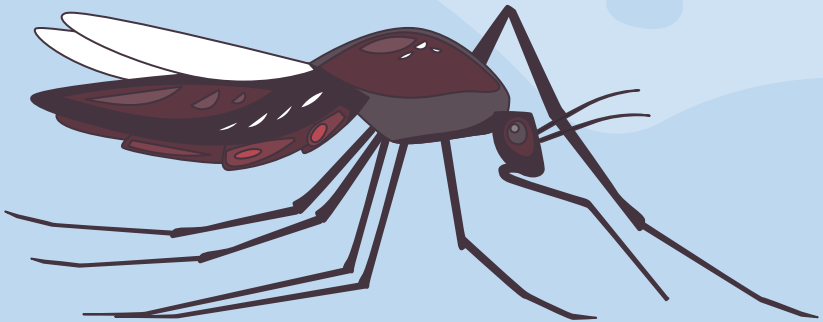
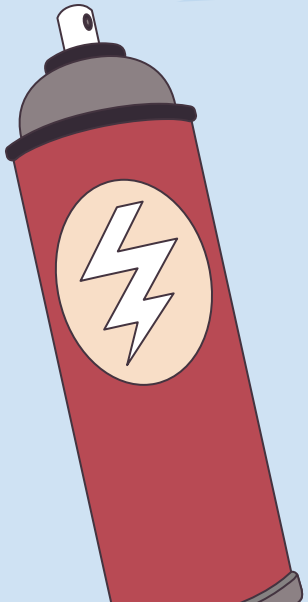
West Nile Virus



Predicting the spread of the virus - a kaggle dataset problem

Ridzuan
Alvin
Mark

Table of contents



01

Introduction

1937

West Nile Virus first appears in Uganda

1999

West Nile Virus makes landfall in USA
in New York City

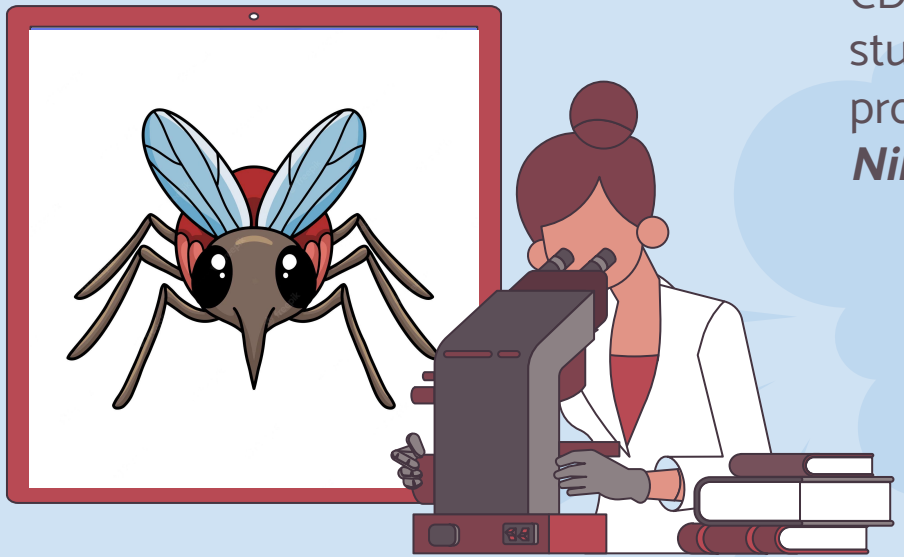
2001

First case in Chicago, IL[1]

[1] https://dph.illinois.gov/topics-services/diseases-and-conditions/west-nile-virus.html

Mosquitoes bad!

Problem Statement



CDPH has contracted us to study the patterns of mosquito propagation, and the ***West Nile Virus***.

Produce **usable insights** to effectively predict the growth of propagation of the virus **through the movement of the mosquito population.**

Our team

Ridzuan
Test Subject

Mark
The Terminator

Alvin
Mad Scientist

Data Summary

Train/Test Data

Train - 2007, 2009, 2011, 2013
Test -2008, 2010, 2012, 2014

Spray Data

When sprays were done in 2011 and 2013

Weather Data

Meteorological data from 2008 to 2014

02

Data Preprocessing & Feature Engineering

Data Treatment

General

- Convert ‘Date’ to Datetime
- ‘Year’, ‘Month’, ‘Day’

Train/Test Data

- Remove duplicate rows (capped at 50)
- Combine ‘*NumMosquitos*’ count for duplicate rows

Spray Data

- Drop ‘Time’ column

Weather Data

- Assign missing values: “M”, “-”, “T”
- Impute Stn 2 missing data from Stn 1
- Drop columns with insufficient data
- Daylight Hours

Data Treatment

Spray Data

- Drop ‘Time’ column

General

Imputing from Station 1 Data	‘Depth’, ‘PrecipTotal’, ‘Snowfall’
Dropping columns with insufficient data	‘Water1’
Filling in Trace "T" data with 0.005	‘PrecipTotal’, ‘Snowfall’
Daylight Hours	‘Sunrise’ + ‘Sunset’

duplicate rows

Weather Data

- Assign missing values: “M”, “-”, “T”
- Impute Stn 2 missing data from Stn 1
- Drop columns with insufficient data
- Daylight Hours

Feature Engineering And Selection

Relative Humidity
Derived from Average Temperature and Dewpoint

Cyclical Transform of Month and Day
Makes more sense for cyclical variable

CodeSum split
Nominal columns for each classification

Clustering
Latlong, resultspd result dir, Tavg Rhumid

Coordinate Tupling
Grouping Latitude/Longitude

A measure of the amount of **moisture** in the air as a proportion of the **maximum amount of moisture** that can exist at a **given temperature**.

$$RH = 100 \times \left[\frac{e^{\frac{17.625 \times D_p}{243.04 + D_p}}}{e^{\frac{17.625 \times T}{243.04 + T}}} \right]$$

RH - Relative Humidity, D_p - Dewpoint (in C), T - Air Temperature (in C)

Relative Humidity

Derived from Average Temperature and Dewpoint

Cyclical Transform of Month and Day

Makes more sense for cyclical variable

CodeSum split

Nominal columns for each classification

Clustering

Latlong, resultspd result dir, Tavg Rhumid

Coordinate Tupling

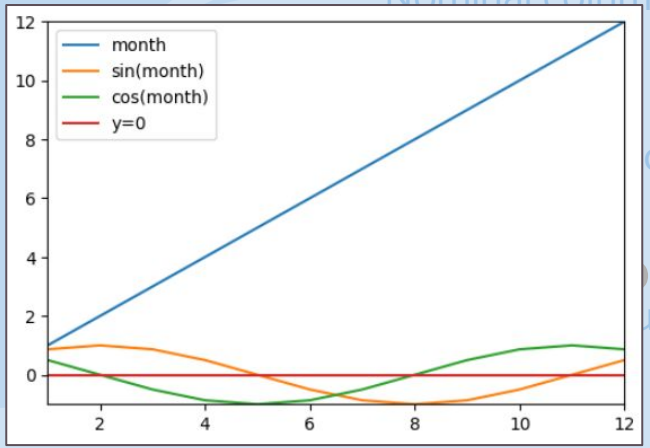
Grouping Latitude/Longitude

Represent **cyclical values** (*in this case, month and day*) as a **function of sin/cos**, to reflect cyclical nature

e.g. sin(December) is close to sin(January)

