Automobile price prediction

Scarlett Xiao

December 3, 2024

1 Introduction

Accurately predicting car prices is crucial for manufacturers, buyers, and sellers in the competitive automotive market. This project uses machine learning models, including regression trees, Random Forests, Gradient Boosting Machine and Principal Component Analysis, to predict car prices based on various vehicle features such as engine size, curb weight, and fuel type. The findings aim to provide actionable insights for car manufacturers to optimize pricing and design strategies.

2 Data Description

2.1 Numerical variables:

In this section, we explore the numerical variables in the dataset to understand their distributions, central tendencies, and relationships with the target variable, price.

Figure 1 illustrates the distribution of several numerical variables in the dataset, offering insights into their central tendencies, spread, and outliers. The normalized losses variable is slightly right-skewed, with most values concentrated between 65 and 150. This suggests that most vehicles have relatively low normalized losses, but a few vehicles exhibit significantly higher losses, which may be due to specific incidents or rare vehicle types. The bore and stroke variables display relatively symmetric distributions centered around their means, with no extreme outliers, indicating that these engine dimensions are relatively consistent across the dataset. This may reflect standard design practices in vehicle manufacturing. The horsepower distribution is right-skewed, with the majority of values below 150 and some outliers above 250, indicating the presence of high-performance or specialty vehicles in the dataset. For peak RPM, the distribution is almost uniform, though there are peaks around 5000–6000 RPM, suggesting that many vehicles are engineered to operate efficiently within this range, reflecting common industry practices. Finally, the price variable is highly right-skewed, with most vehicles priced below $20,000, and a few luxury vehicles priced as high as $45,000. This highlights the variability in pricing strategies and market segmentation.

The correlation analysis in Figure 2 reveals that engine size, curb weight, and horsepower are strongly positively correlated with vehicle price, while fuel efficiency metrics such as city mpg and highway mpg are negatively correlated. Larger dimensions (length, width) and wheelbase also positively correlate with price. Notable interdependencies exist among features, such as strong correlations between city mpg and highway mpg, and between car dimensions (length, width, curb weight). These insights suggest that vehicle size, performance, and efficiency are key determinants of price.

Given the high correlation observed among several numeric variables in the dataset, PCA is a valid approach to address multicollinearity. It transforms these correlated variables into independent principal components while retaining the maximum variance.

2.2 Categorical variables:

In this analysis, we focus on exploring categorical variables to understand their relationship with the vehicle prices. Categorical variables like fuel type, make, and body style can have a significant impact on the pricing of a vehicle, and understanding these relationships is essential for developing predictive models.

Figure 3 provides an overview of the distribution of fuel types across the dataset. Gasoline is the most common fuel type, followed by diesel. Further exploring the relationship between fuel type and vehicle price in Figure 4. From this, we observe that vehicles with a diesel engine tend to be more expensive on average than their gasoline counterparts. However, there are some high-priced gasoline cars as well, showing that fuel type alone doesn't solely determine price. Moreover, in Figure 5, we examine the distribution of body styles across different fuel types. most body styles are more frequently associated with gasoline fuel types, especially hatchback and sedan. And the convertible only have gas fuel.

Moving on to Figure 6 shows how vehicle make influences pricing. Luxury brands like BMW, Jaguar, Mercedes-Benz and prosche tend to have higher average prices compared to other makes. This figure reinforces the idea that vehicle make is a key driver of price, with premium brands commanding a significant price premium.

Figure 7 explores the relationship between engine location and price. Here, we observe that vehicles with front engines tend to have a broader price range, while rear-engine vehicles generally have a more concentrated higher price distribution. This suggests that engine location might correlate with vehicle type (e.g., sports cars or luxury models), influencing price.

These categorical variables have no inherent order, so we'll use One-Hot Encoding.

2.3 Missing values:

normalized.losses has 41 missing values, engine type has 12 missing values and target variable price has 4 missing values while other features have less than 5 missing data.

To ensure a robust approach, we can impute missing values for numeric columns using the mean of the respective column. And for categorical values we impute with mode.

2.4 Outliers:

Since the dataset has only 200 rows, carefully assessing and handling outliers is essential to ensure robust model performance.

1)symboling, height, and bore have no detected outliers.

2)wheel.base, length, curb.weight and engine.size have only a handful of outliers.

3)compression.ratio, stroke and price have several rows flagged as outliers.

Since the outliers reflect valid and critical cases that luxury cars with high prices, and Models like tree-based algorithms are less sensitive to outliers, we don't make adjustment to target variable price. As for rest outlier of compression.ratio, stroke, we apply a logarithmic transformation.

2.5 Principal Component Analysis (PCA)

Given the high correlation among several numeric variables in the dataset, performing PCA helps mitigate multicollinearity by transforming correlated variables into independent components that capture most of the dataset's variance. The PCA rotation matrix in Figure 8 shows the loadings of each feature on the first five principal components, highlighting the contribution of original variables to the new dimensions. Key contributors to PC1, which explains the most variance, include curb weight, length, width, engine size, and price, all positively correlated, reflecting their joint impact on vehicle size and performance. PC2, dominated by height and compression ratio, captures distinct factors such as fuel efficiency or body design. Subsequent components (PC3–PC5) represent a mix of features like stroke, peak RPM, and bore, indicating their more nuanced roles in explaining data variance. This decomposition reduces dimensionality while preserving critical information.

Figure 9 visualizes the relationship between features and their contribution to PC1 and PC2. Horsepower, curb weight, and price are strongly correlated with PC1, while compression ratio, stroke, and engine size are more aligned with PC2. City and highway MPG are negatively correlated with PC1, indicating that vehicles with lower fuel consumption cluster to the left of the plot. Length, width, and wheelbase show a positive correlation with PC1, suggesting that larger vehicles tend to have higher horsepower and price.

The PCA Biplot in Figure 10 further explores car characteristics across different price ranges. It reveals that lower-priced cars typically have smaller engine sizes and lower horsepower, whereas higher-priced cars tend to exhibit better performance and larger dimensions. This demonstrates that price is associated with higher-performing and larger vehicles.

The first few PCA components in Figure 11 account for a significant portion of the variance, suggesting they could enhance model efficiency by reducing feature redundancy. However, due to the complexity of interpreting PCA components and the limited number of variables, we chose to use the original dataset for model training. This decision helps maintain interpretability while addressing multicollinearity through methods such as regularization or feature selection.

However, due to the complexity of interpreting PCA components and the relatively small number of variables in this dataset, we have opted to use the original dataset without PCA for model training and testing. This decision was made to maintain interpretability while still addressing multicollinearity issues through alternative methods, such as regularization or feature selection.

