Predicting West Nile Virus in Mosquitoes across Chicago

DSI-23: Project 4

Ray Tan
Joey Kang
Timothy Chan
Ash Ang

Contents

RAY

Problem Statement
Background
Objectives
Data Sources
Feature Engineering

TIMOTHY

Modelling
Performance
Top Predictors

JOEY

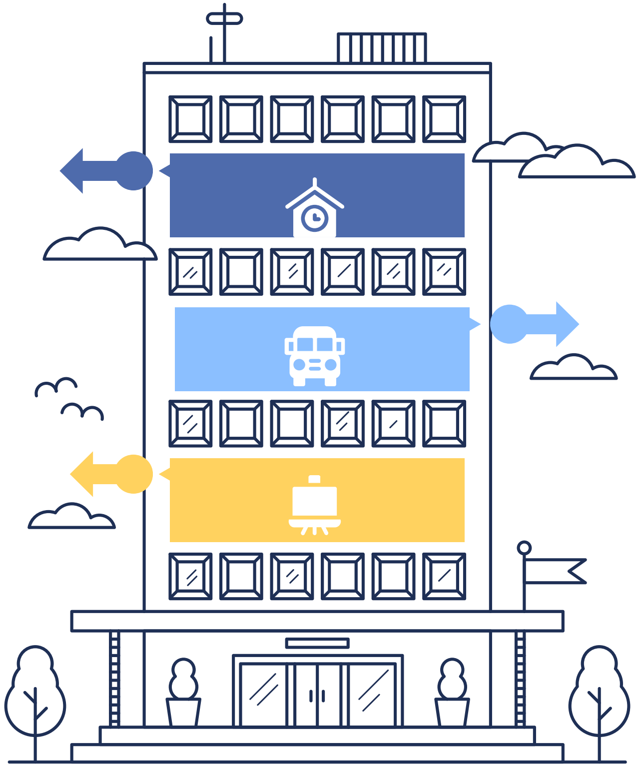
Visualisations

ASH

Cost-Benefit Analysis
Recommendations
Conclusion
Future Steps

Problem Statement

- Due to a recent outbreak of **West Nile Virus** (WNV), the **Illinois Department of Public Health** (IDPH) has set up a **surveillance and control task force** to track and curb the spread of WNV in Chicago
- As part of control efforts, IDPH has engaged our agency to assist the task force in devising a **cost-effective plan** for **deploying pesticide** throughout the city
- We aim to provide IDPH with **insights and predictions** on the spread of WNV in Chicago to help them make **sound policy decisions** surrounding **funding and deployment**



Background

WEST NILE VIRUS

- **Mosquito-borne disease** transmitted to humans by the bite of an infected mosquito
- **Leading cause** of mosquito-borne disease in the United States (US) today

80%

infected typically display few or no symptoms

20%

infected develop a fever, rash, or vomiting

<1%

infected suffer from a neuroinvasive disease

10%

mortality rate for those who had a neuroinvasive disease

USD 778 million

incurred in healthcare expenditures and lost productivity from hospitalized cases between 1999 and 2012

Objectives

We set out to achieve the following:

- Use **classification models** to predict whether a trap will test positive for WNV given **time, trap location, mosquito species, and weather data**
- Evaluate model performance using **ROC AUC, accuracy, and recall** as key metrics
- Recommend a **suitable model** for prediction
- Perform a **cost-benefit analysis** to determine the trade-off between spraying pesticide and the number of WNV cases
- Recommend a cost-effective plan to guide **when, where, and how frequent** should pesticide be sprayed

Data Sources

TRAP (TRAIN) DATASET

Years: ‘07, ‘09, ‘11, ‘13

Features:

- Date of WNV Test
- Species of Mosquito
- Trap ID
- Trap Coordinates
- etc.

Target: Presence or Absence of WNV

WEATHER DATASET

Years: ‘07 to ’14

Features:

- Date of Reading
- Temperature
- Pressure
- Precipitation
- Wind Speed
- etc.

SPRAY DATASET

Years: ‘11, ‘13

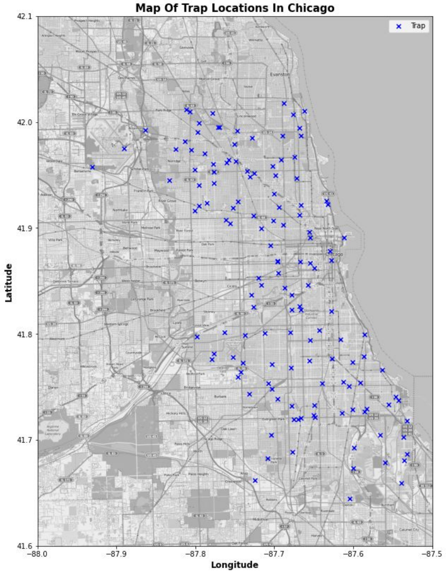
Features:

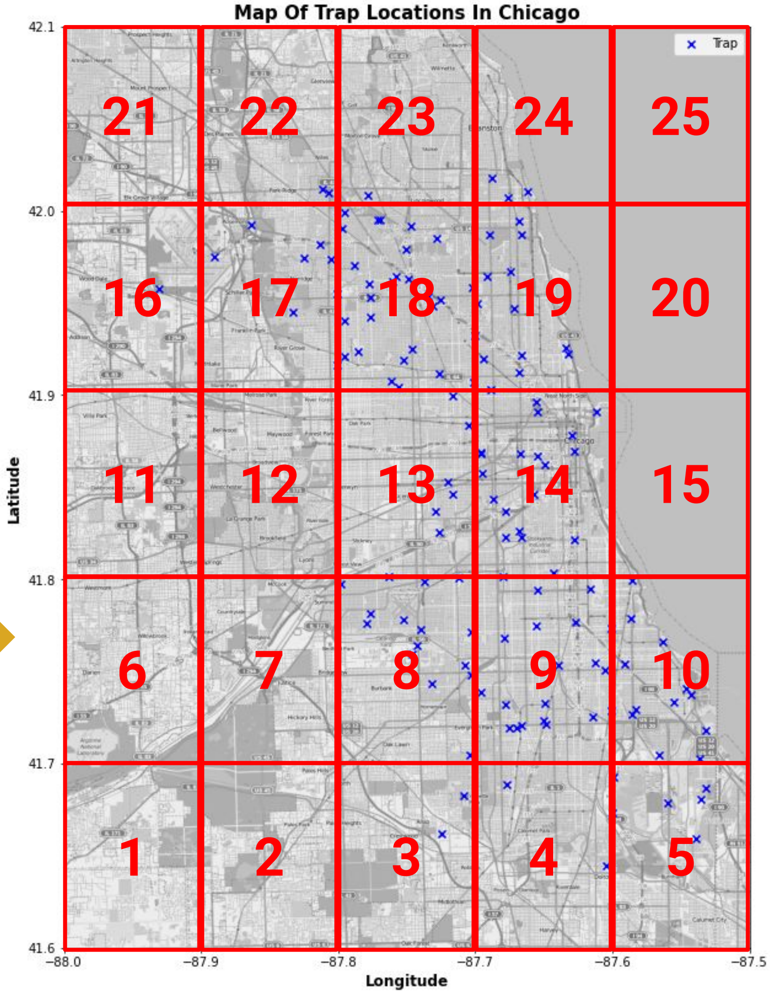
- Date of Spray
- Spray Coordinates
- etc.

