1. Introduction and Context

West Nile Virus (WNV) is most commonly spread to humans through infected mosquitoes. Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever to serious neurological illnesses that can result in death.

In September 2001, the first human cases of West Nile virus were reported in Chicago. By 2004, the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today.

Every year from late-May to early-October, public health workers in Chicago setup mosquito traps scattered across the city. Every week from Monday through Wednesday, these traps collect mosquitos, and the mosquitos are tested for the presence of West Nile virus before the end of the week. The test results include the number of mosquitoes, the mosquito species, and whether or not West Nile virus is present in the cohort. The results of these tests influence when and where the city will spray airborne pesticides to control adult mosquito populations.

2. Problem Statement

Given weather, location, and potentially spray data, we must predict whether or not West Nile Virus (WNV) is present for a given time, location, and species. An accurate method of predicting the presence of West Nile Virus will help the City of Chicago and the Chicago Department of Public Health (CPHD) more effectively and efficiently combat the transmission of this deadly disease.

3. Domain Research

Prior to exploring our dataset, we've engaged in some basic research regarding both West Nile Virus and mosquito behaviors. Domain knowledge will help us better understand our data, make more informed decisions while moving through the data science process, and better interpret our results.

First and foremost, while many species of mosquitoes are known to be carriers of West Nile Virus, of those present in Chicago only 2 species are known carriers - restuans and pipiens. Furthermore, a recent scientific paper has found hybridized carriers may be more likely to carry and transmit the disease. We do see hybridized restuans/pipiens present in the Chicago area.

Mosquitoes have a 4-stage life cycle, and the time from egg to adult takes approximately 8 to 10 days. Under normal circumstances, an adult mosquito lives for 2 to 3 weeks. If examined by sex, male mosquitoes live for roughly a week on average, while female mosquitoes can live for one to two months. This is an important distinction for our model as only female mosquitoes seek a blood meal.

Mosquitoes are also known to be much more active during warmer months. Interestingly, we've also learned the conditions ideal for breeding and conditions ideal for seeking a blood meal are not necessarily the same. Only female mosquitoes seek a blood meal as this is required for

breeding, and they tend to do so when conditions are dry and they are dehydrated. In contrast, mosquitos prefer damp and humid conditions for breeding activity and laying eggs. Consequently, we would expect to see the effects of some of these weather conditions days or weeks after they occur.

Finally, we know mosquitos are adverse to a number of weather conditions, and even the extremes of those they find favorable. Mosquitos will be damaged in high wind and heavy rain, and extreme heat can dehydrate them to the point of death. Extremes on both ends of the spectrum seem likely to affect mosquito behavior.

Our domain research has provided some valuable information to keep in mind as we move into exploring our data. We suspect mosquitoes will breed when it is warm, humid, damp, and relatively low wind. This may take place after a precipitation event or storm, but not during it. We also suspect mosquitos will seek a blood meal when it is hot, dry, and low wind as they may be dehydrated. Importantly, we will keep this knowledge in mind when exploring our data, engineering our features, and interpreting the importance of those features.

4. Data Source and Procurement

The data we are using was procured through Kaggle. The data were provided to Kaggle by the Chicago Department Health, and the competition was further sponsored by the Robert Wood Johnson Foundation.

The data initially consisted of 3 primary sections. The first is GIS data of spraying efforts in 2011 and 2013, spray.csv. This dataset included the data and time of the spray, as well as the latitude and longitude of the spray.

The second section contains weather data from 2007 to 2014, weather.csv. Accompanying this csv is a document of column descriptions from NOAA. As the file name suggests, this dataset effectively describes weather conditions on a daily basis. Temperatures, wind speeds, precipitation and dew point, the date, etc.

The third and final section is split between the training and test set. The training set consists of data from 2007, 2009, 2011, and 2013. The test set asks us to predict test results for 2008, 2010, 2012, and 2014. This data contains information like the date, trap location and Id information, mosquito species, number of mosquitos, and in the case of the training set weather or not west nile virus was present.

