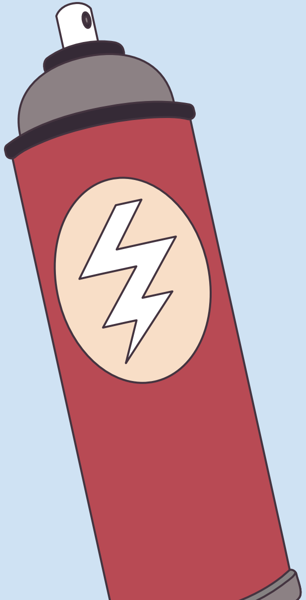
West Nile Virus Prevention

Andre Tan
Nah Wei Jie

Agenda



01 **Introduction**

02 **Problem Statement**

03 **Data Science Process**

04 **Cost Benefit Analysis**

05 **Conclusion & Recommendation**

Introduction (West Nile Virus)

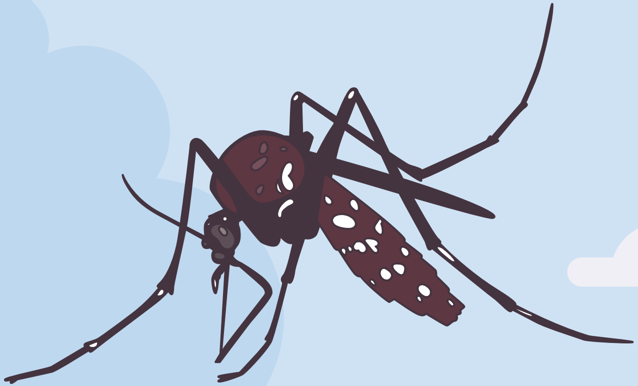
Why?

- 4 of 5 have no symptoms
- 1 of 5 chance of non-fatal symptoms
- No known vaccine / treatment
- 1 in 150 chance of severe neuroinvasive symptoms / fatality

When?

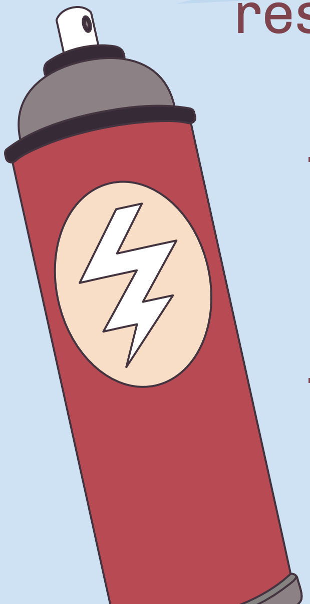
- Summer through autumn

How?



Problem Statement

Generate a model to predict where and when different traps will test WNV+ to predict outbreaks.
This will help the City of Chicago to better allocate resources towards preventing outbreaks.

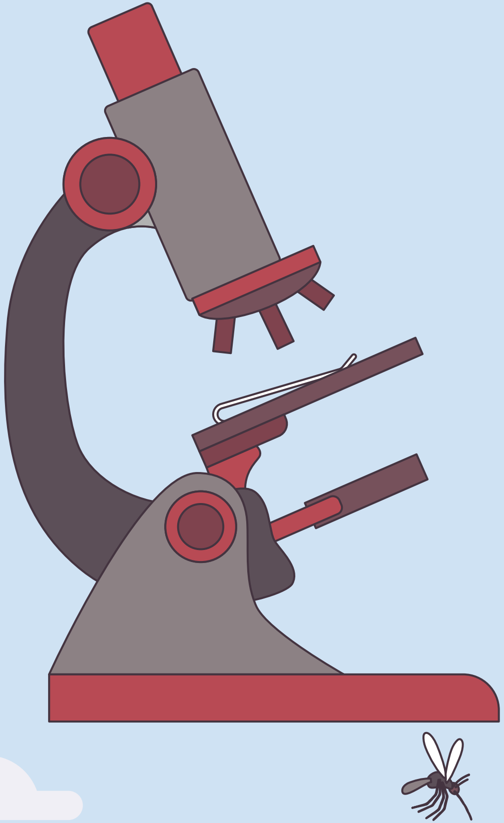


Evaluation Metrics:

- ROC-AUC
- Recall

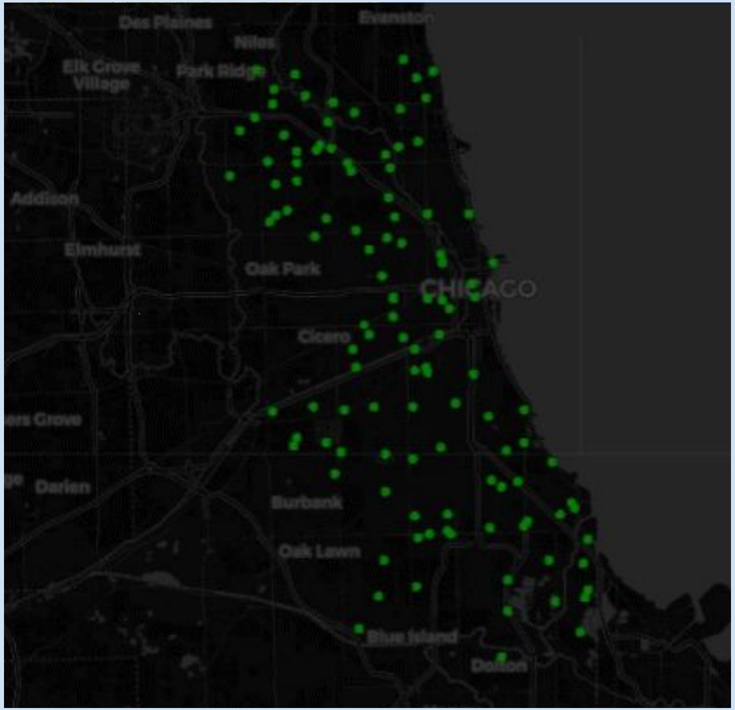
Datasets used :

- Weather
- Spray
- Train
- Test



Exploratory Data Analysis

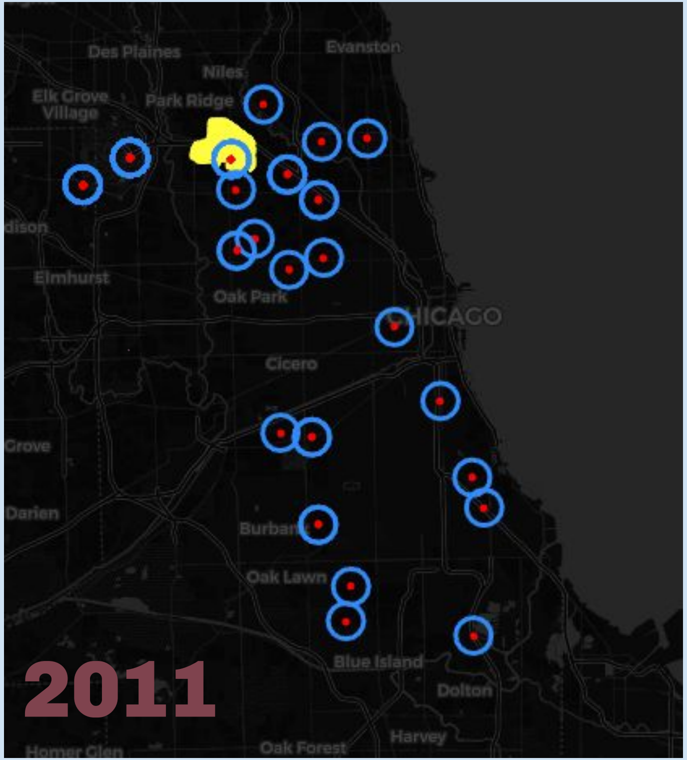
Trap Locations



Legend

- 2007
- 2009
- 2011
- 2013

Spray/Positive Traps (2011, 2013)



