1 Introduction

Recent shifts in the Canadian automotive industry have prompted a reassessment of insurance metrics. The automotive industry, which was once thriving on strong economic factors, faced setbacks because of the COVID-19 pandemic. However, government interventions in 2020 cushioned severe revenue declines, and technological advancements have also transformed traditional operational models. In addition, the automotive insurance sector, which is closely linked to car sales and new vehicle registrations, also navigates this changing landscape. This project aims to predict the relative average loss payment for an insured vehicle per year. This value is normalized for all vehicles within a particular size classification (two-door small, station wagons, sports/speciality, etc.. . .), and represents the average loss per car per year. The model could find direct applications in insurance companies to inform risk management strategies, setting appropriate premiums in an industry marked by evolving consumer preferences and new vehicles coming to the market daily. It uses data, ranging from the vehicle's make and model to engine size and fuel type, offer a detailed view of its attributes. By predicting normalized losses based on these comprehensive vehicle characteristics, the model can serve insurers and consumers alike, guiding informed decisions about vehicle safety, pricing, and insurance coverage.

2 Data Description

2.1 Preprocessing Steps

Prior to analysis and modelling, the data was cleaned and formatted in the following steps. First, the text-based categorical variables were converted into a numeric format. For instance, the number of doors and number of cylinders columns, which originally contained text descriptions like 'two', 'four', or 'six', were transformed into corresponding numeric values like 2, 4, or 6. This conversion was done because numeric values are more easily processed and analyzed by statistical and machine learning models. Additionally, the data was standardized by replacing hyphens in column names with underscores to ensure consistency and ease of access to these columns in the subsequent stages of the analysis. The next step was to handle missing values. In cases where columns had character values such as '?' or blank entries, these were interpreted as missing data and replaced with NA. After identifying and converting numeric columns, all rows with any NA values were dropped to maintain a clean dataset. Next, categorical columns were converted into dummy variables to include categorical data more effectively in predictive models, representing the presence or absence of each category.

2.2 Exploratory Data Analysis

A systematic approach was carried out to visualizing the distribution and relationships of various numeric variables to gain insights that could assist in prediction. Box plots provide a clear summary of the central tendency and spread of the data, highlighting outliers and the range of each variable (ref. Appendix, Figure 1). From the first figure, it is observed that variables like 'price', 'horsepower', and 'engine size' show a wide range of values and several outliers, suggesting significant variability that could influence average loss payment. On further analysis, we can see that outliers in 'price' are indicative of luxury or performance cars that could incur higher insurance costs, which is acceptable. The variable 'symboling' (ranging from -3 to +3), which relates to the risk assessment

of a vehicle relative to its price at the time of manufacture, also shows a widespread, suggesting a diverse set of vehicles in terms of insurance risk rating.

Histograms offer a deeper look at the frequency distribution of each variable (ref. Appendix, Figure 2). For instance, the compression ratio (ratio of volume with different piston position) histogram displays 2 peaks at 10 and 20, which shows 2 classes of cars in terms of engine power. The miles per gallon in city and highway histograms, which measure fuel efficiency, show that most cars in the data are fuel efficient, with few high-performance outliers that have lower mpg ratings. The engine size and horsepower of the car are similarly distributed with the former being centered around 120 cubic centimeters and the latter is centered around 100 hp.

Scatterplots are used to examine the relationship between each numeric variable and average loss payment (target variable) (ref. Appendix, Figure 3). The size of the engine showed a positive relation to the average losses. This could be because of higher risk associated with higher performance cars, but also replacement of larger expensive engines will incur more costs. It is interesting to note that vehicle price does not exhibit an obvious pattern in relation to losses, suggesting that the cost of the car alone is not a sole predictor of insurance cost. Other factors need to be considered. From the above visualizations, we can conclude certain attributes such as engine size, horsepower, height, and the risk symbol associated with the car may have a more pronounced impact on the insurance losses. These insights are valuable for creating a predictive model that estimates average loss payments, in assessing risk levels.

2.3 Correlation Analysis

A correlation matrix (ref. Appendix, Figure 4) is computed, to find the strength and direction of associations between all pairs specifically linked to average loss payment, offering insights into how various vehicle characteristics may relate to the expected insurance payments relative to other vehicles. Looking at the top ten positive correlation coefficients with respect to loss payment we can see that a higher symboling value is associated with higher average loss payment, reflecting the close relation with the initial risk assessment of the vehicle at the time of manufacture. Similarly, horsepower, engine size, and price display a positive correlation, implying that more powerful, larger, and expensive vehicles could potentially incur higher insurance costs due to factors such as increased repair costs or higher value leading to greater insurance payouts. Further, attributes such as curb weight (which is the total weight of a fully fueled car), vehicle width, piston stroke in a rotation cycle, and vehicle length also show positive correlations, also hint that larger vehicles are costlier to repair or replace. Conversely, the negative correlations showed that the height of the vehicle and number of doors, which often denote passenger/ family vehicles, show negative correlations, possibly due to a more cautious driving style or lower risk of theft and damage. More fuel-efficient cars, as denoted by miles per gallon, tend to have lower loss payments, potentially reflecting less aggressive driving habits or lower repair costs when compared to high-performance sports cars and luxury rides. The compression ratio, also an indicator of efficient or diesel engines, also shows a negative correlation, which could be attributed to less performance-oriented driving or more responsible usage patterns. These correlations, while indicative, do not imply causation and only represent the linear aspect of the relationships. These results can serve as a starting point for more sophisticated predictive modeling that can account for complex, non-linear interactions with multiple factors.

