Modeling automative Market – Predicting CAR Price

Predictive Modeling – OPIM 5604



**GROUP 6**

**Marcin Winczura**

**Sarvesh Bangad**

**Lauren Flecha-Rosado**

**Table of Contents**

**1. Problem Statement……………………………………………………………………………………………………………… 2**

**2. Data Set Description……………………………………………………………………………………………………………. 2**

**3. Sample………………………………………………………………………………………………………………………………… 3**

**4. Explore………………………………………………………………………………………………………………………………… 3**

**4.1. Missing Value Analysis…………………………………………………………………………………………. 3**

**4.2. Outlier Analysis……………………………………………………………………………………………………. 3**

**4.3. Multivariate Analysis…………………………………………………………………………………………… 4**

**5. Modify………………………………………………………………………………………………………………………………… 5**

**6. Model and Assess……………………………………………………………………………………………………………….. 5**

**7. Business Recommendation…………………………………………………………………………………………………. 6**

# Problem Statement

Our business goal is to identify the vehicle characteristics that predict MSRP (manufacturer's suggested retail price) using the following sixteen data elements, which are described in the data set section: Vehicle Make, Vehicle Model, Year, Engine Fuel Type, Engine HP, Engine Cylinders, Transmission Type, Driven\_Wheels, Number of Doors, Market Category, Vehicle Size, Vehicle Style, Highway MPG, City MPG, Popularity and MSRP. Using the SEMMA process, we plan to explore, modify, model and assess the data.

# Data Set Description

The dataset that we will be using “Car Features and MSRP” is obtained from the following location: <https://www.kaggle.com/CooperUnion/cardataset>.

The dataset is composed of 11,914 data points across 16 columns. The following data dictionary defines the variables presented within the “Car Dataset”:

|  |  |
| --- | --- |
| **Make** | Manufacture Name. 47 Manufacturers present in the data set. |
| **Model** | Model Name |
| **Year** | Model Year of the vehicle |
| **Engine Fuel Type** | Types: Electric, Diesel, Flex Fuel, Natural Gas, Premium Unleaded, Regular Unleaded |
| **Engine HP** | Engine Horsepower |
| **Engine Cylinders** | Number of engine cylinders |
| **Transmission Type** | Transmission on the vehicle. Types: Manual, Automatic, Direct Drive and Automated Manual. |
| **Driven\_Wheels** | Powertrain setup |
| **Number of Doors** | Number of doors on the car |
| **Market Category** | Manufacturer Description of the type of Vehicle |
| **Vehicle Size** | Overall physical size (3 categories): Compact, Midsize, and Large |
| **Vehicle Style** | Description of the vehicle architecture |
| **Highway MPG** | Miles Per Gallon expected during Highway conditions |
| **City MPG** | Miles Per Gallon expected during City conditions |
| **Popularity** | Popularity Ranking of the car |
| **MSRP** | Manufacturer's Suggested Retail Price |

# Sample

Our first step was to sample the data. We created 3 data sets: Training (50%), Validation (30%) and Test (20%) using Formula Random Function since Target Variable “MSRP” is continuous variable.

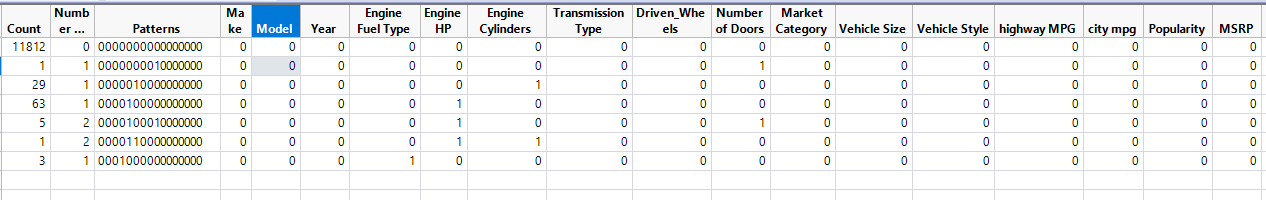
# Explore

After Sampling, we started exploring the data by understanding the range of variables within the dataset and what numbers we could realistically expect. We focused primarily on identifying missing values, outliers and correlations between variables.

## Missing Value Analysis

Total Records = 11,914

1. Missing Data Pattern



From the Missing Data Pattern table, below variables have missing values:

1. Number of Doors = 1 Record Missing
2. Engine Cylinders = 29 Records Missing

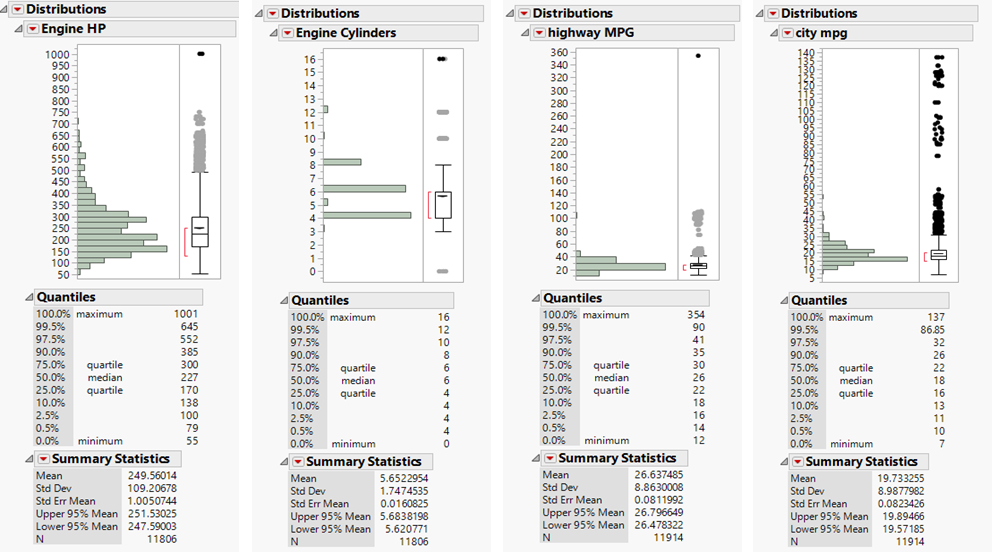
10 records of missing values for Engine Cylinders of electric car have been updated = 0 since electric car do not have engine cylinders.

1. Engine HP = 63 Records Missing
2. Engine HP and Number of Doors = 5 Records Missing
3. Engine HP and Engine Cylinders = 1 Record Missing
4. Engine Fuel Type = 3 Records Missing

Total Missing Records = 93, which account for only 0.78% of data set. Hence, we are ignoring these records for our analysis

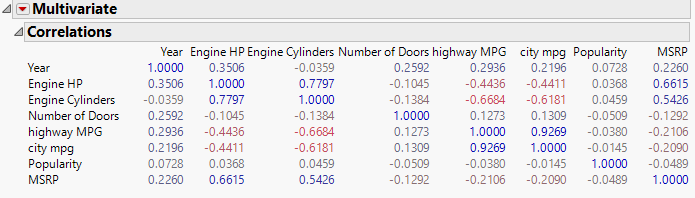
## Outlier Analysis

For finding outliers, we plotted Histogram and Boxplots for Engine HP, Engine Cylinders, Highway MPG and City MPG.



From the boxplots we see that at the outset there are outliers in each of the variables. However, when we look at individual variables in detail, and what they represent, we find that none of these are true outliers. We found that for “Engine HP” there are cars in the market (Bugatti Veyron) which have over 1,000 horsepower. For “Engine Cylinders” there are several brands such as Ferrari, Mercedes, and BMW which are available in configurations with more than 8 engine cylinders. We found that “Highway MPG” had data records above 100 MPG, which were not outliers as our dataset contained electric cars. We did however find a record with 354 MPG, which is not possible, even for electric car. This was manually corrected with a value of 34 MPG as there were other records of the same make and model with this gas mileage.

## Multivariate Analysis



From above Correlation Matrix, we can say:

1. ‘Engine HP’ and ‘Engine Cylinders’ are correlated to MSRP since correlation coefficient > 0.5
2. ‘Engine HP’ and ‘Engine Cylinders’ are also correlated to each other (0.7797).
3. ‘city mpg’ and ‘highway mpg’ are also highly correlated to each other (0.9269)
4. ‘Highway MPG’ being negatively correlated to ‘Engine HP’ and ‘Engine Cylinders’
5. ‘City MPG’ being negatively correlated to ‘Engine HP’ and ‘Engine Cylinders’.

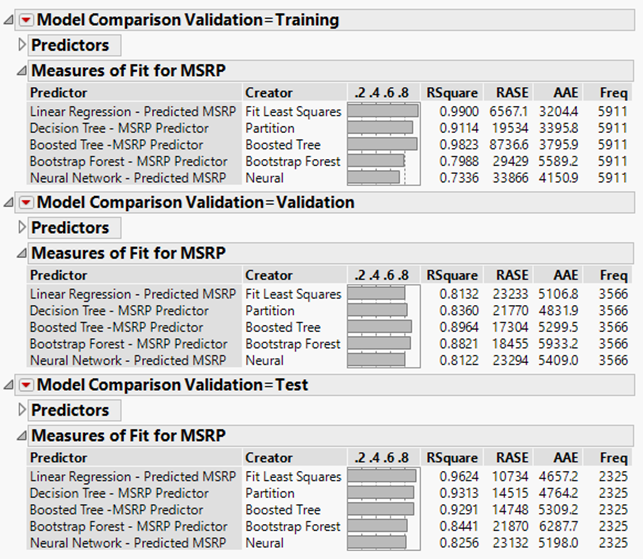
What we gained from these relationships was that larger engines (horsepower and cylinders) should drive a higher MSRP, but this would also imply that the lower the gas mileage the lower the MSRP, although this relationship is only weakly seen in the correlation table.

