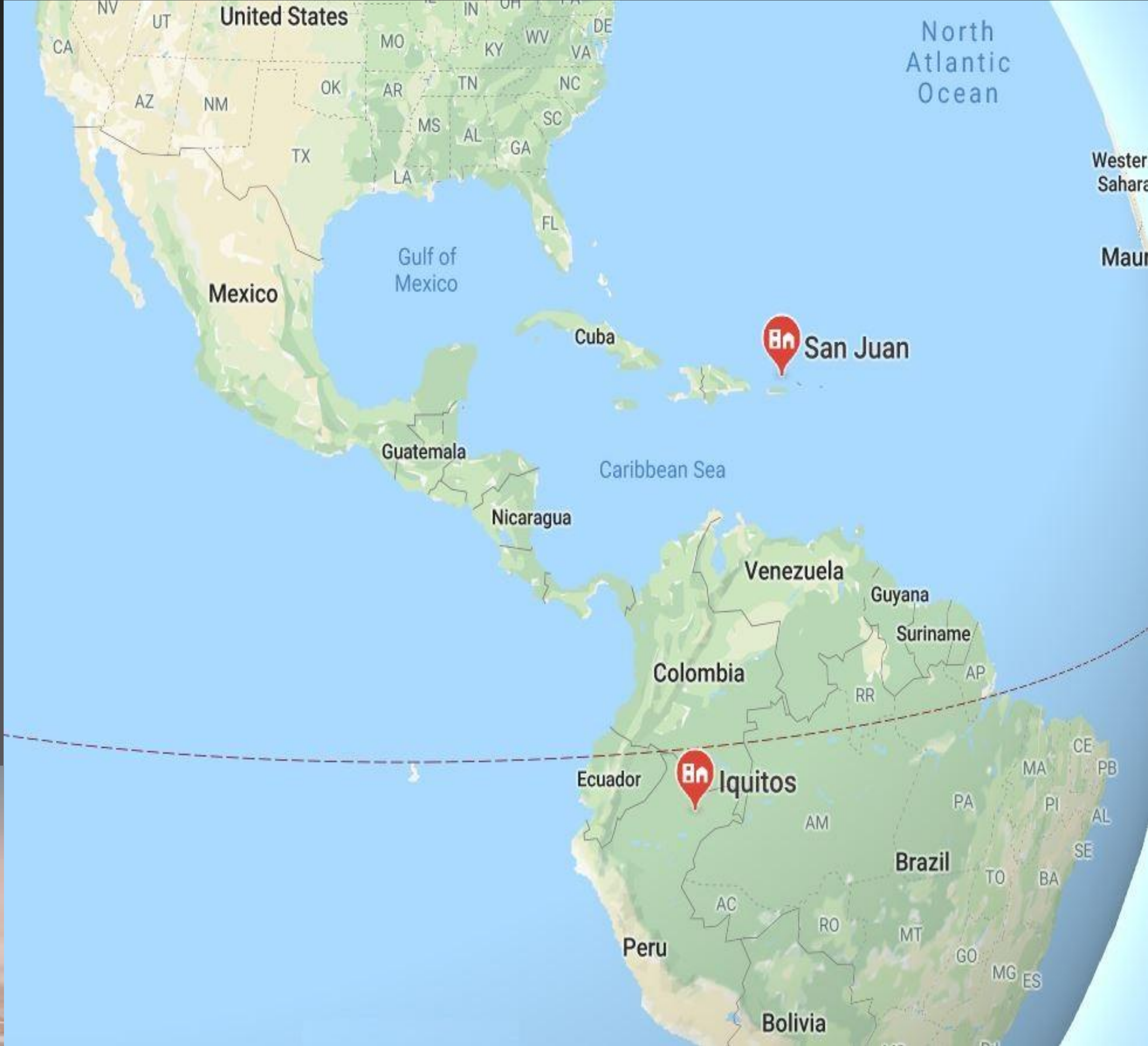
DENGUE PREDICTION

USING ENSEMBLING REGRESSION MODELS



Objective :

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.

Some known facts about Dengue :

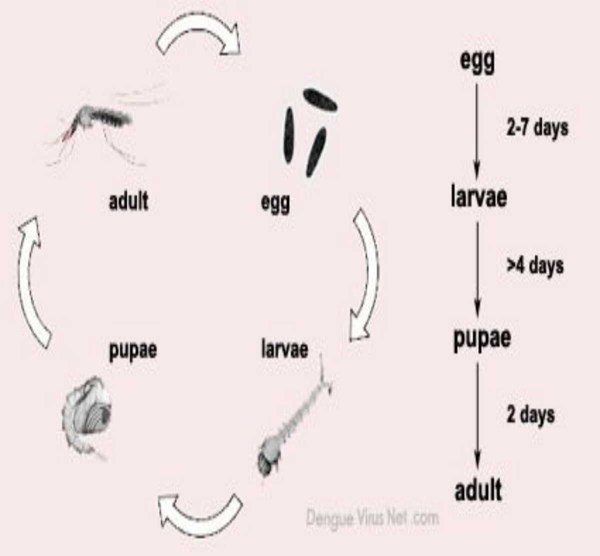
Dengue is a viral disease transmitted by Aedes types of mosquitoes. Mosquitoes become infected with dengue after biting sick humans who have dengue virus in their blood. Between 8 and 12 days later if an infected mosquito bites someone else it can pass on the dengue virus.

