A model for Predicting Dengue Fever Spread of city San Juan and Iquitos from Environmental Factors

Final report prepared for the Data Analysis and Interpretation Specialization

June 10, 2018

Introduction to the Research Question

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation.

In recent years dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands, many of the nearly half-billion cases per year are occurring in Latin America. This study aims to the dataset of city San Juan and Iquitos which is provided by DrivenData, build an Artificial-Intelligence model to predict local epidemics of Dengue Fever from some Environmental Factors:

1. Temperature
2. Humidity
3. Rainfall
4. Vegetation Level

The motivation to do this topic is because: nowadays, human health is one of the hottest topics, which also caught my interest. The prediction model can provide the characteristics of the Environmental Factors which may cause Dengue Fever spread, also might be useful for prearranging public health actions of city San Juan and Iquitos to reduce the effects of major outbreaks. (Data Source: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/)

Methods

Sample

This study is using environmental data of city San Juan and Iquitos collected by various U.S. Federal Government agencies—from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration (NOAA) in the U.S. Department of Commerce. The total sample size is N=1456 and set on a (year, weekofyear) timescale, from (1990, 18) to (2010, 25). The dataset includes information from each of the following sources (as available): “city and date indicators”, “NOAA’s GHCN daily climate data weather station measurements”, “PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)”, “NOAA’s NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)”, “Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA’s CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements”.

Measures

The target variable was the total cases of Dengue Fever for each (city, year, weekofyear) scale for both cities, San Juan and Iquitos, which is a quantitative variable.

Predictors included:

1) 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

2) 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

3) PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- precipitation_amt_mm – Total precipitation

4) 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

5) 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

Analysis

The distributions for the predictors and the total cases target variable were evaluated by examining frequency tables for categorical variables and calculating the mean, standard deviation and minimum and maximum values for quantitative variables.

Distribution plots and Scatterplot matrices were also examined, Pearson correlation is used to test bivariate associations between two pairs of predictors, to see if a dimensionality reduction is needed.

Single-predictor and target Pearson Correlation parameters were used to identify the subset of variables that best predicted total cases (with 0.9 percentile). XGBoost is also used to select important predictors (with > .001 threshold). The final prediction model combined with 10 machine learning models to estimated on a training data set consisting of a random sample of 70% of the total samples (N=1019), and a test dataset included the other 30% of the batches (N=437). All predictors were standardized to have a mean=0 and standard deviation=1 prior to conducting the final machine learning model analysis. Cross-validation was performed using k-fold cross validation specifying 10 folds. The change in the cross-validation mean absolute error rate at each step was used to identify the best subset of predictor variables. Predictive accuracy was assessed by determining the mean absolute error rate of the training data prediction algorithm when applied to observations in the test dataset.

Results

Missing Data Management

As Figure1 shows, many of the predictors have missing values, from 0.7% to 14%, and there is only 7% of the observations which contain missing values have lower dengue fever cases. Therefore, there is not able to delete these observations and missing values were filled with 50% quantile values.

