Why

In 1999, the United States detected the West Nile virus (1).

This virus thrives in a cycle involving birds and mosquitoes, with birds serving as primary hosts. Mosquitoes get infected by the birds and typically infect humans during late summer and early fall, leading to potentially severe illnesses and, in some cases, fatalities (2).

The ability to predict the presence of the virus holds immense importance in unraveling the optimal conditions for mosquitoes to carry and transmit the virus to humans. By harnessing the power of a more accurate machine learning model, the aim is to forecast and potentially prevent outbreaks, consequently curbing the transmission of the West Nile virus.

Audience

The City of Chicago and Department of Public Health (of Chicago).

Data

The data used for this project was the Kaggle dataset on the West Nile virus Prediction. The City of Chicago, in collaboration with the Chicago Department of Public Health, has implemented a comprehensive program to monitor and track the West Nile virus. This involves an extensive system of mosquito traps, checked on a weekly basis from late spring through fall (3).

The datasets used were:

1. Training dataset: This dataset has trap names, dates, specific locations, the number of mosquitoes caught at the traps, the various species of mosquitoes captured at each trap, and whether the West Nile virus was detected.

2. Weather dataset: This dataset comes from two distinct sources, represented by weather stations situated at two airports. It includes information such as dates, temperatures, precipitation levels, sunrise and sunset times, and other relevant weather-related metrics.

Data Wrangling

Both datasets were cleaned before merging the two together. An overview of how each was cleaned is below.

The training dataset:

1. Converted the date object column to datetime format.
2. Collapsed the species categorical column to a binary numeric column to species that do not carry the virus (0) and species that do carry the virus (1).
3. Kept the latitude and longitude columns of the trap location and dropped the other address columns.
4. Dropped trap names.
5. Assigned each trap location to one of the two stations in the weather dataset based upon which one was closer in distance. This had to be done so could merge the two datasets on date and station column.

The weather dataset:

1. Imputed values for the missing values in various columns.
2. Dropped columns that had half to all missing values.
3. Replaced values that represented trace amounts, ‘T’, to 0.005.
4. Changed the numeric columns that were categorized as objects to either ‘int’ or ‘float’.

I merged the two datasets on date and station column.

Exploratory Data Analysis

During the exploratory data analysis step, different relationships were explored in temperature, precipitation, and humidity across the weeks to identify potential patterns connected with the presence or absence of the West Nile virus. First noticed a trend of mosquitoes carrying the West Nile virus exhibited an upsurge beginning around week

32 and reaching their peak between weeks 33-36 (Figure 1). Additionally, a discernible pattern emerged when plotting a 14-day average temperature shift across the weeks. Particularly, it showed a pattern for the rise of mosquitoes (with and without the West Nile virus) when temperatures ranged between 23 to 26 degrees Celsius. This has been noted as ideal temperatures for mosquitoes to thrive (6). This sheds light on how potentially temperature can play an important factor for the rise of the mosquito population in Chicago.

