West Nile Virus Prediction

Predict West Nile virus in mosquitoes across the city of Chicago

Data Science Immersive Course
Project 4

Team Members

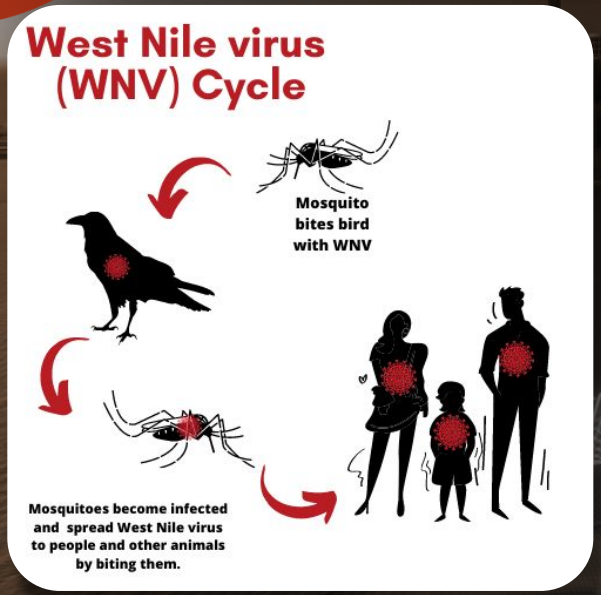
Data Science Immersive Course
Project 4

Jean

Paint

James

Brief Recap



Data is provided by
Chicago Department of Public Health.

The Brief

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 aerial 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.

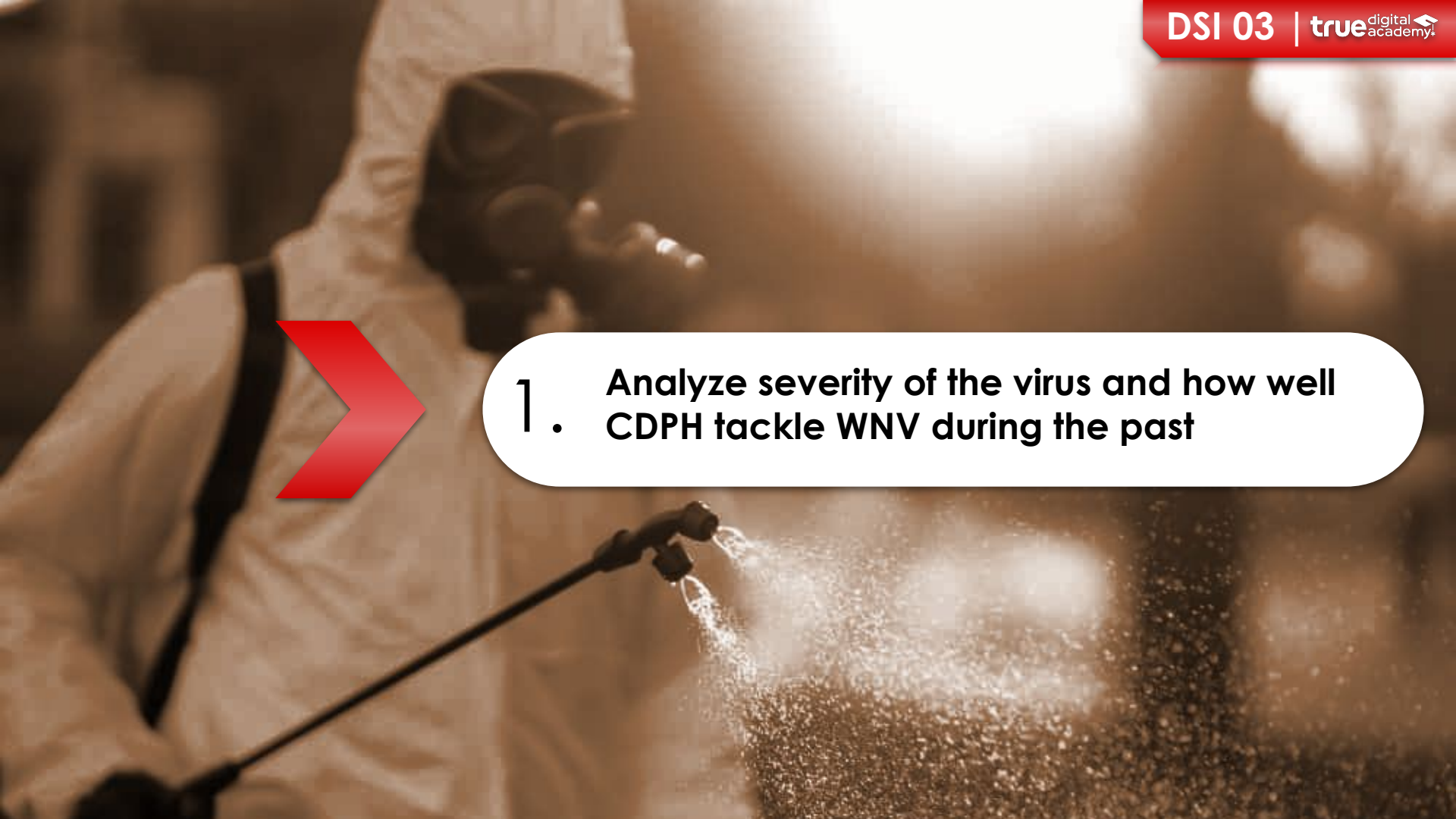
Our Plan

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.

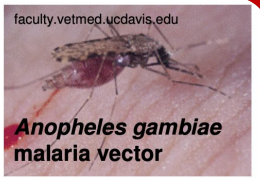
3. Team Suggestion for CDPH next action



1. Analyze severity of the virus and how well CDPH tackle WNV during the past

Knowing WNV

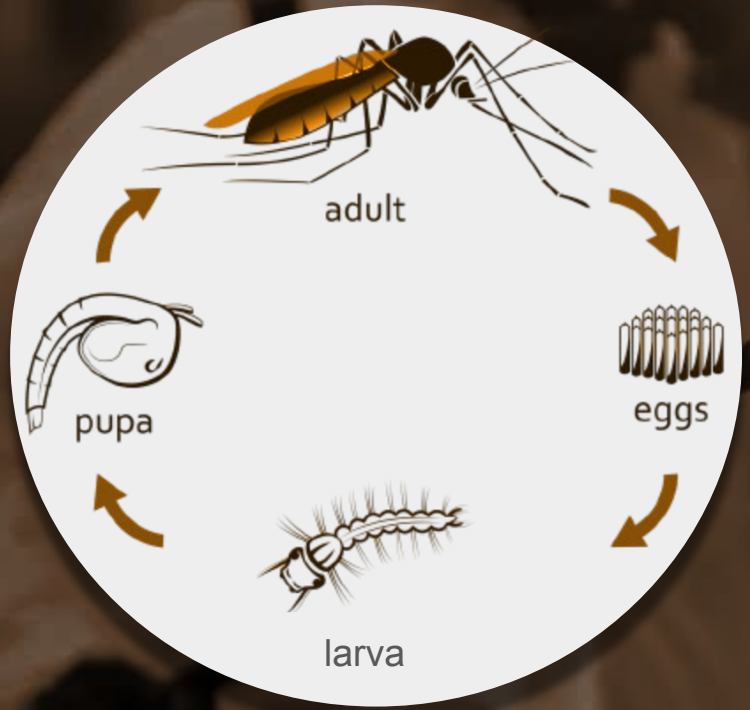
Vector-Borne Disease : Mosquitoes



	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 Disease	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 ~ 88F

