Project 4

West Nile Virus Prediction

Problem Statement

Company

City of Chicago and Chicago Department of Public Health

Context

Great Deal of Variation in West Nile Virus in intensity and duration since 2002 in Chicago.

Up to 29%Fatality Rate, depending on Age.

Difficult to Predict and Allocate Resources.

Proposal

Build a Classification Model:

To Predict Highly Accurately, the Outbreaks of WNV from Mosquitos, based on Environmental and Other Variables.

Exploratory Data Analysis

Main dataset:

Mosquito trap (date, location, species, number of mosquitos, WNV present)

Weather data:

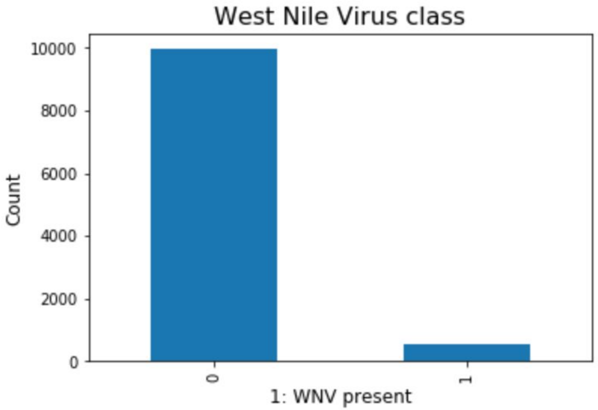
Weather conditions during the months of test

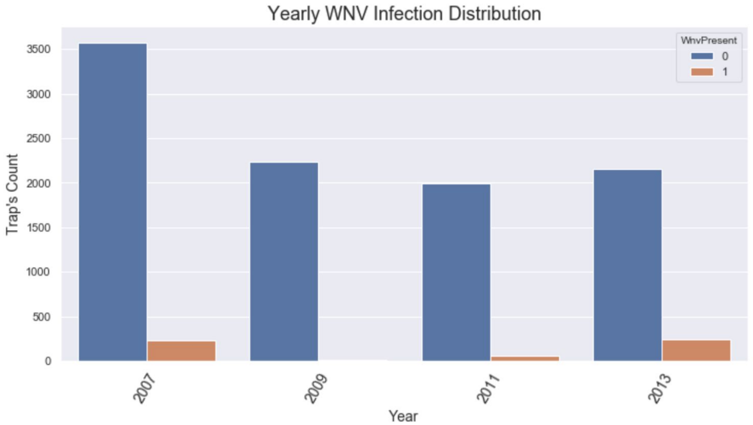
Spray dataset:

GIS data for spray effort

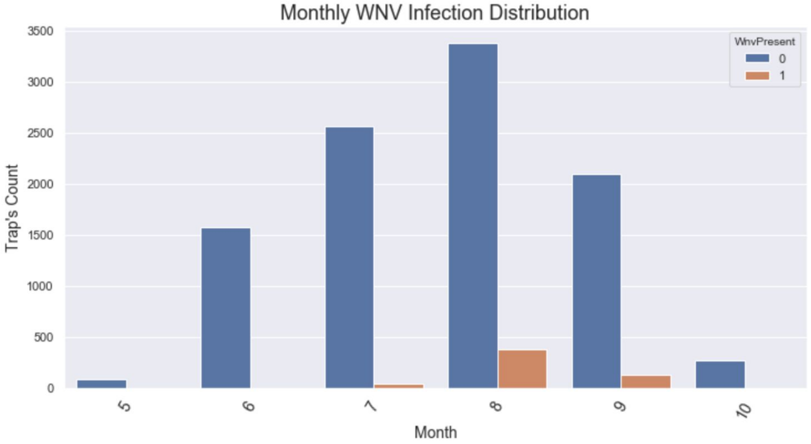
Train data

Wnv Present



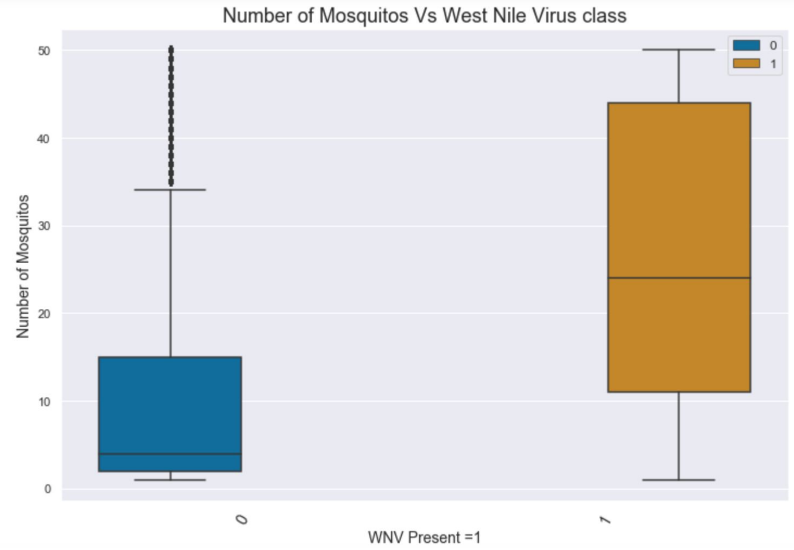


Wnv Infection

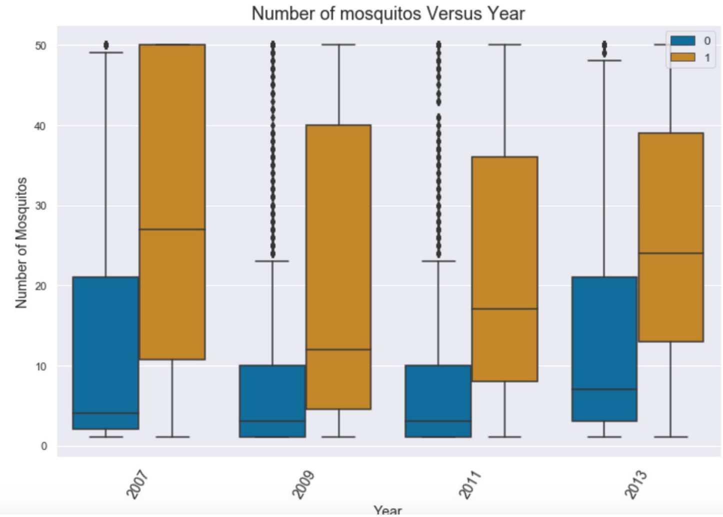


Train data

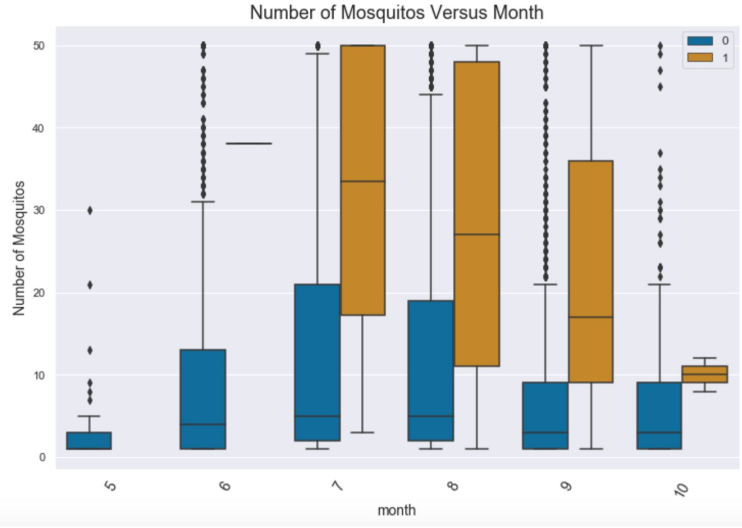
Number of Mosquitos



Number of Mosquitos Vs West Nile Virus class



