SI618 PROJECT - PART II
Predictive Features for MSRP

1. Motivation

The MSRP, or manufacturer's suggested retail price, is quite simply the price that the manufacturer suggests that the dealer should ask for a car. The MSRP is a suggested price, dealers have the freedom to ask more or less than this figure. The vehicle's market price may beyond the suggested price if a car is in high demand, vice versa. A car's make often provides base price on their official website; however, we rarely can find a car with base price from a dealer. Levels of configuration, destination fees (ship from Germany to the U.S.) or warranty coverages are the reasons can vary the actual price of a car. Thus, a car's MSRP is somehow closer to the market price rather than the base price.

In this project, we aim to discover the relationship between MSRP and some car's features. Also, we will fit a model to predict the MSRP of a car by regrading MSRP as the response variable, other features as predictors.

To be specific, we are interested to explore following questions:

1. Exploratory for price (MSRP) levels:
   - Distribution of MSRP by histogram
   - Distribution of MSRP respectively group by make, fuel type, cylinders, driven wheels, vehicle size, to explore the difference between options in each variable?
   - From MSRP levels, find same-level competitor
2. What features most predict MSRP?
   - Clustering, PCA, ANOVA, Random Forest Model
3. What models are the most over-priced for their feature set?
   What makes are the most over-priced?
   - Compute price difference between actual MSRP and predict MSRP
   - Top 10 over-priced models plot
   - Top 10 over-priced makes plot

2. Data Source

Kaggle's dataset: *Car Features and MSRP*, scraped from Edmunds and Twitter.
https://www.kaggle.com/CooperUnion/cardataset

The dataset includes 11914 observations and 16 variables total:

Make	Categorical (48 levels)	Number.of.Doors	Numerical
Model	Categorical (915 levels)	Market.Category	Categorical (Drop)
Year	Numerical (1990-2017)	Vehicle.Size	Categorical (3 levels)
Engine.Fuel.Type	Categorical (11 levels)	Vehicle.Style	Categorical (Drop)
Engine.HP	Numerical	highway.MPG	Numerical
Engine.Cylinders	Numerical	city.mpg	Numerical
Transmission.Type	Categorical (5 levels)	Popularity	Numerical
Driven_Wheels	Categorical (4 levels)	MSRP	Numerical (response)

The data is csv format can be downloaded from Kaggle directly. We will use only 14 variables (highlighted) in our analysis. The data includes records from year 1990 to 2017.

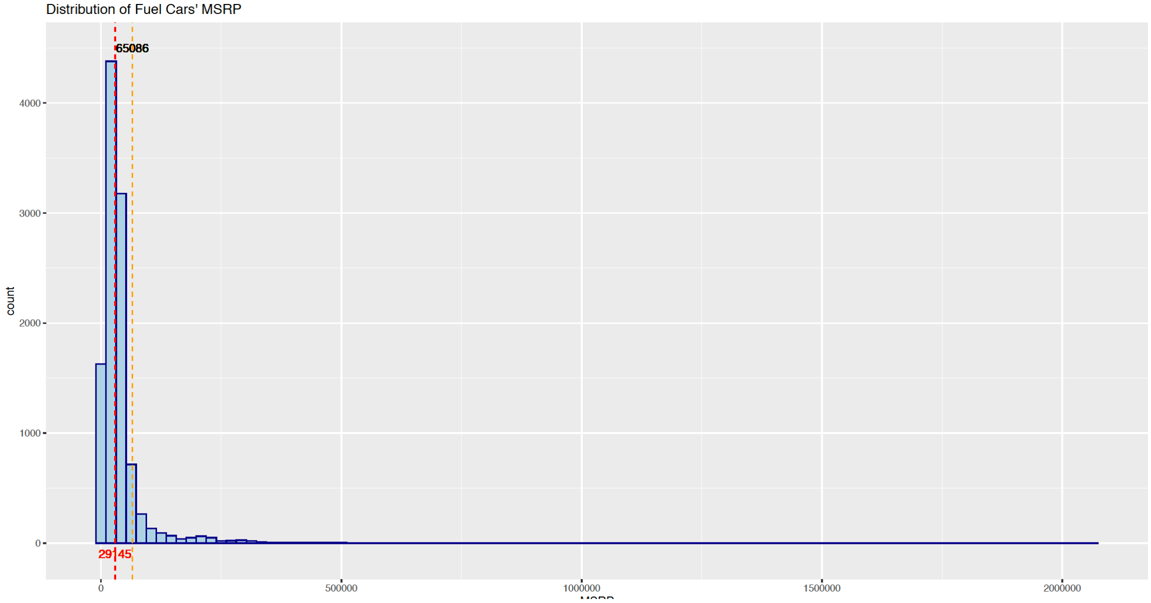
3. Methods and Analysis

Question 1: Exploratory for price (MSRP) levels

First, we need to do data exploration for our original data. Using head() summary(), and str(), we find that variable “market category” and “vehicle style” have either lots of NA values or too much levels (types). Thus, we decide to drop these two variables.