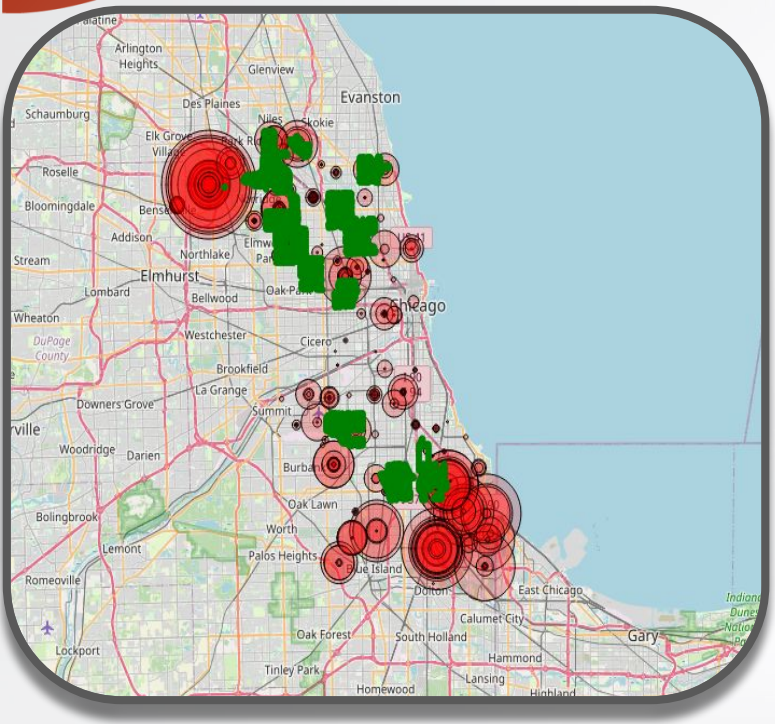
(25 to 31 degrees Celsius)
From temperatures ranging above, mosquitoes are most active and reproduce most efficiently

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

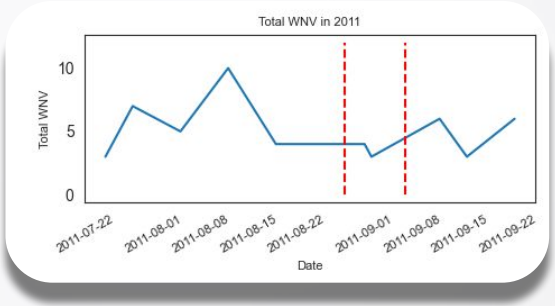
Exploratory Data Analysis: Past Performance

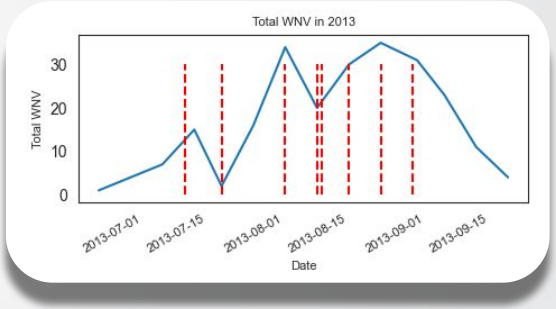
All locations and all sprays in 2007-2014



