Capstone Project
Machine Learning Engineer Nanodegree

Giulia Martielli
February 22nd, 2021

Predicting Dengue Disease Spread

Figure 1 - Image from unsplash.com

Domain Background

In 2019 I had the chance of visiting an astonishing country, Sri Lanka. During our stay there, although we were already aware of it and had taken the necessary precautions, we had a few conversations with local people about the phenomenon of Dengue and the risk that one can run every day by being bitten by a particular species of mosquitoes, *Aedes* mosquitos, that carry this disease. From there my curiosity about this phenomenon grew and I began to wonder if it was possible in some way to use statistical and / or machine learning techniques to analyze and predict it. This project of mine, therefore, is based on a competition held by the site DrivenData.org, called "DengAI: Predicting Disease Spread", which focuses on finding a way of predicting the next dengue fever local epidemic in *San Juan*, Puerto Rico and *Iquitos*, Peru.

Problem Statement

As stated on the DrivenData.org competition page, in the past years dengue fever (DF) has been most prevalent in Southeast Asia and the Pacific Islands. Nowadays, however, it is much more spread around the world and many of the nearly half billion cases per year are occurring in Latin America. According to Salles *et al.* (2018), the virus has become a major threat to American human life, reaching approximately 23 million cases from 1980 to 2017.

DF is a systemic and dynamic infection with a broad clinical spectrum that includes both serious and non-serious clinical manifestations. In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death (WHO,

2009). No effective therapy for dengue exists at the moment; treatment is purely symptomatic, requiring a high level of patient care. This is why predicting upcoming epidemics can be a valuable aid.

Since DF is carried by mosquitos, its transmission is higly correlated with climate variables, especially temperature and precipitation / humidity. The task of the competition is to *predict the number of dengue cases each week* (in each of the two locations – San Juan and Iquitos) based on all the environmental variables provided (temperature, precipitation, vegetation, and more).

Datasets and Inputs

The data for this project are downloaded directly from DrivenData.org (more information about the sources of the data can be found at this link). In addition to the number of cases in the two locations, the data include, as previously mentioned, information on temperature, precipitation, humidity, vegetation, and what time of the year the data was obtained. Below a detailed list and description of datasets provided and features.

Datasets:

- `dengue_labels_train.csv` (1,457 rows) containing four columns being `city` (`sj` for San Juan and `iq` for Iquitos), `year` (ranging from 1990 to 2008 for San Juan and from 2000 to 2010 for Iquitos), `weekofyear` (ranging from 1 to 53) and `total_cases` (the number of cases/week for each city);
- `dengue_features_train.csv` (1,457 rows) containing data for both San Juan (years ranging from 1990 to 2008) and Iquitos (years ranging from 2000 to 2010), plus all the relevant features, for a total of 24 columns;
- `dengue_features_test.csv` (417 rows) containing data for both San Juan (years ranging from 2008 to 2013) and Iquitos (years ranging from 2010 to 2013), plus all the relevant features, for a total of 24 columns; as reported on the DrivenData.org website, "the test set is a pure future *hold-out*, meaning the test data are sequential and non-overlapping with any of the training data".

Here is a list of all the features contained in the train and test dataset:

- City and date indicators
  - `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos
  - `week_start_date` – Date given in yyyy-mm-dd format

- NOAA's GHCN daily climate data weather station measurements
  - `station_max_temp_c` – Maximum temperature
  - `station_min_temp_c` – Minimum temperature
  - `station_avg_temp_c` – Average temperature
  - `station_precip_mm` – Total precipitation
  - `station_diur_temp_rng_c` – Diurnal temperature range

- City and date indicators
  - `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos
  - `week_start_date` – Date given in yyyy-mm-dd format

- PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)
  - `precipitation_amt_mm` – Total precipitation

- NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- `reanalysis_sat_precip_amt_mm` – Total precipitation
- `reanalysis_dew_point_temp_k` – Mean dew point temperature
- `reanalysis_air_temp_k` – Mean air temperature
- `reanalysis_relative_humidity_percent` – Mean relative humidity
- `reanalysis_specific_humidity_g_per_kg` – Mean specific humidity
- `reanalysis_precip_amt_kg_per_m2` – Total precipitation
- `reanalysis_max_air_temp_k` – Maximum air temperature
- `reanalysis_min_air_temp_k` – Minimum air temperature
- `reanalysis_avg_temp_k` – Average air temperature
- `reanalysis_tdtr_k` – Diurnal temperature range

- Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements
  - `ndvi_se` – Pixel southeast of city centroid
  - `ndvi_sw` – Pixel southwest of city centroid
  - `ndvi_ne` – Pixel northeast of city centroid
  - `ndvi_nw` – Pixel northwest of city centroid

Sample values for each feature can be found below.

