Predicting WNV Presence in Chicago through Weather Data

Problem Statement

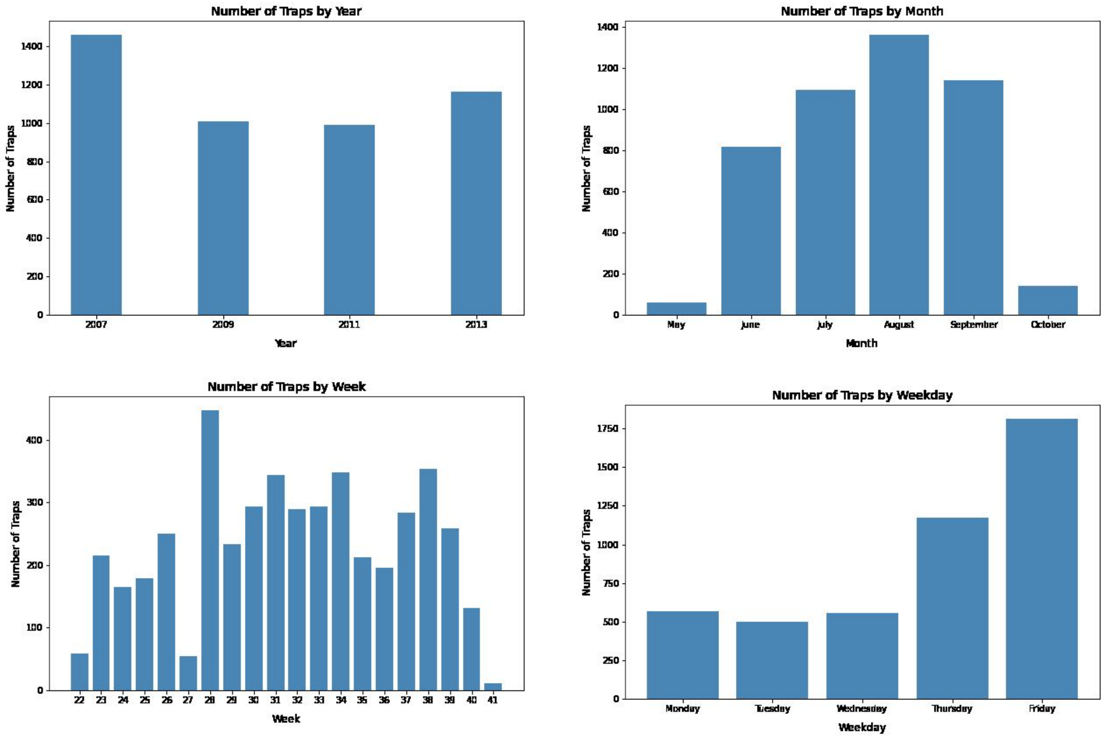
- Data scientists at Disease and Treatment Agency
- Given mosquito collection data by Department of Public Health
- Instructed to come up with a plan on how to manage the West Nile Virus through use of

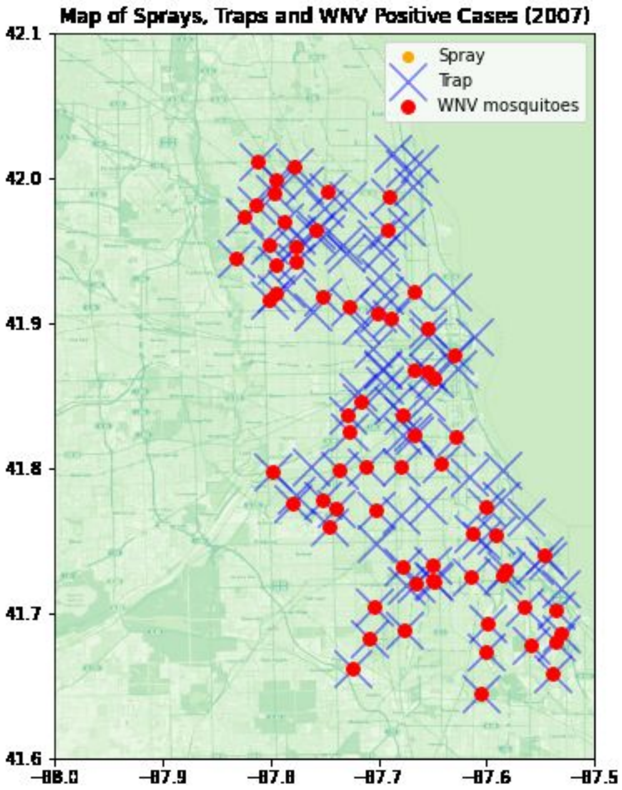
Background Information

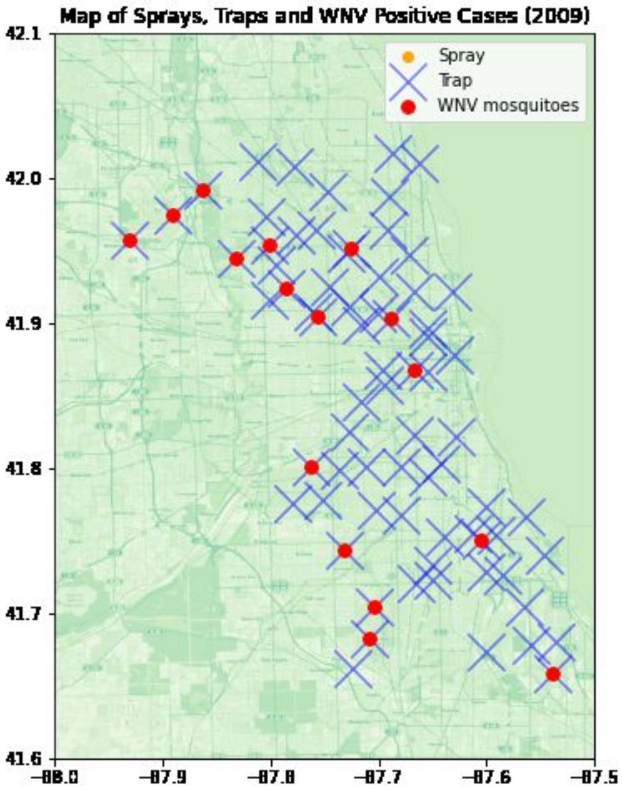
- West Nile Virus is the most dominant vector-borne disease in North America
- Virus stems from Africa, introduced into North America via birds
- Mosquitoes transmit the disease from bird to human through feeding on both species

EDA

- Quality of data
  - Inconsistent sampling quantity
    - Collection date
    - Location