3 Principal Component Analysis

3.1 PCA Results

Principal Component Analysis (PCA) is a method of reducing multi-dimensional datasets into core components that capture the essence of the information. It re-expresses the dataset by prioritizing the directions where the data varies the most. In the PCA plot for the automotive data, the axes, labelled as Principal Component 1 and Principal Component 2, capture the largest and second-largest variance in the dataset, corresponding to 46.12% and 16.67% respectively. Together, these components encapsulate over 62% of the data's variability, offering a high-level snapshot of the dataset's structure. The vectors spreading out from the origin of the PCA plot reflect how each variable influences a principal component. The length of these vectors is indicative of the variable's contribution—the longer the vector, the more significant the role of that variable in that principal component. The direction of these vectors represents the correlation between variables, where vectors pointing in the same direction suggest a positive correlation and those in opposite directions hint at a negative correlation. The insights drawn from this PCA are valuable for reducing dimensionality in data, aiding in the visualization of complex relationships, and informing subsequent predictive modelling. By identifying principal components as new, simplified features that retain the most critical information, we can streamline the modelling process, potentially enhancing the performance of predictive algorithms applied to such datasets.

3.2 PCA Interpretation

From the PCA plot (ref. Appendix, Figure 5), we can discern that engine size, horsepower, curb weight, price have significant positive loadings on Principal Component 1 (PC1). This shows that they are strongly correlated with each other; vehicles with larger engines tend to have more horsepower, be heavier, and are more expensive. These attributes are driving the variability along PC1, which could be thought of as a "size and power" dimension of the vehicles in the dataset. City MPG, Highway MPG are pointing in the opposite direction to the engine size and horsepower on the plot, which suggests that vehicles with better fuel efficiency typically have smaller engines and fewer horsepower. These fuel economy metrics are influencing PC2, which might be considered as an "efficiency" dimension in the data. Symboling relates to the initial risk rating of the car, has a distinct position in the plot, not closely aligned with either PC1 or PC2. This implies that the risk rating symbol is not strongly correlated with the physical attributes of the vehicles represented by PC1 and PC2. Symboling is closely related to the average loss payments, suggesting that it could work as a good predictor for average loss payments. Height and number of doors have smaller loadings on both components compared to the others, which means they contribute less to the variability explained by PC1 and PC2. Their positions suggest that they do not have a strong linear relationship with the "size and power" or "efficiency" dimensions. Wheelbase, length, and width are related to the dimensions of the vehicle and have moderate positive loadings on PC1. This means that larger vehicles, which have longer wheelbases and are wider and longer, also tend to be heavier and more powerful, contributing to the "size and power" aspect of PC1. Peak RPM has a moderate negative loading on PC2, indicating that cars with higher peak RPMs are somewhat less efficient in terms of fuel usage. Bore, stroke, compression ratio are technical engine characteristics that show moderate associations with both PC1 and PC2, suggesting a more complex relationship with the vehicle's overall characteristics.

4 Model Selection

Based on the above analysis, a model is built to predict the average annual loss payment for a particular insured car. The loss payment is set as the target variable, and a range of features, including technical specifications of the cars, make, fuel type, body style, and other categorical variables (having more than 1 instance) that have been converted into binary (dummy) variables, are used as predictors. Random Forest is preferred over traditional regression models because the relationship between the predictors and the target variable is complex and non-linear. Regression models assume a specific form of the relationship (usually linear), and if this assumption is violated, the model's predictions can be significantly off. Random Forest, being a non-parametric method, does not make these assumptions and can capture much more complex relationships between the variables. It also performs feature selection on its own and outputs the most important features that were used to create the model. This makes steps like fixing nonlinearity and dimension reduction techniques unnecessary when working with random forest. Random forest uses the concept of bagging, or bootstrap aggregating. Bagging involves creating multiple datasets from the original data by sampling with replacement, building a decision tree on each of these datasets, and then combining the trees' outputs to make a final prediction. Each tree in a Random Forest is built on a different sample of the data, and at each split in the tree, only a random subset of the features is considered. This process introduces randomness into the model building, which helps to prevent over fitting and contributes to the algorithm's robustness. The final prediction of the Random Forest is typically made by averaging the predictions of all the trees in the case of regression, or by majority voting in the case of classification, leading to improved accuracy and stability compared to a single decision tree. Random Forest is preferred over a single decision tree because a single tree captures noise in the training data as if it were a genuine pattern, leading to poor generalization to new, unseen data.

5 Analyzing Results of a Single Regression Tree

At the top of the tree (ref. Appendix, Figure 6), at the root node, all observations are split based on the symboling of a car. From there, the next decision is based on whether symboling is less than 0. For cars with a symbolling score less than 0, the left path considering horsepower is taken, leading to a node where the predicted value is 73 or 94, which accounts for 14% of the data. For cars with a symboling of 0 or more, the decision path moves right, where the drive wheel feature comes into consideration. If the car has front or all-wheel drive, the next decision is based on height. Cars with a height less than 51 follow the left branch, leading to further decisions based on length between the wheels of the car (wheelbase). If the wheelbase is less than 99, we move to decisions based on stroke, with varying predicted values (92, 100,118) depending on the fuel system (carburetor) has 2 or more barrels. If the wheelbase is 99 or more, the predicted value is 106, covering 51% of the data subset. For cars with rear wheel drive, and bore greater than or equal to 3.5, the predicted value is quite high at 158, but this only represents 16% of the data subset. It suggests that cars with a larger bore size and rear-wheel drive tend to have higher predicted values for the target variable.

This regression tree suggests that a car's symboling, drive type, bore size, and various other dimensional attributes significantly influence its predicted value, with the reliability of each prediction reflected by the percentage of data points at each leaf node (terminal node).

