PREDICTIVE MODELING FOR VEHICLE RISK PROFILING

Rohan Kumar

INTRODUCTION

This project aim to apply data science in the insurance domain, offering tangible benefits in terms of risk management and financial forecasting. It involves developing a predictive model for an insurance company with the aim of assessing the risk associated with new vehicles entering the market. The focal point of the assessment is the 'symboling' attribute, a risk indicator inherent to each car model, which ranges from +3 (signifying high risk) to -3 (indicating low risk). This indicator is pivotal for insurance companies as it influences insurance premium calculations. The development of a reliable risk prediction model stands as a testament to the power of machine learning in transforming industry practices and enhancing decision-making processes.

DATA DESCRIPTION

The dataset under analysis offers a comprehensive view of automotive specifications and their associated risk ratings. Our exploratory analysis revealed several intriguing patterns and relationships within the data. (Note: A description of each variable is present in figure 1 of the appendix.)

Distribution of Symboling Ratings: (figure 2) The distribution of symboling ratings across the dataset is skewed towards the 0 to 1 range, suggesting a moderate risk level for the majority of vehicles. Fewer cars are rated as highly safe (-2 and -1) or highly risky (2 and 3), indicating that extreme ratings are less common in this dataset.

Engine Size by Symboling Rating: (figure 3) A closer look at the engine sizes reveals a discernible trend where cars with higher risk ratings (symboling 2 and 3) tend to have larger engine sizes, while those with lower risk ratings (-2 and -1) generally have smaller engines. This suggests that engine size may be a factor in determining the perceived risk associated with a vehicle.

Average Symboling Score by Car Make: (figure 4) When we consider the average symboling score by car make, we notice that certain manufacturers like Porsche and Saab have higher average symboling scores, implying a tendency towards higher risk ratings. Conversely, makes such as Volvo and Mercesdes have lower average scores, suggesting these are perceived as safer options. This provides valuable insights into how different brands are positioned in terms of risk assessment.

Correlation Heatmap: (figure 5) The correlation heatmap unveils several noteworthy correlations amongst the numeric variables. For instance, there is a strong positive correlation between engine size and curb weight, indicating that heavier vehicles typically have larger engines. On the other hand, negative correlations are observed between highway miles per gallon (mpg) and engine size, as well as city mpg and curb weight, suggesting that fuel efficiency tends to decrease with increasing size and weight of the vehicle.

MODEL SELECTION & METHODOLOGY

To address the dataset's missing values, I employed a methodical approach, beginning with the 'bore' and 'stroke' attributes which are theoretically related by the following:

$$Engine\ Size = \pi \times \left(\frac{Bore}{2}\right)^2 \times Stroke \times Number\ of\ Cylinders$$

I used regression modeling to estimate the 'bore' values using 'engine size' as the predictor, effectively imputing the absent data. A similar approach was taken for the 'stroke' attribute, ensuring consistency in our methodology. Through research it was revealed that 'peak-rpm' and 'engine type' are positively related, with sports cars generally displaying higher peak rpms compared to sedans. This correlation allowed me to impute missing 'peak-rpm' values by computing and using the mean values for each 'engine type'.The 'number of doors' proved to be closely linked with 'body style', prompting me to apply the mode of each 'body style' category for imputation, which mirrored the logical structure of car design—typically, convertibles feature two doors and sedans four.

For the 'normalized losses', I opted for median values to replace the missing data, which mitigated the effect of potential outliers. When confronting missing price information, I resorted to calculating the average price for each 'make', aligning the imputed values with the brand's market valuation. Completing the data cleaning phase, I converted the 'number of cylinders' and 'doors', initially recorded in textual format, into numeric form, thereby streamlining subsequent analytical processes and maintaining data uniformity.

Following the data imputation process, I investigated multicollinearity using the Variance Inflation Factor (VIF) (Results in figure 6). This assessment was carried out by conducting a linear regression analysis, which served solely to identify correlations, not for the development of a predictive model.

MODEL SELECTION & METHODOLOGY

I set a VIF threshold of 10, beyond which variables were considered highly collinear and thus candidates for removal. Based on this criterion, the variables 'length', 'curb_weight', 'engine_type', and 'highway_mpg' were excluded from further analysis. 'Number of cylinders' initially exhibited a VIF greater than 10; however, it was retained because its VIF decreased significantly after the removal of the aforementioned variables. Subsequently, I applied a random forest algorithm to assess feature importance (figure 7 and 8). This analysis led to the exclusion of ‘aspiration’ that showed minimal contribution to the model.

Given that my target variable, 'symboling', is categorical with multiple classes to predict, Linear Discriminant Analysis (LDA) was the chosen method. To facilitate LDA, non-numerical variables were converted into categorical form. Then the LDA model was ran and it achieved an accuracy of 0.7 (Results of the model in figure 9). I also tried out Quadratic Discriminant Analysis (QDA) but I was unsuccessful due to the underrepresentation of the '-1' category in 'symboling'.

