

West Nile Virus Prediction

**Predicting West Nile virus in mosquitoes across the city of
Chicago**

Final Report

Derya Kurt

March 2021

Table of Contents

1	INTRODUCTION	3
1.1	Problem Statement	3
2	DATA AND PRE-PROCESSING	3
2.1	Data Overview	3
2.1.1	Main Data	3
2.1.2	Spray Data	3
2.1.3	Weather Data	4
2.2	Data Processing and Cleaning	4
2.2.1	Main Data	4
2.2.2	Spray Data	4
2.2.3	Weather Data	5
2.2.3.1	Removal of Data	5
2.2.3.2	Imputation	6
3	FEATURE ENGINEERING AND EXPLORATORY DATA ANALYSIS ...	6
3.1	Main Data	6
3.2	Spray Data	8
3.3	Weather Data	9
4	MACHINE LEARNING	10
4.1	Building Model	10
4.2	Metrics and Evaluation	11
4.3	Best Model and Feature Importance	12
5	SHAP (SHapley Additive exPlanations)	12
5.1	The Dependence Plots	13
5.1.1	ResultSpeed_ema50	13
5.1.2	Depart	13
5.1.3	ResultDir_lag28	14
5.1.4	PrecipTotal	14
5.1.5	Species_CULEX RESTUANS	15
5.1.6	Species_CULEX PIPiens	15
5.1.7	FG	16
5.1.8	TS	16
5.1.9	BR	17
5.1.10	spray_day	17
5.2	Force Plot Explanation	18
6	CONCLUSION	18
7	FUTURE WORK	19
8	RECOMMENDATIONS	19

1. INTRODUCTION

1.1 Problem Statement

West Nile virus (WNV) is a mosquito-borne virus and it is first seen in New York City in 1999. It has spread rapidly across the United States. The natural hosts for WNV are birds and mosquitoes. Due to the natural transmission cycle between them, the most significant mosquito species for viral transmission are *Culex* species that feed on birds. West Nile virus is transmitted to birds through the bite of infected mosquitoes. Mosquitoes become infected by biting infected birds and then the WNV is spread to humans through the bite of infected mosquitoes.

WNV was first detected in Chicago in 2001 among dead birds. The following year, Chicago faced its largest epidemic during which 225 human cases were reported, including 22 fatalities.

By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program.

The goal of this project is to predict when and where different species of mosquitoes will test positive for West Nile virus.

The below questions will be analyzed through the given weather, location, and spraying data.

1. Which features play an important role in the presence of WNV?
2. Has spraying been effective to control the mosquito population?
3. What is the best predictive model to predict the WNV?

2. DATA AND PRE-PROCESSING

2.1. Data Overview

3 data sets provided by <https://www.kaggle.com/c/predict-west-nile-virus/overview> were analyzed.

2.1.1 Main Data: It consists of 10506 observations from 2007, 2009, 2011, and 2013 including features about Traps, their locations and tests with the date they are performed and the result showing the number of Mosquitoes caught in trap. The data set also includes target variable West Nile Virus. 1 means WNV is present, whereas 0 means not present.

2.1.2 Spray Data: It consists of 14835 observations from 2011 and 2013 containing spray date and location data.

2.1.3 Weather Data: The data set provided from NOAA(National Oceanic and Atmospheric Administration) contains 2944 observations of 2 weather Stations collected from 2007 to 2014, during the months of tests. It includes features about daily Climatological Data such as temperature, precipitation amount, wind speed and direction.

2.2 Data Processing and Cleaning

In order to perform time series analysis string Date Time of each data set was converted into Python Date time object and got new features ‘Year’ and ‘Month’.

2.2.1 Main Data:

813 out of 10506 observations were duplicate, they were all dropped. Each trap had 7 features describing its location. Since the Latitude and Longitude were enough for location description, I dropped other address information like Street, Block.

Overview After Data Cleaning:

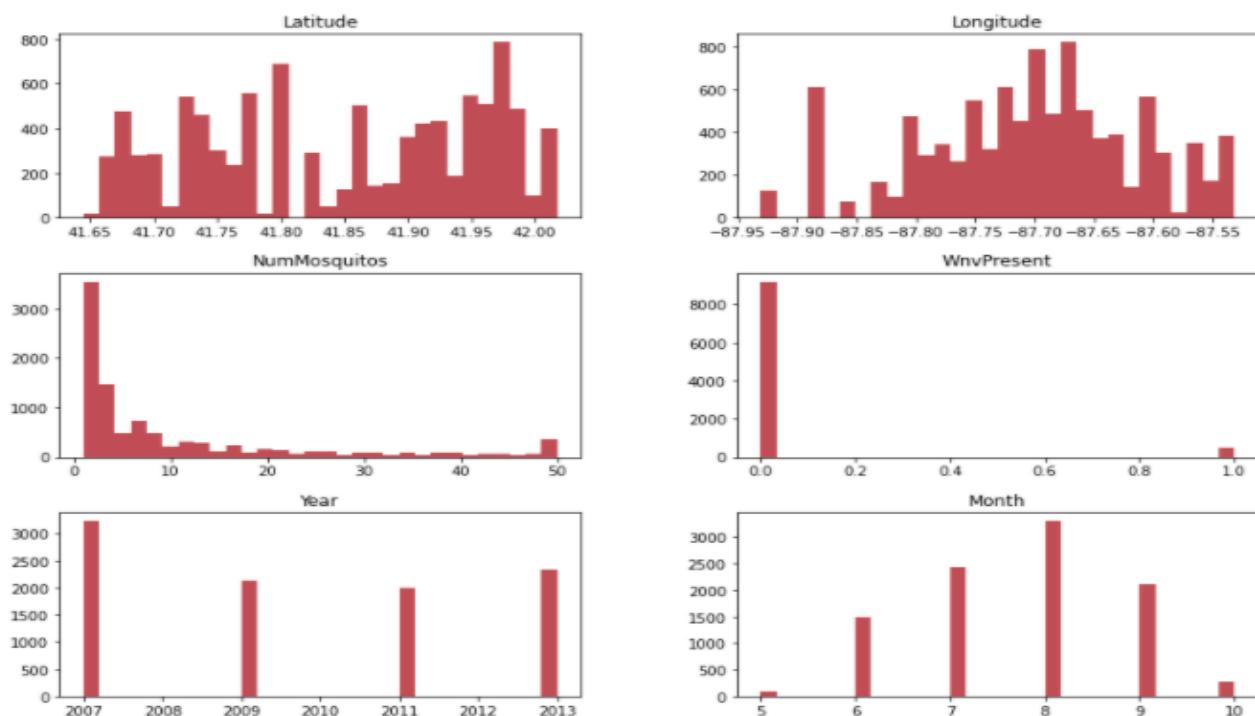


Figure 1

2.2.2 Spray Data:

The data set had 541 duplicates which were dropped. Neither Main Data Set nor Spray Data Set had missing value.

