West Nile Virus Prediction Report

Section 1: Business Problem and Background

Problem Statement

The City of Chicago and the Chicago Department of Public Health (CDPH) are seeking an improved method to predict when, and where, West Nile virus outbreaks will occur in mosquitoes. A solution will enable resources to be allocated more efficiently and effectively in the effort to reduce the transmission of West Nile virus.

Background Information

West Nile virus is transmitted to humans via infected mosquitoes. In severe cases, individuals can develop potentially fatal neurological illnesses. After first observing human West Nile virus cases in 2002, the City of Chicago and the CDPH initiated a program in 2004 to monitor mosquito populations and reduce the transmission of West Nile virus. Mosquito traps were established at different locations throughout Chicago and monitored for the presence of West Nile virus. This information was used to determine the location and timing of mosquito spraying efforts. Making use of mosquito trap data, Chicago weather data, and mosquito spray data the City of Chicago and the CDPH would like to improve their ability to predict the presence of West Nile virus in mosquitoes. An improved method of prediction will allow for greater effectiveness, and a more efficient allocation of resources, in the efforts to reduce transmission of West Nile virus.

Success Criteria

A successful solution will:

1. Improve upon the current method used by the City of Chicago and the CDPH, in its ability to predict the presence of West Nile virus in mosquito populations.
2. Be ready to use prior to the upcoming mosquito season (beginning in May).

References:

- https://www.kaggle.com/c/predict-west-nile-virus

Section 2: Data Wrangling

Data Collection

Mosquito trap data, Chicago weather data, and mosquito spray data will be used to develop our West Nile virus prediction model.

The **mosquito trap data** was gathered from mosquito traps that were placed at various locations throughout Chicago. Each trap was tested for the presence of West Nile virus on a weekly basis. The trap location, mosquito count, mosquito species, and presence of West Nile virus within the trap cohort was recorded.

The **Chicago weather data** was procured from two City of Chicago weather stations: Chicago O'Hare International Airport and Chicago Midway International Airport.

The **mosquito spray data** was a record of the City of Chicago's efforts to control mosquito population and West Nile virus transmission using mosquito spray. The date and location of spraying was provided.

References:

- https://www.kaggle.com/c/predict-west-nile-virus

Data Description and Data Cleaning

1. Mosquito Trap Data

The raw mosquito trap data had 10506 rows and included the following features:

Feature	Datatype	Description
Date	object	The date that the mosquito trap was checked
Address	object	The address for the trap location
Species	object	The mosquito species
Block	int64	The block number of the address (ex. 8159 Sample Street would be given a block number of 81)
Street	object	The street that the trap is located on
Trap	object	A unique identifier for each trap
AddressNumberAndStreet	object	The street number and street name corresponding to the trap's address
Latitude	float64	Latitude for the trap location
Longitude	float64	Longitude for the trap location
AddressAccuracy	int64	The location accuracy for the GIS data
NumMosquitos	int64	The number of mosquitoes in the trap cohort. Number of mosquitoes is capped at 50. When the number of trapped mosquitoes exceeds 50, the data is split into multiple records.

WnvPresent	int64	Indicates whether or not West Nile virus was present in the trap cohort

After analysis, the following steps were taken to clean the mosquito trap data:

- The **Date** column was converted to the datetime datatype
- Since latitude/longitude data was available to represent location, the **Address** column did not provide any unique information. Therefore, the column was dropped.
- Since latitude/longitude data was available to represent location, the **Block** column did not provide any unique information. Therefore, the column was dropped.
- Since latitude/longitude data was available to represent location, the **Street** column did not provide any unique information. Therefore, the column was dropped.
- When analyzing the **Trap** column, it was noted that there were 138 unique addresses but only 136 unique traps. Two traps (T035 and T009) were identified that each had two associated addresses. Each address appeared many times and, without further verification, this could not be assumed to be a data entry error. There were no overlapping dates between the two addresses for each trap. Therefore, it is likely that the traps were moved. Since location was considered likely to be a factor in West Nile prevalence, the trap IDs were updated so that there was a unique trap ID for each location. Trap T035 was renamed T0351 and T0352 for each of its associated addresses. Similarly, trap T009 was renamed T0091 and T0092 for each of its associated addresses.
- Since latitude/longitude data was available to represent location, the **AddressNumberAndStreet** column did not provide any unique information. Therefore, the column was dropped.
- Since the Address columns were dropped, the **AddressAccuracy** column was not considered relevant. Therefore, the column was dropped.

2. Chicago Weather Data

The raw Chicago weather data had 2944 rows and included the following features:

Feature	**Datatype**	**Description**
Station	int64	Weather station indicator (1 = Chicago O'Hare International Airport, 2 = Chicago Midway International Airport)
Date	object	The date
Tmax	int64	The daily maximum temperature (Fahrenheit)
Tmin	int64	The daily minimum temperature (Fahrenheit)
Tavg	object	The daily average temperature (Fahrenheit)

Depart	object	The departure from normal temperature
DewPoint	int64	The daily dew point temperature (Fahrenheit)
WetBulb	object	The daily wet bulb temperature (Fahrenheit)
Heat	object	Heating degree days (HDD)
Cool	object	Cooling degree days (CDD)
Sunrise	object	Sunrise time
Sunset	object	Sunset time
CodeSum	object	Codes corresponding to significant weather types
Depth	object	Snow/ice depth on the ground (inches)
Water1	object	Water equivalent of snow/ice on the ground (inches)
SnowFall	object	Total daily snowfall (inches)
PrecipTotal	object	Total daily precipitation (inches)
StnPressure	object	Average station pressure (inches of Hg)
SeaLevel	object	Average sea level pressure (inches of Hg)
ResultSpeed	float64	Resultant wind speed
ResultDir	int64	Resultant wind direction (degrees)
AvgSpeed	object	Average wind speed

After analysis, the following steps were taken to clean the Chicago weather data:

- The **Tavg** column had 11 values that were labeled ‘M’ to represent missing data. However, the average temperature value was derived by calculating the mean of the Tmax and Tmin columns. Therefore, the Tavg values were imputed accordingly.
- The **Depart** column was renamed to **Tdepart**. Tdepart values were not recorded at weather station 2. Therefore, only values from weather station 1 were used. The two stations are in close proximity (roughly 16 miles apart) so this should not be an issue.
- The**DewPoint** column was renamed to **Tdew_point**.
- The **WetBulb** column was renamed to **Twet_bulb**. There were 4 Twet_bulb values labeled ‘M’ to represent missing data. The mean difference between wet bulb values at the two stations was very small and there were no days where both weather stations were missing a value for Twet_bulb. Therefore, the missing value was imputed using the Twet_bulb value from the same day at the other weather station.
- The **Heat** column was renamed to **HeatDegDay**. It had 11 values labeled ‘M’ to represent missing data. Heating Degree Days (HDD) is calculated by subtracting the

average daily temperature from a base temperature of 65 Fahrenheit. Since average temperature data was available, HDD was calculated using this formula.