I also explored clustering techniques, applying both hierarchical (figure 10) and K-means clustering to the dataset. To prepare for clustering, I converted categorical variables into dummy variables. I then evaluated the optimal number of clusters by analyzing cluster ranges from 2 to 10 and employing the elbow method (figure 11). The analysis indicated that the optimal number of clusters was 7 from K-means, leading to the extraction of centroids for these 7 clusters. (figure 12)

RESULTS

Linear Discriminant Analysis (LDA) Results Summary

In our pursuit to develop a robust model that predicts the insurance risk rating (symboling) for vehicles, a Linear Discriminant Analysis (LDA) was conducted. The analysis yielded an overall accuracy of 70%, significantly higher than the No Information Rate of 33.33%, which is the baseline accuracy that would be achieved by always predicting the most frequent class. This notable improvement in accuracy (p-value < 7.198e-09) suggests that the model is effective in distinguishing between different levels of risk as indicated by the symboling score.

The model displayed a varied range of sensitivity across the symboling classes, with the highest sensitivity observed for the most risky vehicles (symboling 3) at 87.5%, and the lowest for the safest category (symboling -2), which had zero instances correctly identified. This discrepancy could be due to the underrepresentation or absence of vehicles with a -2 symboling in the dataset. Specificity was consistently high across all classes, indicating the model's strong ability to identify true negatives.

The Positive Predictive Value (PPV), or precision, was also variable, with the highest PPV at 87.5% for vehicles with a symboling of 2, suggesting that when the model predicts a vehicle as having a symboling of 2, it is correct 87.5% of the time. The Negative Predictive Value (NPV) was high across the board, which means that the model is reliable in predicting the absence of a specific risk level.

Balanced accuracy, which accounts for both sensitivity and specificity, was also high, particularly for the higher risk categories (symboling 2 and 3), indicating the model's effectiveness at balancing both types of errors across unbalanced datasets.

In conclusion, the LDA model has demonstrated a strong capability in predicting the insurance risk rating for cars, with high specificity and balanced accuracy, particularly for vehicles at the higher end of the risk spectrum. This model can serve as a valuable tool for insurance companies to assess new vehicles' risk ratings and adjust their insurance premiums accordingly.

RESULTS

Clustering Results Summary

The centroids derived from clustering are the mean values of the variables for each cluster, representing the center of the clusters in the multidimensional space of the dataset. These centroids can be interpreted as the average profile of each cluster, which are pivotal to understanding the intrinsic groupings within the data. These centroids provide insight into various groupings of cars based on their features and can be used to identify patterns relevant to risk assessment for insurance purposes.

Clusters with Higher Risk (Higher Symboling Score):

Cluster 1 & Cluster 6 both exhibit higher symboling scores, suggesting they represent riskier vehicles. Cluster 1 stands out with its high engine size and the presence of sports cars (indicated by the make Porsche), while Cluster 6 has a notable presence of Dodge vehicles and higher horsepower, which represents a different category of performance vehicles.

Clusters with Moderate Risk:

Cluster 2 & Cluster 7 are characterized by moderate symboling scores. Cluster 2 is unique with the exclusive presence of Peugeot vehicles and overall larger dimensions. In contrast, Cluster 7 has a broader range of makes including Nissan and Plymouth, indicating a more diverse mix of vehicles.

Clusters with Lower Risk (Lower Symboling Score):

Cluster 3 & Cluster 5 have lower average symboling, suggesting safer vehicle profiles. Cluster 3 is distinguished by the sole presence of Honda cars, which might be attributed to the brand's reputation for safety. Cluster 5, on the other hand, includes luxury vehicles from Mercedes-Benz and has high compression ratios, which could be associated with diesel engines known for lasting for longer time.

RESULTS

Generalized or Mixed Clusters:

Cluster 4 does not show a strong inclination towards high or low risk but has a significant representation from premium brands like Audi and BMW. It features average engine sizes and dimensions, positioning it as a cluster of potentially premium but not necessarily high-risk vehicles.

Each cluster represents a unique combination of vehicle attributes, and this differentiation allows for a nuanced approach to insurance risk assessment. By grouping similar clusters, insurance companies can create more tailored insurance packages and marketing campaigns that resonate with specific customer segments defined by these clusters.

PREDICTIONS AND CONCLUSIONS

In conclusion, the application of Linear Discriminant Analysis (LDA) and clustering techniques to the assessment of vehicle risk ratings offers profound implications for insurance companies. The predictive power of the LDA model serves as a strategic tool, enabling insurers to discern the risk associated with various vehicle attributes and effectively categorize them into appropriate risk brackets. This categorization aids in establishing a pricing strategy that reflects the intrinsic risk of each vehicle.

The clustering analysis complements this by uncovering distinct groups of vehicles with common risk characteristics, thus enabling insurers to segment their market more effectively. The actionable insights derived from the high-risk clusters (1 and 6) and the low-risk clusters (3 and 5) allow for the development of targeted insurance products, from premium policies catering to high-performance vehicles to value-focused offerings for safety-oriented customers.

