

CAR PRICE PREDICTION

Submitted by:

ASHISH CHAND

**ACKNOWLEDGMENT**

This used car data is collected by me from the online website of Second hand cars. The name of the website is [Cartrade.com](https://www.cartrade.com/). This project contains Excel file of scraped data, jupyter notebook for writing the code for the model training.

**INTRODUCTION**

* Business Problem Framing

From the last 1 year, due to covid everything changes,The business all around the globe changes with that. Like some business goes shutdown, Some got work from home and therefore now everything is come-up with the online process. With the covid 19 impact in the market, there 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.

* Conceptual Background of the Domain Problem

In today’s world, everything is now online. Like food delivery to cars, bike delivery to home. Due to work from home and stay in home, people now searching everything Online. Buying a own car is everyone’s dream and due to this covid effect, people can’t go the showroom and also people like to compare first before buying from the shops. So they search online first and get all the information about the car. To buy a second hand cars there are many online websites are there. Our task is to get the data from these online websites.

So, this project is all about to collect the data of used cars online which contain Brand, model of the cars, their transmission and the price. Basically everything which is important for the person to buy a car. In this project, first I have to scrape the car data online and after that train the data for the prediction of upcoming dataset of used cars. Also the data science part will be done like Exploratory Data Analysis, data pre-processing and all.

* Review of Literature

Due to the impact of the corona virus in our life. We do not go outside for buying anything, so now people use online website and resources for the information and for buying and selling anything. This project is all about the price of used car. 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 this project is done two phase - first is to collect the data websites of used car and collect all the important and the necessary details of the car. So that in the second phase, The machine learning model can be trained to predict the price of cars, by giving it some data values. Data science comes as a very important tool to solve problems in the domain to help the client to increase their customers, profits, improving their marketing strategies and focusing on changing trends in car price data. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for Car price prediction. Our client are looking for new machine learning models from new data. So we need to build a model using Machine Learning in order to predict the car price value of the prospective data and give the car price output.

* Motivation for the Problem Undertaken

To buy a home or a new car or second hand car is dream of many peoples and every person life revolve around this across the world and that’s why online car website and real estate markets are increases day by day and contribute a very handy help in increasing the world’s economy. The objective behind this project is that to get the knowledge about this domain. And to know more about the cars and the factors which are important for buying a second hand car. In today’s technology smart world everything is online. Due to covid 19 many offline shops are closed. That’s why they also want to do business online and the marketing of cars selling is buying is very vast. As a data scientist it is very interesting to know about all the information about the cars, what are the factors which are necessary for buying and selling the cars. And how the online car market work. It is a very large market and many companies work in this domain. Data science comes with a very important tool to solve the problems in the domain to help the make the model for car sales and the give the good information regarding the factors which are necessary for buying a car.

Recommendation systems are some of the machine learning techniques used for achieving the business goals for Car selling companies.

