# DengAI Model Results:

The goal of this competition as well as the one hosted by the joint efforts of different US Departments was to use climate data to be able to predict outbreaks of dengue. The project looked at different modeling approaches to try and use these climate features to predict the target variable total\_cases, more models than shown in the notebooks. The scoring metrics used to grade the performance of models submitted to these competitions is the Mean Absolute Error (MAE)

However, I won’t be participating in this competition, and only plan on using the dataset provided to try and predict the case numbers of dengue modeled by climate data for my learning. And as such, I did not consider other aspects of the evaluation the competition set forth such as the probability distribution of forecasts and peak incident probabilities of the predictions. I took an approach of finding a model that would best predict the case numbers that were provided without considering the holdout test set that wasn’t made public which would eventually be used to grade the performance of the models submitted by participants. Since I was limiting the scope of this project to just the years of data that were provided as the training sets to build the models on, I decided to use additional scoring metrics to evaluate my models that researchers studying the same theory[i-iv] found to be useful, such as the Root Mean Squared Error (RMSE), symmetric Mean Absolute Percentage Error (sMAPE), and R2, the coefficient of determination. The MAE and RMSE error scores are easier to digest as they provide the error scores in the same units as the test values, number of cases.

RMSE is the standard deviation of the residuals, the measure of how spread out the prediction errors are. An alternative to the RMSE was also used, though the scores are not presented, known as the Root Mean Squared Log Error. This metric wasn’t discussed in any of the research papers I found but does present some interesting characteristics relating to our task at hand. The RMSLE is similar to the RMSE but takes the log of the predictions and actual values when scoring the prediction error.

RMSLE is an interesting metric to consider for this dataset given the target variable is right-skewed with most of the total case numbers overall being under 100 in a week, and a large number of weeks having no cases of dengue recorded. RMSLE is also more resilient to outliers compared to RMSE, but the reason it was considered for this dataset was for the fact that it incurs a larger penalty for underestimation of the Actual value than for overestimation.[[1]](#footnote-0) More on that later.

The R2 metric, also seen sometimes in the notebooks as adjusted R2, indicates the goodness-of-fit of the model on the target variables. It indicates the proportion of the variance explained by the model over the total variance.

Getting into the results of the models, the first approach was a simple ordinary least squares linear regression model. As the cities display some form of cyclical pattern in the case numbers, the first model uses time data alone to forecast the case numbers. The predictions are cyclical as expected, and the errors for each city contradict each other. The predictions for San Juan, Puerto Rico overestimate the predictions. This may be due to the higher case numbers in the train set the model learned on, having a higher magnitude, and a higher number of outbreak periods than what is present in the test set. Iquitos is the opposite situation, having overall smaller weekly case numbers and smaller magnitude outbreak periods in the train set as those found in the test set.

Using the climate features instead, the seasonal cycle appears to still be present in San Juan, albeit much noisier and not a direct cycle each year. Though the forecasts still don’t capture the true values of the weekly case numbers in the test set, the MAE is lower than modeling off of the months alone, suggesting that the climate features do play a role on the dengue spread on the people of San Juan. On the other hand, modeling off of the climate data for Iquitos, the forecasts have less variance, producing a flatter prediction plot. As explained above, the train set for Iquitos overall provides smaller case numbers per sample for the model to learn from, and the predictions on the test set are on the magnitude of the weeks not belonging to an outbreak period. The two outbreak periods present in the test set for Iquitos are not captured well by the predictions. Due to these two periods and the flatter prediction plot, the predictions off of climate data alone produce a higher MAE than modeling off of the month variables alone.

The competition outline mentions seasonal auto-regressive moving average models as a possible forecasting model to utilize for predictions and was considered in this. The estimates offered by the SARIMAX model are the best forecasts for Iquitos, Peru of all the models tested. That is to say, the forecasts capture the variance in the case numbers the best, but that is not to say that these predicted values align with the true case numbers for the Iquitos test set well. In other models tested, the variance in the case numbers, given there weren’t many outbreak periods to learn from in the train set, was not captured as well but resulted in a much better scoring metrics. The model also had a wide confidence band, indicating there wasn’t a very strong correlation of the selected features to the case numbers, leaving a wide area of “wiggle room” for the predictions to fall in and still be considered by the model.

San Juan fairs worse than Iquitos with the SARIMAX model, having much worse metrics as well as predictions that don’t fit the true values well at all. The predictions are overall over-estimating the case numbers for average weeks and are also under-estimating the outbreak periods case numbers which again I would consider a consequence of the pattern in training data for the case numbers. Indeed the predictions don’t seem able to account for the variance or conditions of the features to meaningful increases in case numbers. Alongside that, the confidence band for San Juan is wider than that of Iquitos; wide enough that predictions well above the peak number of cases in the test set would still be considered possible by the model, and that is reflected by the error scores seen. I’d like to note once again that these models were trained on a subset of features, and the auto\_arima functionality provided by the pmdarima library were both cut short because of the limitations of my system, and the time-complexity required to utilize more features or more complex modeling by the function. And with different approaches and different resources, the results of these models may differ. The SARIMAX or other models in the ARIMA family seem more apt at handling the dataset for Iquitos than for San Juan and the forecasts produced followed the true case numbers better than any other model tried.

