Applied Machine Learning Final Project

Grace Myers
Cornell University
New York, NY 10044
gm586@cornell.edu

Aneesha Tamang
Cornell University
New York, NY 10044
at972@cornell.edu

Kelly Wang
Cornell University
New York, NY 10044
hw796@cornell.edu

Abstract

In this study, we participate in the DengAI: Predicting Disease Spread competition, which focuses on forecasting the number of dengue cases in two Latin American cities. Through an iterative exploration of model architectures, we determined that the Random Forest Regressor outperforms other approaches, delivering the best predictive accuracy. This work contributes to a growing body of research aimed at using machine learning for effective disease surveillance and prevention.

1 Motivation

We are participating in the DengAI: Predicting Disease Spread benchmarking competition, hosted by DrivenData. Dengue fever affects over 390 million people annually, with significant socioeconomic impacts. Predicting its spread is crucial for resource allocation in public health. This benchmarking project uses environmental data to identify effective machine learning models, addressing a real-world challenge where precise forecasting can save lives.

For this project, our goal is to identify the most effective machine learning models and methods to predict the number of cases of dengue fever in two Latin American cities - San Juan, Puerto Rico and Iquitos, Peru - using environmental data. We are leveraging real-world environmental data from these locations provided by DrivenData, such as precipitation (mm) and humidity, to build accurate predictive models. In doing so, our goal is to improve our understanding of dengue fever and its environmental triggers, potentially forming public health strategies to combat dengue and other similar diseases. After exploring numerous models during the earlier portion of the semester, for our final report, we centered our experiments on the Random Forest Regressor due to its superior performance in comparison to other models experimented with.

2 Context

Dengue fever is a mosquito-borne viral illness. As such, its transmission is highly dependent on environmental factors that impact the prevalence of mosquitoes. Therefore, it was critical to ensure that the raw environmental data we had was transformed to enhance its interpretability and robustness.

Preprint. Under review.

2.1 Feature engineering

We applied feature engineering to data for all models. To maximize the predictive power of our models, we engineered features from raw environmental data.
We performed the following feature engineering tasks:

- Imputation of missing values with the mean or using linear interpolation
- Encoded categorical variables (such as city identifiers)
- Standardized numerical features to ensure consistency

2.2 Baseline models

In the initial stages of the project, we experimented with several models, including Linear Regression, Bayesian Ridge Regression, Support Vector Machines (SVM), and Random Forest Regression.

- Linear regression, which served as a simple baseline, achieved a Mean Absolute Error (MAE) of 31.03 cases, highlighting its inability to capture the complex, nonlinear relationships inherent in dengue fever transmission.
- Bayesian ridge regression offered improved interpretability and regularization, but had a slightly higher MAE of 32.80 cases, indicating limitations in accounting for the variability in dengue cases.
- Support Vector Machines, leveraging a support vector regression approach to model nonlinear dependencies, yielded an MAE of 31.24 cases but struggled with the high dimensionality and temporal dependencies of the data.
- Initial implementations of random forest regression yielded an MAE of 26.55, demonstrating its ability to handle complex, nonlinear relationships more effectively than the previous models. Unlike linear models, Random Forest operates by constructing multiple decision trees during training and aggregating their predictions, which allows it to capture interactions between features and non-linear dependencies. Its ensemble nature also made it robust to overfitting and well suited for generalizing to unseen data.

Random Forest Regression emerged as the superior model due to its ability to capture complex, nonlinear relationships inherent in dengue transmission, which simpler models like Linear Regression and Bayesian Ridge Regression failed to do. Unlike Support Vector Machines, which struggled with high dimensionality and temporal dependencies, Random Forest effectively managed these challenges through its ensemble approach, constructing multiple decision trees and aggregating their predictions. This allowed it to model interactions between environmental features and provide robust predictions with a significantly lower MAE of 26.55 compared to the other models. Additionally, its built-in feature importance scores offered valuable insights into the relative contributions of factors such as precipitation, humidity, and temperature. These strengths demonstrated Random Forest's potential as the most effective model, guiding our decision to prioritize it for further experimentation and optimization for the remainder of the project.

