**Car Price Prediction**

**Raja Faizan Nazir1, Fellow, IEEE, Huria Imtiaz2, Member, IEEE**

1University of Management and Technology, Pakistan

2 University of Management and Technology, Pakistan

Corresponding author: Mazhar Javed Awan (e-mail: mazhar.awan@umt.edu.pk).

**ABSTRACT** Travelling from one place to another has evolved over the centuries, back in the 1800s nomads would rely on animals such as the mighty horse or camels to ensure safe travels. As time flew by, science and technology birthed various ideas to improve and enhance the mode of travelling. It wasn’t until 1839 that the essence of automobiles came into existence due to the works of Karl Benz, in Germany. Since then to date, automobiles have changed the face of technology. Not only did they decrease travelling time to just a few minutes but have managed to add value to people’s lifestyle. As time progressed, various versions and automobile companies came into being, which not only made the industry competitive but the booming industry turned out to be heavy for many pockets. Over 1 billion cars run around the streets of this globe, which has put the vehicle saturation to over 18%. The study under evaluation will discuss how valiant features that make up a car can help consumers predict its price and be a better judge of the automobile adding value to their lifestyle.

**INDEX TERMS** Car Price, Automobiles, Vehicle Saturation, Travelling

1. **INTRODUCTION**

A modern world calls for a modern way to enlist pricing of automobiles or cars that consumers want to purchase. Where technology has made life easier, why not predict car’s pricing before running around show rooms to look for the perfect one. In any competitive industry, one must keep in mind what variables will add value to their product and what will sell most out in the real world. Many factors and variables add significance to a vehicle, such as the engine type[1], model, fuel type, horse power, car length etc. Any manufacturing company will have to keep these features along with many other when declaring the price of a car.

In the 21st century, where electric cars are being released and have had a lot of heads turning, the first question that comes to our mind is “How much would this car cost?”. Using modern machine learning and computing methods we have been able to predict car pricing. A common question that arises is what factors contribute A modern world calls for a modern way to enlist pricing of automobiles or cars that consumers want to purchase. Where technology has made life easier, why not predict car’s pricing before running around show rooms to look for the perfect one. In any competitive industry, one must keep in mind what variables will add value to their product and what will sell most out in the real world. Many factors and variables add significance to a vehicle, such as the engine type[1], model, fuel type, horse power, car length etc. Any manufacturing company will have to keep these features along with many other when declaring the price of a car.

In this paper, we have made use of 27 different variables that will help us put a number on the car. The dataset taken into consideration as a reference also consists of 25 dependent and independent variables to support the paper such as, engine size, bore ratio, curb weight, cylinder number, car width, fuel system, wheel base and many more

In this paper, we have made use of 27 different variables that will help us put a number on the car. The dataset taken into consideration as a reference also consists of 25 dependent and independent variables to support the paper such as, engine size, bore ratio, curb weight, cylinder number, car width, fuel system, wheel base and many more

