A Stacking Ensemble Prediction Approach to Weekly Dengue cases Forecasting based on Weather Conditions

Introduction and Problem Definition

The objective of this project is to predict the number of dengue fever cases that will be reported within a particular time span using the environmental data from San Juan and Iquitos cities.
Understanding the environmental factors those spread dengue fever and predicting the count can serve as an early warning system. This helps concerned authorities prepare for abnormally high numbers of cases and create awareness among people. This is an effective way of preventing dengue fever.

Some known facts about Dengue

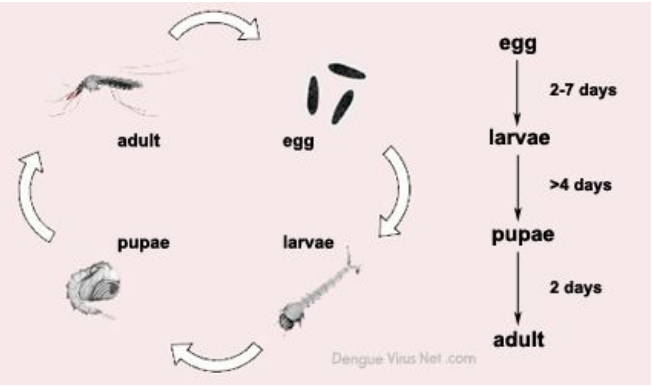
Dengue is a viral disease transmitted by some types of mosquitoes. Dengue outbreaks occur in each year. Dengue does not spread from person to person and is transmitted through the bite of an infected mosquito (Aedes aegypti and Aedes albopictus species of mosquito). Mosquitoes become infected with dengue after biting sick humans who have dengue virus in their blood. Between 8 and 12 days later if an infected mosquito bites someone else it can pass on the dengue virus. This disease causes illness that can range from a mild fever to a severe, even fatal condition.
How to Prevent: Avoiding mosquito bites is the best prevention

Effects of weather factors on dengue fever

Temperature, rainfall and humidity have well-defined roles in the Dengue transmission cycle. Prolonged periods of heavy rain increase the opportunities for the diseases to spread. Longer seasons of mild temperatures may increase the transmission likelihood of dengue diseases. In warmer temperatures, dengue become infectious more quickly and can transmit virus earlier in their lives. In addition, the temperature must be “just right”; if too cold, the virus takes too long to replicate, and if too hot, the dengue virus life spans are decreased. Humidity has been identified as a consistent, substantial weather factor to provide favourable conditions for dengue. But the dengue incidence and weather factors also apparently varies by locality.

Mosquito life cycle

The life cycle of Aedes aegypti can be completed within one-and-a-half to three weeks. Male mosquitoes live three to five days. The females live considerably longer, depending on how much warmth and moisture is in their environment. Under ideal conditions, they may last as long as a month or two.



Data Source

Using environmental data collected by various U.S. Federal Government agencies—from the Canters for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce. This project will assess forecasts using historical data from Iquitos, Peru and San Juan, Puerto Rico. This data is available in DrivenData.
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

Data Description

Datasets collected from DrivenData

1. dengue_features_test.csv - The features for the testing dataset.
2. dengue_labels_train.csv - The number of dengue cases for each row in the training dataset.
3. dengue_features_train.csv - The features for the training dataset.

Features

- City and Date indicators :
    - `city` – City abbreviations: `sj` for San Juan and `iq` for Iquitos
    - year- YYYY format
    - weekofyear - Week number according to the ISO-8601 standard
    - `week_start_date` – Date given in yyyy-mm-dd format
- 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
- PERSIANN satellite precipitation measurements
    - precipitation_amt_mm – Total precipitation
- NOAA's NCEP Climate Forecast System Reanalysis measurements
    - 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
- Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index 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

Data Exploration

As part of the data exploration, followed the below steps to get better understanding of the data.

- Data Format analysis
- Data Dimension checking
- Missing values analysis

- Target variable distribution
- Checking Duplicate entries
- Descriptive Statistics

Data Format

The Train and Test data are in csv format.

Data dimension

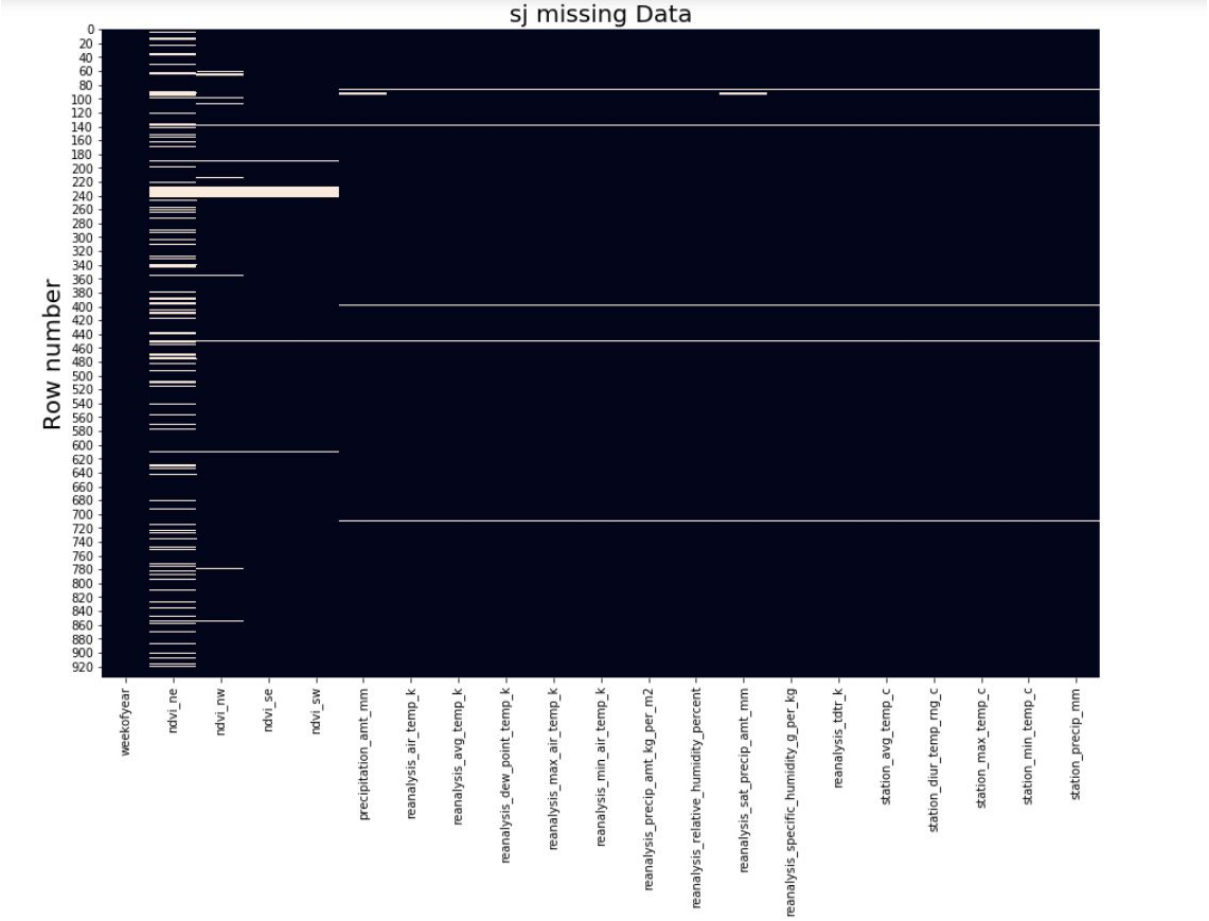
Data Set	Entries	Features	Target Variable
Train Data Set	1456	24	Yes
Test Data Set	416	24	No

Missing Values

Below table represents the missing count of each feature in train and test data.

