Project 4
West Nile Virus Prediction

Vincent Chua
May Gee
Tan Cheng Yeow
Tan Cheng Kiat
Low Zhe Wei

Table Of Contents

Introduction

1. First case in USA in 1999 in New York

2. First Illinois case on September 2001 in Chicago

3. In 2002, Illinois accounted for more human cases (884) and deaths (64) than any other state in USA.

4. City of Chicago and the Chicago Department of Public Health (CDPH) established surveillance and control program that is still in effect today

5. Virus is now an endemic in Chicago

West Nile Virus

Problem Statement

Due to the endemic of West Nile Virus in Chicago, the Department of Public Health has set up a surveillance and control system through which weather, location, testing, and spraying data was collected. CDPH has contacted our team to develop a model to predict the locations where there would be West Nile virus outbreaks.

Using these available datasets, the model will help the City of Chicago and CPHD more efficiently and effectively target spraying of specific neighbourhoods with higher risk of West Nile Virus. This can help the City of Chicago save costs while still keeping the virus at bay. Our model efficacy will be assessed by the Kaggle submission.

Datasets - Kaggle

Spray

- Total of 14835 observations from 4 features
- Date, time and location of spray the pesticides

Train

- Total of 10506 observations from 12 features.
- Additional **NumMosquitos** and **WnvPresent** features
- Date, time, location of trap and mosquitos species and numbers

Weather

- Total of 2944 observations from 22 features
- Date, temperature, windspeed, station pressure etc.

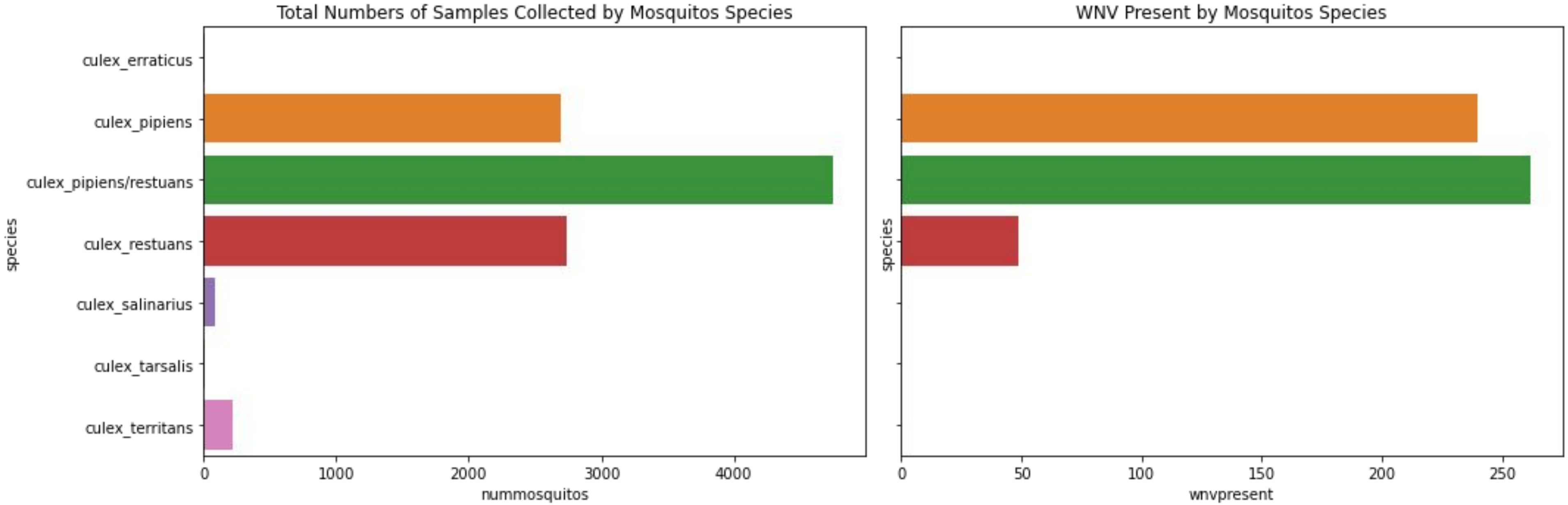
Test

- Total of 11,6293 from 11 features.
- Date, time, location of trap and mosquitos species

Source: https://www.kaggle.com/c/predict-west-nile-virus/

Exploratory Data Analysis (EDA) - Species (Train)

