PM Accelerator Forecasting Report

Mbeleck Berle
mbeleckberle@gmail.com

***Abstract*—This report presents a comprehensive analysis of temperature forecasting using various time-series models, including ARIMA, SARIMA, and LSTM neural networks. The study covers data cleaning, exploratory data analysis (EDA), feature importance evaluation, and anomaly detection, all aimed at enhancing the accuracy and reliability of forecasts. Results indicate that the LSTM model outperforms traditional statistical models, yielding improved forecasting performance.**

I. Introduction

The goal of this study is to leverage advanced forecasting techniques to predict temperature trends accurately. By integrating statistical models with deep learning approaches, we aim to provide robust and actionable insights that support real-world decision-making in environmental and industrial contexts.

II. Data Preparation and Exploratory Data Analysis

A. Data Cleaning

The following steps were undertaken to ensure the quality of the input data:

- Removal of missing values and proper handling of NaNs.
- Conversion of timestamps into standardized datetime formats.
- Standardization of column names and careful feature selection.
- Filtering of anomalous or erroneous data points.

B. Exploratory Data Analysis (EDA)

Key insights from the EDA include:

- **Temperature Patterns (Fig. 1):**
  - Seasonal variations are evident, with temperatures peaking around 20 – 30°C.
- **Correlations (Fig. 2):**
  - Negative correlation between temperature and humidity.
  - Weak correlation between wind speed and pressure.
  - Positive correlation between precipitation and humidity.
- **Outlier Detection:** Identification of extreme values associated with unusual weather conditions.

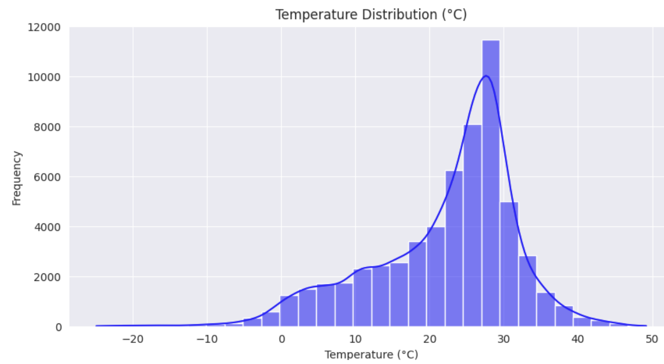

Fig. 1. Temperature Distribution

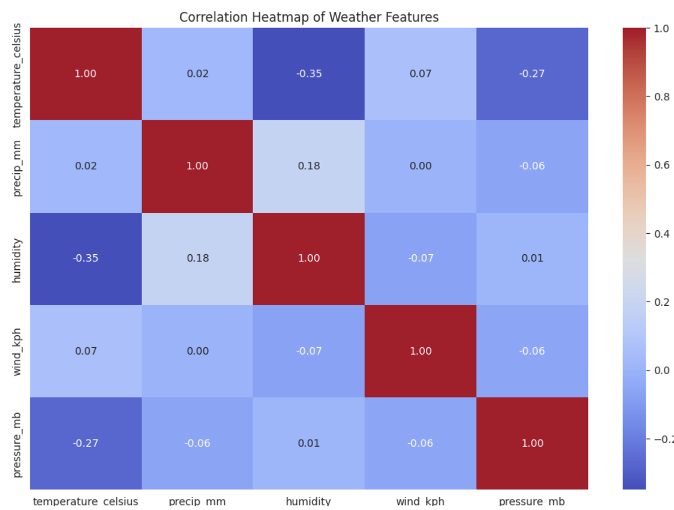

Fig. 2. Correlation Heatmap

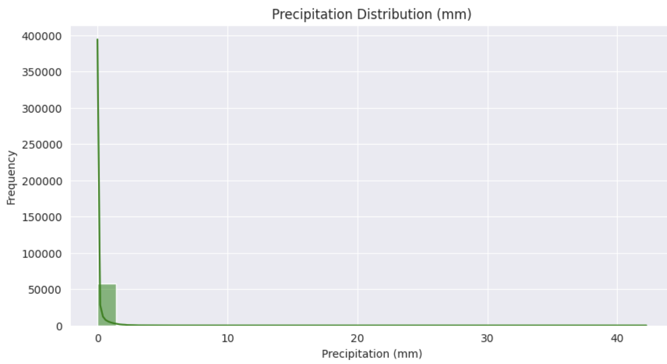

Fig. 3. Precipitation Distribution (mm)

III. Forecasting Models and Evaluation

A. Baseline Models

Performance Metrics

MAE: 8.31
RMSE: 9.78

Fig. 4. ARIMA Model

Performance Metrics

MAE: 9.99
RMSE: 11.58

Fig. 5. SARIMA Model

B. Advanced Model: LSTM Neural Network

The LSTM model was optimized using the Adam optimizer with Mean Squared Error as the loss function. Trained on sequences of historical temperature data, the model demonstrated enhanced performance.

Performance Metrics

MAE: 6.68
RMSE: 8.63

Fig. 6. LSTM Model

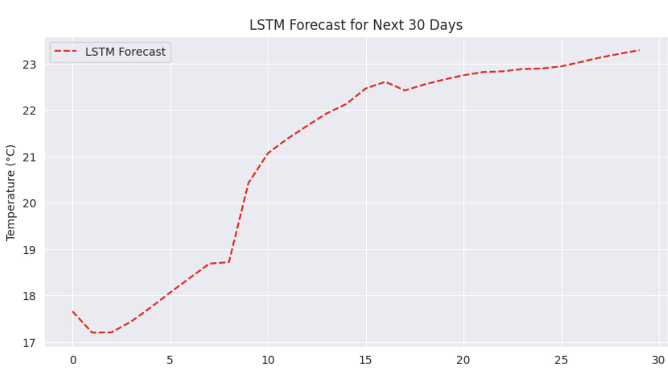

Fig. 7. LSTM Forecast for the Next 30 Days

IV. Advanced Analysis

A. Feature Importance via Random Forest

Key drivers of temperature variations were identified using a Random Forest model. The analysis revealed that:

- Humidity is a major driver.
- Air quality indicators (PM2.5 and PM10) significantly impact temperature readings.
- Wind speed also contributes to the observed variations.

