West Nile Virus Outbreak Prediction
Chicago

Wei Zhe, Kenneth, Shawn, Joanne, Yan Da

Table of contents

01

Background

02

Methodology

03

Exploratory Data Analysis

04

Modelling

05

Cost Benefit Analysis

06

Recommendations

01
Background

Background

The Chicago Department of Public Health (CDPH) runs a surveillance program annually to keep mosquito count low and to protect its residents from the West Nile virus (WNV).

Challenges faced

- Chicago ranked in the Top 5 US cities, with highest no. mosquitoes in 2021

- Priority shifted towards fighting COVID-19

- Tightening of funds due to looming economic recession

Problem statement

Build a model with more than 70% recall to predict the period and location where mosquitoes will test positive for WNV, enabling CDPH to preemptively allocate the city's spraying resources to curb the virus transmission.

Transmission of West Nile virus

- West Nile virus (WNV) is a vector-borne pathogen, carried and spread by Culex Pipiens mosquito species

- WNV passed through a transmission cycle between infected birds and mosquitoes

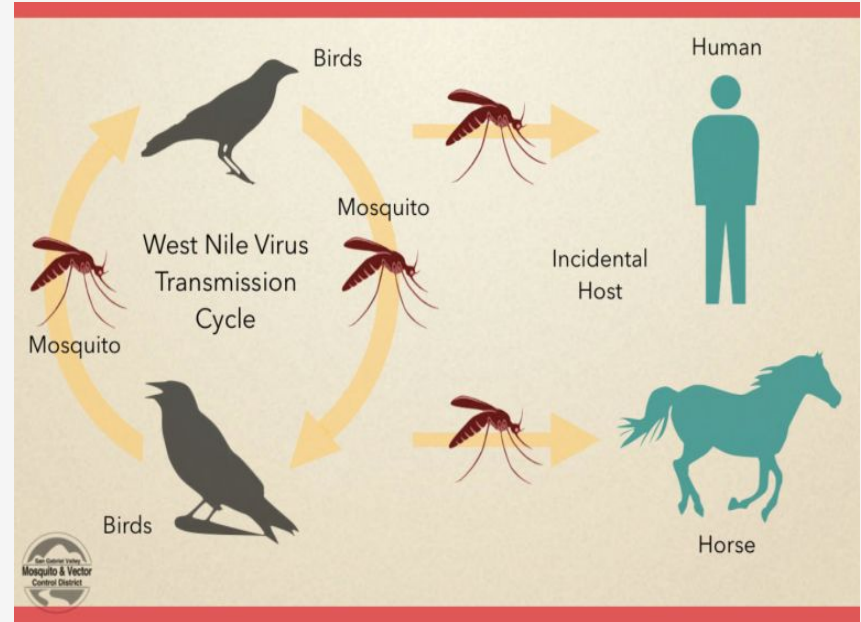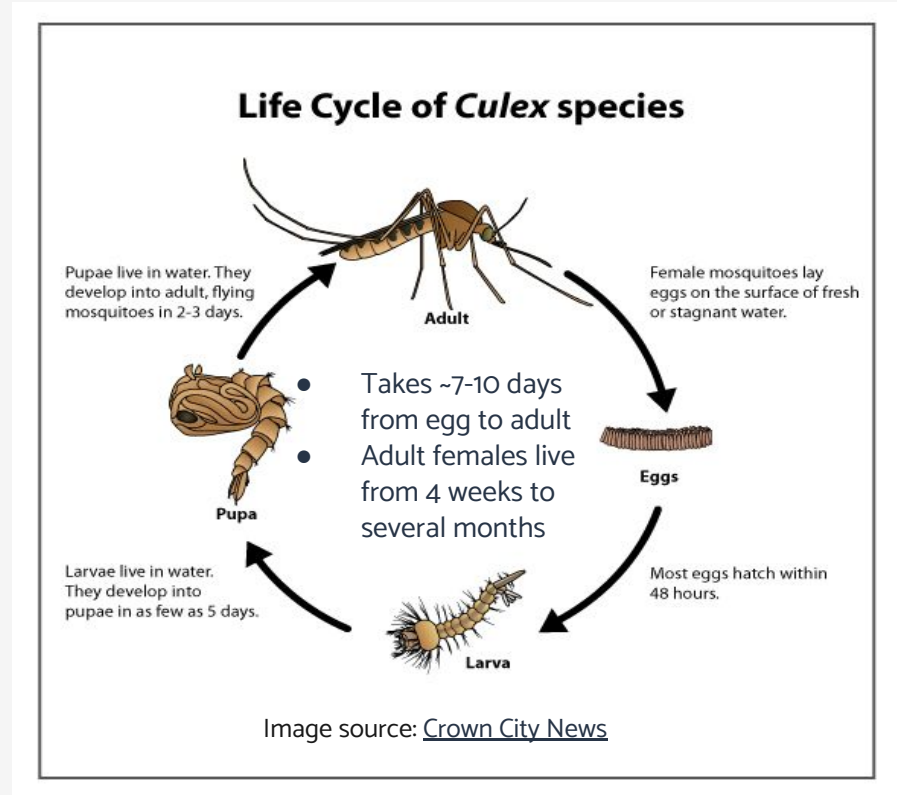- Eventual spread to their incidental hosts through the mosquitoes

