MGSC401 Statistical Foundations of Data Analytics
Final Project

Understanding Risk Factor in Automobile Insurance

Using Machine Learning Techniques

Overview

Insurance companies are constantly looking for ways to accurately assess the risk of insuring different vehicles. This is important, as it allows them to set appropriate premiums and ensure that they can cover potential losses. According to IBC[1], car insurance premiums are affected by not only one's driving record and neighbourhood, but also the characteristics of the vehicle to be insured. This report presents a statistical analysis that uses machine learning to predict the risk rating of automobiles based on their characteristics. More specifically, it will investigate the affect of a car's specifications on the assigned insurance risk rating (symbol) of a vehicle which is a significant factor when calculating insurance premiums. This analysis will not only assist insurance companies but also help consumers become more aware about the key drivers of their insurance rates.

Data Description

The automobile dataset included information about 205 cars from 1985 Ward's Automotive Yearbook. Each car was defined on 26 physical characteristics including the decision variable *symboling*. *symboling* is a rating that stands for the degree to which the vehicle is riskier for the insurance company than what its price indicates. In practice, the rating ranges from -3 (very safe) to +3 (very risky) however in the dataset the minimum value is -2. The adjacent chart shows the frequency distribution of each symbol in the dataset, with 0 having the highest occurrences. Some of the class variables were *make* (brand of the car), *num_of_doors* (number of doors in the vehicle), *body_style* (sedan, hatchback, convertible, etc.), and *fuel_type* (gas or diesel).

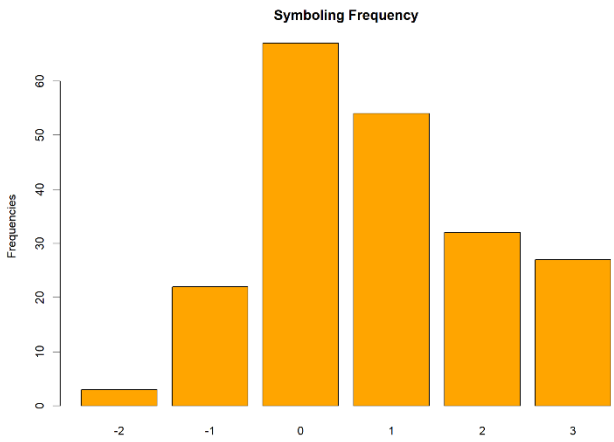


Figure 1 - Symboling Frequency Plot

Before running any exploratory analysis on the dataset, it was essential to ensure that the dataset was clean and free of errors. To begin with, the class variables were converted as factor and others were imported as numeric. Next, the summary command was run on the full dataset (see appendix 1) to get a

[1] How car insurance premiums are calculated. Insurance Bureau of Canada. (n.d.). Retrieved December 14, 2022, from http://www.ibc.ca/sk/auto/buying-auto-insurance/how-auto-insurance-premiums

birds-eye view of all the variables which instantly helped identify missing values and errors in the dataset. There were 41 missing values for the *normalized_losses* predictor. Since the dataset was missing 20% of the values for this predictor it was best to drop the column to retain the size of the dataset. Other columns like *bore*, *stroke*, *horsepower*, *peak_rpm*, and *price* also had 2-4 missing values. In this case these rows were dropped. Two more observations were also dropped due to invalid character ('?') for *num_of_doors*. In the end, the dataset was trimmed down to 193 observations and 25 variables.

To understand how each numeric variable is correlated with each other, a correlation matrix was created (see figure 2). Additionally, a principal component analysis (PCA) was run to visualize the correlation between variables. Although looking at the PCA graph one can say *peak_rpm* is highly correlated to *symboling*, using the correlation matrix it can be verified that the value is 0.23 which is significantly low for causing any collinearity related problems. Other variables like *city_mpg* and *highway_mpg* or *price* and *engine_size* have a strong correlation. Since the model is majorly going to use random forest, collinearity of these features should not pose a problem towards the accuracy of our predictions due to random selection of variables while making each tree.