As known to us, fuel type is a pretty vital component for price. The price for different fuel types can be influenced by export import tax, local environment policy or different technical costs. We definitely want to keep fuel type as a predict variable in our model, but we find there are 11 fuel types in our dataset, such as three types of gasoline (required premium, recommended premium, regular), diesel, electric, natural gas or some hybrid types. Except gasoline and diesel, the rest fuel types have small sample size, which may cause bias in our estimation. Although electric has relatively large sample size with more than one hundred records, almost of them are Tesla. The predict price will be dominated by Tesla price instead of the other makes’ price. What’s more, for a non-petrol car, such as electric car or natural gas car, they may not have cylinders, regular transmissions or mpg information, which can also affect our prediction for MSRP. Consider above reasons, we would like to keep petrol cars (3 gasoline types and diesel) and remove NA values. 10818 records are still remained in our fuel.cars data.

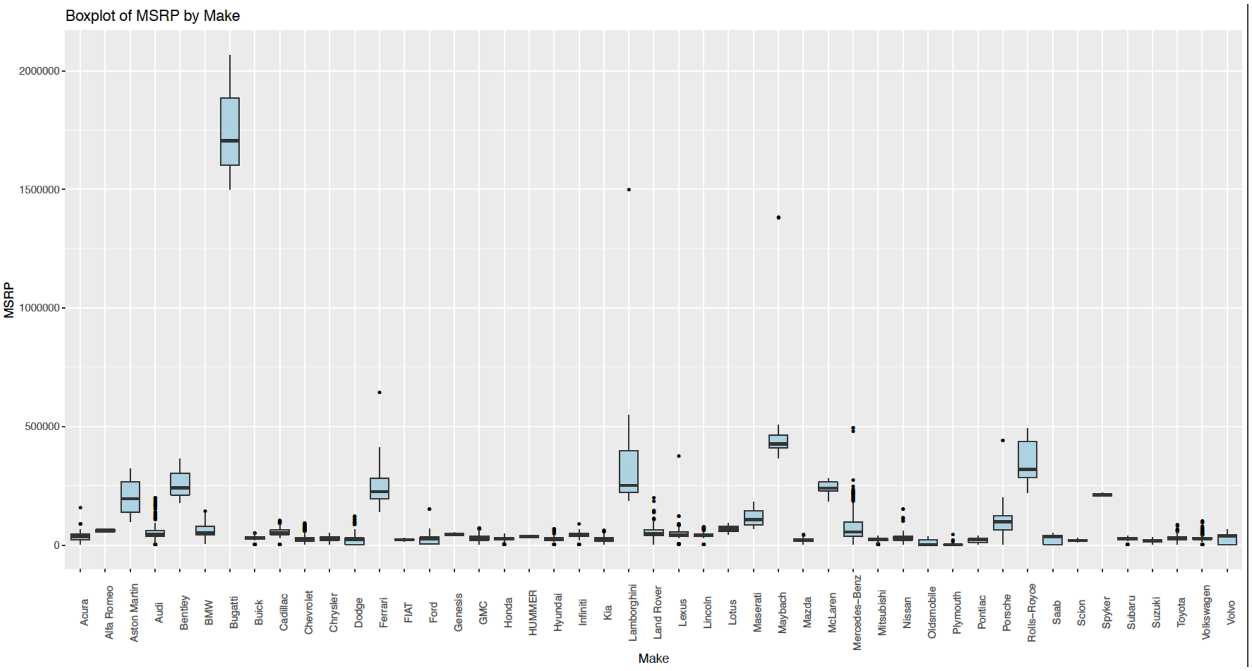
In order to observe the distribution of MSRP, we use “ggplot” to draw the histogram plot of MSRP.



