Capstone 1: Dengue Milestone Report

This project is a study of the effect climate data has on the number of cases of dengue in a week developed off of the data provided by several US departments. The data provided is a compilation of different survey systems recording different features of the climate and geographic surveys of two different cities: San Juan, Puerto Rico, and Iquitos, Peru. Both cities have experienced outbreaks, periods where the reported number of cases in a week is a significant deviation higher than the average number of reported weekly cases. The data spans across several years, with each dataset encompassing different year intervals, but the same features are present in each dataset. This project is an exploration of the interaction that the climate features have on the case numbers, to model the independent features in such a way to be able to get approximate forecasts of the case numbers for a testing set given the appropriate features with attention given to the performance of the models during the outbreak periods for each city. This project was inspired by a competition this coalition of government agencies held in 2013 to build a prediction system that helps with understanding how these features play a role in the case numbers and to be able to better prepare a response or strategy when a forecasted period indicates a spike in case numbers of dengue to provide adequate aid for the cities affected. With a successful prediction model, the response to outbreaks of dengue can prove critical to providing a proper response in aiding the civilians and societies afflicted with necessary actions to either help these cities recover from the widespread plague undercutting the health of these societies or provide guidance for steps and actions that can be put in places by these cities to reduce the number of cases, reduce the number of people exposed, implement strategies for mosquito population control, or build up the inventory of medication and supplies needed to treat those afflicted.

The dataset was obtained from the DengAI: Predicting Disease Spread competition hosted by DrivenData.org, which itself is a representation of the Dengue Forecasting competition hosted by the coalition of different US departments: Dept. of Health and Human Services, Dept. of Defense, Dept. of Commerce, and the Dept. of Homeland Security, with support from the Pandemic Prediction and Forecasting Science and Technology Interagency Working Group. The dataset provided is split into 3 CSV files, a training set of the feature data, a training set of the target variable total\_cases, and a testing set of the feature variables that is used to grade the performance of the submissions to this competition against a testing set of the total cases that is held by the hosts and not made public. This project will utilize the first two CSVs to model and predict since the other CSV doesn’t include the labels with which to compare the model predictions against. The features provided in the dataset includes several climate features including temperature, humidity, dew point, vegetation index, and precipitation variables from the different government agencies in the council that originally hosted this competition as well as some time data associated with when the sample was taken. Some of the data sources include a reanalysis of the same data variables provided by different sources though in different units of measure. There are weeks with missing values for some of these variables to a varying degree. Some preprocessing was necessary to be able to model appropriately.

The month that each sample fell in was extracted from the week\_start\_date feature and from there dummy variables were one hot encoded to obtain binary features rather than having one categorical feature since many models aren’t able to interpret categorical variables. The units of measure were aligned for the temperature features, which may not be a necessary step but it does align the units of similar features which would be easier to interpret. This was done before dealing with the missing values as it is an easier way to judge whether the imputation step was able to estimate a value that was sensical by comparing it to the related feature if that was not also missing. A more in-depth imputer was necessary to provide a more realistic pattern in the climate features rather than using a simple approach like filling with the average value of the feature in question. Other common imputation techniques such as forward filling the previously valid value would be fine for some features with only one or two weeks of missing data for that variable but the resulting pattern of variables that had a longer range of weeks with missing values would be a flat trend in climate features, which is both not realistic and could throw off the models ability to pick up on the variances of these features and potentially result in a worse fit and decrease in prediction accuracy. One other thing to take care of was the years that had entries for the 53rd week of the year, which doesn’t align with the commonly known calendar which only has 52 weeks in a year. After correcting the labeling and ordering of the samples, the basic preprocessing and data cleaning is done; more preprocessing will be performed later on a per-model basis to shape the data to a form that will work best with the model in the notebook.

Mosquitoes are the attack-vector for the dengue viruses, and their entomology is affected by a few factors such as the temperature in their local zone and the water sources available for the larvae to mature in. Thus, the precipitation and temperature measures in our dataset may prove to be useful to understand on a deeper level. We looked at the different climate readings between Iquitos and San Juan since there is not only a difference in the case patterns but of their climate patterns as well. The average air temperature between the two cities, we see a stronger amplitude delta for San Juan, Puerto Rico than in Iquitos, Peru, meaning we have a wider variation in average air temp in San Juan than what is observed in Iquitos. Looking at the distribution of this feature, we see San Juan is bimodally distributed and Iquitos seems more normally distributed. This is good to know for when building models, should this feature be incorporated. To get an idea of how well this is representative of the population, I used bootstrapping to draw replicates from these samples to observe what a mean value of this feature could be for each city. I split each city into its own separate subset since the pattern and volume of case numbers appear to be separate as does the climate data behavior as seen in the EDA portion of this project. Drawing bootstrap samples and utilizing the Central Limit Theorem validated the difference in mean values for the average air temperature between the cities, suggesting that the observed data is representative of the difference of the population from these two cities. I’m glad to see this result because I can be more confident there is a distinct difference between these two cities’ mean average air temperature when modeling, to know that this difference isn’t observed by chance.

Also using this approach, I looked at the precipitation values in these locations as well. Standing water is a hotbed of mosquito activity, which is the vector of spreading the Dengue viruses. I was able to find some information on what happened during those peak outbreak seasons in San Juan, but less so for the peak infection times in Iquitos. I wanted to look at how these precipitation measurements were different between the two cities as I suspected that San Juan, being in the Caribbeans, had precipitation records closer to that of Iquitos than our sample data showed. By utilizing a permutation test of means, the results showed that the two mean replicate precipitation values were much closer together than shown in our data, suggesting that both locations had close to equal amounts of precipitation on average. As seen in the EDA portion, there doesn’t seem to be a feature variable that is strongly correlated with our target variable of total cases per week, so knowing how these two features that can affect mosquito populations strongly are related and represented is important to know