Furthermore, the moderate-risk clusters (2 and 7) highlight the need for balanced insurance solutions that cater to a broad spectrum of consumers, who may prioritize a blend of safety, performance, and affordability. By leveraging the insights from both the LDA model and clustering, insurance companies can refine their risk assessment protocols, personalize their product offerings, and enhance their competitive edge in a market where risk differentiation is paramount.

Ultimately, the integration of these analytical models into the insurance sector's operational framework can lead to more accurate risk pricing, improved customer satisfaction through tailored policies, and a robust risk management strategy that aligns with the evolving dynamics of vehicle safety and performance.

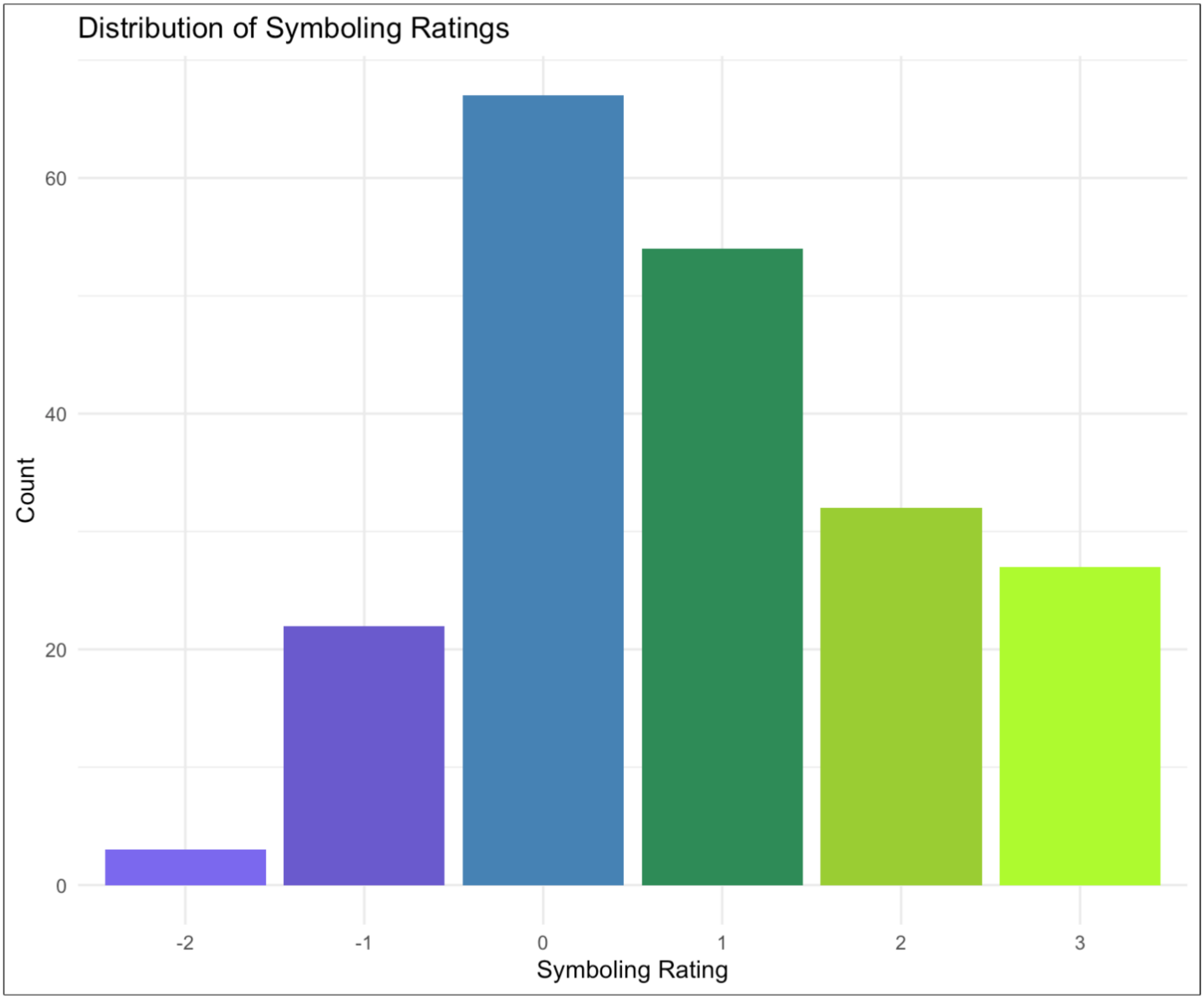
APPENDIX

Figure 1

No.	Variable Name	Description
1	Symboling	Insurance risk rating of the car: +3 indicates high risk, -3 indicates low risk.
2	Normalized Losses	Average loss payment per insured vehicle year (relative to other cars).
3	Make	Manufacturer brand of the car.
4	Fuel Type	Type of fuel the car uses (e.g., gas or diesel).
5	Aspiration	Type of engine aspiration (standard or turbocharged).
6	Number of Doors	Number of doors on the car.
7	Body Style	Design and shape of the car (e.g., convertible, hatchback).
8	Drive Wheels	Type of drivetrain system (e.g., front-wheel drive, rear-wheel drive).
9	Engine Location	Location of the car engine (front or rear).
10	Wheel Base	Distance between the front and rear wheels.
11	Length	Length of the car.
12	Width	Width of the car.
13	Height	Height of the car from the ground.
14	Curb Weight	Weight of the car without passengers or cargo.
15	Engine Type	Type of engine (e.g., overhead camshaft, rotary).
16	Number of Cylinders	Number of cylinders in the car engine.
17	Engine Size	Volume inside the engine cylinders (measured in cubic centimeters).
18	Fuel System	System used to store and deliver fuel to the engine.
19	Bore	Diameter of each cylinder in the engine.
20	Stroke	Distance the piston travels within the cylinder.
21	Compression Ratio	Ratio of the volume of the combustion chamber from its largest capacity to its smallest capacity.
22	Horsepower	Unit of measurement for engine power.
23	Peak RPM	The maximum revolutions per minute of the engine.
24	City MPG	Miles per gallon the car can travel on city roads.
25	Highway MPG	Miles per gallon the car can travel on the highway.
26	Price	Listed price of the car.

