Predicting Chicago West Nile Virus

Bana 273 - Final Project

Team Members

Amy Yeji Lee
Jennifer Siwu
Manodhar Allu
Matthew Littman
Sonal Mendiratta
Vishnu Madan

Content

1. Background……………………………………………………………………..2
2. Project Scope and Goal…………………..……………………………………...2
3. Data Preparation ………..……………………………………………………..2
4. Insights…………………………………………………………………………..3
5. Virus Prediction Model …………………………………………………………...8
6. Final Model Selection…………………………………………………………...…11
7. Next Steps - Time Series………………………………………………………..12
8. Key Learnings and Takeaways……………………………………………………17
9. Appendix………………………………………………………………………...18
10. References……………………………………………………………………….24

Background

On average, around 20% of people who become infected with the West Nile virus develop symptoms ranging from a persistent fever to serious neurological illnesses that can result in death. In 2002, the first human cases of West Nile virus were reported in Chicago. By 2004, the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today.

Every week from late spring through the fall, mosquitoes in traps across the city of Chicago 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.

Project Scope and Goal

In this project, we plan to analyze how we can prevent or West Nile Virus outbreak. Currently, our datasets mainly include information within the city of Chicago in terms of the number of mosquito traps, weather conditions, mosquito spraying, and mosquito species. Most of the data provide all the above information by specific latitude and longitude as well as include information of day and time.

Given the weather, location, mosquito, and spraying data, we are going to predict the presence of the West Nile Virus at a specific location, date, time and weather conditions. With this project, we hope to help the City of Chicago more efficiently and effectively allocate resources to prevent the outbreak of the West Nile Virus.

Data Preparation

The data were collected from Kaggle and contained 3 different files, including Trap data, Spray data, and weather data.

- **Trap** dataset includes information about the trap location, species that are tested, number of mosquitos, and WNV presence. Trap data is recorded for 2007, 2009, 2011,2013.
- **Spray** dataset includes information about the date, time, and location of the spray. Sprays are done 2 times in 2011 and 8 times in 2013 with each spray being separated by one week.
- **Weather** data is captured by 2 stations separated by 17 miles and recorded from 2007 to 2014.

In order to combine all the information together for our prediction, additional features were created from each dataset, including:

- Total mosquitos in a trap, on a given day and location from **Trap** dataset.
- Spray time and distance from **Spray** dataset.
- Relative humidity, Length of day (time between sunset and sunrise), Heat index

(benchmark 65F), and Moving averages from **Weather** dataset.

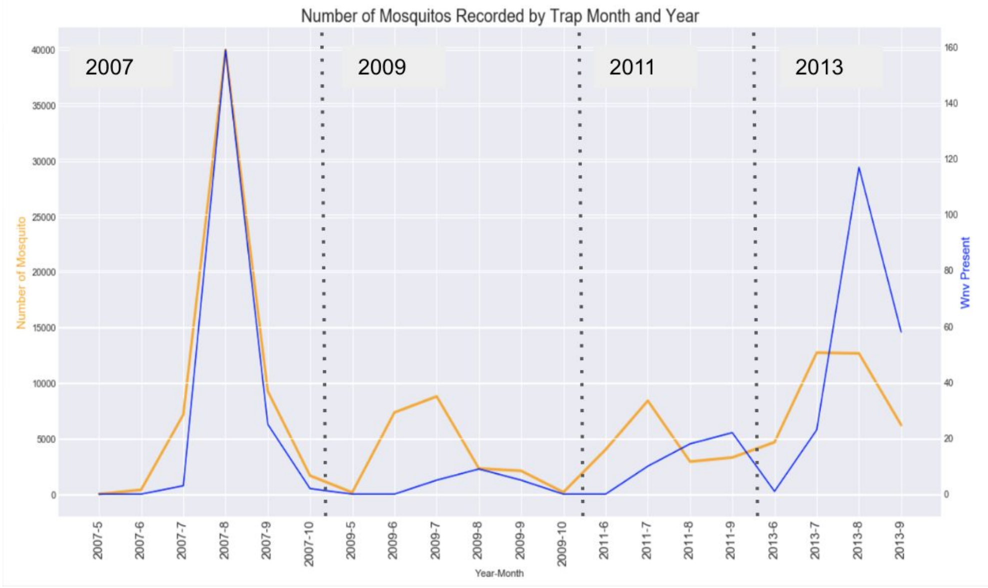
After cleaning the data and creating additional features, we merged the dataset altogether and created a master database file that we used for Exploratory Data Analysis and Modelling.

Insights

To get a better understanding of the data and find trends or pattern, we conducted an exploratory data analysis.

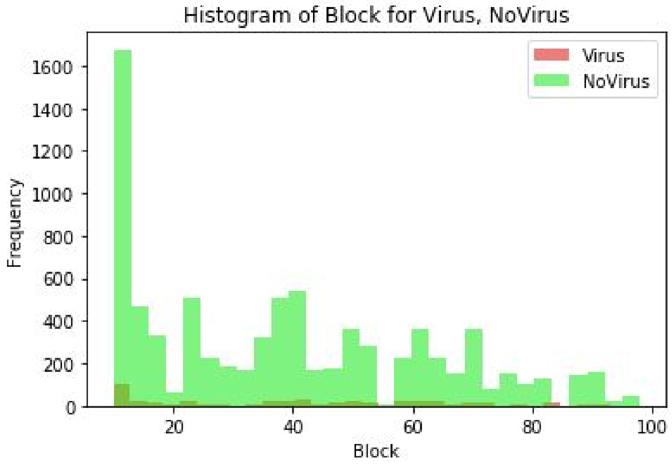
First, we compared the Number of Mosquito vs Virus Presence Distribution by Month and Year from 2017 to 2013. As shown by Appendix I, there was a noticeable decline in both virus presence as well as the number of mosquitoes, and the same base continued until 2011. However, the virus presence saw a prominent increase in 2013 while number of mosquito increase was not as significant.

Appendix I - Mosquito Distribution by Month and Year

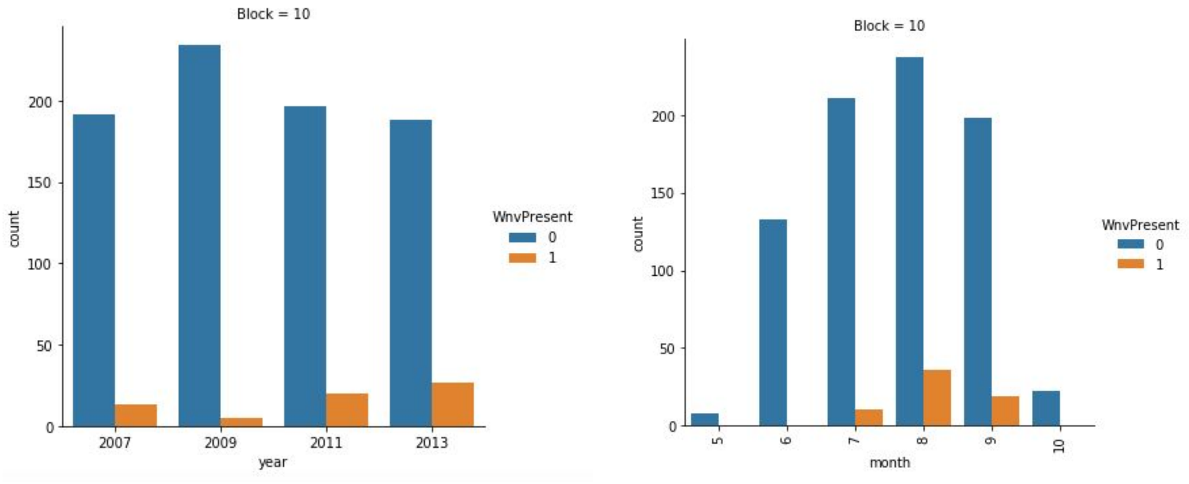


These findings led us to question what could be the indicator of the virus spike in 2013. To begin, we started by zooming through block specific mosquito and virus presence information. According to Appendix IIa, block 10 had the most virus presence compared to other blocks. After zooming through the month (appendix IIb), we also saw that block 10 had the most virus presence in August 2013.

Appendix IIa - Virus Presence in 2013 by Block

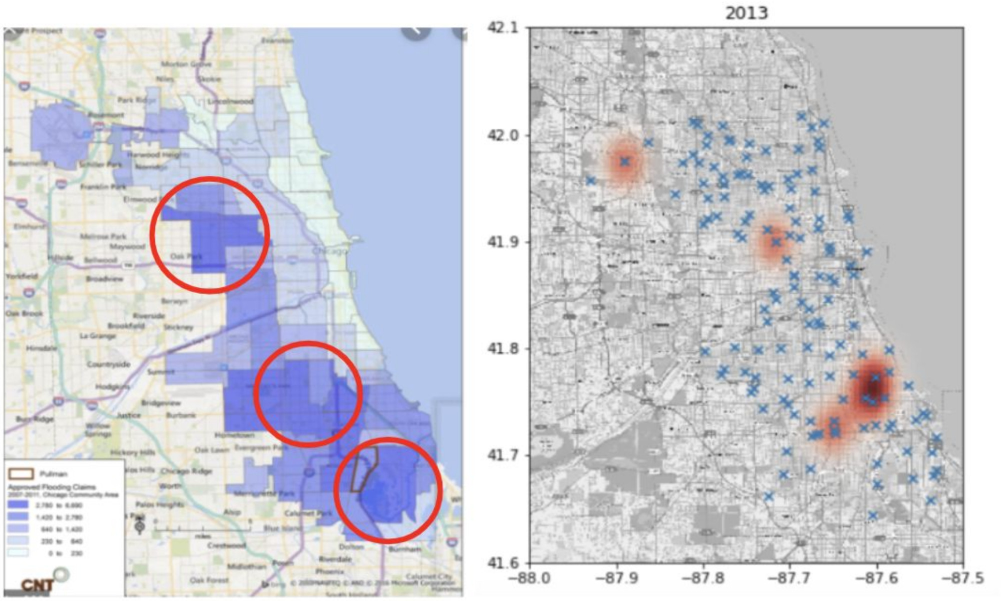


Appendix IIb - Virus Presence in Block 10 by Year and Month



To answer this question further, we looked into what happened at Block 10 in 2013 and we found out that between **April** 15th and 19th **2013**, a slow-moving storm system resulted in record **flooding** across much of the State of Illinois. Appendix III compares the areas damaged by the flood (left) and the areas with virus presence (right). The darker the gradient, the more damaged the area was by the flood. As shown on the graph, the heavily damaged areas were the ones with most west nile virus presence. Additionally, Block 10 showed to be the most damaged area.

