PROJECT - 4

West Nile Virus Prediction

The team

- LARP
- B.B.
- PUNT

Background

West Nile Virus (WNV) is a deadly virus found in mosquitos. One it is infected to human, 20% of people develop symtoms ranging from a persistent fever, to serious neurological illnesses that can result in death.

City of Chicago and CDPH together want to control the spread of mosquitos, hence control the spread of WNV.

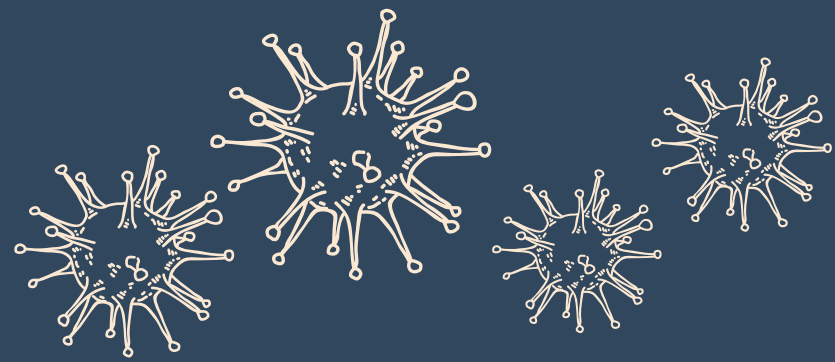
Problem Statement

Our team, as the data scientists from CDC sent to City of Chicago to develop the model to predict the occurence of WNV, so the City of Chicago can use the model to plan for WNV control, and do cost-benefit analysis to justify the plan.

The result will be presented to members of City of Chicago and CDPH, including biostatistician and epidemiologists.

Outline

- Data Study
- Data Cleansing
- Exploratory Data Analysis
- Modeling
  - Pre/Post Feature Re-Engineering
  - Feature Important Analysis
- Plan and Cost Benefit
- Limitation, Conclusion, and Recommend



The Data

features from train data and weather data that have been used for modeling are:

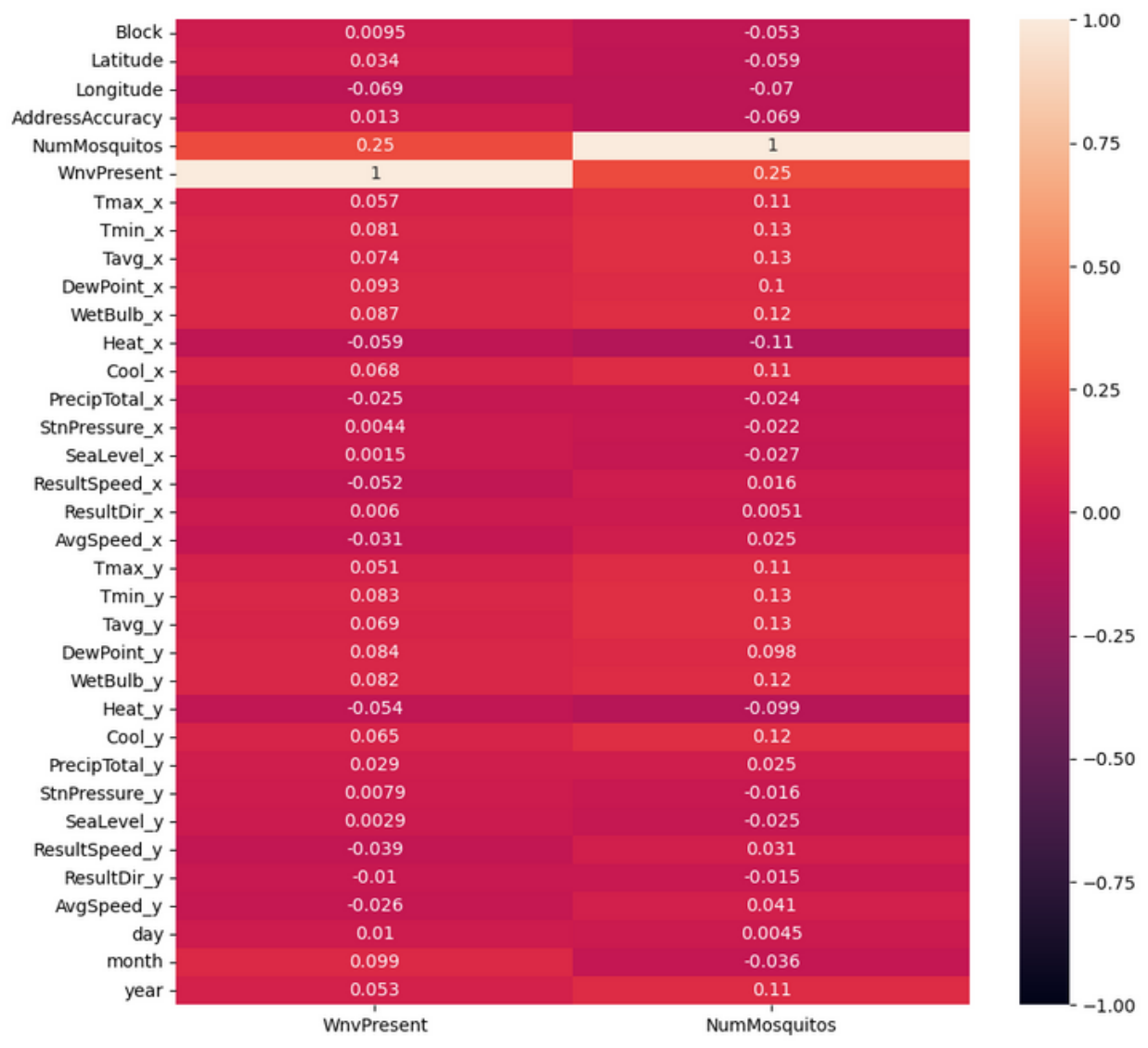
'Species', 'Block', 'Street', 'Trap', 'Latitude', 'Longitude', 'WnvPresent', 'DewPoint', 'Heat', 'Cool', 'PrecipTotal', 'StnPressure', 'SeaLevel', 'ResultSpeed', 'ResultDir', 'AvgSpeed', ‘date’, ‘Tmax’, ‘Tmin’, ‘Tavg’, ‘Wetbulb’

Cleansing

- Null location and treatment (only null found in spray data)
- Drop duplicate record
- Weather data has no "null" value, but virtually, we see blank and missing data. Replace them with "null" before consider dropping them
- Merge weather data to train and test data

	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset	CodeSum	Depth	Water1	SnowFall	PrecipTotal	StnPressure	SeaLevel	ResultSpeed	ResultDir	AvgSpeed
0	1	2007-05-01	83	50	67	14	51	56	0	2	0448	1849		0	M	0.0	0.00	29.10	29.82	1.7	27	9.2
1	2	2007-05-01	84	52	68	M	51	57	0	3	-	-		M	M	M	0.00	29.18	29.82	2.7	25	9.6
2	1	2007-05-02	59	42	51	-3	42	47	14	0	0447	1850	BR	0	M	0.0	0.00	29.38	30.09	13.0	4	13.4
3	2	2007-05-02	60	43	52	M	42	47	13	0	-	-	BR HZ	M	M	M	0.00	29.44	30.08	13.3	2	13.4
4	1	2007-05-03	66	46	56	2	40	48	9	0	0446	1851		0	M	0.0	0.00	29.39	30.12	11.7	7	11.9

EDA



Correlation:

The only significant correlations (>0.5) are among weather data. which isn't usable for any prediction.

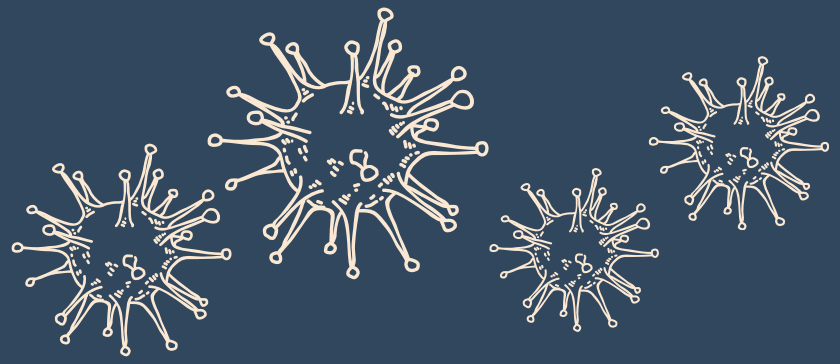
So we tried mapping the correlation toward number of mosquito and the present of virus to check if there is any correlation. Because the present of virus is the outcome that we're predicting and since the virus is coming from the mosquitos, we want to check the correlation from both of them as well.