6 Results and Interpretation

The randomForest result (ref. Appendix, Figure 7) shows how important each feature is in predicting average loss payouts. Two metrics are used to assess importance: %IncMSE and IncNodePurity.

%IncMSE reflects the increase in prediction error (mean squared error) when the data for that feature is not used for prediction. Higher values indicate a greater loss of model accuracy in the absence of that attribute, indicating its significance. IncNodePurity measures the total decrease in node impurities (like variance for regression trees) from splits on that variable, averaged over all trees. Higher numbers indicate that the feature reduces uncertainty and improves node homogeneity. The vehicle's symboling and height have high scores for both criteria, indicating that they are major predictors of loss reimbursements. The 'symboling' score, which represents the car's initial risk evaluation, understandably plays an important role in predicting losses, as riskier cars are more likely to incur bigger losses. 'Height' is an indicator of vehicle stability and safety, which influences loss payouts.

The dimensions of the vehicle like length, width, and the distance between the front and rear wheels are all significant, which could be because larger cars may have different accident profiles and repair costs than smaller ones. Engine size and horsepower are influential as well, which aligns with expectations since more powerful cars could be prone to higher losses either due to increased severity of accidents or higher costs of repair. The brand of the car also provides useful information. The amount of annual loss payments is affected by whether the vehicles are from BMW, Toyota, or Peugeot, demonstrating that brands may have distinct loss profiles, maybe due to brand-specific traits or behaviour patterns associated with their clientele. Interestingly, the miles per gallon in cities and highways emerge as important factors. Higher fuel-efficiency vehicles may have a reduced risk profile or different usage patterns, resulting in fewer or less severe insurance losses. The car's drivetrain, whether rear wheel drive or front wheel drive, is also a significant deciding element, because front wheel drives help drivers maintain traction on slippery roads while also consuming less fuel. Rear wheel drives are more usually linked with performance enthusiasts and truck drivers who desire vehicles that produce the most torque.

Another significant variable in the model is the number of doors on a vehicle. This could be due to underlying patterns associated with certain car types; for example, vehicles with fewer doors, such as two-door sports cars, could be driven more aggressively or more likely to get stolen, resulting in greater insurance claims. Sedans and family automobiles, on the other hand, often have more doors and may be driven more slowly, resulting in fewer or less serious accidents. As a result, the number of doors acts as a proxy for a variety of risk characteristics, such as driving style, vehicle function, and the possibility of theft or damage, all of which are important in deciding insurance losses.

6.1 Effect of Removing Symboling Score From The Model:

Suppose we want to build a model to predict the loss payments based solely on attributes of the car without the help of initial risk score (symboling), we obtain a model with a reduced explained variance of 71.67%.

Comparing the feature importance from two random forest models (ref. Appendix, Figure 7 and 8) (original model with symboling and the alternate model without symboling), we can draw several conclusions. Firstly, in both models, height is the most significant predictor based on the IncNodePurity metric, suggesting that this feature greatly contributes to the purity of the nodes during the tree splits. Its importance increases slightly when symboling is removed (from 26078.76

to 28601.31), indicating that height may compensate for the information previously contributed by symboling. The removal of symboling seems to have a notable effect on number of doors and length in terms of %IncMSE, with both showing an increase in importance in the alternate model. This is due to the redistribution of predictive power among the remaining variables after symboling is excluded, suggesting these features might capture some of the predictive variance symboling accounted for. Interestingly, wheel base and width show a slight decrease in both %IncMSE and IncNodePurity when symboling is removed. This may imply that these variables were partially capturing similar information to symboling, and when symboling is present, it may overshadow their contribution to the model. For engine size, bore, and stroke, their importance in %IncMSE is relatively stable across both models, which suggests these features maintain a consistent predictive power irrespective of the presence of symboling. Overall, the changes in feature importance metrics suggest that symboling does play a substantial role in the model's predictions, and its removal leads to a reassignment of importance to other features. This reassignment isn't uniform, which indicates complex interactions among features.

7 Conclusion

In conclusion, the final Random Forest model predicts the average loss payment using a mix of vehicle design features, performance characteristics, and brand-related variables as key predictors. These insights could guide insurance companies as well as manufacturers in risk assessment and premium calculations. The model explains a 73.26% of the variance in annual loss payments, indicating a strong predictive power.