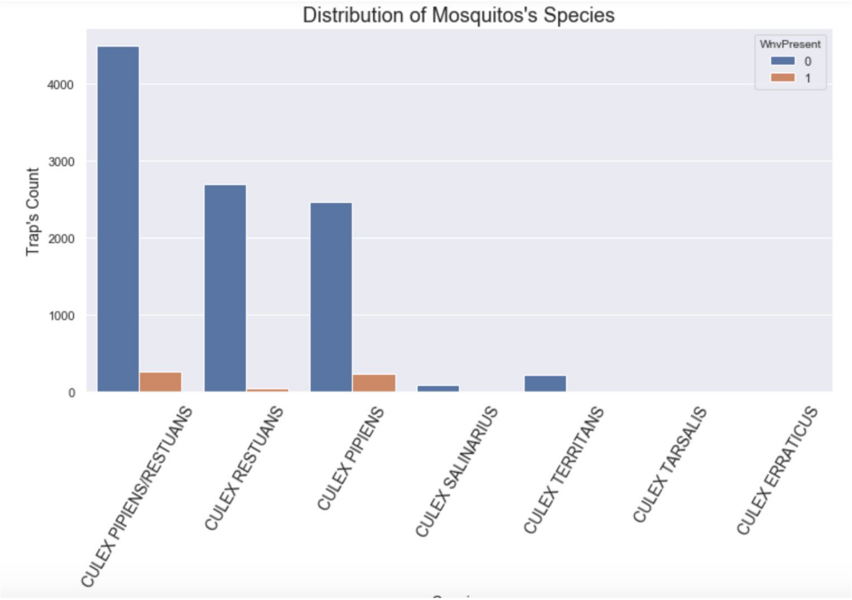
Number of mosquitos Versus Year

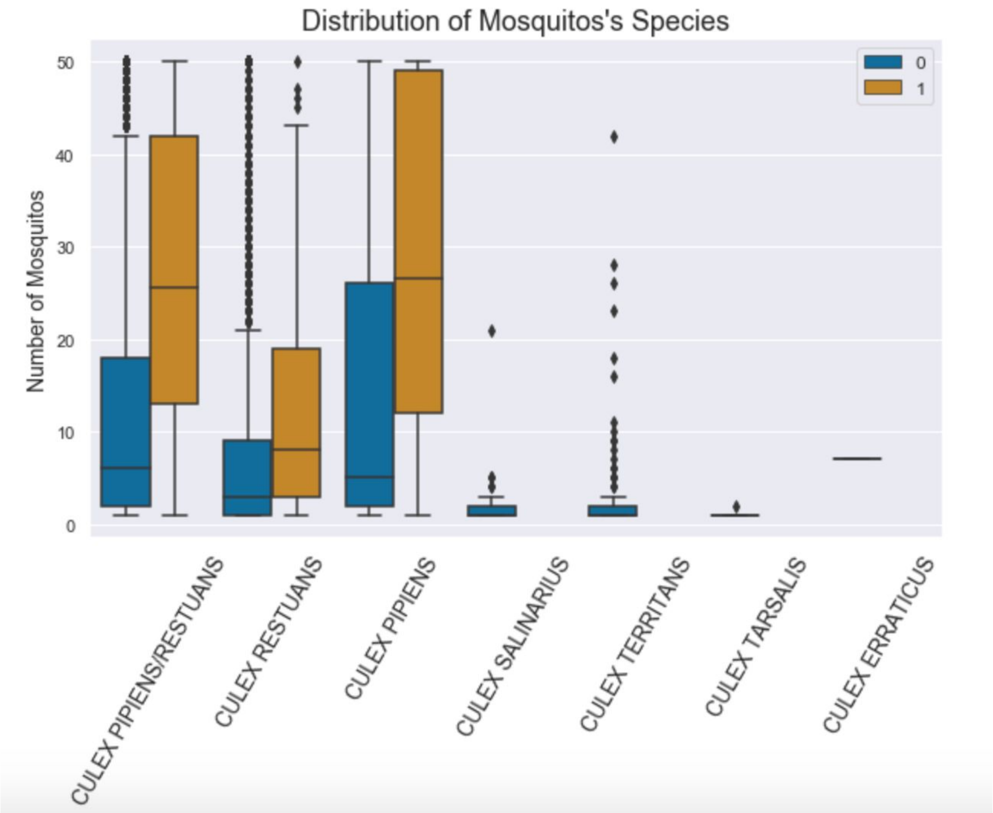


Number of Mosquitos Versus Month

Train data

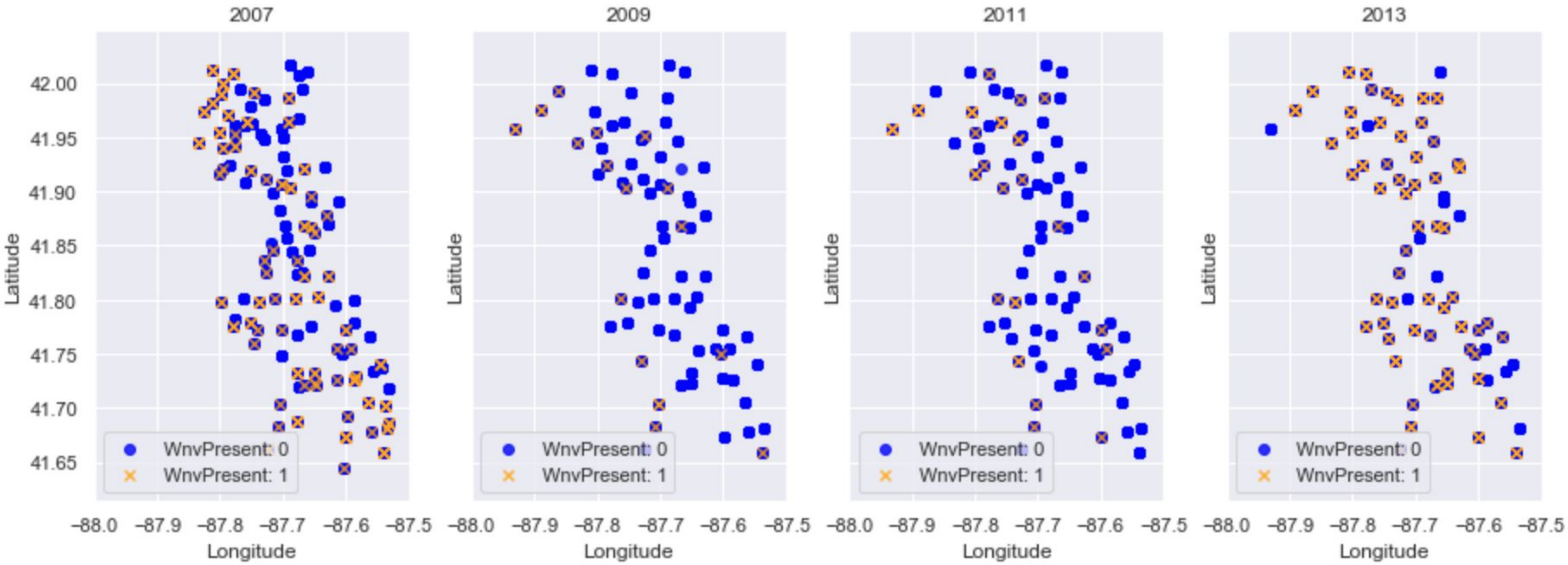
Species

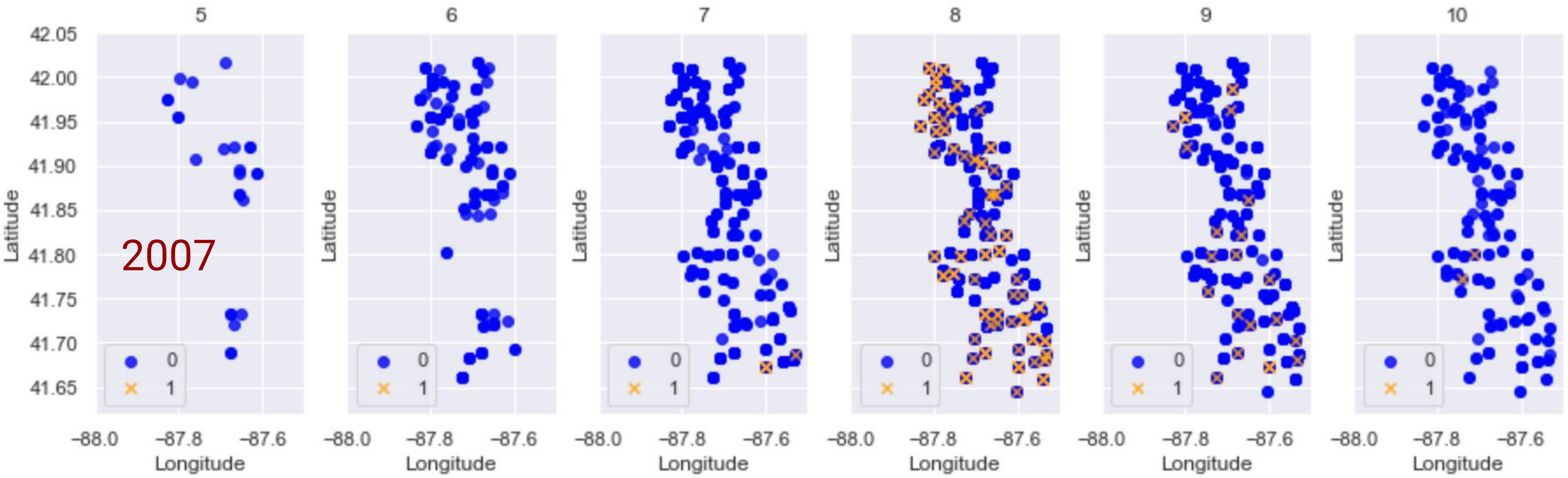




Train data

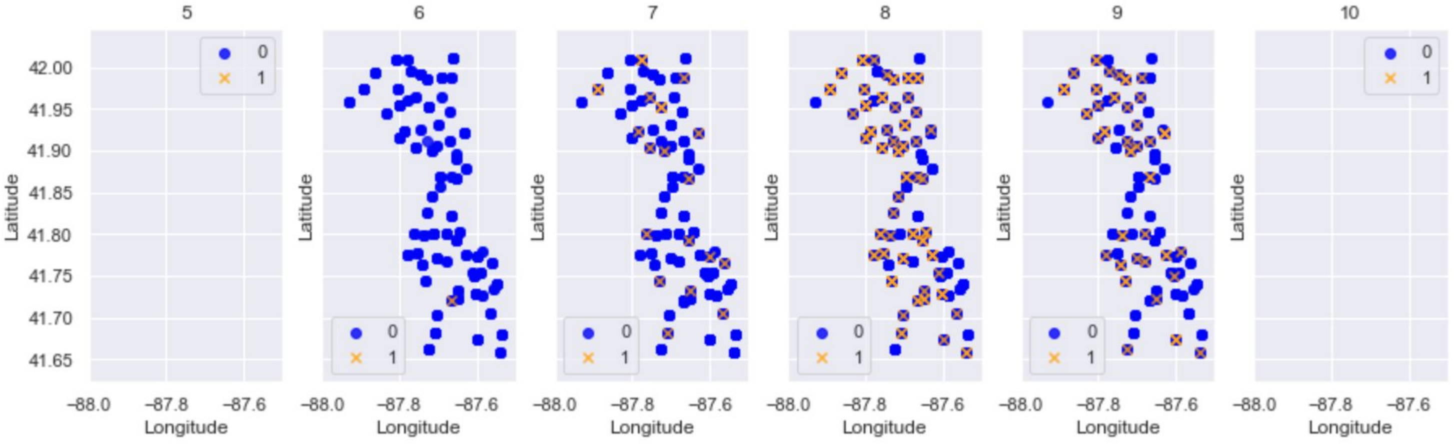
Location





2007

5
6
7
8
9
10
Latitude