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.

Mosquito life cycle

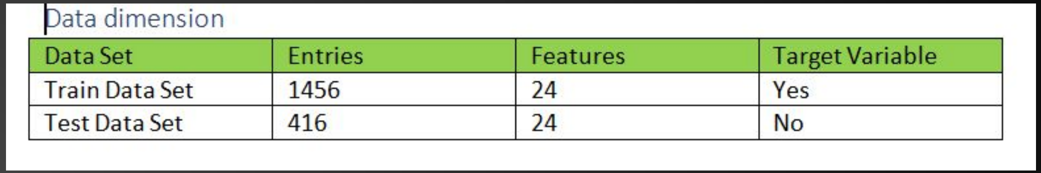
The life cycle of Aedes aegypti can be completed within one-and-a-half to three weeks. Male mosquitoes live three to five days. The females live considerably longer, depending on how much warmth and moisture is in their environment. Under ideal conditions, they may last as long as a month or two.



Datasets : The data has taken from DrivenData.

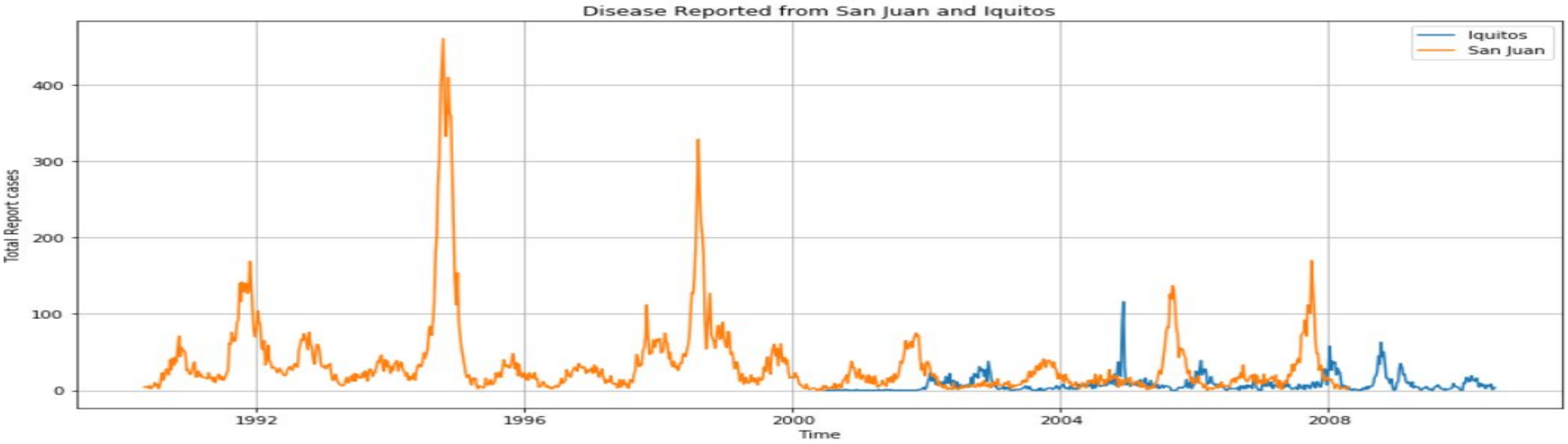
Meta Data:

- **city** – Two cities namely San Juan and Iquitos
- **year**- Years from 1990 to 2006
- **weekofyear** – 52 weeks per year
- **reanalysis_precip_amt_kg_per_m2-** precipitation amount
- **reanalysis_specific_humidity_g_per_kg-** humidity amount
- **reanalysis_avg_temp_k.** - average air temperature
- **reanalysis_max_air_temp_k-** Max air temperature
- **reanalysis_min_air_temp_k**- Min air temperature
- **total_cases**- Total cases recorder per every week.

