AUTOMOBILE PRICES RISK FACTOR ANALYSIS

Multivariate Statistics for Machine Learning
Master of Management in Analytics
McGill University

INTRODUCTION

In the intricate landscape of the automotive industry, dealerships struggle with a pivotal challenge: uncertainty about their clientele's preferences. The consequence often involves maintaining a diverse inventory, encompassing various cars to capture a broad spectrum of customer preferences[1]. This conflict between available vehicles and customer needs leads to inefficiencies in the purchasing process. Thus, the aim of this project is to help dealers decrease this gap of inefficiency and purchase vehicles that are most fitting to their specific customer characteristics. On the customer side, when purchasing a vehicle, the cost of insurance emerges as a paramount consideration. However, the absence of a systematic method to predict insurance ratings based on vehicle features complicates the decision-making process for dealerships when selecting suitable vehicles.

In response, my project attempts to relieve these challenges by developing a classifier that will assess various factors about vehicles and predict the insurance cost tranche that the vehicle will fall into. Secondly, the project employs a clustering algorithm to categorize cars into distinct groups, facilitating an understanding of each car's characteristics and aiding in the identification of models specific to different customer segments. The significance of this project for dealerships is profound. The classifier streamlines the purchasing process by guiding dealers toward vehicles aligned with customers' purchasing power and expectations regarding insurance ability and costs. The clustering algorithm improves the dealership's understanding of each car model's features, enabling tailored marketing strategies for specific customer segments.

DATA DESCRIPTION

For the project, I used a dataset from Kaggle[2] that includes vehicle specifications, normalized losses, and assigned insurance risk. The data is sourced from the 1985 Ward's Automotive Yearbook. Before initiating the model design, a comprehensive data exploration was conducted to understand the distribution of the independent and dependent variables.

The dependent variable is the vehicle's insurance risk score ('symboling') in comparison to its price, ranging from -3 (safest) to +3 (riskiest). I began by visualizing the distribution of the vehicle's risk score through a bar chart(1), which shows the distribution is slightly bell shaped; however, what's clear to see from the histogram is that there's a clear class imbalance, indicating that most of the cars in our dataset tend towards the neutral (0) to very risky (3) section of the risk scores, and no cars having a -3 score (very safe). This problem will be addressed and examined more comprehensively in the next section.

The dataset comprises a range of independent variables, encompassing both numerical and categorical ones. For numerical variables, histograms and boxplots were employed to provide insights into variable skewness, distributions, and potential outliers. For instance, the vehicle price variable exhibited a strong right skew(3), with an average price of $10,000. Only 25% of cars were priced over $16,500, highlighting exotic cars priced over $30,000 as outliers in the boxplot(2). These considerations were considered during model building and analysis. Concerning categorical variables, bar charts were utilized to showcase the distribution and skewness of different variables within each category. The visualizations unveiled an interesting finding: a couple of categorical variables had extreme situations of under-representation of certain categories. For example, the "vehicles fuel type" category consisted of two options—either gas, with 185 instances, or diesel, which was under-represented with only 20 instances.

The relationship between the predictors and target variable was explored; Key findings include:

- **Vehicle Make (Brand):** The brand significantly influences the variance in the insurance risk rating. For instance, Alpha Romeo cars averaged a risk rating of 3, indicating the riskiest category, while Toyota averaged a neutral rating (0). Brands associated with sports vehicles and higher maintenance costs tended to have higher risk.
- **Vehicle Aspiration Engine(4):** The risk score's interquartile range, maximum, and minimum were identical for cars with turbocharged or standard aspiration engines. The boxplots for both categories in Vehicle Aspiration were nearly identical, leading to the removal of this predictor due to its minimal predictive power.
- **Vehicle Price:** The influence of vehicle price on risk rating wasn't as strong as anticipated. A scatter plot of price and risk score(6) obtained a nearly linear line of best fit, while the correlation matrix(5) revealed a negligible very weak positive correlation between them.

MODEL SELECTION & METHODOLOGY

The model selection process began by assessing the usability of each predictor. First, the "normalized losses" column was removed as it contained 20% missing data. Eliminating these rows was avoided to maintain the dataset's potential, given its limited size of around 200 rows. Other than that, missing data was minor and rows containing NAs were removed. Additionally, categorical variables were transformed into binary variables for two-category variables, while those with multiple categories were dummified as factors. Numeric categories like "number of doors" and "number of cylinders," originally represented as number names, were converted to integers. Furthermore, the target variable, "symboling," exhibited class imbalance, with ratings of -1 to 3 having sufficient observations, while -2 had only three instances, and -3 had none. This imbalance was logically attributed to the nature of automobile insurance assessments, where extremely safe (-3 or -2) ratings are rare due to the high costs associated with any damage. Consequently, the decision was made to treat vehicles with a rating of -2 as outliers and remove them, focusing the model on predicting ratings from -1 to 3. This approach ensures the model's robustness and realism in predicting insurance ratings.

