West Nile Virus Prediction

Alicia, Jonathan, Mok, Zaini

Agenda

1. Background
2. Problem Statement
3. Datasets, Feature Engineering, EDA
4. Metric for Modelling
5. Modelling Process
6. Findings
7. Cost Benefit Analysis
8. Recommendations

Background

What's the story?

- WNV was first identified in the US in 2001.
- The following year, Illinois recorded 884 human cases, 67 more than in any other state.

- The battle against the West Nile virus is an annual affair Chicago grapples with:
    - Infection can be asymptomatic or symptomatic in humans, with a 4:1 ratio
    - Can be mild, resulting in flu-like symptoms (West Nile fever **[WNF]**), or severe, causing paralysis and even death(West Nile neuroinvasive disease **[WNND]**)

Background

Who are we?

- Team of Data Scientists at Disease And Treatment Agency

What are we doing?

- Fight the spread through **prediction** of future clusters

- **Collect and analyze data** such as weather conditions, species and population of mosquitos and location data **that may affect WNV propagation**

Problem Statement

1. **Predicting presence** of West Nile Virus

2. What are the **main factors** that may **result in the virus being spread**?

3. What **cost-effective methods** should we adopt to **prevent WNV mosquitoes from breeding**?

Datasets

1. Train & Test Dataset
2. Weather Dataset
3. Spray Dataset

Train and Test Dataset

- Train dataset consists data from 2007, 2009, 2011, 2013
    - Date
    - Species and Sample size of all mosquitos
    - Presence of WNV
    - Trap ID
    - Location details of Trap

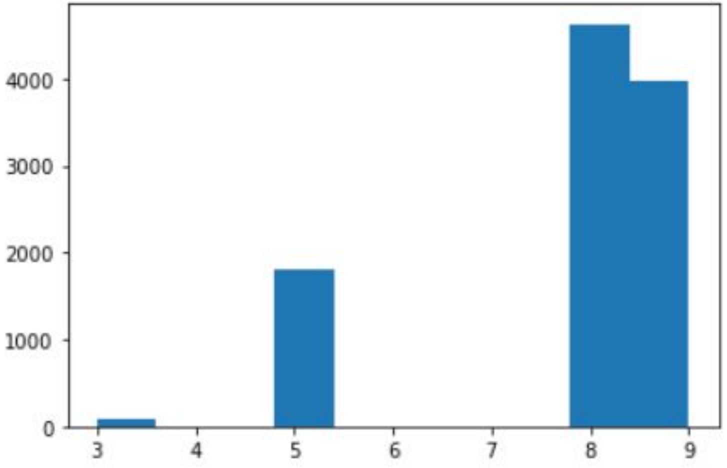
- Test dataset requires us to predict presence of WNV for 2008, 2010, 2012, 2014

Feature Selection/Engineering

- Address accuracy is mostly 8 and 9
- Accuracy score of 5 is accurate to: **“Center of the ZIP or postal code area”**

- Drop all address details other than Longitude and Latitude
- Data points with **same date and location** were **grouped together** and **summed**

```
plt.hist(train['AddressAccuracy']);
```



Feature Selection/Engineering

- Collected data are split into separate rows once mosquito counts exceed 50
- We needed to combine these data entries together

```
train_new = train.groupby(['Date','Trap', 'Latitude', 'Longitude', 'Species']).sum()
train_new
```

Date	Trap	Latitude	Longitude	Species	NumMosquitos	WnvPresent
2007-05-29	T002	41.954690	-87.800991	CULEX PIPIENS/RESTUANS	1	0
				CULEX RESTUANS	1	0
	T007	41.994991	-87.769279	CULEX RESTUANS	1	0
	T015	41.974089	-87.824812	CULEX PIPIENS/RESTUANS	1	0
				CULEX RESTUANS	4	0
...	...	...	...	...	...	...
2013-09-26	T232	41.912563	-87.668055	CULEX PIPIENS/RESTUANS	1	0
	T233	42.009876	-87.807277	CULEX PIPIENS/RESTUANS	5	0
	T235	41.776428	-87.627096	CULEX PIPIENS/RESTUANS	1	0
	T900	41.974689	-87.890615	CULEX PIPIENS	37	0
				CULEX PIPIENS/RESTUANS	43	1

Feature Selection/Engineering

- After grouping and summing up, we ended up with WnvPresent values of more than 1

- Since we are mainly concerned with predicting presence of virus or not (0 or 1), we map all the values above 1 to 1 (virus present)

```
train_new['WnvPresent'].value_counts()
```

```
0     8018
1      409
2       31
3        9
4        2
7        1
6        1
5        1
10       1
9        1
8        1
Name: WnvPresent, dtype: int64
```

