FLIP ROBO

CAR PRICE PREDICTION PROJECT

Submitted by:

JASMINE KAUR

ACKNOWLEDGMENT

This project was my first experience in terms of collecting the data and then making the entire machine learning project based on that data. I learned many things in this process. I discovered different methods of collecting the data and learnt a lot from every single mistake.

This would not have been possible without the kind support of my SME, Ms Khushboo Garg, whose constant guidance helped me in completing this project. I would also like to thank DT Team who taught me the skills to make such projects.

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.

- *Conceptual Background of the Domain Problem*

  It is a difficult task to predict the price of a used car. The re-sale value of a car depends upon a number of factors. Some of the important factors are the age of the car, the model of the car, the mileage of the car, the transmission of the car, the fuel type of the car. There are also some other factors associated with this, like the physical state of the car, air conditioner, sound system, power steering etc. Thus, the price depends on a large number of factors but due to unavailability of data, we will consider major factors to determine the price of the car.

- *Review of Literature*

  The goal of this project is to build a model that predicts the price of the used car. This would help buyers to pay the right amount of price depending on the features available in a given car. On the other side, it will help the sellers know the real worth of their cars. Thus, this model is helpful for both the sellers and the buyers.

Analytical Problem Framing

- ## *Mathematical/ Analytical Modeling of the Problem*

This project has two phases:

1. Data collection phase
2. Model building phase

Data collection phase:

We have scraped 2434 used cars data from cars24.com. We have scraped information like km driven, fuel, transmission, price, year and model of the car.

Model building phase:

After collecting the data, we have cleaned the data. In this process, we have dealt with the null values, skewed data, outliers. We have performed feature engineering in order to get most out of the data. We have checked for multicollinerarity. We performed EDA to get insights about the data. It was necessary to encode the categorical data into numerical. After performing these, we have applied four models and then applied cross validation. The best model comes out to be random forest regressor.

- ## *Data Sources and their formats*

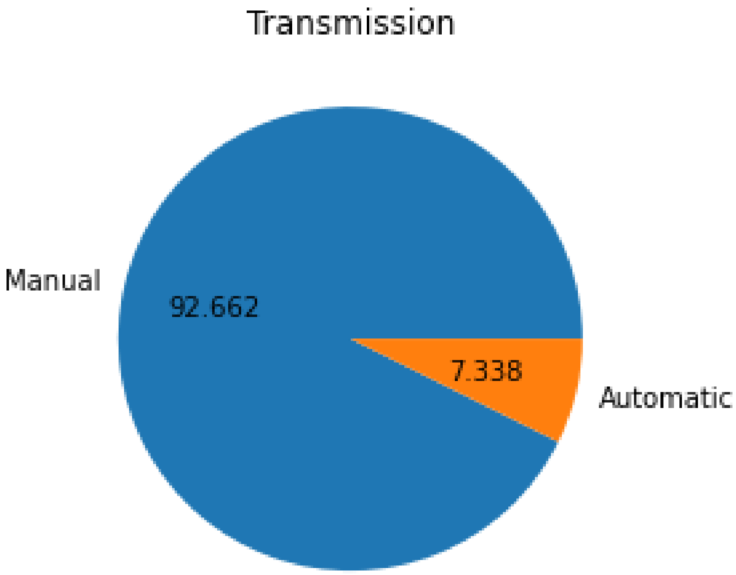
We have collected the data from cars24. The data was scrapped using Selenium.