To ensure that model noise and biases were not generated, feature selection was conducted by examining the data for the existence of multicollinearity. Initially, VIF testing was conducted among numerical predictors, revealing evident collinearity issues. The correlation matrix[(5)] indicated a strong positive correlation between a vehicle's wheelbase length and its curbside weight. This correlation is logically explained by the fact that vehicles with larger wheels are typically higher, often SUVs or 4x4s, and these tend to be the heaviest. Similarly, horsepower and engine size exhibited a strong positive correlation, as higher horsepower engines with greater speed necessitate larger engines and batteries. By recognizing these relationships, I eliminated predictors contributing to data collinearity, reducing all VIF scores to under 4. For categorical variables, those exhibiting severe skewness or underrepresentation of certain classes were removed. For instance, since nearly all cars had their engines in the front, except for three vehicles with rear engine placement, the predictor "engine location" was removed. As a final step, one outlier point was detected and discarded using the Outlier test.

For the classification task, LDAs, Decision Trees, Random Forests, and Gradient Descent were assessed using a combination of training-testing set validation and 5-fold cross-validation on the cleaned dataset. Hyperparameter tuning was performed using cross-validation to ensure the optimal selection of hyperparameters. The models were evaluated using the accuracy metric, and the model achieving the highest accuracy was chosen as the best.

For the clustering task, as these algorithms are highly influenced by noise and outliers, further feature selection was implemented through Random Forest feature importance. The top 10 important features determining a vehicle's insurance risk rank were extracted from this feature importance ranking[7]. Focusing on these top 10 features enhances the clustering algorithm's ability to reveal patterns, creating a more comprehensive understanding of why vehicles receive specific risk ratings. These features include wheelbase, number of doors, price, length, peak RPM, and the brand of the car. Categorical variables like the brand were one-hot encoded to make them suitable for usage with K-Means. Additionally, numerical features were standardized/scaled since clustering algorithms are sensitive to predictor scales, thus ensuring an unbiased cluster representation. Subsequently, K-Means clustering was performed using a combination of the elbow method and the silhouette score to determine the optimal number of clusters (k) to be produced.

RESULTS

For the classification task, various models were assessed, starting with the decision tree. The complexity parameter (cp) was tuned by initially running an overfitted tree model with a very small cp, and then using R's plot cp feature to identify the cp resulting in the highest performing decision tree. Unfortunately, the decision tree emerged as the worst-performing model, achieving an accuracy score of 54.8% on the training-test split validation. Most misclassifications occurred in the neutrally ranked vehicles with a risk score of 0, experiencing a 55% misclassification rate in this class. This is a main drawback of the decision tree, being prone to overfitting and sensitive to training data distribution. To address these drawbacks, a random forest was employed, and the number of trees chosen was the one that plateaued the out-of-bag error scores[12]. The tuned random forest with 500 generated trees achieved a test accuracy score of 83.9%, using the same training-test split—a result significantly better than that of the decision tree.

One significant drawback of random forests is their lack of interpretability, as they are seen as black boxes. Thus, Linear Discriminant Analysis (LDA) was also tested to improve the interpretability of results. Using default LDA parameters, it achieved a test accuracy of 77.4%. Although LDA offers better interpretability, prioritizing the highest accuracy led to the selection of the random forest. Finally, the gradient boosting model was explored, known for its effectiveness in handling class imbalance and its high performance due to the sequential use of weak learners. Cross-validation was used to tune hyperparameters, resulting in an out-of-sample cross-validation accuracy of 82.1%. Comparatively, gradient boosting achieved an 80.6% accuracy on the same training-test split, outperforming the decision tree and LDA, but not the random forest. Thus, the selected model is the random forest, achieving the highest test accuracy score of 83.9%.

For the clustering task, the goal was to find the optimal number of clusters using the k-means algorithm. The elbow method suggested a k value of 4, where the graph plateaued[(8)]. Moreover, the silhouette score[(9)], which assesses point cohesion and cluster separation, yielded a reasonable score of 0.2 for 4 clusters, indicating a visually acceptable cluster representation[(11)].

MANAGERIAL IMPLICATIONS

In the context of dealerships, numerous misclassifications in insurance risk scores pose a significant risk and are highly undesirable. Assuming the dealer has conducted thorough market research, understanding the specific customer segment focus, their tolerance levels, and financial capacity to pay insurance on cars, the dealer should generally be aware of the insurance tranche/risk score that best suits their customers. Misclassification could lead the dealer to purchase a vehicle predicted to be in one risk tranche, only to discover it belongs to another. From the dealer's perspective, this represents a missed opportunity, as a correct prediction could have led to the purchase of a more suitable vehicle in the correct risk tranche, enhancing the likelihood of a successful sale.

