Project 4 : West Nile Virus Prediction

General Assembly DSI 14

July 8, 2020

Song Yuan, Qi Wen, Jin, Eng Seng

Problem Statement

The main goal here for our analytics effort is to put systems in place that reduce people's exposure to mosquitoes that carry WNV

- Gene Leynes, Data Scientist in City of Chicago Dept of Innovation and Technology

Questions?

How to reduce WNV incidences?

Where and how does WNV appear?

How to make prevention methods more effective?

Cost Benefit?

History of West Nile Virus In Chicago

1999

WNV Appears

West Nile Virus Exist In United States

2002

64 Deaths

Total of 884 Cases Reported

2005 & 2006

10 & 13 Deaths

2012 & 2013

12 Deaths

THE GOAL IS 0 DEATHS

TABLE OF CONTENTS

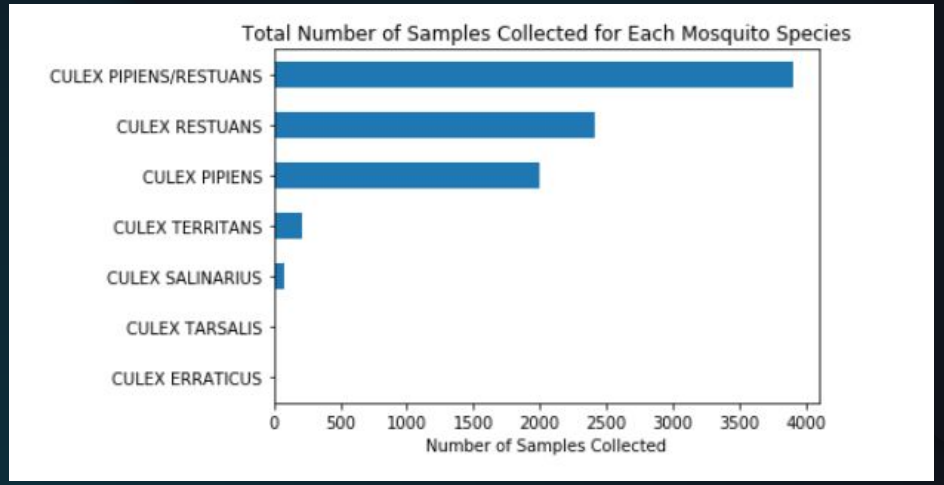
01

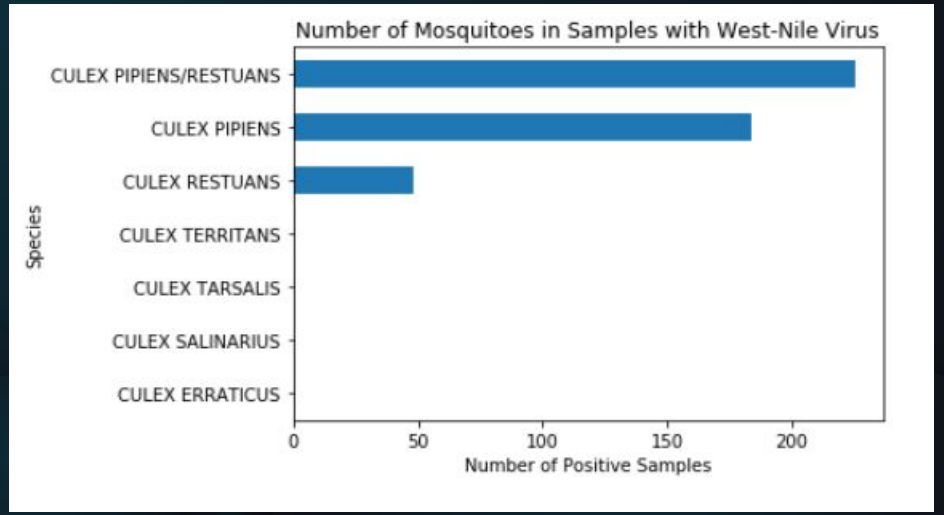
About West Nile Virus Vectors – EDAs

Mosquito Species Samples Collected:

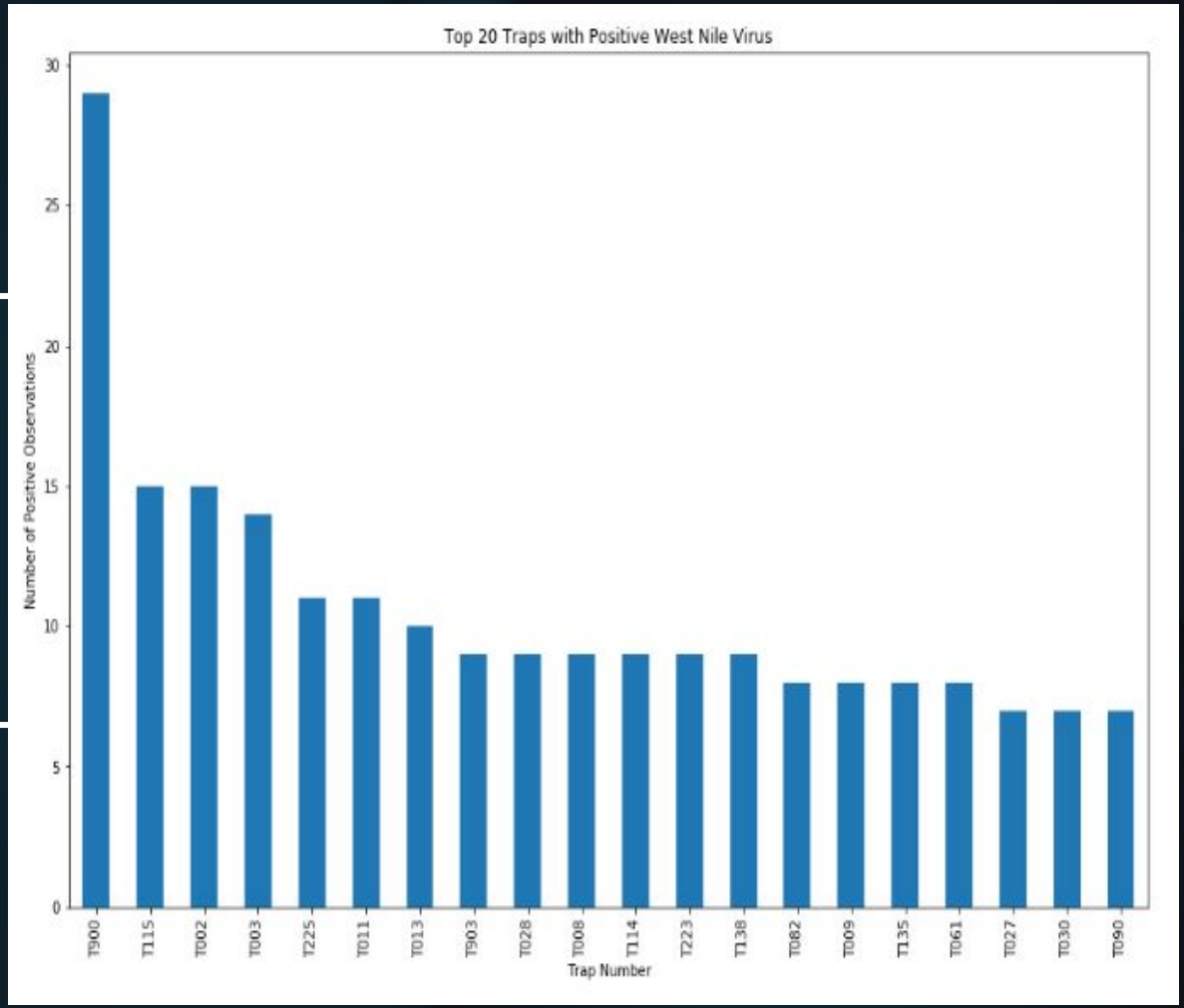
- 3 Major Species

C. Restuans, C. Pipiens, and C. Pipiens/Restuans

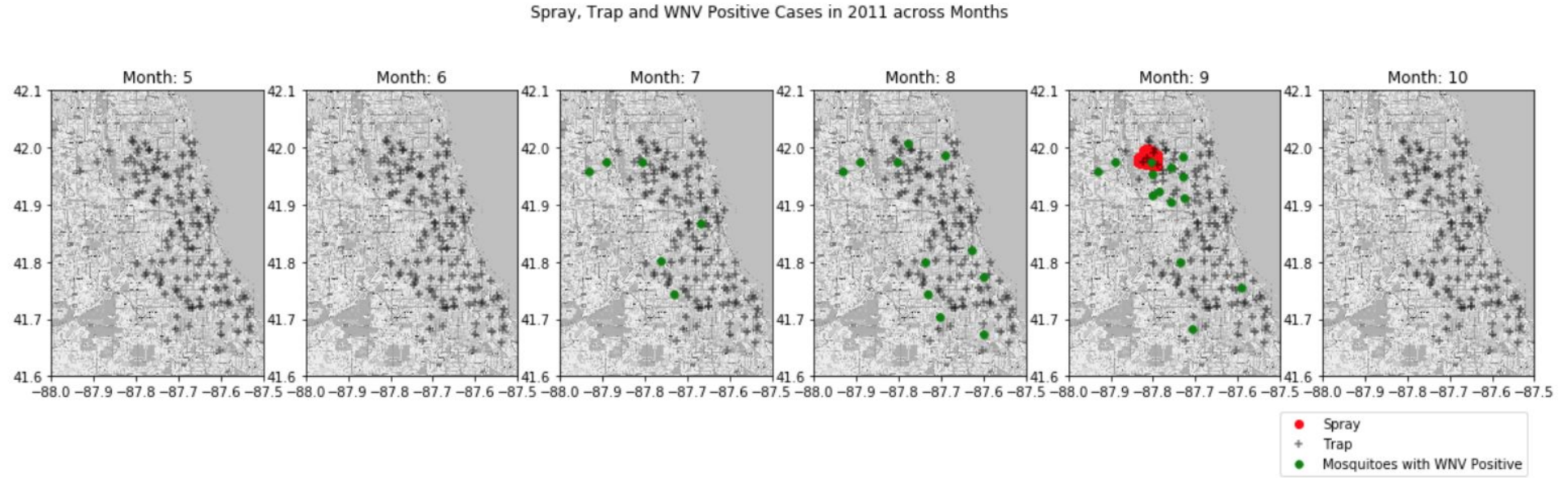




