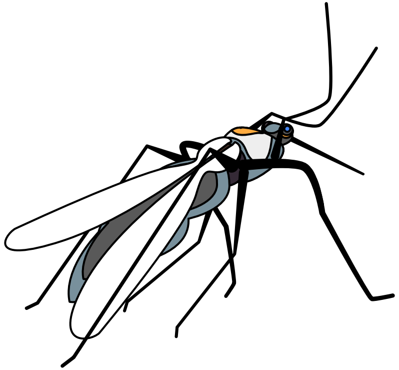
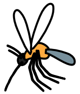
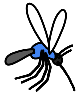


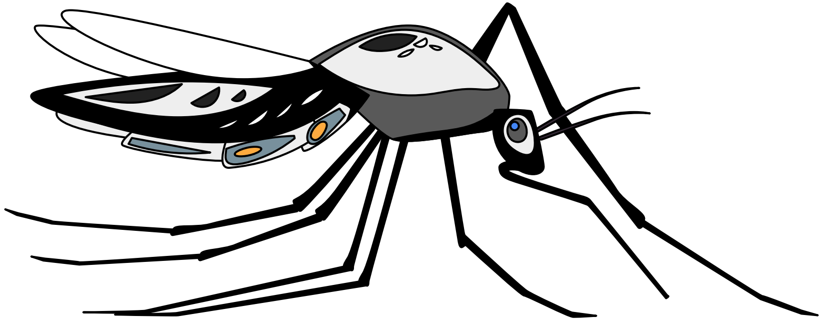
Addressing West Nile Virus (WNV) in Chicago using Data Science





Specially prepared for the Centers for Disease Control and Prevention (CDC)

By: Lee Shi Min, Raymond Onn, Samuel Ng & Jeffrey P



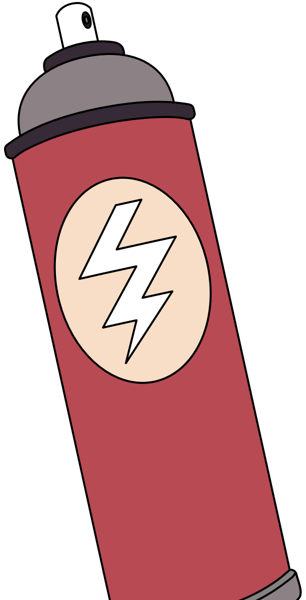
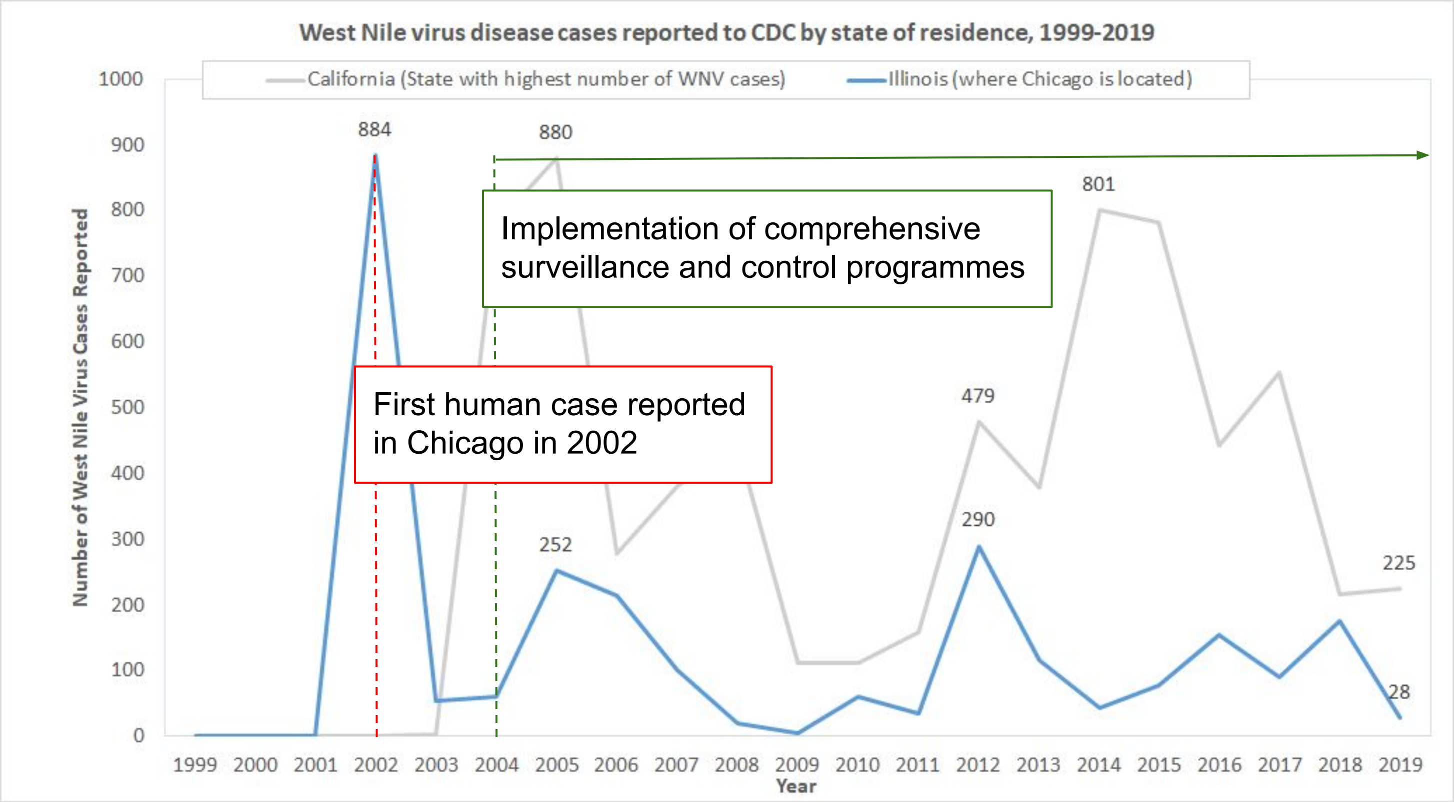
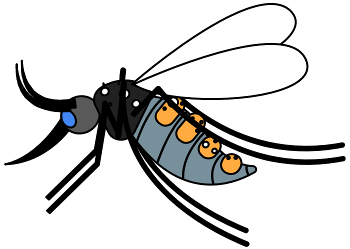