3 Model Selection and Methodology

3.1 Simple regression tree

First, we build a simply regression tree to predict car prices. We split the dataset to 70% training and 30% testing and build the initial tree in Figure 12 with training dataset. To prune the tree, we find the optimal complexity parameter by minimizing the cross-validation error (xerror) and balancing the model's complexity to avoid overfitting. And based on our initial model, the lowest xerror is 0.13015, which occurs when cp = 0.010770. So we prune the tree with optimal cp and get the optimal tree model in Figure 13. To evluate the tree, we make a prediction on test data to evaluate the model and get the R-squared: 0.8414521, meaning that about 84.15% of the variation in car prices is explained by the regression tree model. Mean Squared Error is 8759386 meaning there is some room for improvement in the model's accuracy, but this is normal for complex regression models with high variance in the data.

To further improve the model, we tried other regression models like Random Forests or Gradient Boosting Machines (GBM) to compare performance.

3.2 Random Forest model

The Random Forest model, built with 500 trees, achieved an impressive R-squared of 0.9078, indicating that about 90.78% of the variation in car prices was explained by the model. The Mean Squared Error (MSE) was 5,897,930, which is lower than the MSE from the initial regression tree model, suggesting improved accuracy. The model's high performance is reflected in its ability to capture more complex relationships in the data. Key predictors of car prices, such as engine size, curb weight, and fuel efficiency (city and highway mpg), showed significant importance in the model as showed in Figure 14, reinforcing their role in determining car values. This demonstrates that Random Forests, with their ensemble approach, can outperform simpler models by reducing variance and handling complex data patterns more effectively.

3.3 Gradient Boosting Machines

The GBM model achieved an R-squared of 0.8775, meaning it explained 87.75% of the variation in car prices, which is slightly better than the regression tree but slightly lower than the Random Forest model. The RMSE for the GBM model was 2559.667, showing that it still has room for improvement in terms of accuracy, but it performs better than the initial regression tree model. The GBM model also provided insights into variable importance, as shown in Figure 15, where variables such as engine size, curb weight, and width were found to have the most influence on car prices. The learning curve for the model in Figure 16 indicated that further iterations could potentially enhance the model's performance.

Although this is a strong result, it is slightly lower than the Random Forest's R-squared of 0.9078, which suggests that, in this case, the Random Forest model is performing better in terms of accuracy. Both models show improvement over the simple regression tree, but Random Forest appears to provide a more robust fit to the data.

4 Results

4.1 Final Regression Tree Visualization and Interpretation

The regression tree, shown in Figure 12, was constructed using the training data, where we aimed to predict car prices based on various features. After tuning and pruning the tree with an optimal complexity parameter (`cp` = 0.010770), the pruned tree in Figure 13 showed the most important features influencing price. The final model achieved an R-squared of 0.841, meaning it explained 84.1% of the variation in car prices, which is a strong result for predicting price based on the available features.

In terms of performance metrics, the Mean Squared Error (MSE) was found to be 8,759,386, indicating some error, but this is typical in regression models when dealing with a range of car prices. Despite this, the model demonstrated robust predictive power for car prices.

4.2 PCA Results

The PCA analysis reveals significant patterns in the dataset, with the PC1 primarily influenced by variables like curb weight, length, width, engine size, and price, reflecting their joint impact on vehicle size and performance. The PC2 highlights height and compression ratio, indicating a focus on factors like fuel efficiency and body design. Subsequent components capture more nuanced features such as stroke, peak RPM, and bore. The PCA biplot shows that lower-priced cars tend to have smaller engine sizes and lower horsepower, while higher-priced cars are associated with better performance and larger dimensions. While PCA helps reduce dimensionality and multicollinearity, we choose to use the original dataset for modeling to preserve interpretability.

4.3 Performance Metrics and Variable Importance

Random Forests and GBM were also tested to improve model performance. The Random Forest model achieved an impressive R-squared of 0.9078, which explains 90.78% of the variation in car prices. The MSE for Random Forest was 5,897,930, a noticeable improvement over the regression tree, highlighting the model's ability to handle more complex relationships in the data. Feature importance analysis revealed that engine size, curb weight, and fuel efficiency were among the most significant predictors of car prices (Figure 11).

The GBM model, while slightly lower than the Random Forest, showed an R-squared of 0.8775, meaning it explained 87.75% of the variation in car prices. The RMSE of 2559.667 suggested room for further improvement but still represented a substantial reduction in error compared to the initial regression tree.

5 Interpretation, Conclusions and Recommendations

5.1 Insights on What Drives Car Prices

The key factors influencing car prices, based on the regression models, are engine size, curb weight, and fuel efficiency (both city and highway mpg). Engine size and curb weight emerged as the most significant predictors, reinforcing the idea that larger vehicles or those with more powerful engines tend to command higher prices. Additionally, premium vehicle brands like BMW, Mercedes-Benz, and Porsche are associated with higher prices, confirming the influence of brand on price.

The results from the Random Forest model (R-squared = 0.9078) suggest that cars with larger engines and heavier curb weights are likely to be more expensive, as consumers often associate these characteristics with performance, luxury, and durability. Meanwhile, fuel efficiency metrics, such as city and highway mpg, are inversely correlated with price, suggesting that more fuel-efficient cars tend to be priced lower, likely targeting more budget-conscious consumers.

The PCA results align with the regression and Random Forest models, reinforcing the key factors influencing car prices: engine size, curb weight, and fuel efficiency. Engine size and curb weight are strongly associated with the PC1, indicating their significant role in determining car prices. The PCA also shows that more fuel-efficient cars are linked to lower prices, consistent with the Random Forest model's findings. Overall, PCA confirms that larger, heavier vehicles tend to be more expensive, while fuel-efficient cars are typically priced lower.

5.2 Actionable Recommendations for Manufacturers

Based on these findings, car manufacturers should focus on improving performance and engine specifications, as these factors have the greatest impact on pricing. In particular, investing in cars with better engine sizes and curb weight-to-engine-size ratios could help boost mid-range car sales, as these features are associated with higher prices. Manufacturers aiming at luxury markets should also consider emphasizing engine power and larger vehicle dimensions to cater to consumer preferences for high-performance vehicles.

Additionally, fuel efficiency remains a key concern for budget-conscious consumers, so offering a balance between performance and fuel economy will likely appeal to a broader customer base.

5.3 Limitations

It is important to note that the dataset used in this analysis is not exhaustive, as it only covers a limited set of older car models. This limits the ability to generalize the findings to newer car models or more diverse markets. Furthermore, while tree-based models like Random Forest and GBM performed well, there may be additional features or external factors not captured in the dataset that could influence car prices. Future studies could benefit from a more comprehensive dataset that includes more diverse car types and additional market variables, such as regional preferences, economic conditions, or emerging automotive technologies.

5.4 Conclusion

In conclusion, the analysis provides valuable insights into the factors that drive car prices, with engine size, curb weight, and fuel efficiency emerging as the most influential. The regression models showed promising predictive power, especially with Random Forests. Manufacturers can leverage these insights to adjust production strategies and meet consumer demands effectively.

