İTÜ | FACULTY of COMPUTER ENGINEERING

PREDICTING WEEKLY DENGUE CASES

BLG527E-Machine Learning Term Project

Malik Kalembaşı(704241038)

Tanalp Oğuz(708241019)

14.01.2026

Modeling Disease Dynamics in Two Distinct Climates

San Juan (Puerto Rico)

Environment: Urban, Coastal
Driver: Temperature Range

Strategy:
Separate Models Trained for Each City

Iquitos (Peru)

Environment: Amazon Rainforest
Driver: Vegetation & Humidity

Insight Box

Core Goal: Predict weekly cases using DrivenData meteorological features (Temp, Rainfall, NDVI).

Evaluation Metric: Mean Absolute Error (MAE).

The Baseline: Limits of k-Polynomial Regression

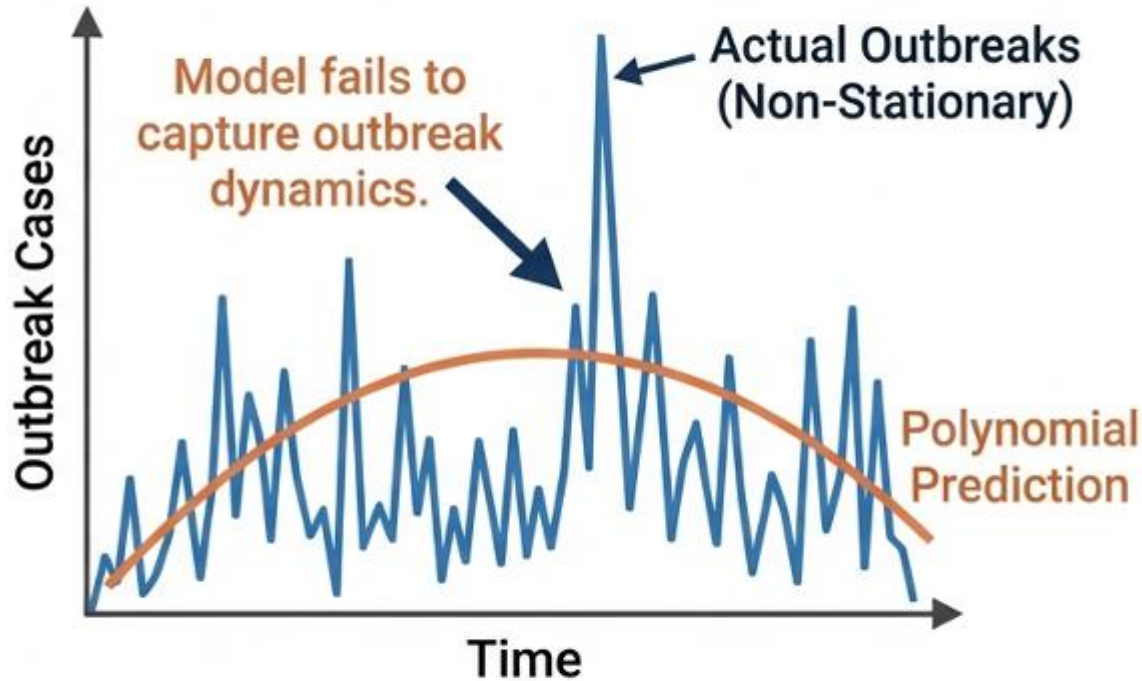
Methodology:

- **Model:** Polynomial regression (degree k) with Ridge regularization.
- **Features:** Seasonal encoding (sine/cosine), lagged variables, rolling means.
- **Inference:** Recursive multi-step forecasting.

The Outcome:

- **Optimal Degree:** k=1 (Linear). Higher degrees caused instability.
- **Result**: MAE 27.19 (Rank 4798).