Legend

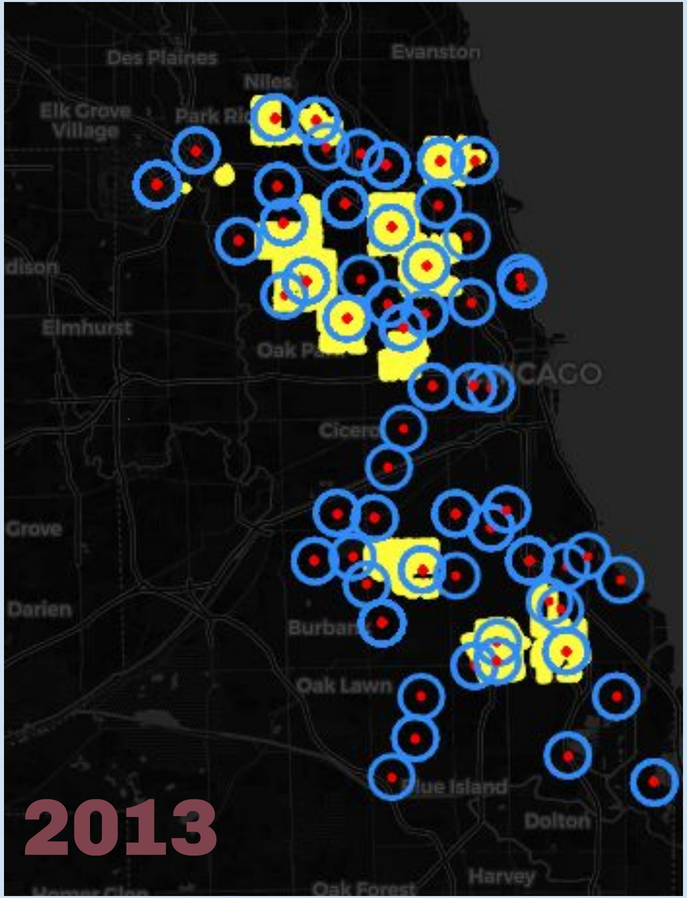
Spray Areas

WNV+ Traps

1.3km Radius

Mosquito Flight range:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278816/

Spray/Positive Traps (2011, 2013)



Legend

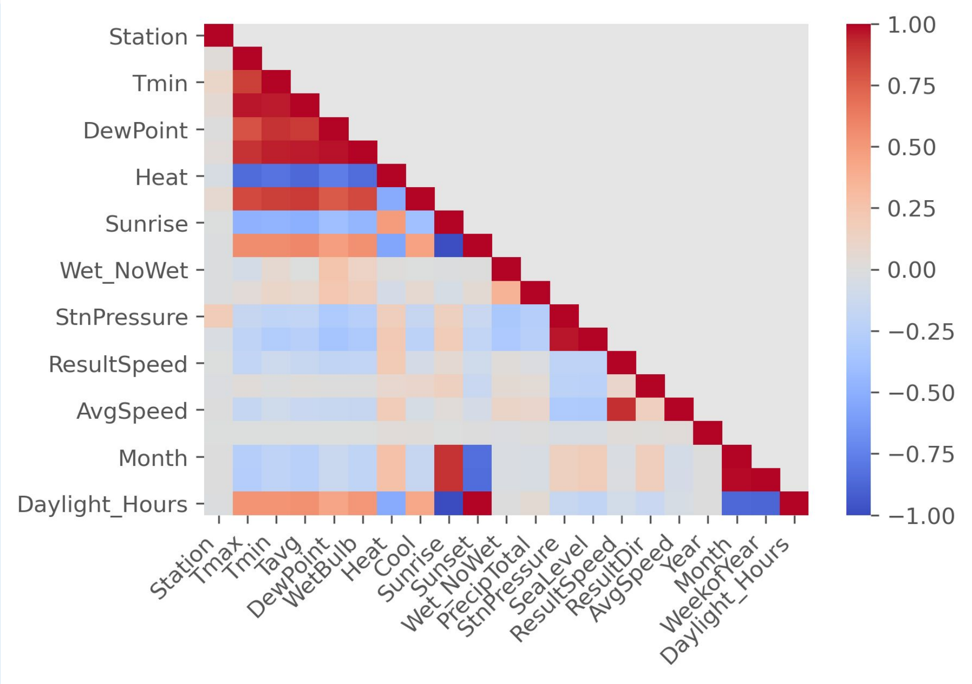
Spray Areas

WNV+ Traps

1.3km Radius

Mosquito Flight range:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278816/