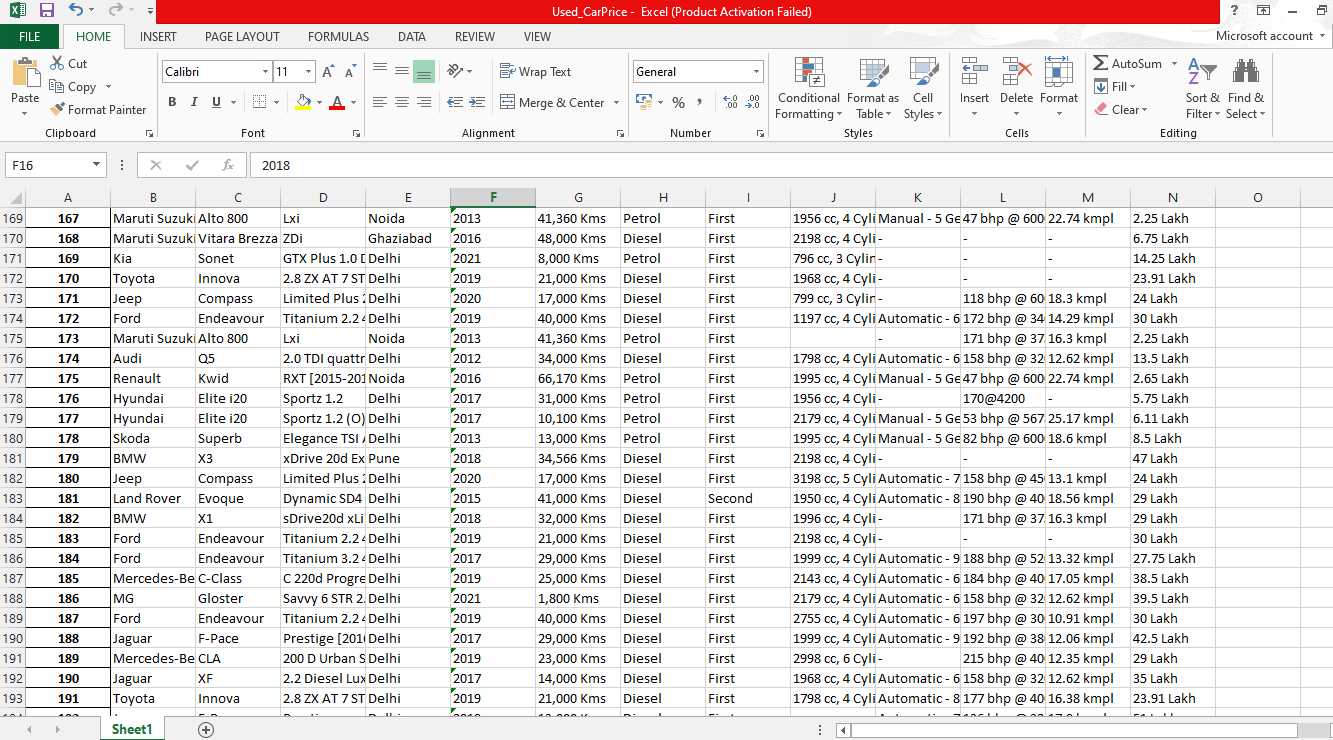
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In this project, as we analyse the dataset we get to know that:

* The first step of this project is collect the data online from the online car websites.
* The data is not normally distributed as the data is not complete and some features were not included, which made the right or left skewed data somewhere.
* Hence we use a log function to make the feature Normally Distributed.
* Also the skew method is used for to reduce the skeness of the data.
* And also we use the zscore method of statistics to remove some outliers from the normal distribution.
* To see the inter relation of features with the target column we use the correlation method, and then visualize with the help of Heatmap to see the correlation between features and with the target column.
* And then I done the Feature selection process where I convert the categorical column into the numerical with the help of label encoding and used the statistical libraries like skew for checking the skewness and norm for the normal distribution and zscore for the outliers removal.
* Here I use the z-score with the threshold value of 3 to check the outliers present in the dataset.
* And finally the dataset get divided into x(without target column) and y(only target column present) for the model training.
* And atlast the model with highest accuracy score is selected for the model prediction.
* Data Sources and their formats

The data is collected online website the cars name <https://www.cartrade.com/>with the help selenium and save it in the form of excel sheet.



* The datasets contains 13columns and 6267 rows –
  + Brand : The brand name of the car.
  + Model : The model name of the brand car.
  + Variant : The variant of the model
  + Location: The location in which the car is being sold or is available for purchase.
  + Manufacturing Year: The year or edition of the model.
  + Kilometers\_Driven: The total kilometres driven in the car by the previous owner(s) in KM.
  + Fuel\_Type: The type of fuel used by the car.
  + Transmission: The type of transmission used by the car.
  + Owner\_Type: Whether the ownership is Firsthand, Second hand or other.
  + Mileage: The standard mileage offered by the car company in kmpl or km/kg
  + Engine: The displacement volume of the engine in cc.
  + Power: The maximum power of the engine in bhp.
  + Price: The price of the used car in INR Lakhs.
* Data Preprocessing Done

In the Data Pre-processing, at first I check the null values present in the dataset, after that there were some missing values present in the engine column. Which were filled by the method fillna of python. After that I done cleaning of the columns i.e. getting the desirable output in the column with the help of ‘Astype’ function, map and the lambda function.

Then the data type of the columns were changed as some numerical columns were present as the object data type which will convert into the float and integer data type. The data type was converted with the help of astype function().

After that I drop the two unnecessary columns which were not useful for the next processes.

Then I fill the null values of the numeric data type with the mean of that column and for the categorical column, the value was filled by the mode value of the column.

