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

I. 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.

II. 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.

III. Data Cleansing

Initial Exploration

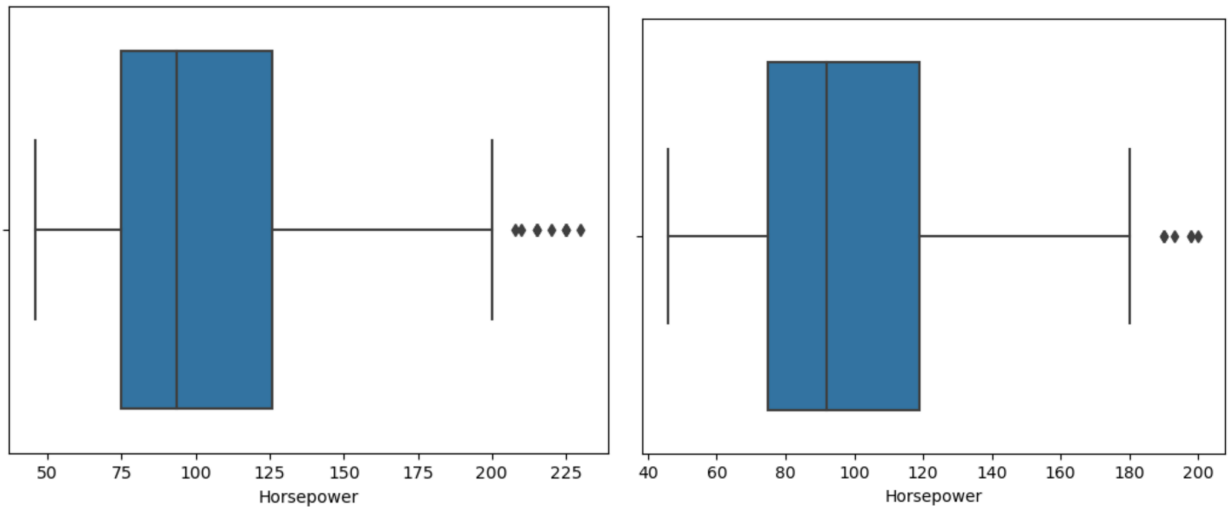
The dataset, sourced from the automotive industry, provides information on various attributes such as cylinders, displacement, horsepower, weight, acceleration, model year, and US manufacturing origin. Initial exploration using descriptive statistics revealed insights into the distribution and central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

	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	US Made
0	18.0	8	307.0	130	3504	12.0	70	1
1	15.0	8	350.0	165	3693	11.5	70	1
2	18.0	8	318.0	150	3436	11.0	70	1
3	16.0	8	304.0	150	3433	12.0	70	1
4	17.0	8	302.0	140	3449	10.5	70	1
...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82	1
394	44.0	4	97.0	52	2130	24.6	82	0
395	32.0	4	135.0	84	2295	11.6	82	1
396	28.0	4	120.0	79	2625	18.6	82	1
397	31.0	4	119.0	82	2720	19.4	82	1

398 rows × 8 columns

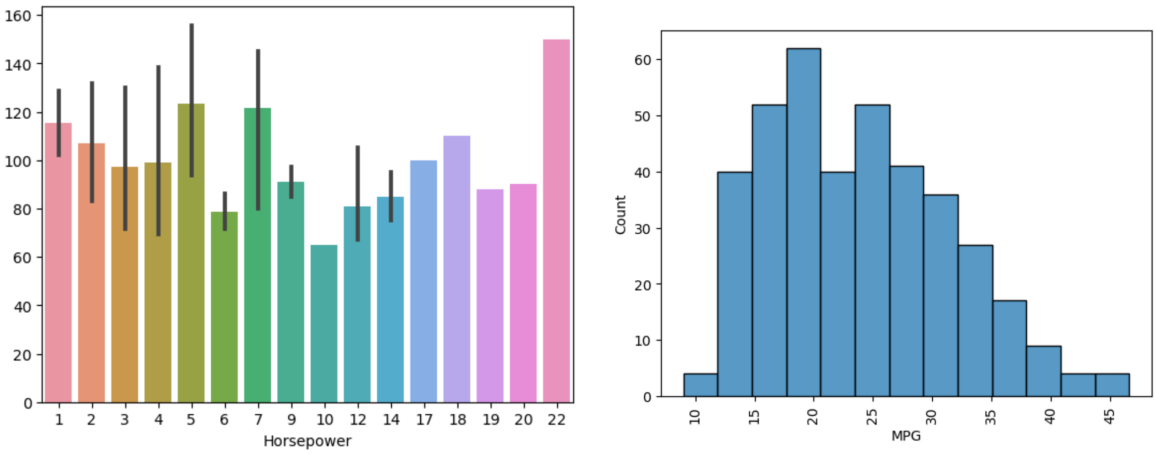
Handling Outliers

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.



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.



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.

```
MPG             1.000000
Model Year      0.579267
Acceleration    0.420289
US Made        -0.568192
Cylinders      -0.775396
Displacement   -0.804203
Weight         -0.831741
Name: MPG, dtype: float64
```

IV. 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.

```
Missing Values:
MPG             0
Cylinders       0
Displacement    0
Horsepower      6
Weight          0
Acceleration    0
Model Year      0
US Made         0
dtype: int64
Mean Squared Error (Imputed): 11.34678760774077
R2 Score (Imputed): 0.8263611993273794
Coefficients (Imputed): [-0.20888761  0.02372668 -0.02429554 -0.00725663  0.03501311  0.77819277
 -2.66219836]
```

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.

```
Mean Squared Error (Pipeline): 12.384830402316544
R2 Score (Pipeline): 0.8104761301670094
```

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.

V. 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.

VI. 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.

VII. 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.