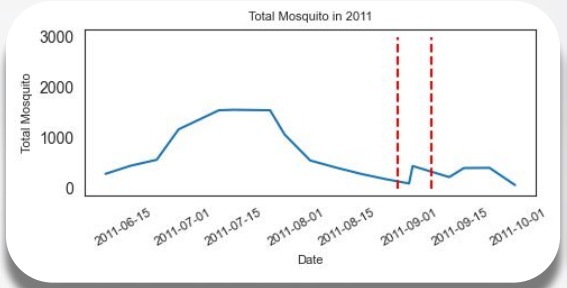
------ Indicates Spray date

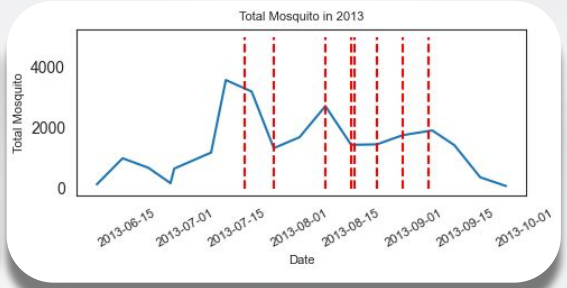
Present of WNV

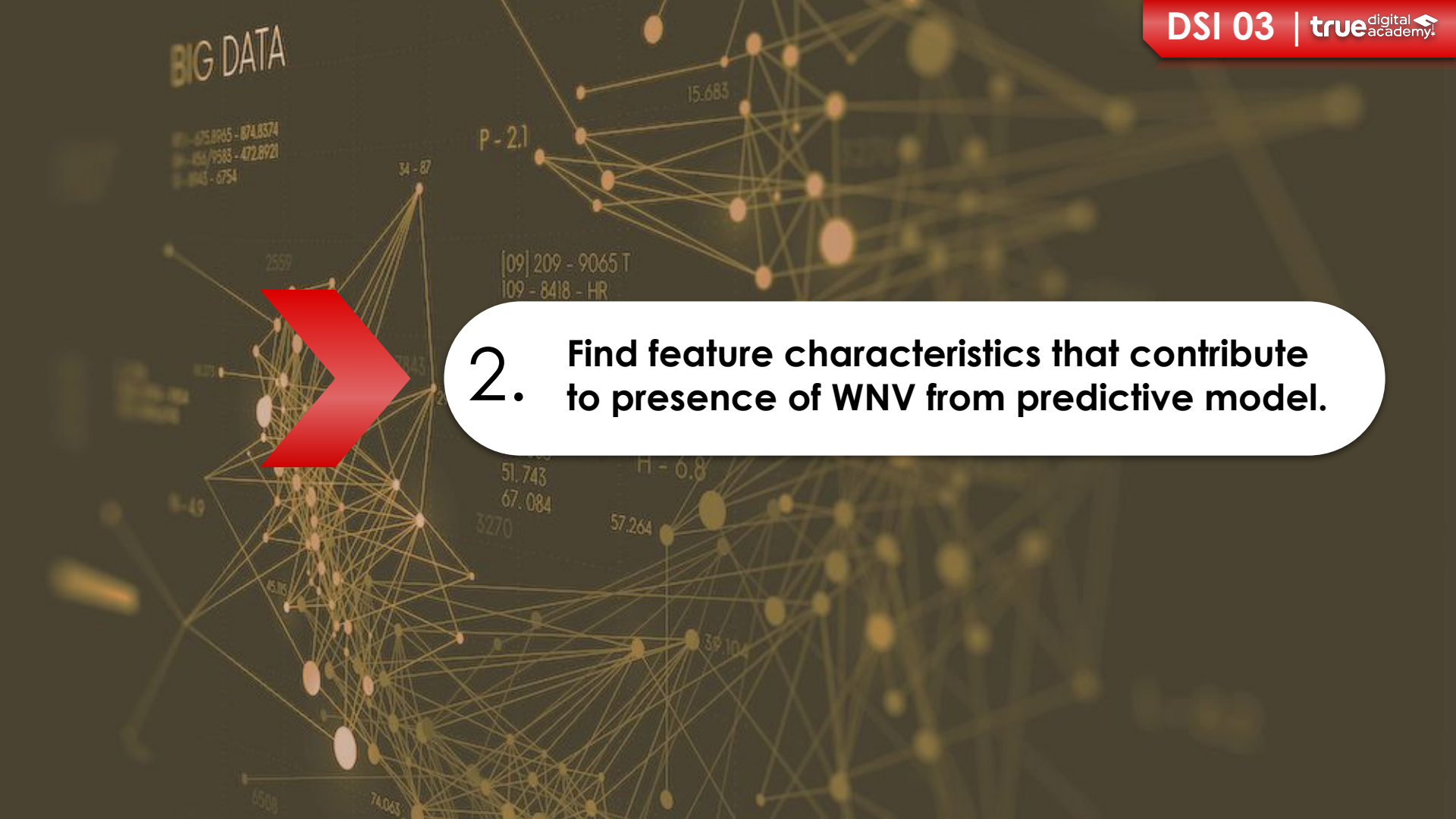




Number of mosquitoes







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

Data Processing Workflow

4 Datasets

Name	(R x C)
Train Set	**(10,506 x 15)**
Test Set	**(16,293 x 14)**
Weather Set	**(2,944 x 22)**
Spray Set	