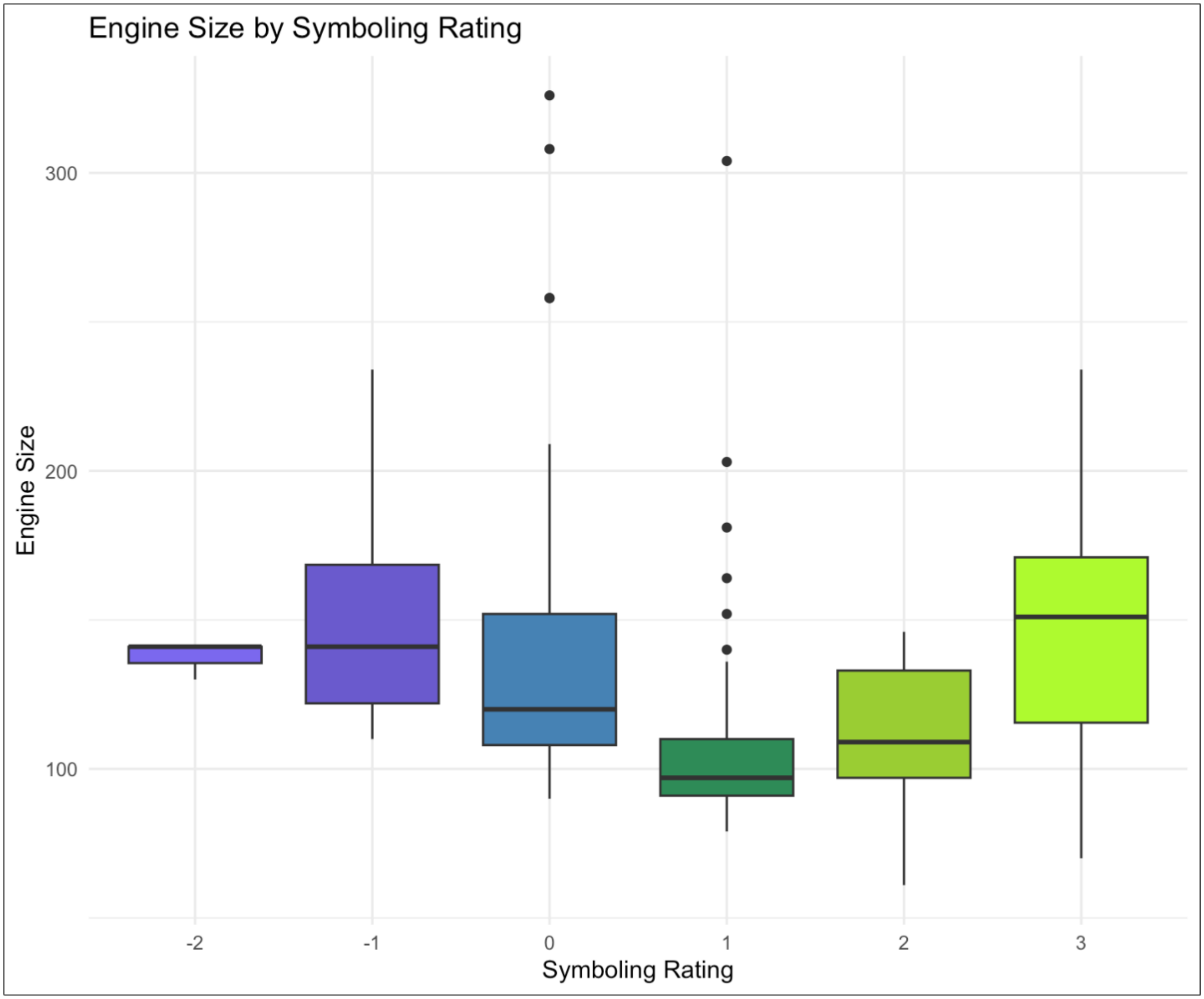
APPENDIX

Figure 2



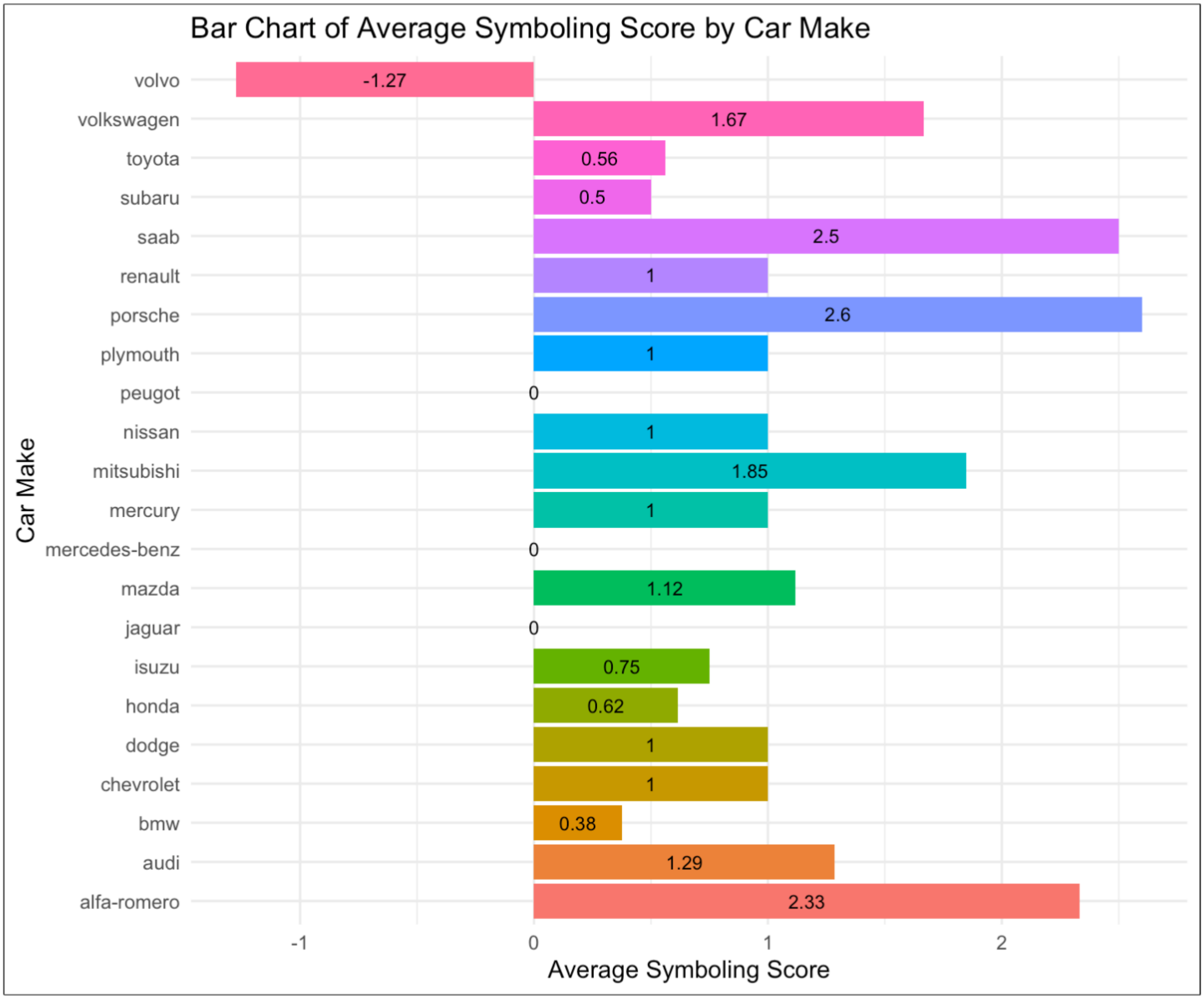
APPENDIX

Figure 3



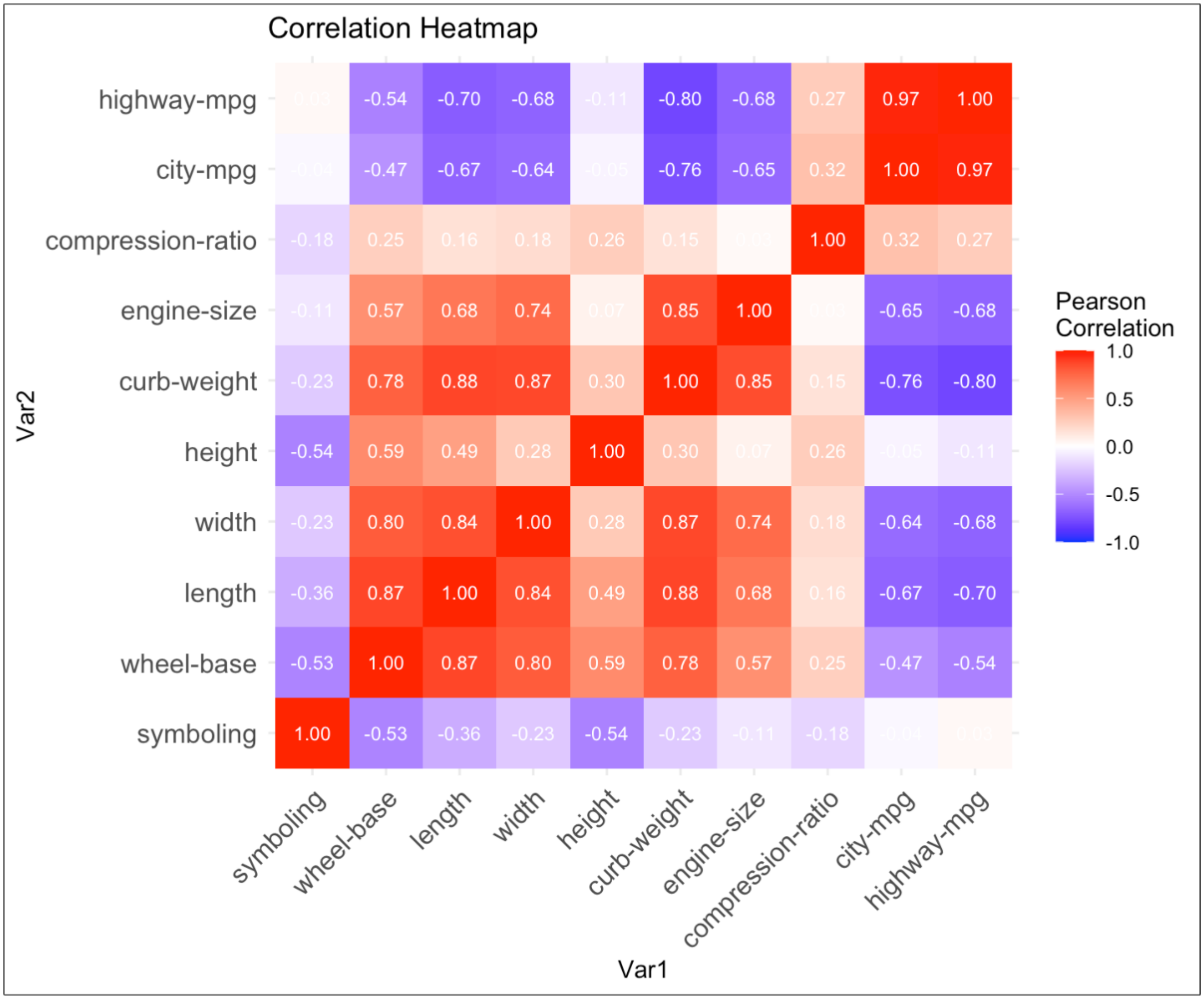
APPENDIX

Figure 4



APPENDIX

Figure 5



APPENDIX

Figure 6

Variable Name	VIF Score
Normalized Losses	1.482
Num of Doors	1.898
Wheel Base	7.673
Length	10.162
Width	6.168
Height	2.677
Curb Weight	17.313
Num of Cylinders	10.547
Engine Size	27.910
Bore	4.008
Stroke	1.984
Compression Ratio	2.335
Horsepower	9.820
Peak RPM	2.255
City MPG	29.109
Highway MPG	25.714
Price	7.160

Figure 7

Variable Name	IncMSE	IncNodePurity
Normalized Losses	21.541	30.402
Make	19.709	17.672
Fuel Type	0.718	0.177
Aspiration	0.411	0.526
Num of Doors	23.513	63.324
Body Style	14.294	27.901
Drive Wheels	7.840	3.881
Engine Location	1.541	0.063
Wheel Base	22.968	44.626
Width	14.952	12.734
Height	17.168	30.993
Num of Cylinders	3.167	1.239
Engine Size	13.469	9.190
Fuel System	6.478	2.488
Bore	15.735	11.336
Stroke	11.571	6.500
Compression Ratio	7.420	5.332
Horsepower	10.744	7.192
Peak RPM	11.643	6.805
City MPG	13.553	14.455
Price	13.991	12.458

