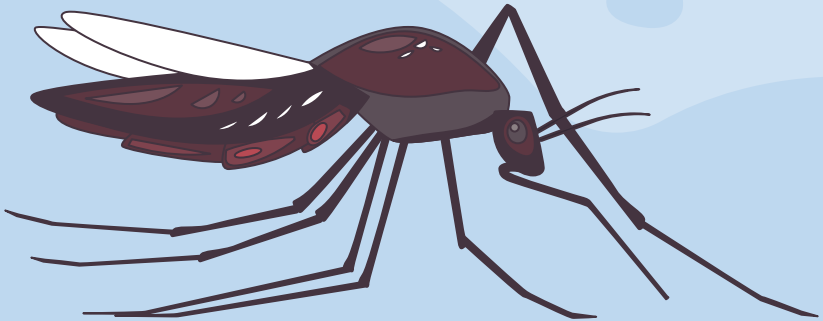
Project 4 : Kaggle - Chicago West Nile Virus

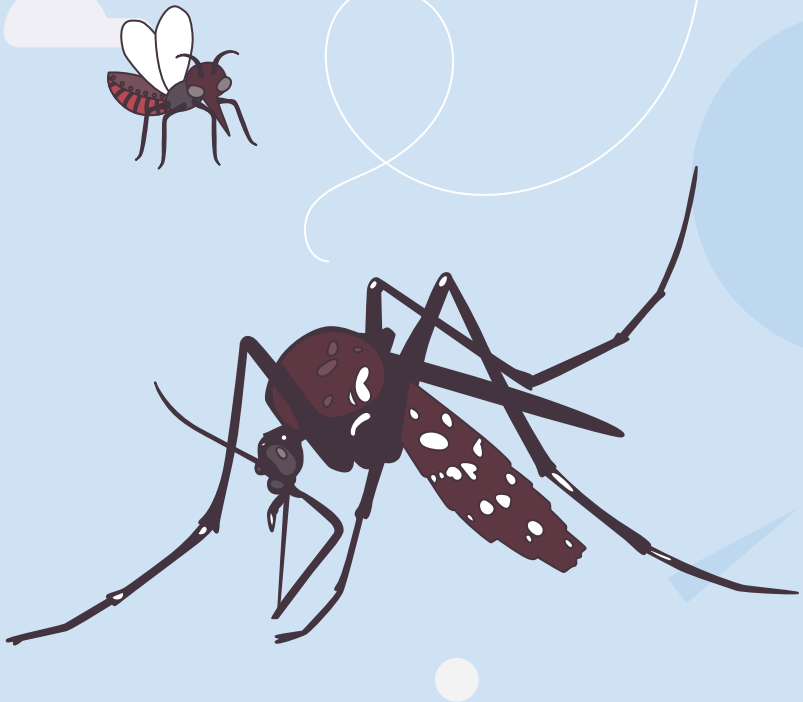


Group 3
DJ | Nazira | Sean | Shuyi

Introduction

Problem Statement

Ultimate aim
Prevent transmission of the mosquito-borne West Nile Virus.



01

Data Cleaning

Data Cleaning

- Break down date column into years, months and weeks
  - Observe any seasonality
- Calculate number of traps set and total number of mosquitoes caught based on year and month
- Identify any duplicate traps

```
# Which addresses is trap T009 associated with?
train[train['trap']=='T035']['address'].value_counts()
```

```
5100 West 72nd Street, Chicago, IL 60638, USA      45
3000 South Hoyne Avenue, Chicago, IL 60608, USA    27
Name: address, dtype: int64
```

```
# Which addresses is trap T009 associated with?
train[train['trap']=='T009']['address'].value_counts()
```

```
9100 West Higgins Road, Rosemont, IL 60018, USA    80
9100 West Higgins Avenue, Chicago, IL 60656, USA   31
Name: address, dtype: int64
```

- Combining duplicate records (for traps capturing more than 50 mosquitoes)

Data Cleaning - Weather

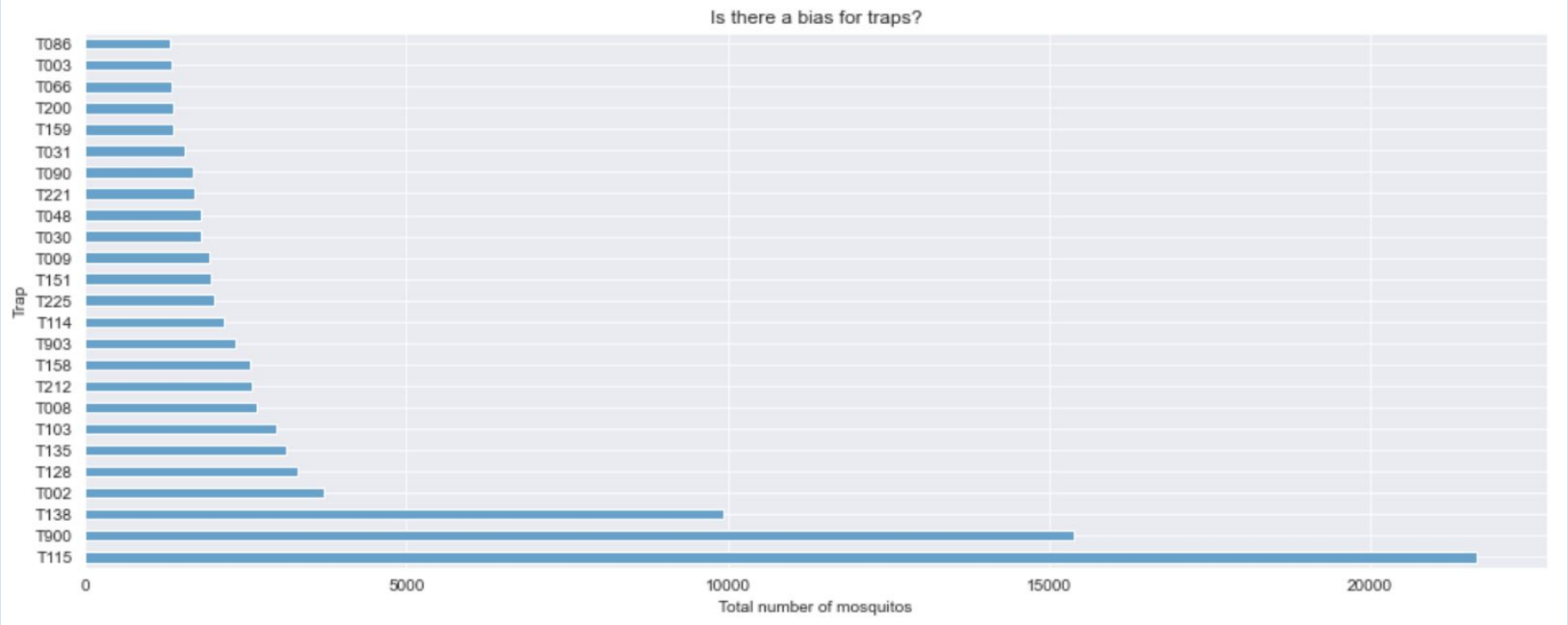
- Imputing all missing values with *np.nan*
- One-hot encoding ‘codesum’ values as there are more than 10 categories
  - Create dummy variables for different conditions
- Tidying ‘stnpressure’ values with median values
- Impute median values from different weather conditions for missing ‘sealevel’ and ‘avgspeed’ values