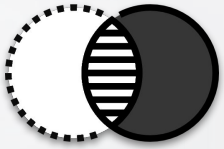
Data Cleansing

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

Train & Test Dataset
- Change type data of **Date** to Datetime

Modeling & Prediction

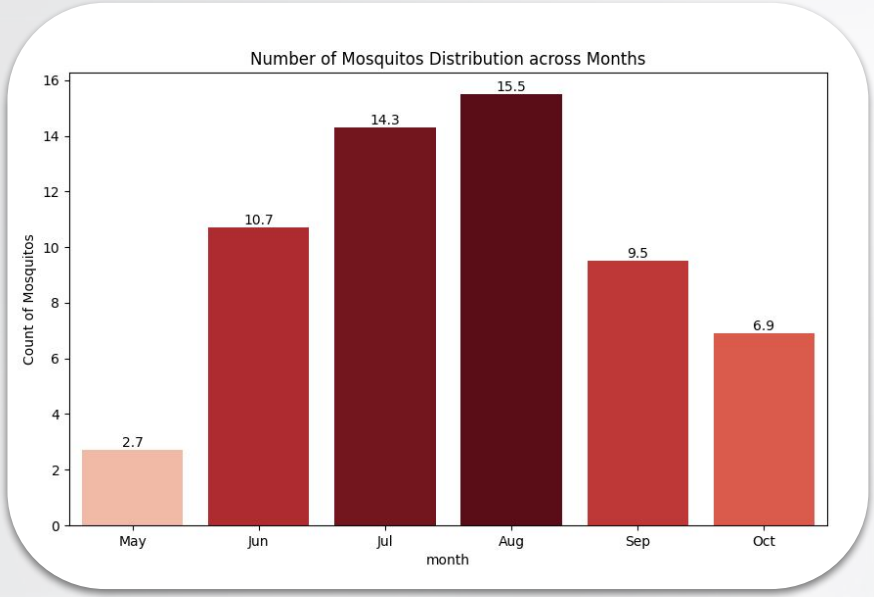


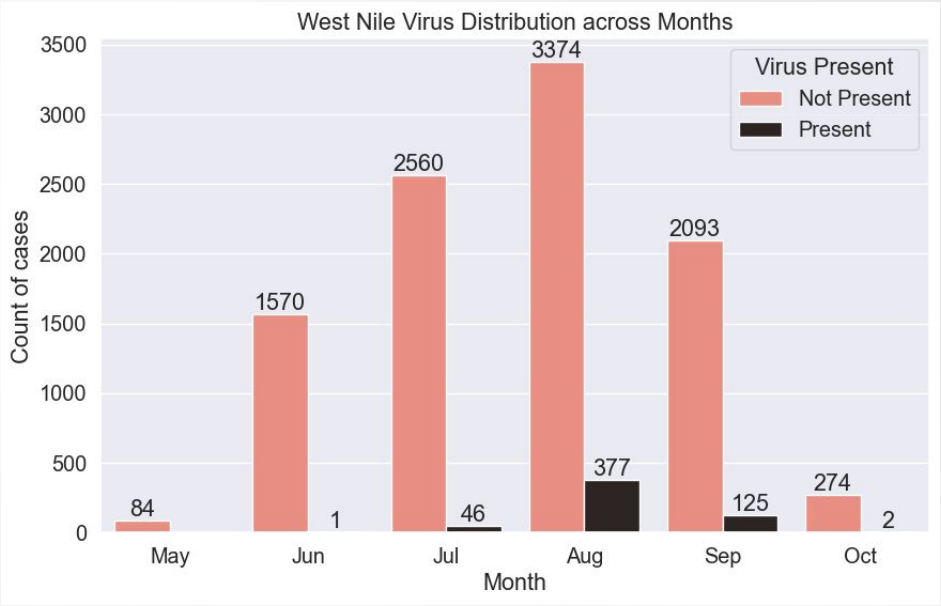
Feature Selection Feature Engineering

EDA

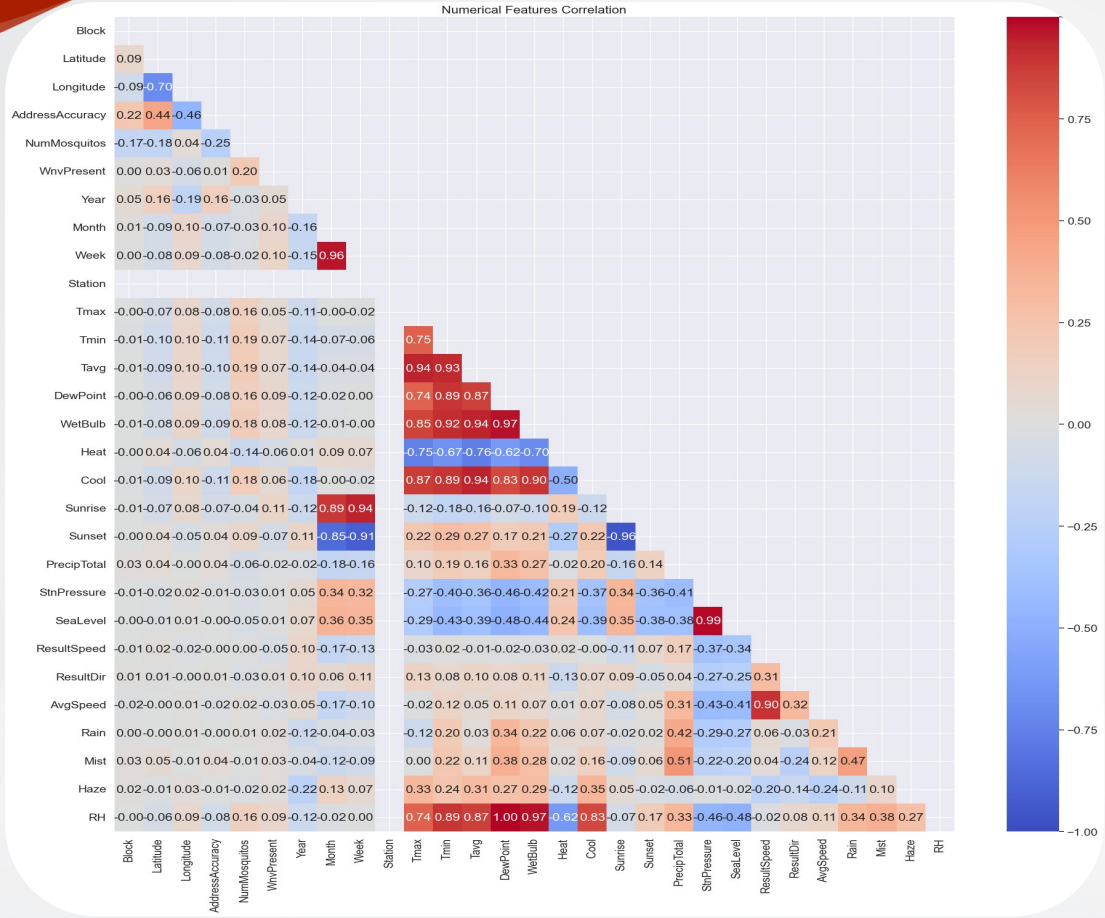
Exploratory Data Analysis

Exploratory Data Analysis: Time factors





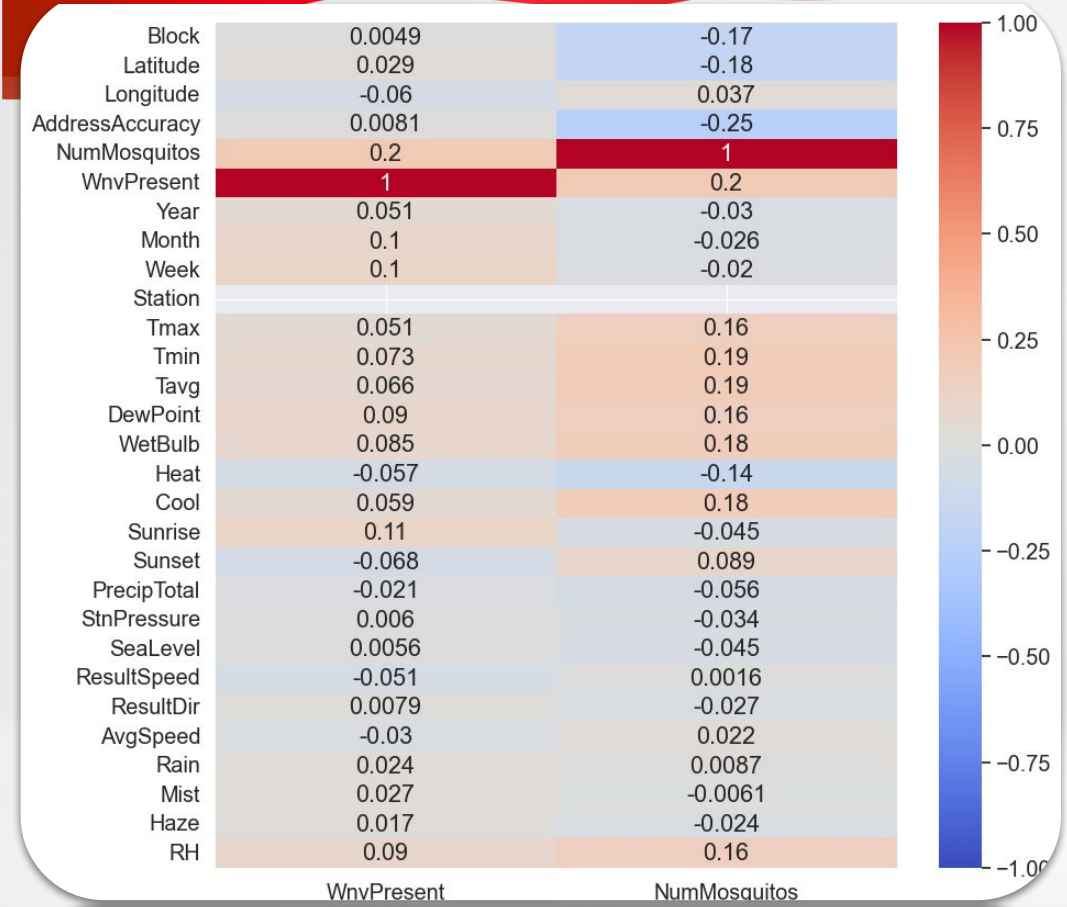
Exploratory Data Analysis:



Tmax
Tmin
Tavg
DewPoint
WetBulb
Cool
Sunrise
RH

***Red color indicates features that have multicollinearity**

Exploratory Data Analysis:



Virus Present

NumMosquitos
Sunrise

NumMosquitos

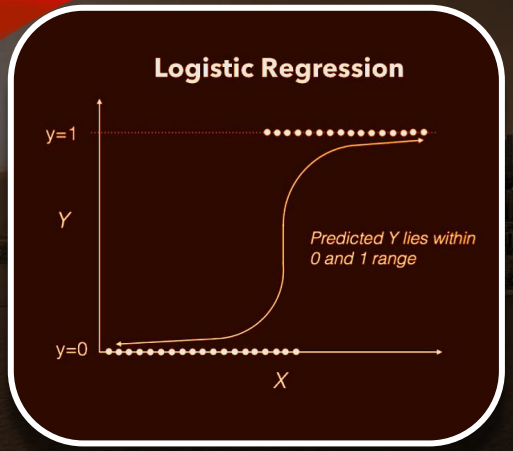
Climate Factors:
Tmax
Tmin
Tavg
Heat
DewPoint
WetBulb
Cool
RH