On the inventory and purchasing side, the dealer now faces significant costs associated with acquiring the misclassified vehicle, along with ongoing monthly storage expenses. This scenario is particularly precarious due to the substantial costs involved. For instance, if the model predicts a vehicle to be rated 1 or 0 but is actually a 3, this presents a serious concern. A 3 rating

denotes the riskiest category, typically associated with high-end, expensive cars catering to a niche customer class, resulting in substantial inventory costs due to special storage needs. Therefore, even with lower accuracy, if the model can avoid such cases, it proves its business value.

The suggested random forest model, boasting an 83.9% average accuracy, proves highly proficient in correct predictions. Examining the confusion matrix(10) of the test set predictions reveal that most misclassification errors occur by predicting -1 or 1-rated cars as 0, a plausible outcome given their proximity. Achieving 80% accuracy in the 3-risk class is acceptable to dealers, either those targeting this segment, or those wanting to refrain from it. Consequently, such a model becomes a powerful tool for dealers, enabling data-driven, well-informed decisions in their car purchases and enhancing their attractiveness to their target customer segment.

To aid dealers in conducting more efficient market research and gaining a better understanding of their customer base, as well as identifying their financial ability, the clustering task proves to be very powerful. This task helps identify the following four car clusters and their respective customer bases:

Cluster 1: Cars in this cluster exhibit a mode risk ranking of 0, signifying that they generally fall on the neutral side of the insurance risk spectrum, having an average insurance price. This moderate insurance cost is highly appealing to a wide customer base. In addition to the attractive insurance price, this cluster displays the highest variability in car sizes. Typically featuring average-sized sedans and hatchbacks, this variety underscores versatility in customer choice, catering to different lifestyles. Cars in this cluster also boast the lowest average price among all clusters, averaging $7,500. The dominant brands in this cluster include Honda, Nissan, and Toyota—brands known for their affordability and appeal to a larger, more general customer base. This cluster targets young professionals and families looking for versatile, moderately priced cars, providing a balanced combination of safety, size, and practicality.

Cluster 2: Cars in this cluster have a mode risk ranking of 1, indicating that they generally fall on the slightly risky range of the spectrum, featuring a slightly higher than average insurance price. Typically, the cluster includes cars of average size, like those in Cluster 1, but is predominantly dominated by sedans. The cars in this cluster are priced slightly higher than those in Cluster 1, yet they are still on the more affordable side, with an average price of $10,100. This cluster caters to practical customers in search of reliable and fuel-efficient options. It is characterized by brands

such as Toyota, Subaru, and Volkswagen, appealing to individuals and families who prioritize everyday usability and value for money.

Cluster 3: Cars in this cluster exhibit a mode risk ranking of 0, positioning them as neutral in terms of risk compared to their price. Dominated by the largest sedan cars in terms of length and wheelbase size, these vehicles suggest a focus on spaciousness and comfort. Premium luxury brands such as BMW, Mercedes-Benz, and Volvo dominate this cluster, justifying the substantial price tags that go with these cars, averaging $22,700. Despite being the most expensive brands, the neutral risk rating can be attributed to the advanced safety features and technology that luxury vehicles often include. While these cars come with higher upfront costs, their emphasis on safety and comfort may offset potential risks associated with ownership. This cluster targets affluent individuals or families in search of both the safety and comfort provided by larger vehicles, as well as the premium driving experience associated with luxury vehicles.

Cluster 4: Cars in this cluster feature a mode risk ranking of 3, making them the riskiest among all cars and resulting in the highest insurance costs. Generally larger in size than regular sedans, this cluster is characterized by convertibles and hardtops. Despite an average price of $16,100, the luxurious and sportier nature of these vehicles contributes to the highest maintenance and repair costs among all cars and the least safety for the driver, thus justifying their elevated risk ranking. Highlighted by the sports lines of Porsche and Nissan, these cars appeal to individuals and enthusiasts looking for a luxurious and performance-oriented driving experience.

CONCLUSION

In summary, this project aimed to address challenges in the automotive industry by developing a classification model for predicting insurance risk ratings and implementing a clustering algorithm to categorize cars. The random forest algorithm emerged as the most proficient tool for risk predictions. The clustering algorithm identified four car clusters, enabling tailored marketing strategies. These insights empower dealerships to optimize inventory, enhance customer satisfaction, and stay competitive in the evolving automotive market.

REFERENCES

[1] https://www.tvi-mp3.com/blog/insights/automotive-industry-customer-segmentation/

[2] https://www.kaggle.com/datasets/toramky/automobile-dataset/data

Notation: [Square brackets] used to reference sources | (Parenthesis) used to reference appendix.

