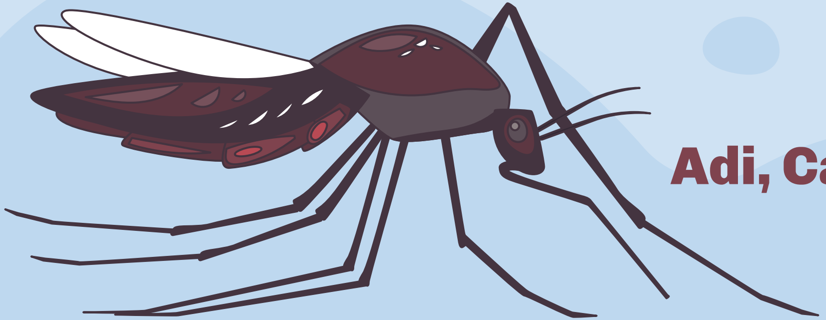
Project 4: West Nile Virus Prediction

DSI-28

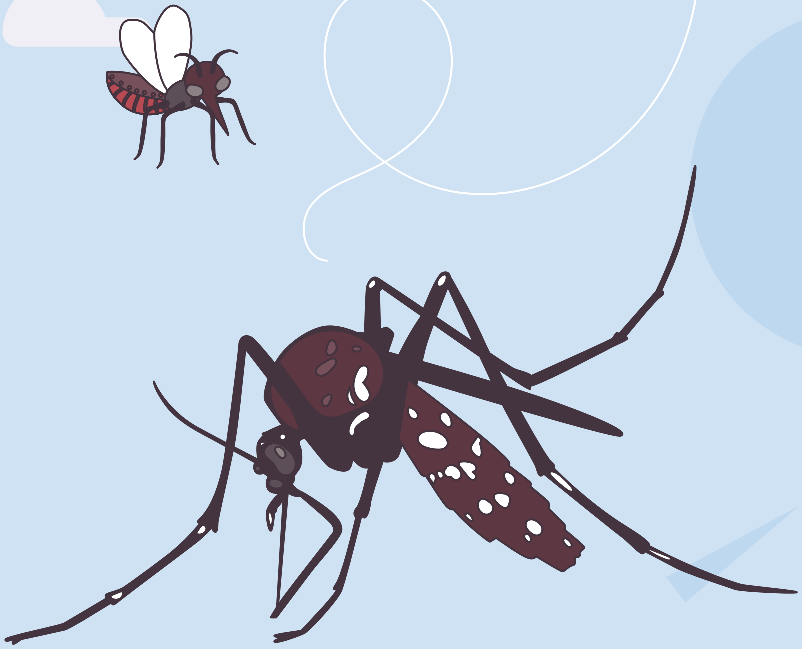
10 June 2022

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



01

Introduction and Problem Statement

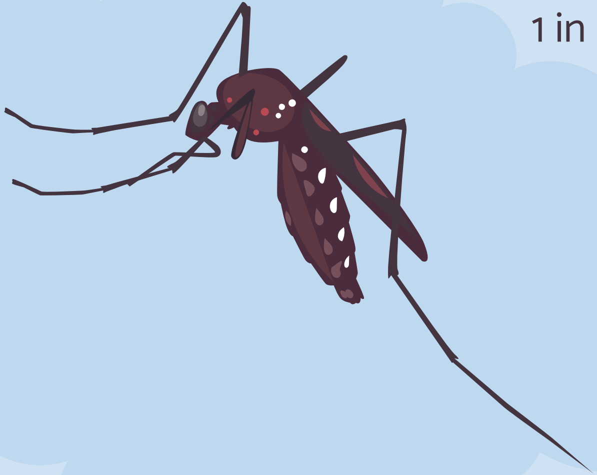
West Nile Virus (WNV)

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

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

Problem Statement

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.

