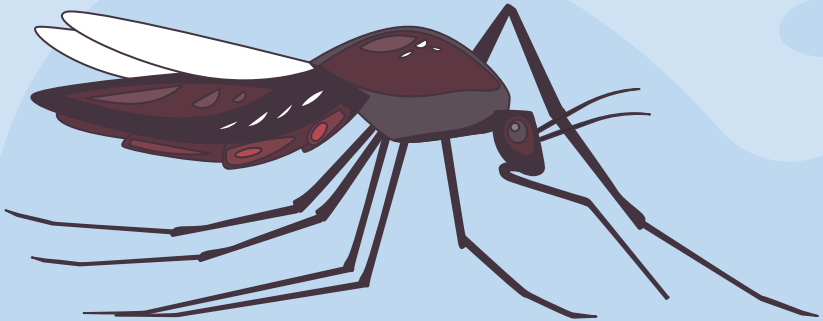


West Nile Virus In Chicago

A status update to members of Centre for Disease Control and Prevention (CDC)

5pm Disease and Treatment Agency

By: Adriel Chen
Andrew Chia
Estebelle Khong
Tay Yi Li

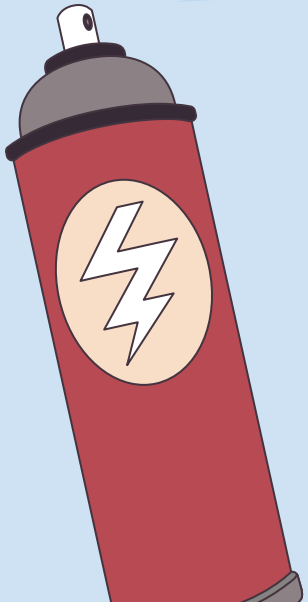
Agenda

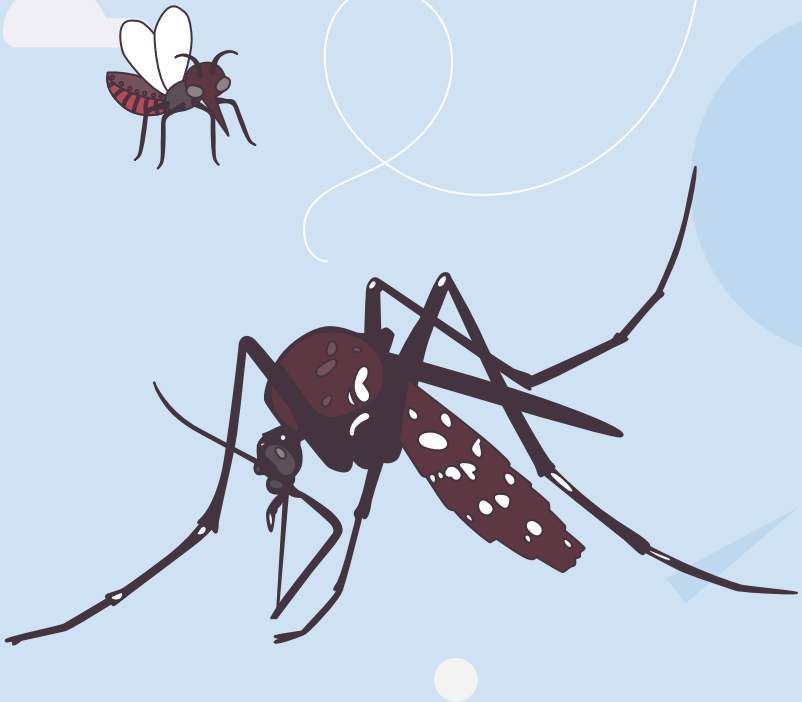
01 Background and Problem Statement

02 Exploratory Data Analysis

03 Modelling and Evaluation

04 Cost-Benefit Analysis, Conclusion and Recommendation





01

Background and Problem Statement

What is West Nile Virus?

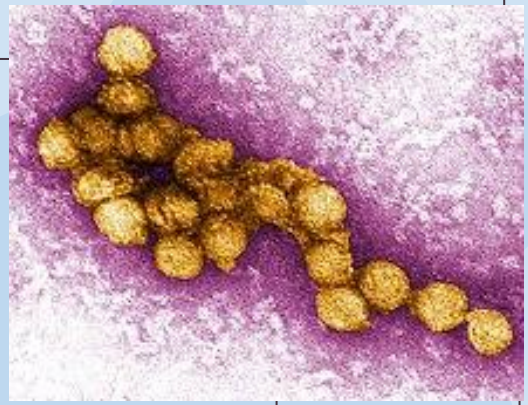
A single-stranded RNA virus that causes West Nile fever.

1. Flaviviridae Family

Family also includes Zika, dengue, and yellow fever.

2. Host: Birds

Virus is found in more than 250 species of birds, especially crows, jays, and ravens.



3. Transmission

Bird -> Mosquitoes -> Human

2 common mosquito types:

Culex Pipiens

Culex Restuans

4. Likely Place of Transmission

Place where birds and human will gather, like farmer markets.

5. Symptoms?

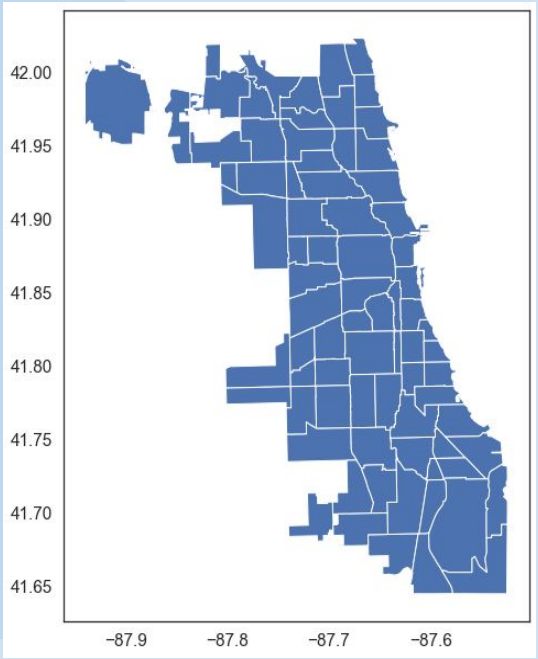
Mammals, especially **Human** and **Horse** are likely to develop a flu-like illness or signs of neurologic disease.

West Nile Virus Situation in US

- An Israeli strain of virus emerged in New York city in fall of 1999.
  - Maybe introduced by infected mosquito via air/sea transport
  - Or through migration of infected wild birds.

- Then, transmitted throughout US by migrating birds.

- In 2002, West Nile Virus reached Chicago for the 1st Time !
  225 human cases reported that summer.



Problem Statement

"This projects aim to use data science methods (classification modellings) to predict the places in Chicago where the West Nile Virus is present, with prediction having the higher AUC performance the better (as close to 1 as possible), so as to enable a more accurate and effective plan in deploying pesticides spraying throughout the state."

Datasets We Used

Main dataset

Records the location of mosquito traps set up, and whether West Nile Virus is present

Train

Time Range:
May to Oct
2007/09/11/13

Test

Time Range:
June to Oct
2008/10/12/14

Spray dataset

Records the date, time and location where the spraying of mosquitoes is conducted.

Time Range:
Sep 2011
Jul-Sep 2013

Weather dataset

Records the weather condition in Chicago from 2 stations.

