Prediction of West Nile Virus in Chicago

Abstract

A machine learning model is developed in order to predict West Nile virus in the city of Chicago. The XGBoost Classifier was selected and its performance was compared to that of a Random Forest Classifier. The final model is able to predict the presence with 96% accuracy and the absence of virus with 44% accuracy. In this case, it is considered acceptable since being accurate on the presence of virus is more important than the absence of virus.

Introduction

The purpose of this project is to expedite and optimize resource allocation by predicting when and where mosquitos carrying West Nile virus will be found. To do this, data regarding weather conditions, locations, and positive cases is utilized. This data is provided by the Chicago Department of Public Health and available publicly on Kaggle.com.

Exploratory Analysis

First it is important to see which locations have the most cases. Using the folium library, the following heatmap was generated.

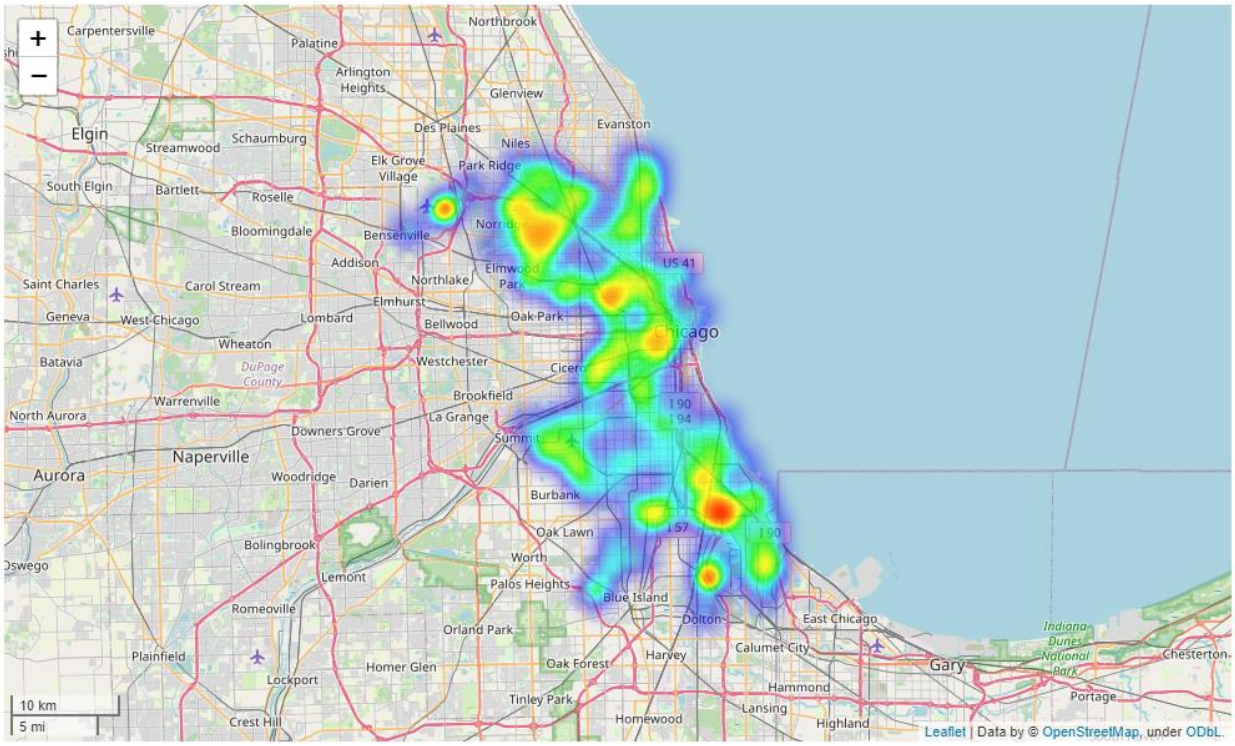

Figure 1: heatmap of collected samples

The heatmap seen in figure 1 shows the number of samples taken throughout the locations within the dataset. However, this does not show the number of positives, which is more important information. For this, a second heatmap was created which shows only samples positive for the virus.

