Contributors
Ben Lee
Eugene
Kumeresh
Richa

DSI 18 Project 4

West Nile Virus

AGENDA

01

Introduction

- Background
- Problem Statement
- Data

02

Data Cleaning and EDA

- Merging Data
- Feature Engineering
- Preprocessing
- EDA

03

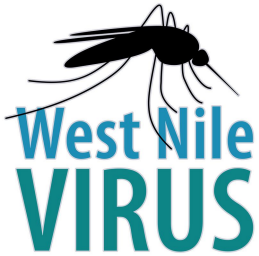
Modeling

- Tuning
- Evaluation
- Model Insights
- Limitations

04

Conclusion

- Cost Benefit Analysis
- Key Findings
- Recommendations



01

Introduction

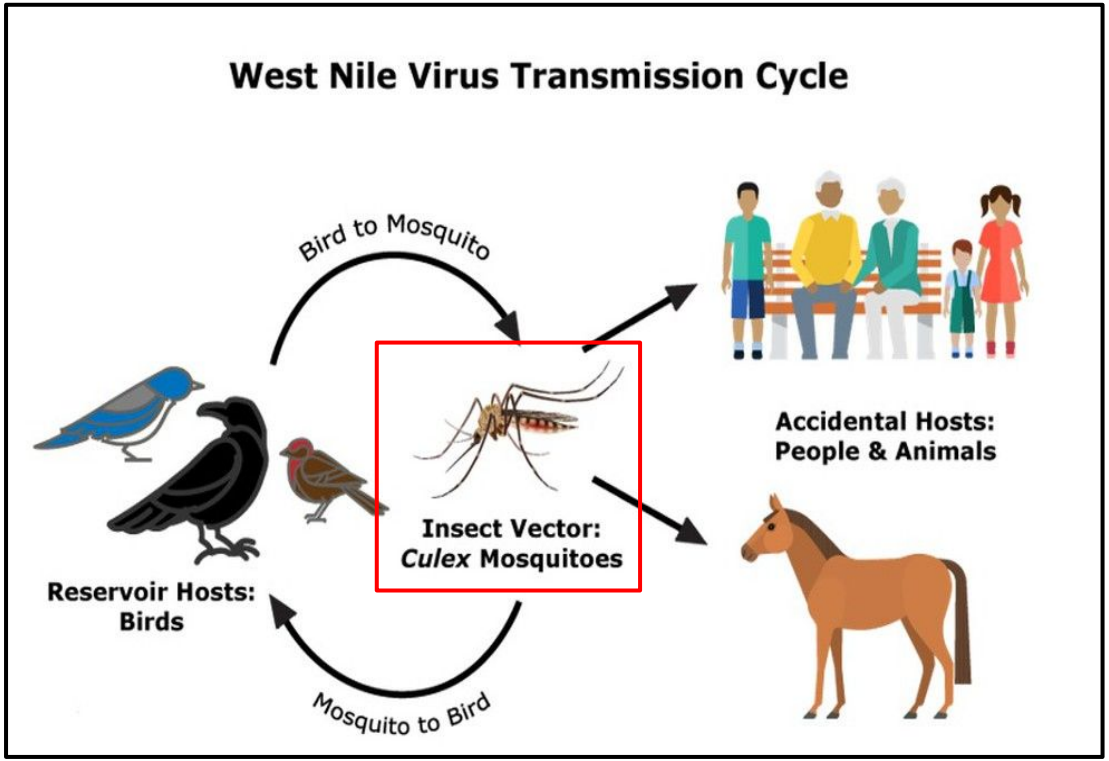
- Background
- Problem Statement
- Data

Background - The Virus

West Nile Virus (WNV) is the leading cause of mosquito-borne disease in the United States.

Around **20% of the population with the virus develops symptoms**, ranging from a persistent fever, to serious neurological illnesses that can result in death.

To date, **no vaccine or specific antiviral treatments** are available.

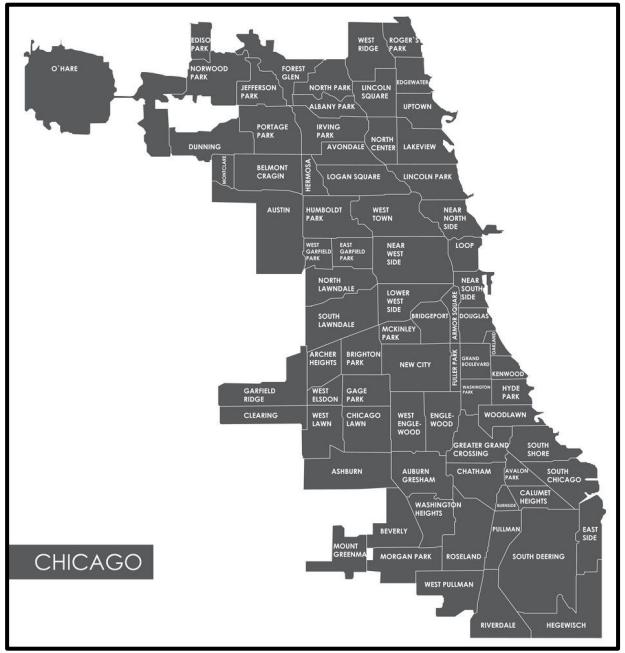


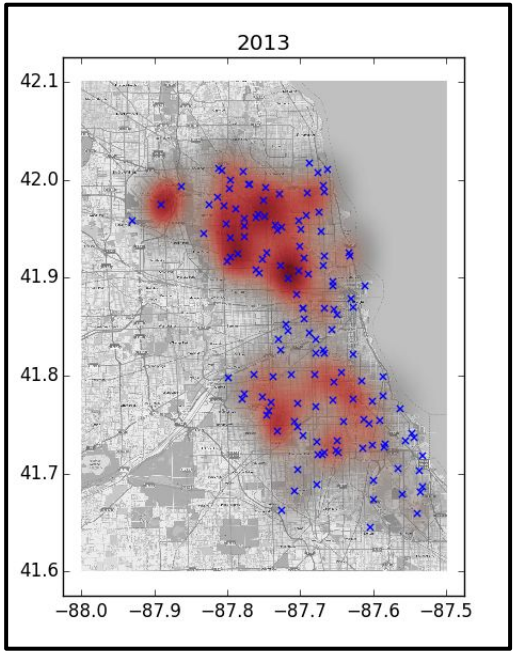
Background - Chicago

WNV first emerged in New York in 1999 and quickly spread across the country.

In **2002**, the **first human case of WNV was reported in Chicago**.

By 2004, the City of Chicago and the Chicago Department of Public Health (CDPH) established a comprehensive surveillance and control program that is still in effect today.



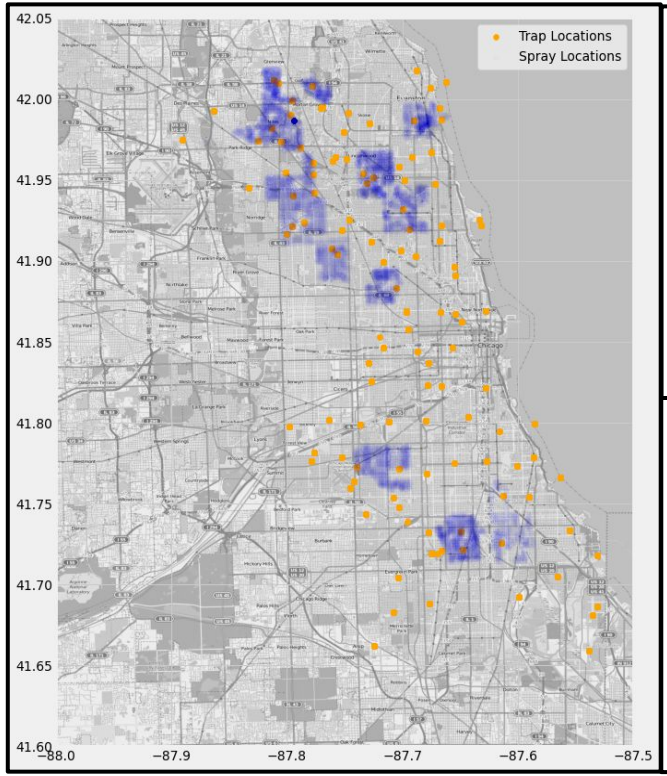


Background - Vector Control

Two methods of vector control were utilized: **(1) Mosquito Traps, (2) Pesticide Spray.**

Every week from late spring through fall (Jun-Sep), **mosquito traps were laid** across Chicago.

The **mosquitos were tested for WNV**, to aid in deciding the areas where the pesticide spray will be released, in order to reduce the number of mosquitos in the area.



Traps

Expense	*Chicken flocks*	*Mosquito traps*	*Dead birds*
Field processing			
Maintenance	$9	$4	$3
Collections	$26	$17	$22
Lab processing			
Preparation of samples	$3	$6	$7
Shipping [b]	$1	$5	$18
Lab testing per unit [c]	$72	$40	$15
Average weekly cost per unit	$111	$72	$65

Spray

FEATURES:

- Ideal for ULV applications, including urban areas
- Versatile formulation can be used diluted or undiluted
- Reduced-risk Etofenprox active ingredient
- Mosquito control adulticide for quick knockdown

Zenivex E20
$1.70/Acre

Problem Statement

As a team of data scientists from the CDPH, we have been tasked with **building a model that can accurately predict when and where mosquitos will test positive for WNV**, using weather, location, testing, and spraying data.

The model will help the City of Chicago and CDPH **prevent transmission of WNV by identifying hotspots and enabling early intervention.**

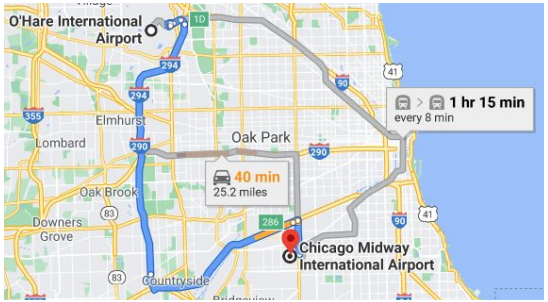
Background - Data

Train/Test

- Train: 10,506 observations with 12 features
- Test: 116,293 observations with 11 features
- No missing data
- Duplicate entries were found in the training set

Weather

- 2944 observations with 22 columns of meteorological data
- Presence of missing data
- Derived from 2 weather stations located ~20 km apart



Spray

- 14,835 observations with 4 features (across 10 days in 2 years)
- Presence of duplicate and missing data
- Does not indicate cost of spray or effectiveness of spray

02

Data Cleaning and EDA

- Merging Data
- Feature Engineering
- Preprocessing
- EDA

Data Cleaning/Feature Engineering/Preprocessing

Individual Datasets

- Spray:
  - Missing values (Time), replaced with mode of Time
  - Duplicate data dropped
- Train/Test
  - Sum of Mosquitos
  - Duplicate data dropped
- Weather:
  - Missing values ('M') were imputed from the other station

Merging Dataset

- Train + Weather + Spray:
  - Each trap record joined with weather data from closer station
  - Joining of spray data handled in Feature Engineering phase

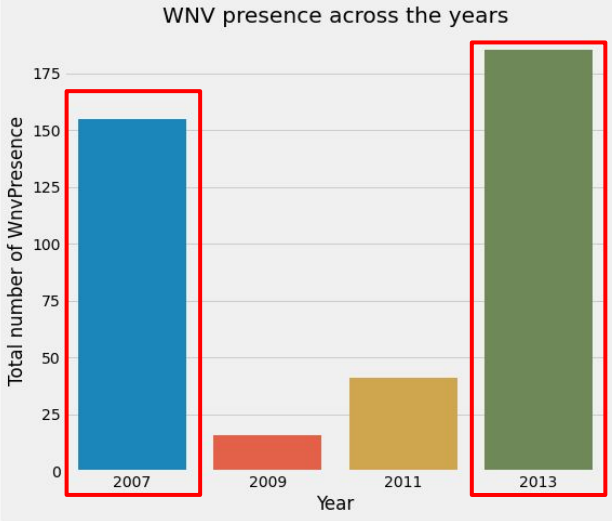
Feature Engineering

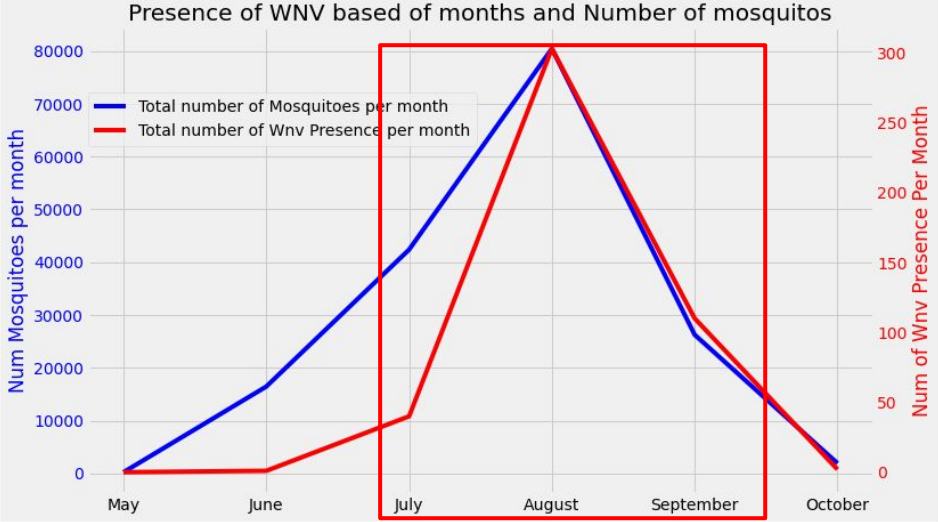
- One-hot encoding of mosquito species, and CodeSum in weather
- Historical weather data columns built with 7/14/21-rolling day aggregated effects (e.g. WetBulb, various CodeSums)
- Historical spray data columns built with arbitrary proximities (25/50/75 meter distances) to trap locations on 7/14-rolling day

Preprocessing

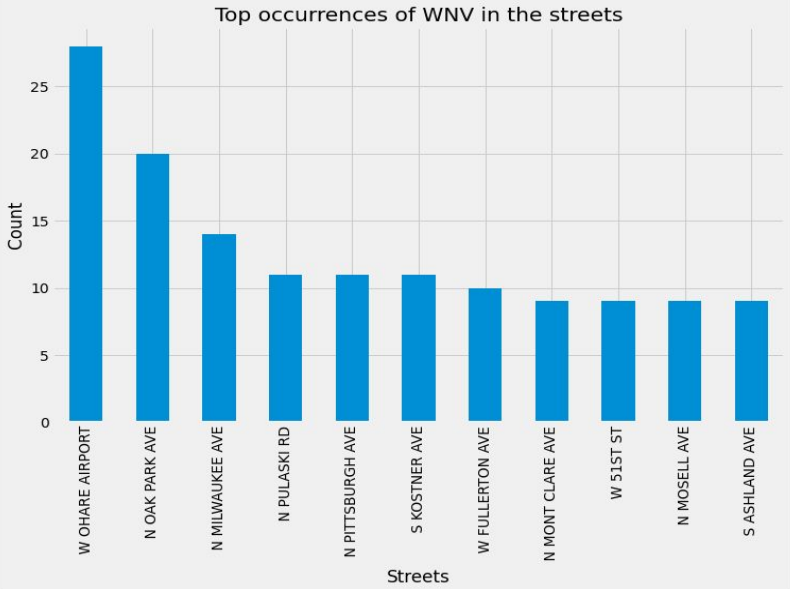
- Standard Scaler + SMOTE

EDA - Peak Periods by Year/Month

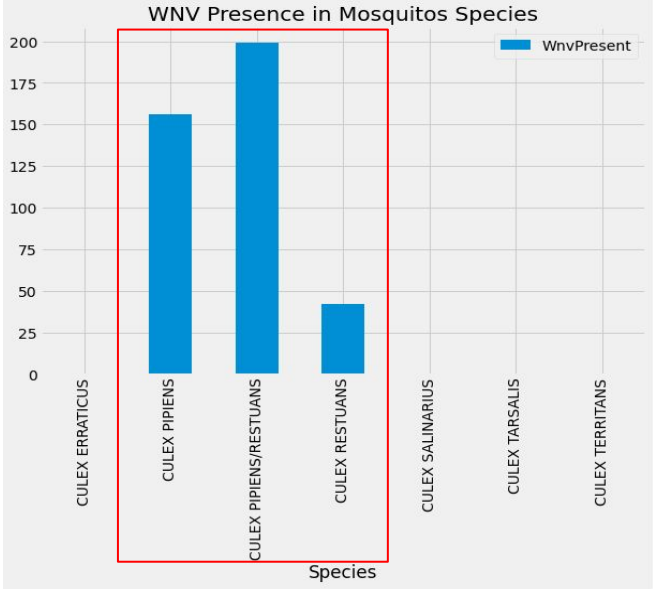




EDA - Top Occurrence and Top Species

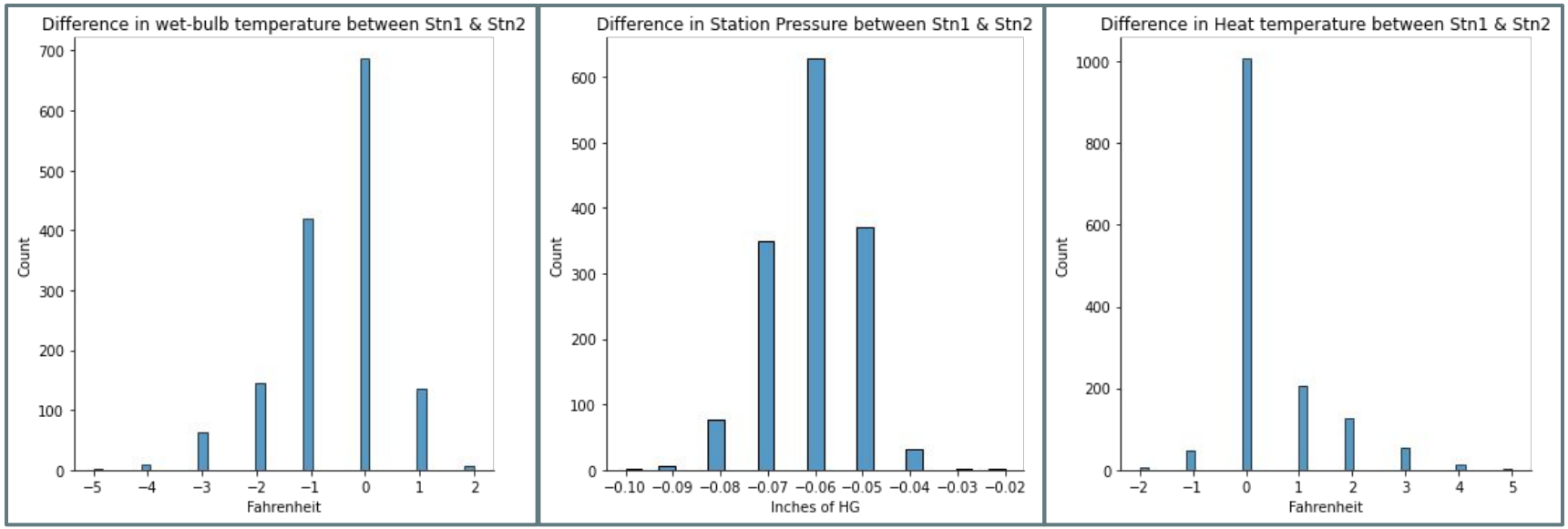


West O'Hare Airport and North Oak Park Avenue have amongst the highest counts of WNV occurrence

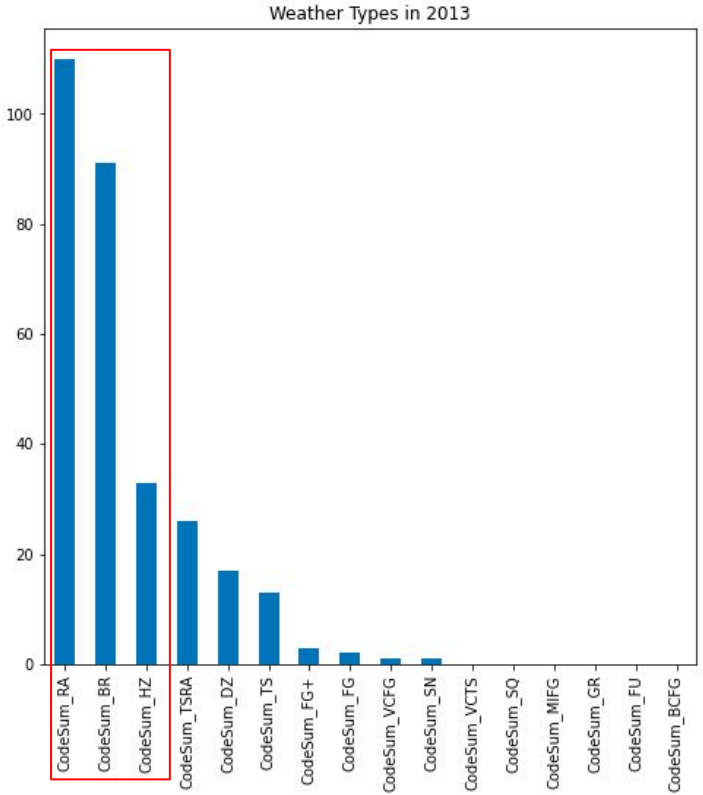


WNV mainly found in 3 species: Culex Pipiens, Culex Pipiens/Restuans, and Culex Restuans

EDA - Close Weather Readings Between Two Stations



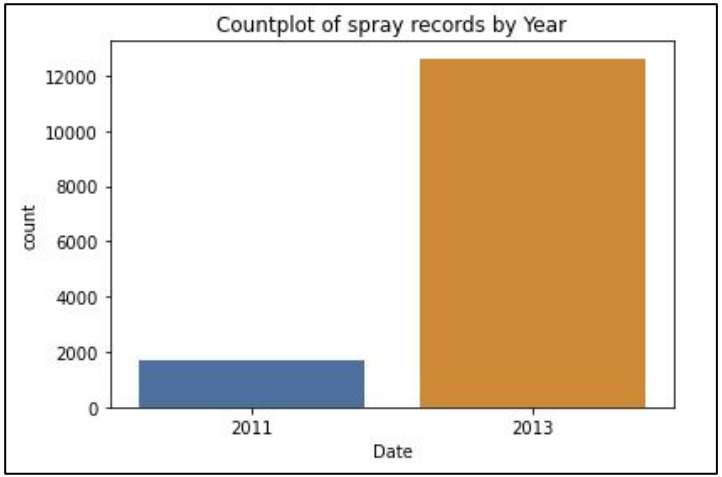
EDA - Weather Codesum

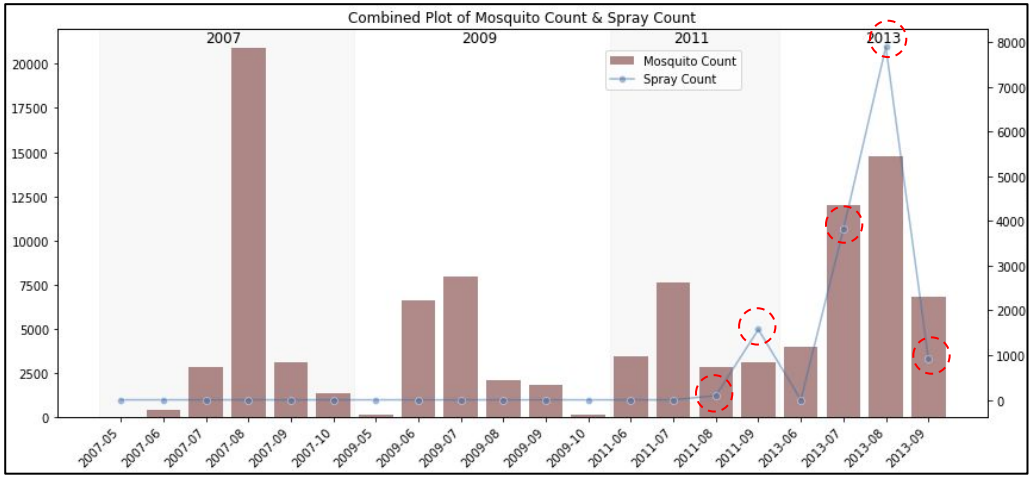


Features Retained:

- Rain
- Breeze
- Haze
- Thunderstorm + Rain
- Drizzling
- Thunderstorm
- Fog+/Fog/Partial Fog
- Snow

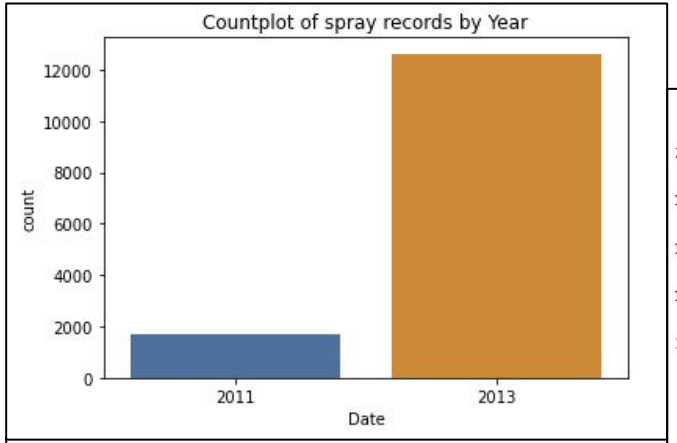
EDA - Spray Dataset

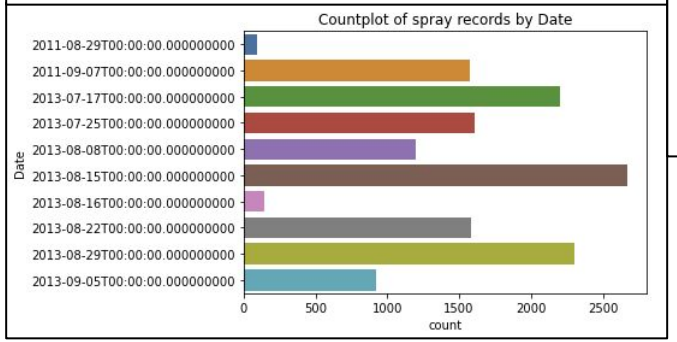


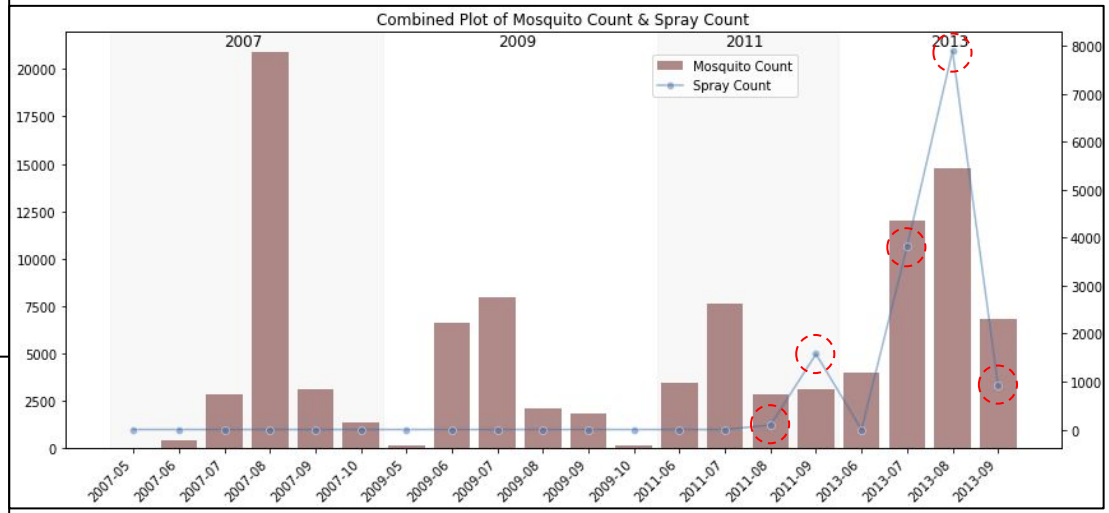


Spray records available for only 5 out of 20 months of trap data in train dataset, with no spray records associated with test dataset.

EDA - Spray Dataset

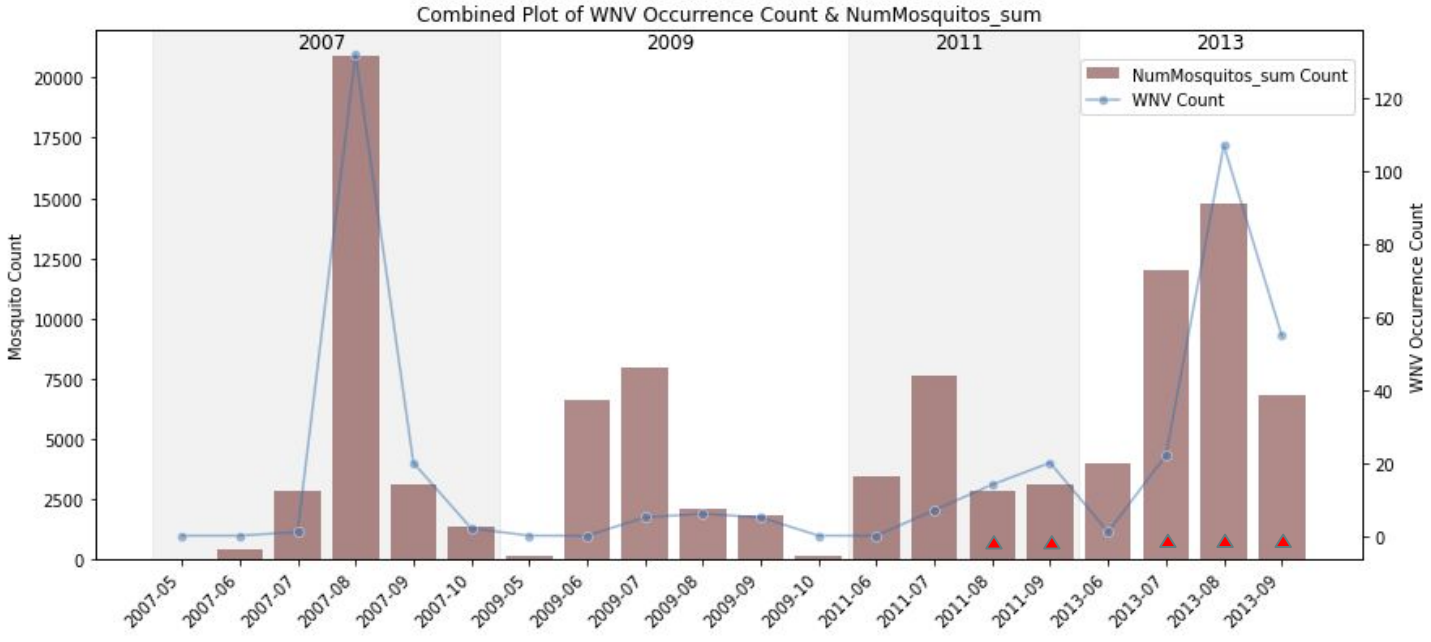






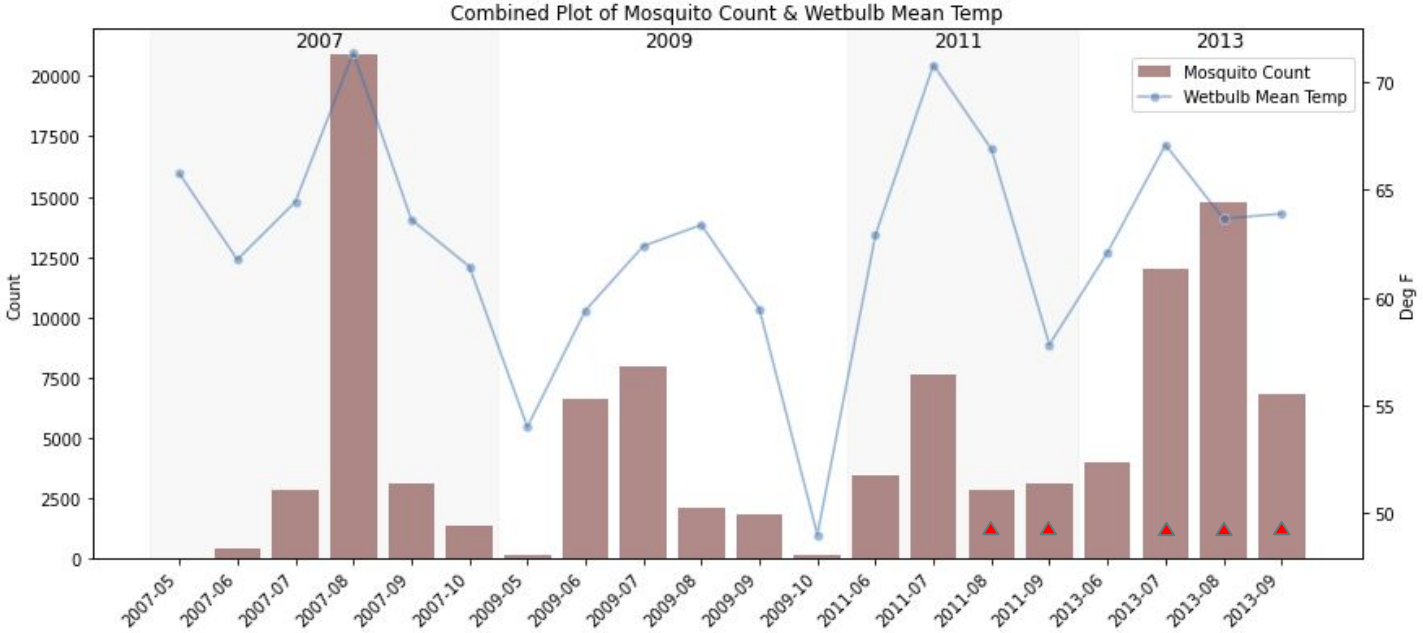
Spray records available for only 5 out of 20 months of trap data, with no spray record associated with test dataset.

EDA - Mosquito Count vs WNV Occurrences



Increased mosquito breeding typically leads to increase in WNV occurrences

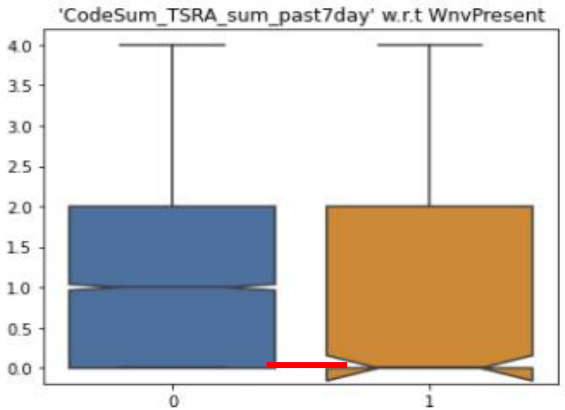
EDA - Mosquito Count vs Mean WetBulb Temperature



Months with highest mean WetBulb temperature tend to coincide with high mosquito breeding

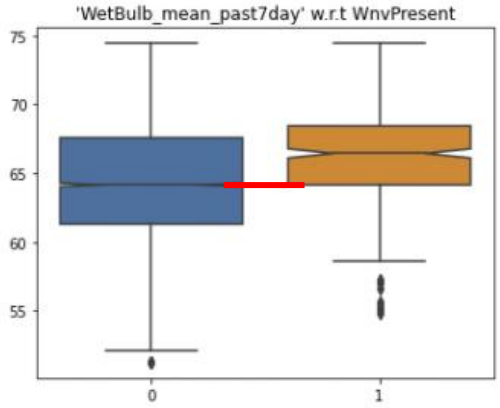
EDA - TSRA (Codesum)/ Wet Bulb vs WNV

Thunderstorm w/ Rain (TSRA) w.r.t WnvPresent



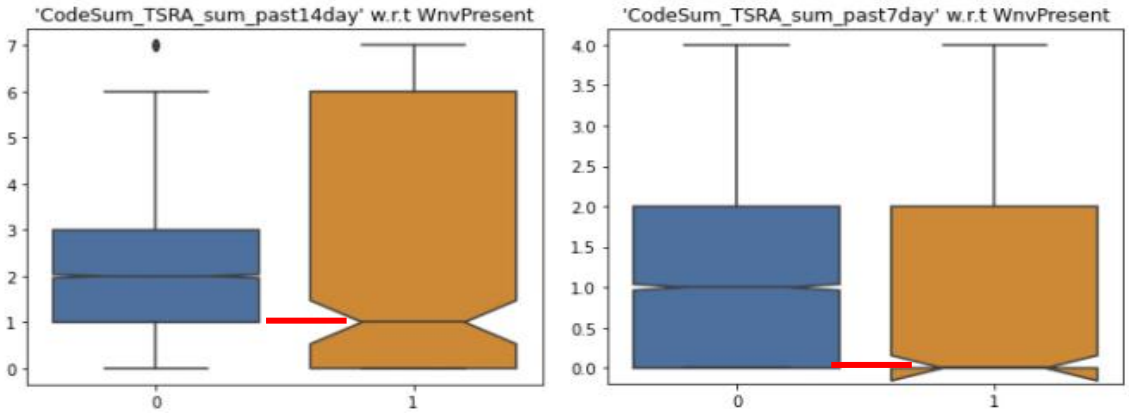
Median of WNV-positive plot in is nearly falling outside of the inter-quartile range of the WNV-negative plot at 25th percentile, giving slight indication that lower incidence of thunderstorms with rain seem to favour the occurrence of WNV in mosquitoes.

WetBulb mean temperature w.r.t WnvPresent



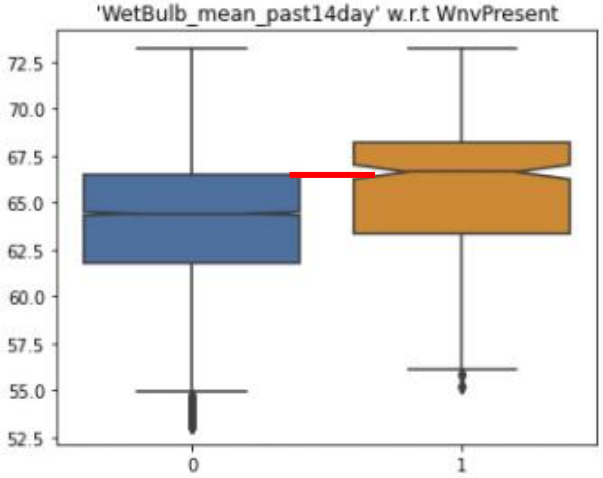
The median for WNV-negative plot is nearly falling outside of interquartile range of WNV-positive plot at 25th percentile mark, giving slight indication that lower WetBulb temperatures may lead to lesser occurrence of WNV in mosquitoes.

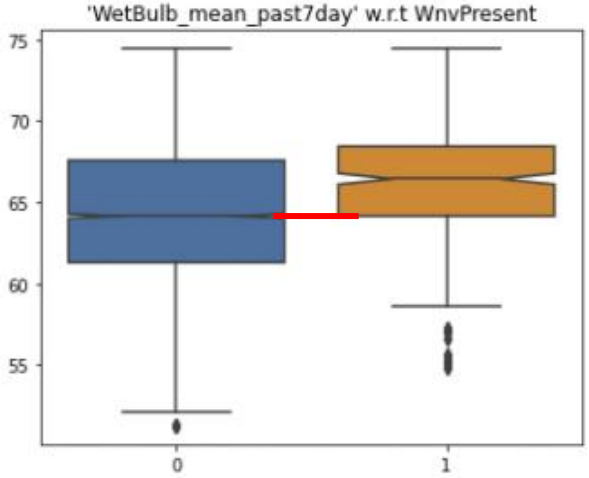
EDA - TSRA (Codesum)/ Wet Bulb vs WNV



Medians of the WNV-positive plot is nearly falling outside of the interquartile range of the WNV-negative plot at the 25th percentile for both past 14-day and 7-day boxplots, giving possible indication that slightly lower incidence of thunderstorms with rain seem to favour the occurrence of WNV in mosquitos

EDA - TSRA (Codesum)/ Wet Bulb vs WNV





03

Modeling

- Tuning
- Evaluation
- Insights
- Limitation

Models

Define Baseline Model

```
train.loc[:, 'WnvPresent'].value_counts(normalize=True)
```

```
0    0.946133
1    0.053867
Name: WnvPresent, dtype: float64
```

model
Logistic Regression
KNN
DT
Bag
RF
GB
ET
AdaBoost
SVC
XGB

+

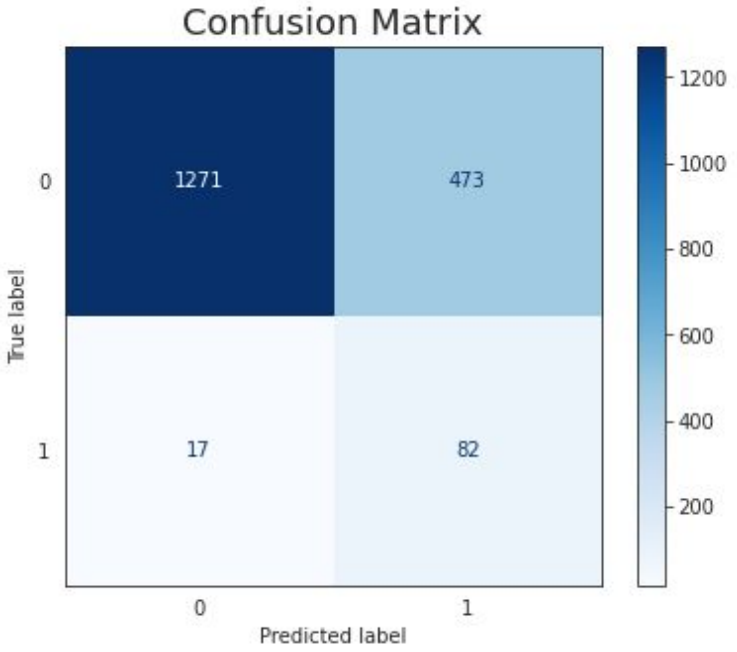
SMOTE()
Standard Scaler()

Model Evaluation

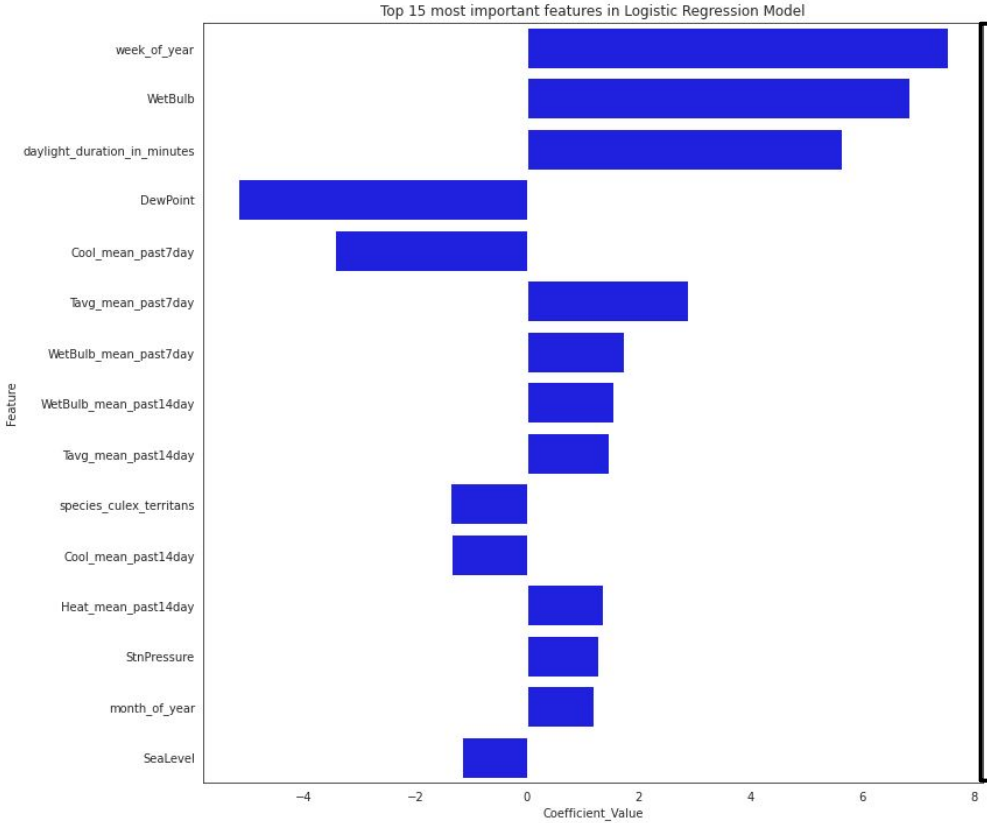
	model	params	train	test	roc	specificity	sensitivity	f_score
0	Logistic Regression	Default	0.761618	0.736300	0.779680	0.148820	0.828283	0.252308
1	Logistic Regression	{'C': 10, 'max_iter': 1000, 'penalty': 'l1', '...	0.773284	0.731416	0.781864	0.147687	0.838384	0.251135
2	KNN	{'metric': 'manhattan', 'n_neighbors': 1, 'p':...	0.994263	0.876289	0.601166	0.155080	0.292929	0.202797
3	DT	{'ccp_alpha': 0, 'max_depth': None, 'min_sampl...	0.936986	0.865979	0.605247	0.147619	0.313131	0.200647
4	Bag	{'max_features': 50, 'n_estimators': 200, 'n_j...	0.994263	0.891481	0.566320	0.141844	0.202020	0.166667
5	RF	{'max_depth': None, 'max_features': 30, 'min_s...	0.799292	0.762887	0.769910	0.156504	0.777778	0.260575
6	GB	{'learning_rate': 0.5, 'max_depth': 7, 'max_fe...	0.993402	0.912642	0.539390	0.139535	0.121212	0.129730
7	ET	{'max_depth': None, 'max_features': 50, 'min_s...	0.937560	0.872491	0.665853	0.193694	0.434343	0.267913
8	AdaBoost	{'learning_rate': 1, 'n_estimators': 100}	0.866036	0.824742	0.778774	0.195652	0.727273	0.308351
9	SVC	{'C': 10, 'gamma': 0.3, 'kernel': 'rbf'}	0.941671	0.858926	0.615811	0.148472	0.343434	0.207317
10	XGB	{'learning_rate': 0.1, 'max_depth': 7, 'n_esti...	0.992637	0.920239	0.552932	0.184211	0.141414	0.160000

Model Evaluation

- In order to prevent transmission, the **chosen model should accurately predict the presence of the WNV**
- **Primary metric** for model evaluation - **Sensitivity**
- **Production Model - Logistic Regression (with SMOTE)**
  - Train Accuracy: 0.77
  - Test Accuracy: 0.73
  - **ROC-AUC Score: 0.78**
  - **Sensitivity: 0.84**
  - Specificity: 0.15



Insights - Coefficients



Top 15 most important features:

1. Week of Year
2. WetBulb
3. Daylight Duration in Minutes
4. DewPoint
5. Cool Mean (Past 7 days)
6. Tavg Mean(Past 7 days)
7. WetBulb Mean (Past 7 days)
8. WetBulb Mean (Past 14 days)
9. TAvg Mean (Past 14 Days)
10. Species Culex Territans
11. Cool Mean (Past 14 days)
12. Heat Mean (Past 14 days)
13. Stn Pressure
14. Month of Year
15. Sea Level

Model Limitations

Limitations

- Average performance in terms of Specificity
- Model was designed based on weather and trap data
- Limited to usage within US

Recommendations

- More spray data could be collect to study the effectiveness of the spray
- Train-test datasets should be visited and recorded more frequently during the peak periods of the WNR (emphasis to Jul - Sep)
- More data can be acquired from different weather stations within Chicago to have an accurate reading in each area of Chicago

04

Conclusion

- Cost Benefit Analysis
- Key Findings
- Recommendations

Cost-Benefit Analysis

$706K

Vector Control Budget

VS

$3.1M

Potential Total Lost

[On average 50 pax for last 3 years]

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/#R6

Cost-Benefit Analysis

Total Budget: **$706K**

Traps: **$80/week**
Spray: **$1.70/acre**

Vector Control Budget

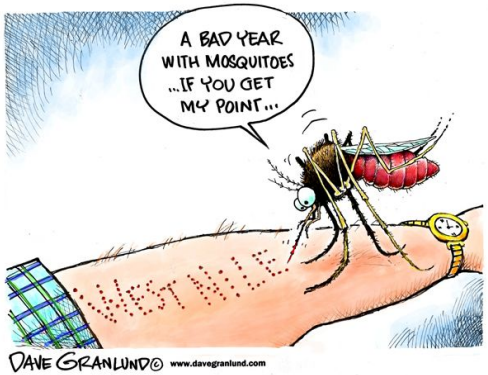
Total Lost (per pax): **$62k**

Treatment: **$46.5k**
Income Lost: **$15.5k**

Total Lost (Per Pax)
[On average 50 pax for last 3 years]

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/#R6

Key Findings & Limitations



Key Findings

- Mosquito Count and the WNV occurrence have consistent trends
- Low precipitation + warmer temperatures = High number of mosquitos = High WNV occurrences
- No clear impact on the number of mosquitos VS the spray count

Limitations

- The model is limited to detection of WNV and not other mosquito-borne diseases
- Official cost/budget were unavailable (Spray & Traps)
- Unable to determine the effectiveness of the spray
- The cost benefit of the spray cannot be measured as the model only measures against certain weather patterns

Conclusion

The **model** works best at **identifying potential hotspots** within the city, hence reducing the transmission of WNV and **allowing early intervention**. Model can be further expanded to aid cost benefit analysis if more spray data can be obtained to study the effectiveness and frequency.

Other than vector control measures, the city can consider eliminating the Reservoir Host through the removal of dead birds during the season. More vigilant monitoring on the ground will also be required during warmer periods (Jul - Sep)


WE SUSPECT THIS BIRD COULD BE CARRYING THE WEST NILE VIRUS...
LAB

Thank You!