DEEP-SCIENCE

A famous man once said 'we need to go deeper'

Desmond | Jonathan | Soon Poh | Wafir

Background

West Nile Virus is the most widespread mosquito-borne virus in the United States

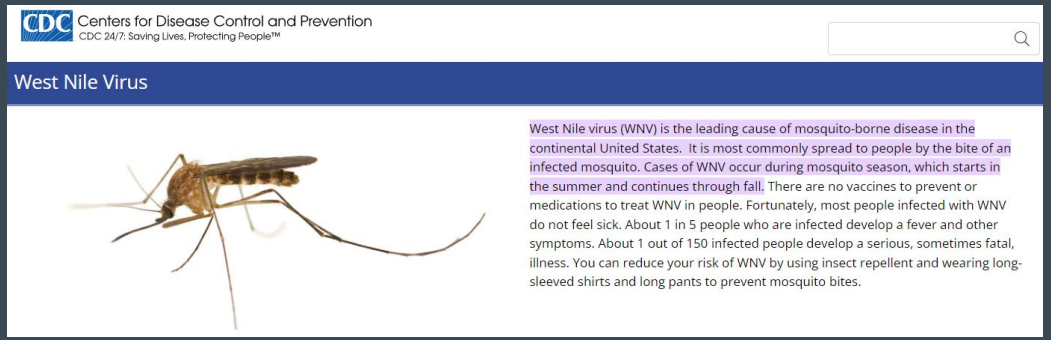


TABLE OF CONTENTS

01 Introduction

02 Data Cleaning and Feature Selection

03 Exploratory Data Analysis

04 Modelling

05 Cost-benefit Analysis, Conclusion and Recommendation

Problem Statement

This project aims to predict West Nile Virus (WNV) in mosquitoes across the city of Chicago for the years 2008, 2010, 2012 and 2014 - to identify potential spray locations and reduce the number of cases

OUR QUESTIONS

1. Are there any significant clusters a.k.a. hot spots
2. Are there any observable trends throughout the year
3. What is the cost-benefit comparison between overcompensating for spray locations or not at all
4. What are the prominent features in predicting WNV

Data Cleaning and Feature Selection

Streamlining Datasets

Train Dataset

- New row duplicated for every 50 mosquitoes vs genuine duplication
- Aggregate number of mosquitoes and number of west nile virus
- Longitude and latitude were the features kept w.r.t. Location

Spray Dataset

- Dropped values beyond the area of train and test datasets

Streamlining Datasets

Weather Dataset (2 stations)

- Missing data ('M', '-', 'T', ' ')
- Impute tavg with mean value
- Impute avgspeed from other weather station
- Impute sealevel and stnpressure from previous and next day values
- Impute preciptotal with median value
- Compute timelag based on average of 9 days for the life cycle of mosquito
- Dropped water1, snowfall, depth, codesum, depart, sunset, sunrise, wetbulb, heat, cool
- Took the average across both stations

Feature Engineering

Train Dataset

- ➔ *Hot Spots
- ➔ Year
- ➔ Month
- ➔ Week of Year

Weather Dataset

- ➔ Relative Humidity

What is Relative Humidity?

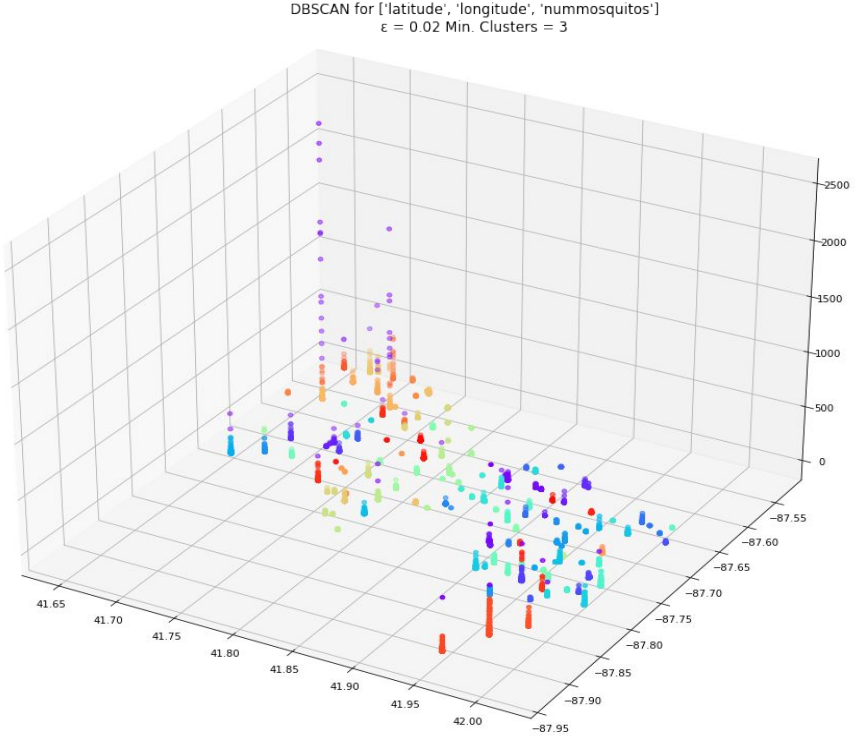
Relative humidity is the amount of water vapor in the air.

