Automobile Characteristics Associated with Risk

Tom Sleigh, Alex Gibson, Matthew Casadonte and Shumeng Zhang

Statistical Analytic Testing

Abstract

The purpose of conducting the following analyses is to begin an in-depth data exploration of automobiles. Automobiles create a risk in society and have many characteristics that influence the danger associated with the vehicles. Multivariate analysis, multiple regression, logistic regression, ANOVA, clustering, decision trees, neural networks, and many more were used to conduct the data exploration process and guide future analyses. The analyses revealed a prevalent relationship between normalized-loss, width, and wheel-base in relation to risk. All three variables are associated with a significant increase in risk for a automobile. The analysis also found automobiles made by Volkswagen contain the lowest risk while automobiles made by Alfa-Romeo were deemed to have higher risk. The overall analysis conducted multiple experiments in order to produce reliable and accurate results.

Introduction

Automobiles were created in the late 19th century by Karl Benz. The invention was influenced by individuals seeking to find a quicker way of transportation. When automobiles were first created only rich and wealthy individuals were able to afford automobiles whereas today a wide variety of individuals own or rent automobiles in order to complete their daily tasks. With an increase in the number of automobiles on the roads comes an increase of total risk for individuals driving and being involved in an accident. Along with the increase of total cars on the road, there is also an increase in the different types of automobiles an individual can purchase and operate on the road. These automobiles have a variety of different features such as size, length, horsepower, etc. Do these different attributes have a significant impact on the risk of an automobile being involved in an accident? This analysis seeks to uncover information regarding what factors of an automobile have the potential to increase the overall risk.

The current analysis was influenced by acknowledging the increasing number of automobile accidents that are prevalent in society today. With the increasing number of accidents, insurance rates will continue to rise, as wills the number of deaths and injuries involving automobiles. The results of this analysis seek to find a relationship between different factors and the risk of automobiles. The analysis has the potential to predict the overall risk of those automobiles. A data set found on UCI Machine Learning contains twenty-six attributes in which range from a variety of automobiles. This data-set will provide the analysis with the foundation needed to discover a meaningful relationship between the risk and the attributes stated above. This analysis will assess risk between the numbers -3 and 3. The automobiles containing the least amount of risk are -3 and the highest amount of risk are 3. Although this data set has the potential to hold reliable and meaningful results, one challenge this analysis will have to overcome is a number of missing values that could be vital to the overall results. Despite the missing values in the data set, the analysis still has the potential to discover significant results. This analysis only considers the attributes of the automobile and not the attributes of an individual driving the car. This problem could be a separate analysis and also yield significant results of which could expand on the current analysis.

In order to achieve the overall goal of this analysis, a variety of statistical tests will be conducted and analyzed. The analysis seeks to discover meaningful, reliable and valid results. Some tests involved in this analysis consist of regression, ANOVA, scatter plots, and many more predictive decision-making tools. The impact this analysis could yield if significant results are found is enormous. First, this analysis could impact the pricing of insurance rates. The more risk associated with a automobile the higher the insurance rates and the lower the risk, the lower insurance rates. Secondly, this analysis has the potential to reduce the number of deaths due to automobile accidents. If individuals can identify the automobiles that have a higher risk of being involved in an accident, then safety precautions could be enforced on these particular automobiles. Lastly, the majority of individuals want to feel safe while driving, if these individuals can identify the characteristics of an automobile of which have a low risk of being in an accident then they could purchase these types of cars. This could also be a marketing technique by dealerships to sell more automobiles. This analysis has the potential to impact the automobile industry for the foreseeable future.

Methodology

Data

205 different instances of automobiles were collected with 26 attributes for each automobile. There are a few missing values in the data set which should have little effect on the overall results of the analysis. Data points were collected and classified as categorical, integer and continuous values. All data was collected through public records of insurance reports and Ward's Automotive Yearbook. There are two variables within the data set that were analyzed in depth and require further explanation. Those variables were Normalized Losses and Symbolizing. Normalized losses is a continuous variable that is calculated on the average amount of money in USD that insurance companies paid out on that specific vehicle per claim in the previous year. The variable is normalized over the body style of the vehicle, as some vehicles are larger. Thus, the larger and more complex vehicles would require more money to when involved in an accident than that of a small and relatively simple vehicle, even if both vehicles had the same amount of physical damage. Symbolizing is an ordinal variable ranging from -3.0 to 3.0, in increments of 1, that is measured by the insurance companies in determining the risk of the vehicle. In our report this variable is sometimes referred to as "risk factor' or simply "risk". In order to analyze the data efficiently, we used JMP Pro Statistical software package.

Procedure

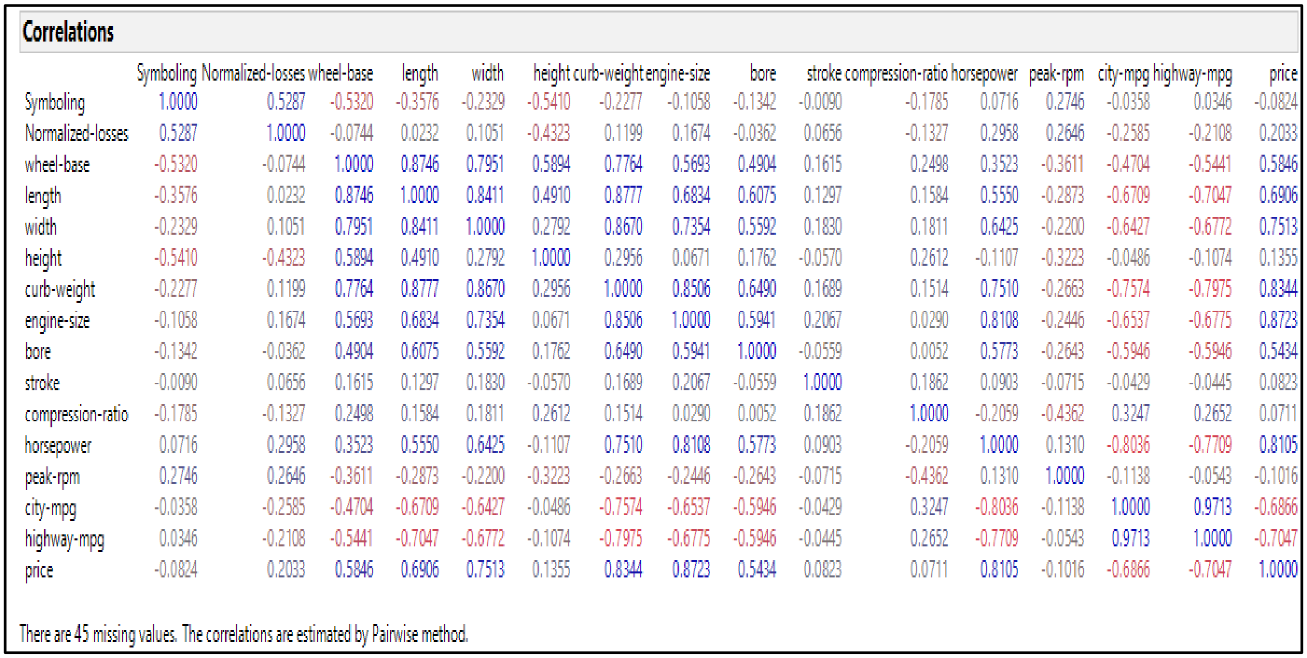
Predictive modeling and correlation relationships are utilized during the analysis. This analysis included multiple regression, ANOVA, principal component analysis (PCA), logistic regression, decision trees, K-Nearest Neighbor, and Naive Bayes. The analysis was started by using the correlation values between variables to assess their relationships to the one another. The multiple regression was completed next and is useful to this study to understand the cause and effect relationship between symbolizing (risk) and the other attributes of an automobile. We then conducted an ANOVA test that assessed the variance between the variables in association with risk. Next, we conducted a principal component analysis to assess the structure of correlations and identify unnecessary variables. By identifying the unnecessary variables, we can create a stronger relationship between our independent and dependent variables. Next, a logistic regression test was built conducted to compare the body style of the vehicle and their associated risk to further explore that relationship. Then a decision tree test was conducted in order to find the automobiles of great or little risk compared with the independent variables. The next analysis conducted were binary model comparisons including ROC, Lift curves and Binary predictive modeling. The binary modeling is important when trying to predict how risky a certain automobile is. Next, the analysis conducted K-Nearest Neighbors (KNN) and Naive Bayes (NB) analysis to further investigate the predictive power of the primary influencing variables and produce easily interpretable results. The analysis concluded by comparing the outputs from each of the binary tests to determine which model best represented our data. It was important to conduct a variety of statistical test in order to provide the public with reliable, meaningful and significant results.

Analysis

Multiple Regression

Figure 1 shows the correlation values of the different variables compared to each other. Wheelbase has a negative correlation valued at -.5320. Using the correlation values an individual can see the lower the wheel-base value the higher the associated symbolizing value. Figure 1 shows that height has a -.5410 correlation with risk. This reveals that the lower the height the higher the risk value, thus the higher associated risk. A third variable is normalized-losses of which also has a correlation with symbolizing. Figure 1 shows that symbolizing and normalized losses have a -.5287-correlation factor.

Many other variables have high correlations with each other. Focusing on variables that have high correlations with wheel-base or height could potentially point to other variables in which play a role in the symbolizing value for a specific vehicle. Length, width, and curb-weight have a .8746, .7951, and .7764 correlation value with wheel-base, respectively. These variables are highly correlated with wheel-base. Height does not exhibit as a strong correlation with the other variables in the model.



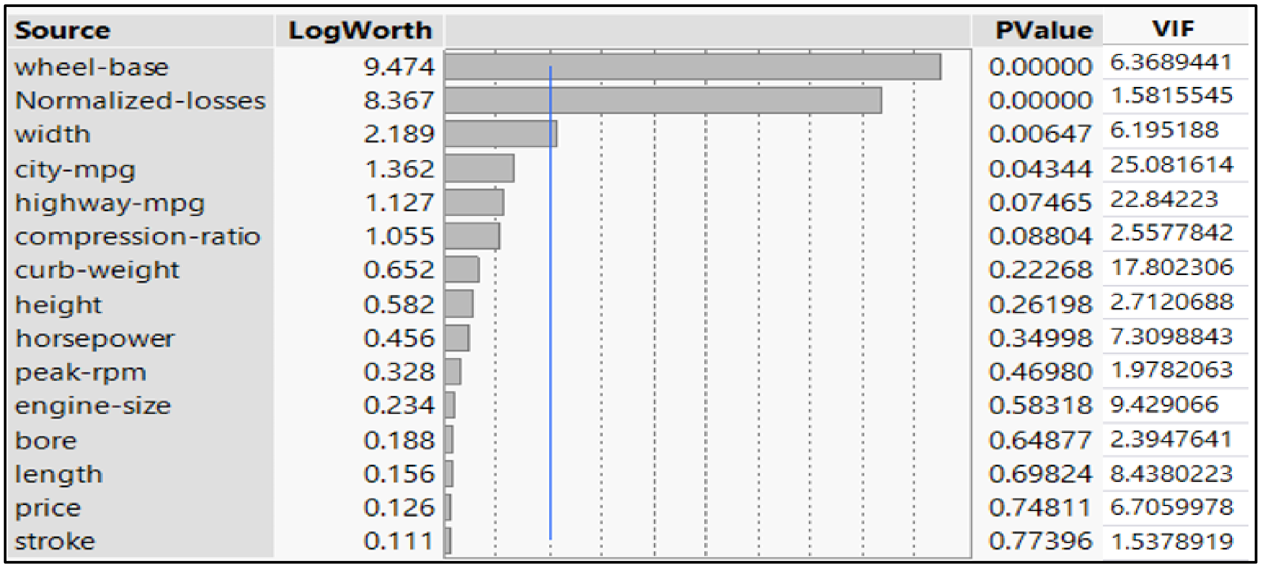
Correlations

	Symboling	Normalized-losses	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
Symboling	1.0000	0.5287	-0.5320	-0.3576	-0.2329	-0.5410	-0.2277	-0.1058	-0.1342	-0.0090	-0.1785	0.0716	0.2746	-0.0358	0.0346	-0.0824
Normalized-losses	0.5287	1.0000	-0.0744	0.0232	0.1051	-0.4323	0.1199	0.1674	-0.0362	0.0656	-0.1327	0.2958	0.2646	-0.2585	-0.2108	0.2033
wheel-base	-0.5320	-0.0744	1.0000	0.8746	0.7951	0.5894	0.7764	0.5693	0.4904	0.1615	0.2498	0.3523	-0.3611	-0.4704	-0.5441	0.5846
length	-0.3576	0.0232	0.8746	1.0000	0.8411	0.4910	0.8777	0.6834	0.6075	0.1297	0.1584	0.5550	-0.2873	-0.6709	-0.7047	0.6906
width	-0.2329	0.1051	0.7951	0.8411	1.0000	0.2792	0.8670	0.7354	0.5592	0.1830	0.1811	0.6425	-0.2200	-0.6427	-0.6772	0.7513
height	-0.5410	-0.4323	0.5894	0.4910	0.2792	1.0000	0.2956	0.0671	0.1762	-0.0570	0.2612	-0.1107	-0.3223	-0.0486	-0.1074	0.1355
curb-weight	-0.2277	0.1199	0.7764	0.8777	0.8670	0.2956	1.0000	0.8506	0.6490	0.1689	0.1514	0.7510	-0.2663	-0.7574	-0.7975	0.8344
engine-size	-0.1058	0.1674	0.5693	0.6834	0.7354	0.0671	0.8506	1.0000	0.5941	0.2067	0.0290	0.8108	-0.2446	-0.6537	-0.6775	0.8723
bore	-0.1342	-0.0362	0.4904	0.6075	0.5592	0.1762	0.6490	0.5941	1.0000	-0.0559	0.0052	0.5773	-0.2643	-0.5946	-0.5946	0.5434
stroke	-0.0090	0.0656	0.1615	0.1297	0.1830	-0.0570	0.1689	0.2067	-0.0559	1.0000	0.1862	0.0903	-0.0715	-0.0429	-0.0445	0.0823
compression-ratio	-0.1785	-0.1327	0.2498	0.1584	0.1811	0.2612	0.1514	0.0290	0.0052	0.1862	1.0000	-0.2059	-0.4362	0.3247	0.2652	0.0711
horsepower	0.0716	0.2958	0.3523	0.5550	0.6425	-0.1107	0.7510	0.8108	0.5773	0.0903	-0.2059	1.0000	0.1310	-0.8036	-0.7709	0.8105
peak-rpm	0.2746	0.2646	-0.3611	-0.2873	-0.2200	-0.3223	-0.2663	-0.2446	-0.2643	-0.0715	-0.4362	0.1310	1.0000	-0.1138	-0.0543	-0.1016
city-mpg	-0.0358	-0.2585	-0.4704	-0.6709	-0.6427	-0.0486	-0.7574	-0.6537	-0.5946	-0.0429	0.3247	-0.8036	-0.1138	1.0000	0.9713	-0.6866
highway-mpg	0.0346	-0.2108	-0.5441	-0.7047	-0.6772	-0.1074	-0.7975	-0.6775	-0.5946	-0.0445	0.2652	-0.7709	-0.0543	0.9713	1.0000	-0.7047
price	-0.0824	0.2033	0.5846	0.6906	0.7513	0.1355	0.8344	0.8723	0.5434	0.0823	0.0711	0.8105	-0.1016	-0.6866	-0.7047	1.0000

There are 45 missing values. The correlations are estimated by Pairwise method.

Figure 1: Correlation Values for Multiple Regression

Using the information gathered from the correlation values multiple linear regression can be conducted on the data set. With symbolizing as the Y variable and the remaining nominal variables used in multiple regression analysis, the Effects Summary, Figure 2, can be generated. Log Worth values greater than 2 correspond to p-values less than .01. Based on the Log worth values, it can be concluded that wheel-base, width, and normalized losses are significant to the model, and thus play an important role in determining the symbolizing, or risk factor of an automobile.



Source	LogWorth	PValue	VIF
wheel-base	9.474	0.00000	6.3689441
Normalized-losses	8.367	0.00000	1.5815545
width	2.189	0.00647	6.195188
city-mpg	1.362	0.04344	25.081614
highway-mpg	1.127	0.07465	22.84223
compression-ratio	1.055	0.08804	2.5577842
curb-weight	0.652	0.22268	17.802306
height	0.582	0.26198	2.7120688
horsepower	0.456	0.34998	7.3098843
peak-rpm	0.328	0.46980	1.9782063
engine-size	0.234	0.58318	9.429066
bore	0.188	0.64877	2.3947641
length	0.156	0.69824	8.4380223
price	0.126	0.74811	6.7059978
stroke	0.111	0.77396	1.5378919

Figure 2: Effect Summary

Wheelbase, width, normalized losses, and city-mpg all have p-values less than .05. Due to the low p-values, we can reject the null hypothesis that the variables contain no relationship. These variables are significant to the model and have some effect on the symbolizing of automobiles.

Multicollinearity can make it difficult to determine the meaning of regression of coefficients of the independent variables. The closer the VIF (Variation Inflation Factor) is to one the less likely that there is collinearity between variables. Figure 2 shows the VIF values for all of the independent variables. Many of the variables have VIF values between 5-10 which raises a concern of the possibility of collinearity within the variables in the model. Variables with a VIF value above 10 indicate high multicollinearity.

ANOVA

A One-Way Analysis (ANOVA) of Symbolizing and Make revealed many attributes about the model. From our data set, we conducted an ANOVA analysis with the Make of the vehicle as the independent and the Risk factor as the dependent variable. In order to simplify the ANOVA test, the vehicles were grouped into American, European, and Asian makes. The One-Way Analysis plot seen in Figure 3 reveals the mean diamond for each of the factors. The diamond reveals the differences in mean between the classes as well as the differences in sample size. These differences can be seen by comparing the shape and structure of the diamond for each class. The American class appears to have the highest mean with its center line being the highest amongst the classes, while the Asian class appears to have the largest sample size with its diamond having the greatest width out of the three classes.

