

Predictive Analysis on West Nile Virus

Derya Kurt

Table of Contents

1 INTRODUCTION & PROBLEM STATEMENT	2 DATA & PREPROCESSING	3 EXPLORATORY DATA ANALYSIS
4 MODEL BUILDING & ANALYSIS	5 CONCLUSION & FUTURE WORK	6 RECOMMENDATIONS

1



INTRODUCTION & PROBLEM STATEMENT

What is West Nile Virus?

- mosquito-borne virus
 - spread rapidly across the United States

2002 - CHICAGO

225 Cases

22 Fatalities



Problem Statement

- ▶ Predicting when and where different species of mosquitoes will test positive for West Nile virus.
- ▶ Finding out the effectiveness of spraying

2

DATA & PREPROCESSING

Data Overview

Main Data

10506 observations

2007, 2009, 2011, 2013

Traps Location, Trap Date,
Number of Mosquitoes,
Species, WNV

Spray Data

14835 observations

2011, 2013

Date and Location of Spray

Weather Data

2944 observations

2 days in 2011, 7 days in 2013

2 Weather Stations, Weather
Conditions, Observation Date

Data Preprocessing

Data Cleaning

DELETION

- Duplicates
- Columns with high missing values
- Highly correlated Features

IMPUTATION

Missing values filled with other station values based on insignificant difference between stations

Data Preprocessing

Feature Engineering

- ✓ Merged Station 1 and Station 2 by taking their average
- ✓ One-Hot Encoding for Species and Weather Features in CODESUM
- ✓ Spray Area within 1 mile of trap calculated by Haverstine Distance
- ✓ Time-Lagged Weather Conditions

Data Preprocessing

Feature Selection & Reduction

- ✓ Feature Selection by WOE (Weight of Evidence) + IV(Information Value)
- ✓ Reduced Multicollinearity by VIF(Variance Inflation Factor)

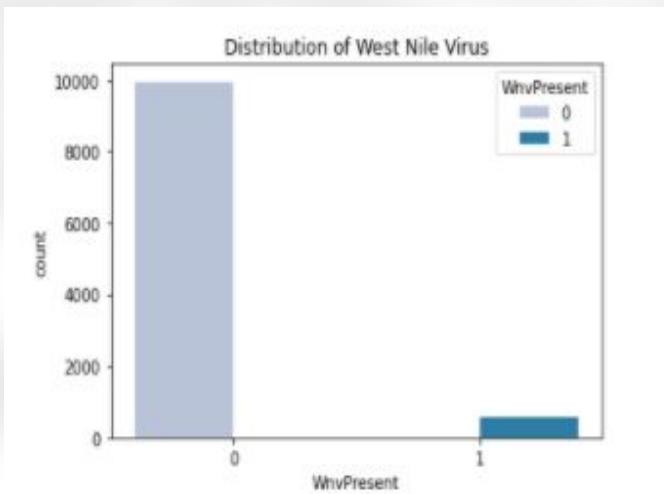
Final Features

1. spray_day
2. FG
3. Species_Culex Pipiens
4. TS
5. PrecipTotal
6. BR
7. ResultDir_lag28
8. ResultSpeed_ema50
9. Species_Culex Restuans
10. Depart

