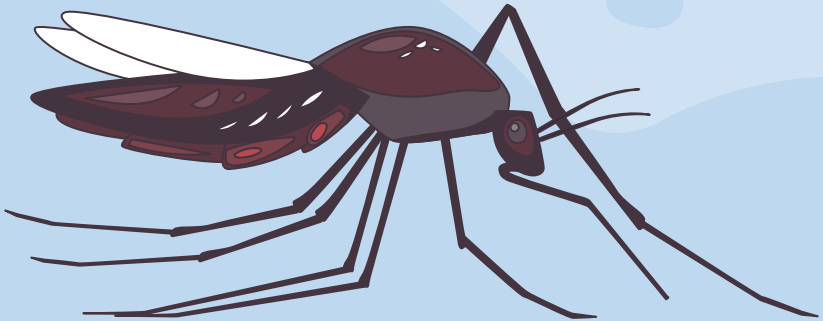
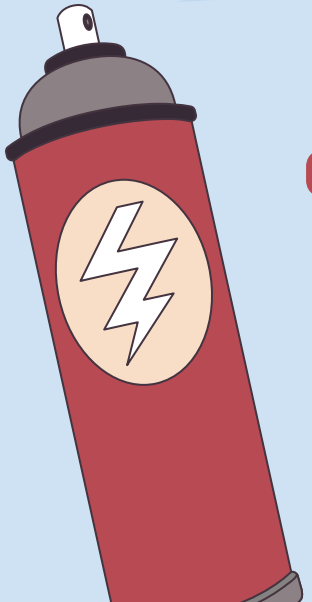


Project FOUR

Group 1: Julian, Terence,
Elang, Henri, Ahmad

Mosquito Map



01 Problem Statement

02 EDA

03 Pre- Processing

04 Feature Engineering

05 Modelling

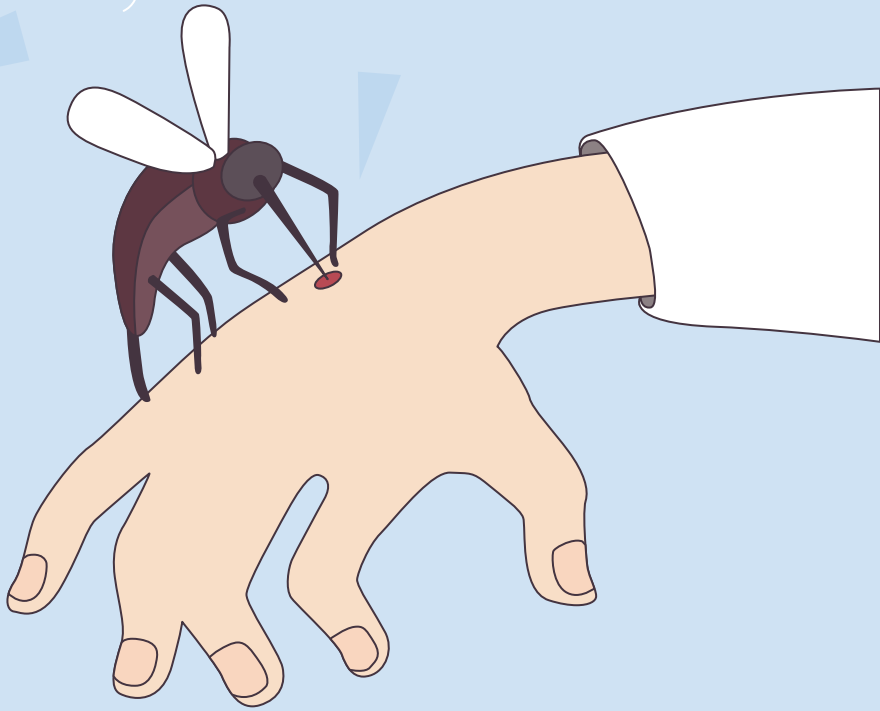
06 Cost Benefit Analysis

01 Problem Statement

1. Given weather, location, testing, and spraying data, **predict** when and where different species of mosquitoes will test positive for West Nile virus.