1. **Previous Work**

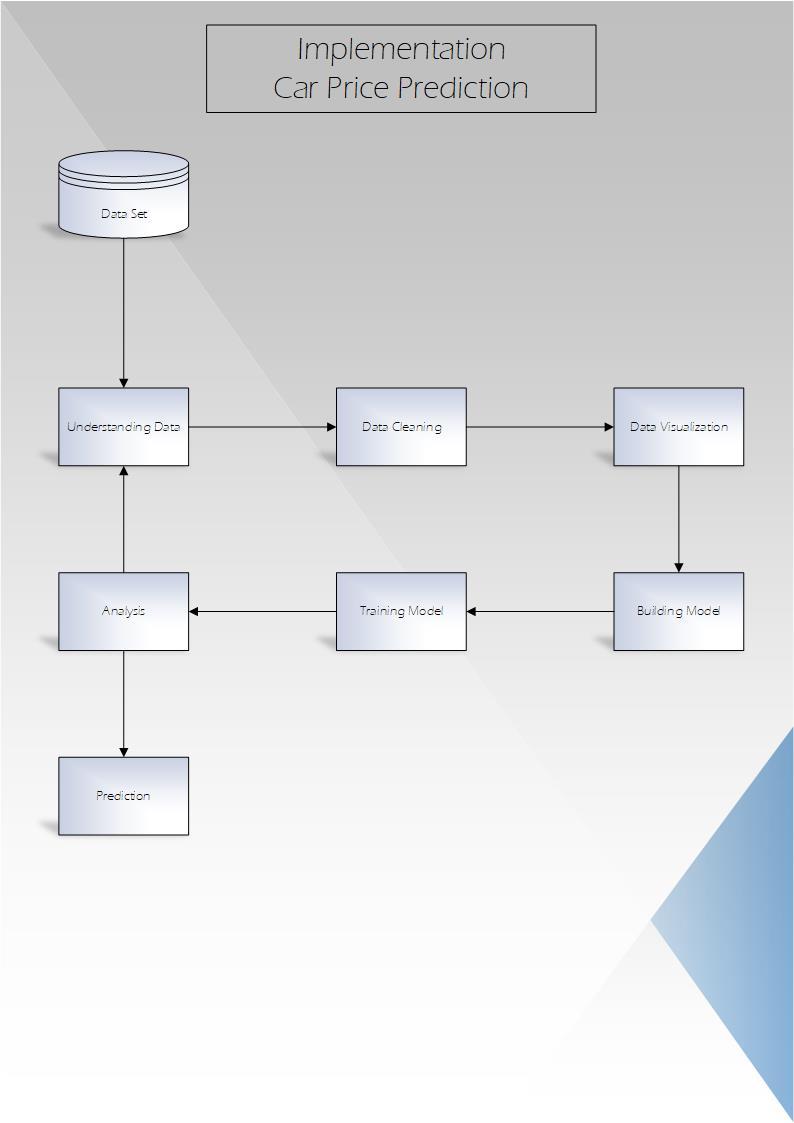
A Post graduate Data Science student [Shalini Goyal](https://www.linkedin.com/in/shalinigoyal/) processed this Dataset using Linear regression, since rows was less so she trained the model multiple time that is why she got 86.1% accurate result. But it was simple linear regression. When we tried linear regression, our accuracy was about 74% because our data was less and we did not train model with random values what she did.

Be used linear regression and another model Random Forrest to get accurate predictions. Although her error rate was more and by error rate with random forest is comparatively less because of the model

1. **PRELIMINARIES**

IMPLEMENTATION

Once the data set has been entered and retrieved, by following the given implementation stages, our dataset will produce results by giving the car pricing predictions.



The figure depicts the steps being followed to reach implementation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| CAR ID | Symboling | Car Name | Fuel Type | Aspiration | …... | Horse Power | Price |
| 1 | 3 | Alfa Romero | gas | STD | …... | 111 | 13495 |
| 2 | 2 | Alfa Romero | gas | STD | …... | 111 | 16500 |
| 3 | 2 | Audi 100ls | gas | STD | …... | 154 | 16500 |
| 4 | 2 | Audi 100ls | gas | STD | …... | 102 | 13950 |
| 5 | 2 | Audi Fox | gas | STD | …... | 115 | 17450 |
| 6 | 1 | Audi 100ls | gas | Turbo | …... | 110 | 15250 |
| 7 | 1 | Audi 5000 | gas | Turbo | …... | 110 | 17710 |

Variables used in Data Set

These are the variables we have in the Dataset

Car Price

We have taken over 200 cars into consideration in order to predict their prices with their corresponding variables. What we learnt from the data set can be simply interpreted through this word cloud. Nissan and Toyota have reigned our data set and have come out triumphed when it comes to pricing.

