**CHAPTER Ⅰ**

**I****ntroduction**

The automobile industry is one of the prominent industries for the national economy. Day by day the car is getting popular for the private transport system. The increase in new cars and customers' economic inability, global sales of old cars are expanding. 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. 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. Accurately predicting used car prices requires expert knowledge due to the nature of their dependence on a variety of factors and features. 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 Indian market, the benchmark dataset containing all key features is scraped.

To solve this problem, different Machine learning algorithms are used to provide continuous values as output rather than classified values. This allows for the prediction of the car's real price rather than its price range. This prediction problem can be considered a regression problem since it belongs to the supervised learning domain. Six Regressor known as linear regression, lasso regression, ridge regression, random forest regression, XG boost, and support vector regression were trained and compared.

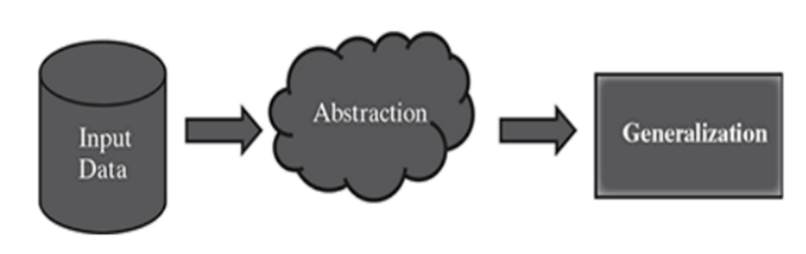
**1.1 Machine Learning**

Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed. The basic machine learning process can be divided into three parts.

**1.** **Data Input:** Past data or information is utilized as a basis for future decision-making

**2.** **Abstraction:** The input data is represented in a broader way through the underlying algorithm

**3.** **Generalization:** The abstracted representation is generalized to form a framework for making decisions



**Fig. 1.1 Schematic Representation of the Machine Learning Process**

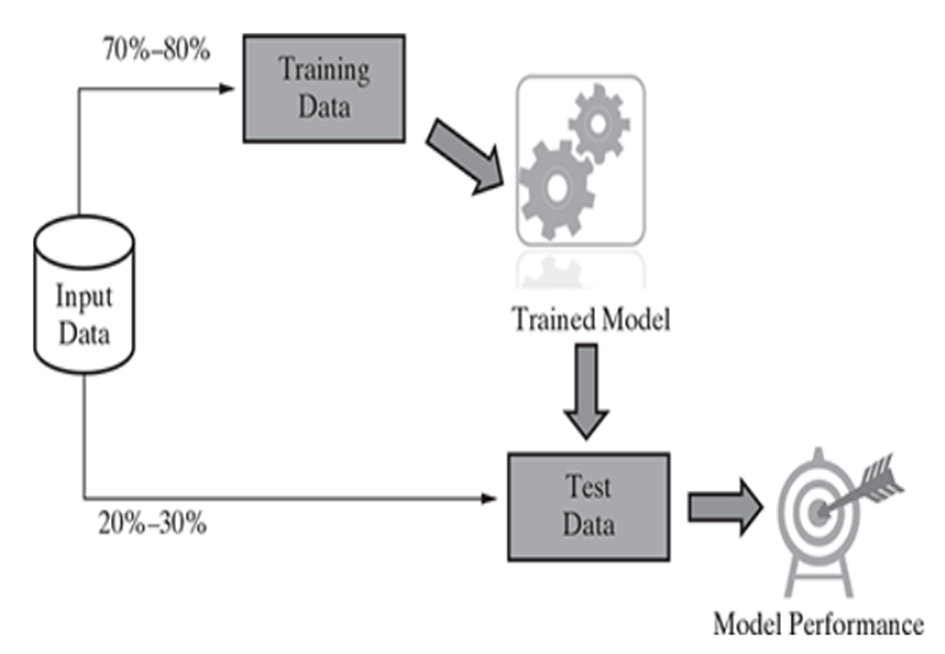
A predictive model is used for tasks that involve the prediction of a given output (or target) using other variables (or features) in the data set. Predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction. The learning algorithm in a predictive model attempt to discover and model the relationships among the target variable (the variable being predicted) and the other features (variables). Machine learning can be classified as supervised learners which construct predictive models, and unsupervised learners which build descriptive models.

In supervised learning, the training data you feed the algorithm includes the target values. Consequently, the solutions can be used to help supervise the training process to find the optimal algorithm parameters. Most supervised learning problems can be bucketed into one of two categories, regression or classification. When the objective of our supervised learning is to predict a numeric outcome, refer to this as a regression problem (not to be confused with linear regression modeling). Regression problems revolve around predicting output that falls on a continuum.

**1.2 Holdout Method:**

The hold-out method for training the machine learning models is a technique that involves splitting the data into different sets: one set for training, and other sets for validation and testing. The hold-out method is used to check how well a machine learning model will perform on the new data. In general, 70%–80% of the input data is used for model training. The remaining 20%-30% is used as test data for validation of the performance of the model. However, a different proportion of dividing the input data into training and test data is also acceptable. To make sure that the data in both the buckets are similar in nature, the division is done randomly. Random numbers are used to assign data items to the partitions. This method of partitioning the input data into two parts – training and test data, which is by holding back a part of the input data for validating the trained model is known as the holdout method.

Once the model is trained using the training data, the labels of the test data are predicted using the model’s target function. Then the predicted value is compared with the actual value of the label. This is possible because the test data is a part of the input data with known labels. The performance of the model is in general measured by the accuracy of prediction of the label value.



**Fig. 1.2 Holdout Method**

**1.3 Underfitting:**

If the target function is kept too simple, it may not be able to capture the essential nuances and represent the underlying data well. Underfitting results in both poor performance with training data as well as poor generalization of test data. Underfitting can be avoided by

1. Using more training data.
2. Reducing features by effective feature selection.

**1.4 Overfitting:**

Overfitting refers to a situation where the model has been designed in such a way that it emulates the training data too closely. In such a case, any specific deviation in the training data, like noise or outliers, gets embedded in the model. It adversely impacts the performance of the model on the test data. Overfitting results in good performance with training data set, but poor generalization and hence poor performance with test data set. Overfitting can be avoided by

1. Using re-sampling techniques like cross validation.
2. Hold back a validation data set.
3. Remove the nodes which have little or no predictive power for the given machine learning problem.

**Bias:**

Bias is the simplifying assumption made by the Machine Learning Model to make the learning process easier. It makes the model faster in terms of learning from the training data but at the same time the flexibility of the model decreases if the Bias is significantly high. A Model with high bias tends to ignore the complexity between the input and the output data and the model becomes too simple and ultimately results in the Underfitting of the data.

**Variance:**

Variance occurs when the model performs too well on the training data, but its efficiency decreases significantly when it works on the testing data. High variance means that the model learns from training data to that extent that it negatively impacts the performance of the model and results in the Overfitting of the data.

