12-2021

Used Cars Price Prediction and Valuation using Data Mining Techniques

Abdulla AlShared
aoa4785@rit.edu

Used Cars Price Prediction and Valuation using Data Mining Techniques

by

Abdulla AlShared

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
DEC 2021

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Abdulla AlShared**

Graduate Capstone Title**: Used Cars Price Prediction and Valuation using Data Mining Techniques**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak Date:

Chair of committee

Name: Dr. Ehsan Warriach Date:

Member of committee

Acknowledgments

I would like to extend my thanks and appreciation to Dr. Erhan Warriach, my mentor, for his continuous support and guidance during the capstone project, we can also never forget the efforts of all the professors that taught me during the Data Analytics program and guided me though the world of data, which was a new realm for me.

Moreover, I would like to extend my gratitude to Dr. Sanjay Modak for his guidance and patience with us during the capstone project.

Abstract

Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend more purchase used cars. A primary objective of this project is to estimate used car prices by using attributes that are highly correlated with a label (Price). To accomplish this, data mining technology has been employed. Null, redundant, and missing values were removed from the dataset during pre-processing. In this supervised learning study, three regressors (Random Forest Regressor, Linear Regression, and Bagging Regressor) have been trained, tested, and compared against a benchmark dataset. Among all the experiments, the Random Forest Regressor had the highest score at 95%, followed by 0.025 MSE, 0.0008 MAE, and 0.0378 RMSE respectively. In addition to Random Forest Regression, Bagging Regression performed well with an 88% score, followed by Linear Regression having an 85% mark. A train-test split of 80/20 with 40 random states was used in all experiments. The researchers of this project anticipate that in the near future, the most sophisticated algorithm is used for making predictions, and then the model will be integrated into a mobile app or web page for the general public to use.

Keywords: Car Price Prediction, supervised learning, linear regression, bagging regression, classification.

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Background

Today, the transportation industry is considered to be one of the backbones of the economy. Automobiles are referred to as the "Industry of Industries" in developed nations. According to industry professionals, the UAE's automotive industry has seen remarkable growth. Besides being the fastest-growing nation in the automobile industry, it represents its global presence. In Dubai, like most other countries, cars are gaining a great deal of popularity among the local population and the ex-pat community who work in the country. There are used cars for sale in the UAE of all makes and models, even cars from well-known brands (Rizvi, 2019).

UAE's auto industry is experiencing constant growth, registered at 27%, with a total industry volume (TIV) of 310,403 cars. Approximately 1.49 million units were sold within the Gulf Cooperation Council (GCC). Compared to the global market, the Gulf Cooperation Council countries are growing at 10% in 2021 (Research, 2020). So far, the market in the UAE has grown by 19%. It is thus the world's largest market in terms of growth rate.

Almost everyone wants their own car these days, but because of factors like affordability or economic conditions, many prefer to opt for pre-owned cars. Accurately predicting used car prices requires expert knowledge due to the nature of their dependence on a variety of factors and features. Used car prices are not constant in the market, both buyers and sellers need an intelligent system that will allow them to predict the correct price efficiently. In this intelligent system, the most difficult problem is the collection of the dataset which contains all important elements like the manufacturing year of the car, its gas type, its condition, miles driven, horsepower, doors, number of times a car has been painted, customer reviews, the weight of the car, etc. It is clear that the price of the product is affected by many factors, but unfortunately, information about these features is not always readily available. Since this project primarily focuses on the Dubai market, the benchmark dataset containing all key features is scraped.

It is necessary to pre-process and transform collected data in the proper format prior to feeding it directly to the data mining model. As a first step, the dataset was statistically analyzed and plotted. Missing, duplicated, and null values were identified and dealt with. Features were chosen and extracted using correlation matrices. To build an efficient model, the most correlated features were retained, and others were discarded. This prediction problem can be considered a regression problem since it belongs to the supervised learning domain. Three Regressor known as random forest, linear regression, and bagging regression were trained and compared. A random forest Regressor outperformed all others in this project, so it was chosen as the main algorithm model.

1.2 Statement of problem

The research objective of this study is to predict used cars prices in Dubai using data mining techniques, by scraping data from websites that sell used cars, and analysing the different aspects and factors that lead to the actual used car price valuation. To enable consumers to know the actual worth of their car or desired car, by simply providing the program with a set of attributes from the desired car to predict the car price.

The purpose of this study is to understand and evaluate used car prices in the UAE, and to develop a strategy that utilizes data mining techniques to predict used car prices.

1.3 Project goals

This project aims to deliver price prediction models to the public, to help guide the individuals looking to buy or sell cars and to give them a better insight into the automotive sector. Buying a used car from a dealer can be a frustrating and an unsatisfying experience as some dealers are known to deploy deceitful sale tactics to close a deal. Therefore, to help consumers avoid falling victims to such tactics, this study hopes to equip consumers with right tools to guide them in their shopping experience.

Another goal of the project is to explore new methods to evaluate used cars prices and to compare their accuracies. Considering this is an interesting research topic in the research community, and in continuing their footsteps, we hope to achieve significant results using more advanced methods of previous work.

1.4 Methodology

The project deals with UAE used cars. Using Parse Hub, the benchmark dataset from dubizzle.ae and buyanycar.com was scraped in order to build the effective intelligent model. The project's methodology is as follows:

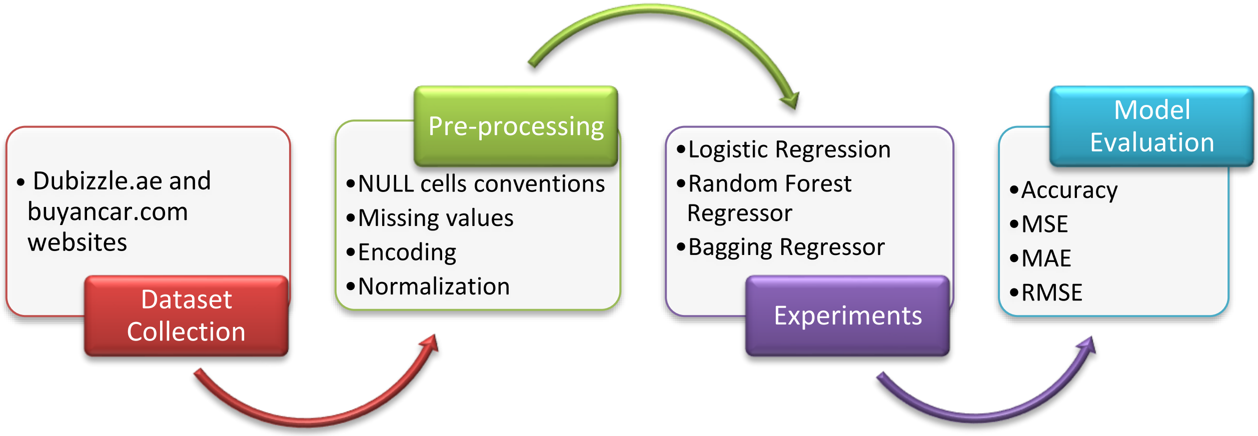


Figure 1 Proposed Methodology

After data collection the dataset was pre-processed to remove samples that have missing value, and remove non-numerical part from numerical attributes, converting categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that doesn't affect the price evaluations if needed to reduce the complexity of the model.

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

Afterwards when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict prices and values of used cars. In this study three models are proposed to be built using Logistic Regression model technique, Random Forest Regressor and Bagging Regressor. Firstly, the data was portioned into section for training and the other part for testing, portioning percentage can be tested with different ratios to analyse different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed.

1.5 Limitations of the Study

In the past year the world of automobiles has seen a drastic change with the semiconductor shortages after the pandemic, which led to spike in used car prices. Hence, there was fast change in car prices during this study which will affect the actual car pricing prediction future. As the current dataset will undervalue the cars in the market. Therefore, a model that is built on real time data can be best integrated into a mobile app for public use would be the idea solution.

Chapter 2 – Literature Review

Several studies and related works have been done previously to predict used car prices around the world using different methodologies and approaches, with varying results of accuracy from 50% to 90%. In (Pudaruth, 2014) the researcher proposed to predict used car prices in Mauritius, where he applied different machine learning techniques to achieve his results like decision tree, K-nearest neighbours, Multiple Regression and Naïve Bayes algorithms to predict the used cars prices, based on historical data gathered from the newspaper.

Achieved results ranged from accuracy of 60-70 percent, the author suggested using more sophisticated models and algorithms to make the evaluation, with the main weakness off the decision tree and naïve Bayes that it is required to discretize the price and classify it which accrue to more inaccuracies. Moreover, he suggested a larger set of data of data to train the models hence the data gathered was not sufficient.

(Monburinon, et al., 2018) Gathered data from a German e-commerce site that totalled to 304,133 rows and 11 attributes to predict the prices of used car using different techniques and measured their results using Mean Absolute Error (MEA) to compare their results. Same training dataset and testing dataset was given to each model. Highest results achieved was by using gradient boosted regression tree with a MAE of 0.28, and MEA of 0.35 and 0.55 for mean absolute error and multiple linear regression respectively. Authors suggested adjusting the parameters in future works to yield better results, as well as using one hot encoding instead of label encoding for more realistic data interpretations on categorical data.

(Gegic, Isakovic, Keco, Masetic, & Kevric, 2019) from the International Burch University in Sarajevo, used three different machine learning techniques to predict used car prices. Using data scrapped from a local Bosnian website for used cars totalled at 797 car samples after pre-processing, and proposed using these methods: Support Vector Machine, Random Forest and Artificial Neural network. Results have shown using only one machine learning algorithm achieved results less than 50%, whereas after combing the algorithms with pre calcification of prices using Random Forest, results with accuracies up to 87.38% was recorded.