02

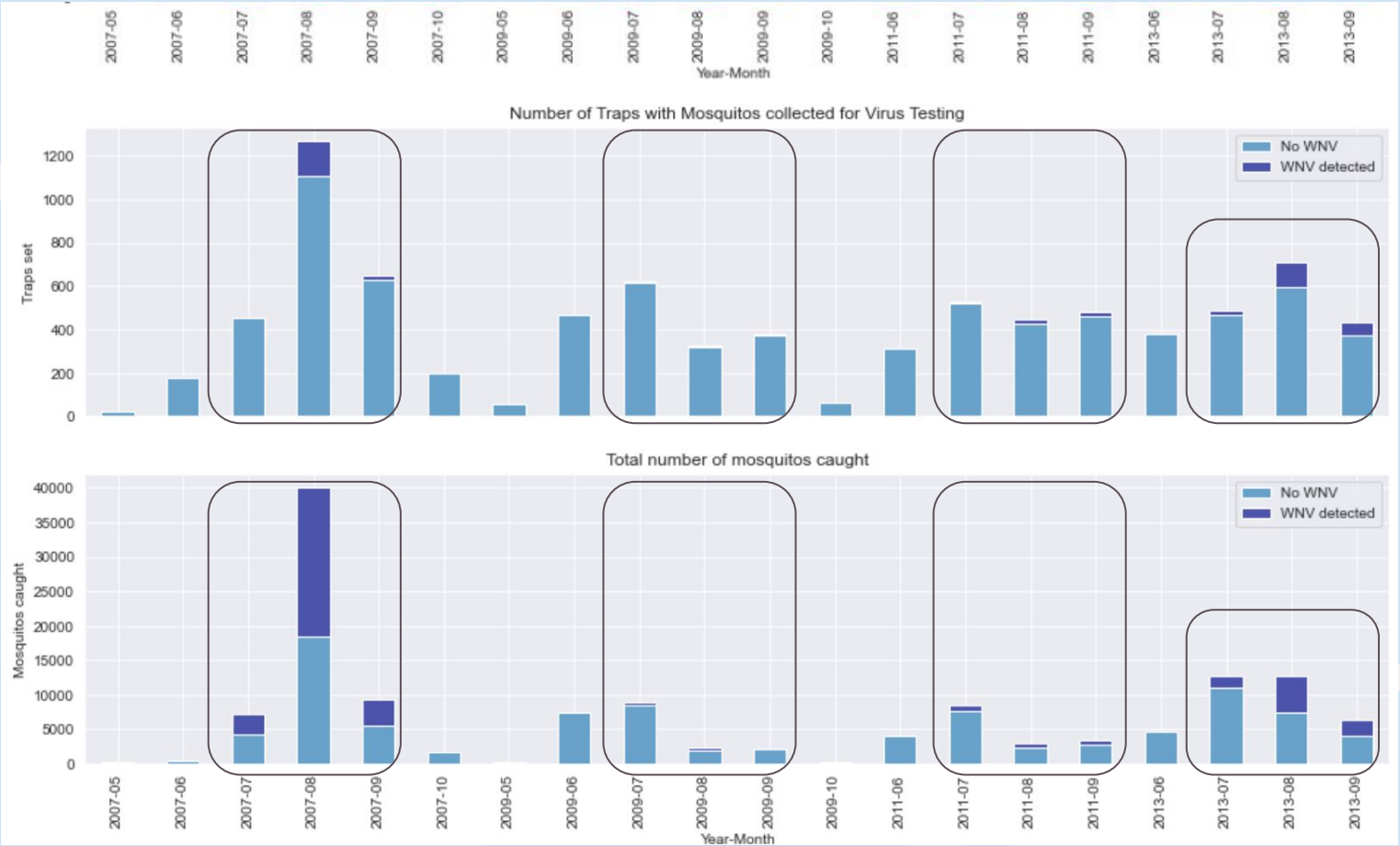
Exploratory Data Analysis

TRAIN

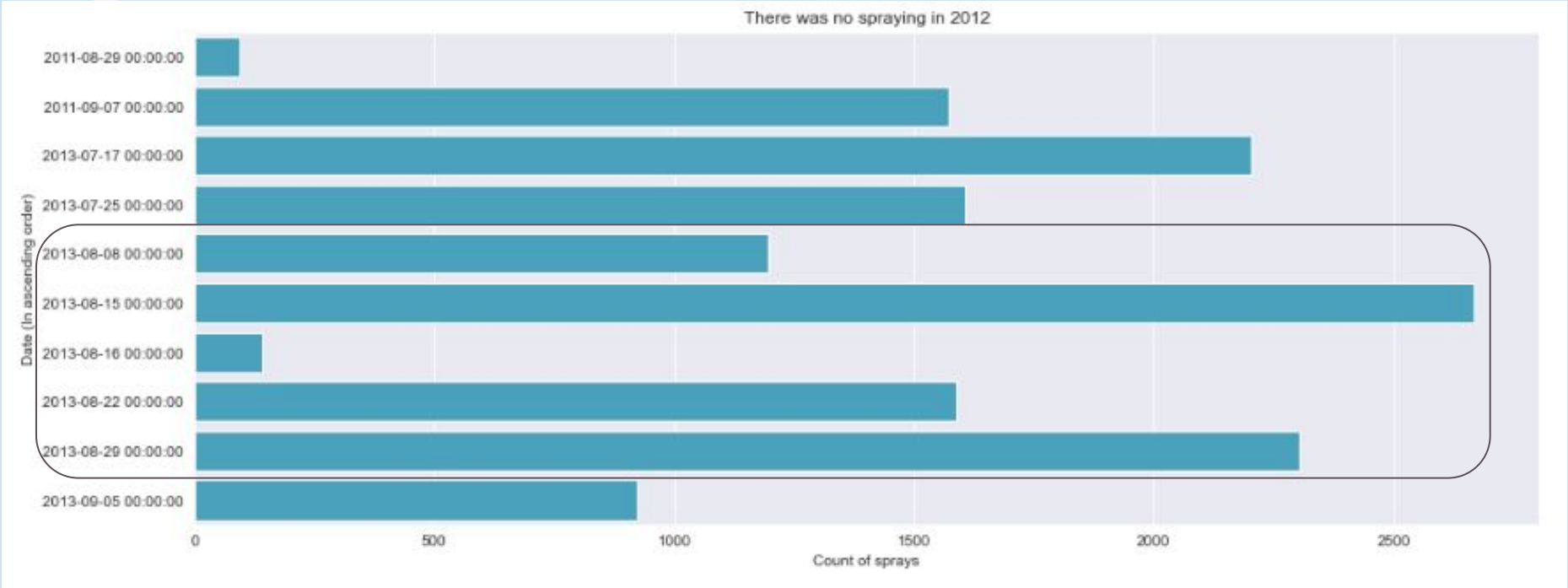
Counting Mosquitoes against Trap Location