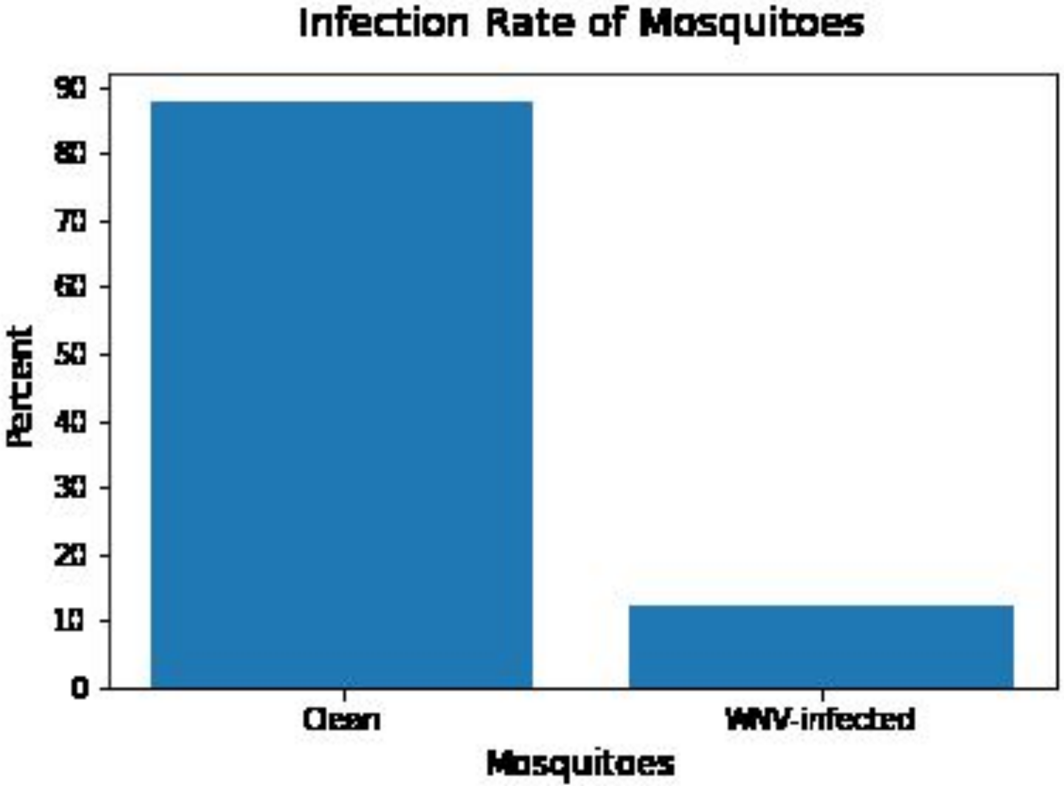



Map of Sprays, Traps and WNV Positive Cases (2007)
Spray
Trap
WNV mosquitoes
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


Map of Sprays, Traps and WNV Positive Cases (2009)
Spray
Trap
WNV mosquitoes
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

EDA

- Large class imbalance
- SMOTE necessary



Modelling Workflow

- Data prep (feature engineering, dual models)

- Balancing Data
  - Scale first then SMOTE (Synthetic Minority Oversampling Technique)

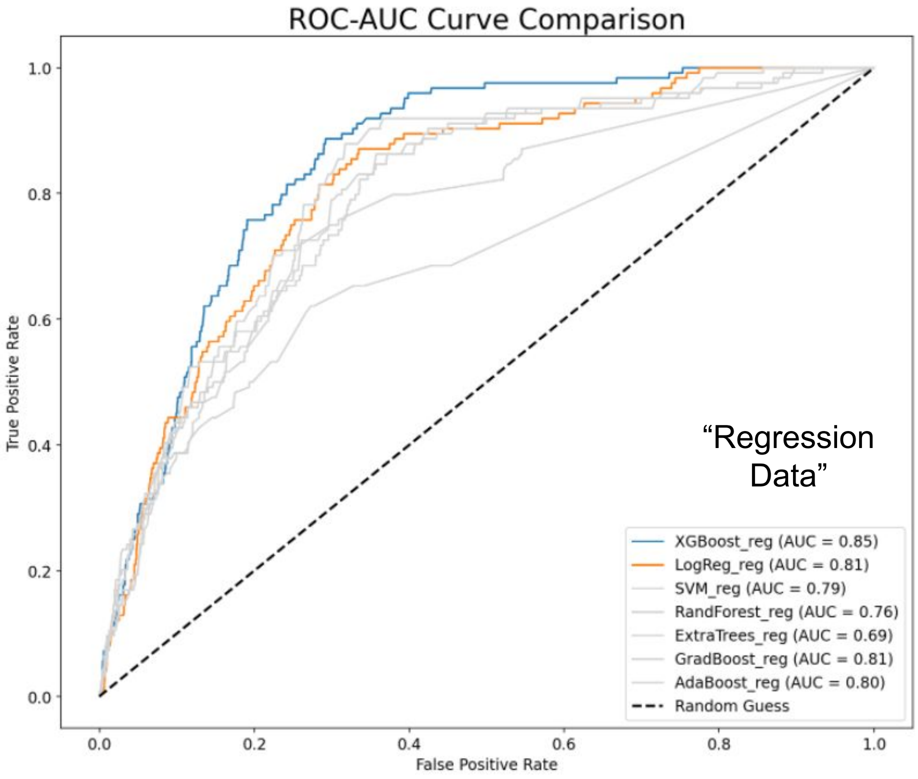
- Final model Tuning and Adding Features

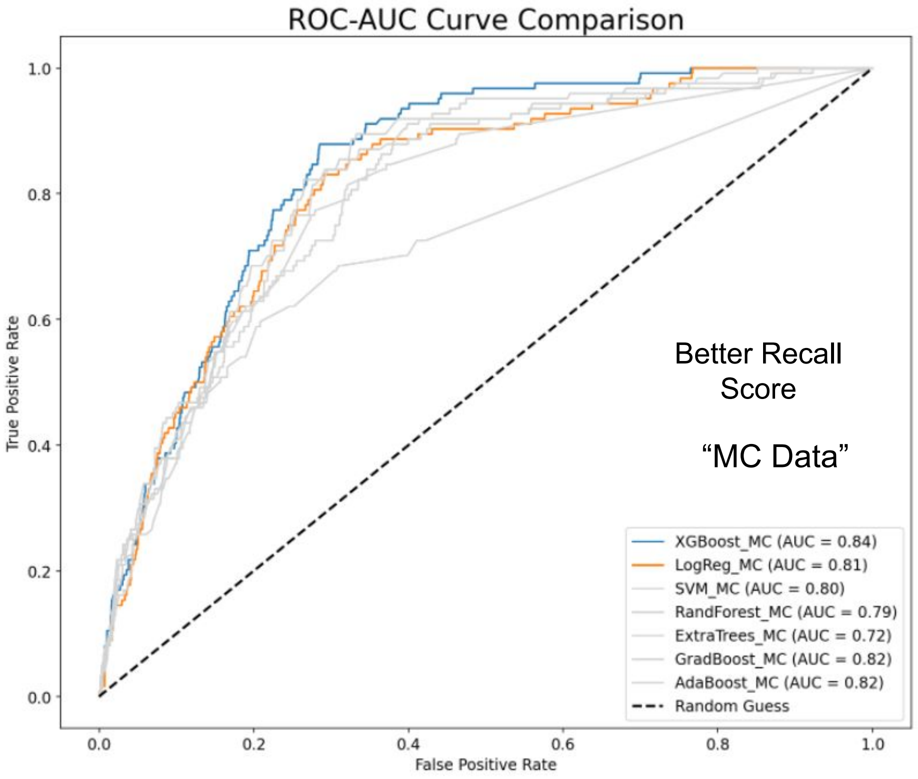
Feature Engineering

- Identified from EDA:
  - Humidity and Temp play a part
  - Locations play a part
    - Used Trap
  - Certain Months/Weeks play a part
- Resorted to Polynomial Features
- Tested 2 Datasets
  - Minimal multicollinearity
  - Features *with* multicollinearity

