CODEJAM

DATA CRUNCH

Time Series Forecasting of Environmental Variables for Harveston's Agriculture

By HackSquad- Data_Crunch_078
General Sir John Kotelwala Defense University

1. Problem Understanding & Dataset Analysis

Forecasting Objective & Expected Outcomes:

The primary objective is to develop time series forecasting models to predict five critical environmental variables for Harveston:

- Average Temperature (°C)
- Radiation (W/m²)
- Rain Amount (mm)
- Wind Speed (km/h)
- Wind Direction (°)

These forecasts will help Harveston farmers make informed decisions about planting cycles, resource allocation, and preparation for weather extremes, thus supporting sustainable agricultural practices and economic stability.

Training Data (train.xlsx)

- Contains multiple columns, including:
    - **Date Information**: Year, Month, Day
    - **Location Information**: kingdom, latitude, longitude
    - **Target Variables**: Avg_Temperature, Radiation, Rain_Amount, Wind_Speed, Wind_Direction
    - **Additional Features**: Avg_Feels_Like_Temperature, Temperature_Range, etc.
- **Potential Issues**:
    - The Avg_Temperature in Atlantis looks pretty high (299.65), suggesting unit inconsistencies.
    - Other kingdoms might have different unit scales as well.

Test Data (test.xlsx)

- Contains only the ID, Year, Month, Day, and kingdom, meaning we need to **predict** the five environmental variables.

Preprocessing Justification:

- **Handling Missing Values:** Imputation strategies (mean, median, forward-fill, or model-based) will be employed depending on the missingness pattern.
- **Unit Standardization:** For consistency, temperature values (°C) will be standardized, particularly converting any Kelvin to Celsius (K - 273.15).
- **Scaling/Normalization:** Will be used for models like neural networks to improve convergence.
- **Smoothing:** Moving averages will be used to reduce noise while preserving essential trends.

- **Outlier Handling:** Z-score or IQR methods will be applied to detect and manage outliers.

2. Feature Engineering & Data Preparation

Feature Creation Techniques:

To capture temporal dependencies and improve model accuracy:

- **Lag Features:** Historical values (e.g., temperature yesterday to predict today's value).
- **Moving Averages/Rolling Statistics:** Features like rolling mean, min, max, and standard deviation over various time windows (e.g., 7-day, 30-day) to capture trends and seasonality.
- **Date/Time Features:** Components like day of the week, month, and cyclic transformations (sin/cos) to help the model capture seasonality.
- **Interaction Features:** Features that represent interactions between variables (e.g., wind speed and temperature).
- **External Variables:** If available, incorporating other external weather variables could enhance model performance.

Feature Selection Justification:

Feature selection techniques will be employed to:

- **Reduce Model Complexity:** Identify the most relevant features, minimizing overfitting.
- **Improve Interpretability:** Ensuring the model remains understandable to non-experts.
- **Enhance Performance:** Methods like correlation analysis, permutation importance, and tree-based model importance will guide the feature selection process.

Data Stationarity Transformations:

Since environmental time series often exhibit trends and seasonality:

- **Stationarity Testing:** Using tests like ADF (Augmented Dickey-Fuller) to assess stationarity.
- **Transformations:** Differencing and log transformations will be applied if needed to achieve stationarity.

3. Model Selection & Justification

Model Evaluation:

The following models will be considered:

- **Baselines:** Simple moving averages, naive forecasting, and seasonal naive.
- **Statistical Models:** ARIMA, SARIMA (for seasonal data).
- **Machine Learning Models:** XGBoost, LightGBM, Random Forest (using lagged features).
- **Deep Learning Models:** LSTMs, GRUs (to capture long-term dependencies).
- **Specialized Libraries:** Prophet (effective for trend and seasonality).

Model Choice Justification:

Model selection will be based on:

- **Dataset Characteristics:** The size of the data, seasonality, and trends.
- **Forecasting Requirements:** Forecasting multiple variables and determining the forecast horizon.
- **Performance Comparison:** Based on validation results using the primary metric (SMAPE).

Hyperparameter Optimization:

Techniques like grid search, random search, or Bayesian optimization will be applied to fine-tune model parameters for optimal performance.

Time Series Validation:

- **Rolling Forecast Origin (Walk-Forward Validation):** Ensures that predictions are made using past data without future lookahead bias.
- **Time-Based Cross-Validation:** Properly splitting data for training and validation to avoid temporal overlap.

4. Performance Evaluation & Error Analysis

Evaluation Metrics:

- **SMAPE:** The primary metric to evaluate model performance.
- **Additional Metrics:** MAE, RMSE, and MAPE will be calculated for a comprehensive understanding.

Model Performance Comparison:

Performance of different models (statistical, machine learning, deep learning) will be compared using the validation set's SMAPE score, and the best-performing model will be selected.

Residual Analysis:

- **Autocorrelation:** Ensuring residuals are uncorrelated (ACF/PACF plots, Ljung-Box test).
- **Normality:** Checking if residuals are normally distributed (Shapiro-Wilk test).
- **Heteroscedasticity:** Assessing if variance of residuals is constant (Breusch-Pagan test).

5. Interpretability & Business Insights

Real-World Application:

The forecasts will enable farmers to optimize their agricultural practices:

- **Optimal Planting/Harvesting:** Based on predicted temperature, rainfall, and wind speed.
- **Irrigation Scheduling & Resource Allocation:** Predicting radiation and temperature.
- **Proactive Measures Against Weather Extremes:** Helping farmers prepare for adverse weather conditions like frost or high winds.

Forecasting Strategy & Deployment Improvements:

- **Probabilistic Forecasting:** Providing uncertainty ranges along with point predictions.
- **Continuous Monitoring & Retraining:** Setting up a feedback loop for model improvement.
- **User-Friendly Deployment:** Developing dashboards or APIs for farmers to access forecasts easily.

6. Innovation & Technical Depth

Novel Approaches:

- **Ensemble Learning:** Combining different model predictions to achieve more robust results.
- **Advanced Feature Engineering:** Leveraging complex interactions like growing degree days.
- **Multi-Location Data Handling:** Using hierarchical models or geographical features effectively.

Unique Techniques:

Innovative methods like optimizing the SMAPE directly, or developing computationally efficient training methods, will be discussed here.

7. Conclusion

Summary of Findings:

The project will summarize the main steps:

- Forecasting task, data preparation, and model evaluation.
- The best-performing model based on SMAPE will be selected.

Challenges & Future Improvements:

Challenges in model optimization, handling multi-source data, and potential improvements (e.g., more granular data, specialized models for each variable) will be addressed.

Team HackSquad

1. S.P.Y.S Sasmika
2. N.D.T.V Nawagamuwa
3. P.V.R Hirushi

Data_Crunch_078