

**Car Price Prediction Project**

Submitted by:

Pooja Mishra

**ACKNOWLEDGMENT**

I would also like to thank Flip Robo Technologies and Datatrained team who has given me such wonderful opportunity to learn more about Machine Learning and their mentors Shubham Yadav, Sajid Chaudhary for their constant encouragement, valuable suggestions, and cooperation.

This acknowledgement will remain incomplete if I fail to express my deep sense of obligation to my family members and God for their consistent blessings and believe in me.

Thank You.

**INTRODUCTION**

* Business Problem Framing

Car Price Prediction is a really an interesting machine learning problem as there are many factors that influence the price of a car in the second-hand market. In this competition, we will be looking at a dataset based on sale/purchase of cars where our end goal will be to predict the price of the used car given its features to maximize the profit.

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. With this project, we are scrapping new data from various car sale/buy portals as well fetching data of used cars from different locations.

As a Data scientist we are required to apply some data science techniques for the price of used cars with the available independent variables. That should help the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the requirement of the cars, the business strategy etc. to meet certain price levels and to have a new machine learning models from new data. We will make car price valuation model.

* Conceptual Background of the Domain Problem

The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in India.

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 to the lack of funds, used cars sales are on a global increase. 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. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling.

* Review of Literature

There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it’s a big corner to pay too much or sell less then it’s market value. 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. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling.

* Motivation for the Problem Undertaken

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. 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 use machine learning algorithms to develop models for predicting used car prices.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

There are websites that offers an estimate value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used car’s market value. The mathematical approach taken in this project is the analysis of used car prices which helps to understand the future prices. Understanding the different markets where used cars are sells and analytics help to get more information about prediction. To understand the factors affecting the pricing of cars in Indian market, since those may be very different from the other country’s market. The prices within India also varies a lot with another city.

* Data Sources and their formats

For this project, we are using the dataset on used car sales from all over the country, scrapped from various used cars selling websites from olx.com and cardekho.com. The features available in this dataset are Brand, Variant, Price, Model, Manufacturing Year, Fuel\_Type, Location and Distance (in km).

It contains most all relevant information that websites provide on car sales including columns like Brand, Variant, Price, Model, Manufacturing Year, and other categories. This dataset has 8 columns and 8163 rows.

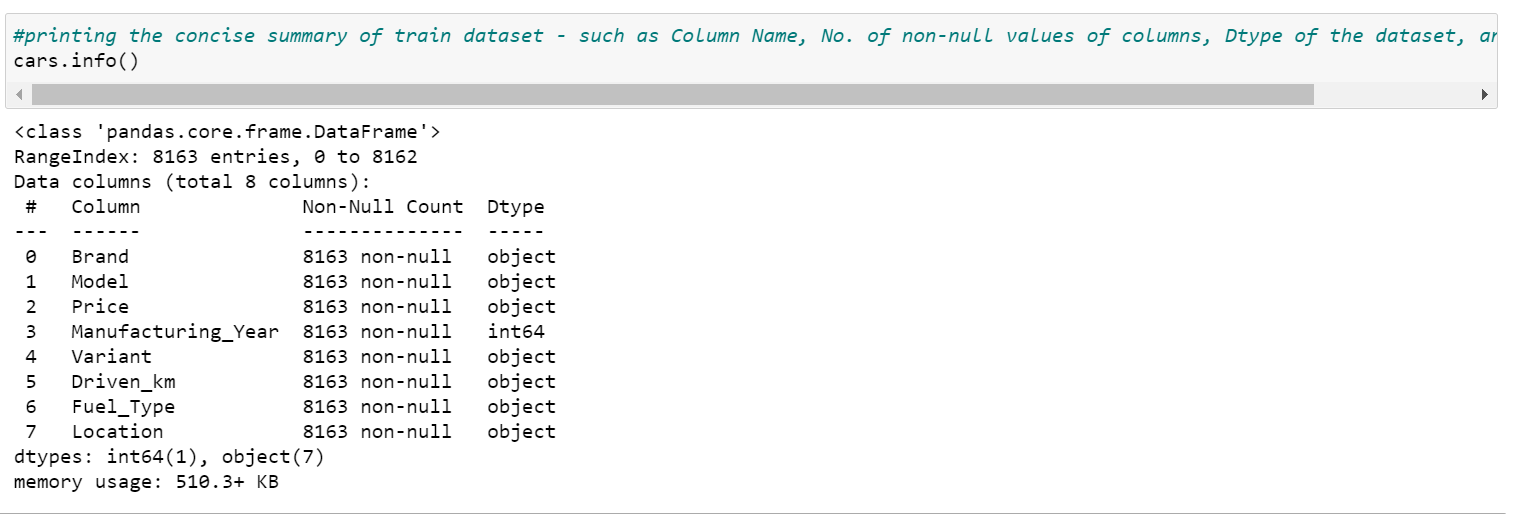
The column “**Price**” is the target variable and rest of the columns are independent variables.

