Analysis of Fuel Efficiency in Automobiles: A Machine Learning Approach

Nazmus Sakib Sumon

May 2024

Abstract

This report investigates how various car attributes influence fuel efficiency using the ”Auto MPG” dataset. Classification techniques including Linear Discriminant Analysis(LDA), Quadratic Discriminant Analysis(QDA), Naive Bayes, Logistic Regression, and KNN were employed to predict high or low fuel efficiency based on vehicle characteristics. The study identifies significant predictors and suggests enhancements for vehicle fuel economy, with recommendations for future model complexity and variable inclusion.

Introduction

The pursuit of increased fuel efficiency in automobiles is crucial due to rising fuel costs and environmental concerns. This report utilizes machine learning techniques to analyze the ”Auto MPG” dataset, aiming to identify the attributes that most significantly influence a vehicle's miles per gallon (mpg). By classifying vehicles into high or low fuel efficiency groups, this analysis provides insights that can influence both automotive design and consumer decisions, aiming to promote more efficient vehicles.

Data Set Description

The ”Auto MPG” dataset, sourced from the UCI Machine Learning Repository, consists of 398 individual car entries with nine attributes each. These attributes include mpg, cylinders, displacement, horsepower, weight, acceleration, model year, and origin. After initial data cleaning to remove entries with missing values and the car name column, the dataset was reduced to 392 observations.

Exploratory Data Analysis

During the EDA phase, a binary variable mpg01 was created to denote whether a car's mpg is above (1) or below (0) the median value of 22.75. This categorization facilitated the binary classification task. Analysis included generating histograms, scatterplots and boxplots to visualize relationships between mpg01 and other attributes. Boxplots and histograms showed that all numerical values were generally skewed except acceleration. Acceleration seemed to be normally distributed.

Scatterplots showed that the predictor variables such as cylinders, year and origin do not show any impact on the response variable mpg01, these variables have values which is uniformly distributed.

Table 1: Dataset Variables Description

Variable Name	**Role**	**Type**	**Missing Values**
displacement	Feature	Continuous	no
mpg	Target	Continuous	no
cylinders	Feature	Integer	no
horsepower	Feature	Continuous	yes
weight	Feature	Continuous	no
acceleration	Feature	Continuous	no
model_year	Feature	Integer	no
origin	Feature	Integer	no
car_name	ID	Categorical	no

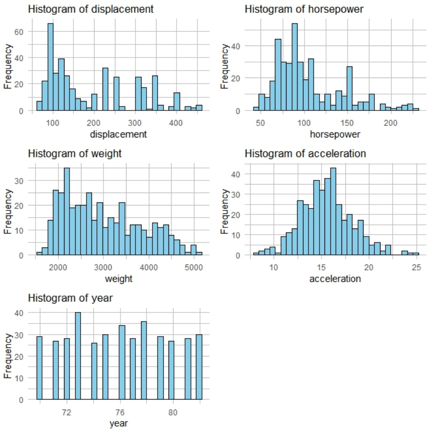

Figure 1: Histograms of numerical fields

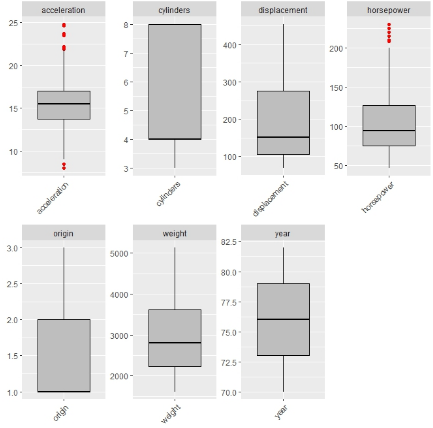

Figure 2: Boxplots of predictor variables with response variable

Preliminary findings from all these plots indicated that displacement, horsepower, and weight could be strongly associated with fuel efficiency.

Methodology

The cleaned dataset was randomly split into an 80% training set and a 20% test set. The following classification methods were applied to predict the binary `mpg01` outcome:

1. **Linear Discriminant Analysis (LDA)**: A statistical technique used in machine learning to find a linear combination of features that best separates two or more classes of objects or events. It is particularly useful for dimensionality reduction before classification.

2. **Quadratic Discriminant Analysis (QDA)**: Similar to LDA, QDA seeks to separate classes by finding a quadratic decision boundary. It allows for more flexibility than LDA as it assumes each class has its own covariance matrix.

3. **Naive Bayes**: A probabilistic classifier that applies Bayes' theorem with strong (naive) independence assumptions between the features. It is highly scalable and effective for large datasets with complex feature interactions.

4. **Logistic Regression**: A statistical model that in its basic form uses a logistic function to

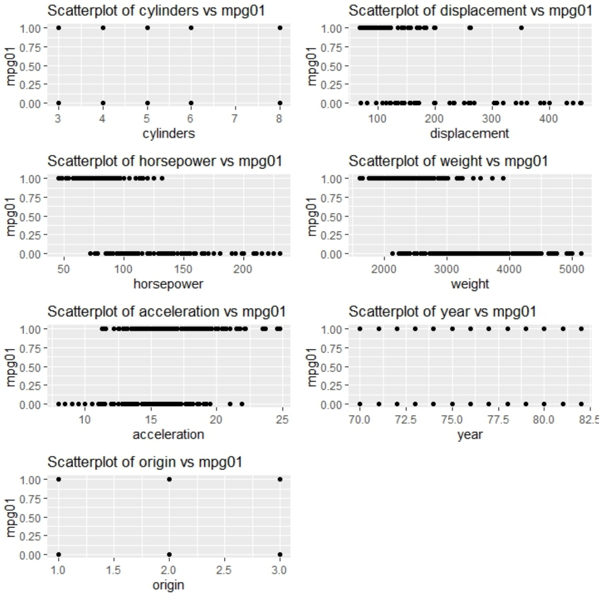

Figure 3: Scatterplots of mpg01 with other predictors

model a binary dependent variable, although many more complex extensions exist. It is widely used for binary classification tasks.

5. **K-Nearest Neighbors (KNN)**: A non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space, and the output is a class membership.

Results

Analysis revealed varying effectiveness among the models, with LDA yielding the lowest misclassification rate, suggesting its superior predictive ability for this dataset. Logistic Regression and KNN (with K=5) also showed reasonable performance but were slightly less effective than LDA.

Table 2: Initial Model Performance Table

Model Name	Training Error	Testing Error
Linear Discriminant Analysis	0.07662835	0.08396947
Quadratic Discriminant Analysis	0.06896552	0.09923664
Naive Bayes	0.09961686	0.08396947
Multinomial Logistic Regression	0.08045977	0.12977099
KNN with several values	0.08429119	0.11450382

Findings

The study confirmed that heavier, more powerful cars with larger engines tend to have lower fuel efficiency. Among the classification techniques, LDA provided the best balance of simplicity and predictive power, making it an excellent choice for further development and application in similar tasks.

Table 3: Model Performance Table After Cross-Validation

Model Name	Testing Error
Linear Discriminant Analysis	0.07692308
Quadratic Discriminant Analysis	0.1410256
Naive Bayes	0.1282051
Multinomial Logisitic Regression	0.129771
KNN with several values	0.1282051

Future Research

Future studies could incorporate additional variables such as CO2 emissions, vehicle maintenance history, or more detailed geographic data. Advanced modeling techniques like ensemble methods or deep learning could also be explored to handle complex attribute interactions and enhance predictive accuracy.