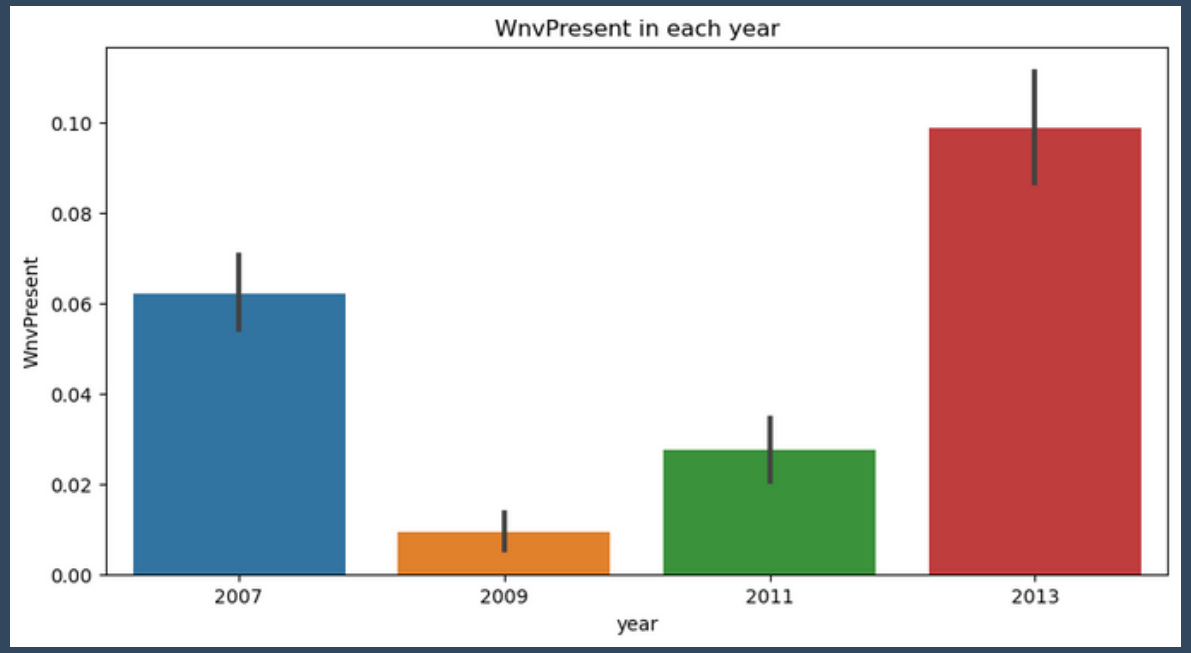
The correlation toward present of virus and number of mosquito are very low. We don't see significant indicator to the present of virus in single feature here

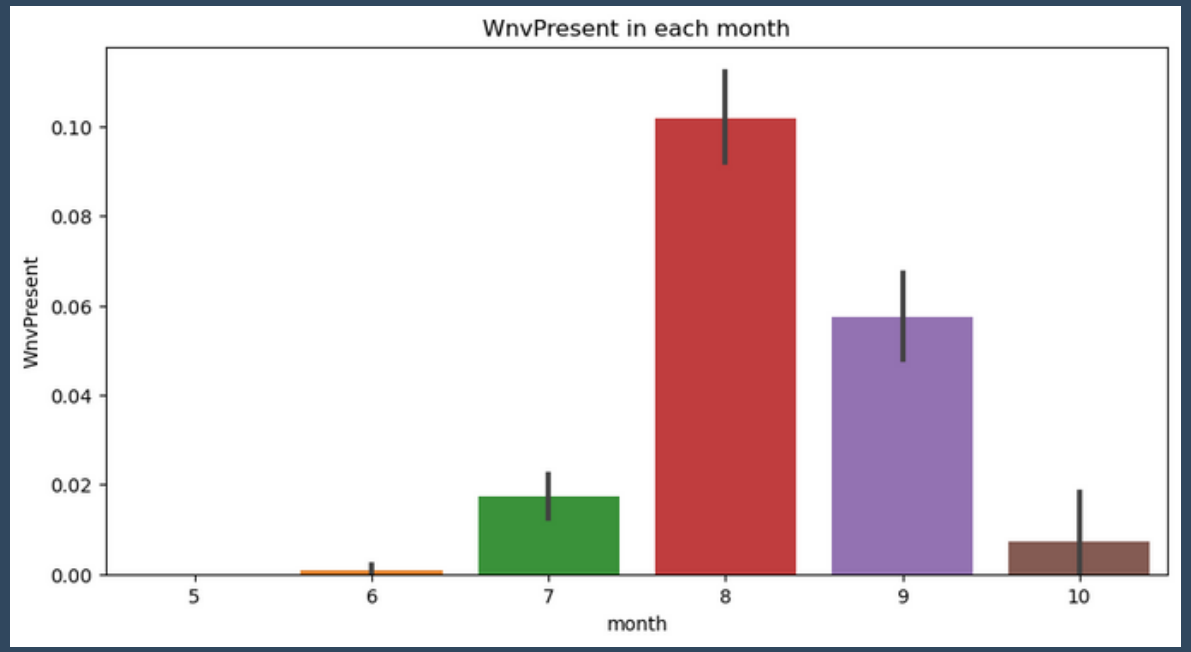
EDA..cont

Train dataset:

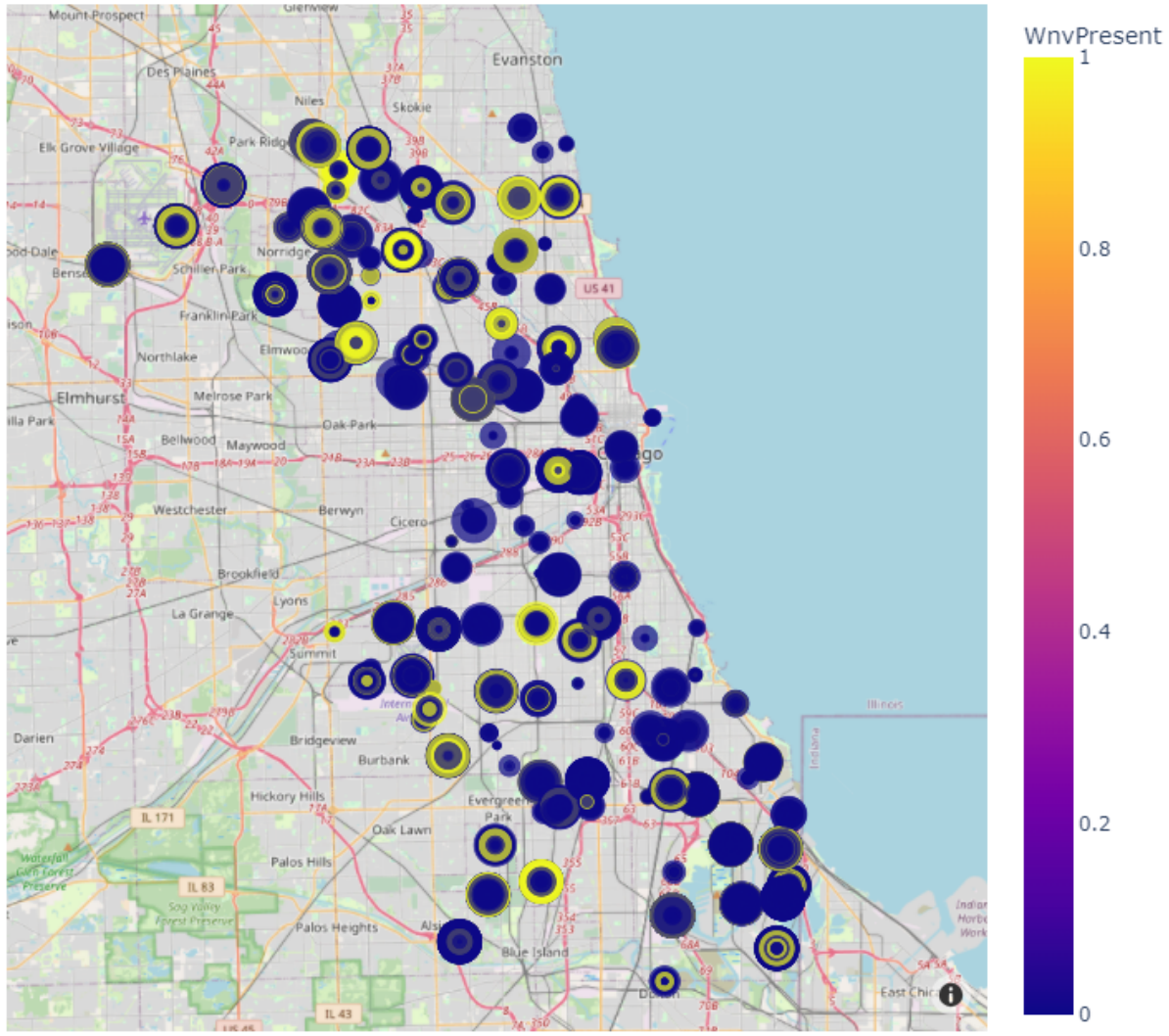
- Virus present clearly during summer
- Virus found more in 2013
- There are some location that virus found more than others, but it looks like it spread all over the city (see next page)

