

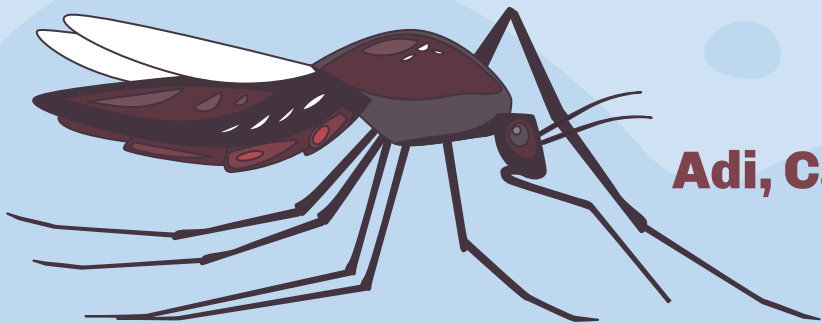


Project 4: West Nile Virus Prediction

DSI-28

10 June 2022

Adi, Calvin, Joel, Priscilla, Yong Lim



Agenda

1. Introduction and Problem Statement
2. Data Cleaning
3. Exploratory Data Analysis (EDA)
4. Feature Engineering
5. Modelling
6. Modelling Results
7. Cost-Benefit Analysis and Recommendations
8. Limitations and Future Steps
9. Conclusions





1

Introduction and Problem Statement

West Nile Virus (WNV)



First case in USA

Illinois, September 2001



Symptoms

1 in 5 will suffer from symptoms ranging from fever to meningitis



West Nile Virus (WNV)



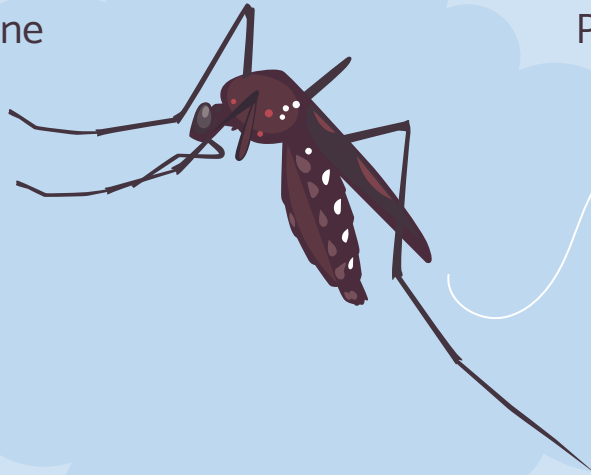
No Treatment

Currently no known medicine
or vaccine for WNV



Prevention

Prevention at personal and
state levels





Problem Statement

Contract Summary Sheet

Contract (PO) Number: 53283

Specification Number: 134997

Name of Contractor: VECTOR DISEASE CONTROL INTERNATIONAL LLC

City Department: DEPARTMENT OF HEALTH

Title of Contract: MOSQUITO ABATEMENT SERVICES

Term of Contract: Start Date: 3/14/2018

End Date: 3/13/2023

Dollar Amount of Contract (or maximum compensation if a Term Agreement) (DUR):
\$6,000,000.00

Brief Description of Work: MOSQUITO ABATEMENT SERVICES

Procurement Services Contract Area: PRO SERV CONSULTING \$250,000orABOVE

Please refer to the DPS website for Contact Information under "Doing Business With The City".

Vendor Number: 56454025

Submission Date: January 22, 2018

To prevent a WNV outbreak in Chicago, the Chicago Department of Public Health (CDPH) has tasked its data science team to develop a *predictive model to detect areas highly likely to have WNV.*

In addition, CDPH has requested for our expertise in advising the *areas and timings to spray pesticide* as part of the terms in the contract. The advice also includes data-driven *benefits* of spraying, and annual *costs* estimates to be used in their price negotiations for the new spraying contract.



The background is a light blue gradient. A large, stylized, light blue cloud is on the right side. On the left, there is a large, detailed illustration of a mosquito with a dark brown body and legs, and a patterned abdomen. Three smaller, simpler fly-like insects are scattered around: one in the upper left, one in the upper right, and one in the lower center. A thin, white, swirling line is also present in the upper left area.

2

Data Cleaning

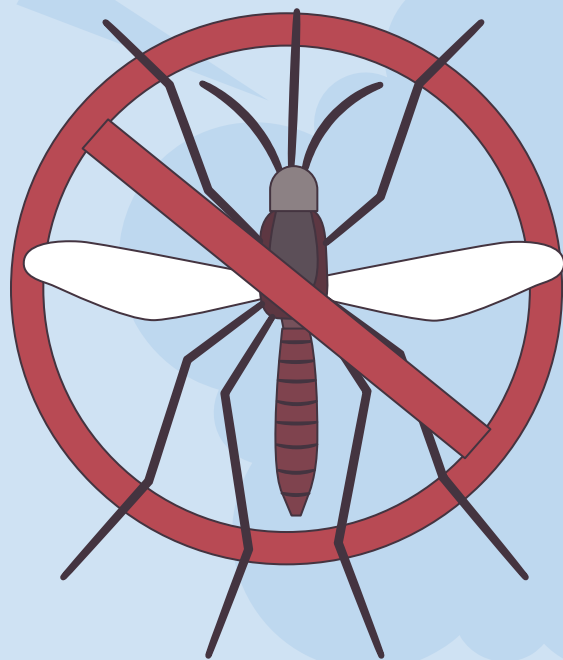
Data Provided



Data	Description	Timeframe
Train	Data set to train predictive model	29 May 2007 to 26 Sep 2013
Test	Data set to produce prediction results	11 June 2008 to 2 Oct 2014
Spray	Time, date and location of previous sprays	29 Aug 2011 to 5 Sep 2013
Weather	Meteorological information of Chicago	1 May 2007 to 31 Oct 2014