42.05
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
Longitude
−88.0
−87.8
−87.6
● 0
× 1



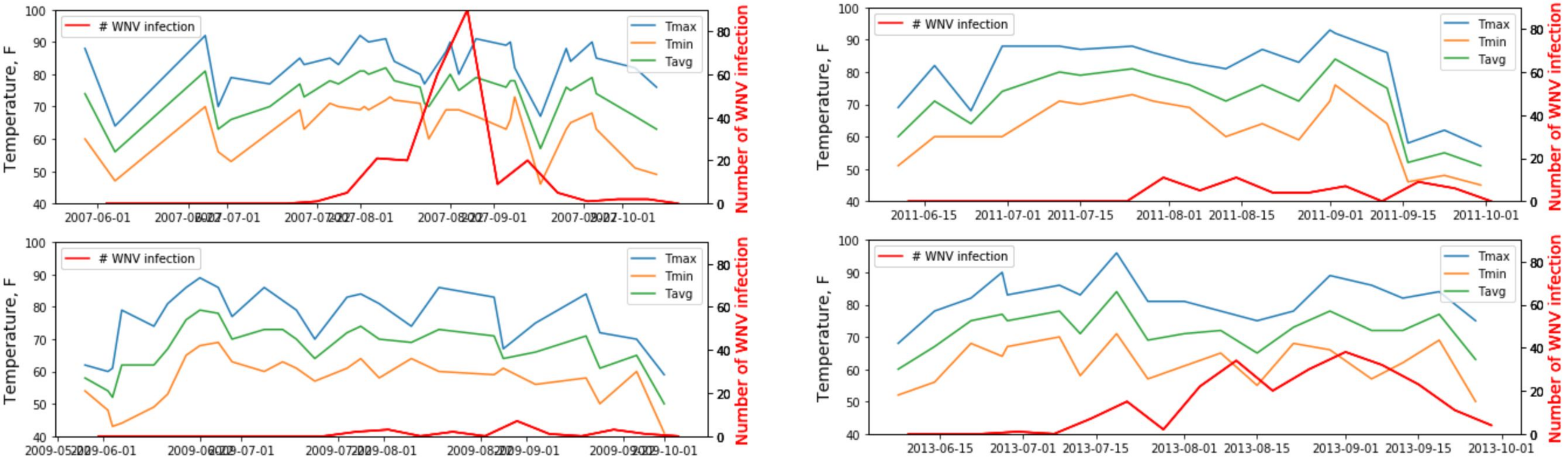
2013

5
6
7
8
9
10
Latitude
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
Longitude
−88.0
−87.8
−87.6
● 0
× 1

Weather data

It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet.