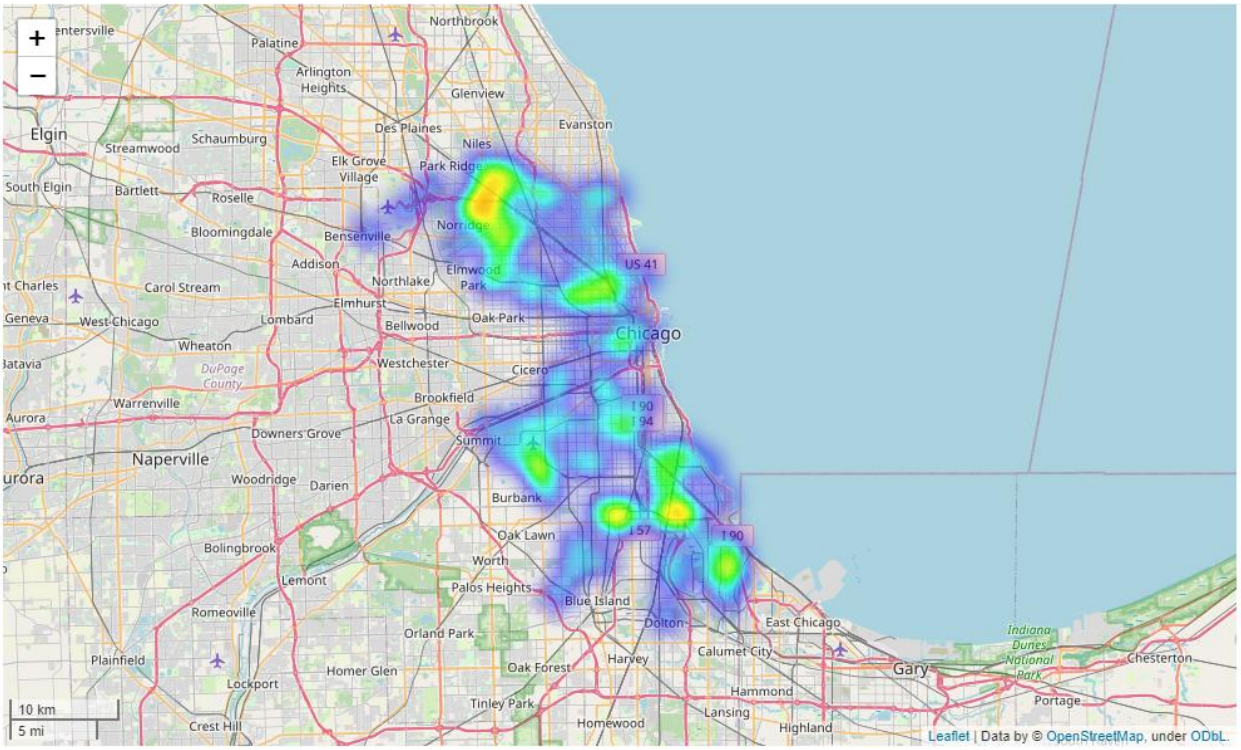

Figure 2: heatmap of positive samples

There appear to be a large concentration of positive samples in the north of the city as well as one smaller area in the south of the city. These heatmaps are further reinforced by positivity rates seen in the dataset.

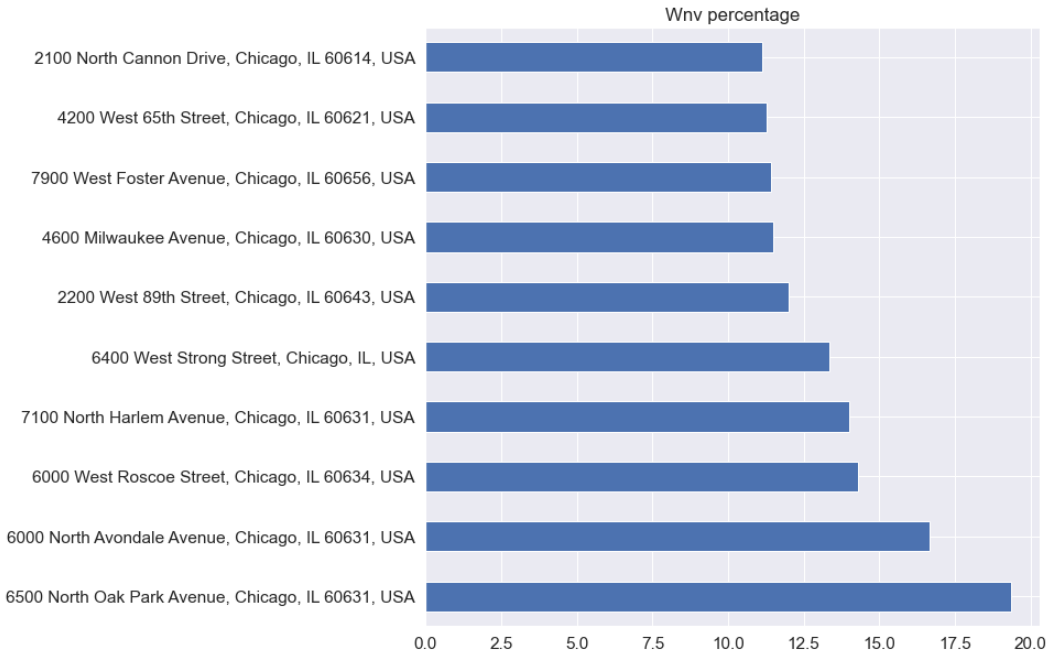

Figure 3: Locations with highest positive rates

Figure 3 reinforces the map shown in figure 2. Two locations have a significantly higher positive rate while the rest have a similar rate throughout. This is shown on the heatmap as similar, cool colors, whereas the clusters are depicted as warmer yellows.

Next, the data presents the presence of several species of mosquito. These can also be plotted on a map. There are seven species of mosquito which are sampled throughout the city of Chicago. These species are then plotted based on the location the sample was taken.

