West Nile Virus Prediction

Group 1

Jerome, Calvin, Jun Pin

Content

Background & Problem Statement

Data Cleaning

EDA

Feature Engineering

Modeling

Results

Conclusions & Recommendations

Background & Problem Statement

West Nile Virus (WNV) is the leading cause of mosquito-borne disease in the continental United States. It is most commonly spread to people by the bite of an infected mosquito. There are no vaccines to prevent or specific medications to treat WNV in people. Based on research, we understand that about 1 out of 150 infected people develop a serious, sometimes fatal, illness. Due to the recent epidemic of West Nile Virus in Chicago, the Department of Public Health has set up a surveillance and control system in which data on the mosquito population has been collected over time. We, as data scientists, have been engaged to analyze the collected data to investigate and predict the presence of WNV. We will also be doing a cost-benefit analysis in relation to the spraying of pesticides to control the spread of WNV.

Datasets

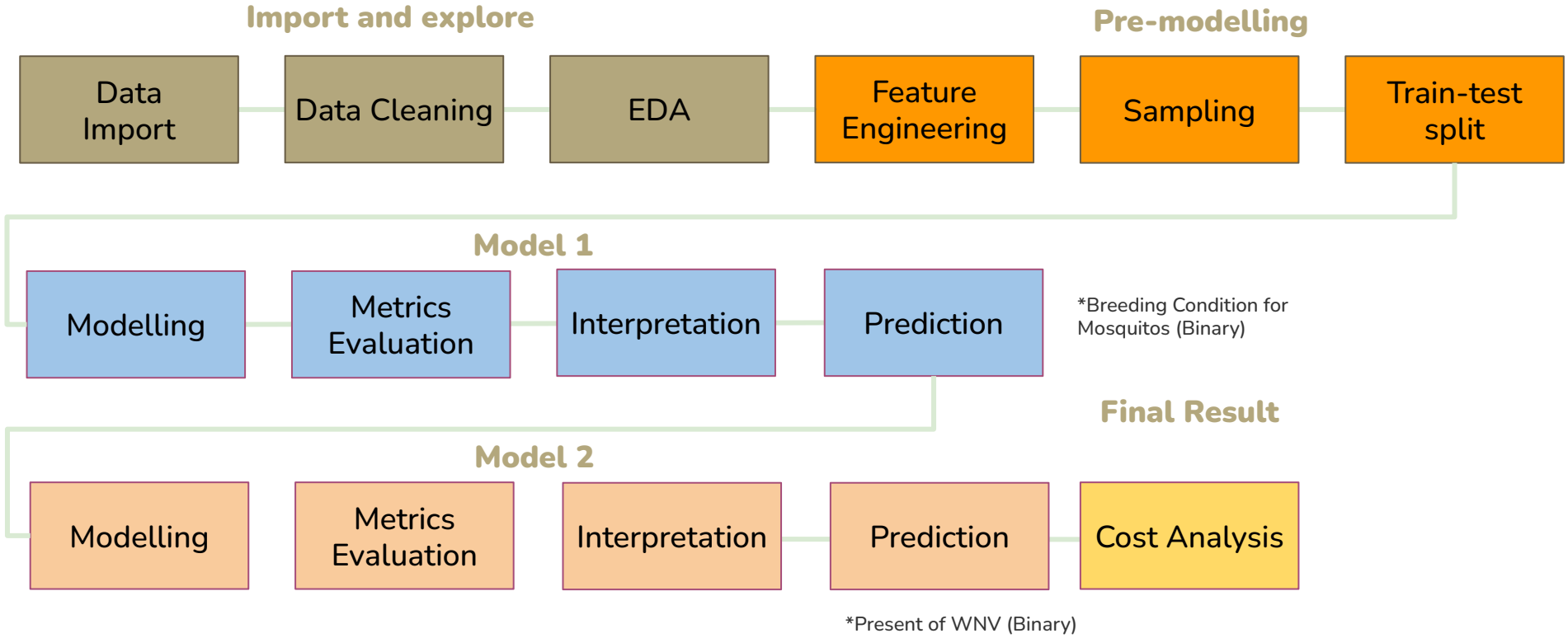
Train Dataset - There are 12 columns (i.e. Date, Address, Trap, NumMosquitos, WNVPresent, etc.) in the dataset and consists of data from 2007, 2009, 2011 and 2013.

Test Dataset - There are 11 columns in the dataset and consists of data from 2008, 2010, 2012 and 2014. The columns are similar to the train dataset with the addition of the ‘Id’ column and removal of the ‘NumMosquitos’ and ‘WNVPresent’ columns.

Spray Dataset - It consists of spraying efforts data in 2011 and 2013 which includes the date, time as well as the latitude and longitude (i.e. location) of the spray.

Weather Dataset - It consists of weather data from 2007 to 2014 and a total of 22 columns in the dataset.

Process Flowchart



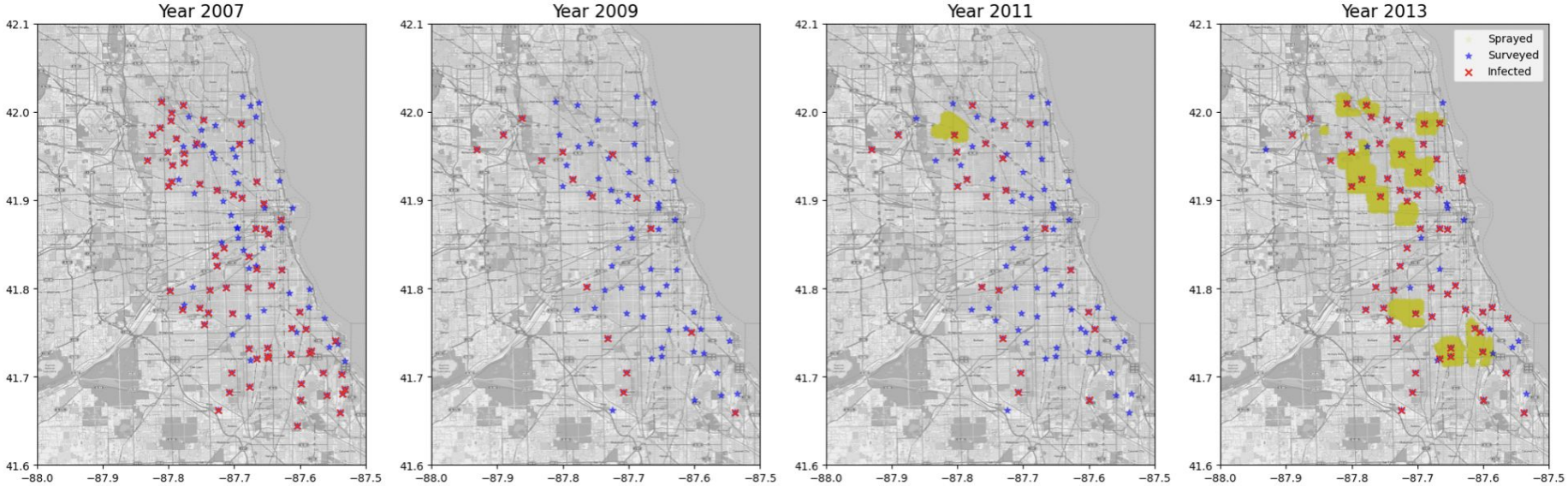
Data Cleaning (Continued)

Station	0
Date	0
Tmax	0
Tmin	0
Tavg	11
Depart	1472
DewPoint	0
WetBulb	4
Heat	11
Cool	11
Sunrise	1472
Sunset	1472
CodeSum	0
Depth	1472
Water1	2944
SnowFall	1472
PrecipTotal	2
StnPressure	4
SeaLevel	9
ResultSpeed	0
ResultDir	0
AvgSpeed	3

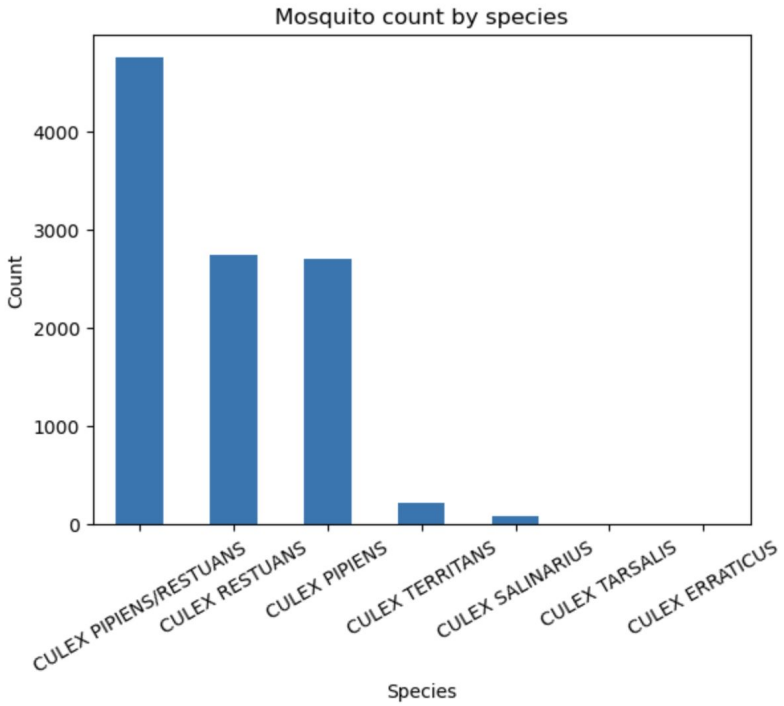
- For the Tavg column, we input the missing values by taking the average of the Tmax and Tmin values.
- For the Wetbulb, Heat, Cool, Sealevel columns, we noted that the values on the same days for both stations are similar. Therefore, we will fill in the missing values as the same values of the corresponding station on the same day.
- We noted that the missing values for 'Sunrise' and 'Sunset' columns are all from Station 2. Therefore, we will input the values as the same values of Station 1 on the same day.
- For the Precipitation column, we input 0 for the ‘NA’ values and based on outside research, we replace 0.005 for the ‘T’ values.
- For the station pressure and average speed columns, we use the median values to input for the missing values.