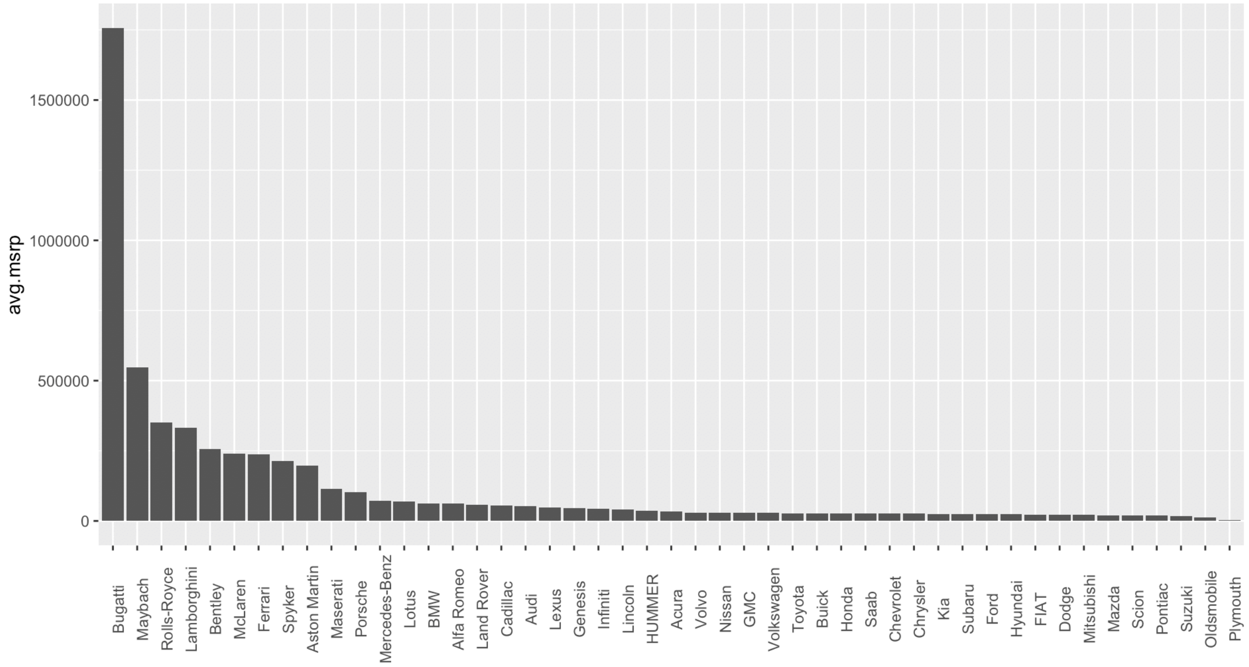
The red vertical line shows that the median MSRP. The orange line indicates that 90 % of MSRP are in range 0 to $65,000. Also, we can see some extreme high MSRP up to more than $2,000,000.

Next, we want to find the distribution of MSRP respectively group by make, transmission types, fuel type, cylinders, driven wheels, vehicle size, to explore the difference between options in each variable by using boxplot.

Makes Plots:



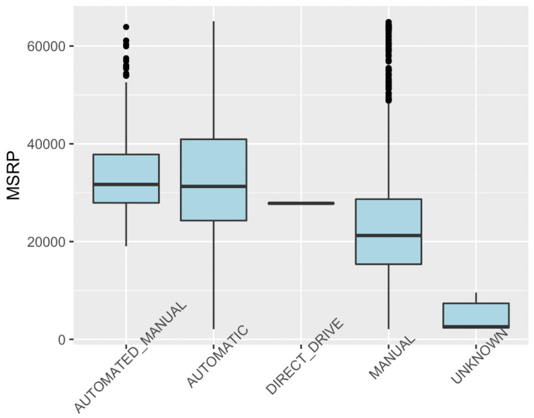
From boxplot of MSRP by make, we can see some makes have higher MSRP than the others. In order to have a clearer visualization, we group each make to get average MSRPs, then we draw a bar plot with sorted average MSRPs.

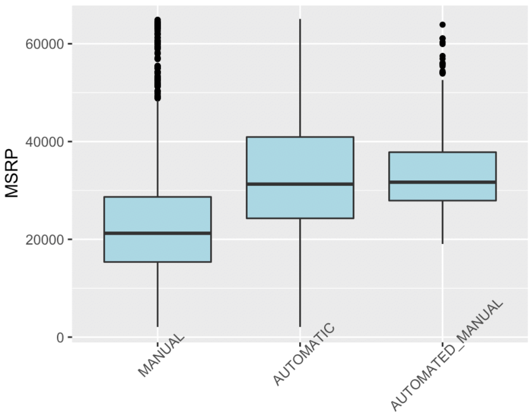


From above plot, we find makes have different levels. Bugatti is the most expensive make in our data without doubt because of its extreme high MSRP values. We can consider Maybach also has a high MSRP far beyond the rest makes. Then, Rolls-Royce and Lamborghini at similar MSRP level, so we may consider them as same-level competitor. Next, Bentley, McLaren, Ferrari, Spyker and Aston Martin are in same level. Maserati and Porsche have similar MSRP. The next groups have many common luxury brands: Mercedes-Benz, Lotus, BMW, Alfa Romeo, Land Rover, Cadillac and Audi.

