**Car Price Prediction**

**Problem Statement**

A Chinese automobile company Geely 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 do those variables describe the price of a car?

Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

**Business Goal**

You are 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.

1. Understand the Existing System and Analyse

Geely Auto is looking into entering the US market by setting up their manufacturing unit and producing cars locally to compete with their US and European counterparts. As a result of this, Geely Auto wants to understand the factors driving the prices of cars in the American market since these factors might differ from the Chinese market. Geely Auto will also like to find out how well these variables describe the price of cars. Currently, the consulting company has gathered a large dataset of different types of cars in the American market to analyse the different variables and its impact on price.

1. Proposed System and its Components

The gathering of large dataset can be haphazard especially if we are looking to improve the model in the future which can be done with more data. Without a proper infrastructure, the collection of data and analysis will be messy, allowing for mistakes to be made in between. As such, it is important to collect the data and store it in a database, making it easier for the consulting firm to query data where necessary. This allows for the storing of data to be made in a more organised and proper manner. Given the plethora of options for database systems, the consulting firm can tap on an open-source software, MongoDB to host the data and build on it as they gather more data. Additionally, where necessary, MongoDB can also be used in conjunction with R which allows the firm to clean, process and improve the machine learning models. As such, this project works under the assumption that the machine learning models created here are preliminary in nature and will improve with more data that will be hosted on a database management system.

To understand the American market better, the management will have to understand the factors driving the prices of cars. This will help management to alter accordingly the strategies required to cater to their American customers. To do this, I have run a regression model to look at the role of 24 independent variables on the dependent (target) variable, price.

In order for the management to cater effectively and efficiently to the needs of the American market, 3 different regression models were done namely, multiple linear, random forest and decision tree regression. The basis of comparison for these models is the root mean square error (RMSE) which broadly refers to the mean differences between the predicted Y value (price) and the actual Y value. RMSE is used to measure the prediction errors and as such, a metric that will consistently be used for all regression models.

Additionally, an interesting result emerges when a correlation test was done between the independent variables. This step is necessary to find highly correlated values – these values will then be removed from before testing the model to see any significant changes to the RMSE value. All 3 models were done twice – one without the highly correlated value, curb weight and another with the curb weight. Generally, the both model provides interesting results with regard to the RMSE values. Subsequently, I will refer model 1 to the model with the highly correlated value and model 2 without the highly correlated value – curb weight.

Regardless of model, it is rather clear that random forest regression provides the best RMSE value of $1664 (model 2) and $2116 (model 1). Despite this, different models provide different interpretations of independent variables that are significant and between model 1 and 2, the RMSE score differs as well if we use the cross validation (CV) result or the test result.

Firstly, I will discuss the comparison for the first method – linear regression for both models.

Table

Description automatically generated

Model 1 linear regression results – without correlated values (curb weight) removed.

Table

Description automatically generated

Model 2 with curb weight removed.

*Evaluation of variables*

Model 1 has Adjusted R square of 0.897. The adjusted R square signals to us how well the model will perform if it were to be replicated on the test set. The estimate signals to management that for every one increase in price, there is a fall or increase in x much. It signals the relationship between the predicted and independent variables.

The values marked with \*, \*\* and \*\*\* are the most significant with p-value of close to 0. From the above two screenshots, it is clear that most of the significant values in model 1 are also in model 2. However, model 2 was able to detect two additional values that were deemed significant – car width and car height. When the linear regression model was used on the test set, model 2 performed much better than model 1.