After that with the help of the seaborn library I draw the Heatmap to check the correlation between the columns. Then I check the distribution in the sales column as it was positively right skewed. So I use the Log transformation function to make it normally distributed.

After that I use the Univariate Analysis and check the columns graphically. And then the bivariate analysis is done between the two columns. Then I perform the feature engineering by first storing the all continuous column into a list and then perform the skew method to remove the positive negative skewness from the data. Then we do the One hot encoding and convert the categorical features into the numerical features .Then the outliers were checked and for that I make the threshold value 3 , above than that the values are outliers and which were removed by the zscore. And after that scaling is done and splitted the data into two parts – 1 is X(without dependent variable) and 2nd is y (Only target variable) and after that model selection and model prediction is done.

* Data Inputs- Logic- Output Relationships

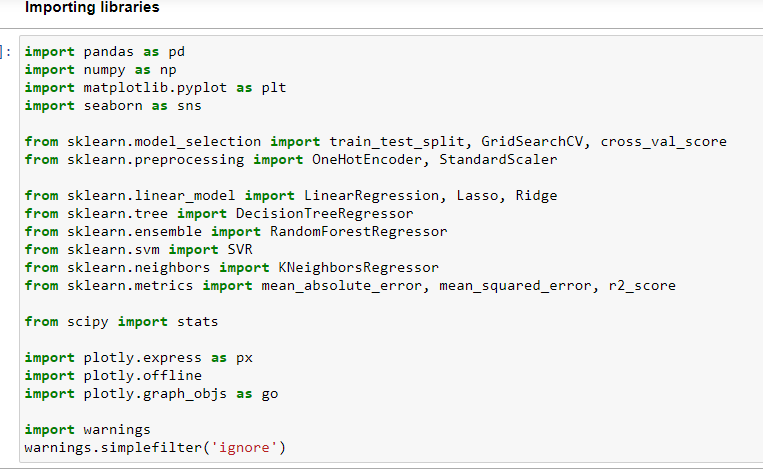
By the visualization of the data we see the relationship between the features and also with the correlation tells about the all features related to the Target variable. By the correlation method we can see the relationship and effects of features on the target variable. And also we see the correlation between the features.

By correlation we came to see that some features(Manufacturing year , Power ,engine of the car) are highly correlated with the target column(Price). That means if the value of dependent feature increases, Price also increases.

* Hardware and Software Requirements and Tools Used

The Project is done on the Window 10, here I use the Software Anaconda platform (Python 3.8.5 64 bit) and the code is written on the Jupyter Notebook where I run different python libraries for the better understanding. The libraries were:

* Pandas – Pandas library is used for the uploading the file and then to create the dataframe from that file and manipulation of dataframe and use some in-built operations.
* Numpy – Numpy is used for the mathematical operations in the dataframe. Here Numpy is used to calculate the mean, median, mode and also the various mathematical operations were used in the dataset, which helps in the better understanding of the different features.
* Matplotlib – Matplotlib.pyplot library is used for the visualization for the dataset. And it helps to understand the data with help of graphs. And also it helps to label the graph, x-axis or y-axis. It also helps to select the size of the graph.
* Seaborn – Seaborn again is used for the visualization and it helps us to visualize the data more efficiently and clearly and understand it properly. Here I used the Heatmap which helps to understand to see the correlation between features and also here I use the various graphs like histogram, Bar graph, Scatterplots, line plots etc.
* Plotly –The Plotly Python library is an interactive open-source library. This can be a very helpful tool for data visualization and understanding the data simply and easily. Plotly graph objects are a high-level interface to Plotly which are easy to use. It can plot various types of graphs and charts like scatter plots, line charts, bar charts, box plots, histograms, pie charts, etc. Here it is used to check the missing values % and represent it graphically. The bar graph is used for understandable and meaningful.
* Scipy – This library is used for the statistical formula and here I used it for calculate the skew for the checking of skewness and visualize it graphically also, norm is used for the normalization and here I used with the visualization for the better understanding of the data and also the boxcox1p is used for the removal of outliers.
* Sci-kit learn – Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy. Scikit-learn library is very important in the machine learning process. Here I use the library for model selection such as Linear regression , Lasso, Ridge Support vector machine,k-nearest neighbour, some ensemble techniques such as RandomForest Regressor, metrices like Mean absolute error, root mean squared error , r2score, Pre-processing tools like Onehot encoder, StandardScaler and model selection like GridsearchCV , train test split and cross-validation value.



