Project 4:

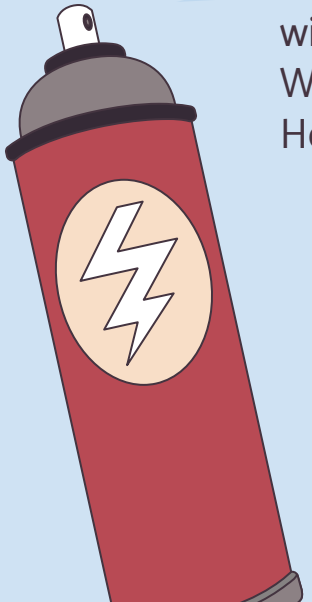
West Nile Virus Prediction

Members: Liubin | Mubina | Wei Hua

Problem Statement

West Nile virus is most commonly spread to humans through infected mosquitos. Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death.

Our client is the Centers for Disease Control and Prevention (**CDC**) who are working with Chicago government to reduce the patients whom are affected by the incurable West Nile Virus. Our team are given a chance to work with Department of Public Health to set up a surveillance control system.



Goals to achieve

As a data scientist from a consultancy firm:

1. our task is to build a model and make predictions to determine the period and location of the sprays.

2. We will also be conducting a cost-benefit analysis which include the annual cost projections for various levels of pesticide and quantity of the pesticide spraying to achieve the maximum benefit.

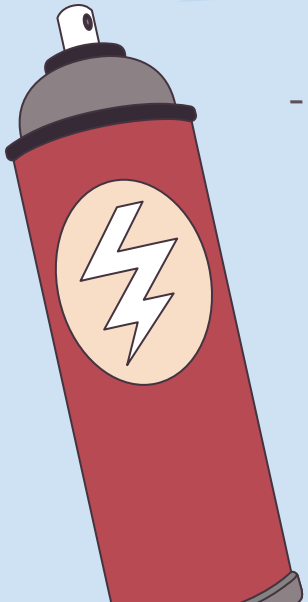
As for the Model:

The model will then be evaluated by ROC AUC score and recall score. The objective of the model is to get a high ROC AUC score and recall score.

Dataset

Main dataset where public health workers in Chicago set up mosquito traps across the city to test for the presence of West Nile virus.

- Spray data which records the details of their spraying such as location and date in order to reduce the number of mosquitoes in the area.
- Weather Data which records the condition of the city. It is believed that hot and dry conditions are more favourable for West Nile virus as compared to cold and wet.
- Map from openstreet map

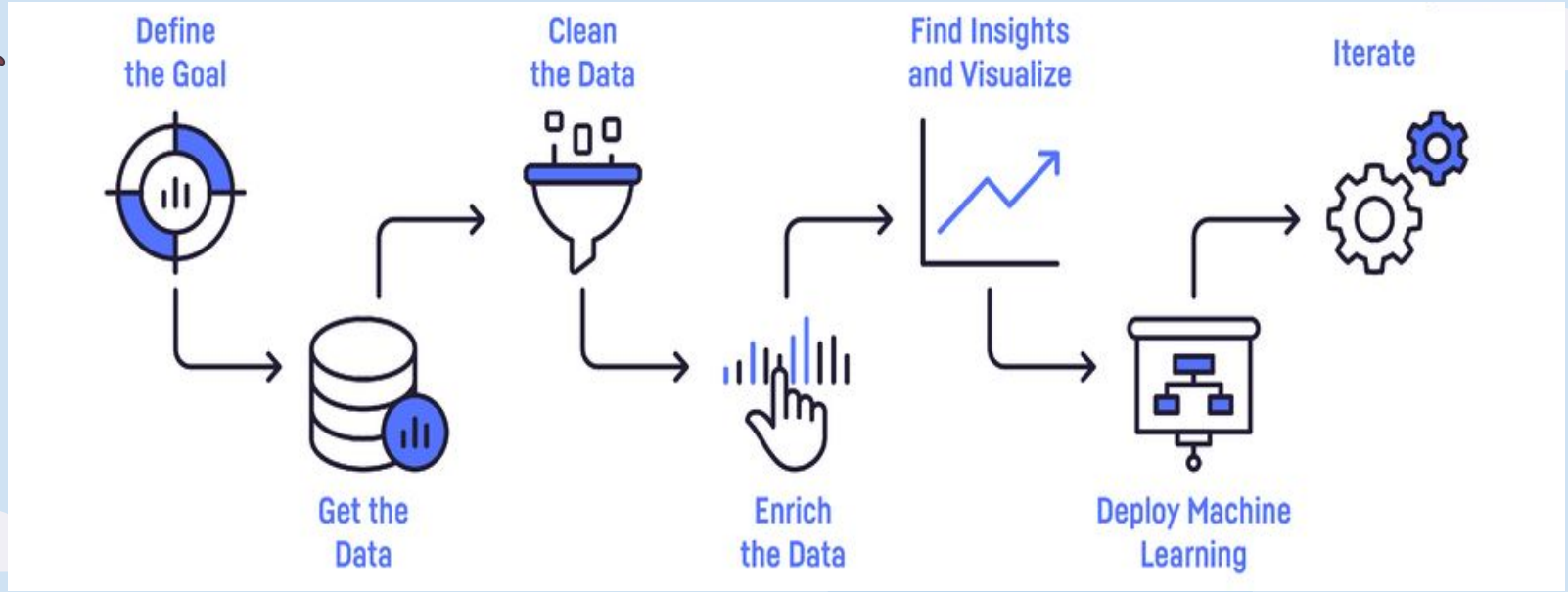


Background

Key facts about West Nile Virus?

