Kaggle
West Nile Virus Prediction

DSi Project 4
Boss, Fong, Gip

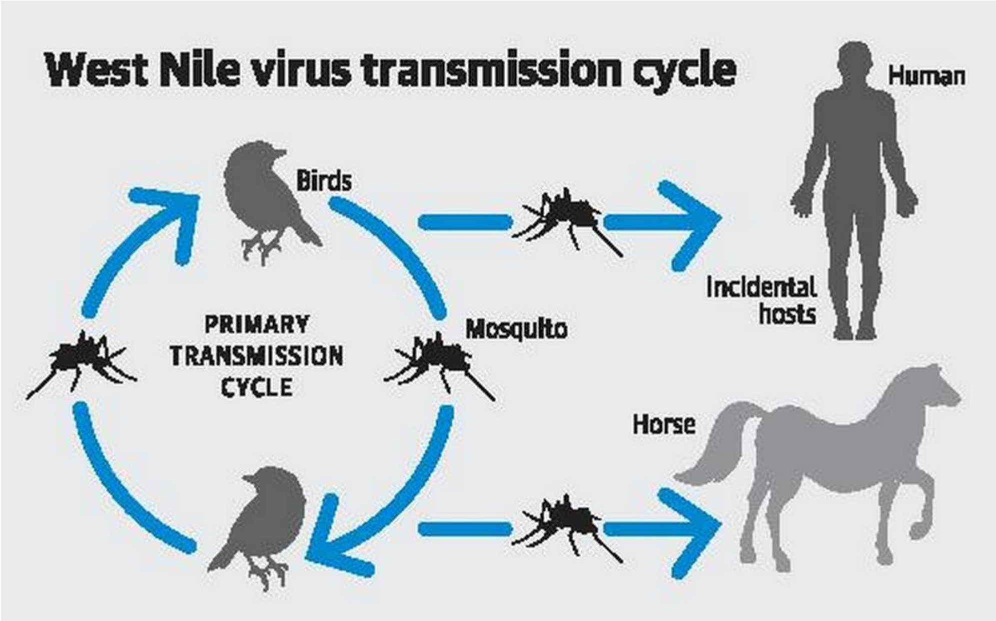
Agenda

- Introduction
- Datasets
- Cleaning & Preprocessing
- EDA
- Model
- Cost-Benefit Analysis
- Conclusion & Recommendations

1. Introduction

West Nile Virus (WNV)

- Most commonly spread to humans through infected mosquitos
- Approximately 20% of infected people develop symptoms ranging from a persistent fever to serious illnesses that can result in death
- First human cases were reported in Chicago 2002.



Chicago & Its Control Program

- In 2004, the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today
- Every year from late-May to early-October, public health workers in Chicago setup mosquito traps scattered across the city. Every week from Monday through Wednesday, these traps collect mosquitos, and the mosquitos are tested for the presence of West Nile virus before the end of the week.
- The test results include the number of mosquitos, the mosquitos species, and whether or not West Nile virus is present in the cohort.

Problem Statement

As a team of data scientists from CDPH, we are tasked with building a model that can help predict when and where different species of mosquitoes will test positive for WNV.

Using weather, location, testing and spray data to evaluate, the model should help the city of Chicago and CDPH more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

2. Datasets

Datasets

Train dataset: *2007, 2009, 2011, 2013*

Test dataset: *2008, 2010, 2012, 2014*

- Train
- Test
- Weather

Modeling

- Spray

Cost Benefit Analysis

Data Types Breakdown

13

Integer

features

- Number of Mosquitoes
- Temperature (Min/Max)
- WetBulb
- etc.

10

Float

features

- Latitude
- Longitude
- Total Precipitation
- etc.

6

DateTime

features

- Trap Dates
- Spray Dates
- Sunrise/Sunset
- etc.

2

String

features

- Trap Name
- Mosquito Species

Training & Testing Datasets

Feature	Variable type	Datatype	Dataset	Description
id	Norminal	int64	test	The id of the record
date	Datetime	datetime	train and test	Date that the WNV test is performed
species	Norminal	object	train and test	Species of mosquito
trap	Norminal	object	train and test	Id of the trap
latitude	Continuous	float64	train and test	Latitude returned from GeoCoder
longitude	Continuous	float64	train and test	Longitude returned from GeoCoder
nummosquitos	Discrete	int64	train and test	number of mosquitoes caught in this trap
wnvpresent	Discrete	int64	train and test	whether WNV was present in these mosquitos. 1 means WNV is present, and 0 means not present.
year	Discrete	int64	train and test	Year that the WNV test is performed
month	Discrete	int64	train and test	Month that the WNV test is performed
weekofyear	Discrete	int64	train and test	Week of year that the WNV test is performed
yearmonth	Discrete	int64	train and test	Year and month that the WNV test is performed

Weather Datasets

Feature	Variable type	Datatype	Dataset	Description
station	Discrete	int64	weather	Station 1 or 2 where weather data is collected
date	Datetime	datetime	weather	Date of weather record
tmax	Discrete	int64	weather	Maximum temperature in Fahrenheit
tmin	Discrete	int64	weather	Minimum temperature in Fahrenheit
tavg	Continuous	float64	weather	Average temperature in Fahrenheit
depart	Discrete	float64	weather	The difference from normal temperatures for the last 30yrs
dewPoint	Discrete	int64	weather	Average Dew Point temperature in Fahrenheit
wetBulb	Discrete	int64	weather	Average Wet Bulb temperature in Fahrenheit
sunrise	Datetime	datetime	weather	Sunrise time
sunset	Datetime	datetime	weather	Sunset time
preciptotal	Continuous	float64	weather	The depth of rainfall/melted snow in inches
resultspeed	Continuous	float64	weather	Resultant wind speed
resultdir	Continuous	int64	weather	Resultant wind direction
avgspeed	Continuous	float64	weather	Average wind speed
daytime	Continuous	float64	weather	Number of hours of sunlight for each day

Spray Dataset

Feature	Variable type	Datatype	Dataset	Description
date	Datetime	datetime	spray	Date of the spray
time	Datetime	datetime	spray	Time of the spray
latitude	Continuous	float64	spray	Latitude of the spray
longitude	Continuous	float64	spray	Latitude of the spray

3. Cleaning & Preprocessing

Data Cleaning

Dealing with **Missing Values**

Missing Weather Station 2 data is imputed using Weather Station 1 data

Data Type **Conversion**

Converting date/time columns to DateTime object

Feature Engineering

Extracting **Weeks/Month/Year** From DateTime

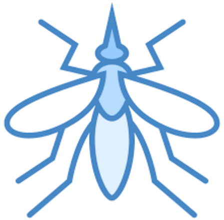
Creating columns for weeks/months/year for training/testing set

