West Nile Virus Final Report
Jonathan Jacobs
Springboard Data Science Career Track
November, 2020

Chicago's Problem with West Nile Virus:

According to the CDC, West Nile Virus (WNV) is the leading cause of mosquito-borne disease in the US. The Chicago area reported 6 cases of WNV in the summer of 2020. Although this figure is small, the disease is dangerous, proving fatal for around 1 in 150 people who become infected. Mosquitos, beyond being a nuisance, are a public health concern, and it's important for densely populated urban areas to control the mosquito population.

The primary method of mosquito control is to spray insecticide over large areas of land. Along with environmental costs, there are significant costs and inconveniences associated with controlling mosquito populations. The city of Chicago Department of Public Health treats 40,000 water basins each year with larvicide and monitors 83 traps around the city each week for mosquitos with WNV. It's costly both in terms of time and resources, and yet there are cases of WNV reported every year.

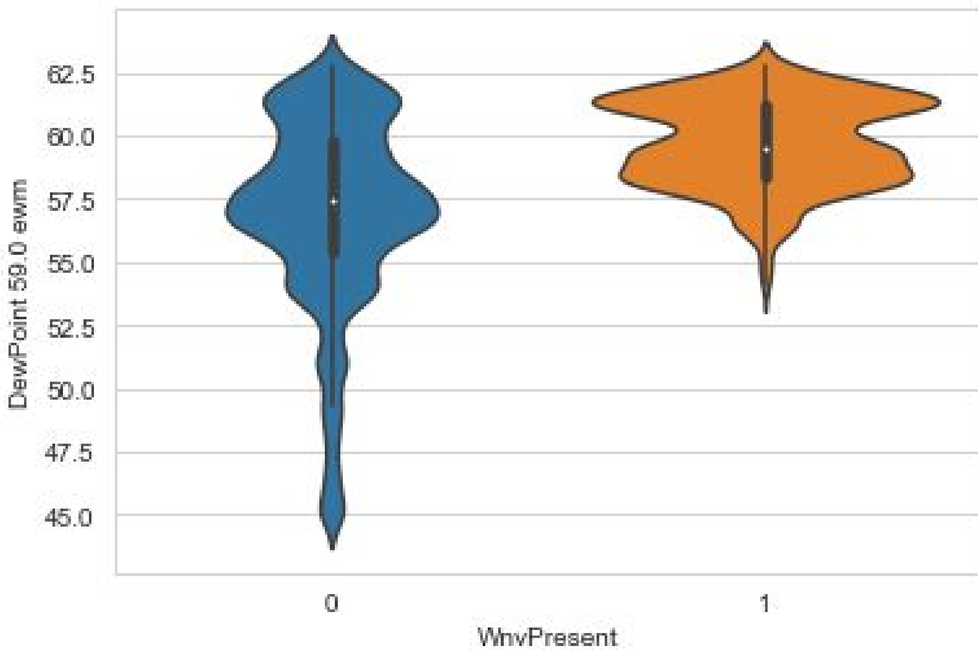
Reactive to Proactive:

Chicago's approach is reactive, kicking into high gear only after WNV has been identified in an area. Here, I suggest a proactive approach based on weather and historical spray data that allows for proactive, targeted intervention to prevent WNV from affecting communities.

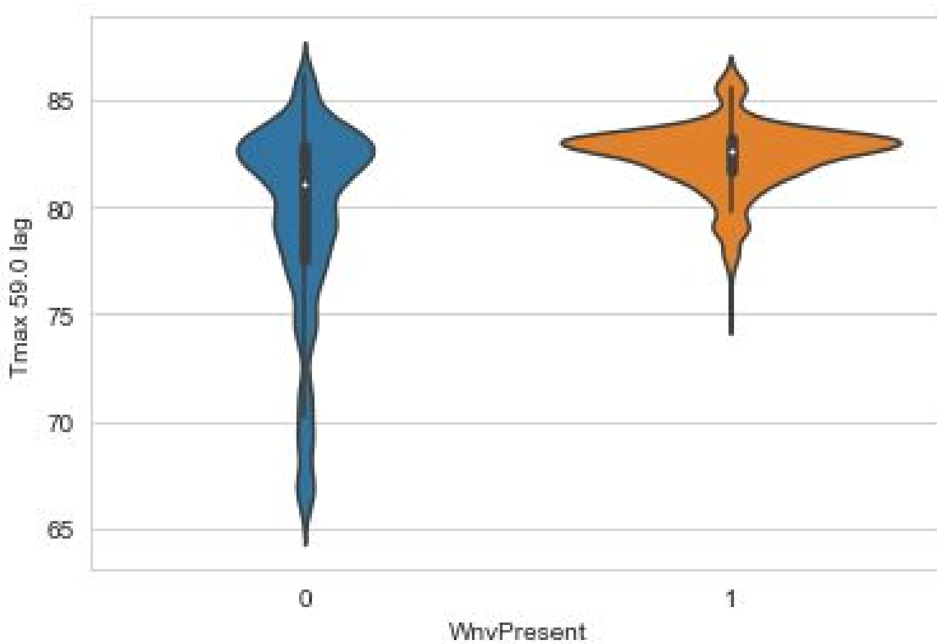
The first step in this process to predict WNV occurrences with a sensitivity of 80% or more.

