Predictive Analysis Case Study

SUBMISSION

Linear Regression Model to Predict the price of cars and determine the factors driving car price

CASE STUDY OVERVIEW

PROBLEM STATEMENT

A Chinese automobile company Geegly Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

- Which variables are significant in predicting the price of a car

- How well those variables describe the price of a car

BUSINESS GOAL

I am required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

DATA UNDERSTANDING

Data Reading and Understanding

The data set contains 205 entries and 29 columns. The names of those columns are:

Company, cars range, Symboling, fuel type, engine type, carbody, door number, engine location, fuel system, cylinder number, aspiration, drive wheel, Car_ID, car length, car width, car height, car volume, curb weight, Horsepower, Bore Ratio, Compression Ratio, Highway miles per gallon (mpg), Engine Size, Stroke, City Miles per gallon (mpg), Fuel economy, Peak Revolutions per Minute (rpm), Wheel Base, Price.

Here, our **target variable** is **PRICE.**

ASSUMPTIONS AND DATA HANDLING

Assumptions and Data Handling

Car Company: We have assumed that volkswagen, vokswagen, and vw are the same companies, and same for others.

Data Cleansing: We have renamed the car company names to their correct names, as per our understanding. We have converted all the data to lower case to avoid any case errors. The **duplicated** function searched for any duplicate values in our data and found none.

There were no missing values in the data and by performing the above steps, we prepared our data for analysis.

A separate data set **corr** was created that dealt only with the correlation of our target variable, price. This was done in order to select the best response variables for our study.

Another dataset **cars** was created that included only the columns that we selected based on our data exploration.

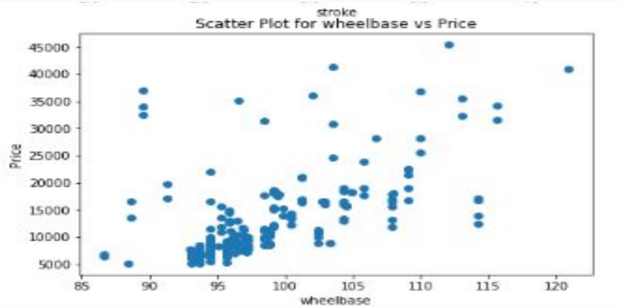
DATA EXPLORATION

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

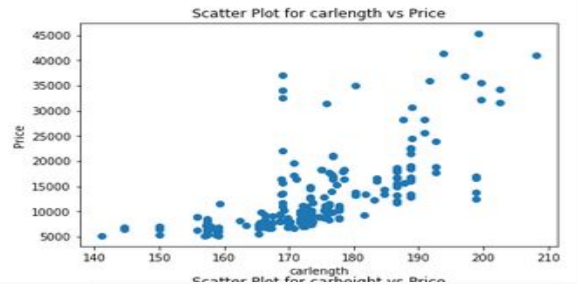
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

1. Wheel Base



It seems to have a significant positive correlation with price.

2. Car Length



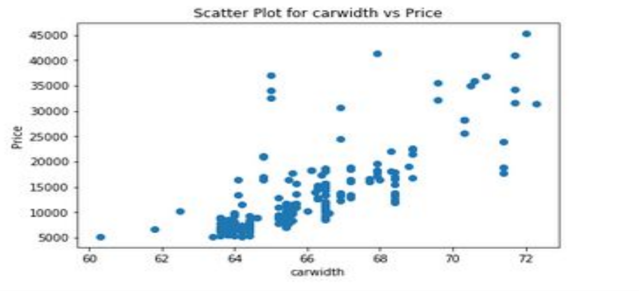
It seems to have a significant positive correlation with price.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

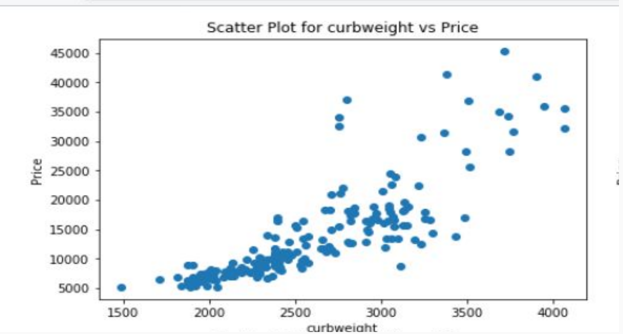
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

3. Car Width



It seems to have a significant positive correlation with price.