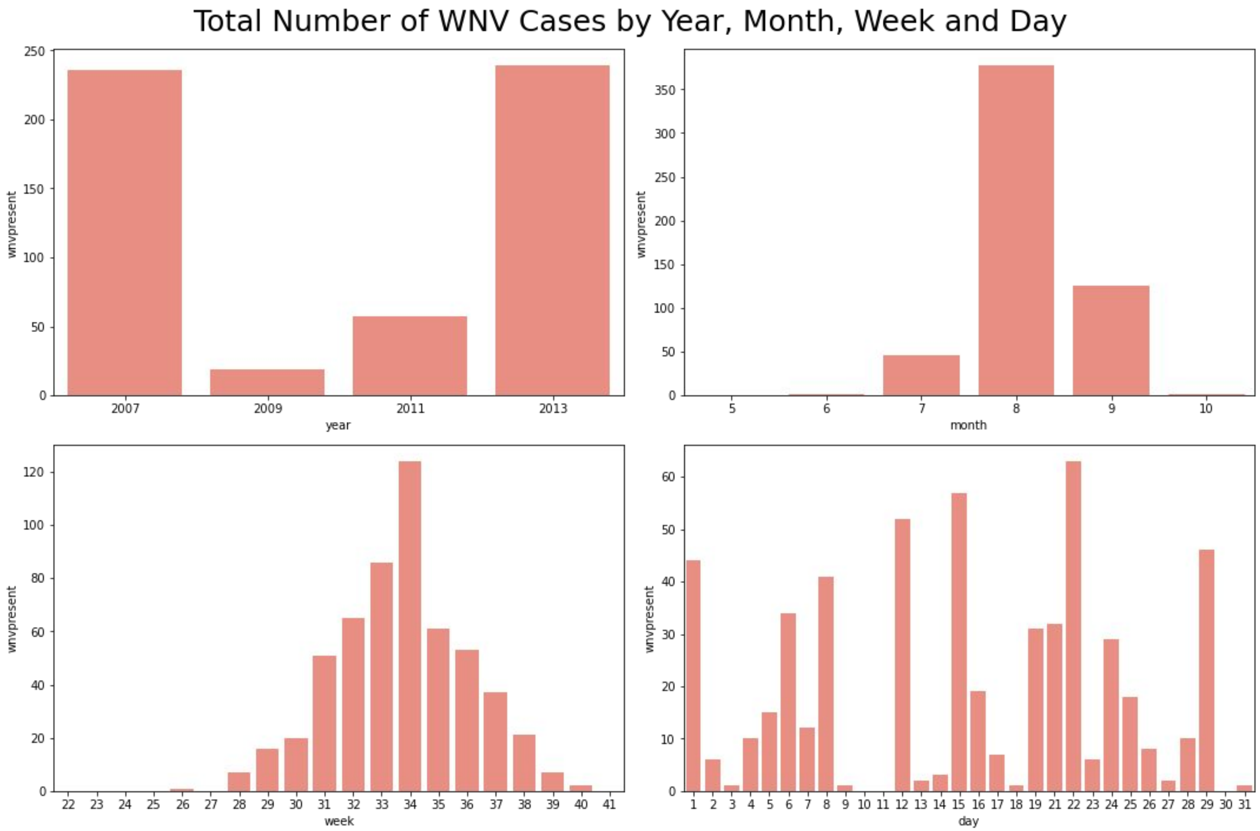
- Imbalanced Class of Datasets with only 5.2% WNV Present cases.
- Majority of mosquitos species are **pipiens & restuans**



EDA - WNV Present

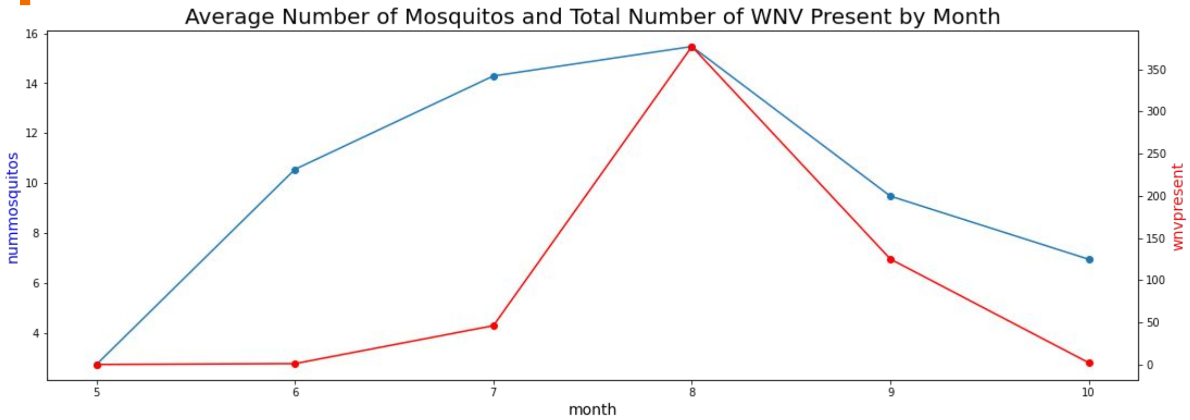
- The months of WNV present cases was observed within july to september and peak in August

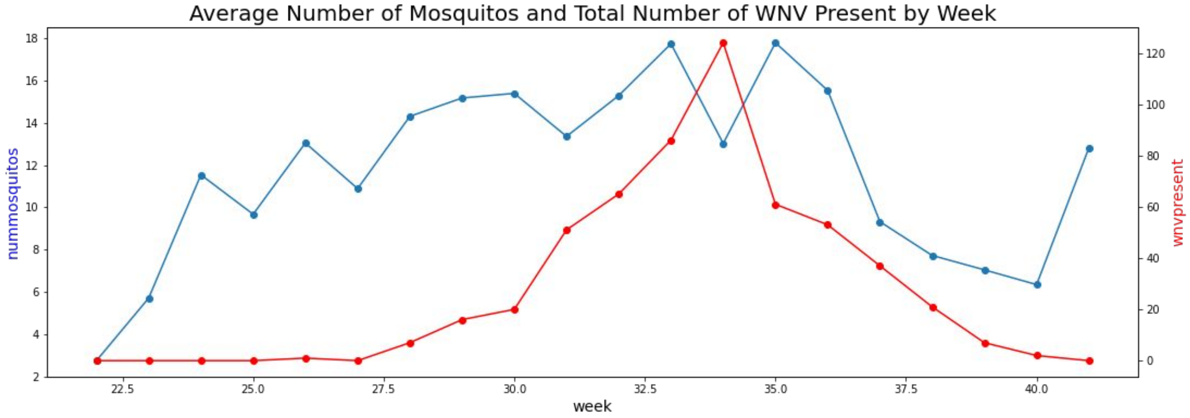
- Years and Day does not show any consistency in case present



