DengAI: Predicting Disease Spread

Group : MLG14

Problem Statement

Using environmental data collected by various U.S. Federal Government agencies from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce predicting the number of dengue fever cases reported each week in San Juan and Iquitos, Peru is the goal of this project. The data is obtained from a competition held by Datadriven.

Methodology

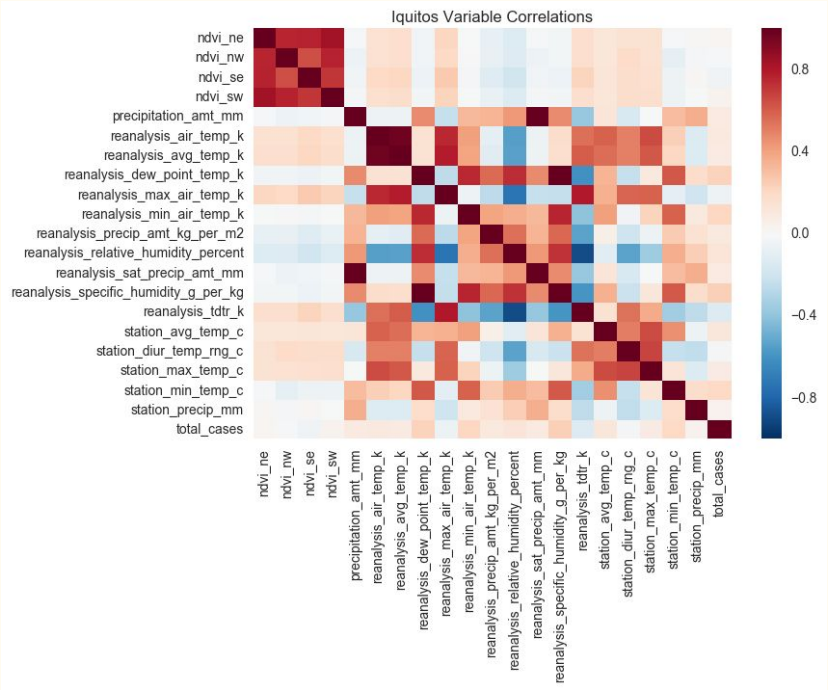
- Data PreProcessing
- Feature Engineering

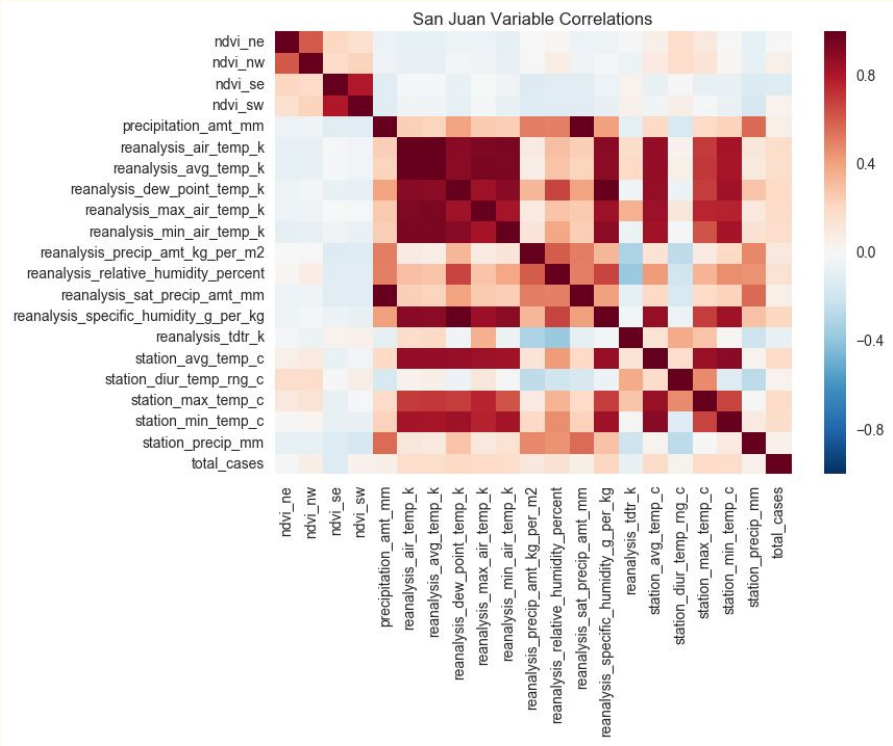
Data Preprocessing

- Seperate two cities
- Filling the missing variable
- Rolling mean values for features