$$\mathrm{var}_{\sin} = \sin\left(x \times \frac{2\pi}{\max(x)}\right)$$

$$var_{cos} = \cos\left(x \times \frac{2\pi}{\max(x)}\right)$$



Relative Humidity
Derived from Average Temperature and Dewpoint

Cyclical Transform of Month and Day
Makes more sense for cyclical variable

CodeSum split
Nominal columns for each classification

Clustering
d result dir, Tavg Rhumid

ordinate Tupling
uping Latitude/Longitude

‘Dummify’ each **CodeSum** into a separate categorical representer

```
array([' ', 'BR', 'BR HZ', 'HZ', 'RA', 'RA BR', 'TSRA RA BR', 'RA VCTS',
       'TSRA RA', 'RA HZ', 'TSRA RA BR HZ', 'TSRA BR HZ', 'RA BR HZ VCTS',
       'TSRA RA HZ', 'TSRA BR HZ VCTS', 'TSRA', 'TSRA BR HZ FU',
       'TSRA RA HZ FU', 'BR HZ FU', 'TSRA RA VCTS', 'HZ VCTS', 'TSRA HZ',
       'VCTS', 'RA BR VCTS', 'TSRA RA BR VCTS', 'TS TSRA RA BR HZ VCTS',
       'DZ BR', 'TS TSRA RA BR HZ', 'TS TSRA BR HZ', 'RA BR HZ',
       'TSRA RA DZ BR HZ', 'TS TSRA RA BR', 'TS RA BR', 'TS TSRA RA',
       'TS TSRA RA BR VCTS', 'TS TSRA BR', 'TS RA', 'RA BCFG BR',
       'TSRA BR', 'RA DZ FG+ BCFG BR', 'RA FG+ MIFG BR', 'RA DZ',
       'RA DZ BR', 'TS TSRA RA HZ', 'TSRA RA FG+ FG BR',
       'TSRA DZ FG+ FG BR HZ', 'TS BR', 'RA BR SQ', 'TS TSRA',
       'TSRA RA BR HZ VCTS', 'BR VCTS', 'TS', 'FG+ BR HZ', 'RA SN',
       'TSRA RA DZ BR', 'DZ BR HZ', 'RA BR FU', 'TS BR HZ', 'DZ',
       'FG+ BR', 'FG+ FG BR', 'FG+ MIFG BR', 'TSRA RA FG BR',
       'TSRA FG+ BR', 'RA DZ BR HZ', 'RA DZ SN', 'FG+ FG BR HZ',
       'TS TSRA RA FG BR', 'BR HZ VCFG', 'TS RA FG+ FG BR',
       'TSRA RA FG+ BR', 'RA DZ FG+ FG BR', 'TS TSRA RA VCTS', 'FU',
       'TS TSRA VCFG', 'TS TSRA HZ', 'TS TSRA GR RA BR', 'RA FG BR',
       'HZ FU', 'RA BR HZ FU', 'MIFG BCFG BR', 'FG+ BCFG BR',
       'TSRA RA FG+ FG BR HZ', 'FG+', 'TSRA BR SQ', 'TSRA DZ BR HZ',
       'RA BR HZ VCFG', 'RA FG+ BR', 'FG BR HZ', 'TS HZ',
       'TS TSRA RA FG BR HZ', 'RA DZ FG+ BR', 'RA DZ FG+ BR HZ',
       'TSRA FG+ BR HZ', 'RA BR VCFG', 'TS RA BR HZ', 'BCFG BR',
       'RA SN BR'], dtype=object)
```

Wow! So many!

Relative Humidity
Derived from Average Temperature and Dewpoint

Cyclical Transform of Month and Day
Makes more sense for cyclical variable

CodeSum split
Nominal columns for each classification

Clustering
Latlong, resultspd result dir, Tavg Rhumid

Coordinate Tupling
Grouping Latitude/Longitude

Tupling coordinate values ***Latitude*** and ***Longitude*** into single, categorical column **'Coordinates'**

Tuples are then **dummified** as specific location vectors

Relative Humidity
Derived from Average Temperature and Dewpoint

Cyclical Transform of Month and Day
Makes more sense for cyclical variable

CodeSum split
Nominal columns for each classification

Clustering
Latlong, resultspd result dir, Tavg Rhumid

Coordinate Tupling
Grouping Latitude/Longitude

Next Section!

Relative Humidity
Derived from Average Temperature and Dewpoint

Cyclical Transform of Month and Day
Makes more sense for cyclical variable

CodeSum split
Nominal columns for each classification

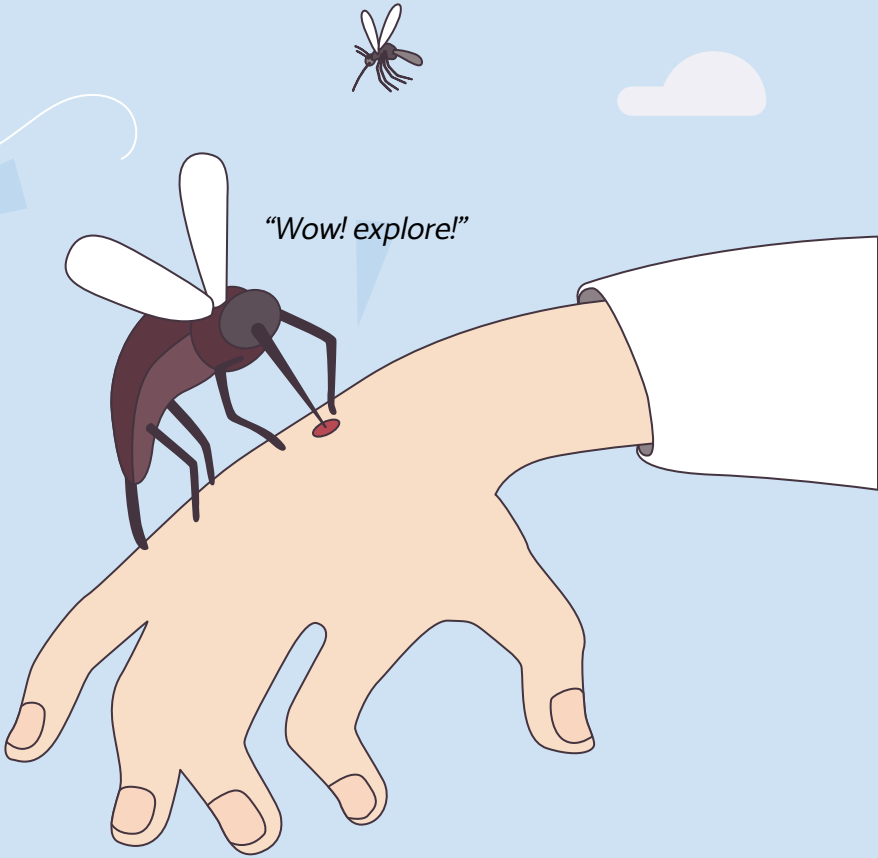
Clustering
Latlong, resultspd result dir, Tavg Rhumid