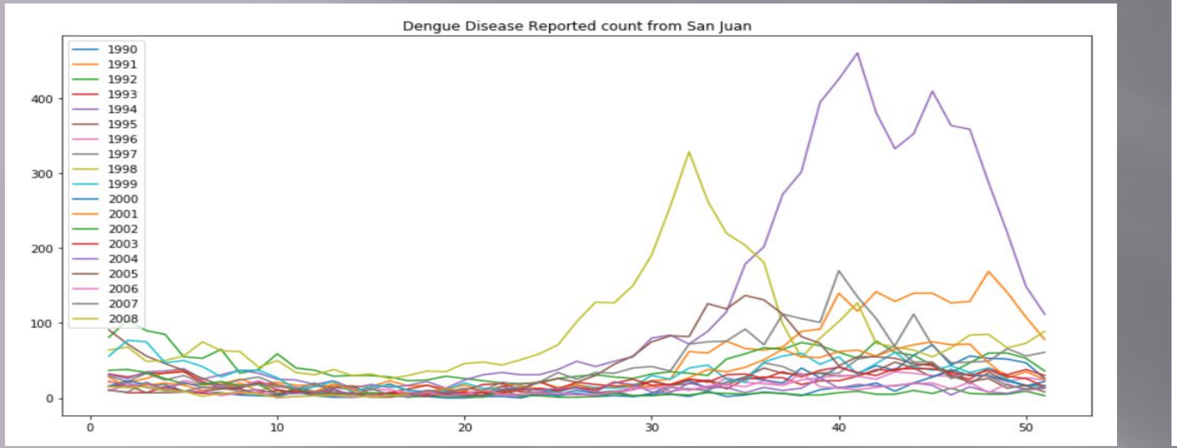


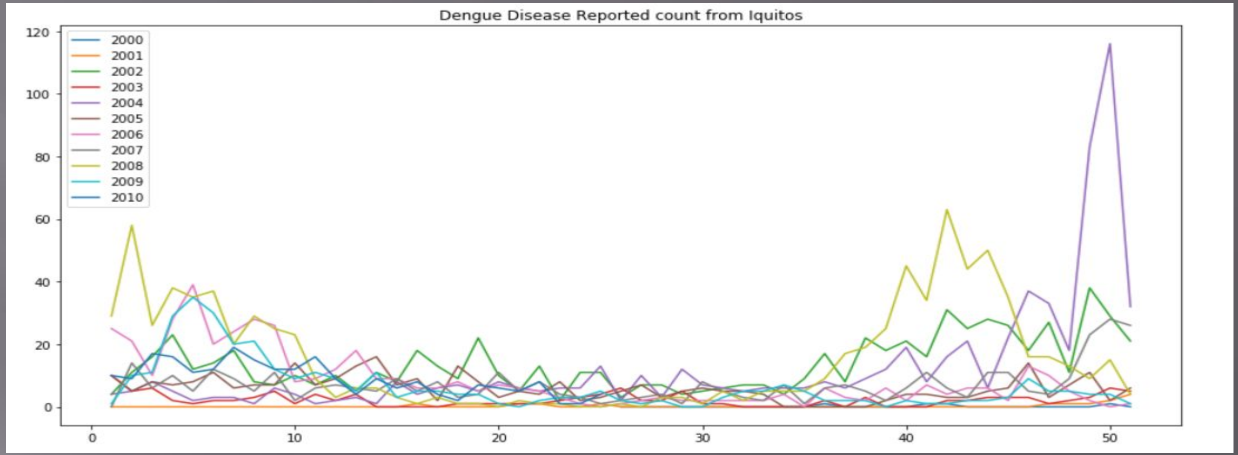
Data dimension

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

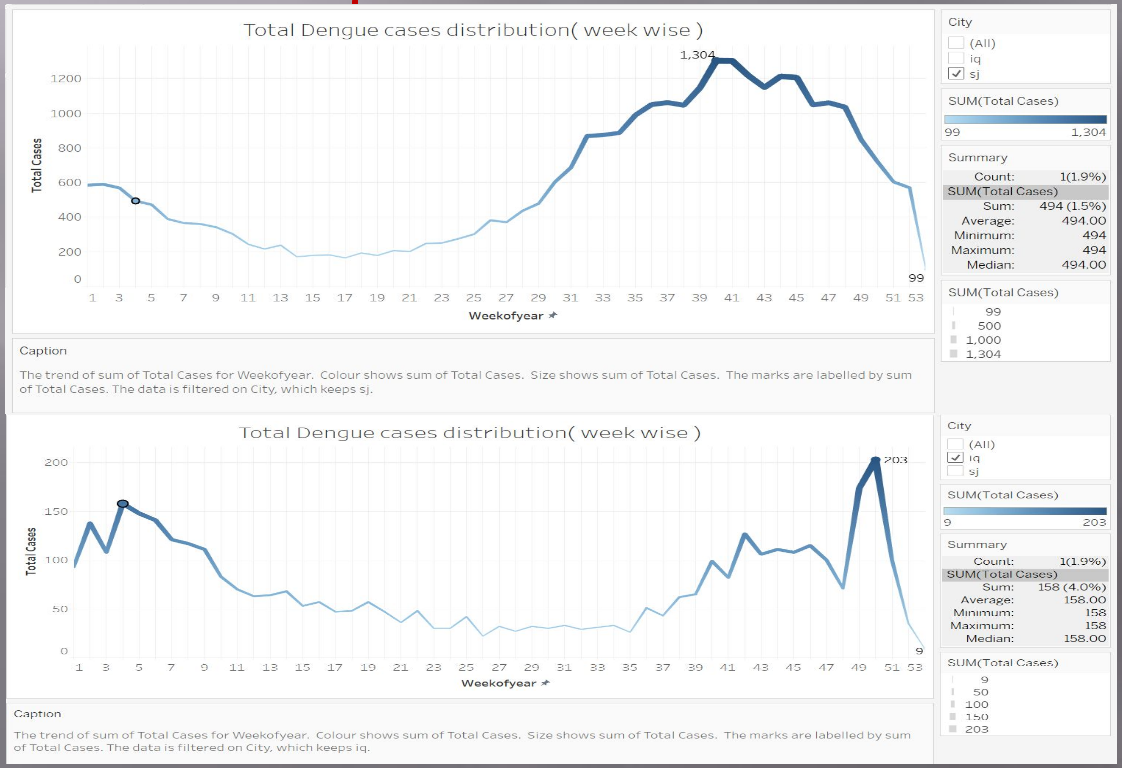


Year wise pattern





Week wise pattern



Summary

Count:	1(1.9%)
SUM(Total Cases)	
Sum:	494 (1.5%)
Average:	494.00
Minimum:	494
Maximum:	494
Median:	494.00

Caption

