Unveiling Automobile Safety Risks:

The Role of Vehicle Characteristics

Arial Huang

McGill University

Desautels Faculty of Management

Master of Management in Analytics

Fall 2024 - MGSC-661

December 6th, 2024

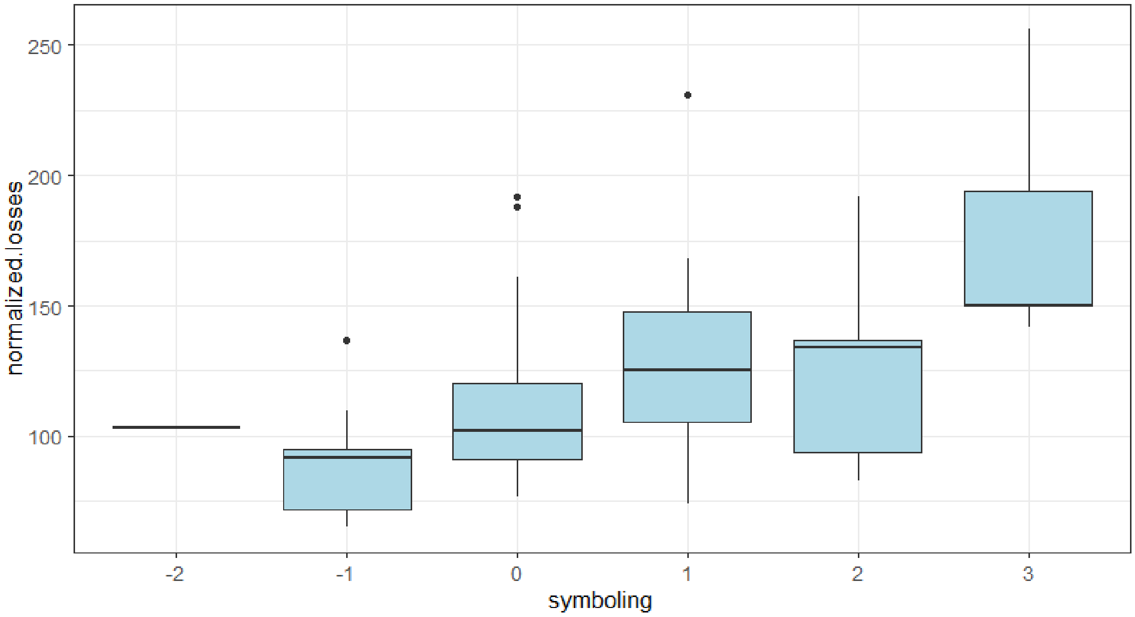
Introduction

Automobiles are an indispensable means of transportation for most people. When selecting a car, factors such as appearance and performance often take precedence; however, safety is undeniably a critical consideration. This report aims to analyze how various features, components, and structural aspects of automobiles influence their safety. Through this exploration, a predictive model will be developed to assess the risk level of a vehicle based on its characteristics, providing valuable insights into the elements that contribute to safer vehicle design and operation.

Data Description

This analysis utilizes a dataset originally collected for insurance purposes. The dataset is composed of 26 columns which can be broadly categorized into four groups: risk indicators, specifications, performance indicators, and manufacturer and price information. The risk indicators will serve as the target variable, while the remaining categories will be used as predictors.

1. Risk indicators: symboling and normalized loss.



In this dataset, vehicles assessed as high-risk are assigned higher "symboling" values. Additionally, the "normalized loss" variable represents the relative loss of a vehicle in an accident and can be interpreted as a safety indicator: the higher the loss, the less safe the vehicle is implied to be. As shown in the figure, the "symboling" ratings appear to be reasonably accurate, as vehicles with higher "symboling" values tend to have higher losses. Therefore, "symboling" will be used as the target variable for this analysis.

2. Specifications: Categorical predictors include fuel type, aspiration, number of doors, body style, drive wheels, engine location, engine type, and fuel system. Numerical predictors include wheelbase, length, width, height, curb weight, number of cylinders, engine size, bore, and stroke.
3. Performance indicators: compression ratio, horsepower, peak rpm, city mpg, and highway mpg. All numerical.
4. Manufacturer and price information: make (categorical) and price (numerical).

As anticipated, the given numerical features exhibit multicollinearity (see Appendix 1), as many of the functionalities are interrelated. For instance, curb weight, number of cylinders, and horsepower are highly correlated with engine size, as they all reflect various aspects of the engine's performance and physical characteristics; additionally, the engine's performance can influence fuel efficiency (mpg), which adds another layer to the multicollinearity issue. This interdependence among variables needs to be addressed to ensure robust analysis.

To address the multicollinearity issue while retaining valuable information, several indicators are derived from the provided features. These indicators not only help mitigate multicollinearity but also offer valuable insights into risk evaluation. These include:

1. Base area: Calculated from the vehicle's width and length, reflecting its footprint and potential stability.
2. Wheelbase-to-length ratio: A measure of proportionality that may influence handling and safety.
3. Engine size ratio: The ratio of engine size to vehicle weight, potentially linked to performance and control.
4. Average mpg: The mean of city mpg and highway mpg.

By adding these additional predictors and removing some of the original ones, the selected features show much lower correlation with each other (see Appendix 2), while ensuring that the critical information from the original dataset is not lost.