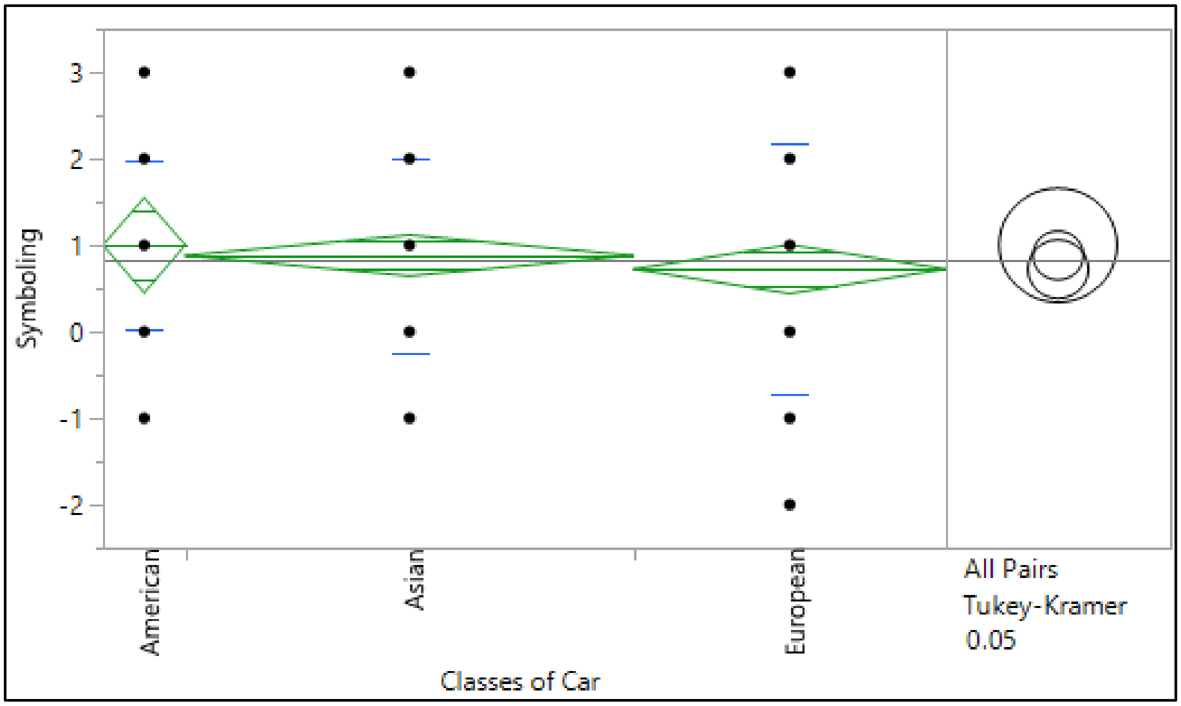
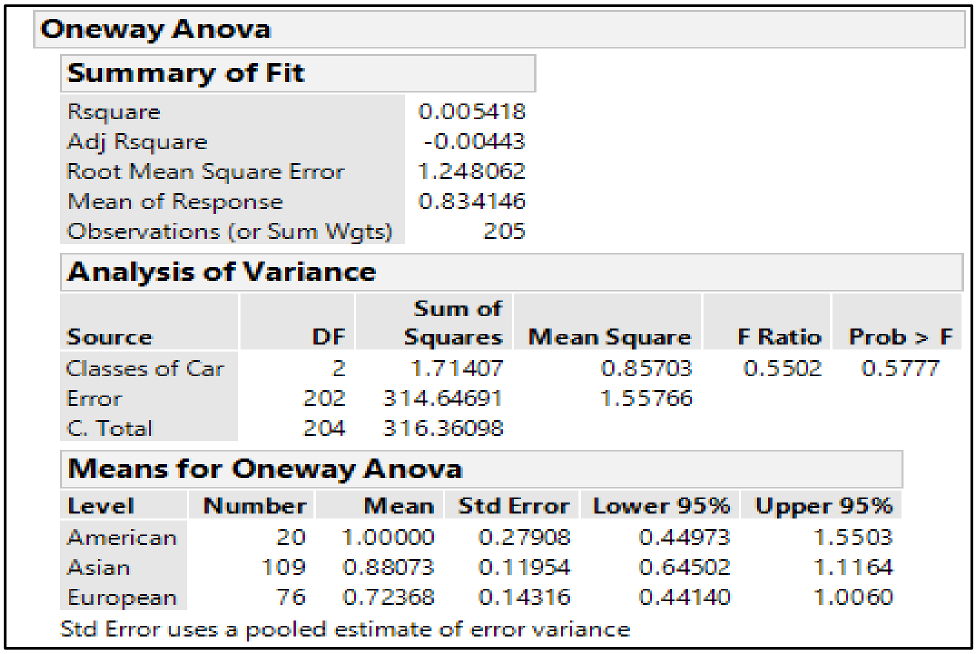


Figure 3: One-Way Analysis of Risk by Classes of Car

Figure 4 represents the results from the ANOVA analysis. The ANOVA analysis failed to reject the null hypothesis because the value for the F-statistic is 0.5777. The test compared the variance between the means of American, European and Asian populations.



Oneway Anova

Summary of Fit

Rsquare	0.005418
Adj Rsquare	-0.00443
Root Mean Square Error	1.248062
Mean of Response	0.834146
Observations (or Sum Wgts)	205

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Classes of Car	2	1.71407	0.85703	0.5502	0.5777
Error	202	314.64691	1.55766		
C. Total	204	316.36098			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
American	20	1.00000	0.27908	0.44973	1.5503
Asian	109	0.88073	0.11954	0.64502	1.1164
European	76	0.72368	0.14316	0.44140	1.0060

Std Error uses a pooled estimate of error variance

Figure 4: ANOVA

The difference in variance between the means of the three populations can be seen in Appendix A. The Levene test has an F statistic value of <.0001, which reveals the variances in the three classes of vehicles are different. The Welch test reveals a p-value of 0.5792 and does not allow for the null hypothesis to be rejected. The means of the different classes of vehicles cannot be deemed significantly different from the Welch test. The Tukey HSD test and the connecting letters report reveals, seen in Appendix A, there is not a significant difference in the symbolizing of the vehicle makes. The ANOVA test revealed that while there are differences in the average symbolizing of the vehicle makes, the differences cannot be deemed statistically significant according to the multiple tests that were conducted.

Principal Component Analysis

Due to the high VIF values, PCA analysis was conducted on all variables with a VIF value lower than 5 to focus in on any remaining collinearity. PCA gave insight into the structure of the data and revealed if any of the remaining independent variables with a low VIF were closely related. The eigenvalues in Figure 5 shows how important the principal components are in relation to one another. Prin1 with a value of 2.0216 is significantly more important than prin2 which only has a value of 1.1847. When examining the horizontal bar graph in Figure 5, an individual can see Prin1 accounts for 33.7% of the variation. The loadings plot shows the contribution made to each of the principal components by the variables. Clustering in the loadings plot reveals which variables may be related to one another in the principal component space. Bore and height are located the closest together in the loadings plot, which points to the two variables being the most correlated. This poses an interesting relationship for these two variables, as one measures the height on the exterior of the vehicle, while bore is a measurement of the individual cylinders inside of the engine. Height and normalized loss are located on opposite sides of the loadings plot which shows the two variables may be negatively correlated to one another.

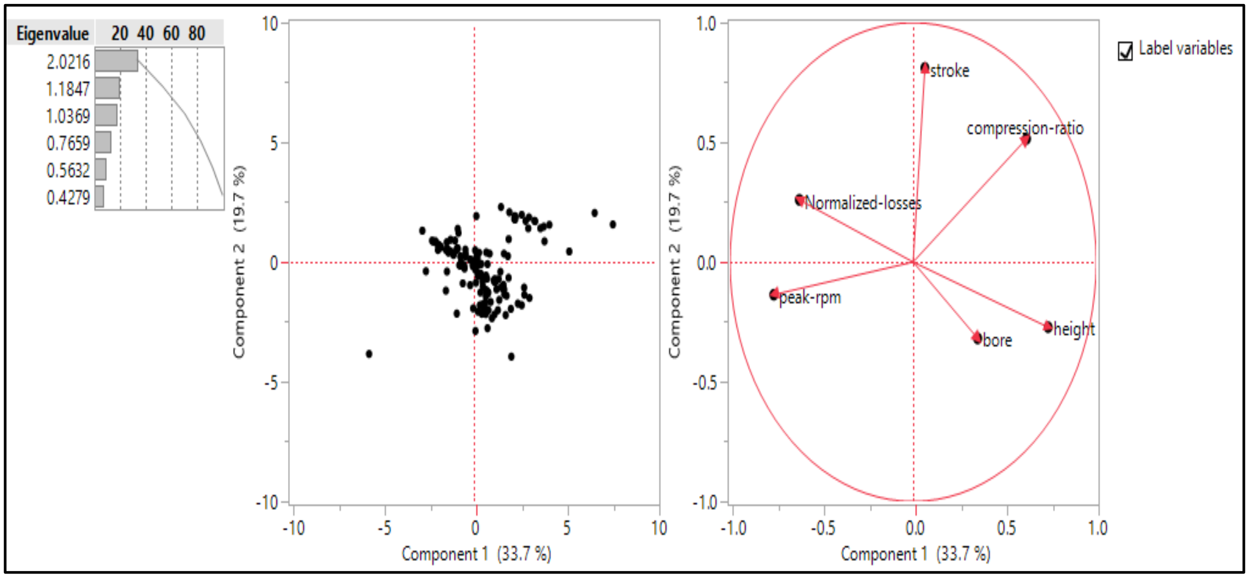
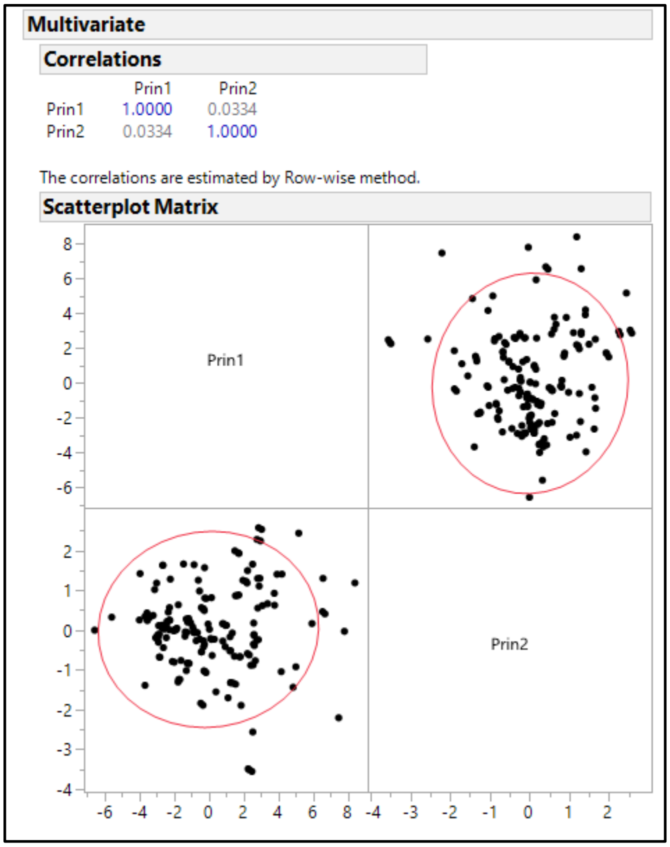


Figure 5: Principal Component Analysis

Figure 6 verifies the lack of correlation between Prin1 and Prin2. Figure 7 allows for the use of the elbow method to determine the number of principal components. The point at which the elbow occurs reveals the number of principal components, in this case, the value is two.



Multivariate

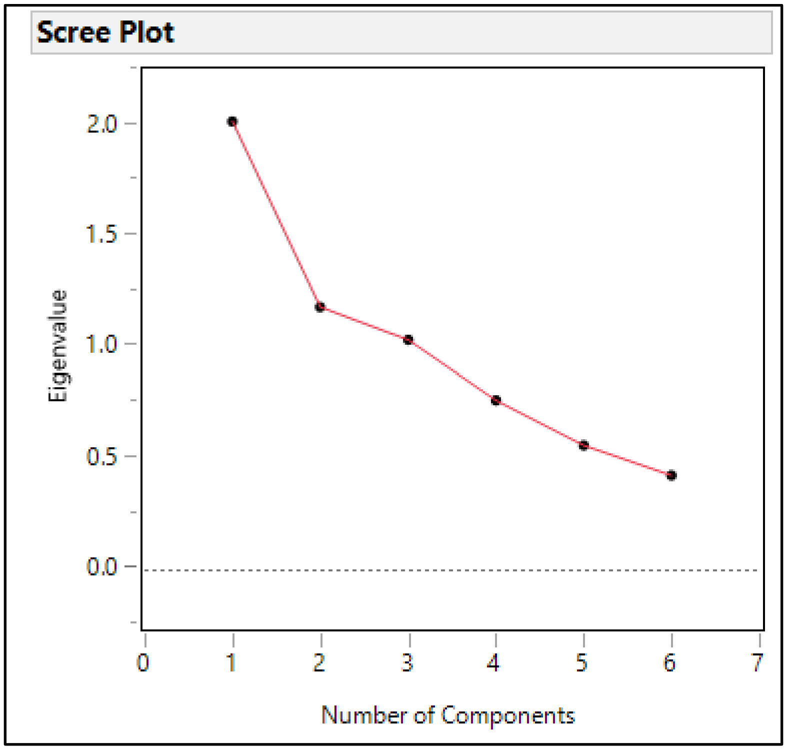
Correlations

	Prin1	Prin2
Prin1	1.0000	0.0334
Prin2	0.0334	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

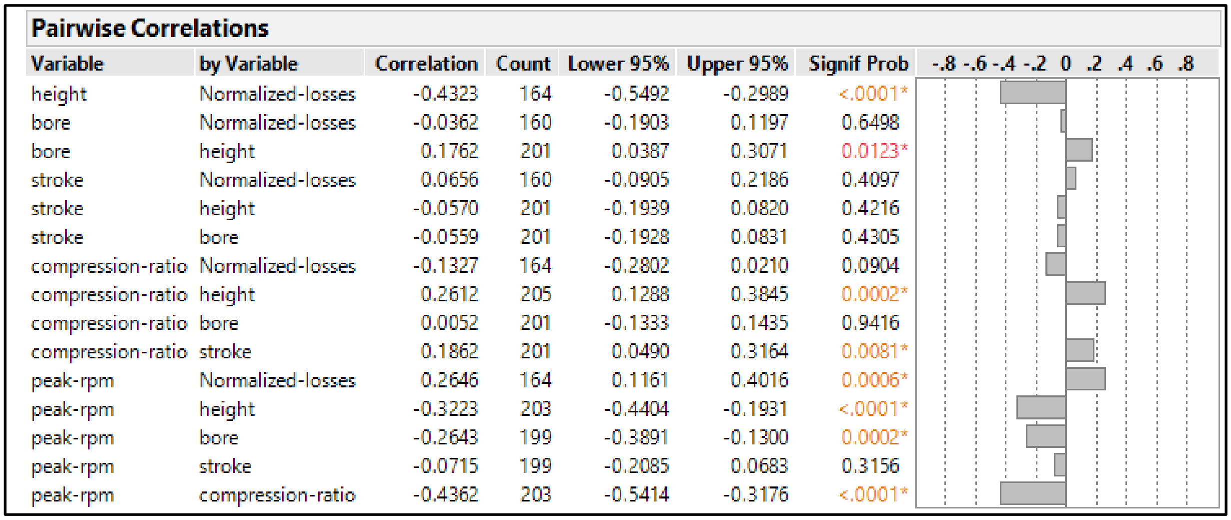
Figure 6: Multivariate of Principle Components



Scree Plot

Figure 7: Scree Plot

Figure 8 depicts the pairwise correlations between variables with a VIF less than 5. The pairwise correlation confirms the clustering that is seen in the loadings plot, such as between bore and height. This can be seen between bore and height on the loadings plot. Variables with a negative correlation are seen on opposite sides of the loadings plot. This can be seen in height and normalized-losses on the loadings plot.



Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
height	Normalized-losses	-0.4323	164	-0.5492	-0.2989	<.0001*	
bore	Normalized-losses	-0.0362	160	-0.1903	0.1197	0.6498	
bore	height	0.1762	201	0.0387	0.3071	0.0123*	
stroke	Normalized-losses	0.0656	160	-0.0905	0.2186	0.4097	
stroke	height	-0.0570	201	-0.1939	0.0820	0.4216	
stroke	bore	-0.0559	201	-0.1928	0.0831	0.4305	
compression-ratio	Normalized-losses	-0.1327	164	-0.2802	0.0210	0.0904	
compression-ratio	height	0.2612	205	0.1288	0.3845	0.0002*	
compression-ratio	bore	0.0052	201	-0.1333	0.1435	0.9416	
compression-ratio	stroke	0.1862	201	0.0490	0.3164	0.0081*	
peak-rpm	Normalized-losses	0.2646	164	0.1161	0.4016	0.0006*	
peak-rpm	height	-0.3223	203	-0.4404	-0.1931	<.0001*	
peak-rpm	bore	-0.2643	199	-0.3891	-0.1300	0.0002*	
peak-rpm	stroke	-0.0715	199	-0.2085	0.0683	0.3156	
peak-rpm	compression-ratio	-0.4362	203	-0.5414	-0.3176	<.0001*	

Figure 8: Pairwise Correlations for Correlations for Variables with VIF<5

Logistic Regression and Fit

To search for variables of which coincide with high-risk vehicles, we fit the nominal variable “body style” against the continuous variable “Risk” in order to determine if a specific body style correlated with higher risk factors. The overall fit of the model returned an R-squared value of 0.1937, however, the analysis we are more interested in is the individual categories and where the midpoints of the lines are within the graph. Figure 9 shows how the line of best fit for each body style passes through the risk variable. Wagons are shown to contain the least risk, with the highest intercept value, approximately 0.45, and the midpoint of the line of best fit being between 0.0 and 1.0. Convertibles were found to be the riskiest, with a midpoint in the line of best-fit crossing over at approximately 3.0 Sedans were shown to have the highest amount of variation in their risk factor, approximately 1.0. This is seen through the large spread of the line that crosses over the midpoint of the entire model. Since the whole model returned a p-value of less than .0001, we can conclude with confidence wagons are the least risky body style and convertibles are the riskiest body style of vehicles.

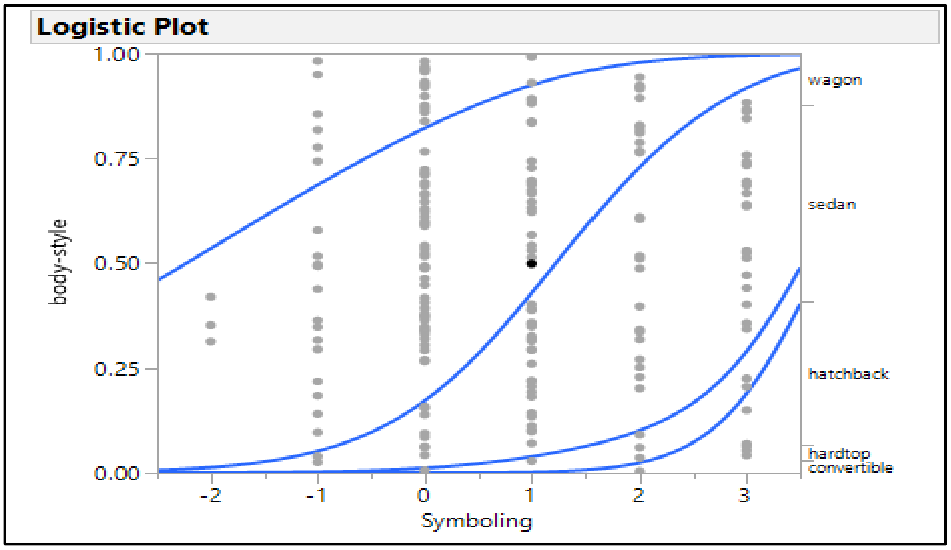


Figure 9: Logistic fit of Body Type by Symbolizing

An alternate finding to this model depicted in Figure 9, showed only 3 vehicles of which returned a Risk value of -2, all of which are made by the brand Volvo. These are the three lowest scoring vehicles within the risk category. This was an unintentional find from this analysis and shows that we should continue to investigate this particular brand and how safe it is compared to its competitors.

Clustering

In order to expand the current analysis, hierarchical clustering was conducted. A constellation plot was created and can be examined in Figure 10. After analyzing the constellation plot, one can see the relatively low-risk automobiles in the same clusters and high-risk automobiles in the same clusters. Some clusters in the constellation plot contain high-risk automobiles ranging from 2 to 3. Other clusters contain average risk and range from -1 to 1. Next, some clusters contain low-risk automobiles ranging from -1 to -3. Lastly and most important, the constellation plot examined multiple characteristics of an automobile. These characteristics include wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, and price. These characteristics were plotted on the constellation plot with the label being symbolizing (risk). From these characteristics and the clusters an individual can see with an increase in any of these characteristics follows an increase in the risk level associated with an automobile. If an individual examines the top right corner of the constellation plot one will discover a cluster of many 1,2 and 3's. These numbers are associated with higher risk vehicles and the top right corner of the plot is associated with the increase of risk in the examined characteristics of the automobiles. However, if an individual examines the bottom left-hand corner of the constellation plot, one will see a cluster of risk levels of which contain 0 and -1's. These risk levels are associated with lower risk automobiles. The bottom left-hand corner of the constellation plot is associated with a decrease in the examined characteristics. The analysis only included five clusters as a result of examining the scree plot in Figure 11. Figure 12 shows the parallel coordinate plot for the constellation plot. After analyzing the parallel coordinate plot one can conclude that price has a less significant effect on the overall risk of a vehicle. One can also see bore, stroke, normalized-loss, width and height all are associated with higher levels of risk.

