1985 VEHICLE RISK PREDICTION

Mohamad Dalati
261113383

1. Introduction

This project utilizes data scientist skills to conduct predictive analysis based on robust statistical foundations. The project shall incorporate our understanding from previous lectures and apply creative ideas to develop better business acumen using the power of analytics. The objective is to predict the insurance risk rating of automobiles and define relationships between variables. The dataset consists of data from the 1985 Ward's Automotive Yearbook.

The scope of the project is to predict the insurance risk rating of vehicles using various Classification models like Multiple Logistic Regression, Classification Trees, Random Fores, and Principal Component Analysis (PCA) to provide clear, professional and sound interpretations of the model.

2 Data Description

The dataset contains 205 observations with 26 variables, 15 of which are numerical variables and 11 categorical. The target variable named "symboling" was renamed to *risk_factor* for a more meaningful definition. risk_factor was converted into a binary categorical variable using the following metrics. "0" for classes between -2 and 0 represents a low insurance risk factor and "1" for classes between 1 to 3 represents a high insurance risk factor.

2.1 Data Cleaning

Several missing values showed a "?" sign, 2 values appeared in the *num_of_doors* categorical variable and several values among *normalized_losses, price, horsepower, peak.rpm, stroke, and bore* numerical variables. Since the dataset contains a limited amount of observations, the best approach is to fill them with a relative value instead. For the categorical variable *num_of_doors*, it was found that both missing values came from the sedan cars of Mazda and Dodge. Using scatter plot 2.1, it was illustrated that most sedans are with four doors, therefore, the values were replaced by "four". On the other contrary, the numerical values were filled using MICE package which uses a variety of techniques, such as predictive mean matching and Bayesian linear regression, to generate values.

2.2 Model checks

The initial step was to convert categorical variables into factors and numerical variables into numerics to plug them into further checks. For outliers, Box plots were created and variables *stroke* and *engine size* seemed were detected. According to [1], strokes range between 2 and 4, so the detected outliers are meaningful since they fall within the range. Referencing to [2], the engine size chart showed that the sizes detected were normal. More importantly, classification

models like decision trees and random forests are less sensitive to outliers and may not require explicit outliers detection and removal.

2.3 Data Visualization

For basic understanding purposes, several plots were made. Plot2.3a illustrates that Toyota was the most frequent brand with ~35 observations and Mercrury was the least in among other brands. Scatter plot 2.3b shows that Wagon & Sedan car styles mostly come in four doors unlike hatchbacks which come in two, and few observations are on the convertibles with two doors. Plot 2.3c shows that a linear relationship exists between horsepower and price, and the majority of the data points are on the lower price margin. Plot 2.3d shows that shorter cars have higher risk ratings. Since they are more like to be sports cars, and more vulnerable in terms if making an accident

3 Model Selection

3.1 Preliminary Feature Selection

The random forest classifier (importance = TRUE) feature was used to provide a rough guide on which predictors to start with. As shown in plots 3.1a and 3.1b, the top 10 variables include *make, num_of_doors, wheel_base, normalized_losses, height, curb_weight, length, body_style, width, and bore* as accuracy measures. For instance, the height of the car, if removed, would decrease 17.5% of the model accuracy. Whereas the Gini impurity captures the impurity of the nodes at the end of the tree, for instance, *wheel_base* has scored a 10% contribution of purity which helps when performing predictions. After that, to gain a better interpretability of the model, Multiple Logistic Regression was used.

3.2 Model Interpretation

The Multiple logistic Regression model was selected due to its high interpretability advantage over other models. After fitting the glm model with each of the features, the coefficient results are combined and displayed in tables 3.2a - 3.2d, Based on glm to predict risk factor probability as a function of the top 10 important features, a vehicle is more likely to be at high insurance risk if, the car brand is Mercury or Saab brand, with a high depreciation value, and shorter distance between the front wheels and back wheels. Also, the vehicle is likely going to be classified as high insurance when it has wider measurements, is shorter in height (like sports cars) with two doors and is lightweight.

3.3 Model Visualization

Since Random Forest branching is challenging to visualize, a classification tree with the top 10 predictors was performed. The optimal CP score was not able to be identified as it is a classification model, but it was created on several Complexity Parameter (CP) values, it was initially set to 0.0001 and kept increasing till 0.05 where it was visually found to be a good balance between not overly complex and not overly simplistic tree. Figure 3.3.1 shows the out-of-sample error as a function of the cp value(lowest error at cp of 0.023). Referencing figure Figure 3.3.1b, shows that the *number_of_doors* was the first important feature followed by the brand name or *make*, and wheelbase distance. For instance, if the car has 4 doors, and is manufactured by either BMW, Chevrolet Isuzu, Jaguar, Mazda, Benz, Peugeot, Renault, Subaru, Toyota, and Volvo. It predicts that it will be classified as 0 (i.e. Low insurance risk factor) and 38% of automobiles fit into this classification. Whereas, a car with 4 doors, and not manufactured by the above-mentioned criteria with a wheelbase distance < 99 inches is likely to be classified as 1 (i.e. high insurance risk) and only 13% fall into this classification.