**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Here we first check the data is supervised or unsupervised by checking the target column and then when we see the output it shows it’s a regression problem.

* First, I import the necessary libraries, such as pandas, numpy ,seaborn , matplotlib, scipy and sklearn.
* After that with the help of pandas I create the dataframe from the given set of data.
* Then I check the information the dataframe which shows that 6267 rows and 13 features and also the data type int64 (1), object (12) were present.
* Here some columns were present as the object data type which was converted later.
* Then I check the missing values in the dataframe and calculate the missing value % and visualize it with the help of seaborn library by drawing the heatmap.
* This gives that Engine feature have the highest number of missing values.
* Then I check the correlation of features and found that manufacturing years , engine and power were negatively highly correlated first.
* Then I came to know that the our target column is positively right skewed as the difference between the mean and median was very high.
* After that I groupby the missing value containing features and the target column to see the relation between them which tell about the null and not value present in the dataset graphically.
* As Our dataset is containing both Object data type and Integer, float data type.
* As some columns were in object datatype, therefore with the help of astype function and lambda or map function we first capture the numerical value and then store them as a numeric value with float data type.
* After that by the help of fill na method missing values of numerical columns were field with mean of that column while the categorical column with the mode of that column.
* Pie graph was used to check the distribution of locations in the dataset.
* Barplot were used to represent the discrete features for the better understanding of dataset.
* Histogram were used to represent the continuous features and it shows that some features were not normally distributed i.e. skewness is present.
* Then I use the log transformation method for reducing the skewness and visualize with the histogram.
* After that I check the outliers with the help of zscore and boxplot for the graphical representation and remove them with the help of zscore.
* After that I fill the missing values with the None.
* Then I check the Target column, which shows that column is normally distributed.
* After that I used the Year column and subtract the yearbuild from the year sold. So this represent the number of years.
* Then I use the label Encoder and convert the categorical columns into the numerical columns.
* After that I remove the skewness by using Z score method.
* And then split the dataset in x and y where x contains all the dependent columns and y contains only target column.
* After that the model training is done and select the best model which have less RMSE and used for the prediction.

Random forest regressor is then used for the testing and for the prediction.

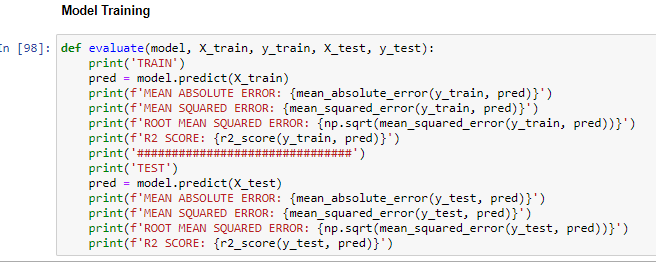
* Testing of Identified Approaches (Algorithms)

Here we use the following Algorithms used for training & Testing:

1. Linear Regression
2. Lasso
3. Ridge
4. Support vector machine
5. K-nearest Neighbors Regressor
6. Random forest Regressor

* Run and Evaluate selected models

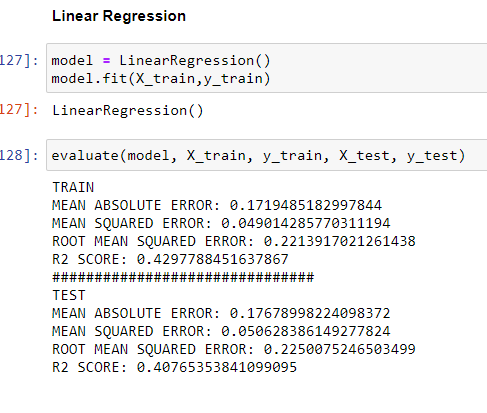
The algorithms used were :



Above is the function used and after that the model were run on it.