4. Curb Weight



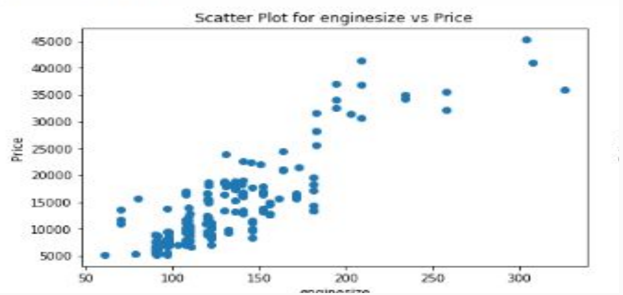
It seems to have a significant positive correlation with price.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

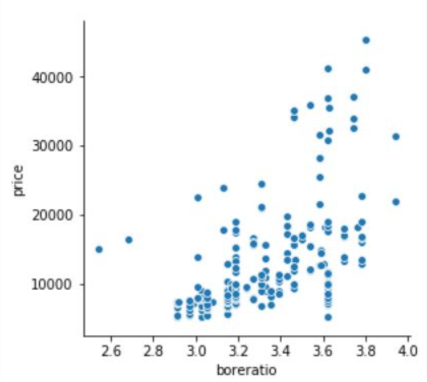
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

5. Engine Size



It seems to have a significant positive correlation with price.

6. Bore Ratio



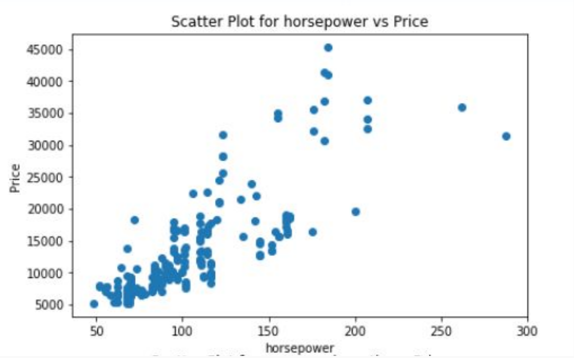
It seems to have a significant positive correlation with price.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

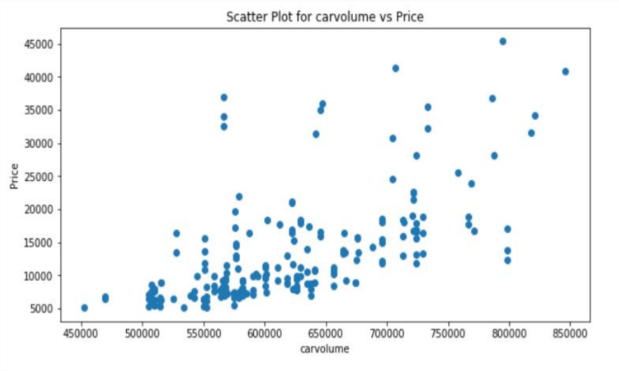
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

7. Horse Power



It seems to have a significant positive correlation with price.

8. Car Volume



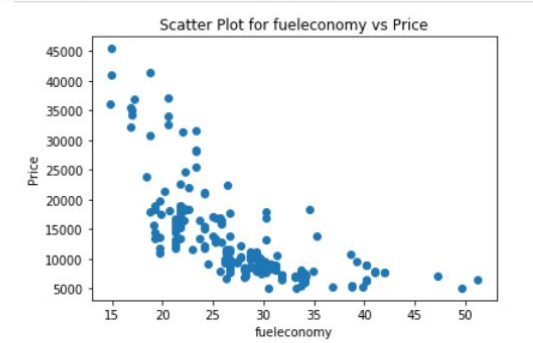
It seems to have a significant positive correlation with price.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

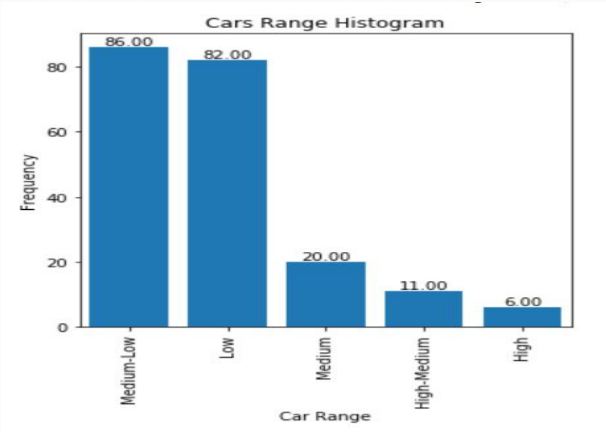
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

9. Fuel Economy



It seems to have a significant negative correlation with price.

