Weather Analysis and Forecasting Report

Project Overview

In this project, I analyzed a global weather dataset containing temperature, precipitation, wind, air quality, and geographical features. My goal were to:

1. Explore and understand global weather patterns through structured exploratory data analysis
2. To build forecasting models capable of predicting daily average temperature.

The dataset includes timestamped weather observations across multiple countries and cities (Date range of data: 2024-05-16 01:45:00 to 2026-02-20 20:00:00). I focused on temperature as the primary target variable while also examining its relationship with environmental and geographical factors.

The *PM Accelerator mission* is centered around access. By making industry-leading tools and education available to individuals from all backgrounds, we level the playing field for future PM leaders. This project reflects that spirit by combining data analysis, forecasting models, and AI-driven techniques to better understand global weather patterns and their environmental impact. Through this work, I aim to apply practical data science methods that align with the growing world of AI product management skills.

Data Cleaning and Preparation

I began by loading the dataset and conducting an initial inspection to understand its structure, variable types, and completeness. This helped identify potential preprocessing steps required before analysis and modeling.

Data Preparation Steps

- Examined dataset shape, column names, and data types
- Checked for missing values across all features
- Parsed datetime-related columns
- Extracted temporal features from timestamps: Month, days of week and hour
- Using last_updated as the primary time index allowed structured time-series modeling.

Extracting these time-based features allowed for more meaningful temporal trend analysis and improved forecasting performance.

Outlier Treatment

The numeric variables showed noticeable skewness and extreme values, which could negatively affect model performance. To address this, I:

- Applied z-score based outlier detection
- Removed observations where |z-score| > 3

```
Original dataset size: 125306 rows
Dataset size after removing outliers: 111130 rows
```

Handling Redundant Variables

Several features represented identical information in different units:

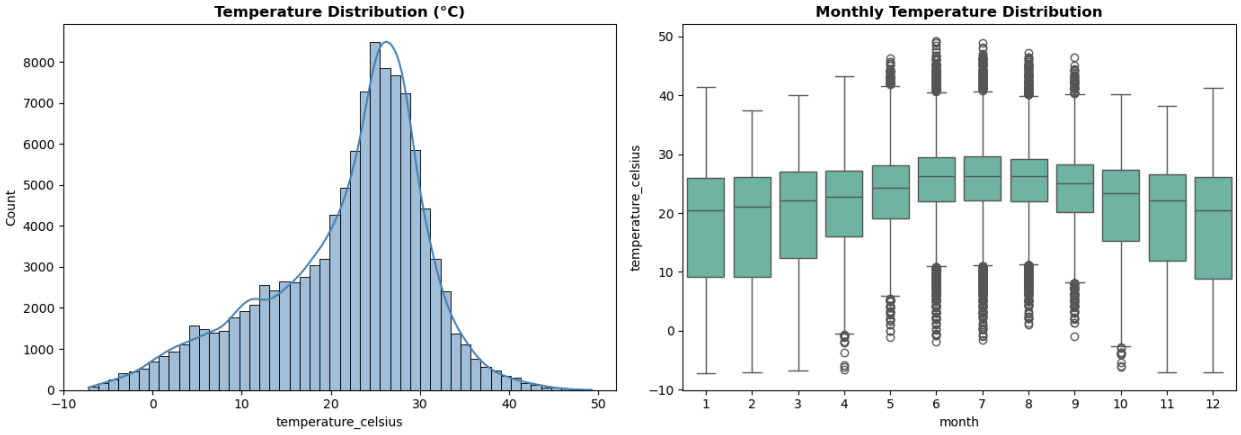
- Celsius vs Fahrenheit
- mph vs kph
- mb vs inches

I removed duplicate unit-based variables to prevent multicollinearity.

This ensured a cleaner dataset and more reliable model interpretation.