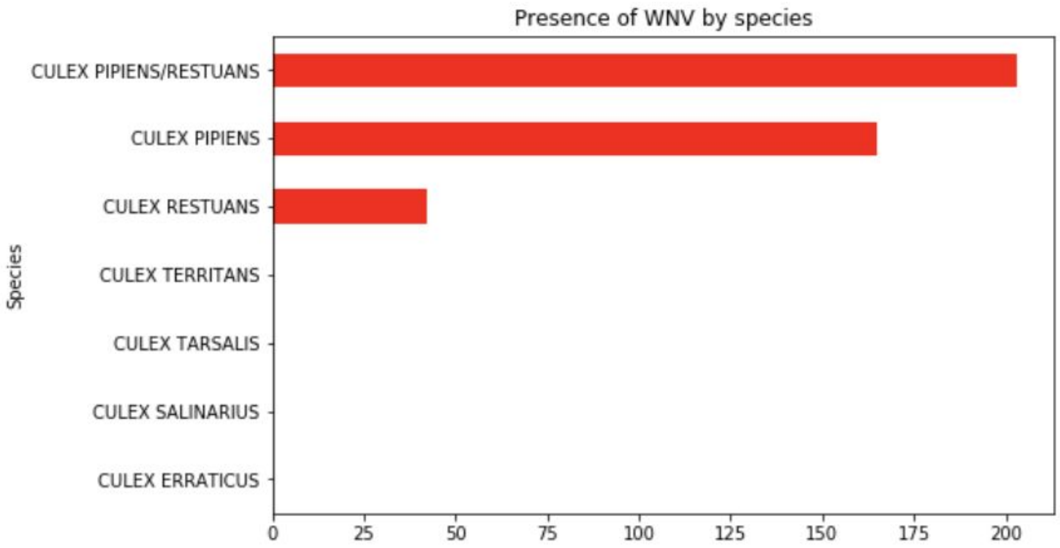
```
train_new['WnvPresent'] = train_new['WnvPresent'].map(lambda x : 1 if x > 0 else x)
```

```
train_new['WnvPresent'].value_counts()
```

```
0    8018
1     457
Name: WnvPresent, dtype: int64
```

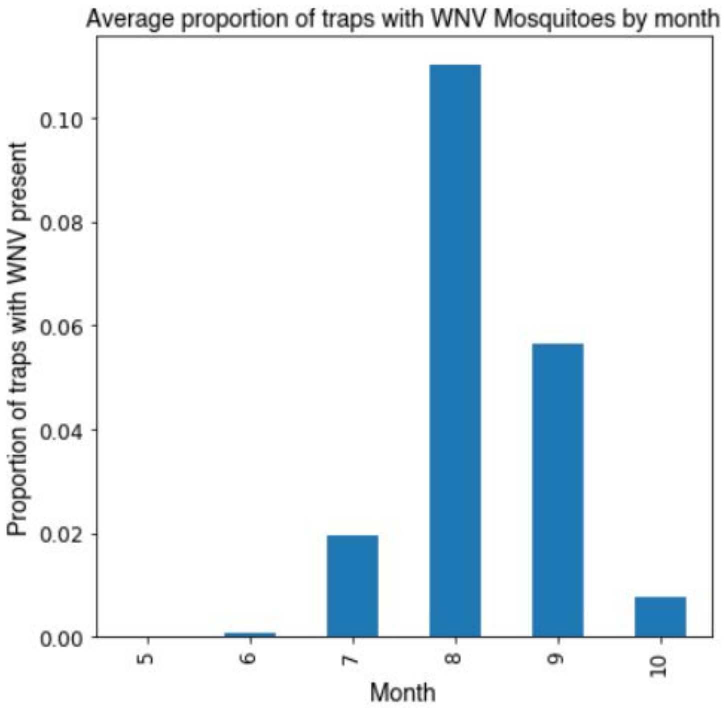
Presence of WNV by Species

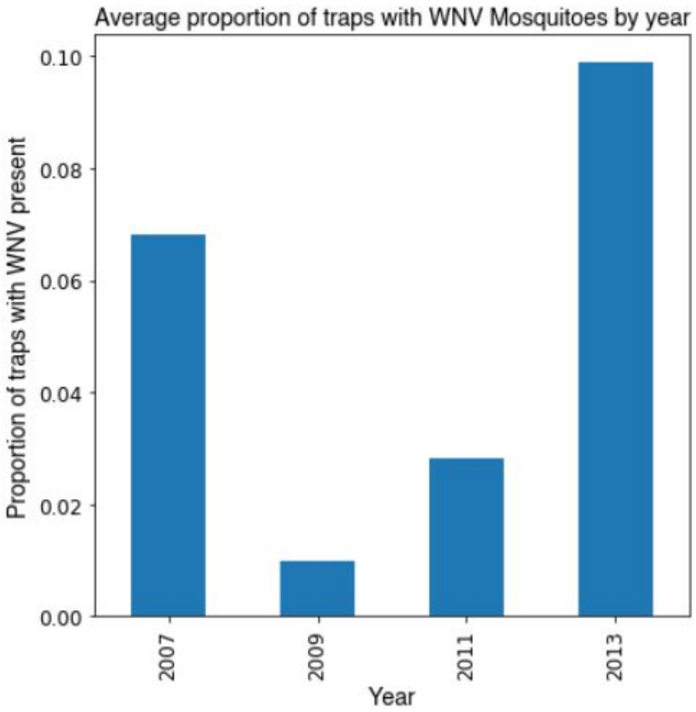
- Not all mosquito species carry the WNV virus
- Culex Pipens/Restuans is **not** a hybrid species, but is recorded as such when tests do are not able to differentiate them adequately.



