West Nile Virus in Mosquitoes across the City of Chicago Prediction

By Yang Gao

Problem

Predict the Probability when and where different species of mosquitoes will test positive for West Nile Virus (WNV) in the City of Chicago.

Approaches

Data Visualization
Data Processing
Machine Learning Models
Issues and Future Work

DATA PROCESS

Raw Data
Clean
ML
F1 Score
Prediction

WNV City of Chicago Data[1]

Visualization

(Seaborn & Matplotlib)

Data Preprocessing

(Check missing value, feature engineering and feature scaling)

Machine Learning Pipeline

(LogisticRegression, Decision Tree, Random Forest,and XGBoost)

Grid Search & Parameter Tuning

Performance Measures

(F1-score,AUC, Precision-Recall)

Probability Prediction

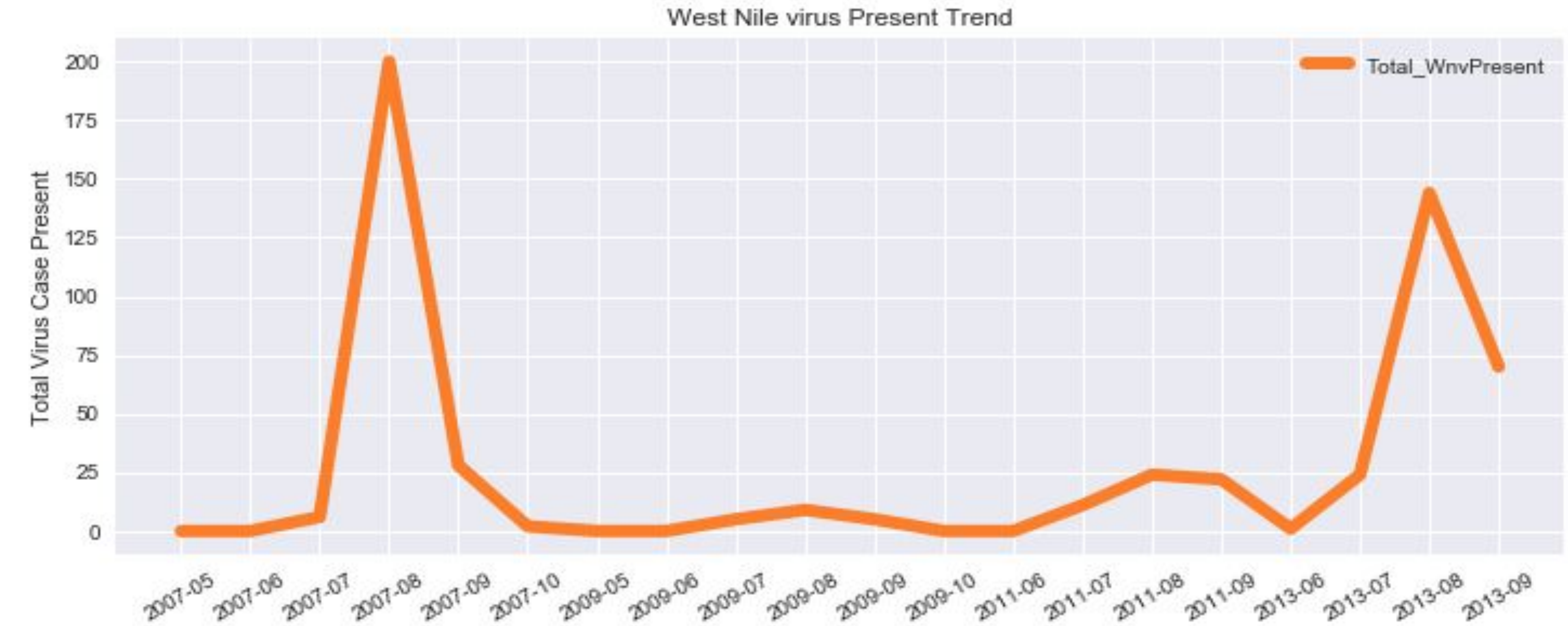
(According to when and where,determine WNV Presents happen)

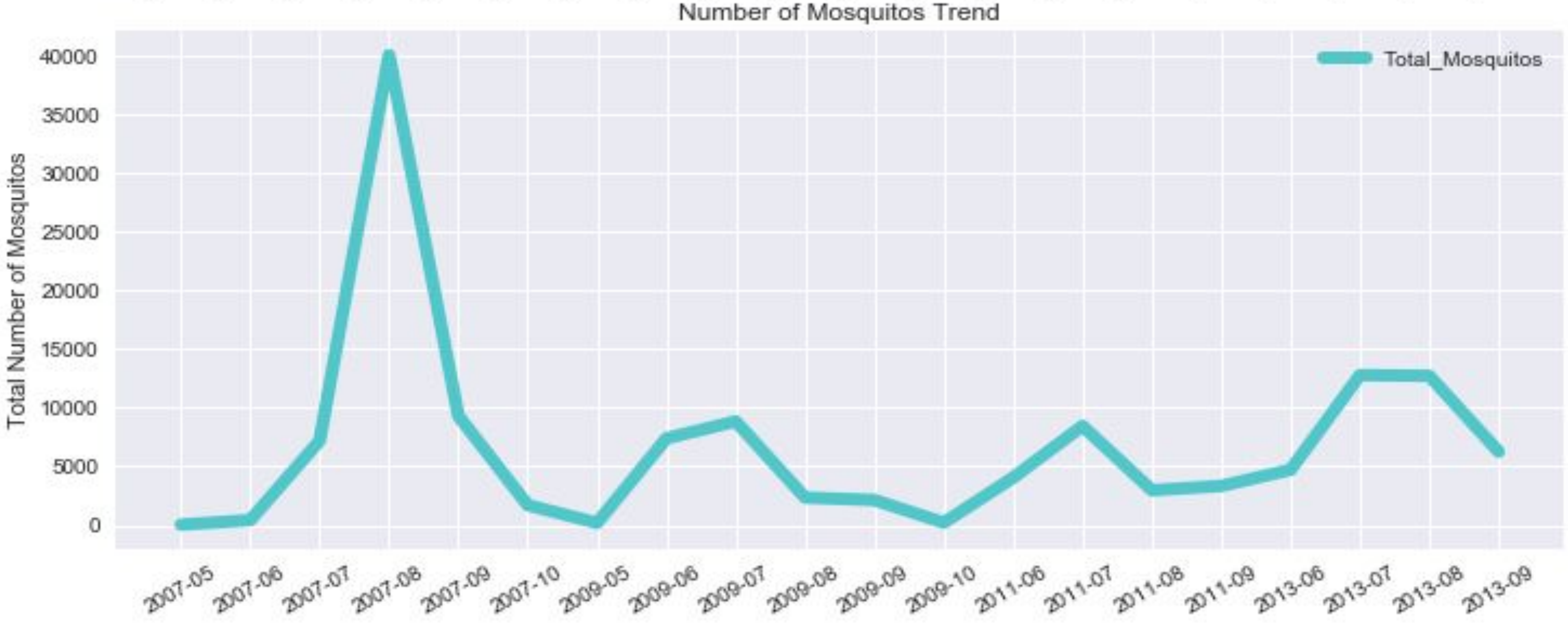
[1] Data Source: https://www.kaggle.com/c/predict-west-nile-virus/data