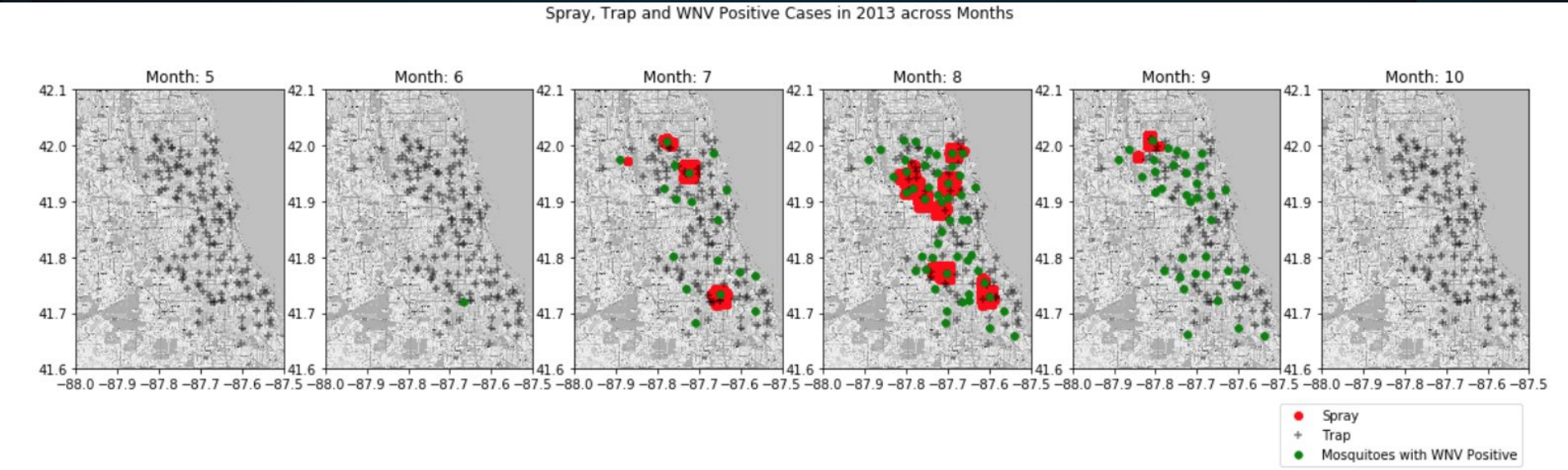
Top 20 Traps With WNV Recorded



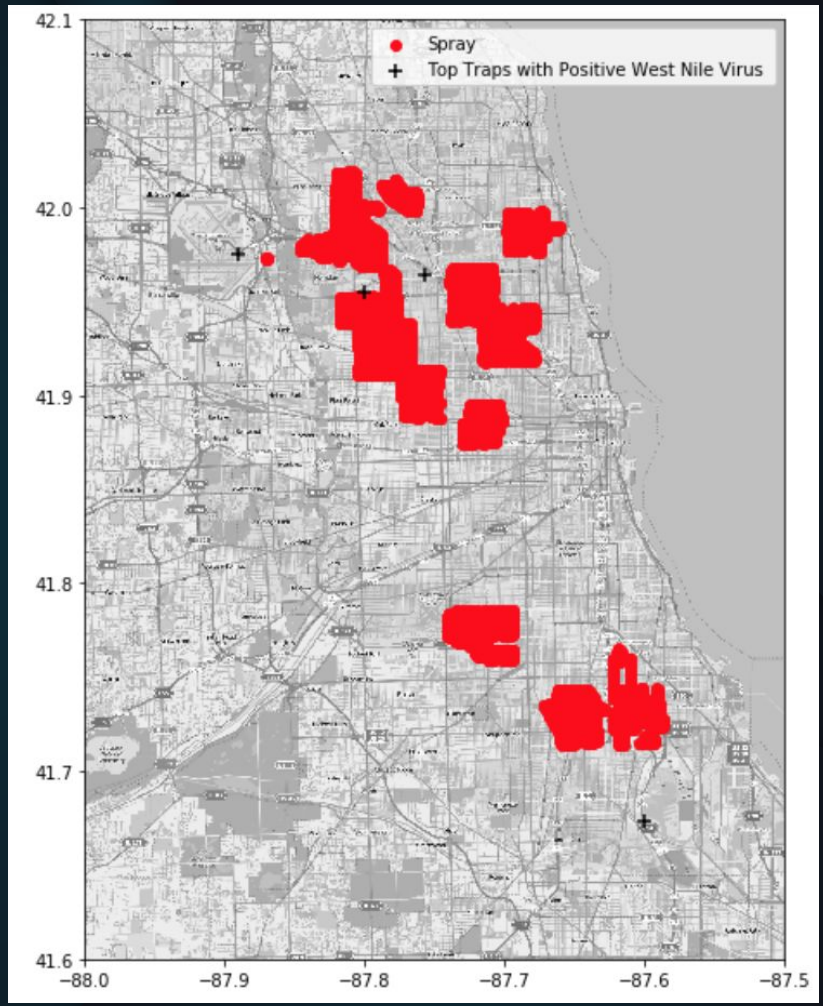
WNV Positive Cases, Spray and Trap Mapping – 2011