Table of contents

Chicago's comprehensive surveillance and control programmes seemed to have helped in keeping the number of cases at the state-level down

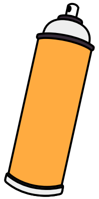




Source: CDC

Chicago’s annual mosquito surveillance & control efforts:

1. Treating catch basins with larvicides	2. Placement of mosquito traps for testing of samples	3. Aerial sprays of pesticides



We aim to ensure improve efficient resource allocation towards virus prevention by way of targeted sprays

1. Machine Learning Solution to Predict incidence of WNV for targeted sprays	
• Use past data for prediction	High **ROC AUC** score
• Identify virus when it is present	High **recall and precision** scores (however, both scores tend to be inversely correlated)
• Precise positive prediction of virus presence	

2. Deep dive into the net benefits of past sprays
Visualise the effect of spray efforts in 2011 & 2013 on virus
Analyse benefits and costs of spraying

Datasets

Spray

CSV

- 14,294 spray observations
- Across 2011 & 2013
- 3 features (Location and Date attributes)

Train

CSV

- 10,505 observations
- Across 2007, 2009, 2011 & 2013
- 10 features (Location, Date, NumMosquitos attributes)
- Target variable: WnvPresent

Weather

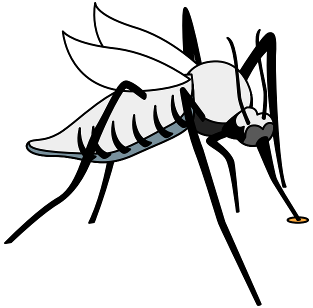
CSV

- Daily weather data collected from 2 weather stations on 1 May to 31 Oct in 2007 to 2014
- 21 features (Station, Date, Weather, e.g. temp, attributes)

Test

CSV

- 116,293 observations
- Across 2008, 2010, 2012 & 2014
- 9 features (Location, Date attributes; missing NumMosquitos)
- Id variable

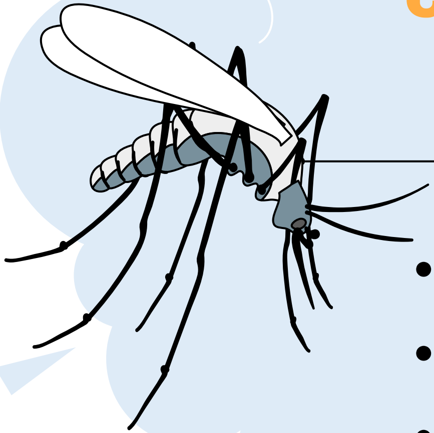


Data visualisation for deep dive into effects of sprays on virus incidence

Merged dataset for model training for prediction of virus incidence

Merged dataset for model prediction on kaggle test dataset

Workflow to develop ML solution



Data Cleaning & EDA

- Removal of outliers
- Impute missing values
- Merge data

Feature Engg

- Lag Variables
- Dummy Variables
- New features (e.g. Relative Humidity)

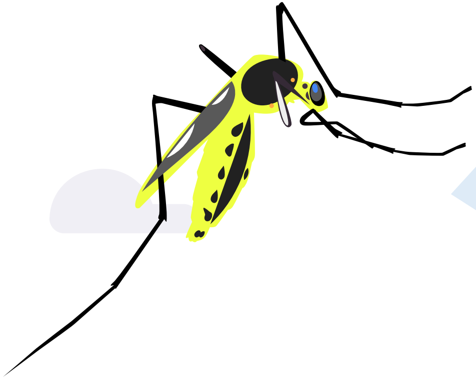
Model Prep & Choice

- SMOTE
- Standard Scaling
- Choose model based on ROC AUC CV, recall and precision score

Model Optimisation & Evaluation

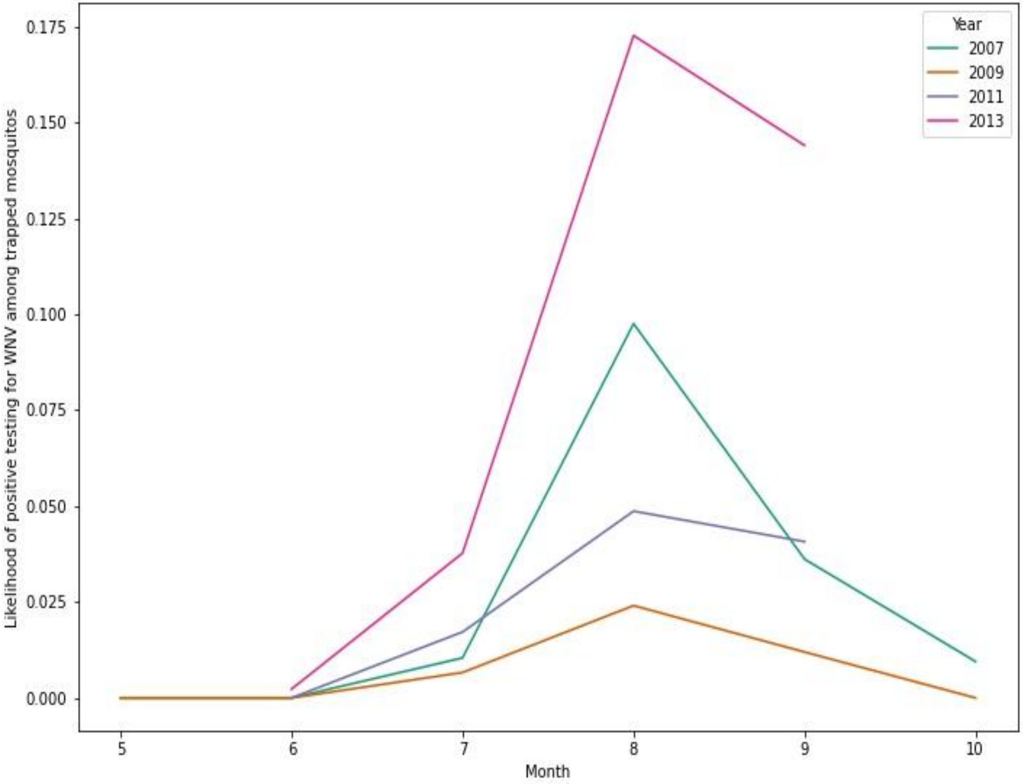
- GridSearchCv
- Confusion Matrix, ROC Curve

