Capstone: West Nile Virus Prediction

When and where the West Nile Virus could be observed in the City of Chicago?

Abstract:

Using XGBoost (eXtreme Gradient Boosting) classifier algorithm, a machine learning binary model is developed to predict West Nile virus for a given time, location, and mosquito species in the City of Chicago using location and weather data. The model 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. The binary model classifies the absence and presence of virus with an AUC of 0.71, which is better than a random classifier.

Introduction:

West Nile Virus is a disease transmitted to human beings through the bite of an infected mosquito. About 20% of infected people develop severe symptoms ranging from a persistent fever to severe neurological illnesses that can result in death. The virus was first reported in 2002 in Chicago. To control the epidemic, the City of Chicago and the Chicago Department of Public Health (CDPH) started a surveillance program to control the mosquitoes by spraying in the region where the potential outbreak may occur. Thus, we need to predict when and where the mosquito will test positive for the virus, analyzing weather, location, and training data. The effective model will help in allocating the resources in the predicted region and time to control the mosquitoes and the outbreak of the virus. Also, it will optimize resource allocation and expenditures associated with controlling the virus.

Exploratory data analysis:

The Chicago Department of Public Health provides data, and it is available in Kaggle. Two sets of data: GIS and weather data are used in this study to build a model. GIS data contains information regarding date, location, mosquito species, and a label indicating the virus's presence and absence. There exist 10506 rows with 11 features in the GIS data. Weather data provides time-series weather parameters recorded by NOAA from two weather stations. There exist 2944 rows with 21 features, and the parameters are missing in some rows and columns in the weather data.

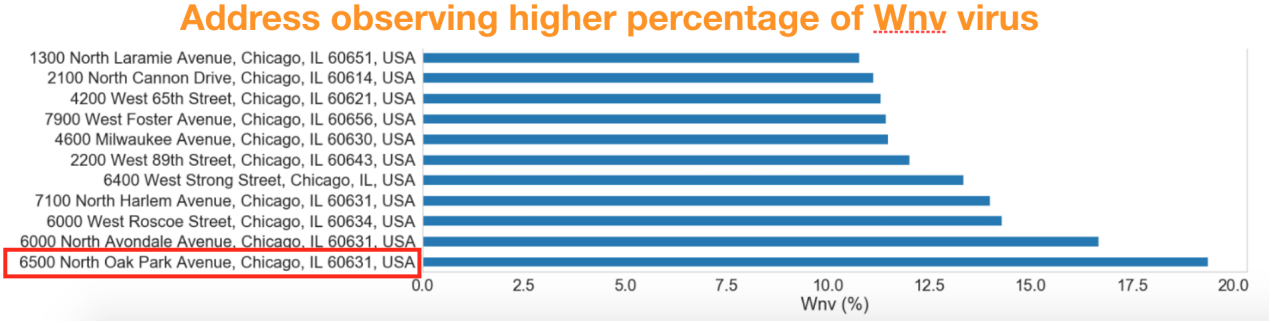

Figure 1: Locations observing higher percentage of virus in the city of Chicago.

GIS data contains data collected every two years apart from 2007 to 2013 across 136 unique locations in the city of Chicago. The latitude and longitude of these locations are provided, along with their coordinate accuracy, street name, block, and address. Mosquitoes are trapped from late May to early October. The presence and absence of virus in the mosquitoes are indicated by labels 1 and 0, respectively. The highest sample was collected from O'Hare International Airport, with 8.8% containing the virus. The highest virus percentage was observed at 6500 North Park Avenue, Chicago IL, 60631, USA (19.35%), as shown in Fig. 1.

Figure 2a shows six species of mosquito in the data with two species, Culex Restuans and Culex Pipiens, being mixed in most of the sample collected. The West Nile virus was observed only in these two species. The average probability of observing a virus in the sample is 5.2%, indicating wide disparities in the presence and the absence of virus in the samples collected (Fig 2b).