- The **Cool** column was renamed to **CoolDegDay**. It also had 11 values labeled 'M' to represent missing data. Cooling Degree Days (CDD) is calculated by subtracting a base temperature of 65 Fahrenheit from the average daily temperature. Since average temperature data was available, CDD was calculated using this formula.
- **Sunrise** time was not recorded at weather station 2. Therefore, only values from weather station 1 were used. The two stations are in close proximity (roughly 16 miles apart) so this should not be an issue.
- Similarly, **Sunset** was not recorded at weather station 2. Therefore, only values from weather station 1 were used. The two stations are in close proximity (roughly 16 miles apart) so this should not be an issue.
- First, the **CodeSum** column was renamed to **WeatherCode**. Next, it was discovered that over half of the records did not contain a WeatherCode value. Therefore, a 'NONE' weather code was created and assigned to these records. Many records contained a list of weather codes in the WeatherCode field. Indicator/flag columns were created for each weather code value and the original WeatherCode column was dropped. Finally, most of the weather codes appeared infrequently and would, therefore, have minimal predictive power. Therefore, only weather codes that appeared 50 or more times were kept. In addition to 'NONE', the remaining weather code columns were: BR (mist), DZ (drizzle), HZ (haze), RA (rain), TS (thunderstorm), TSRA (thunderstorm and rain).
- Half of the **SnowIceDepth** values were missing and the other half were 0 (this is expected since we only have weather data for May to October each year). Therefore, the column offered no useful information and was dropped.
- The **Water1** column was missing all its values. Therefore, the column offered no useful information and was dropped.
- All **SnowFall** values were either missing or 0, with the exception of 13 values that were either labeled 'T' (trace amount) or equal to 0.1. This column did not provide enough information to be useful. Therefore, it was dropped.
- The **PrecipTotal** column had 2 values labeled 'M' (missing) and 318 values labeled 'T' (trace amount). The trace amount values were replaced with the mean of 0 and minimum non-zero numeric value. The missing values were imputed using the most frequent value.
- The **StnPressure** column had 4 values labeled 'M' (missing). The values were imputed using the mean StnPressure value for the corresponding weather station.
- The **SeaLevel** column was dropped. The pressure at sea level is not relevant. The station pressure measurement is more representative of the pressure in our locations of interest.
- The **ResultSpeed** column was renamed to **Wind_ResultSpeed**.
- The **ResultDir** column was renamed to **Wind_ResultDir**.
- The **AvgSpeed** column was renamed to **Wind_AvgSpeed**. It had 3 values labeled 'M' (missing). There was a very small average difference in Wind_AvgSpeed between the two weather stations. Therefore, missing values were imputed using the value recorded at the other weather station on the same day.

- The weather station 1 and weather station 2 data was separated and then merged on the date column. This allowed the weather for **each date to be contained in a single row**.

3. Mosquito Spray Data

The raw mosquito spray data had 14835 rows and included the following features:

Feature	Datatype	Description
Date	object	The spray date
Time	object	The spray time
Latitude	float64	Latitude for the spray location
Longitude	float64	Longitude for the spray location

After analysis, the following steps were taken to clean the mosquito spray data:

- The **Date** column was converted to the datetime datatype.
- The **Time** column was missing several values, all on September 7th 2011. The missing values were imputed using the median time from September 7th 2011.

Section 3: Exploratory Data Analysis (EDA)

1. Trap Data

First, a **class imbalance** was identified, as it was observed that roughly 5.2% of the traps were West Nile virus positive. Our approach to dealing with class imbalance will be discussed below.

The correlations between the explanatory variables and the target variable (WnvPresent) in the mosquito trap data were examined. There was a weak correlation (Pearson correlation coefficient = 0.2) between the number of mosquitoes in a trap and the presence of West Nile virus. Further analysis indicated that the mean number of mosquitoes in a West Nile positive trap was roughly double the mean number of mosquitoes in a West Nile negative trap (see **Figure 1**). This indicated that predictors of **mosquito population numbers** would likely play an important role in predicting when, and where, West Nile virus outbreaks are likely to occur in mosquito populations.

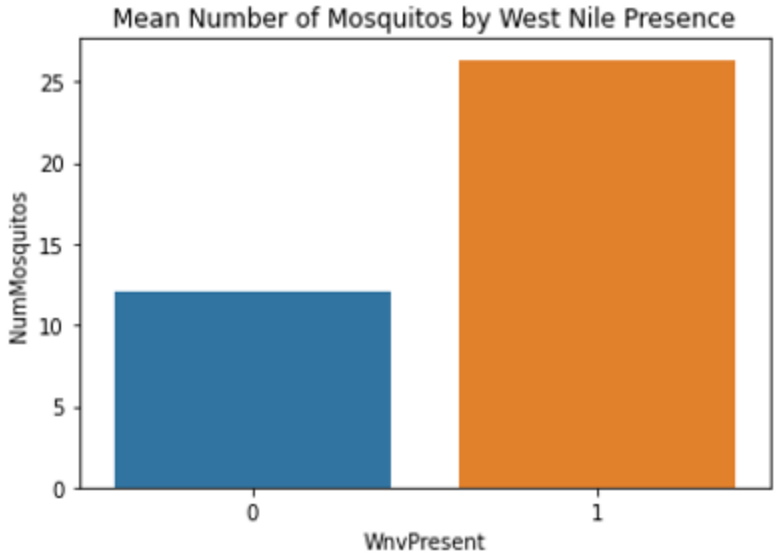


Figure 1: Mean mosquito count in West Nile virus positive and negative traps.

The location of traps in relation to the presence of West Nile virus was another factor that was important to examine. An examination of trap locations with a high (75th percentile) vs. low West Nile proportion, with an extreme West Nile proportion (0.1 or higher), and with no incidences of West Nile virus allowed three zones to be identified: West Nile virus appeared more likely to occur in the **Northwest Zone**, while it appeared less likely to occur in the **Northeast Zone** and **Downtown Core Zone** (see **Figure 2**) .

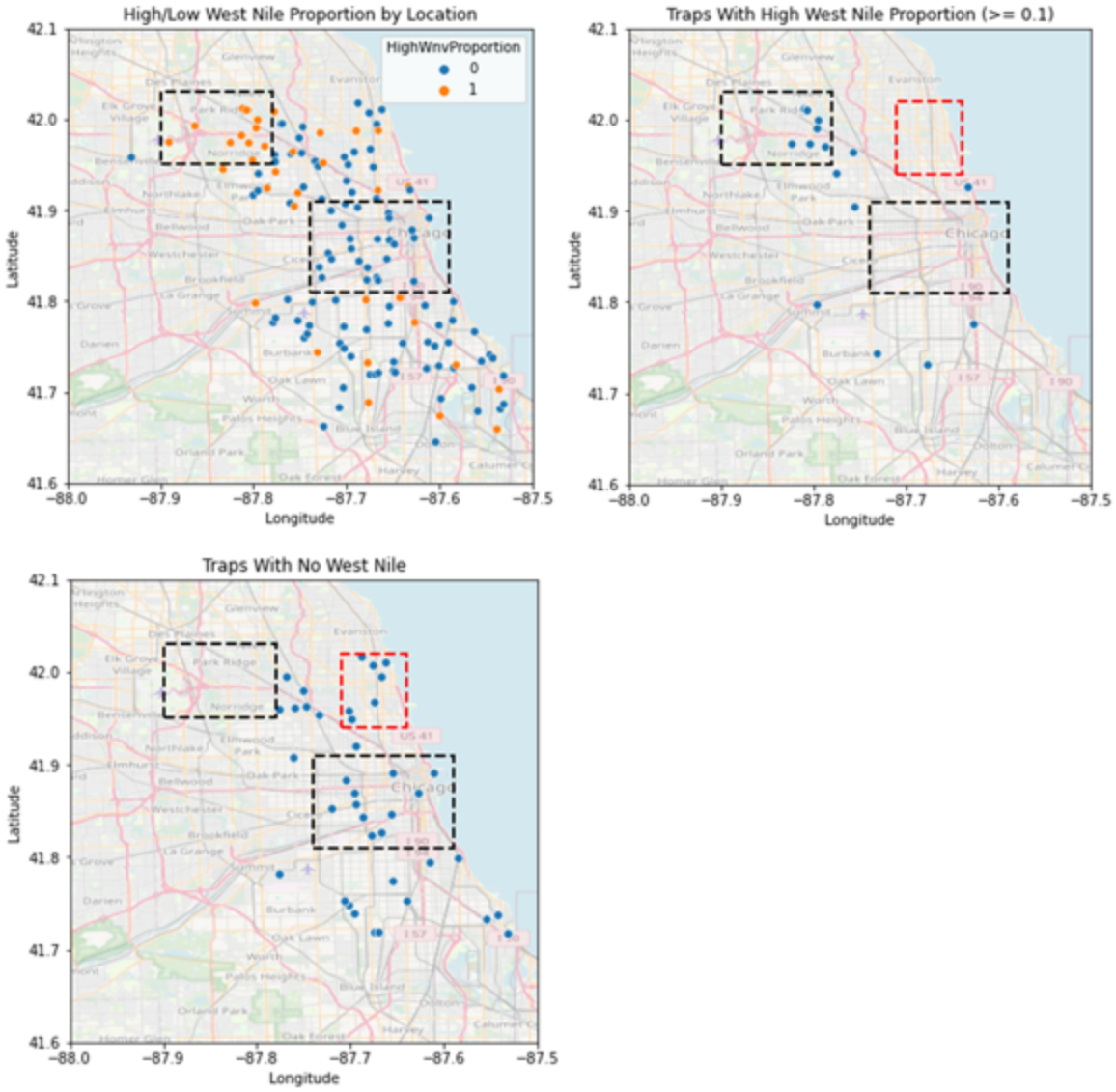


Figure 2: West Nile virus by location: traps above/below the 75th percentile for West Nile virus proportion (upper left), traps with a West Nile virus proportion greater than, or equal to, 0.1 (upper right), and traps with 0 incidences of West Nile (bottom).