3.4 Model Selection

3.4.1 Prediction

Since the target variable is risk classification, model prediction accuracy is paramount. Therefore, Logistic regression and classification tree were not selected, since they are prone to overfitting. Random Forest Classifier achieved high accuracy with minimal overfitting. To improve the accuracy of Random Forest Classifier, tunerf() function was used to find optimal parameters. Number of tree (ntree) and the number of variables tried at each split (mtry) parameters were tuned respectively. Figure 3.4a shows ntree vs OOB error rate, and it shows how the error rate was constant after 400 trees, so optimal ntree = 400 was chosen. Figure 3.4b shows that the lowest OOB error was at 2 variables, so mtry = 2 was optimal. Moreover, to know the optimal number of predictors that shall be considered, trial and error were performed. Several Variables are interchangeably changed and re-fitting was done to try and find which predictors could give the highest OOB rate and at which predictors the OOB becomes almost identical.

3.4.2 Variability

Lastly, a PCA plot was executed to capture the variability of observations and distinguish between what variables contributed to the low-risk factor and vice versa. Seven PCAs were

created as shown in Figure 3.4c, most of the variability found across the data can be explained by the five variables in PC1, as shown below, they consist of: *wheel_base, length, width, curb_weight, and price.* Figure 3.4d shows the percentage of variance explained, ~60% of the variability was captured by PC1, and ~20% was captured for PC out of the seven PC created.

4 Results

After model configuration, the optimal predictors are: *make, normalized_losses, wheel_base, width, height, num_of_doors, length, curb_weight, body_style, and bore.* The model scored an accuracy rate of 92.2% (100% - 7.8% = 92.2%), which means that around 7.8% of the OOB observations got misclassified. The Confusion matrix is shown in Figure 4.1, where the diagonals labelled in dark blue represent the correct predictions the model has made. According to the heat map, 100 observations were classified as class 1(i.e. high risk) correctly. Additionally, 17 predictions were misclassified, for instance, 10 observations that should have been classified as high-risk factors of 1, were classified as a low-risk factor of 0 instead.