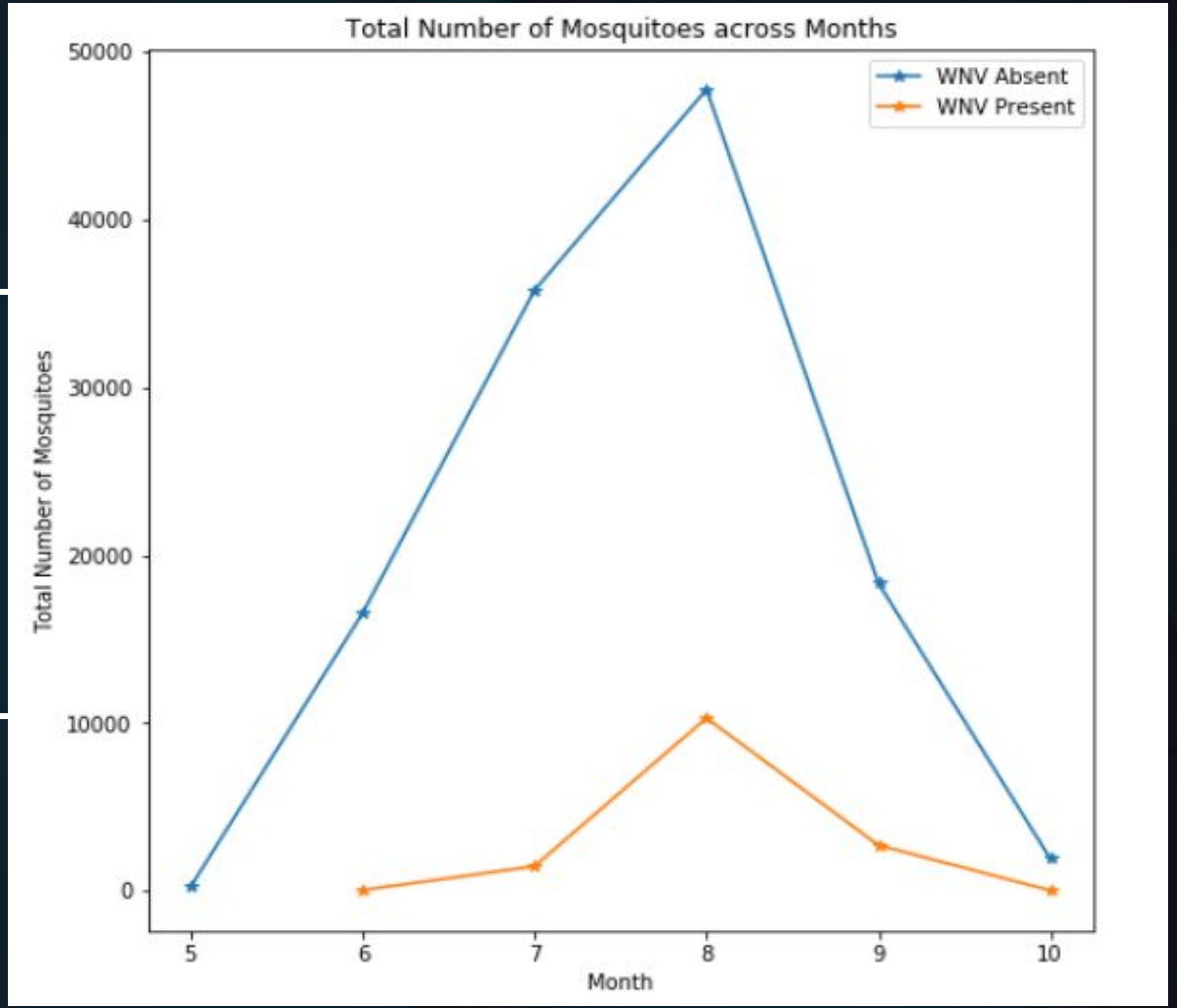
WNV Positive Cases, Spray and Trap Mapping – 2013