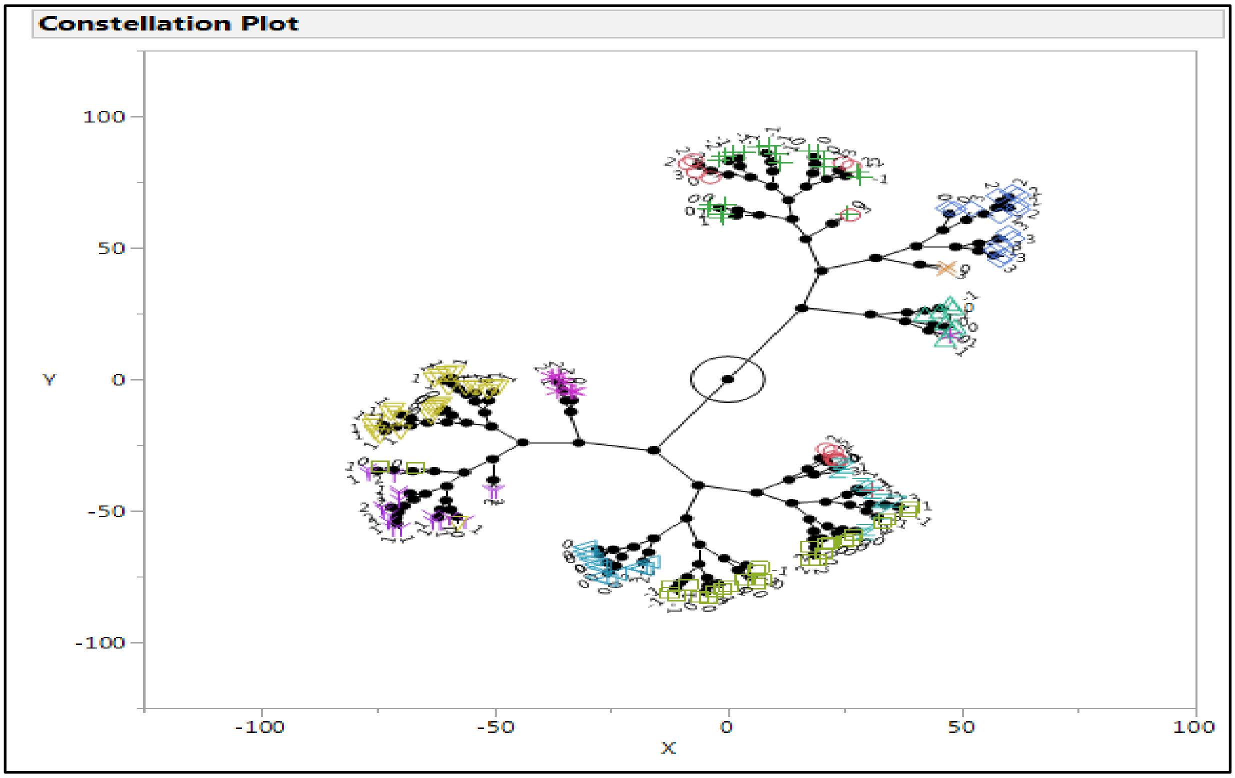


Figure 10: Constellation Plot

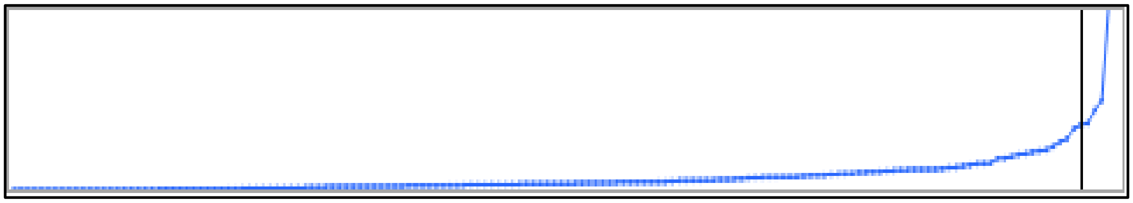


Figure 11: Scree Plot

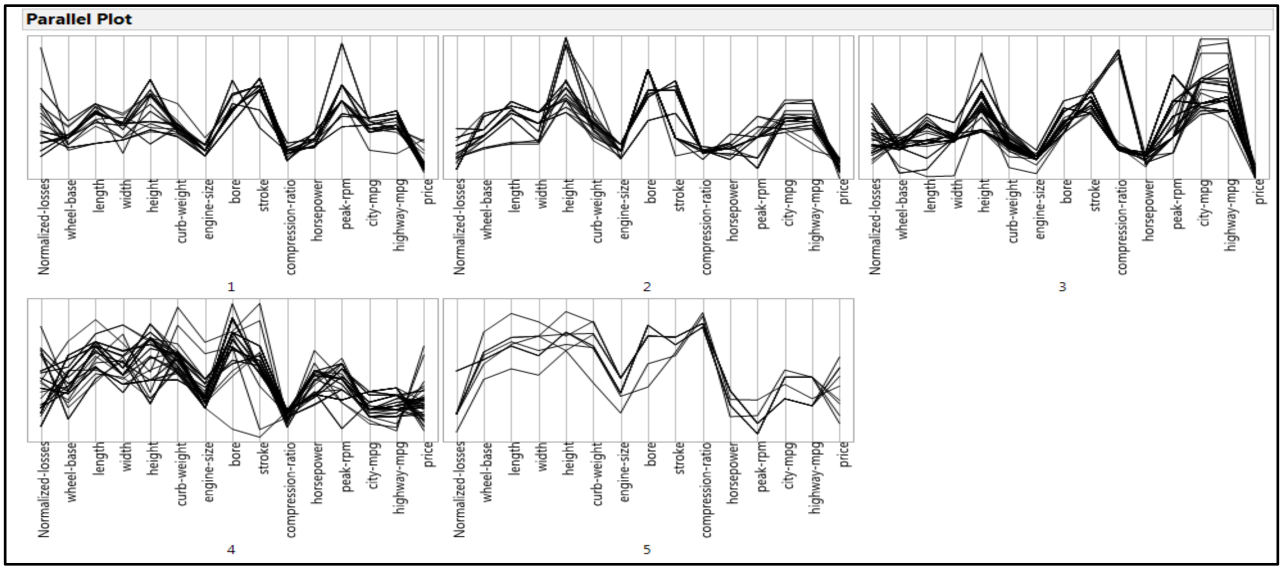


Figure 12: Parallel Plot from Clustering

In order to ensure the right number of clusters were found, a K-Means plot and parallel coordinate plot were constructed. The K-Means plot, Figure 13, shows little to no overlapping of clusters which supports the notion of five being the optimal number of clusters. The k-means plot allowed us to examine what characteristics made up each cluster and revealed interesting results. First, by simply examining the k-means plot an individual can see the higher risk automobiles are associated with clusters 2, 4 and 1. The lower risk vehicles are associated with clusters 3 and 5. Cluster 1 is made up of 46 data points, cluster 2 is made up of 9 data points, cluster 3 is made up of 41 data points, cluster 4 is made up of 39 data points and cluster 5 is made up of 25 data points. Although cluster 2 has a significantly lower amount of data points, we believe it is still an important cluster because points are noticeably different than the other clusters. Our theory that 5 clusters is the optimal number of clusters that best fits our data is also supported by the scree plot which supports cluster 2 being included in the analysis. Saying that the data points in cluster 2 were majority wheelbase and width. This is consistent with previous results as these two characteristics are associated with higher risk automobiles. Next, cluster 4 has automobiles of which contain the highest height. Next, cluster 1 shows vehicles with the highest highway mpg and city-mpg. Cluster 1 also contains automobiles with the smallest wheelbase and width. This is supported by the fact cluster 1 is in the middle-risk level. Next, cluster 3 has the highest normalized-loss and the highest price. Price did not play a major role in the risk factor. Finally, cluster 5 contains almost all of the lowest numbers of characteristics of automobiles. This is supported by the fact cluster five is created by majority low-risk automobiles. The analysis of clustering concluded with a parallel coordinate plot, Figure 14, of the k-means. One can see very similar results in the k-means parallel coordinate plot when compared to the constellation parallel coordinate plot. One noticeable difference between the two coordinate plots is price varies greatly in the k-means coordinate plot and is relatively lower on the plot in the constellation parallel coordinate plot.

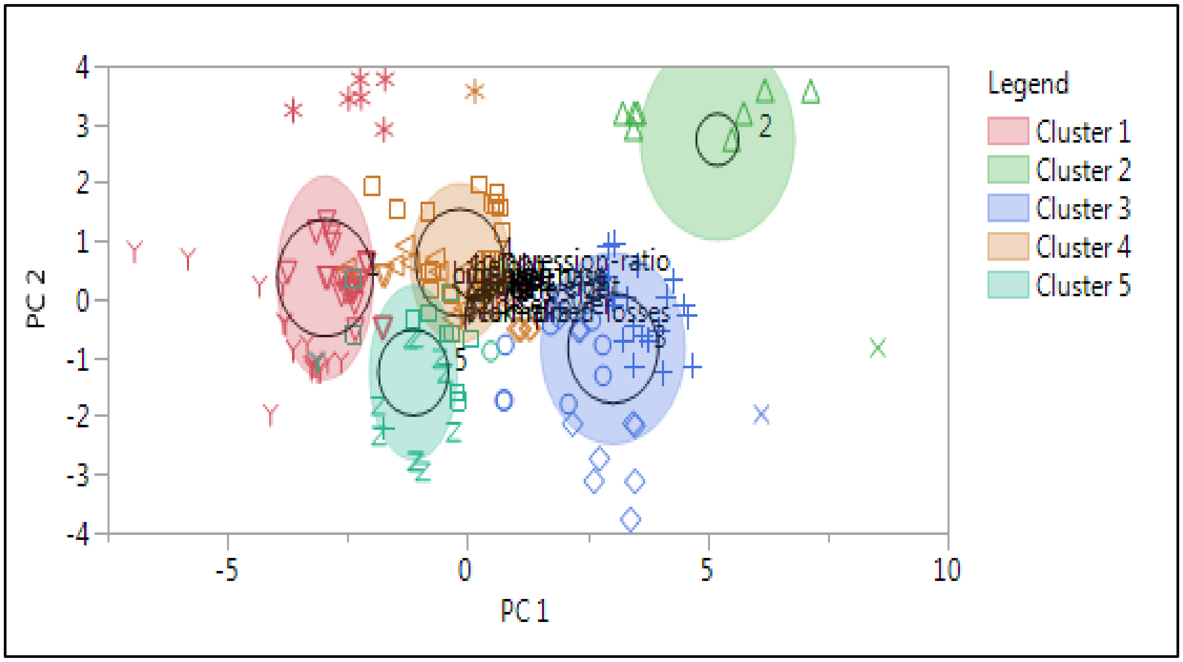


Figure 13: K-Means

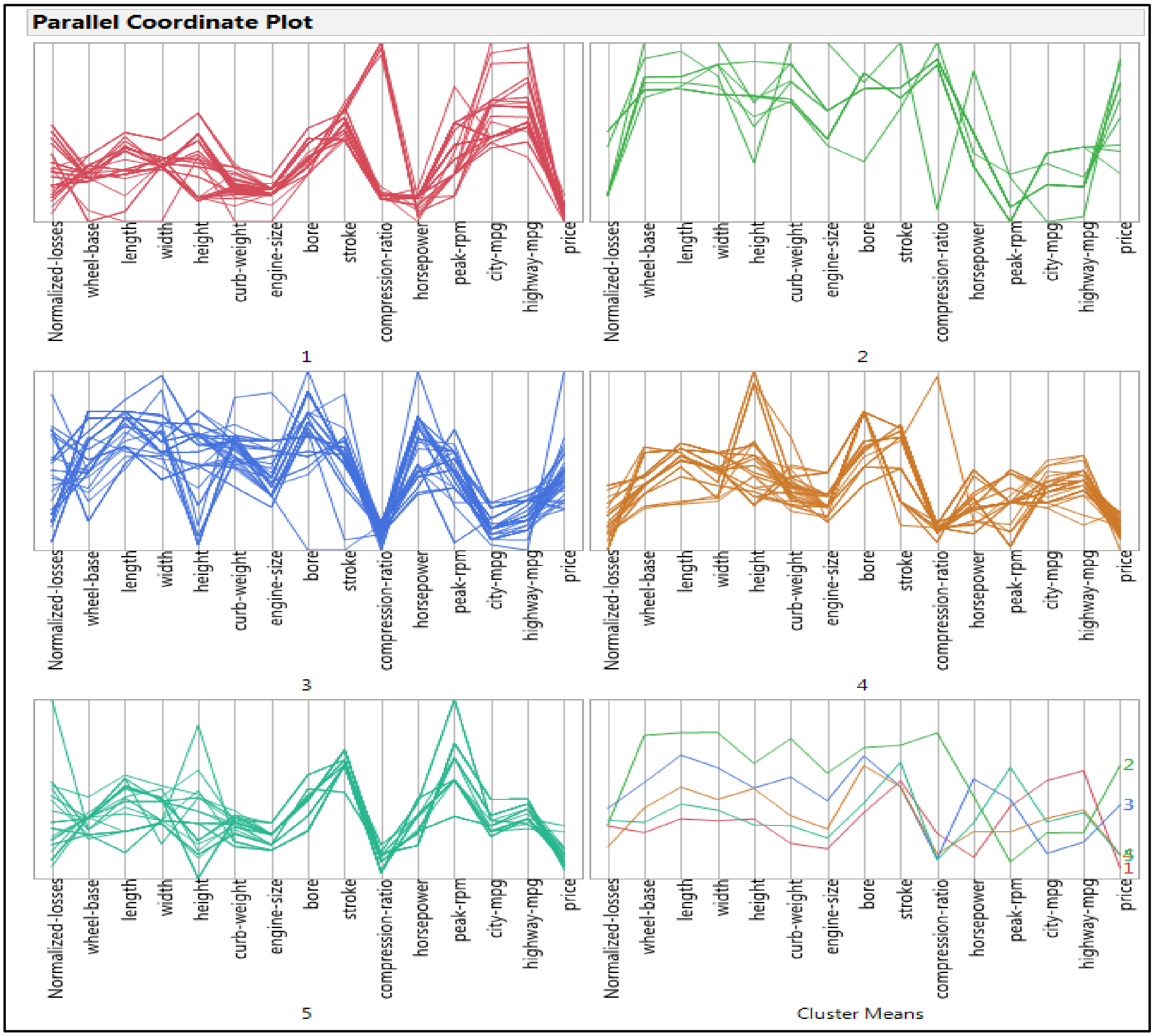
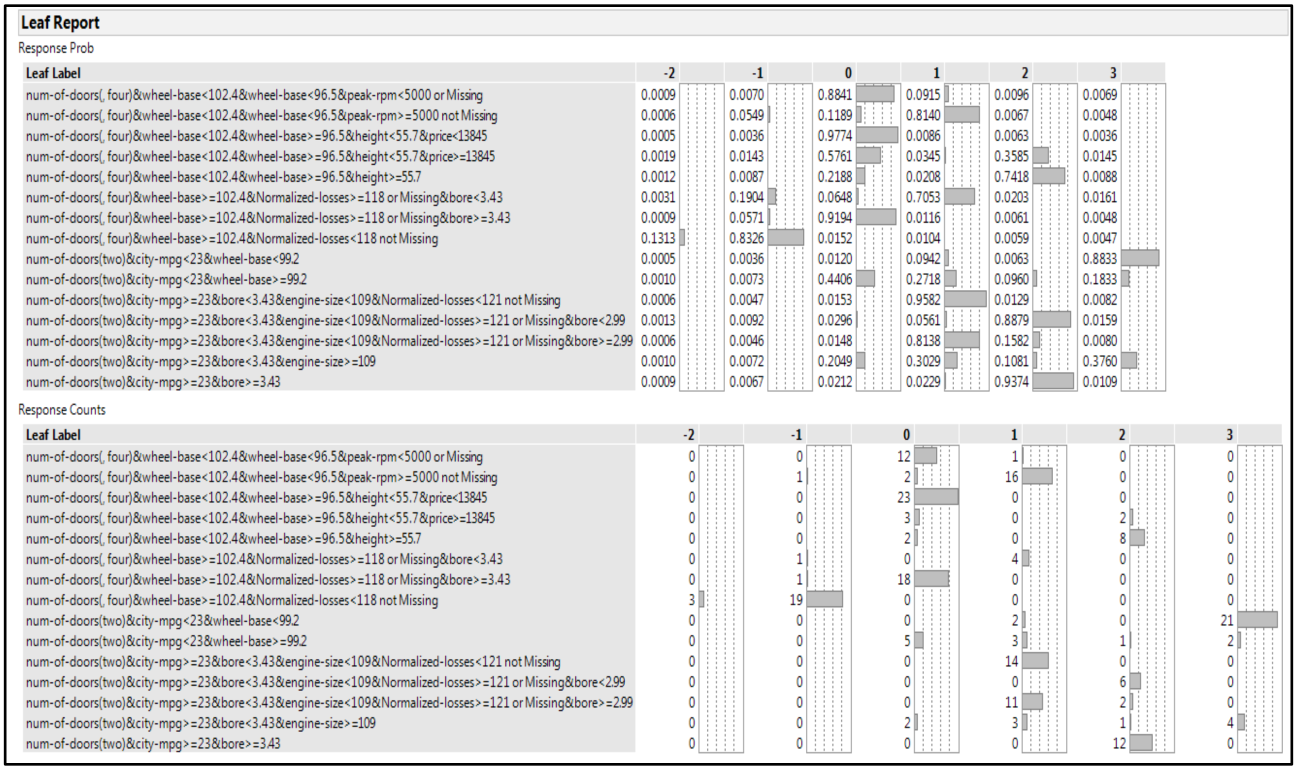


Figure 14: Parallel Coordinate Plot

Decision Tree

Decision Tree analysis was performed in order to support prior findings and uncover any nonlinear relationships of which may have been missed in prior analyses. The decision tree ended with 14 splits and contained an R-squared of 0.743. The analysis was stopped at 14 splits due to the small and insignificant increases in the R-squared value (increases were under .01) and splits after 14 had no significant G^2 effects on the leaves. A leaf report of the decision tree is seen in Figure 15. Findings from the leaf report uncovered relationships between variables that were unnoticed in prior analyses. The major variable of which affected risk was the number of doors. The leaf report revealed that any automobile which had four doors, a wheel-base greater than 102.4, and a normalized-loss greater than 118 posed a significant increase in risk when compared to an automobile with only two doors. Prior analyses revealed width and normalized loss possessed a relationship with an increase in risk. This is the first analysis of which suggested the number of doors may have an increased importance when calculating risk.