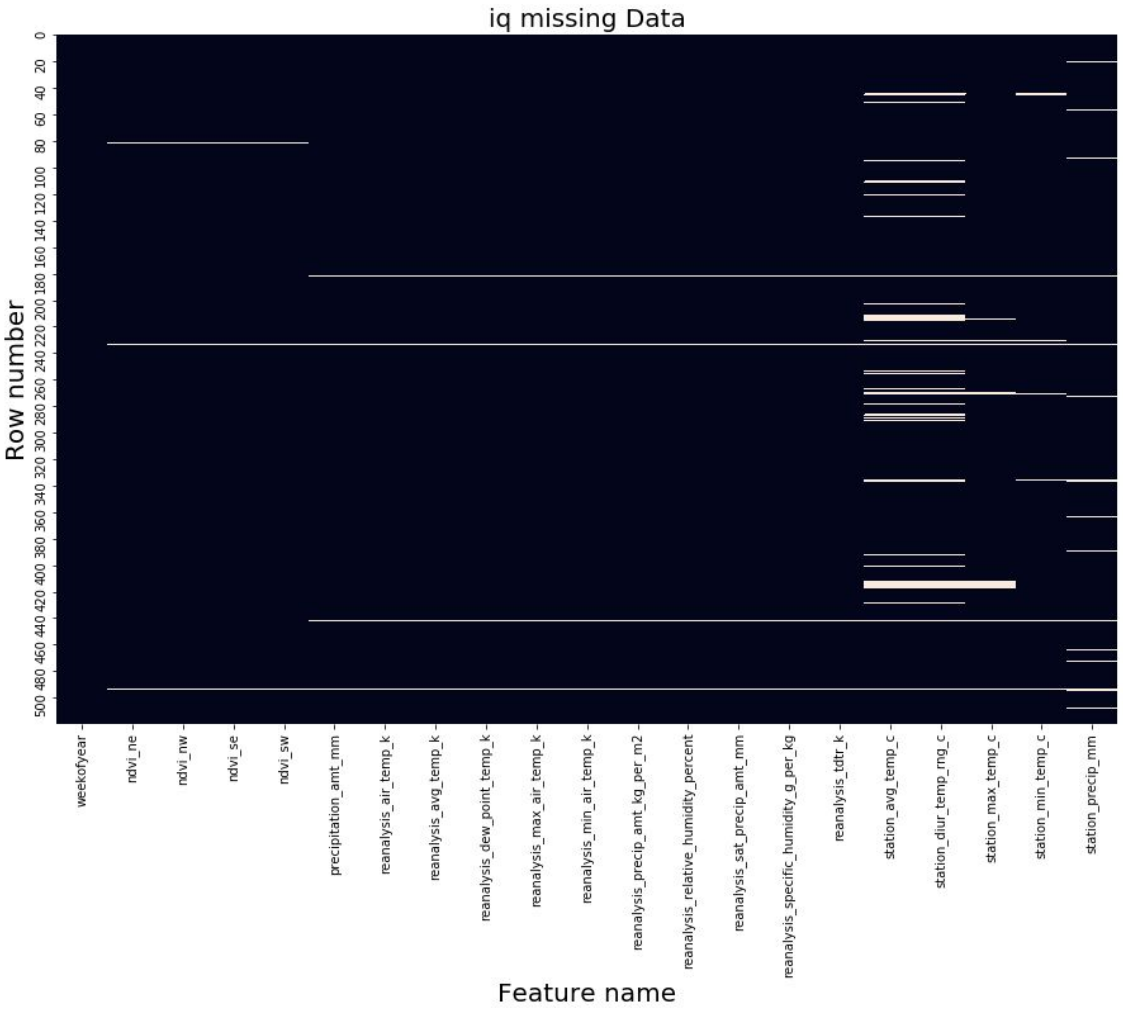
Features	Missing Count(Train)	Missing Count(Test)
City	0	0
Year	0	0
Weekofyear	0	0
week_start_date	0	0
ndvi_ne	194	43
ndvi_nw	52	11
ndvi_se	22	1
ndvi_sw	22	1
precipitation_amt_mm	13	2
reanalysis_air_temp_k	10	2
reanalysis_avg_temp_k	10	2
reanalysis_dew_point_temp_k	10	2
reanalysis_max_air_temp_k	10	2
reanalysis_min_air_temp_k	10	2
reanalysis_precip_amt_kg_per_m2	10	2
reanalysis_relative_humidity_percent	10	2
reanalysis_sat_precip_amt_mm	13	2
reanalysis_specific_humidity_g_per_kg	10	2
reanalysis_tdtr_k	10	2
station_avg_temp_c	43	12
station_diur_temp_rng_c	43	12
station_max_temp_c	20	3
station_min_temp_c	14	9
station_precip_mm	22	5

The below heat maps represent the specific location where data is missing. From this it is quite evident that the features are missing either all through the rows or columns. The frequency of the gap will also help to understand the level of data lost.



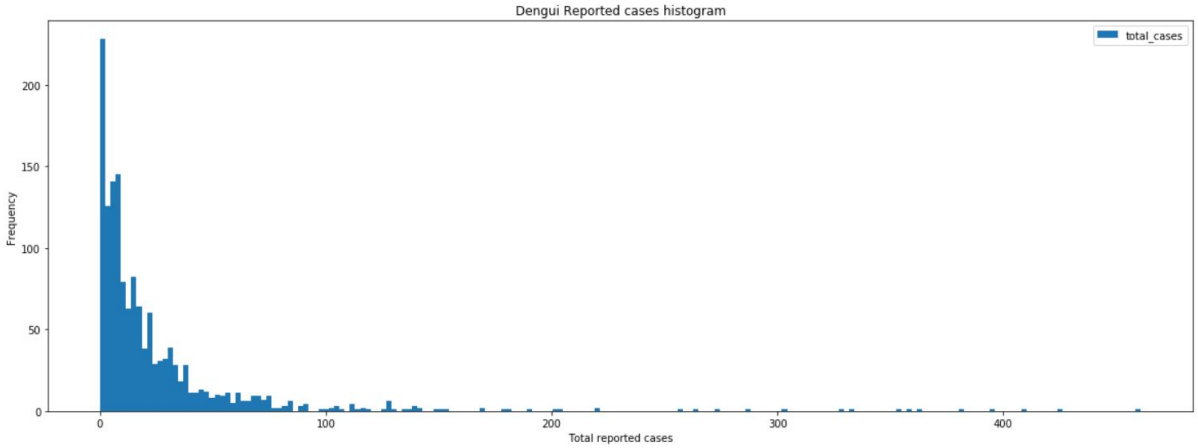
Factors Identified

- nvdi_ne feature data loss is high.
- ndvi_nw feature is also missing data in some rows(but the count of data lost is comparatively less).
- In ~5 rows , all the features are missing data after ndvi features.
- The largest missing count is in ndvi_ne (194 / 1456). It is less than 15% of train data. So it is not necessary to remove the entire column from the dataset.



- 'station_avg_temp_c' , 'station_diur_temp_rng_c'and 'station_precip_mm' features have significant data loss .
- 'station_max_temp_c', 'station_min_temp_c' features also have data loss in some rows (but the amount of data loss is less).
- In 4 rows, more than 15 features are missing data.

Target variable distribution



As per the graph, majority of the Dengue spread cases are reported with less than 100 individual cases. There are few cases reported with more 400 individual cases.

Duplicate entries

There are no duplicate entries in the data set.

Descriptive analysis

	count	mean	std	min	25%	50%	75%	max
year	1456.0	2001.031593	5.408314	1990.000000	1997.000000	2002.000000	2005.000000	2010.000000
weekofyear	1456.0	26.503434	15.019437	1.000000	13.750000	26.500000	39.250000	53.000000
ndvi_ne	1262.0	0.142294	0.140531	-0.406250	0.044950	0.128817	0.248483	0.508357
ndvi_nw	1404.0	0.130553	0.119999	-0.456100	0.049217	0.121429	0.216600	0.454429
ndvi_se	1434.0	0.203783	0.073860	-0.015533	0.155087	0.196050	0.248846	0.538314
ndvi_sw	1434.0	0.202305	0.083903	-0.063457	0.144209	0.189450	0.246982	0.546017
precipitation_amt_mm	1443.0	45.760388	43.715537	0.000000	9.800000	38.340000	70.235000	390.600000
reanalysis_air_temp_k	1446.0	298.701852	1.362420	294.635714	297.658929	298.646429	299.833571	302.200000
reanalysis_avg_temp_k	1446.0	299.225578	1.261715	294.892857	298.257143	299.289286	300.207143	302.928571
reanalysis_dew_point_temp_k	1446.0	295.246356	1.527810	289.642857	294.118929	295.640714	296.460000	298.450000
reanalysis_max_air_temp_k	1446.0	303.427109	3.234601	297.800000	301.000000	302.400000	305.500000	314.000000
reanalysis_min_air_temp_k	1446.0	295.719156	2.565364	286.900000	293.900000	296.200000	297.900000	299.900000
reanalysis_precip_amt_kg_per_m2	1446.0	40.151819	43.434399	0.000000	13.055000	27.245000	52.200000	570.500000
reanalysis_relative_humidity_percent	1446.0	82.161959	7.153897	57.787143	77.177143	80.301429	86.357857	98.610000
reanalysis_sat_precip_amt_mm	1443.0	45.760388	43.715537	0.000000	9.800000	38.340000	70.235000	390.600000
reanalysis_specific_humidity_g_per_kg	1446.0	16.746427	1.542494	11.715714	15.557143	17.087143	17.978214	20.461429
reanalysis_tdtr_k	1446.0	4.903754	3.546445	1.357143	2.328571	2.857143	7.625000	16.028571
station_avg_temp_c	1413.0	27.185783	1.292347	21.400000	26.300000	27.414286	28.157143	30.800000
station_diur_temp_rng_c	1413.0	8.059328	2.128568	4.528571	6.514286	7.300000	9.566667	15.800000
station_max_temp_c	1436.0	32.452437	1.959318	26.700000	31.100000	32.800000	33.900000	42.200000
station_min_temp_c	1442.0	22.102150	1.574066	14.700000	21.100000	22.200000	23.300000	25.600000
station_precip_mm	1434.0	39.326360	47.455314	0.000000	8.700000	23.850000	53.900000	543.300000
total_cases	1456.0	24.675137	43.596000	0.000000	5.000000	12.000000	28.000000	461.000000

The target variable is a count variable. As per the mean and variance, the data shows more dispersed behaviour. The data is not imbalanced, but a number of peaks are there in the total_cases column (after the 75 percentile).