Appendix III - Areas Damaged by Flood vs Areas with WNV Presence

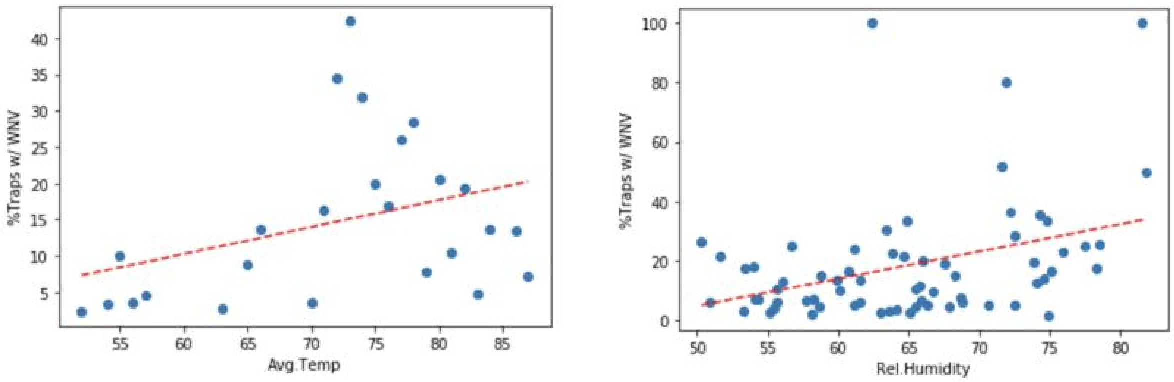


The time gap between the flood incident and the virus presence month were due to temperature. “Mosquitos form 'molestus' achieved complete development between **temperatures** of 15 and 32.5°C, at both constant and fluctuating **temperature** regimes. Survival from egg to adult at both constant and fluctuating **temperature** of 32.5°C was low and **only few individuals pupated and emerged successfully.”**(Wang J., 2011)

This finding concluded our assumption that there was a clear association between weather conditions (temperature and precipitation) and mosquito abundance, which allowed the definition of threshold criteria for temperature and precipitation conditions for mosquito population growth.

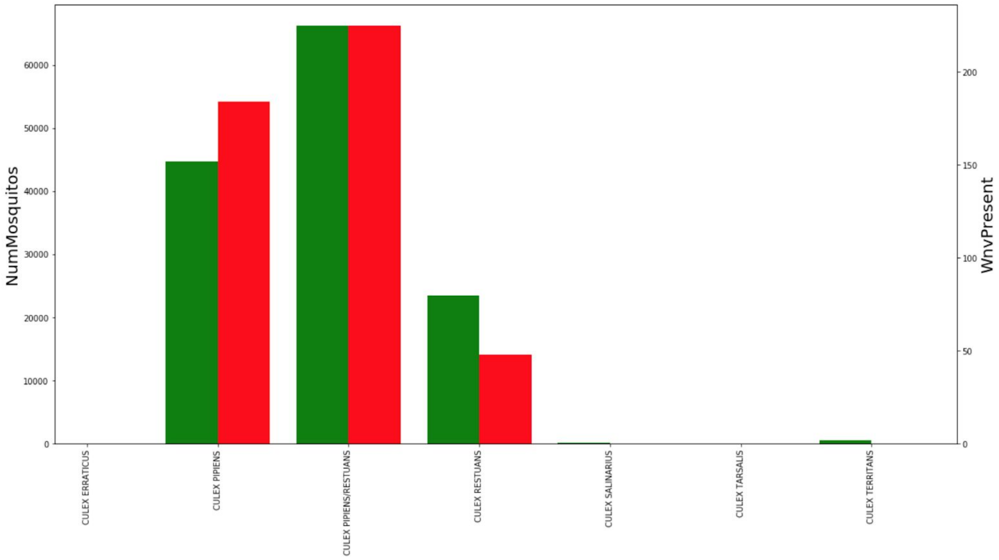
Besides weather conditions, Temperature and Relative humidity also affect virus presence. Based on appendix IV, The percentage of traps with WNV increases with the increase in temperature. About 30-40% of the traps had WNV presence in the temperature range of 72F-82F.

Appendix IV - Impact of Temperature and Relative Humidity on Virus Presence



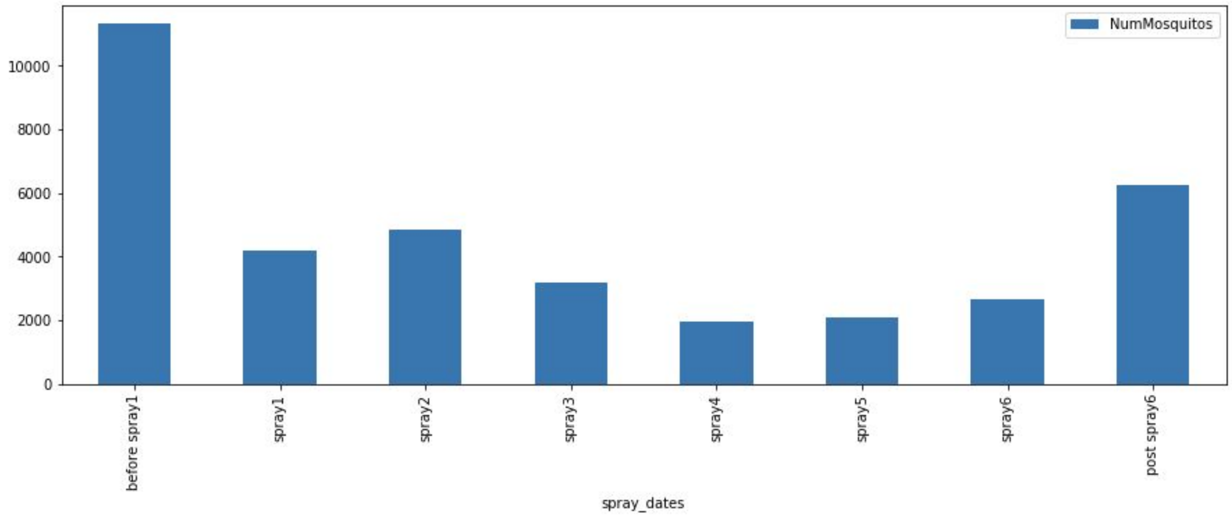
Although there were 6 mosquito types identified by the Trap, only 3 species carried the west nile virus. Based on Appendix V, we confirmed CULEX PIPIENS/RESTAUANS is the main carrier associated with the presence of the virus among those 3 species that spread the virus.

Appendix V - Mosquito Species with West Nile Virus



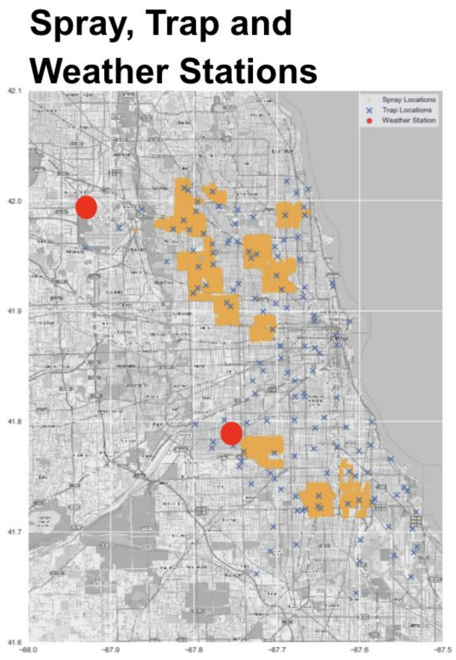
To measure whether or not the spray conducted were effective in limiting the number of mosquitos, we compared the number of days between the sprays to the mosquitos presence at the Trap. Based on Appendix VI, it was shown that In 2013 the number of mosquitoes decreased only during the spray time when compared with number of mosquitoes before and after spray.

