car_price_prediction

December 7, 2020

1 Predicting Carprice

1.1 Introduction

This project aims to predict car prices for a Chinese automobile company. So the main aim would be helping the Chinese automobile company which plans to enter the US automobile company. The company wants to understand the factors affecting the price of cars in the American market, since those may be very different from the Chinese market. So this project will try to answer inquiries like, Which variables are significant in predicting the price of a car and How well those variables describe the price of a car.

The data has been collected from Kaggle and contains data from different market surveys which gathers a large data set of different types of cars across the American market. The business goal can be explained as a model for the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy, etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

This report starts with a detailed understanding of datasets, followed by looking at the data cleaning and feature engineering process done to the data. Then various regression models with regularization will be applied to the data for accurate prediction of the price. After modeling, a careful analysis of the outcome of the various model will lead to a selection of the appropriate model for the prediction.

1.2 Data Cleaning and Feature Engineering

The data named 'carprice.csv' contains different factors about a car in the market and price of cars. The data contains total columns of 25 and 205 rows. The data doesn't contains any null values, which restrains from cleaning the null values.

1.2.1 Log transforming skew variables

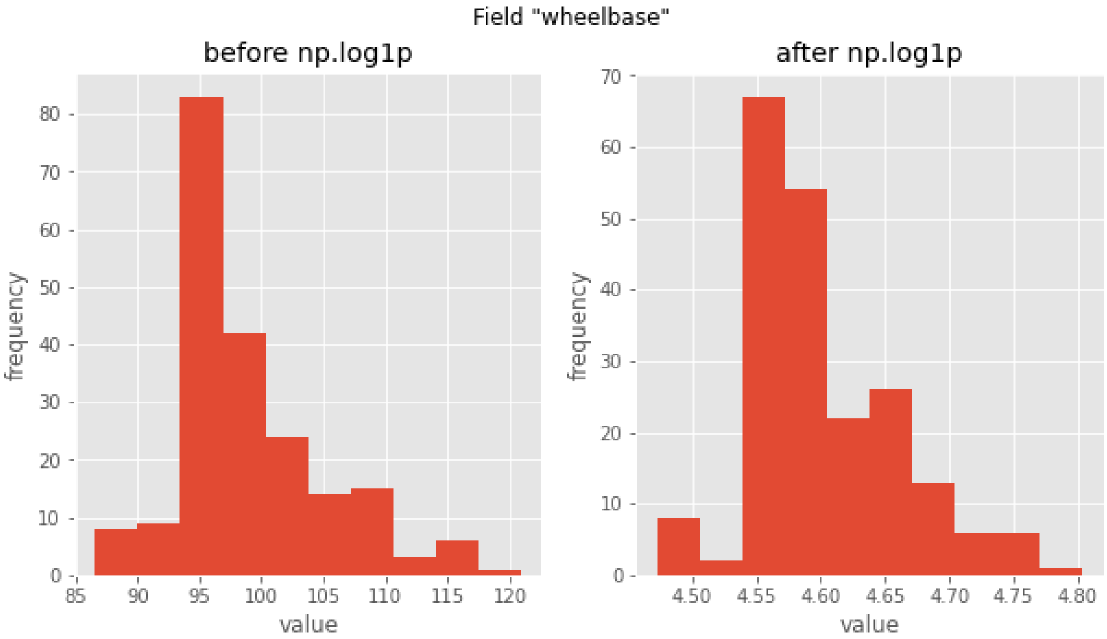
The skewed variables has been log transformed. The skew values of numerical factors are printed below:

```
|                  |         0 |
|:-----------------|----------:|
| wheelbase        |  1.05021   |
| carlength        |  0.155954  |
| carwidth         |  0.904003  |
| carheight         |  0.0631227 |
| boreratio         |  0.0201564 |
| stroke            | -0.689705  |
| compressionratio  |  2.61086   |
| price             |  1.77768   |
```

By setting a skew limit of 0.75, It can be see that the factors, compression ratio, wheelbase and carwidth are over the skew limit. So log transformed these factors for a normal distribution of factors.

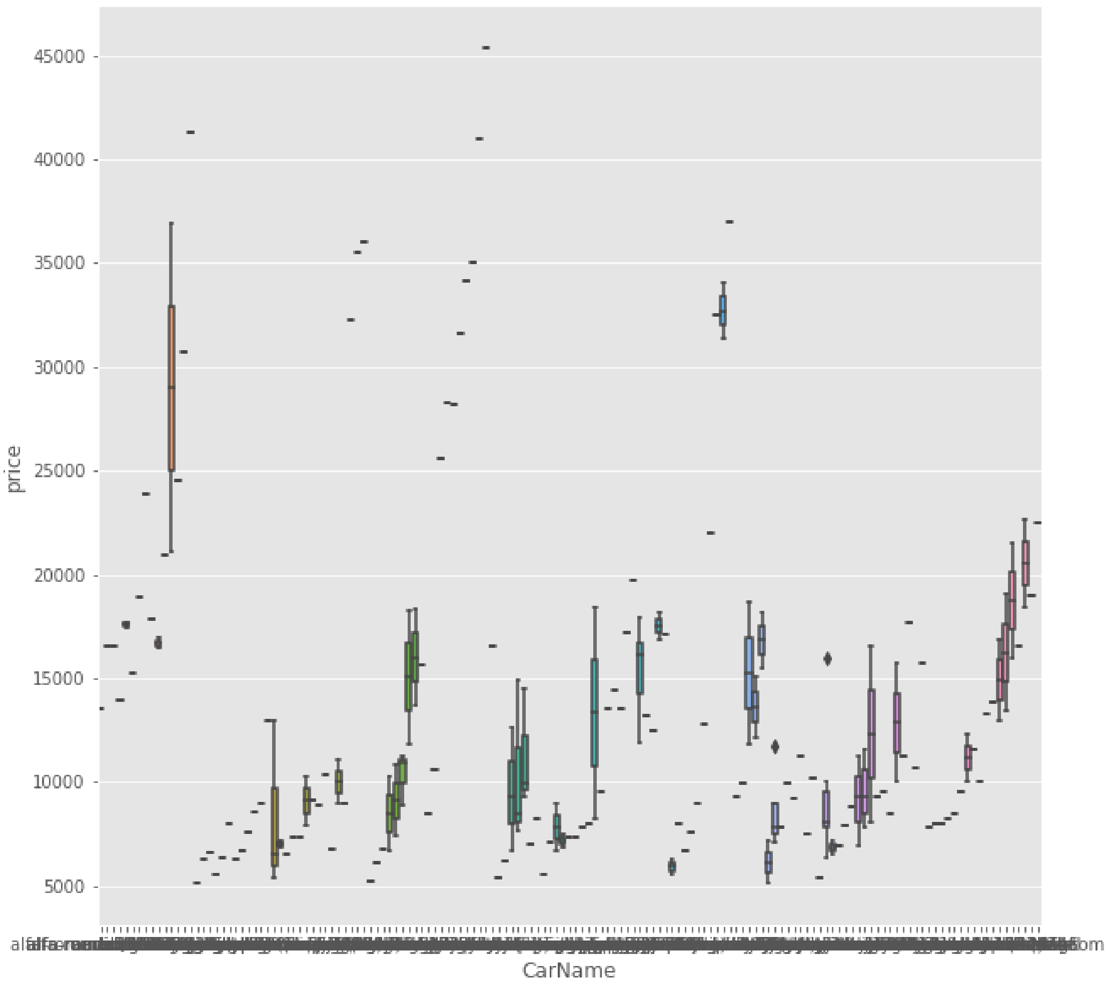
```
|                  |     Skew |
|:-----------------|---------:|
| compressionratio | 2.61086  |
| price            | 1.77768  |
| wheelbase        | 1.05021  |
| carwidth         | 0.904003 |
```