3 Method

We employed a Random Forest Regressor to predict dengue fever cases, configured with 100 decision trees, trained on an 80-20 split of the dataset. This model was chosen for its ability to handle complex, non-linear interactions and its robustness to overfitting.

3.1 Successful approaches

3.1.1 Location-specific models

Given the geographic and environmental differences between San Juan, Puerto Rico, and Iquitos, Peru, we trained separate Random Forest models for each city. This approach allowed us to account for unique environmental factors that influence dengue transmission in each location. For instance, the warmer, wetter climate of San Juan presented distinct seasonal patterns that differed significantly from those in Iquitos.

3.1.2 Time-based features

In the San Juan dataset, we observed distinct seasonal patterns in dengue case counts. Dengue cases dropped on average from weeks 0 to 10, plateaued at low levels from weeks 11 to 29, and spiked from weeks 30 to 52. These patterns aligned with climatic changes: cooler and drier seasons at the start of the year, followed by a warm, rainy period associated with increased mosquito populations. To capture these dynamics, we introduced features corresponding to three seasons: fall (weeks 0–10), plateau (weeks 11–29), and spike (weeks 30–52). These seasonal features helped the model more effectively link climatic conditions to case trends. Iquitos seemed to have the typical four seasons (fall, winter, spring, summer), with case numbers closely aligning with these seasons. Therefore we added features corresponding to these seasons to the Iquitos data.

3.1.3 Feature scaling and interaction terms

Temperature readings from the dataset were provided in mixed units (Celsius and Kelvin). To address this, we scaled and standardized all temperature features using the `MinMaxScaler` from `sklearn`, ensuring consistency and comparability. Additionally, we added interaction terms to capture complex relationships between environmental variables. For example, interactions between temperature, precipitation, and humidity were modeled. These changes reduced the Mean Absolute Error (MAE) from 24.6 to 24.0 and improved our competition ranking from 996 to 803.

3.1.4 Lag features

We added lagged versions of key features (e.g., precipitation, temperature, humidity) to capture temporal dependencies in the data. Missing values in the lagged features were imputed to maintain data integrity.

After the above pre-processing was performed on the data, we dropped unnecessary columns. We only kept columns relating to temperature, humidity, precipitation. Additionally, there were many redundant columns (e.g., there were multiple columns for temperature) that were dropped.

3.2 Unsuccessful Approaches

3.2.1 Climatic aggregates

Initially, we experimented with lagged and rolling climatic aggregates, such as rolling averages and lagged statistics for key variables. However, this approach increased the MAE from 26.55 to 27.6. Even after limiting lagged features to four weeks and rolling windows to four and eight weeks, the MAE only marginally improved to 26.45. As a result, we ultimately decided not to include climatic aggregates in our final implementation. This decision was informed by several challenges: the addition of numerous variables increased model complexity, leading to potential overfitting; some rolling and lagged features introduced redundant information, inflating dimensionality unnecessarily; derived aggregates amplified noise and inconsistencies from the original measurements; and the temporal resolution of these aggregates may not have effectively aligned with the disease dynamics or captured critical environmental triggers.

3.2.2 Peak boosting

We attempted to integrate peak boosting on our predictions for both cities in order to emphasize the sharp peaks that are characteristic of disease outbreaks. Since the competition's performance metric was the MAE, peak boosting could potentially improve our performance if our model underestimated the true peaks in case numbers. MAE penalizes the absolute difference between prediction and true values and bringing our prediction peaks closer to the true peaks would improve the MAE. However, once implemented, peak boosting actually made our performance worse, indicating that our model was not underestimating the peaks in case numbers. As a result, we did not include it in our final implementation.

4 Setup

4.1 Datasets

The dataset provided by DrivenData includes environmental and dengue case data from two Latin American cities, San Juan, Puerto Rico, and Iquitos, Peru. Key features include temperature, precipitation, humidity, and dew point. The data spans multiple years and is accompanied by weekly dengue case counts for each city, which serve as the target variable. To address missing values, we used mean imputation and linear interpolation.