Leaf Report

Response Prob

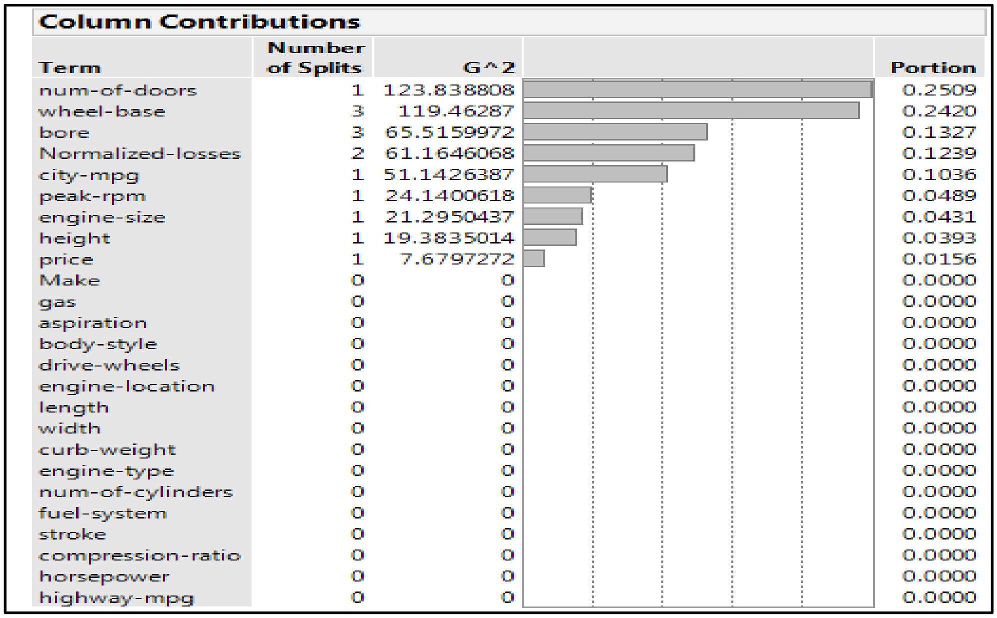
Leaf Label	-2	-1	0	1	2	3
num-of-doors(, four)&wheel-base<102.4&wheel-base<96.5&peak-rpm<5000 or Missing	0.0009	0.0070	0.8841	0.0915	0.0096	0.0069
num-of-doors(, four)&wheel-base<102.4&wheel-base<96.5&peak-rpm>=5000 not Missing	0.0006	0.0549	0.1189	0.8140	0.0067	0.0048
num-of-doors(, four)&wheel-base<102.4&wheel-base>=96.5&height<55.7&price<13845	0.0005	0.0036	0.9774	0.0086	0.0063	0.0036
num-of-doors(, four)&wheel-base<102.4&wheel-base>=96.5&height<55.7&price>=13845	0.0019	0.0143	0.5761	0.0345	0.3585	0.0145
num-of-doors(, four)&wheel-base<102.4&wheel-base>=96.5&height>=55.7	0.0012	0.0087	0.2188	0.0208	0.7418	0.0088
num-of-doors(, four)&wheel-base>=102.4&Normalized-losses>=118 or Missing&bore<3.43	0.0031	0.1904	0.0648	0.7053	0.0203	0.0161
num-of-doors(, four)&wheel-base>=102.4&Normalized-losses>=118 or Missing&bore>=3.43	0.0009	0.0571	0.9194	0.0116	0.0061	0.0048
num-of-doors(, four)&wheel-base>=102.4&Normalized-losses<118 not Missing	0.1313	0.8326	0.0152	0.0104	0.0059	0.0047
num-of-doors(two)&city-mpg<23&wheel-base<99.2	0.0005	0.0036	0.0120	0.0942	0.0063	0.8833
num-of-doors(two)&city-mpg<23&wheel-base>=99.2	0.0010	0.0073	0.4406	0.2718	0.0960	0.1833
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size<109&Normalized-losses<121 not Missing	0.0006	0.0047	0.0153	0.9582	0.0129	0.0082
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size<109&Normalized-losses>=121 or Missing&bore<2.99	0.0013	0.0092	0.0296	0.0561	0.8879	0.0159
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size<109&Normalized-losses>=121 or Missing&bore>=2.99	0.0006	0.0046	0.0148	0.8138	0.1582	0.0080
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size>=109	0.0010	0.0072	0.2049	0.3029	0.1081	0.3760
num-of-doors(two)&city-mpg>=23&bore>=3.43	0.0009	0.0067	0.0212	0.0229	0.9374	0.0109

Response Counts

Leaf Label	-2	-1	0	1	2	3
num-of-doors(, four)&wheel-base<102.4&wheel-base<96.5&peak-rpm<5000 or Missing	0	0	12	1	0	0
num-of-doors(, four)&wheel-base<102.4&wheel-base<96.5&peak-rpm>=5000 not Missing	0	1	2	16	0	0
num-of-doors(, four)&wheel-base<102.4&wheel-base>=96.5&height<55.7&price<13845	0	0	23	0	0	0
num-of-doors(, four)&wheel-base<102.4&wheel-base>=96.5&height<55.7&price>=13845	0	0	3	0	2	0
num-of-doors(, four)&wheel-base<102.4&wheel-base>=96.5&height>=55.7	0	0	2	0	8	0
num-of-doors(, four)&wheel-base>=102.4&Normalized-losses>=118 or Missing&bore<3.43	0	1	0	4	0	0
num-of-doors(, four)&wheel-base>=102.4&Normalized-losses>=118 or Missing&bore>=3.43	0	1	18	0	0	0
num-of-doors(, four)&wheel-base>=102.4&Normalized-losses<118 not Missing	3	19	0	0	0	0
num-of-doors(two)&city-mpg<23&wheel-base<99.2	0	0	0	2	0	21
num-of-doors(two)&city-mpg<23&wheel-base>=99.2	0	0	5	3	1	2
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size<109&Normalized-losses<121 not Missing	0	0	0	14	0	0
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size<109&Normalized-losses>=121 or Missing&bore<2.99	0	0	0	0	6	0
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size<109&Normalized-losses>=121 or Missing&bore>=2.99	0	0	0	11	2	0
num-of-doors(two)&city-mpg>=23&bore<3.43&engine-size>=109	0	0	2	3	1	4
num-of-doors(two)&city-mpg>=23&bore>=3.43	0	0	0	0	12	0

Figure 15: Leaf Report for Decision Tree

Figure 16 shows the contributing columns report, which depicts how each variable contributed to the overall results of the decision tree. The top two variables were number-of-doors and wheel-base. Number-of-doors contained a G^2 of 123.84 and wheel-base contained a G^2 of 119.46. The lowest variable on the contributing columns report is price and contained a G^2 of 7.70.



Column Contributions

Term	Number of Splits	G^2	Portion
num-of-doors	1	123.838808	0.2509
wheel-base	3	119.46287	0.2420
bore	3	65.5159972	0.1327
Normalized-losses	2	61.1646068	0.1239
city-mpg	1	51.1426387	0.1036
peak-rpm	1	24.1400618	0.0489
engine-size	1	21.2950437	0.0431
height	1	19.3835014	0.0393
price	1	7.6797272	0.0156
Make	0	0	0.0000
gas	0	0	0.0000
aspiration	0	0	0.0000
body-style	0	0	0.0000
drive-wheels	0	0	0.0000
engine-location	0	0	0.0000
length	0	0	0.0000
width	0	0	0.0000
curb-weight	0	0	0.0000
engine-type	0	0	0.0000
num-of-cylinders	0	0	0.0000
fuel-system	0	0	0.0000
stroke	0	0	0.0000
compression-ratio	0	0	0.0000
horsepower	0	0	0.0000
highway-mpg	0	0	0.0000

Figure 16: Contributing Columns Report

The analysis concludes with by creating a ROC and Lift curve. The ROC curve, Figure 17, allows an individual to assess the accuracy of the decision tree. All results in the ROC curve are over 0.95 which suggests the decision tree accurately represents the variables and their relationship with risk. The Lift curve, Figure 18, allows an individual to analyze the performance of the decision tree based on the variables that influence the splitting criterion.

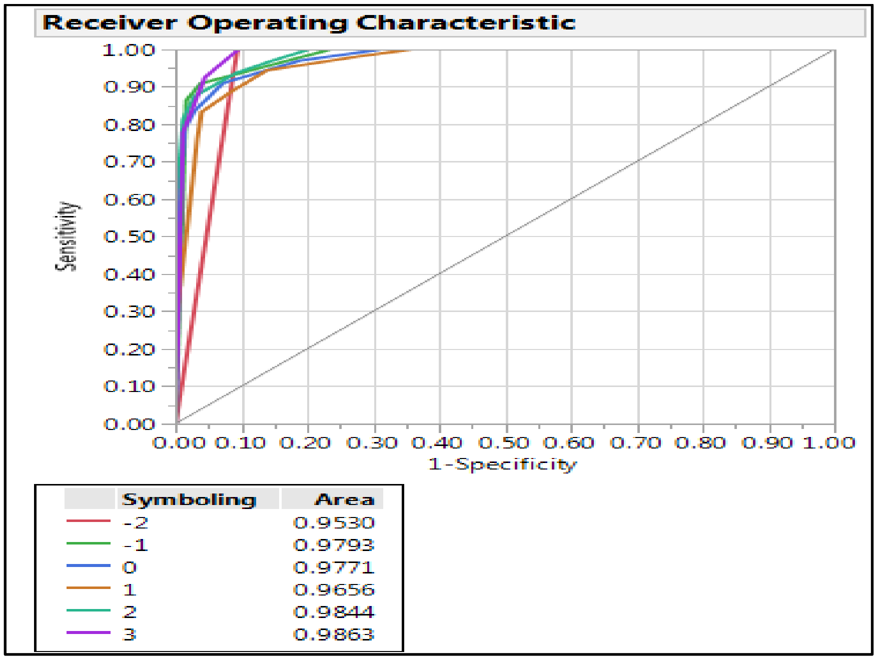


Figure 17: Decision tree ROC curve

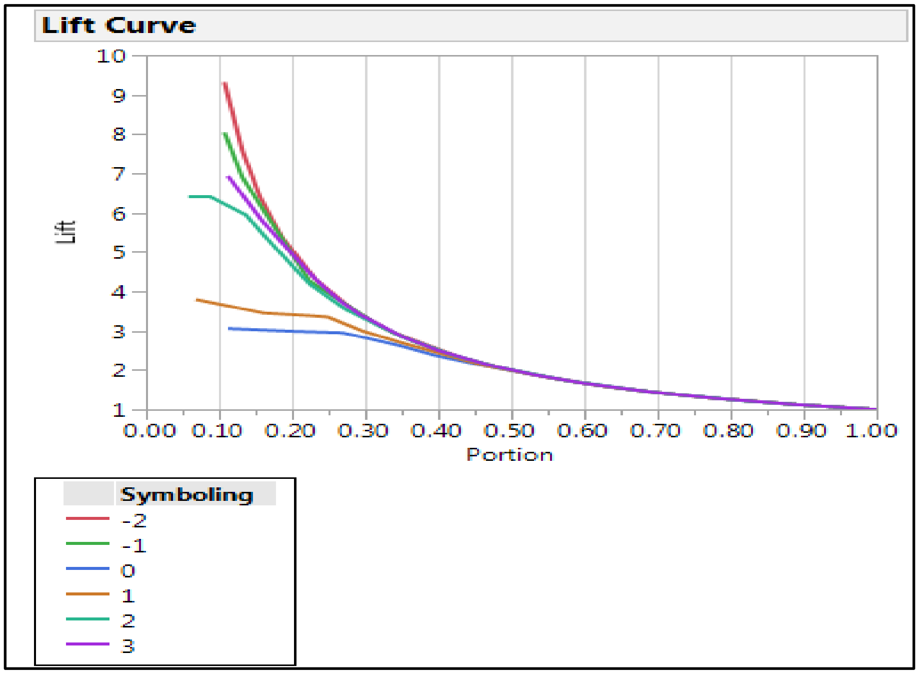


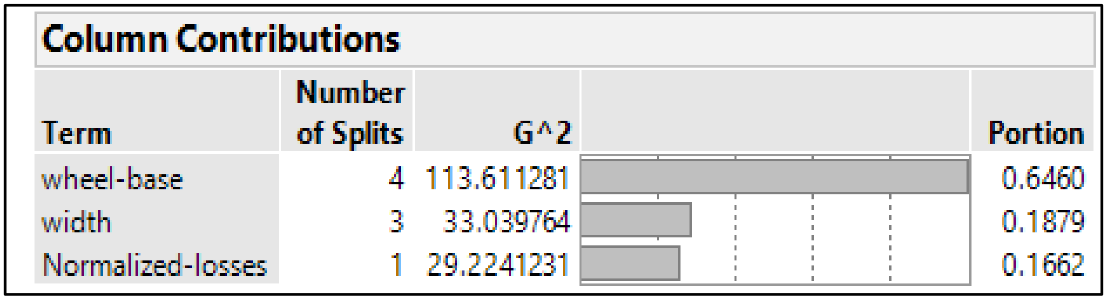
Figure 18: Decision Tree Lift Curve

Model Comparison Data Conversion

Converting the variable symbolizing (risk) into a binary variable was done by splitting the variable into two descriptive characteristics, More Risk and Less Risk. More risk was classified by a symbolizing value of 1 or higher, and less risk was classified as a symbolizing value of 0 or lower. By converting this variable into 2 characteristics, predictive models using decision trees, neural networks, and logistic regression were able to be built. The new variable was created as "Risk Predictor". When attempting to predict this variable, the highest correlating variables were used as input. Within our dataset, these variables were Normalized Losses, Width, and Wheelbase taken from the effects test of multiple regression. There were also several vehicles missing the data for normalized-losses. When conducting our analysis on the converted data, these vehicles were excluded from the analysis. A significant downfall of excluding this data is that there was no value stored for all three Alfa-Romeo models within the data set. This particular model was shown to have a significant effect on our data set as producing vehicles with high-risk factors.

Decision Tree - Binary (Training)

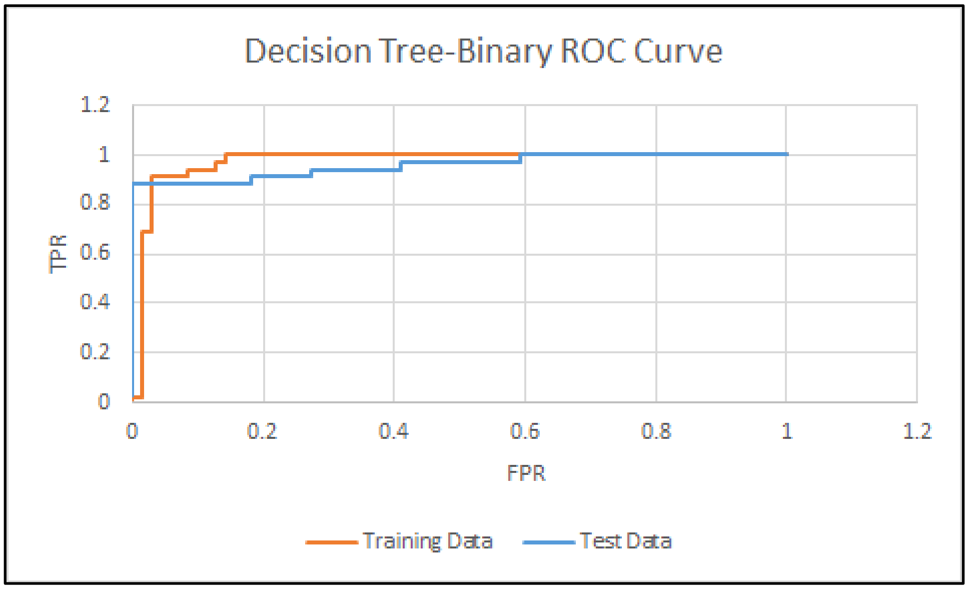
The binary decision tree of Risk Factor seen in Appendix B finds the best fit at 8 splits and has an R-squared value of 0.766. Further splitting does not result in a significant increase in the R-squared value, indicating further splits would result in overfitting. The data was primarily split on the value of wheel-base where it resulted in 4 of our 8 splits. One-third of our data was divided on the first split. This indicates vehicles with a wheelbase of fewer than 95.7 inches are likely to be considered a "less risk" vehicle. We can see on the Leaf Report, located in Appendix B, that the first split had a 97% probability to split the data on this criterion and resulted in a near perfect split. Normalized losses and width contained three and one splits per variable, respectively, which can be seen in Appendix B. Figure 19 shows the column contributions for wheel-base, width, and normalized loss. Wheel-base has a G^2 of 113.61 and accounts for almost 65 percent of the contribution. Width has a G^2 of 33.04 and accounts for 19 percent of the contribution. Lastly, normalized losses has a G^2 of 29.22 and accounts for 16 percent of the contribution. When examining the ROC curve in Figure 20, one can see the line for the training set is close to 90 degrees compared with the curve for the validation set. As we compared training with validation set, training has a higher AUC of 0.9738, and this high AUC is enough to be comfortable with the assumption that our decision tree model is properly predicting our output variable. When examining the Lift curve in Figure 21, an individual can see the lift for test data has a higher value than the training data set in the first 10 quantiles. Thus, test data, in this case, performs better than the training set.

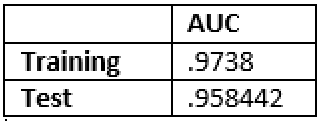


Column Contributions

Term	Number of Splits	G^2		Portion
wheel-base	4	113.611281		0.6460
width	3	33.039764		0.1879
Normalized-losses	1	29.2241231		0.1662

Figure 19: Decision Tree Binary Column Contributions (Training)





	AUC
Training	.9738
Test	.958442

Figure 20: Decision Tree-Binary ROC Curve

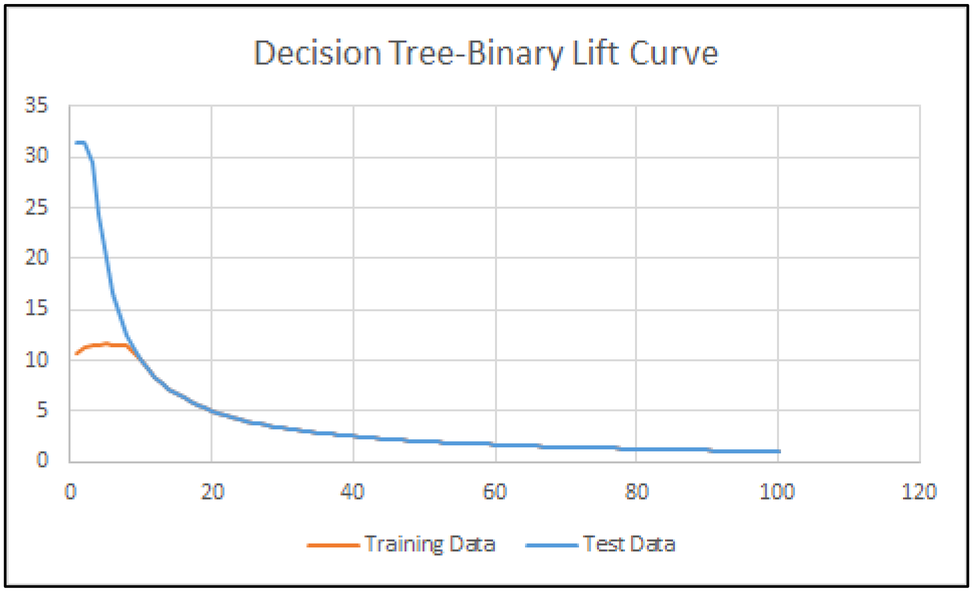
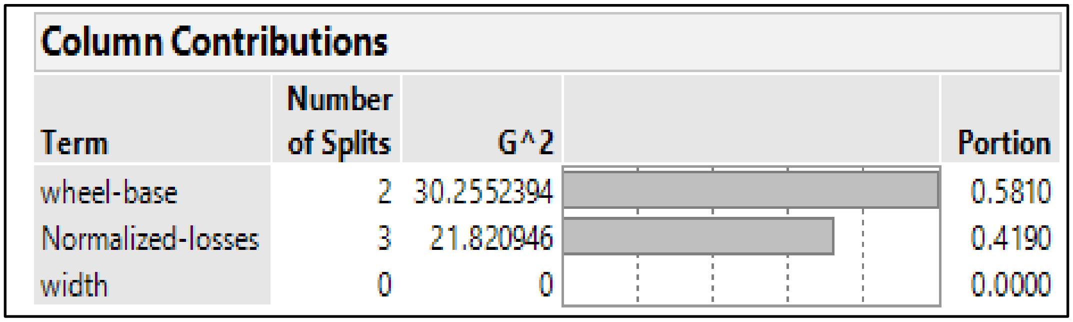


Figure 21: Decision Tree-Binary Lift Curve

Decision Tree - Binary (Holdout)

The binary decision tree of Risk Factor in Appendix C finds the best fit at 7 splits and has an R-squared value of 0.589. Half of the data was divided on the first split. This indicates vehicles with a normalized-losses greater than or equal to 103 are likely to be considered a "more risk" vehicle because insurance rates are higher amongst these vehicles. This tree compared to the Training data, had two splits based on wheelbase, three on Normalized Losses, and zero on width. Figure 22 represents the column contributions of wheel-base, normalized losses, and width. Wheel-base has a G^2 of 30.26 and accounts for roughly 58 percent of the contribution. Normalized losses have a G^2 of 21.82 and account for 42 percent of the contribution. Lastly, width did not account for any splits within this analysis.