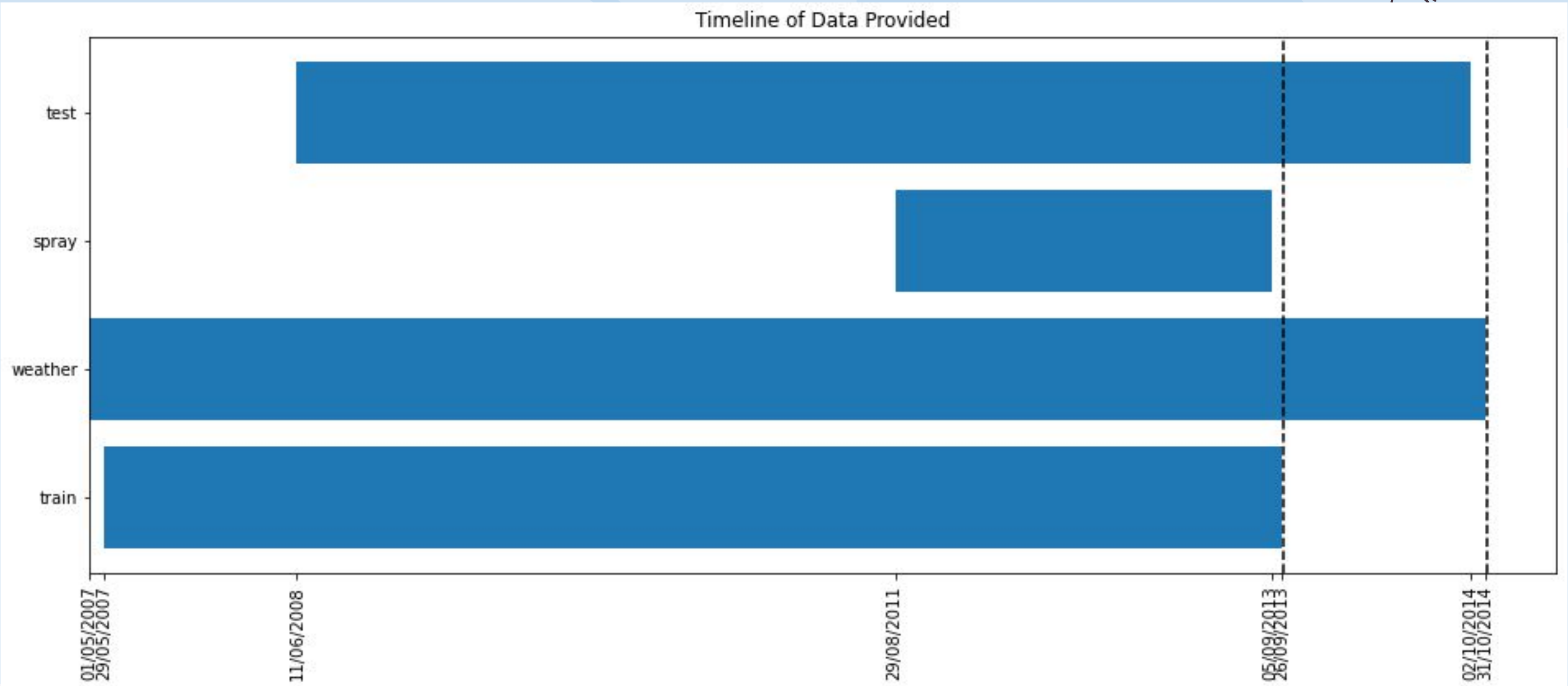
02

Data Cleaning

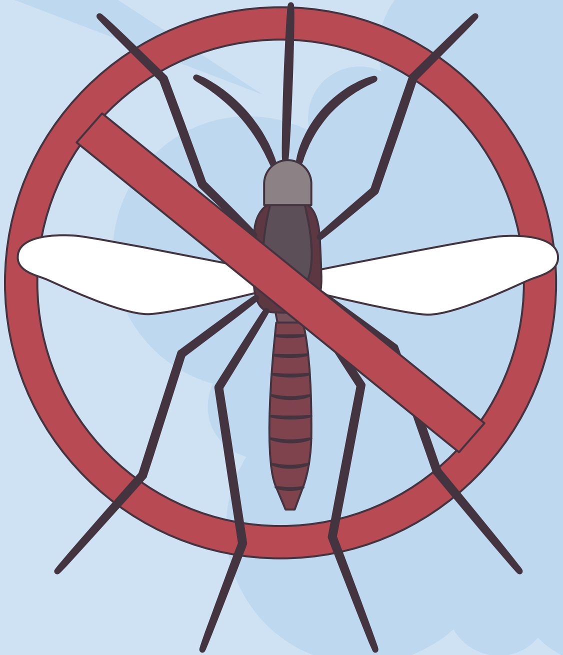
Data Provided

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 Provided



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 WetBulb

Using the ⅓ Rule

Imputing missing PrecipTotal

Using previous observation

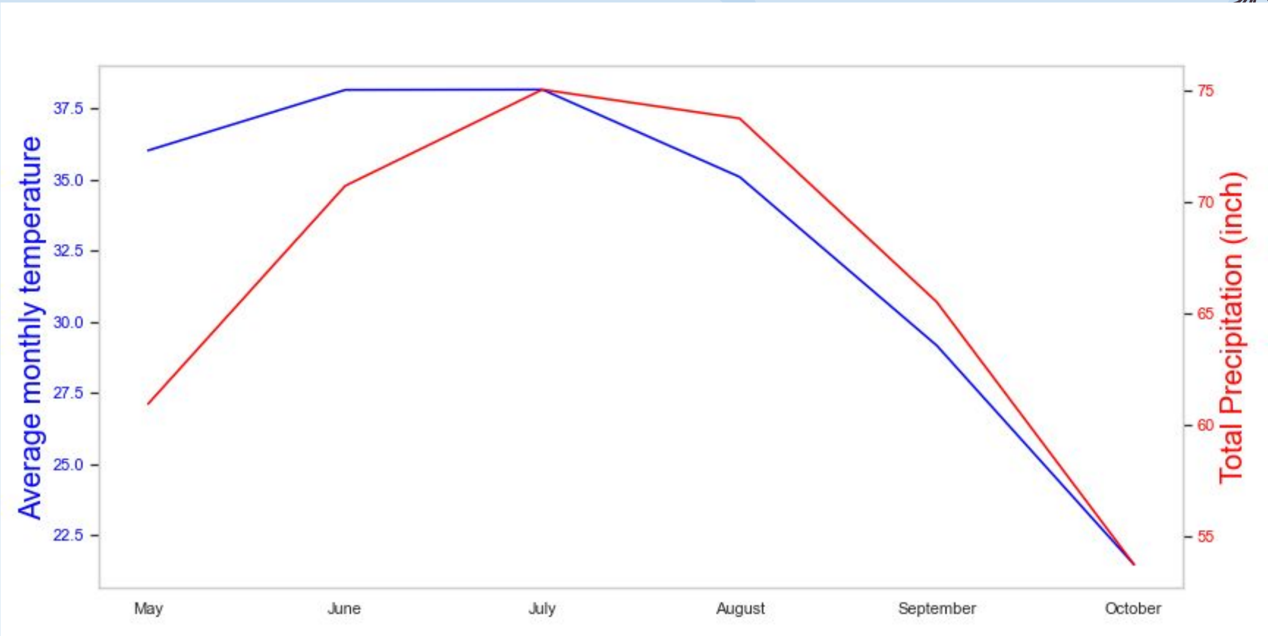
Imputing missing Tavg

Using Tmin and Tmax

03

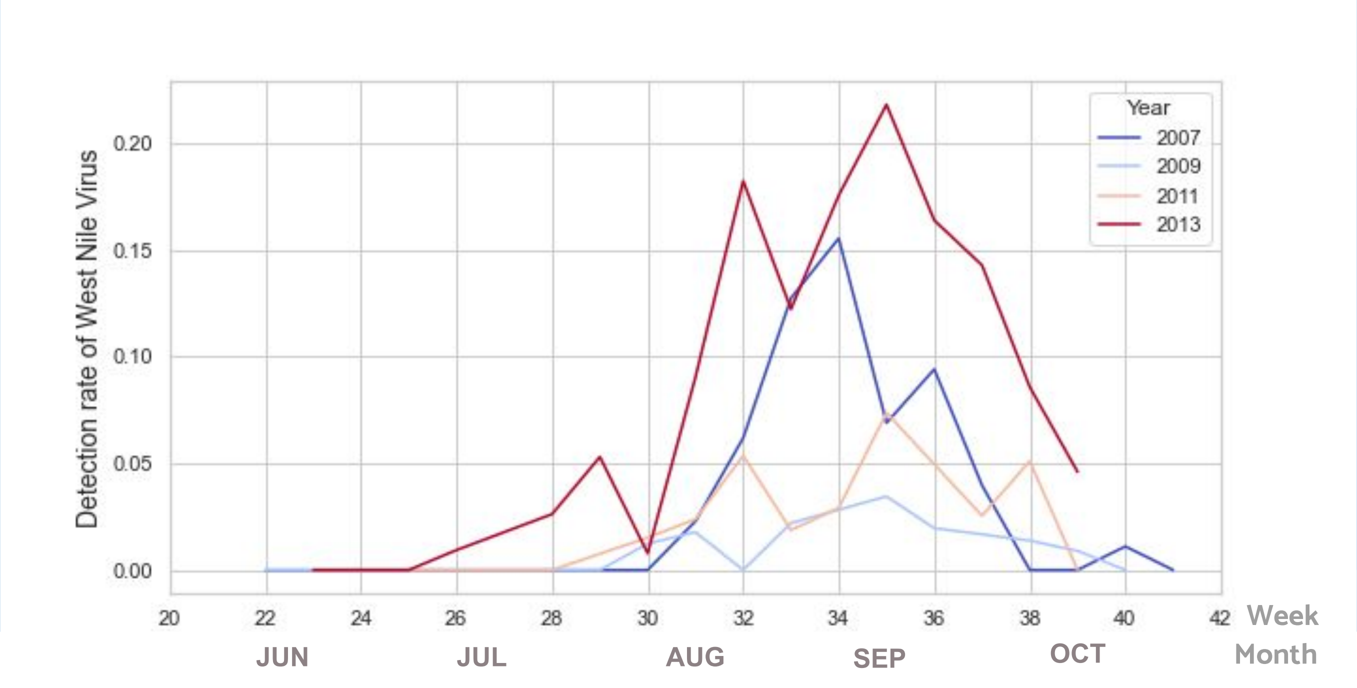
Exploratory Data Analysis (EDA)

Temperature and rainfall



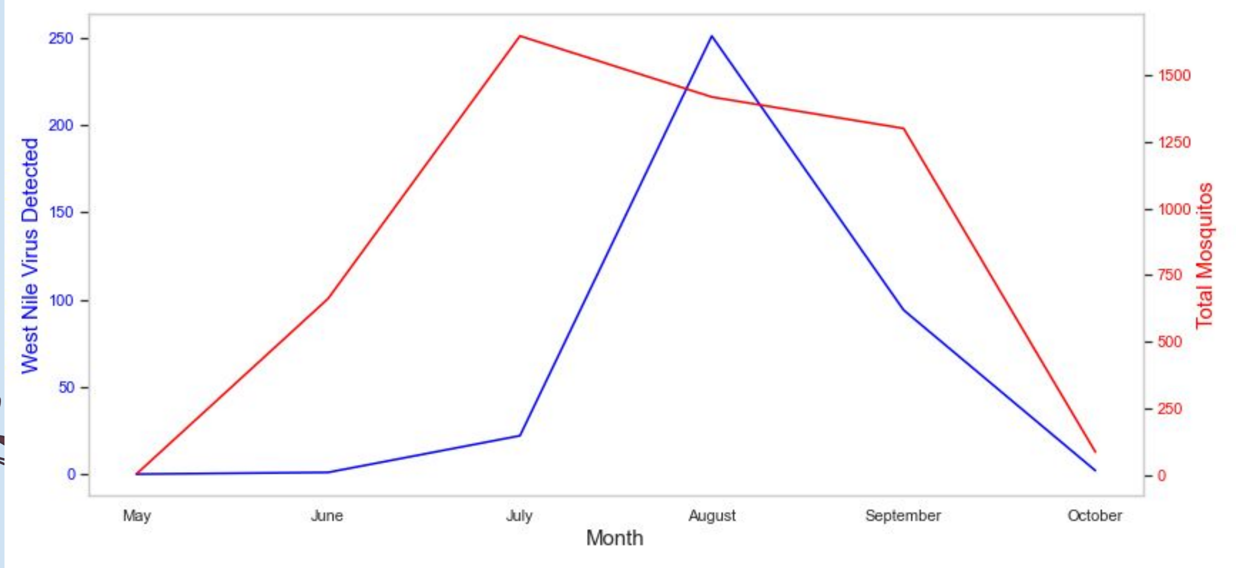
Peak temperature and rainfall occur around June-July-August each year.