4.2 Data splits

The data for each city was split into training and testing sets using an 80-20 split. We opted for city-specific splits to account for the distinct environmental factors affecting dengue transmission in each location. Additionally, we created validation subsets within the training data to fine-tune hyperparameters.

4.3 Feature engineering

We performed extensive feature engineering, including:

- Scaling temperature data using `MinMaxScaler` to ensure consistency across mixed units (Celsius and Kelvin).
- Encoding categorical variables such as city identifiers.
- Adding seasonal features based on observed dengue case trends in San Juan (fall, plateau, and spike seasons).
- Creating interaction terms between features like temperature, precipitation, and humidity.
- Adding lag features for key climatic variables to capture temporal dependencies.

4.4 Metrics

We used Mean Absolute Error (MAE) as the primary metric to evaluate model performance, consistent with the competition's requirements. This metric was chosen for its interpretability and ability to measure prediction accuracy effectively.

4.5 Hyperparameter selection

Hyperparameters for the Random Forest Regressor were tuned using grid search. The parameters explored included: `n_estimators`: [100, 200, 300] (number of decision trees), `max_depth`: [5,10,20,35,50], `max_features`: [2,5,'auto'], and `min_samples_leaf` and `min_samples_split`: [2,3,4].

4.6 Experimental configuration

We trained separate Random Forest models for each city to account for geographic and climatic differences. Each model was trained using 80% of the city-specific data and tested on the remaining 20%. Training was conducted using Python's `scikit-learn` library, leveraging its `RandomForestRegressor` implementation.

4.7 Reproducibility

All experiments were conducted in a Python Jupyter Notebook environment with the following setup:

- Libraries used: `scikit-learn`, `pandas`, `numpy`, `matplotlib`, and `seaborn`.
- Hardware: Experiments were run on a standard laptop with an Intel Core i7 processor and 16GB of RAM.

- Random seeds: To ensure reproducibility, random seeds were fixed for all experiments.

The full implementation, including data preprocessing scripts, feature engineering steps, and model training configurations, is available in our GitHub repository: `https://github.com/aneeshat01/cs5785_project.git`. Readers are encouraged to explore the repository for additional details and to replicate our experiments.

5 Outcomes and results

5.1 Baseline results

- Linear Regression: MAE = 31.03.
- Bayesian Ridge Regression: MAE = 32.80.
- Support Vector Machines: MAE = 31.24.
- Random Forest: MAE = 26.55, Benchmark Place: 3700.

The baseline results highlight the limitations of simpler models like Linear Regression and Bayesian Ridge Regression in capturing the complex, non-linear relationships inherent in dengue transmission. Support Vector Machines showed slight improvement but struggled with the high dimensionality and temporal dependencies in the data. Random Forest Regression initially outperformed these models with an MAE of 26.55 and achieved a benchmark ranking of 3700, making it the most promising candidate for further optimization.

5.2 Random forest results

After iterative feature engineering and model tuning, the performance of the Random Forest model improved significantly. Key results include:

- Incorporating interaction terms reduced the MAE from 26.55 to 24.6.
- Adding lag features and seasonal indicators further reduced the MAE to 24.0.
- These improvements elevated our benchmark ranking from 3700 to 803 in the competition leaderboard.

The model's enhanced accuracy reflects the importance of domain-specific features, such as seasonal patterns and feature interactions, in predicting dengue cases.

5.3 Analysis

The Random Forest model's success lies in its ability to capture non-linear relationships and interactions between key environmental variables. Feature importance analysis revealed that precipitation, temperature, and humidity were the most influential predictors, emphasizing their role in dengue transmission. Interaction terms and lag features further allowed the model to account for temporal dependencies and geographic variability, leading to substantial improvements in performance.

However, some limitations remain. The model's performance differed between San Juan and Iquitos, likely due to differences in data quality, underlying climatic patterns, and sample sizes. Additionally, while the MAE decreased, further experiments with alternative algorithms or additional features could yield even better results.

