Project 4
West Nile Virus Prediction

24 Sep 2022
Wei Hao
Connie
Ethan
Yonghe
Anand

Agenda

Contents	Presenter
Problem Statement & Background	Anand
Data Cleaning	Ethan
EDA	Ethan
Feature Engineering	Weihao
Modelling	Weihao/ Yonghe
Model Evaluation	Yonghe
Cost Benefit Analysis	Anand
Spray Analysis	Connie
Conclusion, Considerations & Recommendations	Connie

Problem Statement

In order to efficiently combat the West Nile Virus in Chicago we aim:

1. To build a model and make predictions that the city of Chicago can use about when and where when it decides to spray pesticides
2. To conduct a cost-benefit analysis that include annual cost projections for various levels of pesticide coverage (cost) and the effect of these various levels of pesticide coverage (benefit)

Background - About Chicago

- City in the State of Illinois
- Third latest city in the US
- Home to 2.7 million residents
- Land size about 600 km^2 (Singapore is about 728 km^2)
- Extensive parklands, including 30km^2 of city parks attract estimated 86 million visitors annually
- Very passionate sports town

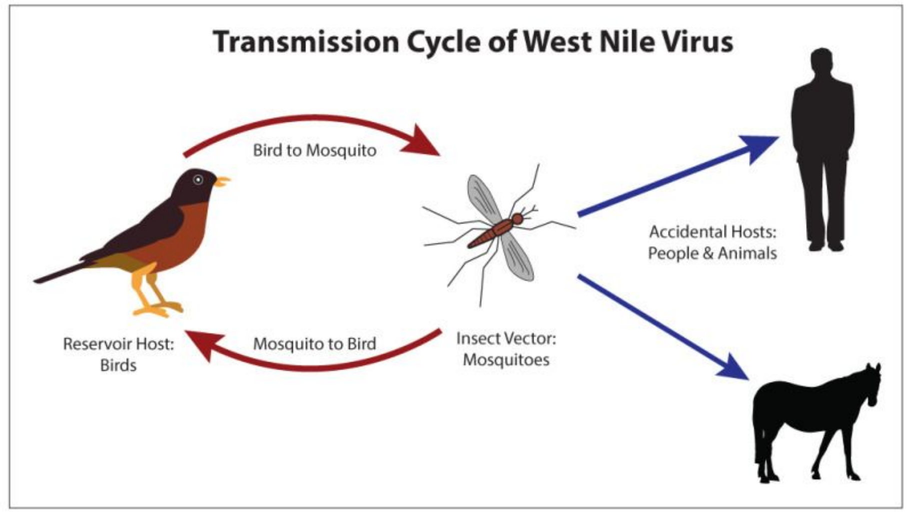
Background - West Nile Virus

What is the West Nile Virus?

- Causes the West Nile Fever infection
- 80% of infections have no symptoms
- 20% of people develop a fever, headache, vomiting, or a rash

Transmission of Virus

- West Nile Virus is found in birds
- Birds transmit the virus to mosquitoes who then infect humans and animals

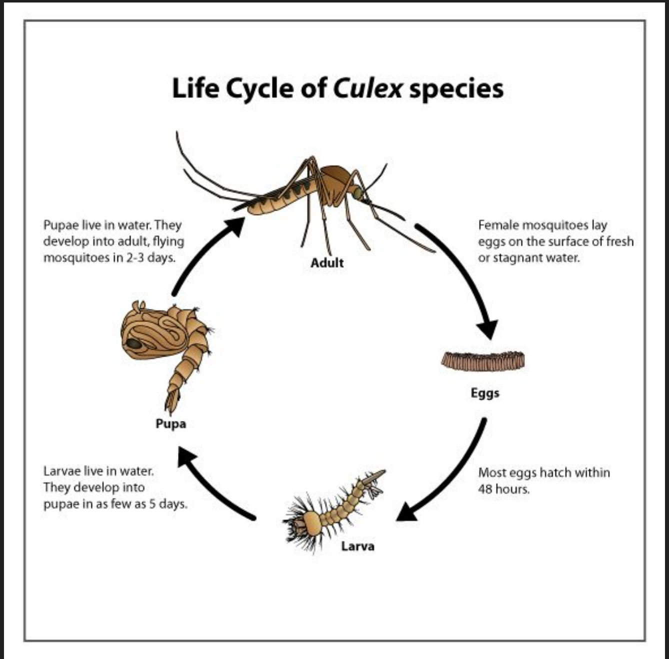


Background - WNV in Chicago

- Chicago has one of the highest death rates of West Nile Virus in the US
- Symptoms include: Headache and bodyache, joint pain, vomiting, diarrhea, etc
- 1 in 150 develop serious symptoms: Encephalitis, Meningitis
- 1 in 10 cases result in death
- No vaccine is available

Background - Life cycle of Culex species

- Eggs to larva within 48 hours
- Larvae live in water, develop into pupae in 5 days
- Pupae also live in water, develop into flying mosquito in 2-3 days
- In total, about <u>7-10</u> days for an egg to develop into an adult mosquito
- Information is crucial for determining the frequency on when to spray to prevent the spread of the West Nile Virus



Data Description

	Years available for each Dataset									
Dataset	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Rows**	**Columns**
Train	✔		✔		✔		✔		10,506	12
Test		✔		✔		✔		✔	116,293	11
Weather	✔	✔	✔	✔	✔	✔	✔	✔	13,710	22
Spray					✔		✔		2,944	4

Data Cleaning - Train

Observations

- No null values in all columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10506 entries, 0 to 10505
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Date                    10506 non-null  object
 1   Address                 10506 non-null  object
 2   Species                 10506 non-null  object
 3   Block                   10506 non-null  int64
 4   Street                  10506 non-null  object
 5   Trap                    10506 non-null  object
 6   AddressNumberAndStreet  10506 non-null  object
 7   Latitude                10506 non-null  float64
 8   Longitude               10506 non-null  float64
 9   AddressAccuracy         10506 non-null  int64
 10  NumMosquitos            10506 non-null  int64
 11  WnvPresent              10506 non-null  int64
dtypes: float64(2), int64(4), object(6)
memory usage: 985.1+ KB
```

Data Cleaning

- Change “Date” data-type to datetime

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10506 entries, 0 to 10505
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Date                    10506 non-null  datetime64[ns]
 1   Address                 10506 non-null  object
 2   Species                 10506 non-null  object
 3   Block                   10506 non-null  int64
 4   Street                  10506 non-null  object
 5   Trap                    10506 non-null  object
 6   AddressNumberAndStreet  10506 non-null  object
 7   Latitude                10506 non-null  float64
 8   Longitude               10506 non-null  float64
 9   AddressAccuracy         10506 non-null  int64
 10  NumMosquitos            10506 non-null  int64
 11  WnvPresent              10506 non-null  int64
dtypes: datetime64[ns](1), float64(2), int64(4), object(5)
memory usage: 985.1+ KB
```

Data Cleaning - Test

Observations

- No null values in all columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116293 entries, 0 to 116292
Data columns (total 11 columns):
 #   Column                  Non-Null Count   Dtype 
---  ------                  --------------   ----- 
 0   Id                      116293 non-null  int64 
 1   Date                    116293 non-null  object
 2   Address                 116293 non-null  object
 3   Species                 116293 non-null  object
 4   Block                   116293 non-null  int64 
 5   Street                  116293 non-null  object
 6   Trap                    116293 non-null  object
 7   AddressNumberAndStreet  116293 non-null  object
 8   Latitude                116293 non-null  float64
 9   Longitude               116293 non-null  float64
 10  AddressAccuracy         116293 non-null  int64 
dtypes: float64(2), int64(3), object(6)
memory usage: 9.8+ MB
```

Data Cleaning

- Change “Date” data-type to datetime

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116293 entries, 0 to 116292
Data columns (total 11 columns):
 #   Column                  Non-Null Count   Dtype 
---  ------                  --------------   ----- 
 0   Id                      116293 non-null  int64 
 1   Date                    116293 non-null  datetime64[ns]
 2   Address                 116293 non-null  object
 3   Species                 116293 non-null  object
 4   Block                   116293 non-null  int64 
 5   Street                  116293 non-null  object
 6   Trap                    116293 non-null  object
 7   AddressNumberAndStreet  116293 non-null  object
 8   Latitude                116293 non-null  float64
 9   Longitude               116293 non-null  float64
 10  AddressAccuracy         116293 non-null  int64 
dtypes: datetime64[ns](1), float64(2), int64(3), object(5)
memory usage: 9.8+ MB
```

Data Cleaning - Spray data

Observations

- 584 null values in “Time” column
- 541 duplicated rows
- All null and duplicate values happen on one single date 2011-09-07