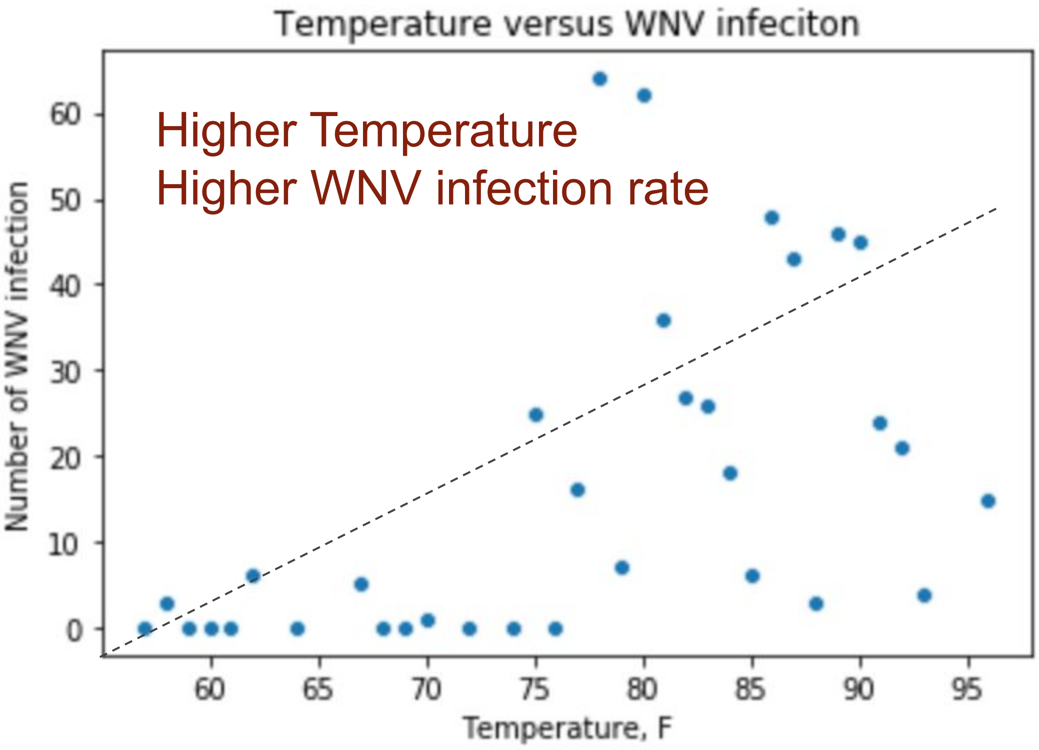
Temperature



Weather data

Temperature

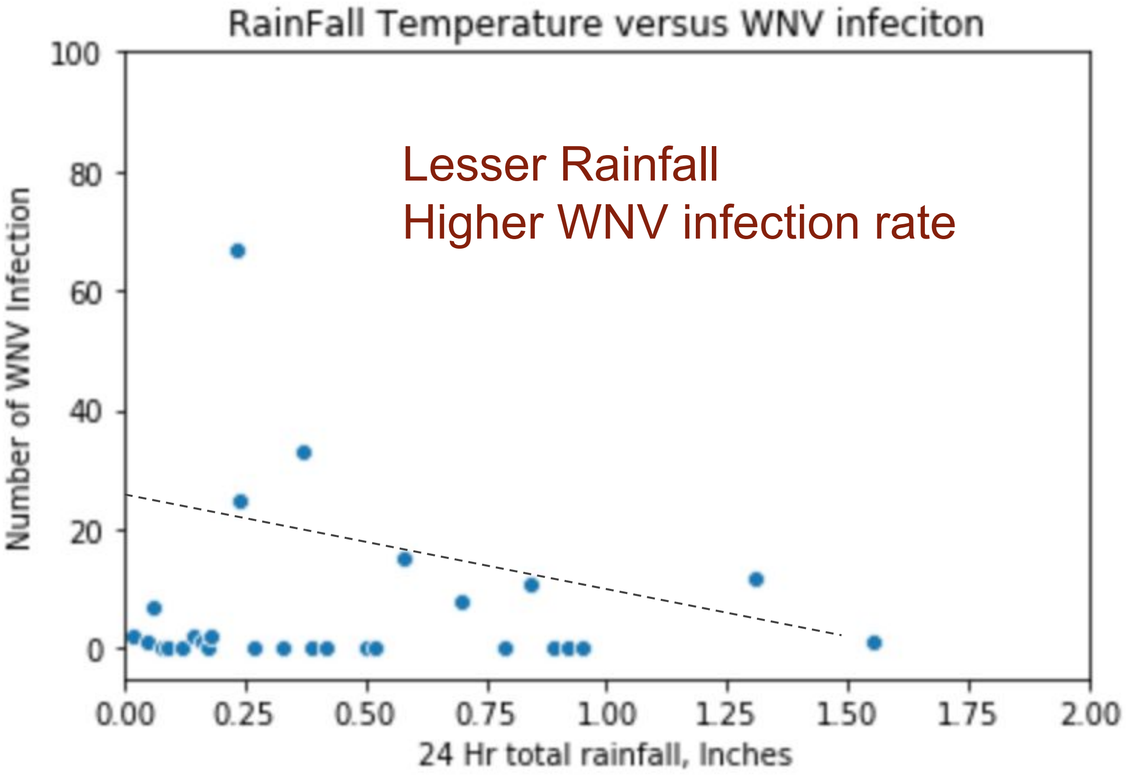
It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet.



Higher Temperature
Higher WNV infection rate

Weather data

Rainfall



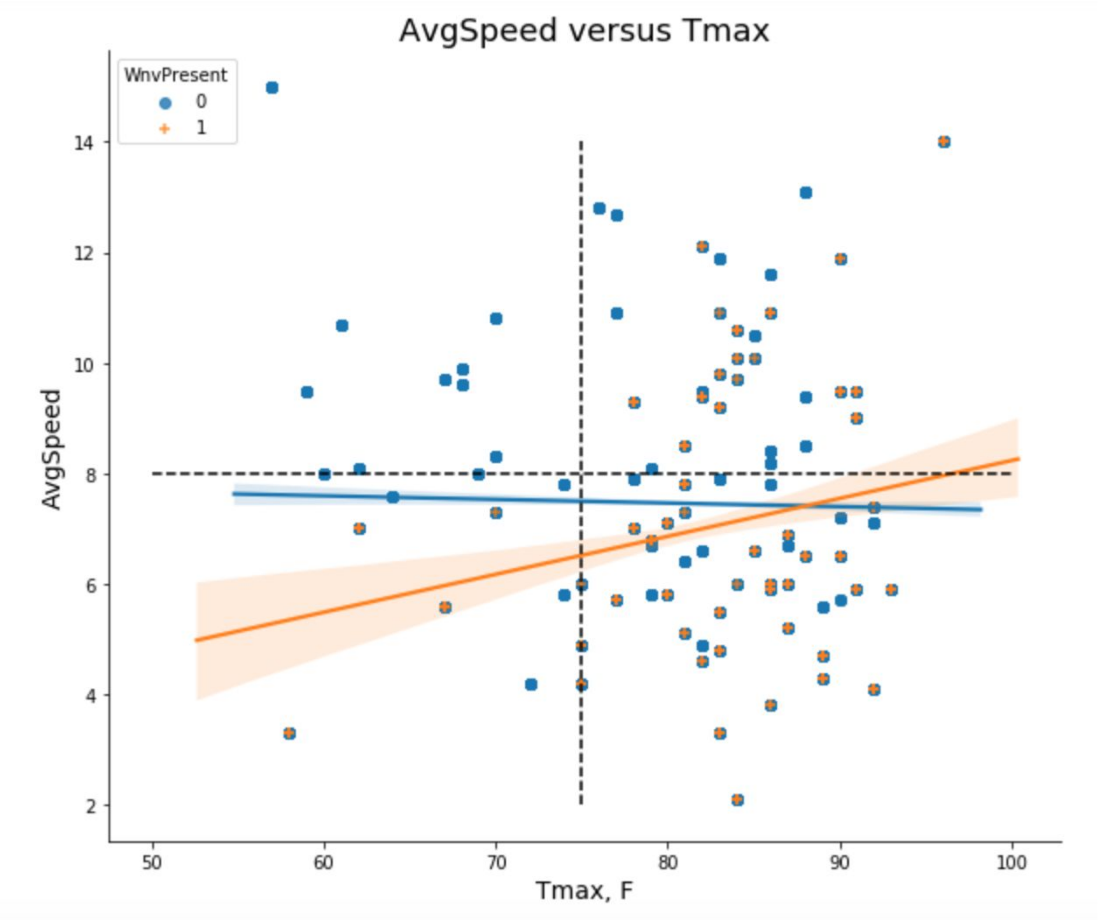
Lesser Rainfall
Higher WNV infection rate