10. Cars Range



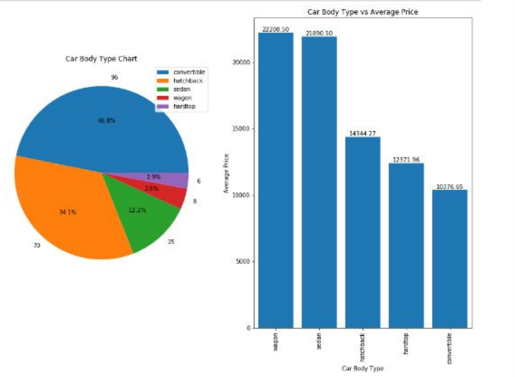
It seems that most people prefer medium-low range and low range cars.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

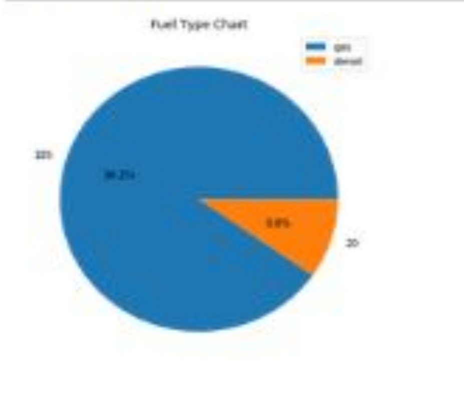
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

11. Car Body



It seems that people prefer convertible due to its low price range

12. Fuel Type



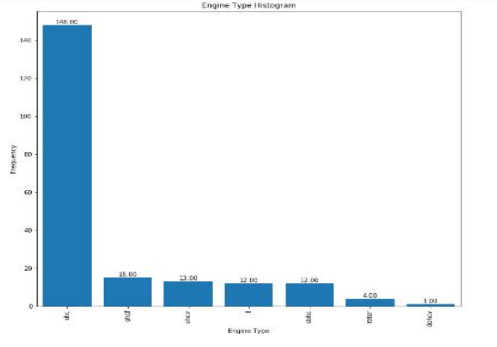
It seems that most people prefer gas cars despite their high price range.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

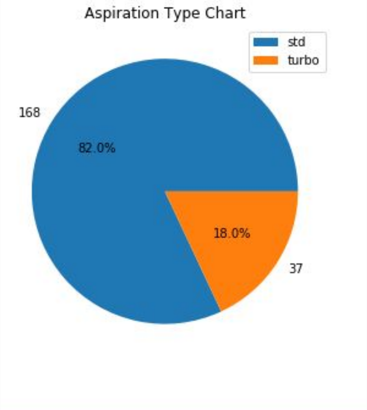
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

13. Engine Type



It seems that people prefer ohc engine type for their cars.

14. Aspiration



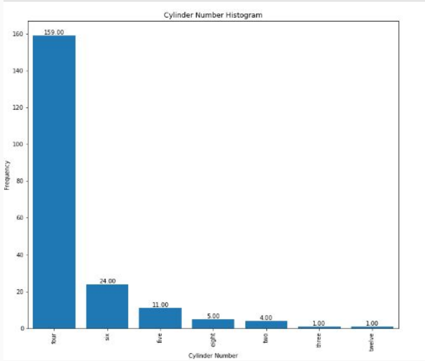
It seems that most people prefer std aspiration for their cars.

Exploratory Data Analysis

Based on our data exploration, we chose the best features that would help us predict the prices of the cars. These are :

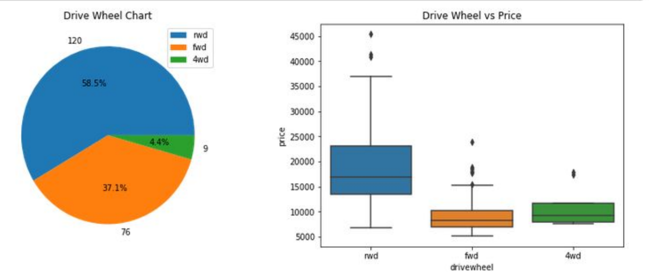
Wheel Base, Car Length, Car Width, Curb Weight, Engine Size, Bore Ratio, Horsepower, Car Volume, Fuel Economy, Cars Range, Car Body, Fuel Type, Engine Type, Aspiration, Cylinder Number, Drive wheel

15. Cylinder Number



It seems that people prefer four cylinders for their cars.

16. Drive wheel



It seems that most people prefer rwd drive wheel despite of its high price range.

TECHNIQUE COMPARISON

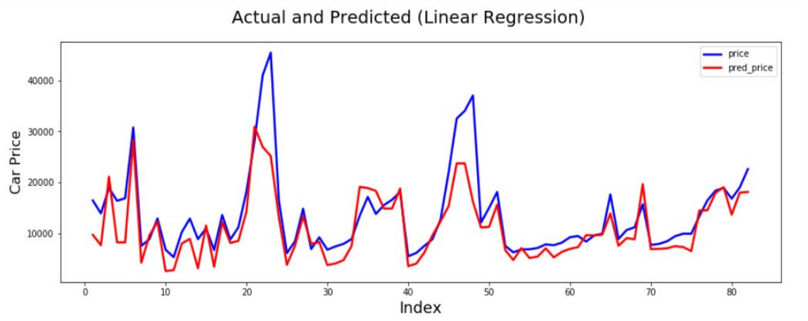
Simple Linear Regression vs Random Forest Regression

