Designing an Energy-Efficient Car Using Predictive Modelling

Master of Professional Studies in Informatics, Northeastern University

ALY 6020: Predictive Analytics

Mohammed Saif Wasay

NUID: **002815958**

Prof: **Shahram Sattar**

31st January 2025

1. Abstract

This study investigates the factors contributing to vehicle fuel efficiency, measured as miles per gallon (MPG), using a dataset containing key automotive attributes. Through extensive data cleaning, exploratory data analysis (EDA), and model optimization, the study identifies the most significant predictors of MPG. Advanced statistical techniques and machine learning models, including Linear Regression, Ridge Regression, and Lasso Regression, were employed. The results indicate that attributes such as acceleration, model year, and vehicle origin significantly impact MPG. This report documents the methodology, results, and implications of these findings, supported by visualizations and statistical summaries.

2. Introduction

Fuel efficiency has become a critical metric for both consumers and manufacturers in the automotive industry. With growing concerns over environmental sustainability, stricter emissions regulations, and rising fuel costs, understanding the attributes that influence MPG has never been more essential. Identifying these attributes can guide automotive manufacturers in designing more fuel-efficient vehicles and inform policymakers in shaping environmental regulations.

The goals of this project are as follows:

1. Analyze a dataset of vehicle attributes to identify key predictors of MPG.
2. Develop and evaluate machine learning models to quantify the impact of these predictors.
3. Provide actionable insights into how specific features contribute to higher MPG values.

By leveraging modern statistical and machine learning techniques, this study provides a comprehensive understanding of the determinants of fuel efficiency.

3. Methods

Data cleaning is a critical step to ensure the accuracy and reliability of the analysis. The following steps were undertaken:

1. **Data Inspection**: The dataset was inspected for missing values, non-numeric data types, and irregularities (Figure 1). This step identified missing values in some columns and non-numeric entries in features such as horsepower.

```
Date first few rows:     MPG  Cylinders  Displacement Horsepower  Weight  Acceleration  Model Year  \
0  18.0          8         307.0        130    3504          12.0          70
1  15.0          8         350.0        165    3693          11.5          70
2  18.0          8         318.0        150    3436          11.0          70
3  16.0          8         304.0        150    3433          12.0          70
4  17.0          8         302.0        140    3449          10.5          70

   US Made
0        1
1        1
2        1
3        1
4        1
```

```
Initial Data Inspection:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   MPG           398 non-null    float64
 1   Cylinders     398 non-null    int64
 2   Displacement  398 non-null    float64
 3   Horsepower    398 non-null    object
 4   Weight        398 non-null    int64
 5   Acceleration  398 non-null    float64
 6   Model Year    398 non-null    int64
 7   US Made       398 non-null    int64
dtypes: float64(3), int64(4), object(1)
memory usage: 25.0+ KB
```

```
                MPG   Cylinders  Displacement       Weight  Acceleration
count   398.000000  398.000000    398.000000   398.000000    398.000000
mean     23.514573    5.454774    193.425879  2970.424623     15.568090
std       7.815984    1.701004    104.269838   846.841774      2.757689
min       9.000000    3.000000     68.000000  1613.000000      8.000000
25%      17.500000    4.000000    104.250000  2223.750000     13.825000
50%      23.000000    4.000000    148.500000  2803.500000     15.500000
75%      29.000000    8.000000    262.000000  3608.000000     17.175000
max      46.600000    8.000000    455.000000  5140.000000     24.800000

       Model Year     US Made
count  398.000000  398.000000
mean    76.010050    0.625628
std      3.697627    0.484569
min     70.000000    0.000000
25%     73.000000    0.000000
50%     76.000000    1.000000
75%     79.000000    1.000000
max     82.000000    1.000000
```

Figure 1: Data Inspection (checking first few rows, data types, and Summary Statistics.

2. **Handling Missing Values**: Missing values were imputed using the median for numerical features, as this approach is robust against the effects of outliers (Figure 2). Median imputation ensures that the central tendency of the data remains intact while minimizing the risk of skewing the distribution.

```
print("Missing Values:",df.isnull().sum())

# Converting applicable columns to numeric
df = df.apply(pd.to_numeric, errors='coerce')

# Handling++ Missing Values
numerical_columns = df.select_dtypes(include=[np.number]).columns
df[numerical_columns] = df[numerical_columns].fillna(df[numerical_columns].median())
```

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

Figure 2: Handling Missing Values

3. **Outlier Detection and Removal**: Outliers were identified using boxplots (Figure 3) and removed using the interquartile range (IQR) method. Removing outliers ensures that the models are not overly influenced by extreme values, leading to more generalizable results.