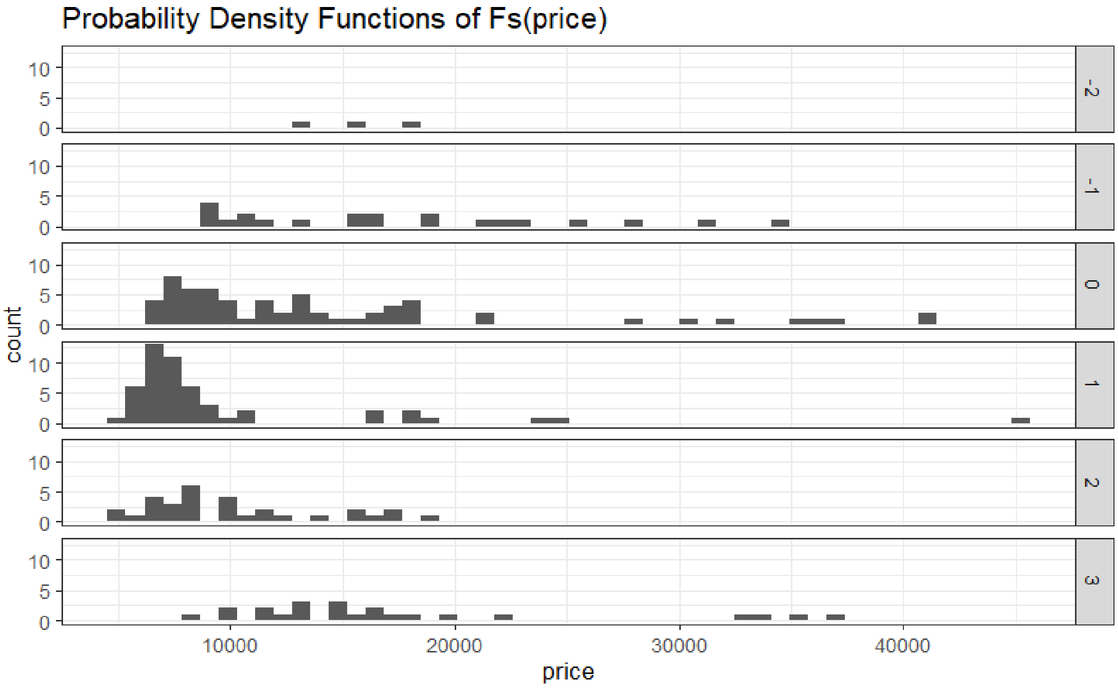
Model Selection & Methodology

For the following analysis, I will focus on linear discriminant analysis and develop a model using the random forest algorithm. Given the relatively small size of the dataset, with only 205 incidents, applying more complex approaches and models, such as quadratic discriminant analysis and boosting, may lead to overfitting. Therefore, linear discriminant analysis and random forest are more appropriate to ensure generalizability and avoid overfitting.

1. Linear Discriminant Analysis

a. Assumption

Take one of the predictors "price" as an example (see Appendix 3 for other numerical predictors). As shown in the following graph, it's reasonable to assume the price of each risk level (i.e., symboling) is normally distributed with different means but with the same standard deviation.



b. Analysis

Feature importance was evaluated by ranking the mean absolute coefficient values across the five linear discriminants (see Appendix 4). The wheelbase-to-length ratio emerged as the most influential feature, with high coefficients across multiple discriminants, highlighting its critical role in distinguishing between classes. Additionally, features related to fuel and make emphasize the importance of brand-specific attributes in defining class boundaries. In contrast, features like base area, peak rpm, and price exhibited minimal coefficients, suggesting their limited contribution to class separation. The weaker impact of these features may be due to low variability or reduced correlation with the target variable. Consequently, the subsequent random forest will focus on more influential predictors, eliminating base area, peak rpm, and price, to improve performance and efficiency.

2. Random Forest

a. Features

After performing preliminary feature selection using LDA, the remaining features were used to develop a classification model with the random forest algorithm. As a tree-based model, random forest inherently accommodates correlations among predictors, making it robust to multicollinearity. Consequently, all predictors retained after the LDA filtering process, including those initially removed, were utilized regardless of their interdependence. This ensures the model leverages all relevant information for accurate classification.

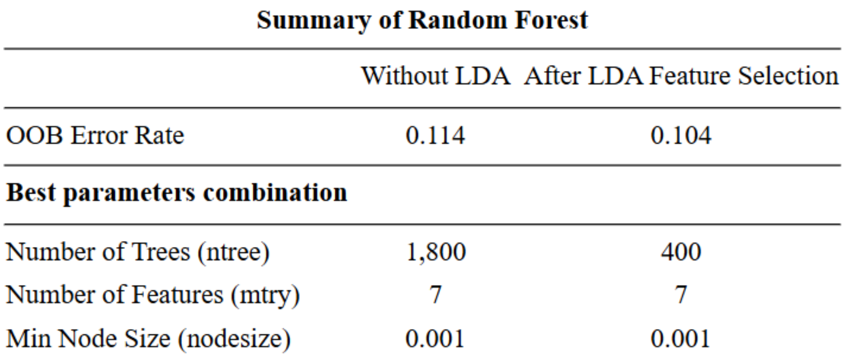
b. Parameters

i. classwt: Given the highly imbalanced distribution of risk levels in the dataset, it is crucial to adjust class weights to prevent the model from favoring majority classes due to their dominance. By employing an inverse frequency approach (see Appendix 5), higher weights are assigned to classes with fewer instances, compensating for their rarity and ensuring a more balanced representation in the model's predictions.

ii. ntree, mtry, nodesize: These parameters impact the complexity and diversity of the model. Proper tuning is essential to ensure the model generalizes well to unseen data while balancing performance with training time and computational efficiency. By carefully adjusting these parameters, an optimal trade-off between underfitting and overfitting can be achieved, enhancing the overall effectiveness of the model.

