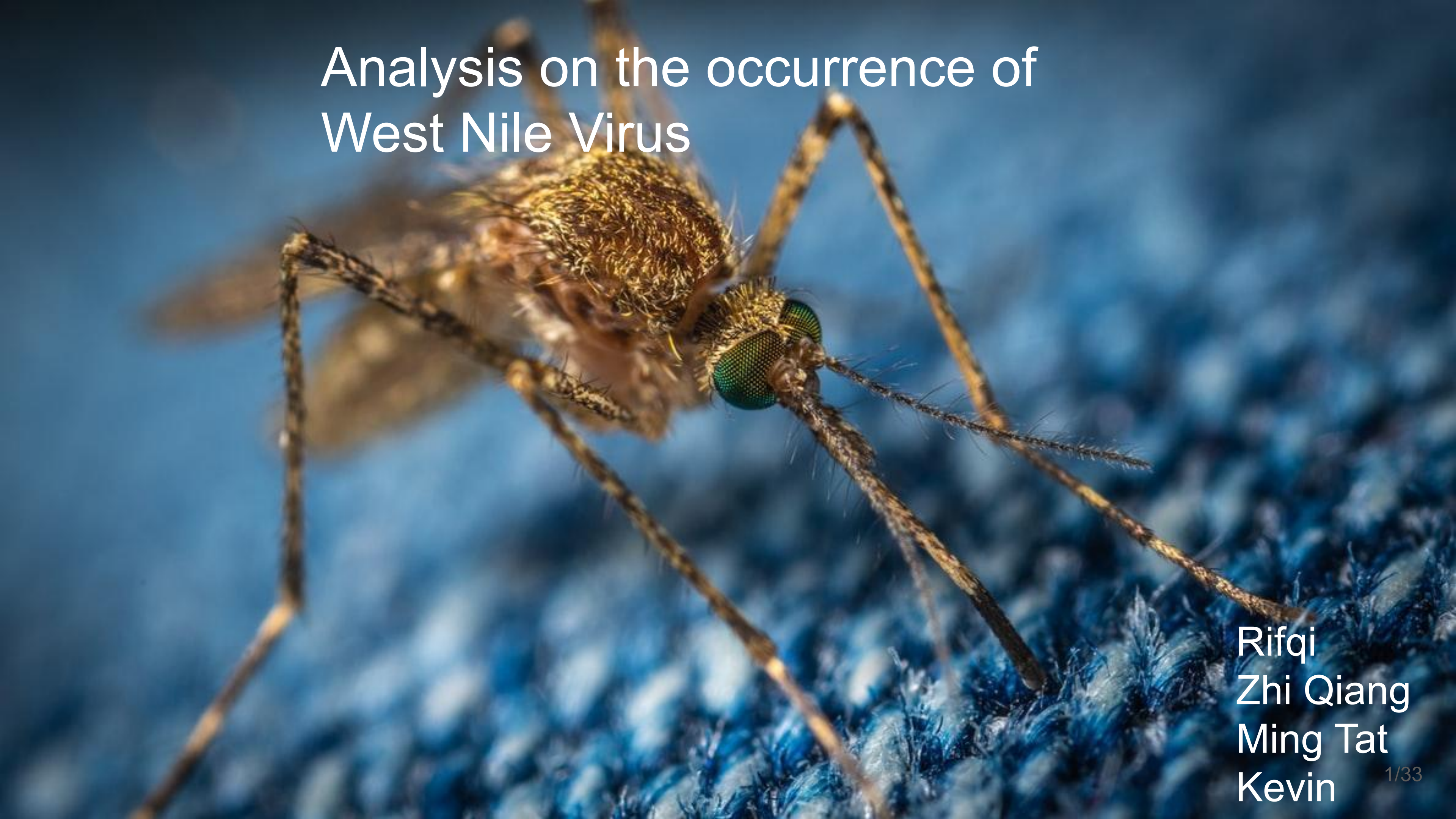
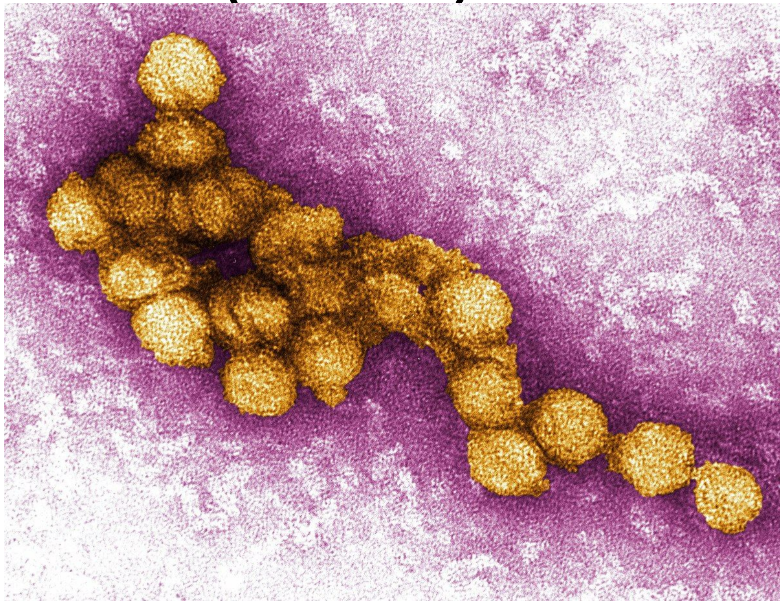


Analysis on the occurrence of West Nile Virus

Rifqi
Zhi Qiang
Ming Tat
Kevin

Background on West Nile Virus (WNV)

- Leading cause of mosquito-borne disease in US
  - May cause death
- No vaccines/medications
- Recent epidemic in Windy City



A micrograph of the West Nile Virus

Information sources:
https://www.cdc.gov/westnile/index.html

https://datasmart.ash.harvard.edu/news/article/predictive-analytics-guides-west-nile-virus-control-efforts-in-chicago-1152#:~:text=Predictive%20Analytics%20Guides%20West%20Nile%20Virus%20Control%20Efforts%20in%20Chicago,-By%20Sean%20Thornton&text=In%202002%2C%20Chicago%20suffered%20its,mitigating%20the%20risk%20of%20transmission.
https://en.wikipedia.org/wiki/File:West_Nile_Virus_Image.jpg

Problem statement

- How can we predict the next possible occurrence of the virus at various locations in view of the recent epidemic to mitigate the spread?
- How to mitigate the epidemic in a cost-effective way?