WNV Presence over time

Train/Test Dataset





- **August is the peak month** for WNV mosquito reproduction
- Number of WNV mosquitoes **dropped after 2007**, but **steadily rose from 2009**

Weather Dataset

- Contains 22 features of weather conditions recorded from 2007 to 2014 from 2 stations: Chicago O'Hare Intl Airport and Chicago Midway Intl Airport

- Station 2 has many missing values, compared to Station 1

```
station1.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1472 entries, 0 to 2942
Data columns (total 25 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Station      1472 non-null   int64
 1   Date         1472 non-null   datetime64[ns]
 2   Tmax         1472 non-null   int64
 3   Tmin         1472 non-null   int64
 4   Tavg         1472 non-null   float64
 5   Depart       1472 non-null   float64
 6   DewPoint     1472 non-null   int64
 7   WetBulb      1469 non-null   float64
 8   Heat         1472 non-null   float64
 9   Cool         1472 non-null   float64
 10  Sunrise      1472 non-null   object
 11  Sunset       1472 non-null   object
 12  CodeSum      1472 non-null   object
 13  Depth        1472 non-null   float64
 14  Water1       0 non-null      float64
 15  SnowFall     1472 non-null   float64
 16  PrecipTotal  1472 non-null   float64
 17  StnPressure  1470 non-null   float64
 18  SeaLevel     1467 non-null   float64
 19  ResultSpeed  1472 non-null   float64
 20  ResultDir    1472 non-null   float64
 21  AvgSpeed     1472 non-null   float64
 22  day          1472 non-null   int64
 23  month        1472 non-null   int64
 24  year         1472 non-null   int64
```

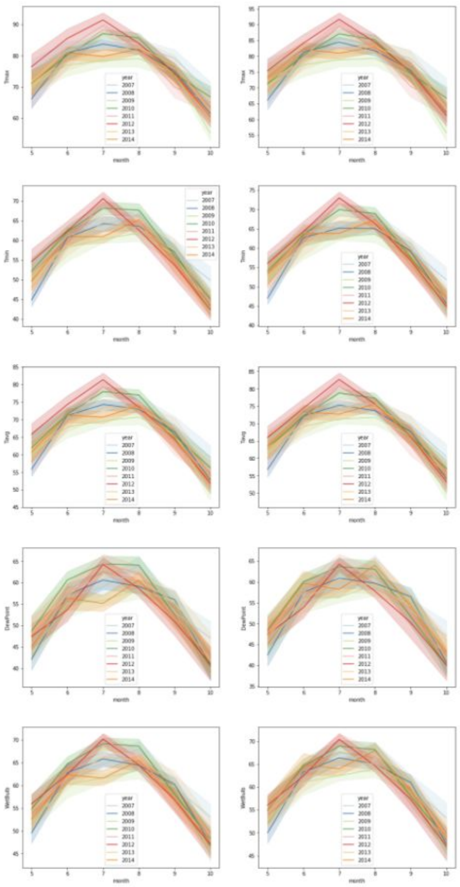
```
station2.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1472 entries, 1 to 2943
Data columns (total 25 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Station      1472 non-null   int64
 1   Date         1472 non-null   datetime64[ns]
 2   Tmax         1472 non-null   int64
 3   Tmin         1472 non-null   int64
 4   Tavg         1461 non-null   float64
 5   Depart       0 non-null      float64
 6   DewPoint     1472 non-null   int64
 7   WetBulb      1471 non-null   float64
 8   Heat         1461 non-null   float64
 9   Cool         1461 non-null   float64
 10  Sunrise      0 non-null      object
 11  Sunset       0 non-null      object
 12  CodeSum      1472 non-null   object
 13  Depth        0 non-null      float64
 14  Water1       0 non-null      float64
 15  SnowFall     0 non-null      float64
 16  PrecipTotal  1470 non-null   float64
 17  StnPressure  1470 non-null   float64
 18  SeaLevel     1468 non-null   float64
 19  ResultSpeed  1472 non-null   float64
 20  ResultDir    1472 non-null   float64
 21  AvgSpeed     1469 non-null   float64
 22  day          1472 non-null   int64
 23  month        1472 non-null   int64
 24  year         1472 non-null   int64
```

Dropping Station 2

- Values from station 1 and station 2 appear to have little significant difference, with high correlation between each other (>0.7) other than PrecipTotal.