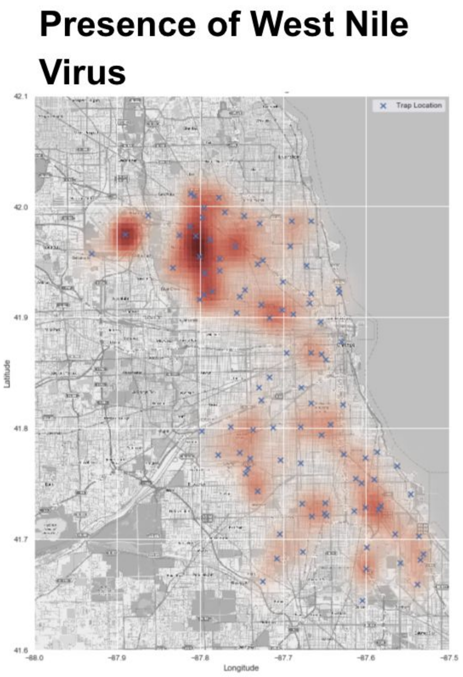
Appendix V - Effect of sprays on number of mosquitos



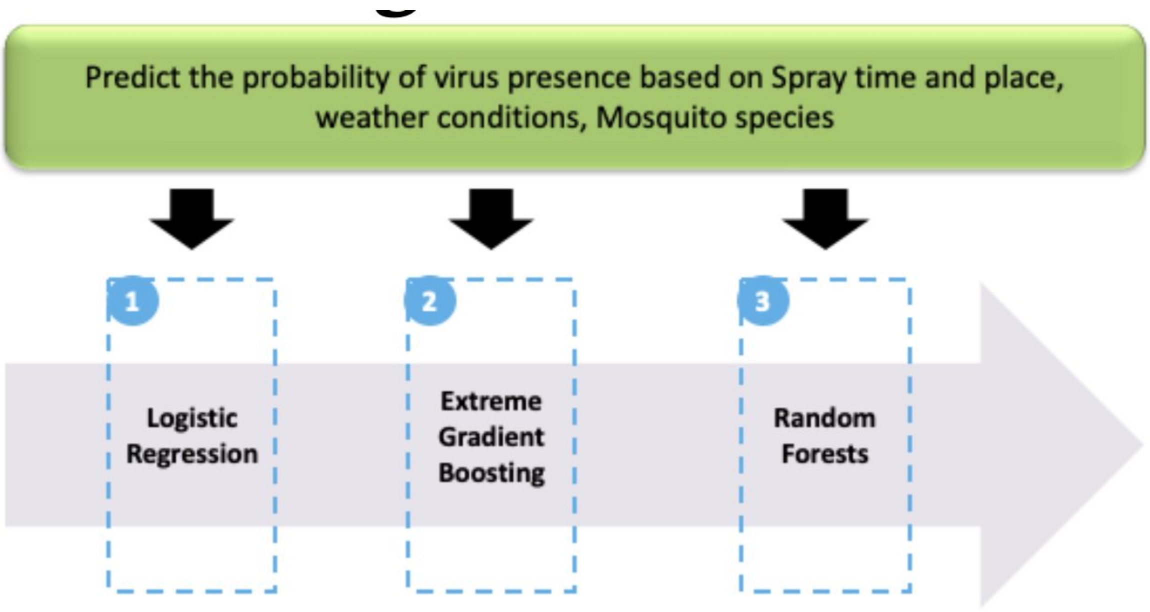
Given that Spray does minimize the number of mosquitoes that potentially have a similar impact to the virus, it was shown that City of Chicago has been spray at the locations where virus were present. (Appendix VI)

Appendix V -Virus Presence vs Spray, Trap & Weather Locations


Spray, Trap and Weather Stations


Presence of West Nile Virus

Virus Prediction Model



Predict the probability of virus presence based on Spray time and place, weather conditions, Mosquito species

1 Logistic Regression

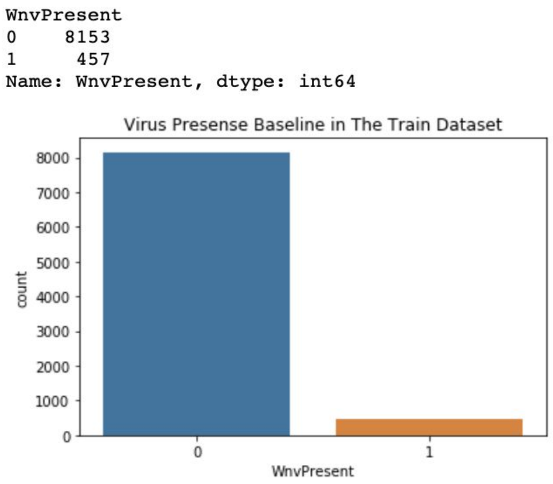
2 Extreme Gradient Boosting

3 Random Forests

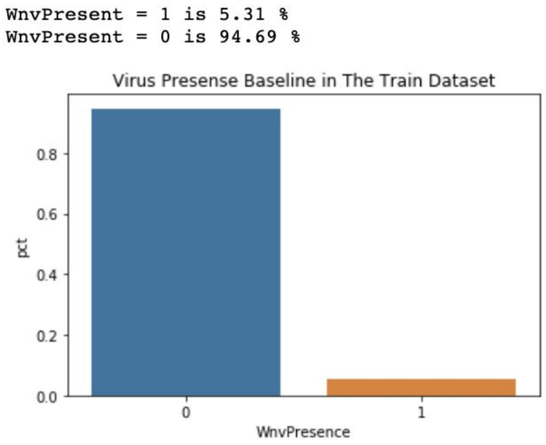
Baseline

Through visualizing the current breakdown between the WNV presence, we noticed that the training dataset was highly imbalanced.

We understand that modeling techniques will have high accuracy with low recall as a model will be biased towards 0s and predicting all not present.



```
WnvPresent
0    8153
1     457
Name: WnvPresent, dtype: int64
```



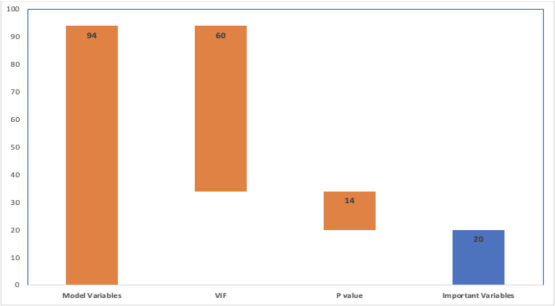
```
WnvPresent = 1 is 5.31 %
WnvPresent = 0 is 94.69 %
```

We started with a linear model and will try to work on tree-based models. The model with a combination of best accuracy and recall will be selected.

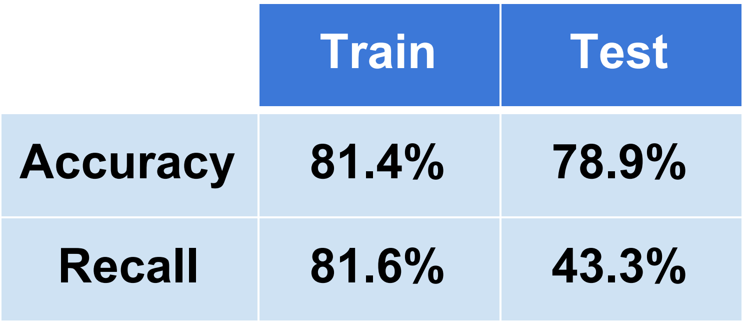
Logistic Regression

We followed the below steps in the modeling process:

- Since the dataset was highly imbalanced, we tried the SMOTE technique to balance the 1s.
- The master dataset has 94 variables that we fill will have multicollinearity, and many of them will not be significant.
- We started variable removal with keeping VIF threshold at 10, that removed 60 variables
- To check the significance of those 34 variables, we did the p-value test and kept the variables with a significance level of 0.05. Post which only 20 variables remain.
- The final model utilizes the 20 most significant variables.

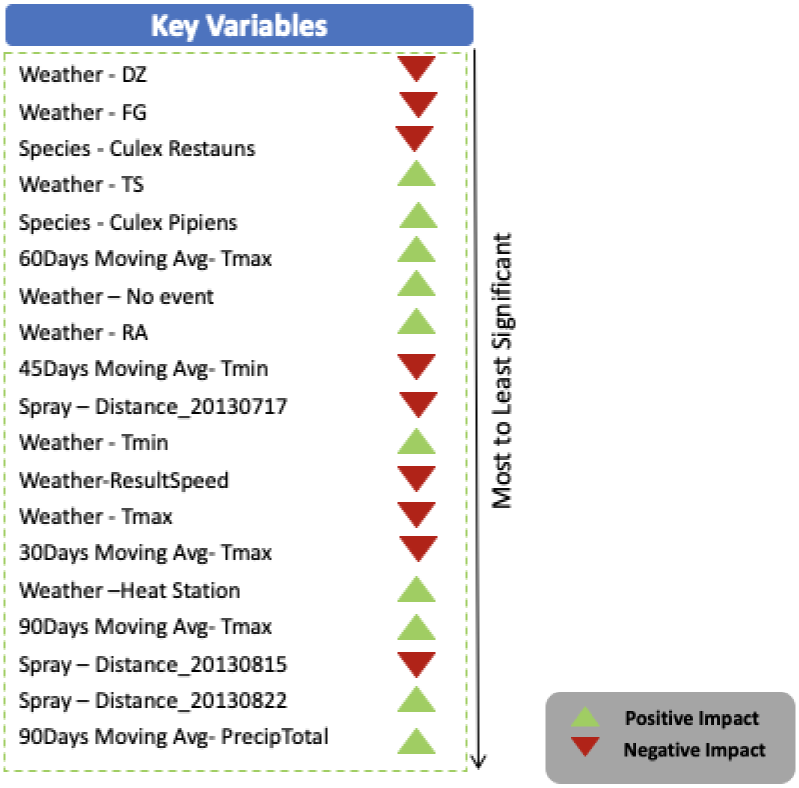


Results-



	Train	Test
Accuracy	81.4%	78.9%
Recall	81.6%	43.3%

Variable Importance



Key Variables	
Weather - DZ	▼
Weather - FG	▼
Species - Culex Restauns	▼
Weather - TS	▲
Species - Culex Pipiens	▲
60Days Moving Avg- Tmax	▲
Weather – No event	▲
Weather - RA	▲
45Days Moving Avg- Tmin	▼
Spray – Distance_20130717	▼
Weather - Tmin	▲
Weather-ResultSpeed	▼
Weather - Tmax	▼
30Days Moving Avg- Tmax	▼
Weather –Heat Station	▲
90Days Moving Avg- Tmax	▲
Spray – Distance_20130815	▼
Spray – Distance_20130822	▲
90Days Moving Avg- PrecipTotal	▲

XGBoost
XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Its name stands for **eXtreme Gradient Boosting.** XGBoost is a scalable and accurate implementation of gradient boosting machines, and it has proven to push the limits of computing power for boosted tree algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources for tree boosting algorithms.

