Predictive Modeling of Dengue Fever Epidemics:
A Neural Network Approach

Carlos Sathler
I590 – Data Science for Drug Discovery, Health and Translational Medicine
Professor Joanne Luciano
December 10, 2017

Abstract

Dengue is a disease caused by four types of related viruses transmitted by a mosquito, most commonly the Aedis aegypti. In its less severe form infected patients will experience flu-like symptoms that vary from mild to intense, but severe dengue or, Dengue Hemorrhagic Fever, can be fatal without proper medical care. The disease is considered an alarming threat to the health of populations spanning millions of people living in tropical and subtropical areas of the globe where the mosquito thrives. A large number of studies have confirmed that the incidence of dengue is positively correlated with climatic conditions, specifically, temperature, humidity and precipitation levels. Many of these studies include quantitative models correlating climate variables with the incidence of dengue cases. The quantitative models invite the question: how well would we be able to predict future cases of the disease based on climate variables that are included in weather forecasts? To answer this question several departments of the U.S. Federal Government have joined efforts to create the Dengue Forecasting project, which makes climate and dengue data available to data scientists at large and challenges them to submit predictive models to help forecast future dengue epidemics. In our project we create several neural network predictive models and evaluate their performance against more common machine learning models.

Introduction

Dengue is a disease caused by four types of related viruses transmitted by a mosquito, most commonly the Aedis aegypti. In its less severe form infected patients will experience mild to intense flu-like symptoms, but severe dengue or, Dengue Hemorrhagic Fever, can be fatal without proper medical care. According to the World Health Organization (WHO), Dengue and Severe Dengue Fact Sheet "severe dengue affects most Asian and Latin American countries and has become a leading cause of hospitalization and death among children and adults in these regions."[1] Dengue incidence is common in tropical and sub-tropical areas of the globe; not only does the main vector for the disease, the Aedis aegypti mosquito, thrive in warm, humid environments with high precipitation levels, but also warm temperatures further increase the incidence of the disease by shortening viral replication in the mosquito. The Centers for Disease Control and Prevention (CDC) on its "Dengue and Climate" web page states that "in countries where transmission does routinely occur, short-term changes in weather, particularly temperature, precipitation, and humidity, are often correlated with dengue incidence."[2] Many researchers have explored and confirmed this correlation between climate and this disease[3,4,5,6]

---

[1] "Dengue and Severe Dengue." Dengue and Severe Dengue: Fact Sheet, World Health Organization, Apr. 2017, www.who.int/mediacentre/factsheets/fs117/en/.

[2] "Dengue and Climate" Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 27 Sept. 2012, www.cdc.gov/dengue/entomologyecology/climate.html.

[3] Hurtado Díaz, M., et al. "Impact of climate variability on the incidence of dengue in Mexico." Tropical medicine & international health 12.11 (2007): 1327-1337.

[4] Johansson, Michael A., et al. "Local and Global Effects of Climate on Dengue Transmission in Puerto Rico." PLoS Neglected Tropical Diseases, vol. 3, no. 2, 2009, doi:10.1371/journal.pntd.0000382.

[5] Morin, Cory W., Andrew C. Comrie, and Kacey Ernst. "Climate and dengue transmission: evidence and implications." Environmental health perspectives 121.11-12 (2013): 1264.

[6] Hii, Yien Ling et al. "Forecast of Dengue Incidence Using Temperature and Rainfall." Ed. Francis Mutuku. PLoS Neglected Tropical Diseases 6.11 (2012): e1908. PMC. Web. 28 Oct. 2017.

and in 2015 the US Department of Commerce released the Dengue Forecasting project[7] inviting data scientists to develop predictive models to forecast dengue using climate related data. In our project we will explore this data and will create predictive models of our own, with a focus on neural network models. We will submit our results to the "DengAI: Predicting Disease Spread" competition from DrivenData[8] to compare our model's performance, particularly the performance of our neural net models, against the results of other data scientists. The WHO in their publication "Global strategy for dengue prevention and control 2012-2020"[9] asserts that "Dengue morbidity can be reduced by implementing improved outbreak prediction and detection through coordinated epidemiological and entomological surveillance".[10] Models that quantitatively predict incidence of the disease based on climate data can potentially serve as one of the many tools to survey the risk of impending Dengue outbreaks.

[7] US Department of Commerce, NOAA, National Weather Service. "Dengue Forecasting." Dengue Forecasting, NOAA's National Weather Service, dengueforecasting.noaa.gov/.

[8] DrivenData. DengAI: Predicting Disease Spread, DrivenData, www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/.

[9] Global strategy for dengue prevention and control World Health Organization - Geneva: World Health Organization, 2012

[10] _____. Page 3.

Problem Statement

Create a predictive model of dengue cases for the cities of San Juan, Puerto Rico, and Iquitos, Peru, based on climate data using a neural network approach.

The problem is defined by the DrivenData DengAI competition as follows: "Your task is to predict the number of dengue cases each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation, and more."[8]

Importance

The Centers for Disease Control (CDC) estimates one third of the world population is exposed to the dengue virus and is at risk of contracting the disease. According to the CDC "As many as 400 million people are infected yearly."[11] The World Mosquito Program states that "500,000 cases develop into dengue hemorrhagic fever… which results in up to 25,000 deaths annually worldwide."[12]

The World Health Organization (WHO) in their publication "Global strategy for dengue prevention and control 2012-2020"[9] asserts that "Dengue morbidity can be reduced by implementing improved outbreak prediction and detection through coordinated epidemiological and entomological surveillance".[10] Models that quantitatively predict incidence of the disease based on climate data can potentially serve as one of the many tools to survey the risk of impending Dengue outbreaks.

---

[11] "Dengue." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Jan. 2016, www.cdc.gov/dengue/index.html.

[12] "Dengue." Dengue, World Mosquito Program, www.eliminatedengue.com/our-research/dengue-fever.

Objectives

The objects for this project are: (1) create a model capable of producing competitive results when compared to results obtained by other data scientists and aspiring data scientists participating in the DrivenData competition "DengAI: Predicting Disease Spread"[8] and (2) do so by using neural networks.

Figure 1 shows the leader board of the competition as of October 24, 2017.

