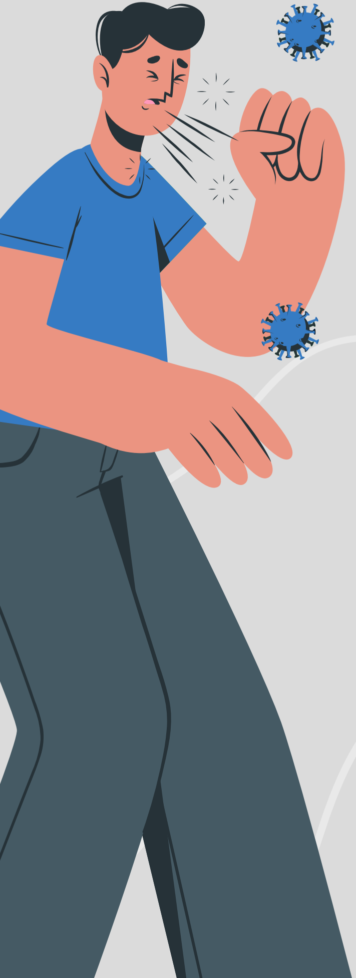


West Nile Virus Prediction

Catherine
Mei Lian
Wenna

Table of Contents

01

Introduction

02

Problem statement

03

Data cleaning & EDA

04

Preprocessing & modelling

05

Cost-benefit analysis Conclusions

Introduction

West Nile virus is most commonly spread to humans through infected mosquitos. Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death.

By 2004, the City of Chicago and CDPH had established a **comprehensive surveillance and control program** that is still in effect today.

Every week from June to early October, mosquitos in traps across the city are tested for the virus. The results of these tests influence **when** and **where** the **city will spray airborne pesticides** to control adult mosquito populations.

Problem Statement

Build a model that predicts outbreaks of the West Nile virus that will help the City of Chicago and CDPH more efficiently and effectively allocate resources towards preventing the transmission of this potentially deadly virus.

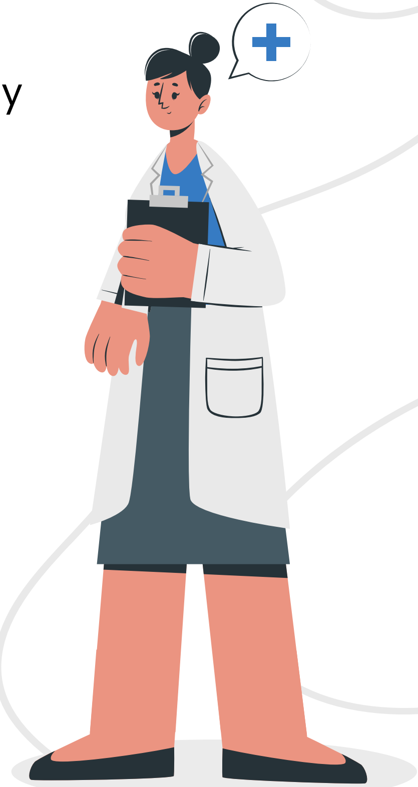
Data Cleaning & EDA

Data Cleaning

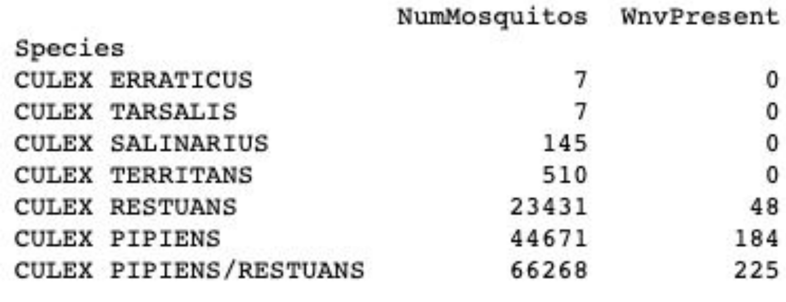
- **Train Dataset**
  - Sum the number of mosquitoes within the same samples if they are collected on the same day.

- **Spray Dataset**
  - Removed 543 duplicated rows. Likely due to data collection error.
  - Dropped 'Time' attribute as it has 584 missing values and too granular for our needs.

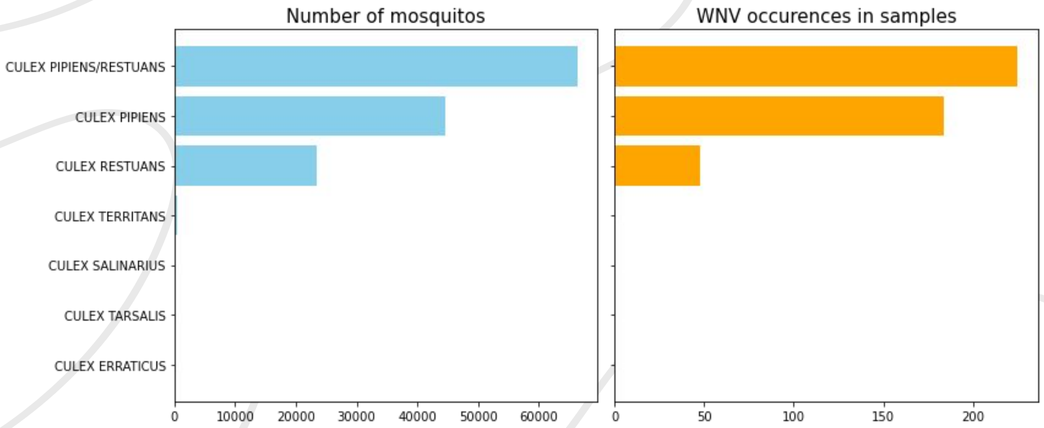
- **Weather Dataset**
  - Dropped 'Depth', 'Water1', 'SnowFall' due to high percentage of missing and zero values (>99.5% null or 'M' or '-' or '0' or '0.0')
  - Imputed missing values in the following attributes: 'Tavg' , 'Depart', 'WetBulb', 'Heat', 'Cool', 'Sunrise', 'Sunset', 'PrecipTotal', 'StnPressure', 'SeaLevel', and 'AvgSpeed'.