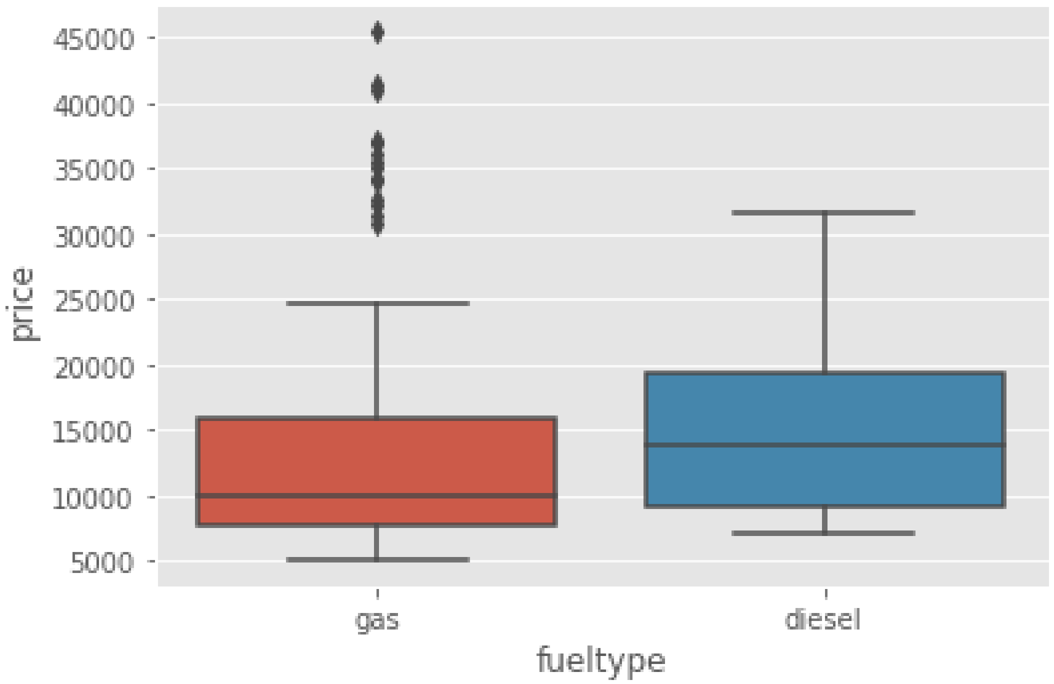
The Distribution of the Field “wheelbase” after and before the transformation is plotted below:

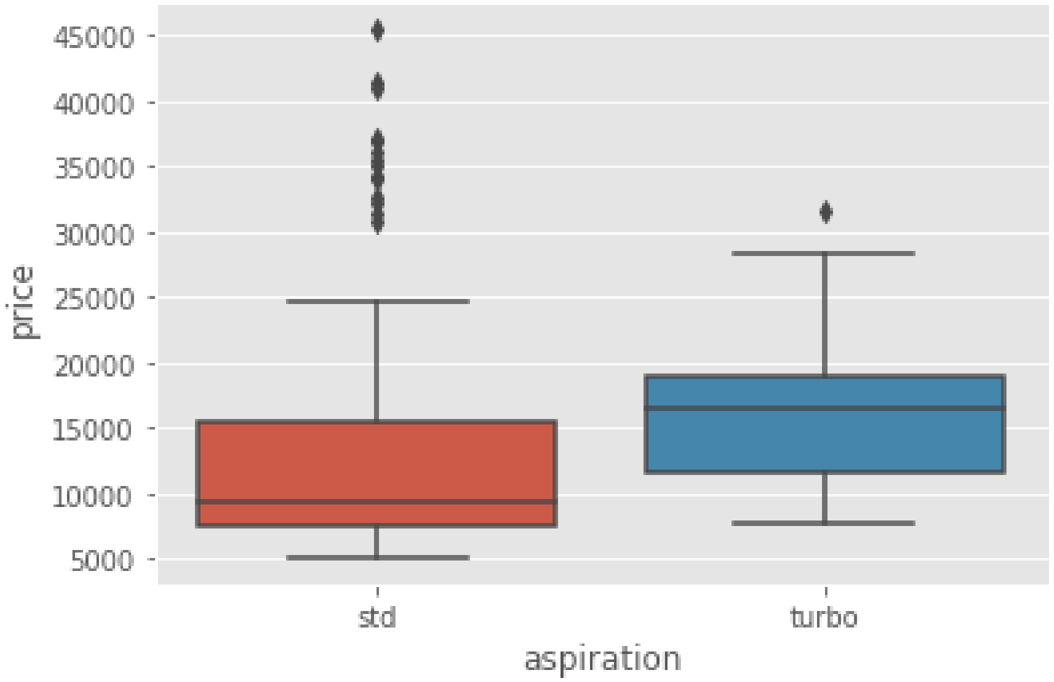


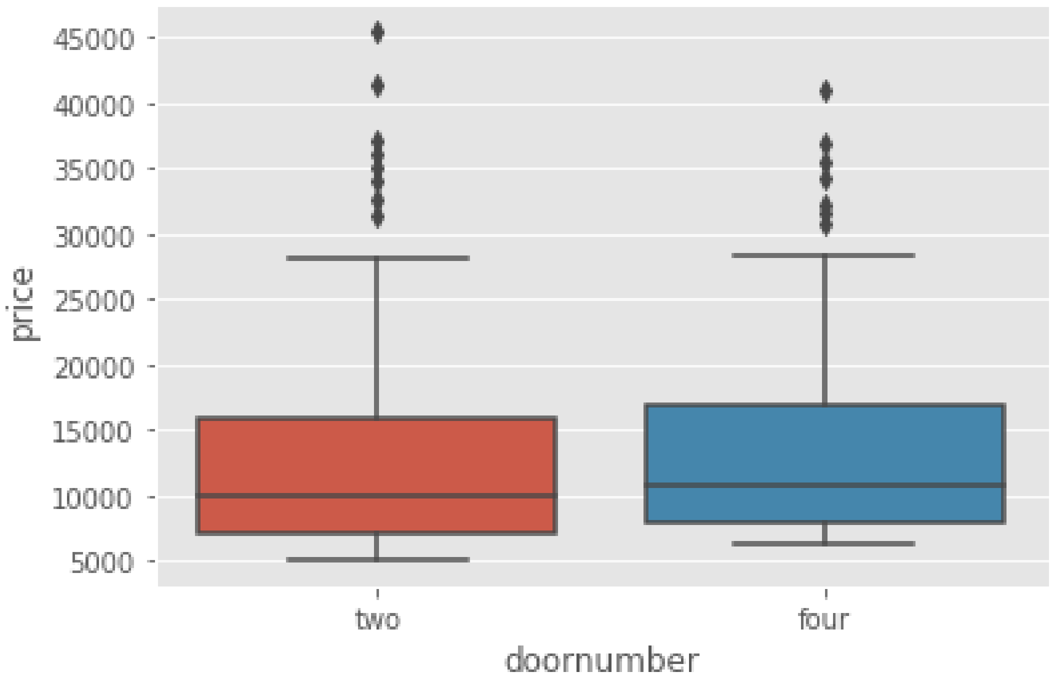
1.2.2 Feature selection

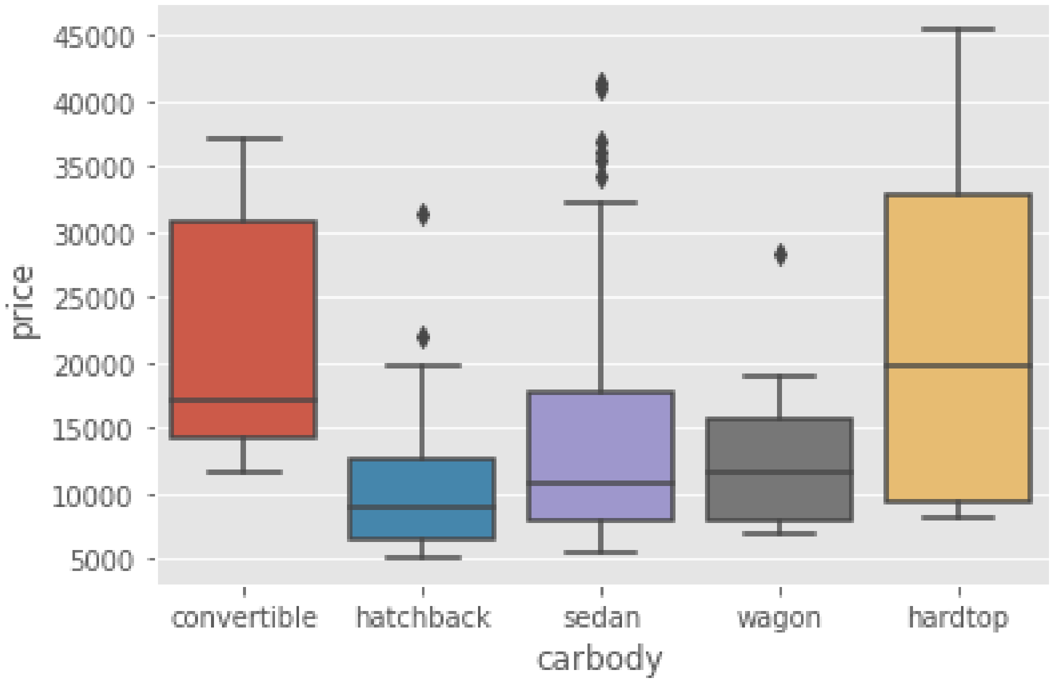
Distribution of numerical features against price Our target column in the dataset is the “Price” column. To find out the correlation of the price column with other factors and factors depends on the price, plotted the box plot of categorical factors with the price. The box plots are shown below:

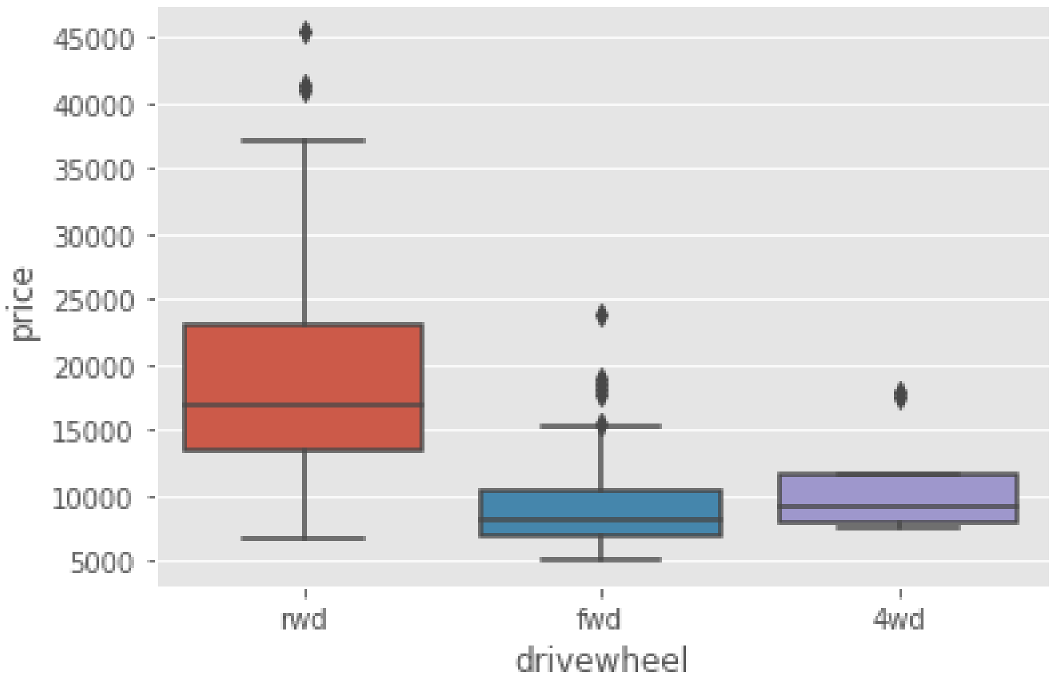

45000
40000
35000
30000
25000
20000
15000
10000
5000
price
CarName

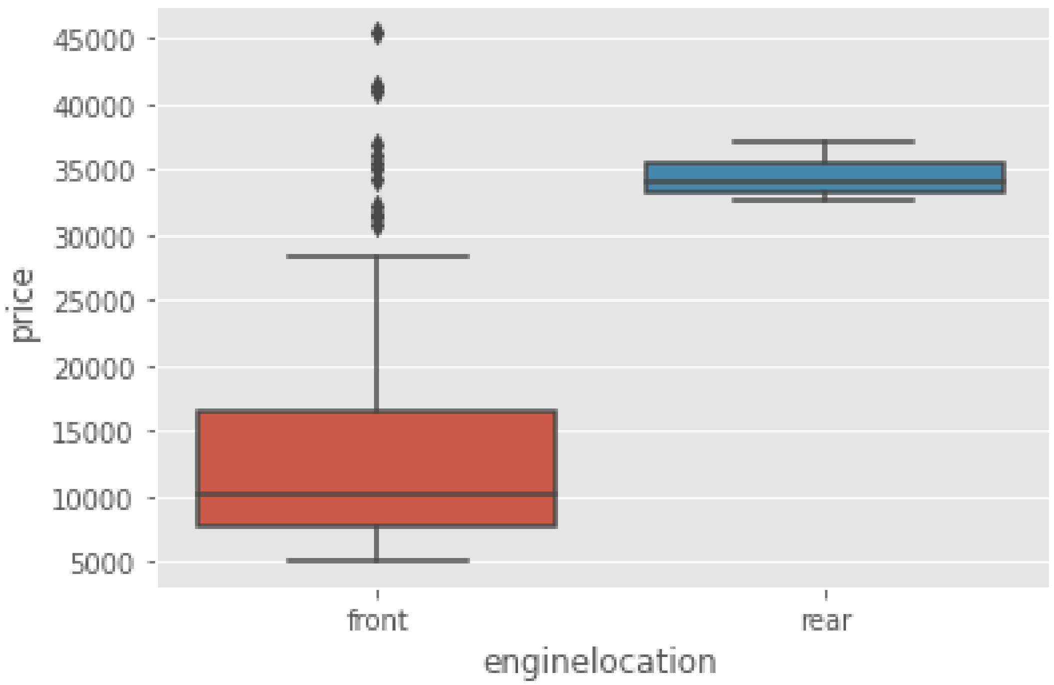

45000
40000
35000
30000
25000
20000
15000
10000
5000
price
gas
diesel
fueltype

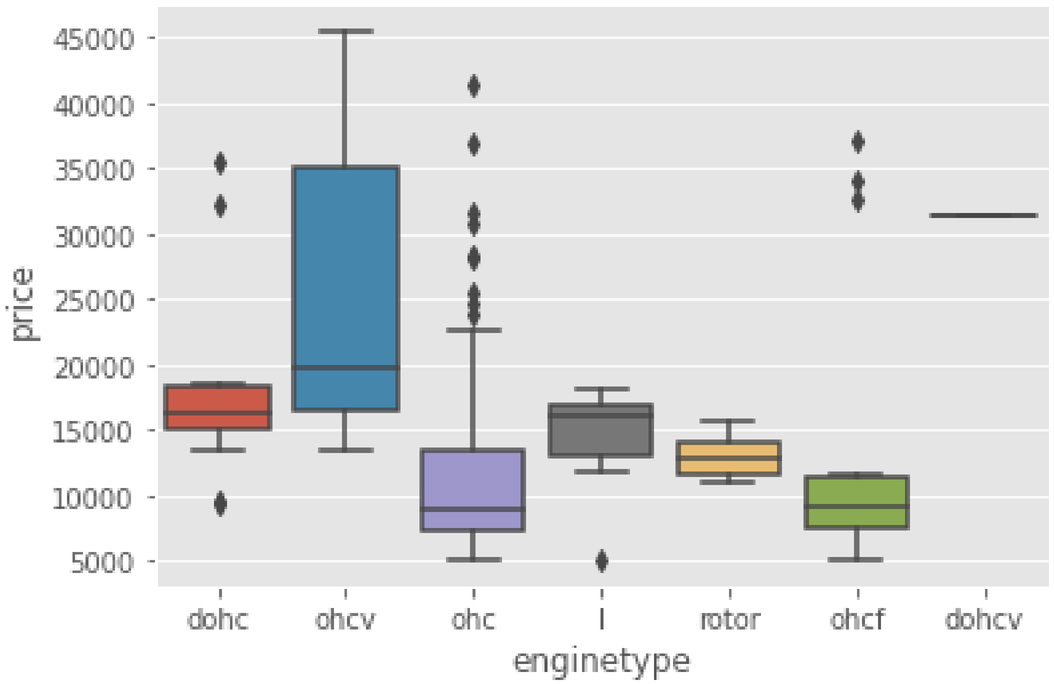

45000
40000
35000
30000
25000
20000
15000
10000
5000
price
std
turbo
aspiration

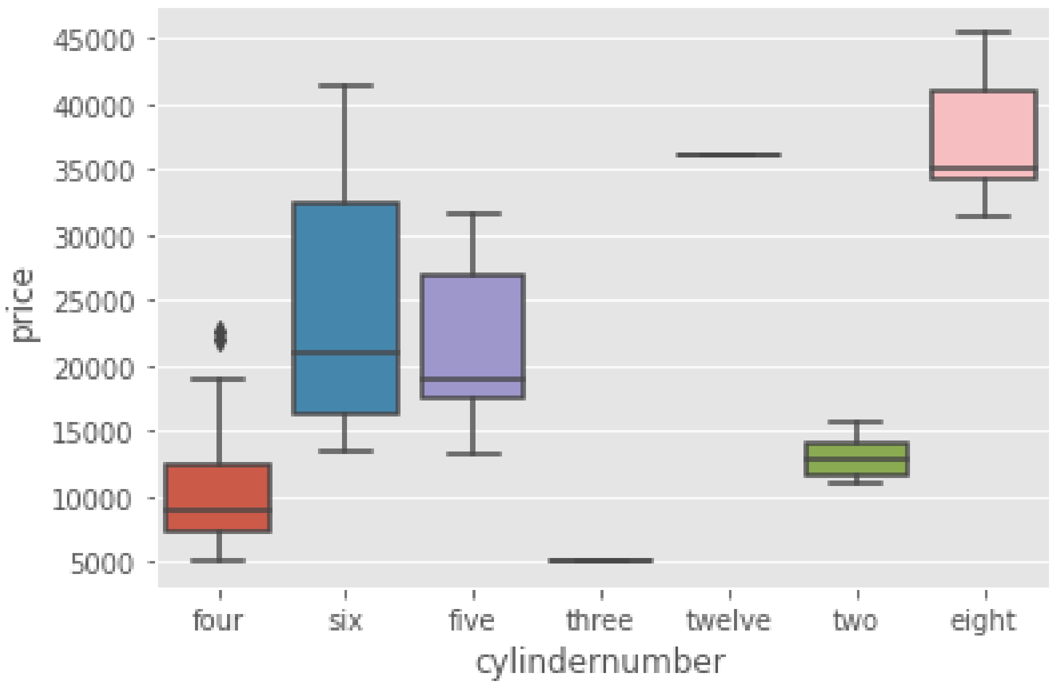

45000
40000
35000
30000
25000
20000
15000
10000
5000
