Predicting West Nile Virus Presence in Mosquitos across the City of Chicago

DSI-SG-27

Yuet Meng
Rashidi
Tiancheng

BACKGROUND

We are a group of data scientists working at Disease And Treatment Agency

Our task is to predict occurrences of the WNV using the dataset provided by the Chicago Department of Public Health (CDPH), and to evaluate the cost and benefit of spraying.

RESEARCH

- West Nile Virus doesn't hop between humans and is only transmitted when specific mosquito species bite an infected bird and then a human.

- The period of greatest concern is around summer through fall especially from June to November where WNV can spread rampantly.

DATA CLEANING

- For the Train dataset, we are provided with observations from years 2007, 2009, 2011, and 2013.

- We are required to predict the observations from years 2008, 2010, 2012, and 2014 in the test dataset. Observations start around May till early October for each year

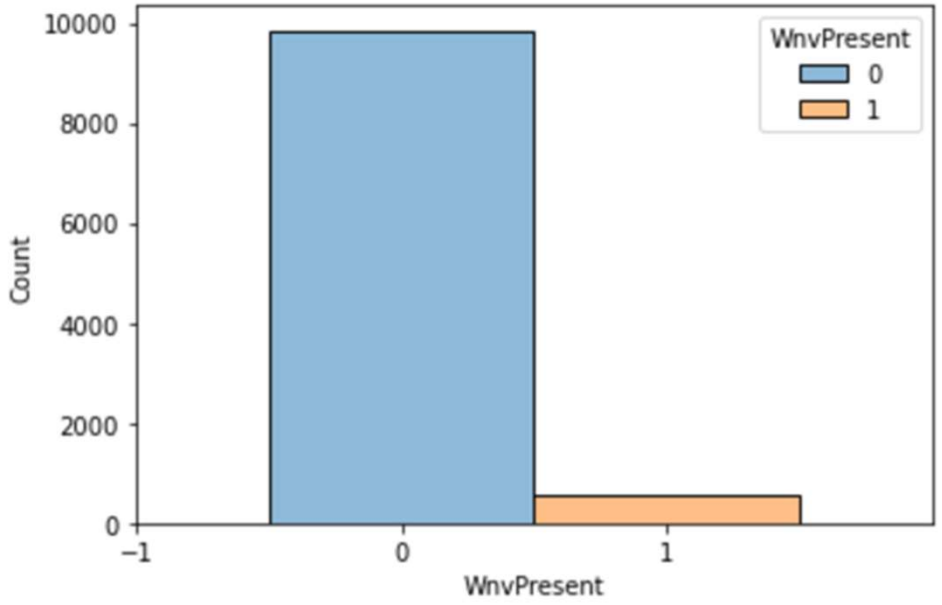
Train Dataset/Test Dataset
- Drop rows with zero number of mosquito observations
- Converting the Date field to DateTime type.
- Drop the duplicated rows

Weather Dataset
- Imputed missing values with values from alternate stations if available or 0
- Converting the Data field to DateTime type
- Convert numeric fields to float types
- Converting the Date field to DateTime type

EXPLORATORY DATA ANALYSIS

Overall Distribution of Negative and Positive Class for WnvPresent

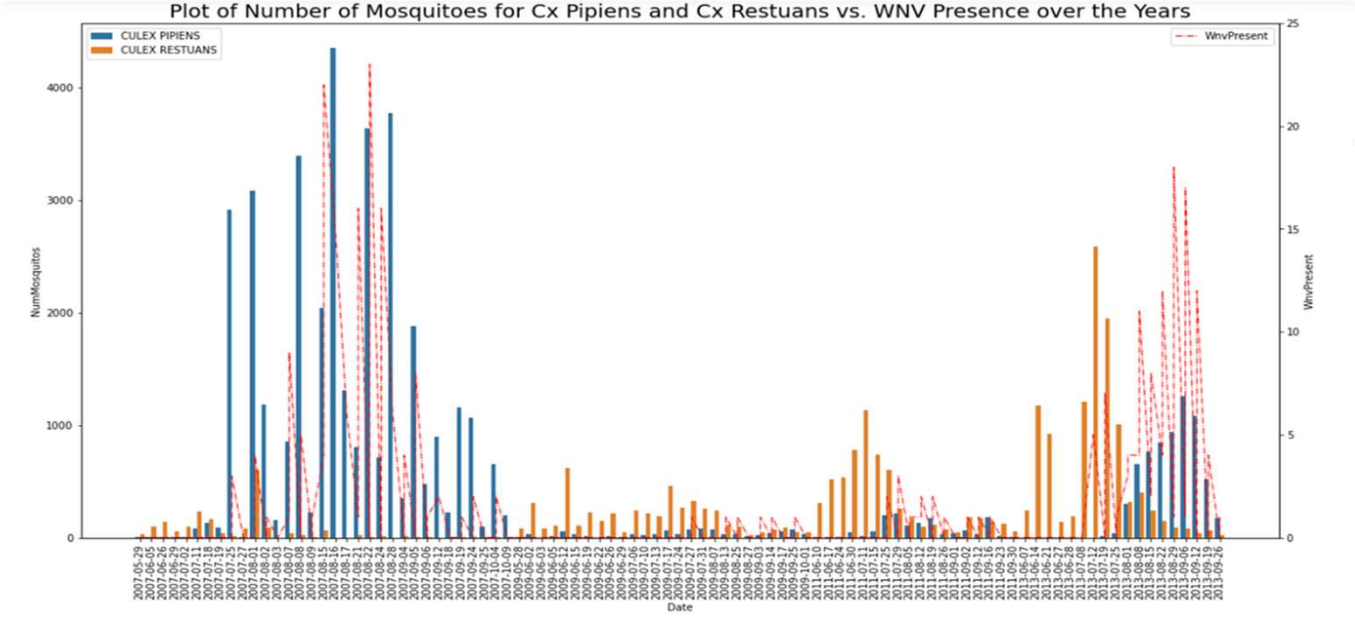
```
0    0.947115
1    0.052885
Name: WnvPresent, dtype: float64
```