OneHotEncoding **Species** with WNV

Encoding mosquitoes species with WNV

Predicting **NumMosquitoes** On Test set

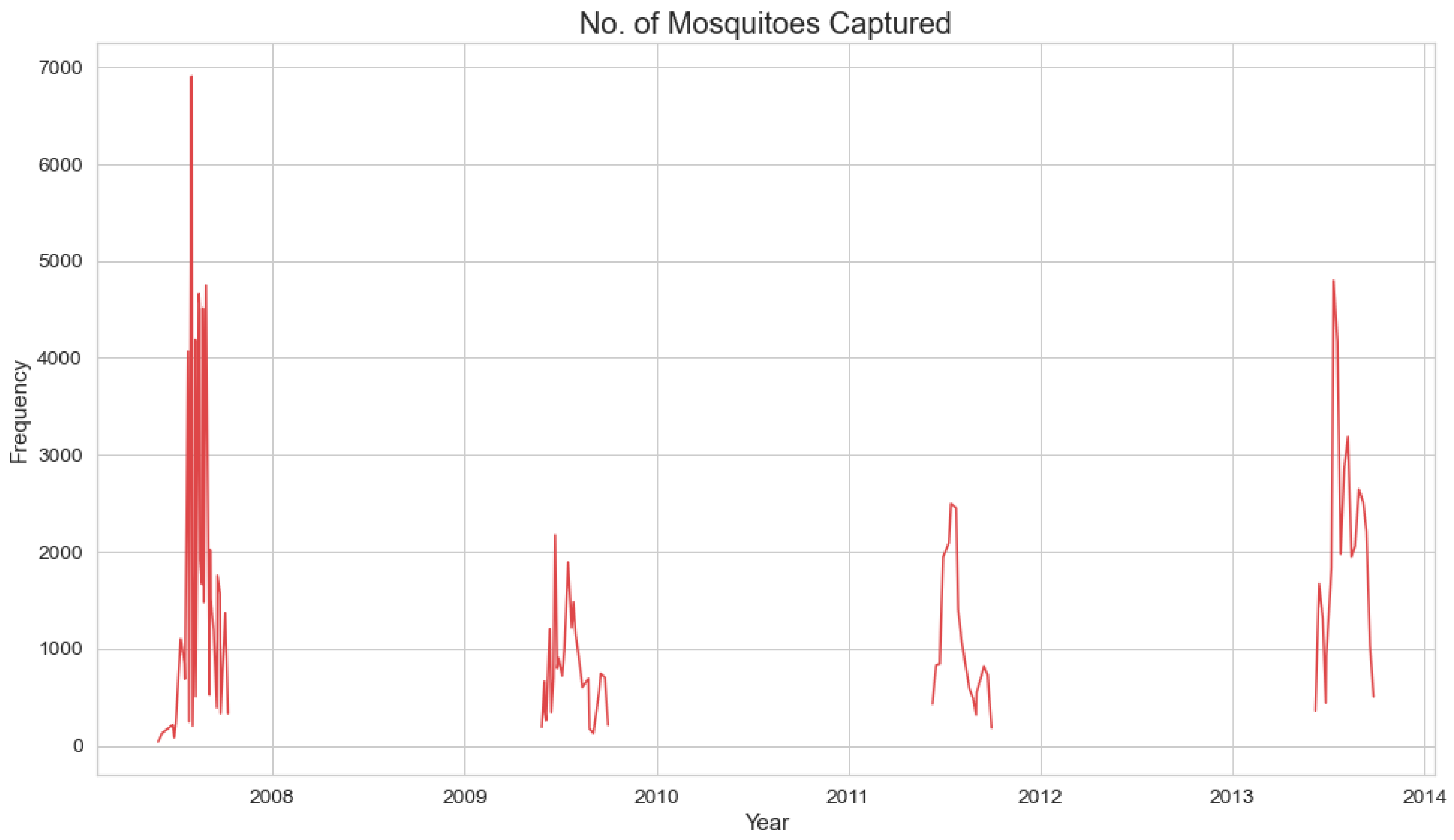
Using kNNRegressor on the training set to predict mosquito count on testing set