Feature Engineering

TRAP DATASET

Placed traps into **25 Coordinate Groups** in total to incorporate location as a predictor of WNV for the model





Feature Engineering

WEATHER DATASET

Engineered a new feature to reflect the wetness/dryness on a given day:

Relative Humidity

from Average Temperature and Dew Point Temperature

RH = 100 - 5 * (Tavg - Tdp)

Reason being to investigate the effect of humidity on the presence or absence of WNV

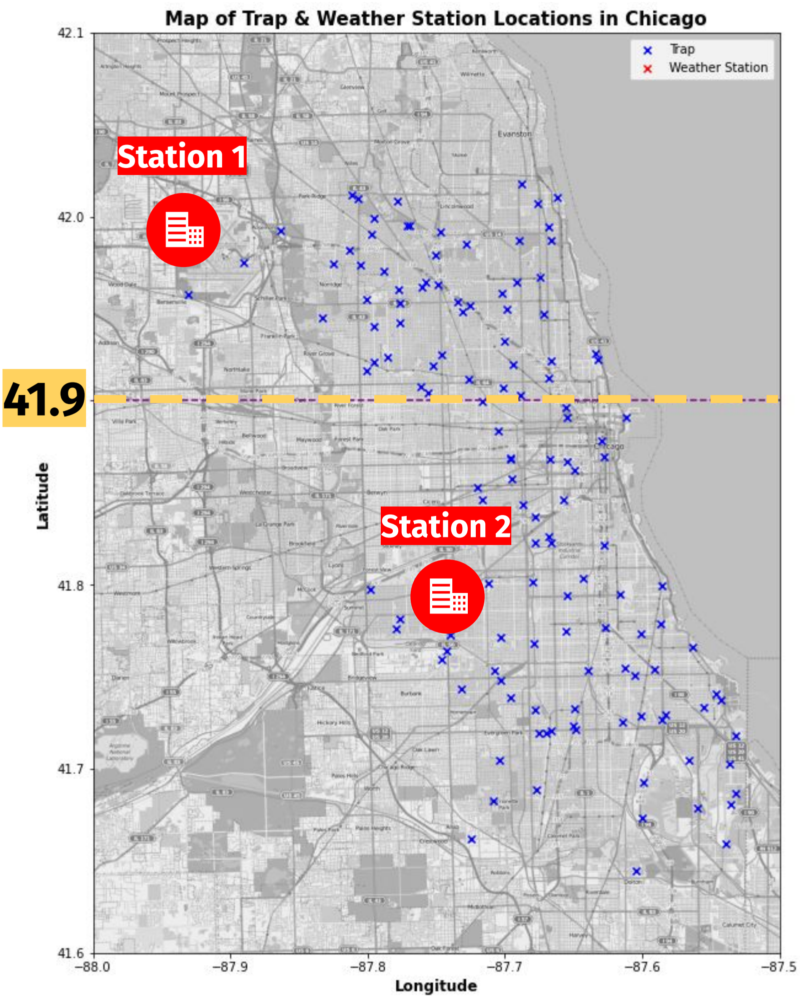
Feature Engineering

COMBINING TRAP DATASET AND WEATHER DATASET

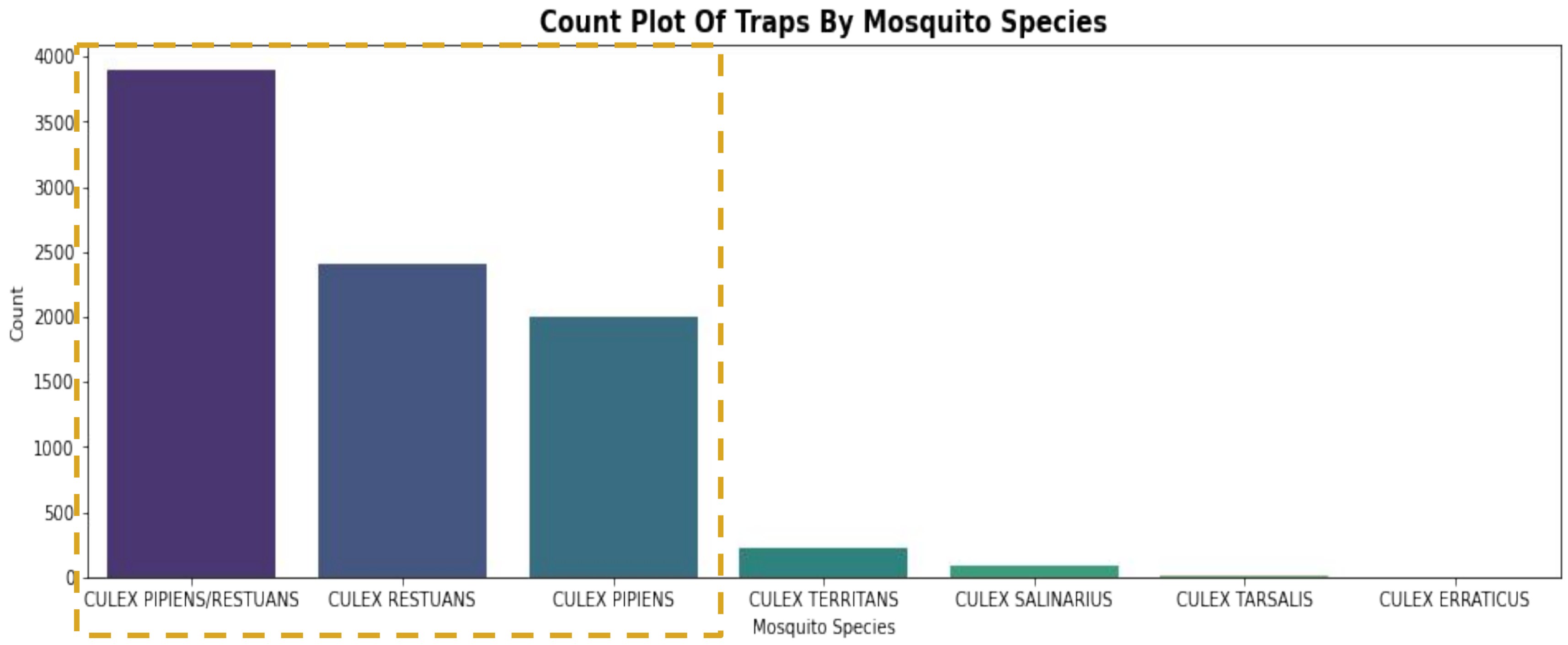
Assigned traps to the **Nearest Weather Station** to get weather reading data for that trap on a specific day

Trap has Latitude > 41.9
→ Assign to Station 1

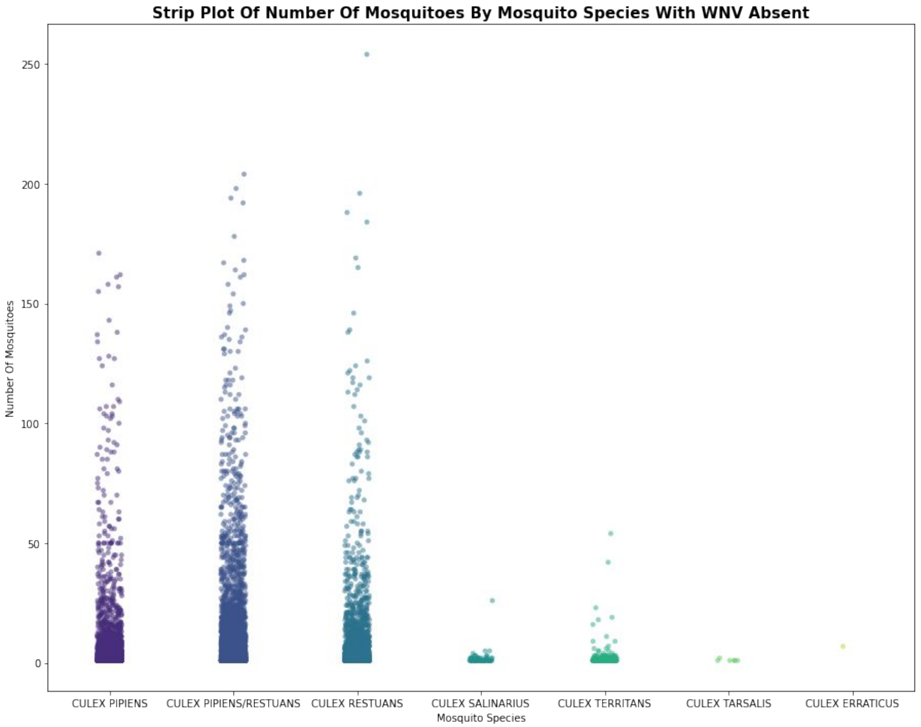
Trap has Latitude <= 41.9
→ Assign to Station 2

