MCGILL UNIVERSITY

IS THE CAR SAFE?

MGSC662
FALL 2022
FINAL PROJECT

SHANSHAN LAO

DECEMBER 15, 2022

Contents

1 Introduction

What type of car has a higher risk? Some may argue cars with higher horsepower are more dangerous because they run faster or cars with more cylinders as they generate more power. Even the size of a car can be an indicator of the safety level of an auto. It may seem intuitive to classify risky cars, but intuition can be misleading, and sometimes inefficient. Using a dataset of 205 automobile insurance reports from Ward's 1985 Yearbook, this research aims to determine the most critical predictors in identifying if a car is safe or risky.

Data pre-processing and exploration were performed with some multivariate analysis methods, such as principal component analysis. Different classification algorithms were tested and evaluated in order to build the most effective model. The best model was built using Random Forest, which showed that the following characteristics are more effective in classifying cars:

`num-of-doors, make, wheel-base, height, length, curb-weight, body-style`

2 Data Description

2.1 Missing Values

Before diving into data exploratory, data quality was assessed. There were 46 observations containing missing values, with 41 of them missing information on `normalized-losses`. This predictor was eliminated for two main reasons. One is that the missing records take up over 20% of the data, simply dropping the observations or replacing them with other values might lead to unexpected bias. Secondly, the value of the losses was normalized and calculated for a particular size classification, meaning it could be inferred by other characteristics.

Missing values in predictor `price` were replaced by the average price of its respective make. The remaining observations that still contain missing values were dropped, leaving us 197 records and 25 predictors in the dataset. Future analysis will be performed based on this cleaned dataset.

2.2 Target Variable – symboling/is_risky

The dependent variable `symboling` is an insurance risk level indicator with 6 levels ranging from -3 to 3 (but it starts from -2 in this dataset). The smaller the value, the safer the auto. Any auto that has a rating larger than 0 was considered more dangerous than its price indicated. For simplicity, `symboling` was converted into a binary variable `is_risky`, where 1 indicates the car is risky and 0 otherwise. Around 55% of the cars were evaluated as risky.

2.3 Numerical values

The 15 quantitative variables were explored by visualizing them with boxplots and histograms. The histograms showed that while some variables had a normal or nearly-normal distribution (`length, stroke`), some others were right-skewed (`engine-size, horsepower, price, compression-ratio`). This distribution tells us that most cars in the market had an average car size between 170 to 180 and an average pistol size between 3 and 3.5, with only a few being extremely large or small. Even though the car price ranged from around $5,000 to $45,000, almost 85% were under $20,000, and half were below $10,000. This demonstrated the consumer affordability of an automobile in 1995.

It is also noted that these variables have huge differences in their range. (`stroke` and `price` for example). Such inconsistency might affect the interpretability of the model and generate biased results, as variables with larger values tend to be weighted more heavily. In an attempt to eliminate this impact, the scale() function was introduced to ensure that every numeric variable is measured in the same way, while retaining the original distribution.

Some extreme outliers were spotted in the boxplot for the variable `compression-ratio` revealed. Meanwhile, its histogram revealed that there was a huge gap in its distribution, which might be something to watch out for. However, no data cleaning was done for this variable at this stage, as it is uncertain if the predictor will have a significant impact on the final prediction.

2.4 Categorical values

The 10 categorical variables were analyzed with bar plots, which display the frequency of each category for each variable. The variables `engine-locations, fuel-type` and `aspiration` only had two subcategories each, and they were primarily centred on one of them. Approximately 98.5%(194 out of 197) of the cars had their engine located in the front, 90.6%(178 out of 197) used gas as fuel rather than diesel, and 81.7% had a standard aspiration system instead of turbo. These three variables were removed because they offered very little variability in the classification model.