Weather variable groups



Temperature:
Tmax, Tmin, Tavg, Heat, Cool

Humidity:
Dewpoint, Wetbulb

Time:
**Daylight_Hours*, Sunrise, Sunset

Pressure:
StnPressure, SeaLevel

Wind Speed:
AvgSpeed, ResultSpeed

Wind Direction:
ResultDir

Wet Weather:
**Wet_NoWet*, PrecipTotal

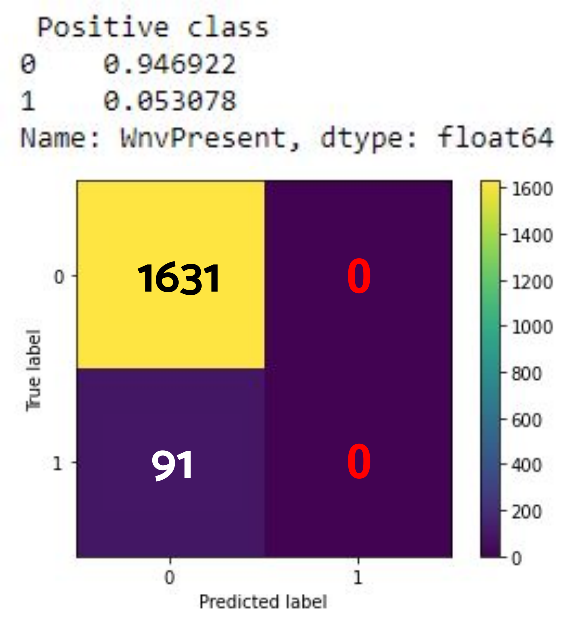
**: Engineered features*

Baseline Model

- Inherent imbalanced class distribution

- Confusion Matrix (LogisticRegression)

