MA 679
Final Project write-up
Guangyu Ding
2025/5/5

Introduction

Dengue fever is a viral disease transmitted by mosquitoes, with seasonal and environmental factors. A reliable short-term forecast of the number of cases per week can provide information for vector control, hospital preparedness, and public warnings. In this project, environmental data collected by various U.S. Federal Government agencies —— the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the US will be used to predict weekly dengue fever cases in San Juan and Lquitos. Some mechanical learning, feature learning, and deep learning models will be considered and ran to make the predictions.

Methods

The key features of the data includes multiple temperature readings like average/ min/max air temperature from ground stations and reanalysis data, several precipitation metrics including station rainfall and reanalysis precipitation estimates, humidity indicators such as relative humidity, dew point and specific humidity and vegetation indices NDVI values from four quadrants around each city. Besides, the dataset provides the total cases of dengue for each week in the training data. These variables allow the models to learn how weather and climate patterns relate to dengue incidence. One thing that need to be noticed is that compared to Lquitos,

San Juan historically has a higher incidence of dengue with some severe outbreaks reaching dozens or hundreds of cases in a week.

Since San Juan and Iquitos have different outbreak patterns and ranges of values, we split the data by city and treated each city's dataset independently. All subsequent processing (feature engineering, modeling) was done separately for each city's time series. To prepare the data for modeling, we performed several preprocessing steps to clean and structure the dataset. First we merged features and target. We combined the feature dataset with the dengue case counts (labels) by matching on city, year, and week. This generated a single training frame for each city containing both input features and the target total cases. Then we need to handle the missing values. Some environmental features, especially NDVI readings had missing values. Rather than discard data, we utilized forward-fill imputation, carrying the last known value forward in time for each city. It is useful here since climate measurements change gradually week to week. Some other EDA steps may includes data type conversions. We converted the date field to a datetime type, and from it we extracted components like month as needed. This made it easier to engineer time-based features like the seasonal indicators.

After the data preprocessing, we explored several modeling approaches. We considered both traditional time-series models and machine learning methods, and select high-performing predictors for each city. The following models were developed and tuned separately for San Juan and Iquitos:Temporal Convolutional Network (TCN) and XGBoost (Gradient Boosting Trees) because TCN dilated causal convolutions

efficiently model long-range and short-term dependencies in sequential data without recurrence or attention mechanisms, and industry-standard for tabular data, handles heterogenous feature sets and is robust to noise when we use XGBoost.

Some Train split strategy of the TCN is that we sort the San Juan (and Iquitos) weekly data by time, then take the first 80 % of the labeled weeks for training and the last 20 % for validation. This preserves the time order and prevents “peeking” into future weeks. As for XGBoost, during hyperparameter tuning we used sklearn.model_selection.TimeSeriesSplit, which repeatedly splits the training portion into expanding training windows and adjacent validation windows. Besides, there were also several Feature Learning strategies applied when running the XGBoost, like Lag features, Rolling Mean of Humidity and Seasonality Encoding to make the result of the prediction better.

Result & Discussion

Finally, two predicted files were generated, containing weekly total case forecasts for two cities from 2008 to 2013. Then they were uploaded to the competition webpage of drivenDATA. The final score for the XGboost model is 27.5721, and the score for the TCN model is 26.4663 (the CSV file and score screenshot of the specific results are on GitHub).

The metric used for this competition is mean absolute error. The absolute error is calculated for each label in the submission and then averaged across the labels. Here is the formula:

$$MAE = 1/n \sum_{\{i=1\}}^{\{n\}} |f_i - y_i|$$

From the results, it can be seen that the TCN model Learns temporal filters that adapt to outbreak onsets and seasonal cycles, yielding tighter error bounds.

But for XGBoost, it tends to regress to the mean in weeks without strong covariate deviations. Some possible limitations may be the lack of micro-climate or vector-surveillance data may hamper predictions during sudden spikes.

Conclusion

This competition on the driveDATA platform gave us an opportunity to practice machine learning models in real cases. We established TCN and XGBoost models to predict the number of cases of Dengue fever in the San Juan and Iquitos region, which helped us gain a better understanding of concepts such as mechanical learning, deep learning, and feature learning. Maybe some future steps are considering additional covariates like Integrate precipitation forecasts, vector index data, and mobility patterns etc, using some methods like DeepAR or quantile for uncertainty intervals to do probabilistic forecasting, and explore transformer models for multi-week predictions for extending horizons.

Source

DrivenData. (2025). DengAI: Predicting Disease Spread. *https://www.drivendata.org*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in Python*