2.2.3 Weather Data Set:

In contrast to previous data sets Weather had many missing values. At first sight its summary didn't indicate any missing value. However we know from NOAA Weather Documentation that the unavailable data were marked as 'M' = MISSING DATA, 'T' = TRACE, '-'. In order to handle these missing values properly I replaced them with NaN and then looked at the pattern of the missing value to decide whether to remove or impute.

Before Data Cleaning:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2944 entries, 0 to 2943
Data columns (total 22 columns):
 #   Column      Non-Null Count  Dtype  
 ---  --          --          --      
 0   Station     2944 non-null    int64  
 1   Date        2944 non-null    datetime64[ns]
 2   Tmax        2944 non-null    int64  
 3   Tmin        2944 non-null    int64  
 4   Tavg        2933 non-null    object  
 5   Depart      1472 non-null    object  
 6   DewPoint    2944 non-null    int64  
 7   WetBulb    2940 non-null    object  
 8   Heat         2933 non-null    object  
 9   Cool         2933 non-null    object  
 10  Sunrise     1472 non-null    object  
 11  Sunset      1472 non-null    object  
 12  CodeSum     2944 non-null    object  
 13  Depth        1472 non-null    object  
 14  Water1      0 non-null     float64 
 15  SnowFall    1460 non-null    object  
 16  PrecipTotal 2624 non-null    object  
 17  StnPressure 2940 non-null    object  
 18  SeaLevel    2935 non-null    object  
 19  ResultSpeed 2944 non-null    float64 
 20  ResultDir   2944 non-null    int64  
 21  AvgSpeed    2941 non-null    object  
dtypes: datetime64[ns](1), float64(2), int64(5), object(14)
memory usage: 506.1+ KB
```

Figure 2

After Data Cleaning:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1472 entries, 0 to 1471
Data columns (total 26 columns):
 #   Column      Non-Null Count  Dtype  
 ---  --          --          --      
 0   Date        1472 non-null    datetime64[ns] 
 1   Tmax        1472 non-null    float64 
 2   Tmin        1472 non-null    float64 
 3   Tavg        1472 non-null    float64 
 4   Depart      1472 non-null    float64 
 5   DewPoint    1472 non-null    float64 
 6   WetBulb    1472 non-null    float64 
 7   Heat         1472 non-null    float64 
 8   Cool         1472 non-null    float64 
 9   PrecipTotal 1472 non-null    float64 
 10  StnPressure 1472 non-null    float64 
 11  Sealevel    1472 non-null    float64 
 12  ResultSpeed 1472 non-null    float64 
 13  ResultDir   1472 non-null    float64 
 14  AvgSpeed    1472 non-null    float64 
 15  FG          1472 non-null    float64 
 16  TS          1472 non-null    float64 
 17  DZ          1472 non-null    float64 
 18  TSRA        1472 non-null    float64 
 19  BR          1472 non-null    float64 
 20  HZ          1472 non-null    float64 
 21  RA          1472 non-null    float64 
 22  Year        1472 non-null    int64  
 23  Month       1472 non-null    int64  
 24  Day         1472 non-null    int64  
 25  Week        1472 non-null    int64  
dtypes: datetime64[ns](1), float64(21), int64(4)
memory usage: 299.1 KB
```

Figure 3

2.2.3.1 Removal of Data:

Due to below reasons I decided to drop the features:

‘Water1’: All values were null.

‘Depth and Snowfall’: All values were either NAN or 0.

‘Sunset, Sunrise’: ‘Sunset’ and ‘Sunrise’ were highly correlated with the ‘Date’. Therefore ‘Date’ was enough for further analysis.

‘CodeSum’ consisted of various weather conditions such as Fog, Mist, Snow. Some observations had more than one weather condition. Since weather conditions play an important role in WNV, I extracted ‘CodeSum’ values to new columns and dropped the ‘CodeSum’. The new features of each observation were labeled as 0 or 1, based on the occurrence of the related feature during the observation.

‘Station’: observations of 2 station didn’t have significant difference. Therefore I merged them by taking the average of the same features for the same date.

2.2.3.2 Imputation:

'Depart' column which indicates the difference from normal temperature for that day of the year, had the highest missing data (1472 missing). The reason for so many missing data was only Station 1 had 'Depart' values. Since there wasn't significant difference in Average Temperature values between the stations I imputed Depart for Station 2 with Station 1 values by using forward fill.

I also replaced missing values of '*PrecipTotal*'(20 missing), '*StnPressure*'(4 missing), '*SeaLevel*' (11 missing), '*AvgSpeed*'(3 missing) with the last observed values by using forward fill.

Both Heat and Cool had 11 missing values. They were imputed by using below calculation methods.

(if $Tavg \leq 65$) Heating Degree = $65 - Tavg$

(if $Tavg > 65$) Cooling Degree = $Tavg - 65$

3. FEATURE ENGINEERING AND EXPLORATORY DATA ANALYSIS

3.1 Main Data:

Our data set was imbalanced. Only 5% of the target feature 'WnvPresent' was positive as shown in below Figure. This would cause the model to bias toward the negative class(non_WNV). To overcome this challenge instead of changing the real data by undersampling, oversampling and generating synthetic data methods, I preferred to enrich the data with additional features.

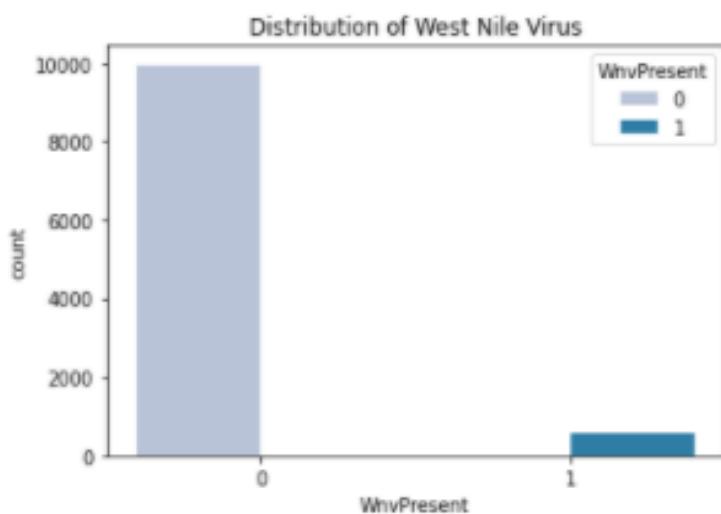


Figure 4

The below Figure 5 shows Culex Pipiens/Restuans, Culex Restuans and Culex Pipiens were the most common mosquito species and the vectors for WNV. Culex Restuans had the lowest WNV rate among them. The highest numbers of Mosquitoes and WNV were observed in August.

