**Logo

Description automatically generated with medium confidence**

**MH6811 Machine Learning in Finance**

|  |  |
| --- | --- |
| **Jonathan Then Tian Meng** | * Holt-Winters * Multiple Linear Regression |
| **Ong Zhen Ying Melissa** | * SARIMA/SARIMAX * CNN |
| **Yong Chai Li** | * Random Forest * LSTM |

Contents

[1 Introduction to problem 1](#_Toc117892459)

[2 Objectives 1](#_Toc117892460)

[3 Description of dataset 1](#_Toc117892461)

[3.1 Number of cases (1 target variable) 1](#_Toc117892462)

[3.2 City and time (4 variables) 2](#_Toc117892463)

[3.3 Climate (20 variables) 2](#_Toc117892464)

[4 Methods 2](#_Toc117892465)

[4.1 Models 2](#_Toc117892466)

[4.2 Train-test split 3](#_Toc117892467)

[4.3 Error function 3](#_Toc117892468)

[5 Data preprocessing 3](#_Toc117892469)

[5.1 Null values 3](#_Toc117892470)

[5.2 Feature selection 3](#_Toc117892471)

[6 Findings 3](#_Toc117892472)

[6.1 Time series analysis: Holt-Winters 3](#_Toc117892473)

[6.2 Time series analysis: Seasonal Autoregressive Integrated Moving Average (SARIMA) / SARIMA with eXogenous factors (SARIMAX) 4](#_Toc117892474)

[6.3 Neural networks: LSTM 4](#_Toc117892475)

[6.4 Neural networks: CNN 5](#_Toc117892476)

[6.5 Machine learning: Multiple Linear Regression 5](#_Toc117892477)

[6.6 Machine learning: Random Forest 6](#_Toc117892478)

[6.7 Comparison between all models 7](#_Toc117892479)

[7 Conclusion 7](#_Toc117892480)

[7.1 Improvements 7](#_Toc117892481)

[Appendix I – Details of the dataset i](#_Toc117892482)

[Appendix II – Variables correlation heatmap iv](#_Toc117892483)

[Appendix III – Grid search results v](#_Toc117892484)

1 Introduction to problem

At the start of the project, Singapore was at the peak of a dengue outbreak. With dengue being endemic to Singapore and the prolific signs of dengue cases plastered around our neighborhoods, our team decided to study the dengue problem.

Unfortunately, data on NEA’s website is limited to current data on their website[[1]](#footnote-2) and archive data is no longer available. In search of a better dataset that may give us more insights, we decided to use the dataset from the DengAI: Predicting Disease Spread competition.[[2]](#footnote-3)

2 Objectives

Our goal is to predict the number of dengue cases weekly accurately for the test data in the dataset. Getting an idea of the number of cases would help in planning the availability of healthcare facilities and skilled labour.

3 Description of dataset

The entire dataset is a time series recording the total number of dengue cases for San Juan, Puerto Rico (“SJ”) and Iquitos, Peru (“IQ”). There are 25 variables in the dataset. The dengue surveillance data is provided by the authorities and universities in the United States of America, in collaboration with the Peruvian government.

3.1 Number of cases (1 target variable)

We assumed that the “total\_cases” variable each week do not include the cases from the previous week, because most people would recover within a week.[[3]](#footnote-4) The number of cases in the training data appears to vary by season, peaking in the later half of the year.

|  |  |
| --- | --- |
| Inserting image... |  |
| Figure 1: San Juan Cases by Year | Figure 2: Iquitos Cases by Year |

For SJ and IQ’s training data, there is a high autocorrelation for the first few lags. However, the high partial autocorrelation at lag 1 and small correlations (less than 0.3) from lag 2 onwards shows that current case numbers are highly correlated with the cases in the last week. With the decomposition of the data, we observe that there is a clear seasonal trend of 1 year as evidenced by its regular repeated shape.

|  |  |
| --- | --- |
|  |  |
| Figure 3: San Juan ACF (Left) and PACF (Right) | |
|  |  |
| Figure 4: Iquitos ACF (Left) and PACF (Right) | |
|  |  |
| Figure 5: San Juan (Left) and Iquitos (Right) Data Decomposition | |