The trend of sum of Total Cases for Weekofyear. Colour shows sum of Total Cases. Size shows sum of Total Cases. The marks are labelled by sum of Total Cases. The data is filtered on City, which keeps sj.

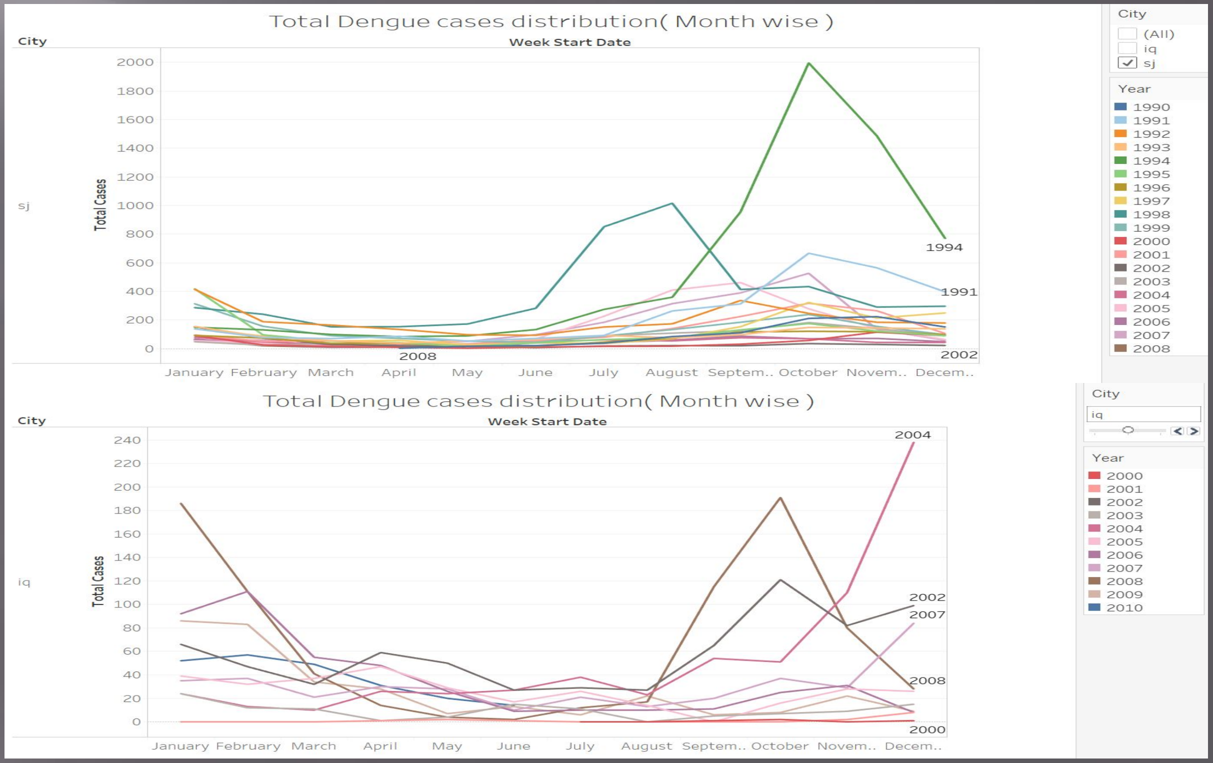
Summary

Count:	1(1.9%)
SUM(Total Cases)	
Sum:	158 (4.0%)
Average:	158.00
Minimum:	158
Maximum:	158
Median:	158.00

