# Project 4
# West Nile Virus Prediction

Clare, Ian, Qingyi, Wan Xian

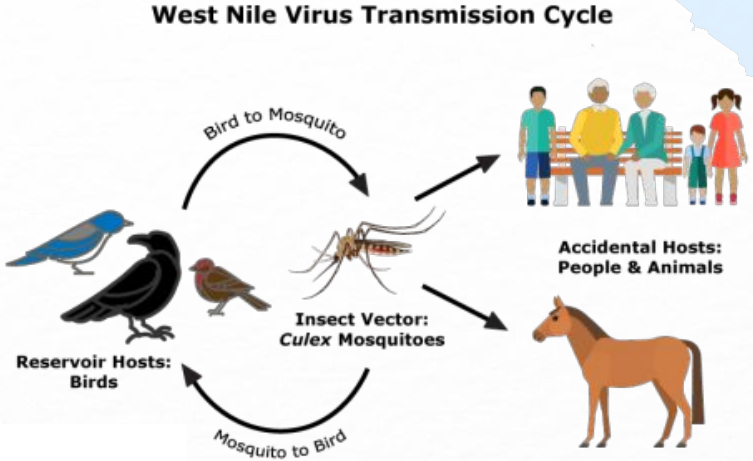
# Pesky Problem

According to the Centers for Disease Control and Prevention (CDC), West Nile virus (WNV) is the leading cause of mosquito-borne disease in the United States. Cases of WNV usually occurs from the start of summer until the end of autumn. Since 2004, the Public Health department of Chicago has been running a surveillance and control program on the trends of WNV cases and the population of Culex Mosquito in the city. As part of the mosquito control plan, pesticides are sprayed in areas of the city with a high influx of WNV cases.

# Understanding the Problem

## West Nile Virus

- Spread to humans through infected mosquitos
- 80% of infections do not develop symptoms
- 20% of infections develop fever with other symptoms such as headache, body aches, joint pains, vomiting, diarrhea or a rash
- 0.67% of infections develop a severe illness that affects the central nervous system, which may lead to death
- No vaccine or medication available to treat WNV



# Understanding the Problem

## Chicago

- City in the state of Illinois
- Population: 2.7 million (Singapore: 5.7 million)
- Area of approx. 600km$^2$ (Singapore: 728km$^2$)
- WNV was first identified in Sept 2001, in 2 dead crows
- First human case was reported in 2002
- In 2006, mosquito control efforts were put in place as an elevated risk of WNV infection was identified

# Our Mission

We are a team of data scientist from the Disease and Treatment Agency.

Our mission is to:

- Develop a machine learning (ML) classification model that predicts the probability of the presence of WNV. With a good prediction model, the agency will be able to allocate resources efficiently and arrange insecticide spraying in high transmission areas.
- Do a cost-benefit analysis of insecticide spraying in previous years. With a good analysis, the agency will be able to convince stakeholders to retain the insecticides spraying regime as part of its mosquito control plan.

# Scope

01 Data Cleaning & EDA

02 Modeling & Evaluation

03 Cost Benefit Analysis

04 Insights & Recommendations

# 01

# Data Cleaning & EDA

# Data Description

| Dataset | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|
| Train | ✓ | | ✓ | | ✓ | | ✓ | |
| Test | | ✓ | | ✓ | | ✓ | | ✓ |
| Weather | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Spray | | | | | ✓ | | ✓ | |

# Data Cleaning

## Train/Test

- Change ‘Date’ datetype to datetime

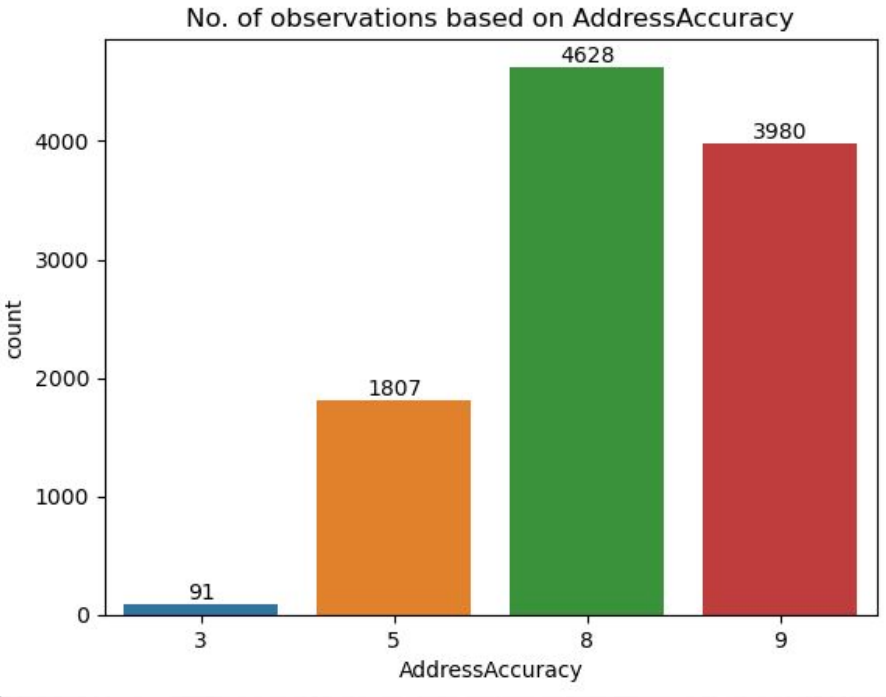
## Spray

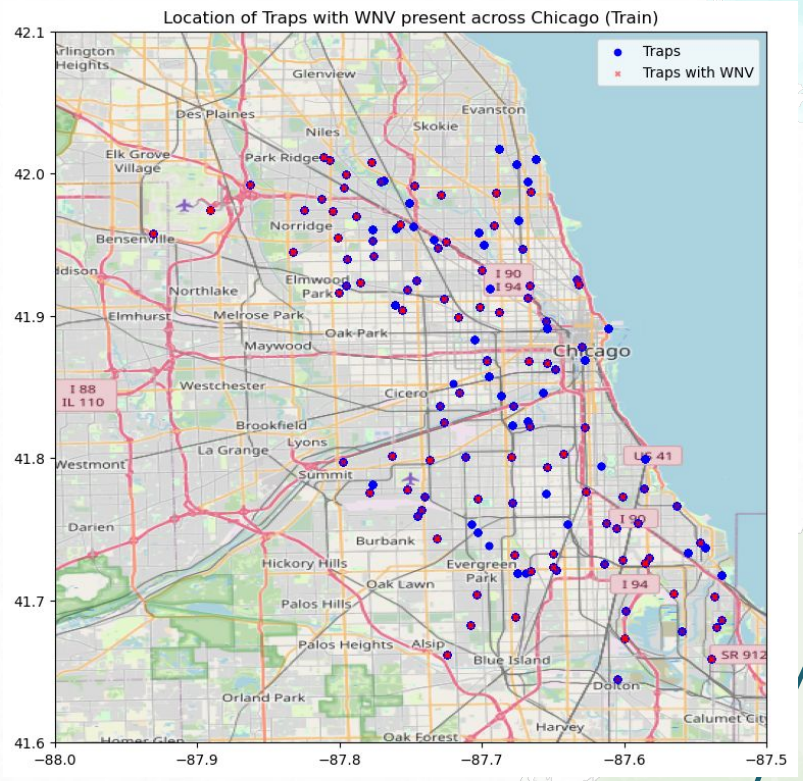
- Drop rows with null value
- Drop duplicates