The independent variables are again divided into Categorical and Numerical variables.

Here is the dataset scrapped for used cars from websites (Olx, cardekho) linked below:

[Car Price Scraped Data.xlsx](Car%20Price%20Scraped%20Data.xlsx)

**DataSet:**



* Data Preprocessing Done

Data pre-processing is very important to get the dataset into the best format before performing algorithm. This is very important step in Machine Learning that should not be skipped. It involves three stages: Data Cleaning, Data Transformation, and Feature Engineering which converts complicated dataset into quality data.

Data Cleaning:

In Data cleaning, we understand the data, we perform several activities to clean the data. This stage comprises of following activities:

* Dealing with Null Values – In this dataset, we found there are no null values.
* Data Transformation:

In Data transformation we recheck that all present columns are numerical or not? if not then we perform encoding type depending on the type of values present in columns.

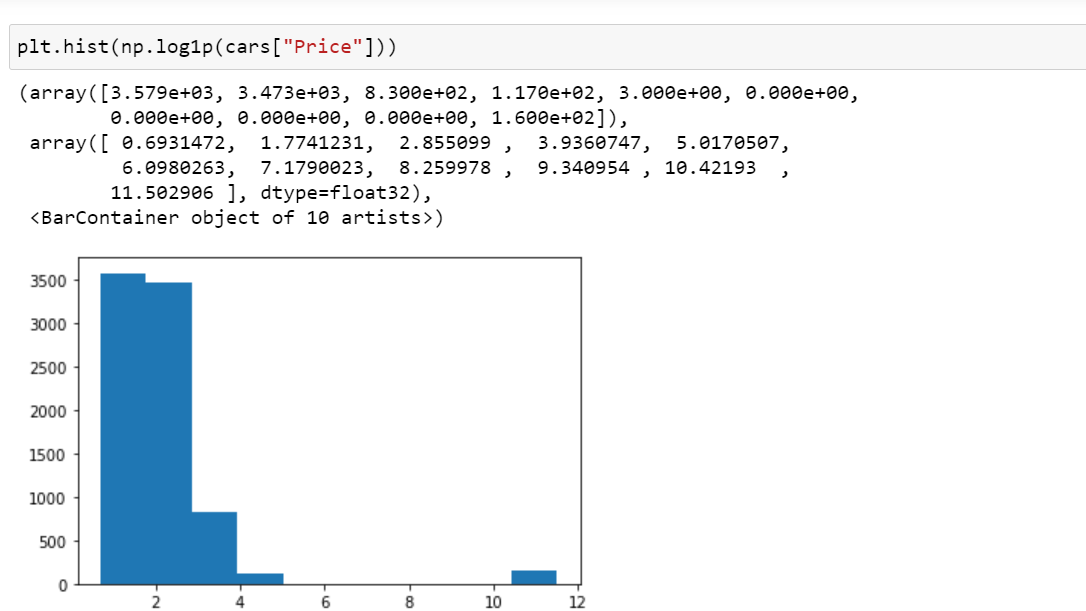
Here in this dataset, we have created dummy features to convert it in numerical.