Some other predictors with more than two levels, such as `num-of-clinders` and `engine-type`, also had a relatively high frequency on specific sub-categories. Around 78.7% of the autos had four cylinders, and 73.1% had an overhead camshaft (OHC) engine. These two variables were re-classified by grouping the other sub-categories together as `Others`. There were 22 brands of cars and 8 types of fuel systems in this dataset. The frequency in each level in the predictor varies, some of them did not have sufficient records. For the levels that had less than 5 observations, they were also renamed as a new category, `Others`.

2.5 Relationship between Variables

Both a correlation matrix and a Principal Component Analysis (PCA) plot were generated to understand the relationship between numerical variables. (See Appendix A)

According to the results from PCA table (See Appendix A), variables `highway-mpg, length, width, engine-size` represent the highest variance. The variables that have the same direction and are close to each other in the PCA plot are considered highly correlated. Here we have `engine-size` and `price`, `bore` and engine-size, `highway-mpg` and `city-mpg`, `curb-weight` and `width`, `wheel-base` and `stroke`. For each pair of these variables, one of them should be dropped to avoid multicollinearity.

In order to decide which variables to keep, a random forest model was built to help identify the feature importance. (see Appendix B) The variables with relatively lower importance were dropped, including `highway-mpg`, `price`, `bore`, `width` and `stroke`.

3 Model Selection

The model was chosen by first deciding the candidate algorithms, then testing each model with different combinations of all remaining predictors, and finally evaluating the performance of each model by comparing their accuracy score.

In terms of predictive modelling, there are a number of algorithms available. Different methods are used to predict different types of information. For example, if the target variable is a continuous variable, then linear and non-linear regression models can be used. In this case, the goal was to predict if the car is risky or not, the target of the problem is a binary outcome. Hence the methods that are suitable for predicting continuous results won't be appropriate for solving this problem. The algorithms for classification tasks on binary categorical variables include Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Classification Trees, and Random Forests.

Another factor to consider is the data type of the predictor. As shown in the feature importance graph (see Appendix B), the top predictors are a mix of quantitative and categorical variables. The type of car (`make`) and the number of doors (`num-of-doors`) were ranked the highest. Dropping either one of them would cause a decrease over 20 in mean accuracy. Therefore, methods that can only deal with numeric input values were ruled out, including both LDA and RDA.

A logistic regression model was first attempted with the remaining predictor. As previously mentioned, outliers were found in some predictors. An outlier test was thus performed on this model, two outliers were detected and removed. The dataset was further split into a training set and a test set with a ratio of 75%-to-15%. For the sake of fair comparison, all three models (Logistic Regression, Decision Tree, and Random Forest) were all trained and tested using the same sets, and the whole training-testing process was repeated 30 times. By doing so, I was able to simulate a K-fold cross-validation, which helped decide the final model more accurately.

4 Results and Conclusions

4.1 Out-of-Sample Performance

The 3 models were compared using their average accuracy rate in classifying the risk of the car. The accuracy was calculated by dividing the number of correct predictions by the total number of tested observations. Note that the logistic regression did not always converge and would return the fitted probabilities instead of the class 0 or 1, its result was thus rounded for better comparison.

Surprisingly, all 3 models performed very well. The logistic Regression model had an approximate 87% average accuracy, the Decision tree model had around 89% accuracy, and the Random Forest model outperformed both models with an average score of 94%. The Random Forest model was again rerun with the entire dataset, returning an OOB of 6.67%, which was very close to the accuracy score calculated before.

