ISYE 6740 - Computational Data Analytics - Summer 2022
Dengue Epidemic Prediction

Krishna Kumar, Korou Meitei, Sumit Machwe

July 30, 2022

Contents

1 Problem Statement

Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world and has been observed to be spreading. Symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death. Since it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. Dengue incidences follows seasonal transmission patterns followed by cyclic larger epidemics. In areas where dengue is endemic, incidence follows seasonal transmission patterns punctuated every few years by much larger epidemics. Because these epidemics are unpredictable and of major consequence to the affected populations, The goal of our project is to develop a model that will forecast total number of cases in a transmission season – by using historical climatic variables (cyclic and seasonal time series from the past) and the number of dengue cases (confirmed or suspected, depending on the location) reported in the transmission season.
Optimal prediction results can be utilized by local policy makers to plan ahead and allocate required resources to manage the epidemic and save lives.

2 Project Objective

Our team will attempt to forecast the total number of Dengue Fever cases each week in San Juan, Puerto Rico and Iquitos, Peru. In order to do this, we will:

- Source the datasets provided by various government organizations
- Curate, standardize and scale the data
- Run algorithms to predict the weekly cases
- Validate the model performance using MSE and $RMSE$ metrics

We will observe the population statistics to identify the probability distribution (Normal, Poisson, Binomial etc). Scaling of data may be required. PCA can also be used for dimensionality reduction.

3 Data Source

We will use datasets published by various Federal Government agencies:

- CDC:Centers for Disease Control and Prevention
- NOAA: National Oceanic and Atmospheric Administration in the U.S. Department of Commerce.
- Department of Defense's Naval Medical Research Unit 6
- Armed Forces Health Surveillance Division
- NOAA's NCEP Climate Forecast System Reanalysis
- Data for Dengue in Iquito, Peru: Timeseries data from 2000/01 to 2008/09
- Data for Dengue in San Juan, Puerto Rico: Timeseries data from 2000/01 to 2008/09.

4 Methodology

Our methodology involves following steps:

4.1 Data Sourcing and Pre-processing

First Step before we build a model is Data Preparation. This involved:

1. Data Exploration and Understanding the data (Categorical vs. Numeric etc.)
2. Cleaning the data and dealing with outliers and missing values
3. Scaling the data. Scaling helps in model convergence and avoiding bias.

We executed below steps for Data Cleansing & Pre-processing of San-Juan and Iquitos data:

- Impute missing climatic data (forward and back filled) SanJuan and Iquitos.
- Weekly mean resampling of daily-climatic data. To conform with ISO calendar, assumed week start as Monday.
- Climatic data follows ISO calendar. There are specific years (e.g, 2005) which ends on Weekends. To avoid bias, we carefully handled week spillovers between years and ensured that each year had exactly 52 weeks of data in our timeseries.

Based on Data Exploration for San-Juan and Iquitos, we have below observations for the Data Distribution:

- $Year$ and $WeekofYear$ are used to construct Timeseries data. They are not used as features.
- Features such as *station_precip_mm*, *precipitation_amount_mm*, *reanalysis_precip_amt_kg_per_m*2, *reanalysis_sat_precip_amt_mm* are tail heavy (not normally distributed). As precipitations are seasonal, it is expected to be tail heavy.

- Other features appear to follow normal distribution.
- From below Histograms (Figure 1 ,and Figure 2), It is observed that there are no significant Outliers for any of the predictors.
- *total_cases* **is the Model Label (Observations)**. It is tail heavy - consistent with outbreak spike.