Coordinate Tupling
Grouping Latitude/Longitude

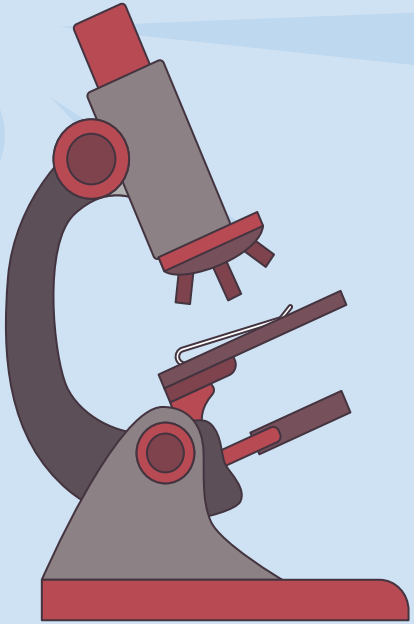
03

EDA & Modeling

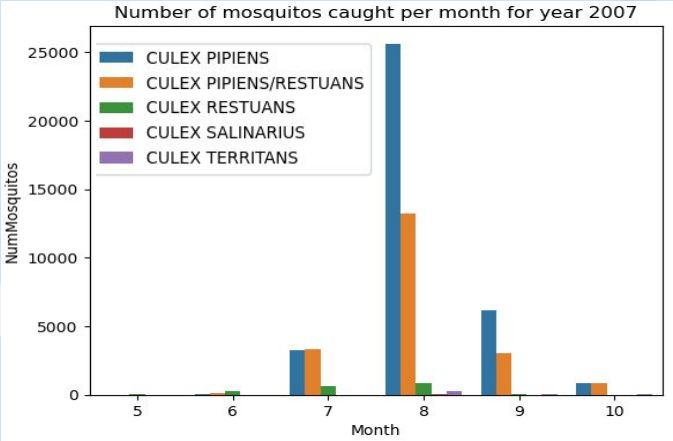
Data insights
Modeling Flow

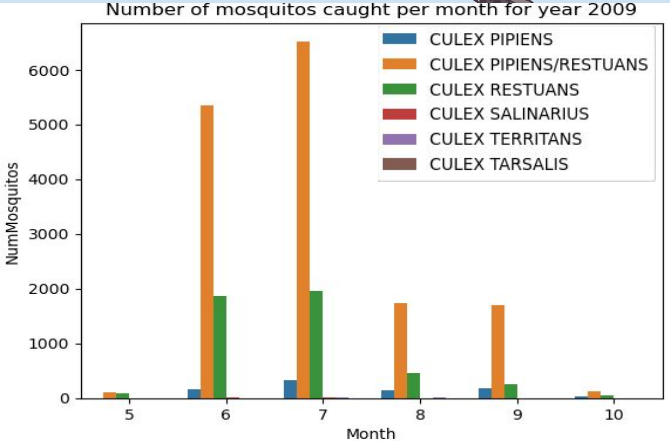


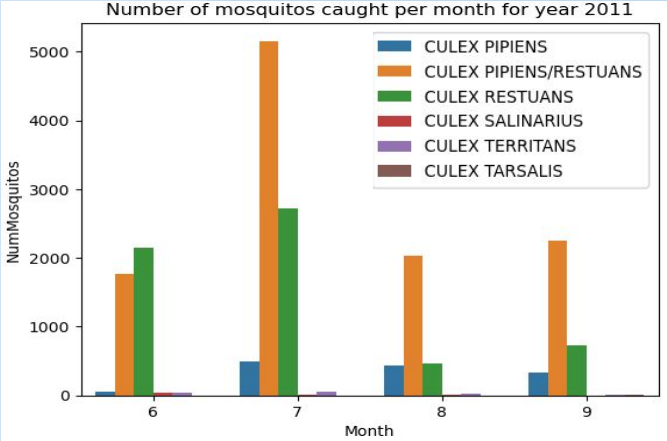
Exploratory Data Analysis

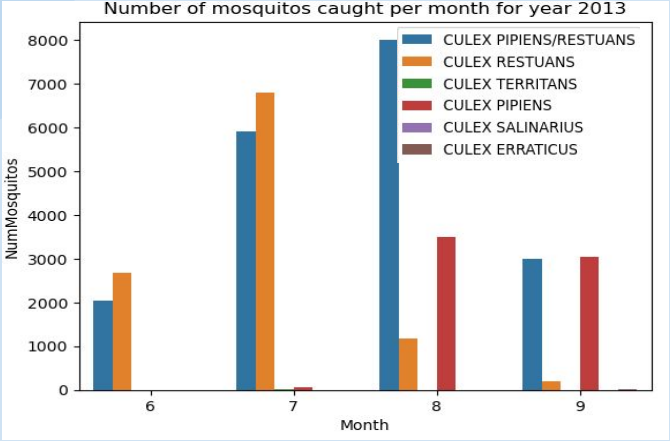


EDA • Number of Mosquito Caught per year

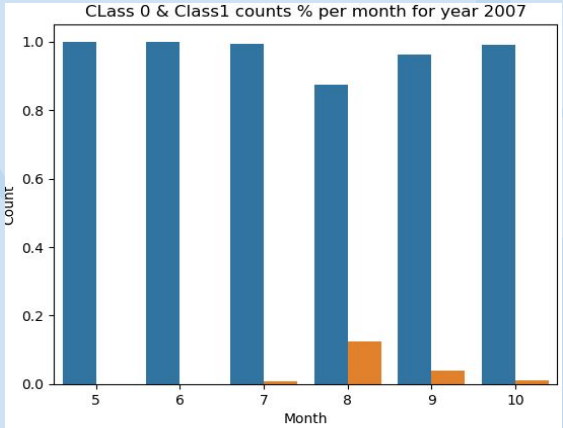


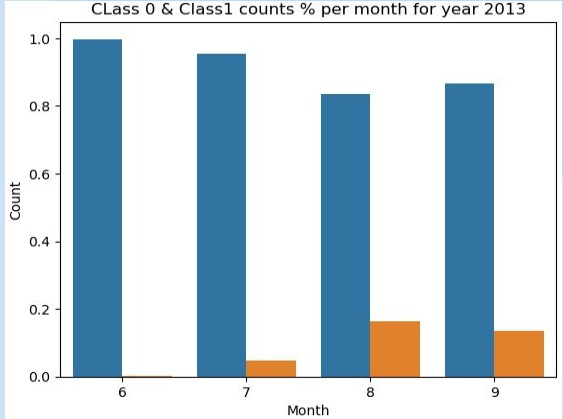


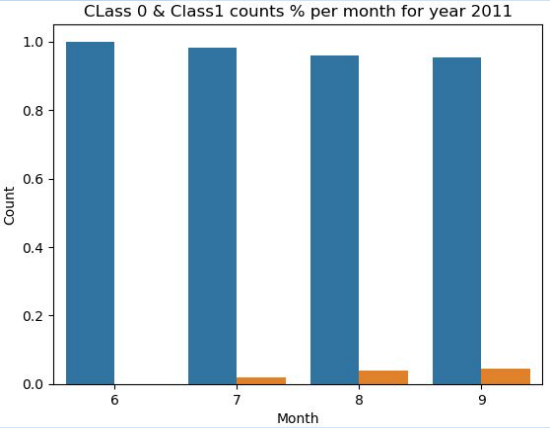


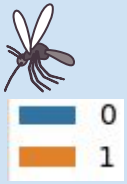


EDA • WNV presences by Months per Year

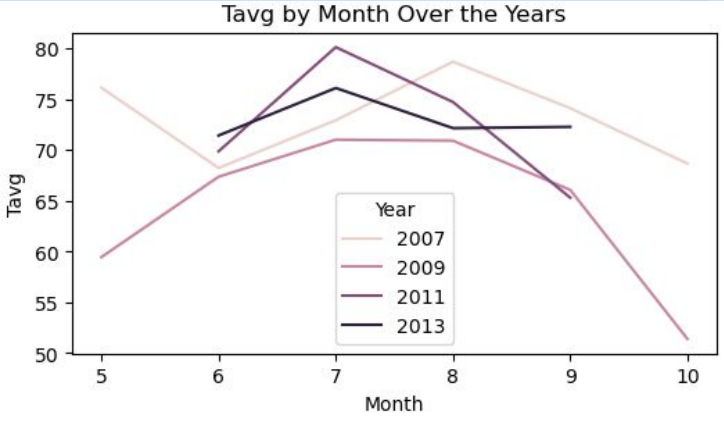


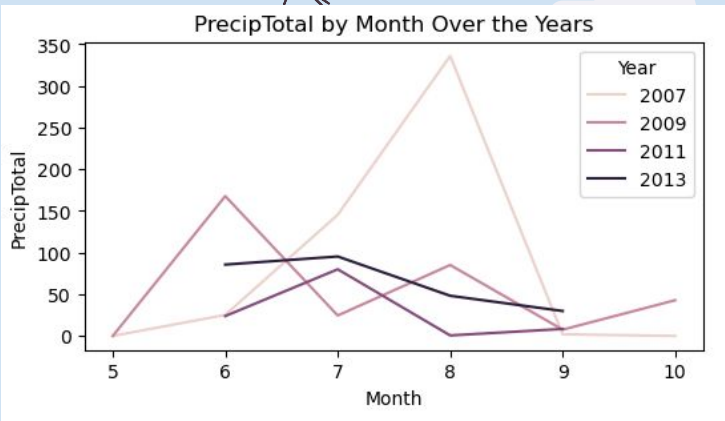


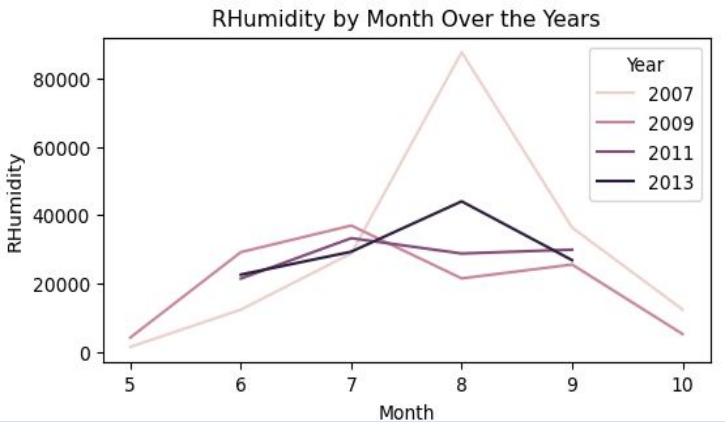




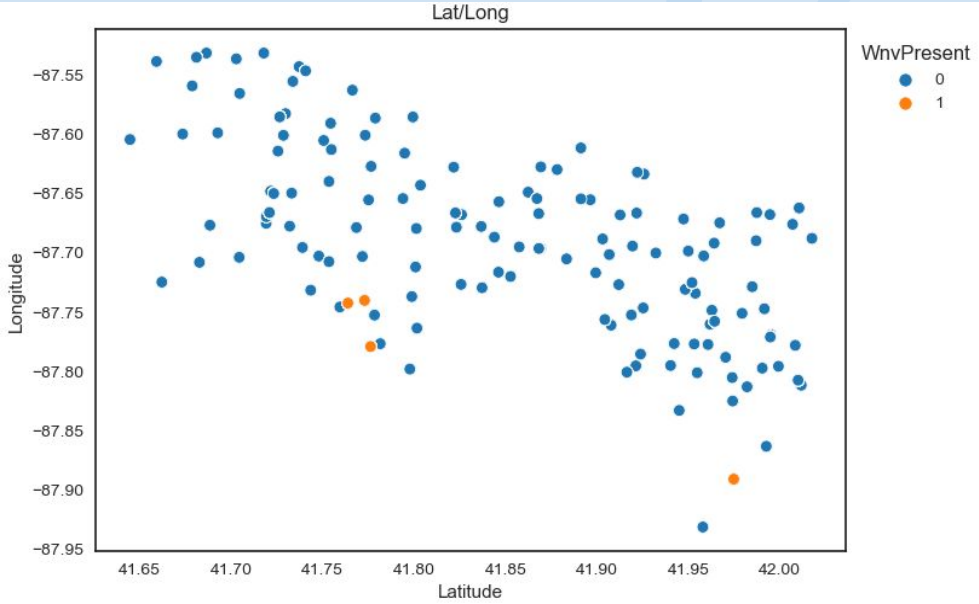
EDA • Weather condition by Years over span of month