Data Visualization

Conjecture:

- Number of Mosquitoes correlated with Time (Month/Day)
- WNV Presents correlated with Time (Month/Day)
- Number of Mosquitoes correlated with WNV Presents (0.20)

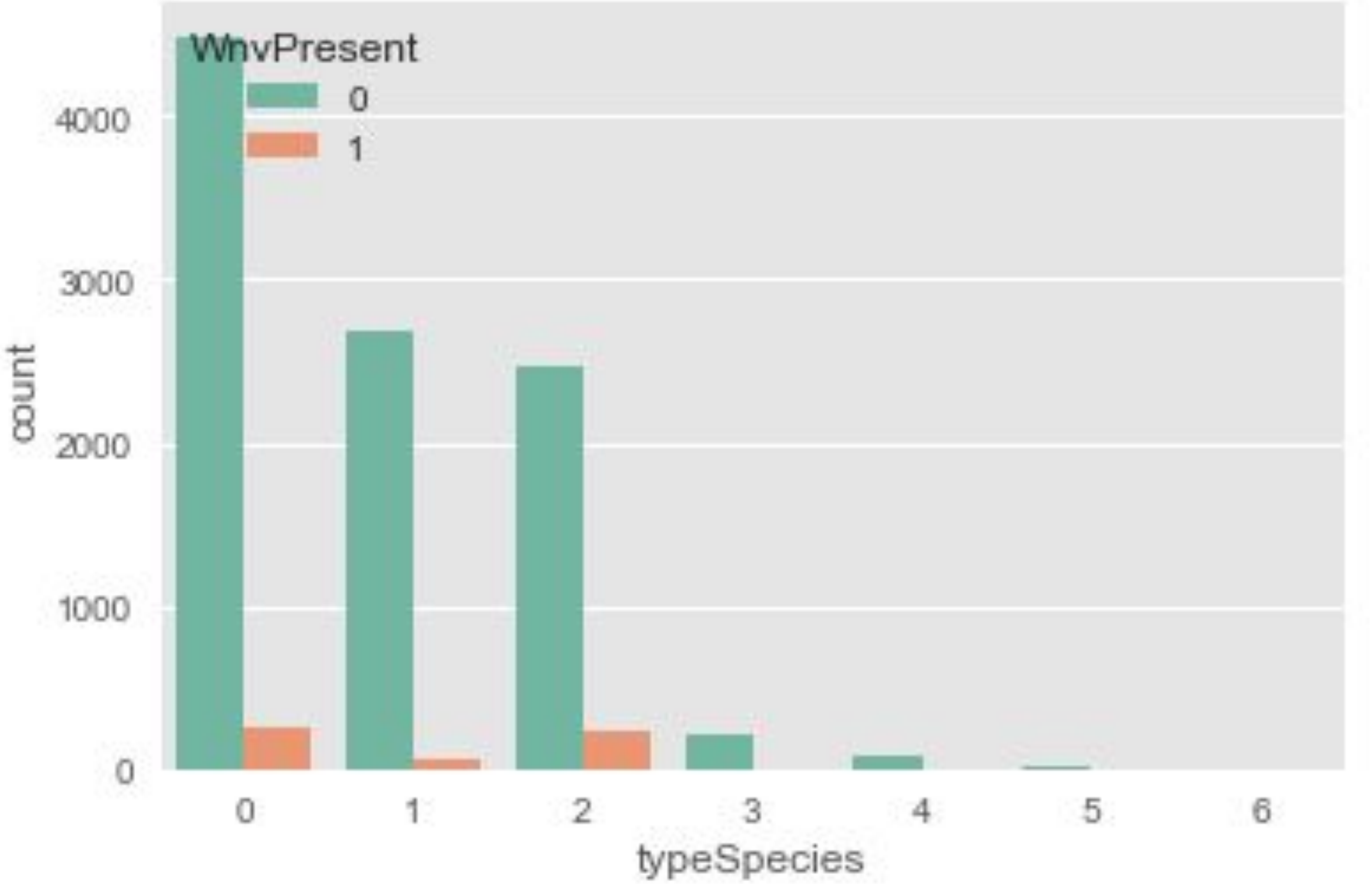