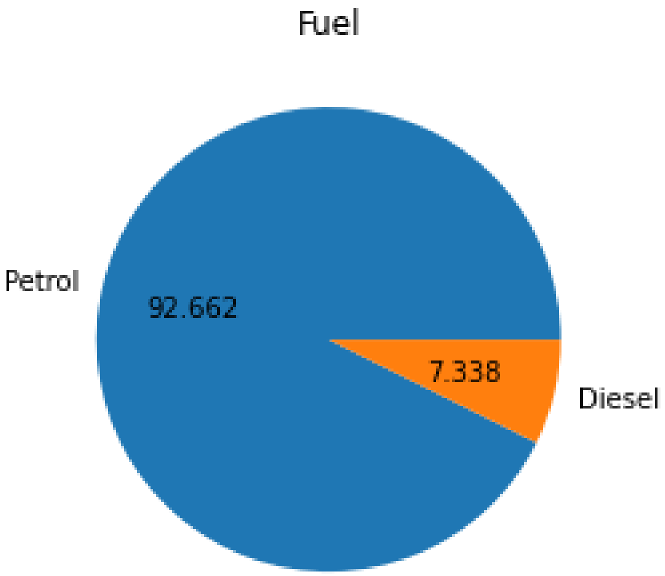
- ## *Data Preprocessing Done*

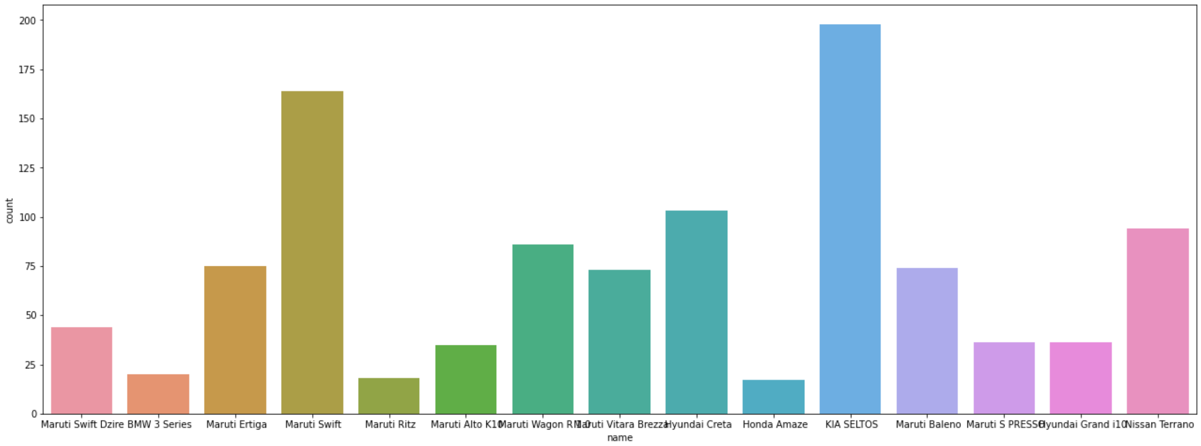
1. Null values: we have checked the null values and dealt with it.
2. Converted the object data type into integer data type
3. Removing unnecessary symbols or units from the data
4. Converted categorical data into numerical data
5. Checked outliers and dealt with them using IQR
6. Checked skewness but it was fine

- ***Data Inputs- Logic- Output Relationships***

  The output variable is the selling price of the used car. Since the price is a continuous variable, thus this is a regression problem. The input variables are fuel, km driven, transmission, fuel, year. These input features work as an independent variable and help in predicting the price of the car.







- ***<u>Hardware and Software Requirements and Tools Used</u>***

The hardware and software requirements along with the tools, libraries and packages used are:

1. Laptop
2. MS Excel
3. Google colab
4. Pandas
5. Numpy
6. Visualization tools
7. Regressors
8. Github

Model/s Development and Evaluation

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

1. Boxplot for summarizing variations and checking outliers
2. Histograms are used to check outliers
3. Correlation matrix was used to analyse the correlation between the features
4. Standard scaler was used to standardize the data
5. We have split the data using train test split

- ***Testing of Identified Approaches (Algorithms)***

The algorithms used for the training and testing are:

1. Random Forest Regressor
2. Linear Regression
3. Gradient Boosting Regressor
4. KNeighbors Regressor

- ***Run and Evaluate selected models***

MODELLING ALGORITHM	R2 SCORE
LINEAR REGRESSION	0.802336
RANDOM FOREST	0.998455
KNEAREST NEIGHBORS	0.916056
GRADIENT BOOSTING	0.991785