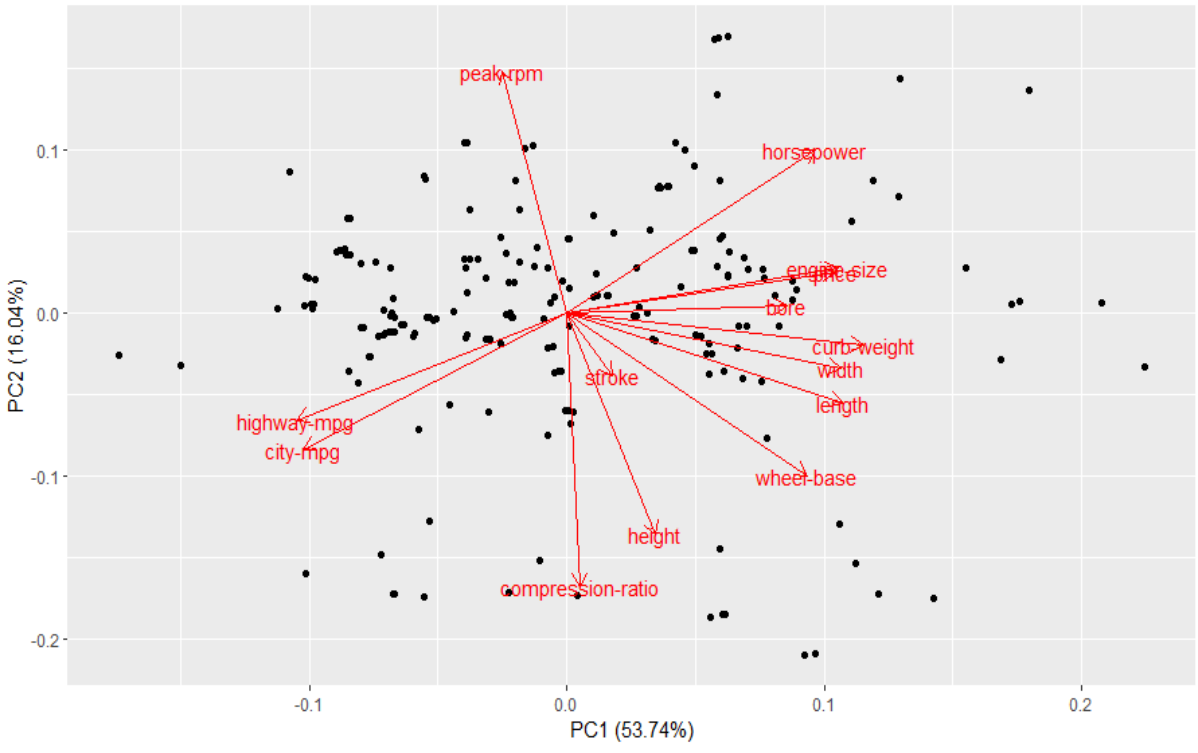
4.2 Model Interpretation and Conclusion

Feature importance was calculated for the new model. (see Appendix D) Now the model tells us the most effective way of identifying if a car is risky or not is to look at its number of doors, its make and wheelbase. It is probably because most two-doors cars were sports/racing cars and sports cars were always much faster than any other car, which makes them more dangerous on the road. The second highest feature is the car brand (`make`). It is because different car companies had a different focus on the type of cars they produced. For example, luxury brands like Audi, Porsche, and BMW tend to make cars with better engines and higher horsepower, and again, higher speed comes with higher risks.

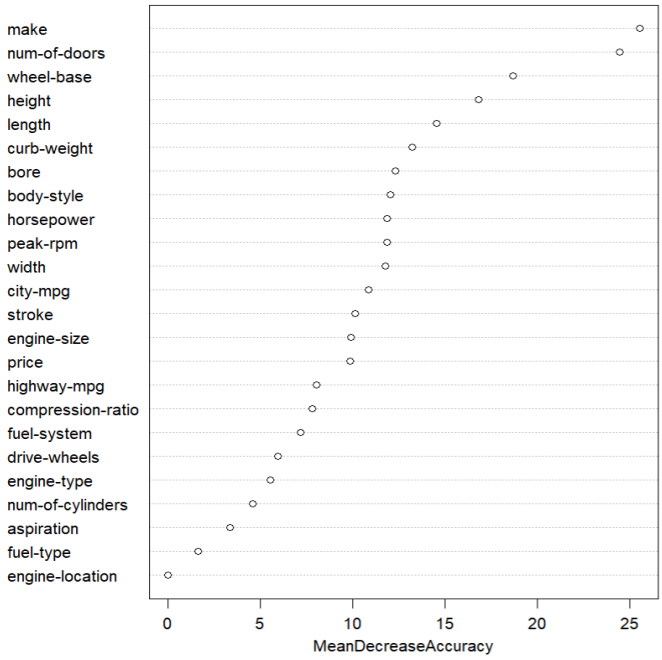
As a driver, such information might be useful for understanding the risk of their own car and potentially reducing insurance costs. As an insurer, it can help the company to improve efficiency in quoting and claiming process, provide a better estimate, and to decide whether to insure the car or not.