Methodology

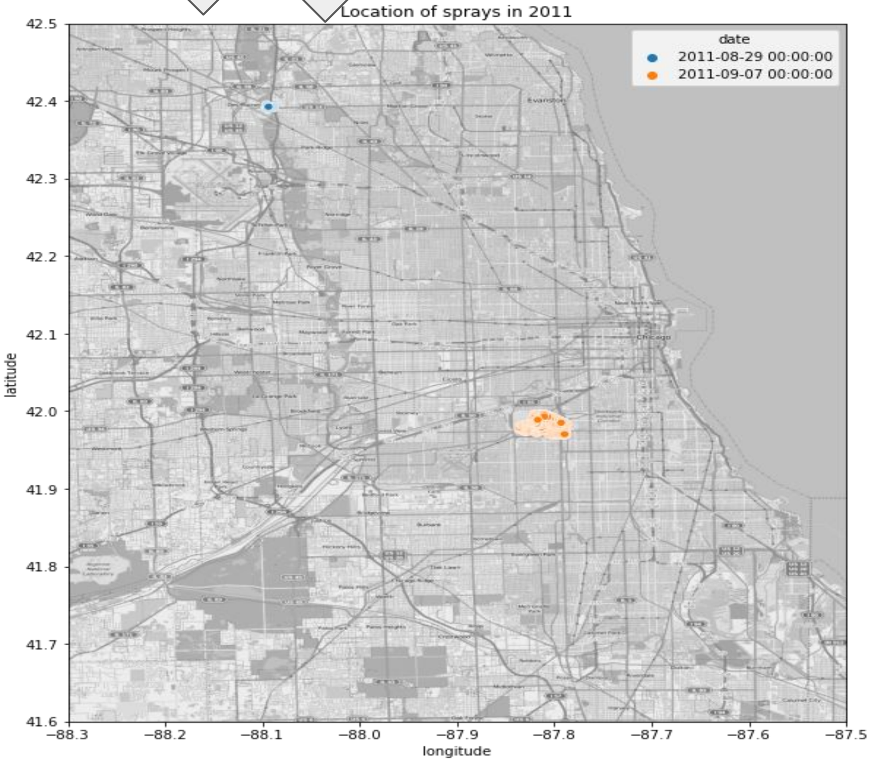
1. Data EDA
2. Data merging
3. Modelling and comparison
4. Cost Benefit Analysis
5. Findings and recommendations