After applying cross validation

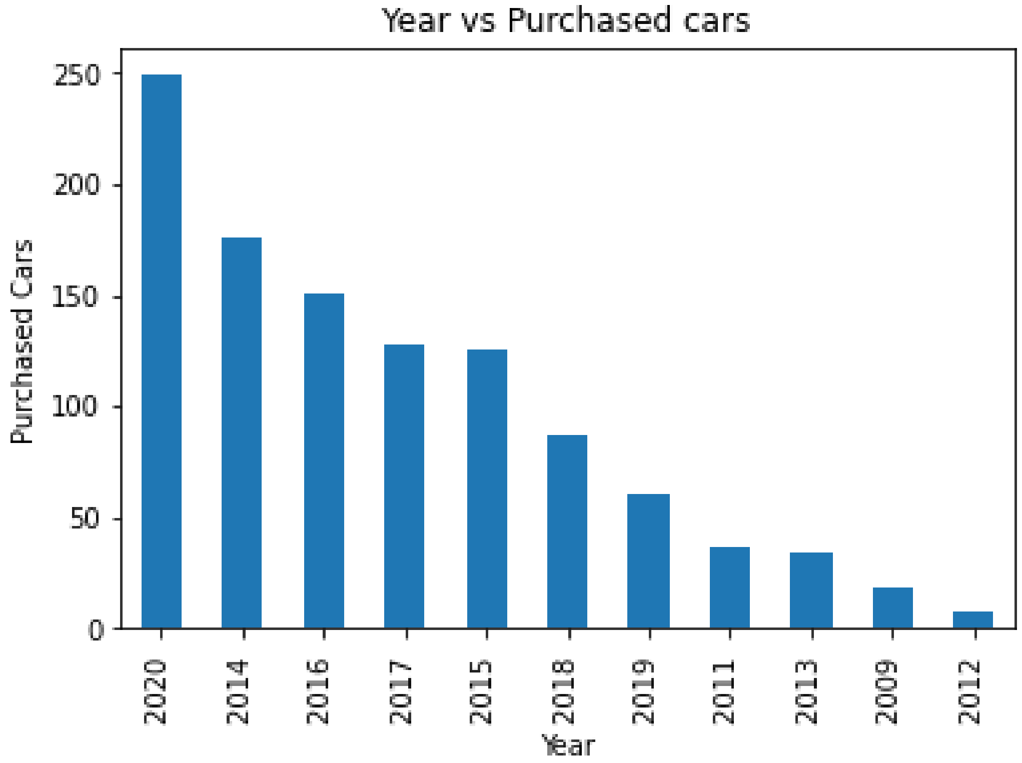
MODELLING ALGORITHM	R2 SCORE
LINEAR REGRESSION	0.657514
RANDOM FOREST	0.989432

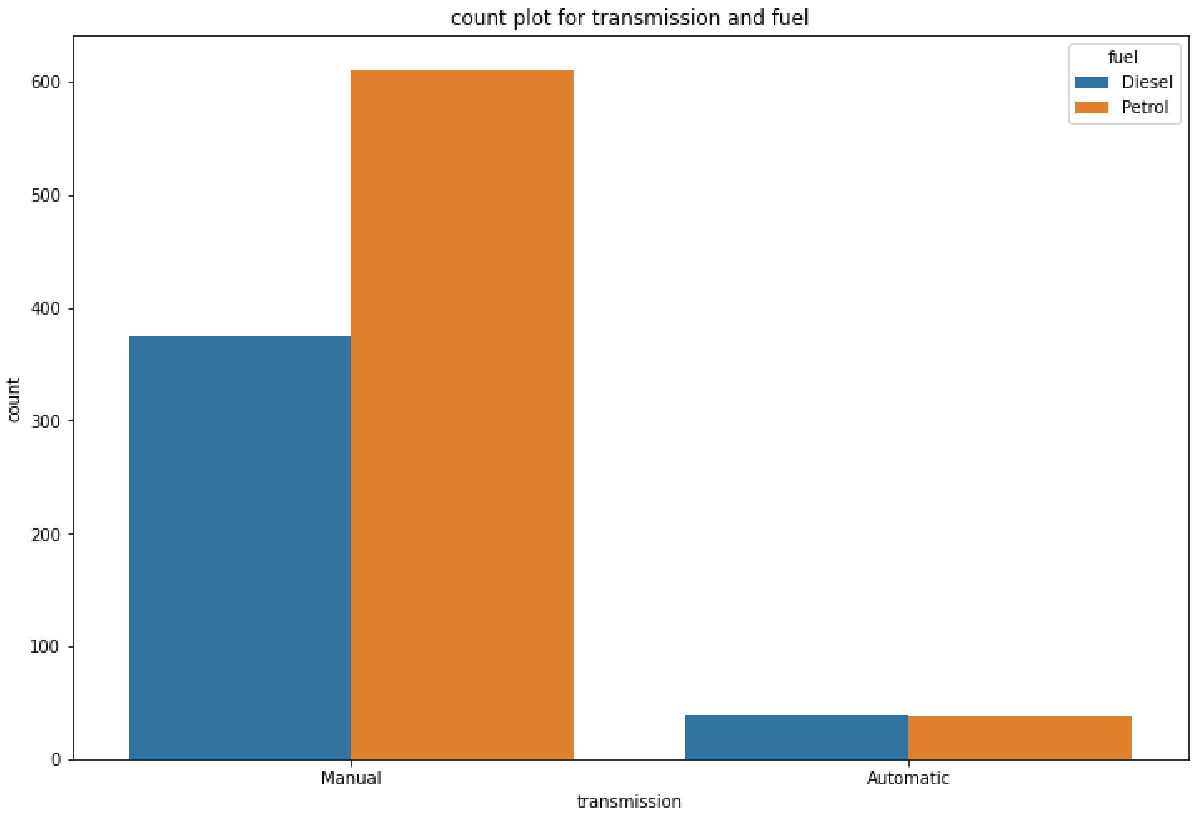
KNEAREST NEIGHBORS	0.762045
GRADIENT BOOSTING	0.979396