Feature Engineering

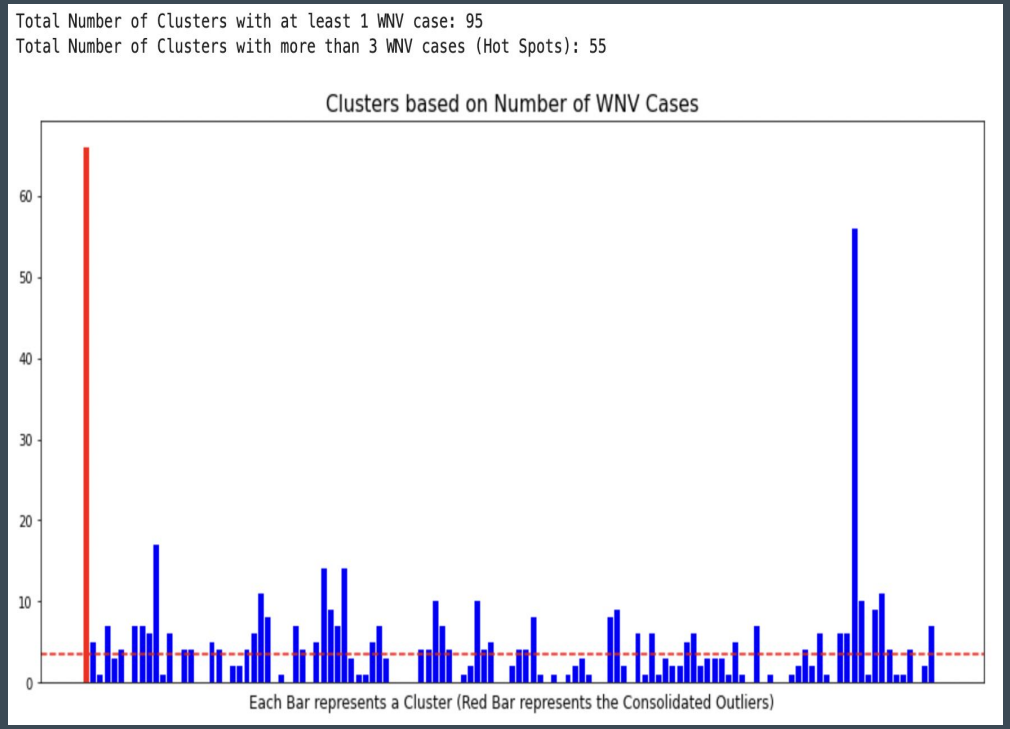
DBSCAN

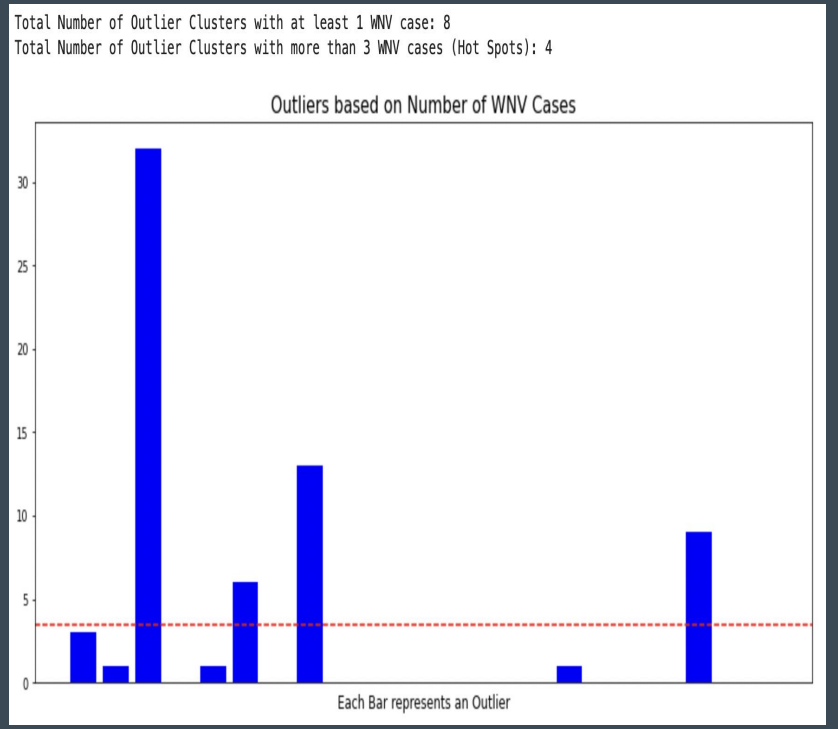
1. Latitude
2. Longitude
3. Number of Mosquitoes

```
Silhouette Score: 0.8538146348419249
Number of outliers: 51 (0.60% of samples)
Number of clusters: 122
```