Figure 1: San-Juan, Puerto Rico - Data Distribution Histogram

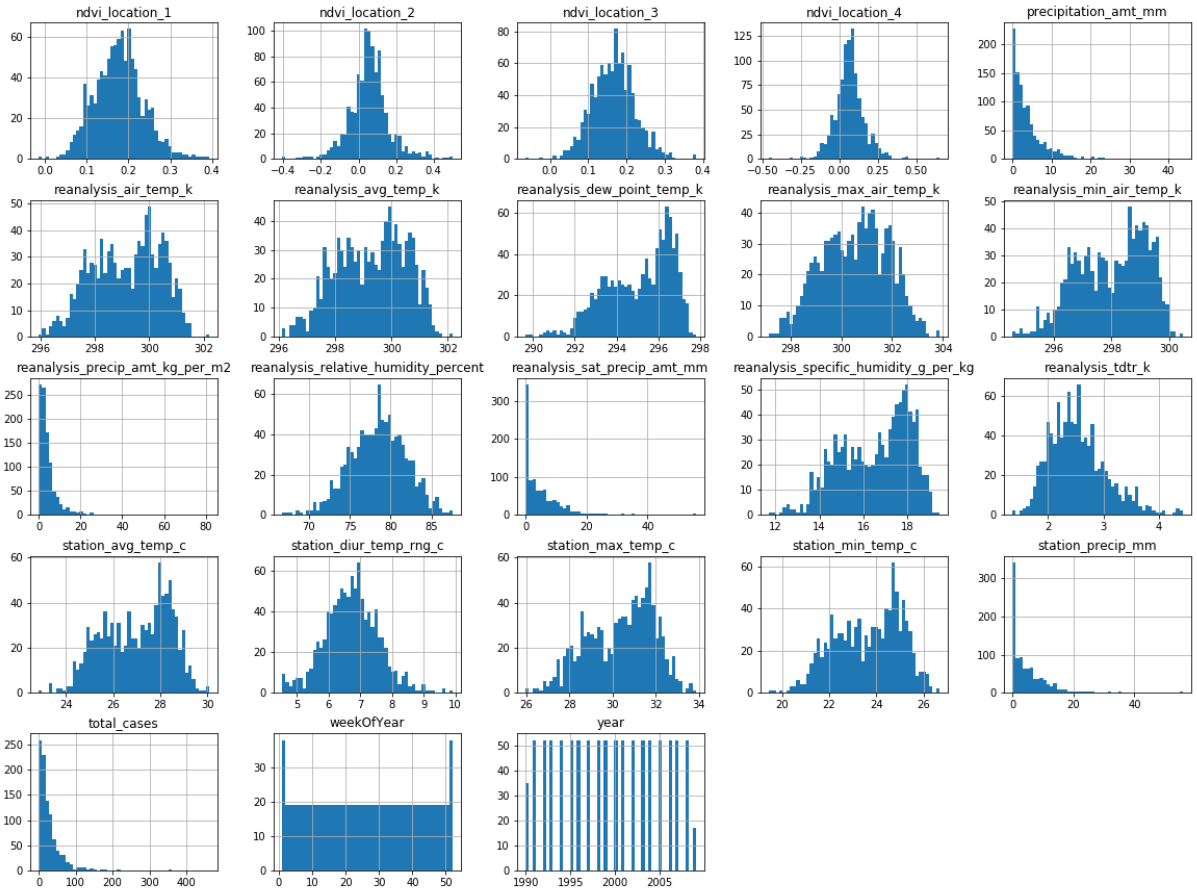
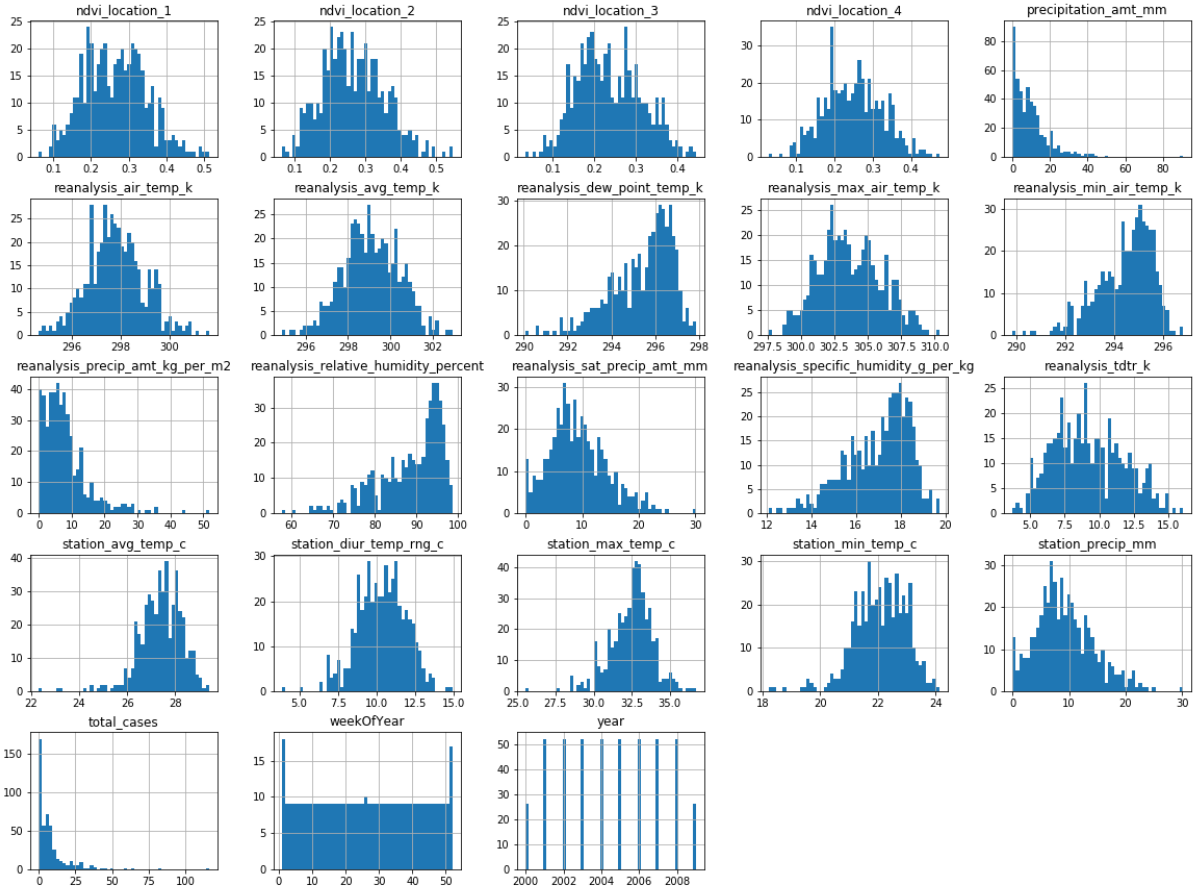


