DengAI : Predicting Disease Spread

Machine Learning Project

Malaka Dayasiri	140097M
Adeesha Jayasooriya	140253N
Bhanuka Mahanama	140381E
Wishmitha Mendis	140392M

Problem Definition & Motivation

DengAI Competition hosted by DrivenData [www.drivendata.org]

The data set provided for the purpose are associated with cities of San Juan and Iquitos in Peru. Through the identified relationship between the factors and the spread of the disease, spread of the disease is required to be predicted for subsequent periods for the two cities.

Predicting dengue cases is helpful for health officials

- To know disease outbreak times beforehand
- Identify the relationship between the factors and the spread of the disease

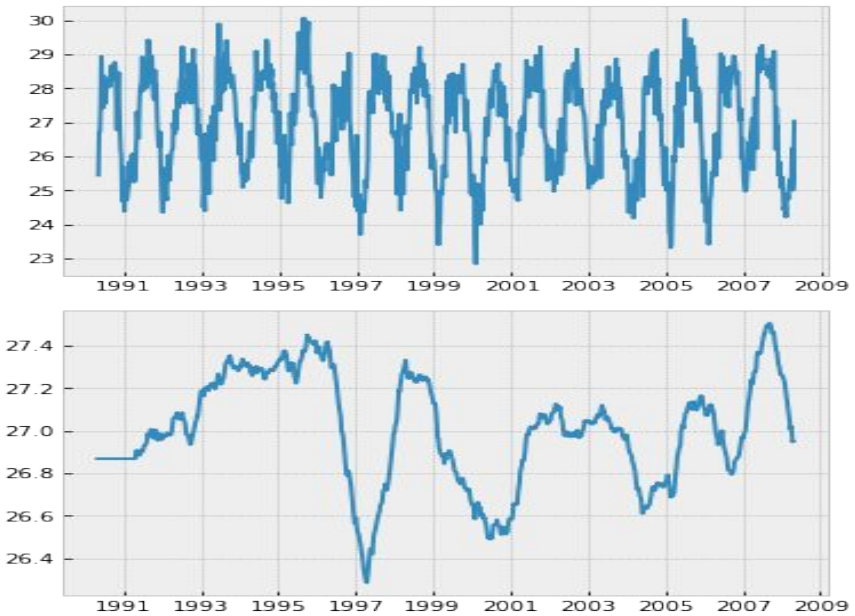
Methodology

- Data Preprocessing
  - Data Cleaning
    - Filling Missing Values
    - Smoothing Noisy Data
  - Data Transformation
- Machine Learning Approaches
  - Two Approaches
  - Combination of Two Models

Data Preprocessing - Data Cleaning

Filling Missing Values

Smoothing Noisy Data



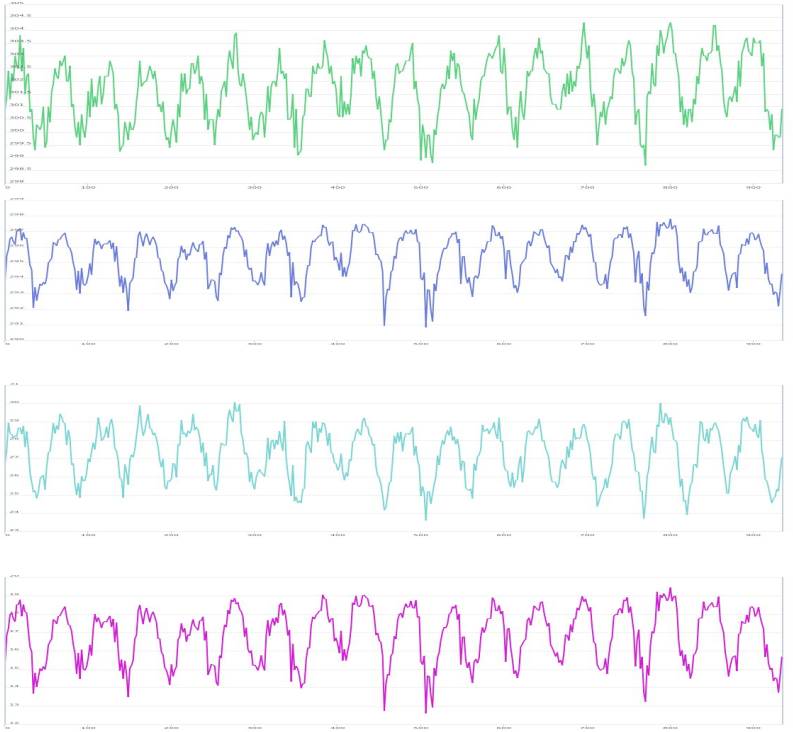
Data Preprocessing - Data Transformation