EXPLORATORY DATA ANALYSIS

3

Imbalanced Data



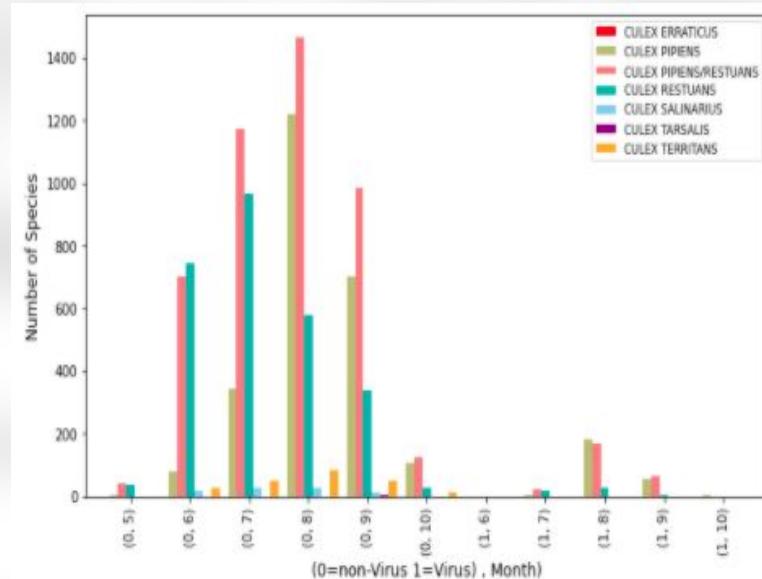
West Nile Virus Absent- 94.8%



West Nile Virus Present- 5.2%

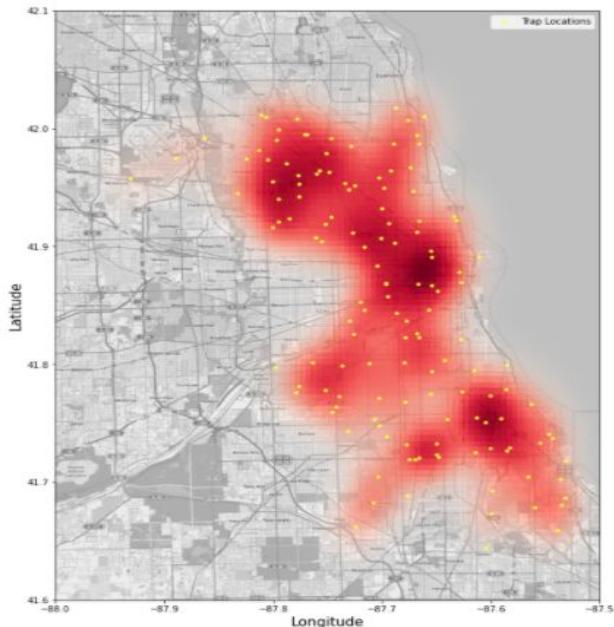
Distribution of Species per Month

- Vectors of the WNV - Culex Restuans , Culex Pipiens
- Culex Pipiens - Main contributor of the WNV
- August- Highest number of Mosquitoes and WNV



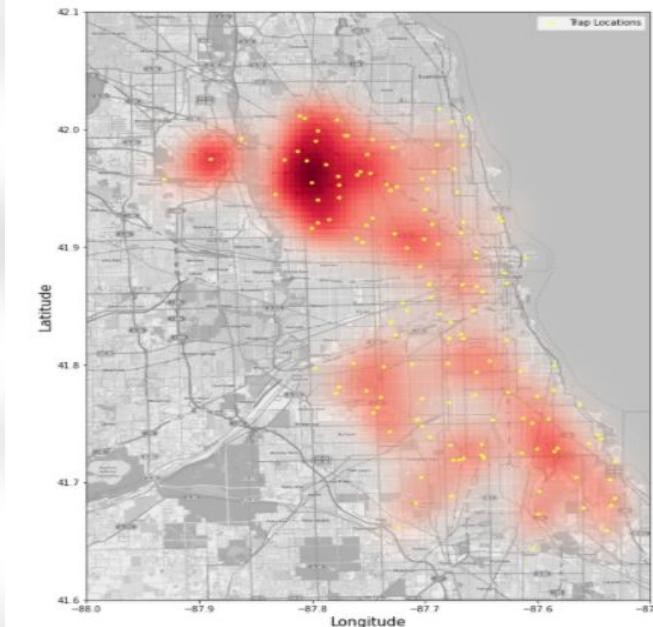
Distribution of Mosquitoes & WNV

Mosquitoes



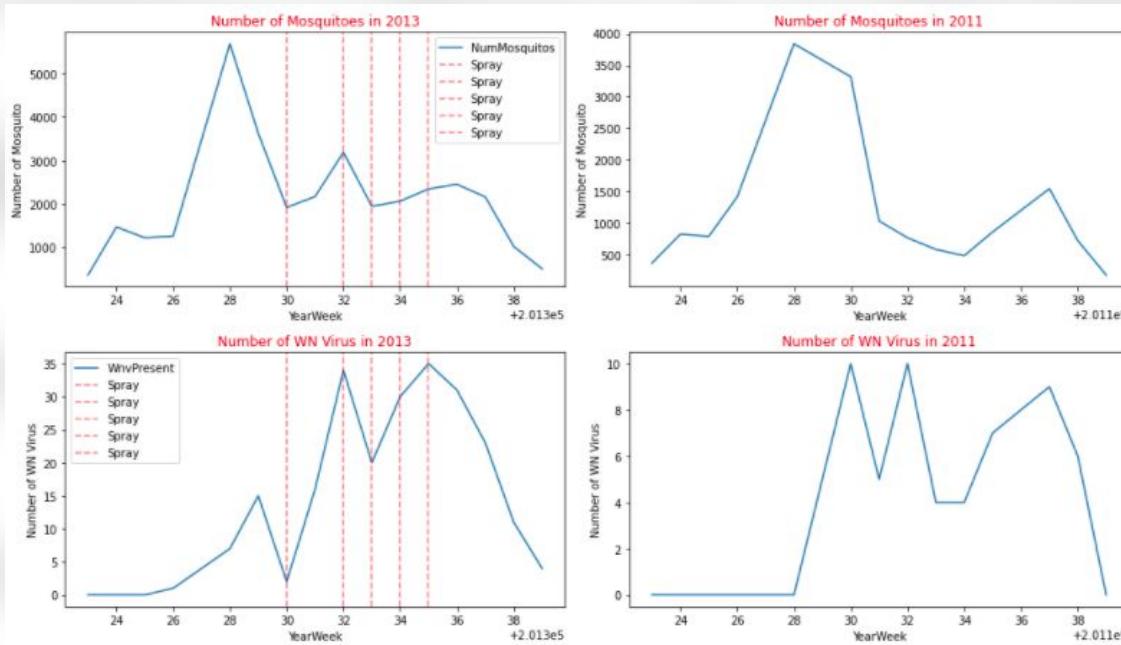
Mosquitoes mostly found close to the Lake Michigan