5. What EDA was performed?

We first explored the spray data. Spray data was available for just two years, occurred only in specific regions, and did not take place during the time frame for which we need to make our predictions. This represented a significant amount of data that could not be reasonably imputed. As a result, we decided to drop the spray data from our analysis.

We then explored our training dataset. First, we checked the balance of the data with relation to our target variable. It quickly became apparent the dataset was imbalanced as West Nile Virus is far more often not detected than detected. Next, we utilized our datetime features to see if there were specific time periods during which we may be more likely to detect the virus. Unsurprisingly, the occurrence of West Nile Virus increased year over year and detection skyrocketed in August and September. Given what we've learned about mosquitoes during our domain research, this makes a lot of sense. Finally, we examined the species captured and whether or not they were carriers. Interestingly, mosquito species captured in the highest numbers were known carriers. Less conclusively, pipiens seemed to be the primary carrier with restuans being less so. The big takeaway, however, was observing hybridized pipiens/restuans presenting as West Nile Virus carriers more so than a typical restuans. This jives with recent studies suggesting hybridization could have significant impact on transmission patterns of WNV.

We then explored the weather dataset, focusing primarily on weather conditions most known to affect mosquito behavior. Average temperature peaked in mid to late July. These are great breeding conditions for mosquitoes, and we suspect this could lead to their increased presence in August. We then compared precipitation totals by month. A nice bump in precipitation leading into July through mid/late July was observed. Naturally, this led to questions regarding damp and humid conditions. Exploring features related to moisture seemed to indicate more damp and humid conditions from June through early August. Our domain research suggests these conditions tend to favor mosquito activity and, more specifically, breeding activity. Finally, we explore our wind features. Average wind speeds were found to be fairly low in late June, July, and September, but lowest in August.

At this point, exploratory data analysis is beginning to tell a story. We suspect increased precipitation along with damp and humid weather are creating ideal conditions for mosquitoes to breed in June and July. This may be leading to increased numbers in August and September, when it is both hotter and drier. We think mosquitoes may be dehydrated and seeking a blood meal during these months. Furthermore, the low wind and lack of extreme weather conditions in August and September does not impede their ability to seek that blood meal.

Finally, we combine our weather and training datasets to see if we can discover any additional relationships. As temperature increases, we find instances of West Nile Virus tend to as well. There is not much activity in May or June, but late July and August sees WNV detected at temperatures of, roughly, 68f and above. What's very interesting is September, where we see instances of WNV at much lower temperatures. We suspect these are the remaining mosquitoes being trapped in cooler weather before they start to die out; almost like a transitionary period. We then explore relationships with regard to our moisture features. As expected, we see more instances of WNV when conditions are more damp and humid. We capped off our EDA by digging into wind speed. We observed more instances of WNV the lower the average wind speed was, suggesting mosquitoes may not be biting when wind conditions are too extreme.

The exploratory EDA seems to corroborate recent research on mosquitoes and West Nile Virus. It also aligns with what we've learned during domain research. Our EDA suggests that specific carriers of West Nile Virus prefer warm, damp, and humid weather to breed. They also seem to come out for blood meals when they are dehydrated and the weather is hot and dry. However, they also seem to dislike extreme conditions on both ends of the spectrum. Weather that is cold, overly rainy, too dry, or too windy does not seem to increase their activity and may, in fact, decrease it.

6. Feature Exploration and Engineering

Domain research and EDA has created a solid roadmap for further feature engineering. First and foremost, we've decided to lag weather features based on the mosquito life-cycle at 1 week, 2 week, and 3 week intervals. A rolling window, rather than a static window, was used with the mean of the period for lagging features to better represent the effect of overall conditions. Missing values created as a result of lagging features were backfilled. Next, we created dummy variables for the species of mosquito in both our training and our test set so our model can recognize their difference. The categorical species feature was then dropped.

Finally, we merged both the training and the test set with our weather data. Additional features determined to not be beneficial to our model, like sunrise and sunset values, were dropped from both combined datasets. The date feature was dropped and replaced with individual year, month, week, and day features to better represent a record's specific place and time. Additionally, we were able to determine there was no correlation between our predictor variables. At this point, our data is clean and ready to begin the modeling process.