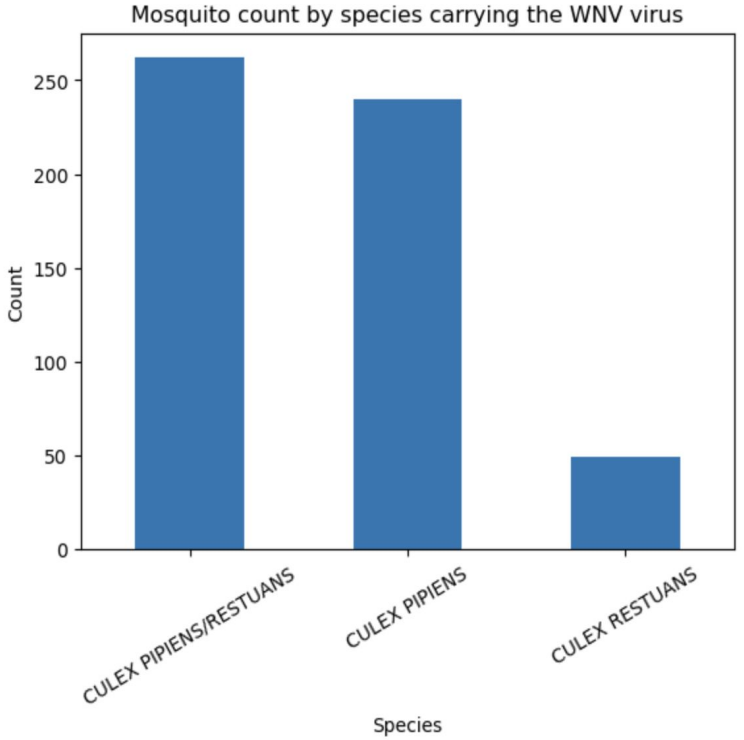
EDA

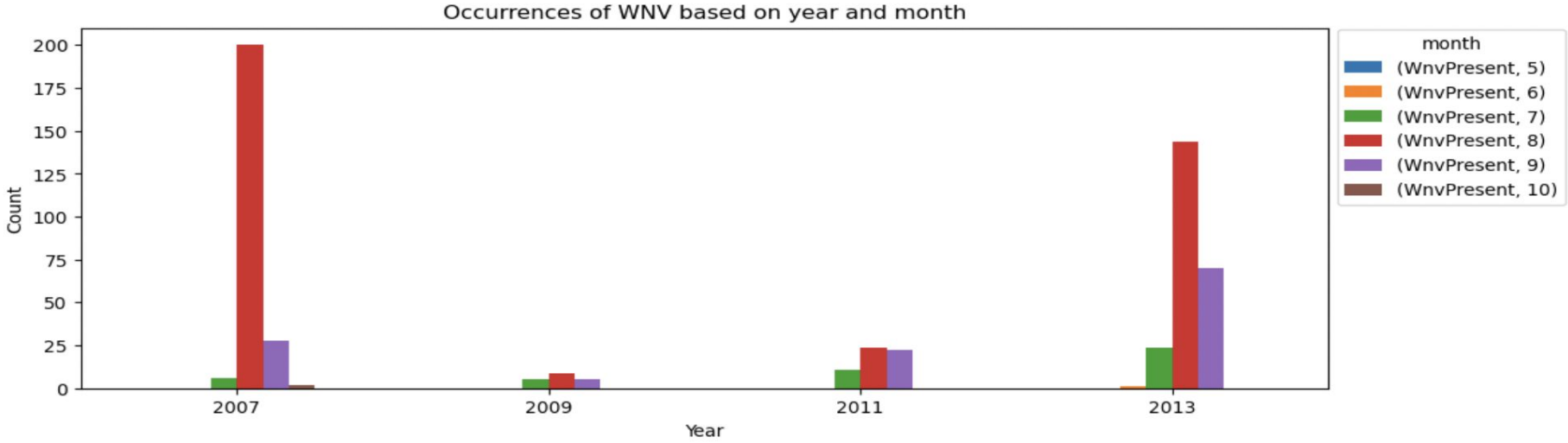
Maps of Surveyed Traps Location, Infected Areas & Sprayed Areas

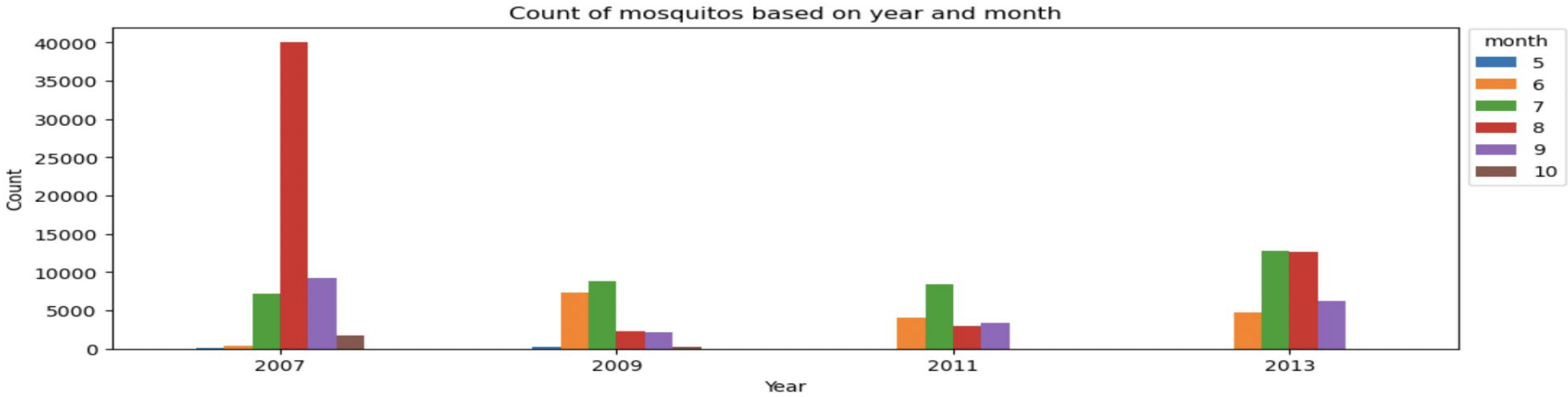


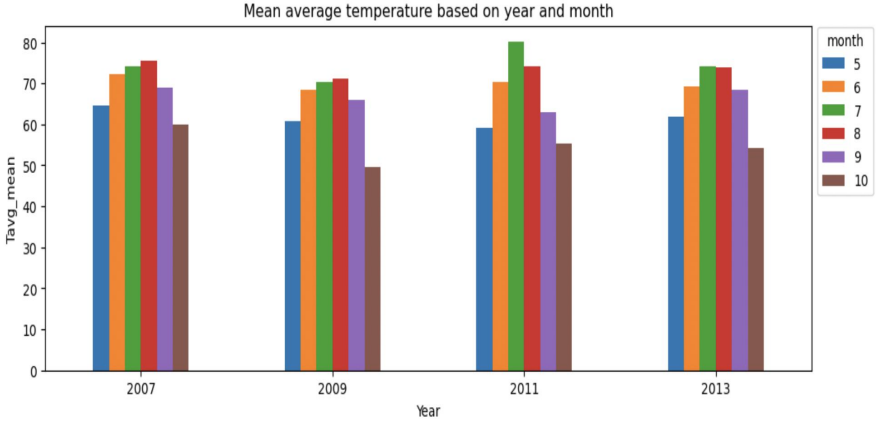
EDA

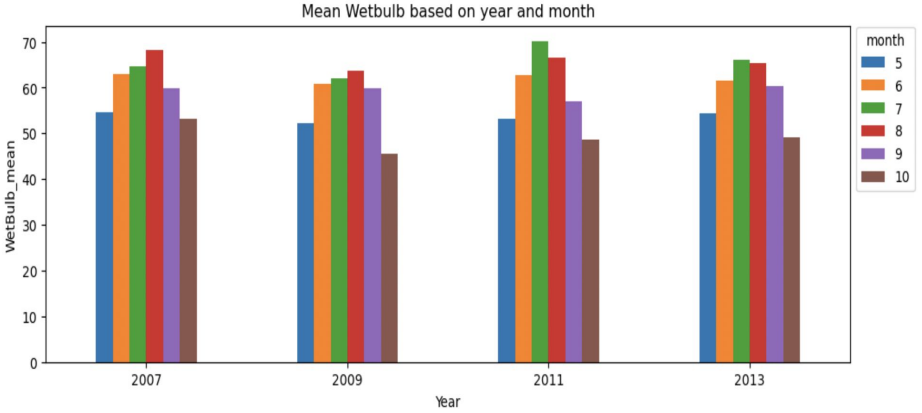

Mosquito count by species
Count
0
1000
2000
3000
4000
CULEX PIPIENS/RESTUANS
CULEX RESTUANS
CULEX PIPIENS
CULEX TERRITANS
CULEX SALINARIUS
CULEX TARSALIS
CULEX ERRATICUS
Species