1. **Linear Regression :**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

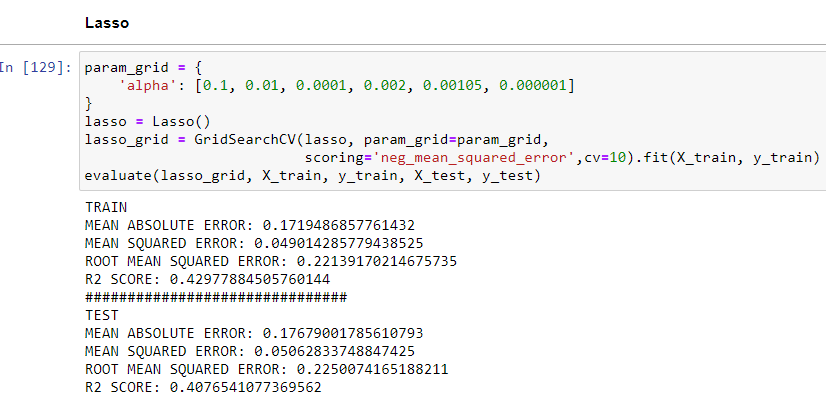


1. **Lasso :**

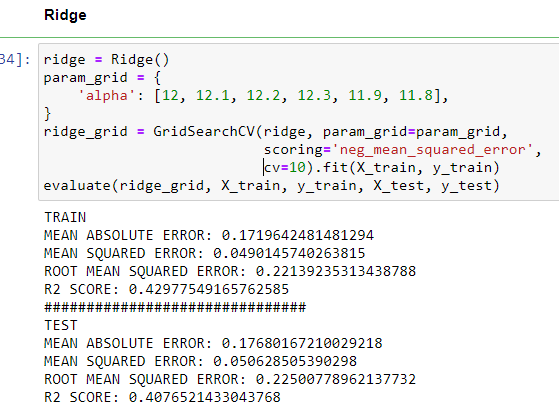
Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso Regression uses L1 regularization.

It is used when we have more number of features because it automatically performs feature selection.

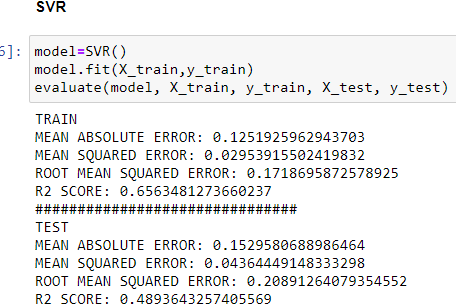


1. **Ridge:** Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

****

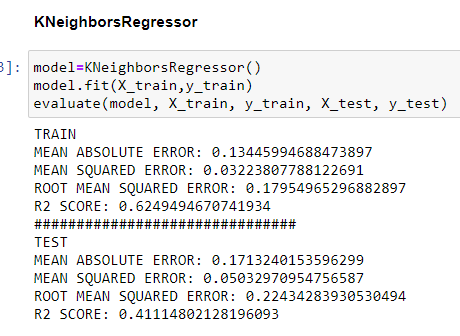
1. **Support Vector machine**

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.



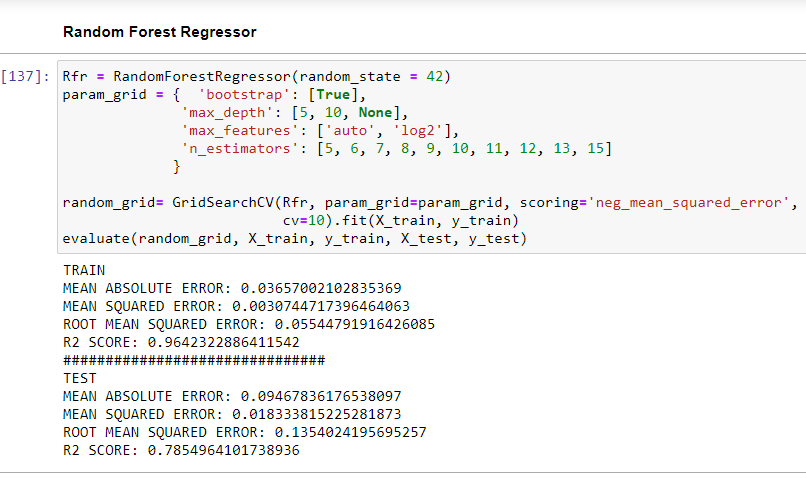
1. **K- Nearest Neighbors regression**

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.