Lag Variable: a variable which has its value coming from an earlier point in time



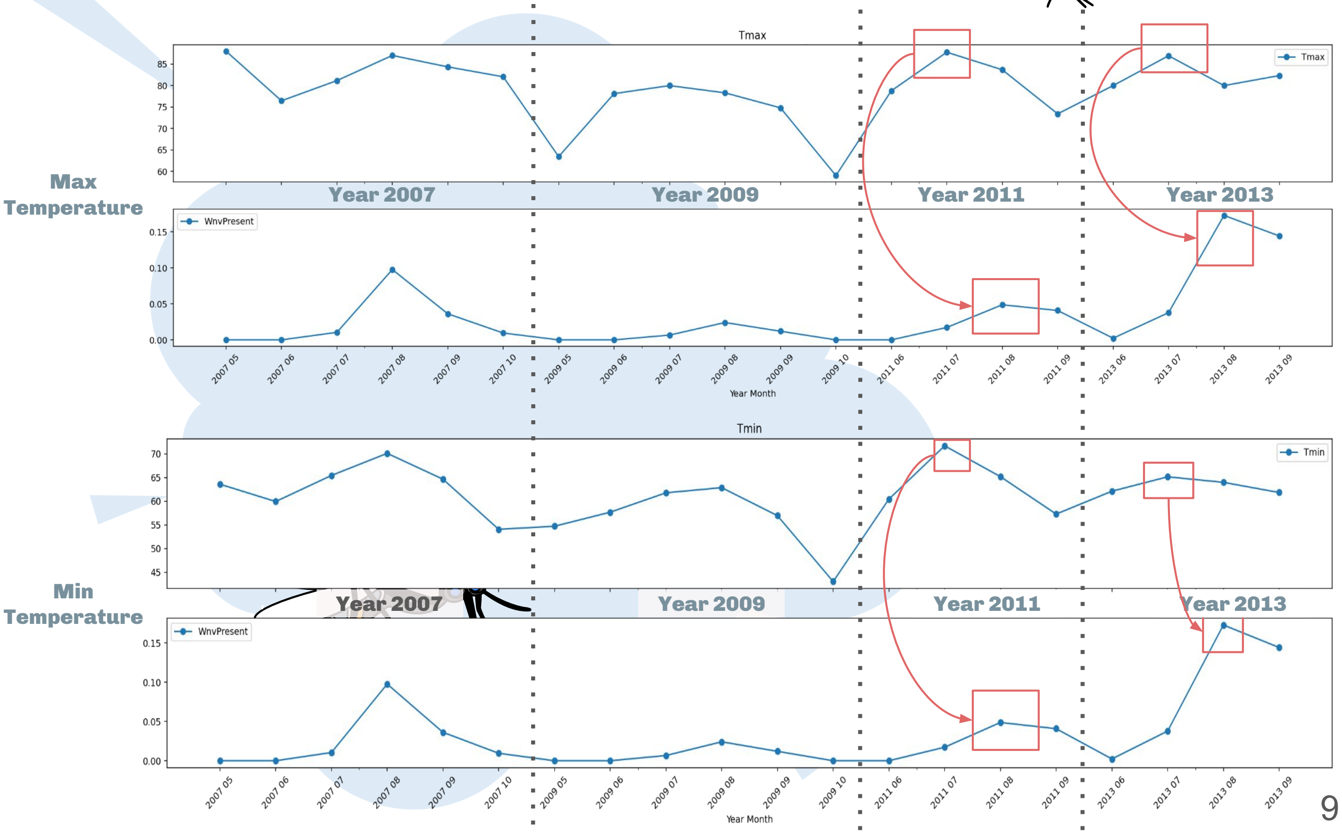
EDA: Time

Peak season for the West Nile Virus falls between **July** and **September**



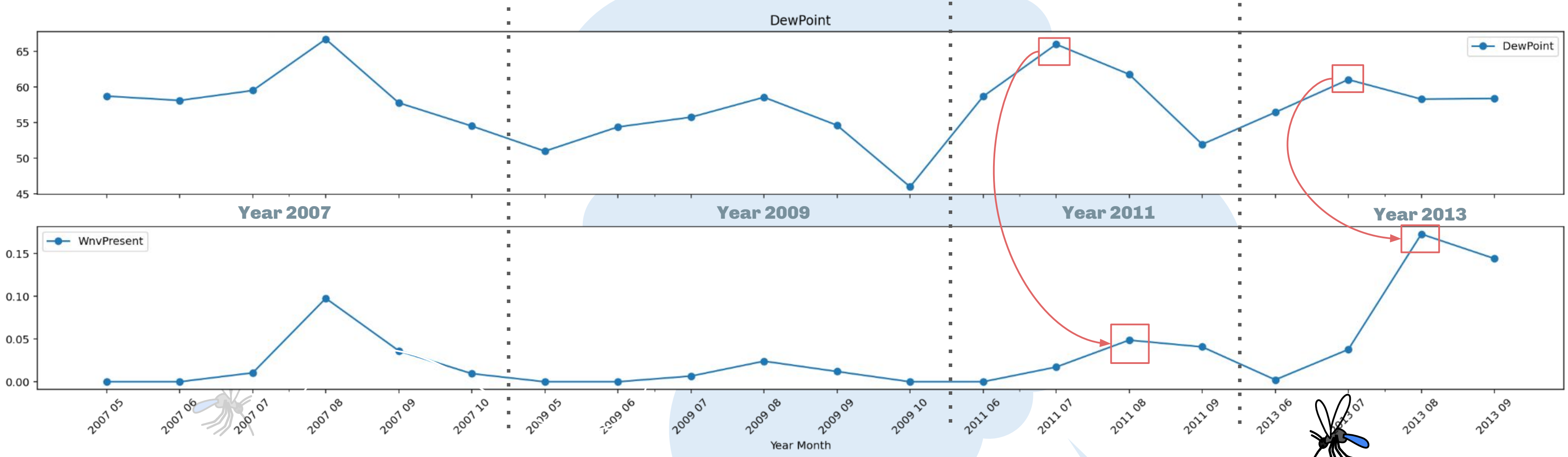
EDA: Temperature

The **higher** the temperature, the **higher** the occurence of virus



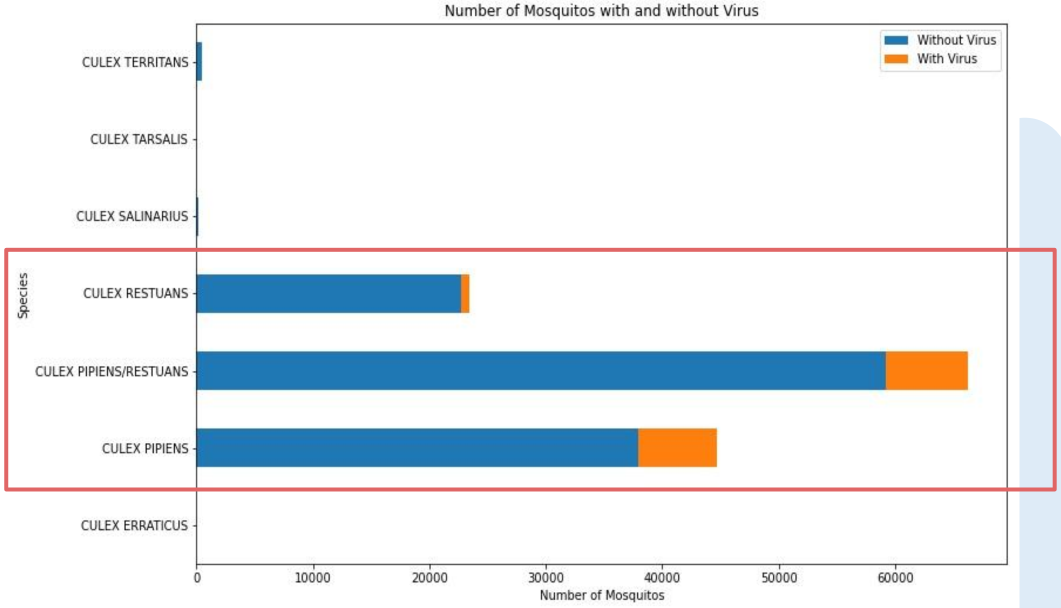