Data Cleaning

- Drop nulls
- Drop duplicates
- Change “Date” data-type to datetime

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14835 entries, 0 to 14834
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       14835 non-null  object
 1   Time       14251 non-null  object
 2   Latitude   14835 non-null  float64
 3   Longitude  14835 non-null  float64
dtypes: float64(2), object(2)
memory usage: 463.7+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13710 entries, 0 to 14834
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       13710 non-null  datetime64[ns]
 1   Time       13710 non-null  object
 2   Latitude   13710 non-null  float64
 3   Longitude  13710 non-null  float64
dtypes: datetime64[ns](1), float64(2), object(1)
memory usage: 535.5+ KB
```

Data Cleaning - Weather data

Observations

- No null values in all columns
- Some non-numeric values in some columns (e.g. “M” in Tavg, “T” in PrecipTotal)
- “-” in Sunset and Sunrise only for Station 2

Data Cleaning

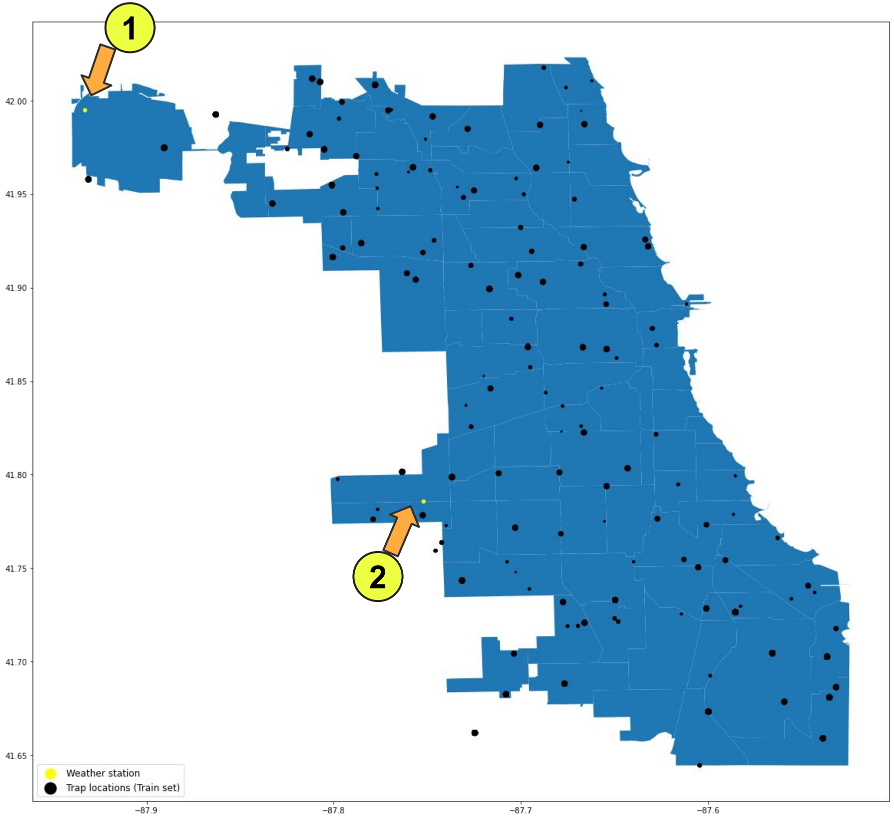
- Change “Date” data-type to datetime
- When reasonably possible, change non-numeric to numeric data
- Update Sunset and Sunrise times for Station 2 data

	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTotal
7	2	2007-05-04	78	51	M	M	42	50	M	M	-	-		M	M	M	0.00
505	2	2008-07-08	86	46	M	M	68	71	M	M	-	-	TS RA	M	M	M	0.28
675	2	2008-10-01	62	46	M	M	41	47	M	M	-	-		M	M	M	0.00
1637	2	2011-07-22	100	71	M	M	70	74	M	M	-	-	TS TSRA BR	M	M	M	0.14
2067	2	2012-08-22	84	72	M	M	51	61	M	M	-	-		M	M	M	0.0[illegible]
2211	2	2013-05-02	71	42	M	M	39	45	M	M	-	-		M	M	M	0.00
2501	2	2013-09-24	91	52	M	M	48	54	M	M	-	-		M	M	M	0.00
2511	2	2013-09-29	84	53	M	M	48	54	M	M	-	-	RA BR	M	M	M	0.22
2525	2	2013-10-06	76	48	M	M	44	50	M	M	-	-	RA DZ BR	M	M	M	0.06
2579	2	2014-05-02	80	47	M	M	43	47	M	M	-	-	RA	M	M	M	0.04
2811	2	2014-08-26	86	49	M	M	68	71	M	M	-	-		M	M	M	T

Date	Station	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum
2007-05-01	2	84	52	68.00	M	51	57	0	3	448	1849	
[illegible]	2	60	43	52.00	M	42	47	13	0	447	1850	BR HZ
2007-05-03	2	67	48	58.00	M	40	50	7	0	446	1851	HZ
2007-05-04	2	78	51	64.50	M	42	50	M	M	444	1852	
2007-05-05	2	66	54	60.00	M	39	50	5	0	443	1853	

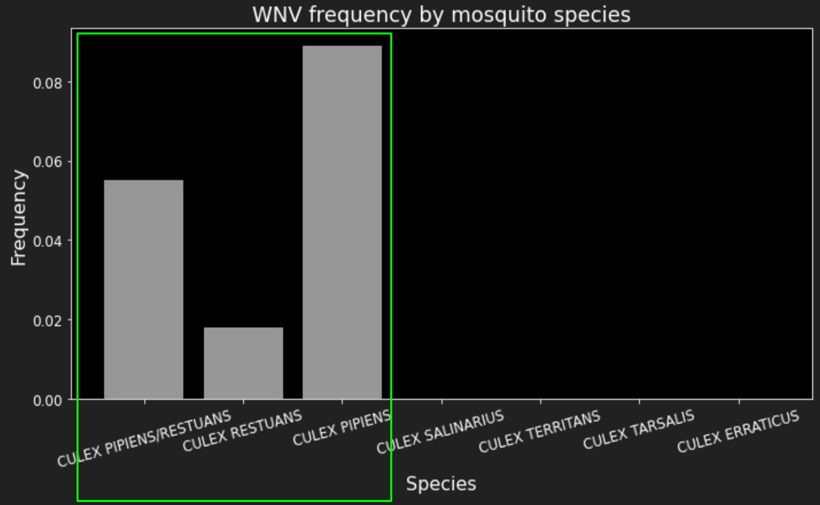
EDA - Trap Locations

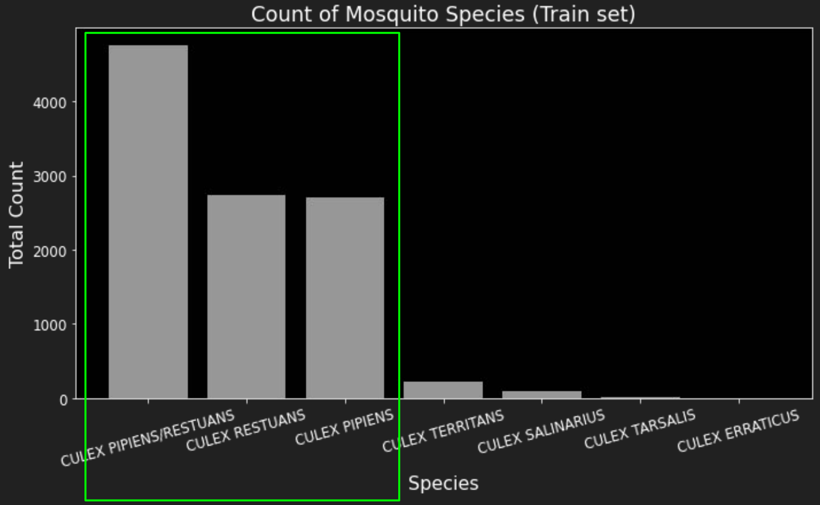
- 136 trap locations are scattered across Chicago, represented by black dots
- Size of black dots represents the number of mosquitoes caught
- Weather stations are represented by yellow dots
- Station 1: Chicago O'Hare International Airport
- Station 2: Chicago Midway International Airport



EDA - Mosquito Species

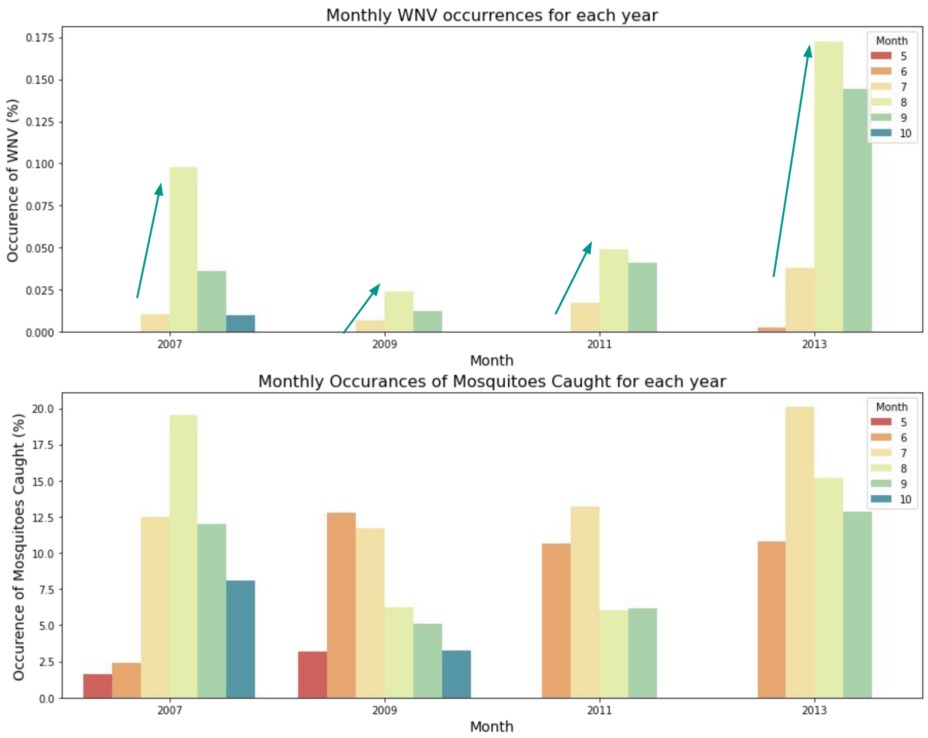
- There are 7 unique species in the **Train dataset**
- Only 3 species are found to spread the West Nile Virus
    - Culex Pipiens/Restuans
    - Culex Restuans
    - Culex Pipiens
- The species that do not spread the virus have low counts in the Train set, but high counts in the Test set





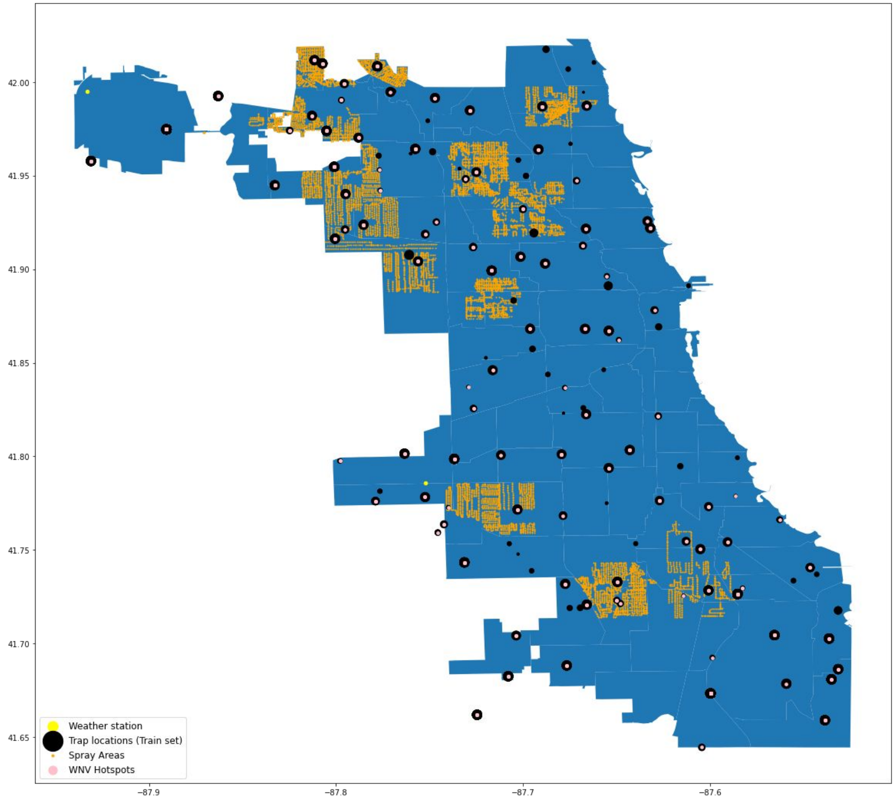
EDA - Seasonality Effects

- WNV cases tend to see a sharp peak in August before dropping
- The number of mosquitoes caught show a similar trend where there is a sharp peak before dropping
- There is likely a time lag between mosquitoes caught and WNV cases
- WNV cases coincides with the summer months of early June to end August