Feature	Correlation
nummosquitos	0.248242
week	0.104807
month	0.101115
dewpoint	0.087043
wetbulb	0.082887
tmin	0.070538
longitude	0.068600
tavg	0.065243
station	0.063496
cool	0.059307
heat	0.054685
year	0.053875
tmax	0.051986
resultspeed	0.045893
latitude	0.033330
avgspeed	0.031990
preciptotal	0.014298
addressaccuracy	0.012268
weekday	0.011084
block	0.009204
sealevel	0.007648
stnpressure	0.006849
resultdir	0.000253

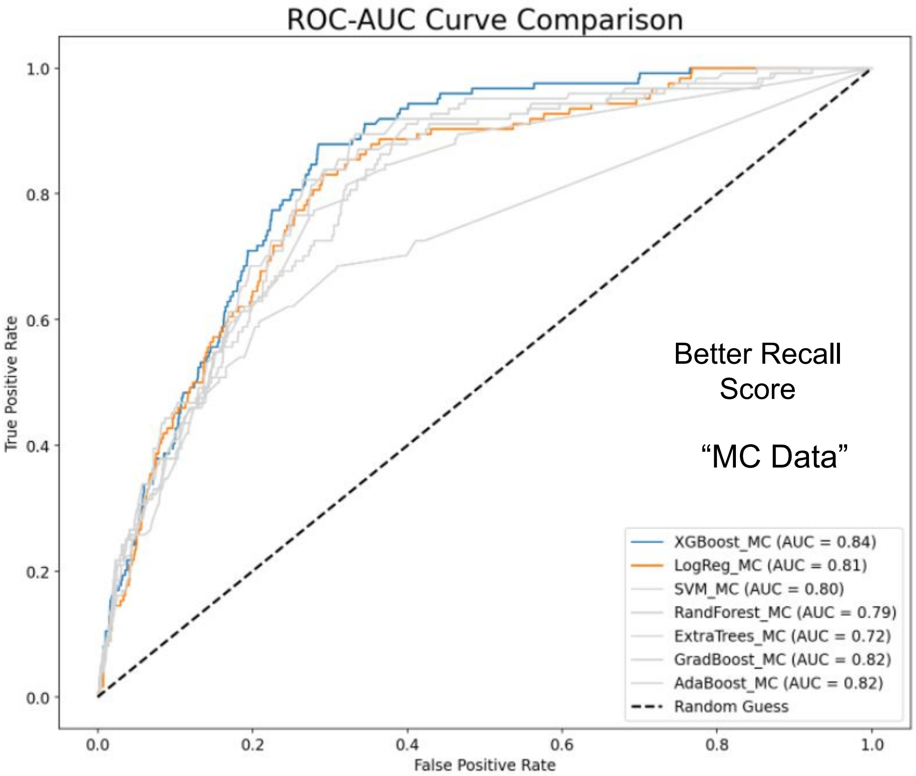
Base Models


ROC-AUC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
“Regression Data”
XGBoost_reg (AUC = 0.85)
LogReg_reg (AUC = 0.81)
SVM_reg (AUC = 0.79)
RandForest_reg (AUC = 0.76)
ExtraTrees_reg (AUC = 0.69)
GradBoost_reg (AUC = 0.81)
AdaBoost_reg (AUC = 0.80)
Random Guess


ROC-AUC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Better Recall Score
“MC Data”
XGBoost_MC (AUC = 0.84)
LogReg_MC (AUC = 0.81)
SVM_MC (AUC = 0.80)
RandForest_MC (AUC = 0.79)
ExtraTrees_MC (AUC = 0.72)
GradBoost_MC (AUC = 0.82)
AdaBoost_MC (AUC = 0.82)
Random Guess

Base Models

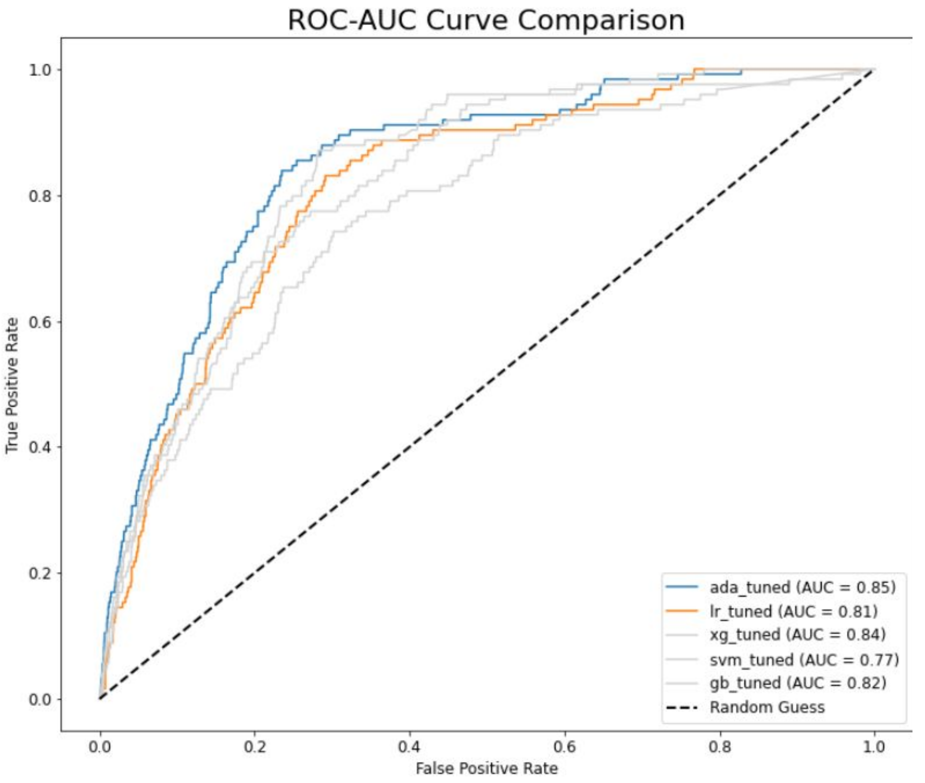
	model_name	model	recall	test_auc
7	lr_base_reg	lr	0.629032	0.813799
0	lr_base_multi	lr	0.620968	0.814525
1	svm_base_multi	svm	0.411290	0.803701
8	svm_base_reg	svm	0.403226	0.793494
6	ada_base_multi	ada	0.225806	0.816133
13	ada_base_reg	ada	0.193548	0.800370
5	gb_base_multi	gb	0.169355	0.822053
12	gb_base_reg	gb	0.096774	0.813685



- Prioritise Recall Score
    - why?