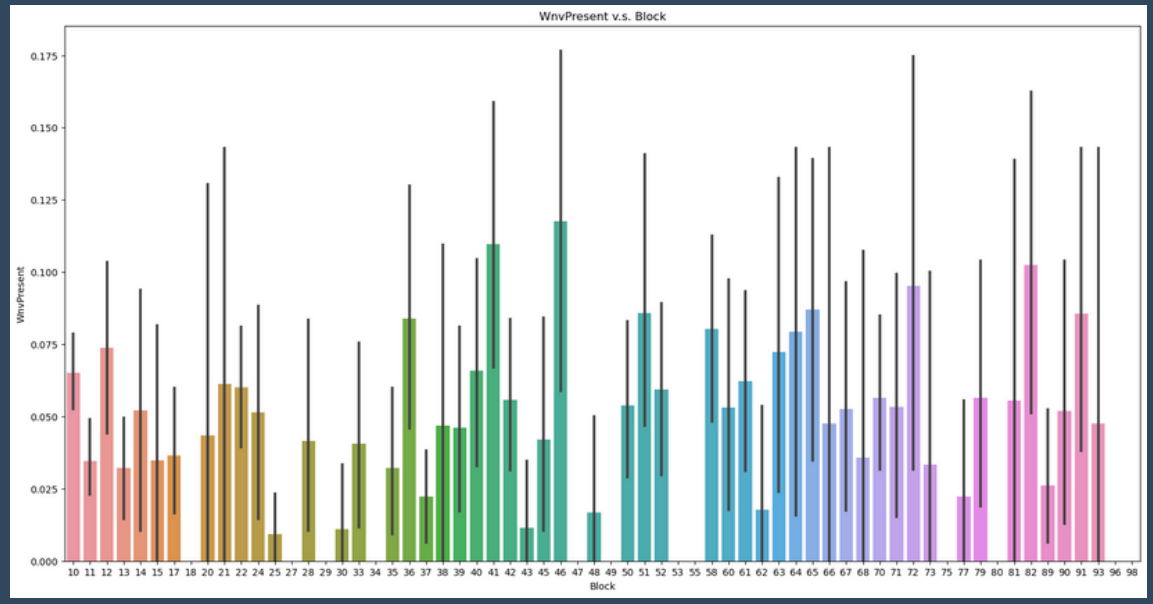






EDA..cont





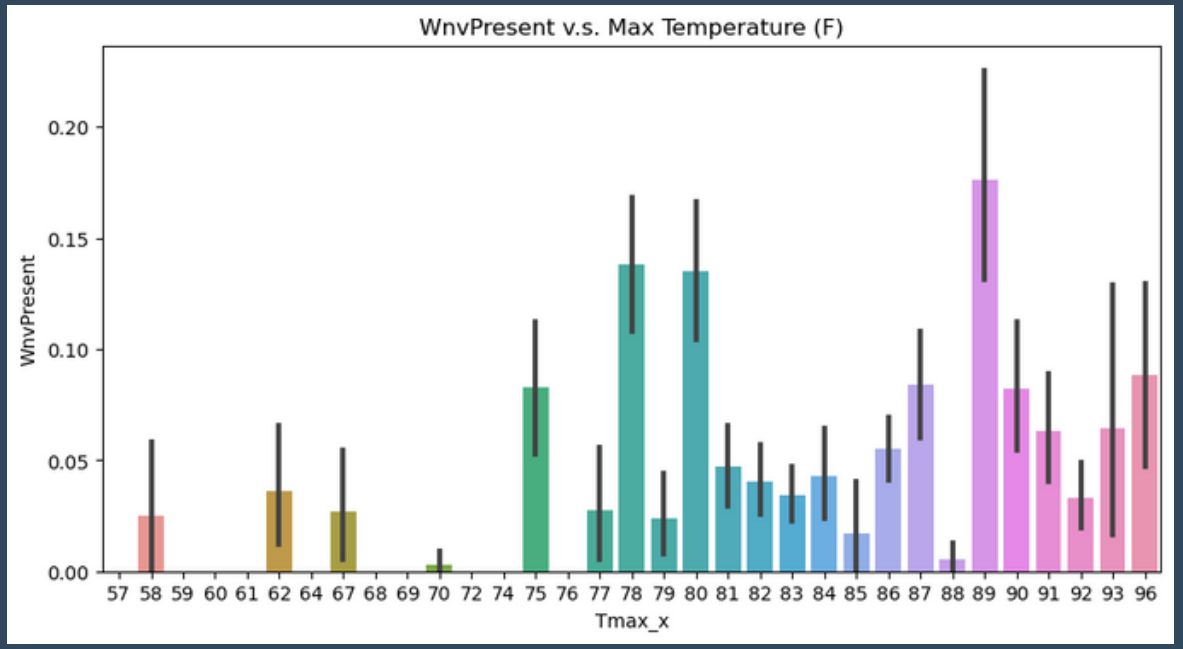
Train dataset (cont.):

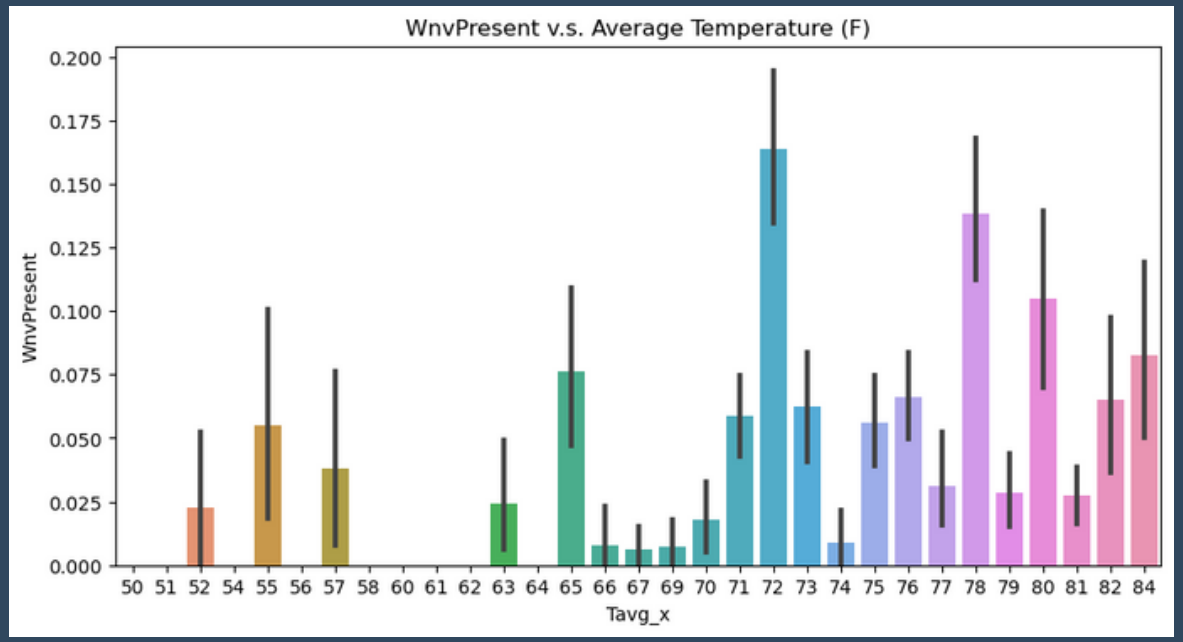
- There are some location that virus found more than others, but it looks like it spread all over the city

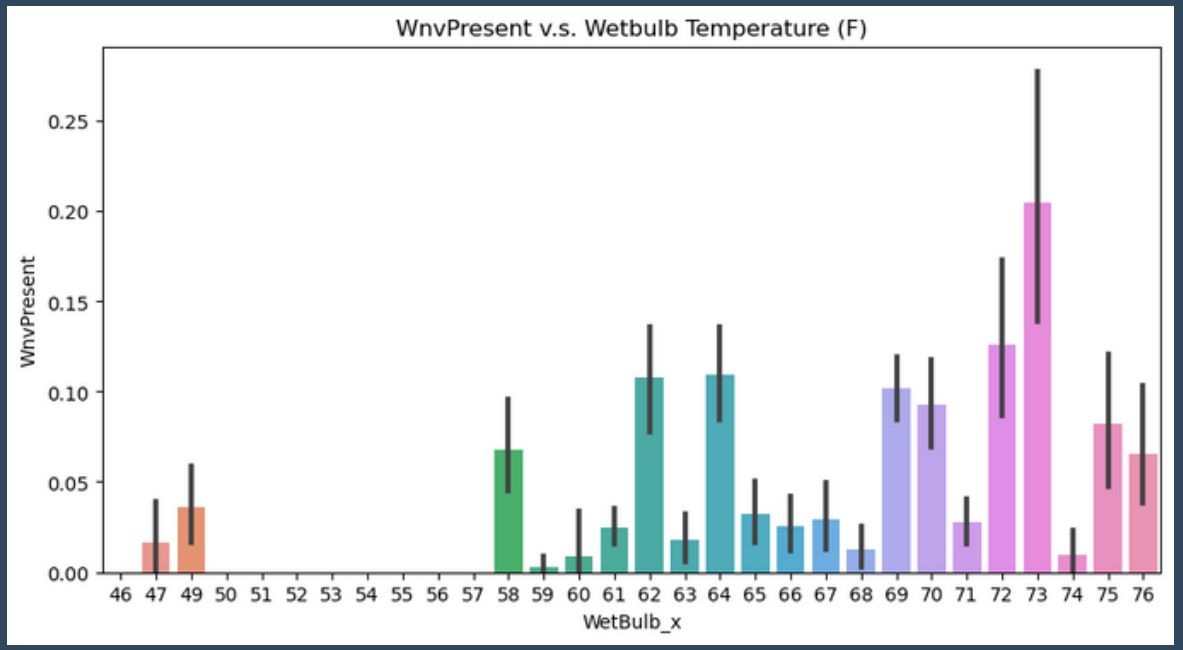
EDA..cont

Weather dataset:

- All temperature indicator (max/min/average/wetbulb temperature, shows the same pattern the the higher the temperature it is like to be more virus present







Modeling

- Pre Feature Re-Engineering
- Feature Important Analysis
- Post Feature Re-Engineering

Modeling

Pre Feature Re-Engineering

Model	Train ROC AUC Score	Test ROC AUC Score
Logistic Regression	0.82	0.81
Decision Tree	1.00	0.61
Random Forest	0.99	0.77
Adaboost	0.87	0.83
Bagged Decision Tree	0.99	0.73
SVM	0.44	0.44
XGBoost	**0.99**	**0.84**