(Noor & Jan, 2017) were able to achieve high level of accuracy using Multiple linear regression models to predict the price of cars collected from used cars website in Pakistan called Pak Wheels that totalled to 1699 records after pre-processing, and where able to achieve accuracy of 98%, this was done after reducing the total amount of attributes using variable selection technique to include significant attributes only and to reduce the complexity of the model.

(K.Samruddhi & Kumar, 2020) Proposed using Supervised machine leaning model using K-Nearest Neighbour to predict used car prices from a data set obtained from Kaggle containing 14 different attributes, using this method accuracy reached up to 85% after different values of K as well as Changing the percent of training data to testing data, expectedly when increasing the percent of data that is tested better accuracy results are achieved. The model was also cross validated with 5 and 10 folds by using K fold method.

(Gongqi, Yansong, & Qiang, 2011) proposed using Artificial Neural Network (ANN) through a combined method of BP neural network and nonlinear curve fit and have achieved accurate value prediction with a feasible model.

(Listiani, 2009) used Support Vector Machines to evaluate leased cars prices, results have shown that SVM is far more accurate in large dataset with high dimensional data than Multiple linear regression. Whereas the computation Multiple linear regression can take several minutes and the SVM would take up to a day to compute the results. Multiple linear regression may be simple, but SVM is far more accurate. Moreover, the study includes Samples with up to 178 attributes which is far more than the proposed variable in our study, hence the use of multiple linear regression may be more suitable in our case.

(Kuiper, 2008) Collected data from General Motor of cars that are produced in 2005, where he as well used variable selection technique to include the most relevant attributes in his model to reduce the complexity of the data. He proposed used Multivariate regression model that would be more suitable for values with numeric format.

In order to predict the price of used cars, researchers (Nabarun Pal, 2018) used a supervised learning method known as Random Forest. Kaggle's dataset was used as a basis for predicting used car prices. In order to determine the price impact of each feature, careful exploratory data analysis was performed. 500 Decision Trees were trained with Random Forests. It is most commonly used for classification, but they turned it into a regression model by transforming the problem into an equivalent regression problem. Using experimental results, it was found that training accuracy was 95.82%, and testing accuracy was 83.63%. By selecting the most correlated features, the model can accurately predict the car price.

In light of the number of works that have been done in this field, another group of researchers (Jian Da Wu, 2017) conducted research on this topic and tried to develop a system that consists of three components: a data acquisition system, a price forecasting algorithm, and a performance analysis. Due to its adaptive learning capability, a conventional artificial neural network (ANN) with a back-propagation network is compared to the proposed ANFIS. In the ANFIS, qualitative fuzzy logic approximation as well as adaptive neural network capabilities are included. Using ANFIS as an expert system in predicting used car prices showed better results in the experiment. Using GUI, the consumer can get accurate and convenient information about used cars' purchasing prices, and experiments proved that the proposed system could provide accurate and convenient price forecasting.

Hence, from all literature review it is concluded that used cars price prediction is an important topic which is the area of many researchers nowadays. So far, the best achieved accuracy is 83.63% on kaggle’s dataset using random forest technique. The researchers have tested multiple regressors and final model is regression model using linear regression.

Chapter 3 – Project Description

3.1 UAE Used Cars

Second-hand car market in UAE is in exponential growth year over year, with rise of individual cars increasing more people are looking into getting their own can with an affordable price, usually looking in the used cars market, "For every new car sold in the UAE in 2019, about 3.5 used cars were sold" (BIELSKI & RAMARATHNAM, 2020). In the coming years the industry will be revolutionized, where markets will be powered by digitalization and new business models to enhance focus efficiency and consumer needs.

Used Car Prices have been on the rise in recent years. Prices increased by 4-10% from 2018 to 2020, and yet dealers are still selling more cars in a shorter time span. Furthermore, Analysts have shown that depreciation costs of cars have only been decreasing in the UAE in a previously unseen phenomena leading to surge in second hand cars demand (Bridge, 2020).

Manheim Used vehicle index have seen an unprecedented increase in used cars value as it has increased 6.81% in the first 15 days of April 2021 compared to the month of March. As well as seen a 52.2% increase in the value compared to April 2020, and the latest trend indicators suggest seeing few more weeks of appreciation. This is caused by stimulus payments and tax refunds due to Covid-19 ramifications, and a decrease in car production globally. (Used Vehicle Value Index, 2021)

Price variations are common and misleading prices on every other website, hence the need for a tool to predict the pricing of used cars based on real data gathered from local websites to give accurate evaluation of cars for consumers. This study will develop a simple interface accurate enough for consumers to evaluate car pricing either for selling or purchasing purposes.

3.2 Machine learning

The goal of machine learning (ML) is to help a computer learn without being explicitly instructed to do so by means of mathematical models of data. Artificial intelligence (AI) is a subset of machine learning. Data is analysed using algorithms to identify patterns, which are then used to create predictive models. Like humans, machine learning becomes more accurate with more data and experience.

With machine learning, you can adapt to situations where data is constantly changing, the nature of the request or task is shifting, or coding a solution isn't feasible.

Three important categories of machine learning are:

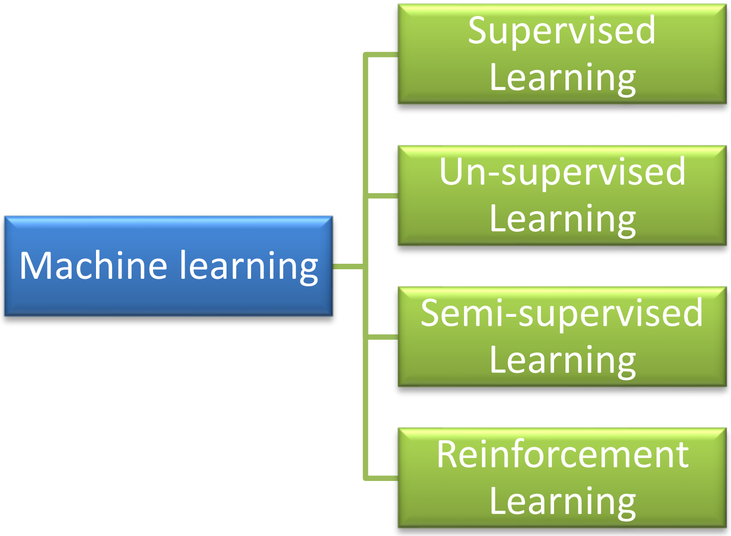


Figure 2 Machine Learning Categories

Supervised and un-supervised learning are commonly used types while reinforcement is sequential decision maker technique. Till date, machine cannot take decision without training (Matthew Botvinick, May 2019). Main categories of supervised and un-supervised machine learning are:

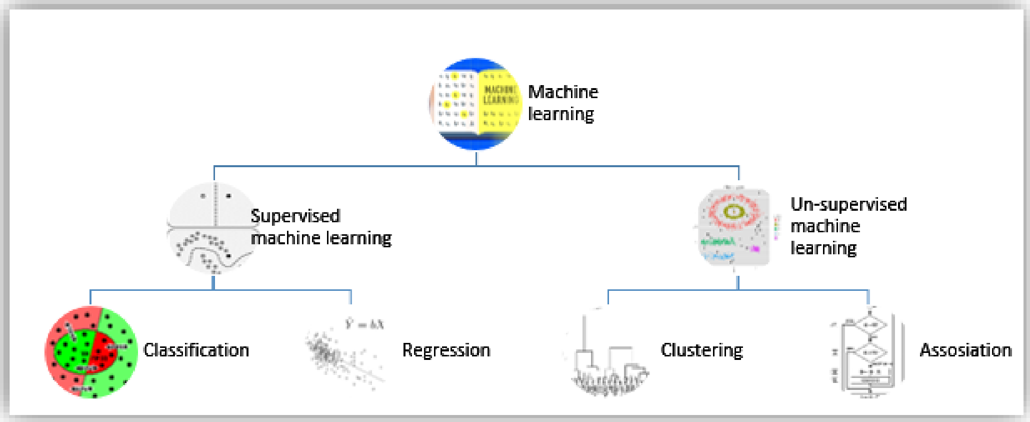


Figure 3 Machine Learning Techniques

The word supervised learning comes from the word supervisor, which means teacher. In this case, the label class will be categorized or predicted. During algorithm training, the correct answers have already been tagged with the corresponding class labels. Support Vector Machines (SVMs), Random Forest Trees, and Decision Trees are commonly used algorithms for supervised machine learning (Singh, Thakur, & Sharma, 2016). Unsupervised learning occurs when input data have no class labels. In order to learn more about the data, we want to model the data's underlying structure. There are two main types: association and clustering (Ceriottia, 2019). Famous un-supervised algorithms are K-means clustering, affinity propagation etc.

3.2.1 Supervised Learning

Working and details of some famous supervised machine learning algorithms which are used in this project are:

Logistic Regression:

Whenever the dependent variable is non-numerical (categorical) and the class should be predicted, not classified, the logistic regression algorithm needs to be abandoned. Machine learning technique Logistic Regression is commonly used to classify binary data. The function of Logistic Regression is to optimize results based on various datasets (Swaminathan, 2018). To predict results, the default label class is always employed, but the results and probability are always calculated after all categorical values have been converted into numerical values and all data has been normalized.

Logistic regression, also referred to as sigmoid regression, was designed by statisticians to explain the properties of the population increasing in the ecological study, growing fast, and maxing out on the capability to wear out the surroundings. With an S-shaped curve, any real-valued range can be mapped right into a number between zero and 1, but not precisely at the limit of 1.

$$\frac{1}{1+e^{-x}}$$

Equation 1 Sigmoid equation

Where e is the base of the logarithms (Euler's wide variety or the EXP() characteristic on your spreadsheet) and price is the real numerical price which needs to be transformed.