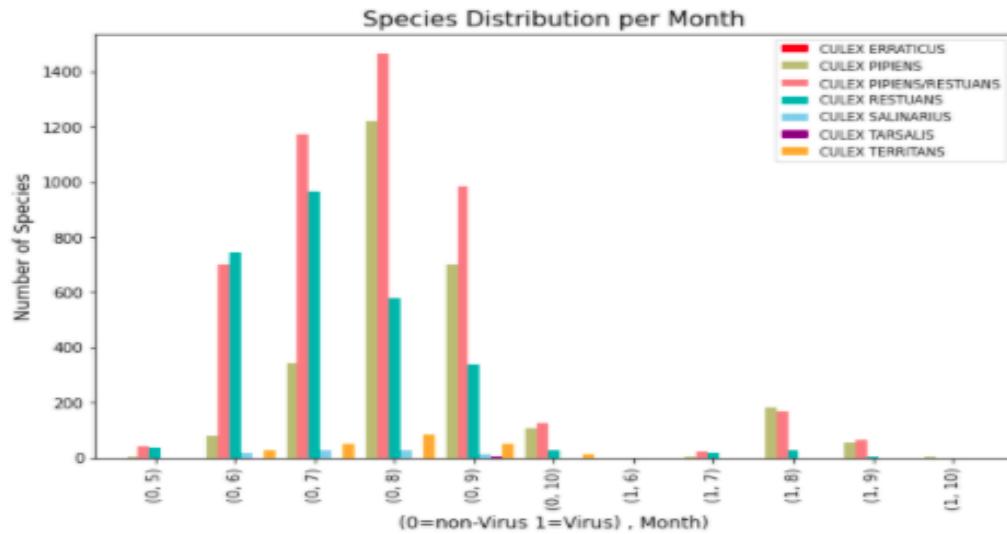
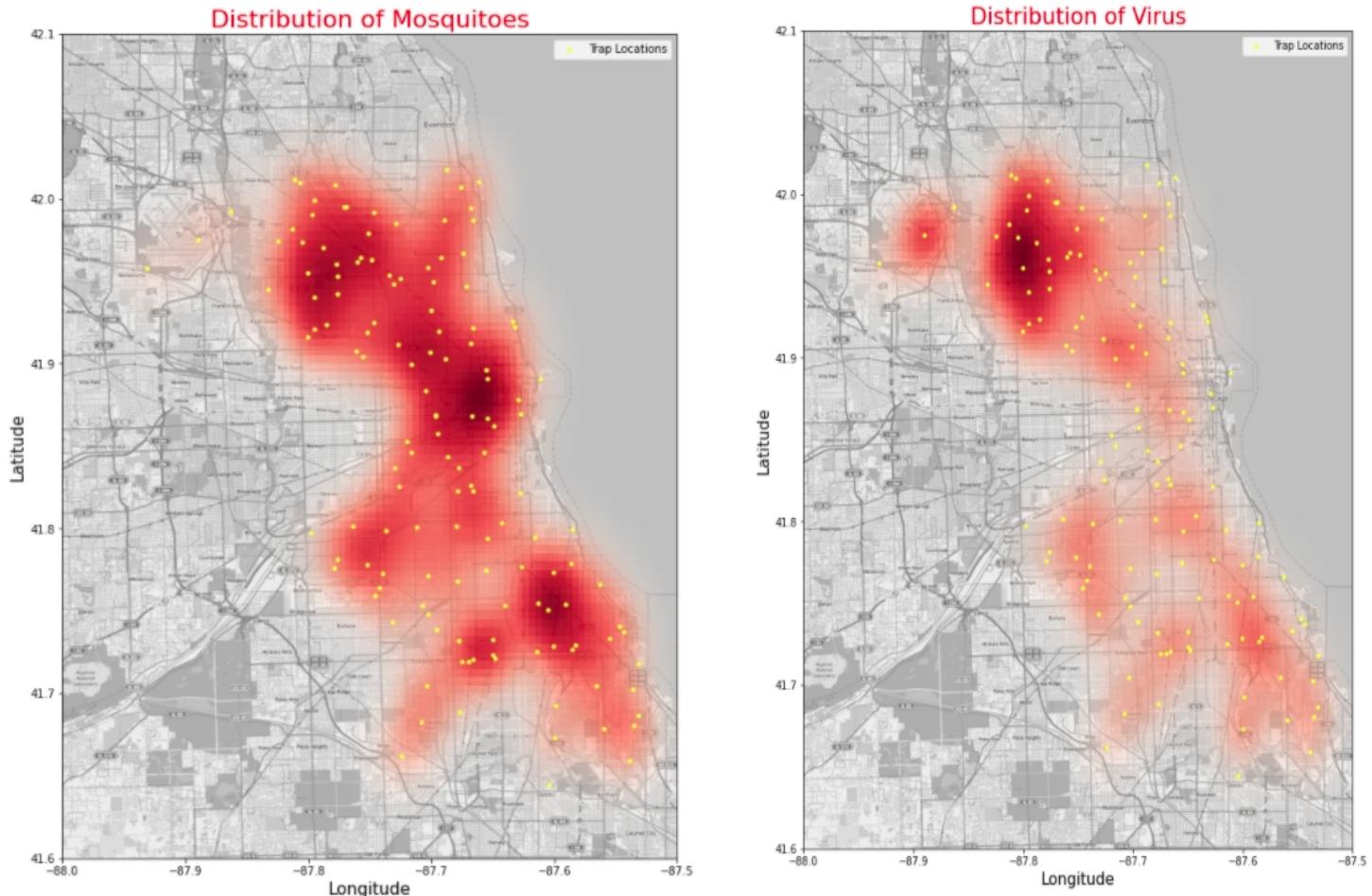


Figure 5

The below maps show the large number of mosquitoes were found close to the Lake Michigan, whereas the viruses were mostly in the southern traps.



3.2 Spray Data:

One of the goal of this project was to find if spray was effective in reducing both mosquito populations and WNV cases. To examine this the ‘*spray_day*’ feature based on whether trap location is within the 1 mile from spray location on that day or not was created by using Haversine distance, an equation for measuring spherical distances on the Earth’s surface.

Spraying was performed only 2 days in 2011, and 7 days in 2013. And the spraying location coincided with traps only in 5 ‘YearWeek’. With this limited spraying it was hard to decide the effectiveness of spraying only by looking at Figure 6. The map also shows the spraying didn’t cover the entire trap and virus locations.