- We retain PrecipTotal feature to investigate its influence

Station 1 Station 2



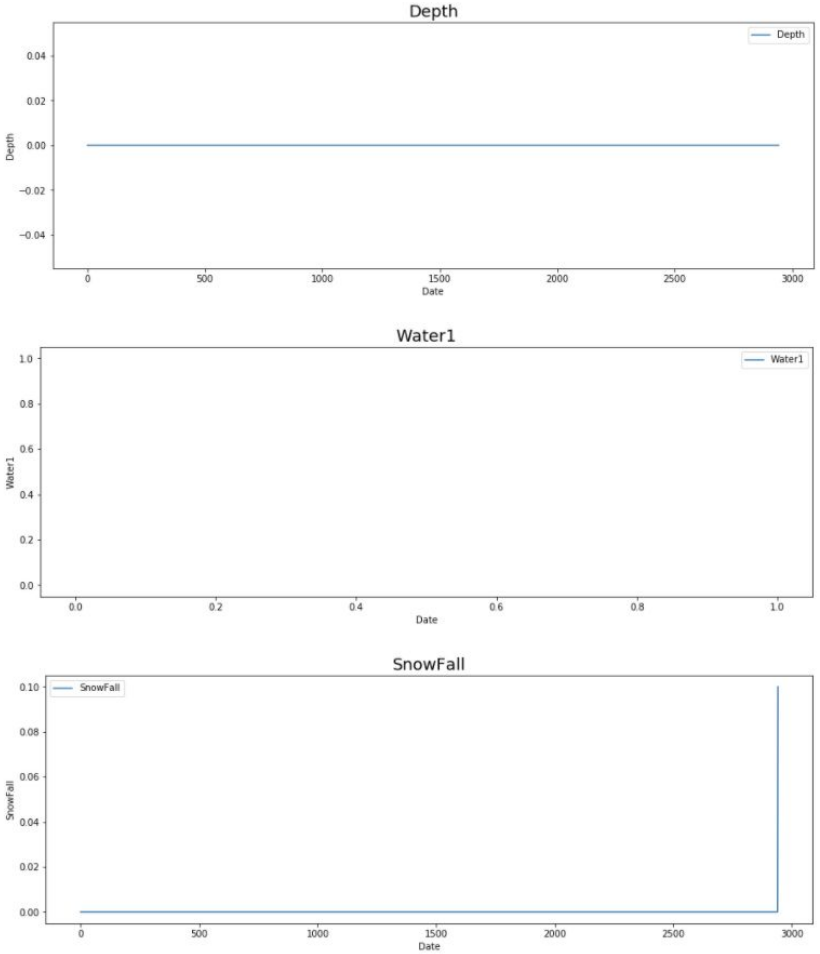
Correlation Values between S1 & S2

	Correlation
StnPressure	0.998212
SeaLevel	0.997670
WetBulb	0.994167
Tavg	0.992288
DewPoint	0.989713
Heat	0.989423
Tmax	0.986896
Cool	0.982518
Tmin	0.977881
AvgSpeed	0.950779
ResultSpeed	0.950507
ResultDir	0.822797
PrecipTotal	0.669457

Dealing with inconsequential features

- Other features had values of 0 or no significant values

- We dropped these features as well

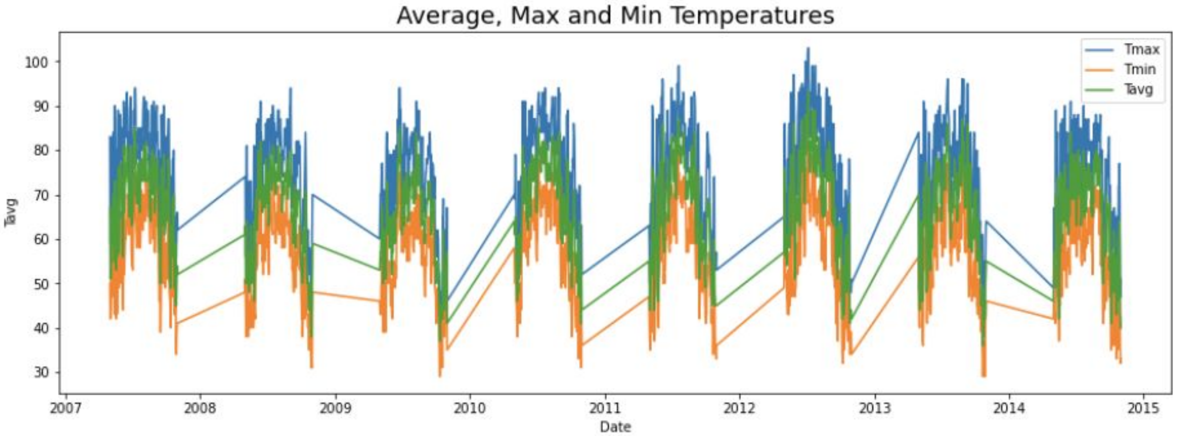


Retaining Aggregated Features

Several features have aggregated variables

- tmin, tmax > tavg
- resultspeed > avgspeed

All variables will be retained to keep any nuances in the data



Feature Engineering

1. Splitting up Codesum

2. Daylight Mins
   a. Aggregating data from sunset and sunrise

3. Rolling weather elements with 7 and 14 days
   a. To investigate the effect of previous weather patterns on WNV
   b. Larvae takes 7-14 days to develop into an adult

