# FINAL

**Nikhil Agarwal**
**PREDICT 413 | Section 57**
**Driven Data screenname: niks**

# Table of Contents

## Problem Description

The intent of this activity was to predict the total number of dengue fever cases in the cities of San Juan, Puerto Rico and Iquitos, Peru. In this dataset, there was five years of information on San Juan and three years of information on Iquitos. This dataset was provided by an organization called Driven Data and consisted of 23 potential predictor variables. Interestingly, the potential predictors were sourced from four different systems: NOAA GHCN, PERSIANN, NOAA NCEP, NOAA CDR Normalized Difference Vegetation Index. Several statistical models were constructed for this activity. This report will discuss five of the models constructed: Negative Binomial, Extreme Gradient Boosting (XGBoost), Random Forests, Arima, and Neural Networks. Since the results were submitted directly to Driven Data, a mean absolute error (MAE) was generated for each submitted model and are available in the Performance/Accuracy section. Prior to submission, the MAE for each model was derived and is discussed in the Performance/Accuracy section.

## Significance

Dengue fever is a common tropical disease that affects millions around the world every year. Typically, the virus that causes the Dengue fever is transmitted via mosquitos (Driven Data, 2016). There are no specific treatments for this disease other than rest and the use of pain killers (Gilbert, 2016). According to an article in the New York Times by Jonathan Gilbert, "On rare occasions, the virus can become fatal. There are four types of dengue, which gives the virus the opportunity to re-establish itself in people who have developed immunity against one of the strains." (Gilbert, 2016). The ability to predict the number of cases that may happen due to a variety of parameters can greatly assist emergency responders and health officials in managing this disease.

## Overview of Data

### Datasets

Although the different predictor variables are from different data sources, Driven Data (the hosts of this online competition) combined all the information into a single CSV file. The response variable, *total_cases*, was provided in a different file which was then combined in the R script. Finally, the test dataset was also provided, but without the actual number of cases. Table 4 (Appendix) provides an overview of the default predictor variables.

A key note to make is that the data were separated by city: San Jose (*sj*) and Iquitos (*iq*). This enabled the opportunity to construct independent models for each city.

### Missing Values and Imputation

Prior to any variable creation, an analysis was done to understand the number of missing values for the default predictor variables. Table 1 summarizes the number of missing values for each of the default predictor variables. Most of them are far below 20% with the exception of the variable *ndvi_ne* for San Juan. Due to the presence of missing values, imputation will be necessary. Overall, less than 2% of all the observations are missing. Therefore, it was decided to impute the missing information prior to further data exploration.

*Table 1: Summary of missing values for each predictor variable*

| Predictor Variable | Number of Missing Values | | Percent Missing by City | | Percent Missing Overall |
|---|---|---|---|---|---|
| | iq | sj | iq | sj | |
| *year* | 0 | 0 | 0.00% | 0.00% | 0.00% |
| *weekofyear* | 0 | 0 | 0.00% | 0.00% | 0.00% |
| *week_start_date* | 0 | 0 | 0.00% | 0.00% | 0.00% |
| *ndvi_ne* | 3 | 191 | 0.58% | 20.41% | 13.32% |
| *ndvi_nw* | 3 | 49 | 0.58% | 5.24% | 3.57% |
| *ndvi_se* | 3 | 19 | 0.58% | 2.03% | 1.51% |
| *ndvi_sw* | 3 | 19 | 0.58% | 2.03% | 1.51% |
| *precipitation_amt_mm* | 4 | 9 | 0.77% | 0.96% | 0.89% |
| *reanalysis_air_temp_k* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_avg_temp_k* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_dew_point_temp_k* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_max_air_temp_k* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_min_air_temp_k* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_precip_amt_kg_per_m2* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_relative_humidity_percent* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_sat_precip_amt_mm* | 4 | 9 | 0.77% | 0.96% | 0.89% |
| *reanalysis_specific_humidity_g_per_kg* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *reanalysis_tdtr_k* | 4 | 6 | 0.77% | 0.64% | 0.69% |
| *station_avg_temp_c* | 37 | 6 | 7.12% | 0.64% | 2.95% |
| *station_diur_temp_rng_c* | 37 | 6 | 7.12% | 0.64% | 2.95% |
| *station_max_temp_c* | 14 | 6 | 2.69% | 0.64% | 1.37% |
| *station_min_temp_c* | 8 | 6 | 1.54% | 0.64% | 0.96% |
| *station_precip_mm* | 16 | 6 | 3.08% | 0.64% | 1.51% |

The imputation method chosen was 'last one carried forward' (LOCF). This method allows the last known values to be copied into the 'missing' value. The LOCF method is fairly simple to understand and it provides acceptable context to variables.

## Descriptive Statistics

Table 5 and Table 6 (in the Appendix) provide a brief summary of the default predictors and the response variable, *total*_cases, separated by city. Note that the response variable, *total_cases*, has also been included in these table. Perhaps the most obvious characteristic of the split dataset is the amount of information available for San Juan versus Iquitos (18 years' worth of information versus 10 years' worth). In terms of the response variable, *total_cases*, note how the number of cases vary widely amongst the two cities (416 max cases for San Juan vs. 116 max cases for Iquitos).