TRAIN

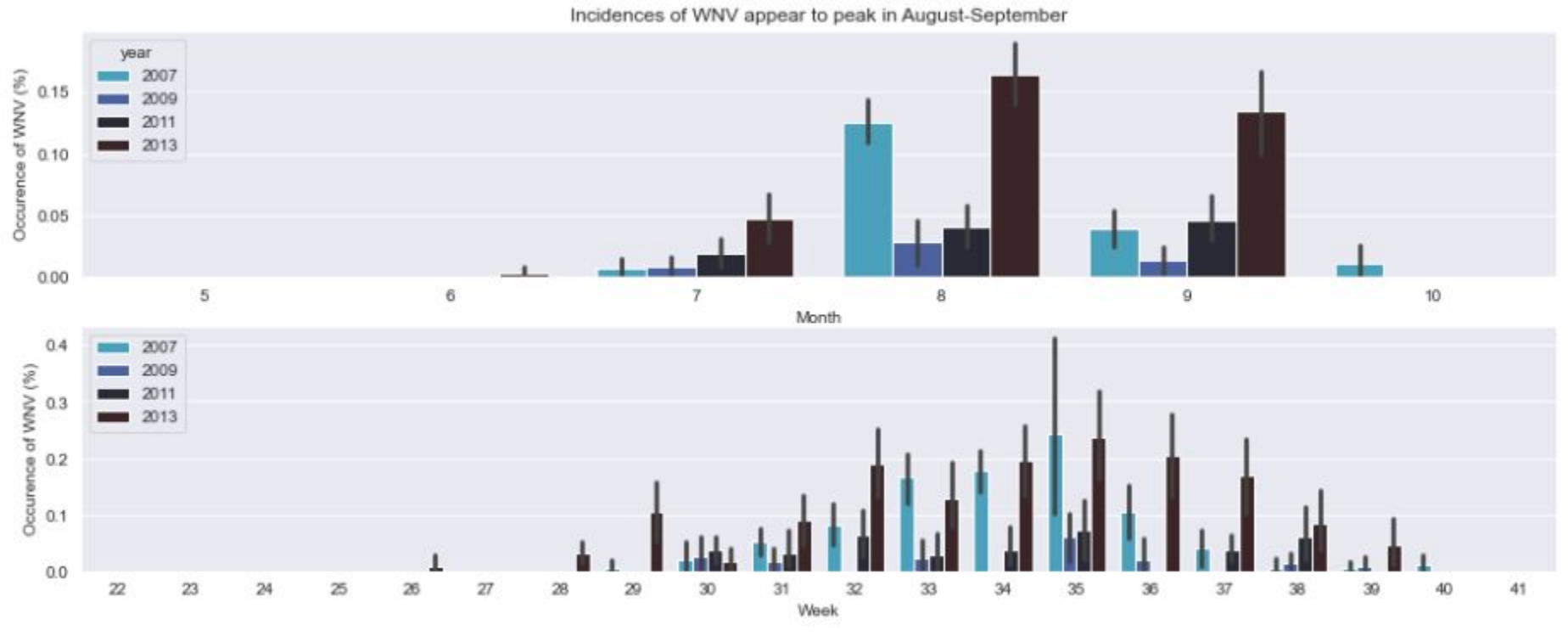

2007-05
2007-06
2007-07
2007-08
2007-09
2007-10
2009-05
2009-06
2009-07
2009-08
2009-09
2009-10
2011-06
2011-07
2011-08
2011-09
2013-06
2013-07
2013-08
2013-09
Year-Month
Number of Traps with Mosquitos collected for Virus Testing
Traps set
0
200
400
600
800
1000
1200
No WNV
WNV detected
Total number of mosquitos caught
Mosquitos caught
0
5000
10000
15000
20000
25000
30000
35000
40000
No WNV
WNV detected
2007-05
2007-06
2007-07
2007-08
2007-09
2007-10
2009-05
2009-06
2009-07
2009-08
2009-09
2009-10
2011-06
2011-07
2011-08
2011-09
2013-06
2013-07
2013-08
2013-09
Year-Month

