

Dengue Time-Series Forecasting

A Structured Machine Learning Approach Using Climate and Temporal Features

Hardik Thapar
Independent Technical Project
hardikthapar1@gmail.com

Abstract

This report presents the development of a structured machine learning pipeline for forecasting weekly dengue cases in San Juan (Puerto Rico) and Iquitos (Peru) using environmental and temporal data. The project began with conventional regression baselines and evolved into a time-aware modeling system incorporating autoregressive lag features and strict chronological validation. Performance improved from baseline MAE values of approximately 30 to 11.8 for San Juan and 4.7 for Iquitos under cross-validation. The emphasis of this work lies in disciplined methodology, proper validation design, and systematic feature engineering rather than algorithmic novelty.

1 Project Motivation and Objective

Dengue fever is a climate-sensitive mosquito-borne disease whose transmission dynamics are closely linked to environmental conditions and seasonal variation. Predicting weekly dengue case counts is a practical forecasting problem with real-world public health implications.

The objective of this project was to build a reliable forecasting model capable of predicting weekly dengue incidence using structured environmental data and historical case information.

The guiding principles of this project were:

- Build from simple baselines upward.
- Respect the temporal structure of the data.
- Evaluate models under realistic deployment conditions.
- Focus on interpretability and robustness.

Rather than pursuing complex deep learning architectures, this work focused on extracting maximum predictive value from structured features and disciplined validation.

2 Dataset Overview

The dataset was sourced from the DrivenData DengAI competition. It consists of weekly records indexed by:

(city, year, weekofyear)

Two cities were modeled independently due to differing climatic and epidemiological patterns:

- San Juan, Puerto Rico
- Iquitos, Peru

Each weekly observation contains:

- Weather station temperature measurements

- Precipitation data
- Reanalysis humidity metrics
- Vegetation indices (NDVI)
- Observed dengue case counts (target variable)

The performance metric used throughout this project is Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

The test set is strictly a future hold-out set, requiring chronological validation during model development.

3 Exploratory Data Analysis

Initial exploratory analysis focused on understanding the structure of the target variable.

Key observations:

- Long periods of low case counts.
- Sudden outbreak spikes.
- Strong temporal continuity between consecutive weeks.

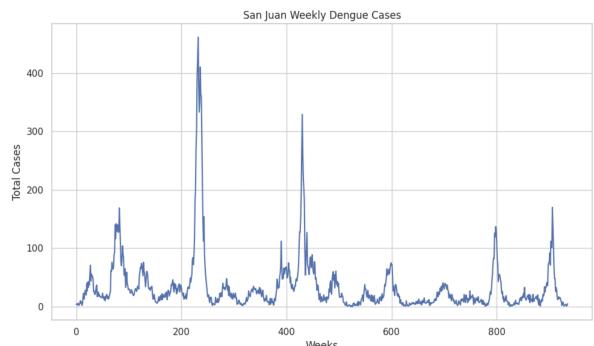


Figure 1: Weekly dengue cases – San Juan

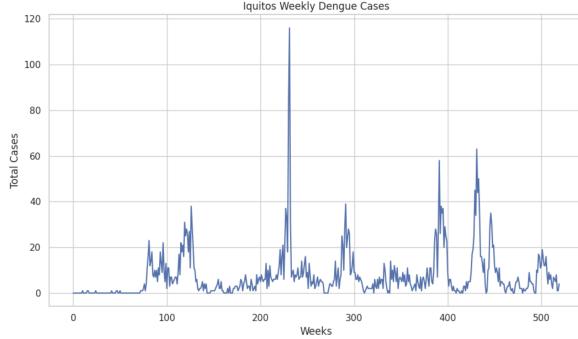


Figure 2: Weekly dengue cases – Iquitos

These plots clearly indicated that weekly cases are not independent observations. Instead, they exhibit outbreak dynamics with momentum and persistence.

This shifted the modeling strategy from conventional tabular regression toward time-aware modeling.

4 Baseline Modeling

Five climate variables were selected based on correlation analysis and domain plausibility:

- Specific humidity
- Dew point temperature
- Average temperature
- Minimum temperature
- Precipitation

Two baseline models were implemented:

- Linear Regression
- Random Forest Regressor

Table 1: Baseline Performance (MAE)

Model	San Juan	Iquitos
Linear Regression	30	31
Random Forest	31	31

The models struggled to capture outbreak spikes, indicating that static climate variables alone were insufficient.

5 Temporal Feature Engineering

To incorporate outbreak momentum, autoregressive features were introduced:

- Lag 1–4 weeks
- Rolling mean (4 weeks)
- Rolling mean (8 weeks)
- Rolling standard deviation (4 weeks)

All lag and rolling features were computed using strictly past observations.

These features allowed the model to capture:

- Short-term outbreak acceleration

- Medium-term seasonal buildup
- Volatility in recent case counts

The addition of temporal features significantly increased predictive capacity.