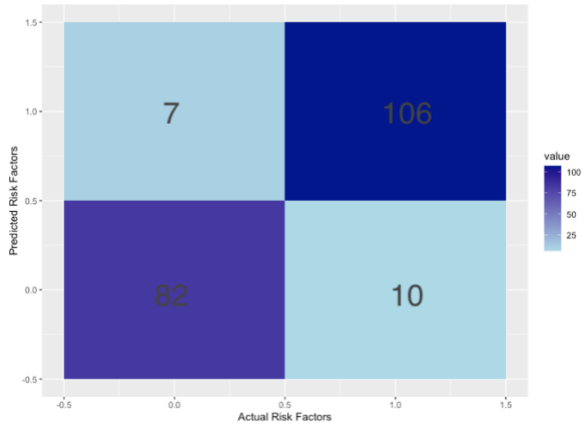

Figure 4.1a

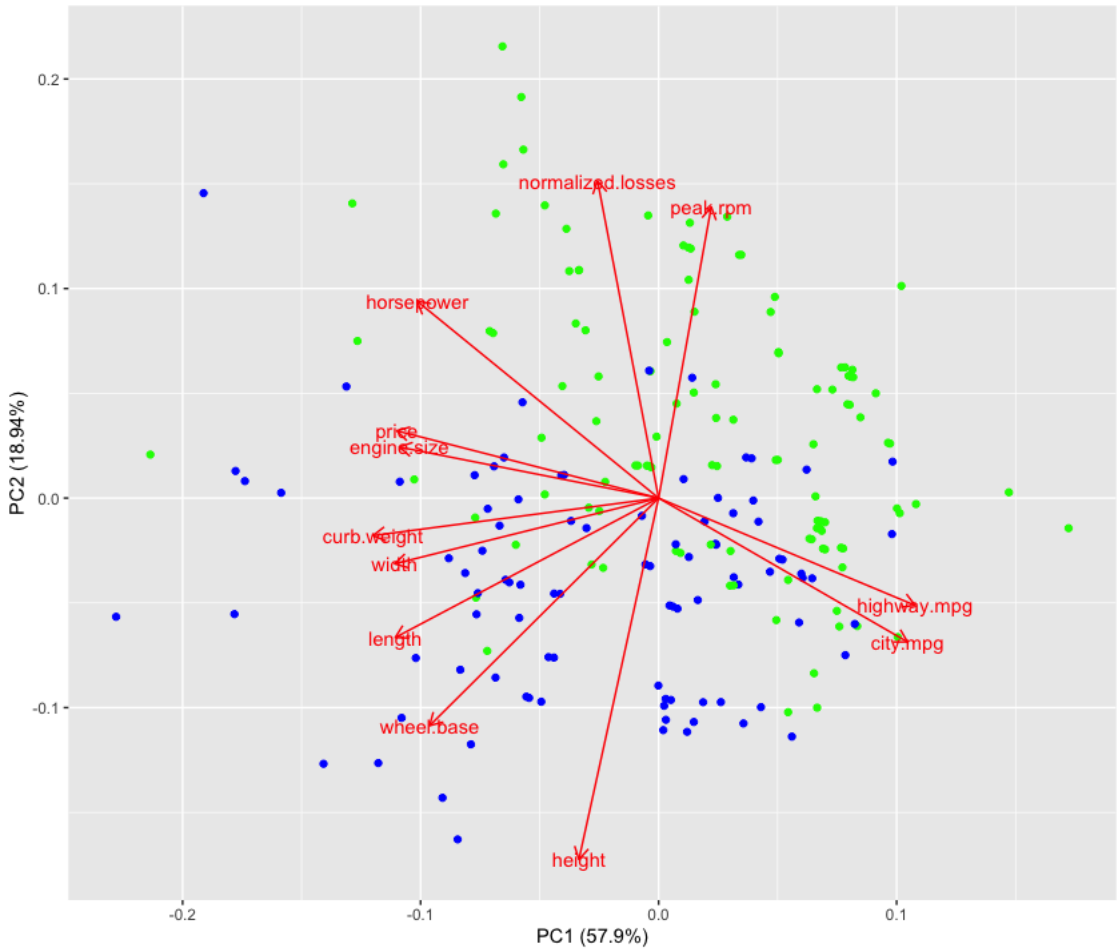

Figure 4.1b

The first two components were shown using autoplot() since they represent the majority of the variance (up to 85%). Blue observations represent the low insurance risk factor and green observations represent the high-risk factor.

Low-risk factors seem to have larger values in variables like weight, length, width of the car compared to the high-risk factor in green. Also, as the weight, width, length, and wheelbase of the car increase, the price increases, in particular, the width and weight of the car are highly correlated. On the contrary, normalized losses (i.e. car depreciation value) and its height are almost negatively correlated. Meaning that taller cars like SUVs for example are not likely to be subjected to depreciation. Moreover, it is easier to detect automobiles with low-risk factors since the variables are pointing towards them with very few observations of high-risk presence. High risk vehicles tend to have high annual loss per year (i.e. car depreciation) and Car engine rotation (directly related to speed of car), Whereas, low risk vehicles are longer & wider. Price and the size of the engine seem correlated to each other, same goes for weight of the car and it's width. High acceleration and city acceleration and correlated as well. Height of the car is

negatively correlated with its acceleration. What doesn't make sense is how horsepower is negatively correlated to highway and city speed, but faster cars on highways will have high horsepower.

5 Classification/Predictions and Conclusions

5.1 Classifications/Predictions

Several interpretations of model relationships were gained using different models like Multiple Logistic Regression, Classification Tree, and PCA. The Random Forest model provided a high accuracy rate of 92.2% whilst minimizing overfitting issues. The error rate of Logistic Regression was not computed since the target variable is of extreme importance and critical to insurance companies. Therefore, logistic regression is not likely to be more accurate than Random Forest or GBT, if it does, the model will suffer from overfitting like our friend Spliney! However, it can be used to help validate an insurance quote.

5.2 Business Insights

Certain brands of vehicles have known reputations for getting into accidents or have higher repair costs, leading to higher insurance imposed on them.

Cars that are heavy weight, wide, and long are more stable on the road, thus, they can obtain a safer insurance risk.

Insurance companies quote cars with high acceleration rates on more premium insurance since speeding can cause accidents which will add repair costs for the insurance company.