EDA - Spray Data

- Pink dots represent locations where WNV cases are present (“hotspots”)
- Black and pink dots tend to coincide
- Orange areas show where spraying takes place - Not all hotspots or locations with mosquitoes are being sprayed
- It is difficult to visualise the relationship now - we will focus on the spray effects at particular times later



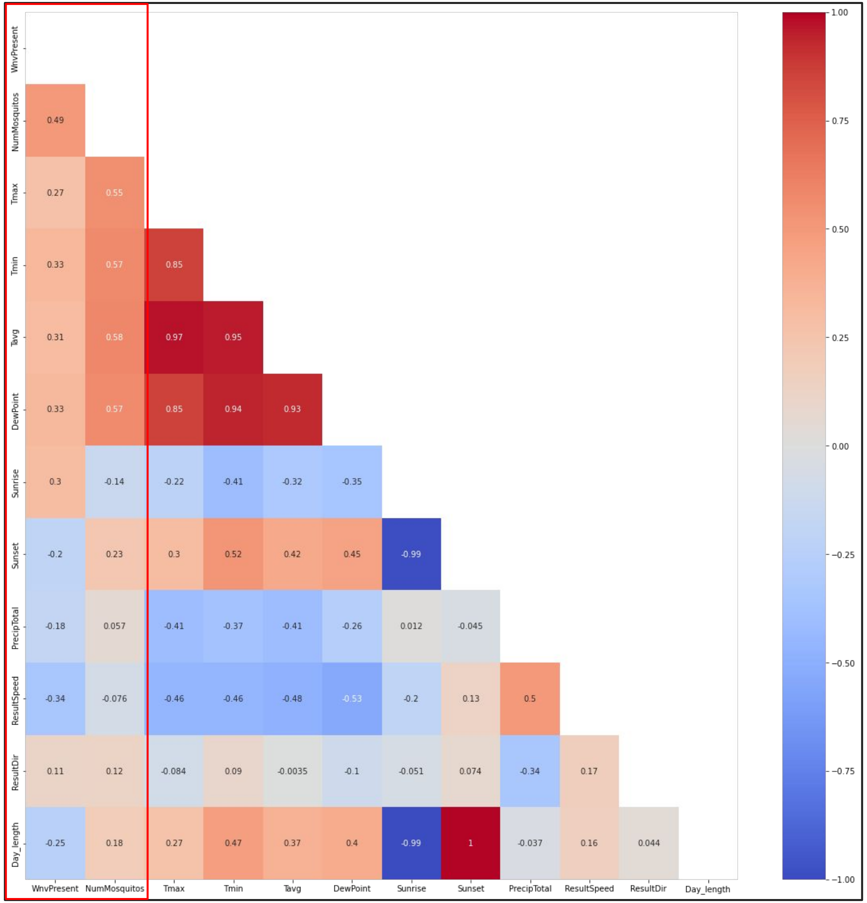
EDA - Weather Data

Correlation table with numeric data

- Max, Min, Average Temperature
- Dewpoint
- Sunset and sunrise
- Precipitation
- Wind Speed (mph)
- Wind Direction
- Day length

Features with higher correlation:

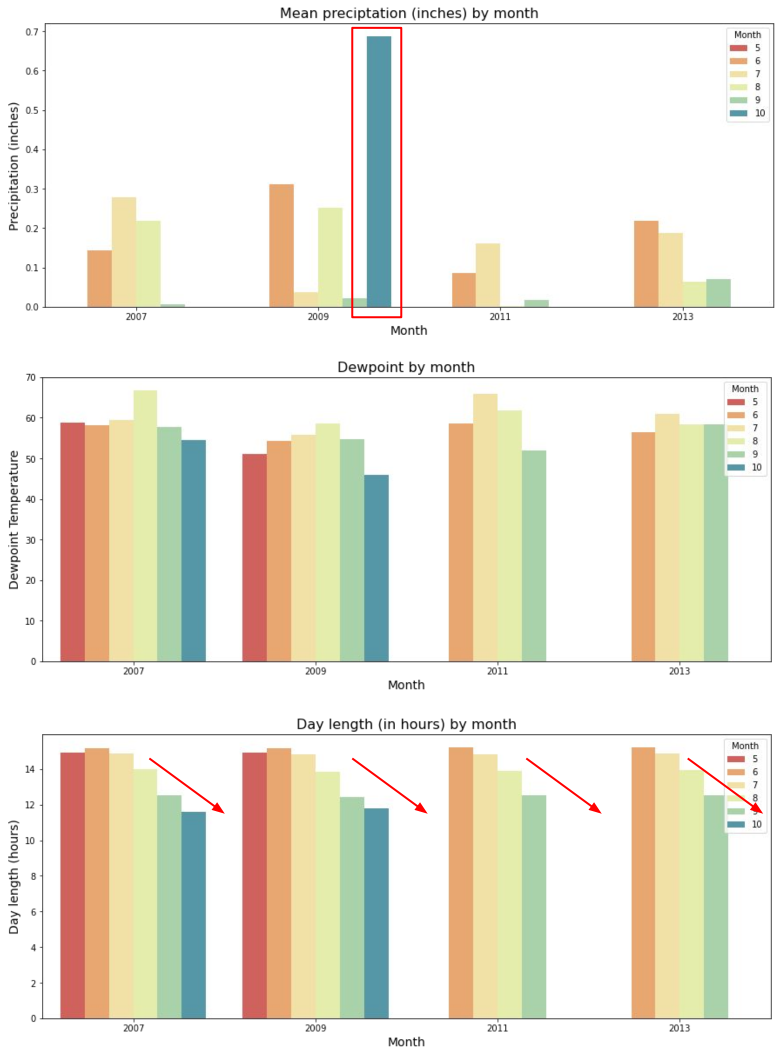
- Temperature and Dewpoint



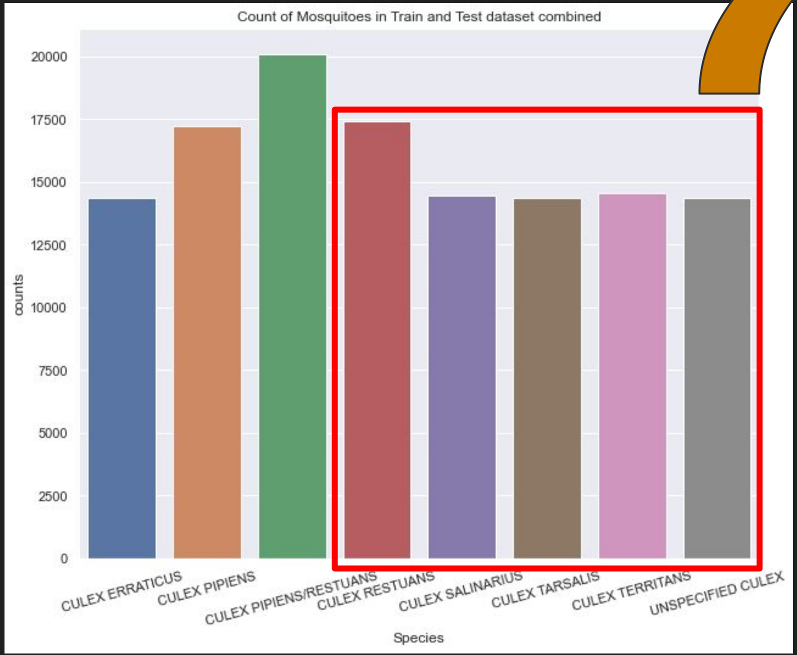
EDA - Selected Weather Data

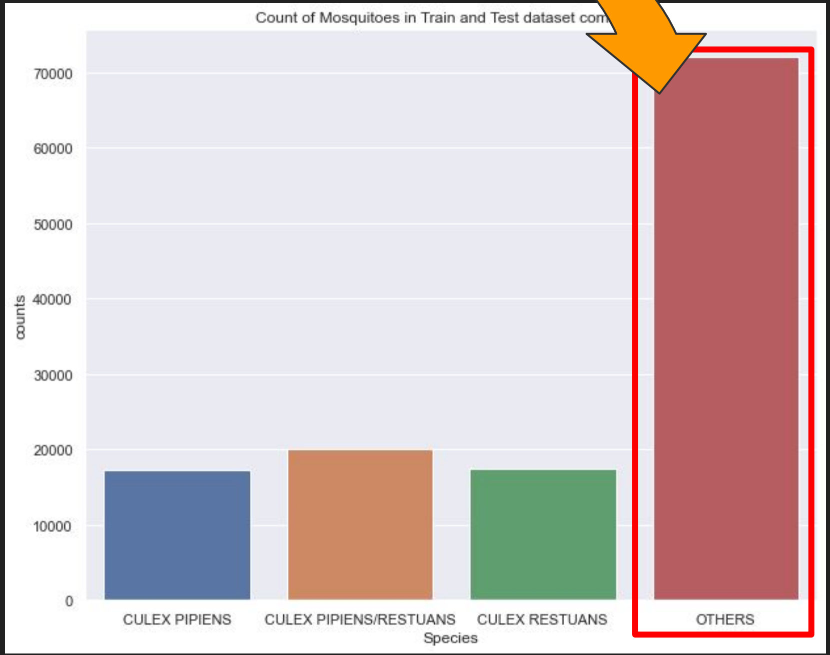
Interesting observations:

- Mean precipitation: Oct 2009 looks like an outlier as there is only one data point in that month
- Dewpoint: Follows a similar seasonal trend to temperature as they are highly correlated
- Day length: Calculated as part of Feature Engineering. Days get shorter after the summer months.



Feature Engineering

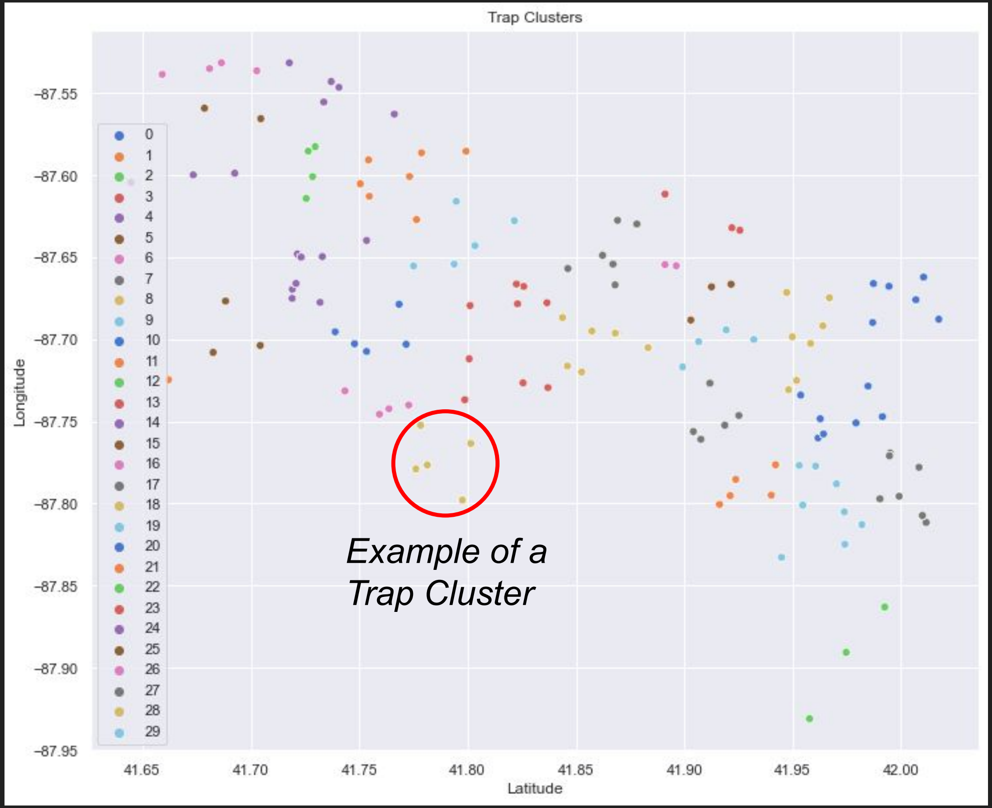




All other species which do not carry the West Nile Virus will be regrouped to 'OTHERS'

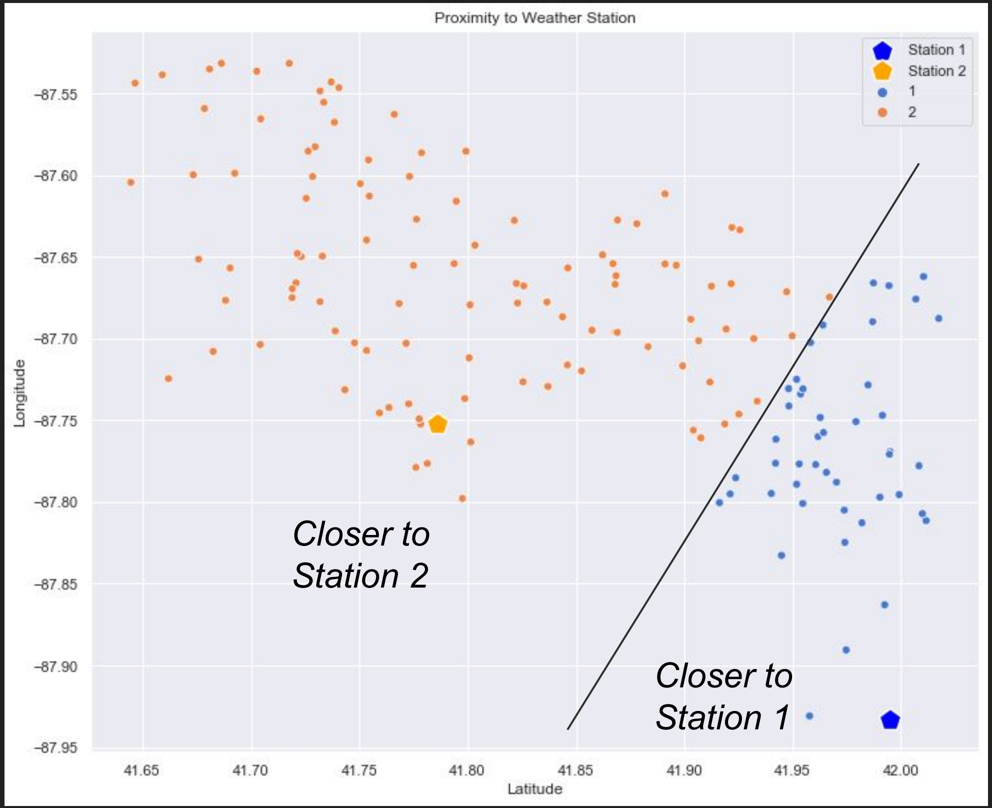
Feature Engineering

- Trap locations are grouped together into 30 clusters using K Means clustering
- These values are then dummified
- Clustering was derived on Train data and subsequently used to predict clustering on Test



Feature Engineering

- Traps are assigned to the nearest weather stations based on proximity
- Weather recorded on the same day is different for both stations
- Blue dots: Closer to Station 1
- Orange dots: Closer to Station 2
- Weather data are then assigned to each trap location



Feature Engineering

Weather features are further transformed with:

1. Shifting feature forward for period of 6 days
2. Looking back 12 days and taking the mean
3. Looking back 14 days and taking the max

Example is shown on the right for Transformation 1 & 2.

	Date	Tavg	TAvg-6	TAvg-Roll-12-Mean	TAvg-Roll-14-Max
7	2007-05-08	68.00	51.00	NaN	NaN
8	2007-05-09	69.00	56.00	NaN	NaN
9	2007-05-10	70.00	58.00	NaN	NaN
10	2007-05-11	61.00	60.00	[illegible]	NaN
11	2007-05-12	55.00	59.00	61.58	NaN
12	2007-05-13	56.00	65.00	60.67	NaN
13	2007-05-14	73.00	68.00	62.50	73.00
14	2007-05-15	69.00	69.00	63.58	73.00
15	2007-05-16	55.00	70.00	63.33	73.00
16	2007-05-17	53.00	61.00	62.75	73.00
17	2007-05-18	57.00	55.00	62.58	73.00
18	2007-05-19	68.00	56.00	62.83	73.00
19	2007-05-20	56.00	73.00	[illegible]	73.00
20	2007-05-21	62.00	69.00	[illegible]	73.00
21	2007-05-22	71.00	55.00	[illegible]	73.00
22	2007-05-23	75.00	53.00	62.50	75.00
23	2007-05-24	76.00	57.00	[illegible]	76.00
24	2007-05-25	63.00	68.00	64.83	76.00
25	2007-05-26	57.00	56.00	63.50	76.00

Max of 14 days

3

2

Add up past 12 values and find the mean

1

Shift 6 dates down

Feature Engineering

After Feature Engineering, dataset would have a total of 86 features, consisting of:

- Time features (Date, year, month, week etc)
- Location features (Address, Block, Street, Longitude, Latitude etc)
- Mosquito Species (3 dummy variables)
- 30 trap clusters
- 9 weather features
- 27 (9x3) transformed weather features