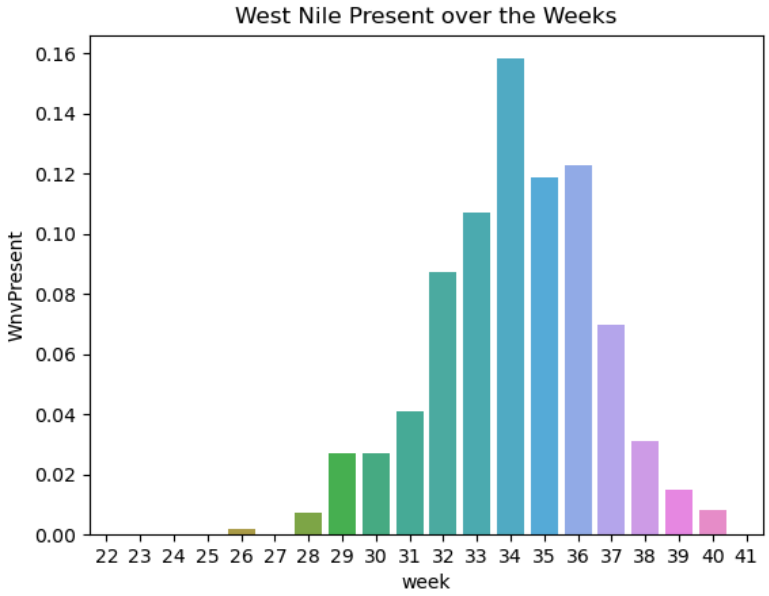

Figure 1: Showing percentage of the West Nile virus present over the total mosquitoes present.

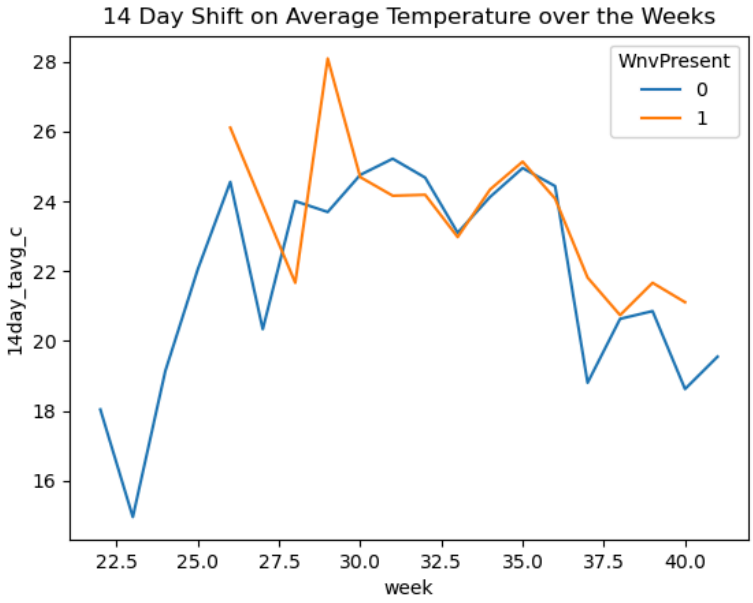

Figure 2: Showing the 14 day shift in average temperature against the weeks for both mosquitoes with the West Nile virus and without the West Nile virus.

Another significant finding from the EDA step was the large number of observations that do not carry the West Nile virus. In the training data, 120,520 observations did not have the West Nile virus whereas 14,519 observations did show to have the West Nile virus.

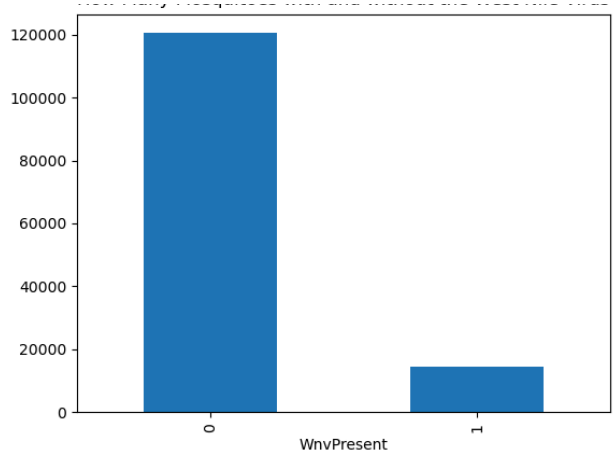

Figure 3: Bar plot to show the number of observations without the West Nile virus (0) and with the West Nile virus (1).

