PM Accelerator

Tech Assessment: Weather Trend Forecasting

Tony Lai

Mar14 2025

Exploratory Data Analysis (EDA)

This section provides an overview of the data and highlights key insights that impact the prediction of temperature using the Ordinary Least Squares (OLS) Regression model.

Data Summary

- Dataset Overview: The dataset comprises 58,410 daily weather observations collected from various global locations, covering the period from May 16, 2024, to March 14, 2025. It encompasses 41 features, with **temperature_celsius** designated as the target variable and **last_updated** serving as the time index. These features include meteorological variables that may influence temperature.
- Target Variable: **temperature_celsius** represents the daily average temperature (°C) across locations, which is the primary focus of this analysis. The goal is to forecast future temperatures by leveraging historical trends and external factors.

Key Features Impacting Temperature

- **temp_t-1**: Lagged temperature, expected to have strong autocorrelation due to time-series dependency.
- **uv_index**: Reflects solar radiation, a key driver of temperature.
- **humidity**: Influences temperature via moisture and cloud cover.

Correlation with Temperature

A correlation heatmap (Figure 1) identified relationships with temperature_celsius:

- temp_t-1: Exhibits a correlation of approximately 0.97 , indicating a very strong autoregressive relationship with the current temperature.
- uv_index: Shows a correlation of 0.53, suggesting a moderate positive influence on temperature prediction.
- humidity: Displays a correlation of -0.35, indicating a moderate negative association, where higher humidity tends to correspond with lower temperatures.

