**West Nile Virus Prediction**

When and where will mosquitos test positive for West Nile Virus in the City of Chicago?

**Mayra Chavez**

Springboard Data Science Career Track

February 2021 Cohort

Mayract84@gmail.com

**Abstract**

The basic aim of this project is to provide a machine learning model to predict when and where the West Nile Virus will test positive in mosquitos in the City of Chicago. XGBoost is a promising classifier algorithm that can be used to help predict the West Nile Virus for a certain time, location, and help determine which mosquito species are more prone to being infected. XGBoost was selected because it performed better than a Random Classifer with AUC scores of 70% and 50%, respectively. Weather data was selected as the features, along with some added features for time lag. Feature selection is discussed.

**Introduction**

The West Nile Virus (“WNV”) is a disease most commonly spread to humans by the bite of an infected mosquito. These occur mostly during mosquito season, which begins in the summer and continues through the fall. Currently there are no vaccines or medications to prevent or stop the virus. Per the Centers for Disease Control and Prevention (“CDC”), 1 in 5 people who are infected by the virus develop a fever and other symptoms and about 1 in 150 people who are infected can develop a serious, sometimes fatal, illness.[[1]](#footnote-2) In 2002 the first human cases of the virus were reported in the City of Chicago. In 2004, the City of Chicago and the Chicago Department of Public Health (CDPH) implemented a control program that is still in effect today. The program consisted of surveillance of mosquitos through traps and spraying the region where potential outbreak may occur.[[2]](#footnote-3) Consequently, we need to predict when and where the mosquito will test positive for WNV. The effective model will help the City of Chicago and the Chicago Department of Public Health (CDPH) effectively allocate mitigating factors, such as spraying, for the predicted time and location.

**Data Wrangling & Exploratory Data Analysis**

The Chicago Department of Public Health provides data and has made it publicly available through Kaggle, a platform for the data science and machine learning community. The datasets consist of the following: weather, spray, and training.

**Weather:** The ***Weather*** data provides data recorded by the National Oceanic and Atmospheric Administration (“NOAA”) from two different weather stations. The two stations are identified as Station 1 and Station 2. The weather data consists of 2944 rows with 21 features. Values are missing, identified as ‘M’, from rows and columns of the weather dataset. Other missing values are just null values. The features are related to temperature, average dew point, average wet bulb, heat, average station pressure, average sea level pressure, and wind speed.

**Spray:** The ***Spray*** dataset displays when and where spraying occurred, which were represented in date, time and GIS coordinates. The dataset consisted of 14835 rows and 4 columns

**Training:** The ***Training*** dataset includes information as it relates to the virus, such as date, traps, location of traps, number of mosquitos caught in traps, species of mosquitos caught in traps, and presence of virus in mosquitos. The target variable is denoted as Wnv\_Present where 1 equates to “Present” and 0 equates to “Not Present”, meaning the virus was or was not present in the caught mosquitos. The dataset consists of 10506 rows and 11 features containing data collected for years 2007, 2009, 2011, and 2013.

During the data wrangling process, it was determined that the ***Spray*** data was not going to be utilized. Only the ***Training*** and ***Weather*** data were going to be used. No obvious anomalies were detected in the ***Training*** dataset. For example, there did not appear to be any missing values, duplicates, or inconsistencies. The ***Weather*** dataset, however, did have missing values and null values for several rows and columns. Therefore, those missing values were filled or replaced. Below is a summary of those that were filled or dropped, and which approach was taken:

**Average**

**Tavg:** Average of Tmax and Tmin was calculated to fill NaNs, consistent with populated fields.

**Median**

**WetBulb, Heat, Cool, SnowFall, PrecipTotal, StnPressure, SeaLevel, and AvgSpeed:** These missing values were filled with Median. This approach was decided because these values were in between the calculated Mode and Average. Also, there weren’t too many fields that needed to be populated, therefore, the risk of skewing the data was minimal.

**Ffill()**

**Depart, Sunrise, and Sunset:** Station 2 did not have information for these features. Because Station 1 and Station 2 are approximately 24 miles apart, I believe it was appropriate to forward fill the missing values with Station 1 numbers. Due to the relatively short distance between the stations, I believe this approach is more accurate than filling with average or median values as it is better aligned with trending data for that day.

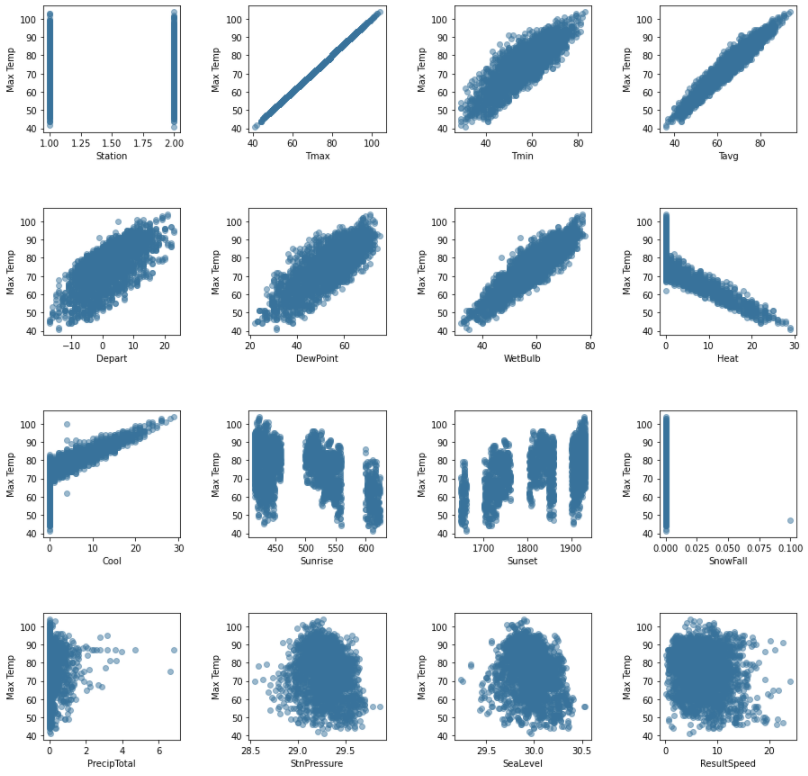
**Deleted**

**Water1 and Depth:** These columns were deleted, as Water1 consisted of only M’s and Depth consisted of M’s and 0. Therefore, they lacked in being informative.

**Time Series**

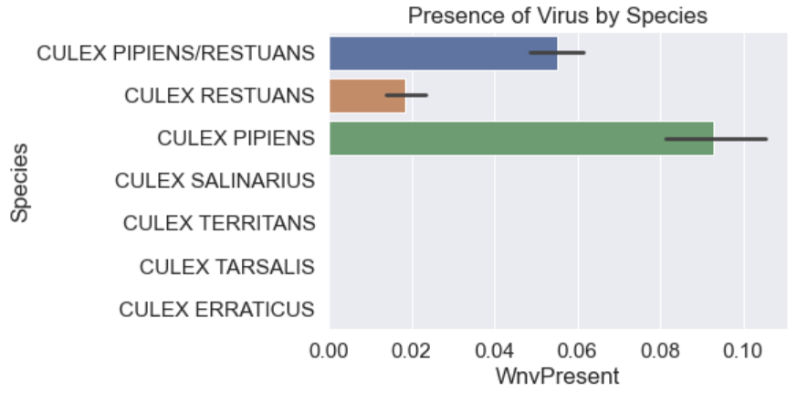
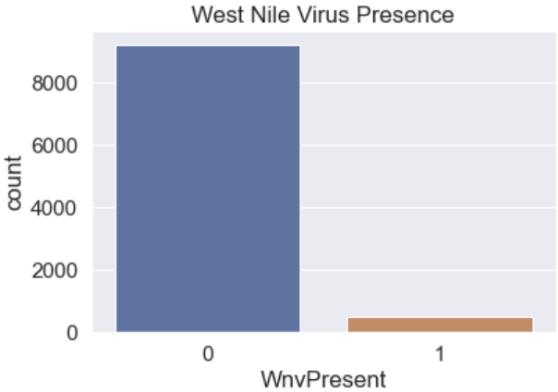
**Date:** For both, the ***Weather*** and ***Training*** datasets, the ‘Date’ column was converted from data type ‘Object’ to ‘DateTime’ to facilitate time series analysis.

After the data was cleaned, analysis was performed on the ***Weather*** data to observe correlation between the features. In the scatterplots below, it can be observed that Tmin, Tavg, Depart, DewPoint, WetBulb, and Cool had a positive correlation with Tmax, while Heat had an inverse correlation.



It was then observed that SnowFall should have been deleted as well, as it provides no additional insight with SnowFall being zero. The column was deleted.

Additionally, I wanted to see how much of the data was showing positive results for the West Nile Virus (Exhibit A), and the correlation with the different species of mosquitos (Exhibit B).



**Exhibit A** **Exhibit B**

It can be observed that only a small amount of the data contained positive results for the West Nile Virus. Of those positive results only 3 of 7 mosquito species carried the virus.

**Feature Engineering**

New features were added to the ***Weather*** dataset. Kaggle provided GIS location for Station 1 and Station 2, therefore, Latitude and Longitude features were added to denote these locations. The ***Training*** and ***Weather*** datasets were joined by Date, for those that had shorter distance between Trap location and Station. For example, if the Trap location in the ***Training*** dataset was closer to Station 1 than Station 2 from the ***Weather*** dataset, then Station 1 was kept and Station 2 was dropped. This was done to avoid duplicated data. After the data was joined more features were added. Below is a summary of features added:

**Relative\_Humidity**: Because the ***Weather*** dataset contained Temperature and DewPoint information it was determined that Relative Humidity could be calculated. Therefore, it was calculated and added as a feature.

**SpeciesDummy\_“SpeciesName”**: The column named ‘Species’ listed the name of the mosquito species. In order to be able to use this data for the models, it was transformed to binary information by using the pd.get\_dummies function.

**“X”\_Shift1[[3]](#footnote-4)**: Weather conditions can have a delayed effect on certain phenomenon. Therefore, Weather data was shifted to show a one-day lag.

**“X”\_Shift2[[4]](#footnote-5)**: Weather conditions can have a delayed effect on certain phenomenon. Therefore, Weather data was shifted to show a two-day lag.

**Month**: Isolated the month from the Date column.

**Season**: Identified the Season from the Date column.

**Year**: Isolated the year from the Date column.

**WeekDay**: Identified the day of the week from the Date column.

**MonthDay**: Identified the day of the month from the Date column.

**Week**: Identified the week number of the year from the Date column.

**Season\_Year**: Combined the Season and Year for year-over-year comparison or season-over-season comparison.

After the additional features were added, further analysis was performed. The relation of virus presence by season can be observed in Exhibit C, which is then broken out by month in Exhibit D. It can be observed that West Nile Virus presence is highest in the Summer, extending into the fall. As I drilled down into the months, it was observed that August, had the highest positive cases.

**Exhibit C**

Chart, bar chart

Description automatically generated

**Exhibit D**

Chart, bar chart

Description automatically generated

Then a year-over-year analysis was performed, to see if the number of positive cases was consistent by season over the years, demonstrated in Exhibit E. It was determined that years 2007 and 2013 observed the highest positive cases, while years 2009 and 2011 had a drop in cases. This could be related to extreme weather condition changes over the years.

**Exhibit E**

Chart, bar chart

Description automatically generated

It can be induced that the West Nile Virus is predominantly present during hotter seasons, which are during late Summer and early Fall.

**Variable Selection and Pre-Processing**

Although correlation between certain variables was observed with additional analysis, there could be other important variables that have not been observed. Information Value (IV) is one of the most useful techniques to select important variables in a predictive model. The IV Statistic formula is written as and calculates values between 0 and 1[[5]](#footnote-6).

The IV scores can be interpreted as follows:

* < 0.02 – Not useful for prediction
* 0.02 to 0.1 – Weak predictive power
* 0.1 to 0.3 – Medium predictive power
* > 0.3 – Strong predictive power

Information value increases as bins/groups increase for an independent variable, therefore, I set bins to max-out at 20. Next an iteration of Variance Inflation Factor (“VIF”) was performed. VIF is a commonly used tool to detect whether multicollinearity exists. It measures how much the variance is inflated due to collinearity. VIF can be calculated by the formula below[[6]](#footnote-7):

Features exhibiting VIF greater than 5 demonstrate extreme multicollinearity and are avoided. This aids in making the features independent and ensures the model can predict the dependent variable. After applying IV and VIF techniques to the dataset, the features were reduced to 12 different features. The features selected as the important features in the dataset were as follow: Depart, Result Speed, Depart (shifted to 2-day lag), Latitude Difference, Block, Result Direction, Cool (shifted to a 1-day lag), Day of the Month, Culex Restuans Species, Day of the Week, Culex Pipiens, and Heat. Feature importance shown in Exhibit F, below.

**Exhibit F**

Chart

Description automatically generated

In order to understand the feature importance better, brief definitions are outlined below:

**Depart**: Departure from normal temperature.

**Result Speed**: Wind speed in miles per hour.

**Latitude Difference**:

**Block**: Block number of addresses.

**Result Direction**: Direction of wind in degrees.

**Cool**: Cooling season begins with January.

**MonthDay**: Day of the week mosquito was trapped.

**Culex Restuans**: One of the species of mosquitos caught in Traps. Culex restuans has a distribution that ranges from central Canada south into Mexico. The mosquito is very common in the eastern and central United States. Populations of this mosquito usually peak by July. A second peak is often evident in the fall to produce the adults for the overwintering generation.[[7]](#footnote-8)

**WeekDay**: Day of the week mosquito was trapped.

**Culex Pipiens**: One of the species of mosquitos caught in Traps. ***Culex pipiens***, commonly referred to as the **common house mosquito**, is considered as the **northern house mosquito**, as it is the most common mosquito to the northern regions of the US. *Culex pipiens*' diet typically consists of vertebrate blood, as they consume human blood, but prefer bird blood of species that are nearly linked to human interaction, such as doves and pigeons. Furthermore, at the end of the summer and the start of the fall season before it is time for them to overwinter.[[8]](#footnote-9)

**Heat**: Heating season begins with July.

The presence of the features is reasonable, as departure in temperature from the average temperatures more than likely escalated to higher temperatures. Mosquitos thrive in high temperatures and heat index, and consequently breed and increase their populations. With higher mosquito populations, the higher the risk of exposure to the West Nile Virus.

**Data Modeling**

The data was separated into training and testing datasets with a 70/30 split, respectively. Initially the training and testing datasets were modeled through a Random Forest Classifier algorithm. However, the model resulted in an AUC score of 51% which is merely a coin toss. Next, the training and testing datasets were modeled through the classification algorithm, eXtreme Gradient Boosting (“XGBoost”). Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems.[[9]](#footnote-10)

Although initially the AUC score didn’t improve significantly with XGBoost (to 55%), an algorithm pipeline was performed to identify the best parameters for the model. Once I got the best parameters, the model was again trained with the selected parameters. The AUC score increased to 70%. Below you can see the Random Forest Classifier (Exhibit G) and XG Boost Model confusion matrices (Exhibit H). The predictive effectiveness of the Random Forest Classifier Model is not very high, with a True Positive rate for Virus Presence being 0.03% and 100% for Virus Absence. The outcome is a result of the highly imbalanced data, reflecting an absence of the virus. Therefore, the model favors this trend and will predict incorrectly.

**Exhibit G** **Exhibit H**

Chart, treemap chart

Description automatically generated Chart

Description automatically generated

With the XGBoost Confusion Matrix it can be seen that it is less biased, therefore, this is the preferred model.

**Results**

The model classifies the presence and the absence of virus with an AUC of 0.70 as seen below (Exhibit I). The XGBoost classifier algorithm predicted the presence of the virus with a probability of 0.65 and the absence of the virus with a 0.75. There is more cost associated with the improper classification of the presence of the virus.

**Exhibit I**

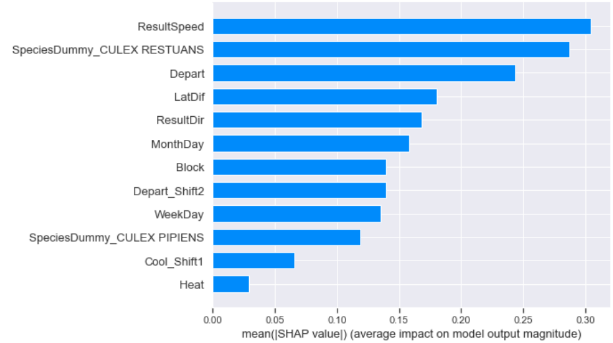
Chart, line chart

Description automatically generated

Additionally, the importance of features and their impact on the model’s output can be explained with SHAP values and summary plots, as shown below. SHAP is an algorithm that was first published in 2017 and it reverse-engineers the output of predictive algorithms (e.g. gradient boosting, neural network, etc.). In summary, SHAP values help understand what decisions complex models are making.[[10]](#footnote-11)

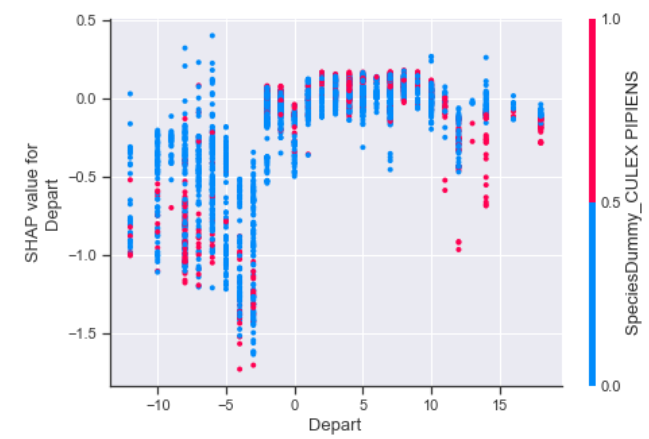
Below can be observed the feature importance with SHAP values (Exhibit J), which are nearly consistent with Exhibit F, previously stated. Colors red and blue show whether the feature has high or low feature value on the model, respectively. The horizontal location shows whether that value has a higher or lower impact on model prediction. Although multiple features were selected as important, some of them had an inverse relation with the WnvPresent feature, which was selected as the feature to be predicted.

**Exhibit J Exhibit K**



It can be observed in Exhibit K that Depart, LatDiff, ResultDir, Block, WeekDay, Culex Pipiens, and Heat had a high value in predicting the presence of the virus. This helps to understand if the model can predict the initial question of when and where the virus will be present. With latitude distance from Station 1 or 2 and the Block information selected as important features, it can be determined where the virus could potentially be present next. In order to answer the question of when, it can be answered when the weather conditions obtained from features Depart, Result, Dir Heat, and the inverse correlation with Cool meet the weather conditions to make vector mosquitos thrive. Per the SHAP values plot, Culex Pipiens is the mosquito mostly susceptible to carrying the virus, which agrees with the previous plot (Exhibit B) performed during initial analysis. Further, it can be inferred that the more temperature departs from the average temperature, the higher the presence of the virus in Culex Pipiens as shown in Exhibit L.

**Exhibit L**



Although, overall, the predictive power of the model is good there are a few observations that can be of concern and should be considered. Per the SHAP values plot in Exhibit K, it is determined that WeekDay has a positive predictive value with virus presence. However, the WeekDay alone provides no additional insight as to the presence of the virus. Seasonality would have been more indicative, as mosquitos thrive in hotter conditions. It is hard to determine if the WeekDay would have been in August or January. Another concern is that the predictive power was average, with a probability of 0.65 for predicting the presence of the virus. To further improve the model, features can be added to identify the WeekDay by Season and percentage values of virus presence by month and season. Lastly, the Spray data was not utilized, but could be included in future modeling. It could help consider current mitigating processes and effectiveness, which could help boost predictive power.

1. [West Nile virus | West Nile Virus | CDC](https://www.cdc.gov/westnile/index.html) [↑](#footnote-ref-2)
2. [West Nile Virus Prediction | Kaggle](https://www.kaggle.com/c/predict-west-nile-virus/overview) [↑](#footnote-ref-3)
3. ”X” representing the Weather dataset column name that was shifted. [↑](#footnote-ref-4)
4. Same as above. [↑](#footnote-ref-5)
5. [Weight of Evidence (WOE) and Information Value (IV) Explained (listendata.com)](https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html#:~:text=Rules%20related%20to%20Information%20Value%20%20%20,Strong%20predictive%20Power%20%201%20more%20rows%20) [↑](#footnote-ref-6)
6. [Variance Inflation Factor (VIF) - Overview, Formula, Uses (corporatefinanceinstitute.com)](https://corporatefinanceinstitute.com/resources/knowledge/other/variance-inflation-factor-vif/) [↑](#footnote-ref-7)
7. [Mosquito Biology: Rutgers Center for Vector Biology](http://vectorbio.rutgers.edu/outreach/species/rest.htm) [↑](#footnote-ref-8)
8. [Culex pipiens - Wikipedia](https://en.wikipedia.org/wiki/Culex_pipiens) [↑](#footnote-ref-9)
9. [XGBoost for Regression (machinelearningmastery.com)](https://machinelearningmastery.com/xgboost-for-regression/#:~:text=Extreme%20Gradient%20Boosting%20Gradient%20boosting%20refers%20to%20a,correct%20the%20prediction%20errors%20made%20by%20prior%20models.) [↑](#footnote-ref-10)
10. [SHAP Values Explained Exactly How You Wished Someone Explained to You | by Samuele Mazzanti | Towards Data Science](https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30#:~:text=SHAP%20Values%20Explained%20Exactly%20How%20You%20Wished%20Someone,6%20Wrapping%20it%20up.%20...%207%20Great%21%20) [↑](#footnote-ref-11)