Feature Engineering

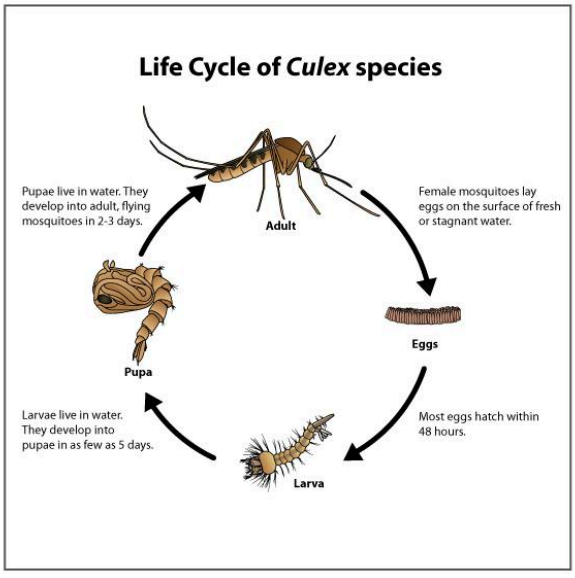

Figure 4: The lifecycle of a Culex mosquito. Image taken from the CDC website (4).

Mosquitoes' life cycle starts with female mosquitoes laying eggs in standing water that can be in drains (4). Eggs hatch especially in more drought-like conditions as the organic matter in the water becomes richer for the eggs (2). In particular, temperatures ranging from 23-26 degrees Celsius are ideal for Culex species' life cycle (6). Once

hatched, it can take 7-8 days until they become adults (4).

The birds are more congested to the standing water in more drought-like weather where the mosquito life cycle is thriving (2). The mosquitoes, which are more active at night (7), acquire the West Nile virus when they bite infected birds. Once infected, they then can infect humans when they blood feed on humans (2).

Based upon these known facts above about mosquitoes new columns were created:

1. Created a binary column if it had not rained that day (0) or if it had rained that day (1).
2. Created a relative humidity column by finding the calculation using the other numbers collected in the weather dataset.
3. Shifted three columns (average temperature in Celsius, precipitation, and relative humidity) by 7, 14, 21 days to see how previous weather affected the number of mosquitoes with the West Nile virus.
4. Two columns were added based upon how much daylight and nighttime there were for each day by extracting that information from the ‘Sunrise’ and ‘Sunset’ columns.

Modeling

For the preprocessing steps, an oversampling technique was applied to the minority group (observations with the West Nile virus) to then have 10% of the majority group. Next, an undersampling technique was applied to the majority group to create the same amount of observations in both groups. .

For the modeling step, three algorithms were tried: Logistic Regression, Random Forest, and XGBoost. Hyperparameter tuning was applied using both the GridSearchCV from the Scikit Learn package and the HyperOpt package (5).

The metrics used to choose the best model were the ROC AUC score and the ROC curve.

The ROC AUC is a single scalar value that quantifies the overall performance of the classifier. It represents the area under the ROC curve. The score ranges from 0 to 1, where a score closer to 1 indicates a better model. An AUC score of .5 shows that the model is no better than random guessing.

The ROC curve is the graphical representation of the performance of the model at various thresholds. It plots the True Positive Rate, TPR aka sensitivity, against the False Positive Rate, FPR, as the discrimination threshold is varied. The curve represents the

model's ability to distinguish between the classes and shows how sensitivity and FPR (which FPR = 1 - specificity) change for different classification threshold values.

Lastly, the SHAP package was used on the best machine learning model on the data to help extract the feature importance and help visualize important relationships in the model.

Best Model

The XGBoostClassifier with the default values produced the best model. The best hyperparameter values were : eta: 0.3, 'max_depth': 6, 'min_child_weight': 1, 'colsample_bytree': 1, 'gamma': 0, 'n_estimators': 100, 'subsample': 1}.

From the best model, the ROC AUC score was .8417 and the ROC curve is shown below.

