**Predicting the price of old car.**

**Project-Based Internship 2020 Report**

Submitted

To

**DataRitz Technologies**

**Duration : Six Weeks**

**By**

**STUDENT NAME (s)**

**Tanya Srivastava**

**Runjhun Khare**

**Swapnil Gupta**

**ABES Engineering College, Ghaziabad**

**APJ Abdul Kalam Technical University**

**Under the guidance of**

**NAME OF MENTOR(s)**

**Gopal Gupta**

**Shashank Shekhar**

**TABLE OF CONTENTS**

Page No.

DECLARATION 1

CERTIFICATE 3

ACKNOWLEDGEMENTS 4

ABSTRACT 5

LIST OF TABLES 8

LIST OF FIGURES 9

**CHAPTER-1INTRODUCTION 10-11**

1.1 Problem Definition 10

1.2 Motivation 10

1.3 Objective of the Project 11

1.4 Scope of the Project 11

1.5 Need of Work 11

**CHAPTER 2 RELATED WORK 12-13**

**CHAPTER 3 PROPOSED METHODOLOGY 14-29**

3.1 Dataset Description 14

3.2 Methods 15

3.3 Hardware / Software Requirements 15

3.3 Our Methodology 15-29

**CHAPTER 4 EXPERIMENT AND RESULT ANALYSIS 30-31**

**CHAPTER 5 CONCLUSION 32**

**5.1 Discussion 32**

**5.2 Future Work 32**

**REFERENCES 33**

## CERTIFICATE

This is to certify that Project Report entitled “Predicting the price of old car” which is submitted by Tanya Srivastava/ Runjhun khare/ Swapnil Gupta in partial fulfillment of the requirement for the summer internship of data analysis and machine learning using python in Department of IT of ABES Engineering college, **AKTU**, is a record of the candidate own work carried out by her under my/our supervision.

**Supervisor 1: Mr. Gopal Gupta**

**Supervisor 2:** **Mr. Shashank Shekhar**

**Date: 08/07/2020**

ACKNOWLEDGEMENT

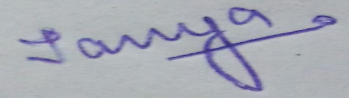
*It gives us a great sense of pleasure to present the report of the Project Based Internship 2020 undertaken during my 2 year and IT branch. We owe special debt of gratitude to “*Mr. Gopal Gupta/ Mr. Shashank Shekhar (Lead Technical Architect/ Project Consultant)” *DataRitz Technologies for their constant support and guidance throughout the course of our work. His constant motivation have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.*

*We also take the opportunity to acknowledge the contribution of team members of DataRitz Technologies for their full support and assistance during the development of the project.*

*We also do not like to miss the opportunity to acknowledge the motivation of it IT department and ABES Engineering college to provide us the opportunity to undergo training at DataRitz Technologies.*

*Signature:*

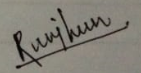
*Name :* Tanya Srivastava

**

*Roll No.: 1803231172*

*Date : 08/07/2020*

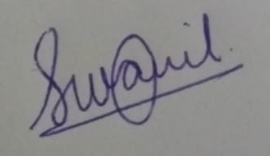
*Name: Runjhun Khare*

**

*Roll No.: 1803210125*

*Date: 08/07/2020*

*Name: Swapnil Gupta*

**

*Roll No.: 1803210156*

*Date: 08/07/2020*

**ABSTRACT**

* This project is to analyze the reusability of old cars, predict the price of the car and provide recommendation for related vehicles with a lower miles, lower price over one that is slightly more expensive using techniques of data analysis and algorithms of machine learning.
* In this project , we have modified the dataset and visualize it according to the

Techniques of data analysis .It has predicted few observations based on graphs and charts . We have used machine learning models like **linear regression.**

* For this project, Linear Regression yielded 67% accuracy for predicting the price of used cars.
* Furthermore, the project showcases that Car Power Engine, Car Mileage, and the Age of the car are the three most important factors in predicting the price of a used car.
* Future work on this project would include predicting the price of each brand and model of the car collections in this dataset. It would be interesting to learn how the inclusion of pictures into the dataset can influence predicting the price of a used car.
* **Keywords**: Data mining, machine learning algorithms

Project Summary

|  |  |
| --- | --- |
| 1. **Region/Unit** | 1. DataRitz Technologies |
| 1. **Location** | 1. Ghaziabad |
| 1. **Program** | 1. DataRitz Technologies.<<programcode>>.<<version>> |
| 1. **Project Number** | 1. DataRitz Technologies. <<projectcode>>.<<version>> |
| 1. **Project Description** | 1. <<project description>> |

Document Control

|  |  |
| --- | --- |
| 1. Prepared by: | 1. Tanya Srivastava/ Runjhun khare/ Swapnil Gupta |
| 1. Title: | Predicting prices of old car |
| 1. College: | ABES Engineering College |
| 1. Department: | 1. IT |
| 1. Location: | 1. Ghaziabad |
| 1. Version date: | 1. <<date of document completion>> |
| 1. Status: | 1. **<<Initial Draft**/Consultation Draft/Approved Document/Minor Revision/Major Revision>> |

Version history

|  |  |  |  |
| --- | --- | --- | --- |
| **Version no.** | **Date** | **Changed by** | **Nature of amendment** |
| 1. **1.0** |  |  |  |
| 1. **1.1** |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Endorsement and Approval

Project Customer

I approve the business requirements specifications in this document.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | <<customer name>> | | |
| 1. Position | <<customer position>> | | |
| 1. Signature |  | 1. Date |  |

The following officers have **endorsed** this document

Project Sponsor

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. <<sponsor name>> | | |
| 1. Position | 1. <<sponsor position>> | | |
| 1. Signature |  | 1. Date |  |

Project Manager (= Component Project Customer)

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. Mr. Gopal Gupta/ Mr. Shashank Shekar | | |
| 1. Position | 1. Lead Technical Architect/ Project Consultant | | |
| 1. Signature |  | 1. Date |  |

Component Project Sponsor

I accept the business requirements specifications in this document.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. Dr B P Sharma | | |
| 1. Position | 1. Country Head – Delivery | | |
| 1. Signature |  | 1. Date |  |
| 1. **Comments** | | | |
|  | | | |

The following officers have **endorsed** this document

Component Program Manager

|  |  |  |  |
| --- | --- | --- | --- |
| 1. Name | 1. Mr. Gaurav Kansal | | |
| 1. Position | 1. Chief Operating Officer | | |
| 1. Signature |  | 1. Date |  |

**LIST OF TABLES**

* Project summary 6
* Document control 6
* Version history 6
* Endorsement and approval 7

