

***Project Title: Predictive & Prescriptive Analysis of Vehicle’s Gas Consumption***

***Course Code: DSCI 726***

***Course Title: Operational Analytics***

***Submitted to: Professor Yi Liu***

***Project Team: Mansu Thapa, Yogesh Siyyadri, Srinivasulu Punugoti***

***Date of Submission: 10/06/2024***

***Project progress – Predictive Analysis***

**Introduction**

This document presents a predictive analysis of vehicle gas consumption, focusing on estimating fuel consumption based on key vehicle specifications. The primary goal of this analysis is to develop a reliable model that forecasts gas mileage (miles per gallon) by evaluating factors such as the number of cylinders, engine displacement, acceleration, horsepower, and vehicle weight. By leveraging statistical and machine learning techniques, this predictive analysis aims to provide insights into how these specifications influence gas consumption, offering valuable predictions for various vehicle types.

**Background Research**

Gas consumption is a critical factor for customers when choosing a vehicle, as it directly impacts running costs and efficiency. Various specifications, such as the number of cylinders, engine displacement, acceleration, horsepower, and vehicle weight, contribute to a vehicle's gas consumption. These factors affect gas mileage differently, with some having a more significant impact than others. Predictive analysis in this project aims to quantify these relationships and forecast gas consumption based on these specifications. By understanding how each of these variables influences fuel efficiency, we can build models that accurately predict gas consumption for different vehicle types.

**Problem Presentation**

As outlined in the Memo sent to General Motors on September 04, 2023, the goal of this predictive analysis is to assist the company in forecasting vehicle fuel consumption based on various specifications. By accurately predicting gas consumption, General Motors can make informed decisions in designing more fuel-efficient vehicles to meet the growing market demand. This will ultimately benefit customers by providing vehicles that offer significant cost savings through improved fuel efficiency.

**Specification and Design**

This analysis focuses solely on the predictive analysis of a vehicle’s gas consumption. The available dataset contains 398 records, which provides a sufficient foundation for developing an accurate predictive model. The primary objective of this project is to build and evaluate different models to forecast the vehicle's miles per gallon (mpg) based on key specifications, including the number of cylinders, displacement, acceleration, and weight. Successful completion of this analysis will result in identifying the most effective predictive model that can accurately forecast mpg based on these vehicle specifications.

**Data Acquisition**

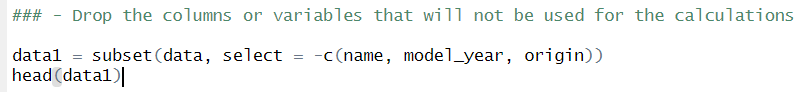
The source of the data is <https://www.kaggle.com/datasets/whenamancodes/automobiles-project-dataset>.

**Data Exploration**

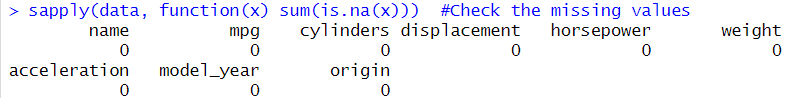
We already explored the data and its structure in the descriptive analysis submitted in the last week

**Data Transformation**

1. **Dropping Irrelevant Variables:**  
   The dataset contained variables such as the vehicle’s name, model year, and origin, which were not directly related to gas consumption. These variables were deemed unnecessary for building an accurate predictive model, and thus, were removed. This step helped reduce complexity and improve the model's focus on more relevant factors affecting fuel consumption.



1. **Handling Outliers:**  
   Outliers in the response variable (miles per gallon, mpg) were identified in the dataset. These extreme values could have skewed the results and negatively impacted the accuracy of the predictive analysis. To ensure the model's reliability, records containing these outliers were excluded from the analysis.
2. **No Missing Data:**  
   Upon reviewing the dataset retrieved from Kaggle, it was confirmed that there were no missing records. This allowed for a smooth analysis process without the need for data imputation or additional handling for incomplete entries.



