

West Nile Virus Prediction

Predict West Nile virus in mosquitoes across the city of Chicago

Data Science Immersive Course
Project 4



Team Members

Data Science Immersive Course
Project 4



Jean



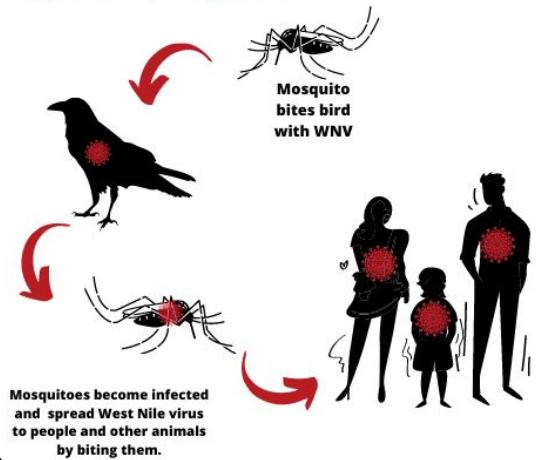
Paint



James

The Brief

West Nile virus (WNV) Cycle



Data is provided by
Chicago Department of Public Health.

The West Nile virus is primarily transmitted to humans through infected mosquitoes, with around 20% of infected individuals exhibiting symptoms ranging from persistent fever to severe neurological illnesses, potentially leading to death. The virus was first reported in humans in Chicago in 2002. In response, the City of Chicago and the Chicago Department of Public Health (CDPH) established an ongoing surveillance and control program in 2004.

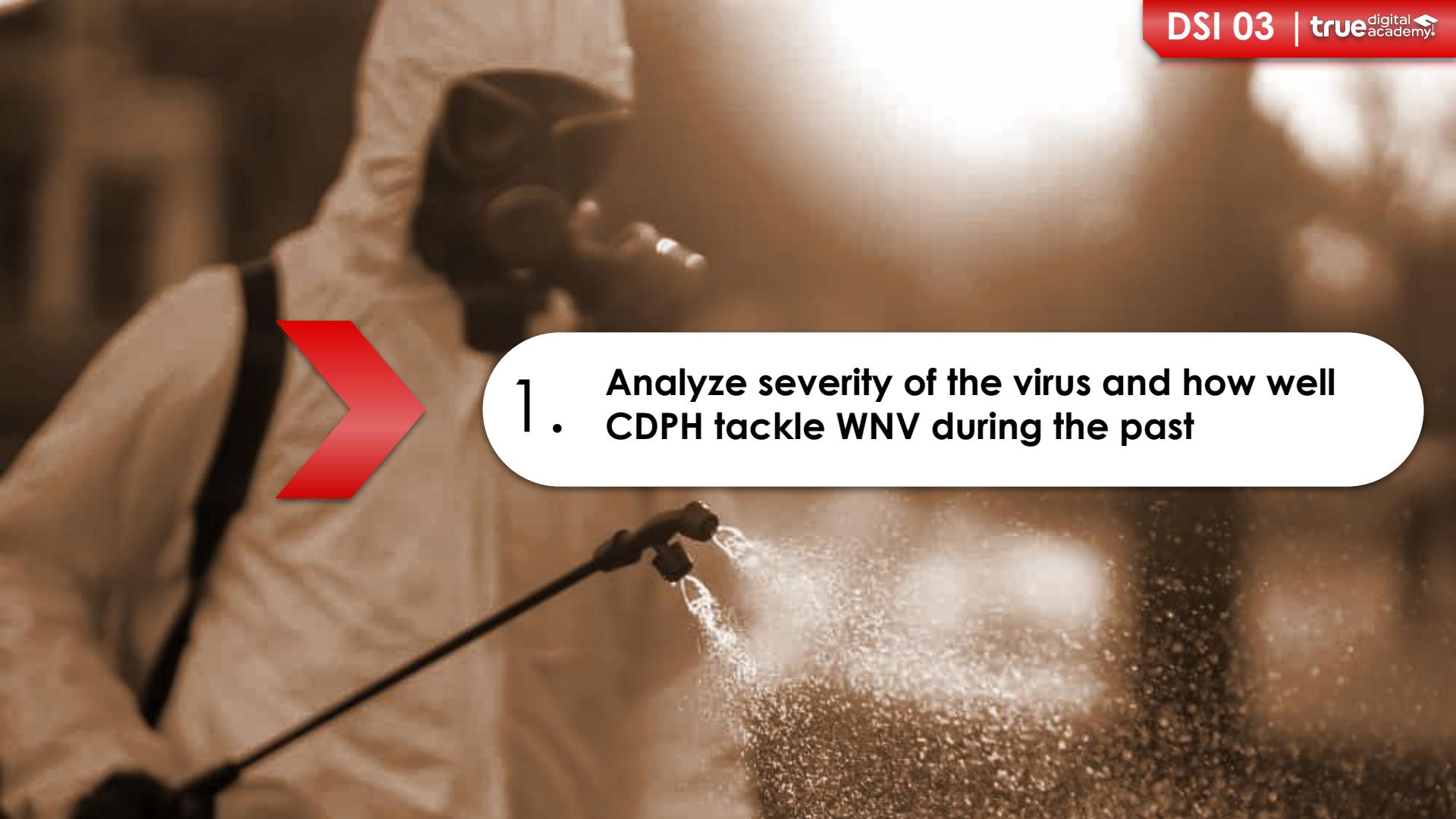
As part of this program, **mosquitoes in traps throughout the city are regularly tested for the West Nile virus from late spring to fall.** (5-10) The results of these tests play a crucial role in determining when and where the city will conduct airborne pesticide spraying to control adult mosquito populations.

There is a competition hosted by CDPH seeks predictive models based on weather, location, testing, and spraying data to forecast when and where various mosquito species are likely to test positive for the West Nile virus. Improving the accuracy of these predictions can aid the City of Chicago and CPHD in allocating resources more efficiently to prevent the transmission of this potentially fatal virus. The initiative aims to enhance public health efforts and contribute to a more proactive approach in addressing West Nile virus outbreaks. Our team decided to take part in this challenge.