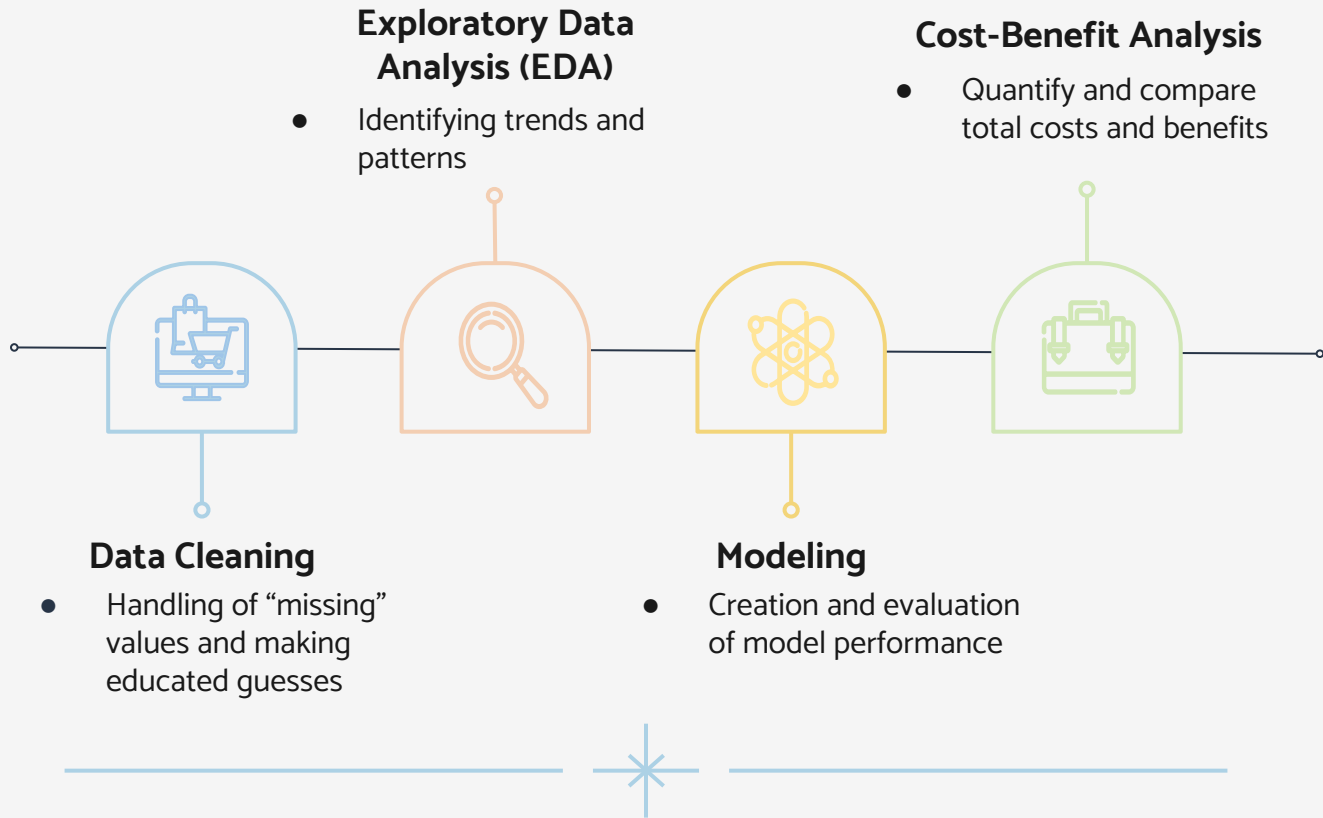Image source: Crown City News

Life cycle of Culex Pipiens

Culex Pipiens Mosquito

Life Cycle of *Culex* species

Adult

Eggs

Larva

Pupa

Pupae live in water. They develop into adult, flying mosquitoes in 2-3 days.

Female mosquitoes lay eggs on the surface of fresh or stagnant water.

Most eggs hatch within 48 hours.

Larvae live in water. They develop into pupae in as few as 5 days.

- Takes ~7-10 days from egg to adult
- Adult females live from 4 weeks to several months

Image source: Crown City News

02
Methodology

Overview of workflow

Exploratory Data Analysis (EDA)

- Identifying trends and patterns

Cost-Benefit Analysis

- Quantify and compare total costs and benefits

Data Cleaning

- Handling of "missing" values and making educated guesses

Modeling

- Creation and evaluation of model performance

03
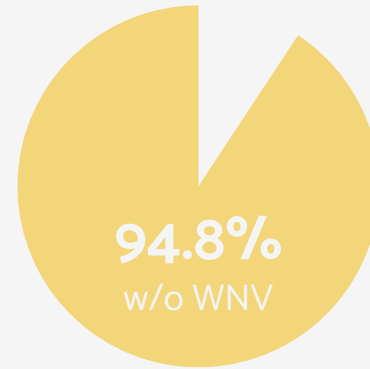# Exploratory Data Analysis

Overview of the datasets

Data period (May - Oct):

- Information of mosquito traps (Year 2007, 2009, 2011, 2013)

- Spray records (Year 2011, 2013)
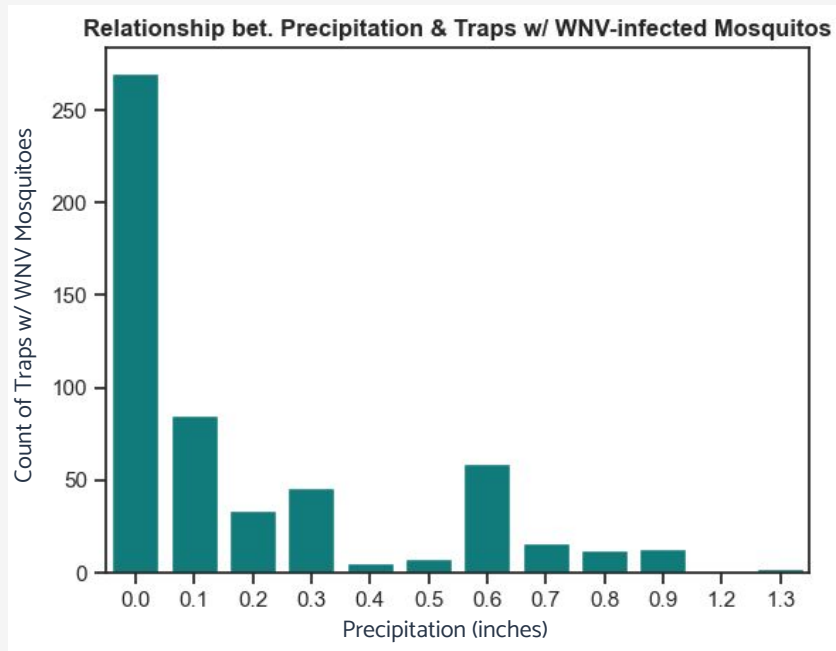
- Weather readings (Year 2007 - 2014)

Processing of weather dataset:

- 7-day rolling average (to align with mosquito life cycle)

94.8%
w/o WNV

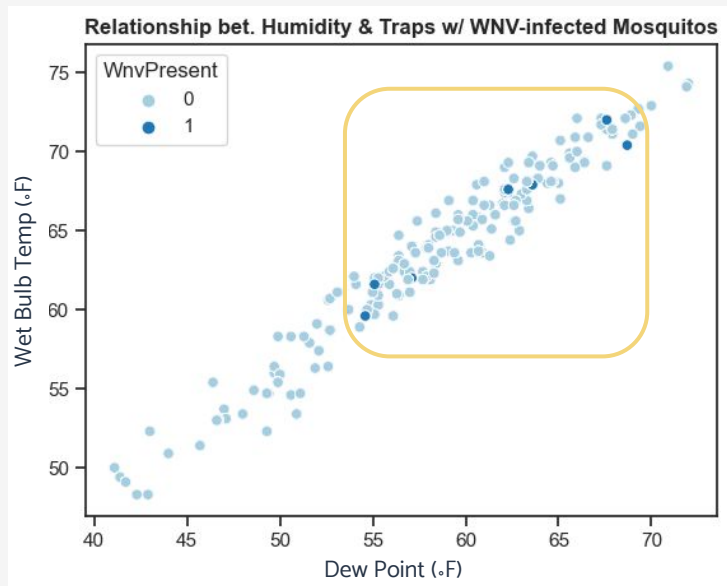Class imbalance with 94.8% of data points (105,06) classified as negative case

Presence of WNV-infected mosquitoes in low precipitation during the warmer seasons
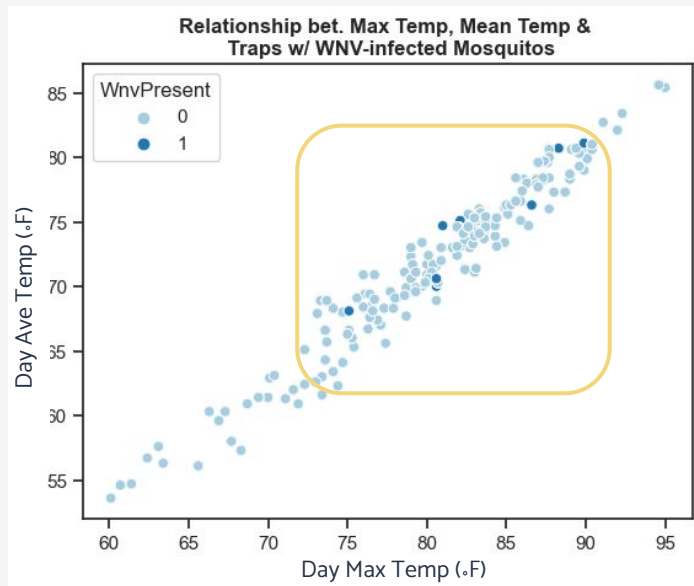

Relationship bet. Precipitation & Traps w/ WNV-infected Mosquitos

- Precipitation (0 inch) accounted for 50% of the data points

- Average precipitation recorded: 0.1-0.2 inches

- Precipitation may not play a key role in facilitating the breeding of mosquitoes in Chicago

