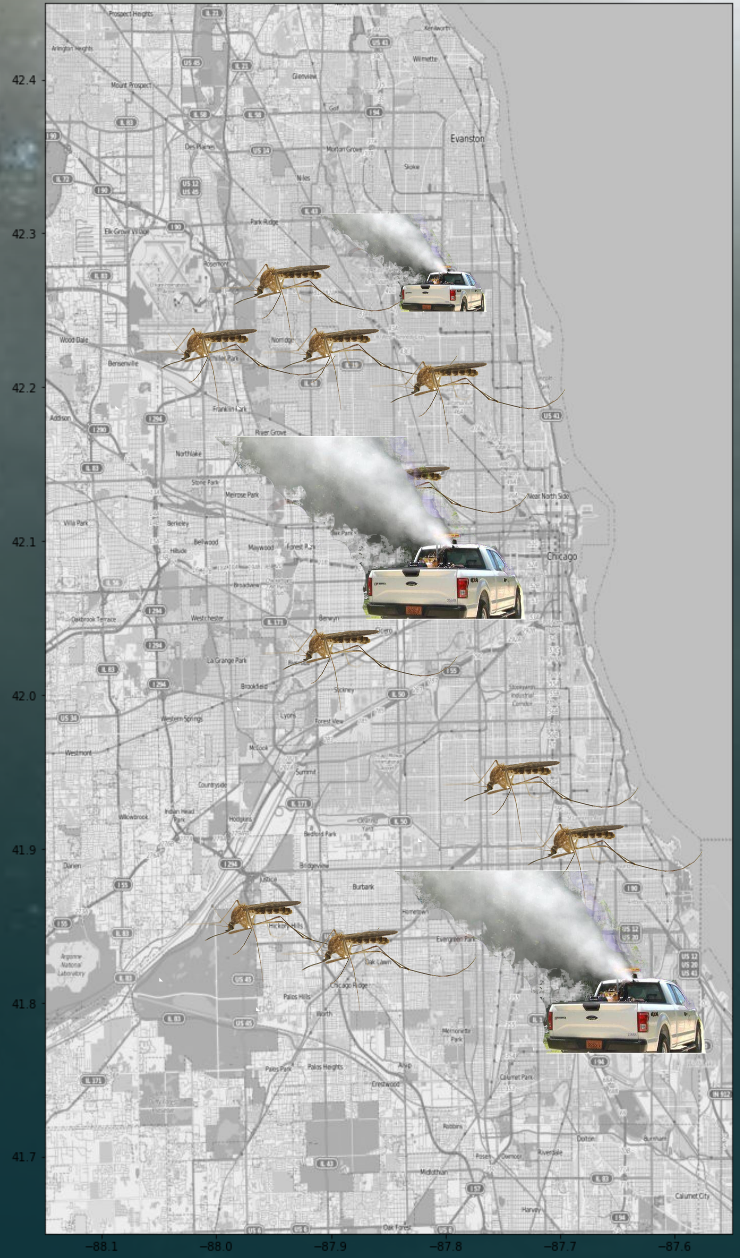


To Spray or Not to Spray?

Predicting West Nile Virus in the City of Chicago

By:
BENEDICT,
MELVIN,
SUFYAN

Problem Statement

We as a team of analysts in the Disease And Treatment Agency in Chicago, have been tasked to :

1. Predict where and when different species of mosquitos will test positive for WNV in the City of Chicago.
2. Perform a cost-benefit analysis on the pesticide coverage (cost) and its effects (benefit).
3. Recommend any improvements to the current mosquito control measures

Background

The West Nile virus (WNV) is the leading cause of mosquito-borne disease in the continental United States since 1999. People infected with the WNV largely do not feel sick, with only 1 in 5 experiencing mild symptoms and 1 in 150 experiencing serious illness.

There is currently no human vaccine available. As such, the only way to reduce infection is to reduce exposure to mosquitos.

Goal

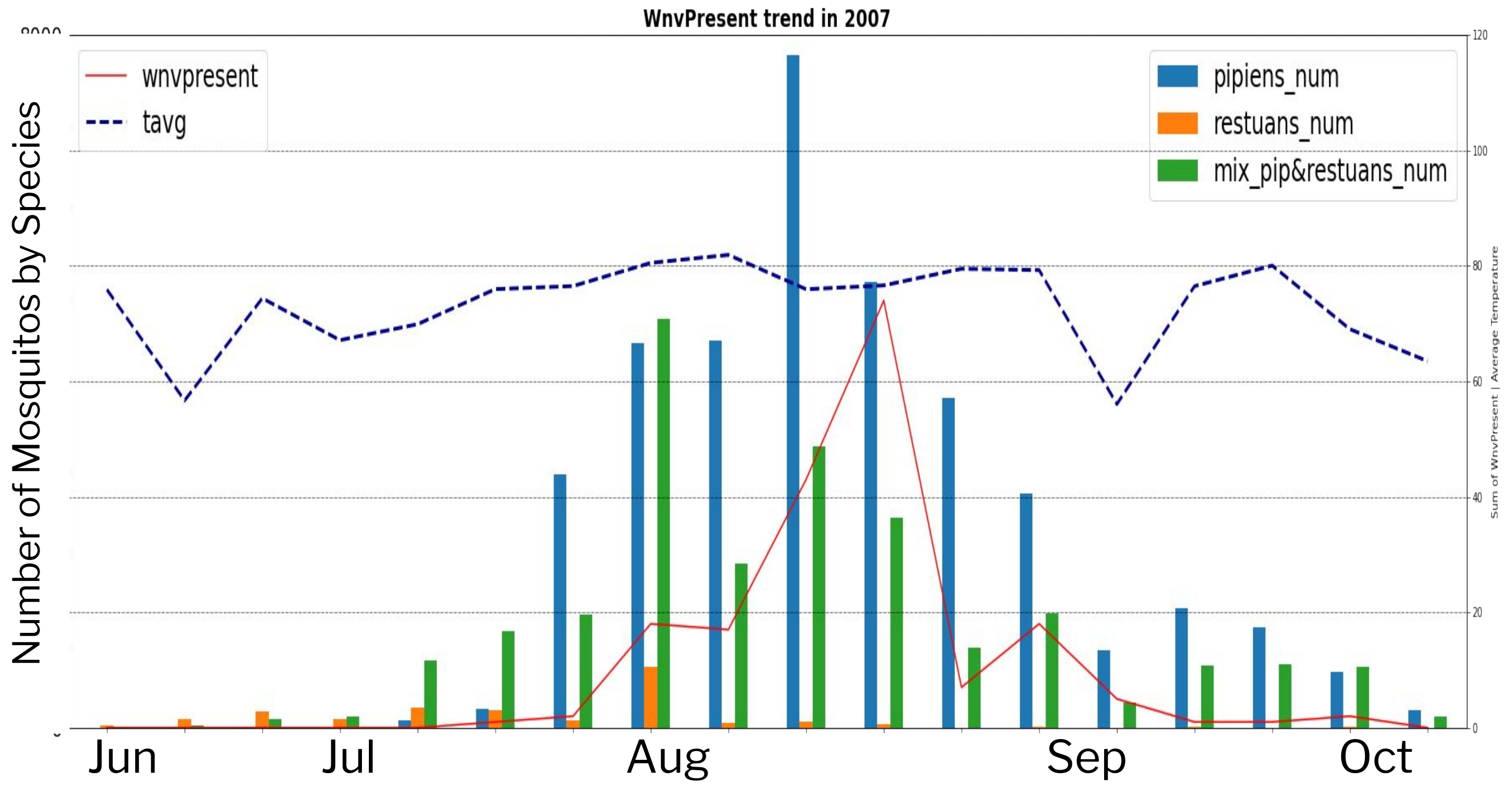
In favour of **reducing human fatalities** and the possibility of an **uncontained WNV outbreak**, the team will focus on minimising false negatives over false positives.