Spraying Trends

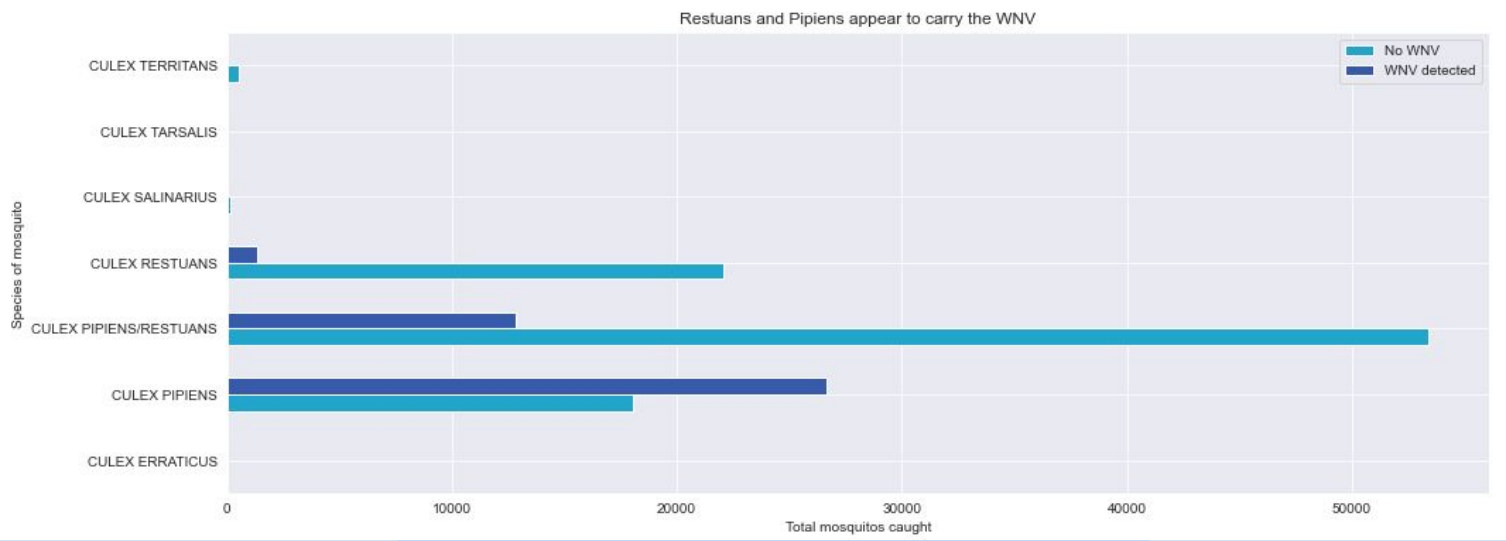


TRAIN

WNV Trends



Significance of Mosquitoes

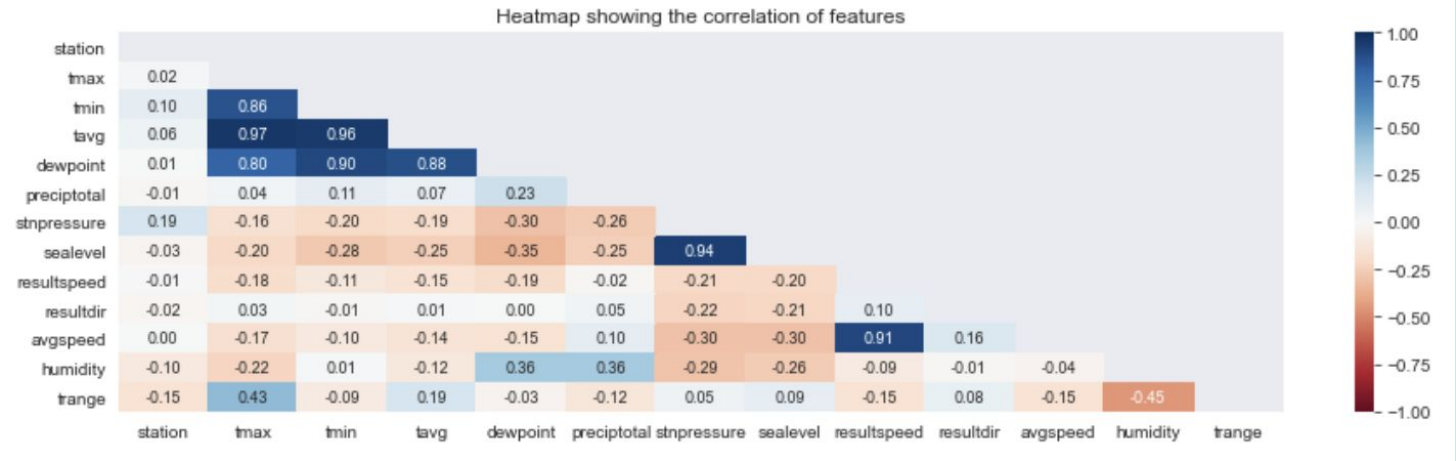



```
test['species'].unique()
```

```
array(['CULEX PIPIENS/RESTUANS', 'CULEX RESTUANS', 'CULEX PIPIENS',
       'CULEX SALINARIUS', 'CULEX TERRITANS', 'CULEX TARSALIS',
       'UNSPECIFIED CULEX', 'CULEX ERRATICUS'], dtype=object)
```

- Drop ‘unspecified culex’

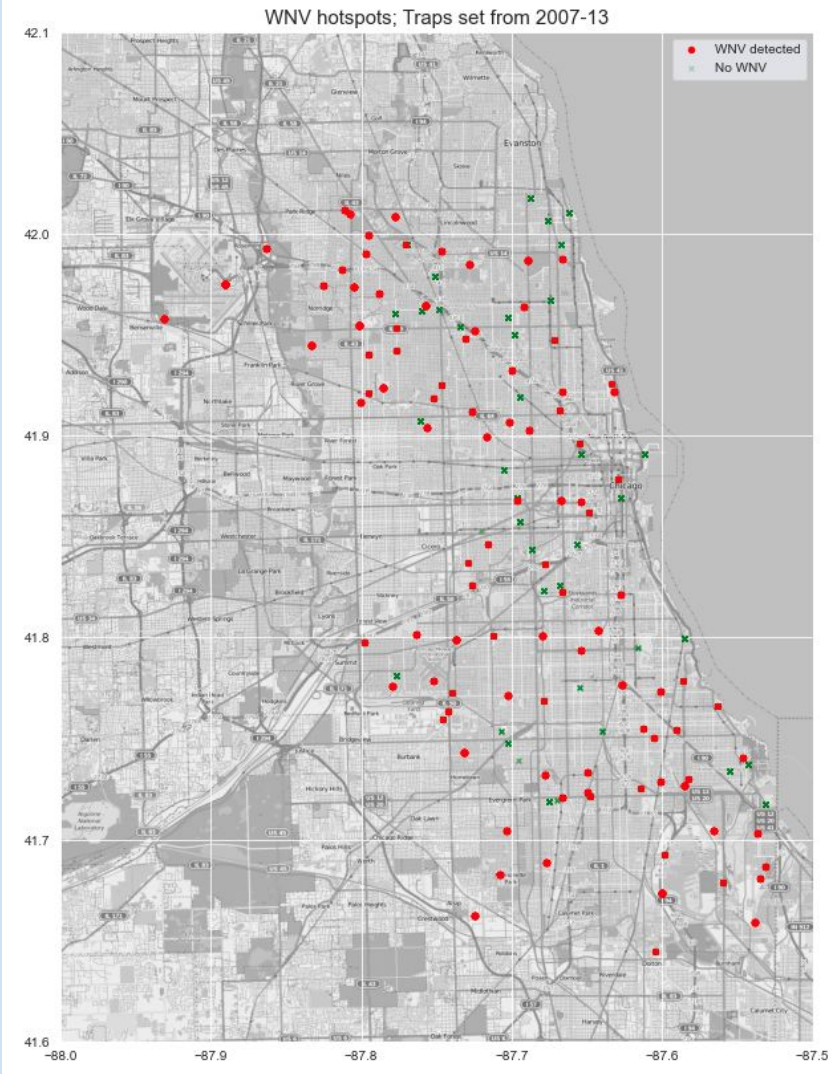
Addressing Multi-collinearity

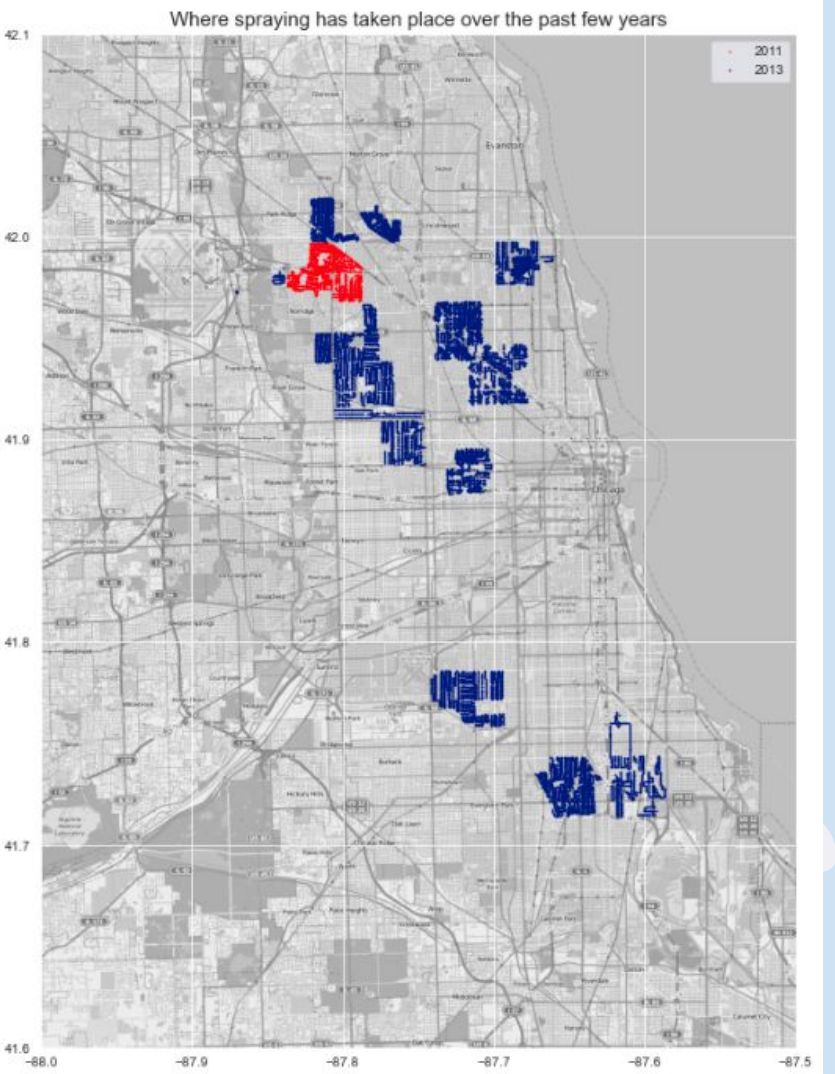


- Dropping features such as ‘tmax’, ‘tmin’ and ‘dewpoint’ → highly correlated with ‘tavg’
- ‘Sealevel’ is highly correlated with ‘stnpressure’
- ‘Avgspeed’ is highly correlated with ‘resultspeed’