From the plot, we find that although the market always consider Mercedes-Benz, BMW and Audi are same level competitor, but from an overall view, Mercedes-Benz has an obvious higher MSRP than Audi, we may think Audi don't have a well-developed high-end market or Audi have a better market share in low-end market. For cheaper makes, we can see some high cost-effective brands, Japanese makes: Honda, Toyota, Nissan, Mazda, and American brands: Chevrolet, Ford, Dodge.

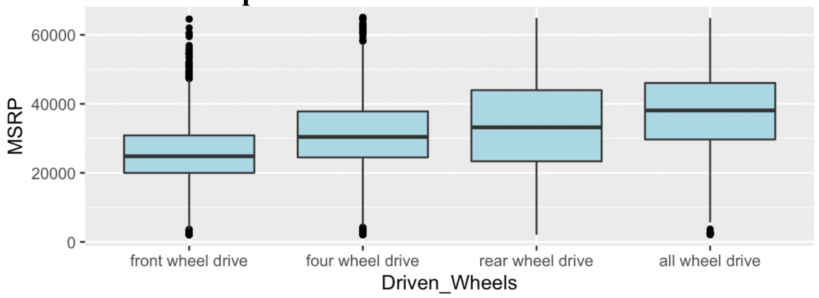
Transmission Type Boxplots:





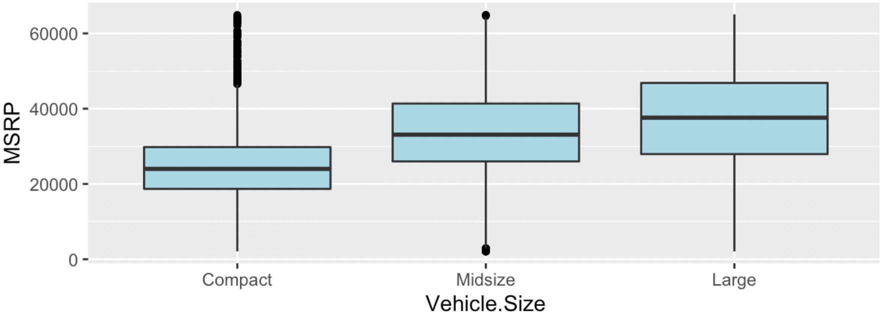
For left transmission type boxplots, we see 5 types of transmission. “Automated-manual” has little higher average than “automatic”, but “automatic” has a larger range. “Manual” has a lower MSRP. “Direct-drive” has a very small sample size. Since “direct-drive” is usually common in electric car, we want to know whether these records are wrong data. We search all “Direct-drive” vehicle from our fuel.car data, then, we get two records with same model called Malibu. However, when we search all Malibu models, the most Malibu have “automatic” transmission. We guess that Malibu with “direct-drive” should be a hybrid version, which is not belongs to our desire types. Thus, we drop “direct-drive” and “unknown” to get the right boxplots. All in all, we believe transmission types have influence on MSRP but not significant since MSRP distribution of “automatic” and “automated-manual” are very similar.

Driven Wheels Boxplot:



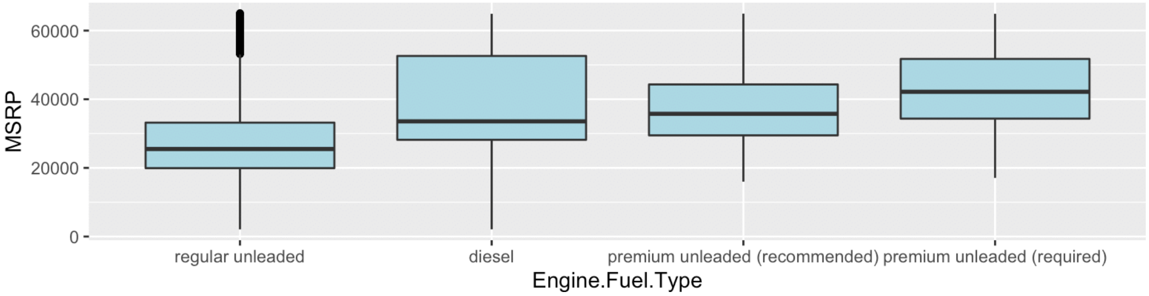
Same principle, we find the MSRP levels for each driven wheel type have following order: all-wheel drive > rear-wheel drive > four -wheel drive > front-wheel drive. So, we may consider driven wheels can help us distinctly estimate the MSRP.