EDA Findings



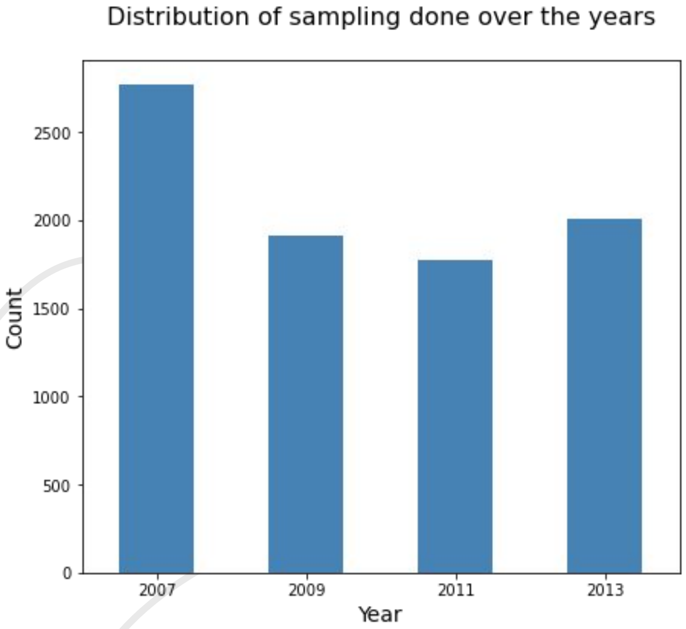
```
                         NumMosquitos  WnvPresent
Species
CULEX ERRATICUS                     7           0
CULEX TARSALIS                      7           0
CULEX SALINARIUS                  145           0
CULEX TERRITANS                   510           0
CULEX RESTUANS                  23431          48
CULEX PIPIENS                   44671         184
CULEX PIPIENS/RESTUANS          66268         225
```

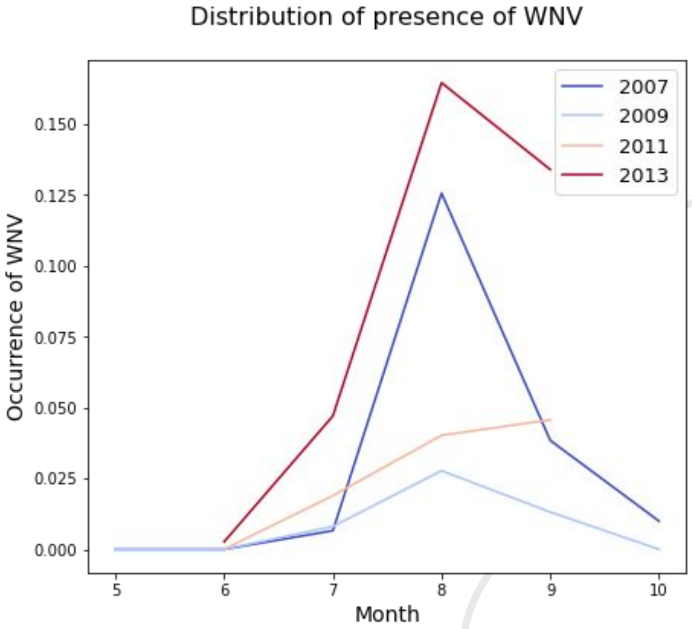


Among the 6 species of mosquitoes identified, only 2 of them (**Culex Pipiens** and **Culex Restuans**) are WNV carriers.

They constitute **99.5%** of the mosquitoes captured in the traps.