Species by location

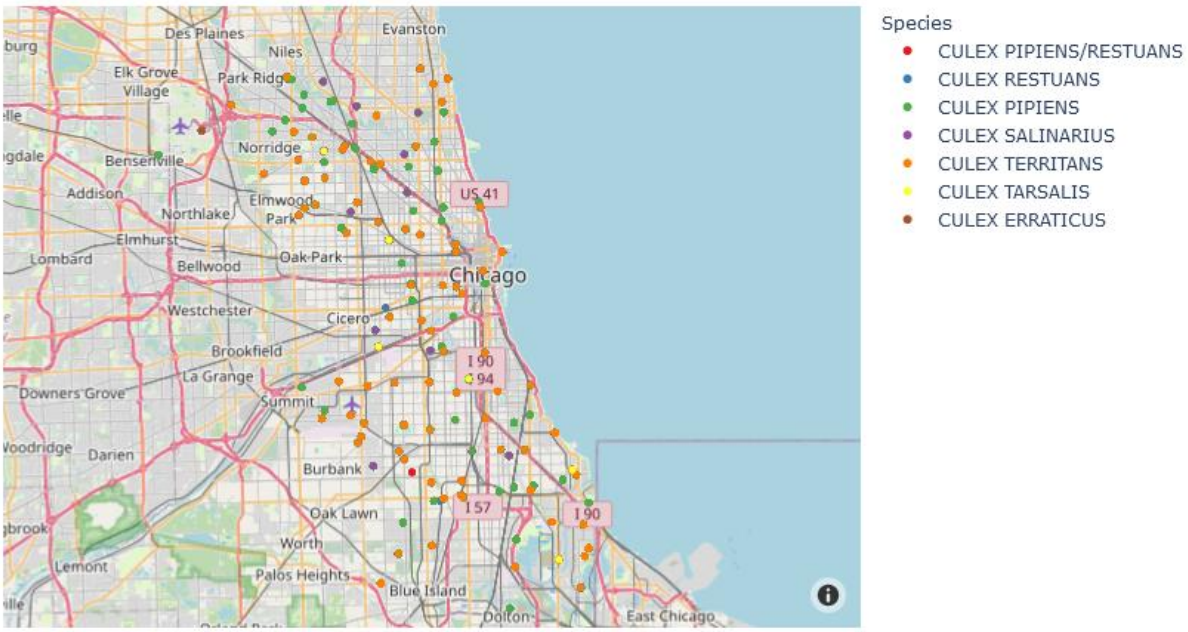

Figure 4: species by location

	Species	Sample	Wnv_sample	Wnv_percent
2	CULEX PIPIENS	2699	240	8.892182
0	CULEX PIPIENS/RESTUANS	4752	262	5.513468
1	CULEX RESTUANS	2740	49	1.788321
3	CULEX TERRITANS	222	0	0.000000
4	CULEX SALINARIUS	86	0	0.000000
5	CULEX TARSALIS	6	0	0.000000
6	CULEX ERRATICUS	1	0	0.000000