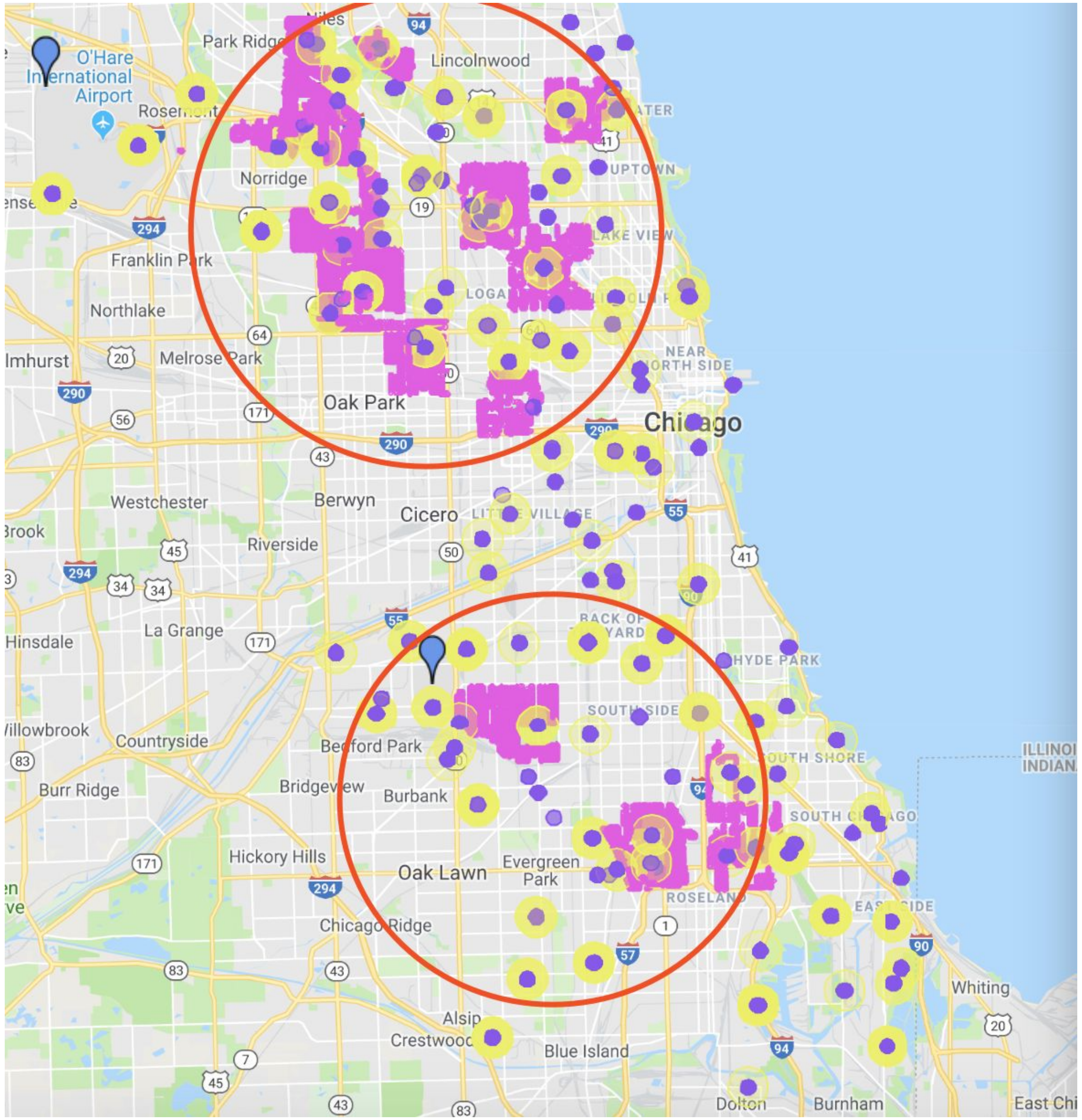


Data Visualization

Conjecture:

- WNV Presents associate with Latitude and Longitude?





Mosquitoes Distribution with WNV Presents VS Spray Location

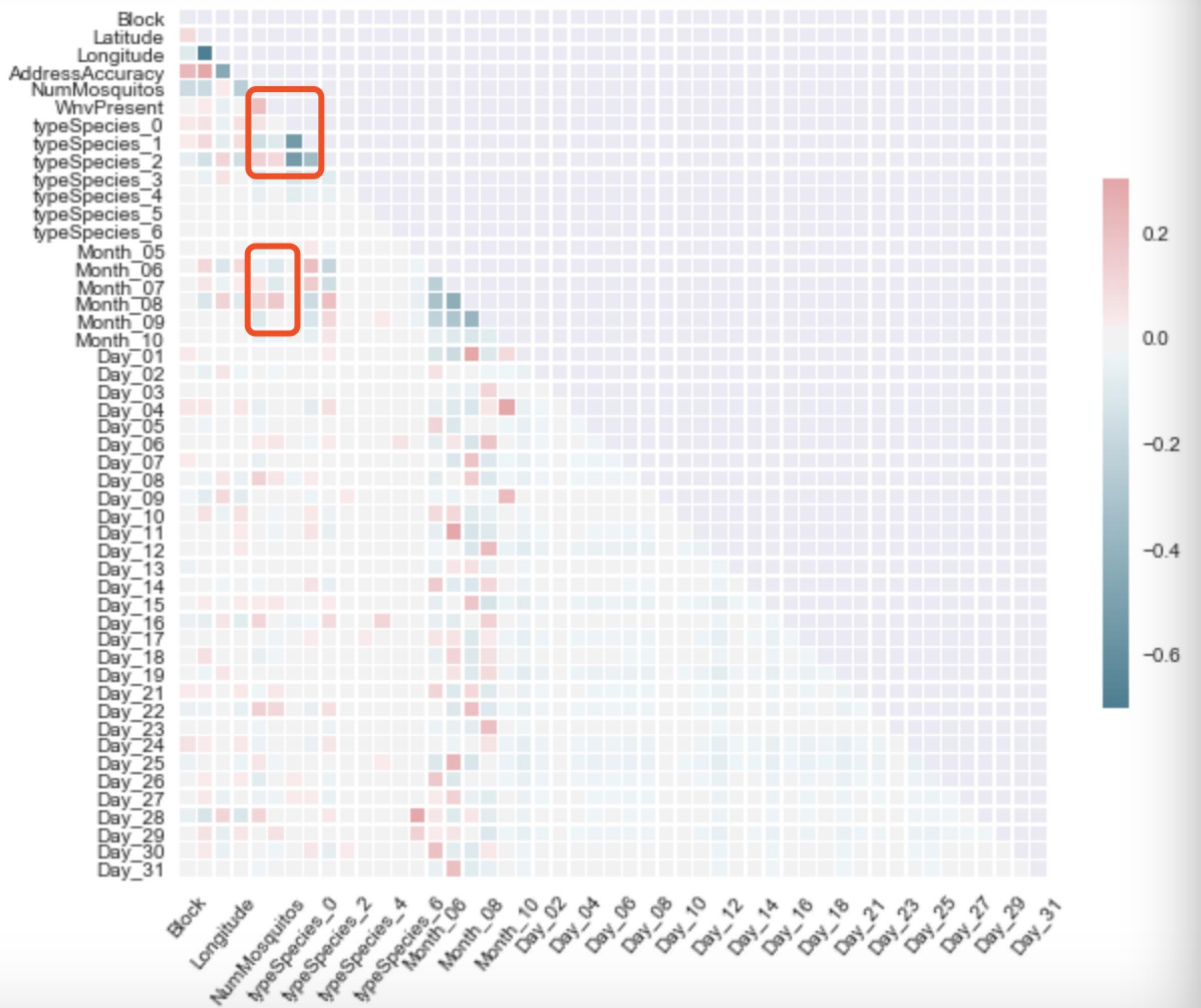
Data Visualization (Train.csv)