**1.5 Language used in Machine Learning:**

Python is majorly used for implementing machine learning concepts during this project as there are a number of inbuilt methods in the form of packaged libraries and modules present in python. The libraries used during the project implementation are the following: advanced mathematical functionalities (NumPy), algorithms and mathematical tools (SciPy), read data (Pandas) and numerical plotting (matplotlib). Built on these libraries, there is a machine learning library named scikit-learn, which has various classification, regression, and clustering algorithms embedded in it.

**1.6 Objective of the study:**

The main objective of this work is to analyze the information about the world’s largest collection of used cars sales data, to use different prediction models to predict the retail price of a used car sales and compare their levels of accuracy using appropriate statistical methods. This objective is studied using the data on the collection of used vehicles for resale data downloaded from Kaggle website and the results are presented in this report.

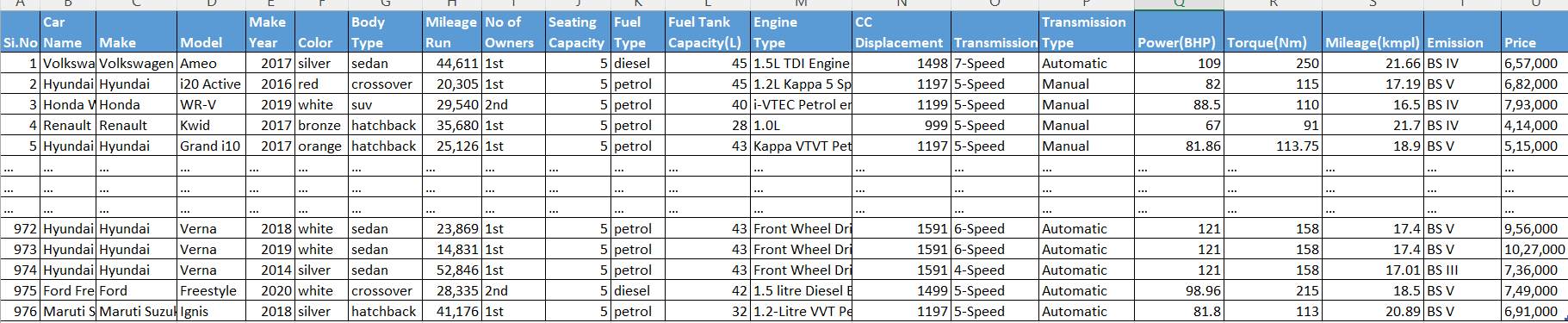
This report consists of four chapters. Chapter II presentation of the data. Using supervised learning method, to predict the prices of used cars. The model has been chosen after careful exploratory data analysis to determine the impact of each feature on price. To find the important features that are influencing each model and the results to the data on the collection of used cars for sale data and its determination is given in Chapter III. The results observed from the statistical analysis and the comparison are summarized in Chapter IV. A comprehensive reference list has also been given at the end, which was referred to during our work.

# 

**CHAPTER Ⅱ**

# **Presentation of Dataset**

The dataset used in this project was downloaded from Kaggle. This is a web scrapped data from a car resale website, which includes more details of the cars along with their respective prices. It was uploaded on Kaggle by Rakkesh Arv who Kaggle.com user. Rakkesh Arv scraped this data from craigslist with non-profit purpose. The database consists of most all relevant information that Craigslist provides on car sales including columns like price, Mileage Run, model, Mileage (kmpl), and 20 other features and 976 rows. The sample data is presented in Fig. 2.1.



**Fig 2.1. The Sample Data**

The features are listed below

**Car Name** : The full name of the car which is displayed in the ad

**Price** : Selling price

**Make Year** : Year of manufacturing

**Make :** Make of the car

**Model** : The model of the car.

**Color :** The color of the car

**Body Type** : Body type of the car

**Fuel :** Fuel type used by the car.

**Odometer :** Distance that the car has traveled after it being bought.

**No of Owners :** No of previous owners.

**Seating Capacity :** Total seating capacity.

**Fuel Tank Capacity (L):** Total Fuel capacity of the car.

**Engine Type :** Engine name, Model and Type

**CC Displacement :** Total Cubic Displacement

**Transmission :** Kind of Transmission

**Transmission Type :** The type of transmission

**Power (BHP) :** Total Max Power

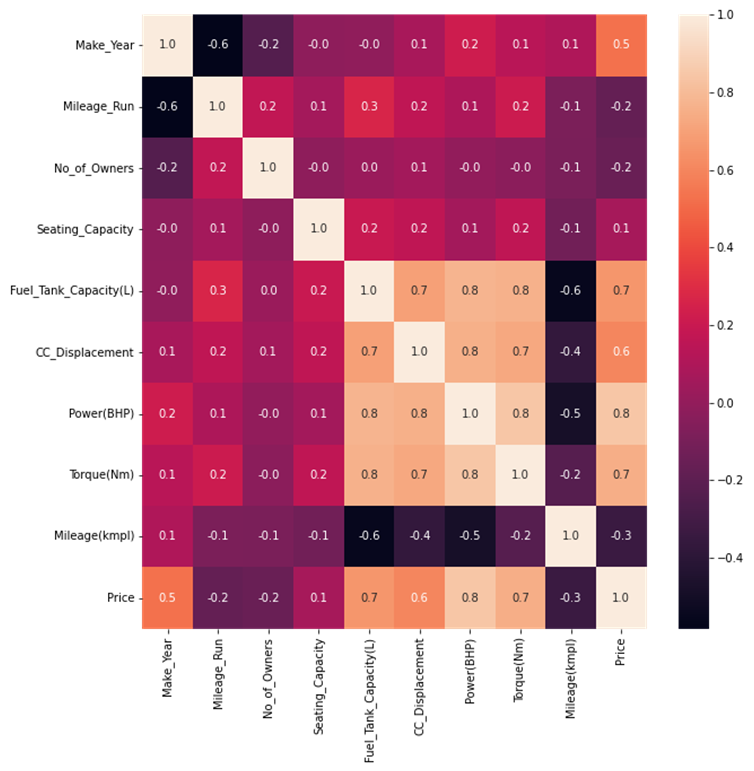
**Torque (Nm) :** Total Max Torque.