While the equation of regression in which intercept and slope are integrated is as follows:

$$y = mx + c$$

Equation 2 Regression equation

Below is a generalized equation for Multivariate regression model:

$$\mathbf{y = \beta0 + \beta1.x1 + \beta2.x2 + \ldots.. + \beta n.xn}$$

There are few steps involved in generating the regression beginning with feature selection, normalizing features, select loss function and hypothesis, set hypothesis parameters, and minimizing the loss function, and finally testing the function of the data.

Random Forest Regressor:

Random Forest is already revealing that it creates forest and then somehow randomizes it. It builds the forest through the ensemble of Decision Trees and most of the time trains it using a method called the Bagging Method. Since it uses the ensemble method, the result is improved. Decision tree and bagging classifier hypermeters are the same. Each feature in the tree can be made random simply by adding thresholds.

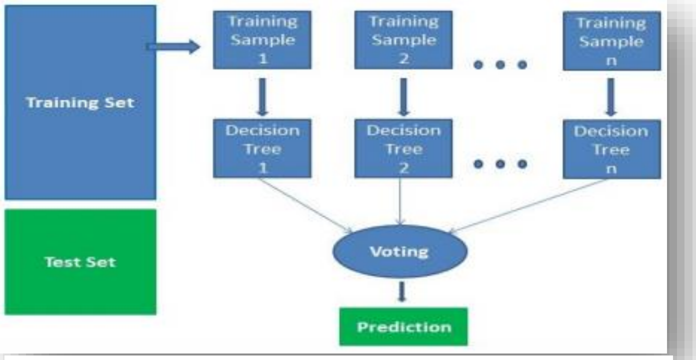


Figure 4 Random Forest Regressor

The following steps enable us to understand how Random Forest operates:

1. First, choose random samples from the given dataset.
2. Following that, each sample will be given a choice tree. Based on those choices, it will get a prediction end result.
3. For each anticipated outcome, voting may occur in this step.
4. In the end, choose the prediction outcome with the most votes because it is the very last prediction outcome.

The fig-4, shows a complete working.

3.2.2 Un-supervised learning

An unsupervised learning algorithm is trained on information that has neither been classified nor labelled, allowing it to act unsupervised on that information. Using this algorithm, unsorted data is grouped using patterns, resemblances, and differences without any prior training.

Unlike supervised learning, the algorithm does not receive any instruction from a trainer. The algorithm, therefore, focuses on discovering the hidden pattern in unlabelled facts by ourselves.

3.3 Dataset description

Data was collected and Scrapped from a website that sell car in the UAE called BuyAnyCar, by using a scrapping tool available for use only called ParseHub, through multiple runs and irritations 21,277 rows of data with 11 variables were collected successfully from the website.

At first, the data types of each attribute were corrected/converted by performing pre-processing on each attribute individually. Details and description of dataset is described in table below:

Sr#	Column Name	Datatype	Unique Values	Mode/ Mean of column
1	Car Name	Object	4009	Mitsubishi Pajero 2016

2	Brand	Object	82	Nissan
3	Model	Object	725	Pajero
4	Production Year	Int	48	2012
5	Body Type	Object	20	Sedan
6	Color	Object	14	White
7	Gear Type	Object	2	Automatic
8	Engine Size	Float	150	2.00 L
9	Mileage	Int	20503	146475.39
10	Specs Origin	Object	5	GCC
11	Fuel Type	Object	4	Petrol
12	Price	Int	684	61301.55

Table 1 Dataset Description

Based on the above information, it can be seen that the dataset has many categorical features that need to be converted to integers or floats. Furthermore, redundant samples need to be removed. As a result, pre-processing is required.

3.4 Exploratory Data Analysis (EDA)

Due to its combination of Model and Year columns, **Car Name** was a redundant attribute, and it was removed. Exploratory Data Analysis composed of following steps:

1. To visualize the missing values, the dataset was plotted using missing matrix as well as missing value heap map.

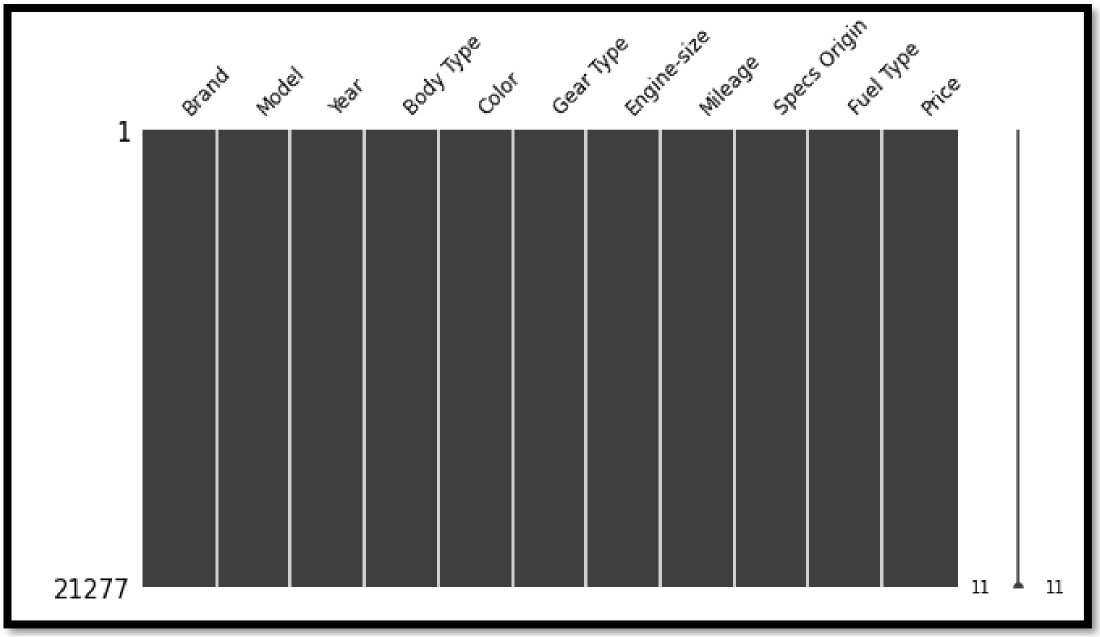


Figure 5 Missing Values Matrix

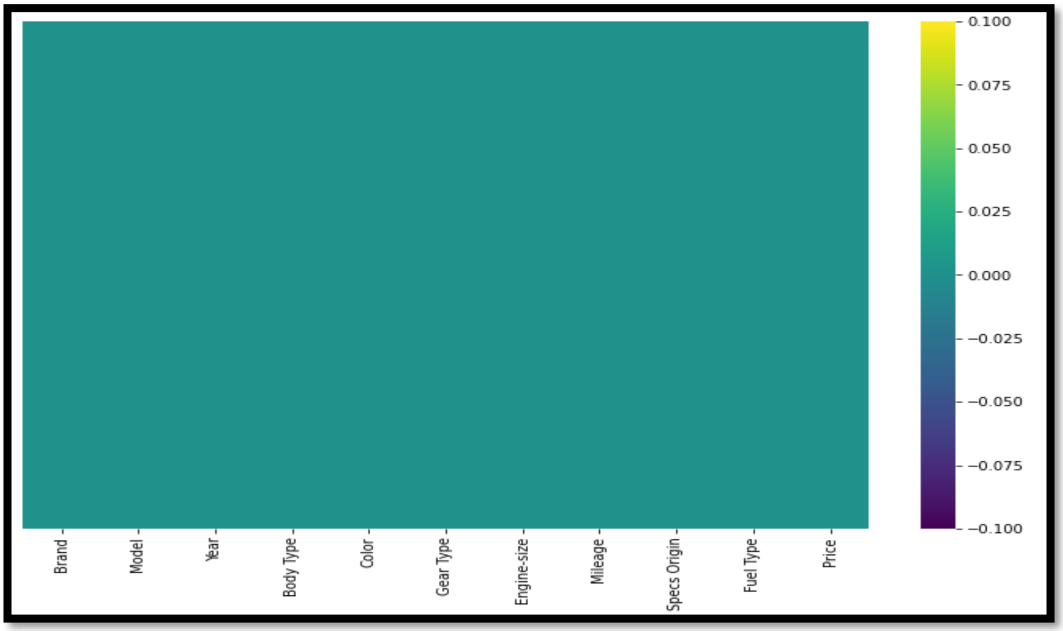


Figure 6 Missing Values - Heat Map

From fig-5 and fig-6, it is concluded that the dataset doesn't have any sample which contain any null/missing value.

2. Box plot is used to identify the outliers of integer attribute. Hence, **Year** attribute is visualized.

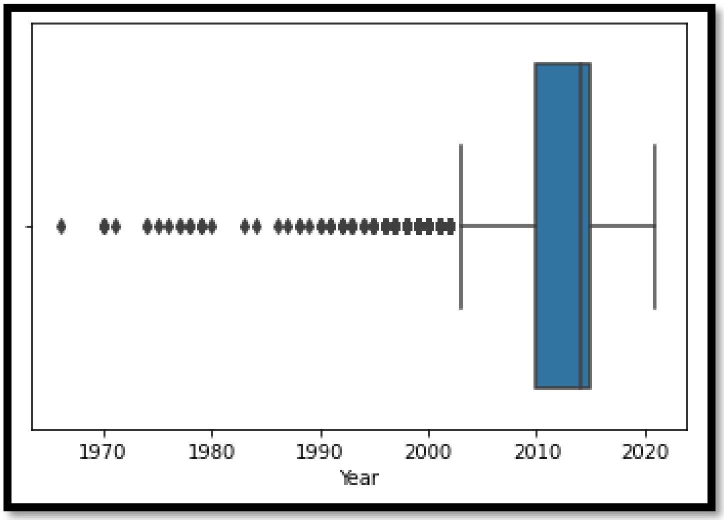


Figure 7 Year - Box Plot

As shown in fig-7, most of the data lies in the range of year 2004-2021. We decided to keep the outliers to have view of all the different classic cars.