5.4 Conclusion

This study demonstrates that a well-tuned Random Forest model, coupled with thoughtful feature engineering, can effectively predict dengue fever cases in two distinct cities. By leveraging interaction terms, temporal features, and domain-specific insights, we significantly reduced prediction error and improved our competition ranking from 3700 to 803. Future work could explore advanced models, such as Gradient Boosting Machines or Neural Networks, to further enhance predictive accuracy and generalizability. Our approach highlights the potential of machine learning in addressing public health challenges.

A Appendix / supplemental material

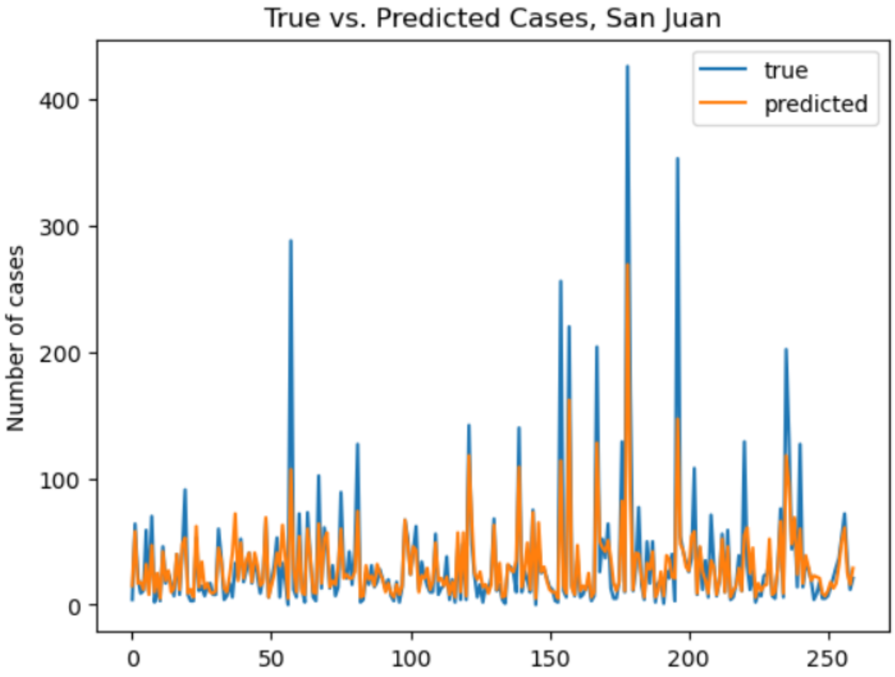


Figure 1: San Juan predicted vs. true case numbers

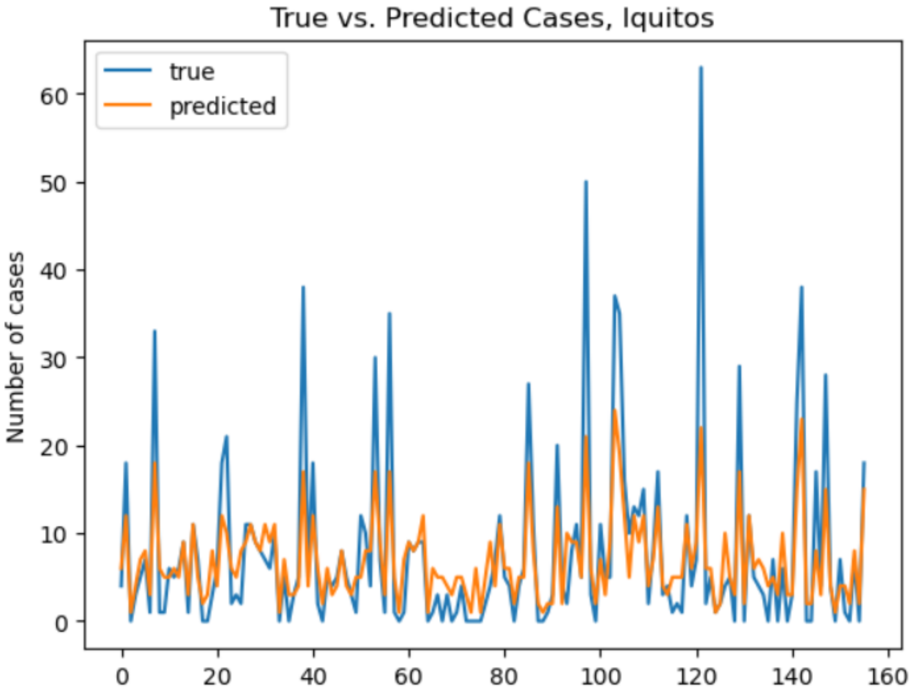


Figure 2: Iquitos predicted vs. true case numbers

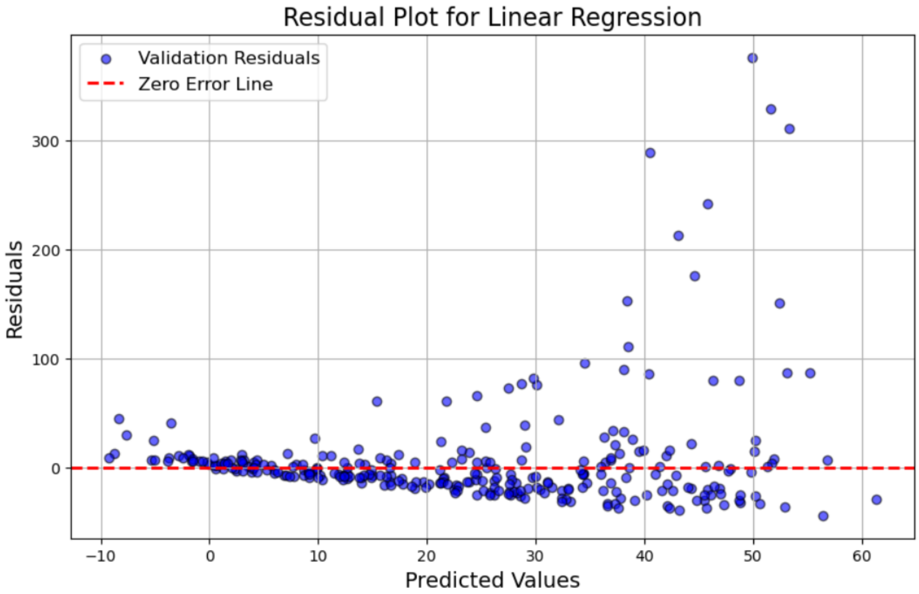
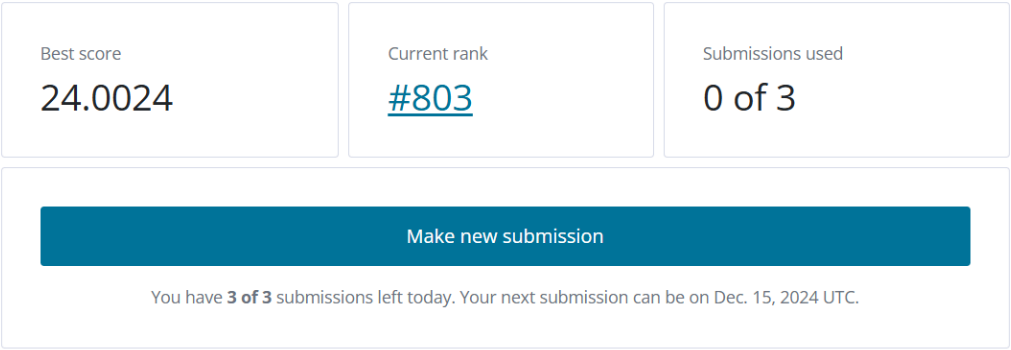


Figure 3: Residual plot for linear regression shows that it is good at capturing linear relationships in the data, but not non-linear ones



Best score
24.0024

Current rank
#803

Submissions used
0 of 3

You have **3 of 3** submissions left today. Your next submission can be on Dec. 15, 2024 UTC.

Figure 4: Best DrivenData score