APPENDIX

Figure 1 - Risk Score Distribution Bar Chart

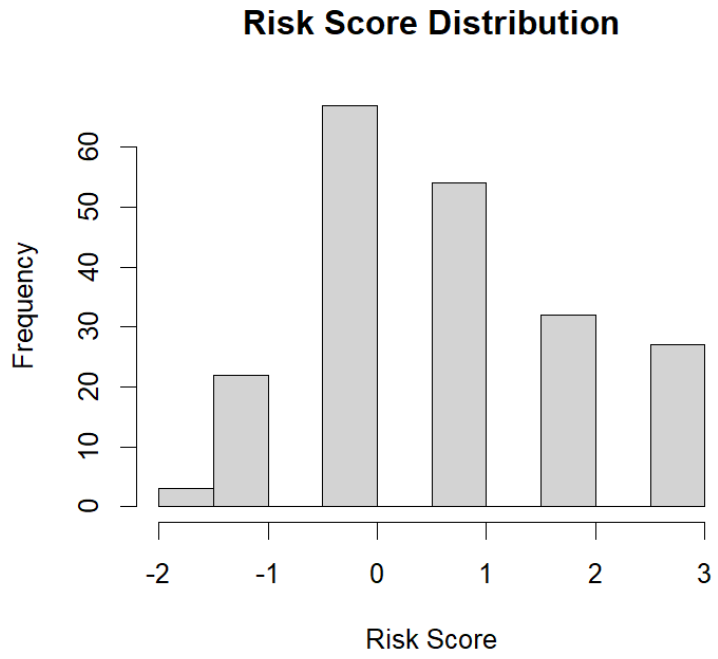
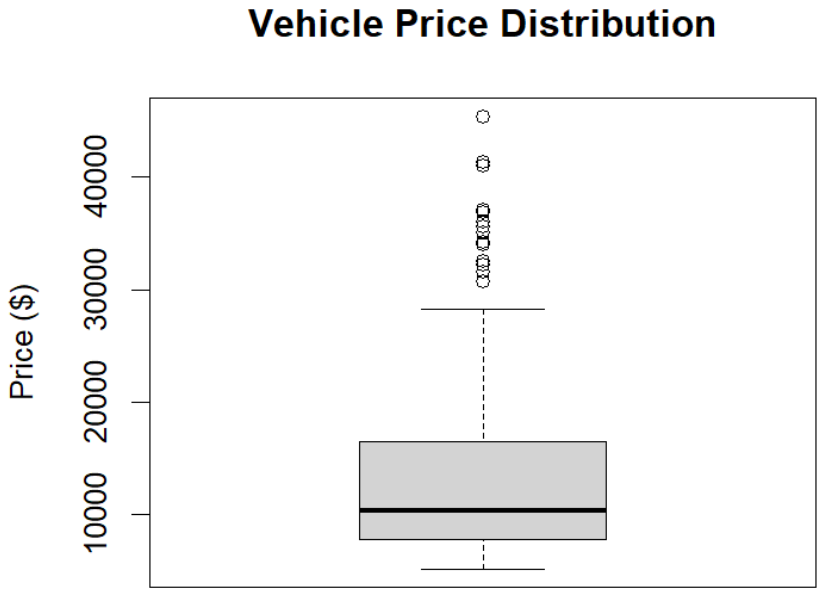


Figure 2 - Vehicle Price Distribution Box Plot



Notation: [Square brackets] used to reference sources | (Parenthesis) used to reference appendix.