We observed that only 5.28% consists of the positive class. Implies imbalance of classes

EXPLORATORY DATA ANALYSIS

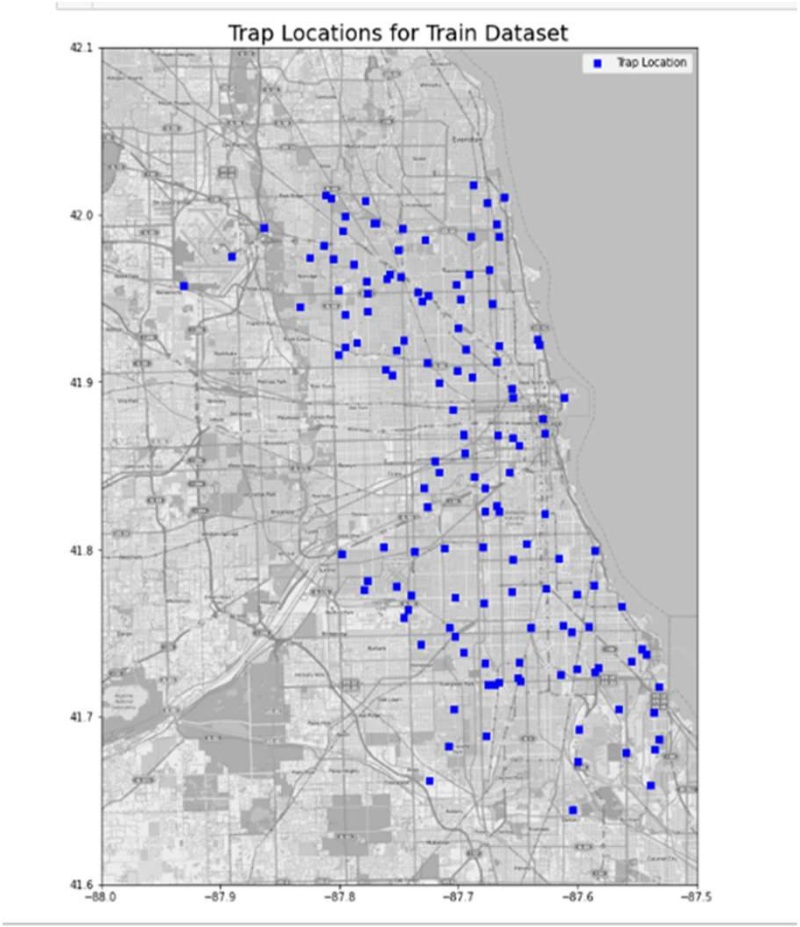
Species	NumMosquitos	WnvPresent
CULEX ERRATICUS	7	0
CULEX PIPIENS	44488	240
CULEX PIPIENS/RESTUANS	65841	261
CULEX RESTUANS	23326	49
CULEX SALINARIUS	144	0
CULEX TARSALIS	7	0
CULEX TERRITANS	508	0



- `CULEX PIPIENS` and `CULEX RESTUANS` are the only two species that were observed having WNV presence
- Number of WNV present observations coincides with the spikes in the number of `CULEX PIPIENS`.
- `CULEX PIPIENS` is a more prominent carrier for the virus`