Zooming in to 'WnvPresent'

- After doing Polynomial Features, there were 49 features.
- Top 5 positively correlated
  - Mosquito Breed (Pipiens)
  - Zooming in to specific week(s)
  - Time period (month)
  - Humidity
  - Time lag
- Top 5 negatively correlated
  - Temperature range
  - Station
  - Mosquito Breed (Restuans)
  - Average Wind Speed
  - Time Lag


WNV hotspots; Traps set from 2007-13
WNV detected
No WNV
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


Where spraying has taken place over the past few years
2011
2013
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

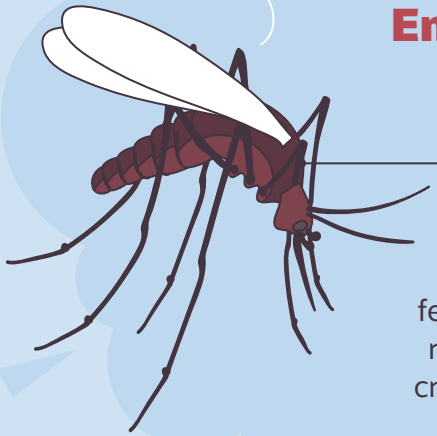
Feature Engineering

- Creating more features:
  - Humidity
  - Interaction between Temperature and Total Precipitation
  - Lagging weather measurements (5/10-day rolling averages)
    - Averaging: Temperature (Min and Max), Dewpoint, Station Pressure, Sea Level, Average Speed and Humidity
    - Sum: Precipitation Total, Raining, Misty, Daylight Minutes
  - Temperature range
  - Daylight Exposure

03

Pre-processing & Modeling

Breaking our work flow down



Feature Engineering

New weather features (rollsum, rollavgs) cluster creation (DBScan)

Pipeline

Standard Scaler, Dummifying data, SMOTE

Building models

LogReg
Extra Trees
Decision Trees
Random Forest
SVC
Gradient Boost
XGBoost
ADABoost

Model selection

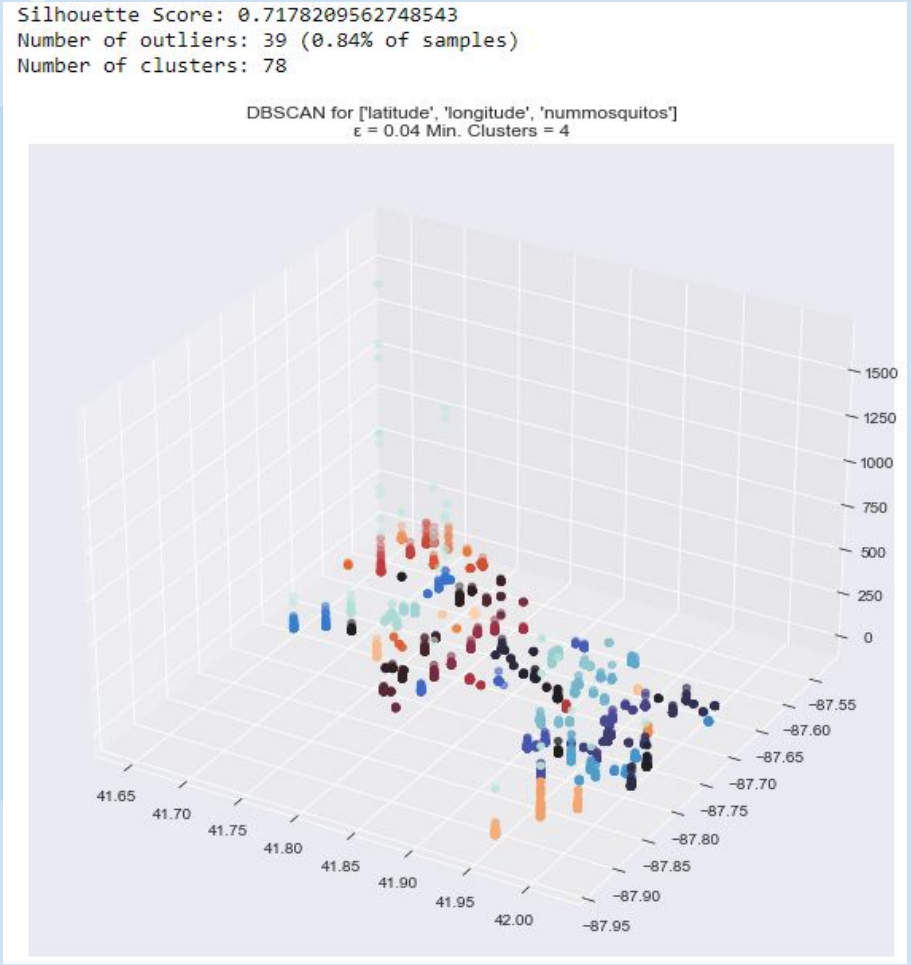
Hyperparameter tuning;
Model insights

Feature engineering

Future considerations

- **Trap bias:** There could be a bias for certain traps (Given that a particular trap may always capture more mosquitos, etc)

- **Time clustering**: Outbreaks will possibly influence the days before and after it (They tend to cluster in the dimension of time too)