Location Factors:
Block
Latitude
Address Accuracy

Time Factors:
Month
Week

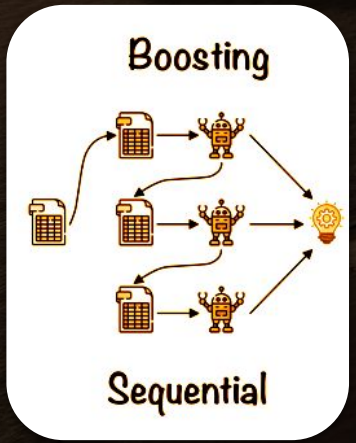
***Orange and Blue colors are significant features >= +/- 0.1**

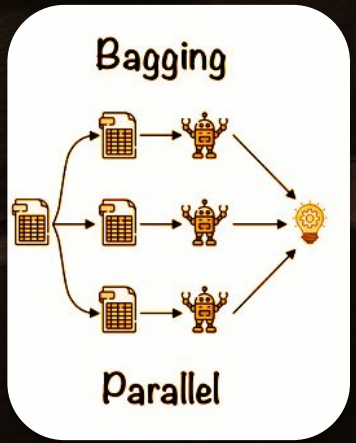
Methodology



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.



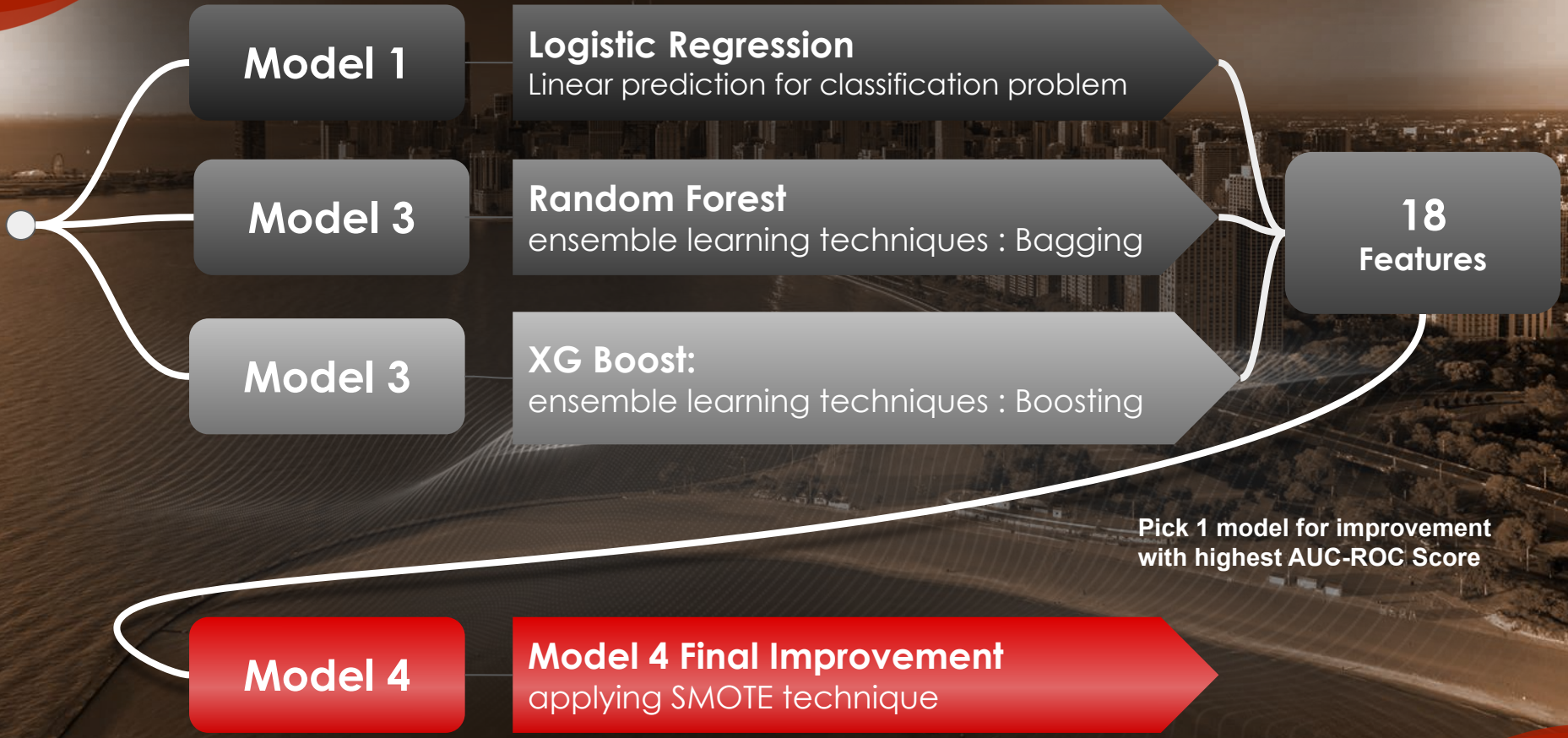


XGBoost (Extreme Gradient Boosting), Random Forest

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**

Predictive Modeling



Feature Selection

Date/Time

- Date **E**
- NEW Month
- NEW Year

Location Associated Features

- Address **E**
- Address No & Street
- Latitude
- Longitude
- Block
- Trap **D**
- NEW Street **D**
- NEW City **D**

Weather Associated Features

- Tmax
- Tmin
- Tavg.
- Heat
- Cool
- DewPoint
- Sunrise
- Sunset
- PrecipTotal
- StnPressure
- ResultDir
- AvgSpeed
- ResultSpeed
- Snowfall
- WetBulB
- Depth
- Sealevel
- CodeSum

Target (Y)

- WnvPresent

Miscellaneous Features

- Address Accuracy
- NumMosquito
- Depart
- Species **D**

Feature Engineering

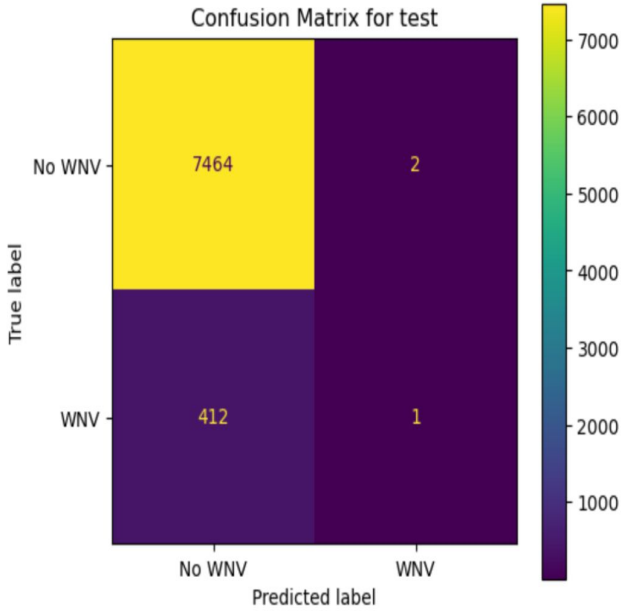
- **D** Dummify
- **C** Count (iterate through items)
- **E** Extract into new features

