West Nile Virus (WNV) Prediction

Where and what season the West Nile Virus could be observed? is aerial spraying effective?

Abstract

West Nile virus (WNV) is an arthropod borne virus and is the leading cause of domestically acquired arboviral disease in the contiguous United States. According to the CDC, about 1 in 5 people who are infected develop a fever and other symptoms. An estimated 70%–80% of WNV infections are asymptomatic. Symptomatic persons usually develop an acute systemic febrile illness. Less than 1% of infected persons develop neuroinvasive disease, which typically presents as encephalitis, meningitis, or acute flaccid paralysis. WNV disease surveillance is important for detecting and monitoring seasonal epidemics and for identifying persons at increased risk for severe disease. Surveillance data can be used to inform prevention and control activities. Various machine learning method have been used to predict the cause and effect of the WNV. The XGBoost model used in this case study is critical in minimizing the cost of false-negative and classifies the presence of the virus with the probability of 0.88. The model performance on the absence of virus is 0.56, which is acceptable as there is more cost of false negative than false positive in the virus outbreak and humans' health.

Introduction

The first case of WNV was reported in 2002 in Illinois. Ever since the virus has consistently been reported to be associated with human cases of illness in Chicago region. Several factors such as temperature, rainfall, and infection status of mosquito are responsible for widespread of virus. Despite the identification of some of these potential risk factors, accurate prediction of human illness cases from WNV remains elusive, especially as it is related to dynamic weather and mosquito infection status. And therefore, it has no particular treatment as of today. Using long-term data related to WNV illness and intensive mosquito surveillance for the Chicago region, we can identify the cause an effect of the widespread of virus. The effective model can help the health department to control the outbreak of WNV virus and to utilize the resources properly.

Data

The dataset is provided by the Chicago Department of Public Health. It is accessible in Kaggle. It includes test, train, weather, and spray datasets. The train data contains information regarding date, location, mosquito species, and a label indicating the virus's presence and absence. The data contains mosquitoes' samples collected from 2007 to 2013 in every 2 years of duration across 136 different locations of Chicago. There exist 10506 rows with 11 features in the train data. Weather data provides time-series weather parameters recorded by NOAA from two weather stations. It contains 2944 rows with 21 features, and the parameters are missing in some rows and columns in the weather data. Similarly, the spray data set

Data cleaning

- Parameters from two weather stations are averaged and the null values are filled by the average of forward and backward filling methods
- Features such as day length, wet-bulb depression, dewpoint depression, and relative humidity are calculated from sunrise and sunset, average temperature, dewpoint, wet-bulb, actual pressure, and saturation pressure respectively.
- The weather parameters are then merged with training data to further explore the relation of weather parameters with the virus observation probability.

- Additional features such as day of the week, day of the month, week of the year, month, season, and year are extracted from the date for gaining seasonality on virus observation.
- Multiple samples were collected each day from several locations. To perform time series analysis, virus observation is resampled in the daily time frame in terms of daily virus observation probability.

Exploratory data analysis (EDA)

The highest sample was collected from O'Hare International Airport. Within the collected sample it appears that 6500 North oak park Ave contains the most viruses 19.35 % followed by 6000 N Avondale Ave 16.67%.