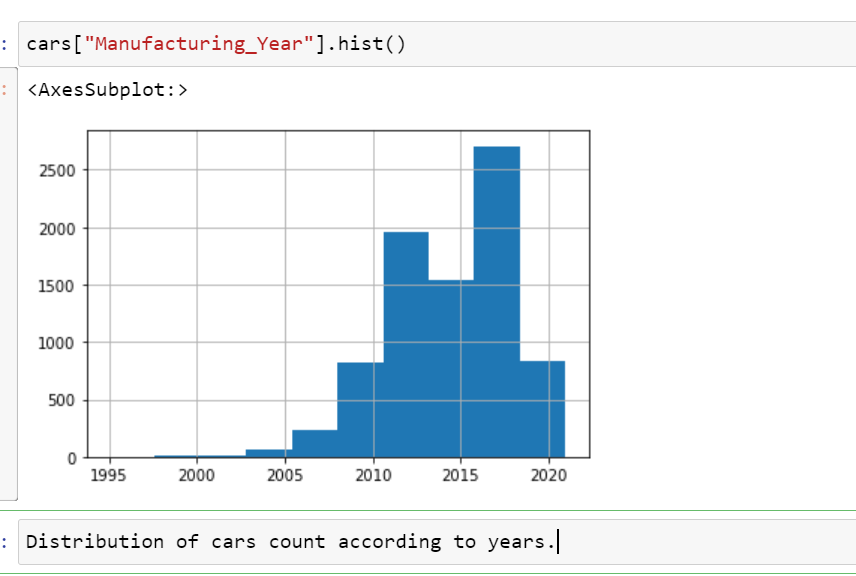
Most important thing we need to keep in mind that all data pre-processing steps will be performed according the dataset present. In order to apply machine learning models, we need numeric representation of the features. Therefore, all non-numeric features were transformed into numerical form.

* Data Inputs- Logic- Output Relationships

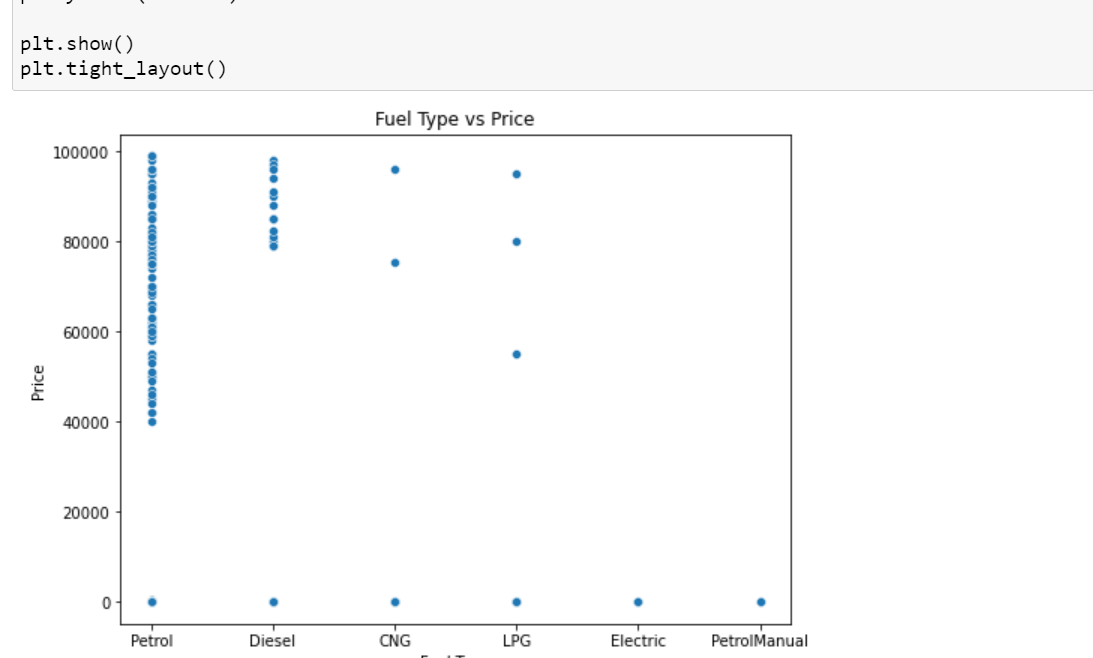
While exploring the data, we will look at the different combinations of features with the help of visuals. This will help us to understand our data better and give us some clue about pattern in data. State the set of assumptions (if any) related to the problem under consideration.

Price is the feature that we are predicting in this study. Before applying any models, taking a look at price data may give us some ideas :





Relationship with Fuel Type



* Hardware and Software Requirements and Tools Used

Machine learning comes with an extensive collection of ML tools, platforms, and software.

Well, in this problem solution we have used following tools that helps to make this project successful as per my possibility.

Jupyter notebook is one of the most widely used machine learning tools among all. It is a very fast processing as well as an efficient platform. Moreover, for this problem solution I have used python programming.

Scikit-Learn is built on top of the three main Python libraries viz. NumPy, Matplotlib, and SciPy. Along with this, it will also help you with testing as well as training your models.

The Libraries are as listed:

1. For Data loading and Visualisation:

Numpy - NumPy is very useful for handling linear algebra, Fourier transforms, and random numbers.