2. Conduct a **cost benefit analysis**

02
EDA



Exploratory Data Analysis

Trap

Species

Spray

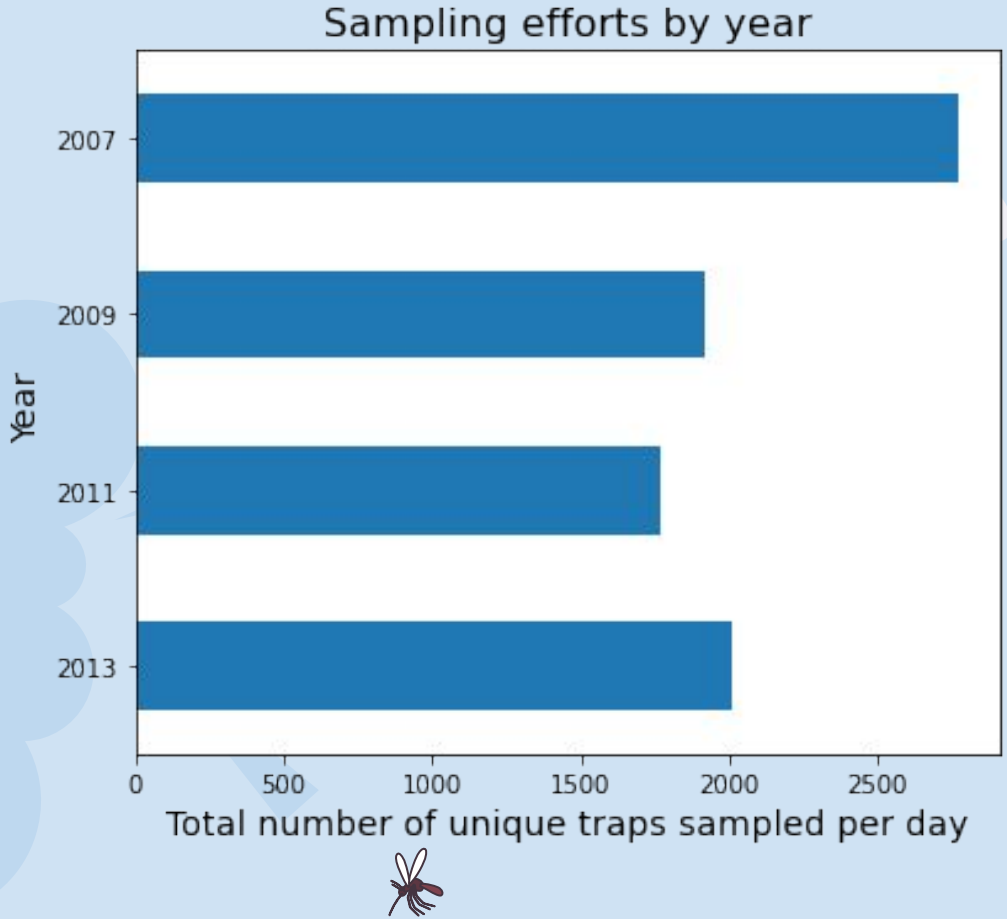
WNV

Weather

Overall

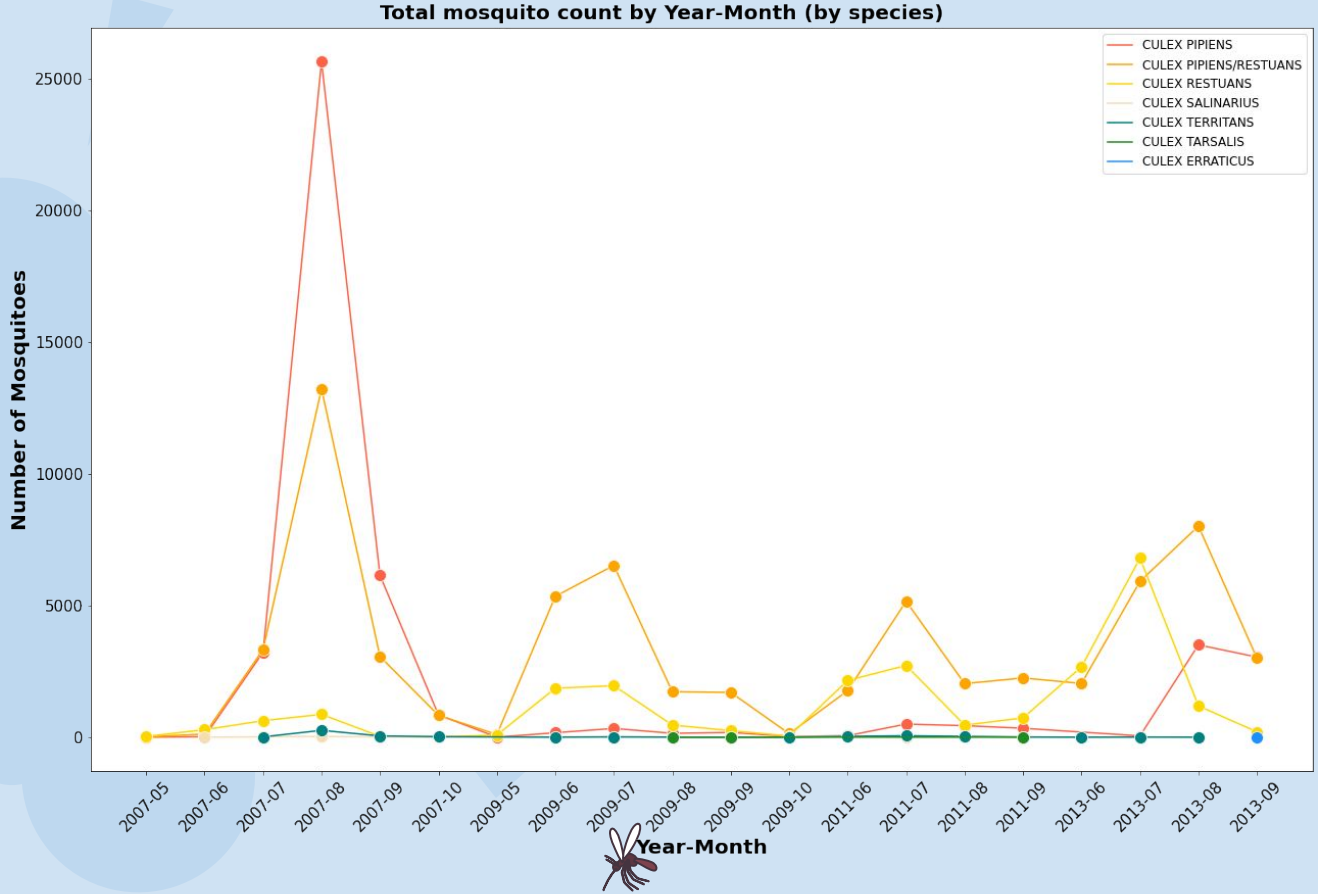
Trap

The number of sampling efforts are inconsistent each year. 2007 has the highest followed by 2013, 2009, 2011



