WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Predicting number of Dengue Cases

Prepared by: Sohit Nayak

Teamate: Vidaan Shankar

Date: 19th Dec, 2022

Business Problem:

The Government of India is concerned with predicting dengue cases in India so that they can set up dedicated dengue clinics in the tropical states of India. Dengue is becoming a pertinent issue in Southern India and the government wants to be prepared for the next outbreak. They want to be able to predict the number of cases in the coming year at the city level so that they have a fair idea of the number of dedicated clinics and doctors to instate in the cities.

Problem Statement:

A mosquito-borne illness called dengue fever is present in tropical and subtropical regions of the world. The flu-like symptoms of fever, rash, and muscle and joint discomfort are present in mild instances. In extreme circumstances, dengue fever may result in fatal bleeding, low blood pressure, and other complications. Since dengue is spread by mosquitoes, climate factors like **temperature** and **precipitation** have an impact on how the disease spreads. Despite the complexity of the connection to climate, an increasing number of scientists contend that global distributional alterations brought on by climate change are likely to have a considerable impact on public health.

Our goal for the competition was to predict the total cases for each case in the test set. There are two distinct cities, San Juan (Puerto Rico) and Iquitos (Peru), with test data for each city spanning 5 and 3 years respectively.

A bit about the Data:

This problem is a regression problem since we are dealing with the number of cases. We received a training and testing dataset for this project. Two files, features and labels, made up the training set. We used 1456 rows and 24 features for the training dataset. The environmental parameters, such as the vegetation index, precipitation, air indices, relative and specific humidity, and diurnal temperature, as well as the week of the year, the city, and the year, made up most of the features. Total number of cases served as the prediction label.

Our Approach:

- We started with data cleaning,
- Exploratory data analysis (univariate and bivariate, outlier analysis, correlation plots),
- feature engineering mostly based on research (created new features),
- Research into other similar studies, and
- Applying predictive models in our manipulated data.

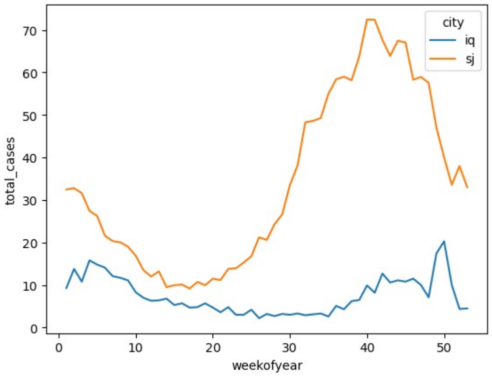
Research Findings:

We reviewed previous comparable dengue-related research projects. We discovered several fascinating publications that connected the transmission of dengue to various crucial environmental elements, including temperature, rainfall, relative humidity, precipitation, wind speed, and vegetation cover. Next, we looked at the Aedes mosquito's life cycle. The information below was pertinent to the study:

- Adult female mosquitoes lay their eggs on the inner, wet walls of containers.
- The larvae then feed on micro-organisms in the water.

A strong conclusion that can be drawn from the above facts is that the mosquitoes need moist weather conditions for their larvae to grow. Hence, **precipitation and relative humidity** play an important role.

One more interesting finding was that the two cities in the dataset had very distinct environmental factors at play. Hence, one challenge was to come up with feature interaction which influence our label. Upon looking up the wet seasons of both the countries in question, we found that the monsoon season in Peru is from November to March while Puerto Rico's monsoon season lasts from April to November and the driest season is from December to March. This fact is supported by the following graph plotting the label with week of year. Although total cases for Iqitos is lower but the pattern is very clear here with regards to our theory mentioned above. An adult mosquito typically develops from an egg within 8 to 10 days. With the monsoon season comes an increase in mosquito population. Therefore, dengue cases are at their lowest in April when the monsoon arrives and at their highest in November. This hypothesis also makes sense for the other city.

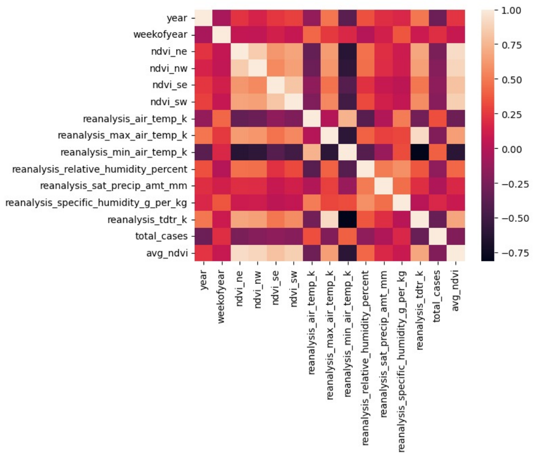


Data Cleaning:

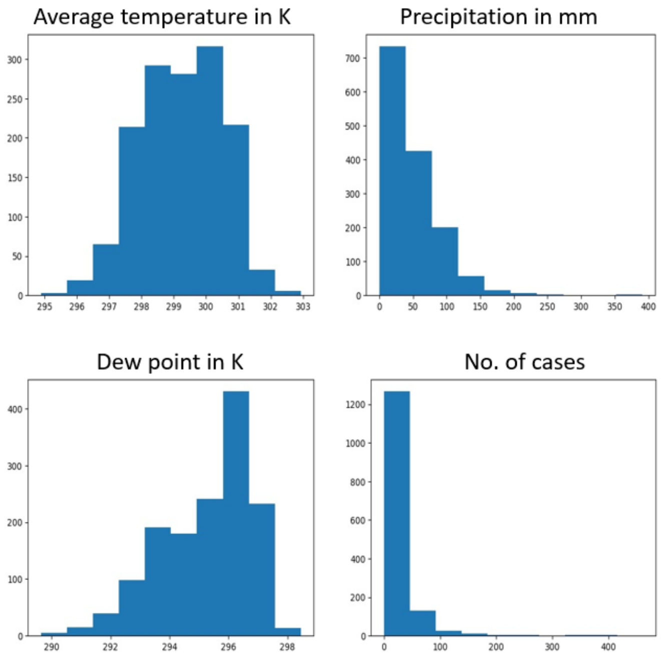
The cleaning was mostly focused on treating the null values. The highest number of null values were present in the ndvi (Vegetation Index) features. There were four features for ndvi features (namely, ne, nw, se and sw). We removed the rows where we had null values for all four features. And then removed the rows having null values using dropna() function.

Exploratory Data Analysis:

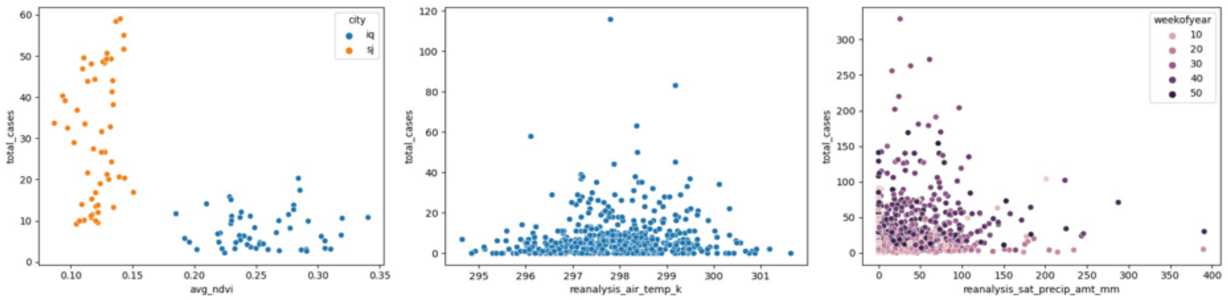
A correlation plot was used to understand the relationship between features. This was used to identify highly correlated features and we got rid of one set of features with correlation more than 0.78.



A univariate analysis was conducted to explore the distribution of our feature variables. Attached below are some illustrations.



The distinctiveness of the environmental features between two features can be clearly seen in the below graphs.



The first graph demonstrates how the two regions vary in terms of the amount of vegetation and the overall number of cases. The second plot demonstrates that at a particular range of air temperature, the cases are high.

Feature Engineering:

Distance from Monsoon:

As the data contained factors pertaining to two cities from different countries we considered the distance from the peak of monsoon to the weeks of the year to bring the two cities onto a common baseline.

Vegetation Index:

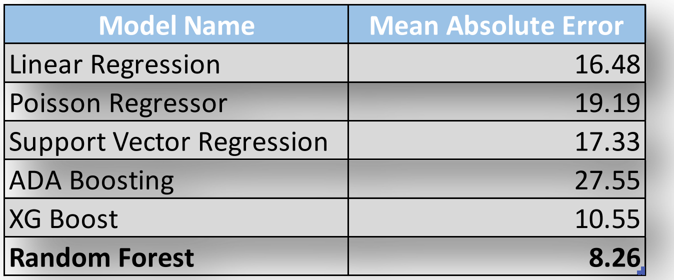
Since there were separate ndvi values for each corner of the city we considered the average of those values to be representative of the city because the values for each corner were almost the same. Some rows had missing ndvi values, so we averaged it based on the available values.

All the data were scaled using the StandardScaler() so that we normalized the actual numerical figures in each feature column.

Models and results:

Our loss function provided in the competition was **Mean Absolute Error (MAE)**.

We ran the below models to evaluate the fit of the model to the training dataset. The **Mean Absolute Error** is calculated for each of these models as it is the governing criteria for the study/competition.



Model Name	Mean Absolute Error
Linear Regression	16.48
Poisson Regressor	19.19
Support Vector Regression	17.33
ADA Boosting	27.55
XG Boost	10.55
Random Forest	**8.26**

Upon performing parameter tuning, the best fit model is the Random Forest with **220 trees** and with **minimum sample split size of 5**.