Appendix

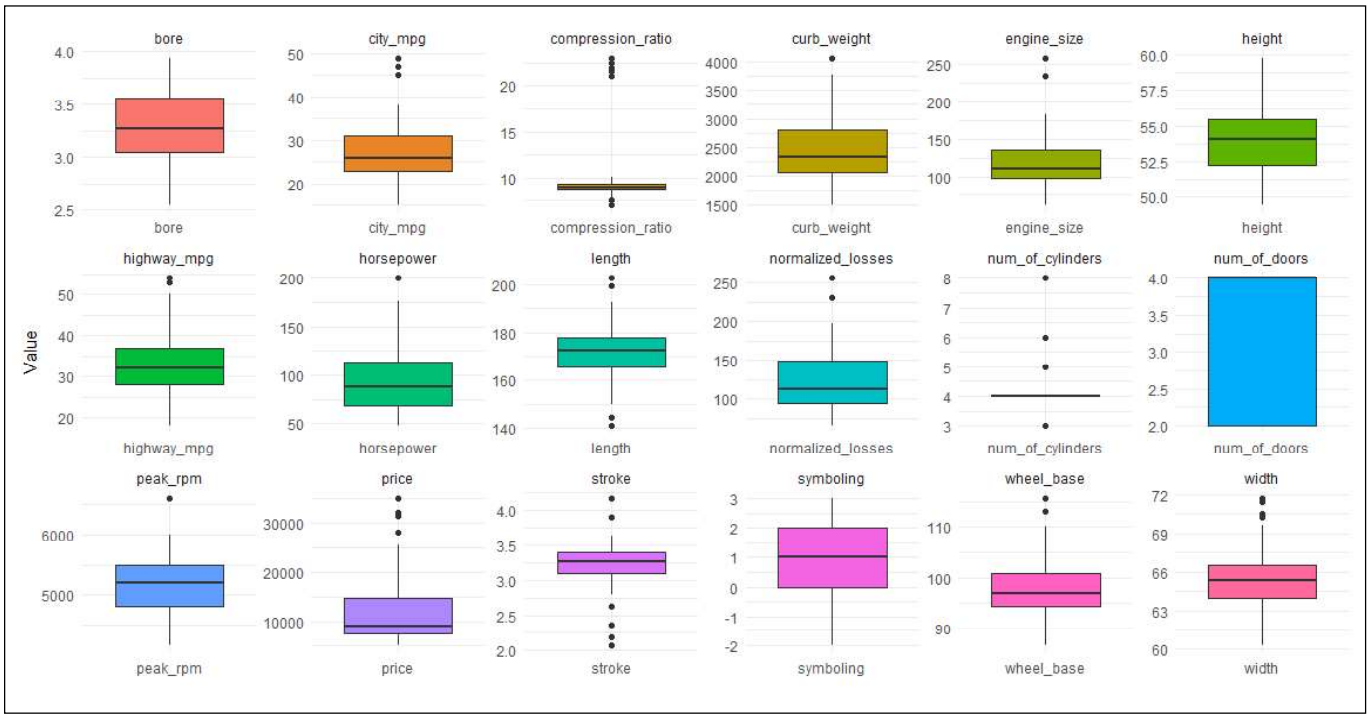

Figure 1: Distribution of Variables

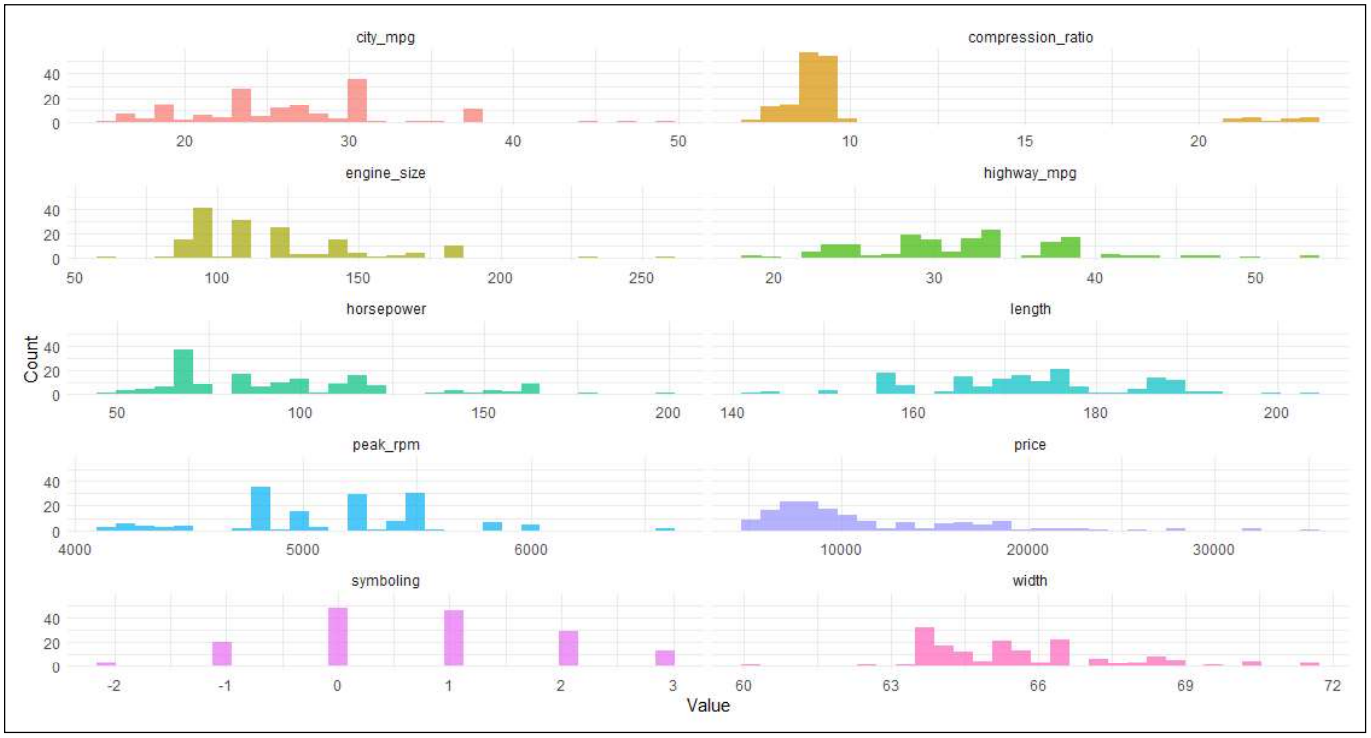