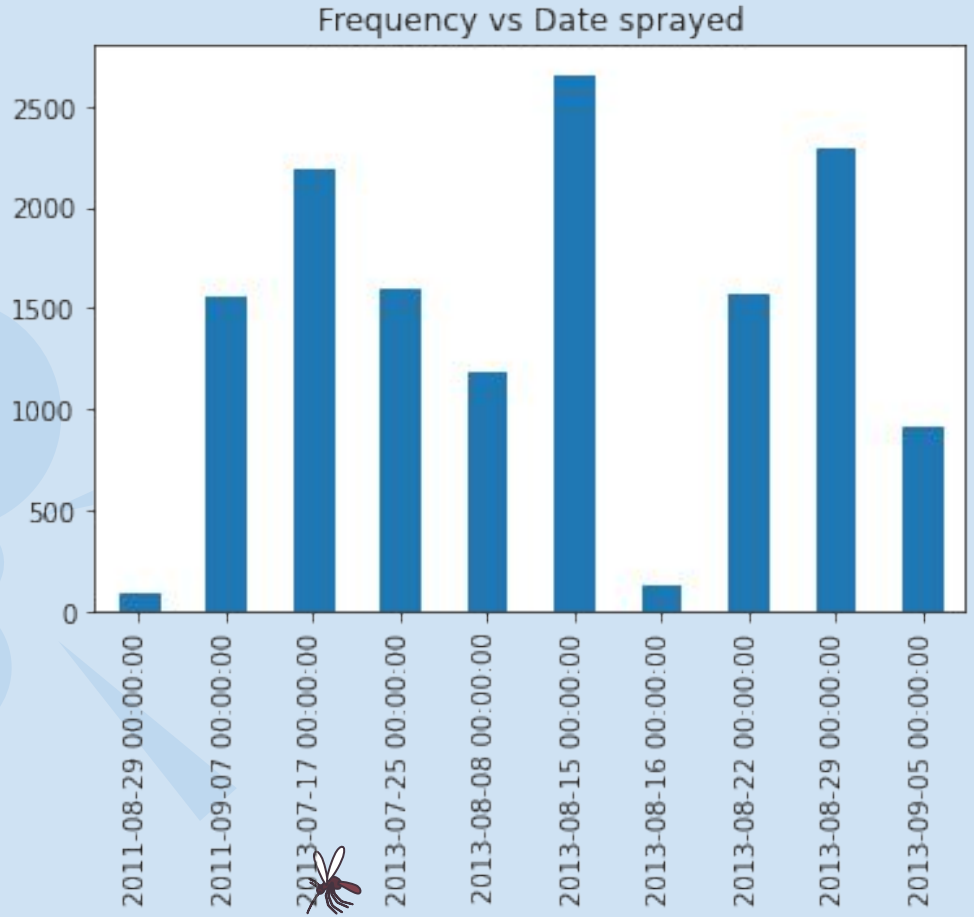
Species

Pipiens and Restuans mostly found species



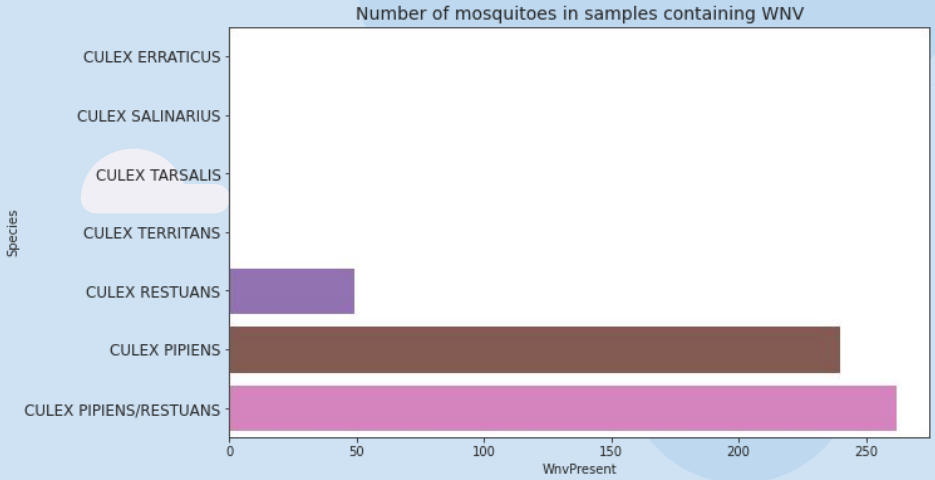
Spray

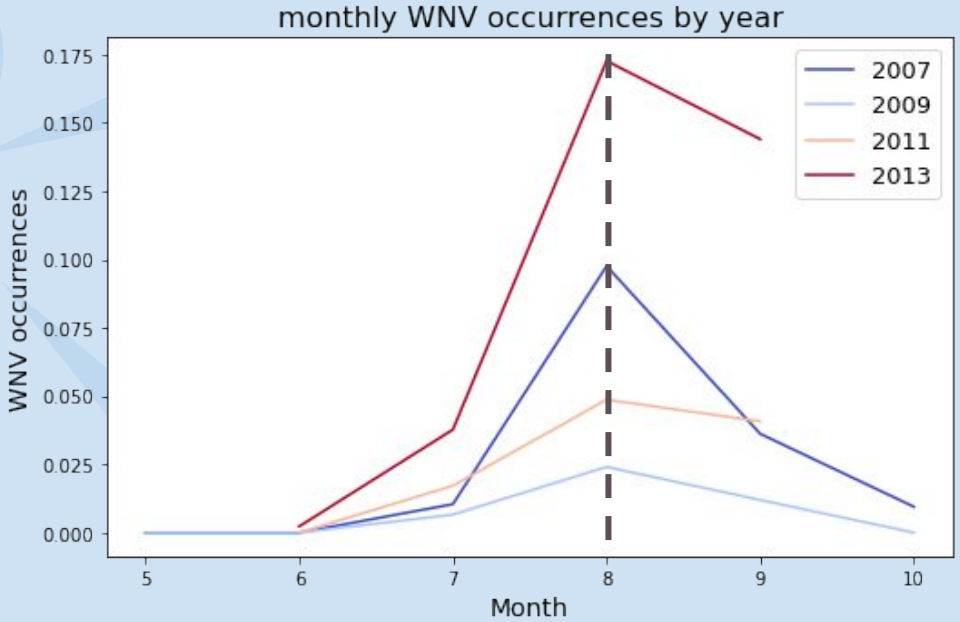
We will go through this later



WNV

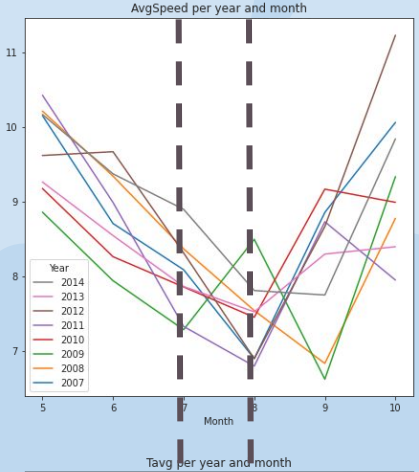
WNV occurrence has a significant increase in certain species and certain month of the year

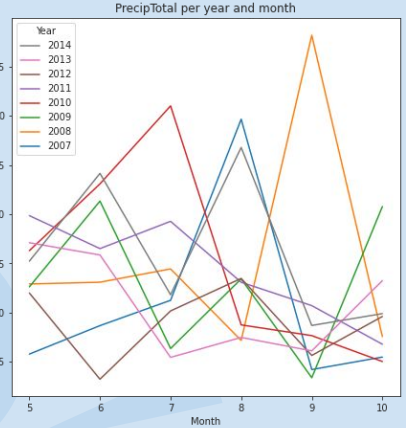


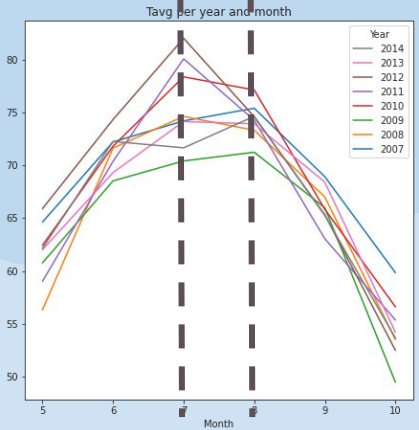


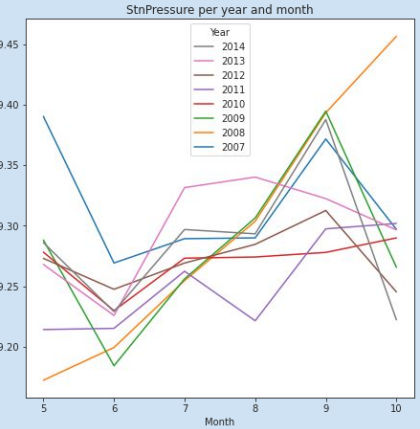
Weather

During 7 - 9th month particularly the Wind Speed is lowest and the Temperature is warmest