Pandas - Pandas are turning up to be the most popular Python library that is used for data analysis with support for fast, flexible, and expressive data structures designed to work on both “relational” or “labeled” data.

Matplotlib - The library helps to generate histograms, plots, error charts, scatter plots, bar charts with just a few lines of code.

Seaborn – Used for visualization.

For Normalization:

Min-max scaler is the standard approach for scaling. We use this library to balance the dataset and make it normal for further algorithm performance.

train\_test\_split: We use it to perform test train spliting.

Algorithm Libraries:

In this problem solution we have used algorithm libraries:

RandomForestRegressor – for regression algorithm

*from sklearn.ensemble import RandomForestRegressor*

Linear Regression– For Linear Regressor algorithm

*from sklearn import linear\_model*

*model = linear\_model.LinearRegression()*

GradientBoostRegressor – To perform gradient Boost Regressor algorithm

*lr = ensemble.GradientBoostingRegressor()*

Root Mean Square Error-

*from sklearn.metrics import mean\_squared\_error*

RandomizedSearchCV – For hyper parameter tuning performance.

*from sklearn.model\_selection import RandomizedSearchCV*

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

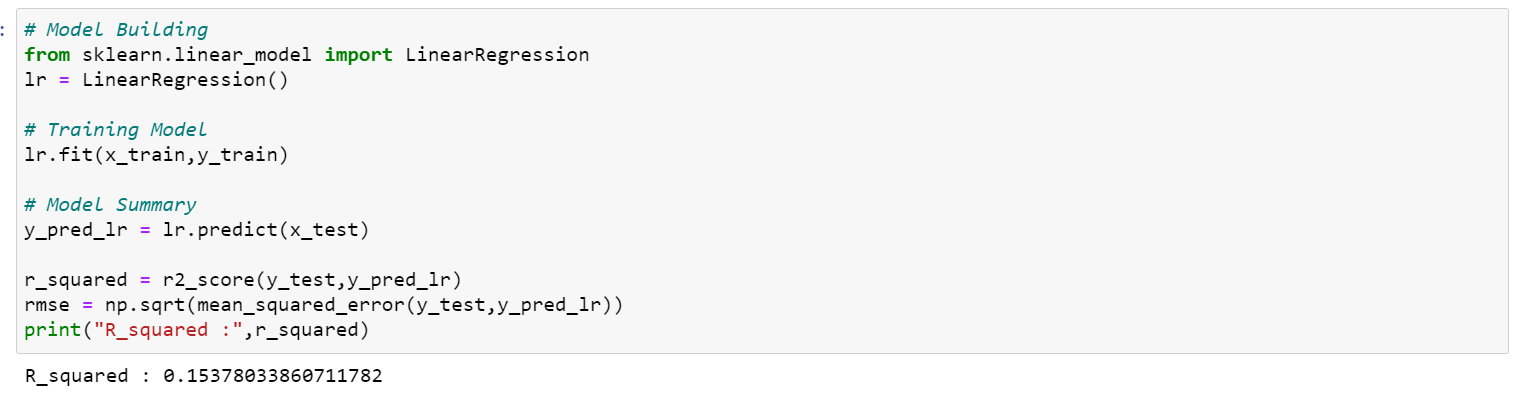
We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. Linear Regression, Random Forest and Gradient Boost were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package was used.

* Testing of Identified Approaches (Algorithms)

This section used applied machine learning models as a framework for the data analysis. The data set is a supervised data which refers to fitting a model of dependent variables to the independent variables, with the goal of accurately predicting the dependent variable for future observations or understanding the relationship between the variables.

In this section, these machine learning models we have be applied following Regression techniques in order:

1. Linear Regression : Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors. No regularization was used since the results clearly showed low variance.