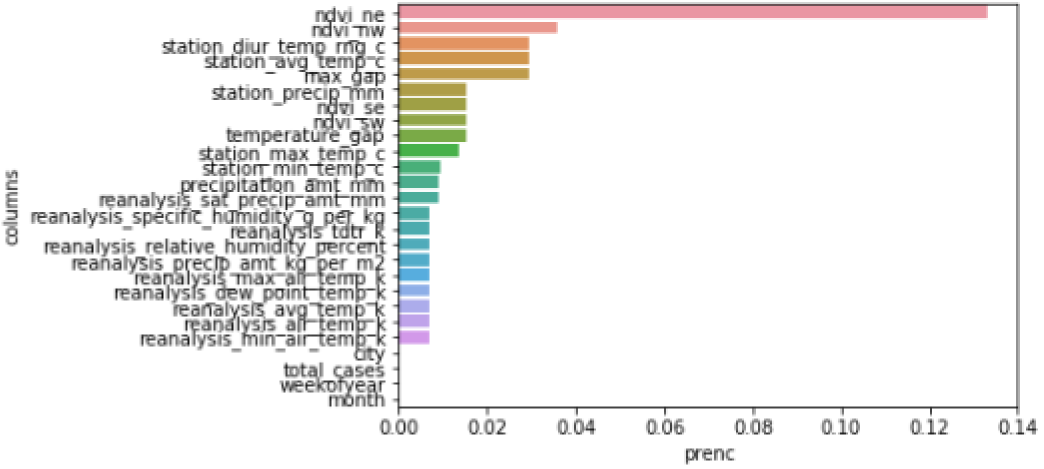


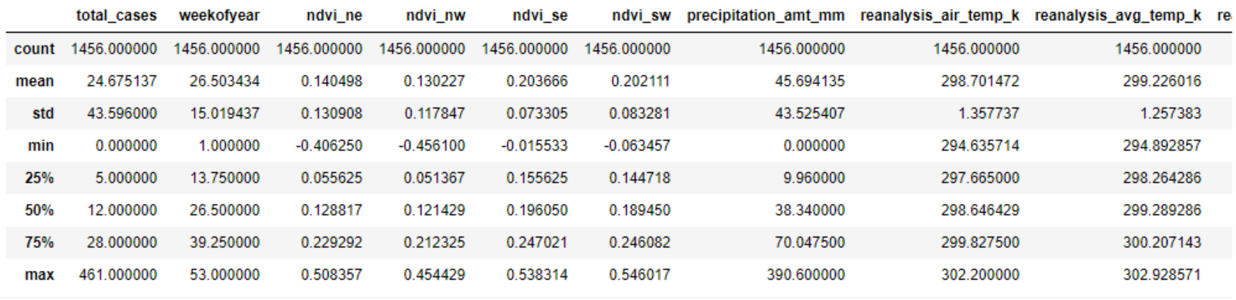
Figure1

Feature Extraction and Normalization

Based on the predictors in the original dataset, the second predictors temperature_gap, max_gap and month were extracted. Temperature_gap is station max temperature average minus station min temperature. Max_gap is station max temp minus station avg temp. Month is the month of every observation.

All quantitative predictors were standardized to have a mean=0 and standard deviation=1, while categorical predictors were encoded by using One Hot Encoding.

Descriptive Statistics



	total_cases	weekofyear	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k	re
count	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	1456.000000	
mean	24.675137	26.503434	0.140498	0.130227	0.203666	0.202111	45.694135	298.701472	299.226016	
std	43.596000	15.019437	0.130908	0.117847	0.073305	0.083281	43.525407	1.357737	1.257383	
min	0.000000	1.000000	-0.406250	-0.456100	-0.015533	-0.063457	0.000000	294.635714	294.892857	
25%	5.000000	13.750000	0.055625	0.051367	0.155625	0.144718	9.960000	297.665000	298.264286	
50%	12.000000	26.500000	0.128817	0.121429	0.196050	0.189450	38.340000	298.646429	299.289286	
75%	28.000000	39.250000	0.229292	0.212325	0.247021	0.246082	70.047500	299.827500	300.207143	
max	461.000000	53.000000	0.508357	0.454429	0.538314	0.546017	390.600000	302.200000	302.928571	

Table1

Table1 shows descriptive statistics for part of the predictors of the training dataset. The average total cases were 24.7 (Std = 43.6), with a minimum total case 0 and a maximum 461. The max air temperature is 302.2, the min air temperature is 294.64, the mean is 298.7 and standard deviation is 1.36. All the pixel of city centroid are between -1 to 1.

The Figure2 shows that the dengue fever incidence was extremely high in years 1994 and 1998. It may contain some important information.

Figure3 and 4 tell that month 9 to 11 and week 36 to 47 have the higher dengue fever incidence than others. Month 10 and week 40 has almost 50 average dengue fever cases. This information can help us to prepare to reduce dengue fever cases.