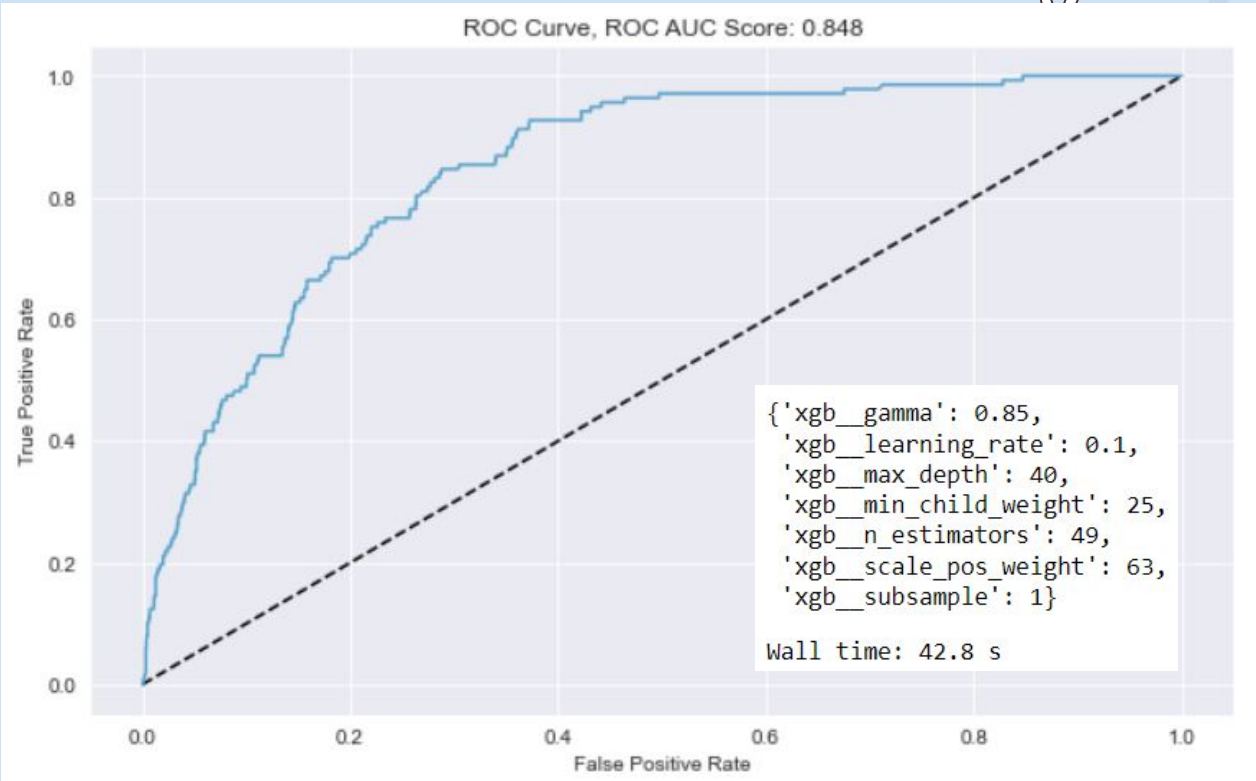
Pipeline; Building our models

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

Class Balance BEFORE

0.0    0.9459
1.0    0.0541
Name: wnvpresent, dtype: float64

Number of rows: 5915

Class Balance AFTER

0.0    0.5
1.0    0.5
Name: wnvpresent, dtype: float64

Number of rows: 11190
```

	model	train_auc_cv	f1	recall	precision	train_auc	test_auc	auc_diff
0	gb	0.981540	0.261780	0.547445	0.172018	0.984820	0.812343	0.172477
1	svc	0.985760	0.240876	0.240876	0.240876	0.990639	0.772468	0.218171
2	ada	0.956751	0.239658	0.613139	0.148936	0.959185	0.791350	0.167834
3	xgb	0.993239	0.217617	0.153285	0.375000	0.998321	0.838572	0.159749
4	lr	0.821628	0.180498	0.635036	0.105200	0.827767	0.739272	0.088494
5	et	0.988801	0.165049	0.124088	0.246377	0.999937	0.770904	0.229033
6	rf	0.997149	0.164948	0.116788	0.280702	0.999936	0.816768	0.183168
7	dt	0.955336	0.153846	0.138686	0.172727	0.999937	0.567881	0.432056