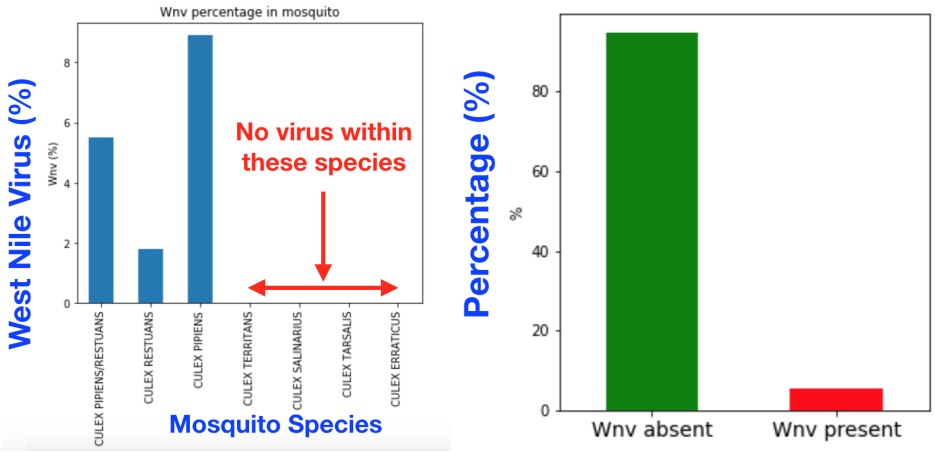

Figure 2: a) Virus percentage in the GIS data. b) Mosquito species and their virus percentage.

Weather data contains weather parameters such as temperature, pressure, speed, sunrise, etc., from two weather stations. Many weather parameters are observed to be correlated in the heat map and scatter plots.

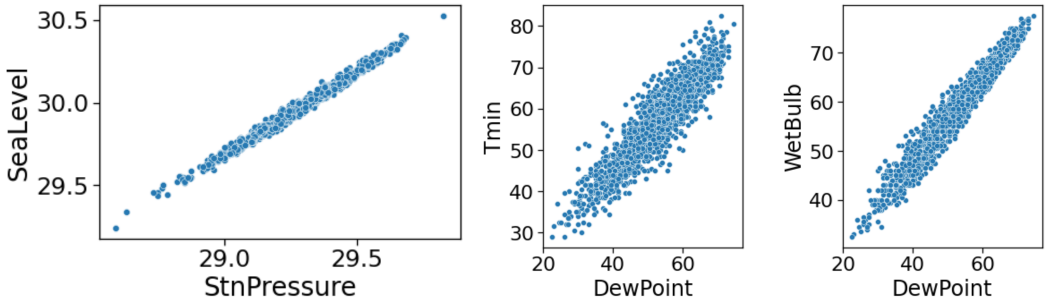

Figure 3: Weather parameters exhibiting strong correlation between them.

The correlated parameters are as follows:-

- average, maximum and maximum temperature

- Sea level and Stnpressure
- Result speed and Average speed
- Dewpoint and bulb

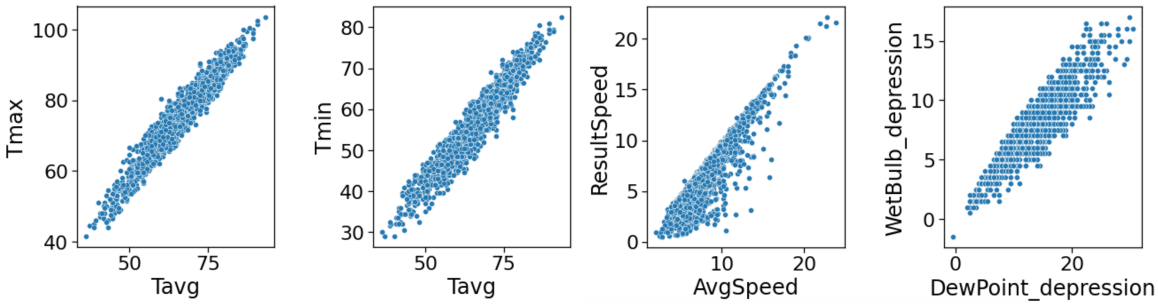

Figure 4: Weather parameters exhibiting strong correlation between them.