Feature Engineering

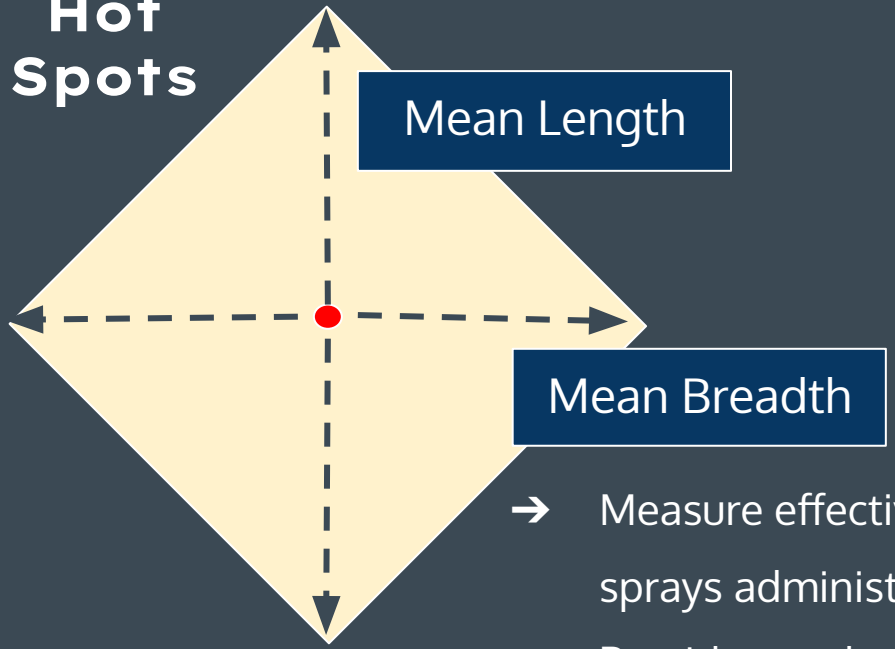




55 Clusters + 4 Outliers = 59 Hot Spots (> once per year on average)

Feature Engineering

Hot Spots

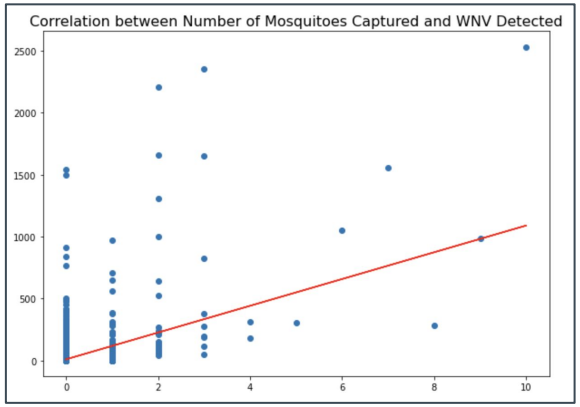


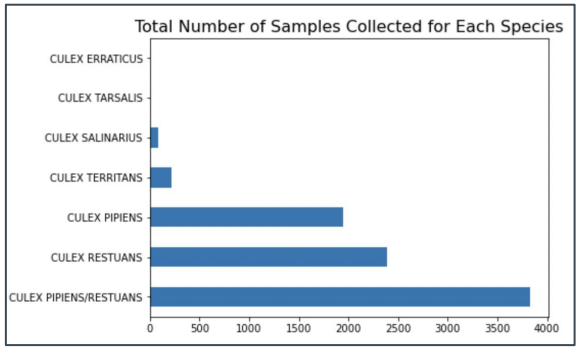
- ➔ Measure effectiveness of sprays administered
- ➔ Provide cost benefit analysis