3.2 City and time (4 variables)

The period of data captured for SJ and IQ and the number of entries is tabled below.

|  |  |  |
| --- | --- | --- |
| **City** | **Time period** | **Number of datapoints** |
| San Juan | Week 18 of 1990 to Week 17 of 2008 | 936 |
| Iquitos | Week 26 of 2000 to Week 25 of 2010 | 520 |
| **Total** | | 1456 |

3.3 Climate (20 variables)

The variables comprise of weather data (temperature, rainfall, and humidity) and the vegetation index[[4]](#footnote-5). Further details of the dataset are in **Appendix I**.

4 Methods

4.1 Models

We first selected models that could predict time series data with seasonality, as the number of cases seemed to vary by season. Due to limited choices, we also included neural networks which could be used in normal time series analysis. Lastly, two machine learning models from the lecture were included as benchmarks for performance against the hypothetically more appropriate models.

|  |  |  |
| --- | --- | --- |
| **Most appropriate** | **→** | **Least appropriate (benchmarks)** |
| * Holt-Winters * SARIMA/SARIMAX | * Long Short-Term Memory (LSTM) * Convolutional Neural Network (CNN) | * Multilinear Regression * Random Forest |

4.2 Train-test split

There are no “total\_cases” label for the test dataset provided. Hence, we will split the training dataset provided into the parts tabled below for the listed purposes. We chose to split the data at the higher end (80%) of the standard range to have more training data because the dataset is small. For the same reason, we chose the Time Series Split Cross Validation instead of Blocked Cross Validation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Train-test split** | | | **Purpose** |
| Training data (80%) | | Test data (20%) | Evaluation |
| Pseudo training data (80%) | Pseudo test data (20%) |  | Model training and 5-fold cross validation |

4.3 Error function

We chose the mean squared error as it is one of the standard evaluation metrics for regression models. It also penalises larger errors more heavily, which complements the aim of predicting total number of cases to plan for healthcare facilities. If there was a spike that the model predicted but did not happen, resources would be wasted unnecessarily.

5 Data preprocessing

5.1 Null values

We filled in the null values using the forward fill method, which propagates the last valid observation to the next, as the backward fill method would use future data to fill in the past gaps, which is not appropriate for time series data. As SJ and IQ have tropical climates, the vegetation, temperatures, and precipitation factors would not change much week on week.

5.2 Feature selection

We anticipate that some features would be highly correlated as they measured the same climate variables. For example, “precipitation\_amt\_mm” and “reanalysis\_sat\_precip\_amt\_mm” both measure total precipitation. Refer to **Appendix II** for the correlation heatmap between the variables.

Using a grid search for correlations ranging from 70% to 95% in increments of 5%, we identified six variables with at least 90% correlation to be dropped. Refer to the **Appendix II** for the six variables and full results of the grid search.

We dropped all variables from both cities’ data that showed a 90% correlation, because even if the variables were from two cities, they were from the same source and measured the same climate variables. Therefore, correlation between the climate variables should be the approximately the same, regardless that they were measured at different locations.

We did not proceed to the next decrease (i.e., 85%), as there would be no more information for temperature in the dataset. Information on the average, minimum and maximum temperature (i.e., station\_avg\_temp\_c, station\_max\_temp\_c, station\_min\_temp\_c) would be removed. Removing these variables would cause our analysis to ignore information showing that temperature affects mosquito activity.[[5]](#footnote-6)

6 Findings

6.1 Time series analysis: Holt-Winters

Holt-Winters is a model used for time series forecasting. It is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). Holt-Winters uses exponential smoothing to encode lots of values from the past and uses them to predict “typical” values for the present and future. We noticed that our data was a time series, and it had a seasonality aspect to it. This led us to choose Holt Winters as one of our clear candidates for our modelling.