We followed the below steps in the modeling process:

- Since the dataset was highly imbalanced, we tried the SMOTE technique to balance the 1s.
- Hyper parameter tuning for the model

Results -

- Accuracy - 92.3%
- Recall - 19%

Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction, and the class with the most votes becomes our model's prediction. A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their errors. While some trees may be wrong, many other trees will be right, so as a group, the trees can move in the correct direction.

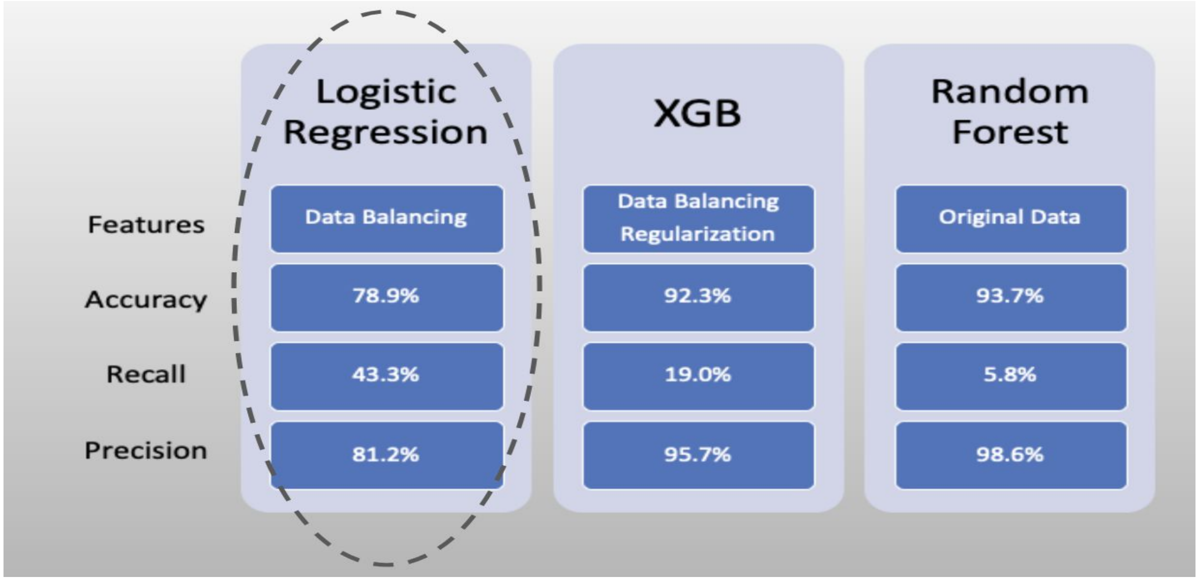
We followed the below steps in the modeling process:

- Since the dataset was highly imbalanced, we tried the SMOTE technique to balance the 1s.
- Tuning RF where the best possible parameters has been selected for the RF models

Results

- Accuracy - 93.7%
- Recall -5.8%

Final Model Selection



	Logistic Regression	XGB	Random Forest
Features	Data Balancing	Data Balancing Regularization	Original Data
Accuracy	78.9%	92.3%	93.7%
Recall	43.3%	19.0%	5.8%
Precision	81.2%	95.7%	98.6%

Based on recall and accuracy we finalize logistic regression.

Next Steps - Time Series

Due to the cyclicality of the virus appearing every summer and increasing in prevalence with heat, it was clear that the time series nature of the data was important. After extensive research into time series forecasting/prediction methods, the number one winner was the nonlinear autoregressive neural network with exogenous variables or *NARX.* Pictured below in figure 1, a NARX neural network accepts many inputs for all of the different exogenous variables, these variables flow into the hidden layer and finally into the output. A NARX architecture is a feed forward model which means that there is no back propagation. The distinguishing feature of the NARX model is that the output of the first run through is added as an input for the next run through.

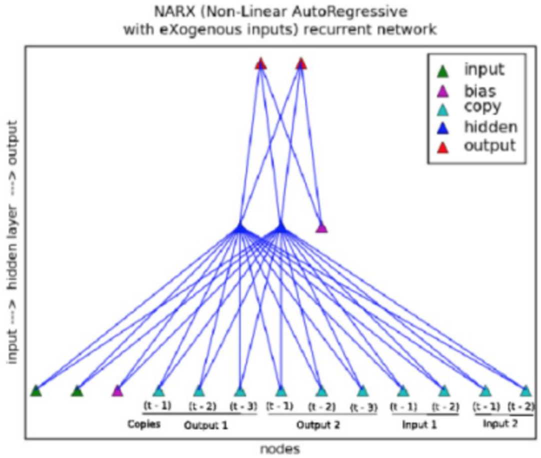


Figure 1. NARX architecture, from the bottom up, one input layer, a hidden layer, and an output layer. (Pyneurgen.sourceforge.net. (2019). *Python Neural Genetic Algorithm Hybrids*. [online] Available at: http://pyneurgen.sourceforge.net/recurrent.html [Accessed 1 Dec. 2019].)

Figure 2 shows that It is possible to add a time delay that will feed the new input through after a certain number of epochs. This ability to add outputs from one run as inputs for the next run helps preserve the memory from a previous date. The NARX architecture, with its feed forward model allows for a regressor to act upon the output before the output gets fed through as a new input. This ability is what gives the architecture the auto regressive power.

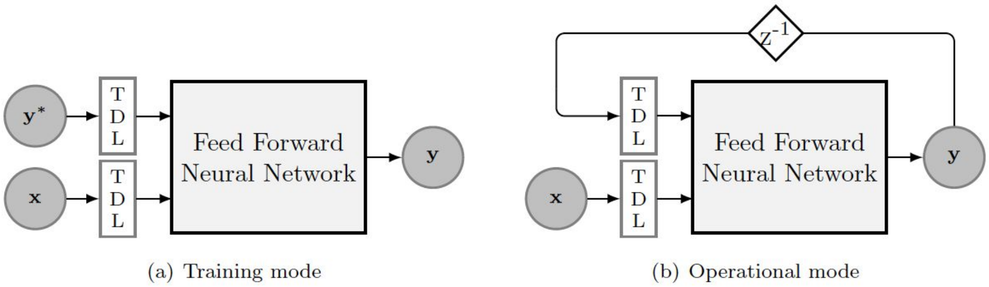


Figure 2. Time delay layer is used to feed the resulting output (z) (Maria Bianchi, F. and Maiorino, E. (2018). *An overview and comparative analysis of Recurrent Neural Networks for Short Term Load Forecasting*. [ebook] UIT Norway. Available at: https://arxiv.org/pdf/1705.04378.pdf [Accessed 1 Dec. 2019].)

Unfortunately, after extensive research into using NARX, it was found that MATLAB is the only place that this structure can be used. There is an implementation with the pyneurgen package, but with very limited documentation, the structure was not capable of being used. Therefore, the closest implementation was a Structural Time Series using tensor flow probability. This model is part of a family of models that accepts many different time series ideas as building blocks that help explain the time series trend. Each exogenous variable can be implemented using a different idea such as:

- Autoregressive Processes
- Seasonality
- Local Linear trends
- Moving Averages

With time series data, the training can not be randomized. Therefore, figure 3 shows that the training data is the first 80% of the data and the last 20% is for testing.

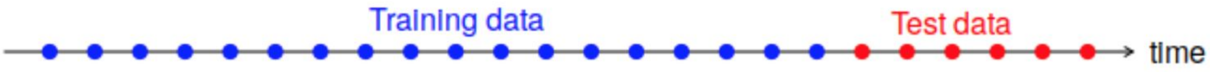
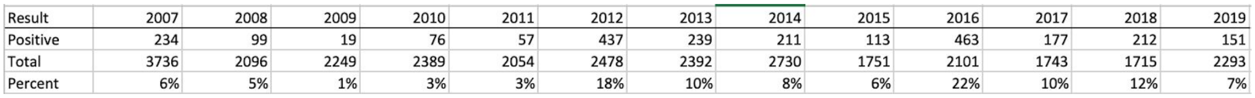


Figure 3. Training/test split for training time series model

The overall goal was to separate the where from the when. If the seasonality of the West Nile Virus numbers could be predicted, then the predicted number could be divided up between the regions that have a higher risk. This is when the spray data can be factored in to lower the risk for that region.

To begin the modeling, each of the observations throughout the city were grouped by date as well as which weather station they are closest to. This left us with two data sets, one for weather station 1 and one for weather station 2. Our data set consists of 4 years of data (2007,2009,2011,2013) and trap information for the summer and fall months. Therefore, the first three years are used for training and 2013 was used for testing. This introduces a problem, as previously discussed, there was a flood that happened in April that affected the West Nile Virus numbers for that year. Therefore, the testing of this method, or any time series model, is going to appear to not fit the trend. To prove that this model is actually a viable option, additional data was collected from the Chicago government that contains all of the trap data from 2007-2019, but not the weather or spray efforts. According to figure 4, 2013 is above the average of 8% and the training for the model was based off of the years that had the lowest percentages of all of the recorded years.