Our Plan

Find out ➤

1. Analyze severity of the virus and how well CDPH tackle WNV during the past
2. Find feature characteristics that contribute to presence of WNV from predictive model.
3. Team Suggestion for CDPH next action

- 
- A large, semi-transparent image of a person in a full-body protective suit and respirator mask is visible in the background, spraying a fine mist from a long black hose. A large red arrow points from the left towards the text box.
1. Analyze severity of the virus and how well CDPH tackle WNV during the past

Vector-Borne Disease : Mosquitoes

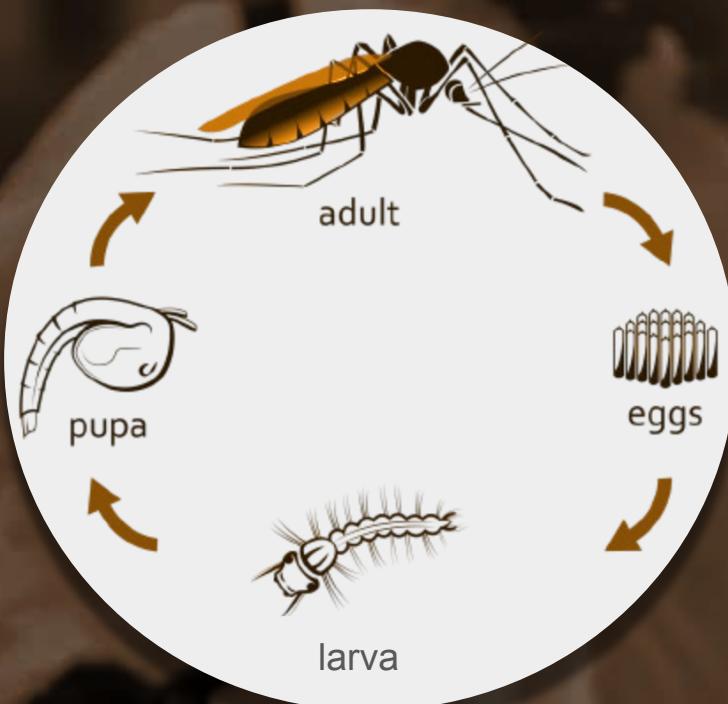


	West Nile Virus (WNV)	Dengue Fever
Transmission	Mosquito-borne (primarily Culex mosquitoes)	Mosquito-borne (Aedes mosquitoes)
Reported Cases	Thousands of cases reported in endemic regions	Millions of cases reported annually worldwide
Severity Spectrum	Asymptomatic, Mild Fever, Severe Neurological	Mild to Severe
Outbreak Frequency	Periodic outbreaks in endemic regions	Frequent in endemic regions
Endemic Regions	North America, Europe, Asia, Africa	Southeast Asia, Pacific Islands, Americas, Africa
Risk Factors for Infection	Mosquito exposure, residing in endemic areas	Mosquito exposure, travel to endemic areas
Vaccines	No specific antiviral treatment; supportive care available	Dengue vaccines available; variable efficacy

Knowing WNV

Mosquitoes are most active and thrive in **warmer temperatures**.

The development of mosquito eggs, larvae, and pupae is influenced by temperature, and their life cycle is typically faster in warmer conditions. They're less active in cooler temperatures and may become dormant in colder climates.



77 ~ 88°F

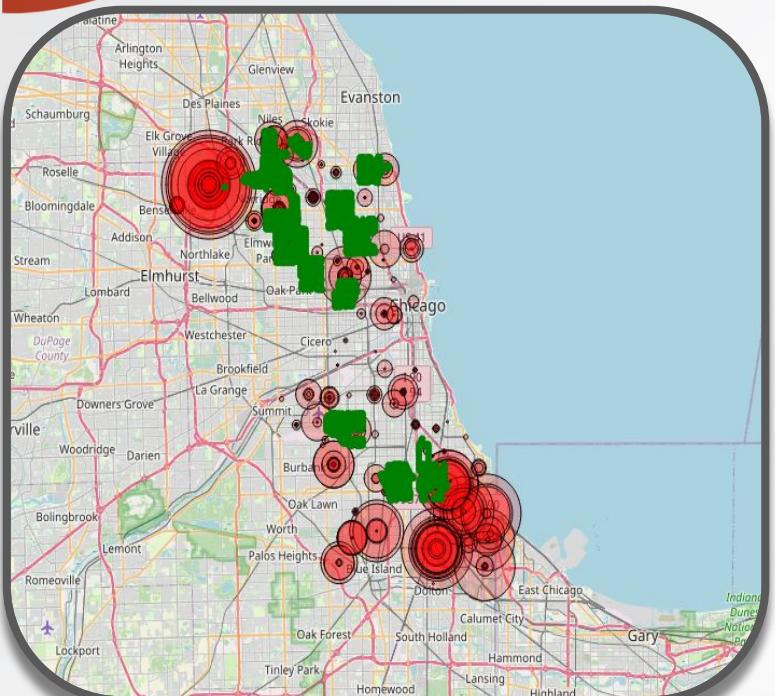
(25 to 31 degrees Celsius)

From temperatures ranging above, mosquitoes are most active and reproduce most efficiently

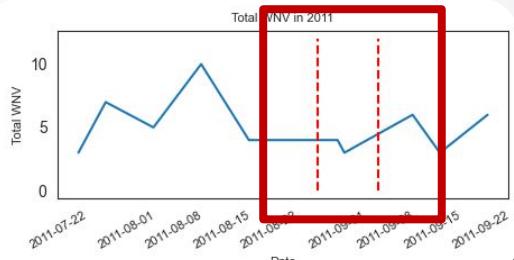
Factors such as **humidity** and the **availability of water** for breeding also play important roles in their life cycle.

Suggested Features: Temperature associated, location based season associated, Species

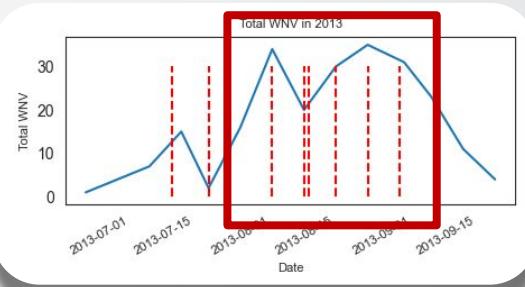
All locations and all sprays in 2007-2014



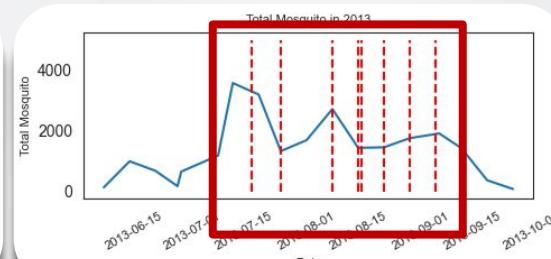
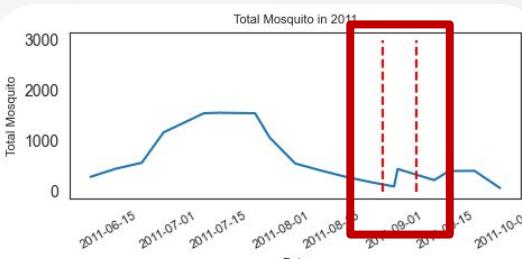
Present of WNV



----- Indicates Spray date



Number of mosquitoes



BIG DATA

- 
2. Find feature characteristics that contribute to presence of WNV from predictive model.