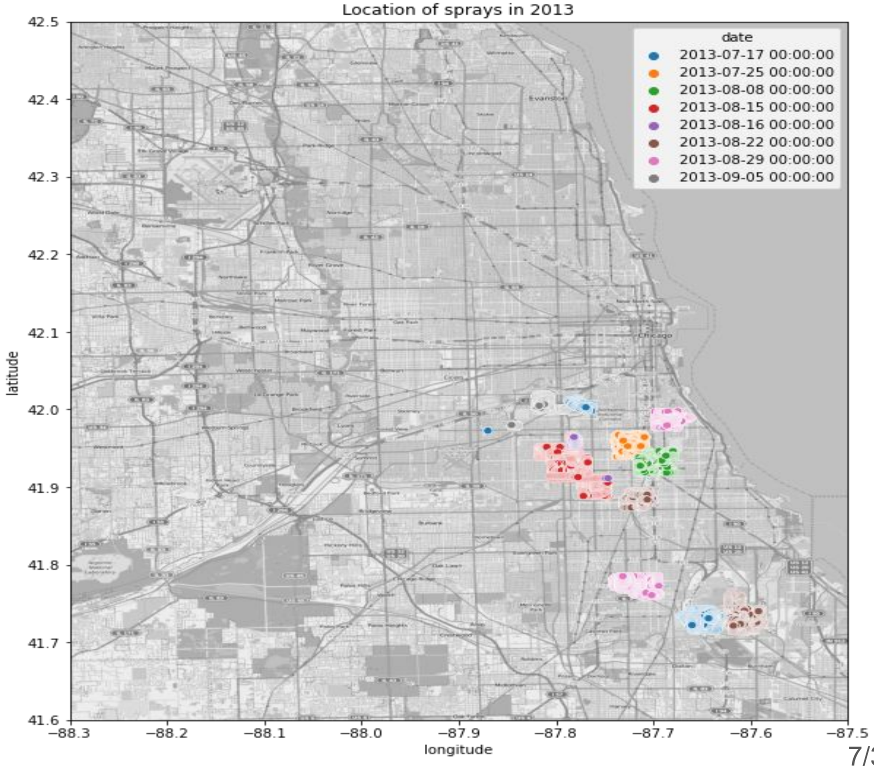
EDA & Feature Engineering

Trap #

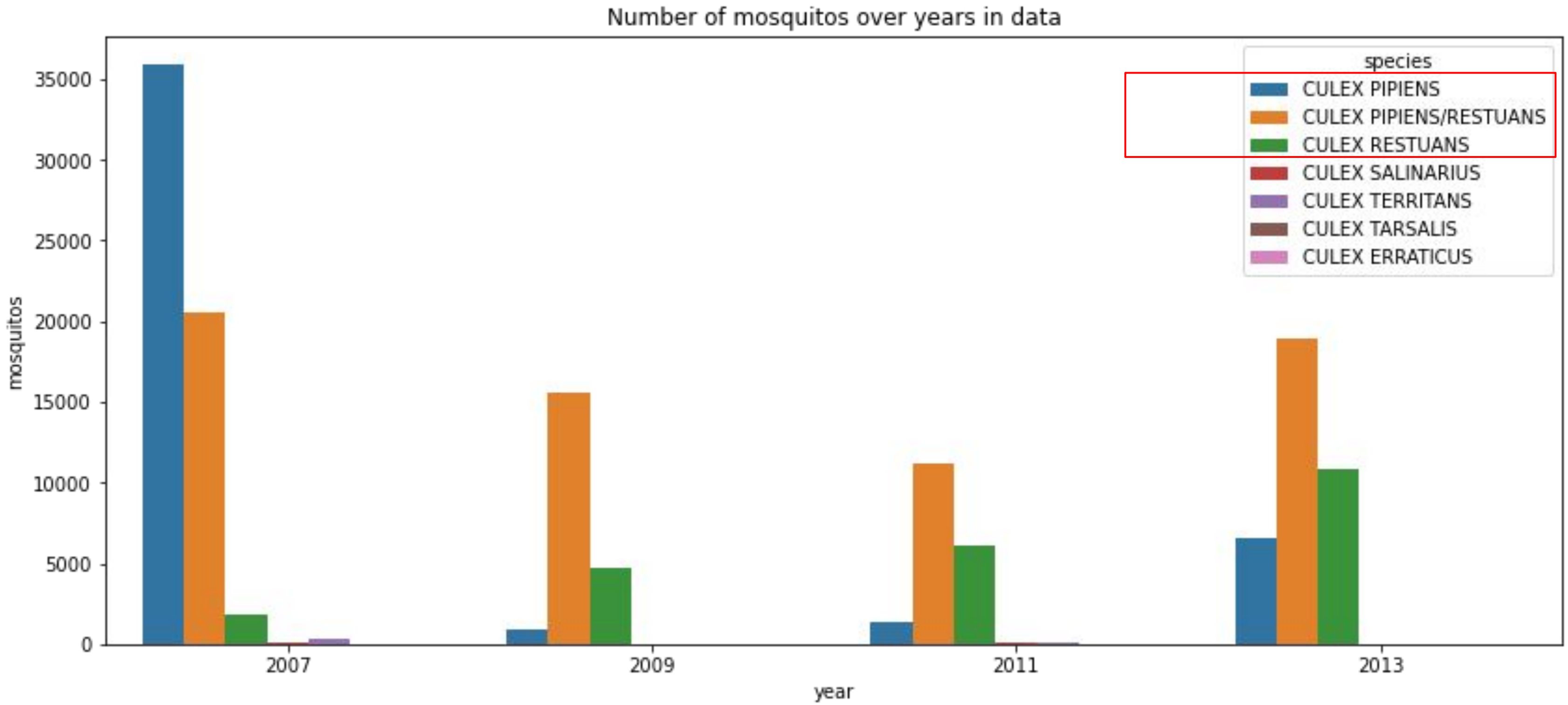
# of mosquitos

Spray

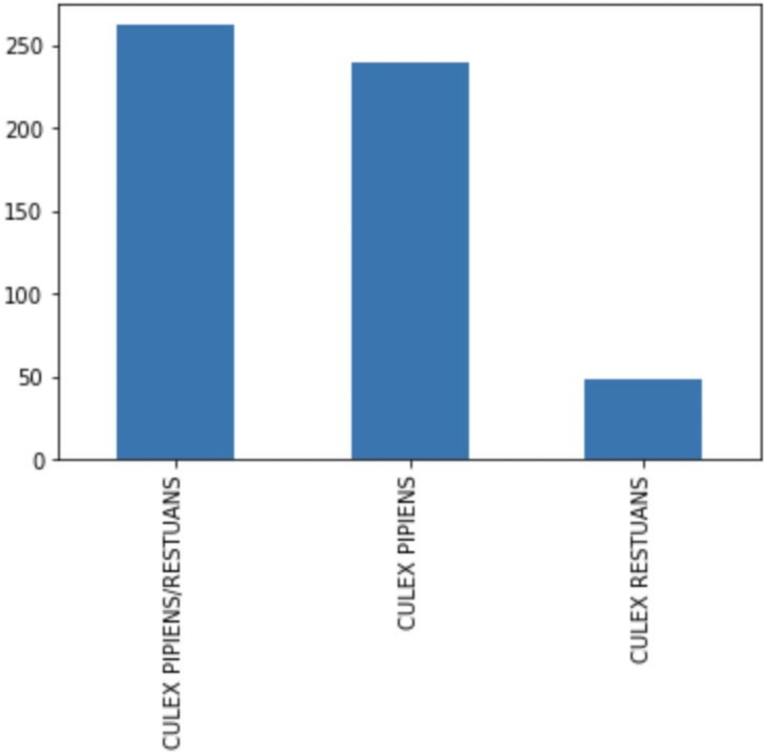

Location of sprays in 2011
date
2011-08-29 00:00:00
2011-09-07 00:00:00
latitude
longitude
42.5
42.4
42.3
42.2
42.1
42.0
41.9
41.8
41.7
41.6
−88.3
−88.2
−88.1
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5
Evanston
Chicago


Location of sprays in 2013
date
2013-07-17 00:00:00
2013-07-25 00:00:00
2013-08-08 00:00:00
2013-08-15 00:00:00
2013-08-16 00:00:00
2013-08-22 00:00:00
2013-08-29 00:00:00
2013-09-05 00:00:00
latitude
longitude
42.5
42.4
42.3
42.2
42.1
42.0
41.9
41.8
41.7
41.6
−88.3
−88.2
−88.1
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5
Evanston
Chicago

Mosquito population over years in data



Number of mosquitoes with West Nile Virus



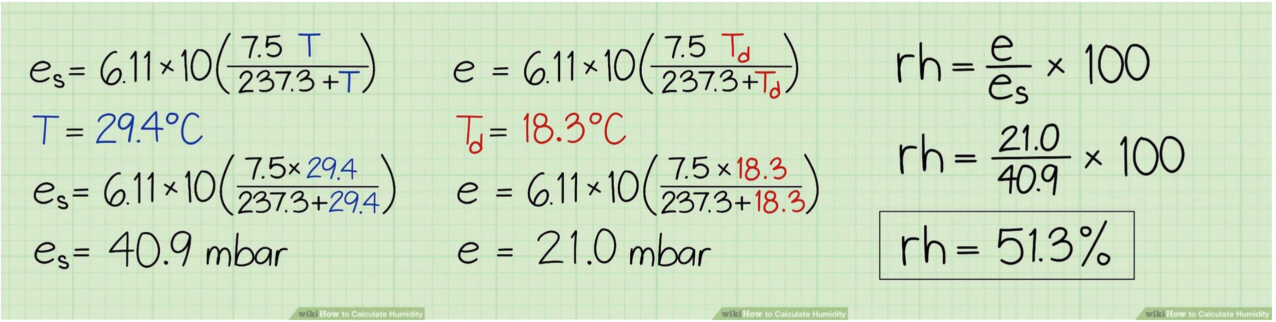
Ordinal category by proportion:

1) Culex Pipiens/Restuans: 2
2) Culex Pipiens: 2
3) Culex Restuans: 1
4) Other species: 0

Encoding the codes to categories

codesum	MI	DZ	FU	FG	BC	FG+	TS	SN	HZ	SQ	BR	RA	GR	VC
{}	0	0	0	0	0	0	0	0	0	0	0	0	0	0
{HZ, BR}	0	0	0	0	0	0	0	0	1	0	1	0	0	0
{HZ}	0	0	0	0	0	0	0	0	1	0	0	0	0	0
{RA}	0	0	0	0	0	0	0	0	0	0	0	1	0	0
{}	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Create relative humidity



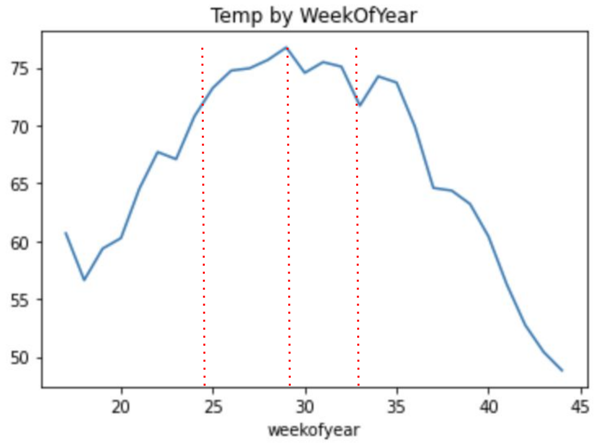
Create rolling temporal features

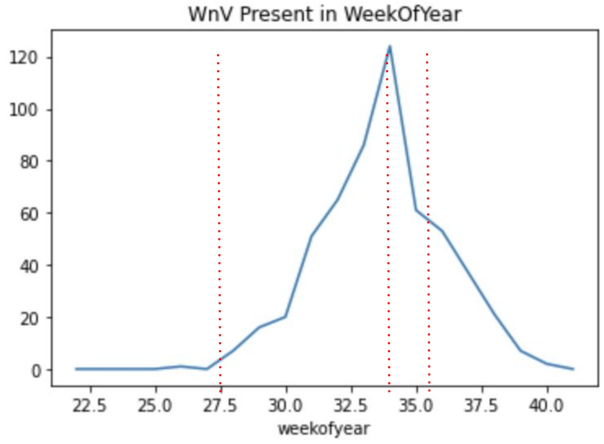
5/14/28 days (Based on mosquito and WNV incubation period)

rel_humid_lag5	rel_humid_lag14	rel_humid_lag28
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN
NaN	NaN	NaN
39.634503	NaN	NaN
...	...	...
37.601956	55.170580	50.068099
34.340974	50.591575	48.476990

- Average temp
- Relative humidity
- Average speed
- Total precipitation

Effect of temperature on WnV

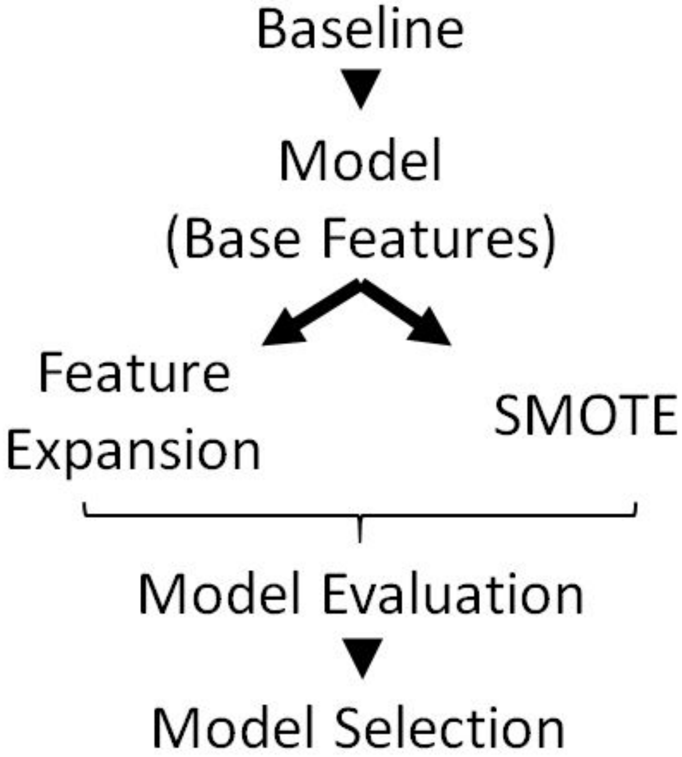




Modelling

Methodology

- 5 Models
  - Logistic Regression
  - Gradient Boost
  - XG Boost
  - Random Forest
  - Extra Trees
- Optimised for ROC-AUC
  - Gathered other metrics as well