**Mileage (kmpl) :** Average mileage of the car

**Emission :** Emission norms of the car.

# **2.1 Exploratory Data Analysis (EDA)**

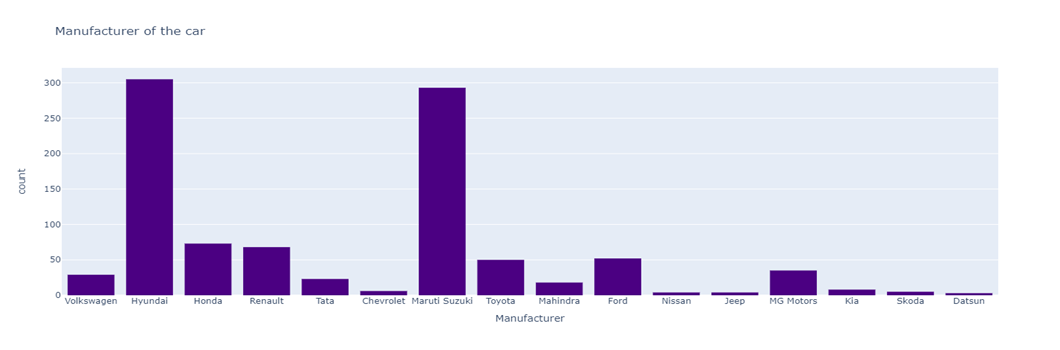
EDA is primarily used to provide a better understanding of dataset's variables and their relationships. Due to its combination of car name and engine type was a redundant attribute, and it was removed. While exploring the data, it will look at the different combinations of features with the help of visuals. This will help us to understand our data better and give us some clue about pattern in data.



**Fig. 2.2 Correlation Heatmap**

The correlation Heatmap is shown in Fig. 2.2. From this figure, it is observed that price is highly influenced by power (BHP) with highest correlation value of 0.8.

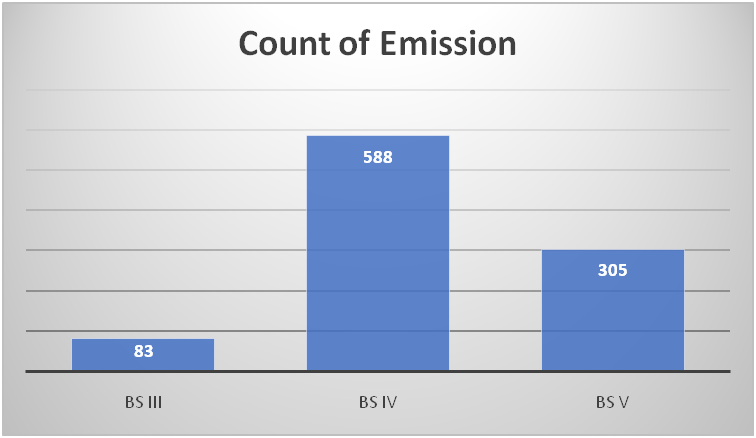
**Manufacturer:**



**Fig. 2.3 Bar Graph for Manufacturer**

The bar graph for manufacturer of car resale data is shown in Fig. 2.3. From this figure it is observed that Hyundai model leads the chart.

**Emission:**

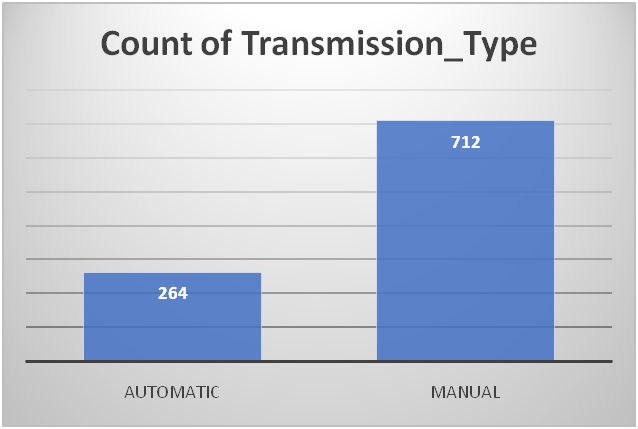


**Fig. 2.4 Bar Graph for Emission**

The bar graph for emission of car resale data is shown in Fig. 2.4. From this figure it is observed that BSIV vehicles are most likely resale car type.

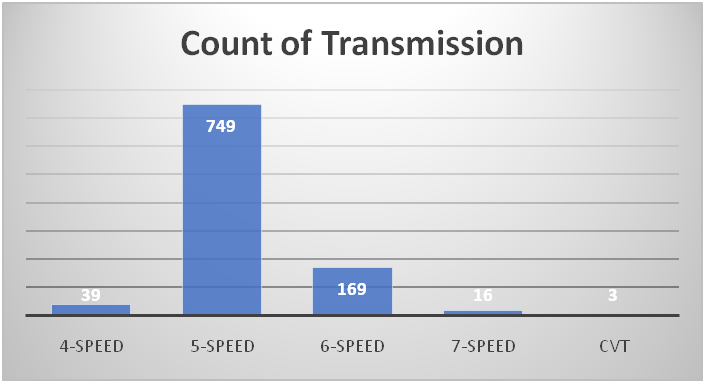
**Transmission Type:**

The bar graph for transmission type is shown in Fig. 2.5. From this figure it is observed that manual transmissiontakes the lead in the resale data.



**Fig. 2.5 Bar Graph for Transmission Type**

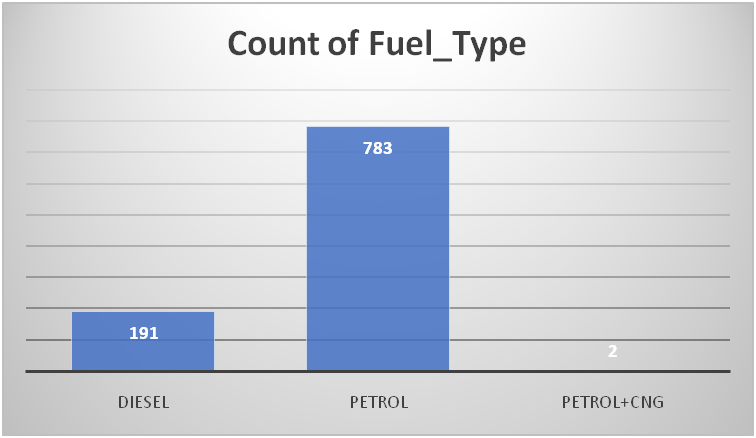
**Transmission:**



**Fig. 2.6 Bar Graph for Transmission**

The bar graph for transmission is shown in Fig. 2.6. From this, it is observed that, 5-speed transmission is in lead with count 749.