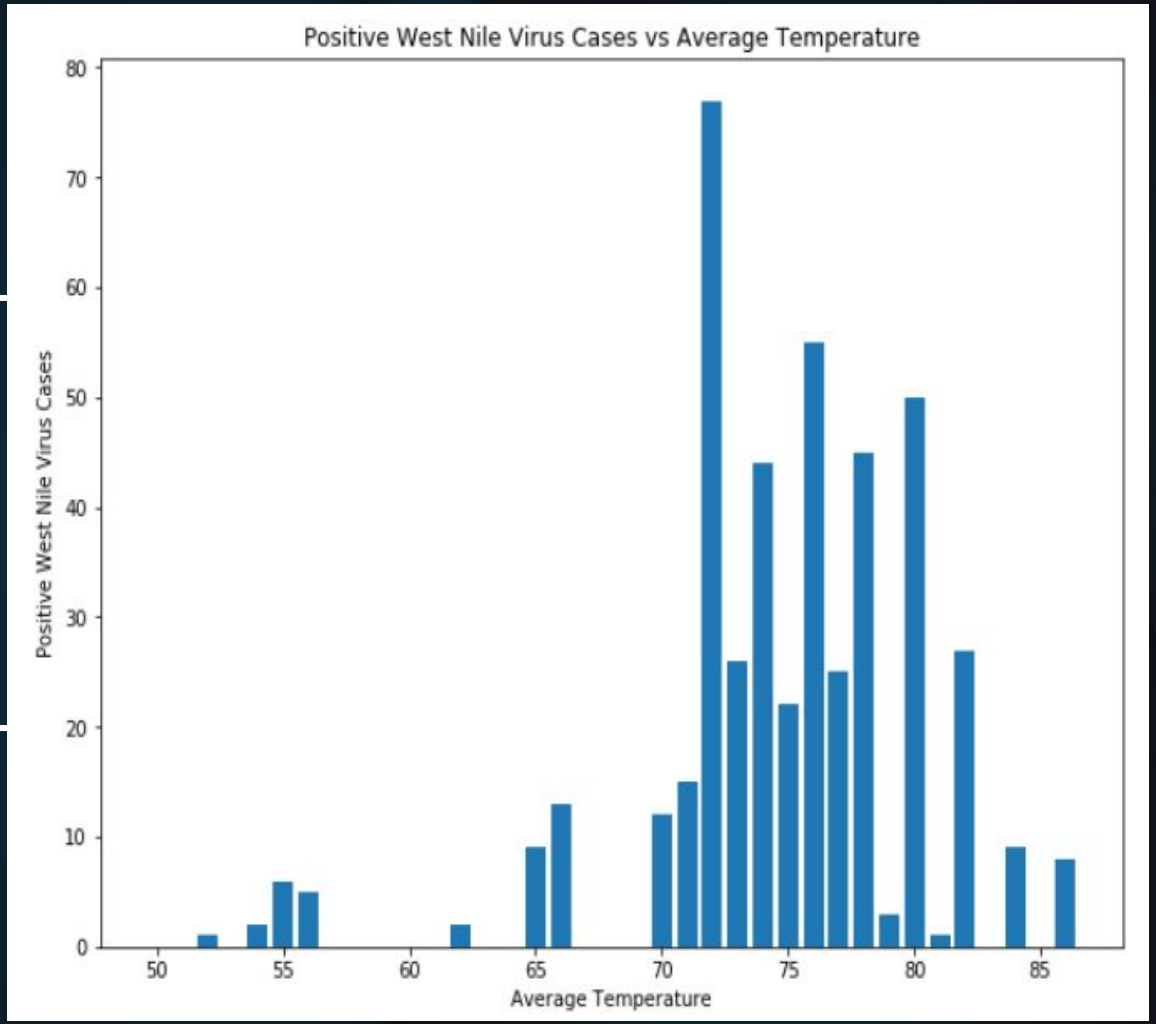
Spray and
Top hotspots with WNV Virus



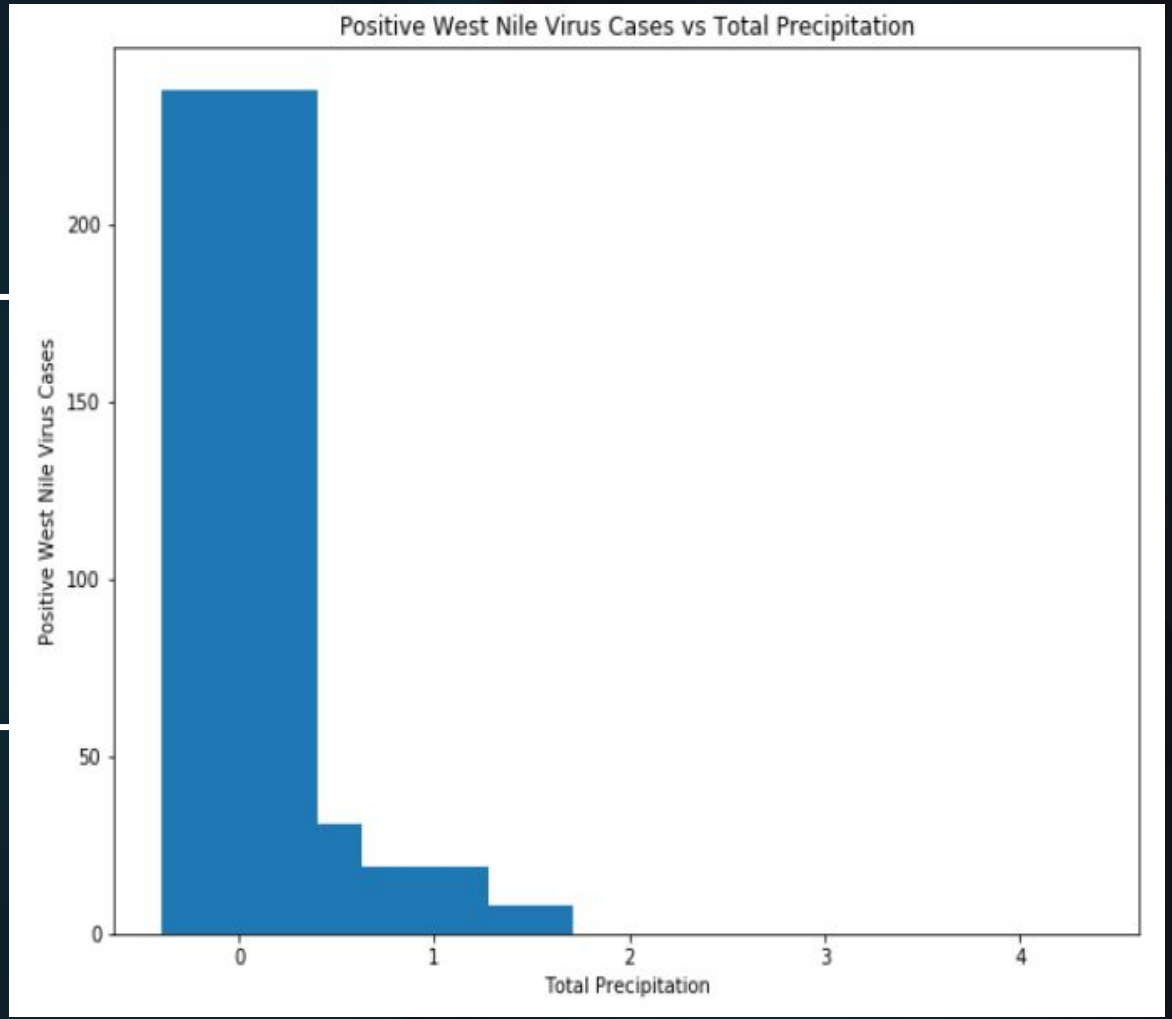
Summer Is the SEASON OF INFECTION