Many of the predictors' average and median values are close to each other – suggesting that the outliers do not necessarily have as much of an impact. However, there are predictors whose average is significantly different from the median. For instance, in San Juan, the variable, *precipitation_amt_mm*, has an average of 35.45 mm versus a median of 20.61 mm. Furthermore, in Iquitos, the variable, *station_precip_mm*, has an average of 62.44 mm versus a median of 45.65 mm. This is an early indication that outliers may need to be handled and that skewness may be present. Figure 5 and Figure 6 (both in the Appendix) provide a histogram and boxplot view for these variables. Note the skewness in both charts as well as the presence of outliers.

## Visual Data Exploration

Numerous visual charts were created to help explore the data visually. However, for succinctness, only a few have been included and discussed in this paper. All of the visual data exploration was conducted after the missing values were imputed. Figure 1 illustrates the number of dengue fever cases for both San Juan and Iquitos. This chart also shows the different date ranges for each city. Note how the dengue fever cases in San Juan in the mid to late 1990's were significantly higher. There is also some distinguishable seasonality for both cities. Intuitively, this makes sense as each city experiences different seasons throughout the year.

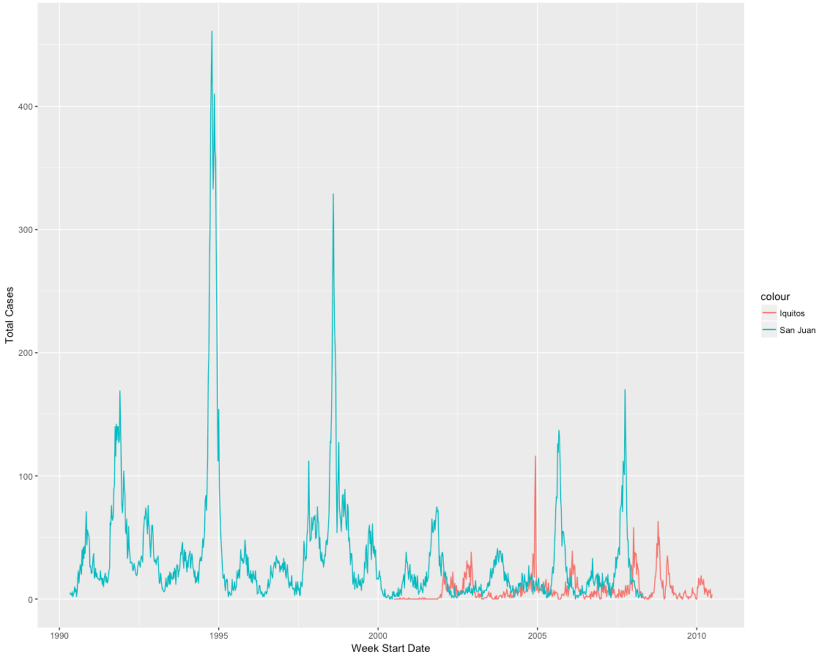

*Figure 1: Plot of total_cases over time by city*

Figure 2 illustrates the total number of cases for each city by week. Interestingly, week 18 (late April) seems to have the lowest total number of dengue fever cases in San Juan. Accordingly, week 30 (late July) seems to have the lowest total number of dengue fever cases in Iquitos.

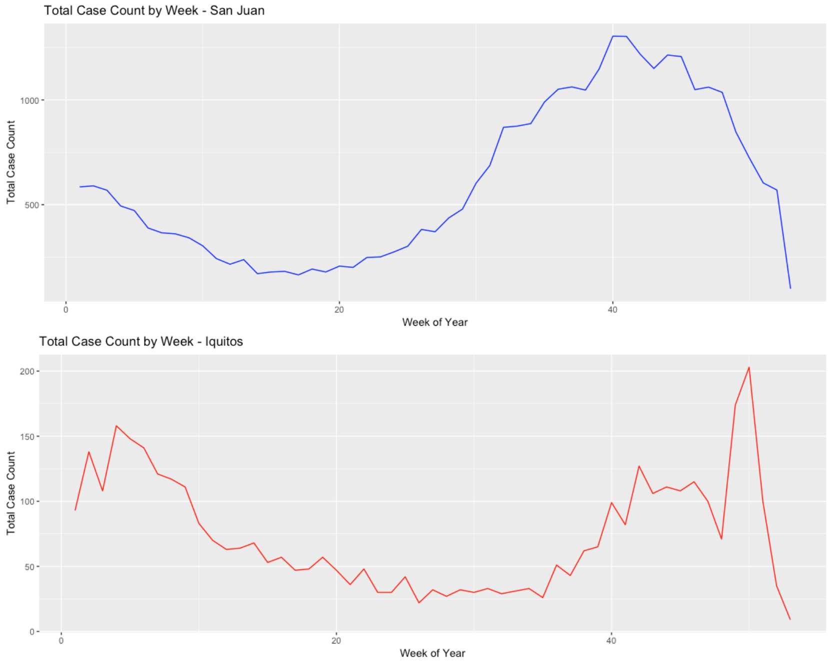

*Figure 2: Total case count by week for each city*

Figure 3 shows the correlation bar plots for each predictor variable against the response variable, *total_cases*.