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Fig.3.1 | Dataset Description | 15 |
| Fig.3.2 | Renaming column | 17 |
| Fig.3.3 | Identifying and handling missing values | 17 |
| Fig.3.4 | Encoding | 18 |
| Fig.3.5 | Data Visualization | 19 |
| Fig.3.6 | Pie Chart | 19 |
| Fig.3.7 | Bivariate analysis | 20 |
| Fig.3.8 | HeatMap | 21 |
| Fig.3.9 | Regression model flow chart | 21 |
| Fig.3.10 | DataSet Splitting | 22 |
| Fig.3.11 | Linear Regresssion Code | 23 |
| Fig.3.12 | Multiple linear calculations | 23 |
| Fig.3.13 | Graph for Multiple Linear | 24 |
| Fig.3.14 | Decision Tree Model | 25 |
| Fig.3.15 | Calculations of Decision tree | 26 |
| Fig.3.16 | Code For Graph | 25 |
| Fig.3.17 | Graph for Decision Tree | 26 |
| Fig.3.18 | Random forest Model | 27 |
| Fig.3.19 | Calculations of Random forest | 28 |
| Fig.3.20 | Code For Random forest | 29 |
| Fig.3.21 | Code for the Graph | 29 |
| Fig3.22 | Graph for Random forest | 30 |
| Fig.3.23 | GUI | 30 |

**CHAPTER 1**

**INTRODUCTION**

* 1. **: Problem Definition**

The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. It is important to know their actual market value while both buying and selling**.**

* 1. **Motivation**
* With the growing economic conditions, the sale of used cars has been increased thus their is a need to design a model.
* Deciding whether a used car is worth the posted price when you see listings online can be difficult, as value of used cars depends on factors like model, manufacturer, mileage, fuel type.
* From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to develop model for predicting used car prices efficiently**.** 
  1. **Objective of the Project:** 
     1. The objective of our project is to predict the price of the old cars using various attributes and features. Some of the expected factors are model name, manufacturer, year of manufacturing, number of cylinders and fuel type .
     2. Additionally, the model will provide recommendation for related vehicles with lower miles, lower price, and one that is slightly more expensive.
     3. We will model the price of cars with the available variables so that it can be used by the management to understand how exactly the prices vary. They can accordingly manipulate the design of the cars, the business strategy etc.
  2. **Scope of the Project:**
  + The purpose of the project is to analysis the expenditure for reusability of old cars.
  + Our project has used dataset from **kaggle** and cleaned the data(removal of unnecessary values) using data cleaning process
  + We have Analyzed various attributes of dataset through data visualization process.
  + This project has used machine learning model or algorithms to predict the prices of old car.
  1. **Need of Work**
  + This project helps to understand to move ahead and to understand what is important for future plan success. It will help with business expenditures

**CHAPTER 2**

**RELATED WORK**

Predicting price of a used cars has been studied extensively in various researches. Listian discussed, in her paper written for Master thesis , that regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression or some simple multiple regression. This is on the grounds that Support Vector Machine (SVM) is better in dealing with datasets with more dimensions and it is less prone to overfitting and underfitting. The weakness of this research is that a change of simple regression with more advanced SVM regression was not shown in basic indicators like mean, variance or standard deviation. Another approach was given by Richardson in his thesis work . His theory was that car producers produce more durable cars. Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time than TEM Journal. This has roots in environmental concerns about the climate and it gives higher fuel efficiency. Wu et al. conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbor machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it . Gonggie proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models. Furthermore, Pudaruth applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior color, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%. Noor and Jan build a model for car price prediction by using multiple linear regression. The dataset was created during the two-months period and included the following features: price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage in kilometer, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model year and model as input features. With the given setup authors were able to achieve prediction accuracy of 98%. In the related work shown above, authors proposed prediction model based on the single machine learning algorithm. However, it is noticeable that single machine learning algorithm approach did not give remarkable prediction results and could be enhanced by assembling various machine learning methods in an ensemble.

**CHAPTER 3**

**PROPOSED METHODOLOGY**

* 1. **Dataset Description**

Our dataset is taken from:

<https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv>”

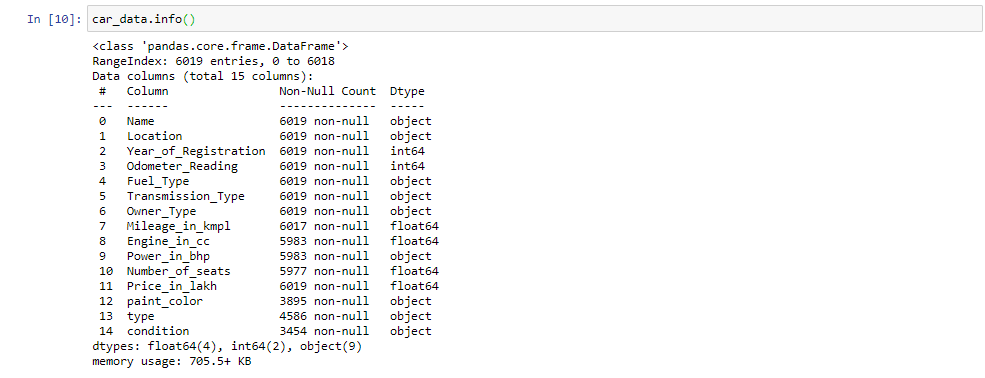


Fig.3.1: Dataset Description

**Attributes Description :**

* **Name -** Name of the car
* **location –** the location or city where the car is currently in.
* **Year\_of\_registration -** At what year the car was first registered - the age of the car..
* **Odometer\_Reading -** the total distance travelled by the car since it was first registered.
* **Fuel\_type -** compressed natural gas(CNG), diesel, petrol, LPG, electric
* **Transmission\_Type –** automatic or manual.
* **Owner\_Type** – first , second, third or above.
* **Mileage\_in\_kmpl -** mileage of the car in kilometre per litre
* **Engine\_in\_cc -**   Cylinder capacity or the cubic centimeters capacity of the combustion cylinder.
* **Power\_in\_bhp -** Car Engine Power in brake horse power.
* **Seats -** describes the number of seats in the car.
* **Price**  - the price in Rs on the advertisement to sell the car.
* **Paint\_color** - Represents the color variants available of the car.
* **Type** – car body type – suv, wegan, sedan, truck, convertible, other etc.
* **Condition** – specify the condition whether it is new, excellent, good, fair etc.

**3.3 Methods**

* This project has used different python packages likes pandas, numpy , seaborn, myplotlib ,etc to apply data analysis and machine learning .
* We have cleaned the dataset in ordered to visualize the dataset better.
* Our second aim was to make appropriate graphs and charts to analyze the dataset.
* Then we have applied machine learning approach to predict the prices of old cars.

**3.3 Hardware / Software Requirements**

**HARDWARE:** I5 processor, 4 GB ram

**SOFTWARE:** Anconda, jupiter notebook,github.

**3.4 Our Methodology:**

**Data preprocessing:**

* Acquiring ,importing and reading of dataset
* Changing the Data Attributes Name:
* If it is required to change the names of the attribute if they are inappropriate then we need to change the names of our attributes.

****

Fig.3.2: Renaming of columns

**Identifying and handling missing values :**  
 In data preprocessing, it is pivotal to identify and correctly handle the missing values, falling to do this, you might draw inaccurate and faulty conclusions and inferences from the data.