Figure 3 - Vehicle Price Distribution Histogram

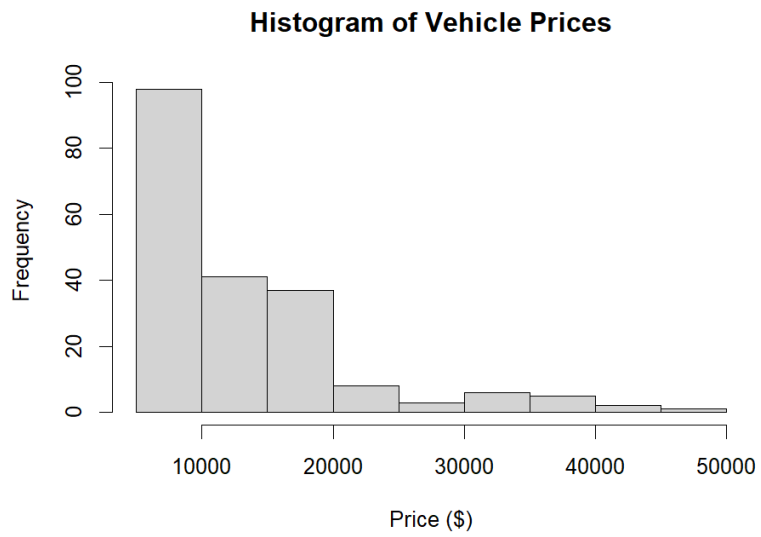
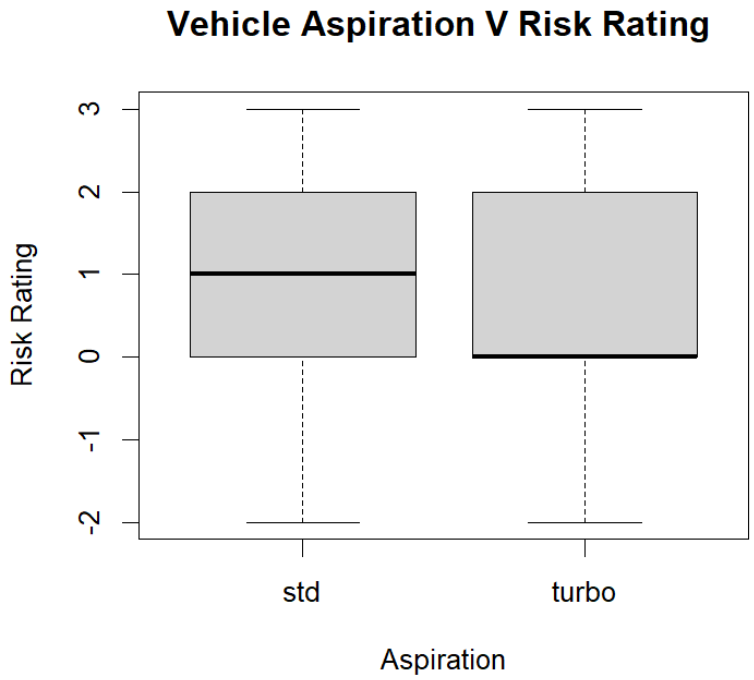


Figure 4 - Vehicle Aspiration V Risk Rating Boxplot



Notation: [Square brackets] used to reference sources | (Parenthesis) used to reference appendix.

Figure 5 - Numerical Variables Correlation Matrix

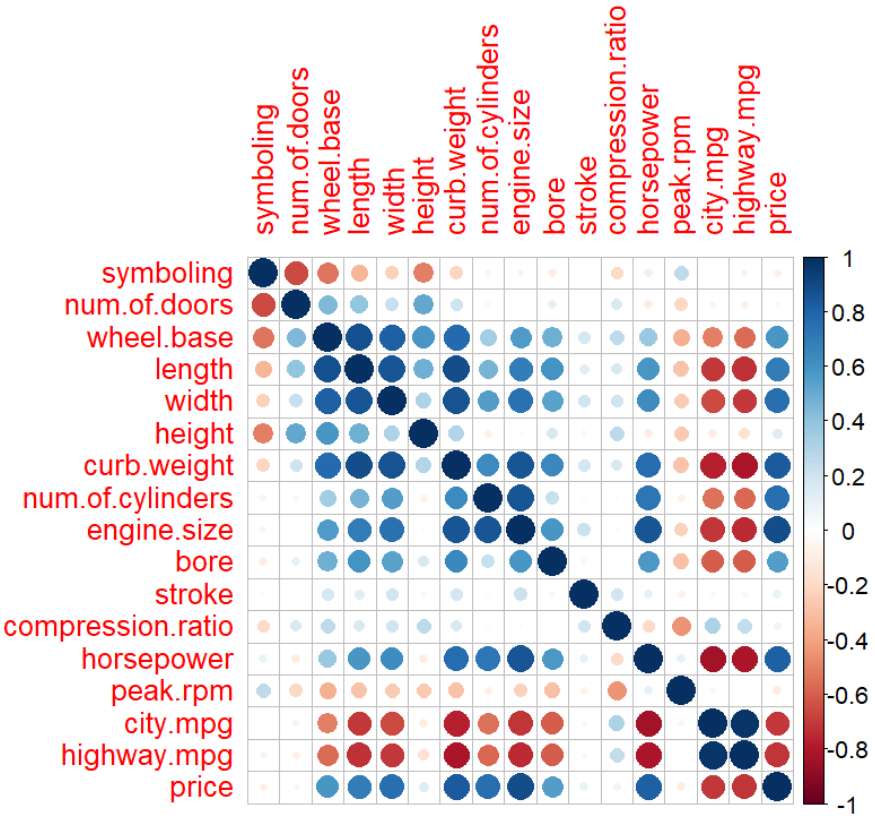
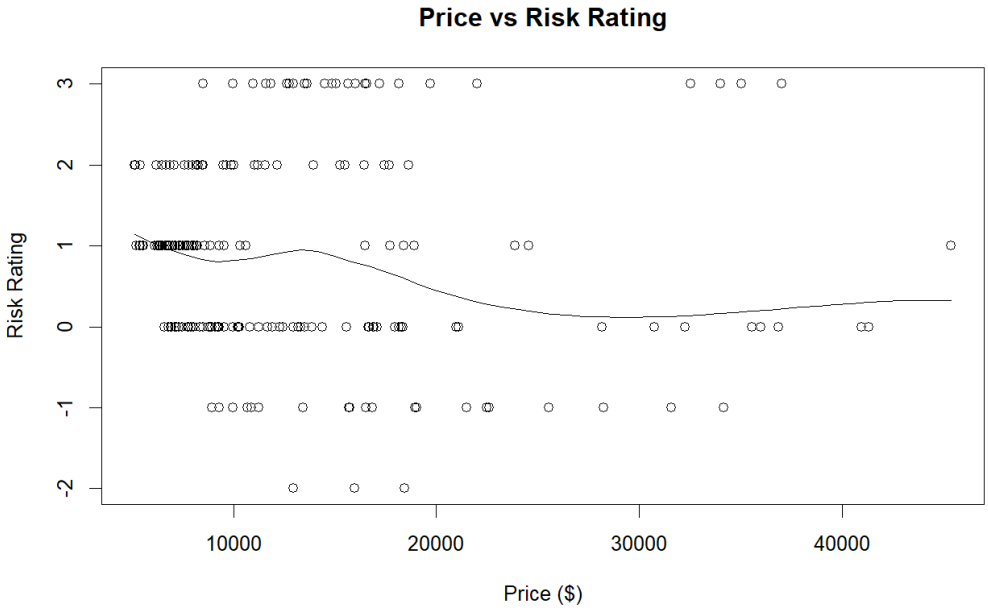