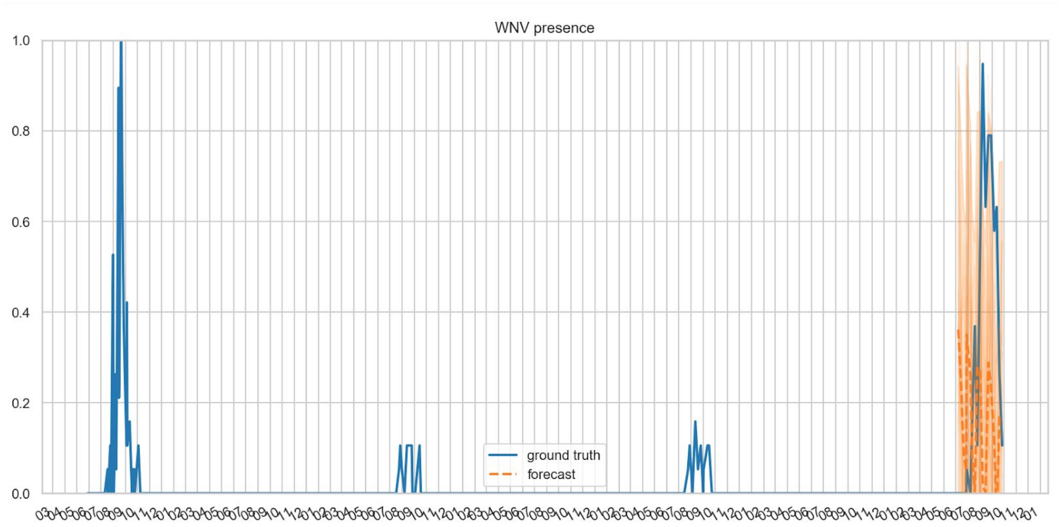


Result	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Positive	234	99	19	76	57	437	239	211	113	463	177	212	151
Total	3736	2096	2249	2389	2054	2478	2392	2730	1751	2101	1743	1715	2293
Percent	6%	5%	1%	3%	3%	18%	10%	8%	6%	22%	10%	12%	7%

Figure 4. Percentage of positive pools to the total number tested for 2007-2019

For station 1 as well as station 2, the x axis is the recorded observations on each date during the years 2007, 2009, 2011, and 2013. The x axis ticks are the month numbers during those years. It is important to point out that the traps are pools that either contain the virus or not, therefore, the ratio of infected mosquitos versus not infected mosquitos wasn't measured. This means that the y axis is a normalized value that shows the total number of infected pools recorded during that date. The blue is the actual observed numbers and the orange dashed lines are the forecasted numbers. The orange shaded region is an error region of the mean forecasted result multiplied by 2. The results for station 1 and station 2 are as follows:

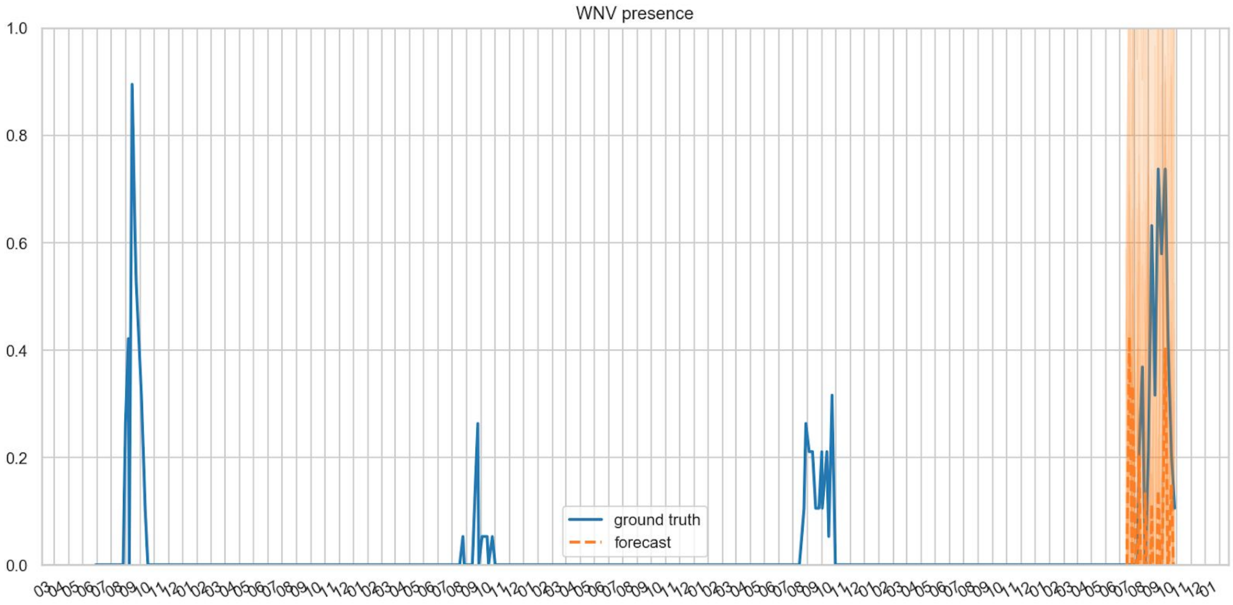
Station 1



```
Inferred parameters:
observation_noise_scale: 0.23898877203464508 +- 0.017843443900346756
temperature_effect/_weights: [-1.4840991] +- [0.13810897]
temperature_max_effect/_weights: [0.70503324] +- [0.13726136]
temperature_min_effect/_weights: [1.0079901] +- [0.14137842]
avg_temp_effect/_drift_scale: 0.020652975887060165 +- 0.010815059766173363
Depart_effect/_drift_scale: 0.001159940380603075 +- 0.00092834368115291
Dew_point_effect/_drift_scale: 0.001665996853262186 +- 0.002860200824216008
Wet_bulb_effect/_drift_scale: 0.0028124512173235416 +- 0.0030560444574803114
Preciptotal_effect/_weights: [2.1972077] +- [2.8792672]
Sea_Level_effect/_drift_scale: 0.0007651039632037282 +- 0.002829749835655093
rel_hum_effect/_drift_scale: 0.0010438390308991075 +- 0.0015077651478350163
num_mosquitos_effect/_drift_scale: 0.007395820692181587 +- 0.014804667793214321
```

NRMSE = 0.3932827

Station 2



```
Inferred parameters:
observation_noise_scale: 0.009671381674706936 +- 0.008544856682419777
temperature_effect/_weights: [-0.297668] +- [0.11473513]
temperature_max_effect/_weights: [0.08619052] +- [0.09613184]
temperature_min_effect/_weights: [0.70601225] +- [0.13693407]
avg_temp_effect/_drift_scale: 0.006463347002863884 +- 0.005139876157045364
Depart_effect/_drift_scale: 0.07947887480258942 +- 0.08370354771614075
Dew_point_effect/_drift_scale: 0.08316729962825775 +- 0.0724836140871048
Wet_bulb_effect/_drift_scale: 0.008678548969328403 +- 0.011841749772429466
Preciptotal_effect/_weights: [10.514013] +- [4.2411613]
Sea_Level_effect/_drift_scale: 0.43226033449172974 +- 0.046366021037101746
rel_hum_effect/_drift_scale: 0.028159229084849358 +- 0.03830115497112274
num_mosquitos_effect/_drift_scale: 0.0021205577068030834 +- 0.005195014178752899
```

NRMSE = 0.40466395

The inferred parameters are similar as expected among the two weather stations, but the main difference is apparent with the preciptotal_effect. The effect for station 1 was 2.19 +- 2.87 and for station 2 was 10.51 +- 4.24. The large effect of the total rain makes sense because mosquitoes need water in order to lay their eggs, so the more rain, the more mosquitoes. The problem is the scale difference between the two stations. This can be attributed to the fact that there were 10 more dates (94 dates where traps were read for station 2) than station 1 (84 dates). As time series data goes, this is a relatively small sample. The following effects were used to build the structural time series model:

- Temperature_effect
- Temperature_max_effect
- Temperature_min_effect
- avg_temp_effect
- Depart_effect
- Dew_point_effect
- Wet_bulb_effect
- Preciptotal_effect
- Sea_Level_effect
- rel_hum_effect
- num_mosquitos_effect

Overall, the structural time series method is a great way to capture the seasonality of some of the variables, remember the past results, all the while addressing each of the variables with a different method due to the different distributions. This analysis covered the forecast of number of cases expected. To predict where the virus would appear, a kernel density estimation technique would be used to determine the likely distributions given a finite set of data. With such a small sample of positive West Nile Virus, the density estimation could smooth the probability for each location to determine it's actual likelihood of containing the virus. Once determined, you would factor in the relative spray distance if the region had been sprayed, as well as how recently. These factors would reduce the peak of the density curve and limit the probability that that specific region would contain a positive test for the West Nile Virus. No meaningful results could be drawn from the estimation due to the limited positive results and underestimated forecast.