Data information: (1) Period: 2007, 2009, 2011, 2013; (2) 7-day rolling average was applied to Precipitation

More WNV-infected mosquitoes in higher humidity and temperature

Relationship bet. Humidity & Traps w/ WNV-infected Mosquitos

Relationship bet. Max Temp, Mean Temp & Traps w/ WNV-infected Mosquitos

High humidity is optimal for mosquitoes to breed

Higher temperature (from 16∘C to 32∘C)* accelerated mosquito population growth

- High humidity leads to increased Dew Point and Wet Bulb temperatures.
- Data information: (1) Period: 2007, 2009, 2011, 2013; (2) 7-day rolling average was applied to the variables

*Source: Goutam Chandra, Devaleena Mukherjee, in *Advances in Animal Experimentation and Modeling*, 2022

Spraying helped reduce mosquito population


Non-infected Mosquito Count by Week
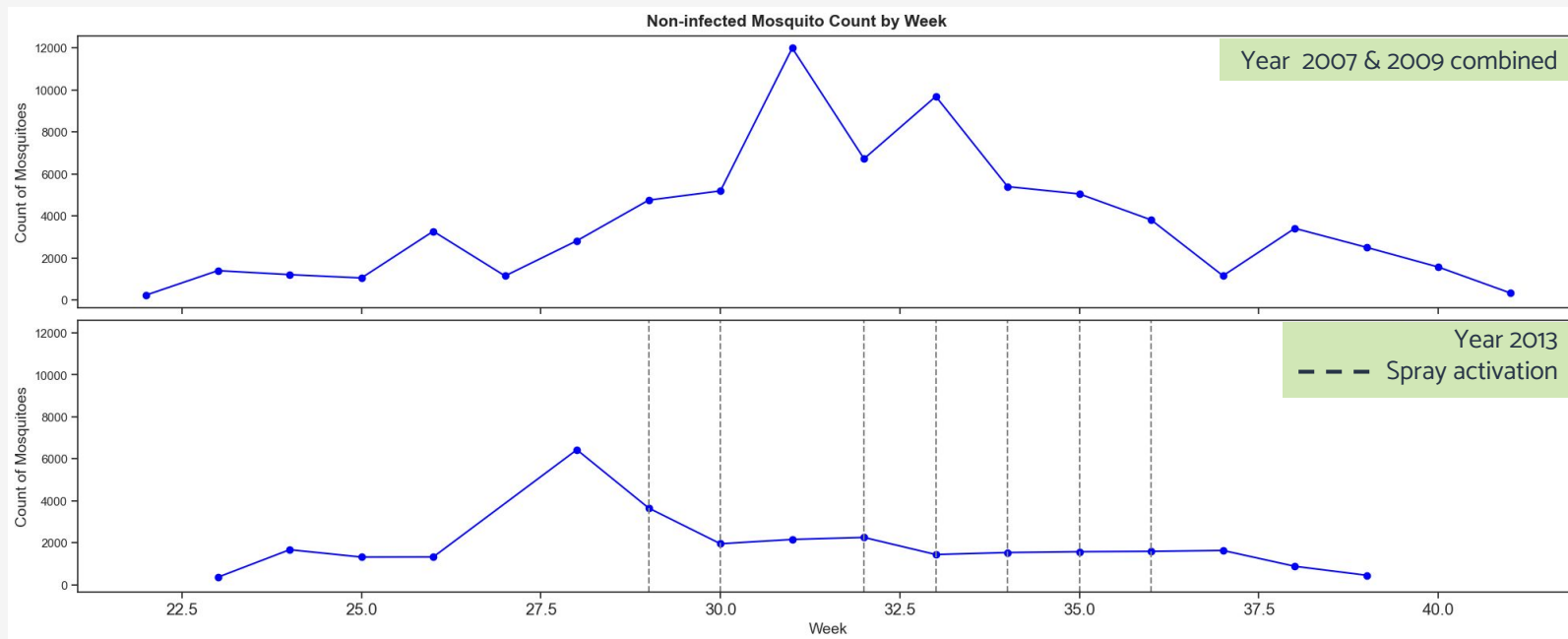
- Without spraying, mosquito population peaked around Week 31

- After spraying around Week 30 period, there was a decline followed by a plateau in the mosquito population

Spraying was not consistently effective in managing WNV-infected mosquitoes

No. of WNV-infected Mosquitos by Week for 2013
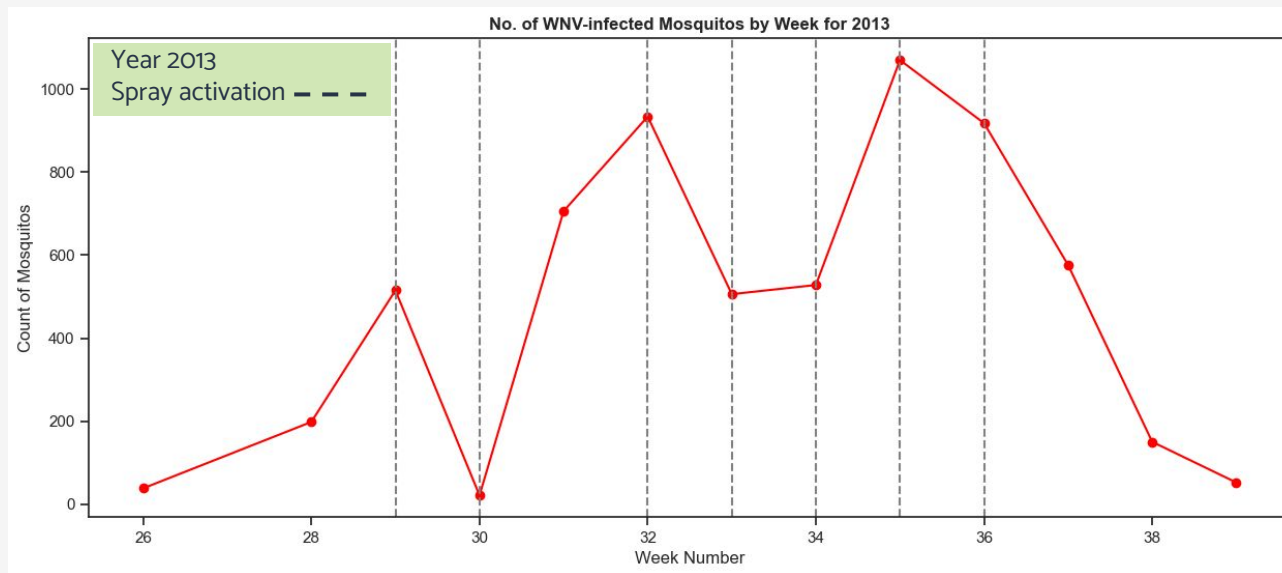
- There was no significant effect on the number of infected mosquitos after each spray

- Takes a while for spray to work; may need to look at spraying earlier and more frequently