The last family of models presented in the notebooks both incorporate the Gradient Boosting framework. Gradient Boosting builds an ensemble of weak models such that the predictors minimize a loss function. Since there aren’t any climate features in particular that seem to correlate the strongest to the number of cases in a week, many weaker models should be able to capture the different magnitudes in the variance of both the target variable and the variances in the independent variables.

The Gradient Boosting Regressor model provided by the scikit-learn library doesn’t perform well with a Grid Search for the optimal hyper-parameters, and manual assistance is needed to find a predictor that returns the best predictions. Iquitos once again proves to be trouble to predict for and results in predictions that lie in between the lower weekly case numbers of non-outbreak weeks, and well below the higher case numbers of the outbreak weeks. These predictions again appear to stay about a mean learned number of cases, and the higher case numbers predicted aren’t strongly associated with spikes in case numbers seen by the true values for Iquitos, which can be seen in some weeks predicted where the increase in case numbers falls in line with a decrease in the true case numbers for that week. As the ARIMA family of models seemed most appropriate for Iquitos, the Gradient Boosting family of models seems to be working modestly well with San Juan data. The overall performance of the predictions for San Juan are noticeably higher for almost all non-outbreak weeks, almost double the true number of cases for the given week, but the first outbreak season appears to be modeled better than any of the models seen previously, though there the predictions vary greatly within that period and aren’t consistent high like ought. The second outbreak period towards the tail end of the test set is still lacking in predictive accuracy with slightly higher case numbers being predicted a few months in advance of what is seen in San Juan. This model does seem to be capturing the small variances of the case numbers more for San Juan, though this doesn’t aid the predictive ability for the outbreak weeks.

This notebook introduced additional features engineered to the models, and reducing the feature space to a smaller subset to bolster their influence seems to hinder the model performance for the outbreak weeks by producing smaller predicted values but also improve the over-predictions of the non-outbreak weeks by reducing these predicted values as well for both cities. This model offers the worst metrics thus far for San Juan, keeping in mind that the ARIMA models were predicting the monthly sum of cases rather than weekly case numbers. It does however seem to produce predictions that are of the same magnitude, to an extent, with the pattern of the case numbers of the test set for the first outbreak period.

The second modeling approach that was tried is provided by the XGBoost library, which stands for eXtreme Gradient Boosting, and as the name implies, is a more robust and advanced package for building parallel Gradient Boosting Trees.

The resulting predictions for Iquitos are very slightly better in terms of prediction metrics, though the model is still not accounting for the outbreak periods well at all. The results don’t seem to exhibit the conflicting pattern as seen prior, with GradientBoosting Regressor, of higher predicted values during weeks of decreasing case numbers. The predicted values are still primarily modeling the values of the non-outbreak weeks, and don’t see any dramatic rise in predicted values for any week. San Juan is drastically different in the predicted values, no longer over-predicting across the whole of the test set, and staying more in line with the weekly case numbers of non-outbreak weeks. This does however nearly eliminate the benefit regular GradientBoosting Trees provided, this being the capability to somewhat predict case numbers in along the same magnitude as in some of the outbreak periods; while the MAE and sMAPE metrics are improved significantly, the RMSE metrics is unaffected. I’ve found that the XGBoost model improves when a sample weight is specified during the fit to the data.

There are different approaches for what sample weight should be specified to the model, and the one that I found seemed to work the best out of what I’ve tried was the square root of the train set case numbers with a small error term to account for the weeks with 0 cases. This approach didn’t aid in predictions for Iquitos, nor did any other approach considered. The results for San Juan, with the sample weights specified, is more prone to over-predicting for the non-outbreak weeks than without, but the first outbreak period predictions are some greater than the surrounding weeks, but are still not able to reach the neighborhood of the very high case numbers that the regular GradientBoosting model was able to achieve. It does nevertheless predict, for the first outbreak periods, that there may be some significant increases in case numbers for San Juan, and these weeks may warrant monitoring and having a response ready if needed. Unfortunately, the second outbreak period later in the test set is still eluding the model, and no significant spike in case numbers for these weeks is predicted.

