Analysis of Automobile Safety: Exploring Car Safety through Random Forest Classification and Clustering

MSGC 661

2023-12-10

Introduction

In a time of rapid technological progress, the world of cars changed dramatically. Cars became more than just a way to get around. They started to represent fashion, speed, and luxury. Because of this, people started to look for different things in cars. Some wanted fast and luxurious cars, while others looked for something more affordable and comfortable.

However, during these varied preferences, one aspect remained consistently important: safety. This brought up important questions. Does spending more on a car mean it's safer? Are bigger cars naturally more secure, or do luxury cars have better safety features? This project explored these questions using a detailed dataset on cars. The project aimed to understand how different aspects of a car relate to its safety, shedding light on what really makes a car safe.

The origin of this project was driven by a key question: Which cars are safer? To answer this, a classification model using the Random Forest technique was utilized, concentrating on the safety aspects of cars. Also, K-Means Clustering was chosen to investigate common beliefs like 'larger cars are safer' and 'more expensive cars, due to their advanced safety features, are naturally safer'.

Data Description

'Dataset 5 - Automobile Data' offered an extensive exploration of car characteristics, featuring 26 unique columns on 205 different automobiles. Each column revealed a different aspect of an automobile, from specific features to safety assessments. This dataset primarily focused on two aspects: the detailed specifications of each car and its insurance risk rating. The insurance risk rating was a distinctive feature of this dataset, showcasing how a car's risk compares to its price. This rating, termed "symboling," adjusted a car's risk level, with a scale where +3 signifies a higher risk and -3 a lower risk. The dataset also provided detailed information on each car's brand, model, and engine details, offering a thorough perspective on various car attributes.

Data Preprocessing and Transformation

Missing Value

While the dataset didn't have typical null values, it did include 59 instances of '?' marks, which were essentially equivalent to missing data. A considerable number of these occurred in the 'normalized.losses' column. Since this particular column wasn't vital for the analysis, I decided to exclude it entirely, along with any rows that contained missing values in other columns. This approach of removing the 'normalized.losses' column was chosen to reduce any potential issues that might arise from its incomplete data.

Data Type Adjustments

In the initial dataset, several columns with numerical values were incorrectly labeled as 'character' types. To ensure precise modeling, these columns needed to be converted to 'numeric' types. For the Random Forest analysis, it was important to handle categorical variables effectively. Therefore, variables like 'make', 'fuel type', 'aspiration', 'number of doors', 'body style', 'drive wheels', 'engine location', and 'symboling' were transformed into factor variables. This transformation was crucial for enabling the classification analysis. Notably, the 'symboling' variable, although not categorically defined in its original form, was also converted into a factor to align with the analysis requirements. This conversion played a significant role in ensuring the accuracy and effectiveness of the modeling process.

Exploratory Data Analysis

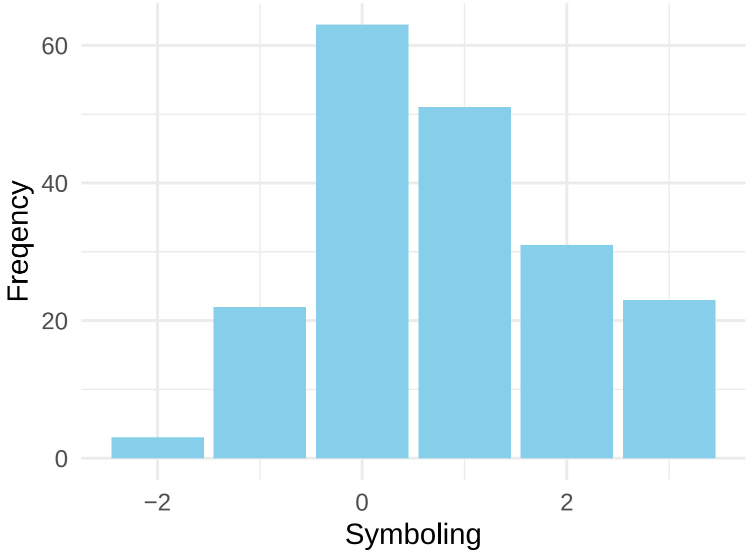

Figure 1: Bar Plot of 'symboling' variable

The 'symboling' data was normally distributed, with only a few automobiles receiving a symboling of -2. This scarcity of data in the -2 'symboling' category might have led to inaccuracies in classification, as the model might not have had enough examples to learn from for this specific rating.

