

**Car Price Prediction**

Submitted by:

R. Rajesh kannan

**ACKNOWLEDGEMENT**

Reference: pandas.pydata.org, seaborn.pydata.org, matplotlib.org

Resource: stackoverflow.com,geeksforgeeks.org

Data Sources are from different car resale website i.e(

Olx, cardekho, Cars24)

Other Resources are Project Use case.

**Introduction**

Business Problem Framing:

Data Sources are web scrape from car resale websites like Olx, cardekho, Cars24.

Problem: 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 model..

Conceptual Background of the Domain Problem:

It is similar to analysis problem with model building because the input data shows condition of car with help of that we have to predict price of that car. Now some cars are in demand hence making them costly and some are not in demand hence cheaper.

Motivation for the Problem Undertaken:

Aim of this project is to predict price of the car from the data which get from car resale websites. To get price for respective car, we have to get the important features like brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location.

**Analytical Problem Framing**

Mathematical/ Analytical Modeling of the Problem:

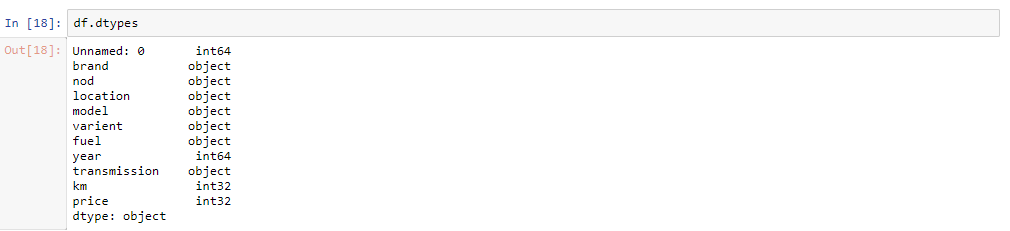
Here some the data are integer data and some are object data.

For analytical purpose I used visualization to analysis the data. Here I used matplotlib.pyplot, seaborn and simple plot function in pandas.

Data Sources and their formats:

Data Sources: From Indian car resale websites.

Formats:



Every data type is object data type except the year, km, price data.

Necessary:

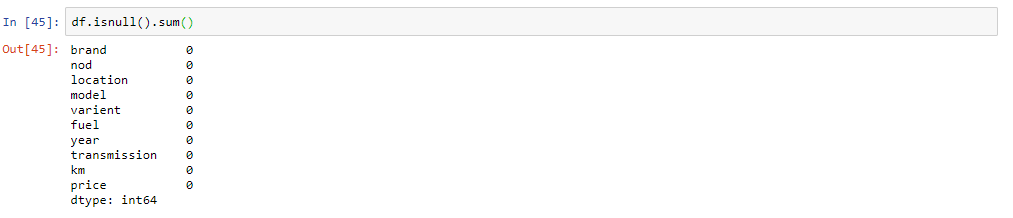
Basically, we need these columns-

Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location etc.

Here all the categorical, integer and ordinal data are necessary, we can treat that necessary variable as our requirement. If the data have more than 60% is null then we have to drop that, if there replacement value for null value in description then treat by replacement, but here there is no presence of null value because I remove null values in scraping part itself.

Data Preprocessing:

There is no presence of null values in the given data set.



While scraping the date in one go without apply exception there may some missing data but I treat there.

Categorical data will not contain outliers and also there will be no presence of skewed in data here all values are scraped perfectly.

Data Inputs and Output Logic Relationships:

There is presence of output features in scraped dataset.

we can predict it by certain models. Output data will be predict by regression models i.e here we have to predict car price .

Visualization techniques like uni-variant visualization, bi varrient and multi variant can also perform to see relationship of input features and also to visualize its % of occupation in total data.

To find:

Apply preprocessing skills to get the data clean by remove the junks like punctuation and text etc. and the train and predict the model.

Hardware and Software Requirements and Tools Used:

Hardware: i5 processor, 8 GB RAM.

Software: OS(windows),

Tools: Jupiter Notebook or Py charm

Libraries: numpy, pandas, matplotlib, seaborn, sklearn ,selenium.