6 Appendices

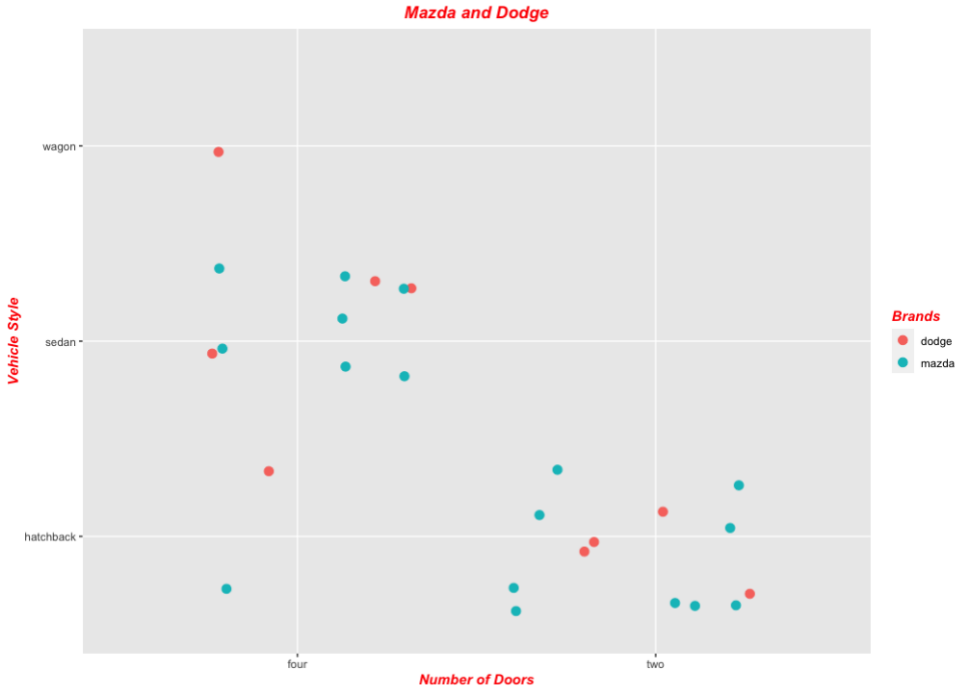

Figure 2.1

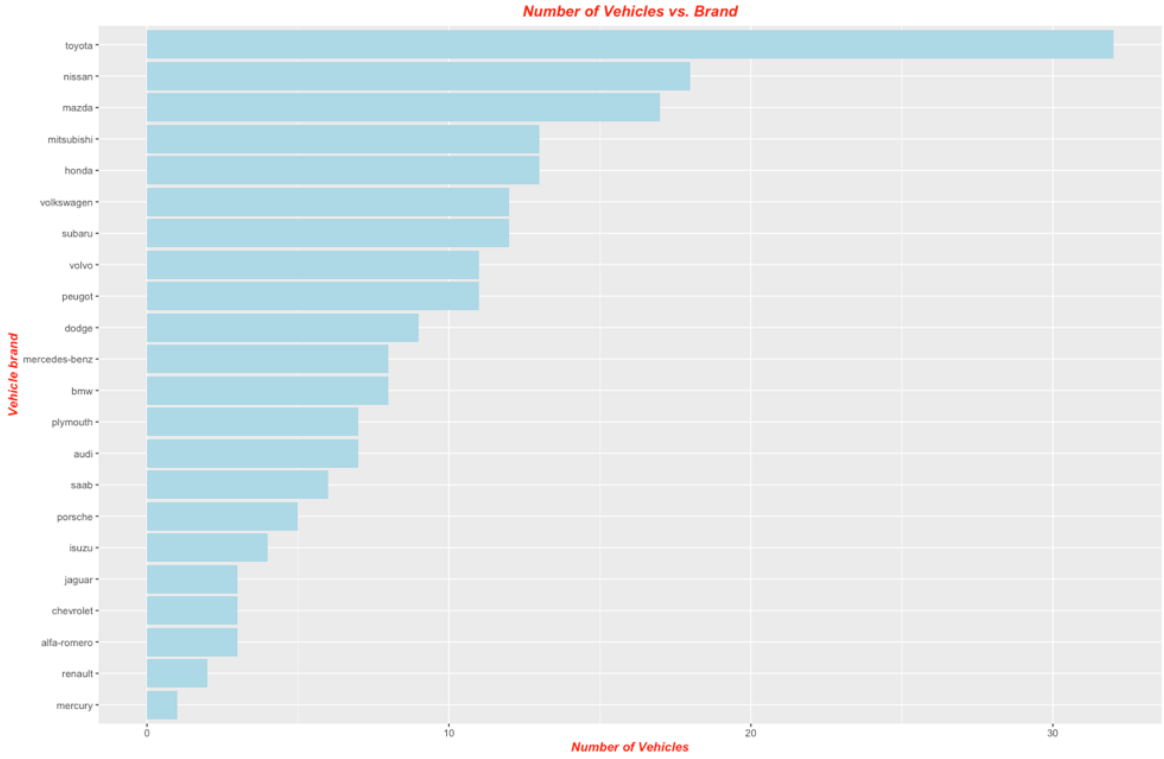

Figure: 2.3a

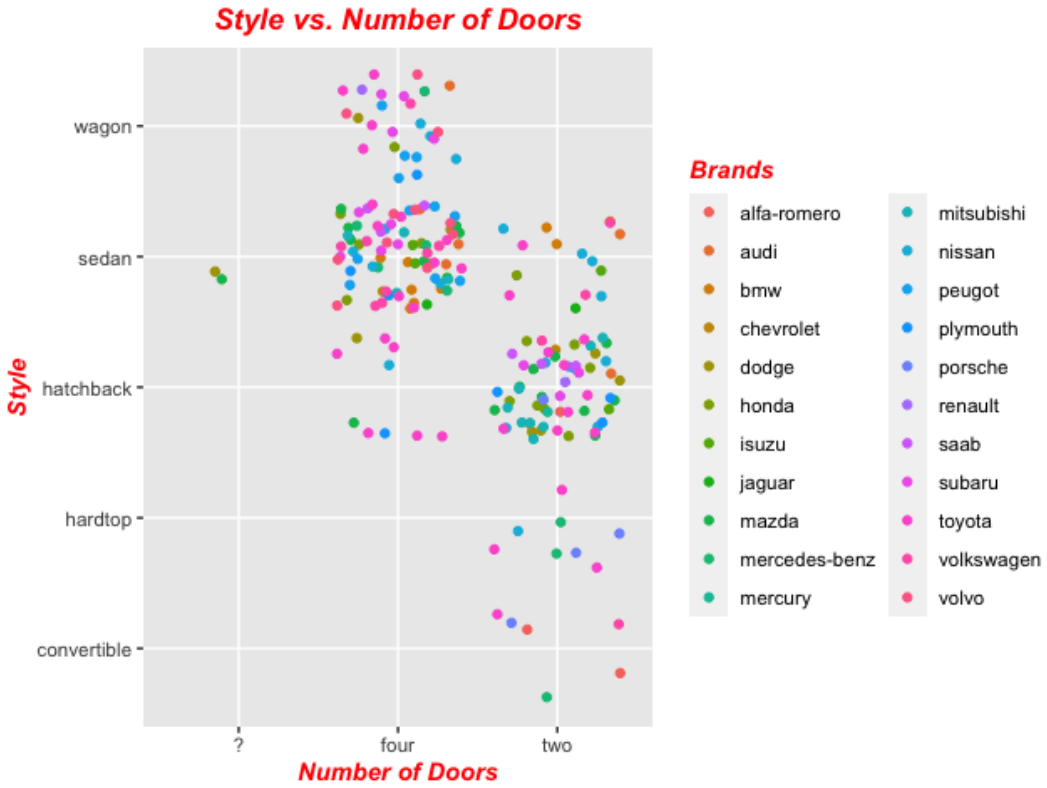

Figure 2.3b

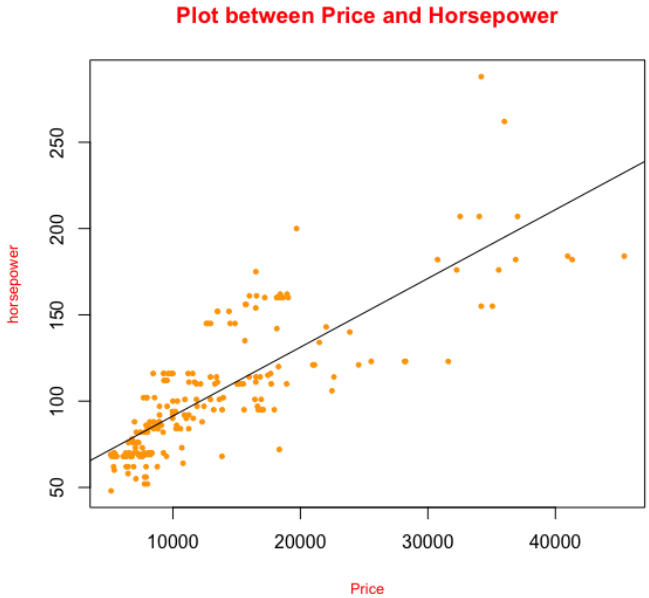

Figure: 2.3c

	0	1	Mean Decrease Accuracy	Mean Decrease Gini
Normalized Losses	14.874	15.669	18.066	7.778
Wheel Base	16.719	14.109	19.923	10.079
Length	11.197	10.985	13.743	5.687
Width	8.727	9.109	11.784	3.253
Height	13.921	11.511	16.249	7.670
Curb Weight	10.731	9.669	14.035	4.198
Engine Size	6.773	7.863	9.763	2.301
Bore	9.968	8.198	11.669	3.710
Stroke	7.494	8.941	10.445	1.814
Compression Ratio	5.027	4.948	7.117	1.496
Horsepower	4.792	5.942	8.537	1.847
Peak Rpm	8.078	9.807	11.171	3.214
City Mpg	2.079	6.704	7.252	1.752
Highway Mpg	2.108	6.393	6.286	1.525
Price	7.174	6.934	9.944	2.883
Make	27.505	16.989	28.780	16.942
Fuel Type	0.245	1.981	1.971	0.148
Aspiration	0.431	3.399	3.611	0.303
Number of Doors	20.910	19.112	24.201	14.921
Body Style	8.401	11.410	12.527	6.318
Drive Wheels	5.316	3.916	5.981	0.718
Engine Location	0	0	0	0
Engine Type	3.000	2.713	3.503	0.242
Number of Cylinders	2.476	2.635	3.259	0.530
Fuel System	3.847	4.799	6.139	1.136