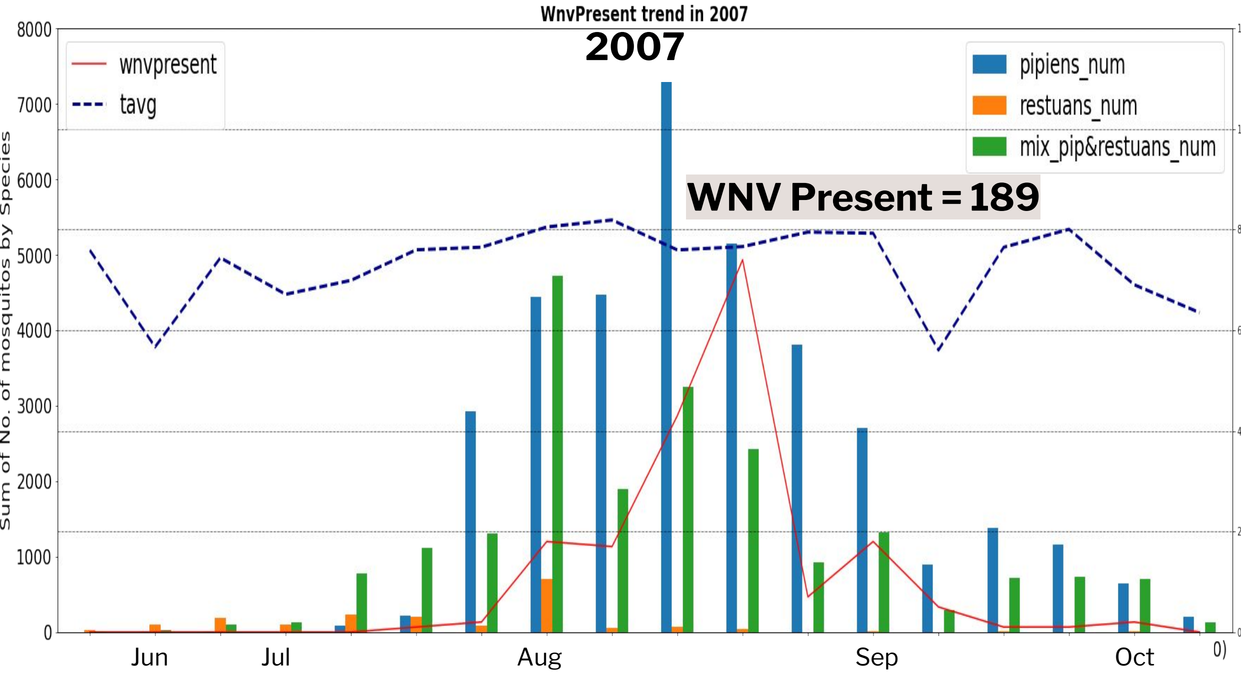
This would mean placing greater emphasis on **Sensitivity** rather than Specificity without sacrificing too much Accuracy.