- Impact of Numbers of Mosquitoes

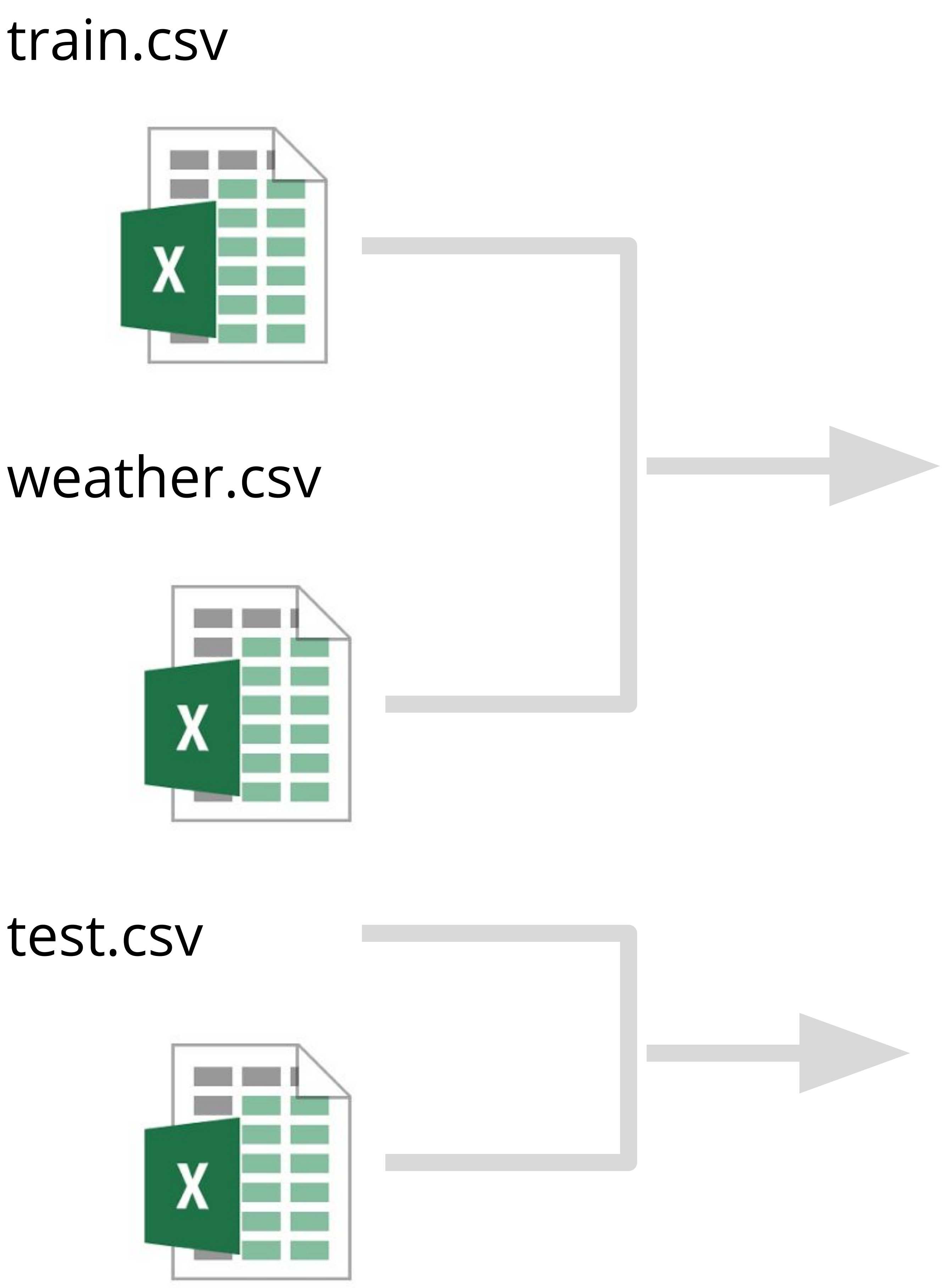
e.g. Month 7 & 8, CULEX PIPIENS /RESTUANS & CULEX PIPIENS (Type 0 &2)