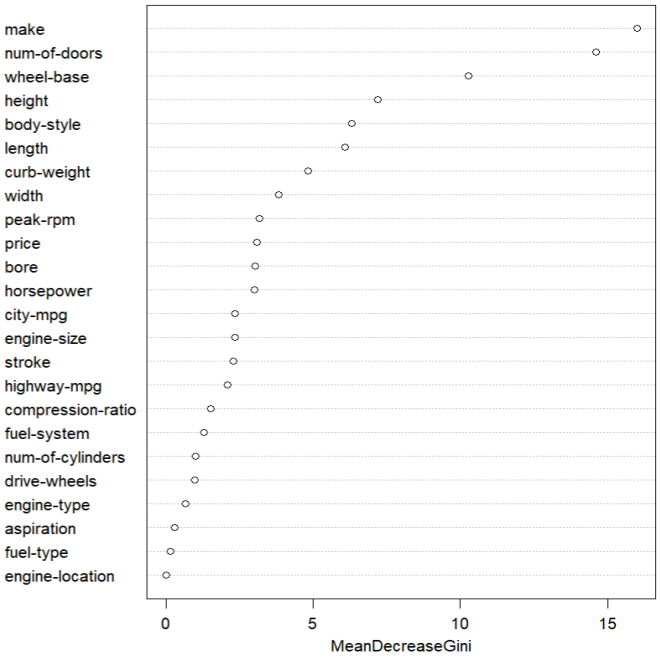
A Principal Component Analysis

	PC1	PC2	PC3
wheel-base	0.286	-0.304	0.131
length	0.328	-0.168	0.133
width	0.325	-0.103	-0.063
height	0.106	-0.414	0.478
curb-weight	0.352	-0.06	-0.055
engine-size	0.322	0.085	-0.245
bore	0.261	0.014	0.155
stroke	0.054	-0.116	-0.689
compression-ratio	0.016	-0.512	-0.309
horsepower	0.295	0.306	-0.144
peak-rpm	-0.077	0.454	0.061
city-mpg	-0.312	-0.255	-0.125
highway-mpg	-0.32	-0.2	-0.13
price	0.319	0.078	-0.138



B Appendix: Feature Importance - Initial Exploration





C Appendix: Feature Importance - Numeric Result

	MeanDecreaseAccuracy	MeanDecreaseGini
make	25.19	15.582
num-of-doors	23.472	14.703
wheel-base	19.514	8.372
height	16.646	7.622
length	14.779	6.716
curb-weight	14.007	5.242
width	12.997	4.137
bore	12.163	3.058
body-style	12.024	6.275
horsepower	11.916	2.864
peak-rpm	11.252	3.13
stroke	10.445	2.114
engine-size	10.189	2.426
price	8.939	3.313
city-mpg	8.798	2.665
fuel-system	8.523	1.28
compression-ratio	7.758	1.702
highway-mpg	7.334	2.123
drive-wheels	7.098	1.018
engine-type	5.61	0.584
aspiration	3.46	0.256
fuel-type	3.329	0.163
num-of-cylinders	2.953	0.822
engine-location	1.001	0.005

D Appendix: Feature Importance - Final Model