WNV



WNV mostly found in the northern traps.

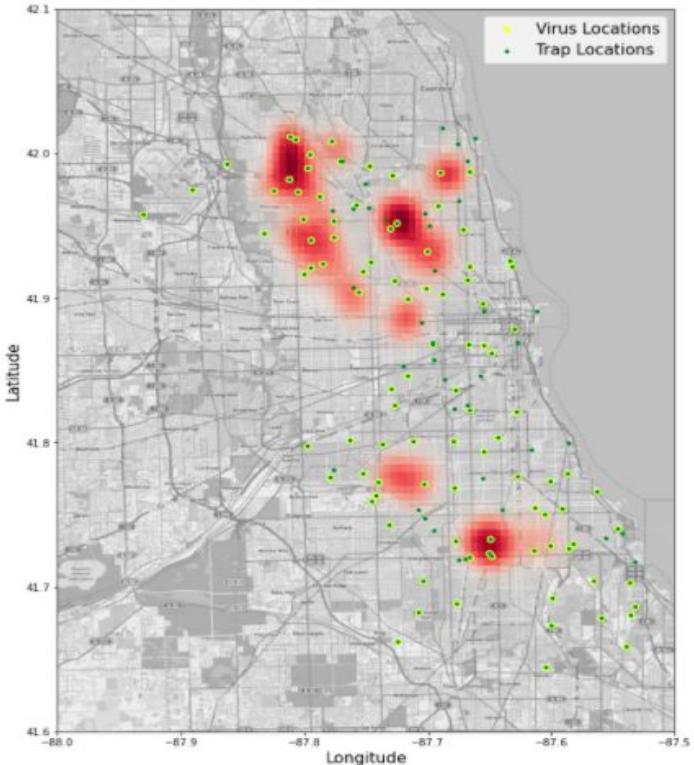
Effect of Spraying



Spraying - 2 days in 2011, and 7 days in 2013

Limited Spraying has insignificant impact on mosquitoes and WNV

Location of Sprayed Traps & WNV



Insufficient spraying

Many trap and virus locations not sprayed

MODEL BUILDING & ANALYSIS

4

Model Building Processing

Hyper-parameter Optimizer	Preprocessing	Metrics	Models	Imbalanced Data Learning
RandomizedSearchCV	StandardScaler()	ROC-AUC F1-Score Confusion Matrix	Logistic Regression Random Forest Gradient Boosting XGBoost	Cost-Sensitive Learning



Imbalanced Data & Evaluation Metrics



because...

ROC-AUC score

- robust to imbalanced data

F1-Score

- balances Precision and Recall

Confusion Matrix

- further insight into the model's performance

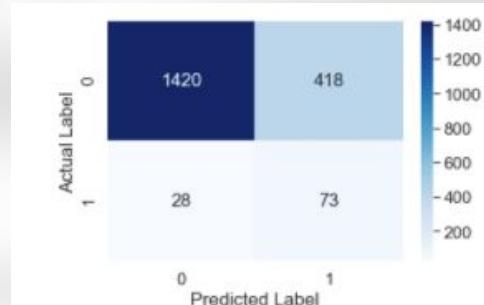
Cost-Sensitive Learning

- minimizes the cost by giving greater weight to minor class WNV

Model Evaluation

Model	F1-Score	Best Score	ROC-AUC	Confusion Matrix
Logistic Regression-1	0.00	0.6902	0.7079	$\begin{bmatrix} 1837 & 1 \\ 101 & 0 \end{bmatrix}$
Logistic Regression-2	0.1491	0.6871	0.6971	$\begin{bmatrix} 1107 & 731 \\ 34 & 67 \end{bmatrix}$
Random Forest	0.00	0.7986	0.8166	$\begin{bmatrix} 1835 & 3 \\ 101 & 0 \end{bmatrix}$
Gradient Boosting	0.0185	0.7956	0.8282	$\begin{bmatrix} 1832 & 6 \\ 100 & 1 \end{bmatrix}$
XGBoost - 1	0.0192	0.7966	0.8228	$\begin{bmatrix} 1836 & 2 \\ 100 & 1 \end{bmatrix}$
XGBoost - 2	0.2466	0.7964	0.8262	$\begin{bmatrix} 1420 & 418 \\ 28 & 73 \end{bmatrix}$