EDA: Humidity

Similarly, the **higher** the dew point, the **higher** the occurence of the virus



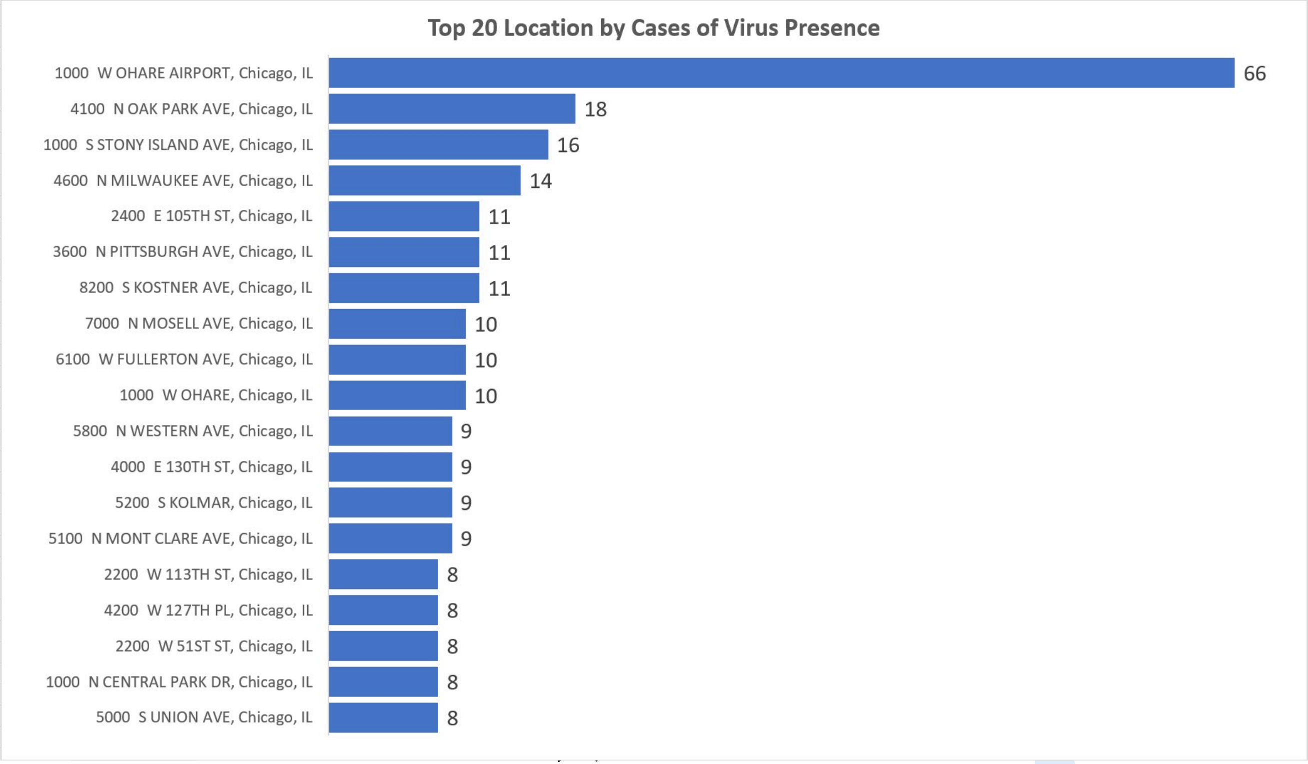
Dew point: The temperature to which air must be cooled to become saturated with water vapor. The measurement of the dew point is related to humidity. A higher dew point means there is more moisture in the air.