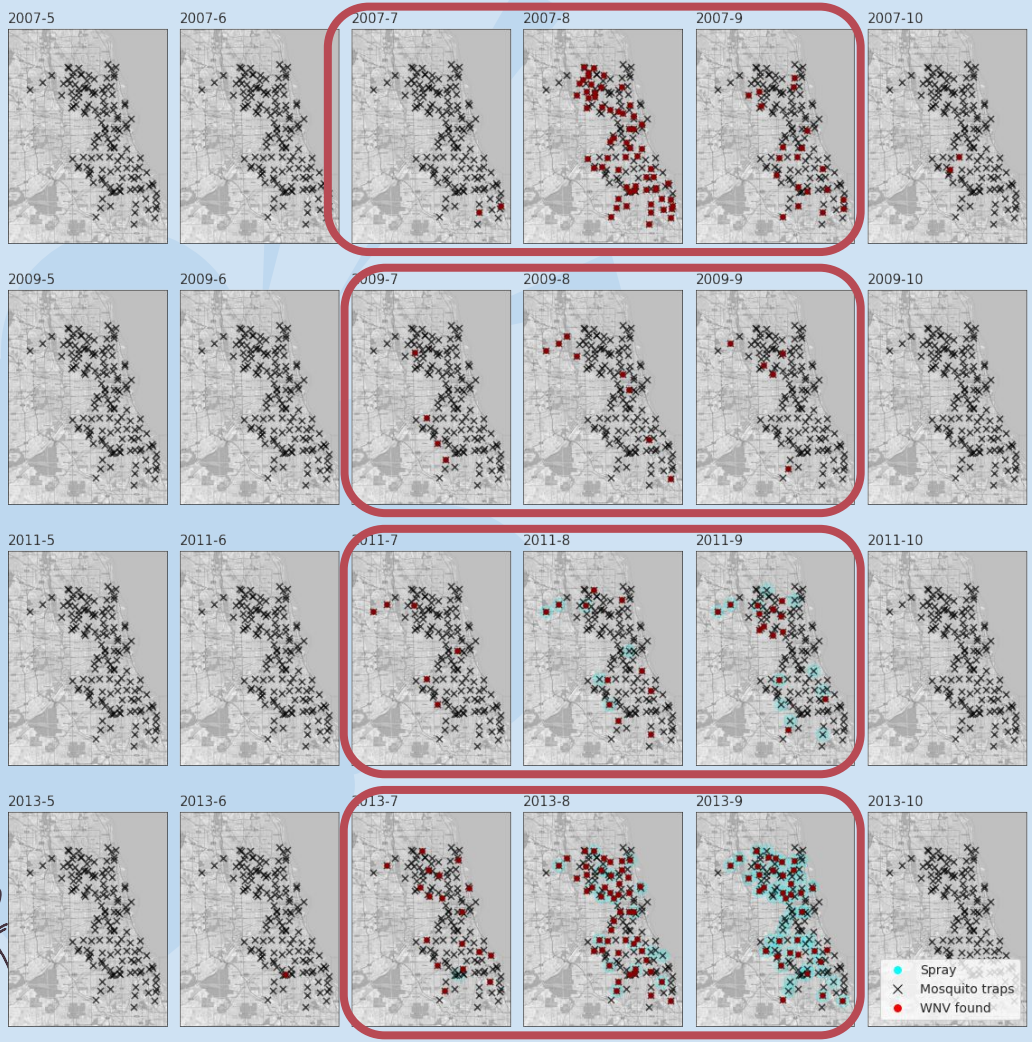




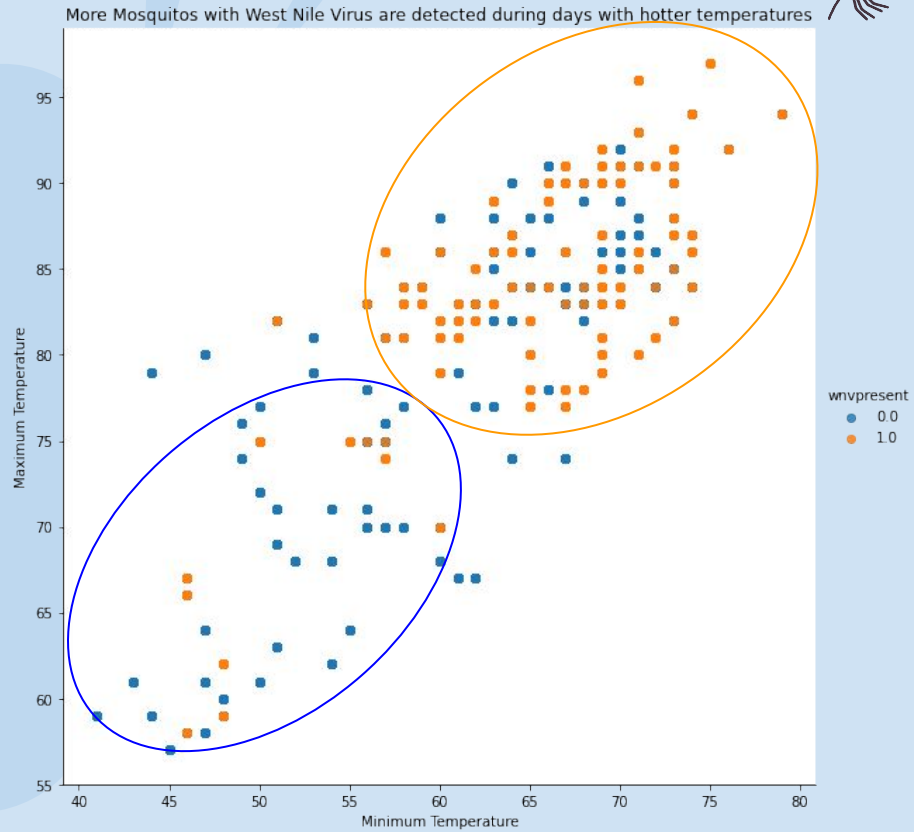




Overall

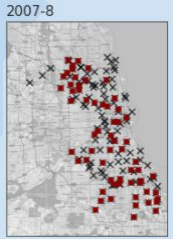

2007-5
2007-6
2007-7
2007-8
2007-9
2007-10
2009-5
2009-6
2009-7
2009-8
2009-9
2009-10
2011-5
2011-6
2011-7
2011-8
2011-9
2011-10
2013-5
2013-6
2013-7
2013-8
2013-9
2013-10
Spray
Mosquito traps
WNV found
Spray
Mosquito traps
WNV found

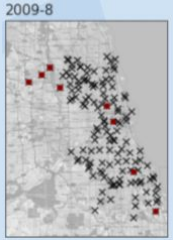
Overall

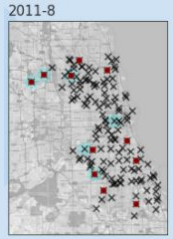

More Mosquitos with West Nile Virus are detected during days with hotter temperatures
Maximum Temperature
Minimum Temperature
wnvpresent
0.0
1.0
95
90
85
80
75
70
65
60
55
40
45
50
55
60
65
70
75
80

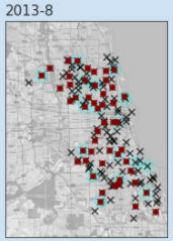
Overall

	Aug 2007	Aug 2009	Aug 2011	Aug 2013
Average Temp	73.46° F	70.15° F	73.58° F	74.28° F
Min Temp	68.87° F	64.68° F	65.9° F	67.9° F









03

Pre-Processing

- Duplicates
- Drop columns
- Split dataset
- SMOTE

Pre-Processing

Duplicates

- Trap #35 appeared **2 times**
- Trap #9 appeared **2 times**
- 4100 N OAK PARK AVE, Chicago, IL appeared **150 times**

Split Dataset

- **Method:** sklearn.model_selection.train_test_split (80:20)

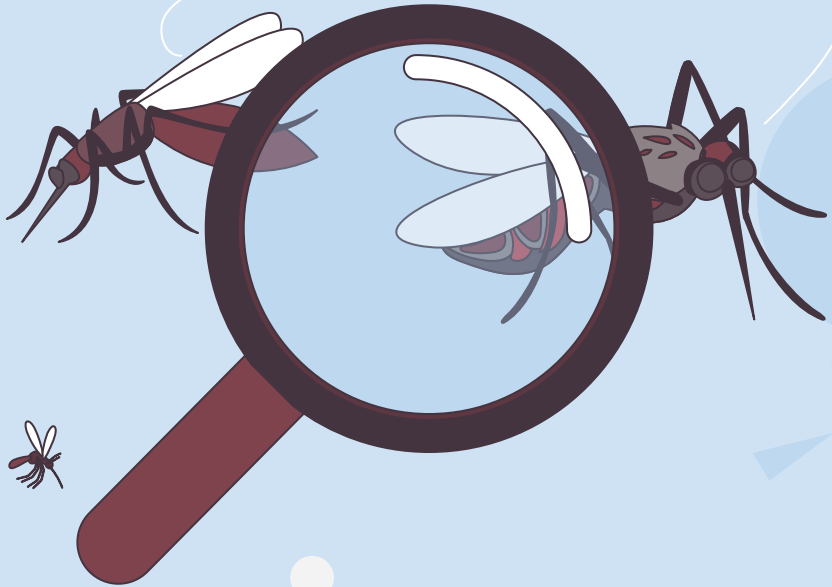
Drop Columns

- 'AddressAccuracy'
- 'Address','Street'
- 'AddressNumberAndStreet'
- 'Water1'
- 'Depart'
- 'Depth'
- 'SnowFall'
- 'AvgSpeed'
- 'ResultSpeed'
- 'ResultDir'
- 'SeaLevel'
- 'StnPressure'
- 'CodeSum'
- 'Heat'
- 'Cool'

Smote

- **Before SMOTE :** [7963 441]
- **After SMOTE :** [6476 6533]
- **Method :** imblearn.combine.SMOTEENN

Data Imbalance



04 Feature Engineering

- Dates
- Sprayed
- Weather

Dates

Separate Month, Year, WeekofYear

From the EDA it was determined that the mosquito population is dependent on the seasonal environments, therefore it is more relevant to process using year, month and week of year.

```
# Create feature Year
# Create feature Month
# Create feature WeekofYear (isocalendar().week)
```

Sprayed

Effectiveness of spray is assumed to be 1000 m radius from the point of spray.

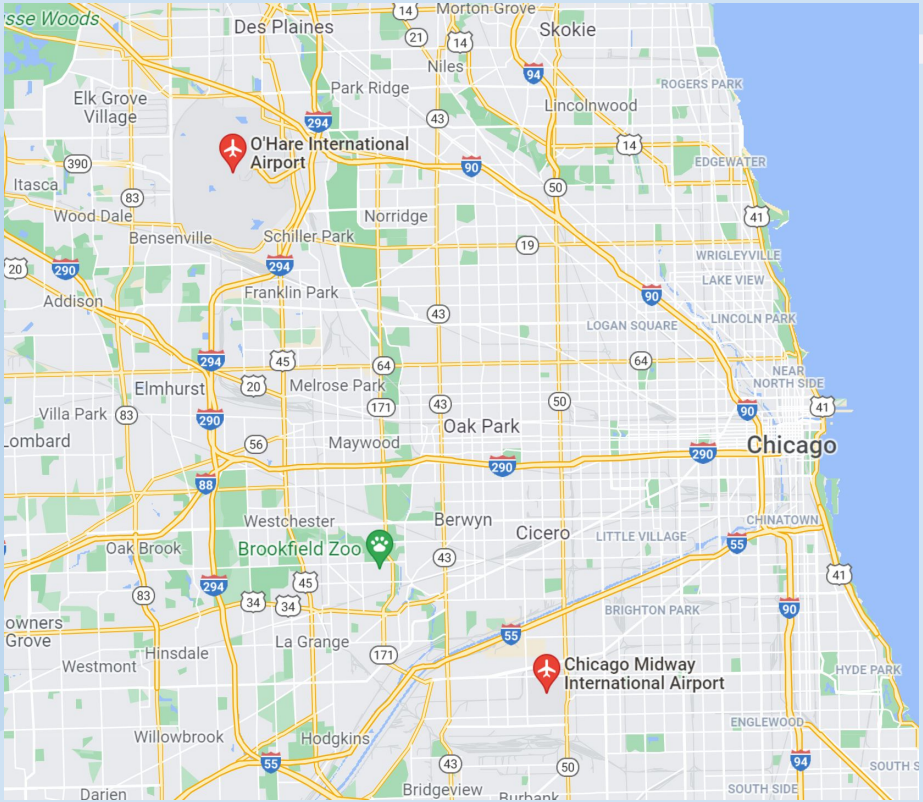
The location of spray determines the affected area, therefore distance features are used to determine which traps were sprayed.