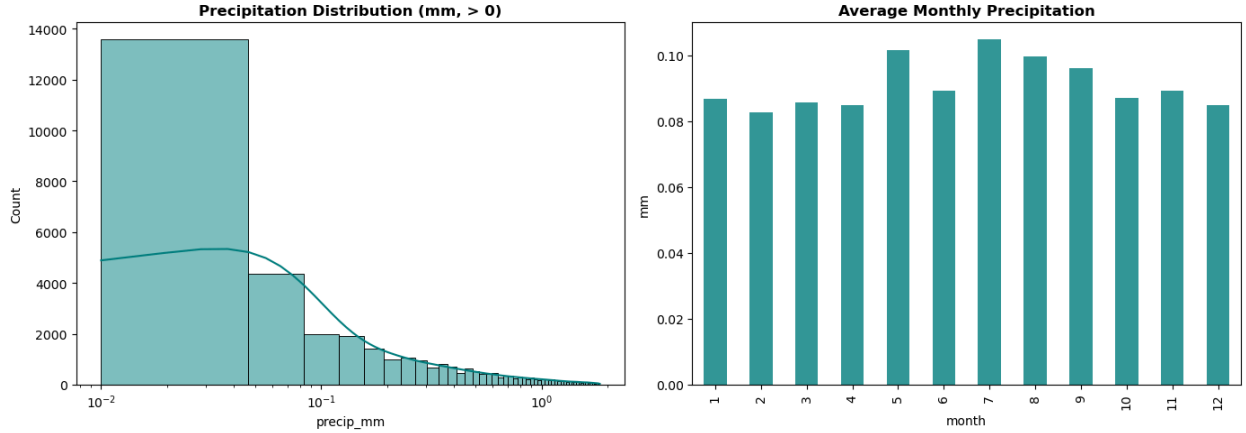
Exploratory Data Analysis

I explored the **global temperature distribution** using histograms and boxplots grouped by month.



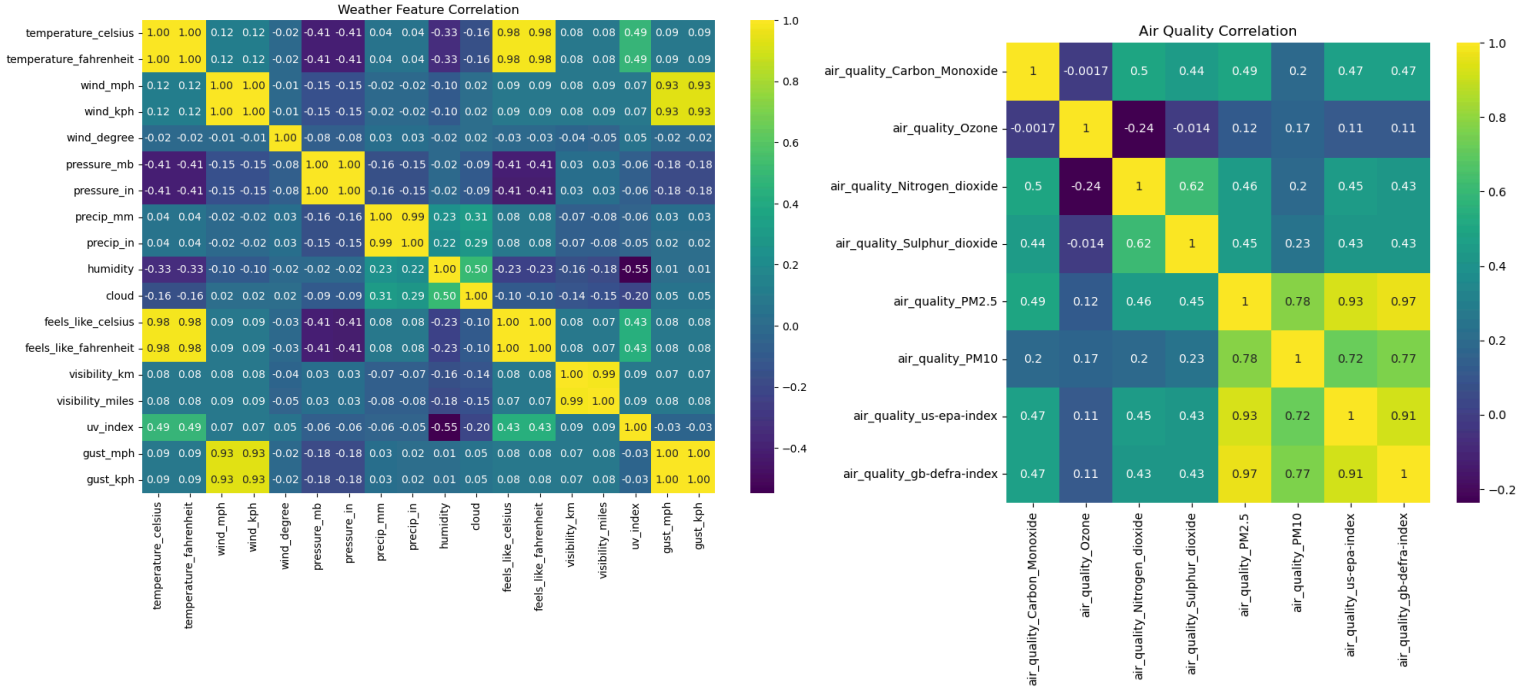
- Temperature shows seasonal variation.
- Higher spread observed in certain months, indicating climate variability.