Figure: 3.1a

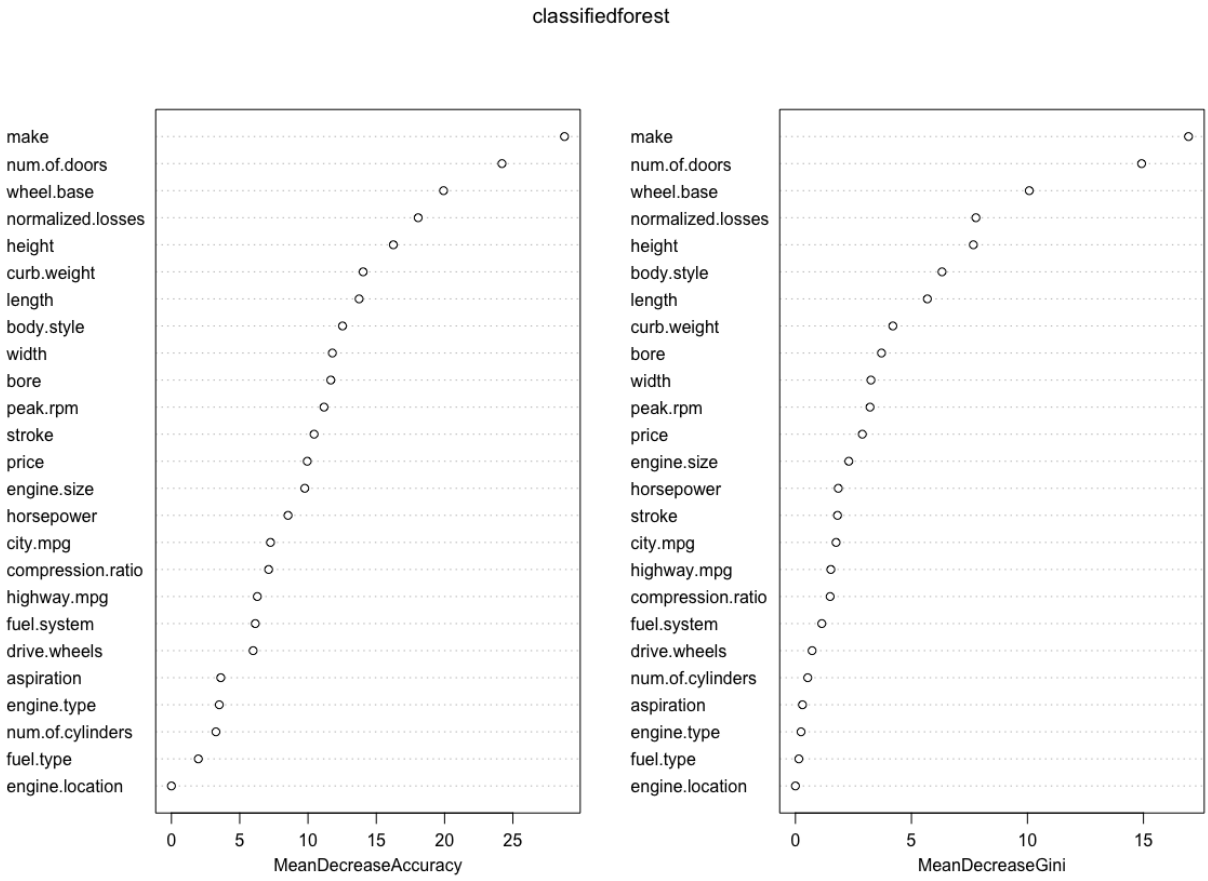

Figure 3.1b

	Dependent Variable:
	Risk Factor
Audi	-16.842
Bmw	-22.356
Chevrolet	-16.645
Dodge	-16.896
Honda	-17.776
Isuzu	-18.680
Jaguar	-38.612
Mazda	-17.796
Mercedes-Benz	-20.348
Mercury	2.481
Mitsubishi	-17.366
Nissan	-17.755
Peugot	-38.798
Plymouth	-16.668
Porsche	-2.221
Renault	-18.772
Saab	0.100
Subaru	-18.205
Toyota	-18.006
Volkswagen	-16.544
Volvo	-35.685
Normalized Losses	0.053*
Constant	12.148
Observations	205
Log Likelihood	-72.875
Akaike Inf. Crit.	191.751
Note:	p<0.1; p<0.05; **p<0.01

Figure 3.2a

	Dependent Variable:
	Risk Factor
Wheel Base	-0.642*
Width	0.917*
Constant	3.097
Observations	205
Log Likelihood	-91.254
Akaike Inf. Crit.	188.508
Note:	p<0.1; p<0.05; **p<0.01

Figure 3.2b

	Dependent Variable:
	Risk Factor
Height	-0.328*
Automobile with Two Doors	2.876*
Constant	16.910*
Observations	205
Log Likelihood	-85.761
Akaike Inf. Crit.	177.521
Note:	p<0.1; p<0.05; **p<0.01

Figure 3.2c

	Dependent Variable:
	Risk Factor
Curb Weight	-0.001
Length	-0.111*
Price	0.0001*
Constant	20.948*
Observations	205
Log Likelihood	-112.018
Akaike Inf. Crit.	232.035
Note:	p<0.1; p<0.05; **p<0.01

