

USED CAR PRICE PREDICTION PROJECT



Submitted by:

Yashna Shah

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Sapna Verma as well as Flip Robo Technologies who gave me the opportunity to do this project on Used Car Price Prediction, which also helped me in doing lots of research wherein I came to know about so many new things especially the data collection part.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1) <https://github.com/>

2) <https://www.kaggle.com/>

3) <https://medium.com/>

4) <https://towardsdatascience.com/>

5) <https://www.analyticsvidhya.com/>

**INTRODUCTION**

* Business Problem Framing

In this project we have to build a used car price prediction model using machine learning models from the collected data. 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.

Cars are more than just a utility for many. We all have different tastes when it comes to owning a car or at least when thinking of owning one. Some fit in our budget and some lauxury brands are heavy on our pockets. But that should not stop us from owning it, atleast used ones. The goal of this project to predict the costs of used cars to enable the buyers to make informed purchase using the data collected from various sources and distributed across various locations in India.

* Conceptual Background of the Domain Problem

One of the most booming markets in the digital space is that of the automobile industry wherein the buying and selling of used cars take place. Sometimes we need to walk up to the dealer or individual sellers to get a used car price quote. However, buyers and sellers face a major stumbling block when it comes to their used car valuation or say their second-hand car valuation. Traditionally, you would go to a showroom and get your vehicle inspected before learning about the price. So instead of doing all these stuffs we can build a machine learning model using different features of the used cars to predict the exact and valuable car price.

* Review of Literature

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.

Impact of COVID-19 on Indian automotive sector: The Indian automotive sector was already struggling in FY20. before the Covid-19 crisis. It saw an overall degrowth of nearly 18 per cent. This situation was worsened by the onset of the Covid-19 pandemic and the ongoing lockdowns across India and the rest of the world. These two years (FY20 and FY21) are challenging times for the Indian automotive sector on account of slow economic growth, negative consumer sentiment, liquidity crunch, low-capacity utilisation and potential bankruptcies.

The COVID-19 crisis has exposed chinks in the automobile business model and it could catalyse a big move towards electric vehicles (EVs). That could be the big positive for auto sector.

* Motivation for the Problem Undertaken

Based on the problem statement and the real time data scrapped from the Cars24 website, I have understood how each independent feature helped me to understand the data as each feature provides a different kind of information and perform root cause analysis to predict the price of the used car.

Based on the different features like the model/brand of the car, kilometres driven, variant type, fuel type, number of owners, EMI etc. I would be able to model the price of used car as this model will then be used by the client to understand how exactly the prices vary with the variables. They can accordingly work on it and make some strategies to sell the used car and get some high returns. Furthermore, the model will be a good way for the client to understand the pricing dynamics of a used car.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In our scrapped dataset, our target variable "Price " is a continuous variable. Therefore, we will be handling this modelling problem as regression.

This project is done in two parts:

Data Collection phase

In this section I scraped the data of used cars from Cars24 website where, I have fetched data for different locations. The features which I scraped are Brand/Model, variant, manufacturing year, driven kilometres, fuel, number of owners, locations, EMI and at last target variable Price of the car. I have tried to include all types of cars in data for example- SUV, Sedans, Coupe, minivan, Hatchback.

Model Building phase:

After collecting the data, we need to build a machine learning model. Before model building, will do all data pre-processing steps. Try different models with different hyper parameters and the select the best model.

Will include all the below steps mentioned:

1. Data Cleaning

2. Exploratory Data Analysis (EDA)

3. Data Pre-processing and Visualisation

4. Model Building

5. Model Evaluation

6. Selecting the best model

* Data Sources and their formats

The dataset is in the form of CSV (Comma Separated Value) format and consists of 8 columns (7features and 1 label) with 3658 number of records.

* Model/Year - This shows the car model names and gives us the year in which the car was made.
* Variant- Gives us the manual or automatic gear shifting mechanism.
* Driven\_kms - Number of kilometres the car the driven reflecting on the Odometer
* Fuel Type - Shows the fuel type used by the vehicle
* Owners – Shows the number of owners who have used the car.
* EMI- shows the monthly instalments to be paid.
* Price - Lists the selling price of the used cars.