WNV occurrence across the year



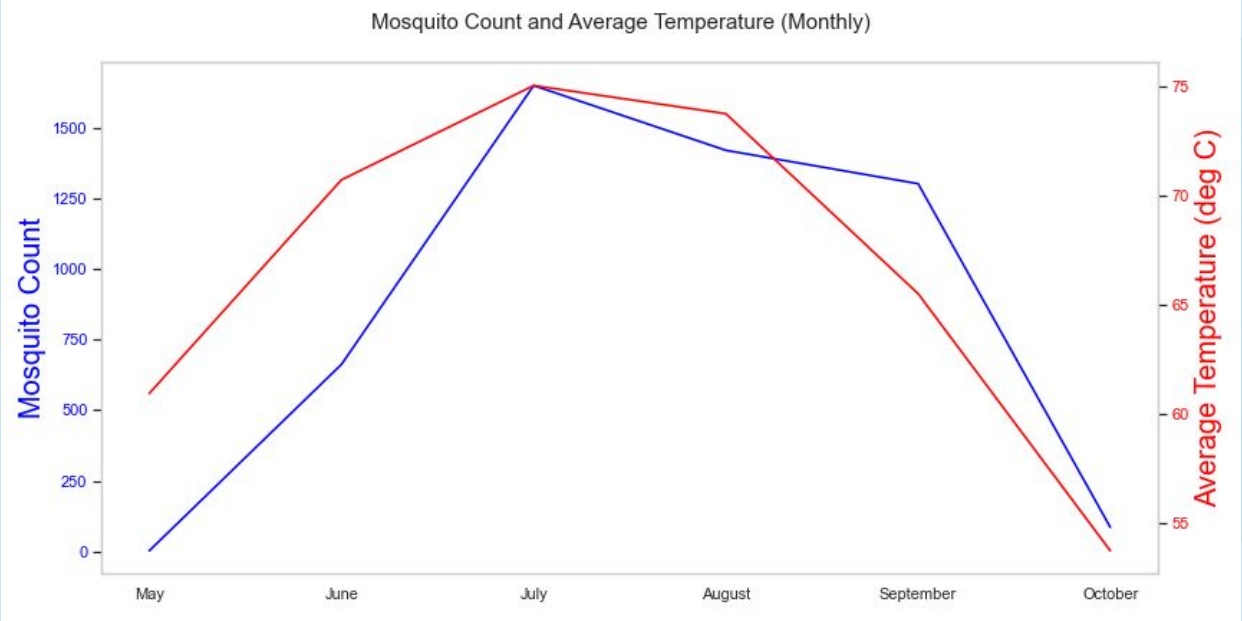
West Nile Virus detection in mosquitoes tend to peak in August and September

West Nile Virus use mosquitoes as transmission vectors



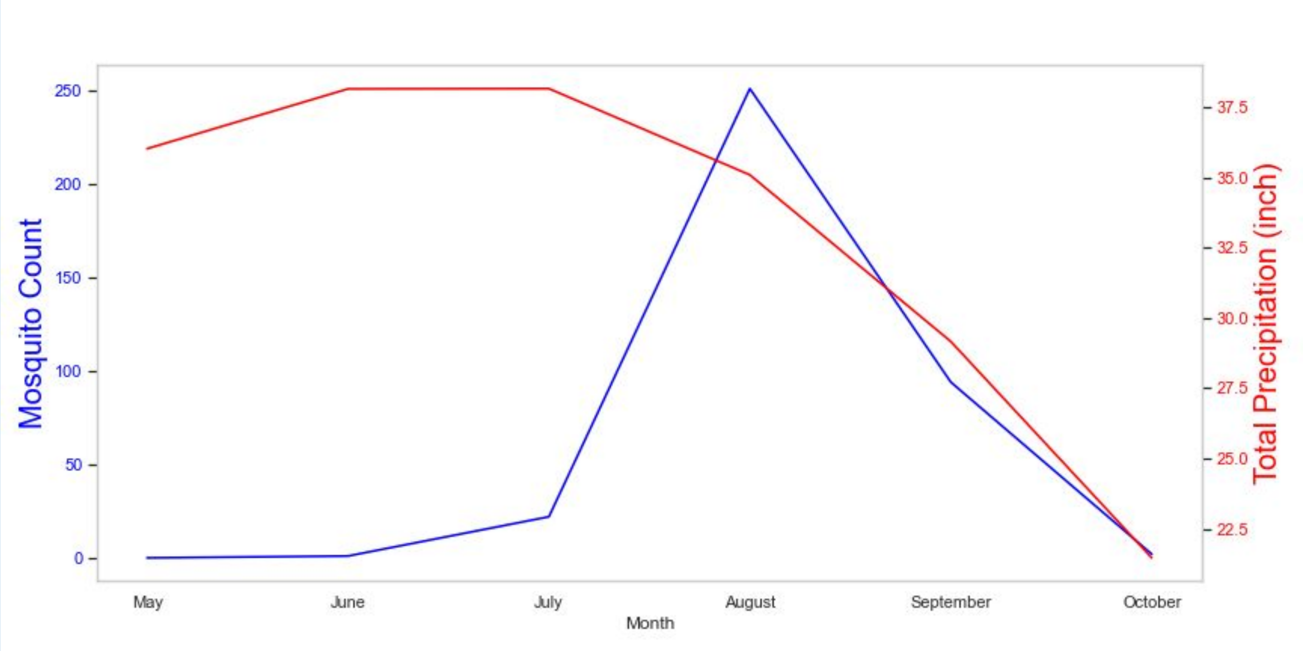
- Both West Nile Virus (WNV) cases and number of mosquitoes peak in July-August.
- Mosquito numbers peak first followed by WNV cases.
- Not surprising as mosquitoes are the vector carriers for the virus.

Mosquito numbers trend closely with temperature over the year



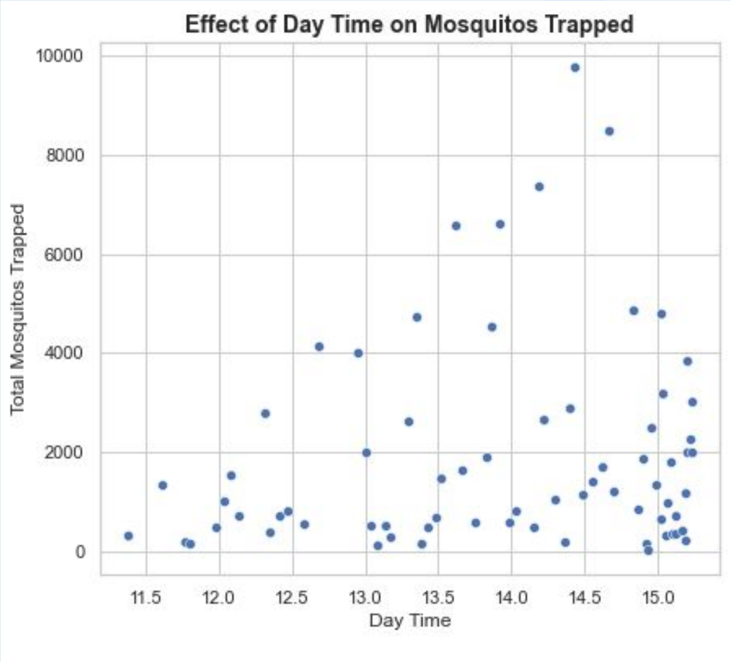
- Mosquito count increases with temperature, peaking in July-August
- Mosquito count decreases in later months as temperature drops
- Slight lag between mosquito count and temperature

Mosquito count tend to peak after rainfall



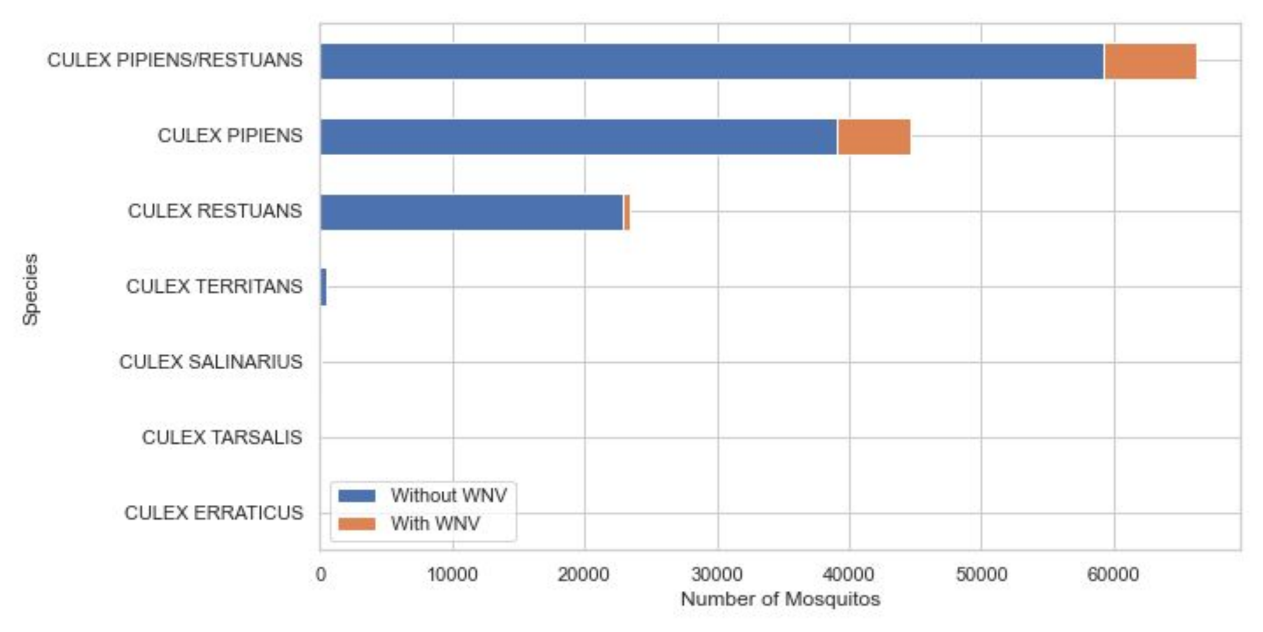
Peak in rainfall followed shortly by peak in mosquito numbers

