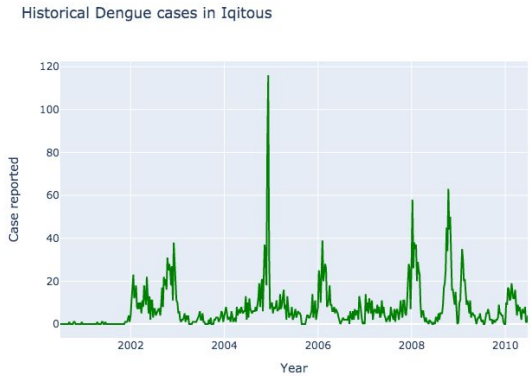
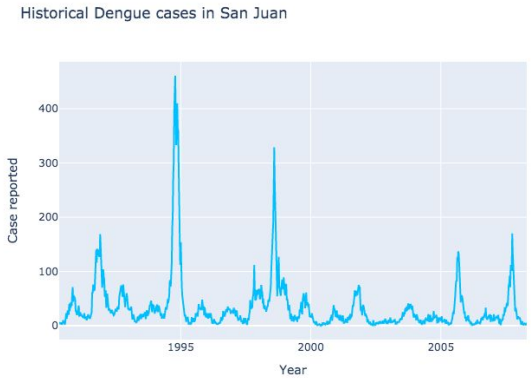
Dengue case prediction using Machine learning

Background

Dengue is a mosquito-borne disease that is spread by dengue virus infected mosquito bites. Seasonal change of climate also appears to be correlated with increased dynamics of virus transmission. The Data also has climate information of these two cities from NOAA- GHCN daily, NOAA-NCEP, and NDVI. Two stage machine learning model is proposed to predict weekly Dengue cases for two cities.

Data exploration





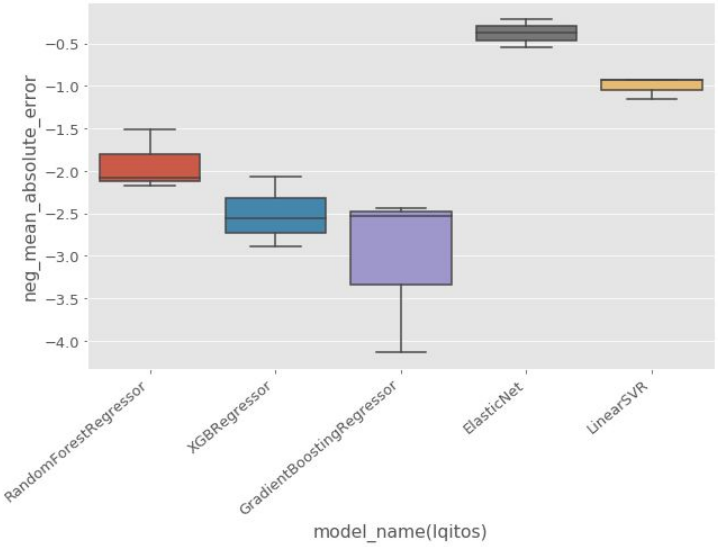
The different measurements of climate such as temperature, precipitation, humidity and vegetation were provided for San Juan and Iquitos from the sources mentioned above. Exploratory data analysis of various feature distribution, correlation plots, average yearly cases reported, and historical cases reported for both cities shows that most of the features are skewed and do not follow normal distribution pattern. The correlation plot for both cases (see DengAI-v1_final.ipynb file) indicates that temperature features are highly correlated to one another for one city, and vegetation features are highly correlated to one another for another city, while none of the features show any significant correlation with the target variable for dengue case reported. Stationarity and decomposition check of reported dengue cases revealed slight seasonality over years, and also few outbreaks reported on specific years. The data also shows that San Juan has more number of dengue cases reported compared to Iquitos.

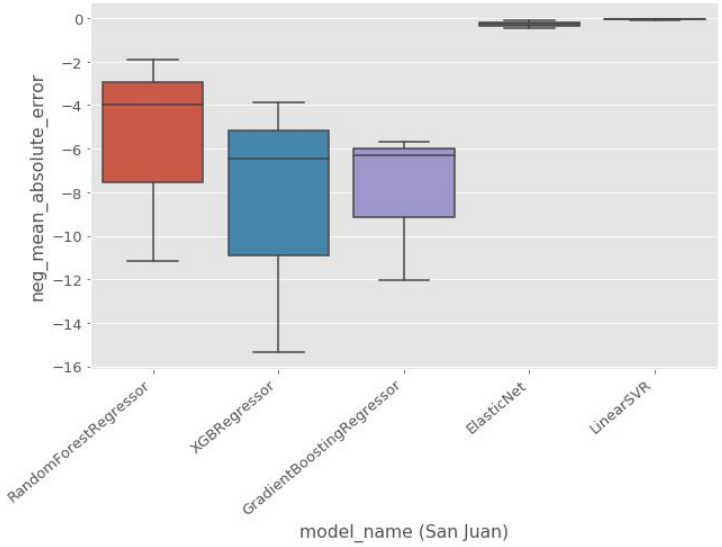
Two stage approach and feature engineering

Exploratory data analysis shows that dengue case dataset contains time-series data, which needs to be transformed in a certain way to apply Machine learning models. Also, both cities have different patterns of the reported cases, so separate models need to be used for each city and to be joined for the final prediction. Since symptoms of the illness may appear a few days after mosquito bite and person would be able to visit the doctor after 2-3 days, in order to account for

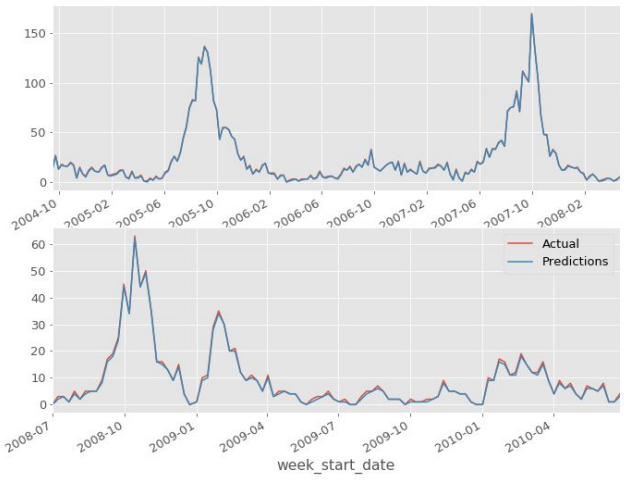
such delay, we need to count previous weeks information. For this reason, the predictive model would need to take a two-stage approach. In the first stage, machine learning model predicts weekly 'case reported' for the test data. In this stage, it is possible to generate extra features based on rolling sum and rolling average for temperature, precipitation, humidity and vegetation variables. In the second stage, 'predicted case reported' variables generated in the first stage are used to create more features using autoregressive and moving average.

Model selection and parameter tuning





In both stages, the dataset was trained through different ML regression models such as RandomForest, GBM, XGB, ElasticNet and SVM, which were compared by the negative absolute error as a scoring criteria. ElasticNet and XGB performed best of all, which is why ElasticNet is chosen as a final model for the grid search for hyper parameter tuning.



Final Prediction and thoughts

The final prediction result using ElasticNet is a fit good with minimum negative mean absolute error for Iquitos (-0.4) and San Juan (-0.2). The test prediction MAE score is 25.8 MAE. This could suggest that my training model may be overfitted, which could be improved with further feature engineering if time permits.