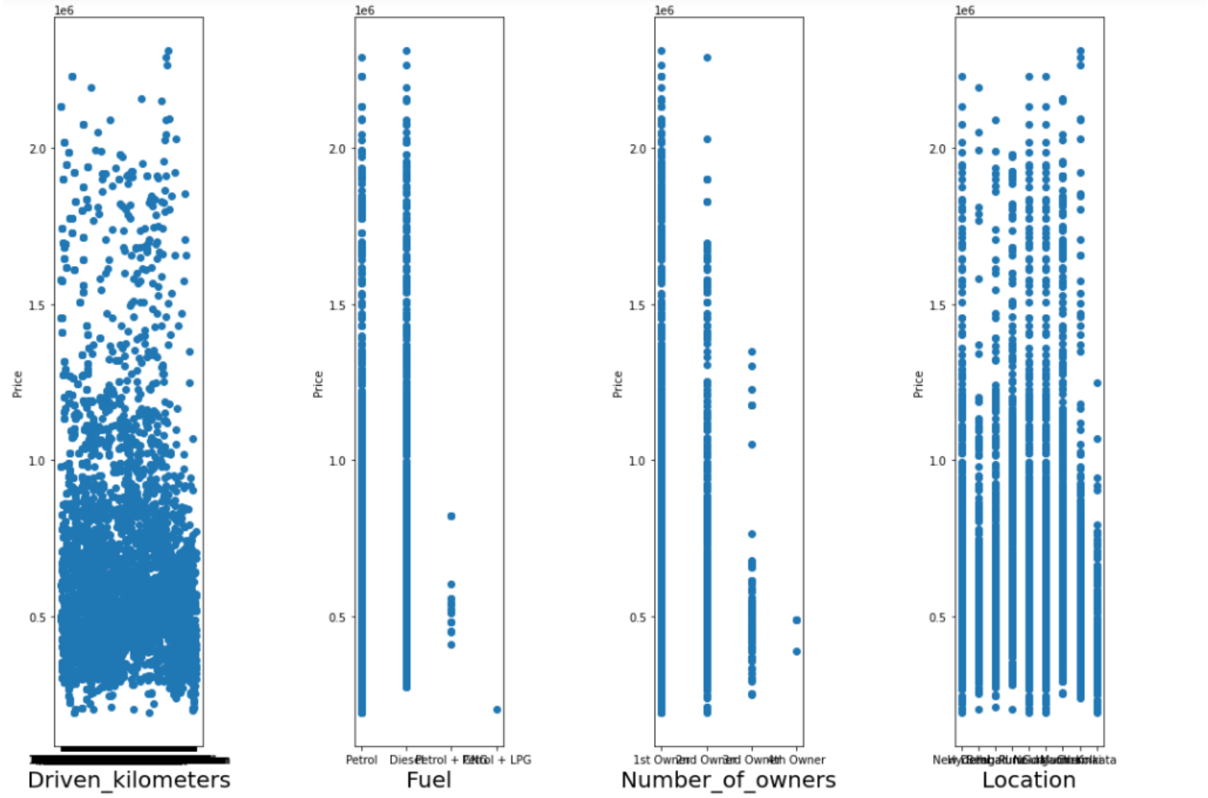
```
df.isnull().sum()
```

```
Brand                  0
Model                  0
Variant                0
Manufacturing_year     0
EMI                    0
Driven_kilometers      0
Fuel                   0
Number_of_owners       0
Location               0
Price                  0
dtype: int64
```

B) Data Visualisation and analysis:

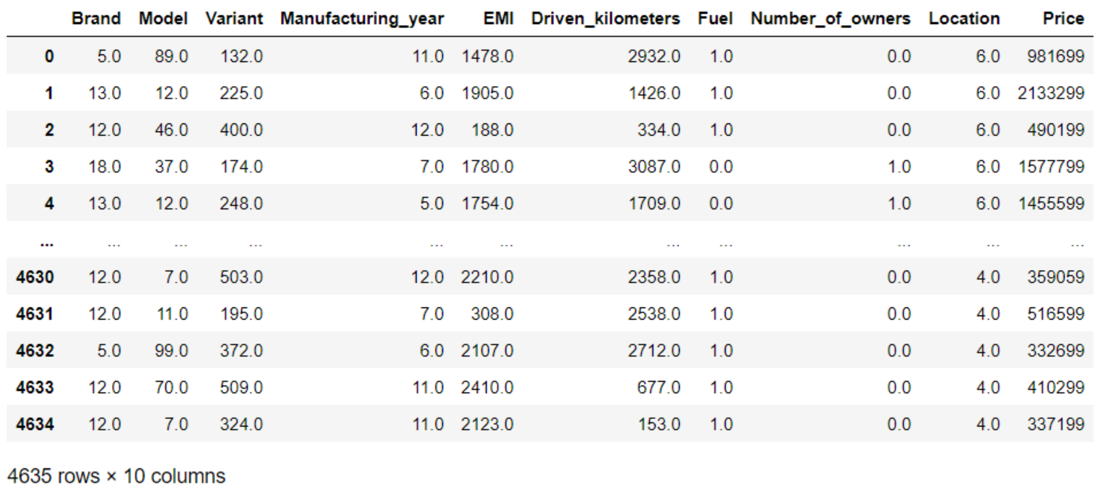
Here we are using scatter plot to visualise all the input features with respect to the Price of the used car.





C) Encoding the data:

TO convert all object type data to numerical of both training and testing data we use Ordinal encoder.And also removing all the unwanted columns.



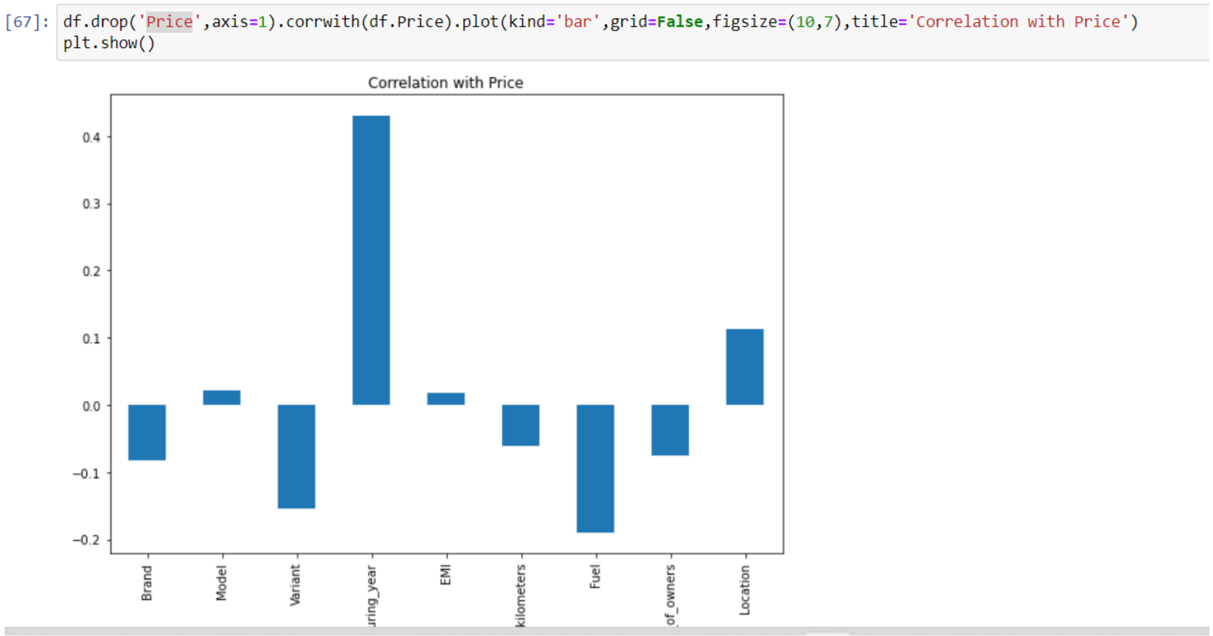
	Brand	Model	Variant	Manufacturing_year	EMI	Driven_kilometers	Fuel	Number_of_owners	Location	Price
0	5.0	89.0	132.0	11.0	1478.0	2932.0	1.0	0.0	6.0	981699
1	13.0	12.0	225.0	6.0	1905.0	1426.0	1.0	0.0	6.0	2133299
2	12.0	46.0	400.0	12.0	188.0	334.0	1.0	0.0	6.0	490199
3	18.0	37.0	174.0	7.0	1780.0	3087.0	0.0	1.0	6.0	1577799
4	13.0	12.0	248.0	5.0	1754.0	1709.0	0.0	1.0	6.0	1455599
...	...	...	...	...	...	...	...	...	...	...
4630	12.0	7.0	503.0	12.0	2210.0	2358.0	1.0	0.0	4.0	359059
4631	12.0	11.0	195.0	7.0	308.0	2538.0	1.0	0.0	4.0	516599
4632	5.0	99.0	372.0	6.0	2107.0	2712.0	1.0	0.0	4.0	332699
4633	12.0	70.0	509.0	11.0	2410.0	677.0	1.0	0.0	4.0	410299
4634	12.0	7.0	324.0	11.0	2123.0	153.0	1.0	0.0	4.0	337199