EDA - Number of Mosquitos

- For both **month** and **week**, when the numbers of mosquitos increasing, the WNV cases tend to increase

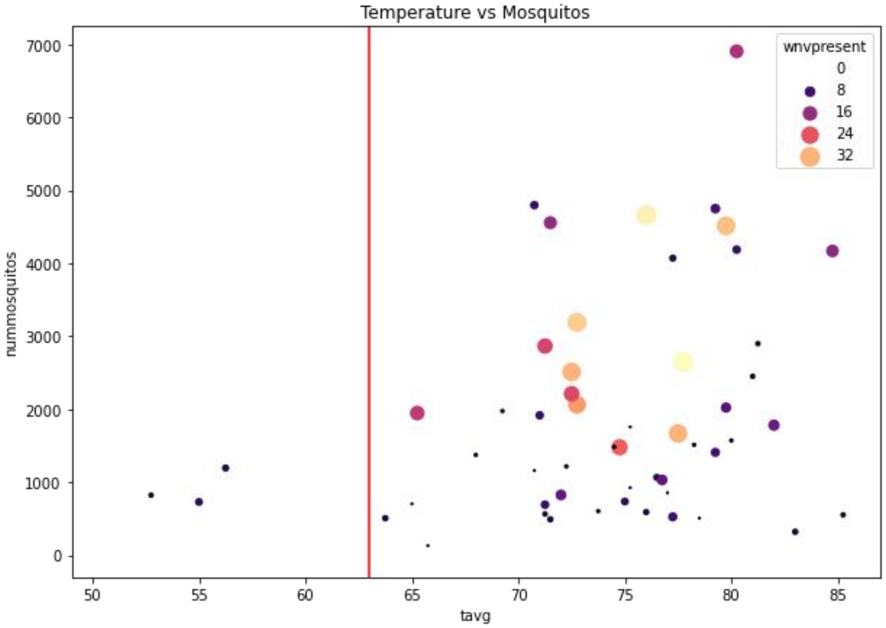


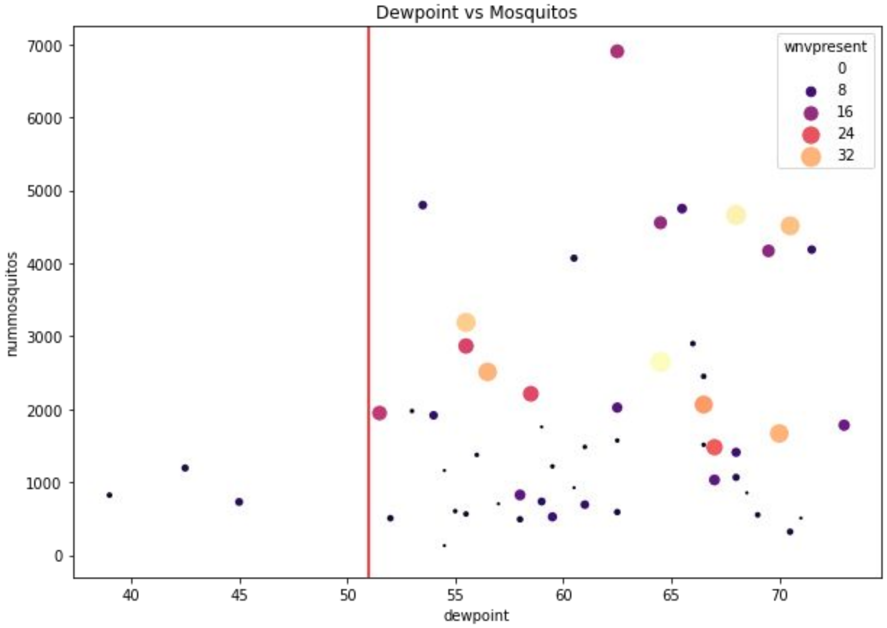


EDA - Temperature & Dewpoint (Weather)

For **average temperatures** above 63°F, we can see that number of mosquitoes and west nile clusters are more prevalent.

For **DewPoint** above 51°F, we can see that number of mosquitoes and west nile clusters are more prevalent.