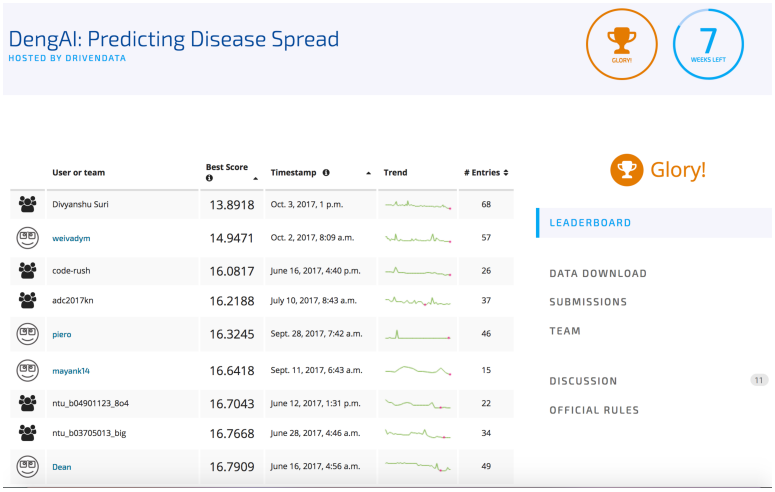


Figure 1 - DengAI: Predicting Disease Spread, Leaderboard[8]

Methodology

We will follow a typical machine learning process, as follows:

1. Obtain data
2. Perform exploratory data analysis
3. Choose a learning model
4. Preprocess the data for machine learning model (feature engineering)
5. Tune the model
6. Run the model on the test dataset
7. Capture predictions
8. Iterate for better results

Because the project is associated with a competition some predictions were loaded to the competition site for scoring and ranking.

Dataset

The project dataset is intended to support the prediction of dengue cases based on climate variables. The dataset is available in the NOAA website Dengue Forecasting project.[7] It can also be found in the Predicting Disease Spread competition website. [8]

Dataset Description and Feature Analysis

The dataset for the project is summarized in the table below. Note the substring "_unit" at the end of several features. It signifies the unit of measurement for the feature. For example, feature 6 is measured in Celsius; feature 16 in square meters.

ID	Feature Name	Feature Description	Type
1	city	City abbreviation ("iq" or "sj")	Location
2	Year	Year of observation	Timescale
3	Weekofyear	Week of the year	Timescale
4	week_start_date	Date given in yyyy-mm-dd format	Timescale
5	station_max_temp_c	Maximum temperature (Celcius)	Temperature
6	station_min_temp_c	Minimum temperature	Temperature
7	station_avg_temp_c	Average temperature	Temperature
8	station_precip_mm	Total precipitation	Precipitation
9	station_diur_temp_rng_c	Diurnal temperature range	Temperature
10	precipitation_amt_mm	Total precipitation	Precipitation
11	reanalysis_sat_precip_amt_mm	Total precipitation	Precipitation
12	reanalysis_dew_point_temp_k	Mean dew point temperature	Temperature
13	reanalysis_air_temp_k	Mean air temperature	Temperature
14	reanalysis_relative_humidity_percent	Mean relative humidity	Humidity
15	reanalysis_specific_humidity_g_per_kg	Mean specific humidity	Humidity
16	reanalysis_precip_amt_kg_per_m2	Total precipitation	Precipitation
17	reanalysis_max_air_temp_k	Maximum air temperature	Temperature
18	reanalysis_min_air_temp_k	Minimum air temperature	Temperature
19	reanalysis_avg_temp_k	Average air temperature	Temperature
20	reanalysis_tdtr_k	Diurnal temperature range	Temperature
21	ndvi_se	Pixel southeast of city centroid	Vegetation index
22	ndvi_sw	Pixel southwest of city centroid	Vegetation index
23	ndvi_ne	Pixel northeast of city centroid	Vegetation index
24	ndvi_nw	Pixel northwest of city centroid	Vegetation index

Initial analysis of the data revealed that the dataset should be partitioned in two sets, one for Iquitos data, and another for San Juan data. Figures 2.1, 2.2, 3.1 and 3.2 show data distribution in these partitions and time series for each feature.

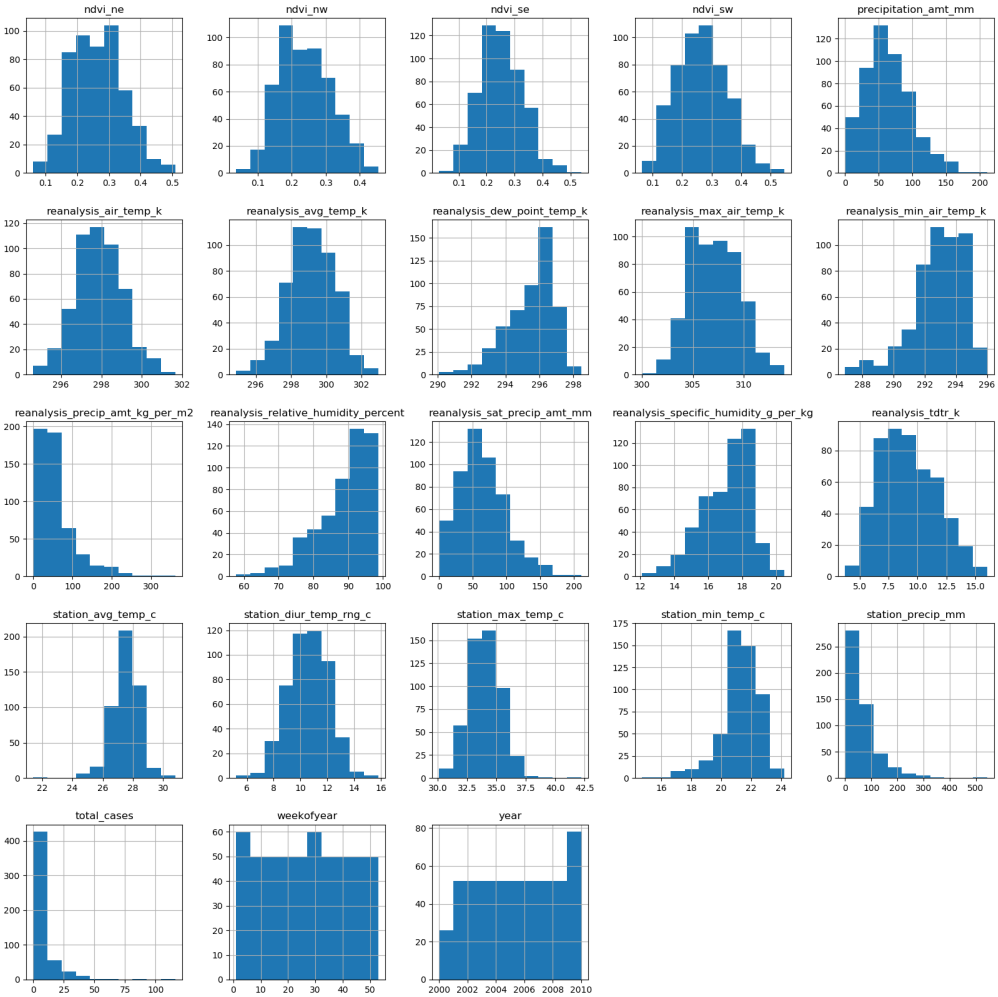


Figure 2.1 – Iquitos Data (Part 1)