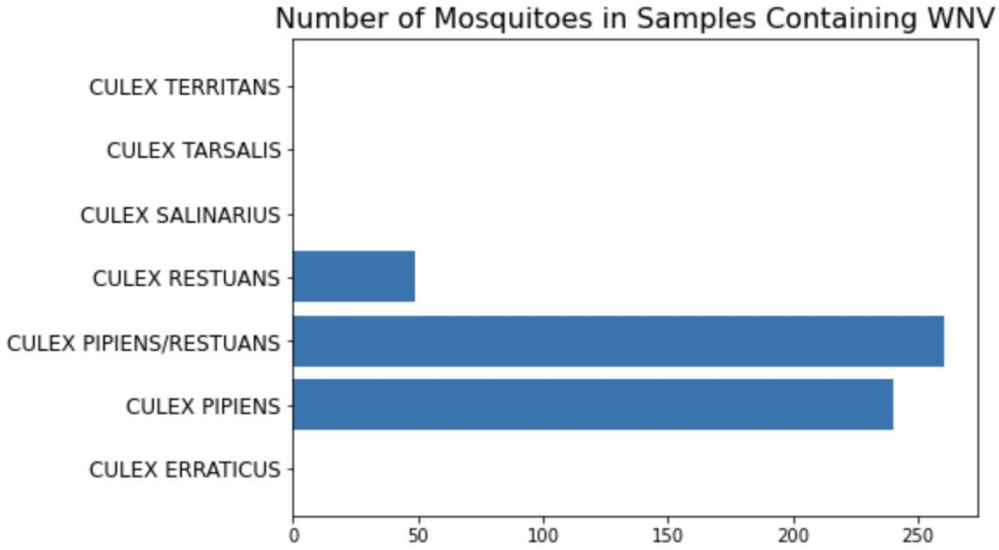
Mean Area

Exploratory Data Analysis

Mosquitoes Species

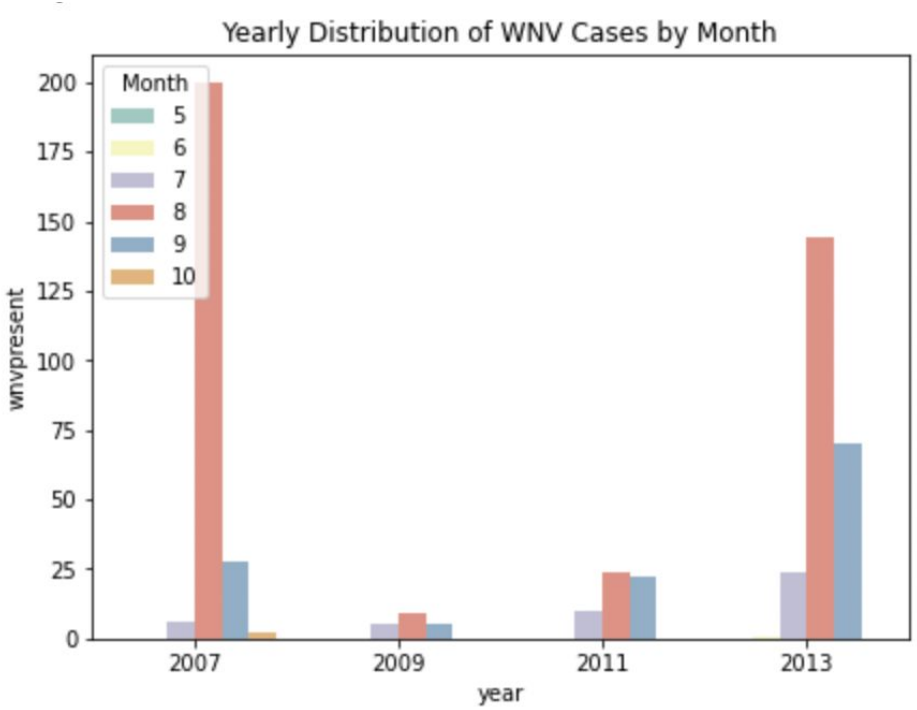


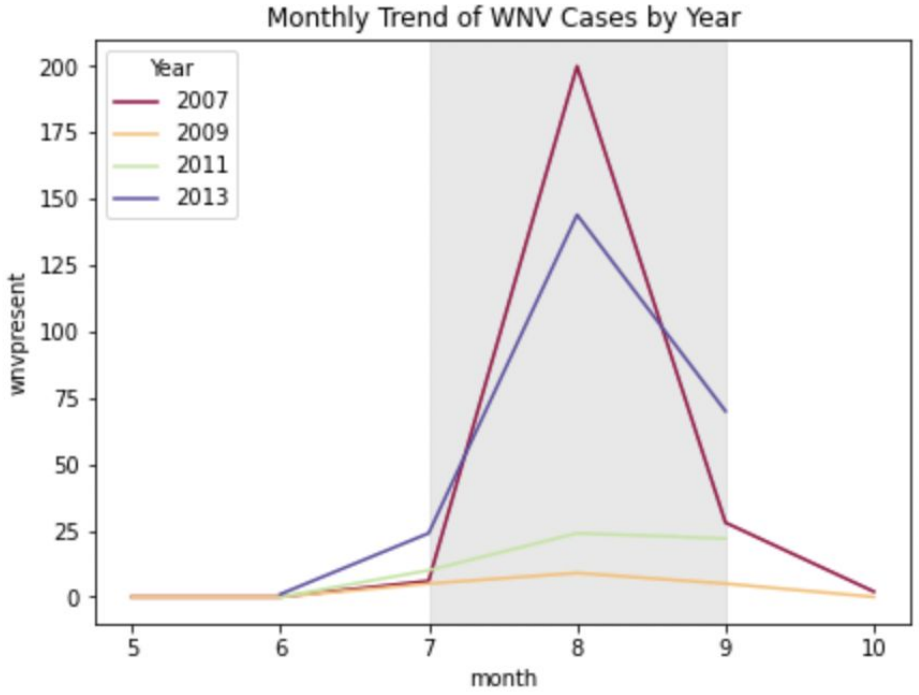




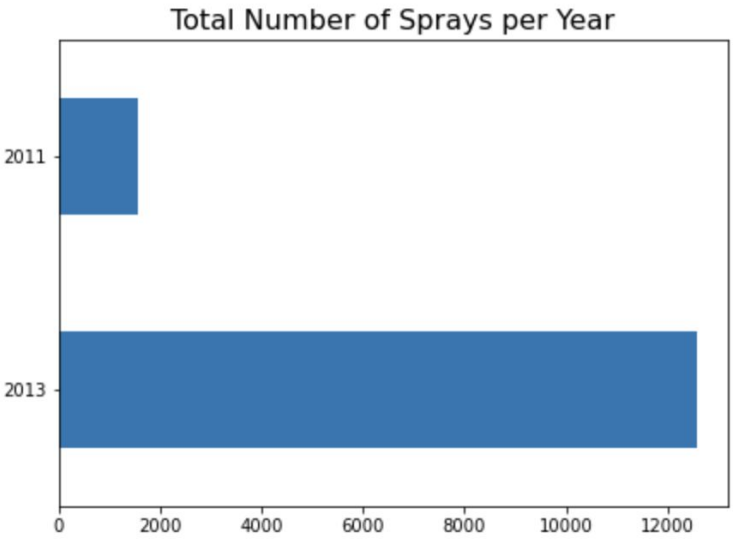
- Three mosquito species to note