Vehicle Size Boxplot:



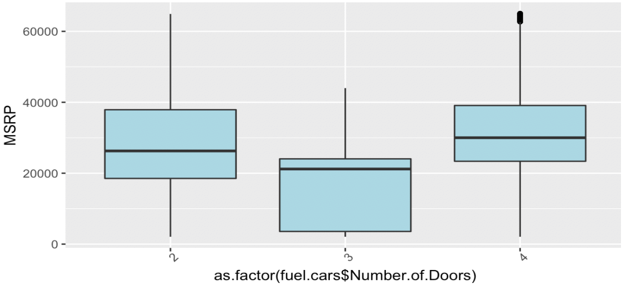
Same principle, we find the MSRP levels for each size have following order: large > midsize > compact. So, vehicle size may be a good predict variable.

Engine Fuel Type Boxplot:



For engine fuel type, we can find the pattern that regular unleaded has the lowest MSRP leve, MSRP level of premium (required) is clearly higher than premium (recommended). Diesel is hard to say, the mean MSRP of diesel car is lower than two premium types, but the 3rd quantile is similar or even higher. However, since most observation in our data are gasoline types, we may still want to use engine fuel type in our estimation.

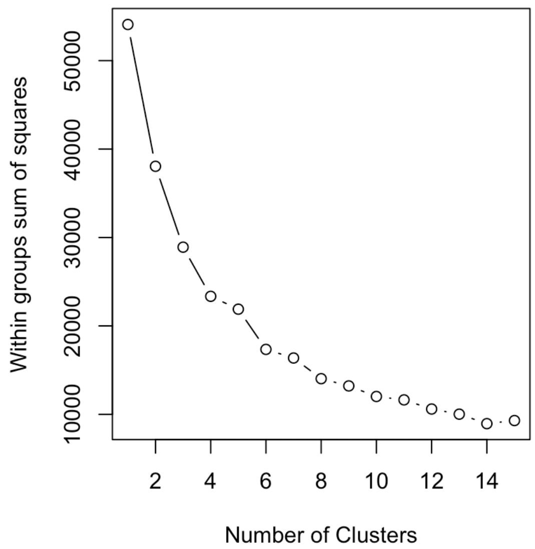
Number of Doors Boxplot:

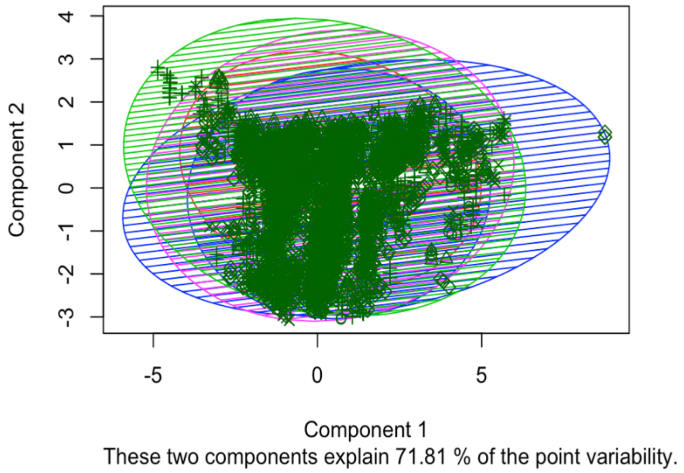


Although the number of doors is a numerical variable, it only has 3 values. Thus, we can treat it as a categorical variable. From the number of doors boxplot, we find "2-doors" and "4-doors" are similar, the "3-doors" has lowest MSRP level.