## Weather

- Change ‘Date’ datetype to datetime
- Drop ‘Water1’ & ‘CodeSum’ (50% missing values)
- Fill null values
  - Impute with mean
  - Impute with values from Station 1
- Average weather statistics between weather stations 1 and 2

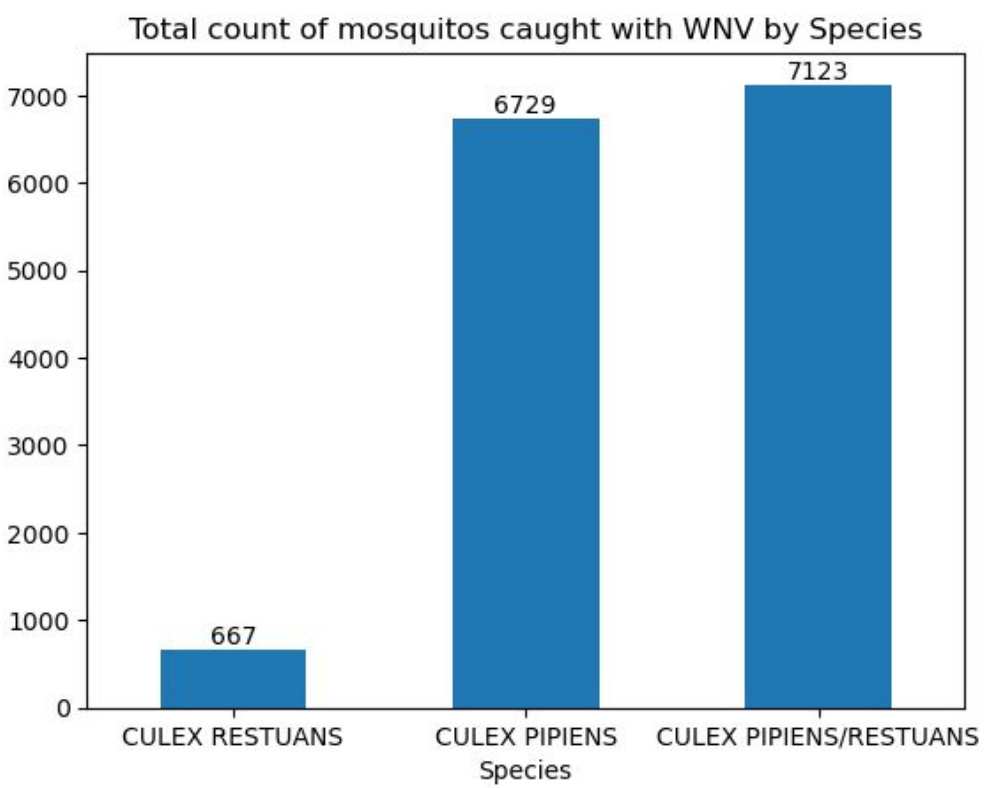
# EDA - Train/Test Dataset



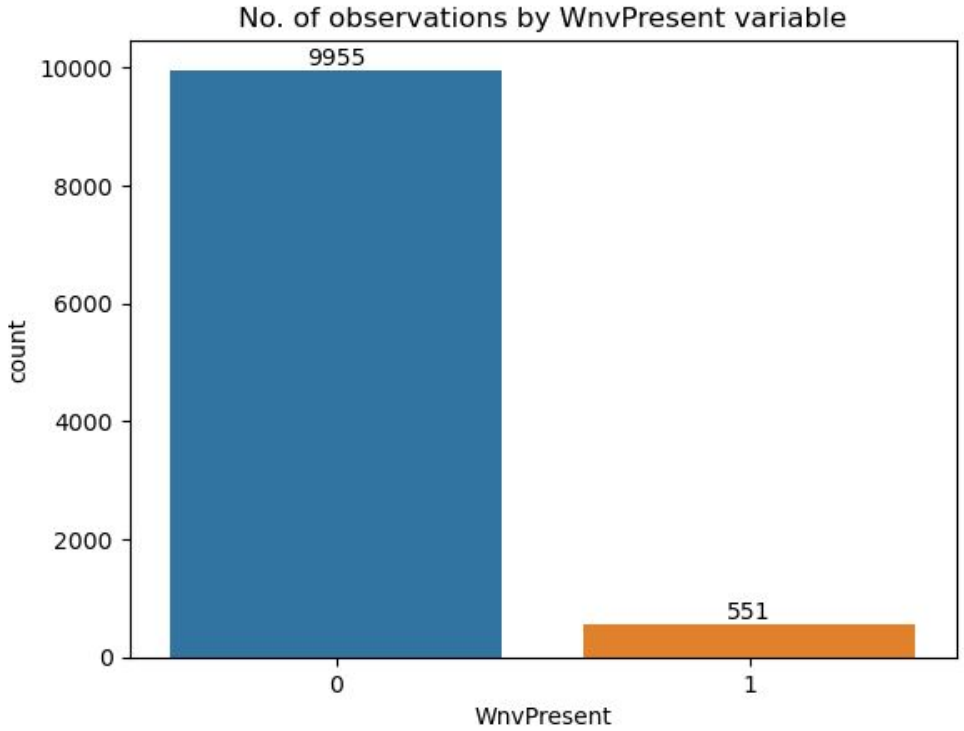



# EDA - Train Dataset

1) Culex Pipiens/Restuans
2) Culex Pipiens
3) Culex Restuans
4) Culex Territans
5) Culex Salinarius
6) Culex Tarsalis
7) Culex Erraticus
8) Unspecified Culex