EDA Findings



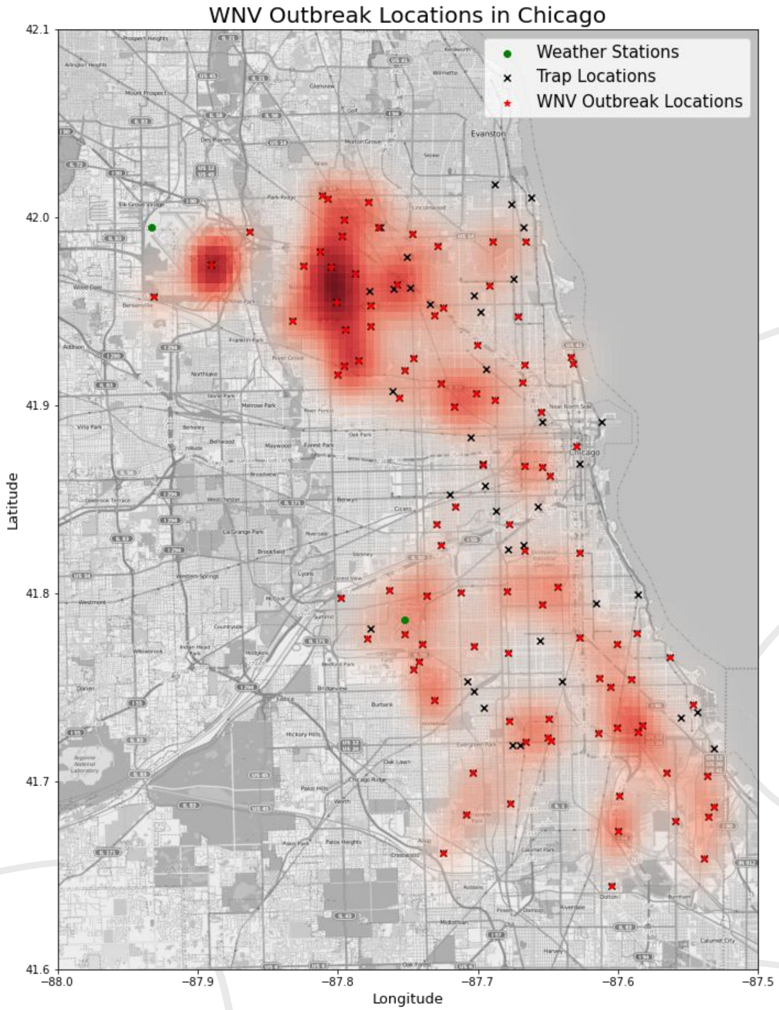


- The total number of the traps implemented each year has generally **reduced** since 2007.
- There is an **increasing prevalence of WNV during summer** - an increasing trend from June/July till it peaks in August, before declining slightly in September.
- In August 2013, the occurrence is the highest - compared to the earlier years during the same month.

EDA Findings

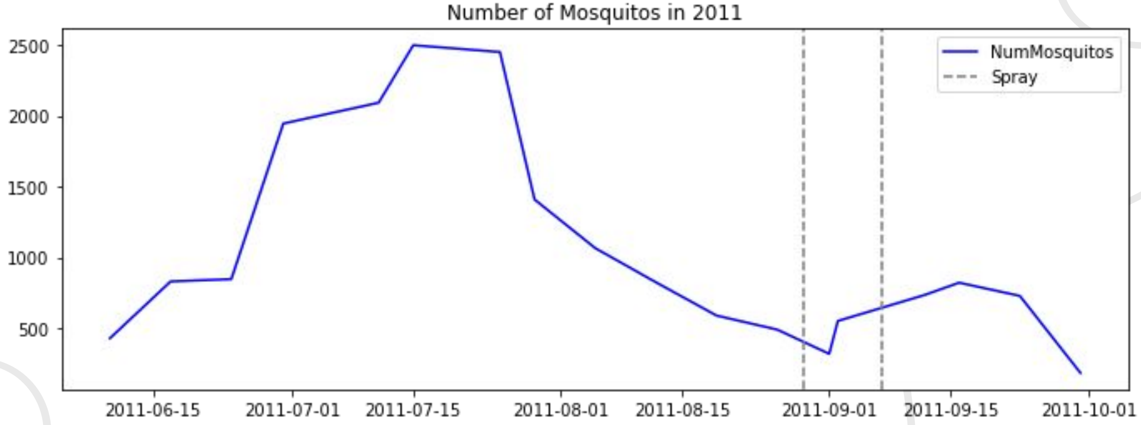
Area with the **darker red** indicates that the region has **more mosquitoes** that carry the West Nile virus.

The occurrences of West Nile virus is more prevalent near bodies of **water** and **O'Hare airport**.

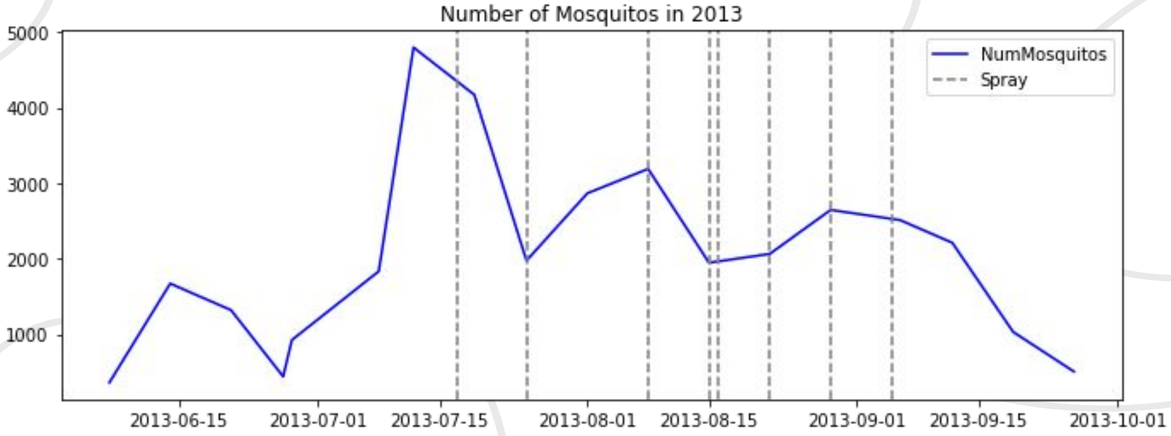


EDA Findings

There were 2 sprays in 2011 and we can see there are **positive effects** at the start, but the number of mosquitoes still increases thereafter.



