Name: Motolani Kay-Salami Karima
Course: COMP 3400
Student ID: 202165668

Predictive Modeling Report: Linear Regression on London Weather Data

Introduction

The objective of this project was to develop a baseline linear regression model for predicting precipitation using various meteorological variables from the London weather dataset. Subsequently, different data preparation techniques were applied to enhance the model's predictive performance. The evaluation metric utilized for assessing the model was the R^2 (coefficient of determination) score calculated on the test set.

Baseline Model

The initial dataset was refined by eliminating the 'date' column and omitting any missing values. An 80/20 split was employed for training and testing the model. An intercept (bias term) was introduced, and parameters were calculated using the normal equation. The baseline model yielded an R^2 score of Baseline R^2: 0.1306

Model Improvement Attempts

1. Transformation

Attempt 1: Standardization
Features were standardized to achieve a zero mean and unit variance. However, the R^2 value remained unchanged:
$R^2 = 0.1306$ (No improvement)

Attempt 2: Min-Max Normalization
Features were scaled to the range [0, 1].
This method was skipped because Standardization did not improve the model; project logic proceeded to the next method only if an improvement occurred.

Attempt 3: Max Abs Scaling
Features are normalized by dividing by their maximum absolute values. Previous attempts were not executed since they did not improve the model, following the notebook's logic.

While only the first attempt (Standardization) was run, all three methods were included to meet project requirements, with their execution depending on the results of previous attempts.

2. Feature Selection (Correlation Threshold = 0.95)

Highly correlated features were identified and removed, resulting in a slight performance improvement, $R^2 = 0.1320$

3. Outlier Removal (Z-Score > 3 in at least one feature)

Outliers were eliminated using z-score analysis, leading to the most significant improvement in performance, $R^2 = 0.1324$

Summary of Model Improvements

Step	R^2 Score
Baseline	0.1306
Standardization	0.1306
Feature Selection (0.95)	0.1320
Outlier Removal (z=3, m=1)	0.1324

Best Model R^2: 0.1324

Challenges Faced

1. One significant constraint was the requirement to avoid using scikit-learn and instead rely solely on NumPy and Pandas. This meant manually implementing processes such as standardization, normalization, correlation filtering, outlier detection, and regression—tasks that are generally simplified with machine learning libraries.
2. Despite employing multiple techniques, the improvements in the R^2 value were minimal. This made it challenging to confidently identify which transformations significantly enhanced the model's performance.
3. Manually identifying and removing highly correlated features for various thresholds (0.95, 0.90, 0.85) involved careful analysis of the correlation matrix and iterative testing.
4. Developing an outlier removal routine that considered z-scores across multiple features without using external libraries was both time-consuming and susceptible to edge case bugs, necessitating meticulous debugging.

Conclusion

This project helped me understand the importance of data preparation beyond model selection. Even without complex algorithms, the predictive quality of a model can change based on how clean and well-prepared the input data is. Through trial and error, I learned to critically evaluate the trade-offs of each preprocessing technique and how they impact model generalization.