Mosquito count increase with daytime length



More mosquitoes are detected as day length increases

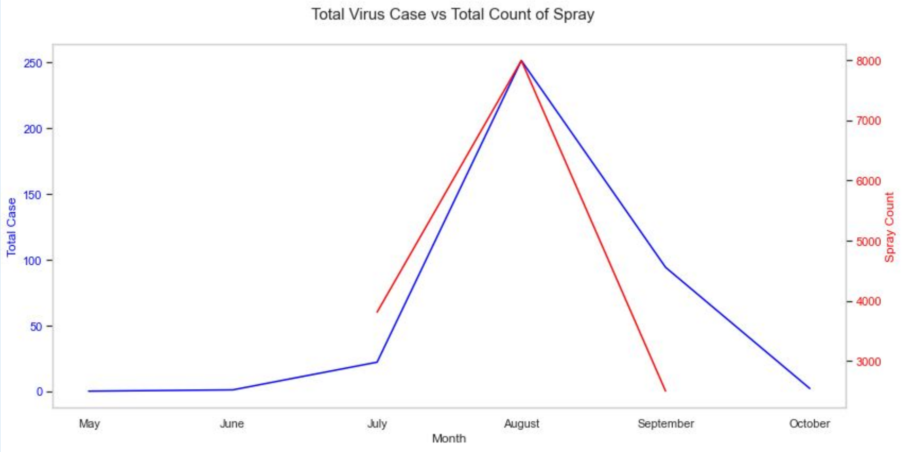
Culex Pipiens and Restuans species carry WNV

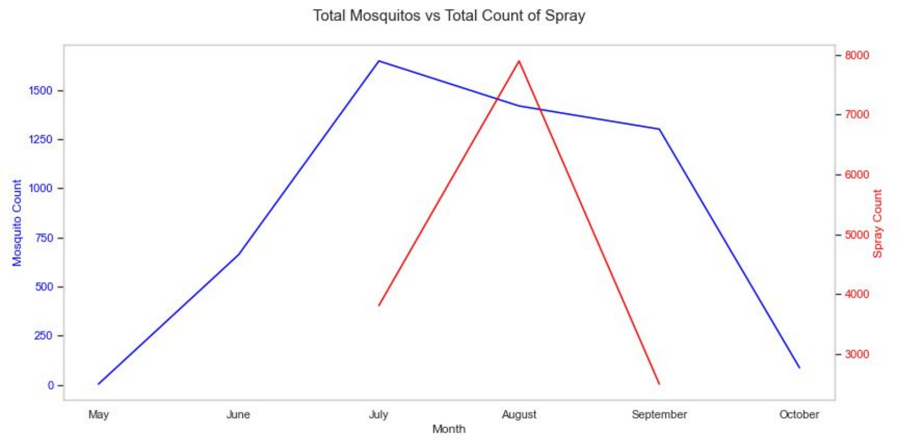


Mosquito species found with West Nile Virus:

- Culex Pipiens/ Restuans
- Culex Pipiens
- Culex Restuans

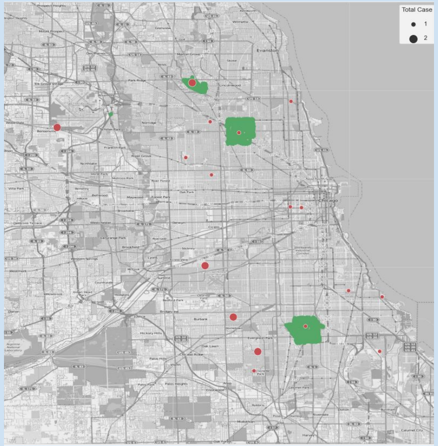
Spraying in response to WNV cases, as opposed to mosquito numbers

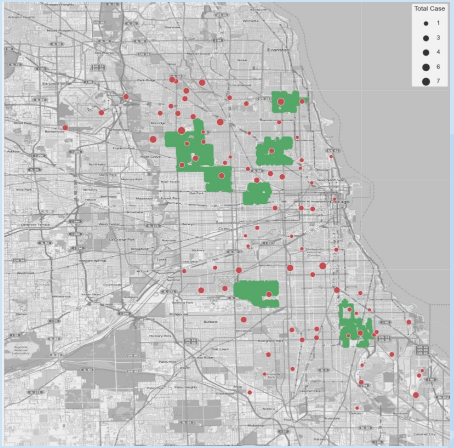


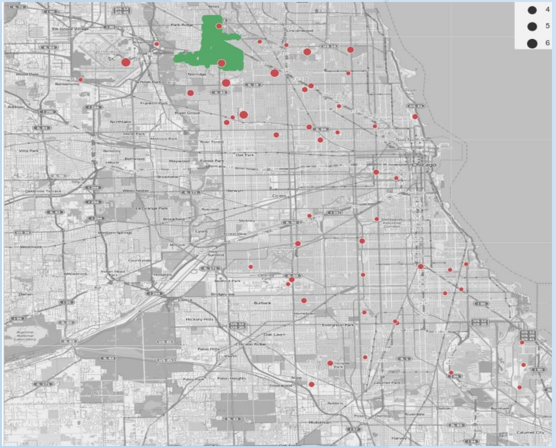


- Spraying counts increased almost instantaneously to emergence of WNV cases in July. (left)
- Increase in number of mosquitoes started much earlier in May. (right)
- Spraying may have started only in response to receiving reports of WNV cases

Disparity in spraying and actual WNV outbreak locations







- Areas sprayed with insecticide (green) and locations with WNV cases (red), from July to September (left to right)
- Sprayed area tend to cover much larger area than the neighbourhood, likely in expectation of WNV mosquito clusters
- However, there were many other areas with WNV cases which did not receive spraying.

04

Feature Engineering

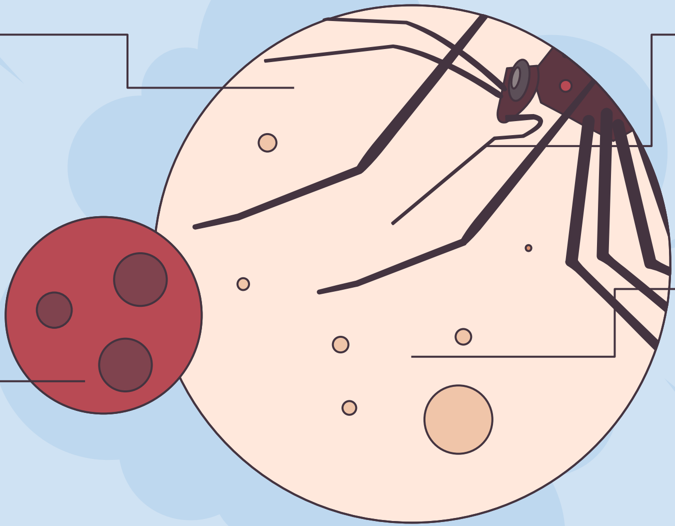
Factors Affecting Mosquito Breeding

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

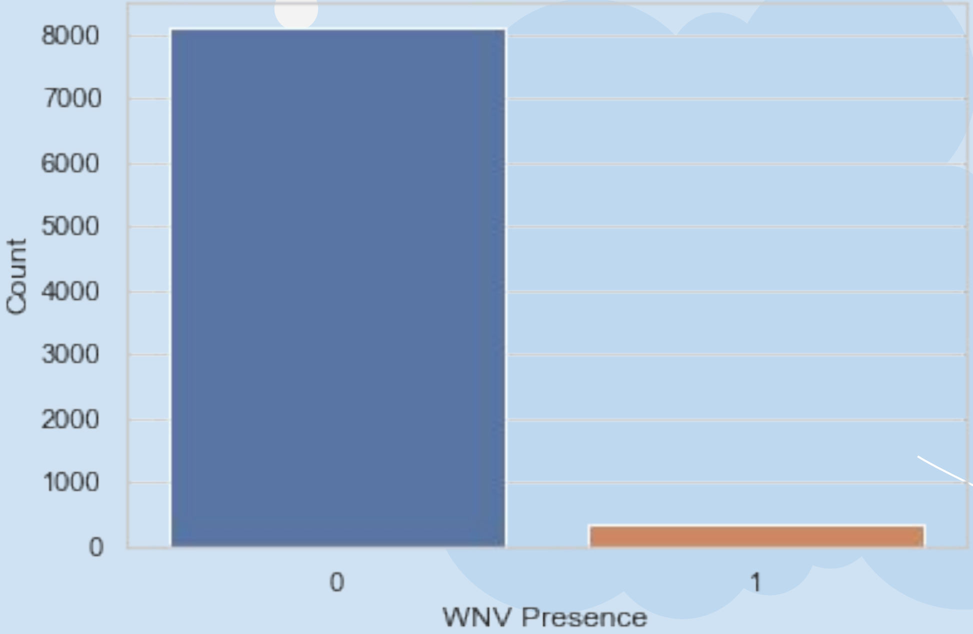
Culex Pipens
Culex Restuan

05

Modelling

- Logistics Regression
- Random Forest
- AdaBoost
- Gradient Boost
- XgBoost
- Support Vector Machine
- KNN