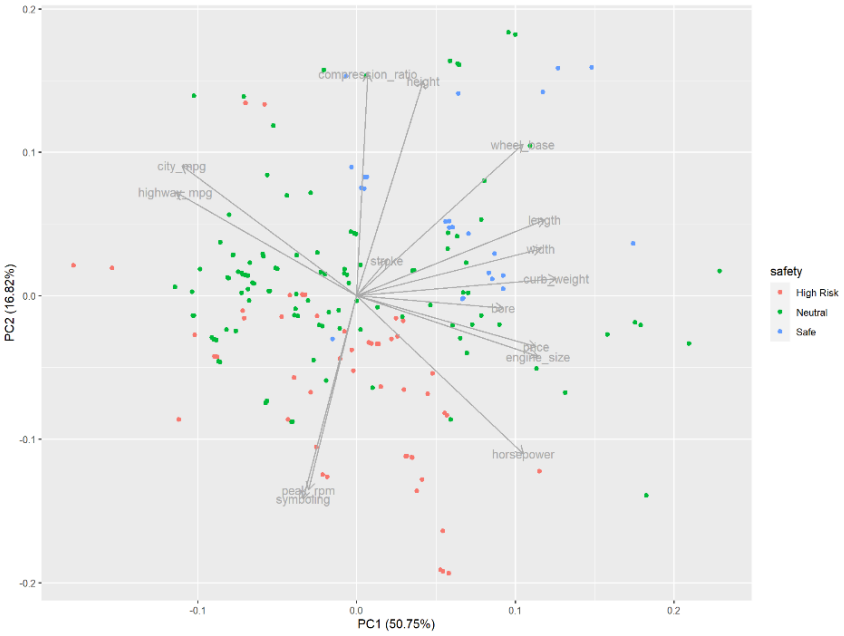


Figure 2 - PCA Graph

Model Selection and Methodology

To better understand the relationship between the predictors (characteristics of a car) and the decision variable (risk symbol), a regression tree was built. Regression tree provided an easy-to-interpret graphical representation of how each predictor effects the risk symbol assigned to a vehicle. The first tree was made with a CP of 0.01, to visualize the most significant factors used to estimate risk. *num_of_doors* was the most important criterion while assigning a risk rating to an automobile. Intuitively, *make* was also a

significant factor when classifying cars as risky or safe. The reasoning behind these factors being significant will be discussed in the results section.

Now that the most significant variables are identified, the next task was to find the optimal value of CP for which the OOS (out of sample) error is minimized. For this an initial tree was made with an extremely small CP of 0.00000001, which was then inspected for its OOS error at different levels of CP. The optimal CP was found to be 0.00066 and a tree making the best possible predictions was made after pruning (see figure 3). Although, a regression tree is a great way to explain the impact of the characteristics on the risk rating, even the best of trees lack accuracy when making predictions.