7. Model Building and Tuning

We've chosen three model types to explore that tend to do well with binary classification problems; logistic regression, random forest, and extreme gradient boosting. Before building and testing these models, however, our data required some additional preprocessing to make it model ready.

We first identify our dependent variable, the presence of West Nile Virus, and separate it from our independent or predictor variables. We then split our data into train and test subsamples. It is important we do the split before transformations like oversampling to avoid data leakage.

We've identified our data as being imbalanced, and have the choice of either undersampling or oversampling. Having tried both and given our somewhat limited number of observations, it was determined oversampling led to the most reliable results. We oversample only the training subsample using SMOTE. Finally, we scale our data to prepare for model creation. A MinMax Scaler was chosen given the non-normal distribution of our data and its original, imbalanced nature.

Prior to tuning the models logistic regression showed an AUC of 0.81, random forest showed an AUC of 0.75, and XG boost showed an AUC of 0.81. Using gridsearchCV to tune

hyperparameters, all models ended up with an AUC very close to 0.82. The highest AUC achieved for this problem is 0.86, so we're in the ballpark with our modeling.

Considering the AUC is highly similar among all models, I would be most inclined to recommend XG boost. Using SHAP to interpret the model's feature importance shows the model puts the most importance on features we've identified during our domain research.

Unsurprisingly, the species of mosquito is among our strongest predictors. This makes sense given only certain species carry the disease. Also among our strongest predictors are lagged weather features relating to moisture, wind, and temperature. This is, again, unsurprising. Mosquitoes need standing water and moderate temperatures to breed. The mosquitoes also need time to grow in population and move through their lifecycle, and our lagged features seem to do a good job of capturing this fact. The day of the year is also a strong predictor, lending credibility to the idea mosquitoes are biting more in August and September. Most important but perhaps least interpretable are our location features, latitude and longitude. These features are given a very high importance, so it may be worth exploring what the landscape looks like in areas the city identifies as having the highest prevalence of West Nile Virus.

8. Recommendation

Based on the fairly strong ability of our model to predict an instance of West Nile Virus and its ability to explain why based on feature importance, we can devise an action plan for the city of Chicago to combat West Nile Virus.

1. Adopt Best Practices City Wide

The city should create a plan to help the average citizen prevent mosquito breeding. The transmission cycle can be broken by removing stagnant water from potential mosquito breeding habitats. This includes breaking up hardened soil, emptying flower pots and buckets, overturning empty vessels, and keeping roof gutters clear. Mosquitoes also like dense vegetation, so efforts can be focused on clearing brush or overgrowth. Finally, citizens can be urged to use bug spray, apply it regularly, and wear loose fitting, long sleeve tops and pants to mitigate bites.

2. Focus Mitigation Efforts

The city can use our model to predict West Nile Virus hotspots and focus mitigation efforts in those areas. Once an area is identified as a potential hotspot, the city can:

- Spray in suspected hotspots, particularly from late May through early September when mosquito activity ramps up. If a specific spray is more effective against pipiens or restuans species, it should be used here.
- Focus government resources toward removing stagnant water and vegetation overgrowth from predicted hotspots. This could include flyovers to identify areas in which potential for standing water or overgrowth is an issue, and informing the owner if it's on private land.

- Offer government assistance in hotspots, particularly in aforementioned areas where the potential for standing water or overgrowth is identified. This may include distributing mosquito repellant to citizens in hotspots, offering a free check and mitigation effort to homes in hotspots (ie, a yard cleanup) and providing resources to help citizens in hotspots keep mosquitoes at bay. I would likely suggest a city-funded WNV task force that undertakes mosquito prevention efforts citizens should be performing themselves (but likely aren't.) We suspect most people will take the free help.

Because West Nile Virus is highly seasonal, these prevention efforts should start in the Spring, before conditions exist for mosquitoes to breed.