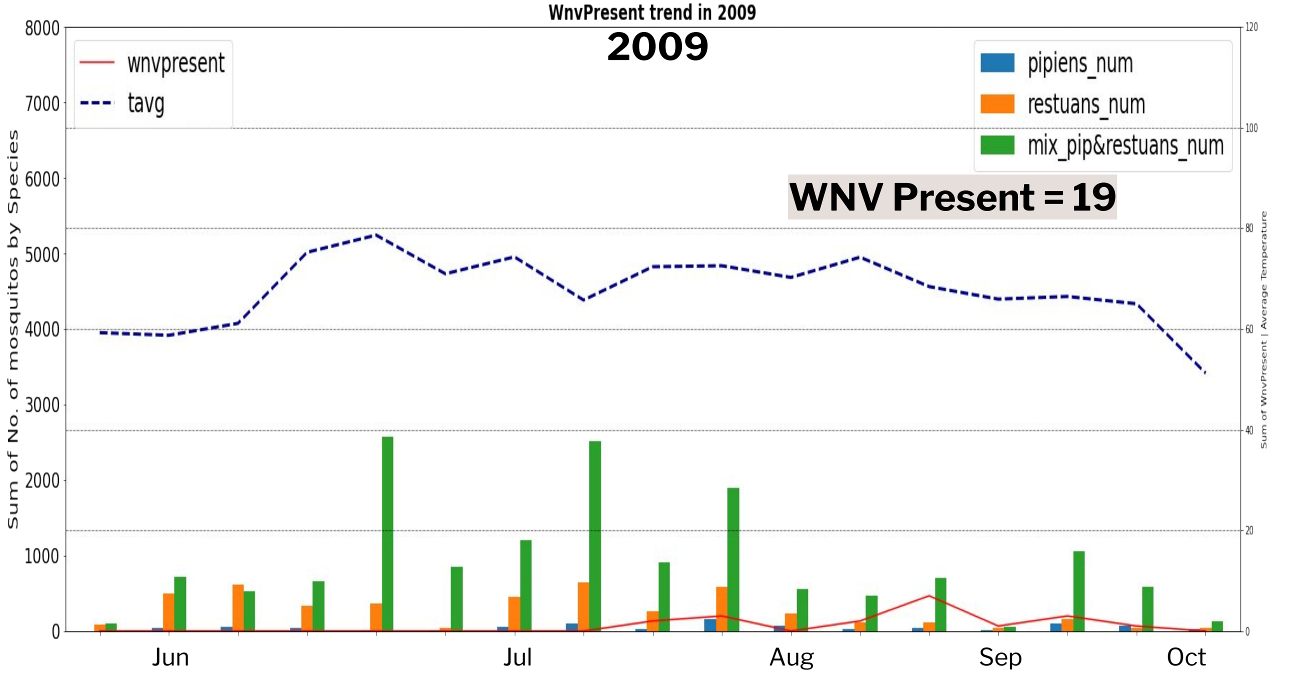
Contents

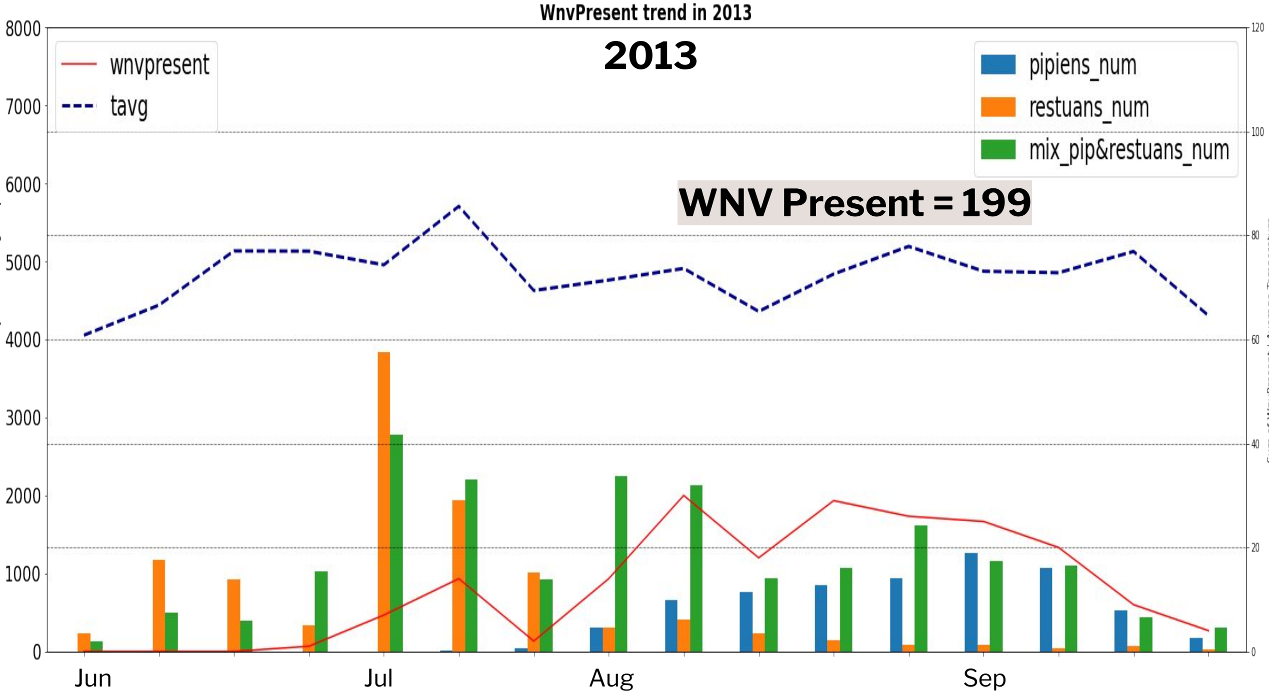
Exploratory Data Analysis

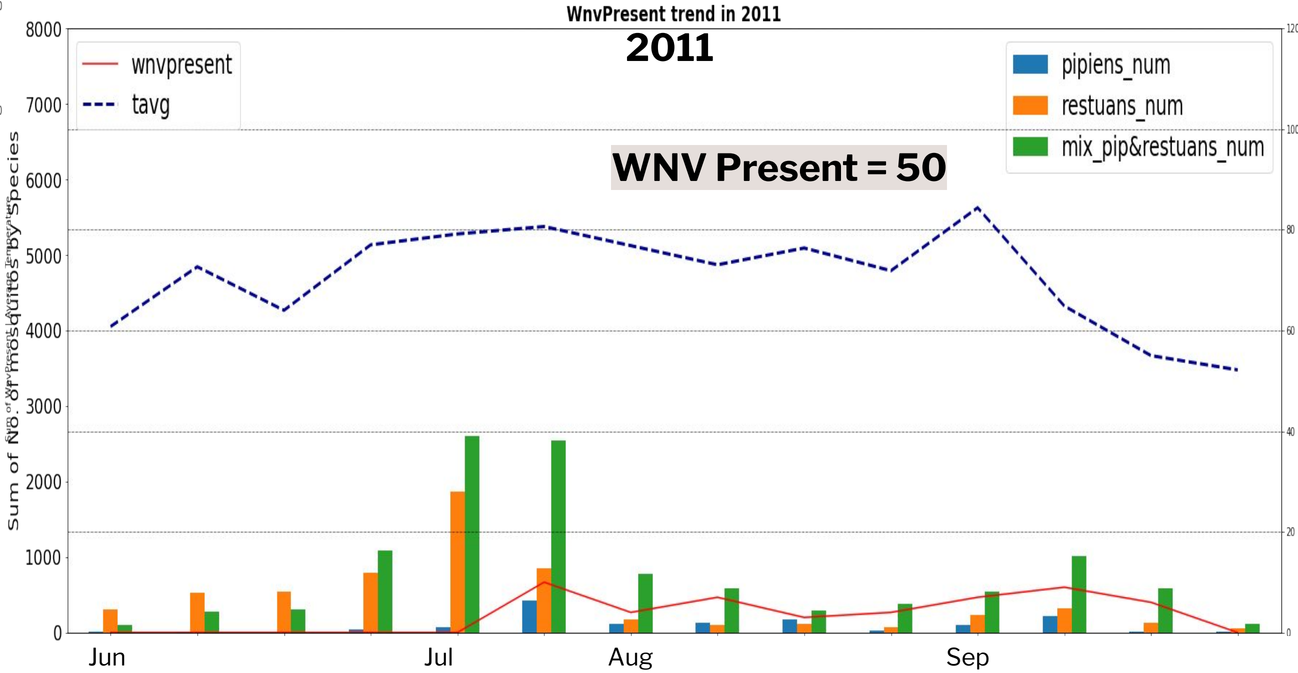
Distribution of Mosquitos across Time




WnvPresent trend in 2007
2007
wnvpresent
tavg
pipiens_num
restuans_num
mix_pip&restuans_num
WNV Present = 189
Sum of No. of mosquitos by Species
Sum of WnvPresent | Average Temperature
Jun
Jul
Aug
Sep
Oct