Results

1. Model Performance



Summary of Random Forest

	Without LDA	After LDA Feature Selection
OOB Error Rate	0.114	0.104
Best parameters combination		
Number of Trees (ntree)	1,800	400
Number of Features (mtry)	7	7
Min Node Size (nodesize)	0.001	0.001

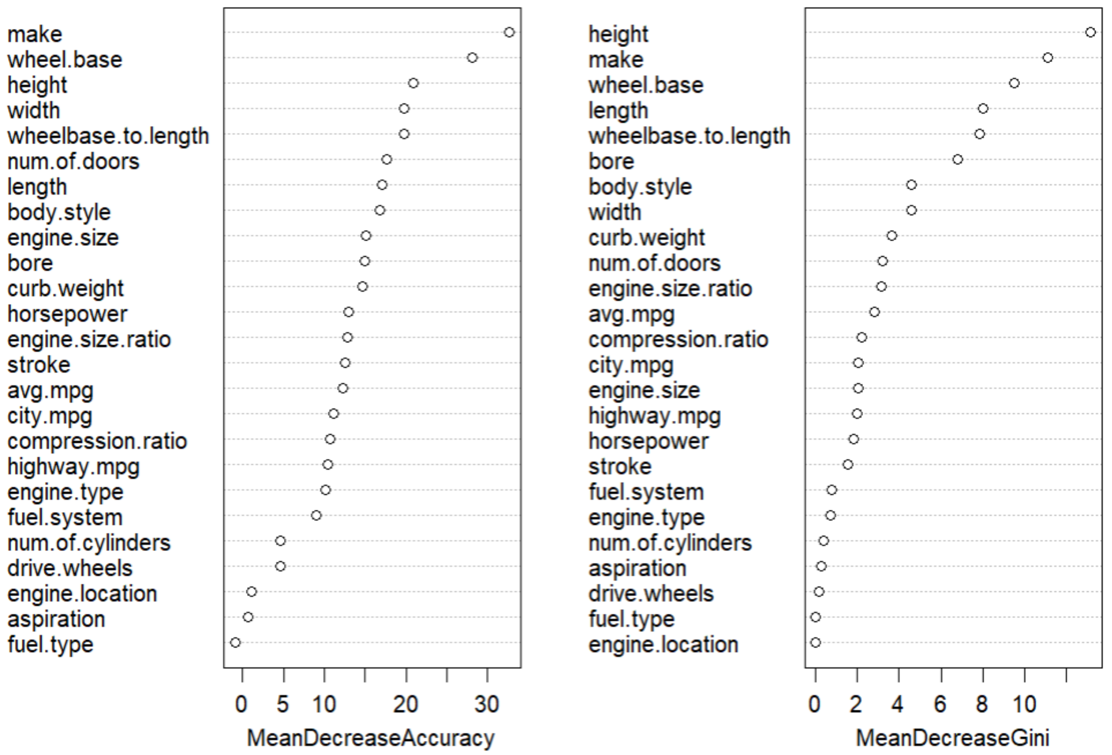
Compared to the model trained using all features without selection through LDA, the performance of the final model is improved, achieving an Out-of-Bag (OOB) error rate of 10.4% (see Appendix 6 for confusion matrix of the prediction). This reduction in error proves the effectiveness of feature selection, as it enables the model to focus on the most relevant predictors.

2. Feature Importance

The feature importance of the random forest model was assessed using two metrics: "Mean Decrease Accuracy" and "Mean Decrease Gini." These metrics measure the contribution of each predictor to the overall model performance.

a. Mean Decrease Accuracy: This metric evaluates the importance of each feature by measuring the reduction in the model's predictive accuracy when the values of that feature are permuted.

b. Mean Decrease Gini: This metric measures the contribution of a feature to the homogeneity of nodes and leaves in the decision trees. A higher value indicates a stronger discriminatory power.



Analyzing these metrics reveals several key insights into the interplay between vehicle characteristics and risk levels:

a. Manufacturer Dominance: The manufacturer of the automobile emerges as the most critical feature influencing risk. This finding underscores the role of brand reputation and design philosophy in vehicle safety. Premium brands may integrate advanced safety technologies or build vehicles with higher quality materials, resulting in lower risk ratings, while budget-oriented manufacturers may prioritize affordability over safety features (see Appendix 7 for detailed risk levels by brand).
b. Wheelbase's Role in Stability: Wheelbase and wheelbase-to-length ratio also significantly contribute to predictive accuracy. A longer wheelbase generally provides better stability and ride quality, especially at higher speeds or during sharp maneuvers. This association reinforces the importance of structural proportions in vehicle safety assessments.
c. Vehicle Dimensions and Risk: Features such as width, height, and length contribute strongly to the model, reflecting their relationship with collision dynamics. For instance, wider vehicles may have lower rollover tendencies, while vehicles with

larger dimensions can better absorb collision forces, potentially reducing injury severity.

d. Marginal Impact of Engine Characteristics: Engine size moderately contributes to risk prediction, as it often correlates with performance-related features like curb weight and horsepower, indirectly affecting vehicle dynamics. One the other hand, other engine-related features, such as engine location, aspiration, and fuel type, show minimal direct impact on safety. This underscores that structural and design attributes play a far greater role in vehicle safety than engine configurations.

Conclusion

This report demonstrates how vehicle characteristics and structural aspects can significantly influence safety, as assessed through predictive modeling. By addressing multicollinearity and optimizing feature selection, the analysis identified key predictors such as manufacturer, wheelbase, and vehicle dimensions that strongly correlate with risk levels. The combination of linear discriminant analysis and random forest models proved effective, achieving a notable reduction in error rates.

Key findings emphasize the critical role of brand design, structural stability, and proportionality in mitigating risks. On the other hand, engine characteristics, though relevant, play a relatively minor role compared to these factors. The methodology of this report not only highlights essential safety considerations but also provides a framework for evaluating risk through data-driven approaches.

These insights offer valuable guidance for consumers seeking informed choices and for manufacturers aiming to design safer vehicles, ultimately contributing to improved road safety.

1. For Consumers:

   a. When selecting vehicles, consider manufacturer reputation as a key indicator of safety performance.

   b. Pay attention to vehicle dimensions and proportions, particularly the wheelbase-to-length ratio.