Analysis details:

- Feature values are dispersed, std is high. There is a big gap between 75% and max value
- Value measured is within the expected range for each feature. But the unit used to measure the data needs to be converted(kelvin to celsius)

Correlation of Features

The two heat maps and pair plots given below describe the dataset's internal correlation for both cities. There is no significant correlation between the features and target variable.

There is no positive or negative correlation to the features. So it was decided to develop new features from the existing to find a better model.

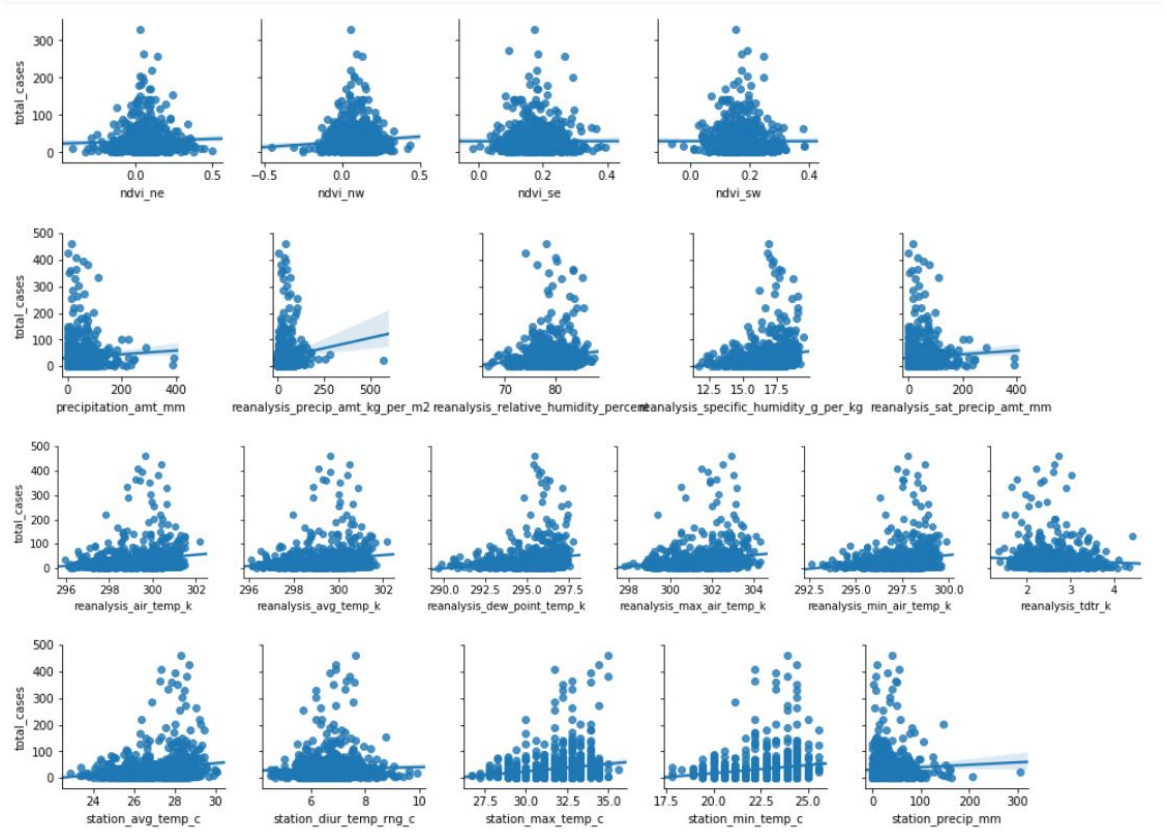


Figure 1 San Juan Pair plot

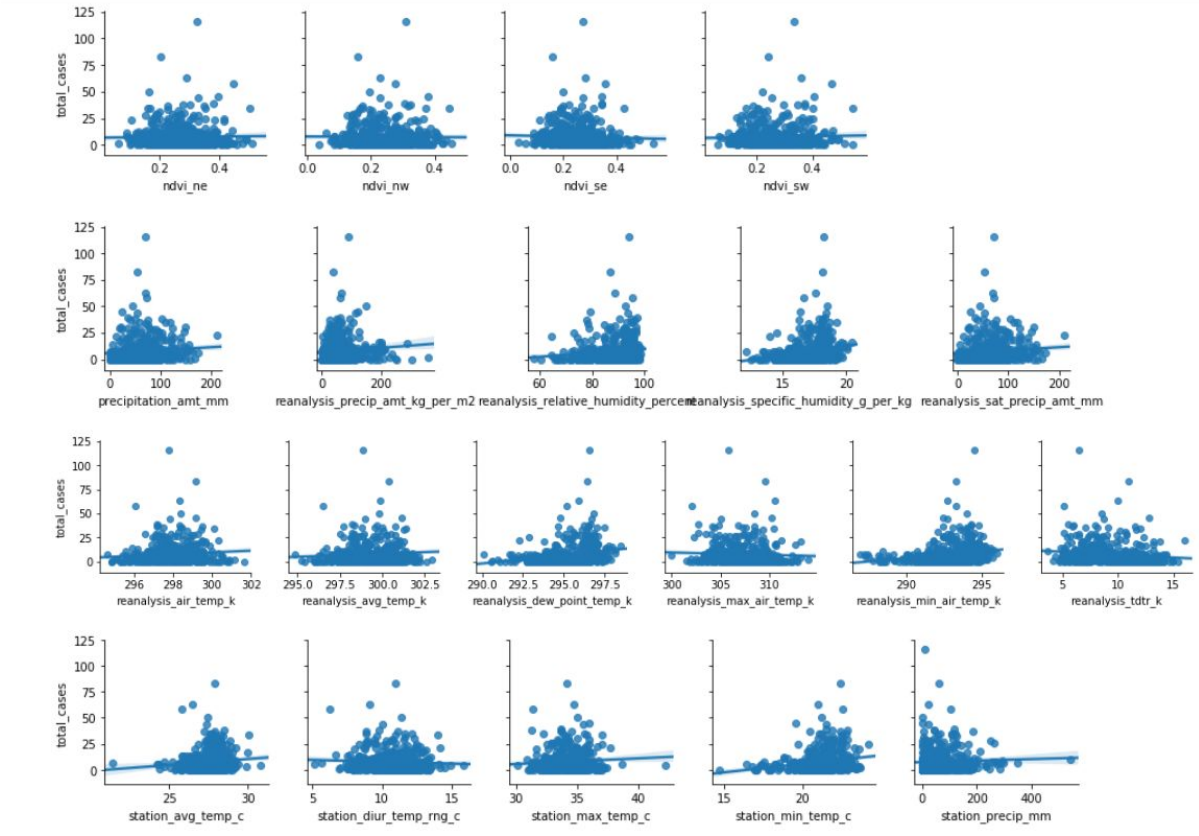


Figure 2 Iquitos Pair Plot

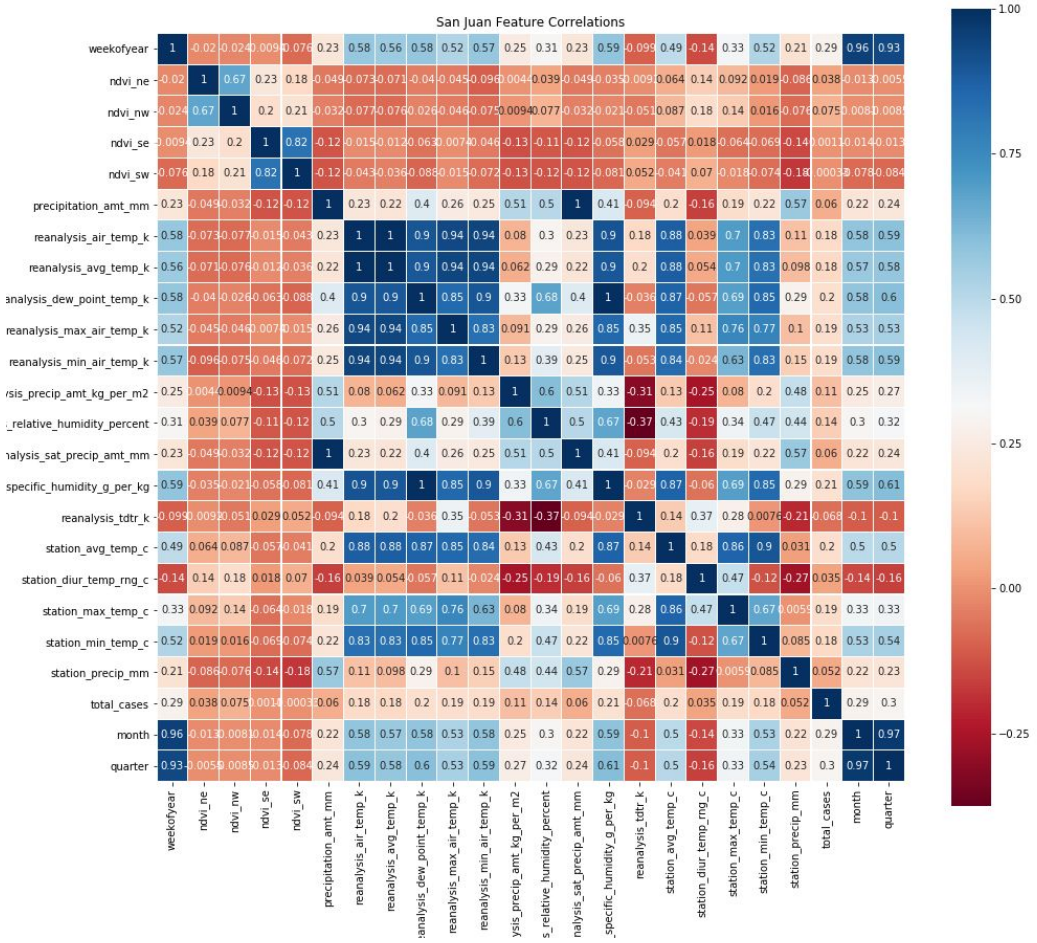


Figure 3 San Juan Feature Correlation

San Juan data

- Features “precipitation_amount_mm” and “reanalysis_sat_precip_amt_mm” were found to be 100% correlated.
- Features “reanalysis_dew_point_temp_k” and “reanalysis_specific_humidity_g_per_kg” were found to be 99.85% correlated.
- Features “reanalysis_avg_temp_k” and “reanalysis_air_temp_k” were found to be 99.75% correlated.
- Features “reanalysis_max_air_temp_k” and “reanalysis_avg_temp_k” were found to be 93.89% correlated.
- Features “reanalysis_min_air_temp_k” and “reanalysis_avg_temp_k” were found to be 93.91% correlated.