Tuned Models

	model_name	recall	test_auc
0	lr_tuned	0.620968	0.814525
1	svm_tuned	0.290323	0.772498
3	gb_tuned	0.161290	0.816846
2	ada_tuned	0.153226	0.848250
4	xg_tuned	0.048387	0.837677



Back to Feature Engineering

nummosquitos	$\boxed{0.248242}$
week	0.104807

- Testing ‘nummosquitos’ feature
  - Made a Significant difference in model performance

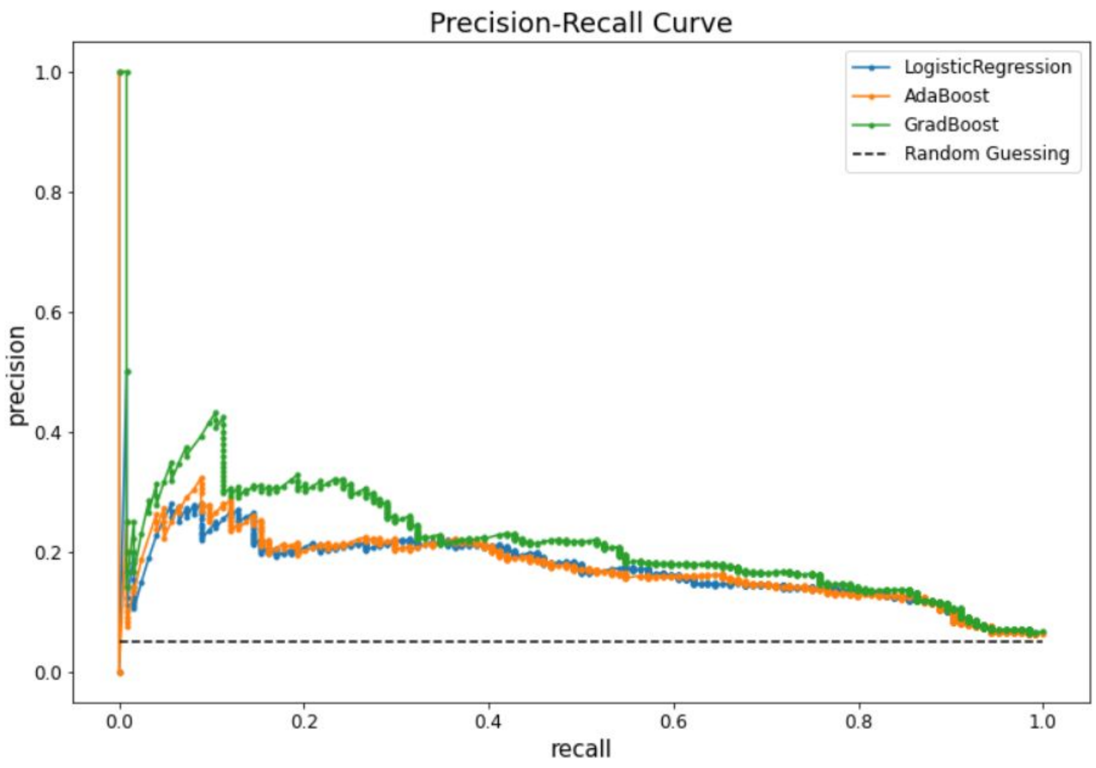
- Secondary Linear Regression model
  - Predicts ‘nummosquitos’ for the test dataset
  - Fairly low root mean squared error, details can be found in the notebooks that will be shared after this presentation

- Increases of up to 6% in model performance
  - Merit to use a prediction model for this