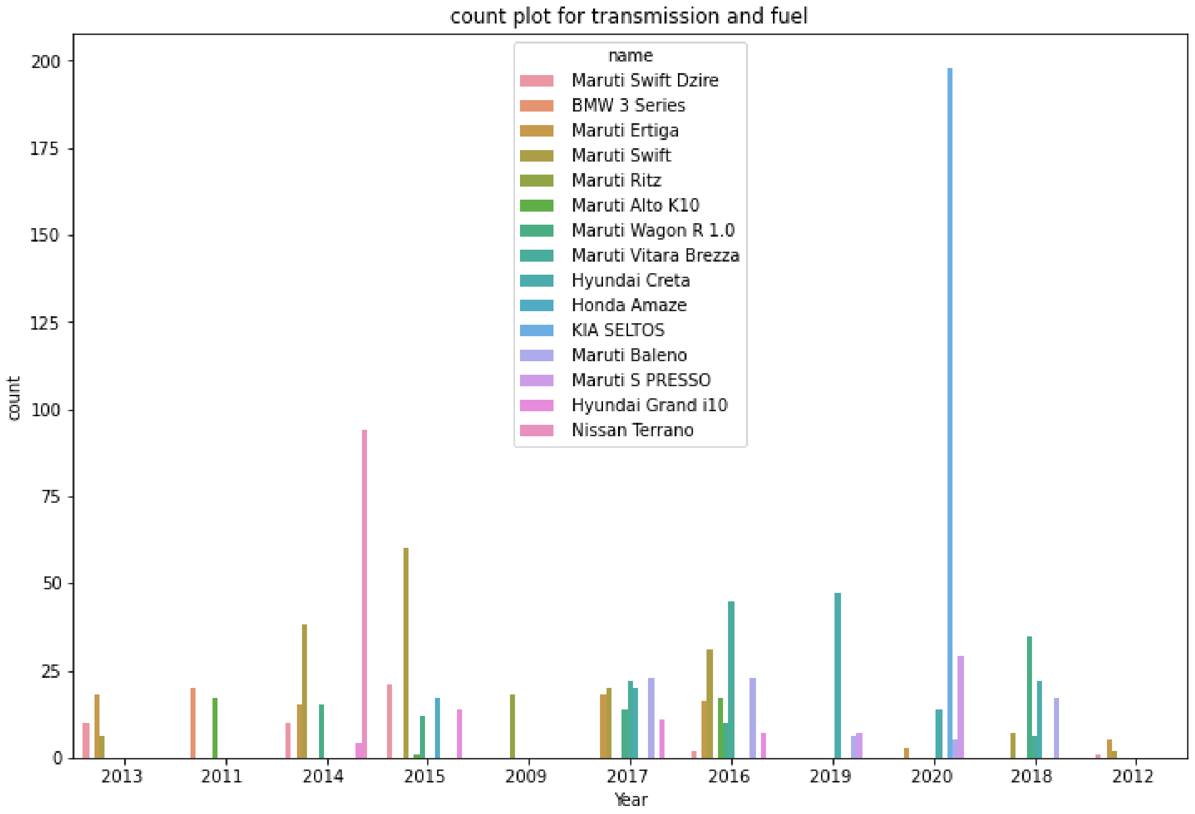
- ***Key Metrics for success in solving problem under consideration***