- Normalized Differential Vegetation Index (NDVI)
  - Include amount of vegetation in the four quadrants (NE, NW, SE, SW) of the city.
  - Rather than using separate four dimensions it would be better use these four as one dimensions.
  - Therefore create **ndvi_mean** as average of ndvi_ne, ndvi_nw, ndvi_se, ndvi_sw values.
- Unit Conversion
  - NOAA's NCEP Climate Forecast System Reanalysis measurements in Kelvin
  - NOAA's GHCN daily climate data weather station measurements in Celsius
  - Therefore NOAA's NCEP Climate Forecast System Reanalysis measurement convert in to Celsius.

Machine Learning Approach - I

Feature Selection

	#	Uni...eg.
reanalysis_specific_humidity_g_per_k		41.788
reanalysis_dew_point_temp_k		40.064
station_avg_temp_c		36.850
reanalysis_max_air_temp_k		36.227
station_max_temp_c		34.390
reanalysis_min_air_temp_k		33.672
reanalysis_air_temp_k		31.387
station_min_temp_c		29.594
reanalysis_avg_temp_k		29.047
reanalysis_relative_humidity_percent		19.891
reanalysis_precip_amt_kg_per_m2		10.805
ndvi_ne		6.946
reanalysis_tdtr_k		4.291
reanalysis_sat_precip_amt_mm		3.408
precipitation_amt_mm		3.408
station_precip_mm		2.521
ndvi_nw		2.103
ndvi_se		1.785
ndvi_sw		1.704
station_diur_temp_rng_c		1.166



Machine Learning Approach - I

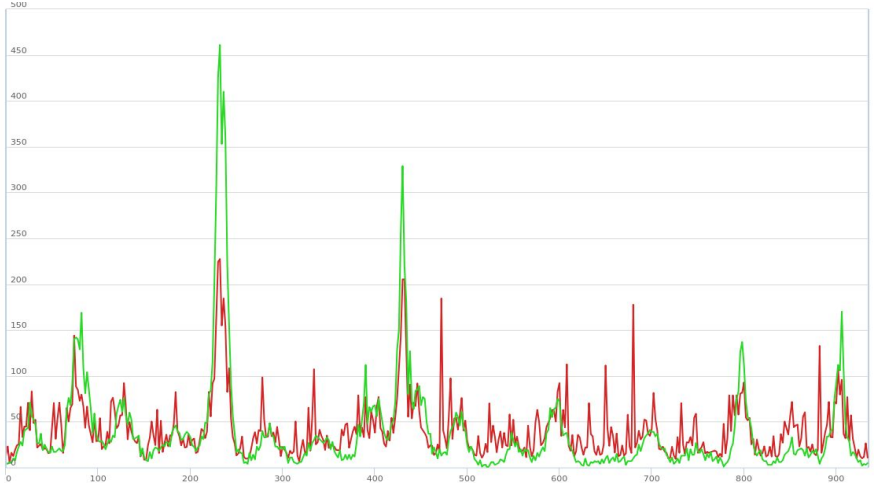
Models

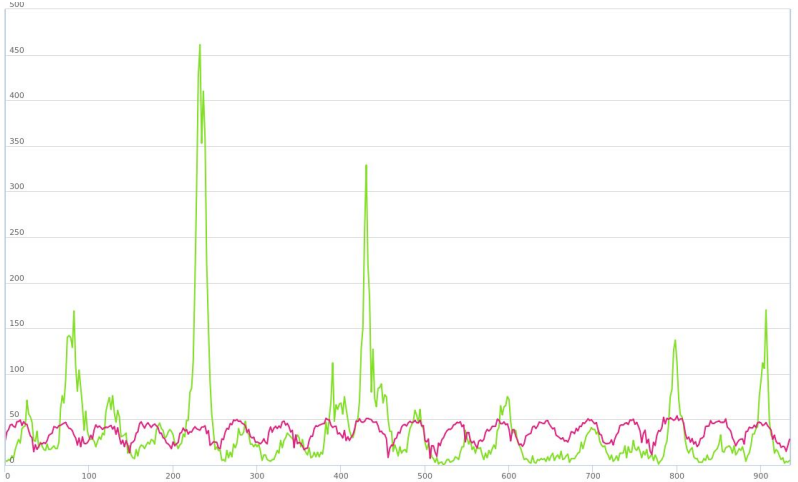
Following models were tried out with the above selected feature for train data with starfied 10-fold cross validation.