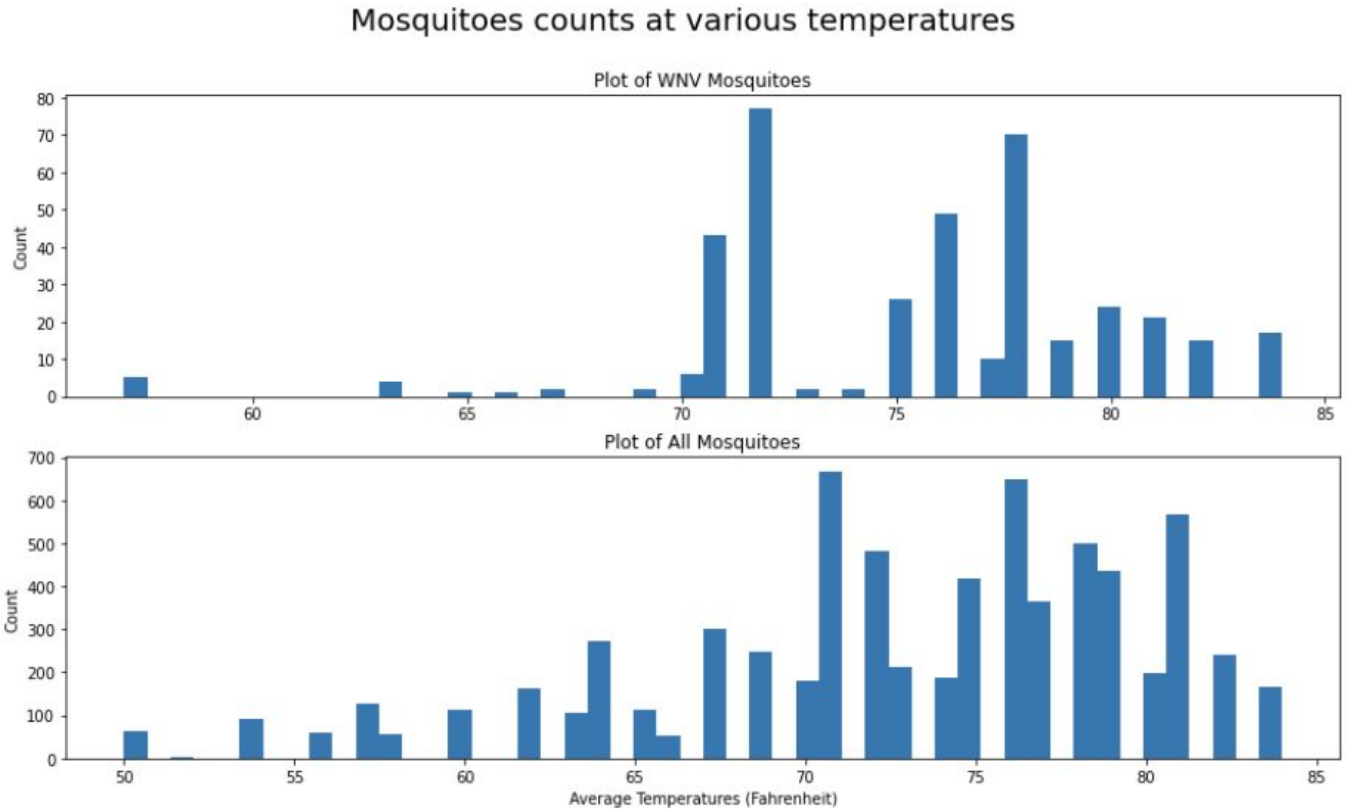
```
CodeSum = pd.DataFrame(X.toarray(),
                       columns=cvec.get_feature_names())
CodeSum.head()
```

	BCFG	BR	DZ	FG	FU	HZ	MIFG	RA	SN	SQ	TS	TSRA	VCTS
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0

```
def minute_converter(time):
    return (int(time[0:2])*60 + int(time[2:4]))
```

WNV Presence over temperature

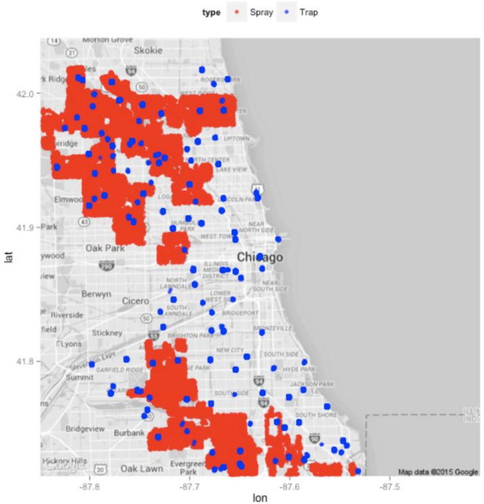
Weather Dataset



WNV carrying mosquitoes thrive in warmer temperatures

Spray Dataset

- Spraying was attempted as part of mosquito control efforts
- We collected data on the Date, Time, and Location of these spraying efforts