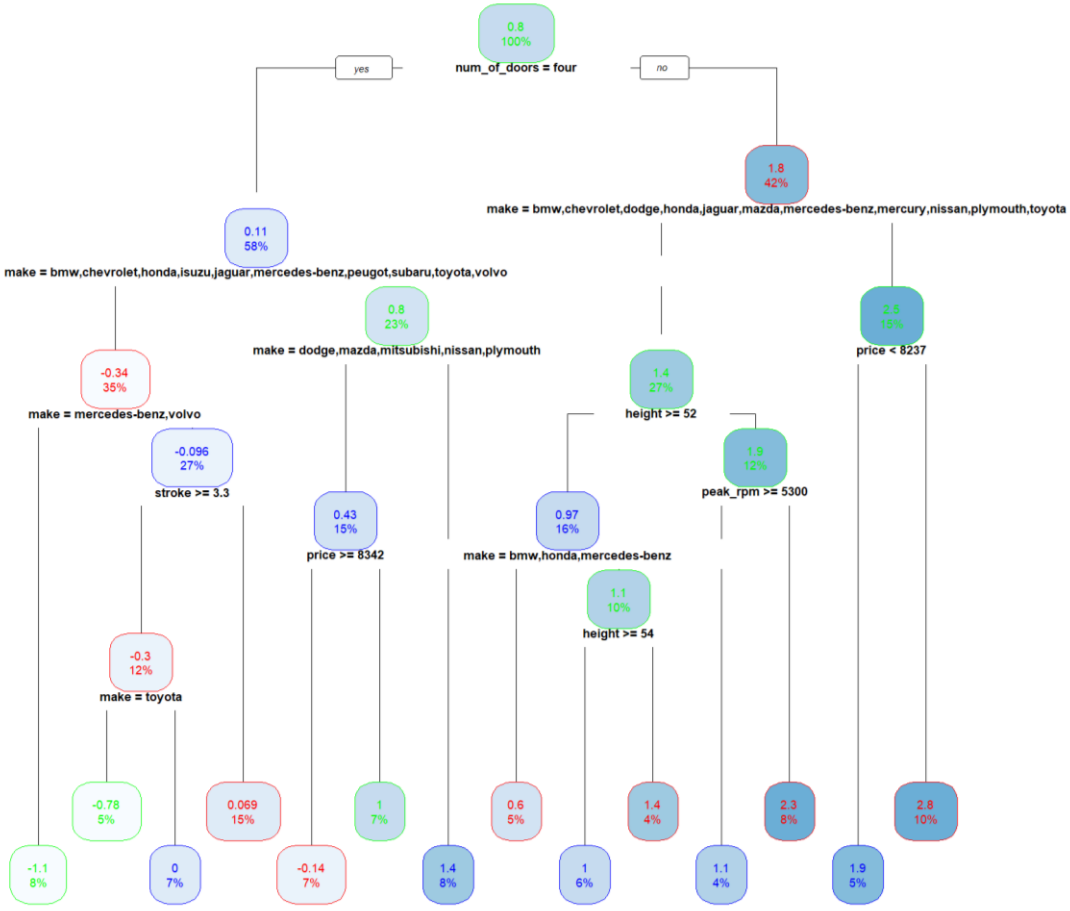


Figure 3 - Optimal Tree (CP=0.0006)

Therefore, random forest was used to reduce the high variance produced by trees and ensure randomness using bagging. The bagging algorithm is designed such that each tree is made on bootstrapped sample

and random forests ensure for each tree only m $(where\ m = \sqrt{p})$ predictors are chosen at random. A random forest was built on the model using 24 predictors where the number of trees were initially set to 10,000 and the out of bag error was traced. After 3000 trees, the MSE was consistent and hence it was not worth to add additional trees to save computational time. The random forest model with 3000 trees explained 81.59% of the variance with an MSE of 0.28. However, to investigate the importance of each predictor the relative predictive importance was evaluated using the *importance()* command (see adjacent table). Interestingly, all variables in the model have a positive value of *X.IncMSE* (% increase in MSE if predictor is removed from the model) and positive IncNodePurity (RSS increase if predictor is removed). In other words, this implies that none of the predictors are having a negative implication on the model. A variable importance plot was also exported which will be discussed in the results section.

	X.IncMSE	IncNodePurity
make	84.51	57.82
num_of_doors	59.48	57.20
wheel_base	46.27	33.72
height	34.35	20.15
body_style	32.27	23.80
curb_weight	31.18	9.93
width	31.07	7.72
bore	30.70	8.26
price	28.24	8.75
length	27.27	12.44
engine_size	27.27	7.16
stroke	23.77	4.70
city_mpg	23.57	8.01
horsepower	21.79	5.87
peak_rpm	20.67	4.04
highway_mpg	17.56	5.13
fuel_system	16.92	3.20
compression_ratio	15.54	3.69
engine_type	14.70	1.96
drive_wheels	13.64	1.57
num_of_cylinders	6.23	1.28
aspiration	4.00	0.43
fuel_type	3.29	0.11
engine_location	1.79	0.04

Figure 4 - Relative Variable Importance for RF

Right off the bat, the random forest yielded a low MSE however this section will explore how boosting effects the accuracy of the model. Before getting into boosting, the dataset was divided in two subsets with 70-30 split between training and test dataset to examine if boosting improved accuracy. Gaussian distribution was used since the symboling values were numeric, with an interaction depth of 5 and number of trees were 20,000.

The first boosted forest was run with all variables, which however resulted in an MSE of 0.48, which was higher than the random forests. Upon running the code multiple times, it was also observed that MSE was highly variable. This was likely due to a different train-test split each time as some characteristics like *engine_location*, *fuel_type*, or *aspiration* had factors that were only observed a few times in the dataset (see appendix 3 for relative variable importance). For example, for variable *engine_location* only 3 observations were "rear" and rest were "front". Splitting the dataset caused the model to overfit for instances when some categorical values were placed only in the test dataset and the model was not trained

for those values. Therefore, it was not feasible to split an already small dataset as the training observations were cut down to only 135, which reduced overall accuracy of the model. Another alternative was to remove these 'problematic' predictors and run a boosted forest on the rest of the variables. Although, this might solve the problem of unstratified sample, removing some predictors cause MSE to naturally increase due to loss of richness in the model.

