

***CAR PRICE PREDICTION***

**Submitted by:**

**Yunus Surty**

# ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I would also like to thank my mentor in Fliprobo, Sapna Verma, for providing me with the problem statement for performing this wonderful task.

Some of the reference sources are as follows:

* Coding Ninjas
* Medium.com
* Analytics Vidhya
* StackOverflow
* Carsdekho.com for data collection

# INTRODUCTION

## BUSINESS PROBLEM FRAMING

The objective of the project was to build the model to predict the price of used cars with the available independent variables. This model can then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market post covid.

## CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM

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

Used cars price depends on various factors. From kilometers driven to condition the car is in. Used cars is a big market as many buyers do prefer to buy the used cars to get the benefit of price.

## MOTIVATION FOR THE PROBLEM UNDERTAKEN

To understand real world problems where Machine Learning and Data Analysis can be applied to help organizations in various domains to make better decisions with the help of which they can gain profit or can be escaped from any loss which otherwise could be possible without the study of data .

# ANALYTICAL PROBLEM FRAMING

## MATHEMATICAL/ ANALYTICAL MODELING OF THE PROBLEM

This is a Regression problem, where our end goal is to predict the Prices of used cars based on data. I have collected the data from carsdekho.com and will be making the model based on the data I have collected.

## DATA SOURCES AND THEIR FORMATS

The data was collected from carsdekho.com in csv format.

Below is the description of the data

**Data Description:**

**1. Brands**

**Name of the company car belongs to**

**2. Model**

**Name of the car model**

**3. Variant**

**Model variant the car has**

**4. Make year**

**The year car was made**

**5. Fuel**

**Type of fuel the car used**

**6. Transmission**

**Whether the car has automatic or manual transmission**

**7. Driven kilometers**

**Number of kilometers the car has run**

**8. Owners**

**Number of owners the car has been sold to**

**9. Location**

**Location from the used car belongs to**

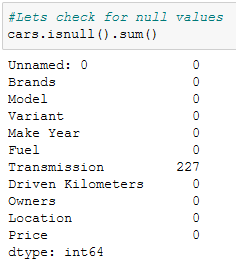
**10. Price**

**This is our traget variable which we needs to predict, the price of the car**

## DATA PREPROCESSING DONE

After loading all the required libraries we loaded the data into our jupyter notebook.

Feature Engineering has been used for cleaning of the data. Some unused columns have been deleted and even some columns have been bifurcated which was used in the prediction. We first done data cleaning. We first looked for the missing values and the same was filled.



We dropped the Null values from the dataset.

Following process was followed to clean the data

1. We need to drop the Unnamed column

2. We will reduce the length of Variant. It will be easy for our machine to understand

3. Remove 'Petrol+' and change LPG to CNG as it is same fuel

4. Treat the Null values in Transmission

5. Remove commas from Driven kilometres and covert it to integer

6. Remove commas from Price and convert it to integer











We created the new column ‘no\_year’. This feature will help us to know how old the car is.

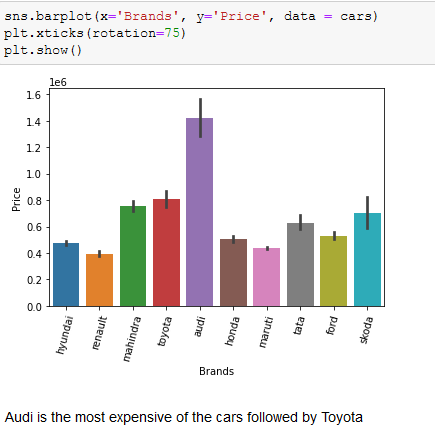




.

## DATA INPUTS- LOGIC- OUTPUT RELATIONSHIPS

Here we check the correlation between all our feature variables with target variable label

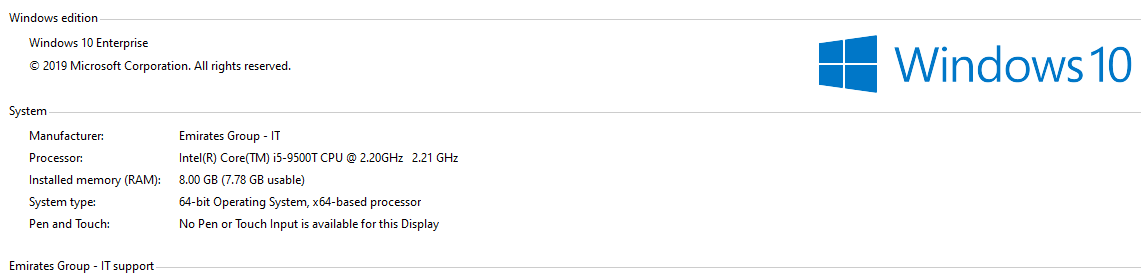


Set of assumptions related to the problem under consideration

By looking into the target variable label we assumed that it was a Regression type of problem.