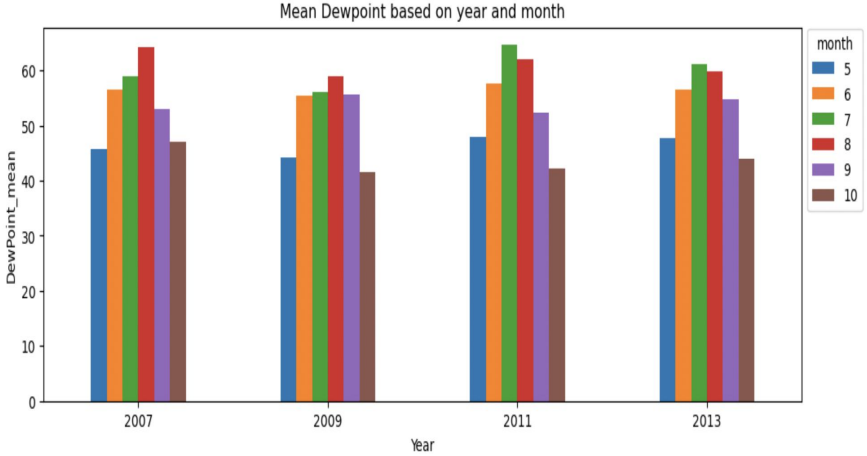

Mosquito count by species carrying the WNV virus
Count
0
50
100
150
200
250
CULEX PIPIENS/RESTUANS
CULEX PIPIENS
CULEX RESTUANS
Species

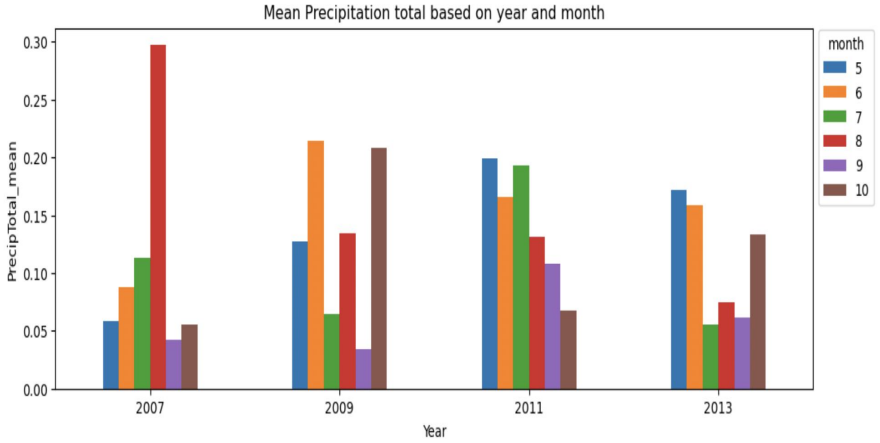

Occurrences of WNV based on year and month
Count
Year
2007
2009
2011
2013
month
(WnvPresent, 5)
(WnvPresent, 6)
(WnvPresent, 7)
(WnvPresent, 8)
(WnvPresent, 9)
(WnvPresent, 10)


Count of mosquitos based on year and month
Count
Year
2007
2009
2011
2013
month
5
6
7
8
9
10

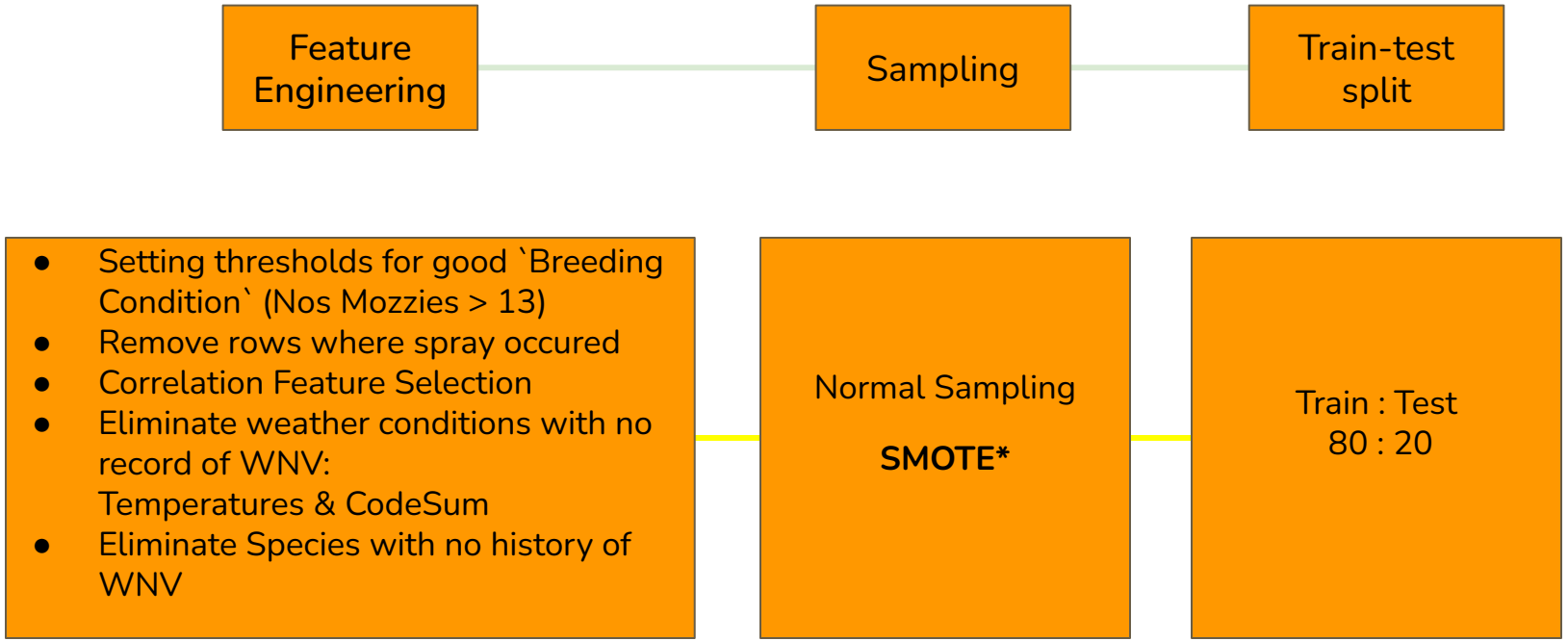

Mean average temperature based on year and month
Tavg_mean
Year
month
5
6
7
8
9
10
2007
2009
2011
2013


Mean Wetbulb based on year and month
WetBulb_mean
Year
month
5
6
7
8
9
10
2007
2009
2011
2013


Mean Dewpoint based on year and month
DewPoint_mean
Year
month
5
6
7
8
9
10
2007
2009
2011
2013

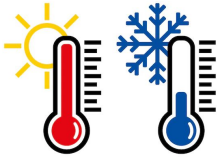

Mean Precipitation total based on year and month
PrecipTotal_mean
Year
month
5
6
7
8
9
10
2007
2009
2011
2013