```
Index(['Date', 'Address', 'Block', 'Street', 'Trap', 'AddressNumberAndStreet',
       'Latitude', 'Longitude', 'AddressAccuracy', 'NumMosquitos',
       'WnvPresent', 'geometry', 'year', 'month', 'week', 'day', 'year_month',
       'Station', 'Species_CULEX PIPIENS/RESTUANS', 'Species_CULEX RESTUANS',
       'Species_OTHERS', 'trap_cluster_1', 'trap_cluster_2', 'trap_cluster_3',
       'trap_cluster_4', 'trap_cluster_5', 'trap_cluster_6', 'trap_cluster_7',
       'trap_cluster_8', 'trap_cluster_9', 'trap_cluster_10',
       'trap_cluster_11', 'trap_cluster_12', 'trap_cluster_13',
       'trap_cluster_14', 'trap_cluster_15', 'trap_cluster_16',
       'trap_cluster_17', 'trap_cluster_18', 'trap_cluster_19',
       'trap_cluster_20', 'trap_cluster_21', 'trap_cluster_22',
       'trap_cluster_23', 'trap_cluster_24', 'trap_cluster_25',
       'trap_cluster_26', 'trap_cluster_27', 'trap_cluster_28',
       'trap_cluster_29', 'Tmax', 'Tmin', 'Tavg', 'DewPoint', 'Sunrise',
       'Sunset', 'Day_length', 'PrecipTotal', 'ResultSpeed', 'Tmax-6',
       'Tmin-6', 'Tavg-6', 'DewPoint-6', 'Sunrise-6', 'Sunset-6',
       'Day_length-6', 'PrecipTotal-6', 'ResultSpeed-6', 'Tmax-avg-12',
       'Tmin-avg-12', 'Tavg-avg-12', 'DewPoint-avg-12', 'Sunrise-avg-12',
       'Sunset-avg-12', 'Day_length-avg-12', 'PrecipTotal-avg-12',
       'ResultSpeed-avg-12', 'Tmax-max-14', 'Tmin-max-14', 'Tavg-max-14',
       'DewPoint-max-14', 'Sunrise-max-14', 'Sunset-max-14',
       'Day_length-max-14', 'PrecipTotal-max-14', 'ResultSpeed-max-14'],
      dtype='object')
```

Modelling

- Handle Imbalance Data
- Evaluation Metrics
- Model Evaluation

Modelling - Handle Imbalance Data

Train Data

WNV Present	Percentage
0	95%
1	5%

Undersampling

Oversampling

Processed Train Data to feed into models

WNV Present	Percentage
0	67%
1	33%

WNV Present	Percentage
0	95%
1	5%

SMOTE

WNV Present	Percentage
0	67%
1	33%

Modelling - Evaluation Metrics

- **Precision-Recall AUC Score:**
  - Also known as the Average Precision Score, it is a way to summarize the Precision-Recall curve into a single value
  - Used when data is heavily imbalance and when you care more about the positive class