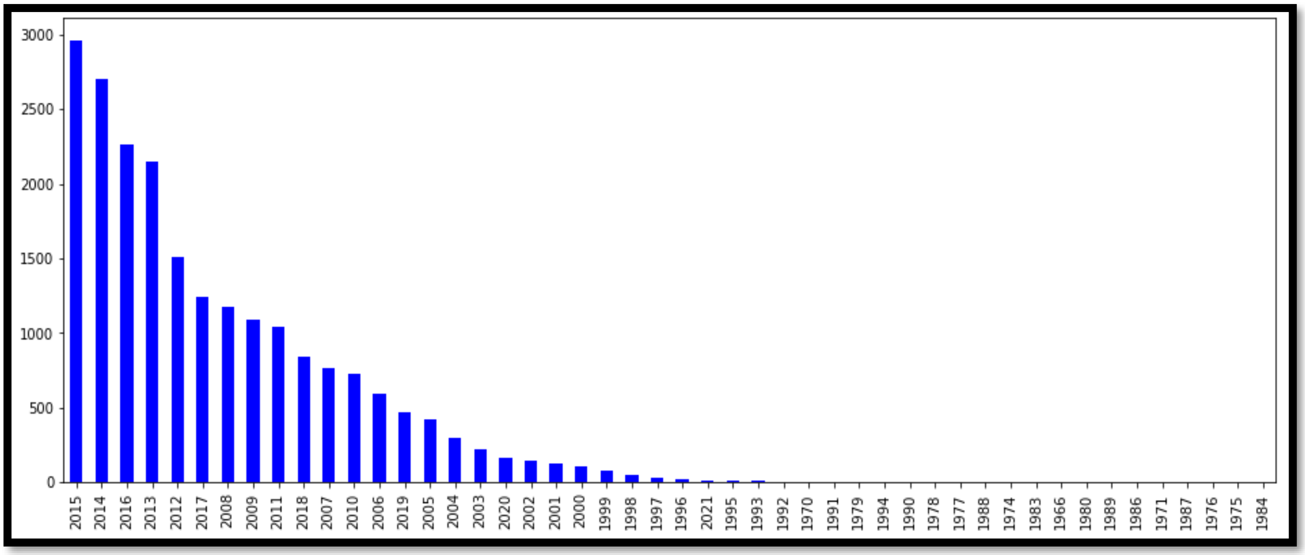


Figure 8 Year - Bar Plot

As seen in fig-8, the mostly samples were collected in year 2015 and the least were in 1993.