```
Wall time: 6min 16s
```

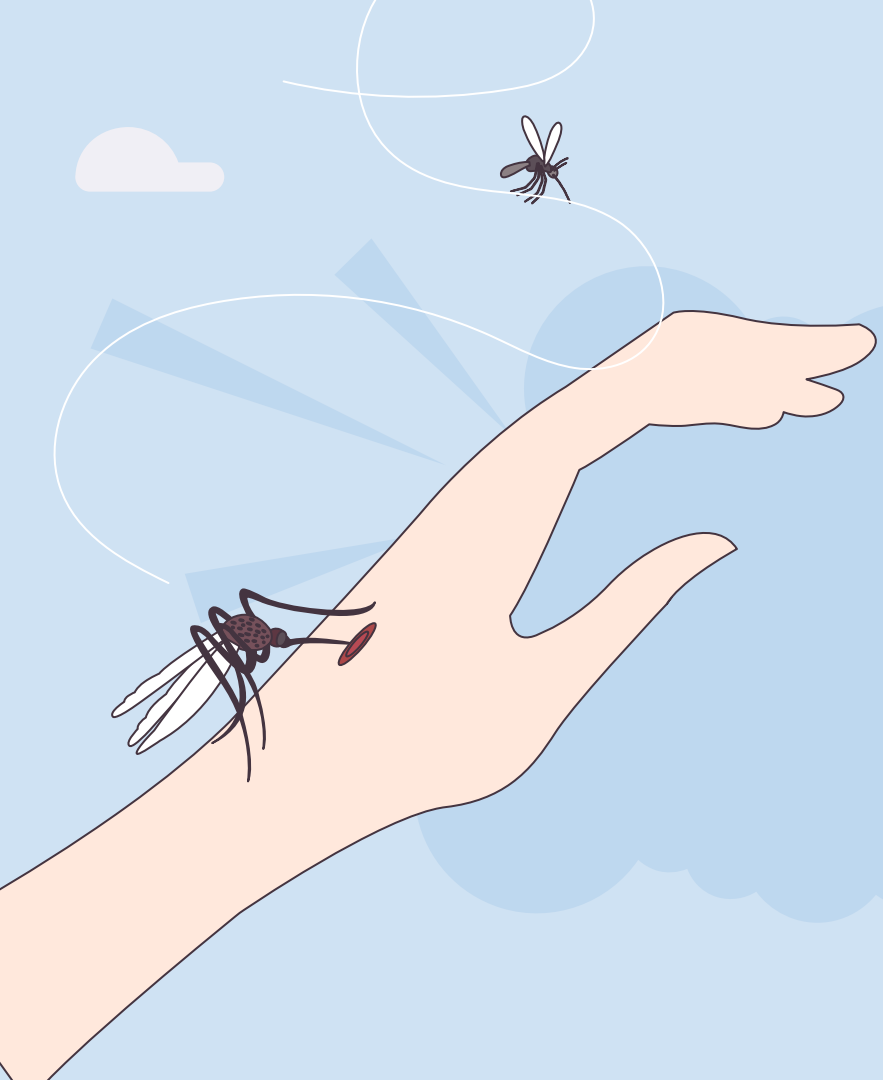
Model selection; Hyperparams tuning



	Features	Importances
1	week	0.130479
0	month	0.053143
33	cluster_31	0.033728
80	PIPIENS	0.031945
6	cluster_4	0.031897
25	cluster_23	0.031414
35	cluster_33	0.030620
62	cluster_60	0.025945
87	sealevel	0.025865
75	cluster_73	0.025526
9	cluster_7	0.024153
20	cluster_18	0.023776
88	resultspeed	0.023362
63	cluster_61	0.022122
43	cluster_41	0.021143

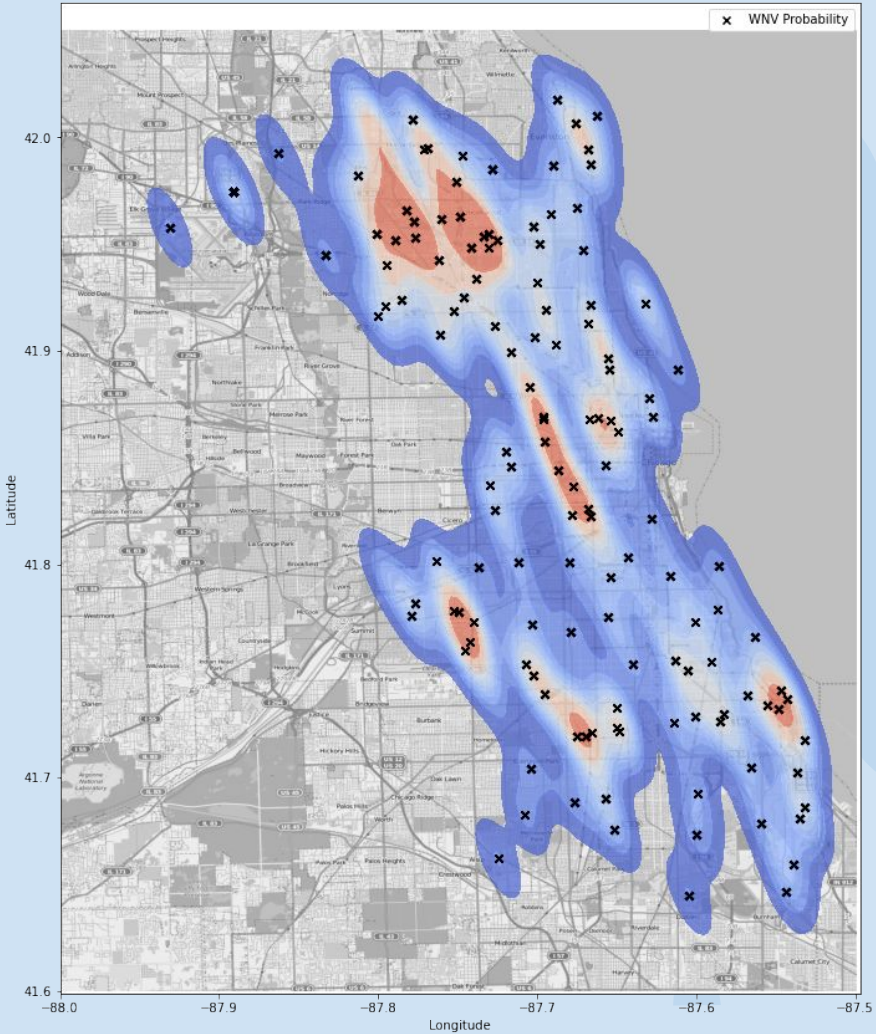
04

Cost Benefit Analysis

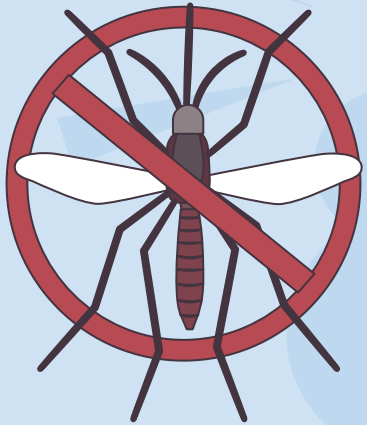


Risk of WNV visualized

Locations with Risk of WNV



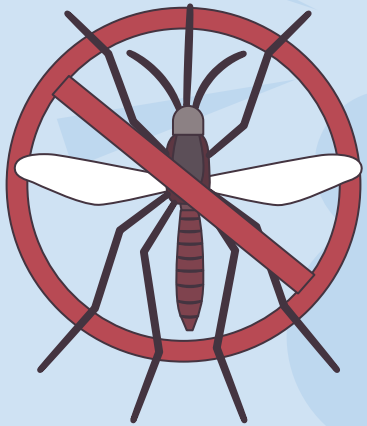
Spraying Entire Chicago



$100,366

Adulticide (Zenivex E4) at USD 0.67 per acre*
Multiplied by
Entire Area of Chicago at 149,800 acres

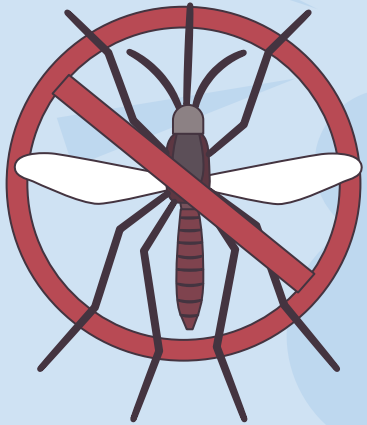
Spraying target areas



$xxx,xxx

Adulticide (Zenivex E4) at USD 0.67 per acre*
Multiplied by
Entire Area of Chicago at 149,800 acres

Approximate Medical Cost



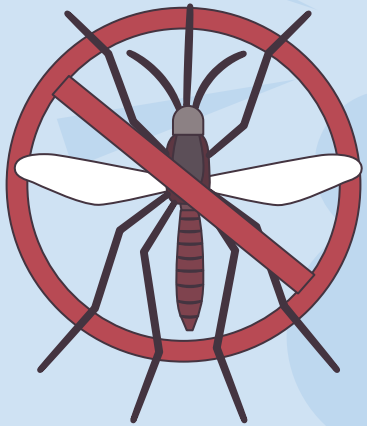
$187,500

of treating 15 Cases of WNV

Acute symptoms:
1. Initial medical cost: $25,000.
2. Long term medical cost: $22,000.

Fever related symptoms:
1. Initial medical cost: $7,500.

Approximate Loss of Productivity Cost

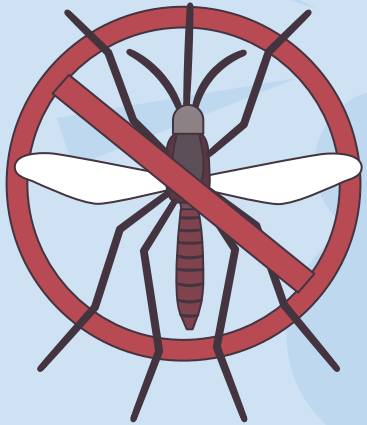


$143,250

of 15 patients unfit for work

For patients <60 y: USD 191 per day
Assuming 2 acute cases, and combined total of 750 days of sick days

Savings



$230,384

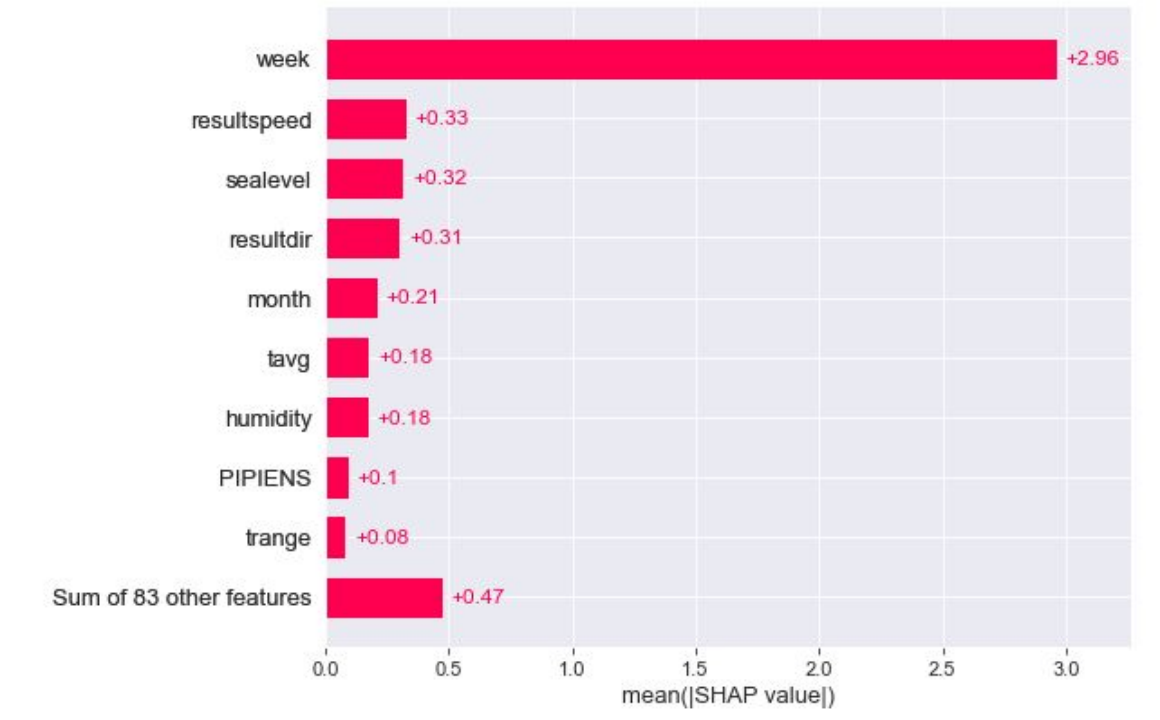
In the scenario that one blanket spray of adulticides prevents 15 cases

05

Recommendations

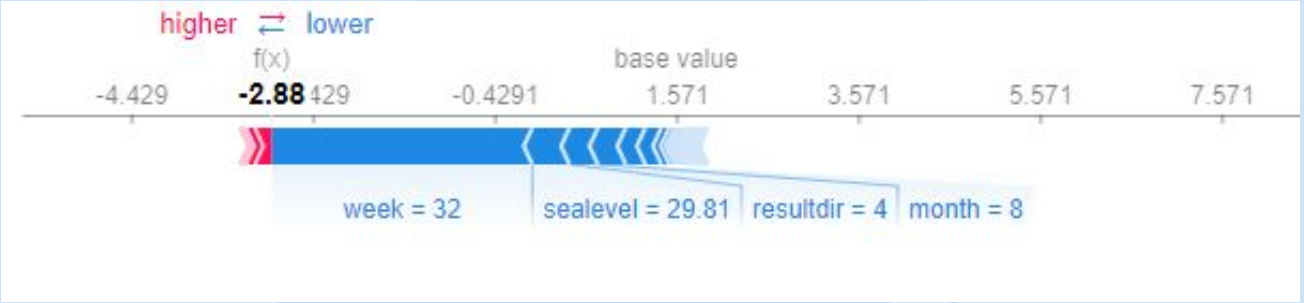
Recommendations

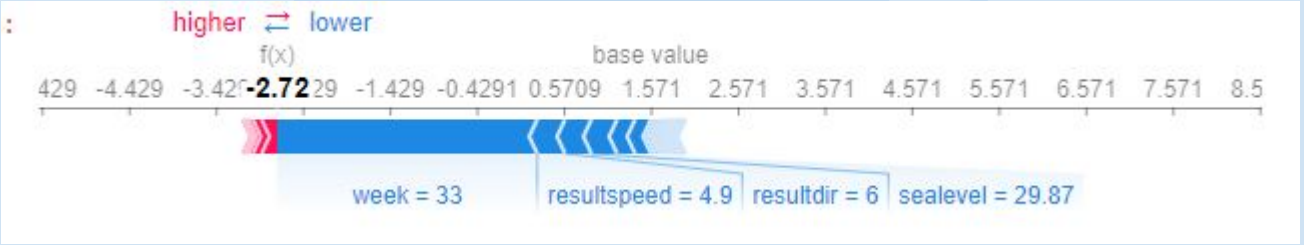
Using SHAP to identify model decision making:



Recommendations

Using SHAP to identify model decision making:





Recommendations

Targeted spraying:

1. From the risk map visualization, we can see that there are clusters where WNV are higher occurring. These locations should be targeted before they start to spike.

2. It might be efficient to start spraying sites based on seasonal spikes in July-August before WNV peaks in August and September.

3. Monitor wind speed to identify periods of slower wind speeds as it is likely to mean that mosquitoes tend to travel over a larger area during this period.

4. Control mosquito breeding grounds through surveillance and education.

06

Conclusion

Conclusion

1. The final model ran to predict WNV risk was
    a. XGBoost

Using ROC AUC score metric for model evaluation, we managed to achieve a score of 0.84.