1. Can cause fatal neurological disease in humans
2. Most of the affected people did not show any symptoms
3. West Nile virus is incurable for human
4. It is mainly transmitted to people through the bites of infected mosquitoes
5. The virus can cause severe disease and death in horses.
6. Vaccines are available for horses.

Data Science Process

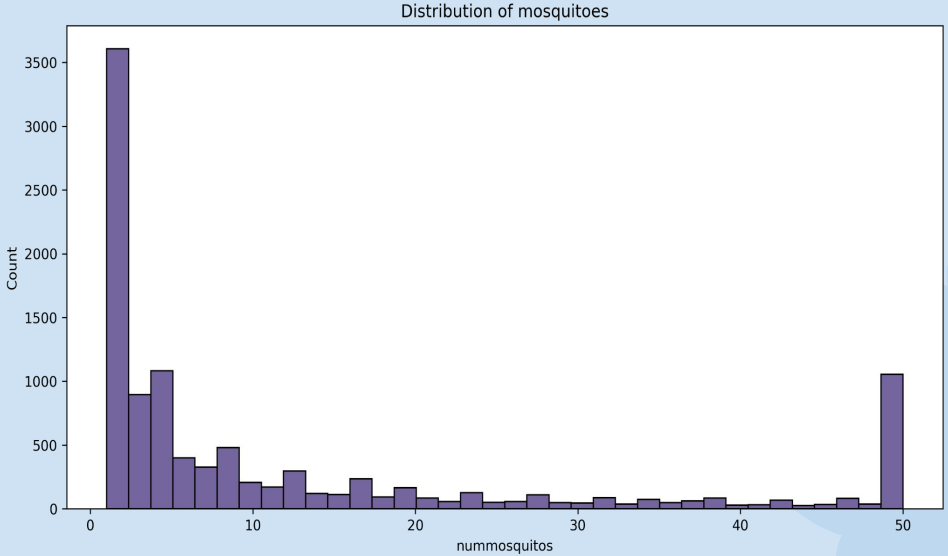


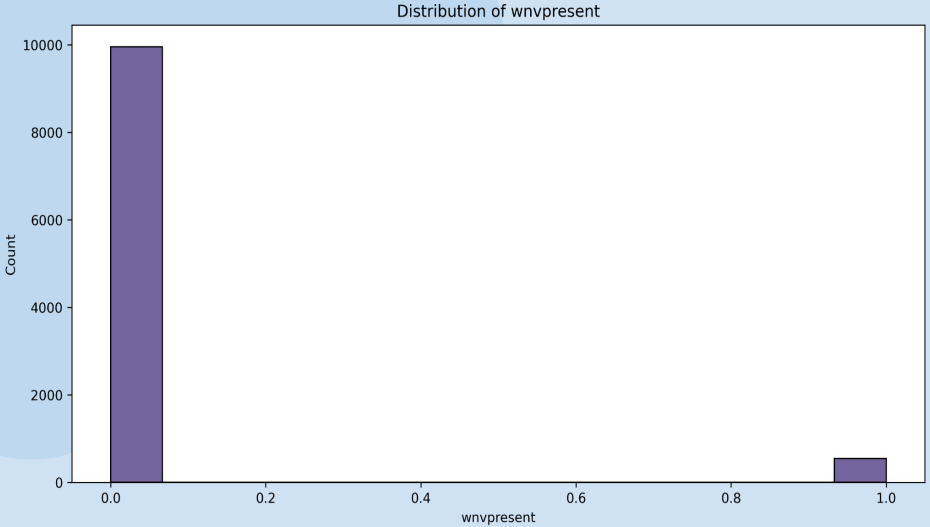
Data Cleaning

Things that have been done in Data cleaning:

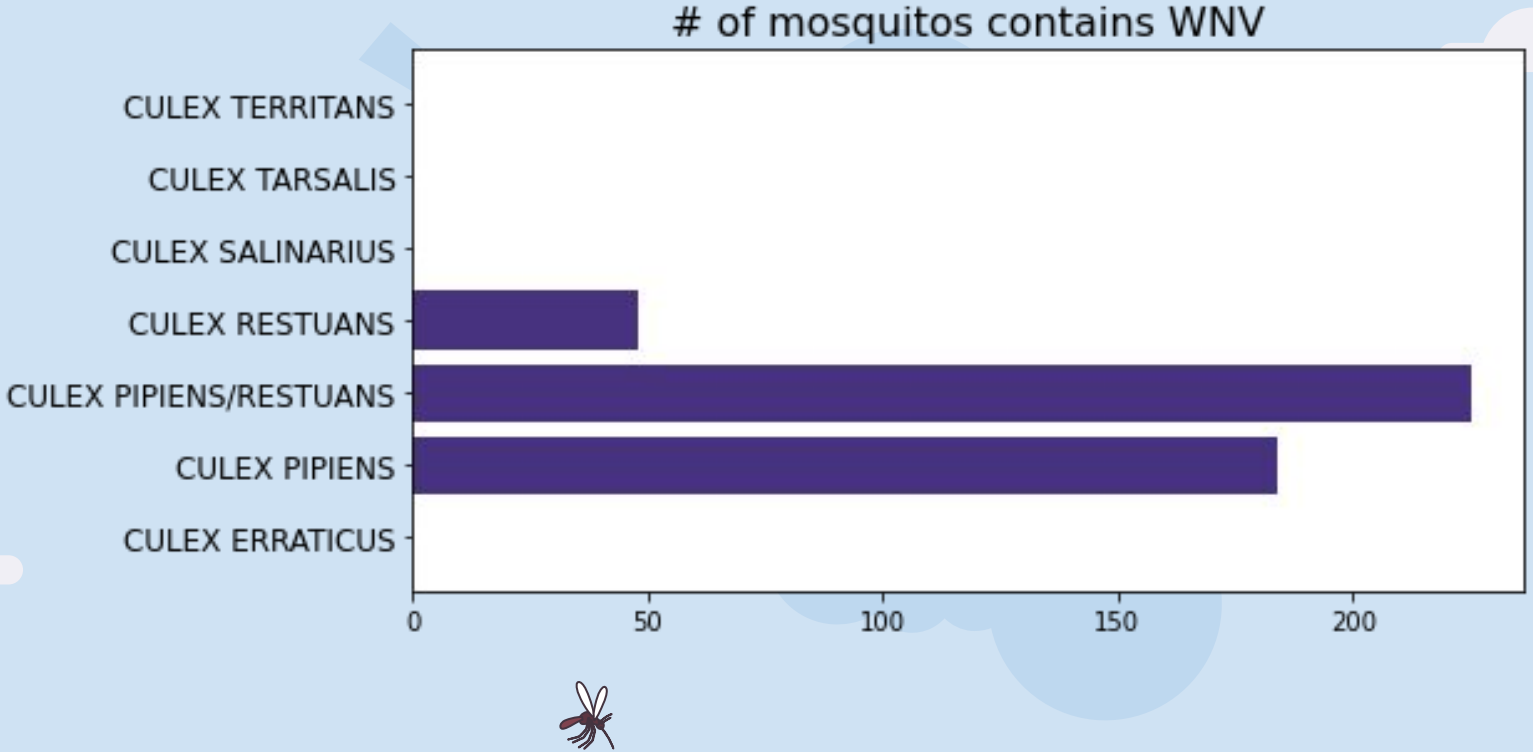
- Convert Date to Datetime type & Split Date into Year, Month, Day
- Missing Values handling, remove duplicates
- Merge Weather Data with Train/Test trap data based on location distance (created new feature-nearest station)
- Reshape the data:
  > Traps with > 50 mosquitos divided into multiple rows
  > Combined these rows

EDA

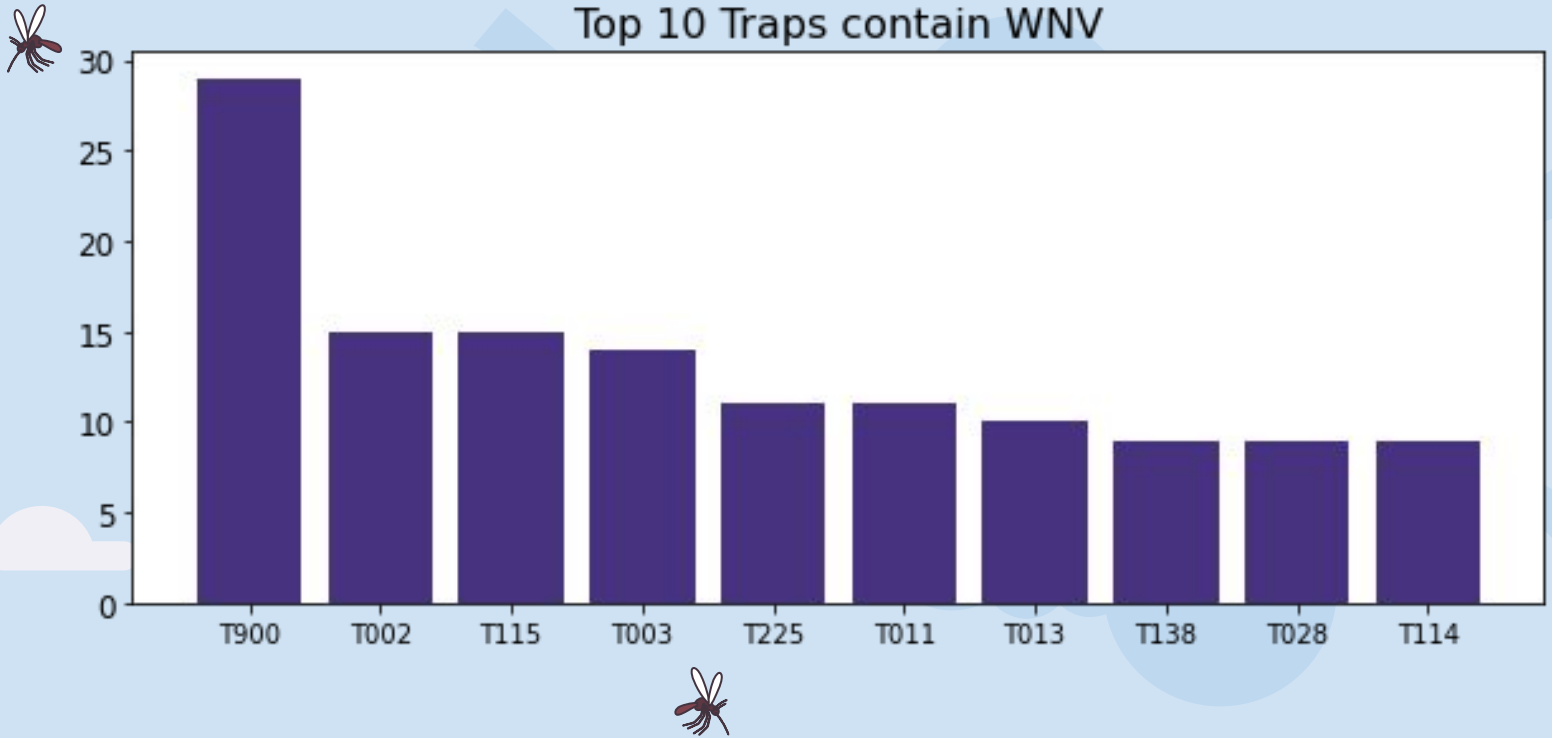

Distribution of mosquitoes
Count
3500
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
nummosquitos