The Underfitting Problem

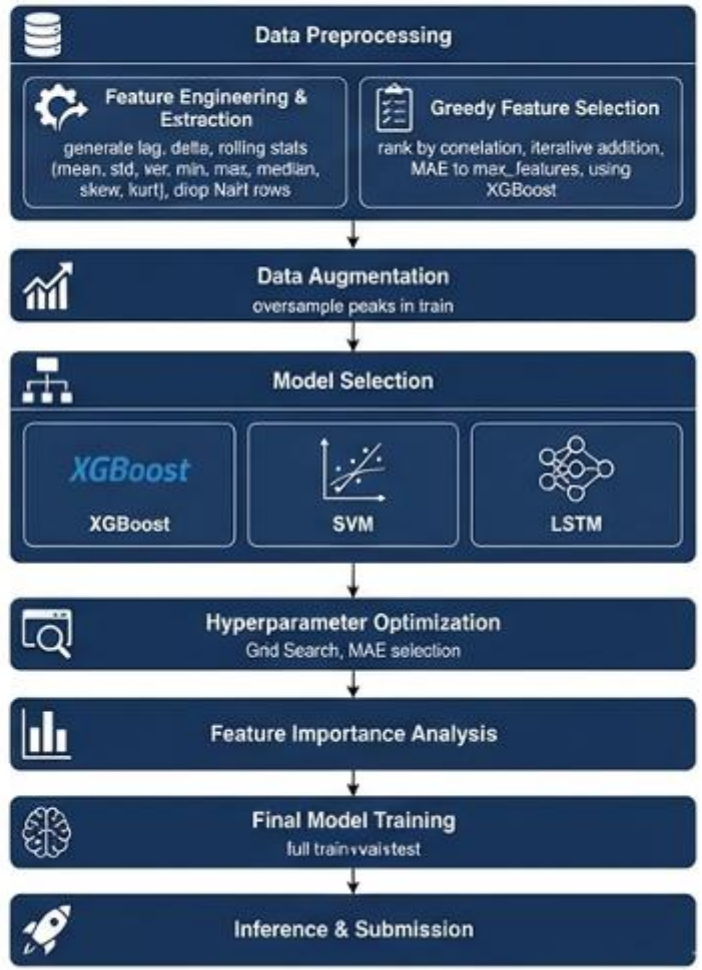


Critical Analysis: Simple curve fitting is insufficient for non-stationary biological data.

The Complexity Trap: A Heterogeneous Ensemble

Strategy: Diversity

Combined sequence learning (LSTM) with non-linear regression (SVM, XGBoost).