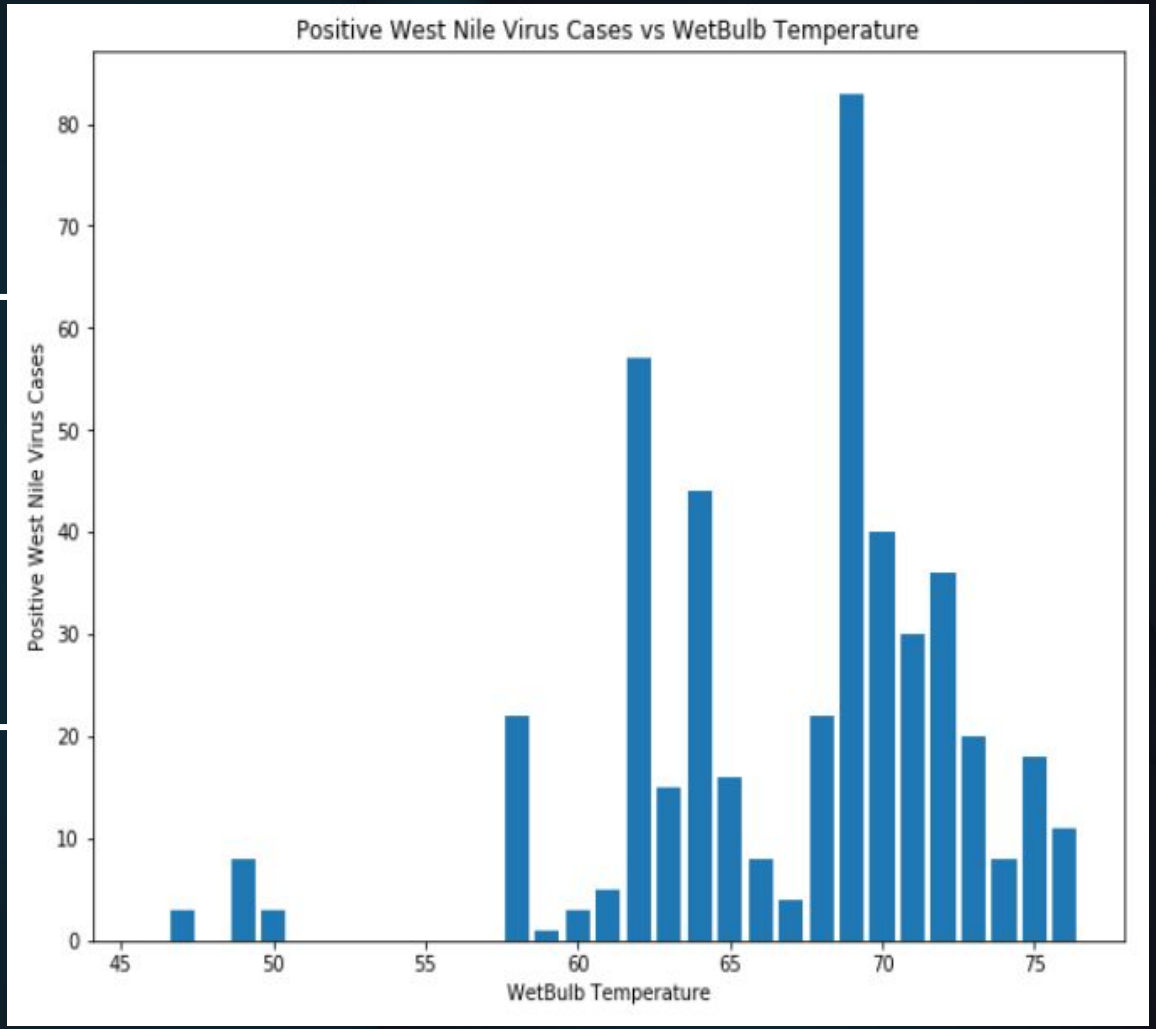
Does Temperature Affect the Existence of WNV?



When the weather precipitation is low, more WNV detected.