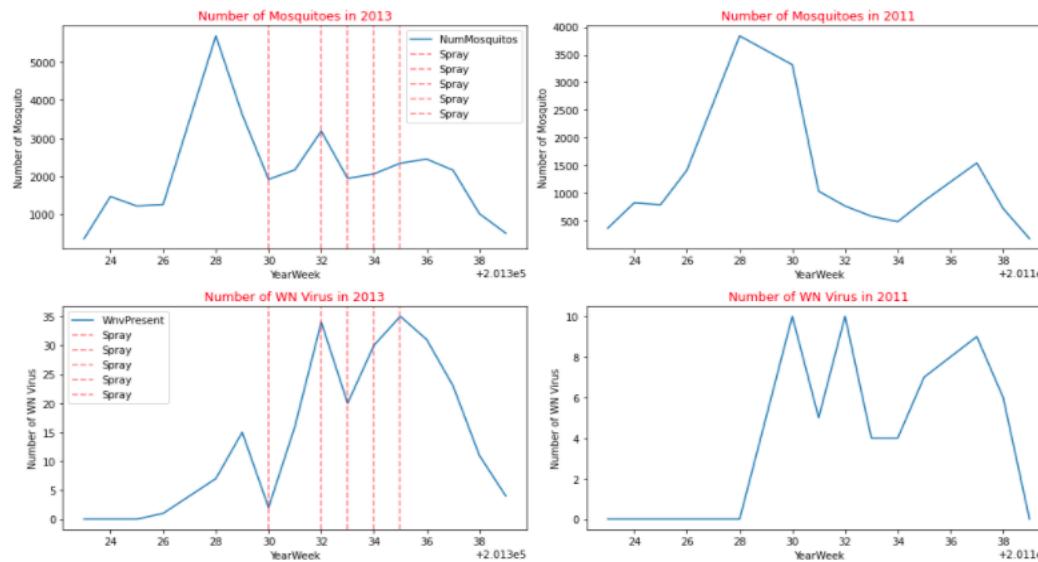
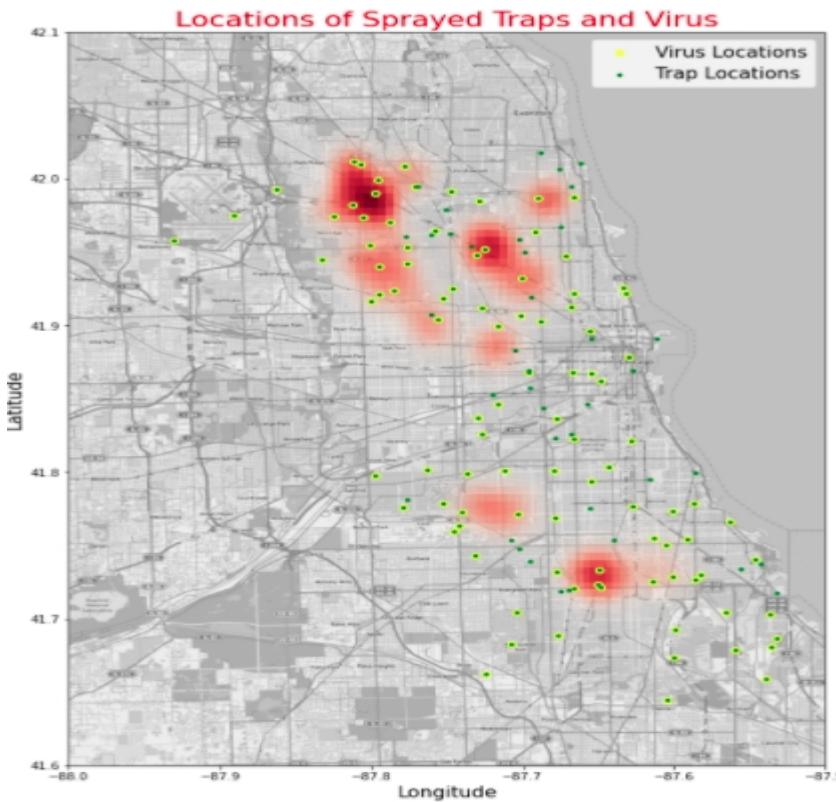


Figure 6



Therefore Permutation Test was performed by shuffling the ‘*spray_day*’ 10,000 times under the assumption Null hypothesis was true. *Null Hypothesis*: Spraying has no effect on virus. The WNV amount difference observed in spraying was due to chance. *Alternative Hypothesis*: Spraying and WNV Presence are related. Figure 7 depicts our observed distance of 0.0185 is nowhere near the distribution generated assuming the null hypothesis is true.

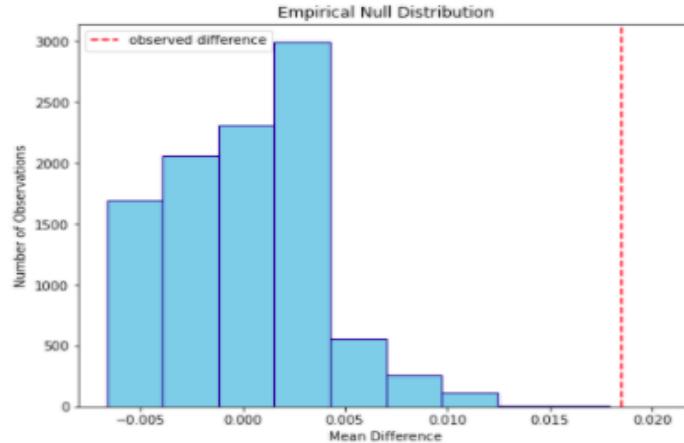


Figure 7

Additionally the P-value was computed as 0, which means our observed data was statistically significant, spraying reduced the number of WNV. Therefore the Null Hypothesis was rejected.

3.3 Weather Data:

According to the researches the cold blooded animals like mosquitoes are sensitive to subtle changes in temperature and humidity which affect the survival and reproduction rates of the WNV. So the ‘*Relative Humidity*’ was added as a feature.

In order to explore the predictive capacity of the features simple moving average and exponentially weighted moving average time-lagged variables were created by going back to two months before observations.

The dummy variables were created from the ‘Species’ column which had 6 different species. Since only ‘*CULEX PIPiens*’, ‘*CULEX RESTUANS*’ and ‘*CULEX PIPiens/RESTUANS*’ were vectors for the WNV the other 3 species were dropped.

After the feature engineering we had 1048 features. Too many variables would cause the algorithm to learn spurious structure. To overcome this challenge *WOE*(Weight of Evidence) technique was applied to measure the strength of features to separate WNV and non-WNV. Then the important features were selected by *IV*(Information Value) method. As per the below shown IV chart the features having the IV statistics less than 0.01 and greater than 0.8 were dropped. The number of features reduced to 376.

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

Extracting new data from raw data improved the model. On the other hand it caused multicollinearity. I fixed this issue by iteration process starting to drop the variable with the highest VIF(Variance Inflation Factor) until no variables with VIF higher than the threshold value of 6 remained. By this method the final data set had 10 features. The below Table ?? shows the rule of thumb for interpreting the variance inflation factor.

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

Figure 8

4. MACHINE LEARNING

4.1 Building Model:

As shown in below table, our data set consisted of 10 variables and 9693 observations.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9693 entries, 0 to 9692
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   spray_day        9693 non-null   int64  
 1   Depart           9693 non-null   float64 
 2   Species_CULEX RESTUANS 9693 non-null   int32  
 3   Species_CULEX PIPiens 9693 non-null   int32  
 4   PrecipTotal      9693 non-null   float64 
 5   ResultSpeed_ema50 9693 non-null   float64 
 6   ResultDir_lag28  9693 non-null   float64 
 7   BR               9693 non-null   float64 
 8   TS               9693 non-null   float64 
 9   FG               9693 non-null   float64 
dtypes: float64(7), int32(2), int64(1)
memory usage: 757.3 KB
```

The data was split into train and test set with the ratio 80%, 20% and stratified to ensure both sets had equal target distribution.

The below function was defined for building and evaluating models.