With some external research, it was discovered that boosted forests (gbm) has a command to run a Cross-Validation Test on the dataset (see appendix 4 for code). Using 50 bags (k=50), the same model with all variables was run on the full dataset. The lowest value of the CV error was found to be 0.35 (MSE). Changing the interaction.depth also did not lower the MSE for boosted forest. Since the MSE of the boosted forest was still higher than the MSE obtained from random forests, it was signaling that boosted forests were overfitting on the dataset. To validate this hypothesis, the initial random forest model was trained on the exact same training dataset as the one used for boosted forests. Then, the predictions were made and MSE was calculated using the same test dataset. The MSE obtained was 0.33, though higher than what was identified initially with the same RF model trained on the full dataset, it was still lower than the boosted forest. On the other hand, when comparing RSS of the two models it was found that in fact boosted forest had a much lower RSS of 3.57e-13 whereas random forest had an RSS of 0.05.

Results

Decision Tree

As discussed in the methodology section, three statistical methods were used to create a model to make conclusions about the risk rating of a car for an insurance company. Initially there was a regression tree, which will help us understand how each feature impacts the decision variable (see figure 3 for optimal tree). *num_of_doors* was at the top of the tree and hence the most important predictor. On average cars with 4 doors had a much lower risk rating whereas all others had a higher risk rating in general. This can be attributed to the fact that many expensive sports cars have two doors, and hence pose a higher risk to the insurance company in case of any claims reported. *make* also had an impact on the risk rating, for example 4 door cars from brands like Mercedes Benz and Volvo which are known for the safety features had a risk rating of -1.1. Another one was *height* amongst the non-four door cars. For instance, the average risk rating of car with the height more than 62 inches was 0.97 whereas if it was less than 62 inches, it was 1.9. Using the same logic, sports cars usually tend to have a lower height than a regular sedan or SUV and hence

have more risk tied to them. Even though it can be interpreted how these top variables affect the decision variable, a tree is only limited to a few predictors. To have a closer look at all the variables the following section presents results from the random forest.

Random Forests

Random forest had a better out of bag performance however, it can also be noted that boosting would yield lower errors on the same dataset (due to its overfitting nature). In other words, random forest is the ideal model to make future conclusions about the *symboling* of an automobile, but boosted forest provides a better explanation of the current dataset in hand. Depending on the managerial objectives, one can be more specific as to which model to choose. For the purpose of this project, the following random forest model was chosen to understand the importance of each characteristic while determining the risk symbol to make recommendations to insurance companies and consumers:

```
Call:
 randomForest(formula = symboling ~ make + fuel_type + aspiration +      num_of_doors +
 body_style + drive_wheels + engine_location +      wheel_base + length + width + height
 + curb_weight + engine_type +      num_of_cylinders + engine_size + fuel_system + bore
 + stroke +      compression_ratio + horsepower + peak_rpm + city_mpg + highway_mpg +
   price, data = auto, ntree = 3000, importance = TRUE, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 3000
No. of variables tried at each split: 8

         Mean of squared residuals: 0.2795487
                   % Var explained: 81.59
```

Figure 5 - Optimal Random Forest Model

The mean square residuals for this model were the lowest at 0.28 which explains 81.59% of variance in the model.

Significance of Predictors

Random forest offers the ability to examine the importance of each predictor in the model. As discussed earlier in the methodology section, all predictors are significant for the model as each variable explains some variance. As also seen from the decision tree, *make* and *num_of_doors* are the most important characteristics. *wheel_base* is the distance between the front and rear wheels of the car, which is also a

significant predictor (see figure). Usually, larger *wheel_base* implies better ride quality which could reduce associated risk hence a lower *symboling* rating. To verify this relationship, a quick clustering analysis was run on just two variables *symboling* and *wheel_base*. The observations were classified in 3 clusters (see appendix 5) and an inverse relationship was identified between these two characteristics. That is, the cluster with the maximum car height had a safer rating than the one with lower car height.