We fitted the model onto our training data and used it to make a prediction for the test data. Since the data was measured on a weekly basis, our seasonal period is 52 weeks. As observed in results from the San Juan dataset in Figure 6, we can see that the model predicted a smooth cyclical pattern. The prediction for the Iquitos dataset was much better, the model was able to predict the sharp increase and decrease. However, we did notice that for both models, at certain points, the predictions went below zero.

We conducted the time series cross validation with 5 splits to get a better understanding of the out of sample performance. The mean squared error for the San Juan and Iquitos datasets are 222.652 and 53.843 respectively.

|  |  |
| --- | --- |
|  |  |
| Figure 6: San Juan (Left) and Iquitos (Right) Holt-Winters Predictions | |

6.2 Time series analysis: Seasonal Autoregressive Integrated Moving Average (SARIMA) / SARIMA with eXogenous factors (SARIMAX)

For implementation of SARIMA, we first used the Augmented Dickey Fuller Test (ADF test) to check for stationarity. We note that the training data for both cities were stationary and hence did not perform any differencing.

We used a gird search minimize the Akaike Information Criterion (AIC) and the best fit models for both cities did not have any seasonal orders. This is despite the data decomposition showing a seasonal effect. The best model for SJ was: (2,1,1)(0,0,0)[52] and the best model for IQ was (3,1,0)(0,0,0)[52]. These resulted in predictions that were fairly constant and almost a flat line as shown in the figures below.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated | Chart, line chart, histogram  Description automatically generated |
| Figure 7: San Juan (Left) and Iquitos (Right) SARIMA Predictions | |

Next, we used the SARIMAX model which considers the exogenous variables. The highly correlated variables as described in the Section 5.2 feature selection were dropped. The training data for exogenous variables were not stationary and hence the data was differenced once. The data was also scaled. We then proceeded to perform the grid search to minimise AIC. The best model for SJ was: (2,0,1)(0,0,0)[52] and the best model for IQ was (3,0,0)(0,0,0)[52]. The predicted results are shown below. Again, with no seasonal orders, the resulting predictions were almost flat lines.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| Figure 8: San Juan (Left) and Iquitos (Right) SARIMAX Predictions | |

6.3 Neural networks: LSTM

We normalized the input variables because natural phenomenon tends to follow a normal distribution. The activation functions of the hidden layers are ReLu because only positive results are wanted. We used the exponential output activation function because the output value should be positive with no known upper bound. The best prediction using 5-fold cross validation for the LSTM model after a grid search is shown in the figure below.

|  |  |
| --- | --- |
|  |  |
| Figure 9: San Juan (Left) and Iquitos (Right) LSTM Predictions | |

The number of time steps considered is capped at two weeks, which is the approximate time taken for a mosquito to become infectious starting from when the mosquito hatches[[6]](#footnote-7),[[7]](#footnote-8). Weather conditions before the hatching would not be relevant. The number of layers was capped at two because studies showed two hidden layers for full generality or one hidden layer with an arbitrarily large number of units is sufficient for the "universal approximation" property[[8]](#footnote-9). The number of neurons was slightly less than a reasonably large amount of 100 because of concerns about overtraining, given the small dataset. The best parameters and grid search results are in **Appendix III**.

6.4 Neural networks: CNN

For CNN, a Conv1D layer was used as time series data has a strong one-dimension locality (time itself)[[9]](#footnote-10). The time step window was set at 2 weeks according to the lifecycle of the mosquitoes and predictions were also outputs of 2 weeks. No pooling layer was added as according to research, pooling layers tend to introduce negative effects for seasonal time series data with trends.[[10]](#footnote-11) The predictions are shown in the figures below. The predictions were erratic and far off from the actual observations, possibly due to the size of the convolution window and lack of training data.

|  |  |
| --- | --- |
|  |  |
| Figure 10: San Juan (Left) and Iquitos (Right) CNN Predictions | |

6.5 Machine learning: Multiple Linear Regression

Multiple linear regression was conducted over the full model using the datasets of SJ and IQ. From Figure 11, we can see that the R-squared values produced by the full models are very low. This indicates that the independent variable, total cases of dengue, is not being explained much by the 14 predictors.