Figure 5: species by sample and positive rate

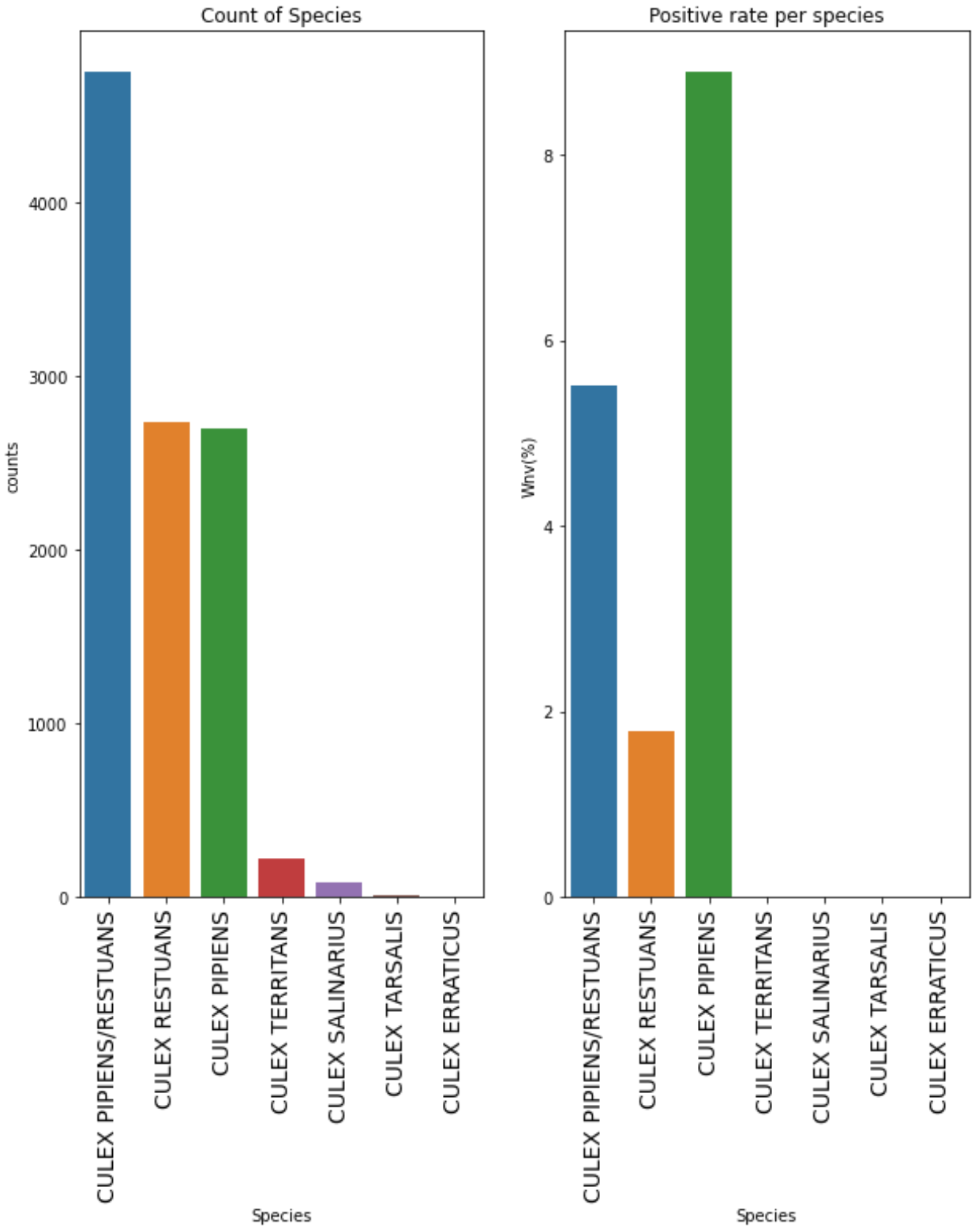

Figure 6: species count and positivity rate

As figures 5 and 6 show, of the seven species, the *Culex Pipiens* species of mosquito has the highest positivity rate followed by the *Culex Pipiens/Restuans* species. In fact, four of the seven species had no positive samples. However, it must also be noted that these species have significantly lower number of samples taken. Therefore, given the dataset, the possibility exists that they have no positive cases due to the low sample size in those species. With this information, an updated map can be generated which only shows species with positive samples.

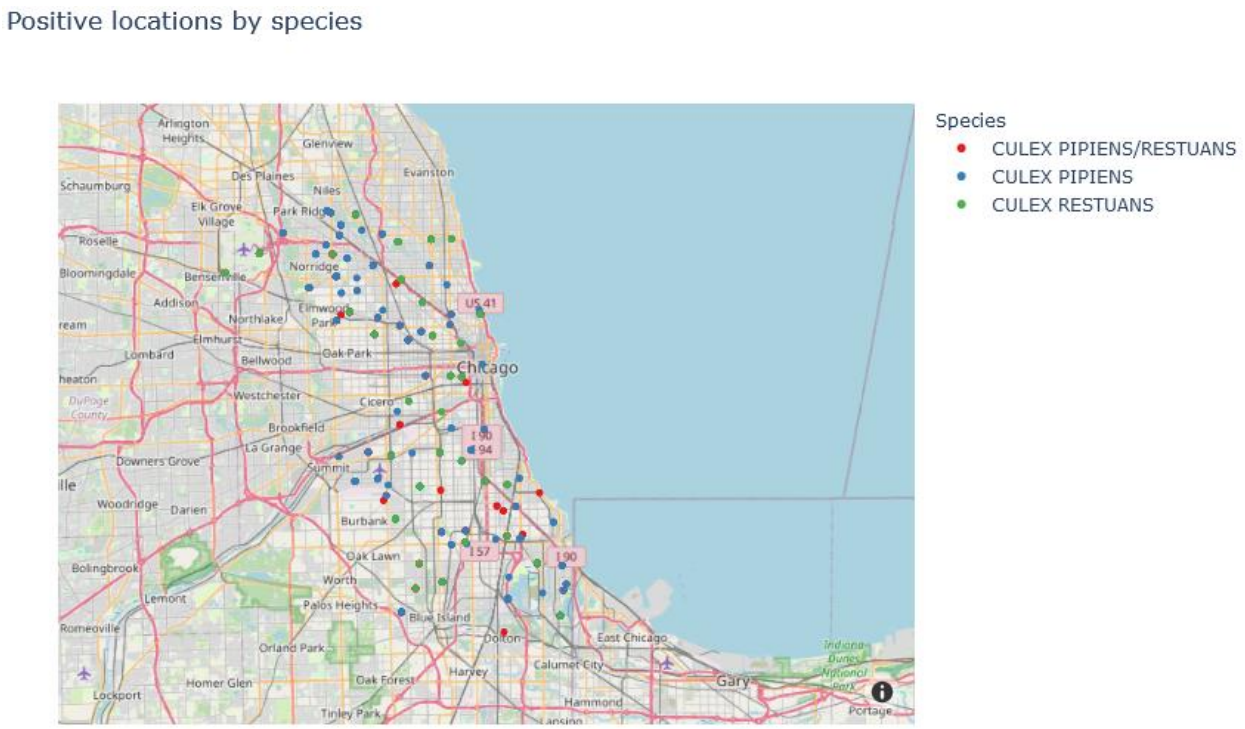

Figure 7: positive species by location

Now that the species carrying West Nile virus (WNV) have been identified, it is possible to find the positivity rate of mosquitoes found in Chicago. Figure 7 shows that only 5.2% of samples taken in Chicago were positive for West Nile virus.