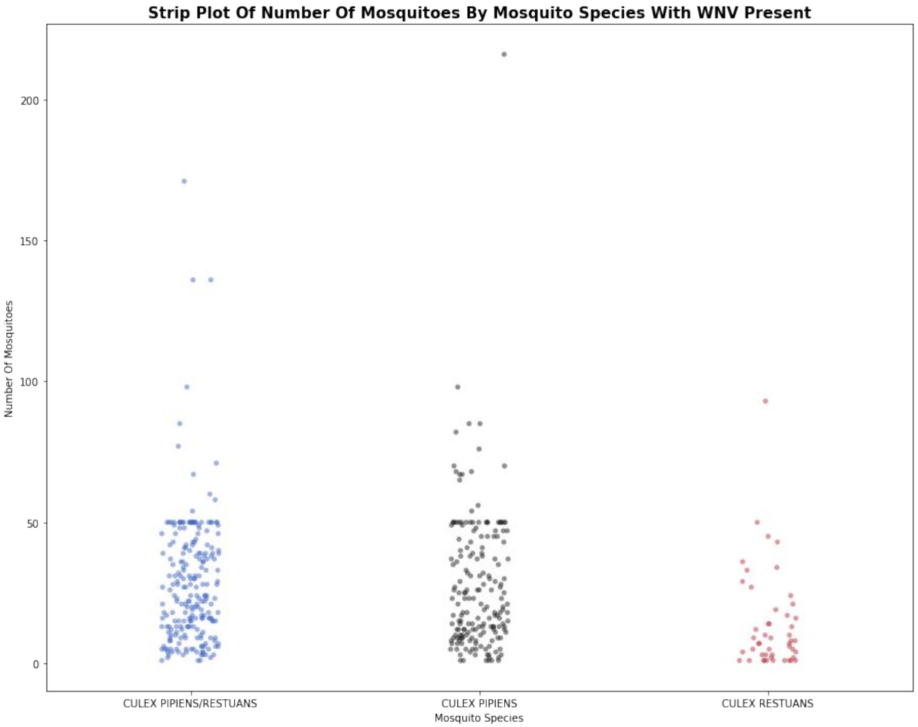


Most of the mosquitoes in the traps were Culex Pipiens/Restuans

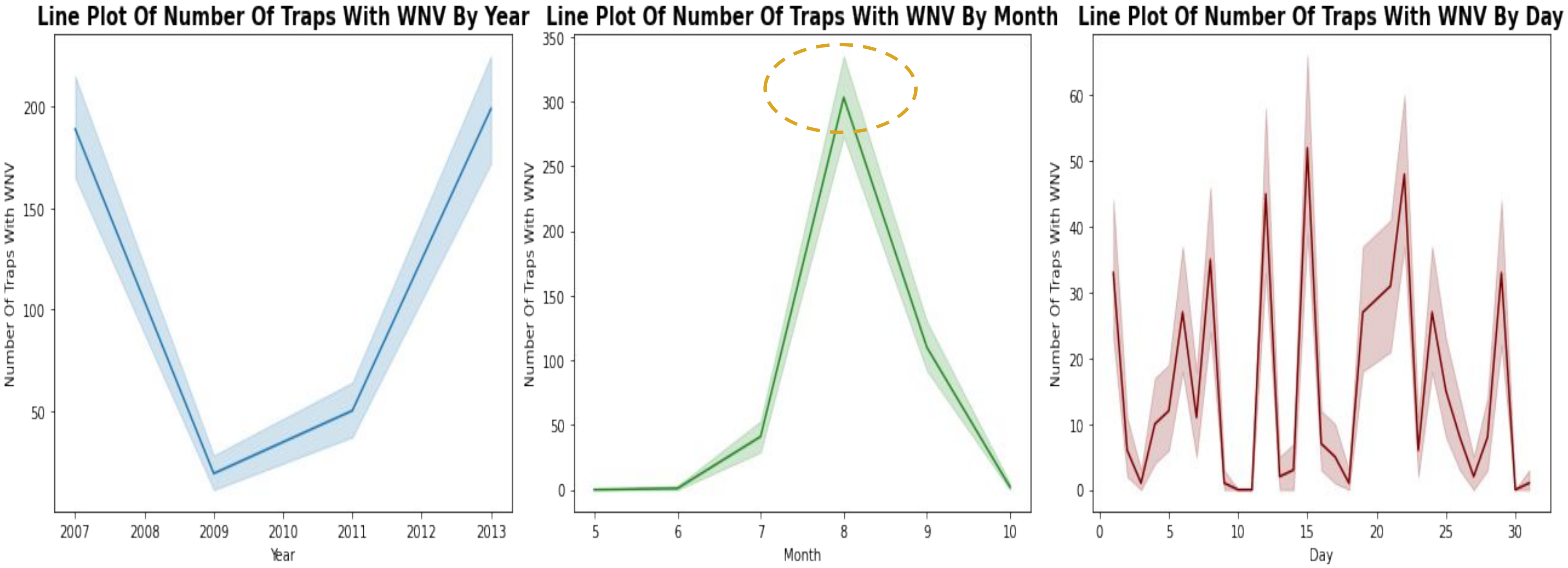


WNV is only present in Culex Pipiens/Restuans

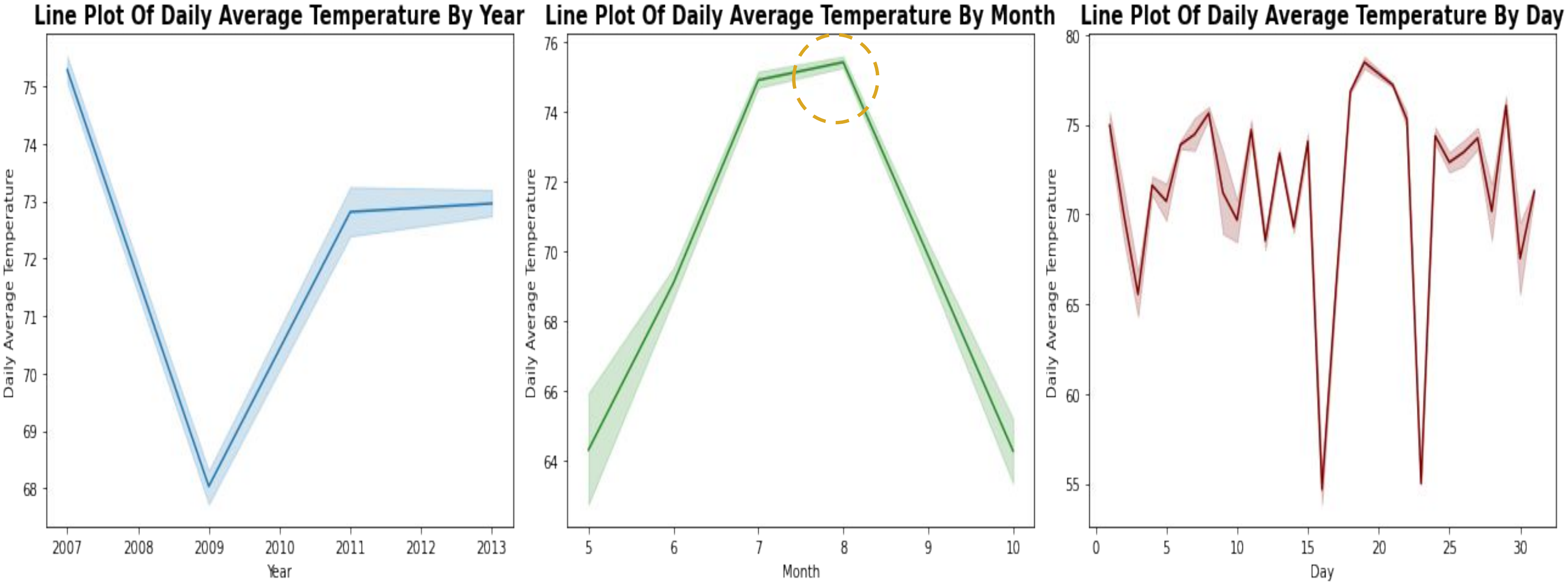