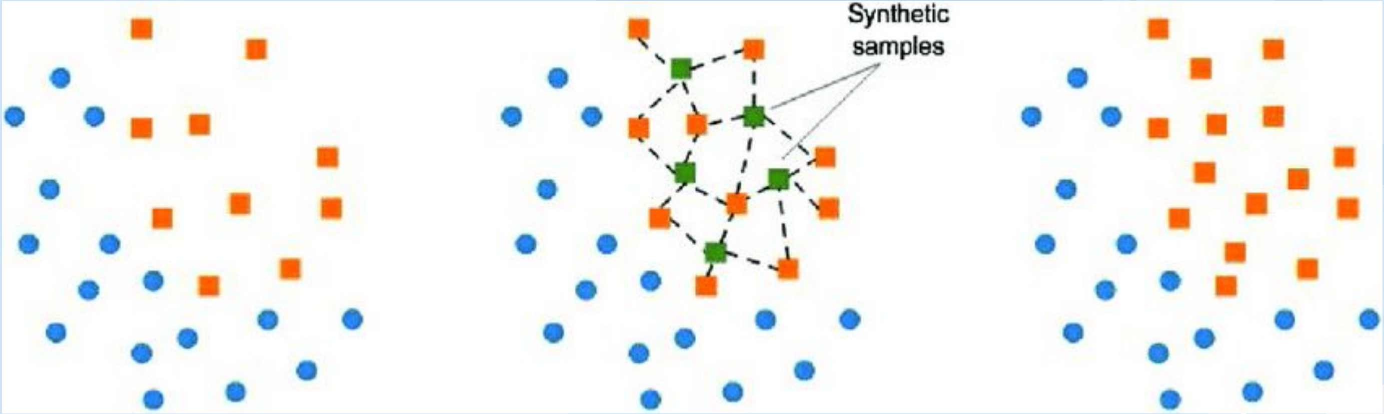
Imbalance Target



0 – No WNV
1 – WNV

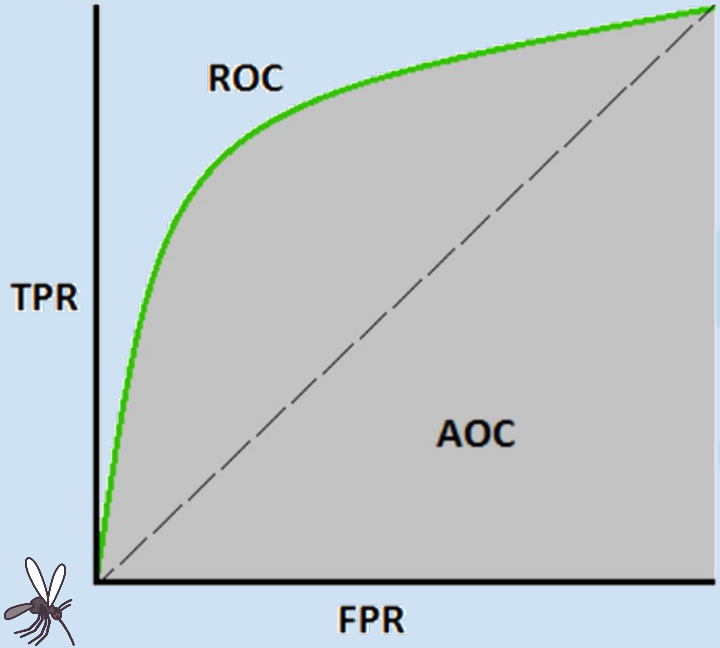
Poor performance on the minority class

Synthetic Minority Oversampling Technique (SMOTE)



Source of illustration:
https://www.researchgate.net/publication/333423855_Evaluation_of_performance_of_drought_prediction_in_Indonesia_based_on_TRMM_and_MERRA-2_using_machine_learning_methods

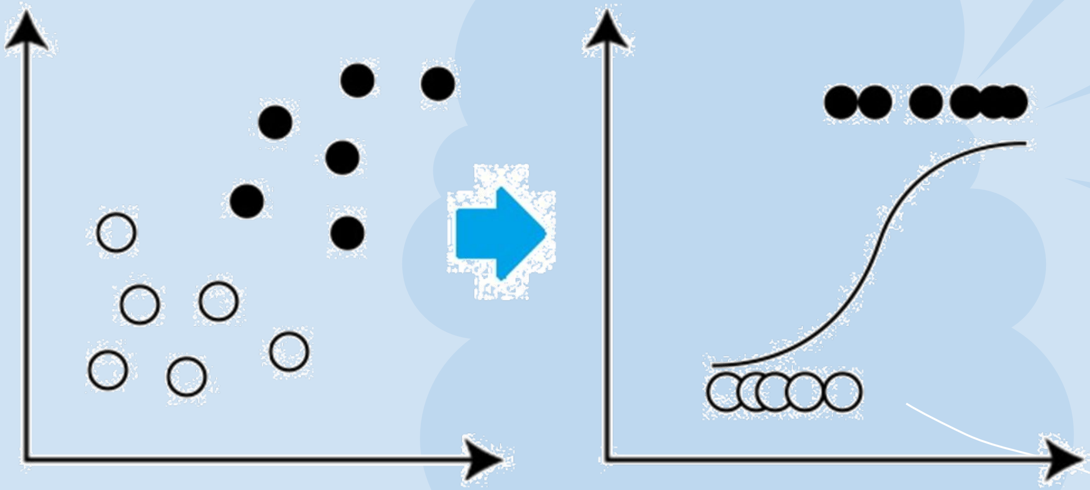
Model Evaluation Metric - AUC ROC



- 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

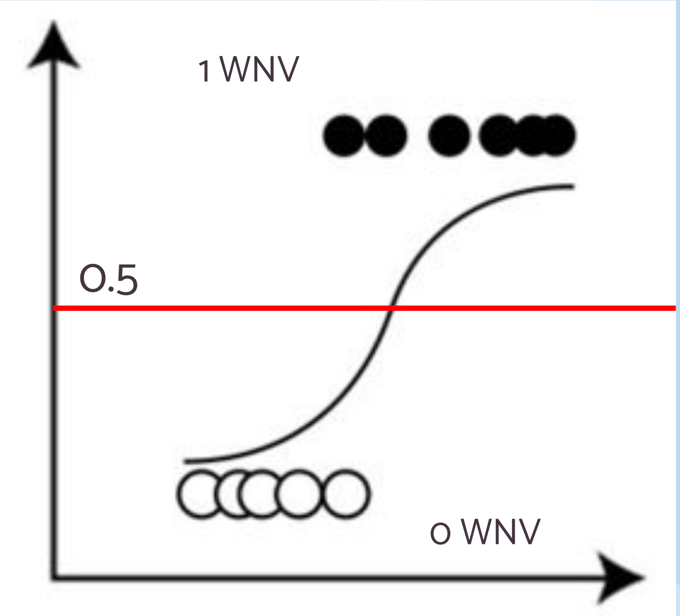
Source of illustration:
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Logistics Regression – Baseline Model



Source of illustration:
https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-logistic-regression-using-python/