1. **Random Forest Regressor:**

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

****

* Key Metrics for success in solving problem under consideration

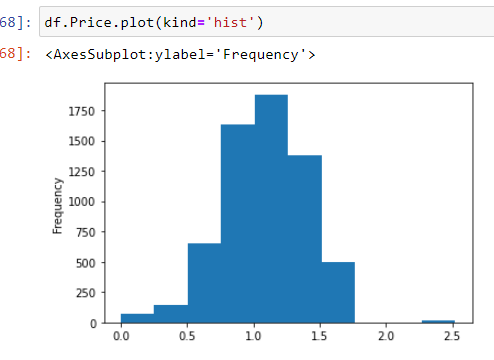
The key metrices used along with the project are :

* **Mean absolute error : (MAE)** represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
* **Mean squared error**: (**MSE**) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
* **Root Mean squared error**: (**RMSE**) is the error rate by the square root of MSE.
* **R2 score : R-squared (Coefficient of determination)** represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.
* Visualizations

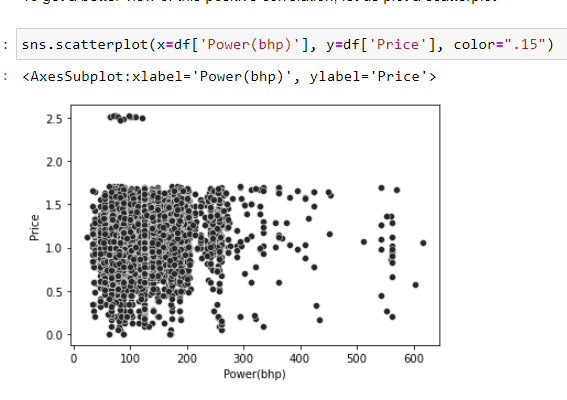
If different platforms were used, mention that as well.

Here we use the following graph:

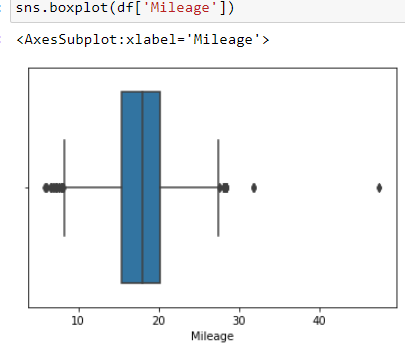
**Histogram-**  for checking the distribution of data of single column. A histogram is basically used to represent data provided in a form of some groups. It is accurate method for the graphical representation of numerical data distribution. It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency



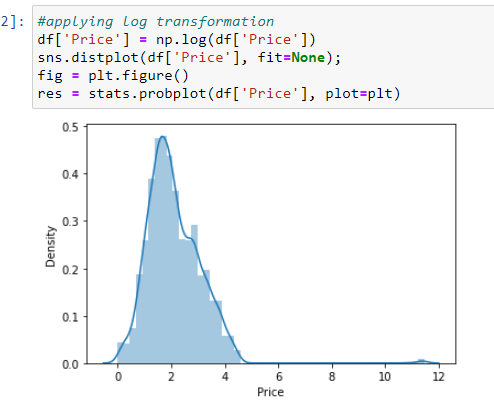
1. **Scatterplot** to see the distribution of data among two columns. Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them. The scatter() method in the matplotlib library is used to draw a scatter plot. Scatter plots are widely used to represent relation among variables and how change in one affects the other.



1. **Box** **plot** : A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median. Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.

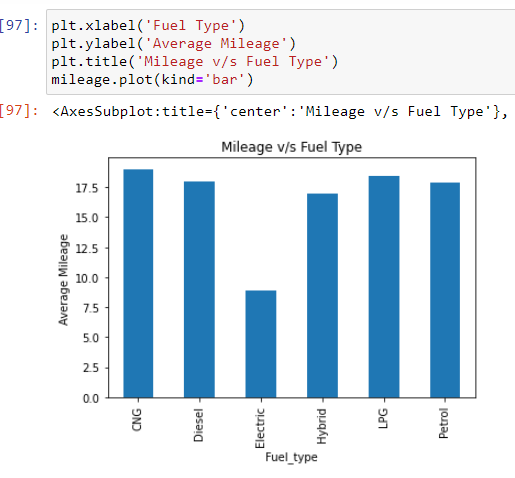
****

1. **Distribution plot:** Python Seaborn module contains various functions to plot the data and depict the data variations. The seaborn.distplot() function is used to plot the distplot. The distplot represents the uni variate distribution of data i.e. data distribution of a variable against the density distribution.