price
two
four
doornumber

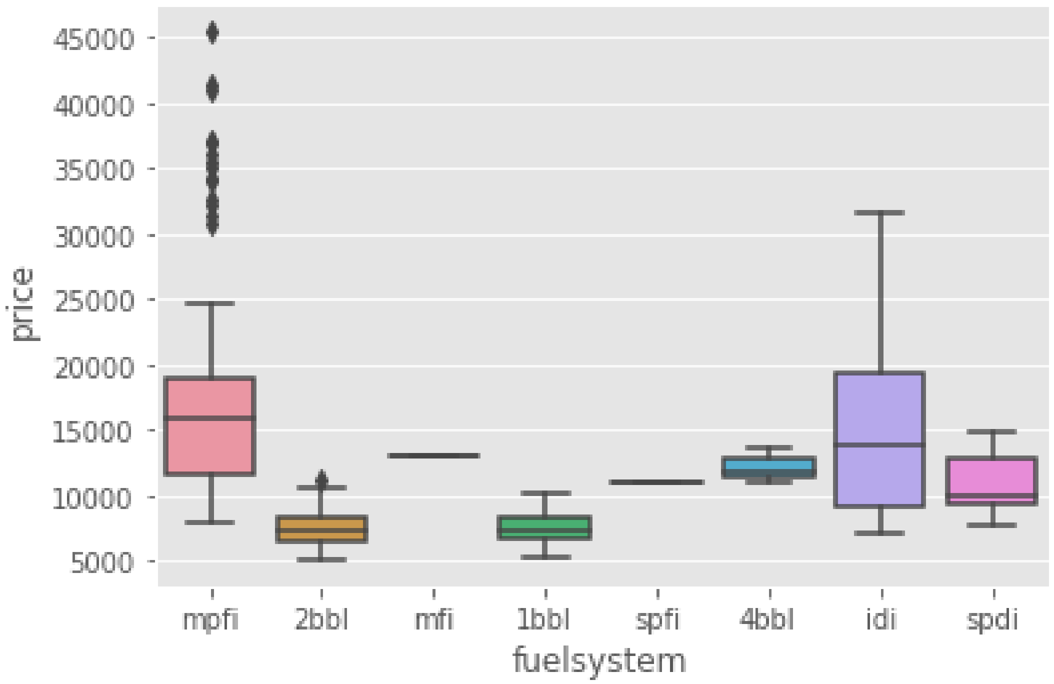

45000
40000
35000
30000
25000
20000
15000
10000
5000
price
convertible
hatchback
sedan
wagon
hardtop
carbody


45000
40000
35000
30000
25000
20000
15000
10000
5000
price
rwd
fwd
4wd
drivewheel


45000
40000
35000
30000
25000
20000
15000
10000
5000
price
front
rear
enginelocation


price
45000
40000
35000
30000
25000
20000
15000
10000
5000
dohc
ohcv
ohc
l
rotor
ohcf
dohcv
enginetype