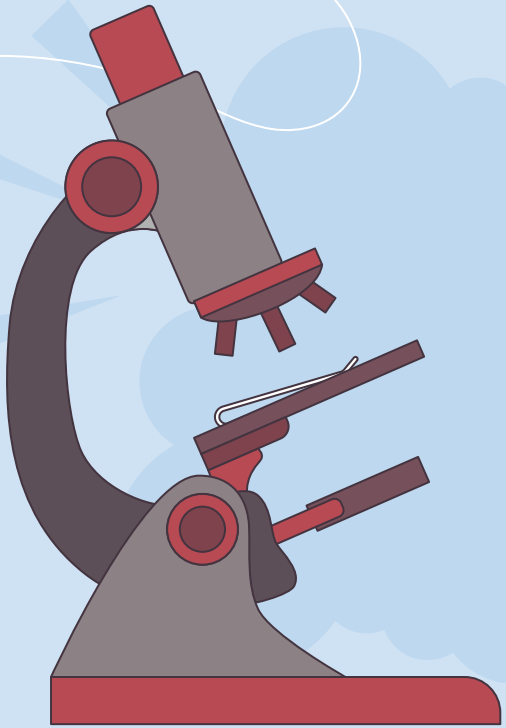
```
# Determine trap coordinate
# Determine spray coordinate
# Determine all trap locations within 1000 meters from the spray coordinate using shapely.nearest_points function, mark this trap as sprayed.
# Otherwise mark the trap as not sprayed.
```

Weather

Merged weather feature
from station 1 and station 2.

The weather records show strong correlation between data from station 1 and station 2. Since station 2 data contains less missing values, it is selected as the weather feature.
Missing data is imputed from the station 1.





05

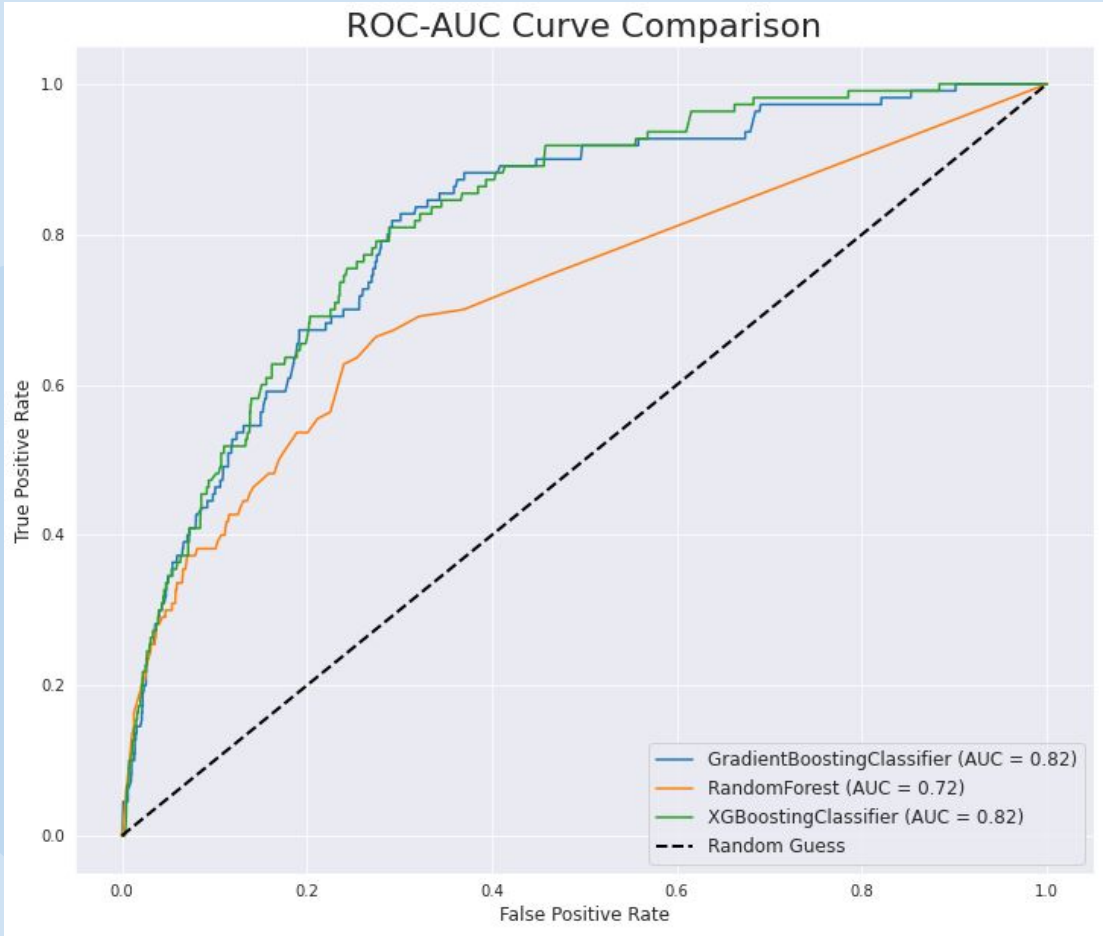
Modelling

- Train -Test Split
- SMOTE Train Only
- StandardScaler
- Grid Search (Appendix)
- Evaluation Metric:
  - ROC
  - F1 , Recall