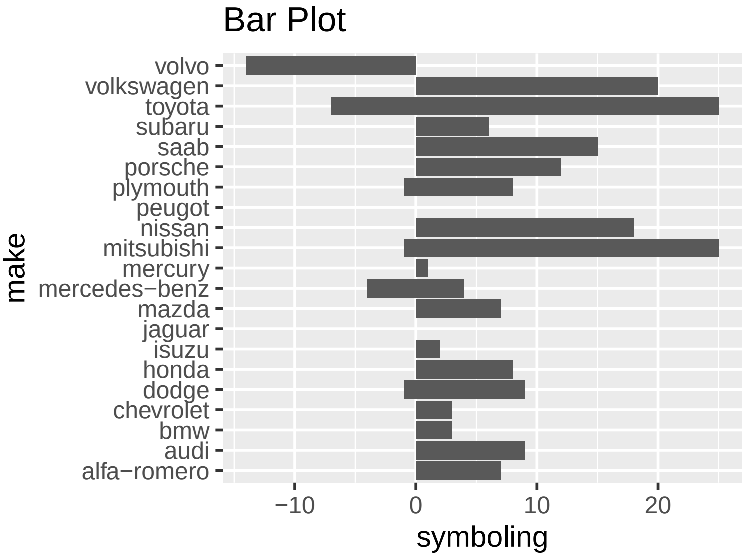

Figure 2: Relationship Between Automobile Brand and Symboling

Upon examining the relationship between automobile brands and their 'symboling' values, it was observed that Volvo cars generally had a very low 'symboling' values. This indicated that cars manufactured by Volvo were perceived as safer compared to other brands. This finding challenged the hypothesis that more expensive cars were inherently safer, as luxury automobile brands like Porsche and Alfa Romeo did not show a significant difference in safety ('symboling' values) compared to other, possibly less expensive, brands.

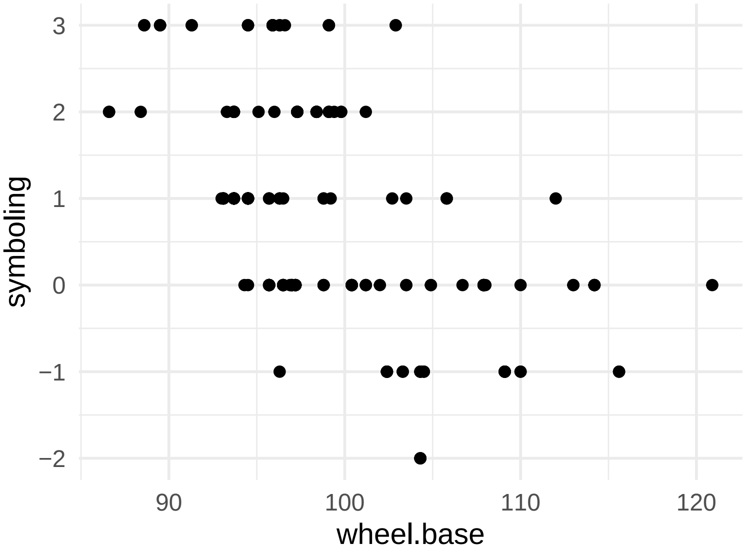

Figure 3: Relationship Between Wheelbase Symboling

Figure 4 illustrated a clear trend: as the wheelbase of a car increased, its ‘symboling’ value decreased. This indicated a negative relationship between ‘symboling’ and wheelbase. In simpler terms, cars with longer wheelbases tended to have lower symboling values, implying they were less risky and, therefore, safer. This observation supported the hypothesis that bigger cars were generally safer.