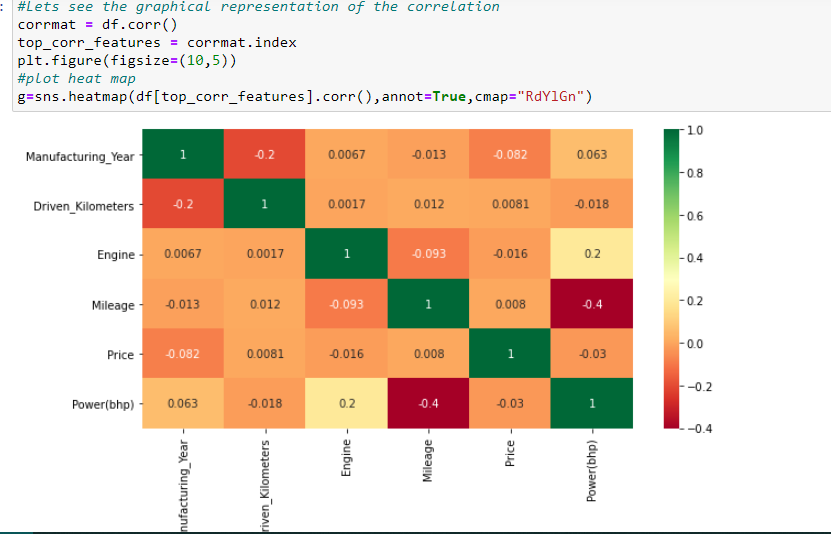


1. **Bar Plot :**

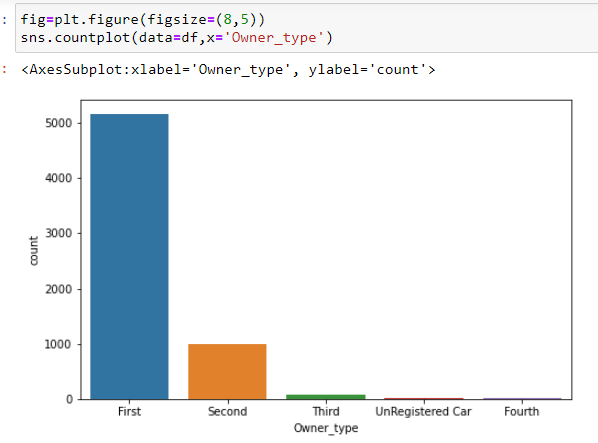
A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories**.**



1. **Heatmap** is used to see the correlation of columns.

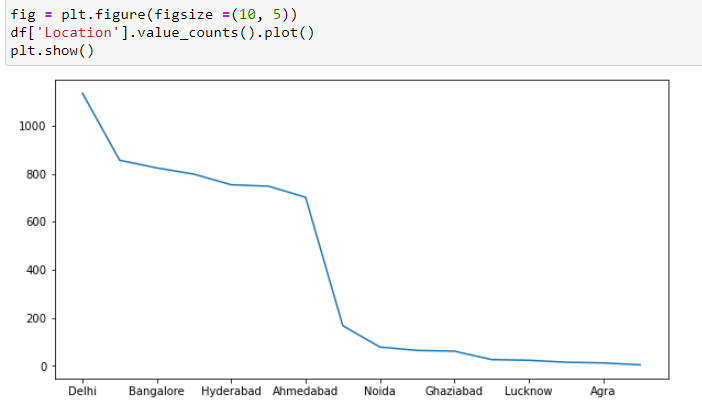


1. **CountPlot -** countplot() method is used to Show the counts of observations in each categorical bin using bars**.**



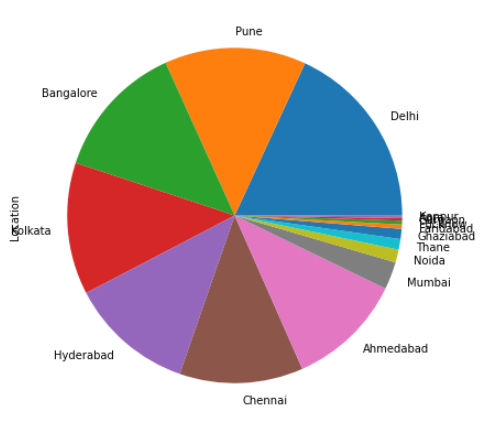
1. **Line plot :**

A line graph—also known as a line plot or a line chart—is a graph that uses lines to connect individual data points. A line graph displays quantitative values over a specified time interval.