Model 1 : Logistic Regression

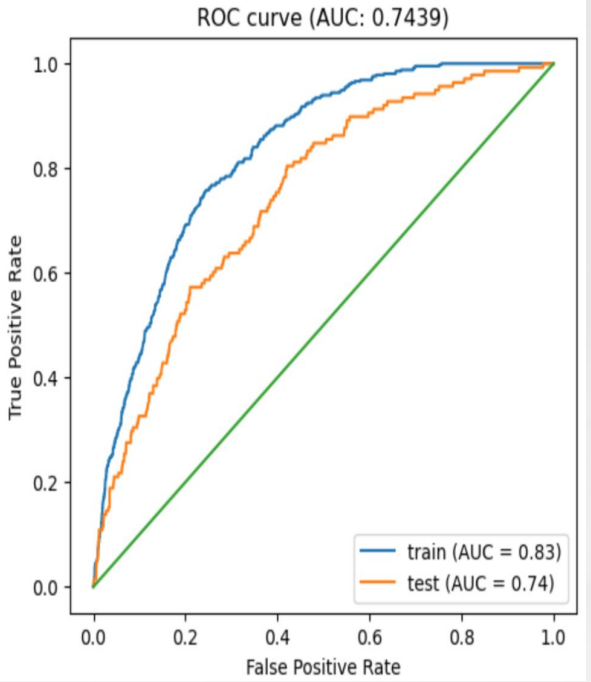
Metric	Score
Recall (WnvPresent:1)	0.01
Recall (WnvPresent:0)	0.99
AUC-ROC Score(Train)	0.83
AUC-ROC Score(Test)	0.74
KAGGLE SCORE	0.66

SMOTE

Confusion Matrix



AUC-ROC

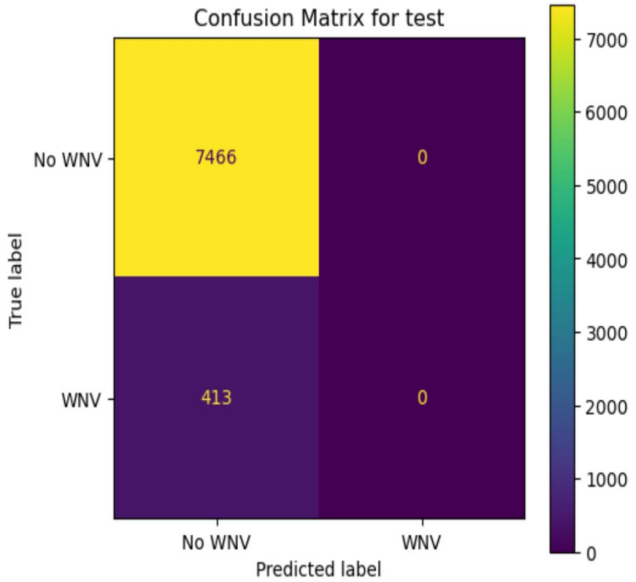


Model 2 : Random Forest Classifier

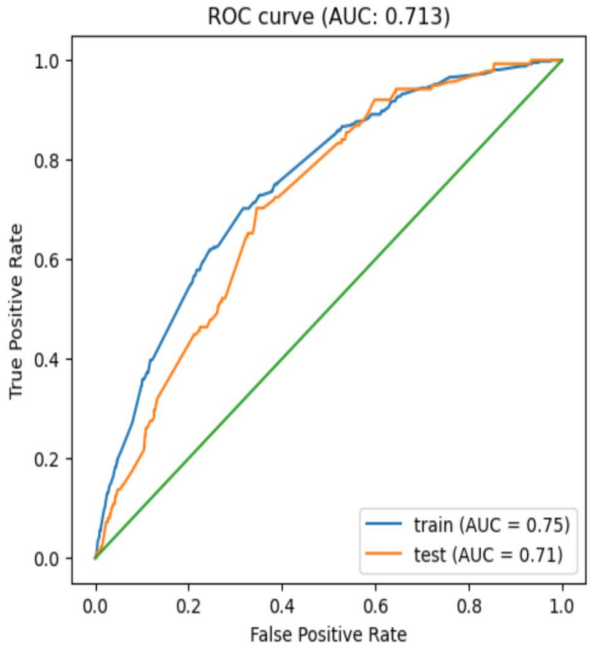
Metric	Score
Recall (WnvPresent:1)	0.00
Recall (WnvPresent:0)	1.00
AUC-ROC Score(Train)	0.75
AUC-ROC Score(Test)	0.71
KAGGLE SCORE	0.68