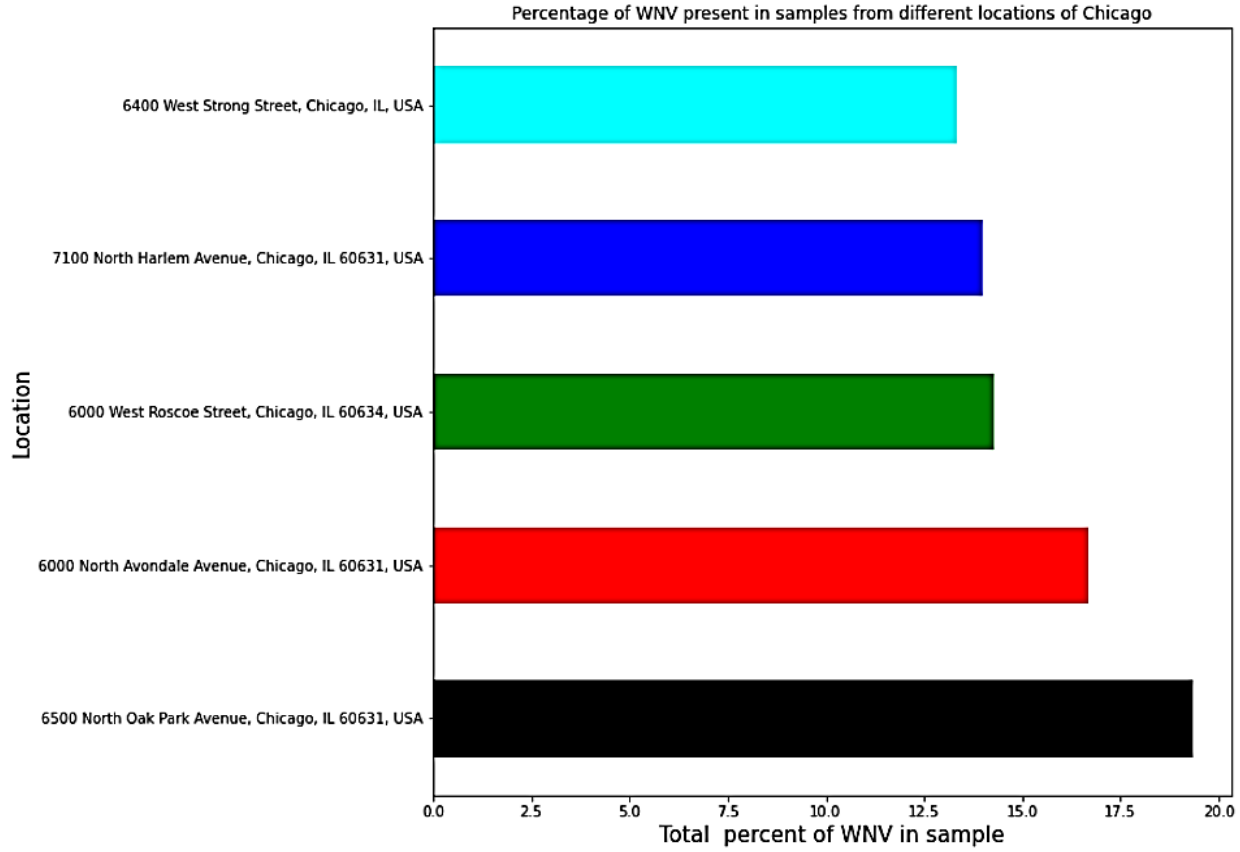

Fig1: WNV present in different locations

It appears that there are altogether 7 different species of mosquitoes in samples. Among them only three of were infected with the WNV virus, while the rest of the species contains none. The following figure demonstrates the number of viruses present in each sample.

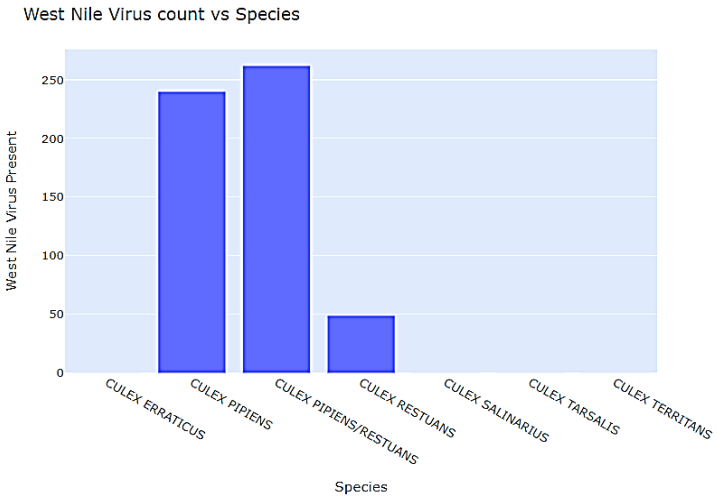

Fig2: percentage of WNV present in each species

Parameters such as sea level and station pressure, result speed and average speed, dewpoint depression and wet bulb depression are highly correlated to each other.