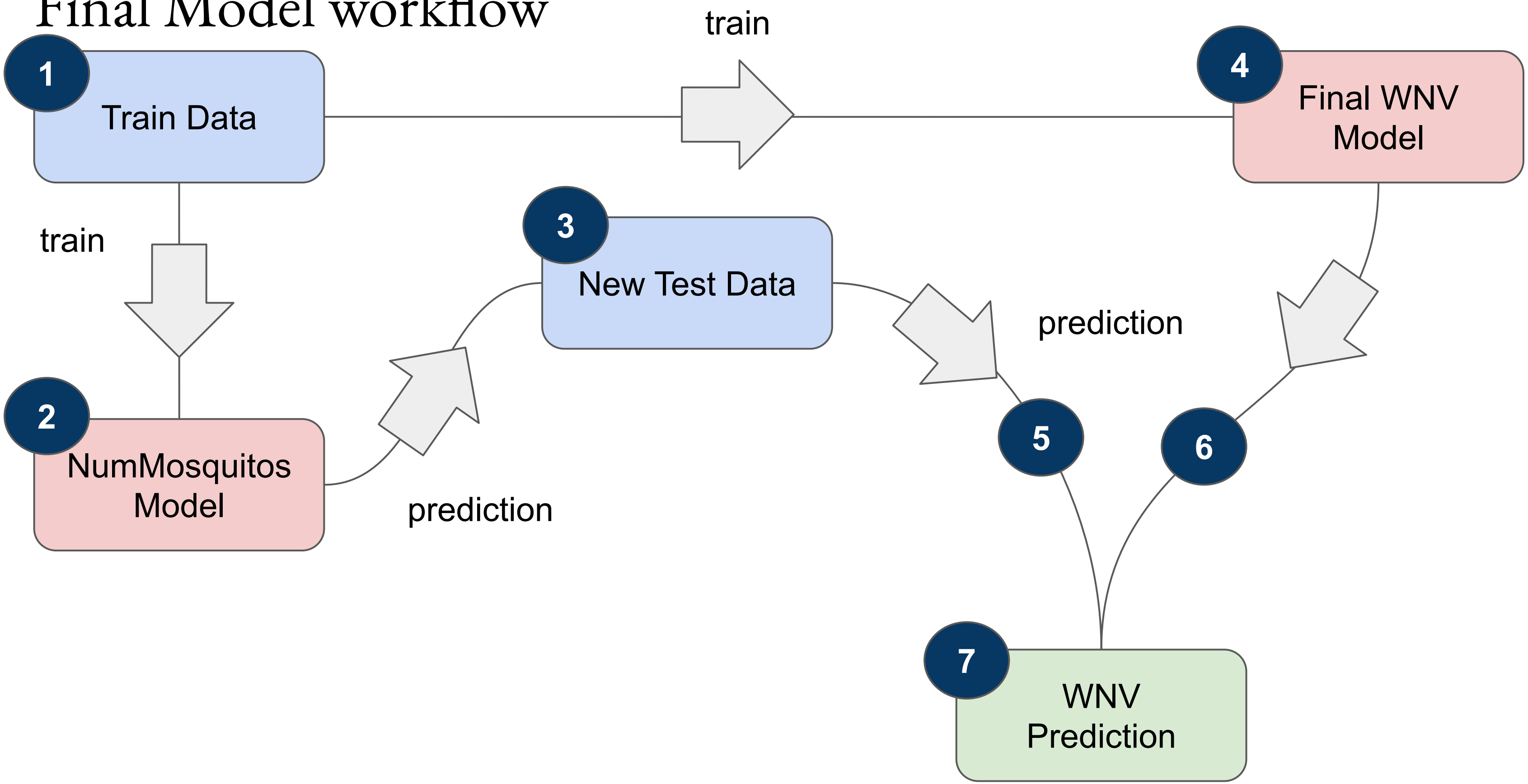
Precision-Recall Curve

- Logistic Regression
  - Tradeoff

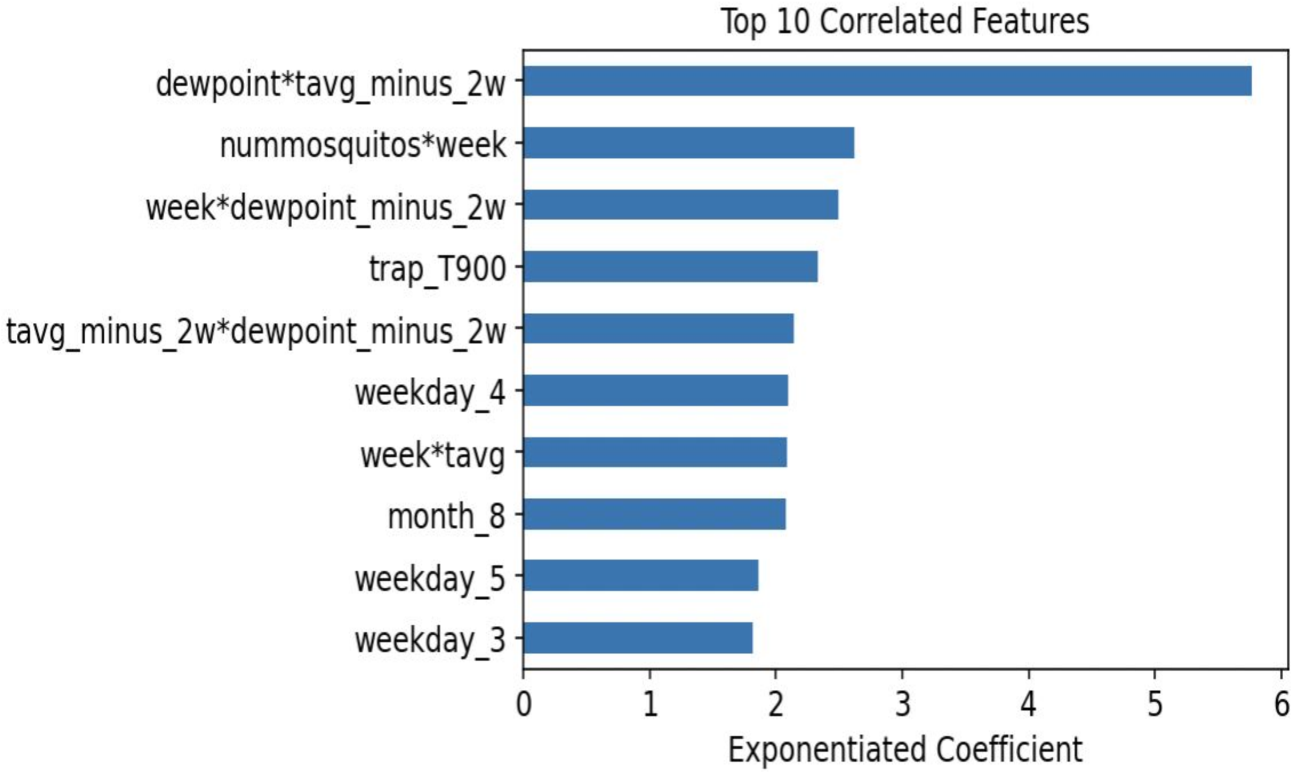
	model_name	recall	test_auc
0	lr_final_tuned	0.685484	0.850978
1	ada_logreg_tuned	0.685484	0.847914
2	gb_logreg_tuned	0.201613	0.860428



