Technical Report: Predictive Modeling of Vehicle Fuel Efficiency

Course: Machine Learning

Dataset: OpenML Cars1 Dataset (Auto MPG)

Developer: Susanna Mkrtchyan

1. Dataset & Motivation

For this project, I chose the **Cars1 (Auto MPG)** dataset. It contains technical data about cars like horsepower, weight, and displacement.

- **Target:** Predict **MPG** (fuel efficiency).
- **Motivation:** This is a classic regression problem where physical features (like weight) have a direct, logical impact on the result, making it perfect for testing the OLS model.

2. Data Preprocessing

To prepare the data, I followed these steps:

- **Splitting:** The data was split into **Train (70%)**, **Validation (15%)**, and **Test (15%)** sets.
- **Scaling:** I used StandardScaler to bring all features to the same scale. This prevents larger numbers (like weight) from over-influencing the model.
- **Feature Selection:** I used **Lasso Regression** to automatically pick the most important features and remove the ones that don't help the prediction.

3. OLS Implementation

The heart of this task was calculating the model weights manually using the **Normal Equation**: **w = (X.T * X)^-1 * X.T * y**

- **Accuracy:** I used **np.linalg.pinv** (pseudoinverse) for the calculation. This is more stable than a regular inverse and prevents errors if features are too similar.
- **Intercept:** I added a column of **1s** to the data so the model could calculate the "baseline" MPG (the Intercept).

4. Results & Comparison

I compared my manual OLS results with the **sklearn** library.

- **Comparison:** Both methods produced identical results (difference < 1e-15).
- **Performance:** The model achieved an R2 score of **0.79**, meaning it explains a large portion of the fuel efficiency variance.

5. Residual Analysis

I checked the errors (residuals) to see if the model is reliable:

- **Scatter Plot:** The errors are randomly spread around zero, which means the linear model was a good choice.
- **Histogram:** The errors follow a normal "bell curve," which is a key requirement for a good OLS model.

6. Interpretation of Features

- **Weight:** This is the most important factor. As weight increases, MPG drops significantly (negative correlation).
- **Horsepower:** Also has a negative impact; more power usually means more fuel used.

7. Limitations & Insights

- **Non-linearity:** Some features might have a curved relationship with MPG, which a simple straight line cannot perfectly capture.
- **Insight:** The manual OLS implementation works perfectly and matches industry standards. Lasso was very helpful in simplifying the model by choosing only the best features.