Feature engineering:

The available parameters from both weather stations are averaged and the null values are filled by the average of forward and backward filling methods. Additional 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.

Only one of the correlated parameters is retained in the final weather data to remove multi-collinearity issues. The weather parameters are then merged with GIS data to further explore the relation of weather parameters with the virus observation probability. The relation of weekly virus observation probability and their corresponding weekly parameters are shown in Fig. 6.

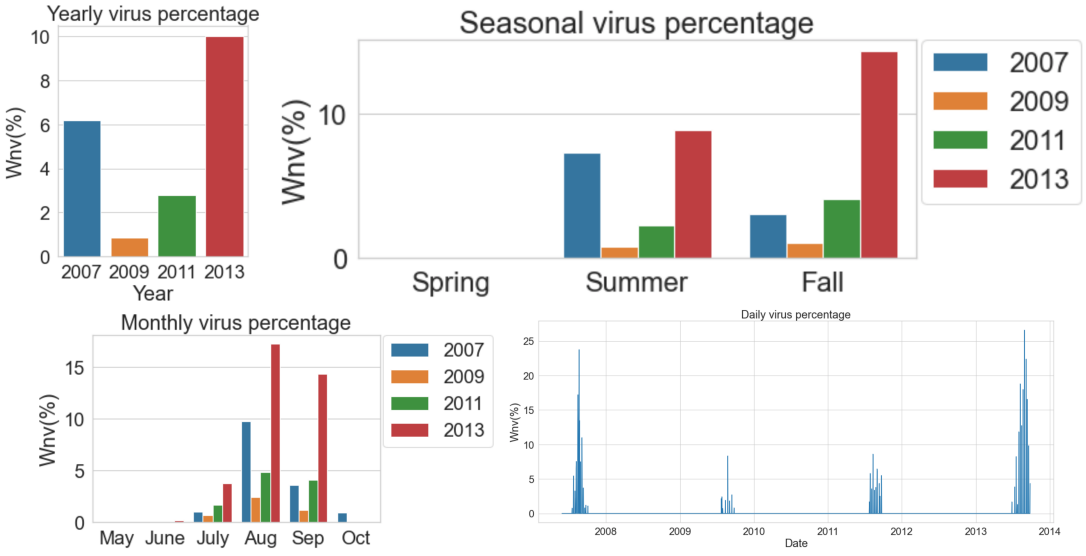

Figure 5: a) Yearly, b) seasonal, c) monthly and d) daily virus percentage.