Figure 8

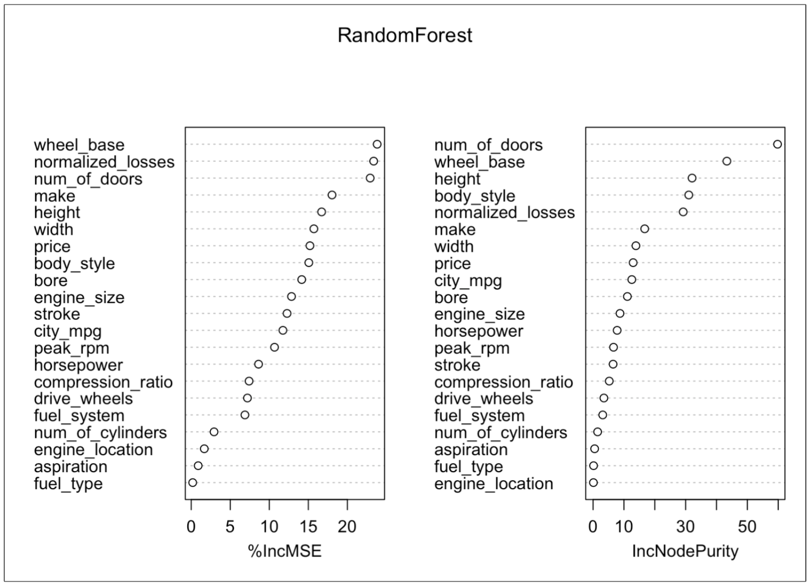


Figure 9

```
Confusion Matrix and Statistics

          Reference
Prediction -2 -1  0  1  2  3
        -2  0  0  0  0  0  0
        -1  1  2  2  0  0  0
        0   2  2 14  0  1  0
        1   0  0  2 12  1  1
        2   0  0  1  0  7  0
        3   0  0  1  4  0  7

Overall Statistics

               Accuracy : 0.7
                 95% CI : (0.5679, 0.8115)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 7.198e-09

                  Kappa : 0.611

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: -2 Class: -1 Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity               0.00   0.50000   0.7000   0.7500   0.7778   0.8750
Specificity               1.00   0.94643   0.8750   0.9091   0.9804   0.9038
Pos Pred Value             NaN   0.40000   0.7368   0.7500   0.8750   0.5833
Neg Pred Value            0.95   0.96364   0.8537   0.9091   0.9615   0.9792
Prevalence                0.05   0.06667   0.3333   0.2667   0.1500   0.1333
Detection Rate            0.00   0.03333   0.2333   0.2000   0.1167   0.1167
Detection Prevalence      0.00   0.08333   0.3167   0.2667   0.1333   0.2000
Balanced Accuracy         0.50   0.72321   0.7875   0.8295   0.8791   0.8894
```