6 Appendix

Variable Name	Description	Data Type
symboling	A numerical code assigned to the vehicle's risk level (e.g., 3 for higher risk)	Numerical
normalized-losses	The number of normalized losses for the vehicle (higher value indicates more damage)	Numerical
wheel-base	The distance between the front and rear axles of the vehicle (measured in inches)	Numerical
length	The overall length of the vehicle (measured in inches)	Numerical
width	The overall width of the vehicle (measured in inches)	Numerical
height	The overall height of the vehicle (measured in inches)	Numerical
curb-weight	The weight of the vehicle without passengers or cargo (measured in pounds)	Numerical
engine-size	The size of the engine in liters (e.g., 130 for 1.3L engine)	Numerical
bore	The diameter of the engine's cylinders (measured in inches)	Numerical
stroke	The length of the engine's pistons' stroke (measured in inches)	Numerical
compression-ratio	Ratio of the cylinder's volume at the bottom of the stroke to the volume at the top	Numerical
horsepower	The engine's power output (measured in horsepower)	Numerical
peak-rpm	The engine's maximum revolutions per minute	Numerical
city-mpg	Miles per gallon the vehicle achieves in the city	Numerical
highway-mpg	Miles per gallon the vehicle achieves on the highway	Numerical
price	The price of the vehicle (in US dollars)	Numerical
make	The manufacturer or brand name of the vehicle (e.g., alfa-romero)	Categorical

Variable Name	Description	Data Type
fuel-type	Type of fuel the vehicle uses (e.g., gas or diesel)	Categorical
aspiration	Type of aspiration in the engine (e.g., std = standard, turbo = turbocharged)	Categorical
num-of-doors	Number of doors in the vehicle (e.g., two or four)	Categorical
body-style	The body style or type of the vehicle (e.g., convertible, sedan)	Categorical
drive-wheels	Type of drive system (e.g., rwd = rear-wheel drive)	Categorical
engine-location	Location of the engine in the vehicle (e.g., front)	Categorical
engine-type	Type of engine used (e.g., dohc = dual overhead camshaft)	Categorical
num-of-cylinders	The number of cylinders in the engine (e.g., four, six, or eight)	Categorical
fuel-system	Type of fuel system used in the vehicle (e.g., mpfi = multi-point fuel injection)	Categorical