```
def model_evaluation(model,params,avg):
    pipe = make_pipeline(StandardScaler(),model)
    model_ran = RandomizedSearchCV(pipe,params,cv=10, n_jobs=-1, scoring = 'roc_auc',random_state = 42)
    model_ran = model_ran.fit(X_train,y_train)
    y_pred = model_ran.predict(X_test)
    y_pred_proba = model_ran.predict_proba(X_test)[:,1]
    f1 = f1_score(y_test, y_pred, average= avg)
    cm = confusion_matrix(y_test, y_pred)
    roc= roc_auc_score(y_test, y_pred_proba)
    print('F1-score: ', round(f1,4))
    print("Best Score: ", round(model_ran.best_score_,4))
    print("ROC AUC: ", round(roc,4), '\n')
    print("Best Parameters: ", model_ran.best_params_)
    print("Confusion Matrix: ", '\n', cm, '\n')
    print("Classification Report: ", '\n', classification_report(y_test, y_pred))
    y_pred_proba=model_ran.predict_proba(X_test)[:,1]
    fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
    plt.plot([0, 1], [0, 1], 'k--')
    plt.plot(fpr,tpr,label= f'{model}' )
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('ROC Curve')
    plt.show()
```

Pipeline object ‘pipe’ was created with StandardScaler() step which enabled features to contribute to model equally. And RandomizedSearchCV with 10 Folds cross validation was used for hyperparameter optimization. 10-fold cross validation split the train set into 10 folder, held out the validation set and calculated 10 accuracy scores, one for each split. Then mean validation score was computed.

The ‘fit’ method ran cross-validation for each randomly chosen parameters, found the best parameter with the highest score, which are stored in `best_params_` and `best_score_` attributes. The method also trained the "full training set" by using the best parameter and stored the best model in `best_estimator_` and evaluated it on the test set.

4.2 Metrics and Evaluation:

The goal of this project was to predict the outbreaks of WNV. However the imbalanced data with small WNV class was one of the challenges to choose the right metrics for model evaluation.

Although Recall was an important metric to evaluate how well the WNV was predicted, it wasn’t the right one for minority positive class. F1 score which balances Precision and Recall was preferred.

ROC-AUC which is invariant to class imbalance and plots the model’s performance on the positive class was used for evaluation. However as seen in the below table the ROC-AUC metric alone wasn’t enough to select the right model. The confusion matrix provided further insight into the model’s performance and see the actual and predicted positive and negative classes.

Logistic Regression, Random Forest, Gradient Boosting and XGBoost Models were analyzed by using above defined function. Below Table depicts the metrics for our models.

MODEL	F1-Score	Best Score	ROC AUC	Confusion Matrix
Logistic Regression-1	0	0.6902	0.7079	<code>[[1837 1] [181 0]]</code>
Logistic Regression-2	0.1491	0.6871	0.6971	<code>[[1187 731] [34 67]]</code>
Random Forest	0	0.7986	0.8166	<code>[[1835 3] [181 0]]</code>
Gradient Boosting	0.0185	0.7956	0.8282	<code>[[1832 6] [100 1]]</code>
XGBoost- 1	0.0192	0.7966	0.8228	<code>[[1836 2] [100 1]]</code>
XGBoost- 2	0.2466	0.7964	0.8262	<code>[[1428 418] [28 73]]</code>

Although “Logistic Regression-1” and “Random Forest” models had high ROC AUC score their F1 Scores were 0. The confusion matrix shows they missed all WNV.

The cost-sensitive learning method which gives higher weight to minority class and lower weight to majority class was applied to “Logistic Regression-2”, “XGBoost-2” by adding the parameters `'class_weight='balanced'` and `'scale_pos_weight:[18]'` respectively. Their F1-scores increased by predicting 67 and 73 WNV correctly respectively.