1. Random Forest : Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees.

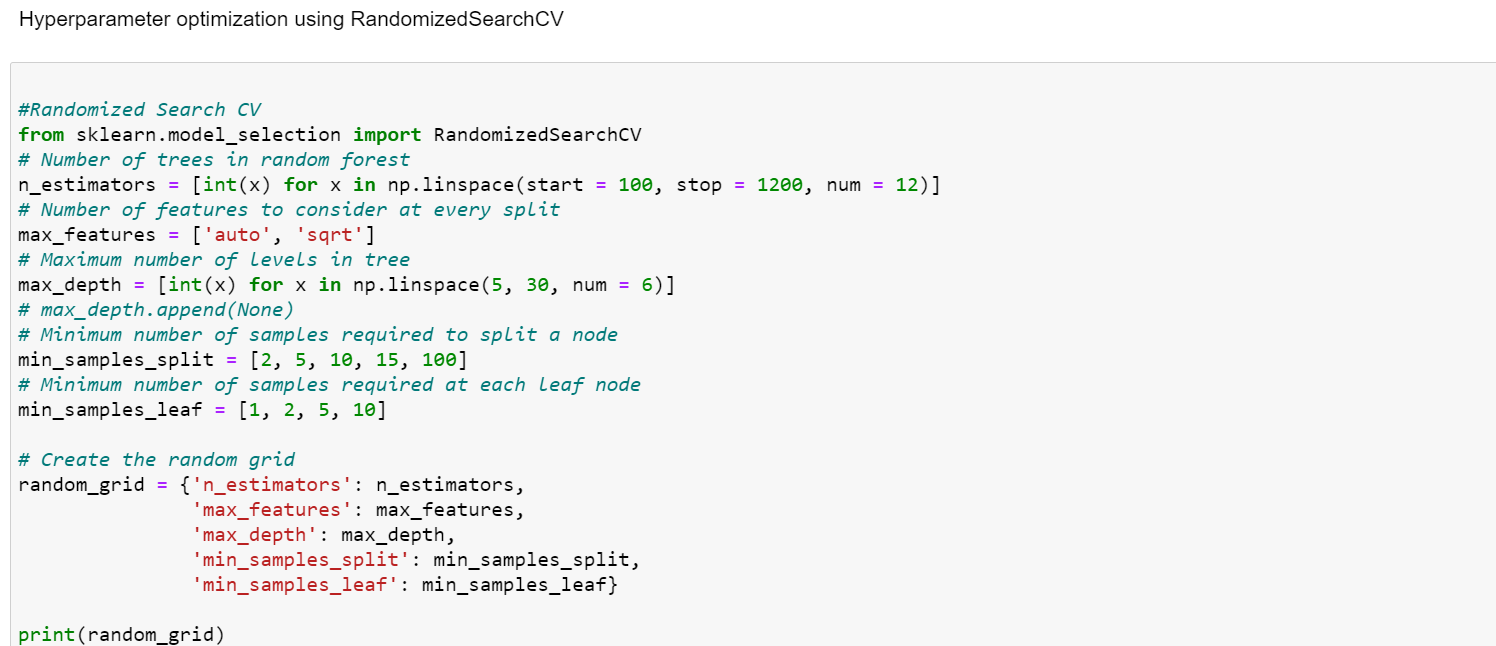


1. Gradient Boost: Gradient Boosting is another decision tree-based method that is generally described as “a method of transforming weak learners into strong learners”. This means that like a typical boosting method, observations are assigned different weights and based on certain metrics, the weights of difficult to predict observations are increased and then fed into another tree to be trained. In this case the metric is the gradient of the loss function. This model was chosen to account for non-linear relationships between the features and predicted price, by splitting the data into 100 regions.



* Run and Evaluate selected models

Random Forest method which marginally outperforms Linear Regression. However Random Forests tend to overfit the dataset due to the tendency of growing longer trees. This was worked upon by restricting the depth of trees to different values and it was observed that beyond limiting depth to 36 resulted in negligible improvement in prediction performance but progressively increased overfitting. As expected GBM performed marginally better than Linear Regression but had a significantly faster training time.