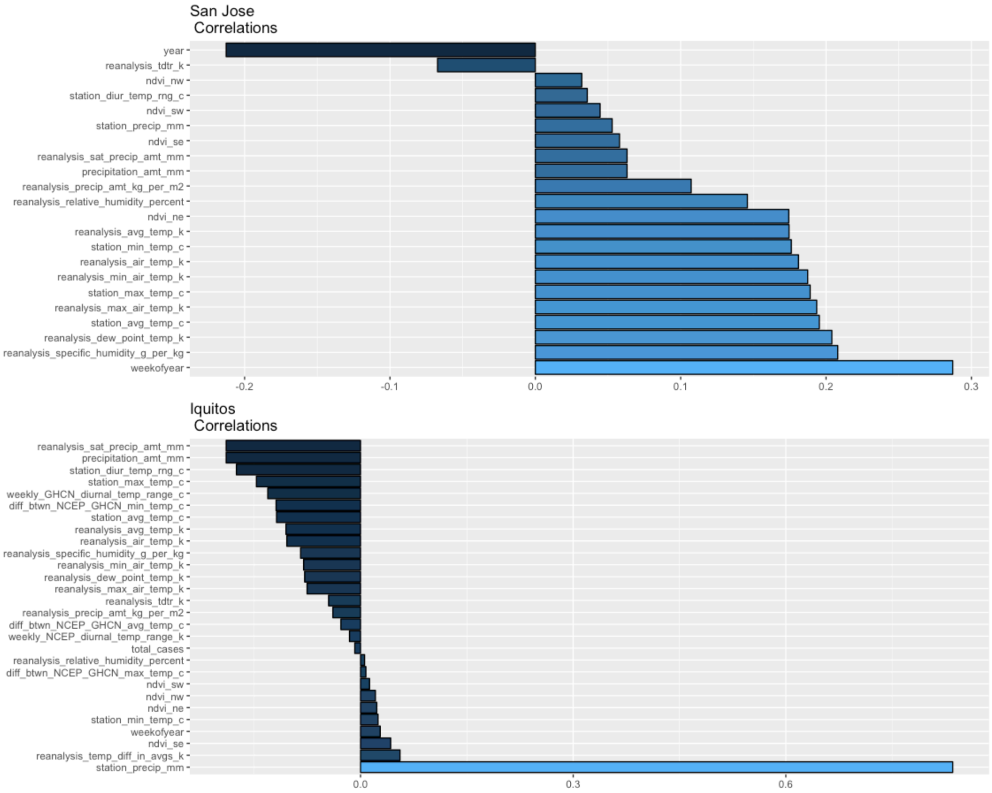

*Figure 3: Correlation bar plot by city*

Note how strongly the variable, *station_precip_mm*, is correlated to *total_cases* for the city of Iquitos. In contrast, this variable has a fairly weak correlation for the city of San Juan. Interestingly, for the city of San Juan, the variable *year* seems to have a negative correlation to *total_cases*. Additionally, all of the predictors for San Juan do not seem to have strong positive or strong negative correlations – they are below 0.3 or above -0.3.

## New Variable Creation

Some of the default predictors tend to have similar names, but they provide information from different sources. Therefore, the new variables constructed attempt to derive the difference between the sources. Table 2 summarizes the new variables that were created.

*Table 2: Summary of derived variables*

| Variable | Description |
|---|---|
| precipitation_diff | station_precip_mm - precipitation_amt_mm |
| weekly_GHCN_diurnal_temp_range_c | station_max_temp_c - station_min_temp_c |
| weekly_NCEP_diurnal_temp_range_k | reanalysis_max_air_temp_k - reanalysis_min_air_temp_k |
| diff_btwn_NCEP_GHCN_max_temp_c | (reanalysis_max_air_temp_k - 273.15) - station_max_temp_c |
| diff_btwn_NCEP_GHCN_min_temp_c | reanalysis_min_air_temp_k - 273.15 - station_min_temp_c |
| diff_btwn_NCEP_GHCN_avg_temp_c | reanalysis_avg_temp_k - 273.15 - station_avg_temp_c |
| reanalysis_temp_diff_in_avgs_k | (reanalysis_air_temp_k) - (reanalysis_avg_temp_k) |

Figure 4 shows the updated correlation bar plot with the new variables.