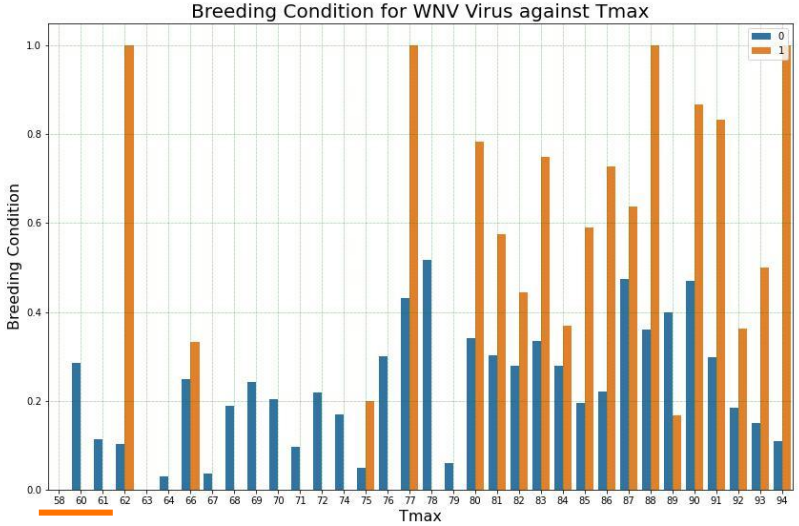
Pre-Modeling

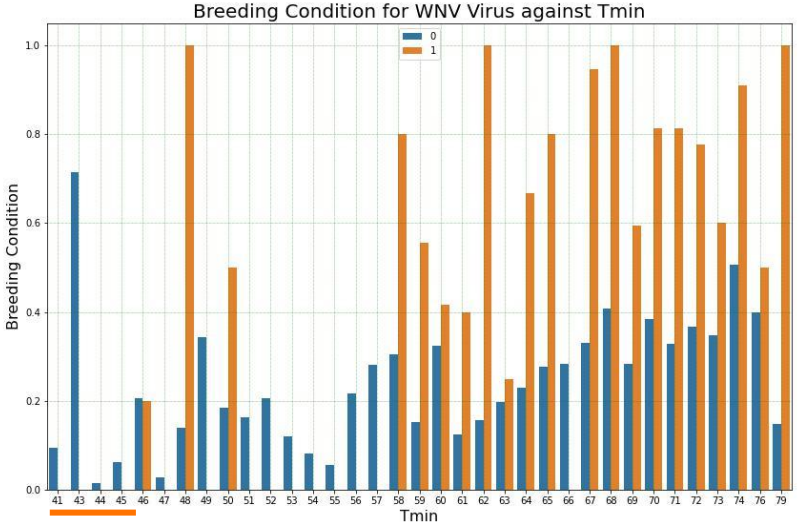


*Synthetic Minority Over-sampling Technique
synthetic samples are generated for the minority class.

Feature Engineering - Temperature



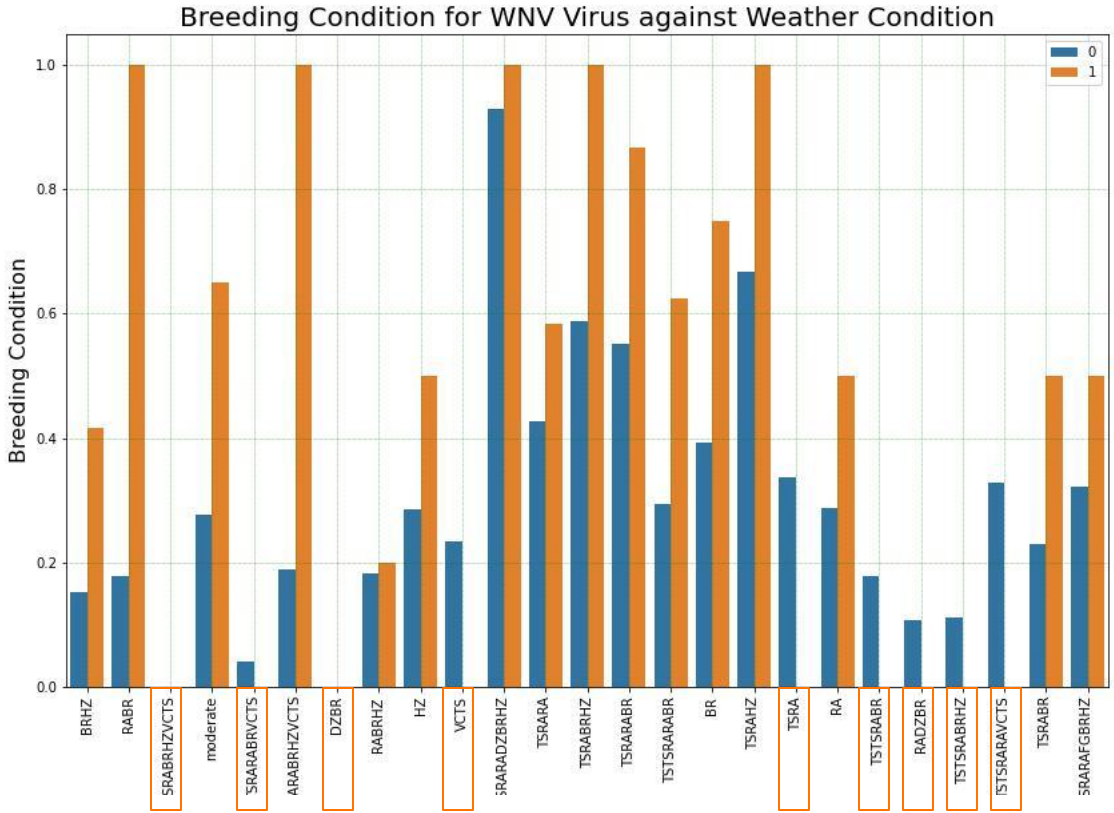




Removed temperature where WNV is not found

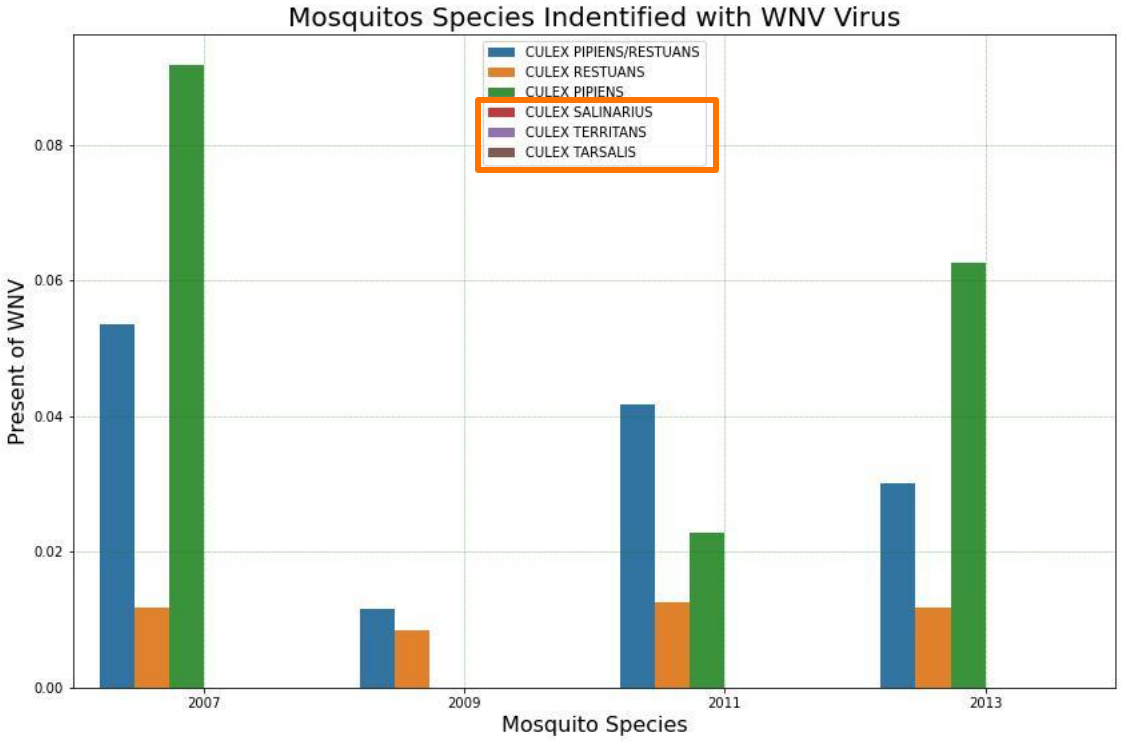
Feature Engineering - Weather Type

Drop Weather condition where WNV was not found



Feature Engineering - Mosquito Species

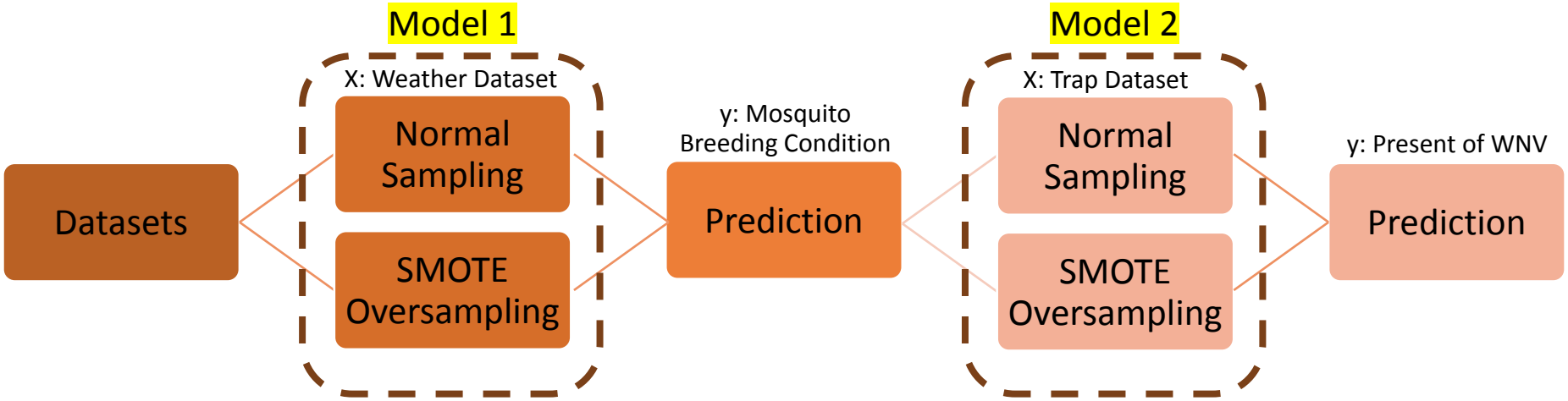
Drop mosquito species where WNV was not found