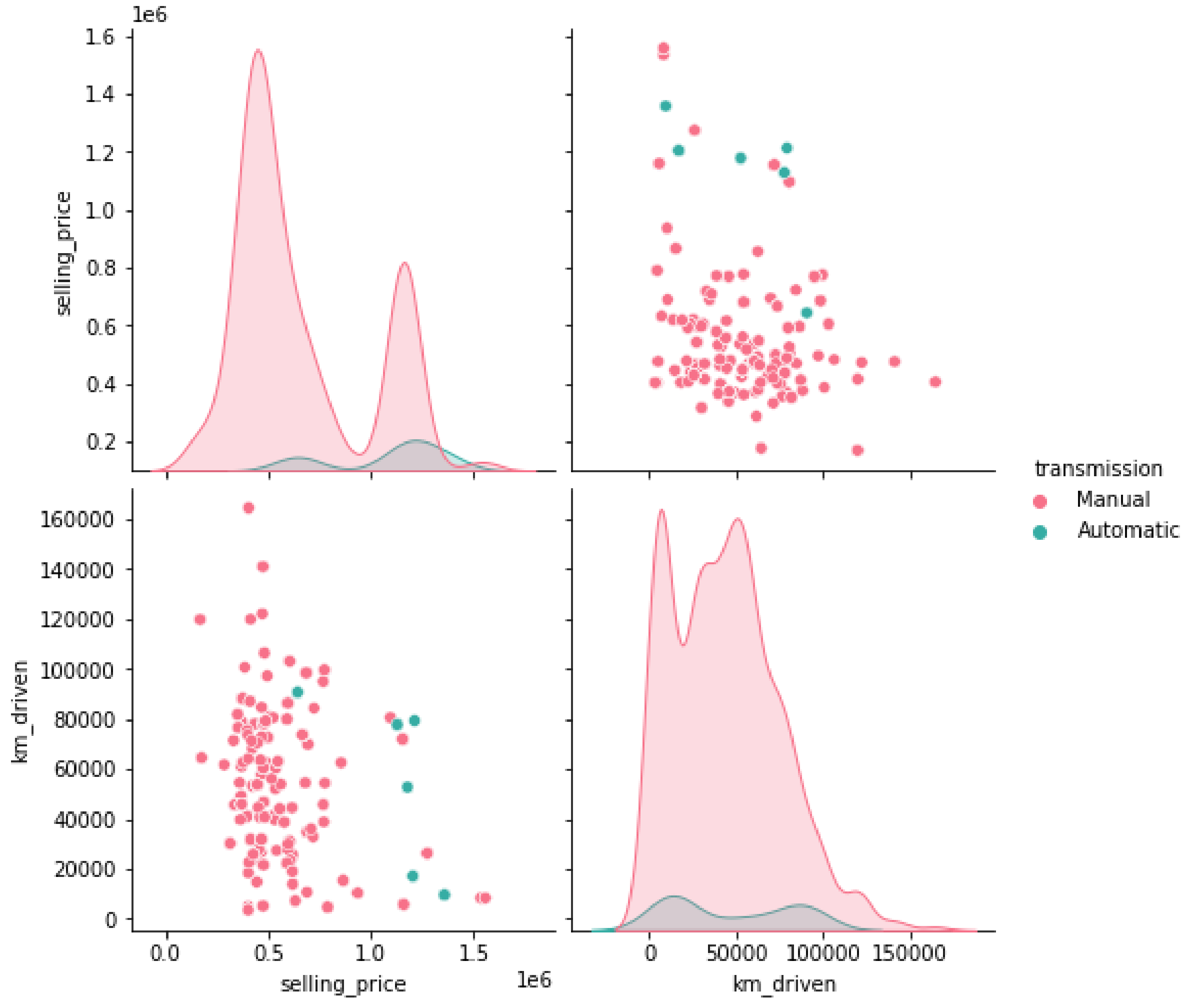
  The key metrics used is r2 score. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