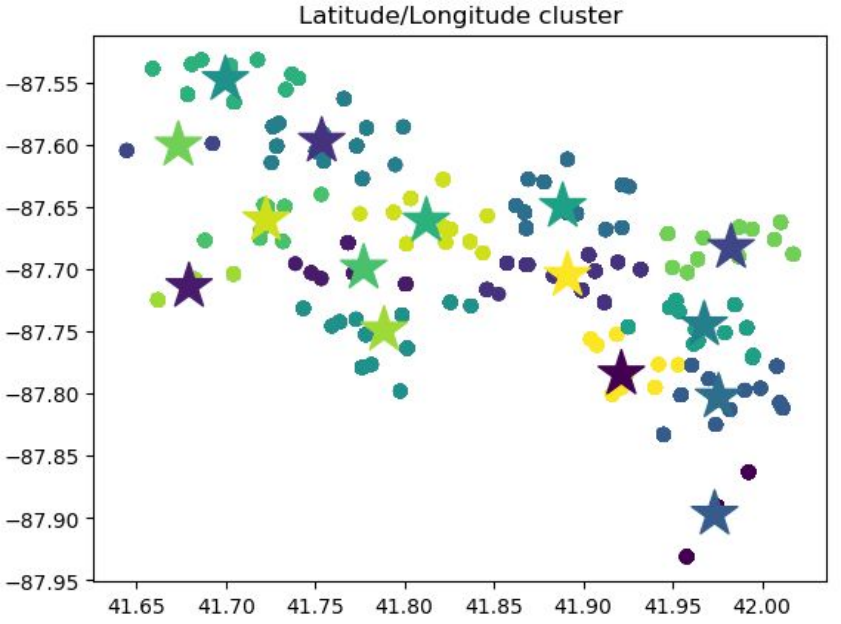




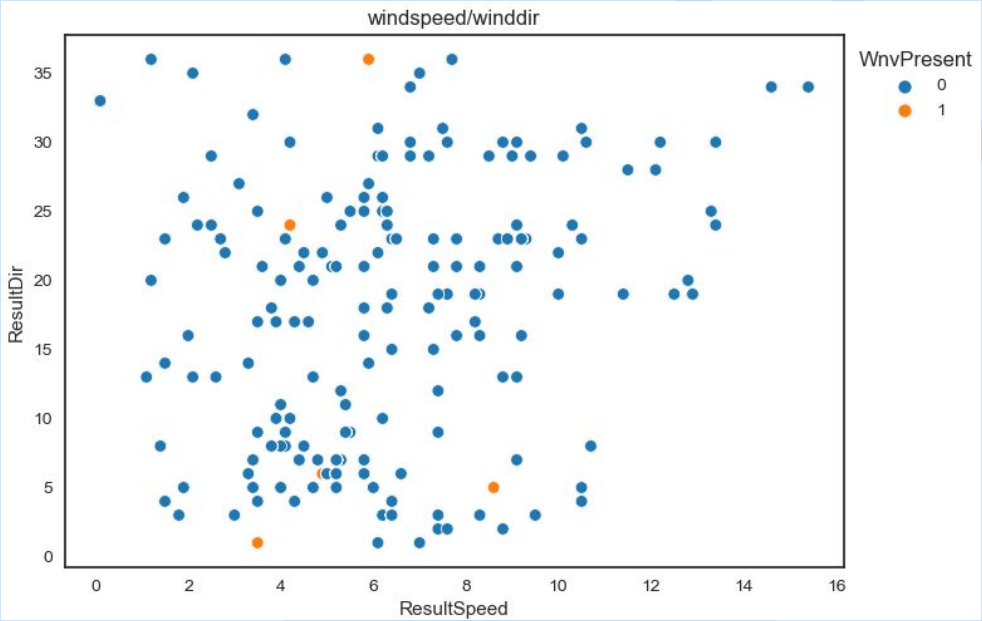


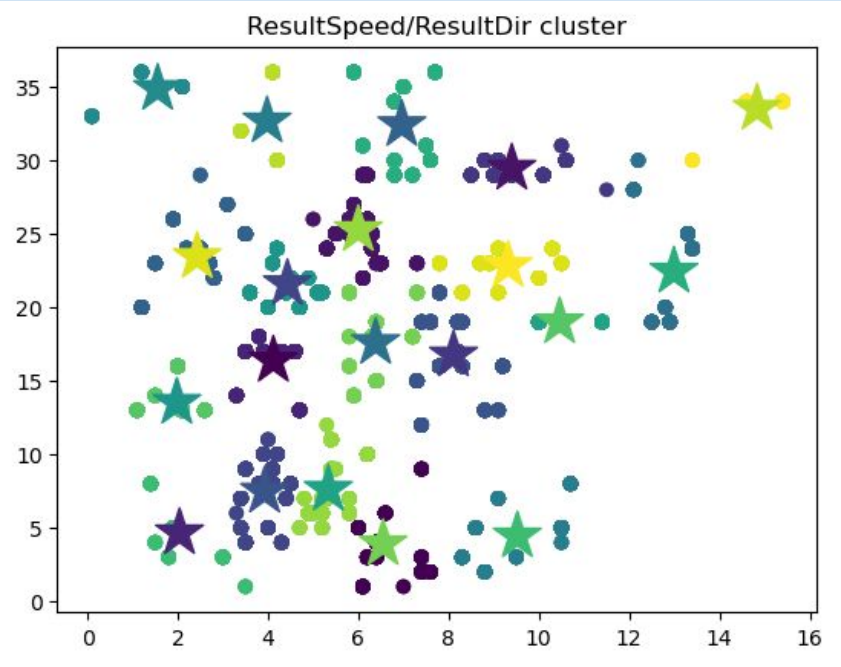
EDA • Clustering for Wnv Presence (Lat/Long)



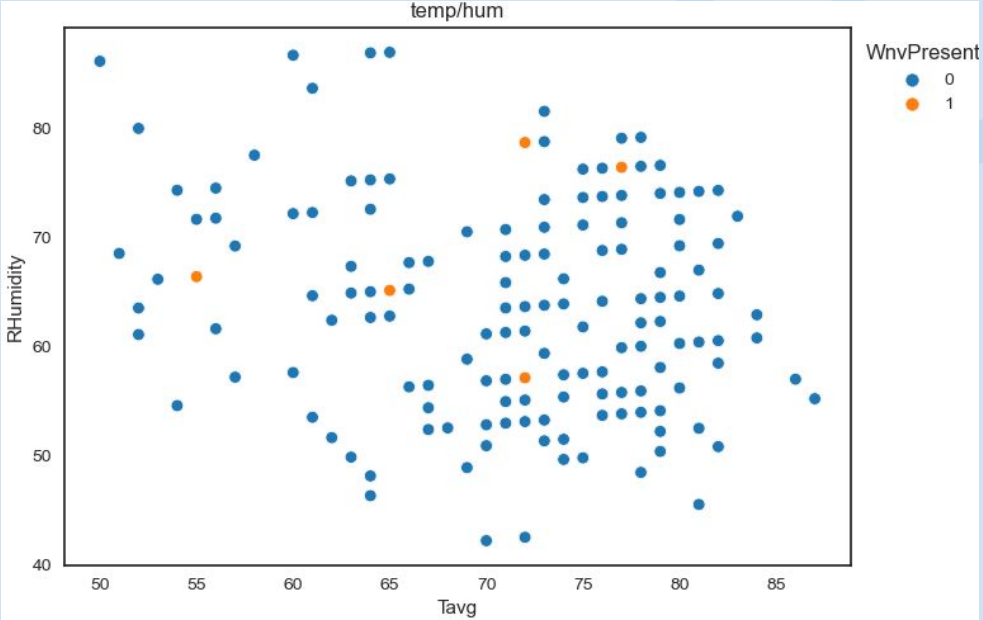


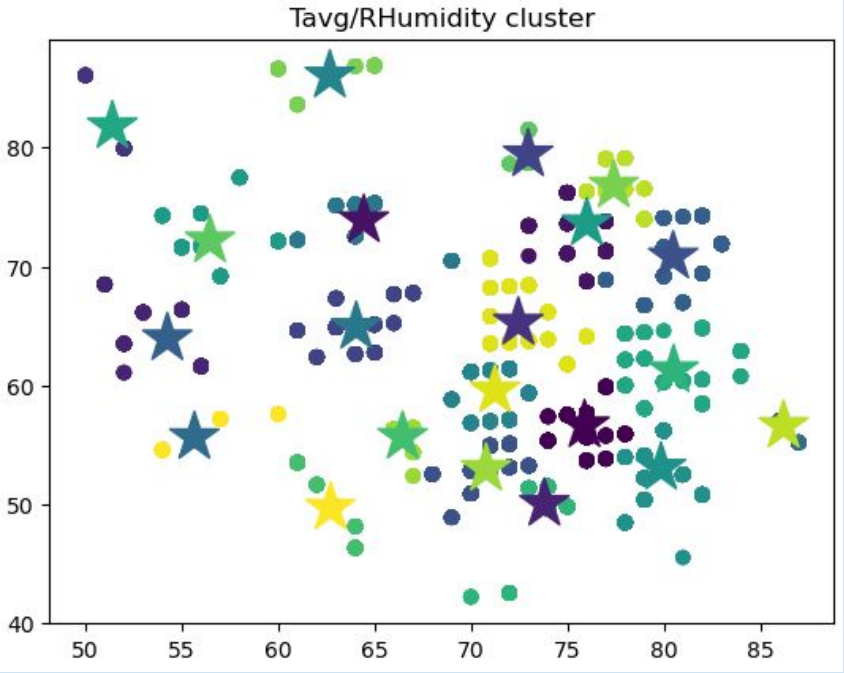
EDA • Clustering for Wnv Presence (Wind)