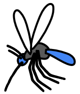
EDA: Species



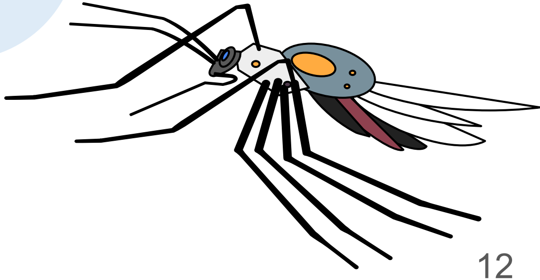
In Chicago, the virus seems to only be carried by 2 species: **Culex Restuans** & **Culex Pipiens**

EDA: Location



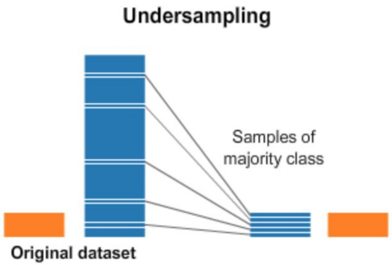


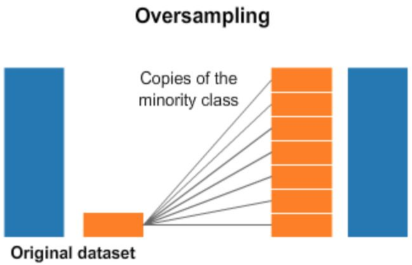
Occurence of West Nile Virus **varies** greatly by **location**



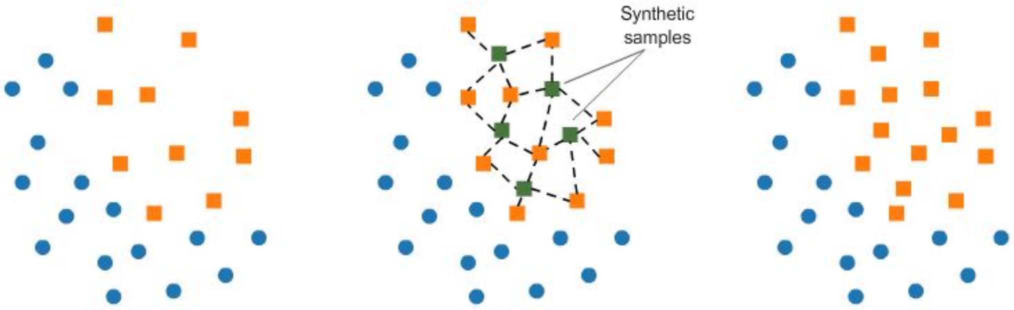
Model Preparation - Preprocessing steps

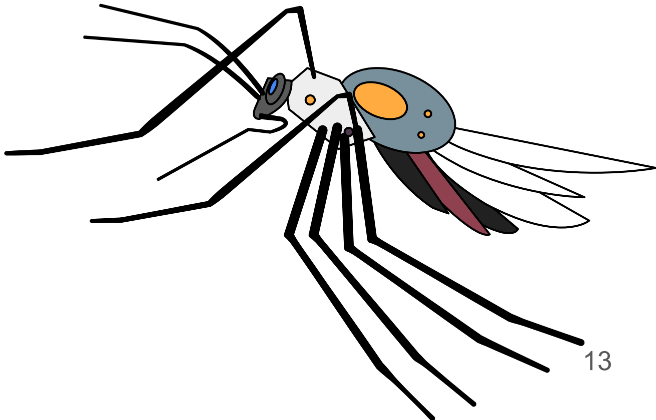
- Undersampling | Oversampling | SMOTE





- Selection - SMOTE

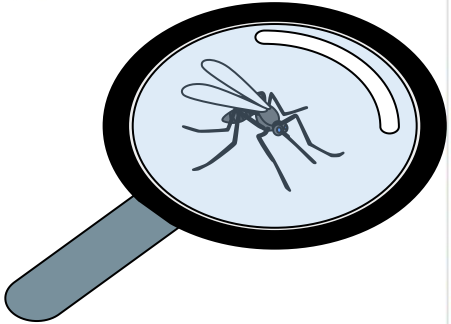




Model Selection & the Trade-off between Recall and Precision

```
Method Used: No sampling ------------------

Class Balance BEFORE

0    0.947087
1    0.052913
Name: WnvPresent, dtype: float64

Number of rows: 7295

Class Balance AFTER

0    0.947087
1    0.052913
Name: WnvPresent, dtype: float64

Number of rows: 7295
```

	model	train_auc_cv	f1	recall	precision	train_auc	test_auc	auc_diff
0	rf	0.767563	0.107143	0.072727	0.20339	0.943761	0.749228	0.194533
1	dt	0.734981	0.104265	0.066667	0.23913	0.944540	0.710449	0.234092
2	et	0.734896	0.102804	0.066667	0.22449	0.944540	0.708432	0.236108
3	lr	0.832778	0.000000	0.000000	0.00000	0.843867	0.813177	0.030690
4	gb	0.855704	0.000000	0.000000	0.00000	0.902317	0.848950	0.053367
5	ada	0.850752	0.000000	0.000000	0.00000	0.879293	0.839028	0.040265
6	svc	0.755371	0.000000	0.000000	0.00000	0.840362	0.747608	0.092754



```
Method Used: SMOTE sampling -------------------------------------------
-

Class Balance BEFORE

0.0    0.947087
1.0    0.052913
Name: WnvPresent, dtype: float64

Nu
Cl
0.
1.
Na
Nu
```

0								
1								
2								
3	rf	0.978472	0.249423	0.327273	0.201493	0.988925	0.748130	0.240795
4	et	0.975039	0.248244	0.321212	0.202290	0.989727	0.707890	0.281837
5	dt	0.973889	0.247086	0.321212	0.200758	0.989727	0.712486	0.277241
6	lr	0.858475	0.228873	0.787879	0.133883	0.861275	0.815409	0.045865

Model Selection Justification

Although the Gradient Boosting model has the strongest ROC AUC score, its recall score (0.527) pales in comparison to that of Adaboost (0.606).