* Key Metrics for success in solving problem under consideration

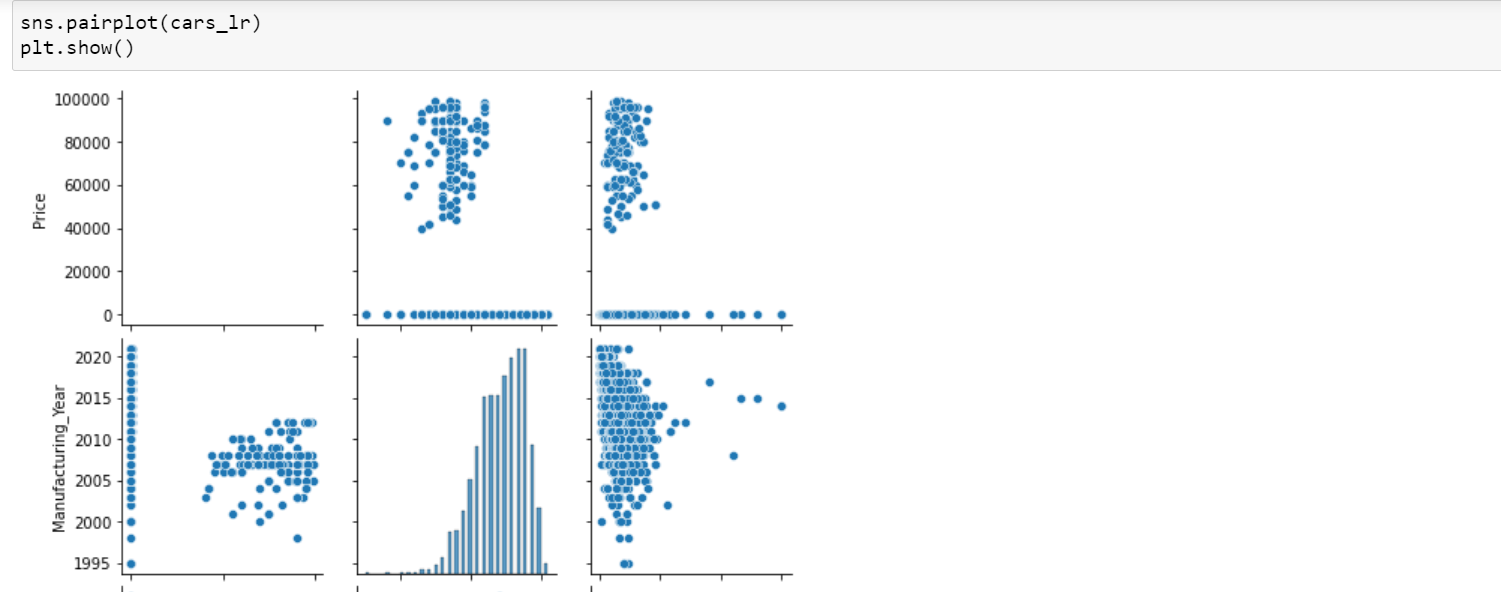
For each one of the techniques mentioned in the previous section (Linear Regression, Random Forest Regressor, Gradient Boost Regressor, etc.), we will follow these steps to build a model:

* Choose an algorithm that implements the corresponding technique
* Search for an effective parameter combination for the chosen algorithm
* Create a model using the found parameters
* Train (fit) the model on the training dataset
* Visualizations

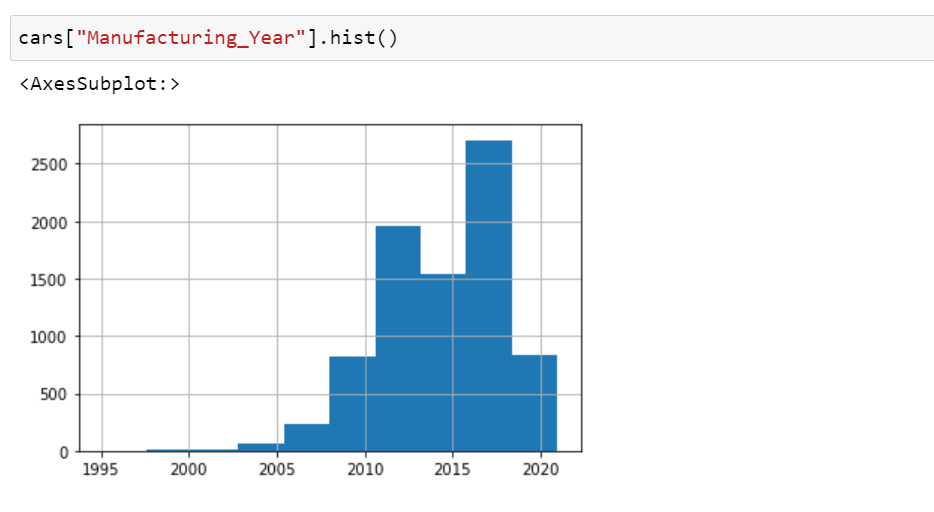
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