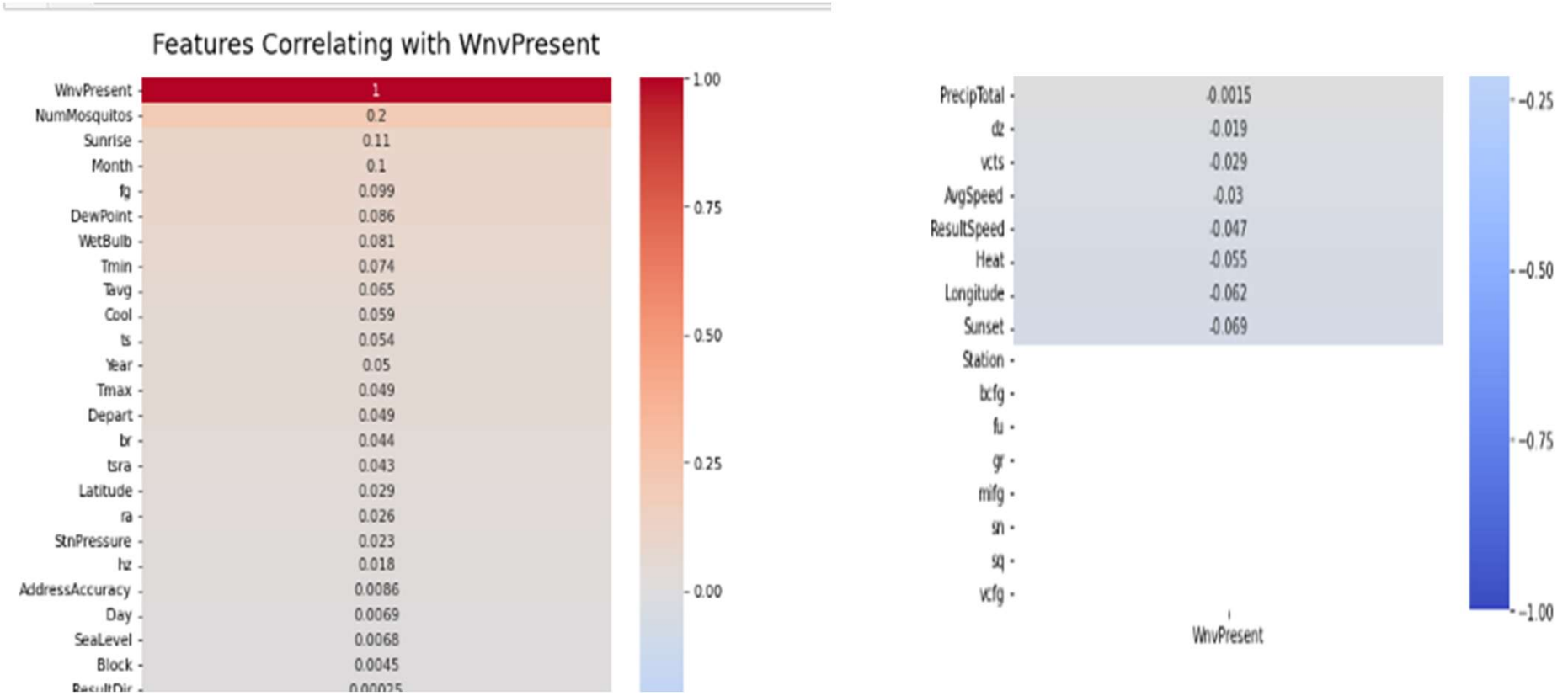
`

EXPLORATORY DATA ANALYSIS



. Based on the plot of trap locations, there are certain trap locations that are in close proximity to each other and thus, may be correlated due to geographical proximity.

EXPLORATORY DATA ANALYSIS



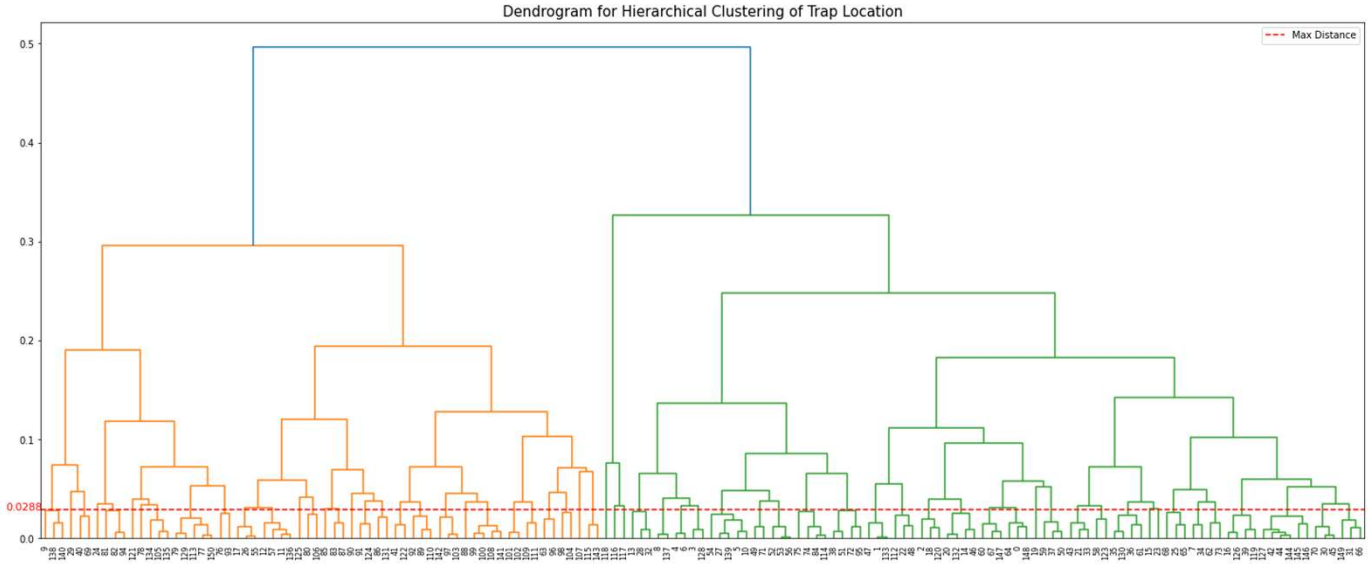
Correlation of weather features with respect to WNV presence is not strong.