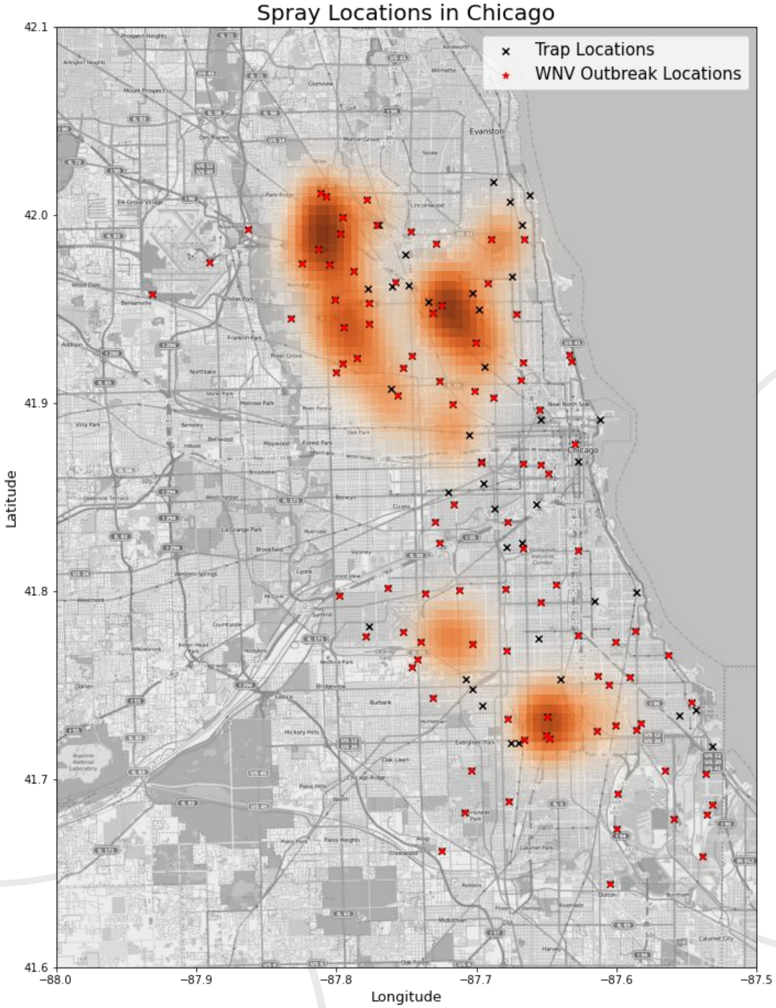
There were 8 sprays in 2013. There are positive effects for **certain sprays**, but not all.



EDA Findings

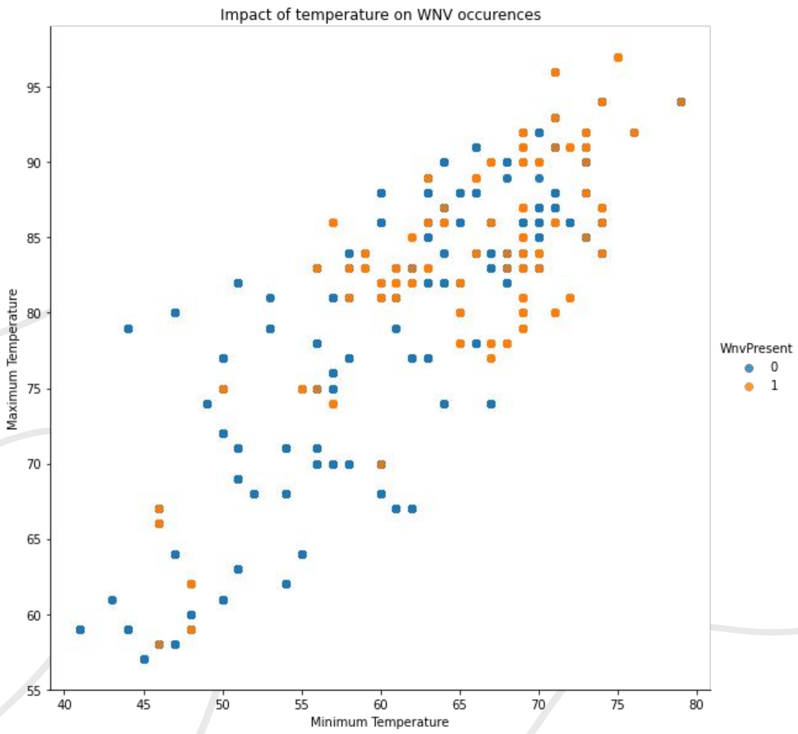
Area with the **darker orange** indicates that the region has more spray concentration area and the area with lighter orange means that less spray concentration area.