Hotspots for WNV over the years
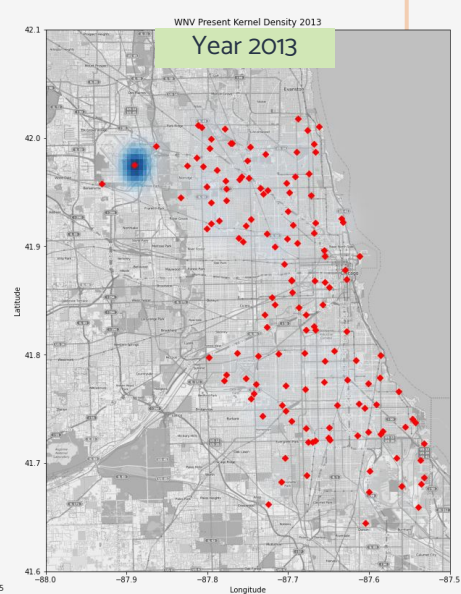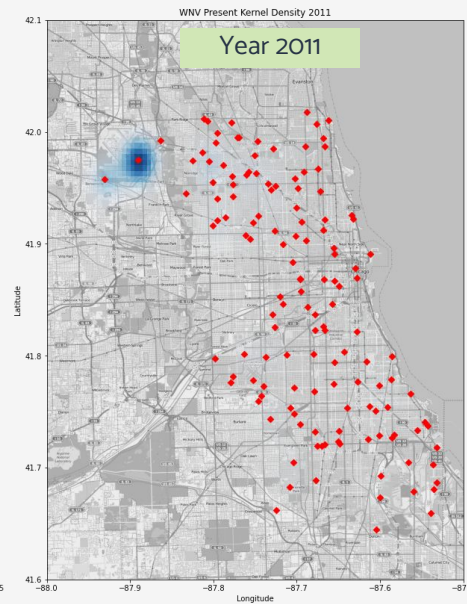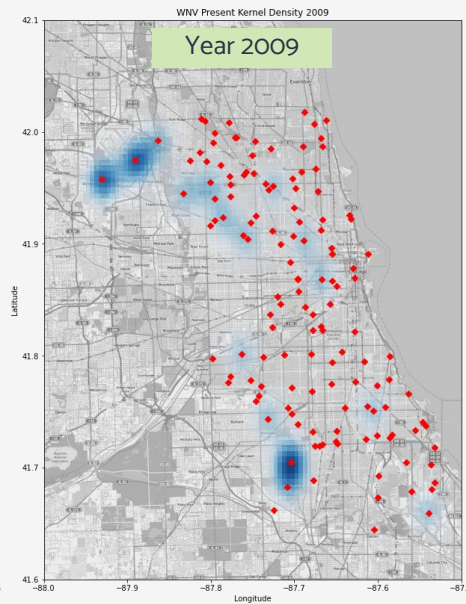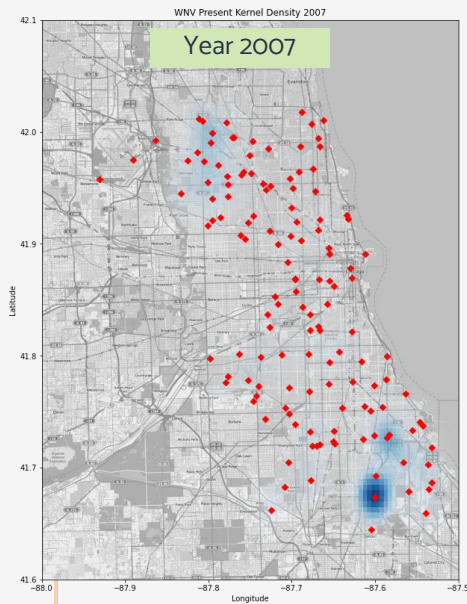
- Hotspots occurred at different locations - important to identify them ahead of time to deploy timely interventions

■ Hotspot for traps with WNV-infected mosquitos

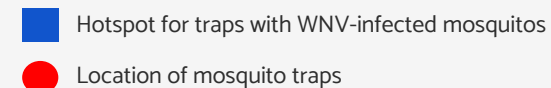● Location of mosquito traps

Targeted spraying is needed to help control population of WNV-infected mosquitoes

WNV Present with Spray 2013

Year 2013
Spray activation
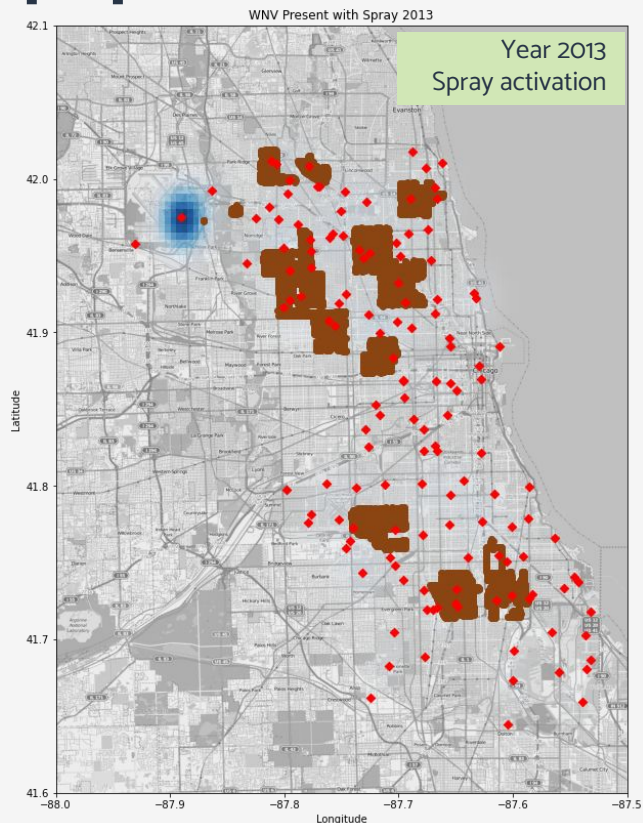
- In 2013, spraying was conducted in various areas with mosquito traps. This could have prevented them from being WNV hotspots

- The WNV hotspot area is situated within the O'Hare International Airport compound where it was not covered by the spraying efforts

■ Hotspot for traps with WNV-infected mosquitos

■ Areas with spray conducted in 2013

● Location of mosquito traps

04
Modelling & Evaluation

Classification model

Location

Inputs

Weather

```
def make_prediction(input):
    # insert random python things
    return model.predict(input)
```

Model

0: No WNV Present

Prediction
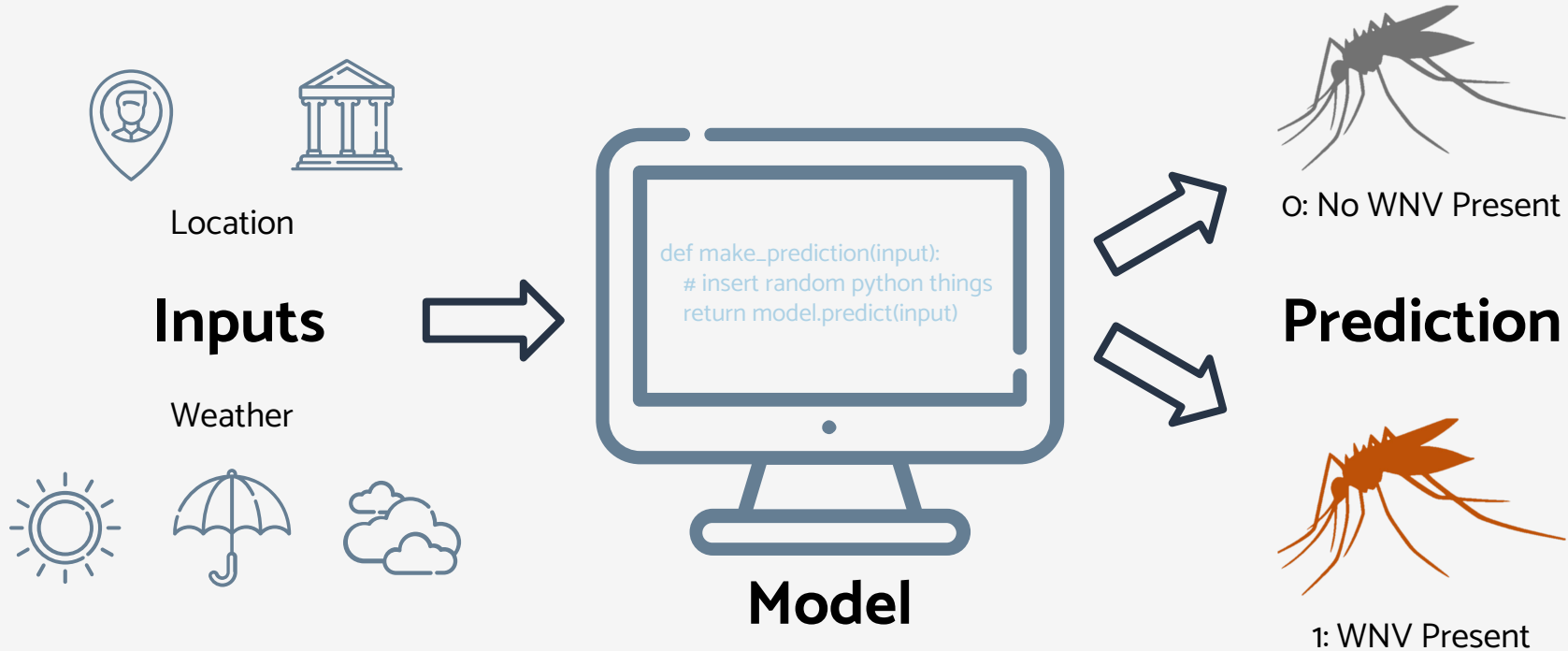
1: WNV Present

Features

40 Features

Weather, Location, Time Period

Binary Target

1: WNV Present
0: No WNV

Train-Validation Split

Stratified by target due to imbalance

GridSearchCV pipeline

1. Vectorisation

OneHotEncoder to vectorise categorical features

2. Scaling

MinMaxScaler to scale our features between (0, 1)

3. PCA

Principal Component Analysis to reduce dimensionality and collinearity

4. Sampling

SMOTE Tomek to address class imbalance

5. Classification

Train classification model to predict the binary outcome

Model selection

Class Imbalance

Accuracy is not an ideal metric
Compare performance using ROC AUC

ROC AUC

Measures the ability of model to distinguish
between classes
- ❌ 0.5 = model with random guesses
- ✅ 1.0 = model which is 100% correct

Generalizability

Minimise the variance between our Train
and Test scores

Model Performance (ROC AUC)
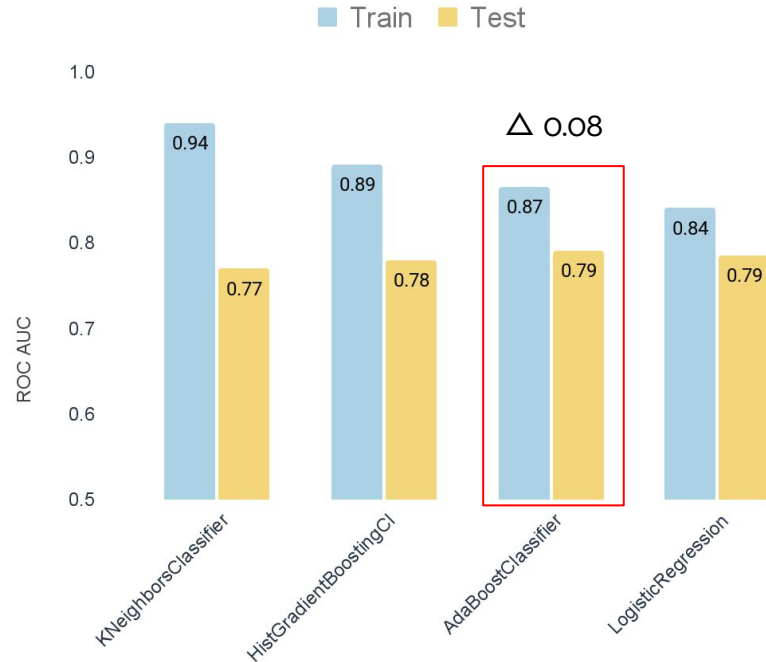
Perfect Metric : 1.0

Baseline (All Predictions 1)

	Predicted 0	Predicted 1
True 0	0	2489
True 1	0	138

AdaBoostClassifier Model

	Predicted 0	Predicted 1
True 0	1721	768
True 1	31	107
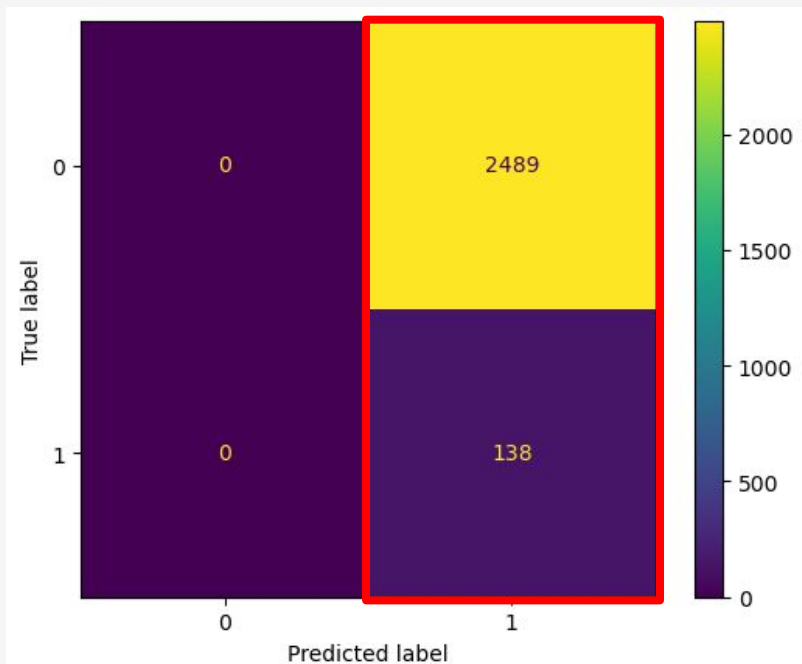
ROC AUC :   0.50
Precision :   0.05
Recall :   1.00
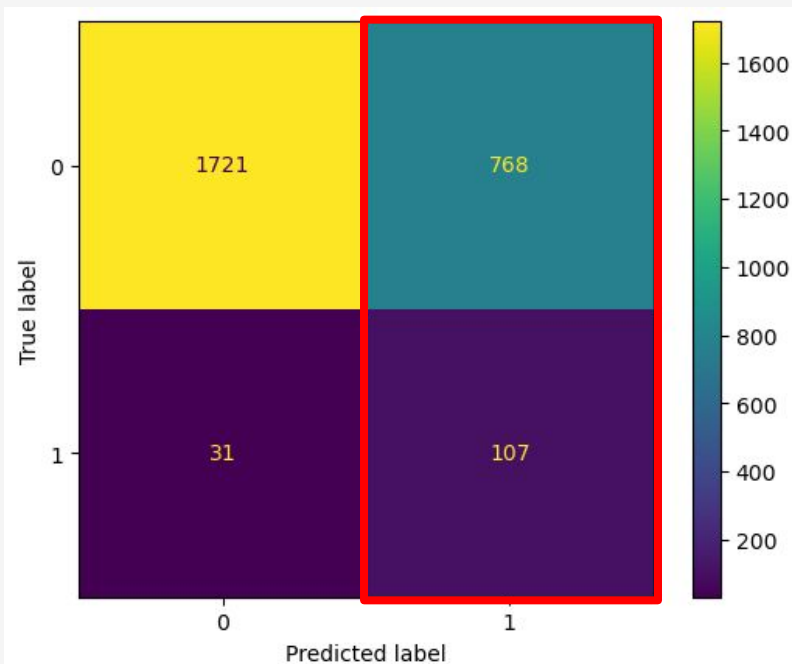
ROC AUC :   0.73
Precision :   0.12
Recall :   0.78

Model optimisation

ROC AUC Score against Threshold
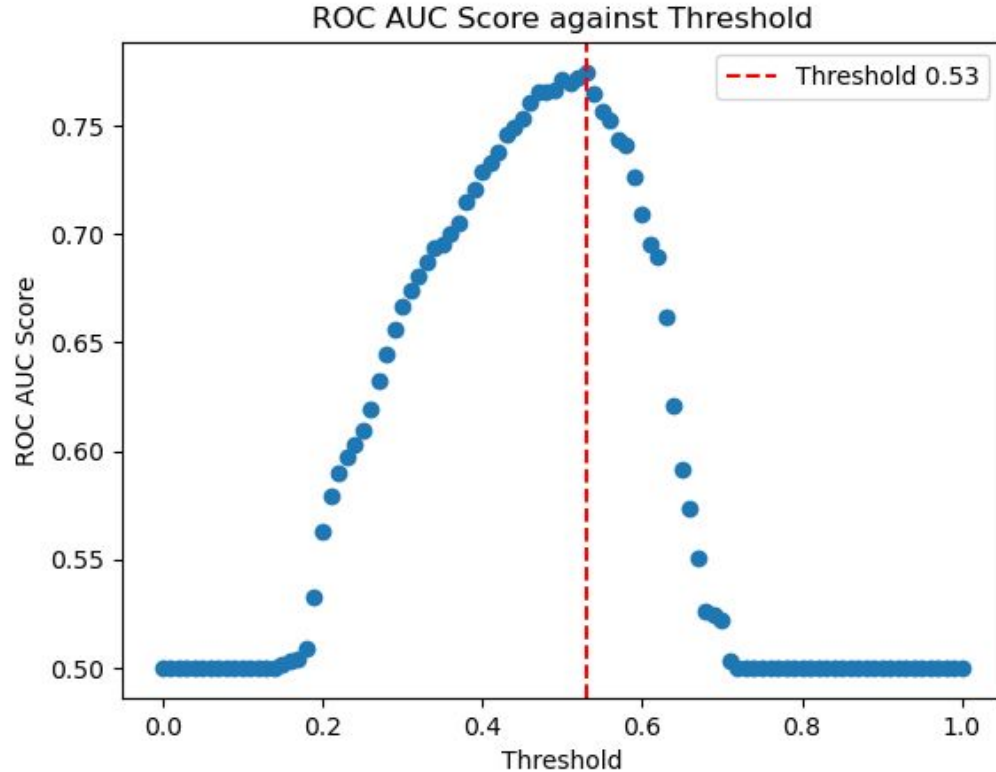
Prediction Threshold

Checked for ROC AUC across various prediction thresholds

Optimal ROC AUC when threshold is set at 0.53

Optimised model
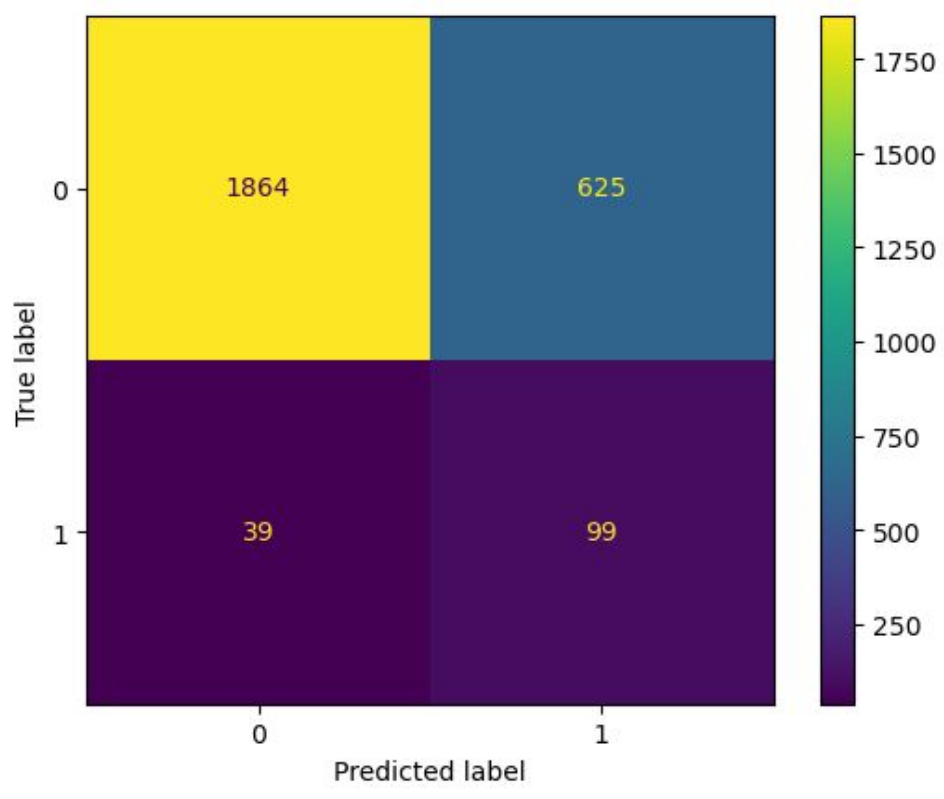
Recall score of 0.72

Adjusting the threshold allows us to maintain a Recall of 0.72, while reducing the number of false positives.

True Positives (7% reduction)
107 -> 99

False Positives (18% reduction)
768 -> 625

05
Cost-Benefit Analysis

Existing surveillance program

Location

- Areas where WNV-infected mosquitoes have been detected

Time

- At night when most mosquitoes are active

Material

- Utilizes Zenivex E4
- Effective in killing adult mosquitoes

Costs for consideration

Spray

67 cents per acre*, leading to:
- $100k for entire Chicago city limits
  - $4-5k for up to 4 neighborhoods

Initial medical costs

Median $4,617^ for less severe cases
Median $25,117^ for more severe cases

Productivity loss

$2,136 median^ loss initially
$6,771 median^ long-term loss

Long-term medical costs

Up to an average of $49,163^ in follow-up
consultations and medical fees

*Taken from _Zenivex flyer_
^ Taken from _Initial and Long-Term Costs of Patients Hospitalized with West Nile Virus Disease_

West Nile virus human cases

These figures include Chicago, Cook and DuPage counties. Approximately 66% of the positive human cases are in Chicago.

Year	Number of human cases	Number of pools tested	Number of positive pools	Total number of mosquitoes tested
2005	181	7,165	1,939	271,235
2006	129	9,428	1,984	318,386
2007	43	12,131	1,259	375,520
2008	10	9,024	587	298,995
2009	1	9,450	298	311,220
2010	47	11,491	2,086	393,279
2011	24	8,911	939	287,774
2012	229	10,162	3,182	323,497
2013	66	11,078	1,967	407,326
2014	31	9,273	990	333,489
2015	36	7,725	1,046	314,363
2016	108	6,144	1,687	219,909

Maximum number of human cases in Chicago is 66% of 181: **119***

* Average figure omitted because of lower no. pools tested

Average human cases with spraying efforts is 66% of 82: **54**

Taken from Table 2 in <u>The drivers of West Nile virus human illness in the Chicago, Illinois, USA area</u>

Option 1: Spray entire city area

0

Approximate no. positive human cases

USD 1,206,740

To spray the entire city once a week, from July to September

USD 0

Estimated initial and long-term medical cost and productivity loss

USD 1,206,740

Total cost & expenditure

- Costly spraying approach
- Likely harmful for the environment with the frequent spraying
- May harm residents with sensitive skin and sense of smell
- Limitation in identifying hot spots
- No guarantees that this will eliminate positive human cases

Option 2: No spray at all

119
Approximate no. positive human cases

USD 0
To spray the entire city once a month, from July to September

USD 5,681,655
Estimated initial and long-term medical cost and productivity loss

$ **USD 5,681,655**
Total cost & expenditure

- Most costly approach
- Higher risk of contracting WNV
- Puts more lives at risk

Option 3: Targeted spraying

15*
Approximate no. positive human cases

USD 50,105
To spray 3-4 targeted locations once a week, from July to September

USD 716,175
Estimated initial and long-term medical cost and productivity loss

+USD 440,459
Cost savings from spraying every area of the city

($) **USD 766,280**
Total cost & expenditure

- Lowest cost approach
- Minimize environmental impact from spraying
- Minimize harm to residents

*Calculated using recall score 72% of the 54 (average no. cases when there's spraying)

Cost-benefit evaluation

Recommended

	Option 1 (Spray all)	Option 2 (No spray)	Option 3 (Targeted spray)
No. cases	0	119	15
Weekly spray cost	USD 1,206,740	USD 0	USD 50,105
Medical & productivity cost	USD 0	USD 5,681,655	USD 716,175
Total cost	USD 1,206,740	USD 5,681,655	USD 766,280
Cost savings	-	-	USD 440,459*

Cost savings compared to spraying every area of the city

06
Conclusions & Recommendations

In a nutshell

Problem statement recap: Build a model with more than 70% recall to predict the period and location where mosquitoes will test positive for WNV, enabling CDPH to preemptively allocate spraying resources to curb the virus transmission.

Model selected:

AdaBoost Classifier

Train ROC AUC: **0.86**

Test ROC AUC: **0.79**

Recall: 0.**72**

Solution:

Optimise through targeted spraying approach in forecasted WNV present areas

Projected savings: **USD 440,459**

Taking preventive measures

Optimizing Insecticide Spray

- Strategically optimise frequency of spraying and target areas

Insect Repellants & Pesticides

- Install mosquito repellant screens
- Distribute insect repellants and pesticides to each home

Mosquito Control & Raise Awareness

- Campaigns to educate the public to prevent vector breeding
- More frequent trash collection (especially disposables)
- Alert neighborhoods with recorded WNV presence

Moving forward

Data collection

- Keep track of spray data
- Include other external factors, e.g. human cases and bird cases
- Control vs Spray areas

Weather conditions

- Current weather data is localised to airports
- Central city areas and neighborhoods could have differing conditions

Thank you!

Do you have any questions?

wnvchicago@mozzie.gov.ua
+91 620 421 838
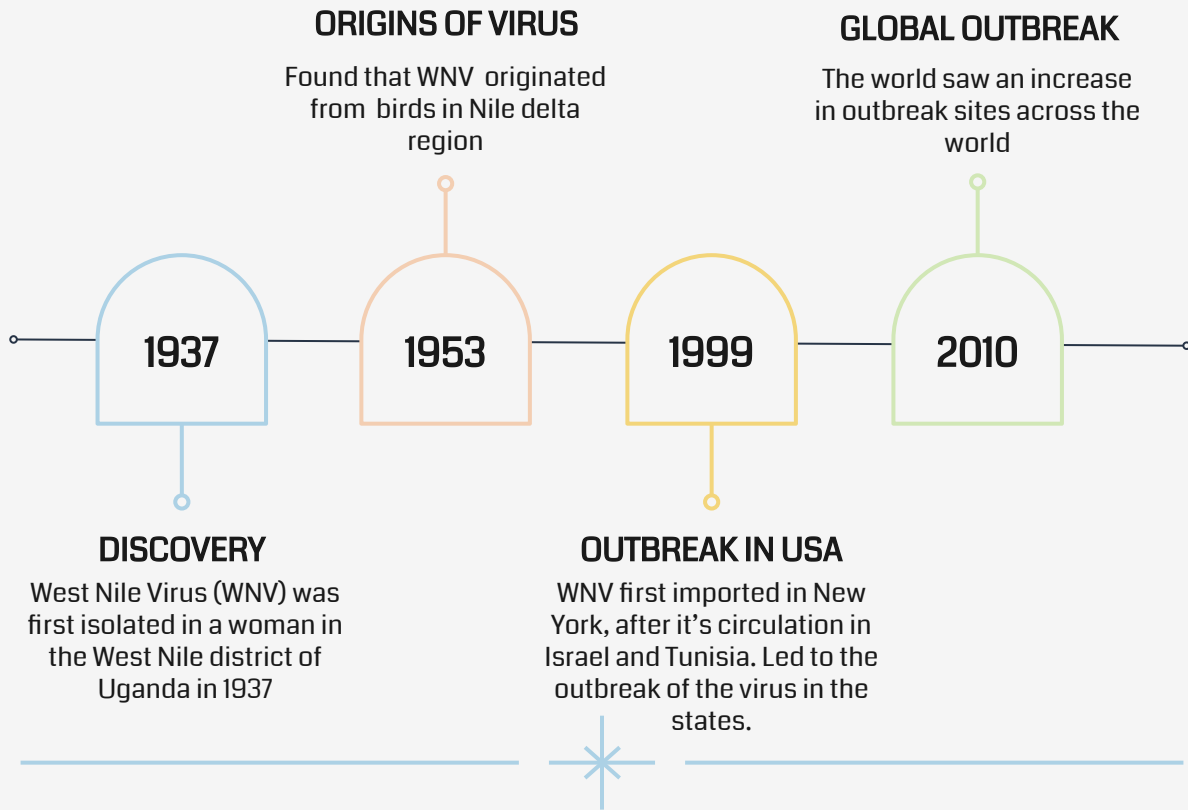wnvbegone.com

07

Appendix

Our tasks at hand

As part of the Disease and Treatment Agency hired by CDPH, we're tasked to:

1. Analyze the years with more severe WNV outbreak

2. Devise a plan to identify locations which are potential WNV hotspots

3. Optimize the use of the city's funds to curb the mosquito population and subsequently, WNV transmission

TIMELINE OF WEST NILE VIRUS (WNV)

ORIGINS OF VIRUS

Found that WNV originated from birds in Nile delta region

GLOBAL OUTBREAK

The world saw an increase in outbreak sites across the world

1937

1953

1999

2010

DISCOVERY

West Nile Virus (WNV) was first isolated in a woman in the West Nile district of Uganda in 1937

OUTBREAK IN USA

WNV first imported in New York, after it's circulation in Israel and Tunisia. Led to the outbreak of the virus in the states.

DATA CLEANING

"Null" Values Replacement	Values Calculation	Dropping columns
- Replacement with station 1 values. Not expecting too much difference across the stations which were only ~20 km apart, e.g. snowfall - Replacement with forward fill method. Not expecting major day to day differences, e.g. Preciptotal.	- Imputation of station 2 Depart values, by finding 30 year normal temperature for station 1.	- Dropping columns with only null or 0 values, e.g. depth, water1.

Principal component analysis
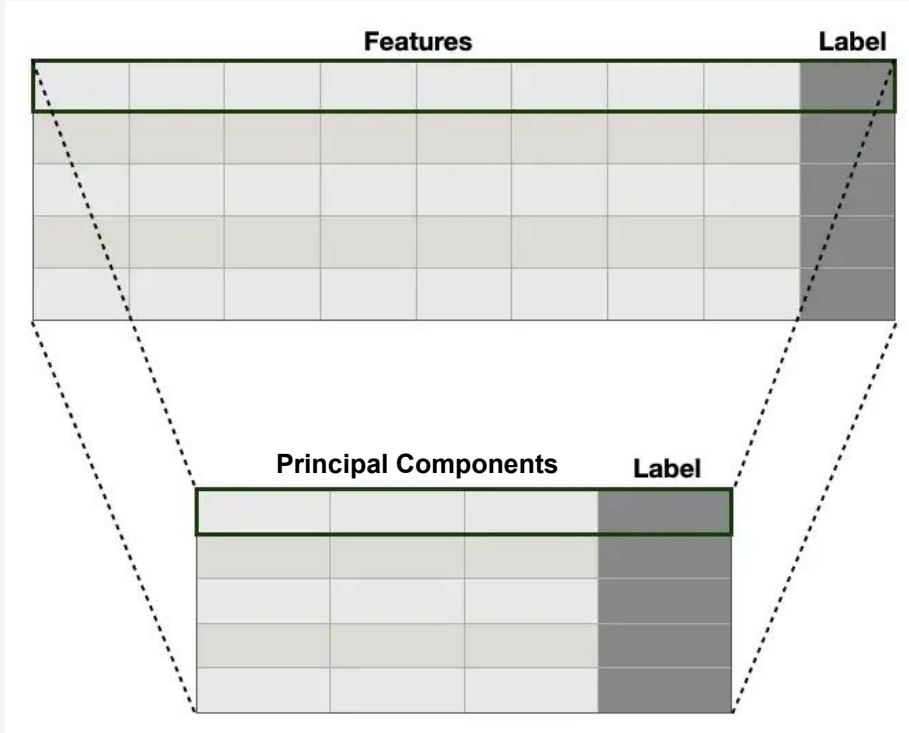
Reduce Dimensionality

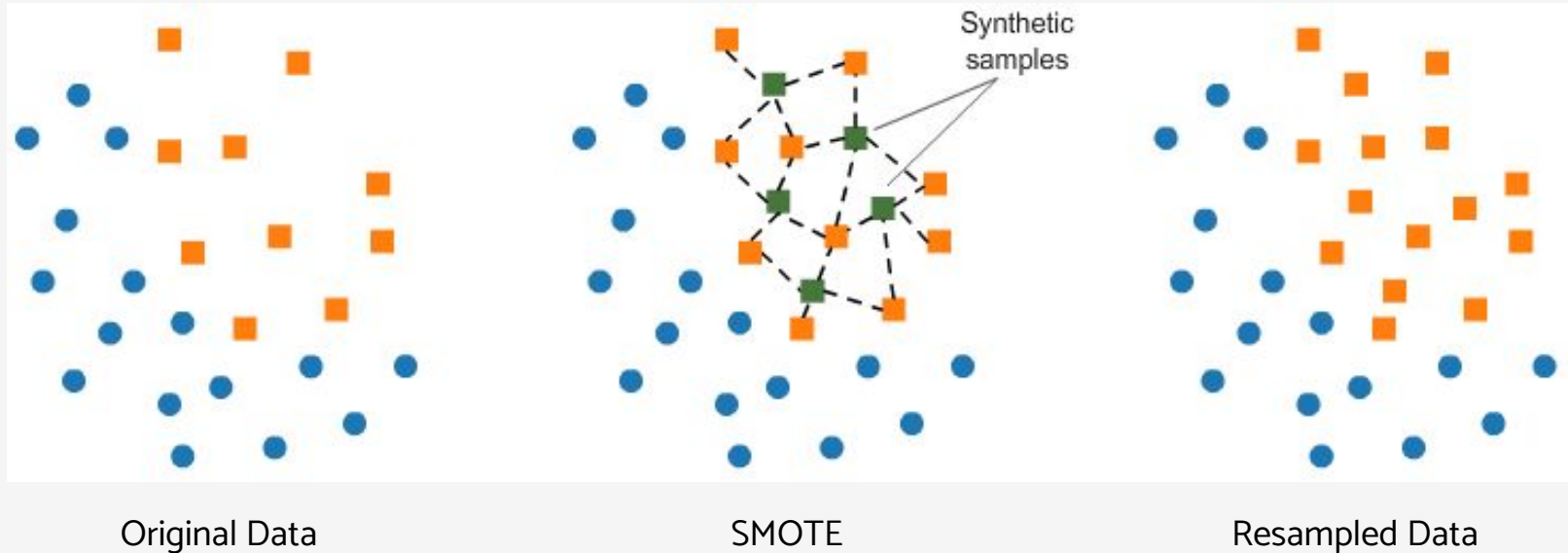Transforms features into principal components

Reduce Collinearity

Removes features that are highly correlated

Reduce Overfitting

Removes unnecessary features

Over-sampling with SMOTE

- Oversample by creating new synthetic samples for the minority class

Original Data   SMOTE   Resampled Data

Under-sampling with Tomek

- Tomek links are close pairs of instances of the opposite class
- Removing the majority instance of each pair increases the space between the two classes

Original Data   Tomek   Resampled Data
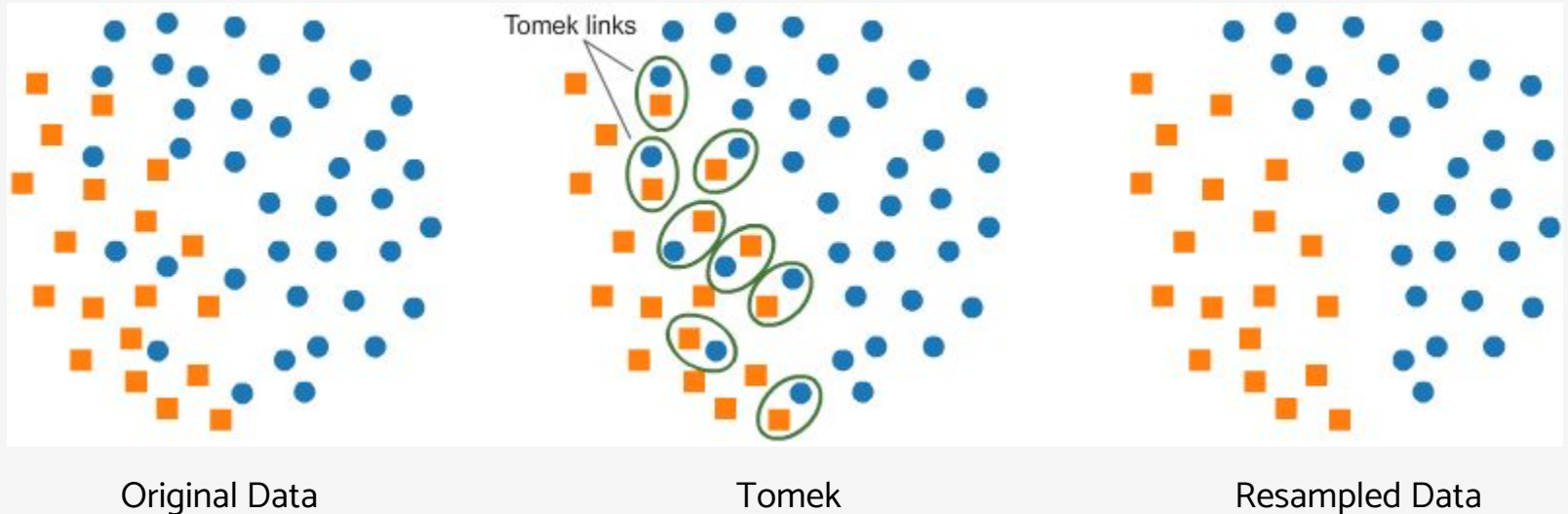
ROC

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (Recall)
- False Positive Rate

True Positive Rate (**TPR**) is the proportion of the positive class that is correctly classified and is therefore defined as follows:

TPR = TP / (TP+FN)

False Positive Rate (**FPR**) is the proportion of the negative class that is incorrectly classified by the classifier and is defined as follows:

FPR = FP / (FP+TN)

A higher TPR and a lower FPR is desirable since we want to correctly classify both classes.