WnvPresent trend in 2009
2009
wnvpresent
tavg
pipiens_num
restuans_num
mix_pip&restuans_num
WNV Present = 19
Sum of No. of mosquitos by Species
Sum of WnvPresent | Average Temperature
Jun
Jul
Aug
Sep
Oct

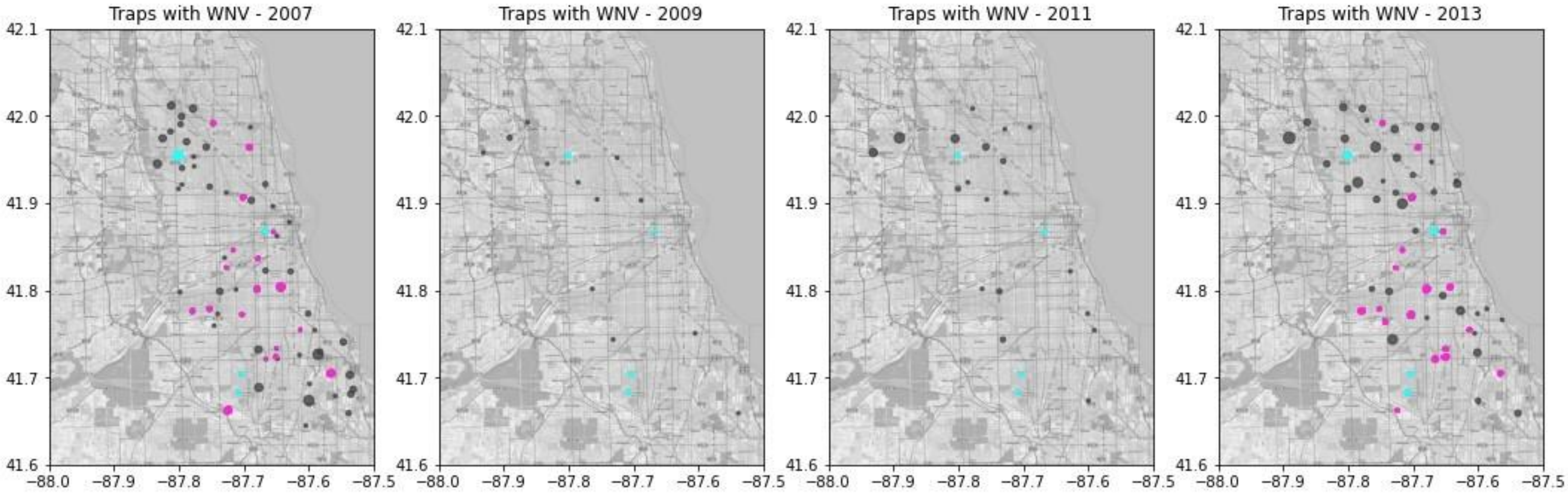

WnvPresent trend in 2013
2013
wnvpresent
tavg
pipiens_num
restuans_num
mix_pip&restuans_num
WNV Present = 199
Jun
Jul
Aug
Sep


WnvPresent trend in 2011
2011
wnvpresent
tavg
pipiens_num
restuans_num
mix_pip&restuans_num
WNV Present = 50
Sum of No. of mosquitos by Species
Jun
Jul
Aug
Sep

EDA

1. Annual trends are not consistent. There might be other factors affecting the presence of mosquitos and WNV.
2. Years with high WNV:
   a. have a corresponding rise in Pipiens counts
   b. from August to September

EDA: Hotspots across years



Appear every year

Only in 2007 and 2013

General Occurrences

1. Hotspots are generally not consistent across every year.
2. There might be hotspots across a subset of years

Key Finding

There are potentially high order interactions between all the features.

Example:

Years with higher temperatures result in a higher population of Culex Pipiens in week 33, resulting in WNV presence at trap_090.

Modelling Implications:

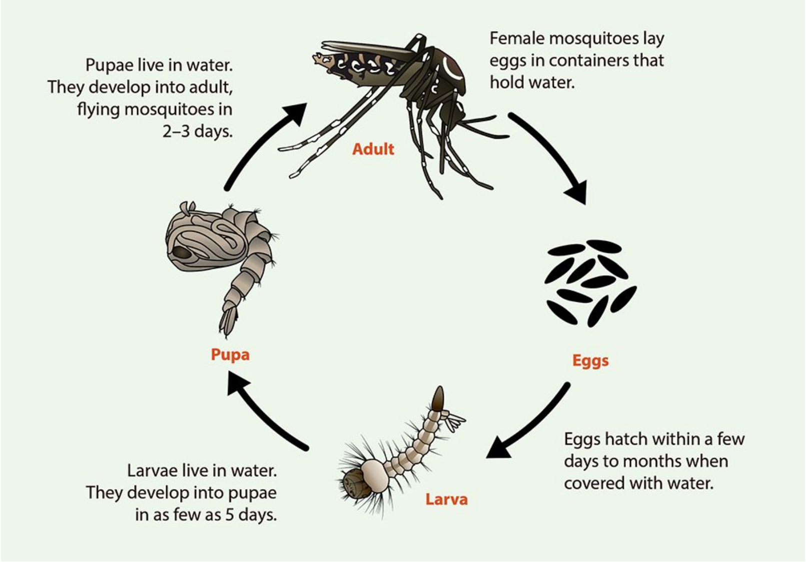
1. To include such variables as dummies so that they can have custom interactions.
2. Models that naturally account for the interactions will likely perform better than those which do not (ie. RandomForest and XGBoost may be better than LogisticRegression).
3. To be conservative in removing features with low feature importance in RF as feature importance does not reflect effect of interactions.