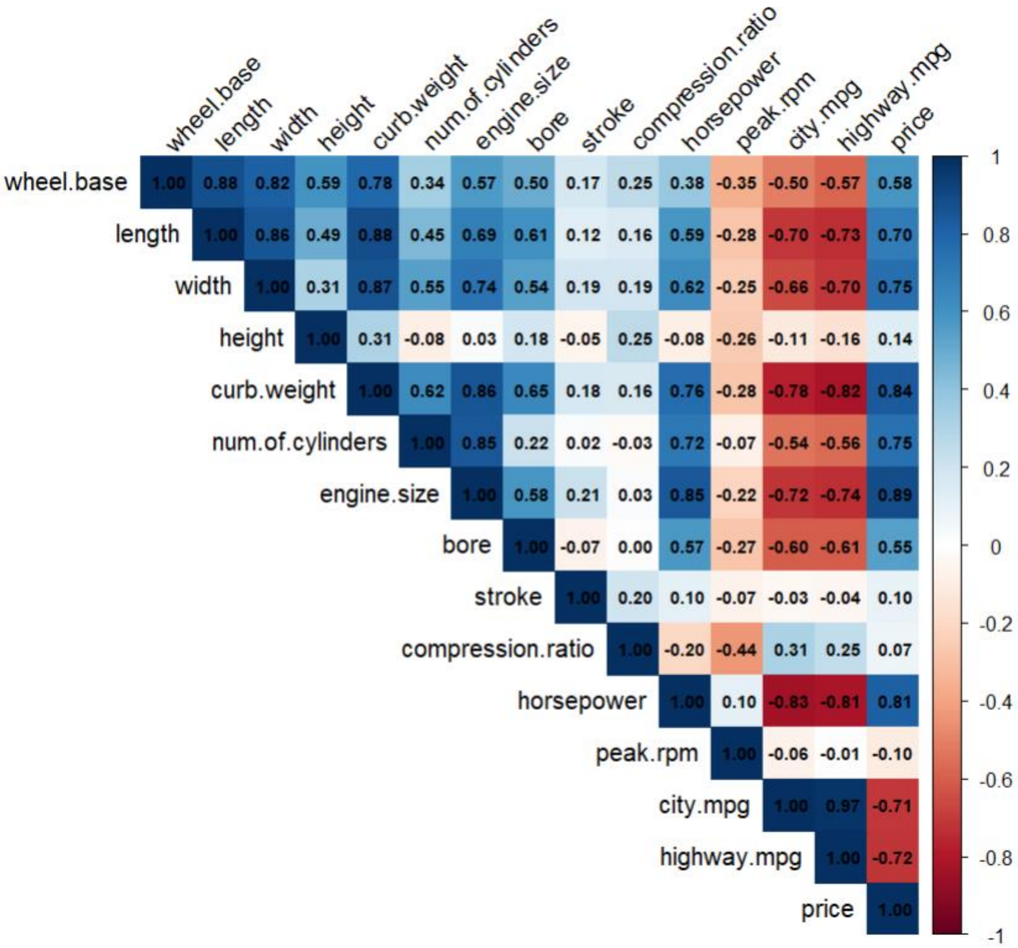
   c. Understand that powerful engines or advanced features may be less relevant to safety than fundamental structural characteristics.

2. For Manufacturers:

   a. Prioritize structural design optimization, particularly focusing on wheelbase-to-length ratios and vehicle dimensions that have been proven to enhance safety.
   b. Invest in maintaining and improving brand-specific safety standards, as manufacturer reputation significantly influences risk assessment.
   c. Consider safety implications during the early stages of vehicle design, as fundamental structural characteristics have been shown to be more important than performance features.

Future research can expand upon this work by incorporating a larger dataset and exploring advanced modeling techniques, such as boosting, to further refine risk assessment. Moreover, integrating external data sources, like real-world accident reports or advanced safety technology metrics, could enhance the robustness and applicability of the predictive model.