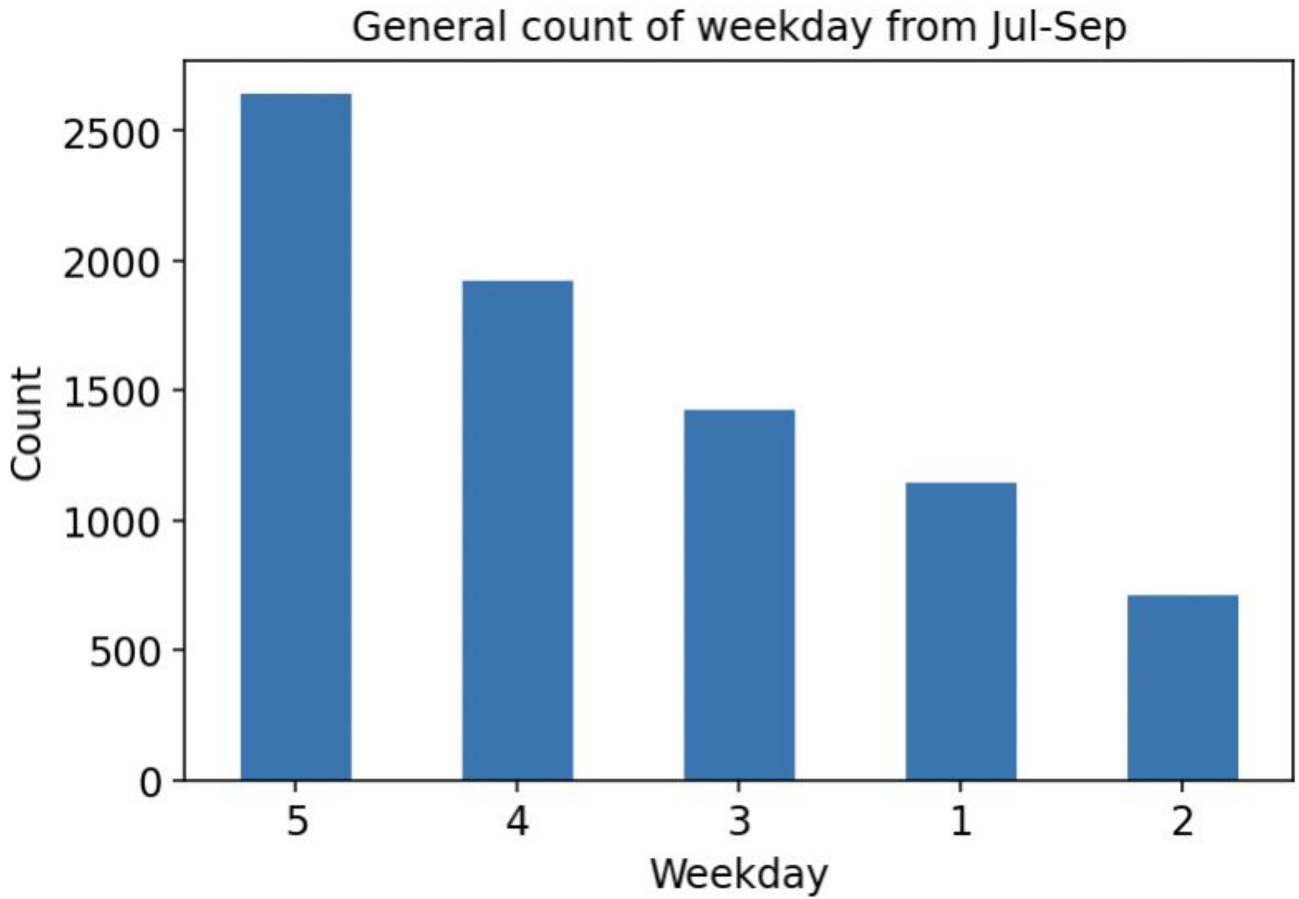
Final Model workflow



Model and Error analysis



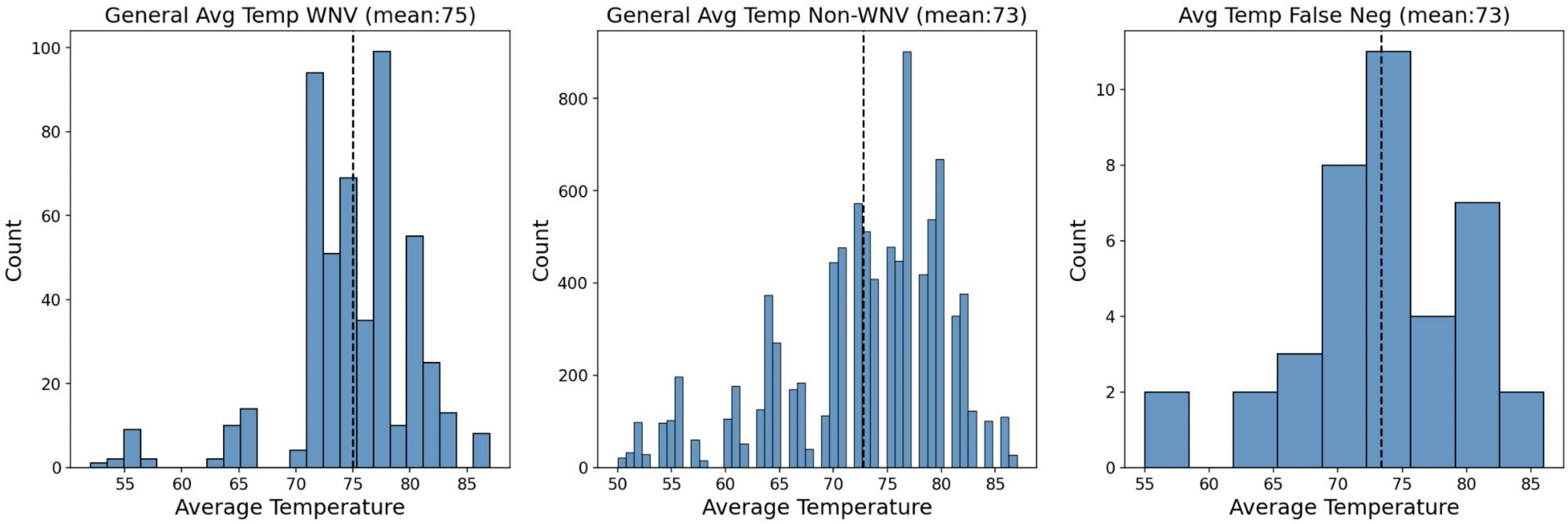
Mode


General count of weekday from Jul-Sep
Count
2500
2000
1500
1000
500
0
5
4
3
1
2
Weekday