- Impact of WNV Presents

e.g. CULEX PIPIENS(Type 2) , Month 8, Latitude



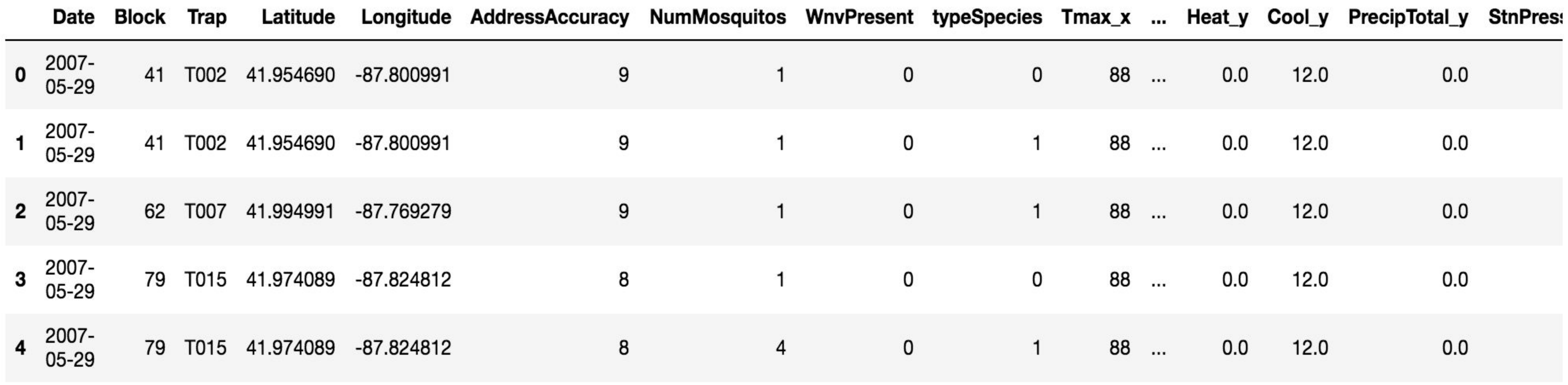
Data



train.csv

weather.csv

test.csv



	Date	Block	Trap	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent	typeSpecies	Tmax_x	...	Heat_y	Cool_y	PrecipTotal_y	StnPress
0	2007-05-29	41	T002	41.954690	-87.800991	9	1	0	0	88	...	0.0	12.0	0.0	
1	2007-05-29	41	T002	41.954690	-87.800991	9	1	0	1	88	...	0.0	12.0	0.0	
2	2007-05-29	62	T007	41.994991	-87.769279	9	1	0	1	88	...	0.0	12.0	0.0	
3	2007-05-29	79	T015	41.974089	-87.824812	8	1	0	0	88	...	0.0	12.0	0.0	
4	2007-05-29	79	T015	41.974089	-87.824812	8	4	0	1	88	...	0.0	12.0	0.0	