Distribution of wnvpresent
Count
10000
8000
6000
4000
2000
0
0.0
0.2
0.4
0.6
0.8
1.0
wnvpresent

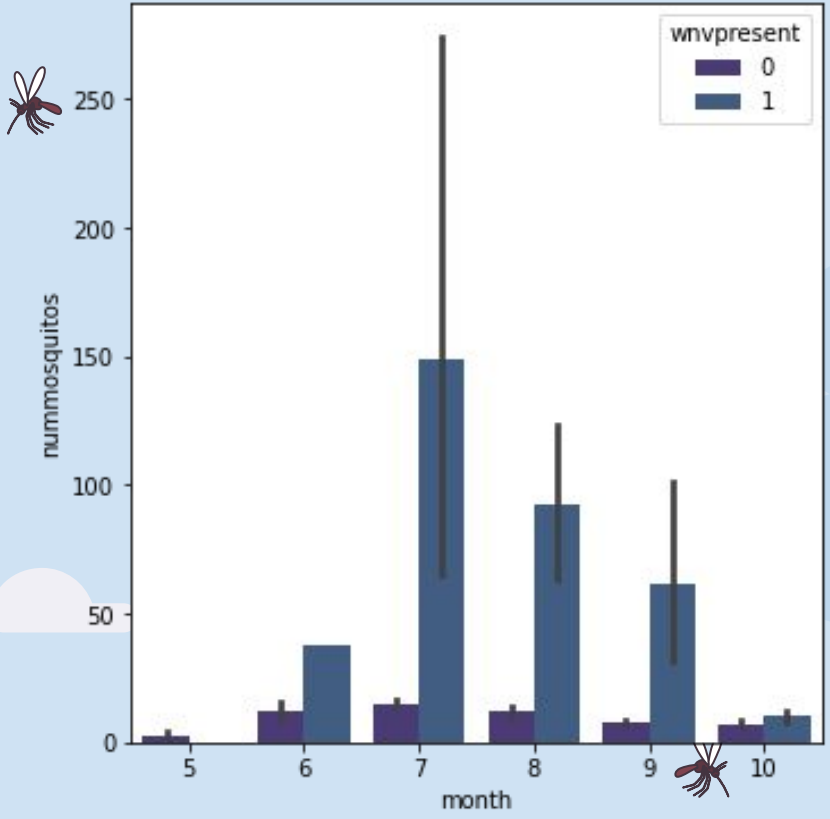
EDA

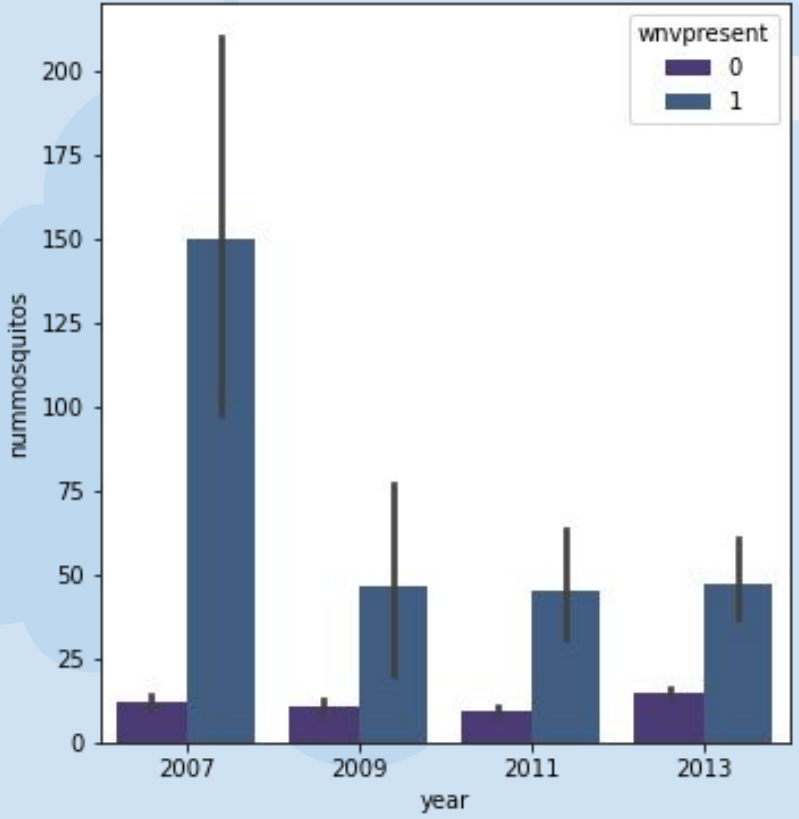

# of mosquitos contains WNV
CULEX TERRITANS
CULEX TARSALIS
CULEX SALINARIUS
CULEX RESTUANS
CULEX PIPIENS/RESTUANS
CULEX PIPIENS
CULEX ERRATICUS
0
50
100
150
200