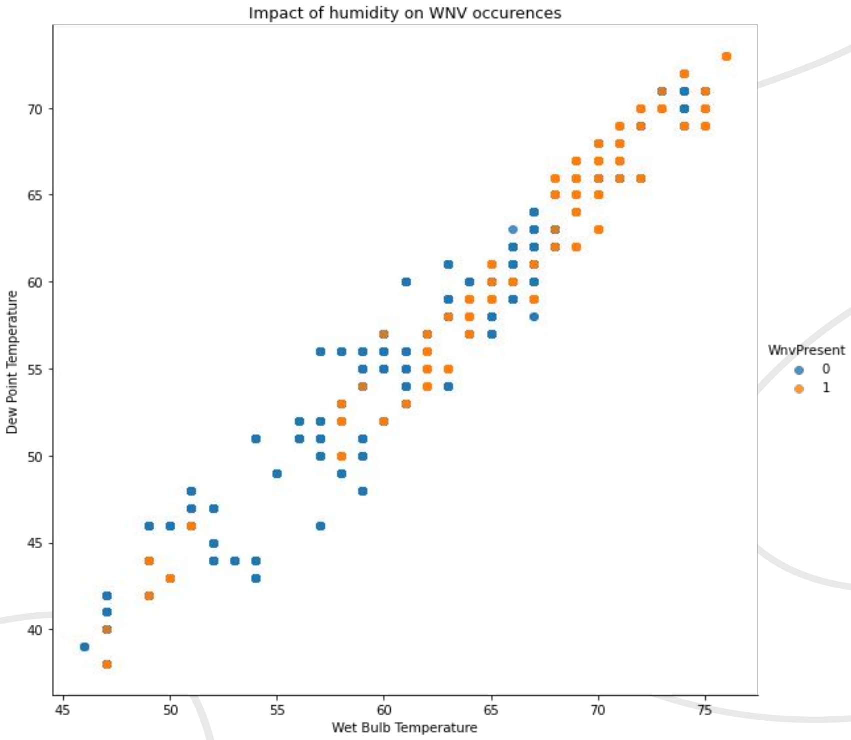
The spray area **fails to fully overlap** with the virus outbreak locations. This is a cause for concern.



EDA Findings

West Nile Virus more prevalent during days with **high temperature** and **high humidity**.





Preprocessing & modeling

Feature engineering

Sun Hours
Duration of daylight between sunrise and sunset

Code Sum Score
Map different weather conditions to ‘Wet’ or ‘Not Wet’

Humidity
A formula that involves Average Temperature and DewPoint to calculate humidity

Time-lagged features
Create lags in days for selected weather features such as Average Speed, Precipitation Total, Resultant Direction

Number of mosquitos
Predict number of mosquitos that is missing from the test dataset

Baseline score

WNV not present: 94.6%
(i.e 94.6% accuracy if we predict WNV to be negative)