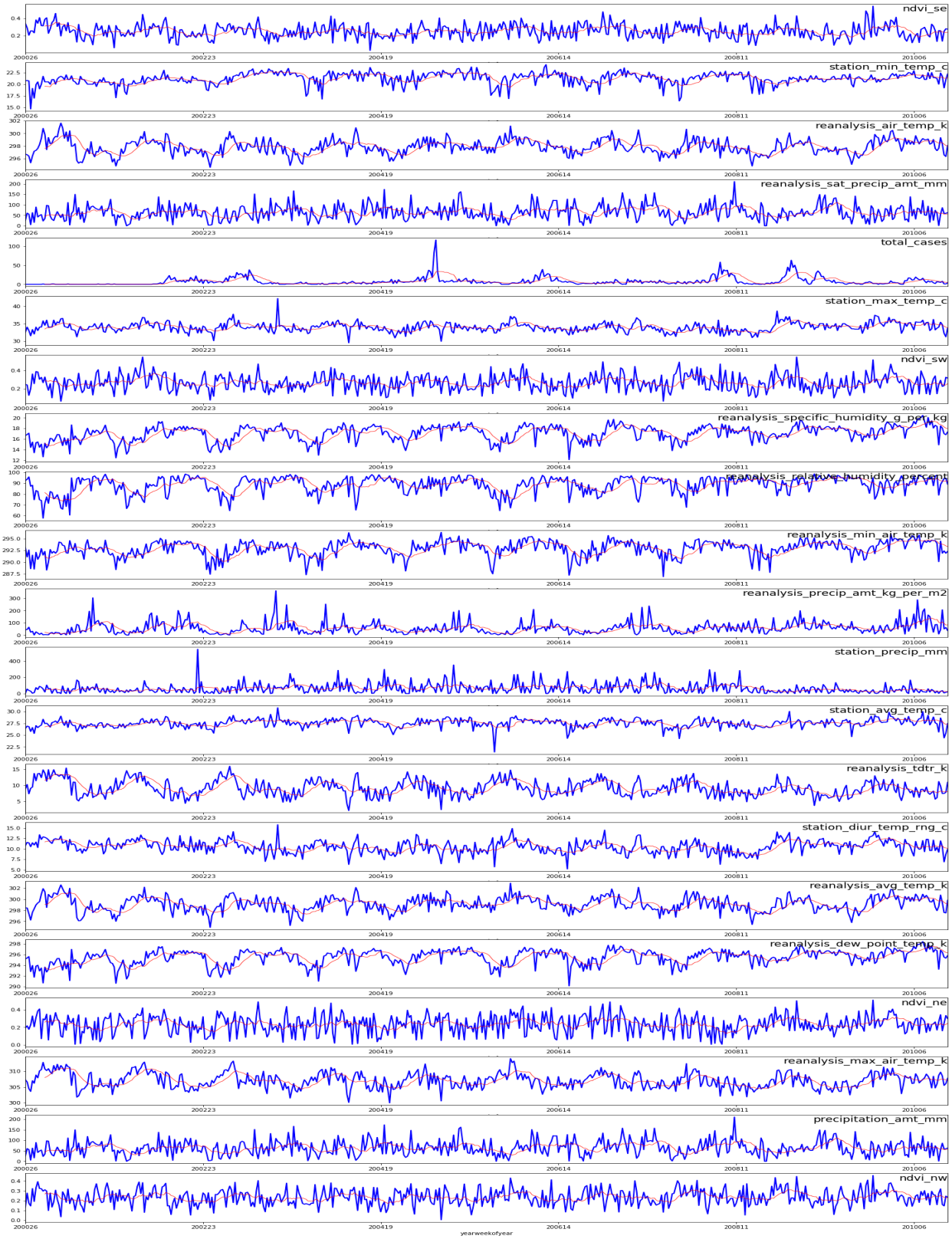


Figure 2.2 – Iquitos Data (Part 2)

A closer look at number of total cases/week for the city of Iquitos:

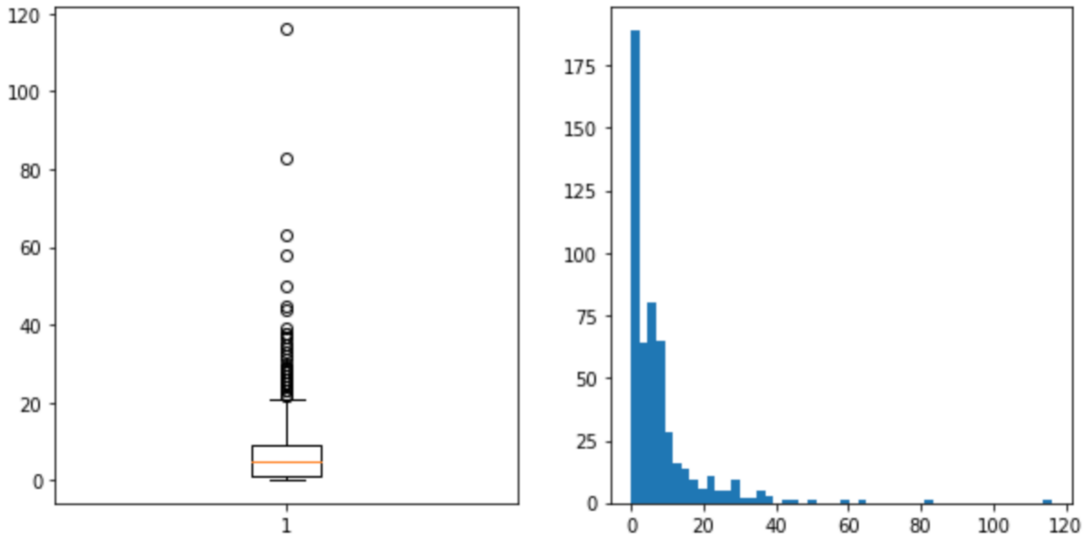


Figure 4 - Box plot and histogram for number of cases/week for Iquitos

Summary statistics for number of cases/week for Iquitos:

```
count    520.000000
mean       7.565385
std       10.765478
min        0.000000
25%        1.000000
50%        5.000000
75%        9.000000
95%       28.000000
max      116.000000
```