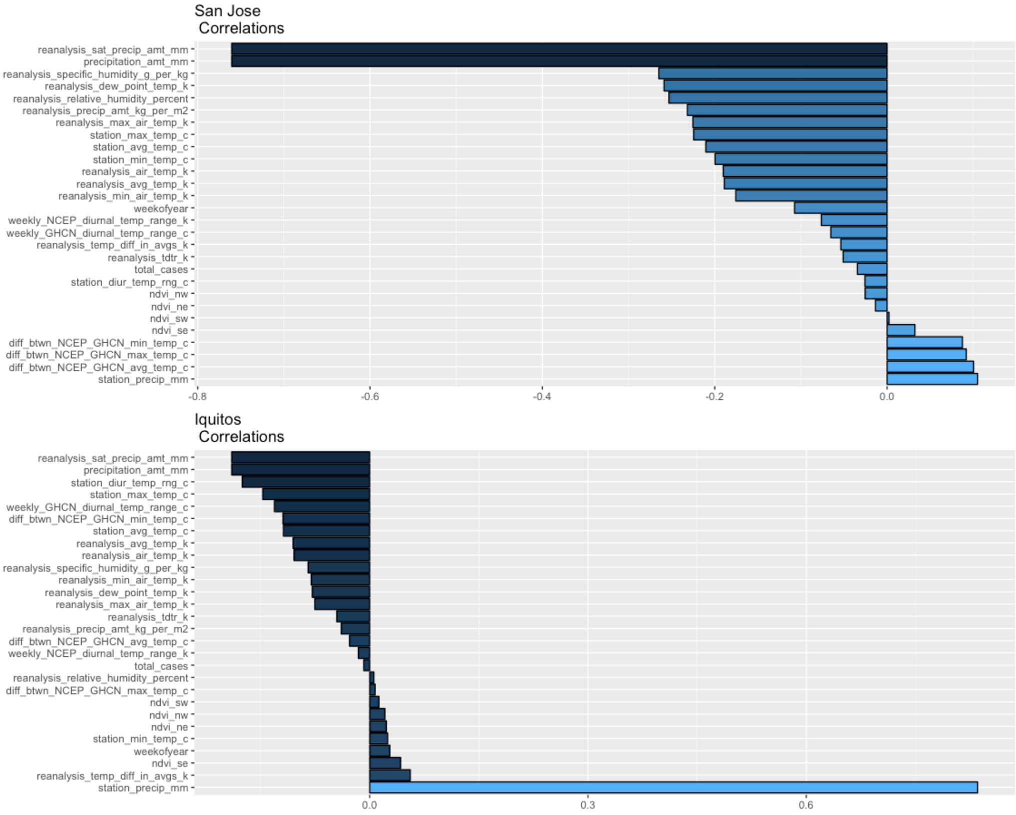

*Figure 4: Updated correlation bar plot with derived variables*

The most glaring difference, for the city of San Jose, is the strong negative correlation found with the predictor *reanalysis_sat_precip_amt_mm*. This change occurred with the creation of the new variables. For the city of Iquitos, the variable, *station_precip_mm*, continues to be strongly correlated to the response variable, *total_cases*.

### Outlier management

For this iteration of the analysis, no outlier management techniques were undertaken. The primary reason for this was to develop models with actual values in place and understand their performance. Outlier management will be reserved for future steps.

## Literature

This analysis is focused on predicting (i.e., forecasting) the number of dengue fever cases in the cities of San Juan and Iquitos. In the following section (Type of Models), five different models are discussed. This section discusses academic (from peer reviewed journals) papers that utilize similar models for classification.

### ARIMA/Random Forest

In the article, "Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks", the authors compare an ARIMA strategy to a random forest strategy. The objective of their research is somewhat similar to what is being undertaken with this activity. During their testing, they found that the random forest model outperformed a retrospective ARIMA model. Nevertheless, they did find that both models were not as satisfactory. For instance, the ARIMA model predicted negative cases – which is not possible. (Kane, Price, Scotch, & Rabinowitz, 2014). However, their research into both ARIMA and random forest suggests that random forests may be able to detect patterns at a more complex level than ARIMA.

### XGBoost

In the article, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring", the authors discuss using decision trees to predict credit scores. The most interesting aspect of this model is the discussion surrounding hyper-parameters. Hyper-parameters are a powerful way to 'simulate' multiple different tuning parameter values to obtain a model that is very good. In their research, the authors found that a Bayesian hyper-parameter optimization is far superior to other approaches, such as a grid search (Xia, Liua, Li, & Liu, 2017).

### Neural Network

In the article, "Comparative study among different neural net learning algorithms applied to rainfall time series", the authors use neural networks to forecast rainfall. Many of the approaches that the authors discuss are quite advanced, but their conclusion was simply, "artificial neural network is a suitable predictive tool for average monsoon rainfall"

(Chattopadhyay & Chattopadhyay, 2008). Neural network models, like decision tree models, are quite complex and very powerful.

### Negative Binomial

In the article, “Using the negative binomial distribution to model over dispersion in ecological count data”, the authors describe the use of a negative binomial regression model in bird migration (as an example). Recall from the section Type of Models, the variance of the variable, *total_cases*, is far higher than the mean. Therefore, a negative binomial model would be more appropriate than a Poisson model. In their research, the authors also describe how over dispersion is a common occurrence in nature (Lindén & Mäntyniemi, 2011).

### Negative Binomial

In the paper, “Modeling Forest Fire Occurrences Using Count-Data Mixed Models in Qiannan Autonomous Prefecture of Guizhou Province in China”, the authors use a negative binomial model, but with a twist. In their research, the authors use mixed-effects to improve the predictability. They also used hurdle models along with zero-inflated binomial regression models to clarify the effects of introducing mixed-effects. This strategy enables randomness to be introduced to the fixed-events and thereby increasing the ability of the model to perform well. According to the authors, the occurrence of forest fires can be unpredictable due to a variety of factors other than just wind (Xiao, Zhang, & Ji, 2014). This article also serves as inspiration for next steps.

## Type of Models

Several different types of models were constructed, but only five of these models were chosen to be submitted on Driven Data. The five models are:

- Negative Binomial
- Extreme Gradient Boosting (XGBoost)
- Random Forests
- ARIMA
- Neural Net

### Negative Binomial

Recall that the response variable is essentially counting the number of dengue fever cases in two cities. According to the author of “Generalized Linear Models An Applied Approach”, “When a dependent variable is over dispersed, many statistical modelers prefer to use…a negative binomial regression model” (Hoffman, 2004). From TABLE XX, the average number of cases for San Juan and Iquitos, respectively, are 19 and 5. However, the variances are (for San Juan and Iquitos respectively) approximately 2640 and 115.9. Therefore, a negative binomial model is a great model to consider.