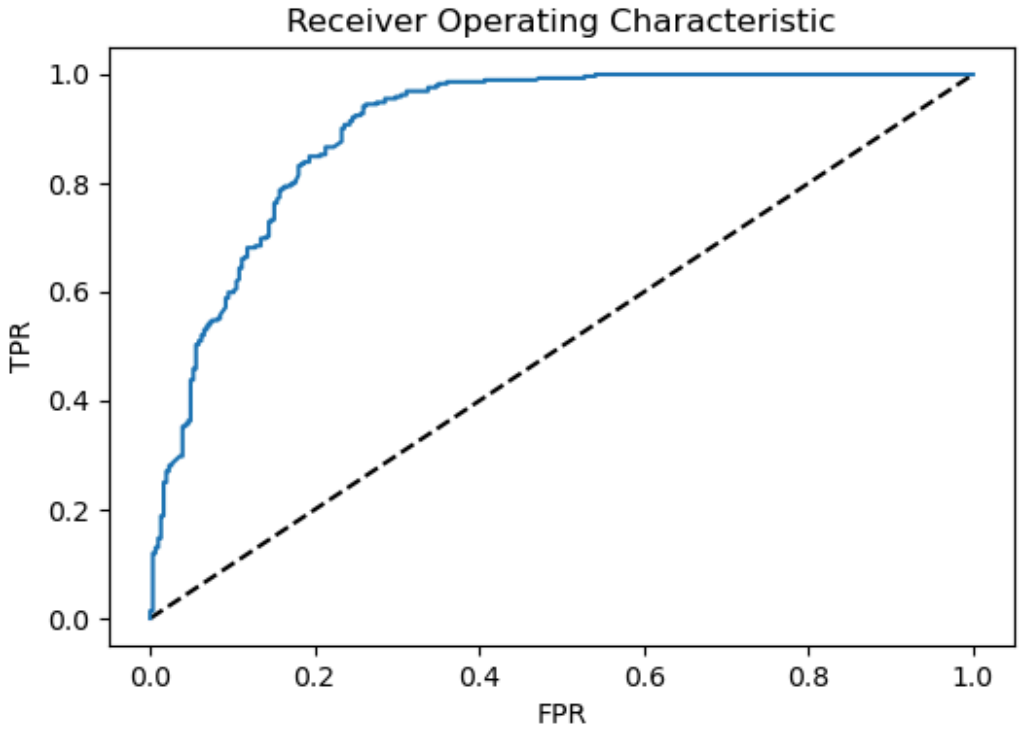

Figure 5: ROC curve from the best model using XGBoost and HyperOpt.

Used the SHAP package to make the bar plot below to help rank the model's feature importance. This bar plot gives the absolute mean value of the SHAP value for each feature. It lists the most important features at the top and decreases as they go down. The top features for the best model were the number of mosquitoes, amount of daylight time, latitude, longitude, year, average speed, month, and week.