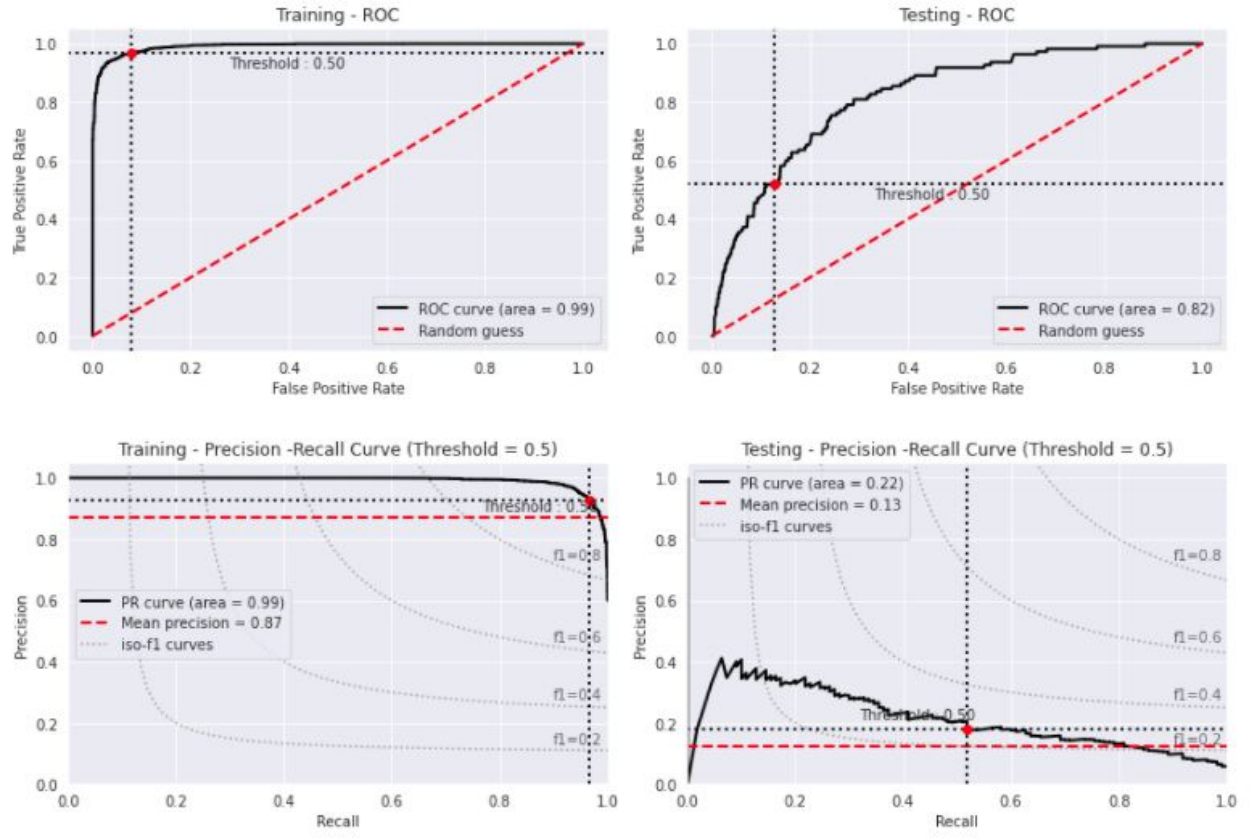
Model Summary

	model	train_auc	test_auc	precision	specificity	recall	f1_score
0	GB (No Smote)	0.89	0.83	0.00	0.99	0.00	0.00
1	GB(Smote)	0.99	0.81	0.17	0.86	0.53	0.26
2	RFC (No Smote)	0.98	0.75	0.33	0.97	0.21	0.26
3	RFC (Smote)	1.00	0.75	0.22	0.92	0.39	0.28
4	XGB (No Smote)	0.88	0.83	0.00	1.00	0.00	0.00
5	**XGB (Smote)**	**0.99**	**0.82**	**0.18**	**0.85**	**0.57**	**0.27**

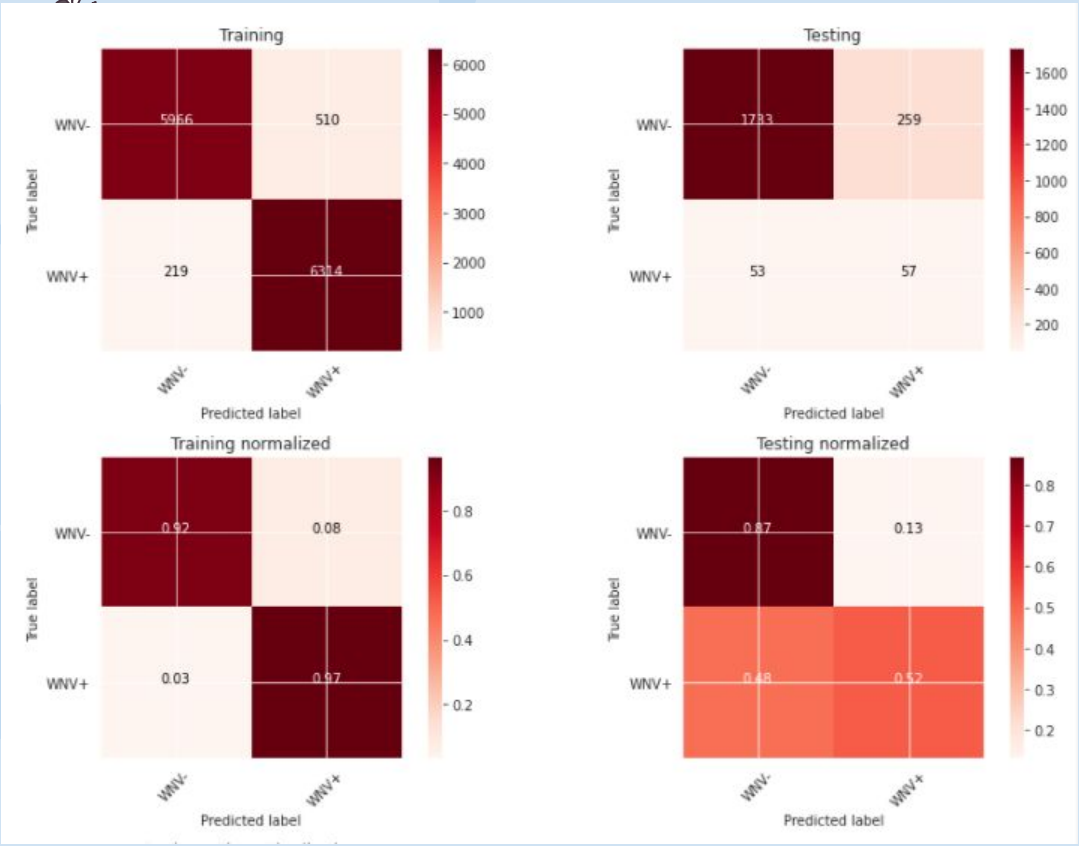
ROC (Smote Model)



XGB Classifier Model



XGB Classifier Model



Kaggle Score

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
df_submission (9).csv	a minute ago	1 seconds	1 seconds	0.62682