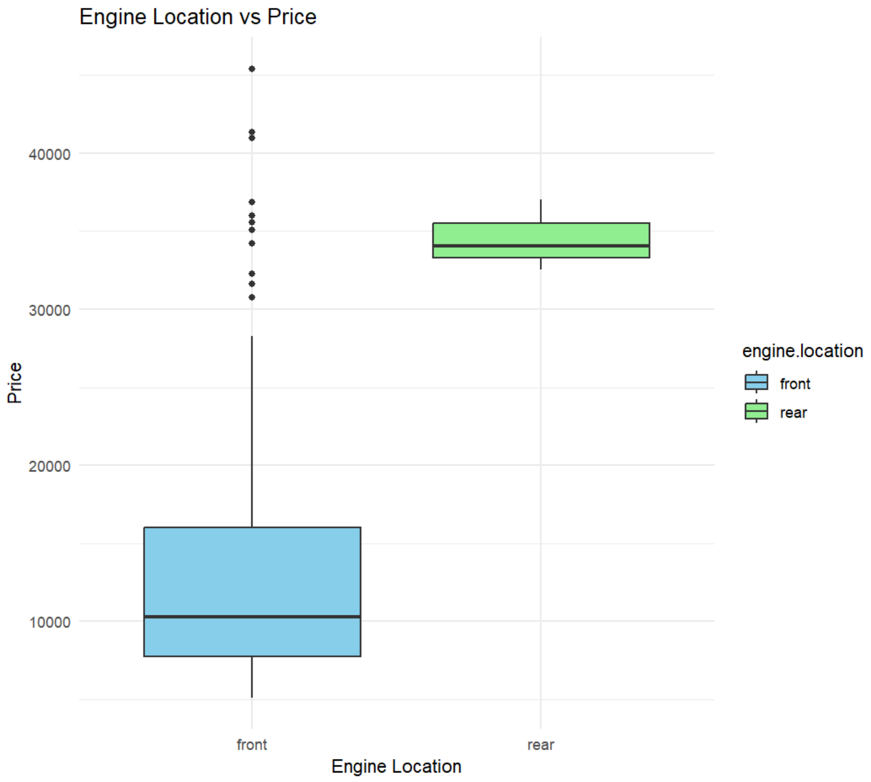


Figure 1: Numerical column distribution

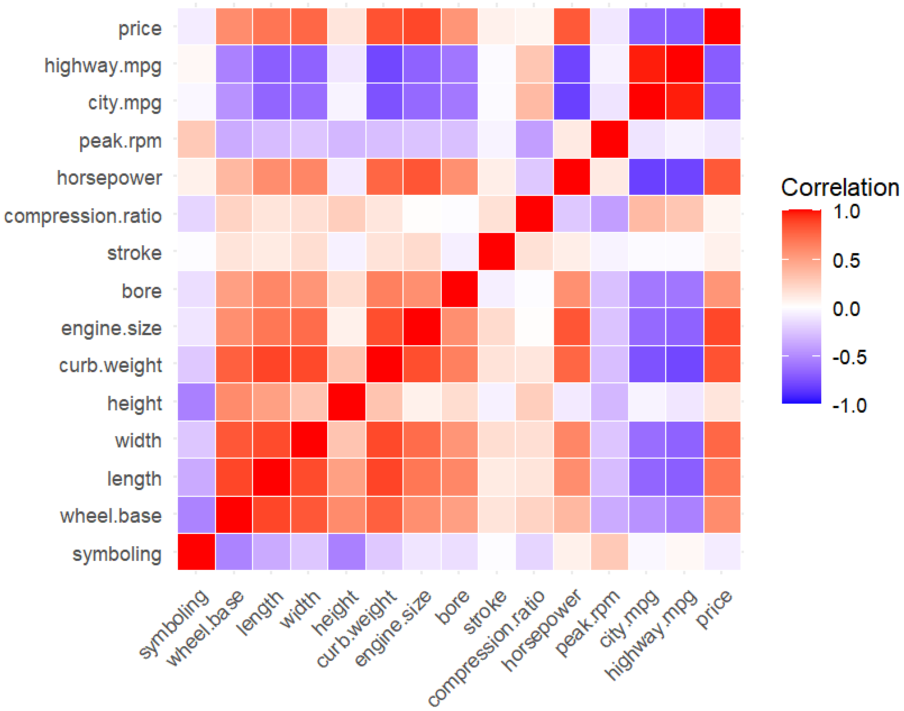


Figure 2: correlation of numeric values

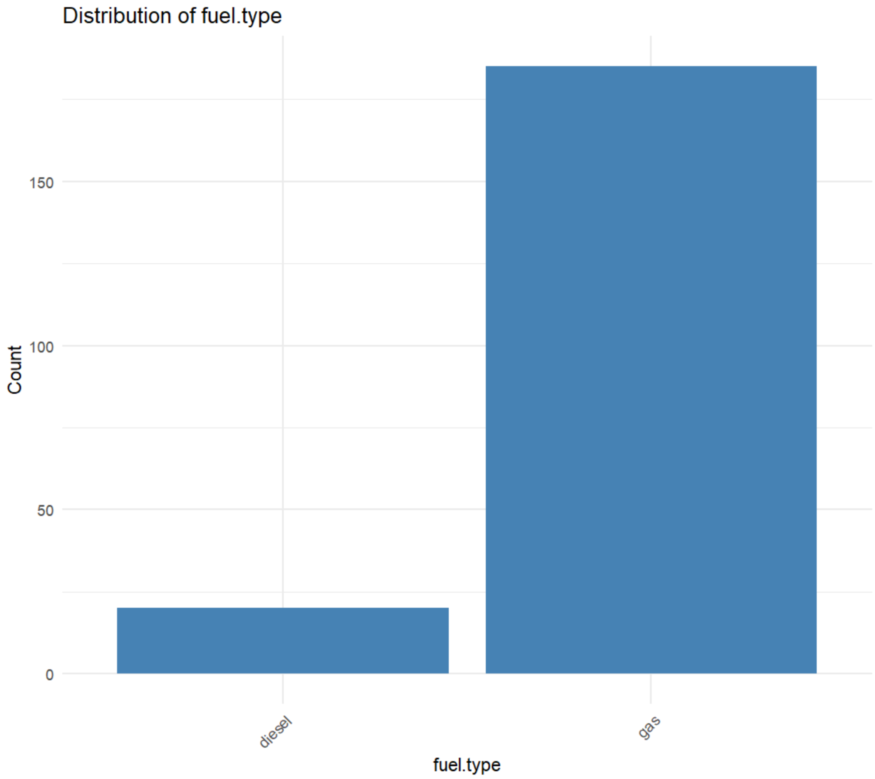


Figure 3: Fuel types

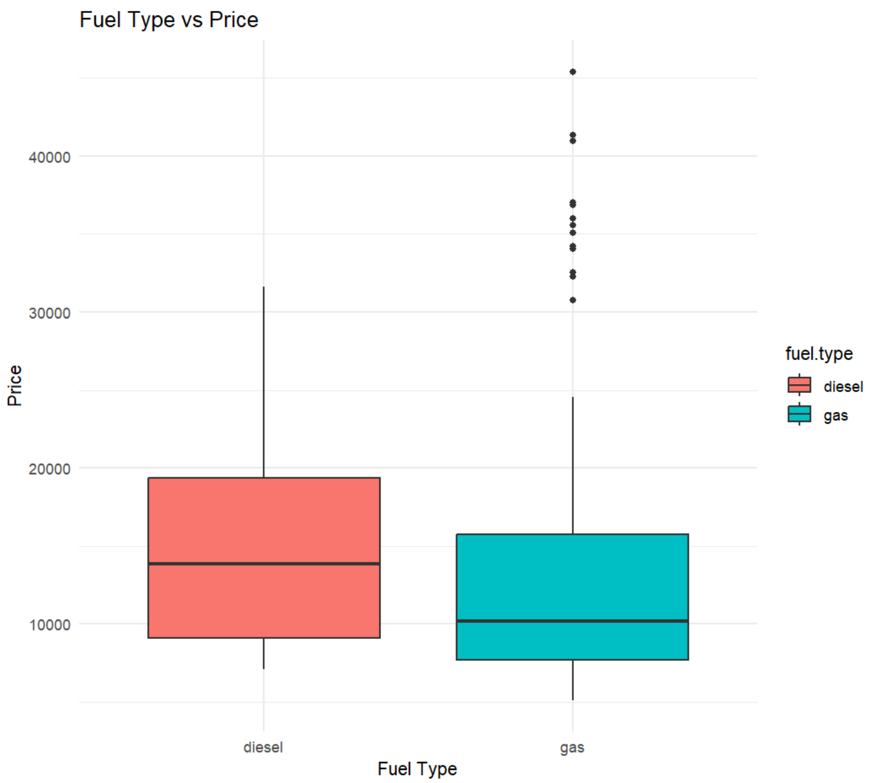


Figure 4: fuel type vs. price

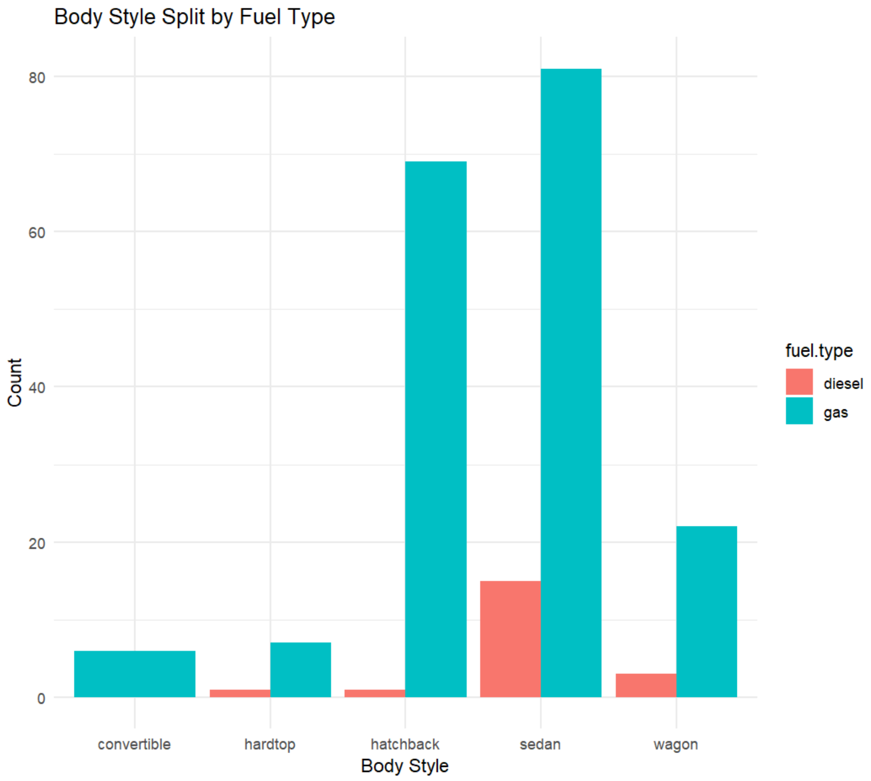


Figure 5: Body style split by fuel types

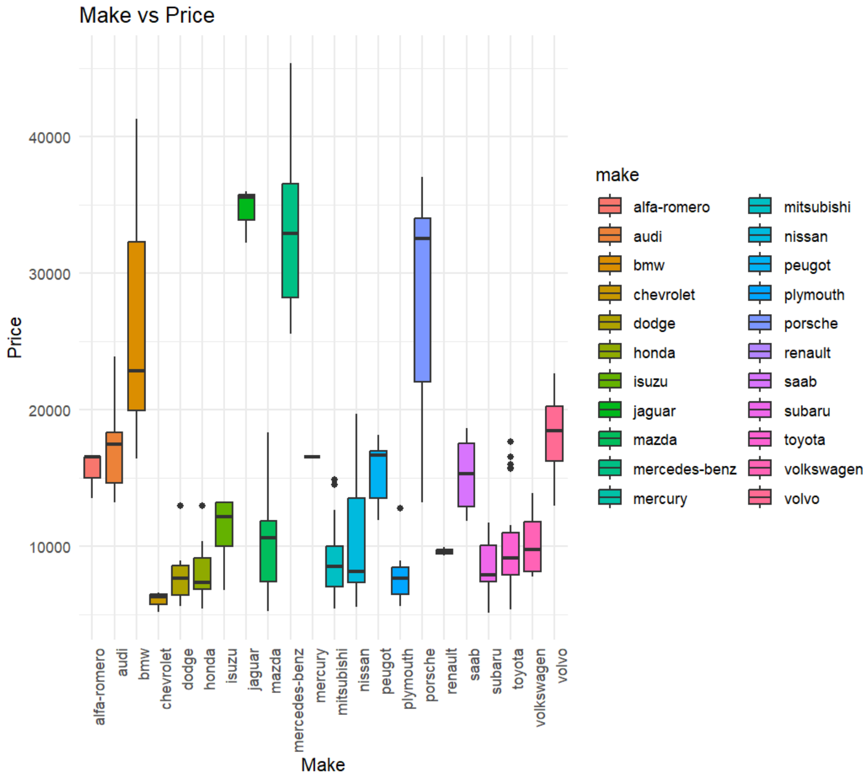


Figure 6: make vs. price

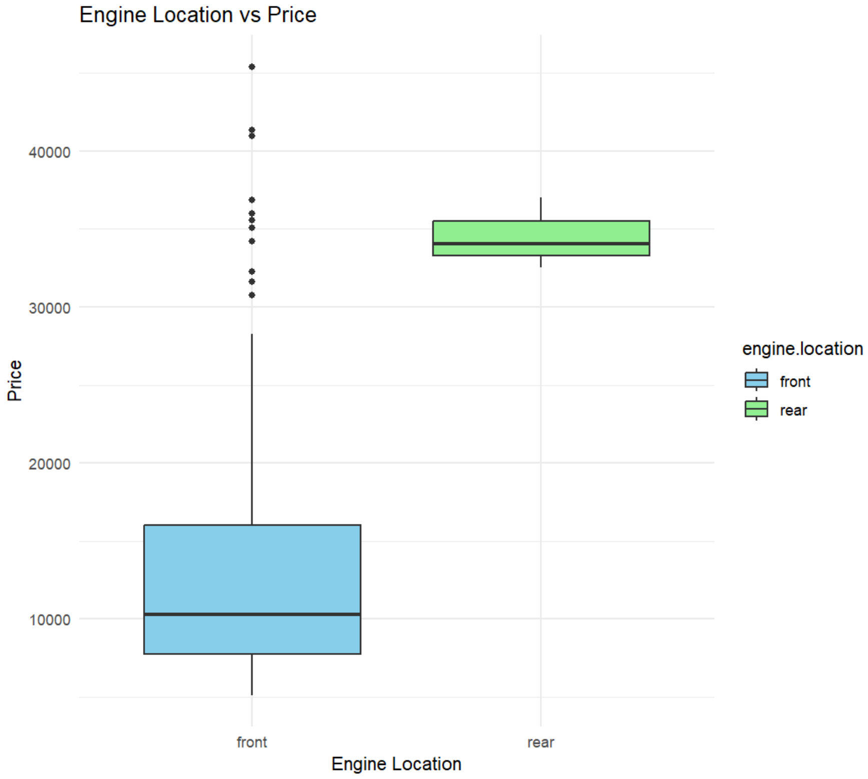
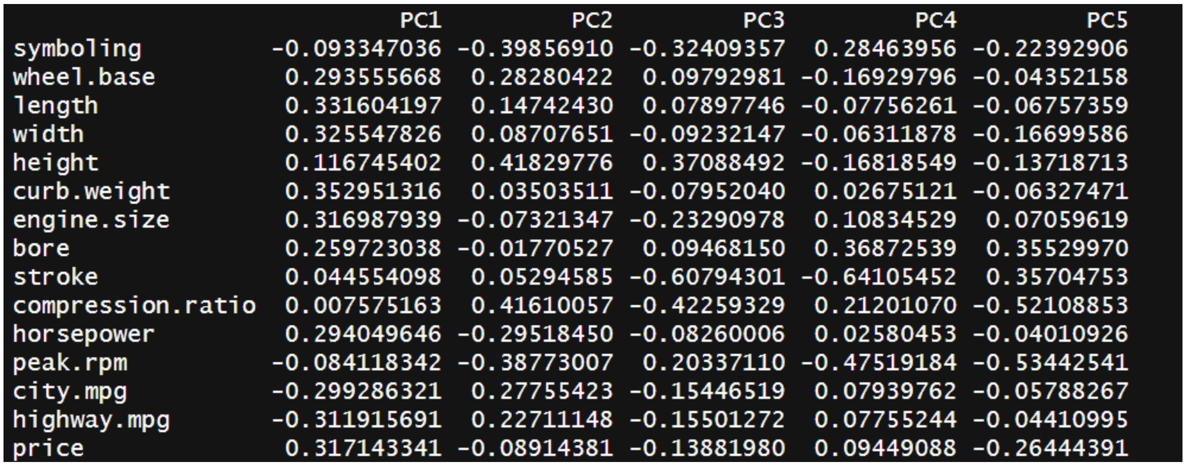


Figure 7: Engine location vs price



```
                            PC1         PC2         PC3         PC4         PC5
symboling         -0.093347036 -0.39856910 -0.32409357  0.28463956 -0.22392906
wheel.base         0.293555668  0.28280422  0.09792981 -0.16929796 -0.04352158
length             0.331604197  0.14742430  0.07897746 -0.07756261 -0.06757359
width              0.325547826  0.08707651 -0.09232147 -0.06311878 -0.16699586
height             0.116745402  0.41829776  0.37088492 -0.16818549 -0.13718713
curb.weight        0.352951316  0.03503511 -0.07952040  0.02675121 -0.06327471
engine.size        0.316987939 -0.07321347 -0.23290978  0.10834529  0.07059619
bore               0.259723038 -0.01770527  0.09468150  0.36872539  0.35529970
stroke             0.044554098  0.05294585 -0.60794301 -0.64105452  0.35704753
compression.ratio  0.007575163  0.41610057 -0.42259329  0.21201070 -0.52108853
horsepower         0.294049646 -0.29518450 -0.08260006  0.02580453 -0.04010926
peak.rpm          -0.084118342 -0.38773007  0.20337110 -0.47519184 -0.53442541
city.mpg          -0.299286321  0.27755423 -0.15446519  0.07939762 -0.05788267
highway.mpg       -0.311915691  0.22711148 -0.15501272  0.07755244 -0.04410995
price              0.317143341 -0.08914381 -0.13881980  0.09449088 -0.26444391
```