Features	Sample Values
city	sj
year	1990
weekofyear	18
week_start_date	30/04/1990
ndvi_ne	0.1226
ndvi_nw	0.103725
ndvi_se	0.1984833
ndvi_sw	0.1776167
precipitation_amt_mm	12.42
reanalysis_air_temp_k	297.572857143
reanalysis_avg_temp_k	297.742857143
reanalysis_dew_point_temp_k	292.414285714
reanalysis_max_air_temp_k	299.8
reanalysis_min_air_temp_k	295.9
reanalysis_precip_amt_kg_per_m2	32.0
reanalysis_relative_humidity_percent	733.657142857
reanalysis_sat_precip_amt_mm	12.42
reanalysis_specific_humidity_g_per_kg	140.128571429
reanalysis_tdtr_k	262.857142857
station_avg_temp_c	254.428571429
station_diur_temp_rng_c	6.9
station_max_temp_c	29.4
station_min_temp_c	20.0
station_precip_mm	16.0

The goal is to predict the `total_cases` label for each triplet (`city`, `year`, `weekofyear`) in the test set. Another important consideration to report is that throughout the datasets, missing values have been filled as `NaNs`. After performing some exploratory analysis to the data I will evaluate whether the missing data might need to be filled / replaced through some sort of data imputation (mean value for continuous features – for instance) and whether there are features which might be removed from the dataset (if not significant or redundant), in order to simplify the analysis.

Solution Statement

The task of predicting the total cases of dengue for the next *X* weeks in the future is a *supervised* (time series, more specifically) problem, since we know that the target variable is a numeric discrete variable (number of cases). For this reason, every forecasting algorithm / model able to support features with different natures – such as the ones we do have for this specific project – as an input might be appropriate. Some examples range from classic statistical forecasting methods (Simple Exponential Smoothing, ARIMA etc.), to machine learning methods, such as Gradient Boosting Trees (XGBoost), and also to Recurrent Neural Network (such as Long Short-Term Memory – LSTM).

Some steps the final solution will include, among others, are:

- Data ingestion and Exploratory Data Analysis (EDA) section, to clearly understand the data before performing any sort of processing
- Pre-processing step (outlier detection, missing values filling, etc.) and data split among the two cities (if supported by the data, meaning that the data distributions of the two cities appear to be very different from each other)
- Feature engineering (if needed) and feature selection
- Trying different supervised learning models (as previously mentioned, there are quite a few possibilities, I would like to try at least two of them before choosing the best model)
- Evaluating the models using the chosen metric (more details can be found in the **Evaluation Metrics** section) and then choosing the best model
- Improve the model, if needed

I would like to point out that the solutions and steps proposed in this document may vary in the final project depending on the preliminary exploratory data analysis that will be performed, and do not represent the final solution that will be evaluated only and exclusively with the support of the data.

Benchmark Model

As a benchmark model I choose to consider the one reported onto the DrivenData.org website itself at this link. The model is a *Negative Binomial regression*, which is usually indicated when the data distribution suggests a much larger variance than the mean. The benchmark model has its flaws and has space for further improvements, as also stated on the website, since by taking a look at its actual vs predicted plot 1) the timing of the seasonality of the predictions has a mismatch with the actual results and 2) the predictions are relatively (too) consistent, they miss the spikes that represent the large outbreaks (hence the most problematic periods). This model, though, will represent a great starting point for comparison.

Evaluation Metrics

The evaluation metric suggested for the DrivenData.org competition is the *mean absolute error (**MAE**)*, which is a model evaluation metric usually used with regression models. "The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set (n). Each prediction error is the difference between the true value (y_i) and the predicted value for the instance ($\hat{y}_i$)" (Sammut C., 2011).

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - \hat{y}_i)}{n}$$

Other graphic evaluation criteria might be considered, such as the actual vs predicted values plot.

Project Design

The project design will follow in broad lines the steps outlined in the **Solution Statement** section, which are, in summary, data analysis, data preparation and features selection, model fitting, performance assessment and, if necessary, model improvement. Please refer to that section for more details.

Again, I would like to point out that the project design proposed in this document may vary in the final project, and does not represent the final solution that will be evaluated only and exclusively with the support of the data.

References

Administration, N. O. a. A., n.d. *Dengue Forecasting.* [Online]
Available at: http://dengueforecasting.noaa.gov/

Ali, S., n.d. *Unsplash.com.* [Online]
Available at: https://unsplash.com/photos/nZgpg4xYhjM

DrivenData, n.d. *DrivenData.* [Online]
Available at: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

Organization, W. H., 2009. *Dengue: guidelines for diagnosis, treatment, prevention and control.* [Online]
Available at: https://www.who.int/tdr/publications/documents/dengue-diagnosis.pdf

Prevention, C. f. D. C. a., n.d. *Dengue.* [Online]
Available at: https://www.cdc.gov/Dengue/

Salles, T. d. E. S.-G. T. d. A. E. e. a., 2018. History, epidemiology and diagnostics of dengue in the American and Brazilian contexts: a review.. *Parasites Vectors,* Issue 11, p. 264.

Sammut C., W. G., 2011. *Encyclopedia of Machine Learning.,* Boston, MA: s.n.