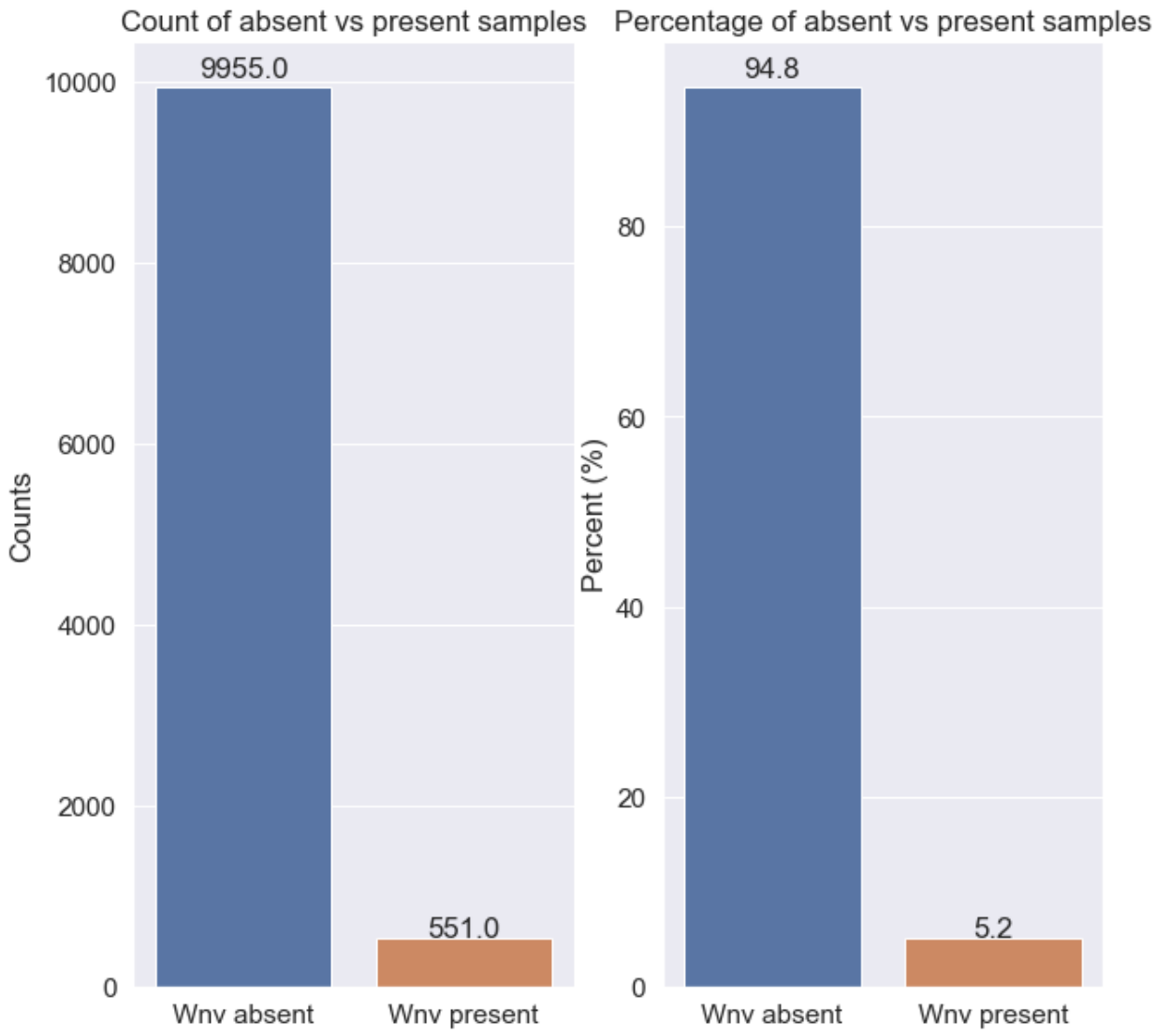

Figure 8: samples and positivity rate

Weather data was also included in the dataset provided by Chicago Department of Public Health. The following correlation matrix is generated from the dataset.