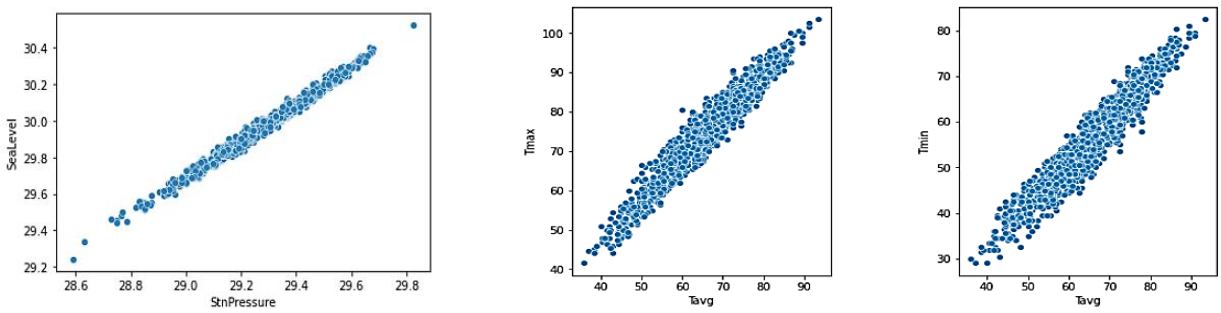

Fig 3: display of strong correlation between parameters.

The observation of the merged dataset of weather and training data provides some interesting output such as maximum sample collection year, virus present in each month/season etc. It appears that maximum sample was collected in 2007, however maximum virus was observed in 2013. Also, the virus appears to active mostly from July to September as shown in the figure below.

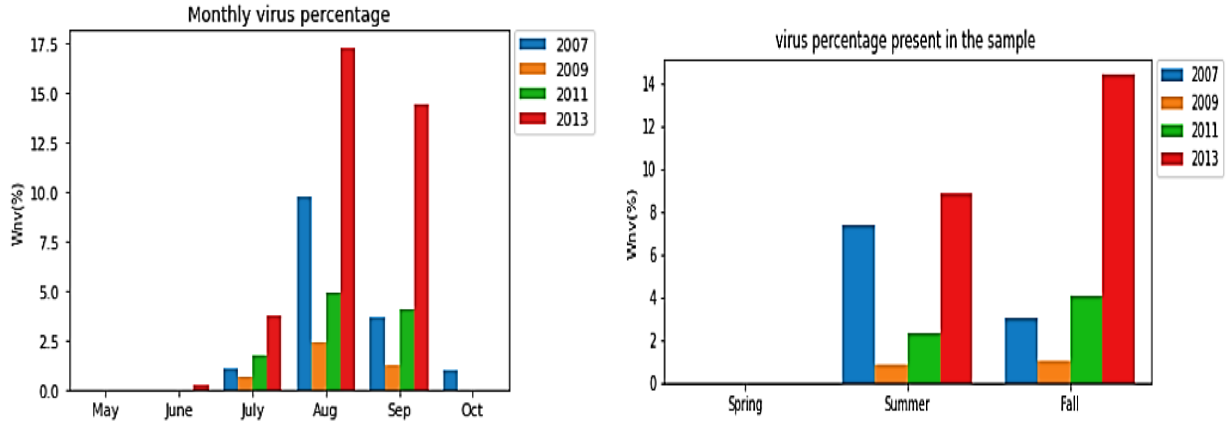

Fig 4: percentage of WNV present in different period

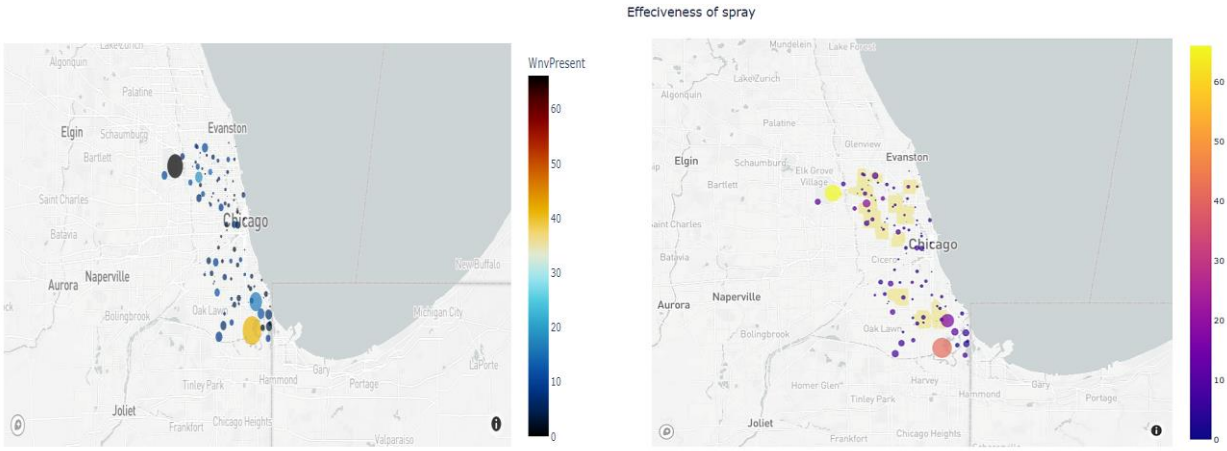

Fig 5: maps showing before and after spraying effect

Spraying is effective. Areas that were sprayed, have very less virus cases but not spraying in correct areas caused more West Nile Virus cases.