Technique Comparison

<table>
<tr><th>Simple Linear Regression</th><th>Random Forest Regression</th></tr>
<tr><td>

In Simple Linear Regression, Our Accuracy Score was **64.377%.**

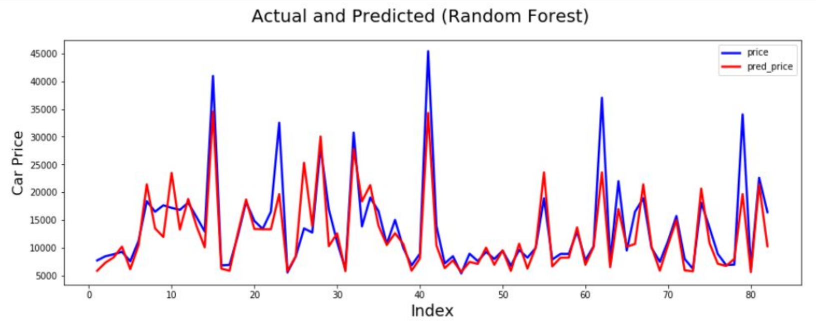
When plotted a graph between actual and predicted price, we could see some significant irregularities.



</td><td>

Using Random Forest, Our Accuracy Score was **76.147%.**

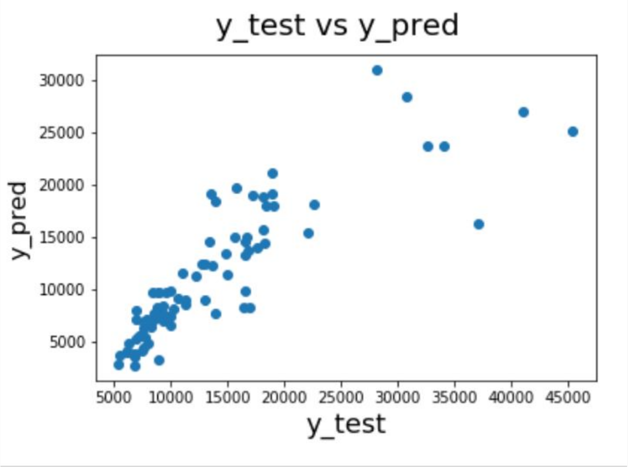
When plotted a graph between actual and predicted price, we noticed negligible irregularities.

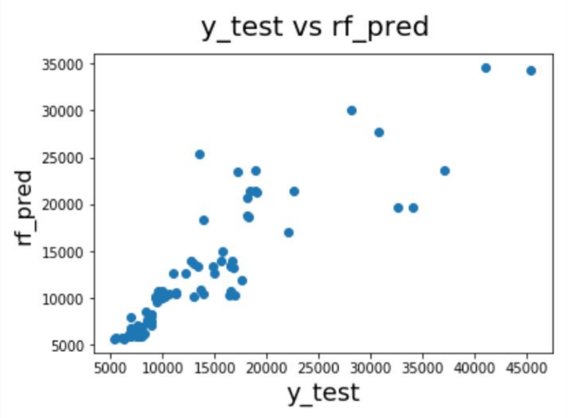


</td></tr>
</table>

Technique Comparison

<table>
<tr><th>Simple Linear Regression</th><th>Random Forest Regression</th></tr>
<tr><td>When checked the price spread between values, we can clearly see some data points scattered too far away from the rest.


</td><td>When checked the price spread between values, we can clearly see data points scattering only after a little while, which is earlier than what we saw in simple linear regression.


</td></tr>
</table>

SUGGESTIONS AND INSIGHTS

Suggestions and Insights

1. Cars Range is one of the most significant variables for deciding price. Hence, the company should select their price range within 18500$ to 30000$ initially.

2. Fuel Economy of the car is very important for the customers these days. The company should make sure the fuel economy of cars is not less than 5 litres/100km.

3. The company should make sure to have cars with drive wheel 'rwd' because people prefer it despite its high price.

4. Most of the cars should have petrol fuel type, because people prefer petrol way more than diesel.

5. Most people prefer convertible despite its price range being medium-high. Hence, company should focus on this as well.

NOTE:

A Car is an Asset to anyone who owns it. Hence, choose your car very carefully, whether for business or for personal use.

Thanks! You can access the project in the following GITHUB Repository:

https://github.com/yashj1301/Python-Projects/tree/master/Car%20Price%20Prediction

Contact details:

Yash Jain
9582617525
Greater Noida, UP-201306

yash.jain106@gmail.com
https://www.linkedin.com/in/yash-j-5537a0146/