We can see our dataset includes a target label " Price" column and the remaining feature columns can be used to determine or help in predicting the price of the used cars

* Data Pre-processing Done

1. Importing the necessary dependencies and libraries.
2. Reading the CSV file and converted into data frame.
3. Checking the data dimensions for the original dataset.
4. Looking for null values and accordingly fill the missing data.
5. Checking the summary of the dataset.
6. Checking unique values.
7. Checking all the categorical columns in the dataset
8. Checking for multi collinearity using VIF.
9. Performed PCA
10. Performed Feature Importance using ExtraTrees Regression.

* Data Inputs- Logic- Output Relationships

1. The input data were initially all object data type so had to clean the data by removing unwanted information like “km” from driven\_kms column, replacing Owners list by (1,2,3,4), removing unnecessary information like “/month” from EMI column and ensuring the numeric data are converted accordingly. I then used Label Encoding method to convert all the categorical feature columns to numeric format.
2. Made use of Z score method to remove outliers that were present on our dataset.
3. To handle the skewness, I made use of Power transformation technique ensuring that at least a bell shape curve closer to normal distribution is achieved.
4. Splitted the dependent and independent variable into x and y
5. Scaled the data using Standard Scaler method and made my data ready for modelling.
6. Checked Statistical Summary of the numeric data.

* Hardware and Software Requirements and Tools Used

Hardware technology being used.

RAM : 16 GB

CPU : 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz

GPU : intel Iris Graphics

Software technology being used.

Programming language : Python

Distribution : Anaconda Navigator

Browser based language shell : Jupyter Notebook

Libraries/Packages specifically being used.