### Extreme Gradient Boosting (XGBoost)

The XGBoost model was chosen due to its veracity and its complex learning abilities. Recall from the VISUAL EXPLORATION section that there appears to be numerous behaviors that affect the number of dengue fever cases. Decision trees and boosted models are able to capture nuances in the datasets and produce accurate models.

### Random Forests

Random forests is an interesting approach as this machine learning staple creates 'tiny' trees that are not very deep (in terms of branches). Furthermore, the algorithms create thousands of trees (each one could virtually use any number of predictors and no tree necessarily has all possible predictors) very quickly. According to the authors of "An Introduction to Statistical Learning with Applications in R", "random forests [are] an improvement over bagged trees by way of a small tweak that decorrelates the trees" (James, Witten, Hastie, & Tibshirani, 2015). This is a clear reason why bagging models were not considered since bagging models will enable the use of all predictors in a tree.

### ARIMA

ARIMA is an interesting concept in which autoregression and moving average strategies are combined (Hyndman & Athanasopoulos, 2013). This approach was used with the assumption that the information in this dataset is a univariate time series. The results of this model will clearly demonstrate some of the limitations of assuming that the total cases of dengue fever of each city is univariate.

### Neural Network

Neural networks, like decision trees, are quite informative and powerful. According to the text, "Applied Predictive Modelling", "the outcome is modeled by an intermediary set of unobserved variables" (Kuhn & Johnson, 2013). In this iteration of the analysis, a neural network model was constructed to help clarify its performance against the other machine learning techniques such as boosting and random forests (both of which are advanced decision trees).

## Formulation

All the models were constructed using R. The following are the key packages used within R:

- Caret
- MASS
- XGBoost
- Forecast

The caret package is quite powerful and enables easier cross-validation for any type of model. One of the control parameters for caret is the trainControl function. With this control parameter, XGBoost and random forest models utilized repeated cross validation, using 10 folds and repeated three times. The cross-validation approach enables a stronger learning capability for the models since a small subset of the training data can be used for validation prior to the validation dataset being used. Since this is done inherently within the caret package, the final model can be considered a decent model with the given parameters.

The forecast package enabled the creation of ARIMA and the neural network model. Both models were created using fundamental principles with no cross-validation.

Different datasets were constructed for use in each model. However, the random forest and xgboost models utilized the same datasets. Recall from earlier that a model was constructed for each city independently. For the machine learning models (xgboost and random forest), the variable *year* and *week_start_date* were removed. No scaling or variable transformation was undertaken. Furthermore, the training and validation split was 80/20 – 80% of the information was used to train the model and 20% of the data were used to validate the model. The 20% of the data (used in the validation) were not reintroduced to the model as that would constitute a potential leakage that could result in false perceptions.

For the negative binomial model, only the variables that had absolute correlation value[1] greater than 0.1. This resulted in a substantially smaller dataset, but the reasoning was that if the variable has a correlation between -0.1 and 0.1, then it may not be as useful. Interaction terms were not considered for the first iteration of this analysis.

For the ARIMA model, all predictors were not considered and the variable *total_cases* was converted into a univariate time series. This is in contrast to the other models where the other variables were used to predict the total number of cases. A grid search was also done to find the optimal *p*, *d*, and *q* values (autoregression, differencing[2], and moving average order respectively). As the grid search commenced, the mean absolute error (MAE) was calculated and the *p* and *q* values were capped at 10 whereas the d value was capped at 2. The MAE, however, was assessed against the validation dataset in the grid search. The thought process was that if the ARIMA values are fitted against the validation dataset, the training dataset would be overfitted, however, the validation dataset would be fitted correctly.

For the neural network model, all of the predictors were used and they were centered and scaled. Recall from TABLE XX that many of the variables have different ranges and the contexts vary greatly. For instance, the temperatures (the ones in Celsius) have a different scale than the precipitation. Therefore, by centering and scaling, the values in each predictor can be close to normal (with a mean of 0).

## Performance/Accuracy

In order to assess the models, the mean absolute error (MAE) was calculated for each city. On Driven Data, the submitted models were also assessed by their MAE. The challenge is that the

[1] The absolute value means that -0.2 is equal to 0.2 (for example).

[2] Differencing is necessary in order to ensure a stationary time series. Figure YY clearly shows that the time series are not stationary for either city.

derived MAE values are not as meaningful since the predicted case counts for each city have to be combined into one file prior to submission on Driven Data.

Table 3 summarizes the results of each model.

*Table 3: Summary of MAE for each model*

| Model | Validation dataset MAE (San Jose, Iquitos) | Driven Data Test MAE |
|---|---|---|
| Negative Binomial | 21.553, 5.568 | 35.0288 |
| XGBoost | 16.592, 5.833 | 27.2212 |
| Random Forest | 17.122, 5.490 | 27.2212 |
| ARIMA | 10.851, 1.656 | 71.9663 |
| Neural Network | 100.78, 4.392 | 33.1971 |

Since the Driven Data MAE is of premier significance, the XGBoost and Random Forest models perform the best. However, based on the validation dataset MAE, it would appear that the ARIMA model performed the best. Surprisingly, the neural network model performed terribly for cases in San Jose, but performed the best of all models for Iquitos.