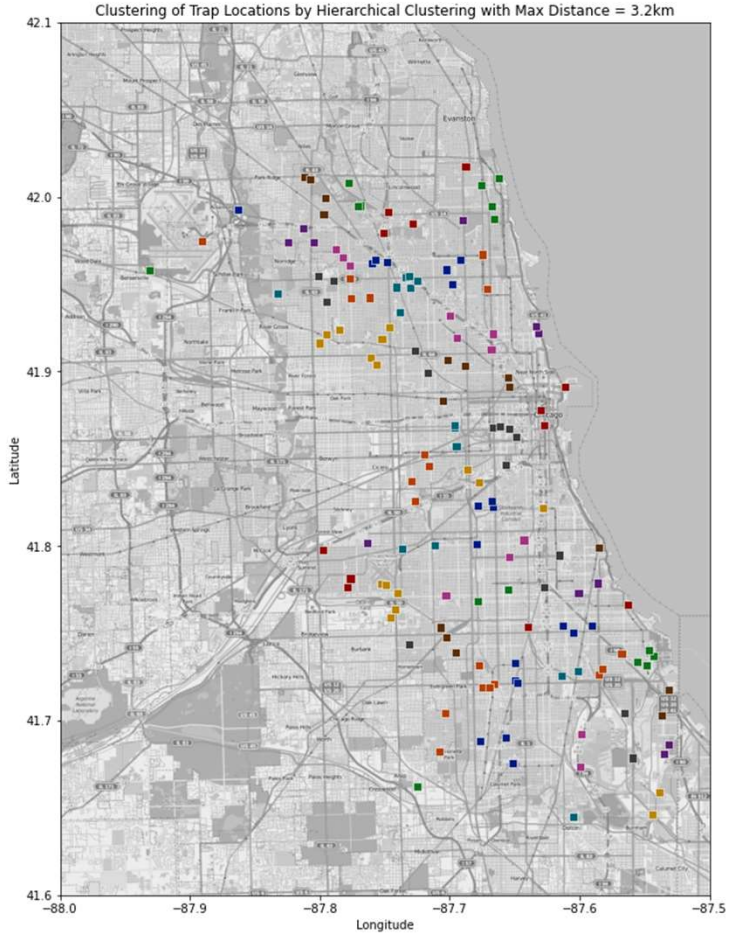
FEATURE ENGINEERING

Train/Test Dataset

1. Added Row_count to address multi-row leakage
   a. To quantify multiple rows if NumMosquito > 50 observed for a given ‘Date’, ‘Address’, and ‘Species’

1. Clustering of trap locations using Hierarchical Cluster Analysis (HCA)
   a. Max distance (cutoff) set based on the flight range for an adult Culex mosquito (3.2km)





FEATURE ENGINEERING

Weather Dataset

1. Dropped one of the features in pairs with >0.95 pairwise positive correlation
   a. Chose features with stronger correlation to WnvPresent

1. Added a 2-day moving average for Weather
   a. Since sexual activity in C. pipiens begins within the first 2–3 days of emergence from the larval development stage

	v1	v2	pair_corr	v1_y_corr	v2_y_corr
0	DewPoint	WetBulb	0.973423	0.085973	0.080827
1	Sunrise	Sunset	0.959904	0.105668	-0.068704
2	Tavg	WetBulb	0.952375	0.064880	0.080827
3	Tmax	Tavg	0.950623	0.049059	0.064880
4	Tavg	Cool	0.950352	0.064880	0.058690
5	Tmin	Tavg	0.938147	0.074372	0.064880
6	Tmin	WetBulb	0.937047	0.074372	0.080827
7	WetBulb	Cool	0.914281	0.080827	0.058690
8	Tmin	Cool	0.904038	0.074372	0.058690
9	ResultSpeed	AvgSpeed	0.901876	-0.046928	-0.030380

MODEL SELECTION

Model		HyperParameter	Sensitivity	Accuracy	Specificity	Precision	F1 Score	Train Score	Test Score	ROC AUC
Base Model	Logistic Regression	Default	0	0.947	1	NaN	0	0.947	0.947	0.79
Model 1	Logistic Regression SMOTE	SMOTE_random_state=100 Logreg_random_state=100 Logreg_solver='liblinear', Logreg__C=1, Logreg_penalty=11, SMOTE__k_neighbors=13, SMOTE__sampling_strategy: auto	0.655	0.758	0.768	0.134	0.225	0.76	0.758	0.78
Model 2	Random Forest SMOTE	'rf__max_depth': None, 'rf__n_estimators': 200, 'sampling__k_neighbors': 13, 'sampling__sampling_strategy': 'minority	0.158	0.929	0.972	0.241	0.19	0.929	0.929	0.78
Model 3	SVM SMOTE	sampling__k_neighbors=13 sampling__sampling_strategy='a uto'	0.503	0.85	0.87	0.177	0.261	0.847	0.85	0.78
Model 4	Xgboost SMOTE	'sampling__k_neighbors': 13, 'sampling__sampling_strategy': 'auto', 'xgb__learning_rate': 0.5, 'xgb__max_depth': None, 'xgb__n_estimators': 50	0.176	0.936	0.978	0.311	0.224	0.933	0.936	0.84

MODEL HYPERPARAMETER TUNING

`scale_pos_weight` [default=1]

Control the balance of positive and negative weights, useful for unbalanced classes.

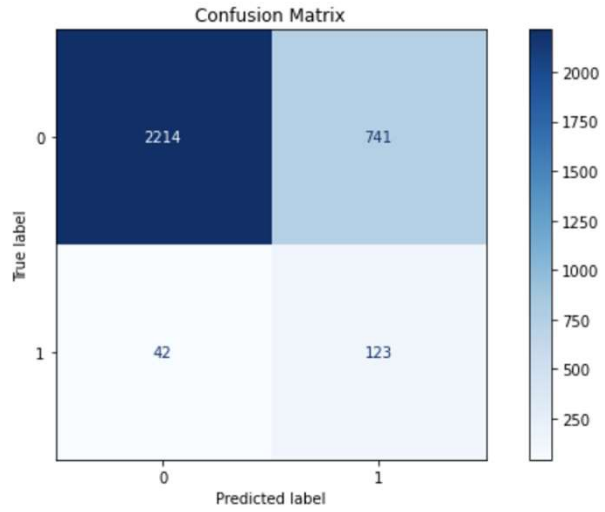
A typical value to consider: `sum(negative instances) / sum(positive instances)`.