4635 rows × 10 columns

D)Data Pre-Processing:

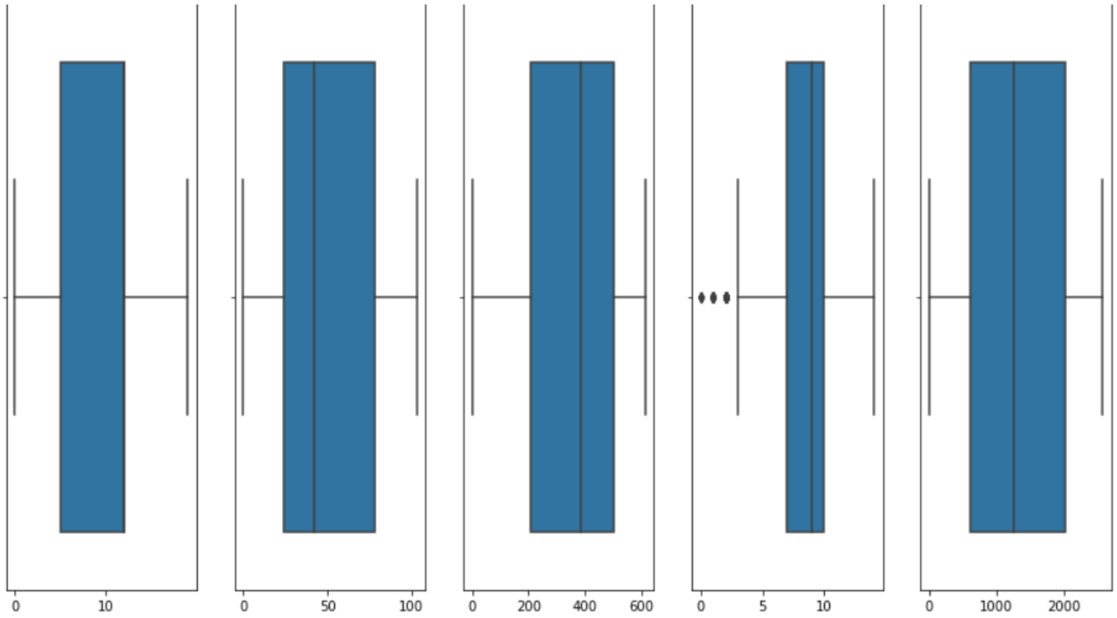
The first step is to Drop all the unwanted column present in the dataset.

Now, since the data sets are ready to process, check the correlation of the training dataset:

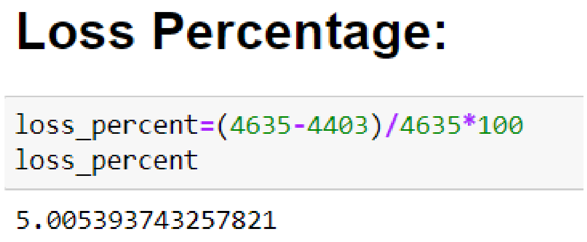


```
[67]: df.drop('Price',axis=1).corrwith(df.Price).plot(kind='bar',grid=False,figsize=(10,7),title='Correlation with Price')
      plt.show()
```

Now lets also check the skewness and outliers if exists lets try to remove using some commonly used techniques.



Removing outliers and skewness is a very difficult process in this case we can use z-score technique with a data loss of 5 percent.