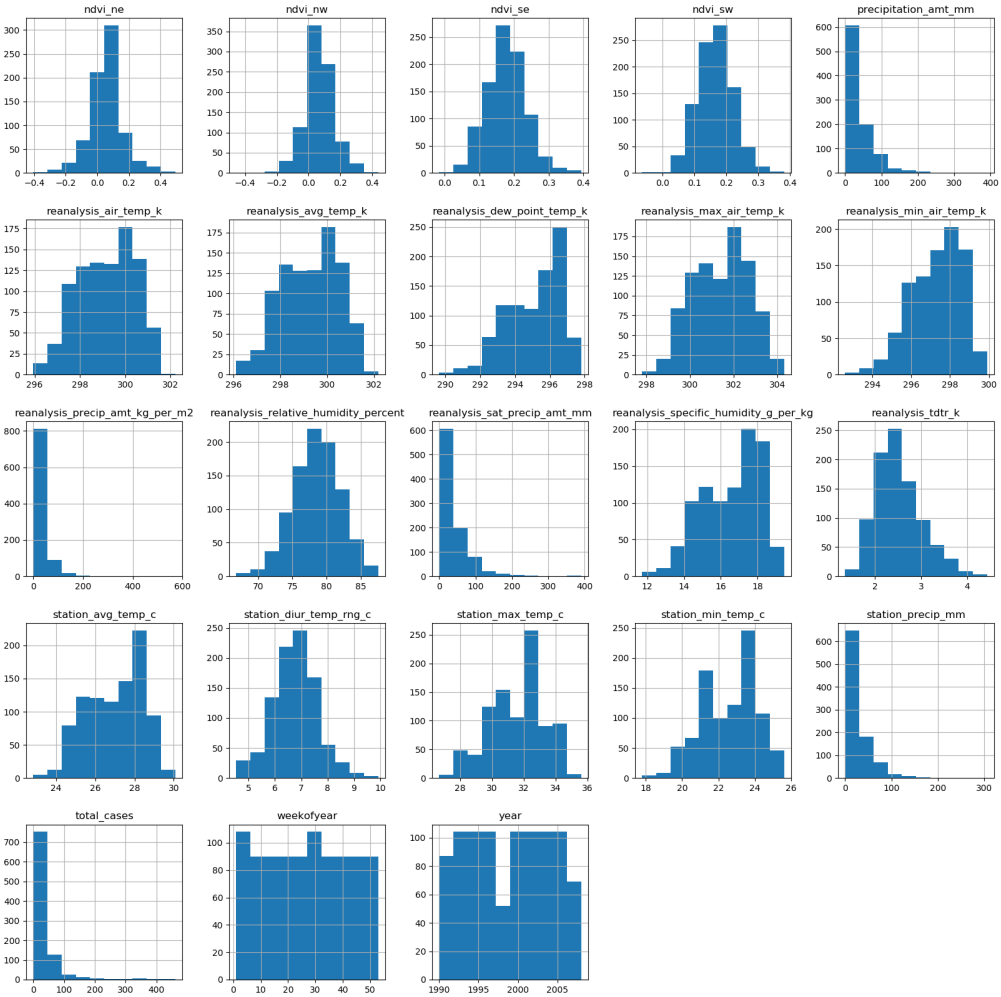


Figure 3.1 – San Juan Data (Part 1)

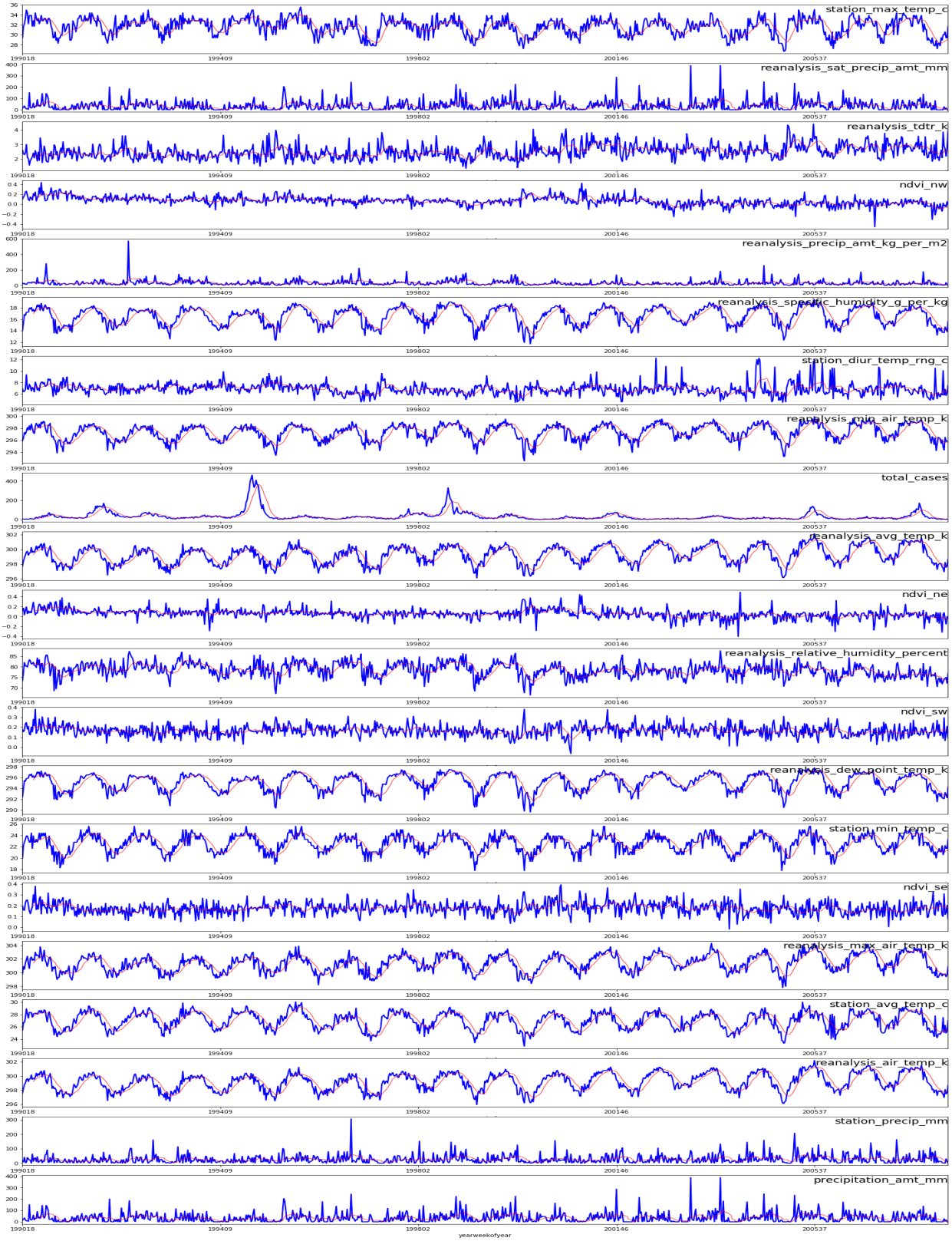


Figure 3.2 – San Juan Data (Part 2)

A closer look at number of total cases/week for the city of San Juan:

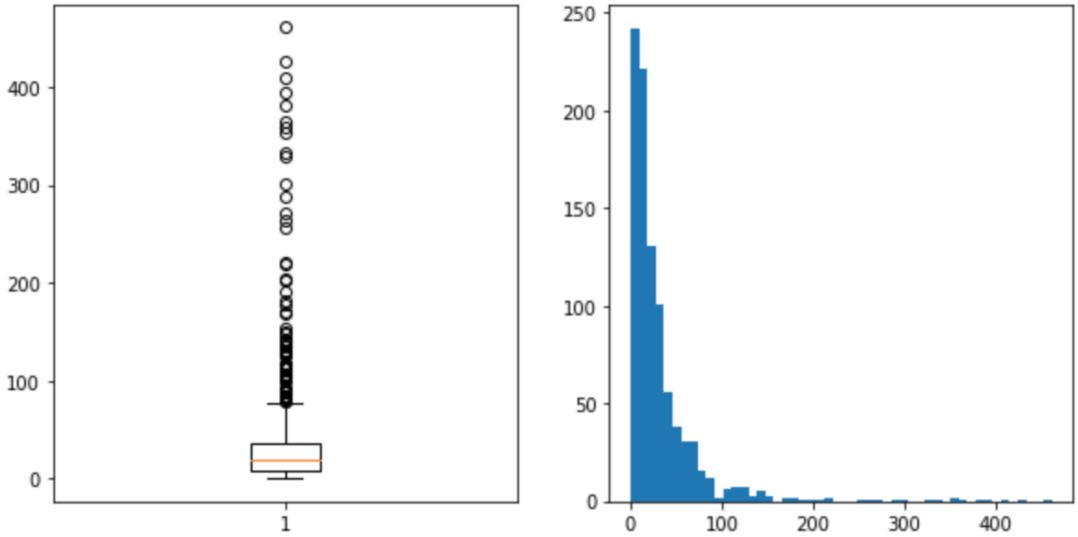


Figure 5 - Box plot and histogram for number of cases/week for San Juan

Summary statistics for number of cases/week for San Juan:

```
count    936.000000
mean      34.180556
std       51.381372
min        0.000000
25%        9.000000
50%       19.000000
75%       37.000000
95%      112.000000
max      461.000000
```

Data Types are as follows:

```
<class 'pandas.core.frame.DataFrame'>
Index: 1456 entries, 199018 to 201025
Data columns (total 25 columns):
city                                     1456 non-null object
year                                     1456 non-null int64
weekofyear                               1456 non-null int64
week_start_date                          1456 non-null object
ndvi_ne                                  1262 non-null float64
ndvi_nw                                  1404 non-null float64
ndvi_se                                  1434 non-null float64
ndvi_sw                                  1434 non-null float64
precipitation_amt_mm                     1443 non-null float64
reanalysis_air_temp_k                    1446 non-null float64
reanalysis_avg_temp_k                    1446 non-null float64
reanalysis_dew_point_temp_k              1446 non-null float64
reanalysis_max_air_temp_k                1446 non-null float64
reanalysis_min_air_temp_k                1446 non-null float64
reanalysis_precip_amt_kg_per_m2          1446 non-null float64
reanalysis_relative_humidity_percent     1446 non-null float64
reanalysis_sat_precip_amt_mm             1443 non-null float64
reanalysis_specific_humidity_g_per_kg    1446 non-null float64
reanalysis_tdtr_k                        1446 non-null float64
station_avg_temp_c                       1413 non-null float64
station_diur_temp_rng_c                  1413 non-null float64
station_max_temp_c                       1436 non-null float64
station_min_temp_c                       1442 non-null float64
station_precip_mm                        1434 non-null float64
total_cases                              1456 non-null int64
dtypes: float64(20), int64(3), object(2)
memory usage: 295.8+ KB
```

Data Visualization

Data visualization is captured in the following notebooks in GitHub:

File	Description	URL
dengueai_dviz.ipynb	Iquitos and San Juan data	https://github.com/csathler/IU_MSDS/blob/master/DSDHT/dengueai/code/dengueai_dviz.ipynb
dengueai_dviz_iq.ipynb	Iquitos data only	https://github.com/csathler/IU_MSDS/blob/master/DSDHT/dengueai/code/dengueai_dviz_iq.ipynb
dengueai_dviz_sj.ipynb	San Juan data only	https://github.com/csathler/IU_MSDS/blob/master/DSDHT/dengueai/code/dengueai_dviz_iq.ipynb

Data Cleaning and Pre-Processing

The following data cleaning steps and pre-processing were performed:

1. All NaN values were set to mean value for the feature based on weekofyear.
2. Feature "week_start_date" was dropped since timescale is set by year and weekofyear features.
3. Unit conversion: Temperatures in Kelvin were all converted to Celsius using the equation Tc = Tk – 273.15 (dataset column names were not changed).
4. Features "precipitation_amount_mm" and "reanalysis_sat_precip_amt_mm" were found to be 100% correlated (pearson), so we dropped the latter.
5. Feature"reanalysis_dew_point_temp_k" and "reanalysis_specific_humidity_g_per_kg" are 99.85% correlated (pearson), so we dropped the latter.

Time Window Considerations

It takes between 8 and 10 days for the Aedis aegypti to develop from egg to a full-grown mosquito. The mosquito eggs will be deposited at the edges on containers and stay there for long periods until exposed to water. The dengue female mosquito (the disease vector) typically lives between 3 and 4 weeks.[13]

A human bitten by a mosquito carrying the dengue virus will be most contagious between 4 and 9 days of the day when he/she was bitten, sometimes longer, up to 12 days. And symptoms of the disease typically appear 5 days after the bite.[13] The disease spreads as a result of a mosquito biting an infected individual, and then biting healthy person. As mosquito control increases, and more patients are treated, the number of cases decline.

With that in mind, we developed a function to "shift" the dataset features one or more weeks (up to 16) since we should expect a lag between weather conditions, number of infected people in prior weeks, and incidence of dengue fever each week. This lag is due to the life cycle of the Aedis aegypti mosquito and the life cycle of the dengue disease. Note that every shifted

[13] "Dengue Transmission." Nature News, Nature Publishing Group, www.nature.com/scitable/topicpage/dengue-transmission-22399758?sa=X&ved=0ahUKEwin843IxKLOAhVDnZQKHSsqDNQQ9QEIEDAA.

week is added to the feature set we input into our model. For example, a 2 period "shift" will create a dataset with F(t-2), F(t-1), Label(t). Where F(t) represents here the entire feature set at time "t". At a minimum all our models perform at least one "shift". All models were tested on a variety of shifts of the dataset.

Feature Set Considerations

We experimented a variety of subsets of the dataset features in our models, which can be summarized as follows:

1. Inclusion or not of feature "weekofyear".
2. Size of time window, which varied from 1 to 16 weeks
3. Autoregression: whether or not we included total number of cases of dengue in prior periods in our feature set

The figures below show autocorrelation of dengue cases to incidence of infections for 4 weeks prior to the current week's count of cases. We notice autocorrelation is particularly strong for San Juan.

	t-4	t-3	t-2	t-1	t
t-4	1.000000	0.725800	0.550666	0.505571	0.438564
t-3	0.725800	1.000000	0.723932	0.547935	0.506178
t-2	0.550666	0.723932	1.000000	0.722041	0.546579
t-1	0.505571	0.547935	0.722041	1.000000	0.721811
t	0.438564	0.506178	0.546579	0.721811	1.000000

Iquitos

	t-4	t-3	t-2	t-1	t
t-4	1.000000	0.955233	0.909080	0.854673	0.798112
t-3	0.955233	1.000000	0.955238	0.909080	0.854669
t-2	0.909080	0.955238	1.000000	0.955237	0.909079
t-1	0.854673	0.909080	0.955237	1.000000	0.955236
t	0.798112	0.854669	0.909079	0.955236	1.000000

San Juan

Type of Prediction

We attempted to predict total cases of dengue as a regression problem and also as a classification problem, following discretization of number of dengue cases per week.