Method	MSE	RMSE	MAE	R2
Neural Network	2532.262	50.322	27.242	0.040
Linear Regression	2532.312	50.322	27.455	0.040
kNN	2652.457	51.502	27.882	-0.006
Random Forest	3259.486	57.092	30.780	-0.236
Tree	3457.277	58.799	31.992	-0.311
AdaBoost	4340.381	65.882	32.395	-0.646
SVM	3304.856	57.488	33.143	-0.253

Machine Learning Approach - I

Random Forest model (Left) is the best at capturing these spike like changes in the total cases whereas Linear Regression model (Right) is capable of capturing the seasonal component of the total cases

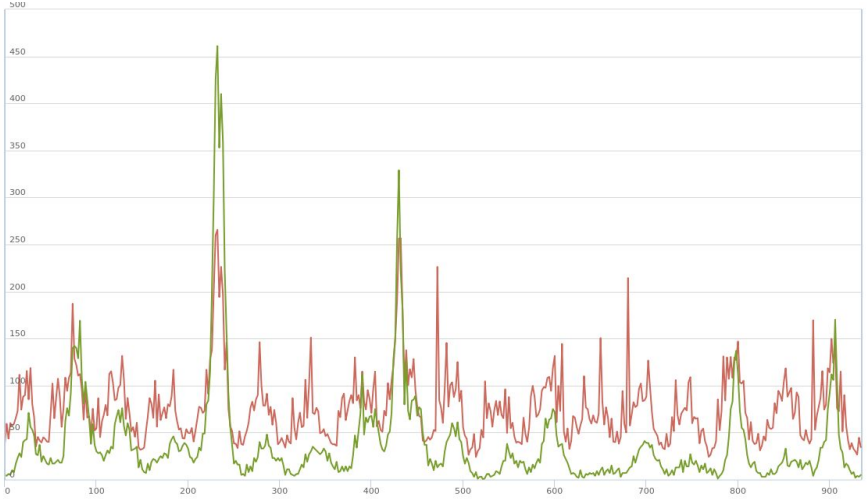




Machine Learning Approach - I

Therefore the summation of this Random Forest model and above Linear Regression model was taken in order to achieve following enhancements.

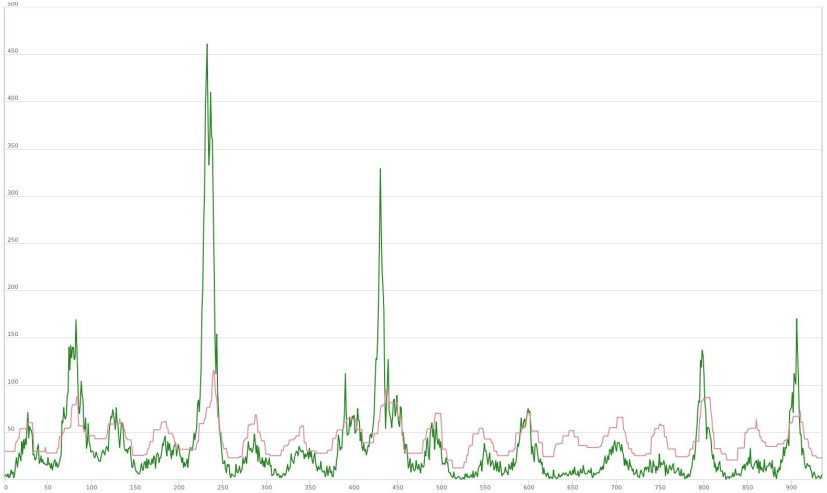
1. Identify sudden spike like variations
2. Reduce the overfitting in Random Forest model
3. Improve the seasonal component of the Random Forest model