```
GridSearchCV(pipe, # what object are we optimizing?
             param_grid = pipe_params, # what parameters values are we searching?
             cv=3, # 3-fold cross-validation.
             n_jobs=-1,
             scoring='average_precision' #'average_precision' = precision_recall_auc_score
             )
```

- **F1 Score**
  - Harmonic mean of the precision and recall
  - Used when you care more about the positive class

Models

1	DummyClassifier always predicting 'WnvPresent' to be 1
2	OverSampling + UnderSampling + GradientBoost
3	OverSampling + UnderSampling + RandomForest
4	OverSampling + UnderSampling + LightGBM
5	Smote + GradientBoost
6	Smote + RandomForest
7	Smote + LightGBM

For sake of time, we will only be covering Models 1, 2 & 7 in this presentation.

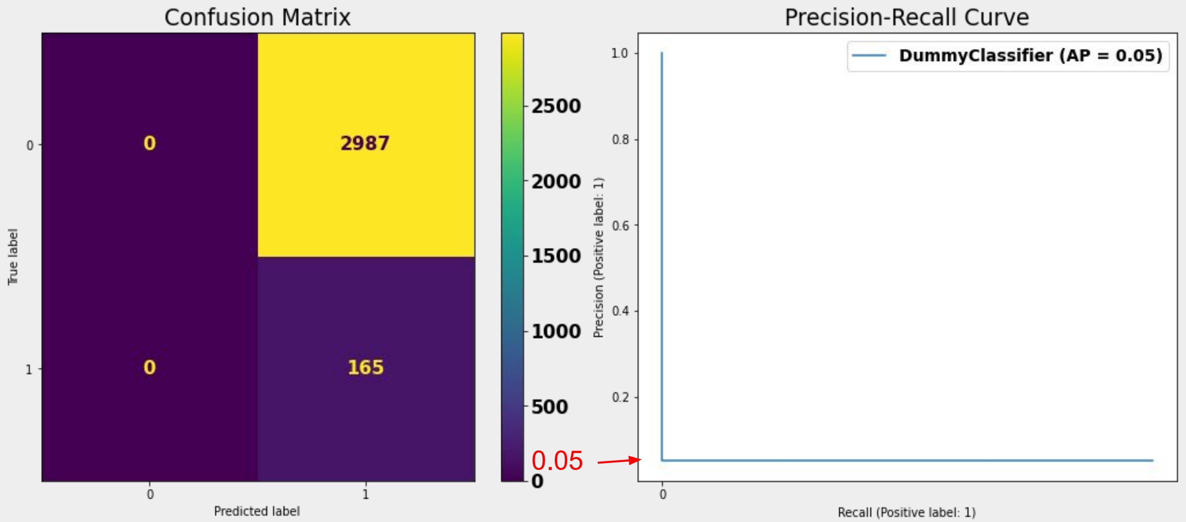
Model 1: Baseline Model

DummyClassifier always predicting 'WnvPresent' to be 1

```
precision_recall_auc_score on training set: 0.052
precision_recall_auc_score on testing set: 0.052
perc_diff: 0.3 %
```

```
f1_score on training set: 0.100
f1_score on testing set: 0.099
perc_diff: 0.3 %
```

Train Data	
WNV Present	Percentage
0	95%
1	5%



Model 2:

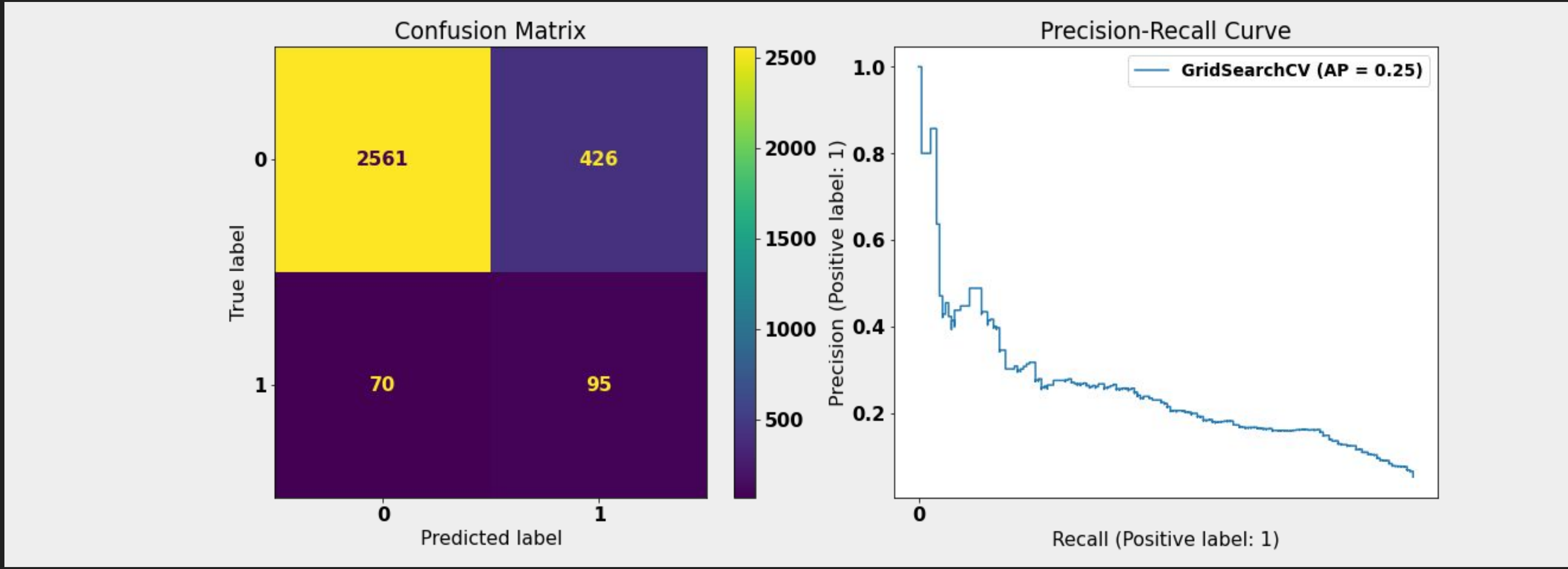
OverSampling + UnderSampling + GradientBoostingClassifier

```
precision_recall_auc_score on training set: 0.281
precision_recall_auc_score on testing set: 0.255
perc_diff: 9.3 %
```

(from 0.05 to 0.25)

```
f1_score on training set: 0.342
f1_score on testing set: 0.277
perc_diff: 19.0 %
```

(from 0.09 to 0.27)



Model 7:

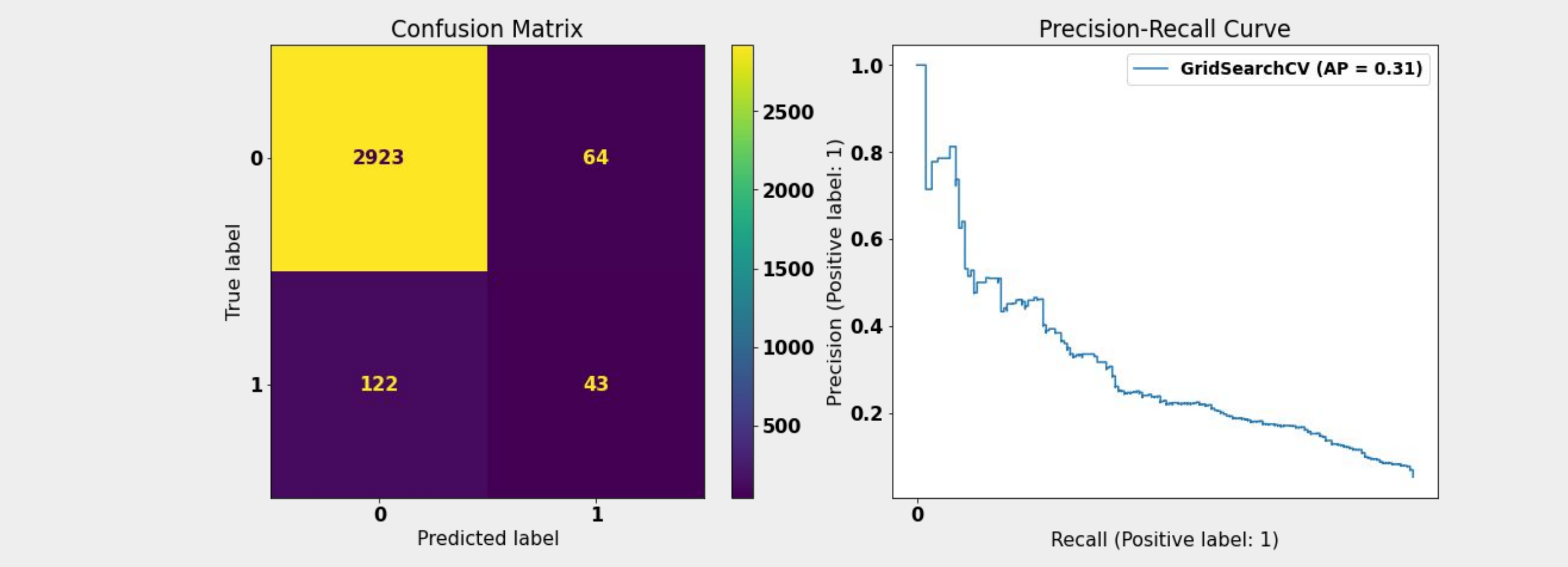
SMOTE + LGBMClassifier