Time Range:
May to Oct
2007-2014

Workflow Outline

Step 1

EDA

Step 2

Modelling and Evaluation

Step 3

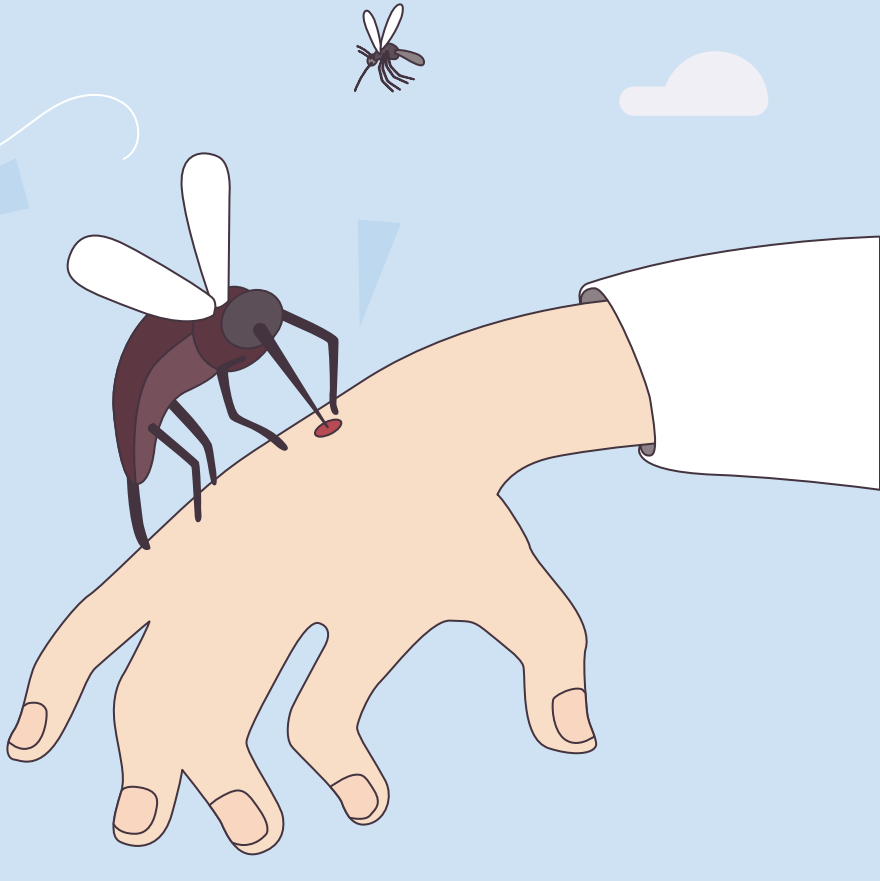
Cost-Benefit Analysis

Step 4

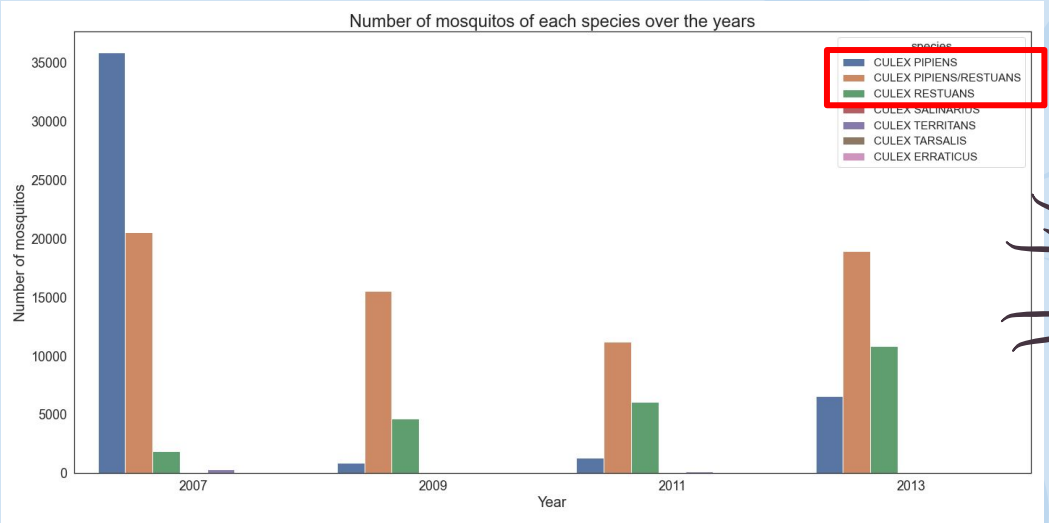
Conclusions and Recommendations

02

Exploratory Data Analysis

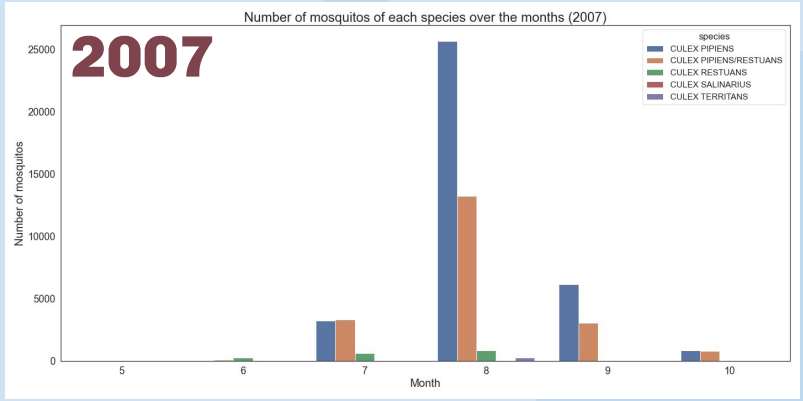


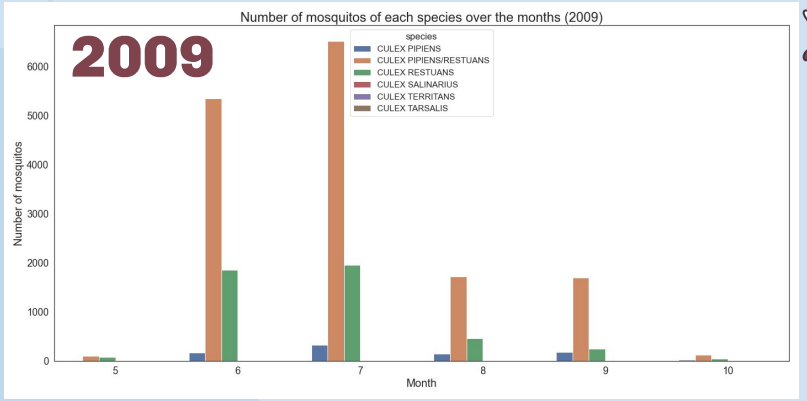
EDA: Overview of Mosquito Data from dataset

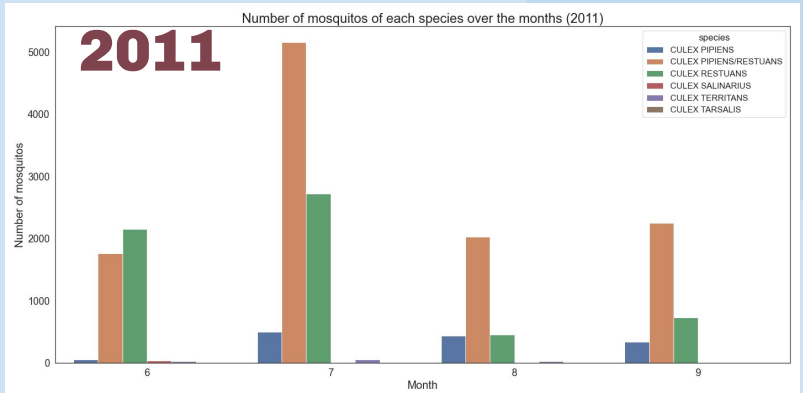


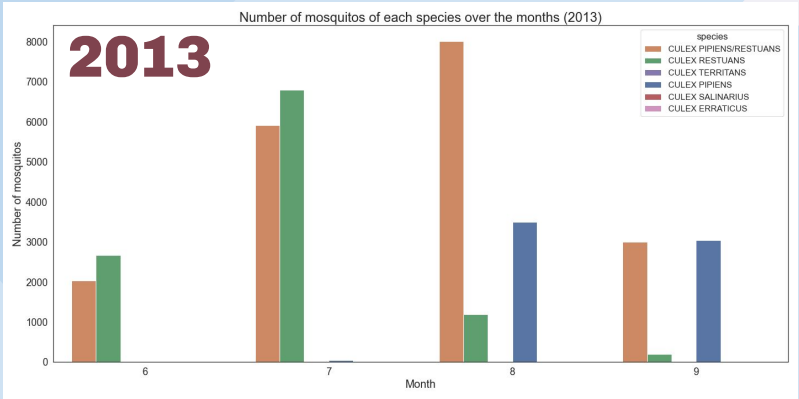
Overall, Culex Pipiens and Restuans are the dominant species