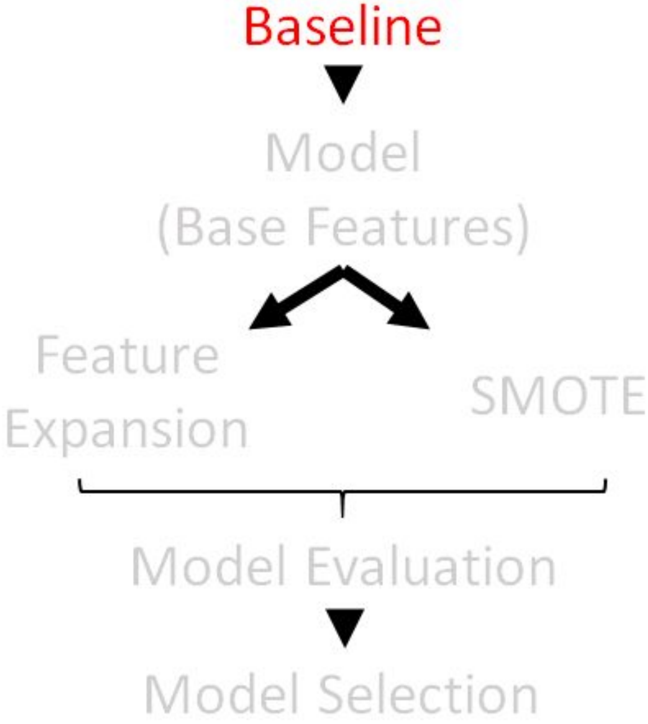


Baseline

```
# Baseline
y.value_counts(normalize=True)
```

```
0    0.946077
1    0.053923
Name: wnvpresent, dtype: float64
```

- Imbalanced data
- ROC-AUC: **0.500**

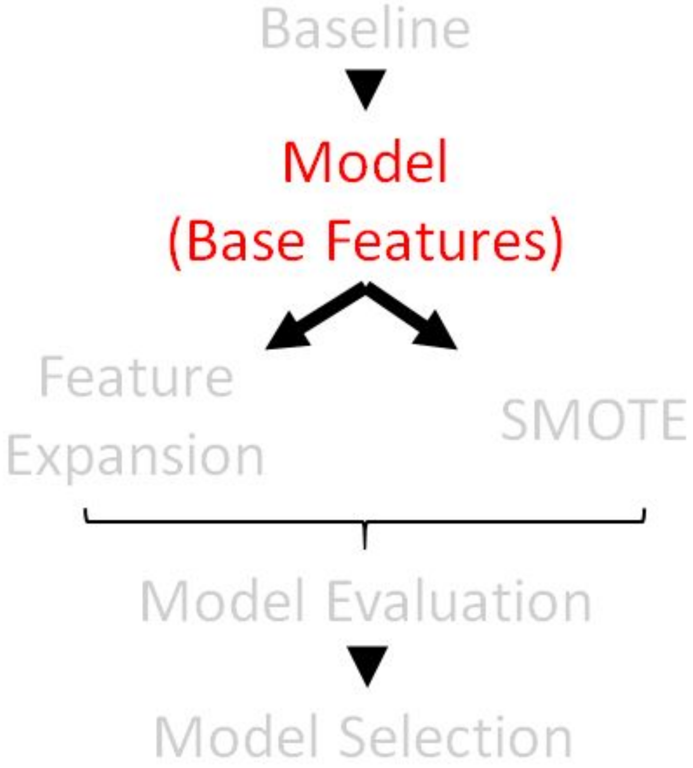


‘Vanilla’ Model with Base Features

- 49 features
- GridSearchCV for each model

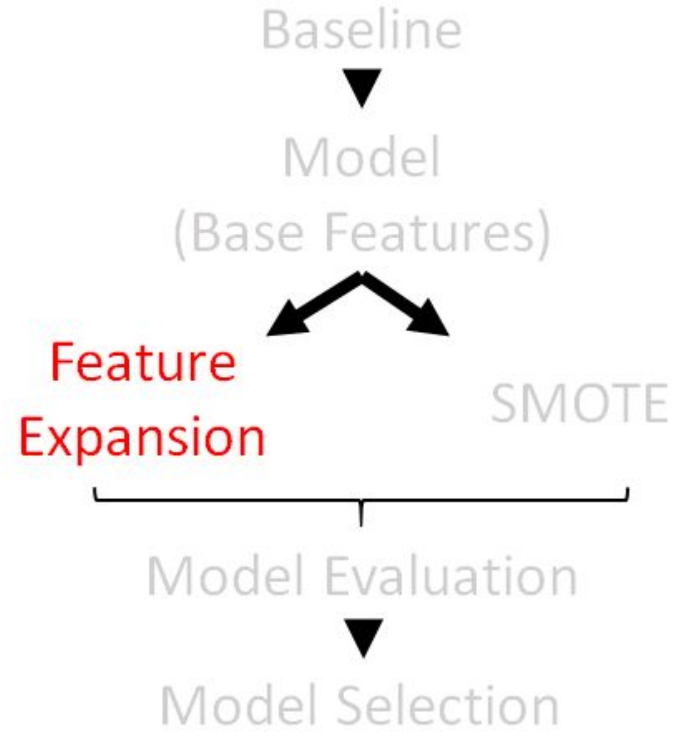
- Top performing model: XGBoost
  - Gamma: 0.3, learning_rate: 0.2, max_depth: 3
  - Validation ROC-AUC – 0.87
  - Recall – 0.051

- All models had poor Recall & F1 Score



Feature Expansion

- Try and improve metrics
- PolynomialFeatures
  - Degree = 2
- 1274 features
  - Might have too many noisy features
- Reduce dimensionality/complexity
  - PCA
  - Filtering

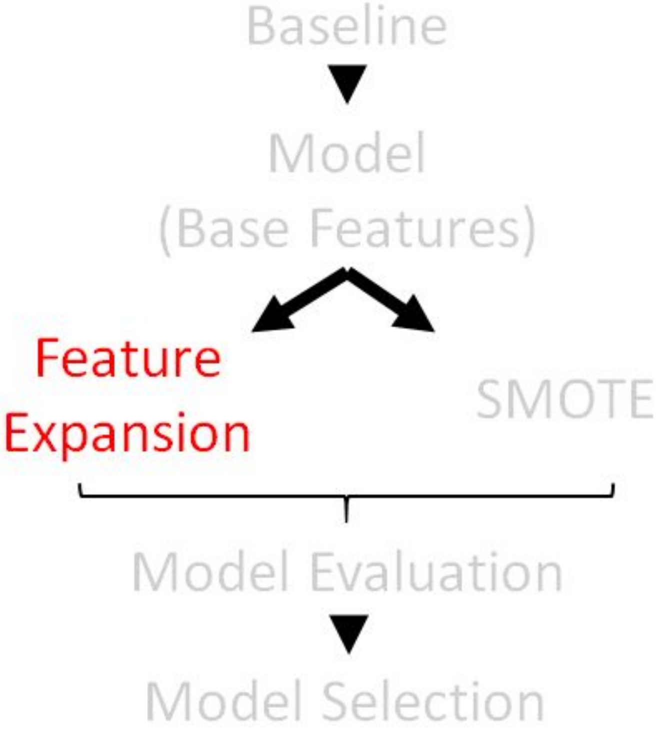


Principal Component Analysis (PCA)