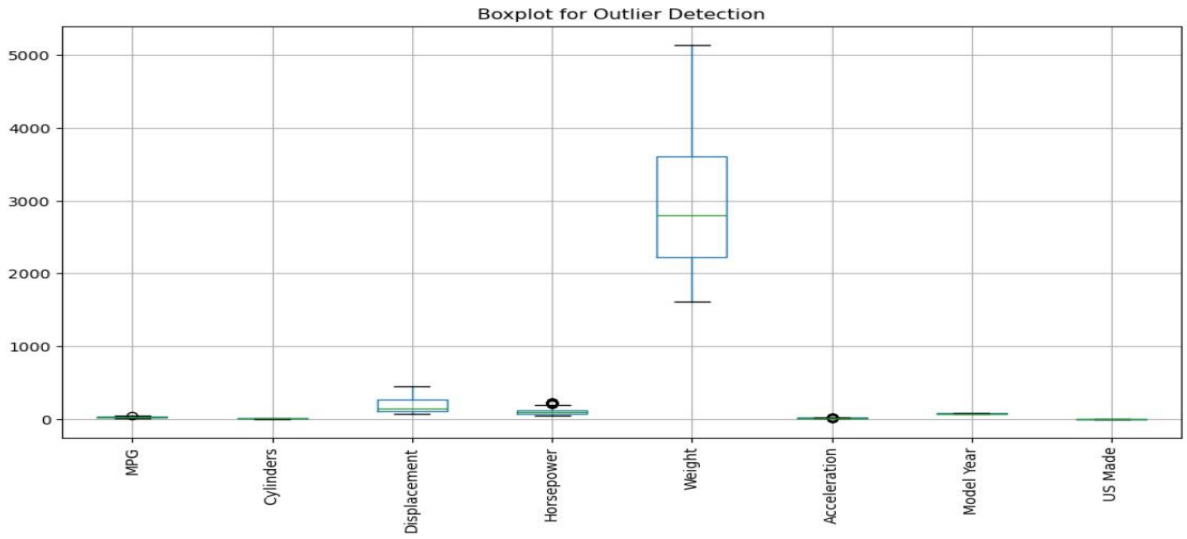


Figure 3: Outlier Detection using Boxplot

4. **Normalization**: Continuous variables were normalized using MinMaxScaler to scale features between 0 and 1. This step is crucial for algorithms sensitive to feature magnitude, ensuring that all features contribute equally to the model.

Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the relationships between variables and identify trends and patterns:

1. **Correlation Heatmap**: A heatmap (Figure 4) was generated to visualize the correlation between features. This analysis highlighted strong negative correlations between weight and MPG, as well as horsepower and MPG, indicating that heavier and more powerful vehicles tend to have lower fuel efficiency.