EDA: Mosquito appearance over the years

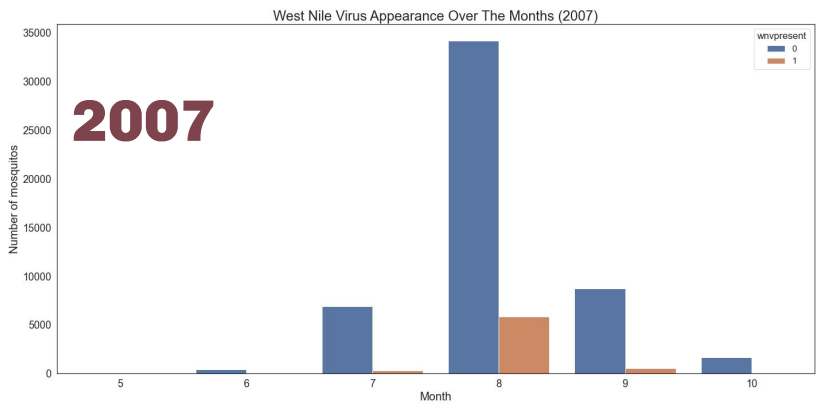


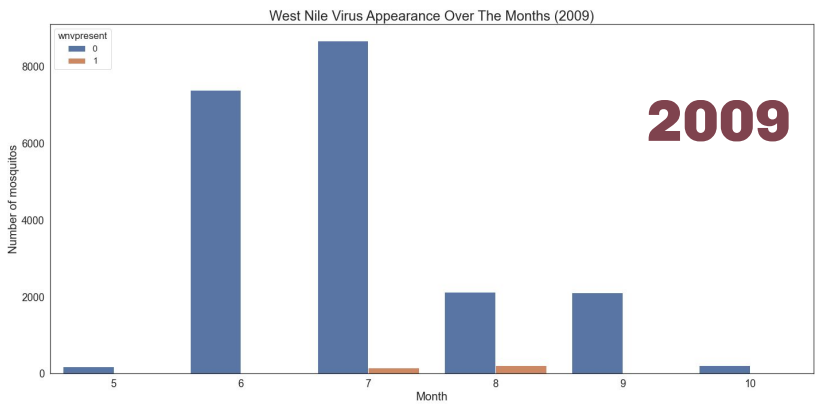


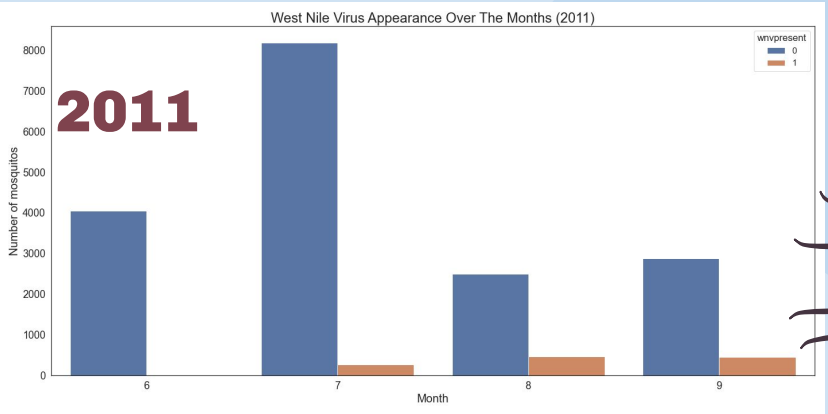


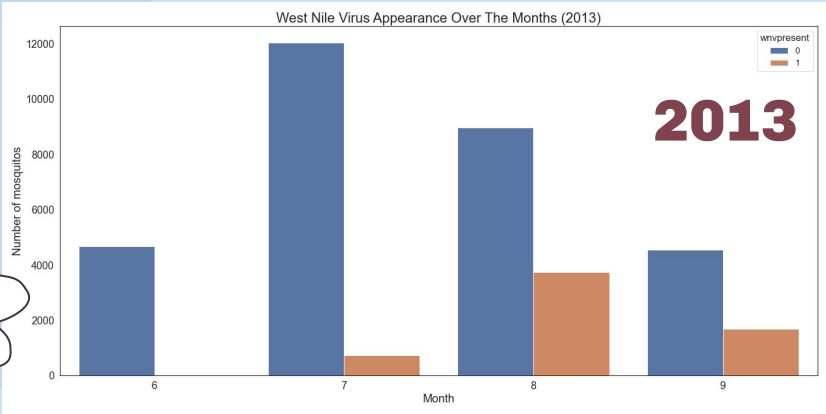


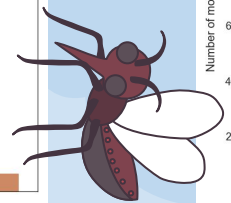
EDA: West Nile Virus appearance over the years