Loss Percentage:

```
loss_percent=(4635-4403)/4635*100
loss_percent
```

```
5.005393743257821
```

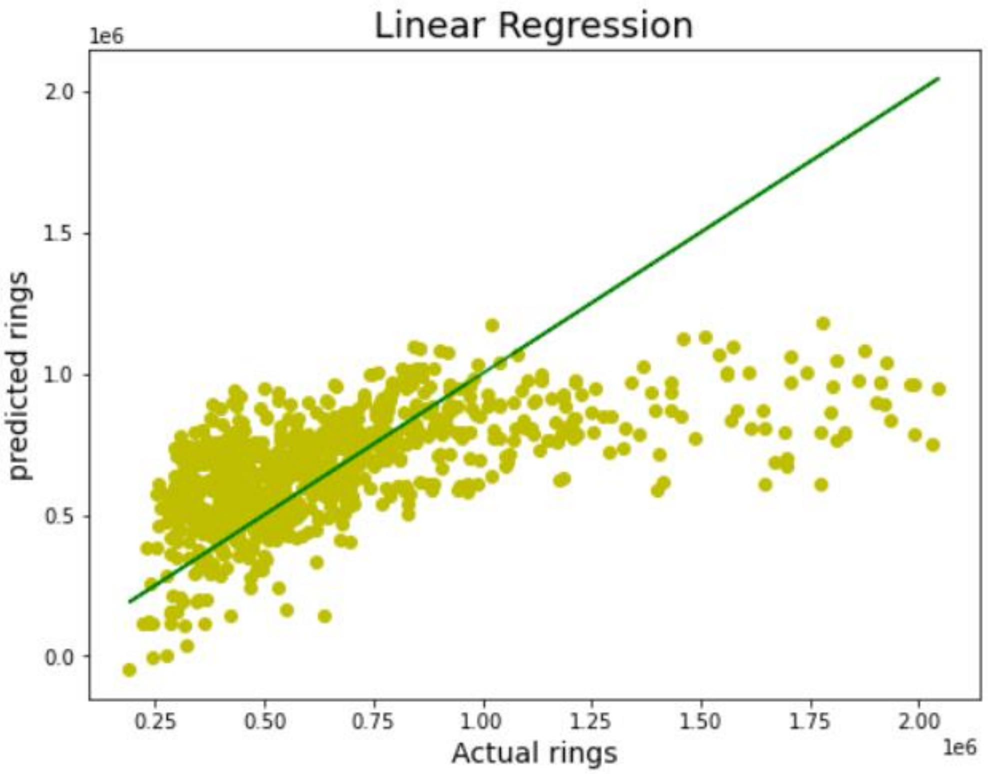
Now once the best features are declared we cannot say that we avoided all the outliers or skewness completely but now by considering these features let’s create a new training and testing data frames.

Model Building:

Now, once the data is pre processed and ready for the building, we first Train the data using various models and then use this trained data for testing or predicting the testing model.

First split the feature variables as ‘x’ data frame and ‘y’ as target. Then, Transform the ‘x’ variable using power transformation to remove any skewness present in the data and then normalise the data using Standard Scaler.
Next create a train test split of random state given as 41and test size as 0.20 and then apply linear regression model to test the data the r2 score of the predicted data is given as **0.3395308**

Linear Regression Curve:

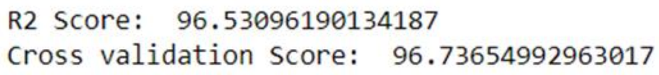


Once the model is linearly regressed then use lasso regularisation technique to get the maximum best output-**0.339523**
Using the best cross validation score we get the output for this model as

0.302473

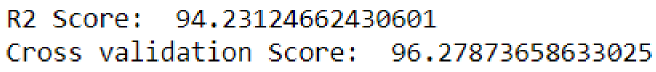
Now we use other Ensemble approaches and further using hyper parameter tuning finding the best parameters gets the better results of

Random Forest Regression:



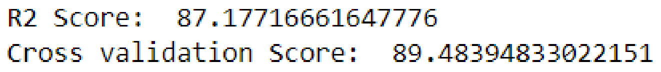
```
R2 Score:  96.53096190134187
Cross validation Score:  96.73654992963017
```

Decision Tree Regression:



```
R2 Score:  94.23124662430601
Cross validation Score:  96.27873658633025
```

K-Neighbors Regression:



```
R2 Score:  87.17716661647776
Cross validation Score:  89.48394833022151
```

Out of all models Random forest regression is having R2 Score: 96.5309and Cross validation Score: 96.73654. So, this model is finalised and further saved.

Saving the model and Loading the data is the further most step, now is that my model is trained it goes to testing phase

FUTURE SCOPE

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

CONCLUSION

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 3 different algorithms for machine learning : Linear Regression, Lasso Regression and Ridge Regression