Feature Engineering

1. Weather lag

- Given favourable conditions mosquitos take time to breed, and for eggs to become mature.
- Weather lag will cater for the interval between favourable conditions and the presence of adult mosquitos

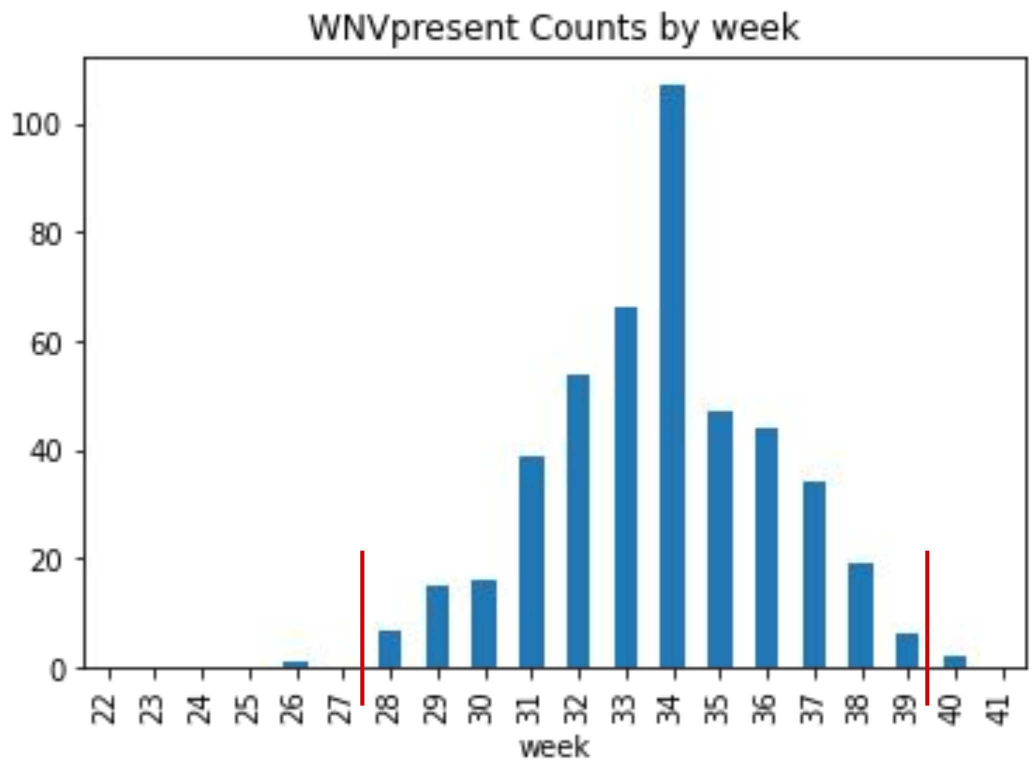
We engineered weather lags varying from 5 to 21 days to account for this.



Feature Engineering

2. **Weeks**
   - Dummy coded weeks 28 to 39.

3. **Species**
   - Dummy coded Pipiens, Restuans and Pipiens/Restuans



Feature Engineering

4. Location (Traps)

Consistent hotspots across years

trap	years_wnvp
T002	4
T090	4
T095	4
T158	4
T028	3
T003	3
T114	3

High WNV traps across years

trap	wnvpresent
T900	29.0
T002	15.0
T115	15.0
T003	14.0
T225	11.0
T011	11.0
T013	10.0
T028	9.0

High WNV traps WITHIN years

trap	wnvpresent
T115	12.0
T138	9.0
T002	7.0
T011	7.0
T086	7.0
T135	7.0
T082	6.0
T016	5.0

trap	wnvpresent
T900	15.0
T013	8.0
T225	8.0
T003	7.0
T030	7.0
T235	7.0
T002	6.0
T027	6.0
T028	6.0