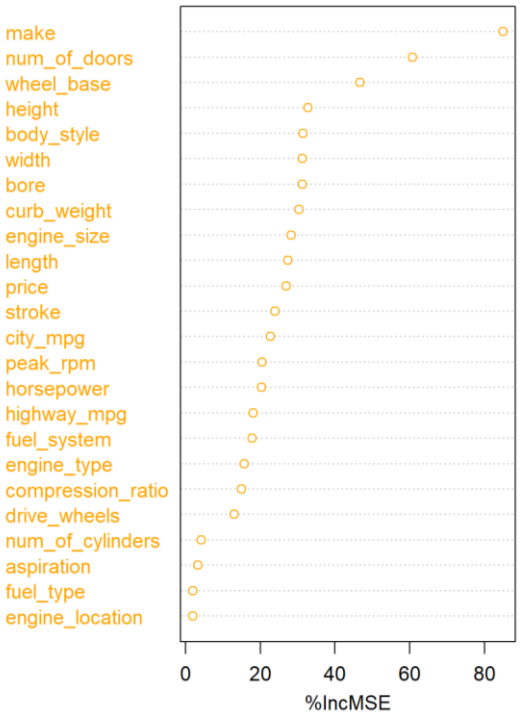


Figure 6 - Variable Important Plot

Factors like *fuel_type* and *engine_location*, are not highly crucial predictors of the risk rating, however they are still contributing positively towards the model's accuracy. If a factor had a negative value for X.Inc.MSE, then it would be better to exclude that variable to improve model's accuracy.

Conclusions

The results of the analysis show that certain characteristics of an automobile such as its make, height, and number of doors are likely key drivers of risk. Additionally, vehicles with certain characteristics like a larger wheel base are less likely to be associated with higher risk. Based on these findings, a model has been developed that can predict the risk associated with insuring a particular vehicle by just inputting a car's specifications. The accuracy of the model was tested using a variety of metrics, and found that it was able to make accurate predictions with a high degree of precision. This suggests that it is a valuable tool for both insurance companies and their customers, as it will help companies to accurately assess the risk of insuring a particular vehicle and to set appropriate premiums. By understanding the factors that contribute to a vehicle's risk, customers can be provided with the best possible coverage at the most competitive rates. Depending upon the richness of data available, the model can be tailored down to a smaller number of predictors using PCA to ensure its usability. As the dataset keeps growing with new car models being added, the algorithm is going to constantly learn and identify new trends and adjust the predictions accordingly. This analysis and report show that machine learning can be a powerful tool for predicting the insurance risk rating of automobiles.