6 Model Selection and Training

CatBoost Regressor was selected due to:

- Strong performance on structured tabular data
- Robust handling of nonlinear feature interactions
- Stable optimization under MAE loss

A log transformation was applied to stabilize variance:

$$y' = \log(1 + y)$$

Predictions were inverse-transformed prior to evaluation.

Hyperparameters were tuned manually based on cross-validation performance.

7 Validation Strategy

A 5-fold TimeSeriesSplit approach was used to simulate realistic forecasting conditions.

Each fold ensured:

- Training data precedes validation data chronologically
- No shuffling of time indices
- Validation mimics future prediction

Table 2: Cross-Validated Performance

City	Mean MAE	Std
San Juan	11.8	2.1
Iquitos	4.7	1.0

These results represent stable, reproducible performance under realistic evaluation.

8 Model Diagnostics and Interpretability

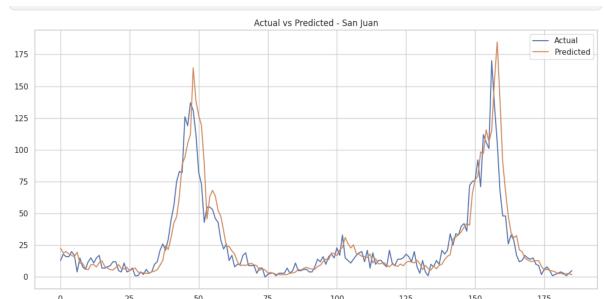


Figure 3: Actual vs Predicted – San Juan

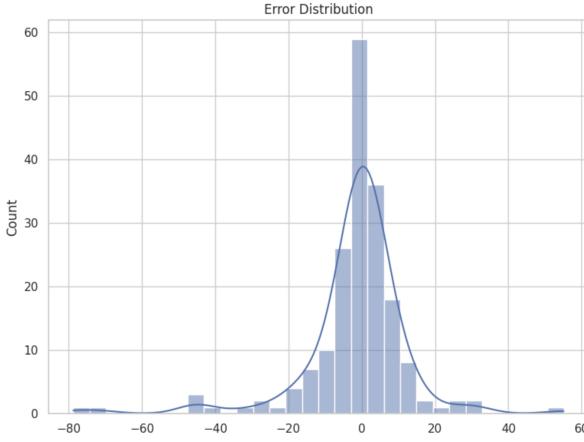


Figure 4: Prediction Error Distribution

The model successfully captured outbreak spikes while maintaining low bias during flat periods.

Feature importance analysis revealed that autoregressive lag variables were among the strongest predictors.

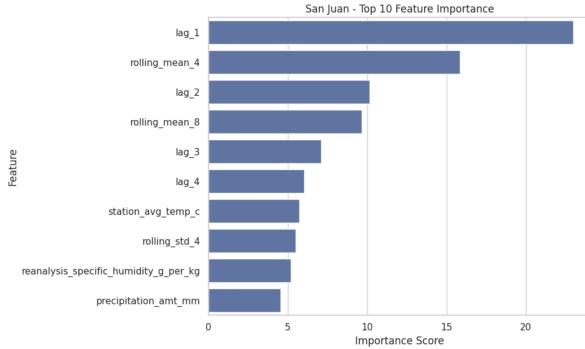


Figure 5: Feature Importance – San Juan

This confirms that temporal momentum is a primary driver of dengue incidence.

9 Workflow Summary

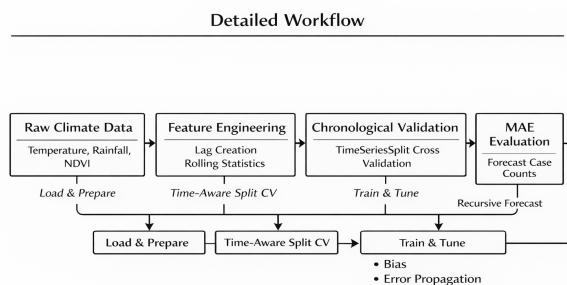


Figure 6: End-to-End Modeling Workflow

The workflow consists of:

1. Data ingestion and preprocessing
2. Feature selection
3. Temporal feature engineering
4. Chronological cross-validation
5. Hyperparameter tuning
6. Final model training

10 Final Results

Performance improved substantially over baseline:

- Baseline MAE: 30
- Final CV MAE (San Juan): 11.8
- Final CV MAE (Iquitos): 4.7

The primary improvement driver was structured temporal modeling rather than architectural complexity.

11 Key Learnings

- Time dependency dominates outbreak forecasting.
- Proper validation design determines credibility.
- Simple feature engineering can outperform naive ensemble upgrades.
- Interpretability and robustness matter as much as accuracy.

12 Conclusion

This project demonstrates that disciplined methodology and temporal awareness are critical in epidemiological forecasting tasks.

By respecting chronological structure and systematically engineering lag features, substantial performance improvements were achieved over baseline regression models.

The final system is reproducible, interpretable, and suitable for practical forecasting applications using structured environmental data.