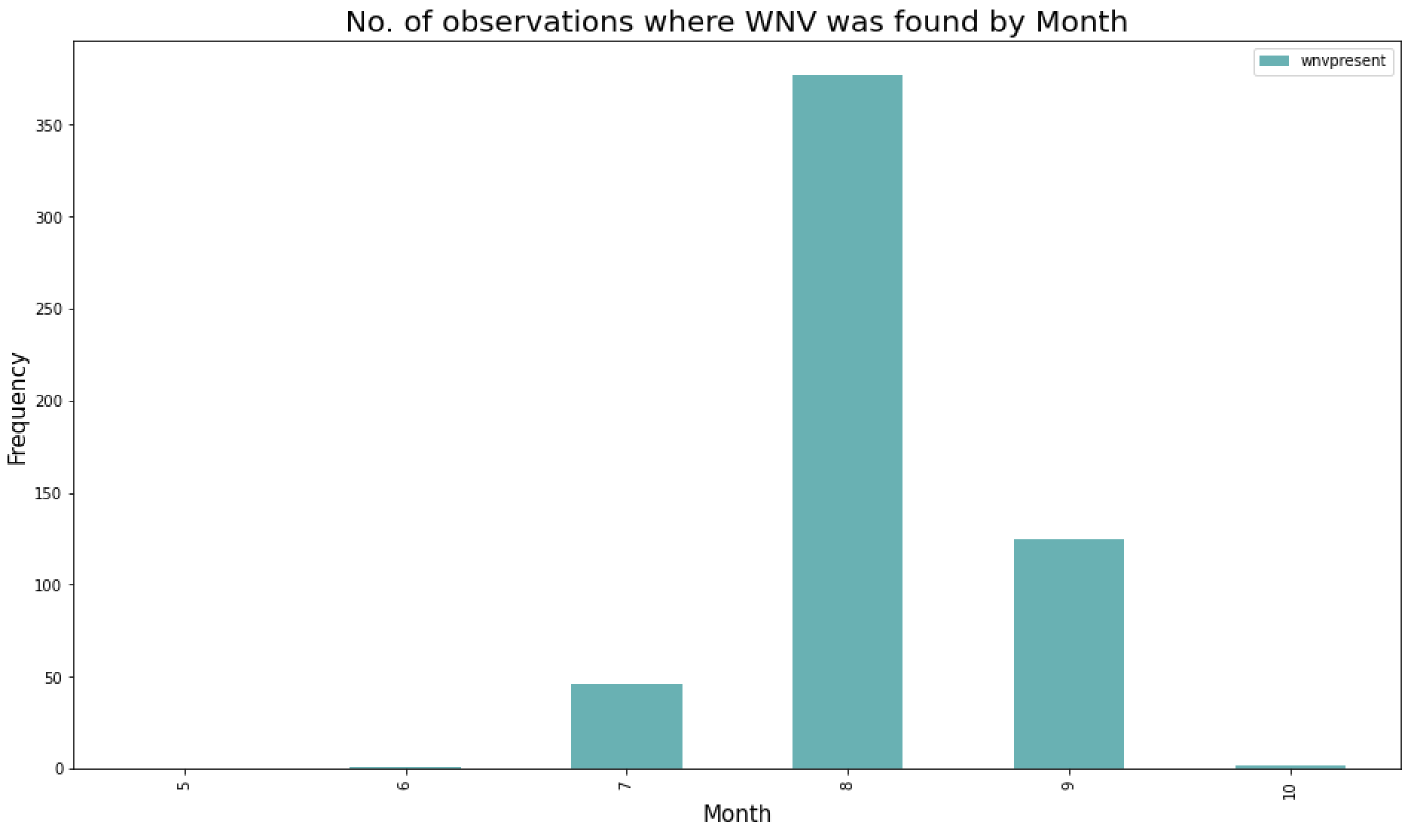
Calculating **Daylight Hours**

Calculating the number of hours of daylight from Sunrise/Sunset

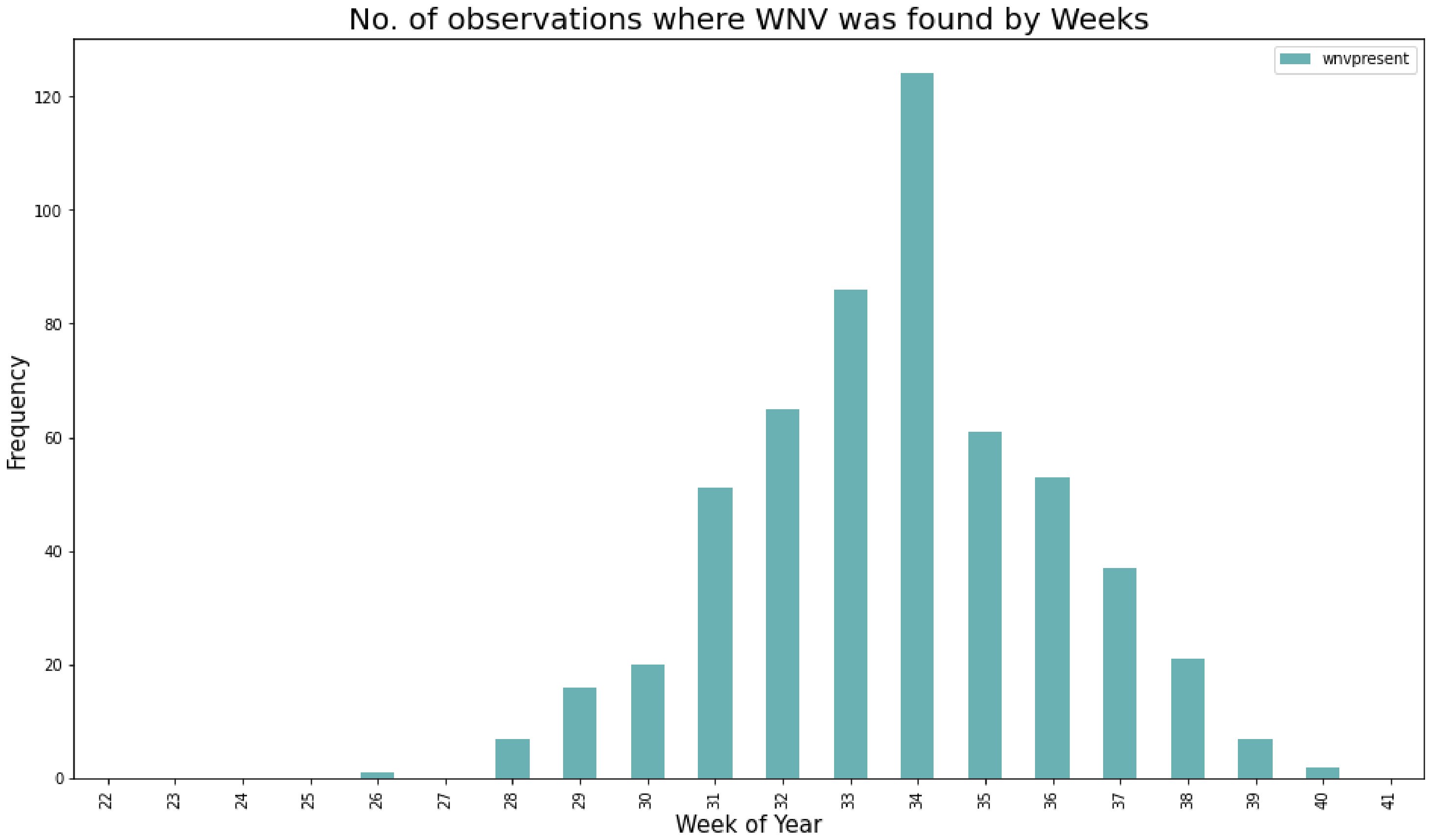
4. EDA


No. of Mosquitoes Captured
Frequency
Year
0
1000
2000
3000
4000
5000
6000
7000
2008
2009
2010
2011
2012
2013
2014