Figure 8: The PCA rotation matrix

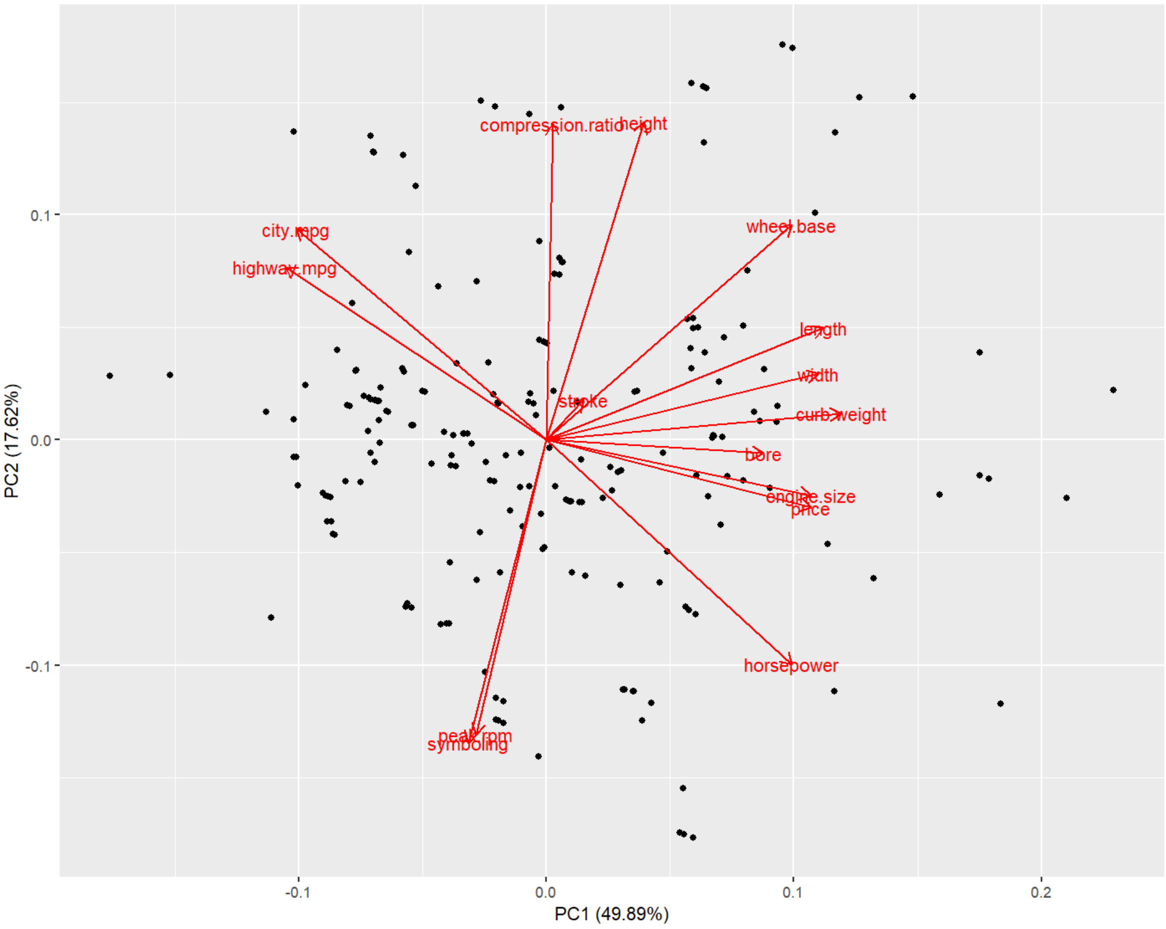


Figure 9: 2D PCA plot

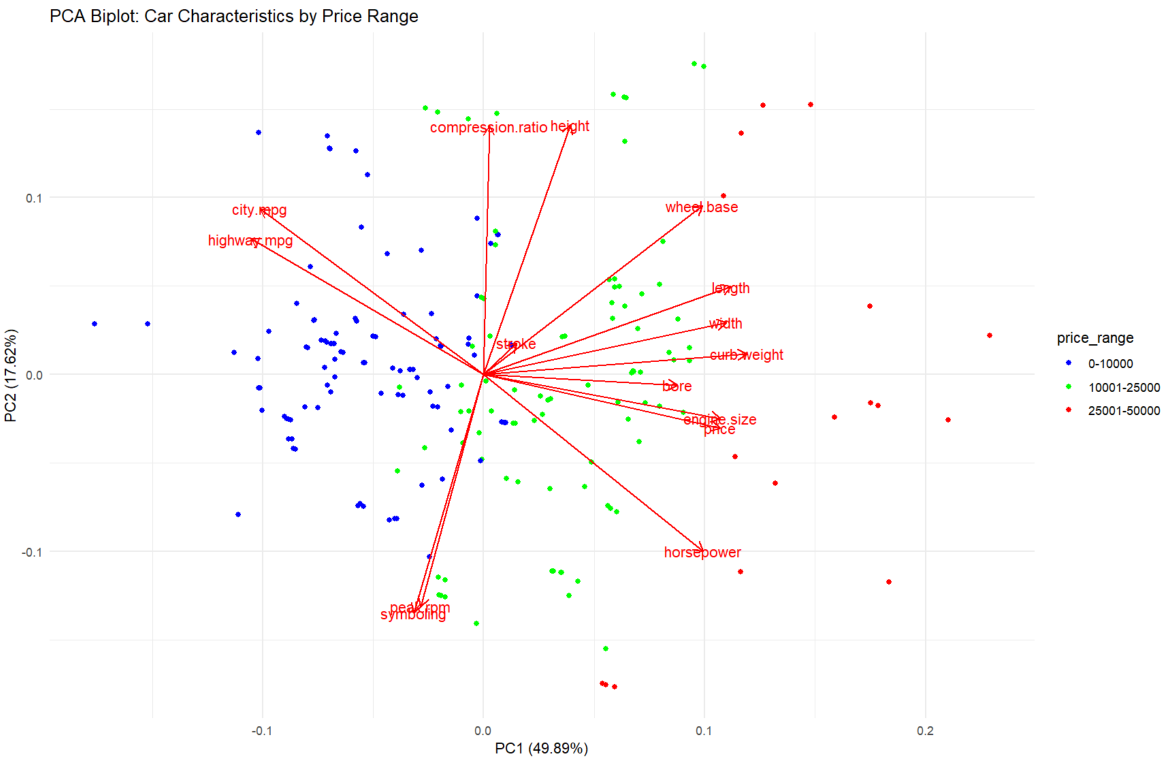


Figure 10: PCA biplot of price range

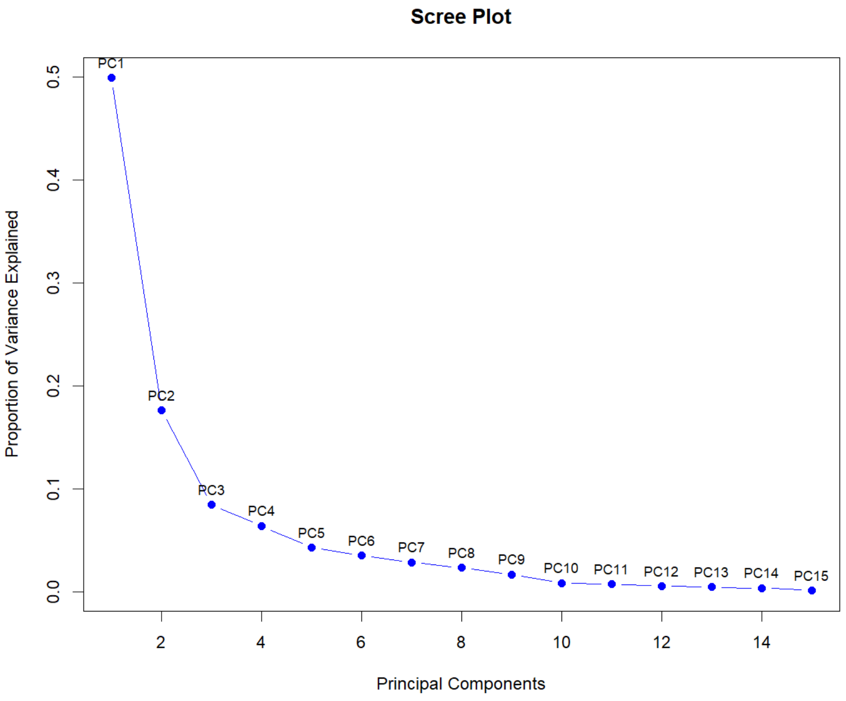