|  |  |
| --- | --- |
|  |  |
| Figure 11: R-squared Values for San Juan (Left) and Iquitos (Right) | |

To determine if there is a better model using a subset of the variables, we decided to perform variable selection using the backward stepwise method with AIC as the selection method, as starting with the full model has the advantage of considering the effects of all variables simultaneously.

|  |  |
| --- | --- |
|  |  |
| Figure 12: Backward Stepwise Selection for San Juan (Left) and Iquitos (Right) | |

The backward stepwise selection on SJ produced a subset model that contained 6 predictors, ndvi\_se, ndvi\_sw, reanalysis\_air\_temp\_k, reanalysis\_tdtr\_k, station\_avg\_temp\_c and station\_max\_temp c. The backward stepwise selection on IQ produced a subset model that contained 5 predictors, ndvi\_se, ndvi\_sw, reanalysis\_air\_temp\_k, reanalysis\_tdtr\_k and station\_avg\_temp\_c. By comparison, the subset model does not do better than the full model.

To estimate the skill of the linear regression models on unseen data, we performed 5-fold cross validation on all four models using the *caret* package. The package provides us with the RMSE, and we can easily calculate the MSE by squaring the smaller RMSE of the two models. The MSE are 2357.44 and 113.97. With that being said, the multilinear regression might not be a good model for this time series data.

|  |  |
| --- | --- |
|  |  |
| Figure 13: Results of the Multilinear Regression Models for San Juan (Left) and Iquitos (Right) | |

6.6 Machine learning: Random Forest

We used the sliding window method to take historical data into consideration. The best prediction using 5-fold cross validation for the random forest model after a grid search is shown in the figure below.

|  |  |
| --- | --- |
|  |  |
| Figure 14: San Juan (Left) and Iquitos (Right) Random Forest Predictions | |

The upper bound of the season parameter was decided as one year, because the decomposition graph showed a potential repeat for each year. The number of time steps considered is capped at two weeks, which is the approximate time taken for a mosquito to become infectious starting from when the mosquito hatches[[11]](#footnote-12),[[12]](#footnote-13). Weather conditions before the hatching would not be relevant. The maximum depth of the tree and minimum leave sample was based on a single tree built with the first training set as a gauge for the maximum of the range. The grid search results are in **Appendix III**.

6.7 Comparison between all models

In contrast to our expectations, the random forest model outperformed the other models. Possible reasons are that it was able to deal with non-linear relationships, for example the spikes in infections in certain years​. These spikes may be caused by weather independent variables, such as new variants of the disease where the population has still yet to develop resistance to. The random forest was also good for this high dimension problems where variables from multiple time periods were concatenated together because it uses only a subset of features each time, resolving issues with leftover highly correlated variables (as high as >80%)​. Thirdly, it avoids potential scaling issues, which is a pre-requisite of other models like the neural networks.

Nevertheless, we note a disadvantage where we were unable to use the model to understand the most important variables contributing to dengue cases.

| **City** | **Model** | **Mean Squared Error (MSE)** | **Model** | **Mean Squared Error (MSE)** |
| --- | --- | --- | --- | --- |
| San Juan | Random Forest | 125.58 | SARIMAX | 1367.53 |
| Holt Winters | 228.29 | Multi linear regression | 2357.45 |
| LSTM | 1002.310 | CNN | 335689.11 |
| SARIMA | 1195.94 |  |  |
| Iquitos | Random Forest | 43.86 | SARIMAX | 150.69 |
| Holt Winters | 54.110 | SARIMA | 209.53 |
| Multi linear regression | 113.9661 | CNN | 16509.94 |
| LSTM | 139.249 |  |  |

7 Conclusion

Contrary to our expectations, the random forest performed the best in predicting the total cases of dengue, out of all the other models with the lowest mean squared error of 125.58 and 43.86 for San Juan and Iquitos respectively.

7.1 Improvements

Due to time constraint, our exploration of the models was limited.

Generally, we could have done grid search on a larger range of values at a smaller denomination.