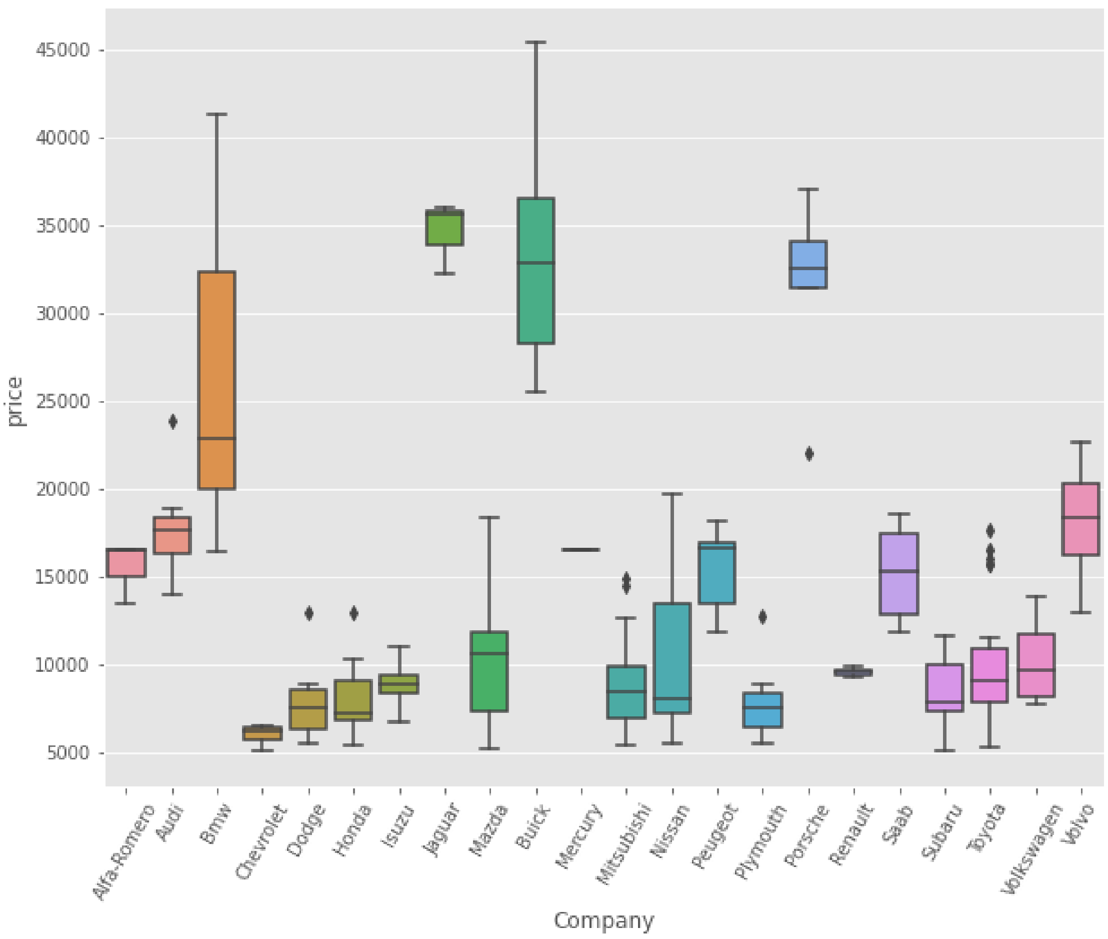

price
45000
40000
35000
30000
25000
20000
15000
10000
5000
four
six
five
three
twelve
two
eight
cylindernumber



From the above box plots, we can see that there is not much price difference when the door numbers are two or four. Therefore assuming a very low correlation we will drop the door number column. For the car name column, we can't see the labels but the plot is somewhat depicting a high correlation with price.

To visualize this better let's split the column and also rename the values in it.

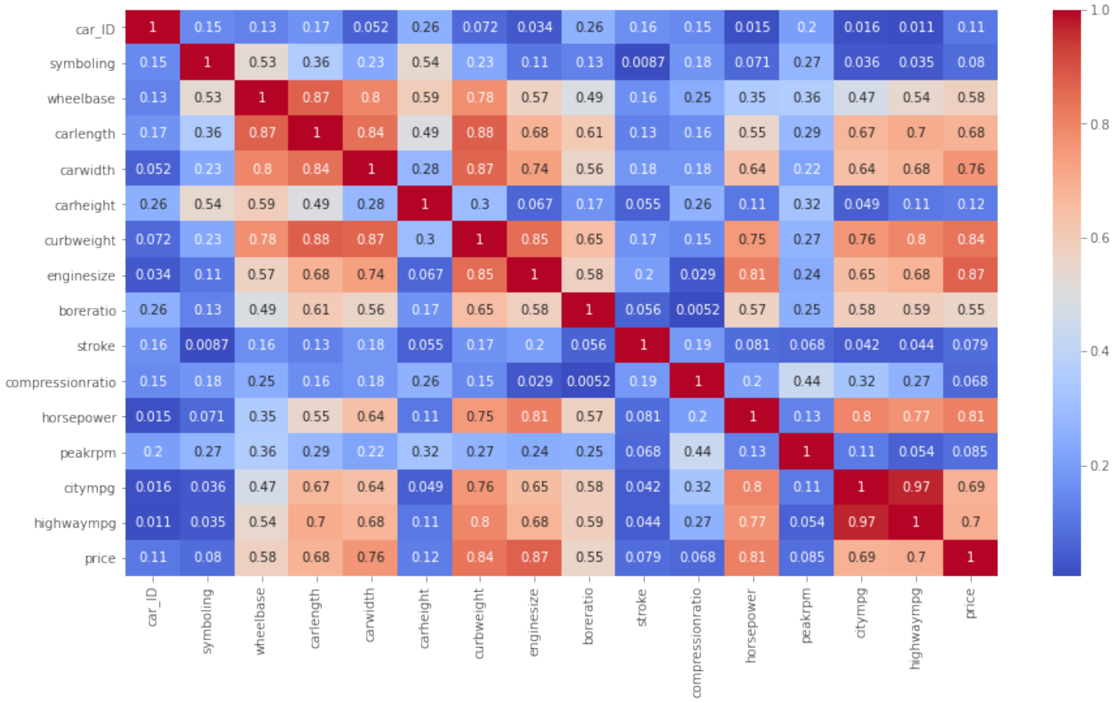
After renaming the car name the boxplot of car name and price looks like:



From the above visualizations, we can see how widely the prices vary from company to company. We therefore can use this feature to train our model and predict the price based on company names rather than using car names (models).

1.2.3 Pearson correlation

To find out the overall correlation among all numerical variables, plotted the heatmap of the Pearson correlation and the heatmap diagram is shown below:



The correlation of factors with the price columns is:

```
|                  |     price |
|:-----------------|----------:|
| highwaympg       | -0.697599  |
| citympg          | -0.685751  |
| car_ID           | -0.109093  |
| peakrpm          | -0.0852672 |
| symboling        | -0.0799782 |
| compressionratio |  0.0505663 |
| stroke           |  0.0794431 |
| carheight        |  0.119336  |
| boreratio        |  0.553173  |
| wheelbase        |  0.568669  |
| carlength        |  0.68292   |
| carwidth         |  0.755968  |
| horsepower       |  0.808139  |
| curbweight       |  0.835305  |
| enginesize       |  0.874145  |
| price            |  1         |
```

From the correlation matrix we found out that car height, stroke, compression ratio, and peak rpm have no noticeable effect on the price of the cars therefore, we will drop these columns. car_ID column is also irrelevant for the prediction of the car price. From the box plot visualizations above we saw carNames and door number attribute can also be dropped.

After feature selection, one-hot encoding for categorical values has been carried out.

Here we haven't removed the outliers. I think the outliers here may represent a real picture for eg the prices of some cars may in the real world be too high. Therefore I believe that removing outliers in the data set at hand would be not a wise thing to do.