Modeling

Pre Feature Re-Engineering

36 Selected Features				
Species	Block	Street	Trap	Latitude
Longitude	AddressAccuracy	Tmax_x	Tmin_x	Tavg_x
DewPoint_x	WetBulb_x	Heat_x	Cool_x	PrecipTotal_x
StnPressure_x	SeaLevel_x	ResultSpeed_x	ResultDir_x	AvgSpeed_x
Tmax_y	Tmin_y	Tavg_y	DewPoint_y	WetBulb_y
Heat_y	Cool_y	PrecipTotal_y	StnPressure_y	SeaLevel_y
ResultSpeed_y	ResultDir_y	AvgSpeed_y	day	month
year				

Note: The columns _x is splited from station 1 data, and _y is from station 2

Target label
WNVPresent

WNVpresent				
Proportion	0	0.94761	1	0.05239

Model

XGBoost

Hyperparameter

Default

Tuning (GridSeachCV)

Train/Test Data

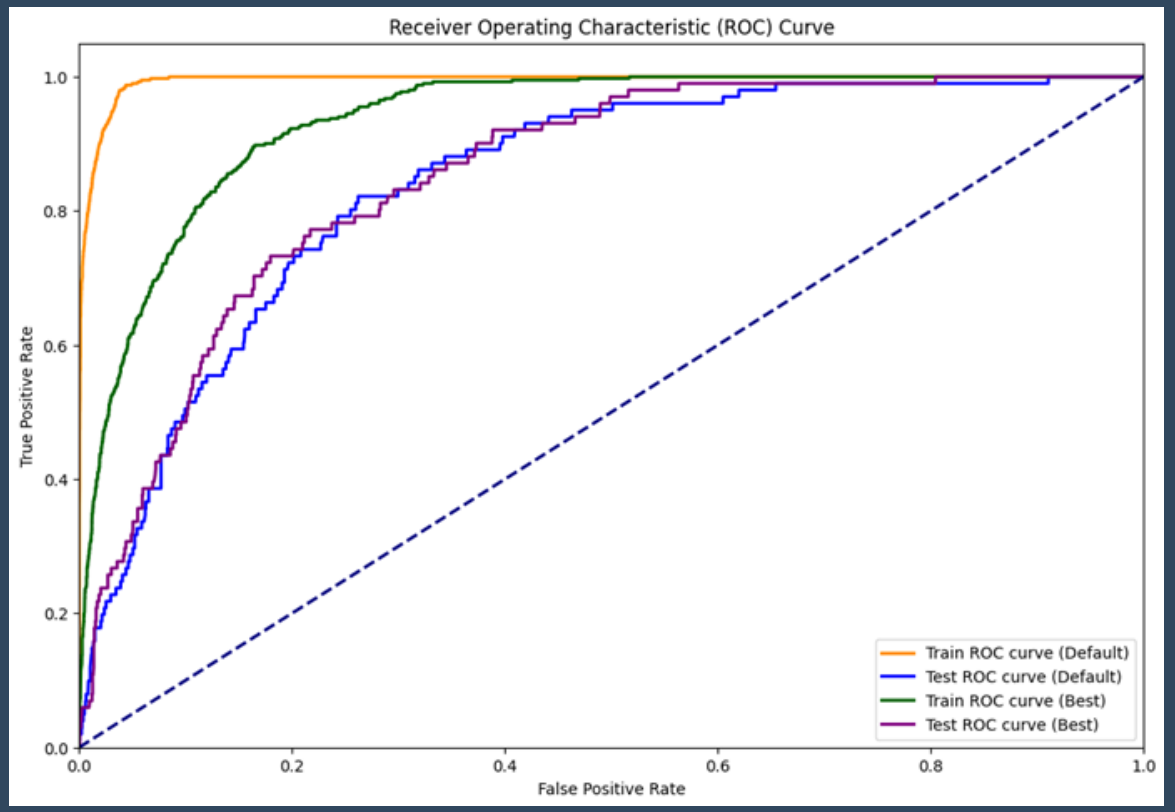
Default

Over Sampling (SMOTE)

Modeling

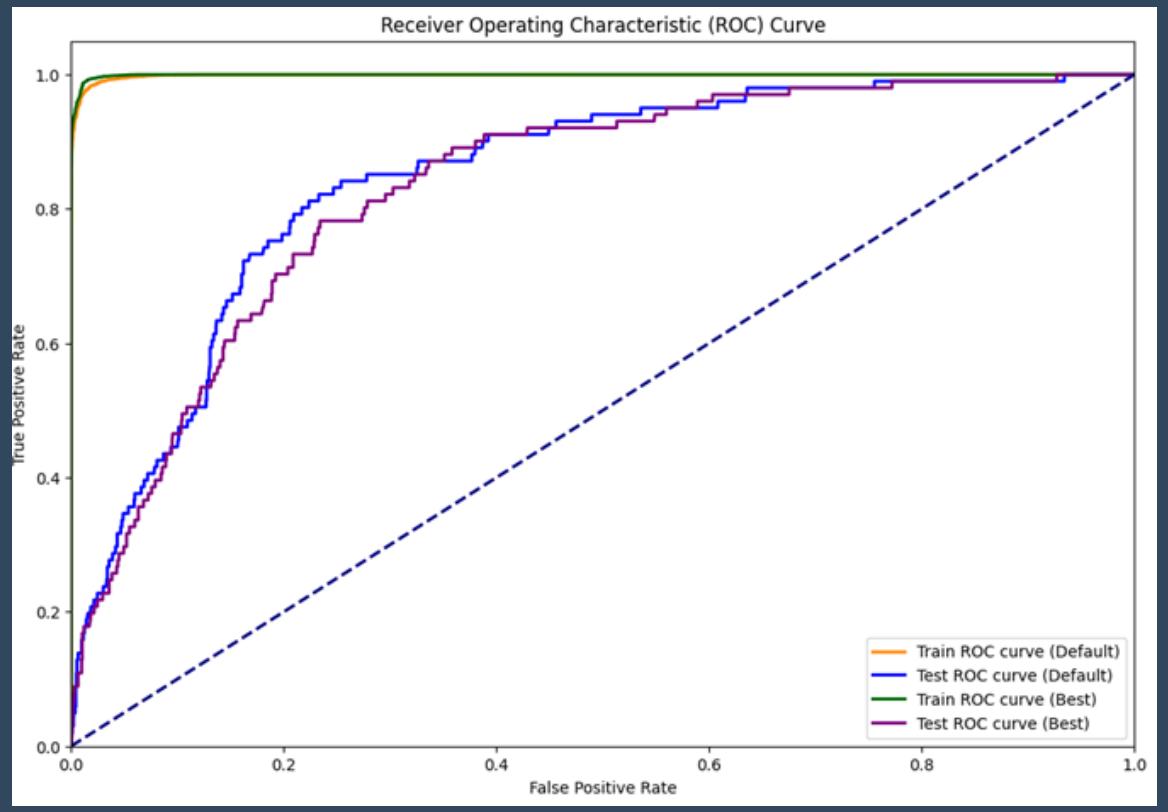
Pre Feature Re-Engineering

Default model and Tuned model (no SMOTE technique)



GridSeachCV	max_depth:3	n_estimators:100	reg_alpha:1.0

Default model and Tuned model (with SMOTE technique)