Objectives:

- Increase Sensitivity
- Reduce False Negatives

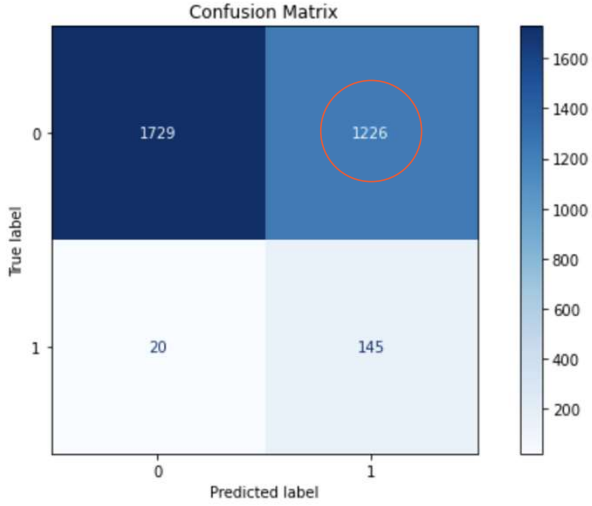
`scale_pos_weight` = 19

```
Sensitivity: 0.7454545454545455
Accuracy: 0.7490384615384615
Specificity: 0.749238578680203
Precision: 0.1423611111111111
f1 score: 0.239067055393586
ROC AUC score: 0.832924165512998
```



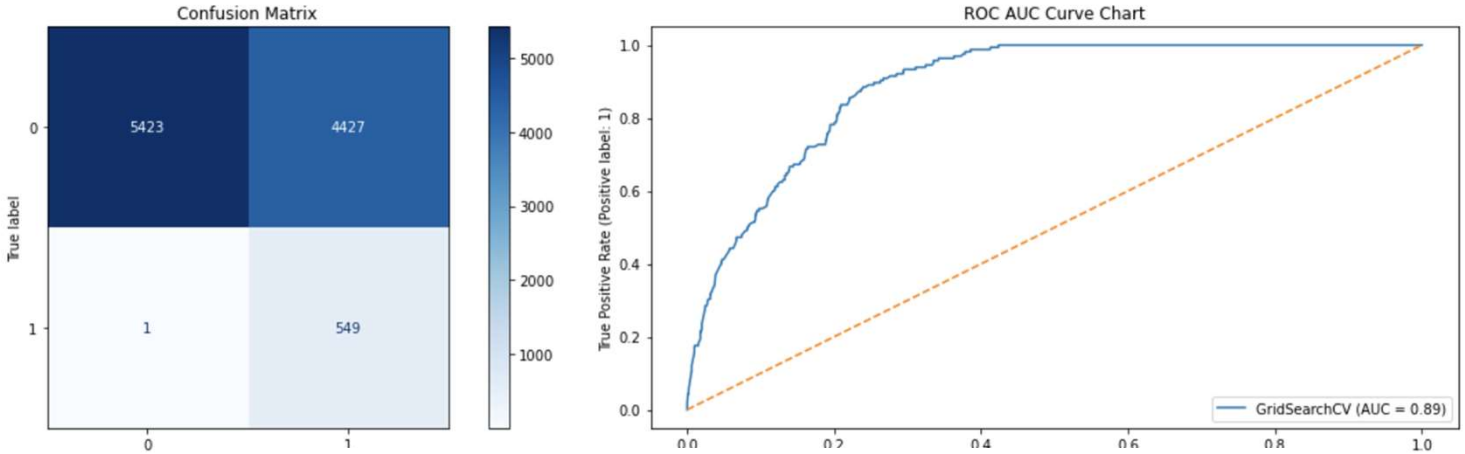
`scale_pos_weight` = 63

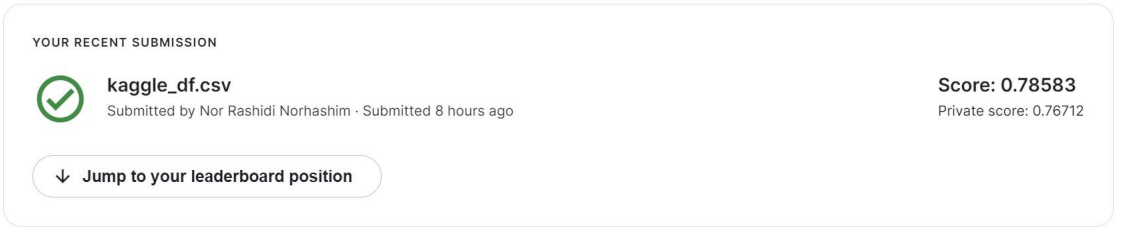
```
Sensitivity: 0.8787878787878788
Accuracy: 0.6006410256410256
Specificity: 0.5851099830795262
Precision: 0.10576221735959154
f1 score: 0.18880208333333334
ROC AUC score: 0.8298046454391632
```



PRODUCTION MODEL

Model		HyperParameter	Sensitivity	Accuracy	Specificity	Precision	F1 Score	Train Score	ROC AUC
Production Model	Xgboost	xgb__gamma=1, xgb__learning_rate=0.1, xgb__max_depth=5, xgb__min_child_weight=25, xgb__n_estimators=60, xgb__scale_pos_weight=63, xgb__subsample=1	0.998	0.574	0.551	0.110	0.198	0.427	0.89