Name (R x C)
Train Set (10,506 x 15)
Test Set (16,293 x 14)
Weather Set (2,944 x 22)
Spray Set (14,835 x 4)
Combine Train + Weather
(Merge : Left join)



Data Cleansing

Data Imputation
(i.e. filling/dropping missing values, filtering out the odd ones)



Weather Dataset

After check M , "-" , T change to Null
- Replace Null in **Tavg** by use Tmax+Tmin/2
- Replace Null in **Preciptotal** with 0.01
Since T stands for not zero but almost Zero
- Replace Null in **Heat** with 0
- Replace Null in **Cool** with 0
- Replace Null in **Sealevel** with Median
- Replace Null in **Stnpressure** with Median
- Replace Null in **Wetbulb** with Median
- Replace Null in **Avgspeed** with Median



Modeling & Prediction

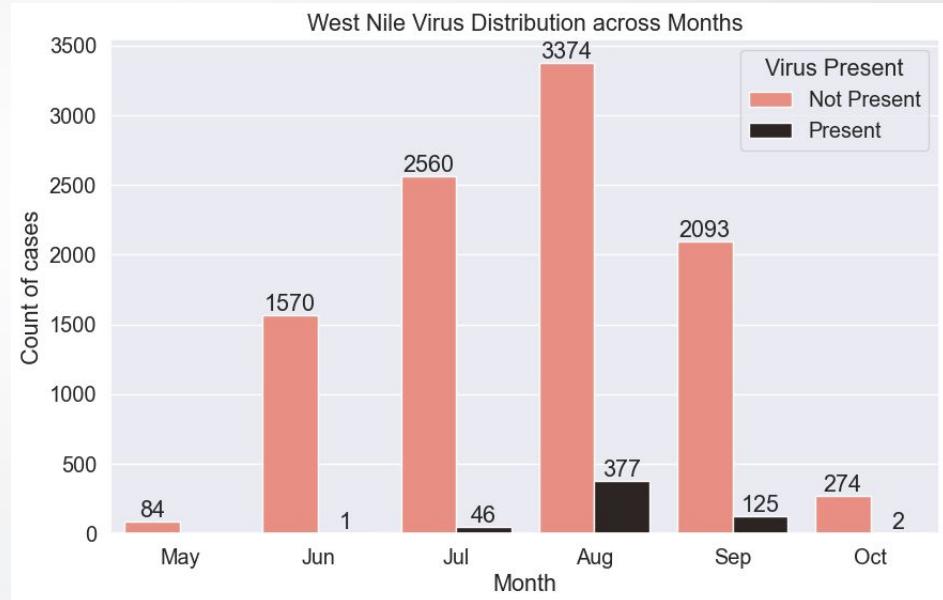
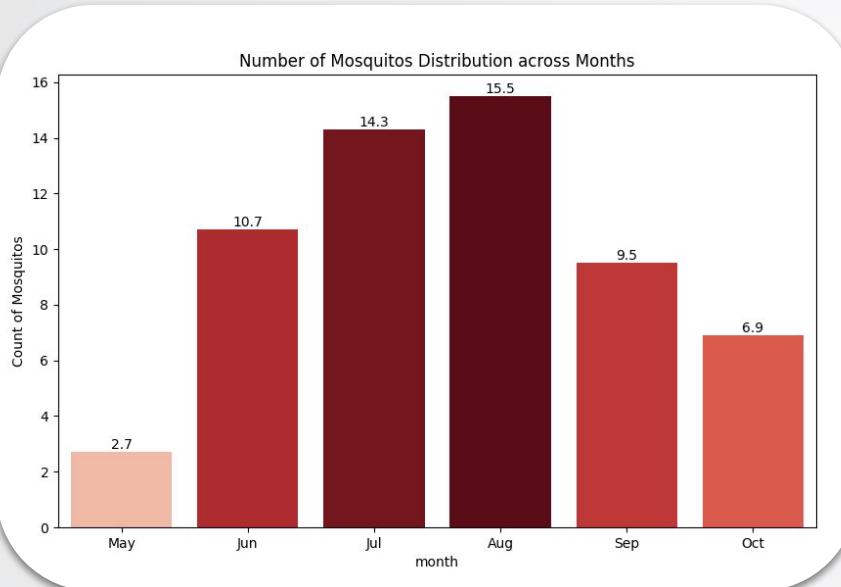