EDA

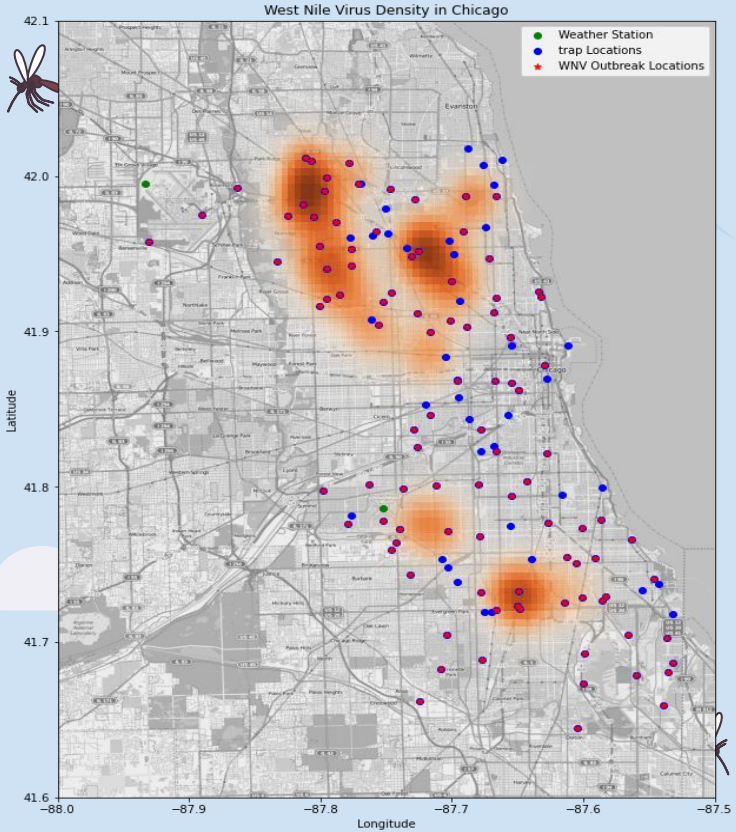

Top 10 Traps contain WNV
30
25
20
15
10
5
0
T900
T002
T115
T003
T225
T011
T013
T138
T028
T114