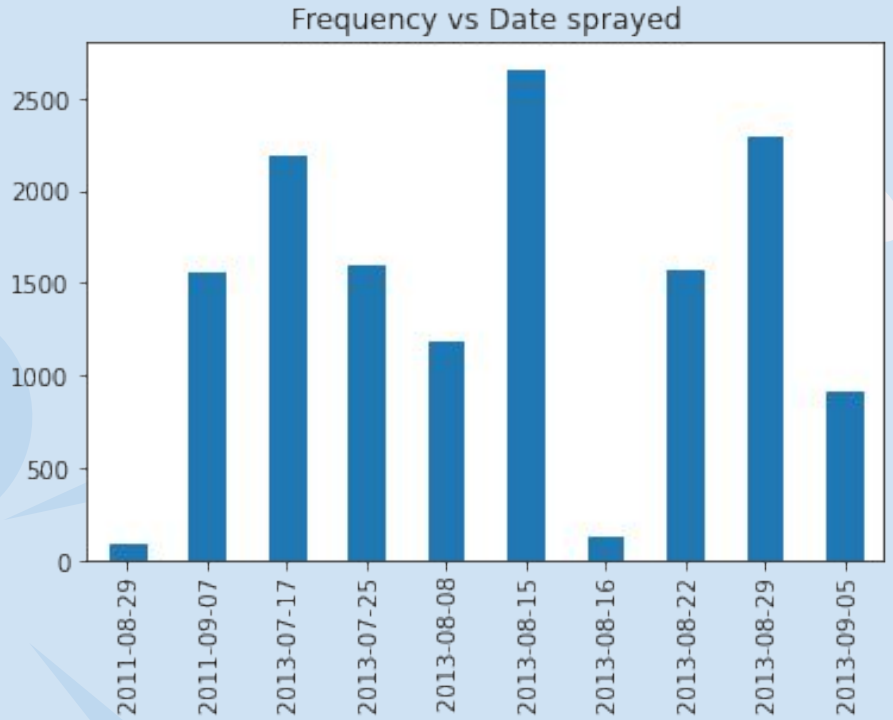
Complete

Jump to your position on the leaderboard

06

Cost Benefit Analysis

Spray Occurrence



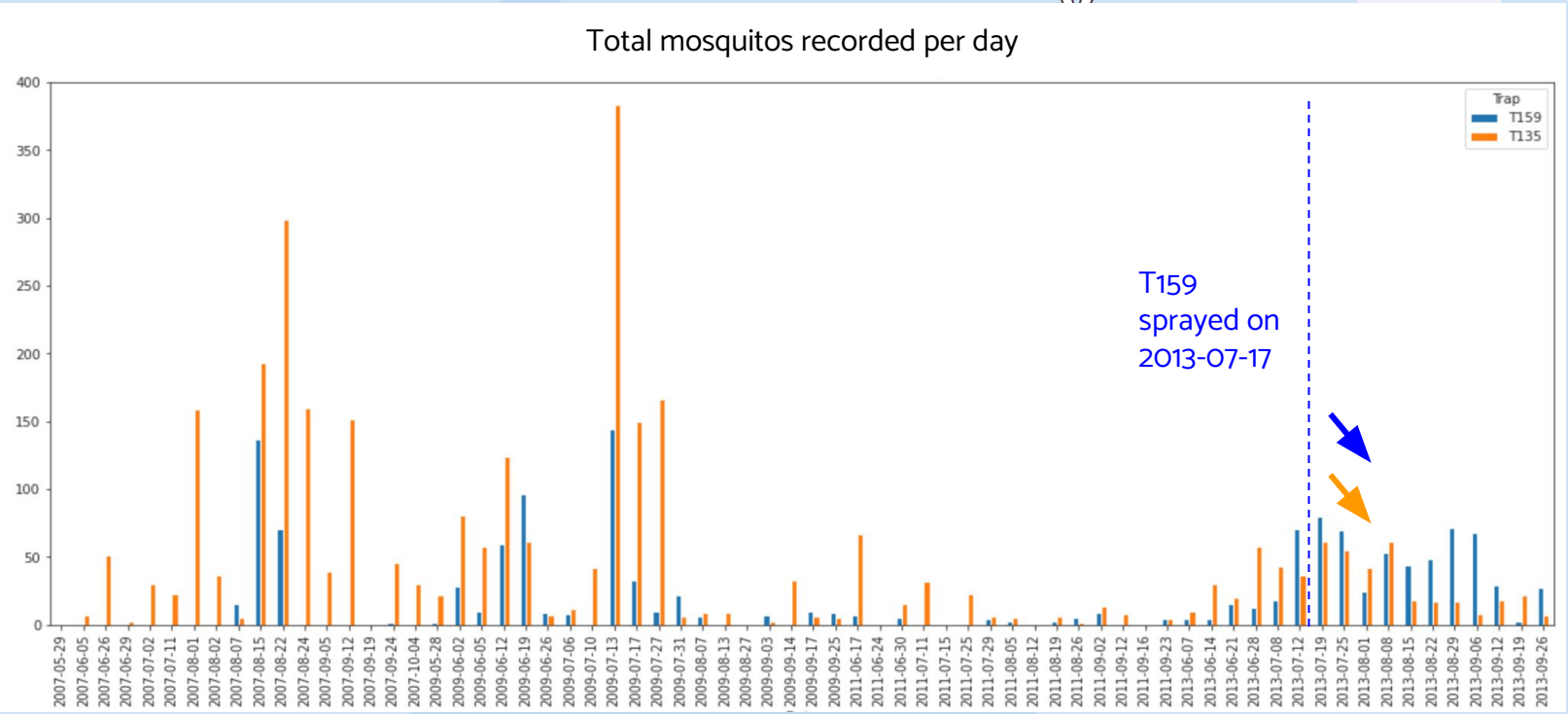
Spray

Only 2 years of spraying with majority in 2013

Spraying is only useful for killing adult mosquitoes. The effects will be short lived because there will still be an uptick of germinated mosquito larvae

- Solutions must be persistent with using adult pesticide

Spray Effectiveness



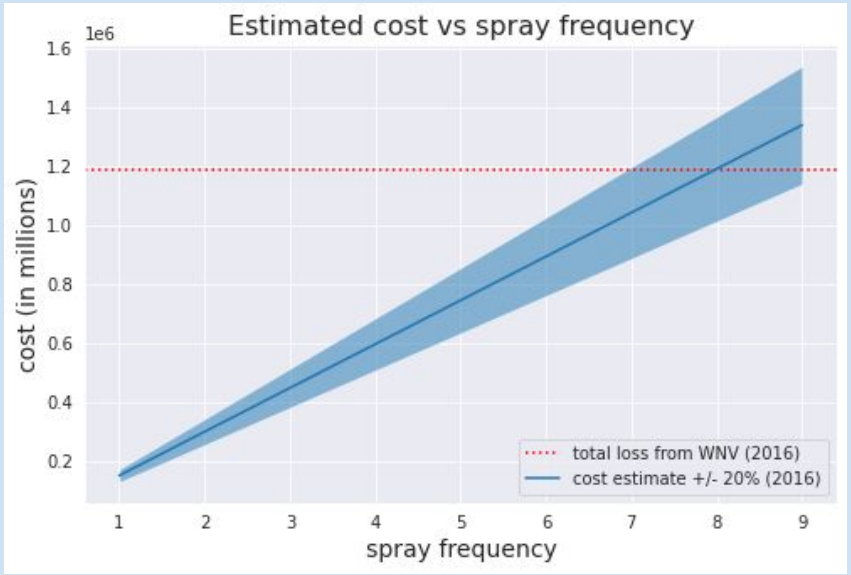
Comparing two specified traps

T159, **sprayed** on 2013-07-17 has **decreasing** trend

T135, **not sprayed** is **decreasing** trend

2013-08: strong resurgence, so even if the pesticide is effective, it is not persistent

Cost Benefit Analysis



Breakeven at 7-8 sprays

- **Costs (using pesticide Zenivex E4[1]):**
  - based on estimates of $500 for a session for 0.5 acre of land[2]
  - Cost is **$149,000 for 0.6 km^2 (Size of Chicago 600km^2)** [3]
- **Benefits:**

○Fewer people dying/falling ill –> increased workplace productivity and healthcare savings (average $11,000)[4]

○108 WNV cases in 2016:[5]

medical bill ~ **$1,190,000**

- Since the benefits outweigh the current prevention cost, therefore the county should socialise the cost

Resources

- Pesticide info
  https://www.cmmcp.org/pesticide-information/pages/zenivex-e4-etofenprox
- Cost of spray
  https://www.callnorthwest.com/2020/05/how-much-does-a-mosquito-treatment-cost/
- Chicago spray strategy
  https://www.chicago.gov/city/en/depts/cdph/provdrs/healthy_communities/news/2020/august/city-to-spray-insecticide-thursday-to-kill-mosquitoes.html
- WNV cost
  https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-019-4596-9
- No. of WNV 2016 Chicago
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7241786/

Appendix : GridSearch

4.4 Modelling With SMOTE and GridSearch

In [46]:
```
# gb_param_grid = {
#     "gb__loss":["deviance"],
#     "gb__learning_rate": [0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2],
#     "gb__min_samples_split": np.linspace(0.1, 0.5, 12),
#     "gb__min_samples_leaf": np.linspace(0.1, 0.5, 12),
#     "gb__max_depth":[3,5,8],
#     "gb__max_features":["log2","sqrt"],
#     "gb__criterion": ["friedman_mse",  "mae"],
#     "gb__subsample":[0.5, 0.618, 0.8, 0.85, 0.9, 0.95, 1.0]
#     }

# gb_smote = run_model('gb' , mod_params = gb_param_grid,  grid_search = True)
# prediction_list,proba_list= predictions(gb_smote)
# evaluation_plot(gb_smote, prediction_list,proba_list)
```