- Features "reanalysis_dew_point_temp_k" and “reanalysis_air_temp_k” were found to be 90.33% correlated.

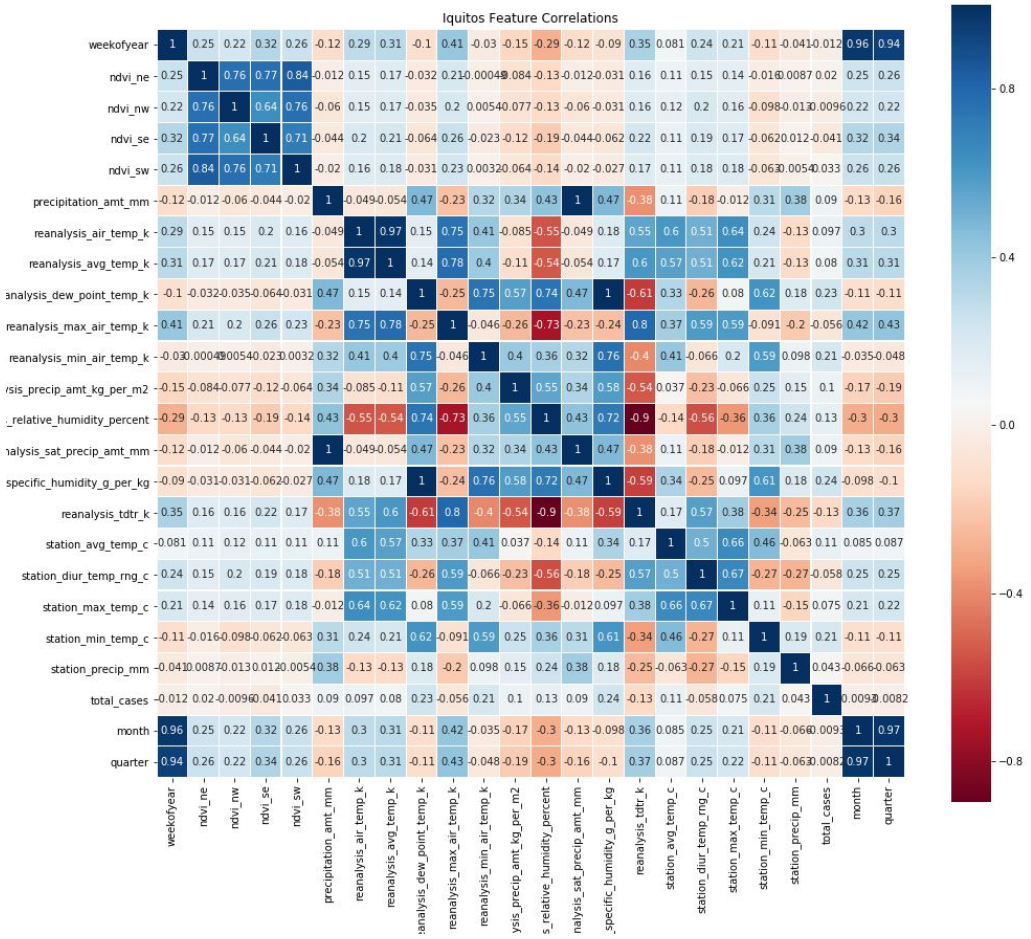


Figure 4 Iquitos Feature Correlation

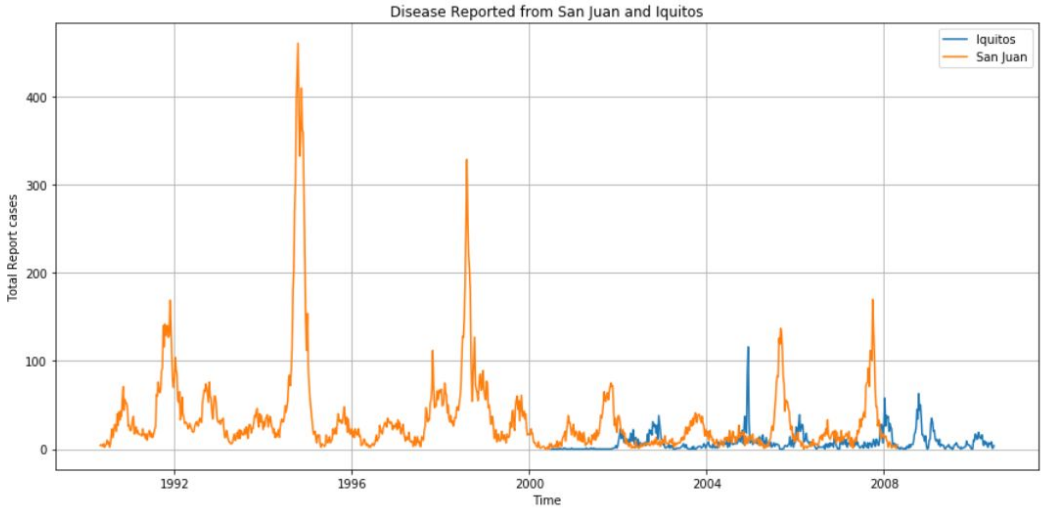
As per the feature correlation analysis, some features are correlated more than 90%.

Iquitos data

- Features “precipitation_amount_mm” and “reanalysis_sat_precip_amt_mm” were found to be 100% correlated.
- Features “reanalysis_dew_point_temp_k” and “reanalysis_specific_humidity_g_per_kg” were found to be 99.77% correlated.
- Features “reanalysis_avg_temp_k” and “reanalysis_air_temp_k” were found to be 97.33% correlated.

Data Visualization

Dengue reported cases over time



- As per the graph, there is dengue outbreak happened in 1991-92, 1994, 1998, 2007-08.
- In Iquitos, the reported cases are less compared to San Juan. The max we can see is in 2005-2006, 2007-2008.
- There is no yearly pattern for this data.
- As per the online media data analysis, during the high epidemic, DEN-2 was the predominant serotype isolated, followed by DEN-4 and DEN-1. But based on the weather features, it is difficult to identify the reason behind the DEN-2 predominant scenario.

Year wise pattern of each city

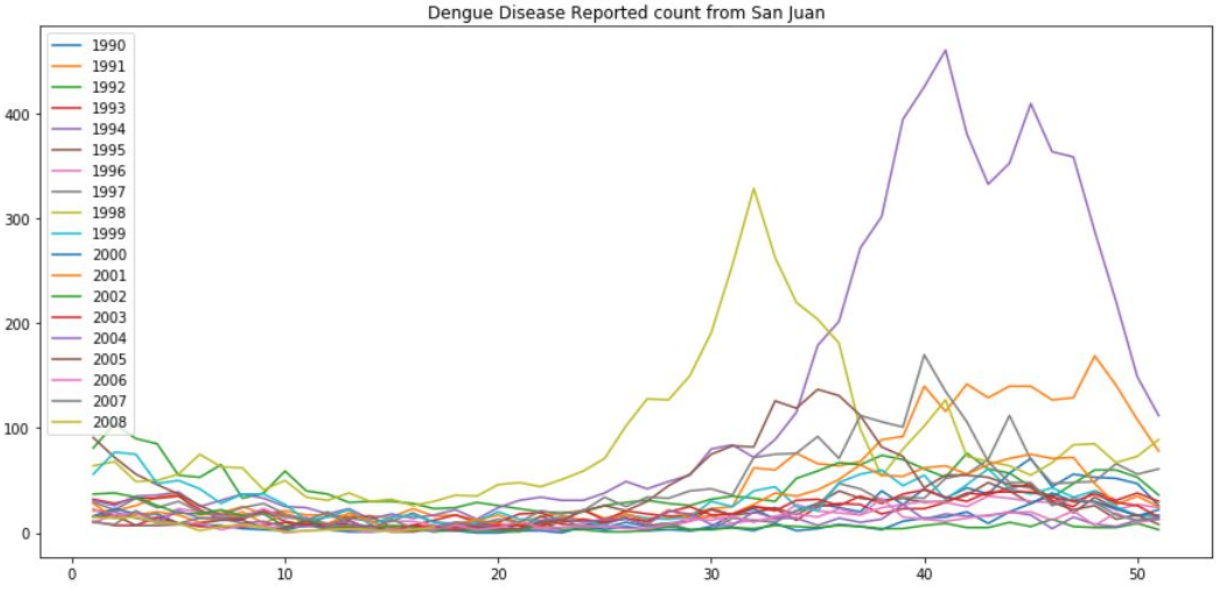


Figure 5 Dengue year wise pattern (San Juan)

- As per the graph, more number of dengue cases are reported from week 35 to end of year. And it extends to 5th week of next year.