SMOTE

Confusion Matrix



AUC-ROC

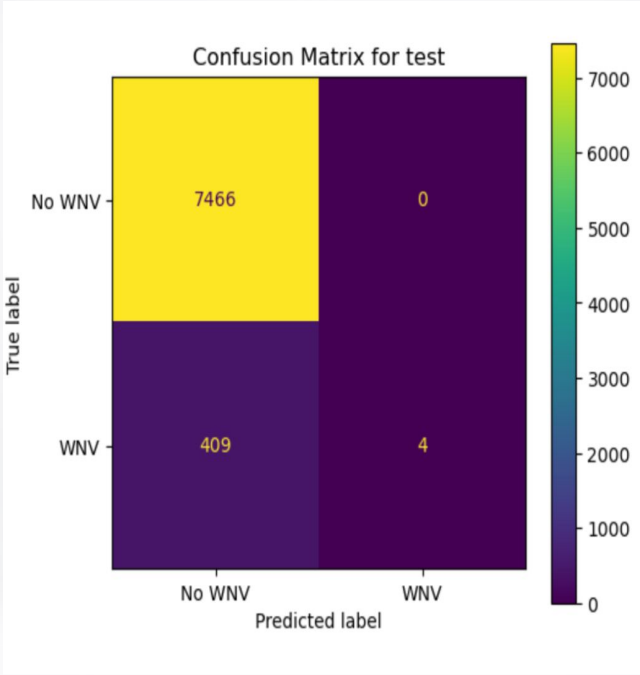


Model 3 : XGboost Classifier

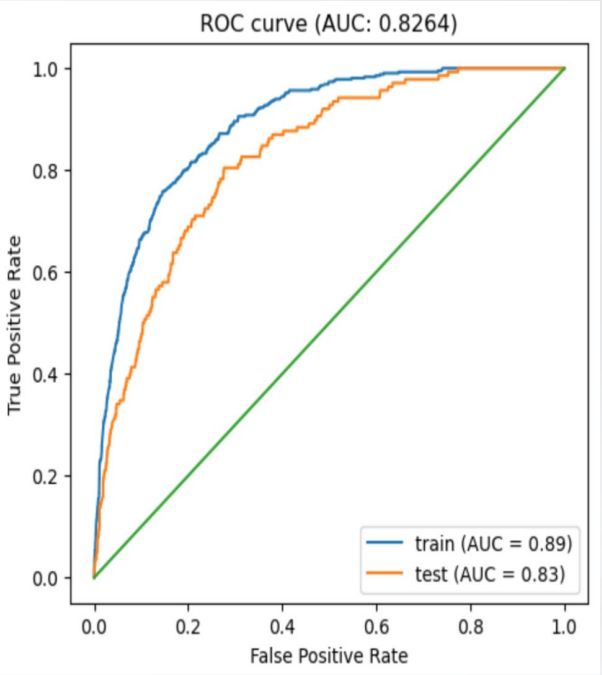
Metric	Score
Recall (WnvPresent:1)	0.00
Recall (WnvPresent:0)	1.00
AUC-ROC Score(Train)	0.89
AUC-ROC Score(Test)	0.83
KAGGLE SCORE	0.697

SMOTE

Confusion Matrix



AUC-ROC

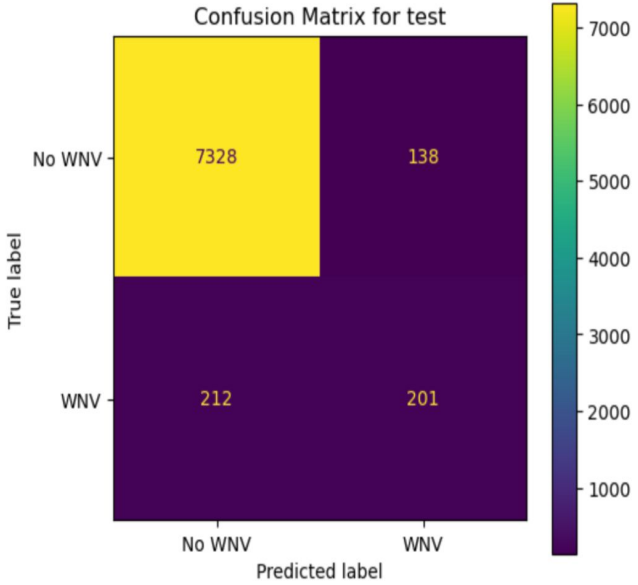


Model 4 : Model Improvement

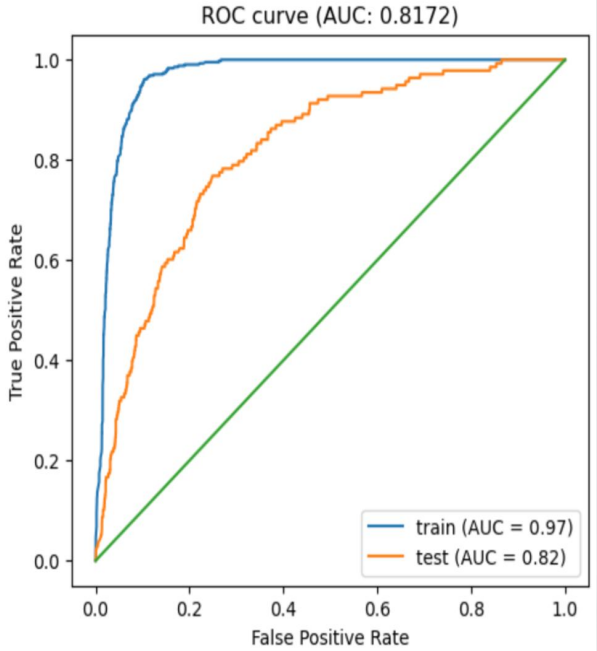
Metric	Score
Recall (WnvPresent:1)	0.167
Recall (WnvPresent:0)	0.972
AUC-ROC Score(Train)	0.97
AUC-ROC Score(Test)	0.82
KAGGLE SCORE	0.73

SMOTE

Confusion Matrix



AUC-ROC



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
AUC-ROC Score(Train)	0.83	0.75	0.89	0.97
AUC-ROC Score(Test)	0.74	0.71	0.83	0.82

