# Capstone 2 Final Report:

## Predictive modelling for mitigation of West Nile Virus in Chicago, Illinois

Blaine Murphy

7/27/2021

*Summary*

I created a predictive model to assist the City of Chicago's efforts in mitigating the risk of West Nile Virus to its citizens. The final model is an XG boost classifier as it outperformed 4 other models on this dataset. The classifier had a ROC AUC of 84% and a recall of 83% on a hold-out test set of data demonstrating its predictive capacity. Important predictive features include the historical percent of positive cases by week of the year, temperature, and the precipitation and wind speed from the previous days before the sample was collected. The model is ready for deployment.

*Context*

In 2002, the first case of West Nile Virus (WNV) occurred in Chicago, and it became such a problem that the city instituted a system of traps around the city for detection of the virus. Every summer traps are placed around the city and each week mosquito specimens are collected and test for WNV. The Chicago Department of Public Health (CDPH) then makes decisions on where to spray pesticides to decrease the mosquito populations. My task is to create a predictive model based on weather data that will enable the CDPH to proactively spray for mosquitos instead of reacting to positive tests that could be more than a week delayed.

WNV is a potentially deadly pathogen that primarily infects humans through the bites of mosquitos carrying the virus. It is the leading cause of mosquito born disease in the United States. According to the CDC, about 20% of the people who are infected will develop symptoms, and "1 out of 150 infected people develop a serious, sometimes fatal, illness." Given the risk to the citizens of Chicago, the CDPH is seeking a predictive model to assist in the pesticide spraying regime.

*Data*

The data available for this project includes data from the mosquito traps, weather data from two stations in the city of Chicago, and a dataset containing the pesticide spraying operations. The mosquito traps have location information including latitude and longitude as well as the number and species of mosquitos collected and the result of a WNV test. The traps are (mostly) sampled weekly and the data set is imbalanced as only ~5% of records have positive WNV tests. Some of the records needed to be aggregated because they were split by mosquito species or they were split due to a large number of specimens collected. After cleaning there are 4616 records in the traps data set.

The weather data is collected daily from two stations located at Chicago's airports. The data includes temperature, precipitation, dew point, wind speed and more. I added the location of the two weather stations to the data set and dropped columns that did not appear to have any value like snow fall (there is no snow in the summer). There were a small number of temperature null values that I filled in using the other weather station. There is a column called 'Depart' that represents the difference in average daily temperature and the historical average, and it is null for all records in the second of the two

stations. To remedy this, I calculated the daily historical average temperature based on the first station and then used that value to find the departure in temperature at the second station. I am glad I did this, as 'Depart' turned out to be an important feature in modelling. Finally, I filled in a small number of missing values in other columns and converted "sunrise" and "sunset" to useable numbers.

The spraying data set includes the date, time, and location of several spray runs in different areas of the city. Spraying is a reactive measure, so for this project I focus on the weather data for predicting WNV cases.

### *EDA and Processing*

I explored traps and weather data set and noticed a few important phenomena that influenced the features I created for modelling. Firstly, there are certain times of the summer when the probability of mosquitos is greater. I added columns based on the historical presence of positive WNV cases by year, month, and week of year. I also noticed there were general trends in the data that correlated to the presence of mosquitos, such as warmer temperatures and lower wind speeds. These observations and intuition of the problem led me to creating lagged variables. For several features in the weather data set I created time lagged features up to 10 days before. I also created features aggregating values from previous days. Then, I merged the traps and weather data sets on date and which weather station was closest to the trap.

Next, I divided the data into an 80% training set and 20% testing set, ensuring that there is an equal proportion of positive cases in each. Using only the training set, I calculated the information value of each feature relative to the target and dropped those with very low value or suspiciously high value reducing the number of features from 89 to 64. To reduce overfitting in the modelling phase I checked collinearity of the features by calculating the variance inflation factor among the predictors. If a feature can be very accurately predicted from other features in the dataset, then it gets dropped. Performing this step reduced my number of predictive features to 13 for modelling.

Finally, I scaled all of the final predictor features to mean zero and standard deviation one based only on the training set.

### *Modelling and results*

I attempted 5 different models for predicting positive cases, logistic regression, random forest, gradient boosted trees, XG boosted trees, and a support vector classifier. The XG boost model performed the best on this data set with an ROC AUC of 84% and a recall of 83% on the hold out testing set as shown in Figure 1. I chose the model based on highest AUC that also had a strong recall score. High model recall is a priority for this project, because I'd rather the model make a false positive than a false negative. The random forest and support vector classifier also performed well but had performance metrics slightly below the XG boost model.

