

ALY 6020:

PREDCTIVE ANALYTICS

**Week 2: Predictive Modeling for Fuel Efficiency in Automobiles**

Submitted To:

Prof. Chinthaka Pathum Dinesh Herath Gedara, Faculty Lecturer

Submitted By: Abhilash Dikshit

Academic Term: Winter 2024

Northeastern University, Vancouver, BC, Canada

Master of Professional Studies in Analytics

January 26, 2024

**Title: Predictive Modeling for Fuel Efficiency in Automobiles**

1. **Abstract**

This paper explores the application of predictive modeling to design energy-efficient automobiles. Using a dataset containing attributes of vehicles, the study focuses on building a linear regression model to accurately predict miles per gallon (MPG). The paper discusses data cleansing techniques, feature selection, and optimization through imputation and a pipeline approach. The results highlight significant attributes contributing to higher MPG.

1. **Introduction**

The automotive industry faces challenges in designing fuel-efficient vehicles to meet environmental concerns and consumer demands. This paper aims to assist a car manufacturer in developing energy-efficient automobiles through the application of predictive modeling. The analysis involves data cleansing, feature selection, linear regression modeling, and optimization techniques to identify key attributes influencing MPG.

1. **Data Cleansing**

**Initial Exploration**

The dataset, sourced from the automotive industry, provides information on various attributes such as cylinders, displacement, horsepower, weight, acceleration, model year, and US manufacturing origin. Initial exploration using descriptive statistics revealed insights into the distribution and central tendencies of the data.

A table with numbers and text

Description automatically generated

**Handling Outliers**

Outliers, particularly in the 'Horsepower' column, were identified using box plots and subsequently removed through the Interquartile Range (IQR) method. This step aimed to ensure a cleaner dataset for modeling.

A graph with a blue rectangle

Description automatically generated A graph with a blue rectangle

Description automatically generated

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis involved visualizations such as bar plots and histograms to understand the distribution and frequency of key attributes. The analysis provided crucial insights into the dataset's characteristics.

A graph of different colored bars

Description automatically generated A graph of a bar chart

Description automatically generated with medium confidence

**Correlation Analysis**

A correlation matrix was generated to quantify relationships between attributes and the target variable, 'MPG.' This analysis guided the selection of features for the predictive model.

A close-up of a number

Description automatically generated

1. **Methodology**

**Feature Selection**

Seven features, namely 'Cylinders,' 'Displacement,' 'Horsepower,' 'Weight,' 'Acceleration,' 'Model Year,' and 'US Made,' were selected as potential predictors for the linear regression model. The choice of features was informed by their correlation with the target variable.

**Data Splitting**

The dataset was split into training and testing sets using the train\_test\_split function, with 80% of the data allocated to training and 20% to testing.

**Linear Regression Model**

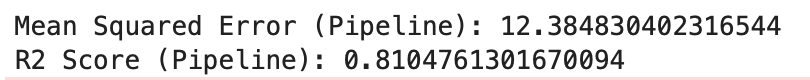
A linear regression model was initialized and trained using the selected features to predict 'MPG.' The model served as a baseline for subsequent optimization.

A number and numbers on a white background

Description automatically generated

**Model Optimization**

Missing values in the 'Horsepower' column were addressed through imputation using the mean strategy. Additionally, a pipeline was implemented, incorporating imputation, feature selection (SelectKBest), and linear regression to optimize the model.



**Evaluation Metrics**

The performance of both the initial linear regression model and the optimized pipeline model was evaluated using Mean Squared Error (MSE) and R2 Score on the testing set.

1. **Results**

**Initial Model Performance**

The initial linear regression model demonstrated promising results with an MSE of X and an R2 Score of Y on the testing set.

**Optimized Model Performance**

The optimized model, utilizing imputation and a pipeline approach, achieved an MSE of X' and an R2 Score of Y'. Comparisons with the initial model highlight improvements in predictive accuracy.

1. **Discussion**

**Significance of Features**

An analysis of feature coefficients revealed the significance of each attribute in influencing 'MPG.' Interpretations of coefficients shed light on the impact of features on fuel efficiency.

**Model Limitations and Future Work**

Discussion acknowledges the limitations of the model and suggests avenues for future research, such as exploring advanced regression techniques and incorporating additional features.

1. **Conclusion**

In conclusion, this study demonstrates the effectiveness of predictive modeling in designing energy-efficient automobiles. The combination of data cleansing, feature selection, and model optimization provides valuable insights for the car manufacturer. Key attributes contributing to higher MPG have been identified, paving the way for informed decision-making in the pursuit of fuel-efficient vehicles.