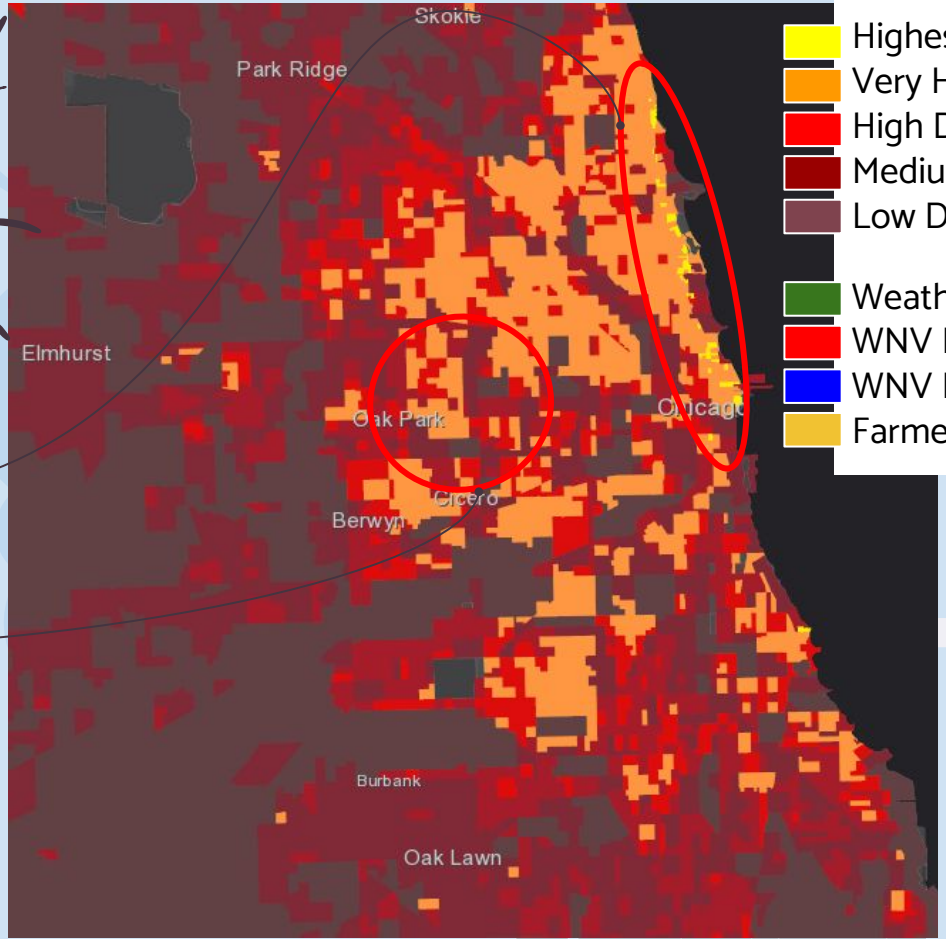




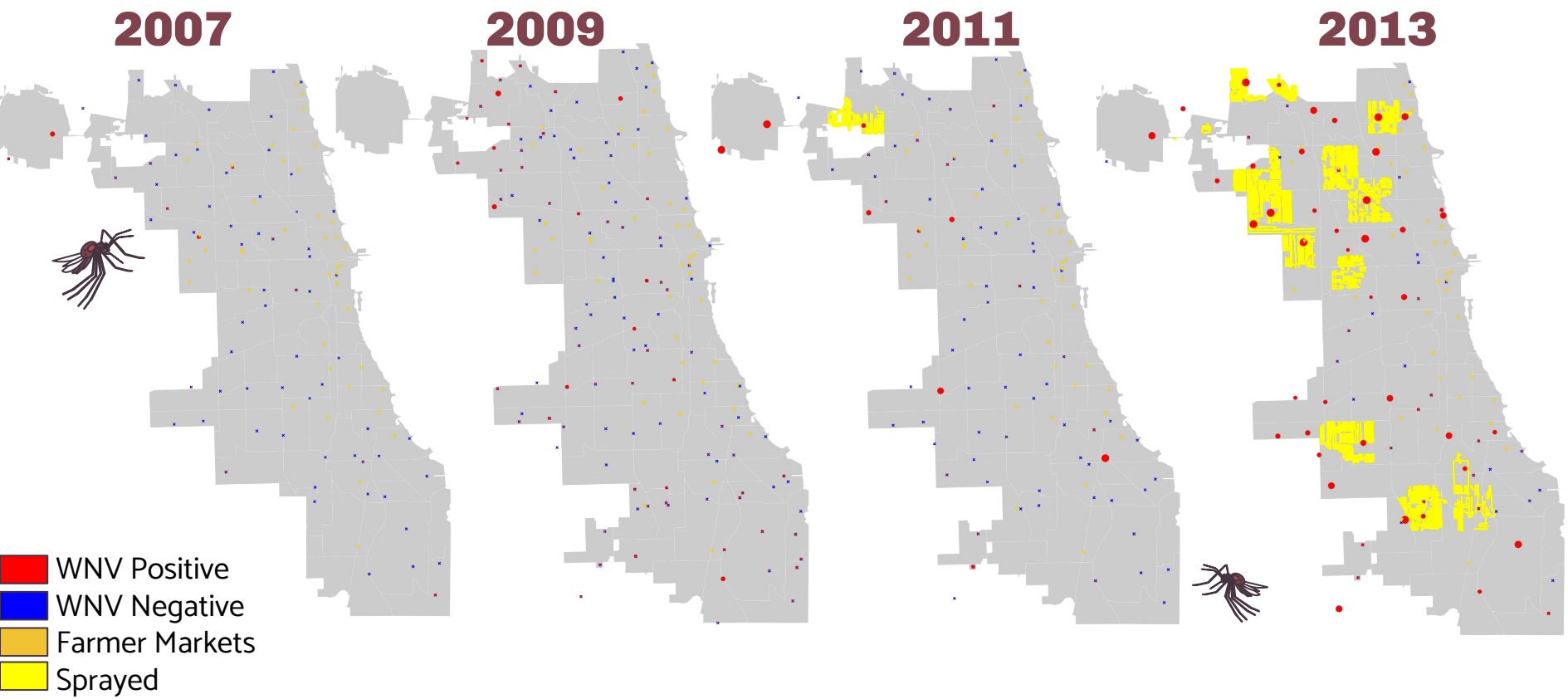
EDA: Mapping of traps and population density