```
precision_recall_auc_score on training set: 0.328
precision_recall_auc_score on testing set: 0.307
perc_diff: 6.7 %
```

(from 0.25 to 0.30)

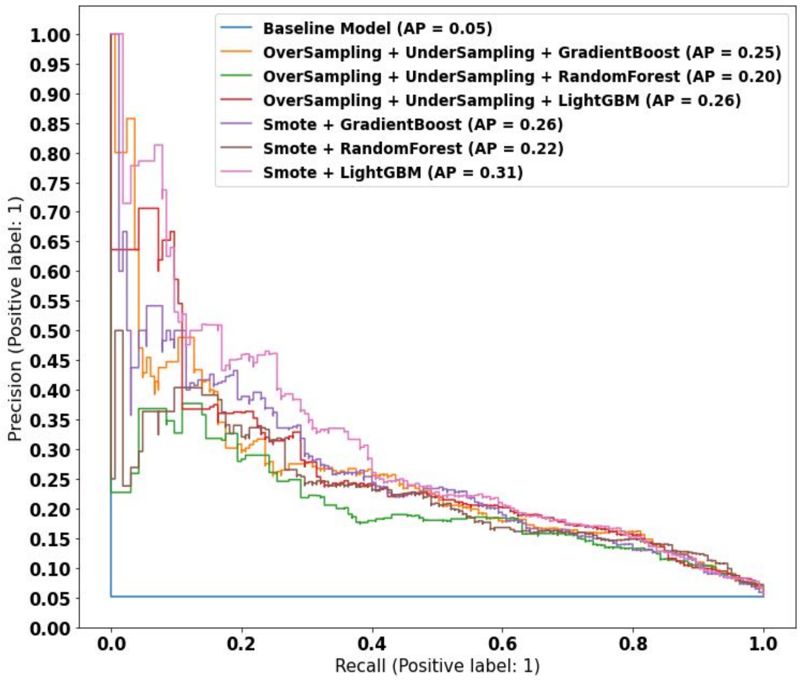
```
f1_score on training set: 0.313
f1_score on testing set: 0.316
perc_diff: 1.0 %
```

(from 0.27 to 0.31)

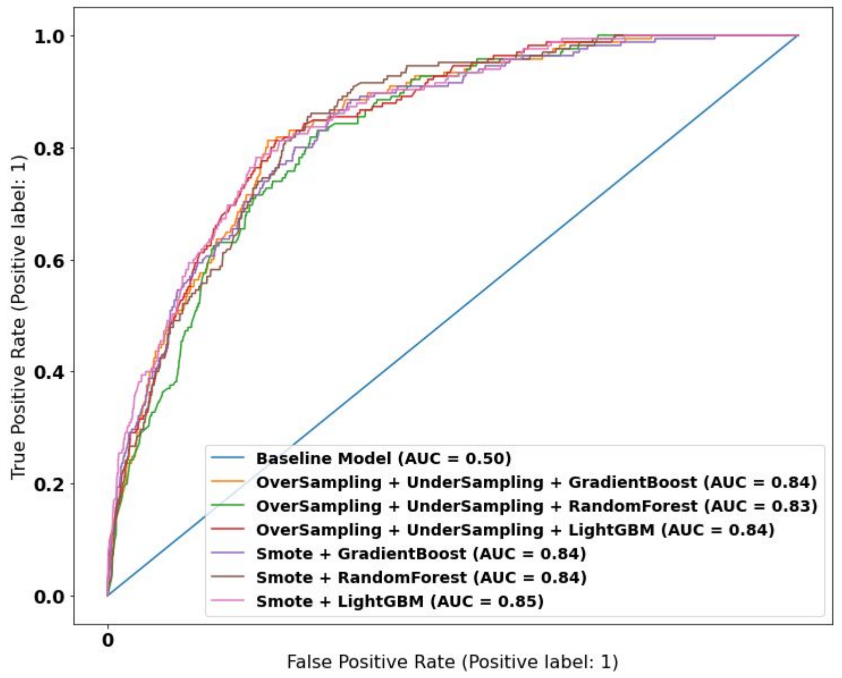


Model Evaluation

Precision-Recall Curve



ROC curve

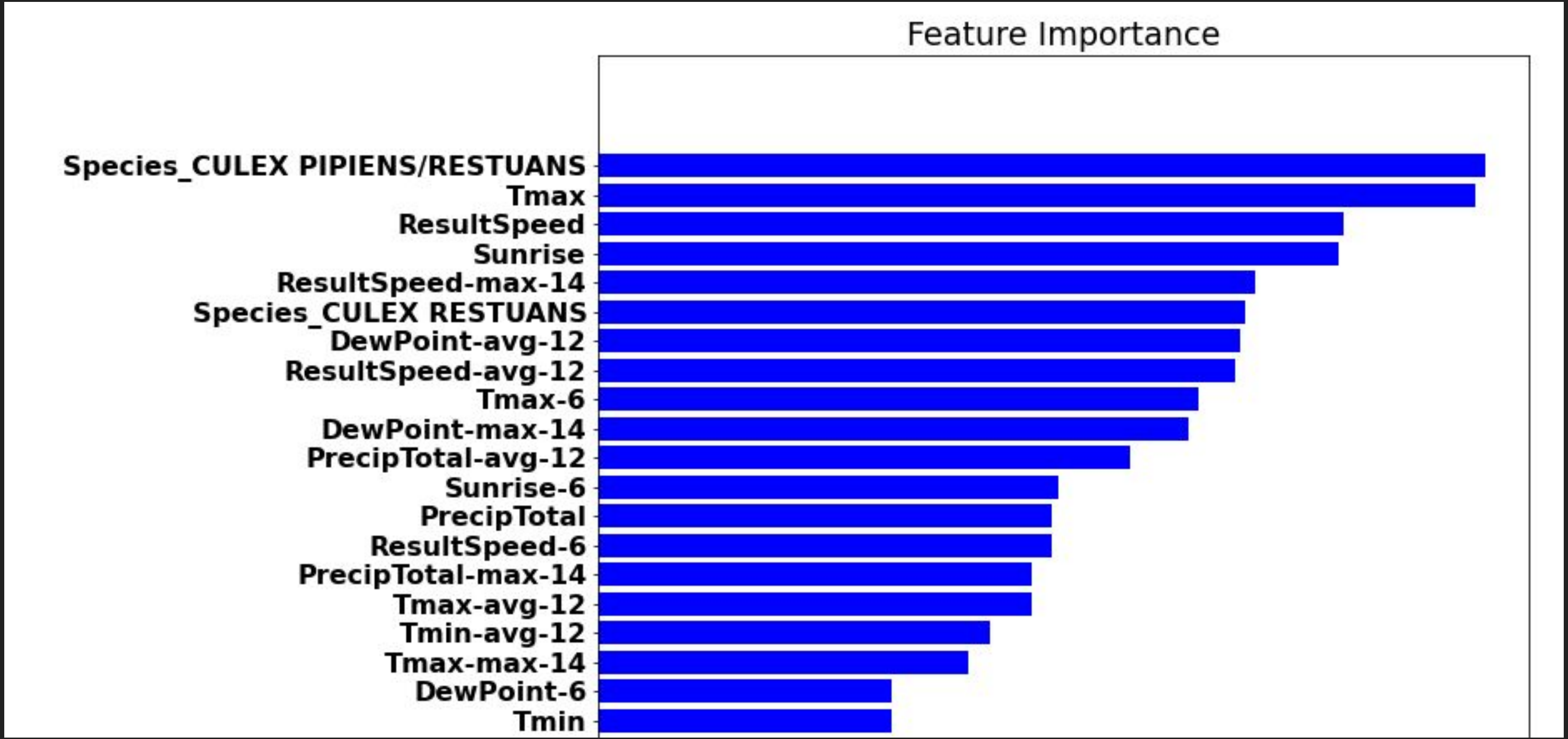


Model Evaluation

Models	PR_AUC_Train	PR_Auc_Test	Generalization	F1_train	F1_test	Generalization
Baseline Model	0.05	0.05	0.27	0.10	0.10	0.25
O/S + U/S + GradientBoost	0.28	0.25	9.33	0.34	0.28	18.95
O/S + U/S + RandomForest	0.21	0.20	4.59	0.17	0.16	5.62
O/S + U/S + LightGBM	0.28	0.26	4.67	0.32	0.29	9.15
Smote + GradientBoost	0.30	0.26	11.83	0.24	0.27	10.82
Smote + RandomForest	0.24	0.22	6.71	0.24	0.24	0.77
Smote + LightGBM	0.33	0.31	6.67	0.31	0.32	0.96

Production Model chosen for best score and generalization

Feature Importance (Top 20)



Cost-Benefit Analysis

Inaccuracy Costs	
Impact of False Positive indication of West Nile Virus	**Impact of False Negative indication of West Nile Virus**
1. Unnecessary Spraying 2. Loss of Productivity of Civil Servants 3. Causes disruption to daily life in affected communities 4. Increased burden on taxpayers	1. Increased proliferation of West Nile Virus disease 2. Increased strain on health care resources due to rise in cases 3. Public Health reputational and political risk

Cost-Benefit Analysis

Economic and Social Costs without Spraying			
Medical and Productivity Costs (included)			Total Costs Before Model
In-Patient cost	$33000/person	39 outpatients	$1,287,000
Out-Patient cost	$6300/person	45 out patients	$283,500
No. of deaths per year (mean of 8 years)	5 deaths/year	65,000/person per year	$3,250,000
Cost of Spraying			
Cost of pesticide spray per acre	1000/acre		
Total Acres being sprayed	1.5 fluid ounces per acre		
Chicago Area	607km2		