# Modify

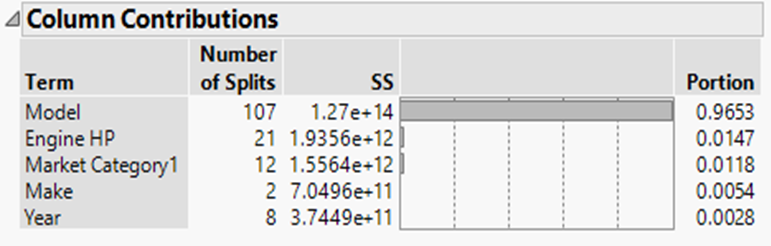
1. Market Category has 3742 records with “N/A” value. Hence, we created new column “Market Category1” in which “N/A” values have been replaced with new market category as “Unknown”.

# Model and Assess

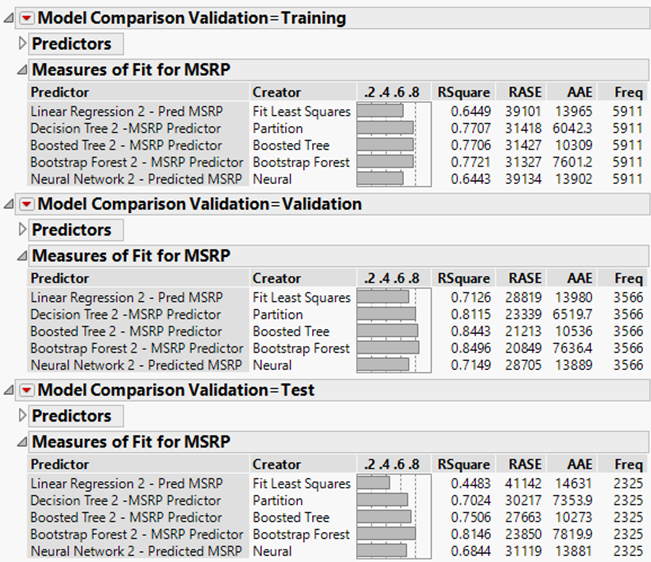
Our first approach was to use all the data elements (columns) to create the following types of models: Linear Regression, Decision Tree, Boosted Tree, Bootstrap Forest, and Neural Network. The final results can be seen as follows:



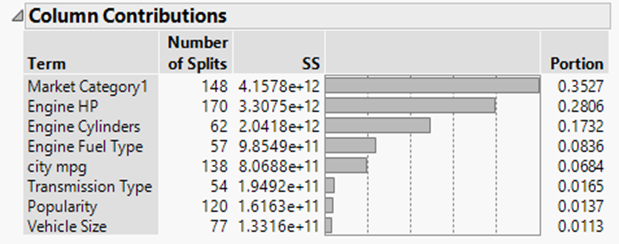
The data shows that the boosted tree model was the best at predicting MSRP as the RSquare within the validation dataset was 0.8964. Investigating the details of the Boosted Tree model showed that the “Model” of the vehicle was the single most important factor in predicting the MSRP of a vehicle.



This however, is a flawed result, as the “model” of the car (its designation) is only a name. There are various other features which define what a “model” is, not just the name. Granted, this information is interesting in that it suggests that it is possible to sell (or not to sell) a car just based on its model designation (i.e. sell a Rolls Royce Phantom simply because it is a “Phantom”).  
 We decided we would try a second approach, this time eliminating make, model, and year from our predictive models. Our business was as a new auto manufacturer so our cars would only belong to a single brand, the cars would all be brand new, and they could be named in any which way, therefore making these parameters non-essential to our goals. Re-running the various models with these certain predictors removed produced the following results:



This time the bootstrap forest model performed the best in the validation dataset with an RSquare of 0.8496.



We can say that the following features/characteristics of the car are the most important in predicting the price of the car: Market Category, Engine HP, Engine Cylinders, Engine Fuel Type, and City MPG. This result would explain a broader category of vehicles and be more useful as an automotive manufacturer to use in pricing our vehicles.

# Business Recommendations

When we saved the predicted formula from the Bootstrap forest model, we found that the formula had 6,860 terms in it. This made it extremely difficult to determine what characteristics would lead to the highest MSRP. So instead we decided to apply the prediction formula to our dataset and determine which characteristics came out on top. We found that the Lamborghini Aventador was the vehicle for which our model predicted the highest MSRP. This car, along with several others in the top MSRP segment were part of the “Exotic, High-Performance” category (or “Exotic, Luxury, Performance” as well), all of these cars had more than 600 horsepower, had 12 cylinders or more, all used premium unleaded fuel, and all had a City MPG of less than 11 miles per gallon. This result would make sense as exotic vehicles would demand the highest retail price. So our recommendation for our business, if we wanted to maximize our MSRP, would be to go after the “exotic” car market. If we were to ensure that our vehicle had characteristics similar to those mentioned before, and we marketed it correctly, then we would surely be able to demand a large premium.