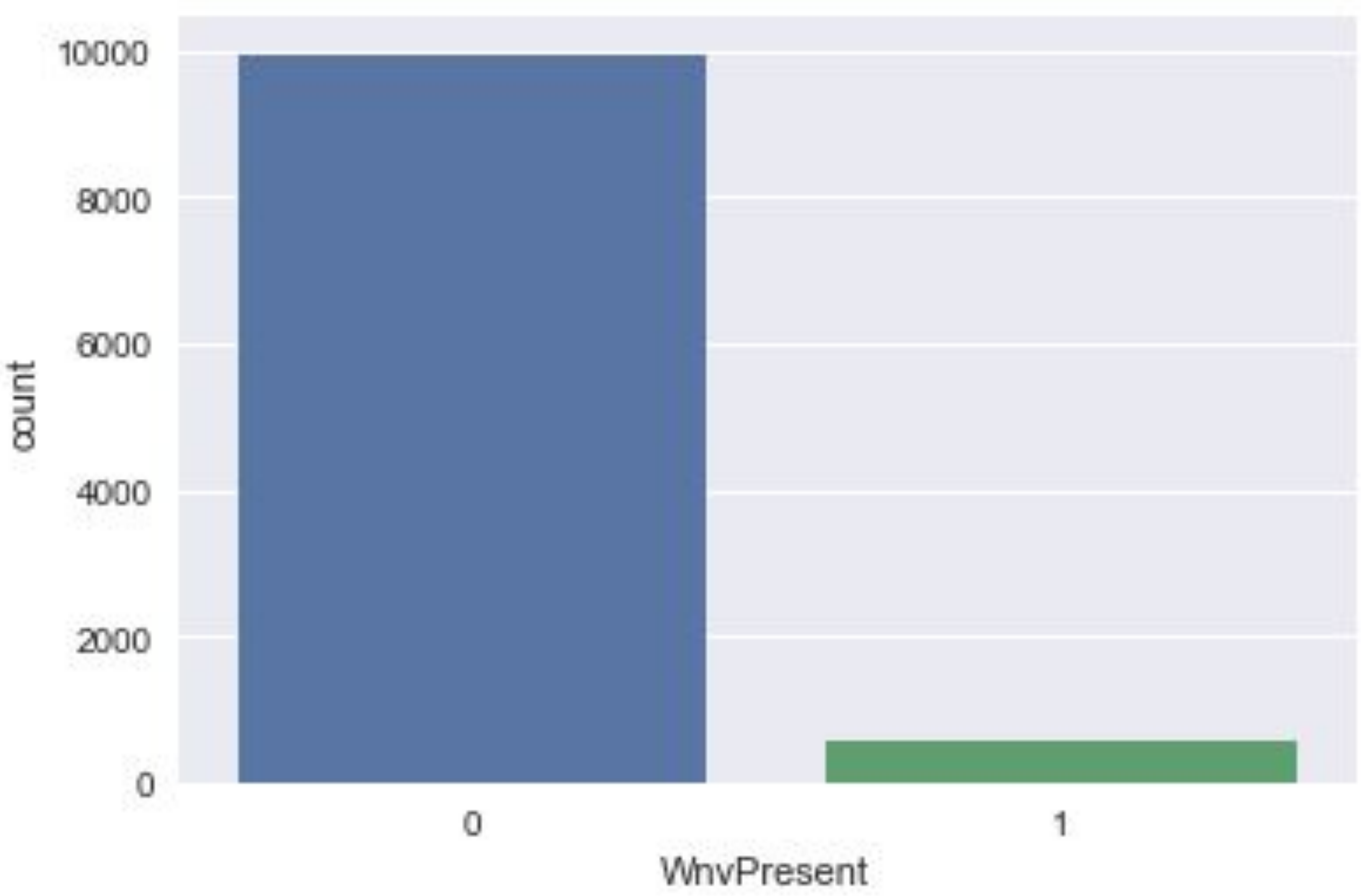
NEW Train Data

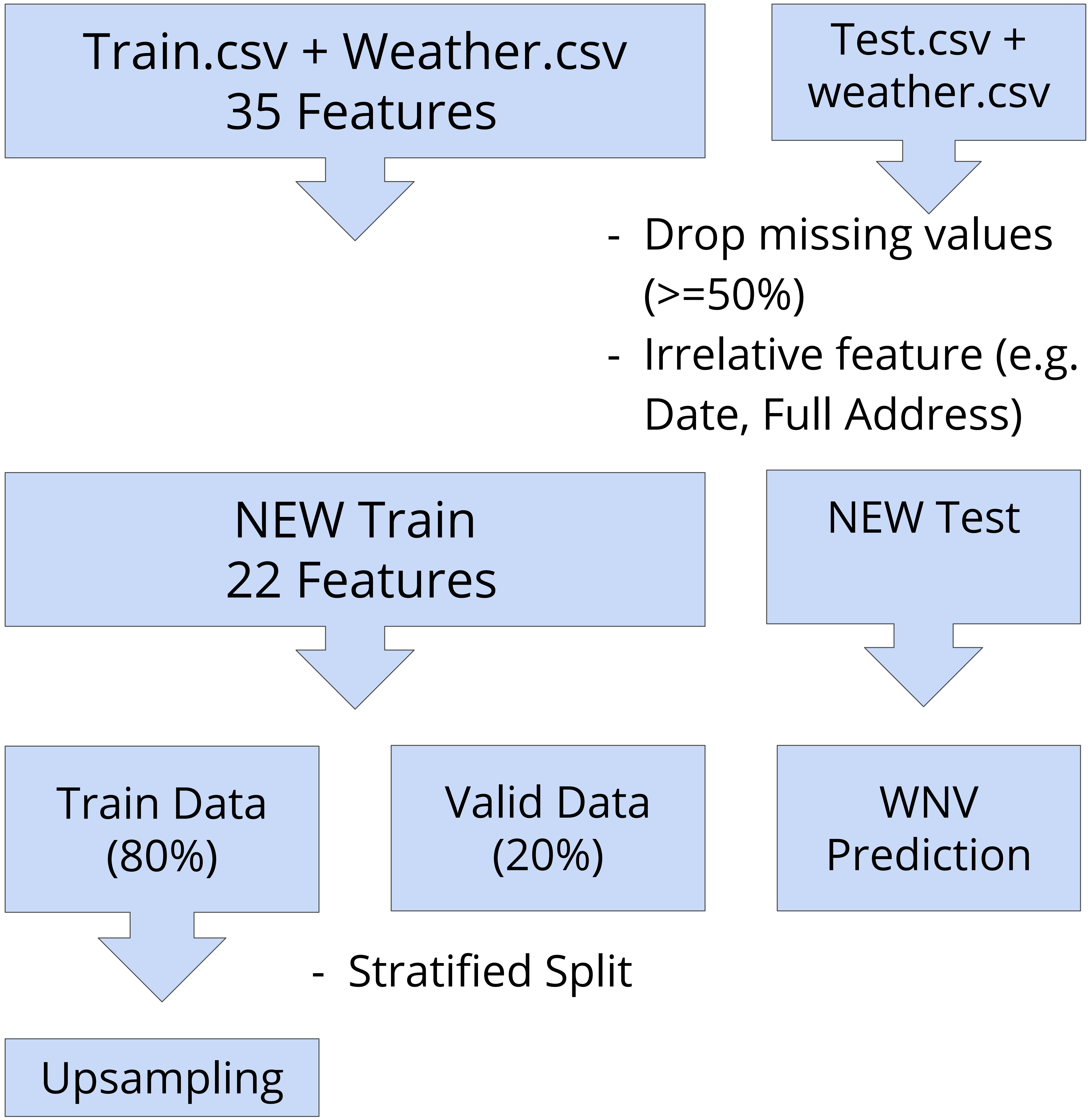
NEW Test Data

WNV Present?

Class Distribution

- Imbalance
- Resample (Upsampling Minority Train Data)





Train.csv + Weather.csv
35 Features

Test.csv + weather.csv

- Drop missing values (>=50%)
- Irrelative feature (e.g. Date, Full Address)

NEW Train
22 Features

NEW Test

Train Data (80%)