**List of predictive models used in this analysis:**

1. **Multiple Linear Regression (MLR):**

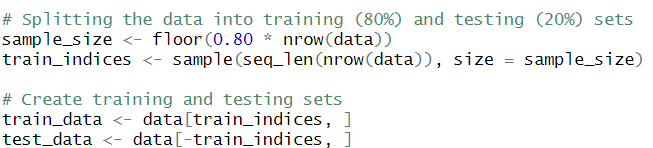
This method was chosen due to its simplicity and interpretability. It helped us understand the linear relationships between the predictors and the response variable (MPG).

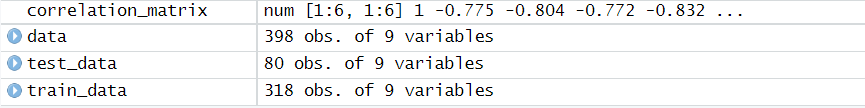
1. **Regression Tree:**

We also employed a regression tree to capture non-linear relationships and interactions between variables that MLR might not fully address.

**Data Partitioning**

1. Training and Testing: The dataset was split into training and testing sets to ensure that the predictive models could generalize well to unseen data. Models were trained on the training set and their performance evaluated on the test set.





1. Overfitting: To prevent overfitting, 80% of the data was used for training and 20% for testing, ensuring the model performs well on new data.

**Model Building & Evaluation**

In this analysis, two predictive models were developed to estimate vehicle gas consumption based on its specifications: **Multiple Linear Regression (MLR)** and **Regression Tree**. The models were trained on 80% of the dataset and evaluated on the remaining 20% using **Mean Squared Error (MSE)** and **R-Squared** as performance metrics.

* **MSE** quantifies the average squared difference between actual and predicted values.
* **R-Squared** measures how well the model explains the variation in the response variable (gas consumption).

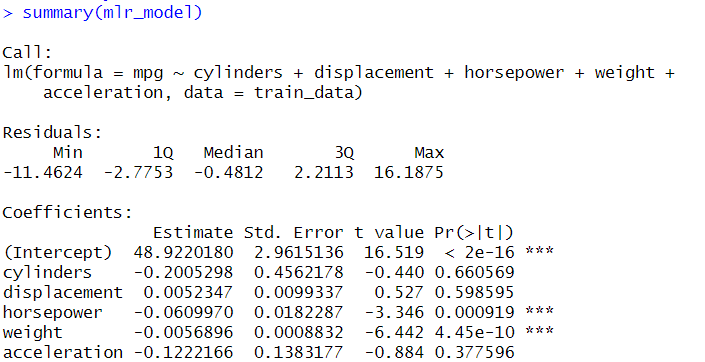
**Results**:

1. **Multiple Linear Regression (MLR)**:
   * Training Set: MSE = 17.62, R-Squared = 71.7%, Testing Set: MSE = 20.21, R-Squared = 63.40. A close up of numbers

     Description automatically generated A close-up of a sign

     Description automatically generated

Multiple linear regression has a higher **Training MSE** (17.62) and **Testing MSE** (20.21), and lower **R-squared** values than the Regression Tree. However, the difference in performance between training and testing is less extreme, suggesting that MLR might generalize slightly better or overfit less than the regression tree.



A graph of a number of dots

Description automatically generated with medium confidence

1. **Regression Tree**:
   * Training Set: MSE = 14.42, R-Squared = 76.8%
   * Testing Set: MSE = 18.93, R-Squared = 65.7%

A close up of a white background

Description automatically generated

A close-up of a sign

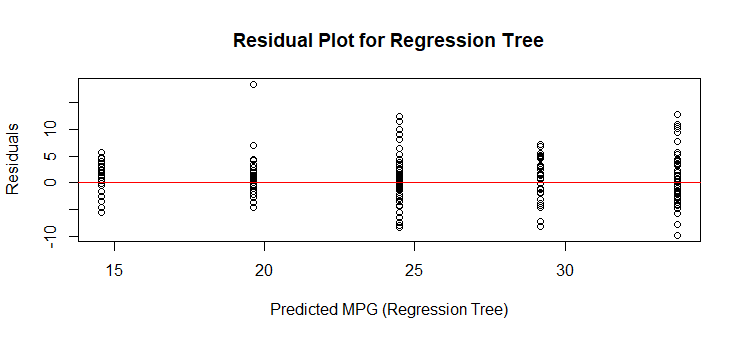
Description automatically generated

Regression Tree performs better on both the training and testing sets compared to the MLR model.