1.3 Data Modelling

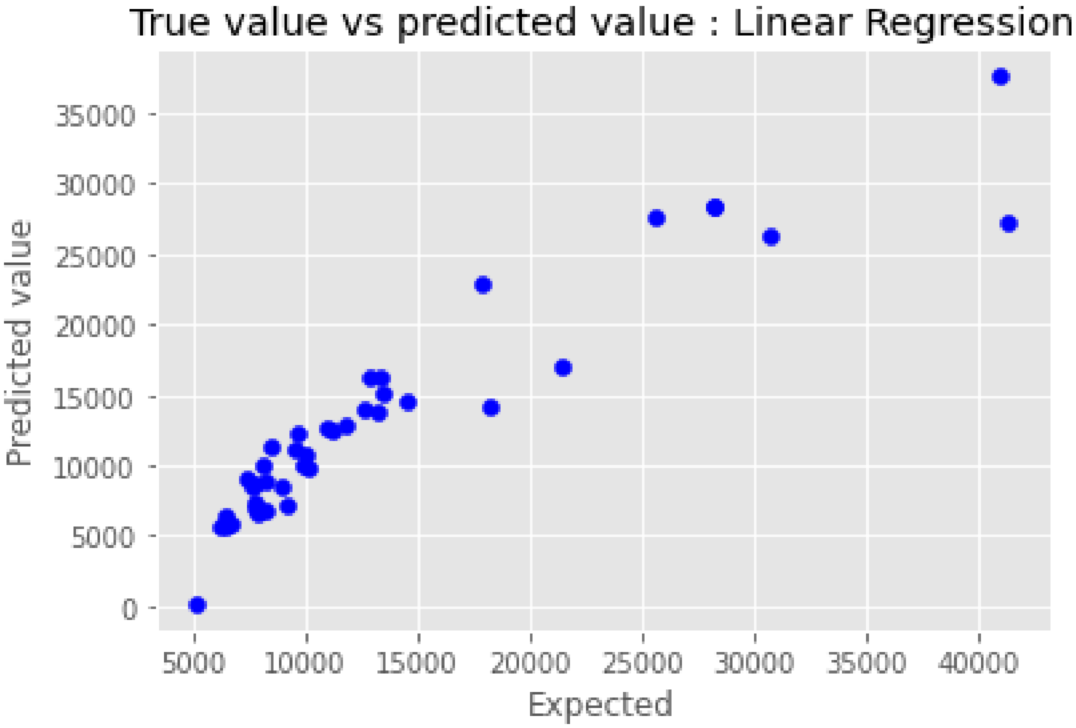
In this section, we will try different models for predicting the car price and finds out which model is best suitable for the prediction of price.

Before splitting the data into train and test sets, the data set transformed into arrays. Then split data to train and test data with 80 and 20 percent respectively. After that, the standard scaling of the X data set has been carried out.

1.3.1 Linear Regression

We start by applying the linear regression model to the data. The accuracy of the model has been measured through Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and R^2 values.

The scatterd plot which visualize the relation between the actual values and predicted values is given below:



The above plot shows approximately a straight line, but the accuracy of prediction would not be that good. The metrics of the predictions are:

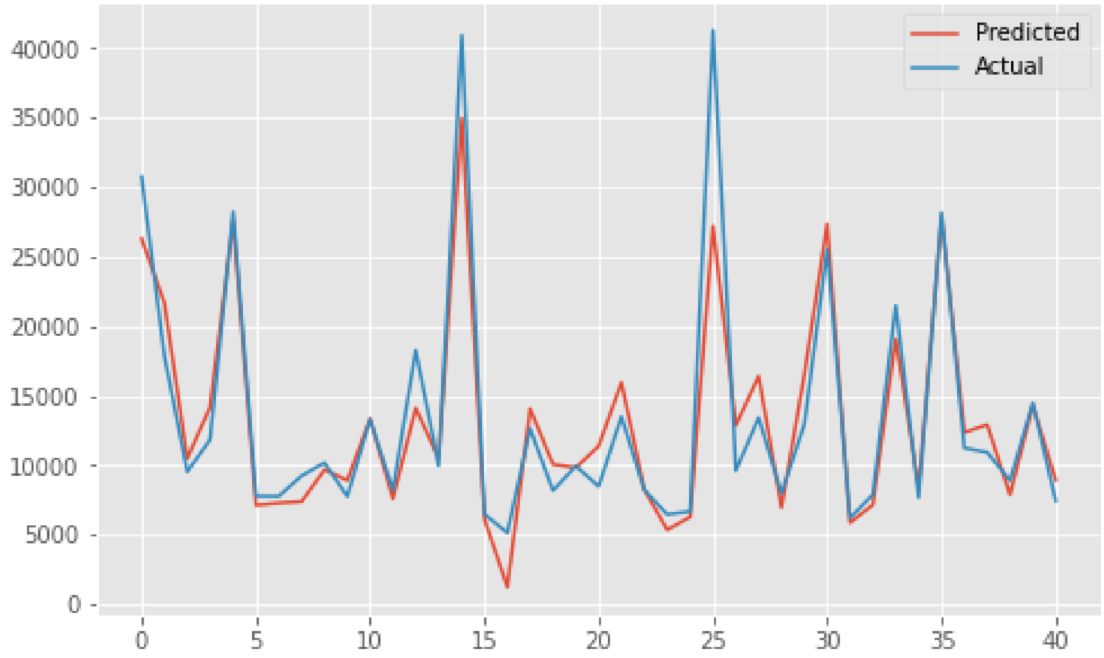
```
Mean Squared Error is: 9501168.015570093
Mean Absolute Error is: 1970.039894410287
Root Mean Squared Error is: 3082.396472806523
The R2 value is: 0.8796467685620716
```

1.3.2 Ridge Regression

The second model we will be using is ridge regression. We used GridSearchCV for finding the best alpha parameter for the prediction.

The best alpha value after the cross-validation is 5 for ridge regression with the best R2 score of 0.91.

The plot comparing actual and predicted values after the model prediction is shown below:



The metrics of the Redge regression is given below:

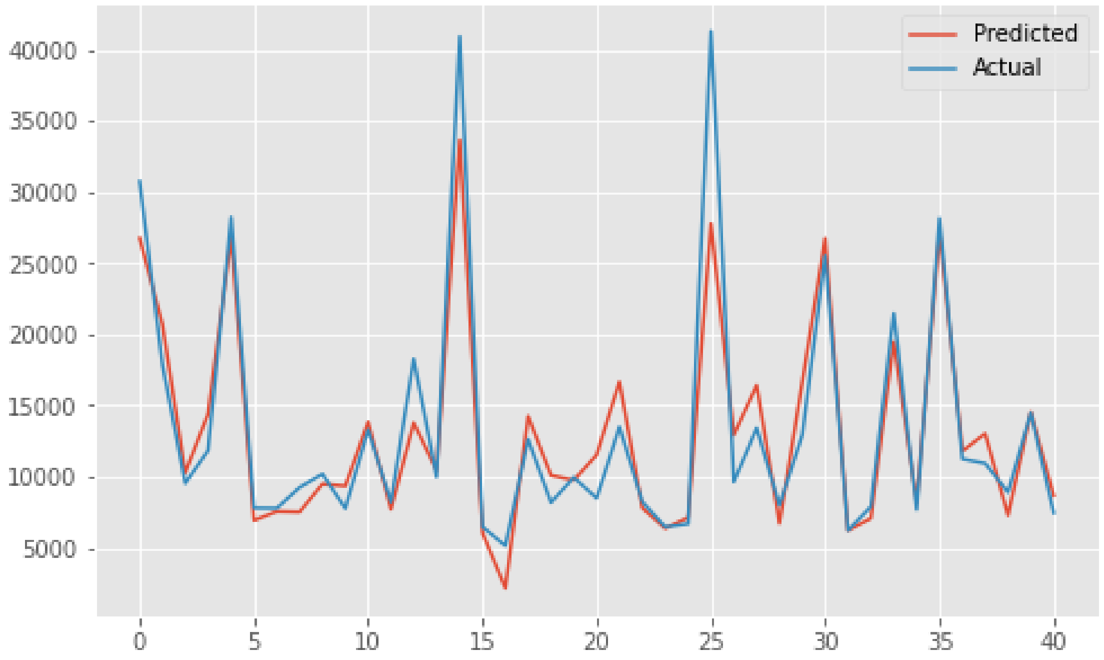
```
Mean Squared Error is: 9468258.115130369
Mean Absolute Error is: 1945.7484939724832
Root Mean Squared Error is: 3077.053479406942
The R2 value is: 0.8800636449774482
```

From comparing the R2 value of the predictions, It can be seen that Ridge regression is slightly better than the basic linear regression.

1.3.3 Lasso Regression

After the ridge regression, the lasso regression has been carried out. Like the ridge, using GridSearchCV, the cross-validation for the alpha parameter has been carried out. The best alpha value was found to be ‘alpha= 40’ with an R2 score of 0.91. The plot below shows the variation of predicted values from the actual values:

```
Mean Squared Error is: 9339487.532622132
Mean Absolute Error is: 1952.2430540498651
Root Mean Squared Error is: 3056.057514612926
The R2 value is: 0.8816948081874465
```

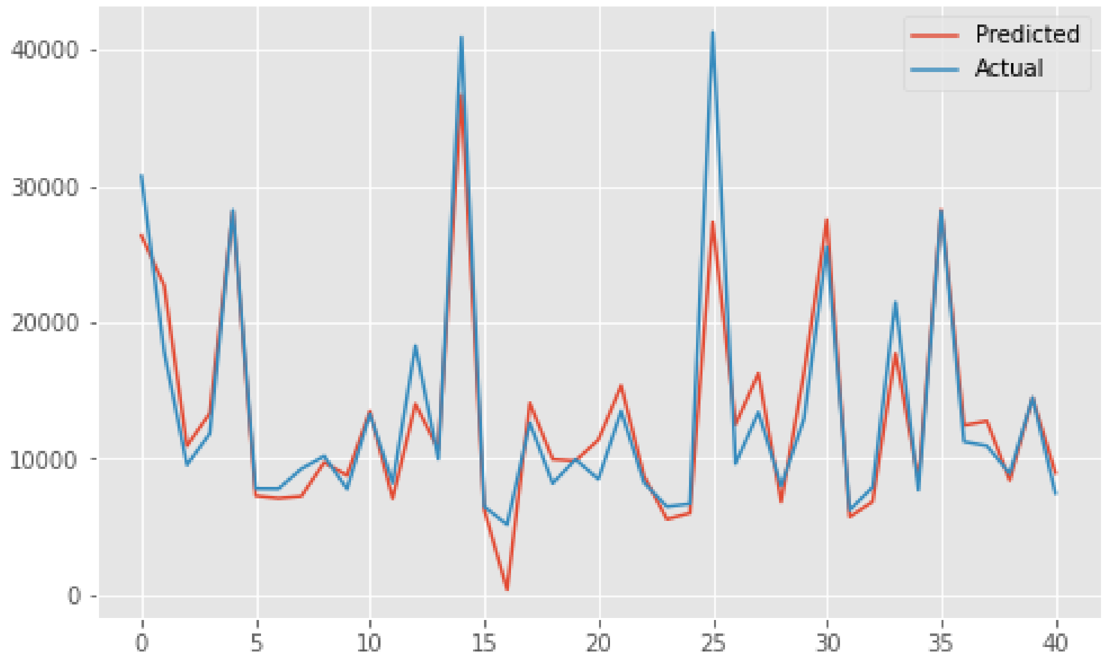


By comparing the R2 values of the ridge regression and lasso, it seems that lasso is slightly better in predicting the price. Since lasso consumes much more time than the ridge because of the iteration, relatively ridge regression will be better.

1.3.4 Elastic Net Regression

Apart from ridge and lasso, the Elastic net regression was also carried out and the metrics of the predictions have been calculated. The best alpha is ‘0.01’

The plot showing the relation for elastic net regression is shown below:



The metrics of the prediction is:

```
Mean Squared Error is: 9369553.879408361
Mean Absolute Error is: 1955.9020316413844
Root Mean Squared Error is: 3060.972701513093
The R2 value is: 0.8813139516456691
```

1.3.5 Model Evaluation

The summarised metrics of all the models is given below:

```
[174]:                     Model  Mean Square Error  Mean Absolute Error  \
       0        Linear Regression       9.501168e+06          1970.039894
       1         Ridge Regression       9.468258e+06          1945.748494
       2         Lasso Regression       9.339488e+06          1952.243054
       3   Elastic Net Regression       9.369554e+06          1955.902032

          Root Mean Square Error  R2 score
       0             3082.396473  0.879647
       1             3077.053479  0.880064
       2             3056.057515  0.881695
       3             3060.972702  0.881314
```

So from the table above, the Lasso regression has the best R2 score relatively. But for a large data set, this model consumes comparable much time than other models. So in that case I would prefer Ridge for predicting the Price of the car. The difference in the accuracy between ridge, lasso, and the elastic net is not that drastic.

1.4 Conclusion

In this project, we tried to predict the car price for an automobile company. Figured out the factors which are affecting the price of a car. Then applied various regression models such as Linear, Ridge, Lasso, and elastic net. From the careful examination of metrics of the different predicted models, concluded that the Ridge regression model performs relatively better in terms of accuracy and time taken for the prediction.