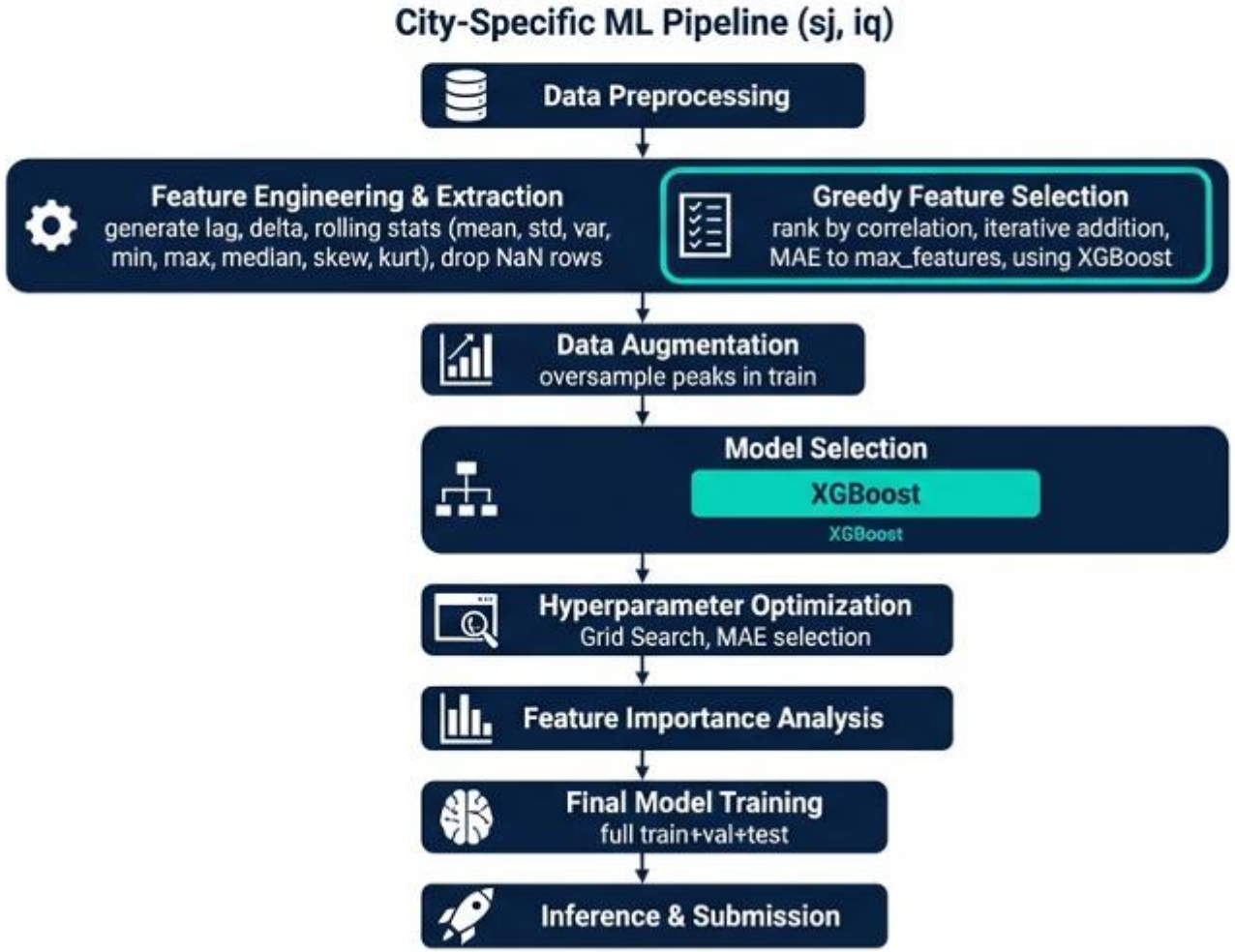


Performance

MAE: 23.31 (Rank 1463)

Insight: The ensemble outperformed buselt was weighed down by weaker learners (SVM/LSTM).

The Winning Solution: Feature-Engineered XGBoost

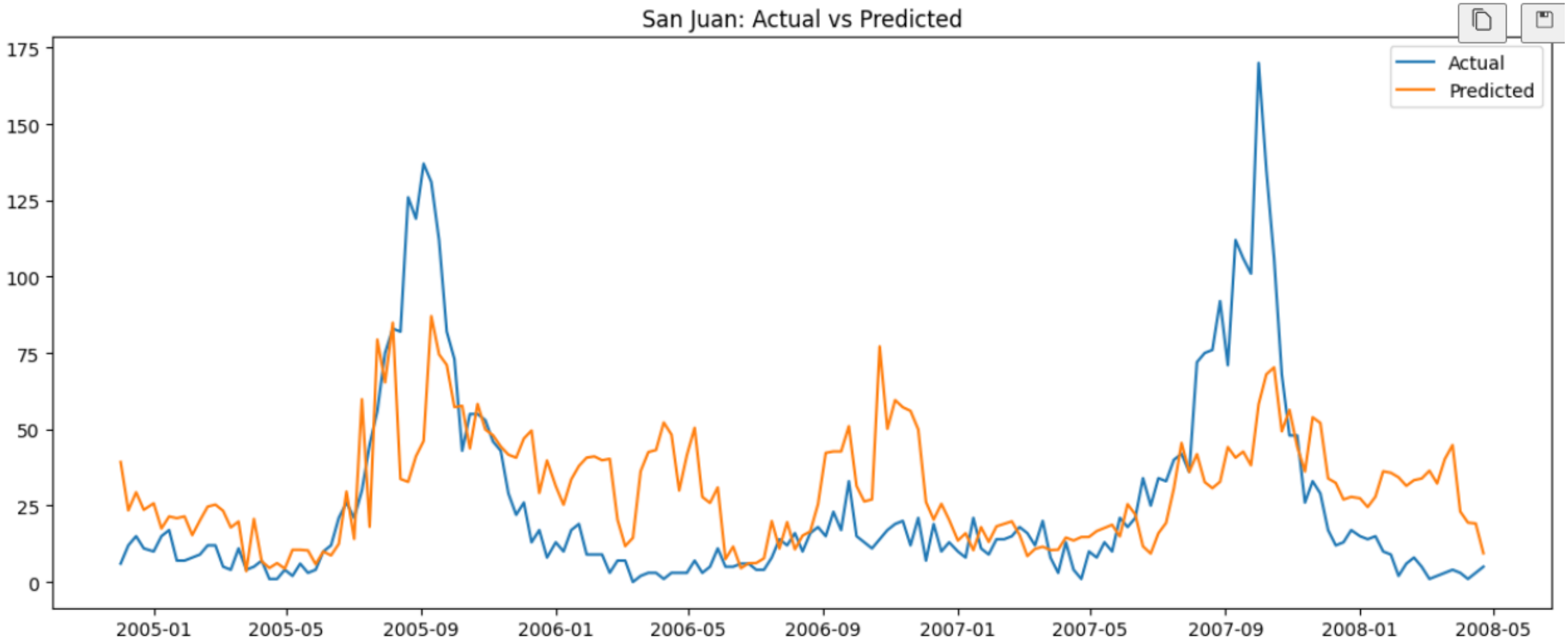


Feature Explosion: Expanded set includes Skewness, Kurtosis, Deltas, and Multi-period Lags.