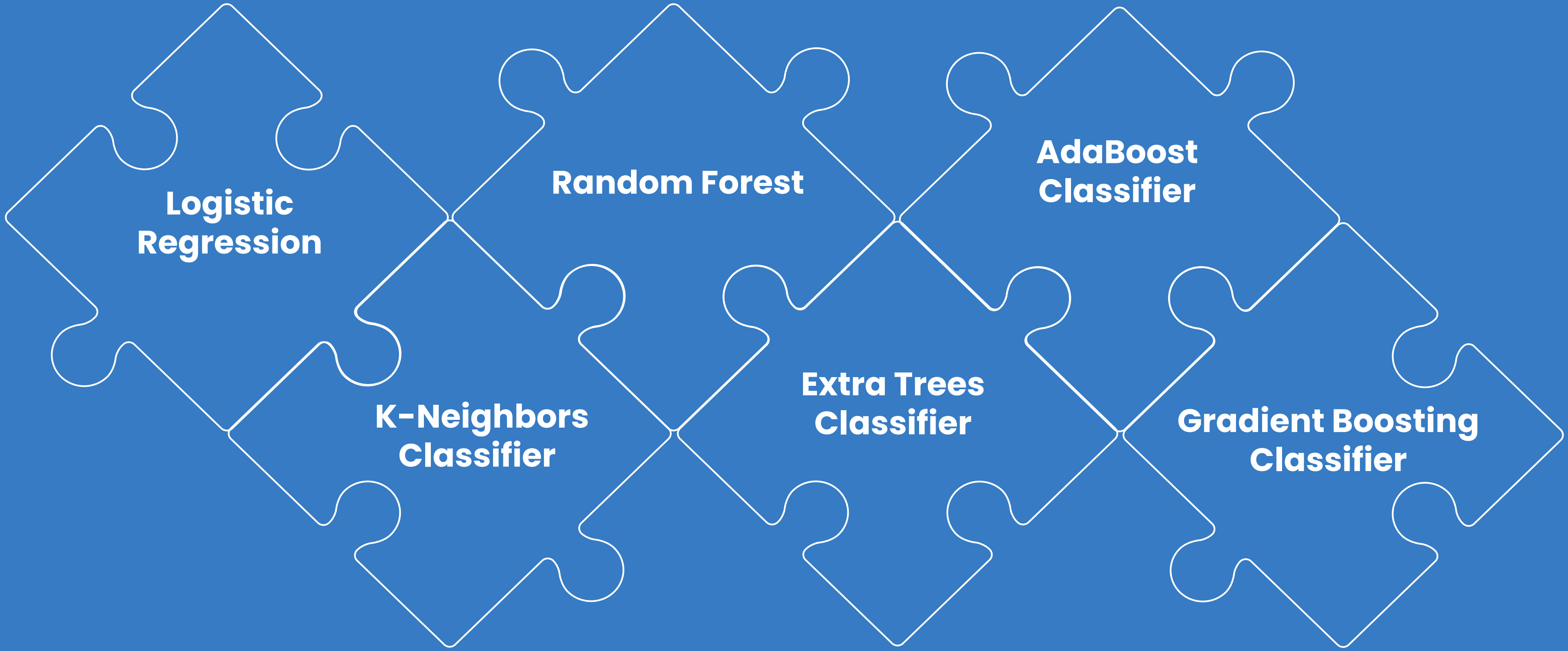
Imbalance dataset

Oversample minority class use SMOTE to make it more balanced - 50:50

Metrics

Use ROC-AUC score (50.0%) to optimise instead of accuracy

Modeling



Logistic Regression

Random Forest

AdaBoost Classifier

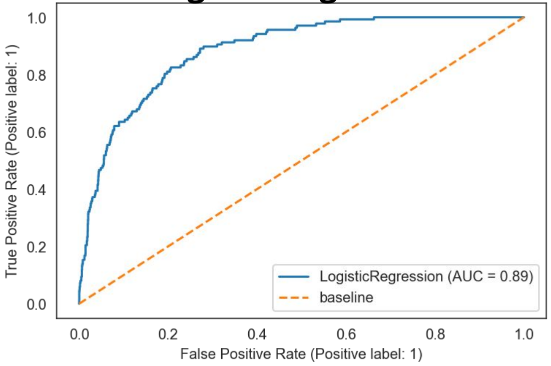
K-Neighbors Classifier

Extra Trees Classifier

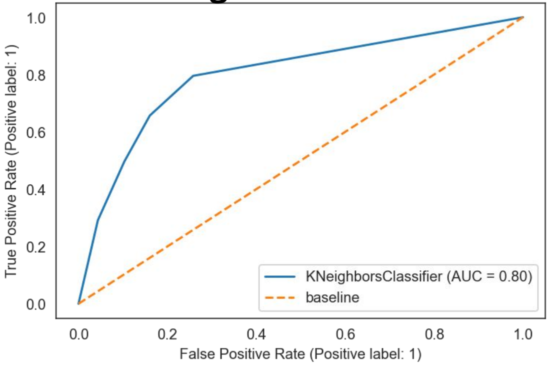
Gradient Boosting Classifier

ROC-AUC graph

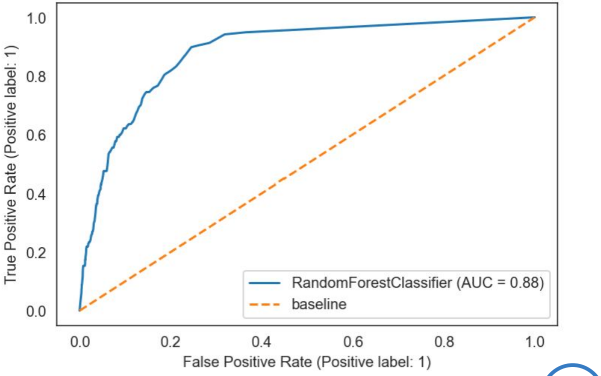
Logistic Regression



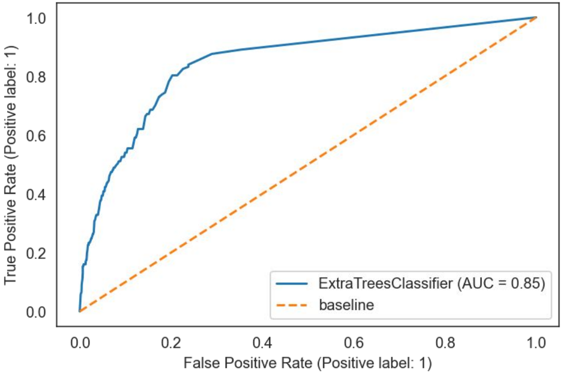
K-Neighbors Classifier



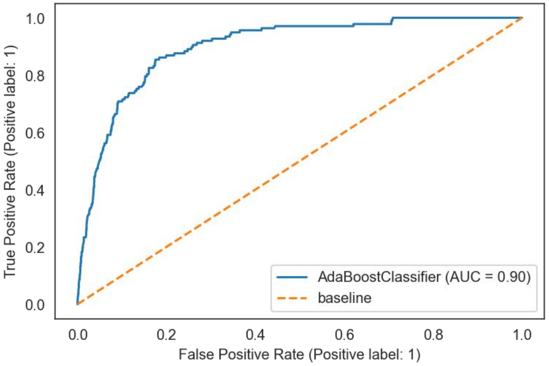
Random Forest Classifier



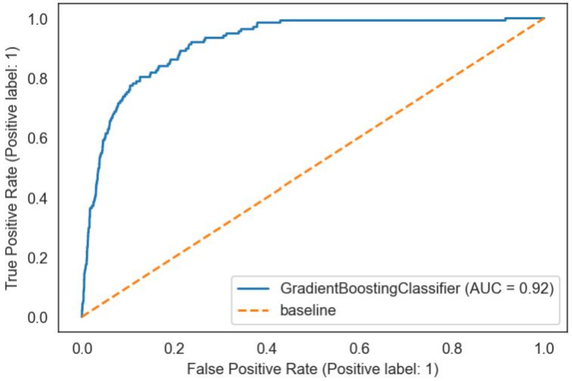
Extra Trees Classifier



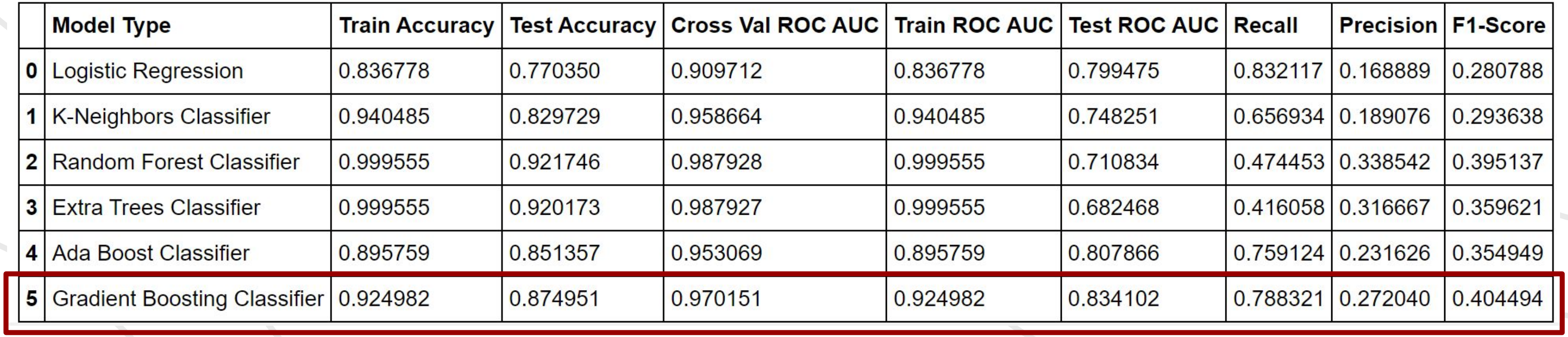
Ada Boost Classifier



Gradient Boosting Classifier



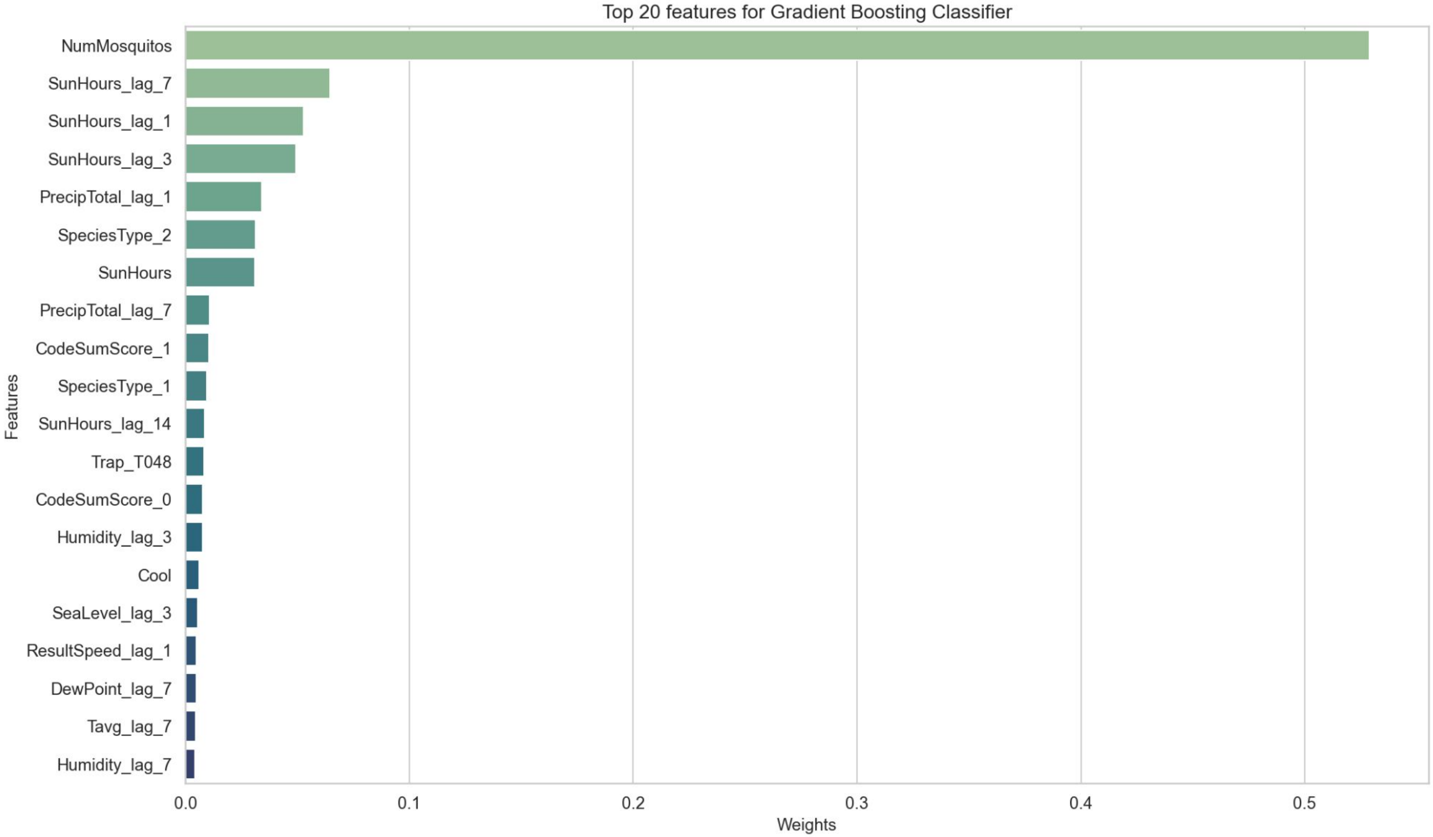
Modeling results



	Model Type	Train Accuracy	Test Accuracy	Cross Val ROC AUC	Train ROC AUC	Test ROC AUC	Recall	Precision	F1-Score
0	Logistic Regression	0.836778	0.770350	0.909712	0.836778	0.799475	0.832117	0.168889	0.280788
1	K-Neighbors Classifier	0.940485	0.829729	0.958664	0.940485	0.748251	0.656934	0.189076	0.293638
2	Random Forest Classifier	0.999555	0.921746	0.987928	0.999555	0.710834	0.474453	0.338542	0.395137
3	Extra Trees Classifier	0.999555	0.920173	0.987927	0.999555	0.682468	0.416058	0.316667	0.359621
