DrivenData Username: Kimberly Kaminsky
DrivenData Display Name: KimKam

Kimberly Kaminsky
Predict 413 Section 55
Final – DengAI competition

DrivenData DengAI Competition

Problem Definition and Significance

DrivenData hosted a competition on its website, DengAI: Preventing Disease Spread. The purpose of the competition is to predict local epidemics of Dengue Fever. Dengue fever is a mosquito-borne disease that, in severe cases, can cause severe bleeding, low blood pressure, and even death. It occurs in tropical and sub-tropical parts of the world as weather patterns in these parts of the world are conducive to the spread of mosquitoes. It is believed that the spread of Dengue is related to the climate variables such as temperature and precipitation. For this competition, environment data was collected from several governmental agencies including the Centers for Disease Control and Prevention, the National Oceanic and Atmospheric Administration, and the US Department of Commerce. An understanding of the relationship between environmental variables and the spread of Dengue will facilitate the ability to allocate resources for further research initiatives which can help prevent further outbreaks or contribute to the education of affected populations.

Data Exploration

There were three datasets provided for this competition: dengue_features_train, dengue_labels_train, and dengue_features_test. The dengue_labels_train dataset contains actual data about the number of cases of dengue fever in two cities: San Juan, Puerto Rico and Iquitos, Peru. The number of cases are recorded by year and the week of the year. There were 936 records for San Juan and 520 records for Iquitos for a total of 1436 records. The dengue_features_train dataset provides environment information that have been gathered by several government agencies. Again, this is recorded by year and week of the year. There are an additional 23 predictor variables provided, in addition to the year and week of the year. Table 1 contains a list of these variables. The test dataset provides the predictor variables for 260 weeks of data for San Juan and 156 weeks of data for Iquitos. In order to train the data, the dengue_features_train and dengue_labels_train datasets were merged on the year and week of year columns.

An initial inspection of the data indicated that the train dataset is missing at least some of the values for all of the predictor variables, except year, weekofyear, city and week_start_date. The variable ndvi_ne was missing the largest number of variables, 194 out of 1436. This only represents 13% of the data and so none of the variables were considered to be missing too many values to be excluded from the dataset. Only week_start_date was excluded as it did not provide any additional information considering that year and weekofyear were provided and these provided all the information necessary to conduct time series analysis.

An initial analysis of the correlations between the predictor variables and the response variable were conducted at this point. Figure 1, below, displays the correlation matrix for all numeric variables. All rows with missing variables were excluded for this analysis. None of the variables showed particularly strong correlations. The three variables with the strongest correlations with the response variable, total_cases, were reanalysis_min_air_temp_k, reanalysis_tdtr_k, and reanalysis_air_temp_k.