3. From fig-9, it can be visualized that Nissan is the most occurred Brand of dataset.

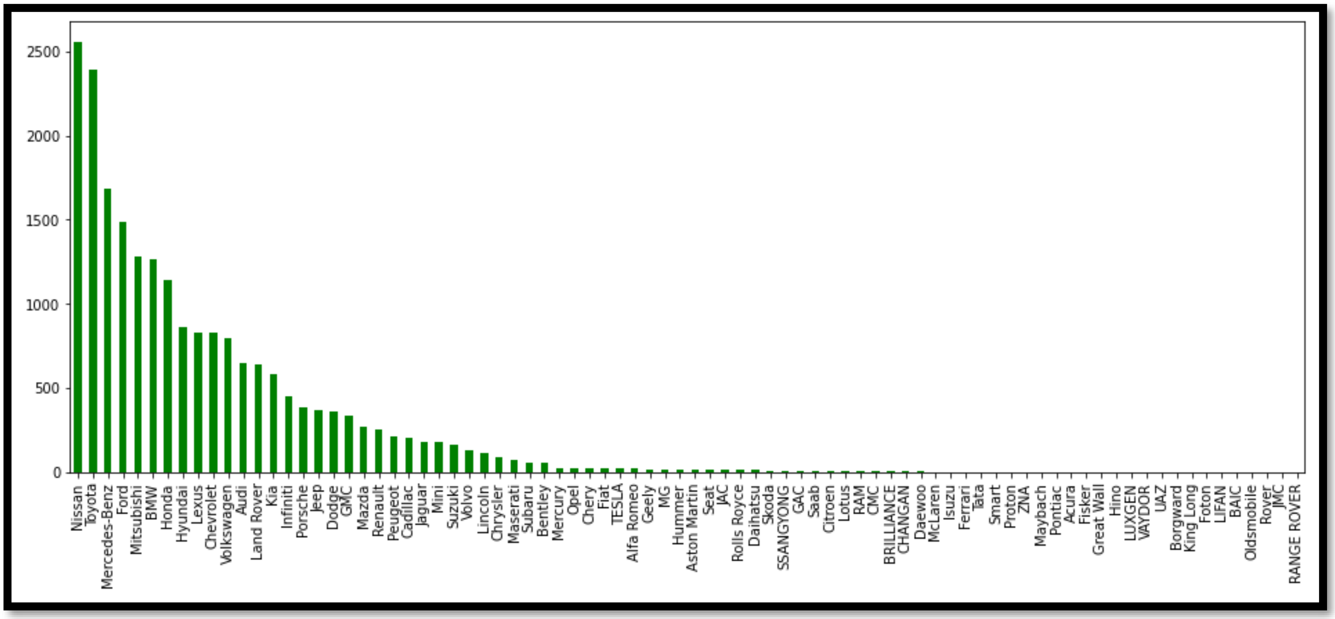


Figure 9 Attribute Brand - Bar plot

4. Fig-10 shows that the mode of attribute **Body Type** is sedan.

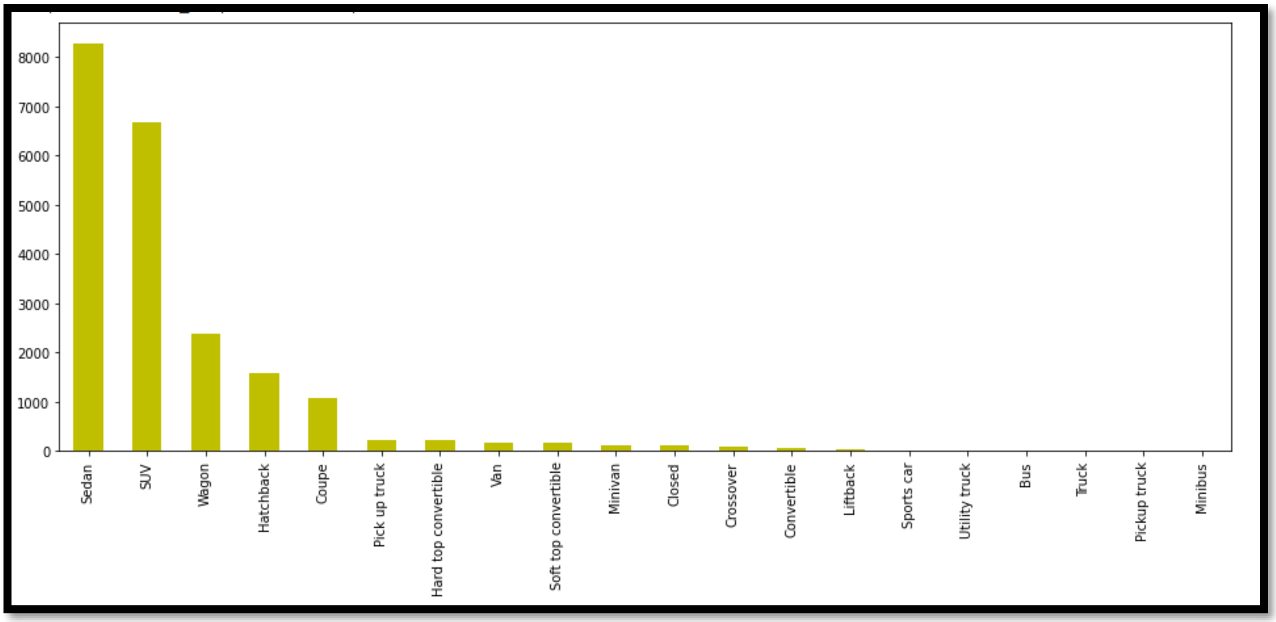


Figure 10 Attribute Body Type - Box Plot

5. Fig-11 shows that the most occurred car color is white.

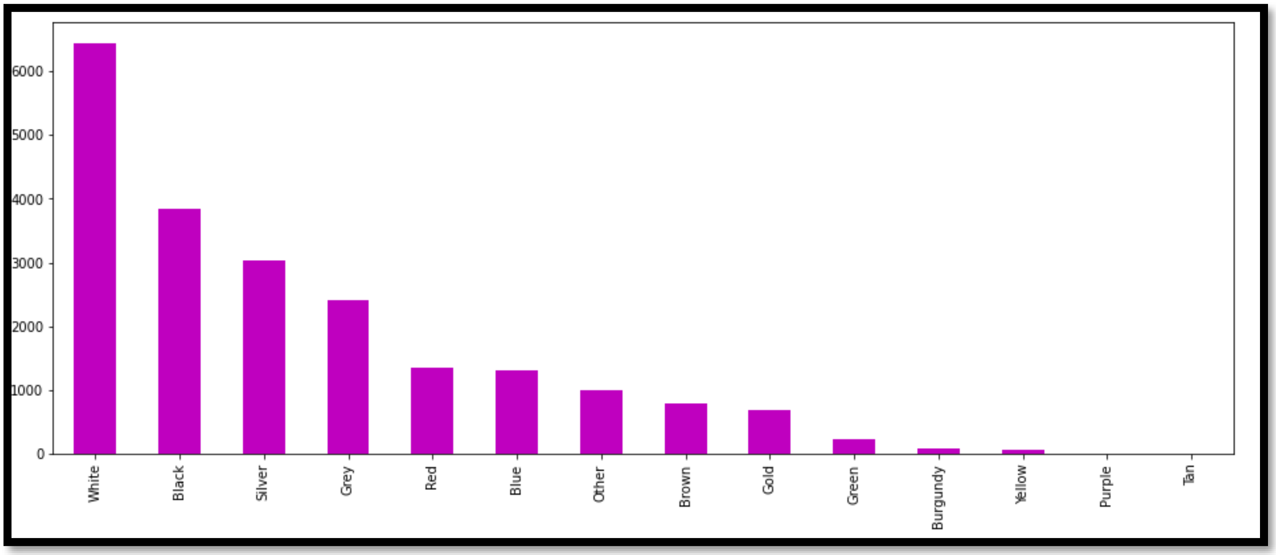


Figure 11 Attribute Color - Box plot

6. Gear type is the binary attribute, its visualization shows that it is highly imbalanced.

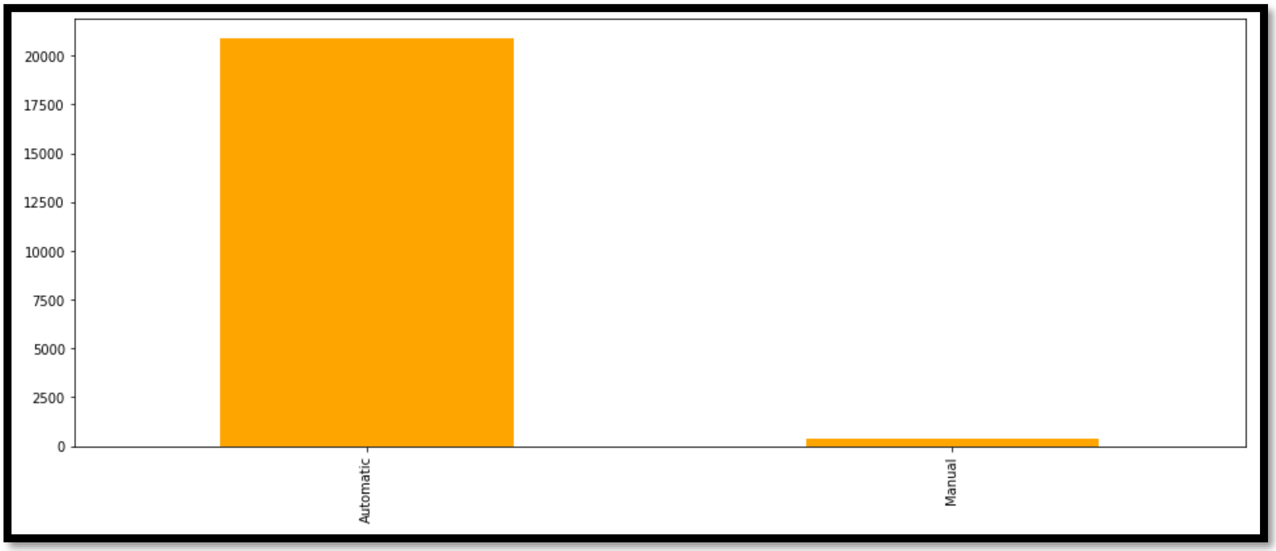


Figure 12 Attribute Gear Type - Box Plot

7. On the comparison of body type vs. price following results are achieved.

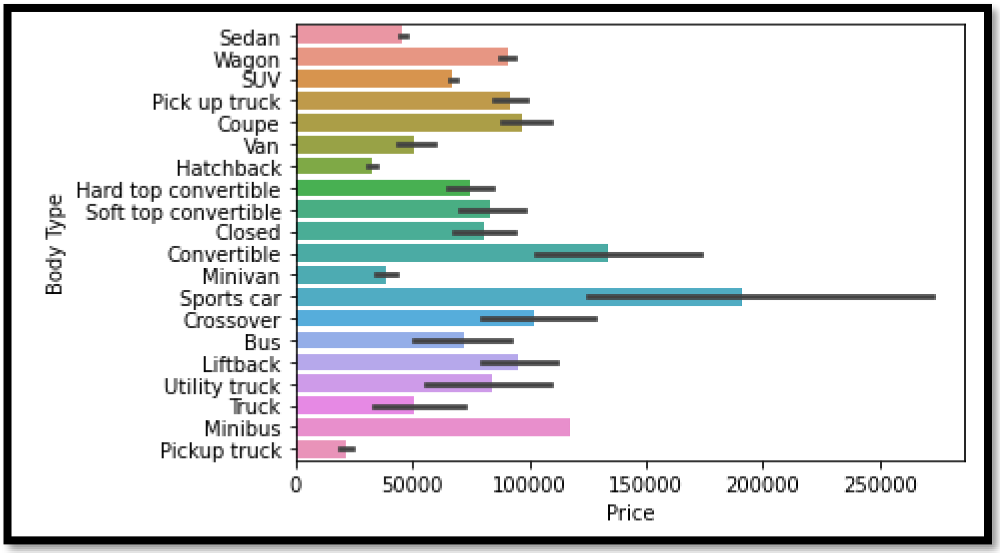


Figure 13 Body Type VS Price

From fig-13, it can be concluded that sports cars are the most expensive of all categories.

8. Based on a comparison of mileage and fuel type, the following results were obtained.

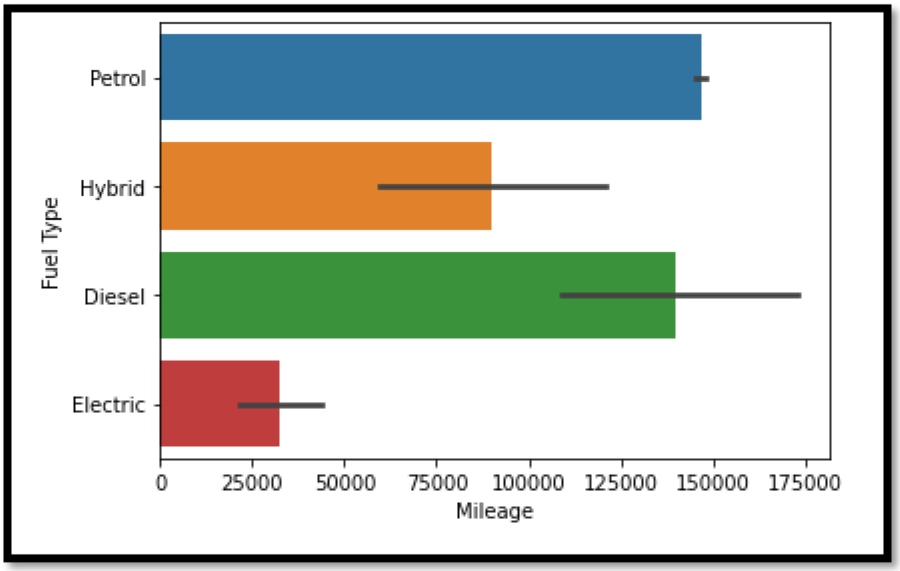


Figure 14 Fuel Type VS Mileage

Fig-14 reveals that petrol cars are the most fuel-efficient followed by diesel cars, while the least amount of mileage is given by electric cars.

9. The last step in visualization is to check the skewedness of all attributes that whether they normally distributed or not.

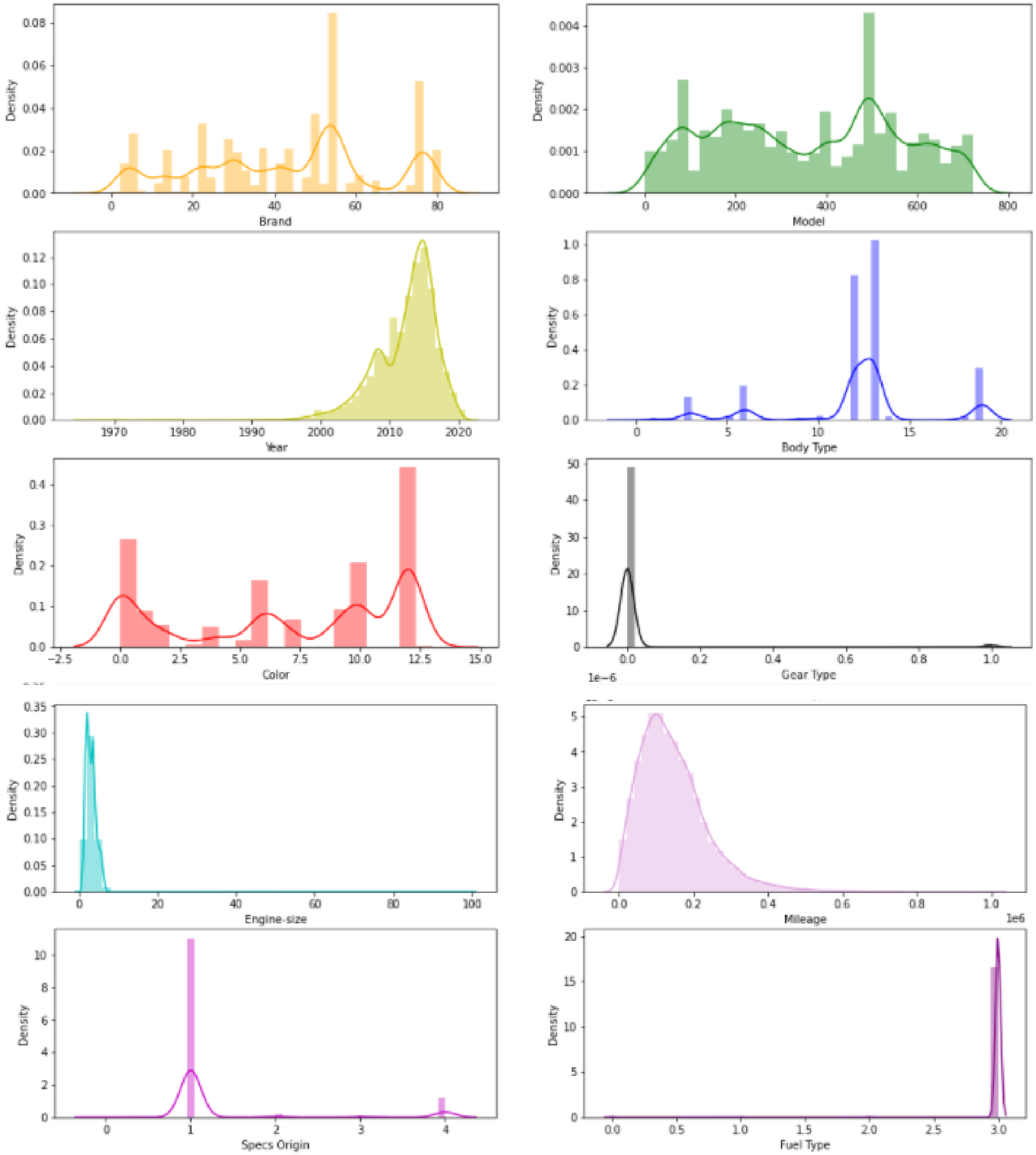


Figure 15 Attribute Distributions

Fig-15 explains the Skewness of all attributes of dataset. The model attribute is perfectly normally distributed followed by specs origin and color attributes. Two columns known as Year and Mileage are highly left and right skewed.