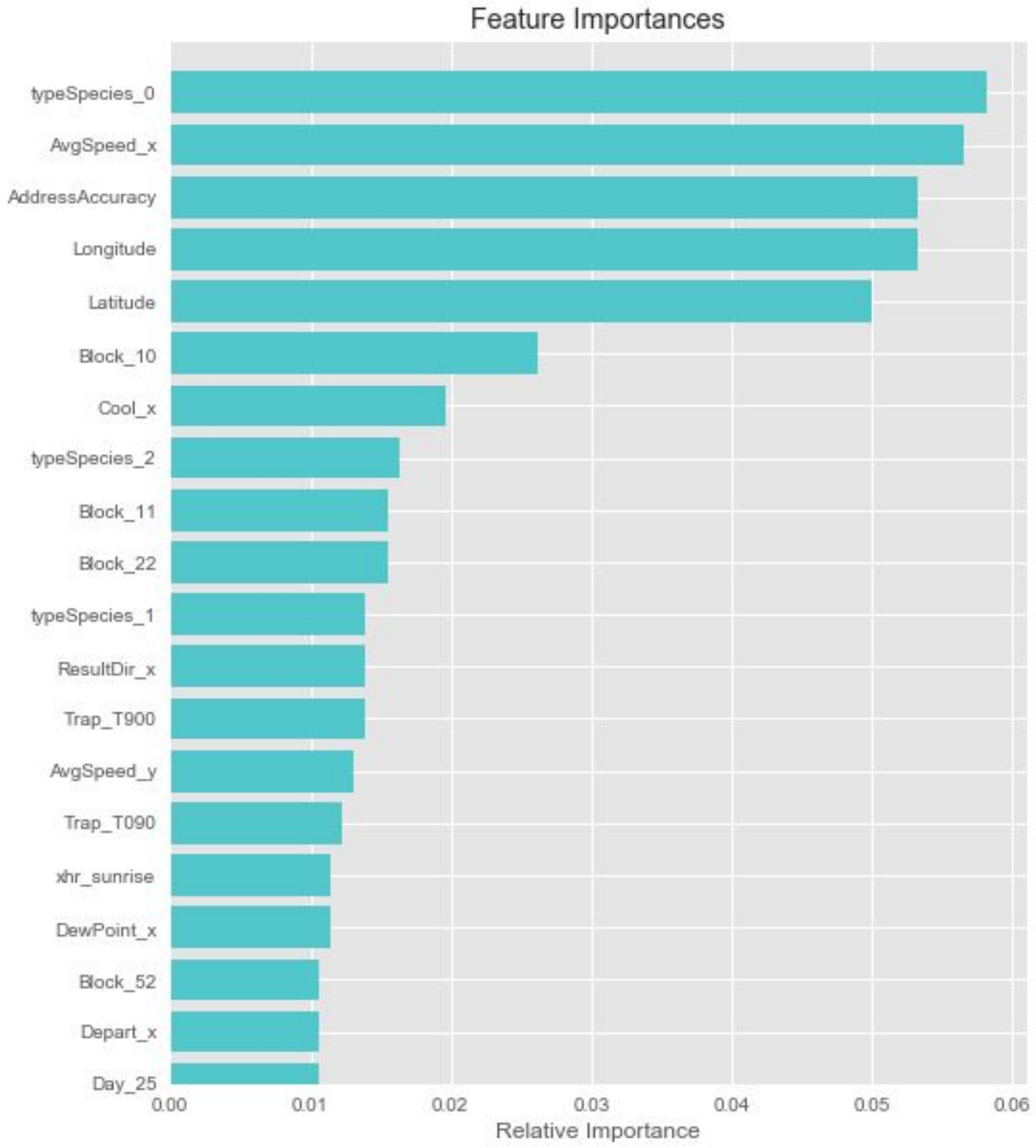
Valid Data (20%)