Precipitation distribution was highly skewed, so I visualized it on a log scale to better understand distribution patterns.



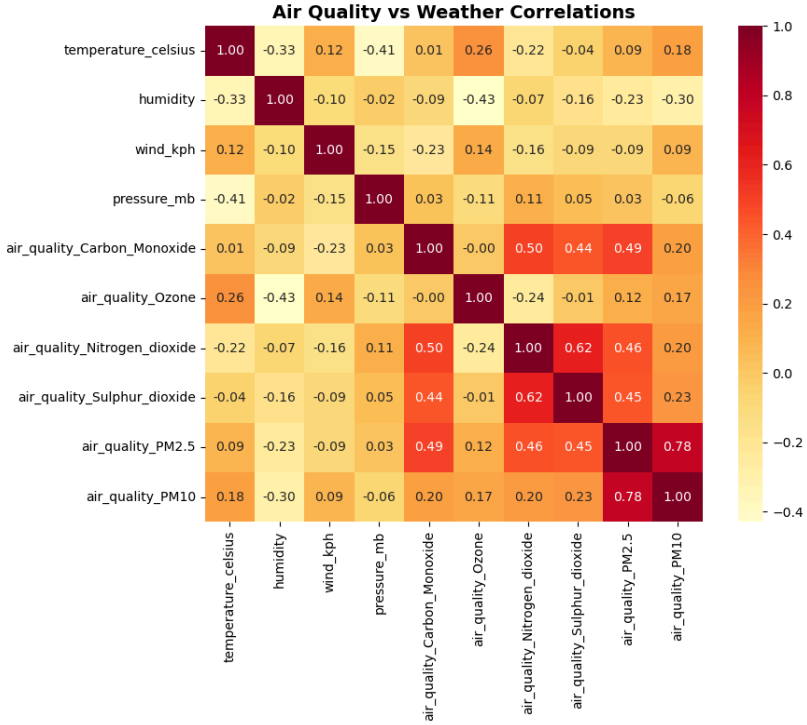
- Precipitation is highly skewed.
- Some months show concentrated rainfall patterns.

Correlation heatmaps helped identify relationships between weather variables and air quality metrics.



- Strong correlation between temperature and UV index.
- Latitude shows clear inverse relationship with temperature.
- Air pollutants show mild but noticeable relationships.
- PM2.5 and ozone vary across weather conditions but are not primary temperature drivers.

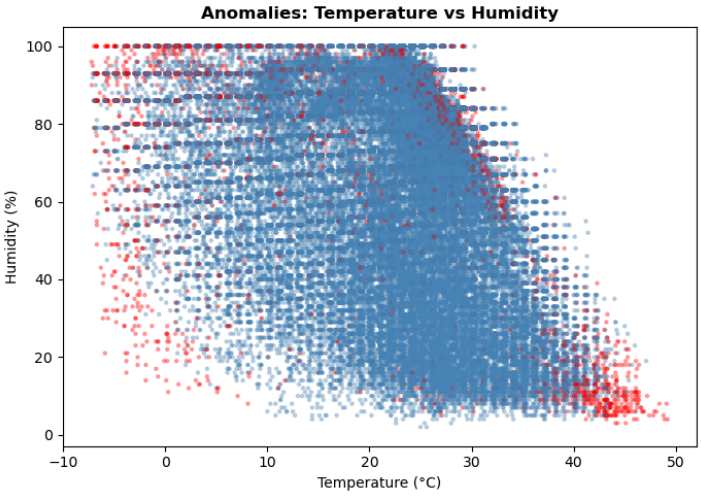
To understand environmental impact, I examined the relationship between air quality and weather variables.