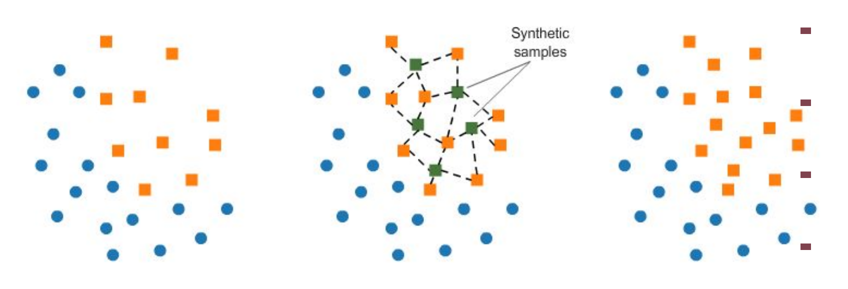
- Use SMOTE to overcome class imbalance



Pipeline steps

Standard Scaler

SMOTE



Classifiers

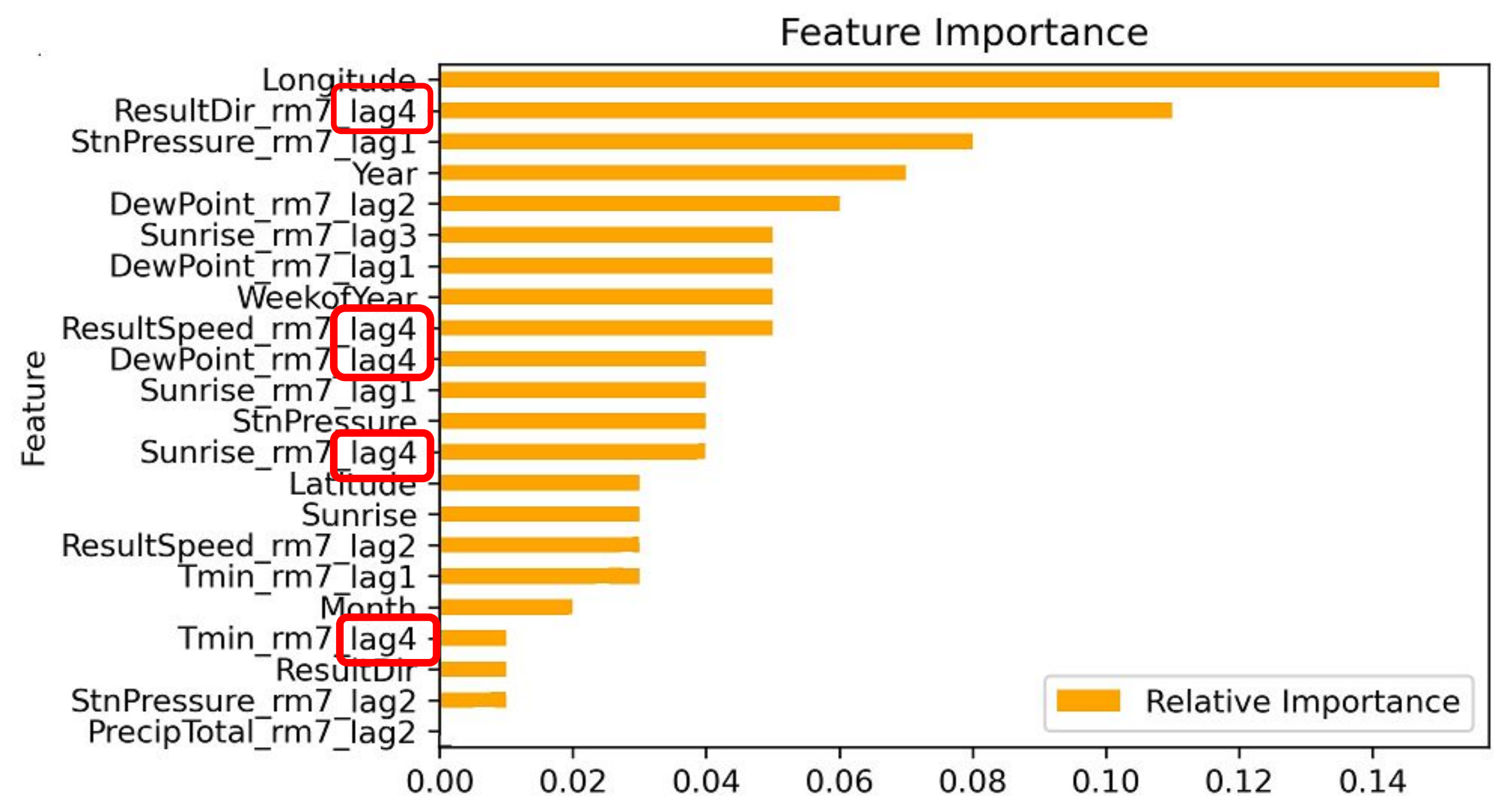
- Logistic Regression
- K Nearest Neighbours
- Decision tree
- Random Forest
- AdaBoost

GSCV Model Scores

	Classifier	Train ROC-AUC Score	Val ROC-AUC Score	Best Parameters	F1	Precision	Recall	Accuracy
0	AdaBoostClassifier()	0.769356	0.77956	{'clf__learning_rate': 0.1, 'clf__n_estimators': 100, 'clf__random_state': 42}	0.222826	0.12	0.901099	8
1	RandomForestClassifier()	0.791749	0.768264	{'clf__max_depth': 8, 'clf__n_estimators': 30, 'clf__n_jobs': -1, 'clf__random_state': 42}	0.228826	0.132302	0.846154	0.698606
2	LogisticRegression()	0.756251	0.749917	{'clf__C': 0.5005, 'clf__l1_ratio': 0, 'clf__max_iter': 750, 'clf__n_jobs': -1, 'clf__penalty': 'l1', 'clf__random_state': 42, 'clf__solver': 'liblinear'}	0.215827	0.124172	0.824176	0.683508
3	DecisionTreeClassifier()	0.755334	0.749776	{'clf__criterion': 'gini', 'clf__max_depth': 5, 'clf__min_samples_leaf': 8, 'clf__random_state': 42}	0.200247	0.112813	0.890110	0.624274
4	KNeighborsClassifier()	0.886681	0.707777	{'clf__n_neighbors': 7, 'clf__p': 2}	0.242291	0.151515	0.604396	0.800232