Import & Explore

Data Import	Data Cleaning	EDA
Traps Test Spray Weather	• Remove null values • Converting index to date • Identify the Tavg from Tmin and Tmax • Replace null values in weather dataset with values from another station • Inpute missing values with median values • Remove duplicate posts (spams) • Regex on Weather Condition to replace null	• Location of Traps • Number of Traps over the Year • Percentage of WNV • Mosquitos Counts by Species w/ WNV • Occurence of WNV • Weather features by Year

Modelling Process

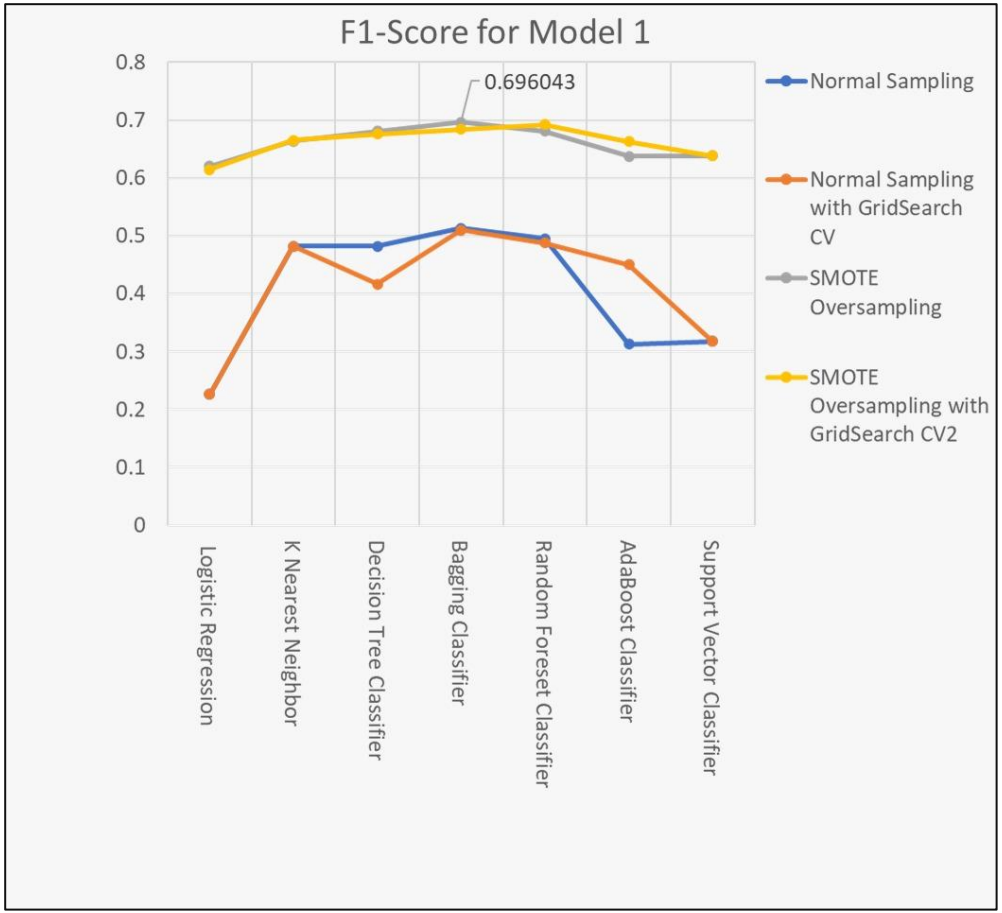


Classification Methods:
Logistic Regression
K Nearest Neighbour
Decision Tree
Bagging
Random Forest
AdaBoost
Support Vector

Modelling:
Default Setting
GridSearchCV(scoring = 'f1_micro')

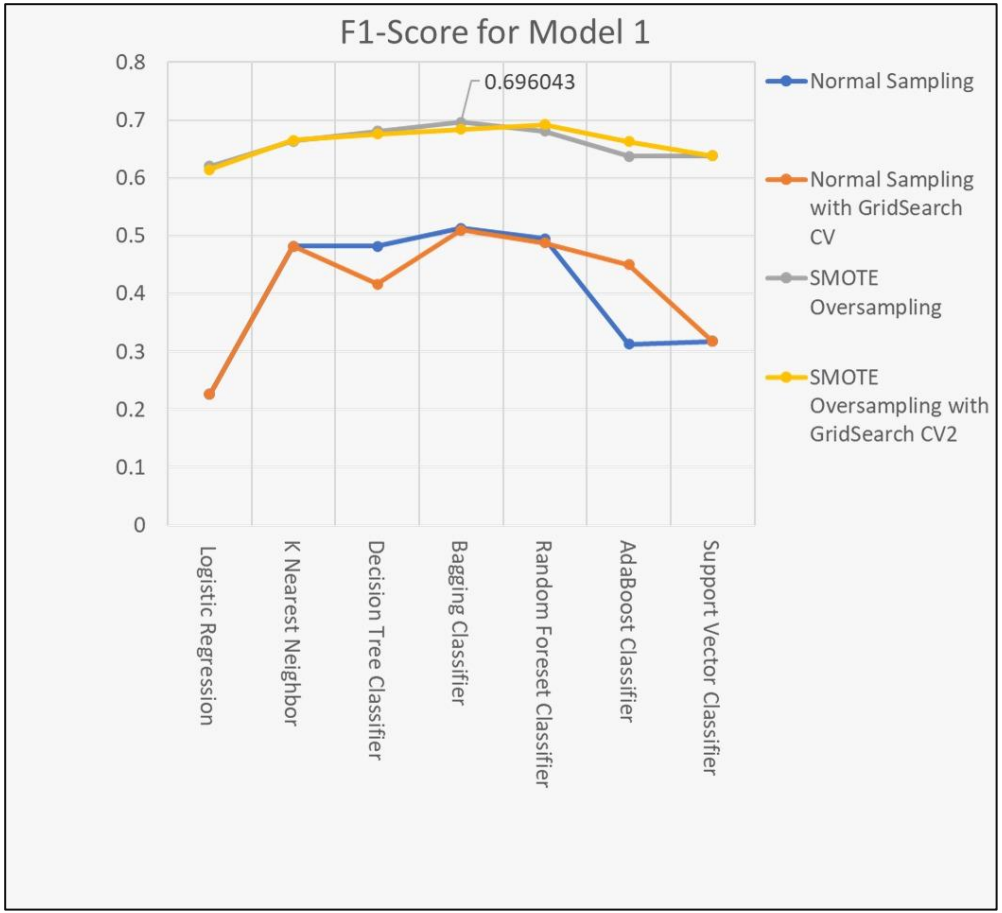
Model 1 - Prediction of Breeding Condition for Mosquito

No	Model	Remark	Train Score	Test Score	Precision	Recall	F1 Score
1	Base Model | Logistic Regression	None	0.714533	0.720477	0.714286	0.134298	0.226087
2	K Nearest Neighbor	None	0.747997	0.738065	0.604361	0.400826	0.481988
3	Decision Tree Classifier	None	0.774548	0.766332	0.739316	0.357438	0.481894
4	Bagging Classifier	None	0.774234	0.762563	0.681507	0.411157	0.512887
5	Random Foreset Classifier	None	0.774548	0.761935	0.696629	0.384298	0.495340
6	AdaBoost Classifier	None	0.738256	0.734925	0.738462	0.198347	0.312704
7	Support Vector Classifier	None	0.732129	0.738693	0.769841	0.200413	0.318033
8	Logistic Regression	GSCV	0.211714	0.226087	0.714286	0.134298	0.226087
9	K Nearest Neighbor	GSCV	0.512165	0.481988	0.604361	0.400826	0.481988
10	Decision Tree Classifier	GSCV	0.468694	0.416667	0.744681	0.289256	0.416667
11	Bagging Classifier	GSCV	0.546431	0.509603	0.670034	0.411157	0.509603
12	Random Foreset Classifier	GSCV	0.524046	0.487805	0.708661	0.371901	0.487805
13	AdaBoost Classifier	GSCV	0.486860	0.449857	0.733645	0.324380	0.449857
14	Support Vector Classifier	GSCV	0.294580	0.318033	0.769841	0.200413	0.318033
15	Logistic Regression	SMOTE	0.636754	0.620767	0.620596	0.620596	0.620596
16	K Nearest Neighbor	SMOTE	0.668586	0.660045	0.656085	0.672087	0.663989
17	Decision Tree Classifier	SMOTE	0.713625	0.695711	0.715852	0.648600	0.680569
18	Bagging Classifier	SMOTE	0.712383	0.694808	0.692927	0.699187	0.696043
19	Random Foreset Classifier	SMOTE	0.713625	0.695711	0.715852	0.648600	0.680569
20	AdaBoost Classifier	SMOTE	0.678180	0.652370	0.665358	0.612466	0.637817
21	Support Vector Classifier	SMOTE	0.641156	0.632054	0.627622	0.648600	0.637939
22	Logistic Regression	SMOTE & GSCV	0.630977	0.614964	0.621198	0.608853	0.614964
23	K Nearest Neighbor	SMOTE & GSCV	0.673599	0.664882	0.656966	0.672990	0.664882
24	Decision Tree Classifier	SMOTE & GSCV	0.674957	0.676143	0.588235	0.794941	0.676143
25	Bagging Classifier	SMOTE & GSCV	0.701561	0.684137	0.709709	0.660343	0.684137
26	Random Foreset Classifier	SMOTE & GSCV	0.712760	0.691858	0.689068	0.694670	0.691858
27	AdaBoost Classifier	SMOTE & GSCV	0.682716	0.662438	0.664545	0.660343	0.662438
28	Support Vector Classifier	SMOTE & GSCV	0.645952	0.637939	0.627622	0.648600	0.637939