- To reduce dimensionality
- GridSearchCV + Pipeline
  - Each model
  - 30/40/50 Principal Components

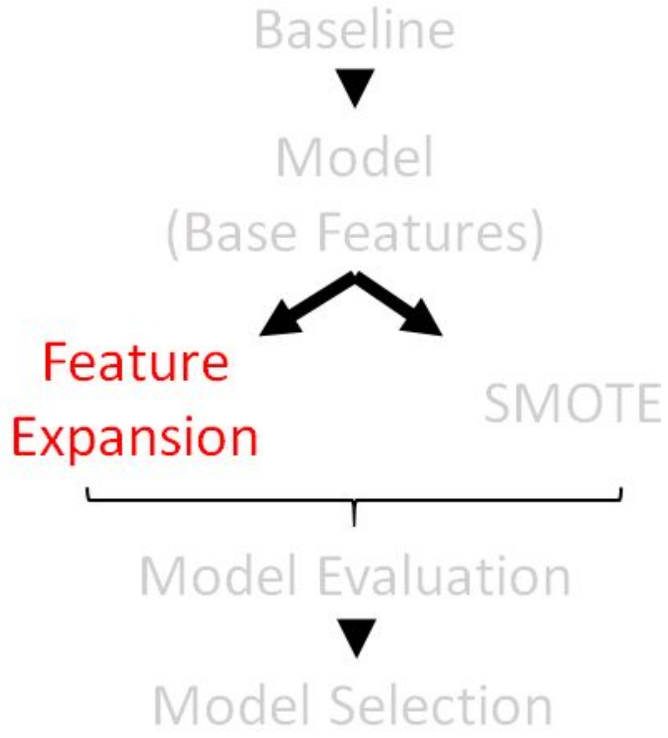
- Top performing model: RandomForest
  - PC: 50, max_depth: 6, n_estimators: 200
  - Validation ROC-AUC – 0.846
  - Recall – 0.014599

- Not much difference from vanilla model



Filtering

- By Pearson's Correlation with target
- GridSearchCV + Pipeline
  - Each model
  - Different correlation cutoff points
- Top performing model: XGBoost
  - Cutoff: 0.01
  - Gamma: 0.2, learning_rate: 0.1, max_depth: 3
  - Validation ROC-AUC – 0.857
  - Recall – 0.014599
- Not much difference from vanilla model

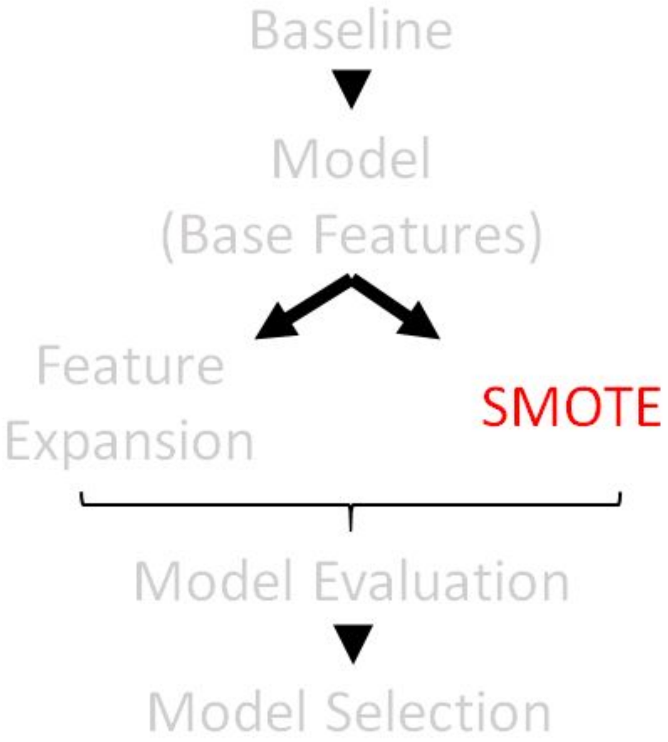


SMOTE

- To address class imbalance
  - Improve recall
- GridSearchCV for each model

- Top performing model: XGBoost
  - Gamma: 0.3, learning_rate: 0.2, max_depth: 4
  - Validation ROC-AUC – 0.847
  - Recall – 0.54

- Slightly lower ROC-AUC, much better recall!

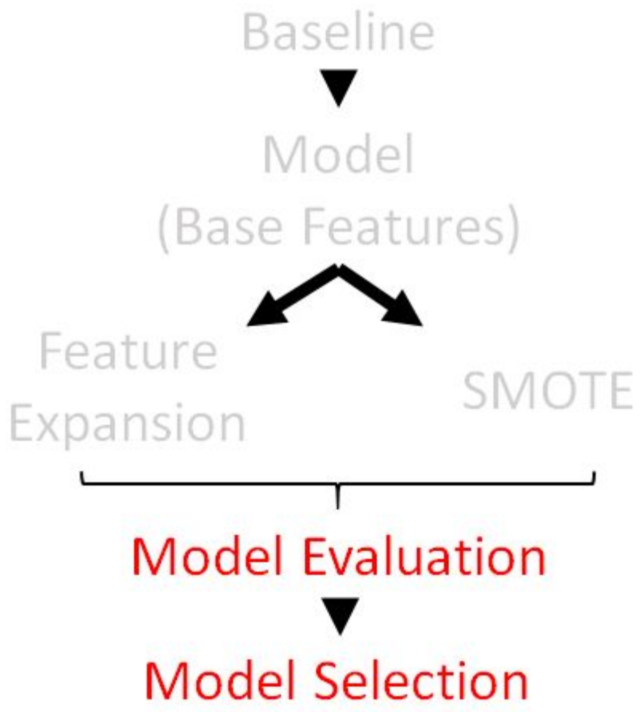


Model Evaluation & Selection

- Compared top 2 models from each method
- Important metrics:
  - ROC-AUC
  - Recall
- Predictive performance > interpretability