Figure 11: PCA Scree plot

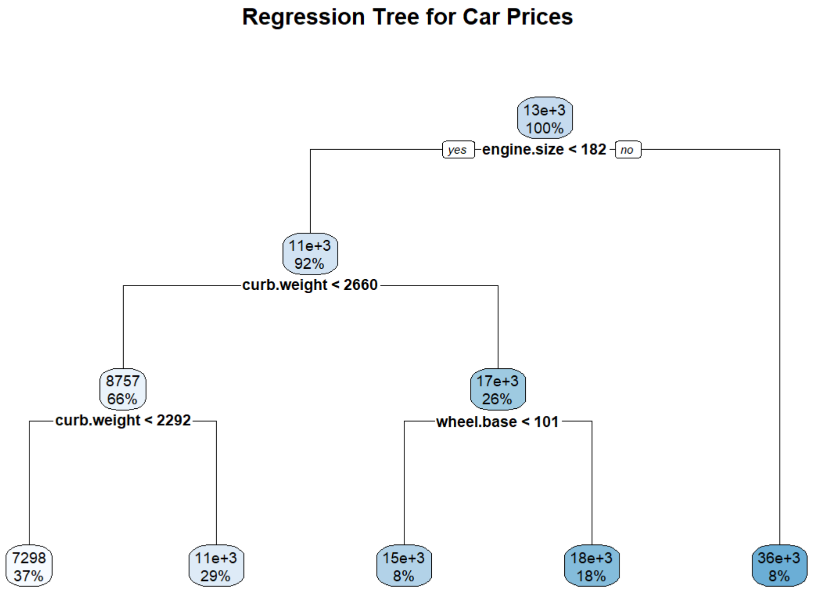


Figure 12: Initial tree model

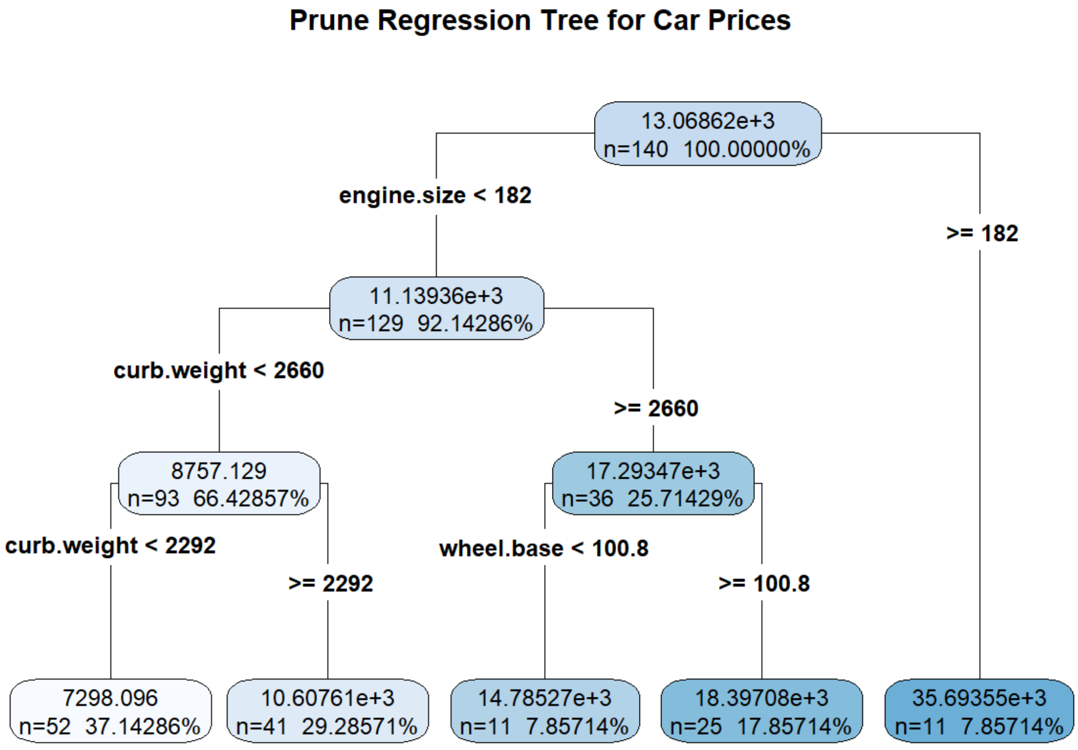


Figure 13: Prune tree model