Amount of pesticide sprayed	44.4 ml per 0.00405 km2 = 6,667 litres in total		
Cost of Labour to Spray	60 men contracted at $1,000/year		$60,000
Cost of Sprayer Trucks	$200/day for 20 trucks 4 times a year		$16,000
No. of Trucks needed	20 trucks		
Cost of spray pesticide	$55/16 oz of Zenivex		
	$116/ltre		$773,372
			$5,669,872

- The costs of spraying are a fraction of the Medical and Productivity costs (not to mention the lives lost), which makes the effort well worth the financial investment
- Benefit could save up to $4.8M in Medical and Productivity cost with less than $1M investment
- Usage of our model would assist in a more target usage of pesticide spray which could also further reduce costs
- Money saved for the taxpayer could engender more fiscal confidence in public health system

Negative Externalities due to West Nile Virus

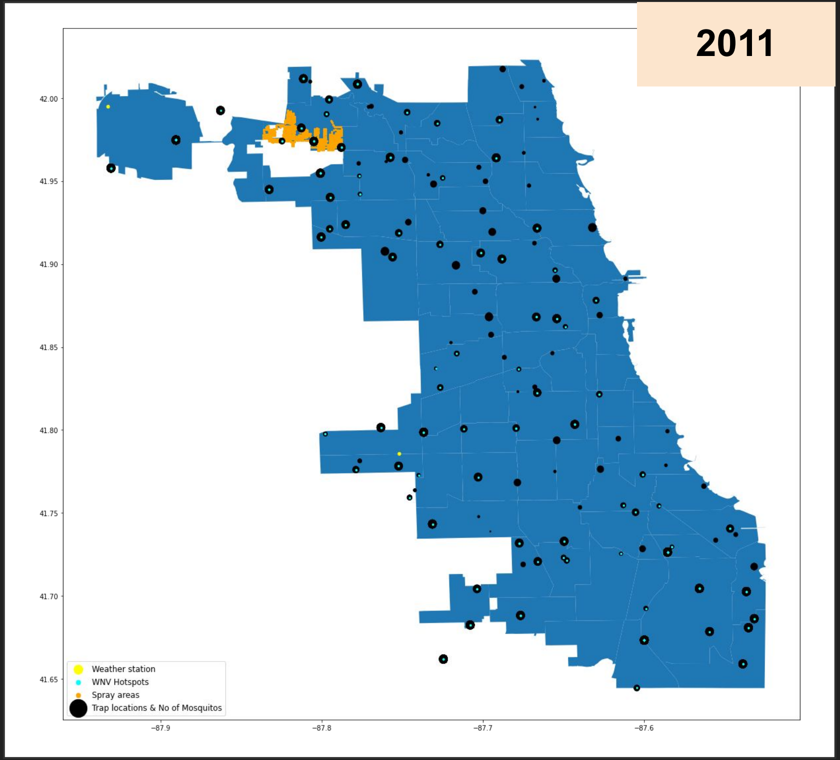
- Work absenteeism
- Public health impact and cost
- Government and Public Health Officials reputational loss
- Impact to families (financial burden, caregiver costs for most vulnerable, etc.)
- Decreased tourism
- Increased death risk amongst population might incur public outrage

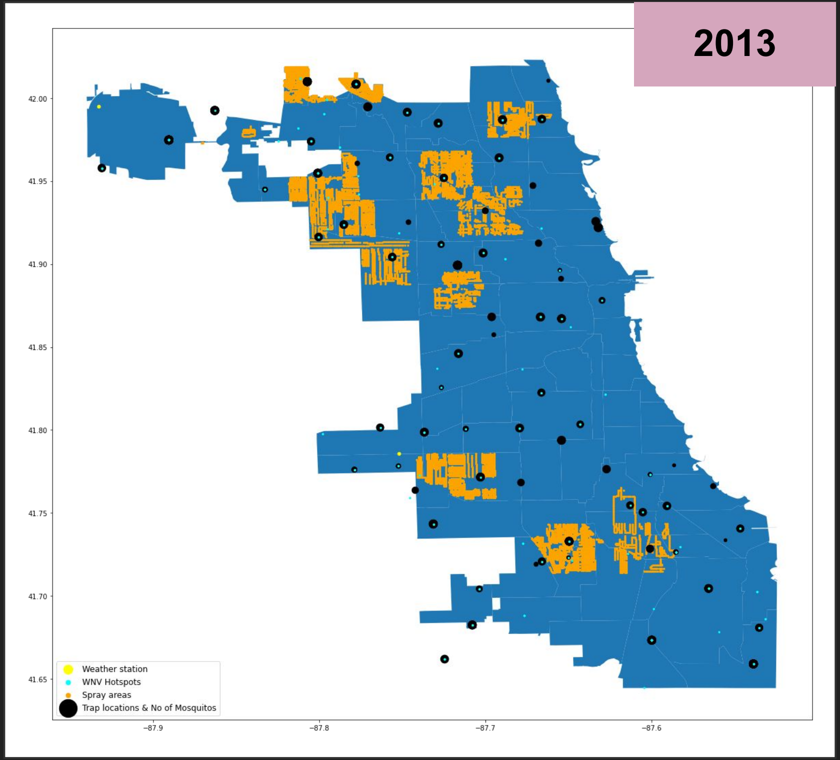
Recommend to : **SPRAY**

Spray Data Analysis

There are a total of 9 spray dates in dataset, 1 in 2011 and 8 in 2013.

Spraying is done indiscriminately, regardless of whether there is a WNV hotspot or not.





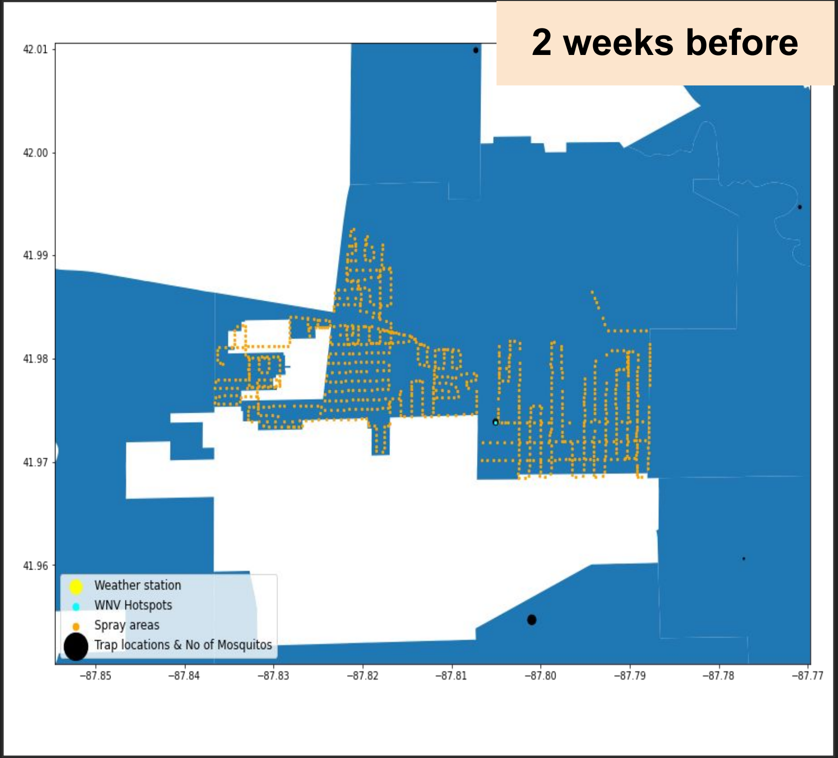
Spray Data Analysis - 7 Sep 2011

To test the effectiveness of the spray, we look at number of mosquitoes two weeks before and after spray (based on life cycle of a mosquito).

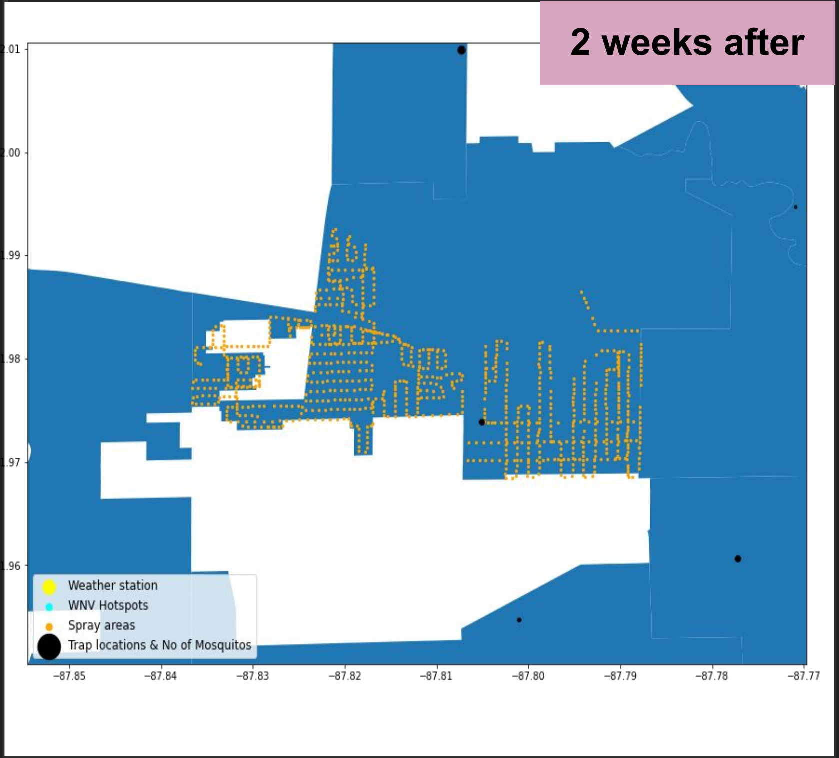
1 WNV Hotspot

0 left

2 weeks before



2 weeks after

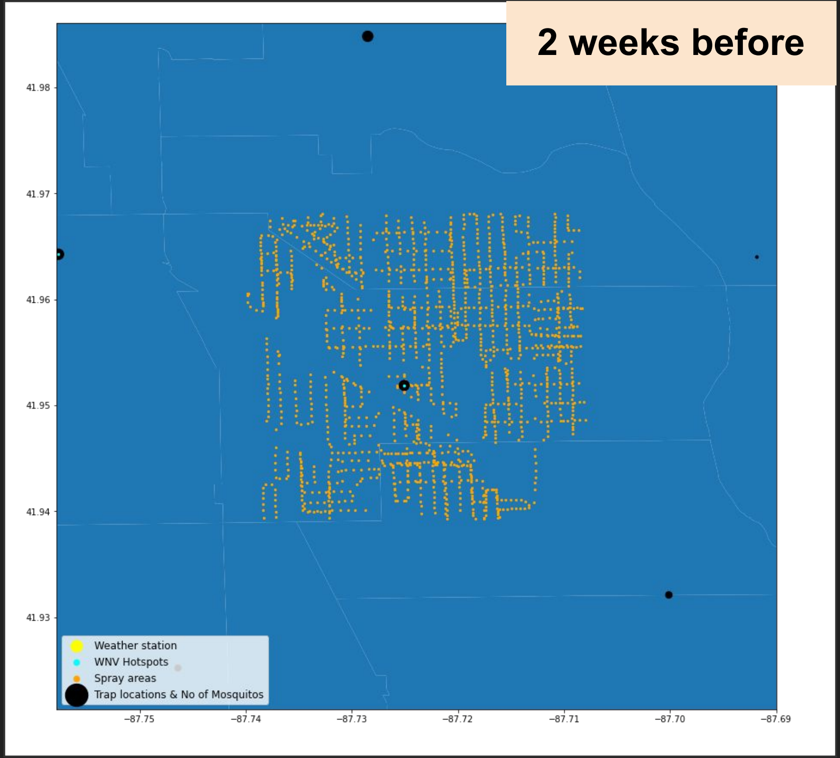


Spray Data Analysis - 25 Jul 2013

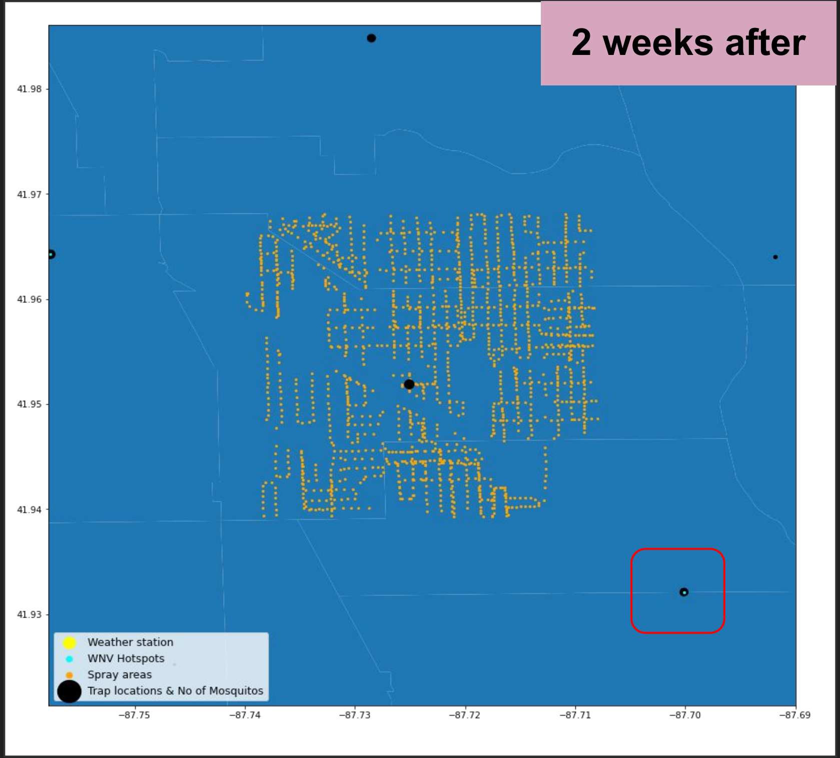
1 WNV Hotspot

0 left

2 weeks before



2 weeks after

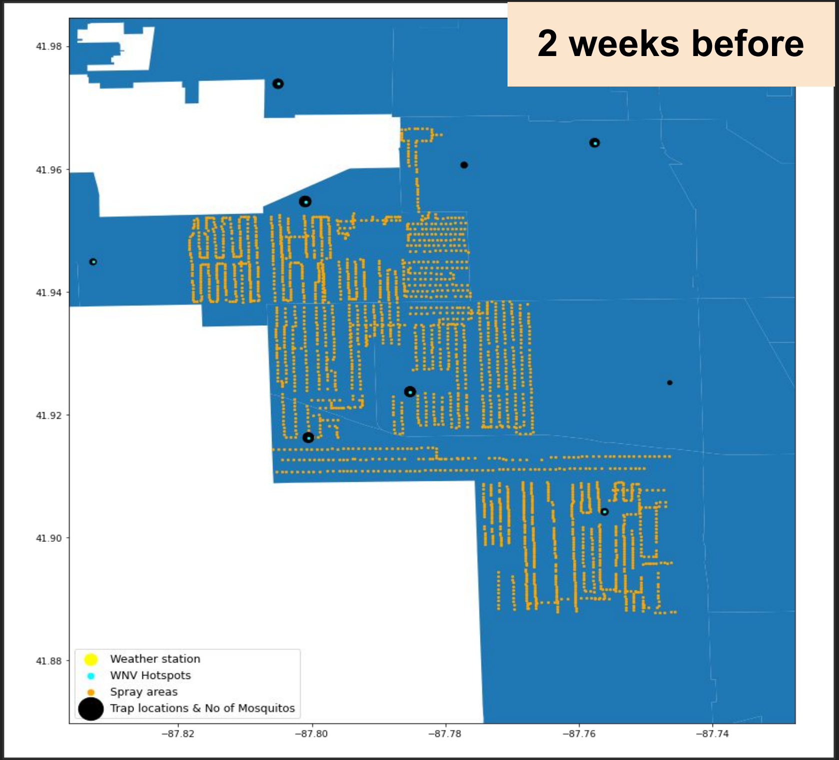


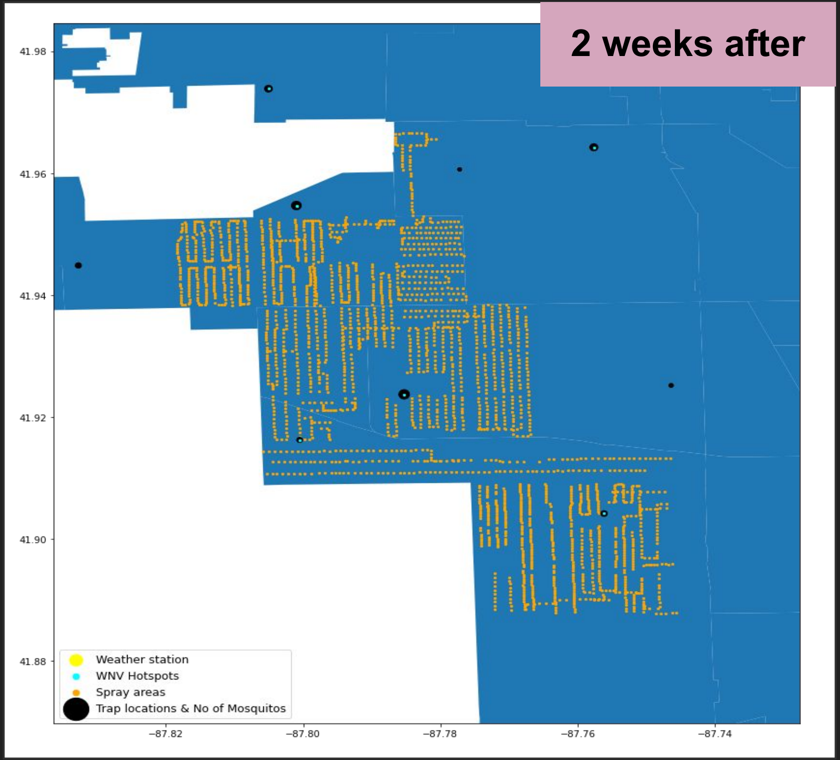