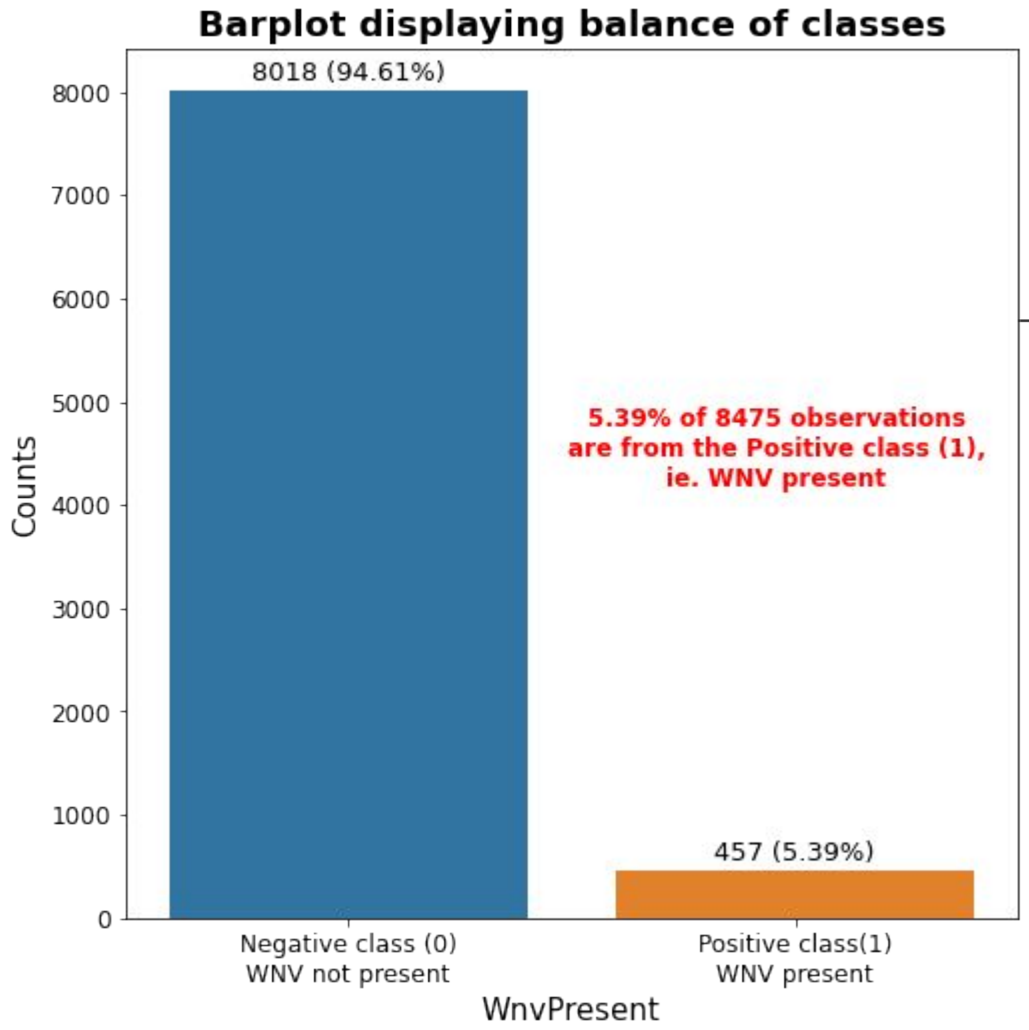
Preliminary Feature Selection

- Weather lag Choices : 7 days | 21 days
- Feature groups : g1 | g2

	weather_lag	train_acc	test_acc	sensitivity	specificity	roc_auc_score
0	0	79.63	88.82	71.05	81.40	84.84
1	5	81.21	88.77	71.93	83.44	85.20
2	7	80.36	89.81	72.81	81.90	85.52
3	14	82.06	88.49	71.05	83.44	85.45
4	21	80.90	88.35	73.68	82.04	84.65

	feature_grp	weather_lag	train_acc	test_acc	sensitivity	specificity	roc_auc_score
0	g1	7	77.09	83.25	74.56	75.06	82.75
1	g2	7	77.00	83.25	74.56	74.76	82.73
2	g3	7	76.68	82.96	74.56	74.66	82.64
3	g1	21	76.81	81.93	76.32	75.11	82.77
4	g2	21	76.73	82.07	77.19	74.56	82.78
5	g3	21	76.57	81.69	77.19	74.26	82.59

Weather **lag_7** and **lag_21** were chosen for further tuning



Classification modelling

Dataset is skewed towards the negative class, where 5.39% of the total observations are from the Positive Class

Classification model

General takeaway and tradeoffs

1) To bump up Sensitivity, we have to make trade off for other matrices

CLASSIFIER	BALANCING TECHNIQUE	TRAIN ACC	TEST ACC	SENSITIVITY	SPECIFICITY	PRECISION	ROC_AUC
Logistic Regression	SMOTE	0.865	0.811	0.737	0.746	0.141	0.811
RandomForest	SMOTE	0.623	0.623	0.711	0.618	0.096	0.720
RandomForest	Class_weight: 'Balanced Subsample'	0.731	0.725	0.763	0.723	0.136	0.812
SVC	SMOTE	0.621	0.591	0.772	0.581	0.095	0.746
XGBoost	Scale_pos_weight: 19	0.691	0.679	0.851	0.669	0.127	0.821

15%

Classification model

General takeaway and tradeoffs

1) With smote, we might see bigger trade offs, and harder to tune for a good balance.
2) Using RandomForest and XGBoost weightage method, it is easier to tune sensitivity while maintaining relatively high scores for other matrices

CLASSIFIER	BALANCING TECHNIQUE	TRAIN ACC	TEST ACC	SENSITIVITY	SPECIFICITY	PRECISION	ROC_AUC	
Logistic Regression	SMOTE	0.865	0.811	0.737	0.746	0.141	0.811	
RandomForest	SMOTE	0.623	0.623	0.711	0.618	0.096	0.720	SMOTE
RandomForest	Class_weight: 'Balanced Subsample'	0.731	0.725	0.763	0.723	0.136	0.812	Weighted
SVC	SMOTE	0.621	0.591	0.772	0.581	0.095	0.746	SMOTE
XGBoost	Scale_pos_weight:19	0.691	0.679	0.851	0.669	0.127	0.821	Weighted

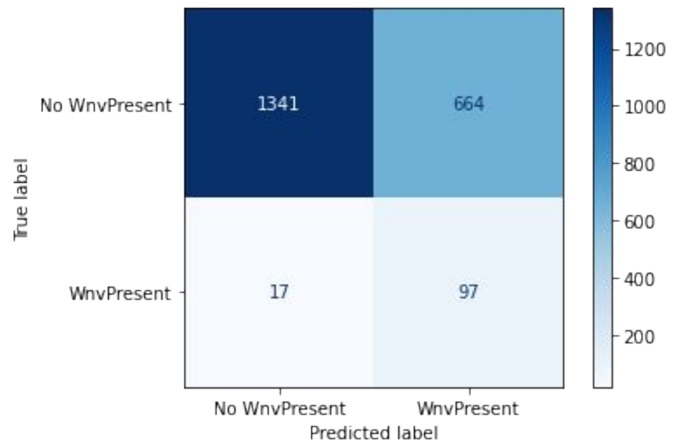
Classification model

CLASSIFIER	BALANCING TECHNIQUE	TRAIN ACC	TEST ACC	SENSITIVITY	SPECIFICITY	PRECISION	ROC_AUC
XGBoost	Scale_pos_weight: 19	0.691	0.679	0.851	0.669	0.127	0.821

- Time consuming
  - First pass GridSearchCV - for good AUC score
  - Fine tune for good Sensitivity score

- Scale_pos_weight = 19 (Proportion of Negative Class / Positive Class)