Treatment of Outliers

We observed several features in the dataset show a long tail distribution, some of them with very extreme outliers. We wanted to make sure outliers were not impacting our models' ability to learn, so during the second phase of the project we created a different version of the dataset by artificially "shrinking" outliers to see if our results improved.

For features with long tail distributions, outliers were updated to match value at 95% percentile point. For symmetrically (normally) distributed features, we used the Grubbs' test to identify outliers, which were then updated so all data would fall within 95% interval of original distribution. We used the Python stats.skewtest module to identify features with a normal distribution.

Scaling

We experimented our models against scaled and non-scaled features. Feature scaling was performed with skilearn MinMax method. For neural net models we used [-1,+1] range; for all other models we used [0,1] range.

Results

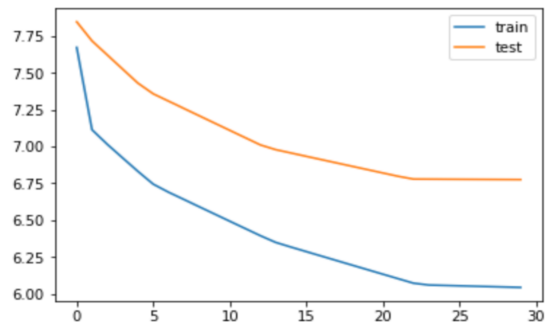
A summary of our results and submissions to the competition can be found at the end of this section. Model performance is measured with a mean absolute error score. Not all submissions are included in the table. Our best result was 22.8077, obtained with Bayesian Ridge regression for San Juan predictions, and weekly average for Iquitos. The code for the project can be found in the follow GitHub repository:

https://github.com/csathler/IU_MSDS/tree/master/DSDHT/dengueai/code

Phase I - Benchmarking

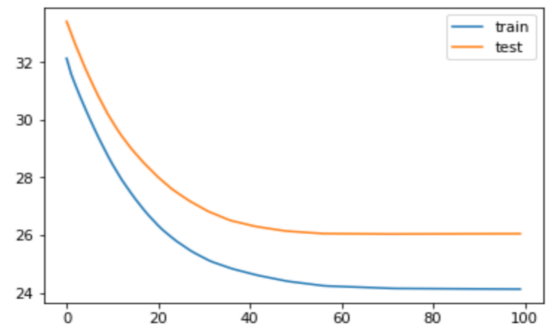
In the first phase of the project we compared results of Machine Learning linear regressors (available from Python's scikit-learn), long-short term memory recurrent neural networks (LSTM), and a simple average of dengue cases by week of year. We found Bayesian Ridge regression to be the most effective regressor. LSTM did not perform well in any of the many architectures (number of nodes and layers) we tried. Our best recurrent model implementation was with Gated Recurrent Unit (GRU) networks. But typically learning plateaued rather quickly for both Iquitos and San Juan data even for GRU, as seen in the plots below.

```
Final loss train: 6.042249753965554
Final loss valid: 6.775002200262887
```



Iquitos

```
Final loss train: 24.128345670133687
Final loss valid: 26.050581795384907
```

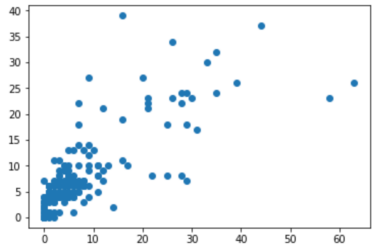


San Juan

For Adaboost with Random Tree Regressor we included total cases of dengue in the feature set for the first time. The results highlighted the predictive relevance of total cases relative to other features in the feature set. The figures below show the results. Y axis is prediction; X axis is actual number of cases. Time window is one week only. The last feature in the list under "Feature Importances" is total cases. Note that for Iquitos total cases is 64% relevant and for San Juan it is 95% relevant.

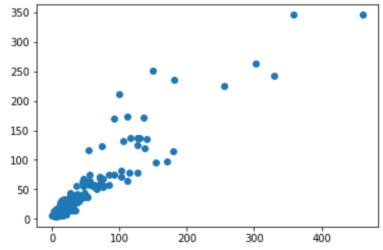
We analyzed R2 for our predictions on the validation partition of the training dataset. Notice the low R2 score for the city of Iquitos (R2=0.6), which highlights the poor performance of our model on that city's data.

```
Train mea: 2.0829016783136276
Valid mea: 4.1395348837209305
Variance score: 0.60
Feature Importances:
[ 0.03740463  0.03716446  0.02476863  0.01200519  0.00854032  0.01005905
  0.00945026  0.0520139   0.02178538  0.0193987   0.01466117  0.01476109
  0.01709193  0.01087095  0.01313661  0.01545407  0.02400761  0.01213547
  0.64529057]
```



Iquitos

```
Train mea: 4.908545512855319
Valid mea: 9.388349514563107
Variance score: 0.88
Feature Importances:
[  1.61865377e-03   2.36557338e-03   4.21717343e-03   2.26094640e-03
   3.57893507e-03   2.03580492e-03   1.71370576e-03   7.84346221e-03
   1.29074525e-03   1.73147293e-03   3.46304332e-03   2.13319763e-03
   1.87544713e-03   2.63107601e-03   3.28629603e-03   7.85804588e-04
   1.04498575e-03   2.64285283e-03   9.53480824e-01]
```



San Juan

Phase II – Dataset Without Outliers

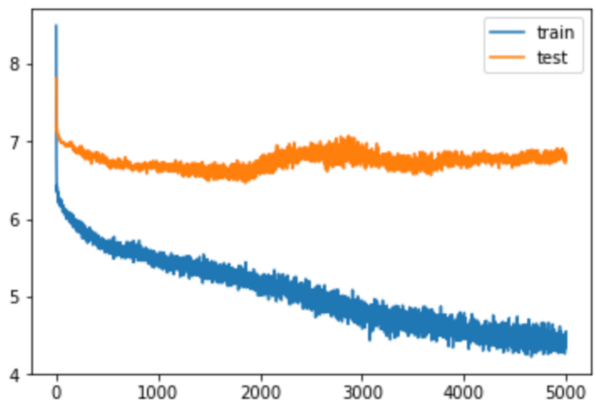
In the second phase of the project we compared results of our models against dataset versions with and without changes to outliers. There was not significant performance improvement for models fed data without outliers, so we abandoned the idea of removing outliers.