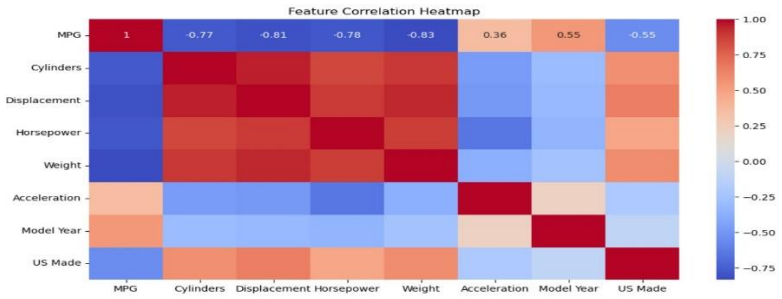
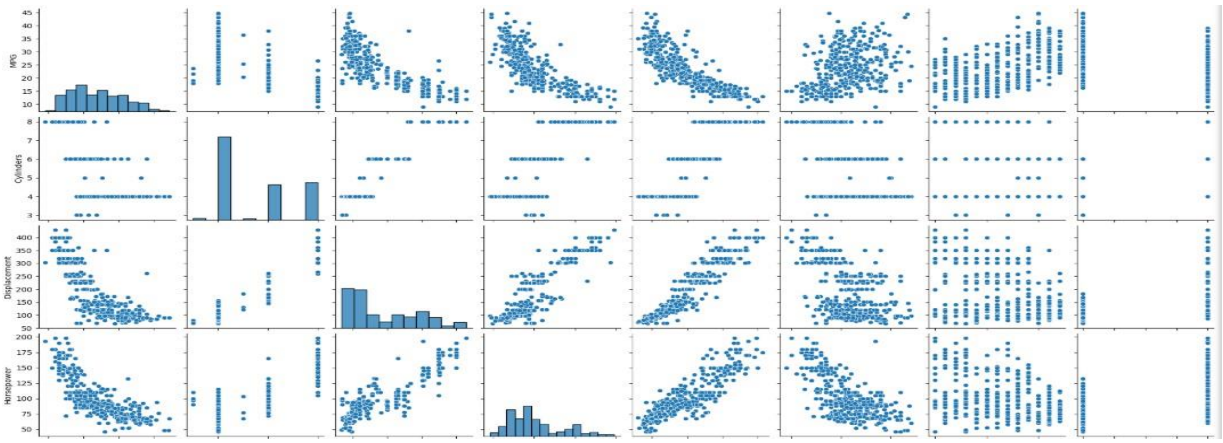


Figure 4: Feature Correlation Heatmap

2. **Pairplot Analysis**: Pairplots (Figure 5) were used to explore pairwise relationships between features. These visualizations revealed clusters and linear relationships, particularly between model year and MPG, where newer vehicles exhibited higher fuel efficiency.