Appendices

Appendix 1. Summary of Raw Dataset

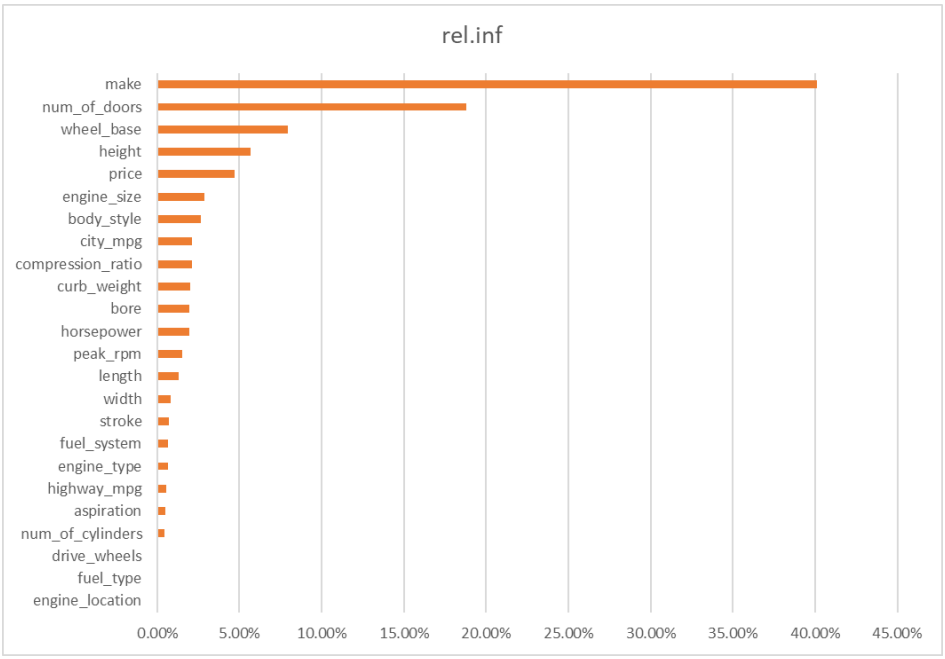
```
   symboling       normalized_losses     make            fuel_type          aspiration
 Min.   :-2.0000   Min.   : 65       Length:205         Length:205         Length:205
 1st Qu.: 0.0000   1st Qu.: 94       Class :character   Class :character   Class :character
 Median : 1.0000   Median :115       Mode  :character   Mode  :character   Mode  :character
 Mean   : 0.8341   Mean   :122
 3rd Qu.: 2.0000   3rd Qu.:150
 Max.   : 3.0000   Max.   :256
                   NA's   :41
 num_of_doors        body_style        drive_wheels       engine_location      wheel_base
 Length:205         Length:205         Length:205         Length:205         Min.   : 86.60
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 94.50
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 97.00
                                                                             Mean   : 98.76
                                                                             3rd Qu.:102.40
                                                                             Max.   :120.90

     length          width           height       curb_weight   engine_type        num_of_cylinders
 Min.   :141.1   Min.   :60.30   Min.   :47.80   Min.   :1488   Length:205         Length:205
 1st Qu.:166.3   1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   Class :character   Class :character
 Median :173.2   Median :65.50   Median :54.10   Median :2414   Mode  :character   Mode  :character
 Mean   :174.0   Mean   :65.91   Mean   :53.72   Mean   :2556
 3rd Qu.:183.1   3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935
 Max.   :208.1   Max.   :72.30   Max.   :59.80   Max.   :4066

  engine_size     fuel_system          bore           stroke      compression_ratio   horsepower
 Min.   : 61.0   Length:205         Min.   :2.54   Min.   :2.070   Min.   : 7.00     Min.   : 48.0
 1st Qu.: 97.0   Class :character   1st Qu.:3.15   1st Qu.:3.110   1st Qu.: 8.60     1st Qu.: 70.0
 Median :120.0   Mode  :character   Median :3.31   Median :3.290   Median : 9.00     Median : 95.0
 Mean   :126.9                      Mean   :3.33   Mean   :3.255   Mean   :10.14     Mean   :104.3
 3rd Qu.:141.0                      3rd Qu.:3.59   3rd Qu.:3.410   3rd Qu.: 9.40     3rd Qu.:116.0
 Max.   :326.0                      Max.   :3.94   Max.   :4.170   Max.   :23.00     Max.   :288.0
                                    NA's   :4      NA's   :4                         NA's   :2
    peak_rpm       city_mpg      highway_mpg        price
 Min.   :4150   Min.   :13.00   Min.   :16.00   Min.   : 5118
 1st Qu.:4800   1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7775
 Median :5200   Median :24.00   Median :30.00   Median :10295
 Mean   :5125   Mean   :25.22   Mean   :30.75   Mean   :13207
 3rd Qu.:5500   3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16500
 Max.   :6600   Max.   :49.00   Max.   :54.00   Max.   :45400
 NA's   :2                                      NA's   :4
```

Appendix 2. Correlation Matrix

	symboling	wheel_base	length	width	height	curb_weight	engine_size	bore	stroke	compression_ratio	horsepower	peak_rpm	city_mpg	highway_mpg	price
symboling	1.00	-0.54	-0.36	-0.25	-0.52	-0.23	-0.07	-0.14	-0.01	-0.18	0.07	0.23	0.02	0.09	-0.08
wheel_base	-0.54	1.00	0.88	0.82	0.59	0.78	0.57	0.50	0.17	0.25	0.38	-0.35	-0.50	-0.57	0.58
length	-0.36	0.88	1.00	0.86	0.49	0.88	0.69	0.61	0.12	0.16	0.59	-0.28	-0.70	-0.73	0.70
width	-0.25	0.82	0.86	1.00	0.31	0.87	0.74	0.54	0.19	0.19	0.62	-0.25	-0.66	-0.70	0.75
height	-0.52	0.59	0.49	0.31	1.00	0.31	0.03	0.18	-0.05	0.25	-0.08	-0.26	-0.11	-0.16	0.14
curb_weight	-0.23	0.78	0.88	0.87	0.31	1.00	0.86	0.65	0.18	0.16	0.76	-0.28	-0.78	-0.82	0.84
engine_size	-0.07	0.57	0.69	0.74	0.03	0.86	1.00	0.58	0.21	0.03	0.85	-0.22	-0.72	-0.74	0.89
bore	-0.14	0.50	0.61	0.54	0.18	0.65	0.58	1.00	-0.07	0.00	0.57	-0.27	-0.60	-0.61	0.55
stroke	-0.01	0.17	0.12	0.19	-0.05	0.18	0.21	-0.07	1.00	0.20	0.10	-0.07	-0.03	-0.04	0.10
compression_ratio	-0.18	0.25	0.16	0.19	0.25	0.16	0.03	0.00	0.20	1.00	-0.20	-0.44	0.31	0.25	0.07
horsepower	0.07	0.38	0.59	0.62	-0.08	0.76	0.85	0.57	0.10	-0.20	1.00	0.10	-0.83	-0.81	0.81
peak_rpm	0.23	-0.35	-0.28	-0.25	-0.26	-0.28	-0.22	-0.27	-0.07	-0.44	0.10	1.00	-0.06	-0.01	-0.10
city_mpg	0.02	-0.50	-0.70	-0.66	-0.11	-0.78	-0.72	-0.60	-0.03	0.31	-0.83	-0.06	1.00	0.97	-0.71
highway_mpg	0.09	-0.57	-0.73	-0.70	-0.16	-0.82	-0.74	-0.61	-0.04	0.25	-0.81	-0.01	0.97	1.00	-0.72
price	-0.08	0.58	0.70	0.75	0.14	0.84	0.89	0.55	0.10	0.07	0.81	-0.10	-0.71	-0.72	1.00