Caption

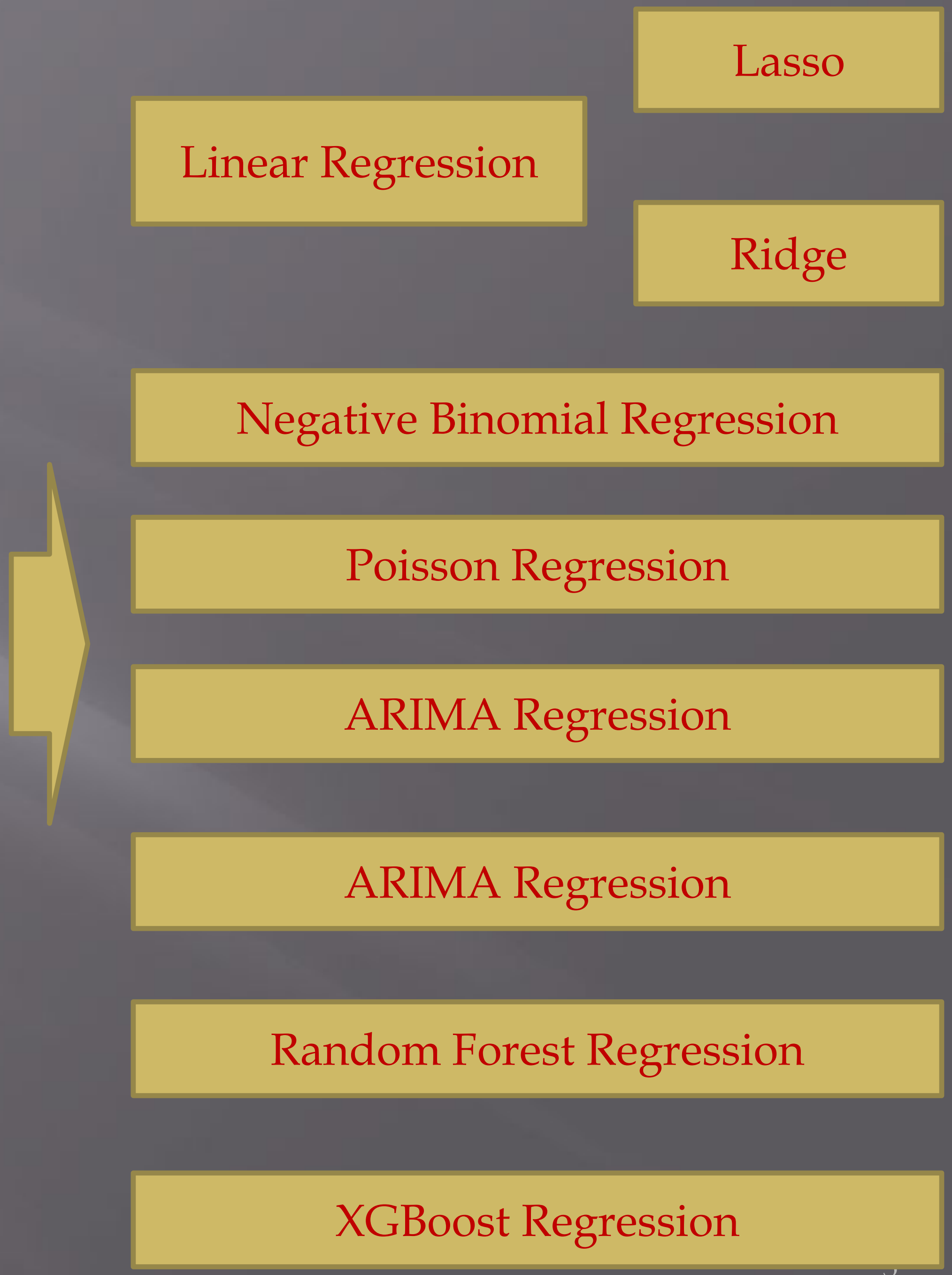
The trend of sum of Total Cases for Weekofyear. Colour shows sum of Total Cases. Size shows sum of Total Cases. The marks are labelled by sum of Total Cases. The data is filtered on City, which keeps iq.

Month wise pattern



Work Flow of individual modelling

- Data Exploration(data format, dimension, missing values, duplicate entries, descriptive analysis, correlation ...)
- Filling Missing Entries (interpolate), Unit-Conversion
- Data Visualization(Target variable distribution, year wise, month wise, week wise data distribution analysis)
- New Feature Exploration (Mean of vegetation , rolling lag window)
- Separated the dataset for each city