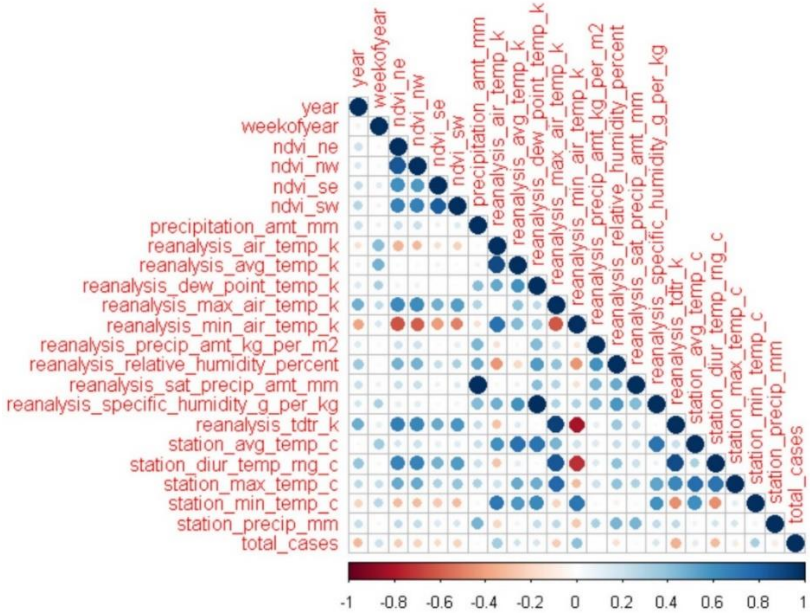

Figure 1:Correlation Matrix for predictor and response variables

At this point, the training and test data was split into two datasets: one for San Juan and Iquitos. Since this data has uninterrupted data for each week of the year for their respective time periods, a time series graph was constructed for each of the cities. Both of the graphs show seasonality though the trend is clearer for San Juan. Both cities show a good deal of volatility in the data though in this case there is more volatility in the Iquitos data.

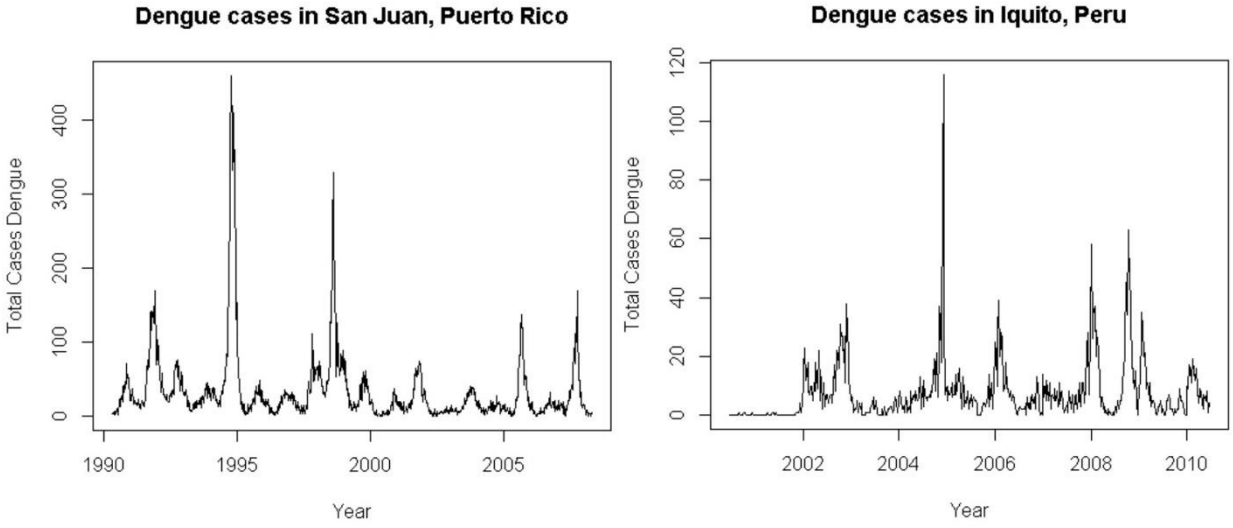

Figure 2: Time Series Representation of Total Dengue Cases in Each City

Since there appears to be a strong time series component to the data, it seems reasonable that time series analysis could be used to impute some of the missing values in the predictor variables. In order to test this theory, time series graphs were created for each of the variables and some of the variables seemed as though they could be imputed utilizing the mean value for the value preceding and following the missing value. Other variables had too much volatility in the data for this to be a reasonable method. Figure 3 shows a variable, Air Temperature, that was imputed using the time series method versus a variable, relative humidity, that was unable to use the time series method. The remaining variables, that were not imputed using the time series method, were imputed using the MICE package in R.