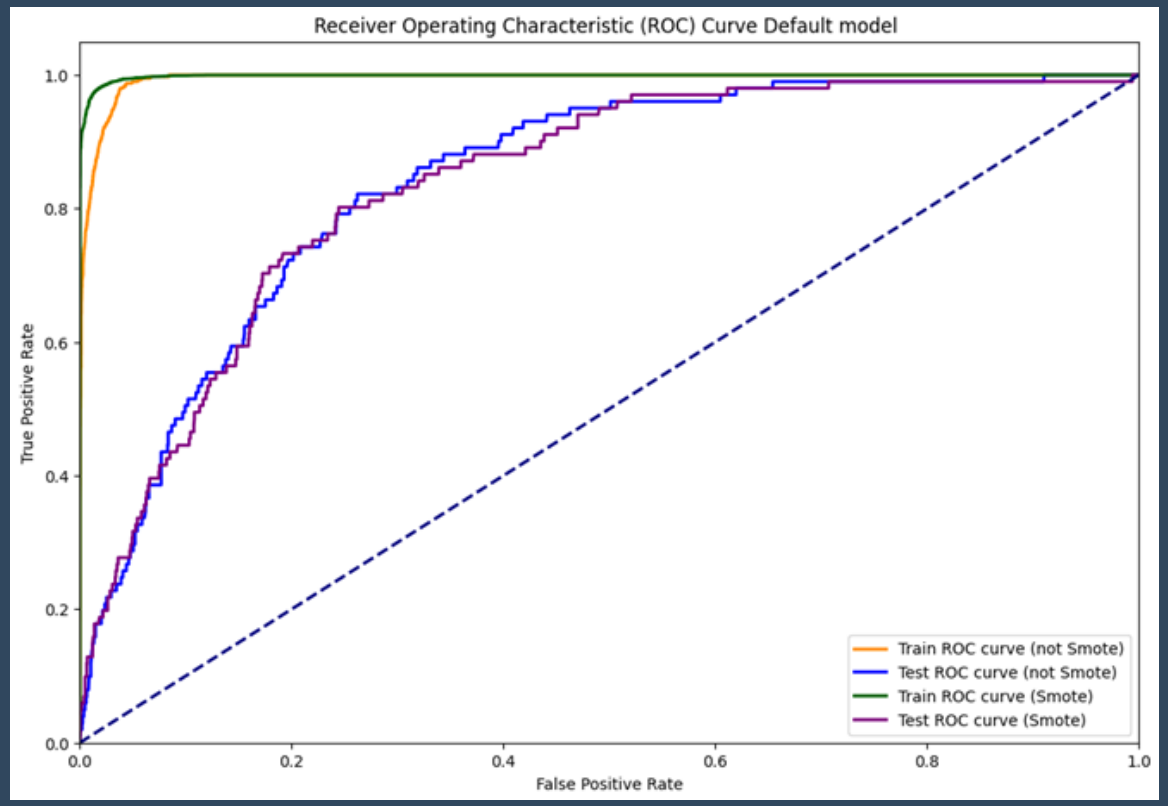


GridSeachCV	max_depth:5	n_estimators:300	reg_alpha:1.0

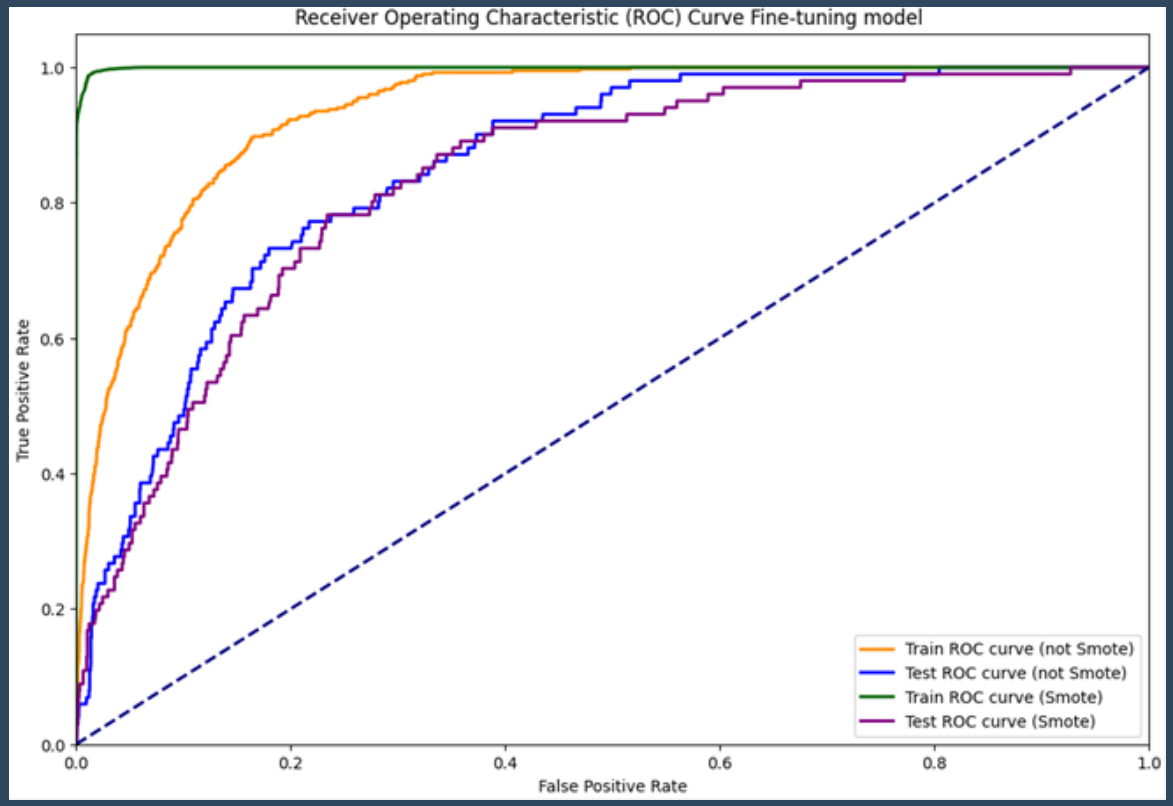
Modeling

Pre Feature Re-Engineering

Default model (with SMOTE and no SMOTE)



Tuned model (with SMOTE and no SMOTE)



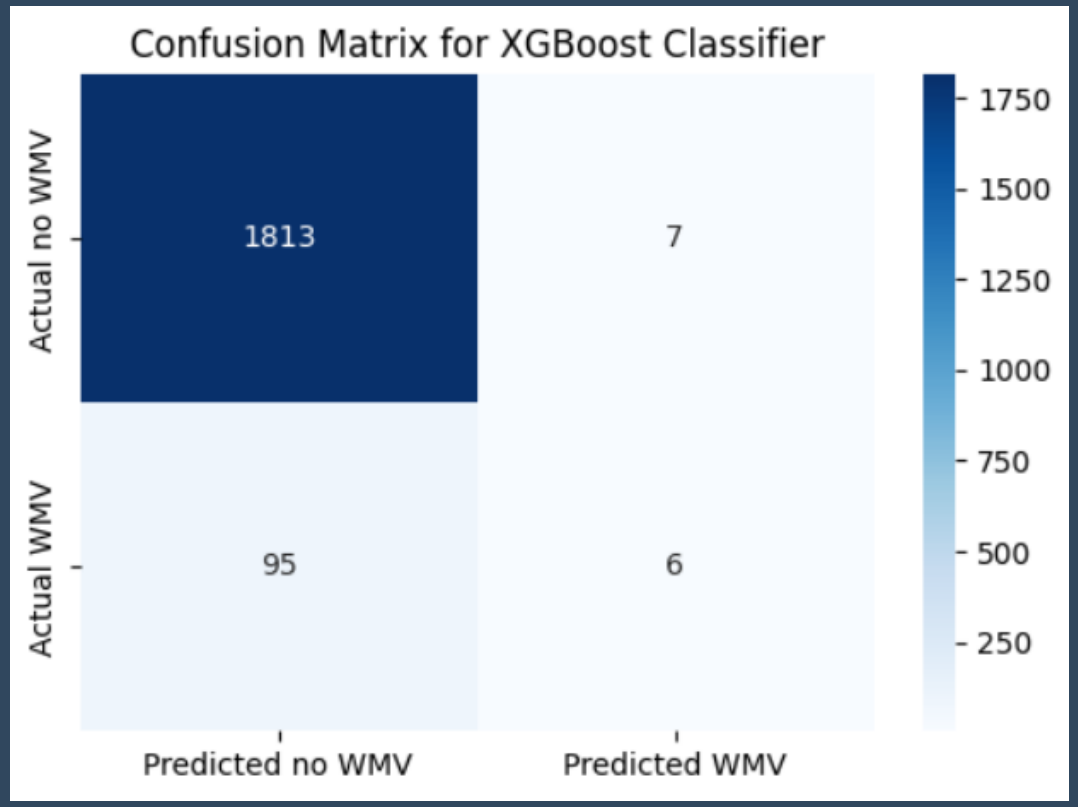
Modeling

Pre Feature Re-Engineering

Model	Train ROC AUC Score	Test ROC AUC Score	Kaggle Score
• **Pre Feature re-engineering**			
XGBoost	0.99	0.84	
XGBoost + Fine-tuning hyperparameter	0.94	0.85	0.71
SMOTE + XGBoost	1.00	0.84	
SMOTE + XGBoost + Fine-tuning hyperparameter	1.00	0.84	0.68

Modeling

Pre Feature Re-Engineering

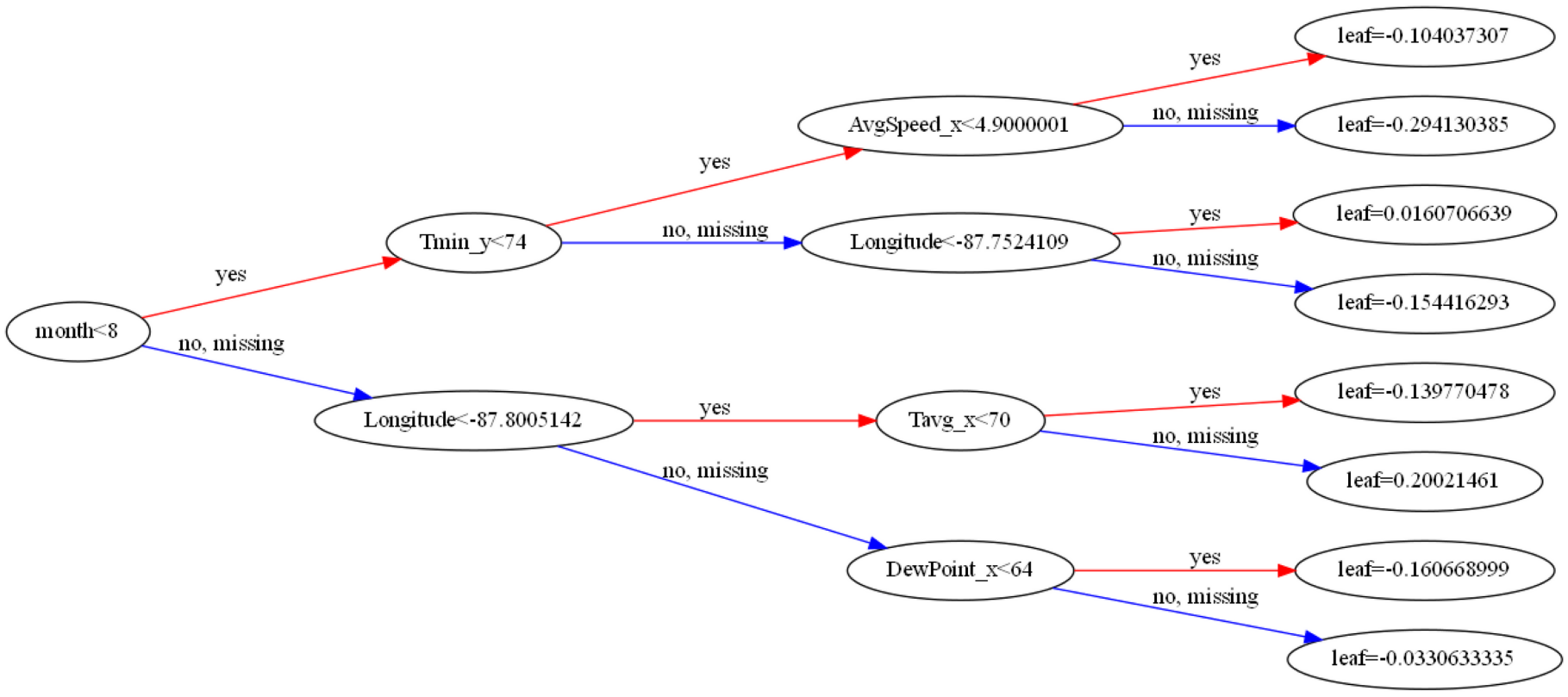


0	Negative	Predicted No WMV
1	Positive	Predicted WMV

Accuracy Score:	0.9469
Sensitivity:	0.0594
Specificity:	0.9962
Precision:	0.4615
F1 Score:	0.1053