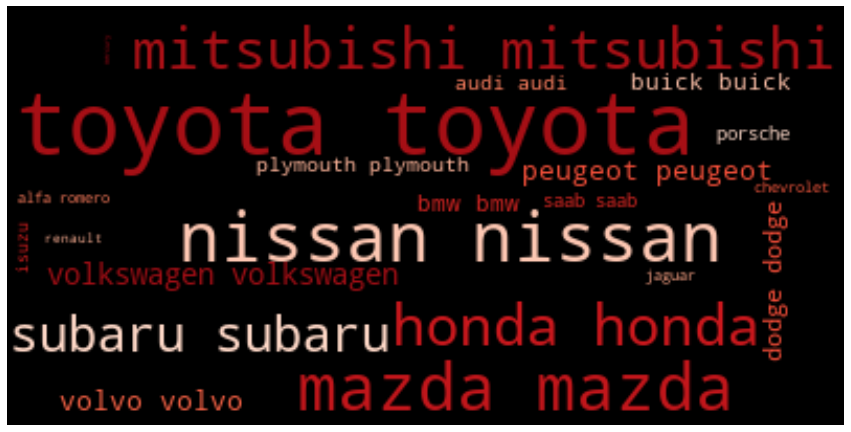
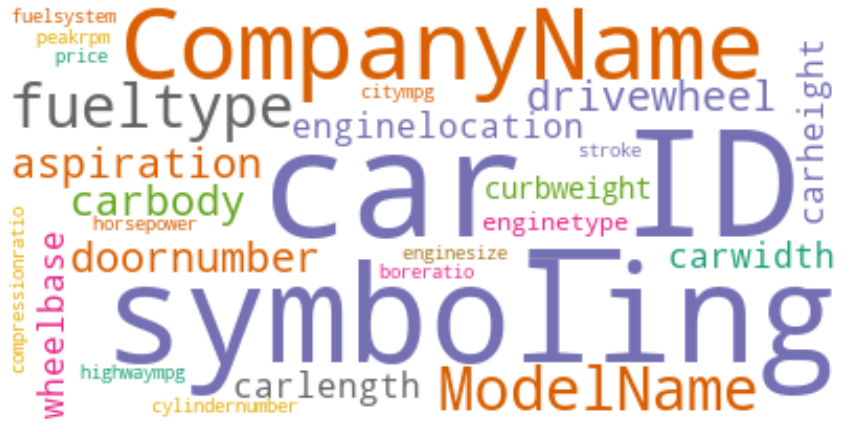


Figure 2.1 Word Cloud of the data set

The following table shows the variables being used in the study and the corresponding values obtained against them. We have shown in the table but this is a easy to read way to get idea of all companies



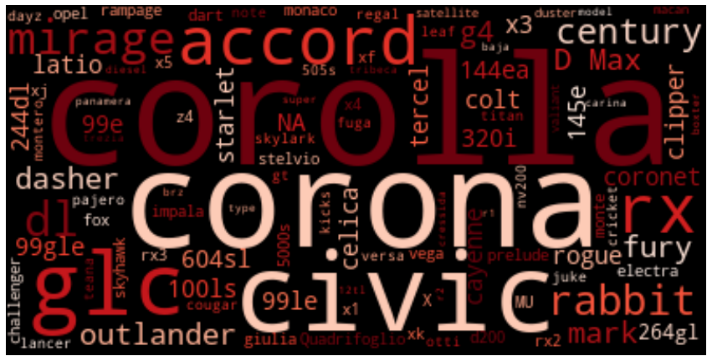
Column in Images (Easy to Read)

We have CAR ID in rows, Company name, Model name

1. **PREPROCESSING**

In the preprocessing we had to remove some rows and split some rows.

Like in the Data set, Name of Car were and Model were merged in one column so we had to split Company Names and Model names



Thanks to this visualization, I found out that in a lot of columns name of model ‘Corolla’ were written as ‘Corona’ and also some other models name were not correct so we dropped this column.