Figure 6 - Price V Risk Rating Scatter Plot



Notation: [Square brackets] used to reference sources | (Parenthesis) used to reference appendix.

Figure 7 - Random Forest Feature Importance for Insurance Risk Rating

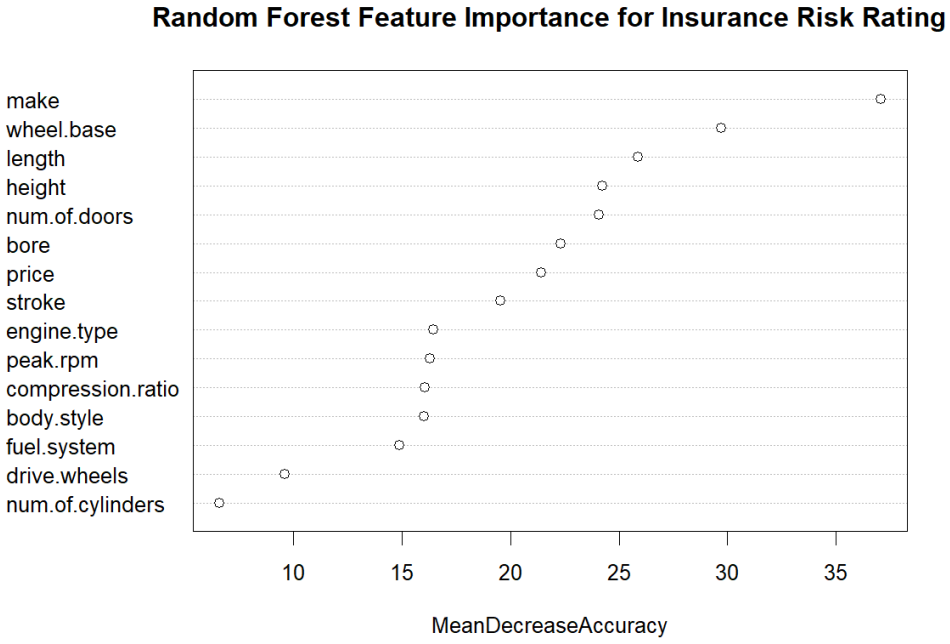
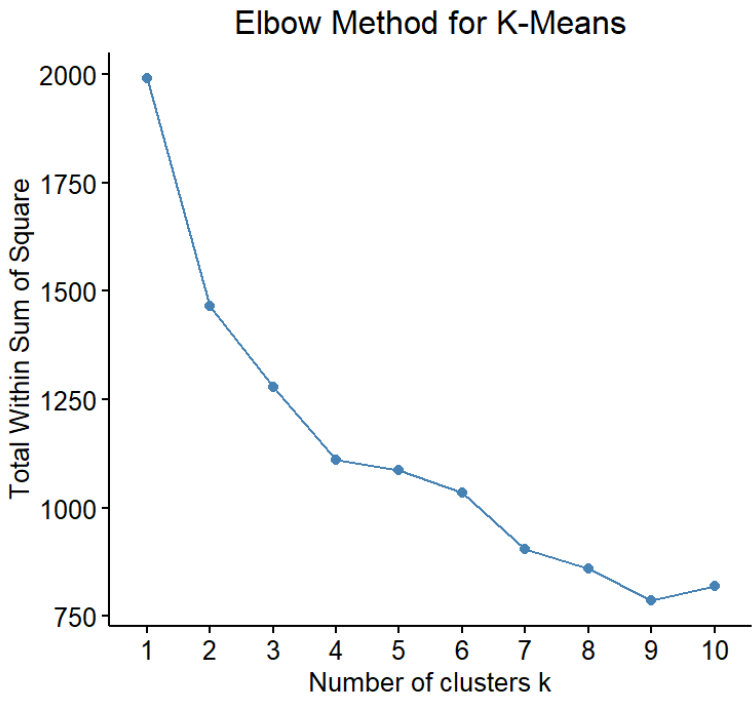


