Northeastern University

Module 2 Assignment – Building the Car of the Future

Professor: Justin Grosz

Course: ALY6020: Predictive Analytics (CRN 20466)

Submitted by :

Sameeksha Bellavara Santhosh

Date: 21-01-2024

Introduction

The goal of this project is to identify the characteristics that lead to higher gas mileage (MPG) to help a struggling automaker design an energy-efficient vehicle. Using a variety of car attributes and MPG readings, we wish to create a linear regression model that will allow us to accurately estimate a vehicle's MPG based on its attributes. A few key procedures, like feature selection, model optimization, and data cleansing, are needed for this project.

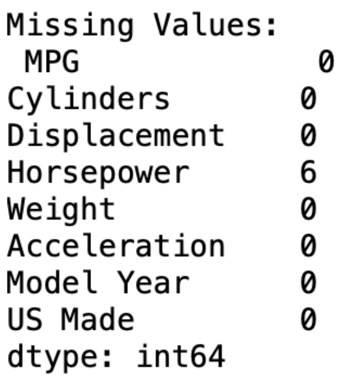
Part 1: Data Cleaning

Several steps were taken in the data pre-processing stages to improve the dataset's integrity and quality. To commence, the datatype of the 'Horsepower' column had to be converted to an integer. To facilitate numerical computations and analyses, it was decided that this conversion was required to guarantee that the values in the column are represented as whole numbers.

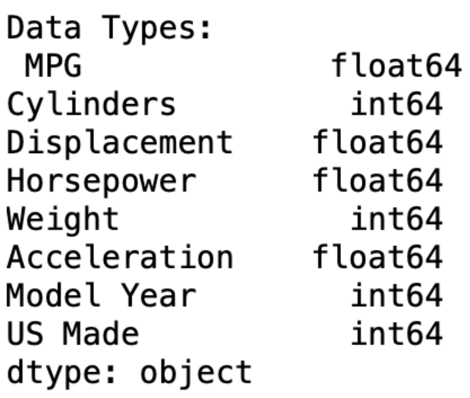
The 'Horsepower' column had six missing entries, as discovered when a check for missing values was done across all columns. To remedy this, the mean of the corresponding column was used to impute the missing values. Imputation prevents important information from being lost by allowing data points to be retained. Data integrity was preserved by identifying and eliminating duplicate rows. If duplicate entries exist, they may cause biases in machine learning models and statistical analysis. Accurate insights require a clean dataset with distinct observations.

After completing these procedures, a data type confirmation was carried out to confirm that the data type modifications and imputations were appropriate. Consistency and compatibility with later analyses and model building are guaranteed by this verification.

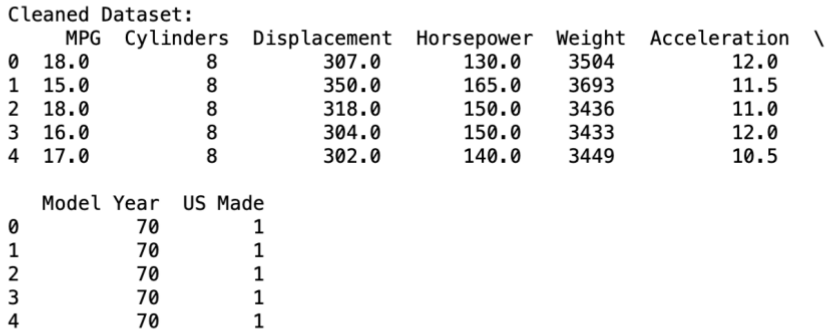
Furthermore, an approach for detecting and managing outliers was employed for the 'Horsepower' column. Rows that contained outliers were eliminated after they were identified using the interquartile range (IQR) method. To avoid skewed results and model inaccuracies, handling outliers is crucial. To guarantee consistency in representation, the categorical variable "US Made" was lastly examined and changed to the integer type. A more complete and trustworthy dataset for future analysis was also produced by replacing the missing values indicated by '?' in the 'Horsepower' column with NaN and filling them using mean imputation.



```
Missing Values:
 MPG             0
Cylinders       0
Displacement    0
Horsepower      6
Weight          0
Acceleration    0
Model Year      0
US Made         0
dtype: int64
```