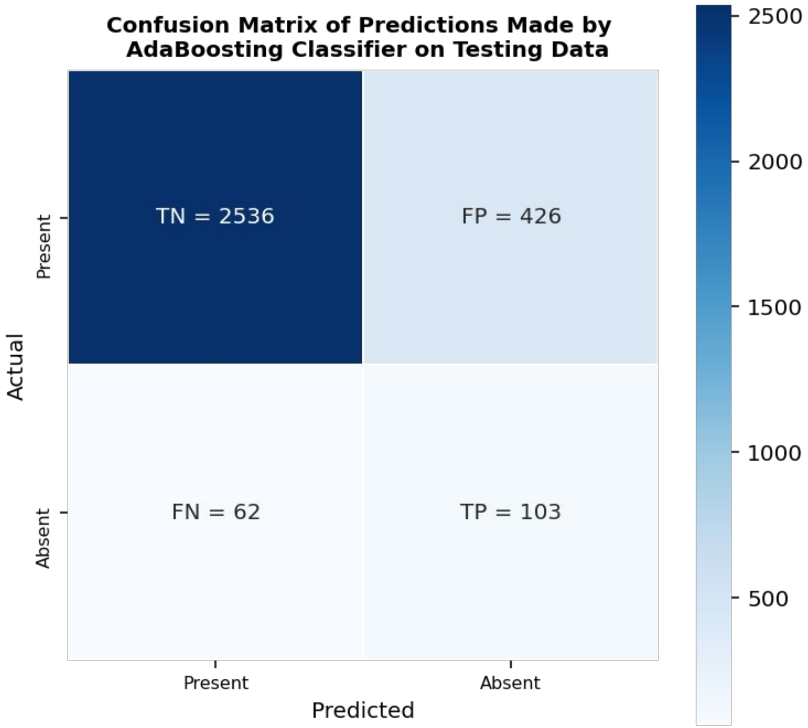
This means that we are likely to have **fewer False Negatives using Adaboost**.

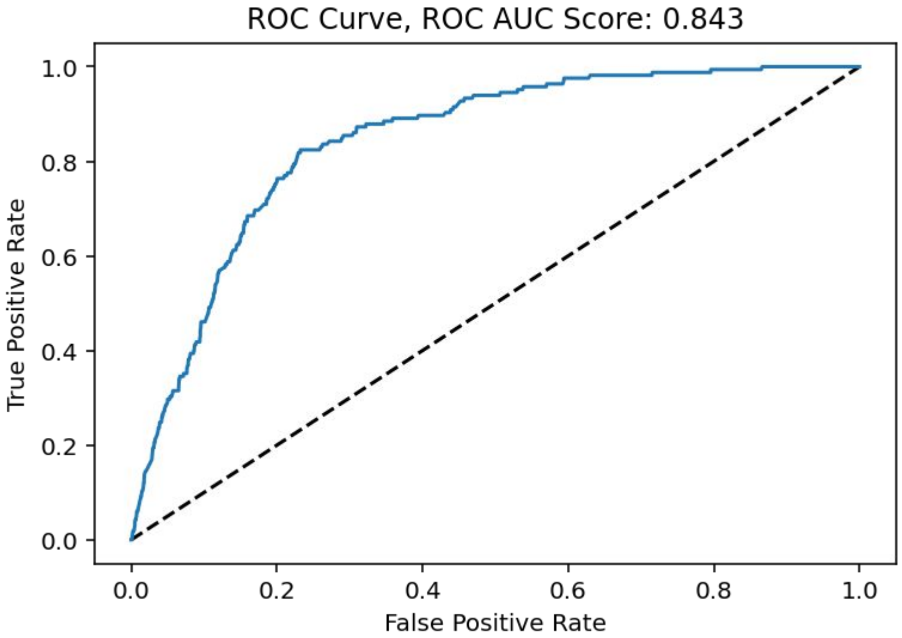
Support Vector Classification is also a possible consideration, but it fared worse in terms of the ROC AUC score and precision score compared to Adaboost.

AdaBoost seems like the best model for this use case as it is important to ensure a relatively high recall score that does not compromise the ROC AUC and/or precision score.