Spray Data Analysis - 15 Aug 2013

Not as effective

4 WNV Hotspot

4 left



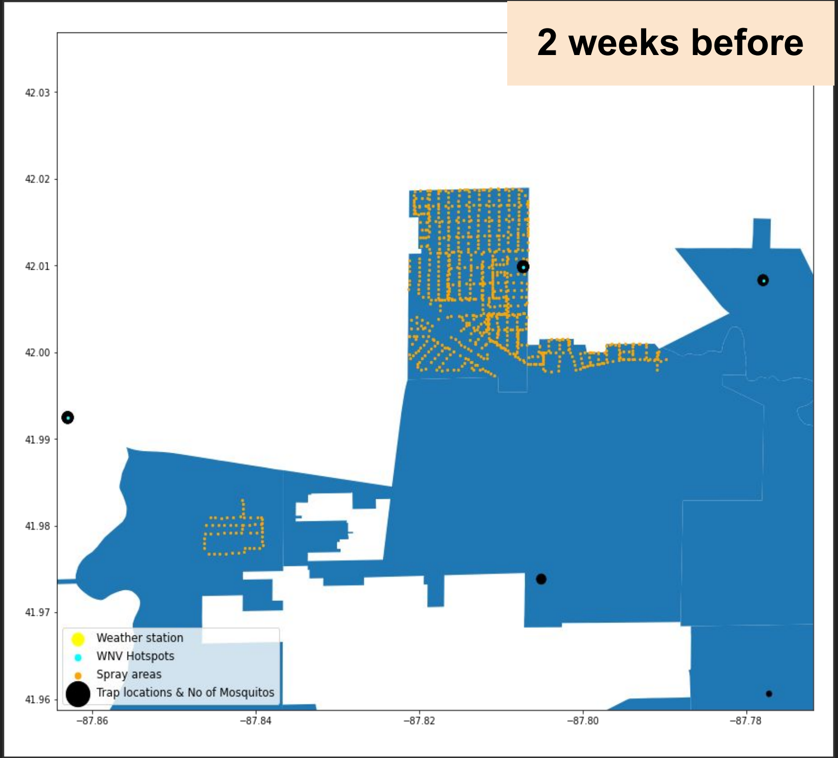


Spray Data Analysis - 5 Sep 2013

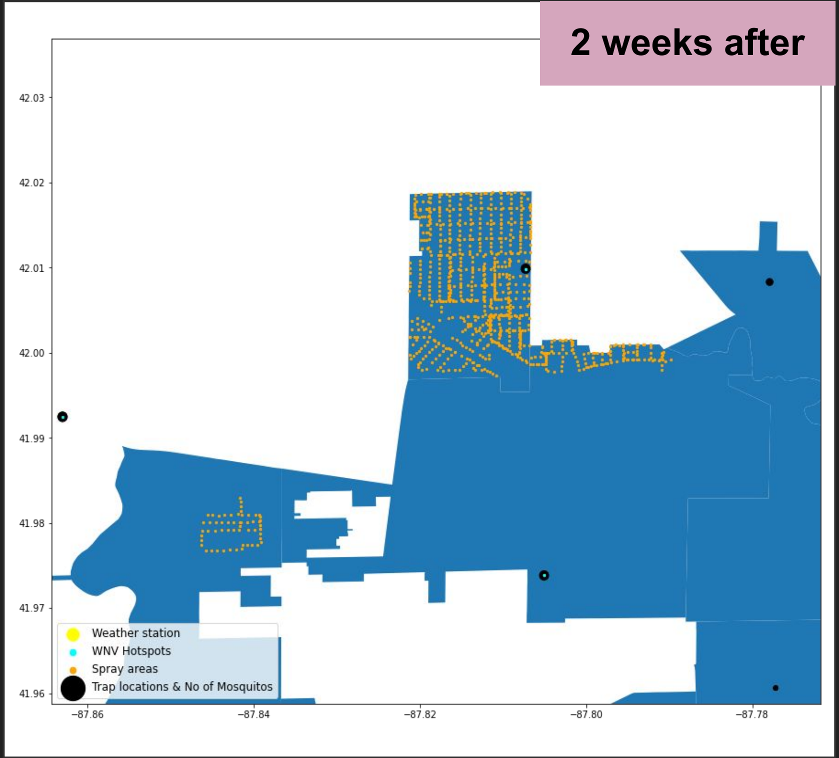
Not as effective

1 WNV Hotspot	1 left

2 weeks before



2 weeks after



Spray Data Analysis - Findings

1. **Spraying done in an ad-hoc manner**
    - Data from 2011 and 2013 seems to suggest that it was done without prior research
    - For e.g. in 16 Aug and 22 Aug 2013, spraying was not done on WNV hotspot areas or areas where trap locations are found
2. **Spray not effective with time**
    - Number of mosquitoes did not drop within spraying area.
    - Effectiveness of spraying seemed to reduce later on in the months, perhaps due to mosquitoes developing resistance to pesticides over time
3. **Spraying not effective in curbing virus**
    - WNV hotspots still remain 2 weeks after spraying
    - Assuming adulticide sprays are applied, which only kills adult mosquitoes, it is not truly effective in reducing virus as mosquito larvae is still alive

Conclusions, Considerations and Recommendations

WNV is more prevalent under certain conditions:

- Longer daylight hours
- Higher average temperatures
- Humid weather

Spraying efforts should be focused during June to early July

- Current spraying efforts are ineffective
- Suggest to spray in early June to July, considering the gestation period of mosquitoes resulting in peak WNV cases in August

Health issues related to spray chemicals

- Pregnant women and children have a greater risk of getting sick from pesticides

Consider different methods / alternatives to spraying

- Consider larviciding catch basins, which involves dropping tablets in storm drains along the public roads which will slowly dissolve over a five-month period to prevent mosquito larvae from hatching
- Eliminating standing water by ensuring that swimming pools and construction sites are regularly maintained

Q & A