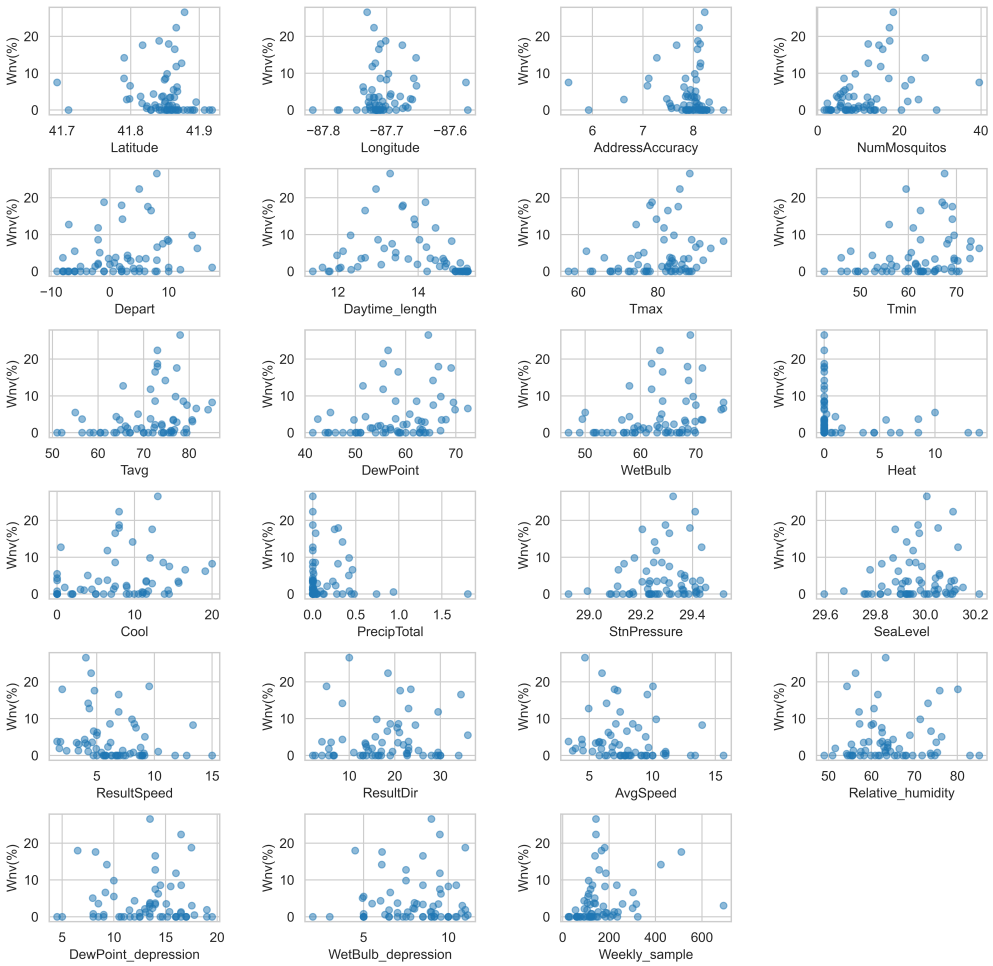

Figure 6: Relation between weekly virus observation probability and weather parameters

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. Although the maximum sample was collected in 2007, the maximum virus was observed in 2013. The virus appears in June, becomes critical in August, and then disappears in October, as shown in Fig. 5.

There are multiple samples 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 (percentage). While plotting the autocorrelation of daily virus percentage, the correlation with prior 7, 14, and 21-day observations are significantly higher than other days, as shown in Fig. 7. As a result, these lags are further added as separate features in the dataset. With the addition of new features, the total features resulted in 45 features.

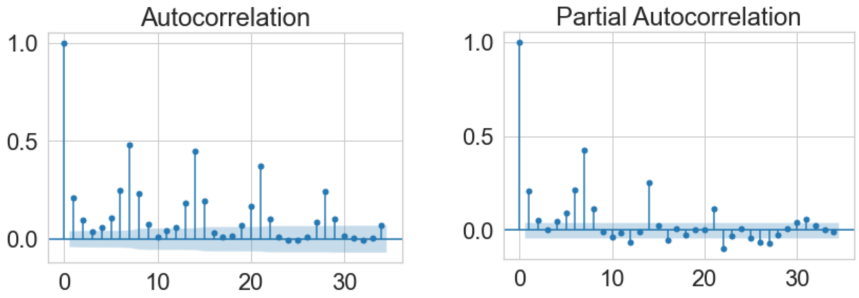

Figure 7: a) Autocorrelation and b) partial correlation of daily virus observation probability.

The categorical columns representing mosquito species, month, and season are further hot-encoded. The hot encoded month and mosquito species that do not contain are dropped. Furthermore, the date column and the object columns describing locations are dropped, retaining only latitude and longitude.

Model pre-processing

As the important features in a binary classification model can be selected using a technique called information value (IV), this method is employed in the selection of features for the prediction of the virus. Only the features exhibiting IV statistics within a range of 0.1-0.8 are selected for building a model. While the IV statistics less than 0.1 are not useful for modeling, a value greater than 0.8 could lead biased and suspicious relationship with a dependent variable. To reduce the degree of multicollinearity between the identified features, a technique called variance inflation factor (VIF) is further used in feature selection. Features exhibiting VIF greater than 5 exhibit extreme multicollinearity and are avoided. This helps 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.