PCA and Correlation

	PC1	PC2	PC3	PC4	PC5
symboling	-0.094	-0.394	-0.342	0.261	-0.243
wheel.base	0.292	0.294	0.097	-0.155	-0.004
length	0.330	0.148	0.087	-0.064	-0.052
width	0.323	0.091	-0.081	-0.066	-0.153
height	0.117	0.417	0.382	-0.144	-0.166
curb.weight	0.350	0.034	-0.081	0.019	-0.065
engine.size	0.318	-0.118	-0.232	0.072	0.006
bore	0.257	-0.024	0.090	0.398	0.344
stroke	0.053	0.069	-0.572	-0.639	0.403
compression.ratio	0.020	0.431	-0.433	0.165	-0.497
horsepower	0.292	-0.307	-0.078	0.001	-0.071
peak.rpm	-0.084	-0.376	0.245	-0.519	-0.493
city.mpg	-0.306	0.252	-0.156	0.053	-0.098
highway.mpg	-0.317	0.201	-0.151	0.055	-0.090
price	0.315	-0.098	-0.134	0.057	-0.299

Table 1: Principal Component Table (First 5 Components)

In the first principal component (PC1), a notable correlation emerged among features like wheelbase, length, width, curb weight, price and engine size. This correlation suggested that these variables collectively characterized aspects related to a car’s size and engine capacity. Conversely, the second principal component (PC2) unveiled an intriguing inverse relationship between two sets of attributes. On one side of this spectrum were factors such as height and compression ratio, and on the other, symboling and peak RPM. The inverse relationship suggested that vehicles designed for higher performance and RPMs might possess different safety ratings (symboling) and structural characteristics compared to those emphasizing engine efficiency and structural height.

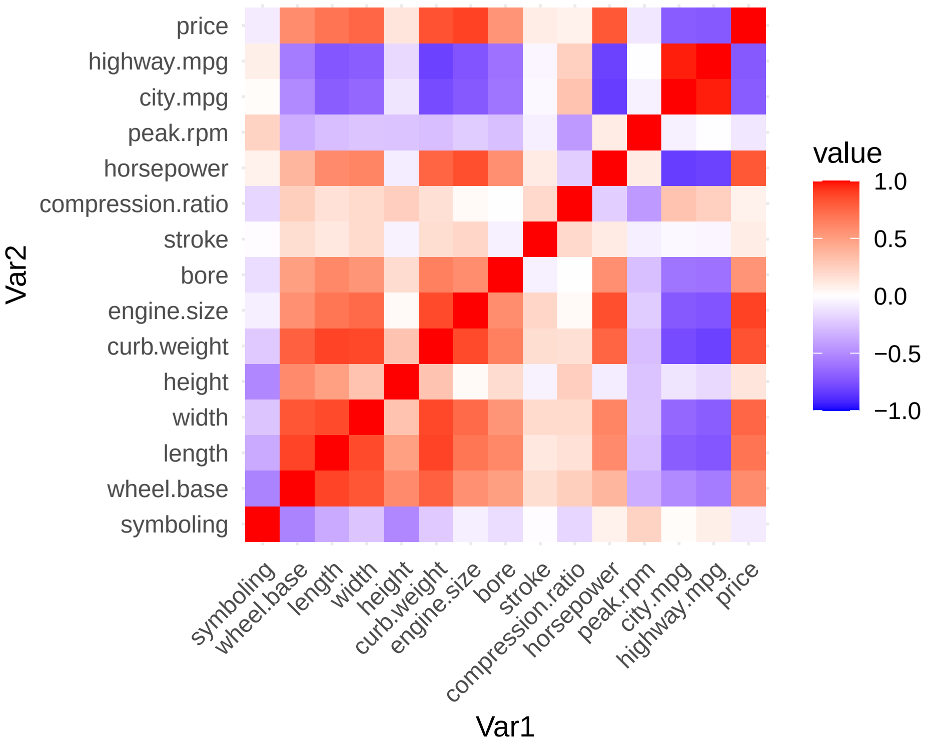

Figure 4: Correlation Matrix

Feature Correlations:

Principal Component Analysis (PCA) was pivotal in our study for uncovering potential correlations between various car attributes, with a specific emphasis on two key areas: the size of the automobile and its engine characteristics. To validate these connections, a correlation matrix was employed, where any correlation value above 0.80 was considered significant, denoting a strong relationship between the variables.

The first notable set of correlations revolved around the size of the automobile. Variables such as 'wheelbase', 'length', 'width', and 'curb weight' were found to be highly correlated. This indicated that cars with a longer wheelbase typically also have increased length and width, along with greater curb weight. These correlations suggested a trend where larger cars, in terms of physical dimensions, also tend to be heavier.

Secondly, engine characteristics also showed a clear pattern. Heavier cars with larger engines were more powerful but less fuel-efficient and pricier. Larger engines correlated with increased curb weight and horsepower, impacting both fuel efficiency and price. This highlighted a trade-off in car design, where more power and size come at the cost of fuel efficiency and higher costs.

And Curb Weight had significant correlations with 'engine size', 'highway mpg', and 'price'. Heavier cars, as suggested by these correlations, are usually longer, wider, have larger engines, consume more fuel on highways, and come with a higher price tag.

Model Selection & Methodology

Predictor Selection for Classification Model

The selection of predictors for the classification model was a crucial step to ensure the effectiveness of the analysis. To avoid multicollinearity issues, it was essential to first examine the correlations among predictors. Following the insights gained from the exploratory data analysis, variables such as ‘length’, ‘width’, ‘engine size’, ‘highway mpg’, ‘city mpg’, and ‘price’ were excluded due to their high correlations. The remaining variables were then subjected to Random Forest analysis. The selection of predictors was guided by the output of this analysis, specifically focusing on metrics such as the Out of Bag (OOB) estimate of error rate, Mean Decrease Accuracy, and Mean Decrease Gini. These metrics provided a robust basis for selecting the most relevant predictors for the model.

Why Random Forest Classification?

Random Forest was chosen for its robustness and effectiveness in handling large datasets with multiple predictors. It is particularly adept at managing overfitting, a common issue in complex models. Additionally, Random Forest can handle both categorical and numerical data, making it an ideal choice for this diverse dataset.

Model Selection

Random Forest Classification

The primary objective was to uncover the relationship between an automobile’s safety and its specifications. The Random Forest model was tailored for this purpose. During the model tuning process, it was observed that after 450 trees, the Out of Bag (OOB) error rate started to increase. Therefore, the final model was set with ntrees=450. To further validate the model, the dataset was split into two parts: one for training the model and the other for testing its performance. 30% of the dataset was allocated for testing. The model’s accuracy was the key metric for validation.

K-Means Clustering for Additional Insights

In addition to the Random Forest model, K-Means clustering was utilized to explore the relationships between ‘symboling’ and ‘wheel base’, as well as ‘symboling’ and ‘price’. This method was particularly suited for this analysis since the interest was in understanding the characteristics of car safety in relation to its price and size. By comparing the characteristics of each cluster, deeper insights could be gleaned about how these factors interplay in determining a car’s safety profile.

Result

The analysis of the Random Forest model and the K-Means clustering yielded valuable insights into the relationship between various automobile characteristics and their safety ratings.

Random Forest Model Analysis

Key Predictors

The model identified a set of variables - make, wheel base, number of doors, height, curb weight, bore, horsepower, and body style - that had a strong influence on predicting automobile safety. Notably, attributes such as wheel base, height, and curb weight, which relate directly to a car's size, were significant predictors. This supported the initial hypothesis that larger cars might be safer. However, due to the inherent characteristics of Random Forest models, it was difficult to determine the exact nature of these relationships, such as whether they were positively or negatively associated with safety.