## Limitations

There are two key limitations to the models constructed complexity and optimization. Random forest, XGBoost, and neural networks are complex models that are not very easy to explain. For instance, the random forest and XGBoost approaches constructed thousands of trees, but there is no distinctive equation that can be documented and easily explained. Rather, the explanation of each tree would result in a convoluted explanation of the logic and the randomness. Similarly, the neural network model is also convoluted and is essentially a network of non-linear equations to model the final response.

One key area to explore would be the optimization of each of these algorithms. For instance, the shrinkage parameter would enable the reduction of predictors and could improve a fit of a model while also reducing the likelihood of overfitting a model. Although it is difficult to overfit a model using boosting methods, it is still possible and the choice of parameters could affect the accuracy. Another avenue would be to investigate the type of external regressors used and their optimizations. It is very likely that not all features have been created that could better explain the different number of dengue cases.

Finally, the different data points seem to measure different things. For instance, it is not clear what the diurnal range is. Therefore, it was prudent to create a new diurnal temperature range that looked at the weekly maximum and minimum. On a personal note, this limitation could be overcome by focusing on the source of the data and how they are compiled. The Driven Data website was quite sparse on the background of the information, however, they did link to the NOAA websites – which would prove to be useful.

# Future Work

There are four areas of improvement for the next step:

- New data sources
- Feature creation
- Outlier management
- Model Optimization

## New Data Sources

The Driven Data competition did not explicitly discourage the use of external data sources. Two interesting data sources would be the GPS coordinates of the reported cases and the location of the different clinics treating the reported cases. This would enable a greater understanding of the impact of transmission and dengue fever localization. Another great resource would be the outcome of the reported case along with healing time. For instance, is it possible that within Iquitos (for example), a particular region is more susceptible for dengue fever cases?

## Feature creation

As previously mentioned, the weather data provided are quite convoluted and not easy to understand. Through increased research on the weather information itself, new features could be created. Furthermore, with the ability to bring in external data sources, that could result in even more meaningful information. Another feature would be to understand abnormal weather patterns or other catastrophic events of nature that may have influenced the number of dengue fever cases.

## Outlier Management

As previously mentioned, the presence of outliers was quite glaring. The risk with elimination of outliers is that the model may work well with both training and validation datasets, it may not work as well with the test dataset. Furthermore, the outliers themselves need a greater understanding. For instance, one of the weeks in San Juan had over 400 reported cases. Why is that? One of the benefits of machine learning algorithms is their inherent ability to account for outliers. However, if outliers are managed properly, these algorithms may experience improvements in their accuracy.

## Model Optimization

Earlier, a note was made about the lack of interaction terms in this first analysis. Interaction terms would be an excellent way to understand how other variables (e.g., ones with low correlation) can be instrumental when combined with other variables. This type of optimization could provide even more insight than the default predictors. Another approach would be to leverage mixed-effects (the introduction of randomness) to help improve the overall model's ability to forecast dengue fever cases. Other approaches include the ARCH-GARCH method and Hurdle models.

## Learning

The key takeaway of this activity has been the pleasure of working with weather data. One of my projects at work this summer will be incorporating weather data to understand machine (i.e., combine) behavior. I was actually quite ecstatic to use weather data in this project. However, I am also overwhelmed with the amount of information that exists regarding weather and the fact that there are so many different type of weather data available. I am looking forward to continuing the study of weather data as it will greatly help me at work.

Another important lesson for me was the difference between univariate time series along with using neural network with external regressors. This past quarter in PREDICT 413 has been an incredible learning experience and I have absolutely enjoyed applying my skills on real world data. The incorporation of ARIMA and neural network in this paper was definitely fun and a great learning experience.

## Works Cited

Chattopadhyay, S., & Chattopadhyay, G. (2008, April 18). Comparative study among different neural net learning algorithms applied to rainfall time series. *Royal Meteorological Society, 15*(2), 273-280.

Driven Data. (2016). *DengAI: Predicting Disease Spread*. Retrieved May 28, 2017, from Driven Data: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/

Gilbert, J. (2016, February 16). *Argentina Battles Major Outbreak of Dengue as Mosquito Population Swells*. Retrieved May 28, 2017, from New York Times: https://www.nytimes.com/2016/02/18/world/americas/argentina-battles-major-outbreak-of-dengue-as-mosquito-population-swells.html?action=click&contentCollection=Health&module=RelatedCoverage&region=EndOfArticle&pgtype=article

Hoffman, J. P. (2004). *Generalized Linear Models An Applied Approach.* New York, New York: Pearson Education Inc.

Hyndman, R. J., & Athanasopoulos, G. (2013). *Forecasting: principles and practice*. Retrieved May 27, 2017, from Forecasting: principles and practice: http://otexts.org/fpp/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R* (6th Edition ed.). New York, New York: Springer Science + Business.

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014, August 13). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, 276.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modelling* (5th Edition ed.). New York, New York: Springer Science+Business Media.

Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecological Society of America*, 1414-1421.

Xia, Y., Liua, C., Li, Y., & Liu, N. (2017, February 9). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Elsiever Expert Systems With Applications*, 225-241.

Xiao, Y., Zhang, X., & Ji, P. (2014, June 10). Modeling Forest Fire Occurrences Using Count-Data Mixed Models in Qiannan Autonomous Prefecture of Guizhou Province in China. *PLOS One*.