Machine Learning Approach - I

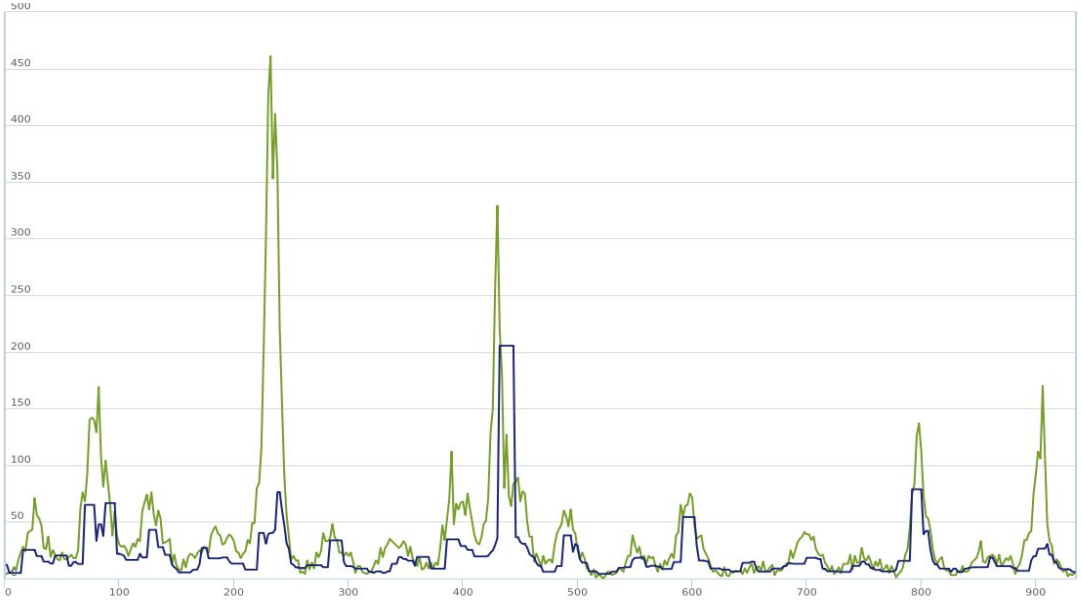
Summation of Random Forest model and Linear Regression model with minimum for 13 week rolling window for training data set of San Juan. This distribution (red) is much better than the previous distribution for the following two reasons.

1. Predicted value is much closer to the expected value
2. The distribution is smoothened since a rolling window operation is carried out



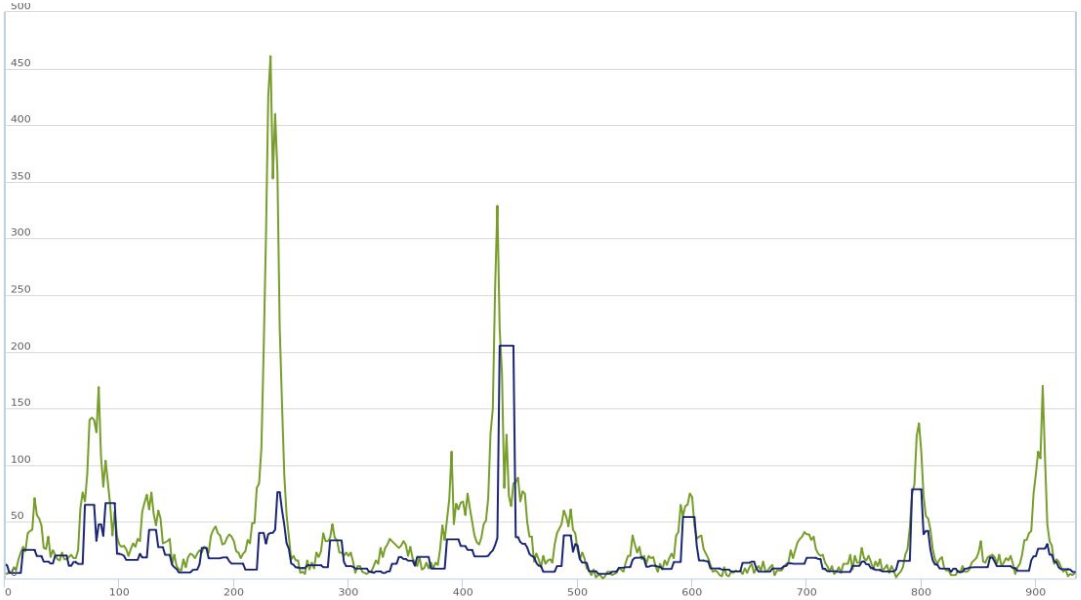
Machine Learning Approach - I

Mode of Random Forest model with 13 week rolling window for training data set of San Juan. In this distribution mode remains in a constant high value near most of the sudden increments.