Inspecting **annual variations**, it was discovered that both mosquito population numbers and the proportion of traps containing West Nile virus were much higher in 2007 and 2013, compared to 2009 and 2011 (see **Figure 3**). This was likely due to differences in weather conditions (which will be discussed below). Next, **seasonality** was examined. West Nile virus tended to first appear in mosquito traps in mid-July, so July 15th was considered the beginning of West Nile season. August was by far the worst month for West Nile virus and there was also a high West Nile virus presence in September (see **Figure 4**). For a finer grained look, we also examined West Nile presence by week of the year. Peak West Nile season tended to occur between weeks 31 and 37.

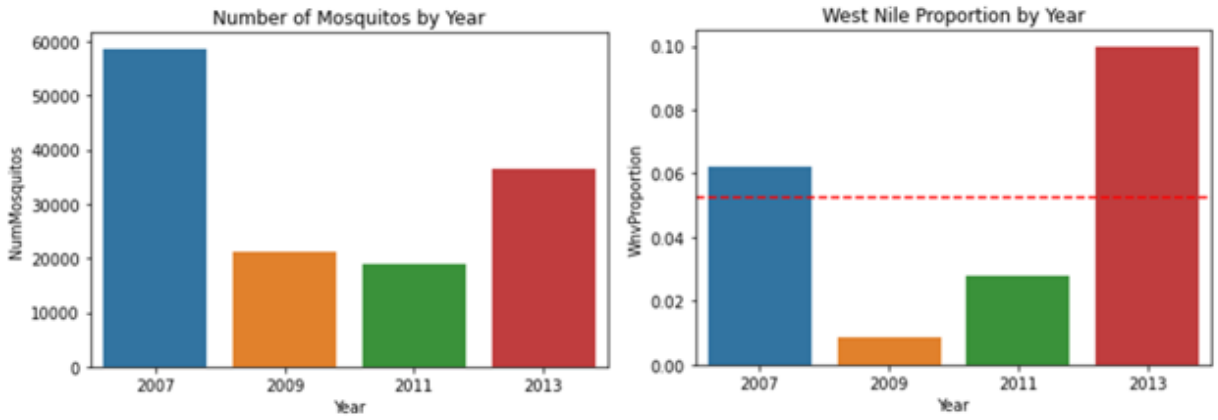


Figure 3: Annual variation in mosquito population numbers (left) and West Nile virus proportion (right). The dashed red line represents the dataset's overall West Nile virus proportion.

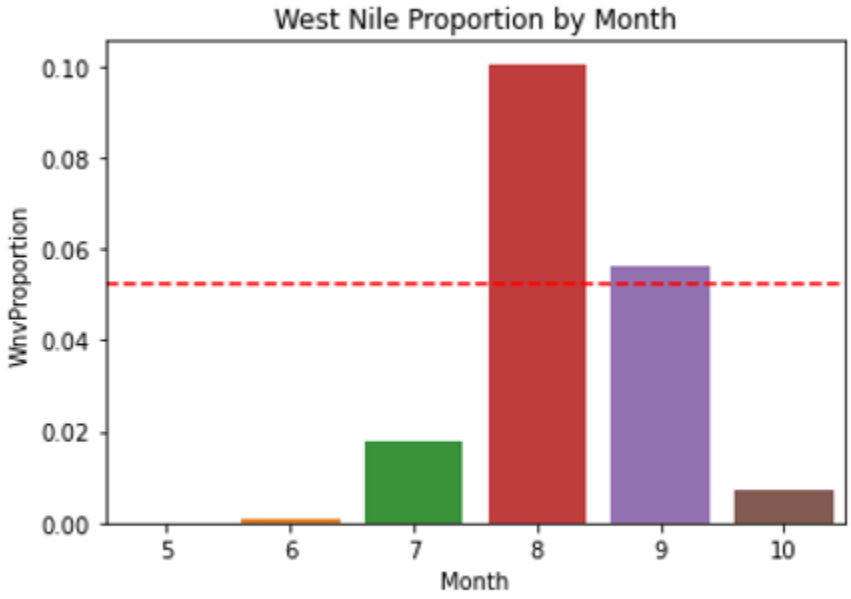


Figure 4: West Nile virus proportion by month. The dashed red line represents the dataset's overall West Nile virus proportion.

There were 6 unique **species** of mosquitoes trapped. However, West Nile virus was not found in 4 of the species (see **Figure 5**). It is not known whether this was due to the species of mosquito or due to the fact that very few mosquitoes of these species were caught, resulting in a small sample size. These species were combined into a single category labeled "Other". West Nile virus was found in the Culex Pipiens and Culex Restuans mosquito species, and was more likely to occur in Culex Pipiens (see **Figure 6**).

Species	NumMosquitos	WnvPresent
CULEX ERRATICUS	7	0
CULEX TARSALIS	7	0
CULEX SALINARIUS	145	0
CULEX TERRITANS	510	0
CULEX RESTUANS	23431	1
CULEX PIPIENS	44671	1
CULEX PIPIENS/RESTUANS	66268	1

Figure 5: Number of trapped mosquitoes and West Nile virus presence by species.

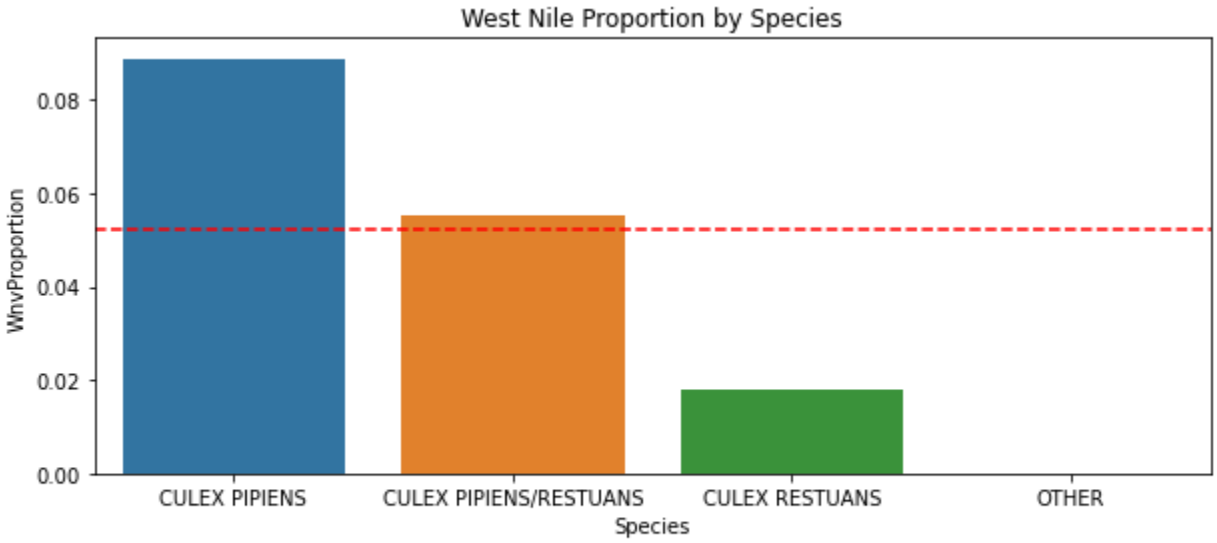


Figure 6: West Nile virus proportion breakdown by species. The dashed red line represents the dataset's overall West Nile virus proportion.

