Technical Report

Team Details

- Data_Crunch_117
- De Zoysa A S I
- University of Moratuwa

1 Problem Understanding & Dataset Analysis

1.1 Problem Statement

The objective of this project is to predict weather parameters, including:

- Average Temperature
- Radiation
- Wind Speed
- Wind Direction
- Rain Amount

These predictions assist in weather forecasting and decision-making for various applications.

1.2 Dataset Overview

The dataset consists of the following features:

- **ID** - Unique identifier for each record.

- **Year** - The year of the recorded data.
- **Month** - The month of the recorded data.
- **Day** - The day of the recorded data.
- **Kingdom** - The geographical region where the data was recorded.
- **+ More:** Location, temperature, precipitation, wind, and radiation-related parameters.

The target variables are **Avg_Temperature, Radiation, Wind_Speed, Wind_Direction, and Rain_Amount**.

1.3 Data Preprocessing

- **Categorical Encoding**: One-hot encoding for the categorical feature 'Kingdom'.
- **Feature Scaling**: StandardScaler is used for numerical features.
- **Handling Missing Values**: Assumed minimal or handled in preprocessing.

2 Feature Engineering & Data Preparation

2.1 Feature Selection

- Used **Month, Day, and Kingdom** as input features.
- Predicted **Avg_Temperature, Radiation, Wind_Speed, and Wind_Direction** first.
- Used these predictions as additional features to predict **Rain_Amount**.

2.2 Transformations Applied

- **Standardization**: Applied to numerical features to improve model performance.
- **Rain_Amount Scaling**: Standardized using StandardScaler before model training and inverse transformed after prediction.

3 Model Selection & Justification

3.1 Choice of Model: XGBoost Regressor

- Chosen for its effectiveness in tabular data and ability to capture nonlinear relationships.
- Handles missing values internally and provides robust predictions.
- Supports parallel processing for efficiency.

3.2 Hyperparameter Optimization

- **n_estimators**: 200
- **objective**: reg:squarederror
- **random_state**: 42

3.3 Validation Strategy

5-Fold Cross-Validation was used to ensure reliable performance estimation and reduce overfitting.

4 Performance Evaluation & Error Analysis

4.1 Evaluation Metric

sMAPE was used for evaluation. Cross-validation was performed using a custom **sMAPE scorer**.

4.2 Model Comparison

- Performance of each model was assessed individually before using predictions as features for **Rain_Amount**.
- Additional tuning could be explored to refine model accuracy.

4.3 Limitations & Improvements

- Predictions for **Rain_Amount** rely on intermediate variables, which may introduce compounding errors.
- Future enhancements could include additional meteorological factors or time-series-based modeling.

5 Interpretability & Business Insights

5.1 Practical Applications

- Assists in agriculture, energy planning, and disaster preparedness by forecasting essential weather conditions.
- Businesses reliant on weather forecasts (e.g., tourism, logistics) can optimize operations using these predictions.

5.2 Deployment Considerations

- Models can be deployed as a REST API for real-time weather predictions.

- Continuous model retraining using updated weather data can improve accuracy.

6 Innovation & Technical Depth

6.1 Novel Approaches

- **Multi-Stage Prediction**: Instead of training a single model, predictions from the first four targets were used to enhance **Rain_Amount** forecasting.
- **Feature Engineering**: Effective use of categorical encoding and numerical transformations to optimize performance.
- **Scalability**: The pipeline structure allows easy integration with real-time weather data sources.

7 Conclusion

This project effectively utilizes **XGBoost** to predict key weather parameters with a structured preprocessing and model training pipeline. The approach balances performance and interpretability while being adaptable for real-world deployment.