Again using a subset of features, the resulting predictions for Iquitos are nearly identical as with the full feature set, and result in slightly worse metrics scores, though this difference is inconsequential. The already spotty predictions of San Juan XGBoost are made even more erratic with using a feature subset. The results are more erratic throughout the test set, but also capture the severity of the first outbreak better, but the inconsistent magnitude of case numbers predicted is still present. The second outbreak season is still lacking in the scale of case numbers predicted by the model, but the reduced number of features learned improves the sum number of cases predicted during that period.

There are many different approaches to modeling this dataset such as: modeling the log-transformed case numbers for each city, scaling the features and target variables, applying different weights to models on the train sets, and many different models were attempted for predicting the case numbers. In the end, there were no off the shelf modeling approaches that best estimated the case numbers for the test set of climate data for either city; some that resulted in close approximations but most were lacking prediction accuracy for the outbreak periods for each city. Overall, it seems that different approaches will need to be considered for each city, for example, the ARIMA family of models was seen to perform the best for the Iquitos dataset, while a tree-based approach such as the GradientBoosting models seemed better for the San Juan data, which is not to say that they performed well.

One of the issues present in both datasets is the fact that the training set contains different patterns in case number history as to that found in the testing set. Iquitos contains a smaller number of outbreak weeks that matched the severity of dengue spread compared to the outbreak weeks in the test set, and San Juan offers more outbreak periods that had more dengue cases than the spikes in case numbers that were found in the testing set. A smaller testing set can be used to allow for the models to learn some of these differences better, but that would limit the ability to test the predictions against the true values in the test set. This project took the approach of modeling the data as a time series, which doesn’t allow for the data to be shuffled but instead be modeled chronologically. Shuffling the samples so the models can learn from different points in time could aid in capturing the impact different features have on the case numbers instead of considering the trend in the variables.

MAE doesn’t seem to be the most informative metric to consider when the model performance during the outbreak periods plays importance. Many models performed better than others in the MAE scores while having predictions that didn’t match up well with the true number of cases, and were severely under-predicting the case numbers during the outbreaks. When trying to model the data to minimize this error score, the predictions are more inclined to result in lower values, ones more in line with the average case numbers of the non-outbreak weeks since there were much more of these weeks and far fewer weeks of outbreak cases. As explained above, when the goal of the modeling exercise is to predict outbreaks and the prediction accuracy of the lower non-outbreak week case numbers, it seems the RMSE and RMSLE metrics were more appropriate to try to minimize instead.

An ARIMA family model seems to be more appropriate for Iquitos and a tree-based model with a proper scoring metric specified was a more appropriate approach for the San Juan dataset. Something which was not attempted but may prove beneficial to higher prediction accuracy would be to build an ensemble of models, learning the residuals from one of these models and trying to learn from the differences between the true case numbers and the initially predicted case numbers to build a deeper learning pipeline.

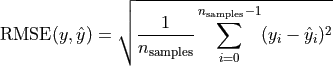
Studies done on this subject, including studies for different countries:

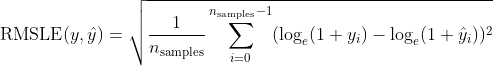
1. Rigau-Pérez, J G, et al. “The Dengue and Dengue Hemorrhagic Fever Epidemic in Puerto Rico, 1994-1995.” The American Journal of Tropical Medicine and Hygiene, 1 Jan. 2001, www.ajtmh.org/content/journals/10.4269/ajtmh.2001.64.67. p. 67-74
2. Campbell, Lindsay P et al. “Climate change influences on global distributions of dengue and chikungunya virus vectors.”Philosophical transactions of the Royal Society of London. Series B, Biological sciences vol. 370,1665 (2015): 20140135. doi:10.1098/rstb.2014.0135
3. Ewing, D.A., Purse, B.V., Cobbold, C.A. et al. Uncovering mechanisms behind mosquito seasonality by integrating mathematical models and daily empirical population data: Culex pipiens in the UK. Parasites Vectors 12, 74 (2019). https://doi.org/10.1186/s13071-019-3321-2
4. Stoddard ST, Wearing HJ, Reiner RC Jr, Morrison AC, Astete H, et al. (2014) Long-Term and Seasonal Dynamics of Dengue in Iquitos, Peru. PLOS Neglected Tropical Diseases 8(7): e3003. <https://doi.org/10.1371/journal.pntd.0003003>

Brief reference to formulas:

\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \left| y_i - \hat{y}_i \right|

\text{MSE}(y, \hat{y}) = \frac{1}{n_\text{samples}} \sum_{i=0}^{n_\text{samples} - 1} (y_i - \hat{y}_i)^2





R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}

\text{sMAPE}(y, \hat{y}) = {\frac{1}{n_\text{samples}} \sum_{i=1}^{n} \frac{|\hat{y_i} - y_i|}{|\hat{y_i}| + |y_i|}}

1. <https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a> [↑](#footnote-ref-0)