APPENDIX

Figure 10

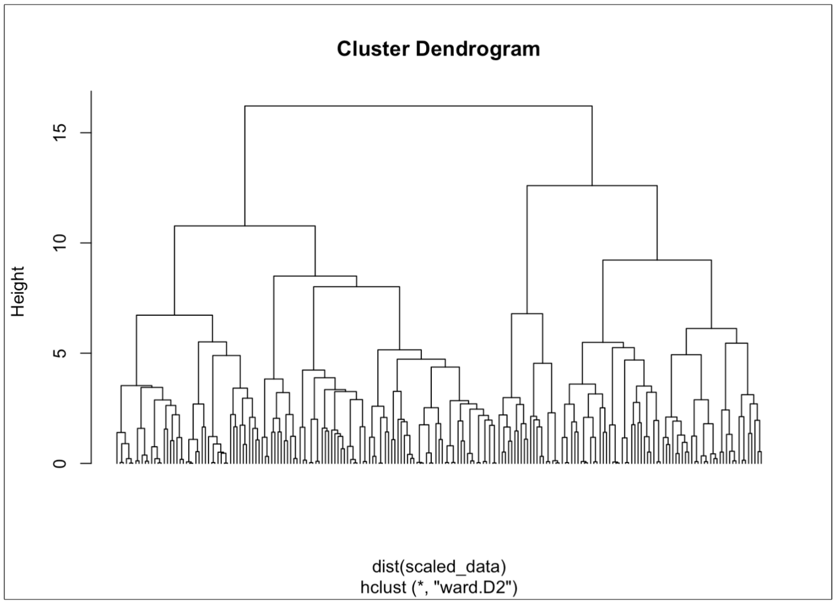


Figure 11

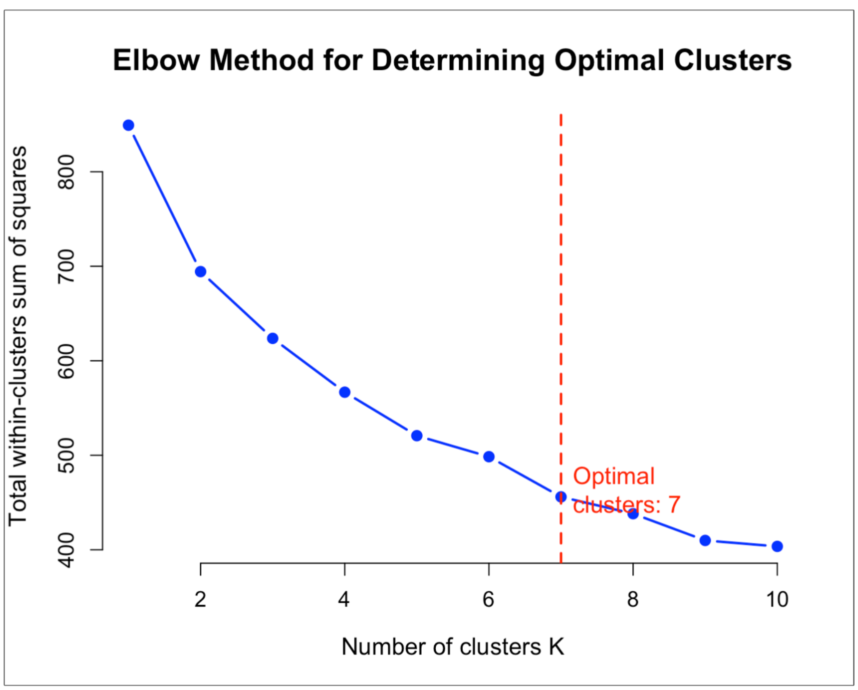


Figure 12

	symboling	normalized_losses	makealfa.romero	makeaudi	makebmw	makechevrolet	makedodge	makehonda	makeisuzu	makejaguar	makemazda	makemercedes.benz	makemercury	makemitsubishi	makenissan	makepeugot	makeplymouth	makeporsche	makerenault	makesaab
1	1	0.831	0.423	0.094	0.031	0	0	0	0.031	0.031	0.125	0.062	0.031	0	0.094	0	0.031	0.156	0	0
2	2	0.400	0.428	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	3	0.533	0.207	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	4	0.435	0.269	0.115	0.154	0	0.019	0.019	0	0.038	0.019	0.038	0	0	0.058	0	0	0	0	0.038
5	5	0.427	0.190	0	0	0	0	0	0	0	0.133	0.267	0	0	0.067	0	0	0	0	0
6	6	0.700	0.305	0	0	0.045	0.114	0	0	0	0.136	0	0	0.205	0.045	0	0.068	0	0.023	0.068
7	7	0.487	0.261	0	0	0.026	0.077	0	0.077	0	0.103	0	0	0.103	0.231	0	0.077	0	0	0.128

	makesubaru	maketoyota	makevolkswagen	makevolvo	fuel_typegas	aspirationturbo	num_of_doors	body_stylehardtop	body_stylehatchback	body_stylesedan	body_stylewagon	drive_wheelsfwd	drive_wheelsrwd
1	0	0.312	0	0	1	0.125	0	0.188	0.594	0.062	0	0	0.969
2	0	0	0	0	0.545	0.545	1	0	0	0.636	0.364	0	1
3	0	0	0	0	1	0	0.333	0	0.583	0.333	0.083	1	0
4	0.019	0.077	0.115	0.192	1	0.173	0.904	0	0	0.808	0.192	0.462	0.481
5	0	0.200	0.267	0.067	0	0.533	0.800	0.067	0.067	0.800	0.067	0.600	0.400
6	0.068	0.182	0.045	0	1	0.205	0.205	0	0.977	0	0	0.955	0.023