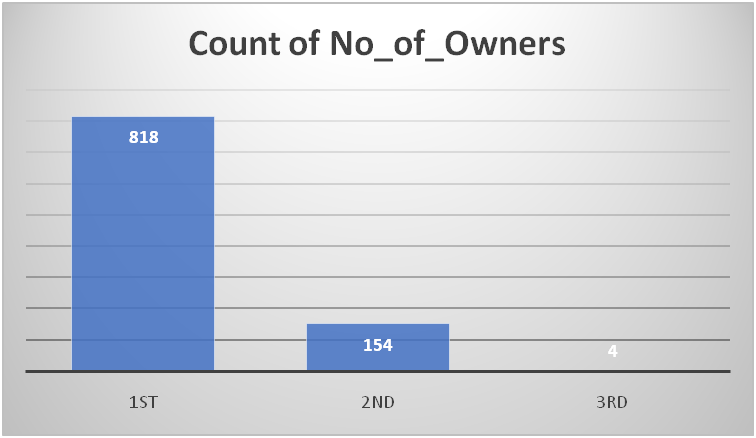
**Fuel-Type:**



**Fig. 2.7 Bar Graph for Fuel Type**

The bar graph for fuel-type is shown in Fig. 2.7. From the bar graph, it is observed that most likely fuel type is petrol**.**

**No. of Owners:**

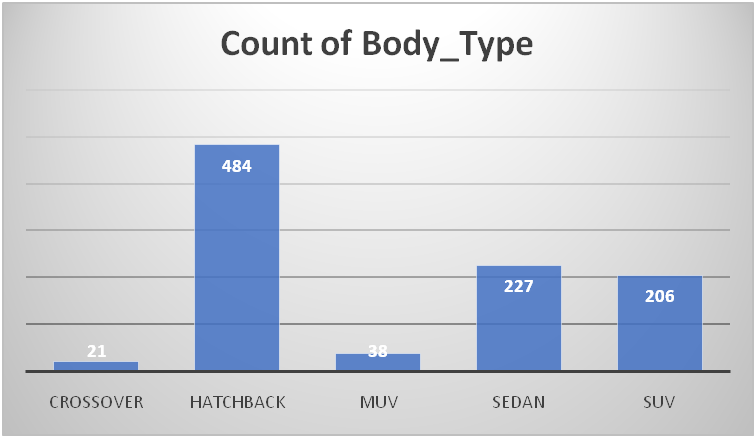


**Fig. 2.8 Bar Graph for No. of Owners**

The No. of owners of the car has a very prominent effect on its resale value. The bar graph for no. of owners is shown in Fig. 2.8. From this, it is clear that most of the car resales have single owner.

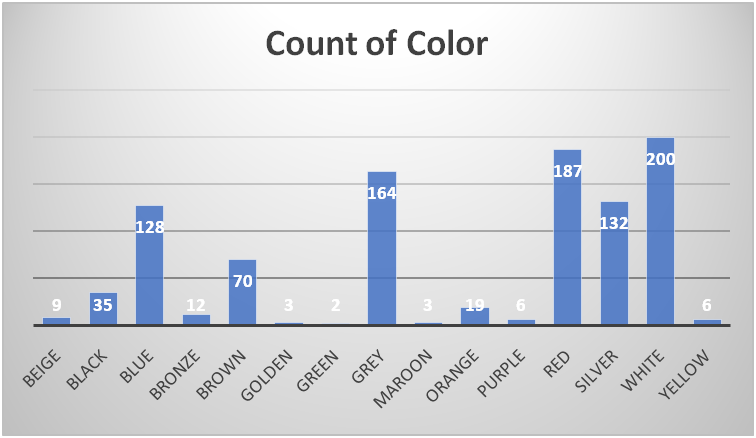
**Body Type:**

The bar graph of car body type is shown in Fig. 2.9. From this, it is observed that, the most likely body type is Hatchbacks.



**Fig. 2.9 Bar Graph for Body Type**

**Color of the Cars:**



**Fig. 2.10 Bar Graph for Color**

From the above Fig. 2.10, it is observed that white color leads the chart.

# **2.2 MODEL EVALUATION PARAMETERS:**

In this section, performance of the models will be tested by using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and

R-squared method (R2).

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

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

**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 the 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.

**4. R-squared method:**

* R-squared is a statistical method that determines the goodness of fit.
* It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
* The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
* It is also called a coefficient of determination, or coefficient of multiple determination for multiple regression.
* It can be calculated from the below formula:

**CHAPTER Ⅲ**

**Machine Learning Regression Models and**

**Its Statistical Analysis**

Due to the increased price of new cars and the inability of customers to buy new cars due to the lack of funds, used cars resale 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. Even though there are websites that offer this service, their prediction method may not be the best. Besides, different models and systems may contribute to predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling. To study price prediction via machine learning methods, it is observed that, when buying a used car, people pay serious attention to the manufacturer of the car. It is shown in Fig. 2.2. Hyundai and Maruti Suzuki are one dominant manufacturers in India. Transmission is another feature that has a dominant subcategory in the used car market. According to Fig. 2.5, manual transmission has a strong effect on people’s preference for cars.

In this chapter, we will apply machine learning models as a framework for the data analysis. In relation to the data set, literature suggests below five methods that can be appropriate. Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regression and XGBoost.

**Pre-processing the data**

**Label Encoding:** In the dataset, there are 20 predictors. 9 of them are numerical variables while the rest of them are categorical. To apply machine learning models, we need numeric representation of the features. Exploratory Data Analysis is done by using statistical graphics and other visualization methods to summarize the main characteristics of data. Due to its combination of car name and engine type was a redundant attribute, and it was removed. Then the features allocation of data is done where the dependent feature (price) and independent features (Make Year, Body type, No of Owners, Fuel Type, Transmissionetc.) are allocated for further procedure. **Train-Test Split:** After the allocation of dependent and independent features is completed, proceed further with the splitting of dataset into training and testing data. Using 70% of data for training our model and 30% data for testing purposes.

**Scaling the Data:** Without scaling, the machine learning models will try to disregard coefficients of features that have low values because their impact will be so small compared to the big value features.

# **3.1 Linear Regression**

Linear regression, a staple of classical statistical modeling, is one of the simplest algorithms for doing supervised learning. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

The linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (*x*) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, represent a linear regression as:

Here,

Dependent variable (Target Variable)

Independent variable (Predictor Variable)

Intercept of the line (Gives an additional degree of freedom)

Linear regression coefficient (scale factor to each input value)

Random error

The values for *x* and *y* variables are training datasets for Linear Regression model representation.

## **Assumptions of Linear Regression:**

* Linear relationship between the features and target
* Small or no multicollinearity between the features
* Homoscedasticity Assumption
* Normal distribution of error terms
* No autocorrelations

## **Results of Linear regression are:**

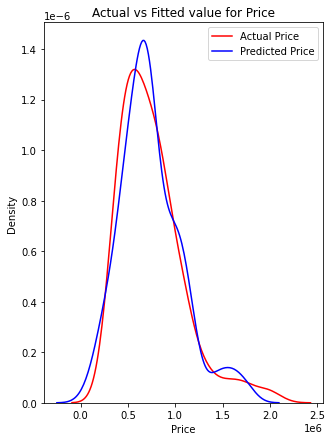
RMSE : 119223.04089032744

R2 : 0.8779081743258501

MAE : 95325.73 degrees.

Accuracy : 85.8 %.

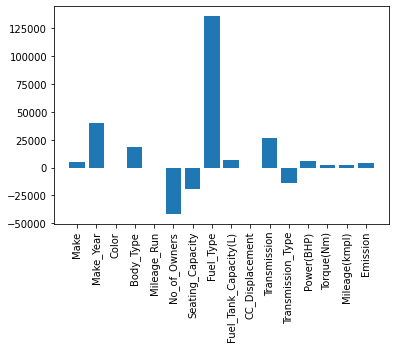
The below Fig. 3.1 shows the difference between actual and predicted values. From that there is more difference between predicted and actual values.