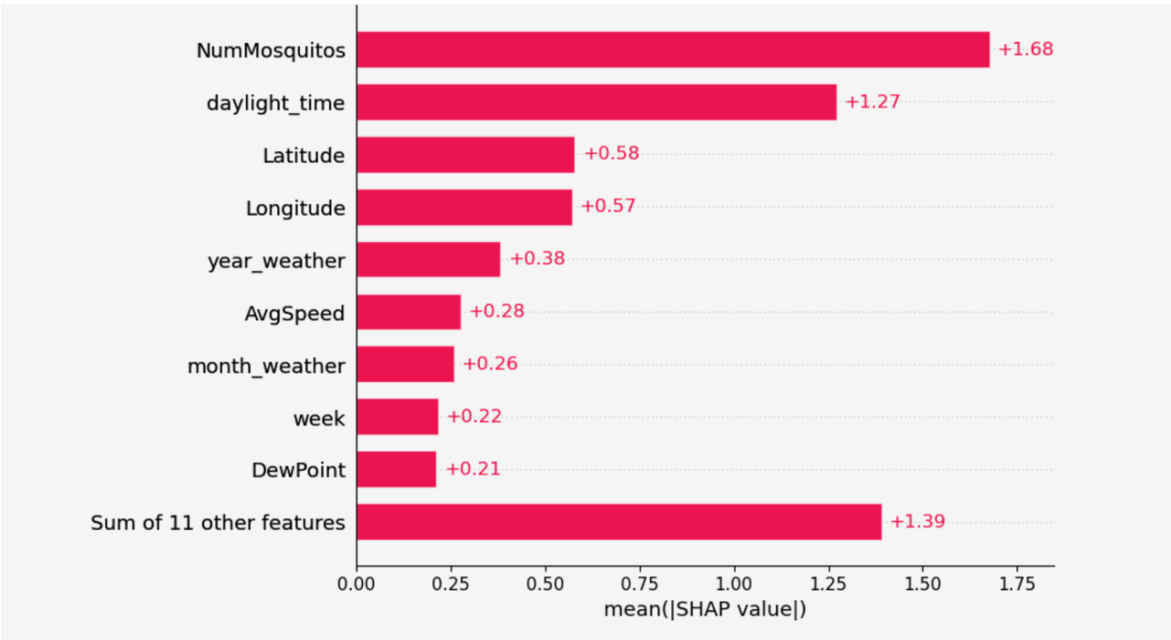

Figure 6: Bar plot from SHAP package to help show important features on the best model.

The beeswarm plot from the SHAP package shows the same ordering of the features as the bar plot above. This plot visualizes the distribution of the SHAP values for each feature. The points on the graph reveal the relationships of what features have an impact on the model's predictions. For the number of mosquitoes, this graph reveals low negative SHAP values have low feature values whereas higher SHAP values have higher feature values.

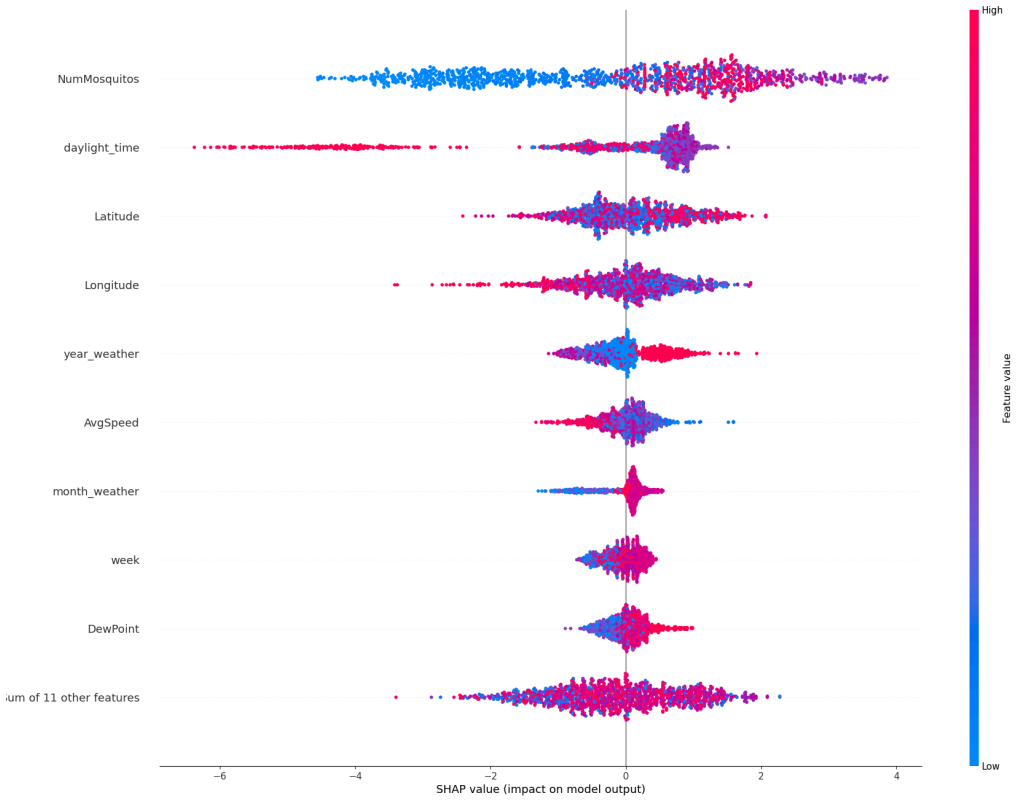

Figure 7: Beeswarm plot to plot each SHAP value for each feature and their respective feature value.