Model Performance

The Out of Bag (OOB) estimate of error for the model stood at 11.85%, demonstrating a robust performance. When applied to a separate testing dataset, the model achieved an accuracy rate of 79.3%, further affirming its effectiveness.

Clustering Analysis

In the clustering analysis, different patterns and relationships were observed, which provided additional perspectives on car safety in relation to various features. The clustering process, especially when examining attributes like wheel base and price in relation to safety ratings, offered a nuanced view of how these factors interplayed in determining a car's safety profile. The clustering results complemented the findings from the Random Forest model, enriching the overall understanding of the factors contributing to automobile safety.

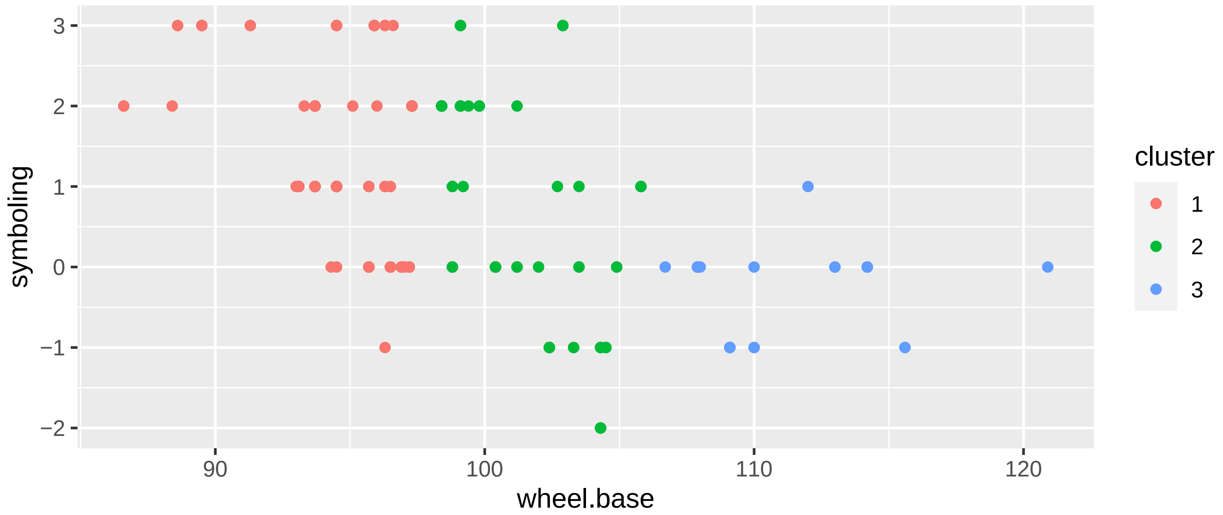

Figure 5: K-Means Clustering (Wheelbase and Symboling)