- Pearson correlation coefficient *(r)* = 0.49

WNV Trends

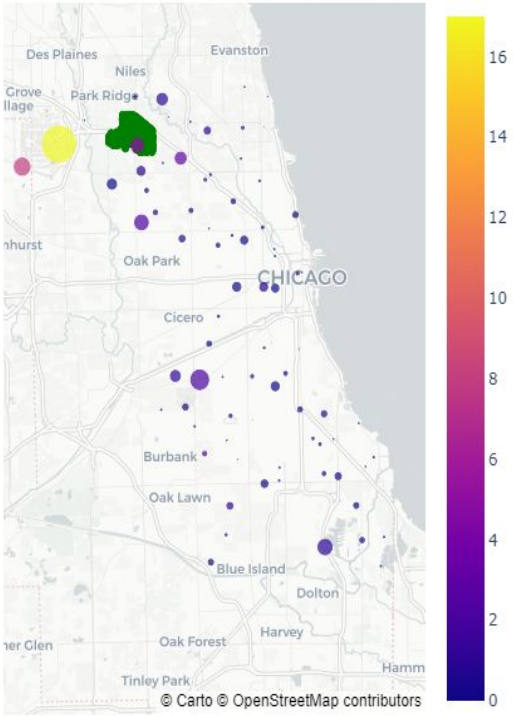

Yearly Distribution of WNV Cases by Month
Month
5
6
7
8
9
10
wnvpresent
0
25
50
75
100
125
150
175
200
2007
2009
2011
2013
year