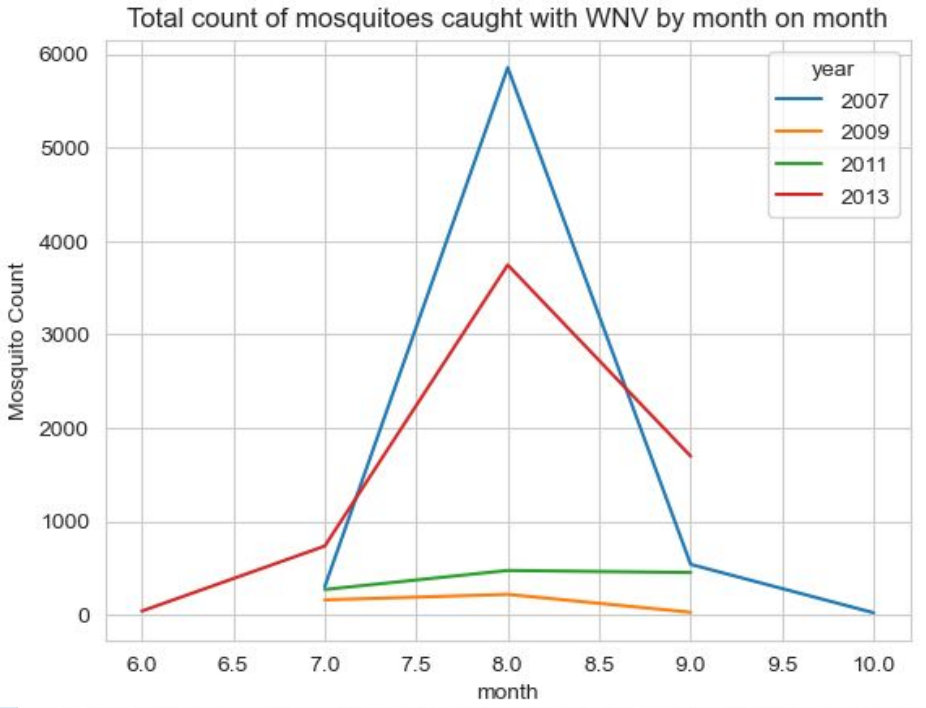


# EDA - Train Dataset




- Imbalance dataset
  - 5% of records are WNV Present
  - 95% of records are WNV Absent

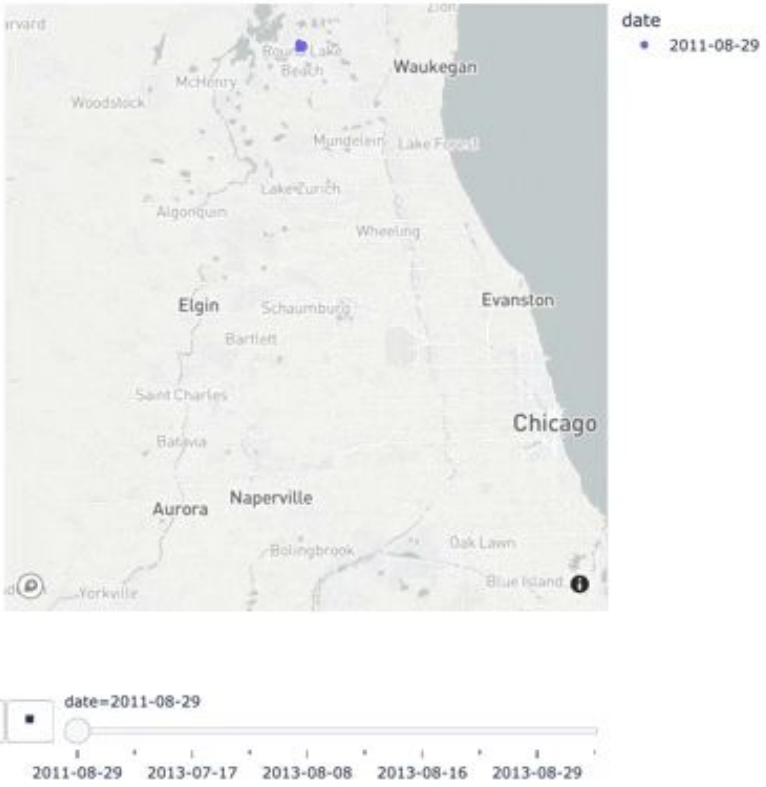
# EDA - Train Dataset



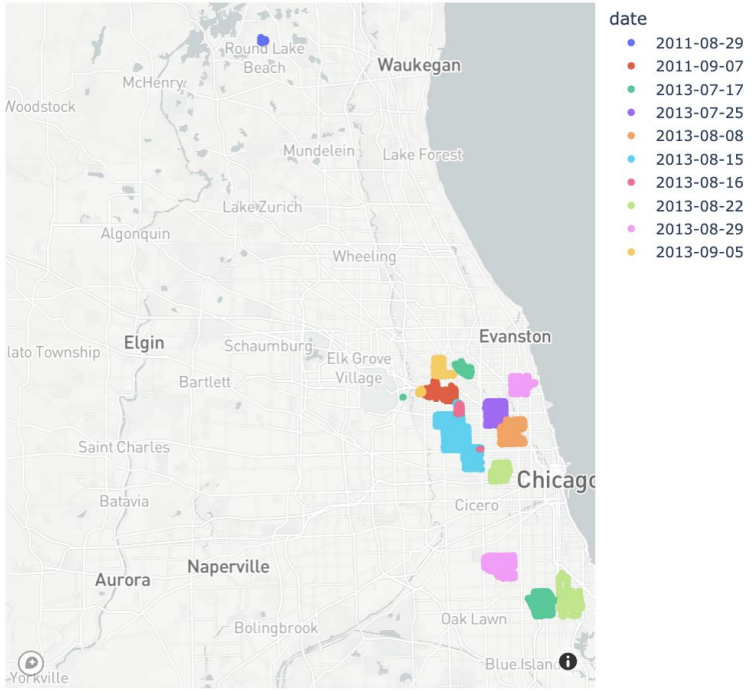

- Peak month for mosquitoes with WNV is August
- For year 2009 & 2011, the general population of WNV vector mosquitoes are much lower than that of 2007
- However, the figures spike upwards in 2013

# EDA - Spray Dataset

Timeline of sprayed locations




Locations sprayed on each date



# 02

# Modeling & Evaluations

# Models

Train Data is extremely imbalanced

- About 5% records are present with WNV
- Remaining 95% records are absent of WNV

Methodology

- Resampling methods
  - SMOTE
  - Random Over Sampling/Random Under Sampling
- Model algorithms
  - Naive Bayes
  - Random Forest
  - k-Nearest Neighbors
  - Gradient Boosting