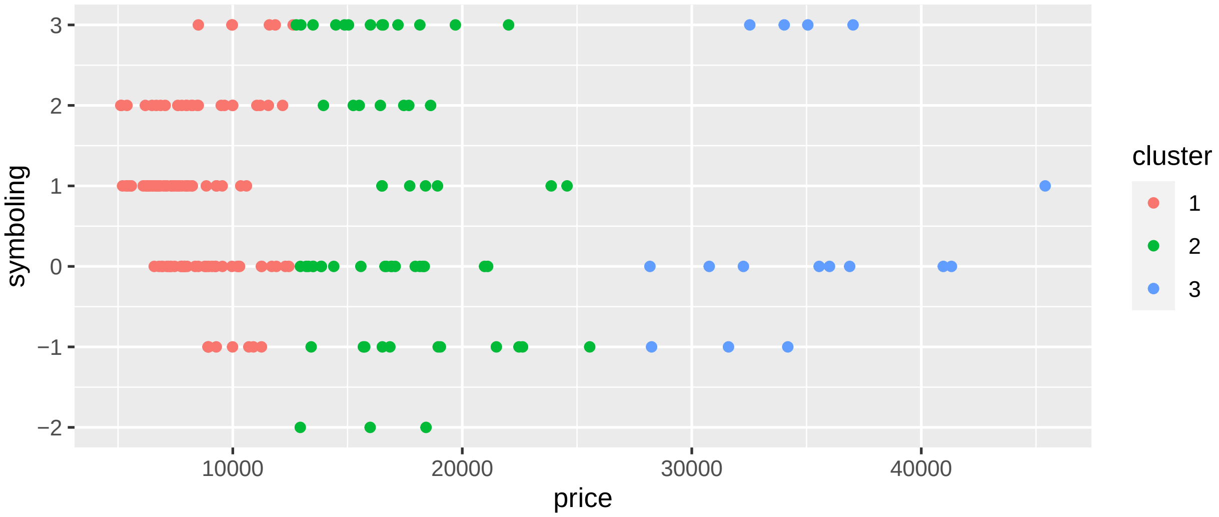

Figure 6: K-Means Clustering (Price and Symboling)

Table 2: Average of Wheel.Base and Symboling by Cluster

cluster	wheel.base	symboling
1	94.69541	1.2018349
2	101.48276	0.5344828
3	110.94231	-0.3076923

Table 3: Average of Price and Symboling by Cluster

cluster	price	symboling
1	8311.379	0.8706897
2	17047.984	0.7049180
3	34997.688	0.6250000

The clustering analysis, conducted using K=3, uncovered some intriguing patterns:

Cluster Analysis on Wheel Base

The clusters characterized by a lower average wheel base value exhibited higher average symboling values. This pattern suggested that smaller cars tended to have higher risk ratings on average, leading to the inference that larger cars were generally safer.

Cluster Analysis on Price

Interestingly, the clusters with a higher average price did not demonstrate a notable difference in their average symboling values when compared to clusters with a lower average price. This finding indicated that while the size of the car, as denoted by the wheel base, seemed to have a clear connection with safety, the price of the car did not exhibit a strong correlation with its safety ratings.

These insights from the clustering analysis provided a deeper understanding of the factors influencing automobile safety. The results particularly highlighted the significance of a car's physical dimensions over its price in terms of safety implications.

Conclusions

From the combined analysis using Random Forest and K-Means clustering, it can be concluded that: The size of the car (as indicated by factors like wheel base, height, and curb weight) is a significant predictor of its safety. The price of the car, however, does not show a strong correlation with safety ratings. The Random Forest model, with an accuracy of 79.3% on the testing dataset, effectively captures the relationship between car specifications and safety. This analysis offers a nuanced understanding of what factors contribute to the safety of automobiles, highlighting that size plays a more critical role than price in determining a vehicle's safety.