Lack of traps in highest density areas

Lack of traps near farmer markets

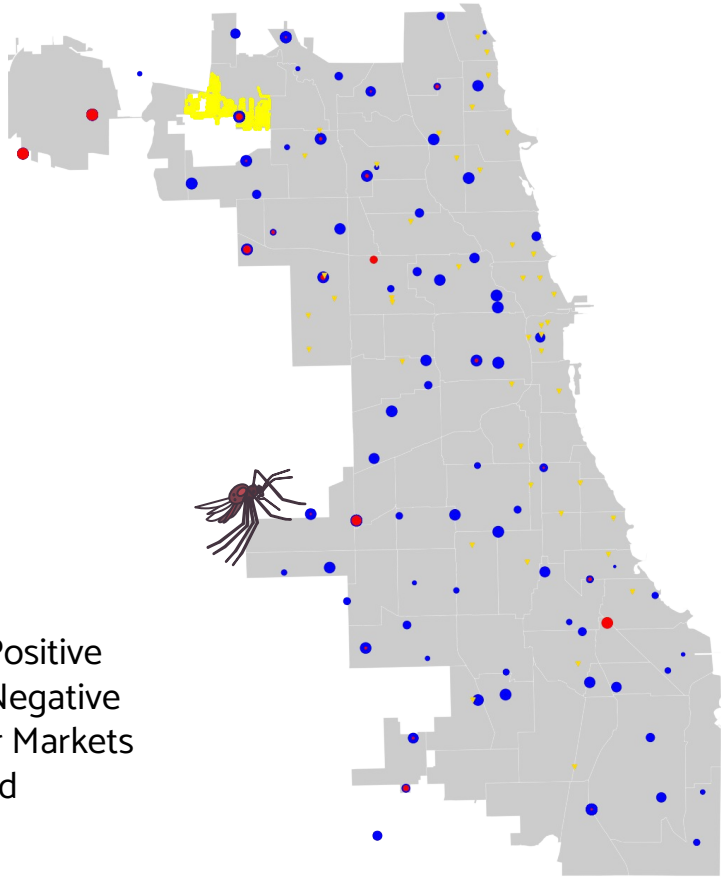


Mapping of WNV Appearance and Spraying

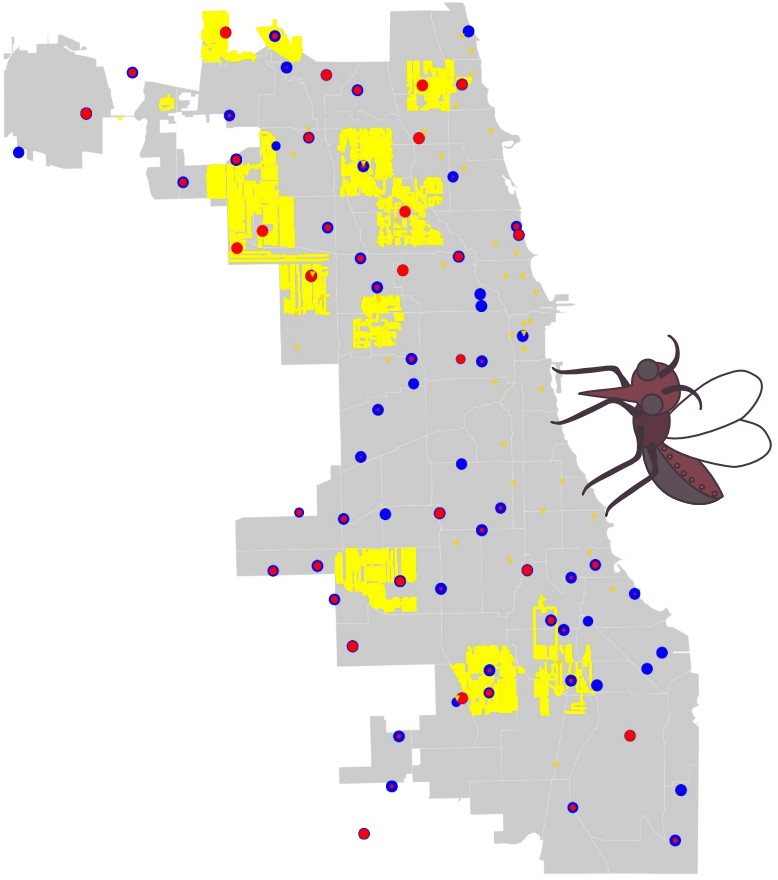


Mapping of Mosquito Appearance and Spraying

2011

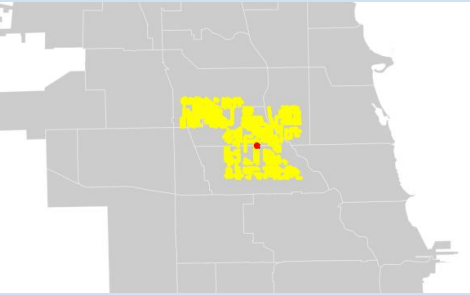


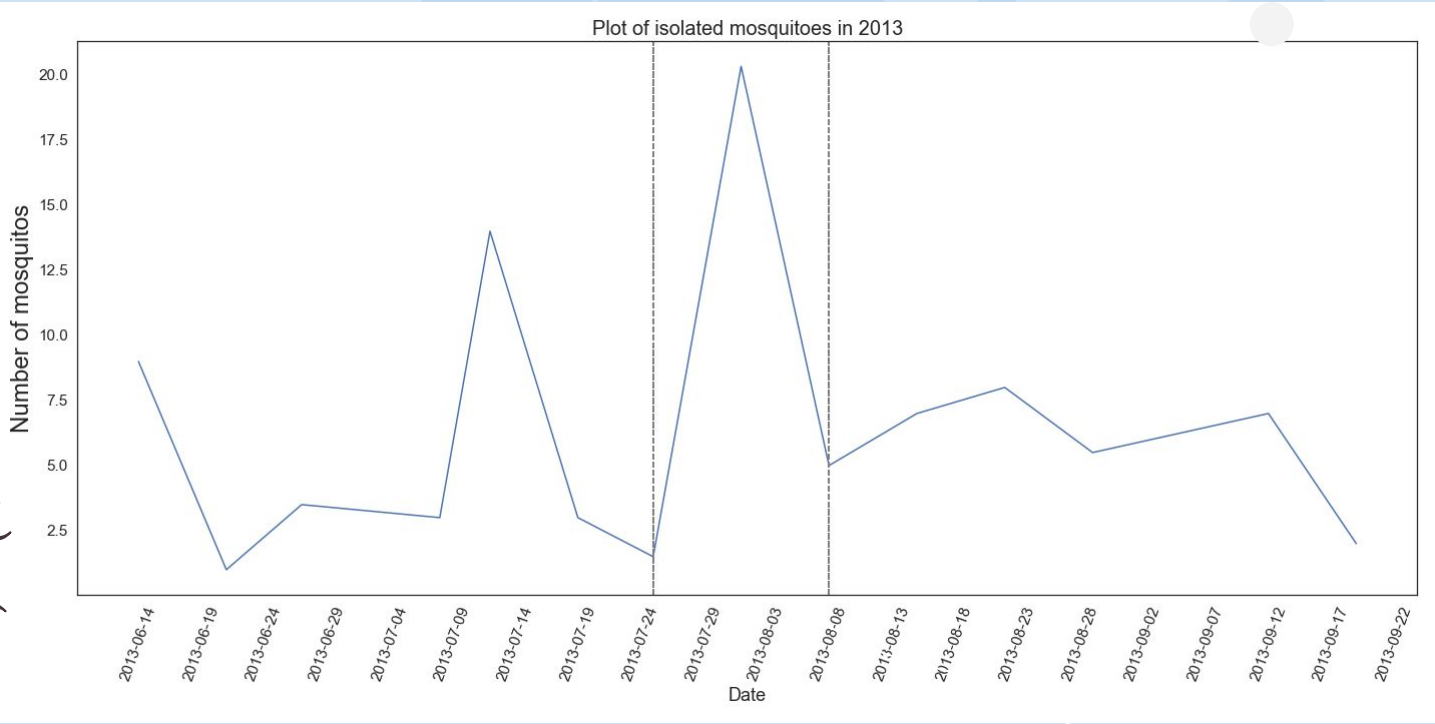
2013

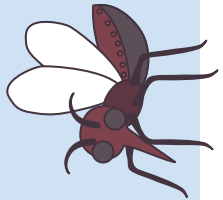


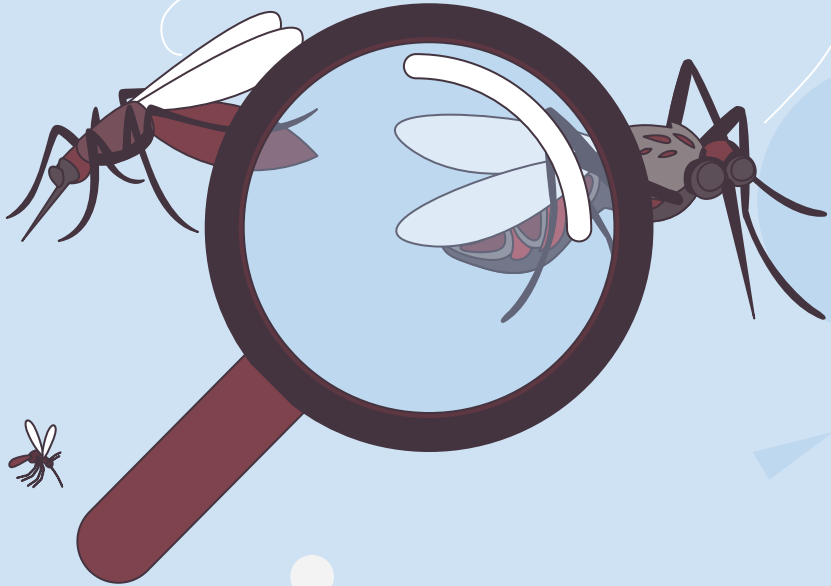
WNV Positive
WNV Negative
Farmer Markets
Sprayed

EDA: Effectiveness of spraying





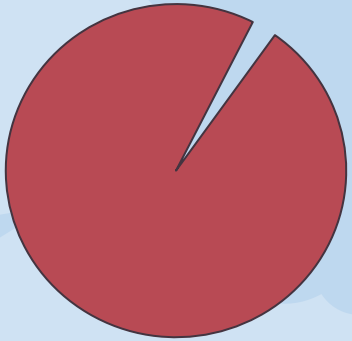




03 Model Evaluation

Baseline

95% Accuracy



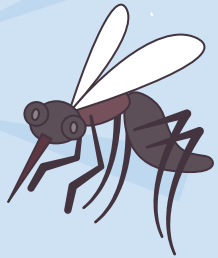
Predict: No West Nile Virus

Project's Purpose..

1. Evaluate data collected
2. Identify places with West Nile Virus
3. Deploy pesticides

..but the Baseline..

1. Ignores collected data
2. Predicts no West Nile Virus
3. No pesticides will be deployed



Accuracy score should <u>not</u> be main metric used for model evaluation

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

- Overview of model's ability to predict both majority and minority classes combined
- Not good for extremely imbalanced datasets
- High Accuracy score due to large number of correct predictions in majority class
- Accuracy score masks model's inability to correctly predict and classify minority class

Test and Train Accuracy to be compared to check for overfitting/underfitting

Area Under Curve (AUC) and F1-score for evaluation

Area Under Curve (AUC)

- Ability to distinguish between Positive (virus present) and Negative (virus not present)
- Good model = High AUC

F1 score

- Weighted average of Precision-Recall score, used to compare the performance of classifiers
- Good model = High F1 score