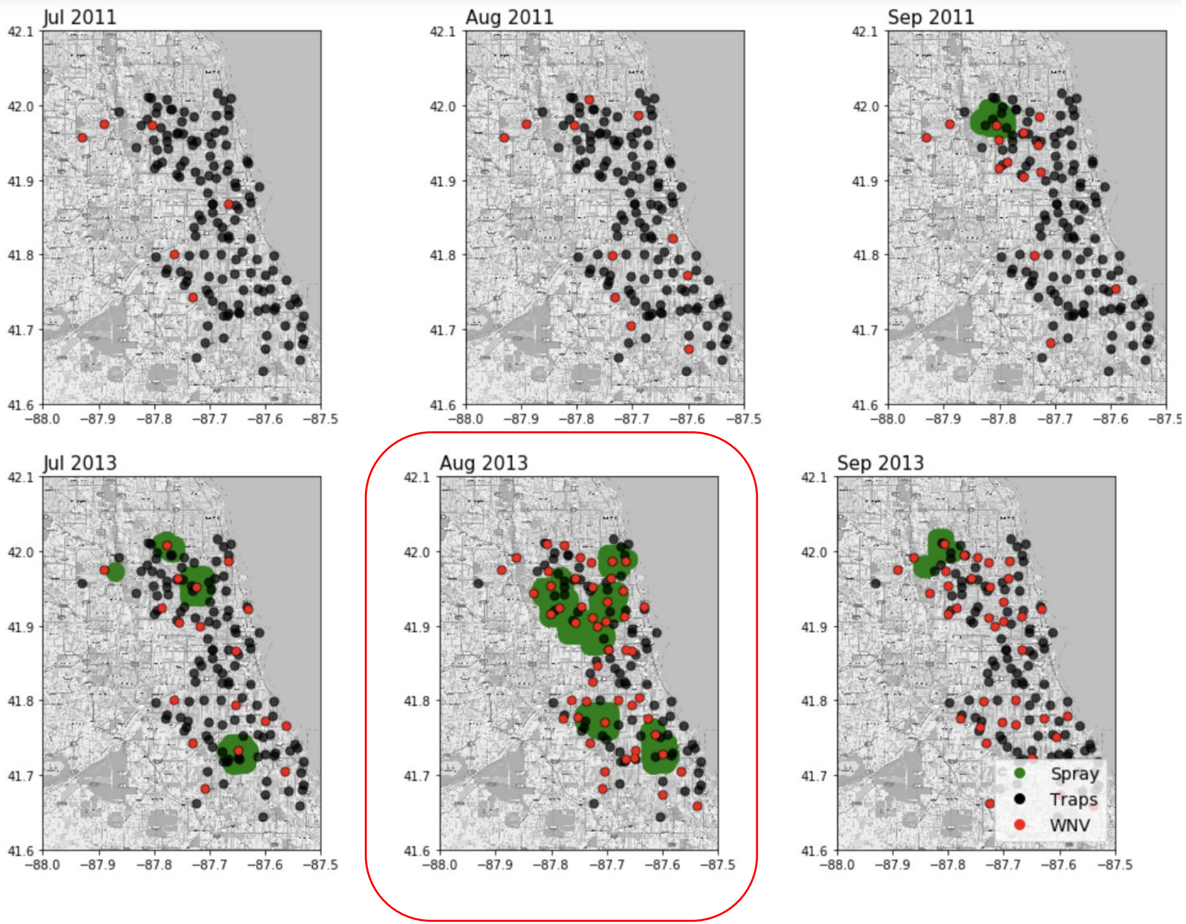


Locations where Sprays are conducted
(with Traps visualized)

Sprays were done in 2011 and 2013, and after sunset hours

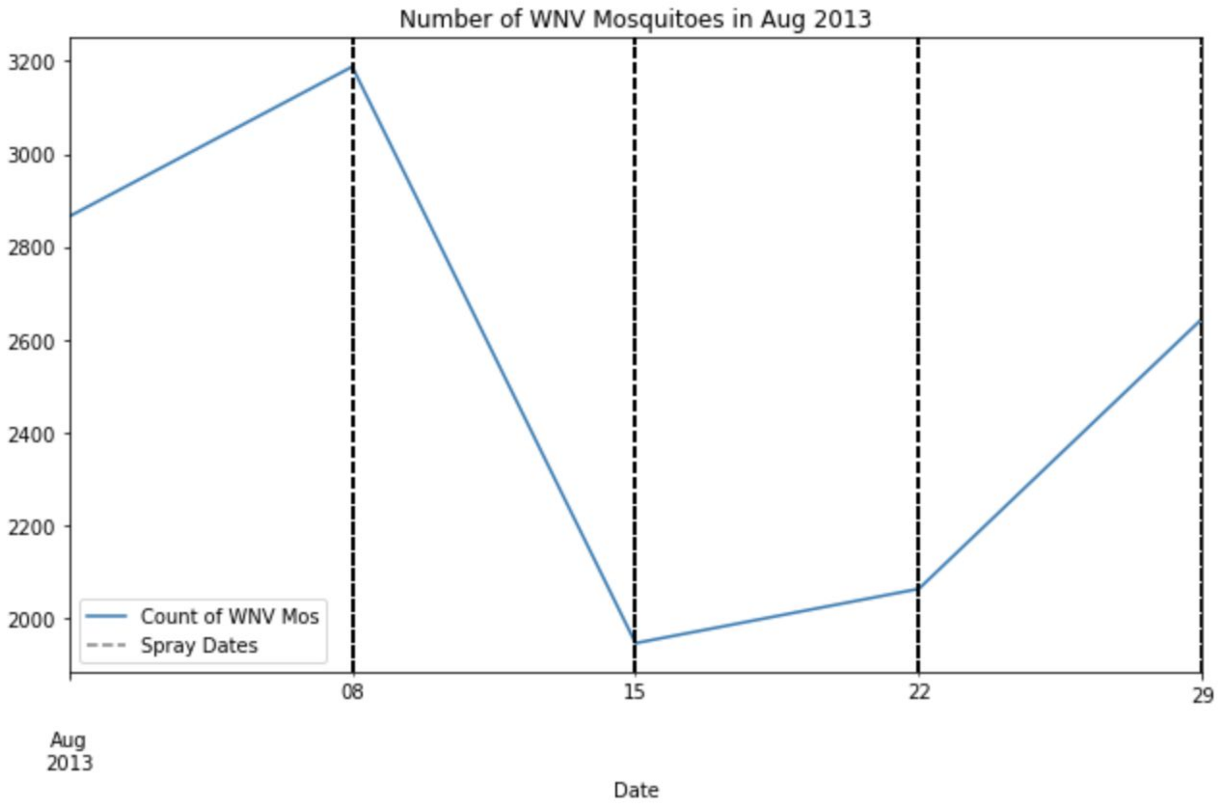
Investigating the impact on Mosquito Counts after spraying