Appendix

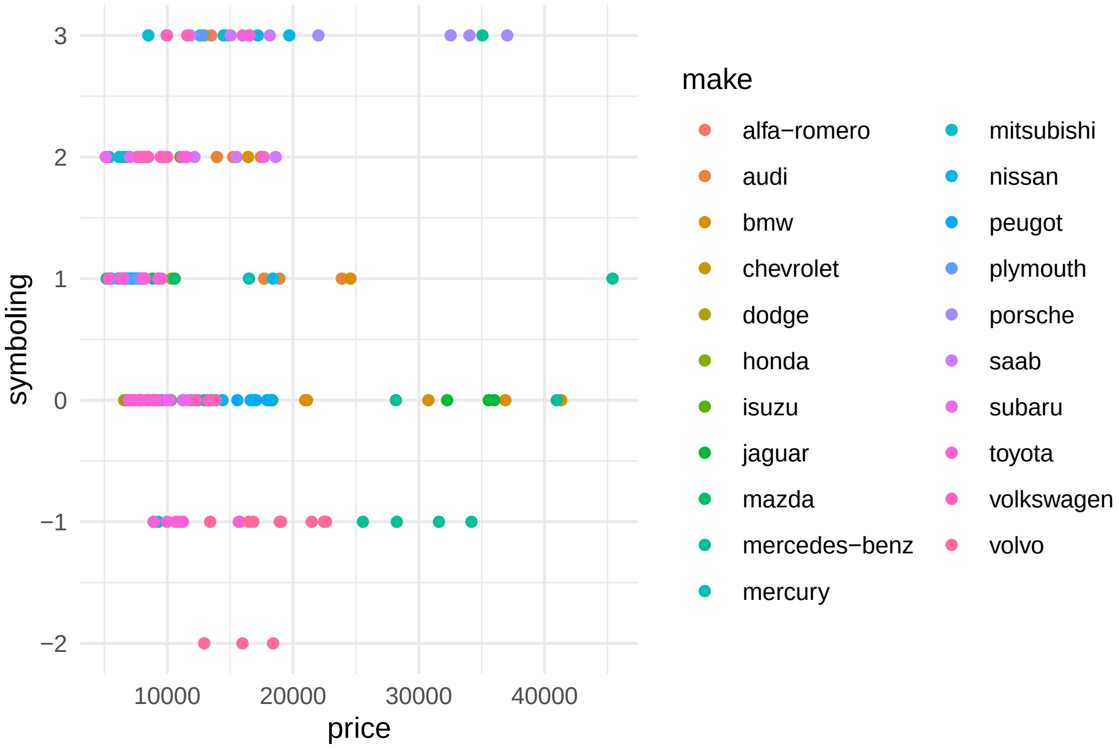

Figure 7: Relationship Between Price and Symboling with manufacturer

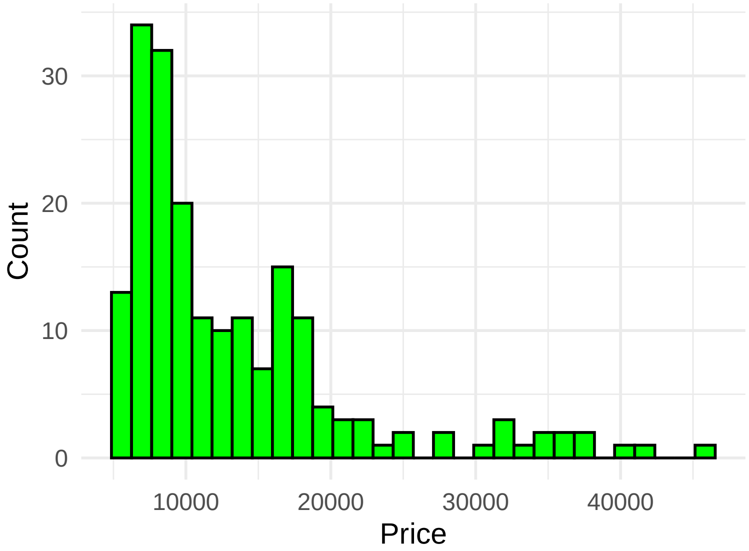

Figure 8: Histogram of Price