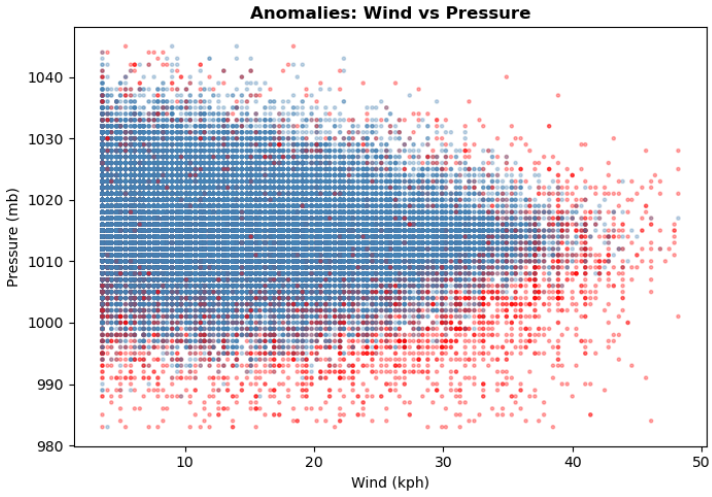


There were mild but visible relationships between temperature and certain pollutants, especially ozone.

Advanced EDA: Anomaly Detection

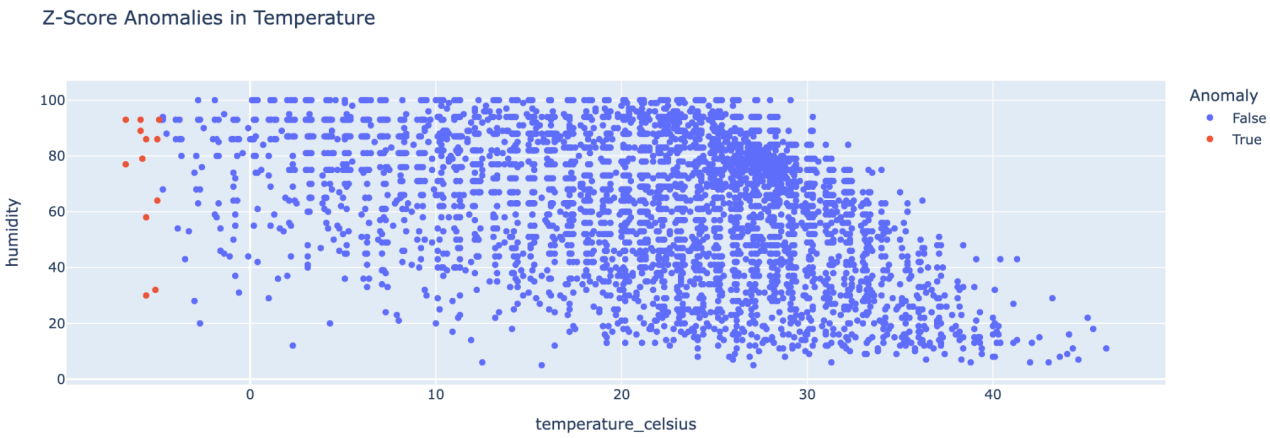
I applied Isolation Forest to detect anomalous weather conditions using temperature, humidity, wind speed, pressure, and precipitation.





Approximately 5% of observations were flagged as anomalies. These likely represent extreme weather events or rare environmental conditions.