Data Cleaning - Train Dataset



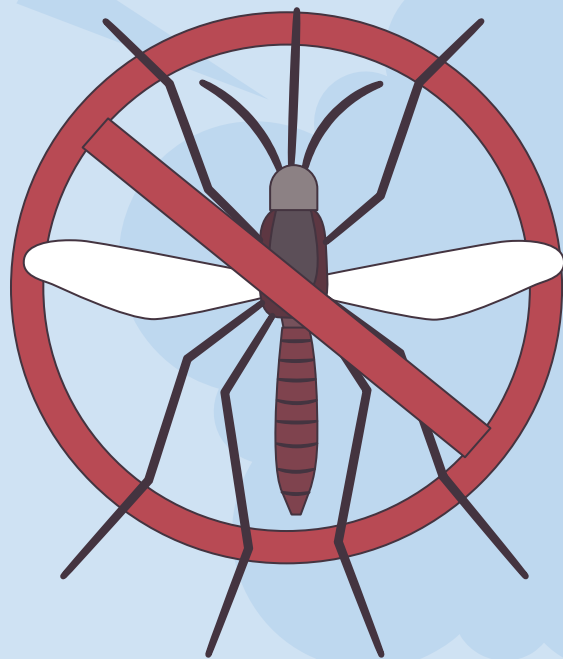
**Removing
duplicate
records**

Based on Date, Species
and Trap





Data Cleaning - Weather Dataset



Replacing T

Trace replaced with
value 0 for Total
Precipitation

Imputing missing PrecipTotal

Using previous
observation

Imputing missing WetBulb

Using the $\frac{1}{3}$ Rule

Imputing missing Tavg

Using Tmin and Tmax

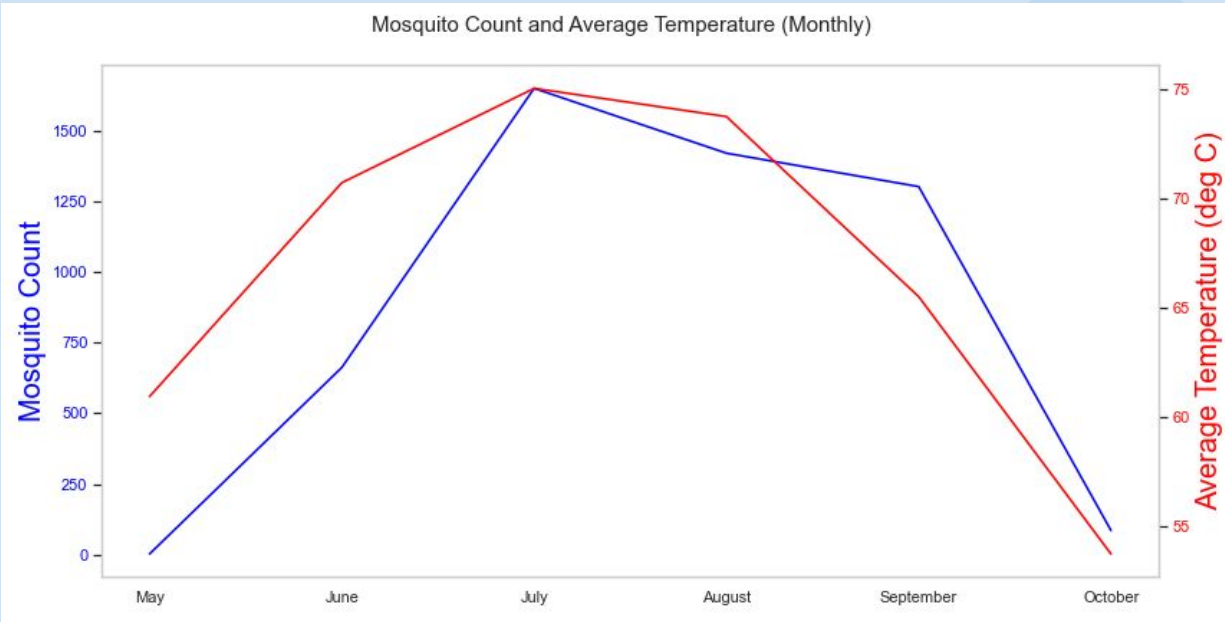




3

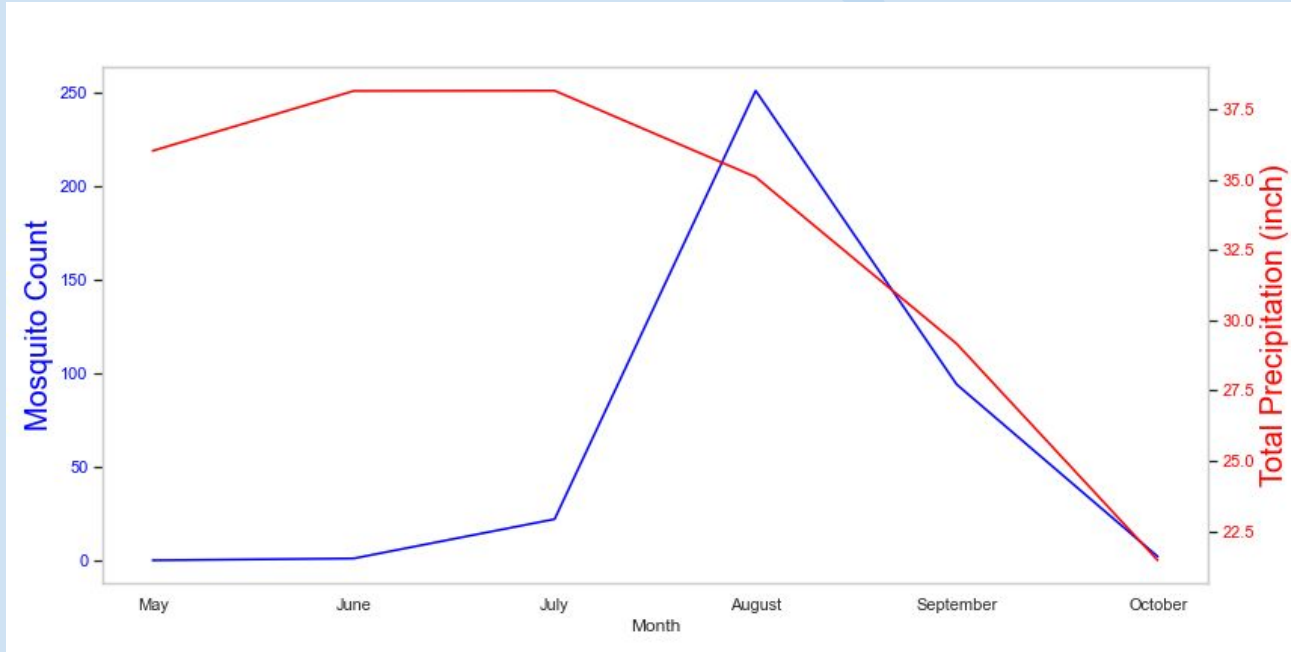
Exploratory Data Analysis (EDA)

Mosquito numbers trend closely with temperature over the year



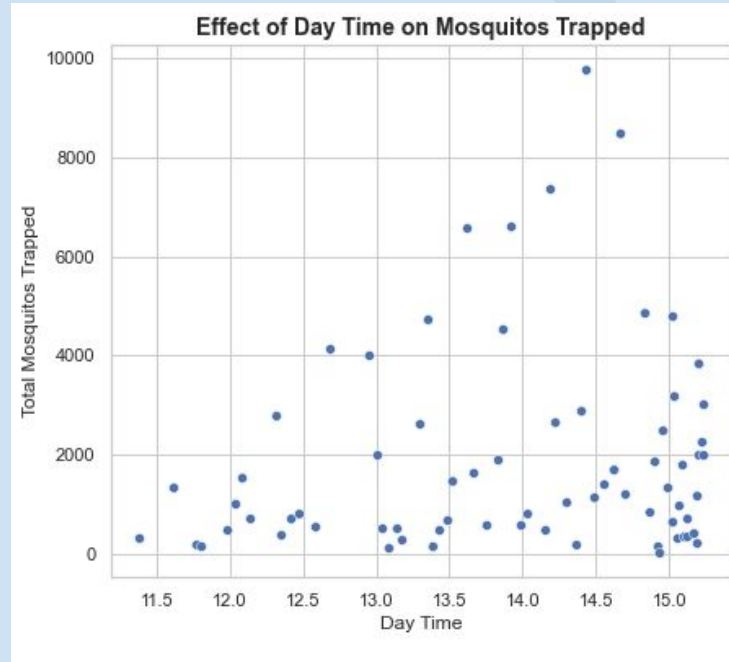
- Mosquito count increases with temperature, peaking in July-August
- Mosquito count decreases in later months as temperature drops
- Slight lag between mosquito count and temperature