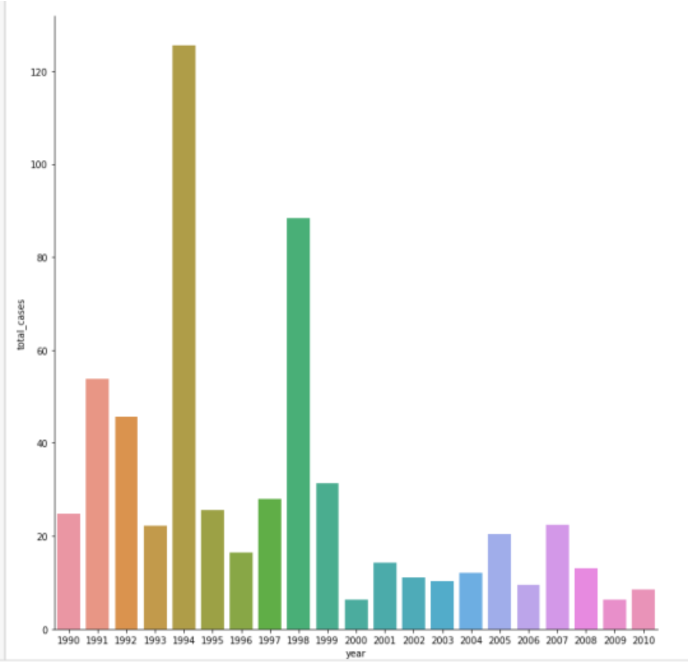


Figure2

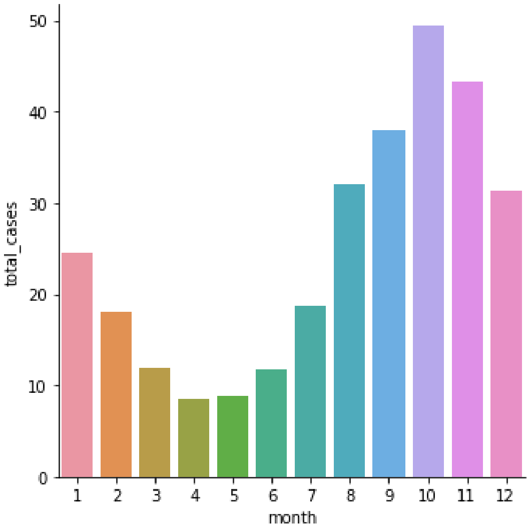


Figure3

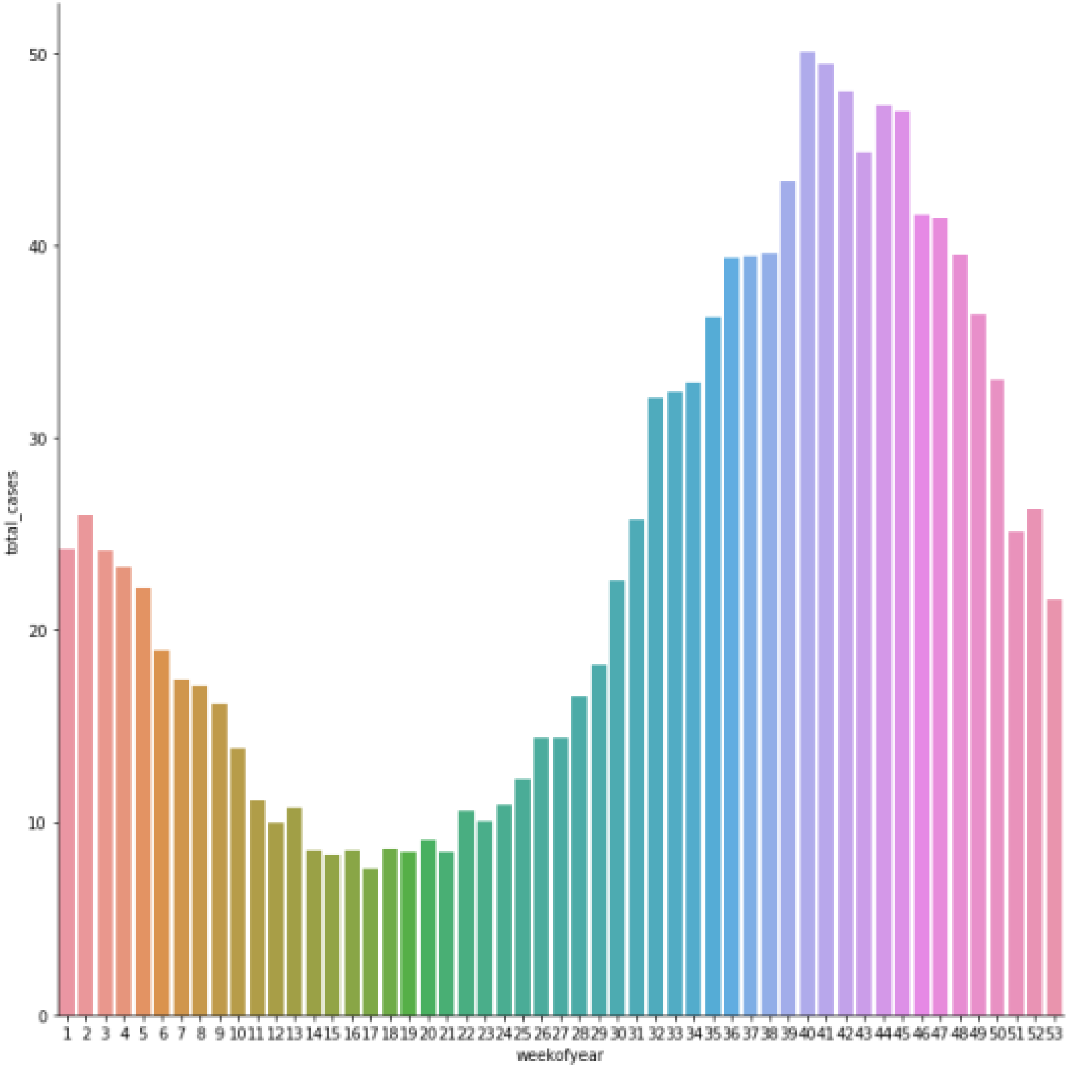


Figure4

Bivariate Analysis

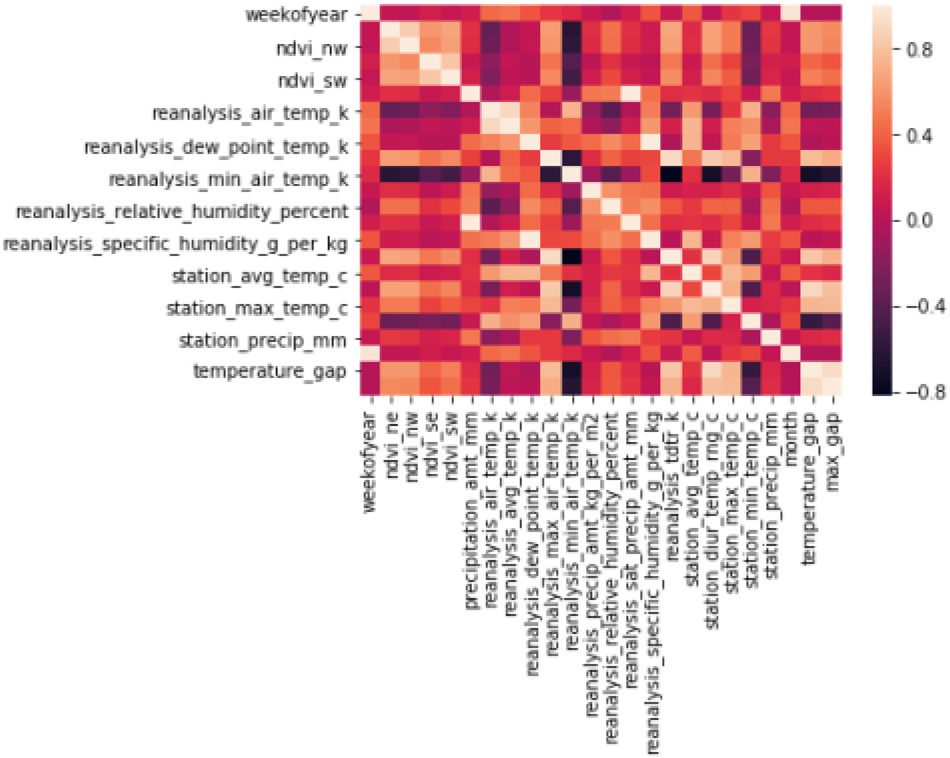
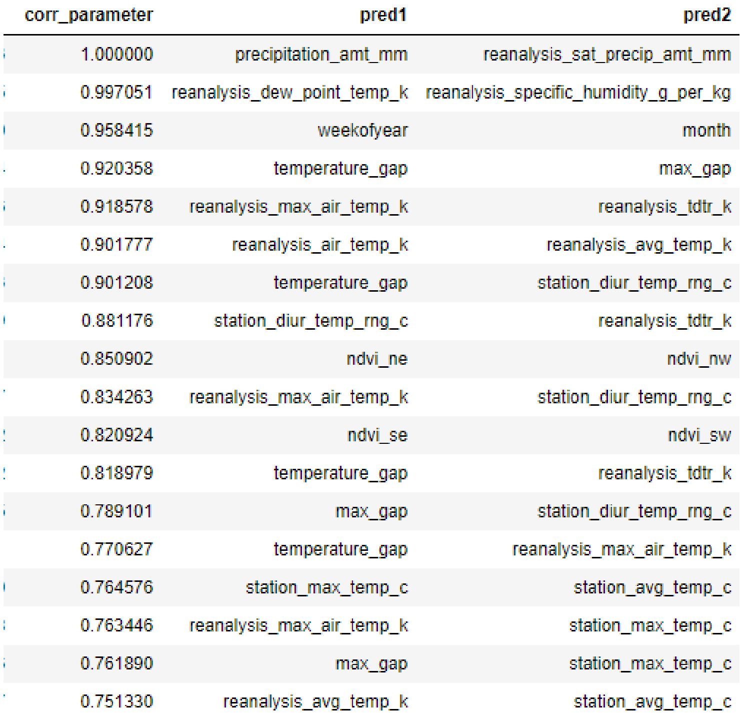


Figure5



corr_parameter	pred1	pred2
1.000000	precipitation_amt_mm	reanalysis_sat_precip_amt_mm
0.997051	reanalysis_dew_point_temp_k	reanalysis_specific_humidity_g_per_kg
0.958415	weekofyear	month
0.920358	temperature_gap	max_gap
0.918578	reanalysis_max_air_temp_k	reanalysis_tdtr_k
0.901777	reanalysis_air_temp_k	reanalysis_avg_temp_k
0.901208	temperature_gap	station_diur_temp_rng_c
0.881176	station_diur_temp_rng_c	reanalysis_tdtr_k
0.850902	ndvi_ne	ndvi_nw
0.834263	reanalysis_max_air_temp_k	station_diur_temp_rng_c
0.820924	ndvi_se	ndvi_sw
0.818979	temperature_gap	reanalysis_tdtr_k
0.789101	max_gap	station_diur_temp_rng_c
0.770627	temperature_gap	reanalysis_max_air_temp_k
0.764576	station_max_temp_c	station_avg_temp_c
0.763446	reanalysis_max_air_temp_k	station_max_temp_c
0.761890	max_gap	station_max_temp_c
0.751330	reanalysis_avg_temp_k	station_avg_temp_c

Table2

The Pearson Correlation Heatmap for the association between each pair of quantitative predictors (Figure5) revealed that some pairs of predictors are significantly associated with each other, so may need dimension reduction in next steps (the darker colors means these two pair of predictors are more associated with each other).