FEATURE ANALYSIS

	Features	Importances
69	Sunrise	0.140950
10	cluster_9	0.032977
71	StnPressure	0.030755
75	AvgSpeed	0.029902
0	Species_CULEX PIPIENS	0.027692
11	cluster_10	0.027684
65	Tavg	0.027466
67	DewPoint	0.027369
26	cluster_25	0.025914
70	PrecipTotal	0.025743
73	ResultSpeed	0.025019
64	Tmin	0.024436
2	Row_count	0.023839
37	cluster_36	0.021818
1	Species_CULEX RESTUANS	0.021616

Train Dataset Specific Features

- Feature Engineered features such as 'cluster' and 'Row_count' appearing
- 'Species_CULEX PIPIENS' and 'Species_CULEX RESTUANS'
    a. due to its relatively higher correlation to WnvPresent

Weather Specific Features

- 'Sunrise' is the feature of highest importance.
    a. Good indicator of seasonality.
- 'StnPressure', drop in atmospheric pressure is associated to inclement weather.
- 'AvgSpeed' and 'ResultSpeed', in relation to wind speed, by increasing flight distance/spread of mosquitoes.
- 'PrecipTotal' is directly related to rainfall.
    a. High rainfall can affect flight speed and patterns.
    b. Low rainfall promotes breeding as standing water becomes richer in nutrients required for their larvae survival and proliferation.
- Temperature related feature such as 'DewPoint', 'Tavg', 'Tmin'.
    a. Generally, when 'DewPoint', 'Tavg', 'Tmin' increases, the number of mosquitoes increases and in turn, WNV observations increases.

COST BENEFIT ANALYSIS (OVERALL FINDINGS)

- Net benefit from spraying from model is $5,560,469 annually.
- Return on investment of 604%, an indicator we use to model whether it was with economical sense.

- Why the big number?
  - Mostly attributable to benefit of prevention of death.
  - Cost of each death around 294k, large number as it calculates total lifetime lost productivity caused by deaths from Staples et al (2014)
  - There were 22 fatalities reported in Chicago in 2002.

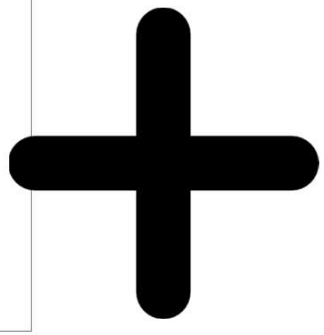
BENEFITS (ASSUMPTIONS AND INPUTS)

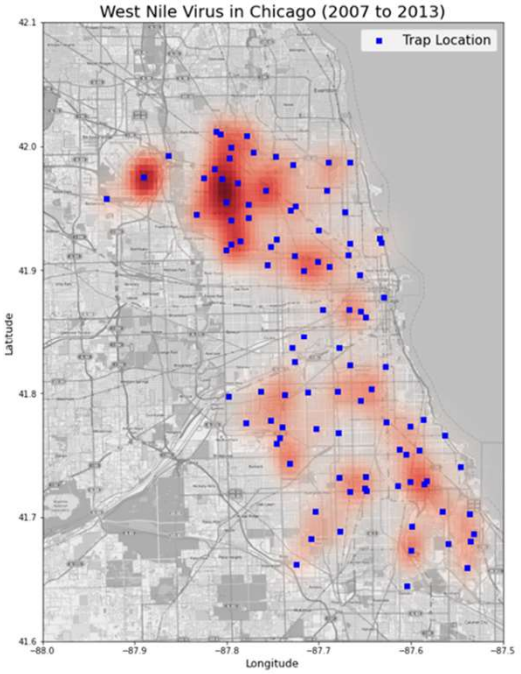
- Assume max benefit of preventing an epidemic in Chicago of the likes of 2002, which we model as the worst case scenario.
- Assume that 225 cases in 2002 were all the cases there were in Chicago.
- Assume that 1 in 5 rate of West Nile Fever and 1 in 150 rate of West Nile neuroinvasive disease stands.
- Actual deaths of 22 happened during epidemic.

SN	Item	Input	Basis
1	Number of cases in an epidemic	225	Based on 2002 Chicago real number of WNV cases
2	Number of WNF cases	45	Assume 1 in 5 rate as suggested by CDC
3	Number of WNND cases	2	Assume 1 in 150 rate as suggested by CDC
4	Number of deaths	22	Based on actual deaths reported in 2002 for Chicago

COST (ASSUMPTIONS AND INPUTS)

- Assume max cost of spraying the entirety of Chicago land area
- Using Zenivex, the chosen adulticide
- Assume that 11 sprays are made between July to September





COST BENEFIT ANALYSIS

Annual estimates	Benefit/(Cost) per number	Number	Benefit/(Cost)	Assumptions
Costs				
Cost of spray			(1,104,026)	1) Zenivex used at rate of 0.67 cents per acre, 2) Land area of Chicago 149,800 acres, 3) 11 city wide sprays in total
Benefits				
Reduction of economic cost of West Nile fever (WNF)	1,170	45	52,630	Barber et al (2010)
Reduction of economic cost of West Nile neuroinvasive disease (WNND)	10,539	2	21,078	Barber et al (2010)
Reduction of medical cost of West Nile neuroinvasive disease (WNND)	61,833	2	123,667	Barber et al (2010)
Reduction of cost of death	293,960	22	6,467,119	22 deaths as reported in 2002 Chicago epidemic, 2005 cost of death from Staples et al (2014)
		Subtotal for benefit	6,664,495	
		Net benefit/(cost)	5,560,469	
		Return on Investment	604%	