Mosquito count tend to peak after rainfall



Peak in rainfall followed shortly by peak in mosquito numbers

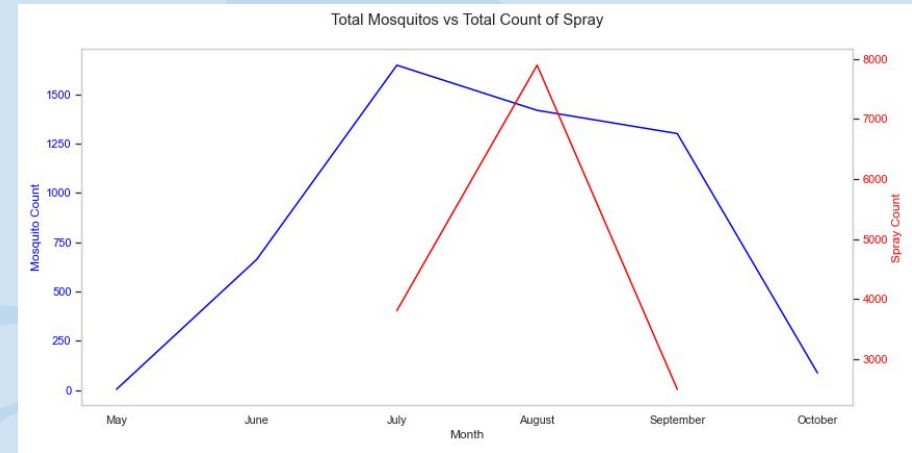
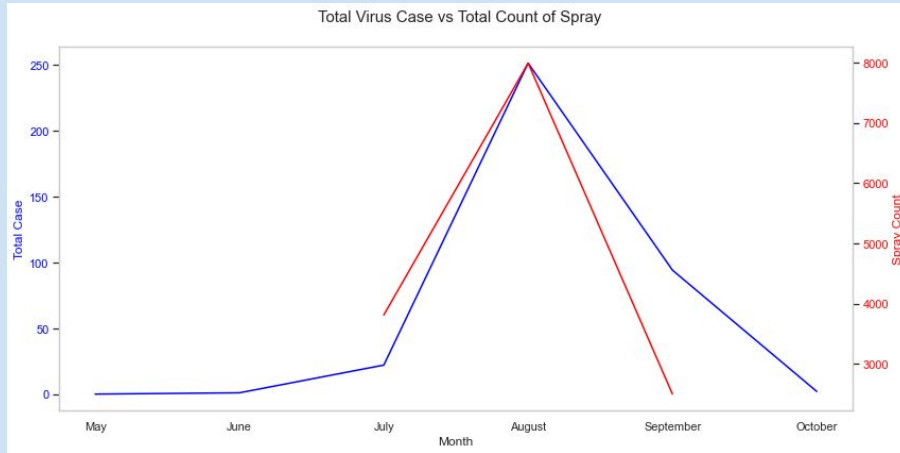
Mosquito count increase with daytime length



More mosquitoes are detected as day length increases



Spraying in response to WNV cases, as opposed to mosquito numbers



- Spraying counts increased almost instantaneously to emergence of WNV cases in July. (left)
- Increase in number of mosquitoes started much earlier in May. (right)
- Spraying may have started only in response to receiving reports of WNV cases

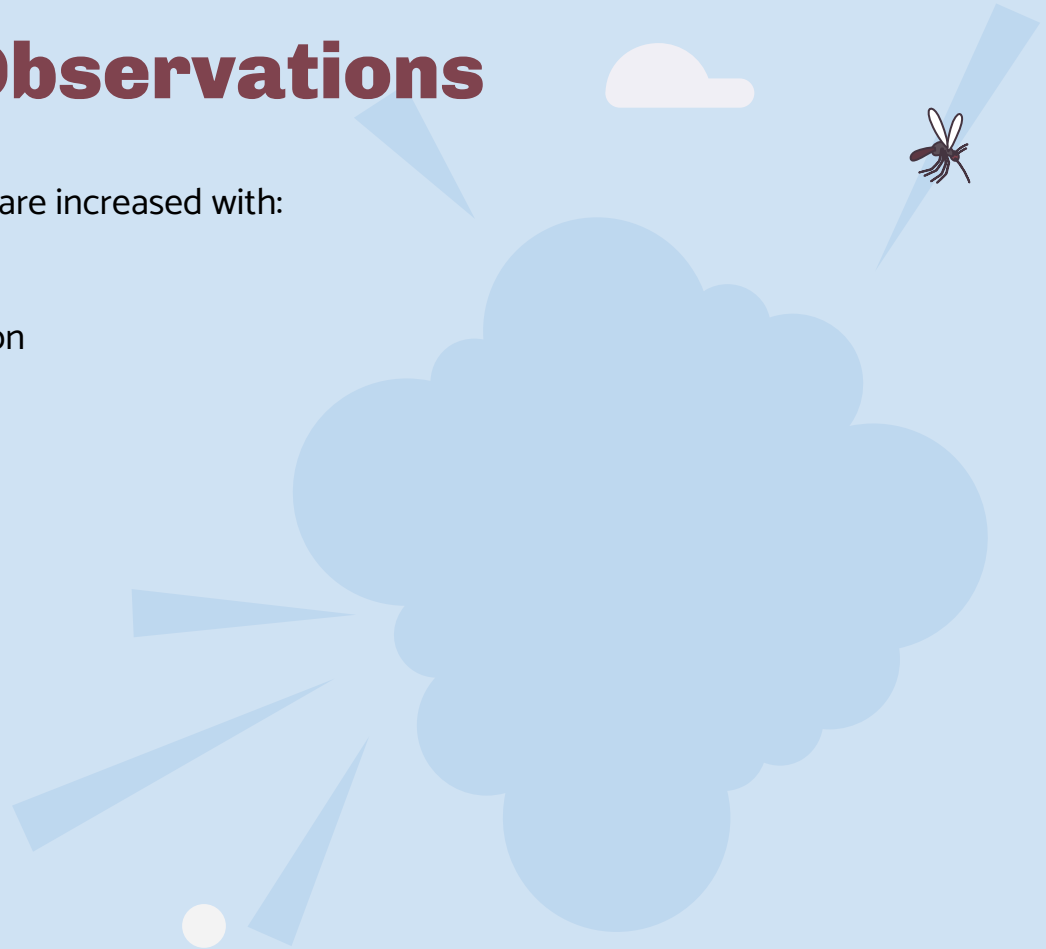
Summary of Observations

Mosquito count and Virus count are increased with:

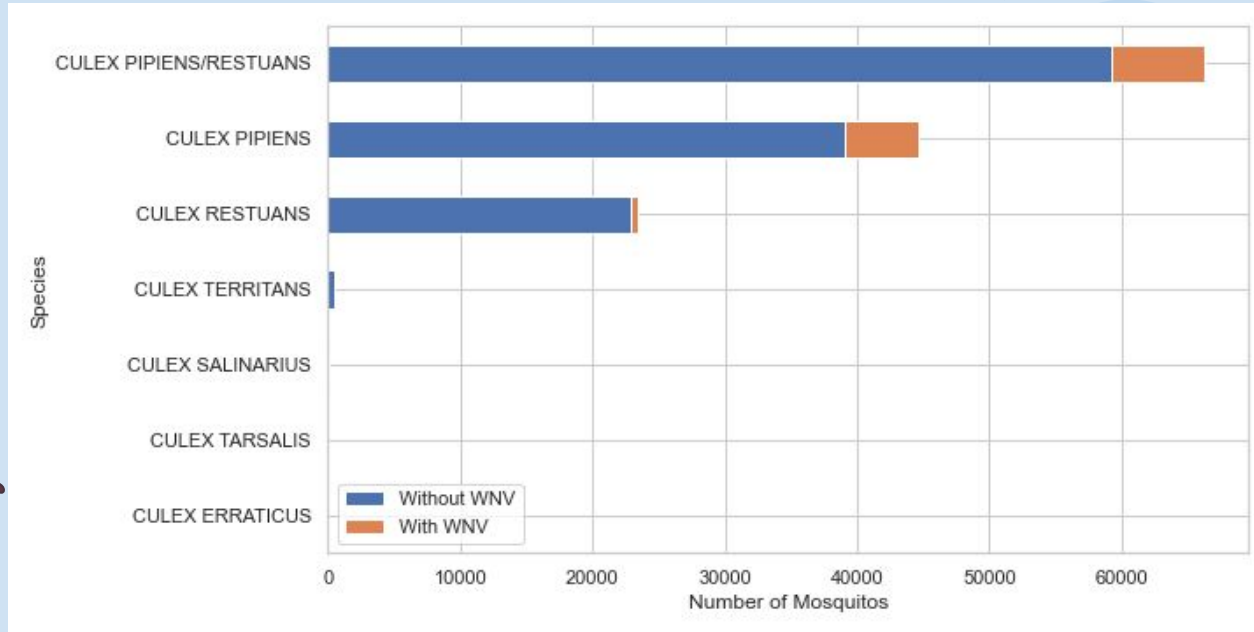
- Higher temperatures
- Higher rainfall/ precipitation
- Longer daytime

And are effectively lowered by:

- Spraying insecticide



Culex Pipiens and Restuans species carry WNV



West Nile Virus found exclusively in:

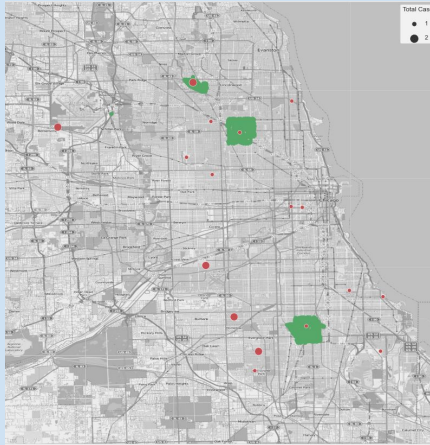
- Culex Pipiens
- Culex Restuans



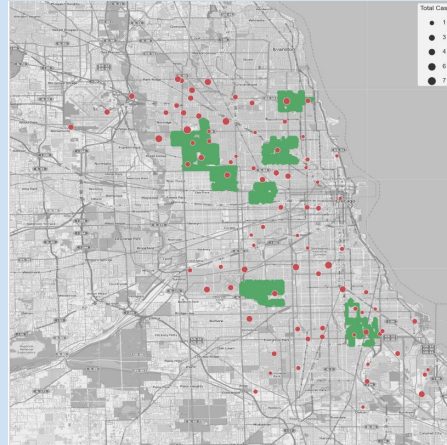
Disparity in spraying and actual WNV outbreak locations



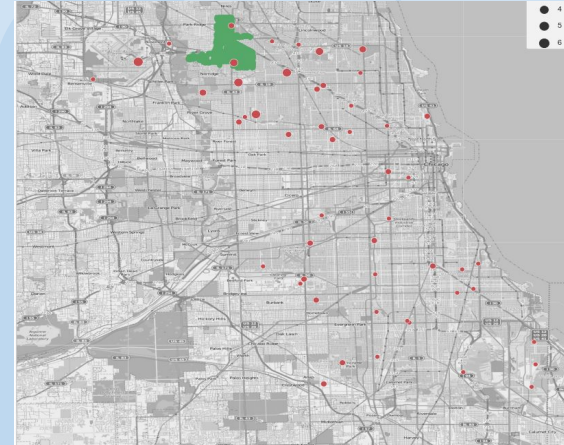
July



August



September



- Areas sprayed with insecticide (green) and locations with WNV cases (red)
- Sprayed area tend to cover much larger area than the neighbourhood - wastage
- Problem areas not receiving spraying



4

Feature Engineering

Factors Affecting Mosquito Breeding



Environment

Temperature, Relative Humidity, Precipitation / Rain, Wind

Time

Week, Night / Day, 7 days lag

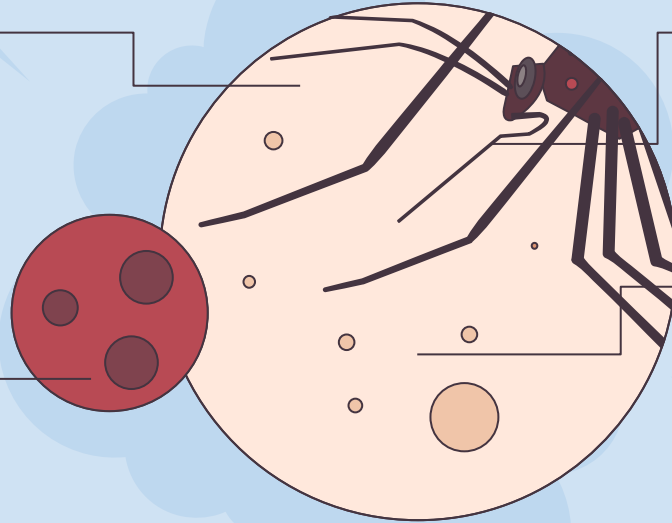


Location

Latitude, Longitude , WNV risk

Species

Culex Pipens, Culex Restuan



Source 1: Predicting Culex pipiens/restuans population dynamics by interval lagged weather data

Source 2: When are mosquitos most active

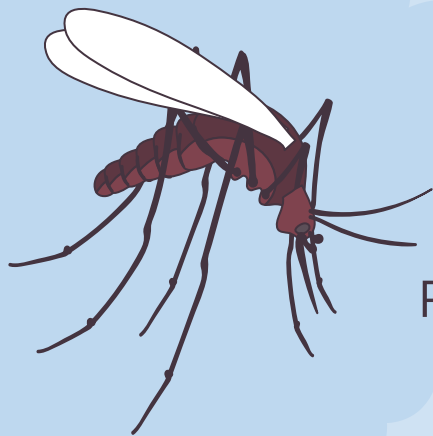
Environment Added Features



Relative Humidity

$$RH = 100 \times \left[\frac{e^{\frac{17.625 \times D_p}{243.04 + D_p}}}{e^{\frac{17.625 \times T}{243.04 + T}}} \right]$$