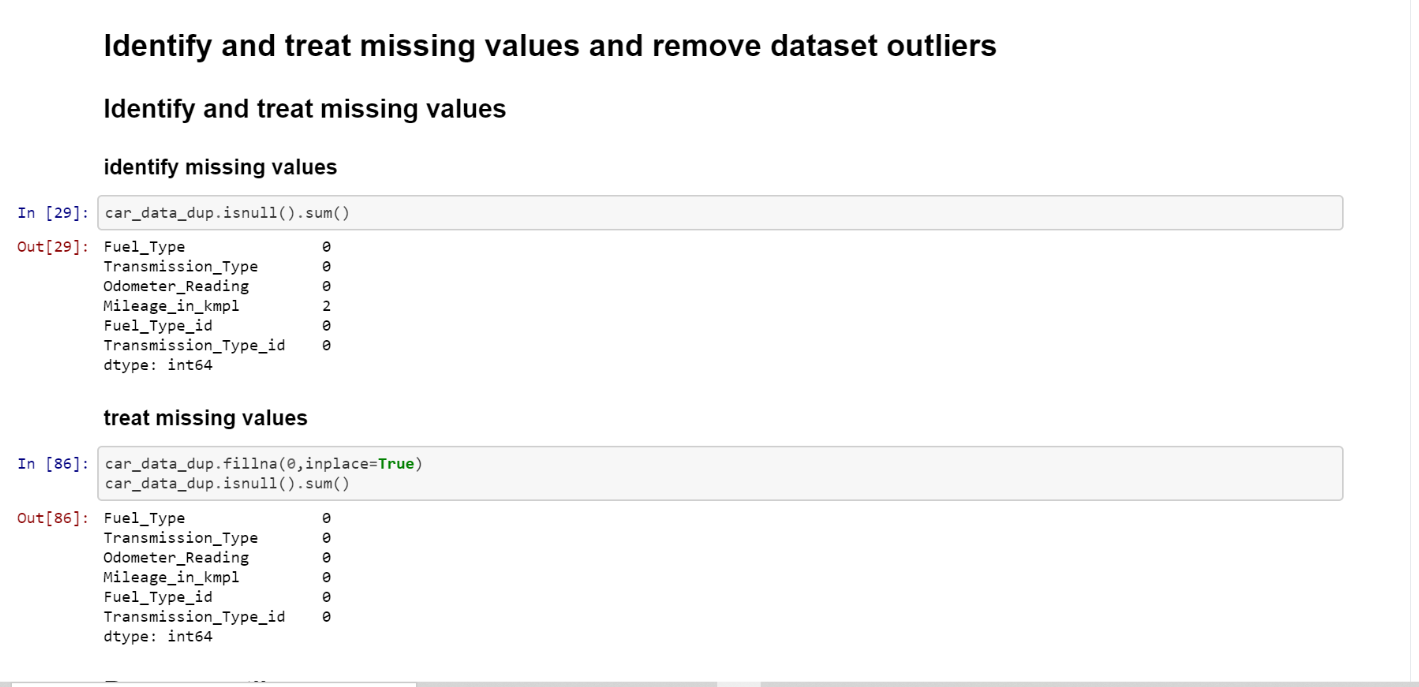


Fig.3.3: Identifying and Handling missing values

**Encoding :**

Categorical data refers to the information that has specific categories within the dataset.

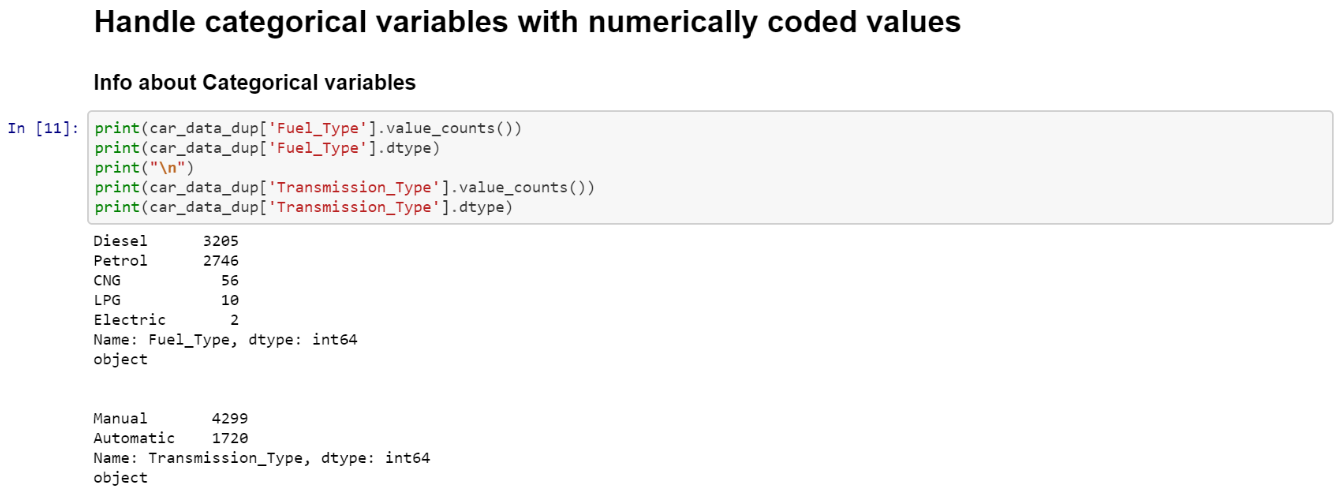
****

Fig.3.4: Encoding

* For Fuel\_Type, create new column "Fuel\_Type\_id" which handles categorical data as numerical data.
* 0 <- CNG 1 <- Diesel 2 <- Electric 3 <- LPG 4 <- Petrol
* 2. For Transmission\_Type, create new column "Transmission\_Type\_id" which handles categorical data as numerical data.
* 0 <- Automatic 1 <- Manual

**Data Visualisation:**Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. Few of the observation we made are listed below