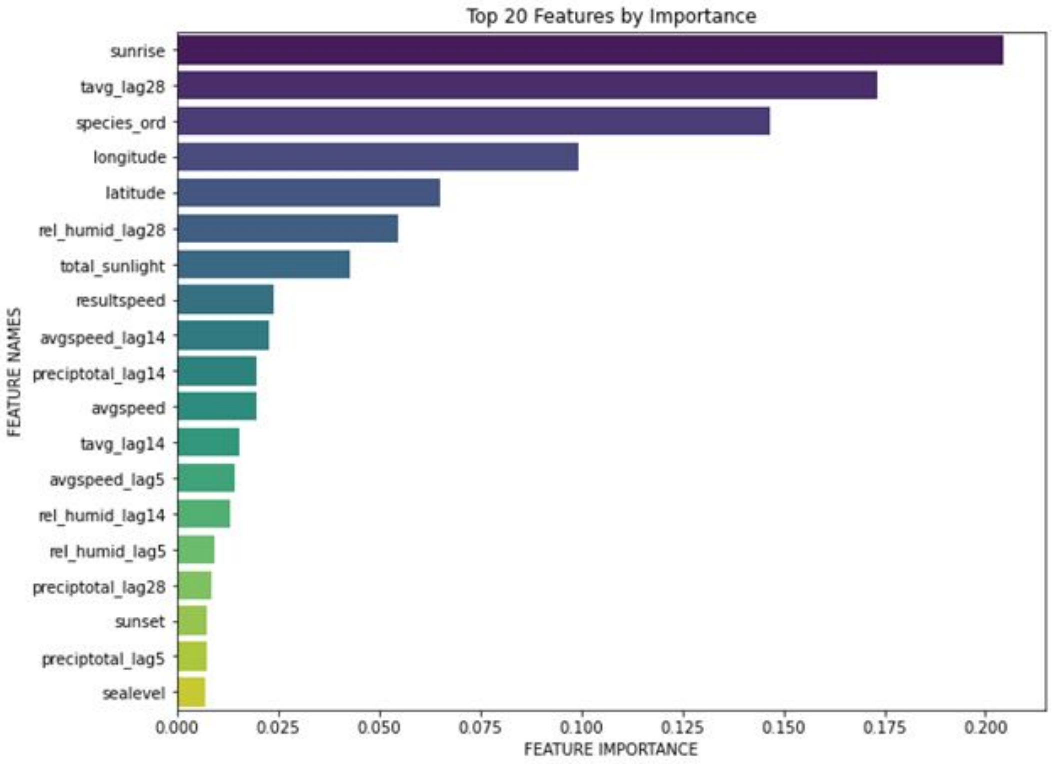
- Best model: GradientBoost with SMOTE
  - max_depth: 3, learning_rate: 0.15
  - Validation ROC-AUC – 0.842
  - Recall – 0.657

- Kaggle submission score: **0.727**



Feature Importance

- Many temporal features
- Features related to time of year
- Location also important



Cost Benefit Analysis

Cost Benefit Analysis Model

Direct Costs: Procuring spray chemicals
Indirect Costs: Productivity loss incurred from seeking treatment
Maximum Benefits Scenario: Department of Public Health adopts aggressive spraying approach.

Computing Direct Costs

Cost per gallon (128 ounces) of Zenivex: ~ US$ 300
Spray rate: 1.5 ounces per acre
Chicago land area: 145,300 acres
Assumption: Department of Public Health decides to conduct spraying efforts biweekly in the summer (~6 times) for the whole of Chicago

Estimated annual direct cost
~ US$ 1.36 million

Computing indirect costs

Mean productivity loss (2012)

Fever	Meningitis	Encephalitis	Acute Flaccid Paralysis
$546	$684	$53,234	$12,357

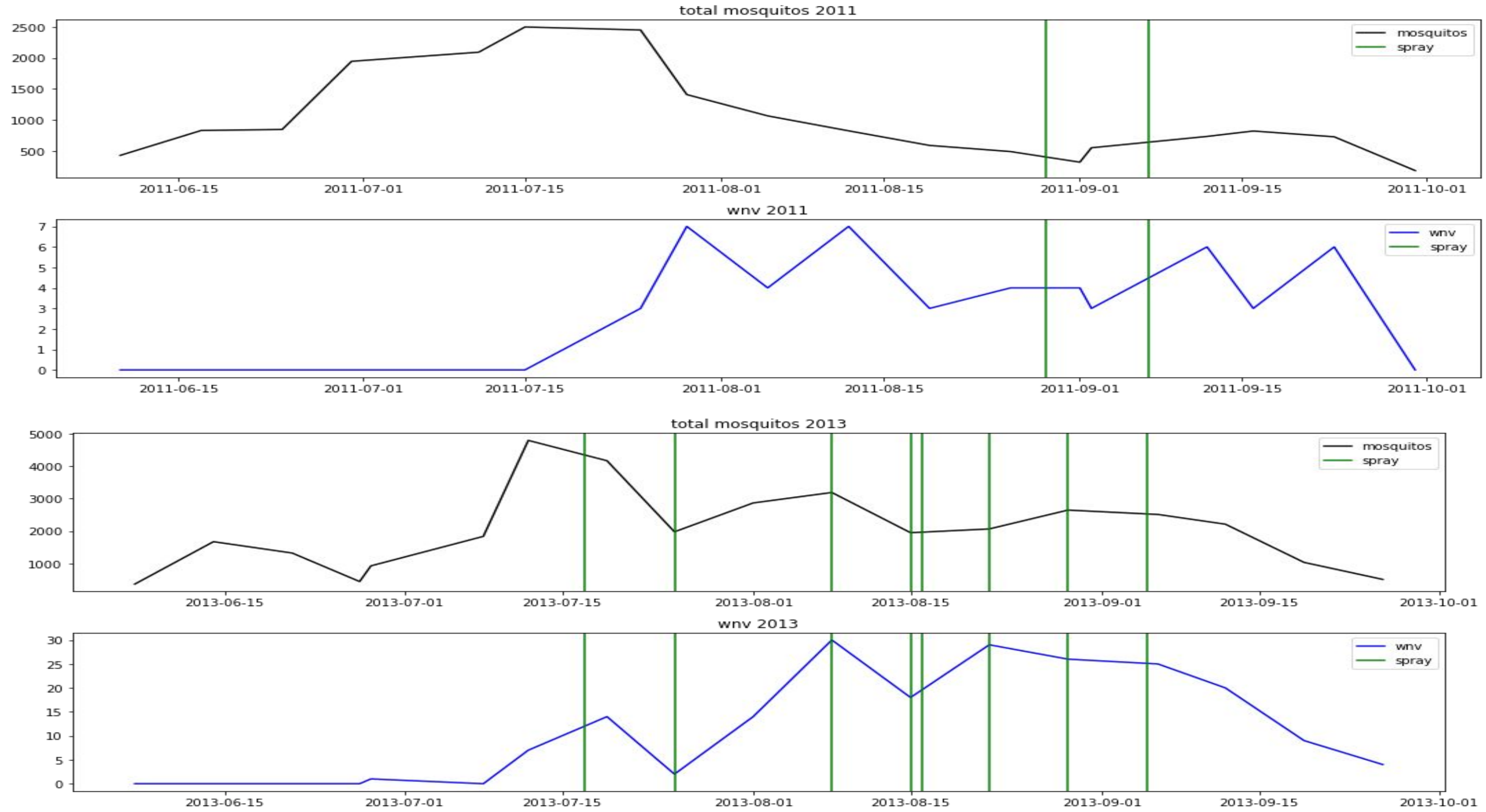
Source: American Journal of Tropical Medicine and Hygiene