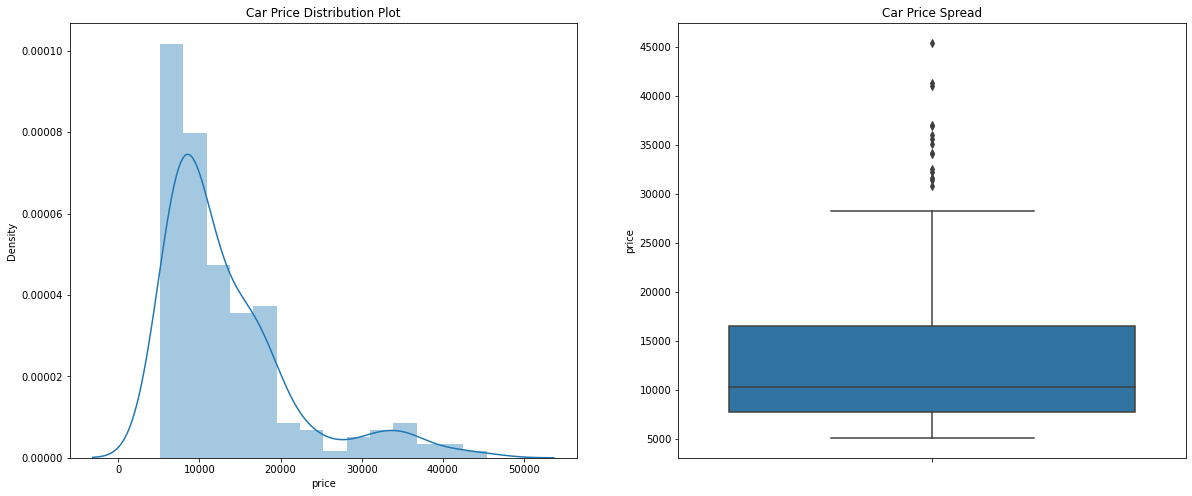


Figure 2.3 Density of price

The above plot signifies the density of pricing, which is around 7800 people per km2. The box plot seems to be right skewed, which means that most of the prices in our data sets is below 15000. We can easily see a great difference between the price distribution in the plots. High variance in car pricing can be spotted given the distance between the data points.

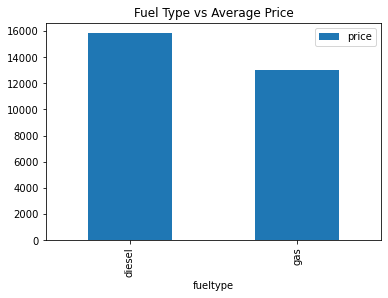


Figure 2.4 Fuel Type vs Average price

Price is directly proportional to the type of fuel used. The more expensive fuel type used, the more is the price, we can see in the graph above.

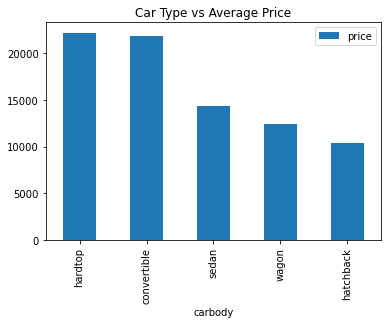


Figure 2.5 Car Type vs Average Price

According to the data set sorted, we see that convertibles and hardtop cars have been prices higher than sedans, wagons or hatchbacks.

We have also added a word cloud depiction of popular car type using our data set



Figure 2.6 Word Cloud of Car Type

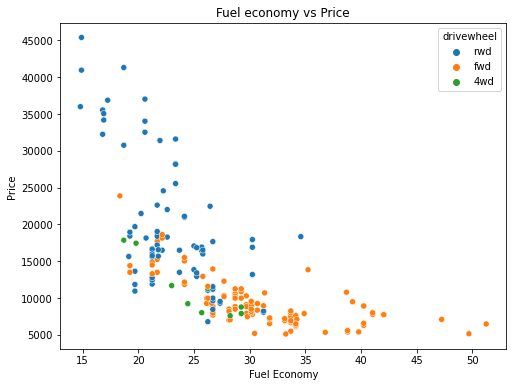


Figure 2.7 Fuel economy vs Price

Data set revealed that fuel economy has a negative correlation with price.