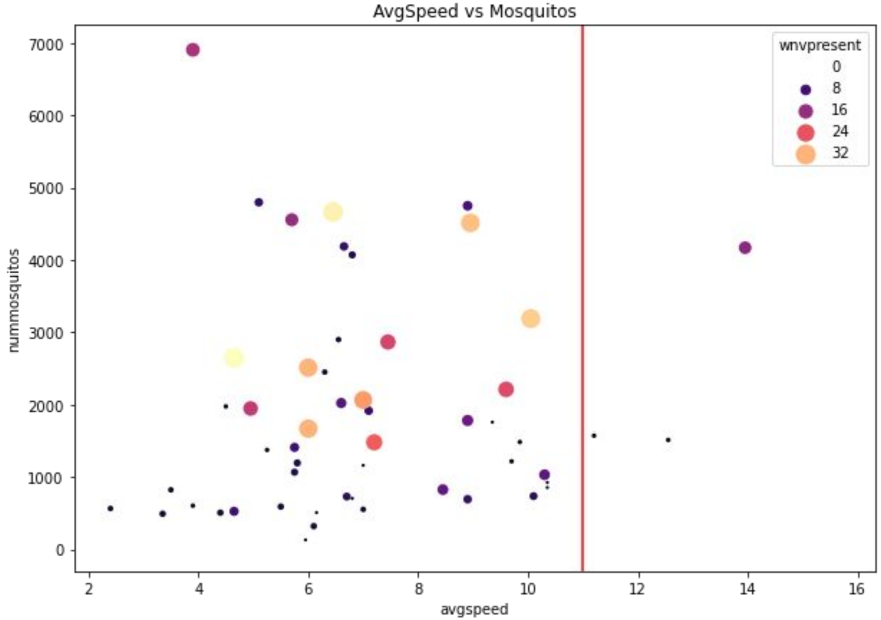


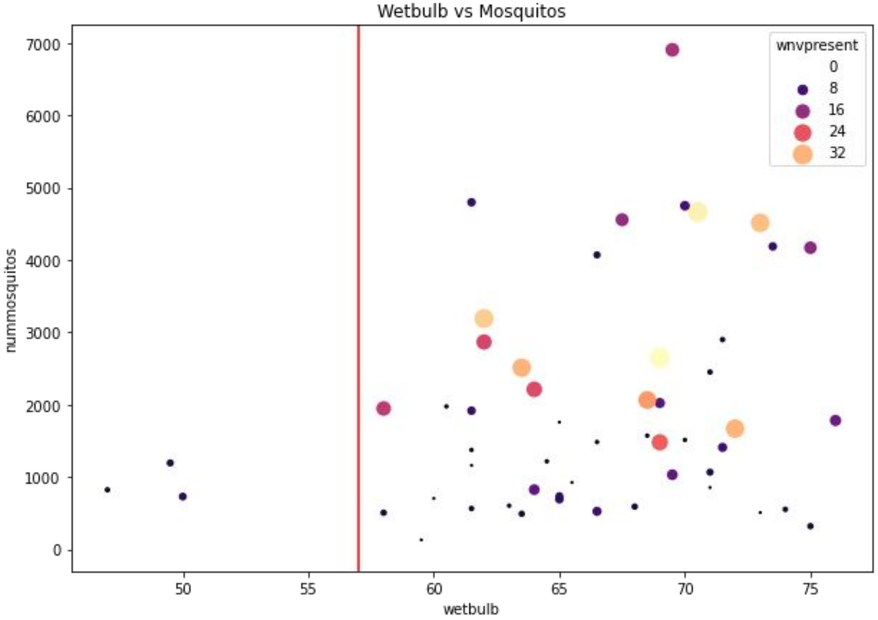


EDA - Wind Speed & Wet bulb (Weather)

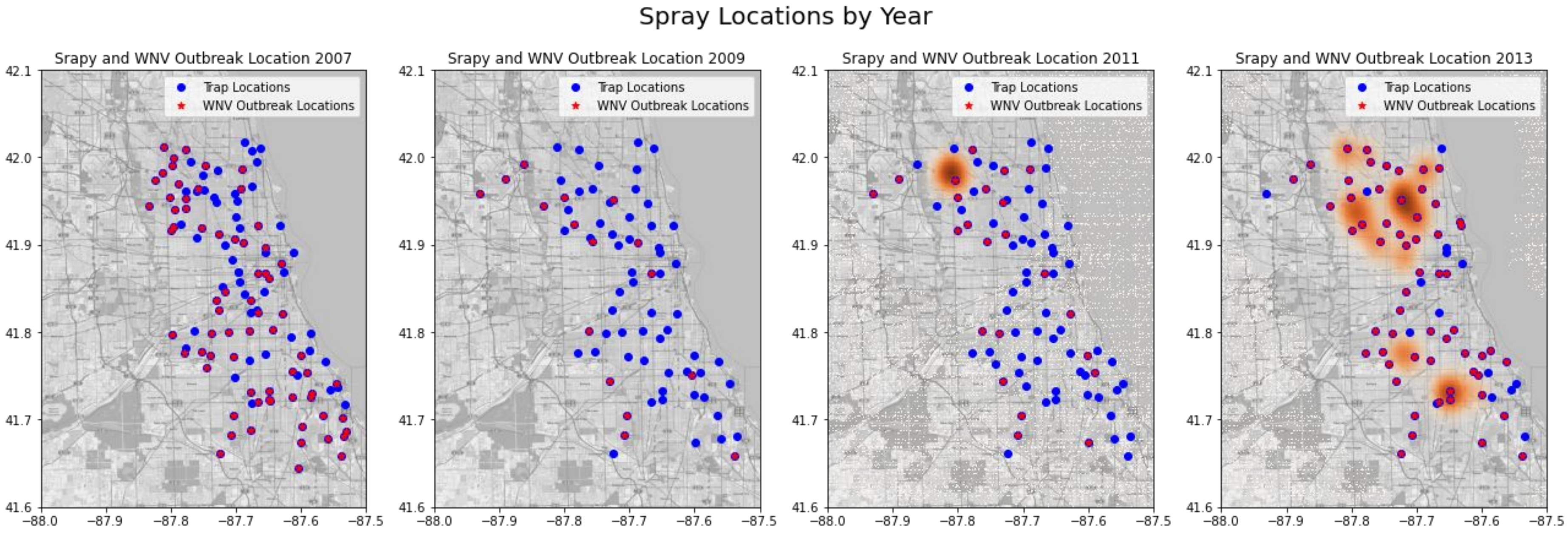
From the graph, we can see that number of mosquitos and wnv clusters are more prevalent at **avgspeed** below 11 miles/hour.

For **WetBulb** above 57°F, we can see that number of mosquitos and west nile clusters are more prevalent.