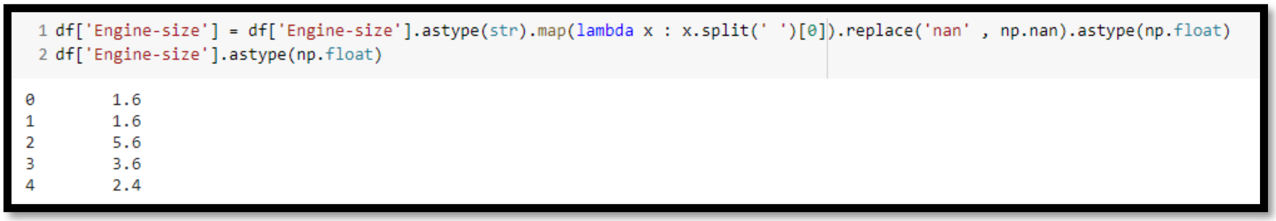
3.5 Dataset pre-processing

Pre-processing is a Data Mining technique that involves converting raw data into a comprehensible format. There is often a lack of specific activity or trend data, and many inaccurate facts are included in real-world data. Consequently, this may result in poor-quality data collection, and, in turn, poor-quality models constructed from the data. Such problems can be resolved by pre-processing the data.

Pre-processing in Machine Learning is the process of modifying, or encoding, data so that the machine can parse it more easily. Thus, the algorithm can now properly interpret the data.

In this project, following steps are preformed to pre-process the dataset.

1. Engine size attribute values are in the varchar format and need to be converted into int or float.



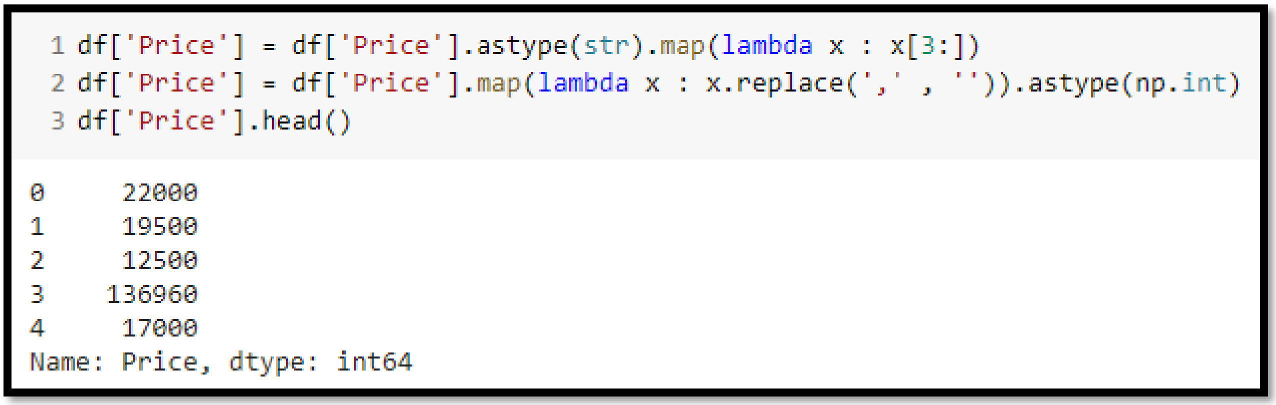
```
df['Engine-size'] = df['Engine-size'].astype(str).map(lambda x : x.split(' ')[0]).replace('nan' , np.nan).astype(np.float)
df['Engine-size'].astype(np.float)
```

```
0       1.6
1       1.6
2       5.6
3       3.6
4       2.4
```

Figure 16 Pre-processing of attribute Engine-Size

In the beginning, the values were like 1.8L, 3.2L, etc. It was decided to remove the convention L and change the data type to float.

2. The price was also in the format "AED167635". The price was converted into int from the string after the AED was removed.

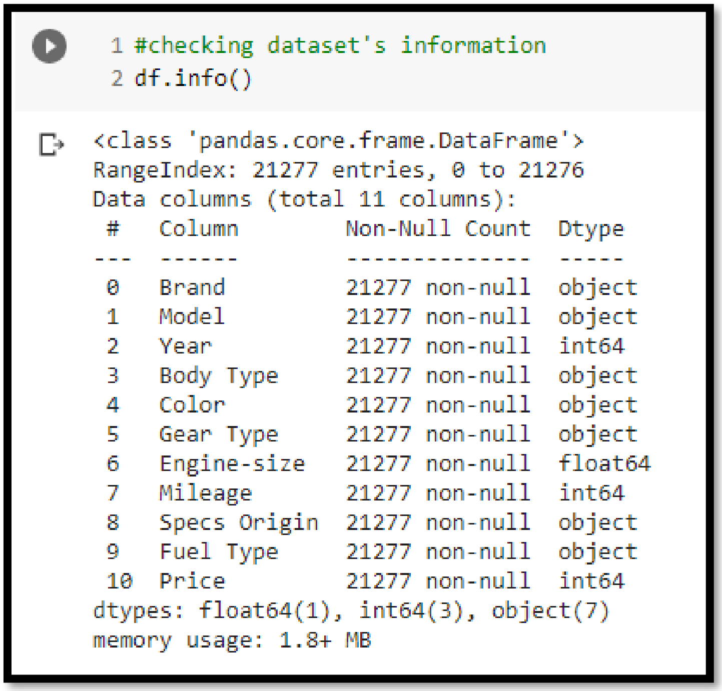


```
df['Price'] = df['Price'].astype(str).map(lambda x : x[3:])
df['Price'] = df['Price'].map(lambda x : x.replace(',' , '')).astype(np.int)
df['Price'].head()
```

```
0      22000
1      19500
2      12500
3     136960
4      17000
Name: Price, dtype: int64
```

Figure 17 Pre-processing of attribute Price

3. From fig-18, it can be concluded that now all the attributes are in appropriate format and data types.



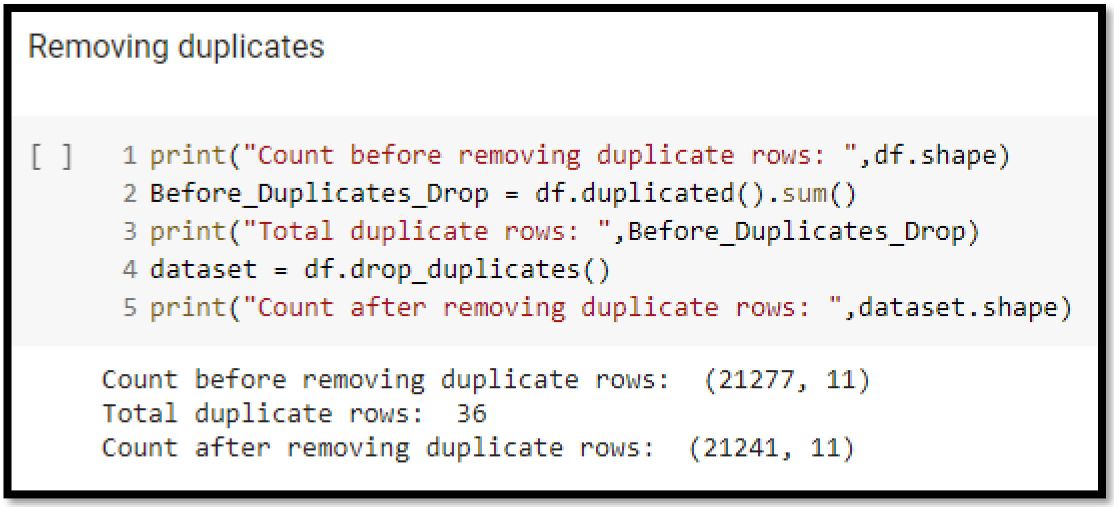
```
#checking dataset's information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21277 entries, 0 to 21276
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Brand         21277 non-null  object 
 1   Model         21277 non-null  object 
 2   Year          21277 non-null  int64  
 3   Body Type     21277 non-null  object 
 4   Color         21277 non-null  object 
 5   Gear Type     21277 non-null  object 
 6   Engine-size   21277 non-null  float64
 7   Mileage       21277 non-null  int64  
 8   Specs Origin  21277 non-null  object 
 9   Fuel Type     21277 non-null  object 
 10  Price         21277 non-null  int64  
dtypes: float64(1), int64(3), object(7)
memory usage: 1.8+ MB
```

Figure 18 Attribute data types

Now, all the attributes are ready for further processing.