The presence of weather parameters, heat index, and departure of temperature from the average of 30 years, is reasonable as mosquitoes become active in higher temperatures ($>80^0$ F) and breed rising their populations. Also, at an extreme heat index without humidity, mosquitoes die out. Virus host on mosquito and are observed only in Culex Pipiens and Culex Restuans. While a female mosquito's lifespan is longer, the average lifespan of a male mosquito is 10 days. The lifespan of mosquito species Culex Pipiens is 7 days. Thus, the presence of a fourteen-day lag in the daily virus observation probability is sensible. The presence of the datetime variable, month, also makes sense as mosquitoes appear since June, becomes maximum in August, and disappear in October.

Similarly, mosquitoes are more active in higher temperatures ($>80^0$ F). Thus, the higher heat index creates a favorable environment for the mosquitoes to survive.

Data modeling:

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. XGBoost is a popular machine learning algorithm that implements an optimized

gradient-boosting method in a parallel mode outperforming single algorithm methods' speed and performance.

Using a grid search method in the XGBoost classifier, multiple model parameters are tuned with a five-fold cross-validation and roc_auc scoring function to predict the target feature representing a virus's absence or presence. The best model parameters are selected with a score of 0.81. The model is again trained with the selected parameters.

Results and discussion:

When the model is applied to the test set, a slightly lower score of value 0.71 is obtained (Fig 8a), which is obvious as the model is built on the train set. The model performs better than a random classifier having an AUC of 0.5. The dataset in this project is highly imbalanced, with only 5% of data containing the virus. 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). This implies that the magnitude of metric "recall" should be maximum.

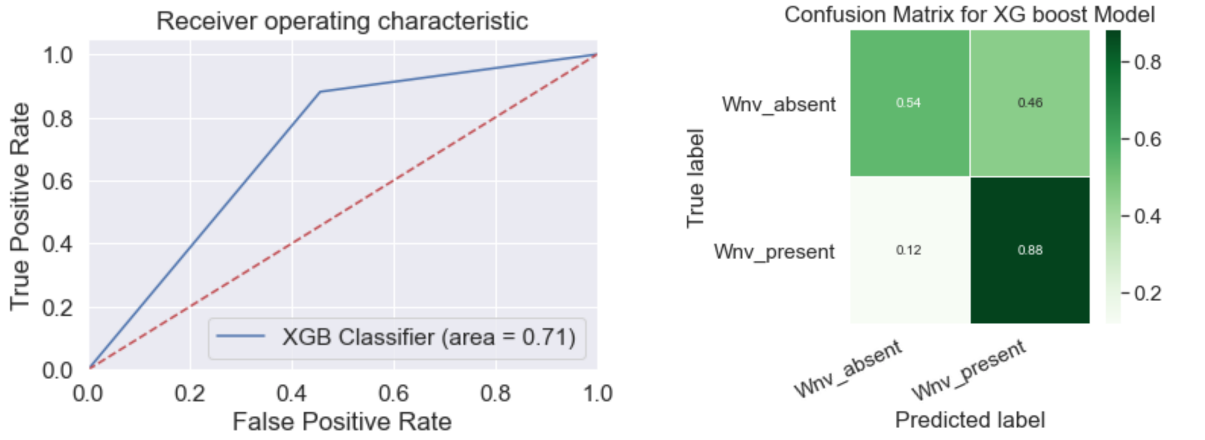

Figure 8: XGB classifier's a) AUC curve and b) Confustion matrix.

To evaluate the performance of the model in detail, the confusion matrix (Fig 8b) and the classification reports are generated. The model predicted the presence of virus with a probability of 0.88 (recall) and the absence of virus with 0.54. Thus, the model is efficient in predicting the presence of the virus. Although the model performance on the absence of virus is not better, it is acceptable as there is more cost associated with false-negative than false positive in the virus outbreak and human's health.

The importance of features and their impact on the model's output can be explained with shap values and their summary plots (Fig 9) . Comparing the shap summary plot, fourteen-day lag and seasonal observation are the most important features affecting the classification decision or model output. Other features with their importance in descending order are a departure from the average of 30-year temperature, June, mosquito species Culex Restuans and Culex Pipiens, Aug, July, September, heat index, and October.