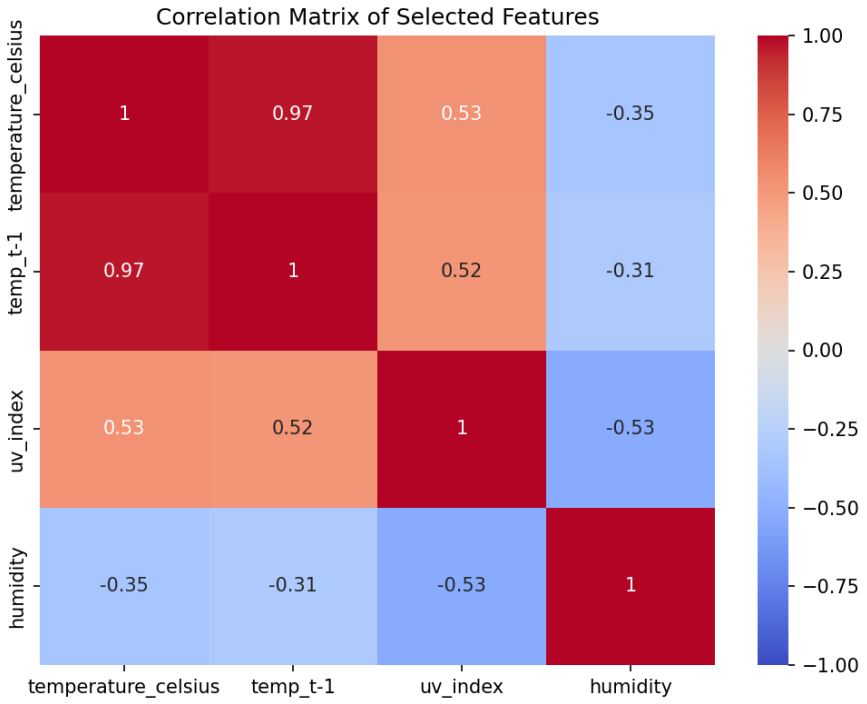

Figure 1. Feature Correlation Heatmap

Visualizations of Key Features

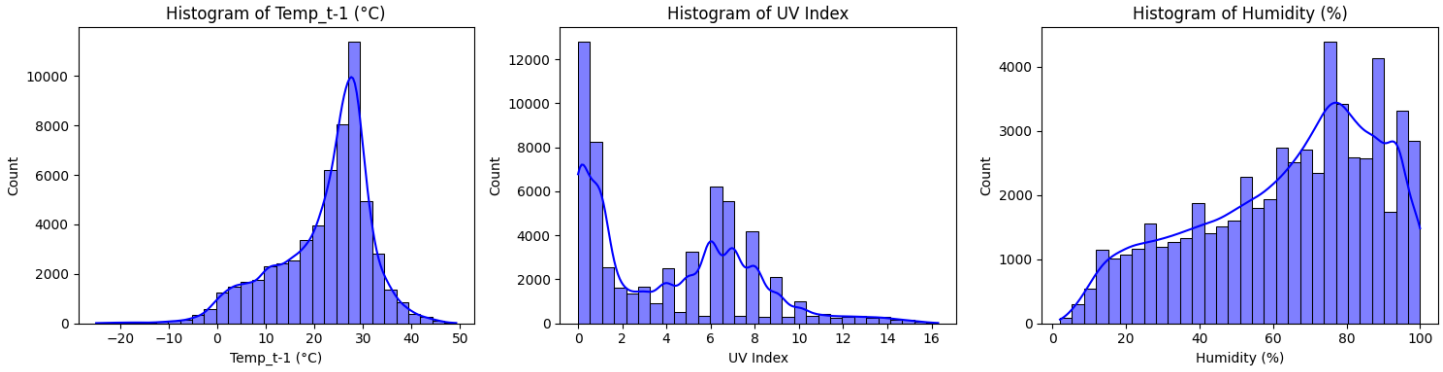

Figure 2. Histograms of Feature Distributions

- Temp_t-1 (°C): Slightly right-skewed, mostly between 10°C and 30°C, with rare extremes below 0°C and above 40°C, reflecting diverse global climates.
- UV Index: Heavily right-skewed, peaking at 2–4, with a tail to 16, indicating low to moderate UV levels, higher in equatorial regions.
- Humidity (%): Left-skewed, peaking at 70%–90%, with fewer values below 40%, suggesting prevalent high humidity, likely in coastal or tropical areas.

Time Series Trends and Cycles

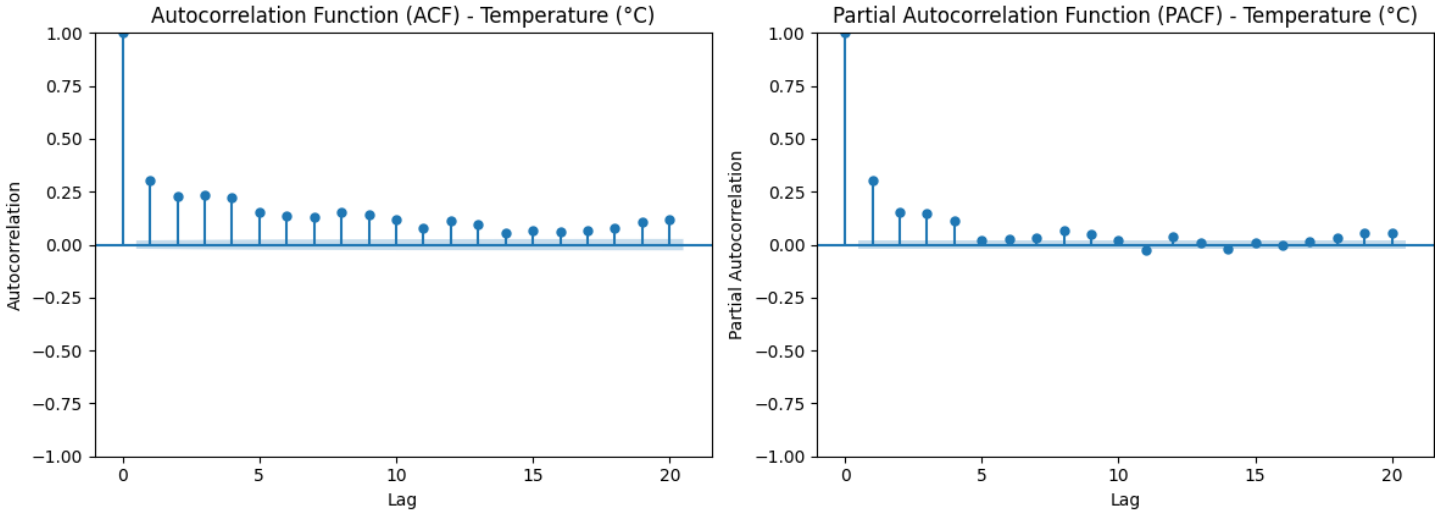

Figure 3. ACF & PACF Plots

- Autocorrelation Function (ACF) - Temperature (°C): High initial autocorrelation (0.75 at lag 1), gradually decreasing, showing strong persistence in temperature trends over time.
- Partial Autocorrelation Function (PACF) - Temperature (°C): Significant spike at lag 1 (0.75), near-zero afterward, indicating a direct autoregressive effect primarily from the previous day.

