

CAR PRICE PREDICTION PROJECT

Submitted by:

Kamran Ashraf

ACKNOWLEDGMENT

I would like to thank FlipRobo Technologies for giving me the opportunity to work on this project. I am very grateful to DataTrained team for providing me the knowledge which helped me a lot to work on this project.

Reference sources are:

1. Google
2. GitHub
3. DataTrained Notes

**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.

* + Review of Literature

The aim of this project is to build a model which can be used to predict the car price. This project is more about data exploration, finding the better insights from data and using the skills and techniques to build an efficient model which can predict the prices after the recession faced during Covid-19 crisis. Since we scrape a good amount of data that are related to the price of car, we can do better data exploration and derive some interesting features using the available columns.

## Analytical Problem Framing

* + Mathematical/ Analytical Modelling of the Problem

In this project we have worked to predict the price of the used car. ‘Price’ is our target column and it is continuous in nature, so we’re dealing with regression type of problem.

This project is based on 2 main phases-

1. **Data Collection:**In this section, we need to scrape the data of used cars from websites (Olx, car Dekho, Cars24 etc.). Web scraping is required for this. We have to fetch data for different locations. Generally, the features columns are Brand, model, variant, manufacturing year, driven kilometres, fuel, number of owners, location and at last target variable Price of the car. This data is to give a hint about important variables in used car model. We can make changes to it, we can add or you can remove some columns, it completely depends on the website from which we are fetching the data. We have tried to include all types of cars in our data for example- SUV, Sedans, Coupe, minivan, Hatchback.
2. **Model Building:** After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model. Follow the complete life cycle of data science. Include all the steps like.
   1. Data Cleaning
   2. Exploratory Data Analysis
   3. Data Pre-processing
   4. Model Building
   5. Model Evaluation
   6. Selecting the best model
   * Data Sources and their formats

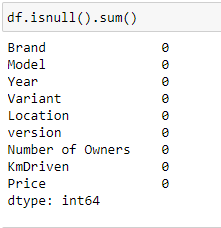
Data sources for our model are online Car Re-sale websites like Car24. We scraped our data using the web scrapping tool Selenium. Let’s have a look at our dataset.



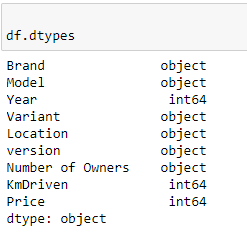
* Data Pre-processing Done

Checking for the null values in the dataset:

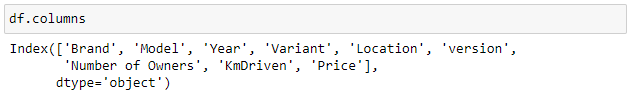
After loading the dataset, we checked for the presence of null values or missing values in the dataset. We found that there were no null values in the dataset.



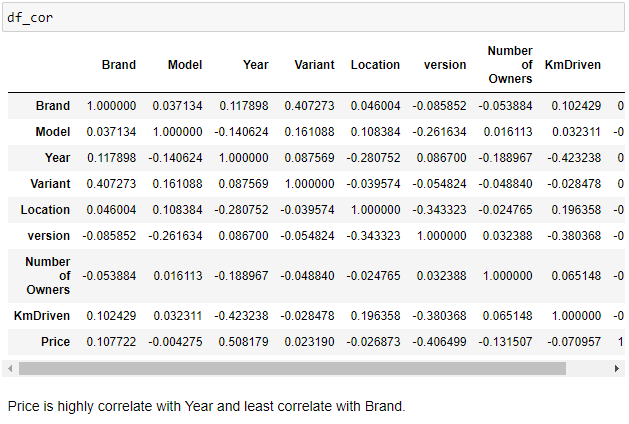
* Let’s have a look at the data formats:

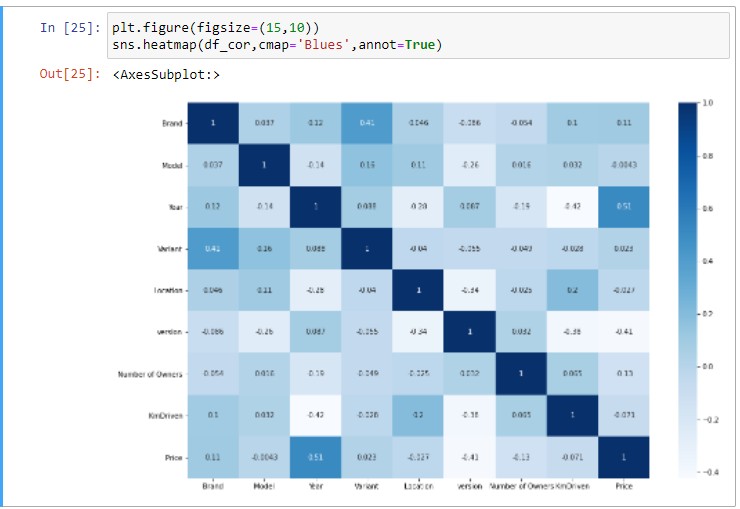


* Now, let’s check the feature columns of the dataset.

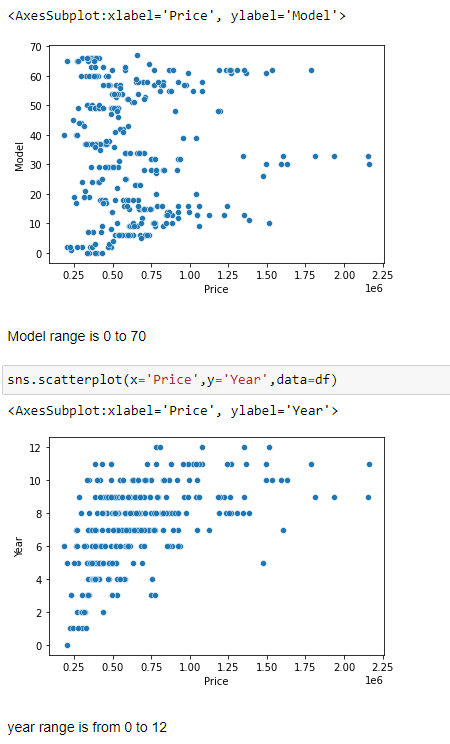


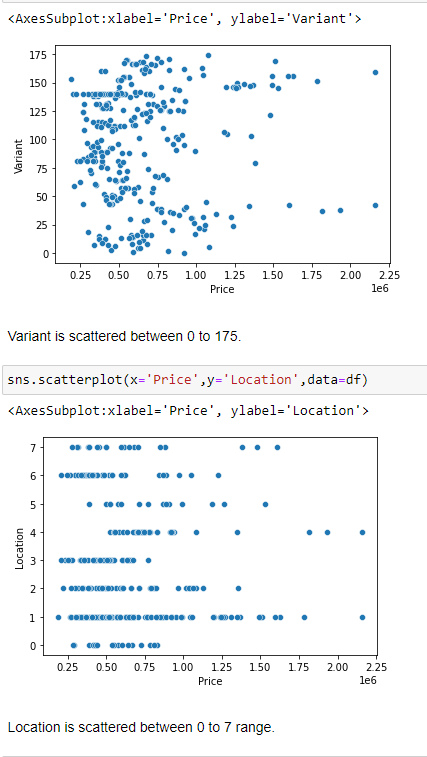
* + Checked for the correlation of the features columns with the target column:

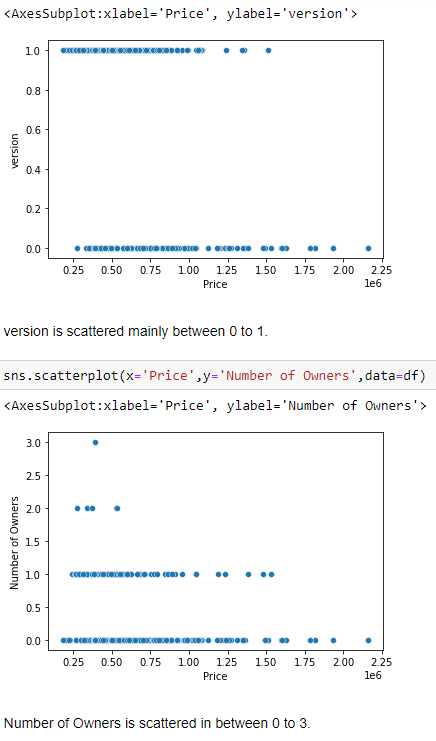


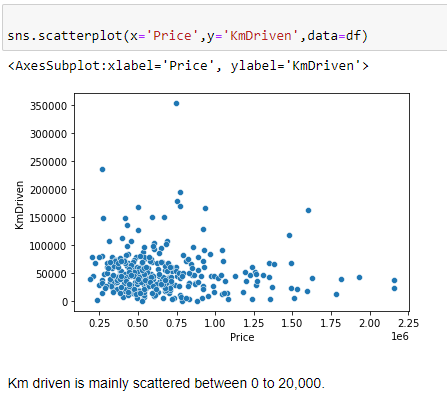


# Data Visualization

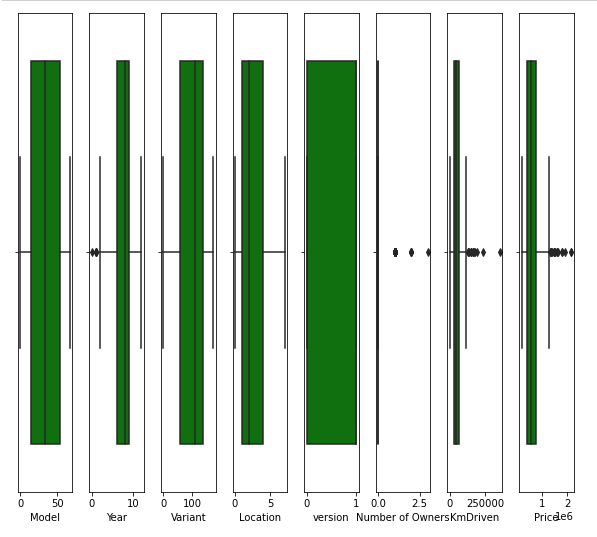




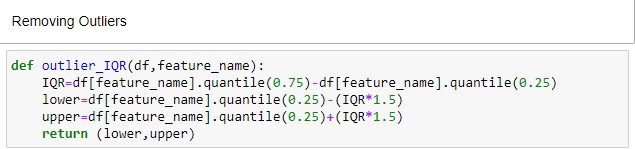




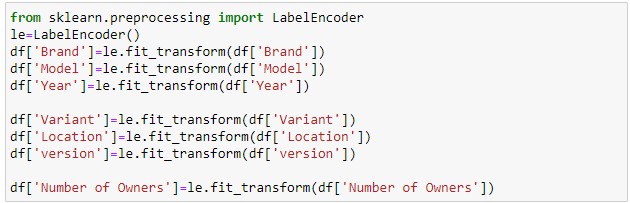
* Checked for the outliers



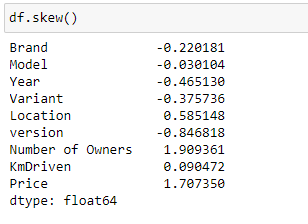
* Removing Outliers



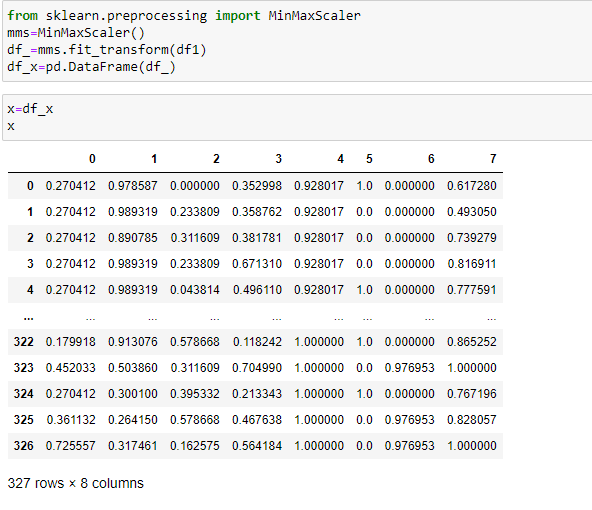
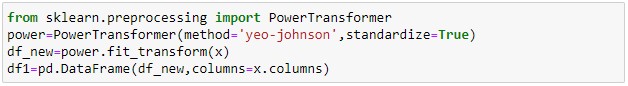
We are using LabelEncoder to convert the categorical data into numerical data.



Skewness in the dataset:



Removing Skewness:



* Hardware and Software Requirements and ToolsUsed.