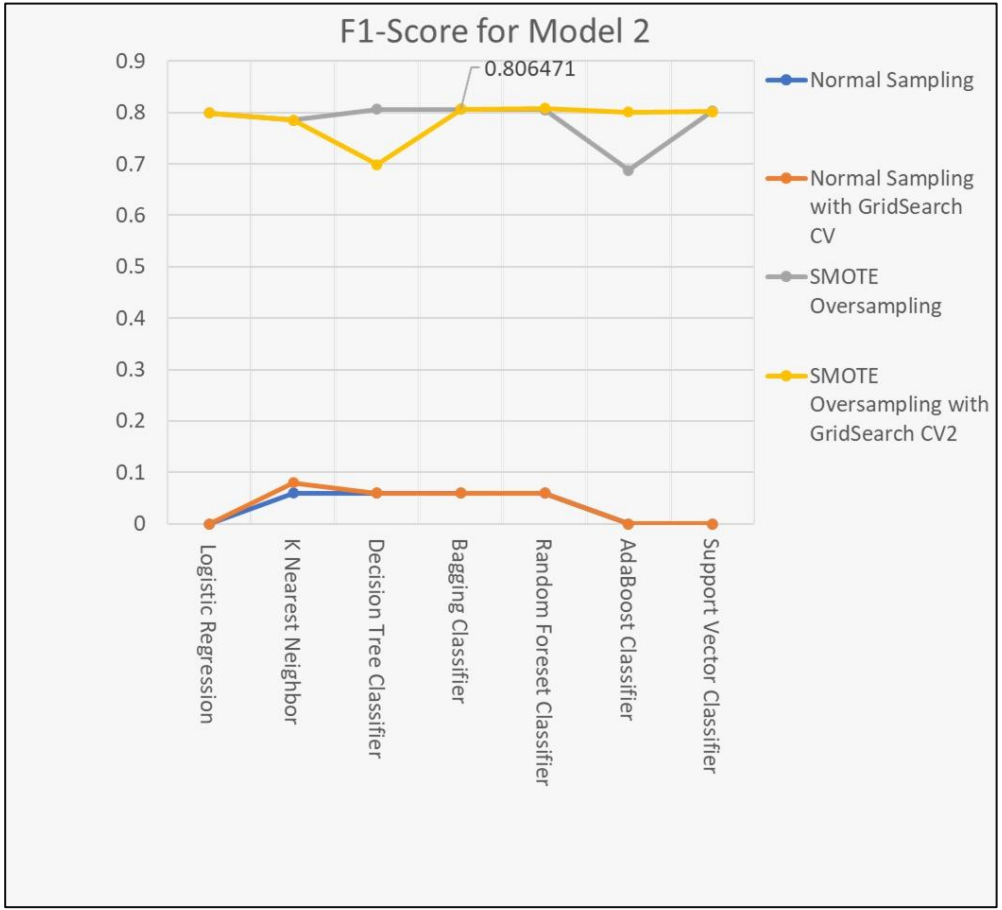
Model 1 - Prediction of Breeding Condition for Mosquito

- SMOTE Oversampling outperforms Normal Sampling in F1-Score for all Methods
- Logistic regressions method has worst F1-Score
- Bagging Classifier outperforms most models
- Selected Model for Prediction:
  SMOTE Oversampling Bagging Classifier



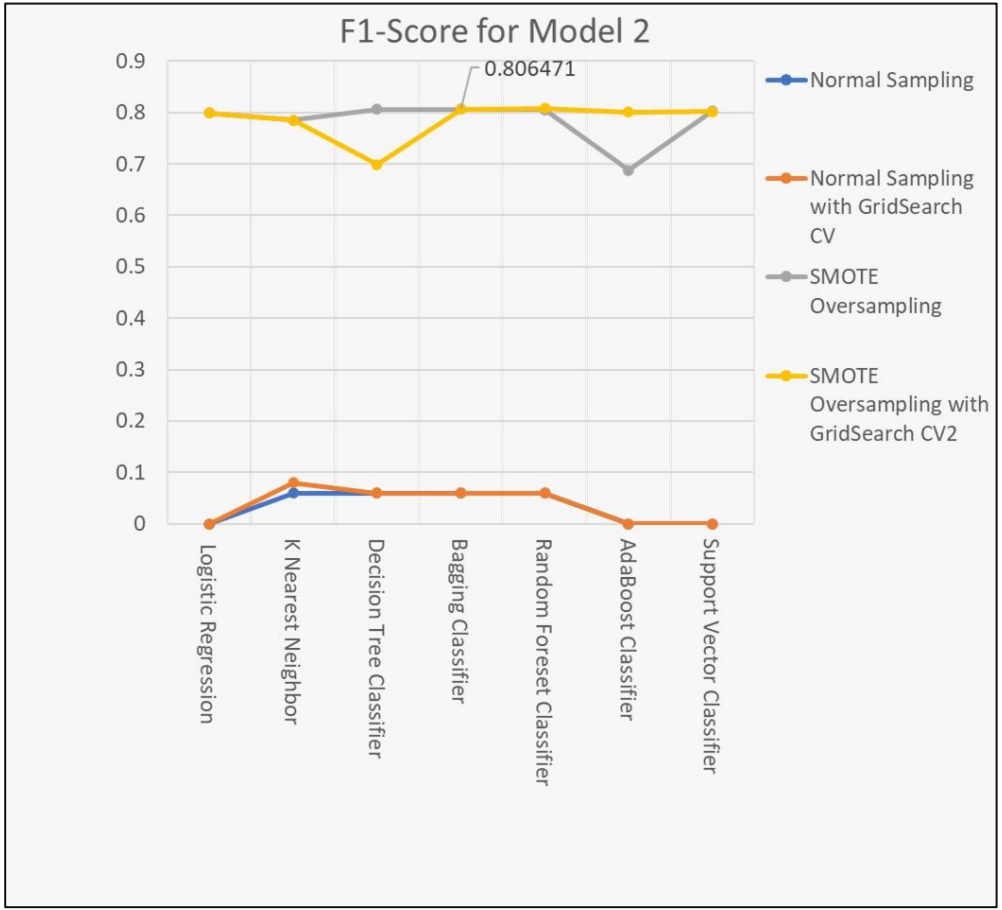
Model 2 - Prediction of Present of West Nile Virus

No	Model	Remark	Train Score	Test Score	Precision	Recall	F1 Score
1	Base Model | Logistic Regression	None	0.958995	0.959171	0	0	0
2	K Nearest Neighbor	None	0.959309	0.960427	1	0.030769	0.059701
3	Decision Tree Classifier	None	0.959466	0.960427	1	0.030769	0.059701
4	Bagging Classifier	None	0.959466	0.960427	1	0.030769	0.059701
5	Random Foreset Classifier	None	0.959466	0.960427	1	0.030769	0.059701
6	AdaBoost Classifier	None	0.958995	0.959171	0	0	0
7	Support Vector Classifier	None	0.958995	0.959171	0	0	0
8	Logistic Regression	GSCV	0.958995	0.959171	0	0	0
9	K Nearest Neighbor	GSCV	0.958052	0.956658	0.3	0.046154	0.08
10	Decision Tree Classifier	GSCV	0.959309	0.960427	1	0.030769	0.059701
11	Bagging Classifier	GSCV	0.959466	0.960427	1	0.030769	0.059701
12	Random Foreset Classifier	GSCV	0.959466	0.960427	1	0.030769	0.059701
13	AdaBoost Classifier	GSCV	0.958995	0.959171	0	0	0
14	Support Vector Classifier	GSCV	0.958995	0.959171	0	0	0
15	Logistic Regression	SMOTE	0.783766	0.787095	0.755245	0.849279	0.799506
16	K Nearest Neighbor	SMOTE	0.764682	0.771372	0.740698	0.834862	0.784966
17	Decision Tree Classifier	SMOTE	0.801048	0.792663	0.755289	0.865662	0.806718
18	Bagging Classifier	SMOTE	0.800557	0.792335	0.754857	0.865662	0.806471
19	Random Foreset Classifier	SMOTE	0.800967	0.792991	0.758382	0.859764	0.805897
20	AdaBoost Classifier	SMOTE	0.701859	0.707173	0.736173	0.645478	0.687849
21	Support Vector Classifier	SMOTE	0.796052	0.79168	0.76115	0.849934	0.803096
22	Logistic Regression	SMOTE & GSCV	0.79736	0.799506	0.755245	0.849279	0.799506
23	K Nearest Neighbor	SMOTE & GSCV	0.78114	0.78545	0.74156	0.834862	0.78545
24	Decision Tree Classifier	SMOTE & GSCV	0.693832	0.697347	0.695709	0.70118	0.698433
25	Bagging Classifier	SMOTE & GSCV	0.800885	0.792663	0.755289	0.865662	0.806718
26	Random Foreset Classifier	SMOTE & GSCV	0.799246	0.792663	0.752402	0.872215	0.807891
27	AdaBoost Classifier	SMOTE & GSCV	0.7818	0.784802	0.744789	0.866317	0.800969
28	Support Vector Classifier	SMOTE & GSCV	0.795806	0.79168	0.764252	0.843381	0.801869