Appendix 1: Correlation Matrix & VIF Results of Given Features



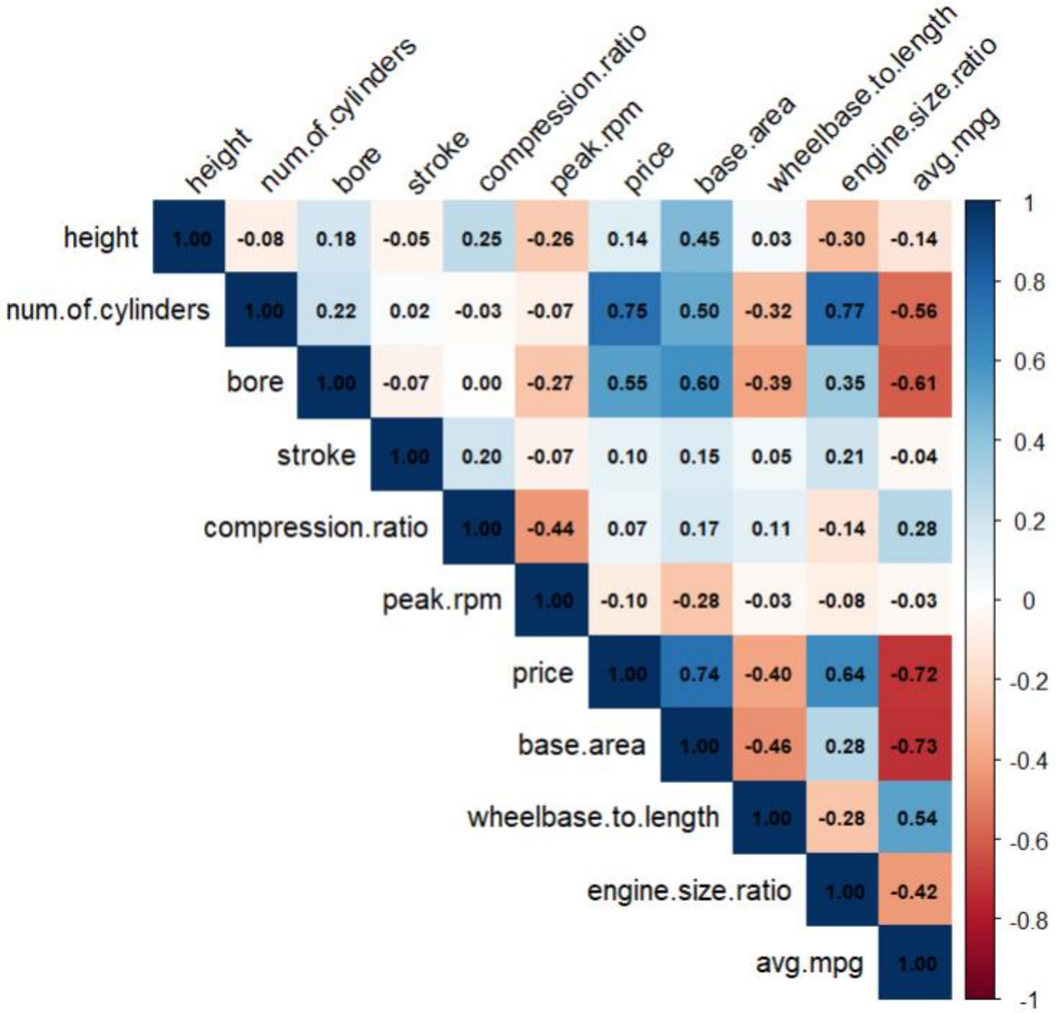
	wheel.base	length	width	height	curb.weight	num.of.cylinders	engine.size	bore	stroke	compression.ratio	horsepower	peak.rpm	city.mpg	highway.mpg	price
wheel.base	1.00	0.88	0.82	0.59	0.78	0.34	0.57	0.50	0.17	0.25	0.38	-0.35	-0.50	-0.57	0.58
length		1.00	0.86	0.49	0.88	0.45	0.69	0.61	0.12	0.16	0.59	-0.28	-0.70	-0.73	0.70
width			1.00	0.31	0.87	0.55	0.74	0.54	0.19	0.19	0.62	-0.25	-0.66	-0.70	0.75
height				1.00	0.31	-0.08	0.03	0.18	-0.05	0.25	-0.08	-0.26	-0.11	-0.16	0.14
curb.weight					1.00	0.62	0.86	0.65	0.18	0.16	0.76	-0.28	-0.78	-0.82	0.84
num.of.cylinders						1.00	0.85	0.22	0.02	-0.03	0.72	-0.07	-0.54	-0.56	0.75
engine.size							1.00	0.58	0.21	0.03	0.85	-0.22	-0.72	-0.74	0.89
bore								1.00	-0.07	0.00	0.57	-0.27	-0.60	-0.61	0.55
stroke									1.00	0.20	0.10	-0.07	-0.03	-0.04	0.10
compression.ratio										1.00	-0.20	-0.44	0.31	0.25	0.07
horsepower											1.00	0.10	-0.83	-0.81	0.81
peak.rpm												1.00	-0.06	-0.01	-0.10
city.mpg													1.00	0.97	-0.71
highway.mpg														1.00	-0.72
price															1.00

VIF

wheel.base 7.989	length 10.119	width 6.113	height 2.328
curb.weight 16.811	num.of.cylinders 11.232	engine.size 29.957	bore 4.566
stroke 2.100	compression.ratio 2.424	horsepower 9.564	peak.rpm 2.035
city.mpg 26.778	highway.mpg 24.967	price 7.144	