WNV Prediction

- Stratified Split

Upsampling

Feature Importance



Machine Learning models
LR
Decision Tree
Random Forest
GBM

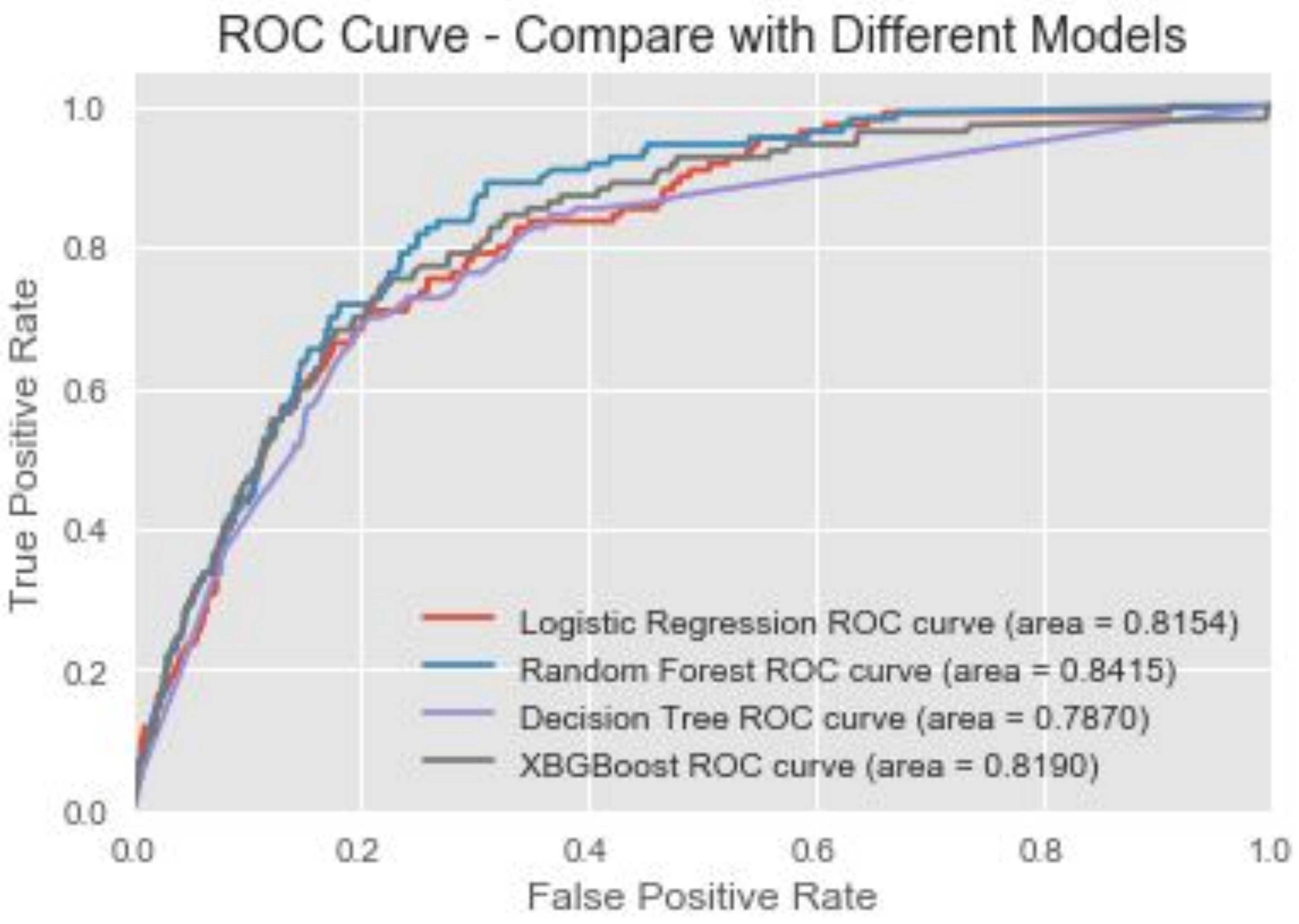
Grid Search

- Parameter Tuning on Each Model
- Feature Importance

Performance Measures

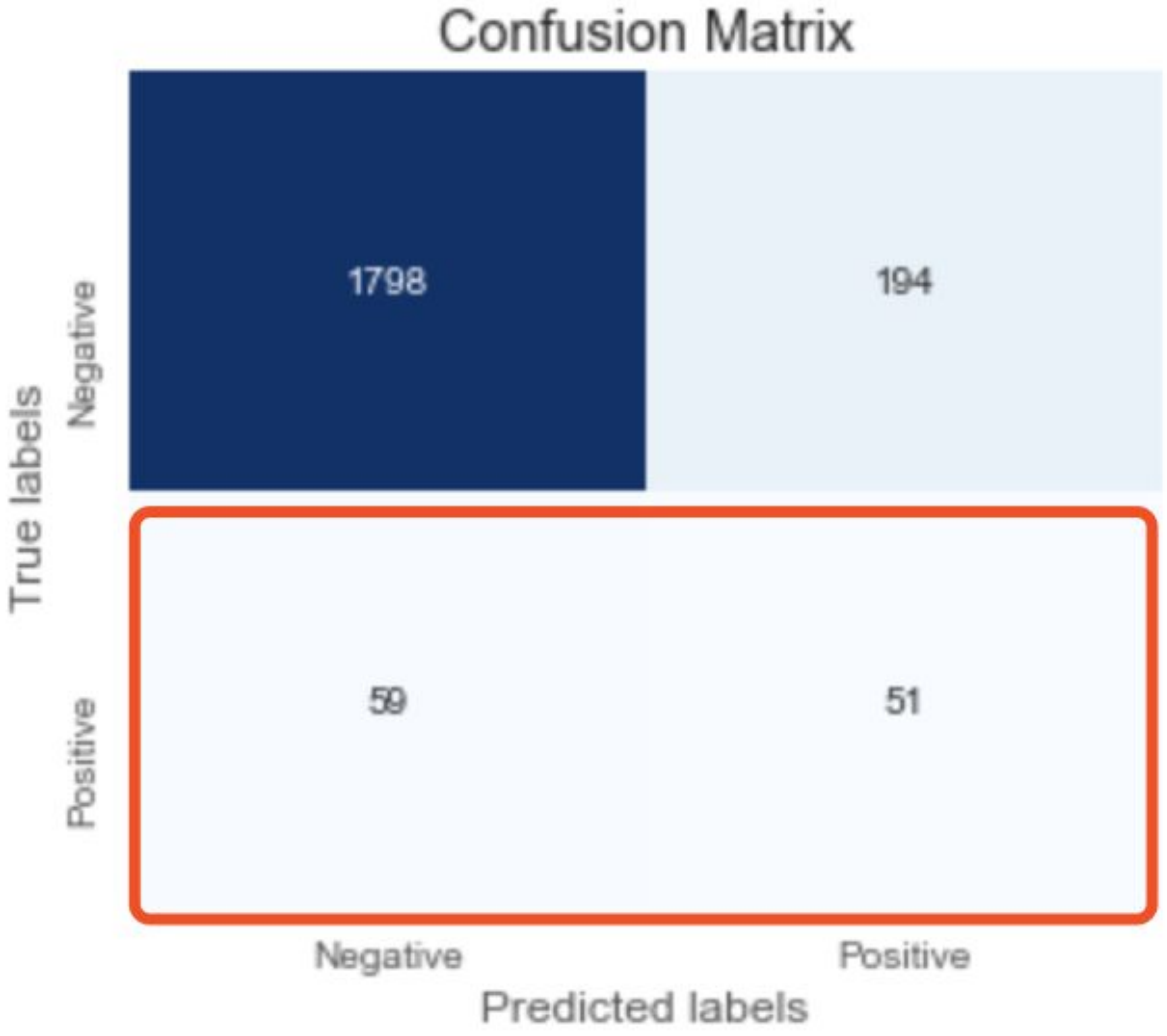
- F1 score
- AUC/Precision-Recall

Outcome & Suggestion



Model	F1 Score	AUC
Logistic Regression	0.5248	0.8154
Decision Tree	0.5812	0.7870
Random Forest	0.6024	0.8415
XGBoost	0.6107	0.8190

Limitations & Future Work



GBM (XGBoost)

Reduce False Negative (FN)

Detect True Positive as much as possible

- Try LightGBM (Speed Up Computing)
- Try Complexity Neural Networks
- Add More Data (Sanitation Level, Spray ,Population and Economic,Twitter #hashtag)

Thanks!

Any questions?