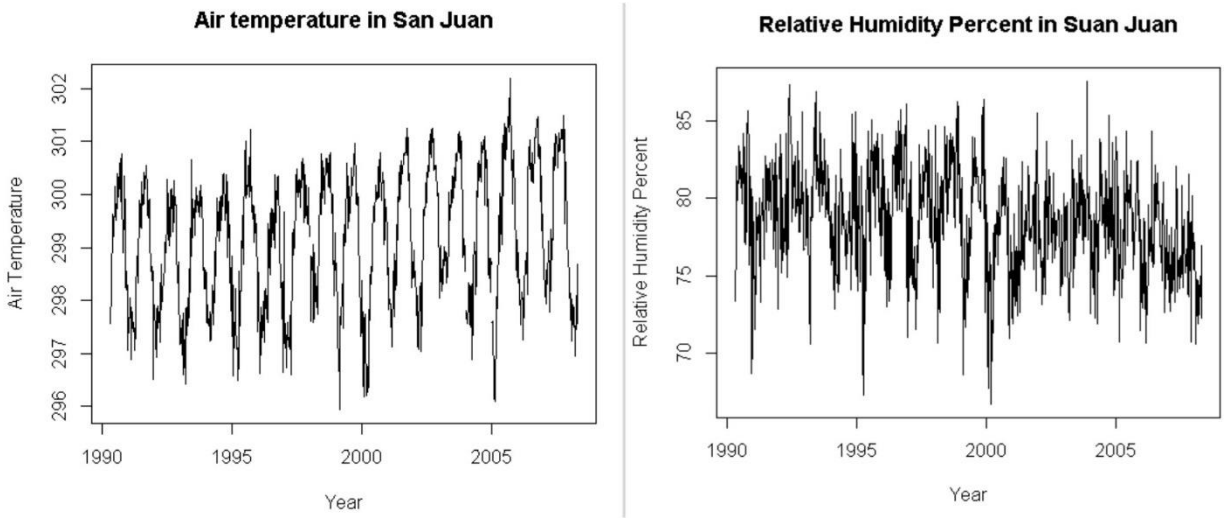

Figure 3: Variable Imputed using Time Series vs. Variable That Cannot Use Time Series

The variables were also analyzed to determine normality and to find outliers. Variables were either transformed or capped, or both. One example is the total_cases variable for Iquitos. Figure 4 shows the histograms of the data before and after the data transformation.