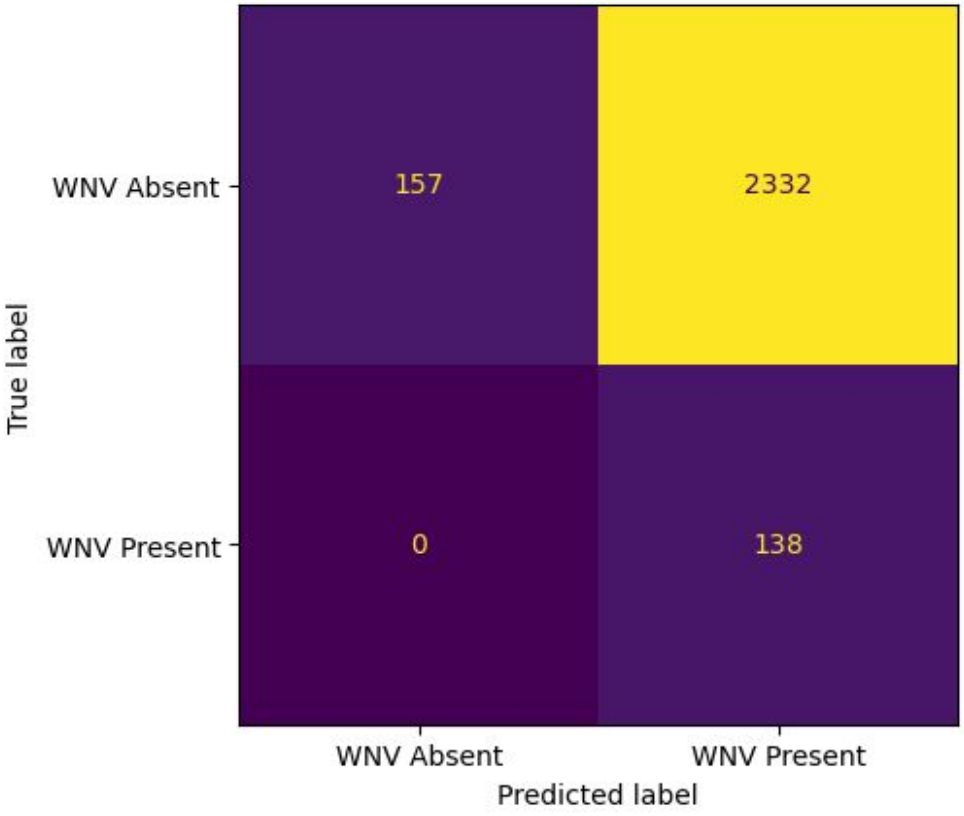
Key Evaluation Metrics

- F1 score
- ROC AUC score

# Models

| Model Algorithm | Resampling method |
|---|---|
| Naive Bayes (Baseline) | Random Over Sampler |
| Random Forest | SMOTE |
| Random Forest | Random Under Sampler |
| k-Nearest Neighbors | SMOTE |
| k-Nearest Neighbors | Random Under Sampler |
| Gradient Boosting | SMOTE |
| Gradient Boosting | Random Under Sampler |

# Baseline - Naive Bayes + ROS




| | |
|---|---|
| F1 Score (Train) | 0.10 |
| F1 Score (Test) | 0.11 |
| ROC AUC | 0.75 |
| Precision | 0.06 |
| Recall | 1.00 |
| Average Precision | 0.11 |

Baseline performed fairly

- Predicted a huge proportion of traps as WNV Present

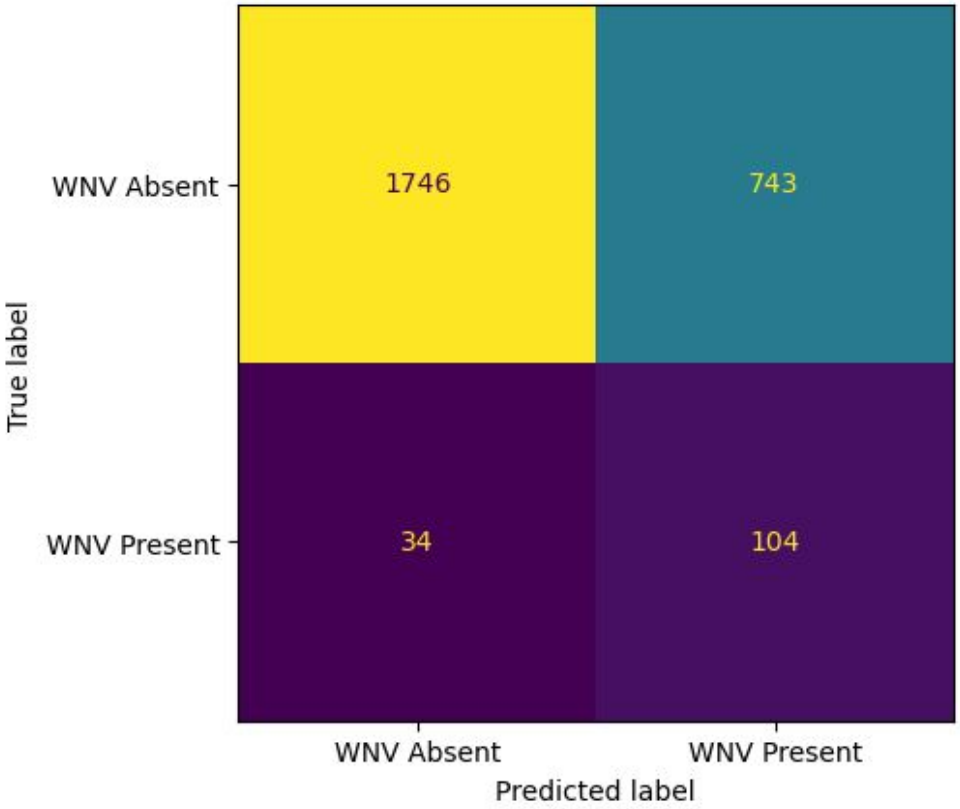
# Model Metrics so far

| Model | Resampling method | F1 (Train) | F1 (Test/Hold-Out) | ROC AUC | Precision | Recall | Average Precision |
|---|---|---|---|---|---|---|---|
| Naive Bayes (Baseline) | Random Over Sampler | 0.10 | 0.11 | 0.75 | 0.06 | 1.00 | 0.11 |
| Random Forest | SMOTE | 0.27 | 0.24 | 0.80 | 0.17 | 0.40 | 0.20 |
| Random Forest | Random Under Sampler | 0.22 | 0.20 | 0.80 | 0.12 | 0.71 | 0.20 |
| kNN | SMOTE | 0.22 | 0.21 | 0.73 | 0.14 | 0.41 | 0.12 |
| kNN | Random Under Sampler | 0.18 | 0.18 | 0.74 | 0.10 | 0.68 | 0.11 |
| Gradient Boosting | SMOTE | 0.30 | 0.26 | 0.81 | 0.19 | 0.41 | 0.21 |
| **Gradient Boosting** | **Random Under Sampler** | **0.22** | **0.21** | **0.80** | 0.12 | 0.75 | 0.19 |

Gradient Boosting + Random Under Sampler

- Small difference in F1 scores
- High ROC AUC score

# Chosen - Gradient Boosting + RUS




| | |
|---|---|
| F1 Score (Train) | 0.22 |
| F1 Score (Test) | 0.21 |
| ROC AUC | 0.80 |
| Precision | 0.12 |
| Recall | 0.75 |
| Average Precision | 0.19 |

Chosen model performed well

- Predicted 75% of traps with WNV correctly
- Predicted 29% of traps without WNV wrongly