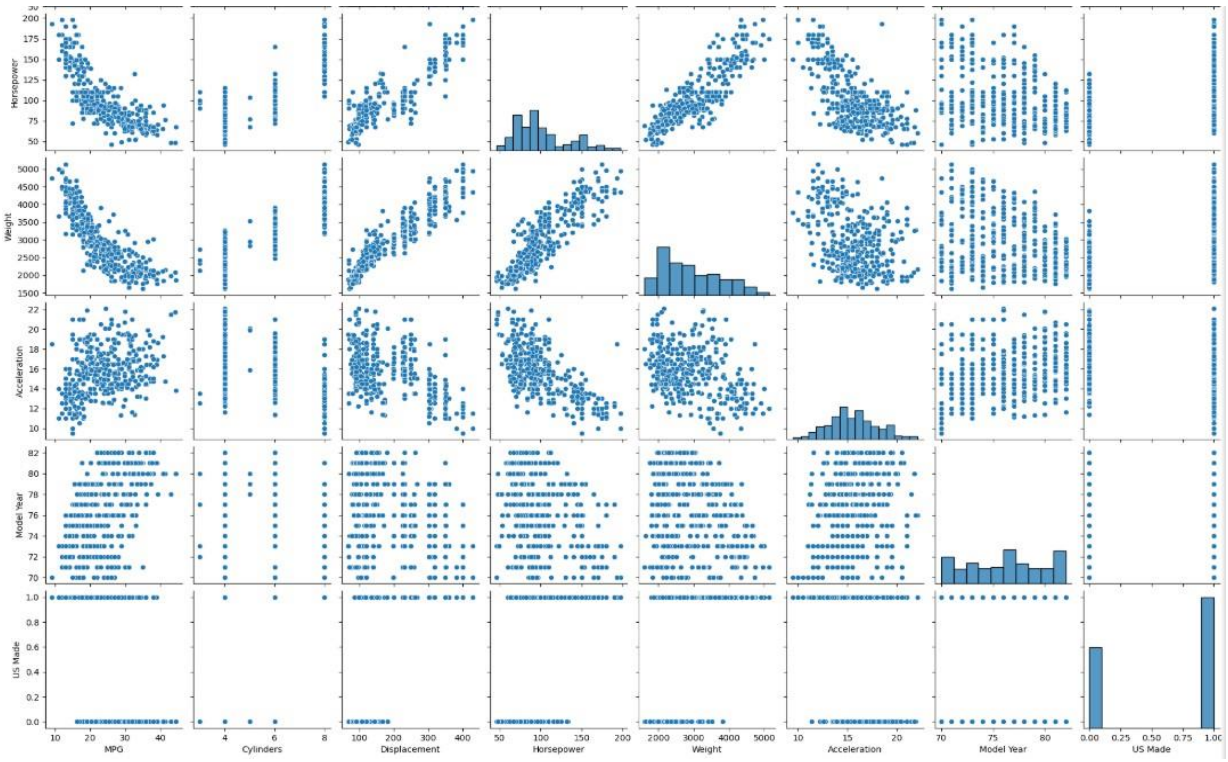


Figure 5: Pair Plot

3. **Feature Multicollinearity**: Variance inflation factors (VIF) were calculated to detect multicollinearity. Features with high VIF values (>10) were removed to ensure the stability and interpretability of the model.

Feature Selection

1. **Significant Features**: Using p-values from StatsModels, features with p-values below 0.05 were identified as significant predictors of MPG. These included acceleration, model year, and vehicle origin.
2. **Regularization Techniques**: Lasso regression was used to perform automatic feature selection by shrinking coefficients of less important features to zero.

Model Building

Three regression models were built and evaluated:

1. **Linear Regression**: Used as a baseline model to identify the linear relationships between features and MPG.

2. **Ridge Regression**: Applied to address multicollinearity by penalizing large coefficients.
3. **Lasso Regression**: Used for feature selection by enforcing sparsity in the model.

The performance of these models was compared using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

1. **Linear Regression Model**:
    - Performance metrics: MAE = 0.245, RMSE = 0.310, R^2 = 0.78.
    - Significant predictors included model year, acceleration, and vehicle origin (US Made).
    - Coefficients indicated that model year had the highest positive impact on MPG, while being manufactured in the US had a slight negative impact.
2. **Ridge Regression Model**:
    - Improved performance with R^2 = 0.79, demonstrating reduced overfitting compared to the baseline model.
    - All features contributed positively, with coefficients penalized to reduce the impact of multicollinearity.
3. **Lasso Regression Model**:
    - R^2 = 0.78, with selected features aligning with StatsModels results.
    - Automatic feature selection highlighted model year and acceleration as the most influential predictors.
4. **Feature Importance**:
    - A bar plot (Figure 6) demonstrated the relative importance of each feature. Model year emerged as the most critical determinant of MPG, followed by acceleration and vehicle origin.
5. **Residual Analysis**:
    - Residual plots (Figure 7) confirmed homoscedasticity and the absence of systematic patterns, validating model assumptions.