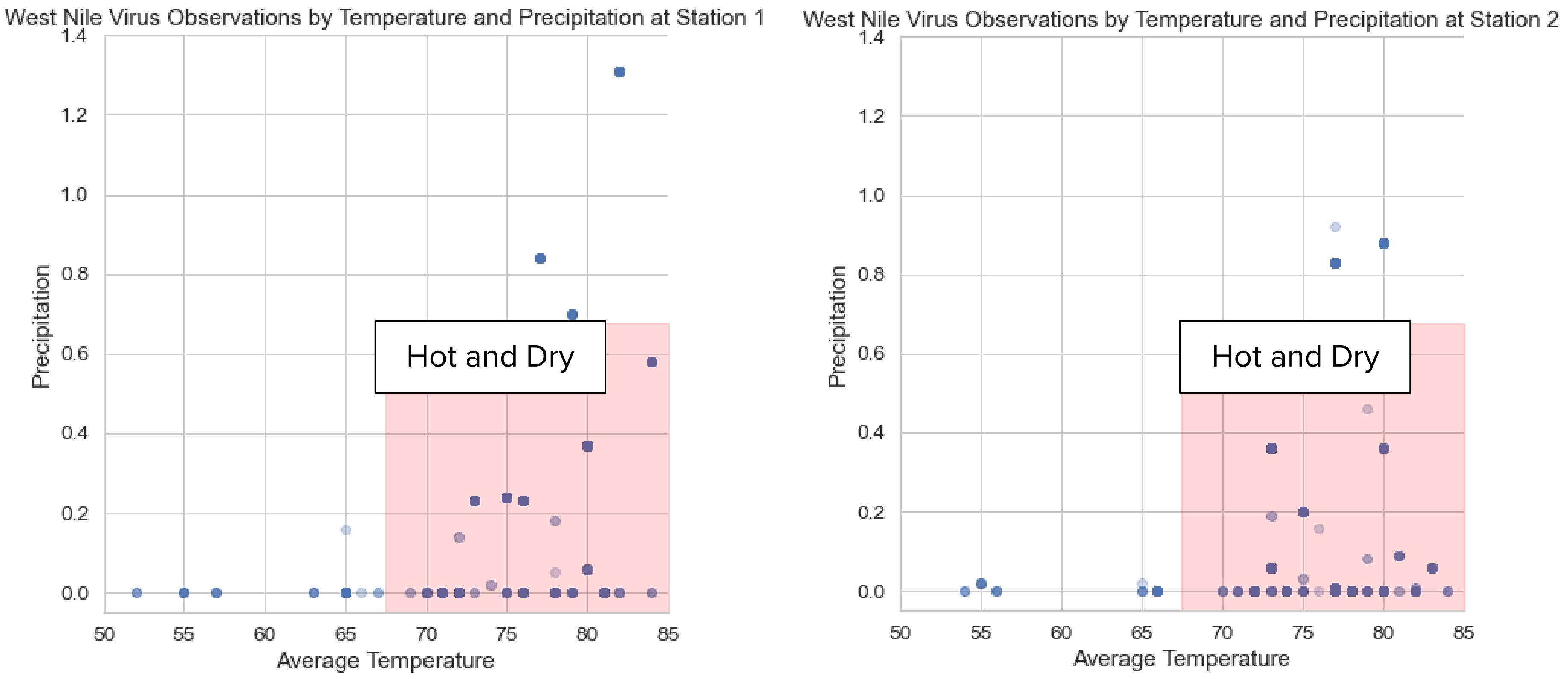
Summer is the season of infection



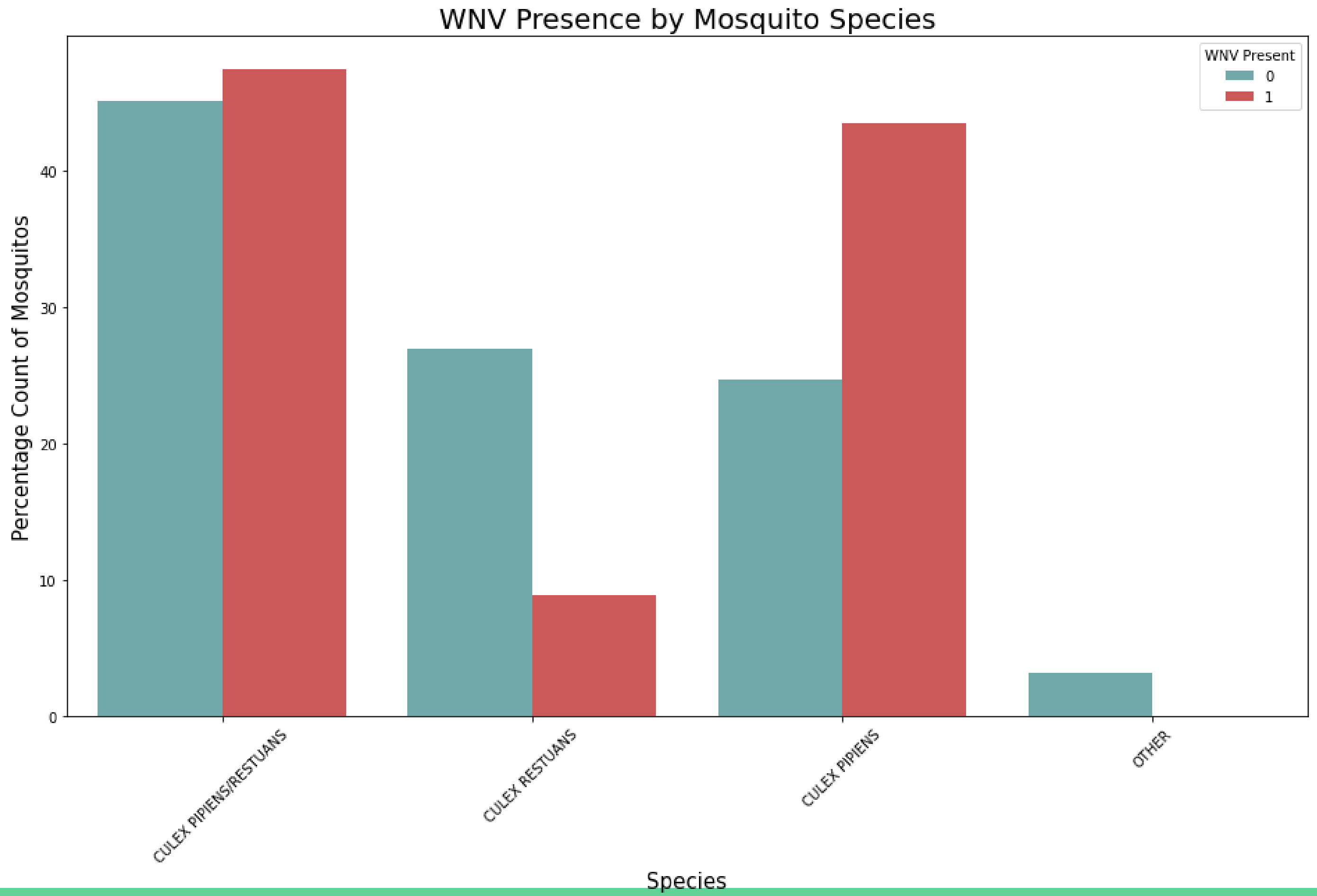
Summer is the season of infection

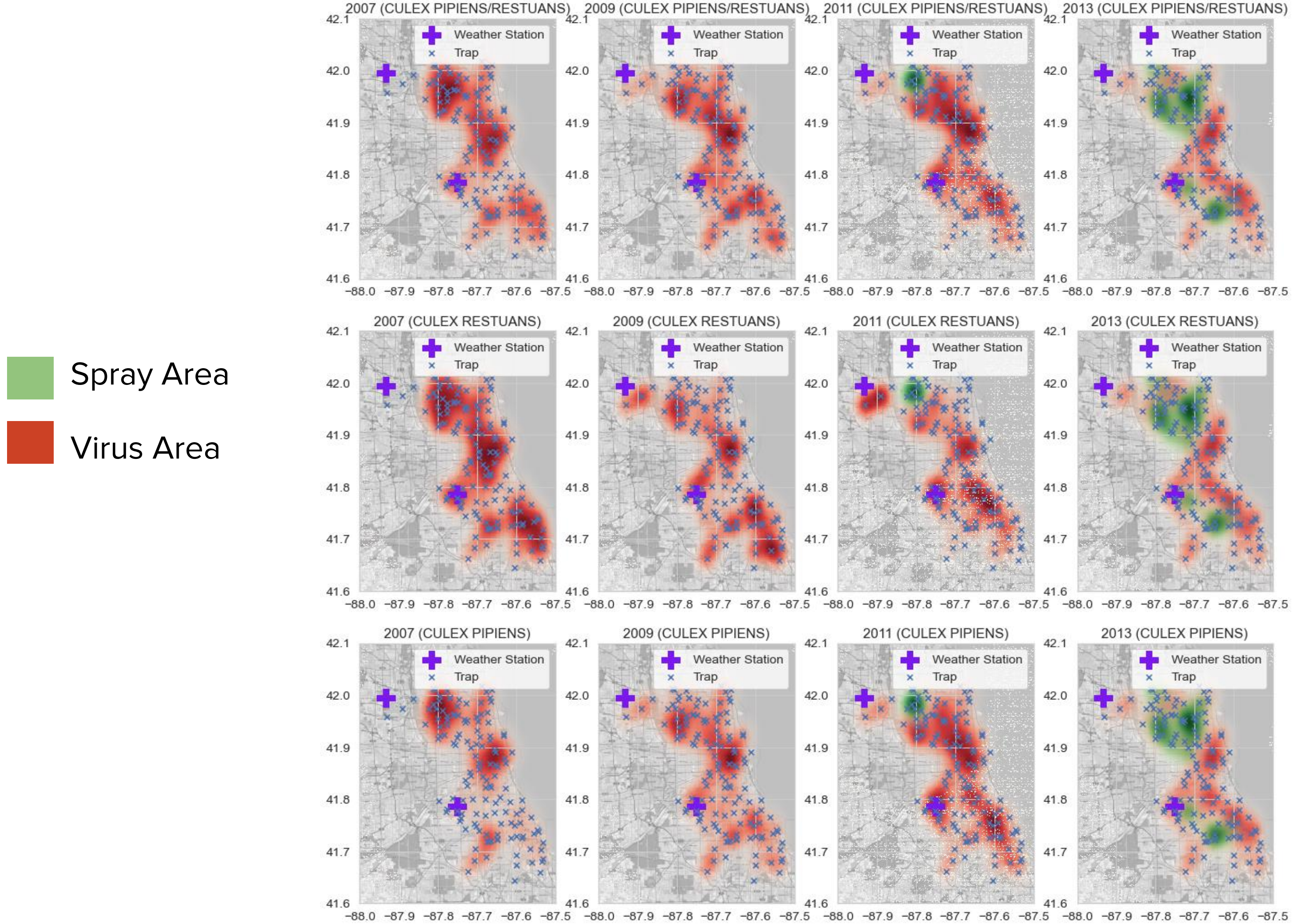


WNV is prevalent in hot and dry conditions

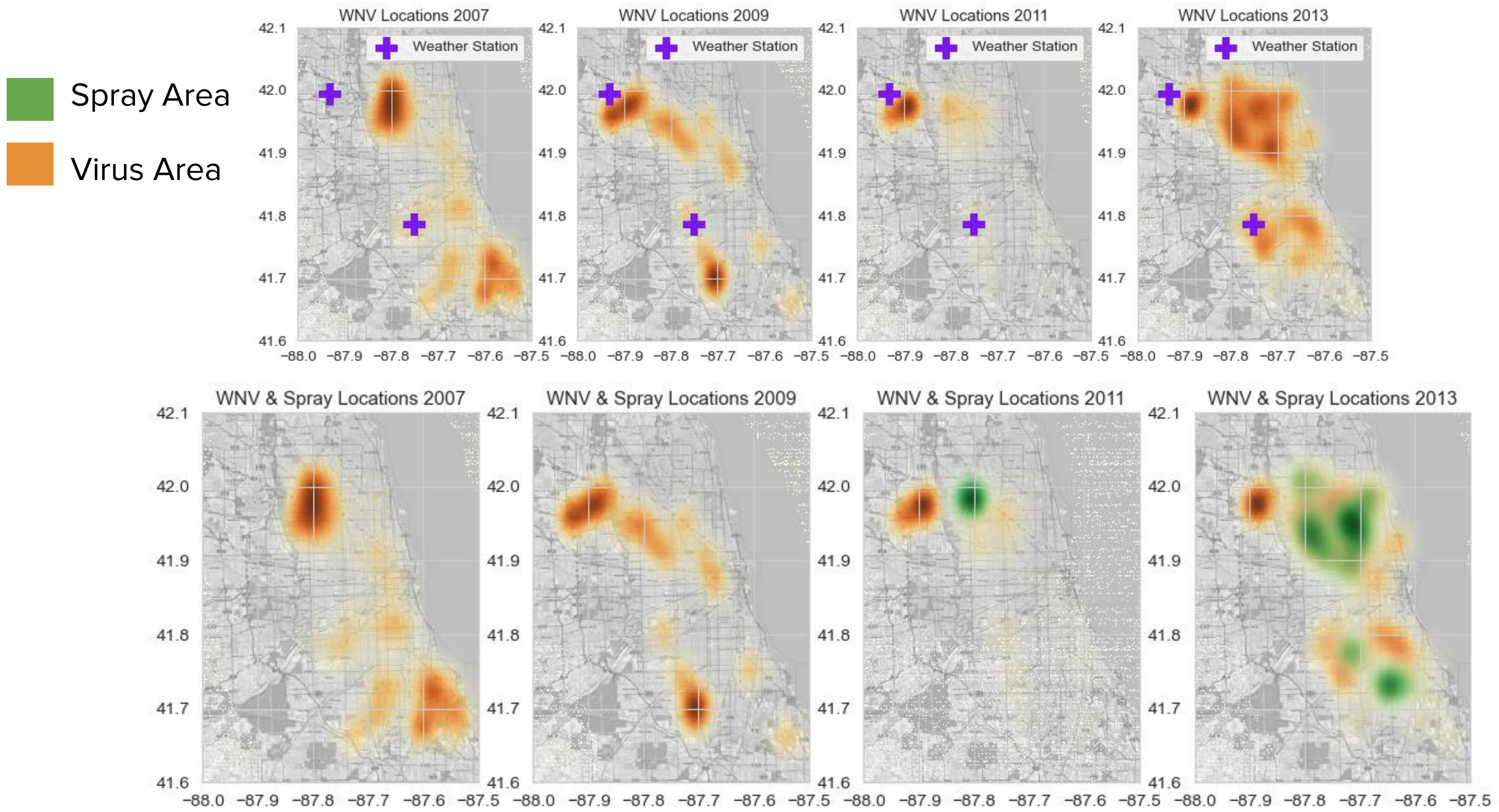


The virus carrying mosquitoes




Spray Area
Virus Area
2007 (CULEX PIPIENS/RESTUANS)
2009 (CULEX PIPIENS/RESTUANS)
2011 (CULEX PIPIENS/RESTUANS)
2013 (CULEX PIPIENS/RESTUANS)
2007 (CULEX RESTUANS)
2009 (CULEX RESTUANS)
2011 (CULEX RESTUANS)
2013 (CULEX RESTUANS)
2007 (CULEX PIPIENS)
2009 (CULEX PIPIENS)
2011 (CULEX PIPIENS)
2013 (CULEX PIPIENS)
Weather Station
Trap
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

More positive case of WNV case in 2007 and 2013



5. Modeling

Models

Model 1
Logistic Regression

PIPELINE

1. *StandardScaler*
2. *SMOTE*
3. *LogisticRegression*

BEST PARAMS

'lr__C': 0.1
'lr__penalty': 'l1'
'sm__k_neighbors': 1

Model 2
Random Forest

PIPELINE

1. *SMOTE*
2. *RandomForest*

BEST PARAMS

'rf__max_depth': 5
'rf__min_samples_split': 4
'rf__n_estimators': 50
'sm__k_neighbors': 1

Model 3
Extra Tree

PIPELINE

1. *SMOTE*
2. *Extra Tree*

BEST PARAMS

‘et__class_weight': 'balanced'
'et__max_depth': 4
'et__min_samples_leaf': 3
'et__n_estimators': 200
'sm__k_neighbors': 1