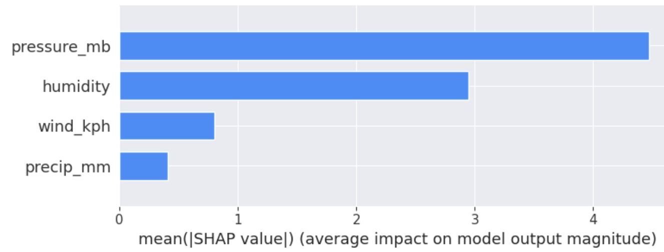

Fig. 8. Average Impact on Model Output Magnitude

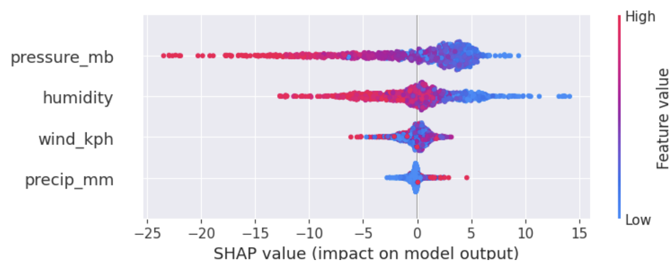

Fig. 9. Top 4 Impacts on Model Output

B. Anomaly Detection

Anomaly detection was performed using the Isolation Forest algorithm, alongside complementary methods:

- **Z-Score:** Detected 1088 anomalies.
- **Isolation Forest:** Detected 1174 anomalies.
- **Autoencoder:** Also identified 1174 anomalies.

These techniques helped filter out data points that might skew the forecasting performance.

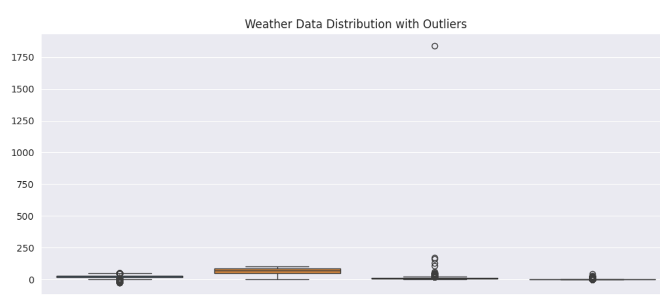

Fig. 10. Anomaly Detection Using Isolation Forest

V. Additional Climate and Geographical Analysis

A. Climate Analysis

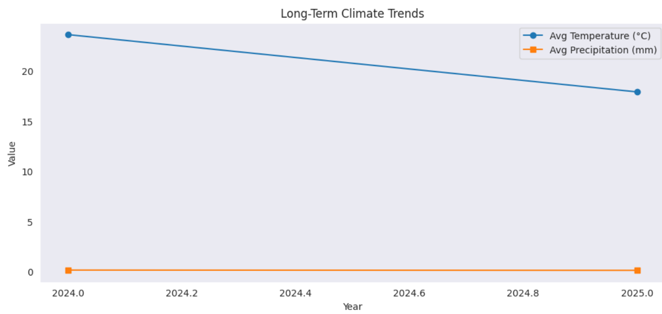

Fig. 11. Long-Term Climate Trends

Temperature Trend:

- The blue line represents the average temperature over time.
- There is a noticeable decline in average temperature from 2024 to 2025.

Precipitation Trend:

- The orange line represents average precipitation.
- Precipitation remains constant over the same period.

B. Geographical Patterns

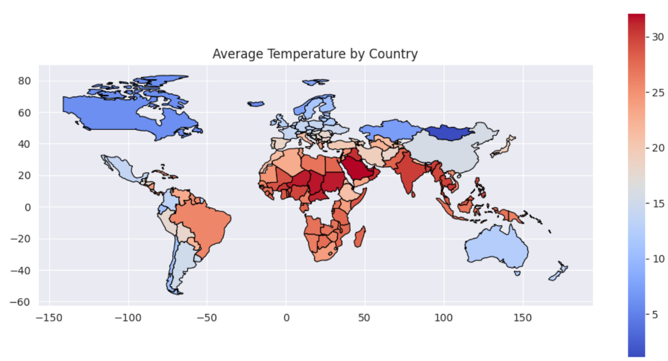

Fig. 12. Long-Term Climate Trends

Color Coding:

- The map uses a color gradient to represent average temperatures, with red indicating higher temperatures and blue indicating lower temperatures.

Regional Observations:

- Warmer regions are concentrated around the equator, particularly in Africa and parts of Asia.
- Cooler regions are found in higher latitudes, such as North America and Northern Europe.

VI. Insights and Recommendations

Based on our analysis, the following conclusions and recommendations are offered:

1) **LSTM Model Superiority:** The LSTM model delivers superior accuracy compared to traditional ARIMA/SARIMA models.
2) **Enhanced Feature Engineering:** Incorporating additional meteorological factors could further improve prediction accuracy.
3) **Importance of Anomaly Detection:** Filtering out extreme values enhances the reliability of forecasts.

VII. Repository and Further Resources

The full project, including code, models, and detailed visualizations, is available on GitHub:

GitHub Repository