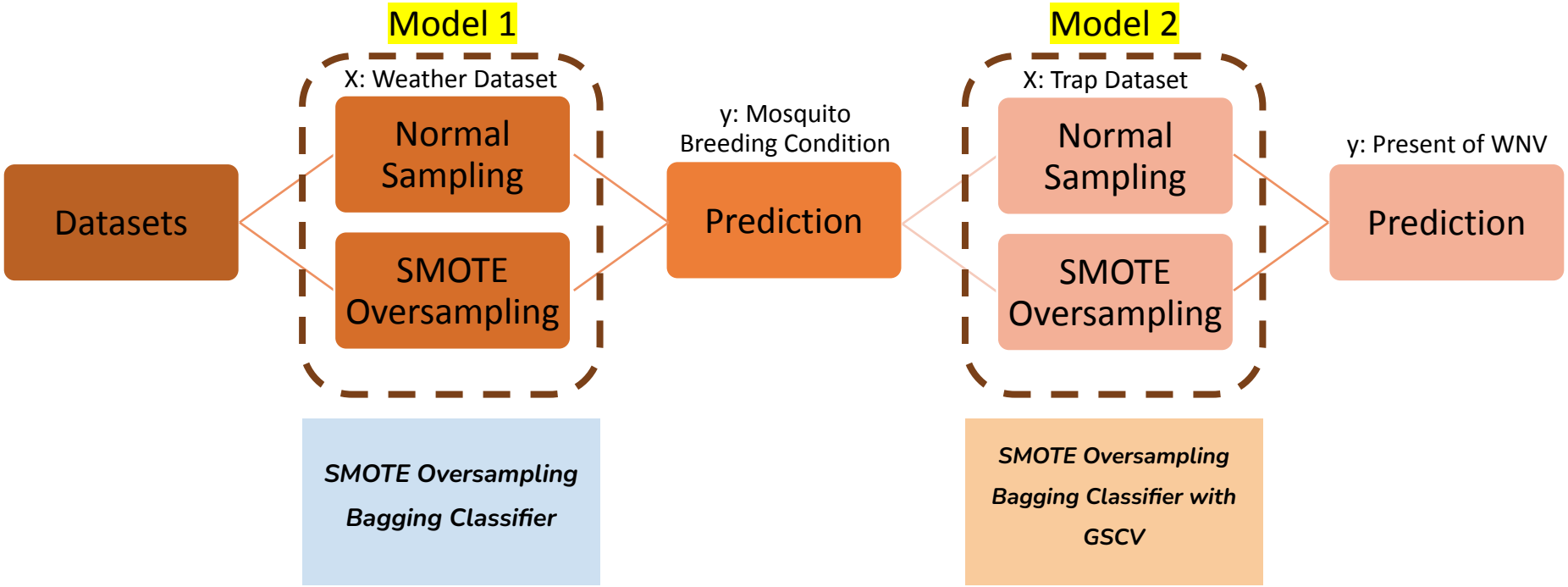


Model 2 - Prediction of Present of West Nile Virus

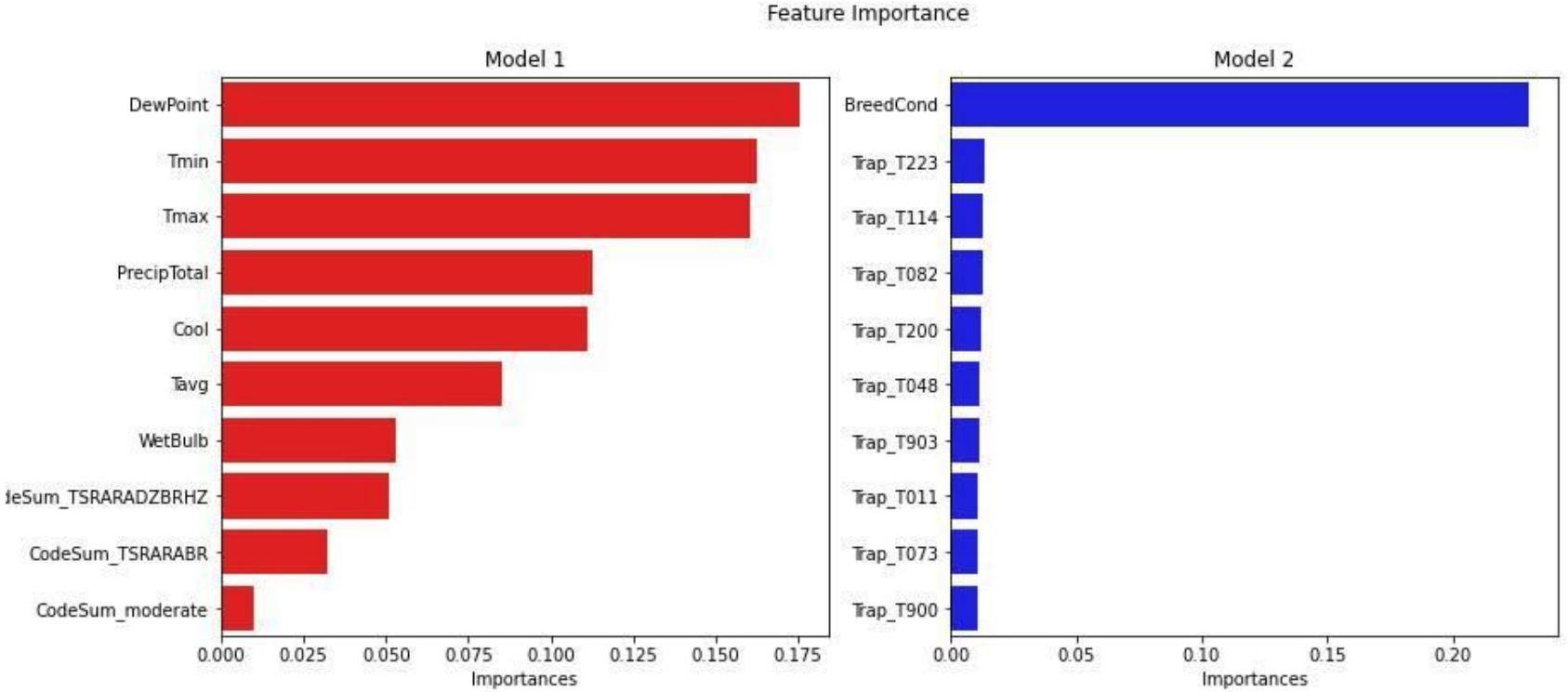
- SMOTE Oversampling outperforms Normal Sampling in F1-Score for all Methods
- Datasets are overly unbalanced with some F1 score being 0
- Selected Model for Prediction: *SMOTE Oversampling Bagging Classifier with GSCV*