2. Mosquito Spray Data

When first examining spray data on a Chicago map, **outlier** spray locations were identified that were not in close proximity to any of the mosquito traps. These outliers were removed (see **Figure 7**).

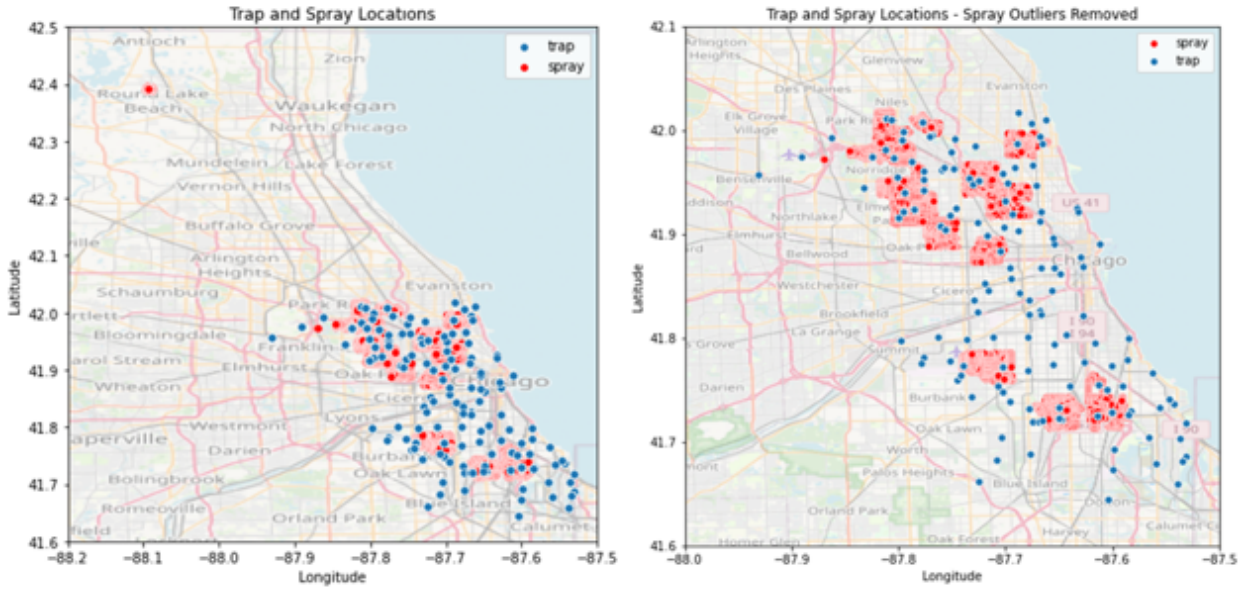


Figure 7: Mosquito trap and mosquito spray locations before (left) and after (right) removing spray outliers.

Next, the impact of **spray on mosquito population** numbers was examined. It appeared that spraying impacted not only the spray zones but also the areas between spray zones (see **Figure 8**). This is likely because mosquitoes do not remain in a single location and may be impacted if they fly in and out of spray zones.

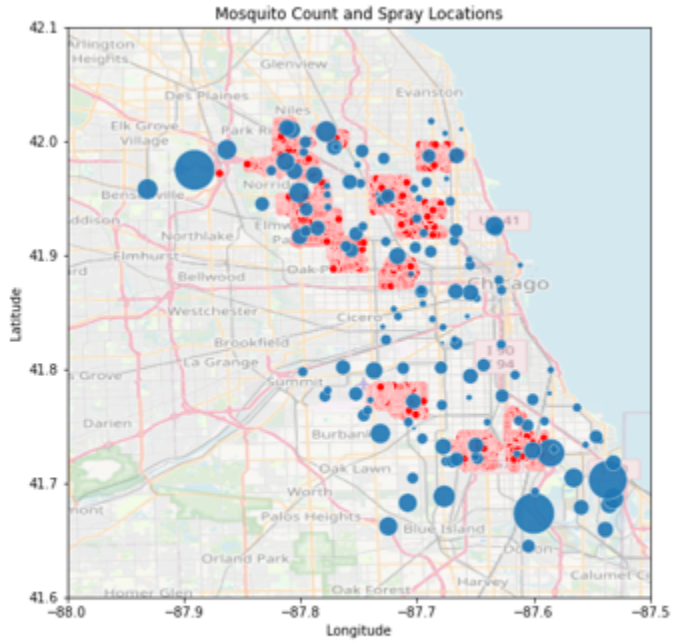


Figure 8: Spray locations and mosquito counts. The size of the blue circles represents the number of mosquitoes. The red areas are mosquito spray zones.

3. Chicago Weather Data

Each mosquito trap was mapped to the nearest weather station (see **Figure 9**). This way, when combining the weather and trap data, the weather at each trap was as accurate as possible.

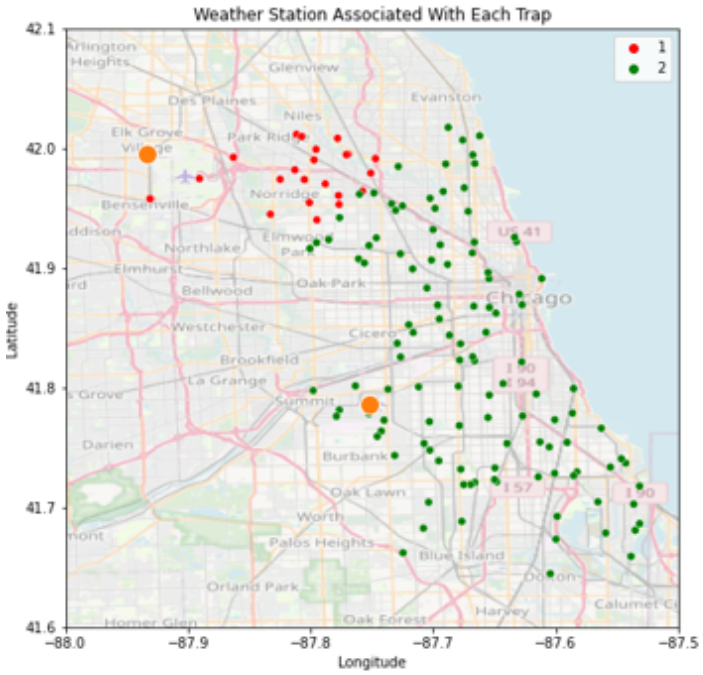


Figure 9: Traps were mapped to their nearest weather station. Orange circles represent weather stations. The red dots represent traps closer to station 1 while green dots are the traps that were closer to station 2.

4. Combined Analysis (Trap, Spray, and Weather Data)

a. Trap and Weather Data

Several key observations were made when analyzing the weather and trap data together. West Nile virus presence increased sharply when **temperatures** were at, or above, 70F (see **Figure 9**). In terms of **total daily precipitation**, a range of 1.25-1.5 inches was associated with a high proportion of West Nile virus cases. When precipitation was below this range, the proportion of West Nile virus cases was moderate. When precipitation was above this range, West Nile virus was not found in any of the traps. West Nile virus proportion increased steadily with **total weekly precipitation**, up to a level of 5 inches (see **Figure 9**). West Nile virus was not present above this level. Since moisture levels were also expected to be relevant to mosquito population levels, the number of **days since precipitation** was also considered. West Nile virus levels tended to be highest 7 days after precipitation, moderate when less than 7 days had passed since precipitation, and low when there had been no precipitation for more

than 7 days. **Wind Speeds** of 4-6 mph were observed when West Nile virus levels were greatest. There were no observed West Nile virus cases when wind speeds were below 2 mph or above 14 mph.