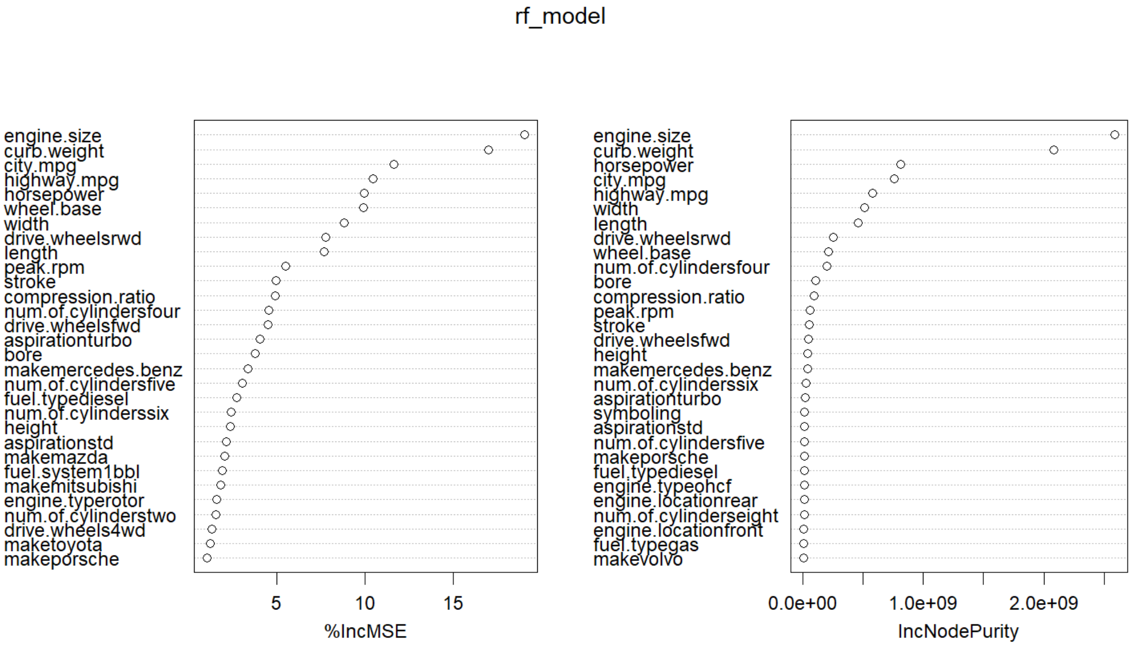


Figure 14: rf model variable importance

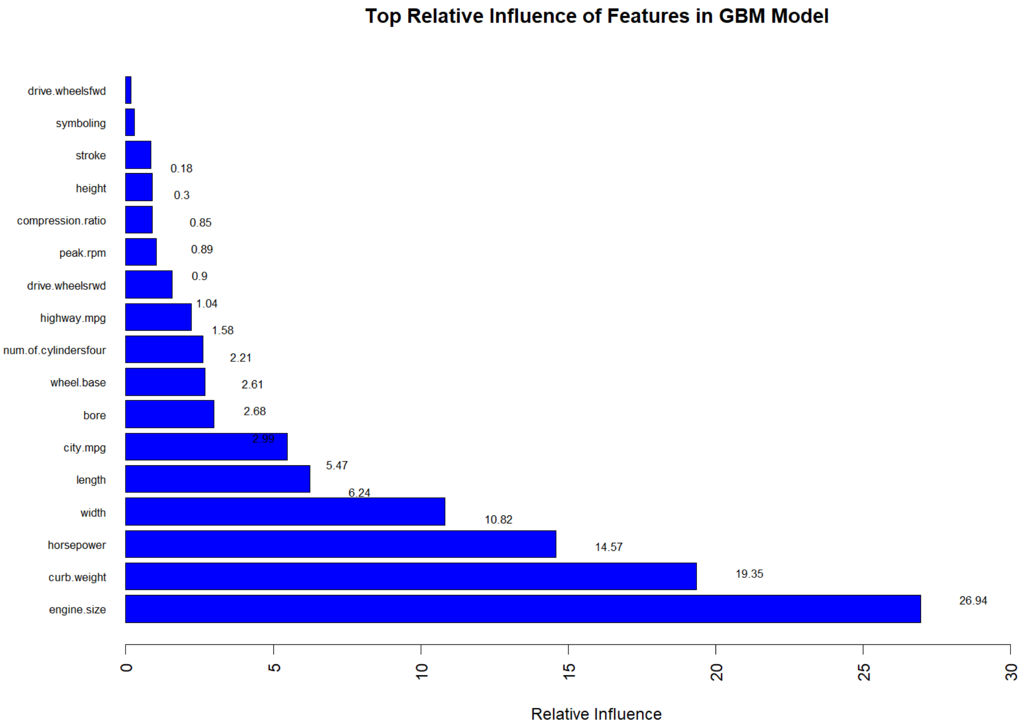


Figure 15: GBM relative influence

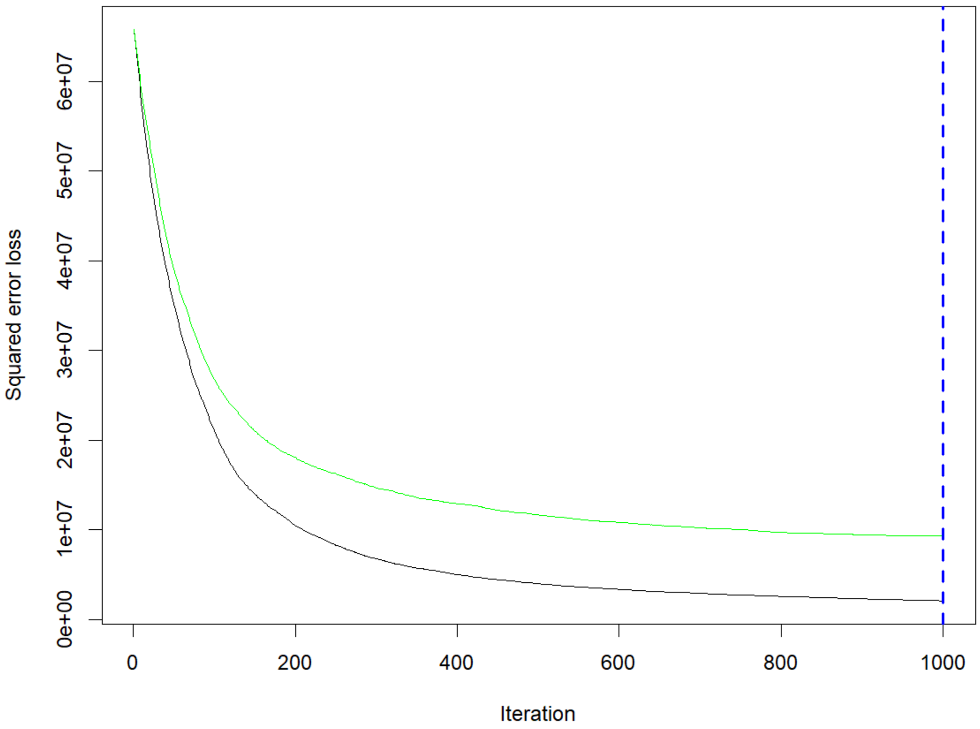


Figure 16: Learning curve