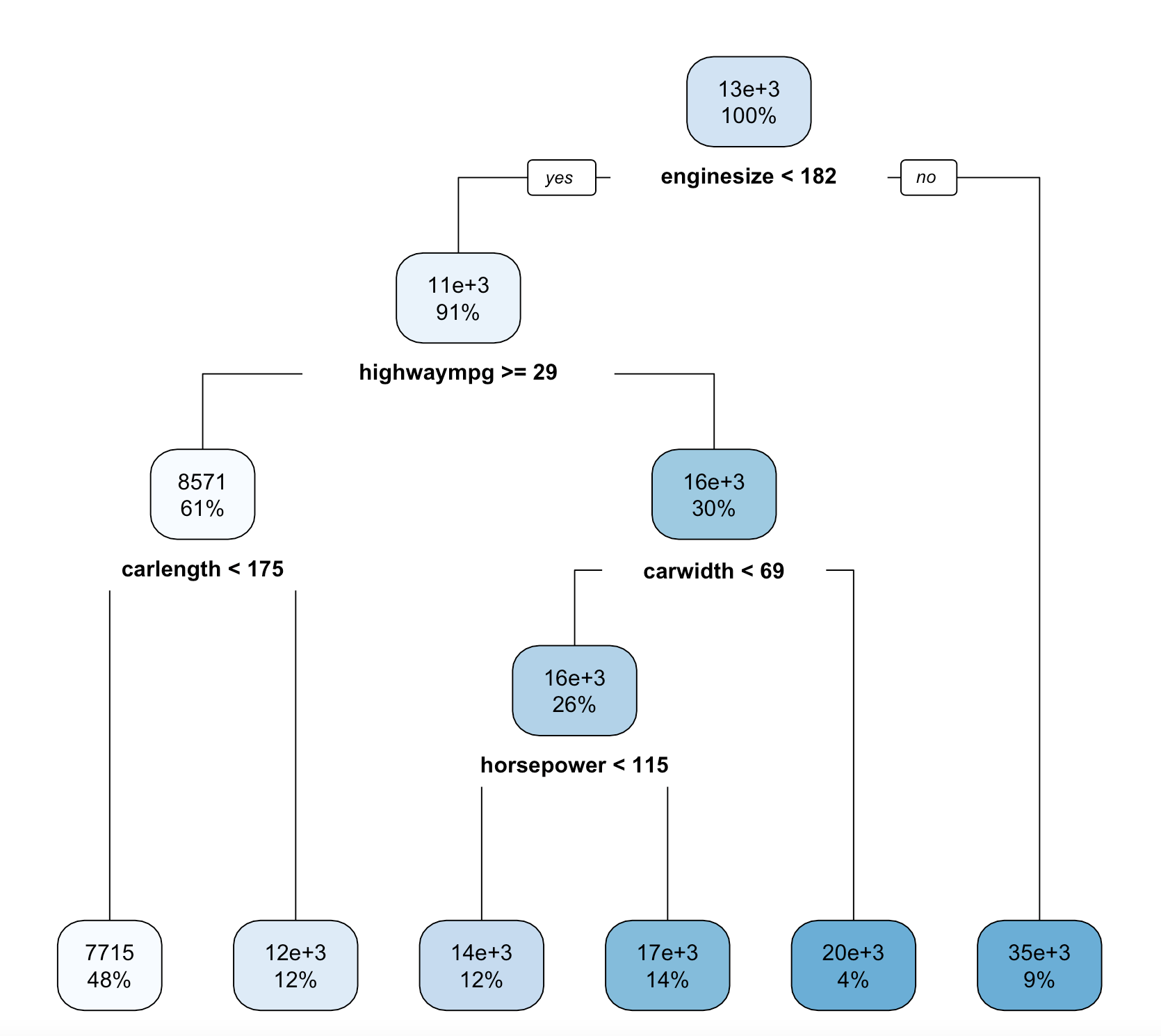
The RMSE score are $3494.49 and $2107.5 respectively. Naturally, it is safer to assume that Model 2, by virtue of lower RMSE score is better than Model 1. When comparing the train set and test set, the training set seems to perform much better than it’s test set. The RMSE score of the train set after cross validation is much lower - $2378.853. This high RMSE score is dangerous because it signals to us and management that we could have been overfitting the model.

On the contrary, model 2’s RMSE score is lower for the train set than it is for test set ($2516.566). Given this, model 2, with curb weight removed, seems better since it is able to account for much lesser prediction errors than model 1 when comparing the test set.

Within the training set, k-Fold cross validation was used and when we compare the RMSE scores of two models, the RMSE score was higher in model 2 than it is with model 1. However, given the smaller margin of discrepancy, comparing within model RMSE score is more valuable. As such, model 2’s performance is better.

The second machine learning method that was used was the decision tree. This machine learning method was used in both models and model’s performance was compared using the RMSE score.

The RMSE score for model 1’s test set was at $3024.522. Using cross-validation on our train set, model 1’s RMSE score is much lower at $2441.838. This is similar to the linear regression model that was done on model 1 – lower score for train than test set. When we compare this with model 2, the RMSE score for the test set, it is much lower - $2559.699. The trend is similar to the train set after cross-validation at $2422.93. Even though we see that the model performs better in the training set than testing set, it is clear that model’s 2 performance is better given the smaller differences between training and test set. Additionally, even when we compare the score of the cross-validation train-set, model 2’s RMSE score is lower. How about the variables that are important? This is depicted in the following screenshots:



1) root 184 11977760000 13154.780

2) enginesize< 182 168 3525974000 11114.370

4) highwaympg>=28.5 112 714149700 8570.705

8) carlength< 174.8 89 189363900 7714.719 \*

9) carlength>=174.8 23 207234600 11883.000 \*

5) highwaympg< 28.5 56 637827000 16201.700

10) carwidth< 68.6 48 429134400 15506.670

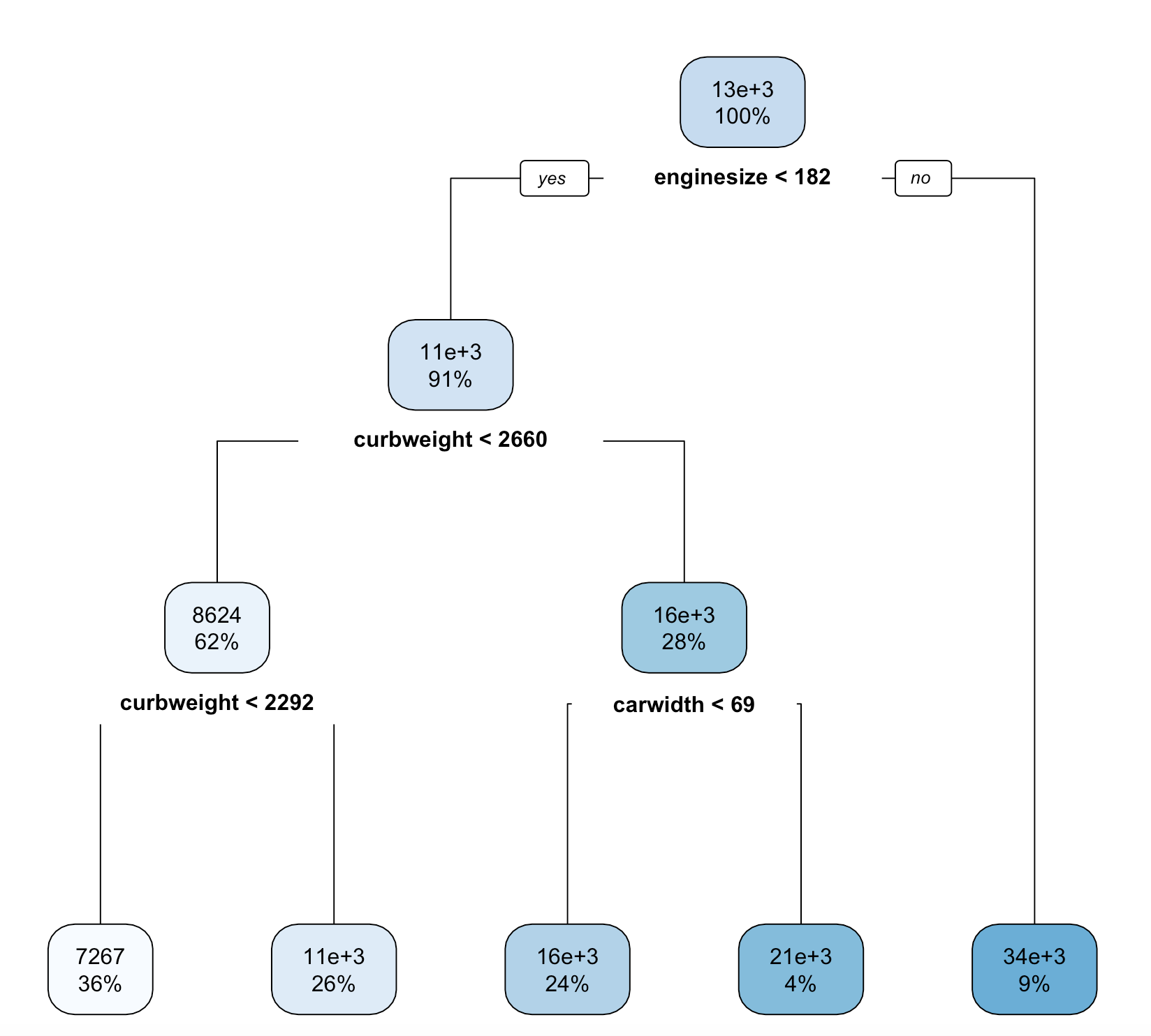
20) horsepower< 114.5 22 87468080 13709.680 \*

21) horsepower>=114.5 26 210512400 17027.200 \*

11) carwidth>=68.6 8 46382600 20371.880 \*

3) enginesize>=182 16 408358800 34579.060 \*

*Information on model 2*



1) root 184 12237450000 13209.280

2) enginesize< 182 167 3354013000 11044.020

4) curbweight< 2659.5 115 589850400 8624.348

8) curbweight< 2291.5 67 75951480 7266.985 \*

9) curbweight>=2291.5 48 218149900 10519.000 \*

5) curbweight>=2659.5 52 601823700 16395.210

10) carwidth< 68.65 45 417051700 15747.140 \*

11) carwidth>=68.65 7 44370490 20561.430 \*

3) enginesize>=182 17 409133700 34479.740 \*

*Information on model 1*

From the above 2 screenshots, it is clear that engine size is the most significant factor. This is seen in how engine size is the root node (primary split) for both models. Similarly, this is also the same for car width. However, interestingly, car length and horse power became the next node of significance for the second node.