EDA - Trap / WNV / Spray Location

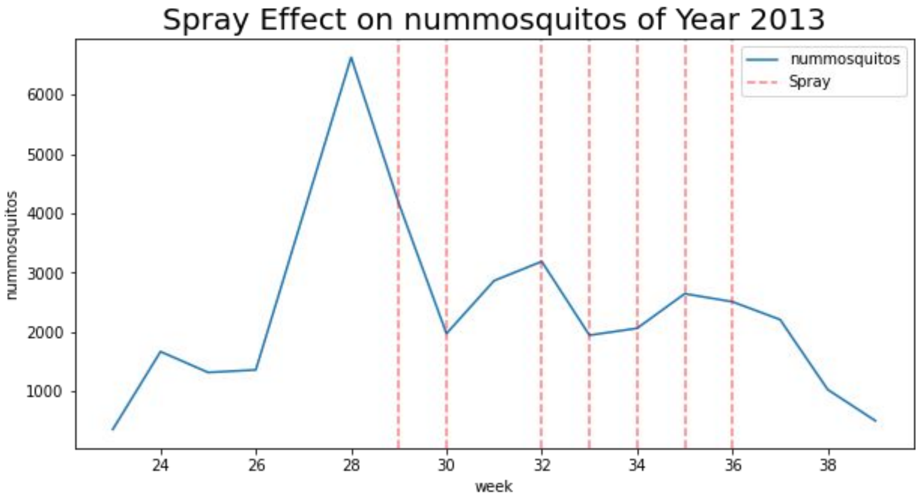


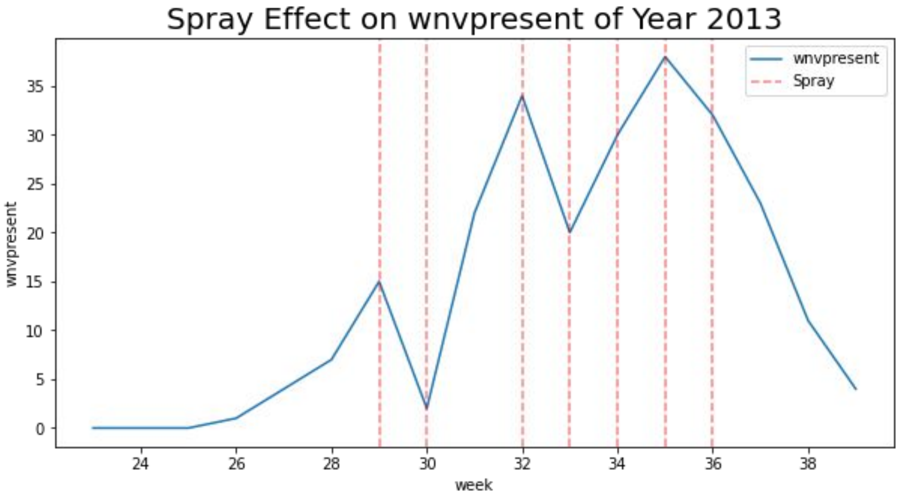
- In 2007 and 2009, there were no spray data shows in City of Chicago. Spray datasets shows only in 2011, and in 2013, the spraying areas were expanded.

EDA - Spray Effect

- Spray dataset limited, only have 2 years of data where only 2 sprays carried out in 2011.

- Spray does have the effect on reducing number of mosquitos in both year 2011 and 2013, we cannot conclude that it have strong effect on reducing WMV present





Feature Engineering

1 **Extract Year / Month / Week / Day**

From Date where WNV test was performed

2 **Label Encoding Mosquitos Species**

Culex pipiens, culex restuans and the rest of species

3 **Clustering Trap Locations**

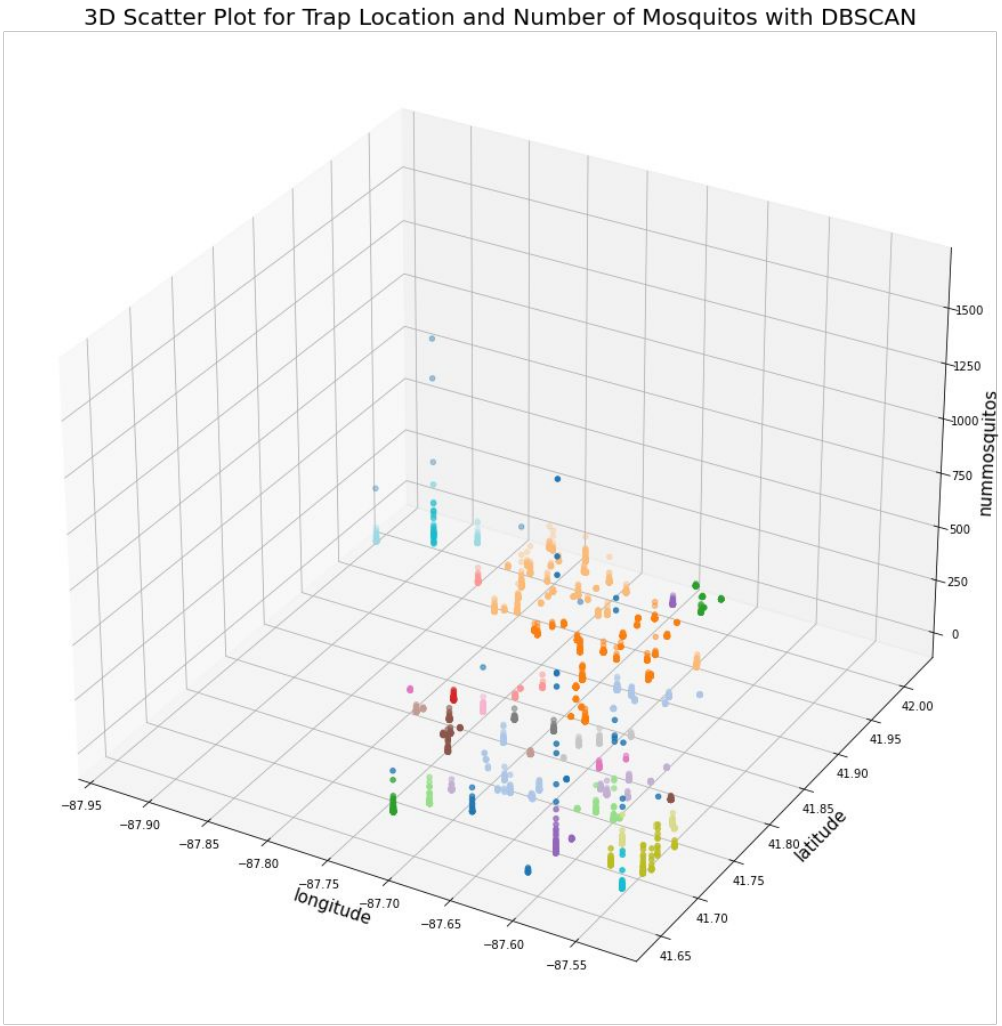
Based on Trap Geographic Coordinates and Numbers of Mosquitos, perform DBSCAN clustering