**Fig. 3.1 Linear Model- Actual vs Predicted**

## **Feature importance:**

Fig. 3.2 is a simple bar plot of the feature importance to illustrate the disparities in the relative significance of the variables. From the above figure, it is clearly shown linear regression suggests that Make Year, Body type, No of Owners, Fuel Type and Transmissionare the five important features.



**Fig. 3.2 Linear Model-Feature Importance**

**3.2 Ridge Regression**

To study about Ridge regression, we have to know about the technique known as Regularization.

**Regularization:**

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.Sometimes the [machine learning](https://www.javatpoint.com/machine-learning) model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when dealing with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be dealt with the help of a regularization technique.

  It mainly regularizes or reduces the coefficient of features to zero. In simple words, “In regularization technique, it reduces the magnitude of the features by keeping the same number of features.”

* Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
* Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called L2 regularization.
* In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
* The equation for the cost function in Ridge regression will be:

* In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.
* As we can see from the above equation, if the values of λ tend to zero, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ, the model will resemble the linear regression model.

**Assumptions of Ridge Regression:**

The assumptions of ridge regression are the same as that of linear regression: Linearity, Constant variance and Independence. However, as ridge regression does not provide confidence limits, the distribution of errors to be normal need not be assumed.

## 

## **Results of Ridge Regression**:

RMSE : 119299.9995

R2 : 0.878

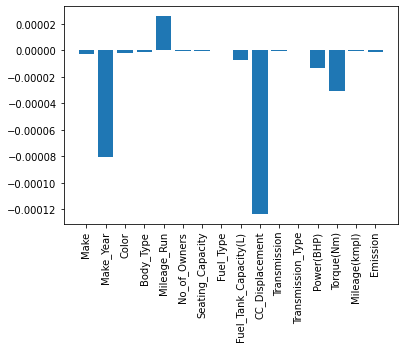
MAE : 95381.13 degrees.

Accuracy : 85.8 %.

## 

## **Feature importance:**

From the below Fig. 3.3, Ridge regression suggests that Make Year, Mileage Run, CC Displacement, Power (BHP), and Torque (Nm) these five variables are the most important.



**Fig. 3.3 Ridge Regression Model- Feature Importance**

‍

# **3.3 Lasso Regression:**

* Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator.
* It is like the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.
* Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
* It is also called L1 regularization. The equation for the cost function of Lasso regression will be:
* Some of the features in this technique are completely neglected for model evaluation.
* Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

## **Results for Lasso Regression:**

RMSE value : 119223.30089013484

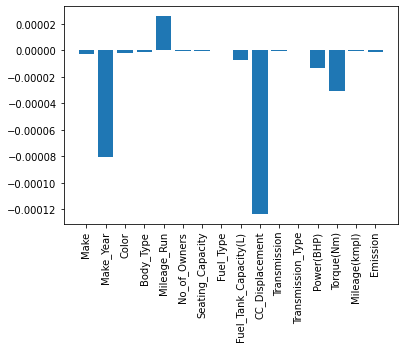
R2 : 0.8779076418132495

MAE : 95326.28 degrees.

Accuracy : 85.8 %.

**Feature importance:**

From the below Fig. 3.4, Lasso regression suggests that Make Year, Mileage Run, CC Displacement, Power (BHP), and Torque (Nm)these five variables are the most important.



**Fig. 3.4 Lasso Model- Feature Importance**

‍

# **3.4 Random Forest Regression:**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. Random forest is a set of multiple decision trees. Deep decision trees may suffer from overfitting, but random forest prevent overfitting by creating trees on random subsets. That’s why it’s a good model for analysis. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Preparing the Data as with the classification task, in this section, it will divide our data into attributes and labels and consequently into training and test sets.

Here the *x* variable contains all the columns from the dataset, except the ‘price’ column, which is the label. The y variable contains values from the ‘price’ column, which means that the *x* variable contains the attribute set and y variable contains the corresponding labels.

## 

## **Results of Random Forest Regression:**

In the analysis, 1000 trees were created. In general, the more trees give the better results. As a result, 64569.56 RMSE and 43503.0648 mean absolute error (MAE) is obtained.