Logistics Regression – Baseline Model

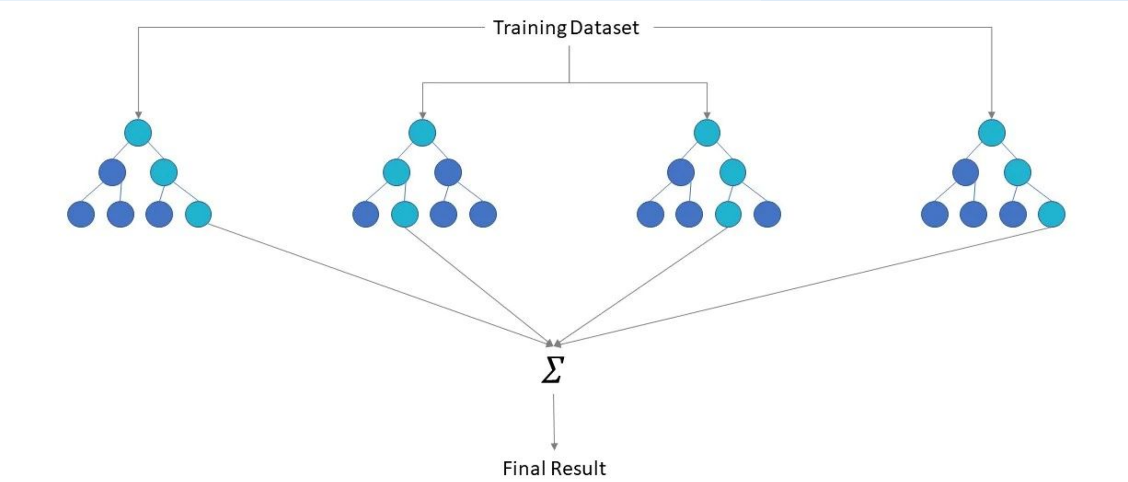


ROC
0.7771

ROC
0.8554

Source of illustration:
https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-logistic-regression-using-python/

Random (Decision) Forest



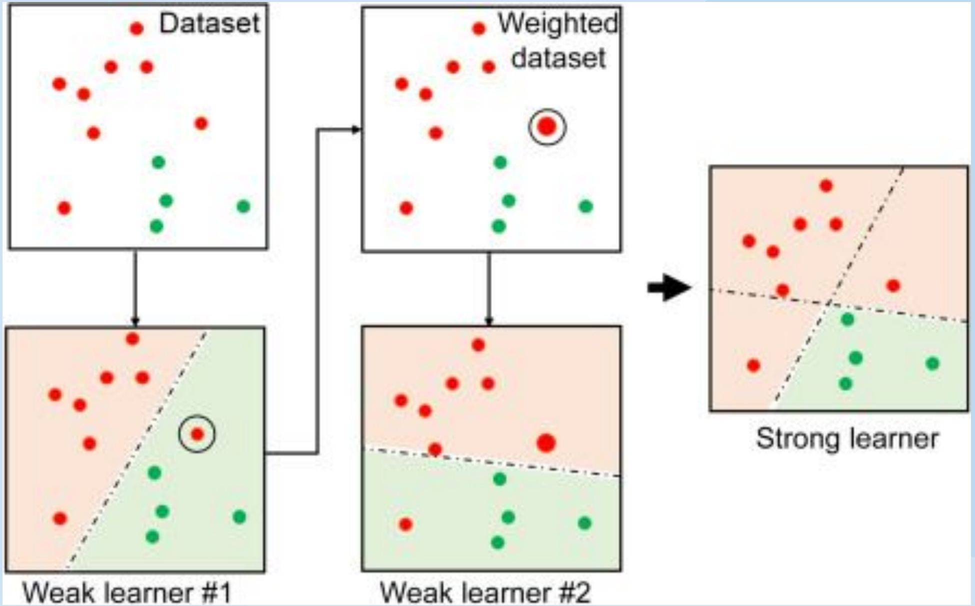
ROC AUC (Train)

0.9169

ROC AUC (Validation)

0.8790

Adaptive Boosting (a.k.a. AdaBoost)



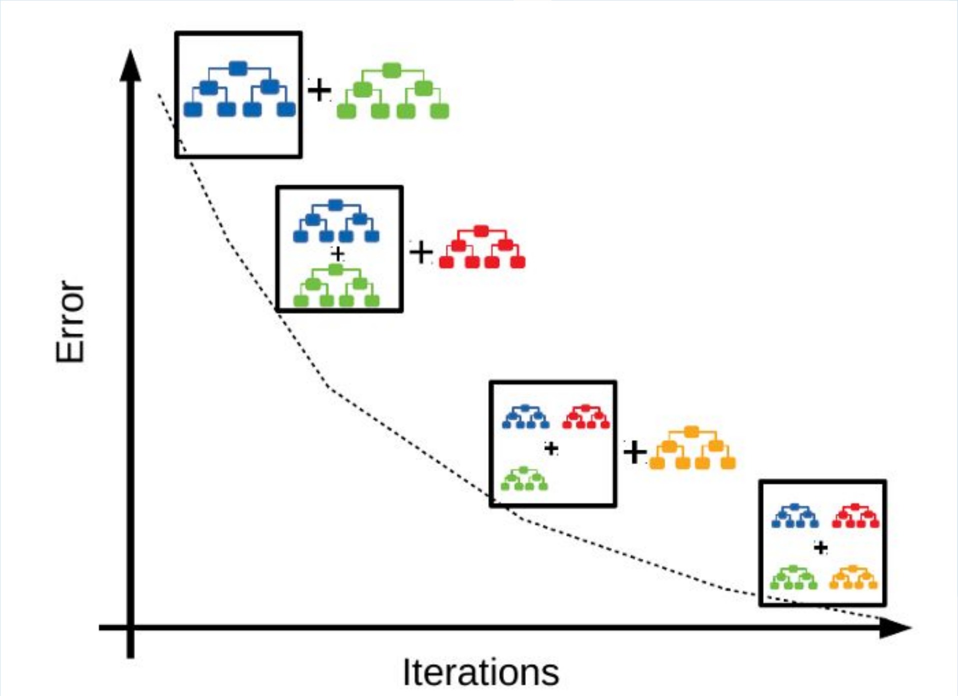
ROC AUC (Train)

0.8752

ROC AUC (Validation)

0.8809

Gradient Boosting



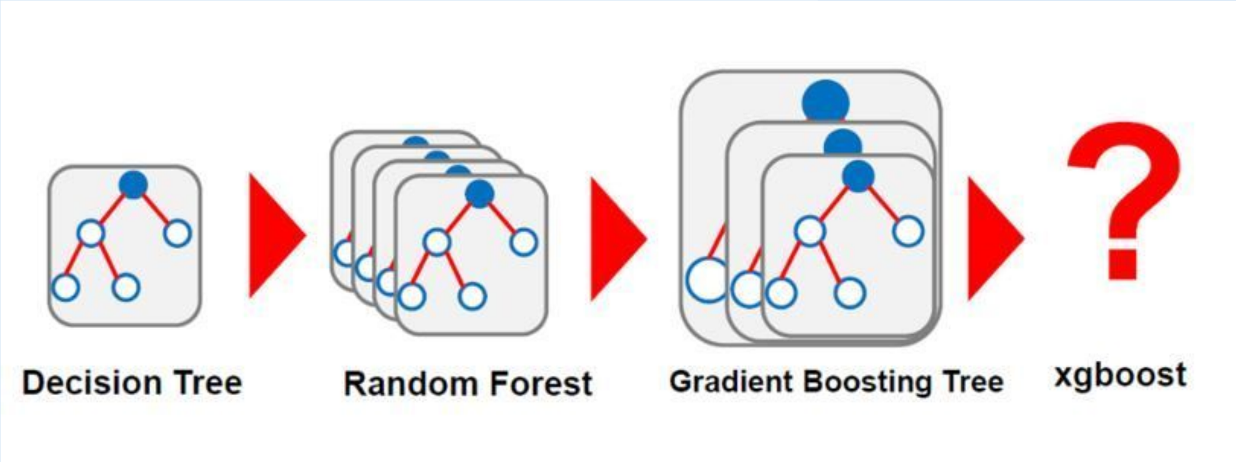
ROC AUC (Train)

0.9065

ROC AUC (Validation)

0.8808

eXtreme Gradient Boosting (a.k.a. XGBoost)



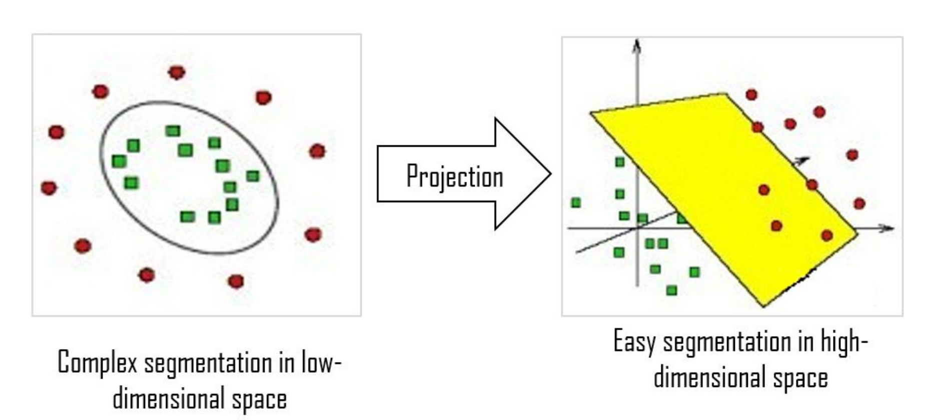
ROC AUC (Train)