Figure 2: Iquitos - Data Distribution Histogram

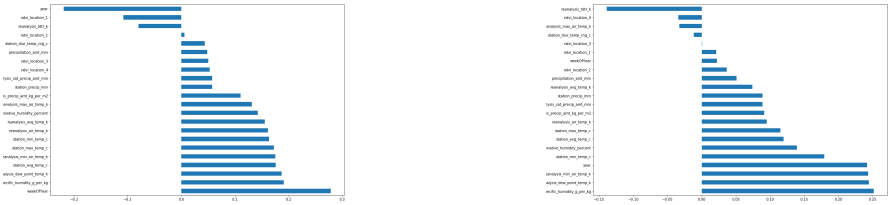


4.2 Data Visualization and preparing input for running the Models

1. Visualize the data (used *matplotlib* and *seaborn* python libraries) to find correlation among features
2. To avoid Multicollinearity in highly correlated feature space, we applied dimensionality reduction using Principal Component Analysis (PCA).
3. We analyzed SanJuan and Iquitos data separately - as climatic condition and features are quite different across cities.

We plotted feature correlation against *total_cases* for San Juan and Iquitos, after data cleansing and before feeding the data to the models. Below are some observations:

- We did not observe any feature having significant correlation with *total_cases* (observed cases)
- Many of the temperature data are strongly correlated, which is expected. But the *total_cases* variable doesn't have many obvious strong correlations.
- Many of the climatic variables are more strongly correlated. Interestingly, the *vegetation index* has weak correlation with other variables.
- **The wetter the better** *(for Dengue and model prediction).* The correlation strengths differ for each city, but it looks like *reanalysis_specific_humidity_g_per_kg* and *reanalysis_dew_point_temp_k* are strongly correlated with *total_cases*. This makes sense as we know that mosquitoes thrive in *wet climatic conditions.*
- **Hot and heavy**: As temperature and humidity increases, the *total_cases* of Dengue fever tend to rise as well.
- **Sometimes it rains, so what?**: Interestingly, the *precipitation* measurements alone does not have strong correlation with *total_cases*. Leading us to believe that precipitation only does not lead to rise in Dengue cases (refer to Figure 4 and 5)
- **Manual feature selection:** Observation shows that none of the features correlate strongly with *total_cases*. There could potentially be ****non-linear**** relationship. **Based on above, we leverage below target features to a machine learning model**
  - *reanalysis_specific_humidity_g_per_kg*
  - *reanalysis_dew_point_temp_k*
  - *station_avg_temp_c*
  - *station_min_temp_c*
- Please refer to Figure 3, 4 and 5 with sorted Correlation of data for San-Juan and Iquitos and respective Heatmaps.