4. The dataset contained no null values, but there were duplicate values as shown in fig-19.



Removing duplicates

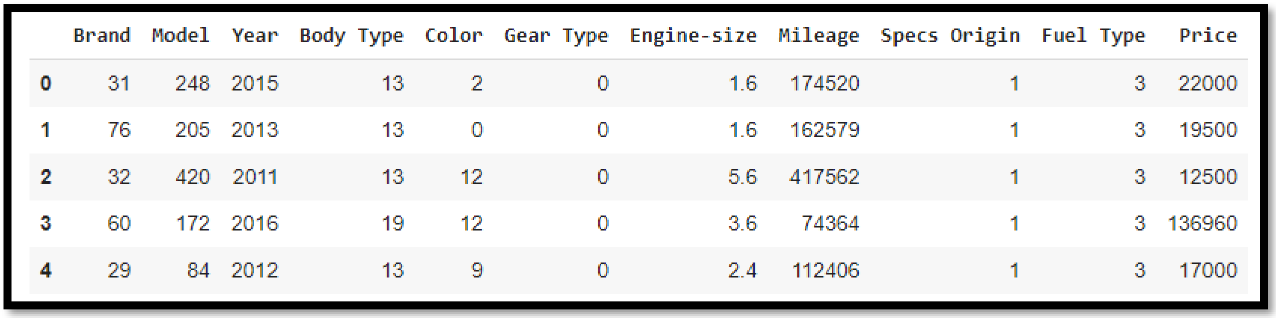
```
[ ]
print("Count before removing duplicate rows: ",df.shape)
Before_Duplicates_Drop = df.duplicated().sum()
print("Total duplicate rows: ",Before_Duplicates_Drop)
dataset = df.drop_duplicates()
print("Count after removing duplicate rows: ",dataset.shape)
```

```
Count before removing duplicate rows:  (21277, 11)
Total duplicate rows:  36
Count after removing duplicate rows:  (21241, 11)
```

Figure 19 Removal of Redundant Values

After dropping duplicate samples, the size of dataset reduced to 21241 from 21277.

5. The encoding Technique is used to turn categorical data into numerical data since machine learning algorithms cannot process categorical data. LabelEncoder is used in this project.



	Brand	Model	Year	Body Type	Color	Gear Type	Engine-size	Mileage	Specs Origin	Fuel Type	Price
0	31	248	2015	13	2	0	1.6	174520	1	3	22000
1	76	205	2013	13	0	0	1.6	162579	1	3	19500
2	32	420	2011	13	12	0	5.6	417562	1	3	12500
3	60	172	2016	19	12	0	3.6	74364	1	3	136960
4	29	84	2012	13	9	0	2.4	112406	1	3	17000

Figure 20 Label Encoding Results

Fig-20 validates that all categorical features have been converted into numerical features.

6. After data transformation, it is time to check for relationships among attributes. A correlation matrix was used to determine whether relationships exist.

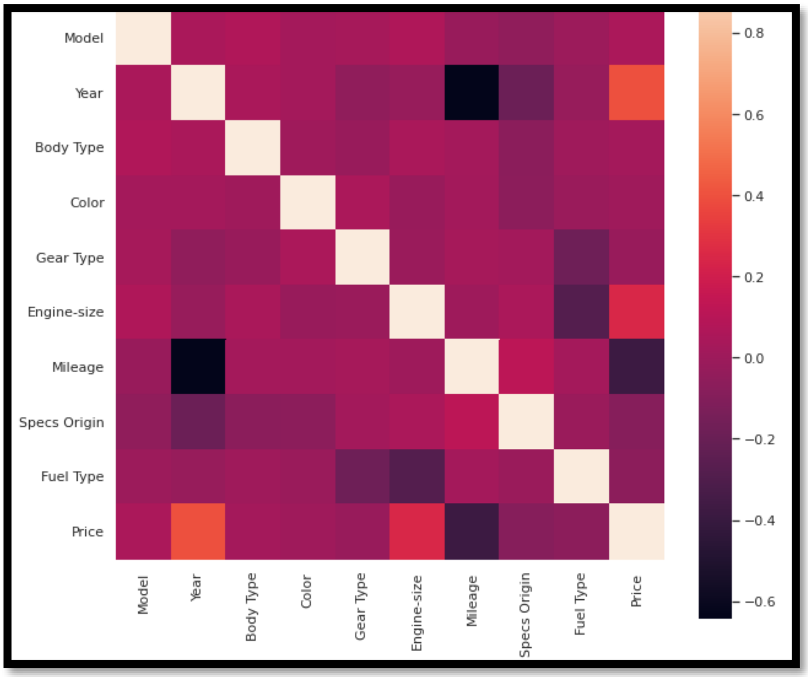


Figure 21 Correlation matrix

It is evident from this table that the attributes Gear type, fuel type, specs origin, and Milage are negatively correlated with the target class. Hence, negatively affecting the price.

In this step, the pre-processing is complete, and now it is ready for the prediction models.

3.6 Model Evaluation Parameters:

The regression model can be evaluated on following parameters:

1. **Mean Square Error (MSE):**

MSE is the single value that provides information about goodness of regression line. Smaller the MSE value, better the fit because smaller value implies smaller magnitude of errors.

$$\boldsymbol{MSE = \frac{1}{N}\sum_{i=1}^{N}|yi - y|^2}$$

Equation 3 MSE equation

2. **Root Mean Square Error (RMSE):**

RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.

3. **Mean Absolute Error (MAE):**

This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset.

$$\boldsymbol{MAE = \frac{1}{N}\sum_{i=1}^{N}|yi - y|}$$

Equation 4 MAE equation

Chapter 4 – Project Analysis

4.1 Experimental results & Analysis

Following machine learning classifiers are implemented.

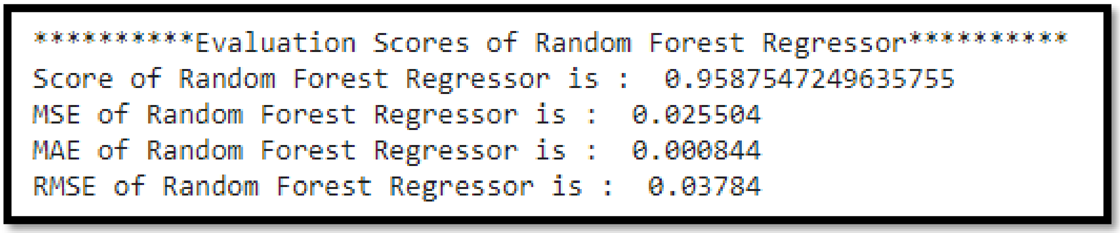
1. Random Forest Regressor
2. Linear Regression
3. Bagging Regressor

The implementations, selected parameters and accuracies of classifiers are:

4.1.1 Random Forest Regressor

Normally, random forests or random decision forests are used for classification, regression, and other tasks where they construct a multitude of decision trees at training time and output the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. A random decision forest corrects decision trees' habit of overfitting their training set. While random forests are generally better than decision trees, they are not as accurate as gradient boosted trees. They are, however, affected by data characteristics.

In this project, the random forest Regressor was trained with the intercept property. Score, MSE, RMSE and MAE errors are used to evaluate the model. Following results are achieved from there.



```
**********Evaluation Scores of Random Forest Regressor**********
Score of Random Forest Regressor is :  0.9587547249635755
MSE of Random Forest Regressor is :  0.025504
MAE of Random Forest Regressor is :  0.000844
RMSE of Random Forest Regressor is :  0.03784
```

Figure 22 Evaluation scores of Random Forest Regressor Model

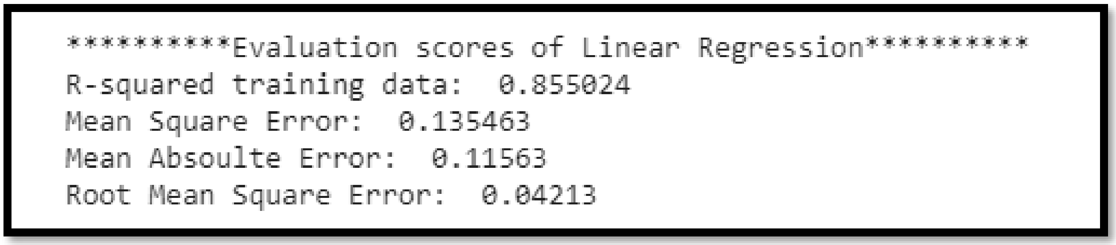
The R-square of training data is almost 95% while the RMSE is almost 0 which indicates that model is trained efficiently.

4.1.2 Linear Regression

By applying a linear equation to the observed data, linear regression attempts to illustrate the relationship between two variables. There should be one independent variable and one dependent variable. As an example, the weight and height of a person are linearly related. Therefore, the weight and height of the person have a linear relationship. As height increases, the weight also increases.

There is no requirement that one variable cause another, but there are some critical relationships between the two variables. When this is the case, we use a scatter plot to show the number of relationships between the variables. There is no pattern of increasing or decreasing on the scatter plot if there is no correlation or link between the variables. Hence, linear regression is not suitable for the given data in such cases.

In this project, the Linear Regression was trained with the intercept property. Score, MSE, RMSE and MAE errors are used to evaluate the model. Following results are achieved from there.



```
**********Evaluation scores of Linear Regression**********
R-squared training data:  0.855024
Mean Square Error:  0.135463
Mean Absoulte Error:  0.11563
Root Mean Square Error:  0.04213
```

Figure 23 Evaluation scores of Linear Regression model

The R-sqaure of training data is almost 85% while the RMSE is almost 0 which indicates that model is trained efficiently.