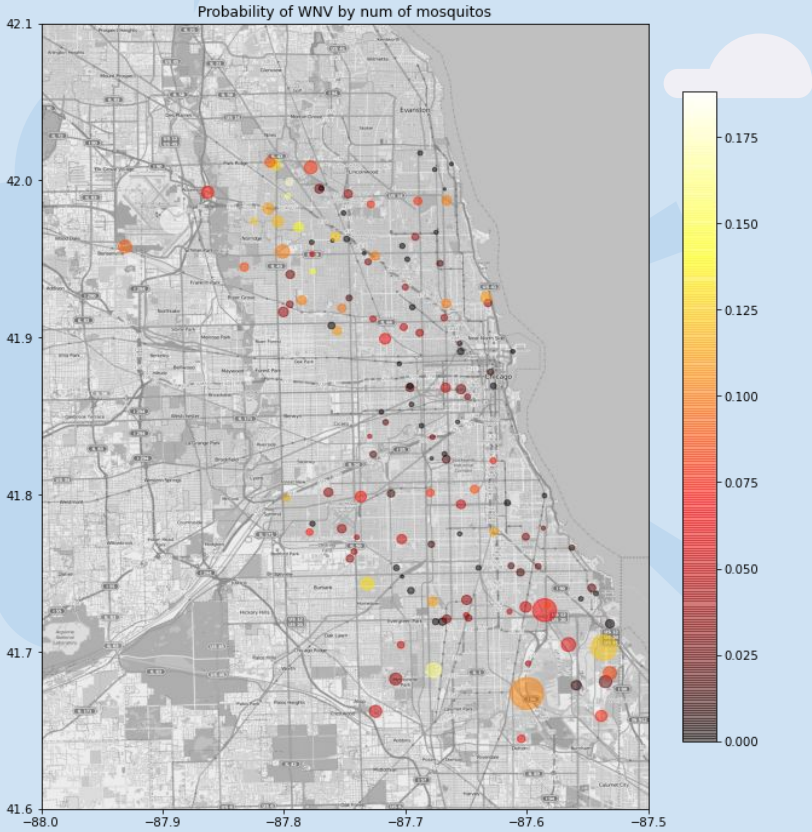
EDA


wnvpresent
0
1
nummosquitos
0
50
100
150
200
250
5
6
7
8
9
10
month

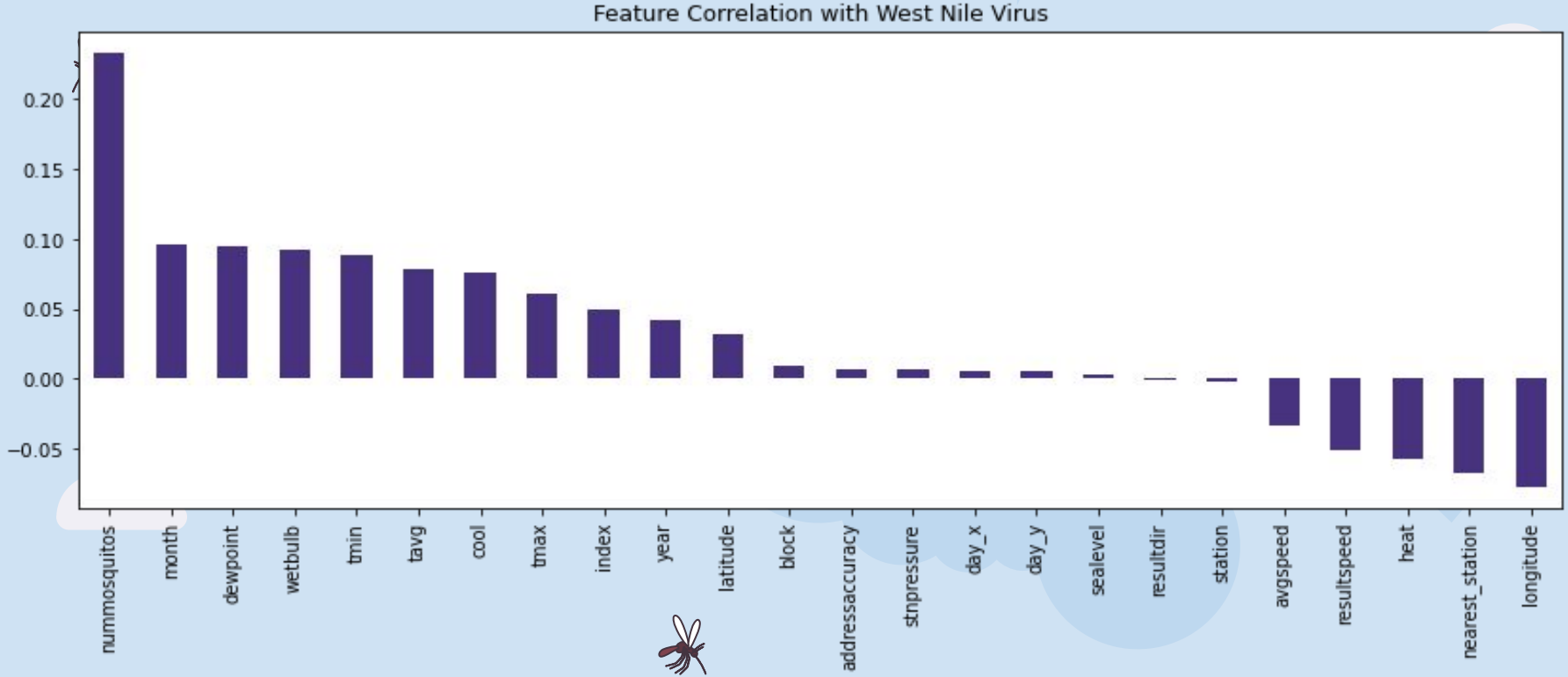

wnvpresent
0
1
nummosquitos
0
25
50
75
100
125
150
175
200
2007
2009
2011
2013
year

EDA


West Nile Virus Density in Chicago
Weather Station
trap Locations
WNV Outbreak Locations
Latitude
Longitude
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


Probability of WNV by num of mosquitos
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000

EDA


Feature Correlation with West Nile Virus
0.20
0.15
0.10
0.05
0.00
−0.05
nummosquitos
month
dewpoint
wetbulb
tmin
tavg
cool
tmax
index
year
latitude
block
addressaccuracy
stnpressure
day_x
day_y
sealevel
resultdir
station
avgspeed
resultspeed
heat
nearest_station
longitude

Modelling - Methodology

3 Datasets to see the effect of weather on our predictions:

1. Trap data
2. Trap data with weather data from the nearest station
3. Trap data with up to 21 day moving average and time lagged weather data

3 classification models using gridsearch with emphasis on ROC AUC:

1. Logistic Regression
2. KNN Classifier
3. Random Forest Classifier

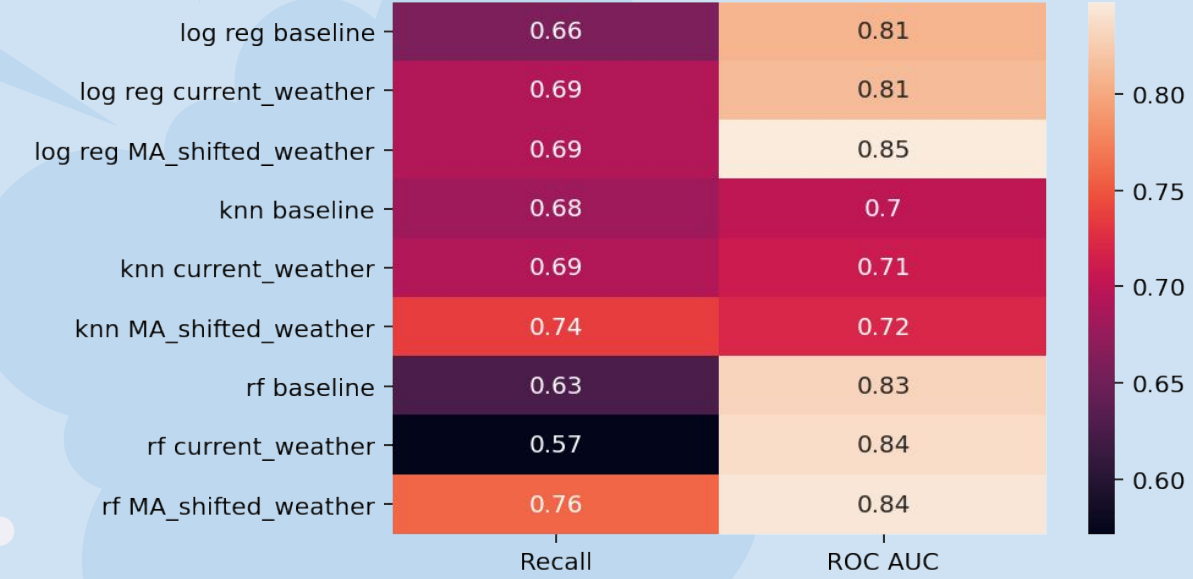
Modelling - Baseline

95% of the data is for areas where WNV is not present

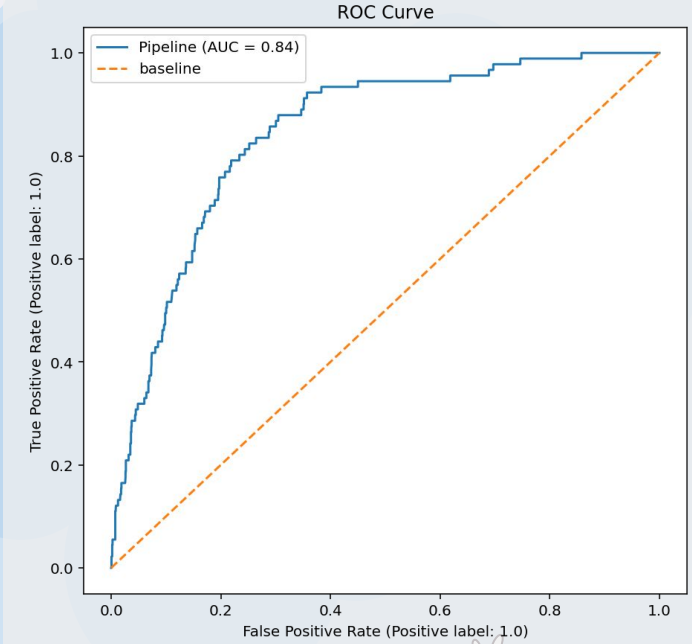
If we predict WNV is not present for 100% of the cases:

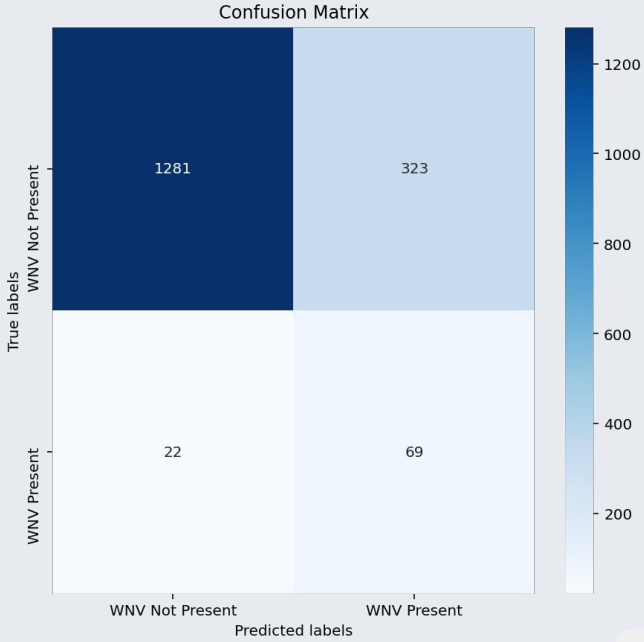
- Accuracy would be 95% however,
- ROC AUC would be 0.5
- Recall would be 0