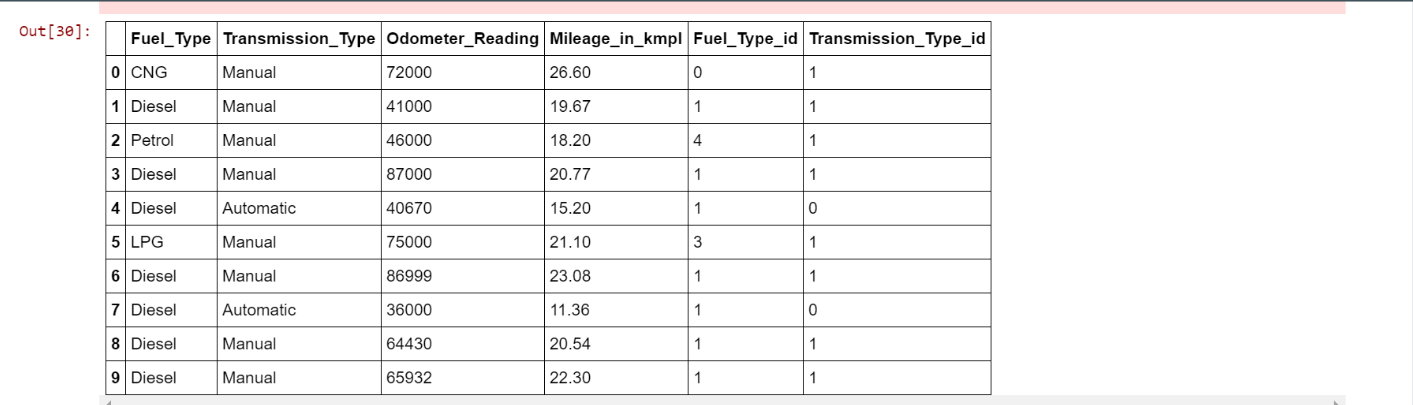
****

Fig.3.5: Data Visualization

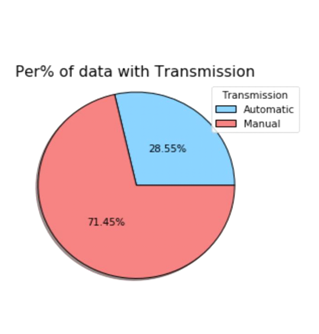
****

Fig.3.6.Pie Chart of Transmission

* As we can see from figure, in our dataset, in attributes Transmission most of car are manual. It implies people do not prefer automatic cars.

**Bivariate Analysis with respect o miles in kmpl**.

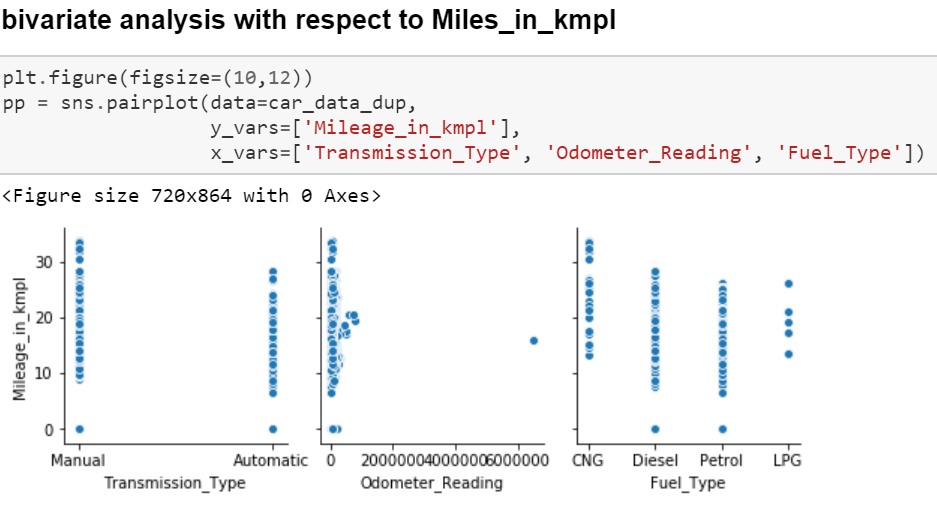
****

Fig.3.7.Bivariate Analysis

* In this graph for manual it ranges from 10 to 30 and for automatic transmission type it ranges from 10 but is more than 20 but less than 30.

**Correlation matrix**

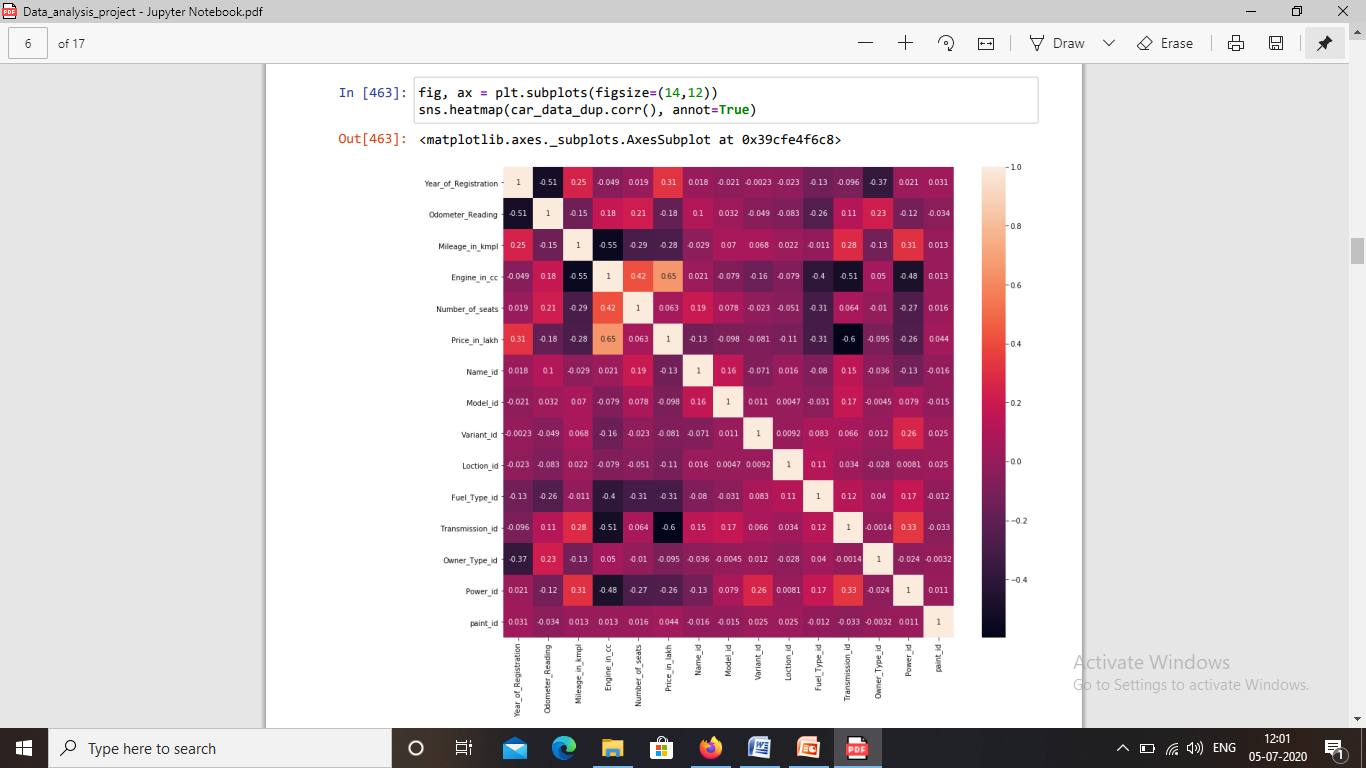


Fig.3.8: Heatmap

**MACHINE LEARNING ALGORITHMS :**

**Regression Analysis:**

* Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).
* This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.
* Here in our model we are using four types of regression models:
* Multiple Linear Regression
* Decision Tree Regression
* Random Forest Algorithm
* To perform regression analysis we first need to separate the dataset into input and target variables.
* Split the dataset into test and train.
* In our model we have splitted them into test size of 0.3 and random state of 42**.**

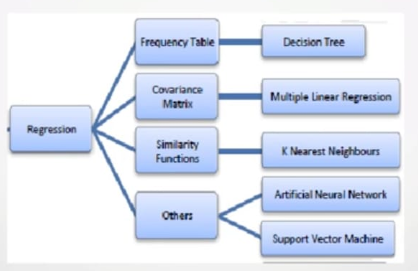
****

Fig.3.9: Regression Model Flowchart

Splitting of Data set into two parts test data and train data.