Clustering Trap Location

★ There are total of 136 unique trap location in train datasets

★ Unsupervised Learning - DBSCAN to cluster **Longitude, Latitude & Numbers of Mosquitos** into different trap clusters

★ After our clustering, reduce the trap cluster to 38 for our modeling



Modeling - Logistic Regression & XGBoost Classifier

Using logistic regression as the base model and XGBoost classifier as the model, we achieve an AUC score of 0.82 and a kaggle score of 0.71

```
====== XGBClassifier's Metrics ======
Train Score: 0.961035664424193
Test Score: 0.9440426341834792
Precision Score: 0.42105263157894735
Recall Score: 0.17391304347826086
Average Precision: 0.11662205280666621
f1-Score: 0.24615384615384617
roc_auc Score: 0.8254857605347589
```

Modeling - Using PyCaret

	Data Type
species	Numeric
latitude	Numeric
longitude	Numeric
month	Categorical
week	Categorical
day	Numeric

PyCaret wrongly detects species and day as numeric variables instead of categorical variables so we define our own set of categorical variables below.

```
cat_features = ['species', 'month', 'week', 'day', 'cluster_0', 'cluster_1',
                'cluster_2', 'cluster_3', 'cluster_4', 'cluster_5', 'cluster_6',
                'cluster_7', 'cluster_8', 'cluster_9', 'cluster_10', 'cluster_11',
                'cluster_12', 'cluster_13', 'cluster_14', 'cluster_15', 'cluster_16',
                'cluster_17', 'cluster_18', 'cluster_19', 'cluster_20', 'cluster_21',
                'cluster_22', 'cluster_23', 'cluster_24', 'cluster_25', 'cluster_26',
                'cluster_27', 'cluster_28', 'cluster_29', 'cluster_30', 'cluster_31',
                'cluster_32', 'cluster_33', 'cluster_34', 'cluster_35', 'cluster_36',
                'cluster_37']
```

Modeling - PyCaret Results

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9153	0.8296	0.3442	0.2741	0.3043	0.2601	0.2624	0.7050
xgboost	Extreme Gradient Boosting	0.9162	0.8234	0.3314	0.2728	0.2983	0.2545	0.2562	1.8970
lda	Linear Discriminant Analysis	0.7210	0.8093	0.7547	0.1319	0.2245	0.1471	0.2314	0.3710
gbc	Gradient Boosting Classifier	0.8534	0.8089	0.5201	0.1882	0.2760	0.2144	0.2489	1.9600
lr	Logistic Regression	0.7293	0.8085	0.7318	0.1327	0.2246	0.1476	0.2276	2.6710
ada	Ada Boost Classifier	0.8107	0.7922	0.5605	0.1529	0.2400	0.1706	0.2183	0.5470
rf	Random Forest Classifier	0.9111	0.7743	0.2958	0.2382	0.2630	0.2165	0.2183	0.7930
knn	K Neighbors Classifier	0.7565	0.7304	0.5890	0.1239	0.2047	0.1279	0.1832	0.3070
nb	Naive Bayes	0.4369	0.6845	0.9208	0.0808	0.1485	0.0560	0.1521	0.0820
et	Extra Trees Classifier	0.9107	0.6795	0.2932	0.2349	0.2598	0.2131	0.2151	1.0240
dt	Decision Tree Classifier	0.9102	0.6362	0.2701	0.2253	0.2435	0.1967	0.1986	0.2170
qda	Quadratic Discriminant Analysis	0.3330	0.6332	0.9693	0.0721	0.1342	0.0389	0.1330	0.2310
dummy	Dummy Classifier	0.9467	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0710
svm	SVM - Linear Kernel	0.6261	0.0000	0.7136	0.1100	0.1766	0.1055	0.1772	0.4670
ridge	Ridge Classifier	0.7214	0.0000	0.7624	0.1330	0.2264	0.1492	0.2350	0.0800

Baseline Model - Dummy Classifier with AUC of 0.5.