Figure 8 - K-Means Elbow Method



Notation: [Square brackets] used to reference sources | (Parenthesis) used to reference appendix.

Figure 9 - K-Means Clusters Silhouette Scores

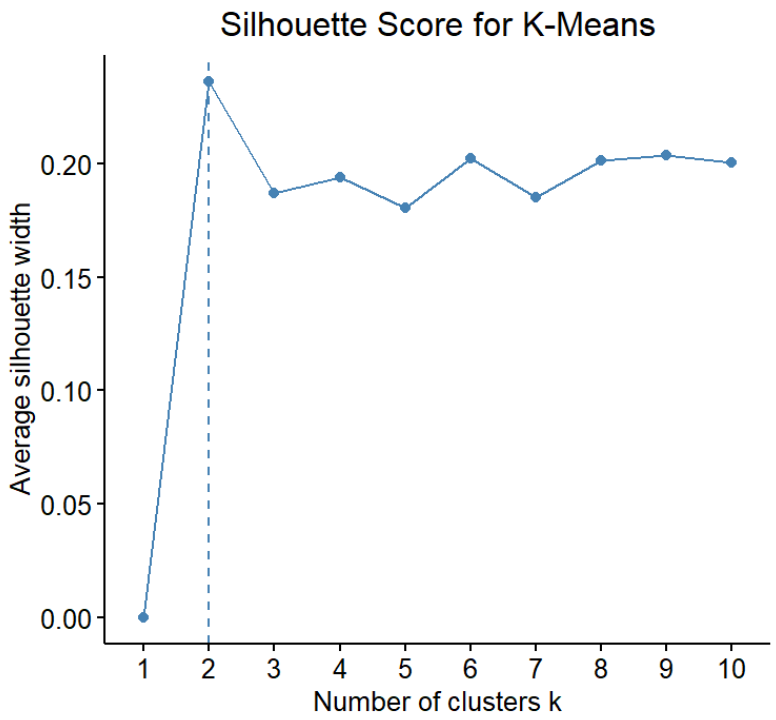
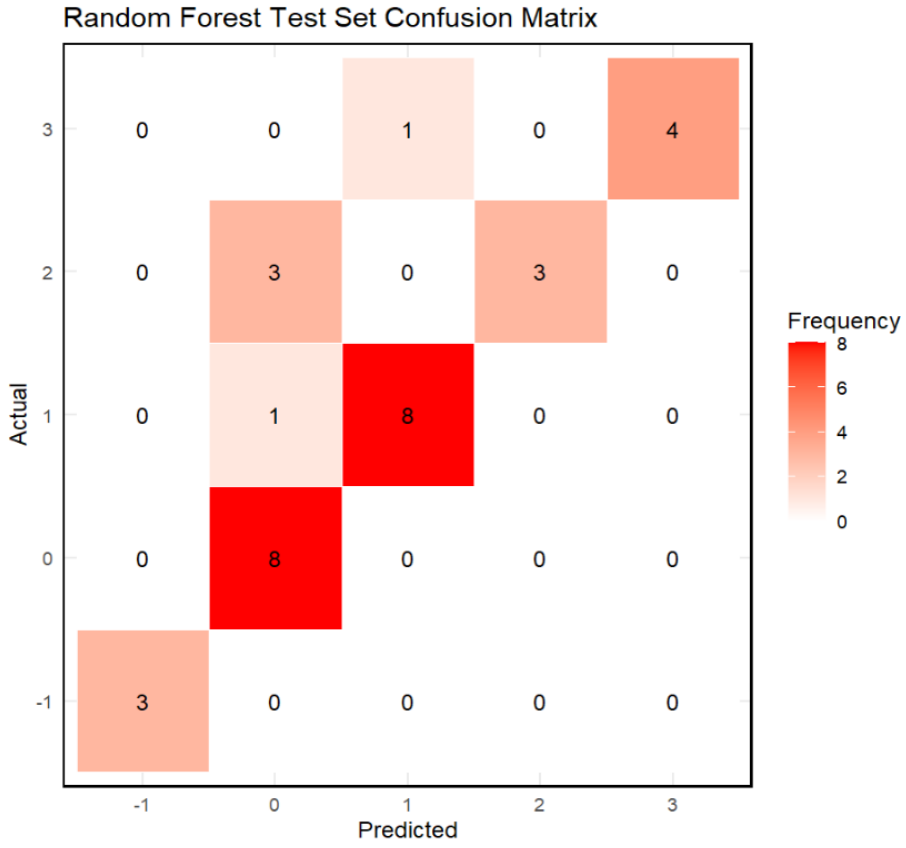


Figure 10 – Random Forest Test Set Confusion Matrix



Notation: [Square brackets] used to reference sources | (Parenthesis) used to reference appendix.