Pandas, NumPy, matplotlib, seaborn, scikit-learn

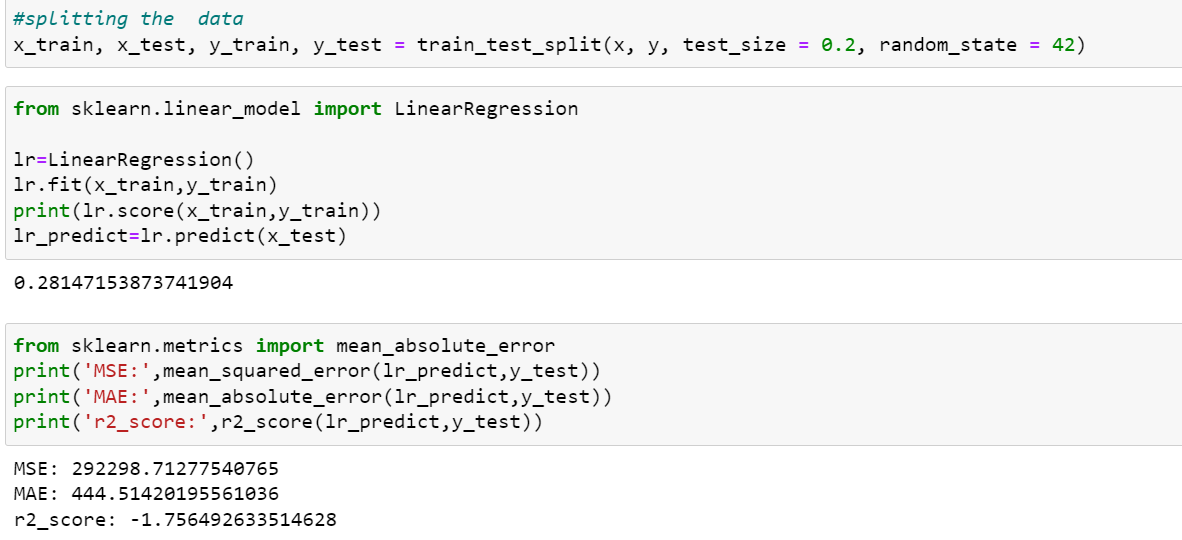
**Model/s Development and Evaluation**

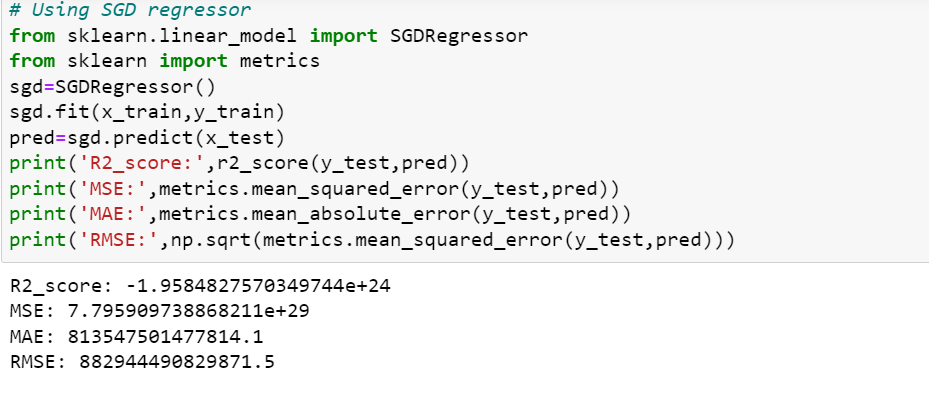
* Identification of possible problem-solving approaches (methods)
* Clean the dataset from unwanted scraped details.
* Impute missing values with meaningful information.
* Encoding the categorical data to get numerical input data.
* Compare different models and identify the suitable model.
* R2 score is used as the primary evaluation metric.
* MSE and RMSE are used as secondary metrics.
* Cross Validation Score was used to ensure there are no overfitting our underfitting models.

* Testing of Identified Approaches (Algorithms)

All the regression machine learning algorithms used are:

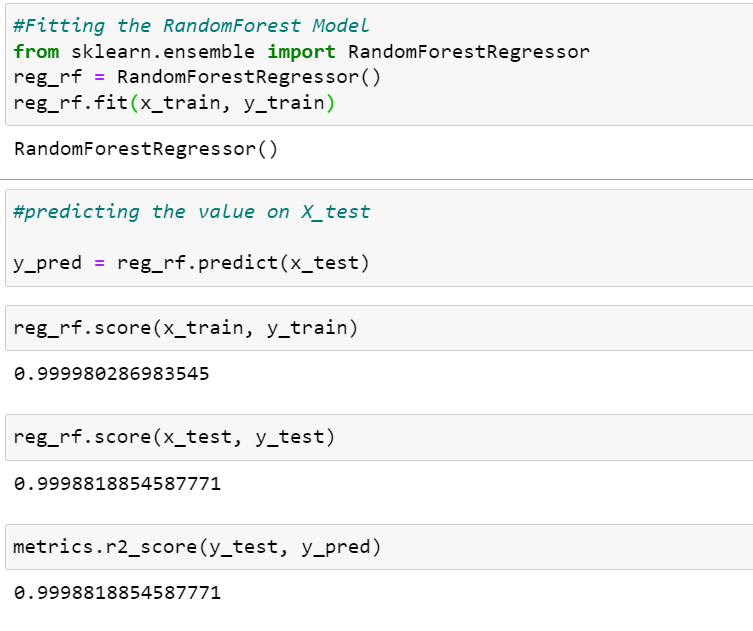
* Linear Regression Model
* Ridge Regularization Model
* Lasso Regularization Model
* SGD Regressor
* Random Forest Regression Model
* Run and Evaluate selected models