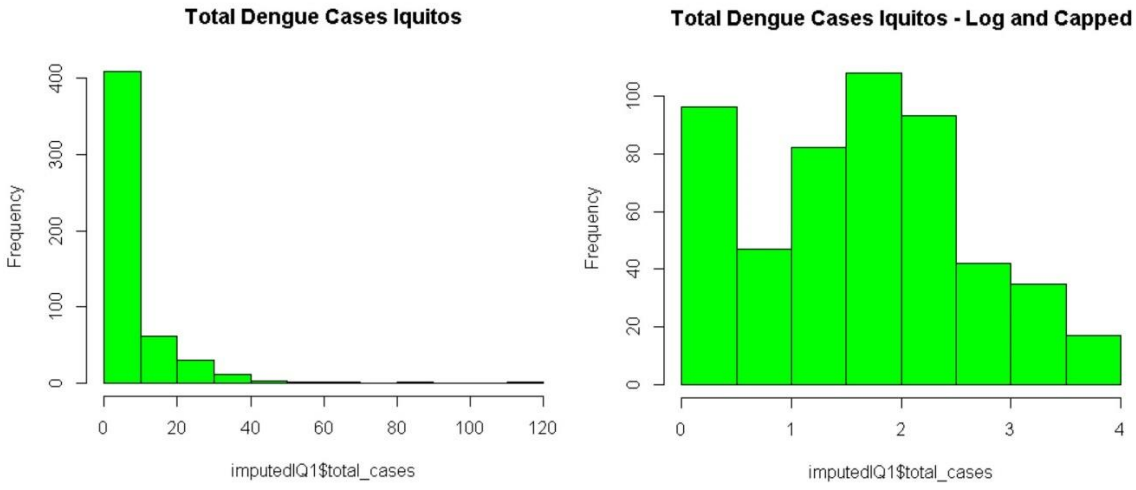

Figure 4: Transformed Total Cases for Iquitos

Model Selection\Implementation

Not surprisingly, since this data had a clear time series element to it, most of the current research in predicting epidemics of mosquito transmitted diseases involved time series algorithms. The value of additive vs. multiplicative models, depending on topology, was used to predict malaria epidemics (Githeko et al 2014). Seasonal Arima models were used to successfully predict dengue incidents in Brazil (Martinez et al. 2011), Mexico (Johansson et al. 2016), and in the Phillipines (Dela Cruz et al. 2012). Linear regression has also been used, but in conjunction with lag variables which essentially makes it a time series (Ramadona et al. 2016). As the preponderance of research regarding prediction of dengue fever has utilized seasonal Arima models, a decomposed Auto Arima was run. As mentioned earlier, the data exhibited a great deal of volatility so a GARCH model was also implemented. Although, GARCH has not been used in previous research to predict mosquito transmitted diseases, it has been used successfully to predict rainfall (Yusof et al. 2012). Neural Networks have also been utilized, although without the time component (Aburas et al. 2010). Since the time series methods have been very successful, neural networks incorporating the time element were tested.

For each method, two models were built: one which incorporated external climatic regressors and one that did not. This was done because there were conflicting findings within the research. Johansson found that climate data did not significantly affect the predictive power of the models (Johansson et al. 2016). Yet, Dela Cruz found that inclusion of climatic data enhanced the model (Dela Cruz et al. 2012). The GARCH and Neural Network models both used seasonally decomposed data. Each model was run separately for San Juan and Iquitos as their time series graphs for the total cases looked quite different. All work was done utilizing the R programming language. The San Juan and Iquitos training sets were broken out into train/test datasets based upon an 80/20 split of the data.

Model Performance

The DengAI competition evaluated the results based on Mean Absolute Error, MAE, between the total predicted cases of dengue fever for the city against the actual total cases of dengue fever for the city. Table 1 compares the MAEs between the best model produced using each of the different modeling methods for San Juan, Iquitos and the total MAE once it was submitted to the DengAI competition. Perhaps what is most interesting in the results is that it is not clear cut whether adding in the climate data helped the models or not. The test model that performed the best for the San Juan data was the neural network that incorporated climate data. The test model that performed the best for Iquitos was the neural network that did not incorporate climate data. Opposite to what might be expected, the GARCH model that did not incorporate climate data performed better for San Juan and the GARCH model that did incorporate climate data performed better for Iquitos. Although, the GARCH model, (without climate data), did not perform best on the test data for either San Juan or Iquitos it was clearly the winner once it was submitted to the DengAI competition.

Modeling Method	San Juan	Iquitos	DengAI
Seasonal AA w/ Climate Data**	13.4616	34.9956	29.7236
Seasonal AA**	13.0084	25.796	30.8389
GARCH	13.478	7.7509	26.6154
GARCH w/ Climate Data	14.6438	7.5736	28.3702
Neural Net w/ Climate Data	12.048	7.0255	30.8389
Neural Net	13.0013	6.6414	29.5769

Table 1: Model Comparison of Mean Absolute Error against San Juan & Iquitos test sets and Final DengAI result
***** Note that the model submitted to DengAI was the Seasonal AA for San Juan and the AA without Seasonality for Iquitos***

Model Limitations

Each of the models above have their limitations. The seasonal auto Arima without climate features has a very simple model with 5 coefficients, 2 autoregressive variables, 2 moving average variables and 1 seasonal moving average variable. The Arima model with climate data improved the score a bit but included an additional 29 variables, increasing its complexity. However, the model does not perform as well as the other models. The neural networks performed very well on the data but were among the worst performers against the actual data. Also, the model is a black box and difficult to explain. The GARCH model, which performed best is explainable but contains 44 variables due it being an ARMA(2,40) + GARCH(1,1) model.