The fourteen-day lag is positively correlated with the presence of the virus. In other words, the high value of virus observation probability in the last 14 days causes higher predictions, and low value causes low predictions. This is also accounted by the ACF and PACF plot. Thus, the pattern of virus observation repeats after every two weeks, and this may be related to the life span

Figure 9: a) Feature importance and b) their impact on model's output. Color red and blue show whether the feature has high or low value and the horizontal location shows whether the effect of that value has a higher or lower impact on model prediction.

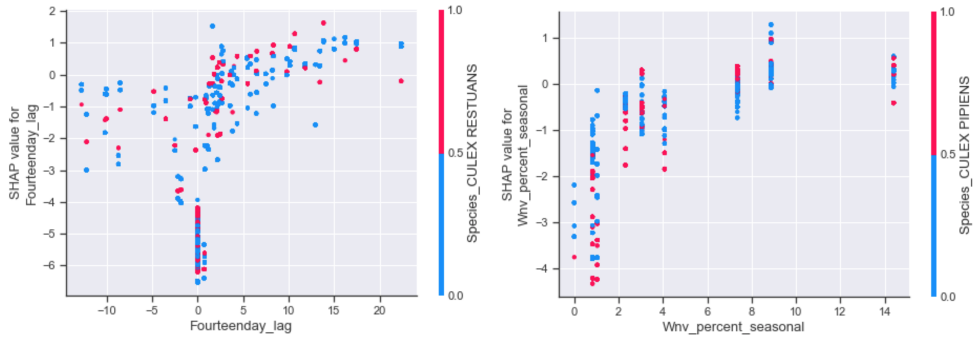

Figure 10: The partial dependence plot of fourteen day lag and seasonal virus observation probability.

of mosquitoes, as described earlier in the prior section. The partial dependence plot of fourteen-day lag exhibits some sort of linear relationship with the target variable, and the spread suggests its interaction with the mosquito species, Culex Restuans (Fig 10a). The larger probability in the fourteen-day lag increases the prediction for the presence of the virus if the mosquito species is Culex Restuans.

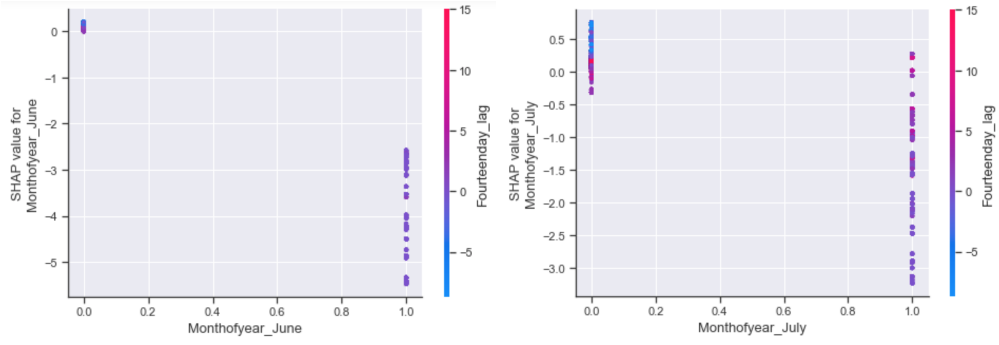

Figure 11: The partial dependence plot of months a) June and b) July.

Similarly, the season having higher virus observation probability is positively correlated with

the virus's presence, and it frequently interacts with the mosquito species, Culex Pipens (Fig 10b). The month of June and July are negatively correlated with the presence of the virus. The interaction of both months with the fourteen-day lag indicates that the presence of the virus increases if the lag probability is greater than 5 (purple color) (Fig 11b). In August, the presence of the virus increases if the mosquito species is Culex Restuans. The model dependence on the months of the year is clear from the exploratory data analysis, where the probability of virus observation starts from July, reaches a maximum in August, and disappears in October. Mosquito species alone are negatively correlated in virus prediction. However, their interaction with fourteen-day lag, season, and months of the year states that virus prediction increases when there are higher virus observation probability in these features (Fig 12).