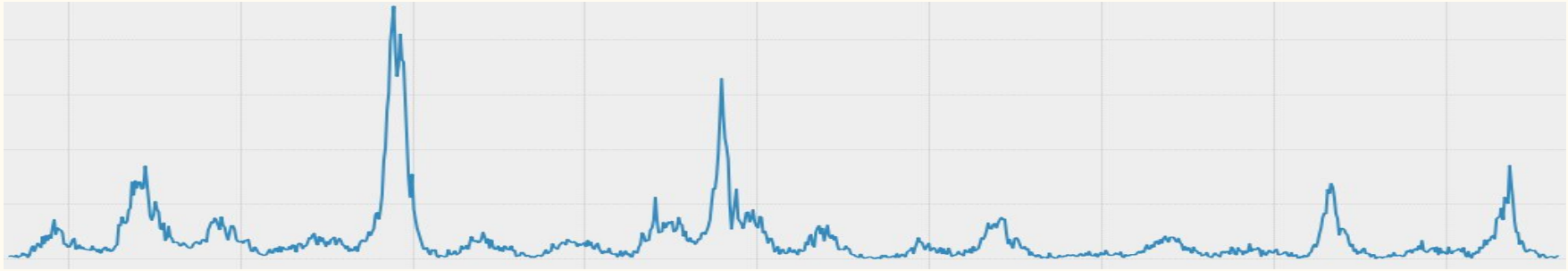
Feature Engineering

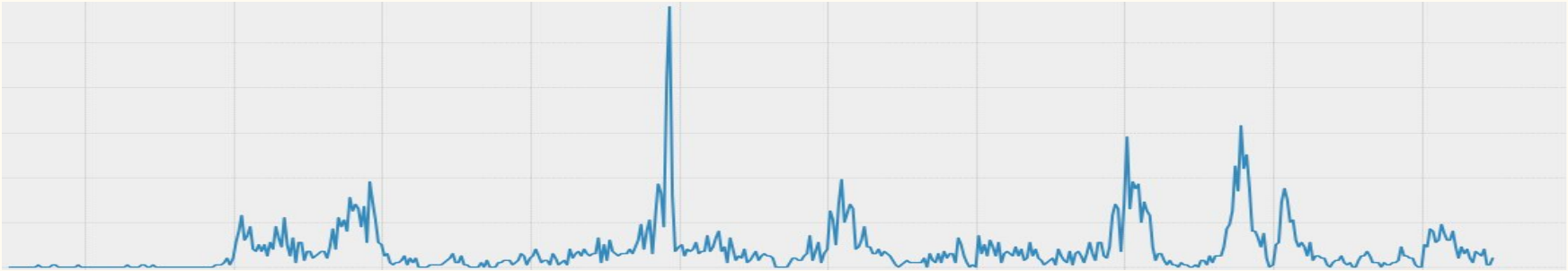
- Eliminate Highly correlated variables (Reducing dimensions)





- Monthly trend for dengue cases(Separated by cities)





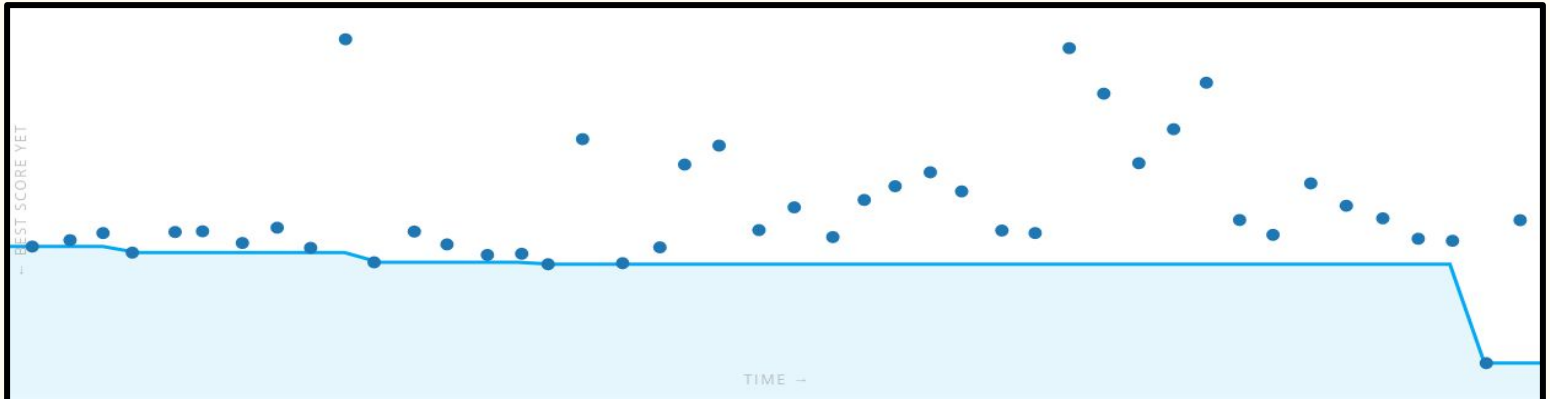
- Rolling mean for weather data
- Residual analysis on monthly trend with total cases
- Combine monthly trend of total cases with temperatures

Models and performances

Model	Test(mean absolute error)
Linear Regression(without trend analysis)	27
Random Forest	25.3894
Negative Binomial	25.3053
LSTM	26
Linear Regression with trend analysis	20.9159

Issues

- Limited domain knowledge
- Models couldn’t predict the outbreaks
- Overfitting (Huge variation from train error and test error)



DengAI: Predicting Disease Spread

HOSTED BY DRIVENDATA

5
MONTHS LEFT

HOME ABOUT PROBLEM DESCRIPTION

Submissions

Glory!

BEST	CURRENT RANK	# COMPETITORS	SUBS. TODAY
20.9159	168	4162	2 / 3

EVALUATION METRIC

LEADERBOARD

DATA DOWNLOAD

SUBMISSIONS 27

Conclusion

In a nutshell we learned how to tune the data and identify the model to predict useful informations.