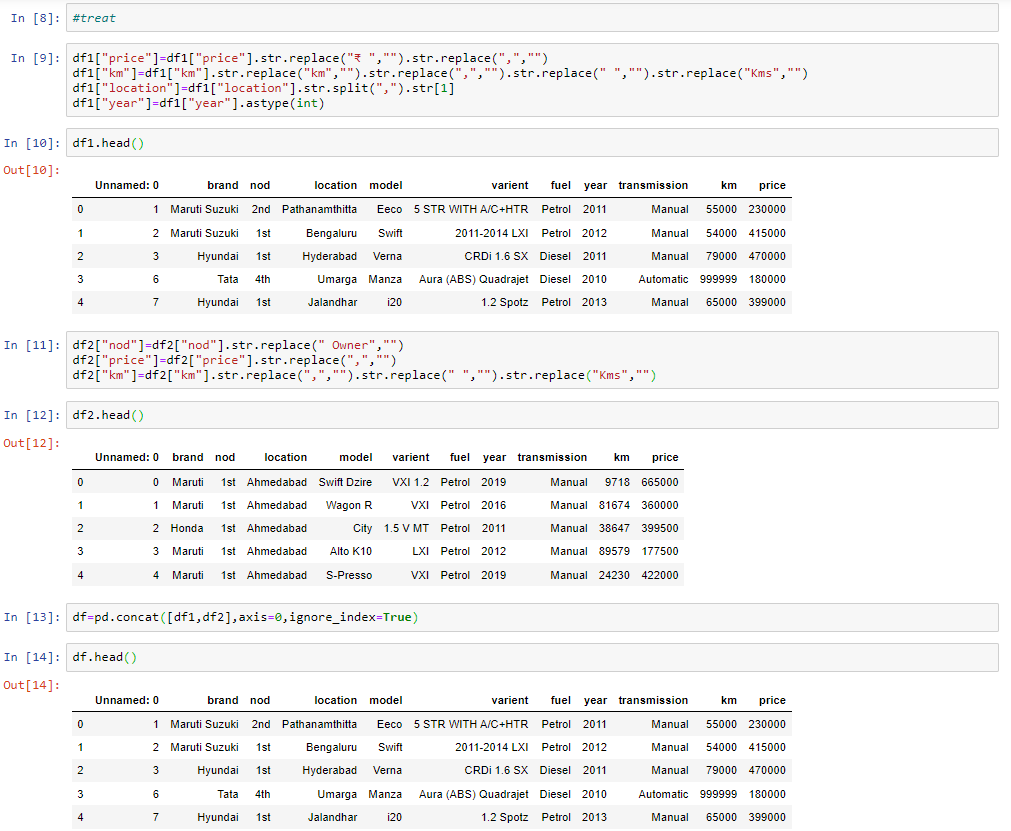
Packages: Pyplot, metrics, model\_selection, and respective model packages.

**Analysis by Visualizations:**

**Here by visualize data we can able to get which feature is more important than other to predict car price.**

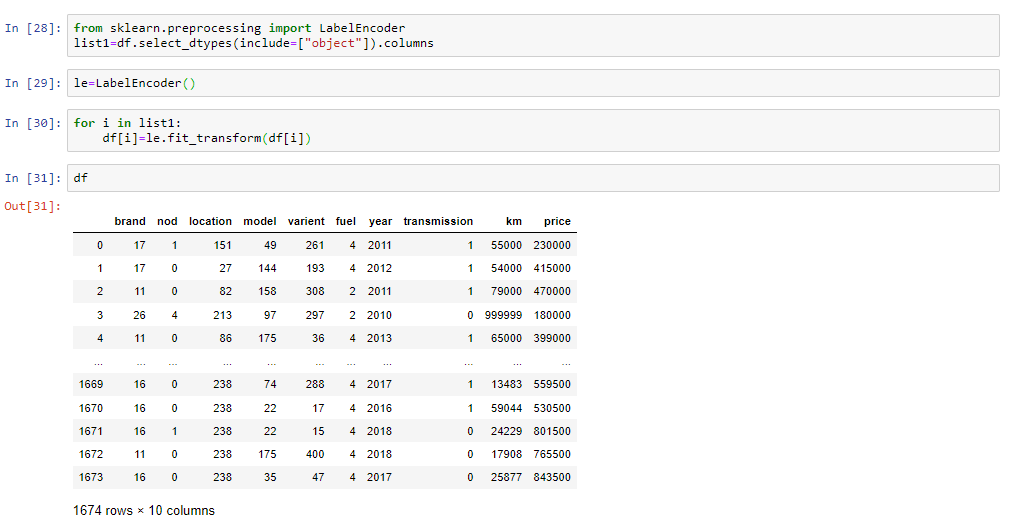
**Pre-Processing Text Data**

Apply preprocessing skills to get the text data clean by remove the junks like punctuation, numbers etc. and the train and predict the model.