- But the pattern is not similar in all the year.

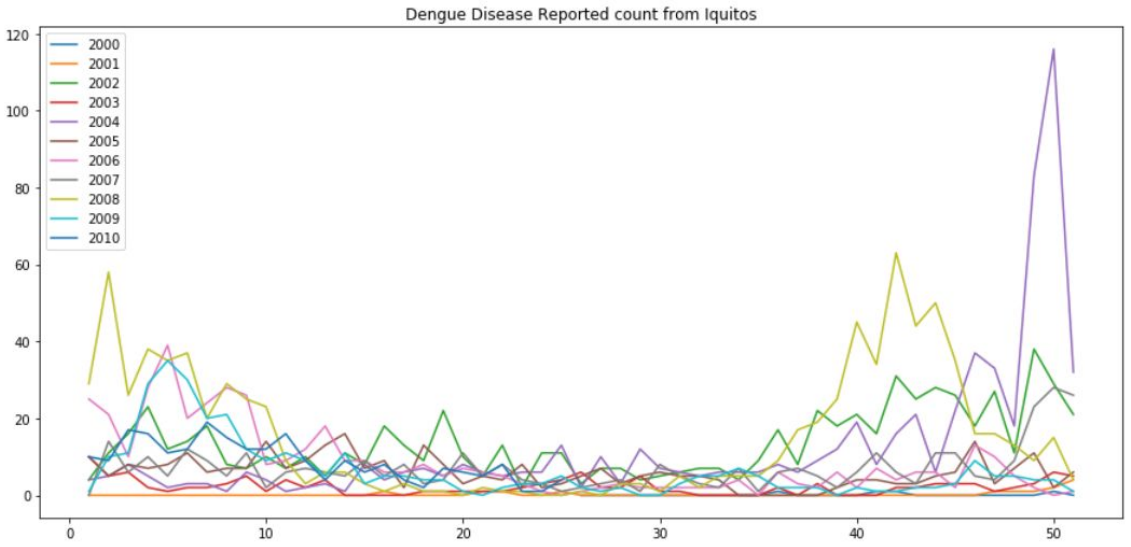


Figure 6 Dengue year wise pattern (Iquitos)

Looks like the San Juan analysis details are applicable to Iquitos pattern too.

Week wise pattern of each city

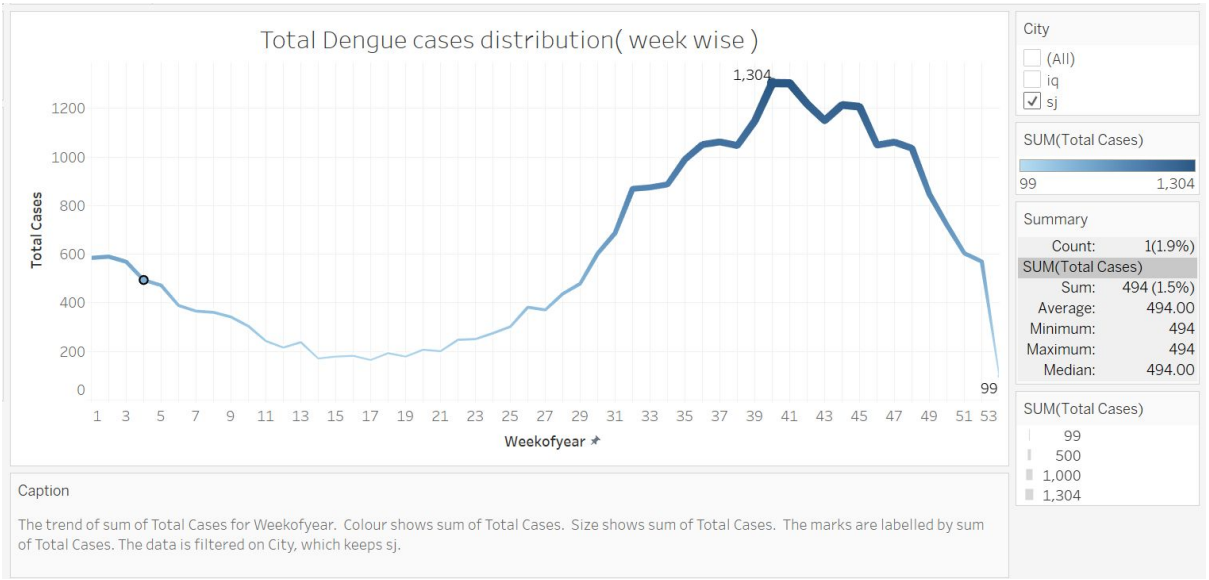


Figure 7 week wise distribution (San Juan)

- As per the graph, less number of dengue cases were reported from week 7 to 25.
- Week 53 is not applicable in all years, that's the reason for the low value
- And maximum cases were reported in week 40 and 41.

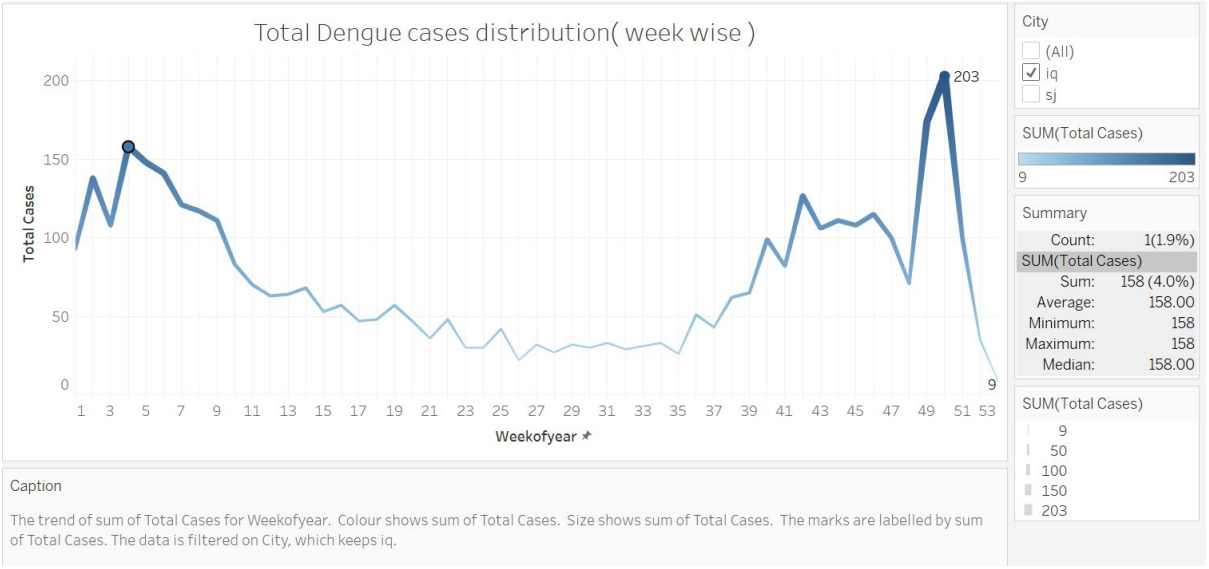


Figure 8 Total Dengue cases week wise distribution Iquitos

- Iquitos week wise pattern is not same as San Juan. We can see a dip in dengue case reported in 48th week. And the maximum is in 50th week.
- Week 53 is not applicable in all years, that's the reason for the low value