Appendix 2: Correlation Matrix & VIF Results

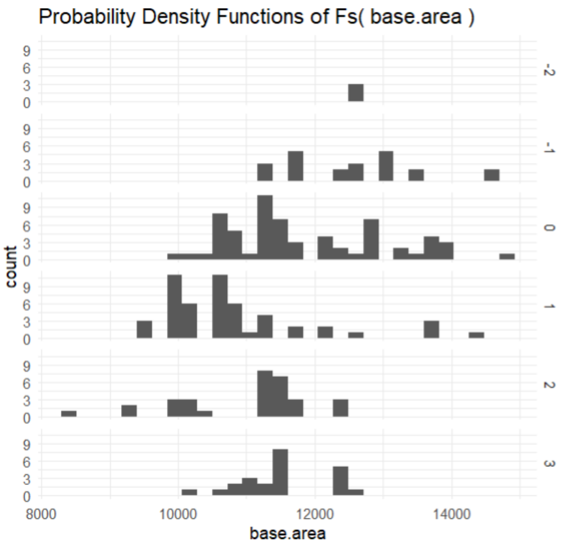
After Adding and Dropping Features

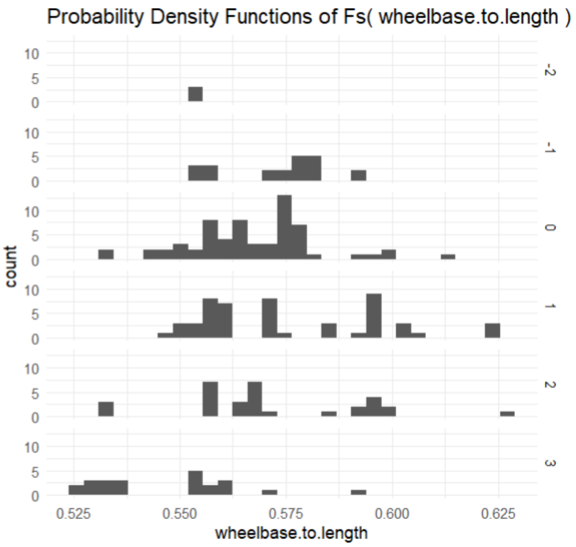


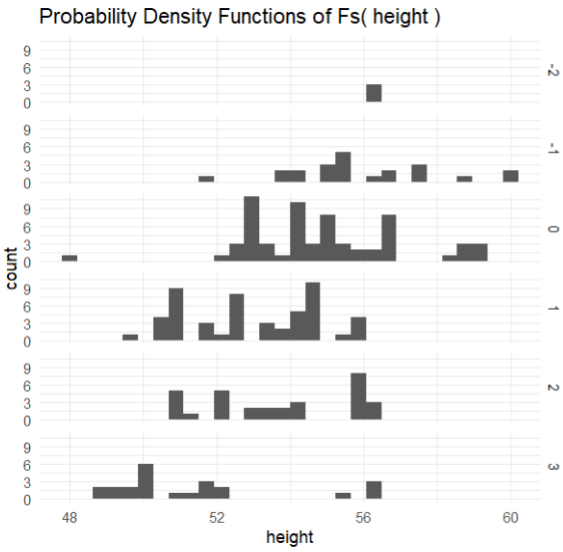
VIF

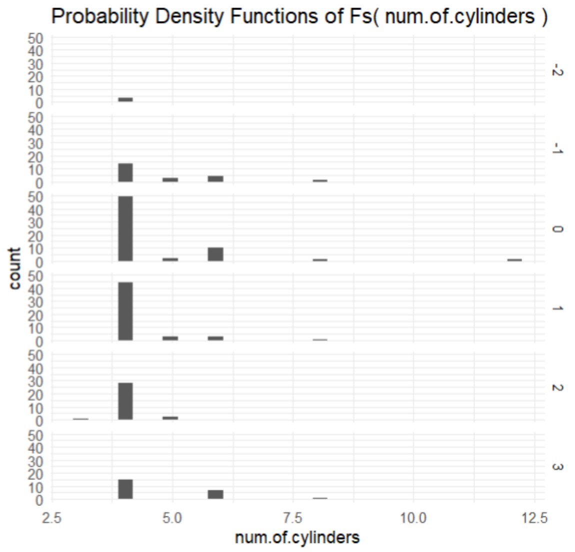
height 1.861	num.of.cylinders 6.777	bore 3.567	stroke 1.720
compression.ratio 2.193	peak.rpm 1.599	price 5.376	base.area 6.214
wheelbase.to.length 1.602	engine.size.ratio 6.189	avg.mpg 5.399	

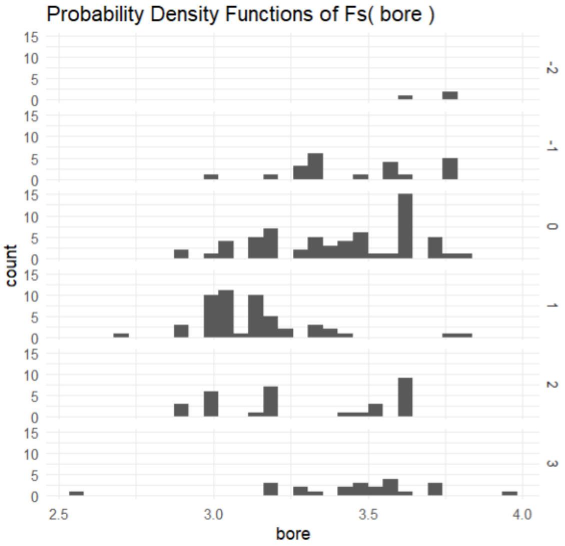
Appendix 3: Probability Density Functions of Numerical Predictors