Weather data

Temperature

+

Wind Speed

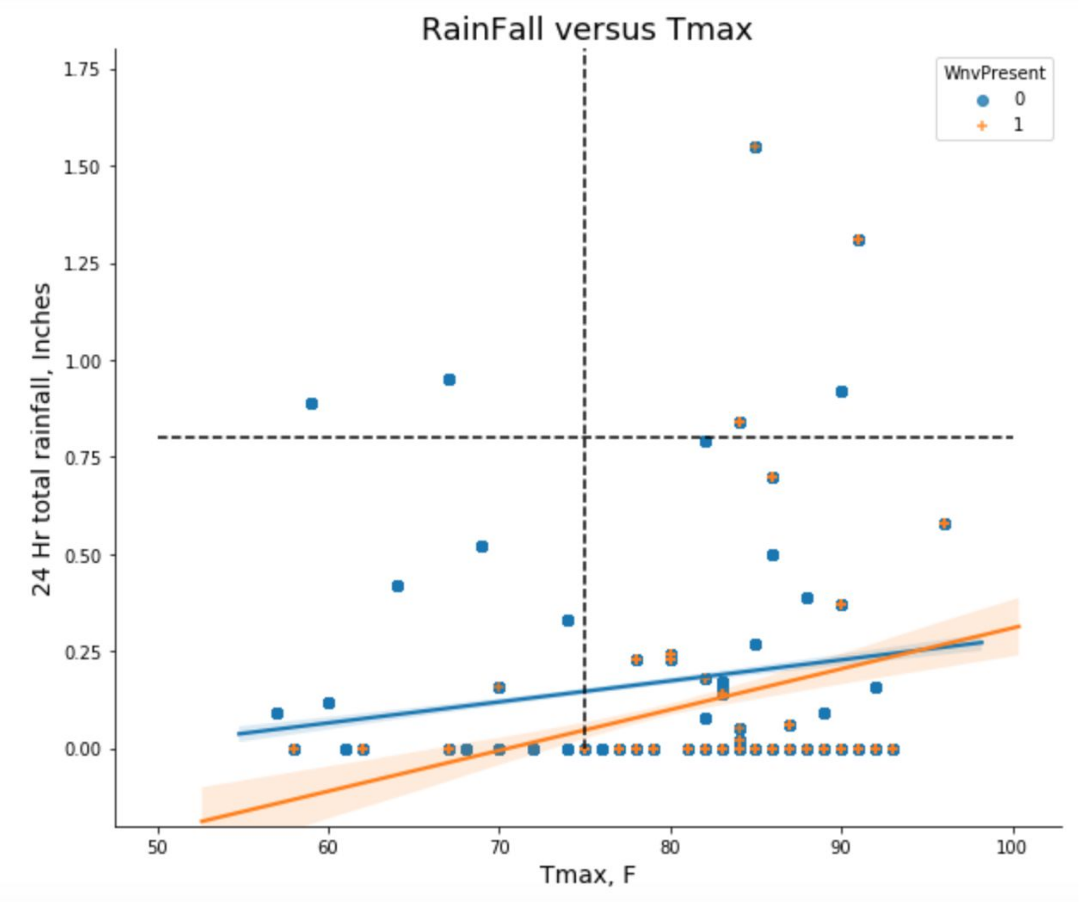


Weather data

Temperature

+

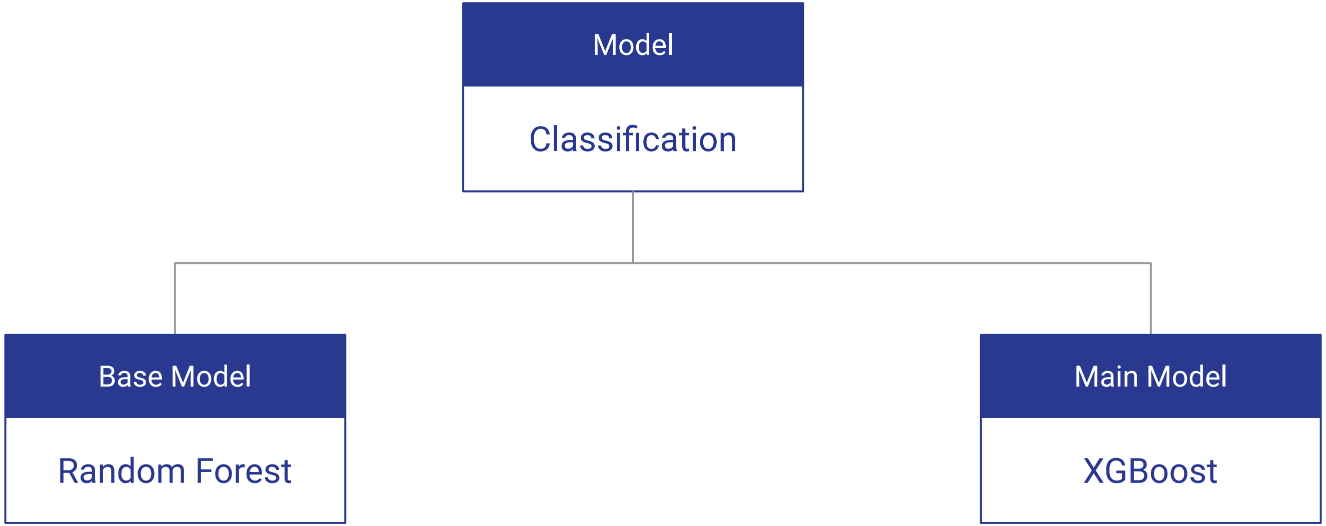
Rainfall



Predictive Model

Predictive Model

WNV present (1) or NOT present(0)



Model

Classification

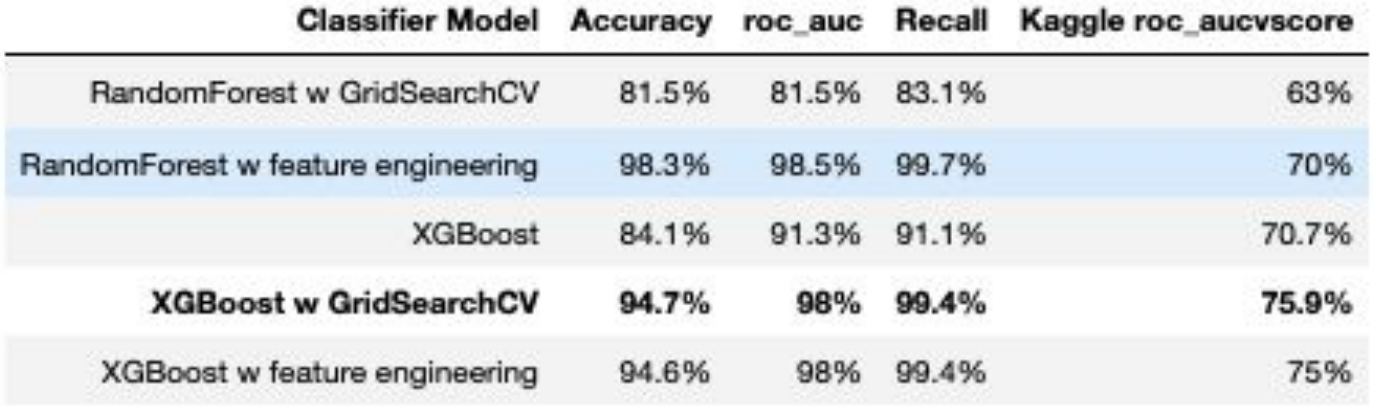
Base Model

Random Forest

Main Model

XGBoost

Model Performance



