Disease Spread

Predicting the Spread of Dengue using Machine Learning

Capstone II Project: Proposal

Springboard Data Science Career Track

Mark Rojas

January 2019

PROBLEM

Dengue fever is a disease caused by any one of four closely related dengue viruses (DENV 1, DENV 2, DENV 3, or DENV 4). The viruses are transmitted to humans by the bite of an infected *Aedes aegypti* or *Aedes albopictus* mosquito. In mild cases, symptoms are similar to the flu, however, in severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death. As many as 400 million people are infected yearly with an estimated 500,000 people hospitalized with severe dengue, also known as dengue hemorrhagic fever (DHF)[1].

Dengue cannot be spread directly from person to person. Outbreaks of dengue occur primarily in areas where *Ae. aegypti* or *Ae. albopictus* mosquitoes live. This includes most tropical urban areas of the world. Dengue viruses may also be introduced into areas by travelers who become infected while visiting other areas of the tropics where dengue commonly exists. Because there is no vaccine for preventing dengue, the best preventive measure for residents living in areas infested with *Ae. aegypti* is to eliminate the places where the mosquito lays her eggs.

Since dengue is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as **temperature** and **precipitation**. Although the relationship to climate is complex, **a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide**.

More recently, dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the **nearly half-billion cases per year** are occurring in **Latin America**. And the problem is only growing worse. Over the past 50 years, dengue cases have increased thirtyfold, and currently half of the world's population, some 3.9 billion people, are at risk.

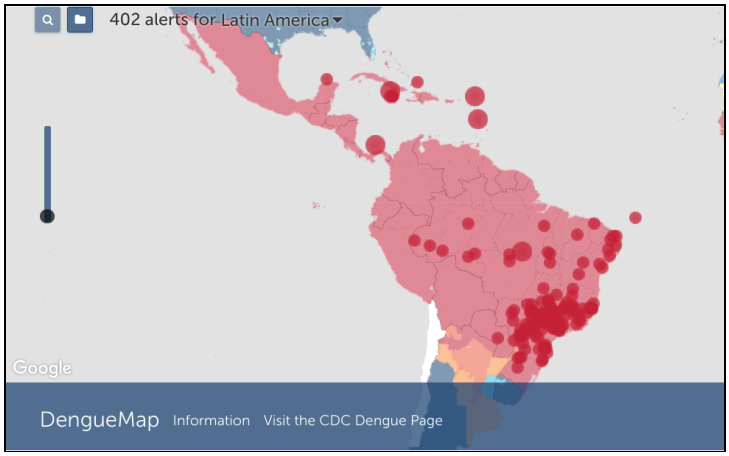

Figure: Dengue Notifications/Alerts for Latin America
402 alerts for the month of January 2019
https://www.healthmap.org/outbreaksnearme/

1. Dengue Fever Outbreaks

GOAL

Use environmental data collected by various U.S. Federal Government agencies to **predict the number of dengue fever cases reported each week in San Juan, Puerto Rico and Iquitos, Peru** based on environmental variables describing changes in temperature, precipitation, vegetation, and more.

CLIENT

Accurate dengue predictions would help several departments in the U.S. Federal Government (Department of Health and Human Services, Department of Defense, and Department of Commerce), public health workers and people around the world to take steps to reduce the impact of these epidemics.

DATA

The data for this competition comes from multiple sources aimed at supporting the **Predict the Next Pandemic Initiative**.

Dengue Surveillance Data from:

- Centers for Disease Control and Prevention
- Department of Defense's Naval Medical Research Unit 6
- Armed Forces Health Surveillance Center
- Peruvian government and U.S. universities

Environmental and Climate Data from:

- National Oceanic and Atmospheric Administration in the U.S. Department of Commerce

File	Description
Training Data Features	The features for the training dataset.
Training Data Labels	The number of dengue cases for each row in the training dataset.
Test Data Features	The features for the testing dataset

Table: Data available from DrivenData.org

https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/data/

For a complete list of Features, click here.

METHODOLOGY

To predict the number of Dengue cases reported each week, I plan to use supervised learning for classification and forecasting on time series data. I will review multiple time series forecasting methods such as Autoregressive Moving Average, Autoregressive Integrated Moving Average, and Seasonal ARIMA. In addition, I hope to apply neural networks and compare model performance.

The performance will be evaluated according to the **Mean Absolute Error (MAE)**, such that given any test data set, the MAE of the model will refer to the mean of the absolute values of each prediction error on all instances of the test data set. *Note, the prediction error is the difference between the actual value and the predicted value for that instance*.

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

DELIVERABLES

- Milestone Report
- Final Report
- Collection of Resources used
- Presentation available on Slide Deck
- Code available through GitHub