- Used 2 research papers to justify the benefit numbers

BREAKEVEN POINT

Factors	Number of each factor to breakeven	Likelihood
WNF cases	944	More likely for other cases to happen
WNND cases	15	Possible
Death cases	4	22 happened before. Likely

Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005

Loren M. Barber, Jerome J. Schleier, III, and Robert K.D. Peterson

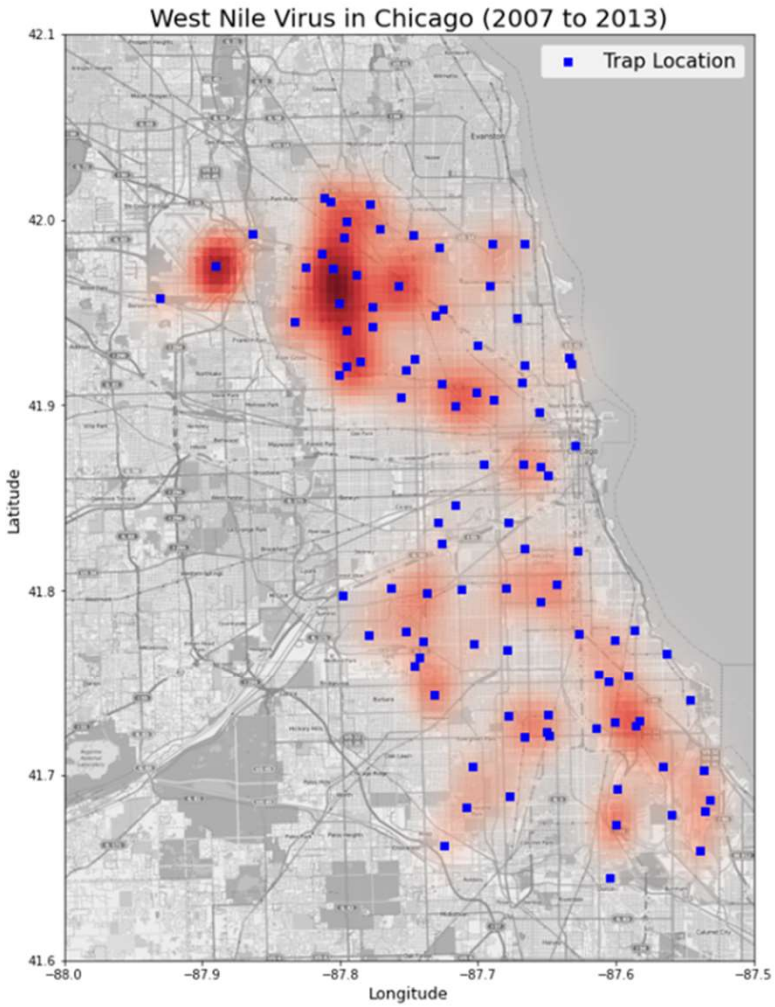
Abstract

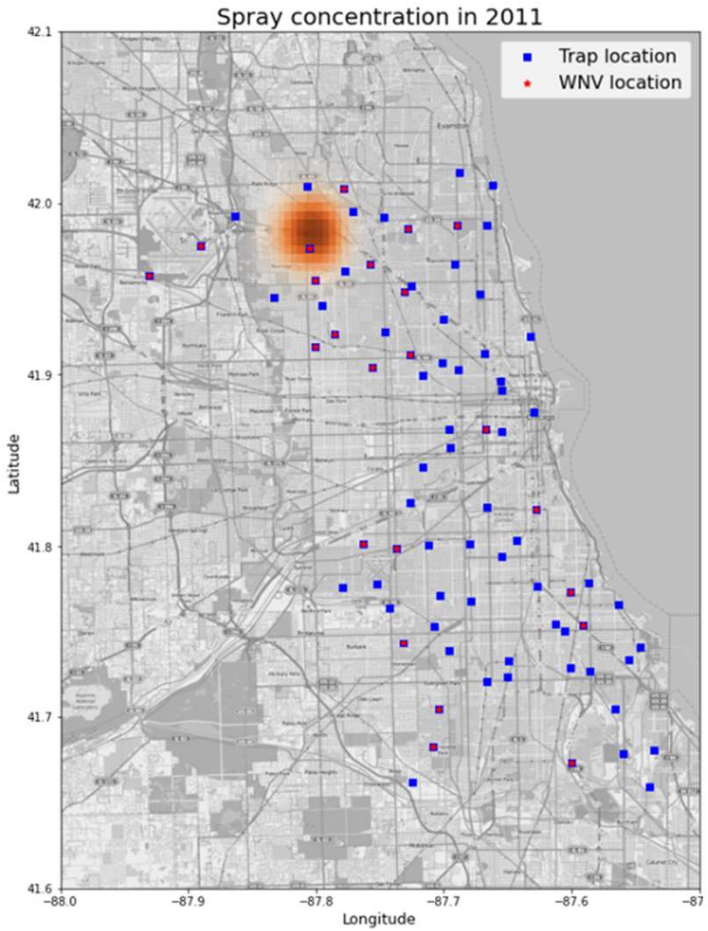
Go to:

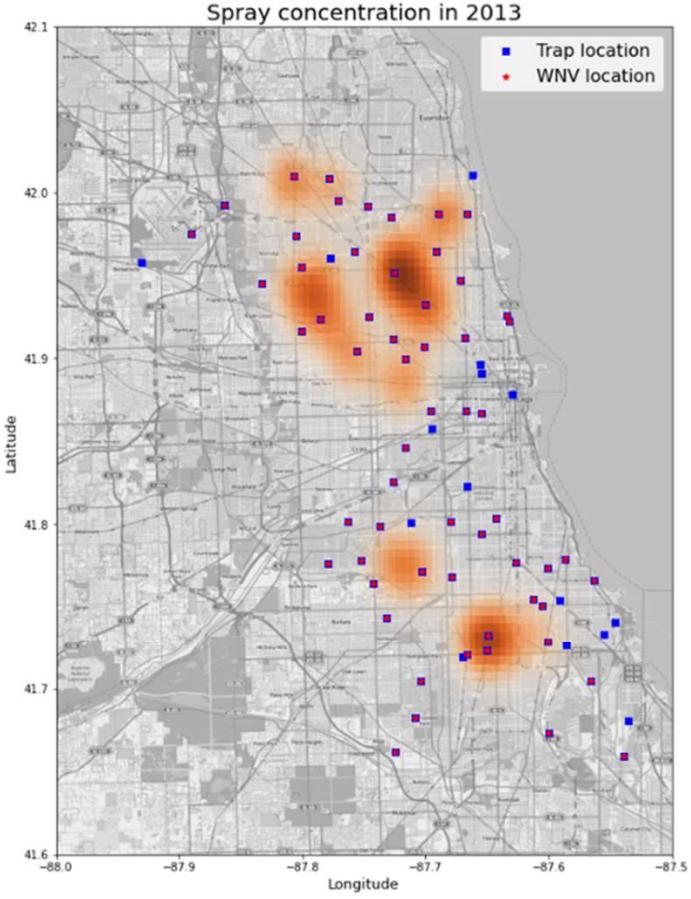
In 2005, an outbreak of West Nile virus (WNV) disease occurred in Sacramento County, California; 163 human cases were reported. In response to WNV surveillance indicating increased WNV activity, the Sacramento-Yolo Mosquito and Vector Control District conducted an emergency aerial spray. We determined the economic impact of the outbreak, including the vector control event and the medical cost to treat WNV disease. WNV disease in Sacramento County cost ≈$2.28 million for medical treatment and patients' productivity loss for both West Nile fever and West Nile neuroinvasive disease. Vector control cost ≈$701,790, including spray procedures and overtime hours. The total economic impact of WNV was $2.98 million. A cost-benefit analysis indicated that only 15 cases of West Nile neuroinvasive disease would need to be prevented to make the emergency spray cost-effective.

- Current cost benefit analysis is sound - cross checked against research paper (Barber et al 2010).
- Just need 15 to have WNND, or 4 to die, and breakeven scenario occurs for spraying.
- Low bar to pass in the justification of spraying.

WHAT HAPPENS WITH MORE TARGETED SPRAYING?







SENSITIVITY ANALYSIS WHEN TARGETING TOP 20 CLUSTERS

				Efficacies in reduction of WNV					
Annual estimates	**Benefit/(Cost) per number**	**Number**	**Benefit/(Cost)**	**20%**	**33.33%**	**50%**	**70%**	**100%**	**Assumptions**
Costs									
Cost of spray								(368,009)	1) Zenivex used at rate of 67 cents per acre, 2) Land area of Chicago 149,800 acres, spray 20/60 clusters 3) 11 city wide sprays in total
Benefits									
Reduction of economic cost of West Nile fever (WNF)	1,170	45	52,630	10,526	17,543	26,315	36,841	52,630	Barber et al (2010)
Reduction of economic cost of West Nile neuroinvasive disease (WNND)	10,539	2	21,078	4,216	7,026	10,539	14,755	21,078	Barber et al (2010)
Reduction of medical cost of West Nile neuroinvasive disease (WNND)	61,833	2	123,667	24,733	41,222	61,833	86,567	123,667	Barber et al (2010)
Reduction of cost of death	293,960	22	6,664,495	1,332,899	2,221,498	3,332,248	4,665,147	6,664,495	22 deaths as reported in 2002 Chicago epidemic, 2005 cost of death from Staples et al (2014)
		Subtotal for benefit		1,372,374	2,287,290	3,430,935	4,803,309	6,861,870	
		Net benefit/(cost)		1,004,365	1,919,281	3,062,926	4,435,300	6,493,862	
					Likely range				
		Return on Investment		373%	**622%**	**932%**	1205%	1865%	

- Potential increase in ROI above 600% (622 to 932% likely)

RECOMMENDATION

•Spraying is a pillar of good vector control. However, big positive number in model is not a clear mandate/ not a blank cheque to support indiscriminate spraying.

•A multi-pronged approach for vector control can bring further cost savings and better returns on spending.

1) Citizen involvement and education

•Avoid outdoors during time between dawn and dusk.

•Apply insect repellent when outdoors and wear covered clothing.

•Close windows and doors when indoors.

•Remove stale water.

2) Use of larvicide

3) Surveillance and testing for WNV

•eg. O'Hare International Airport and South Doty Avenue can be focused. More economies as seen from ROI

AREAS OF FUTURE WORK

Explore a leading weather indicator to know where to spray up to 2 weeks from today's date.

- Allows preemptive action to plan for sprays in high risked areas.