Future Work

As mentioned earlier it was strange that in some cases the climate data was helpful to the model but in other cases it was not helpful. Other research has found that utilizing lags of up to two months on rain variables helps predictions of dengue fever epidemics (Ramadona et al. 2016). Adding lags on the rain may help as it probably takes some ramp up time before the mosquitos are born and have time to transmit the disease after a heavy rainfall. Also, variable selection techniques were not employed since there were so few variables. The correlation matrix shown above indicates that there may be collinearity between some of the variables. Employing variable selection technique may help improve the results. Finally, time series models were used exclusively. Other types of models may perform better or could be combined with time series techniques to produce better results.

Conclusions

The object of this competition was to build a model that minimized the mean absolute error in predicting dengue fever cases. Some basic data cleanup which included imputing missing values and trying out a variety of time series models yielded a decent result, which placed 616 out of 2269 in the DengAI competition. The model that provided the best results on the test set submitted to DengAI was the GARCH model without climatic data included. Overall, all the models did a good job with an MAE ranging from 26.6154 to 34.8056. This shows that there wasn't a great deal of difference between the models. All the models can likely be improved

either through feature engineering or parameter tuning. In order to make a definitive conclusion about which type of model performs best on this type of problem, a great deal more rigor and experimentation would be needed.

References

Ramadona AL., Lazuardi L., Hii YL., Holmner Å., Kusnanto H., Rocklöv J. (2016) Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data. *PLoS ONE* 11(3):

Dela Cruz AC, Lubrica JA, Punzalan BV, & Martin MC. (2012) Forecasting Dengue Incidence in the National Capital Region, Philippines: Using Time Series Analysis with Climate Variables as Predictors. *Acta Manilana*; 60:19–26

Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M. (2016) Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Scientific Reports* 6, 33707

Yusof F., Kane I. (2013) Volatility modeling of rainfall time series *Theoret. Appl. Climatol.*, 113 (1–2) (2013), pp. 247-258

Martinez EZ, Silva EA. (2011) Predicting the number of cases of dengue infection in Ribeirao Preto, Sao Paulo State, Brazil, using a SARIMA model. *Cad Saude Publica*., 27: 1809-1818.

Githeko AK, Ogallo L, Lemnge M, Okia M, Ototo EN. (2014) Development and validation of climate and ecosystem-based early malaria epidemic prediction models in East Africa. *Malar J.* 13:329.

Appendix A – DengAI Variables

Variables Names	Variables Description
year	Year
weekofyear	Week of the Year
city	City abbreviations: sj for San Juan and iq for Iquitos
week_start_date	Date given in yyyy-mm-dd format
station_max_temp_c	Maximum temperature
station_min_temp_c	Minimum temperature
station_avg_temp_c	Average temperature
station_precip_mm	Total precipitation
station_diur_temp_rng_c	Diurnal temperature range
precipitation_amt_mm	Total precipitation - PERSIANN
reanalysis_sat_precip_amt_mm	Total precipitation - NOAA's NCEP
reanalysis_dew_point_temp_k	Mean dew point temperature
reanalysis_air_temp_k	Mean air temperature
reanalysis_relative_humidity_percent	Mean relative humidity
reanalysis_specific_humidity_g_per_kg	Mean specific humidity
reanalysis_precip_amt_kg_per_m2	Total precipitation
reanalysis_max_air_temp_k	Maximum air temperature
reanalysis_min_air_temp_k	Minimum air temperature
reanalysis_avg_temp_k	Average air temperature
reanalysis_tdtr_k	Diurnal temperature range
ndvi_se	Pixel southeast of city centroid
ndvi_sw	Pixel southwest of city centroid
ndvi_ne	Pixel northeast of city centroid
ndvi_nw	Pixel northwest of city centroid
total_cases	Total weekly cases of dengue