The higher humidity the more frequent WNV was detected.



02
Modeling

Dataset

Initial Dataset

Training Data - trap locations, number of mosquitoes, presence of West Nile Virus (2007, 2009, 2011, 2013)

Weather Data (from NOAA) - Weather conditions from 2007 to 2014, during the months of the tests

Spray Data - in 2011, 2013

Testing Data - predict results for 2008, 2010, 2012, and 2014

Cleaning & Preprocessing

Weather Data
- Drop correlated columns, columns with more than 50% missing values

Train & Test Data
- Drop address columns
- Create new columns for total number of mosquitoes

Spray
- Drop column with many duplicates and missing values

Fusing Data: Match the nearest weather station to each trap

One Hot Encoding: Species and Trap number (168 features)

Final Dataset

Split train dataset into training and validation data

- Training: SMOTE oversampling to 11414 rows
- Validation: 2583 rows

Testing Set: 116293 rows

How We Built the Models...

- The key output is score which indicating the risk that a specific point/area could test positive for WNV in particular time.
- With Variety models tested, we keep the features which will improve the predictive ability
- For each trap, the model will predict the probability the WNV would be exist

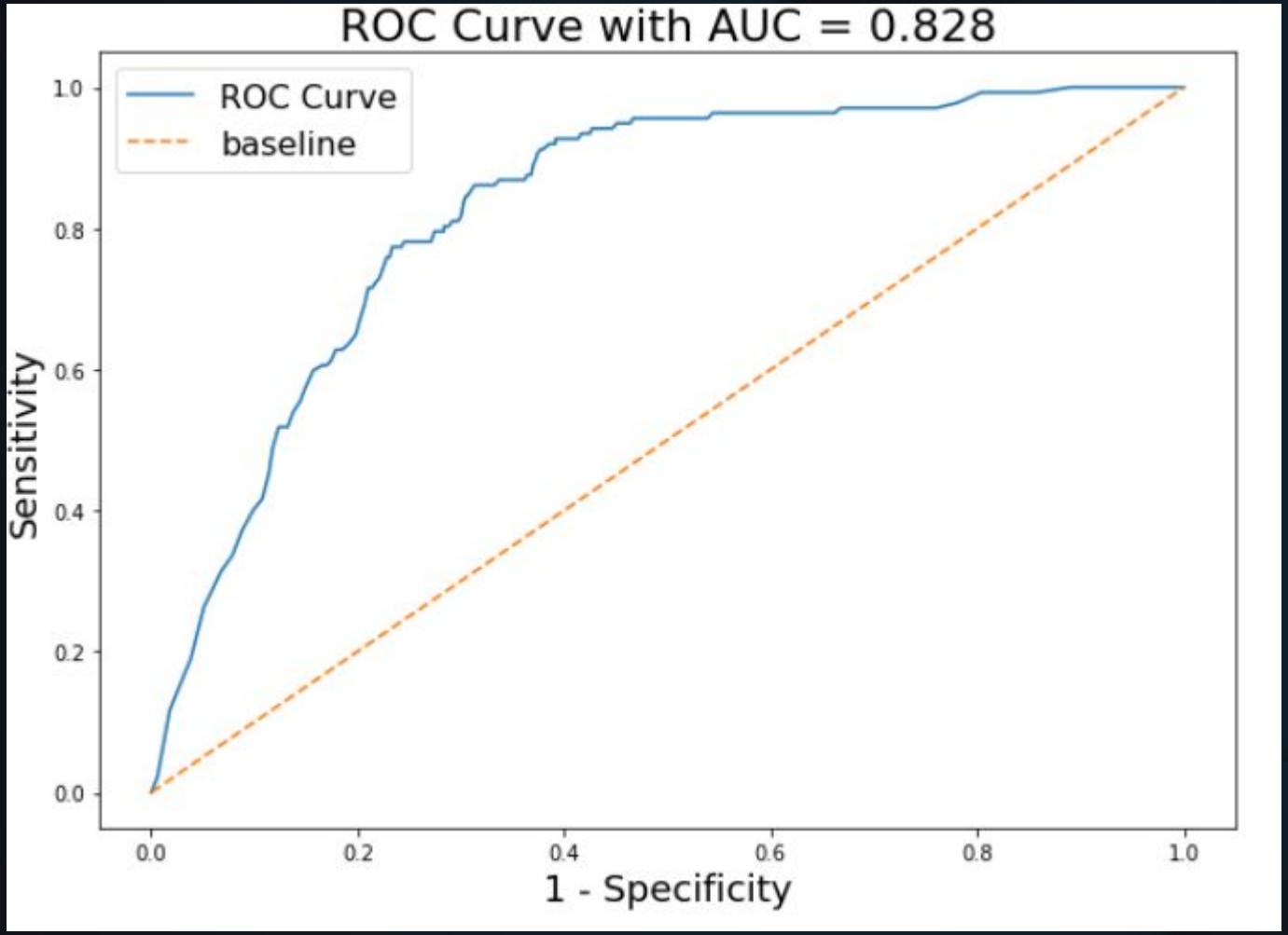
Modeling Scoring

MODELS (SMOTE)	BEST PARAMS	KAGGLE PRIVATE SCORE	KAGGLE PUBLIC SCORE
Logistic Regression (lbfgs)	C = 1, Penalty = l2, max_iter = 1000	0.5719	0.5022
Logistic Regression (liblinear)	C = 1, Penalty = l1, max_iter = 1000	0.5719	0.5022
Decision Tree	max_depth = 10, min_samples = 4, min_samples_split = 20	0.5505	0.5873
Random Forest	max_depth = 20, max_leaf_nodes = 20, min_samples_leaf = 10, n_estimator = 100	0.5765	0.5882
AdaBoost	learning_rate = 1.0, n_estimators = 100	0.5463	0.5481
XGBoost	eval_metric = ‘auc’, scale_pos_weight = 18, subsample = 0.5, eta = 0.2	0.68208	0.6974