**Feature Selection
Feature Engineering**

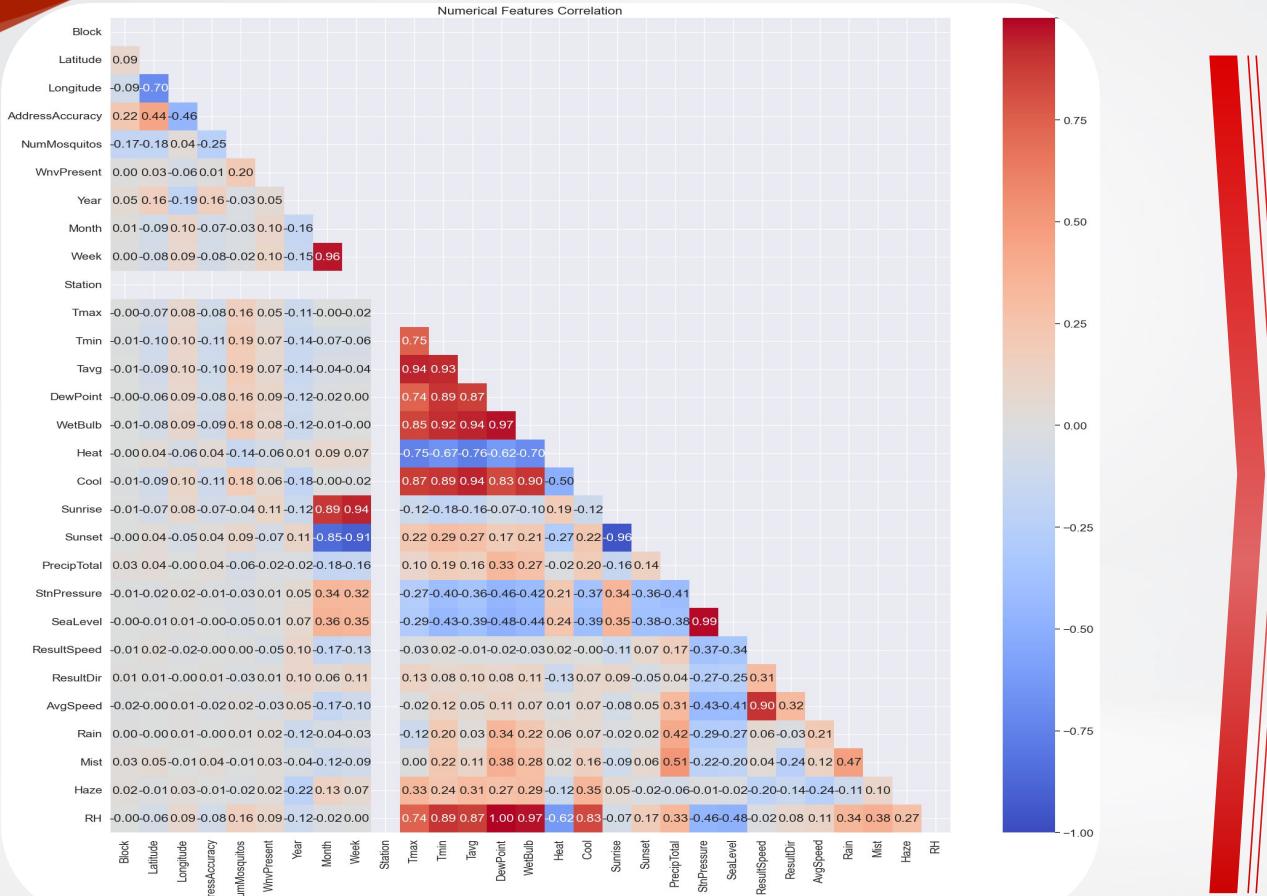
EDA

**Exploratory
Data Analysis**

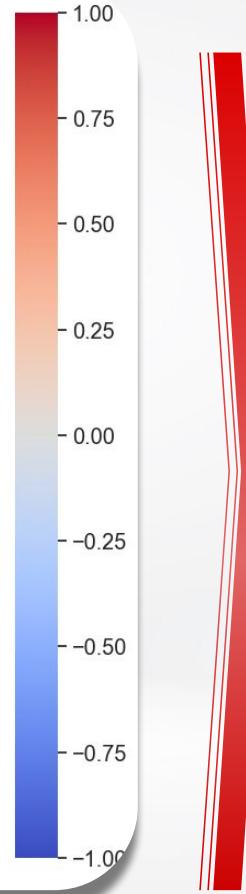
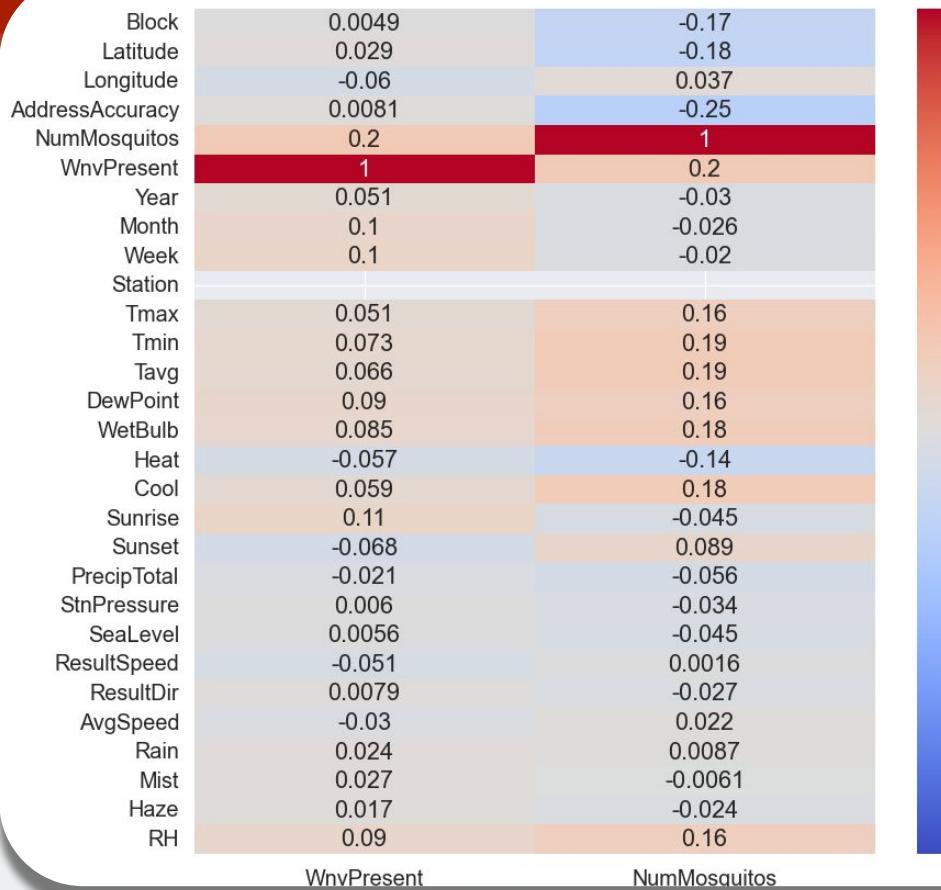




Exploratory Data Analysis:



*Red color indicates features that have multicollinearity



Virus Present

NumMosquitos
Sunrise

NumMosquitos

Climate Factors:

Tmax
Tmin
Tavg
Heat
DewPoint
WetBulb
Cool
RH

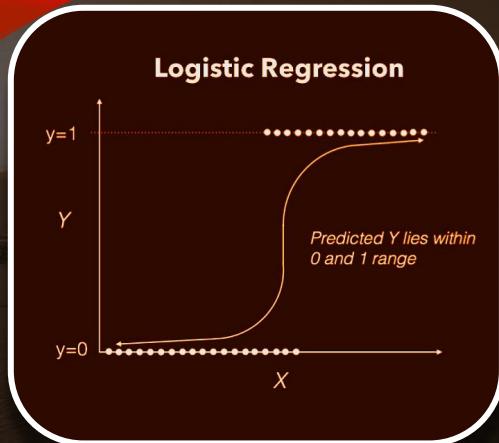
Location Factors:

Block
Latitude
Address Accuracy

Time Factors:

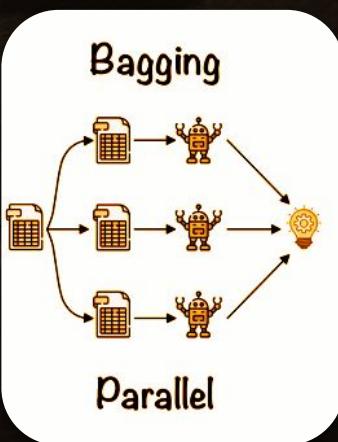
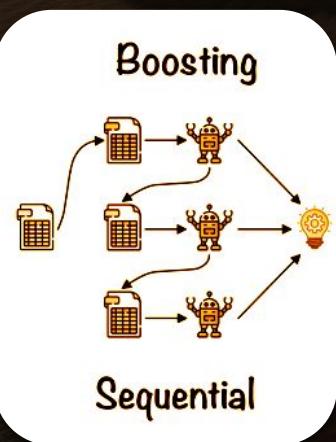
Month
Week

*Orange and Blue colors are significant features
>= +/- 0.1



Logistic Regression

logistic regression is a versatile and powerful tool for classification problems. Its interpretability, simplicity, robustness, performance, probabilistic output, and versatility make it a popular choice for a wide range of applications.