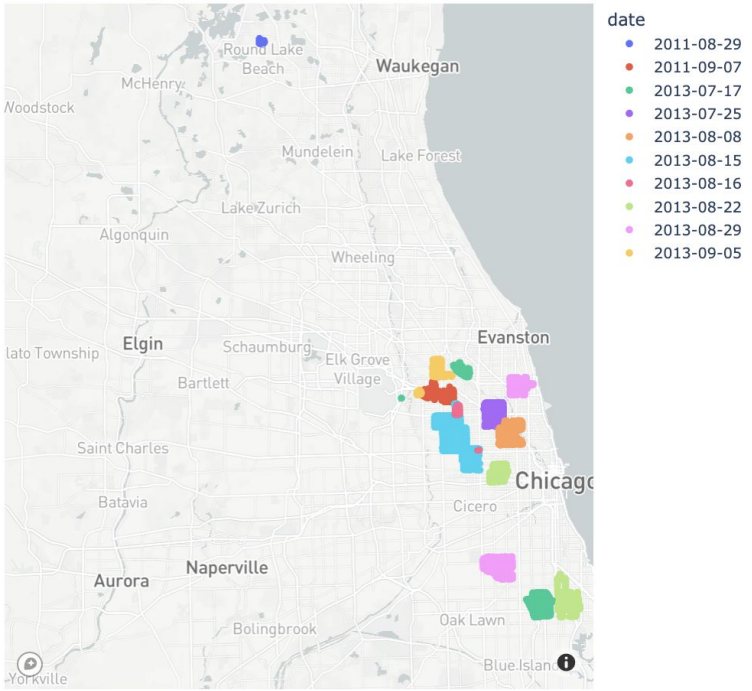
# 03

# Cost Benefit Analysis

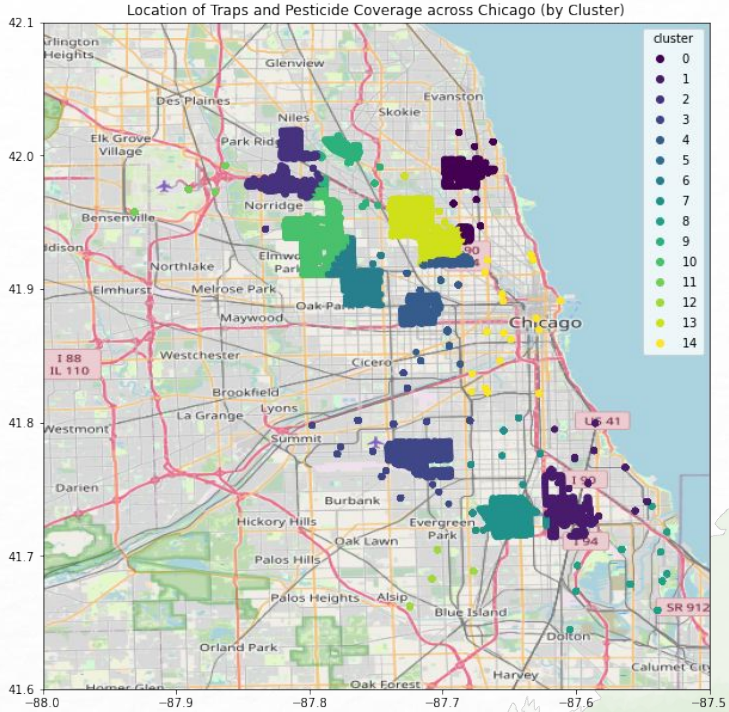
# Spraying Locations

## By Dates

Locations sprayed on each date

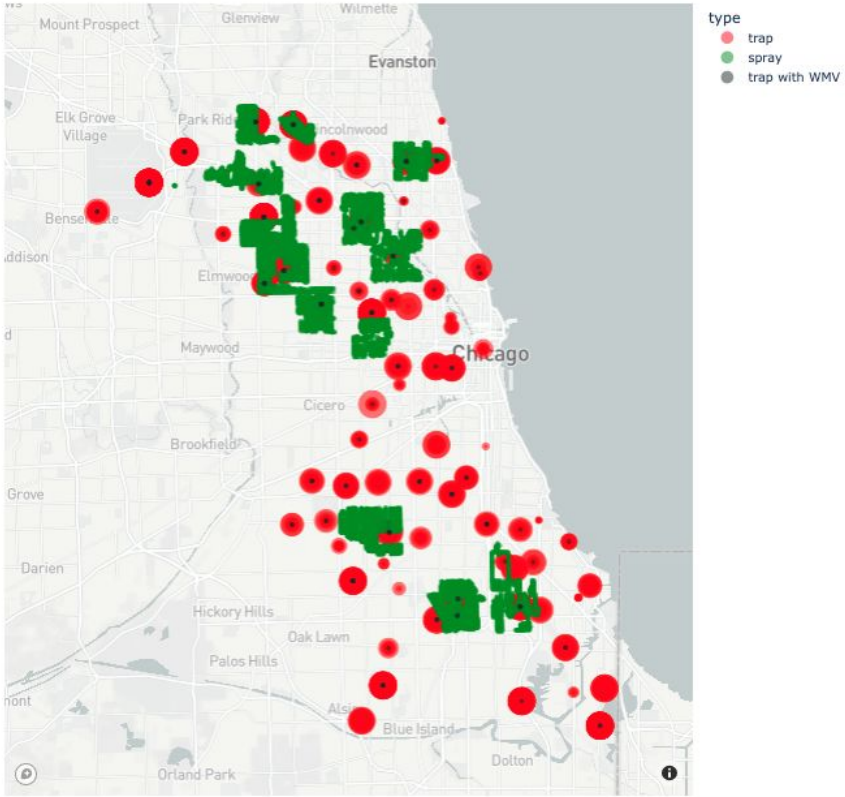



## By Clusters



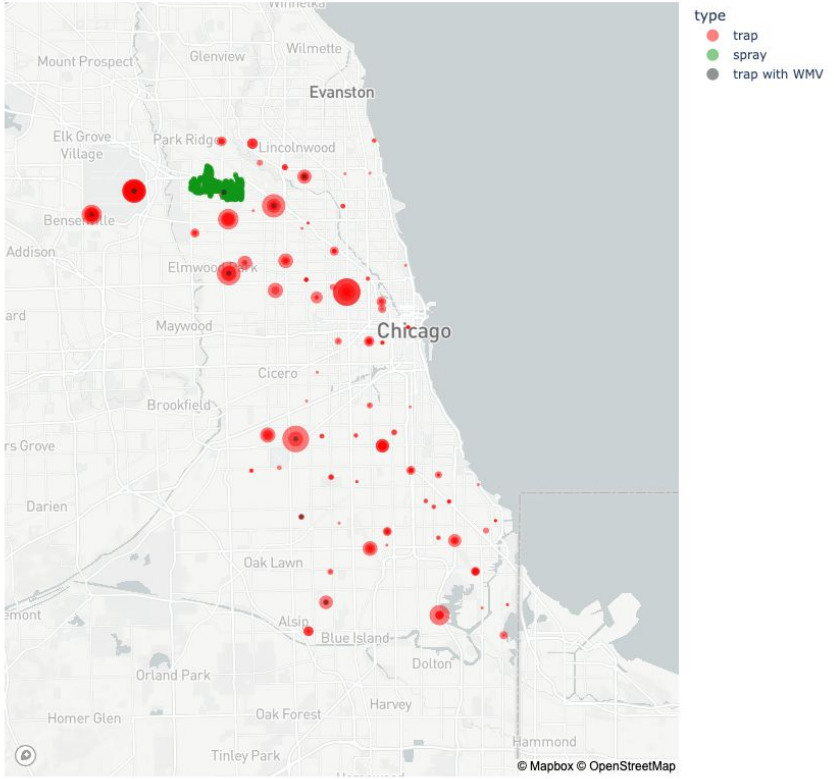
# Pesticide Effectiveness (by Dates)