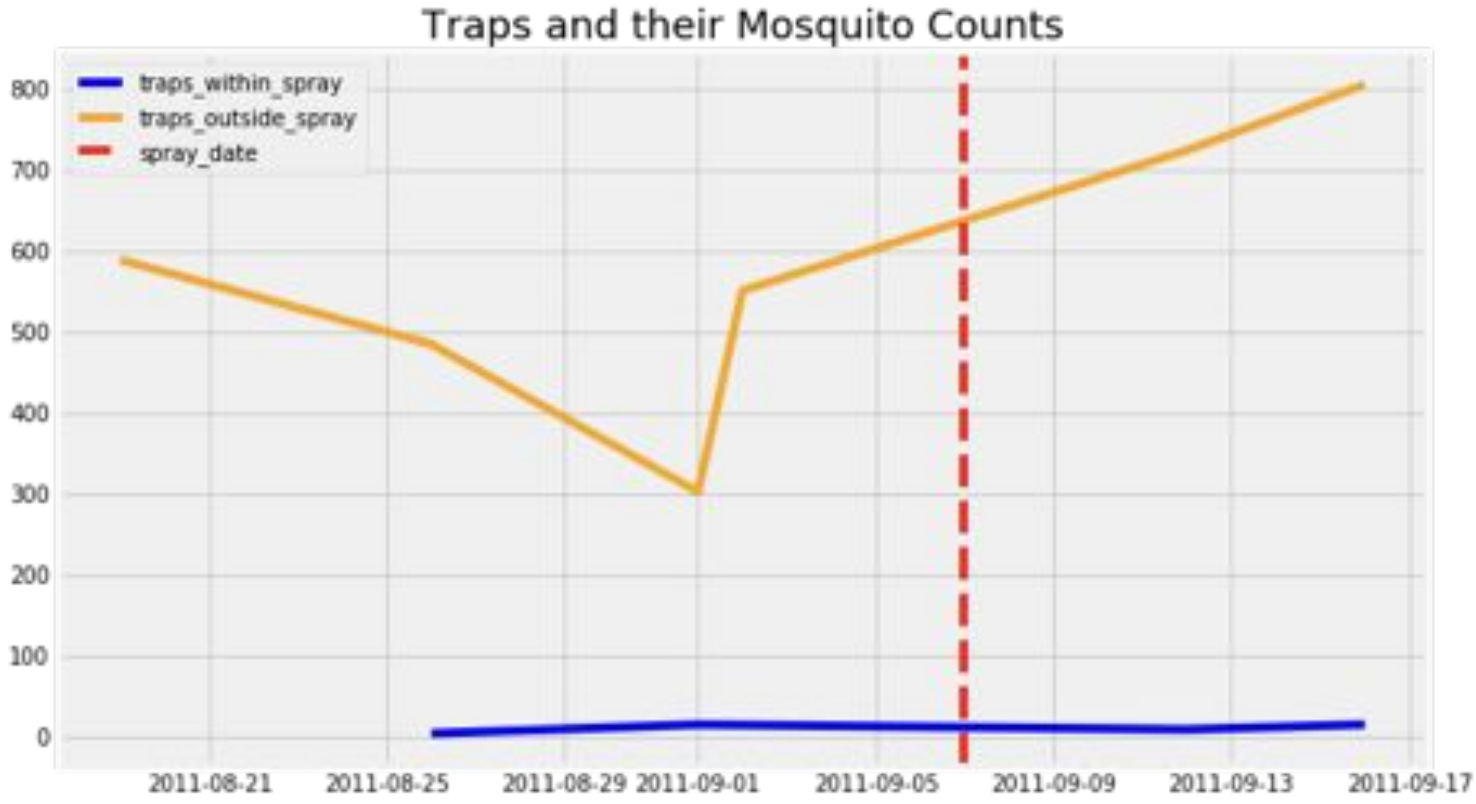
Classifier Model	Accuracy	roc_auc	Recall	Kaggle roc_aucvscore
RandomForest w GridSearchCV	81.5%	81.5%	83.1%	63%
RandomForest w feature engineering	98.3%	98.5%	99.7%	70%
XGBoost	84.1%	91.3%	91.1%	70.7%
XGBoost w GridSearchCV	**94.7%**	**98%**	**99.4%**	**75.9%**
XGBoost w feature engineering	94.6%	98%	99.4%	75%