XGBoost (Extreme Gradient Boosting), Random Forest

are powerful ensemble learning techniques, but they differ in their strategies for combining multiple decision trees. XGBoost is designed for boosting, while Random Forest is designed for bagging. Each approach has its strengths and may be more suitable for different types of datasets and problems. They are commonly used for classification tasks due to several key advantages:

- **Robustness to Overfitting**
- **Ensemble Methods**
- **Tree-Based Models**

Model 1

Logistic Regression

Linear prediction for classification problem

Model 2

Random Forest

ensemble learning techniques : Bagging

Model 3

XG Boost:

ensemble learning techniques : Boosting

Model 4

Model 4 Final Improvement

applying SMOTE technique

**18
Features**

Pick 1 model for improvement
with highest AUC-ROC Score

Feature Selection

Date/Time

Date

E

Month

NEW

Year

NEW

Target (Y)

WnvPresent

Location Associated Features

Address

E

Address No &
Street

Latitude

Longitude

Block

Trap

D

Street

D

City

D

NEW

Miscellaneous Features

Address
Accuracy

NumMosquito

Depart

Species

D

Feature Engineering

D

Dummify

C

Count (iterate through items)

E

Extract into new features

Weather Associated Features

Tmax

Tmin

Tavg.

Heat

Cool

DewPoint

Sunrise

Sunset

PrecipTotal

StnPressure

ResultDir

AvgSpeed

ResultSpeed

Snowfall

WetBulB

Depth

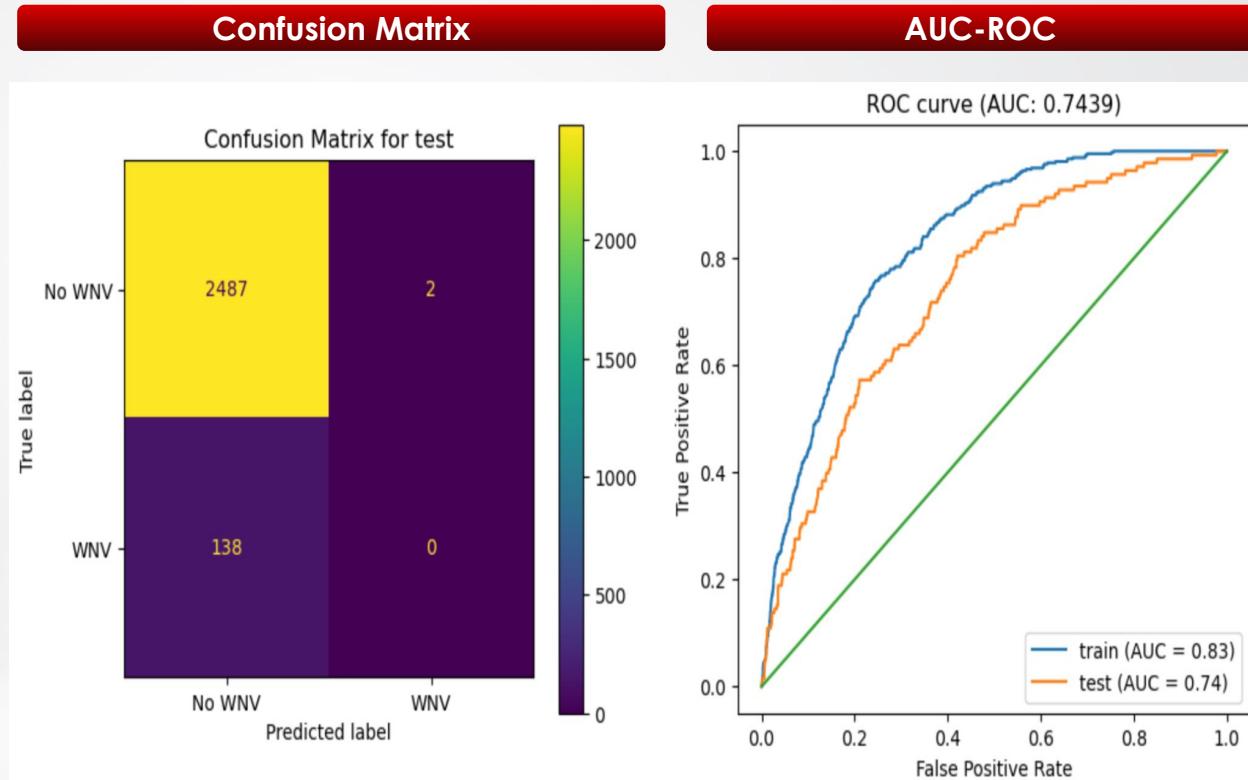
Sealevel

CodeSum

Model 1 : Logistic Regression

Metric	Score
Recall (WnvPresent:1)	0.01
Recall (WnvPresent:0)	0.99
ROC -AUC Score(Train)	0.83
ROC -AUC Score(Valid)	0.74
ROC -AUC Score(Test)	0.66