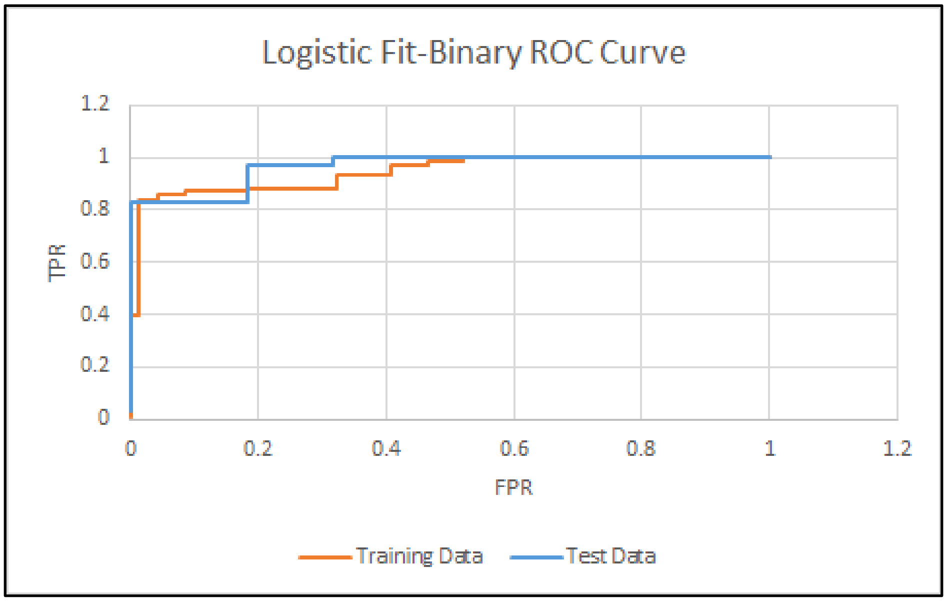
Column Contributions

Term	Number of Splits	G^2	Portion
wheel-base	2	30.2552394	0.5810
Normalized-losses	3	21.820946	0.4190
width	0	0	0.0000

Figure 22: Decision Tree Binary Column Contributions (Test)

Logistic Fit - Binary (Training)

Appendix D pictures the logistic fit of the binary variable against the top three input variables wheelbase, width, and normalized losses. The model revealed a whole model fit R Squared value of 0.72. This low fit value is likely the result of the input from the width variable, which returns a low LogWorth value of 5.136. The remaining variables wheelbase and normalized-loss return high LogWorth values and low p-values, thus they are more accurately predicting our binary variable. We can see on the ROC curve, in Figure 23, an AUC value of 0.943 showing that our model is strong when predicting the Risk Predictor variable. The ROC curve also reveals the line for training is relatively close to 90 degrees. With that and the AUC value of 0.943813 in mind, one can be comfortable with the assumption that our logistic fit model is properly predicting our output variable. From these results an individual can assume the current model is the optimal model and is properly predicting our binary Risk Predictor variable. Checking against the confusion matrix one can see the number of correct false predictions, 81, compared to the number of false positives, 8, which returns an FPR of 0.098. This is crucial when predicting the risk level of a vehicle and in our data, we incorrectly predicted 8 vehicles as being “less risk” when in fact they were actually “more risk”. A better model could be demanded to more accurately predict risky vehicles. An FPR of 10% is likely not going to be accepted by insurance companies when pricing premiums based on the wheelbase, width, and normalized loss of the vehicle. It is extremely important to consider a model with a lower FPR for this case because of the ramifications that can come from improperly classifying a vehicle as less risky, to both the consumer and the insurance companies. When examining the ROC curve for both training and validation set in Figure 23, we can see the validation ROC curve is closer to the perfect classifier, and it has a higher AUC of 0.964935 compared with the training set.



	AUC
Training	.943813
Test	.964935

Figure 23: Logistic Fit-Binary ROC Curve

When examining the Lift curve for both training and validation set in Figure 24, an individual can see the lift for test data has a higher value than the training data set in the first 10 quantiles. Thus, test data, in this case, performs better than the training set.

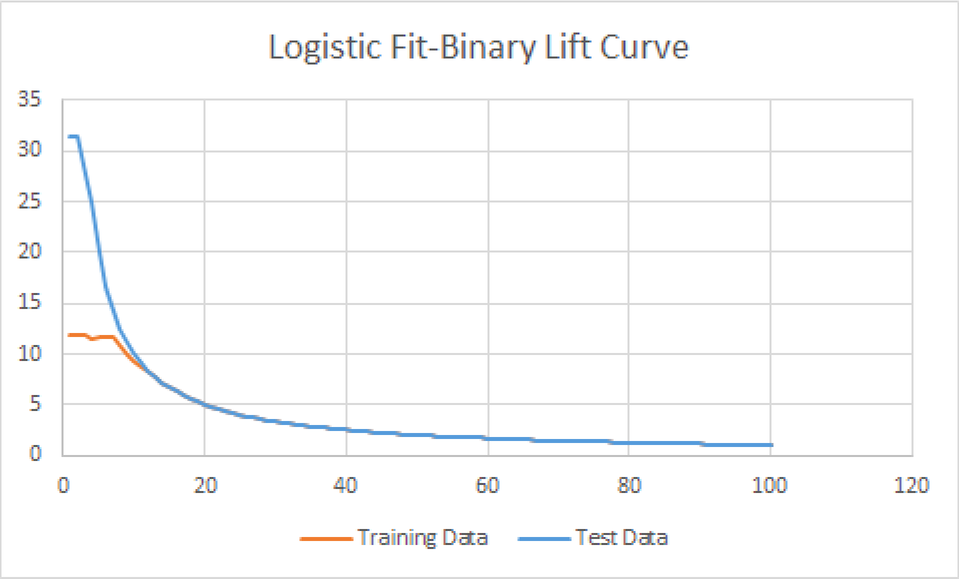


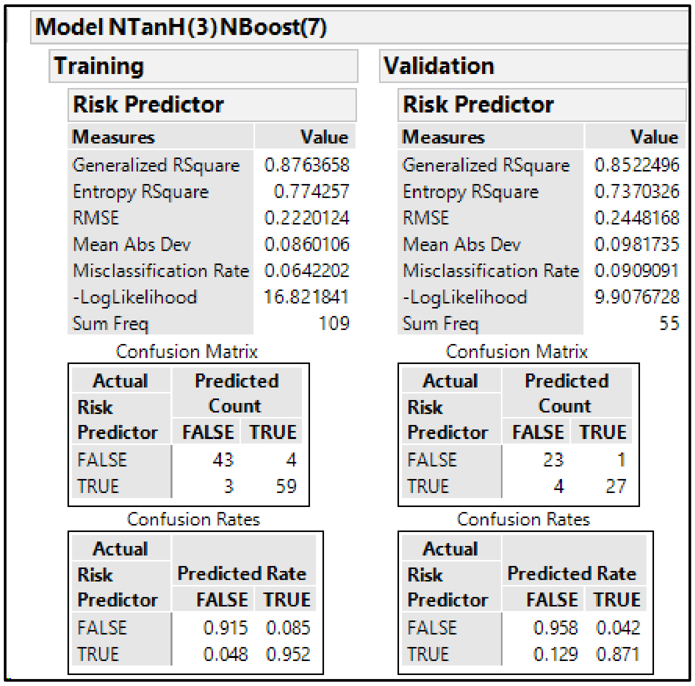
Figure 24: Logistic Fit-Binary Lift Curve

Logistic Fit - Binary (Holdout)

Appendix E shows the logistic fit of the binary variable against the top three input variables wheelbase, width, and normalized losses run against our one-third holdout, or test, data. The model was found to have a whole model fit R Squared of 0.64. This low fit value is likely the result of the input from the width variable, which returns a low LogWorth value of 1.184, confirming the results found in the training data. The remaining variables wheelbase and normalized loss return high LogWorth values and low p-values, thus they are more accurately predicting our binary variable. Checking against the confusion matrix one can see the number of correct false predictions, 29, compared to the number of false positives, 4, which returns an FPR of 0.18. This is crucial when predicting the risk level of a vehicle and in our data, we incorrectly predicted 8 vehicles as being “less risk” when in fact they were actually “more risk” and the model returned an FPR of 10%. As previously mentioned, it is extremely important to consider a model with a lower FPR for this case because of the ramifications that can come from improperly classifying a vehicle as less risky, to both the consumer and the insurance companies.

Neural networking - Binary (Training)

Construction of the binary neural network model was done with the same variables as the logistic regression and classification tree models, with wheelbase, width, and normalized losses as input and risk predictor as the output variable. The neural network was run ten times to allow us to compare outputs and determine if there was any high fluctuation in results. Overall, the variability of results was low, and the experiment found an average R-squared of 84.8 for the training data and an average R-squared of 79.6 for the test data. From the low variability and high R-squared in both the training and test data, one can assume successful and reliable results. Most of the tests resulted in a higher R-squared for the training data when compared to the test data. The goal is to have the best-fit model, and because of this, the NN model with the highest R-squared and of which had a supporting validation was chosen and is shown in Figure 25, as well as it’s respective confusion matrix. Next, the chosen model was used to analyze the ROC and Lift Curve. When examining the ROC curve for both training and validation set in Figure 26, we can see the training ROC curve is closer to the perfect classifier, and it has a higher AUC of 0.983644 compared with the validation set. When examining the Lift curve for both training and validation set in Figure 27, an individual can see the lift for test data has a higher value than the training data set in the first 10 quantiles.



Model NTanH(3)NBoost(7)

Training

Risk Predictor

Measures	Value
Generalized RSquare	0.8763658
Entropy RSquare	0.774257
RMSE	0.2220124
Mean Abs Dev	0.0860106
Misclassification Rate	0.0642202
-LogLikelihood	16.821841
Sum Freq	109

Confusion Matrix

Actual	Predicted	
Risk	Count	
Predictor	FALSE	TRUE
FALSE	43	4
TRUE	3	59

Confusion Rates

Actual		
Risk	Predicted Rate	
Predictor	FALSE	TRUE
FALSE	0.915	0.085
TRUE	0.048	0.952

Validation

Risk Predictor

Measures	Value
Generalized RSquare	0.8522496
Entropy RSquare	0.7370326
RMSE	0.2448168
Mean Abs Dev	0.0981735
Misclassification Rate	0.0909091
-LogLikelihood	9.9076728
Sum Freq	55

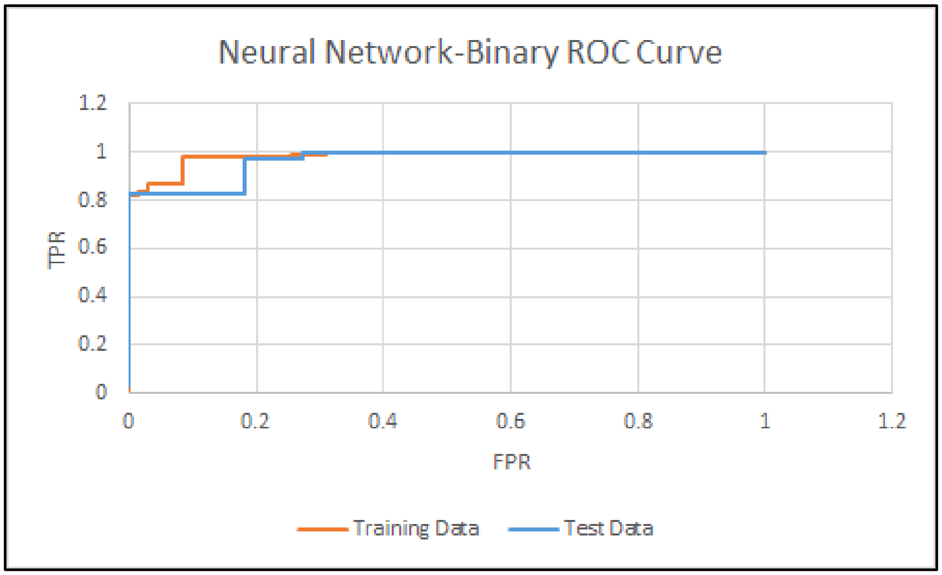
Confusion Matrix

Actual	Predicted	
Risk	Count	
Predictor	FALSE	TRUE
FALSE	23	1
TRUE	4	27

Confusion Rates

Actual		
Risk	Predicted Rate	
Predictor	FALSE	TRUE
FALSE	0.958	0.042
TRUE	0.129	0.871

Figure 25: Neural Network Results (Training and Validation)



	AUC
Training	.983644
Test	.966234

Figure 26: Neural Network-Binary ROC Curve

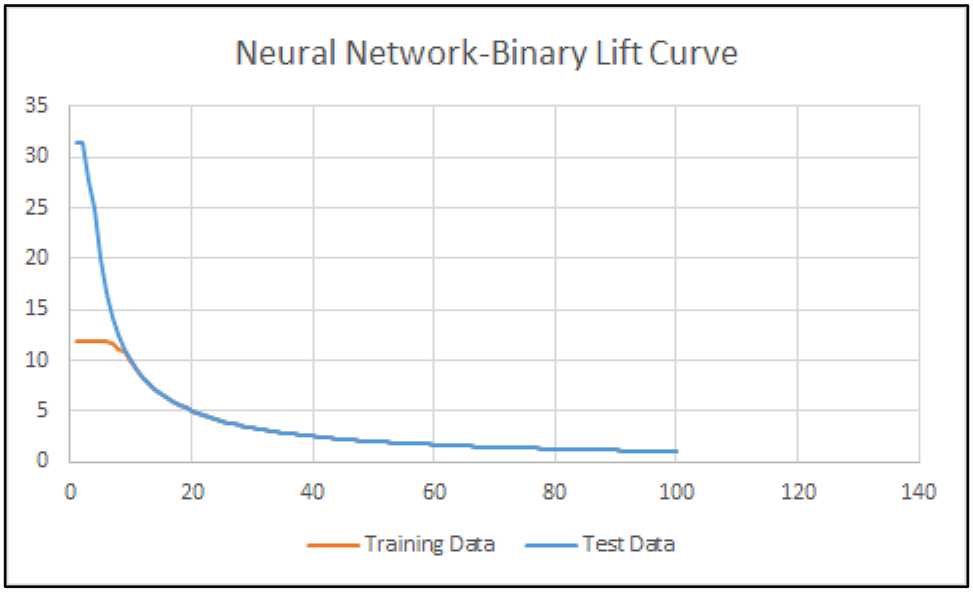
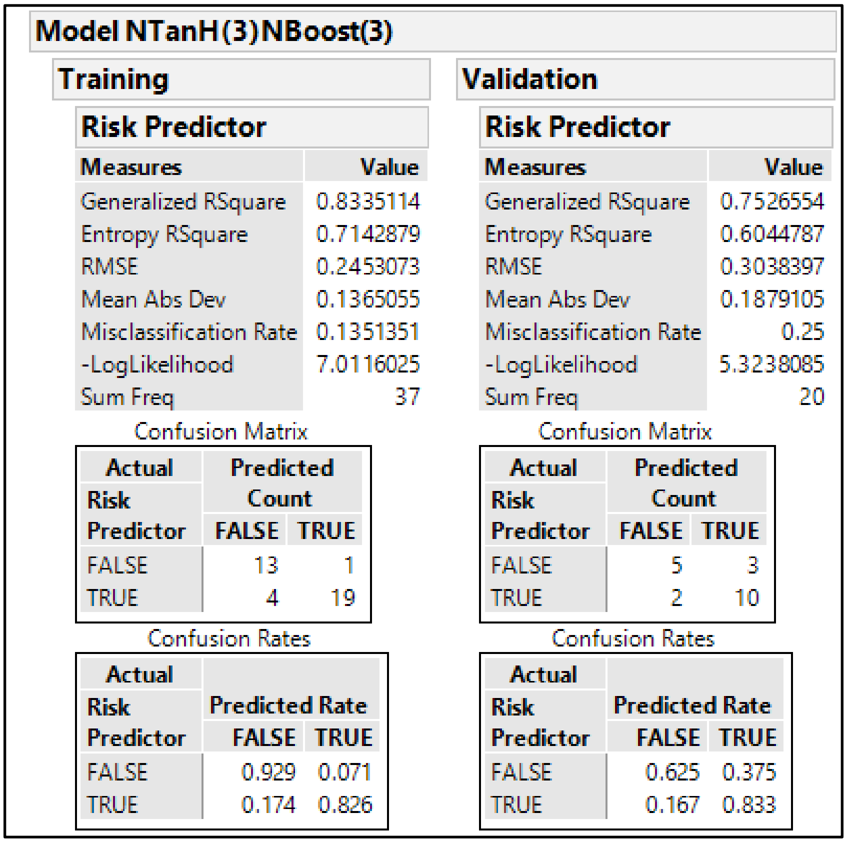


Figure 27: Neural Network-Binary Lift Curve

Neural networking - Binary (Holdout)

Construction of the binary neural network model was done with the same variables as the logistic regression and classification tree models, with wheelbase, width, and normalized losses as input and risk predictor as the output variable. The neural network was run ten times to allow for the comparison of outputs and to determine if there was any high fluctuation in results. From the low variances and high R-squared in both the training and test data, one can assume successful and reliable results. The majority of the tests resulted in a higher R-squared for the training data when compared to the test data. The training data of the experiment contained an R-squared anywhere from 0.75 to 0.8. The goal is to have the best-fit model, due to this goal, and therefore the model with the highest R-squared and of which had a supporting validation was chosen. An example of the test can be seen in Figure 28.



Model NTanH(3)NBoost(3)

Training

Risk Predictor

Measures	Value
Generalized RSquare	0.8335114
Entropy RSquare	0.7142879
RMSE	0.2453073
Mean Abs Dev	0.1365055
Misclassification Rate	0.1351351
-LogLikelihood	7.0116025
Sum Freq	37

Confusion Matrix

Actual	Predicted	
Risk	Count	
Predictor	FALSE	TRUE
FALSE	13	1
TRUE	4	19

Confusion Rates

Actual		
Risk	Predicted Rate	
Predictor	FALSE	TRUE
FALSE	0.929	0.071
TRUE	0.174	0.826

Validation

Risk Predictor

Measures	Value
Generalized RSquare	0.7526554
Entropy RSquare	0.6044787
RMSE	0.3038397
Mean Abs Dev	0.1879105
Misclassification Rate	0.25
-LogLikelihood	5.3238085
Sum Freq	20

Confusion Matrix

Actual	Predicted	
Risk	Count	
Predictor	FALSE	TRUE
FALSE	5	3
TRUE	2	10

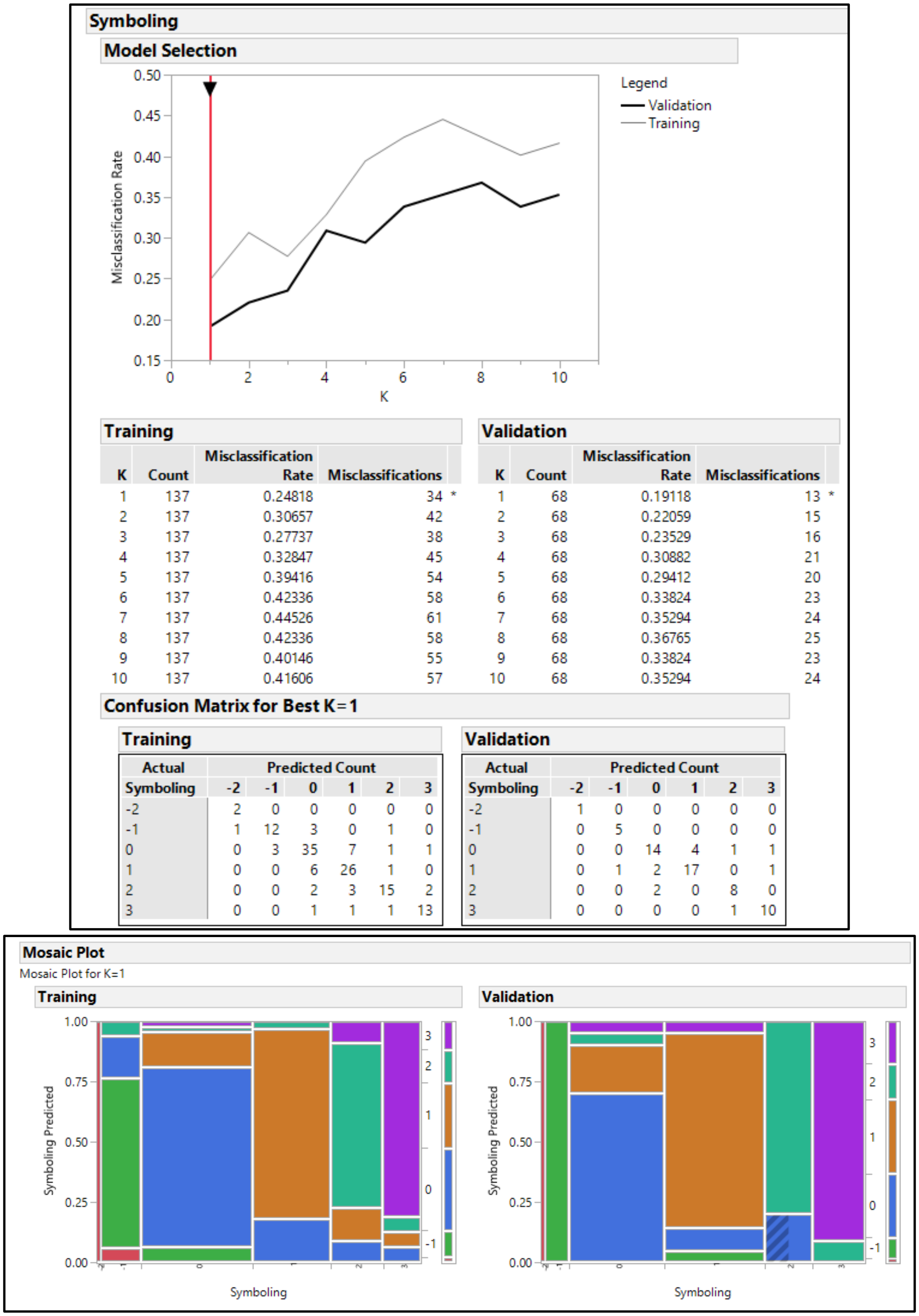
Confusion Rates

Actual		
Risk	Predicted Rate	
Predictor	FALSE	TRUE
FALSE	0.625	0.375
TRUE	0.167	0.833

Figure 28: Neural Network Results - Binary (Test and Validation)

K Nearest Neighbors (Training)

When attempting to conduct K Nearest Neighbors analysis, the highest correlating variables were used as input. Within our dataset, these variables were Normalized Losses, Width, and Wheelbase. As can be seen from model selection graph referring to Figure 29, the misclassification rate begins to have trend since k = 1, which ranges from 0.15 to 0.5. This indicates when k = 1, our model has the smallest misclassification rate, and this model should be considered as the best performer for our data set. When we look at training results, when k = 1, the results indicate misclassifications and misclassification rates are smallest, which are respectively 0.24818 and 34. When we check the validation result, the test set verifies that the single nearest neighbor model is the best performer for our data set. Looking at the confusion matrix for best k = 1 training set, the correct classifications are all on the main diagonal from upper left to lower right. We can see the model has trouble distinguishing between symbolizing level 0 and 1.Thus, the single nearest neighbor model is the best performer for the training dataset.