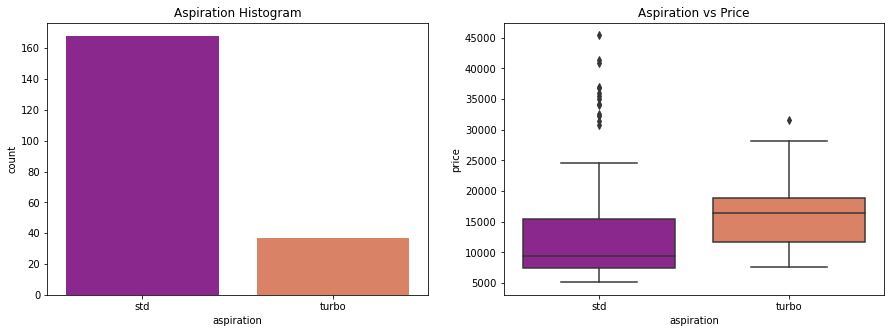


Figure 2.8 Aspiration vs Price

Aspiration with a turbo[2] engine are pricier than that with standard ones. Aspiration refers to the natural combustion of the engine without the need of any turbo charger etc.

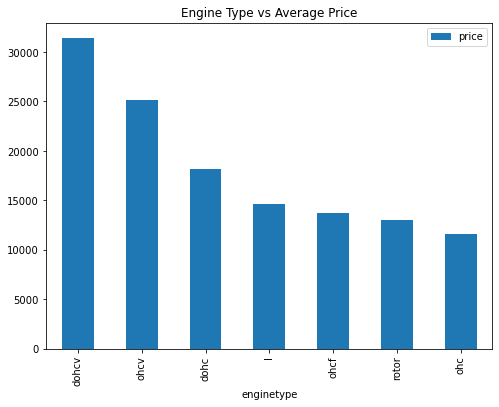


Figure 2.9 Engine Type vs Price

According to the revelations of our data set ohc engine type has been favored the most out of all types.