For the time series models, there were certain points where the models predicted negative values which should not be the case. If given more time, it would be imperative to try converting the data labels with 0 total cases to 0.01 so that the dataset will be strictly positive. With a strictly positive dataset, usage of Box-Cox or logarithms would be viable in preventing negative value predictions from occurring. Additionally, as an extension to Holt-Winters, we can try out the Brutlag algorithm for anomaly detection.

For neural networks, we would have liked to do a grid search on more model parameters such as batch size and epochs and try out autoregressive neural networks which use the predictions to feed back into the model as input.

For machine learning models, we would have liked to try out other models given that the random forest performed well. For example, doing another type of tree algorithm to try to diagnose the main variables that cause dengue cases would provide more value for policy makers.

Appendix I – Details of the dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| <class 'pandas.core.frame.DataFrame'> | |  |  |  |
| RangeIndex: 1456 entries, 0 to 1455 | |  |  |  |
| Data columns (total 24 columns): | |  |  |  |
| # | Column | Non-Null Count | Dtype |  |
| --- | ------ | -------------- | ----- |  |
| 0 | city | 1456 non-null | object | City abbreviations: sj for San Juan and iq for Iquitos |
| 1 | year | 1456 non-null | int64 |  |
| 2 | weekofyear | 1456 non-null | int64 |  |
| 3 | week\_start\_date | 1456 non-null | object | Date given in yyyy-mm-dd format |
| 4 | ndvi\_ne | 1262 non-null | float64 | Pixel northeast of city centroid |
| 5 | ndvi\_nw | 1404 non-null | float64 | Pixel northwest of city centroid |
| 6 | ndvi\_se | 1434 non-null | float64 | Pixel southeast of city centroid |
| 7 | ndvi\_sw | 1434 non-null | float64 | Pixel southwest of city centroid |
| 8 | precipitation\_amt\_mm | 1443 non-null | float64 | Total precipitation (PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)) |
| 9 | reanalysis\_air\_temp\_k | 1446 non-null | float64 | Mean air temperature |
| 10 | reanalysis\_avg\_temp\_k | 1446 non-null | float64 | Average air temperature |
| 11 | reanalysis\_dew\_point\_temp\_k | 1446 non-null | float64 | Mean dew point temperature |
| 12 | reanalysis\_max\_air\_temp\_k | 1446 non-null | float64 | Maximum air temperature |
| 13 | reanalysis\_min\_air\_temp\_k | 1446 non-null | float64 | Minimum air temperature |
| 14 | reanalysis\_precip\_amt\_kg\_per\_m2 | 1446 non-null | float64 | Total precipitation (NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)) |
| 15 | reanalysis\_relative\_humidity\_percent | 1446 non-null | float64 | Mean relative humidity |
| 16 | reanalysis\_sat\_precip\_amt\_mm | 1443 non-null | float64 | Total precipitation (NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)) |