The Training MSE (14.42) and Testing MSE (18.93) are both lower than the corresponding values for the MLR model.

The Training R-squared (0.768) and Testing R-squared (0.657) are higher than those of the MLR, indicating that the regression tree has more accuracy in predicting.

While both models performed well, **Regression Tree** offers better generalization and simplicity, making it the preferred model for predicting vehicle gas consumption based on specifications.



The plot likely shows random scattering of residuals, meaning the model's predictions are reasonably good without systematic errors.

The residuals span a range from -10 to 10, indicating some variability in the model's predictions, but no significant issues like overfitting or bias based on this plot alone.

A residual plot with no distinct pattern confirms that the model is performing adequately, though the range of errors could be explored further if the variability seems large.

In summary, this residual plot suggests that the Regression Tree model has a decent fit for predicting MPG but does have some errors, as shown by the spread of residuals around 0. The horizontal red line helps visualize how far off the predictions are from the actual values.

A graph of a tree model

Description automatically generated

The Variable Importance plot reveals how each predictor influences the response variable (MPG) in the Regression Tree model:

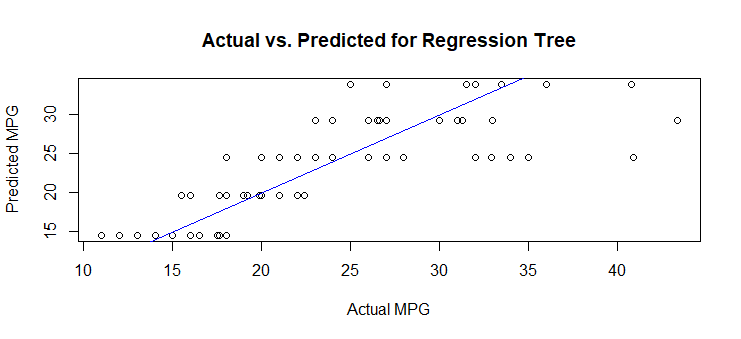
Displacement and Weight are the most significant predictors, both exceeding an importance score of 12,500, indicating they substantially impact gas consumption.

Cylinders also contribute notably, with an importance score between 10,000 and 12,500.

Horsepower shows moderate influence, slightly above 10,000.

Acceleration is the least important predictor, with a score around 5,000, suggesting it has a minor effect on fuel efficiency predictions.

This plot identifies the key predictors to focus on for optimizing vehicle gas consumption.



Approximately 70% of the data points are clustered close to the blue line, indicating a good alignment between predicted and actual values for these cases.

However, some points deviate from the line, suggesting that the model struggles to accurately predict MPG for certain vehicles.

Overall, this plot demonstrates that while the model performs reasonably well, there are instances where it could improve, particularly for specific ranges of actual MPG values.

A graph with red and blue dots

Description automatically generated

* Most of the blue and orange points are clustered near the reference line, suggesting that both models perform similarly in predicting MPG.
* There are only a few points that deviate significantly from the line, indicating minor discrepancies in predictions.
* However, the orange points (MLR) show a slightly lower frequency of outliers compared to the blue points (Regression Tree), implying that the Multiple Linear Regression model may have a slight edge in maintaining prediction accuracy.

Overall, this plot highlights that both models perform well, with the regression tree model showing slightly better predictive consistency.

**Conclusion:**

In the predictive analysis aimed at estimating gas consumption based on vehicle specifications, the Regression Tree model demonstrated superior performance compared to the Multiple Linear Regression (MLR) model. The results indicate that the Regression Tree achieved a Mean Squared Error (MSE) of 18.93 and an R-Squared value of 65.7% on the testing set, while the MLR model had an MSE of 20.21 and an R-Squared of 63.40%.

The findings underscore the importance of selecting an appropriate modeling technique for predictive tasks. The Regression Tree effectively captured the non-linear relationships within the data, making it more adept at predicting miles per gallon (mpg) for unseen data compared to the MLR approach.

Overall, the analysis provided valuable insights into the factors influencing gas consumption, with the Regression Tree model being adopted for further applications and considerations in optimizing vehicle efficiency.