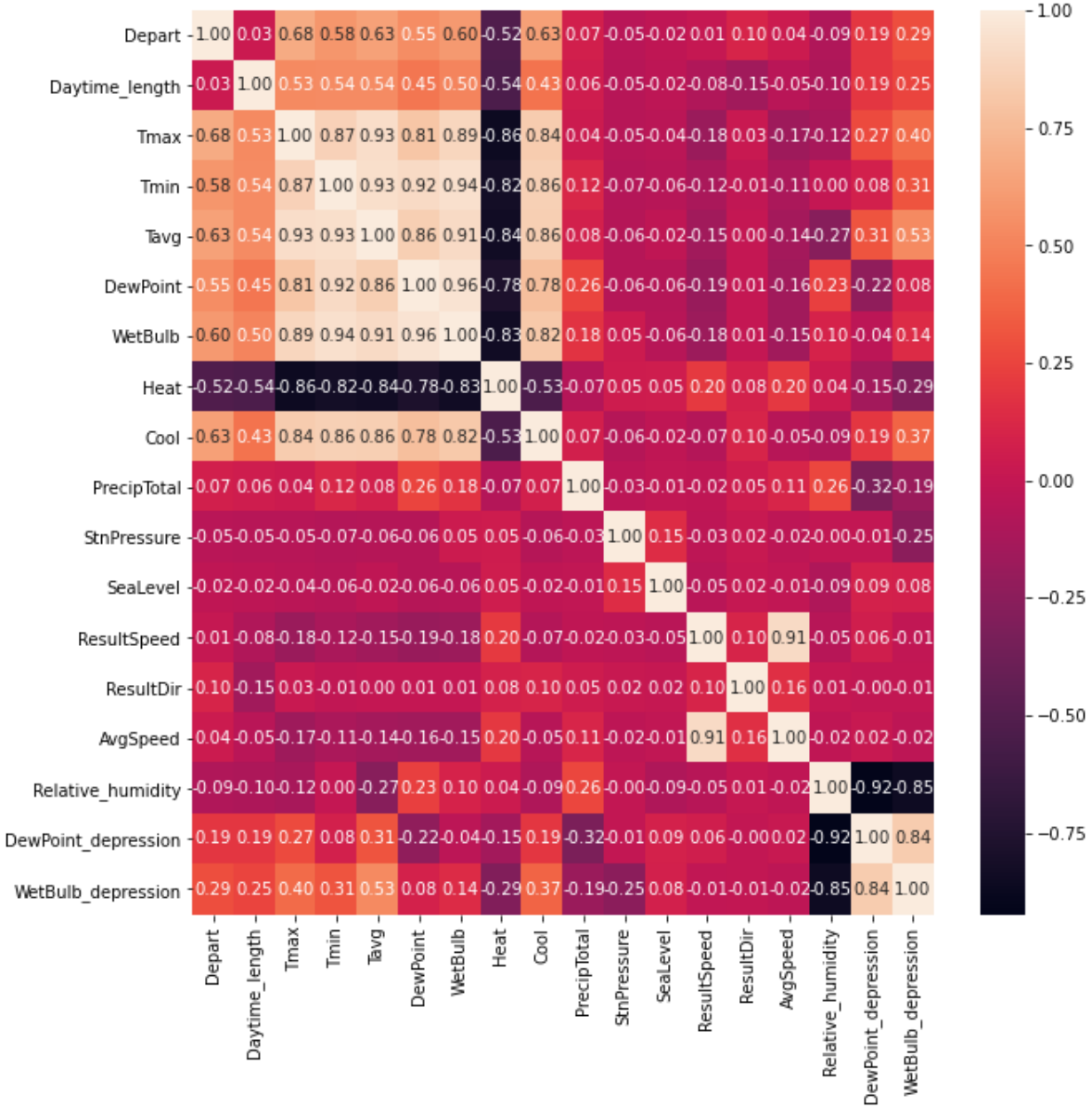

Figure 9: correlation matrix of weather data

Following this, it is now possible to differentiate data based on yearly data, such as season, month, and even week. Figures 10, 11, and 12 present the data in yearly, seasonal, and monthly intervals, respectively. Figure 13 then presents data based on year and season.

	year	sample	WnvSample	WnvPerc
0	2007	3811	236	6.192600
1	2013	2392	239	9.991639
2	2009	2249	19	0.844820
3	2011	2054	57	2.775073

Figure 10: yearly data

	season	sample	WnvSample	WnvPerc
0	Summer	7928	424	5.348133
1	Fall	2494	127	5.092221
2	Spring	84	0	0.000000

Figure 11: seasonal data

	total_count	Wnv_count	WnvPerc
Aug	3751	377	10.050653
Sep	2218	125	5.635708
July	2606	46	1.765157
Oct	276	2	0.724638
June	1571	1	0.063654
May	84	0	0.000000

Figure 12: monthly data

	year	season	sample	WnvSample	WnvPerc
0	2007	Fall	985	30	3.045685
1	2007	Spring	25	0	0.000000
2	2007	Summer	2801	206	7.354516
3	2009	Fall	483	5	1.035197
4	2009	Spring	59	0	0.000000
5	2009	Summer	1707	14	0.820152
6	2011	Fall	540	22	4.074074
7	2011	Summer	1514	35	2.311757
8	2013	Fall	486	70	14.403292
9	2013	Summer	1906	169	8.866737

Figure 13: yearly-seasonal data

As shown in figure 11, the majority of positive samples are taken during the summer. However, it should be noted that the percentage of positive cases in the summer and the spring were roughly the same at 5%. This implies, that further study may be required with more samples being taken in both the summer and spring. Figure 13 further elaborates that resources are mostly devoted to taking samples during the summer. However, according to figure 11, it may be worthwhile to expand resource usage during the fall in order to determine whether the similarity in positivity rate between summer and fall is factual or merely a coincidence.

With exploratory analysis complete, it is now possible to begin modeling. First a model based on the random forest classifier is designed in order to provide comparison with the XGBoost (XGB) classifier. The random forest model is able to predict the absence of virus with 97% accuracy and the presence of virus with only 23% accuracy. This is an unacceptable outcome, since the cost of a false negative is very high. In this case, it results in a person who is infected, not knowing they are infected. This, in turn, could result in furthering the spread of disease. Figure 14 shows the confusion matrix for the random forest model.