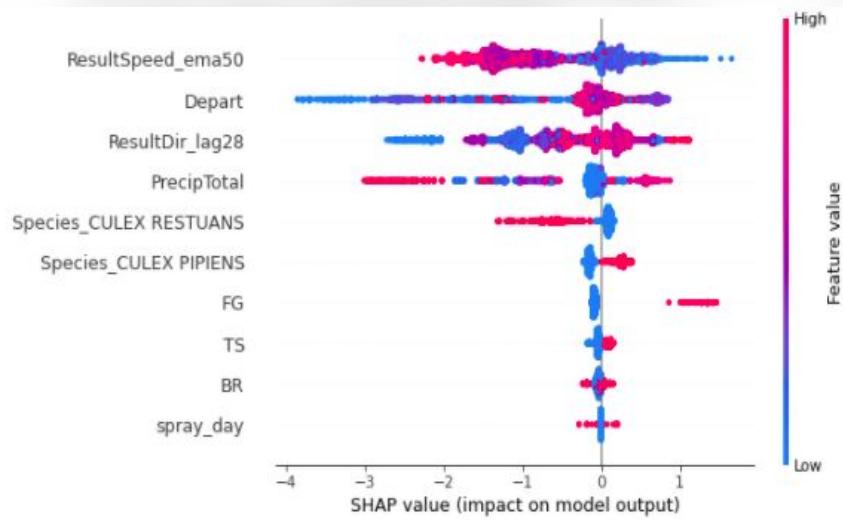
The Best Model - XGBoost



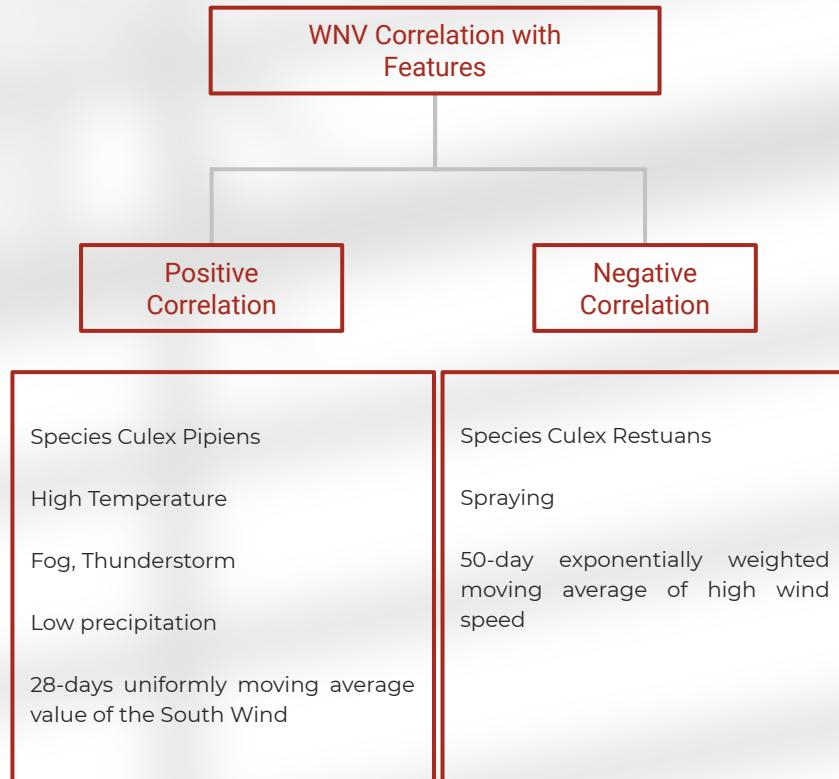
Cost-sensitive learning method improved our model in predicting 73 WNV correctly.

SHAP(SHapley Additive exPlanations)

SHAP- provides to interpret the impact of the features



SHAP Analysis



5

CONCLUSION & FUTURE WORK

CONCLUSION

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

FUTURE WORK

- Spraying, Mist, Wind Direction require more detailed study.
- Different sampling approaches like Upsampling, SMOTE, Downsampling can be applied to handle the imbalanced data.

6

RECOMMENDATIONS

Recommendations

New Features

Based on bird-mosquito transmission features about bird life cycle and bird migration can be useful to improve our model.

Spraying

Many traps were missed in spraying. Spraying area should be revised accordingly. More Spray data included.

Alternatives to Spraying

Spraying has significant impacts on human health and other organisms in the environment

Decreasing the number of breeding sites for mosquitoes, and releasing natural predators like Mosquitofish in small ponds, and canals should be analyzed.