Figure 3.2d

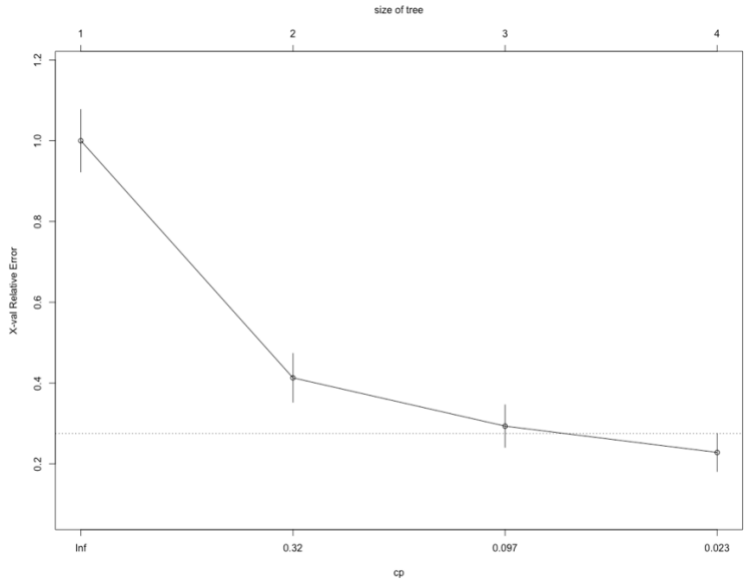

Figure 3.3.1

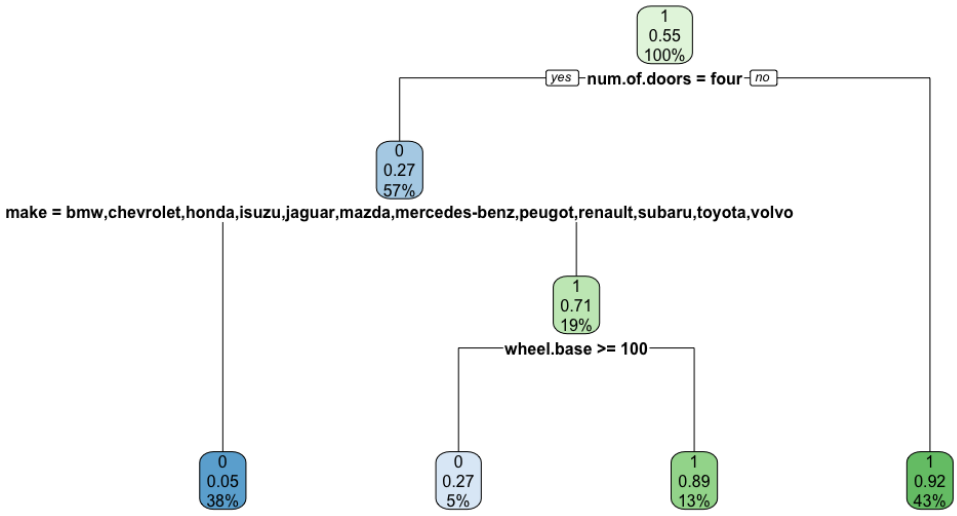

Figure 3.3.1b

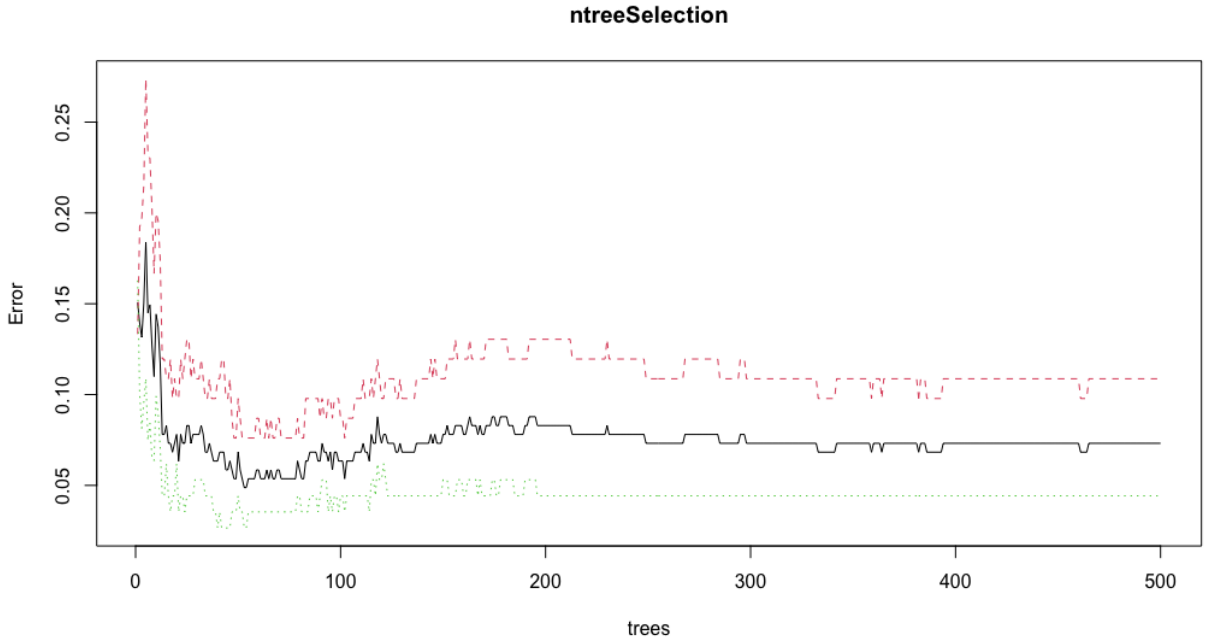

Figure 3.4a

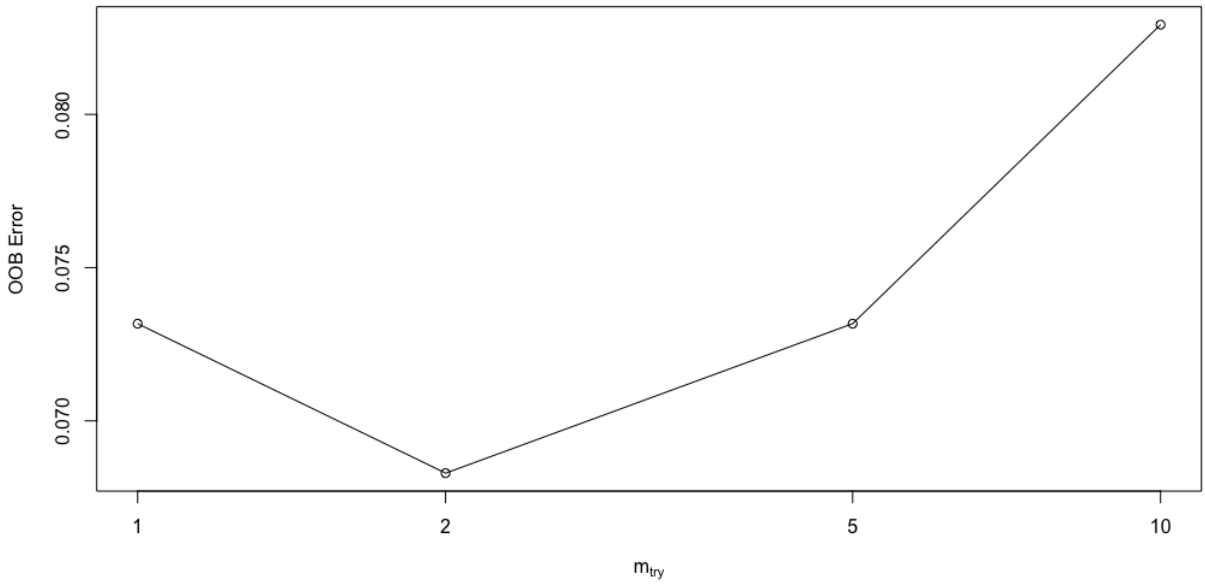

Figure 3.4b

```
Rotation (n x k) = (7 x 7):
                          PC1         PC2        PC3         PC4         PC5         PC6         PC7
normalized.losses -0.02510656 -0.69687652  0.7102082 -0.00151745  0.09039115 -0.03078044 -0.01472460
wheel.base        -0.43391144  0.18552183  0.1777658  0.38727443  0.15385083  0.74642603 -0.12205914
length            -0.45484322  0.05142943  0.1072251  0.23042330 -0.53741845 -0.20894340  0.62721974