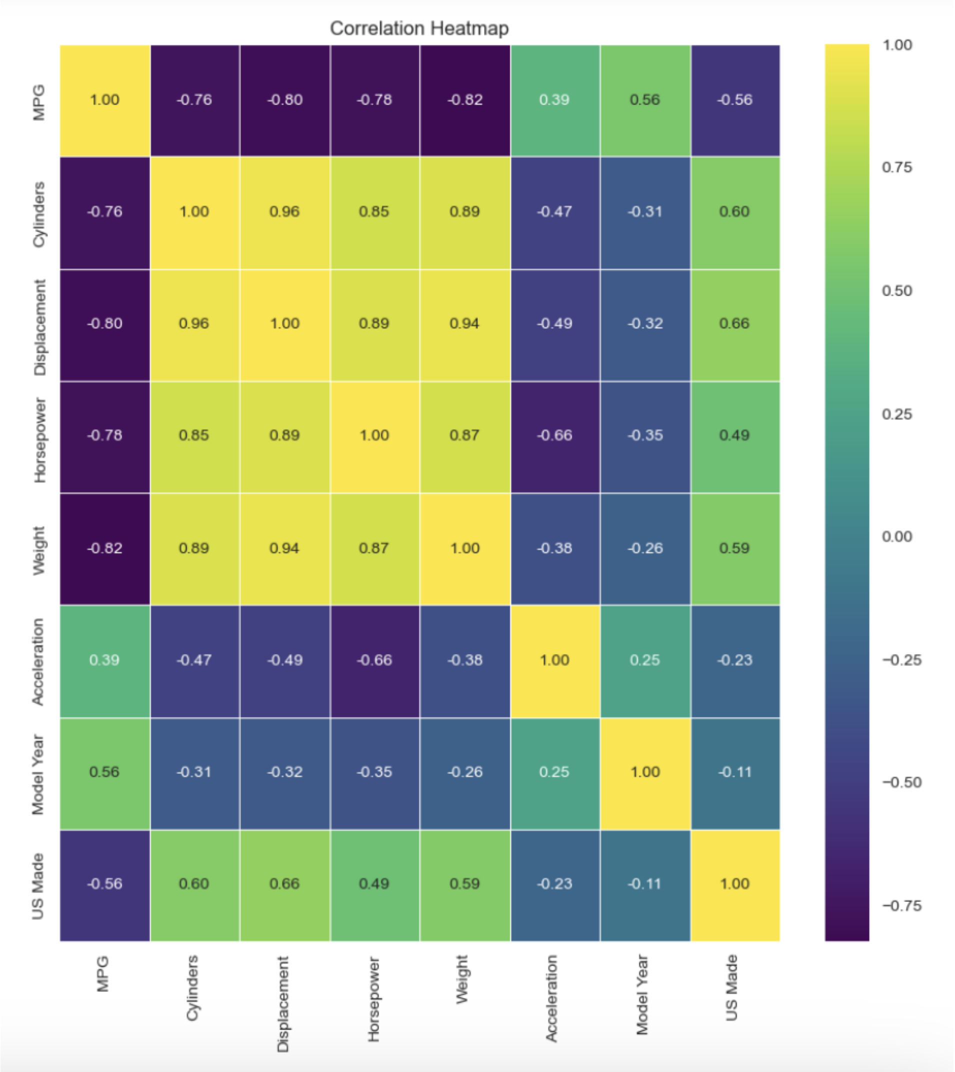


```
Data Types:
 MPG             float64
Cylinders         int64
Displacement    float64
Horsepower      float64
Weight            int64
Acceleration    float64
Model Year        int64
US Made           int64
dtype: object
```



```
Cleaned Dataset:
    MPG  Cylinders  Displacement  Horsepower  Weight  Acceleration  \
0  18.0          8         307.0       130.0    3504          12.0
1  15.0          8         350.0       165.0    3693          11.5
2  18.0          8         318.0       150.0    3436          11.0
3  16.0          8         304.0       150.0    3433          12.0
4  17.0          8         302.0       140.0    3449          10.5

   Model Year  US Made
0          70        1
1          70        1
2          70        1
3          70        1
4          70        1
```



	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	US Made
MPG	1.00	-0.76	-0.80	-0.78	-0.82	0.39	0.56	-0.56
Cylinders	-0.76	1.00	0.96	0.85	0.89	-0.47	-0.31	0.60
Displacement	-0.80	0.96	1.00	0.89	0.94	-0.49	-0.32	0.66
Horsepower	-0.78	0.85	0.89	1.00	0.87	-0.66	-0.35	0.49
Weight	-0.82	0.89	0.94	0.87	1.00	-0.38	-0.26	0.59
Acceleration	0.39	-0.47	-0.49	-0.66	-0.38	1.00	0.25	-0.23
Model Year	0.56	-0.31	-0.32	-0.35	-0.26	0.25	1.00	-0.11
US Made	-0.56	0.60	0.66	0.49	0.59	-0.23	-0.11	1.00

The resulting heatmap provides a thorough summary of the correlation patterns found in the cleaned dataset ('df'). The purpose of the heatmap is to give the correlation matrix a clear visual representation so that it can quickly assess how numerical features relate to one another.