Dp - Dewpoint Temp
T - Ave Temp



Environment Dataset



Average Temperature

Dew Point

Precipitation

Wind Speed / Direction

Station Pressure

Sea Level

Rain / Thunderstorm / Mist



Location Added Features



WNV Risk

Low – 0 to 2 WNV cases
Medium – 3 to 5 WNV cases
High – Above 5 WNV cases



Location Dataset



Latitude
Longitude



Time Added Features



Night Time
Day Time
Week

7 Days Lag for temp,
dewpoint & precipitation

Species



Culex Pipens
Culex Restuan





5

Modelling

Logistic Regression

Random Forest

AdaBoost

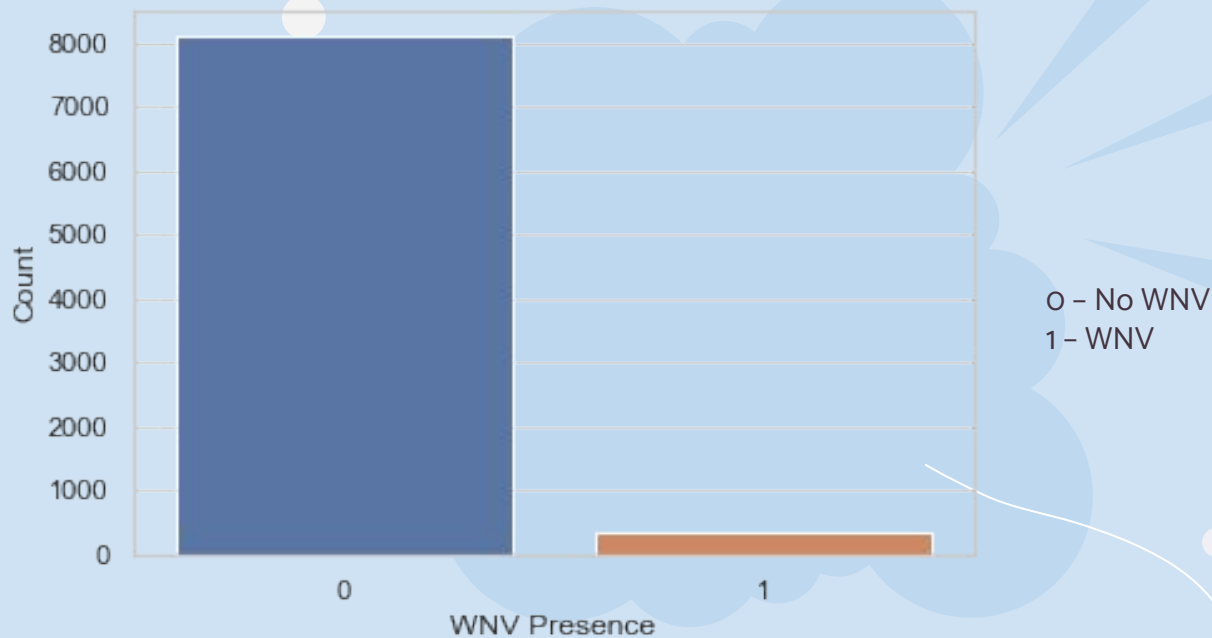
Gradient Boost

XGBoost

Support Vector Machine

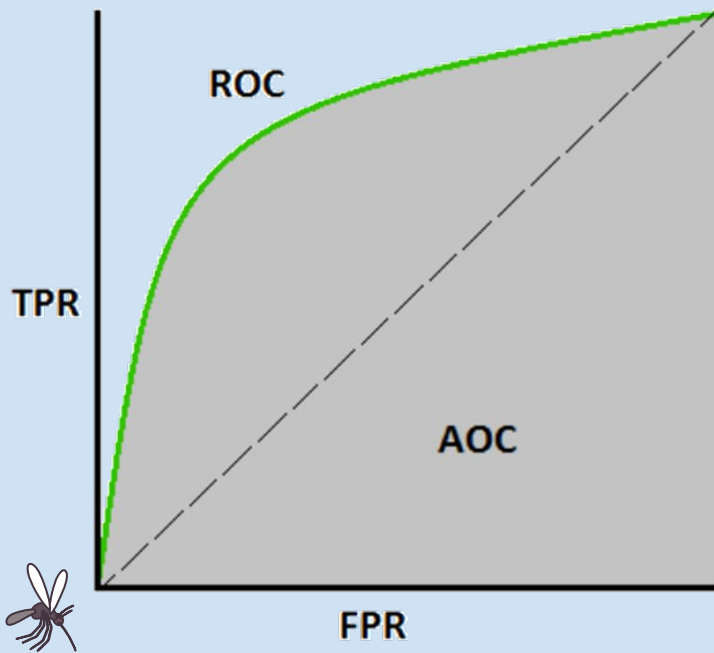
K-Nearest Neighbours

Imbalanced Target Variable



Generally lead to poor performance on the minority class

Model Evaluation Metric - AUC ROC

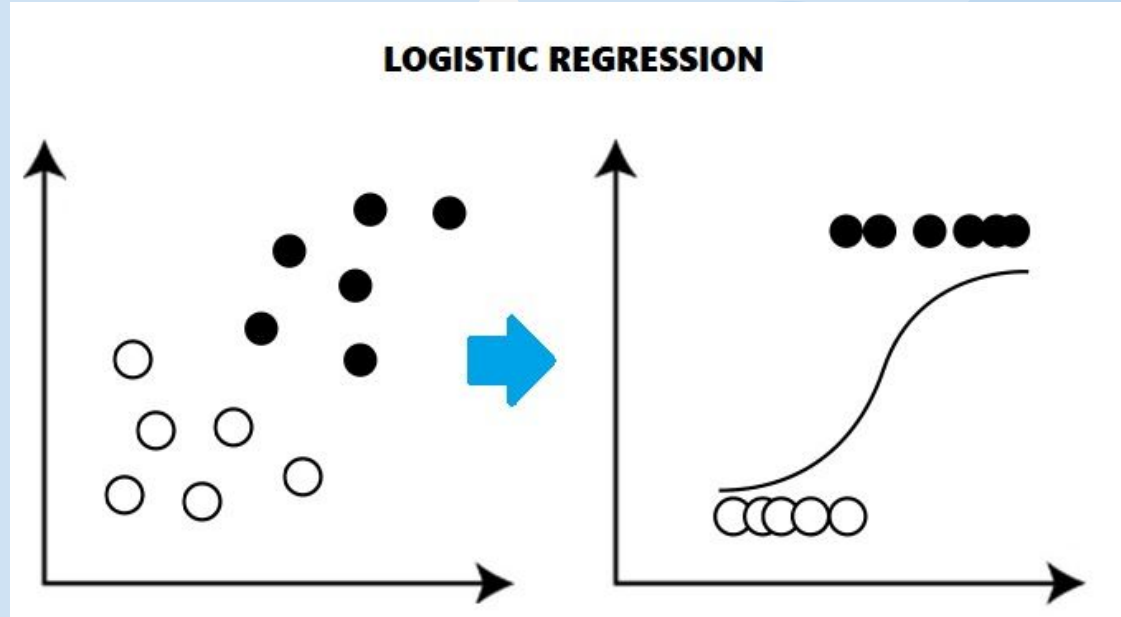


- Evaluates how good the model distinguish the classes
- Higher AUC -> Better in predicting 0 class and 1 class
- Good Model -> Accurately predict the presence of the virus

Source of illustration:

<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

Logistic Regression



Used as baseline
model

**ROC AUC
(Train)**

0.7771

**ROC AUC
(Validation)**

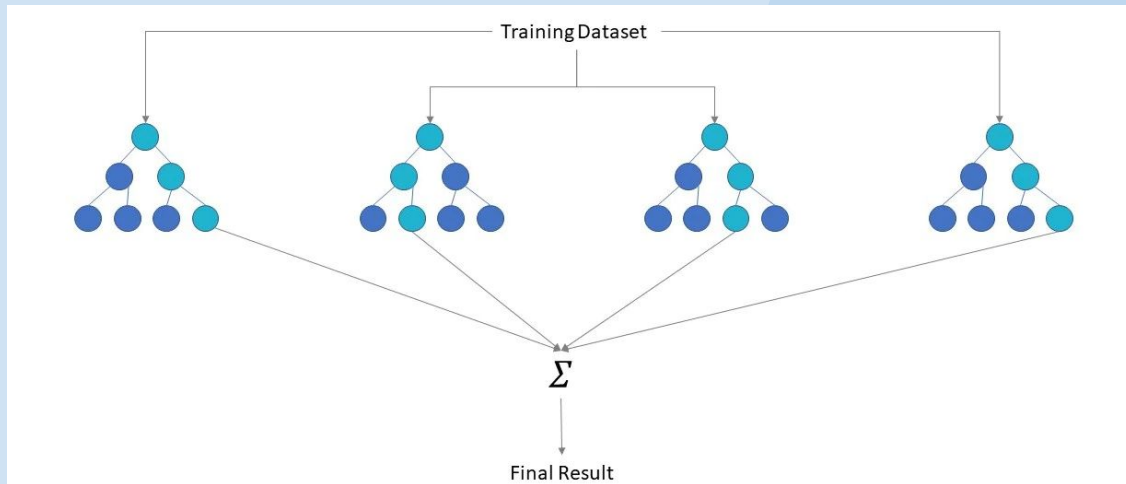
0.8554



Source of illustration:

<https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-logistic-regression-using-python/>

Random Forest



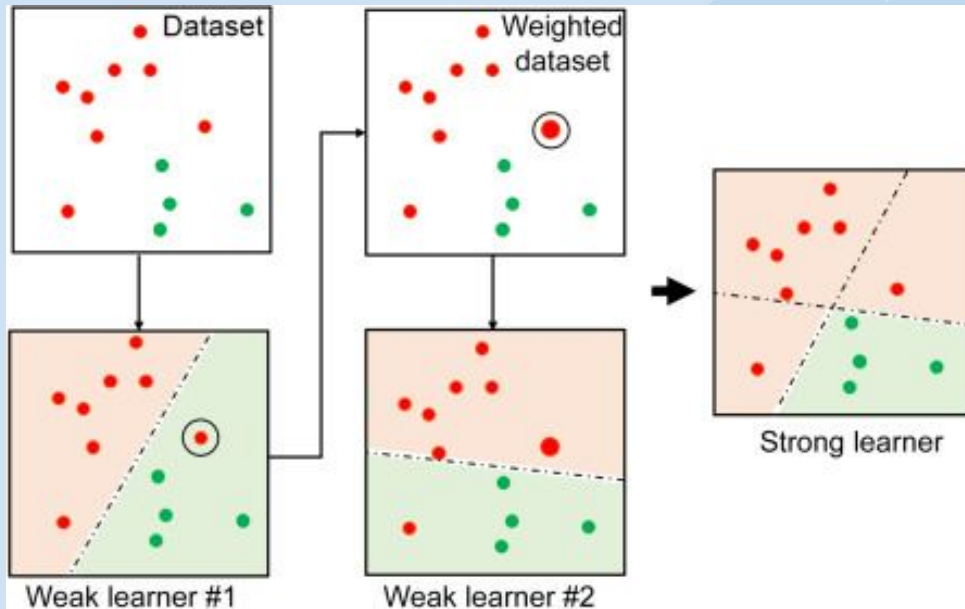
**ROC AUC
(Train)**

0.9169

**ROC AUC
(Validation)**

0.8790

Adaptive Boosting (AdaBoost)



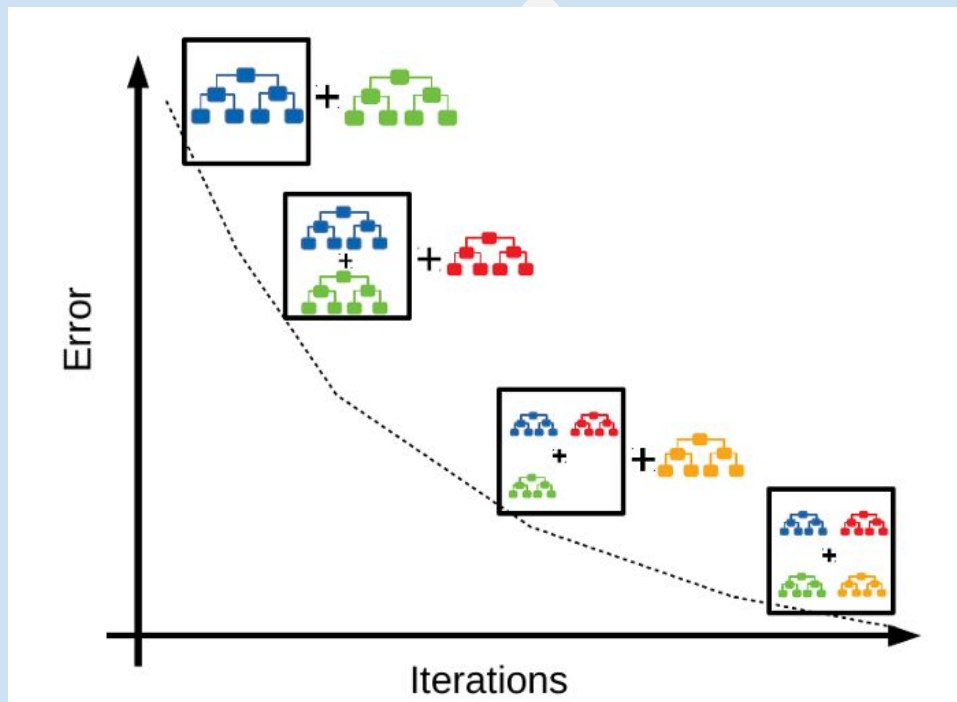
**ROC AUC
(Train)**

0.8752

**ROC AUC
(Validation)**

0.8809

Gradient Boosting



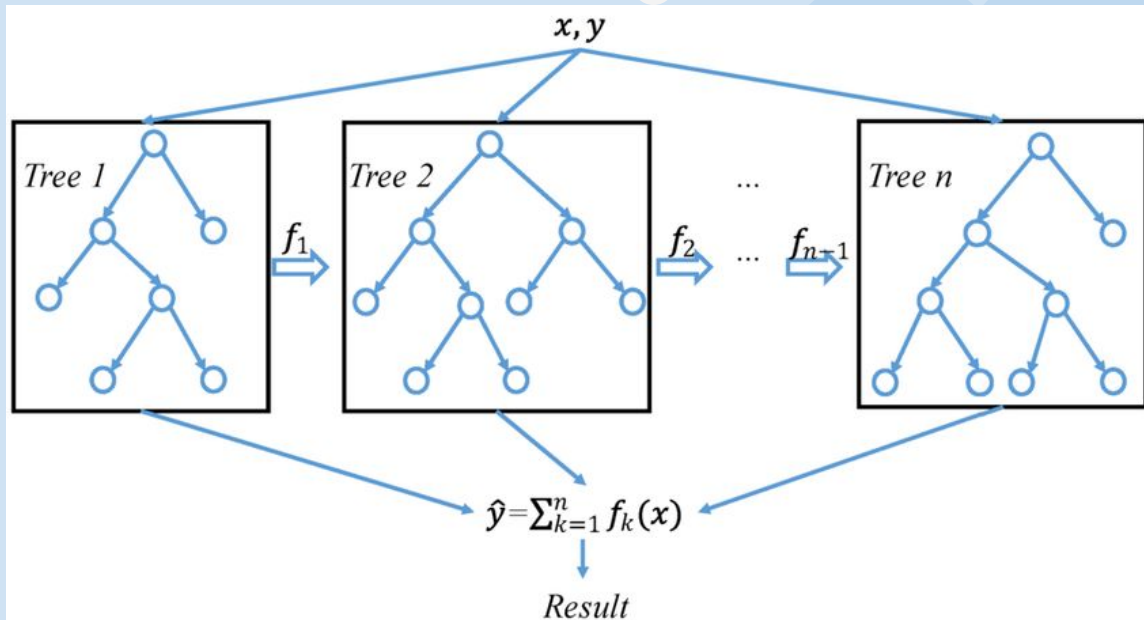
**ROC AUC
(Train)**

0.9065

**ROC AUC
(Validation)**

0.8808

eXtreme Gradient Boosting (XGBoost)