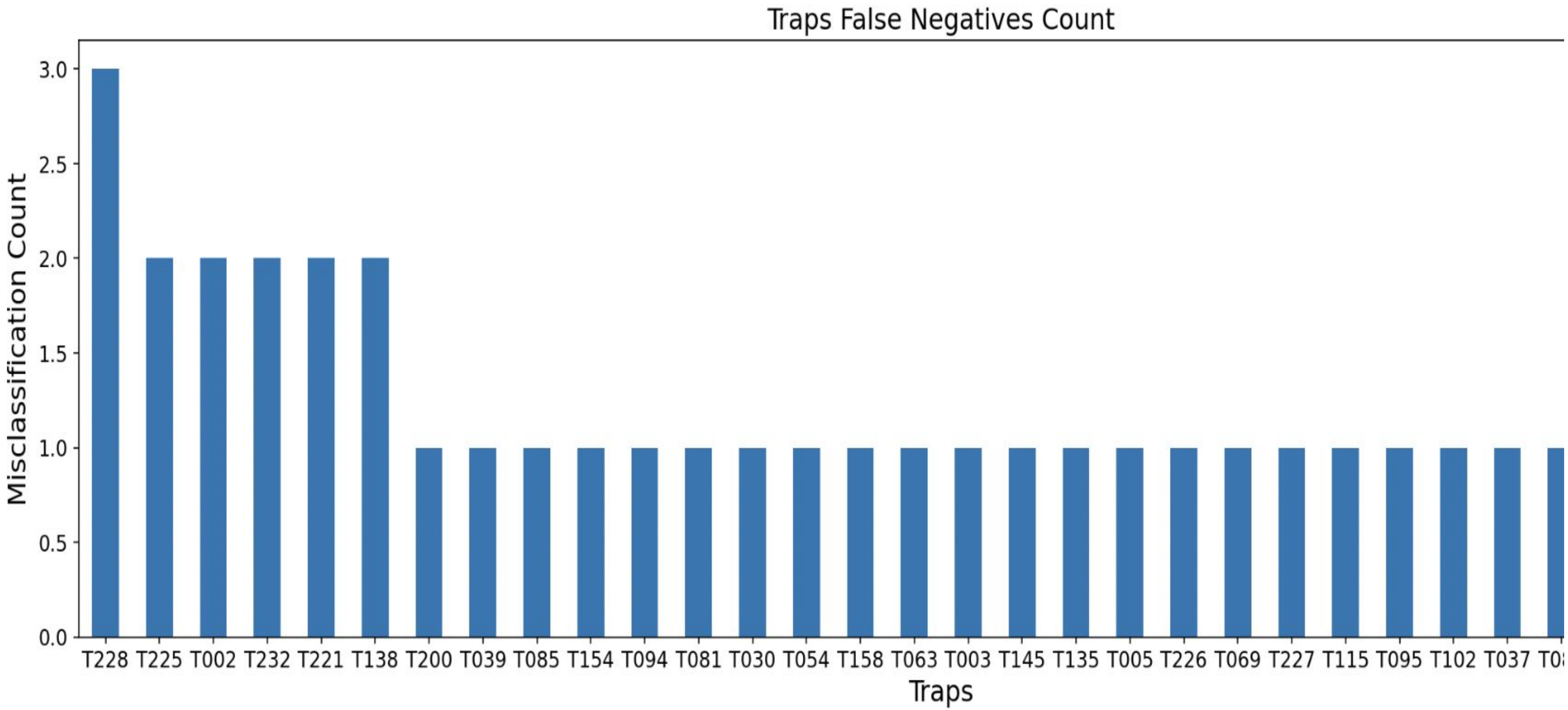
Confusion Matrix

	Predict No WNV	Predict WNV
Actual No WNV	1952 (80.5%)	348 (14.3%)
Actual WNV	39 (1.6%)	85 (3.5%)

Error analysis on Temperature



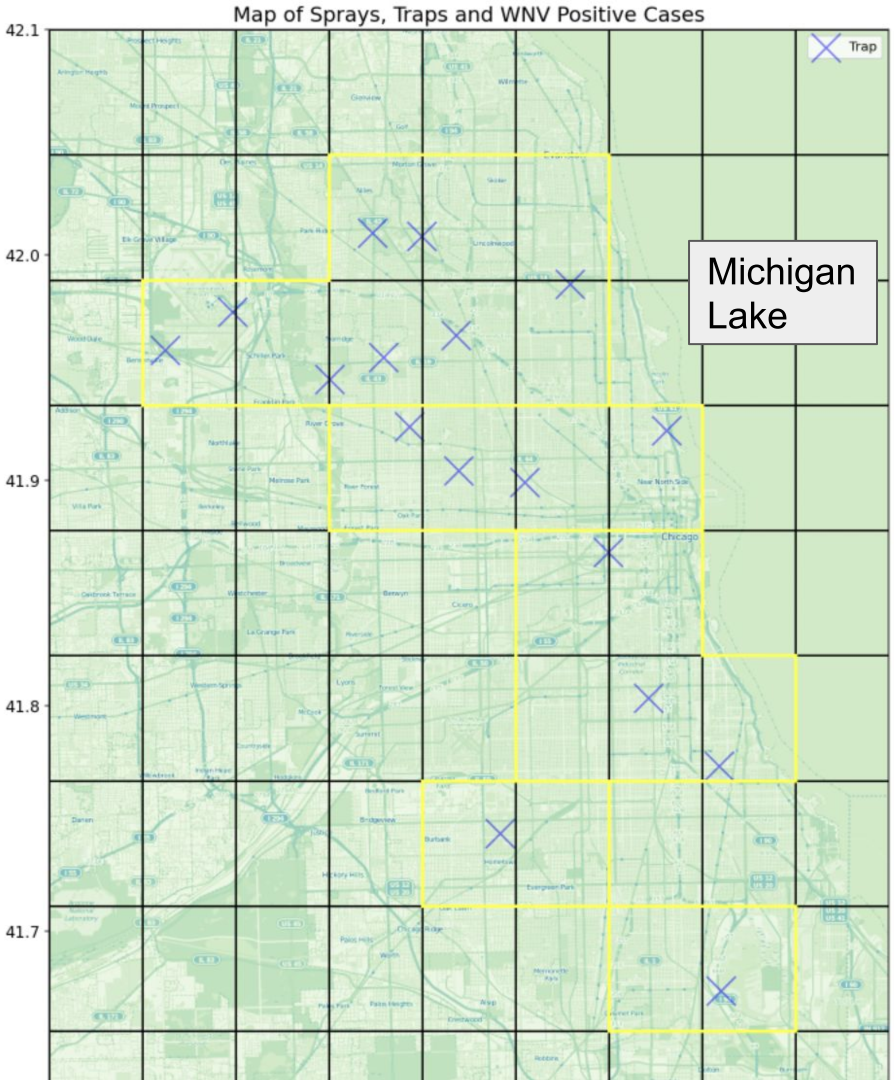
Error analysis on traps



Deep dive into Misclassified Trap T228

General WNV Rate	Trap T228 WNV Rate
5%	9%

	nummosquitos	month	tavg	dewpoint	trap
7955	29	7	71.0	53	T228
4989	2	9	71.0	55	T228
8123	17	7	84.0	70	T228



Map of Sprays, Traps and WNV Positive Cases
42.1
42.0
41.9
41.8
41.7
Trap
Michigan Lake
Chicago

Cost Benefit Analysis - Spray Cost [1]

Item	Cost Calculation
Spray Cost Per km sq	$245.21
Total km sq	238,211 km sq
Frequency	12
Total Spray Cost	$2,858,536

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

Cost Benefit Analysis - Estimated Reduction in WNV (2013)

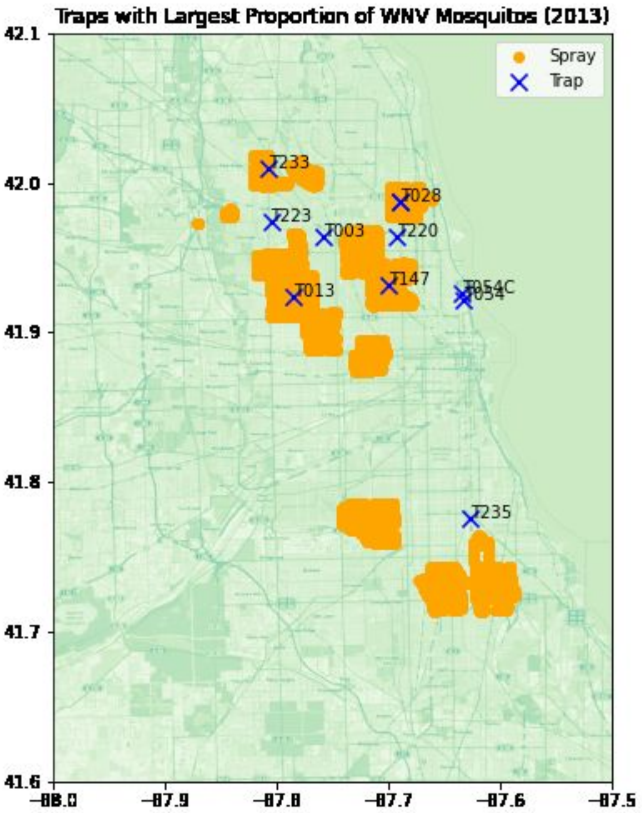
Item	Cases
WNV Incidences	113
Reduction from Spray	57.5%
Predicted Reduction of WNV instances	67

Cost Benefit Analysis - Cost Savings (2013)