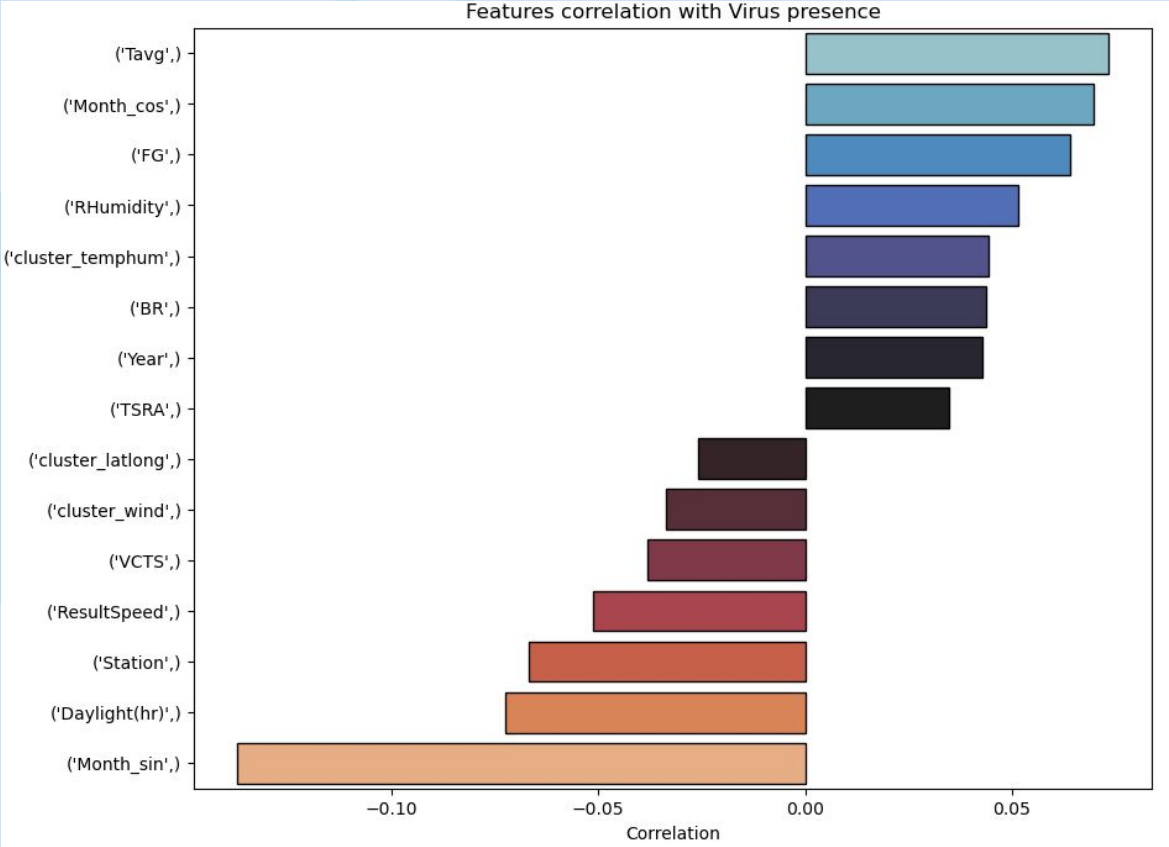


EDA • Clustering for Wnv Presence (Temp/Humidity)

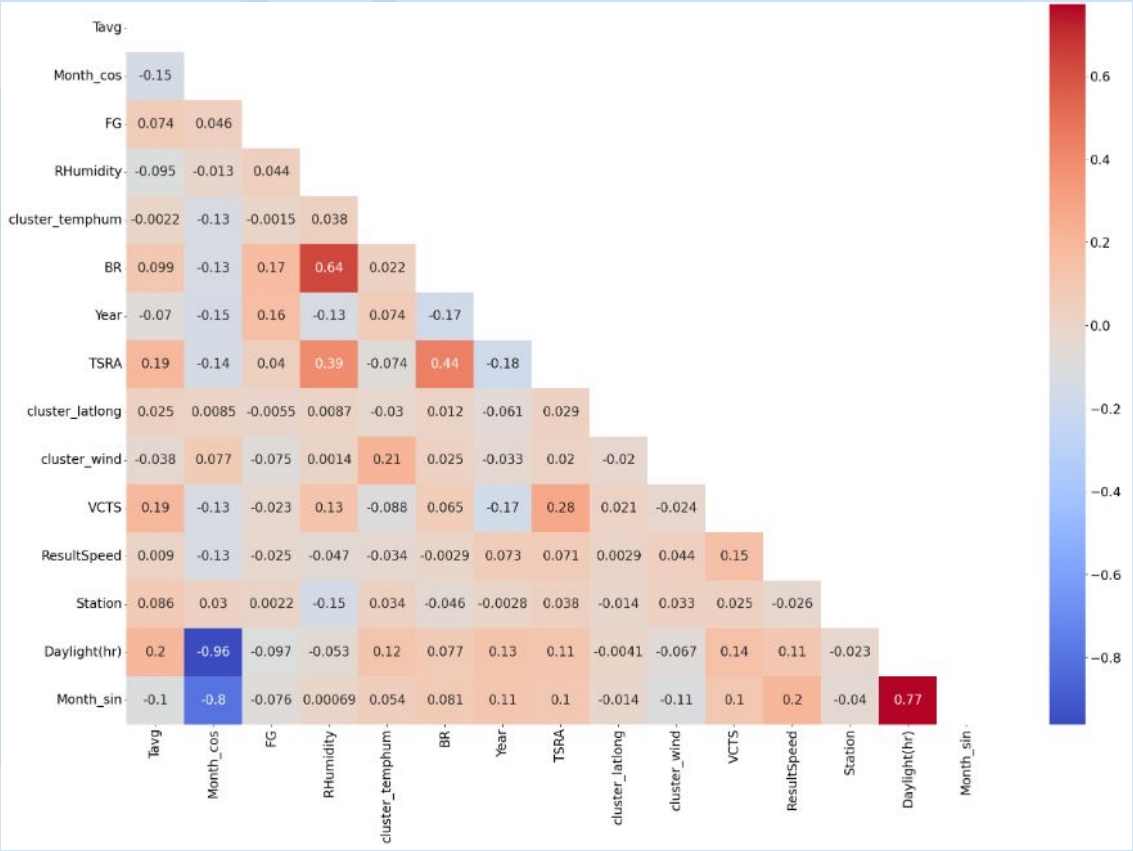


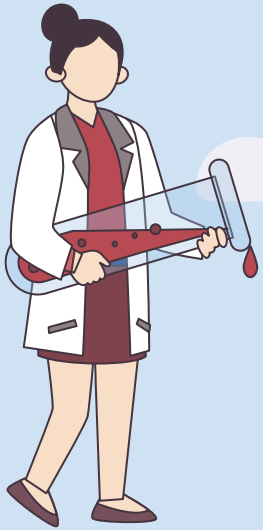


EDA • Feature correlation with Virus presence (>= 0.025 & <= -0.025)



EDA • Feature correlation with each other





MODEL

Baseline Model

- **Oversampling and PCA comparison**
  - Log Regression baseline with no resampling
  - SMOTENC was chosen for oversampling technique to help with our imbalance class
  - PCA

Score	LR no resampling	LR SMOTE resampling	LR ADASYN resampling	LR SVMSMOTE resampling	LR SMOTENC resampling	LR SMOTENC PCA
Acc (train)	0.87	0.99	0.99	0.99	0.99	0.75
Acc (test)	0.75	0.73	0.73	0.73	0.74	0.73
MisclassRate(test)	0.25	0.27	0.27	0.27	0.26	0.27
Recall (test)	0.69	0.68	0.68	0.69	0.68	0.68
Spec (test)	0.75	0.73	0.73	0.74	0.74	0.74
Precision (test)	0.14	0.13	0.13	0.13	0.13	0.13
F1 (test)	0.23	0.22	0.22	0.22	0.22	0.22
ROC_AUC (test)	0.72	0.71	0.71	0.71	0.71	0.71

Pycaret Modeling

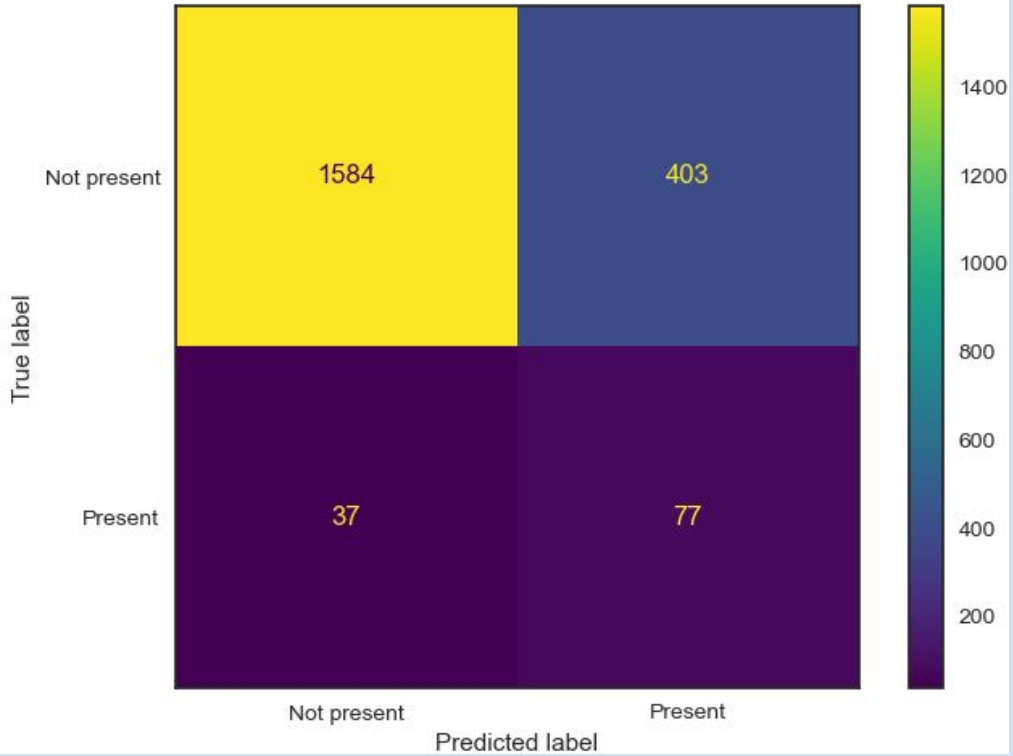
- **Use Pycaret Best model function (sort by AUC)**
  - Random Forest Classifier
  - Extra Trees Classifier
  - Logistic Regression
  - Extreme Gradient Boosting (Self add in)