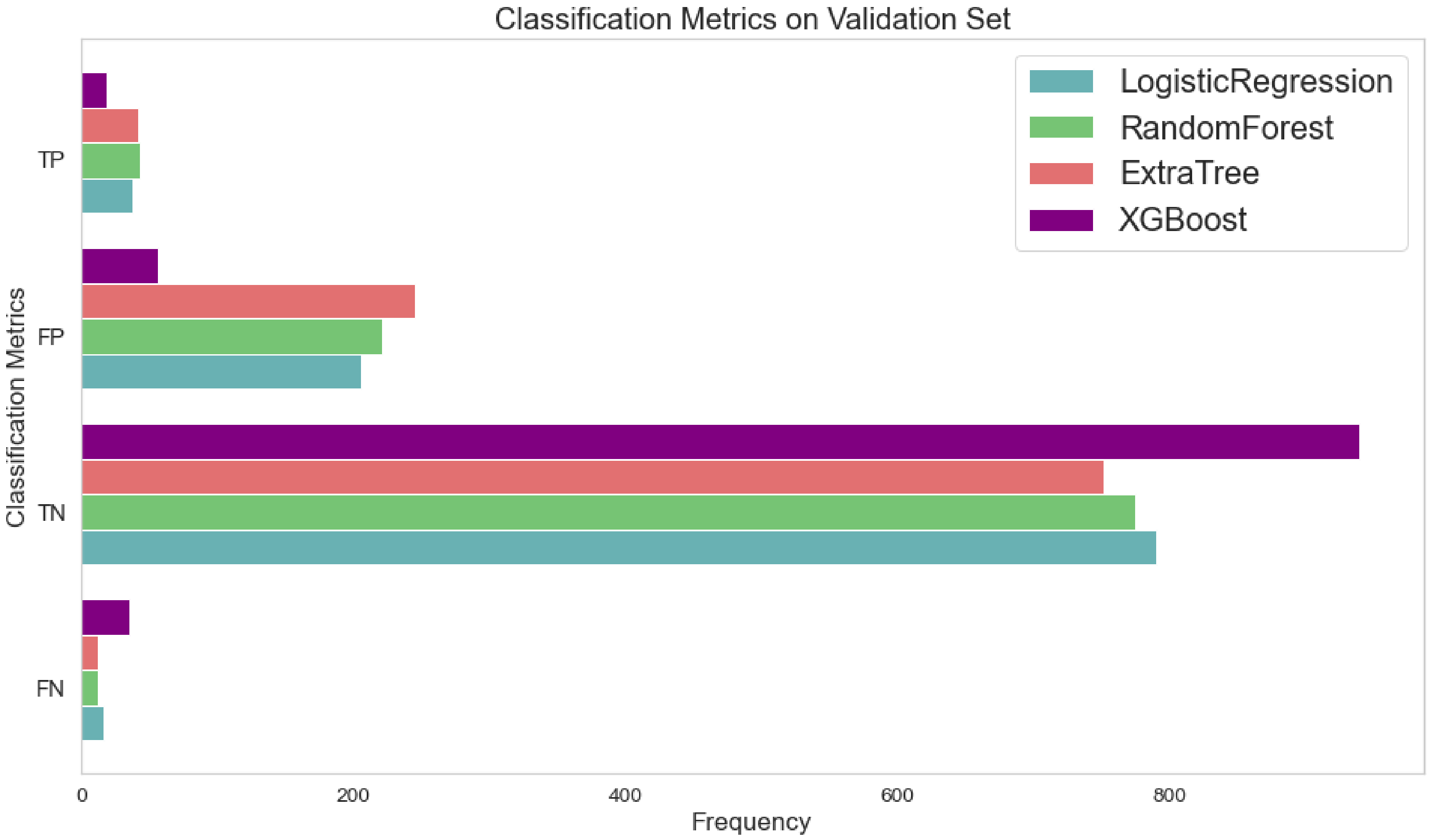
Model 4
XGBoost

PIPELINE

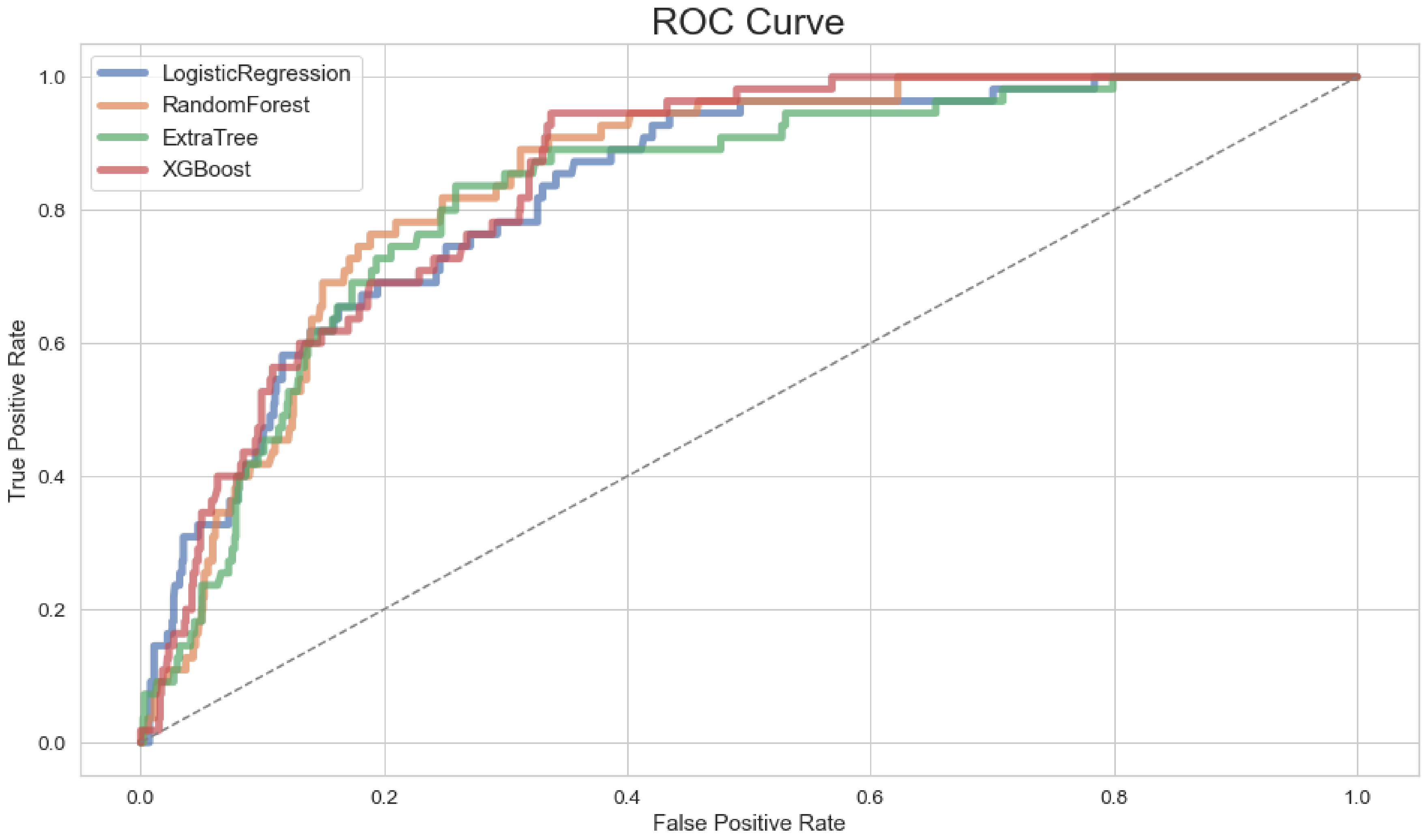
1. *SMOTE*
2. *XGBoost*

BEST PARAMS

'sm__k_neighbors': 5,
'xgb__colsample_bytree': 0.5,
'xgb__eval_metric': 'auc',
'xgb__gamma': 0.1,
'xgb__learning_rate': 0.1,
'xgb__n_estimators': 100,
'xgb__objective':
'binary:logistic',
'xgb__reg_alpha': 0.01,
'xgb__subsample': 0.5



TP: *Correctly Predicting Virus* **FP:** *Incorrectly Predicting Virus* **TN:** *Correctly Predicting No Virus* **FN:** *Incorrectly Predicting No Virus*