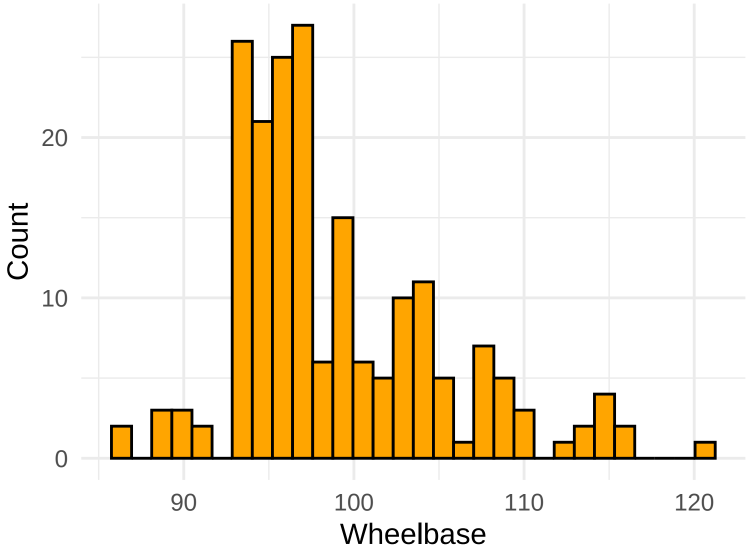

Figure 9: Histogram of Wheeelbase

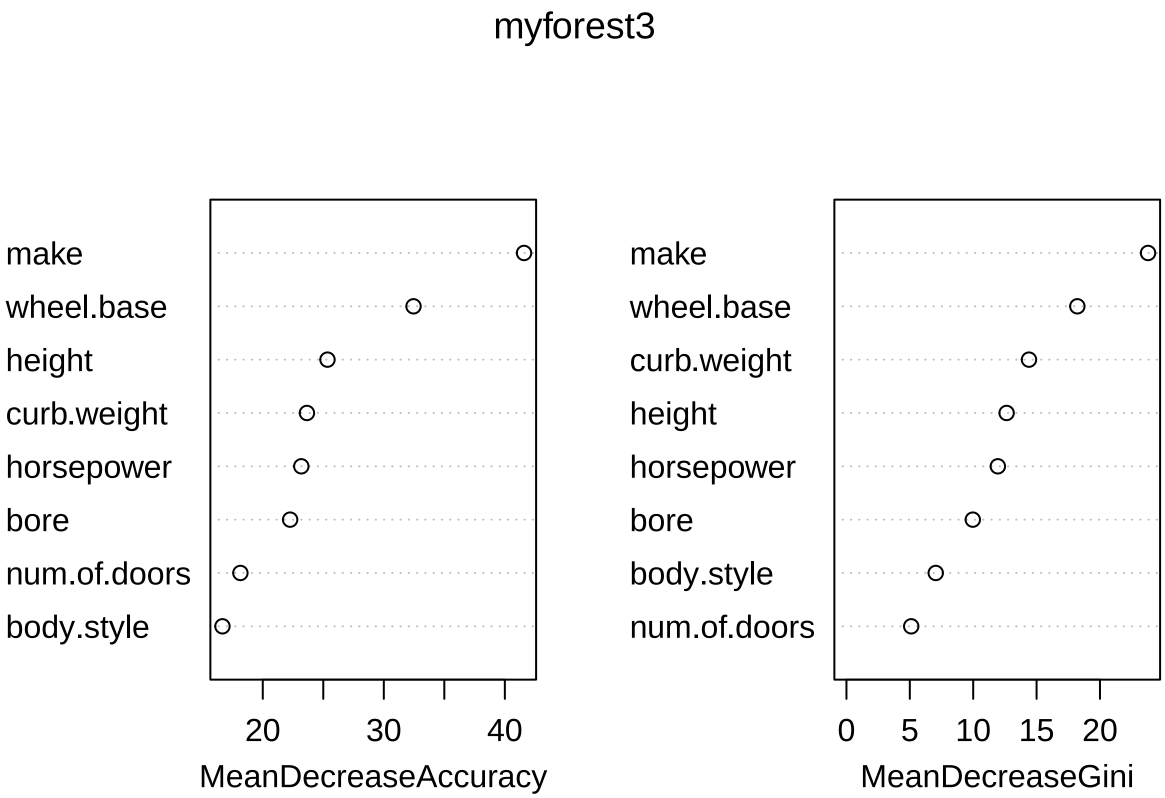

Figure 10: Mean Decrease Accuracy and Mean Decrease Gini of the Classification Model