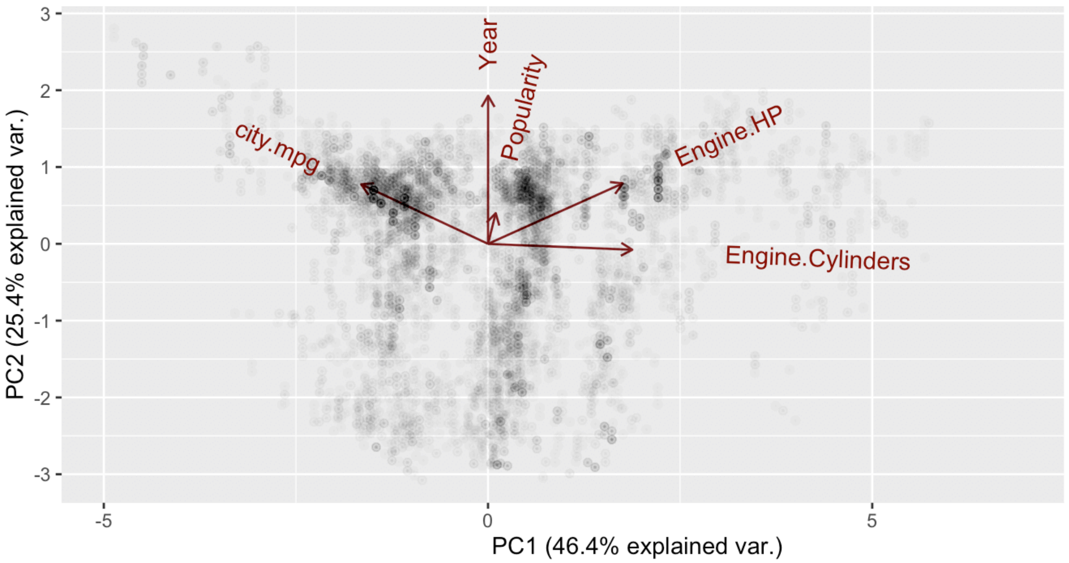
Question 2: Most predictive Features for MSRP:

For this question, we would like to apply the clustering only on numerical variables first. Because "city.mpg" and "highway.MPG" are similar things, we will keep one to represent cars' MPG. Here, we choose "Year", "Engine.HP", "Engine.Cylinders", "city.mpg", and "popularity" to do clustering. After removed the NA values, we center and scale our data with scale function. Though computing the within groups sum of squares, we figure out the appropriate cluster numbers is 5 from left plot.





Then, we use clusplot function to get plot on the right. As we can see, 5 clusters are overlap with each other's in large ratio, which indicates strong correlation between these variables. One idea to reduce the variables' correlation is using PCA to do dimension reduction. However, the **challenge** we meet here is that PCA usually apply on numerical data, while our data contains more than half categorical variables. It's hard to compute the categorical data's distance. We can only use PCA apply on our numerical data first.



Though biplot (ggbiplot), we find that "Engine.Cylinders" and "Engine.HP" have same direction on PC1 and shows strong correlation relationship. Also, both of them have opposite direction with "city.mpg". "Year" and "Popularity" have very low or even null variance on PC1, but with same direction on PC2. However, "Popularity" can explain very less variance on PC1 and PC2. Thus, we may drop this variable in further estimation.

Since the categorical variables are not included in PCA, we cannot use PC here to fit the model. We will select variables from our previous exploratory results.

Training data and Validation data

We randomly divide our data into two parts: train set (70%) and validation set (30%). We will use train set for further model fitting, and we use validation set to predict the "MSRP" and check models' accuracy.

Model fitting

First, we try to apply lm() to do the linear regression with "MSRP" as response variable, the rest variables as predict variables. Using ANOVA table to check which variables are significant. The "Engine.Cylinders" , "Number.of.Doors", "Vehicle.Size" is not very significant as the others, so we can drop these variables and fit a new linear model. After getting the predicted result, we use it to calculate mean absolute error [mean_error01 = sum(abs(testy-pred01))/length(testy)] and mean percent error [mean_error01/mean(testy)]. These two value point out all model accuracy. The mean absolute error is 10221.78, and mean percent error is 25%. The error rate is pretty high; thus, we will abandon this model.

Next, we decide to use random forest, which is a popular technique is used to improve the predictive performance of Decision Trees by reducing the variance in the Trees by averaging them. We apply to random forest model with different selected variables.

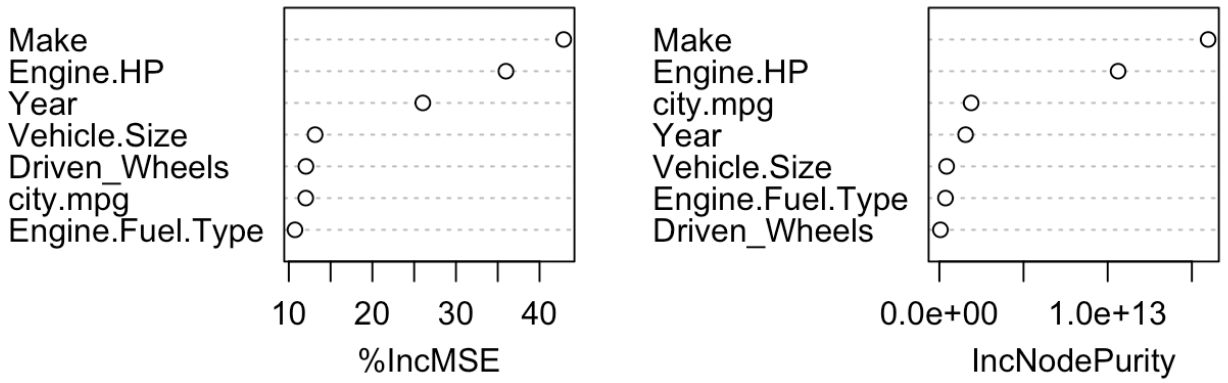
First random forest model with same variables selected from ANOVA:

randomForest(MSRP ~ Make + Year + Engine.Fuel.Type + Engine.HP + Transmission.Type + Driven_Wheels + city.mpg, data = TrainSet, ntree = 500, mtry = 3, importance = TRUE)
We get 4119.629 as Mean absolute error, 10.15% as mean percent error.

Second random forest model with selected variables from our previous exploration results:

randomForest(MSRP ~ Make + Year + Engine.Fuel.Type + Engine.HP + Vehicle.Size + Driven_Wheels + city.mpg, data = TrainSet, ntree = 500, mtry = 5, importance = TRUE)
We get 3876.729as Mean absolute error, 9.5% as mean percent error.

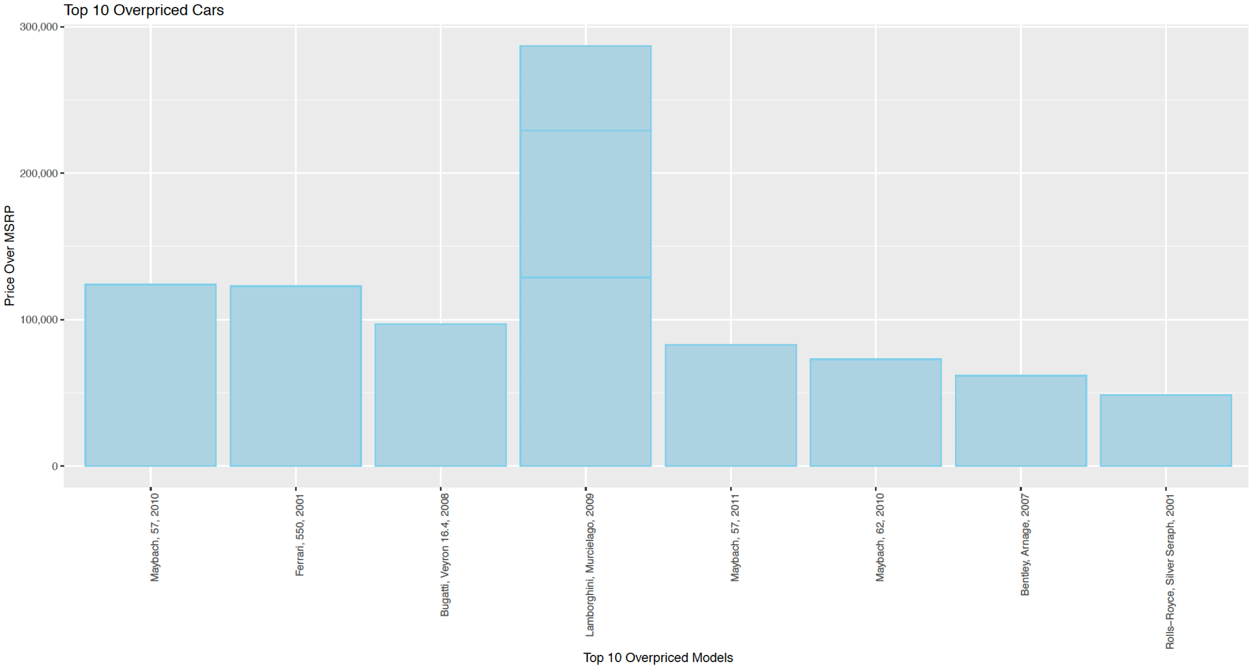
Thus, we choose the second model with lower error. Here, we compute the importance of variables (Mean Decrease in Accuracy) in model 2.