ROC Curve
LogisticRegression
RandomForest
ExtraTree
XGBoost
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

Models Evaluation

ROCAUC Score	LogisticRegression	RandomForest	ExtraTree	XGBoost
Training	0.888	0.884	0.861	0.930
CV	0.837	0.856	0.839	0.867
Validation	0.830	0.846	0.825	0.846
Testing/Kaggle	0.756	0.709	0.720	0.695

Model 1
Logistic Regression

PIPELINE

1. *StandardScaler*
2. *SMOTE*
3. *LogisticRegression*

BEST PARAMS

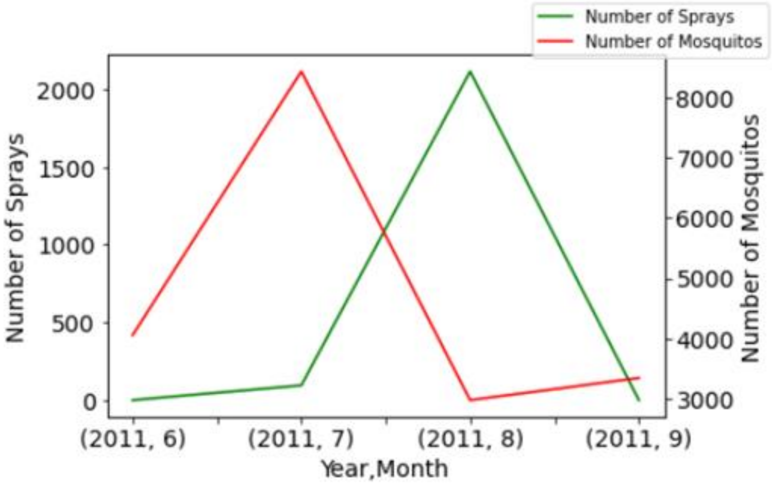
'lr__C': 0.1
'lr__penalty': 'l1'
'sm__k_neighbors': 1