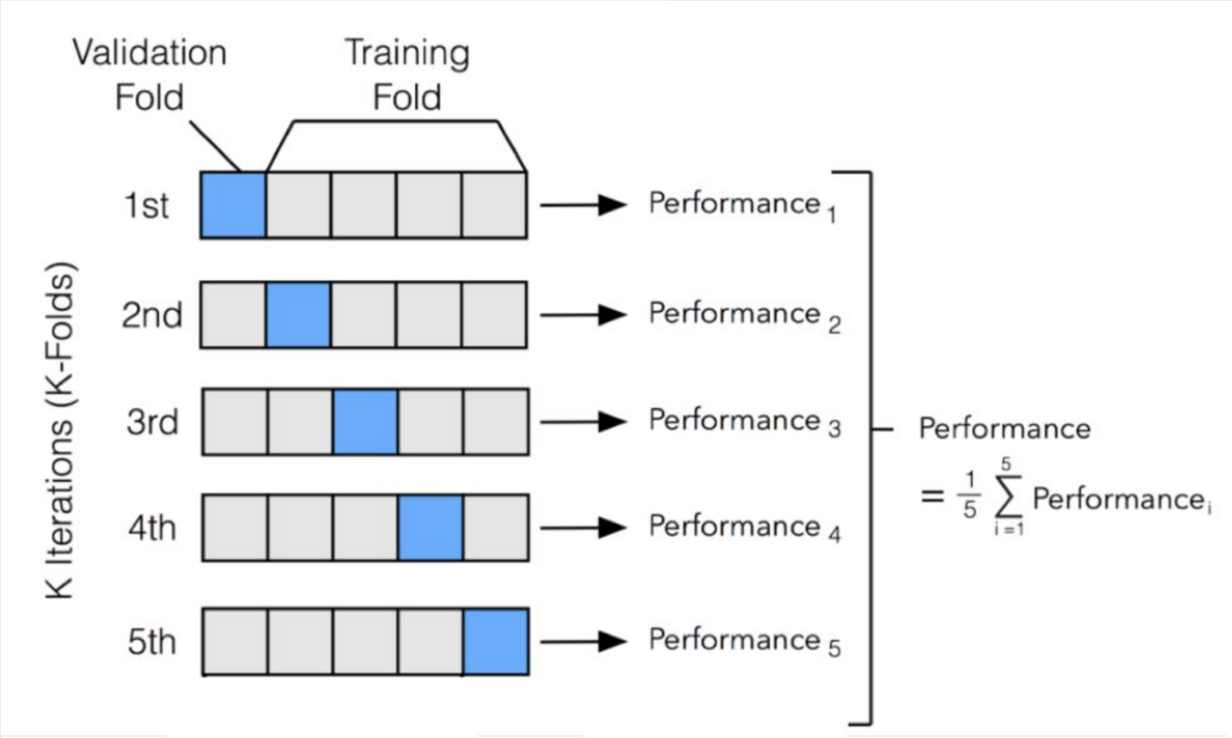
- Best Kaggle Model (ROC-AUC)
- Best business case model (Recall)