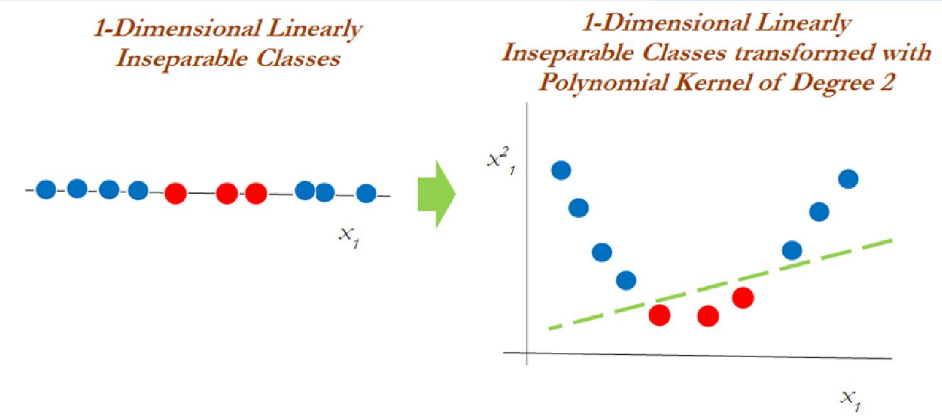
0.8762

ROC AUC (Validation)

0.8799

Support Vector Machine (a.k.a. SVM)





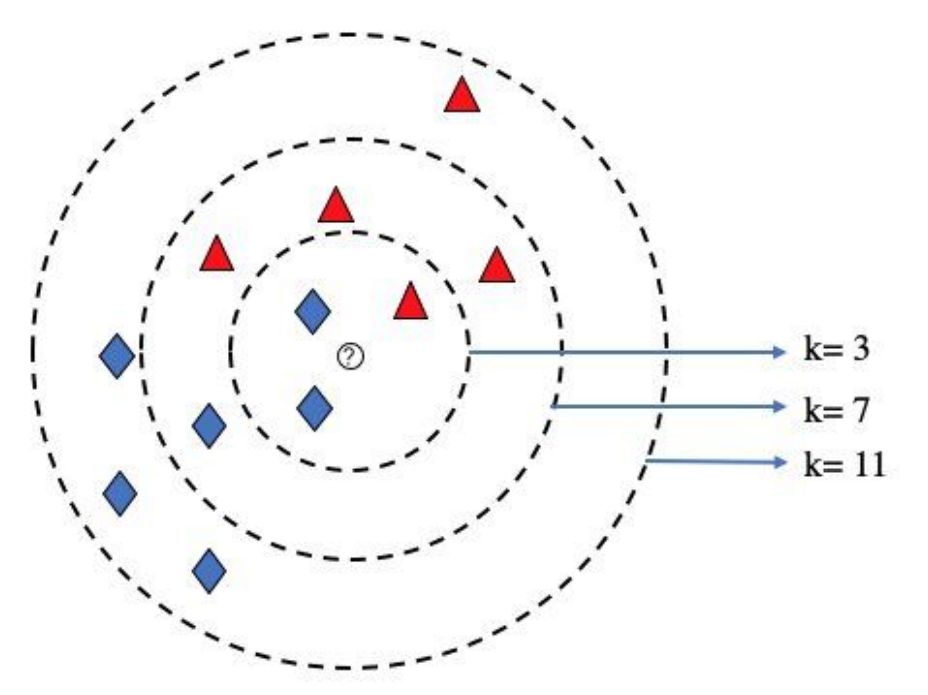
ROC AUC (Train)

0.8930

ROC AUC (Validation)

0.8772

k-Nearest Neighbours (a.k.a. k-NN)



ROC AUC (Train)

0.9361

ROC AUC (Validation)

0.8592

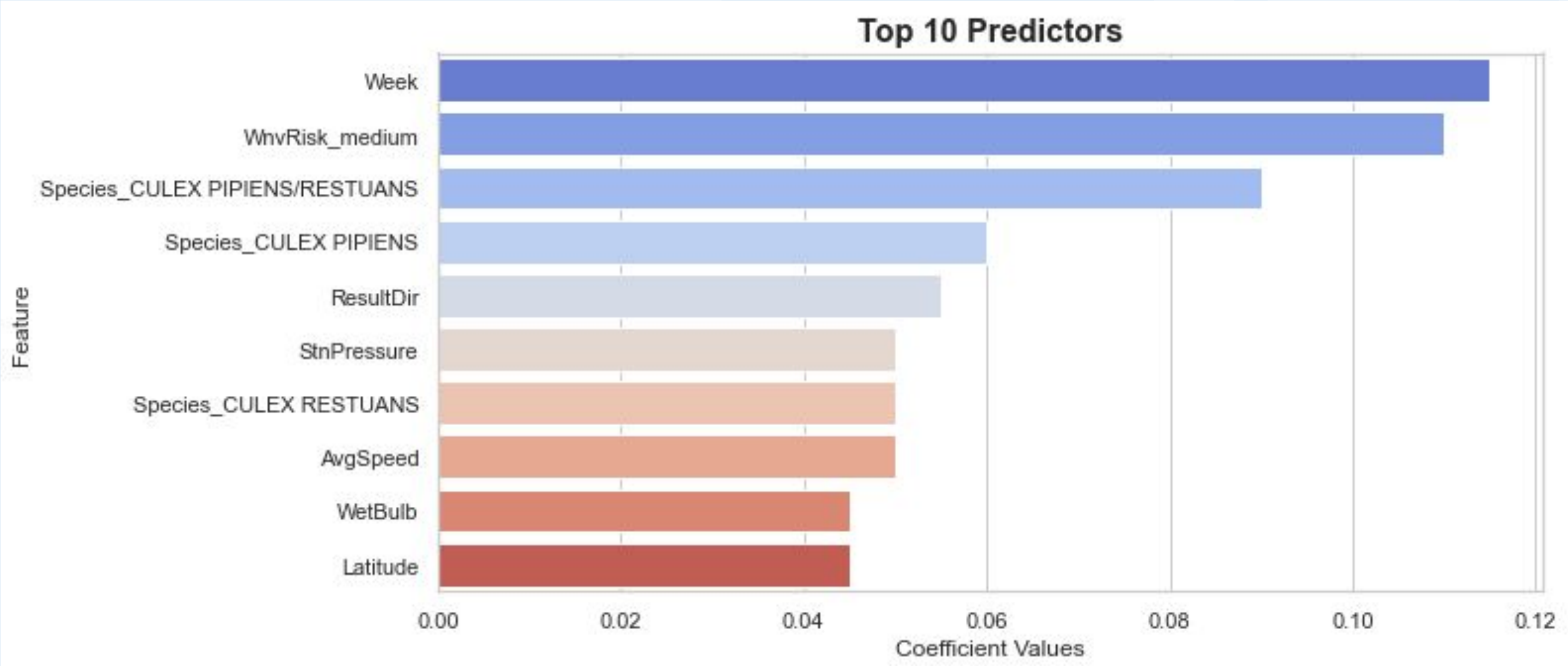
06

Modelling Results

Models Performance Summary

Model	ROC AUC			
	Train (A)	Validation (B)	delta (A) - (B)	Test (Kaggle)
Logistic Regression	0.8388	0.8542	-0.0154	**0.7228**
Random Forest	0.9169	0.8790	0.0379	0.6336
Ada Boost	0.8752	**0.8809**	-0.0057	0.6514
Gradient Boost	0.9065	0.8808	0.0257	0.6378
Extreme Boost	0.8762	0.8799	-0.0037	0.6638
Support Vector Machine	0.8930	0.8772	0.0158	0.6550
k-Nearest Neighbours	0.9361	0.8592	0.0769	0.6208

Top 10 AdaBoost Predictors



Why Logistic Regression outperforms all other models in this Test dataset?