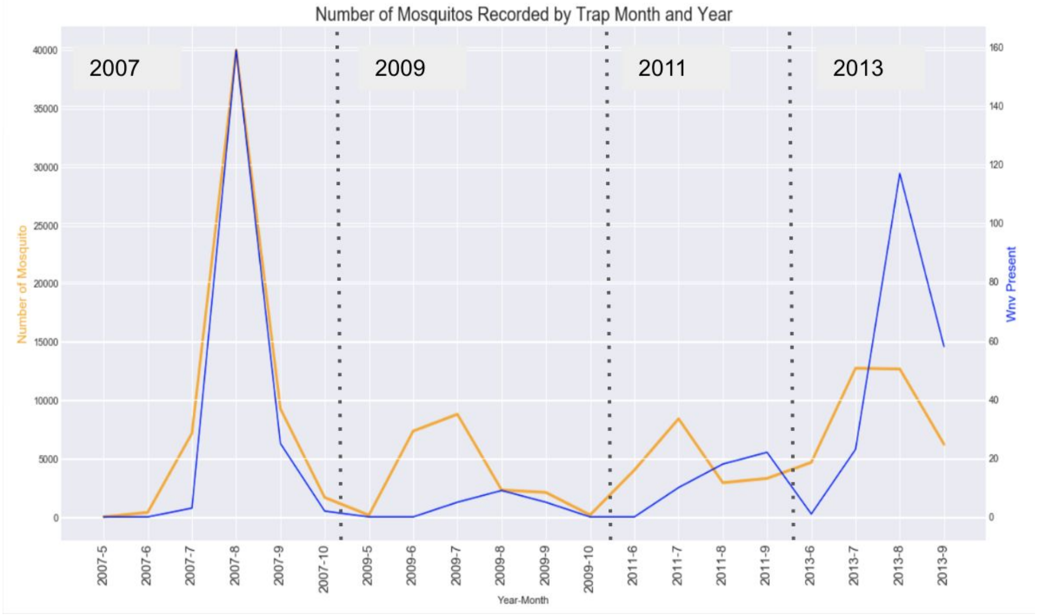
Key Learnings and Takeaways

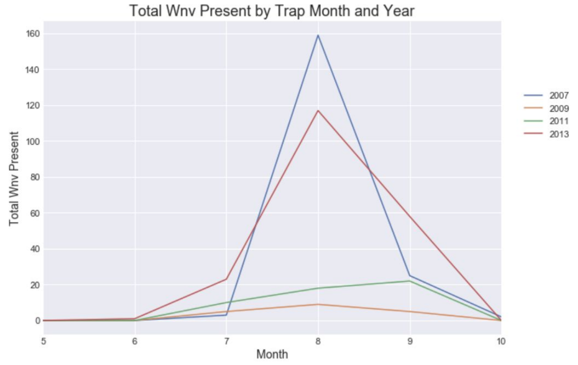
Despite the fact that we had the highest accuracy when working with random forest algorithm, we decided to use logistic regression because we understand that accuracy may or may not be a good measurement to judge models. To achieve our goal, instead of looking at accuracy, we needed to focus on recall because the higher recall rate, the lower number of false-negative in the model, which it does significantly matter in our project goal.

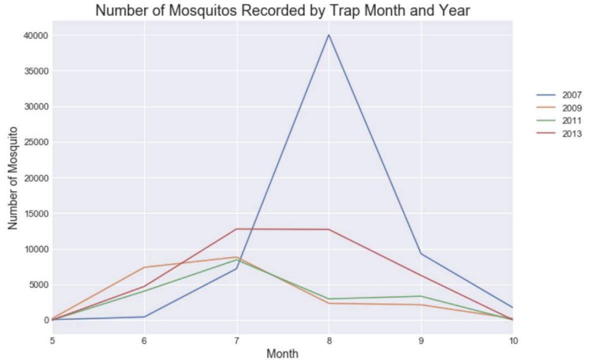
1. Accuracy is not the best measure of assessing the performance of the model.
2. Having context background is key
3. Assess your prediction based on the context knowledge
4. It is essential to look at the big picture to impact bigger scope
5. Most importantly, understanding the main objective is the first step

Appendix

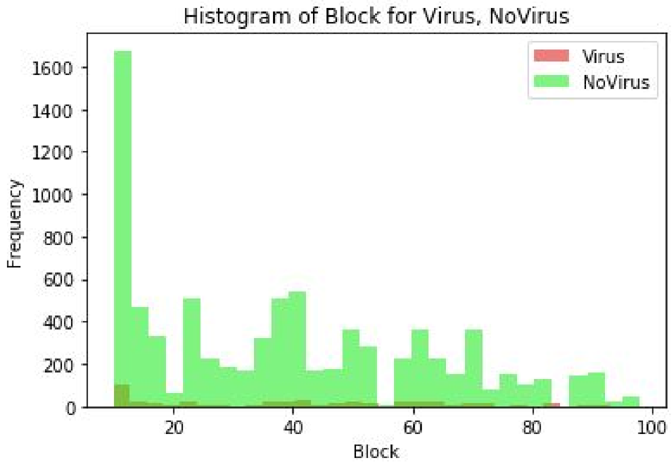
Appendix I - Mosquito Distribution by Month and Year



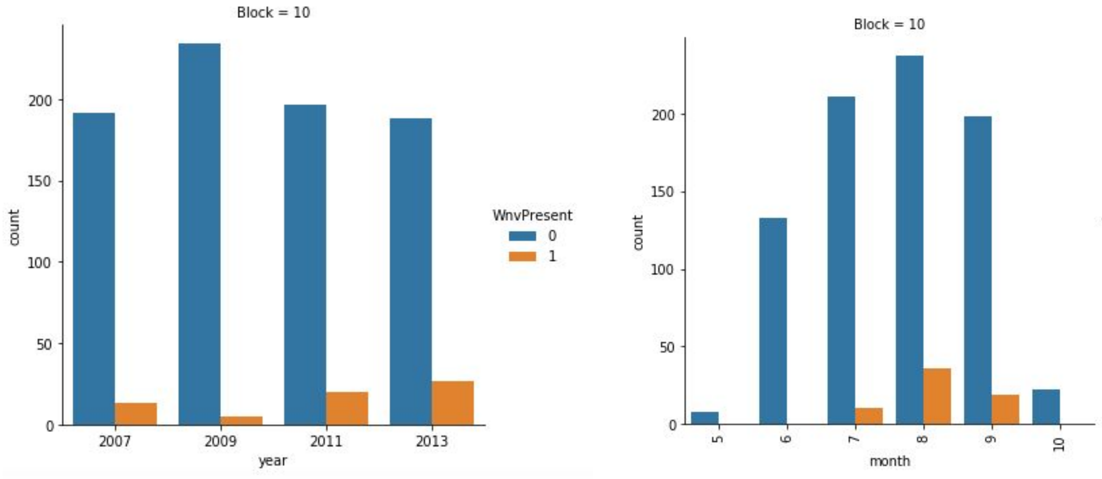




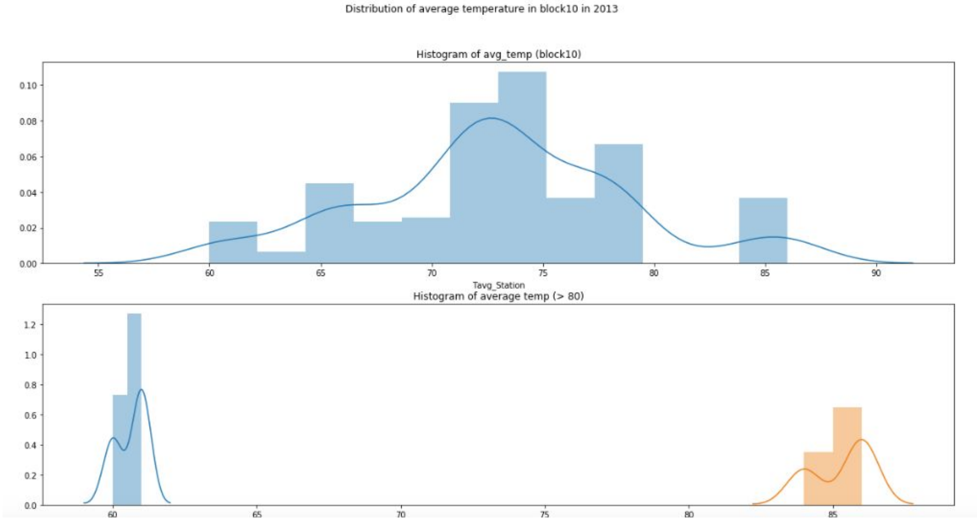
Appendix IIa - Virus Presence in 2013 by Block