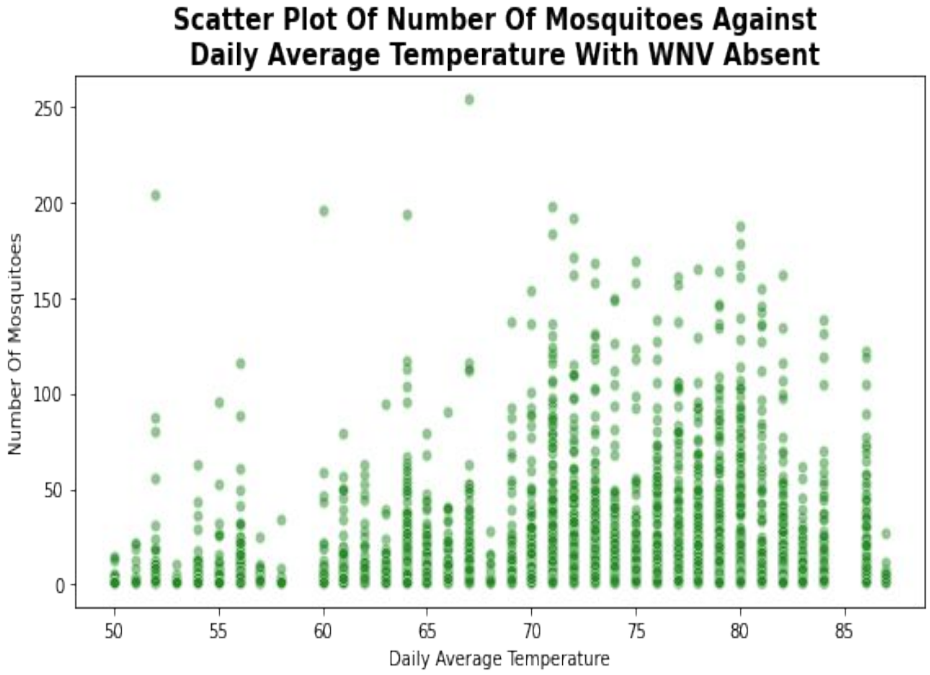
Incidence of WNV is highest in August

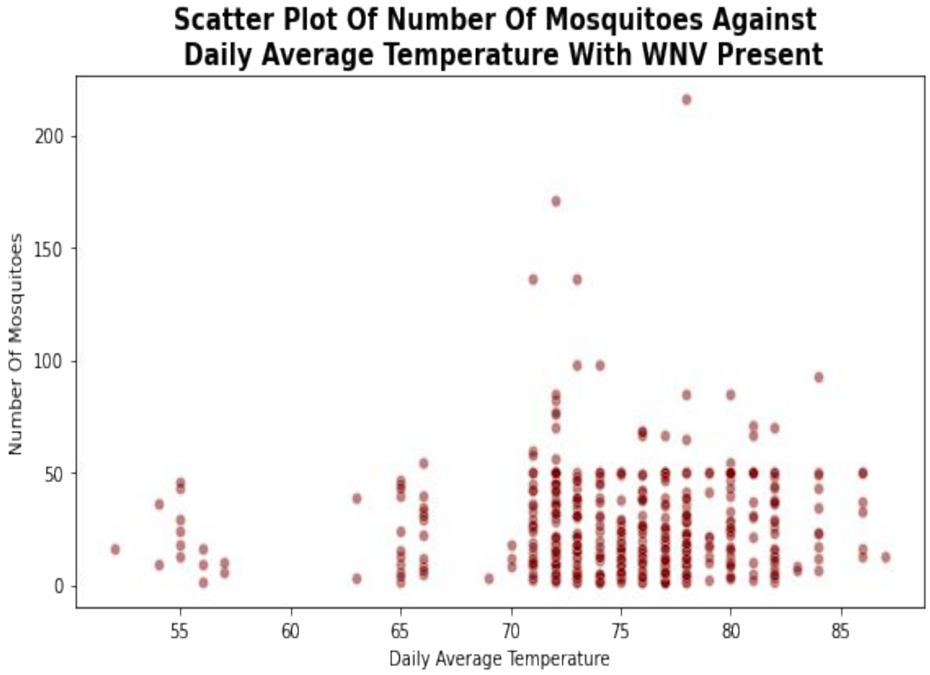


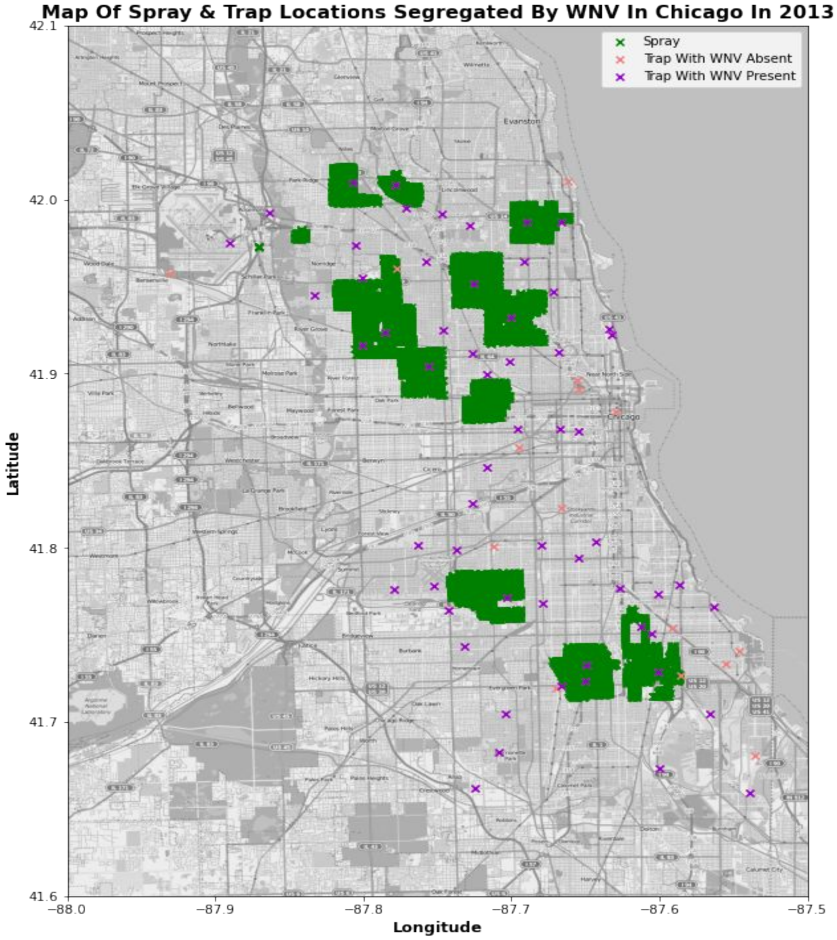
Daily Average Temperature is also highest in August