West Nile Virus In Chicago: 29-08-2011 to 05-09-2013

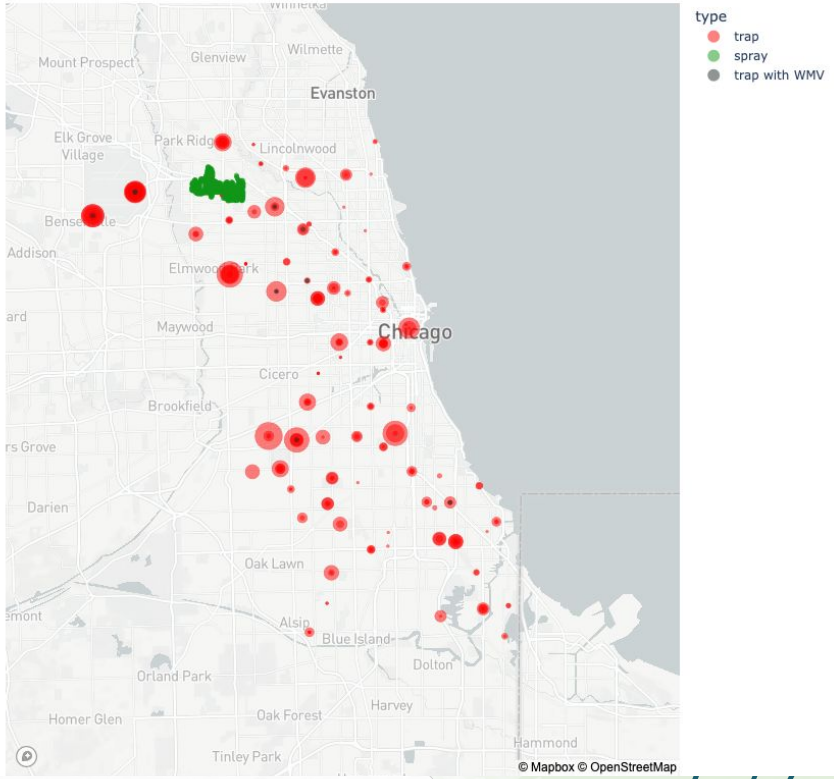


# Pesticide Effectiveness (by Dates)

West Nile Virus In Chicago: 2011-08-24 (2 weeks before 2011-09-07 spray)

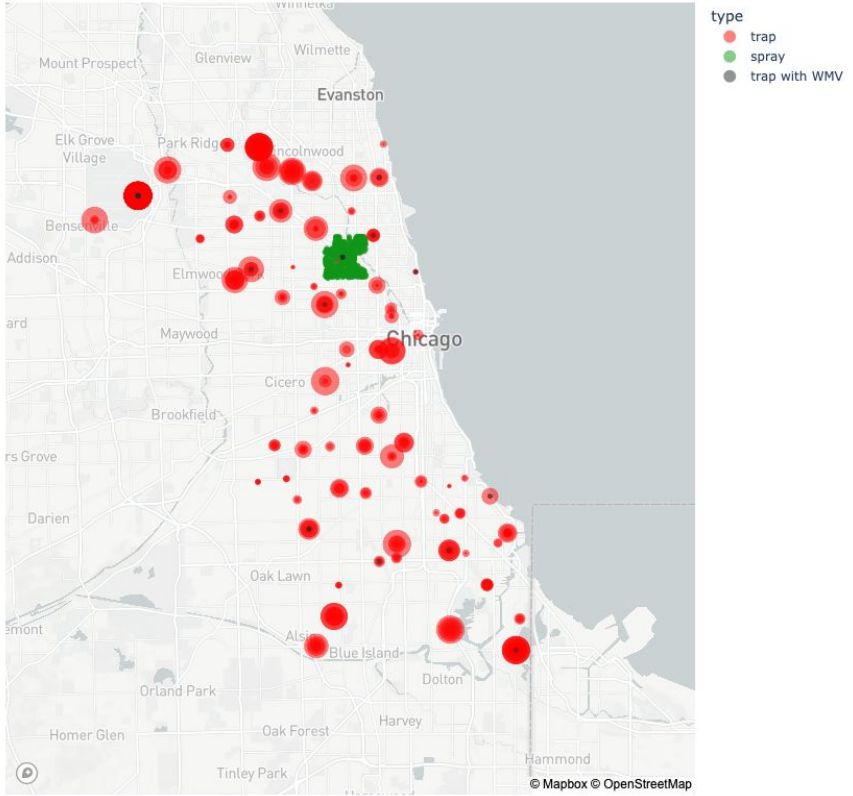



West Nile Virus In Chicago: 2011-09-21 (2 weeks after 2011-09-07 spray)

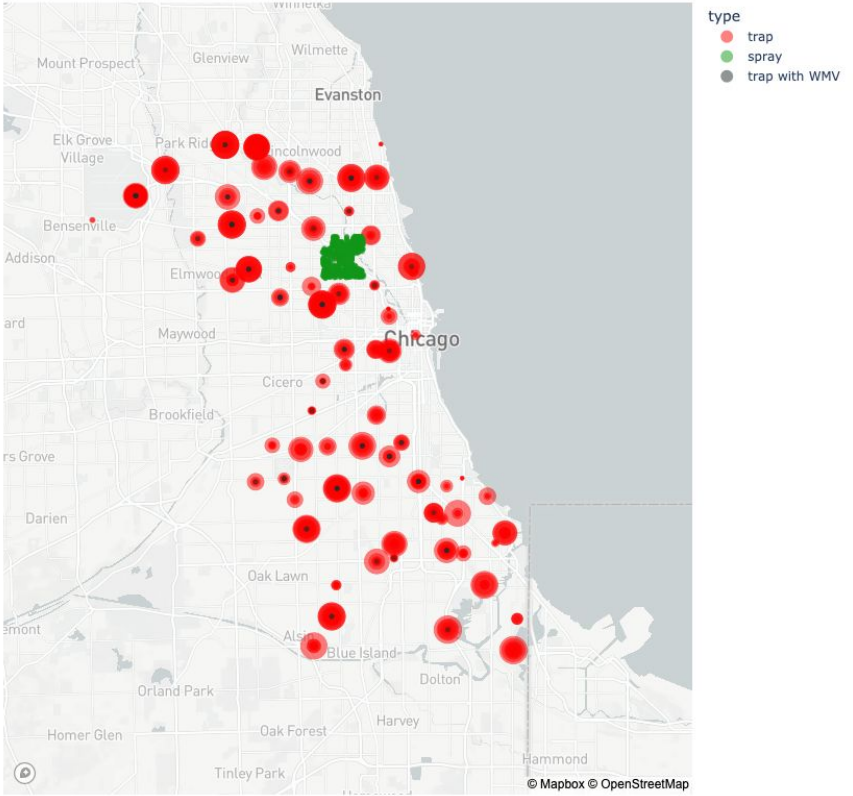


# Pesticide Effectiveness (by Dates)

West Nile Virus In Chicago: 2013-07-25 (2 weeks before 2013-08-08 spray)