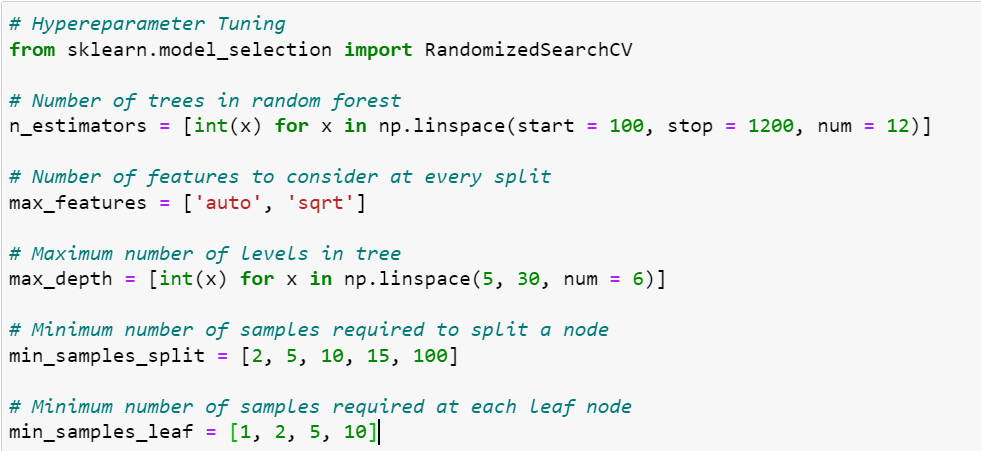


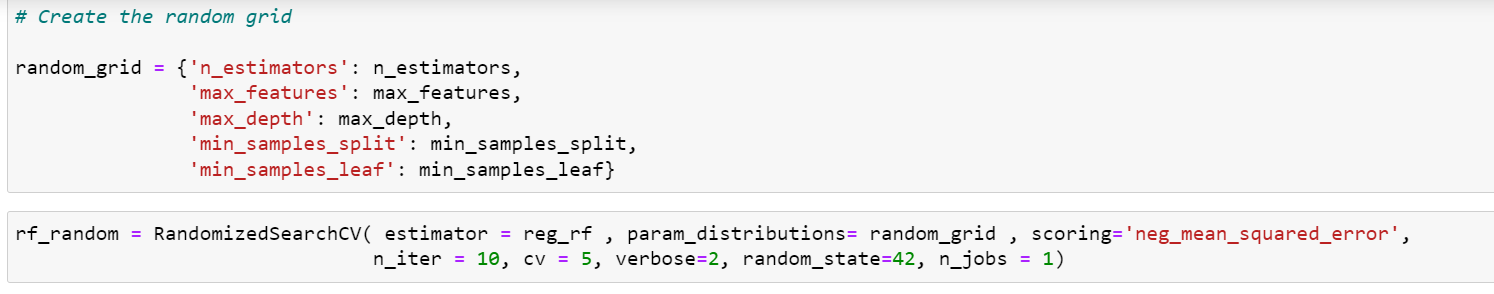


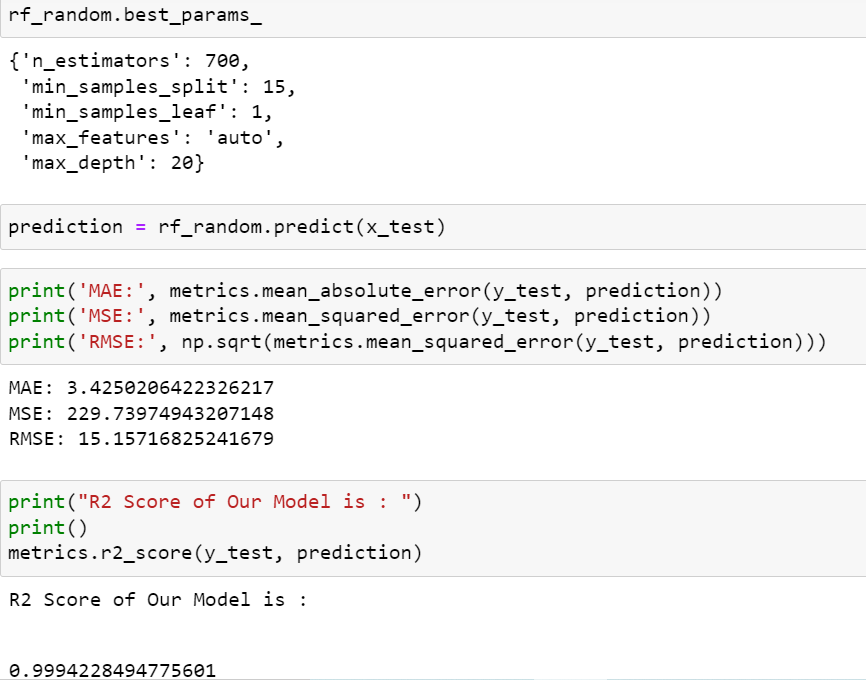












* Key Metrics for success in solving problem under consideration

Will go with Random Forest Regressor as it gives us the best accuracy score of 99%.

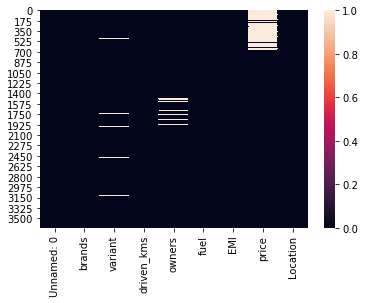
Reasons:

1. Random Forest reduces overfitting in decision tree and helps to improve accuracy.
2. It is flexible for both classification and regression tasks.
3. It also works well with both continuous and categorical variables.
4. It is a rule based approach.
5. It automates missing values present in the data.