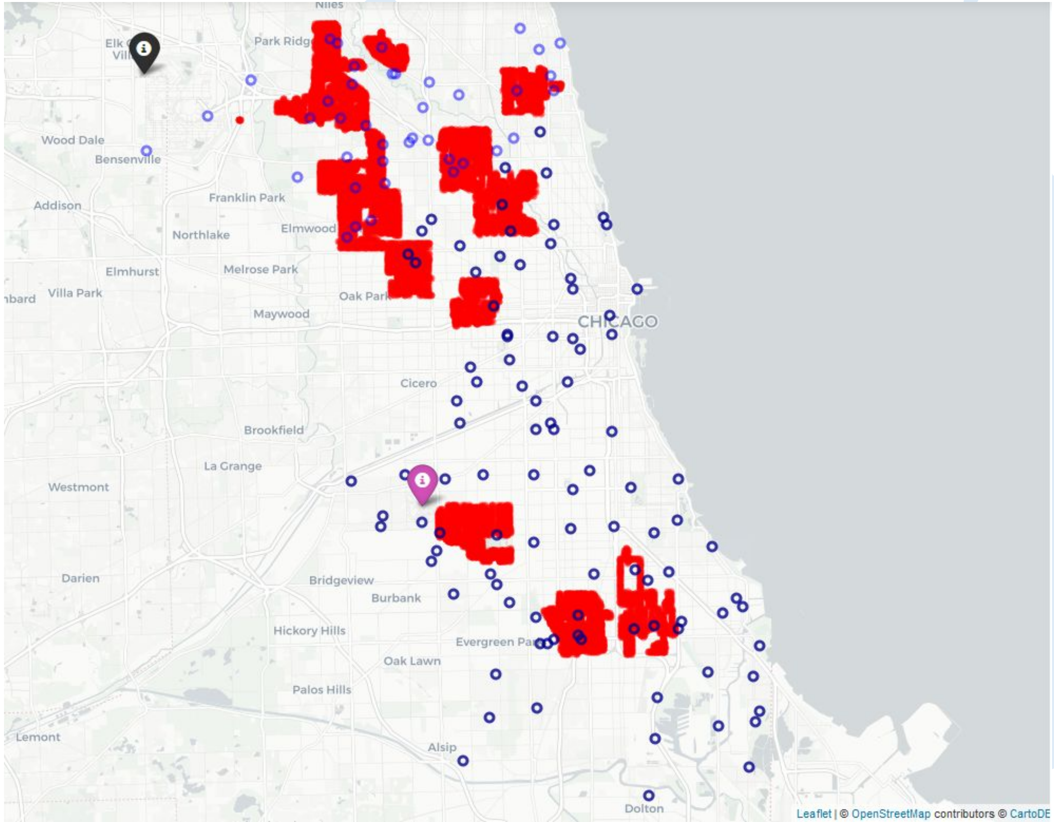
	model	train_auc_cv	f1	recall	precision	train_auc	test_auc	auc_diff
0	gb	0.976043	0.309609	0.527273	0.219144	0.978288	0.837721	0.140567
1	ada	0.962699	0.307220	0.606061	0.205761	0.963814	0.837294	0.126520
2	svc	0.955815	0.285714	0.636364	0.184211	0.962178	0.828141	0.134037

Model Evaluation





Spraying Effects



Visualization of sprayed areas and trap locations

Spraying Effects

Visualization of sprayed areas and WNV

Cost Benefit Analysis - Spraying

The Chicago Department of Public Health (CDPH) has been combatting WNV since 1999 through 2020.

They use an insecticide called Zenivex E4.

	Price (gallon)	Pounds AI/gallon	Price per Pound	Application Rate/Acre	Cost/Acre	Annual Acres Treated	Annual Cost
275 gal Zenivex® E20	$282.00*	1.48	$190.54	.0035	$0.67	20,000	$13,338
275 gal Zenivex® E4	$78.85*	.3	$262.83	.0035	$0.92	20,000	$18,398
2.5 gal Zenivex® E20	$296.00*	1.48	$200.00	.0035	$0.70	20,000	$14,000
2.5 gal Zenivex® E4	$80.75*	.3	$269.17	.0035	$0.94	20,000	$18,842

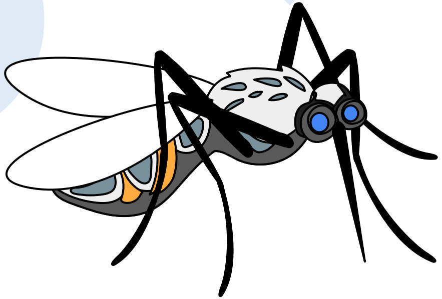
Cost Benefit Analysis - Spraying

Total land area size in Chicago = 145,545 acres

Cost of Zenivex per acre = $0.92

Cost of spraying the entirety of Chicago in a year:

$0.92 x 145,545 acres x 12 months = $1,606,816.80



Cost Benefit Analysis - Spraying

Our model predicted the WNV to be present in 149 unique traps.

Trap area = π $(700\text{ft})^2$ = 1,538,600sqft (35.32 acres)

