HW Assignment 3 (A3) Part 1: Regression Models
CS6140: Machine Learning Fall 2024
Due Date: Thursday, _____
(__ points)

Introduction

Regression is a fundamental technique in machine learning used for predicting continuous outcomes based on input features. In ML, regression models learn the relationship between a dependent variable (target) and one or more independent variables (features), enabling predictions about future data points. Multiple regression algorithms, such as linear regression, decision trees, and ensemble methods like random forests and boosting, allow for flexibility in handling complex datasets with varying degrees of non-linearity and noise. In applications such as predicting fuel efficiency, CO2 emissions, or stock prices, regression models provide valuable insights and accurate predictions by capturing trends and patterns within the data.

The Environmental Protection Agency (EPA) plays a critical role in monitoring and regulating air pollutants emitted by vehicles on American roads. Through its testing programs, the EPA assesses various pollutants such as carbon dioxide (CO2), nitrogen oxides (NOx), carbon monoxide (CO), and particulate matter (PM). The agency enforces standards to ensure that vehicle manufacturers meet stringent emission guidelines that help reduce the environmental impact of transportation. The EPA also measures an approximate mileage per gallon (mpg) for vehicles. These regulations are part of broader efforts to combat climate change, improve air quality, and protect public health by minimizing the harmful effects of pollutants generated by millions of vehicles in operation. The EPA's data is invaluable for research, policymaking, and machine learning applications aimed at modeling and predicting vehicle emissions and fuel efficiency trends.

You are given last ten years of EPA data. The first task (Part 1) of this assignment is to perform EDA and prepare the data for Regression Analysis.

Part 1: Exploratory Data Analysis (EDA)

1. Initial Exploration and Summary Statistics:

Load the datasets for all the years and compare the basic summary statistics. What are the key summary statistics (mean, median, standard deviation, etc.) for continuous variables such as "Test Veh Displacement (L)" and "Set Coef A (lbf)" in each dataset? Identify any outliers or anomalies.

2. Data Structure and Missing Values:

Explore the structure of each dataset. Are there any missing values? If so, which columns have the most missing data, and how do you plan to handle them?

3. Comparing Column Names Across Years:

Compare the column names across the datasets. Are there any inconsistencies in column naming between the datasets (for example, different names for similar variables)? How would you standardize these column names to prepare the datasets for stacking?

4. Distribution of Key Variables:

Plot the distribution of key continuous variables (such as "Test Veh Displacement (L)" over the years. How do the distributions of these variables change over time? Are there any visible trends?

5. Correlation Analysis:

Perform a correlation analysis on the numeric variables in each dataset. Are there any strong correlations between certain variables, particularly those related to vehicle performance or emissions? What insights can be drawn from these correlations?

6. Combining Datasets:

Standardize the column names and stack the datasets for all the years (2015 onward) into a single dataset. After stacking, check for any discrepancies or issues that arise (e.g., data types, missing values).

7. Standardizing Continuous Variables:

Standardize the continuous variables (such as vehicle displacement and coefficients) across the years. How does standardization affect the interpretation of the data? Why is it important to standardize these variables before performing regression analysis?

8. Temporal Analysis:

Investigate how certain variables evolve over time. Plot these variables against the "Model Year" and describe any noticeable patterns or shifts in vehicle design or emissions standards.

9. Vehicle Manufacturer Analysis:

Identify the top 10 vehicle manufacturers by the number of test records in the dataset. How does the performance (in terms of coefficients or displacement) vary across these manufacturers? What insights can be drawn from these differences?

10. Creating a Clean Dataset for Regression:

After performing the EDA, create a clean dataset that can be used for regression analysis. What transformations did you perform (e.g., removing outliers, imputing missing values, standardizing variables)? Provide a brief summary of your cleaning process.

Expected Output

Please submit a fully executed Jupyter notebook clearly identifying question number and steps. Make sure to add proper commentary to your solution.