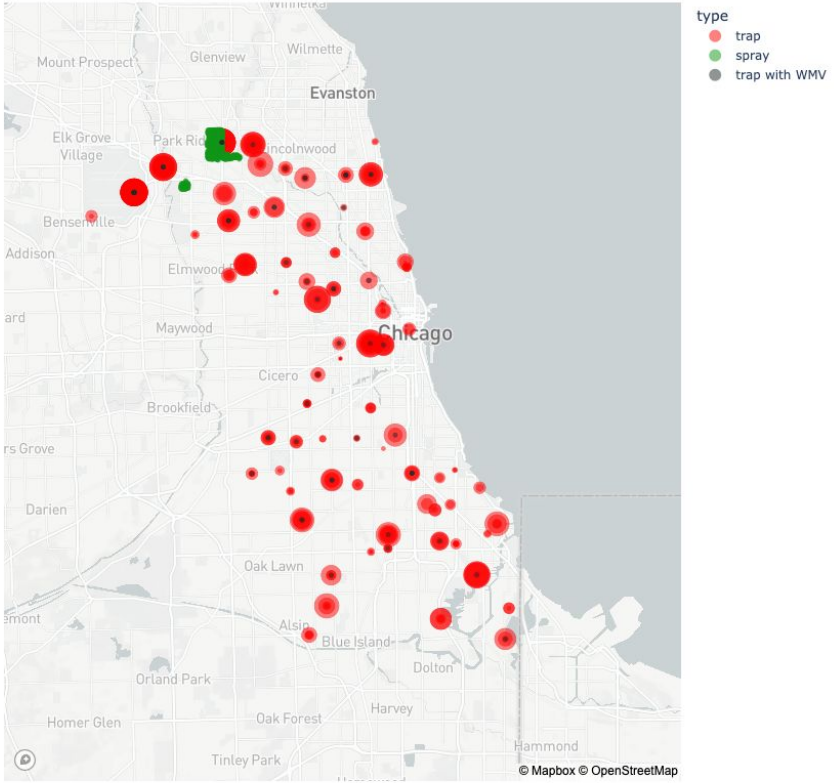



West Nile Virus In Chicago: 2013-08-22 (2 weeks after 2013-08-08 spray)

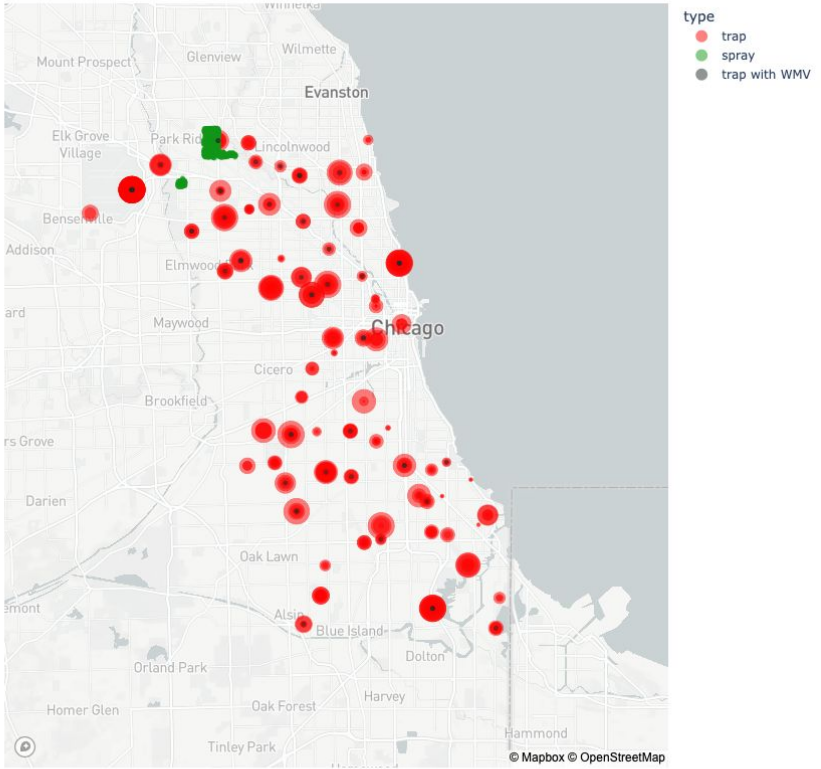


# Pesticide Effectiveness (by Dates)

West Nile Virus In Chicago: 2013-08-22 (2 weeks before 2013-09-05 spray)

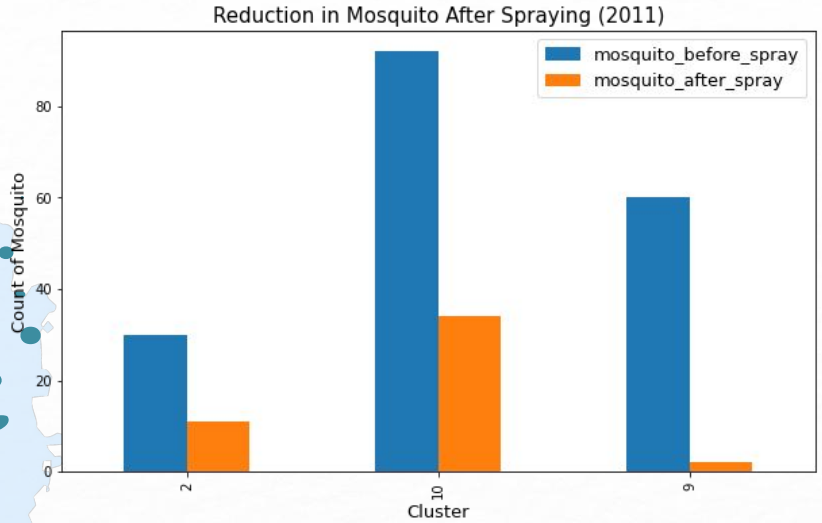



West Nile Virus In Chicago: 2013-09-19 (2 weeks after 2013-09-05 spray)

