REPORT ON BUILDING THE CAR OF THE FUTURE: A DATA-DRIVEN APPROACH TO FUEL EFFICIENCY

Daniel Makala Mackaemba

Northeastern University – Master of Professional Studies in Data Analytics

Data Analytics and Predictive Modeling

Date: January 21, 2026

Abstract

This report presents a data-driven analysis aimed at helping a large automobile manufacturer improve vehicle fuel efficiency. Using historical vehicle data and multiple linear regression modeling, the study identifies the vehicle attributes that most strongly influence miles per gallon (MPG). After applying rigorous data cleansing techniques, a predictive model was developed and optimized using feature selection methods. The final model explains approximately 82% of the variance in fuel efficiency, with vehicle weight, model year, and engine displacement emerging as the most influential factors. The findings provide actionable recommendations to guide the design of more energy-efficient vehicles.

1. Introduction

1.1 Background and Problem Statement

The automotive industry is experiencing increased pressure to produce vehicles that are both environmentally sustainable and cost-efficient. A manufacturer traditionally focused on producing large vehicles is facing declining sales due to poor fuel economy. This project applies predictive analytics to identify which vehicle characteristics most significantly affect MPG, enabling the company to redesign future models with improved fuel efficiency.

1.2 Objectives of the Study

The objectives of this analysis are to:

- Identify vehicle attributes that significantly impact MPG
- Develop a regression-based model to predict fuel efficiency
- Optimize the model using feature selection techniques
- Provide data-driven recommendations for fuel-efficient vehicle design

1.3 Analytical Framework

This analysis follows the CRISP-DM methodology, including business understanding, data preparation, modeling, evaluation, and interpretation. This structured approach ensures analytical rigor and alignment with business goals.

2. Data Description and Exploration

2.1 Dataset Overview

The dataset consists of 398 historical vehicle observations with the following variables:

- MPG: Miles per gallon (dependent variable)
- Cylinders: Number of engine cylinders
- Displacement: Engine displacement (cubic inches)
- Horsepower: Engine power output
- Weight: Vehicle weight (pounds)
- Acceleration: Time to accelerate from 0–60 mph
- Model Year: Year of manufacture
- Origin: Geographic manufacturing origin

2.2 Initial Observations

Exploratory analysis revealed strong negative correlations between MPG and both vehicle weight and engine size, while model year showed a positive relationship with fuel efficiency. Data quality issues, particularly missing values in the horsepower variable, were identified and addressed prior to modeling.

3. Data Cleansing and Preparation

3.1 Data Quality Assessment

Initial inspection identified missing and improperly formatted values in the horsepower column, along with potential outliers in weight and engine-related variables.

3.2 Data Cleaning Decisions

Handling Missing Values

Non-numeric values in the horsepower column were converted to missing values and removed using listwise deletion. Only 1.5% of observations were affected, preserving the integrity of the dataset while ensuring reliable modeling.

Data Type Conversion

Horsepower values were converted from string to numeric format to support statistical analysis and modeling.

Outlier Considerations

Although some extreme values were present, all observations were retained because they represent realistic vehicle configurations relevant to the manufacturer's market.

3.3 Final Dataset Summary

After cleaning, the dataset contained 392 complete observations with no missing values. This clean dataset served as the foundation for regression modeling.

4. Linear Regression Model Development

4.1 Model Specification

A multiple linear regression model was developed using MPG as the dependent variable and all remaining attributes as predictors.

4.2 Model Training and Evaluation

The dataset was split into training (80%) and testing (20%) subsets. The model achieved the following performance on the test data:

R^2 Score: 0.82

RMSE: Approximately 3.4 MPG

These results indicate strong predictive performance and minimal overfitting.

4.3 Interpretation of Key Predictors

Weight: The strongest negative predictor of MPG; heavier vehicles consume more fuel.

Model Year: Strong positive relationship with MPG, reflecting technological advancement.

Displacement: Larger engines are associated with lower fuel efficiency.

Horsepower and Cylinders: Negatively related to MPG, though partially redundant due to correlation with displacement.

5. Model Optimization and Feature Selection

5.1 Feature Selection Approach

Recursive Feature Elimination (RFE) was applied to reduce multicollinearity and improve interpretability while maintaining predictive accuracy.

5.2 Optimized Model Results

The optimized model retained four key features:

- Weight
- Model Year
- Displacement
- Origin

The reduced model showed only a marginal decrease in performance, confirming that these variables capture most of the predictive power.

5.3 Model Effectiveness

The optimized model satisfies the project's objectives by providing accurate predictions and clear guidance for vehicle design decisions.

6. Business Implications and Recommendations

Based on the analysis, the following recommendations are proposed:

1. Reduce Vehicle Weight: Use lightweight materials and optimized designs to improve MPG.
2. Leverage Modern Technology: Incorporate newer engine and transmission technologies.
3. Optimize Engine Size: Reduce displacement while maintaining performance through advanced engineering solutions.
4. Benchmark Efficient Designs: Study best practices from historically fuel-efficient manufacturers.

Implementing these strategies could result in a 15–20% improvement in fuel efficiency.

7. Limitations and Future Work

While the model explains a large portion of MPG variability, it does not account for factors such as aerodynamics, tire resistance, or driving behavior. Future studies could incorporate non-linear models or additional variables to further improve accuracy.

8. Conclusion

This study demonstrates that linear regression modeling is an effective tool for identifying the primary drivers of vehicle fuel efficiency. Vehicle weight, model year, and engine displacement were found to be the most influential factors affecting MPG. By applying data-driven insights, the manufacturer can make informed design decisions that improve fuel efficiency, competitiveness, and sustainability.

References

U.S. Environmental Protection Agency. (n.d.). Fuel economy data. https://www.epa.gov