Understanding Factors Relevant to WNV Carrying Mosquitos.

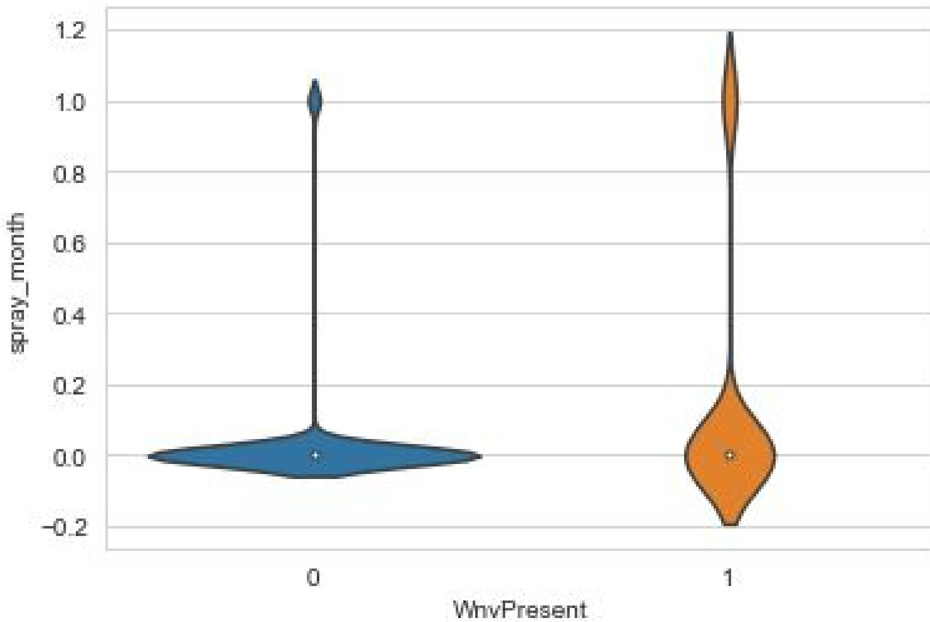
Weather and spray data published by the City of Chicago provide a strong dataset for the prediction of WNV carrying mosquitos. Below we see a comparison of dew points averaged over two months, a proxy for humidity and general precipitation, as it relates to WNV sightings (1) or non-sightings (0). WNV is rare, so the distributions are not on opposite ends of the spectrum, but it's clear that mosquito populations are hurt by low-humidity and below a dew point of 53 degrees F, there have been no observed WNV carrying mosquitos.



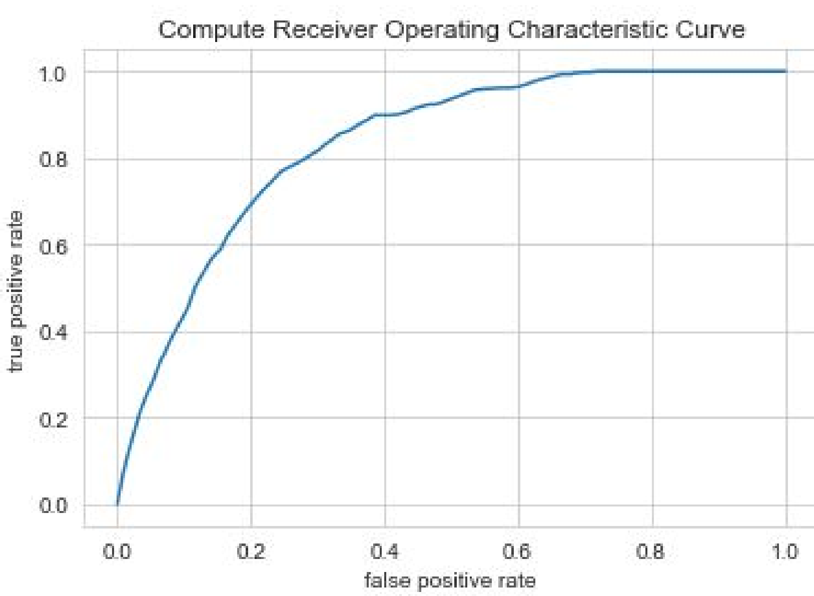
Similarly, we see that prolonged high temperatures are related to WNV sightings.