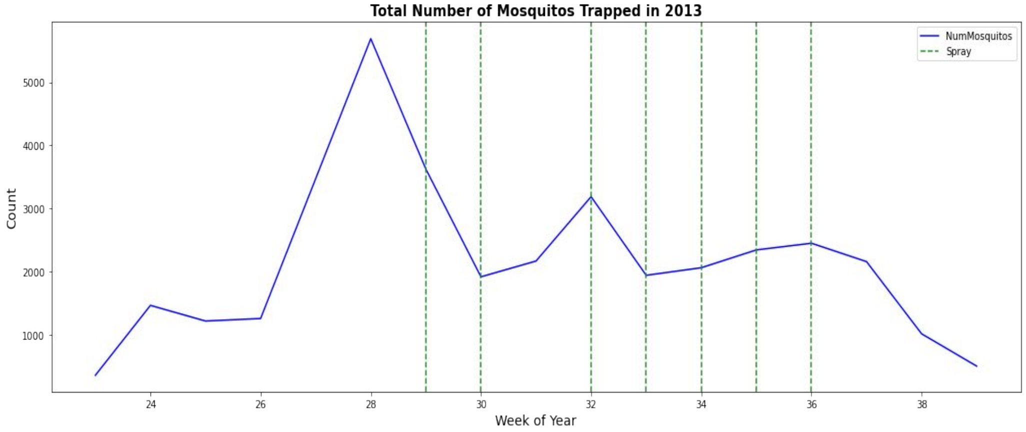


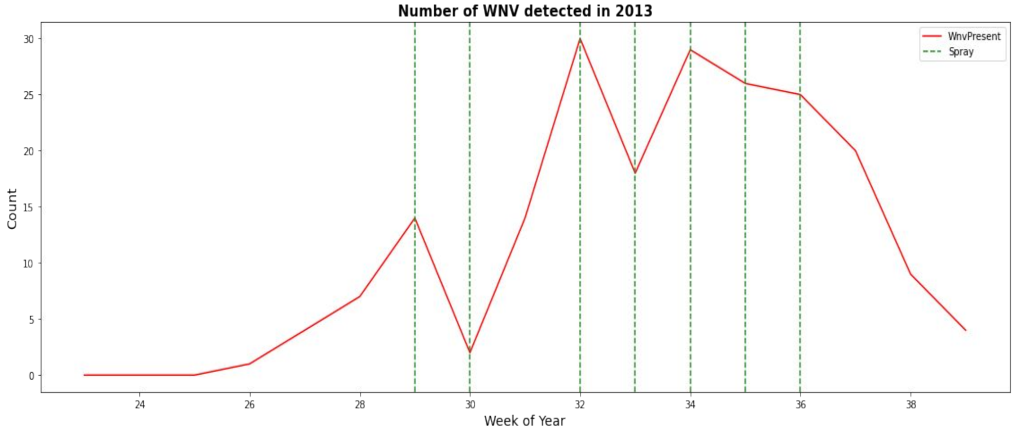
WNV appears to be present more frequently at higher daily average temperatures



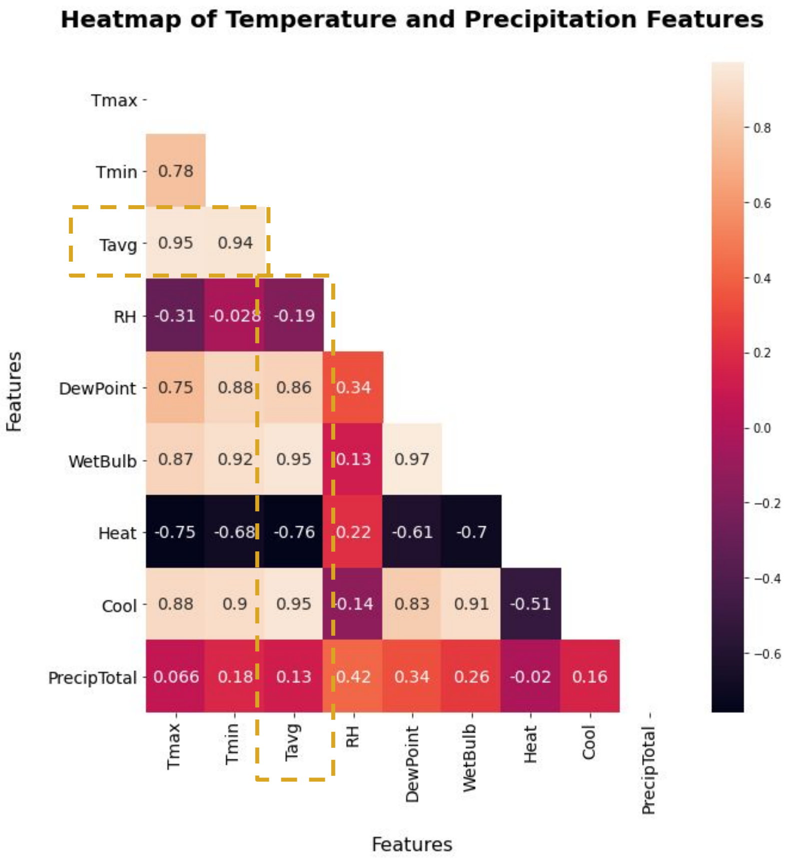



Map Of Spray & Trap Locations Segregated By WNV In Chicago In 2013
Spray
Trap With WNV Absent
Trap With WNV Present
Latitude
Longitude
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5
Evanston
Chicago


Total Number of Mosquitos Trapped in 2013
NumMosquitos
Spray
Count
Week of Year
5000
4000
3000
2000
1000
24
26
28
30
32
34
36
38


Number of WNV detected in 2013
WnvPresent
Spray
Count
Week of Year
30
25
20
15
10
5
0
24
26
28
30
32
34
36
38

Feature Selection



Dropped the following variables due to strong multicollinearity with Tavg:

- Tmax
- Tmin
- DewPoint
- WetBulb
- Heat
- Cool

Kept:
- Tavg
- RH
- PrecipTotal

Classification Model Choices

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- K Nearest Neighbours (KNN)
- Random Forest (RF)

Modelling Process

Partition data using 70/30 train-test split

Fit and train the classification model using train data

Predict classification results for ‘WnvPresent’ using test data

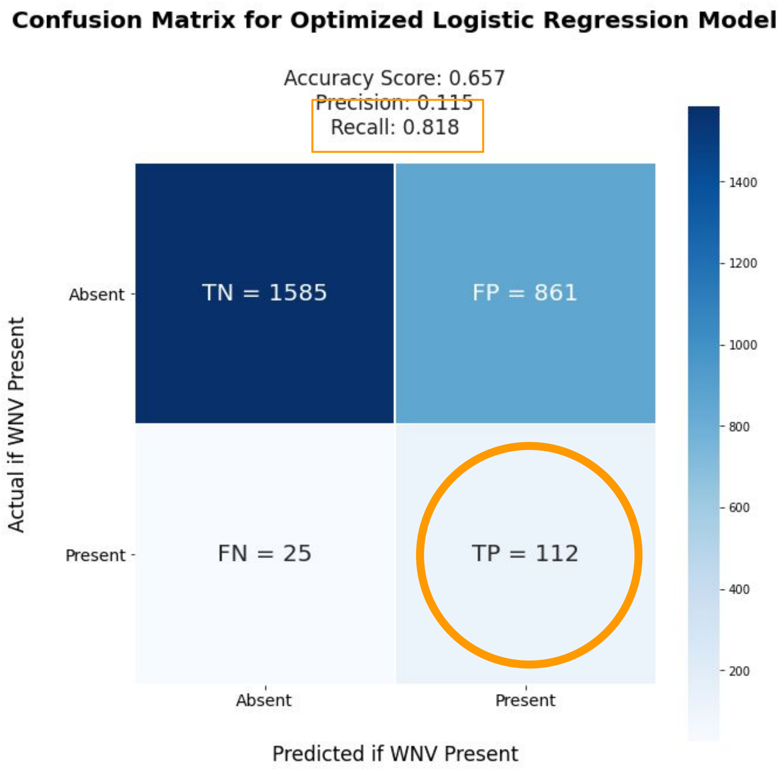
Perform hyperparameter tuning using GridSearchCV