Month wise pattern of each city

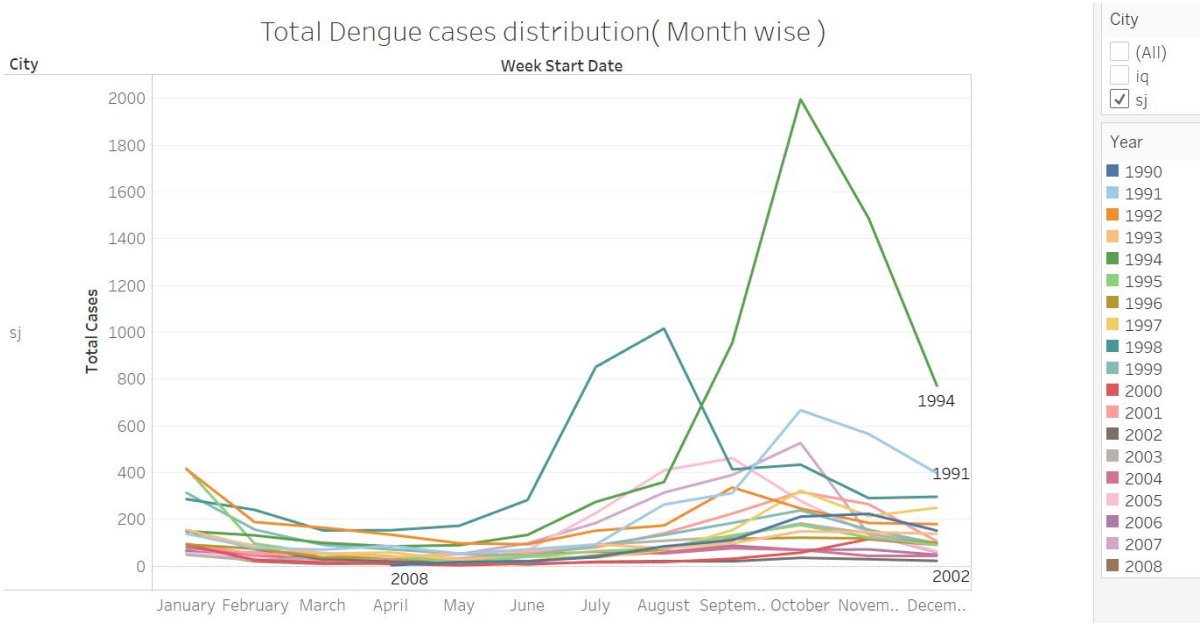


Figure 9 Month wise dengue case distribution

- From February to August, the dengue cases reported are less and stable.
- In some years, peaks are present because of specific dengue DEN-2 spread.

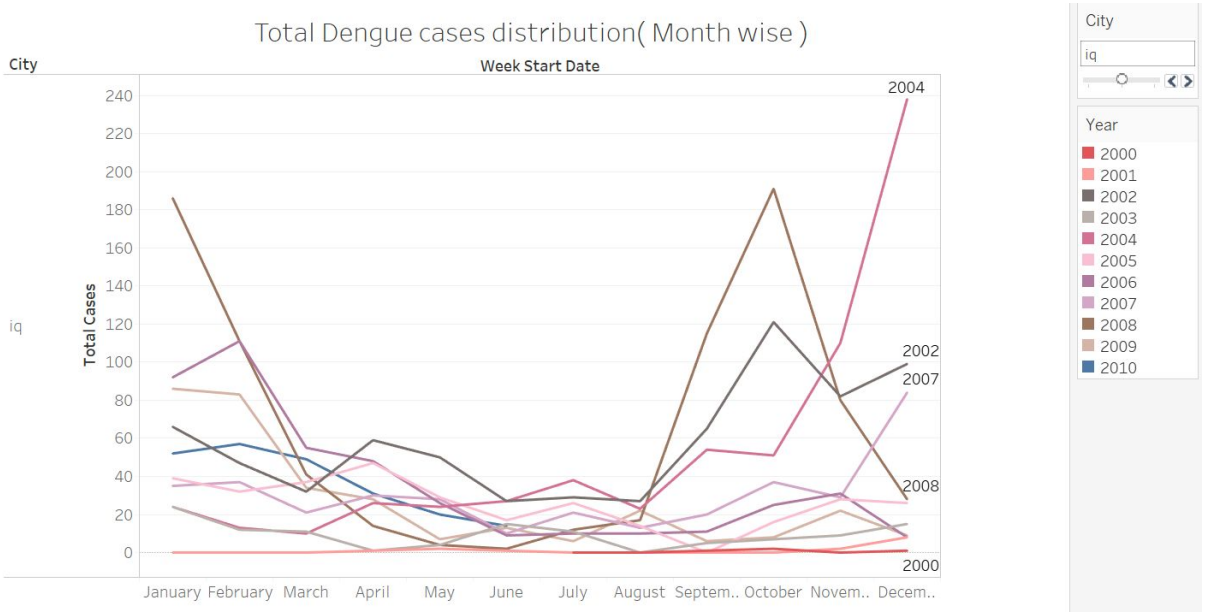
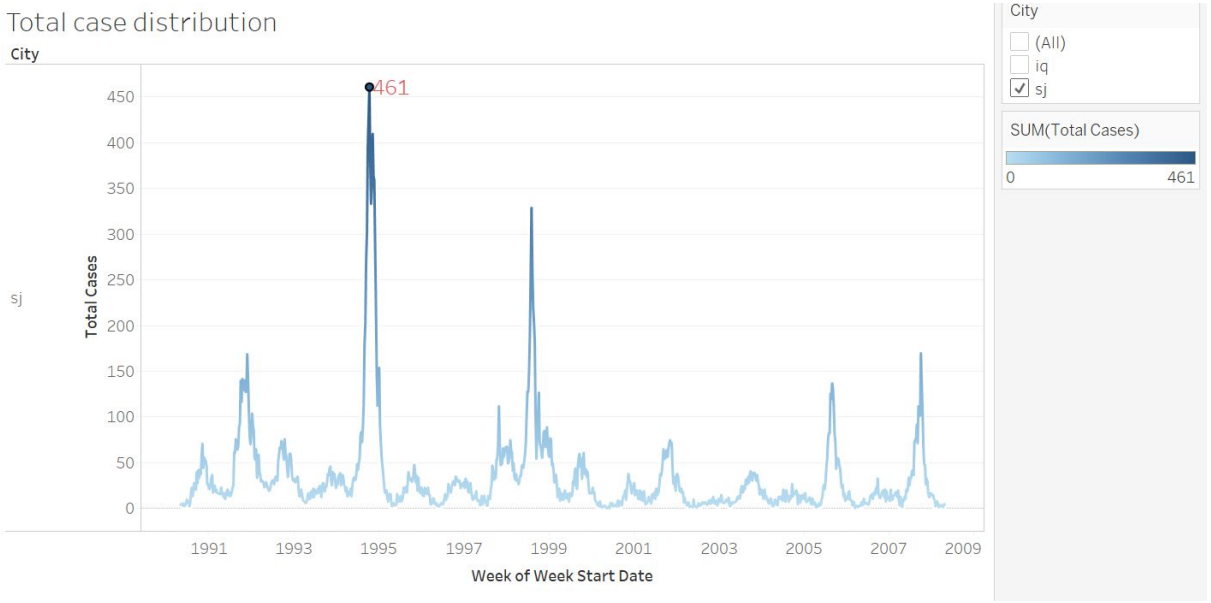


Figure 10 Month wise dengue case distribution(Iquitos)



Filling Missing Values

As per the missing data analysis, the missing features are not continuous in the data set. It is less than 15% of entire column. So it was decided to fill the missing values instead of dropping the entire column. The feature values depend on the target variable. So filling with mean value is not an appropriate solution. As there are no missing entries corresponding to date, there is no need to interpolate with time. Direct interpolation is enough to fill the missing data. It was tried with forward fills and backward fills too. But interpolate filling gave good result compared to all.

Merge Target Variable with train data set

The features city, year, week of year in dengue_features_train.csv and dengue_labels_train.csv are matching while comparing. Based on this the target variable is merged with feature data set. to get better understanding about the feature relationship with the target variable, added total_cases column to the main dataframe.

Unit conversion (Kelvin to Celsius)

To set same type of units for all similar features, Temperatures in Kelvin are converted to Celsius using the equation: Temperature Celsius = Temperature Kelvin – 273.15. The column names are updated after conversion.

New Feature Exploration

- Month and quarter are identified from date field. In most of the models, the quarter field is irrelevant. But in some cases, the model selected quarter as an important feature. Considering these features as categorical did not add value to the model performance, so it was decided to go with integer data type.

- New features from Satellite vegetation - Normalized difference vegetation index (NDVI): Statistical descriptive analysis of ndvi_se, sw, nw and nw, helped to get an understanding about the dispersed behaviour of ndvi features. So a mean variable is generated using all the four features.

- From the disease information, there got an understanding about the mosquito life cycle and weather factors affecting dengue fever. So the weather factors of the earlier weeks were also considered to identify the actual reason behind the disease spread. For that, created different rolling window with lag 1 to 6(1 week to 1.5 month) and tried the basic models to confirm the window lag. As per the trials, the four weeks rolling features for San Juan got good results. And for Iquitos, five weeks rolling features are good. New rolling window features are created using aggregate functions Sum, Mean and Variance.

- With the existing and new features, the feature count reach to 86.