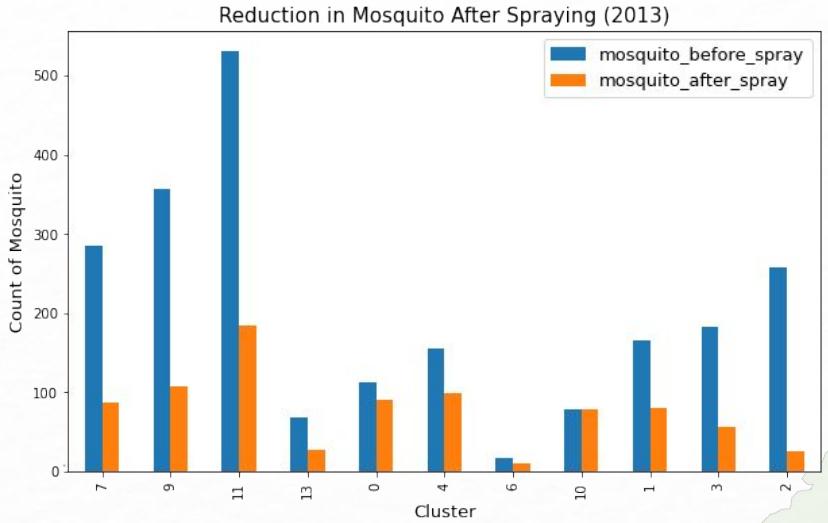


# Pesticide Effectiveness (by Cluster)

## 2011




## 2013

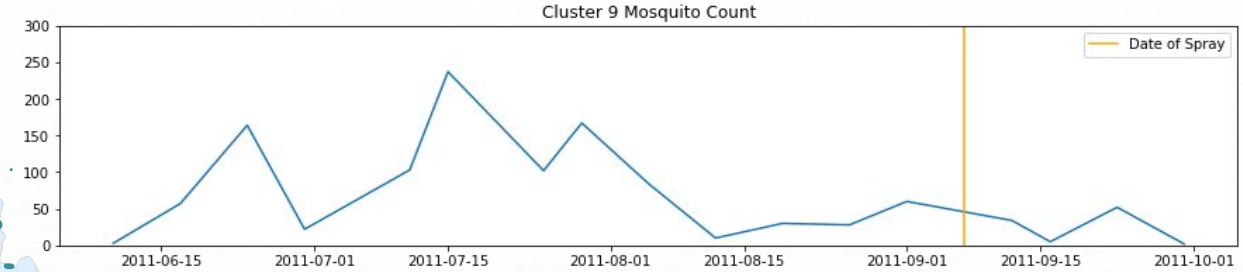



Overall, effective in reducing number of mosquito

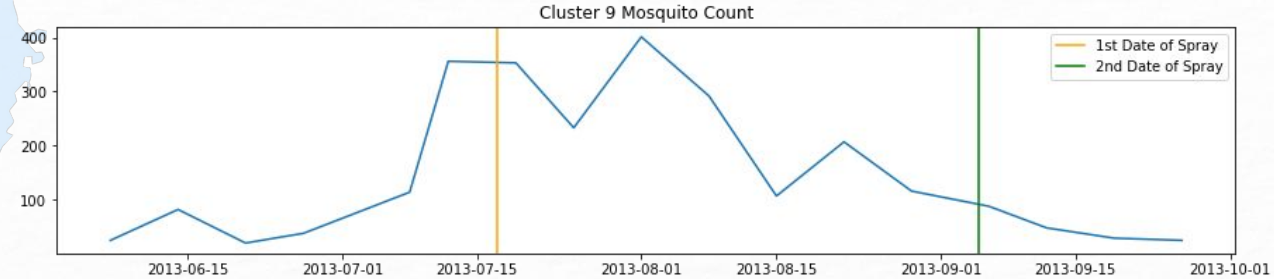
# Compare Pesticide Effectiveness in 2011 & 2013

Using cluster 9 as an example:

## 2011




## 2013

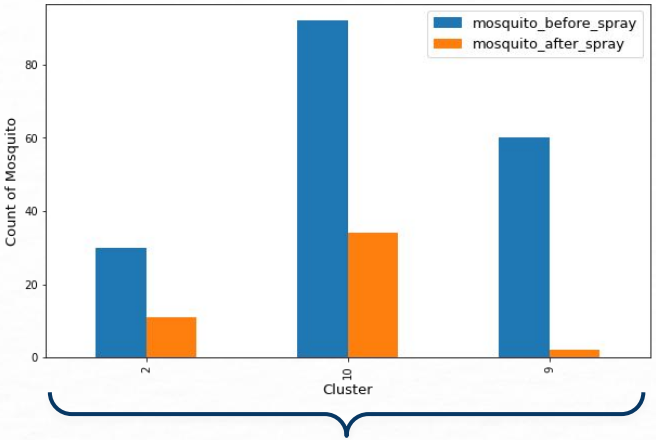



**2013**

- Increased frequency.
- More timely. Spraying before peak of mosquito breeding.

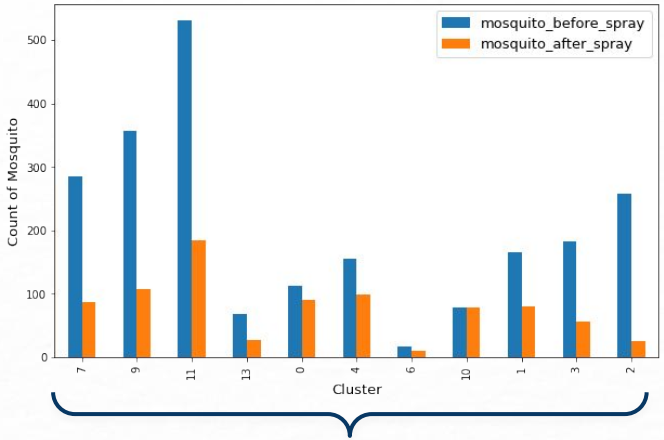
# Compare Pesticide Effectiveness in 2011 & 2013

## 2011




3 clusters

## 2013




11 clusters

**2013**

- Increased frequency
- More timely. Spraying before peak of mosquito breeding.
- Greater coverage

# Pesticide Effectiveness - Cost Savings

**Cost of Each Positive Human Case**

- Medical costs
- Indirect cost due to lost productivity
- $39,000 per positive case on average (source)

| Year | Mosquito Reduced | Estimated Human Cases Reduced | Estimated Cost Savings ($) |
|---|---|---|---|
| **2011** | 135 | 0.4 | 13,700 |
| **2013** | 1362 | 4.3 | 166,500 |

# 04

# Insights & Recommendations

# Conclusions

## Best model

- Gradient Boosting + Random Under Sampler
- F1 score = 0.21, AUC = 0.8

## Working with Imbalanced Dataset

- **Preprocessing**: Resampling techniques (undersampling, oversampling)
- **Choice of Algorithm**: Tree based algorithm tends to perform better
- **Metrics**: F1 score, AUC instead of accuracy score

# Insights and Recommendations

**Maximise benefits of pesticide**

- Increase coverage
- Increase frequency
- Optimise timing, push forward the spraying dates before peak of mosquito breeding (~July)

**Minimise excessive spraying**

- Use model to generate predictions, target areas that are predicted to have virus

# Next Steps

## Better data for modelling

- **More balanced**: More records of traps with WNV present
- **More features**:
  - Air humidity
  - CO2 concentration
  - Distance to nearby water bodies
- **Better record-keeping**:
  - Same trap, same date, 1 record instead of multiple records
  - Include time for each record

“Spray more, but not too much.”

**–ANONYMOUS DSI 32 STUDENT**

# THANK YOU

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