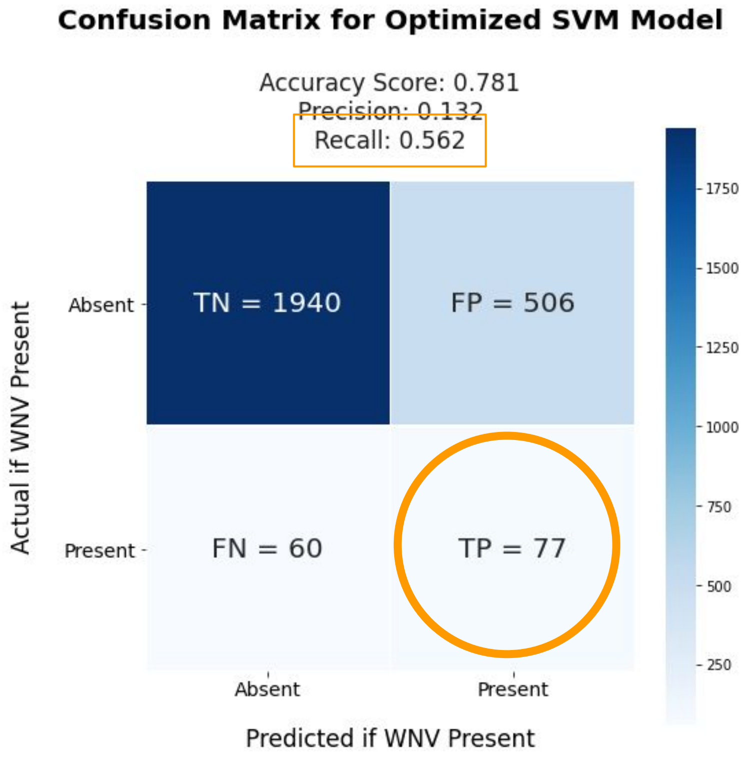
Model Performance Table

Optimized Model	Training Accuracy	Testing Accuracy	Training Recall	Testing Recall	(Testing) AUC Score
LR	0.640	0.657	0.816	0.818	0.79
SVM	0.831	0.781	0.991	0.562	0.78
KNN	0.947	0.947	0.009	0.000	0.73
RF	0.901	0.863	0.772	0.307	0.68

- Logistic Regression is the best performing model
- The model has the highest recall score and AUC score
- Recall is more important than accuracy in evaluating our model performance

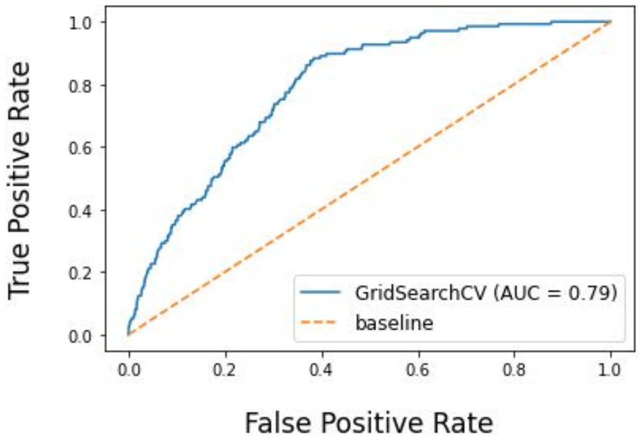
Confusion Matrix for LR & SVM



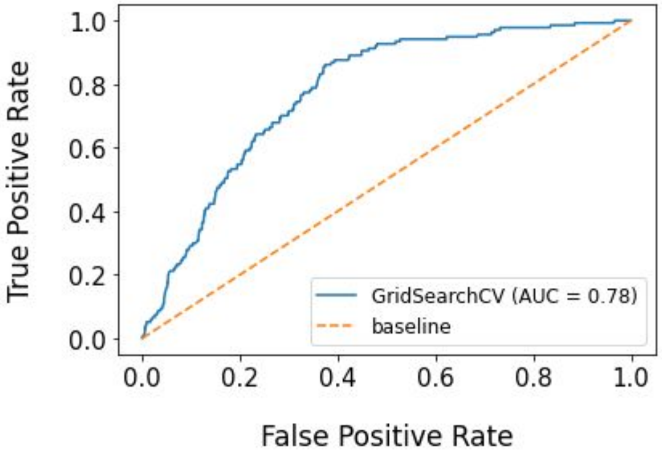


ROC Curve for LR & SVM

LR



SVM



- The AUC for LR is 0.79 and SVM is 0.78
- Both models perform better than the baseline AUC of 0.50
- There is a high chance that the classifier will be able to distinguish the positive class values from the negative class values

Top 3 Features by Variables Category

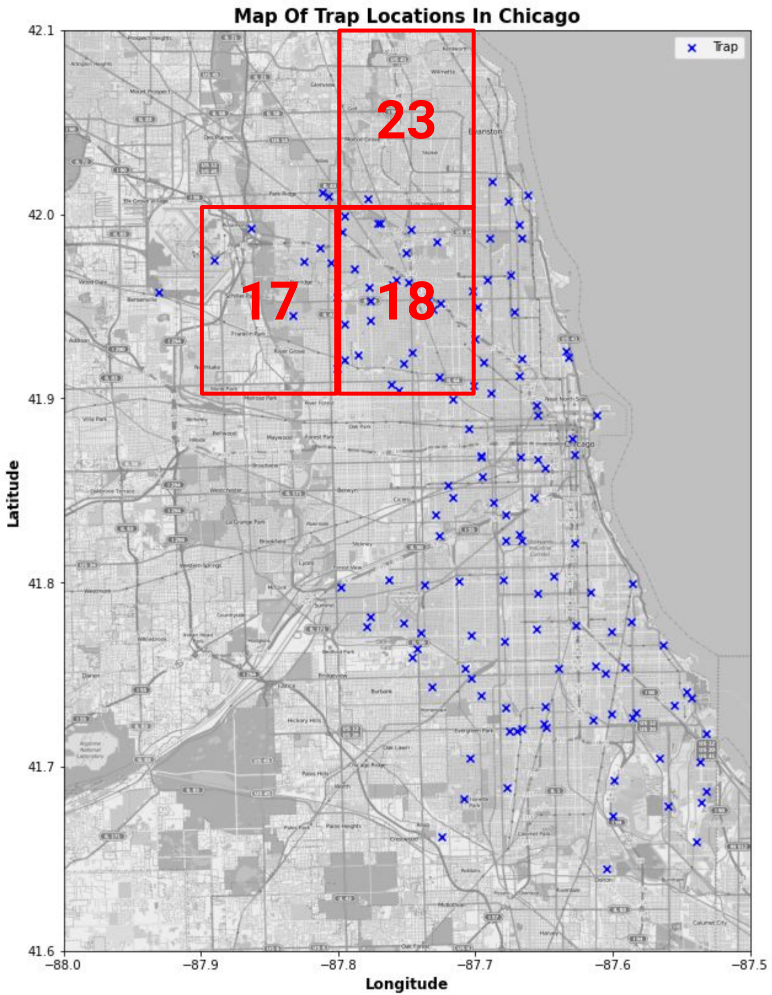
Month	Odds Coef
August	2.892771
September	2.010299
July	1.218872

Species	Odds Coef
Culex Pipiens	1.265604
Culex Pipiens / Restuans	1.232219

Weather	Odds Coef
Average Temp	1.484479
Sea Level	1.174961
Wind Direction	1.14185

Top 3 Features by Variables Category

Coordinate Group	Odds Coef
Group 17	1.396158
Group 18	1.120353
Group 23	1.117992



Cost-Benefit Analysis

Definitions

Cost	Total expenditure associated with spraying pesticide on adult mosquitoes incurred in a year.
Benefit	Benefits are measured by direct and indirect cost savings in the form of healthcare costs and productivity lost associated with the potential reduction in number of human WNV cases from pesticide spraying.