Phase III – Time Window and Autocorrelation Experiments

In the third phase of the project we systematically ran our best neural network model (GRU) against a variety of dataset configurations with time windows ranging from 1 to 16 weeks, and we experimented with different autocorrelation configurations. We experimented our best models against versions of the dataset containing weather related data only, total cases only, and a mix of both. Results were not sensitive to these changes.

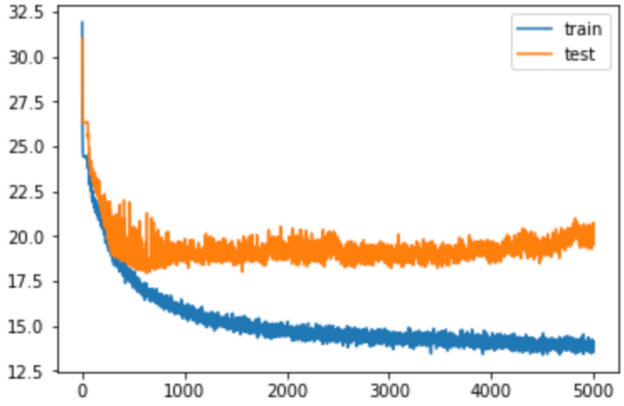
Additionally, we tested our best models on large number of epochs, up to 5000. For large number of epochs results improved on the training set, but overfitting invariably occurred, even with regularization and dropout, as seen below.

```
Final loss train: 4.548939554615224
Final loss valid: 6.773314711025783
```



Iquitos

```
Final loss train: 13.777417920757067
Final loss valid: 20.05314829450477
```



San Juan

Phase IV – Classification

In the last phase of the project we discretized total cases and approached the forecasting problem as a classification problem as opposed to regression. With the time available to complete the class we were not able to fully explore this strategy but initial results were not promising.

Summary of Results

Ref	Notebook and supporting script	Model	Hyperparameter tuning	Features	Datafile (.csv)	Score
Phase I - Benchmarking						
1	N/A	Monthly average	N/A	Total cases only	dengue_means_per_week	26.2524
2	dengueai_regr1 myutil_regr	BayesianRidge regression	None	All numerical features Minmax scaling	submission_20171122_regr	23.9639
3	dengueai_lstm1 myutil_lstm	LSTM	Number of cells, layers, epochs, batch size	All features Minmax scaling	submission_20171122_lstm	32.4832
4	dengueai_dense myutil_dense	Fully connected MLP	Number of cells, layers, epochs, batch size	All numerical features Minmax scaling	submission_20171207_dense	31.3080
5	dengueai_gru1 myutil_gru	GRU	Number of cells, layers, epochs, batch size	All features Minmax scaling	submission_20171123_gru	31.7981
6	dengueai_ada1 myutil_ada	Adaboost with RandomForestRegressor	Regressor=Treeregressor, max depth, number of estimators, learning rate, loss	All numerical features Minmax scaling Total cases	submission_20171123_ada	26.7332
Phase II – Outliers vs. No-outliers. No significant improvements in results						
Phase III – Time Window and Autocorrelation Experiments. No significant improvements in results						
7	dengue_all1 dengue_all2 dengue_all3 myutil_all	GRU	Time windows from 1 to 16.	Weather features only, some weather features, no weather features (total cases only)	Several files	30+ range
8	dengueai_gru1 myutil_gru	GRU	Large number of epochs, regularization and dropout	All features Minmax scaling	submission_20171123_gru	30+ range
Phase IV – Modified problem into classification problem.						
9	dengueai_gru_classification1, 2 and 3.ipynb myutil_gru_class	GRU	Time windows from 1 to 16.	All features Minmax scaling Discretized total cases of dengue	Several files	30+ range

Discussion

We submitted a total of 45 predictions to the DengAI competition[8]. Most of these were predictions obtained with neural network models. Our models included multilayer perceptron implementations (MPL), Long-Short Term Memory (LSTM) recurrent networks and Gated Recurrent Unit (GRU) networks. We experimented with architectures with hidden layers ranging from 1 to 5 and each time with a wide range of node numbers, epochs, learning rates, batch sizes, optimizer options, activation functions, regularization settings, and dropout factors. We experimented different time windows ranging from 1 to 16 weeks; we explored models with auto regression, and we tried a few different pre-processing options, including the removal of outliers. We also ran classification models by discretizing dengue cases into 10 and 50 categories, with little impact to our results, which always scored above the 30 Mean Absolute Error (MAE) mark.

The leading score in the competition at the time of this report is 13.8918. Our best score, 22.8077, was obtained with Bayesian Ridge regression for San Juan predictions, and weekly average for Iquitos. That placed us in 160th place in the competition, out of 2,362 participants, as seen in the figure below. If we had only submitted results obtained with neural network models our ranking would be much worse, probably around 850th place. Therefore, the performance of neural network models for the problem at hand was very disappointing and not competitive.

In his article "Neural networks for time series processing"[14] Georg Dorffner shows that neural networks are theoretically capable of approximating any "reasonable" time series function but that limited data, and a variety of other problems, such as local minima and overfitting will

[14] Dorffner, Georg. "Neural networks for time series processing." Neural network world. 1996.

impair neural networks' ability to produce results comparable to those obtained by linear models [14 p.11-12]. Not surprisingly, in their article "Climate change and dengue: a critical and systematic review of quantitative modeling approaches"[15] Naish, Suchithra, et al. analyze in detail 16 studies that model dengue incidence based on climate data, and that only two of them employs a non-linear model [15 p. 4-6]. Fourteen out of the 16 reported studies used analytical approaches, such as wavelet time series, linear regression, and poisson regression models. One of the two non-linear models studied used a data set spanning 39 years [15 p.8], which is far more data than what was available for our project