Monthly Trend of WNV Cases by Year
Year
2007
2009
2011
2013
wnvpresent
0
25
50
75
100
125
150
175
200
5
6
7
8
9
10
month

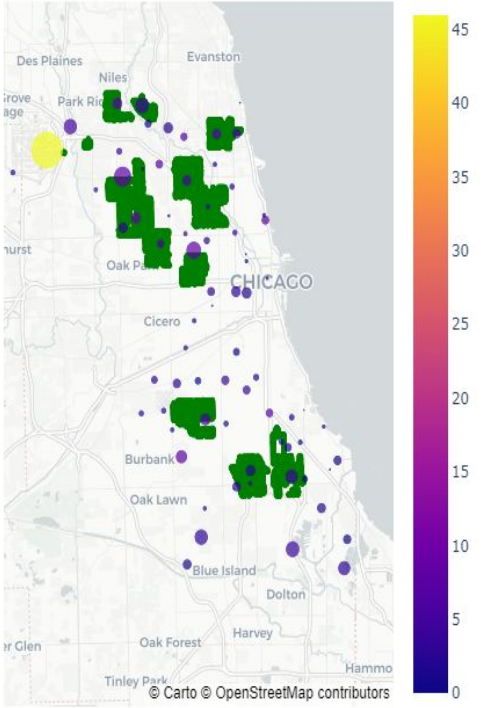
Map Visualisation



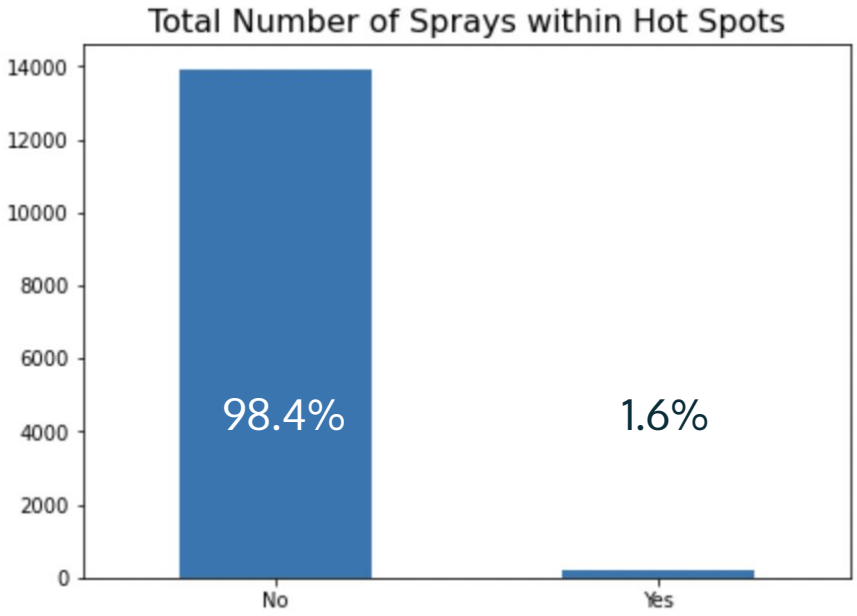
Mosquito and Spray Area in 2011



Mosquito and Spray Area in 2013



Effectiveness of Sprays



Of the 1.6%....

12

Hot Spots were found to be sprayed.

To measure the effectiveness...

It took an average of 17.4 days for first WNV case to occur from the date of spraying.