Best Model is lightgbm with AUC score of 0.8296

AUC score selected as evaluation as AUC measures the performance of the model at distinguishing between the positive and negative classes.

For this problem, we want to clearly identify the true positive and the true negatives so we optimize AUC.

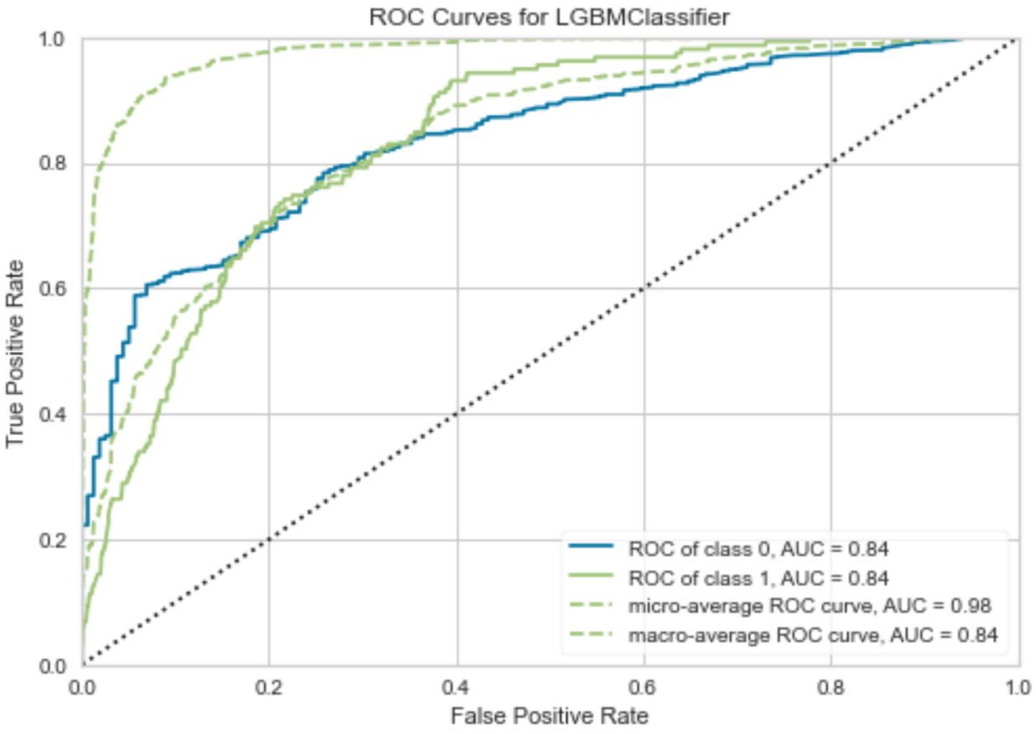
Modeling - Hyperparameter Tuning

```
best = automl(optimize = 'AUC')
best
```

```
LGBMClassifier(boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
               importance_type='split', learning_rate=0.1, max_depth=-1,
               min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0,
               n_estimators=100, n_jobs=-1, num_leaves=31, objective=None,
               random_state=1, reg_alpha=0.0, reg_lambda=0.0, silent='warn',
               subsample=1.0, subsample_for_bin=200000, subsample_freq=0)
```

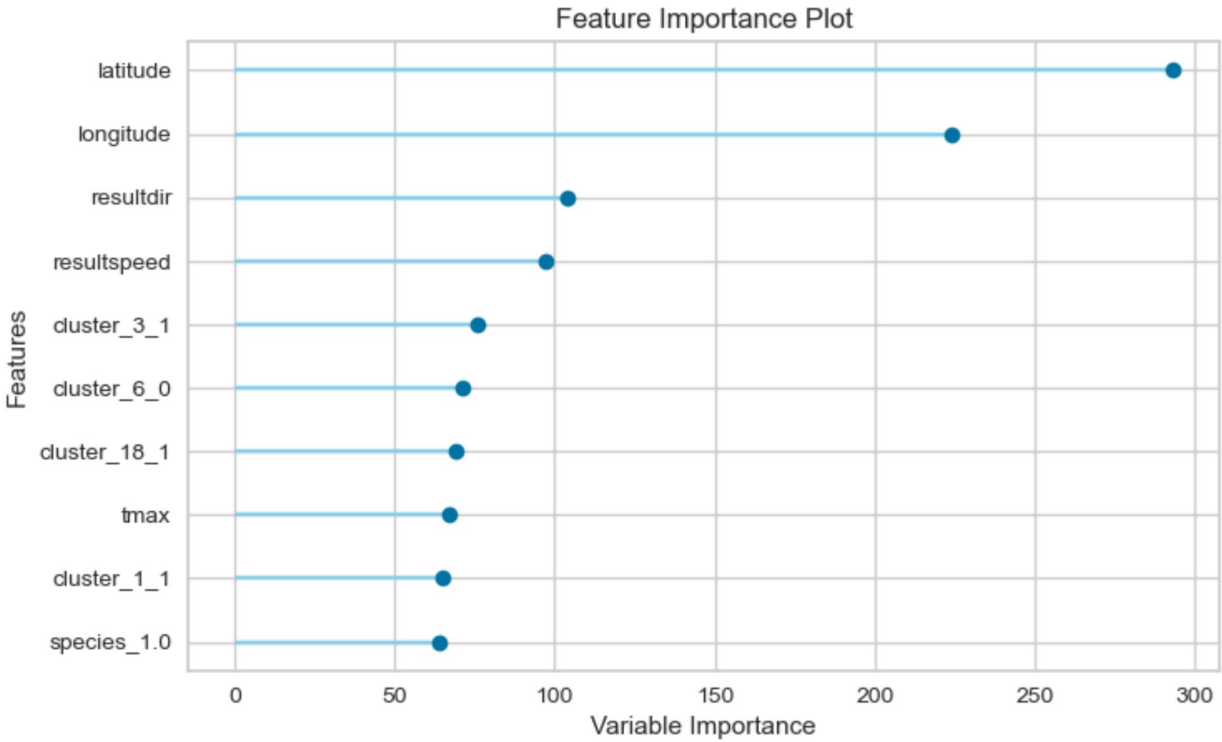
Using automl to perform hyperparameter tuning