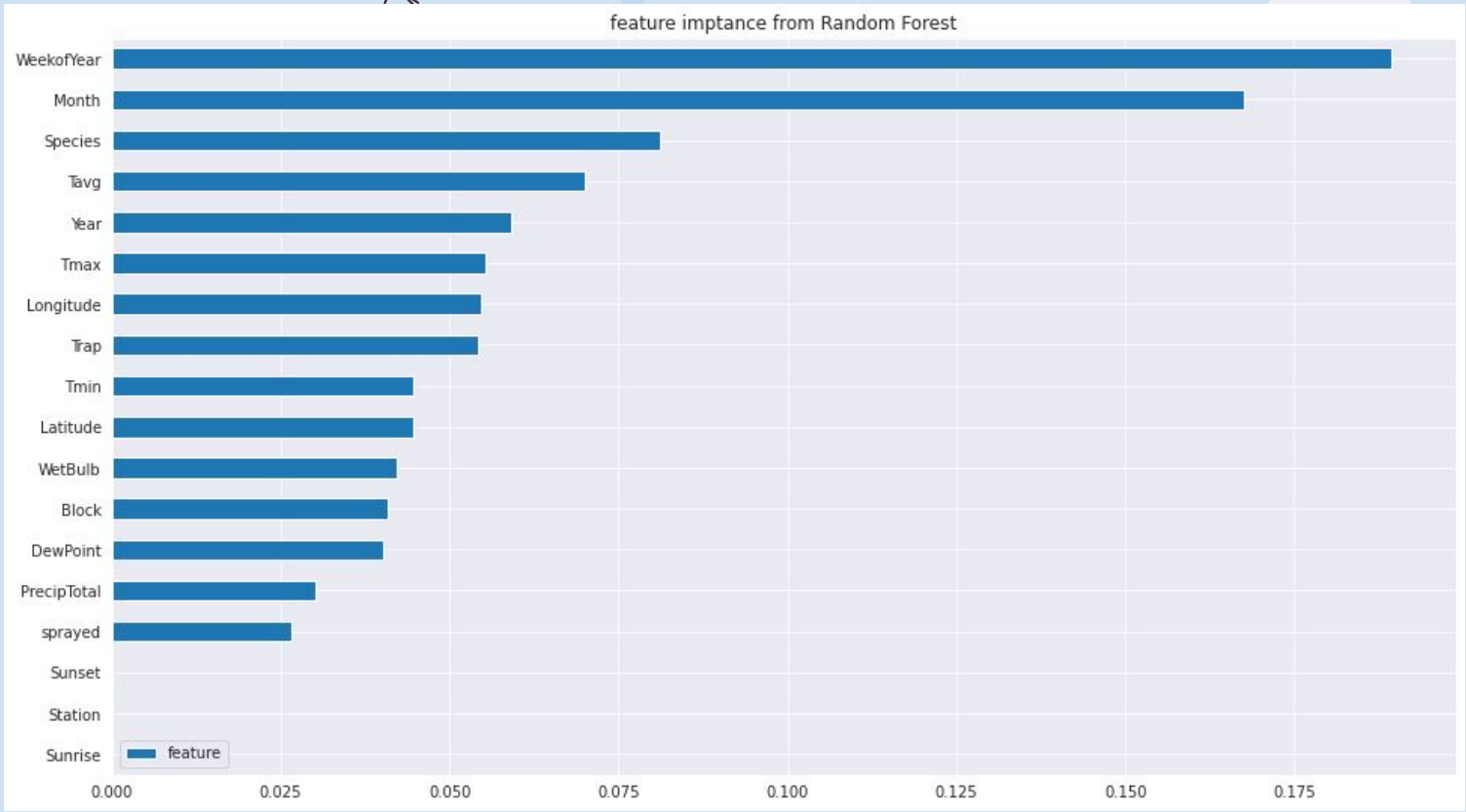
In [ ]:
```
# rf_param_grid = {'rf__min_samples_split' : np.arange(2 , 10 , 2),
#                  'rf__min_samples_leaf' : np.arange(1, 5 , 1),
#                  'rf__n_estimators' : np.arange(50 , 80 , 10),
#                  'rf__max_features' : ['log2' , 'sqrt' , 'auto'],
#                  'rf__max_depth' : [None , 3 , 5]}


# rf_smote = run_model('rf' , mod_params = rf_param_grid ,  grid_search = True)
# prediction_list,proba_list= predictions(rf_smote)
# evaluation_plot(rf_smote, prediction_list,proba_list)
```

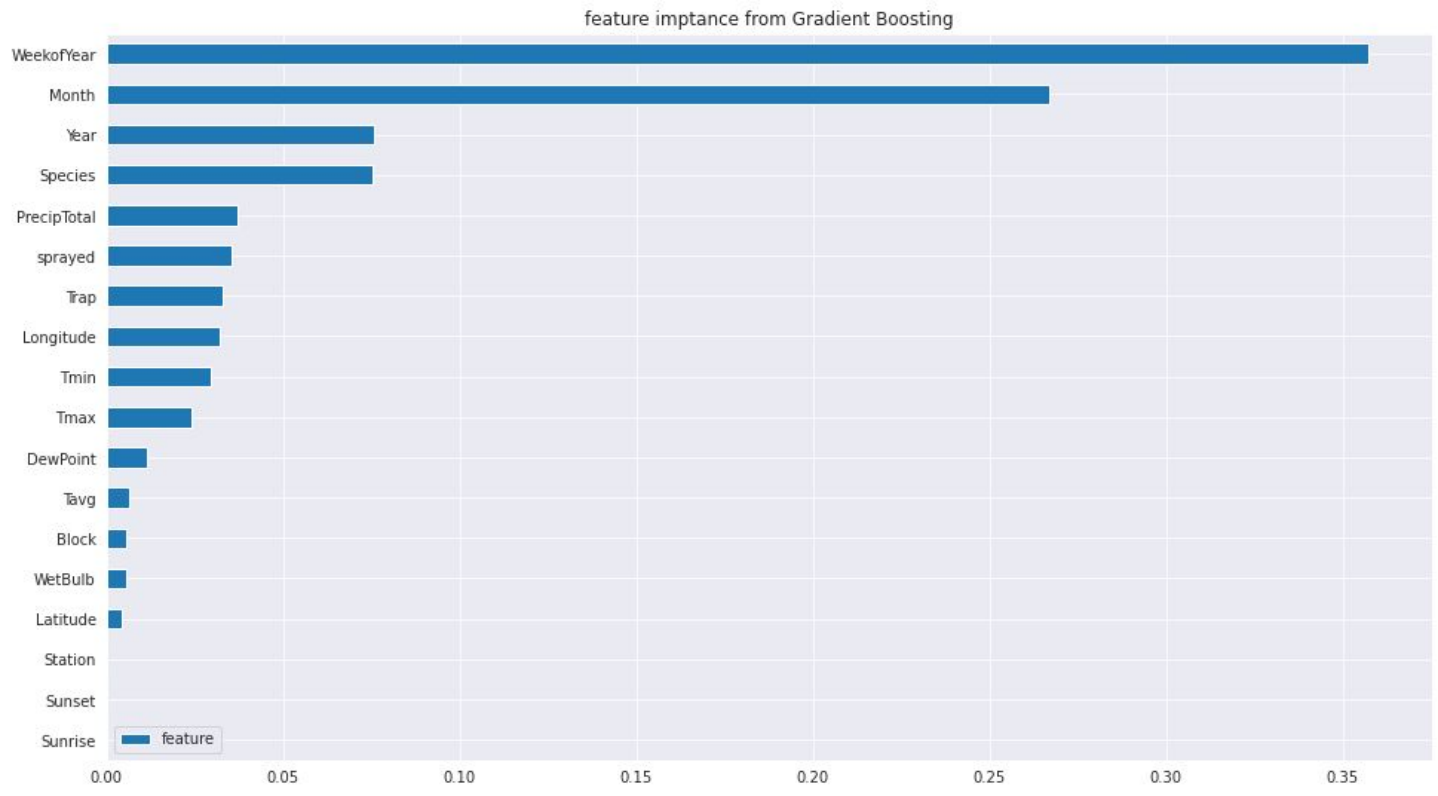
In [49]:
```
# xgb_param_grid = {
#     'xgb__n_estimators': [100, 200, 500],
#     'xgb__learning_rate': [0.01,0.05,0.1],
#     'xgb__booster': ['gbtree', 'gblinear'],
#     'xgb__gamma': [0, 0.5, 1],
#     'xgb__reg_alpha': [0, 0.5, 1],
#     'xgb__reg_lambda': [0.5, 1, 5],
#     'xgb__base_score': [0.2, 0.5, 1]
# }

# xgb_smote = run_model('xgb' , mod_params = xgb_param_grid , grid_search = True)
# prediction_list,proba_list= predictions(xgb_smote)
# evaluation_plot(xgb_smote, prediction_list,proba_list)
```

Appendix : Feature Impt (RFC)



Appendix : Feature Impt (GB)



Appendix : Feature Impt (XGB)