DengAI : Predicting Disease Spread

140441N - N.C.N.C. Pathirana

140552F - R.S.Samarawickrama

140575D - I.S.Seneviratne

140701M - S.L.S.wolff

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.
  - Temperature
  - Climate changes
  - Uncontrolled urbanization
  - Poor infrastructure

Introduction ctd.

- Data set for the prediction was based on two cities; San Juan and Iquitos .
- Data set included the following features.
  - City and Data Indicators
    - City - sj for San Juan and iq Iquitos
    - Week_start_date - Date given in yyyy-mm-dd format
  - NOAA’s GHCN Daily Climate Data weather station measurement
    - station_max_temp_c - maximum temperature
    - station_min_temp_c - minimum temperature
    - station_avg_temp_c - average temperature
    - station_percip_mm - total precipitation
    - station_diur_temp_rng_c - Diurnal temperature range

Introduction ctd.

- 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_temp_air_k– Minimum air temperature
    - reanalysis_avg_temp_k– Average air temperature
    - reanalysis_tdtr_k– Diurnal temperature range

Introduction ctd.

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

- 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

Methodology

- Data Cleaning
  - Checked all the features and samples for missing values.
  - Missing values were filled with front fill method
- Feature Analysis
  - 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.

Methodology ctd.

    - 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.
- XGBoost Prediction
    - XGBoost regressor was used to predict the target cases using average NDVI score.
- Monthly Trend Prediction
    - A Simple time series model using only month feature was tested.

Methodology ctd.

- Monthly Trend and Residual Prediction
  - A linear regression model was trained to predict the monthly trend
  - residual was calculated using the target.
  - residual was then predicted using temperature and vegetation data.
  - predicted residual was added to the predicted monthly to get the final target prediction

Results & Analysis

- Prediction from Monthly Trend resulted in a mean absolute error of 26.5
- After smoothing with rolling mean of window size 3, the mean absolute error was 25.8
- 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.

Results & Analysis ctd

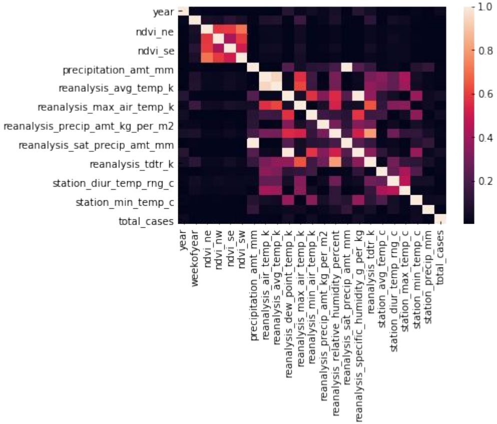


Figure 1

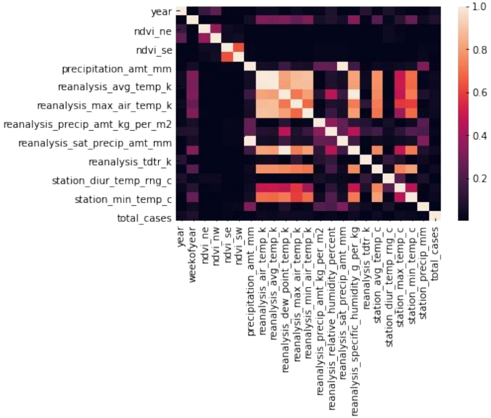
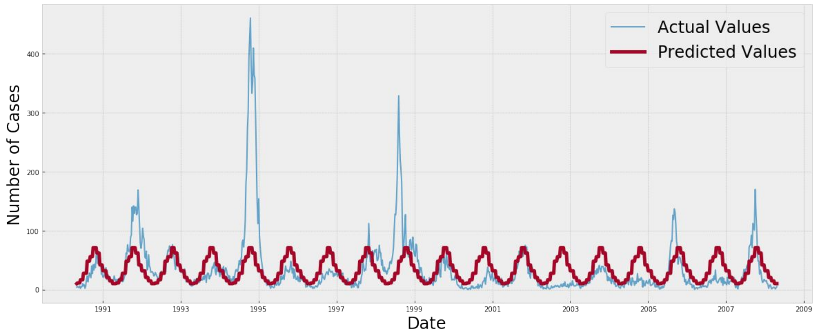


Figure 2

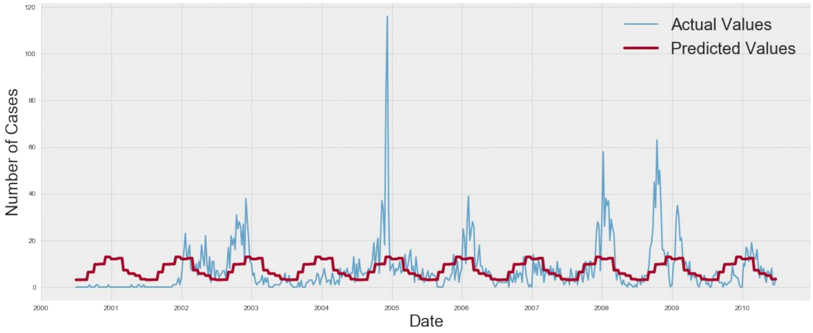
Results & Analysis ctd

- 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.
- It resulted in a mean absolute error of 20.77

Results & Analysis ctd



Actual and Predicted Value - Iquitos



Actual and Predicted Value - San Juan

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.

Thank You !