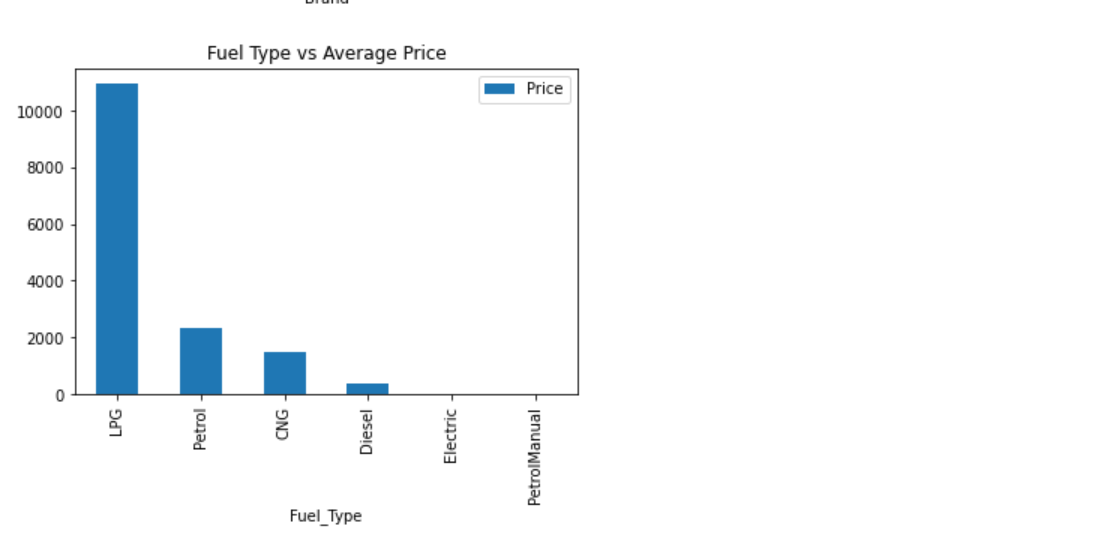
* Pairplot: Analysing the relationship between target feature Price with all other significant features.



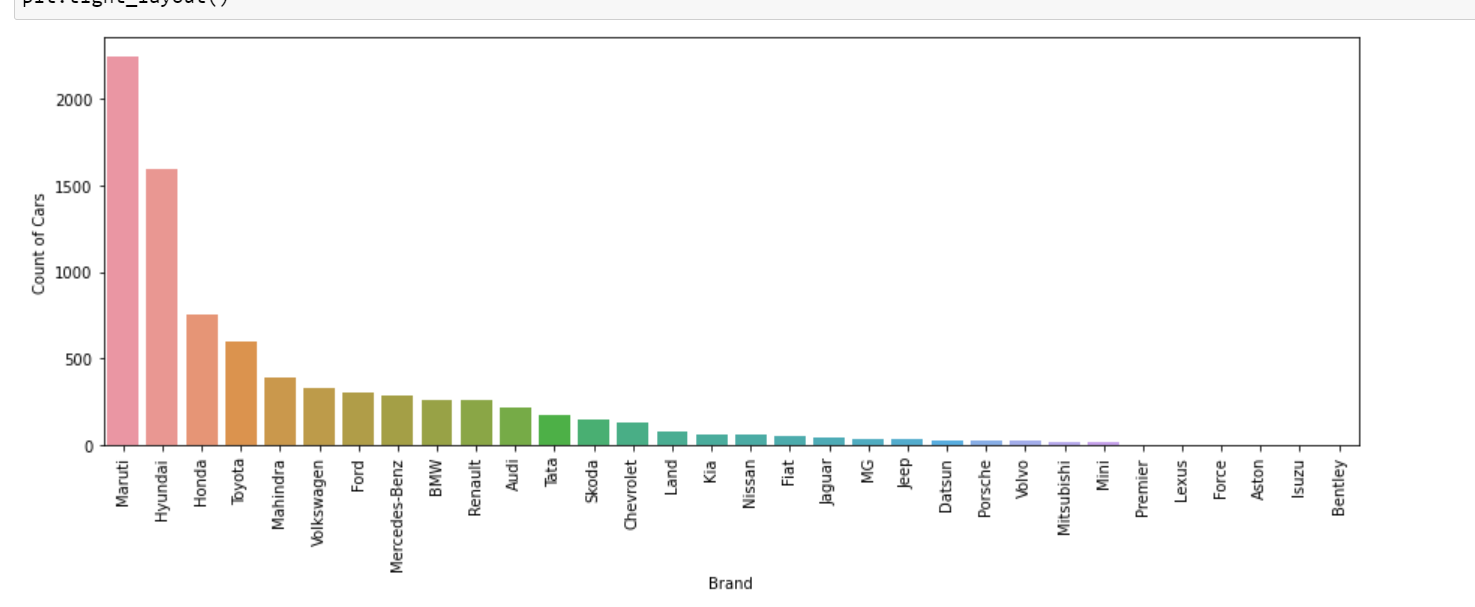
-Hist Plot

Analysing manufacturing year relationship by count histogram plot.