Modeling - AUC Graph



AUC Score of 0.84

Modeling - Feature Importance

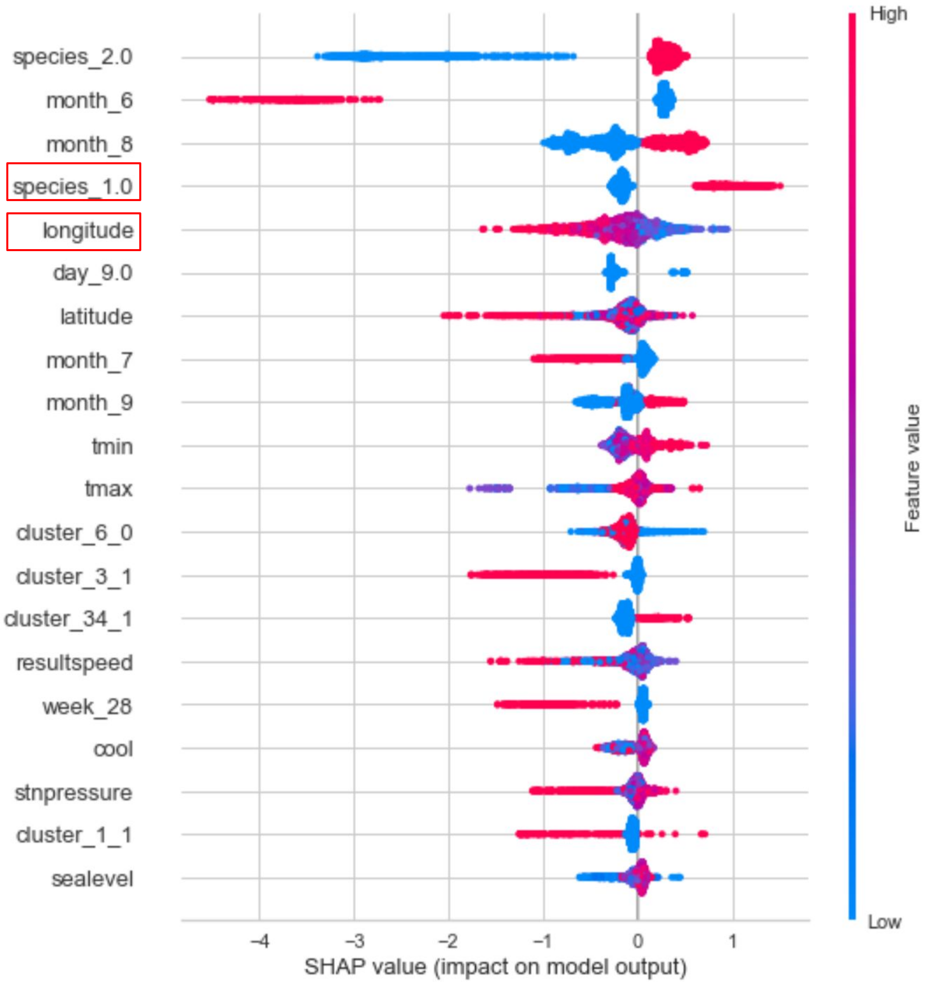


The most important features are latitude and longitude which is the location of the virus.

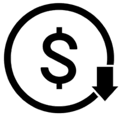
Other important features include resultdir, resultspeed, tmax, speices_1 and a few clusters.

Modeling - SHAP Values

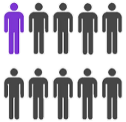
The features that contribute highest to a positive SHAP values are longitude and spieces_1.0 which is the location of the virus and the species.



The Cost of West Nile Virus

$800 million 

That’s how much in hospitalisation and lost productivity the West Nile Virus has cost the USA from 1999 - 2013.

1 in 150 

patients with the West Nile Virus that will develop severe symptoms.

$7,500 

The mean hospitalisation and lost productivity cost for mild cases.

$80,000 

The mean hospitalisation and lost productivity cost for severe cases.

Cost Benefit Analysis

Assumptions

Spray Cost	Cost of not Spraying
• Zenivex costs $0.92/acre • Pest control worker earns $20/hour • 8pm - 1am (5 hour spray window) • 149 traps - all will have spray operations • 1 worker per trap • 1km radius spray per trap • Spray will be 7 times a year	• All cases are non-severe • Mean cost for non-severe cases - $7,500 • 200 additional cases if no spray conducted
$840,000	**$1,500,000**

$660,000 savings annually

Conclusion

Best model: Light Gradient Boosting Machine with AUC score 0.8260

Further research

- Insight number of mosquito caught per trap
- the life cycle of mosquito
- Weather pattern

Better Adoption

Technology solution

Drone used in mosquito control

Personal precautious

- Reduce the standing water
- Use repellent
- Wear covered clothes


Q&A