Models

Random Forest

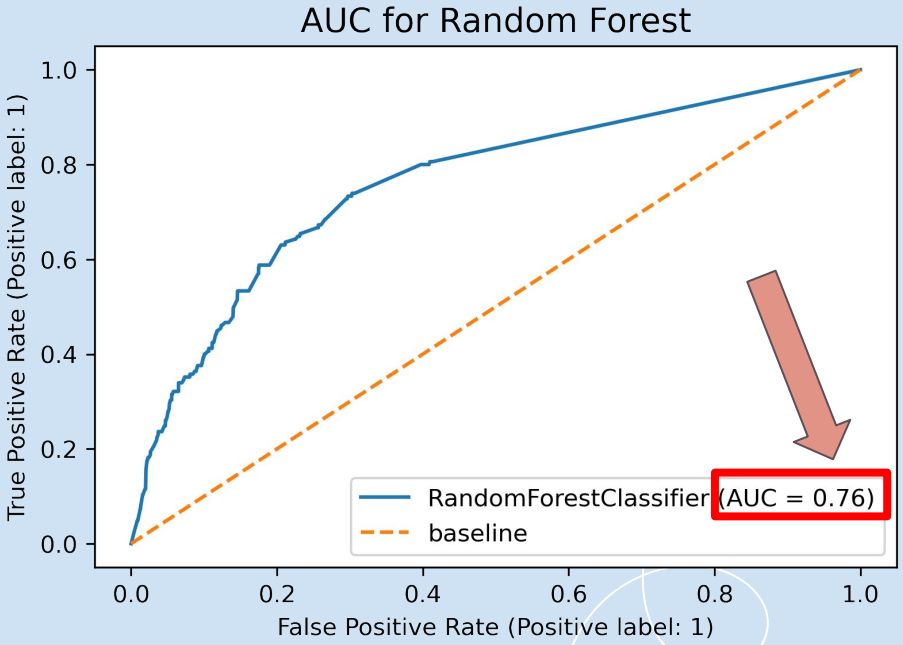
Support Vector Classifier

AdaBoost

Gradient Boosting

Random Forest

Area Under Curve



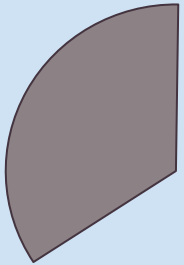
AUC closer to 1

Better at distinguishing between classes

Better Model

Random Forest

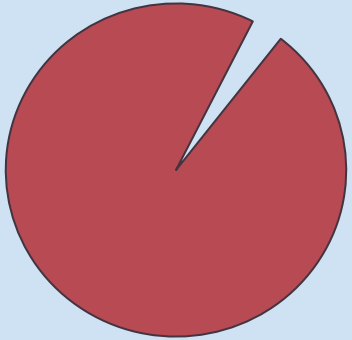
31%



F1-score

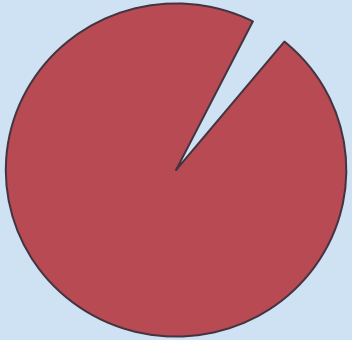
Used to compare classifier performance

95%



Train Accuracy

94%

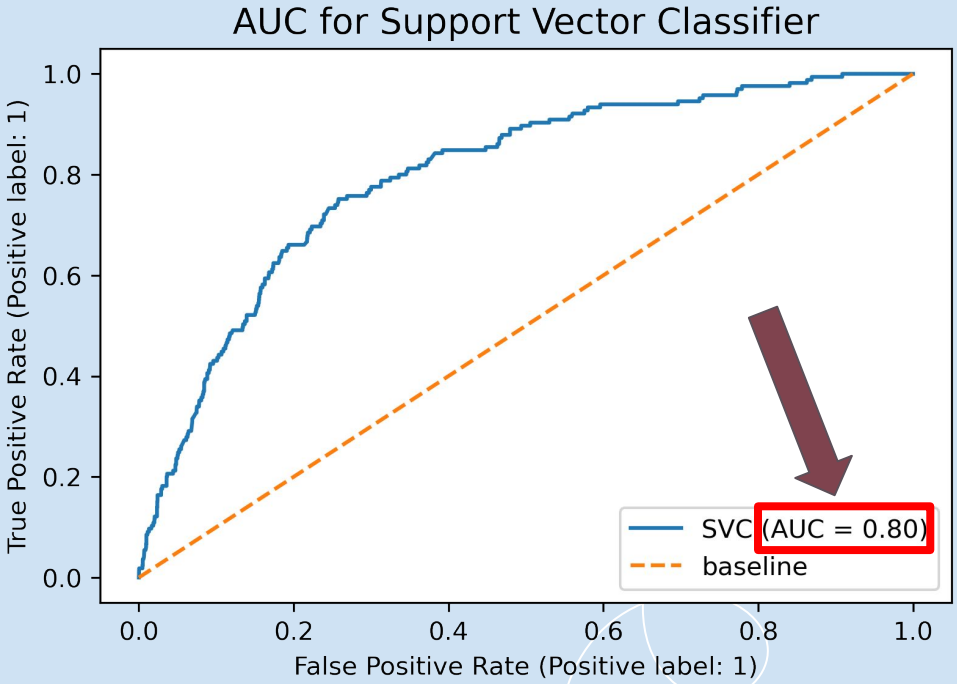


Test Accuracy

Train and Test Accuracy compared to check for overfitting/underfitting

Support Vector Classifier

Area Under Curve



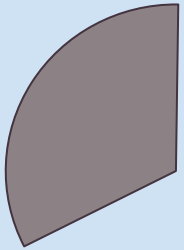
AUC closer to 1

Better at distinguishing between classes

Better Model

Support Vector Classifier

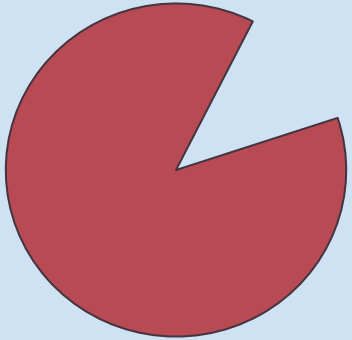
29%



F1-score

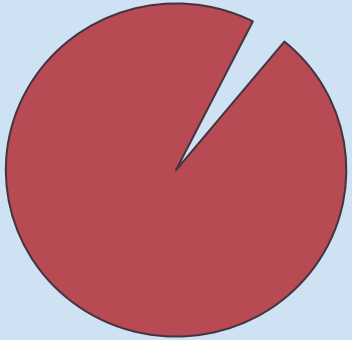
Used to compare classifier performance

83%



Train Accuracy

95%

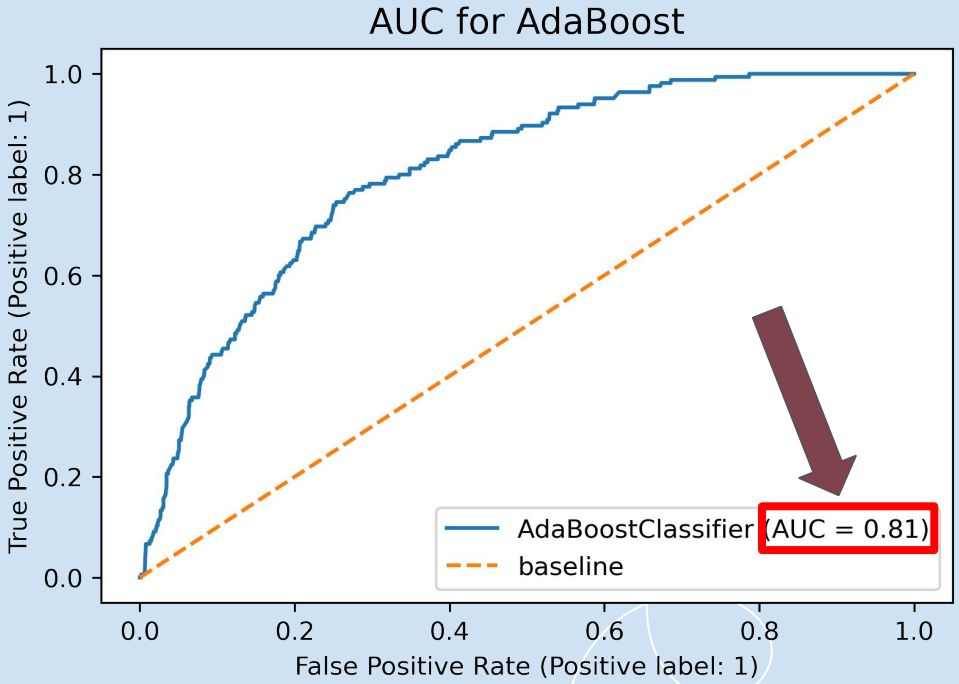


Test Accuracy

Train and Test Accuracy compared to check for overfitting/underfitting

AdaBoost

Area Under Curve



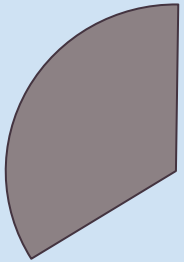
AUC closer to 1

Better at distinguishing between classes

Better Model

AdaBoost

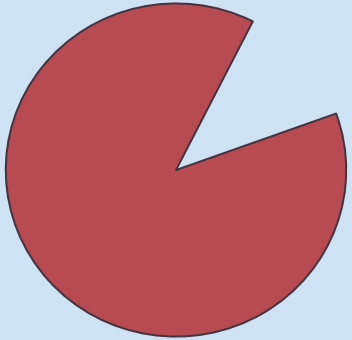
31%



F1-score

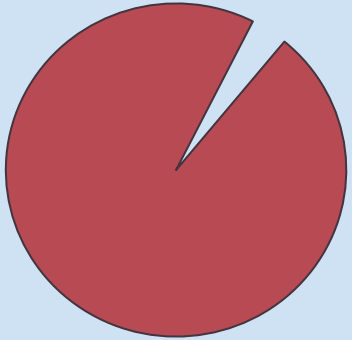