As mentioned above, mosquito population numbers and the proportion of traps containing West Nile virus were much higher in 2007 and 2013, compared to 2009 and 2011. This led to an examination of **annual weather variation**. Temperatures were lower in 2009 than the other years. In 2011, precipitation levels were lower than the other years.

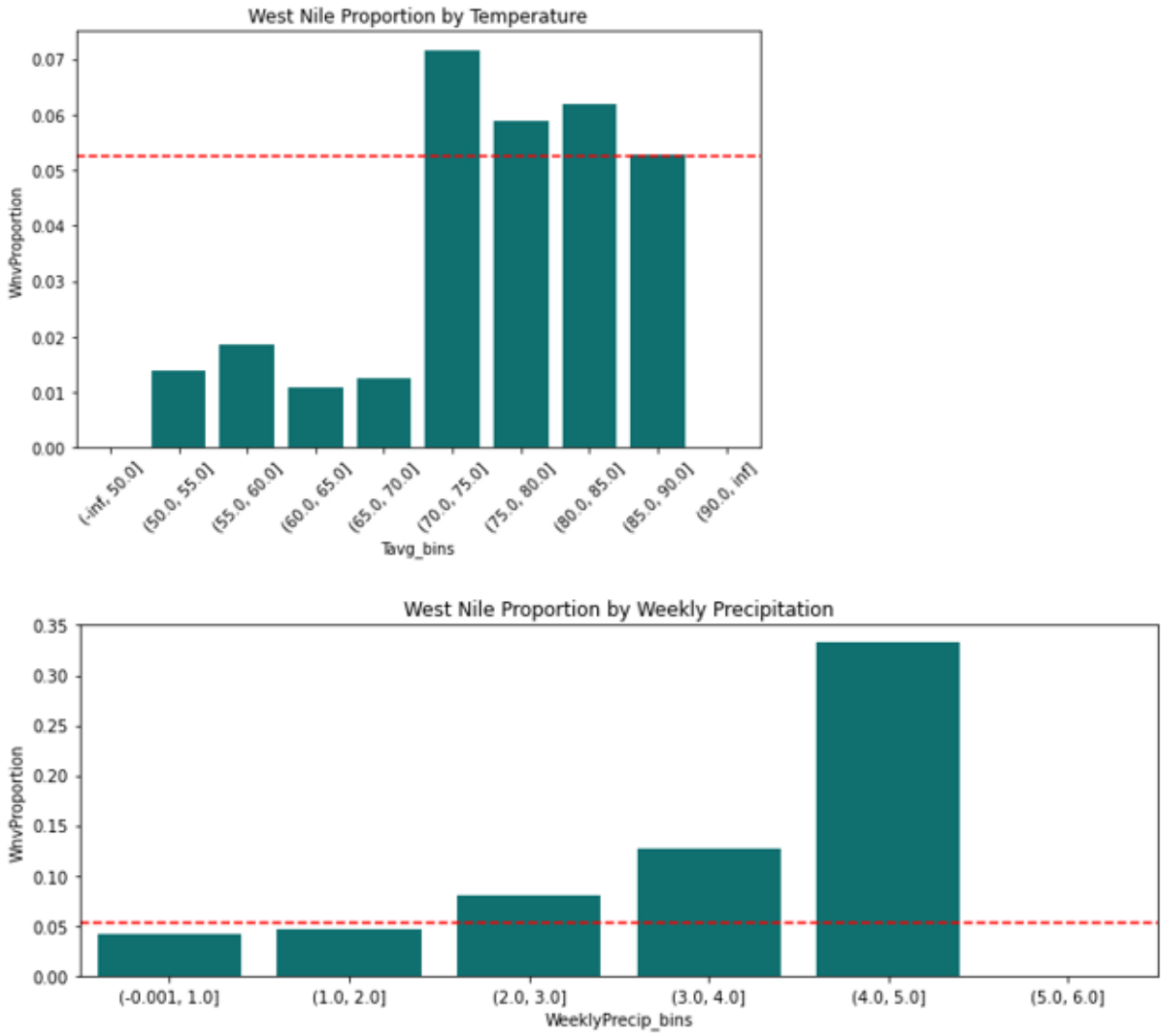


Figure 9: West Nile virus proportion by Temperature (top) and Weekly Precipitation (bottom). The dashed red line represents the dataset's overall West Nile virus proportion.

b. Trap and Spray Data

Looking at trends in mosquito numbers **before and after spraying**, there did not seem to be a predictable impact of spraying on mosquito numbers. This may have been due to the

fact that traps were not checked daily. The spray only killed adult mosquitoes - not eggs or larvae. It is possible that by the time traps were checked, there was time for eggs to hatch and larvae to mature, allowing mosquito populations to recover. However, the West Nile virus proportion did appear to be lower after spraying. It is possible that, by the time traps were checked, there had been time for mosquito numbers to regenerate but not enough time for West Nile virus to propagate through the mosquito population.

The **days since spray** and a trap's **proximity to a spray zone** was also examined. West Nile virus proportions tended to be higher close to spray zones and 0-1 days after spraying. This is likely a case of correlation, but not causation, because the City of Chicago is targeting areas that already have high West Nile virus proportions. This is why developing a predictive method that allows proactive, rather than reactive, actions to be taken should be beneficial. It should be noted that when there were 0 days since spraying, the traps may have been checked prior to spraying (trap checking times were not provided).

5. Feature Selection and Feature Engineering

After analysis, several original features were selected to be used as model inputs. New features were also engineered to provide the model with additional, potentially important, inputs. The selected and engineered features are outlined below (a description is provided for engineered features):

- **Mosquito Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Species	Original	
NumMosquitos	Original	
WnvPresent	Original	

- **Location Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
IsDownTownCore	Engineered	Downtown Chicago area with low West Nile virus proportion (0 = False, 1 = True)
IsNorthWestZone	Engineered	Area to the northwest with a high proportion of West Nile virus (0 = False, 1 = True)
IsNorthEastZone	Engineered	Area to the northeast with a high proportion of West Nile virus (0 = False, 1= True)

TrapZone	Engineered	The Chicago map was broken into a grid of zones: each zone is 0.5 degrees in latitude and 0.5 degrees in longitude. TrapZone indicates which of these zones the trap is located within.

- **Seasonality Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Month	Engineered	The month of the year
Week	Engineered	The week of the year
IsInSeason	Engineered	Indicates that West Nile season (July 15th or later) has begun (0 = False, 1 = True)
IsPeakSeason	Engineered	Indicates that the week was between 31 and 37 (inclusive), indicating peak West Nile season (0 = False, 1 = True)
DaylightMinutes	Engineered	The total minutes of daylight on that date

- **Weather Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Tmax	Original	
Tmin	Original	
Tavg	Original	
Tdepart	Original	
Tdew_point	Original	
Twet_bulb	Original	
HeatDegDays	Original	
CoolDegDays	Original	
PrecipTotal	Original	
StnPressure	Original	

Wind_AvgSpeed	Original	
Wind_ResultDir	Original	
WeatherCode_BR	Engineered	Weather code indicating mist (0 = False, 1 = True)
WeatherCode_DZ	Engineered	Weather code indicating drizzle (0 = False, 1 = True)
WeatherCode_HZ	Engineered	Weather code indicating haze (0 = False, 1 = True)
WeatherCode_RA	Engineered	Weather code indicating rain (0 = False, 1 = True)
WeatherCode_TS	Engineered	Weather code indicating thunderstorm (0 = False, 1 = True)
WeatherCode_TSRA	Engineered	Weather code indicating thunderstorm and rain (0 = False, 1 = True)
WeatherCode_NONE	Engineered	Indicates no weather was present (0 = False, 1 = True)
DaysSincePrecip	Engineered	Number of days since the last precipitation event
PrecipWeekly	Engineered	Total precipitation for the week
IsOptimalTemp	Engineered	Indicates temperature is greater than or equal to 70F (0 = False, 1 = True)
PrecipConditions	Engineered	Categorizes daily precipitation conditions into 3 categories (based on West Nile virus proportion): ● Optimal (1.25 <= daily precipitation <= 1.5) ● Moderate (daily precipitation < 1.25) ● Poor (daily precipitation > 1.5)
PrecipWeekly_Score	Engineered	Assigns a score, from 1-5, based on weekly precipitation (1 is most likely to find West Nile, 5 is least likely)
MoistureConditions	Engineered	Categorizes the days since the last precipitation event into 3 categories (based on West Nile virus proportion): ● Optimal (DaysSincePrecip = 7) ● Moderate (DaysSincePrecip < 7) ● Poor (DaysSincePrecip >7)
WindConditions	Engineered	WindConditions**:** categories average daily wind speed into 3 categories (based on West Nile virus proportion): ● Optimal (4 < Wind_AvgSpeed <= 6)