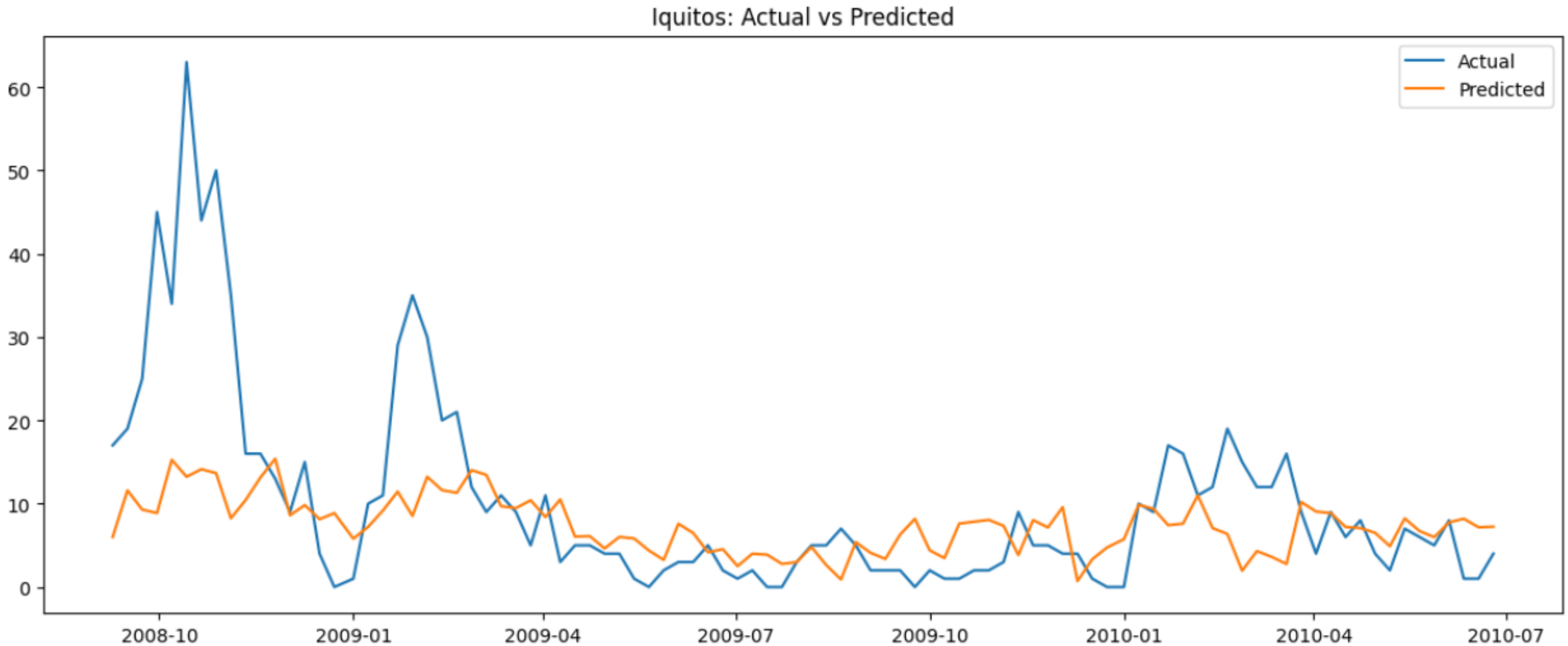
Greedy Selection: Iterative ranking by correlation. Retained only features that reduced Validation MAE (Top 300 capped).

Handling Outbreaks: Targeted oversampling of high-incidence peaks in training data.