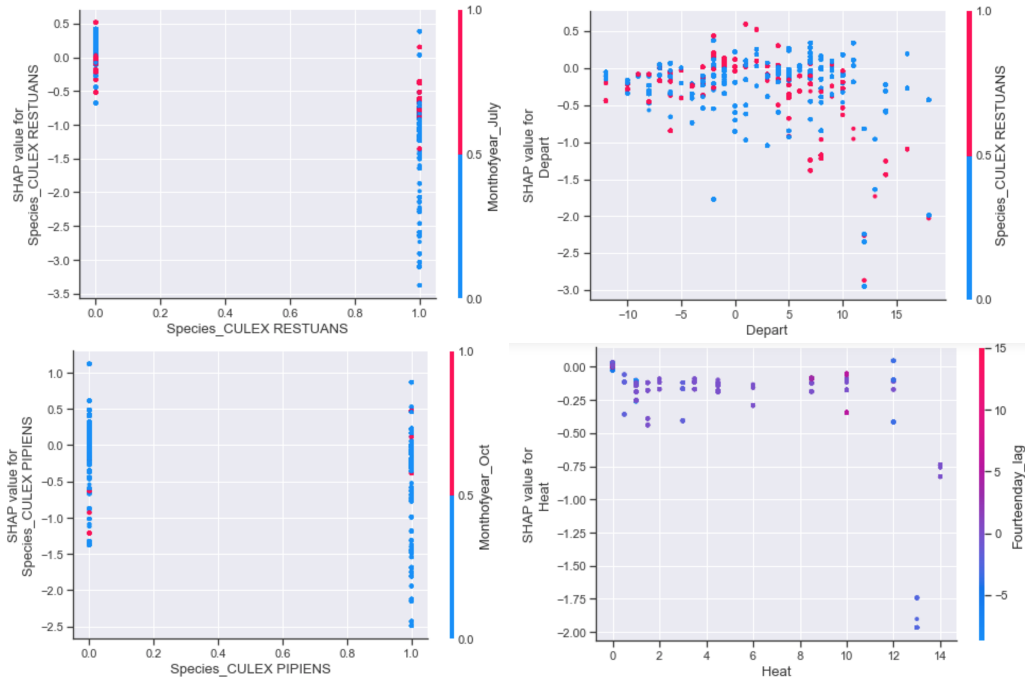

Figure 12: The partial dependence plot of mosquito species a) Culex Restuans, b) Depart, c) Pipiens and d) Heat.

The plot also indicates that the weather parameters departure and heat index do not affect the prediction, but when they interact with the mosquito species, Culex Restuans, and the fourteen-day lag, the prediction of the virus increases (Fig 12).

Conclusions and Discussions:

There is more cost associated with the improper classification of the presence of the virus. As a result, the model predicting the virus should be very accurate in properly classifying the presence of virus than the absence of virus. The binary model developed using XGB classifier algorithm predicted the presence of virus with a probability of 0.88 and the absence of virus with 0.54. Thus, the model is much critical in classifying the presence of virus and the model reasonably fulfills the objective set for predicting the presence of virus. The model classifies the presence and the absence of virus with an AUC of 0.71.

The virus classification is largely dependent on the fourteen-day lag value of daily virus observation probability. The fourteen-day lag value is positively correlated with the shap value for virus classification. This implies the virus presence increases with the higher magnitude of fourteen-day lag value. Also, the virus observation increases in the season, having a higher observation probability which is as expected. The virus observation does not depend on the number of mosquitoes but depends on the mosquito species. When the mosquito species, Culex Pipens and Culex Restuans, interacts with the month, season and the fourteen-day lag values having the higher virus observation probability, the observation of virus increases. However, the mosquito species alone are not sufficient for the prediction of the presence of virus as they are negatively correlated in the model output. The probability of virus observation starts from July, reaches maximum in August and disappears in October. Weather parameters do not play a major role in virus classification. While the model relationship with the temperature departure from 30-year average temperature is complex, the model predicts the presence of virus with the low heat index.

Thus, the time series analysis of virus observation probability and their lag values are the most important features for the prediction of the virus. The model performance may be improved by creating rolling features and percentile of virus observation over multiple rolling time windows.