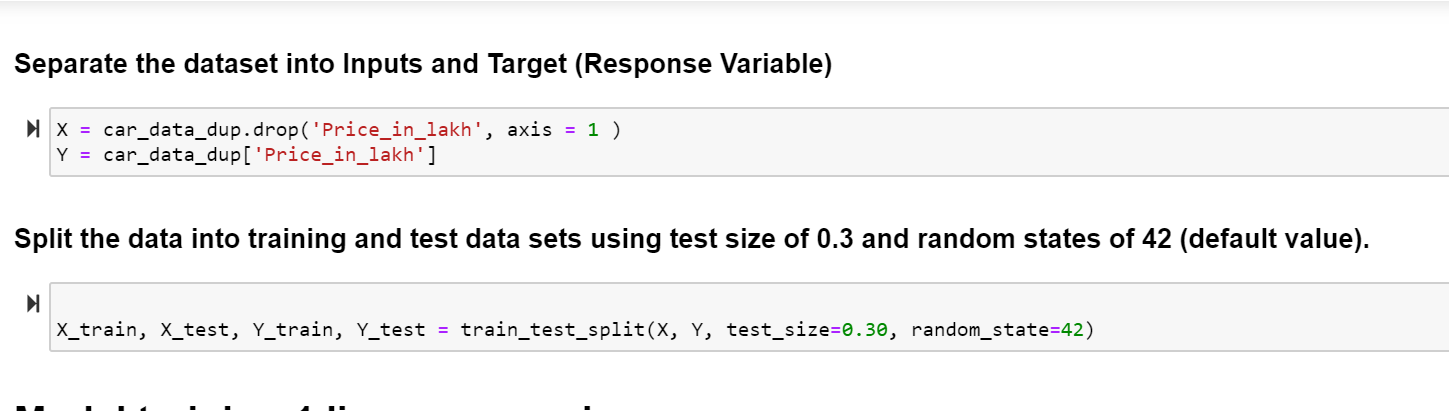
****

Fig.10: Splitting of Dataset

**Multiple Linear Regression:**

* Multiple Linear Regression is the most common form of linear regression analysis.
* As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.
* The independent variables can be continuous or categorical (dummy coded as appropriate).

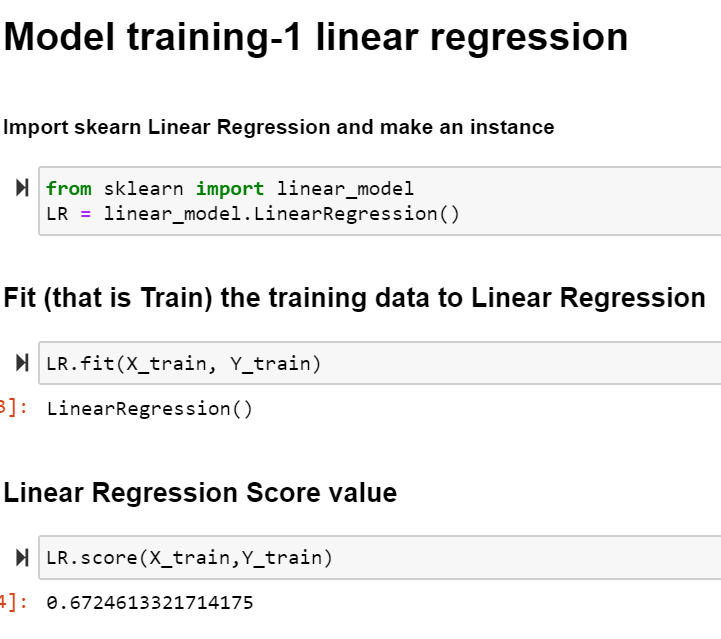
****

Fig.3.10 Linear Regresssion Code

* Separate the dataset into input and target variables.
* Split the dataset into test and train.
* Apply Linear regression model.
* Fit the Regression Model
* After fitting the regression model we obtained 67% accuracy.
* R square value of multiple linear regression is 0.67

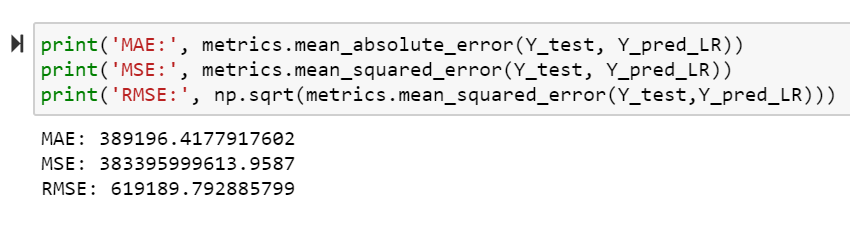


Fig.3.11: Multiple linear calculations

* Values obtained:
* Mean absolute error- : 389196.4177917602
* Mean squared error- 383395999613.9587
* Root mean squared error-619189.792885799
* Graph is linearly plotted.
* Most of the graph is scatered in the region from 0 to 2000000.

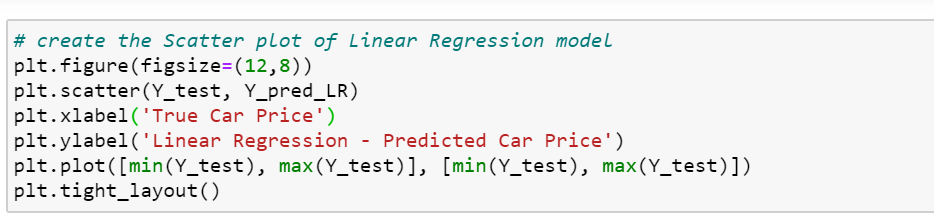


Fig.3.12: Code for plot of multiple linear regression

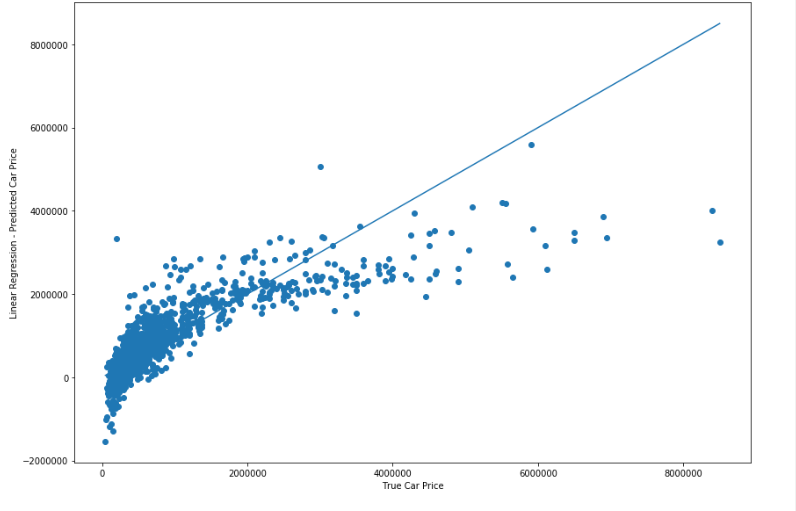


Fig.3.13: Graph for multiple linear regression

**Decision Tree Model:**

* Decision tree builds classification or regression models in the form of a tree structure.
* It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
* The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node

