**Weather Forecasting Models Comparison: Complete Report**

**PM Accelerator Mission**

"By making industry-leading tools and education available to individuals from all backgrounds, we level the playing field for future PM leaders. This is the PM Accelerator motto, as we grant aspiring and experienced PMs what they need most – Access. We introduce you to industry leaders, surround you with the right PM ecosystem, and discover the new world of AI product management skills."

PM Accelerator is committed to empowering aspiring and experienced Product Managers through quality education, real-world opportunities, and a supportive community. In line with this mission, we also extend our support to disadvantaged youths through PMA Kids – fostering the next generation of leaders and innovators.

**Introduction**

This report focuses on the importance of precise weather forecasting, using various statistical and machine learning models to predict weather conditions accurately.

The objective is to compare several forecasting models to determine the most effective approach for weather prediction.

**Data Overview:**

The dataset consists of weather readings from various locations around the globe, as shown in the sample data frame. Each entry includes data on temperature, humidity, air quality, and more, detailed with specific timestamps and geographic coordinates.

This dataset is pivotal for developing accurate and reliable weather forecasting models, encompassing a broad spectrum of climatic conditions.

**Exploratory Data Analysis (EDA):**

Statistical Summary: The dataset provides comprehensive statistics for various weather parameters, including temperature (in Celsius and Fahrenheit), humidity, wind speed, and air quality indices. A describe() function was used to generate descriptive statistics which help in understanding the central tendency, dispersion, and shape of the dataset’s distribution.

Histograms: The distribution of key numerical features such as temperature, humidity, and wind speed is visualized through histograms. These plots reveal the frequency of data points for different ranges, helping identify patterns or anomalies in weather conditions.

A graph of a growth

Description automatically generated with medium confidence

**Advanced EDA: Anomaly Detection**

Anomaly Detection Technique: The Isolation Forest algorithm, an effective method for identifying anomalies in large datasets, was applied to the temperature data. This model isolates anomalies instead of profiling normal data points.

Visualization of Anomalies: Anomalies in temperature were visualized on a scatter plot against the 'last\_updated' timestamp, highlighting the data points classified as anomalies (temperature outliers) distinct from the normal range.

Results Summary: A total of 586 temperature anomalies were detected. The descriptive statistics of these anomalies show significant deviations from the typical temperature readings, with some extremities reaching as high as 49.2°C, which could indicate data errors or unusual weather conditions.

A close-up of a temperature

Description automatically generated

A screenshot of a computer

Description automatically generated

**Forecasting with Multiple Models**

**Model Setup and Forecasting**

Data Preparation: The dataset was indexed by date to ensure the time series analysis could be appropriately conducted. Data cleaning included removing duplicates and ensuring the data was sorted by the date.

ARIMA Model: An ARIMA (AutoRegressive Integrated Moving Average) model was chosen for its capabilities in handling non-stationary data, which is typical in weather-related time series. The model was fitted using parameters that best captured the seasonal variations observed in preliminary analyses.

Prophet Model: Developed by Facebook, the Prophet forecasting tool is used for its robustness in dealing with missing data and its ability to model yearly, weekly, and daily seasonality. It requires the time series to be in a DataFrame with two columns: 'ds' (the date) and 'y' (the value).

**Ensemble Forecasting**

Combining Predictions: To leverage the strengths of both the ARIMA and Prophet models, an ensemble method was employed. This involved averaging the predictions from both models to potentially reduce forecast error and improve reliability.

**Model Evaluation**

Root Mean Square Error (RMSE): This metric was used to measure the average magnitude of the forecasting errors. It provides a good indication of how accurately the models are predicting the temperature. The lower the RMSE, the better the model's performance.

RMSE for ARIMA: 9.97542

RMSE for Prophet: 9.68874

RMSE for Ensemble: 9.19994

**Visual Analysis**

Temperature Forecast Comparison: The plot illustrates the actual temperatures against the forecasts made by the ARIMA model, the Prophet model, and the ensemble approach. Each model's forecast is represented in a different color (Actual in blue, ARIMA in red, Prophet in green, and Ensemble in purple), providing a clear visual comparison of each model's predictive accuracy over time.

A graph of a graph showing a number of different colored lines

Description automatically generated with medium confidence

**Unique Analyses**

**Climate Analysis: Temperature Distribution by Country**

Overview: This analysis investigates the temperature distribution across different countries using historical temperature data.

Method: A boxplot was used to visualize the range of temperatures for each country, displaying medians, interquartile ranges, and outliers. This visualization helps identify countries with extreme temperature variations and general climate behavior.

Findings: Countries near the equator show less variability in temperatures, while those with higher latitudes present a wider range of temperatures, reflecting seasonal changes more prominently.

A screenshot of a graph

Description automatically generated

**Environmental Impact Analysis: Air Quality vs. Climate Parameters**

Overview: This section explores the relationship between various air quality indices and climate parameters such as temperature, humidity, and wind speed.

Method: A correlation matrix was employed to quantify and visualize the strengths of the relationships between air quality components (PM2.5, PM10, ozone, nitrogen dioxide) and weather parameters.

Findings: The analysis reveals moderate correlations between air quality and humidity, suggesting that higher humidity levels might be associated with increased particulate matter due to condensation and moisture retention in the air.

A screen shot of a chart

Description automatically generated

**Feature Importance in Weather Prediction**

Overview: This analysis identifies the most important features affecting temperature predictions.

Method: A Random Forest Regressor was utilized to assess the importance of each feature in predicting temperature.

Findings: Atmospheric pressure (pressure\_mb) and latitude are among the top predictors of temperature, indicating that geographic and atmospheric conditions significantly influence temperature variations.

A graph with blue bars

Description automatically generated

**Spatial Analysis of Weather Patterns**

Overview: This analysis visualizes global temperature patterns to provide a spatial understanding of climatic conditions.

Method: Temperature data were plotted on a global map using latitude and longitude coordinates to show temperature distributions across different geographical locations.

Findings: The visual highlights how temperature varies significantly with geographical location, with colder temperatures prevalent in higher latitudes and warmer temperatures near the equator.

A map of the world

Description automatically generated

**Geographical Patterns: Temperature Differences by Continent**

Overview: This analysis aggregates temperature data by continent to explore regional climate patterns.

Method: Temperature data for countries were mapped to their respective continents, and a boxplot was created to show temperature distribution by continent.

Findings: The plot illustrates significant differences in temperature distributions across continents, with Europe showing a wider range of temperatures compared to more consistent conditions in continents like Asia and South America due to their larger equatorial regions.

A diagram of a graph

Description automatically generated

**Conclusion**

The comparison of different models illustrates the advantages of using advanced machine learning techniques in weather forecasting.

Future research should focus on integrating more diverse data sources and exploring hybrid models that can leverage the strengths of both statistical and machine learning approaches.