**ROC AUC
(Train)**

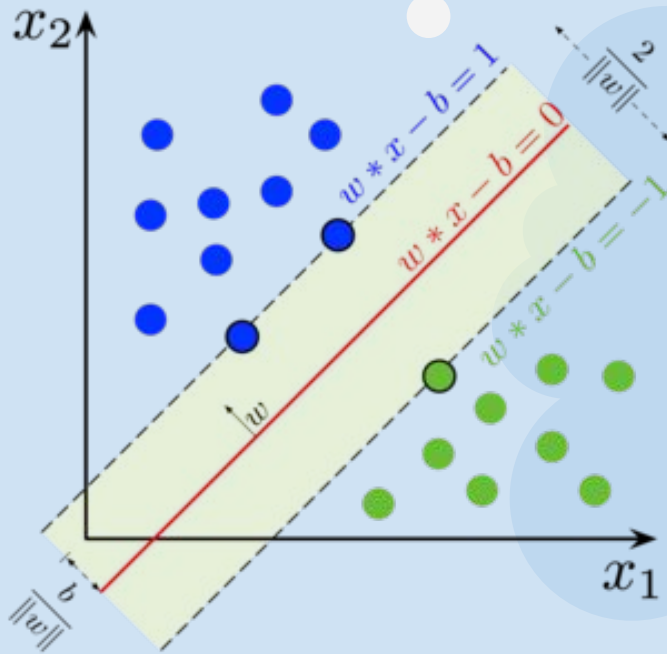
0.8762

**ROC AUC
(Validation)**

0.8799



Support Vector Machine (SVM)



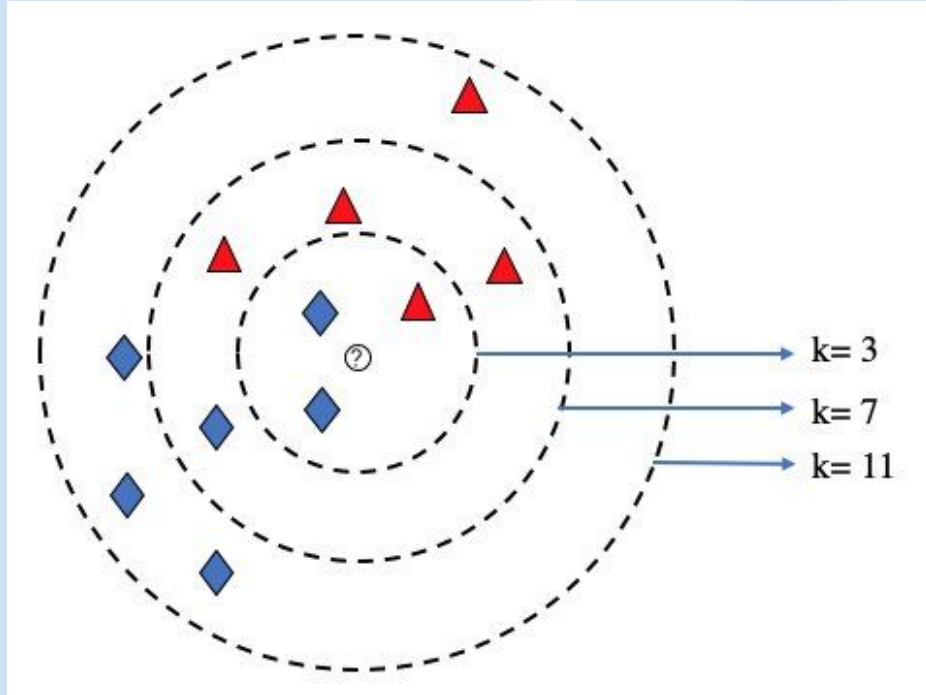
**ROC AUC
(Train)**

0.8930

**ROC AUC
(Validation)**

0.8772

k-Nearest Neighbours (KNN)



**ROC AUC
(Train)**

0.9361

**ROC AUC
(Validation)**

0.8592



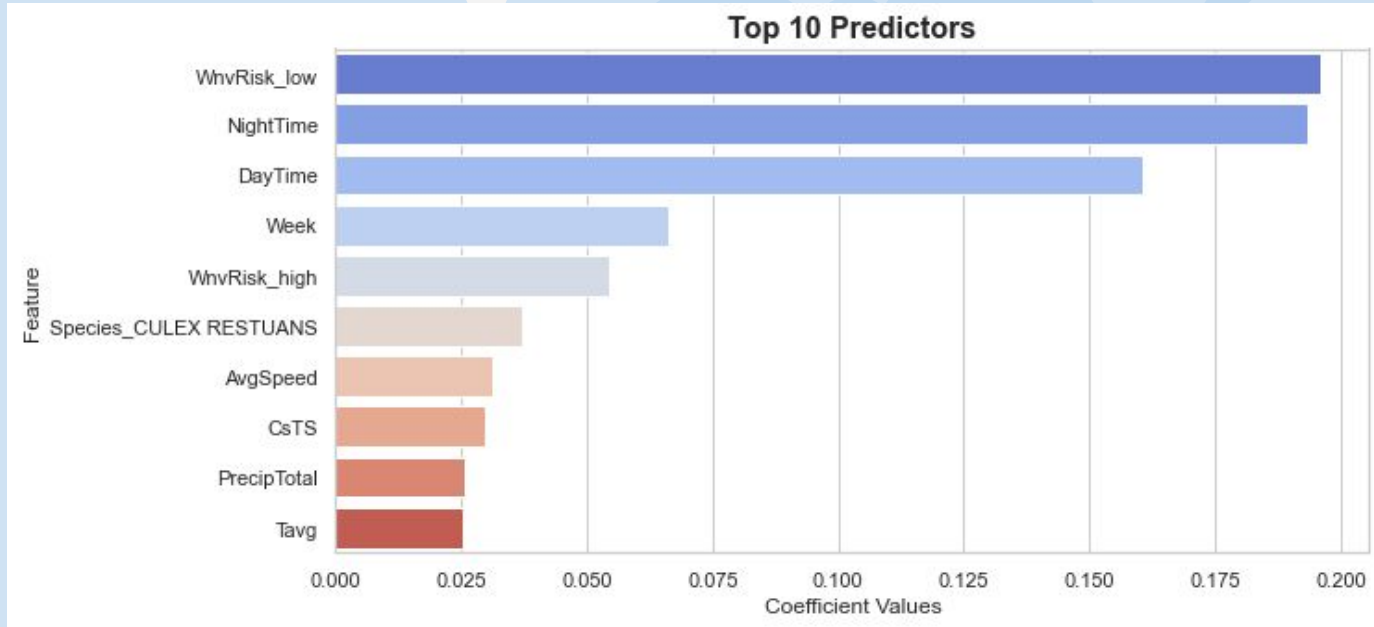
6

Modelling Results

Models Performance Summary

Model	ROC AUC	
	Train (A)	Validation (B)
Logistic Regression	0.8385	0.8574
Random Forest	0.9177	0.8781
AdaBoost	0.8771	0.8781
Gradient Boost	0.9072	0.8737
Extreme Gradient Boost	0.8762	0.8805
Support Vector Machine	0.8941	0.8767
k-Nearest Neighbours	0.9353	0.8550