Features	Importance
Month	0.036464
Trap_T095	0.028762
Trap_T225	0.018460
CULEX TERRITANS	0.017817
DATE	0.017134
Trap_T094	0.016628
Trap_T158	0.015939
Trap_T080	0.015407
Trap_T063	0.015346
Trap_T102	0.015210

Results

The most predictive feature



ROC CURVE WITH AUC

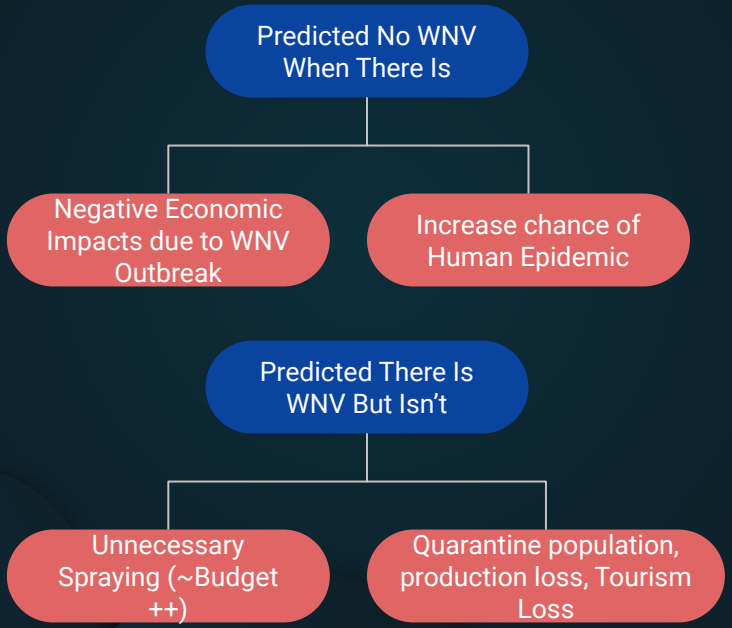
Scoring : 0.828

03

Cost Benefit Analysis

What are the real cost and benefits of vector control?

The Confusion Matrix



The Confusion Matrix

```
Best Score: 0.9890168225013817
Best Params: {'xg__eta': 0.2, 'xg__eval_metric': 'auc', 'xg__scale_pos_weight': 18, 'xg__subsample': 0.8}
Score of training data: 0.8966181881899422
Score of validation data: 0.7642276422764228
Precision score of the model: 0.15395894428152493
Sensitivity score of the model: 0.7664233576642335
              precision    recall  f1-score   support

         0.0       0.98      0.76      0.86      2446
         1.0       0.15      0.77      0.26       137

    accuracy                           0.76      2583
   macro avg       0.57      0.77      0.56      2583
weighted avg       0.94      0.76      0.83      2583
```

	Predict No WNV	Predict WNV
Actual No WNV	1869	577
Actual WNV	32	105

906 Cases

Reported From Chicago from Year 2005 to 2016

$ 12,977,000

Estimated Medical Cost from Year 2005 to 2016

$ 702,000

49 Cases in 2016. Costing $ 702k in medical expenses and lost productivity

148,510 Acres

Chicago city Area

$ 137,000 estimated

Estimated for worst case scenario of spraying whole city

5 Times

Medical Expenses & Loss Productivity cost incurred 5 times more expensive than spraying in the city !

04

Conclusion

Key Findings

- Mosquito numbers follow the natural onset of the season, begin in May and end in October.

- Peak at August and decreasing at October (died / hibernation due to winter)

- Spray Operations by City of Chicago and Department of Public Health do not have noticeable effect.

- Past spray operations had missed locations that reported the most number of positive cases

- Strong Correlation between temperature and precipitation with number of WNV positive!

- Most WNV positive samples found between 71F - 80F, zero precipitation.

Recommendations

- Better spraying regime that is informed by weather and trap data.
- Concentrate efforts to combat WNV prone areas which tend to fall into two categories
  - Inner suburbs, fairly high income, 1940s - 1960s housing
  - City, Low income, young .
- More public awareness campaigns. Urging people to minimize exposure with necessary precautions steps.

Limitations & Further Research

- The dataset was extremely imbalanced and we had to generate synthetic samples to balance the dataset
- Spraying data was limited to only 2 years with no indications of dosage
- Due to computational power limitations, we had only tuned limited combinations of hyperparameters
- Other mosquito control efforts have not been accounted for in our model and that hindered our efforts to determine the actual effectiveness of spraying
- For a more complete analysis on economical and costings, we should collect the data in terms of healthcare expenses and mosquito control efforts.

THANKS!

ANY QUESTIONS ???