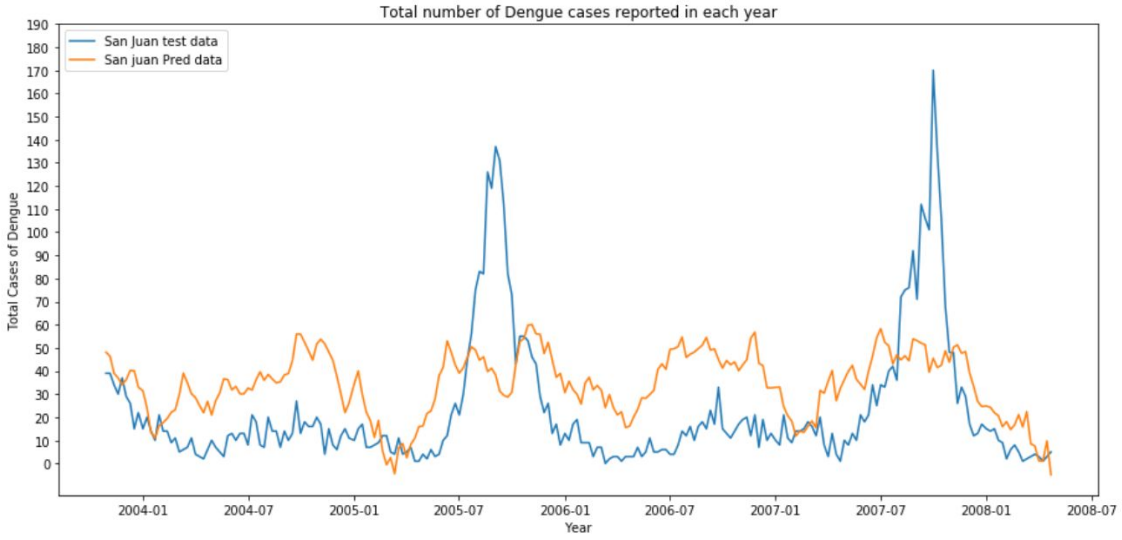
Separated the dataset for each city

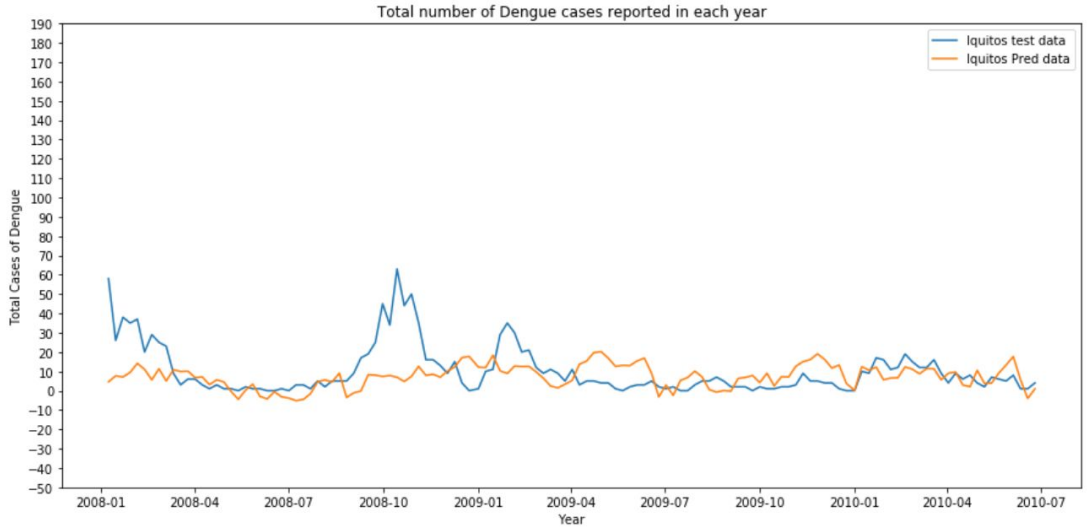
Separated San Juan and Iquitos data to produce two data sets, one to predict weekly dengue fever cases in San Juan and the other in Iquitos. This is because weather and vegetation behave differently in relation to time between both locations, because these locations are separated by distance, climate, population, and ecosystem.

Model Evaluation

Linear Regression

As per the target variable distribution and the correlation of features with the target variable, the linear regression is not a recommended method. But still to identify the simplest method's result, linear regression with regularization Lasso and Ridge were tried. FFS (Forward Feature Selection) and RFE (Reverse Feature Elimination) are used for the feature selection and model comparison.





Model	City	MAE	Comments
Linear Regression	San Juan	27.09	
Linear Regression	Iquitos	9.11	
Linear Regression with FFS	San Juan	25.78	
Linear Regression with FFS	Iquitos	5.59	Good result
Linear Regression with RFECV	San Juan	28.06	
Linear Regression with RFECV	Iquitos	7.80	
Ridge L2 Regularisation	San Juan	28.59	
Ridge L2 Regularisation	Iquitos	7.88	
Lasso L1 Regularisation	San Juan	22.82	Good Result
Lasso L1 Regularisation	Iquitos	8.02	

Negative binomial regression

Negative binomial regression is a type of generalized linear model in which the dependent variable is a count of the number of times an event occurs. In this case, the target variable is a count of dengue cases reported in each week. And the data has high variance compared to mean, and dispersed data, so this regression model is matching for the data set.

Dataset	City	MAE
Train	San Juan	24.98
Train	Iquitos	5.34
Test	San Juan	19.27
Test	Iquitos	7.97

Poisson Regression

The data is count type of distribution, so we are trying Poisson model too. As per the statistical description of mean and variance, the data is dispersed. If the data is dispersed, this model is not good for prediction. Still this model was used to study the result.

Dataset	City	MAE
Train	San Juan	26.28
Train	Iquitos	5.35
Test	San Juan	19.51
Test	Iquitos	7.98

Compared to Poisson model, Negative Binomial model is good based on MAE (Mean Absolute Error). But still Poisson model is good compared to linear regression models.

Linear Regression (OLS)

Using stats model implementation, OLS (Ordinary Least Squares) regression model was tried, even though the linear regression was done earlier.

Dataset	City	MAE
Train	San Juan	28.18
Train	Iquitos	4.80
Test	San Juan	21.70
Test	Iquitos	8.72

Log transformation and normalisation of variables also were tried, but there was no improvement in the result.

Time Series

Before analyzing the data in time series mode, it was verified whether the time series is stationary or not.

Check whether the data Time series is stationary or not?

- As per data plot (based on year), there is no trend or seasonality.
- As per data plot (based on month), can observe a change in trend.
- As per the summary statistics of the data, the mean and standard deviation are not same for each year's data. But in this case time series is not stationary.

Augmented Dickey-Fuller test

A statistical test was designed to explicitly comment on whether a univariate time series is stationary or not. The Augmented Dickey-Fuller test is a type of statistical test called a unit root test. The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend.

- Null Hypothesis (H0): it is non-stationary.
- Alternate Hypothesis (H1): it is stationary

This result was interpreted using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests rejecting the null hypothesis (stationary). A p-value above the threshold suggests we fail to reject the null hypothesis (non-stationary).

- As per the output, it can be seen that our statistical value of -5 is less than the value of -3.449 at 1%.
    - This suggests that we can reject the null hypothesis with a significant level of less than 1%.
    - Rejecting the null hypothesis means that the time series is stationary or does not have time-dependent structure.

Autoregressive Integrated Moving Average ARIMA(p, d, q) Model

Different values of p, d and q are tried. Got the below result for 3, 1, 1.

Dataset	City	MAE
Train	San Juan	20.12
Train	Iquitos	7.13
Test	San Juan	22.24
Test	Iquitos	9.71

The results are not promising compared to other models.

Random Forest

In this bagging model, tuned the hyper parameters using GridSearchC and got a good result compared all other linear models.

Dataset	City	MAE
Train	San Juan	21.73
Train	Iquitos	4.57

Test	San Juan	17.22
Test	Iquitos	7.37
Train	Total	15.62
Test	Total	13.72

XGBoost

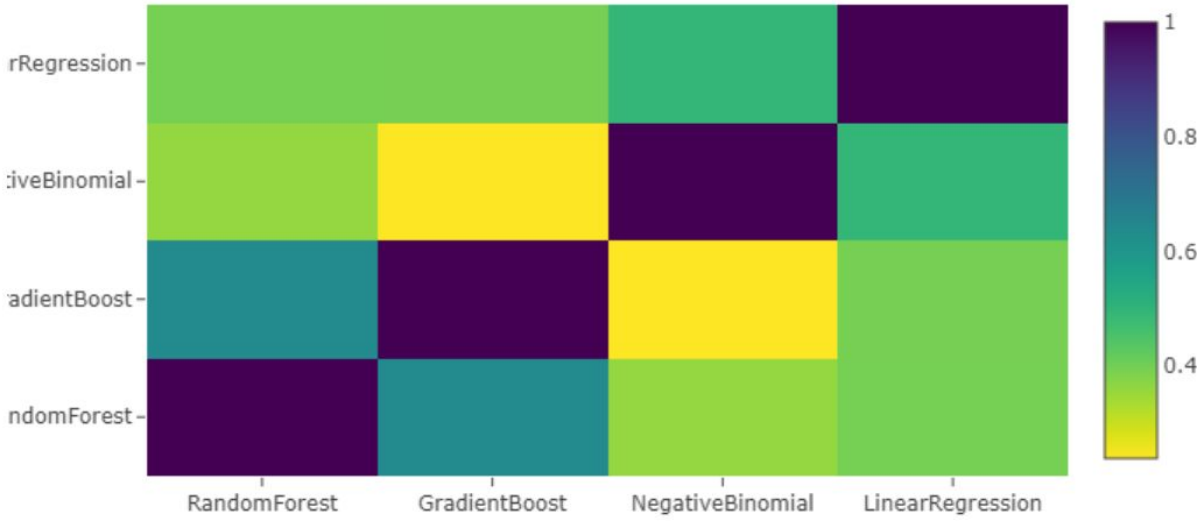
In this boosting model, tuned the hyper parameters using GridSearchCV and got a good result compared all other linear models.