The key metrics used here were R2 Score, MSE, MAE, RMSE and Cross Validation Score. We also tried to find the best parameters using Hyper parameter tuning and Random Grid Search to increase the accuracy.

* Visualizations

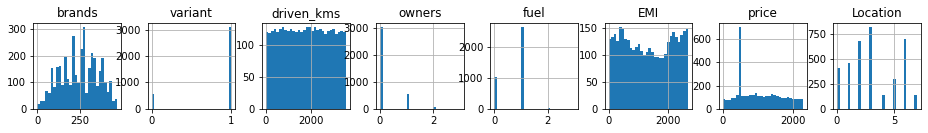
1. Heatmap with null values in our dataset



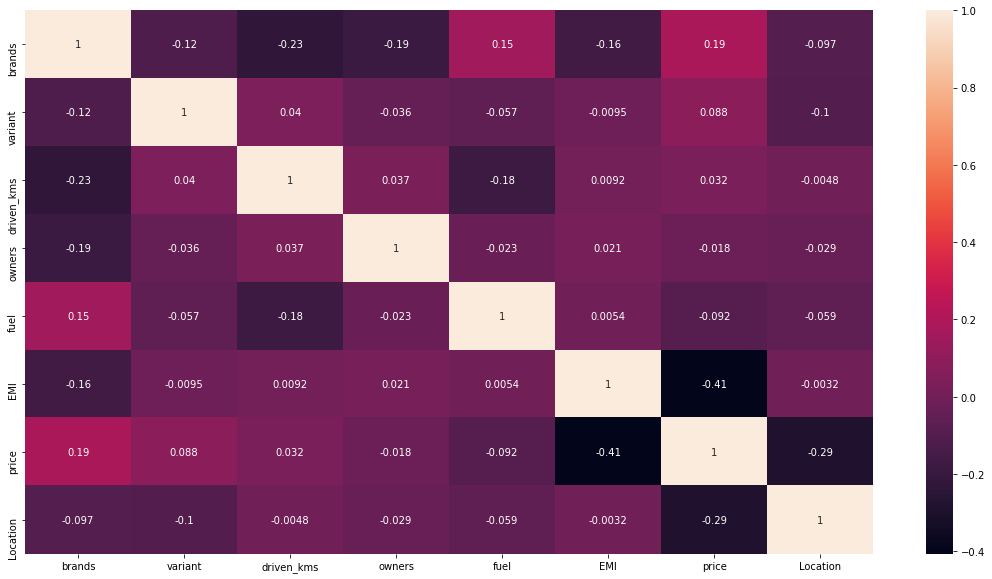
1. Heatmap with no null values in our dataset