Colour intensity becomes an important visual cue when interpreting the heatmap. Warmer hues, like yellow, denote positive correlations, which means that rising values of one variable typically follow rising values of the other. Cooler hues, such as purple, on the other hand, indicate negative correlations, which imply that a rise in one variable is connected to a fall in another. Each cell's numerical annotations, which provide precise correlation coefficients rounded to two decimal places, enhance the visual representation's accuracy. This data facilitates a more quantitative comprehension of the connections between feature pairs. A useful tool for determining possible multicollinearity between variables, helping with feature selection for modelling, and learning about the underlying patterns in the dataset is the heatmap. All things considered, it provides an informative and user-friendly representation of the complex interactions among the numerical characteristics in the examined dataset.

Part 2: Model Development

The chosen variables were used to fit the multiple linear regression model, and summary statistics, R-squared, and AIC were used to evaluate the model's performance.

Model Summary: Based on the R-squared value of 0.820, the model can account for roughly 82.0% of the variability observed in the target variable (MPG). This indicates a reasonably good fit, indicating that a significant amount of the observed variation in the dependent variable can be explained by combining the selected independent variables.

Adjusted R-squared: Considering the number of predictors, the adjusted R-squared is 0.816. Considering the number of variables in the model, it is a more dependable measure of the model's goodness of fit than R-squared, despite being somewhat lower.

AIC (Akaike Information Criterion): The model's quality is determined by dividing its goodness of fit and complexity into equal parts, with an AIC value of 1616.0. In this instance, lower AIC values indicate that the model is reasonably successful at explaining the variation in the data when using the selected set of variables. Lower AIC values are preferred.

Coefficients: The predicted change in the target variable (MPG) for a one-unit change in the corresponding predictor, holding other variables constant, is indicated by the coefficients associated with each independent variable.

The importance of coefficients: Each coefficient's p-value tests the null hypothesis, which states that the corresponding coefficient equals zero. In case the p-value is lower than the selected significance level, which is usually 0.05, the null hypothesis is declared invalid.

'Weight,' 'Model Year,' and 'US Made' in this model have p-values less than 0.05, indicating that these variables' changes are statistically significant in predicted 'MPG.'

As the R-squared and adjusted R-squared values show, the model has a respectable amount of explanatory power. The model's efficacy is supported by the AIC, although potential multicollinearity requires care. It might be necessary to perform additional diagnostic tests and sensitivity analyses to make sure the model is reliable and understandable.

Part 3: Analysis of Variable Significance

The significance of each variable in the multiple linear regression model offers important insights into how various features affect fuel efficiency when it comes to predicting miles per gallon (MPG) for automobiles. Now let's talk about the importance of each variable and how it affects solving the business problem:

Weight:

Significance: The 'Weight' coefficient exhibits a statistically significant p-value of less than 0.05, indicating that heavier cars typically have lower MPG.

Impact on Business: The automotive industry is aware that heavier cars typically use more fuel, which is supported by this finding. Investigating methods of vehicle weight reduction, perhaps through material innovation or design modifications, would be one way to address the business problem of increasing fuel efficiency.

Model Year:

Significance: The positive coefficient for "Model Year" indicates a highly significant result (p-value < 0.001). Greater MPG is correlated with newer model years.

Impact on Business: This finding suggests that advances in technology and gradual gains in engine efficiency have a positive impact on fuel efficiency. The business objective of increasing fuel efficiency is aligned with promoting the use of newer vehicles or investing in technologies that improve engine efficiency.

US Made:

Significance: The 'US Made' coefficient has a negative coefficient and is statistically significant (p-value < 0.05), indicating that cars manufactured in the United States generally have lower MPG.

Impact on Business: Despite the negative coefficient's seeming paradox, it might be impacted by elements like the size and kind of vehicles that are popular in the US market. To solve the business problem, it might be necessary to investigate the unique qualities of cars built in the United States and take into account technological or design modifications that would increase fuel economy.

Displacement:

Significance: The 'Displacement' coefficient shows a positive correlation with a statistically significant p-value less than 0.05, suggesting that higher MPG is correlated with engine displacement.

Business Impact: It's possible that in some situations, larger engine displacements yield better results. Knowing how engine displacement and fuel efficiency relate to one another and possibly looking into engine technologies that maximize performance is necessary to address the business issue.

Horsepower, acceleration, and cylinders:

Significance: The 'Cylinders,' 'Horsepower,' and 'Acceleration' coefficients do not exhibit statistical significance (p-value being greater than 0.05).

Business Impact: These variables' effects on MPG are not statistically significant within the framework of the model. To learn more about alternative variables and their impact on fuel efficiency, it might be wise to investigate them further or carry out additional research.

