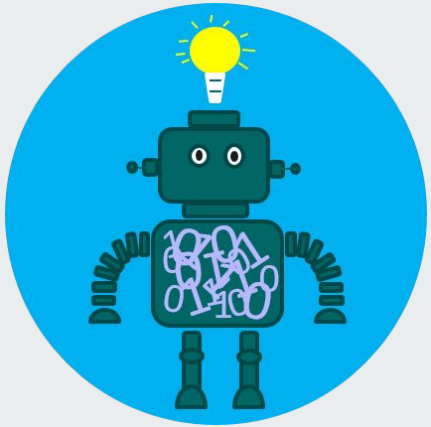
Predicting Dengue Fever

Machine Learning Project

Group Members:
140045E - S. Arunan
140338F - A.B.P.R.Lakshani
140571L - S.P.A. Senanayake



Problem ?

Predict **Total Dengue Cases** per week for San Juan and Iquitos cities by using past weather data

Develop machine learning models for predictions such that Mean Squared Error is at minimized.

Dengue Fever

- Mosquito-borne disease
- Life cycle spanning to 1.5 - 3 weeks
- Transition dynamic dependant on climate (eg: temperature , precipitation)
- Possibility of outbreak increase with climate

Due to this timely relationship and climatic relationship with dengue transmission indicates possibility of predicting dengue cases using Machine Learning

Feature Set

- Temperature
  - Max, min, average, diurnal range,dew point
- Precipitation
  - Total rainfall
- Humidity
  - Mean relative and mean specific
- Vegetation
  - Level of vegetation in NW, NE, SW and SE

Approach

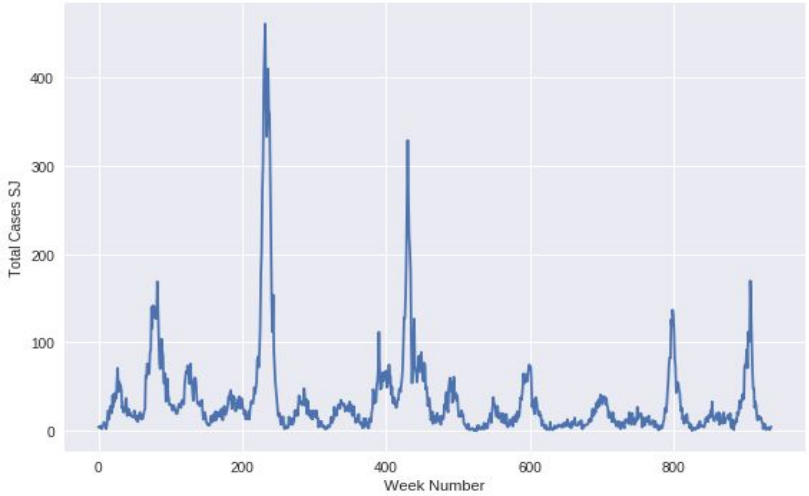
In order to build the final machine learning model, we incrementally followed given steps

- Exploratory Data Analysis (EDA)
  - Visualize distribution of features, patterns in features and *total_cases*
- Data Imputation
  - To fill missing values in features
- Feature Engineering
  - Find optimal set of features that shows significant relationship for predictions

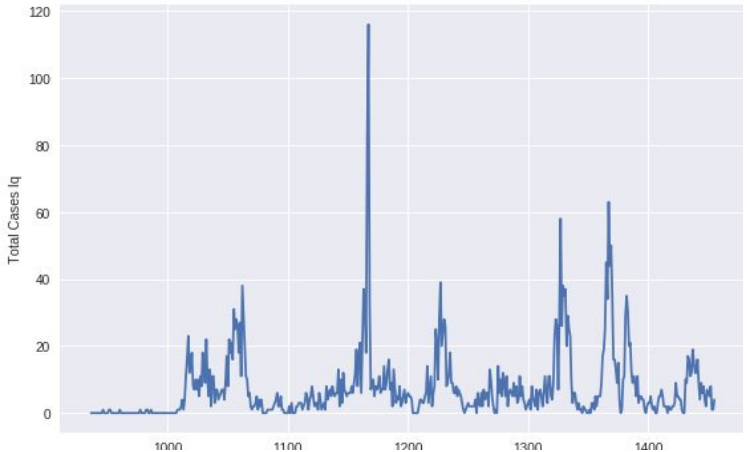
Approach cont.

- Evaluate machine learning models
  - To find the optimal machine learning model giving minimum error
- Results
  - Analysing results

Distribution of Total cases over Time



Total_case over time for San Juan



Total_case over time for Iquitos

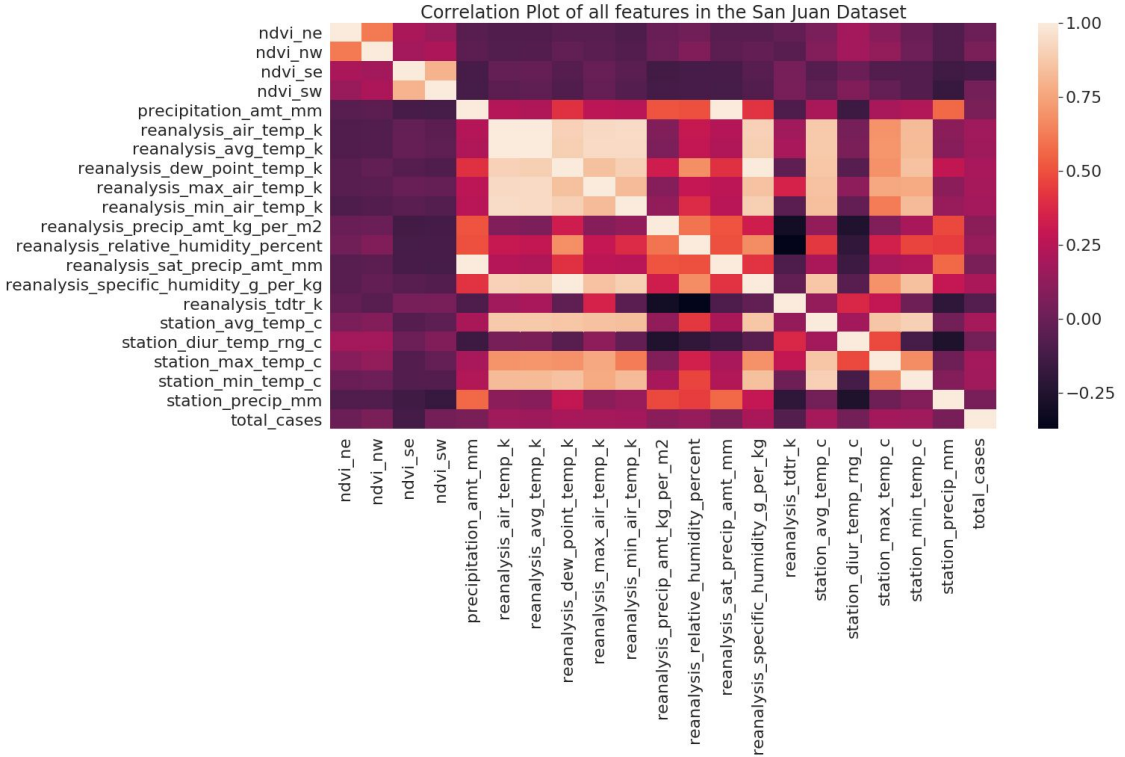
Data Imputation

- Almost all the features contained missing values
- Imputation using
  - Forward fill
    - Fill data from predecessor values
  - Mean value
    - Fill data from mean value
  - Redundant values
    - Use *reanalysis* features to impute *station* values in features

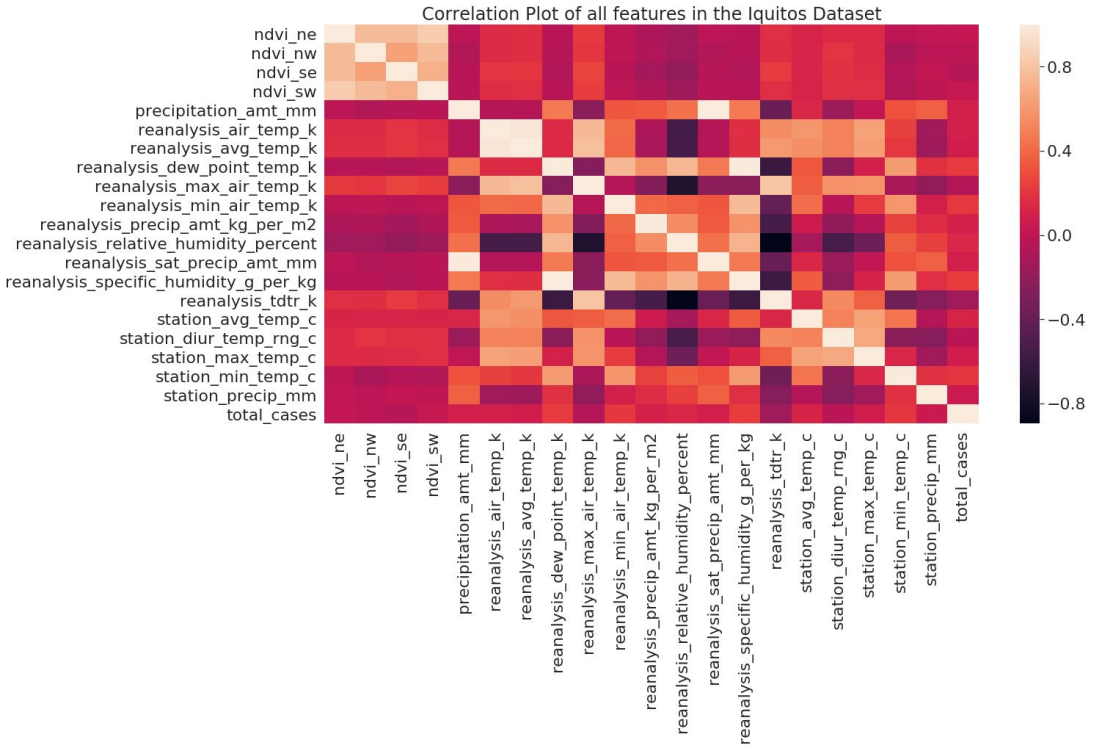
Feature Engineering

- Use correlation matrix to find strength of relationship between features and *total_case*
- Add time-series features
- Drop unnecessary features
- Normalize features
    - To ensure all the numeric values of values are in the same range

Correlation matrix for San Juan Data set



Correlation matrix for Iquitos dataset



Time Series data

- Mosquitoes precise weather conditions to reproduce
- Hence cases for a given week will be results of **past** weather conditions
- Required to analyze past weather records for current predictions
- Use **Moving Average**
    - Smooth short term fluctuations
    - Highlight long term trends
    - For features - *recent_mean_dew_point* , *recent_mean_spec_humid* , *recent_sum_precip* use moving average of 100 for San Juan , 30 for Iquitos

Drop Unnecessary features

Features which does not have a significant correlationship with total_cases was dropped to ensure predictions are based on features with strong relationship

Sample dropped feature:

- ndvi_ne , ndvi_nw , ndvi_se , ndvi_sw
- Precipitation_amt_mm
- reanalysis_air_temp_k etc

Machine Learning Model

In-order to find the Machine learning model having minimum squared error for predictions, first we used five different models.

- Linear regression
- KNN
- SVM
- Gradient Boosting
- Random Forest
- MLP

Cross Validation

Before the training process

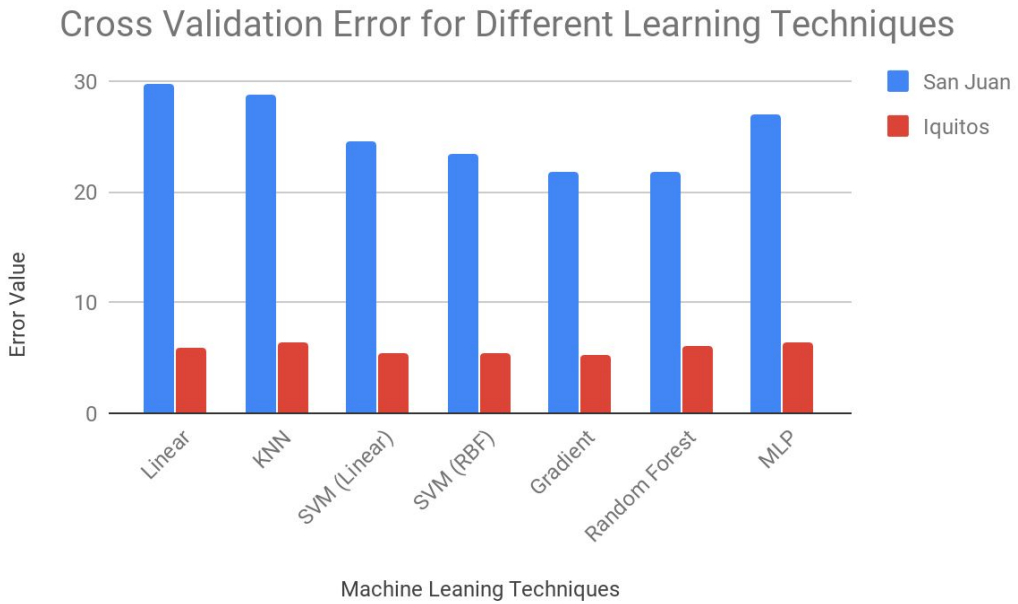
- Training data was splitted to training and cross-validating

After the training process

- Cross validation is used to minimize **overfitting** and **underfitting** of data

Results

Minimum error from Gradient boosting for San Juan and SVM for Iquitos



Challenges

- Model cannot not predict outbreaks, just increased caseloads
- Overfitting the data. Received better validation scores, but worse test scores after Submitting to DrivenData

Additional Information

Final code here - https://github.com/arunans23/DengAI-datadriven