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9391	0.9908	0.9697	0.9136	0.9408	0.8782	0.8799	0.9460
et	Extra Trees Classifier	0.9429	0.9907	0.9632	0.9255	0.9440	0.8859	0.8867	0.9000
lr	Logistic Regression	0.9610	0.9891	0.9503	0.9710	0.9605	0.9221	0.9223	0.6220

Pycaret Modeling

- **Random Forest**
  - Able to predict 68% of class 1 and 80% of class 0
  - AUC score 0.74

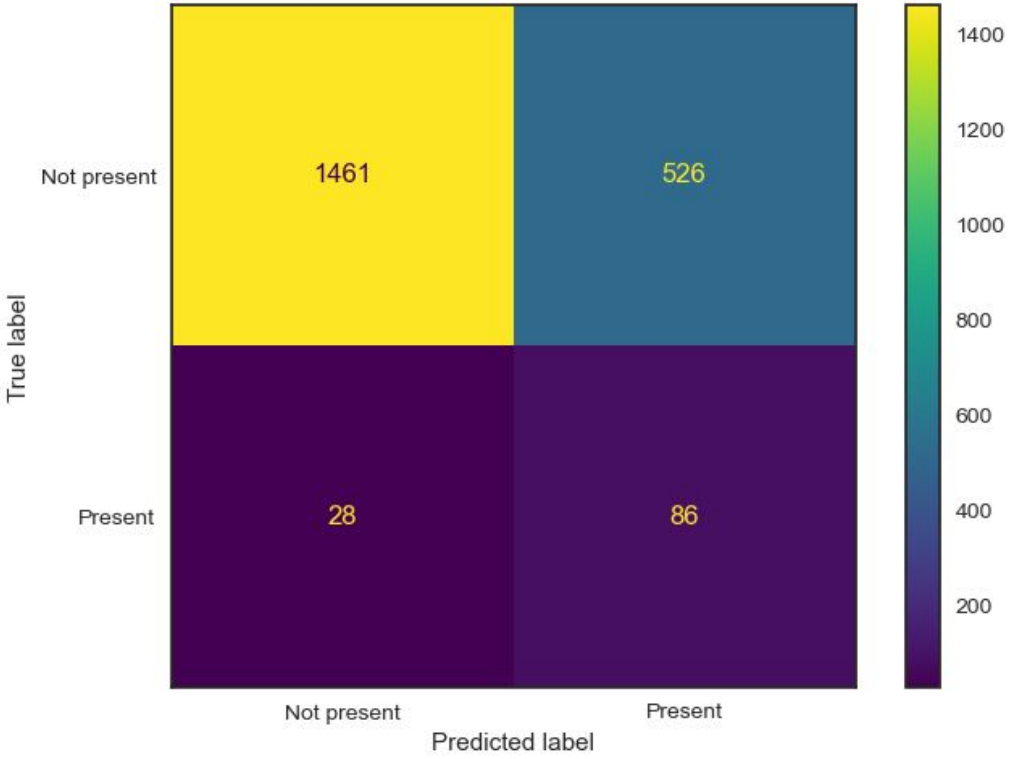
```
Accuracy_score(test):0.79
MisclassificationRate_score(test):0.21
Recall_score(test):0.68
Specificity_score(test):0.80
Precision_score(test):0.16
F1_score(test):0.26
ROC_AUC_score(test):0.74
```



Pycaret Modeling

- **XGBoost**
  - Able to predict 75% of class 1 and 74% of class 0
  - AUC score 0.74

```
Accuracy_score(test):0.74
MisclassificationRate_score(test):0.26
Recall_score(test):0.75
Specificity_score(test):0.74
Precision_score(test):0.14
F1_score(test):0.24
ROC_AUC_score(test):0.74
```

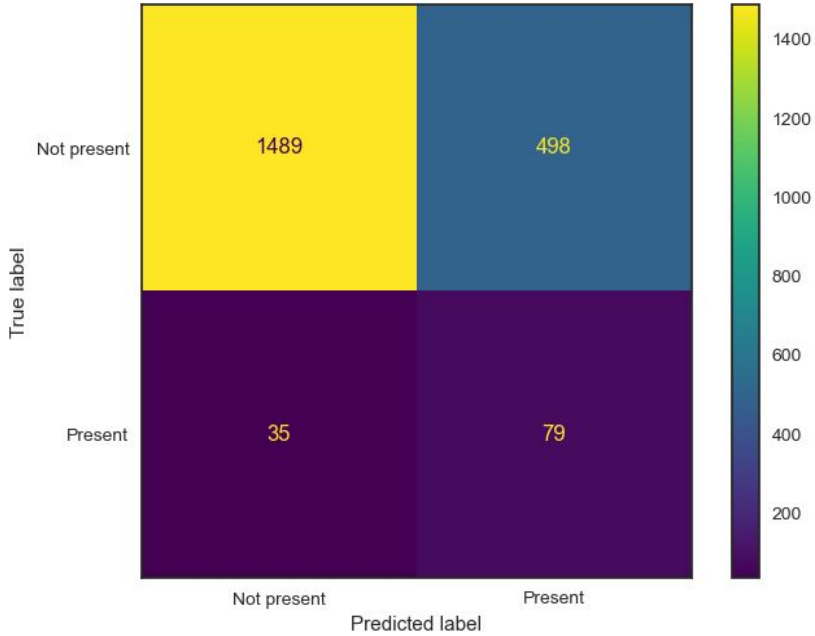


Pycaret Modeling

- **Baseline vs Best Model**

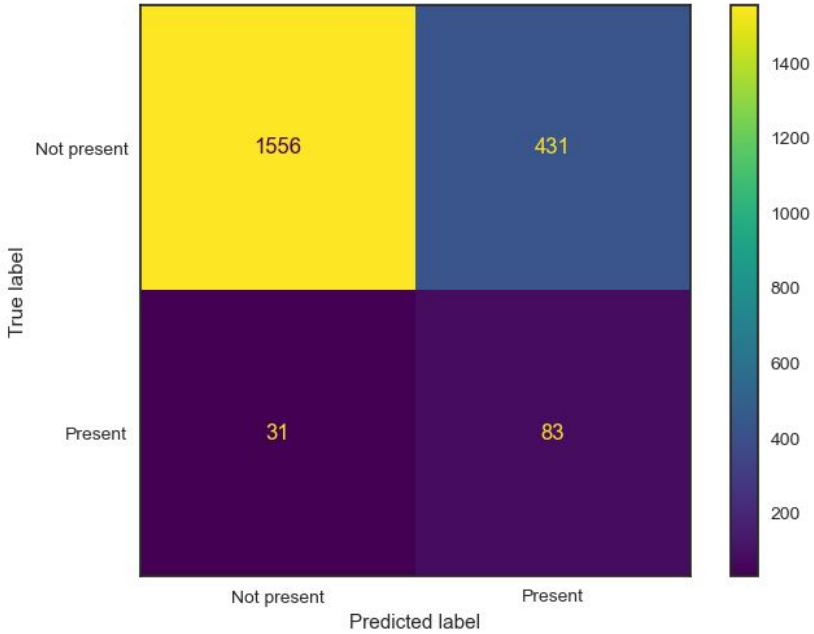
```
Accuracy_score(test):0.75
MisclassificationRate_score(test):0.25
Recall_score(test):0.69
Specificity_score(test):0.75
Precision_score(test):0.14
F1_score(test):0.23
ROC_AUC_score(test):0.72
```