To sum up, the model suggests that when tackling the business problem of increasing fuel efficiency, variables like vehicle weight, model year, and country of manufacture should be taken into account. Prioritizing technological developments, encouraging the use of more recent models, and looking into ways to lighten vehicles could all be considered calculated moves. Unexpected correlations, like the positive coefficient for "Displacement," call for more research to improve our comprehension of how particular variables affect fuel efficiency in a particular business setting.

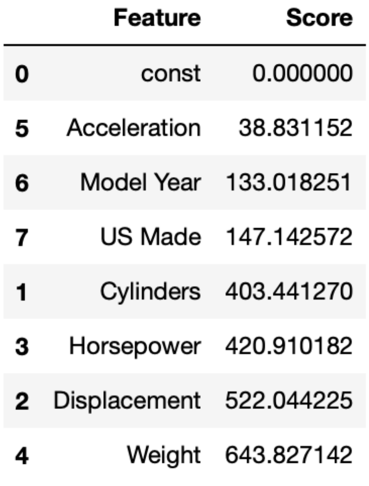
Part 4: Comparison with the Stepwise Selection Model

The selection of feature 2 through the forward selection approach, characterized by the lowest p-value, suggests an enhancement in the model's predictive capacity with the addition of this feature. In contrast, the backward selection approach excluded two features due to their high p-values, leading to an empty response. The stepwise selection approach introduced six attributes, namely "Weight," "Model Year," "US Made," "Acceleration," and "Displacement," indicating their significance in predicting MPG.

Comparing these results to the model from Part 2, notable changes in the selected features are observed. Despite the inclusion of 'Cylinders' and 'Horsepower' in the Part 2 model, feature selection algorithms deemed them insignificant. This implies that these characteristics might not significantly impact MPG prediction.

The enhanced models from Part 3, shaped by feature selection, potentially offer a more precise and efficient means of predicting MPG. By focusing on the most influential attributes, the models allow for simplified yet improved performance. This has strategic implications for the auto manufacturer, facilitating better resource allocation and concentrated efforts in designing and building energy-efficient cars, aligning with their commercial objectives.

Furthermore, the optimized models align with the business issue highlighted in Part 1. Identifying critical features that enhance MPG allows manufacturers to prioritize these elements in vehicle design and production. This strategic focus enables the creation of more fuel-efficient cars that meet consumer preferences and regulatory standards. Ultimately, by delivering vehicles aligned with energy efficiency goals, the optimized models empower the manufacturer to make informed decisions, allocate resources judiciously, and strengthen their competitive position in the market.



	Feature	Score
0	const	0.000000
5	Acceleration	38.831152
6	Model Year	133.018251
7	US Made	147.142572
1	Cylinders	403.441270
3	Horsepower	420.910182
2	Displacement	522.044225
4	Weight	643.827142

Weight and Displacement emerge as the two variables with the highest coefficients, underscoring their substantial impact on predicting MPG. These factors are deemed pivotal in influencing a car's fuel efficiency. This observation aligns seamlessly with previous findings. In Part 2, the linear regression model identified Weight and Displacement as noteworthy variables. Furthermore, in Part 3, Weight emerged as a selected feature in both forward and stepwise selection procedures. The significance of Weight lies in its representation of the overall mass of the vehicle. Larger vehicles demand more energy to move, resulting in lower fuel efficiency. Conversely, Displacement pertains to the total volume swept by all the pistons within an engine's cylinders, often correlated with the engine's size and power. Larger displacement engines may exhibit lower MPG and higher fuel consumption.

Part 5: Business Focus and Recommendations

The two most important factors in greatly increasing MPG are weight and displacement.

Weight: The weight of an automobile has a big impact on how fuel-efficient it is. Since lighter cars need less energy to move, they typically have higher MPG ratings. Focusing on weight reduction, using lightweight materials, and putting effective design strategies into practice can help manufacturers achieve their goal of producing energy-efficient cars.

Displacement: MPG and fuel efficiency are typically higher in engines with smaller displacements. By decreasing displacement while keeping desired performance levels, engine builders can increase fuel efficiency. This can be accomplished by utilizing technologies like hybrid powertrains and turbocharging while also optimizing engine architecture.

References

1. Gupta, A. (2023, December 21). *Feature selection techniques in machine learning (Updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
2. What is a linear regression model? - MATLAB & Simulink. (n.d.). https://www.mathworks.com/help/stats/what-is-linear-regression.html
3. Hayes, A. (2022, January 10). *Stepwise Regression: Definition, uses, example, and limitations*. Investopedia. https://www.investopedia.com/terms/s/stepwise-regression.asp#:~:text=Stepwise%20regression%20is%20the%20step,statistical%20significance%20after%20each%20iteration.