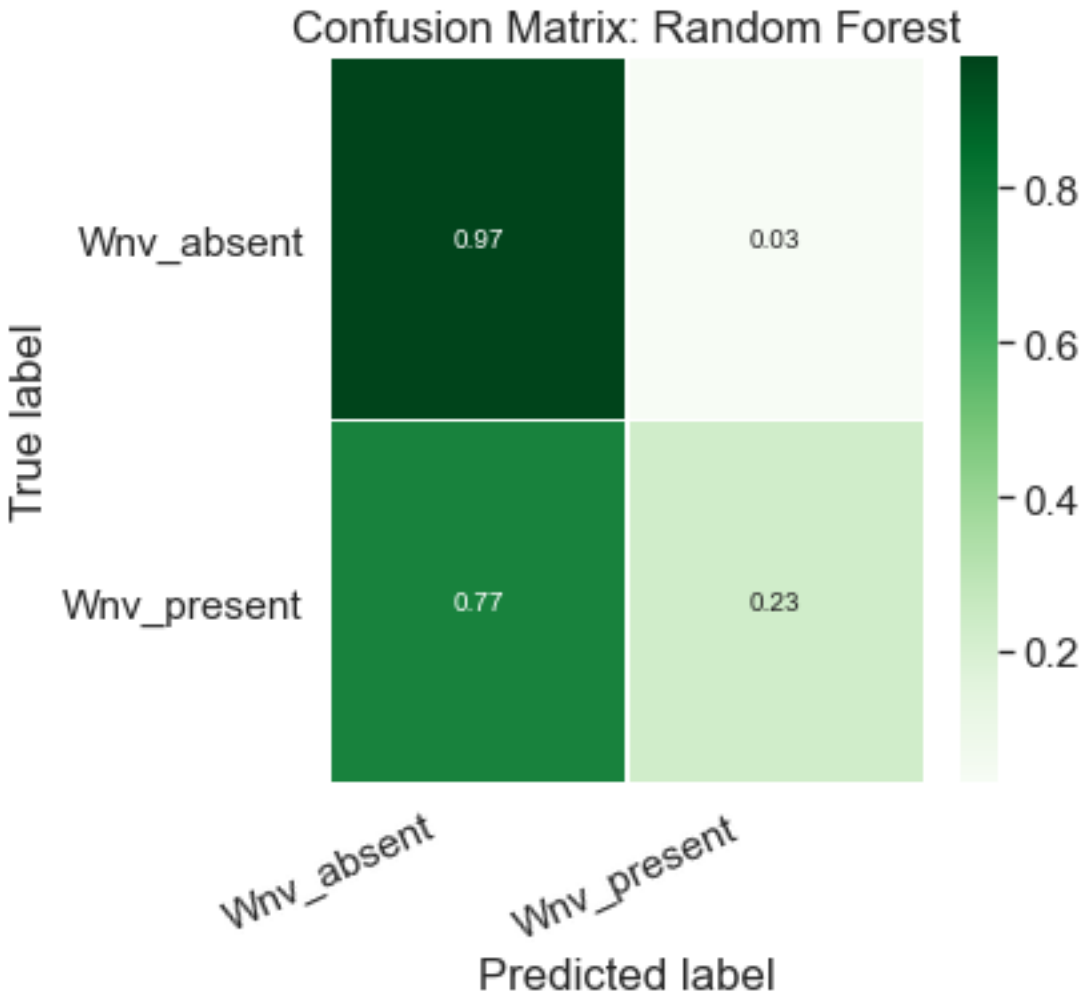

Figure 14: confusion matrix for random forest model

On the other hand, the XGB classifier performs significantly better in this case. The first iteration results in a model with can predict the presence of virus with 47% accuracy and the absence of virus with 91% accuracy. This is a marked improvement of the random forest model, however, it still falls short of the accuracy desired in this application.