Modeling

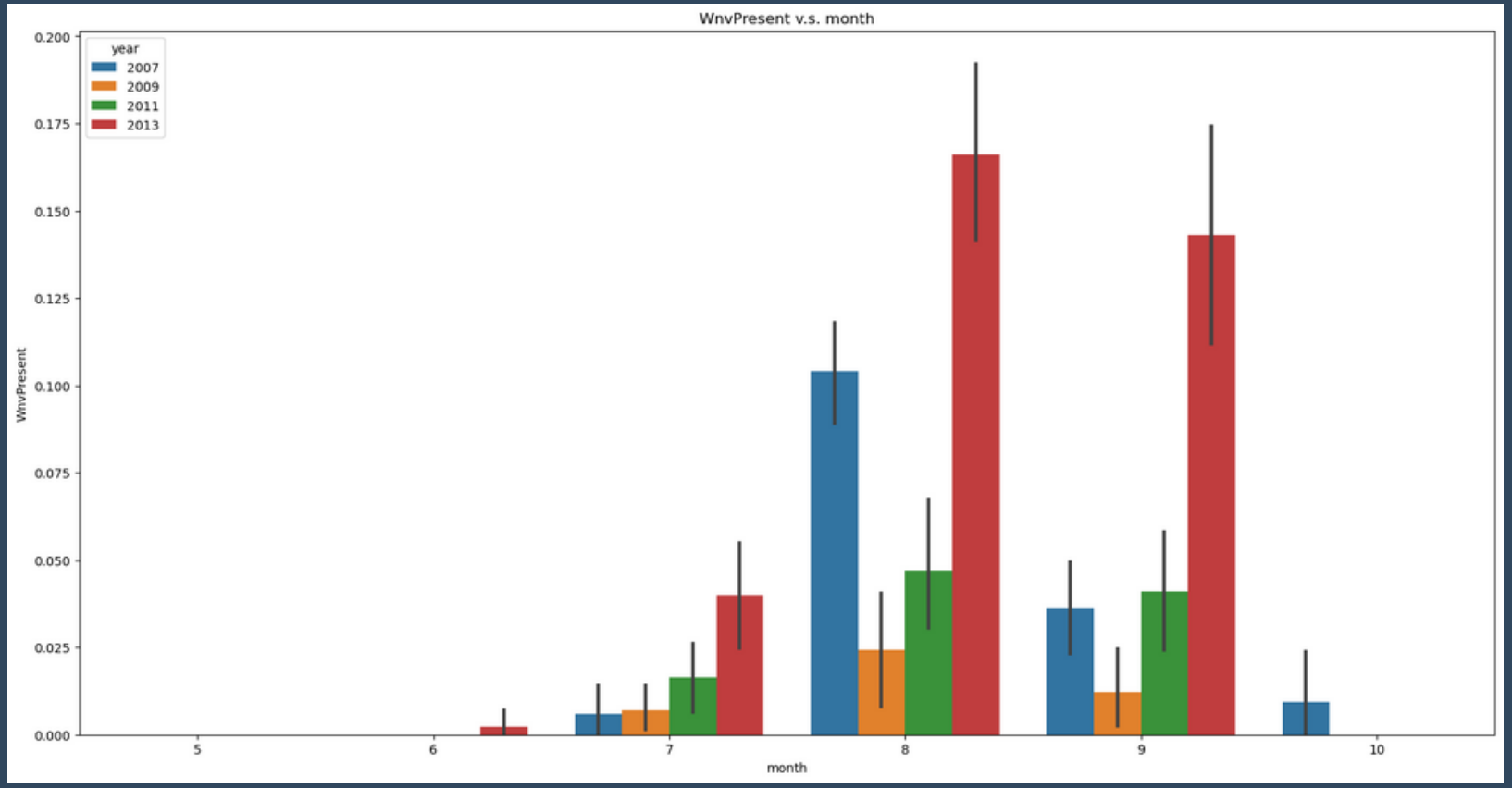
Feature Important Analysis



Modeling

Feature Important Analysis

After reviewing a spray file, it was done only in 2011 and 2013, but from the data we have. The Virus has the the most frequently occurring number in 2013. This looks like pesticides spray wasn't effective, so let's leave this out and be more focus on weather data. Finally, we ended up with more features creation from creating the bin for every 8 degree F of Tmax, Tmin, Tavg, and Wetbulb.



Modeling

Post Feature Re-Engineering

71 Selected Features				
Id	Species	Block	Street	Trap
Latitude	Longitude	DewPoint_x	Heat_x	Cool_x
PrecipTotal_x	StnPressure_x	SeaLevel_x	ResultSpeed_x	ResultDir_x
AvgSpeed_x	day	month	year	Tmax_x_bin_bin2
Tmax_x_bin_bin3	Tmax_x_bin_bin4	Tmax_x_bin_bin5	Tmax_x_bin_bin6	Tmax_x_bin_bin7
Tmax_x_bin_bin8	Tmax_x_bin_bin9	Tmax_x_bin_bin10	Tmax_x_bin_bin11	Tmax_x_bin_bin12
Tmax_x_bin_bin13	Tmax_x_bin_bin14	Tavg_x_bin_bin2	Tavg_x_bin_bin3	Tavg_x_bin_bin4
Tavg_x_bin_bin5	Tavg_x_bin_bin6	Tavg_x_bin_bin7	Tavg_x_bin_bin8	Tavg_x_bin_bin9
Tavg_x_bin_bin10	Tavg_x_bin_bin11	Tavg_x_bin_bin12	Tavg_x_bin_bin13	Tavg_x_bin_bin14
Tmin_x_bin_bin2	Tmin_x_bin_bin3	Tmin_x_bin_bin4	Tmin_x_bin_bin5	Tmin_x_bin_bin6
Tmin_x_bin_bin7	Tmin_x_bin_bin8	Tmin_x_bin_bin9	Tmin_x_bin_bin10	Tmin_x_bin_bin11
Tmin_x_bin_bin12	Tmin_x_bin_bin13	Tmin_x_bin_bin14	WetBulb_x_bin_bin2	WetBulb_x_bin_bin3
WetBulb_x_bin_bin4	WetBulb_x_bin_bin5	WetBulb_x_bin_bin6	WetBulb_x_bin_bin7	WetBulb_x_bin_bin8
WetBulb_x_bin_bin9	WetBulb_x_bin_bin10	WetBulb_x_bin_bin11	WetBulb_x_bin_bin12	WetBulb_x_bin_bin13
WetBulb_x_bin_bin14				
Note: The columns _x is splited from station 1 data, but station 2 data is dropped out				

Target label
WNVPresent

Model

XGBoost

Hyperparameter

Default

Tuning (GridSeachCV)