4	Ada Boost Classifier	0.895759	0.851357	0.953069	0.895759	0.807866	0.759124	0.231626	0.354949
5	Gradient Boosting Classifier	0.924982	0.874951	0.970151	0.924982	0.834102	0.788321	0.272040	0.404494

Final model: Gradient Boosting Classifier, as it has the highest ROC AUC score and 2nd highest Recall score.

Top 20 features for predicting WNV



Cost-benefit analysis

Cost-benefit analysis

Spraying

- If The Chicago Department of Public Health (CDPH) detects the presence of West Nile Virus in the mosquitoes, they will spray at dusk till approximately 1am
- An insecticide call Zenivex™ is used that cost $0.67 per acre
- It takes about 8 - 10 days for an egg to develop into an adult mosquito, recommended to spray every 10 days for 4 months i.e. 12 times in total

Medical and productivity cost

- About 1 in 5 people who are infected will develop a fever with headache and joint pains
- About 1 in 150 people develop a nervous system illness that requires hospitalisation
- Using Sacramento County's estimated economic impact in 2005 as reference:
    - West Nile fever - 117 people
    - West Nile neuroinvasive disease - 46 people
    - Fatal case - 1
    - Total medical and productivity cost: Approx $2.98 million

Cost-benefit analysis

The ratio of cost of spraying to the average cost of treatment per person is approximately 1:7