Figure 2: Histograms of Specific Numeric Columns

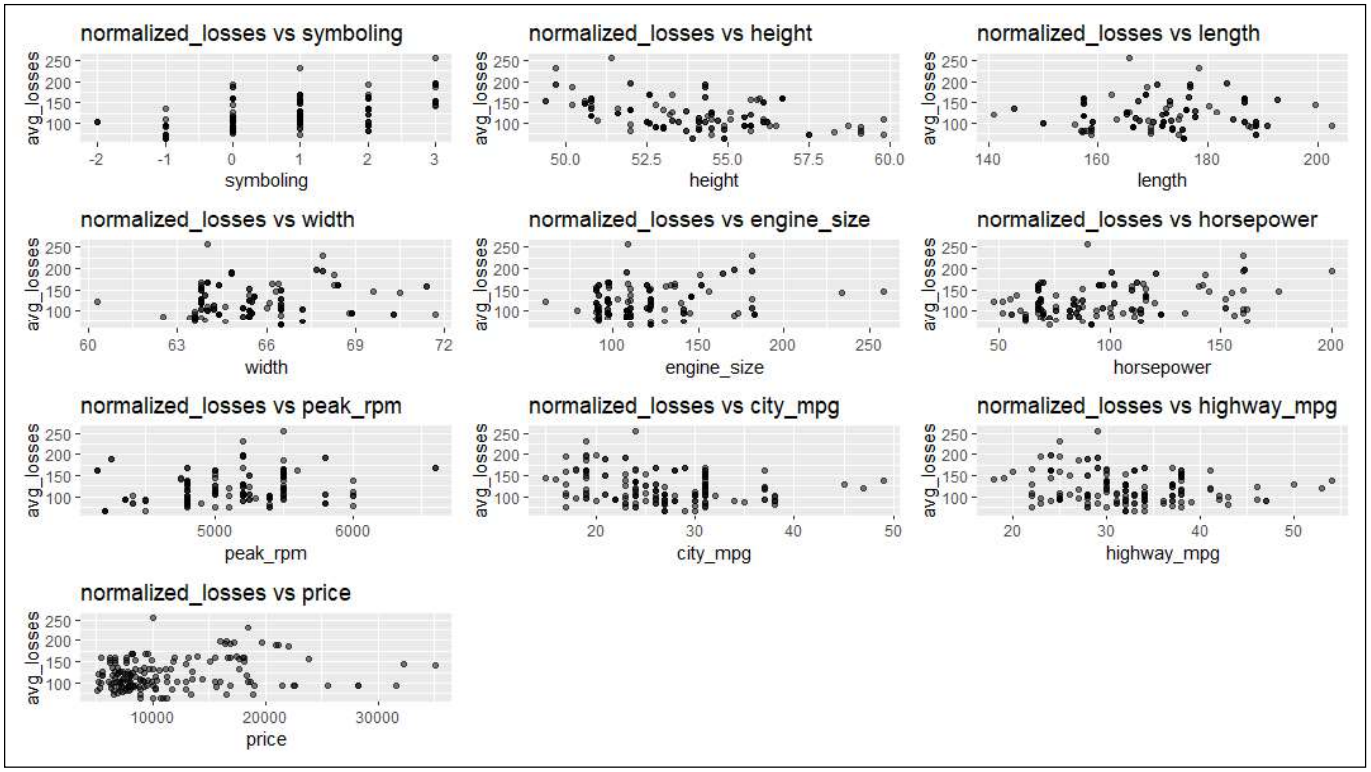

Figure 3: Scatterplots of Numeric Data