		• Poor (Wind_AvgSpeed <= 2 or Wind_AvgSpeed > 14) • Moderate (Wind_AvgSpeed is not Poor or Optimal)
Rel Humidity	Engineered	The relative humidity (calculated using average temperature and dewpoint temperature)

- **Rolling Average Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Tavg_**n**days	Engineered	Rolling average daily temperature (7, 14, 21, and 28 days)
Wind_Avgspeed_**n**days	Engineered	Rolling average daily wind speed (7, 14, 21, and 28 days)
PrecipTotal_**n**days	Engineered	Rolling average precipitation (7, 14, 21, and 28 days)
RelHumidity_**n**days	Engineered	Rolling average relative humidity (7, 14, 21, and 28 days)

Note: for each rolling average feature name, replace “n” with the number of days

- **Time Lag Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Tavg_lag**n**	Engineered	Time lagged average temperature (1-28 days back)
Wind_AvgSpeed_lag**n**	Engineered	Time lagged average wind speed (1-28 days back)
PrecipTotal_lag**n**	Engineered	Time lagged precipitation (1-28 days back)
DaylightMinutes_lag**n**	Engineered	Time lagged daylight minutes (1-28 days back)
RelHumidity_lag**n**	Engineered	Time lagged relative humidity (1-28 days back)

Note: for each time lag feature name, replace “n” with the number of days

- **Time Lag Features - Mean Lagged Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Tavg_lag**n**_mean	Engineered	Time lagged mean temperature (7-14 days back, 14-21 days back, 21-28 days back, 28-35 days back)
Wind_AvgSpeed_lag**n**_mean	Engineered	Time lagged mean wind speed (7-14 days back, 14-21 days back, 21-28 days back, 28-35 days back)
PrecipTotal_lag**n**_mean	Engineered	Time lagged mean precipitation (7-14 days back, 14-21 days back, 21-28 days back, 28-35 days back)
DaylightMinutes_lag**n**_mean	Engineered	Time lagged mean daylight minutes (7-14 days back, 14-21 days back, 21-28 days back, 28-35 days back)
RelHumidity_lag**n**_mean	Engineered	Time lagged mean relative humidity (7-14 days back, 14-21 days back, 21-28 days back, 28-35 days back)

Note: for each mean lagged feature, replace “n” with the lower bound on the number of days. For example, replacing “n” with 7 would give the name of the feature for the 7-14 days back time period.

- **Spray Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
IsSprayed	Engineered	Indicates the trap has been sprayed within the last 30 days (0 = False, 1 = True)
TrapSprayDistance	Engineered	The trap’s distance from mosquito spray on the most recent spray date
DaysSinceSpray	Engineered	The number of days since the last spray date
InSprayBounds	Engineered	Indicates that the trap is within the max/min longitude/latitude boundaries

		of all spray zones, up to that point in the year (0 = False, 1 = True)
IsNearSprayZone	Engineered	TrapSprayDistance is less than, or equal to, 0.2 (0 = False, 1 = True)
IsRecentlySprayed	Engineered	Indicates that it has been 1 day, or less, since the most recent spray date (0 = False, 1 = True)
IsDayAfterSpray	Engineered	Indicates that spraying was done 1 day ago and the trap was near the spray zone (0 = False, 1 = True)

- **Trap Features:**

Feature	**Original or Engineered**	**Description (if Engineered)**
Prev_Check	Original	Number of days since the trap was last checked

Section 4: Preprocessing

Train/Test Split

For modeling purposes, the data was split into training and test sets. An allocation of 70% training data and 30% test data was used. The dataset had imbalanced classes, as roughly 5.2% of the traps were West Nile virus positive. Therefore, **stratified sampling** was used when performing the train/test split in order to preserve this ratio in both the training and test sets.

Feature Reduction

After feature selection and feature engineering there were 223 features in the dataset. **Weight of Evidence (WOE)** was used as a feature reduction method. WOE is a technique for calculating the predictive power of an explanatory variable, given a specific dependent variable. **Information Value (IV)** is calculated using WOE values. IV is a useful way to interpret WOE values, and was used to select features based on their predictive power. Features with an IV between 0.05 and 0.8 (inclusive) were selected, while all others were eliminated. An additional

benefit of this form of feature reduction is that, unlike some methods (ex. PCA), we do not lose feature interpretability. After this process, 105 features remained. Further feature reduction was carried out by removing features with multicollinearity. This process will be discussed below.

References:

- https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

Categorical Feature Encoding

One Hot Encoding was used to transform categorical features into a numerical format appropriate for modeling. The first encoded column for each categorical feature was dropped in order to remove redundant information.

Scaling of Continuous Features

Standard scaling was performed on all continuous features.This ensures that none of the features will be more or less influential based on their original value range.

Multicollinearity

Feature redundancy was minimized by removing features with multicollinearity. This was a form of further feature reduction that helped to simplify the model inputs, while minimizing loss of valuable information, which should result in a more robust model. In order to remove features with multicollinearity, the **Variance Inflation Factor (VIF)** was calculated for each feature and the feature with the maximum VIF value was dropped. This process was repeated until all remaining features had a VIF value less than 2. Similar to the WOE/IV method of feature reduction (discussed above), this feature reduction approach has the benefit of maintaining feature interpretability. After this process was complete, 44 features remained. The final set of features used as model inputs is outlined below.

References:

- https://towardsdatascience.com/everything-you-need-to-know-about-multicollinearity-2f21f082d6dc

Final Set of Features

After feature reduction (using the WOE/IV and Multicollinearity techniques discussed above) the final list of features selected for modeling was the following:

Feature
InSprayBounds_1
IsDayAfterSpray_1

IsNorthWestZone_1
MoistureConditions_Optimal
MoistureConditions_Poor
NumMosquitoes
PrecipTotal_lag1
PrecipTotal_lag23
PrecipTotal_lag27
PrecipTotal_lag28_mean
PrecipWeekly_Score_3
PrecipWeekly_Score_5
RelHumidity_lag10
RelHumidity_lag4
Species_CULEX RESTUANS
Species_OTHER
Tavg_lag11
Tavg_lag23
Week_24
Week_25
Week_26
Week_27
Week_28
Week_29
Week_30
Week_31
Week_32
Week_33
Week_34

Week_35
Week_36
Week_37
Week_38
Week_39
Week_40
Week_41
WindConditions_Optimal
WindConditions_Poor
Wind_AvgSpeed_lag15
Wind_AvgSpeed_lag18
Wind_AvgSpeed_lag23
Wind_AvgSpeed_lag5
Wind_AvgSpeed_lag6
Wind_AvgSpeed_lag7_mean

Section 5: Modeling

Class Imbalance in the Dataset

The dataset had highly imbalanced classes, as roughly 5.2% of the traps were West Nile virus positive with all remaining traps being West Nile virus negative. The first step to dealing with the class imbalance was using **stratified sampling** when performing the train/test split (discussed above). The next step was to use **undersampling** of the majority class (the West Nile virus negative class) in order to create balanced classes. Undersampling was preferred to oversampling methods because it does not require the generation of synthetic data, which may reduce the model's ability to generalize to unseen data.