Conclusion

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

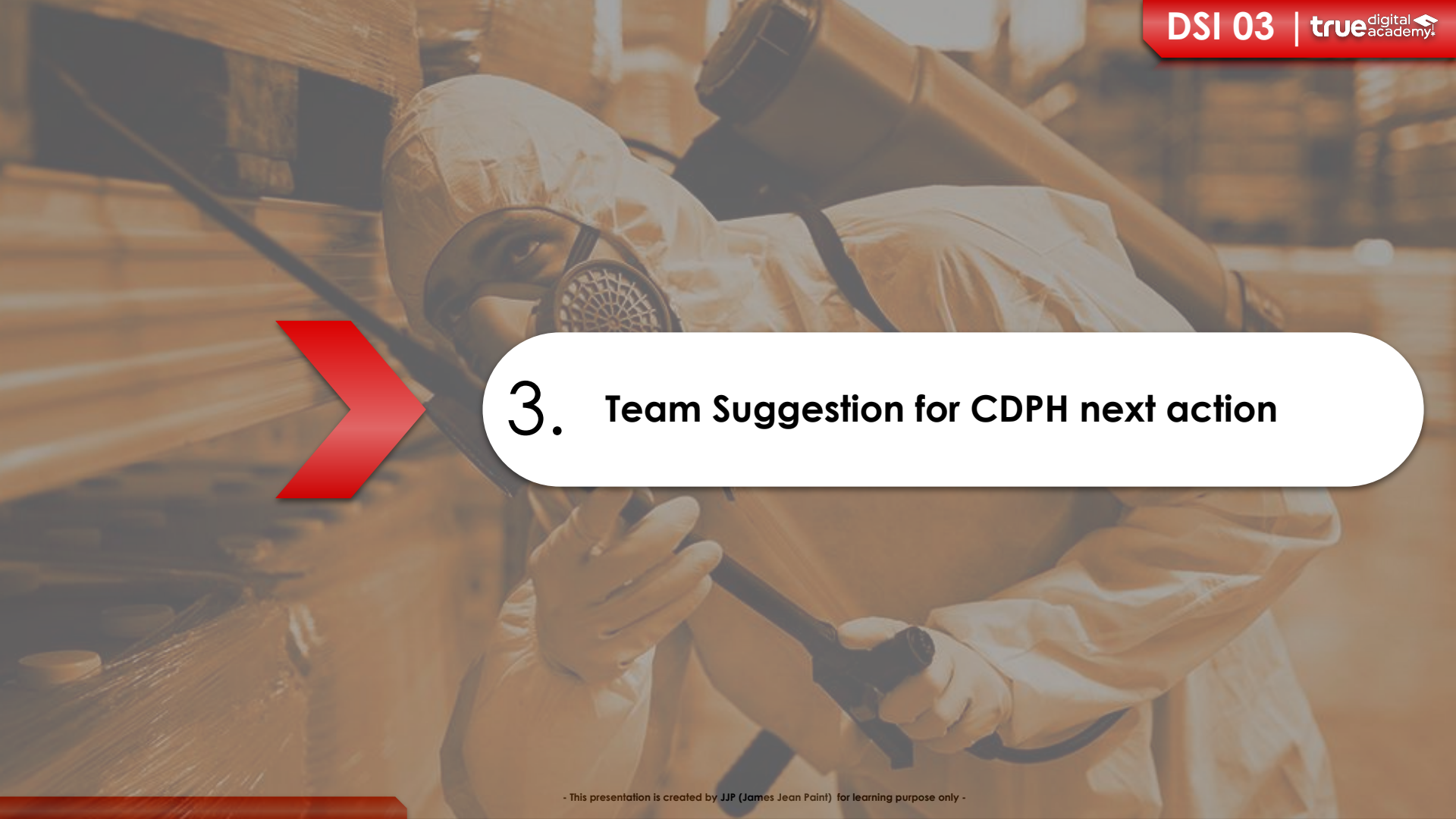
Key Findings

High Kaggle Score and AUC Score

- XGBoost with SMOTE achieves a notable Kaggle Score of 0.729 and an AUC of 0.817. 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 0.173, 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.

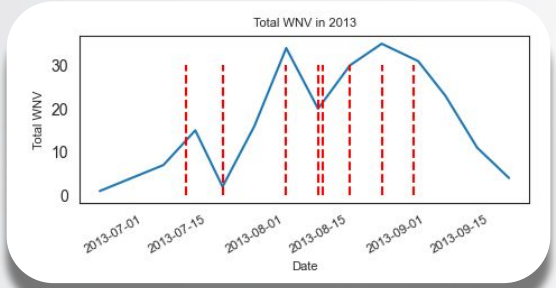


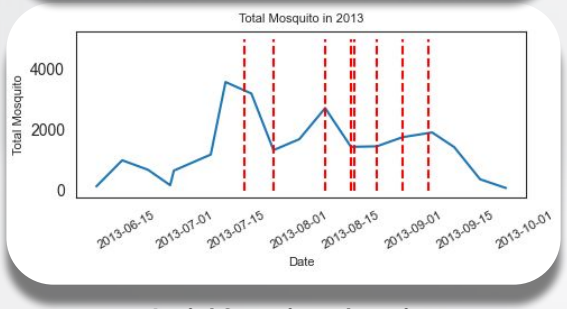
3. Team Suggestion for CDPH next action

Team Suggestion

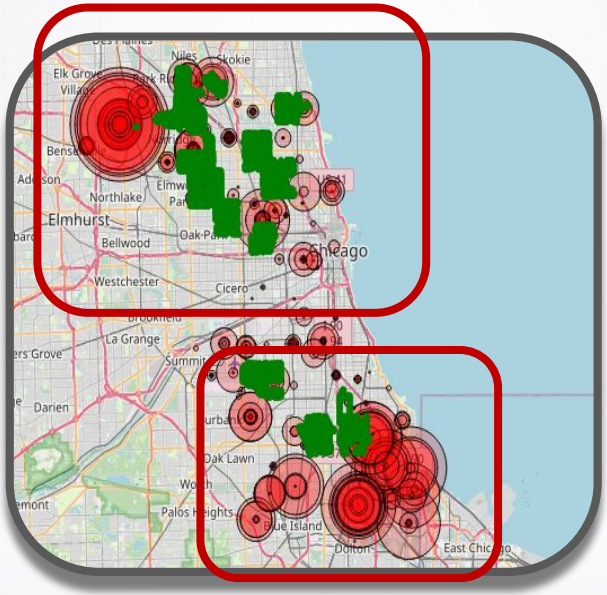
More Mosquitos = More chance for WNV presence

Key Action: Eliminating mosquito breeding sites & On-Ground spraying





Aerial Spraying alone is an ineffective method



CDPH may consider changing to On-Ground spraying instead

Recommended Location

Northern and Southern of Chicago

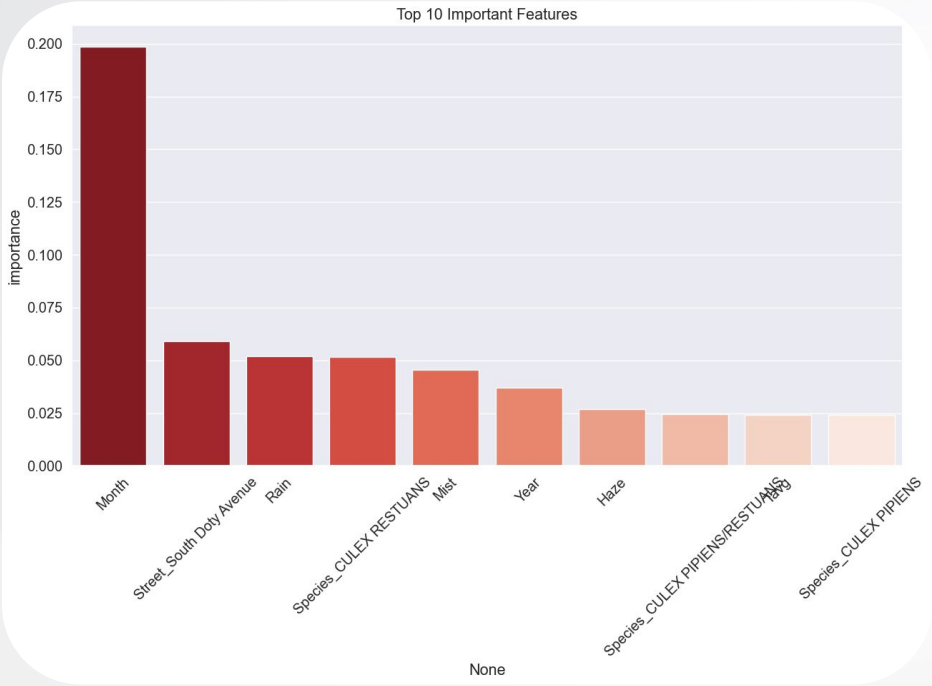
- Chicago O'Hare Int'l Airport
- South Doty Ave.
- North Oak Park Ave.
- South Stony Ave.
- Milwaukee Ave.
- And some more..

Vulnerable Target Residing

- Hospitals
- Elderly care center

Team Suggestion

Key Action: Extra Surveillance & Control



- Take extra Cautious Surveillance During **June to September** every year.
- **Warmer temperature** plays key role in mosquitoes activity & reproduction
- **Historical infected location** tends to reoccur its spread and should be in focus areas (i.e.)

Thank you