17.4 **Days**

Modelling

PYCARET

1. Setup environment
2. Compare models
3. Generate model
4. Tune model
5. Finalize and deploy model for prediction

Top 5 Models

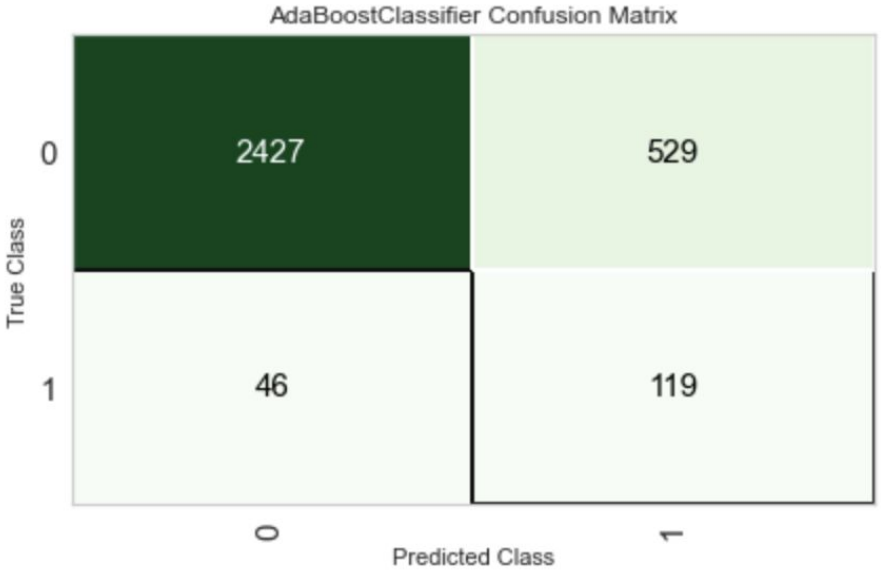
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9026	0.8406	0.3738	0.2344	0.2861	0.2370	0.2455	0.1260
gbc	Gradient Boosting Classifier	0.8281	0.8385	0.5970	0.1739	0.2685	0.2034	0.2543	0.7000
ada	Ada Boost Classifier	0.8060	0.8314	0.6391	0.1620	0.2580	0.1897	0.2502	0.2060
xgboost	Extreme Gradient Boosting	0.9141	0.8311	0.2755	0.2335	0.2511	0.2062	0.2078	0.8820
lr	Logistic Regression	0.7170	0.8268	0.7948	0.1346	0.2300	0.1534	0.2454	0.6080

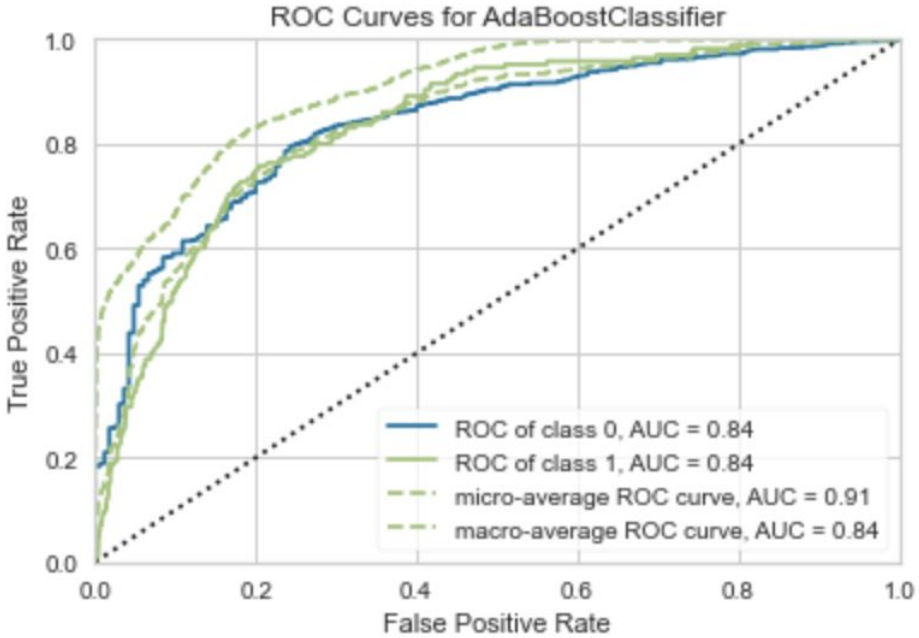
AdaBoost Classifier

- Combines multiple models (weak learners) to reach the final output (strong learners)
- Hyperparameter tuning optimizing AUC to achieve best scores in the scoring metrics (AUC and Recall)
- **Accuracy score - 0.82**
- **AUC score - 0.84**
- **Recall score - 0.72**