Global Temperature Trend Over Time

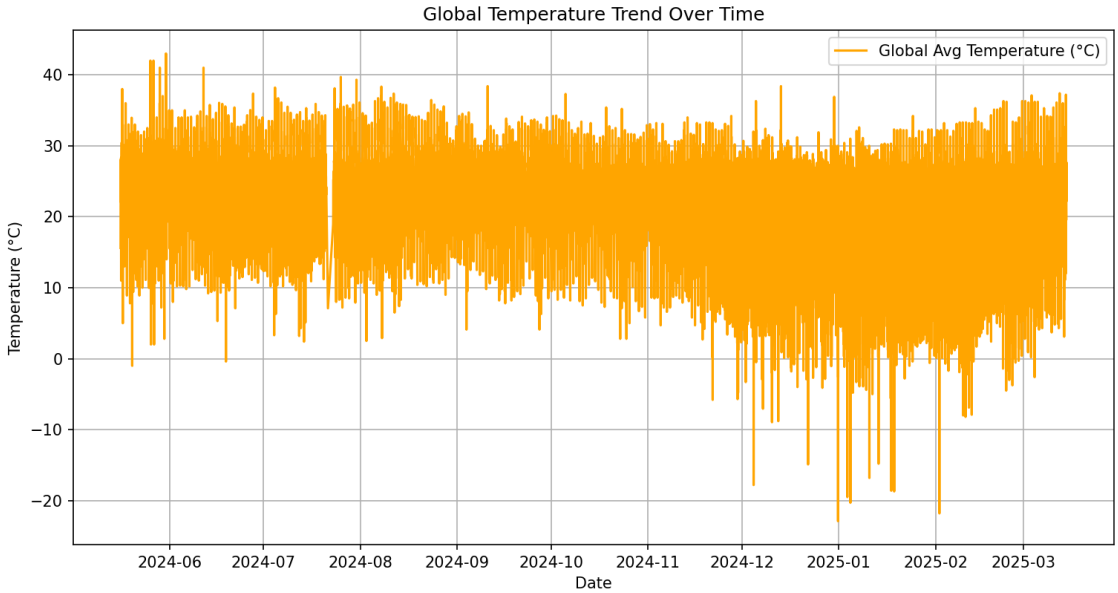

Figure 4. Global Temperature Over Time

- Trends: The global temperature fluctuates over time, showing seasonal variations.
- Cycles: Temperature patterns exhibit periodic changes, likely due to seasonal influences.
- Variability: Short-term fluctuations indicate sensitivity to external factors such as weather anomalies.

Model Development

Three models were developed to forecast power generation: OLS Regression, Holt-Winters Exponential Smoothing, and ARIMA.

•1. Ordinary Least Squares (OLS) Regression

- Technique: A multiple linear regression model minimizing squared residuals to fit temperature against lagged temperature (Temp_t-1), UV index, and humidity.
- Purpose: Captures linear relationships between past temperature values and external factors, enhancing predictive accuracy.

2. Holt-Winters Exponential Smoothing

- Technique: Applies additive trend smoothing to temperature, adjusting for gradual changes while dampening short-term fluctuations.
- Purpose: Suitable for trend-driven data but limited by its exclusion of external weather factors.

3. ARIMA

- Technique: An ARIMA(1,1,2) model with autoregression (p=1), first-order differencing (d=1), and moving average (q=2) components.
- Purpose: Models time-series dependencies in temperature trends but lacks external predictors like humidity or UV index.

Model Evaluation

Three models were evaluated for predicting power generation. Table 1 presents a summary of each model's features, evaluation metrics, and overall performance.

Model Type	Features	Scaling or smoothing or other treatments	RMSE	Backtest RMSE (Last 6 Days)
OLS Regression	'temp_t-1', 'uv_index', 'humidity'	None	2.241009	1.912048
Holt-Winters	temperature_celsius	Additive trend smoothing	3.861684	3.261261
ARIMA	temperature_celsius	Differencing (1,1,1)	3.048725	3.177727

Table1. Comparison of Models

Analysis of RMSE Results

- OLS Regression: Lowest RMSE across all tests (1.91–2.24), demonstrating superior precision.
- Holt-Winters: Higher errors (3.26–3.86), struggling with short-term variability due to its reliance on smoothing.
- ARIMA: Moderate performance (3.05–3.18), outperformed by OLS due to its exclusion of external factors.

Model Performance Explanation

- OLS: Excels by leveraging multiple correlated features, capturing both autoregressive and external influences.
- Holt-Winters: Limited to trend smoothing, missing critical external drivers, resulting in poor adaptability.
- ARIMA: Effective for autocorrelation but less robust without additional predictors.

Conclusion

OLS Regression outperforms Holt-Winters and ARIMA, achieving the lowest RMSE and effectively integrating lagged temperature values and external factors (UV index, humidity). Its consistent accuracy across tests makes it the optimal choice for forecasting temperature trends.