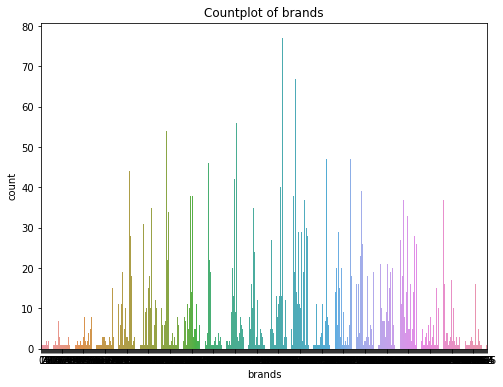
1. Plotting a histogram



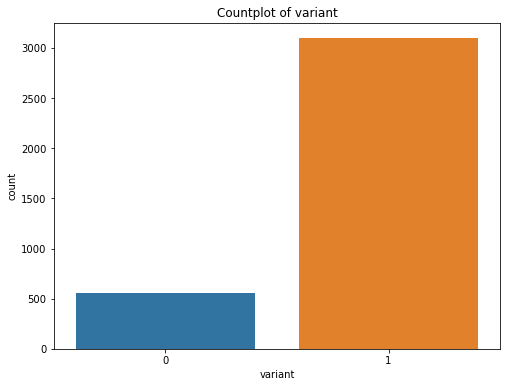
1. Correlation Matrix



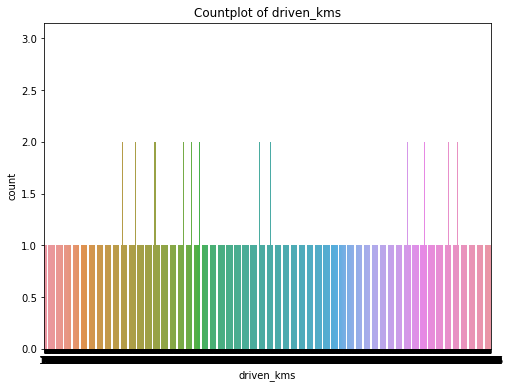
1. Countplot of column Brands



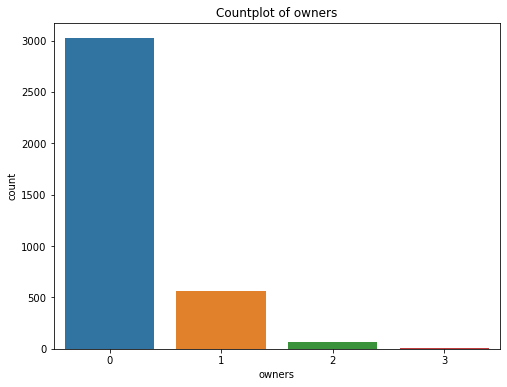
1. Countplot of column Variant



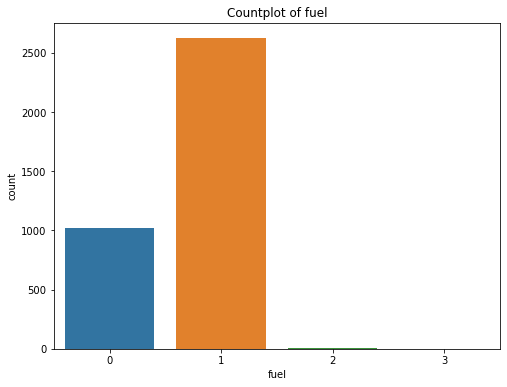
1. Countplot of column driven\_kms



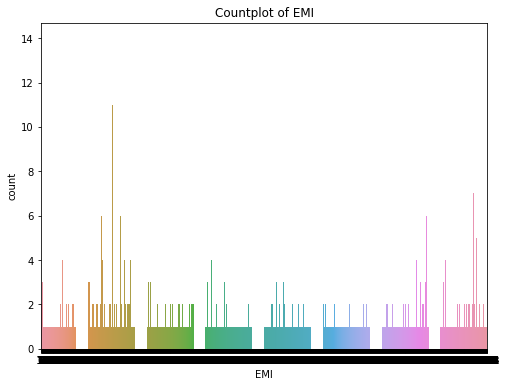
1. Countplot of column Owners



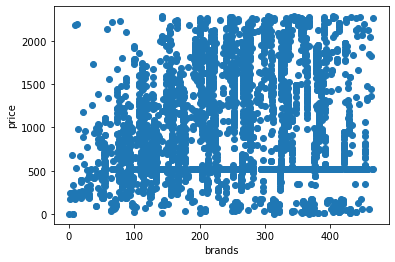
1. Countplot of column Fuel

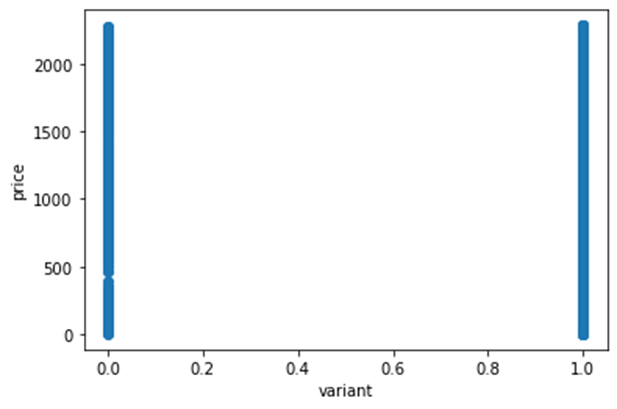


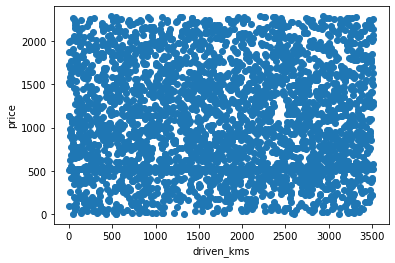
1. Countplot of column EMI

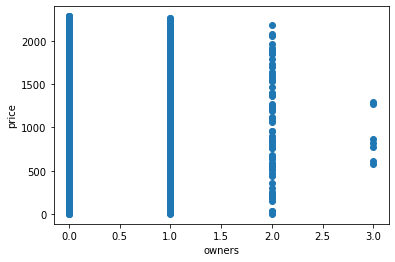


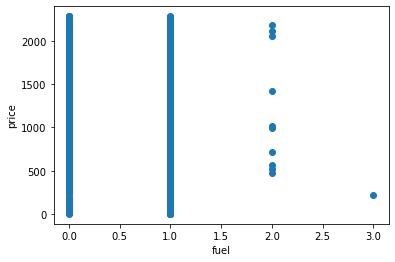
1. Plotting the Scatter plot between all feature variables and target variable

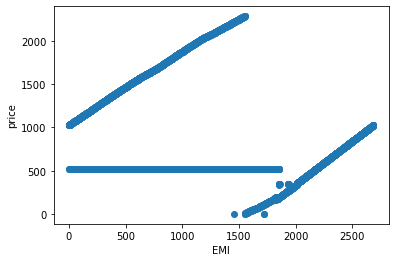


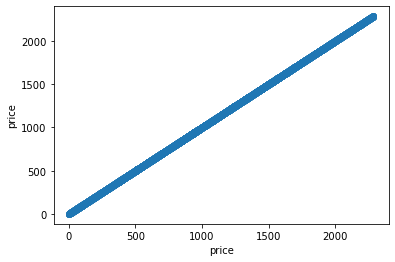


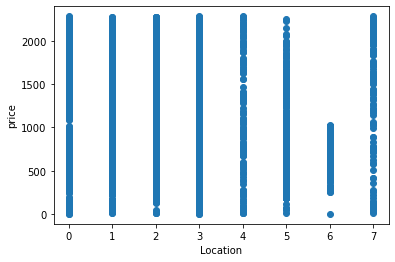




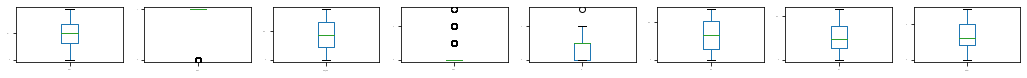




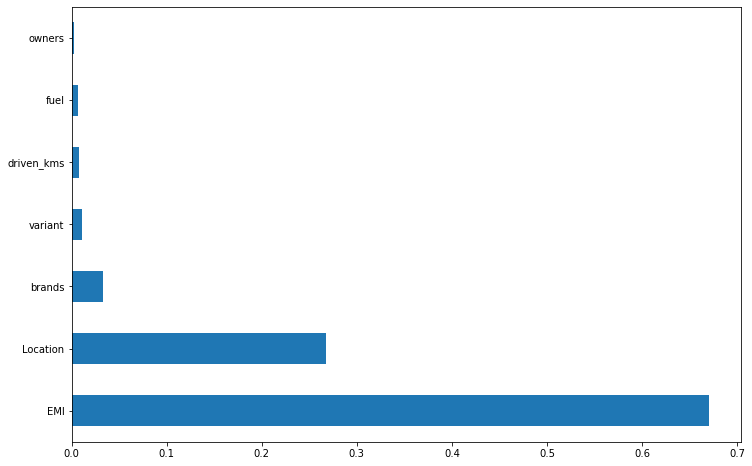




1. Plotting boxplots for checking outliers



1. Feature Importance



Observations:

1. Correlation helps in feature selection and making the feature ready for the model. From the above model we observe that EMI, location and Price are highly correlated to each other.
2. From the counplots we can observe that there are more of 0 owners of the used car, the more of diesel fuel was used compared to others.
3. By the Feature Importance plot, we can observe that EMI has the highest importance.
4. From the boxplots we can observe that there are few outliers which we need to remove before building the model.

**CONCLUSION**

* After the completion of this project, we got an insight on how to collect data, pre-processing the data, analyzing the data and building a model. First, we collected the used cars data from different website Cars 24 and it was done by using Web Scraping.
* The framework used for web scraping was Selenium, which has an advantage of automating our process of collecting data. We collected almost 3500 of data which contained the price and other related features of used cars. Then the scrapped data was combined in a single data frame and saved in a csv file so that we can open it and analyze the data and build the model with best accuracy.

Saving the model

