CS4622 – Machine Learning Project – DengAI – Predicting Disease Spread

Group Members:

1. 140441N – N. C. N. V. Pathirana
2. 140552F – R. S. Samarawickrama
3. 140575D – I. S. Seneviratne
4. 140701M – Shane Wolff

Git Repository – https://github.com/shanewolff/CS4622-Machine-Learning-Project-DengAI

Introduction

Dengue is a mosquito borne disease caused by dengue viruses. Over the years it has spread rapidly causing deaths due to different factors that are congenial for it. Dengue is primarily caused by mosquitoes. Hence, the factors like climate changes, temperature, population growth, unplanned and uncontrolled urbanizations, poor infrastructure which provide a favorable environment for the outspread of mosquitoes causing dengue should be monitored in order to reduce the effect of dengue. Through the identification of how the outbreak of dengue varies with these factors, this can be done.

This report is based on DengAI, predicting disease spread competition hosted by Driven Data. with the task of predicting the number of Dengue cases in each week in each location. To obtain an accurate prediction, machine learning and data mining techniques can be used while considering both the direct and indirect factors affecting the outspread of dengue.

Disease prediction using machine learning techniques has become popular lately as there are different models/approaches which will be useful to identify and prevent fatal diseases before they spread and cause more harm to human lives. Using machine learning approaches, diseases can be predicted easily and with lesser medical equipment if there is sufficient data. Through the predictions in this DengAI competition, a proper understanding about the relationship between climate and dengue dynamics can be obtained which paves the path for new research areas, as well as is important in eliminating life threatening pandemics.

Methodology

Complete methodology can be summarized as follows.

1. **Data Cleaning:**
   Checked all the features and samples for missing values. Missing values were filled with front fill method which repeats the same last known value for all consecutive missing values.
2. **Feature Analysis:**
   All the features of the dataset were analyzed by general statistical analysis which includes count, mean, standard deviation, minimum, maximum etc. Dataset was divided into two sets based on the city. Feature correlation was analyzed and found out that target does not correlate to any individual feature. Satellite imagery scores of vegetation growing have a strong correlation in city Iquitos but not well in San Juan. Temperature features are strongly correlated in San Juan but not in Iquitos.
3. **XGBoost Prediction:**
   XGBoost regressor was used to predict the target cases using average NDVI score.
4. **Monthly Trend Prediction:**
   A Simple time series model using only month feature was tested.
5. **Monthly Trend plus Residual Prediction:**
   A linear regression model was trained to predict the monthly trend and the residual was calculated using the target. The residual was then predicted using temperature and vegetation data. The predicted residual was added to the predicted monthly trend in order to output the final target prediction.

Results and Analysis

The model alone did not result a good prediction. The secret lies in feature engineering. Samples of missing values cannot be neglected since they all together represent a great portion of the prediction. Therefore the front fill method for filling missing values did considerably good in prediction.

Predicted monthly trend resulted in a mean absolute error of 26.5 and when the prediction was smoothed with rolling mean of window size 3 resulted in a mean absolute error of 25.8, which was an improvement. Therefore, data smoothening plays a vital role in prediction tasks since smoothening reduces outlier effects.

Correlation analysis revealed that individual features do not have any correlation to target prediction but satellite imagery score of vegetation data and temperature features tend to correlated in Iquitos (figure 1) and San Juan (figure 2) respectively. Therefore average scores of NDVI and temperature were used instead.

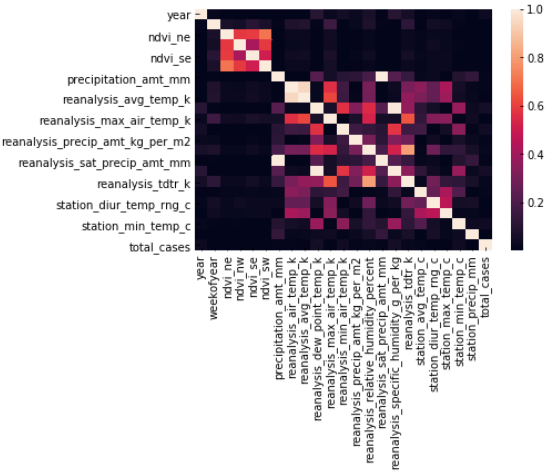

Figure 1: Correlation Heat Map – Iquitos

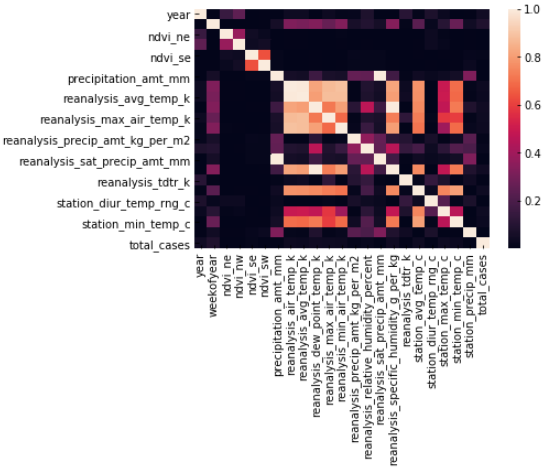

Figure 2: Correlation Heat Map – San Juan

For XGBoost prediction average NDVI score was used since they tend to correlate each other. The prediction was not improved as expected. Mean absolute error for XGBoost approach was 27.37.

Finally a residual between the actual target and the predicted monthly trend was calculated. The residual was predicted with weather features and added to the predicted trend and resulted in a mean absolute error of 20.77.

Conclusion

- Front fill method of substitution for missing values did a fairly good job.
- Averaging NDVI and temperature scores is better than using component features alone since they show a degree of correlation.
- A prediction model alone could not achieve the desired accuracy, some feature selection and methodology is required to hit a lower error.
- The residual between the predicted monthly trend and the actual target has been fairly modeled by weather data.