**Results for Random Forest Regression:**

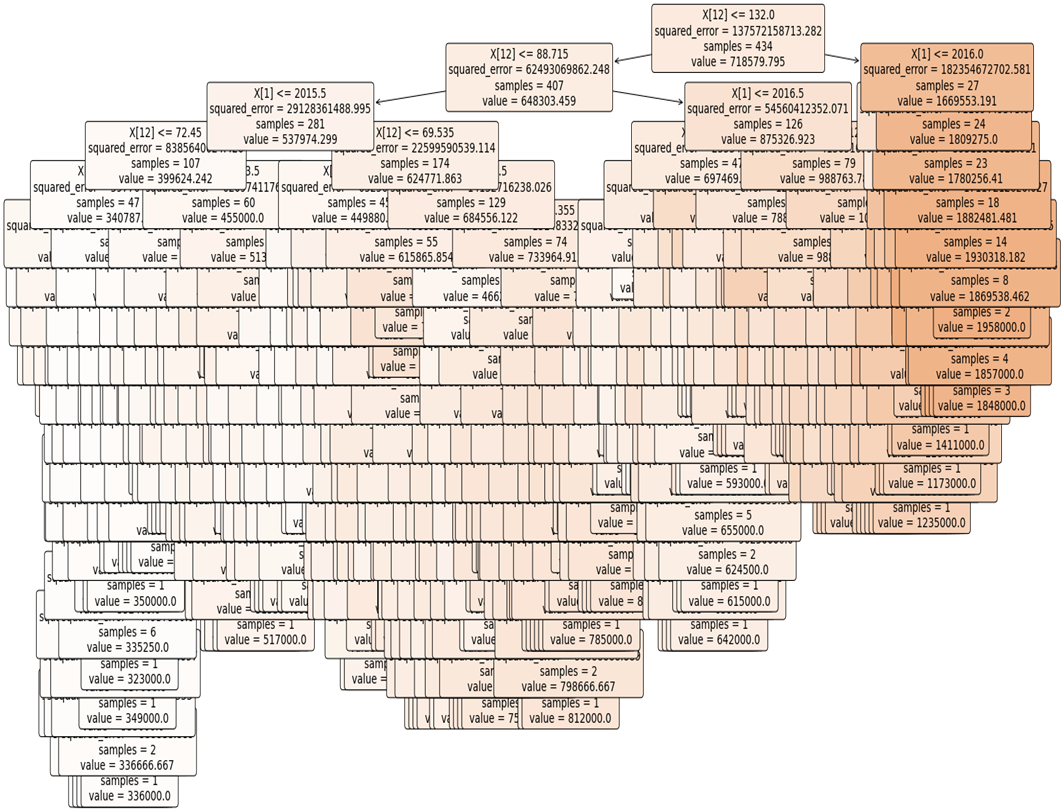
RMSE : 64569.56830879093

R2 : 0.9688532169892883

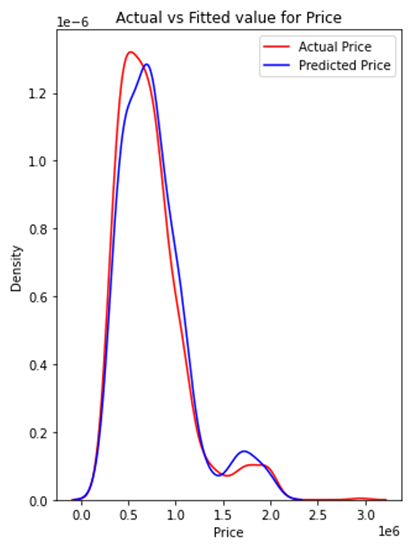
MAE : 43503.0648 degrees

Accuracy : 93.7%.

Look at Fig. 3.5 below, it is observed that there are multiple leaf nodes. To classify a new point, simply move down the tree, using the features of the point to answer the questions until you arrive at a leaf node where the class is the prediction.



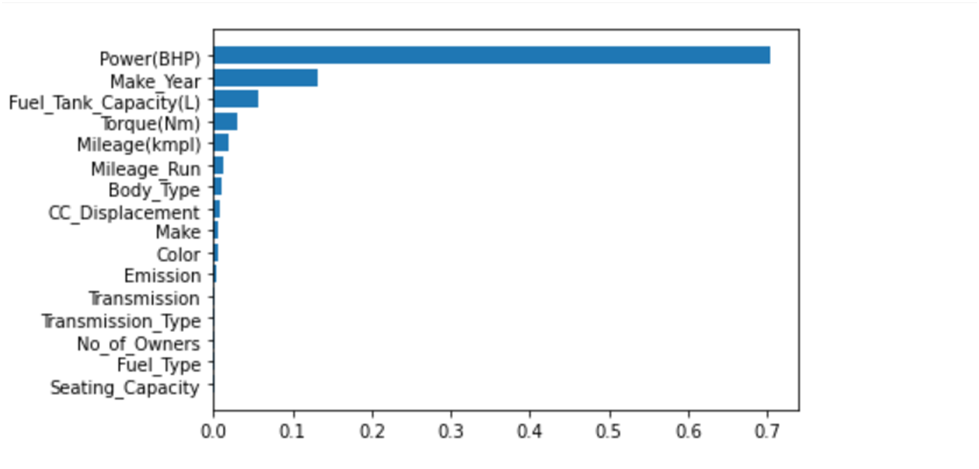
**Fig. 3.5 Random Forest Tree**



**Fig. 3.6 Random Forest Model- Actual vs Predicted**

The above Fig. 3.6 shows the difference between the actual and predicted values where we can see that there is less difference between our predicted and actual values.

## **Feature Importance:**



**Fig. 3.7. Bar plot for Feature Importance**

To quantify the usefulness of all the variables in the entire random forest, it can look at the relative importance of the variables. Fig. 3.7 is a simple bar plot of the feature importance to illustrate the disparities in the relative significance of the variables. The number of features importance is 7 (Fig. 3.7). The features are Power (BHP), Make Year, Fuel Tank Capacity, Torque (Nm), Mileage Run, Mileage (kmpl) and CC Displacement. The ultimate purpose of modelling is to get a smaller number of features that can give us a strong prediction. At this point, the model was run with only these six important features. In addition, this performance was obtained just by using 7 features instead of 20. Therefore, it can be considered as an improvement in both prediction power and computational cost.

# **3.5 XG Boost Regressor:**

XG Boost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

When using gradient boosting for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of the leaves that contains a continuous score. XG Boost minimizes a regularized (L1 and L2) objective function that combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity (in other words, the regression tree functions). The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It’s called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

**Unique features of XG Boost:**

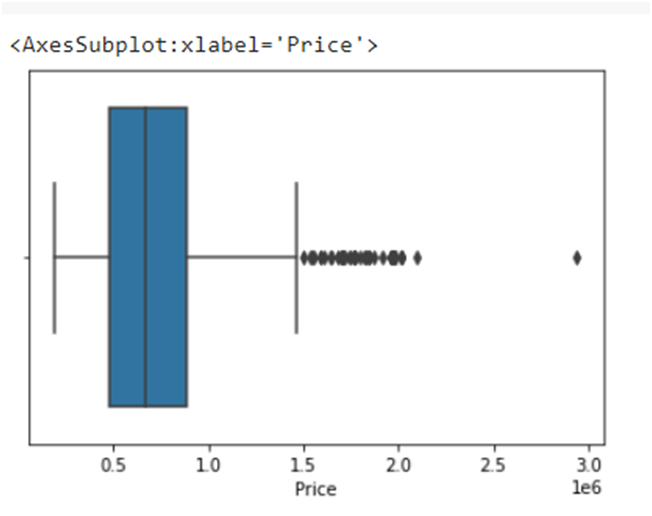
Regularization, handling sparse data, Weighted quantile sketch, Block structure for parallel learning, Cache awareness, Out-of-core computing and identifying and removing outliers with box plots.

**Outlier Detection**

An outlier may be due to variability in the measurement, or it may indicate an experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. Outliers present in data can skew the model and disrupt its learning process. Hence, now we will detect these outliers in our dataset and remove them using the Box plot (Inter-Quartile Range) method.

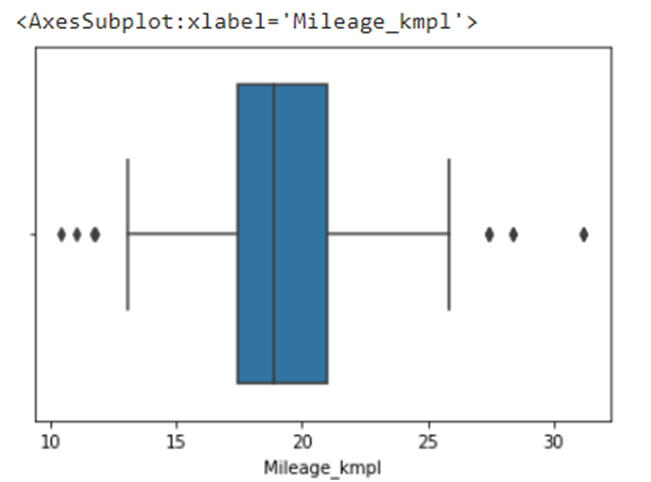
The Box-Whisker plots are drawn based on median quartiles and extreme values. The Whiskers are the lines extended from the box to the maximum and minimum values in the data, excluding outliers. The plot consists of extreme points which occurred in above values of the observations of the data. A line across the box indicates the median. For positively (negatively) skewed distribution, the Whisker will be too long from the box to the maximum (minimum) value in the data. For symmetrical reasons, the Whiskers will be of equal length on both sides of the box.