7	0.179	0.179	0	0	1	0.026	0.846	0.026	0	0.744	0.231	0.846	0.051

	engine_locationrear	wheel_base	length	width	height	curb_weight	engine_typedohcv	engine_typel	engine_typeohc	engine_typeohcf	engine_typeohcv	engine_typerotor
1	0.094	0.287	0.502	0.530	0.281	0.523	0.031	0	0.344	0.094	0.219	0.125
2	0	0.688	0.747	0.674	0.782	0.672	0	1	0	0	0	0
3	0	0.220	0.275	0.335	0.447	0.224	0	0	1	0	0	0
4	0	0.474	0.635	0.568	0.607	0.533	0	0	0.712	0.077	0.115	0
5	0	0.460	0.563	0.567	0.623	0.492	0	0	1	0	0	0
6	0	0.256	0.363	0.366	0.401	0.283	0	0.023	0.886	0.068	0	0
7	0	0.268	0.405	0.340	0.506	0.270	0	0	0.872	0.128	0	0

	num_of_cylinders	engine_size	fuel_system2bbl	fuel_system4bbl	fuel_systemidi	fuel_systemmfi	fuel_systemmpfi	fuel_systemspdi	fuel_systemspfi	bore	stroke	compression_ratio	horsepower	peak_rpm	city_mpg	highway_mpg	price
1	0.297	0.360	0	0.094	0	0	0.844	0.031	0.031	0.662	0.546	0.130	0.425	0.483	0.182	0.246	0.343
2	0.200	0.282	0	0	0.455	0	0.545	0	0	0.745	0.519	0.438	0.216	0.212	0.263	0.280	0.257
3	0.200	0.141	0.083	0	0	0	0	0	0	0.336	0.658	0.140	0.127	0.653	0.498	0.529	0.066
4	0.281	0.327	0	0	0	0	1	0	0	0.655	0.567	0.108	0.322	0.424	0.219	0.269	0.315
5	0.240	0.261	0	0	1	0	0	0	0	0.519	0.669	0.959	0.137	0.155	0.513	0.537	0.266
6	0.198	0.175	0.636	0	0	0.023	0.205	0.136	0	0.476	0.559	0.106	0.163	0.387	0.414	0.467	0.088
7	0.200	0.156	0.949	0	0	0	0	0.051	0	0.483	0.522	0.126	0.118	0.374	0.451	0.505	0.068

For better visibility, I have attached a HTML file through which you can view this table.