Train/Test Data

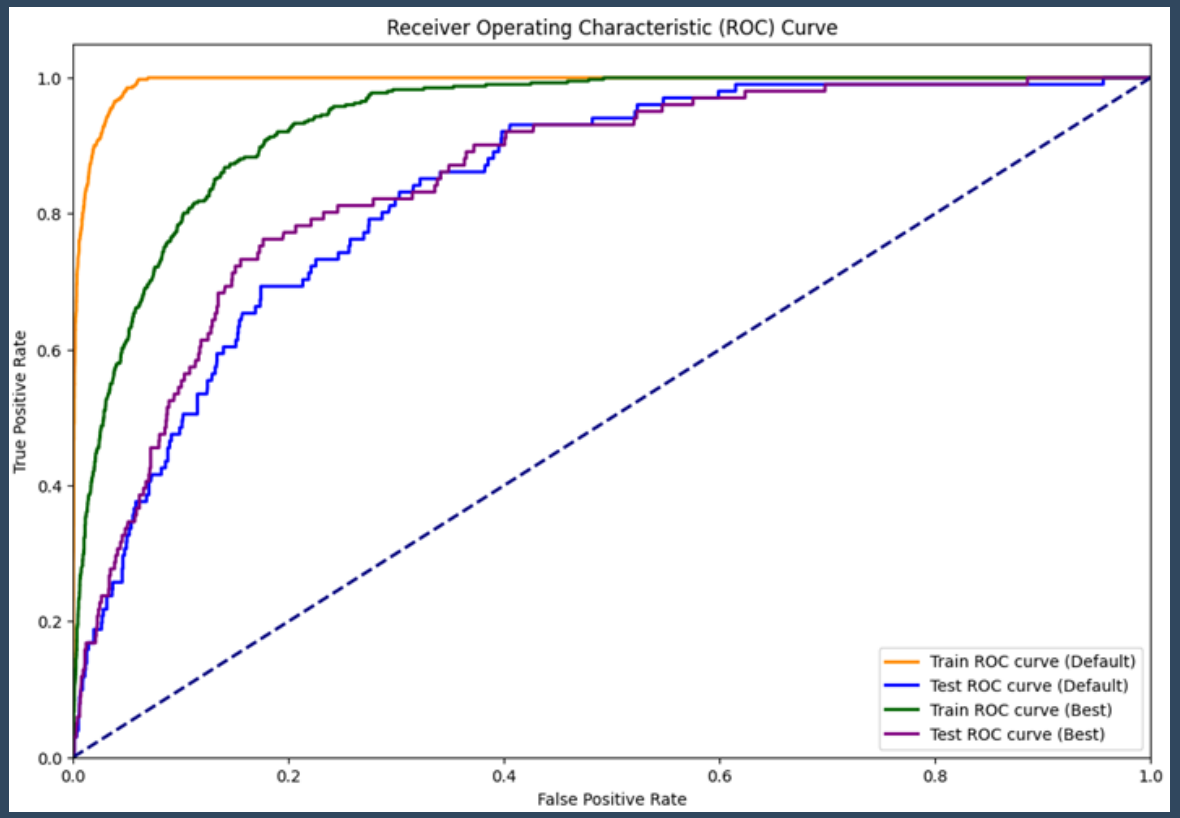
Default

Over Sampling (SMOTE)

Modeling

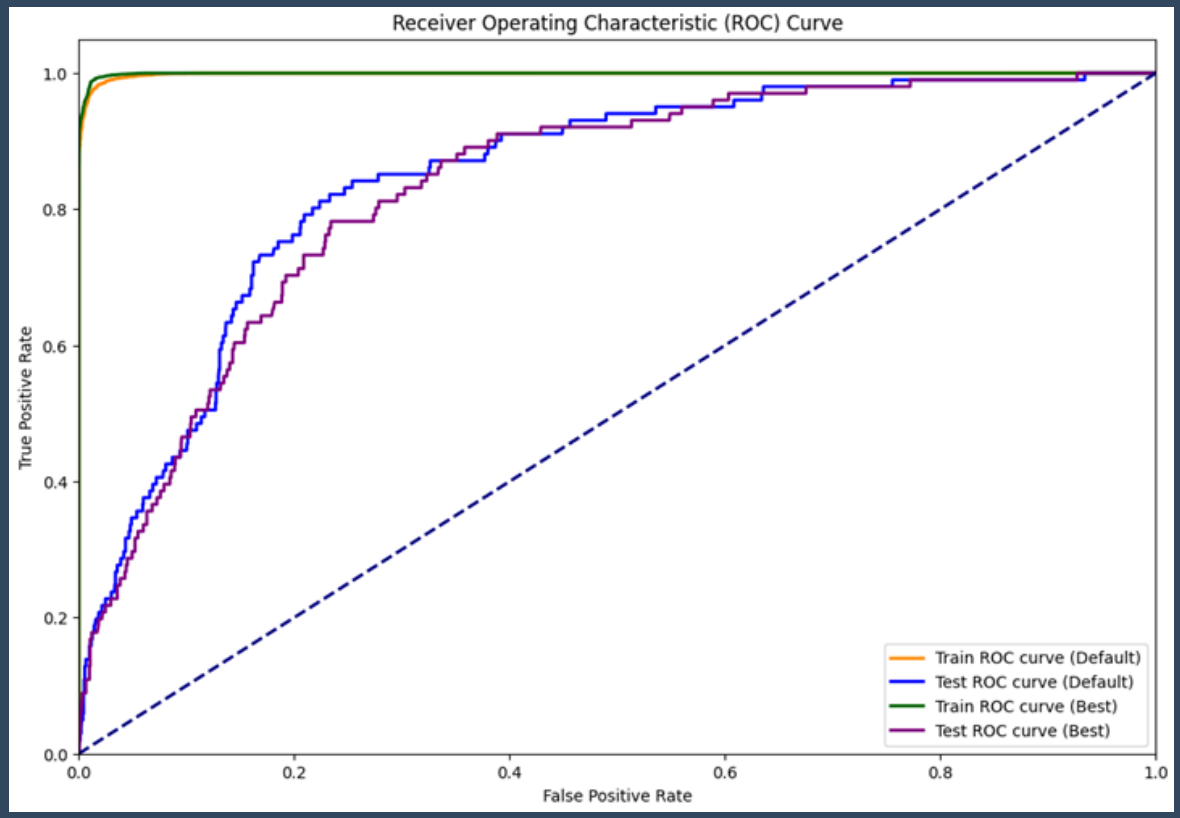
Post Feature Re-Engineering

Default model and Tuned model (no SMOTE technique)



GridSeachCV	max_depth:3	n_estimators:100	reg_alpha:1.0

Default model and Tuned model (with SMOTE technique)