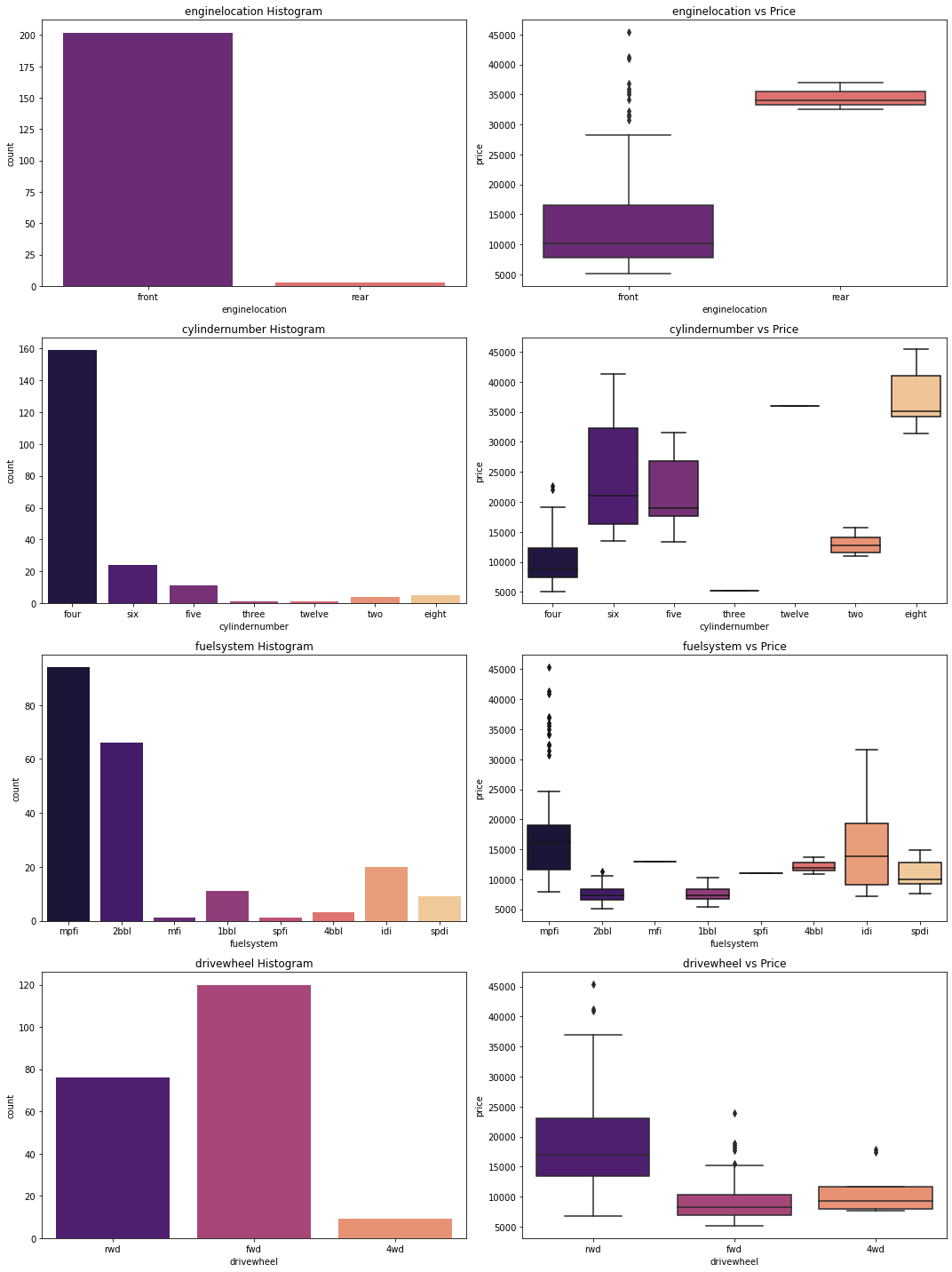


Figure 2.10 Fuel System

There aren't enough data points for engine location categories to draw any conclusions. The most common cylinder counts are four, six, and five. The pricing of an eight-cylinder engine is the most expensive. The most prevalent fuel systems are mpfi and 2bbl. The most expensive options are mpfi and idi. However, there is insufficient data for other groups to make any useful inferences. There's a big variation in the drivewheel category. The majority of high-end vehicles appear to prefer rear-wheel drive.

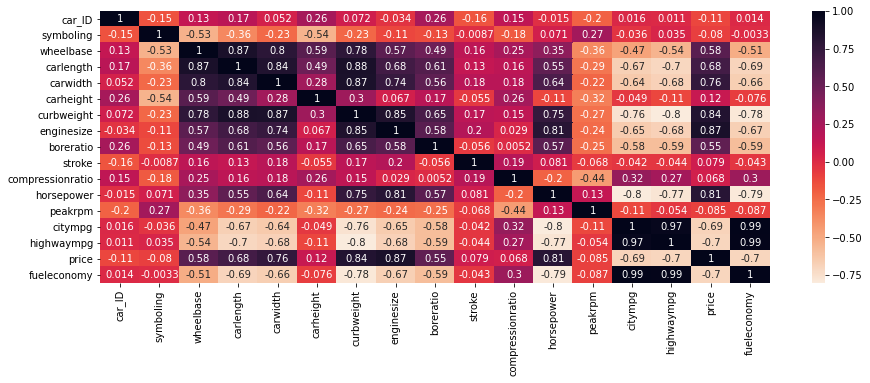


Figure 2.10 Correlation using heatmap

Data set suggests that the most highly co related variables with price are horsepower, enginesize, curbweight, carwidth and highend.