For model 1 and 2, the model splits it primarily on engine size – whether it is less than 182 or not. In model 2, the second node further splits the model into highway mpg – whether it is above or equal to 29. If it is, the model splits it further with the car length. We see that 91% of cases have engine size of more than 182. This is the similar to model 2 where 91% of cases have engine size of >182. For model 2, we see that those cars that have engine size of <182, highwaympg >=29 and car length of <175, have prices of $7715. This accounts for 36% of the observation. In model 1, for cars that have engine size of <182, curb weight of <2660 and <2292 average around $7267 in price. This accounts for 36% of the set. In both models, the \* indicate the nodes.

Let’s take a look at the performance of the last model, random forest before concluding.

Just like the linear regression and decision tree, cross validation was done on the train set and was compared to the test set capped at 10%. The RMSE of model 1’s test set is $2116. Conversely, the RMSE of model’s 2 test set is $1664.088. Additionally, model 1’s RMSE score for the cross-validated test set is $1012.916 and model 2’s is $1016.816. As seen in this case, both models have low RMSE score for the cv train-set but perform rather poorly in the test-set. However, model 2’s RMSE for the test set is lower than model 1, allowing us to conclude that model 2 has better performance than model 1.

For random forest, both models have specified n-trees of 500. Each tree will then calculate the average of the Y score (predicted value) after the model has split according to the data points. After n-trees have calculated/predicted the Y value, the average will be calculated. As such, for a new predicted value, it will assign the average in relation to the data point in question.

In the case of random forest, these data points are picked at random will not use all the data points in the model. In the case of both models, the default minsplit (the minimum number of data points required to attempt to split before it is forced to create a terminal node) was used, which is 20. The maxdepth (maximum number of internal nodes between the root node and the terminal nodes/leaf) for these models are also unspecified, allowing R to use the default which is 30, thus, allowing for large trees to be built.

To investigate which variables are important, executing the importance() function from the caret package can explain for this.

%IncMSE

symboling 2.796045

CarName 8.938828

fueltype 3.227989

aspiration 4.967632

doornumber 1.398699

carbody 2.176773

drivewheel 4.538980

enginelocation 4.214834

wheelbase 9.907694

carlength 9.287599

carwidth 9.898576

carheight 4.536266

curbweight 16.807140

enginetype 3.191833

cylindernumber 2.472485

enginesize 21.057694

fuelsystem 6.735197

boreratio 4.461730

stroke 3.650510

compressionratio 6.999581

horsepower 13.405691

peakrpm 7.598103

citympg 8.799287

highwaympg 9.488635

*Results from model 1*

%IncMSE

symboling 1.9350413

CarName 6.0881289

fueltype 3.1436846

aspiration 5.2291790

doornumber 1.1012313

carbody 0.4849986

drivewheel 5.7762647

enginelocation 4.1425277

wheelbase 12.3094553

carlength 12.5341172

carwidth 13.2103393

carheight 5.8226340

enginetype 2.7041423

cylindernumber 3.7405576

enginesize 22.3408830

fuelsystem 6.8892811

boreratio 6.9283393

stroke 6.4263289

compressionratio 6.3740462

horsepower 14.4968220

peakrpm 7.5762821

citympg 11.3048957

highwaympg 11.0890336

*Results from model 2.*

By calling the importance function, and specifying type 1, the model will show the % of MSE which is how much model accuracy decreases if the variable is taken out. Figures highlighted in yellow indicate variables that are significant and thus have a bigger impact if it is left out (variables with threshold of >10% are highlighted). By comparing the decision tree and random forest of model 2, it’s obvious that some variables that are important converge. The variables that are important are engine size, horsepower, car width, car length, wheelbase, city mpg and highway mpg. Some of these variables are similar to that of the decision tree in model 2: engine size, highway mpg, car width, car length and horsepower.

However in model 1, there are very similar variables have values of close to 10% such as care length, car width and highway mpg. The decision tree in model 1, interestingly is able to capture three variables: engine size, curb weight and car width. Additionally, model 1’s decision tree was not able to capture another significant variable – horsepower which was present in random forest.

Conclusion

For the management to understand which variables are significant in predicting the price of cars, this largely depends on the model of which we are using. To conclude, since this report uses RMSE as a metric to measure model’s performance, random forest seems to have the best performance with the lowest RMSE score. This is regardless of whether the model uses the RMSE of the CV train set or the test set. Consistent with the three models, variable engine size is the most significant variable in predicting the price of car. Using the random forest regression, other variables that are important include horsepower and car width, to name a few. Removing these variables will result in a 10-20% decrease in model accuracy.