Symboling

Model Selection

Training

K	Count	Misclassification Rate	Misclassifications	
1	137	0.24818	34	*
2	137	0.30657	42	
3	137	0.27737	38	
4	137	0.32847	45	
5	137	0.39416	54	
6	137	0.42336	58	
7	137	0.44526	61	
8	137	0.42336	58	
9	137	0.40146	55	
10	137	0.41606	57	

Validation

K	Count	Misclassification Rate	Misclassifications	
1	68	0.19118	13	*
2	68	0.22059	15	
3	68	0.23529	16	
4	68	0.30882	21	
5	68	0.29412	20	
6	68	0.33824	23	
7	68	0.35294	24	
8	68	0.36765	25	
9	68	0.33824	23	
10	68	0.35294	24	

Confusion Matrix for Best K=1

Training

Actual	Predicted Count					
Symboling	-2	-1	0	1	2	3
-2	2	0	0	0	0	0
-1	1	12	3	0	1	0
0	0	3	35	7	1	1
1	0	0	6	26	1	0
2	0	0	2	3	15	2
3	0	0	1	1	1	13

Validation

Actual	Predicted Count					
Symboling	-2	-1	0	1	2	3
-2	1	0	0	0	0	0
-1	0	5	0	0	0	0
0	0	0	14	4	1	1
1	0	1	2	17	0	1
2	0	0	2	0	8	0
3	0	0	0	0	1	10

Mosaic Plot

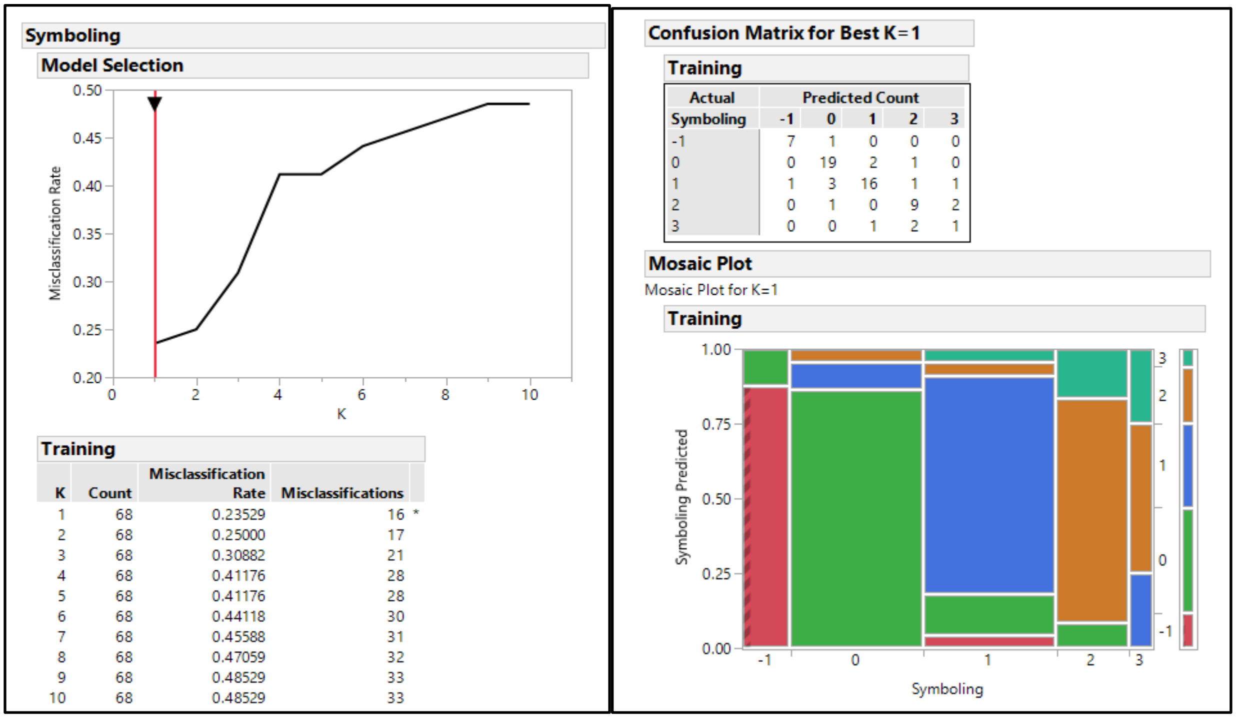
Training

Validation

Figure 29: K - Nearest Neighbors (Training)

K Nearest Neighbors (Holdout)

When conducting K Nearest Neighbors analysis the highest correlating variables were used as input as like our previous binary models. As can be seen from model selection graph referring to Figure 30, the misclassification rate begins to have trend since k = 1, which ranges from 0.15 to 0.5, and it the upward trend reaches its peak when k = 10. This indicates when k = 1, our model has the smallest misclassification rate, and this model should be considered as the best performer for our data set. Looking at the confusion matrix for best k = 1 training set, the correct classifications are all on the main diagonal from upper left to lower right. We can see the holdout data model performs well on predicting the true count for symbolizing because we have fewer data in holdout data set. Thus, the single nearest neighbor model is the best performer for both K Nearest Neighbors models.



Training

K	Count	Misclassification Rate	Misclassifications	
1	68	0.23529	16	*
2	68	0.25000	17	
3	68	0.30882	21	
4	68	0.41176	28	
5	68	0.41176	28	
6	68	0.44118	30	
7	68	0.45588	31	
8	68	0.47059	32	
9	68	0.48529	33	
10	68	0.48529	33	

Confusion Matrix for Best K=1

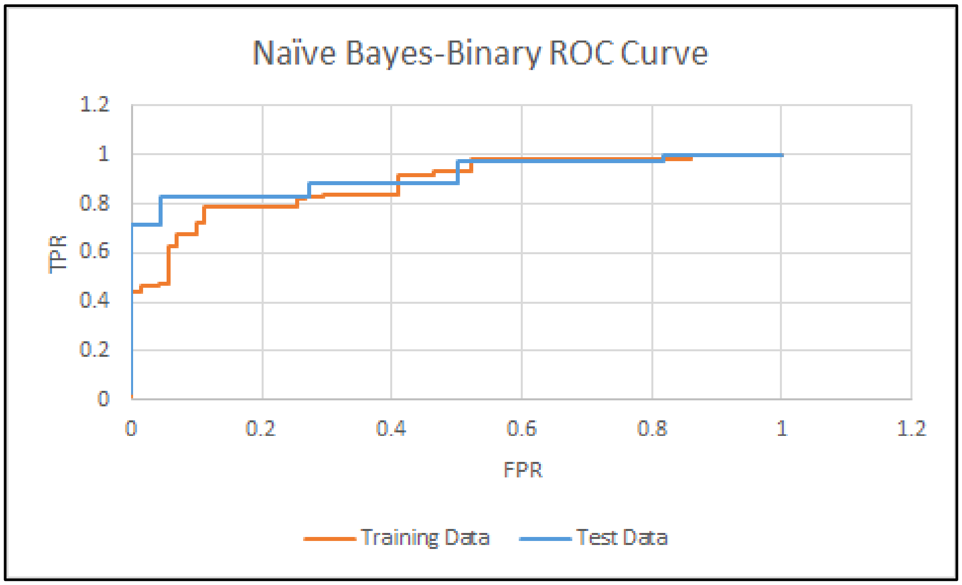
Training

Actual	Predicted Count				
Symboling	-1	0	1	2	3
-1	7	1	0	0	0
0	0	19	2	1	0
1	1	3	16	1	1
2	0	1	0	9	2
3	0	0	1	2	1

Figure 30: K-Nearest Neighbors (Holdout)

Naive Bayes Prediction

Naive Bayes predictions allow us to attempt to predict the outcome of whether the vehicle will be classified as "More Risk" or "Less Risk" based on the three primary influencing variables in our data set (Normalized Losses, Wheelbase, and Width). Using this method, it can be seen how accurately one can predict the classification of the vehicle and whether or not this method could potentially be recommended to insurance companies when predicting risk. The same Test and Training data sets from previous analysis were used, and any entry that was missing a value for "symbolizing" was excluded. The results of the analysis returned a ROC curve with an AUC of .8796 for the training data seen in Figure 31, while the test data returned a ROC of .9130 seen in Figure 31. Thus, we can be close to 90% confident in the predictive modeling ability of these three variables in relation to predicting risk. The lift curve for the training data and the holdout data in Figure 32 show the effectiveness of the model.



	AUC
Training	.8796
Test	.912987

Figure 31: Naive Bayes-Binary ROC Curve

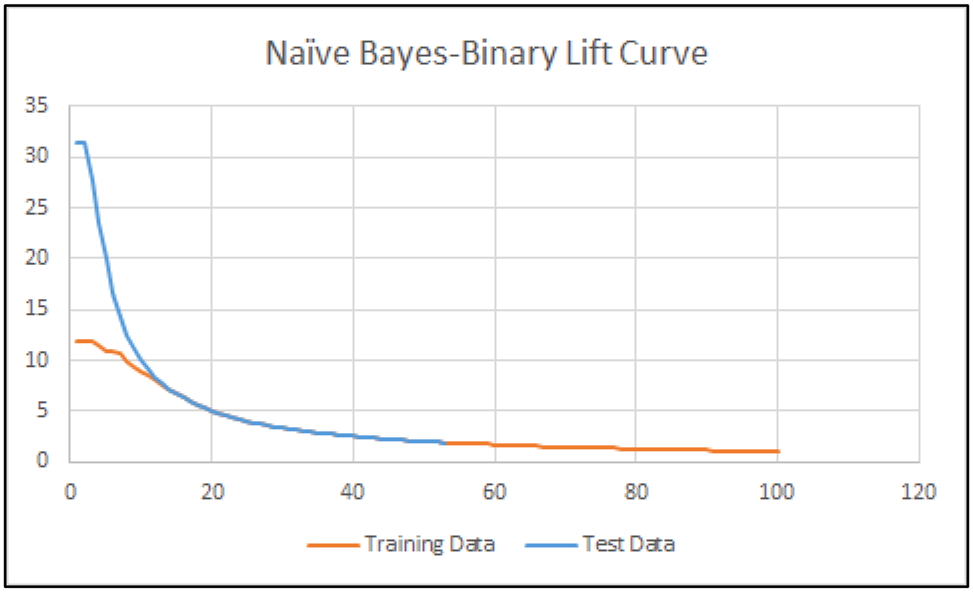
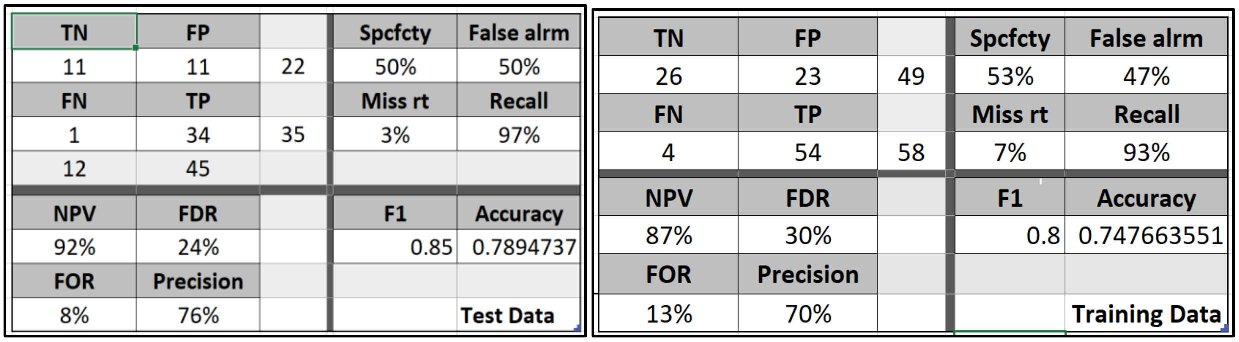


Figure 32: Naive Bayes-Binary Lift Curve

<u>Naive Bayes Model Comparison</u>

To fully understand the predictive ability of these three variables in relation to the response variable, the performance of the predictor must be calculated for training and test datasets. The test data set returned a specificity and false alarm of 50% and a miss rate of 3% and a recall of 97%. Thus, resulting in an accuracy rating of 79%. For the training data, the probability matrix returned a specificity of 53%, false alarm of 47% respectively, while miss rate returned 7% and recall of 93%. Resulting accuracy of training data was calculated to be 75% pictured in Figure 33. The models returned very similar values for all measurements and therefore we can be confident in the model's ability to predict the output variable and not overfit the data. While the accuracy rate for these tests returns a value lower than an insurance company might want to see if they were to use this model, they can take solace in the extremely low miss rate values returned from the model. The miss rate, in this case, would result as classifying a vehicle as "Less Risk" when in fact it is it "More Risk". These classifications would likely result in major losses for an insurance company due to the unaccounted risk. This model, however, is more biased to classify a vehicle as being "More Risk" when in fact it is "Less Risk", as seen by the high false alarm percentage. This is not ideal for a consumer, as it would likely lead to higher than necessary insurance premiums for a vehicle deemed riskier than it actually is. However, the insurance company would benefit from less legal liability in under classifying the safety of a vehicle.



TN	FP		Spcfcty	False alrm
11	11	22	50%	50%
FN	TP		Miss rt	Recall
1	34	35	3%	97%
12	45			
NPV	FDR		F1	Accuracy
92%	24%		0.85	0.7894737
FOR	Precision			
8%	76%			Test Data

TN	FP		Spcfcty	False alrm
26	23	49	53%	47%
FN	TP		Miss rt	Recall
4	54	58	7%	93%
NPV	FDR		F1	Accuracy
87%	30%		0.8	0.747663551
FOR	Precision			
13%	70%			Training Data

Figure 33: Performance Figures of Naive Bias

Validation and Training Model Comparison

After conducting a complete spectrum of binary variable analysis, we can compare the models against each other to determine which model should be recommended to best predict risk. As we can see from figure 34, on the left, the training ROC curves show that neural network and decision tree models compete closesly for the highest modeling statistics. However, the neural network is closer to the perfect classifier. In figure 37, we can see neural network has the highest AUC, 0.983644, compared to other models from the training dataset. On the right, looking at the validation ROC curves, all the four models are relatively close to 90 degrees. However, as we compared each models AUC individually, neural network had returned the highest AUC of 0.966234.

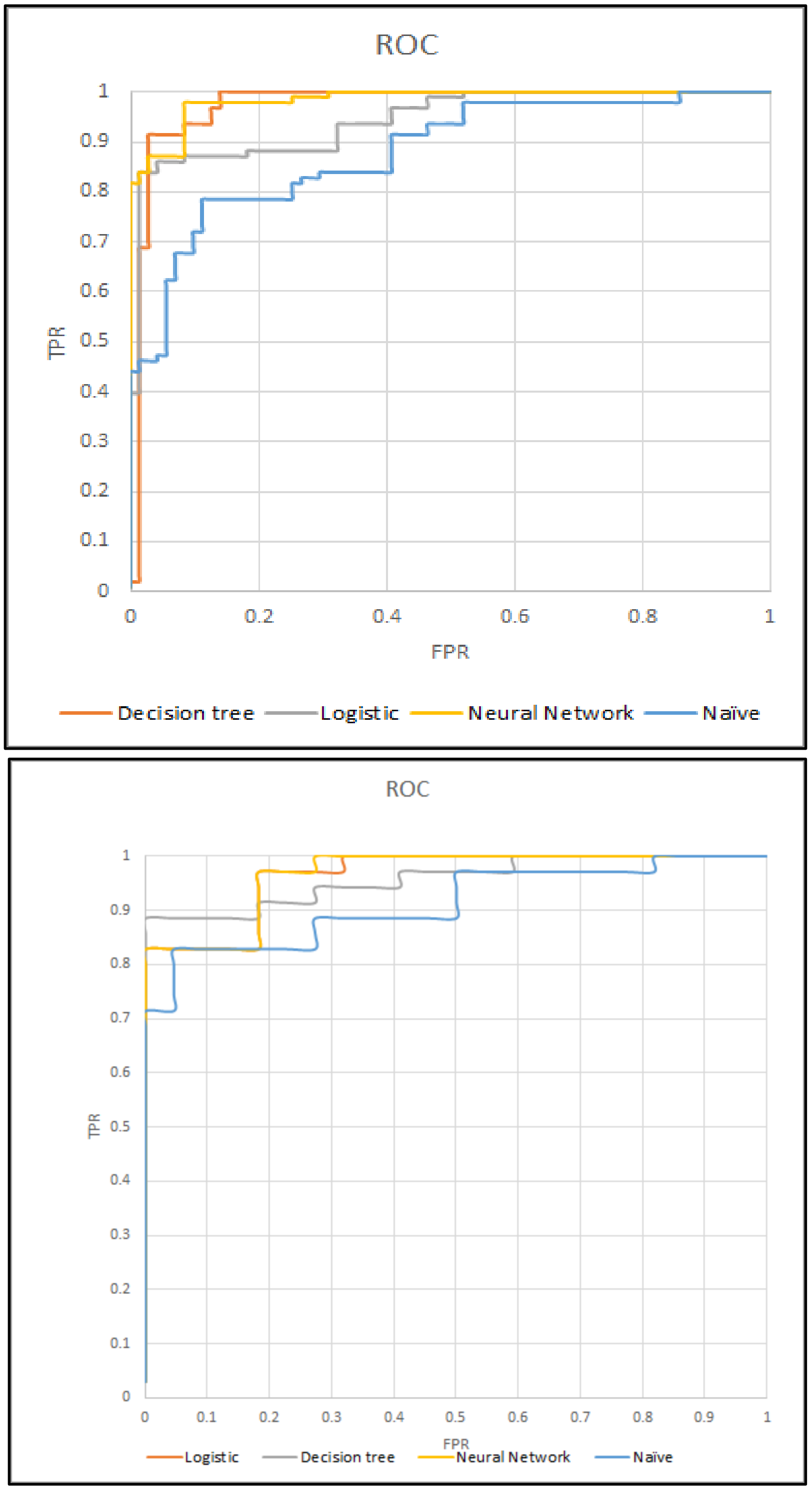


Figure 34: ROC Curves Comparison (Training on the left, validation on the right)

On the left of Figure 35, by comparing different training lift curves, neural network has the highest lift in the first 10 quantiles. So, suppose we need to know the car that is that is likely to have more risk, and 10% of positive responses are expected. Based on this figure, neural network is the superior option as it has the highest lift compared to any of the other models. On the right of Figure 35, for validation lift curves, since there are just a few datasets for the holdout the lift curves for all four models tend to be very similar.