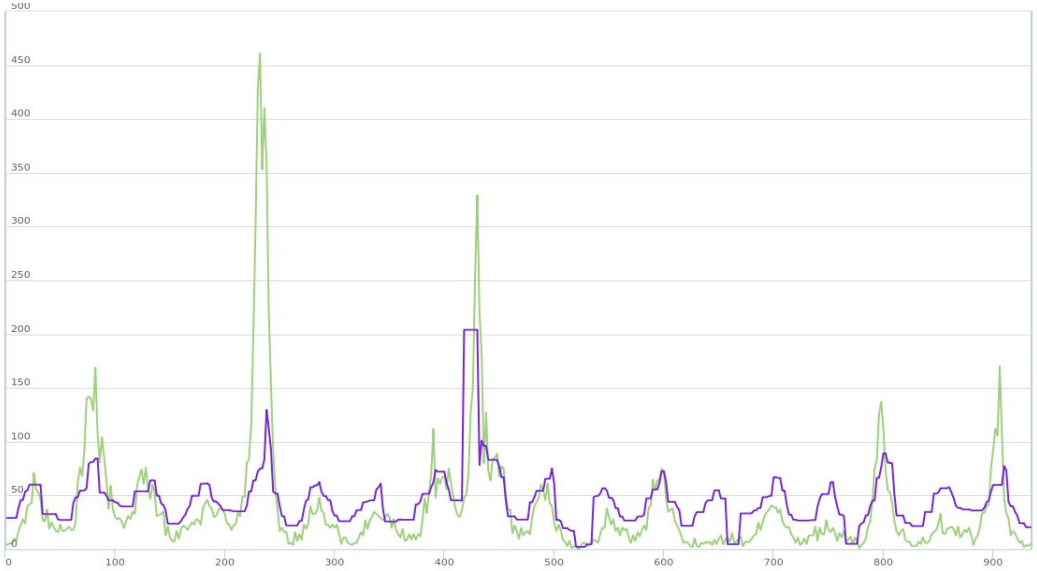
Machine Learning Approach - I

Mode of Random Forest model with 13 week rolling window for training data set of San Juan. In this distribution mode remains in a constant high value near most of the sudden increments.



Machine Learning Approach - I

Then the above two distributions were combined in the final model. The average error was **19.9639**.



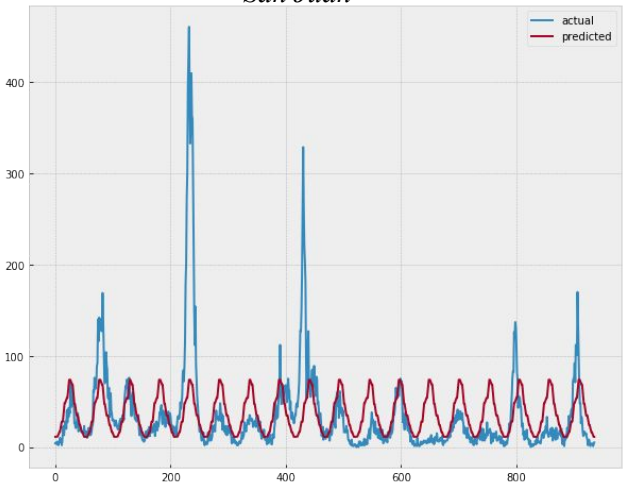

```
if(mode[i]>60):
out[i] = max(base[i],mode[i])

if(mode[i]<6):
out[i] = min(base[i],mode[i])

else :
    out[i] = base[i]
```

Machine Learning Approach - II

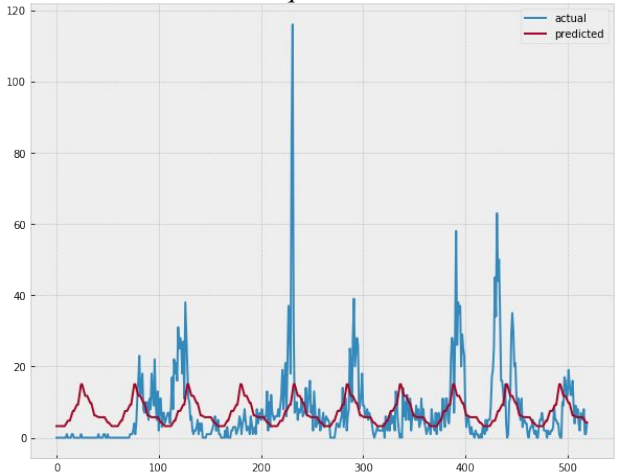
Identifying Annual Trend

predict the total cases of San Juan using linear regression model and month variable as the only feature.