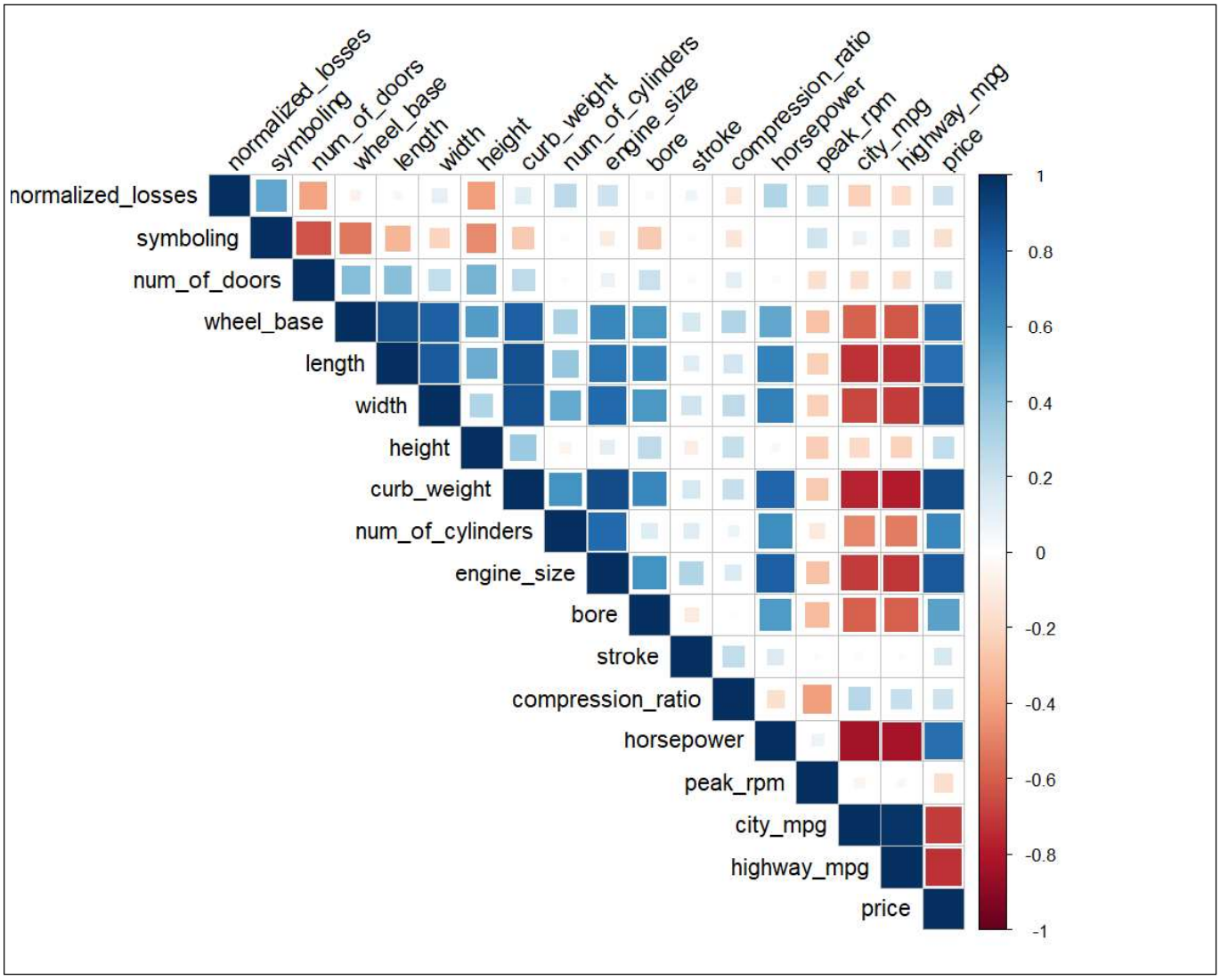

Figure 4: Correlation Heatmap

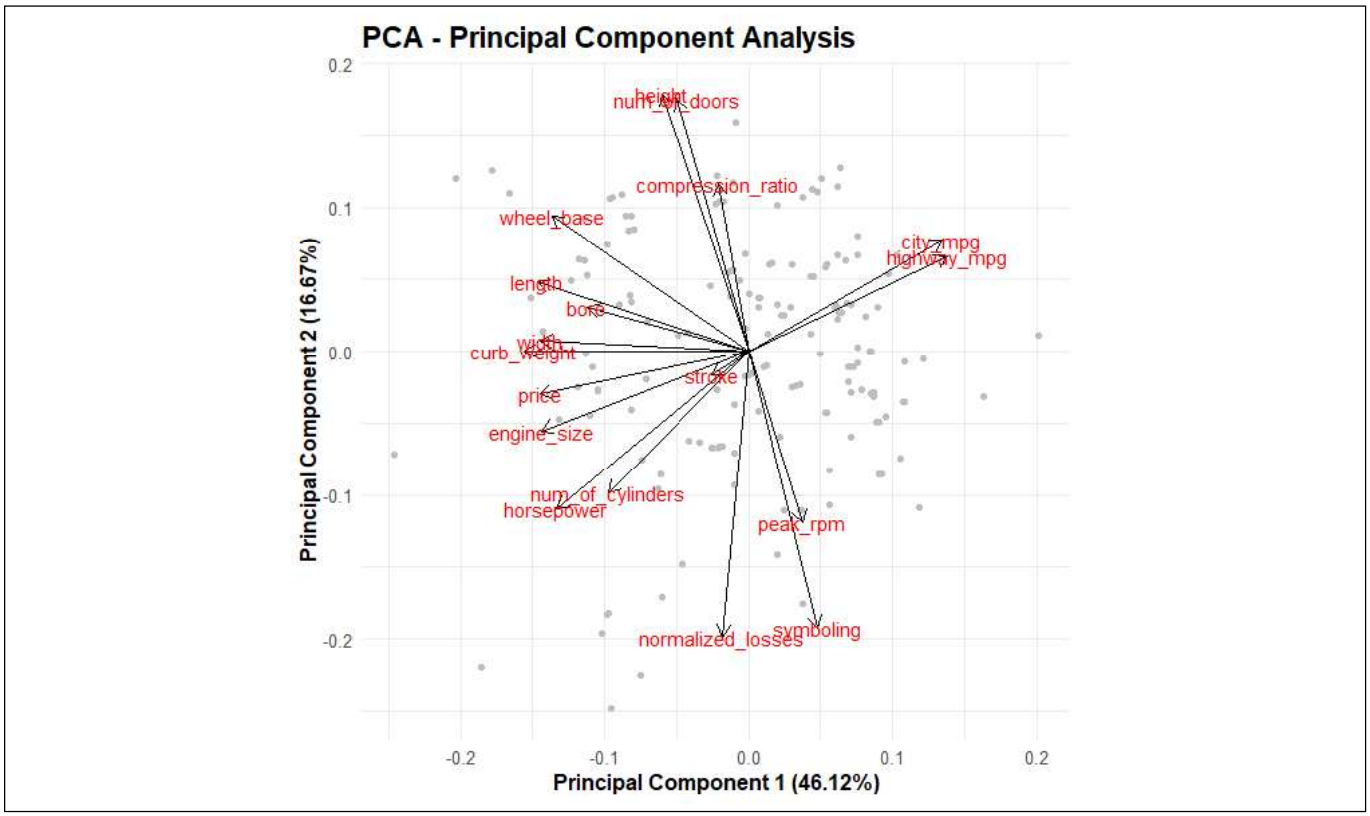

Figure 5: PCA Visualization