4.3 Best Model and Feature Importance:

The above metrics depict XGBoost-2 model performed the best.

The Figure 9 shows the best features identified by the model but it doesn't indicate the direction of the impact. For example both the species Culex Pipiens and Culex Restuans have high feature importance. However we know from our data that Culex Restuans had the lowest WNV rate. So it should have negative impact on the WNV.

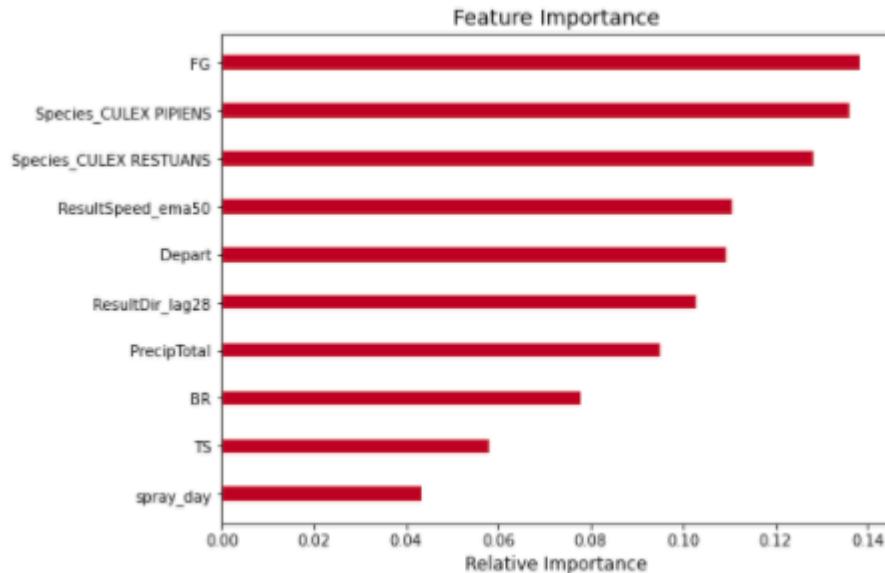
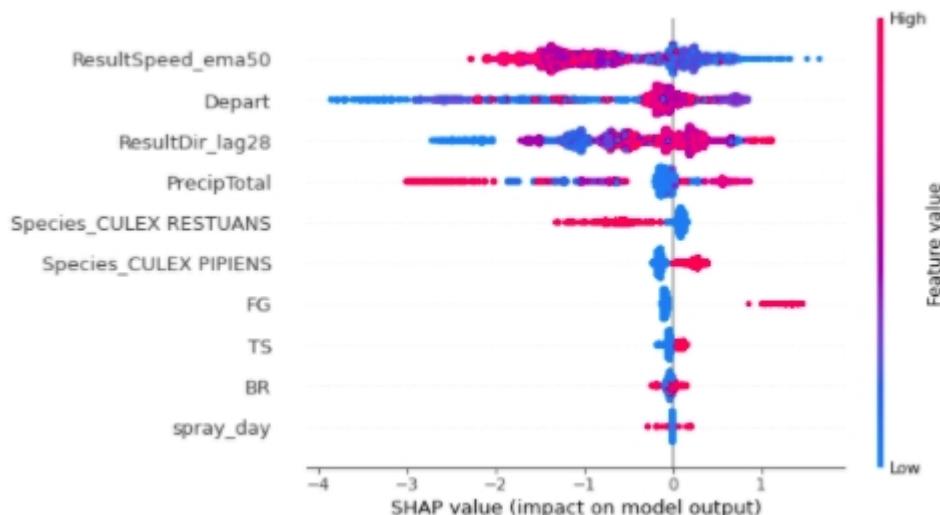


Figure 9

5.SHAP

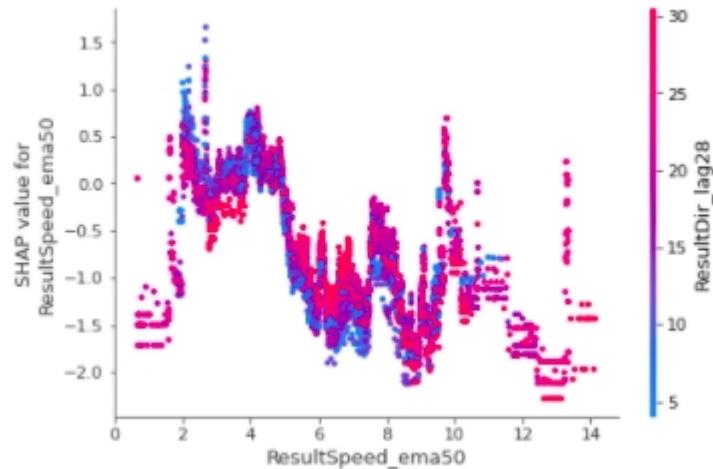
In order to interpret the impact of the features correctly SHAP (SHapley Additive exPlanations) which measures the influence of a feature by comparing model predictions with and without the feature was used.

The blow summary plot shows the positive(Red color) and negative(Blue color) relationships of the features with the target. For example high 'ResultSpeed_ema50' has a negative impact on the presence of the WNV, whereas 'FG' has positive impact. To see each feature on the target, dependence plot was used.

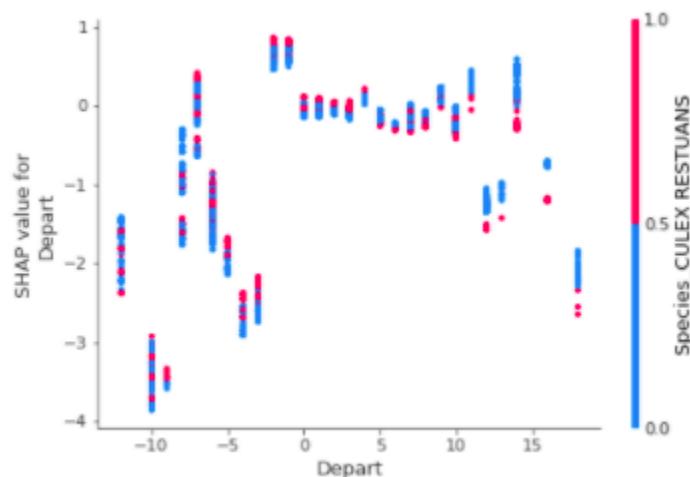


5.1 The Dependence Plots:

5.1.1 ResultSpeed_ema50: is 50-day exponentially weighted moving average of wind speed where recent values are given a higher weight. There is negative correlation between the presence of WNV and high wind speed. The presence of WNV is associated with the low wind speed ranged between 2-5 mph after a lag of 50 days, while the wind direction is between 0-10 degree after a lag of 28 days. (ResultDir_lag28) According to researches the natural transmission between birds and mosquitoes impacts the WNV. Mosquitoes pick the virus up when they feed on infected birds. Wind speed and direction affect bird migration which impact WNV spreading also.

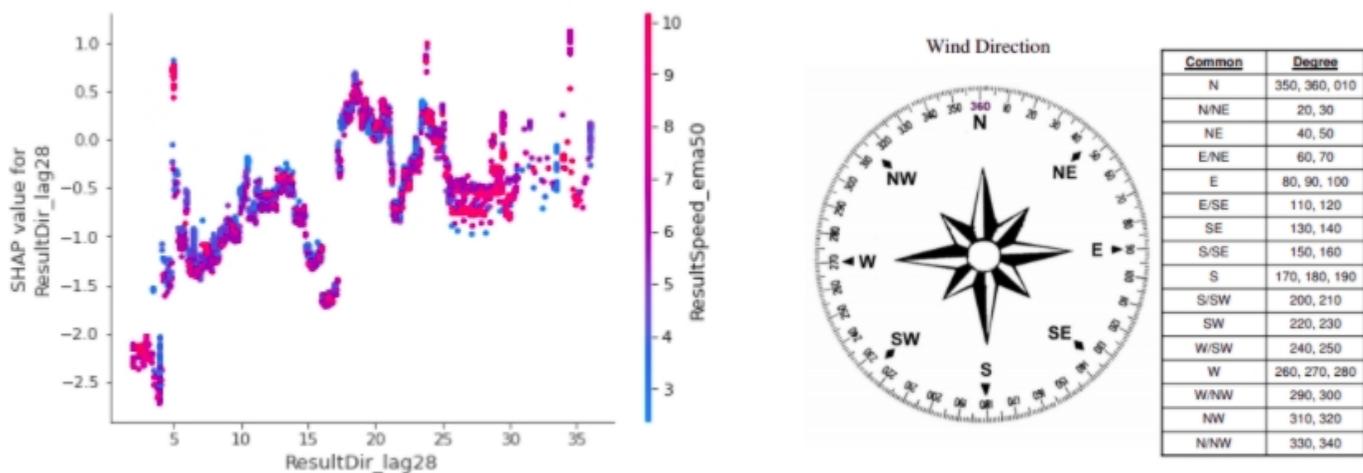


5.1.2 Depart: (Departure From Normal) is the difference between Average Temperature and the 30 year normal temperature for this date. A minus (-) indicates that the average for that day was below normal. The plot indicates temperature below normal has negative correlation with the WNV presence, while temperature above normal is positively associated. However temperature difference above 10 F has negative impact on the virus presence. The excessively hot temperatures slow down the mosquito activity and increases mortality. The interaction between Depart and Species_Culex Restuans is negative. The blue color means the species is not Culex Restuans. The higher the depart value, less the Culex Restuans. Our data set indicates the virus is transmitted primarily by Species CULEX PIPiens. Therefore the positive SHAP value mostly consists of species other than Culex Restuans.

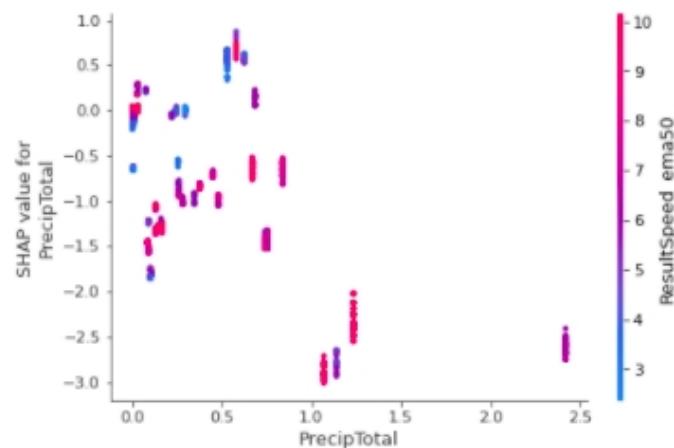


5.1.3 ResultDir_lag28: is 28 days average value of the resultant wind direction (tens of degrees) prior the observation. The Wind direction is the direction from which the wind blows. For example, a north or northerly wind blows from the north to the south. We know that wind speed and direction affect most animal flight, including mosquitoes and birds.

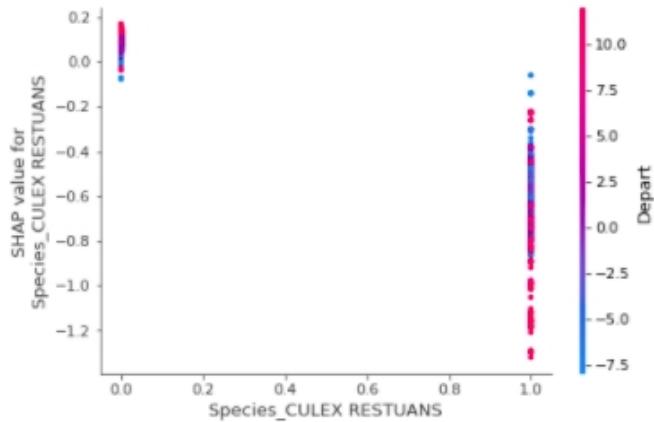
Wind direction 180, 230, 280 degrees has positive correlation with the WNV. As per the below chart the related degree range shows the wind is from South, Southwest and West, from the land towards the Michigan Lake. Since the wind direction from the land towards the Michigan Lake belongs to 28 days average prior the WNV presence, it refers back to the life cycle of mosquitoes and indicates how the wind direction helped providing habitat for adult females to lay eggs near the water. However SHAP value is negative for the ranges 200-230, 240-280. This may require more detailed study on Wind direction.