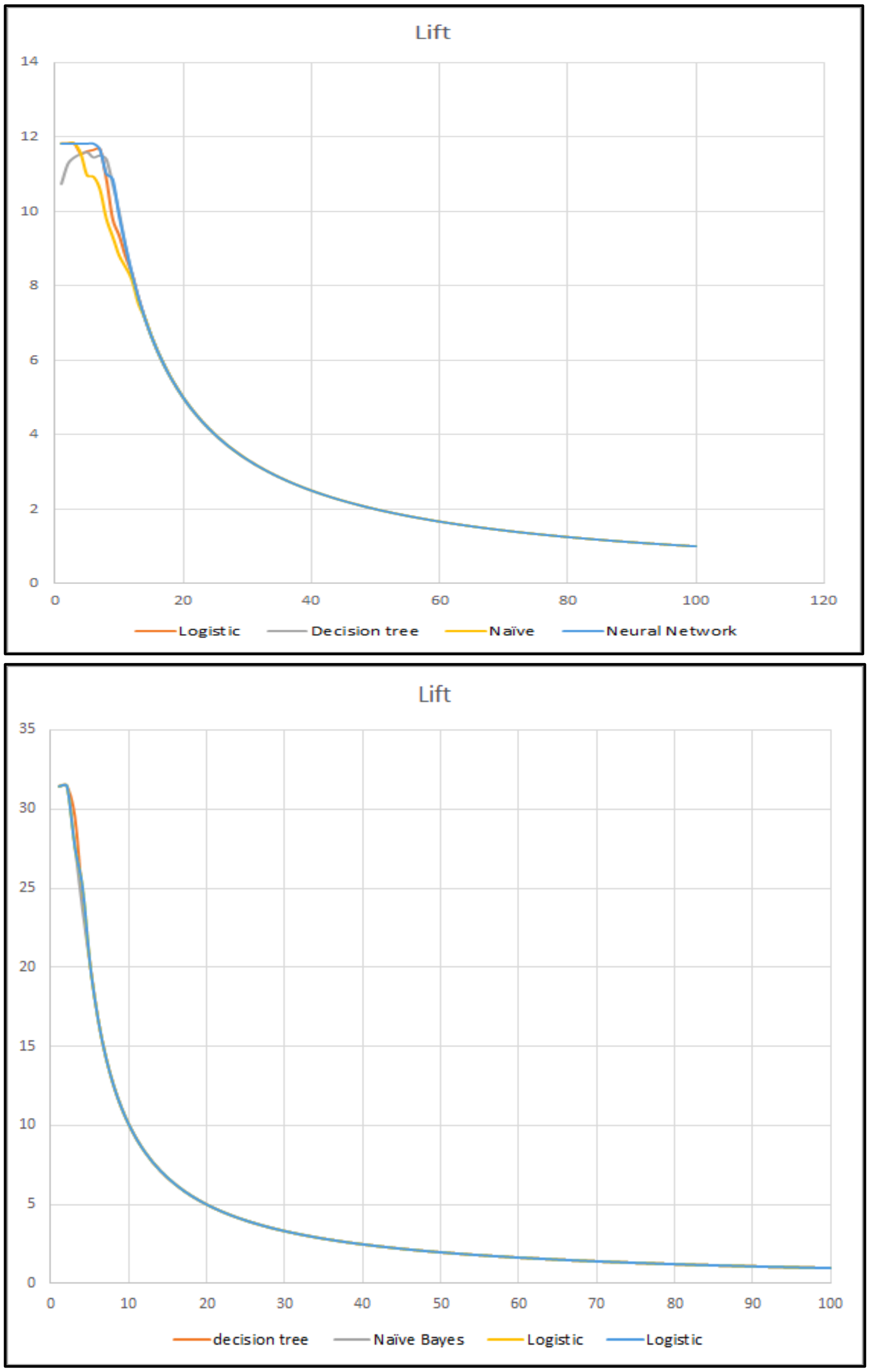
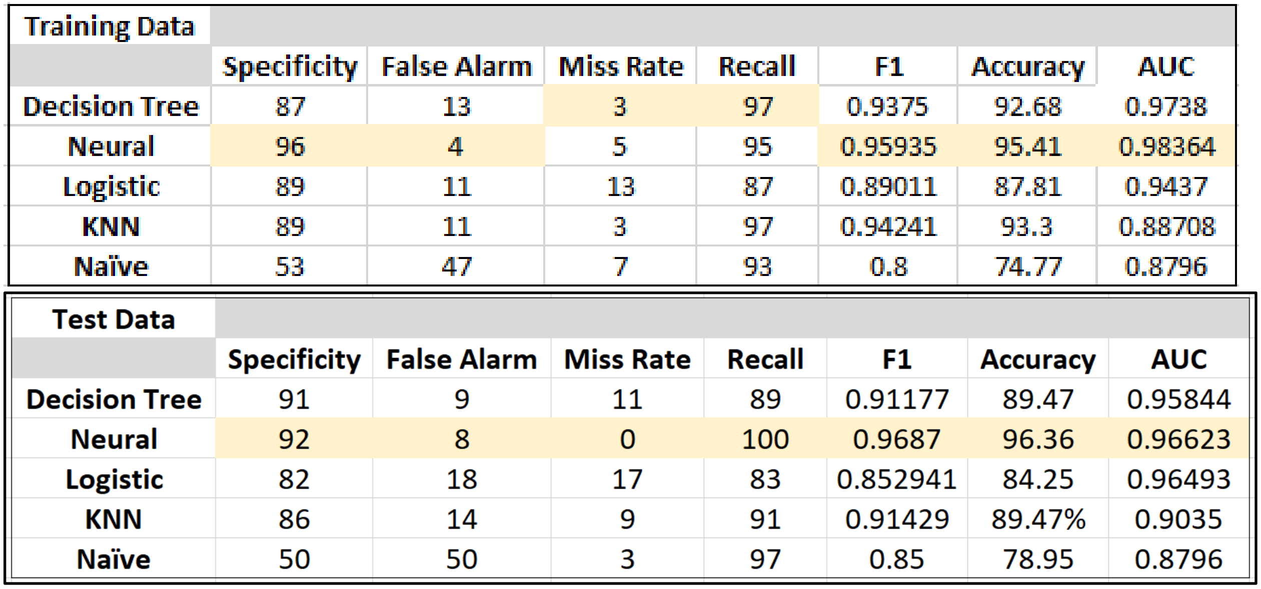


Figure 35: Lift Curves Comparison (Training on the left, validation on the right)

After completing all relevant tests, a comparative performance table was built to compare all five binary models based on training data and test data. In Figure 36 one can see a model comparison between the decision tree, neural network, logistic regression, K Nearest Neighbors, and Naive Bayes analysis. The decision tree analysis resulted in an accuracy of 92.6, precision of 91 and a specificity of 87 percent. The neural network resulted in an accuracy of 95%, precision of 97 and a specificity of 96%. The logistic regression resulted in an accuracy of 87, precision of 91 and a specificity of 89%. KNN returned accuracy of 93 and an F1 of .943, and specificity of 89%. Naive Bayes returned extremely lopsided results when comparing False Alarm and Specificity, 53 and 47 for Training and 50:50 for Test data respectively. This analysis resulted in the lowest accuracy, F1 statistic, and ROC from any of our binary comparison models. The neural network analysis returned a 5% miss rate, .959 F1 rate, and an AUC of .98364. This is the model that we feel best predicts our data, even though it falls second to the decision tree analysis results in Miss Rate, Recall, and AUC. This is substantiated by the fact that the test and validation data for NN returned higher R Squared for our model, while the decision tree returned lower R Squared than the training data, allowing us to conclude that the Neural model is more accurate and less prone to overfitting and random error. We can also see below that the test data returned the higher performance statistics than any of the other models. Thus, the model that we would recommend to any firm or agency to predict risk based on the three primary influencing factors would be our Neural Network model.



Training Data							
	Specificity	False Alarm	Miss Rate	Recall	F1	Accuracy	AUC
Decision Tree	87	13	3	97	0.9375	92.68	0.9738
Neural	96	4	5	95	0.95935	95.41	0.98364
Logistic	89	11	13	87	0.89011	87.81	0.9437
KNN	89	11	3	97	0.94241	93.3	0.88708
Naïve	53	47	7	93	0.8	74.77	0.8796

Test Data							
	Specificity	False Alarm	Miss Rate	Recall	F1	Accuracy	AUC
Decision Tree	91	9	11	89	0.91177	89.47	0.95844
Neural	92	8	0	100	0.9687	96.36	0.96623
Logistic	82	18	17	83	0.852941	84.25	0.96493
KNN	86	14	9	91	0.91429	89.47%	0.9035
Naïve	50	50	3	97	0.85	78.95	0.8796

Figure 36: Training and Validation Model Comparison

Conclusion

If an individual would compare the navigational roads and how individuals use these roads when the first automobile was invented to today's society, one would see drastic changes. These changes include but are not limited to the number of individuals operating an automobile, size of automobiles, distance driven, how fast these automobiles can go etc. Due to these drastic changes, an individual could assume more risk is associated while operating an automobile. The current analysis found a significant increase in risk when examining many characteristics of an automobile.

An individual has the potential to examine an automobile and hypothesize if there is more or less risk associated with its characteristics. The current analysis was extremely interesting and intriguing because we were able to identify the specific characteristics of which are associated with an increase in risk. We were able to identify these specific characteristics by conducting statistical analysis tests such as regressions, ANOVA, PCA, scatter plots, neural networks and many more. The most interesting aspect of the analysis is the process used to verify the results. Seen in the table and figure section, an individual can identify multiple confirmation analyses to the overall results such as residual plots, Levene tests, test for equal variance, Welch's and many more. By conducting multiple statistical tests, reliable and valid results were able to be presented.

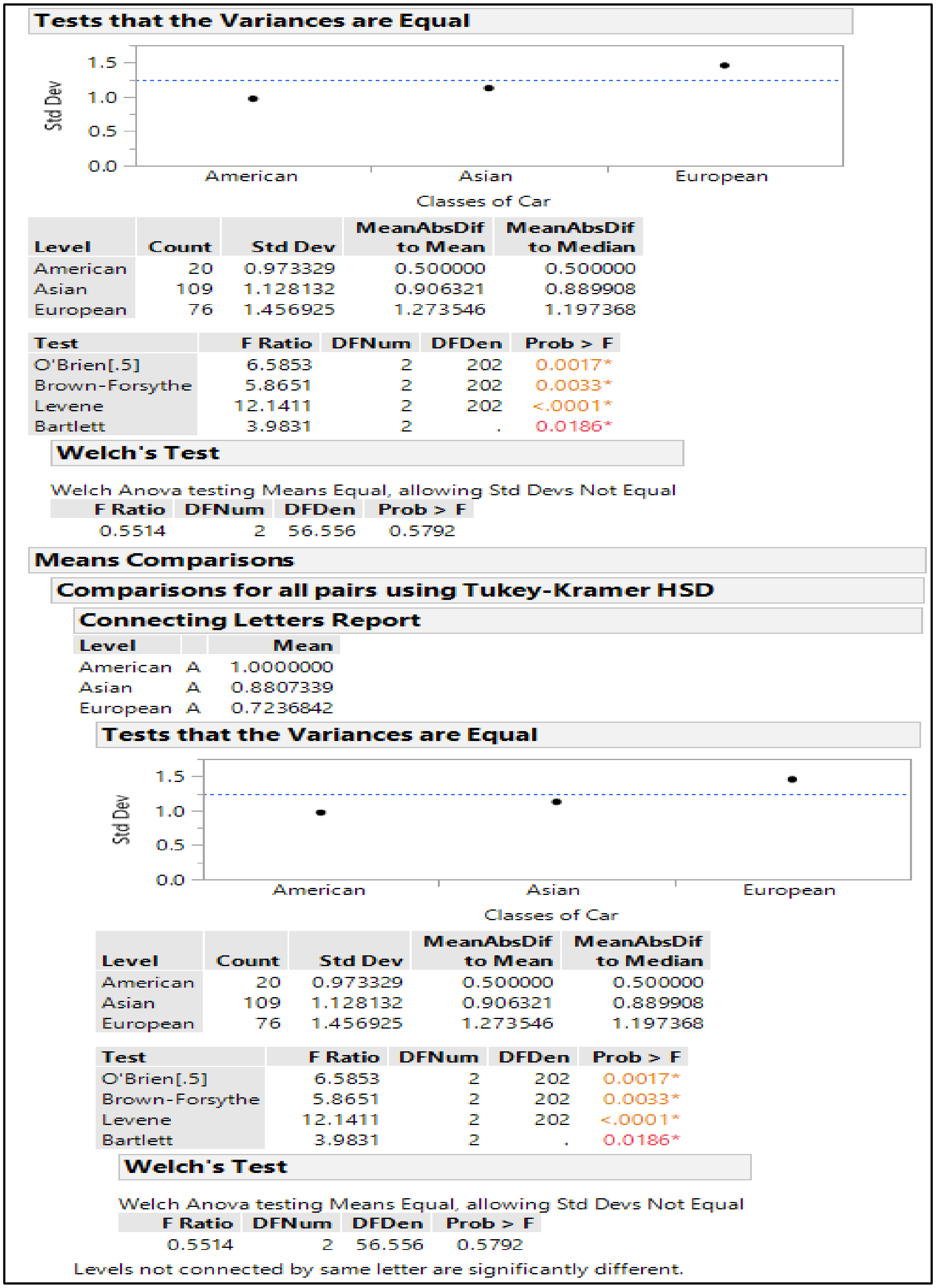
The current analysis found three major characteristics of an automobile of which are associated with the increase or decrease in risk. The three major characteristics are the width, normalized-loss, and wheelbase of an automobile. These three characteristics showed significant results in multiple analyses. Based on these results an individual could assume an increase in any one of these categories would also result in increased risk. If an individual wanted to have less risk while driving, we suggest focusing attention on these three categories to minimize potential risk. Furthering the investigation, a significant increase or decrease in risk was found when comparing the manufacturer brand (make) of automobiles. An individual has the option to purchase a variety of different makes of automobiles. From the collected sample, an increase in risk was found when examining the Alfa-Romeo models. A decrease in risk was also found when examining the Volkswagen model. From this analysis, if an individual has the desire to have the feeling of less risk when driving it is suggested to purchase a Volkswagen model. If risk is not a concern for the driver, then any model in the sample is of interest. The analysis also showed that there were only three vehicles that scored the lowest on the risk scale with a score of negative two. These three vehicles were all manufactured by Volvo. Thus, if a consumer is looking for the safest vehicle possible from the data set, it is suggested to choose an automobile from the Volvo brand. Next, a binary comparison model was conducted to predict risk associated with an automobile. The results from this comparison support the previous results in that wheel-base, normalized loss and width all have significant roles when predicting the risk associated with any vehicle. Model comparison revealed the advantageous nature of the different models. The model comparison results can be a great tool for insurance companies and car users when determining the potential risk of a vehicle. Lastly, the analysis of which produced the best results was the neural networks. Results from the neural networks show how normalized-loss, width and wheelbase influence the overall risk factor of an automobile. We emphasized the importance of these results especially the miss rate to anyone who is attempting to assess the risk level associated with a specific vehicle.

Although this analysis was thorough and provided reliable results, there were some obstacles and drawbacks to our analysis. One obstacle that was overcome when examining data was the missing values. There were few missing values and it was assumed that these missing values had little to no effect on the overall results. Along with missing data values, the analysis only examined attributes of the automobile and not characteristics of the operator. In future investigations, it would be interesting and worthwhile to examine driver characteristics associated with risk. Driver characteristics could include age, height, education level etc. Continued analysis is highly recommended type because society is continually changing, and more automobiles are being introduced to the navigational roads every day.

References

"UCI Machine Learning Repository: Automobile Data Set." *UCI Machine Learning Repository: Flags Data Set*, archive.ics.uci.edu/ml/datasets/Automobile

Appendices



Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
American	20	0.973329	0.500000	0.500000
Asian	109	1.128132	0.906321	0.889908
European	76	1.456925	1.273546	1.197368

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	6.5853	2	202	0.0017*
Brown-Forsythe	5.8651	2	202	0.0033*
Levene	12.1411	2	202	<.0001*
Bartlett	3.9831	2	.	0.0186*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
0.5514	2	56.556	0.5792

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Connecting Letters Report

Level		Mean
American	A	1.0000000
Asian	A	0.8807339
European	A	0.7236842

Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
American	20	0.973329	0.500000	0.500000
Asian	109	1.128132	0.906321	0.889908
European	76	1.456925	1.273546	1.197368

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	6.5853	2	202	0.0017*
Brown-Forsythe	5.8651	2	202	0.0033*
Levene	12.1411	2	202	<.0001*
Bartlett	3.9831	2	.	0.0186*

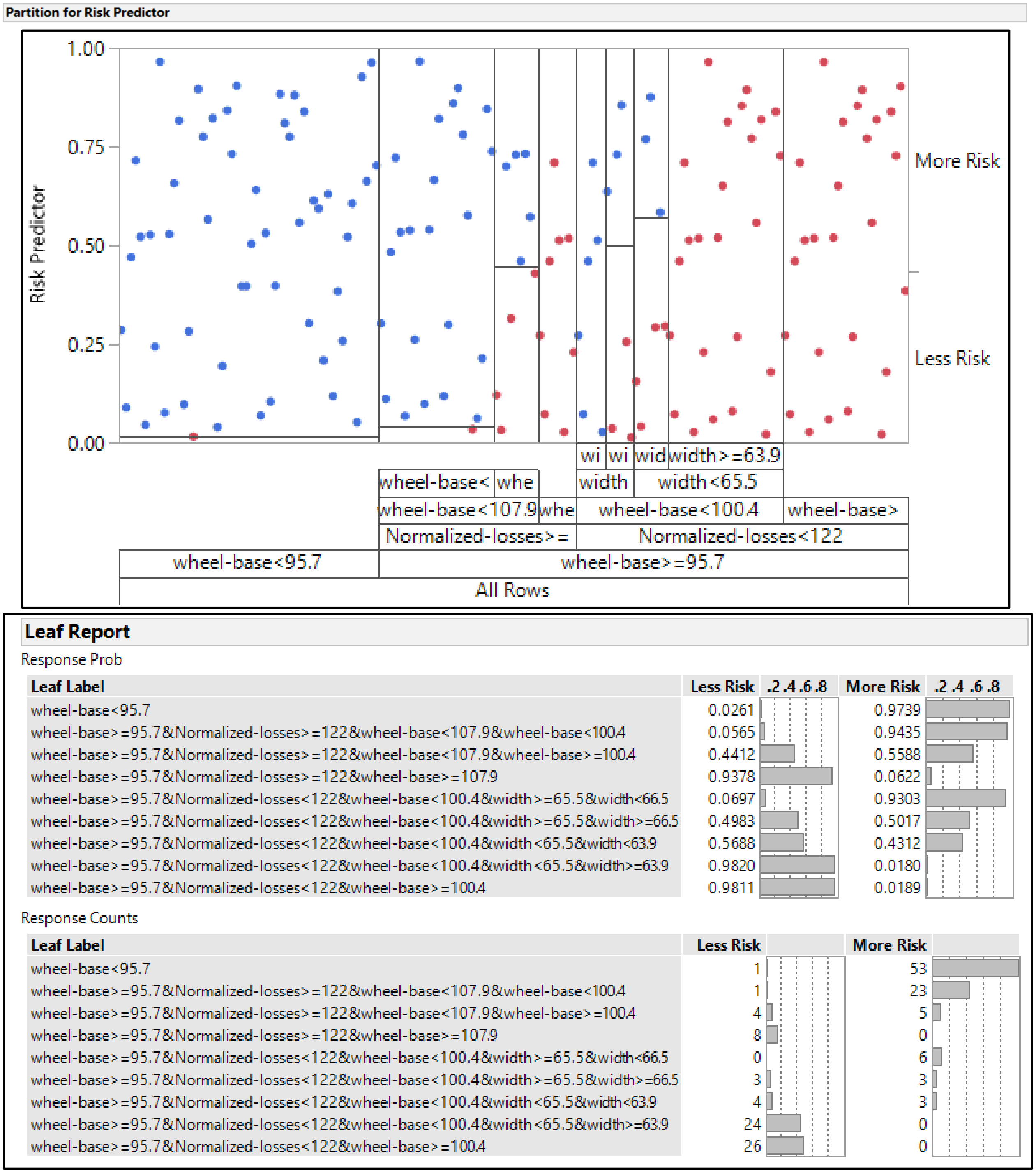
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
0.5514	2	56.556	0.5792

Levels not connected by same letter are significantly different.