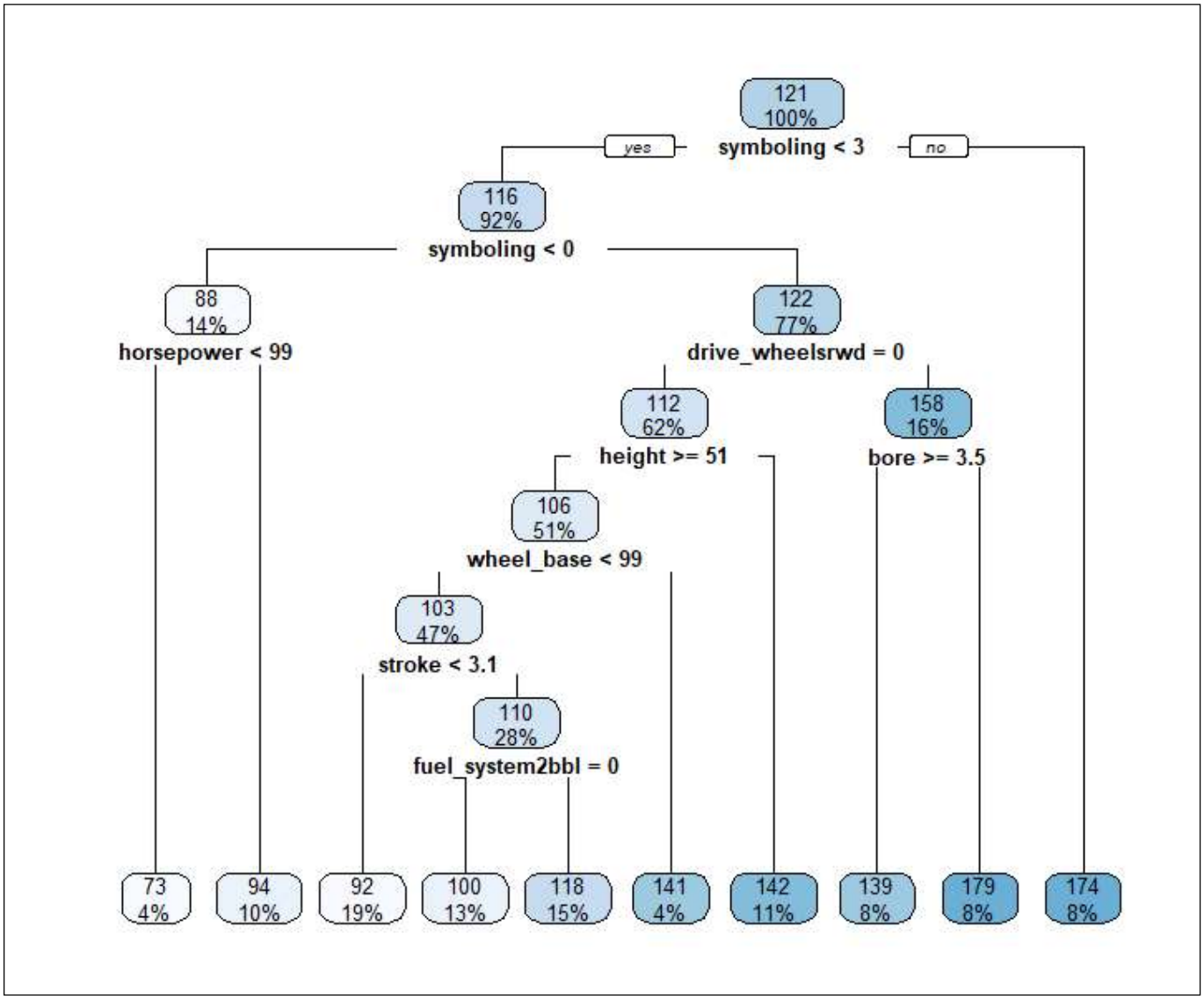

Figure 6: Analysis of Single Regression Tree

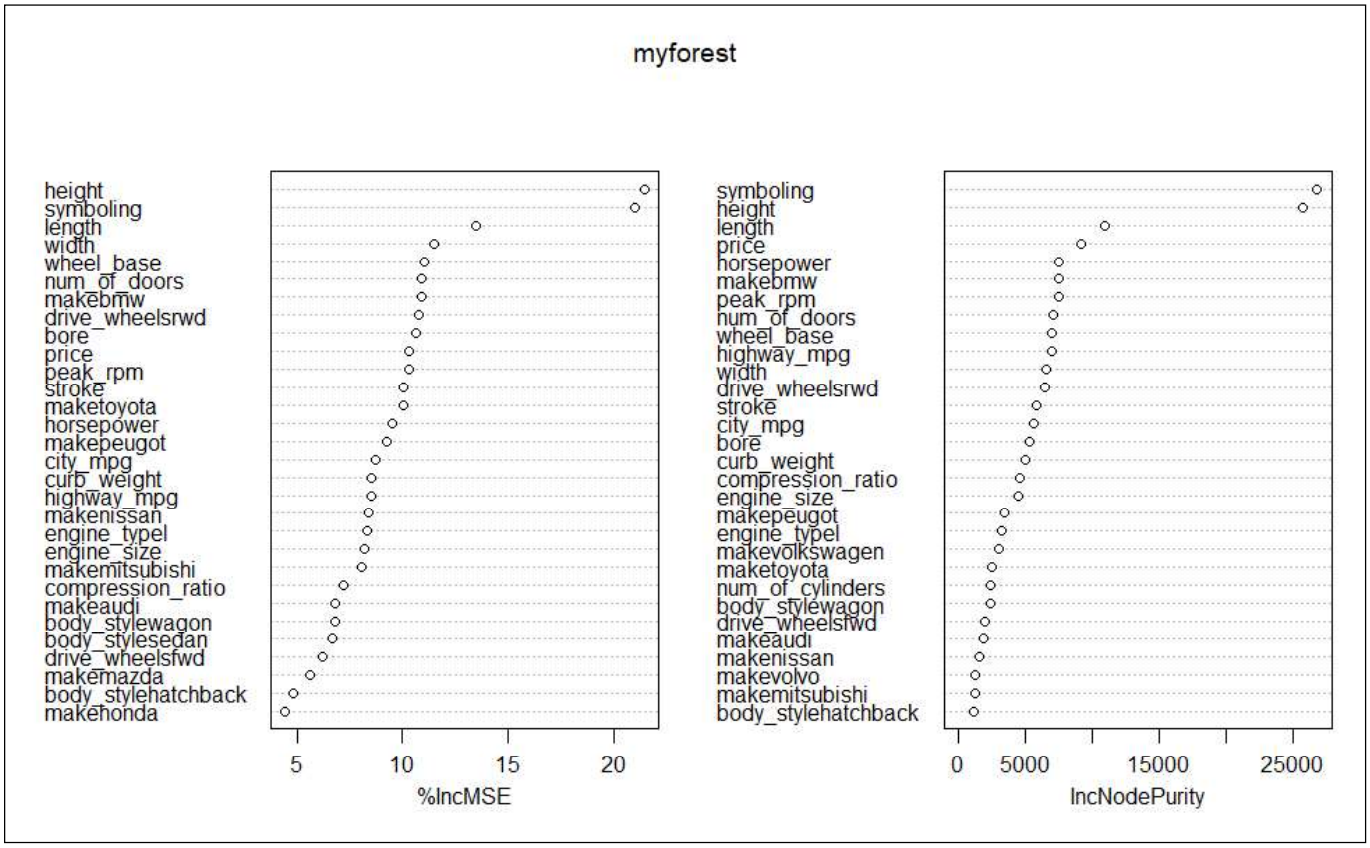

Figure 7: Feature Importances