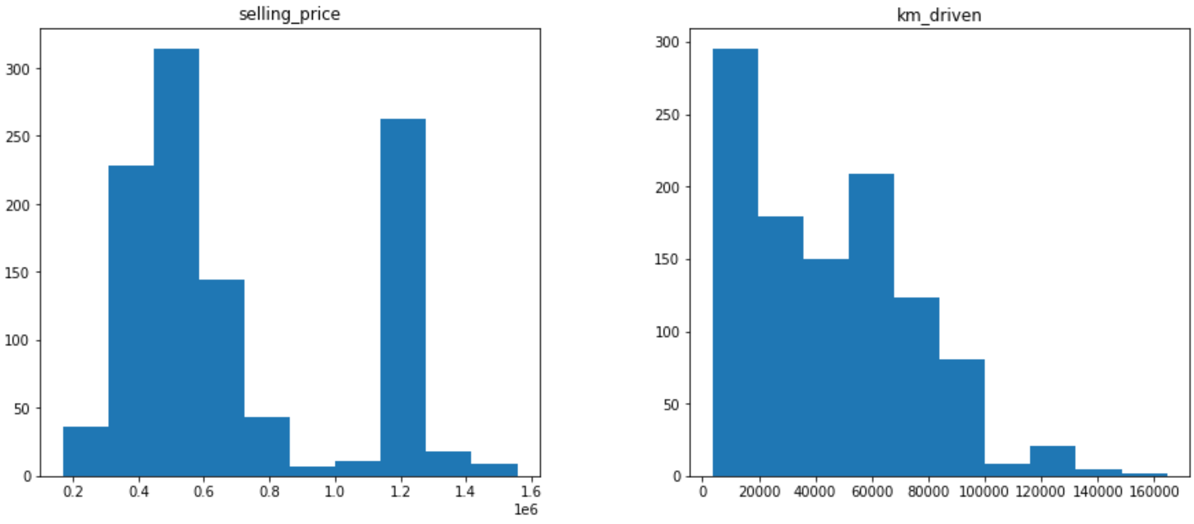
- ***Visualizations***

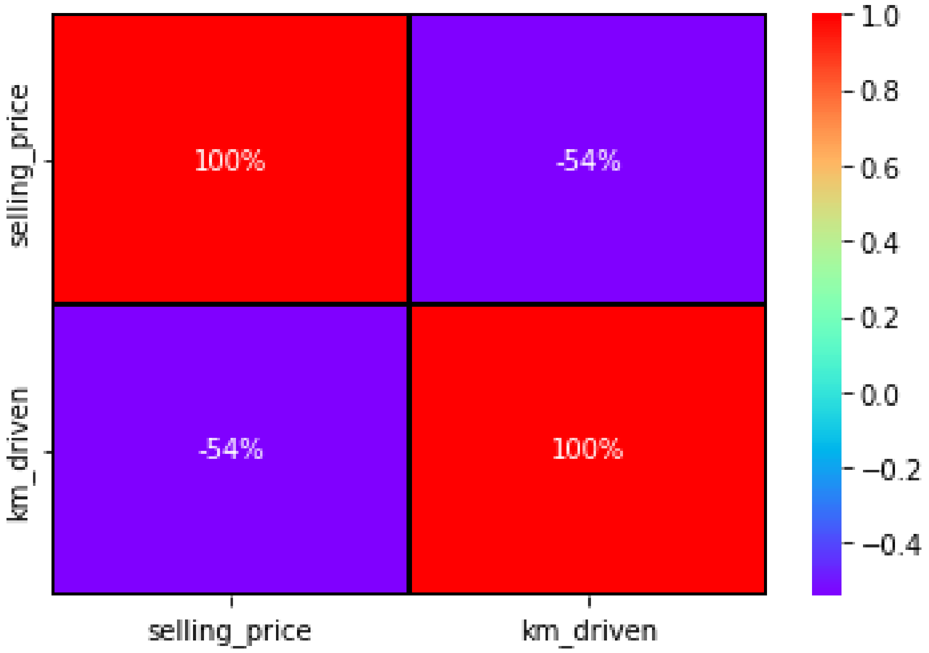


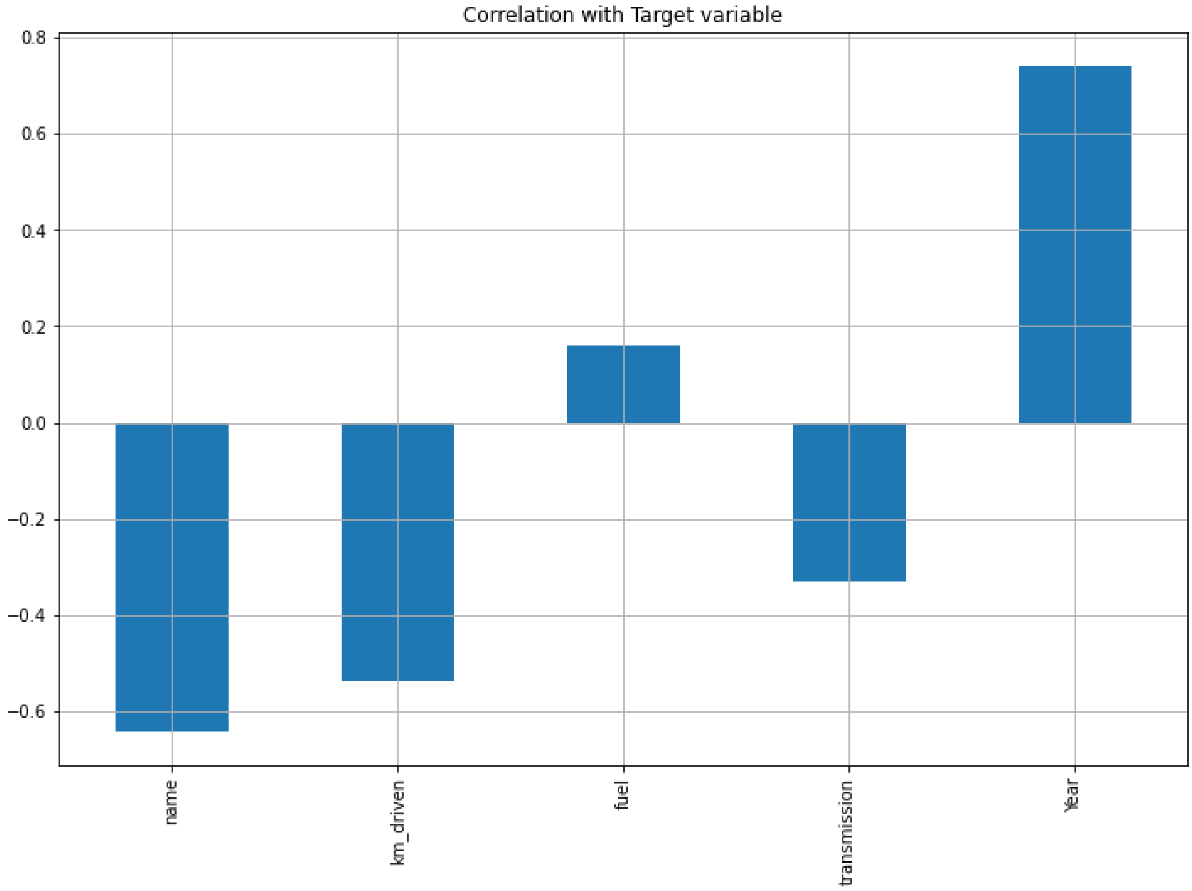

count plot for transmission and fuel
fuel
Diesel
Petrol
count
0
100
200
300
400
500
600
Manual
Automatic
transmission