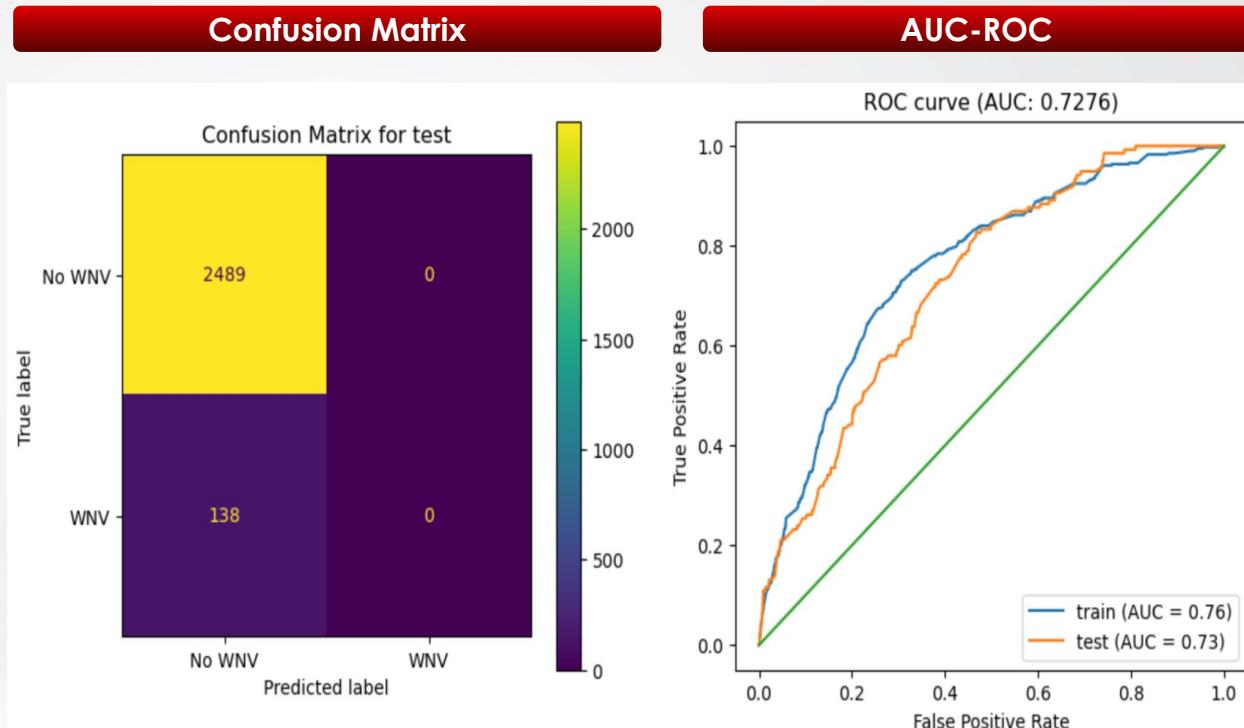
SMOTE



Model 2 : Random Forest Classifier

Metric	Score
Recall (WnvPresent:1)	0.00
Recall (WnvPresent:0)	1.00
ROC -AUC Score(Train)	0.76
ROC -AUC Score(Valid)	0.73
ROC -AUC Score(Test)	0.68

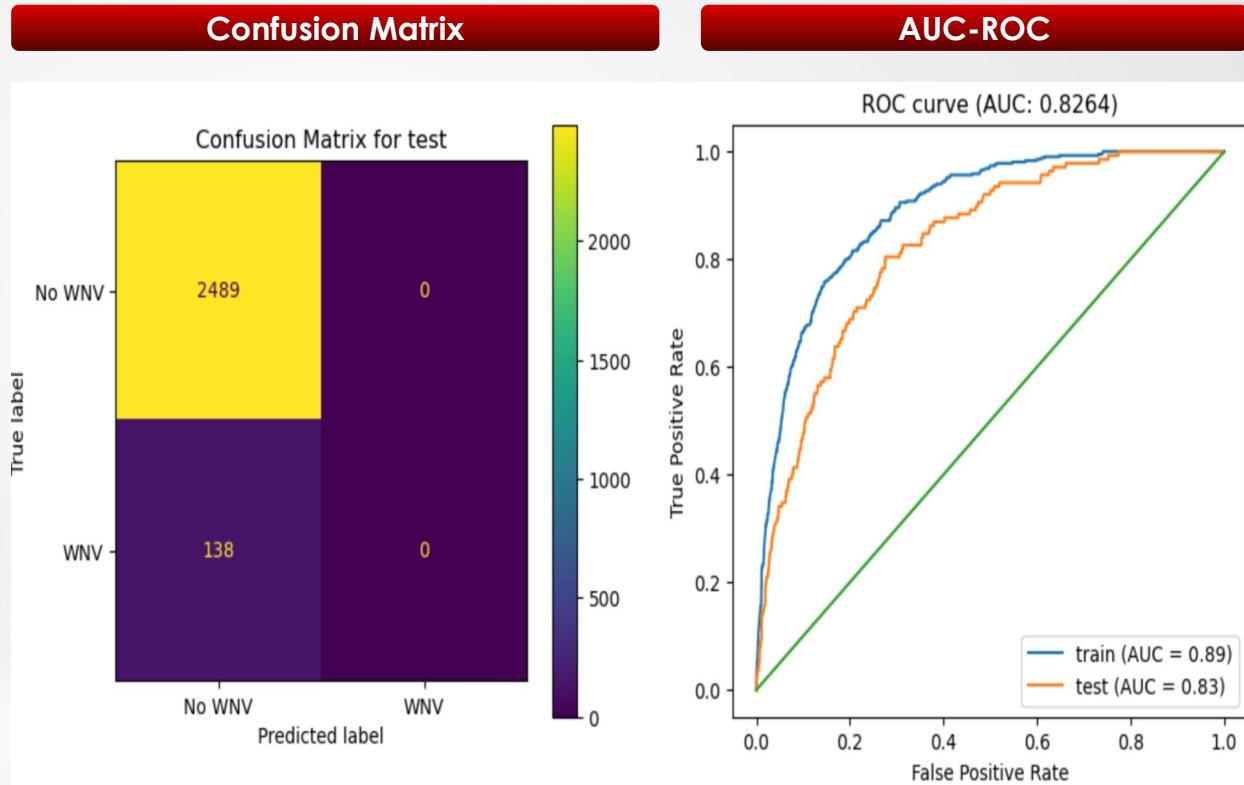
SMOTE



Model 3 : XGboost Classifier

Metric	Score
Recall (WnvPresent:1)	0.00
Recall (WnvPresent:0)	1.00
ROC -AUC Score(Train)	0.89
ROC -AUC Score(Valid)	0.83
ROC -AUC Score(Test)	0.697