Dataset	City	MAE
Train	San Juan	1.08
Train	Iquitos	2.75
Test	San Juan	15.35
Test	Iquitos	4.86
Train	Total	1.67
Test	Total	11.63

Stacking Ensemble

Stacking ensemble is a learning technique that combines multiple regression models through a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features. In the base level model following are the regressors considered.

1. Linear Regression
2. Negative Binomial
3. Random Forest
4. XGBoost



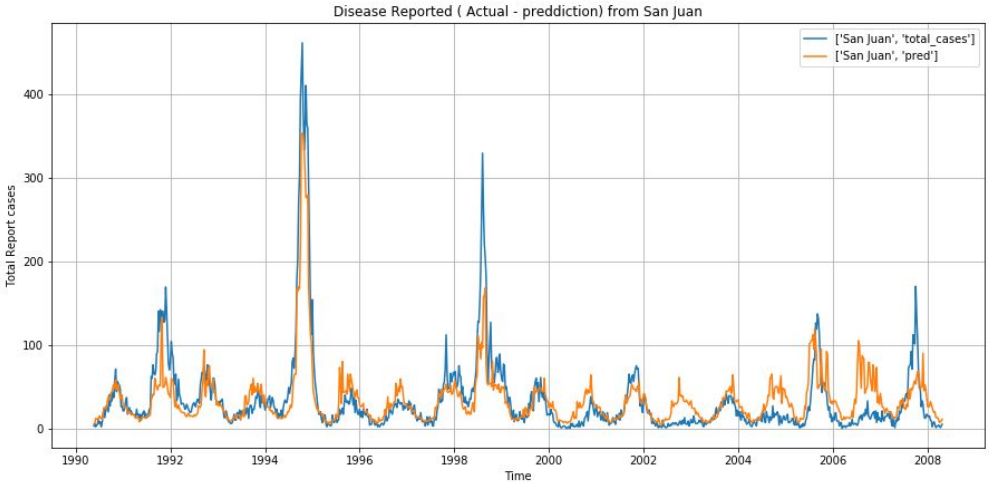
The above picture depicts the correlation between the first level model's output. Due to the different levels of correlation, the ensemble gave a good result. The meta-model considered for this

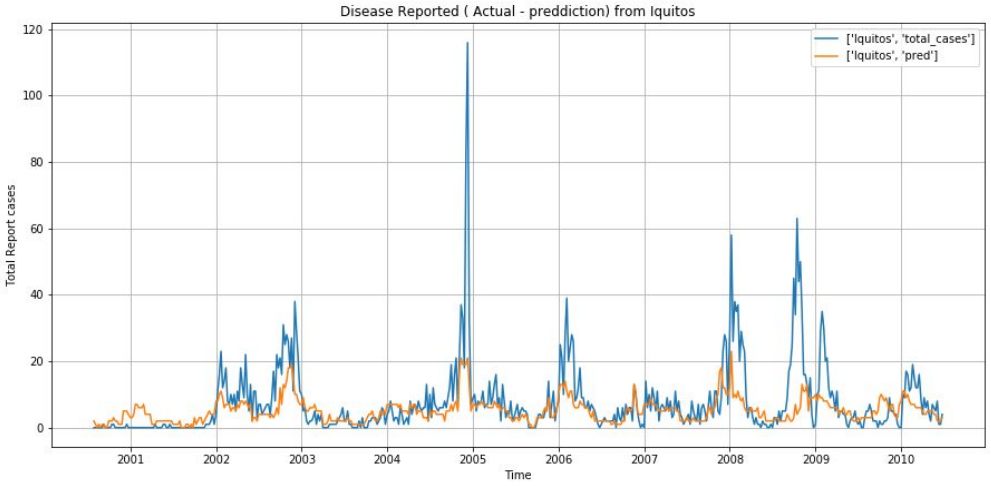
is XGBoost regressor. Then the output of base level models is considered for second level modelling as features.

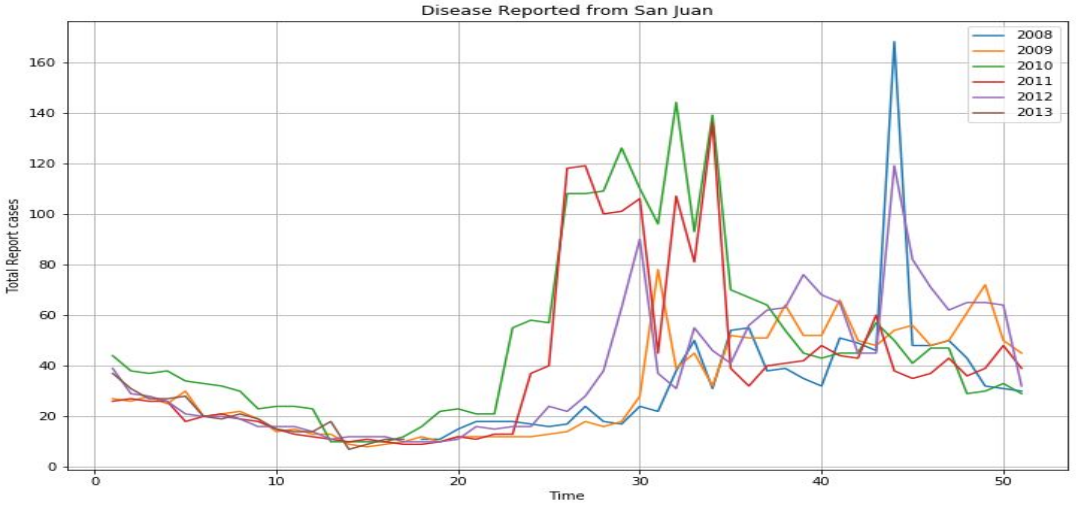
Dataset	City	MAE
Train	San Juan	16.73
Train	Iquitos	24.83
Test	San Juan	7.28
Test	Iquitos	4.59

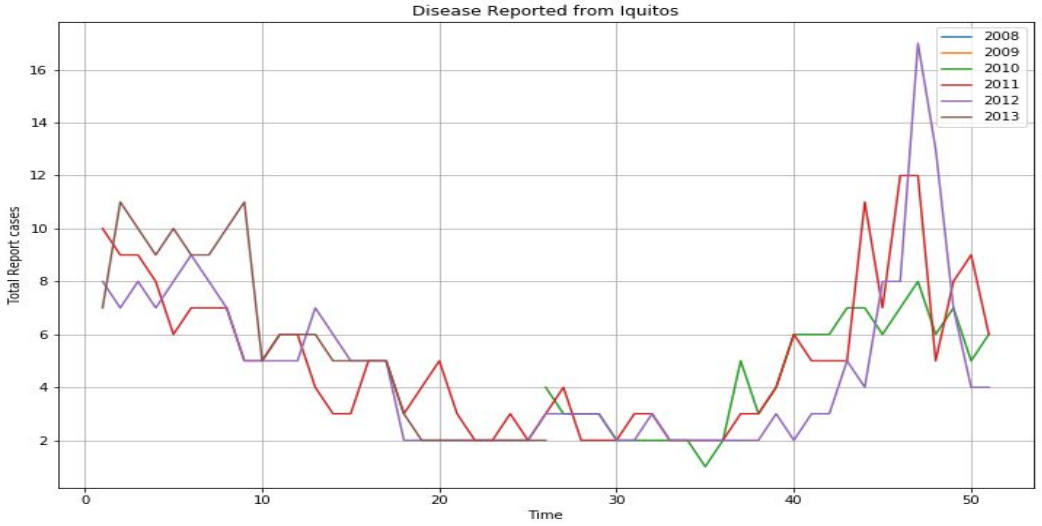
Conclusion

The best predictive models for cases of dengue fever in San Juan, Puerto Rico and Iquitos, Peru were found using current weather-vegetative features and weather data from the previous weeks too(lagged data) as explanatory variables using random forest model. Expected Stacking model didn't improve as much. Each of the models had very similar scores ranging from 16 to 25(for San Juan) and 4 to 8 (for Iquitos). The only major difference created was separating the cities datasets during the initial stage.




Disease Reported (Actual - preddiction) from Iquitos
['Iquitos', 'total_cases']
['Iquitos', 'pred']
Total Report cases
Time


Disease Reported from San Juan
2008
2009
2010
2011
2012
2013
Total Report cases
Time


Disease Reported from Iquitos
2008
2009
2010
2011
2012
2013
Total Report cases
Time

References

1. https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06
2. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
3. https://www.kaggle.com/mmueller/allstate-claims-severity/stacking-starter/run/390867

Issues Faced

- Over fitting the data - Got good validation scores in training, but same performance is not repeated in competition submission.
- As per the media, there is dengue outbreak reported in San Juan 2010-2012. But models are not predicted the outbreaks, but just increased the reported cases count compared to previous years.

Future Areas of Work

- Same window period was tried for all weather features. Instead of that, it should have been tried each lag window for each variable and fine tune it. Might be this may take time, but surely it will improve the model performance.
- Currently we modelled for entire city. Instead of that, if separate datasets for north, south, west and east are available for each city, more accurate prediction will come.
- MAE is only used as metric in this project for fine tuning. Other metrics like R2 is also need to be considered for modelling.