| 17 | reanalysis\_specific\_humidity\_g\_per\_kg | 1446 non-null | float64 | Mean specific humidity |
| 18 | reanalysis\_tdtr\_k | 1446 non-null | float64 | Diurnal temperature range |
| 19 | station\_avg\_temp\_c | 1413 non-null | float64 | Average temperature |
| 20 | station\_diur\_temp\_rng\_c | 1413 non-null | float64 | Diurnal temperature range |
| 21 | station\_max\_temp\_c | 1436 non-null | float64 | Maximum temperature |
| 22 | station\_min\_temp\_c | 1442 non-null | float64 | Minimum temperature |
| 23 | station\_precip\_mm | 1434 non-null | float64 | Total precipitation |
| dtypes: float64(20), int64(2), object(2) | |  |  |  |
| memory usage: 273.1+ KB | |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | year | weekofyear | ndvi\_ne | ndvi\_nw | ndvi\_se | ndvi\_sw | precipitation\_amt\_mm | reanalysis\_air\_temp\_k | reanalysis\_avg\_temp\_k | reanalysis\_dew\_point\_temp\_k | reanalysis\_max\_air\_temp\_k | reanalysis\_min\_air\_temp\_k | reanalysis\_precip\_amt\_kg\_per\_m2 | reanalysis\_relative\_humidity\_percent | reanalysis\_sat\_precip\_amt\_mm | reanalysis\_specific\_humidity\_g\_per\_kg | reanalysis\_tdtr\_k | station\_avg\_temp\_c | station\_diur\_temp\_rng\_c | station\_max\_temp\_c | station\_min\_temp\_c | station\_precip\_mm |
| count | 1456 | 1456 | 1262 | 1404 | 1434 | 1434 | 1443 | 1446 | 1446 | 1446 | 1446 | 1446 | 1446 | 1446 | 1443 | 1446 | 1446 | 1413 | 1413 | 1436 | 1442 | 1434 |
| mean | 2001.032 | 26.50343 | 0.142294 | 0.130553 | 0.203783 | 0.202305 | 45.76039 | 298.7019 | 299.2256 | 295.2464 | 303.4271 | 295.7192 | 40.15182 | 82.16196 | 45.76039 | 16.74643 | 4.903754 | 27.18578 | 8.059328 | 32.45244 | 22.10215 | 39.32636 |
| std | 5.408314 | 15.01944 | 0.140531 | 0.119999 | 0.07386 | 0.083903 | 43.71554 | 1.36242 | 1.261715 | 1.52781 | 3.234601 | 2.565364 | 43.4344 | 7.153897 | 43.71554 | 1.542494 | 3.546445 | 1.292347 | 2.128568 | 1.959318 | 1.574066 | 47.45531 |
| min | 1990 | 1 | -0.40625 | -0.4561 | -0.01553 | -0.06346 | 0 | 294.6357 | 294.8929 | 289.6429 | 297.8 | 286.9 | 0 | 57.78714 | 0 | 11.71571 | 1.357143 | 21.4 | 4.528571 | 26.7 | 14.7 | 0 |
| 25% | 1997 | 13.75 | 0.04495 | 0.049217 | 0.155087 | 0.144209 | 9.8 | 297.6589 | 298.2571 | 294.1189 | 301 | 293.9 | 13.055 | 77.17714 | 9.8 | 15.55714 | 2.328571 | 26.3 | 6.514286 | 31.1 | 21.1 | 8.7 |
| 50% | 2002 | 26.5 | 0.128817 | 0.121429 | 0.19605 | 0.18945 | 38.34 | 298.6464 | 299.2893 | 295.6407 | 302.4 | 296.2 | 27.245 | 80.30143 | 38.34 | 17.08714 | 2.857143 | 27.41429 | 7.3 | 32.8 | 22.2 | 23.85 |
| 75% | 2005 | 39.25 | 0.248483 | 0.2166 | 0.248846 | 0.246982 | 70.235 | 299.8336 | 300.2071 | 296.46 | 305.5 | 297.9 | 52.2 | 86.35786 | 70.235 | 17.97821 | 7.625 | 28.15714 | 9.566667 | 33.9 | 23.3 | 53.9 |
| max | 2010 | 53 | 0.508357 | 0.454429 | 0.538314 | 0.546017 | 390.6 | 302.2 | 302.9286 | 298.45 | 314 | 299.9 | 570.5 | 98.61 | 390.6 | 20.46143 | 16.02857 | 30.8 | 15.8 | 42.2 | 25.6 | 543.3 |