Condition for pesticide spraying to be cost-effective:

Benefit-Cost Ratio >= 1

Benefits

Method to Project Cost Savings

- Limited published data on the medical costs and economic burden for WNV
- Reference ***Initial and Long-Term Costs of Patients Hospitalized with West Nile Virus Disease*** published in American Journal of Tropical Medicine and Hygiene, 2014 to estimate direct and indirect medical costs of hospitalization
- Paper studied a cohort of 80 patients in Colorado in 2003; 38 followed for 5 years to determine long-term medical and lost-productivity costs
- Paper estimated total costs of 18,256 hospitalized WNV cases in the US from 1999 to 2012 using 10,000 outputs from the Monte Carlo simulation of the findings from the cohort

Benefits

Deriving Unit Cost of Different Hospitalized Cases

$351,542
Average cost of death due to WNV

$20,687
Average cost of a case with neuroinvasive disease

$16,904
Average cost of a case with non-neuroinvasive disease

$49,607
Weighted-average cost of a hospitalized case

Costs of Spraying Pesticide

Assumptions

- Pesticide is sprayed by a truck-mounted fogger using very small amount of pesticide in a process known as Ultra Low Volume (ULV) fogging
- Apply at rate **1.5 ounces per acre** to ensure there is no significant risk to present to individuals and the community
- Zenivex E20 adulticide is an attractive option for professionals looking for advanced mosquito control
- **3,800 Ounces of Zenivex E20 costs $10,899.95**
- Therefore, cost per acre is $4.30

Costs of Spraying Pesticide

Assumptions

- Without our classification model, assumed spray coverage is up to the entire area of Chicago, spanning about **149,976 acres (~150,000 acres)**

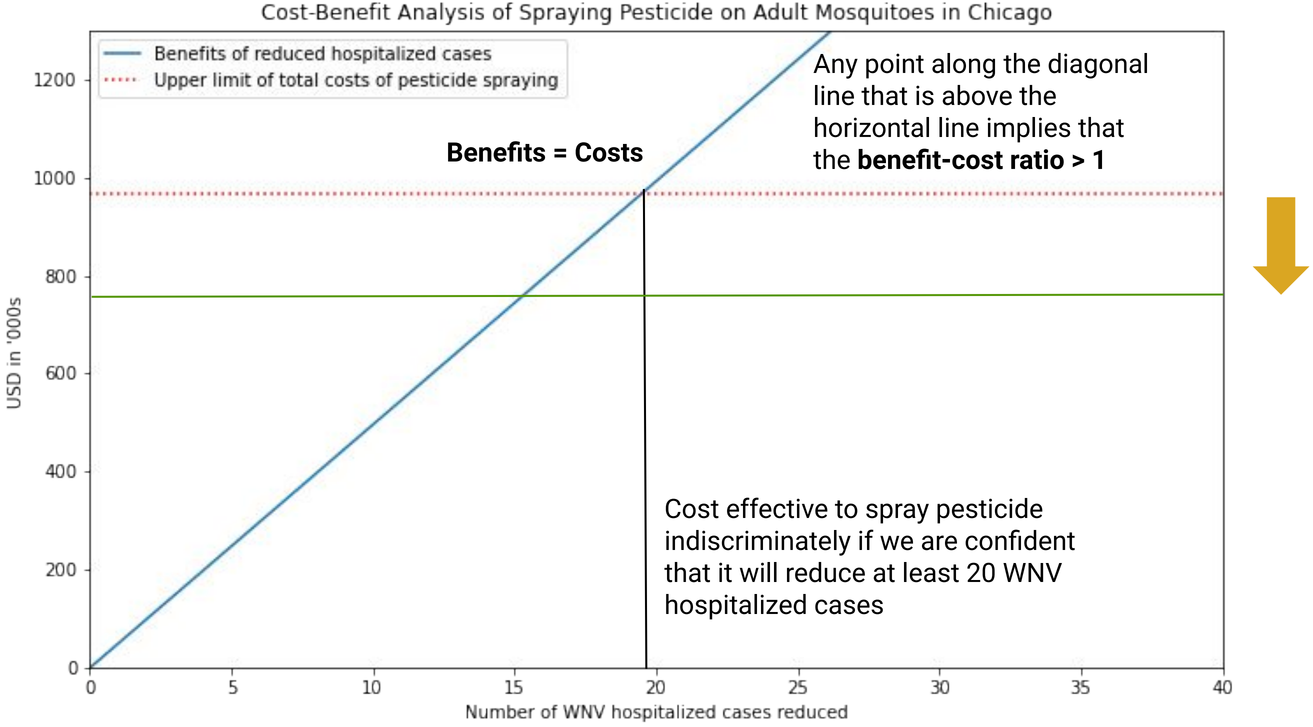
- Cost of pesticide = \$150,000 x 4.30
  = \$645,000

- Assume labor and overheads costs are 50% of pesticide cost:
  Total cost of pesticide spraying = 1.5 x \$645,000
  = **\$967,500**

- With our classification model and more targeted spraying, we assumed a 20% reduction in pesticide coverage, bringing total costs down to **\$774,000**

Cost-Benefit Analysis

Spraying Pesticide Seems to be Cost-Effective



Assume a 20% reduction in pesticide coverage: Brings total costs down to **$774,000**

Recommendations

- Reconciling findings from our EDA, selected model, and the cost-benefit analysis, we strongly recommend IDPH to control mosquitoes population in Chicago by **spraying pesticide in a more targeted fashion**

- First, IDPH should **commence pesticide spraying from the start of July to the end of August,** when higher average temperature accelerates growth of adult mosquito population → Spraying **frequency should be weekly** for pesticide to take effect

- Second, repeated spraying can be performed in 'hot' zones in **coordinates Groups 17, 18, and 23**

- Finally, it would be **helpful to determine and target locations where majority of the Culex Pipiens and Culex Restuans populations** are found

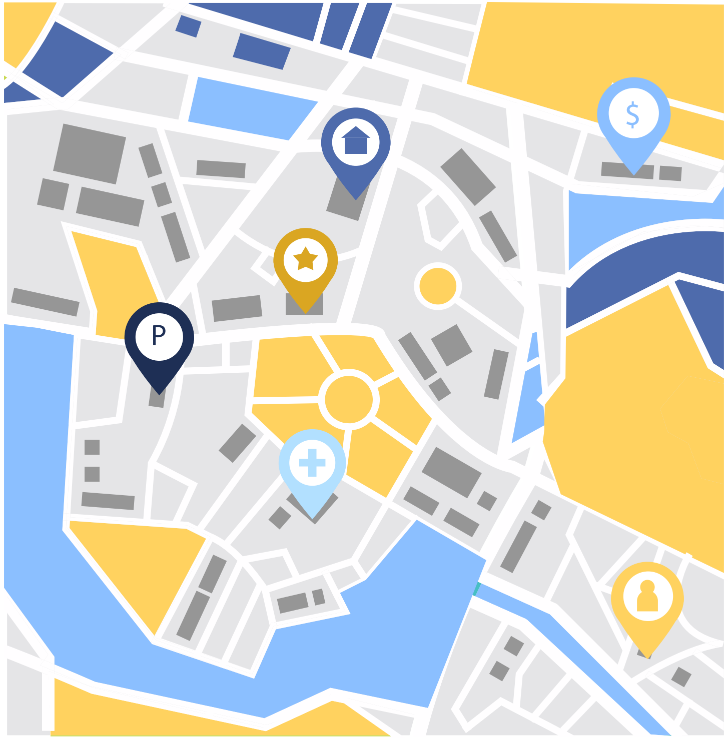
Conclusion

- Best performing model for predicting WNV is the **logistic regression**
    - Model was able to correctly predict **81.8% of positive cases**
    - Good guide for targeted spraying of pesticide at **high-risk areas**
- Model can be further **improved**
    - Perform **over-sampling** on the **minority class** (positive cases)
    - Retrain model on new dataset to get a **stronger logistic regression**
- Spraying **pesticide** is only one of the many possible control measures
    - **Drawbacks**: Repeated applications, potential negative ecological repercussions, health risk to the human population
- Use **minnows** (small freshwater fishes)
    - **Benefits**: Low maintenance, cost effectiveness, environmental friendliness, minimal implications on public health