To calculate the cost-benefit, these are the factors that we consider:

- The threshold of the probability of the predicted output that WNV=1
- Confusion matrix and the respective cost for TN, FP, FN, TP

Threshold	Total cost/savings
0.33	$7,363 savings
0.5	$5,533 savings
0.8	$9,375 costs

Conclusion

&

Recommendations

Conclusion

Observations

- Culex Pipiens and Culex Restuans species are the two main carriers of WNV.
- Increasing prevalence of WNV during summer (June to August).

Modelling & Predictions

- Handled imbalance data with SMOTE.
- Selected Gradient Boosting Classifier as our chosen model, with highest ROC AUC score of 0.8342.
- Top features include number of mosquitos, duration of sun hours, total precipitation, etc.
- Cost benefit analysis revealed an expected savings of $7,683.

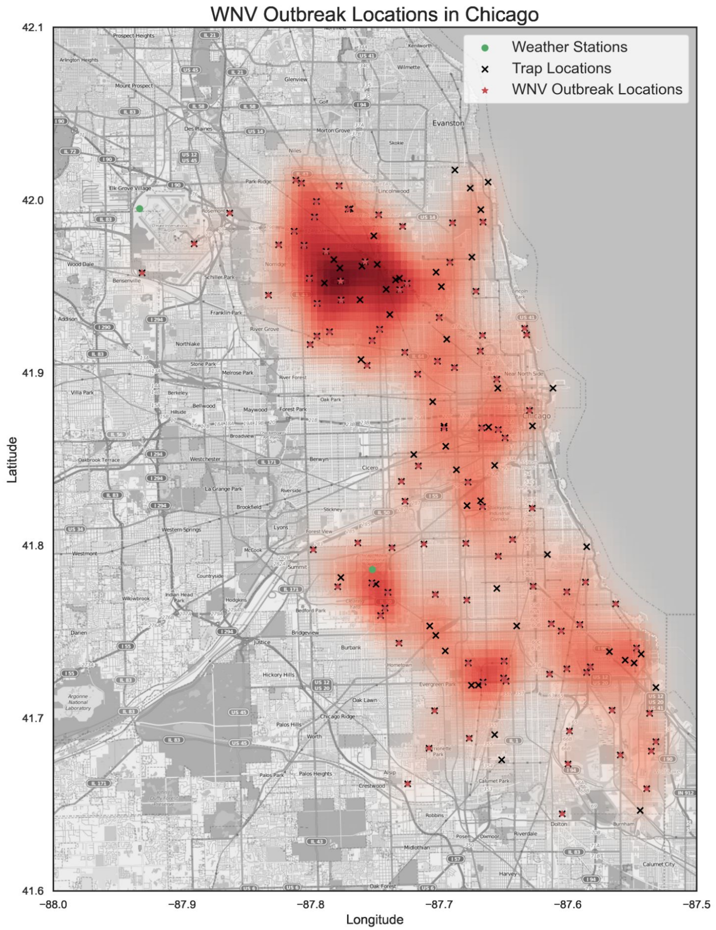
Recommendations

What can be done:

- Recommend CDPH to ramp up on its spraying efforts, by focusing on regions in the northern part of Chicago, particularly areas near water bodies.

Future explorations:

- Employing bagging-based or boosting-based techniques for handling imbalanced data;
- Gathering additional years of pesticides spraying data to examine the effectiveness of the spray;
- Looking into more recent years of trap, weather and spray information (from 2014 onwards).



Thank You!

Do you have any questions?