# Appendix

*Table 4: Data dictionary for default predictor variables*

| Predictor Variable | Description | Source |
|---|---|---|
| *year* | Year of the date | N/A |
| *weekofyear* | Week number of the year | N/A |
| *week_start_date* | Date given in yyyy-mm-dd format | N/A |
| *station_max_temp_c* | Maximum temperature | NOAA's GHCN |
| *station_min_temp_c* | Minimum temperature | NOAA's GHCN |
| *station_avg_temp_c* | Average temperature | NOAA's GHCN |
| *station_precip_mm* | Total precipitation | NOAA's GHCN |
| *station_diur_temp_rng_c* | Diurnal temperature range | NOAA's GHCN |
| *precipitation_amt_mm* | Total precipitation | PERSIANN satellite |
| *reanalysis_sat_precip_amt_mm* | Total precipitation | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_dew_point_temp_k* | Mean dew point temperature | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_air_temp_k* | Mean air temperature | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_relative_humidity_percent* | Mean relative humidity | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_specific_humidity_g_per_kg* | Mean specific humidity | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_precip_amt_kg_per_m2* | Total precipitation | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_max_air_temp_k* | Maximum air temperature | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_min_air_temp_k* | Minimum air temperature | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_avg_temp_k* | Average air temperature | NOAA's NCEP Climate Forecast System Reanalysis |
| *reanalysis_tdtr_k* | Diurnal temperature range | NOAA's NCEP Climate Forecast System Reanalysis |
| *ndvi_se* | Pixel southeast of city centroid | NOAA's CDR Normalized Difference Vegetation Index |
| *ndvi_sw* | Pixel southwest of city centroid | NOAA's CDR Normalized Difference Vegetation Index |
| *ndvi_ne* | Pixel northeast of city centroid | NOAA's CDR Normalized Difference Vegetation Index |
| *ndvi_nw* | Pixel northwest of city centroid | NOAA's CDR Normalized Difference Vegetation Index |

*Table 5: Descriptive statistics for default predictor variables for San Juan*

| Variable | Minimum | Maximum | Average | Median |
|---:|---|---|---|---|
| **year** | 1990 | 2008 | 1999 | 1999 |
| **weekofyear** | 1 | 53 | 26.5 | 26.5 |
| **week_start_date** | 1990-04-30 | 2008-04-22 | 1999-04-26 | 1999-04-26 |
| **ndvi_ne** | -0.40625 | 0.4934 | 0.058466 | 0.05935 |
| **ndvi_nw** | -0.4561 | 0.4371 | 0.06599 | 0.06595 |
| **ndvi_se** | -0.01553 | 0.39313 | 0.17796 | 0.17759 |
| **ndvi_sw** | -0.06346 | 0.38142 | 0.16614 | 0.16737 |
| **precipitation_amt_mm** | 0 | 390.6 | 35.45 | 20.61 |
| **reanalysis_air_temp_k** | 295.9 | 302.2 | 299.2 | 299.2 |
| **reanalysis_avg_temp_k** | 296.1 | 302.2 | 299.3 | 299.4 |
| **reanalysis_dew_point_temp_k** | 289.6 | 297.8 | 295.1 | 295.4 |
| **reanalysis_max_air_temp_k** | 297.8 | 304.3 | 301.4 | 301.5 |
| **reanalysis_min_air_temp_k** | 292.6 | 299.9 | 297.3 | 297.5 |
| **reanalysis_precip_amt_kg_per_m2** | 0 | 570.5 | 30.45 | 21.3 |
| **reanalysis_relative_humidity_percent** | 66.74 | 87.58 | 78.57 | 78.67 |
| **reanalysis_sat_precip_amt_mm** | 0 | 390.6 | 35.45 | 20.61 |
| **reanalysis_specific_humidity_g_per_kg** | 11.72 | 19.44 | 16.54 | 16.83 |
| **reanalysis_tdtr_k** | 1.357 | 4.429 | 2.514 | 2.45 |
| **station_avg_temp_c** | 22.84 | 30.07 | 27 | 27.21 |
| **station_diur_temp_rng_c** | 4.529 | 9.914 | 6.753 | 6.757 |
| **station_max_temp_c** | 26.7 | 35.6 | 31.6 | 31.7 |
| **station_min_temp_c** | 17.8 | 25.6 | 22.59 | 22.8 |
| **station_precip_mm** | 0 | 305.9 | 26.8 | 17.8 |
| **total_cases** | 0 | 461 | 34.18 | 19 |