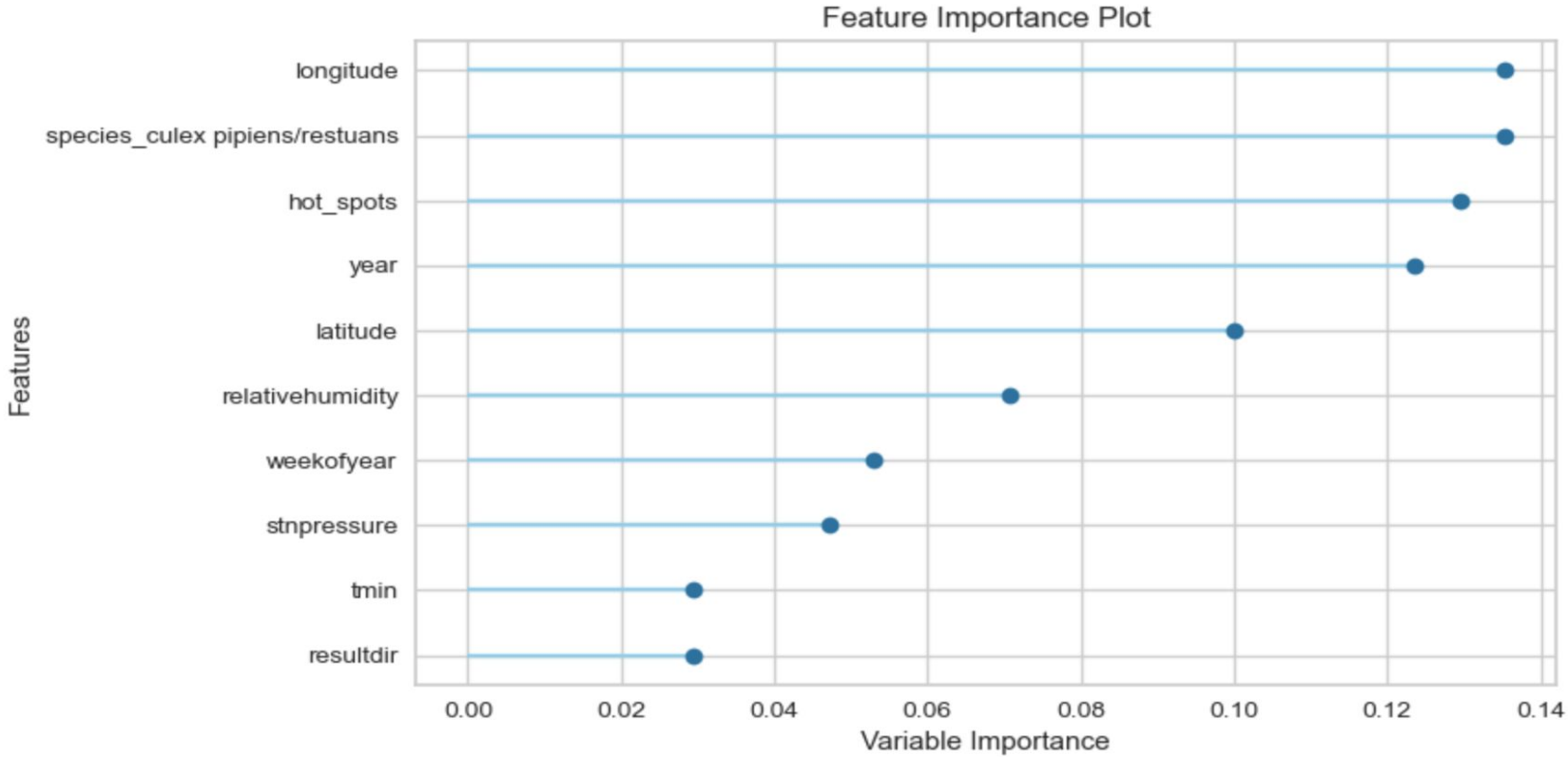
0.7793

Recall and AUC





Feature Importance



Conclusion and Recommendations

Cost-Benefit Analysis

Cost of Spray

- Estimation based on Zenivex
- **USD $300 per gallon**
- Approx **USD $136** per cluster
- **USD $8,024** to spray all the hot-zone spots once
- Recommend to spray once a fortnight for 14 weeks
- Total cost of **USD $56,168** to spray the hot spots in Chicago from June to August

Cost of Medical Care

- **1 in 5** infected people develops fever
- **1 in 150** infected people develops serious nervous system illness
- Costs of additional related medical care and missed worked in the 5 years after initial infection is estimated to cost **USD $ 56 million** across United States annually.
- Mean cost for acute infection **USD $1,177**
- **USD $180** for continuing care
- Expected 1 year cost were **USD $13,648**, adjusted for survival

Conclusion

AdaBoost Classifier

- ROC-AUC score of 0.77 from Kaggle
- Accuracy score of 0.82
- AUC score of 0.84
- Recall score of 0.72

Cost of spraying is low

- Reducing false negative is important
- 2011 and 2013 only covered **1.6% of the cluster hotspots** identified
- Effect of spraying has a significant impact on reducing WNV
- Caveat that there is no control to weather and climate change

Recommendation

Spraying of insecticides

- Location - focus on areas where the model predicts high probability
- Time - focus on months with higher number of mosquitoes, eg July and August
- Wind direction - further research required to determine the inefficiency caused by wind

Public education

Increase checks and patrolling during hotspot season

Project Wolbachia

- Adoption of project as proven effective in countries such as Singapore, Australia and Brazil


THANK
YOU