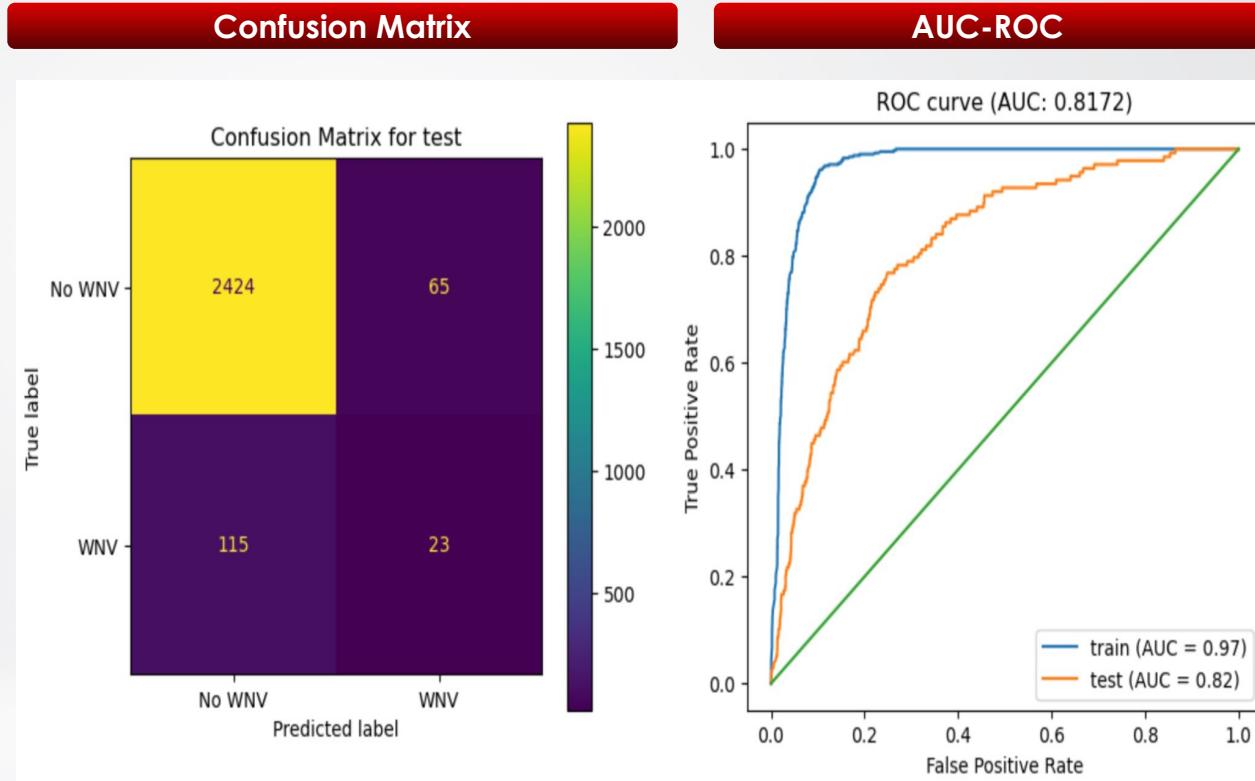
SMOTE



Model 4 : Model Improvement

Metric	Score
Recall (WnvPresent:1)	0.167
Recall (WnvPresent:0)	0.972
ROC -AUC Score(Train)	0.97
ROC -AUC Score(Valid)	0.82
ROC -AUC Score(Test)	0.73

SMOTE



Model Comparison

Metric	Model1 (LR)	Model2 (RF)	Model3 (XG)	Model4 (XG)
Recall (WnvPresent : 1)	0.01	0.00	0.00	0.167
Recall (WnvPresent : 0)	0.99	1.00	1.00	0.972
ROC-AUC Score(Train)	0.83	0.76	0.89	0.97
ROC-AUC Score(Valid)	0.74	0.73	0.83	0.82
ROC -AUC Score(Test)	0.66	0.68	0.697	0.73

XGBoost with SMOTE for Improved Classification Performance

After thorough evaluation and experimentation, the decision has been made to utilize XGBoost in combination with the Synthetic Minority Over-sampling Technique (SMOTE) for addressing class imbalance. This decision is based on the compelling performance metrics observed during model evaluation.