of Random Forest Model

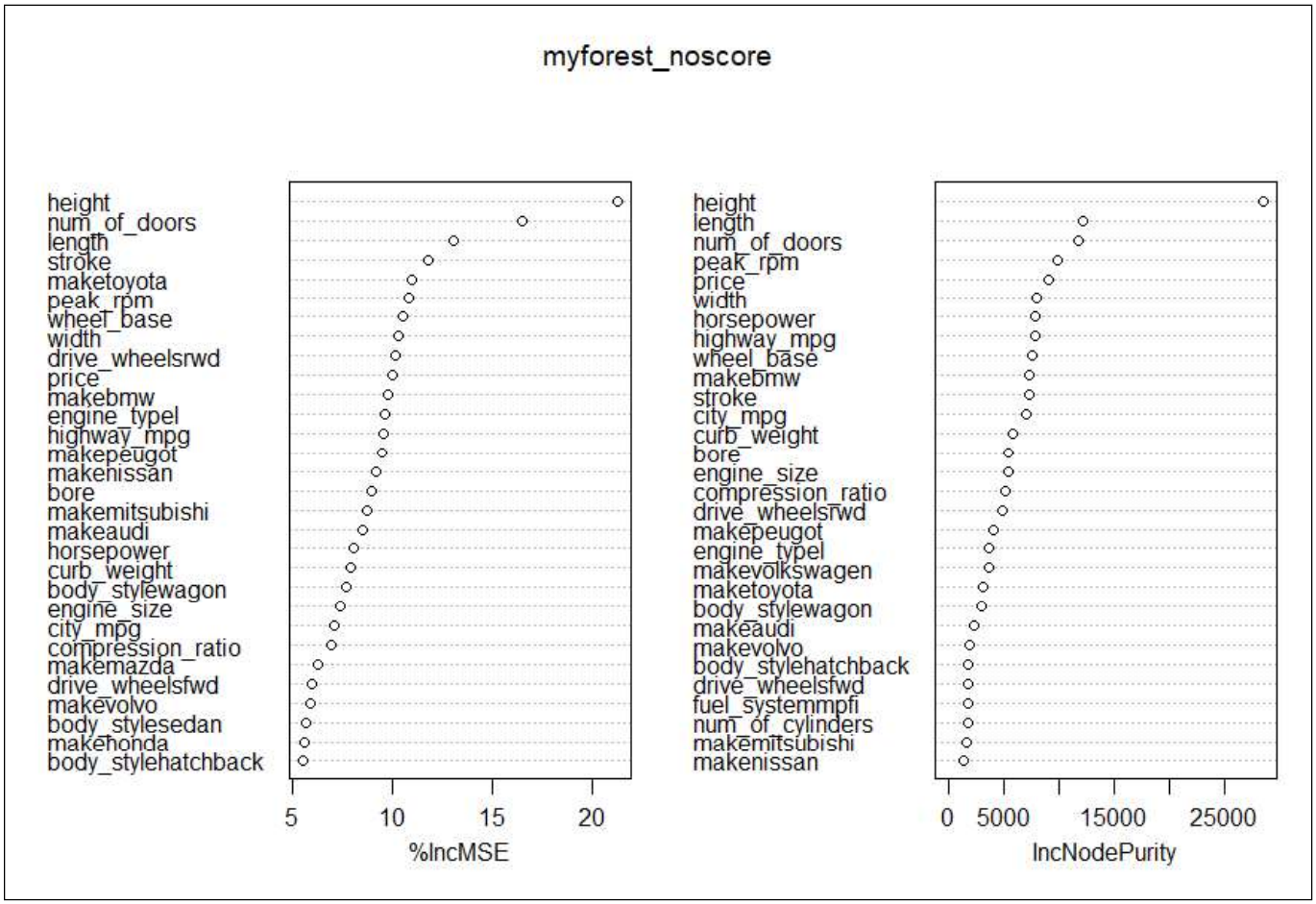

Figure 8: Feature Importances of Random Forest Model Without Symboling Score