There are 17 pairs of predictors have high correlation parameter. Precipitation_amt_mm and reanalysis_sat_precip_amt_mm have 100% correlation with each other. Before we use machine learning model to analyze these predictors, we are not able to decide if there is a need to delete some of the predictors.

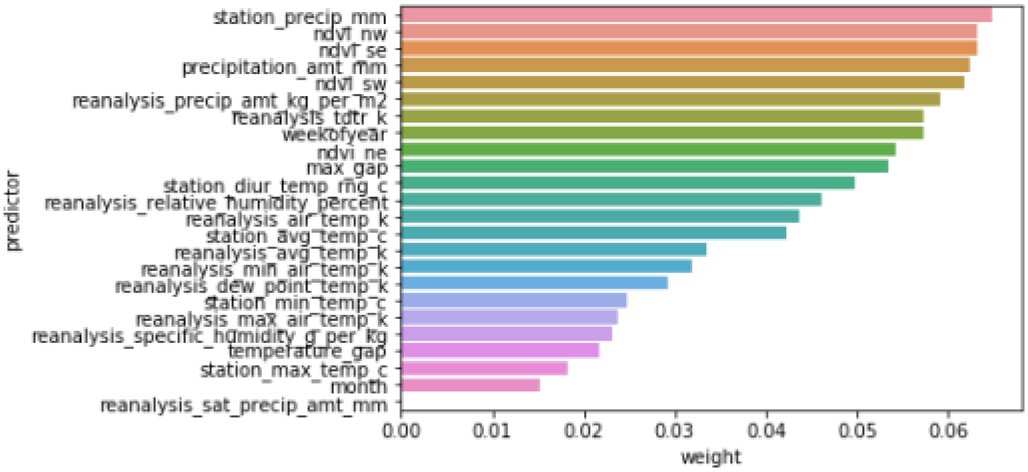
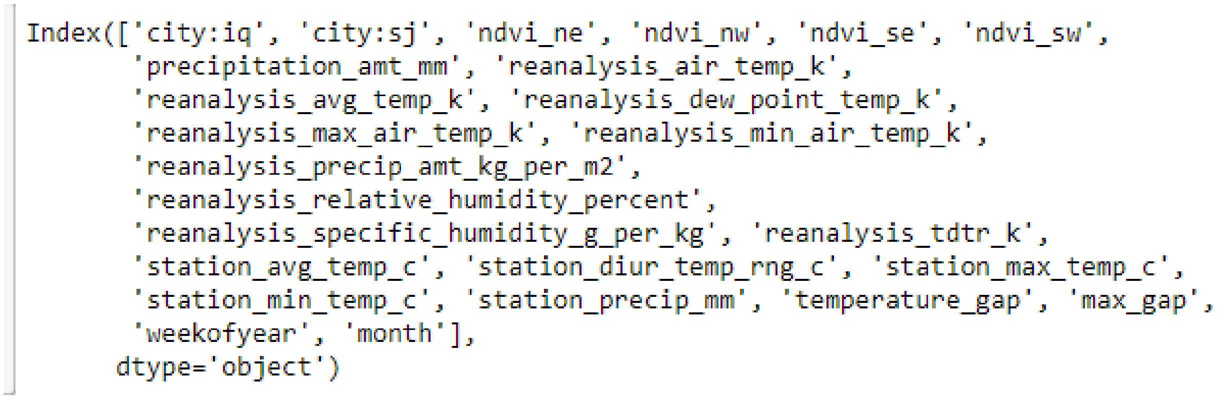


Figure6

XGBoost model is used for feature selection, as Figure6 shows, all the predictors have a weight more than 0.001 except predictor reanalysis_sat_precip_amt_mm (from 0.015 to 0.065), so delete the reanalysis_sat_precip_amt_mm predictor and don't do any dimension reduction. In the result, the number of 25 predictors were selected for future analysis:



```
Index(['city:iq', 'city:sj', 'ndvi_ne', 'ndvi_nw', 'ndvi_se', 'ndvi_sw',
       'precipitation_amt_mm', 'reanalysis_air_temp_k',
       'reanalysis_avg_temp_k', 'reanalysis_dew_point_temp_k',
       'reanalysis_max_air_temp_k', 'reanalysis_min_air_temp_k',
       'reanalysis_precip_amt_kg_per_m2',
       'reanalysis_relative_humidity_percent',
       'reanalysis_specific_humidity_g_per_kg', 'reanalysis_tdtr_k',
       'station_avg_temp_c', 'station_diur_temp_rng_c', 'station_max_temp_c',
       'station_min_temp_c', 'station_precip_mm', 'temperature_gap', 'max_gap',
       'weekofyear', 'month'],
      dtype='object')
```