1. **Pie Chart :**

Pie Charts help show proportions and percentages between categories, by dividing a circle into proportional segments. Each arc length represents a proportion of each category, while the full circle represents the total sum of all the data, equal to 100%.



* Interpretation of the Results

The data is collected online from the car website. Then I check the missing values and represent through the heatmap. After that I check the describe method check the mean, median , Standard deviation, max values which shows that the there are outliers present and the data is not normally distributed as the data is not complete and some features were not included, which made the right or left skewed data somewhere. Hence we use a log function to make the feature Normally Distributed. And also we use the zscore method to remove some outliers which are very-very far from the normal distribution. To see the inter relation of features with the target column we use the correlation method, and then visualize with the help of **Heatmap** to see the correlation between features and with the target column.

As the dataset was not in the proper data type , First I convert the object datatype numerical column to the Integer and Float data type columns. And the I drop two columns which were not so important. After that I fill the missing values with the help of fillna method and store the mean value of the column and mode value of categoricial column. After that I use the statistical method which is describe method, this tells about that the Price column is not normally distributed and it has the positively right skewed data. Which I solve with the help of the log transformation. And to check distribution in each columns I use **Pie chart , line chart , bar graph , count plot, Boxplot, distribution plo**t for different columns.

And then to check the relationship between the Target column i.e price and other columns I used the **scatter plots, bar plots and line plot** for the better understanding and meaningfulness.

And then I did the Feature selection process where I convert the categorical column into the numerical with the help of one hot encoding and used the statistical libraries like skew for checking the skewness and norm for the normal distribution and zscore method for the outliers removal. And to check the outliers I used the **Boxplot** which shows the outliers present in some columns.

Here I use the z-score with the threshold value of 3 to check the outliers present in the dataset.

And finally the dataset get divided into x(without target column) and y(only target column present) for the model training.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The prediction of Car price we find out that the Price is Correlated with the -

1. Manufacturing Year(The year or edition of the model)

2. Power(in bhp) of the car and

3. Engine of the car

That Means if the manufacturing year of the car is recent its price is will be high, i.e. the newer the car the higher will be the prices.

The Power of the Lamborgini is highest among the all, while the Maserati is the expensive car producer.

Second the Power and engine of the car are also important for buying the used car.

Expensive company cars are sold before being driven much, while Mahindra-Renault cars are sold after driving many kilometers.

Maruti and hyundai car brands has the most number of selling cars in India.

In this project we use different models like Linear Regression, Lasso, Ridge, SVR, KNN Regression and Random Forest Regression.

Here we select the RandomForestRegressor model for our final model training and testing as it gives the very less root mean squared error value and also its R2 score is highest among the all models we choose.

* Learning Outcomes of the Study in respect of Data Science

In this project , I came to know about the car price sell market working, which factors are important for the buying and selling of the car and with the help of the visualization tool like matplotlib, seaborn and plotly , it get easier for the understanding.

The study shows different-different regression algorithms when predicting Car prices in India .The results were good for the data due to it being with features and having correlation.

Hence, the data needs more features to be added preferably with a strong correlation with the car price. However, Random Forest gave the best RMSE score, got the best R2 score overall. The final results of this study showed that Random Forest makes better prediction compared to other used algorithms.

* *Which variables are important to predict the price of variable?*

🡪 The manufacturing year of the car, the engine and the power of the cars are very important as they were highly correlated with the price. That means if the manufacturing year of the car is recently the car price will be high, similarly the more the power, the more the cost for e.g. Lamborghini.

* **Limitations of this work and Scope for Future Work**

Future work on this study could be divided into five main areas to improve the result even further. Which can be done by:

* The used pre-processing methods do help in the prediction R2score. However, experimenting with different combinations of pre-processing methods to achieve less Root Mean squared error and Good R2 score.
* Make use of the available features and if they could be combined as binning features has shown that the data got improved.
* Training the datasets with different regression methods such as XGBoost and Catboost. In order to expand the comparison and check the performance.
* The correlation has shown the association in the data. Thus, attempting to enhance the data is required to make rich with features that vary and can provide a strong correlation relationship.
* The factors that have been studied in this study has a weak correlation with the price. Hence, by adding more factors to the local dataset that affect the car color, such as body type,number of seats, etc. In order to increase the number of factors that have an impact on car prices.