Feature	Exp(coef)
tavg_station1	159.954790
nummosquitos	3.125598
month	2.433159
weekofyear	2.222087
dewpoint_station1	2.089268
avgspeed_station2	1.702942

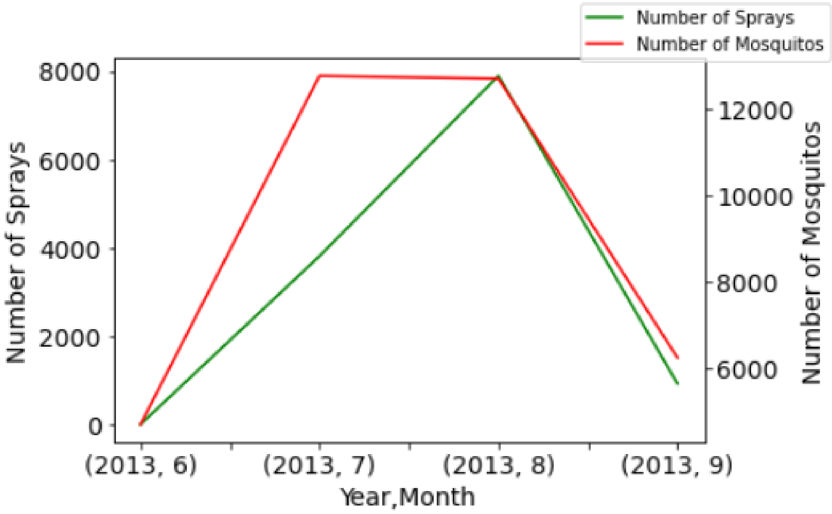
6. Cost Benefit Analysis

Spray should be used to reduce the number of mosquitos

2011



2013



The Cost - Benefit Analysis

Spraying

- **Spray cost**: approximately $900,000 [1] to spray the Chicago, including spray procedures and overtime hours ($1471/km2)

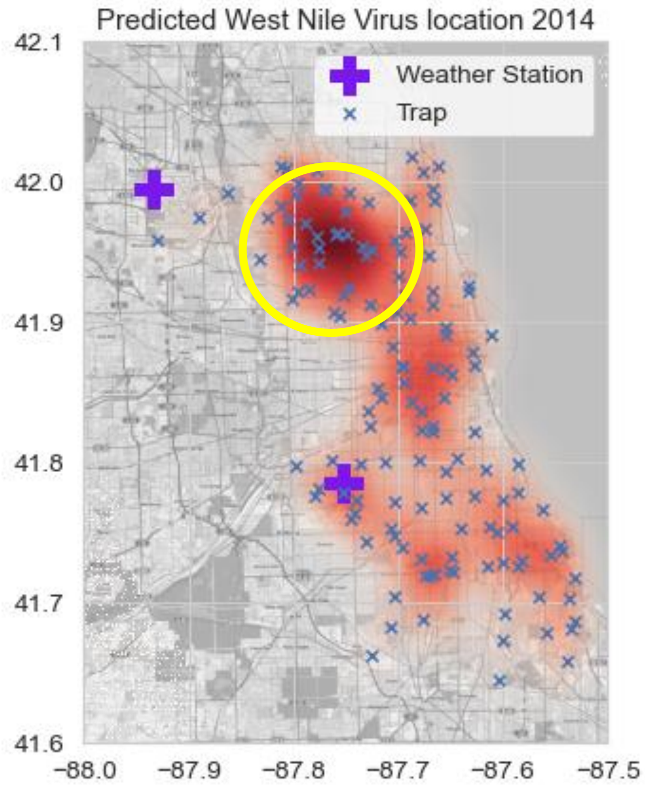
Not Spraying

- **Medical cost**: $7,500 / non severe cases
  Severe cases involving acute flaccid paralysis, a partial- to whole-body paralysis caused by WNV infection could cost an average of **$25,000** [2]
- **2.71 million** people in Chicago have risk of WNV
- **Economic loss**: $56 million / year between 1999 and 2012 [3]

Cost Effectiveness

- Chicago’s GDP Per Capita: $ 61,170 [1]
- Chicago’s Daily GDP Per Capita $ 170
- Recovery Period from severe WNV could be anywhere from a several weeks to several months [2]
- Assuming the best case scenario of 3 weeks for severe cases…
- Per person total cost would be $3570 economic cost + $25,000 medical cost = $ 28,570
- With the spraying/overtime cost of $900,000, only 30 severe WNV cases would be required to make the spraying cost effective in terms of the medical and economic impacts.

Using model to recommend the target areas for sprays



Target Areas & Population

- Albany Park (50,343)
- Irving Park (56,665)
- Portage Park (64,954)
- Avondale (37,909)
- Belmont Cragin (80,648)
- Logan Square (72,724)
- Hermosa (25,489)

Almost 400,000

Estimated Cost

- Total area = 50.5 square kilometer
- Spray cost = $74,300

7. Conclusion & Recommendations

Conclusion

Model

- Our best performing model is Logistic Regression and we achieved an ROC_AUC of 0.756.
- Average temperature was the top predictor with the exponential coefficient at about 160.
- Location was not a strong predictor in our best model, but weather and week of year were more important features.
- Chicago Council should spray more in summer (August) because this month had more risk of WNV in human from the virus carrying mosquito.

Cost-Benefit Analysis

- We found that the spray cost for Chicago would be $900,000 and the total cost for the infected person would be $28,570. The spray cost covers only 30 severe WNV cases.
- The target areas from the model have the total area at 50.5 square kilometer which requires $74,300 of spray cost.

Recommendations

Recommendations

- The sub-urban in Chicago has more risk from WNV due to the poor sanitation system in the older houses compared to new houses [3]. Therefore, the Chicago council should give the sanitation maintenance to the old houses.
- We recommend to always spray to prevent the high medical cost and economic loss, and the spray help Chicago to prevent the unpredictability of WNV outbreaks in people

Next Steps

- Further improve the model
    - Consider the human behaviour
    - Examine the type and efficiency of spray
    - Consider the income of people in each area
    - Built model to predict outbreaks more reliability in humans