Algorithms & Machine Learning

For the machine learning model, binary classification model was applied using a technique of Information Value (IV). Because the method can be used in the different features such as prediction of the virus. The IV value less than 0.1 were eliminated as they are not useful for the model. Similarly, value 0.8 were also opted out to prevent the bias and suspicious relationship with other dependent variables. Therefore, only the features exhibiting IV statistics within a range of 0.1 to 0.8 were selected for building a model.

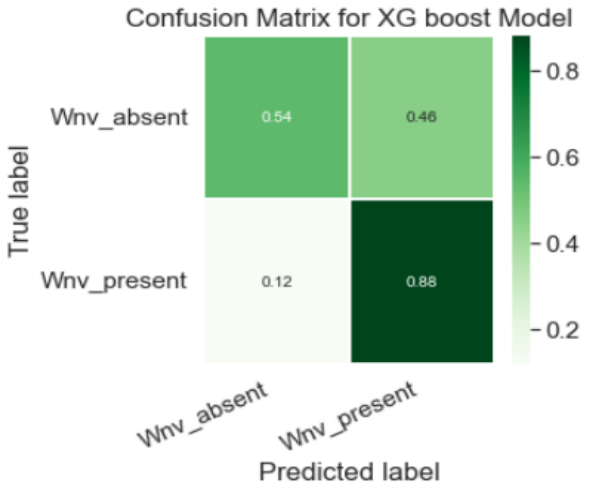

Fig 6: Confusion matrix

Another approach of using variance inflation factor (VIF) was executed to reduce the degree of multicollinearity between the identified features. Features exhibiting VIF greater than 5 exhibit extreme multicollinearity and were eliminated avoided to make features independent and ensure that the model can easily predict the dependent variable. These techniques reduce the number of features to 11 related to the time lag, weather, mosquito species and time periods. Of the 11 features, seven are encoded categorical variables describing months of the year and mosquito species. Data are partitioned into train and test sets with a size ratio of 7/3. Data modeling is performed on a train set with a supervised learning technique, an eXtreme Gradient Boosting (XGBoost) classifier algorithm.

Effectiveness of machine learning model

It appears that the 'train' dataset used in this project is highly imbalanced, with only about 5 percent of data being virus infected and rest of them with no virus at all. Also, the cost of incorrectly classifying the presence of the virus (false negative) is riskier for virus outbreak than incorrectly classifying the absence of the virus (false positive). I deployed the confusion matrix to evaluate the effectiveness of model in detail. As a result, the model showed that the virus presence probability is 0.88 and the absence of virus 0.54. Thus, the model is efficient in predicting the presence of the virus.

Findings and conclusion

The partial dependence plot of biweekly lag exhibits some sort of linear relationship with the target variable. The larger probability in the biweekly lag increases the prediction for the presence of the virus if the mosquito species is Culex species.