Linear Regression

Lasso

Ridge

Negative Binomial Regression

Poisson Regression

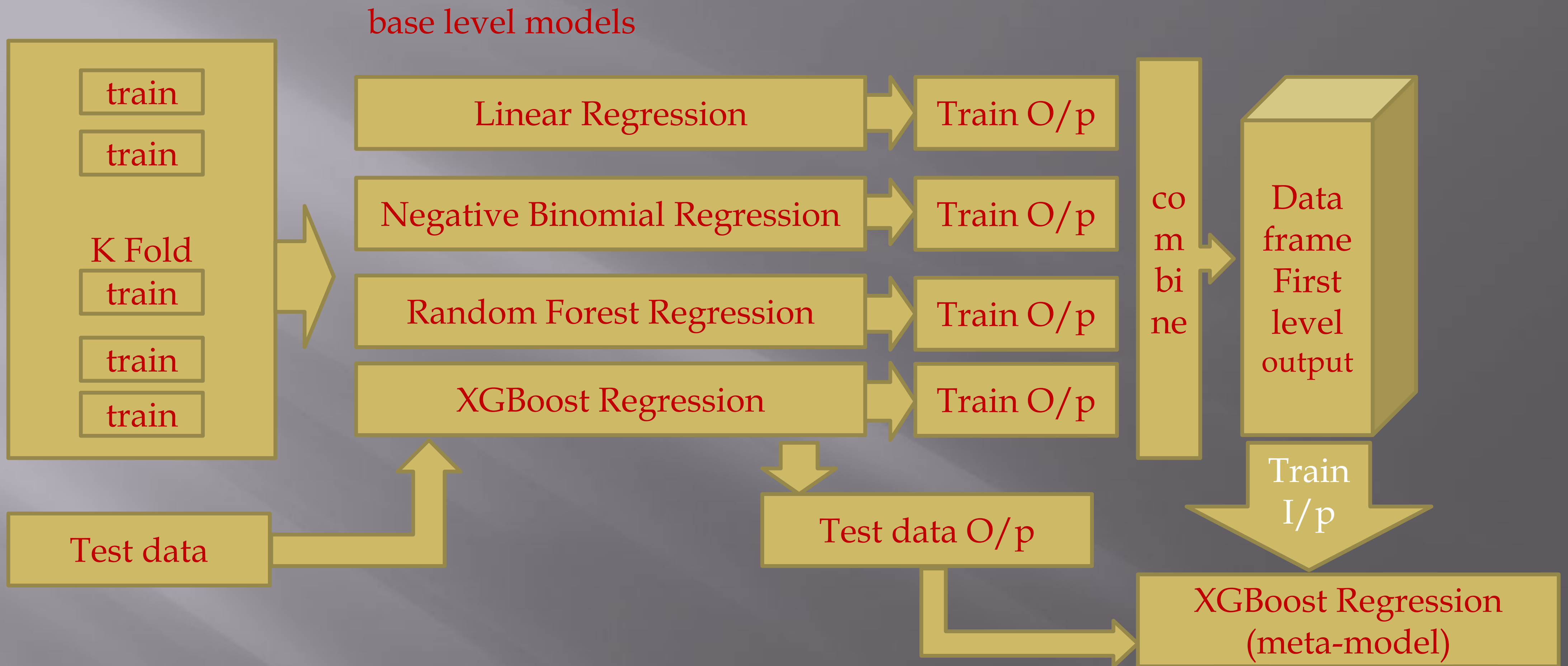
ARIMA Regression

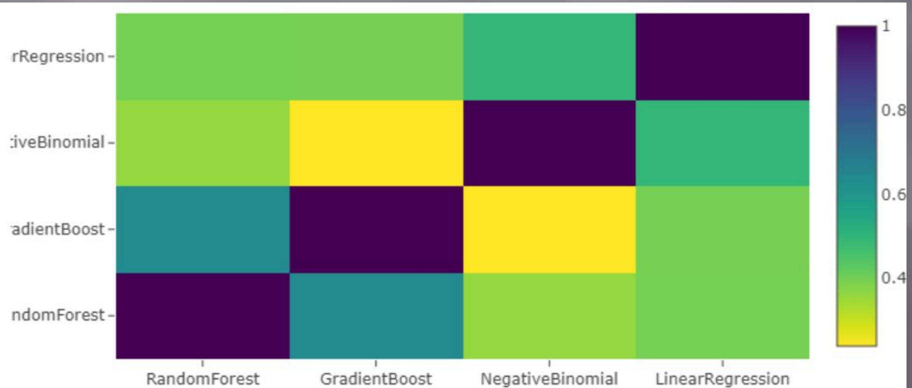
ARIMA Regression

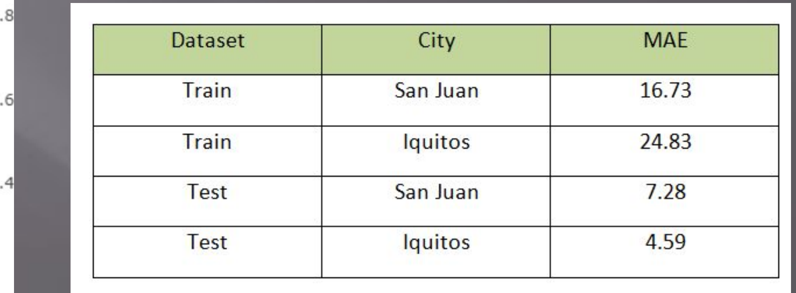