LogReg
No resampling



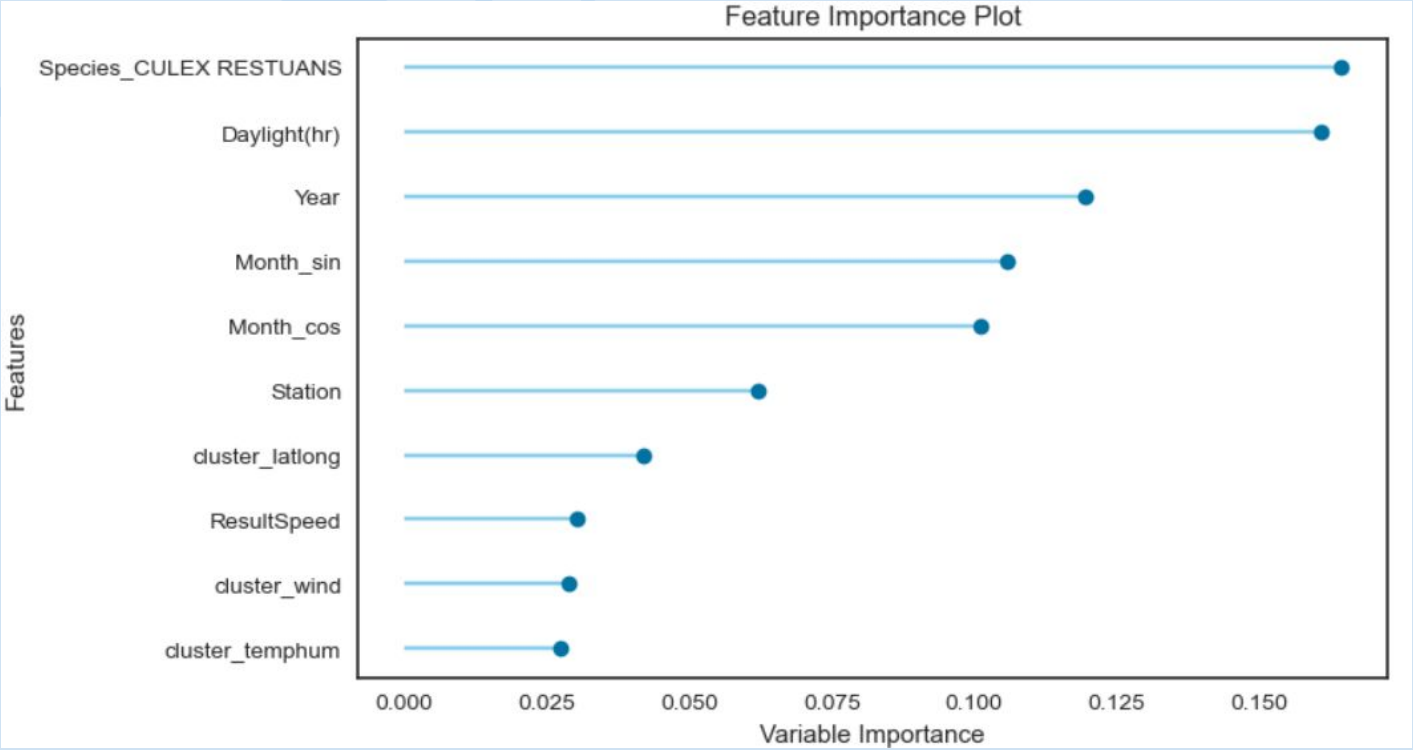

```
Accuracy_score(test):0.78 (+3%)
MisclassificationRate_score(test):0.22
Recall_score(test):0.73 (+4%)
Specificity_score(test):0.78 (+3%)
Precision_score(test):0.16 (+2%)
F1_score(test):0.26 (+3%)
ROC_AUC_score(test):0.76 (+4%)
```

ET
SMOTENC



Extra Trees Model interpretation

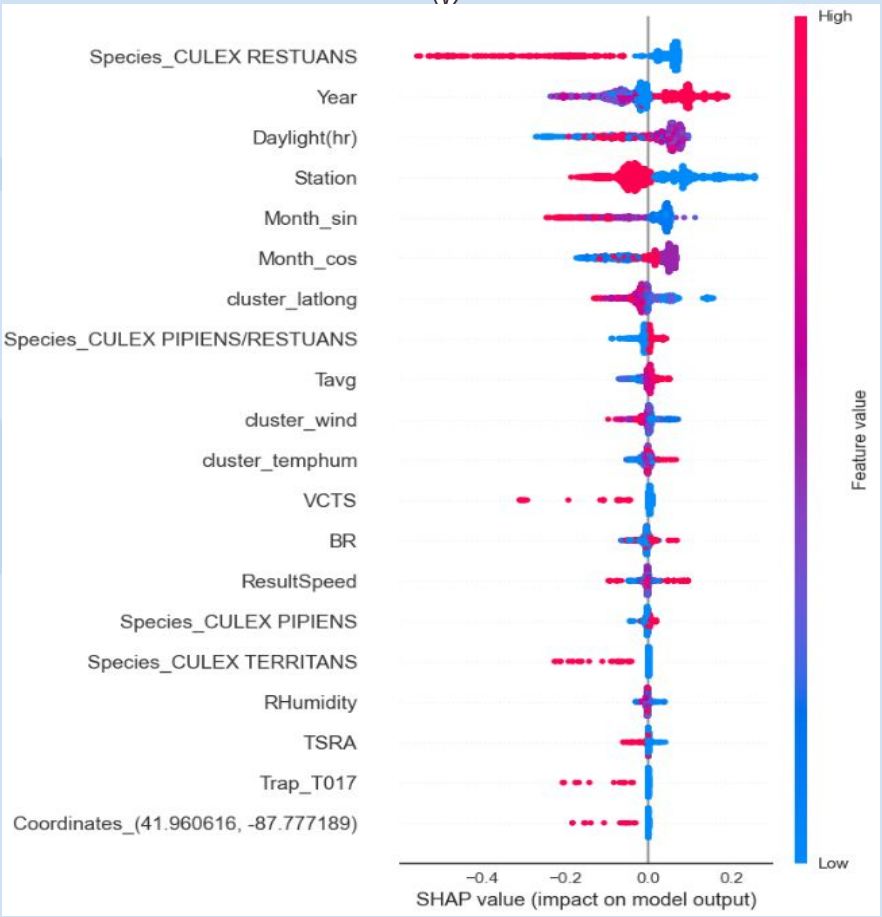
- **Feature Importance Plot**



Extra Trees SHAP Values

- High Feature values with Positive impact:
  - Year
  - CULEX PIPIENS/RESTUANS mix
  - TEMP Average

- Low Feature values with Positive impact:
  - CULEX RESTUANS only
  - STATION
  - Month_Cyclical
  - Cluster_latlong

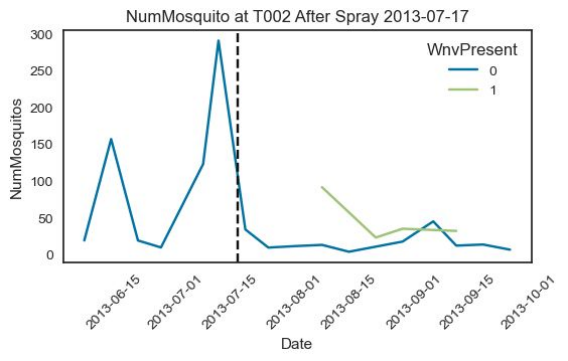


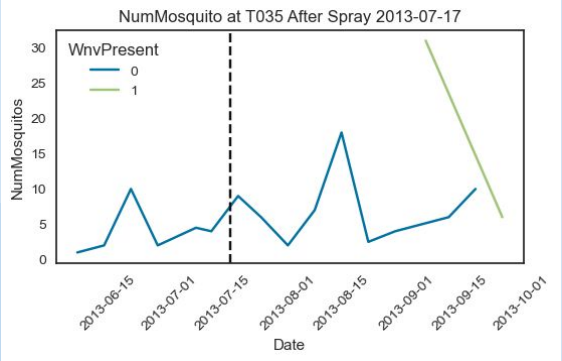
04

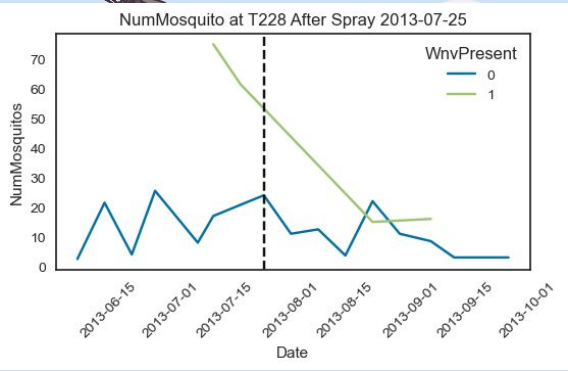
Cost Benefit Analysis

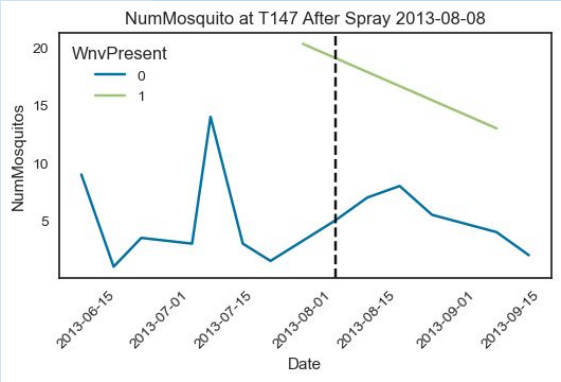

Is spraying pesticide an effective countermeasure?
Is the spray regime cost-effective?