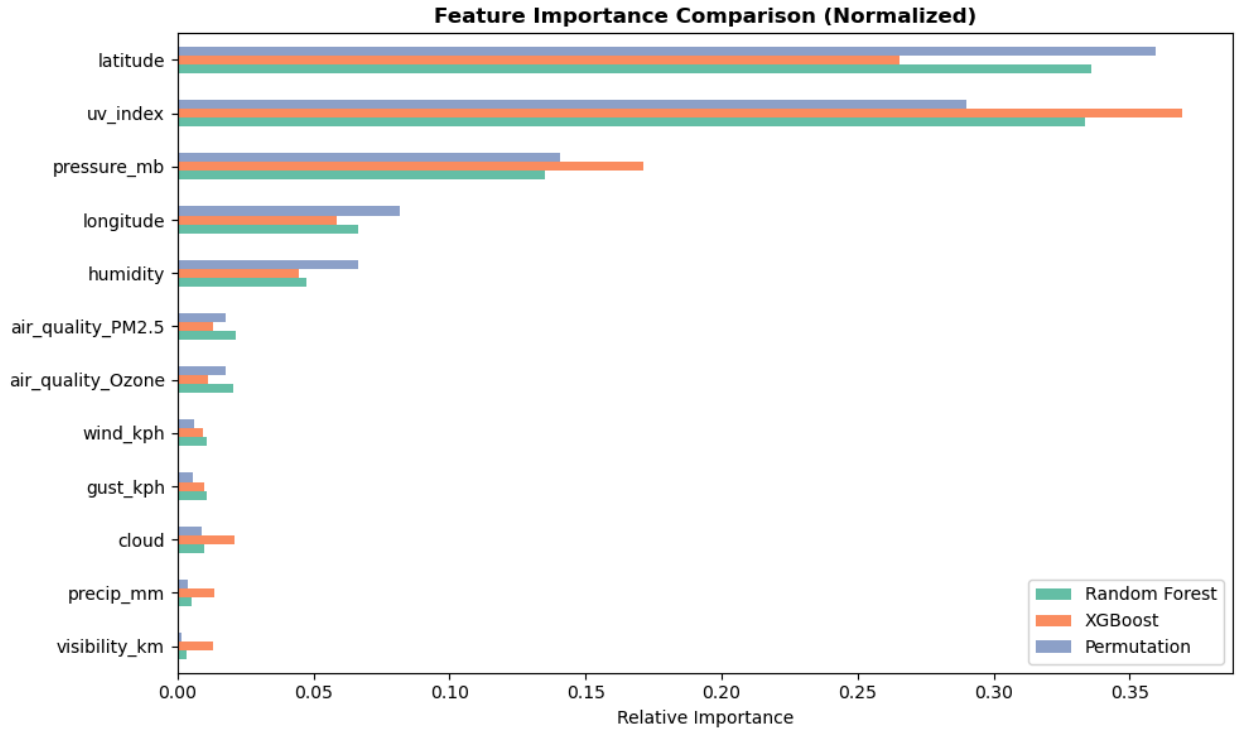
I also applied a z-score based anomaly check specifically on temperature to identify extreme deviations.



Both methods identified overlapping extreme points, confirming consistency.

Feature Importance Analysis

To quantify feature influence, I trained both Random Forest and XGBoost models using weather and environmental features to predict temperature. I also computed permutation importance for robustness.