## HARDWARE AND SOFTWARE REQUIREMENTS AND TOOLS USED

***HARDWARE:***

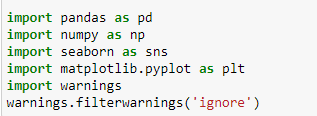


***SOFTWARE:***

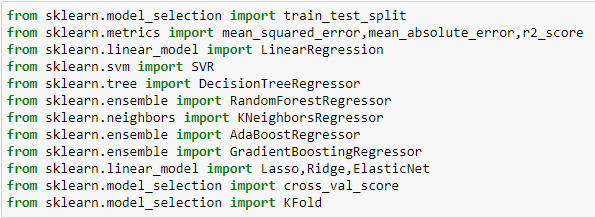
Jupyter Notebook (Anaconda 3) – Python 3.8.5

***LIBRARIES:***

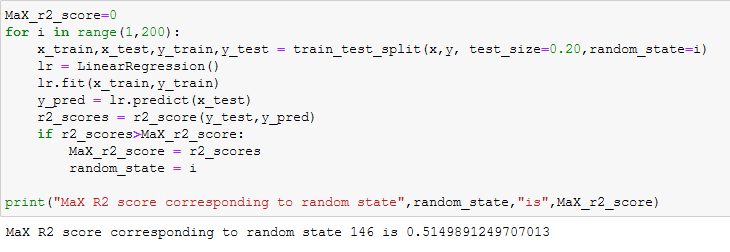
The tools, libraries and packages we used for accomplishing this project are pandas, numpy, matplotlib, seaborn, scipy stats, sklearn.decomposition pca, sklearn standardscaler, GridSearchCV, joblib.







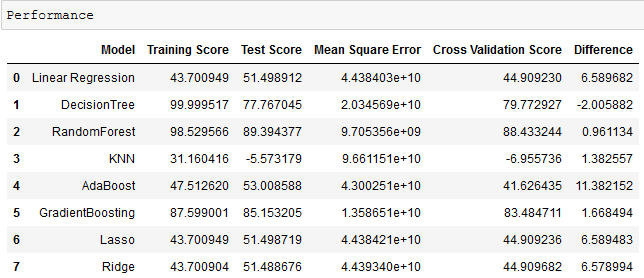
# MODEL/S DEVELOPMENT AND EVALUATION



## RUN AND EVALUATE SELECTED MODELS



## KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

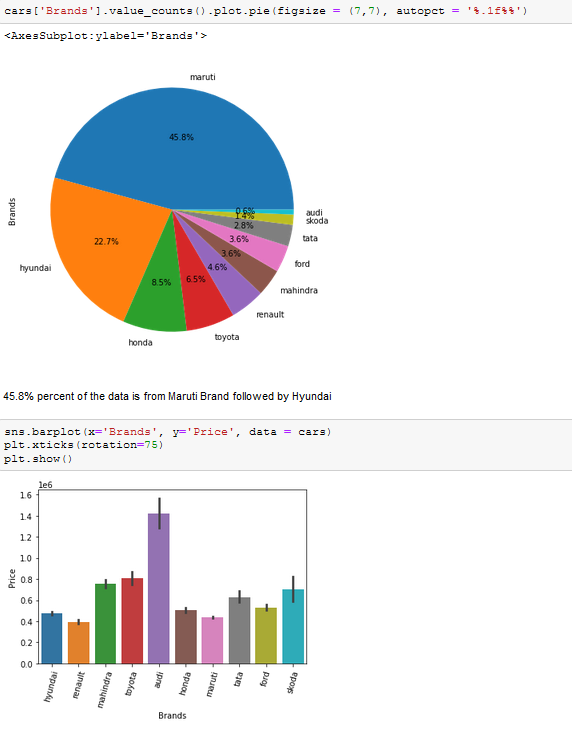


We used the R2 score and difference between test and cross validation to select the model.



## VISUALIZATION

Below technique was used for data visualization

******

## INTERPRETATION OF THE RESULTS



Our model performance increased slightly after Hypertuning.

# CONCLUSION

## KEY FINDINGS AND CONCLUSIONS OF THE STUDY

In this project we have tried to show how the used car prices vary and what are the factors related to the changing of cars prices. The Random forest regressor model has performed well in predicting the prices..

## LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This project has demonstrated the importance of modelling and predicting data.

Through different powerful tools of visualization we were able to analyse and interpret different hidden insights about the data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

## LIMITATIONS OF THIS WORK AND SCOPE FOR FUTURE WORK

As with any project there is room for improvement here. The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project.