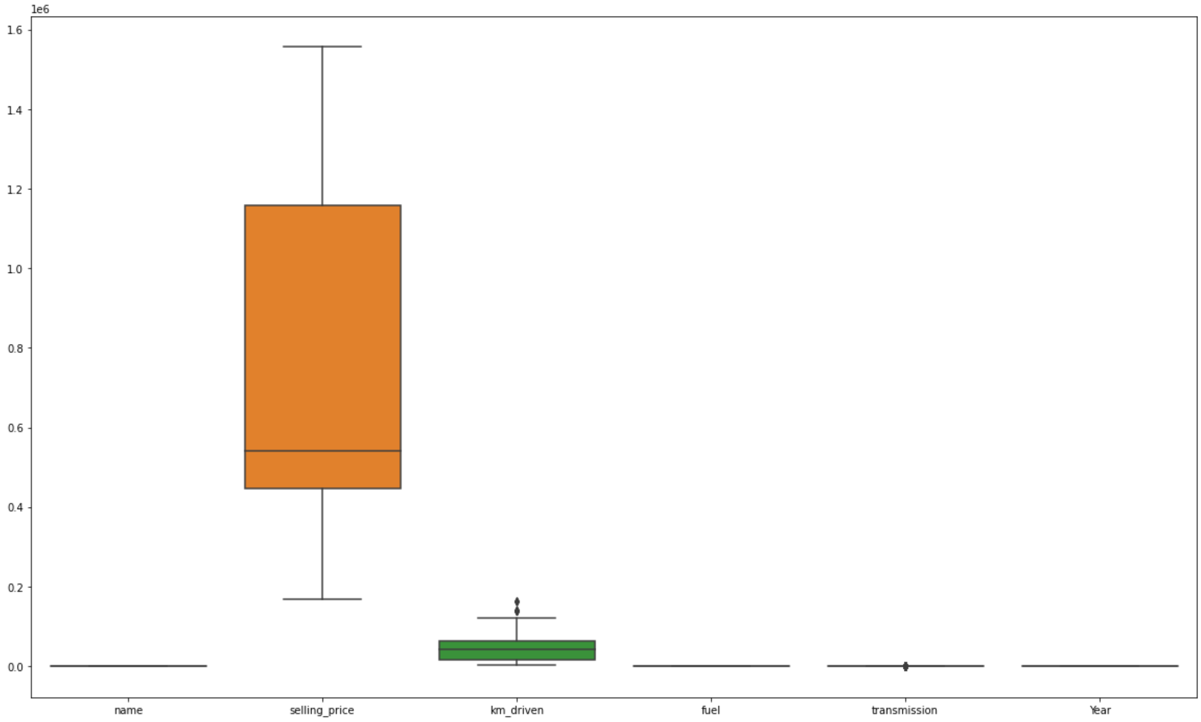

count plot for transmission and fuel
name
Maruti Swift Dzire
BMW 3 Series
Maruti Ertiga
Maruti Swift
Maruti Ritz
Maruti Alto K10
Maruti Wagon R 1.0
Maruti Vitara Brezza
Hyundai Creta
Honda Amaze
KIA SELTOS
Maruti Baleno
Maruti S PRESSO
Hyundai Grand i10
Nissan Terrano
count
0
25
50
75
100
125
150
175
200
2013
2011
2014
2015
2009
2017
2016
2019
2020
2018
2012
Year


1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
selling_price
160000
140000
120000
100000
80000
60000
40000
20000
0
km_driven
0.0
0.5
1.0
1.5
1e6
selling_price
0
50000
100000
150000
km_driven
transmission
● Manual
● Automatic


selling_price
300
250
200
150
100
50
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e6
km_driven
300
250
200
150
100
50
0
0
20000
40000
60000
80000
100000
120000
140000
160000


100%
-54%
-54%
100%
selling_price
km_driven
selling_price
km_driven
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Correlation with Target variable
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
name
km_driven
fuel
transmission
Year



Interpretation of the Results

After comparing the r2 score and the cross validation score, we have reached at a conclusion that the random forest regressor is the best model. We have applied hyperparameter tuning but it reduced the r2 score by 10%, so we retained the default parameters and saved the model.

CONCLUSION

- ## *Key Findings and Conclusions of the Study*

  We have got insights about collecting the data and then using to solve real life problems.

  Performing data cleaning, data pre-processing are difficult but interesting steps in the process of reaching the end goal.

  Machine learning models help us in predicting the dependent variable with the help of independent variables.

  We have finalised Random Forest as the best model and saved it in pkl format.

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

  Collecting data is one of the most difficult steps. As it is also the most important step in building a model, we need to be very careful in scraping it.

  Scraping data is a time consuming step, so we need to pay attention to every small detail.

  Handling unclean data and cleaning the data is the base of any model. We should perform this task with care.

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

  We could have used more data but due to limited time, we are able to scrap this much data only.