Based on most optimistic scenario: Aug 2013 where sprayed occured over the most areas where WNV presence was detected



Investigating the impact on Mosquito Counts after spraying

Count of mosquitoes did not seem to drop after spray dates



Pre-Modeling

Imbalance of classes

- Since there is a strong imbalance in the classes, SMOTE is used before modeling

```
combined_df['WnvPresent'].value_counts(normalize=True)
```

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

```
# SMOTEing features and target variables

sm = SMOTE(random_state=42)

X_train_sm, y_train_sm = sm.fit_sample(X_train_ss, y_train)
```

Metric of Focus

ROC-AUC, with emphasis on Sensitivity

- Since there is a strong imbalance in the classes, ROC-AUC will be a better metric for modelling because it measures the degree of separability. It tells how much model is capable of distinguishing between classes.

- Unnecessary spraying might have detrimental effects on public health and ecosystems

- High sensitivity focuses on true positive areas, and penalizes false positive to reduce indiscriminate spraying

$$\text{Sensitivity} = \frac{\textit{True Positives}}{\textit{All Positives}} = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Modelling

- Log Reg, SGD and AdaBoost all have similar ROC-AUC scores
- However, AdaBoost seems to prioritise Specificity over Sensitivity
- We selected Log Reg as it has the highest ROC-AUC and Sensitivity Score

	Score	tn	fp	fn	tp	Sensitivity	Specificity	Accuracy
lr	0.775758	1766	640	25	112	0.817518	0.733998	0.738498
rf	0.659193	2171	235	80	57	0.416058	0.902328	0.876131
sgd	0.768459	1766	640	27	110	0.80292	0.733998	0.737711
ada	0.770175	1985	421	39	98	0.715328	0.825021	0.819111
bag	0.65792	2200	206	82	55	0.40146	0.914381	0.886748
xgb	0.702989	2171	235	68	69	0.50365	0.902328	0.880849
svc	0.661606	1726	680	54	83	0.605839	0.717373	0.711365

Feature Importance

Strong Predictors

- Dewpoint (Rolling 7, 14)
- Wetbulb (Rolling 7, 14)
- Thunderstorm, rain (Rolling 7, 14) (within vicinity as well)
- Airpressure (Stn level, sea level)
- Fog
- Species of mosquitoes

Log Reg

	Weight
DewPoint_7	898.824
StnPressure	73.806
Species_CULEX PIPIENS/RESTUANS	24.410
AvgSpeed_14	20.879
Tmin	20.457
Species_CULEX PIPIENS	15.866
DewPoint_14	14.611
BCFG_7	13.808
Species_CULEX RESTUANS	11.906
Tmin_14	6.764
MIFG_7	6.245
Depart	5.891
Heat_7	5.612
month_7	5.383
VCTS_14	3.602
SeaLevel_14	3.597
WetBulb_14	3.248
RA_7	3.081
VCTS	2.738
FG	2.573

SGD

	Weight
DewPoint_14	2.552
Species_CULEX PIPIENS/RESTUANS	2.142
DewPoint_7	2.091
Species_CULEX PIPIENS	1.990
WetBulb_14	1.725
TSRA_7	1.662
SeaLevel_14	1.609
WetBulb_7	1.501
BCFG_7	1.473
month_8	1.457
StnPressure_14	1.430
FG	1.413
StnPressure	1.405
DZ_14	1.398
ResultDir_7	1.394
TS_14	1.391
VCTS	1.374
RA_7	1.344
ResultSpeed_14	1.332
Tmax	1.327

Inferences

Most predictors give an emphasis on humidity of weather

Rainy weather also seems to be an important predictor

Warm temperatures are a factor, but not as strong as initially thought

August and September seem to be prime months for WNV propagation

Log Reg

	Weight
DewPoint_7	898.824
StnPressure	73.806
Species_CULEX PIPIENS/RESTUANS	24.410
AvgSpeed_14	20.879
Tmin	20.457
Species_CULEX PIPIENS	15.866
DewPoint_14	14.611
BCFG_7	13.808
Species_CULEX RESTUANS	11.906
Tmin_14	6.764
MIFG_7	6.245
Depart	5.891
Heat_7	5.612
month_7	5.383
VCTS_14	3.602
SeaLevel_14	3.597
WetBulb_14	3.248
RA_7	3.081
VCTS	2.738
FG	2.573