Combined Model Analysis/Multivariable analysis

The EVALUATION METRIC of the model is mean absolute error. The absolute error is calculated for each label in the submission and then averaged across the labels. The goal is to minimize MAE.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

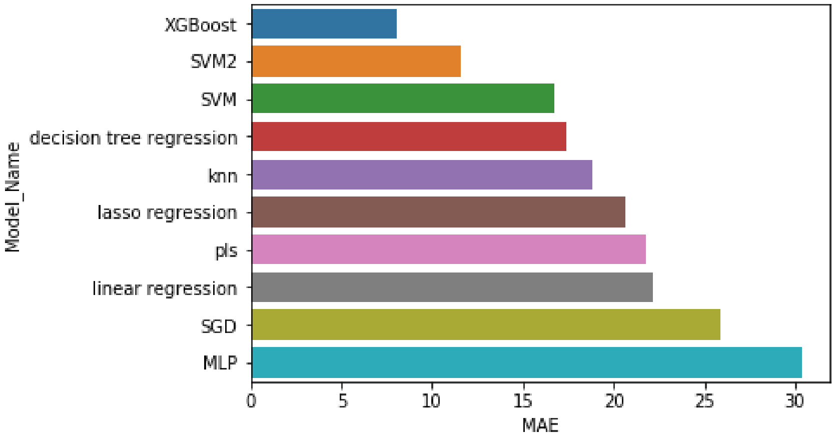
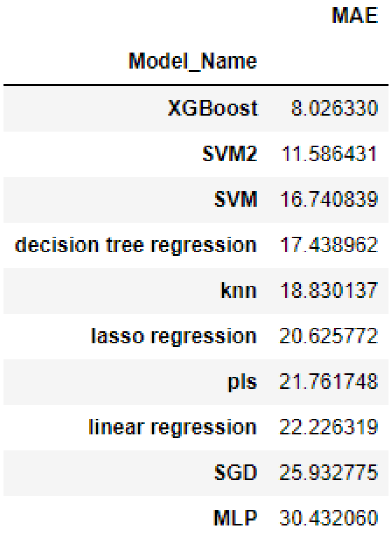


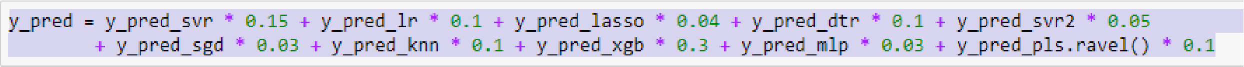
Figure7



	MAE
Model_Name	
XGBoost	8.026330
SVM2	11.586431
SVM	16.740839
decision tree regression	17.438962
knn	18.830137
lasso regression	20.625772
pls	21.761748
linear regression	22.226319
SGD	25.932775
MLP	30.432060

Table3

By following Figure7 and Table3, which can know XGBoost model has the lowest mean absolute error, it was given the highest weight when building the final model. The final model was combined with following equation:



```
y_pred = y_pred_svr * 0.15 + y_pred_lr * 0.1 + y_pred_lasso * 0.04 + y_pred_dtr * 0.1 + y_pred_svr2 * 0.05
        + y_pred_sgd * 0.03 + y_pred_knn * 0.1 + y_pred_xgb * 0.3 + y_pred_mlp * 0.03 + y_pred_pls.ravel() * 0.1
```

The models used to build the final regression prediction model are list on Figure7, this model is using Python programming language: XGBoost (XGBRegressor), SVM (used twice, SVR), Decision Tree (DecisionTreeRegressor), KNN (KNeighborsRegressor), Lasso (Lasso), PLS(PLSRegression), Linear Regression (LinearRegression), SGD (SGDRegressor), MLP (MLPRegressor).