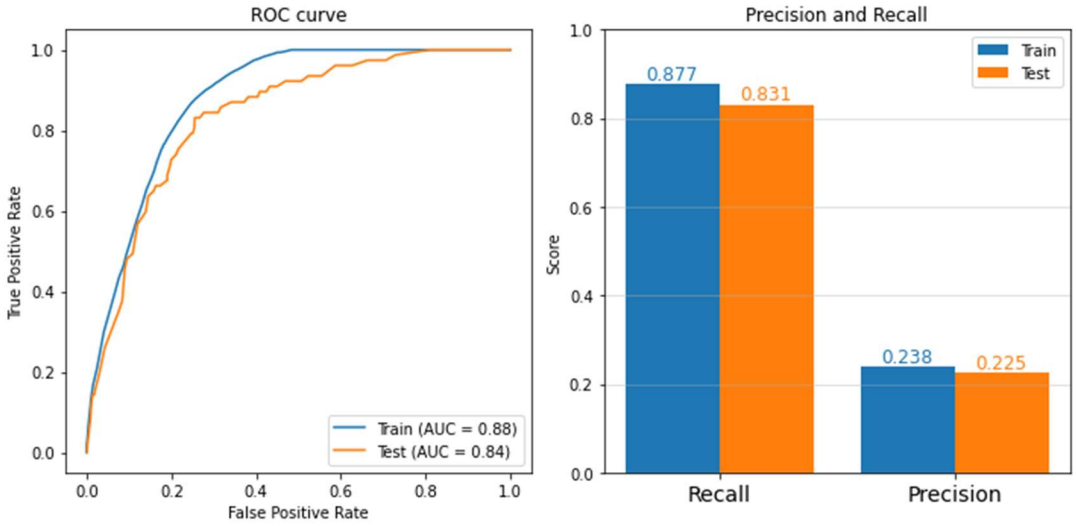


*Figure 1. The ROC curve shown on the left is steep for both training and testing, and the high AUC indicates that model's predictive power. The curves are close together indicating the model is not overfit to the training data. The plot on the right shows both the training and testing precision and recall. High model recall is a priority for this project.*

Figure 2 depicts the features of the model and their relative importance to the output. The most import feature in the model is the historical weekly percentage of positive WNV cases. The second and third most impactful features are the wind speed lagged 5 days and 3 days. When these values are high there is a lower probability of WNV being present at the trap location. The fourth most important feature is the temperature departure from normal. When the temperature is warmer than normal there is a greater probability of a positive WNV. Most of the other features that impact the model are lagged precipitation values. From my experience living in Houston, TX, mosquitos are most prevalent on hot, still days after a recent rainfall, so the model and its features pass my stupidity check.

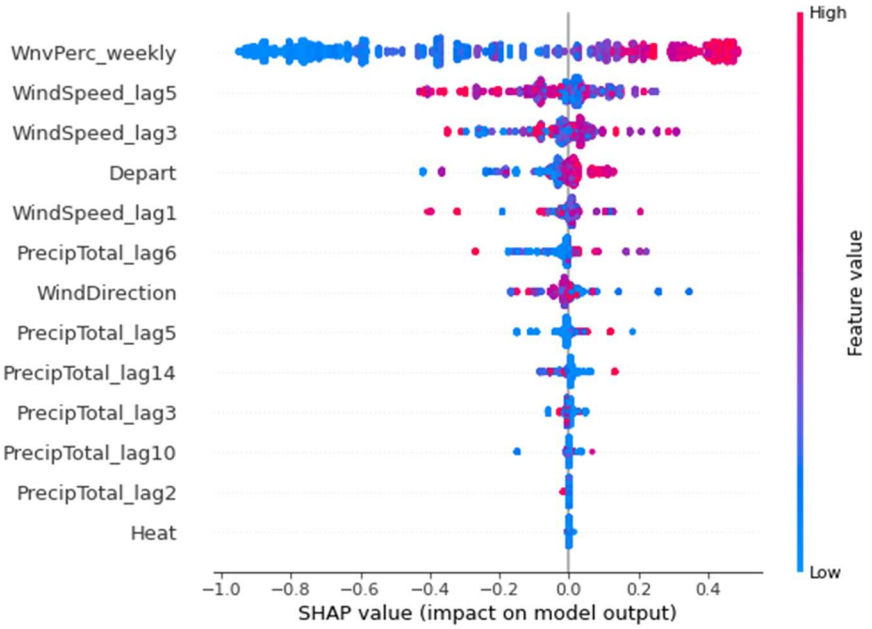


*Figure 2. The plot shows the relative importance of each feature to the output of the model. The historical weekly WNV positive rate, the lagged wind speed, and the temperature departure from normal have the greatest impact on the model output.*

### *Recommendations*

I chose the XG boost model for its excellent performance on the testing hold out set. I then trained the model on all of the available data and it is ready for deployment. I recommend utilizing up to date weather data and running this model daily to predict locations in the city in which there are a greater probability of mosquitos carrying WNV. When the model predicts there will be outbreaks of WNV, deploy a pesticide sprayer to suppress mosquito populations and thereby the risk of WNV to Chicagoans. The model performance indicates there will be false positives and project costs are always a constraint. I recommend focusing pesticide spraying efforts on those locations that have the greatest predicted probability value calculated by the model.

### *Future work*

After seeing the aerial distribution of traps and weather stations I began to suspect that the spatial sampling of Chicago weather is not dense enough for the task. I would like to see if the accuracy of the model decreases as a function of the traps distance to the nearest weather station. If it does, then I would recommend adding another weather station during the summer months for gathering more data.

On the next iteration of this project, I would like to include the pesticide spraying data in the predictive model. I can see a situation in which the spray causes what would normally be a positive case based on weather conditions to result in a negative test due to the elimination of mosquitos in the area. If that situation occurs then the predictive accuracy of the model could be improved by including the spraying data.