Item	Cost
Cost Per Case	$17,000
Total Case Reduction	67
Human Infection Per WNV instance	3 (Minimum profitability)
Estimated Total Cost Savings	+ $3.41 mil
Total Spray Cost	- $2.85 mil
Total Savings	+ $0.56 mil

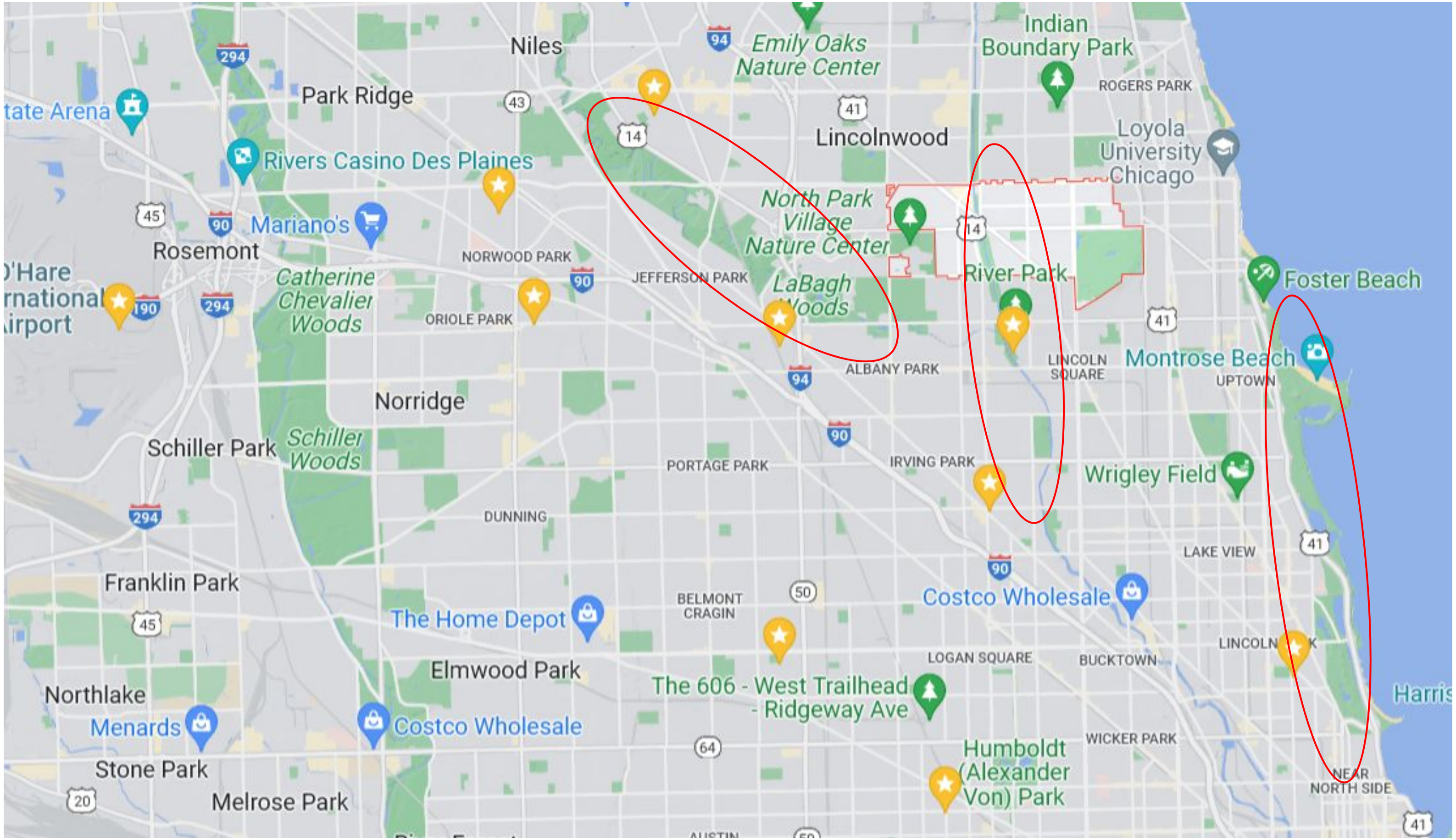
Cost Benefit Analysis - Cost Savings (2013)

Human Infection Per WNV instance	**Cost Savings** (+$ 1.13 mil per additional case)
1	- $1.72 mil
2	- $0.59 mil
3	+ $ 0.56 mil
4	+ $ 1.69 mil
5	+ $ 2.82 mil



Conclusions and Recommendations

	Details
Where to spray	Trap locations based on highest WNV infection rate
When to Spray	July - September
What to spray with	Malathion


Niles
Park Ridge
Emily Oaks Nature Center
Indian Boundary Park
ROGERS PARK
Lincolnwood
Loyola University Chicago
Rivers Casino Des Plaines
Mariano's
Rosemont
North Park Village Nature Center
River Park
NORWOOD PARK
JEFFERSON PARK
LaBagh Woods
Foster Beach
Catherine Chevalier Woods
ORIOLE PARK
LINCOLN SQUARE
Montrose Beach
UPTOWN
ALBANY PARK
Norridge
Schiller Park
Schiller Woods
PORTAGE PARK
IRVING PARK
Wrigley Field
DUNNING
LAKE VIEW
Franklin Park
BELMONT CRAGIN
Costco Wholesale
The Home Depot
LOGAN SQUARE
BUCKTOWN
Elmwood Park
The 606 - West Trailhead - Ridgeway Ave
Northlake
Menards
Costco Wholesale
WICKER PARK
Humboldt (Alexander Von) Park
Stone Park
Melrose Park
NEAR NORTH SIDE

Conclusions and Recommendations

Consider…

- Using alternate insecticides
- Using alternate insecticide dispersal methods
  - Community-based larvicide application
- Optical sensor-based counting system to ensure consistent sampling

Future Improvements

- Fine-tune Linear Regression to better predict number of mosquitos

- Further fine-tune XGBoost to explore its potential performance

- Explore other classification techniques such as Neural Networks etc

External Research

1.https://academic.oup.com/jme/article/56/6/1456/5572378

2.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

3.https://academic.oup.com/jme/article/56/6/1456/5572378

4.https://www.chicago.gov/city/en/depts/cdph/provdrs/healthy_living/news/2021/august/city-to-spray-insecticide-wednesday-to-kill-mosquitoes.html

5.https://www.centralmosquitocontrol.com/-/media/files/centralmosquitocontrol-na/us/specimen%20labels/zenivex%20e20%20specimen%20label.pdf

6.https://pubmed.ncbi.nlm.nih.gov/23883850/

7.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2661378/

8.https://link.springer.com/article/10.1007/s00340-019-7361-2

9.https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179673

10.https://www.science.org/content/article/after-40-years-most-important-weapon-against-mosquitoes-may-be-failing

11.https://www.cmmcp.org/pesticide-information/pages/zenivex-e4-etofenprox

12. http://npic.orst.edu/factsheets/archive/malatech.html