1. **Materials and Methods**

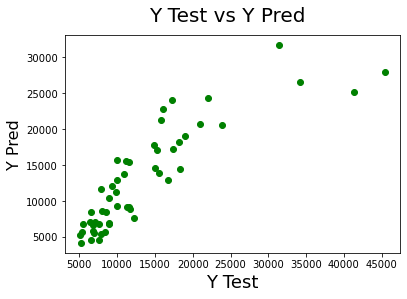
* Models

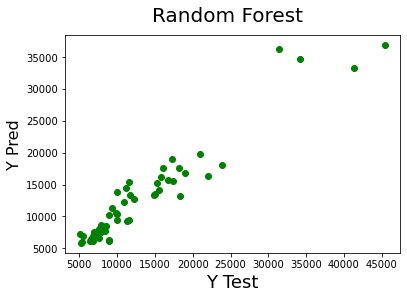
Linear Regression

Random Forest

1. **Results**

The following graph depicts the tested vs predicted pricing results, we used a training model to predict and test the results to ensure accuracy.

 Figure 3.1 Test vs Pred using Linear Regression

 Figure 3.2 Test using Random Forrest

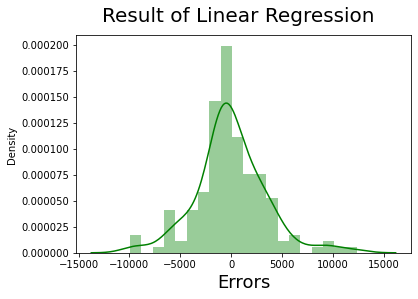


Figure 3.2 Result using **LR**

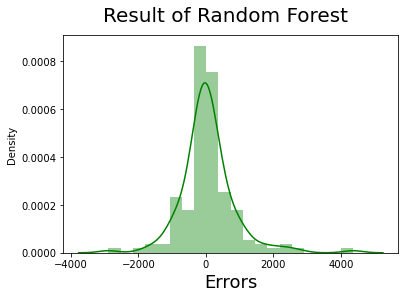


Figure 3.3 Result using **RF**

Both models have shown accuracy but only one managed to reach optimum accuracy which was the goal of the study, to release car pricing predictions to it’s utmost accuracy. The following table gives the percentages of accuarate results obainted from both the models.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Random Forest | 90% |
| Linear Regression | 74% |

As we can see in the above graph that the error terms seem to be distributed evenly across the board, we can safely say that the assumptions on the linear model were fulfilled

.

1. **Discussion**

In this research we were able to study how various variables can add to the prediction of getting a quote on a car. We learnt that however value the asset, it will only add more significance to the pricing being laid. For instance, cars that use diesel as a fuel type, keeping in view diesel is expensive than gas, are found to be costlier. By using the linear model, we were able to draw comparisons and reach to our utmost accurate pricing predictions for cars. Using the word cloud, we figured which companies were more popular than the rest in the dataset used for references. We have also s As an updated edition to this paper, we can also predict the taxes implied by Governments these cars will be manufactured or imported to so that the exact price of the vehicle can be quoted to the consumer or buyer.

1. **Author Contribution**

“Conceptualization, A.H. and S.A.; methodology, S.A.; software, A.H.; validation, A.H. and S.A.; formal analysis, A.H.; investigation A.H.; resources, A.H.; data curation, A.H. and S.A.; writing—original draft preparation, S.A.; writing—review and editing, S.A.; visualization, A.H.; supervision, S.A.; project administration, S.A.; funding acquisition, None. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not Applicable for this study.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Data set used is available at following google drive link:

**Conflicts of Interest:** The authors declare no conflict of interest.

1. **References**

[1] A. Kapustin and V. Rakov, "Methodology to evaluate the impact of hybrid cars engine type on their economic efficiency and environmental safety," *Transportation Research Procedia,* vol. 20, pp. 247-253, 2017.

[2] R. Capata and E. Sciubba, "The LETHE©(Low Emissions Turbo-Hybrid Engine) city car of the University of Roma 1: Final proposed configuration," *Energy,* vol. 58, pp. 178-184, 2013.