* FuelType analyses with price of cars.

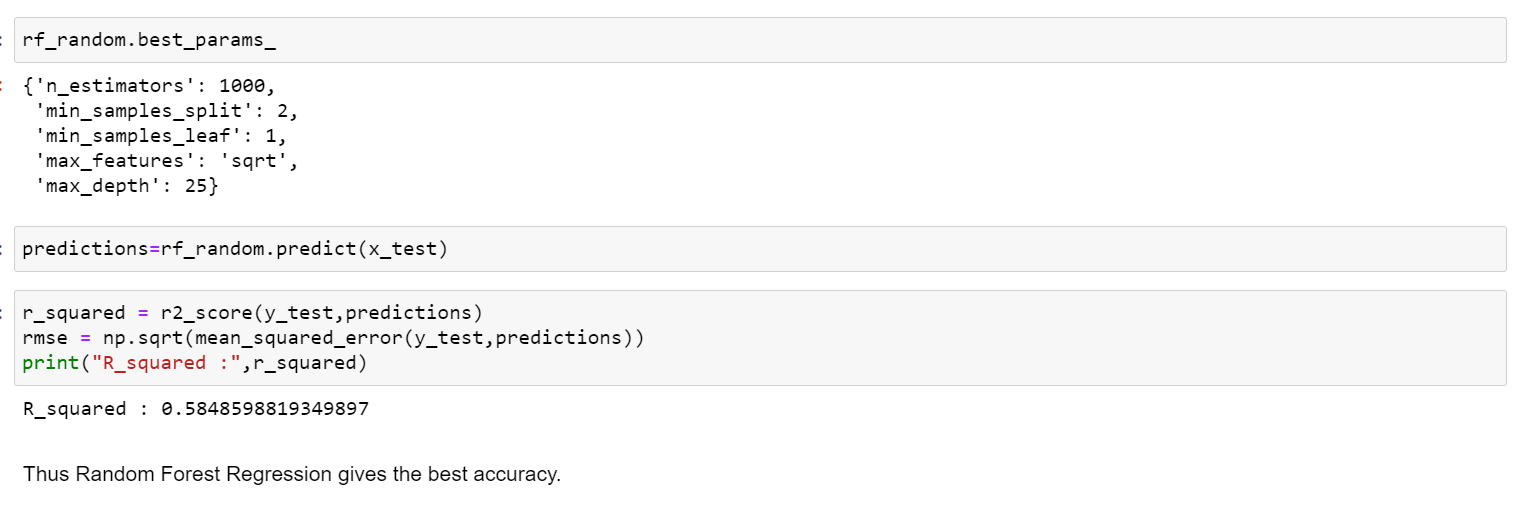


* Car Brand Analysis

Analysing car brand count.

* Interpretation of the Results

Compared to Linear Regression, most Decision-Tree based methods did not perform comparably well. This can be attributed to the apparent linearity of the dataset. We believe that It can also be attributed to the difficulty in tuning the hyperparameters for most gradient boost methods. The exception to this is the Random Forest method which marginally outperforms Linear Regression..



**CONCLUSION**

* Key Findings and Conclusions of the Study

By performing different models, it was aimed to get different perspectives and eventually compared their performance. With this study, it purpose was to predict prices of used cars by using a dataset that has predictors and multiple observations. With the help of the data visualizations and exploratory data analysis, the dataset was uncovered, and features were explored deeply. The relation between features were examined. At the last stage, predictive models were applied to predict price of cars: random forest, linear regression, , Gradient Boost.

* Learning Outcomes of the Study in respect of Data Science

To learn data pre-processing and preparation techniques in order to obtain clean data.

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.

* Limitations of this work and Scope for Future Work

This study used different models to predict used car prices. However, there was a relatively small dataset for making a strong inference because number of observations was only around 8000. Gathering more data can yield more robust predictions. Secondly, there could be more features that can be good predictors. For example, here are some variables that might improve the model: number of doors, gas/mile (per gallon), color, mechanical and cosmetic reconditioning time, used-to-new ratio, appraisal-to-trade ratio.

Another point that that has room to improvement is that data cleaning process can be dome more rigorously with the help of more technical information.

Thank You.