Modelling Process

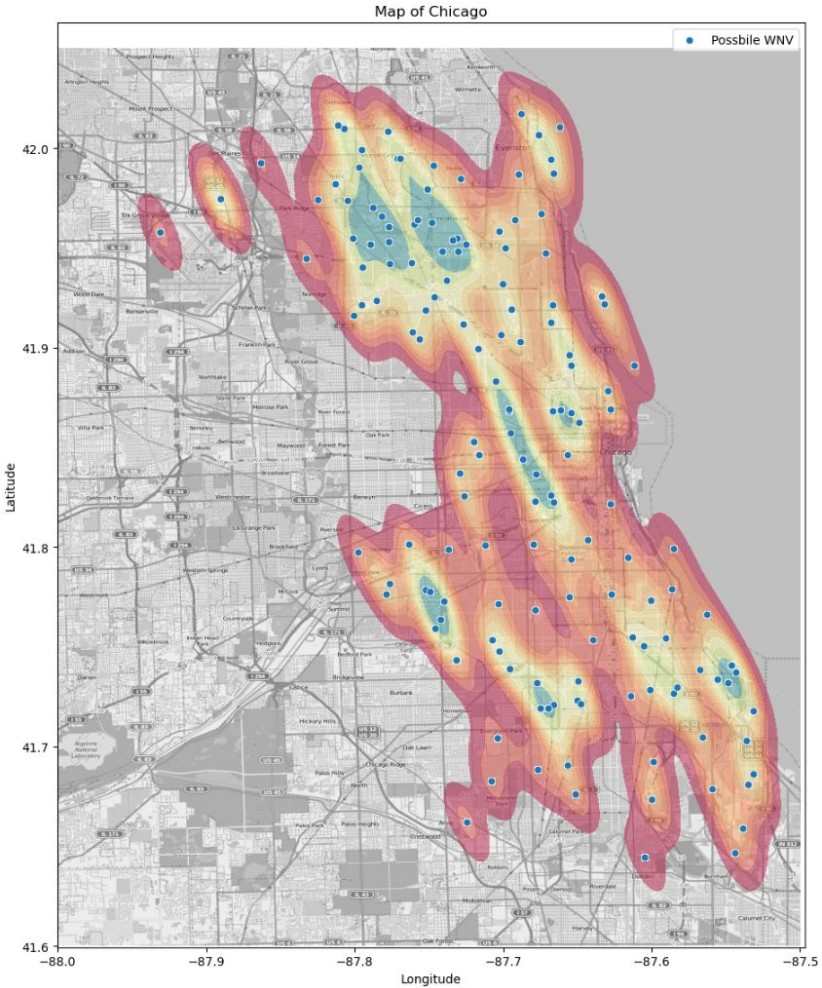


Results



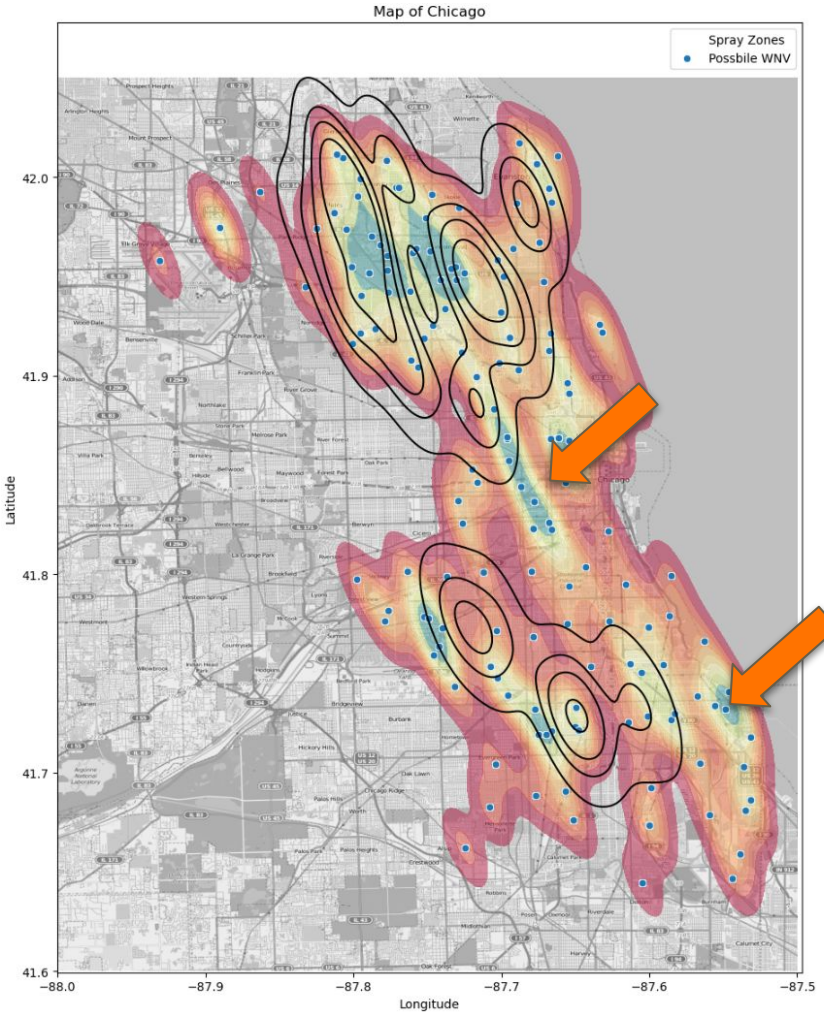
Prone to Temperature, DewPoint, Precip

Results



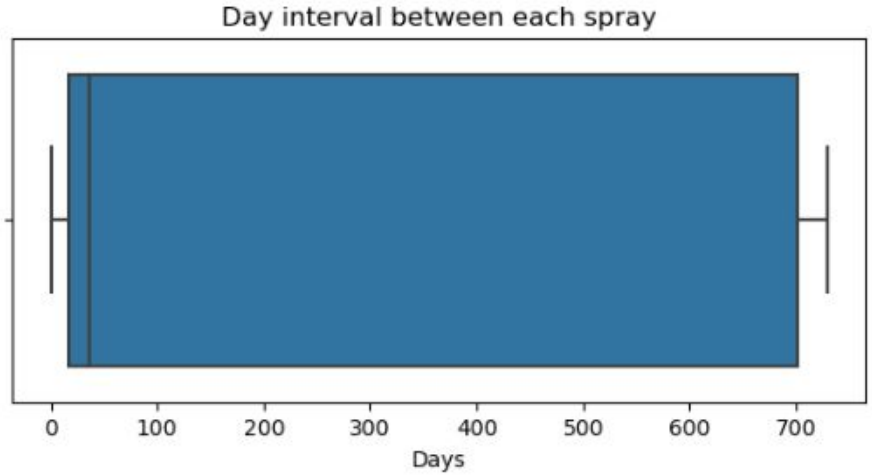
- Predicted cases of WNV infected mosquitoes

Results



- North Area of Chicago is predicted to have high presence of WNV infected mosquitoes
- Ineffective with high concentration of spraying in the upper region
- Possible missing spray spots

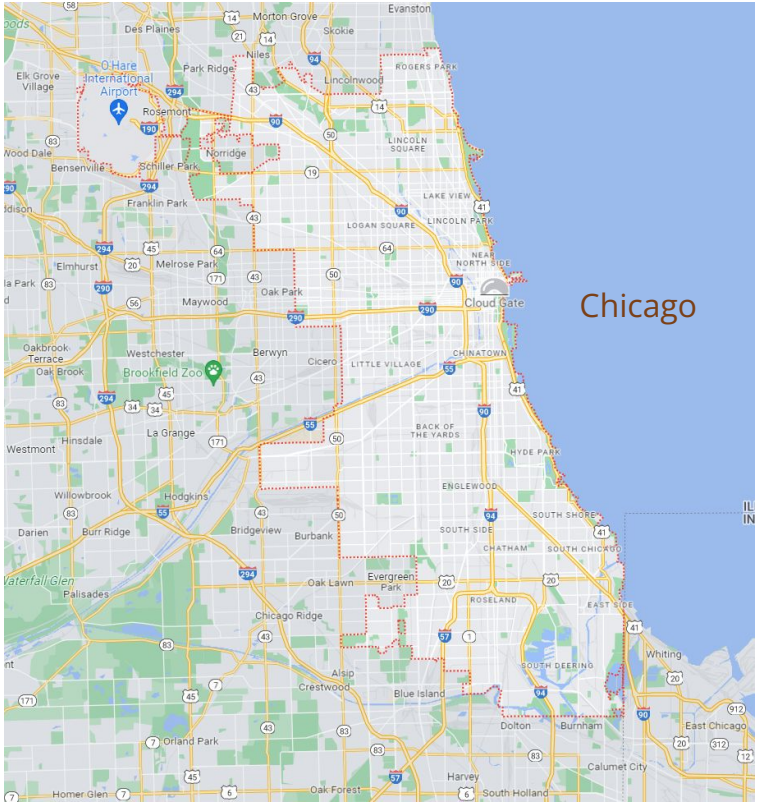
Cost Benefit Analysis



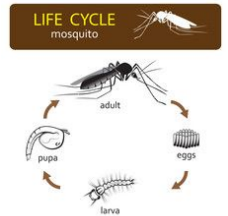
Possible Reasons to why upper region of Chicago have high predicted

1. Ineffective Adulticide for mosquitoes; biodegrades in presences of sunlight or other microorganisms
2. Low Frequency of sprays

Cost of Spraying



- Approximated Land Area is 607.4 km^2. Which equates to 150,000 Acres of land
- Spray used is **Zenivex**, costing US$0.67 per acre. Total cost of one spray = US$0.67 * 150,000 = **US$100,500**
- Half life of Zenivex at 1.5 days, mosquito breeding cycle at 8-10 days. We recommend to spray twice a month
- Annual Cost = **US$2.42 million**



Medical Cost breakdown

Severity of Symptom	Description	Medical Cost (USD)
Moderate	Non-neuroinvasive diseases. Fever or Flu like symptoms	$4,467
Severe	Acute Flaccid Paralysis (AFP) and Speech Impediment	$20,774

Total WNV Cost breakdown

Description	Cost
Total Medical Cost Incurred (225 cases and 22 fatality reported in 2002)	US$1.64 million
Total Productivity Cost (Average patient misses 50 days. Lifetime productivity cost at US$1.2 million	US$28.7 million
Total	US$30.3 million

Conclusion

1) US$30.3 million vs US$2.42 million, benefits of spraying far outweighs the cost of it. Lives saves and long lasting neuroinvasive disease prevented.
2) Our predicted WNV cases could be useful to convince the Chicago state to increase spending on the prevention.
3) Good practices of removing stagnant water

Future works

- Try Deep Learning, could potentially find better predictors
- Expand to other region of US since WNV hit other states too