And lastly, as one might expect, if an area has been sprayed in the past month, it's less likely that WNV carrying mosquitos will be found there.



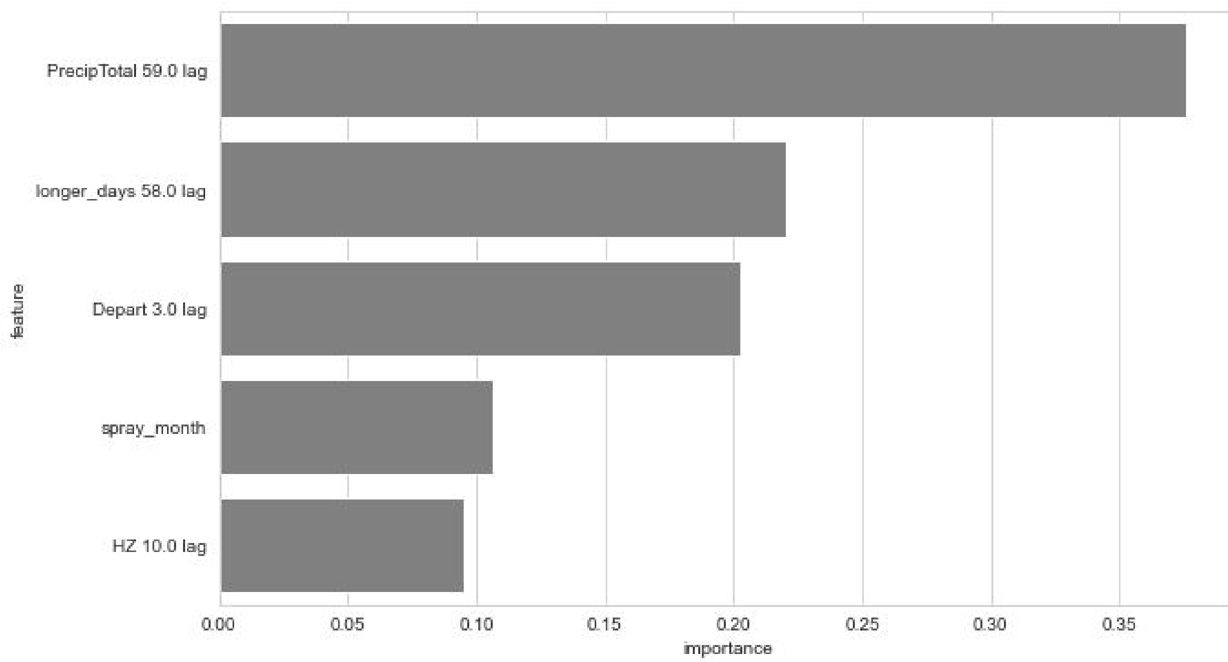
Modeling and Prediction:



Using a random forest classifier with 750 estimators and a max depth of 9, we're able to achieve good predictive power and model performance. The Compute Receiver Operating Characteristic Curve shows how the true positive rate (sensitivity) and false positive rate vary based on the probabilistic threshold to assign an observation as having WNV mosquitoes or not.

What this figure shows is that we're able to predict over 90% of WNV carrying mosquitos correctly. And the model is able to do so without drowning out those positive signals with many false positives. With a threshold of roughly 4.5%, the false positive rate is around 50%, meaning that on the data set available one of every two positive indications of WNV carrying mosquitos is accurate. Given the preventative approach, and the inherent inaccuracy of models, this is a great result.

We expect that the model will be resistant to sampling error. Information value and variance inflation factors were used to cull features from the dataset that we suspected were artificially good predictors due to sampling error, and we end with only five metrics to track for prediction. Below are the metrics with their relative importance for the model.



Future Work and Recommendations

1. Validate the model's predictive capacity through integration with existing mosquito monitoring. Public health rides on the decisions on how to control the mosquito population, so a measured and careful approach is necessary to validate the efficacy of the model as a predictive tool. For the mosquito season in 2021, the model's predictive power can be assessed by predicting which of the 83 traps are expected to have WNV

carrying mosquitoes. Field validation, and a sensitivity of 90% or more, will build confidence in the technology's ability to act as a tool for the health department without endangering public health.

2. Integrate the model into mosquito control decision making for the 2022 season. Once the model's efficacy has been demonstrated and the sensitivity has been shown to be significant, the model can be used to effectively target locations for mosquito control.