Model Evaluation

In order to perform model evaluation, the **AUC-ROC Curve (AUC).** ROC is effective at assessing model performance for binary classification problems. The AUC assesses the model's ability to differentiate between the positive and negative class.

References:

- https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Modeling Options

The following **8 machine learning models** were selected, trained, and used to make predictions on the test data:

- Logistic Regression
- Support Vector Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- LightGBM Classifier
- AdaBoost Classifier
- Extra Trees Classifier

Additionally, a Dummy Classifier was used to establish a **baseline model**. The Dummy Classifier makes predictions by randomly drawing a class from a uniform distribution of the unique classes.

Model Selection

After initial modeling using the default hyperparameters for each of the 8 models, the test **AUC scores** (in descending order) were as follows:

Model	TrainAUC	TestAUC
Gradient Boosting Classifier	0.906843	0.838901
Logistic Regression	0.868972	0.835116
AdaBoost Classifier	0.879392	0.834506
Support Vector Classifier	0.880611	0.823008
LightGBM Classifier	0.916291	0.813316
XGBoost Classifier	0.914583	0.807397
Random Forest Classifier	0.928445	0.797384
Extra Trees Classifier	0.942395	0.767761
Baseline Model	0.500000	0.500000

Since the **Gradient Boosting Classifier** had the best initial performance, in terms of AUC score on the test set, it was selected as the starting point to build a final model.

HyperParameter Tuning

In an attempt to improve upon the Gradient Boosting Classifier's initial performance, hyperparameter tuning was performed. The following hyperparameters were tested using a **Randomized Search** with 15-fold cross validation and 30 iterations:

Hyperparameter	**Test Values**
learning_rate	0.001, 0.01, 0.1, 1
max_depth	3, 6, 7, 8, 9, 10
subsample	0.25, 0.5, 0.75, 1
max_features	'sqrt', 'log2', 1

After hyperparameter tuning, the model's **AUC score improved to 0.8425** (see **Figure 10**). This represented a 0.3425 improvement over the baseline model. The **best parameters** for the model were:

Hyperparameter	**Best Value**
learning_rate	0.1
max_depth	3
subsample	0.75
max_features	'log2'

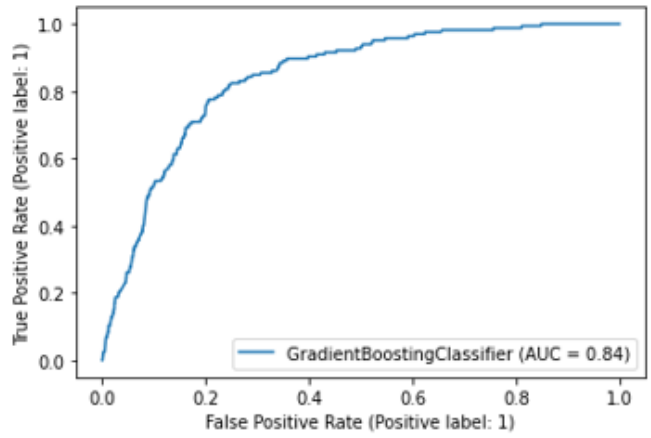


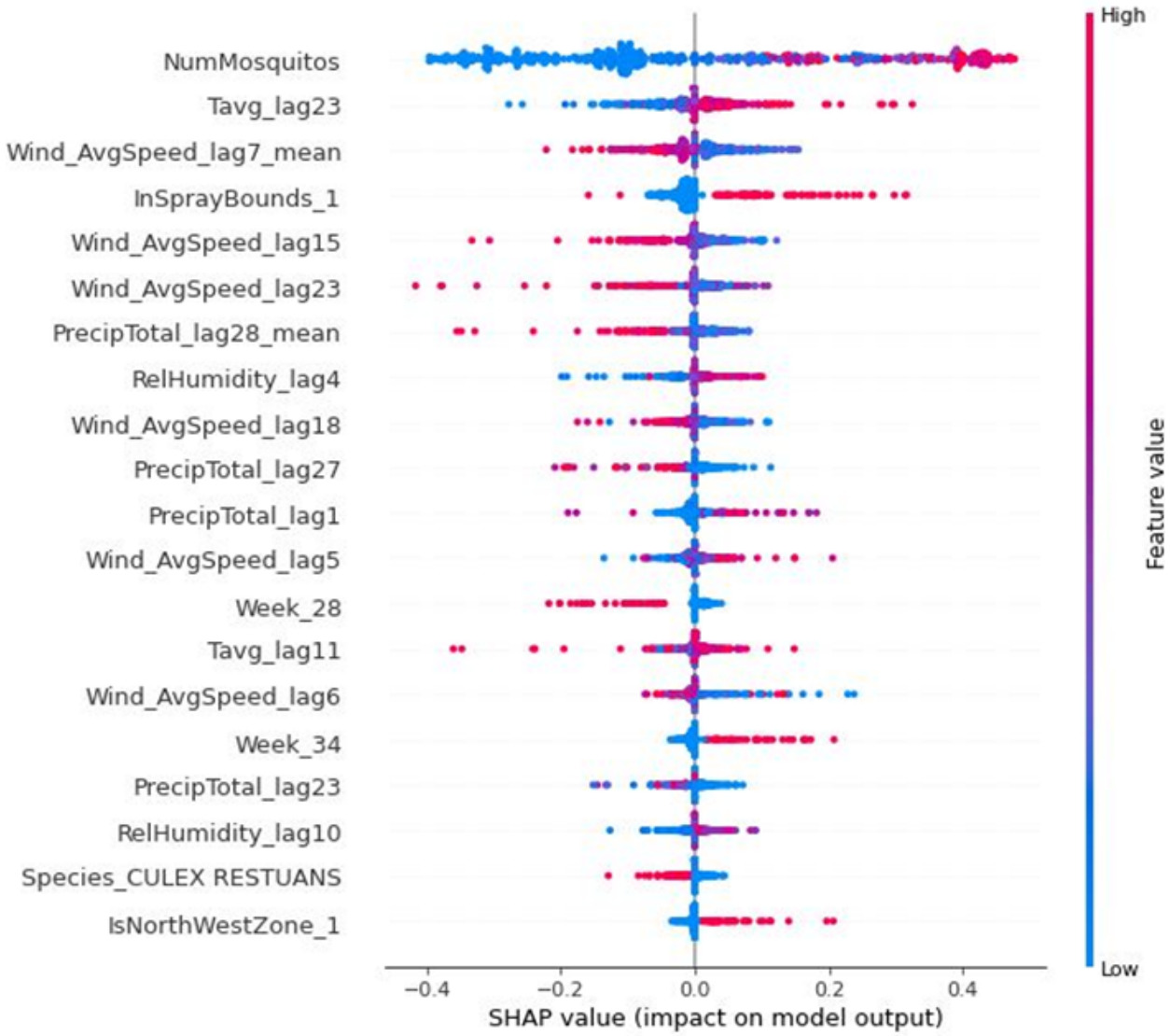
Figure 10: ROC Curve and AUC Score for the final model.

Therefore, a Gradient Boosting Classifier with the above parameters was selected as the **final model**.

Section 6: Conclusions

SHAP Analysis

SHAP analysis is a technique that is useful for analyzing feature importance. SHAP analysis uses Shapley values to identify the contribution that each feature made to the model's prediction. The SHAP analysis output is shown in the following figure:



In the SHAP analysis output, the top 20 features are listed in descending order of importance. The x-axis indicates whether the feature's value is associated with a higher or lower prediction, as well as the magnitude of the effect. Red dots indicate that the feature's value was high for the given record, while blue dots indicate a low value.