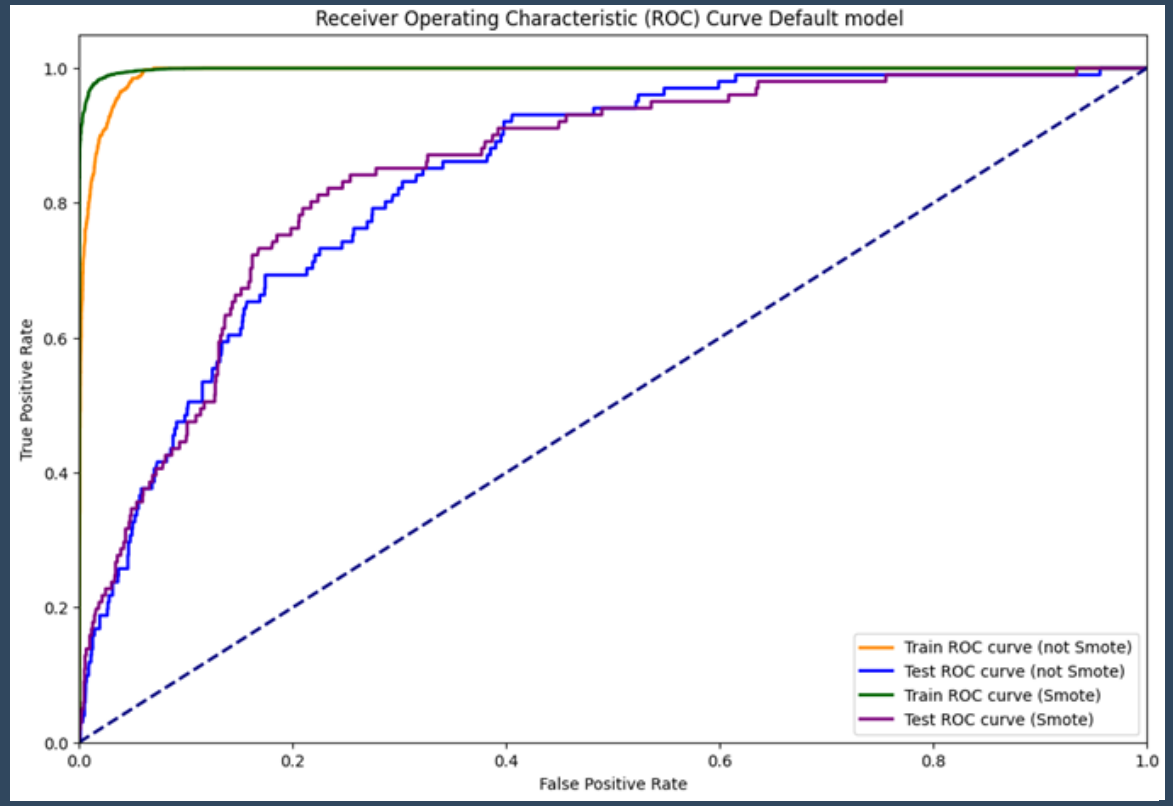


GridSeachCV	max_depth:5	n_estimators:300	reg_alpha:1.0

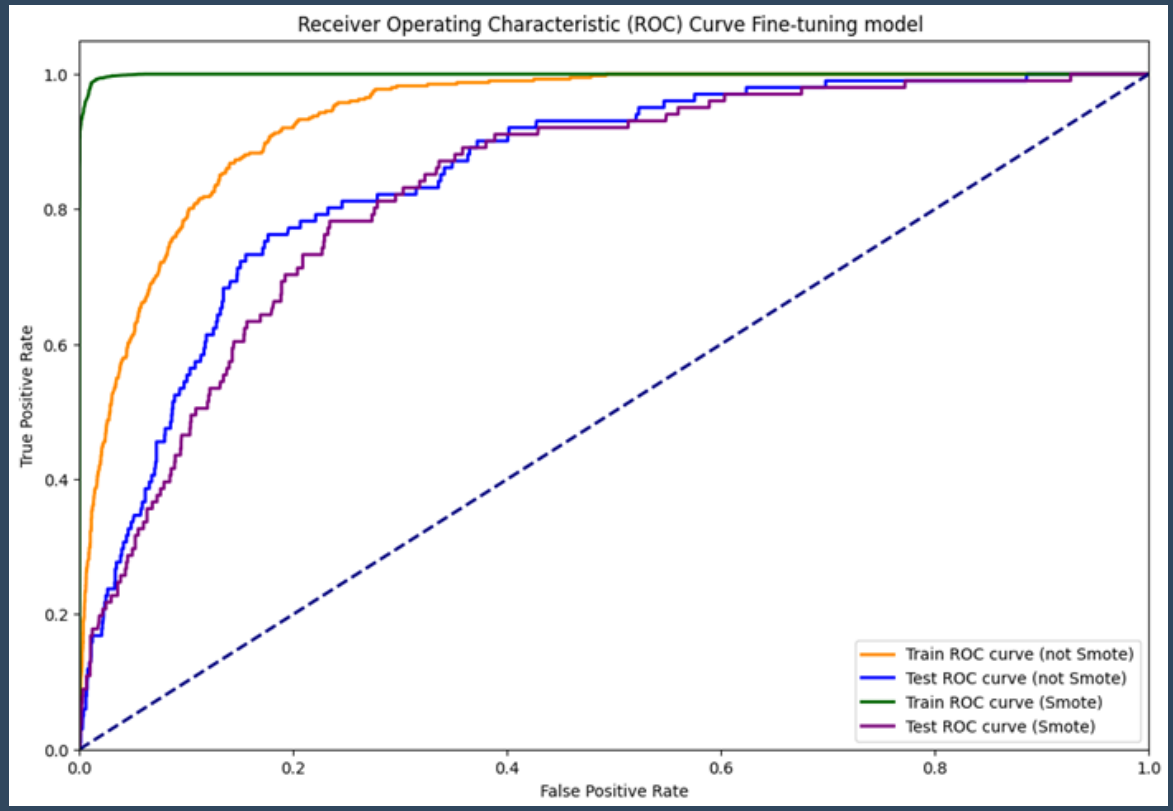
Modeling

Post Feature Re-Engineering

Default model (with SMOTE and no SMOTE)



Tuned model (with SMOTE and no SMOTE)



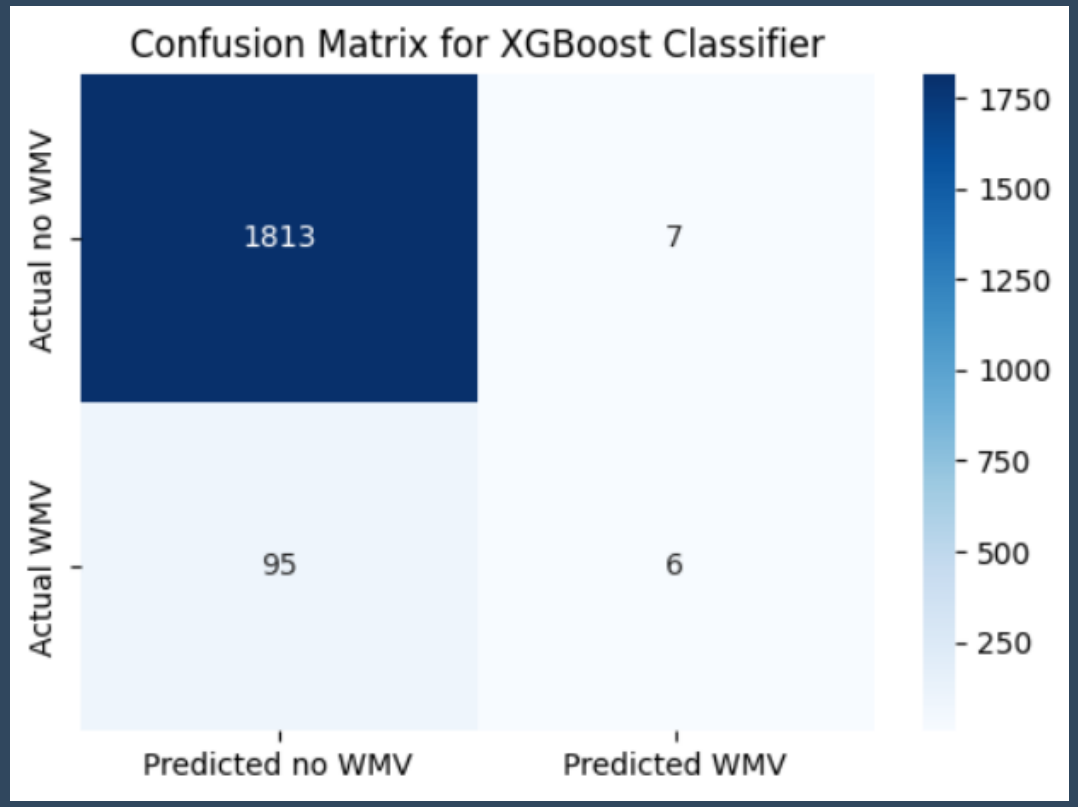
Modeling

Post Feature Re-Engineering

Model	Train ROC AUC Score	Test ROC AUC Score	Kaggle Score
• **Pre Feature re-engineering**			
XGBoost	0.99	0.84	
XGBoost + Fine-tuning hyperparameter	0.94	0.85	0.71
SMOTE + XGBoost	1.00	0.84	
SMOTE + XGBoost + Fine-tuning hyperparameter	1.00	0.84	0.68
• **Post Feature re-engineering**			
XGBoost	0.99	0.84	
XGBoost + Fine-tuning hyperparameter	0.94	0.85	0.72
SMOTE + XGBoost	1.00	0.84	
SMOTE + XGBoost + Fine-tuning hyperparameter	1.00	0.83	0.72

Modeling

Pre Feature Re-Engineering



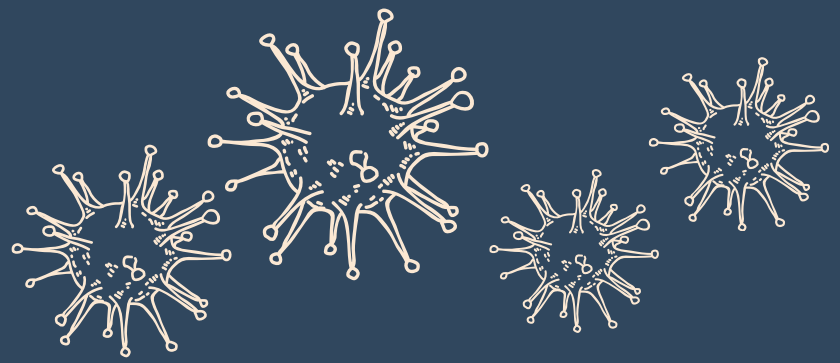
0	Negative	Predicted No WMV
1	Positive	Predicted WMV

Accuracy Score:	0.9469
Sensitivity:	0.0594
Specificity:	0.9962
Precision:	0.4615
F1 Score:	0.1053

Plan

Report

we can use our prediction to check predicted locations where you see water sitting stagnant for more than a week such as roadside ditches, flooded yards, and similar locations that may produce mosquitoes.



Reduce

we can use our prediction to make sure that in the predicted areas ...

- - sprays are used preventively
- - doors and windows have tight-fitting screens, and be repaired or replaced for those that have tears or other openings
- - doors and windows are shut
- - all sources of standing water where mosquitoes can breed, including water in bird baths, ponds, flowerpots, wading pools, old tires, and any other containers are taken care of

Repel

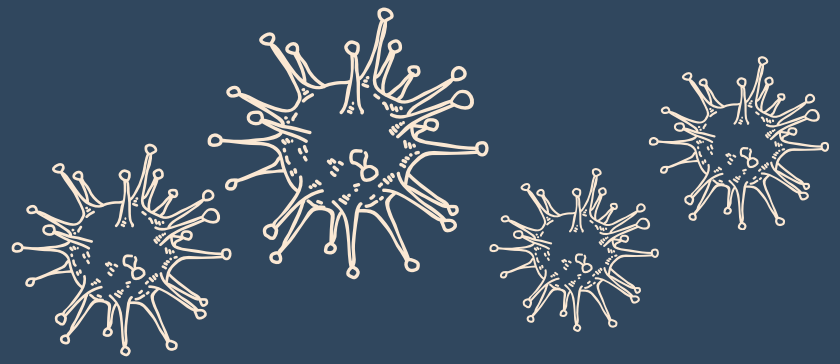
- we can use our prediction to make sure that in the residents in the predicted areas wear shoes and socks, long pants and a light-colored, long-sleeved shirt, and apply an EPA-registered insect repellent that contains DEET, picaridin, oil of lemon eucalyptus, IR 3535, para-menthane-diol (PMD), or 2-undecanone according to label instructions. Consult a physician before using repellents on infants.

Cost-Benefit

Benefit:

1. Improve the public health
   a. Based on statistics in 2022, there were 34 human cases (which are significantly under-reported) and 8 deaths attributed to the disease in the state in 2022, the most in any year since 2018, when there were 17 deaths

2. Reduce the in direct economic cost



Cost:

1. purchasing and applying larvicide,
2. working with local municipal governments and local news media for WNV prevention and education,
3. investigating mosquito production sites and nuisance mosquito complaints.
4. collecting mosquitoes for West Nile virus testing and also collect sick or dead birds for West Nile virus testing.

Data & Model Limitation

- The impact of pesticide spray is not significant. But we're not certain whether it is due to the spray which was done far away from Traps.
- The train data was strongly imbalance.
- The time data is in train data is not yearly consecutive

Conclusion & Recommendation

From our result, the key predictor is likely to be weather. When we focus our feature engineer on weather, we got a better result. we expect that spary might be good predictor as well but it is not usable.

We recommend the City

- to plan pesticide spraying to cover the trap areas especilly in predicted area.
- to always treat the imbalance data with care
- to rearrange train data to yearly consecutive data instead

Limitation, Conclusion, & Recommend