Top 10 AdaBoost Predictors





7

Cost-Benefit Analysis and Recommendations

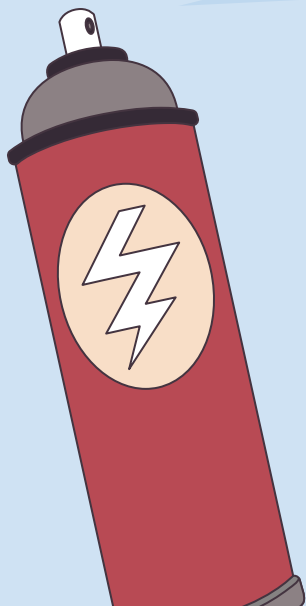
Cost of Mosquito Abatement Program 2023



~USD 520,698*

which includes:

- Weekly Environmental Surveillance (~ 147 gravid traps)
- Conduct Larviciding (~ 190 acres)
- Conduct Adulticiding (~ 100 miles)



* based on Contract (PO) Number 17068
**"SLE Vector Mosquito Abatement
Program"** awarded to Vector Disease
Control International (VDCI)

Average Total Economic Cost for 2023: ~USD 2,800,100*

Average Cost Per Person: ~USD 176,071*



Assumption based on:

- Average of people infected of 17 throughout 2012-2021 ⁺
- Average death rate of 2 throughout 2014-2016 ⁺

* Forecasted from data source: "**Initial and Long-Term Costs of Patients Hospitalized with West Nile Virus Disease**" ([*Source*](#)) paper by [*Centers for Disease Control and Prevention \(CDC\)*](#) dated 05 Mar 2014

⁺ **Source:** [*West Nile Virus Surveillance Reports*](#)




Benefits > Costs on Average by 4 times → Continue Mosquito Abatement Program



	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Human Cases	22	1	6	16	49	6	42	2	11	10
Human Cases (Fatality)	-	-	-	3	2	1	-	-	-	-

Source: [West Nile Virus Surveillance Reports](#)

Worst case scenario based on 2016 records, with **49** cases reported and **2** casualties,

 the total lost instead would be: **~USD 8,434,382**, a whopping **16 times** from the 2023 abatement cost

Recommendations

Increase Larviciding Initiation

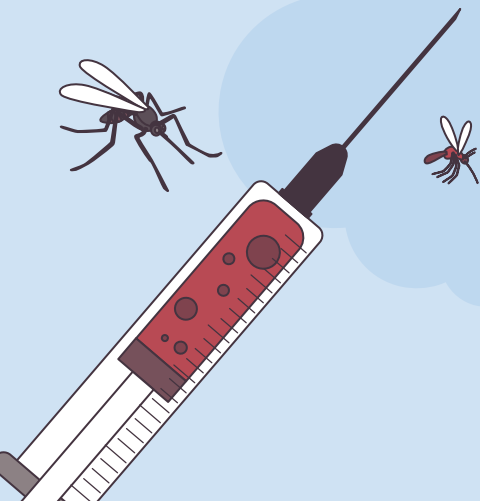
At the location which has a high risk of West Nile Virus emergence (WnV case > 5), additional 800 acres => USD 77,920

Lower the Threshold to Activate Adulticiding

For the month of June and July, to suppress the population of mosquitoes , additional 100 miles => USD 12,060

Conduct Awareness Roadshow

Before the breeding season start. Helps to reduce potential breeding location, estimated cost ~USD 15,000

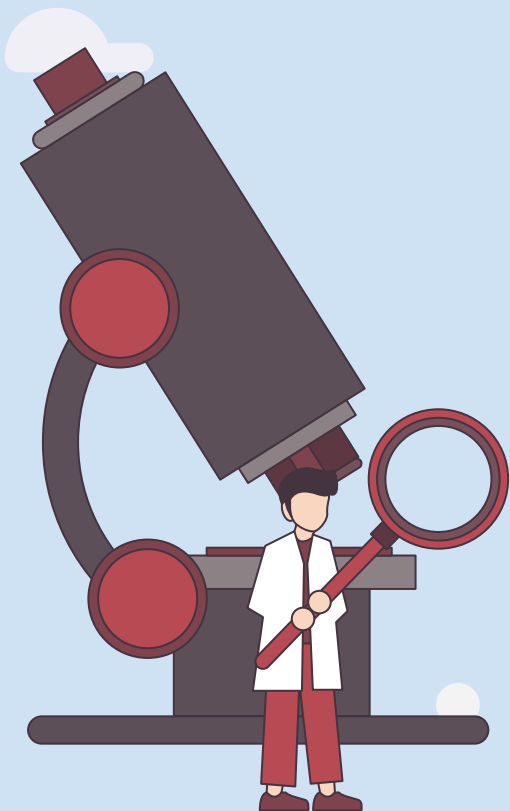




8

Limitations and Future Steps

Limitations



Train & Weather Dataset Range

→ Only includes data from 2007 to 2014. Weather conditions may have changed since 2014

Train Dataset Size

→ Train dataset size is comparatively smaller compared to test dataset

Time Constraint

→ Limited time for hyperparameter tuning to obtain better performing model



Future Works



Effect of the New Spraying Schedule

→ To update the model with latest data and check if it is effective or the trend still persists

Data on Location and No of Larvae

→ To study the trend on the larvae found to have a better plan on early prevention

New Technique on Treating Features

→ PCA can be tested for treating the collinearity between existing features





9

Conclusion

Conclusions

Recommended Model

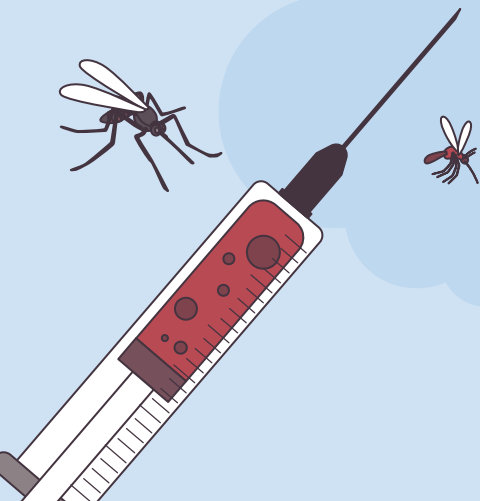
AdaBoost

Estimated Costs

~USD 625,678

Targeted Mosquito Abatement Efforts

1. Increase Larviciding Initiation at high risk areas
2. Lower the Threshold to Activate Adulticiding during June and July
3. Conduct Awareness Roadshow before May



Thank You

**Be Ready,
Stay Vigilant,
&
Abolish !**

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