Ensemble performs **no better** than best-performing member of the ensemble

- 1 top-performing model; and
- Other members **do not offer** any benefit or Ensemble is **unable to harness** their contribution effectively

Ensemble performs **worse** than best-performing member of the ensemble

- 1 top-performing model whose predictions are **made worse** by 1 or more poor-performing other models; and Ensemble is **unable to harness** their contributions effectively

— Jason Brownlee, PhD

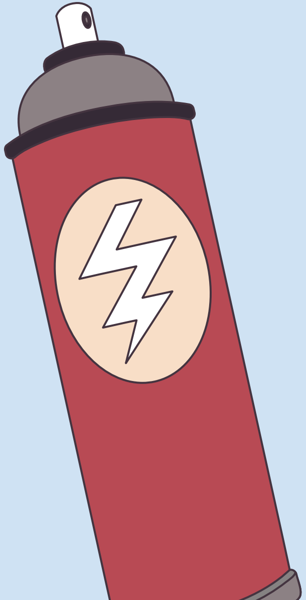
Source: https://machinelearningmastery.com/why-use-ensemble-learning/

07

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

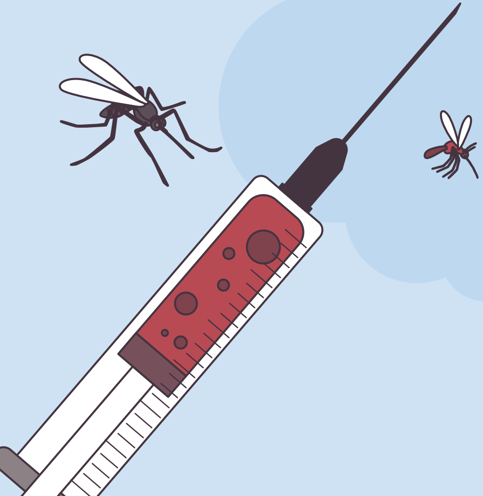
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Human Cases	22	1	6	16	**49**	6	42	2	11	10
Human Cases (Fatality)	-	-	-	3	2	1	-	-	-	-

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 whooping **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

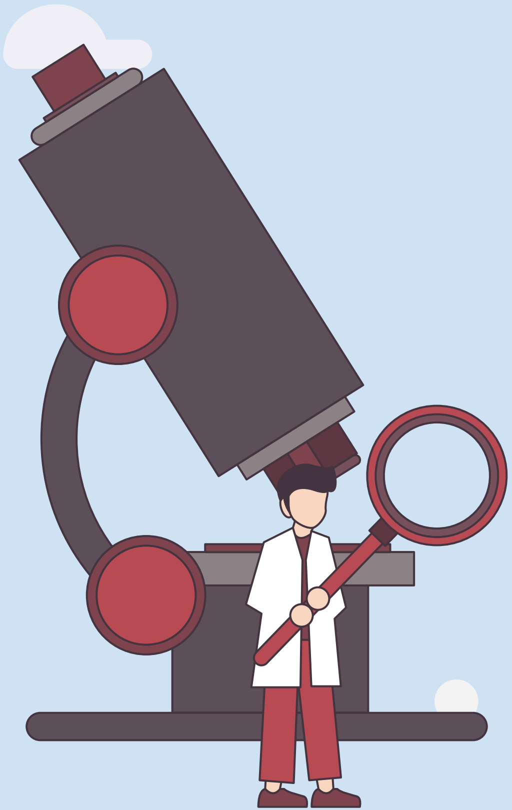
Conduct Awareness Roadshow

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

08

Limitations and Future Steps

Limitations



Train & Weather Dataset Range

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

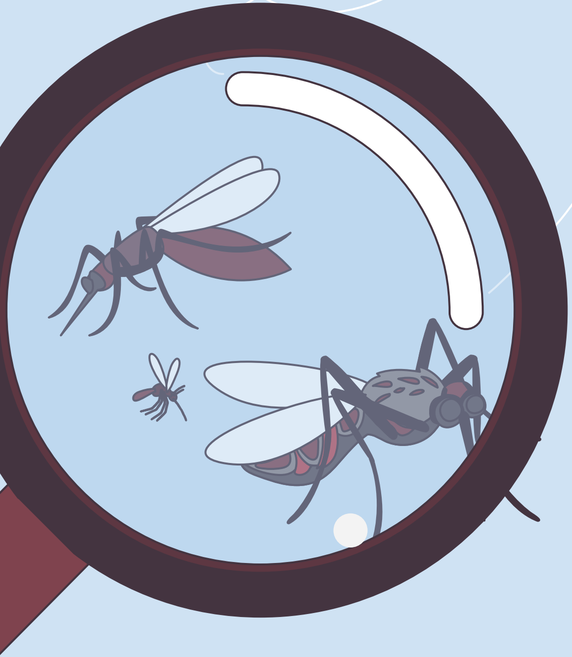
Train Dataset Size

→ Train dataset size is comparatively smaller compare to test dataset

Time Constraint

→ Limited time for hyperparameter tuning to obtain better performing model

Future Works



Effect of the New Spraying Schedule

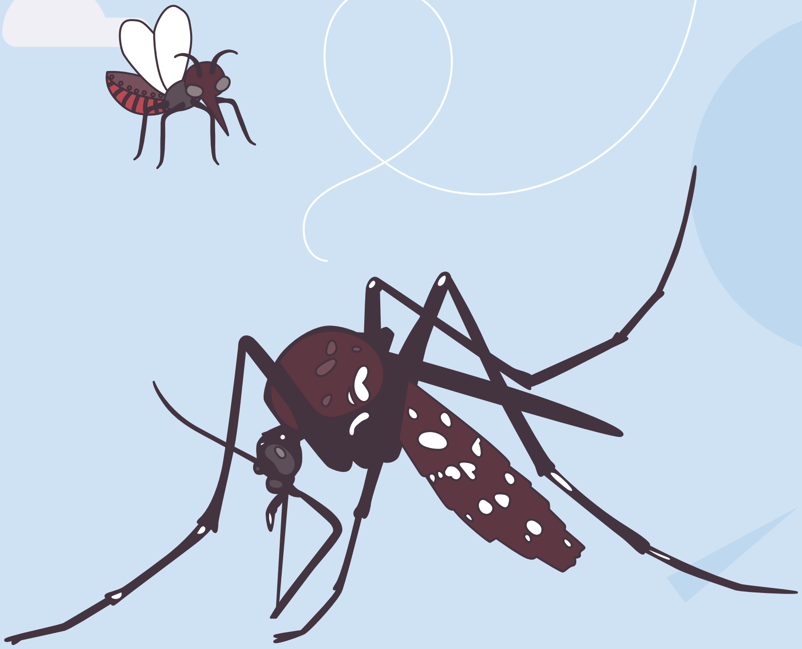
→ To update the model with latest data and check if it is effective or the trend still persists

Data on Location and No of Larvaes

→ To study the trend on the larvaes 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



09

Conclusion

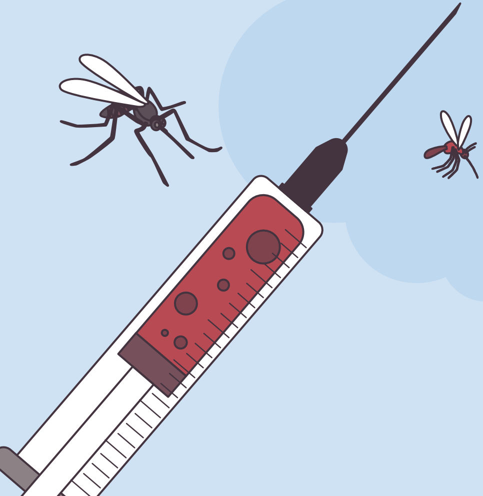
Conclusions

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

Benefits outweigh Costs by At Least 2x → Continue 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

- From 2012 to 2021, most number of Human Cases resulting in Fatality = **3** (in year 2015)
- Assume all things constant,
  - Cost of Mosquito Abatement Program's efforts for 2023 = **USD 520,698**
  - Assumed worst scenario of all Human Cases result contracted Acute Flaccid Paralysis (AFP) & eventual fatality,

    Benefit or economic burden caused by the West Nile Virus alleviated = **USD 3,003,756** (= 3 x USD 1,001,252)

Recommendation: Continue with Mosquito Abatement Program's efforts for 2023 and beyond.

Mosquito Abatement Program Costs for 2023: USD 520,698

Budget	Year 1	Year 2	Year 3	Year 4	Year 5	Option Year 6	Option Year 6
	2018	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
Estimated Annual Fee (USD)	448,819.70 *	462,284	476,153	490,808	505,532	**520,698**	**536,319**
CPI %		3.00%	3.00%	3.08%	3.00%	3.00%	3.00%

Source:

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

Total Economic Cost per Individual for 2023: ~USD 1,001,252

	Acute Flaccid Paralysis (AFP)			
Initial Costs	Min	Median	Mean	Max
Total inpatient hospital costs (USD)	7,013	28,756	97,154	365,681
Total lost productivity (USD)	321	2,957	17,105	201,752
Total initial costs (USD)	**7,454**	**34,768**	**114,257**	**392,266**
Long-Term Costs	Min	Median	Mean	Max
Medical appointments (USD)	626	5,082	6,212	16,740
Additional care costs (USD)	0	385	1,824	8,470
Medicines, equipment, or modifications (USD)	147	817	60,015	591,107
Subtotal of long-term medical costs (USD)	864	7,368	68,053	608,987
Lost productivity (USD)	0	9,373	38,184	197,991
Total long-term costs (USD)	**864**	**31,322**	**106,236**	**608,987**
Total Costs (USD)	**8,318**	**66,090**	**220,493**	**1,001,252**

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