HW Assignment 3 (A3) Part 2: Regression Models
CS6140: Machine Learning Fall 2024
Due Date: Thursday, _____
(__ points)

Part 2: Regression

Regression is a fundamental technique in machine learning used for predicting continuous outcomes based on input features. In machine learning, 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 comprehensive testing programs, the EPA assesses various pollutants such as carbon dioxide (CO2), nitrogen oxides (NOx), carbon monoxide (CO), and particulate matter. The agency enforces standards to ensure that vehicle manufacturers meet stringent

emission guidelines that help reduce the environmental impact of transportation. 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, policy-making, and machine learning applications aimed at modeling and predicting vehicle emissions and fuel efficiency trends.

Task:

In this assignment, you will explore multiple machine learning algorithms to predict two different outcomes: mileage per gallon (RND_ADJ_FE) and CO emissions. You will run six different algorithms, analyze their performance, and identify the most important features from each model.

Step 1: Data Preprocessing and Exploration

Handle missing values, standardize variable names across the years, perform outlier analysis, and multiple imputations.

Standardize the features to ensure they are on the same scale.

Create training and test datasets for both target variables.

Step2: You will build models using the following six algorithms:

1. Multiple Linear Regression (MLR)
2. K-Nearest Neighbors (KNN)
3. Random Forest
4. Gradient Boosted Trees
5. XGBoost
6. CatBoost

For each algorithm, train two models:

One to predict RND_ADJ_FE (mileage per gallon) and the other to predict CO emissions.

Step3: Model Evaluation

For each of the 12 models (6 algorithms x 2 target variables), evaluate the performance using the following metrics:

1. R-squared: To measure the proportion of variance explained by the model.

2. Mean Absolute Error (MAE): To measure the average magnitude of errors.

3. Root Mean Squared Error (RMSE): To account for larger errors and compare on the same scale as the target variable.

Step 4: Feature Importance

For models that provide feature importance scores (such as Random Forest, Gradient Boosted Trees, XGBoost, and CatBoost), identify the top features that most influence the predictions.

Step 5: Results Comparison

After training and evaluating the models, compare the results:

1. Compare performance metrics (R-squared, MAE, RMSE, F-statistic) across all models.

2. Compare feature importance across the algorithms that provide it. Discuss how different algorithms may weigh features differently for each target variable.

3. Summarize findings: Which algorithms performed the best for each target variable? Were the top features consistent across models? How did the algorithms handle the data differently?

This framework ensures that students explore a wide range of algorithms and performance metrics, while also learning to interpret the results and feature importance from multiple machine learning models. Let me know if you need any further adjustments!

Expected Output

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