Modelling - Comparison

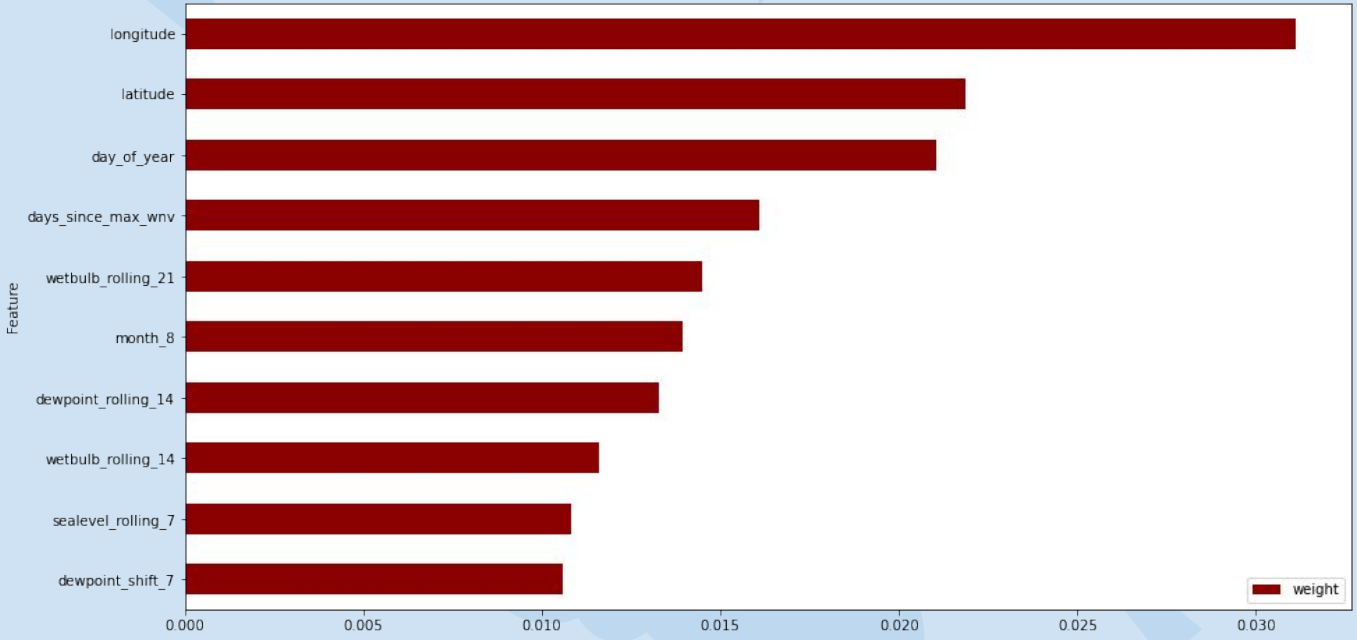


Random Forest Classifier





Top 10 Features - Random Forest



Model Evaluation

Model	Recall score	ROC AUC	TP	FP	TN	FN
Logistic Regression	0.725	0.848	66	274	1330	25
KNN classifier	0.736	0.720	67	622	982	24
Random forest classifier	0.747	0.844	69	323	1281	22

Conclusion

Feature Engineering

Created 3 different train datasets:

1. Dataset with no weather features
2. Datasets with weather features
3. Dataset with moving average and time lagged weather features

Handling imbalance train dataset

SMOTE is used to address imbalance train dataset by oversample the minority class

Choice of metrics

1. ROC AUC score
   - binary classification problems
2. Recall score
   - to focus on false negative

Conclusion

Best Model

Random Forest Classifier

- Least cross validation score difference
- Highest recall score with least False Negative ~ 75%
- High ROC AUC score ~84%

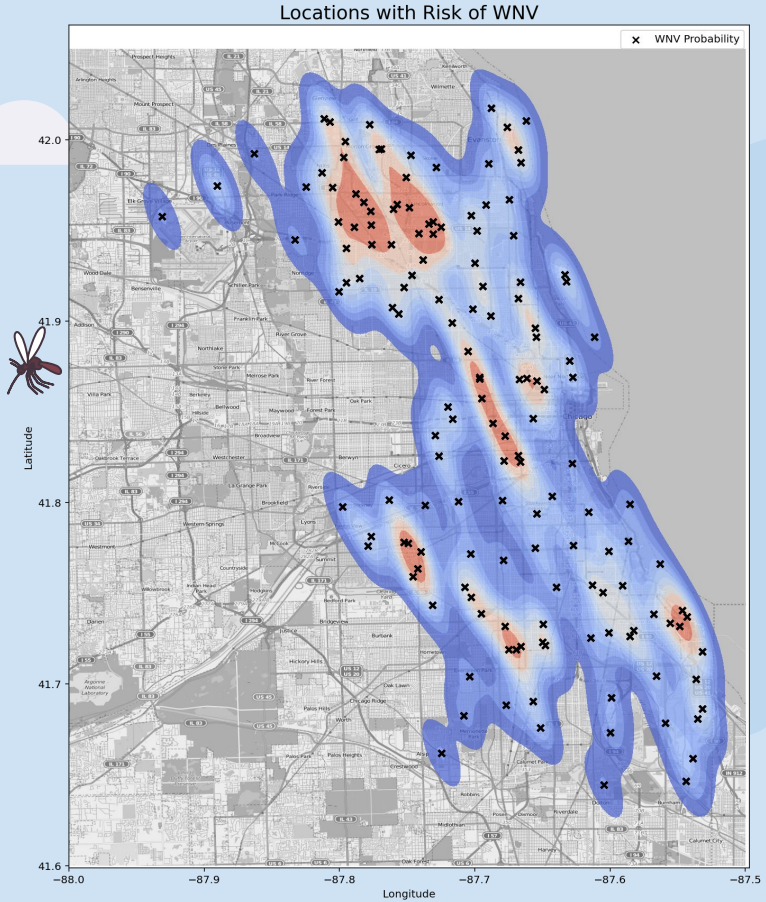
Top Features identified

1. Location - Longitude and latitude
2. Period of the year - days since max wnv, day of year, month
3. Environment conditions - wetbulb temperature, sealevel and dew point

Benefit of the model

1. Primary Stakeholder - To optimise the spray location and period
2. Secondary Stakeholder- To prevent virus outbreak in the community

Cost Benefit Analysis



Economic Cost Breakdown		
Medical Cost	Inpatient Cost	$ 33,143
	Outpatient Cost	$ 1,424
Productivity Cost	Productivity cost per day	$ 191
	No of days recuperating	30
	Productivity cost per day	$ 11,460
Total Cost per person		$ 46,027
Estimated Economic Cost	664 cases as of May 25,2021	$ 30,561,928

Cost of Spray		
	Sacramento County	Chicago
Area (km2)	2,574	606
Sprayed Area	477	606
Sprayed $Cost per Area	$1,662	
Spraying Cost	$792,774	$1,007,172

Recommendations and Limitations

Recommendations:

1. Strategise the location and spray period : North of Chicago and August
2. Educate the public to avoid any exposed stagnant water by putting up more poster for awareness and to introduce mosquitoes control as part of education
3. Promote the use of insect repellent

Limitations:

1. Data is imbalanced
2. Model is trained based on outdated data (2007 to 2013)
3. Model did not take population density into considerations

Thank You