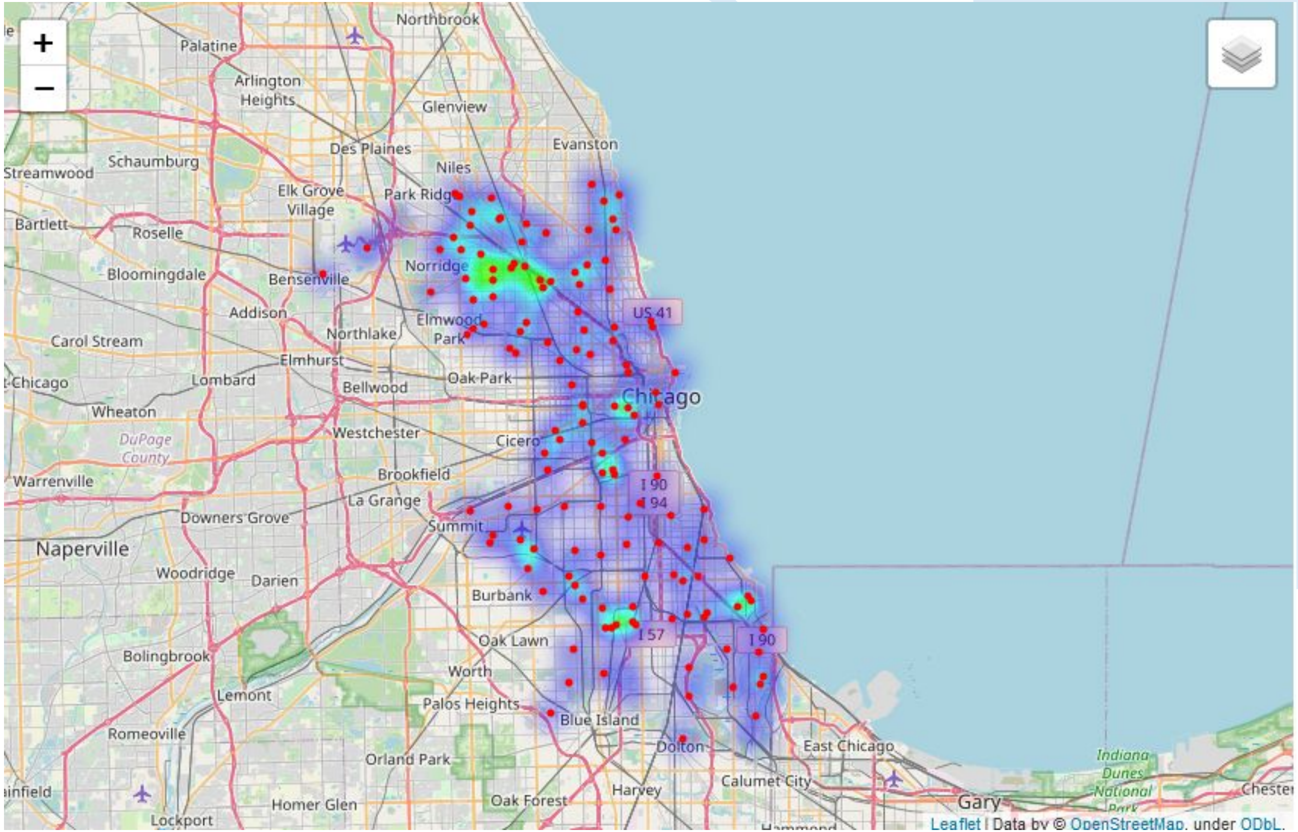
Trap cost = 35.32 x \$0.92 = \$32.50

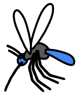
Total cost:

\$32.50 x 149 traps x 12 months = \$58,102.28

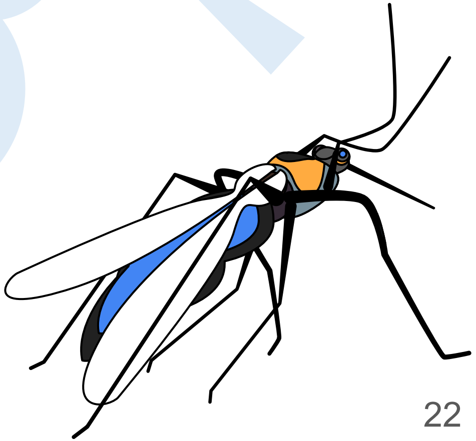
Cost savings on spraying = \$1,606,816.80 - \$58,102.28 = **\$1,548,714.51 (96%)**

Cost Benefit Analysis - Spraying





Visualization of model's prediction of WNV clusters



Cost Benefit Analysis - Hospitalization & Lost Productivity

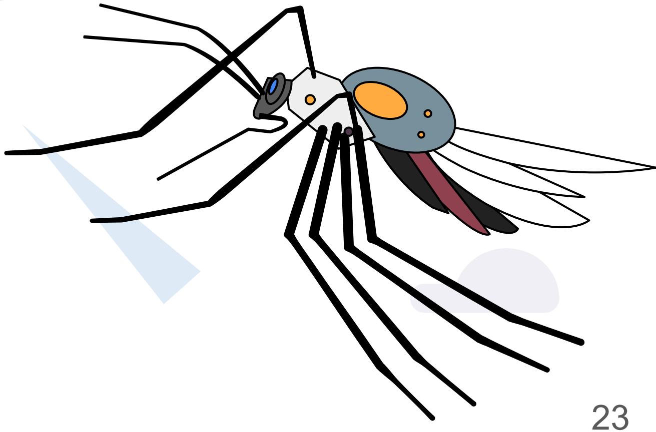
From 1999 through 2012, health care expenses and lost productivity in the US totalled up to $800 million.

4% died and 49% of the total cases were hospitalized.

In Chicago, the worst year in 2002 reported 225 cases.

A. Initial costs

	Fever $N = 18$	Meningitis $N = 19$	Encephalitis $N = 16$	AFP $N = 27$
Total inpatient hospital costs*				
Median (Range)	$4,467 (419–23,374)	$7,261 (337–13,633)	$15,136 (3,734–207,303)	$20,774 (5,066–264,176)
Mean (SD)	$6,955 (6,282)	$6,961 (3,300)	$27,020 (49,012)	$70,186 (80,133)
Total lost productivity*†				
Median (range)	$328 (92–2,729)	$682 (68–1,592)	$1,380 (113–307,871)	$2,136 (232–145,750)
Mean (SD)	$546 (659)	$684 (376)	$53,234 (97,583)	$12,357 (33,089)
*Total initial costs**				
Median (range)	$4,617 (538–24,010)	$7,942 (1,057–14,569)	$20,105 (3,965–324,167)	$25,117 (5,385–283,381)
Mean (SD)	$7,501 (6,762)	$7,644 (3,495)	$80,254 (104,785)	$82,542 (94,388)



Cost Benefit Analysis - Hospitalization & Lost Productivity

Estimated yearly hospitalization costs:

$7,500 x 225 = $1,687,500

Through our model, we are confident to predict 60% of the WNV cases (recall = 0.6), and thus we would be able to save **~$1,000,000**.

Conclusions & Recommendations

The final selected model was AdaBoost, with a test AUC of 0.837 and recall score of 0.606.

Our model was able to achieve significant cost-savings. However, the WNV prediction rate could be better. More data points would be helpful.

The cost analysis was over-simplified and not performed on a macro level. Further efforts beyond spraying and trapping could be explored. For instance, we can investigate if a neighbourhood’s proximity to nearby water bodies (e.g. ponds) can affect the incidence of West Nile Virus.

References

https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

https://www.calcunation.com/calculator/humidity-calculator.php

https://www.chicago.gov/city/en/depts/cdph/provdrs/healthy_communities/news/2020/august/city-to-spray-insecticide-thursday-to-kill-mosquitoes0.html

https://www.chicago.gov/content/dam/city/depts/cdph/Mosquito-Borne-Diseases/Zenivex.pdf

https://www.callnorthwest.com/2019/05/how-long-does-a-mosquito-treatment-last/

https://datasmart.ash.harvard.edu/news/article/predictive-analytics-guides-west-nile-virus-control-efforts-in-chicago-1152

https://www.cdc.gov/mosquitoes/about/life-cycles/culex.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

https://www.cityofchicago.org/dam/city/depts/cdph/comm_dis/general/Communicable_Disease/CD_CDInfo_Jun07_WNV.pdf