Decision trees can handle both categorical and numerical data

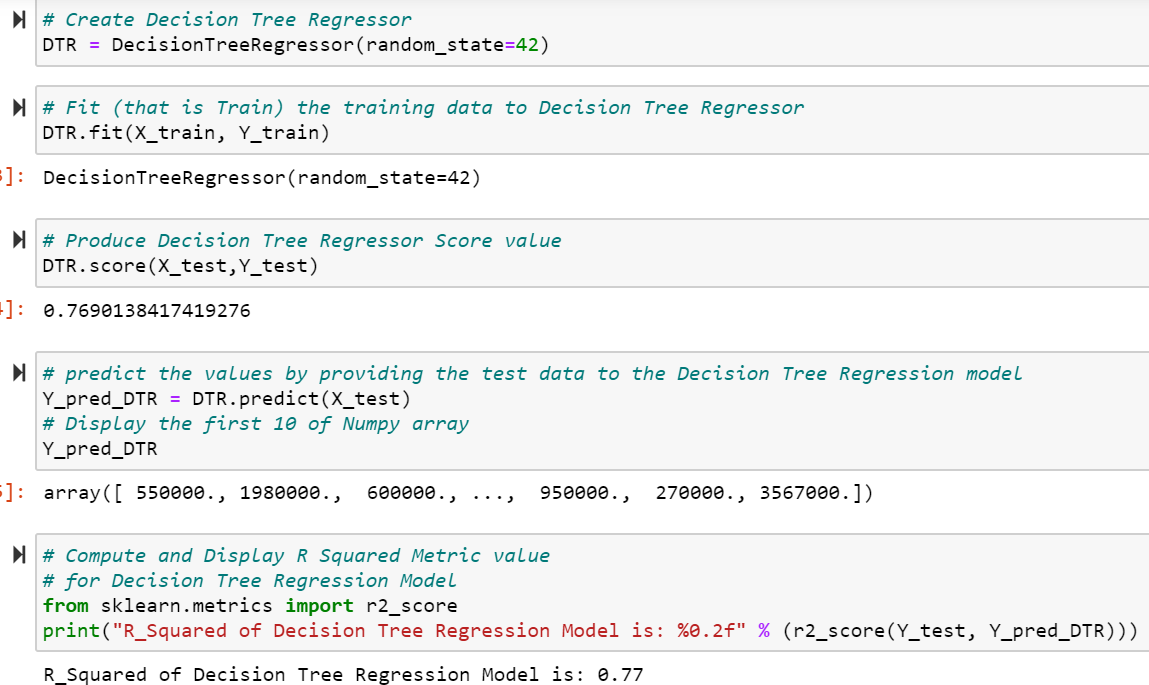


Fig.3.14: Decision Tree Model

* Import Decision Tree Regressor
* Apply Decision Tree Regressor model
* Fit the model.
* After fitting the model score obtained is 0.76.. i.e., 76% accuracy
* Predict the values providing the test data to Regression Model
* R\_Square value is 0.77

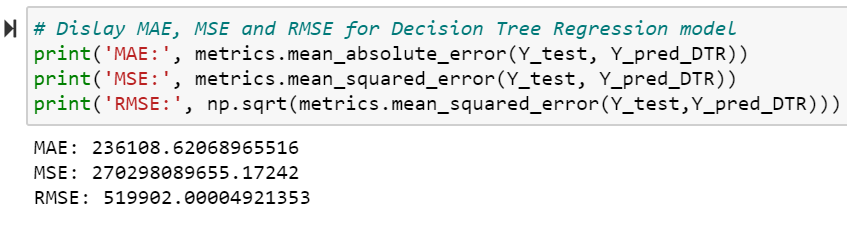


Fig.3.15: Calculations of Decision Tree

* Values Obtained:
* Mean absolute error- 2.317905172413793
* Mean squared error- 1.7949274999999998
* Root mean squared error-4.273482708540924

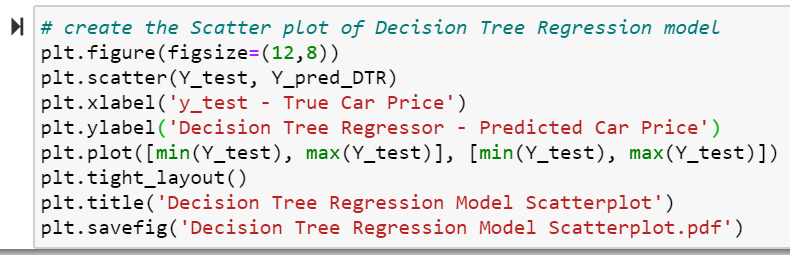


Fig.3.16: Code for the plot of Decision Tree

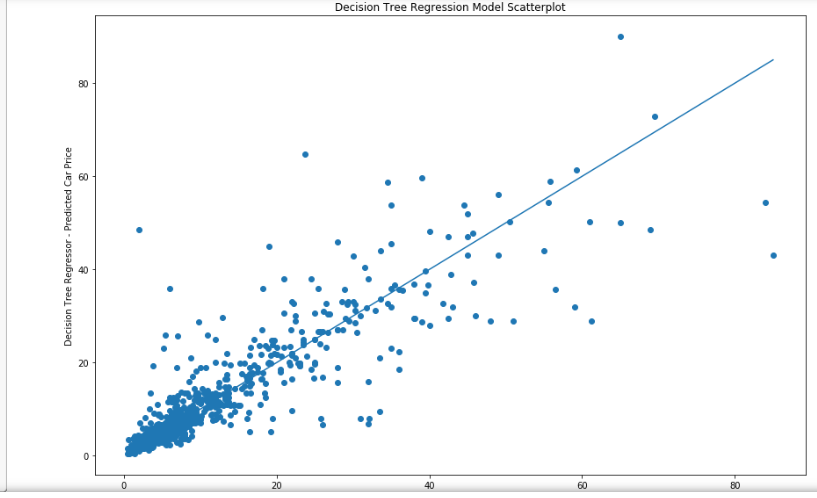


Fig.3.17: Graph for Decision Tree

**Random Forest Algorithm**:

* It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
* As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
* The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