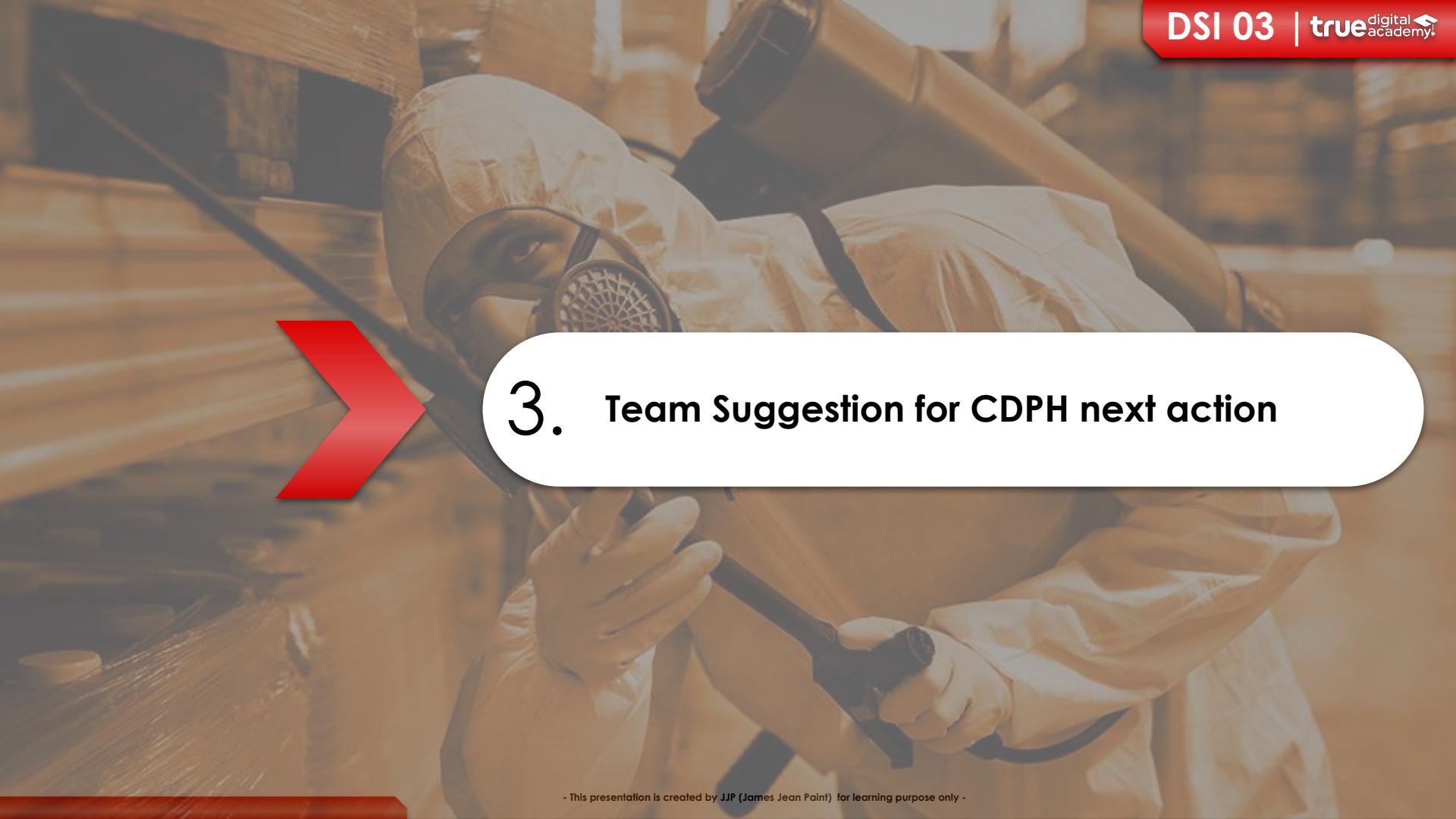
Key Findings

High Test Score(Kaggle) and ROC-AUC Score

- XGBoost with SMOTE achieves a notable Kaggle Score of **0.73** and an ROC-AUC Score of **0.82**. These high scores indicate not only effective performance on the provided dataset but also a strong potential for generalization to unseen data.

Improved Recall for Class 1(Present WNV)

- The recall for Class 1 with XGBoost and SMOTE is 0.167, showcasing a substantial improvement compared to the recall of 0 observed without SMOTE. This enhancement is particularly crucial in scenarios where correctly identifying instances of Class 1 is of paramount importance.

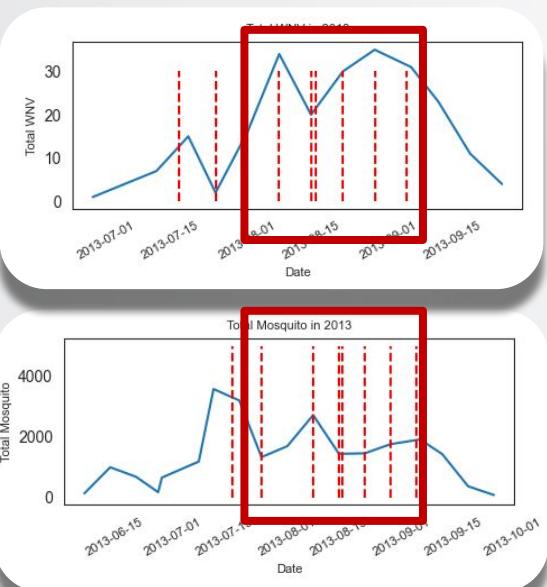


A large red arrow points from the left towards a white speech bubble containing the section title.

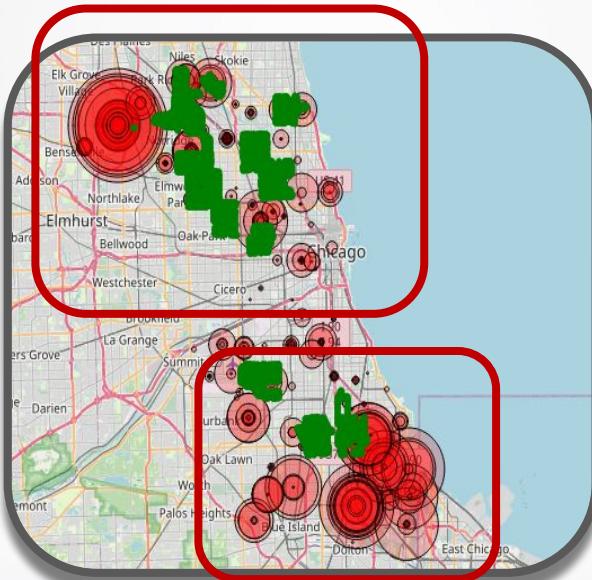
3. Team Suggestion for CDPH next action

More Mosquitos = More chance for WNV presence

Key Action: Eliminating mosquito breeding sites & Wide coverage spraying



Spraying alone is
an ineffective method



CDPH may consider revamp strategy to
eliminating mosquito breeding sites & wide coverage spraying.

Recommended Location

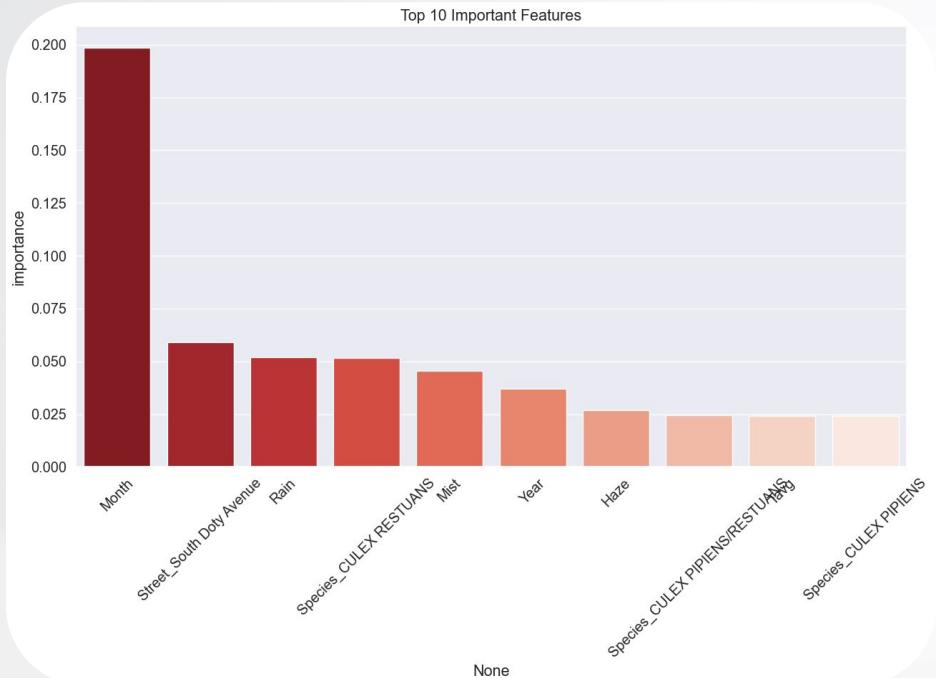
Northern and Southern
of Chicago

- Chicago O'Hare Int'l Airport
- South Doty Ave.
- North Oak Park Ave.
- South Stony Ave.
- Milwaukee Ave.
- And some more..

Vulnerable Target Residing

- Hospitals
- Elderly care centers

Key Action: Extra Surveillance & Monitoring



- Take extra Cautious Surveillance During **June to September** every year.
- **Warmer temperature** plays key role in mosquitoes activity & reproduction
- **Historical infected locations** tend to reoccur the outbreak spread and need be check whether they're hosting favorable mosquitoes breeding site

Thank you