San Juan



Iquitos

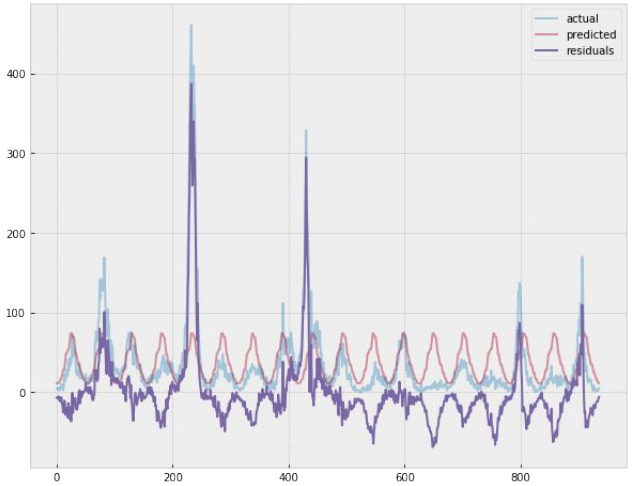


Machine Learning Approach - II

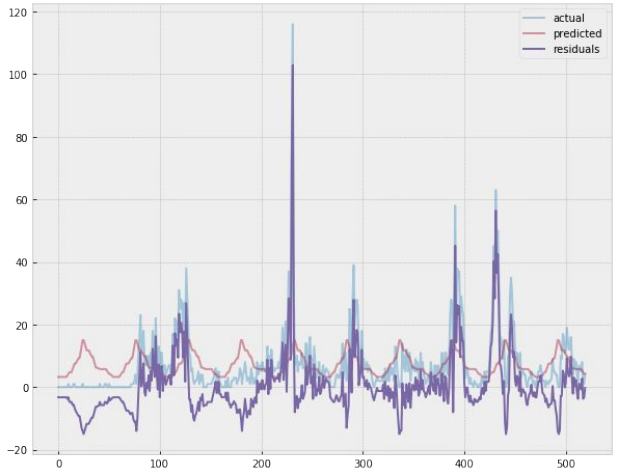
Identifying residuals

Residual = Actual - Predicted

San Juan



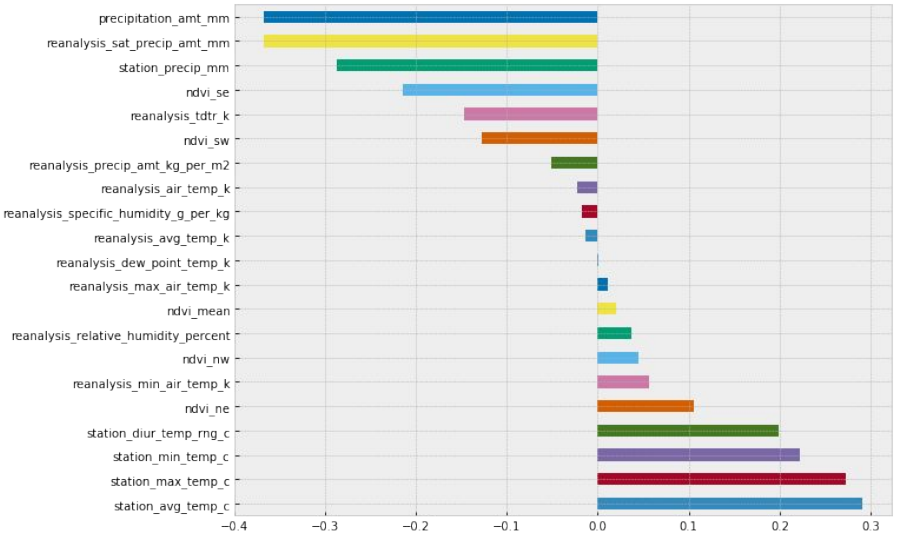
Iquitos



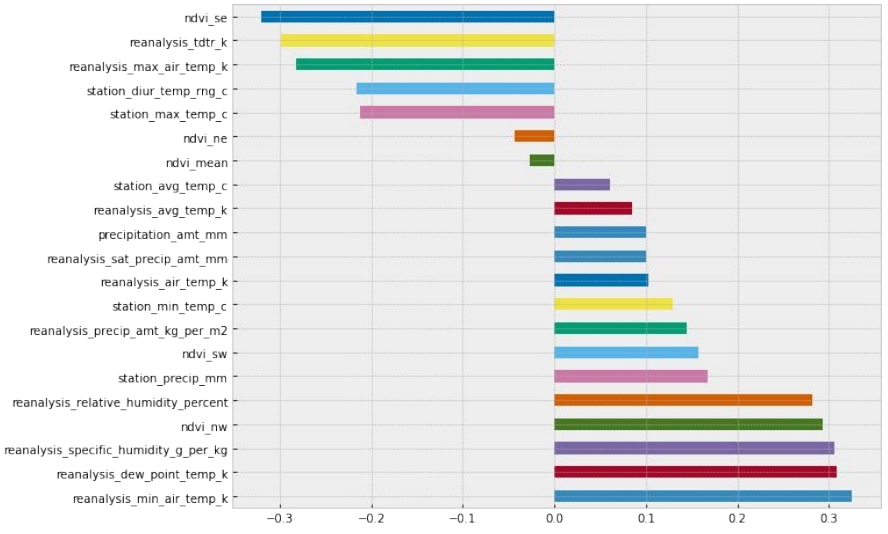
Machine Learning Approach - II

Feature selection

San Juan



Iquitos



Machine Learning Approach - II

Selected Features

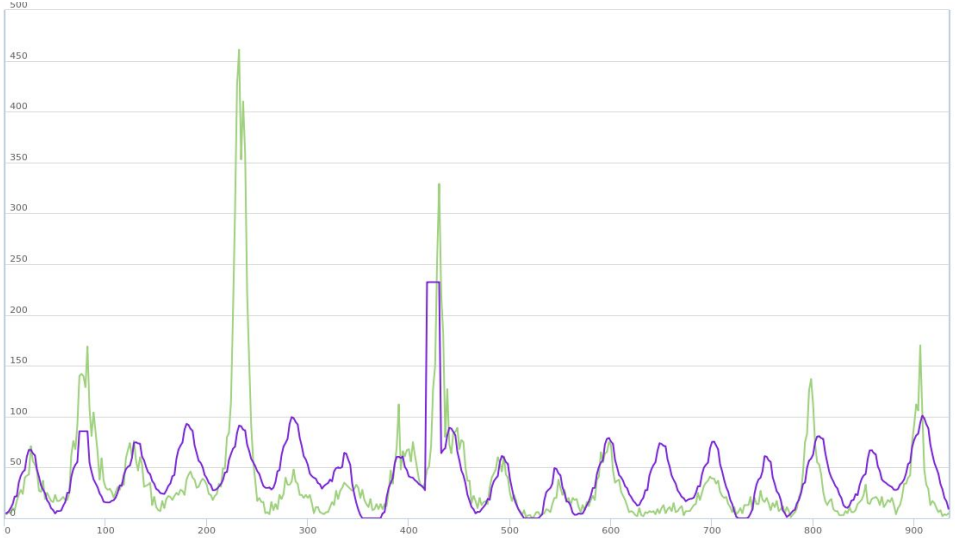
- Pixel southeast of city centroid (ndvi_se) that having higher negative correlation with both cities.
- Average temperature (station_avg_temp_c) that having higher positive correlation with San Juan.

Prediction model

- **Linear regression** used to predict the residuals and annual trend.
- First partitioned the given training data set as training set and test set.
- Then annual trend and residuals predict using Linear regression.
- After that get the final predictions using sum of annual trend prediction and residuals predictions.
- Mean absolute error for San Juan 22.7222 and for Iquitos 6.3751.
- Final submission based on this model gave mean absolute error of **21.4591**.

Model Combination

Developed model distributions were combined using previous algorithm.



Results

- Evaluation Methodology
  - Mean Absolute Error(MAE)
- Results Comparison

Model	Score (MAE)
First Approach	19.9639
Second Approach	21.4591
Combined Approach	17.7668

- Final Result

Thank You