****

**Object to integer:**

We can not able to input data as object format so we have to convert it to integer data type.

****

Data Inputs and Output Logic Relationships:

Relationships between target variable and inputs can be find by visualization and correlation of data

Correlation:

If near to 1 is strong +ve correlation

If near to 0 is normal correlation

IF near to -1 is strong –ve correlation

Input Variables which is strong correlated with output variable is the necessary variable.

By use of heat map in seaborn library we came to know the correlation of data between each variables.

Also by a code function df.corr()



Here Year, fuel, transmission, no of owners are highly correlated with price of cars. so these are important variables to predict the Sale Price of car.

So these are important variables to predict the Sale Price

Change in these variables leads to changes in target variable

Visualization techniques like bi-variant visualization can also perform to see relationship between input and output variable.

**Models Development and Evaluation**

Identification of Possible Problem solving approaches:

Model Selection: In Superwised there are classification and regression models. But here output variable have many unique values, so it will come under linear regression model.

Testing of identification Approaches:

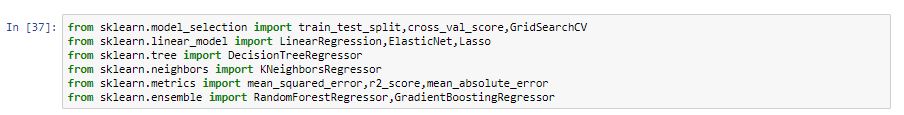
To find the best random state we have to run the for loop by iterating some range of numbers in a model Train Test Split. Random state with high accuracy will be considers as best.

After the selection the best random state we have to find best model by getting good accuracy score, through train and test the data.

Run and Evaluate selected models:

Import the required model through respective packages in sklearn library,

Import the required model through respective packages in sklearn library,



Split the data as input train and test, output train and test,



13.JPG

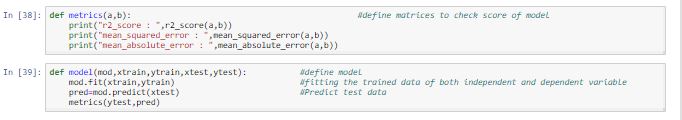
Firstly have to fit the model,

Secondly have to train the input and output data,

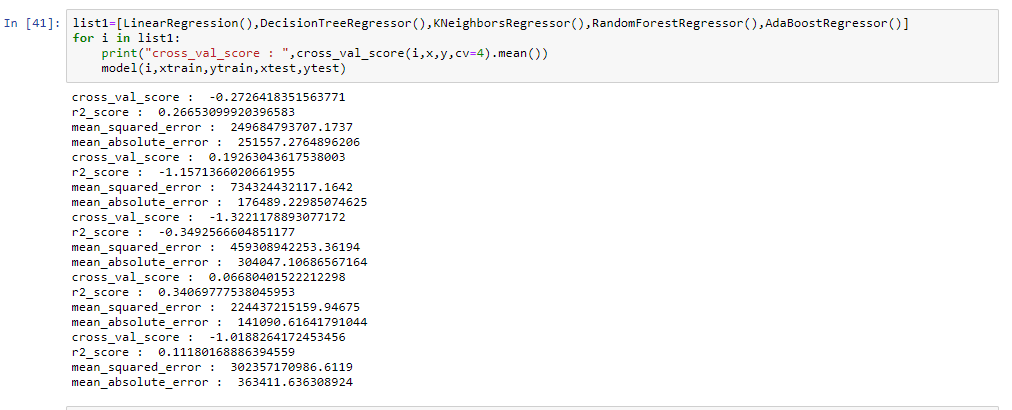
Then have to predict certain test input data,

Finally have to check the accuracy score by compare output test data with predicted output data.