I have used i3 processor with 4GB RAM as hardware. Software:

Jupyter Notebook (Anaconda 3) Python 3.9 Libraries used:

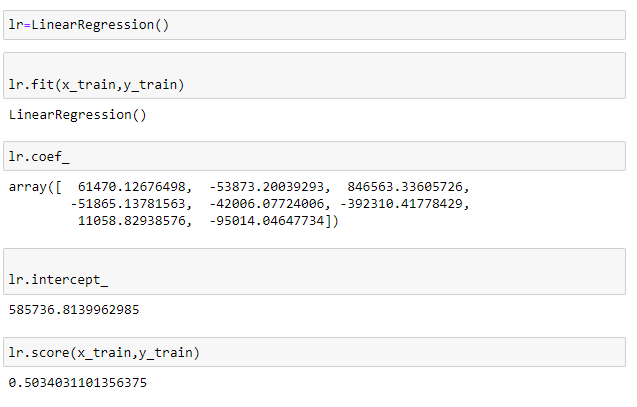
1. Pandas
2. NumPy
3. Matplotlib
4. Seaborn
5. Scipy
6. Selenium

## Model/s Development and Evaluation

* Identification of possible problem-solving approaches (methods).

As we already mentioned above we have continuous target variable so based on the type of variable we are using Regression approach for our model and will use some algorithms like:

Linear Regression Random Forest Regression

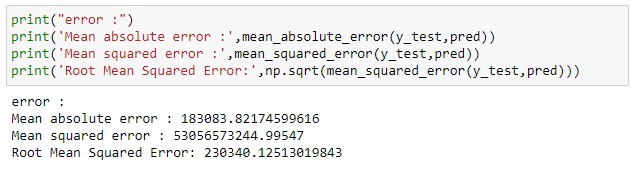


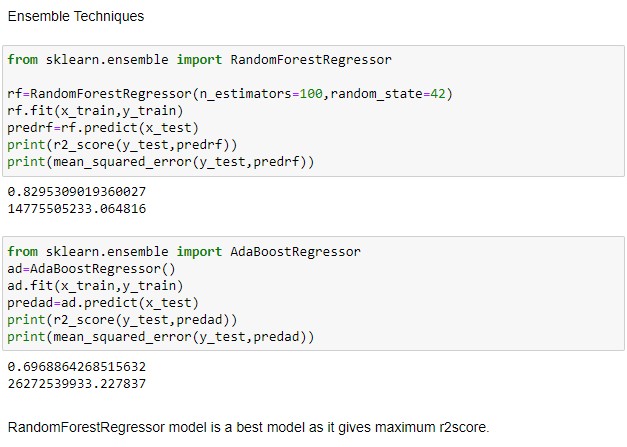
Linear Regression

* Key Metrics for success in solving problem under consideration

In this project we have used the following methods to decide the best model:

* + R2 Score
  + Mean Absolute Error
  + Root Mean Squared Error
  + Cross Validation Score
  + Standard deviation





Ensemble Technique:

### CONCLUSION

Key Findings and Conclusions of the Study

After working on this project ,we got a lot of insights of how to collect data , pre-process it ,remove unwanted data and how to tackle null values to create a good model. In this project we find various factors responsible for the sale price of a car. Let’s take each variable to describe our observations from this project ü Brand:

After analyzing the dataset we draw the outcome that in the Indian market the brands like ‘Maruti’, ‘Hyundai’, ‘Honda’ are covering major portion of the market. If we talk about the valuations of the brands then the premium category brands like Mercedes, Audi, KIA, JEEP are having more value.

Purchase Year:

Most of the cars which are listed for the re-sale belongs to the year 2010 to 2018. There was less number of cars who was listed from the year 2020, this may be due to the effect of COVID-19 pandemic on the automobile sector.

Owner:

We can say that as the number of owners increases the price of the car decreases. This depends on the purchase year of the car also. If the car is recently purchased then the resale value is more.

### Learning Outcomes of the Study in respect of Data Science

By working on this projects, we got to know more about the Indian car market and its resell values. We also come to know about the different features which decide the valuation of the car. This project also given us the exposure to the different types of techniques used in the Machine Learning world specially the Regression type of problems.

* Limitations of this work and Scope for Future Work

This project was a challenging task for me and this also helped me to sharpen my knowledge. After working on the project, I find some difficulties like the website was not very much updated and there were some missing values and outliers in the different feature outcomes.

# Thank you.