*Table 6: Descriptive statistics for default predictor variables for Iquitos*

| Variable | Minimum | Maximum | Average | Median |
|---:|:---:|:---:|:---:|:---:|
| **year** | 2000 | 2010 | 2005 | 2005 |
| **weekofyear** | 1 | 53 | 26.5 | 26.5 |
| **week_start_date** | 2000-07-01 | 2010-06-25 | 2005-06-28 | 2005-06-28 |
| **ndvi_ne** | 0.06173 | 0.50836 | 0.26355 | 0.26355 |
| **ndvi_nw** | 0.03586 | 0.45443 | 0.23844 | 0.2325 |
| **ndvi_se** | 0.02988 | 0.53831 | 0.24991 | 0.24976 |
| **ndvi_sw** | 0.06418 | 0.54602 | 0.26637 | 0.26187 |
| **precipitation_amt_mm** | 0 | 210.83 | 64.19 | 60.47 |
| **reanalysis_air_temp_k** | 294.6 | 301.6 | 297.9 | 297.8 |
| **reanalysis_avg_temp_k** | 294.9 | 302.9 | 299.1 | 299.1 |
| **reanalysis_dew_point_temp_k** | 290.1 | 298.4 | 295.5 | 295.9 |
| **reanalysis_max_air_temp_k** | 300 | 314 | 307.1 | 307.1 |
| **reanalysis_min_air_temp_k** | 286.9 | 296 | 292.9 | 293.1 |
| **reanalysis_precip_amt_kg_per_m2** | 0 | 362.03 | 57.59 | 46.45 |
| **reanalysis_relative_humidity_percent** | 57.79 | 98.61 | 88.65 | 90.94 |
| **reanalysis_sat_precip_amt_mm** | 0 | 210.83 | 64.19 | 60.47 |
| **reanalysis_specific_humidity_g_per_kg** | 12.11 | 20.46 | 17.1 | 17.43 |
| **reanalysis_tdtr_k** | 3.714 | 16.029 | 9.202 | 8.964 |
| **station_avg_temp_c** | 21.4 | 30.8 | 27.52 | 27.57 |
| **station_diur_temp_rng_c** | 5.2 | 15.8 | 10.54 | 10.55 |
| **station_max_temp_c** | 30.1 | 42.2 | 33.99 | 34 |
| **station_min_temp_c** | 14.7 | 24.2 | 21.2 | 21.35 |
| **station_precip_mm** | 0 | 543.3 | 62.44 | 45.65 |
| **total_cases** | 0 | 116 | 7.565 | 5 |

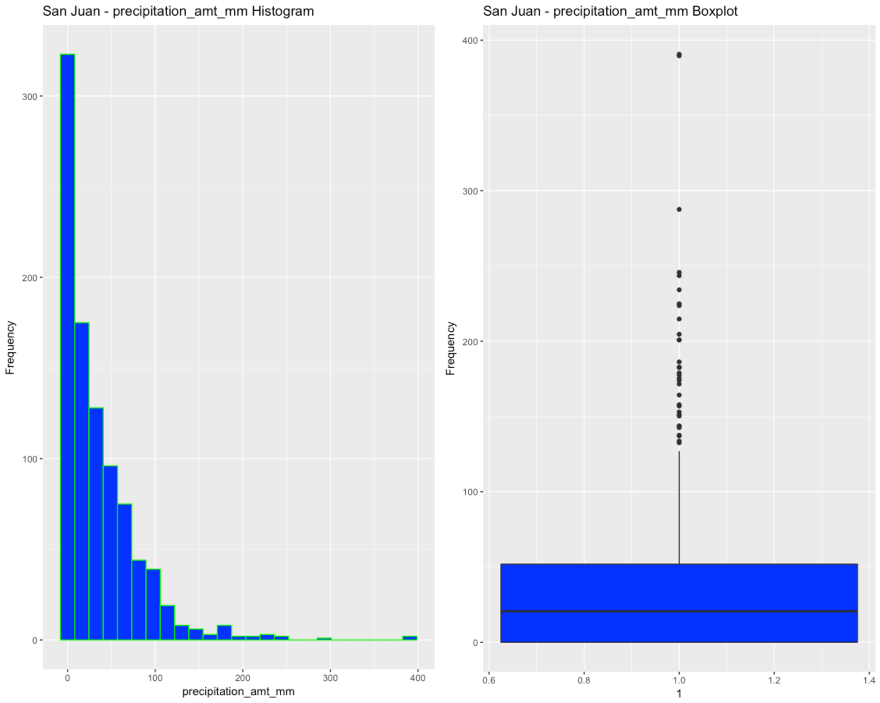

*Figure 5: Histogram and boxplot for variable precipitation_amt_mm for San Juan*

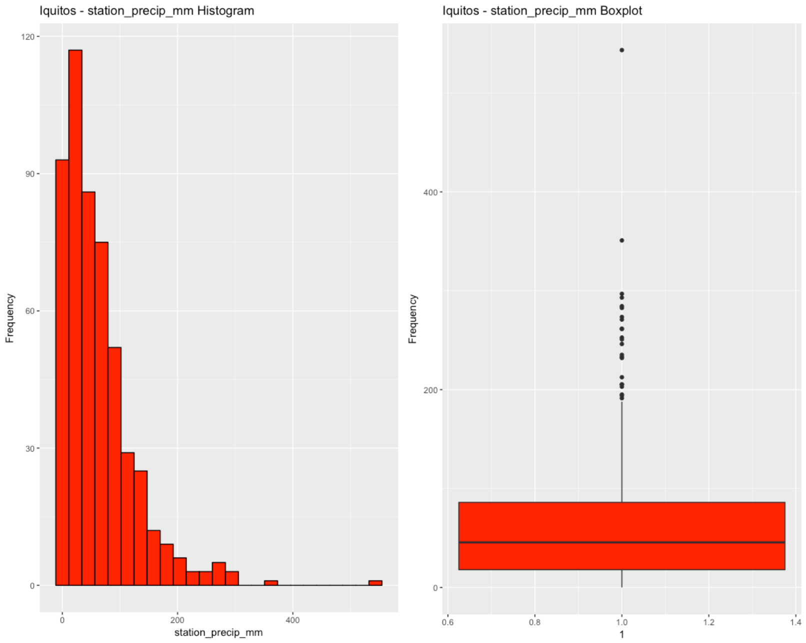

*Figure 6: Histogram and boxplot for variable station_precip_mm for Iquitos*