I define 2nd,3rd and final steps in model user define function and metrics in matrix user define function

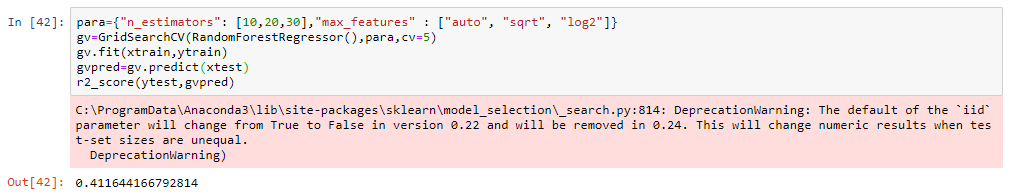


Not only by this process can find the best model because there may occur over fitting due to presence of high bias and high variance in it. So we have to do Cross Validation on it for right result.



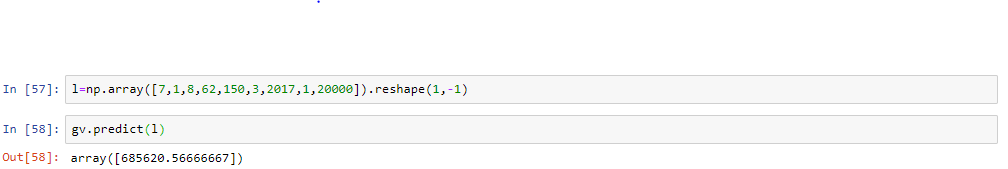
So these models in the list have to iterate through cross validation, model and matrix user define function.

Hyper Parameter Tuning: For to improve the accuracy score for selected best model or all model, we have to apply Grid Search or Randomized search, by apply all the required parameters of respective model inside Grid Search we can get better accuracy.



* Random Forest Regression gives a better result at 41%.
* And also came to know the best parameters .

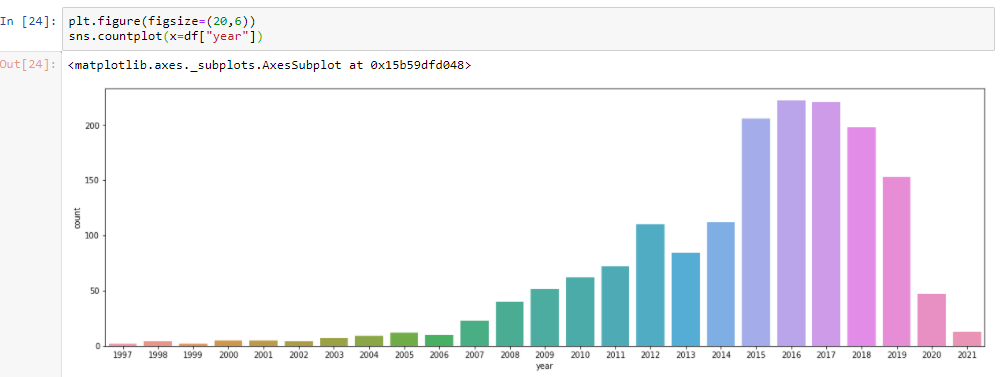
Predict the review data for test input by fit train data in Random Forest Regression model.



**Visualization on user details:**

**Uni variant visualization:**

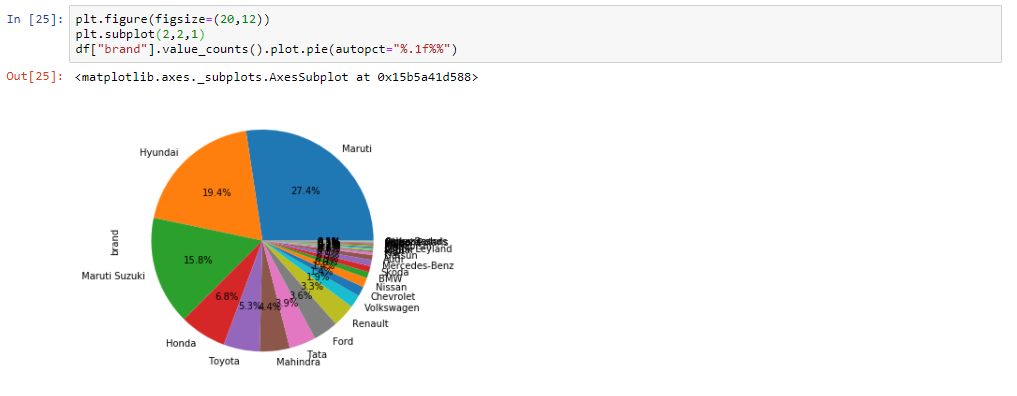
**Count by manufacture year:**



It shows most of the cars are made between year 2010 to 2020 particularly in 2015 to 2019.