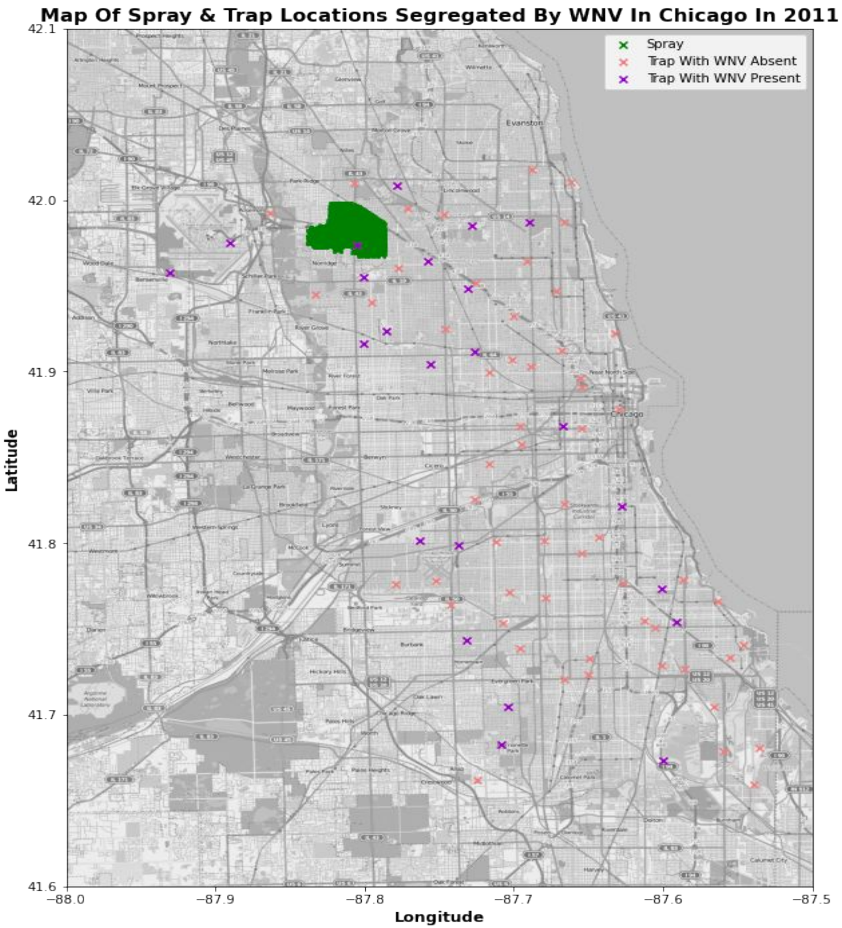
Future Steps

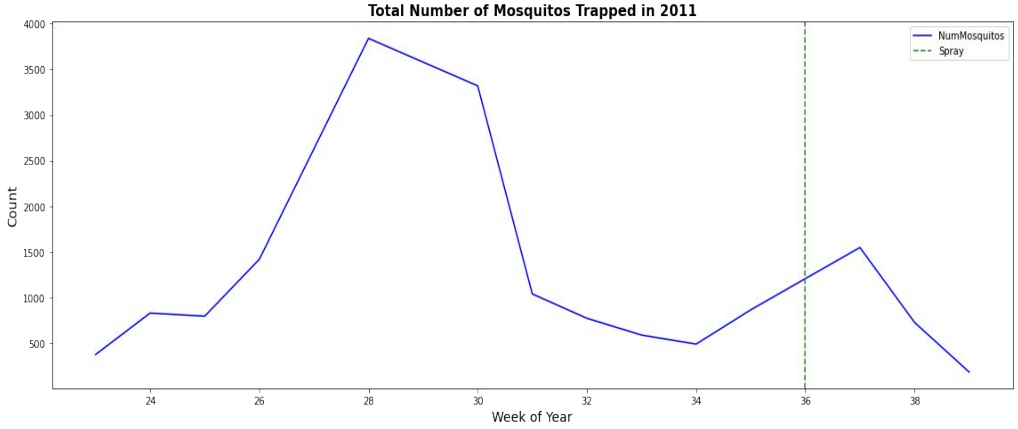
1. Factor in benefit derived (in terms of the cost saved) from reducing cases of **other forms of mosquito-borne diseases** such as Zika, Chikungunya, dengue, and malaria
2. Analyse cost-benefit of employing **other forms of mosquito eradication measures** such as use of minnows
3. Modify and apply predictive model to help predict and direct spray efforts in **other cities of the US**

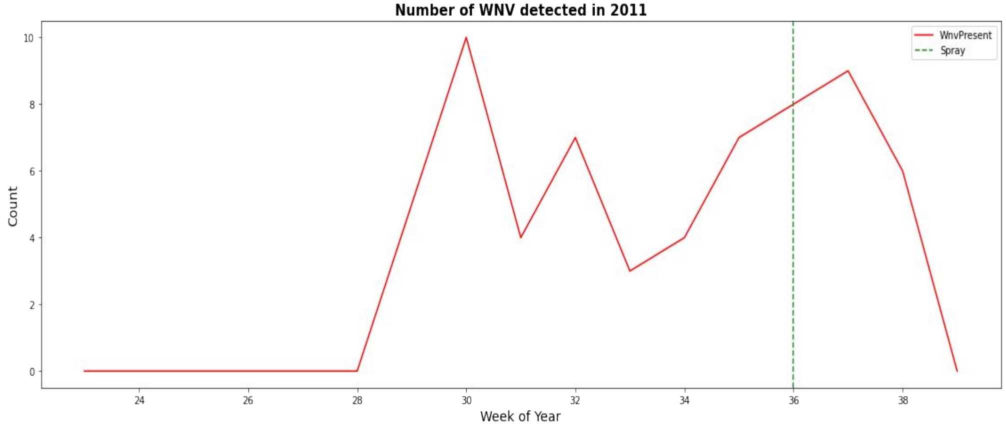
Thank You



Appendix


Map Of Spray & Trap Locations Segregated By WNV In Chicago In 2011
Spray
Trap With WNV Absent
Trap With WNV Present
Latitude
Longitude
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5
Evanston
Chicago


Total Number of Mosquitos Trapped in 2011
NumMosquitos
Spray
Count
Week of Year
4000
3500
3000
2500
2000
1500
1000
500
24
26
28
30
32
34
36
38


Number of WNV detected in 2011
WnvPresent
Spray
Count
Week of Year
10
8
6
4
2
0
24
26
28
30
32
34
36
38

Benefits

Cost Categories

Cost Category	Description	Mean (USD)
(A) Total acute medical care	Inpatient costs associated with hospital-based care	252,115,100
(B) Total acute lost productivity	Assumed for hospitalized patients who missed work had a work schedule of 5 out of every 7 days. Time lost from work, by age and sex, using estimates from Grosse and others. Does not include death.	22,081,260
(C) Total long-term medical care	Includes costs such as medical appointments and institutional care costs, drug costs and durable medical equipment incurred in the 5 years after initial hospitalization.	27,570,280
(D) Total long-term lost productivity	Similar to how (B) was estimated. For persons who retired early as a result of WNV, we valued their indirect costs as the number of potential years and months of lost employment(65 minus age at early retirement).	26,866,800
(E) Total lifetime lost productivity caused by deaths	Derived from the lifetime production value discounted at 3% for their age and sex.	449,464,800
(F) Grand total	Sum of (A) to (E)	778,098,240

Lifetime lost productivity was calculated directly from Grosse and others based on age and sex for the 1,524 WNV disease case-patients reported to CDC who died.