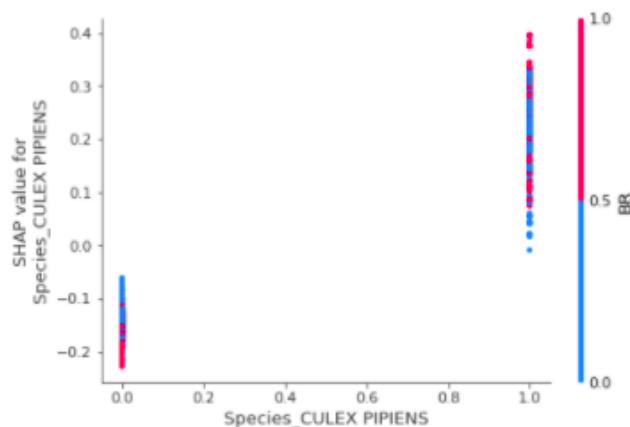
5.1.4 PrecipTotal: is the total daily rainfall (inch). The plot shows the low precipitation (0-0.7 inches) has positive impact on the WNV presence when the 'ResultSpeed_ema50' is low. When the 'ResultSpeed_ema50' is high, low precipitation has no effect on the prediction. According to researches heavy rainfall may dilute the nutrients for larvae, thus decrease the development rate. And plot shows High precipitation has negative correlation with the WNV.



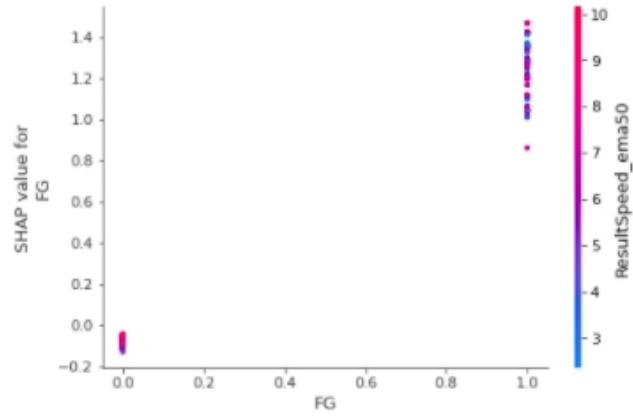
5.1.5 Species_CULEX RESTUANS: The x-axis value 1 means the species of mosquito is Culex Restuans whereas 0 means the species other than Culex Restuans. Culex Restuans has negative correlation with WNV, while the species other than Culex Restuans with high ‘Depart’ have positive SHAP value. This supports the article which states “Temperature is known to influence mosquito population dynamics, increasing the reproduction rate, the number of blood meals, and the duration of the breeding season.” (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3924437/>)



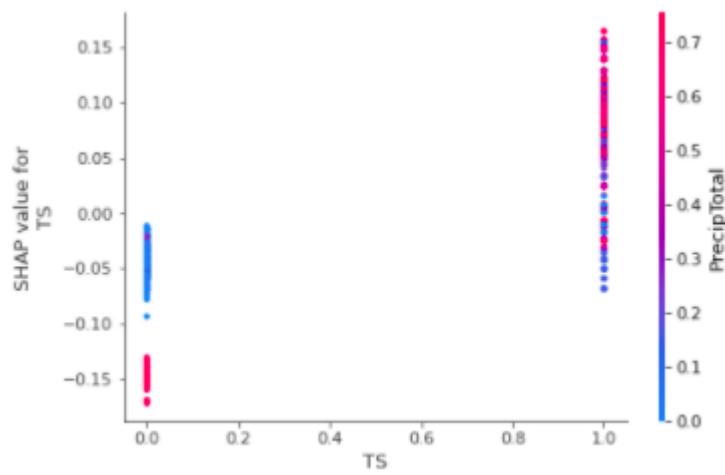
5.1.6 Species_CULEX PIPiens: The above plot indicated there is highly positive correlation between Culex Ppipens and WNV. The academic researches state the Species Culex Pipiens mosquitoes are major vectors of West Nile Virus. (<https://www.ecdc.europa.eu/en/all-topics-z/disease-vectors/facts/mosquito-factsheets/culex-pipiens-factsheet-experts>)



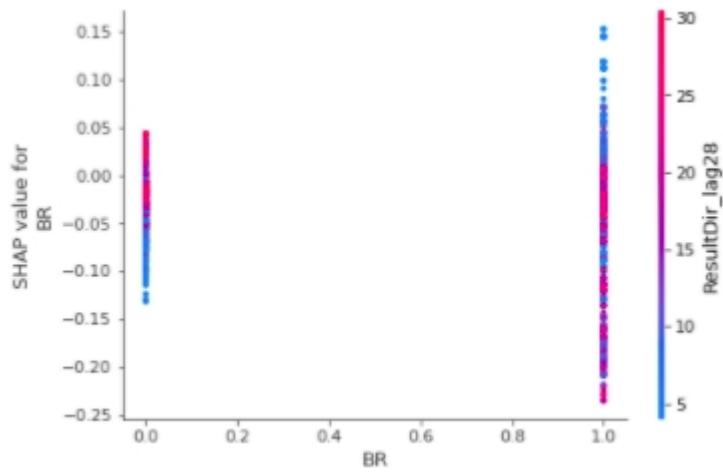
5.1.7 FG: Fog forms when the relative humidity increases to 100% by the change of temperature. It is heavily influenced by nearby bodies of water, topography, and wind conditions. The location of traps are close to the Lake Michigan where fog occurs. Therefore the fog is highly correlated with the presence of WNV. It has an interaction with the wind.



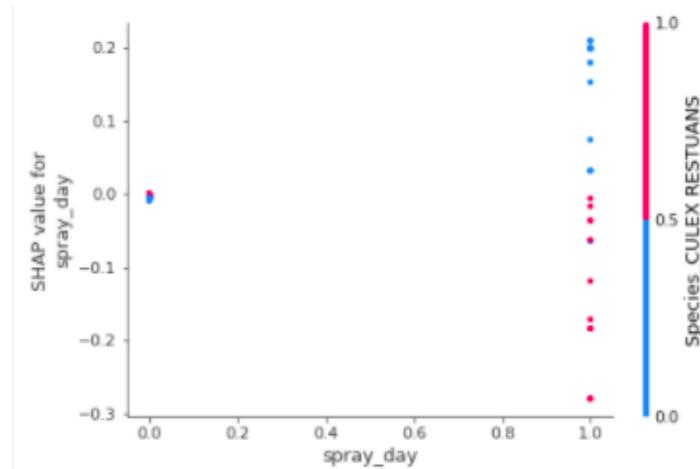
5.1.8 TS: Thunderstorm is almost always associated with lightning, thunder, dense clouds, heavy rain, and strong gusty winds. Thunderstorm with high precipitation has positive correlation with the WNV.



5.1.9 BR(Mist): Like fog, mist is the result of the suspension of water droplets, but simply at a lower density. And the plot shows the interaction of mist with the wind from North at 28 day-lag has positive correlation with WNV.

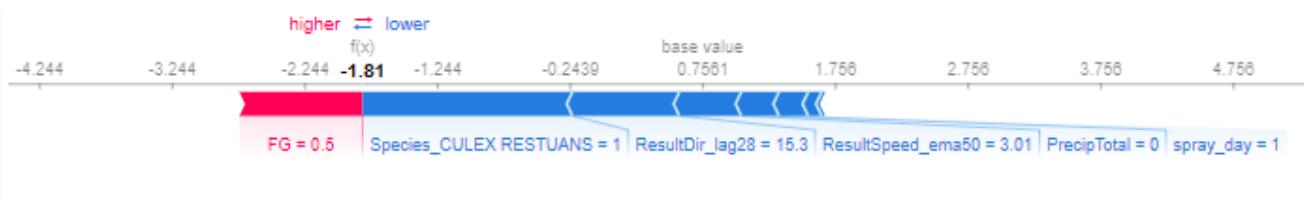


5.1.10 spray_day: is the day the location within 1 mile of trap is sprayed. Red ones are Culex Restuans, blue one are other species. This plot is not well interpretable. The reason is spraying day is the feature that occurs after WNV occur. It kills mosquitos, reduces their population, and "in the long term" causes WNV to disappear. Therefore it is hard to see the effect on the same day. The summary plot also shows the spraying day is the least influential variable for the prediction of WNV. In addition to this we have only 10 spray days which makes it hard to make inference about the effect of spraying. As we mentioned above Culex Pipens are vectors for WNV and spray killing this species (blue ones) has positive correlation with the WNV.



5.2. Force Plot Explanation:

To better understand the impact of features on predictions let me explain SHAP on the below shown individual observation. The base value, 0.7561 is stored in the expected_value attribute of the explainer object and it is the average of the predictions in a featureless model. The output value is -1.81, which means when the above features are added to our base model, the model classifies this observation as no-WNV. The red feature (presence of FOG) pushes the prediction toward WNV. The blue ones pushes the prediction toward no-WNV. Spraying has a negative impact on the WNV. When it is sprayed it pushes the prediction toward no-WNV. Presence of Species_CULEX RESTUANS has a negative impact on the WNV also. We know from our data the most of the major vector of WNV is Culex Pipiens. Precipitation has a positively relation with the WNV. In this observation having no precipitation decreases the probability of observing WNV and pushes the prediction value towards -1.81. The size of the arrow shows the magnitude of the feature's effect. Species has the biggest impact on decreasing the prediction towards.



Temperature is known to influence mosquito population dynamics, increasing the reproduction rate, the number of blood meals, and the duration of the breeding season. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3924437/>)

6. CONCLUSION

In this project I tried to find out the reasons of WNV outbreaks and the effectiveness of spraying on WNV based on the surveillance data.

The exploratory analysis indicated the main factors contributing to the WNV spread included precipitation, wind, species Culex Pipiens, spray and indirect effect of humidity. Also time lagged wind speed and direction which had correlation with bird migration and bird-mosquito transmission cycle increased the WNV spread.

In order to better understand the impact of spraying I examined the mapping of spraying area and the variation of mosquitoes and virus after spraying. One of the challenges in this project was the insufficient spraying data which made difficult to interpret its impact in reducing the WNV. Therefore I used permutation test to measure the effect of the spray, which concluded spray had negative impact on WNV occurrence.

Based on the above mentioned features different models were built. However imbalanced data was one of the other challenges in choosing the right metrics. To overcome this cost-sensitive learning method was applied to models. Since our goal was to predict the positive minor class F1 score and ROC-AUC which plots the model's performance on the positive class were used for evaluation. XGBoost with 0.8262 ROC-AUC score and 0.2466 F1-score had the best performance.

To understand the impact of the features in our model's prediction SHAP was used. This method laid bare the required improvements on the model. For example *ResultDir_lag28*(wind direction at the 28 lag) correlation wasn't clear in the plot.

7. FUTURE WORK

The relationship between the WNV and ResultDir_lag28, spray_day, mist were not well interpretable. They require more detailed study.

Although our model had high ROC-AUC score, it couldn't detect the WNV well. One of the reasons should be the imbalanced data. For imbalanced data other approaches like resampling, SMOTE can be applied.

8. RECOMMENDATIONS

Based on the results the below recommendations would help to improve our model.

The map indicates spraying missed many traps with the virus. Therefore spraying areas should be changed accordingly. Also adding more spray data would help to see its impact more clearly.

Taking into consideration the bird-mosquito transmission features about bird migration should be included.

Although spray is effective in WNV control, it has significant impacts on human health and other organisms in the environment. Cost-effectiveness analysis can be useful to decide to continue spraying or not. Furthermore other options like decreasing the number of breeding sites for mosquitoes, releasing natural predators like Mosquitofish in small ponds, and canals should be analyzed.