The pie graph shows the percentage for brand count.

Count of Product:

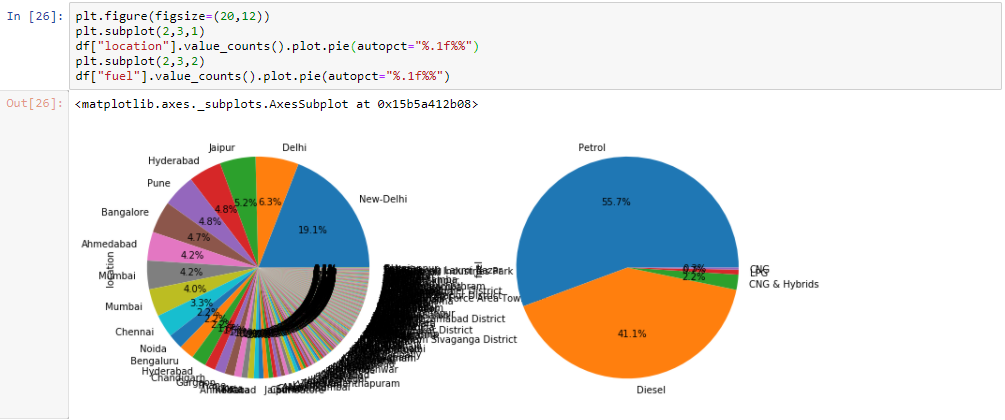


It shows number of count the given brand occur in data set.

The pie graph shows the percentage for different brand count.

Maruti, Hyundai are in maximum count

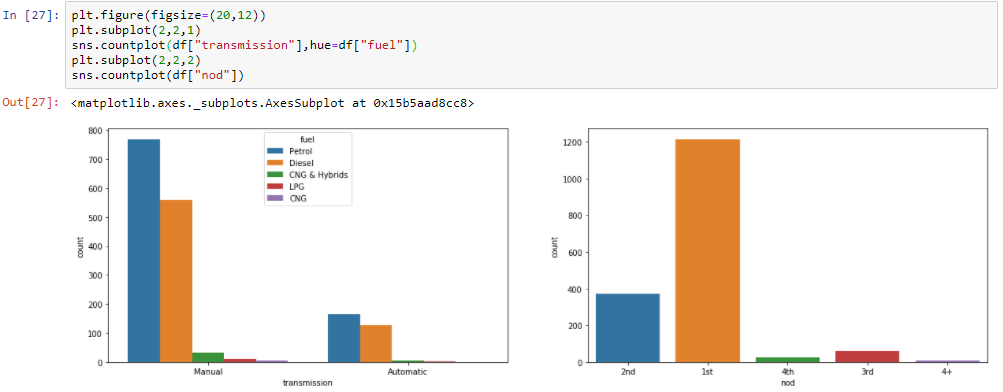
Count of fuel and location:



Most cars are from New Delhi.

Most cars are petrol and Diesel based.

Bi-Varient visualization:



Manual Transmission vehicle use Petrol the most than diesel, For Automatic both are nearly in equal amount.

Most of the Cars have no of owner is 1.

**CONCLUSION**

Key finding: Analysis in review after clean the data and predict price by train the model.

Inferences: From the report it concluded that there are some wrong data. By clean it and prediction was lead to get good model.

Observations: Sale Price of car are mostly depend on important variables (Year of buying, transmission, no of owner, fuel type). Based on it Sales of house get affected.

Learning Outcomes of the study in respect of Data Science

* I learned by visualize also can get important numbers from mixed data and also find how to clean text information.
* Learned to analyse and predict in price features.

Limitations and Future work:

Limitations: Not able to scrape the data in balanced way to get good accuracy

Steps to follow further: Here I scrape all data by one go and also did cleaning in one go , I need to do better model training.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |