**The West Nile Virus Analysis Report**

**Capstone 2**

By Patric Oh

# CONTEXT

The West Nile Virus is mostly spread to humans through infected mosquitos and roughly 20% of people who become infected with the virus develop symptoms ranging from a persistent fever to serious neurological illness that can result in death.

In 2002, the first human cases of the 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 week from late spring through the fall, mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray airborne pesticides to control adult mosquito populations.

# CRITERIA FOR SUCCESS

In order to be considered successful in this project, we must complete the following:

* Conduct exploratory analysis and determine what factors are affecting the presence of the West Nile Virus in mosquitos
* Build a logistic regression model with an AUC score of over 50%
* Identify what threshold is required of this logistic model to identify actual cases of the virus at a 100% rate

The goal is to create an improved detection model that will ultimately save many lives in the city of Chicago.

# SCOPE OF SOLUTION SPACE

The entire data analysis has been completed using Python for data analysis and visualization in Jupyter Notebook.

1. Data Preparation

After importing all the necessary Python packages (Numpy, Pandas, Matplotlib, etc.) for analysis, we imported the datasets and examined the data. We immediately saw that the data had to be cleaned and prepared after viewing the shape and information of the datasets.

Some integer value columns were set as objects due to having “M” or “-“ inputs for missing values. These missing values were replaced to null and have been filled using forward fill to estimate the missing values as the dataset was in order by date.

There were also “T” inputs for traces of precipitation which we replaced with 0.005 to resemble extremely low amounts of precipitation.

After the dataset was cleared of null values, the datatypes that were “object” were changed to “numeric” so that analysis can be properly conducted.

1. Data Exploration

First thing we checked was the data balance of virus count to non-virus count. The dataset was highly imbalanced at a ratio of approximately 95 to 5.

Chart, bar chart

Description automatically generated

We grouped the sum of the virus by species to observe if there were any particular species that are more prone to carrying the virus. Out of the recorded 7 species, only 3 of them have been recorded carrying the virus. We will only focus on these 3 types (Culex Pipiens/Restuans, Culex Pipiens, Culex Restuans) of mosquitos for the remainder of this analysis.

Chart, histogram

Description automatically generated

Text

Description automatically generated

To view the chances that these 3 species will carry the virus, the positive cases of each species were divided by the number of mosquitos of each species. This concluded that just based on the species alone, there was anywhere between 0.2-0.5% chance that they have been carrying the virus.

Text

Description automatically generated

Time series analysis was done for each of the 3 types of mosquitos. We noticed that there was an abnormal number of cases in Aug 2007. During this month, there was also an unusually high number of mosquitos recorded as well. The highest cases of the West Nile Virus were recorded in August of every year. 2009 and 2011 had very low virus rates compared to 2007 and 2013.

Chart, line chart

Description automatically generated

We used the Geopandas package to visually plot cases on a map of Chicago when there were no cases of the virus and when there were cases. Based on this geographical visualization, we saw that there were significantly less cases of the virus in the middle region of Chicago.

Chart, scatter chart

Description automatically generated

The geographical coordinates were used to split up the region into 6 categories: Northwest, Northeast, Midwest, Mideast, Southwest, and Southeast. The Northwest region of Chicago had the most tests and the highest cases of the virus. There were more tests conducted in the Middle region of Chicago, but more virus cases have been found in the South. We can conclude that the Middle region of Chicago is safer, in terms of the West Nile Virus, compared to the North or South.

Chart, bar chart

Description automatically generated

Until now, we have been conducted analysis based on the Training dataset only. The Weather dataset consisted of 2 stations which each had records on the same date. The Weather dataset was self-joined on the date to unify the date column, then it was joined to the Training dataset on the date column.

We were given the station 1 coordinates (41.995, -87.933) and the station 2 coordinates (41.786, -87.752). As each record of the Training dataset included coordinates, the Haversine package was used to identify which station was closer in distance and used the closest station’s weather information as “true” weather.

The range of temperatures in 2009 and 2011 when the virus was present is smaller than when the virus is not present. Because there were less cases during those years, there were less chances for the range to be larger.

But in 2009, there was a big difference as temperature was roughly between 70-73 degrees when the virus was detected. In comparison, the average temperature range when the virus was not present during that year was between 52-80 degrees.

The average temperature during 2007 and 2013 doesn’t seem to play a factor with the presence of the virus.

Chart, box and whisker chart

Description automatically generated

We viewed the percent change of all the weather codes when the virus was not present to when the virus was present. To do this, we normalized the weather codes under both conditions and then ran a percent change from when WNV=0 to WNV=1. The visualization below shows that under certain weather codes, there has been up to an approximate of 300% change when the virus was present.

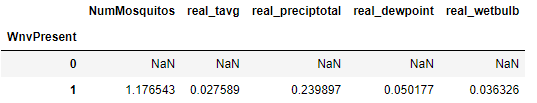
Chart, bar chart

Description automatically generated

There were other averages that we did want to look into such as the number of mosquitos (NumMosquitos), average temperature (real\_tavg), precipitation total (real\_preciptotal), dewpoint (real\_dewpoint), and wet bulb (real\_wetbulb). All of these features showed an increase when the virus was found. In particular, we see the number of mosquitos increase by 117.65% and the precipitation total increase by 23.99% on average.

Table

Description automatically generated



1. Logistic Regression

To create a logistic regression model, we had to import packages such as SciKit-Learn and Imbalanced-Learn. Not all of the columns will be used so we sliced the data and only brought in the relevant features for the logistic regression model. Because some of these features are categorical, we created dummy variables which brought the total number of columns for this data frame to 59.

The dataset was split to dependent (y) and independent variables (x). The dependent variable is “WNVPresent” which identifies if the virus is present or not. Then Feature Importance was ran to see which features are useful in predicting the presence of the virus. The features that had no effect on the presence of the virus was dropped which brought the total number of columns to 34.

To train the logistic regression model, the dataset was split into 80% train and 20% test. The scale of the variables was standardized so that the mean of each column equaled to zero. Then we fitted our first logistic regression model.

Being that our data was highly imbalanced, this model was predicting that there were no cases of the virus 99.38% of the time. This resulted in a large number of false negatives which makes this model unusable. For a total of 103 virus occurrences in the test data, the model was able to predict only 7% correctly.

Chart

Description automatically generated

The accuracy was high at 0.95 but our recall for 1 was only 0.07. In order for us to build a logistic regression model that will perfectly predict actual virus occurrences, the recall must be 1.00.

Table

Description automatically generated with medium confidence

In the ROC Curve, the AUC (Area Under the Curve) was 0.8642. We have already achieved our objective of creating a logistic regression model with an AUC score of over 0.50 but we will see if we can make further improvements on the score.

Chart, line chart

Description automatically generated

Because the previous model was unusable, we utilized under sampling to balance the data by using RandomUnderSampler. Under sampling was used instead of over sampling because we wanted to keep the actual virus data without inflating the numbers which will create fictional records. Balancing the data to about a 6:4 ratio (772 negatives, 448 positives) made the new model a lot more usable with a recall score for 1 of 0.68.

Table

Description automatically generated with low confidence

Unfortunately, the AUC score of this new model dropped a bit.

Chart, line chart

Description automatically generated

We attempted to increase the AUC score by tuning the hyperparameters of “penalty” and “C”.



With the optimized hyperparameters, we were able to increase the AUC score from 0.8576 to 0.8642 (approximately 0.76%).

Chart, line chart

Description automatically generated

In our scenario, it is more important that we are able to predict all actual cases of the virus, even if that means compromising our precision. The current threshold of our model is at 0.5. By lowering the threshold to 0.04, we were able to achieve a recall score of 1.00 that correctly predicted every case of the virus in the test dataset. Achieving this recall score for 1 compromised our recall score for 0 and the precision score for 1.

Table

Description automatically generated

This new model with the adjusted threshold is acceptable in our case because although we will be over spraying areas that our model predicted a false positive, we can be confident that we will also be spraying areas where the virus is present. Our final model will help eliminate the virus in the area of Chicago.

Chart, treemap chart

Description automatically generated

Using SHAP, we created a bar plot of the summary where we are able to see which features were the most impactful in our logistic regression model. The plot is in descending order by the level of impact, regardless of positive or negative impact.

Chart, bar chart

Description automatically generated

The SHAP summary plot displays the following information:

* Feature importance: Variables are ranked in descending order.
* Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
* Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.
* Correlation: A high level of the “number of mosquitos” has a high and positive impact on the positive prediction of the virus. The “high” comes from the red color, and the “positive” impact is shown on the x-axis. Similarly, we will say the “month of June (month\_6)” is negatively correlated with the target variable.

We can see that the "month of June (month\_6)" had the highest impact on our prediction model and it was negatively correlated. This means that our model predicted the non-presence of the virus mostly when it was the "month of June".

The following few features, from the "number of mosquitos (NumMosquitos)" to the "month of September (month\_9)", were all positively correlated to the prediction of the presence of the virus in descending order. For example, the "number of mosquitos" had the highest impact on the prediction of the presence of the virus followed by the "Culex Pipiens/Restuans species" and the "Culex Pipiens species".

Chart

Description automatically generated

# CONCLUSION & RECOMMENDATIONS

The features that had the most impact towards the prediction of the virus, other than the number of mosquitos, were the type of species and the month of year. If the species are the Culex Pipiens/Restuans, Culex Pipiens, or Culex Restuans, the model is influenced to predict the presence of the virus. The same goes for the month of August and September.

The Chicago City Council can utilize this model to understand the likelihood of where the breakout will happen next. They will be able to see which factors impact the presence of the virus and which are less likely to do so. With this data, the city can take preventive action or measures to reduce the spread of the virus and hopefully get rid of it altogether. The model will also show approximately how much over spraying will need to be done in order to be confident that all incidents of the virus will be sprayed.

# CONSTRAINTS WITHIN SOLUTION SPACE

Being that there are multiple large datasets involved, there is a possibility that there may be incomplete, missing, or erroneous data that were not identified. Some missing data elements had to be estimated from the previous day in the Weather dataset. This may compromise the accuracy of the weather element of the analysis, but it should stay somewhat correct as weather generally doesn’t make a drastic change from one day to the next.

# STAKEHOLDERS TO PROVIDE KEY INSIGHT

Stakeholders:

* Chicago City Council
* Data Scientist (Rahul Sagrolikar)
* Data Analyst (Patric Oh)

# KEY DATA SOURCES

Data source: <https://www.kaggle.com/c/predict-west-nile-virus>

Datasets:

* Train data (train.csv): Training set in 2007, 2009, 2011, 2013
* Weather data (weather.csv): Weather information from 2007 to 2014