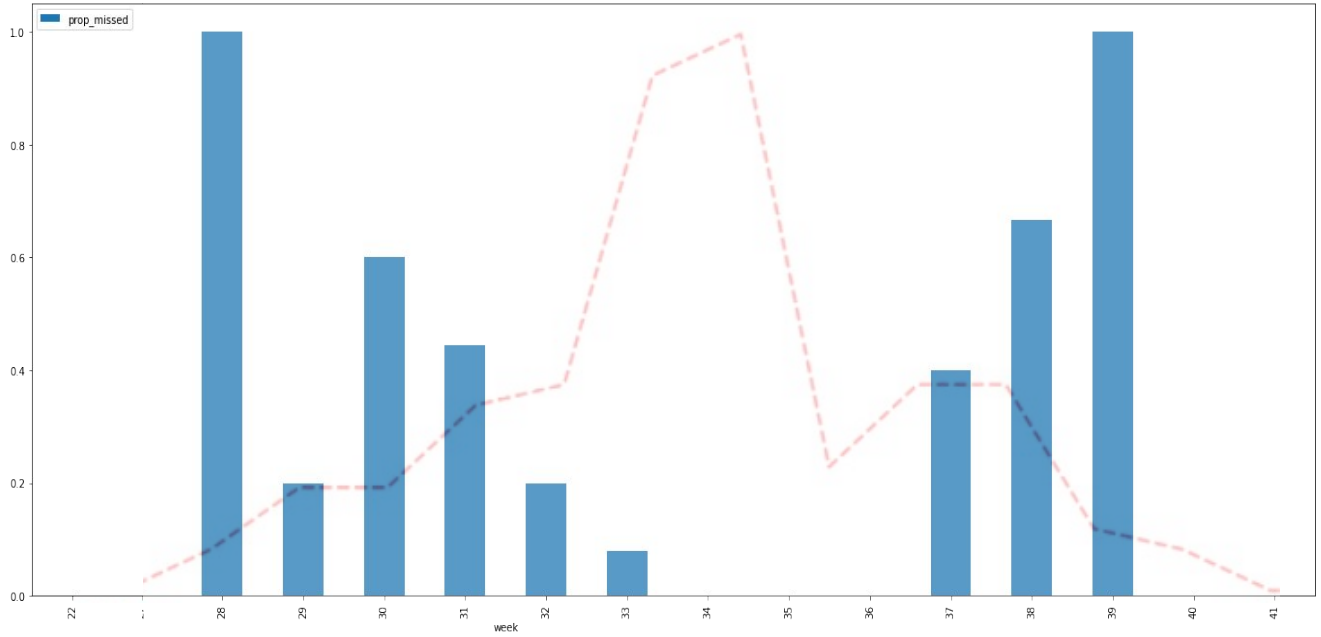
- Hyperparameters tuned:
  - learning Rate
  - subsample
  - reg_alpha
  - max_depth
  - min_child_weight
  - gamma



Error Analysis

Focusing on Sensitivity, we will be looking at the False negatives.

- Plotted the proportion of False negatives (FN/Total WNV) against each week
- Found that there is higher proportion of FN at the tails of the WNV outbreak across the weeks



	wnv present	False negatives	Proportion FN
Culex pipiens/restuans	56	9	0.16
Culex pipiens	45	3	0.066
Culex restuans	13	5	0.38

There are higher proportions of FN in mixed species and restuans, as naturally Pipiens generally carry higher number of WNV.

Recommendations

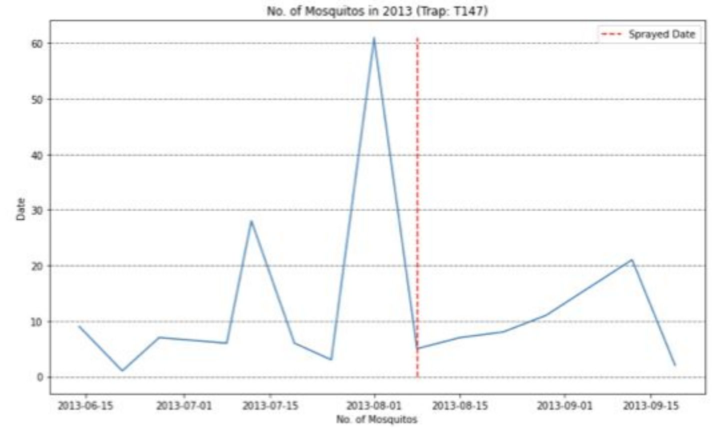
1) Assuming the mixed species has both restuans and pipiens, the model is expected to **perform better if distinguished well.** As such, we recommend taking extra time to ensure that the species are properly identified.

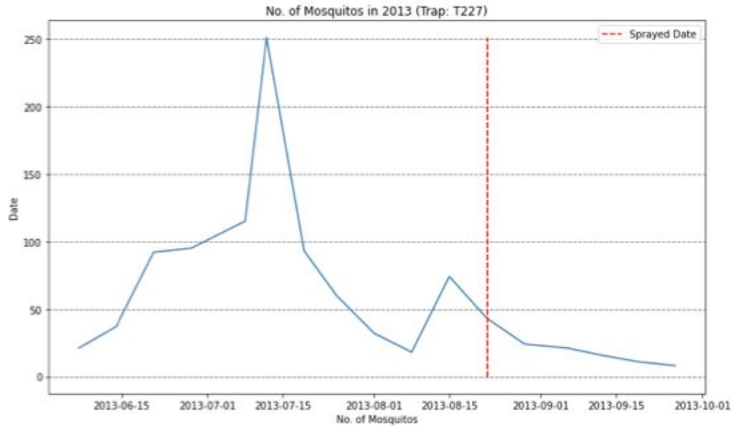
2) While the inclusion of trap locations are sufficient leaving out the rare wnv cases, we suggest for future models to also model **location clusters as hotspots** to better capture high risk areas.

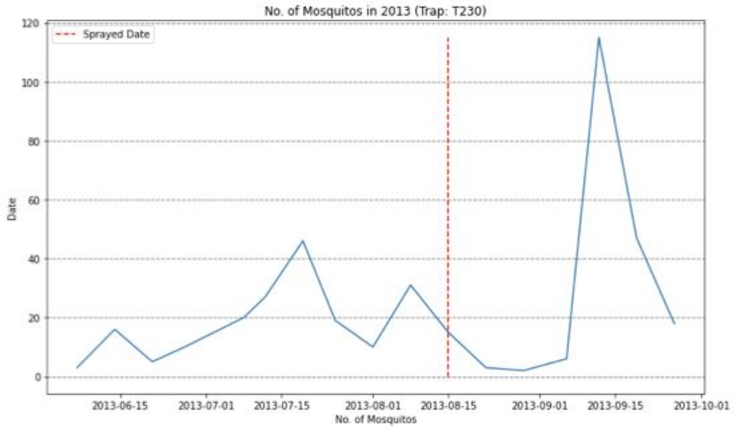
3) The model currently is effective at picking out WNV presence at the **peak** of the wave rather than the **start** and **end of the wave**. We recommend to build future models to be more sensitive to the **onset of the WNV wave** as that would be the best time to tackle the issue