**Comparison of weather conditions in Singapore with San Juan and Iquitos**

|  |  |
| --- | --- |
|  |  |

Appendix II – Variables correlation heatmap

|  |  |
| --- | --- |
| A picture containing chart  Description automatically generated | Chart  Description automatically generated with medium confidence |
| San Juan correlations of variables | Iquitos correlation of variables |

| **Dropped variables** | **Correlated variable remaining (higher of correlation between San Juan and Iquitos)** |
| --- | --- |
| reanalysis\_avg\_temp\_k | reanalysis\_air\_temp\_k (0.997) |
| reanalysis\_dew\_point\_temp\_k | reanalysis\_air\_temp\_k (0.905) |
| reanalysis\_max\_air\_temp\_k | reanalysis\_air\_temp\_k (0.930) |
| reanalysis\_min\_air\_temp\_k | reanalysis\_air\_temp\_k (0.939) |
| reanalysis\_sat\_precip\_amt\_mm | precipitation\_amt\_mm (1) |
| reanalysis\_specific\_humidity\_g\_per\_kg | reanalysis\_air\_temp\_k (0.906) |

**Grid search to drop correlated variables**



Appendix III – Grid search results

**Grid search results for random forest**

Best model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **City​​** | **Season (52-week time step)** | **No. of time steps** | **Max depth​** | **Min leaf sample​** | **Mean squared error​​** |
| San Juan​​ | 0 | 2​ | 15​ | 10​ | 123.326 |
| Iquitos​​ | 0 | 1​ | 5​ | 4​ | 42.413​ |
| Grid search range​ | [0,1] | [1,2]​ | [5,20]​ | [2,10]​ | ​ |

Full grid search results



**Grid search results for LSTM**

Best model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **City​​** | **No. of time steps considered​** | **No. of layers​** | **No. of neurons in 1st layer​** | **No. of neurons in 2nd layer, if any​** | **MSE​** |
| San Juan​​ | 2​ | 1​ | 25​ | NA​ | 1002.310​ |
| Iquitos​​ | 2​ | 2​ | 50​ | 50​ | 139.249 |
| Grid search range​ | [1,2]​ | [1,2]​ | [25,75]​ | [25,75]​ | ​ |

Full grid search results



1. Dengue Clusters. (n.d.). Retrieved September 2, 2022, from https://www.nea.gov.sg/dengue-zika/dengue/dengue-clusters [↑](#footnote-ref-2)
2. DengAI: Predicting Disease Spread. (n.d.). DrivenData. Retrieved September 2, 2022, from https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/ [↑](#footnote-ref-3)
3. Dengue Symptoms and Treatment| CDC. (2021, September 20). Centers for Disease Control and Prevention. https://www.cdc.gov/dengue/symptoms/index.html [↑](#footnote-ref-4)
4. A vegetation index is a single number that quantifies vegetation biomass and/or plant vigor for each pixel in a remote sensing image. Extracted from: What does vegetation index mean in remote sensing technology? – Geospatial Technology. (2019, August 21). https://mapasyst.extension.org/what-does-vegetation-index-mean-in-remote-sensing-technology/ [↑](#footnote-ref-5)
5. How does Weather Affect Mosquito Activity? (2016, July 5). Preventive Pest Control. https://www.preventivepestcontrol.com/weather-affect-mosquito-activity/ [↑](#footnote-ref-6)
6. Aedes Mosquito. (n.d.). Retrieved October 17, 2022, from https://www.nea.gov.sg/dengue-zika/prevent-aedes-mosquito-breeding/aedes-mosquito [↑](#footnote-ref-7)
7. Mosquito-borne Transmission. (n.d.). Retrieved October 17, 2022, from https://www.cdc.gov/dengue/training/cme/ccm/page45915.html [↑](#footnote-ref-8)
8. comp.ai.neural-nets FAQ. Retrieved October 20, 2022, from http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-9.html [↑](#footnote-ref-9)
9. Verma, S. (2019, September 20). Understanding 1D and 3D Convolution Neural Network | Keras. https://towardsdatascience.com/understanding-1d-and-3d-convolution-neural-network-keras-9d8f76e29610 [↑](#footnote-ref-10)
10. Nonpooling Convolutional Neural Network Forecasting for Seasonal Time Series With Trends. (2020, August 1). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/8823990 [↑](#footnote-ref-11)
11. Aedes Mosquito. (n.d.). Retrieved October 17, 2022, from https://www.nea.gov.sg/dengue-zika/prevent-aedes-mosquito-breeding/aedes-mosquito [↑](#footnote-ref-12)
12. Mosquito-borne Transmission. (n.d.). Retrieved October 17, 2022, from https://www.cdc.gov/dengue/training/cme/ccm/page45915.html [↑](#footnote-ref-13)