Used to compare classifier performance

86%



Train Accuracy

94%

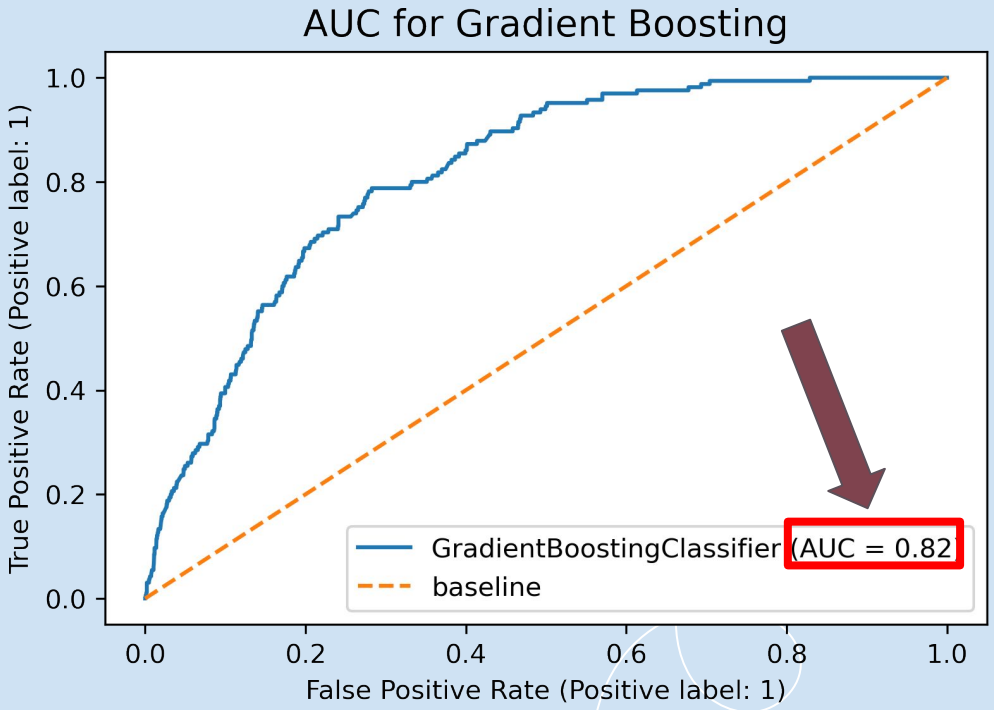


Test Accuracy

Train and Test Accuracy compared to check for overfitting/underfitting

Gradient Boosting

Area Under Curve



AUC closer to 1

↓

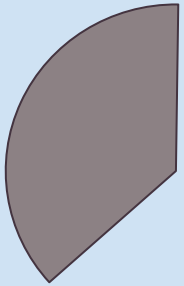
Better at distinguishing between classes

↓

Better Model

Gradient Boosting

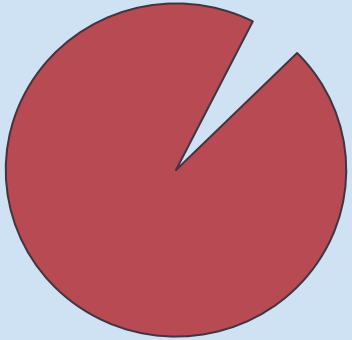
34%



F1-score

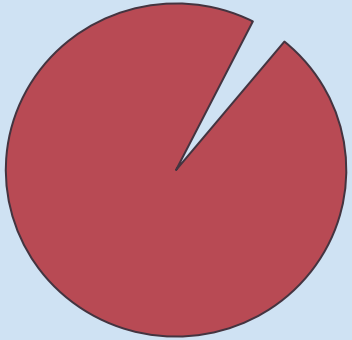
Used to compare classifier performance

90%



Train Accuracy

94%



Test Accuracy

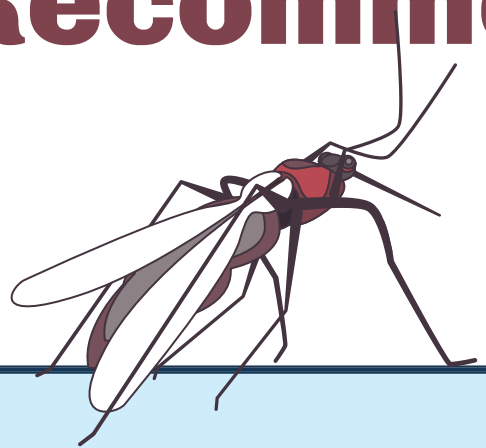
Train and Test Accuracy compared to check for overfitting/underfitting

Models Summary

	AUC	F1-score	Train Accuracy | Test Accuracy
Random Forest	76%	31%	95% | 94%
Support Vector Classifier	80%	29%	83% | 95%
AdaBoost	81%	31%	86% | 94%
Gradient Boosting	**82%**	**34%**	90% | 94%

04

Conclusion & Recommendation



Our goal:

To predict the places in Chicago where the West Nile Virus is present through classification modelling to enable an **effective plan in deploying pesticides spraying** throughout the city.

Cost-Benefit Analysis

Conclusion

Recommen-dation

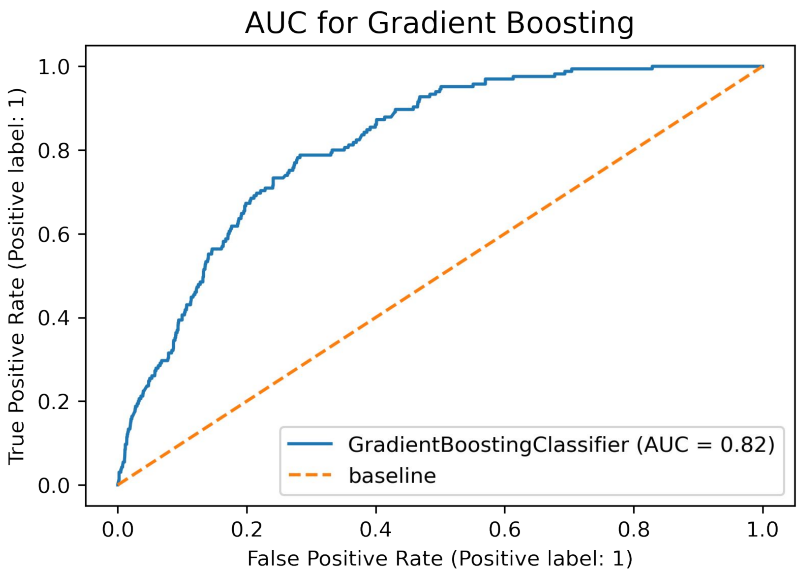
The Best Model

Gradient Boosting is the best model among our classification models tested.