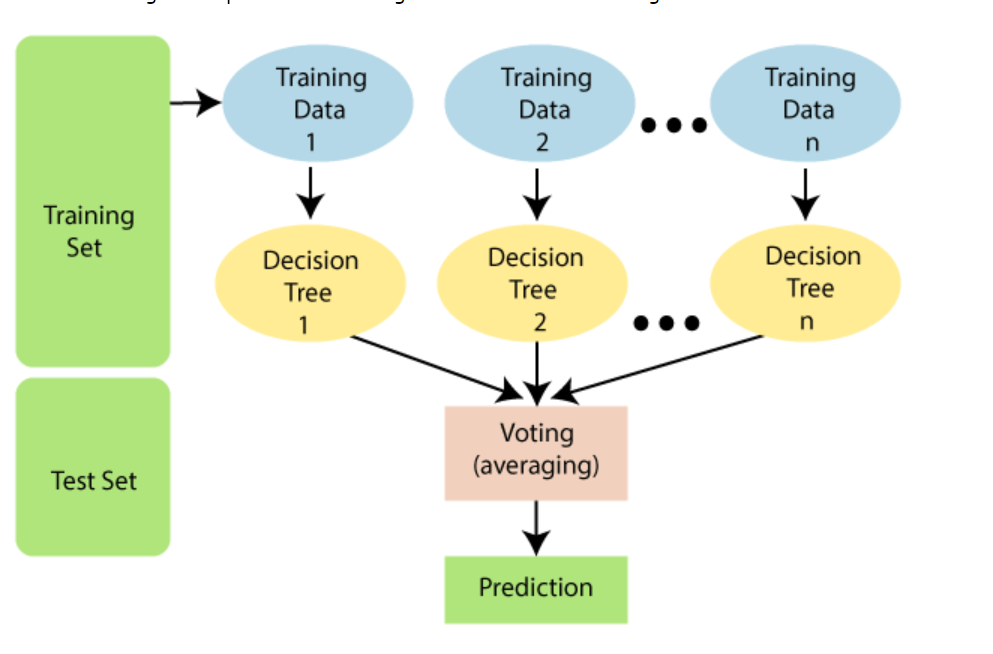


Fig.3.18: Random forest flowchart

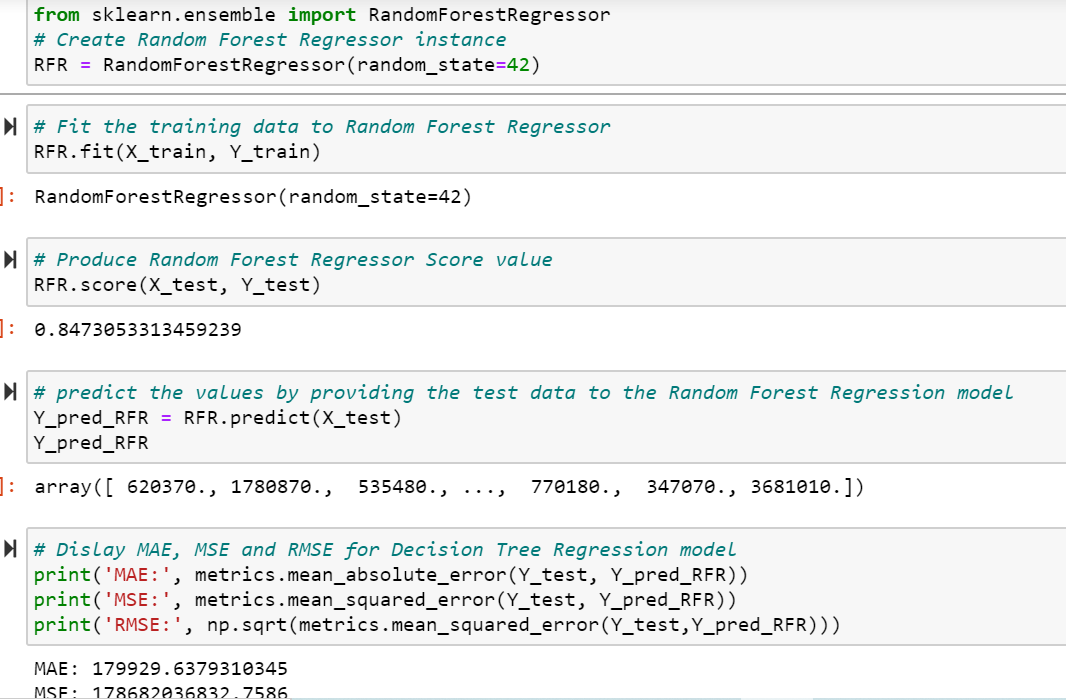


Fig.3.19: Code for Random Forest Algorithm

* Importing the Random Forest Regressor.
* Applying Random Forest Regression.
* Fit the model
* After fitting the model score obtained is 0.847..i.e., accuracy of 84%.
* Predict the test data to Random Forest Regressor model.
* Values Obtained:
* Mean absolute error-1.7949274999999998
* Mean squared error- 1.7949274999999998
* Root mean squared error-4.273482708540924

Fig.3.20:Calculations and Graph for Random Forest Algorithm

* Values Obtained:
* Mean absolute error-1.7949274999999998
* Mean squared error- 1.7949274999999998
* Root mean squared error-4.273482708540924

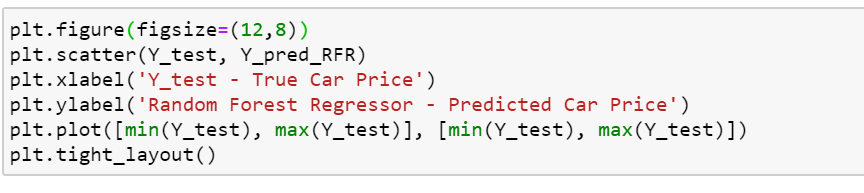


Fig3.21. Code for The graph of Random Forest

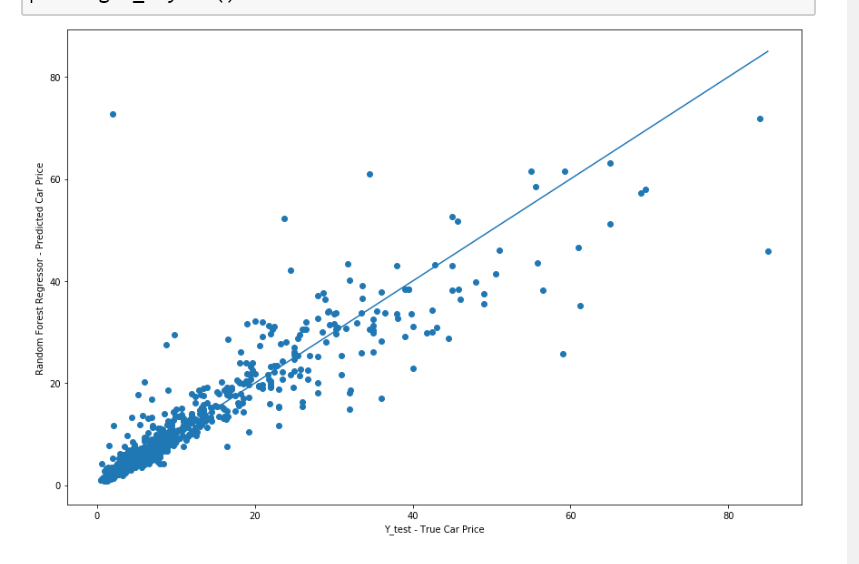


Fig.3.22:Graph for Random Forest Algorithm

**GRAPHIC USER INTERFACE:**

****

Fig.3.23: GUI..