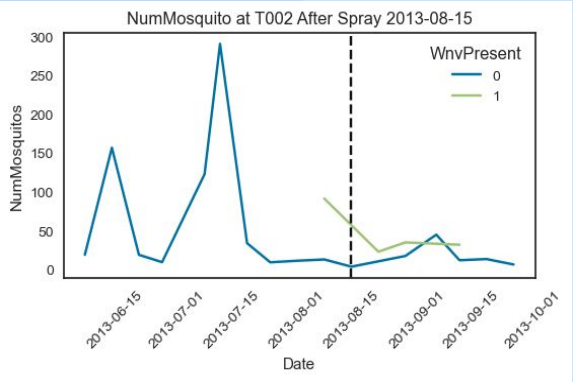
Cost Benefit Analysis

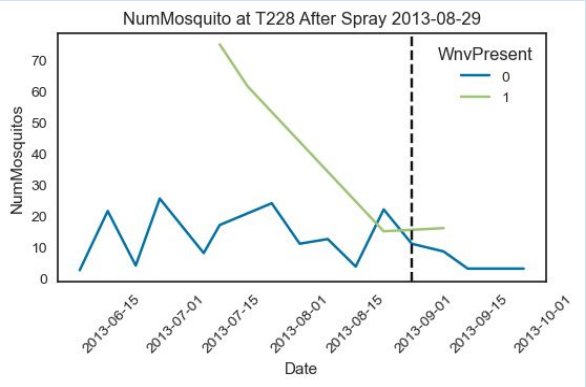


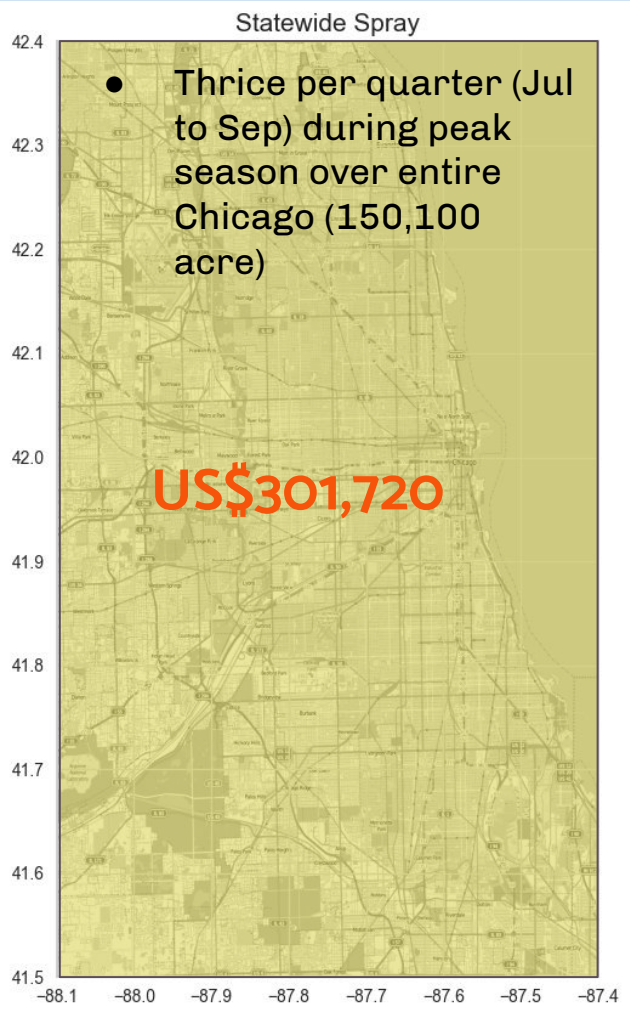




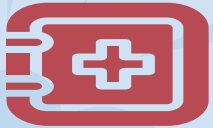






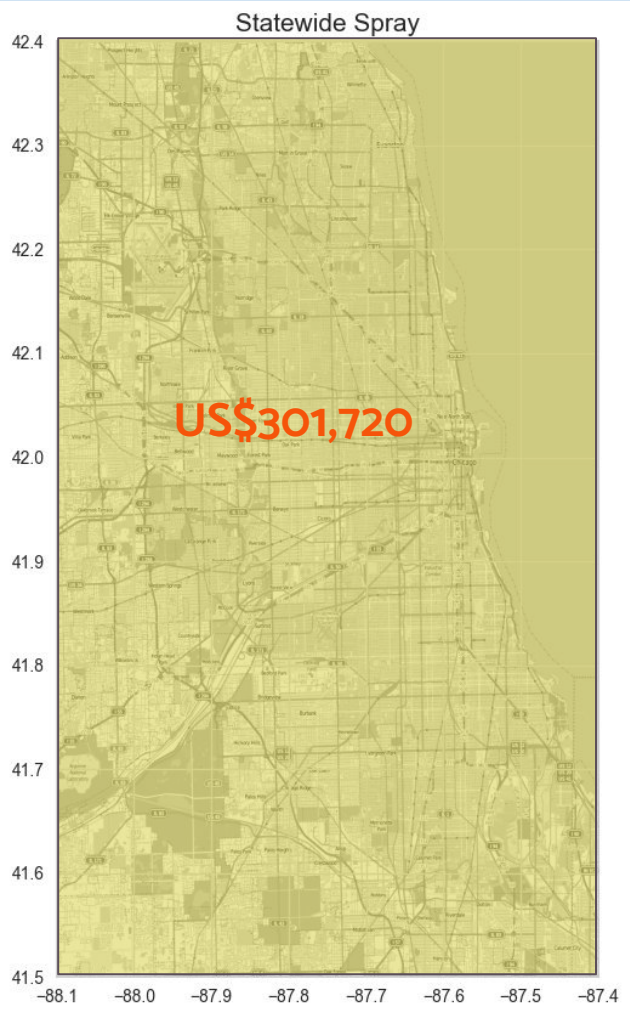


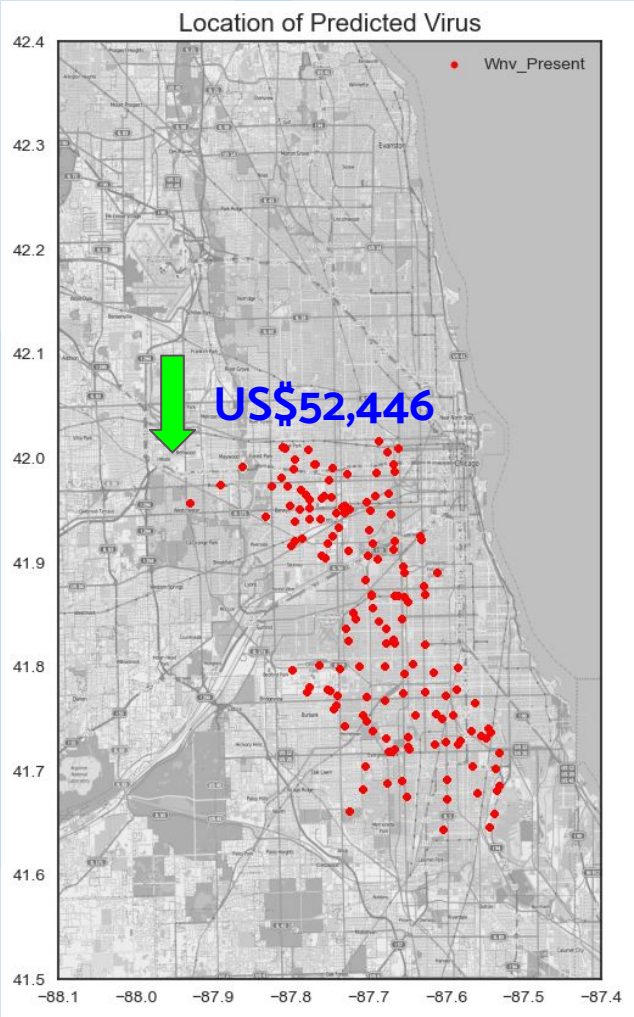
Cost Benefit Analysis



- **Medical Cost**
  - Personnel who get serious illness may need to be hospitalised
  - Substantial cost incurred for treatment of such patient (estimated US$21,000 per patient).
  - 15 cases need to be prevented to cover the cost of spray programme.

- **Impact to Workforce / Productivity**
  - Personnel may be absent from work affecting Chicago's workforce productivity.
  - Significant impact to businesses if West Nile Virus is not under control
  - Estimated loss of US$281 for each man-day loss.
  - 358 cases need to be prevented to cover the cost of spray programme





- Based on our model, 12,058 positive observations of presence of Wnv in Year 2012 (highest in test data across 2008, 2010, 2012 and 2014).

- Selective spray of pesticide using prediction from model..

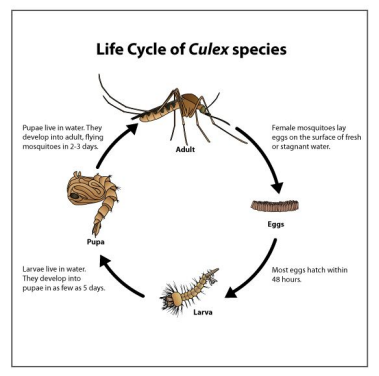
05

Conclusion and Recommendation

Conclusion and Limitations

- Pesticide spraying is an effective means for prevention of West Nile Virus.
- Recommend CDPH to adopt Extra Trees model (our best model) to predict presence of Wnv carrying mosquito to derive spray regime.
- Model and prediction is limited to:
  - Chicago only (or locations with similar weather conditions)
  - 6 known mosquito types

Recommendations



Life Cycle

Weather

Larvicide

THANK YOU

Together we can achieve a West Nile Virus Free-Day...