Random Forest Regression

XGBoost Regression

Stacking Ensemble

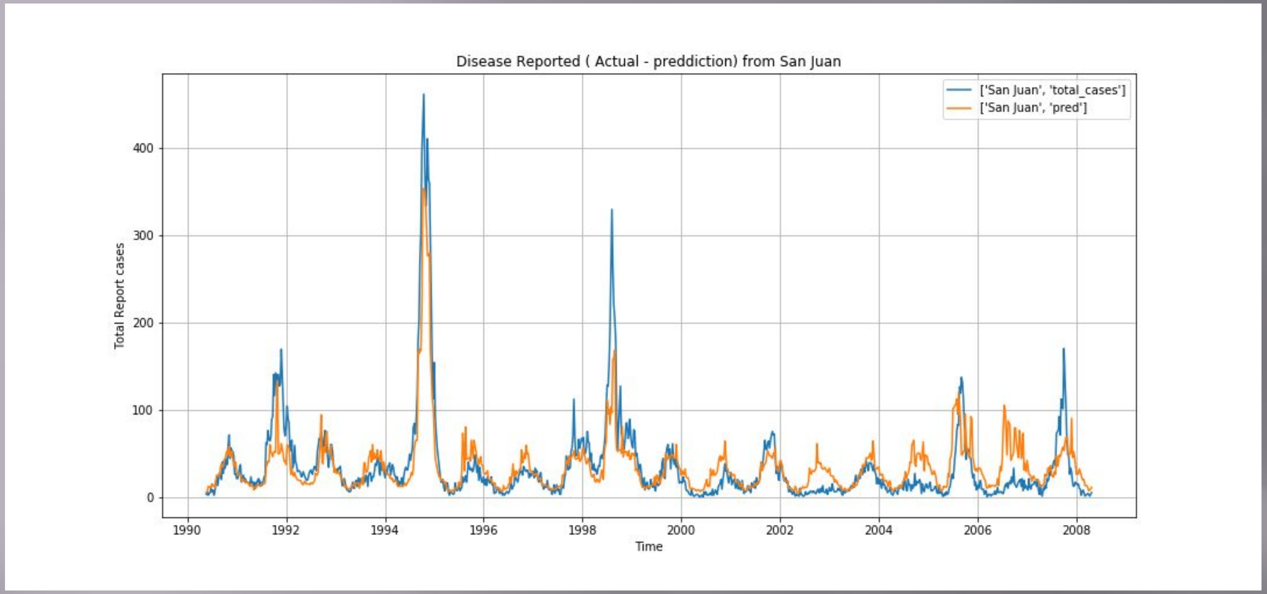




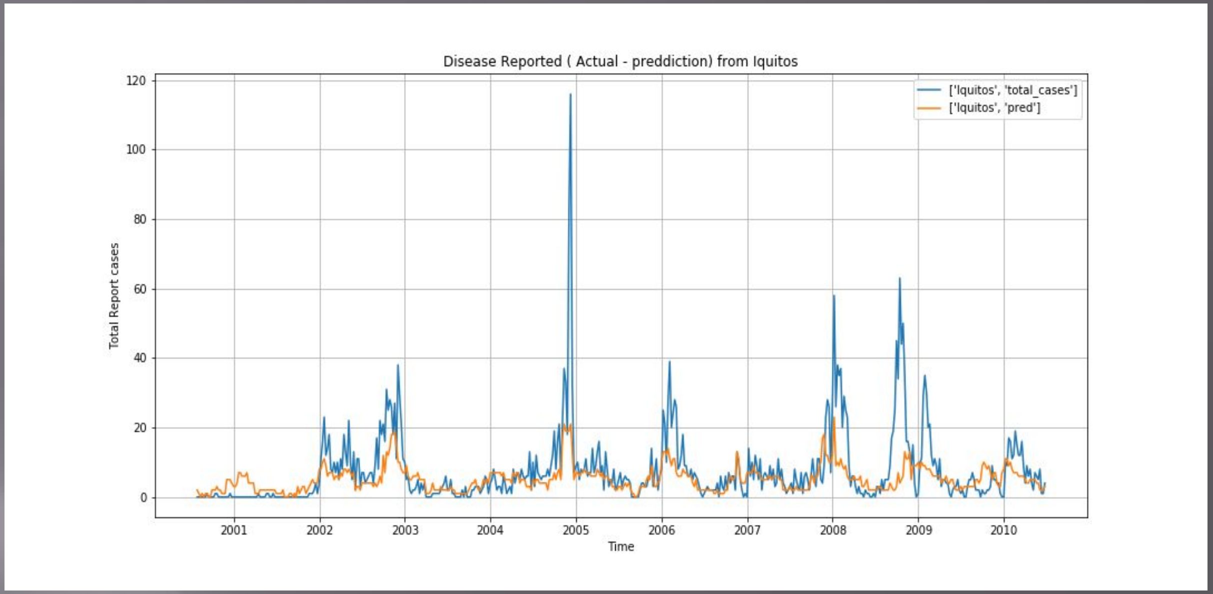


Dataset	City	MAE
Train	San Juan	16.73
Train	Iquitos	24.83
Test	San Juan	7.28
Test	Iquitos	4.59