Feature Importance



Limitations of modelling process

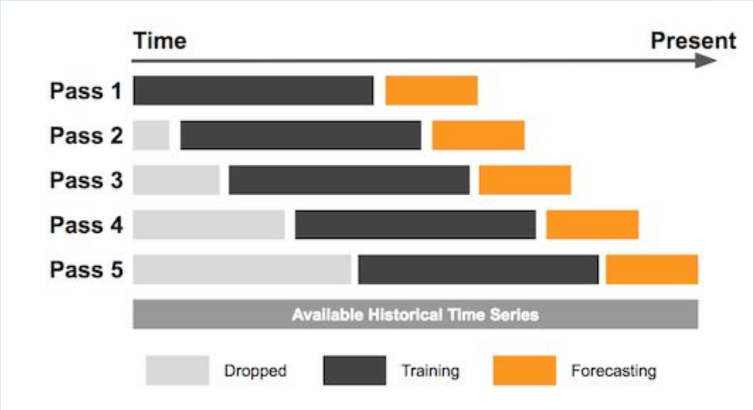
Look ahead bias and CV's role



2007 → 2013

Walk forward validation

Expanding Window



Sliding Window

Cost Benefit Analysis

Cost Benefit Analysis

Item	Result	Remarks
emergency spray cost	$899,048	excluding overtime
sprayed area	477km^2	
spray cost per km^2	$1,884.70	
spray area per trap (1.3km radius)	5.31km^2	
spray cost per trap	**$10,007.76**	excluding overtime
Total medical costs accrued by all WNV patients	$ 2,140,409.00	lculated using spray area per trap * population density
Number of patients	46 pax	1 in 150 cases are severe
Medical cost per patient	$ 46,530.63	$63,383 per pax (emergency spray cost / 46 cases)
------------------------------	----------	-------------------
number of traps to spray	4051	total predicted positive, averaged accross 4 years 2008, 2010, 2012, 2014
total spray cost incurred	$40,541,423.60	

Source:https://wwwnc.cdc.gov/eid/article/16/3/09-0667_article

Cost Benefit Analysis

Item	Result	Remarks
emergency spray cost	$899,048	excluding overtime
sprayed area	477km^2	
spray cost per km^2	$1,884.70	
spray area per trap (1.3km radius)	5.31km^2	
spray cost per trap	**$10,007.76**	excluding overtime
number of people covered within spray area	**77,993**	calculated using spray area per trap * population density
likelihood of severe infection	522	1 in 150 cases are severe
cost for medical treatment	$33,085,926	$63,383 per pax (emergency spray cost / 46 cases)
--------------------------------	----------	--------------------
number of traps to spray	4051	total predicted positive, averaged accross 4 years 2008, 2010, 2012, 2014
total spray cost incurred	$40,541,423.60	

ROI for every $1 spent on spray : $~3306.03

Sources: Refer to reference section

Conclusion

Where

- Targeted spraying using median flight radius of 1.3km, hotspots can be identified using count of overlapping radii of WNV+ traps

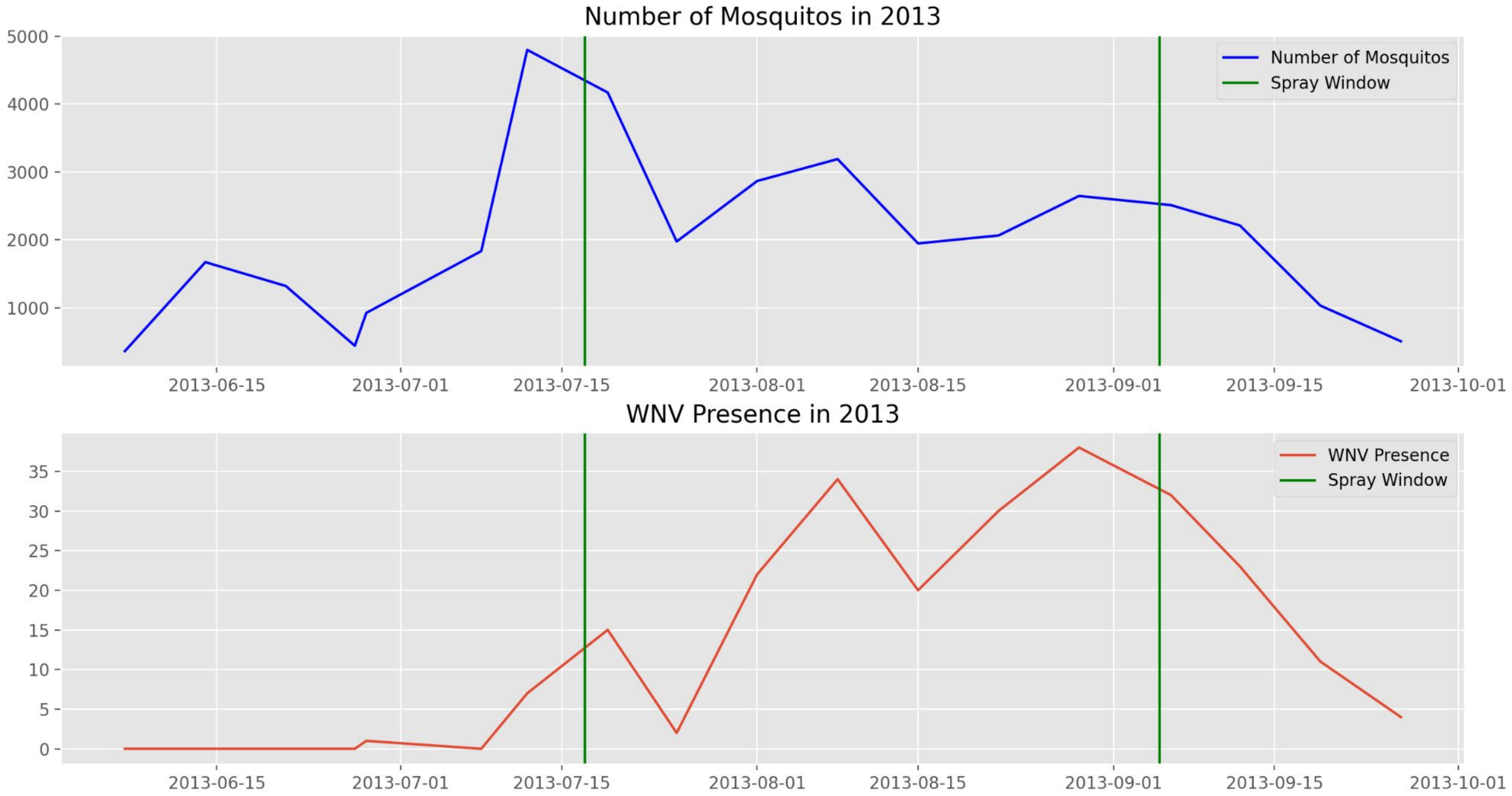
< 1 : Low (0.2km) < 3 : Moderate (1.3km) < 5 : High (2km)