width             -0.44168867 -0.10124098 -0.2175523  0.31245763  0.65151982 -0.47446015  0.01058746
height            -0.22960596  0.59973559  0.5461357 -0.45113723  0.17304566 -0.22568508 -0.07021968
curb.weight       -0.45179249 -0.14196274 -0.1204456 -0.07535969 -0.45427593 -0.18334786 -0.71802565
price             -0.39023573 -0.29531991 -0.3039837 -0.70006261  0.13620309  0.29762521  0.26622624
```

Figure 3.4c

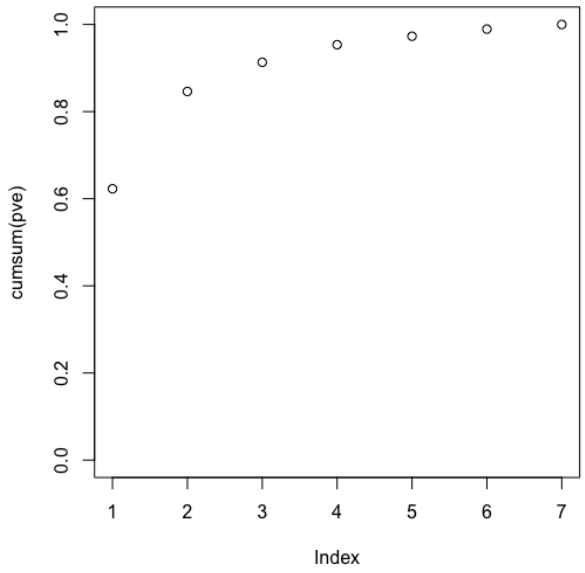

Figure 3.4d

[1] https://www.berrymanproducts.com/two-stroke-vs-four-stroke-engines/

[2] https://www.cjponyparts.com/resources/engine-size-chart