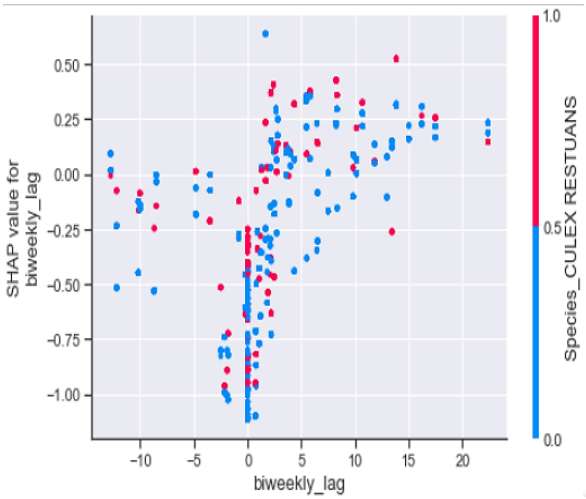

Fig 7: The partial dependence plot of biweekly lag

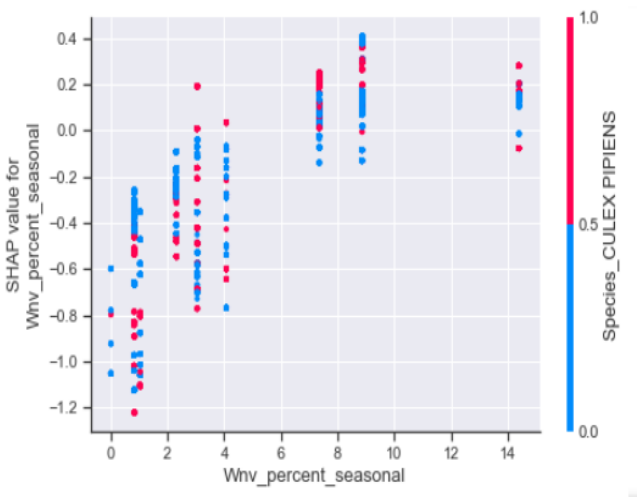

Fig 8: Seasonal virus observation probability

The mean temperature is generally higher in the year 2007 and 2013 when the mosquitos were majorly found to be WNV carriers.

Year	Mean Temperature (F)
2007	75-80
2009	65-70
2011	50-55
2013	75-80

The association between higher temperature and WNV human illness has also been observed. The presence of weather parameters, heat index, and departure of temperature is reasonable as mosquitoes become active in higher temperatures and breed rising their populations.

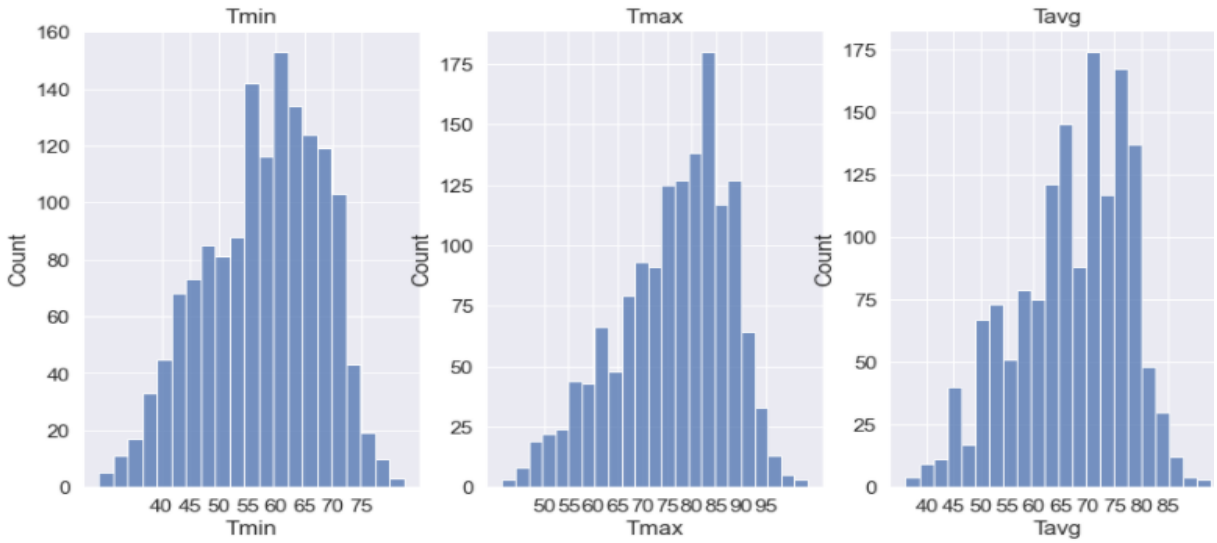

Fig 10: change in number of virus with temperature

Hence, from the rigorous study using machine learning model we can conclude that the virus classification is largely dependent on the biweekly (14 days) lag value of daily virus observation probability. This recommends that Chicago health department should increase the surveillance on mosquito's species to collect more sample and to get the better predictions. As it is clear that aerial spraying on mosquitoes on infected areas is effective way to minimize the mitigation, people should use mosquito repellant in the region where the potential outbreak may occur. This analysis helps to better understand the major causes of WNV and ways to minimize the effect. This information can be used as a guideline to develop threshold for public health safety measures by Chicago city.