4.1.3 Bagging Regressor

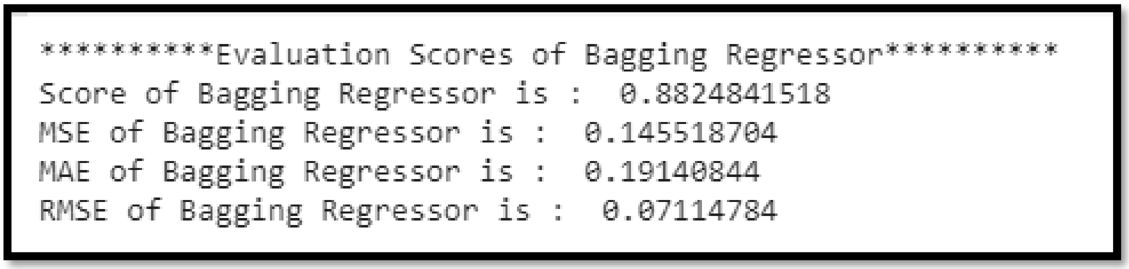
Bagging stands for Bootstrap Aggregating or simply Bootstrap + Aggregating.

- In Bagging, bootstrapping is the technique of generating multiple subsets from the whole (set) by using the replacement procedure.
- In Bagging, all possible outcomes of a prediction are gathered and randomized.

A better model is formed by combining many weak models.

Parallel ensemble method Bagging is a method of constructing models independently. When reducing variance is the goal, bagging is used.

In this project, the Bagging Regressor of decision trees was trained with the intercept property. Score, MSE, RMSE and MAE errors are used to evaluate the model. Following results are achieved from there.



```
**********Evaluation Scores of Bagging Regressor**********
Score of Bagging Regressor is :  0.8824841518
MSE of Bagging Regressor is :  0.145518704
MAE of Bagging Regressor is :  0.19140844
RMSE of Bagging Regressor is :  0.07114784
```

Figure 24 Evaluation scores of Bagging Regressor model

The R-sqaure of training data is almost 88% while the RMSE is almost 0 which indicates that model is trained efficiently.

4.2 Results comparison

The comparison of all the experiments shown in the table-2.

SR#	Algorithm	Accuracy	MSE	MAE	RMSE
1	Random Forest Regressor	0.95	0.025	0.0008	0.03
2	Linear Regression	0.85	0.13	0.11	0.04
3	Bagging Regressor (Decision tree)	0.88	0.14	0.19	0.07

Table 2 Results Comparison

Given the evaluation parameters the Random Forest Regressor outperformed as it has the highest accuracy of the three different algorithms, as well as the lower error in all three-evaluation parameter. Second in accuracy in the Bagging regressor with 88% accuracy, even through it has a higher error parameter than linear regression. Least accurate was the Linear regression with 85% accuracy thought it had a lower error value than Bagging Regressor.

Graphical illustration of accuracies all errors are:

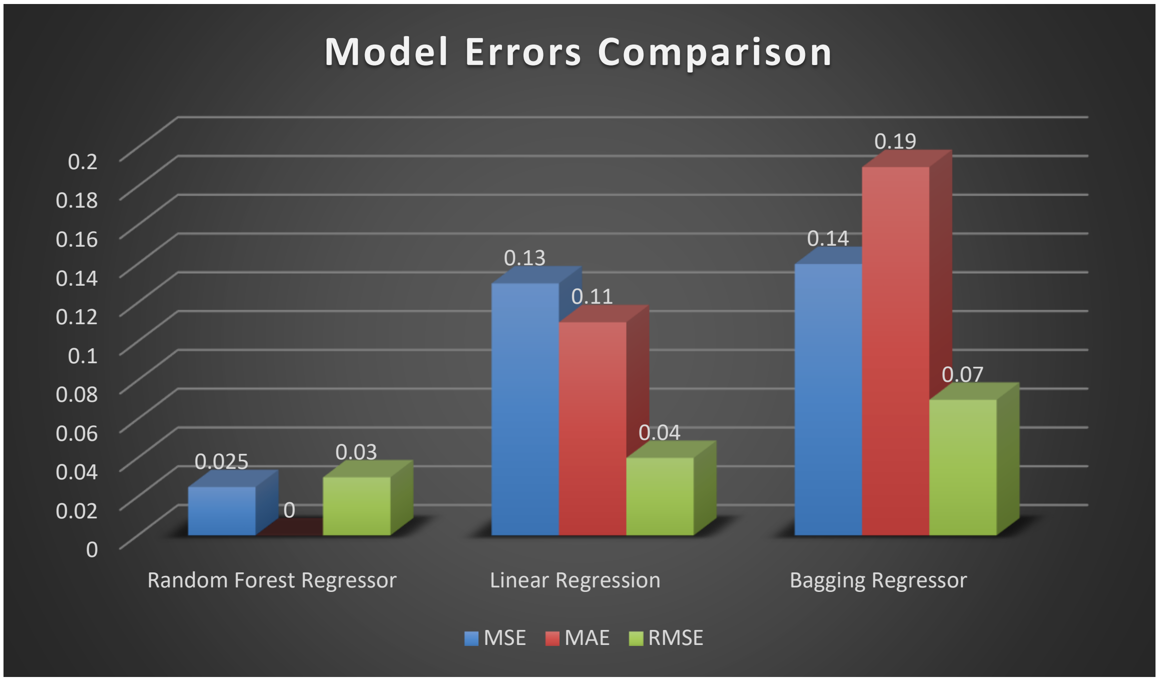


Figure 25 Model Errors Comparison

From here this can be concluded that the errors of Random Forest Regressor are lower in comparison with Linear and Bagging regressor. Hence, Random Forest outperformed.

Chapter 5 – Conclusion

5.1 Conclusion

Using data mining and machine learning approaches, this project proposed a scalable framework for Dubai based used cars price prediction. Buyanycar.com website was scraped using the Parse Hub scraping tool to collect the benchmark data. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 95% accuracy followed by Bagging Regressor with 88%. Each experiment was performed in real-time within the Google Colab environment. In comparison to the system's integrated Jupyter notebook and Anaconda's platform, algorithms took less training time in Google Colab.

5.2 Recommendations and Future Work

In the future, more data will be collected using different web-scraping techniques, and deep learning classifiers will be tested. Algorithms like Quantile Regression, ANN and SVM will be tested.

Afterwards, the intelligent model will be integrated with web and mobile-based applications for public use. Moreover, after the data collection phase Semiconductor shortages have incurred after the pandemic which led to an increase in car prices, and greatly affected the secondhand market. Hence having a regular Data collection and analysis is required periodically, ideally, we would be having a real time processing program.

.

Bibliography

BIELSKI, V., & RAMARATHNAM, S. (2020, July 16). *UAE's used car sales set to surge past 1 million mark by 2025.* Retrieved from gulfbusiness: https://gulfbusiness.com/uaes-used-car-sales-set-surge-past-1-million-mark-2025/#:~:text=For%20every%20new%20car%20sold,crossing%20the%201%20million%20mark.

Bridge, S. (2020, January 10). *Why the value of used cars is rising for the first time in the UAE.* Retrieved from arabianbusiness: https://www.arabianbusiness.com/retail/435520-why-the-value-of-used-cars-is-rising-for-the-first-time-in-the-uae

Ceriottia, M. (2019). Unsupervised machine learning in atomistic simulations, between predictions and understanding. 150-155.

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019, February). Car Price Prediction using Machine. *TEM Journal, 8*(1), 113-118. doi:10.18421/TEM81-16

Gongqi, S., Yansong, W., & Qiang, Z. (2011). A New Model for Residual Value Prediction of the Used Car Based on BP Neural. *Third International Conference on Measuring Technology and Mechatronics Automation* (pp. 682-685). Shanghai: IEEE. doi:10.1109/ICMTMA.2011.455

Great Learning Team. (2020, August 17). *Introduction to Multivariate Regression Analysis*. Retrieved from mygreatlearning: https://www.mygreatlearning.com/blog/introduction-to-multivariate-regression/#:~:text=Multivariate%20Regression%20is%20a%20supervised,try%20to%20predict%20the%20output.

Jian Da Wu, C.-c. H.-C. (2017). "An expert system of price forecasting for used cars using adaptive. *ELSEVEIR, 16*, 417-957.

K.Samruddhi, & Kumar, D. R. (2020, September). Used Car Price Prediction using K-Nearest Neighbor Based Model. *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), 4*(3), 686-689.

Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth? *Journal of Statistics Education*. doi:10.1080/10691898.2008.11889579

Listiani, M. (2009). Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. *Master Thesis*. Hamburg: Hamburg Univesity of Technology .

Matthew Botvinick, S. R.-N. (May 2019). Reinforcement Learning, Fast and Slow. *Trends in cognitive sciences, 23*(5), 408-422.

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of Prices for Used Car by Using Regression Models. *5th International Conference on Business and Industrial Research (ICBIR)*, (pp. 115-119). Bangkok.

Nabarun Pal, P. A. (2018). How much is my car worth? A methodology for predicting used cars prices using Random Forest. *Future of Information and Communications Conference (FICC) 2018* , 1-6.

Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. *International Journal of Computer Applications*, 27-31.

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning. *International Journal of Information & Computation Technology*, 754-764.

Research, F. -M. (2020, Feburary 25). *Automotive Industry in Dubai*. Retrieved 10 24, 2021, from https://www.feedbackme.com/automotive-industry-in-uae/

Rizvi, R. (2019, April). *Car Production is on the Rise in Dubai*. Retrieved September 10, 2019, from https://propakistani.pk/2019/04/08/car-production-is-on-the-rise-in-dubai/

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*.

Swaminathan, S. (2018, March 15). *Logistic regression - detailed overview*. (towards Data science) Retrieved October 27, 2020, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

uae.buyanycar. (n.d.). Retrieved 4 20, 20202, from buyanycar: https://uae.buyanycar.com/en/home

Used Vehicle Value Index. (2021, April). Retrieved from manheim: https://publish.manheim.com/en/services/consulting/used-vehicle-value-index.html