Conclusion:

Several key insights were derived using the best model and SHAP analysis. Primarily, the quantity of mosquitoes emerged as the most influential feature for the model, which aligns with statistical reasoning—increasing mosquito numbers typically correlate with a higher likelihood of encountering mosquitoes carrying the West Nile virus. Additionally, notable features within the model encompassed variables associated with daylight duration. This correlation makes sense considering that mosquitoes, predominantly active during the night, have their activity periods integrated into the model, allowing for a longer window for these insects to bite infected birds and contract the West Nile virus.

Moreover, findings from the exploratory data analysis, as illustrated in Figure 1, revealed a distinct relationship between the prevalence of West Nile virus in mosquitoes and the timing of the year, notably peaking around weeks 33 to 36. This observation notably featured as one of the primary elements in the SHAP analysis, further solidifying its significance in explaining the model's predictions.

Finally, the location through longitude and latitude showed to be significant for the model. This insight is crucial for the City of Chicago and the Chicago Department of Health as it could narrow down where the mosquitoes with the West Nile virus are more

likely present and focus on those areas to curb transmission.

Future Improvements

Future work could be done to this project. Firstly, exploring the removal of the feature of the "number of mosquitoes" could significantly streamline the West Nile virus prediction process for the City of Chicago and Department of Public Health.

Further work could be manipulating the existing dataset features by shifting them over various days, such as sunrise, sunset times, and daylight duration.

Additionally, a crucial extension to this work would be to include additional data from other years. This could potentially unlock trends from how existing years play a role into the next year's prediction.

References:

1. Colpitts, T. M., Conway, M. J., Montgomery, R. R., & Fikrig, E. (2012). West Nile Virus: biology, transmission, and human infection. *Clinical Microbiology Reviews*, *25*(4), 635–648. https://doi.org/10.1128/cmr.00045-12
2. D'Amore, C., Grimaldi, P., Ascione, T., Conti, V., Sellitto, C., Franci, G., Kafil, S. H., & Pagliano, P. (2023). West Nile Virus diffusion in temperate regions and climate change. A systematic review. *Le Infezioni in Medicina : Rivista Periodica Di Eziologia, Epidemiologia, Diagnostica, Clinica E Terapia Delle Patologie Infettive*, *31*(1). https://doi.org/10.53854/liim-3101-4
3. Wendy Kan. (2015). West Nile Virus Prediction. Kaggle. https://kaggle.com/competitions/predict-west-nile-virus
4. *Culex Mosquito Life Cycle | CDC*. (2022, July 12). Centers for Disease Control and Prevention. https://www.cdc.gov/mosquitoes/about/life-cycles/culex.html
5. Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. TProc. of the 30th International Conference on Machine Learning (ICML 2013), June 2013, pp. I-115 to I-23.
6. Shocket, M. S., Verwillow, A. B., Numazu, M. G., Slamani, H., Cohen, J. M., Moustaid, F. E., Rohr, J. R., Johnson, L. R., & Mordecai, E. A. (2020). Transmission of West Nile and five other temperate mosquito-borne viruses peaks at temperatures between 23°C and 26°C. *eLife*, *9*. https://doi.org/10.7554/elife.58511
7. Vector Disease Control International. (2023, May 3). *West Nile virus: Education, public health, mosquito management*. https://www.vdci.net/vector-borne-diseases/west-nile-virus-education-and-mosqui

to-management-to-protect-public-health/#:~:text=West%20Nile%20virus%20is%20spread,feed%20from%20evening%20to%20morning