Here the Box-Whisker plot for price of the car resale data is shown in Fig. 3.8. It is observed that the outlier detected above price value Rs. 1500,000 car resale data.

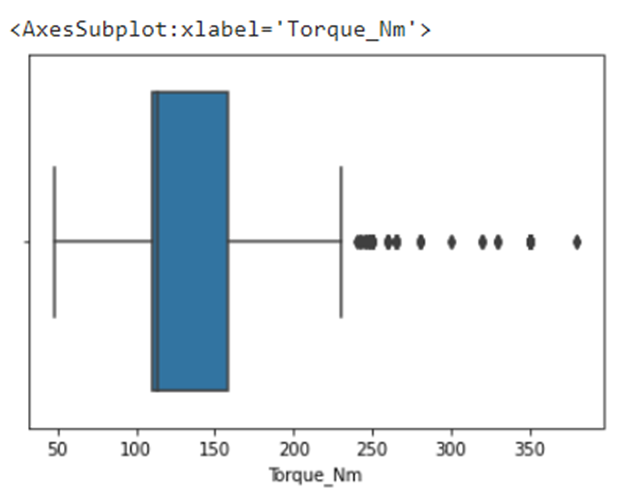


**Fig. 3.8 Box-Whisker Plot for Price**

The Box-Whisker plot for mileage (in kmpl) of the car resale data is shown in Fig. 3.9. From this Figure, it is observed that the outlier lies between 13 to 25 kmpl.

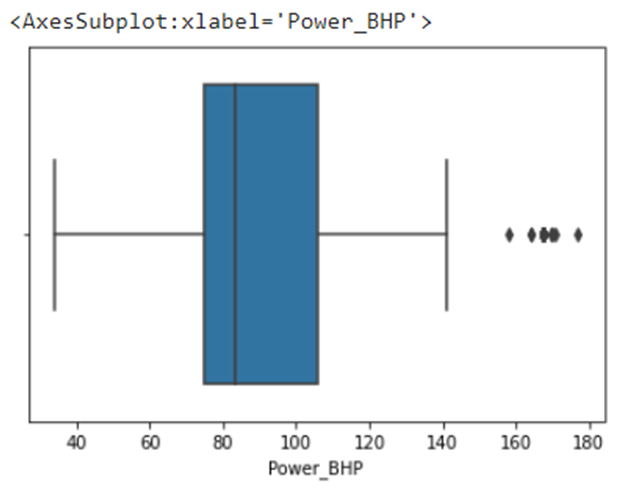


**Fig. 3.9 Box-Whisker’s Plot for Mileage (in kmpl)**



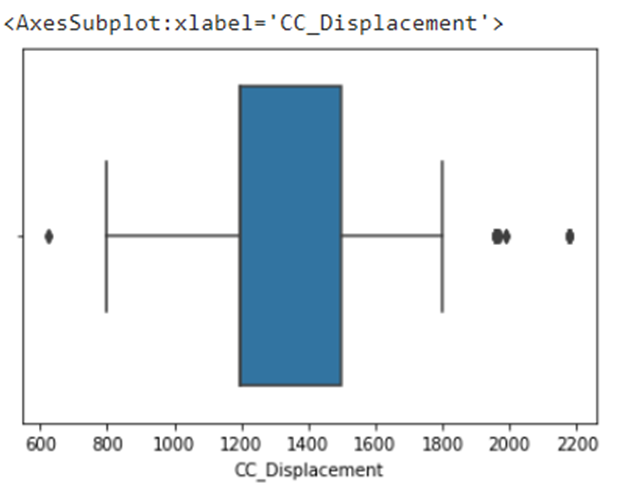
**Fig. 3.10 Box-Whisker Plot of Torque (in Nm)**

The Box-Whisker plot for torque (in Nm) of the car resale data is shown in Fig. 3.10. From this Figure, it is observed that the outlier lies after 240 Nm values of the data.



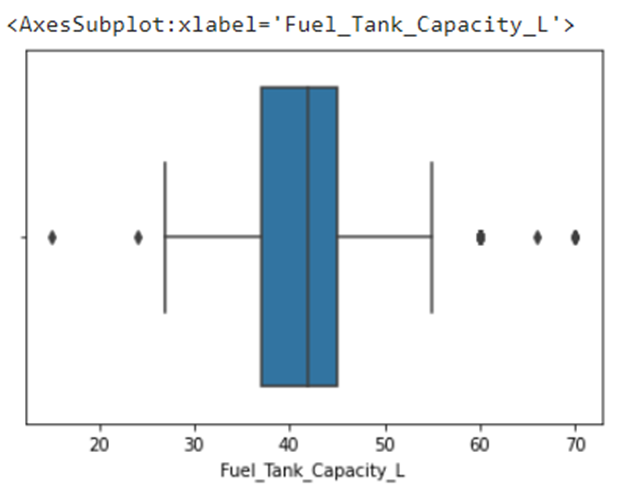
**Fig. 3.11 Box-Whisker Plot of Power (in BHP)**

The Box-Whisker plot for power (in BHP) of the car resale data is shown in Fig. 3.11. From this Figure, it is observed that the outlier lies after 155 BHP values of the data.



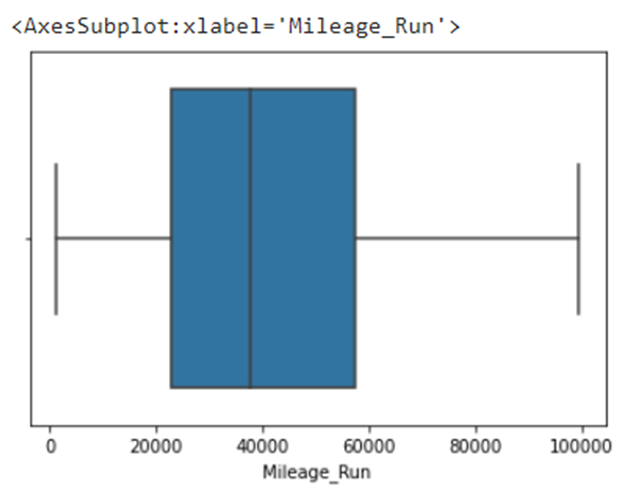
**Fig. 3.12 Box-Whisker Plot for CC Displacement**

The Box-Whisker plot for CC displacement of the car resale data is shown in Fig. 3.12. From this Figure, it is observed that the outlier lies between 800 to 1800 values of the data.



**Fig. 3.13 Box-Whiskers Plot for Fuel Tank Capacity (in L)**

The Box-Whisker plot for price of the car resale data is shown in Fig. 3.13. From this Figure, it is observed that the outlier lies between 28L to 55L values of the data.