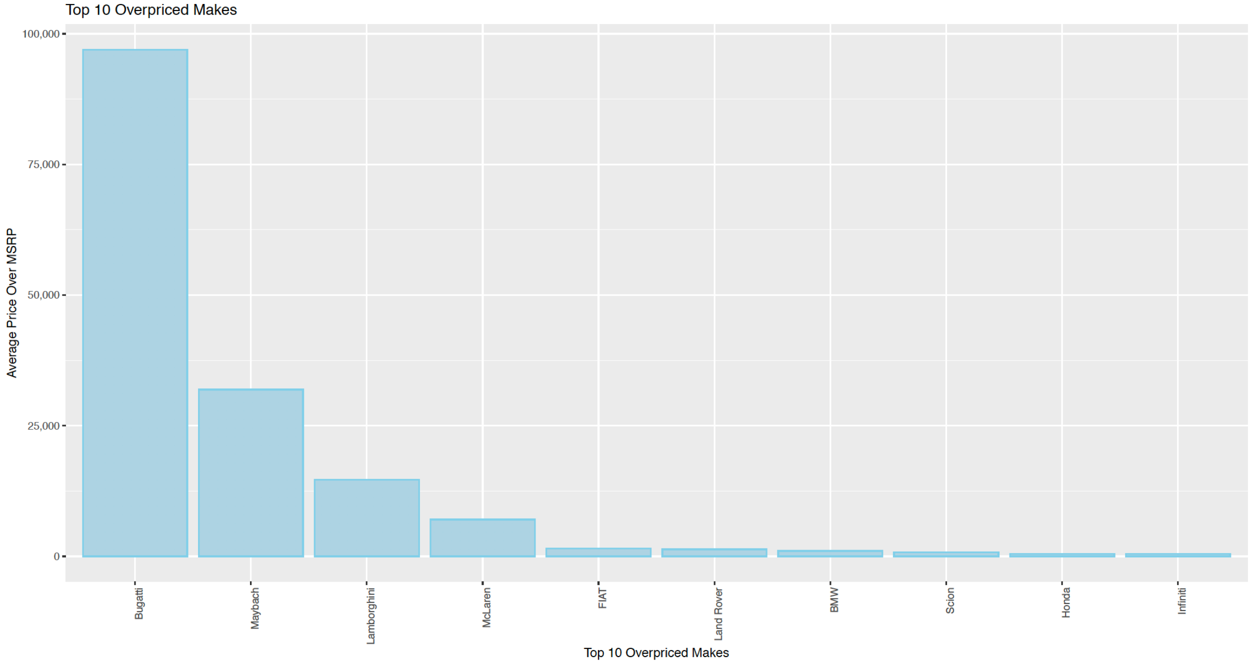


From the importance plots, we can clearly say that "Make" and "Engine.HP", "Year", and "city.mpg" are most important predictors, in other words, they are the most predictive features for MSRP.

Question 3: Top 10 Over-priced Model and Top 10 Over-priced Makes:

How to determine a vehicle is overpriced? Here, we use error as price difference, which is a car's actual price minus predict price. From our fitted model and validation data, we can generate car's predicted MSRP. Then, use it to minus the actual MSRP to get the price difference. If the price difference is positive, we consider the car is overpriced. We have to sort the price difference with decreasing order, then merge it with the original cars data by their index number. Further, we can find the top 10 overpriced model. By grouping makes, we can find the median value of price difference for each make.





The top 10 overpriced models are from following makes: Maybach, Ferrari, Bugatti, Lamboghini, Bentley and Rolls-Royce. The result is reasonable since super luxury brands have their potential brand effects and collect values. The same in top 10 overpriced makes, the top 4 makes, Land Rover and BMW are luxury brands, rest are some popular brands. We can't guarantee that all Honda models are highly cost-effective; moreover, the average overprice value is pretty low which may vary with change of validation data.

Some challenges we meet here:

First, when we fit the model, we know that factor "Make" will largely influent the MSRP which is called brand effect. For a perfect prediction of cars' real value, we need more specific data, such as car's dimension measurements, engine and transmission performance parameters, brake quality, technologies, exterior and interior material, repair rate, design standard, number of productions, etc. Because we don't have these detailed data, "Make" becomes the alternative features to predict the MSRP. However, we have to understand that using "Make" and current variables may cause some bias or error if a brand has wide range levels of cars, such as Mercedes-Benz has price range from 32K to 2.8 million. However, using "Make" can still produce meaningful prediction since we can consider brand effect as a predictor which is a real estimate standard in industry.

Second, since training and validation data are randomly sampled, some extreme expensive cars may only show in the training data not testing data. In this case, we may produce different results, especially for the top 10 overpriced models. It is inevitable since we cannot determine the observations in training and validation set.

It is noteworthy to mention that we use median value of price difference for each make instead of median. If a brand has lots of low-end models, which has low probability to be overpriced, can lower its average overprice level. For example, if a make only have one or two overpriced car, but the rest models have high cost-effective, using mean will increase its average overprice level. Thus, we use median here.