This score indicates a high level of class separability and shows that the probability of making a correct class prediction is high.

Our model also successfully allowed us to identify patterns for targeted insecticide spraying for cost savings and efficiency.

Thanks

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

List of references (West Nile Virus)

1. The drivers of West Nile virus human illness in the Chicago, Illinois, USA area: Fine scale dynamic effects of weather, mosquito infection, social, and biological conditions

Surendra Karki,William M. Brown,John Uelmen,Marilyn O'Hara Ruiz,Rebecca Lee Smith

Published: May 21, 2020 https://doi.org/10.1371/journal.pone.0227160

2. https://cookcountypublichealth.org/communicable-diseases/west-nile-virus/

3. Chicago NBC News https://www.nbcchicago.com/tag/west-nile-virus/

4. Public surveillance https://www.chicago.gov/city/en/depts/cdph/provdrs/healthy_communities/svcs/report_standing_water.html

5. Hopkins Medicine https://www.hopkinsmedicine.org/health/conditions-and-diseases/west-nile-virus

List of references (other data modelling)

1. https://towardsdatascience.com/a-go-at-kaggle-723447f8d95f

2. https://medium.com/@vijay.swamy1/where-in-chicago-will-the-west-nile-virus-occur-8b6b6d50c94a

3. https://github.com/zql321/DSIFProjects/tree/main/ML%20Prediction%20West%20Nile%20Virus%20Project%204

4. https://github.com/zzeniale/West-Nile-Virus-prediction

5. https://github.com/xbno/DSI-Projects/tree/master/Unassigned%20Project

6. https://github.com/zql321/DSIFProjects/tree/main/ML%20Prediction%20West%20Nile%20Virus%20Project%204

List of references (Cost Benefit Analysis)

1. https://thebottomlinegroup.com/20-cost-saving-ideas-for-the-workplace/

2. https://www.cmmcp.org/pesticide-information/pages/zenivex-e4-etofenprox#:~:text=Zenivex%20is%20an%20insecticide%20that,of%20sunlight%20and%2For%20microorganisms

3. Area of Chicago: 149,800 acres https://www.chicago.gov/city/en/about/facts.html

4. Spray used by Chicago is Zenivex E4
https://www.fox32chicago.com/news/chicago-to-spray-insecticide-to-protect-against-west-nile-virus

5. Cost of Spray
https://www.centralmosquitocontrol.com/-/media/files/centralmosquitocontrol-na/us/resources-lit%20files/zenivex%20cost%20comparison%20fact%20sheet.pdf

6. Medical Costs in Chicago https://www.sciencedaily.com/releases/2014/02/140210184713.htm

7. Population Affected by WNV in Chicago, 2014
https://www.nbcchicago.com/news/local/illinois-reports-first-west-nile-virus-deaths-of-2014/63948/

8. Productivity Costs in Chicago https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/