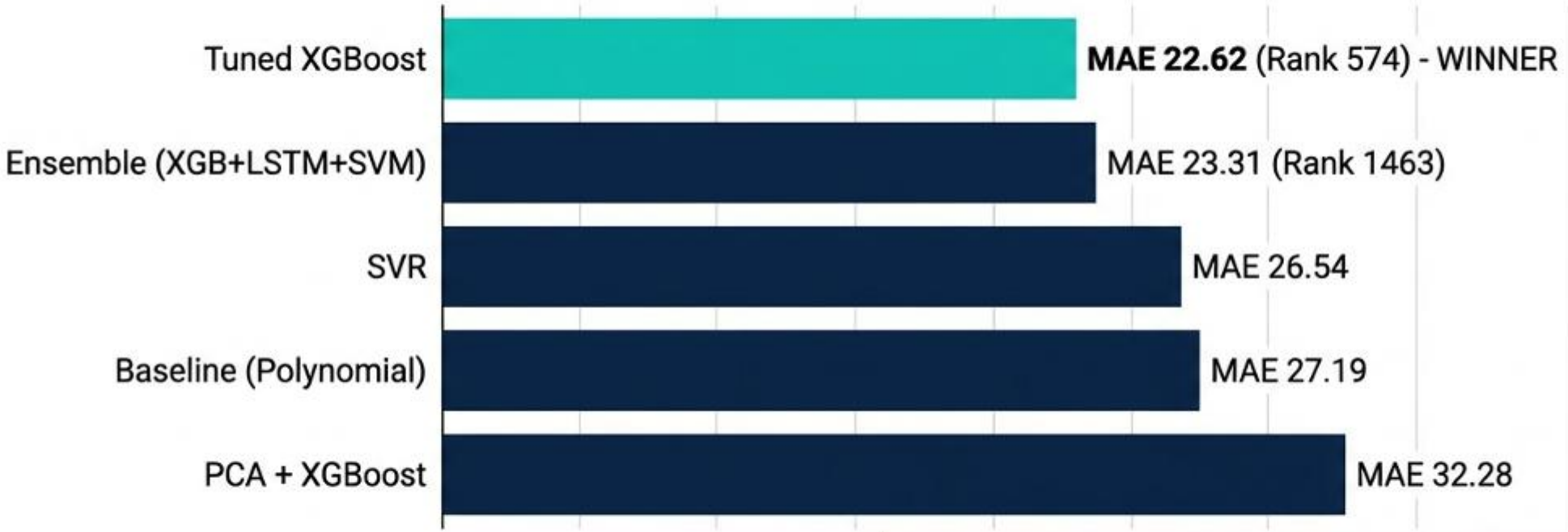
Visualizing the Model Fit: Actual vs. Predicted Cases



Visualizing the Model Fit: Actual vs. Predicted Cases



Leaderboard Comparison: Simplicity vs. Complexity

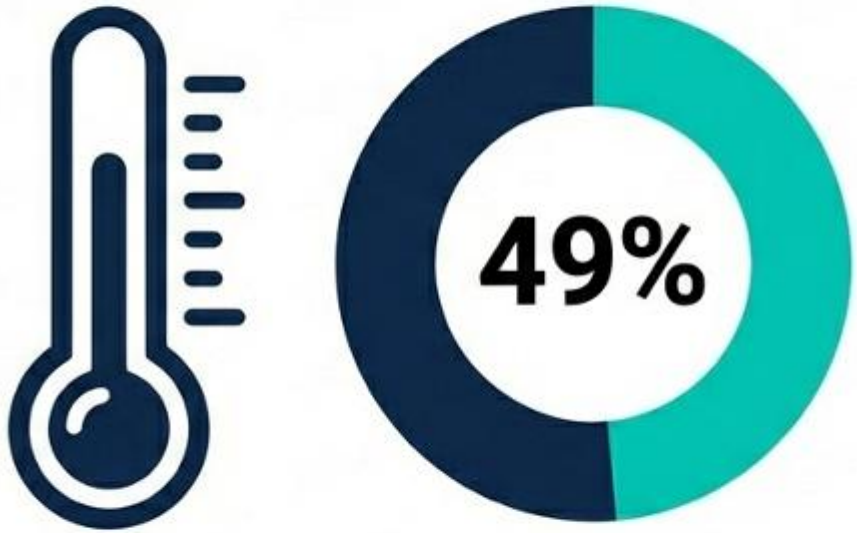


Key Takeaway:

Linear dimensionality reduction (PCA) destroyed signal. Feature engineering on a single strong learner (XGBoost) beat the complex ensemble.

Feature Importance: What Drives the Outbreaks?

San Juan (Urban/Coastal)



Temperature Driven

Key Feature: Diurnal Temperature Range (49.15%)

Interpretation: Daily temp fluctuations trigger outbreaks.