**Fig. 3.14 Box-Whisker Plot for Mileage Run**

There is no outlier in the Box-Whisker plot for mileage run.

Then these outliers are removed. Here the X variable contains all the columns from the dataset, except the ‘Price’ column, which is the label. The y variable contains values from the ‘Price’ column. Now standardize the variables and split it into train (70%) and validation (30%) sets.

**Training and Testing using XG Boost model:**

XG Boost model to train on the data and predict the output. It is also use the GridSearchCV from sklearn module to identify the best parameters for the model. The parameters tuned using GridSearchCV are:

* n\_estimators: This defines the number of estimators (trees) used by the XGBoost model.
* max\_depth: This defines the maximum depth allowed by each tree and is used to control overfitting.
* eta: This is the learning rate of the model which shrinks the weights of the nodes to make it more robust.

To get an unbiased estimate of the model performance, use the K-fold cross validation method with cv =5. The best parameters are identified by using the RMSE as the error metric. Since we need to find the model which produces the least RMSE, pass the negative of the RMSE to the scoring argument. After the parameter search is complete, it can now view the best parameters identified and their corresponding metric score.

Using these parameters identified by GridSearchCV it is built the final model and use it for prediction. Here employing RMSE and R² (coefficient of determination) as the error metrics. The score () function of XG Boost returns the R² value for a regression problem.

## **Results of XG Boost:**

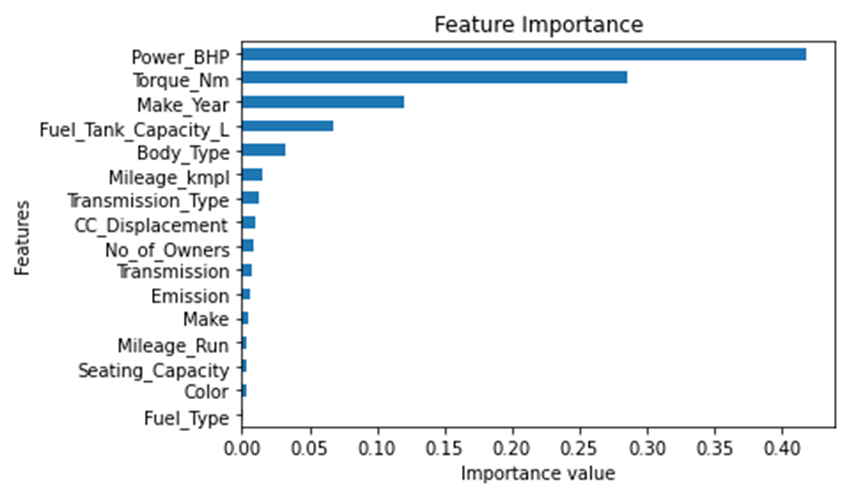
RMSE : 65023.88789801809

R2 : 0.929755087642493

MAE : 45125.27 degrees

Accuracy : 93.13 %.

F**eature Importance:**



**Fig. 3.15 XG Boost Model- Feature Importance**

In order to quantify the usefulness of all the variables in the entire XG Boost model, it can look at the relative importance of the variables. Fig. 3.15 is a simple bar plot of the feature importance to illustrate the disparities in the relative significance of the variables. The number of features importance is 7 (Fig. 3.15). The features are Power (BHP), Torque (Nm), Make Year, Fuel Tank Capacity, Body type, Mileage (kmpl) and Transmission Type. The ultimate purpose of modelling is to get a smaller number of features that can give us a strong prediction. At this point, the model was run with only these seven important features. In addition, this performance was obtained just by using 7 features instead of 20. Therefore, it can be considered as an improvement in both prediction power and computational cost.

**3.7 COMPARISON:**

Given the model evaluation parameters the Random Forest Regressor outperformed as it has the highest accuracy of 93.27% and R2 value is 0.9689 of the five different algorithms, as well as the lower error in all three-evaluation parameters. Followed by Random Forest Regression, XG Boost Regressor has high accuracy of 93.13% and the error is also less compared to all the algorithm except Random Forest.

**Fig. 3.17 Graphical Illustration of Accuracies**

**CHAPTER Ⅳ**

**Summary of Results**

As per the reports, the growth of new cars in the upcoming 5 years is an average of 3.5% while, for used cars 5% and the ratio of used cars to new cars is increasing continuously. So, the used cars market is increasing day by day and a consumer wants the best resale price of the car.

In this chapter the results obtained in the preceding chapter are summarized. By performing different models, it was aimed to get different perspectives and eventually compare their performance.

With this study, we predict the prices of used cars by using a dataset that has 20 predictors and 976 observations.

With the help of the data visualization and exploratory data analysis, the dataset was uncovered, and features were explored deeply. At the last stage, predictive models were applied to predict price of cars in an order: Linear regression, Ridge regression, Lasso, Random Forest regression, and XG Boost. Below are the results observed in all the five models in the table.

**Table 1. Comparison of Regression Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | R2 | RMSE | MAE Value | Accuracy |
| Linear Regression | 0.8779 | 119223.04 | 95325.73 | 85.8%. |
| Lasso Regression | 0.8779 | 119223.30 | 95326.28 | 85.8%. |
| Ridge Regression | 0.878 | 119299.99 | 95381.13 | 85.8%. |
| Random Forest Regression | 0.9689 | 64569.57 | 43503.07 | 93.27%. |
| XG Boost | 0.9298 | 65023.89 | 45125.27 | 93.13%. |

By considering all four metrics from the above table 1, it can be concluded that random forest is the best model for the prediction for used car prices. Random Forest as a regression model gave the least MAE and RMSE values. According to random forest, the most important features are Power (BHP), Make Year, Fuel Tank Capacity, Torque (Nm), Mileage Run, Mileage (kmpl) and CC Displacement.

The value of a car drops right from the moment it is bought and the depreciation continues with each passing year. In fact, in the first year itself, the value of a car decreases by 20 percent of its initial value. The make and model of a car, total kilometers driven, overall condition of the vehicle and various other factors further affect the car’s resale value.

From these above results it would suggest that Random Forest is perhaps worthy of further study on this problem. Among all the methods it is the best of all which gives the best results. It has the lowest MSE and RMSE. It refers to random forest for better predictions. So, one can choose Random Forest Regression as the better model to study this dataset.

This project is more helpful for all e-commerce companies who act as mediators for selling and buying pre-owned cars. The customer can easily be convinced in taking a decision to buy a pre-owned car out of various car models with various features. The seller can easily convince the buyer by comparing and analyzing various models. The seller and buyer are both satisfied with this process. This model reduces time and cost and is also more user friendly because of which there is improvement in business by selling more cars. Here we are also conducting a comparative study on performance of regression based on supervised machine learning models. Each model is trained using data of used car market collected from e-commerce websites. As a result, Random Forest regression gives the best performance with R2 = 0.9689, followed by XG Boost and Ridge Regression.