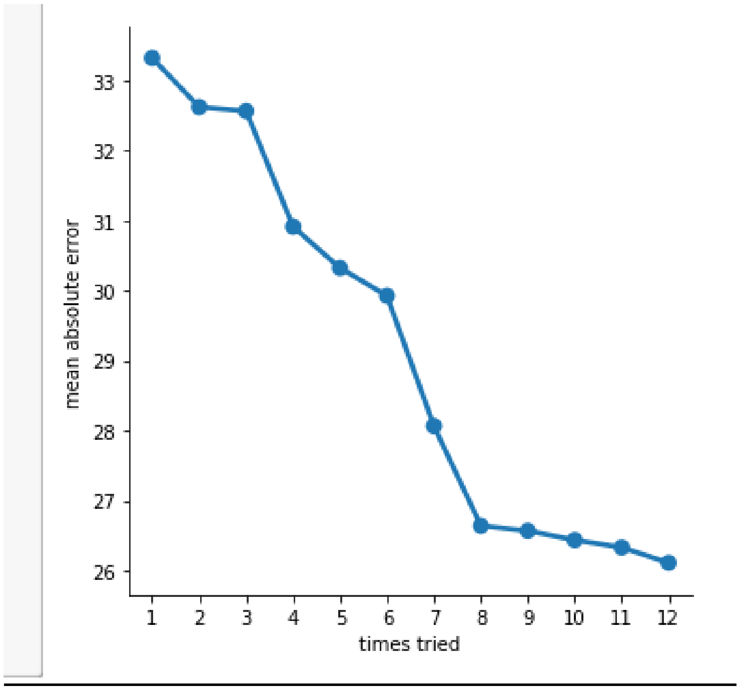
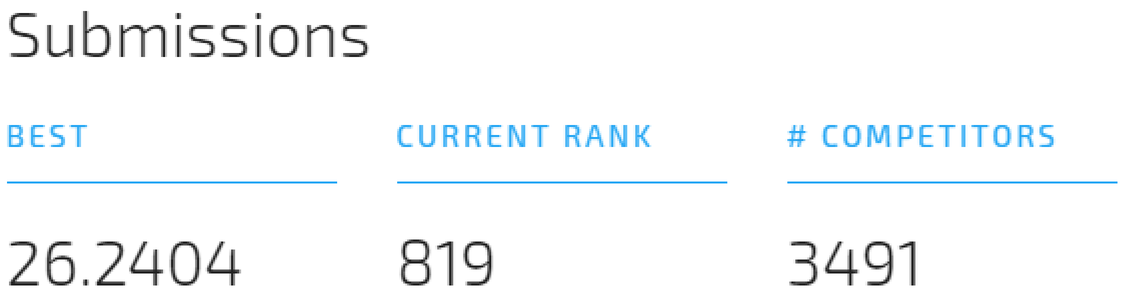


Figure8

Optimized the final model by tweaking the parameters of each model and the weight of every model in the final model. After uploading the submission file 12 times, the Mean Absolute Error is from 33.3221 to 26.2404 (Figure8), at rank top 819 of 3491 competitors (Figure9).



Submissions

BEST	CURRENT RANK	# COMPETITORS
26.2404	819	3491

Figure9

Main Characteristics of Environmental Factors When Dengue Fever Rate is High

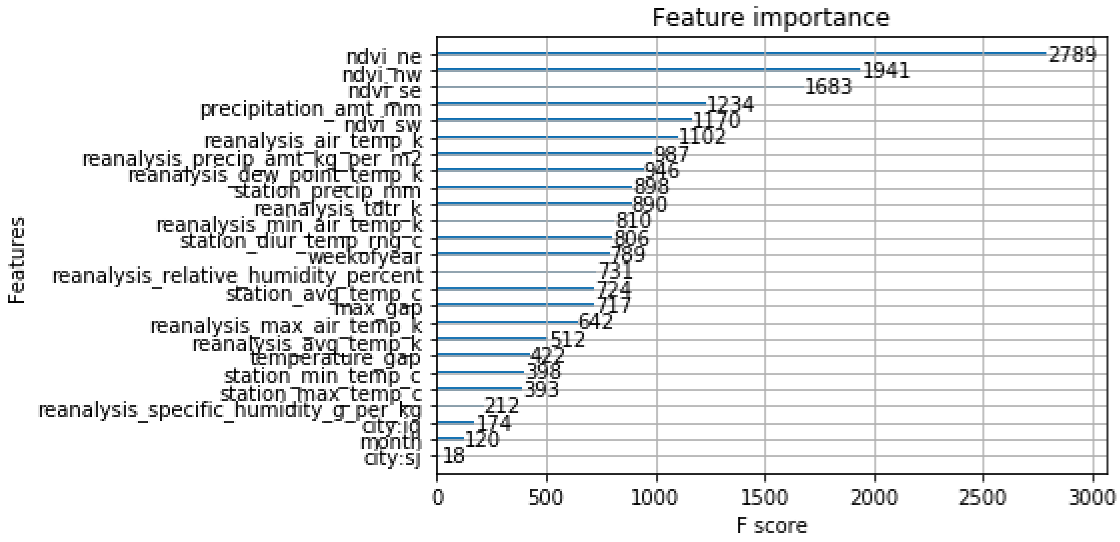


Figure10

Based on the final best model, NDVI (Normalized difference vegetation index) is the most important predictor of Dengue Fever cases: the pixel of city centroid. Dengue Fever has a higher rate when the pixel of northeast and northwest are between 0 to 0.2, the pixel of southeast and southwest are around 0.2.