**FEATURES OF GUI**

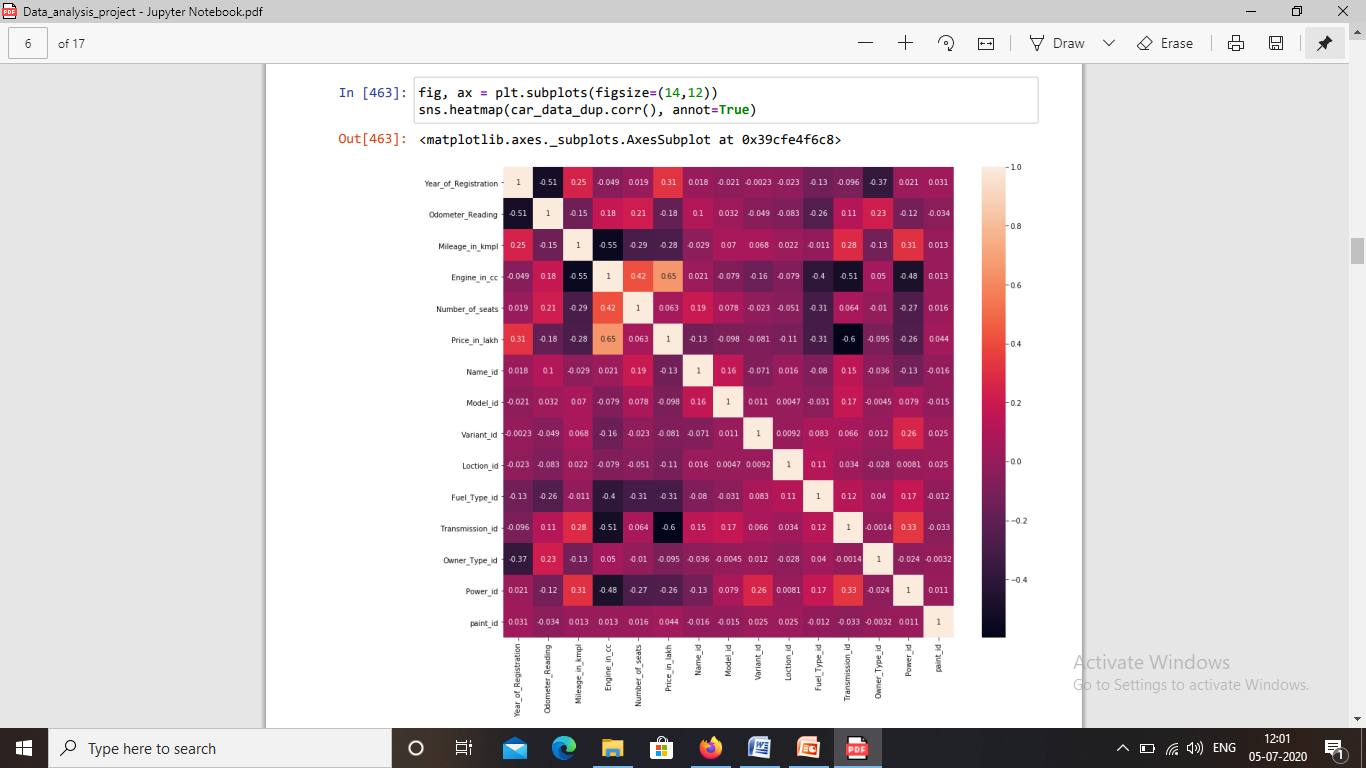
* All the drop down lists in our GUI are arranged in sorted manner.
* Only the most required entries are asked by the user which play an important role in predicting the price of car.
* Our GUI takes all the entries and give us the price of the car.

**CHAPTER 4**

**EXPERIMENT AND RESULT ANALYSIS**

**EXPERIMENT**

* We have preprocessed the data by altering the names of attributes and giving them more relatable names. Removal of standard NULL (NaN ) or missing values is done.
* To visualize the dataset we have drawn various graphs that shows relationship between various attributes of the dataset.



This is a correlation matrix that shows relation between all the attributes used.It is also color coded.

* Here we visualize that all columns in black color have negative relationship. That implies we can have to compromise on either attributes depending upon individual need. For example, we cannot get engine and power both the facilities in a single car.
* Shades of pink and purple shows positive relationship .That implies we can get both qualities in a single car. It various in range from 0 to +1.
* +1 indicates the relationship with itself. The white color boxes show this relationship.
* We have a applied Linear Regression model, decision tree model, and random forest algorithms to analyze the dataset and predict the prize of old cars.

**RESULT ANALYSIS**

On applying data visualization, we came across following conclusions:

* We see that most of attributes are important in deciding the prize of the car
* After drawing above conclusion we applied several machine learning algorithms.
* We have applied logistic regression, decision tree model. After applying above models and analyzed our data by testing it on each model we got following results
* For this project the accuracy for linear regression is **67%,** for decision tree repressor it is **76%,** and for random forest algorithm it is **84%.**
* After analyzing the project we come to conclusion that **Random Forest Algorithms** is best suited with 84% accuracy.

**CHAPTER 5**

**CONCLUSION**

* 1. **Discussion**
* In our project after importing the data set and performing data preprocessing and data visualisation we obtained our first conclusion that Fuel\_Type , Transmission\_Type,Paint\_color,Owner\_Type,type,condition are categorical data which are replaced by numerically coded values.
* Second , In transmmission we come to the conclusion that people prefer more manual cars in comparison to automatic cars,simiilarly we come to know in terms of fuel type deisel cars are preferred more over petrol cars.
* Third, most of the outliers were contained in Odometer\_Reading and Mileage\_in\_kmpl.
* Fourth, mileage\_in\_kmpl and Transmission\_type\_id are correlated with an approximate value of 0.33
* Fifth, On performing multiple linear regression we obtained an accuracy of 67%
* Sixth,On performing Decision Tree Regression we obtained an accuracy of 79%
* Seventh,On performing Random Forest Algorithm we obtained an accuracy of 84%
  1. **Future Work**
* The proposed model is GUI based, it is more reliable and expandable as modifications can be done in future according to our ease. We can have a model which can give the selling prize with more accuracy with focus on profit and efficiency.
* all non-numeric features can be converted to nominal data instead of ordinal data . This may cause a serious change in performance of predictive models
* Handle categorical variables more precisely in comparison to numerically coded value
* Making our model more accurate to provide better results in comparison to vehicles having lower price and lower miles

**REFERENCES**

* <https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv>
* NATIONAL TRANSPORT AUTHORITY. 2014. Available from:

http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January.

* MOTORS MEGA. 2014. Available from:

http://motors.mega.mu/news/2013/12/17/auto-market-8-decrease-sales-new-

cars/ [Accessed 17 January 2014].

* YOUTUBE
* Python\_cousre.eru