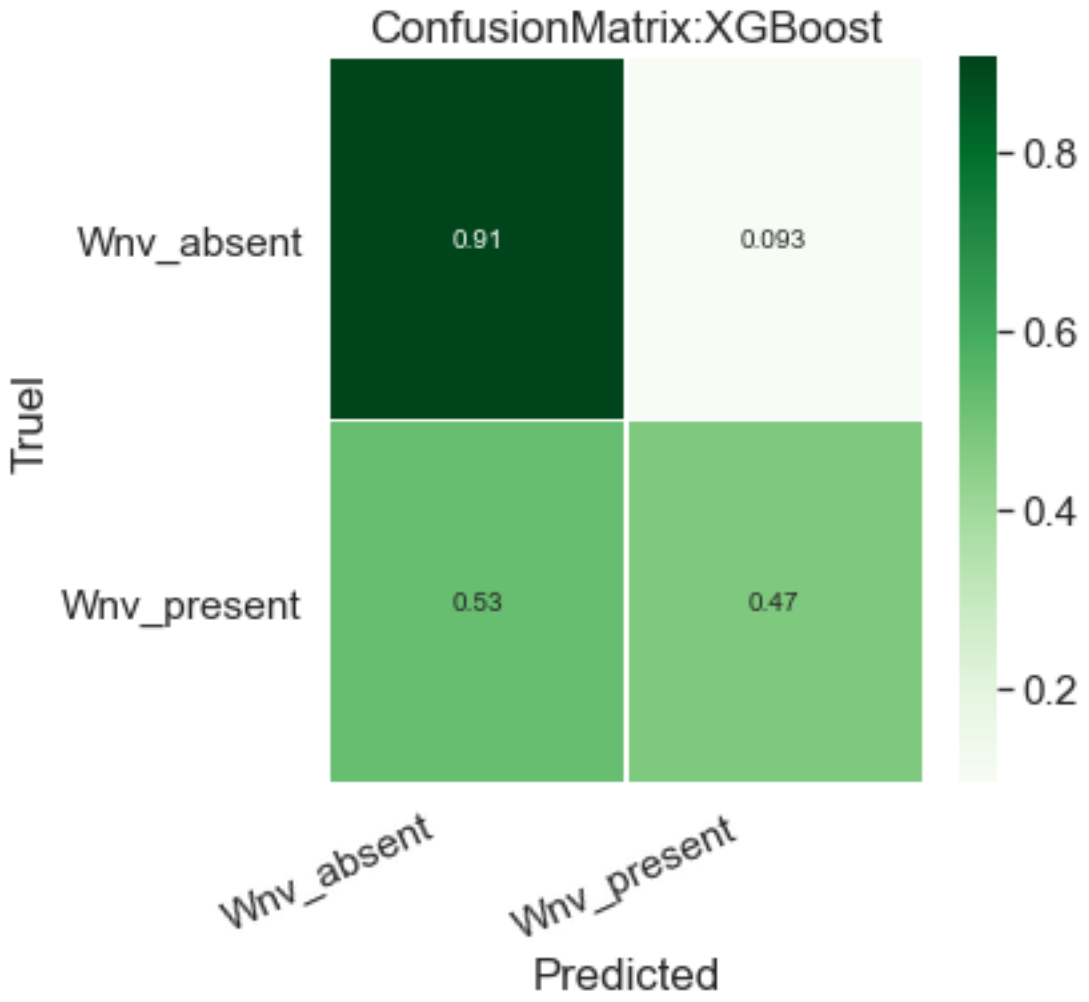

Figure 15: XGB model iteration 1

The model is further refined by utilizing a grid search with rough values for parameters through the model. This resulted in a model which can predict the presence of virus with 61% accuracy and the absence of virus with 86% accuracy. This iteration is in the more acceptable range, however, it may be possible to improve results further. This is achieved by further processing the data and by utilizing a second grid search. However, the second grid search is done using more specific possible values for parameters. This makes it possible to pinpoint a very precise parameter in order to get the best results possible. Following this, iteration three of the XGB model was prepared. This model performed remarkably well for the current usage case.

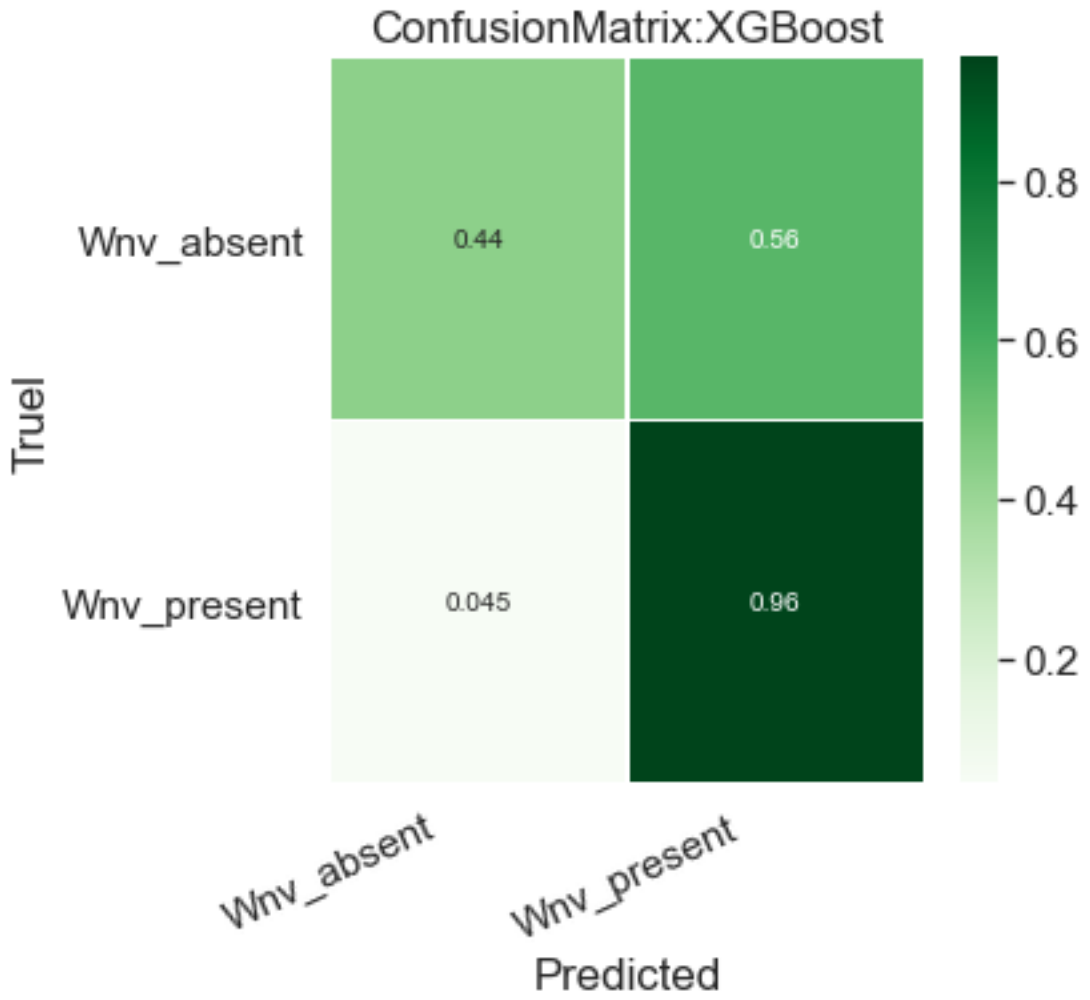

Figure 16: final (3rd) iteration XGB model

The final version of the XGB model (figure 16) was able to predict the presence of virus with 96% accuracy. However, it should be noted that the absence of virus has fallen to 44% accuracy. This is far more acceptable since a false positive, in the use case of improving resources to combat disease, is far less costly. A false positive does not result in the disease continuing to spread further such as the case with a false negative. Therefore, it can be determined that with the XGB classifier, it will be possible to properly deploy resources needed to combat West Nile virus in Chicago.

Conclusions

Throughout this project, it has been established that the XGBoost (XGB) classifier is excellent for use cases in which the cost of a false positive is less than that of a false negative. In the current case, it is possible to utilize the XGB classifier to properly deploy resources to combat

and prevent the spread of West Nile virus. The developed model is able to predict the presence of West Nile virus in a sample with an accuracy of 96% and the absence of West Nile virus in a sample with an accuracy of 44%. With the current use case, these values are quite excellent in helping to properly deploy necessary resources.