It has the highest test **AUC** score of

0.82

Other performance indicators:
F1 **0.34** | Accuracy **0.94**



AUC – ROC curve is the performance measurement for the classification problems. The higher the AUC, the better the model is at distinguishing between the presence of virus or no virus.

Cost Projection

Based on the confusion matrix of our model, we will take the percentage of **TP, FN, and FP** to assume the presence of WNV. This is to ensure all actual positives and predicted positives are covered.

Test set size = 3152
TP+FN+FP = 618 (**20%**) *see appendix*

Taking the total number of traps set up in 2013 ~80, we estimated that we only need to target **17 traps with radius of 2 km each, which will cover about an area of 12.56 km²***,
with overlapped coverage if traps are close to each other.

Legend: TP (True Positive), FN (False Negative), FP (False Positive)
*Note: *According to CDC, Culex mosquitoes can fly up to 3km²*

Cost Analysis

(spraying with spray truck)

Pesticide: Zenivex E4 *(minimum risk to human and non-targeted species)*

Cost of Pesticide: *USD 227.24 per km²*

Cost of full coverage spraying

Entire city of Chicago: *606.10 km²*
~ USD **137, 730** *each time,*
~ USD **1,377,300** *twice a month for 5mths*

Cost of targeted spraying

Targeted area of Chicago: *213.5 km²*
~ USD **48, 516** *each time,*
~ USD **485,160** *twice a month for 5mths in a year*

Benefit Projection

If the West Nile Virus is eliminated, or significantly reduced with effective approaches, this will certainly give the residents a peace of mind.

Quality of life

- Fewer people falling ill or dying
- Higher productivity at work

Risk of monetary loss

- Hospital bill can be avoided, ~ $25K per patient
- Loss of income or labour during recovery ~ $1.5K/patient* can be saved

20 persons' monetary loss > cost of spray $485K/year

Eliminate fear of body damage

- Prolonged fatigue
- Severe neuro-invasive diseases, such as meningitis

(Peterson, 2019)

*Note: *per capita income is $37K, if loss of income for 2 weeks will cause $1.5K loss to a personal or to a business (https://www.census.gov/quickfacts/fact/table/chicagocityillinois/LND110210)*

Conclusion

As shown in our cost-benefit analysis, the **cost per infection** and the overall well being of the city outweigh the total **cost of spray**. **Spraying is necessary**, it instills confidence and a peace of mind, at the minimum..

Should there be an outbreak of virus, people would be subjected to **mental health illness**, or even **death**, as there is no effective medication.

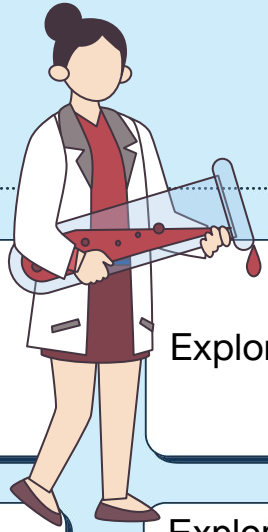
Based on our EDA, it shows that the current spraying schedule was not very effective:

- The sprays were done infrequently and during the window of which less favorable for decreasing the breeding of mosquitoes.

- We need more details to understand how the spraying was carried out.

The data used for the modelling is grossly imbalanced which may result inaccurate prediction of the virus.

Recommendation



Further **hyperparameters tuning** of the model.

More data is required, preferably a more balanced dataset.

Explore other **modelling techniques**

Lavriciding technique

Uses biological pesticides, chemicals or fish to kill mosquito larvae in waters before they hatch

Explore **Wolbachia** (Singapore) Mosquito Suppression Strategy (with care)

When male *Wolbachia*-carrying *Aedes aegypti* (*Wolbachia-Aedes*) mosquitoes mate with urban female *Aedes aegypti* that do not carry *Wolbachia*, their resulting eggs do not hatch.

More traps to be placed near **Farmer Market**

Areas where birds and humans would gather, thereby drawing mosquitoes, especially in the summer.

Thanks

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

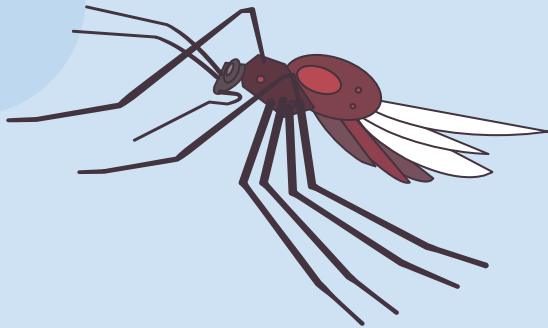
Data Cleaning & Preparation

Train	Dataset from Kaggle and was exceptionally clean	Converted date to datetime for further processing
Spray	Dataset from Kaggle had missing time values	Removed as it only affected 4% of dataset
Weather	Dataset from Kaggle had no null values but had imputed values for null	Converted date to datetime for further processing Cleaned wrongly imputed values

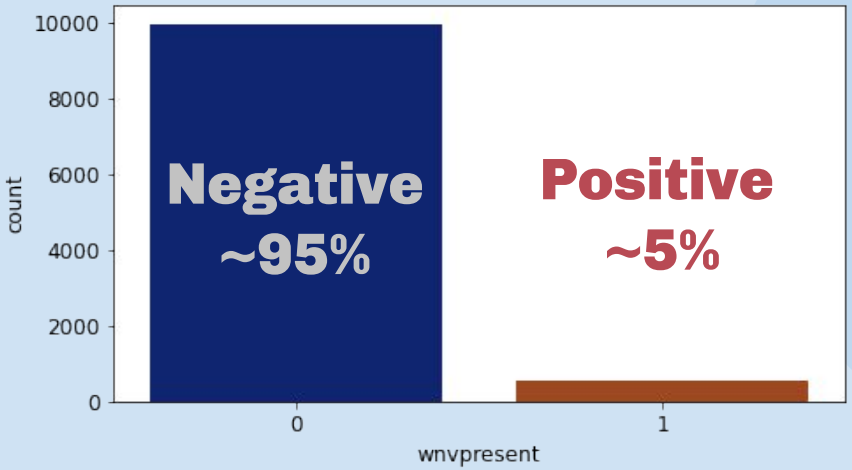
Engineered

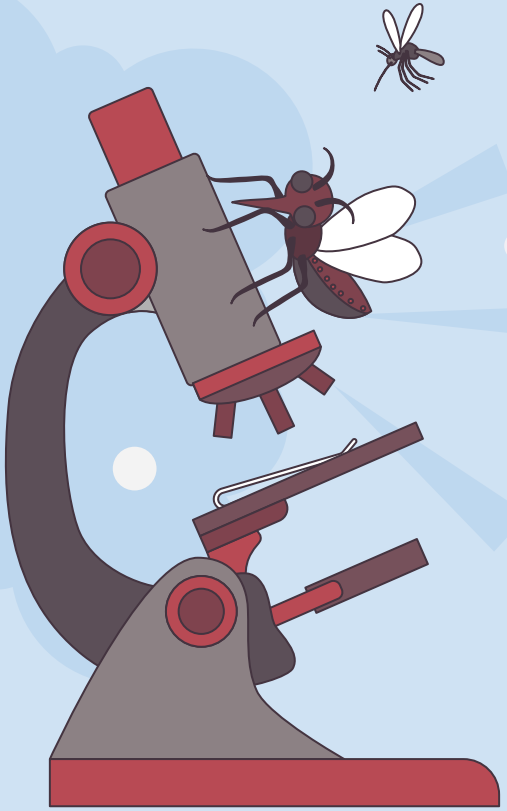
Codesum: As not all weather conditions were available, we grouped the remainder into wet or dry variables

Date: Created YearMonth and YearWeek, to see if the months and weeks had any correlation with the virus



EDA: WNV Presence





Precision-Recall

Precision

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Ratio of True Positives to All Predicted Positives
- High Precision = More number of places predicted as virus-present actually does indeed have the virus

Recall

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- Ratio of True Positives to All Actual Positives
- High Recall = More number of places that actually have the virus present is correctly classified by model

Confusion Matrix Of Gradient Boosting Model