WNV cases (2012): 60
Assumption:
-Most cases are asymptomatic, only 1 in 5 cases develop fever and 1 in 150 develop more serious illnesses, e.g. Encephalitis and Acute Flaccid Paralysis
- Aggressive spraying will convert the indirect costs into gains (no one contracts WNV).

Estimated indirect costs
~ US$ 59,786

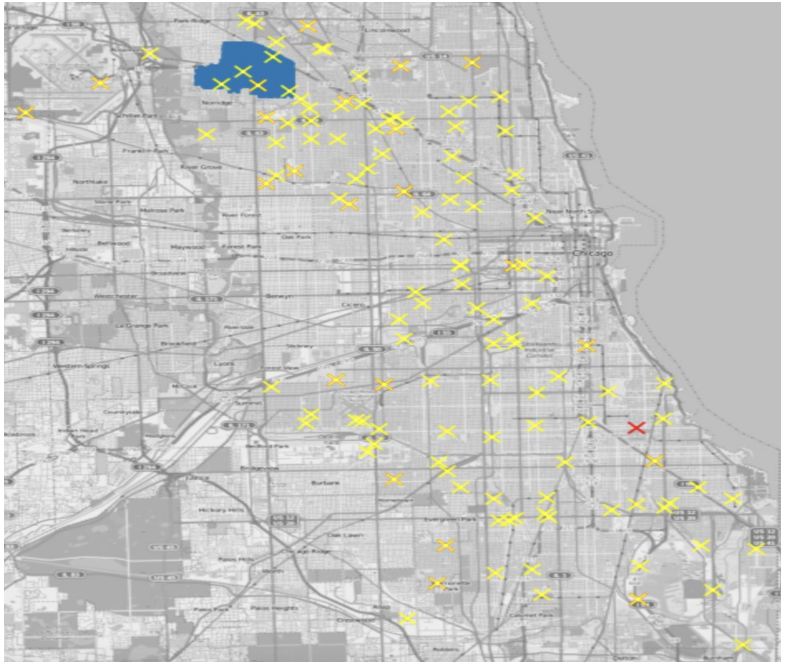
The investment in spraying does not seem to justify the gains. A targeted approach in spraying would be more cost effective.

Effect of sprays on wnv and mosquito population

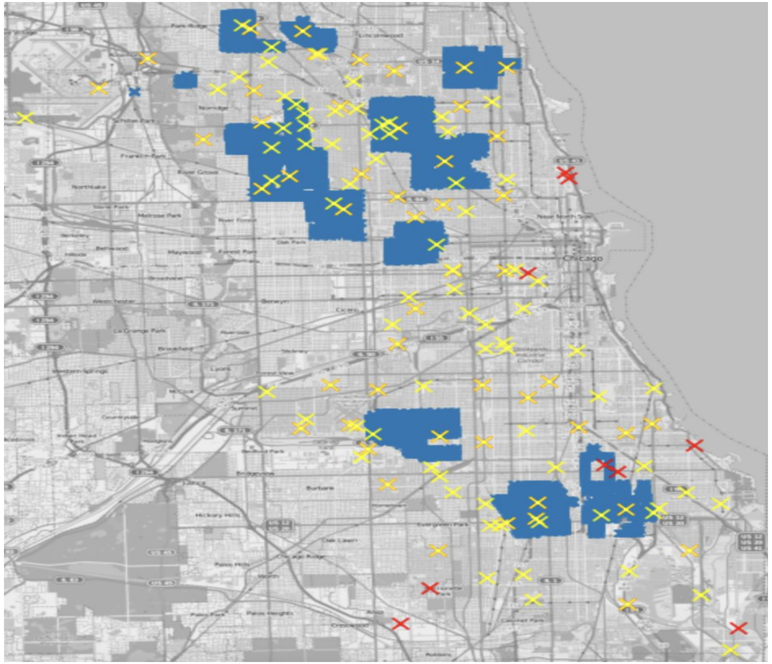


Spraying is not effective

WNV Cluster in 2011/2012



WNV Cluster in 2013/2014



Conclusions and Recommendations

Conclusions

- GradientBoosting with SMOTE: Successful in answering our main problem statement
  - Model predictions can be utilised to improve existing efforts

- Existing spraying efforts ineffective and uneconomical
  - Negative impact to health/environment [BeyondPesticides] [CDC]
  - Cost incurred exceeds potential benefit

Recommendations for epidemic mitigation

- Optimize Spraying Approach
- Public Education
- Enforcement checks

End