Figure 11 - K-Means Cluster Representation on the 2-Dimensional PCA Reduced Data

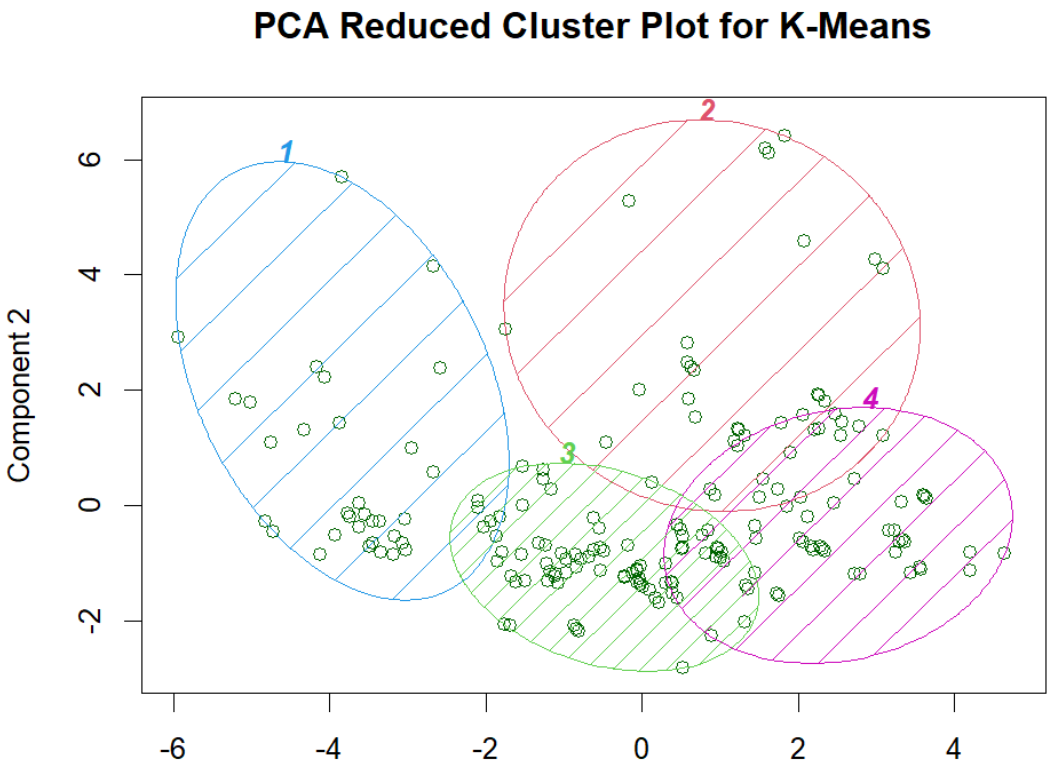


Figure 12 - Random Forest Number of Trees Out of Bag Error

ntree	OOB	1	2	3	4	5
100:	15.19%	15.79%	9.09%	17.07%	20.00%	22.22%
200:	14.56%	15.79%	9.09%	19.51%	16.00%	16.67%
300:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%
400:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%
500:	13.92%	10.53%	9.09%	19.51%	16.00%	16.67%
600:	13.29%	10.53%	9.09%	19.51%	16.00%	11.11%
700:	13.29%	10.53%	9.09%	17.07%	16.00%	16.67%
800:	13.29%	10.53%	9.09%	17.07%	20.00%	11.11%
900:	13.92%	10.53%	9.09%	19.51%	20.00%	11.11%
1000:	13.92%	10.53%	9.09%	19.51%	20.00%	11.11%
1100:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%
1200:	15.19%	10.53%	9.09%	19.51%	20.00%	22.22%
1300:	14.56%	10.53%	9.09%	17.07%	20.00%	22.22%
1400:	14.56%	10.53%	9.09%	17.07%	20.00%	22.22%
1500:	14.56%	10.53%	9.09%	17.07%	20.00%	22.22%
1600:	15.19%	10.53%	9.09%	19.51%	20.00%	22.22%
1700:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%
1800:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%
1900:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%
2000:	14.56%	10.53%	9.09%	19.51%	20.00%	16.67%