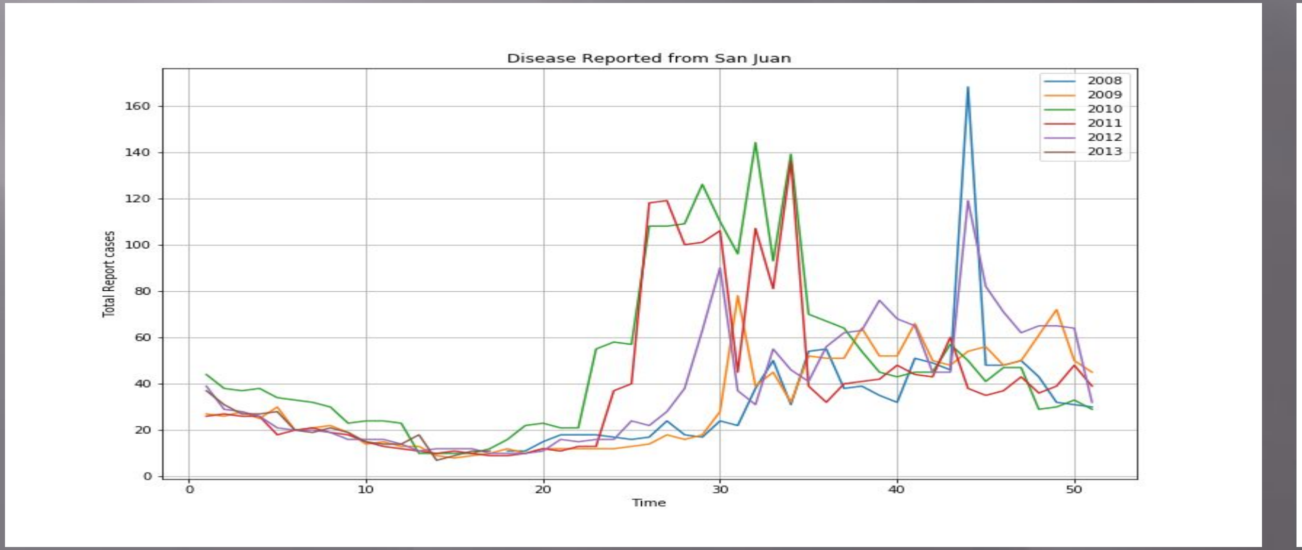
Best Model Output prediction



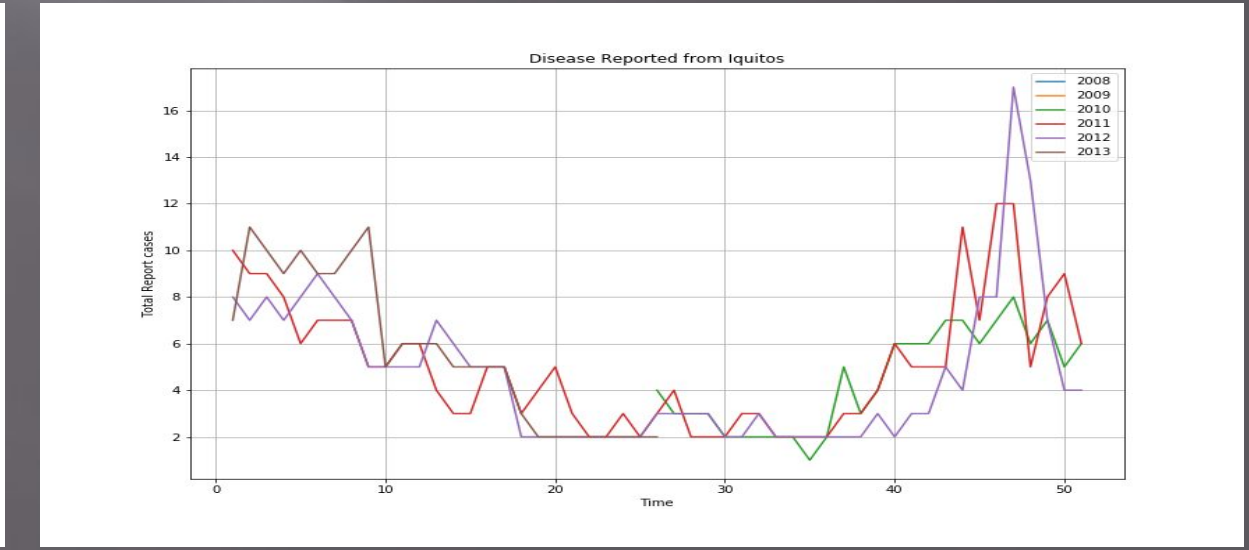
Disease Reported (Actual - preddiction) from San Juan



Disease Reported (Actual - preddiction) from Iquitos



Disease Reported from San Juan



Disease Reported from Iquitos

Issues Faced

- Overfitting 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.

Questions

Thank You