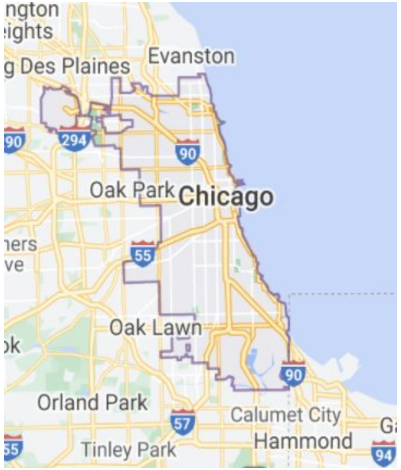
SGD

	Weight
DewPoint_14	2.552
Species_CULEX PIPIENS/RESTUANS	2.142
DewPoint_7	2.091
Species_CULEX PIPIENS	1.990
WetBulb_14	1.725
TSRA_7	1.662
SeaLevel_14	1.609
WetBulb_7	1.501
BCFG_7	1.473
month_8	1.457
StnPressure_14	1.430
FG	1.413
StnPressure	1.405
DZ_14	1.398
ResultDir_7	1.394
TS_14	1.391
VCTS	1.374
RA_7	1.344
ResultSpeed_14	1.332
Tmax	1.327

Case Study: Sacramento 2005

Case study: Economic Cost Analysis of WNV Outbreak in Sacramento County, 2005

- Chicago's infected cases were estimated based on population ratio
    - Helps to estimate costs on social landscape, and medical resources
- Chicago's spraying costs were based on area ratio
    - Helps to estimate cost of intervention



Cost Benefit Analysis

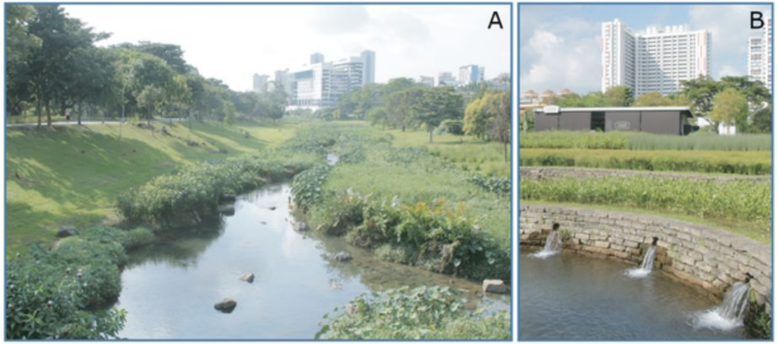
	Socio-economic	Medical Costs	Ecology Costs
WNF	• Productivity Loss per Sick Day reported	• Diagnostic tests • Physician visits • Treatment costs	• Death in horses and birds more commonly
WNND	+ Cost of human lives	+ Inpatient and Outpatient costs	

Estimated $ of Virus

	Socio-economic	Medical Costs	Ecology Costs
Sacramento	USD 586,305	USD2.86 Million	**Requires further study*
Chicago	~USD1.21 Million	~USD5.9 Million	**Requires further study*

Case Study: Singapore

1. Natural Suppression Technology
   a. Wolbachia-Aedes
2. Intra- and Inter-sectoral collaboration
   a. Drains were designed with sufficient gradient to prevent water pooling
   b. Regular removal of floating vegetation in parks
3. Vector Surveillance
   a. Monitoring traps every 2 weeks
   b. Insecticides
   c. Monitoring construction sites
4. Engagement of communities
   a. Activations in dormitories, shopping mall
   b. Education via the ‘5-Step Mozzie Wipeout’



Cost Benefit Analysis

	Socio-economic	**Production Costs**	**Ecology Costs**
Aerial Spraying	Manpower hours Business Costs	Spray Cost	Environmental Pollution
Community Engagement	Manpower hours	Advertising/ Campaign Costs	Nil
Natural Suppression	Manpower hours	**Requires further study*	**Minimal with proper mitigation*

Estimated $ of Intervention

	Aerial Spraying	Education Costs	Ecology Costs
Sacramento/ Singapore	USD701,790	SGD5 Million	**Requires further study*
Chicago	~USD0.165 Million	~USD0.391 Million	**Requires further study*

Recommendations

- Cost of life is valued at $9.1 million per citizen, much valued over the cost of spraying

- **Spraying**
    - Propose re-examination of insecticide spraying as it does not seem to be very effective
    - Consider oil treatment of stagnant water bodies to kill off larvae
    - When: 1-2 weeks prior to warm and/or humid seasons (e.g. August)

- **Educational/Awareness campaign**
    - E.g. Clearing stagnant water in the community

- **Smart Engineering for water infrastructure projects**
    - E.g. Prevention of water pooling in drainage infrastructure