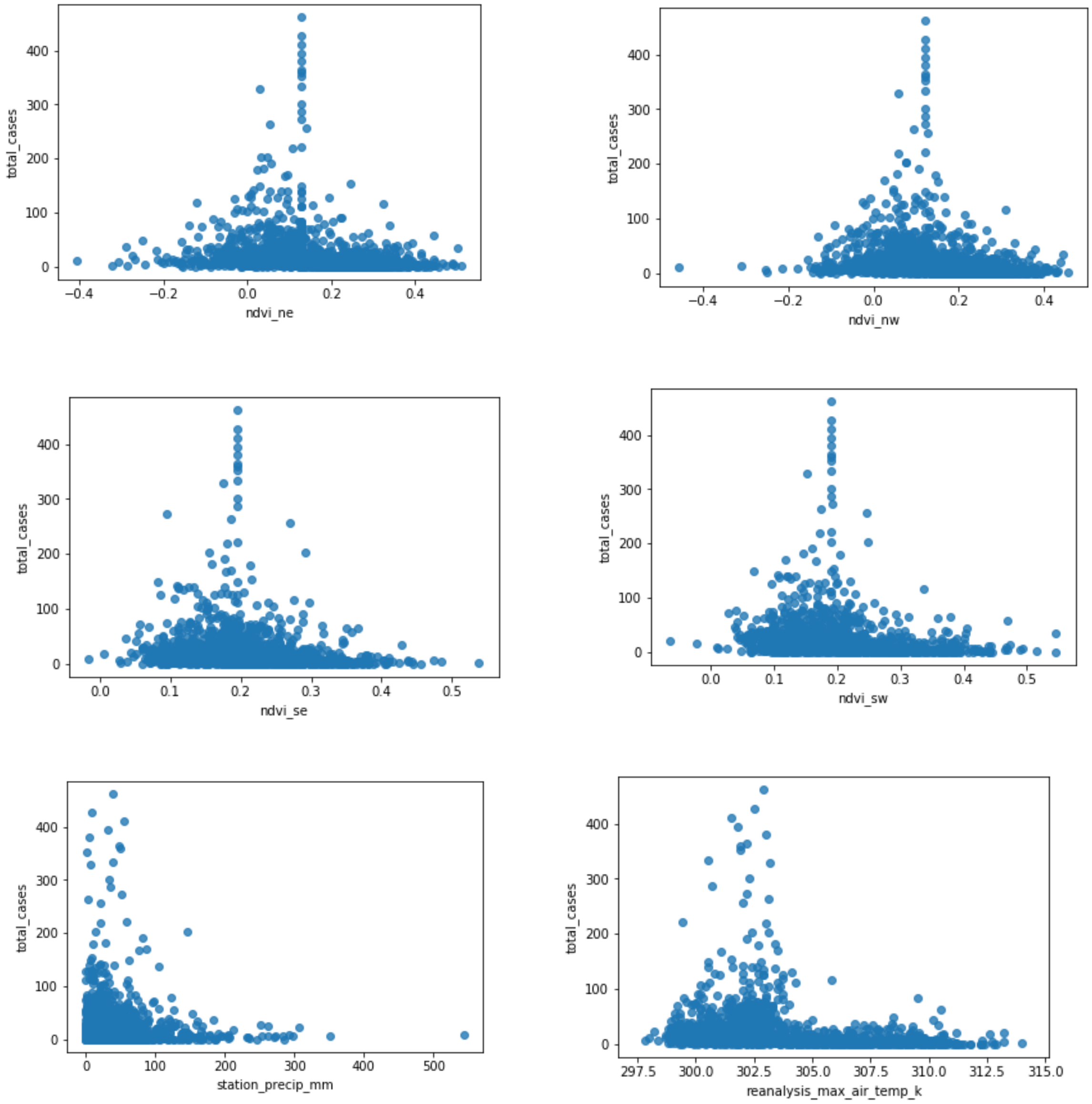


Figure11

City San Juan has a higher dengue fever rate. Dengue fever has a higher rate when total precipitation is around 50 millimeters, average temperature is around 299.5 Kelvins, mean dew point temperature is around 295.5 Kelvins, diurnal temperature range is around 3 Kelvins, minimum air temperature is around 298 Kelvins, minimum temperature is around 24 Celsius, maximum air temperature is around 303 Kelvins, maximum temperature is around 33 Celsius, mean relative humidity is around 80.

Conclusions/Limitations

Overview of key findings

Modelling dengue fever using environmental and NDVI factors with ten traditional machine learning approaches can help predict dengue fever incidence rates in San Juan and Iquito with a power greater than 70 percent.

Four model runs, total precipitation, air temperature, normal temperature, previous dengue cases, week of year, and rainfall as the most influential factors that predict dengue fever outbreak occurrences. NDVI factors, including pixel of city centroid, were of most importance.

Implications

Understanding the meaning of the environmental factors can help improve the effectiveness of early warning systems in this region and mitigate the disease. Dengue fever has a higher rate around October, so we can do some preparation before October.

Limitations

The limitation of our study is that dengue fever is effected by the medical development very easily, so in order to reduce the dengue fever cases, hospitals cannot only rely on our machine learning model.

Future directions

Further studies are needed to incorporate vector and dengue fever virus dynamics into models, as these can help improve the skill of simulations and understand similar diseases that depend on climate and environmental changes.