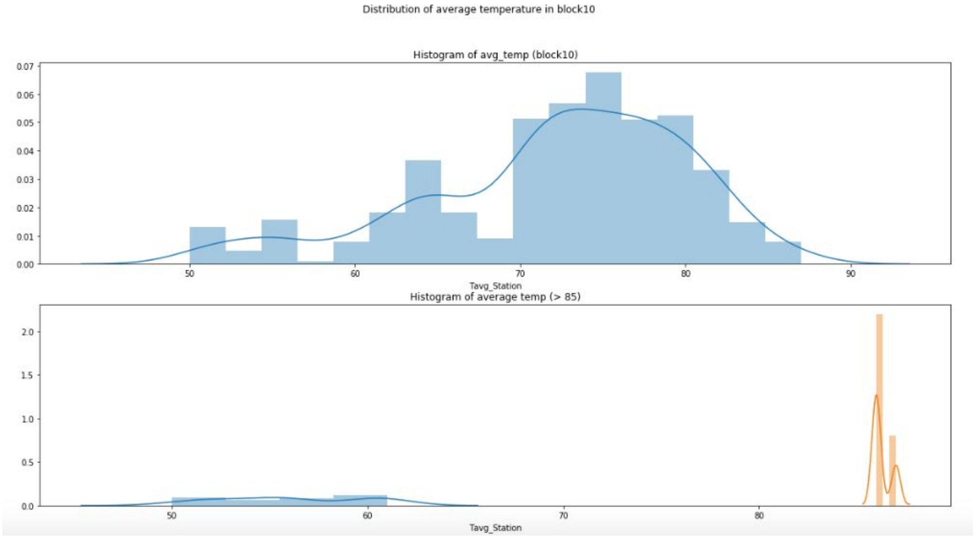


Appendix IIb - Virus Presence in Block 10 by Year and Month

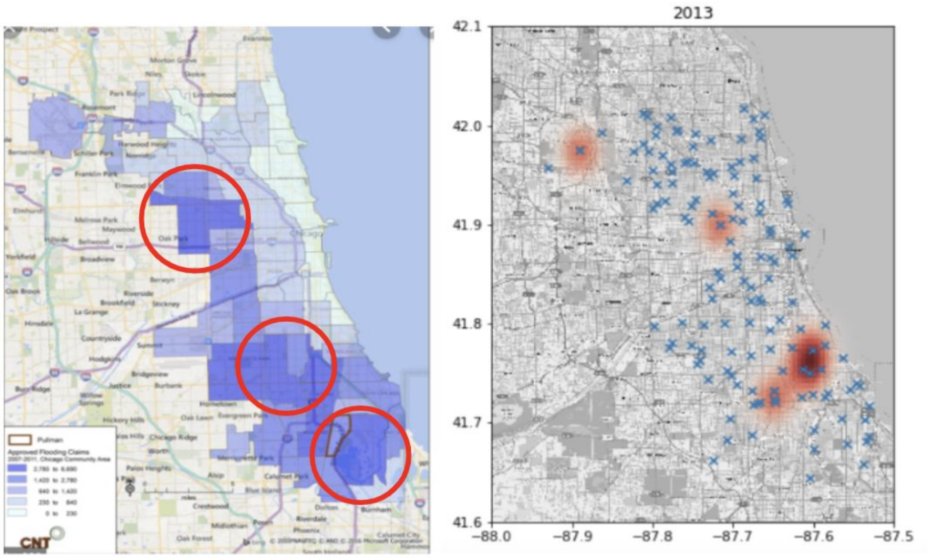


Appendix IIc - Average Temperature at Block 10 in 2013

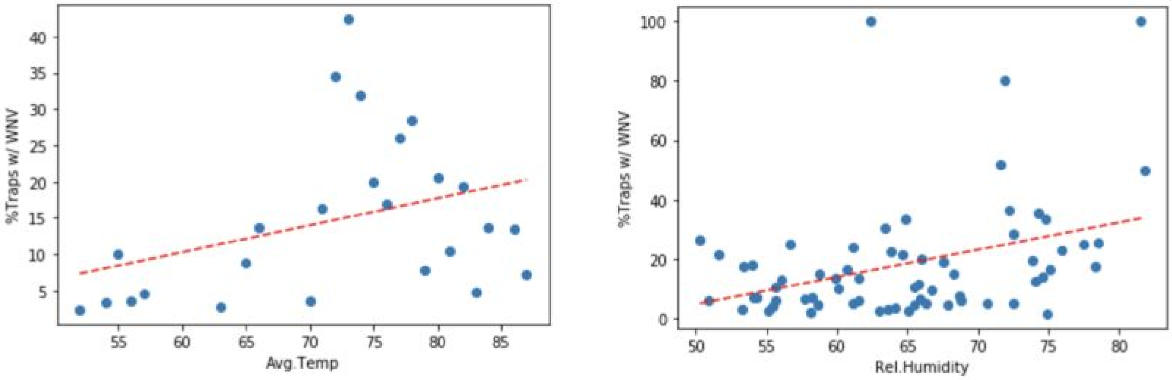




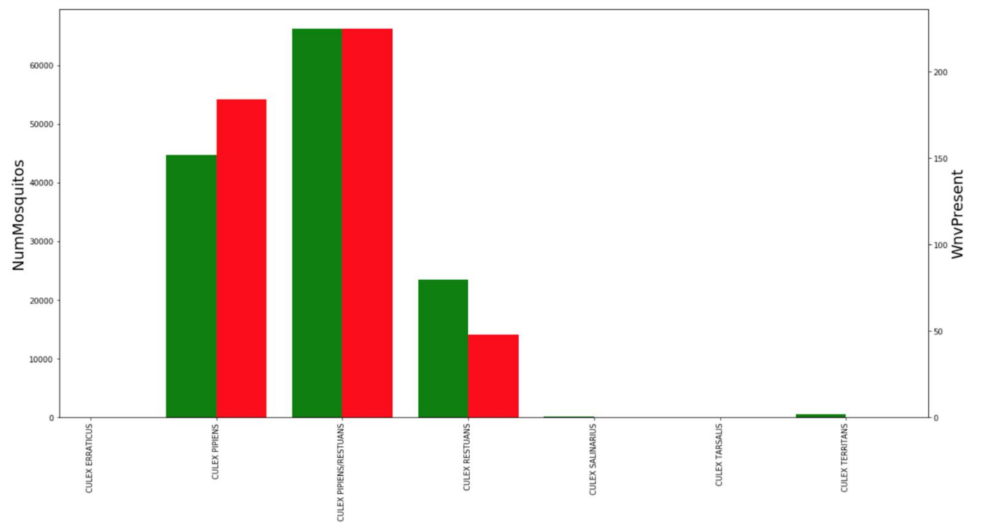
Appendix III - Areas Damaged by Flood vs Areas with WNV Presence

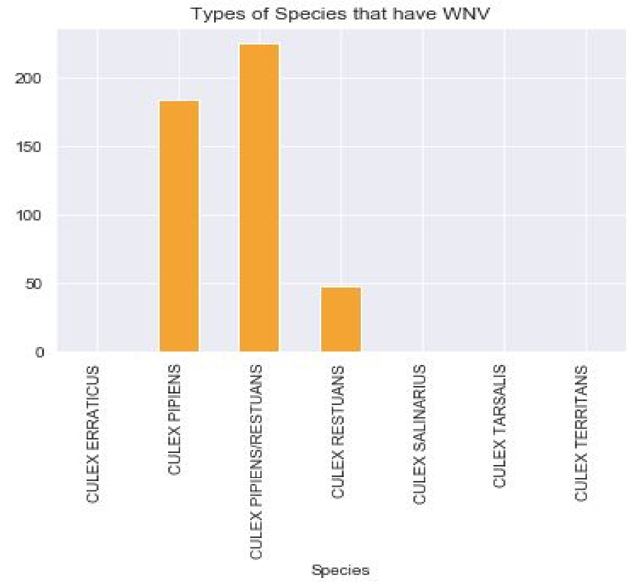


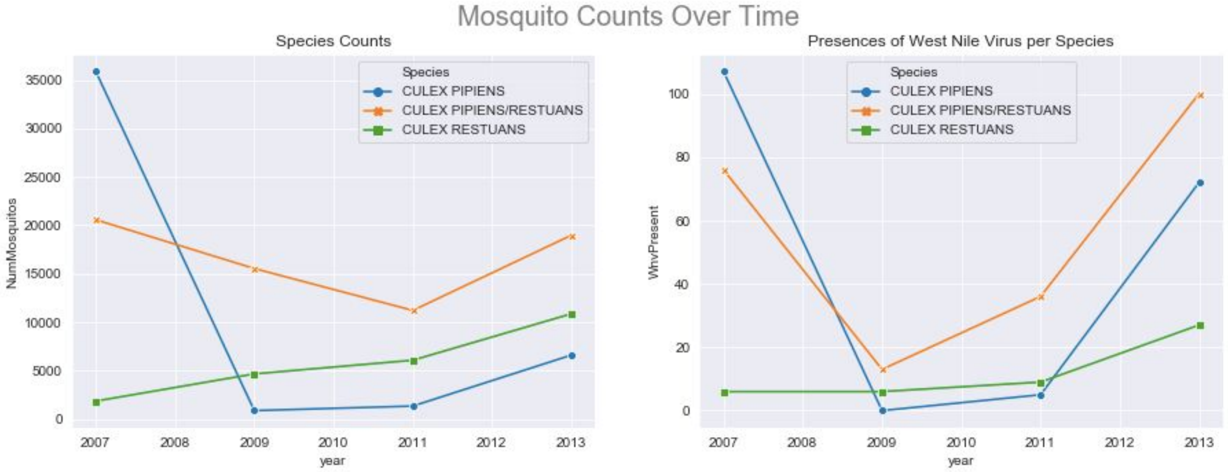
Appendix IV - Impact of Temperature and Relative Humidity on Virus Presence



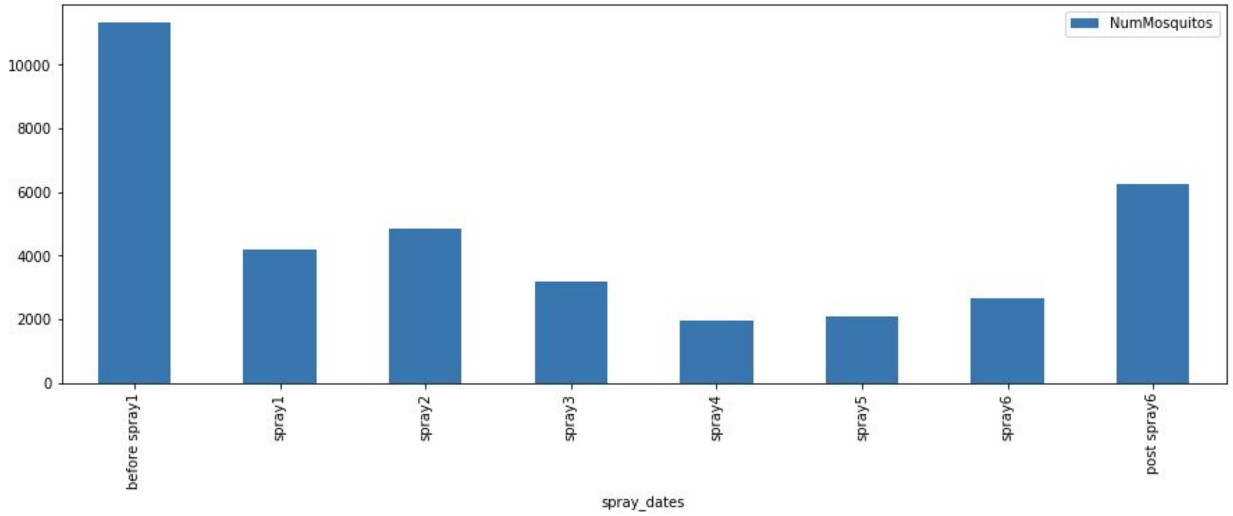
Appendix V - Mosquito Species with West Nile Virus