Confusion Matrix

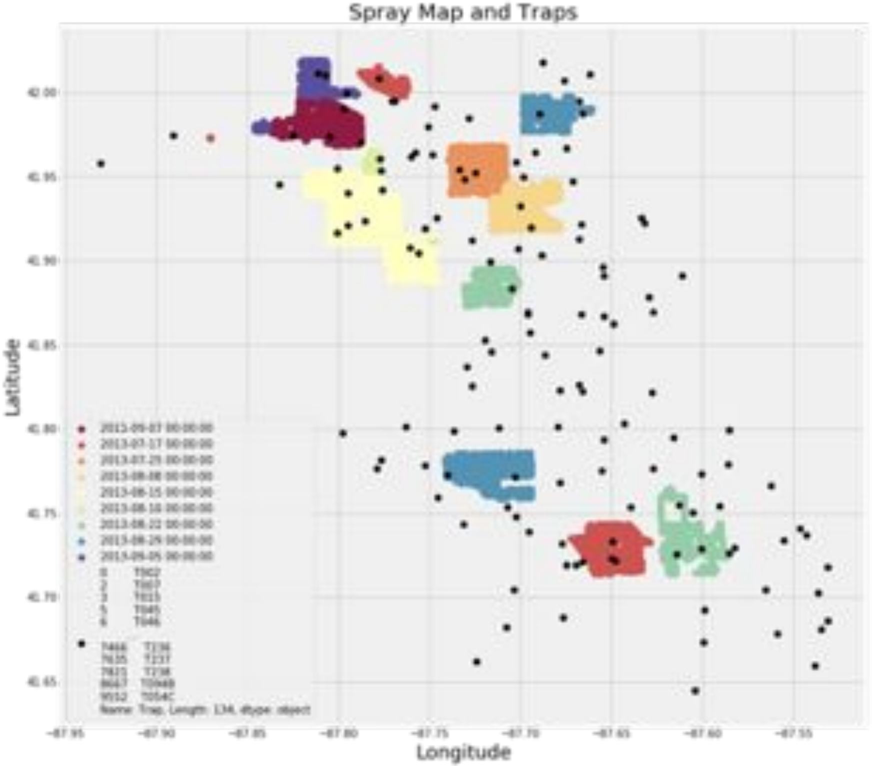
(XGBoost w GridSearchCV)

TP	3316	- Correctly predicted as West Nile Virus carrier.
TN	2963	- Correctly predicted as non West Nile Virus Carrier
FP	331	- Mispredicted someone as a West Nile Virus carrier. - TYPE I error.
FN	21	- Mispredicted someone as a non West Nile Virus carrier. - Type II error.

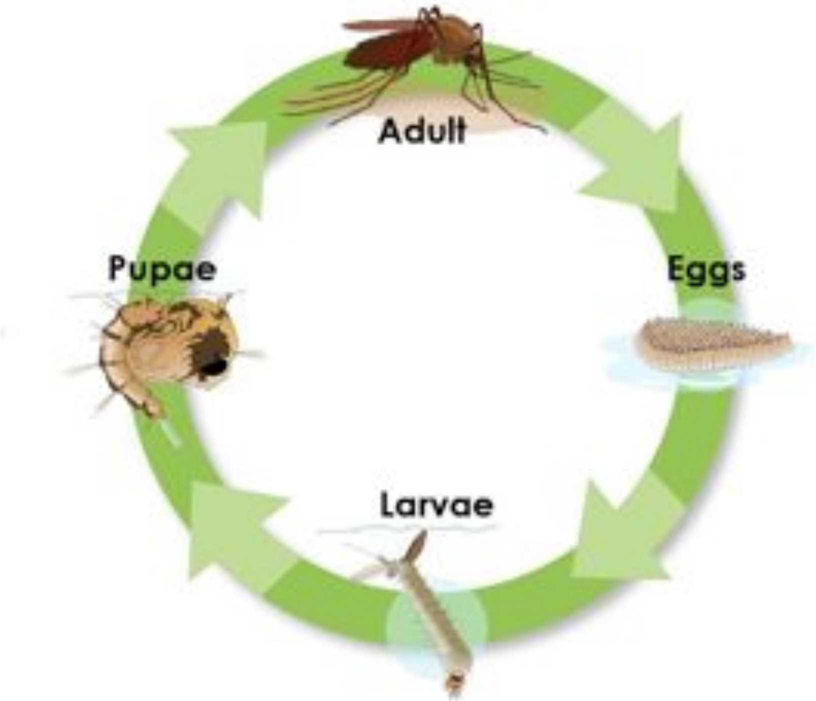
Conclusion – XG Boost Works! Need more Data!



Future – Better Geo-Libraries, ‘Dist from Spray Border’, Time-lag Stationarity Tests!



Other Recommendations – 7 Day Cycle


Adult
Eggs
Larvae
Pupae