Based on the SHAP analysis, the following conclusions were drawn about feature importance:

1. Top 5 Features

- The **number of mosquitoes** caught in a trap was the most important predictor of whether West Nile virus would be present. This makes sense because, even when West Nile conditions are optimal, there is not necessarily a high likelihood that any individual mosquito will be West Nile positive. As more mosquitos are caught in a trap, it becomes more likely that at least one mosquito will be West Nile positive.
- As the **average daily temperature with a 23-day lag** increased, it was more likely that West Nile virus would be present. Mosquitos tend to thrive in warmer environments so this was not surprising. The warmer temperatures likely led to improved mosquito breeding conditions and, therefore, higher mosquito population numbers. However, the effect will not be seen immediately. The length of the mosquito life cycle tends to be 2-4 weeks (although it can be shorter depending on weather conditions). This is a likely explanation for the importance of the 23-day lag.
- As the **mean wind speed from 7-14 days prior** increased, the presence of West Nile virus decreased. Higher wind speeds likely make it more difficult for mosquitos to fly, reducing both their ability to fly greater distances and to accurately fly to their target destination. Therefore, higher wind speeds may result in reduced mosquito breeding. The importance of the 7-14 day lag is likely due to the length of the mosquito life cycle which tends to be 2-4 weeks, but can be as short as one week in peak summer conditions.
- Traps were considered **within the spray boundaries** when they were inside the box created by the northernmost, southernmost, easternmost, westernmost spray location (up to that point in the year). The trap locations within the spray boundaries tended to have a much higher West Nile presence. This is likely due to correlation rather than causation. Spraying may have been a reactive response in areas that were already known to have either high levels of West Nile presence or high mosquito counts. It makes sense that spray boundaries, not only the spray zones, were important because individual mosquitos are not necessarily confined to a small area. Mosquitos have a maximum flight distance of 50m to 50km, depending on the species. Therefore, if there is a high mosquito count in one area, it is likely that there will be a high mosquito count in neighbouring areas (assuming relatively similar environments). This should hold true for West Nile presence as well.
- As **average wind speed with a 15-day lag** increased, the presence of West Nile virus decreased. As mentioned above, higher wind speeds likely reduce both the distance mosquitos are able to fly and the ability of a mosquito to accurately fly to a target destination. This can be expected to result in reduced mosquito breeding. A delayed impact on mosquito population numbers would be observed, since the mosquito life cycle takes 2-4 weeks. Again, this explains the importance of the 15 day lag.

2. Other Observations Related to Feature Importance

- A **23-day lag** seemed to be especially significant. Temperature and average wind speed lagged by 23 days both showed up in the top 6 most important features. Total precipitation lagged by 23 days also showed some importance. The typical mosquito life cycle is 2-4 weeks and varies based on environmental factors including temperature and moisture levels. It is possible that a life cycle of roughly 23 days tends to be common in the conditions created by the Chicago climate.
- Higher **relative humidity** tended to go along with higher West Nile virus presence. Importance in relative humidity was seen for both a 4-day and 10-day lag. With the 4-day lag time, it is likely that mosquito activity increased with humidity, resulting in a higher number of mosquitoes being caught in the traps. Since traps were not checked daily, mosquitos may be caught in a trap for a few days before being discovered. With the 10-day lag time, it is likely that increased humidity improved breeding conditions for mosquitos. The mosquito life cycle is shorter in peak summer heat (which is often accompanied by high humidity levels). This may explain why a 10-day lag was important with relative humidity rather than the longer lag times associated with other weather features.
- The mosquito **species** showed some importance in the model. Only two species - Culex Restuans and Culex Pipiens - were caught in significant numbers. These were also the only two species that were observed to be West Nile virus positive. **Culex Restuans** mosquitoes were less likely to test positive for West Nile virus.
- Among **location** based features, the **Northwest Zone** identified during the EDA phase was positively correlated with the presence of West Nile. It is possible that this area provides an optimal environment for either mosquitoes or for bird species that are susceptible to West Nile virus (since mosquitoes become West Nile virus carriers by biting birds infected with the disease).
- **Seasonality** also played a role in West Nile virus prediction, as week of the year indicators appeared in the top 20 most important features. Weeks 28 and 34 had negative and positive correlations with the presence of West Nile virus respectively.
- **Precipitation** had a varied effect, but the two most important precipitation features had long lag times (mean precipitation with a lag of 28-35 days, and precipitation with a lag of 27 days). In these two cases, precipitation was negatively correlated with West Nile virus. This may be due to the fact that water sources become more scarce in dry conditions. Culex species tend to disperse more widely under these conditions, as adequate breeding grounds become more scarce. Additionally, contact between birds and mosquitoes tends to increase in dry conditions, as they both rely on the few remaining water sources. This increases the likelihood that mosquitoes will bite a bird that is infected with West Nile virus. On the other hand, mosquito numbers tend to increase in rainy conditions. This may explain why precipitation can have a varied impact on the presence of West Nile virus.

References

- https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernel explainer-79de9464897a
- https://www.in.gov/health/erc/zoonotic-and-vectorborne-epidemiology-entomology/pests/culex-species-mosquitoes/
- https://www.sciencedirect.com/science/article/pii/S0075951113001011
- https://www.orkin.com/pests/mosquitoes/mosquito-life-cycle-facts
- https://www.medicalnewstoday.com/articles/west-nile-virus-in-the-us-a-case-study-on-climate-change-and-health#The-West-Nile-virus-in-the-U.S.

Recommendations to the City of Chicago and CDPH

A trained Gradient Boosting Classifier will be provided to the City of Chicago and CDPH, which will provide an improved ability to predict where, and when, West Nile virus will occur. Weather data and mosquito population trends can be monitored, and the Gradient Boosting Classifier can be used to model different scenarios to proactively determine if, and when, different areas are at risk for the presence of West Nile virus. The following are recommendations on how the City of Chicago and CDPH can make use of this capability:

1. Mosquito Spraying

Mosquito spraying can be done in locations where West Nile virus is predicted to be present. The model's predictive ability should allow for mosquito spraying efforts to be more effective. Spraying can be done proactively when conditions are ideal for West Nile virus. Therefore, mosquito population numbers can be reduced prior to the occurrence of a West Nile virus outbreak. This should be more effective in limiting the West Nile virus risk to the human population than the current reactive approach to spraying.

2. Public Awareness

Modeling can also be used to inform the public of West Nile virus risks at different times and locations throughout mosquito season. The predictive ability of the model will allow the public to be given advance notice of the risks, giving individuals time to plan accordingly.

3. Insights from EDA and Feature Importance Analysis

In addition to making use of the model, several important insights can be taken from the EDA and Feature Importance Analysis stages of the exploration. West Nile virus season tends to begin around July 15th, and the presence of West Nile virus is at its peak in August. The risks are especially high in areas with a high concentration of Culex Pipiens mosquitoes. In peak West Nile virus season it should be noted that the Chicago downtown core appeared to be especially safe, while West Nile virus presence in the Northwest Zone (identified during EDA) was very high. In terms of weather, high temperatures, high humidity, and low wind speeds

appeared to create favourable conditions for the appearance of West Nile virus 1-3 weeks later. The impact of precipitation was more varied.

Recommendations for Future Investigation

- Incorporating **bird population** data into a model may be beneficial. Mosquitoes become West Nile virus carriers by biting birds infected with the disease. Analyzing susceptible bird species, bird migration patterns, bird habitat information, and West Nile virus presence in birds may provide additional value.
- An analysis of Chicago **geography/terrain** may be valuable. This could provide a finer grained understanding of which locations provide favourable habitats for mosquitoes. Identifying which areas are forested, grasslands, swampy, or urban could be beneficial for understanding and modeling.
- **Checking mosquito traps daily** may provide a more accurate picture of mosquito population and West Nile virus trends. However, City of Chicago resource allocation will be a factor here.
- Adoption of the model should lead to more **proactive spraying**. Analysis of these efforts may allow for a better understanding of the impact of spraying.
- Weather data was only provided for May through October. Analyzing **annual weather** data may be beneficial in determining the impact of winter and early spring weather conditions on mosquito numbers and/or West Nile virus presence later in the year.
- The model found that **mosquito population numbers** were an important West Nile virus predictor. Therefore, it may be worth doing further research into factors that impact mosquito populations. It may be possible to collect data on these factors and incorporate them into a future model.