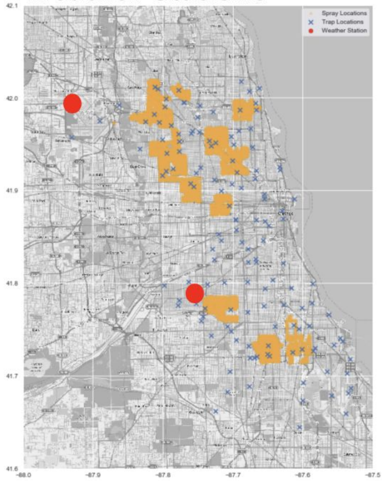


Appendix V - Effect of sprays on number of mosquitos

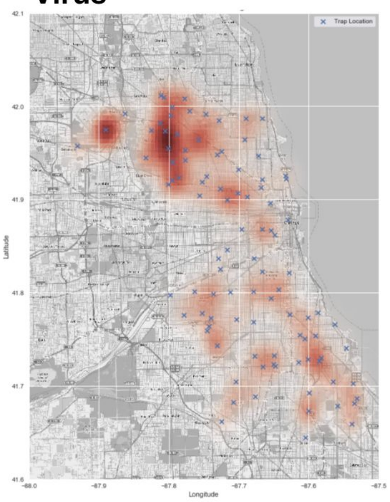


Appendix V -Virus Presence vs Spray, Trap & Weather Locations

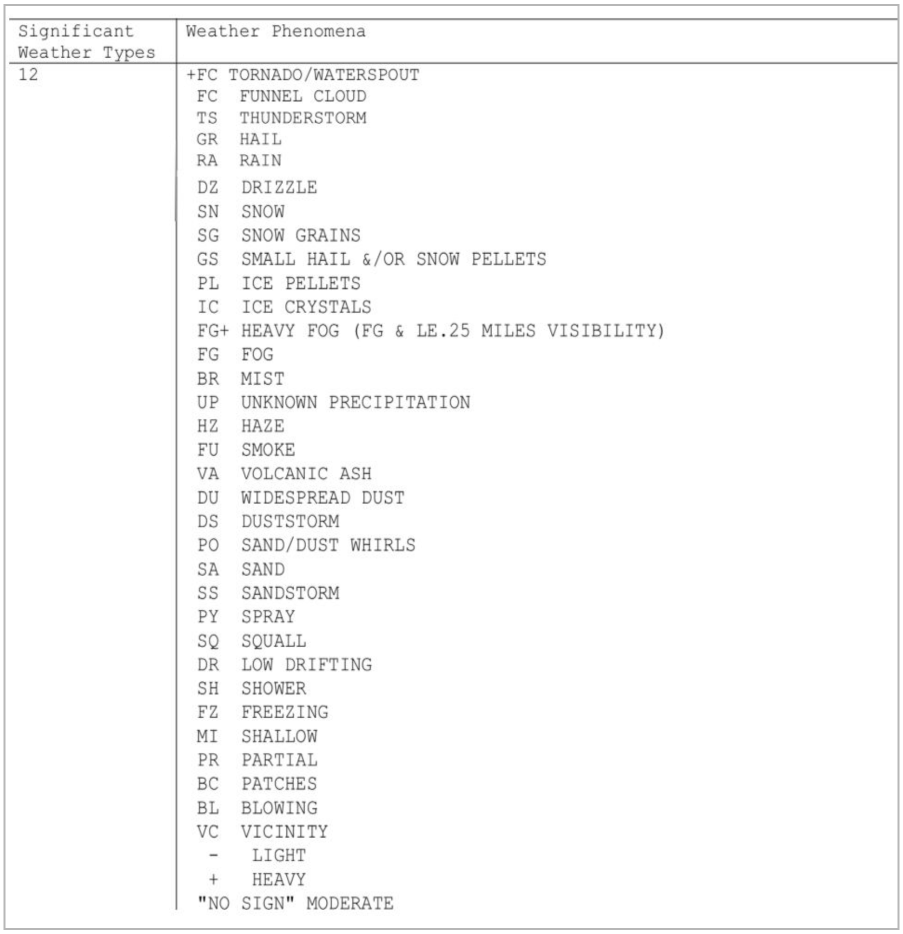
Spray, Trap and Weather Stations



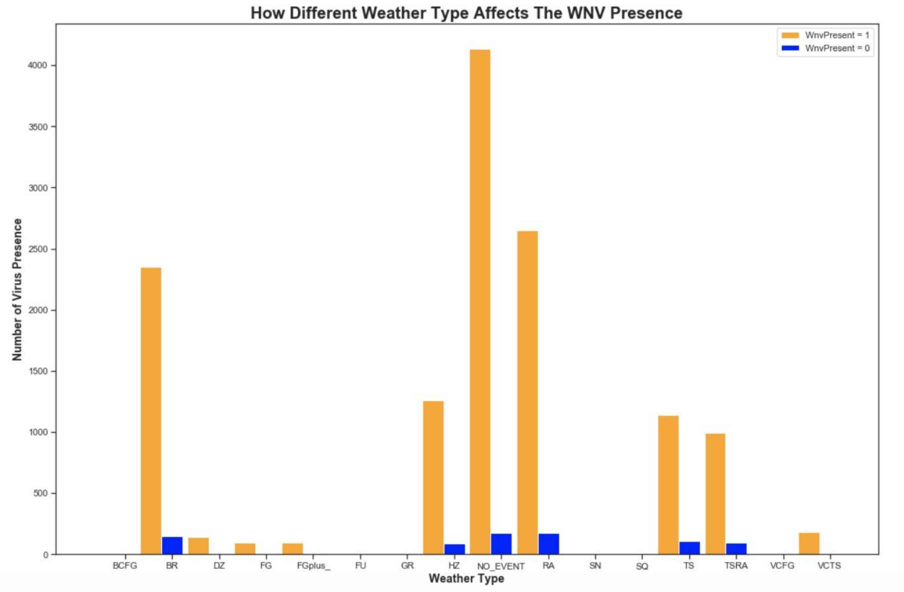
Presence of West Nile Virus



Appendix VI - Weather Type vs Virus Presence



Significant Weather Types	Weather Phenomena
12	+FC TORNADO/WATERSPOUT
	FC FUNNEL CLOUD
	TS THUNDERSTORM
	GR HAIL
	RA RAIN
	DZ DRIZZLE
	SN SNOW
	SG SNOW GRAINS
	GS SMALL HAIL &/OR SNOW PELLETS
	PL ICE PELLETS
	IC ICE CRYSTALS
	FG+ HEAVY FOG (FG & LE.25 MILES VISIBILITY)
	FG FOG
	BR MIST
	UP UNKNOWN PRECIPITATION
	HZ HAZE
	FU SMOKE
	VA VOLCANIC ASH
	DU WIDESPREAD DUST
	DS DUSTSTORM
	PO SAND/DUST WHIRLS
	SA SAND
	SS SANDSTORM
	PY SPRAY
	SQ SQUALL
	DR LOW DRIFTING
	SH SHOWER
	FZ FREEZING
	MI SHALLOW
	PR PARTIAL
	BC PATCHES
	BL BLOWING
	VC VICINITY
	- LIGHT
	+ HEAVY
	"NO SIGN" MODERATE


How Different Weather Type Affects The WNV Presence
WnvPresent = 1
WnvPresent = 0
Number of Virus Presence
4000
3500
3000
2500
2000
1500
1000
500
0
BCFG
BR
DZ
FG
FGplus_
FU
GR
HZ
NO_EVENT
RA
SN
SQ
TS
TSRA
VCFG
VCTS
Weather Type

References

Jiafeng Wang, Nick H. Ogden, Huaiping Zhu, The Impact of Weather Conditions on *Culex pipiens* and *Culex restuans* (Diptera: Culicidae) Abundance: A Case Study in Peel Region, *Journal of Medical Entomology*, Volume 48, Issue 2, 1 March 2011, Pages 468–475, https://doi.org/10.1603/ME10117

West Nile Virus Activity --- United States, 2009. (2019). Retrieved 13 December 2019, from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5925a1.htm

1. Reimann CA, Hayes EB, DiGuiseppi C, et al. Epidemiology of neuroinvasive arboviral disease in the United States, 1999--2007. Am J Trop Med Hyg 2008;79:974--9.
2. CDC. Surveillance for human West Nile virus disease---United States, 1999--2008. Surveillance Summaries, April 2, 2010.MMWR 2010;59(No. SS-2).
3. Mostashari F, Bunning ML, Kitsutani PT, et al. Epidemic West Nile encephalitis, New York, 1999: results of a household-based seroepidemiological survey. Lancet 2001;358:261--4.
4. Komar N. West Nile virus: epidemiology and ecology in North America. Adv Virus Res 2003;61:185--234.
5. Hayes EB, Komar N, Nasci RS, Montgomery SP, O'Leary DR, Campbell GL. Epidemiology and transmission dynamics of West Nile virus disease. Emerg Infect Dis 2005;11:1167--73.

Kaggle.com. (2019). *West Nile Virus Prediction | Kaggle*. [online] Available at: https://www.kaggle.com/c/predict-west-nile-virus/data [Accessed 13 Dec. 2019].

Extremes | National Centers for Environmental Information (NCEI). (2019). Retrieved 13 December 2019, from https://www.ncdc.noaa.gov/extremes/