Appendix 3. Relative Variable Importance for Boosted Forest

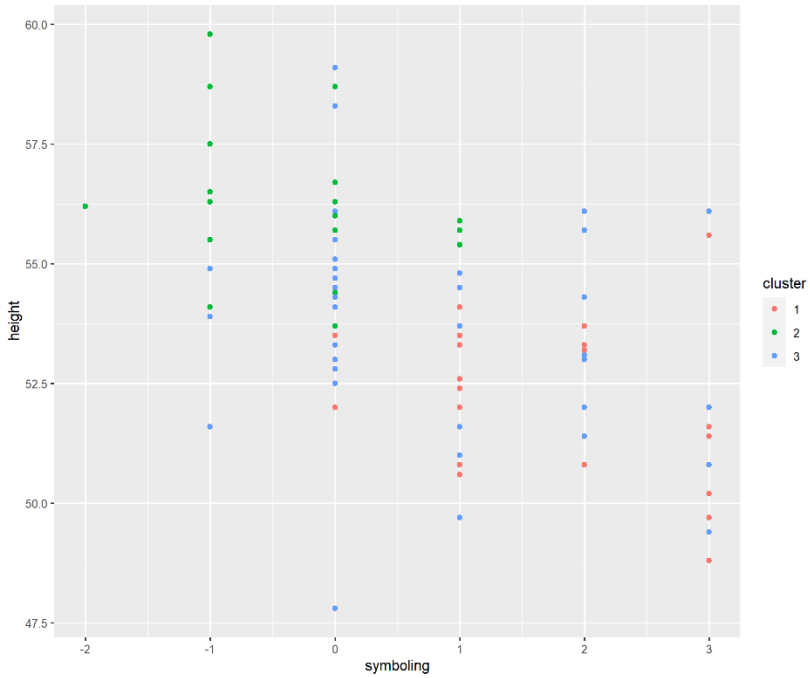
var	rel.inf
make	40.10%
num_of_doors	18.78%
wheel_base	7.92%
height	5.65%
price	4.70%
engine_size	2.87%
body_style	2.64%
city_mpg	2.13%
compression_ratio	2.10%
curb_weight	2.02%
bore	1.96%
horsepower	1.93%
peak_rpm	1.54%
length	1.33%
width	0.82%
stroke	0.69%
fuel_system	0.67%
engine_type	0.65%
highway_mpg	0.57%
aspiration	0.50%
num_of_cylinders	0.42%
drive_wheels	0.02%
fuel_type	0.00%
engine_location	0.00%



Appendix 4. Code for Cross-Validation for Gradient Boosting Model

```
# all variables with full dataset using cross validation
boosted3=gbm(symboling~.,data=auto, distribution="gaussian",
             n.trees=20000, interaction.depth=5, cv.folds=50)
summary(boosted3)

  #### finding MSE
  best=which.min(boosted3$cv.error)
  mse=boosted3$cv.error[best]
  mse
```

Appendix 5. Symboling and Wheel Base Clustering Results




```
K-means clustering with 3 clusters of sizes 57, 44, 92

Cluster means:
   symboling wheel_base
1  1.4912281   93.10351
2 -0.3863636  108.26364
3  0.9347826   98.06304
```