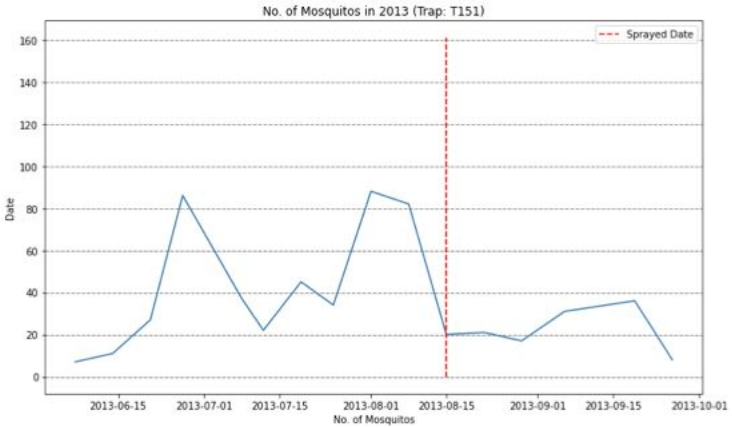
Cost Benefit Analysis

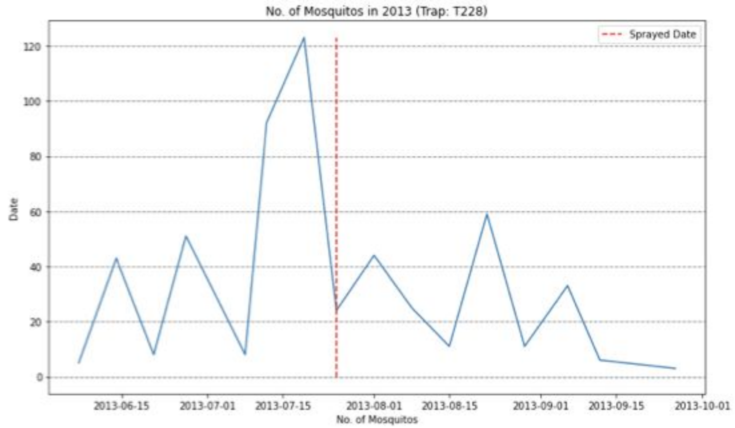
Effect of Spray in Previous Years

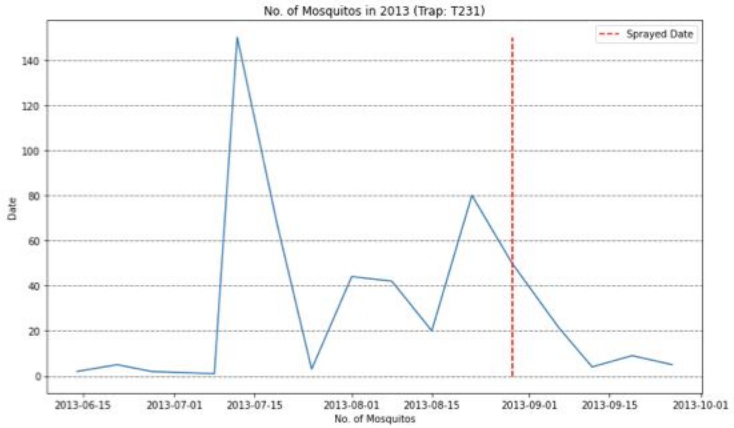












Cost Benefit Analysis

2 different Severities:

1) West Nile Fever (WNF)
   a) Less severe
   b) Outpatient cost = ~$167
   c) Higher percentage of all cases = 71.8%

2) West Nile Neuroinvasive Disease (WNND)
   a) More severe
   b) Hospitalisation cost = ~$46,000
   c) Lower percentage of all cases = 28.2%

[Source]

Average price per spray

= $1907

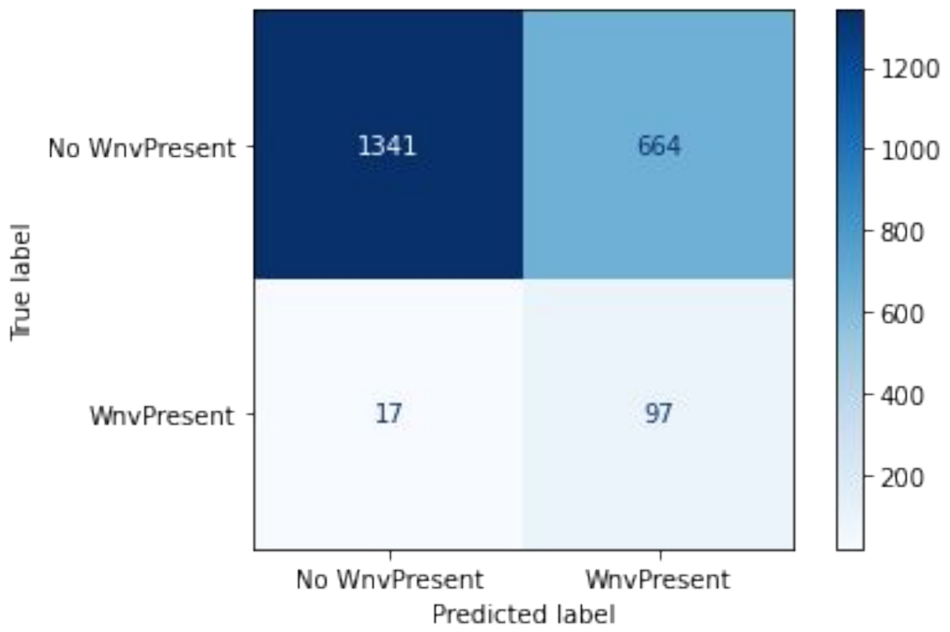
Cost of renting fogging truck

= $11,095

Overall cost per spray session

= $13,002

Cost Benefit Analysis - Predicted



Options	1: Spray All WNV	2: Spray with Average Rate	3: No Spraying
Spraying Cost	$9,894,560.05	$403,063.55	$0
Medical Cost	$25,120.13	$162,705.95	$168,452.66
Total expected Cost	$9,919,680.18	$565,768.5	$168,452.66

$$\text{Average Spray Rate} = \frac{\text{Number of Spray}}{\text{Sum of WnvPresent}}$$
$$= 4\%$$

Cost Benefit Analysis - Recommendation

- Option 2: To spray at average spray rate
  - Other intangible benefits and costs

Future improvements:

1) Reduce the triggering criteria for spraying a hotspot from 2 consecutive weeks to only 1 week for earlier months
2) Pre-emptive spraying in late July(weeks 29 / 30)

Wake me up when September ends