[San-Juan] [Iquitos]

Figure 3: Correlation plots of features against Total Cases

Figure 4: San-Juan - HeatMap

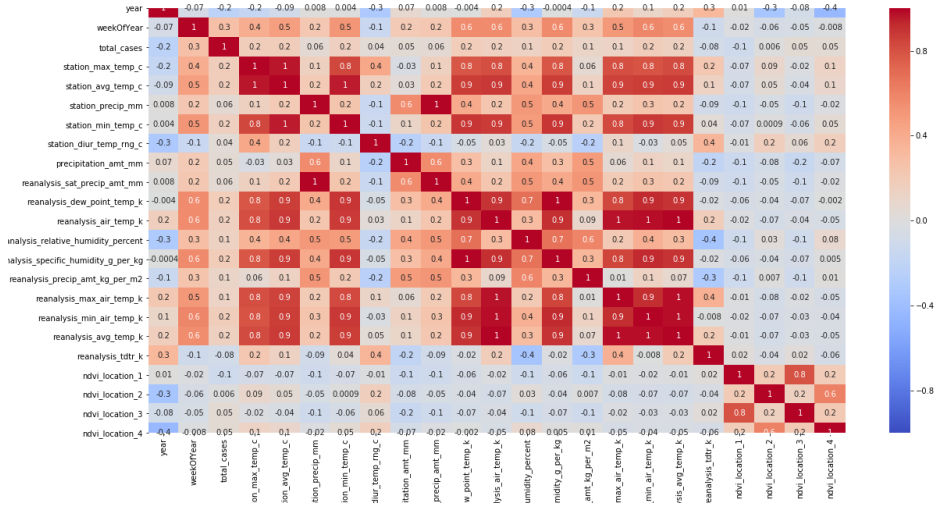
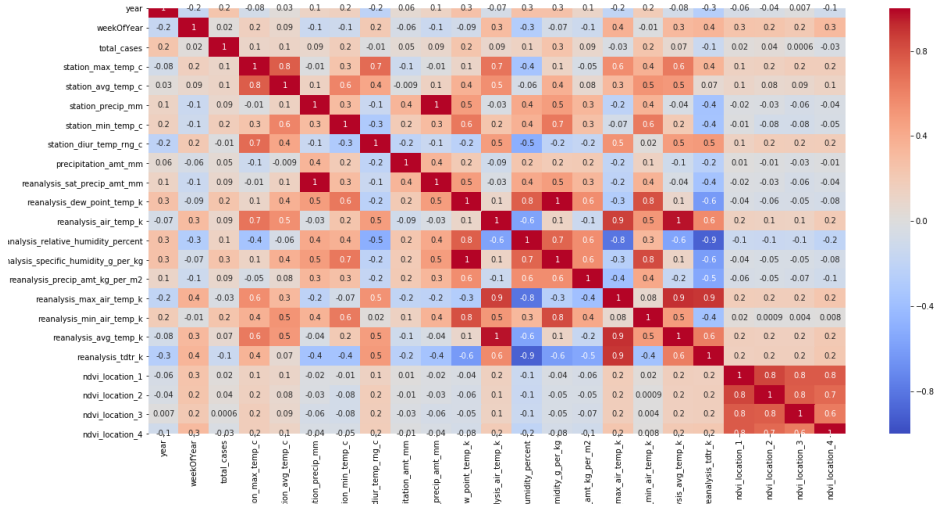


Figure 5: Iquitos - HeatMap



4.3 Data Split and Model building & Training

1. We split the data in ratio of $train : test = 80 : 20$. We leveraged cross validation - to remove any biasedness due to data distribution.
2. We trained below models :
   - Linear regression
   - Decision Tree
   - Random Forest,
   - PCA, and
   - Negative Binomial Distribution
3. For regression models (Linear Regression, Decision Tree and Random Forest) we predict $total\_cases$ of Dengue (continuous variable) using Climate independent variables. We tuned hyper parameters ($n\_estimators$, max_depth) for Random Forest for model performance optimization.
4. As there are correlation among features. We used PCA (leveraging all features in feature space) to reduce dimensions (up to 95% accuracy) and fit PCA dimensions to a Linear Regression model.
5. We leveraged Negative Binomial Distribution - in our pursuit of finding an optimal model.

4.4 Evaluation and Final Results

- Since the predicted variable is a continuous number, Forecasts were quantitatively evaluated for each target using two metrics **MAE** and **RMSE**.
- Mean absolute error (MAE) is the mean absolute difference between predictions $\hat{y}$ and observations y over n data points

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1}$$

- Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

  where n is the number of data points, y_i is the i-th measurement, and $\hat{y}_i$ is its corresponding prediction.

  RMSE is NOT scale invariant and hence comparison of models using this measure is affected by the scale of the data. For this reason, RMSE is commonly used over standardized data.

- For each model, RMSE and MAE were calculated across seasons and time (week of the season) to identify model strengths.
- The primary comparisons was made for the testing period. However, forecasts were also compared between the training and testing periods to assess how forecast accuracy changes when training data was excluded from the model.
- Based on above, we selected the best model that has optimal performance.
- For the best selected model, we leveraged the $test$ data for prediction and conclusion.

San-Juan - Observations on Model Performance

1. Among the three regression models (Linear Regression, Decision Tree and Random Forest), Linear Regression performed the best (measured by MAE and RMSE).
2. PCA with Linear Regression is the second best performing model - as expected (It has eliminated correlation among features and has used only first 2 Principal Components with 95% of explained variance ratio). Reduced PCA dimensions were fit with Linear Regression to find the best fit.
3. Decision Tree was over-fitting and hence the worst performing model.
4. Random Forest had mediocre performance results. Using Random Forest with tuned hyper-parameters did not improve results significantly. Further tree Randomization may help.
5. **Based on model runs, we observed that Negative Binomial model has best performing metrics.**
6. Below are rationale for applying **Negative Binomial Distribution Model**
   - It is based on the on the hypothesis that if we know underlying distribution of data, we may be able to model it better.
   - Target variable, *total_cases* is a non-negative integer, which means we are looking to make some ****count predictions****. Standard regression techniques for this type of prediction include:
     (a) Poisson regression
     (b) Negative binomial Distribution regression
   - Which model will perform better depends on many things, but the choice between Poisson regression and negative binomial regression is straightforward. Poisson regression fits according to the assumption that the mean and variance of the population distribution are equal. In this case, they are not equal - hence we eliminate Poisson regression model.
   - Based on current data distribution of San-Juan, we observed that label *total_cases* variance $\sigma^2 = 2521.44$ is quite large than the mean $\mu = 33.43$. When the variance is much larger than the mean, the negative binomial approach is better and **hence it is a classical case for applying Negative Binomial Distribution model** - which has best performing metrics for San-Juan.

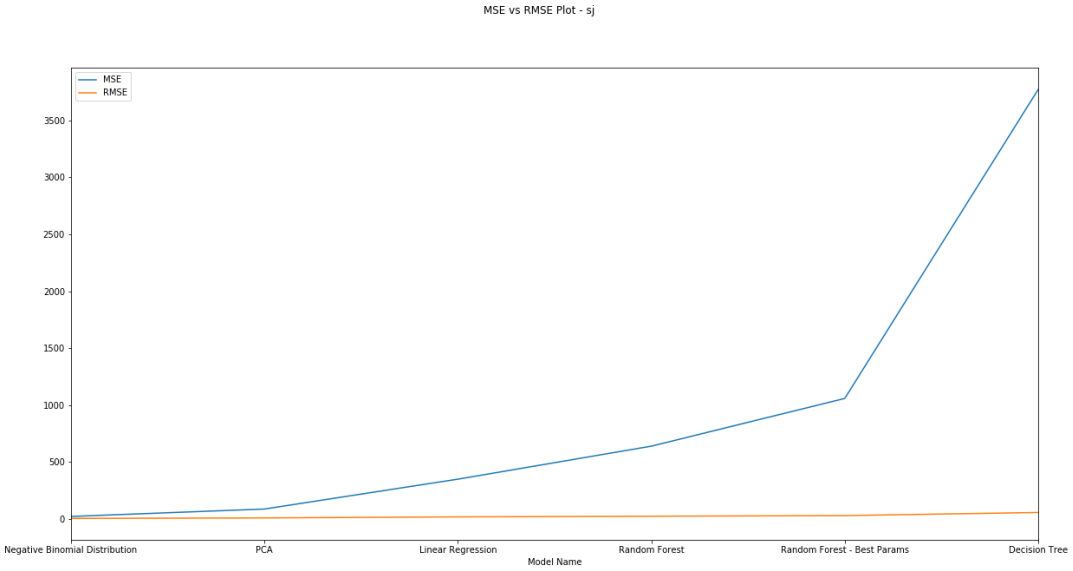
7. **Please refer (Table 1) Model Performance data for San-Juan**

Table 1: Model Performance data for San-Juan

Sr. No	Model Name	MAE	RMSE
1	**Negative Binomial Distribution**	21.9545	4.6856
2	PCA with Linear Regression	87.2259	9.1020
3	Linear Regression	349.3889	18.1871
4	Random Forest	639.1883	23.9841
5	Decision Tree	3771.5598	56.9017

8. **Please refer (Figure 6) Model Performance Plot for San-Juan**

Figure 6: Performance Metrics Plot - San-Juan



Iquitos - Observations on Model Performance

1. For Iquitos data distribution - we observed that PCA with 95% of explained variance ratio fitted with Linear Regression is the best performing model.
2. We observed that *Negative Binomial Distribution* is not the best performing model.
3. We observed that Decision Tree has the worst performance (as Expected, due to over-fitting) and Random Forest has mediocre performance.

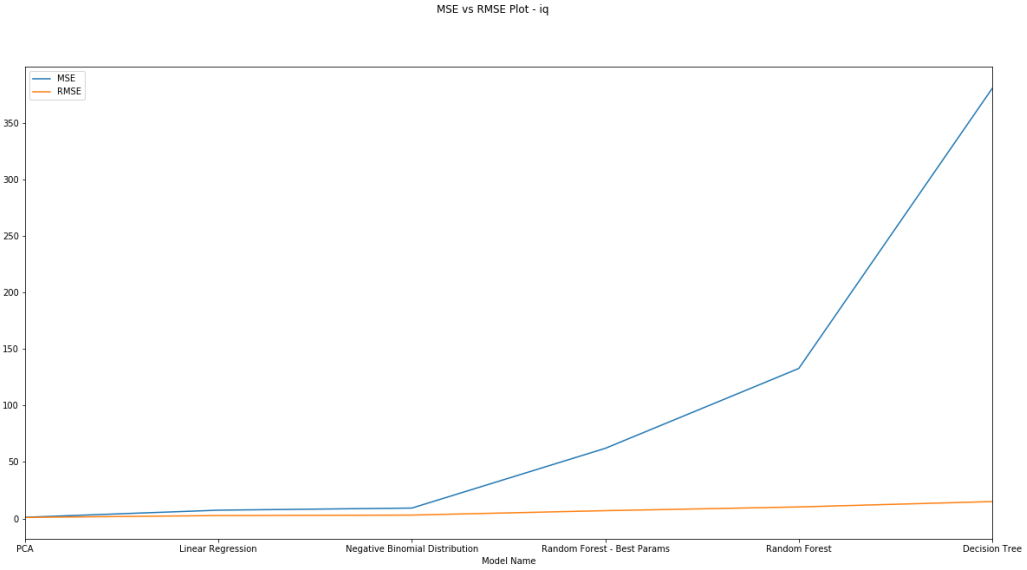
Please refer (Table 2) Model Performance data for Iquitos

Table 2: Model Performance data for Iquitos

Sr. No	Model Name	MAE	RMSE
1	Negative Binomial Distribution	9.2553	3.0423
2	**PCA with Linear Regression**	1.1126	0.9808
3	Linear Regression	7.3920	2.6282
4	Random Forest	62.1000	6.9653
5	Decision Tree	380.3814	15.0104

Please refer (Figure 7) Model Performance Plot for Iquitos

Figure 7: Performance Metrics Plot - Iquitos



4.5 Final Observations - Reflecting on Model's performance

- **Negative Binomial Distribution is the best performing model for San Juan and PCA is the best performing model for Iquitos.**
- We observe that the model in blue (Figure 8 and Figure 9) does track the seasonality of Dengue cases. However, the timing of the seasonality of our predictions has a mismatch with the actual results. One potential reason for this could be that our features don't look far enough into the past. That is to say, we are asking to predict cases at the same time as we are measuring predictions.
- Because dengue is mosquito borne, and the mosquito life-cycle depends on water, we need to take both the life of a mosquito and the time between infection and symptoms into account when modeling dengue. This is a critical avenue to explore when improving this model.
- The other important error is that our predictions are relatively consistent. We miss the spikes that are large outbreaks.

Figure 8: Best Performance Model (Negative Binomial Distribution) - San-Juan

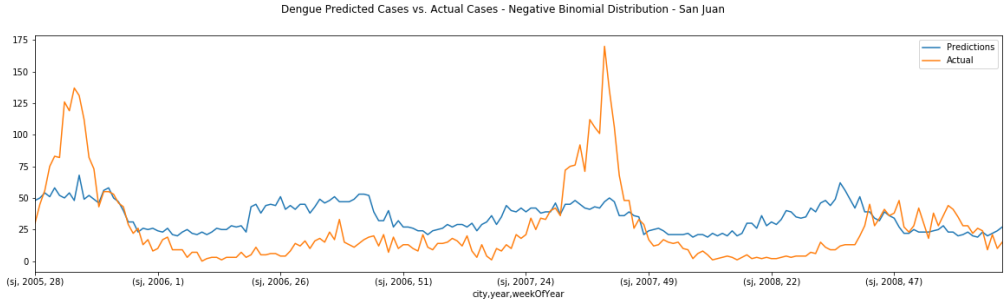
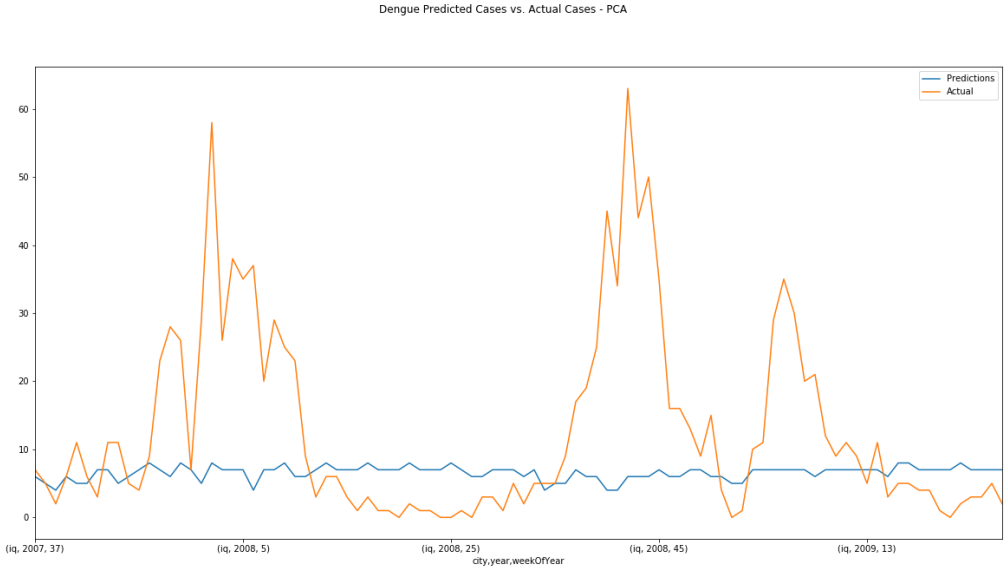


Figure 9: Best Performance Model (PCA With Linear Regression) - Iquitos



5 Further Improvements

- This is a time Series model and there is *trend and seasonality* in $total\_cases$. We should try exponential smoothing model to remove trend and seasonality, test for time series being stationary and explore any further performance improvement.
- We failed to account for contagiousness of Dangue. A possible way to account for this is to build a model that progressively predicts a new value while taking into account the previous prediction (Naïve Bayes model). By training on the dengue outbreaks and then using the predicted number of patients in the week before, we can start to model this time dependence that the current model misses.
- We should have created train data set (to fit the model), validation data set (to find the best performing model), and test data set (to measure performance). However, due to small data set we did not follow this approach. Instead of splitting $train : validation : test = 60 : 20 : 20$, we split the data in $train : test = 80 : 20$.
- We predicted only point estimates of $total\_cases$. We should have observed and plotted probability distribution of peak cases on 4 weeks cycle.

6 References

- https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/
- https://c3.ai/glossary/data-science/root-mean-square-error-rmse/
- https://rda.ucar.edu/datasets/ds093.0/#metadata/detailed.html?_do=y
- https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0002159
- https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003003