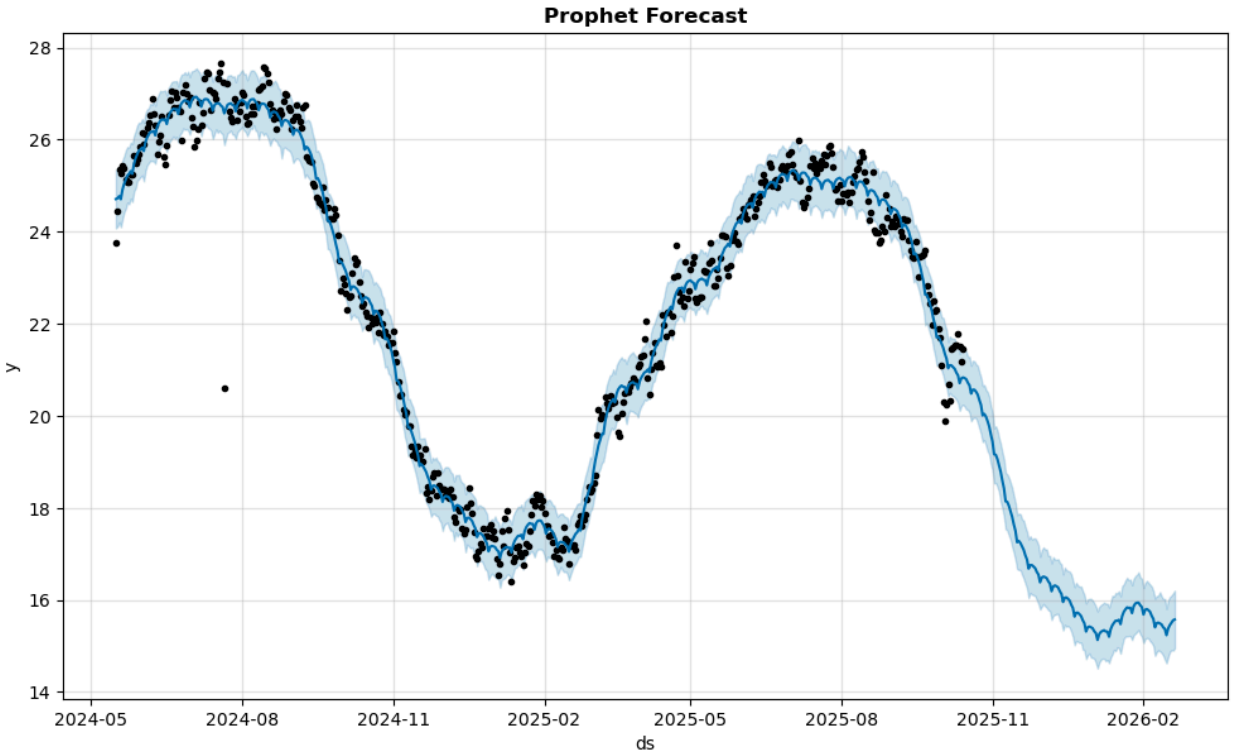
Latitude and **UV index** emerged as the most important predictors. This confirms that geographical location and solar radiation are dominant drivers of temperature. Pressure contributed moderately, while wind, precipitation, and pollutant variables had lower relative importance.

Time Series Forecasting

For forecasting, I evaluated multiple models including Prophet, ARIMA, SARIMA, Exponential Smoothing, Random Forest with lag features, and XGBoost.

I used MAE, RMSE, and R-squared as evaluation metrics. All forecasting used last_updated as the time index with proper time-based split (80/20).

Time-Series models: Prophet, ARIMA, SARIMAX and Exponential Smoothing



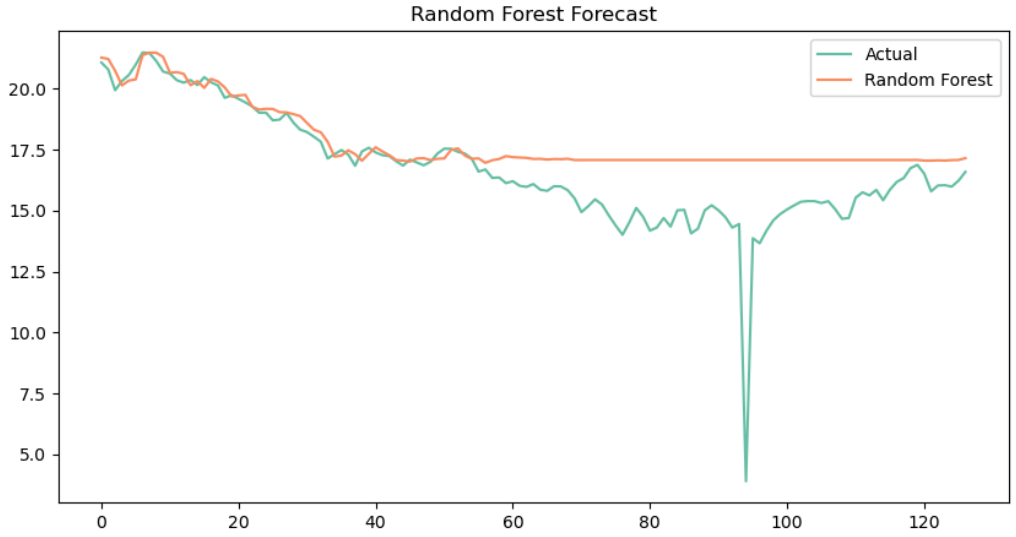
Classical statistical models like ARIMA and Exponential Smoothing captured trend behavior reasonably well.