When

- Spraying can be done in advance to curb mozzie population before it spirals out of control:

  Optimal window to spray in advance : 2-4 weeks before peak season (Summer)

  Continue sprays at intervals of every two weeks till mid-summer


Number of Mosquitos in 2013
5000
4000
3000
2000
1000
Number of Mosquitos
Spray Window
2013-06-15
2013-07-01
2013-07-15
2013-08-01
2013-08-15
2013-09-01
2013-09-15
2013-10-01
WNV Presence in 2013
35
30
25
20
15
10
5
0
WNV Presence
Spray Window
2013-06-15
2013-07-01
2013-07-15
2013-08-01
2013-08-15
2013-09-01
2013-09-15
2013-10-01

Recommendations

- Monitor other known vectors (Avians, Equine Species) for WNV+
- Include breakdown of male-female species (males don’t bite/lay eggs)
- Explore feasibility of Sterile Insect techniques before peak season
- Public outreach / education programs to reduce breeding spots (Can affect trap accuracy)

“Treatment without prevention is simply unsustainable”

— Bill Gates

References

Sacramento Country Medical Costs
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2600250/table/T1/?report=objectonly
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/
https://wwwnc.cdc.gov/eid/article/16/3/09-0667_article

Chicago Population
https://www.macrotrends.net/cities/22956/chicago/population

Mosquito Flight range
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278816/

Inflation Calculator:
https://smartasset.com/investing/inflation-calculator#hdQLsKK2RU

Q&A

Problems Implementing Walk forward validation

So can we use it in this project?

- Cold start problem

- Ground truths for test years not available to use for training in subsequent periods

- Merged test data will be incorporating results from 4 different models

So can we use it in this project?

Train - test Split

- Train 1 (**2005**, **2006**, 2007)
    - Test 1 (2008)
- Train 2 (2007, **2008**, 2009)
    - Test 2 (2010)
- Train 3 (2009, **2010**, 2011)
    - Test 3 (2012)
- Train 4 (2011, **2012**, 2013)
    - Test 4 (2014)

Yields

- Model 1 (**2005**, **2006**, 2007)
- Model 2 (2007, **2008**, 2009)
- Model 3 (2009, **2010**, 2011)
- Model 4 (2011, **2012**, 2013)

No ground truth available, cannot use prior predictions