***Appendix A:** Analysis of Variance (Levene test, Welch's Tukey-Kramer HSD and Connecting Letter Report)*



Leaf Report

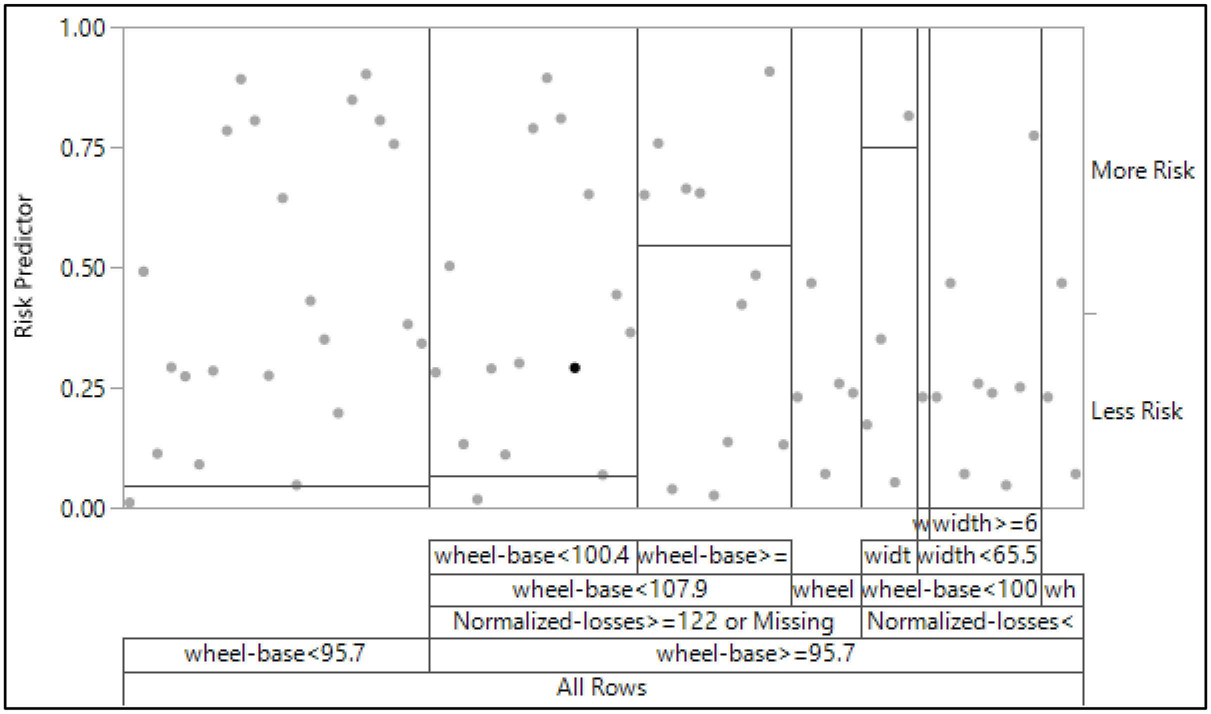
Response Prob

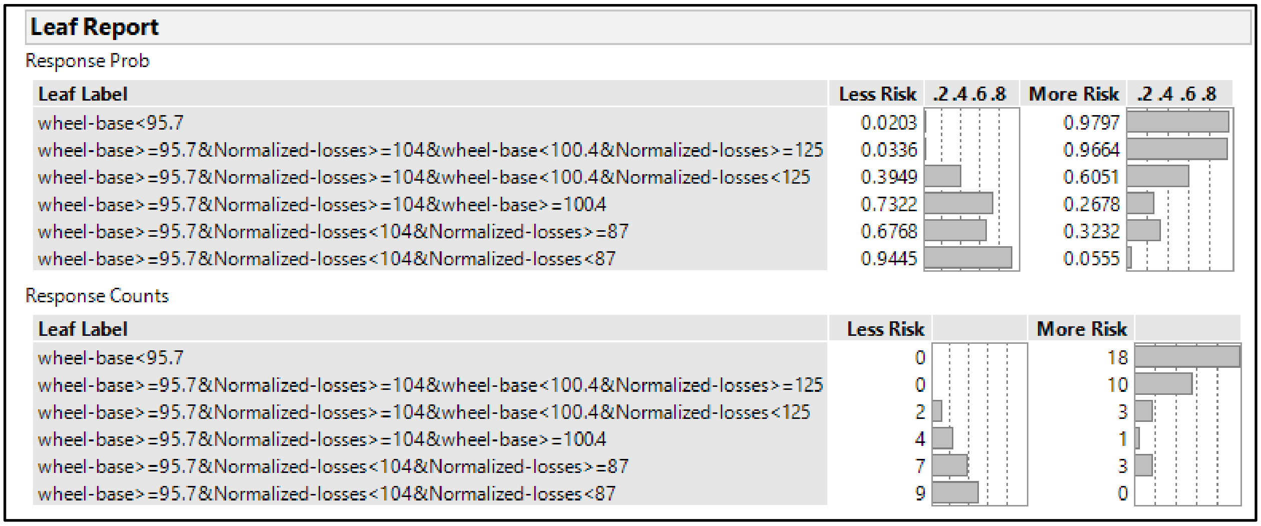
Leaf Label	Less Risk	.2 .4 .6 .8	More Risk	.2 .4 .6 .8
wheel-base<95.7	0.0261		0.9739	
wheel-base>=95.7&Normalized-losses>=122&wheel-base<107.9&wheel-base<100.4	0.0565		0.9435	
wheel-base>=95.7&Normalized-losses>=122&wheel-base<107.9&wheel-base>=100.4	0.4412		0.5588	
wheel-base>=95.7&Normalized-losses>=122&wheel-base>=107.9	0.9378		0.0622	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width>=65.5&width<66.5	0.0697		0.9303	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width>=65.5&width>=66.5	0.4983		0.5017	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width<65.5&width<63.9	0.5688		0.4312	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width<65.5&width>=63.9	0.9820		0.0180	
wheel-base>=95.7&Normalized-losses<122&wheel-base>=100.4	0.9811		0.0189	

Response Counts

Leaf Label	Less Risk		More Risk	
wheel-base<95.7	1		53	
wheel-base>=95.7&Normalized-losses>=122&wheel-base<107.9&wheel-base<100.4	1		23	
wheel-base>=95.7&Normalized-losses>=122&wheel-base<107.9&wheel-base>=100.4	4		5	
wheel-base>=95.7&Normalized-losses>=122&wheel-base>=107.9	8		0	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width>=65.5&width<66.5	0		6	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width>=65.5&width>=66.5	3		3	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width<65.5&width<63.9	4		3	
wheel-base>=95.7&Normalized-losses<122&wheel-base<100.4&width<65.5&width>=63.9	24		0	
wheel-base>=95.7&Normalized-losses<122&wheel-base>=100.4	26		0	

Appendix B: Binary Decision Tree (Training)





Leaf Report

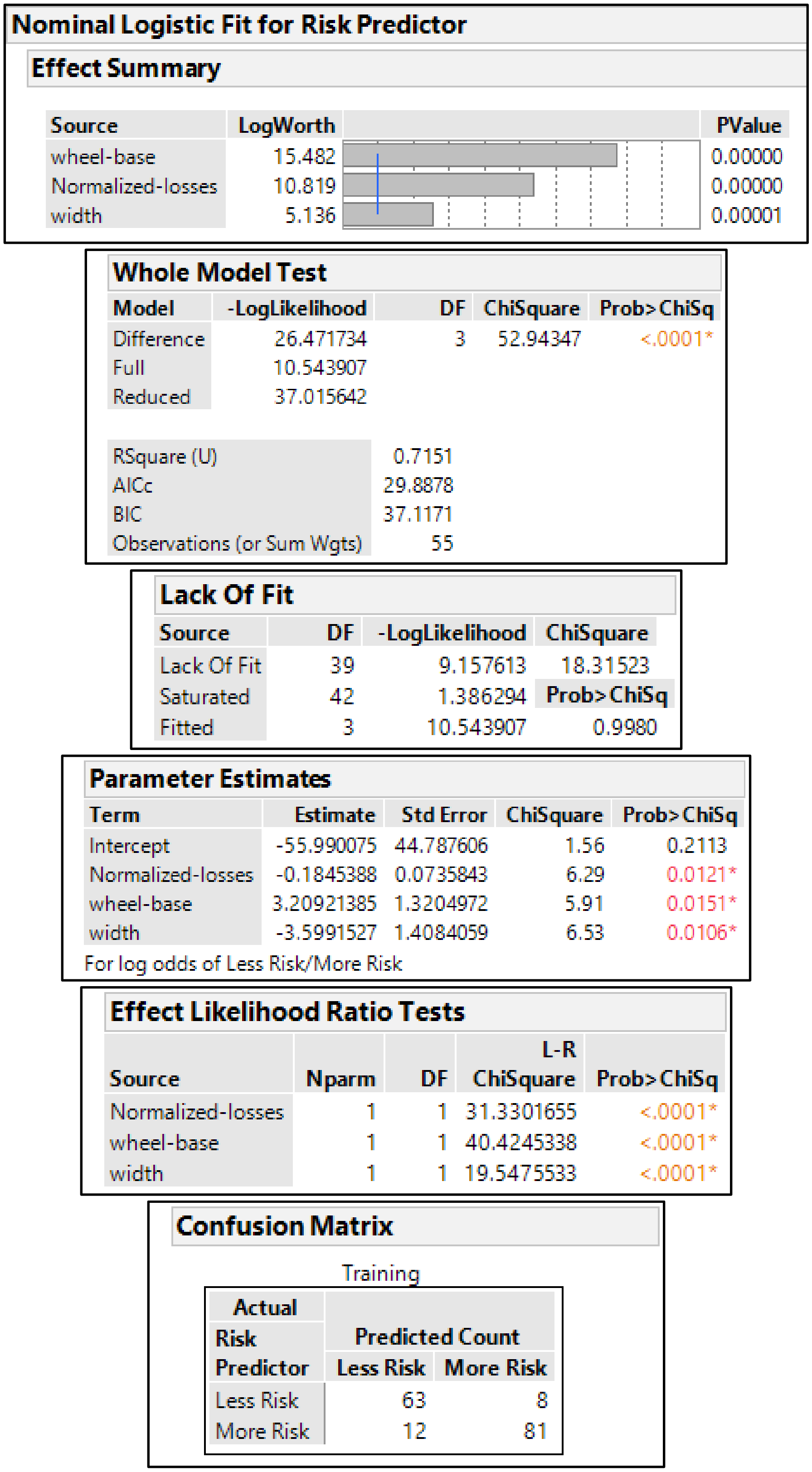
Response Prob

Leaf Label	Less Risk	.2 .4 .6 .8	More Risk	.2 .4 .6 .8
wheel-base<95.7	0.0203		0.9797	
wheel-base>=95.7&Normalized-losses>=104&wheel-base<100.4&Normalized-losses>=125	0.0336		0.9664	
wheel-base>=95.7&Normalized-losses>=104&wheel-base<100.4&Normalized-losses<125	0.3949		0.6051	
wheel-base>=95.7&Normalized-losses>=104&wheel-base>=100.4	0.7322		0.2678	
wheel-base>=95.7&Normalized-losses<104&Normalized-losses>=87	0.6768		0.3232	
wheel-base>=95.7&Normalized-losses<104&Normalized-losses<87	0.9445		0.0555	

Response Counts

Leaf Label	Less Risk		More Risk	
wheel-base<95.7	0		18	
wheel-base>=95.7&Normalized-losses>=104&wheel-base<100.4&Normalized-losses>=125	0		10	
wheel-base>=95.7&Normalized-losses>=104&wheel-base<100.4&Normalized-losses<125	2		3	
wheel-base>=95.7&Normalized-losses>=104&wheel-base>=100.4	4		1	
wheel-base>=95.7&Normalized-losses<104&Normalized-losses>=87	7		3	
wheel-base>=95.7&Normalized-losses<104&Normalized-losses<87	9		0	

Appendix C: ***Binary Decision Tree (Holdout)***



Nominal Logistic Fit for Risk Predictor

Effect Summary

Source	LogWorth		PValue
wheel-base	15.482		0.00000
Normalized-losses	10.819		0.00000
width	5.136		0.00001

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	26.471734	3	52.94347	<.0001*
Full	10.543907			
Reduced	37.015642			

RSquare (U)	0.7151
AICc	29.8878
BIC	37.1171
Observations (or Sum Wgts)	55

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	39	9.157613	18.31523
Saturated	42	1.386294	Prob>ChiSq
Fitted	3	10.543907	0.9980

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-55.990075	44.787606	1.56	0.2113
Normalized-losses	-0.1845388	0.0735843	6.29	0.0121*
wheel-base	3.20921385	1.3204972	5.91	0.0151*
width	-3.5991527	1.4084059	6.53	0.0106*

For log odds of Less Risk/More Risk

Effect Likelihood Ratio Tests

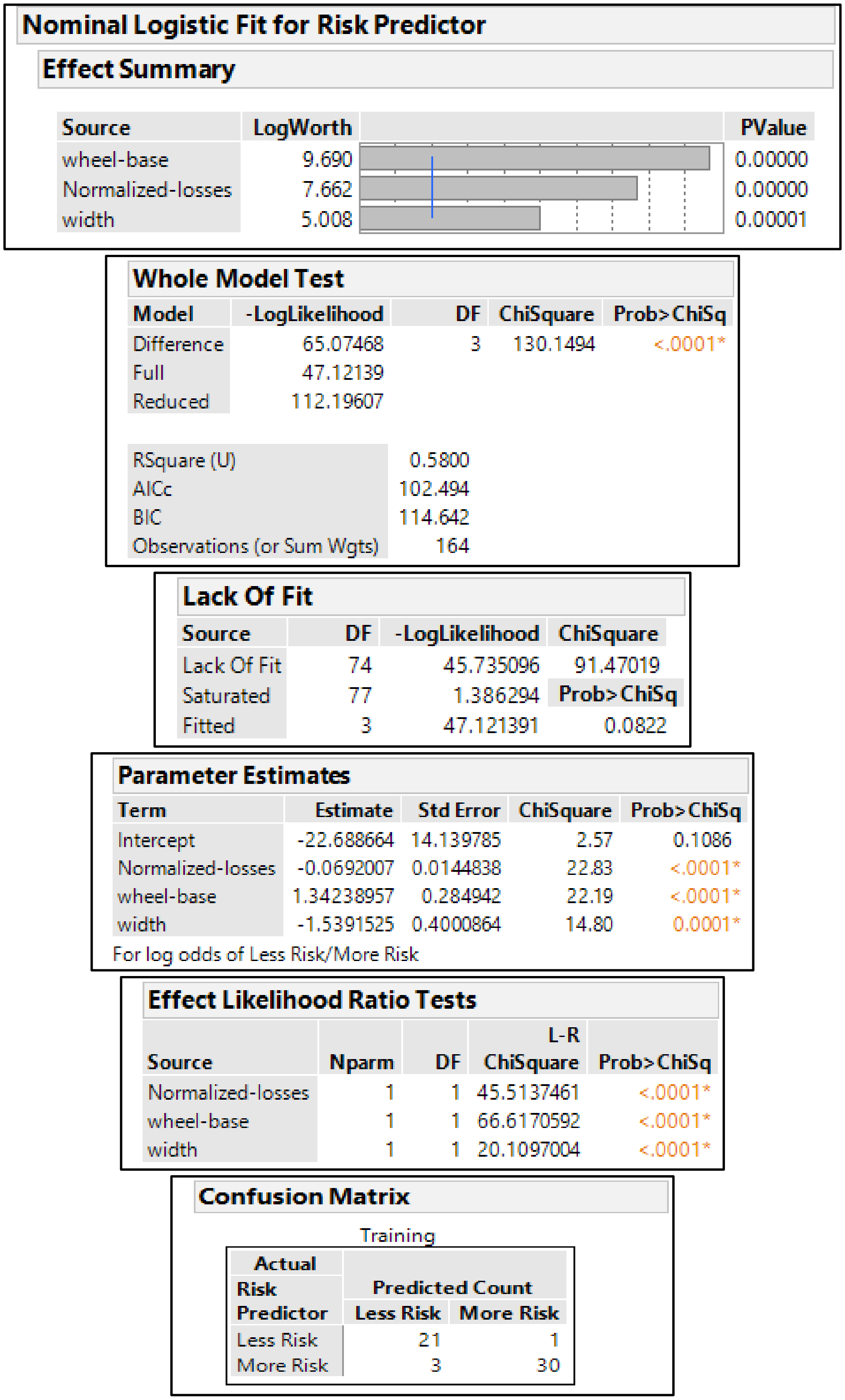
			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Normalized-losses	1	1	31.3301655	<.0001*
wheel-base	1	1	40.4245338	<.0001*
width	1	1	19.5475533	<.0001*

Confusion Matrix

Training

Actual		
Risk	Predicted Count	
Predictor	Less Risk	More Risk
Less Risk	63	8
More Risk	12	81

Appendix D: Binary Logistic Regression (Training)



Nominal Logistic Fit for Risk Predictor

Effect Summary

Source	LogWorth		PValue
wheel-base	9.690		0.00000
Normalized-losses	7.662		0.00000
width	5.008		0.00001

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	65.07468	3	130.1494	<.0001*
Full	47.12139			
Reduced	112.19607			

RSquare (U)	0.5800
AICc	102.494
BIC	114.642
Observations (or Sum Wgts)	164

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	74	45.735096	91.47019
Saturated	77	1.386294	Prob>ChiSq
Fitted	3	47.121391	0.0822

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-22.688664	14.139785	2.57	0.1086
Normalized-losses	-0.0692007	0.0144838	22.83	<.0001*
wheel-base	1.34238957	0.284942	22.19	<.0001*
width	-1.5391525	0.4000864	14.80	0.0001*

For log odds of Less Risk/More Risk

Effect Likelihood Ratio Tests

			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Normalized-losses	1	1	45.5137461	<.0001*
wheel-base	1	1	66.6170592	<.0001*
width	1	1	20.1097004	<.0001*

Confusion Matrix

Training

Actual		
Risk	Predicted Count	
Predictor	Less Risk	More Risk
Less Risk	21	1
More Risk	3	30

Appendix E - Binary Logistic Regression (Holdout)

Figures and Tables

Attribute	**Attribute Range**
Symbolizing (Risk)	-3, -2, -1, 0, 1, 2, 3
Normalized-Losses	Continuous from 65 to 256
Make	alfa-romeo, audi, bmw, chevrolet, dodge, honda,isuzu, jaguar, mazda, mercedes-benz, mercury,mitsubishi, nissan, peugeot, plymouth, porsche ,renault, saab, subaru, toyota, volkswagen, volvo
Fuel-type	Diesel, gas
Aspiration	std, turbo
Number of Doors	Four, two
Body-Style	Hardtop, wagon, sedan, hatchback, convertible
Drive-Wheels	4wd, fwd, rwd
Engine-Location	front, rear.
Wheel-base	continuous from 86.6 120.9.
Length	continuous from 141.1 to 208.1.
Width	continuous from 60.3 to 72.3.
Height	continuous from 47.8 to 59.8.
Curb-weight	continuous from 1488 to 4066.
Engine-Type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor
Number of Cylinders	eight, five, four, six, three, twelve, two
Engine Size	continuous from 61 to 326.
Fuel System	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
Bore	continuous from 2.54 to 3.94
Stroke	continuous from 2.07 to 4.17
Compression-Ratio	continuous from 7 to 23
Horsepower	continuous from 48 to 288

Peak-rpm	continuous from 4150 to 6600
City-mpg	continuous from 13 to 49
Highway mpg	continuous from 16 to 54
Price	continuous from 5118 to 45400

Table 1: Variables and Variable Ranges