Machin-learning Models: XGBoost and Random Forest

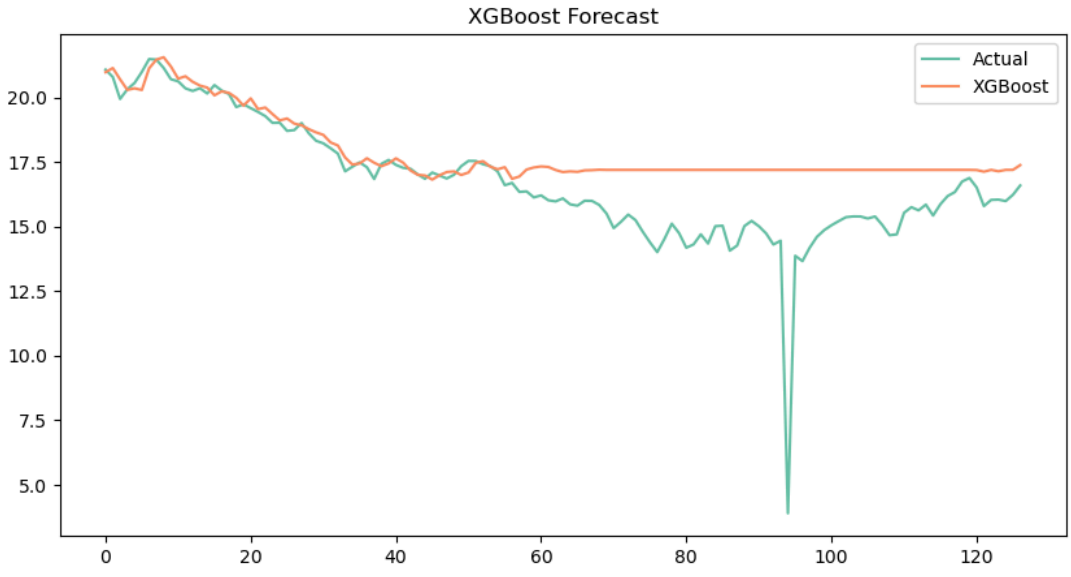
In addition to statistical time-series models, tree-based machine learning models were implemented to capture potential nonlinear relationships in the data. Lag features, rolling statistics, and calendar-based variables (e.g., day-of-week) were used as predictors.

However, tree-based models with lag features, particularly XGBoost, achieved lower error metrics and better generalization.

The Random Forest model improved predictive accuracy by averaging multiple decision trees, reducing variance and overfitting. It successfully captured nonlinear interactions between lagged values and seasonal indicators. However, because Random Forest does not inherently model temporal dependence, its performance depends heavily on feature engineering.



Feature importance analysis indicated that recent lag values and weekly seasonality indicators were the most influential predictors.



Model Comparison Summary

Across all experiments, machine learning models with engineered features consistently outperformed classical time-series models in predictive accuracy. Prophet performed well for seasonal structure, but XGBoost provided better flexibility in capturing nonlinear interactions.

The best model was selected based on lowest MAE.

Model	MAE	RMSE	R^2
Prophet	**0.732527**	**1.283841**	**0.710902**
ARIMA(5,1,2)	4.608747	5.189174	-3.723013
SARIMA	3.599509	4.106020	-1.957092
Exp. Smoothing	1.884905	2.397579	-0.008251
Random Forest	1.177486	1.847621	0.378798
XGBoost	1.231689	1.911673	0.334981

Key Insights

- Temperature is primarily driven by geography and solar exposure. Environmental pollutants show weaker direct influence but may interact indirectly through broader climate patterns.

- Global weather variability differs significantly by region. Some countries exhibit higher temperature variance, suggesting stronger seasonal or continental climate behavior.

- Machine learning approaches offer improved forecasting performance when temporal lag features are included. Classical models remain useful for interpretability and baseline comparisons.

Conclusion

In this project, I combined exploratory analysis, environmental impact study, anomaly detection, and forecasting models to build a comprehensive understanding of global weather behavior. The results demonstrate how structured EDA combined with both statistical and machine learning models can produce reliable forecasting systems.

This project reflects applied AI modeling in a real-world setting and aligns with the broader mission of expanding access to advanced tools and data-driven decision making.