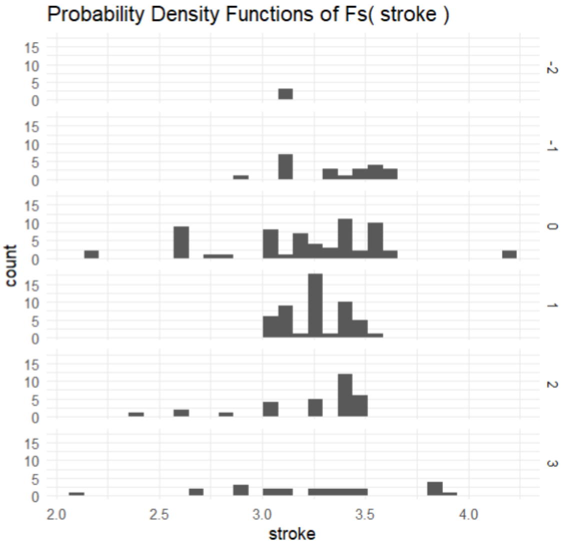


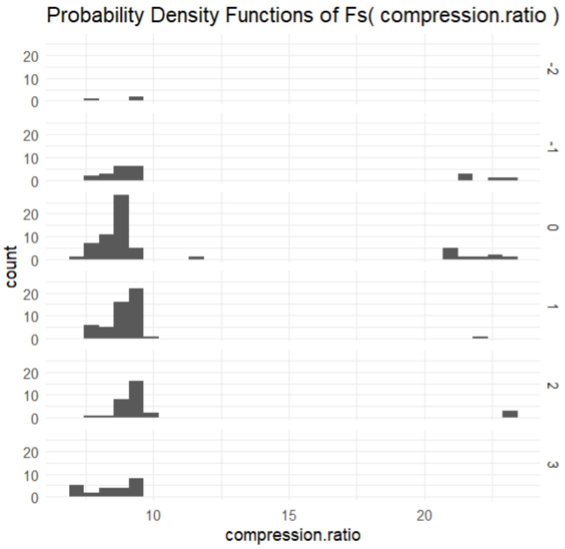


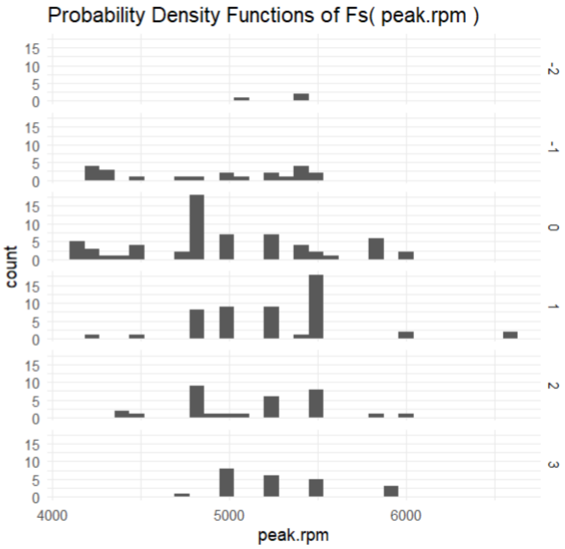




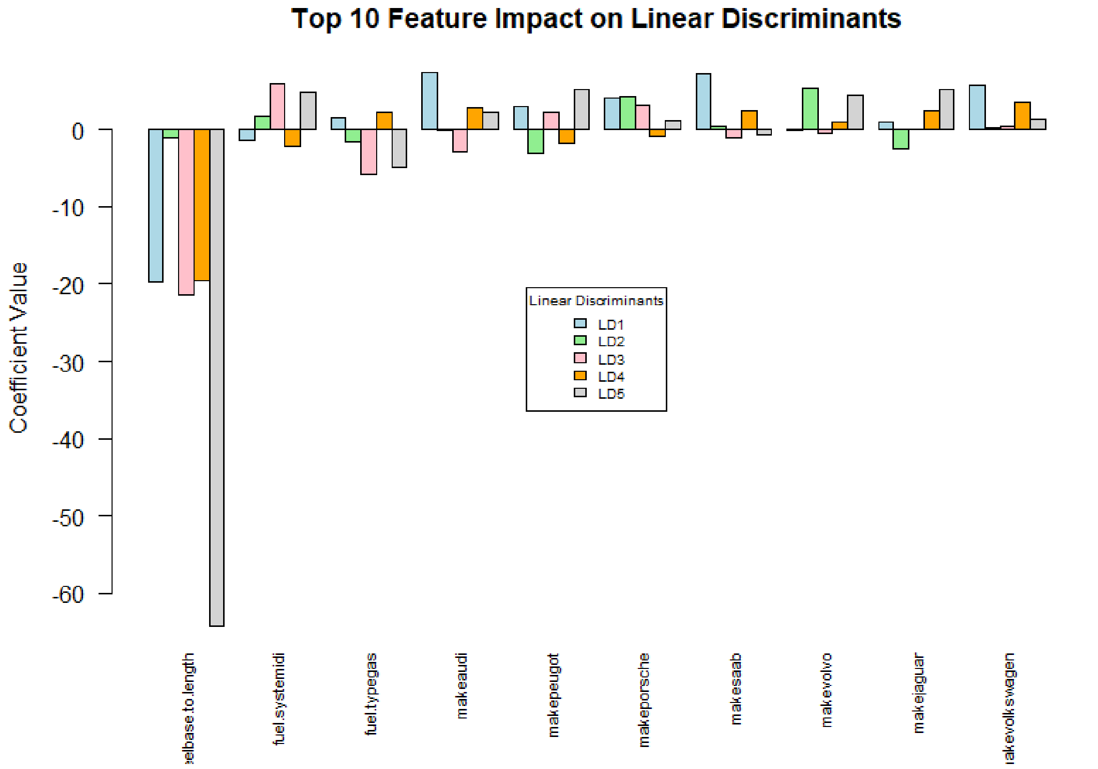

Probability Density Functions of Fs(bore)

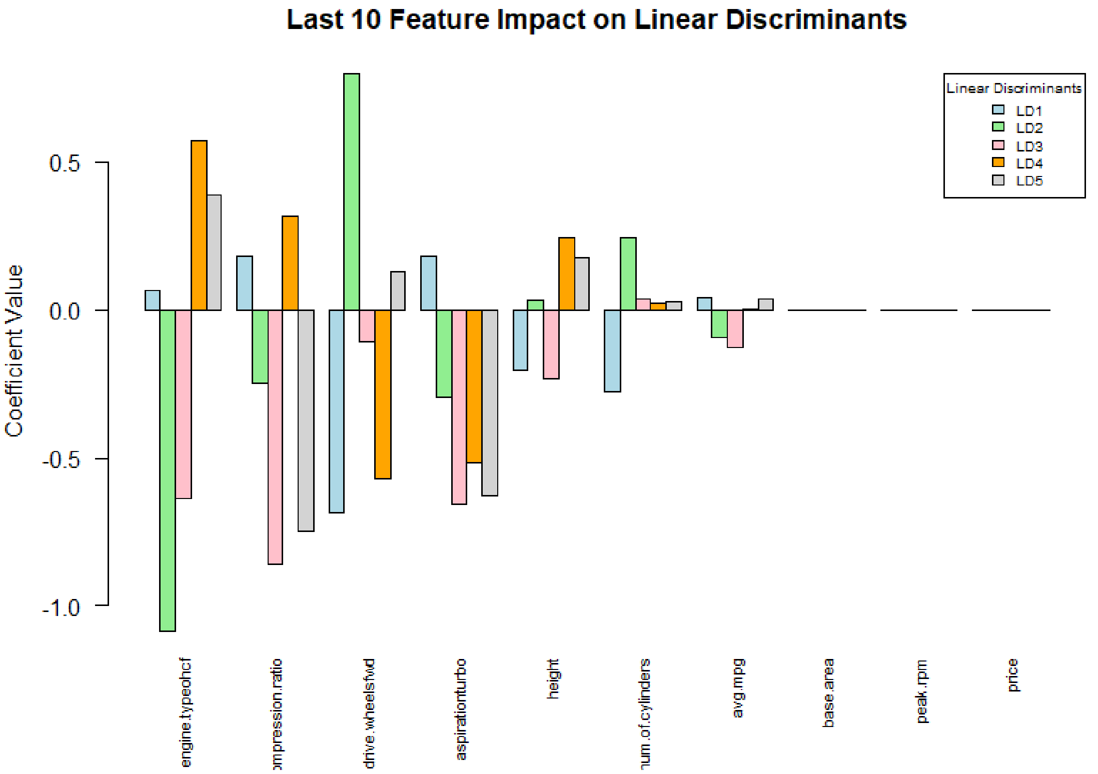

Probability Density Functions of Fs(stroke)