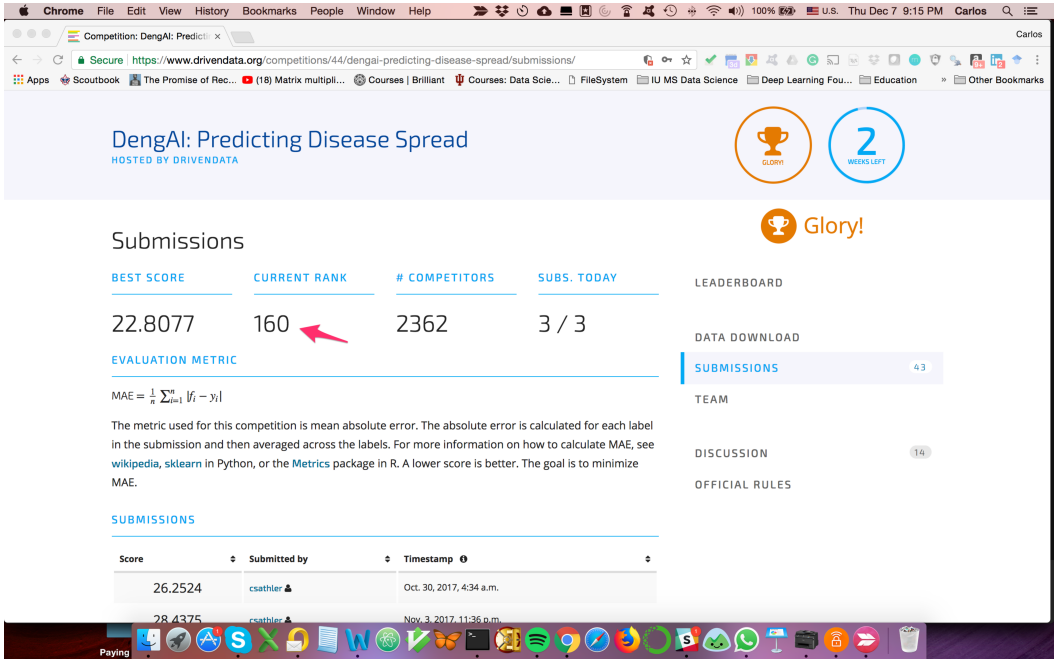


Figure 6 – DengAI competition results and ranking after 45 submissions

Should we have more time for this project, it would make sense to switch to an analytical approach, and experiment with some of the models studied by Naish, Suchithra, et al.[15] Still, it

[15] Naish, Suchithra, et al. "Climate change and dengue: a critical and systematic review of quantitative modelling approaches." BMC infectious diseases 14.1 (2014): 167.

would be well worthy it to evaluate more advanced neural network architectures on the project data. For example, it would be interesting to frame our problem as a sequence-to-sequence (seq2seq) problem, which would allow us to explore the effect of different time windows not only for past periods, but also for predictions beyond a single week.

The figure below shows total cases of dengue per week for the city of San Juan. Notice that the time series shows spikes corresponding to periodic epidemics (e.g. 2005, week 37). On a model that would include total cases of dengue as a feature (autoregressive model), it would be useful to evaluate the performance of models predicting dengue cases for several weeks at a time (i.e. a "sequence" of weeks). Such seq2seq models perhaps will better learn to predict the spikes in the time series and yield overall better results. It is important to remember that dengue spreads when a mosquito bites and infected person, and the more infected people the higher the risk others will be infected. Therefore an autoregressive model is consistent with the disease life cycle.

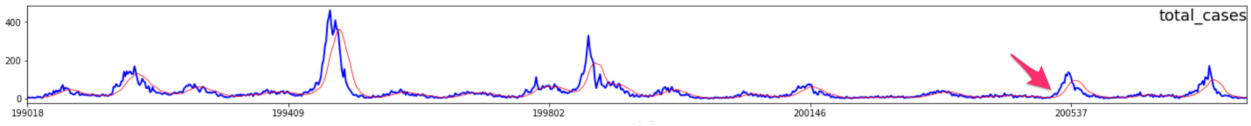


Figure 7 – Number of cases of dengue per week for the city of San Juan

Finally, a full exploration of neural networks wouldn't be complete without experiments with deep learning and convolutional networks. Particularly Undecimated Fully Convolutional Neural Networks (UFCNN) have been reported to outperform other neural networks,[16] such as the ones we explored in our project, for time series problems. It would be interesting to analyze the performance of an UFCNN implementation for dengue prediction.

[16] Gamboa, John Cristian Borges. "Deep Learning for Time-Series Analysis." *arXiv preprint arXiv:1701.01887* (2017).

Conclusion

The CDC estimates one third of the world population is exposed to the dengue virus and is at risk of contracting the disease.[11] Every year 400 million people contract the disease leading to 25,000 deaths yearly.[12] Meanwhile the WHO asserts that "dengue morbidity can be reduced by implementing improved outbreak prediction and detection…".[9] The importance of good predictive models is therefore of paramount significance to contain this disease.

In our project we created predictive models of dengue epidemic in connection with the US Department of Commerce Dengue Forecasting project[7] and the "DengAI: Predicting Disease Spread" competition from DrivenData[8]. We used neural network models to predict dengue cases from climate data and compared our results with more common machine learning models, such as decision tree regression and several multivariate linear regression models. We submitted our results to the DengAI competition and found that more traditional machine learning approaches outperformed the following neural network models: Multi-Layer Perceptron, Long-Short Term Memory and Gated Recurrent Unit networks.

While our results were not competitive, we believe our project offers a positive contribution to a well worthy cause, mainly through judicious testing and documentation of several neural network approaches to dengue prediction. Further work would be needed to determine if more sophisticated neural network architectures (such as seq2seq and UFCNN) are capable of producing better, more competitive results for dengue prediction.

Works Cited

“Dengue.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Jan. 2016, www.cdc.gov/dengue/index.html.

“Dengue.” Dengue, World Mosquito Program, www.eliminatedengue.com/our-research/dengue-fever.

“Dengue and Climate” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 27 Sept. 2012, www.cdc.gov/dengue/entomologyecology/climate.html.

“Dengue and Severe Dengue.” Dengue and Severe Dengue: Fact Sheet, World Health Organization, Apr. 2017, www.who.int/mediacentre/factsheets/fs117/en/.

“Dengue Transmission.” Nature News, Nature Publishing Group, www.nature.com/scitable/topicpage/dengue-transmission-22399758?sa=X&ved=0ahUKEwin843IxKLOAhVDnZQKHSsqDNQQ9QEIEDAA.

Dorffner, Georg. "Neural networks for time series processing." Neural network world. 1996.

DrivenData. DengAI: Predicting Disease Spread, DrivenData, www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/.

Gamboa, John Cristian Borges. "Deep Learning for Time-Series Analysis." arXiv preprint arXiv:1701.01887 (2017).

Global strategy for dengue prevention and control World Health Organization - Geneva: World Health Organization, 2012

Hii, Yien Ling et al. “Forecast of Dengue Incidence Using Temperature and Rainfall.” Ed. Francis Mutuku. PLoS Neglected Tropical Diseases 6.11 (2012): e1908. PMC. Web. 28 Oct. 2017.

Hurtado Díaz, M., et al. "Impact of climate variability on the incidence of dengue in Mexico." Tropical medicine & international health 12.11 (2007): 1327-1337.

Johansson, Michael A., et al. “Local and Global Effects of Climate on Dengue Transmission in Puerto Rico.” PLoS Neglected Tropical Diseases, vol. 3, no. 2, 2009, doi:10.1371/journal.pntd.0000382.

Morin, Cory W., Andrew C. Comrie, and Kacey Ernst. "Climate and dengue transmission: evidence and implications." Environmental health perspectives 121.11-12 (2013): 1264.

Naish, Suchithra, et al. "Climate change and dengue: a critical and systematic review of quantitative modelling approaches." BMC infectious diseases 14.1 (2014): 167.

Nature News, Nature Publishing Group, www.nature.com/scitable/topicpage/dengue-transmission-22399758.

US Department of Commerce, NOAA, National Weather Service. “Dengue Forecasting.” Dengue Forecasting, NOAA's National Weather Service, dengueforecasting.noaa.gov/.