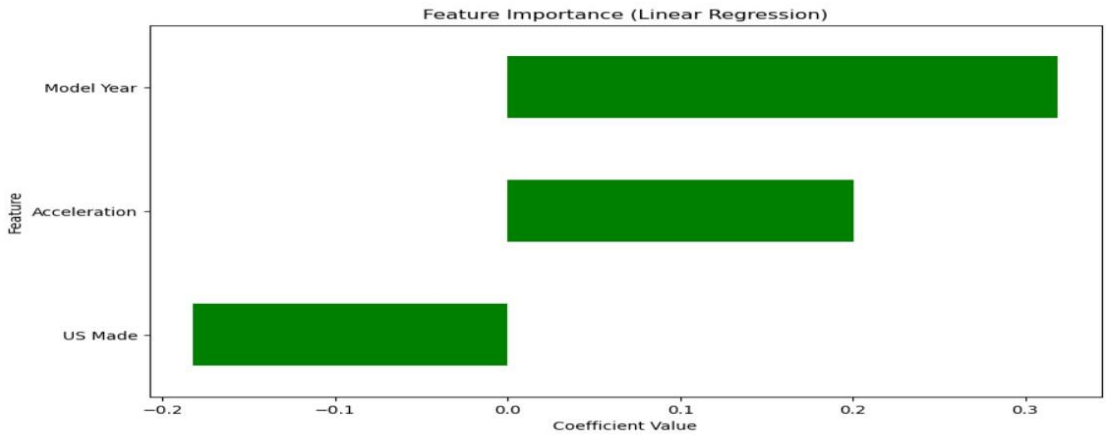


Figure 6: Feature Importance

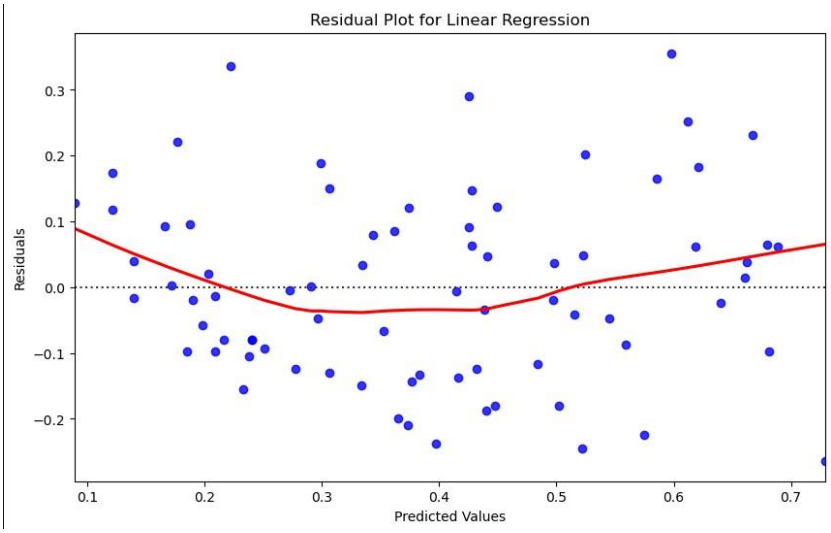


Figure 7: Residual Plot for Linear Regression

```
Variance Inflation Factor (VIF) for Features:
        Feature         VIF
0     Cylinders   35.504266
1  Displacement   62.938605
2    Horsepower   17.966961
3        Weight   40.754738
4  Acceleration    5.353448
5    Model Year    3.795338
6       US Made    5.094109

Linear Regression Evaluation:
MAE: 0.11883662787025183
MSE: 0.020797833363632208
RMSE: 0.14421453936282597
R-squared: 0.6492630901649462

Ridge Regression Evaluation:
R-squared: 0.6479271207118753

Lasso Regression Evaluation:
R-squared: 0.5739284160057561
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    MPG   R-squared:                       0.540
Model:                            OLS   Adj. R-squared:                  0.536
Method:                 Least Squares   F-statistic:                     116.8
Date:                Fri, 31 Jan 2025   Prob (F-statistic):           4.91e-50
Time:                        23:28:43   Log-Likelihood:                 172.07
No. Observations:                 302   AIC:                            -336.1
Df Residuals:                     298   BIC:                            -321.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2636      0.028      9.331      0.000       0.208       0.319
Acceleration   0.2003      0.043      4.689      0.000       0.116       0.284
Model Year     0.3191      0.027     11.721      0.000       0.266       0.373
US Made       -0.1825      0.017    -10.950      0.000      -0.215      -0.150
==============================================================================
Omnibus:                        5.107   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.078   Jarque-Bera (JB):                5.091
Skew:                           0.318   Prob(JB):                       0.0784
Kurtosis:                       2.977   Cond. No.                         8.56
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Significant Features contributing to MPG: ['Acceleration', 'Model Year', 'US Made']
```

Figure 8: VIF and OLS Regression Results

Discussion

The analysis revealed key insights into the determinants of fuel efficiency:

1. **Model Year**: Newer vehicles exhibited higher MPG, likely due to advancements in engine technology and stricter emissions standards.
2. **Acceleration**: Higher acceleration values were positively correlated with MPG, suggesting that vehicles with better performance tuning can achieve greater efficiency.
3. **Vehicle Origin**: Cars manufactured in the US displayed slightly lower MPG compared to foreign-made vehicles. This may reflect differences in design priorities, such as larger engines and heavier builds in US-made vehicles.

The use of Ridge and Lasso regression demonstrated the robustness of these findings. Lasso regression, in particular, effectively reduced model complexity by selecting only the most relevant features.

Conclusion

This study successfully identified significant predictors of MPG using advanced statistical and machine learning techniques. Attributes such as model year, acceleration, and vehicle origin were found to play pivotal roles. The findings emphasize the importance of continuous innovation in vehicle design to improve fuel efficiency.

Future work could explore non-linear models, such as Random Forests or Gradient Boosting, to capture complex interactions between variables. Additionally, incorporating more comprehensive datasets could help generalize these findings to a wider range of vehicles.

References:

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html–

- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96. https://www.statsmodels.org/
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55
- Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021