Probability Density Functions of Fs(compression.ratio)


Probability Density Functions of Fs(peak.rpm)

Appendix 4: Bar Plots of Coefficients for Top and Least Important Features


Top 10 Feature Impact on Linear Discriminants

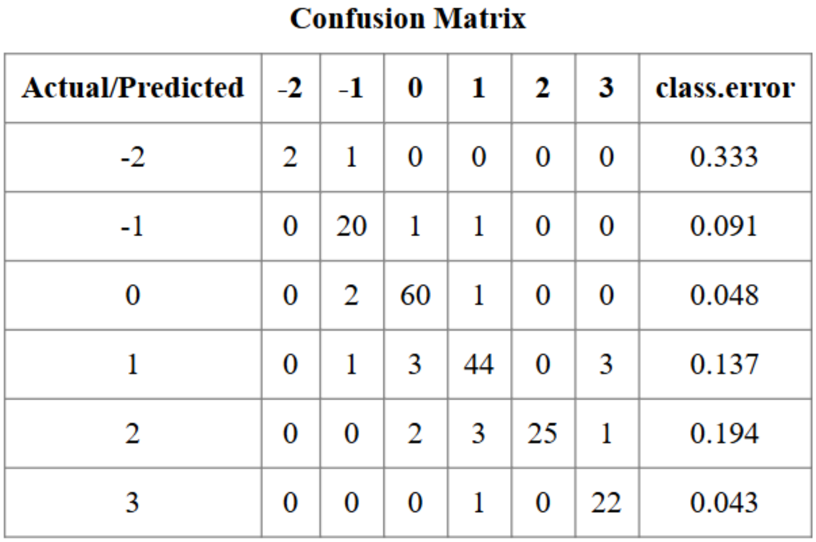

Last 10 Feature Impact on Linear Discriminants

Appendix 5: Inverse Frequency Approach

$$Weight_i = \frac{Total\ Observations}{Number\ of\ Classes\ \times Class\ Size_i}$$

I chose this approach because it ensures proportional and fair weighting across all classes, effectively accounting for differences in class distribution. This method avoids the need for complex resampling techniques while still addressing the imbalance in the dataset.

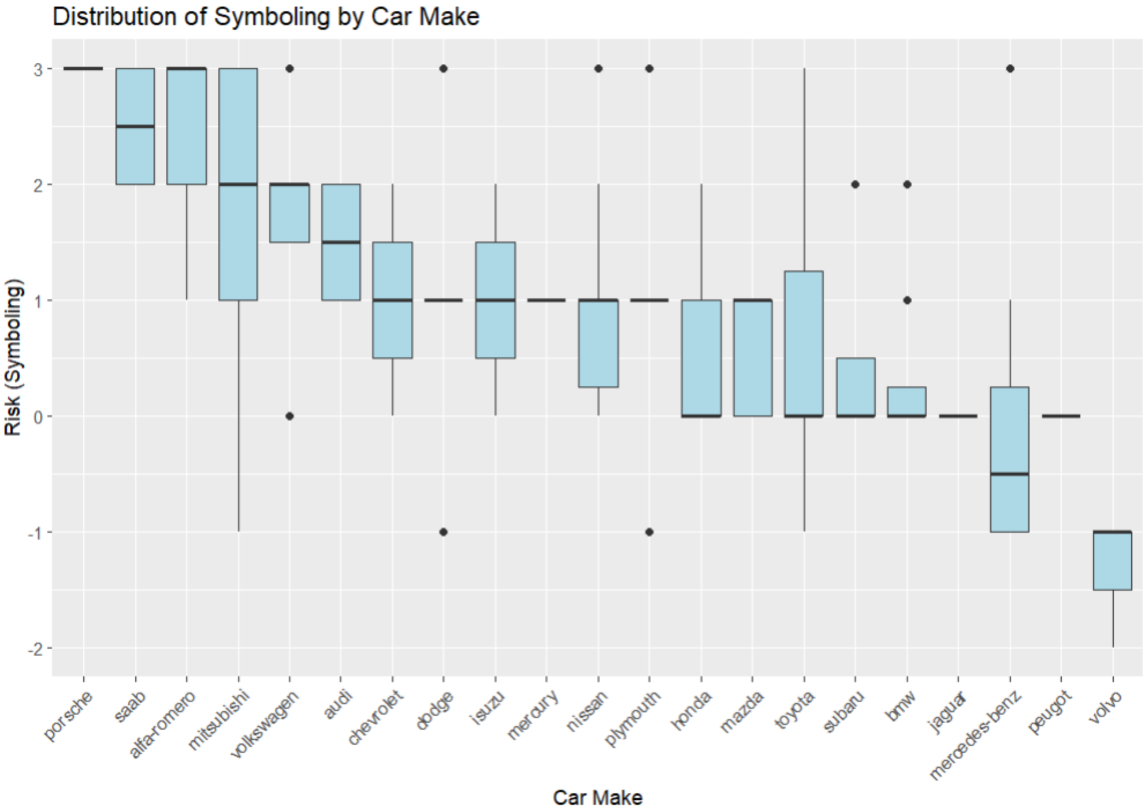
Appendix 6: Confusion Matrix of Random Forest Result



Confusion Matrix

Actual/Predicted	-2	-1	0	1	2	3	class.error
-2	2	1	0	0	0	0	0.333
-1	0	20	1	1	0	0	0.091
0	0	2	60	1	0	0	0.048
1	0	1	3	44	0	3	0.137
2	0	0	2	3	25	1	0.194
3	0	0	0	1	0	22	0.043

Appendix 7: Risk Levels by Brand



Code

Please refer to Automobile_Risk_Level_Classification.R.