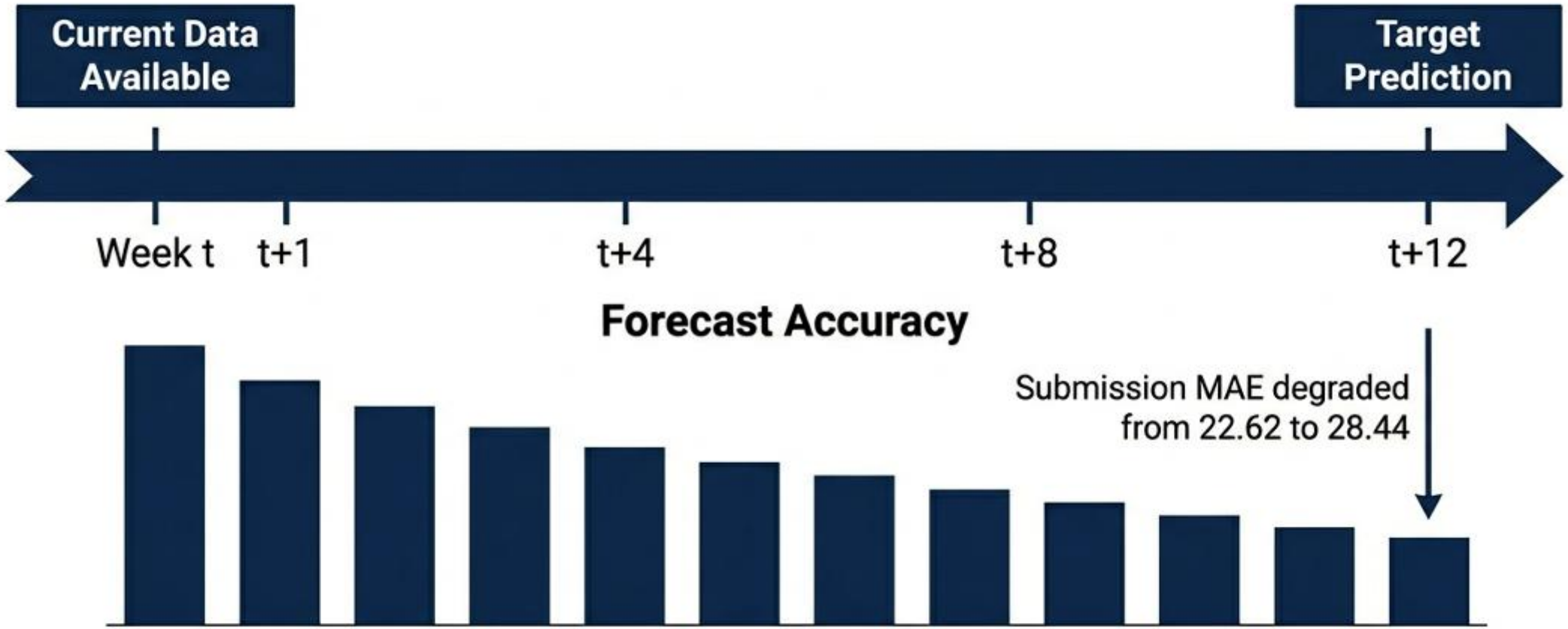
Iquitos (Rainforest)

Vegetation & Rain Driven

Key Feature: NDVI (Vegetation Index) (33.18%)

Interpretation: Ecologically sensitive model reflecting Amazonian location.

Task 2: Ahead-of-Time Forecasting



The Challenge: Predicting n weeks in advance. Accuracy naturally degrades as the horizon extends, forcing a trade-off between early warning and precision.

Optimal Prediction Horizons by City

San Juan: Long-Term Trends

Optimal Horizon: **t + 12 Weeks**

Validation MAE: 10.12

Analysis: Strong seasonal temperature dependence allows for long-range planning.

Iquitos: Immediate Volatility

Optimal Horizon: **t + 2 Weeks**

Validation MAE: 7.39

Analysis: High volatility in the rainforest makes long-term prediction unreliable. Best for immediate response.

Navigation

- Home
- About
- Problem description
- Official rules
- Leaderboard
- Discussion (56)
- Data download
- **Submissions** (23)
- Share your work
- Team

Submissions

- To help you track your progress during the competition, each submission is scored against publicly available test data to give a "public score".
- The primary evaluation metric is Mean Absolute Error. **Show more**.

Best score	Current rank	Submissions used
22.6202	#576	3 of 3

No submissions remaining

You have **0 of 3** submissions left today. Your next submission can be on Jan. 15, 2026 UTC.