PROJECT - 4

West Nile Virus Prediction

The team

- LARB
- B.B.
- PUNT

Background

West Nile Virus 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 West Nile Virus.

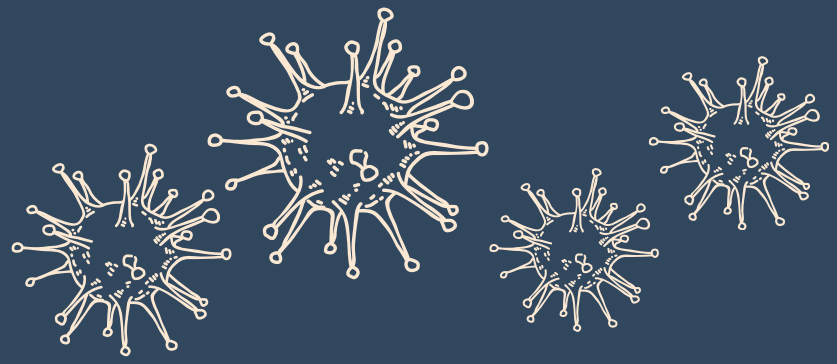
Problem Statement

Our team, as the amature data scientist group, we enter the competition hosted by City of Chicago to develop the model to predict the occurence of virus, so the City of Chicago can use them when they want to plan pesticides spraying as well as the cost-benefit analysis to justify the spraying plan.

The result will be presented to members of CDC, including biostatistician and epidemiologists.

Approach


Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment



The data

4 set of data were given for model development.

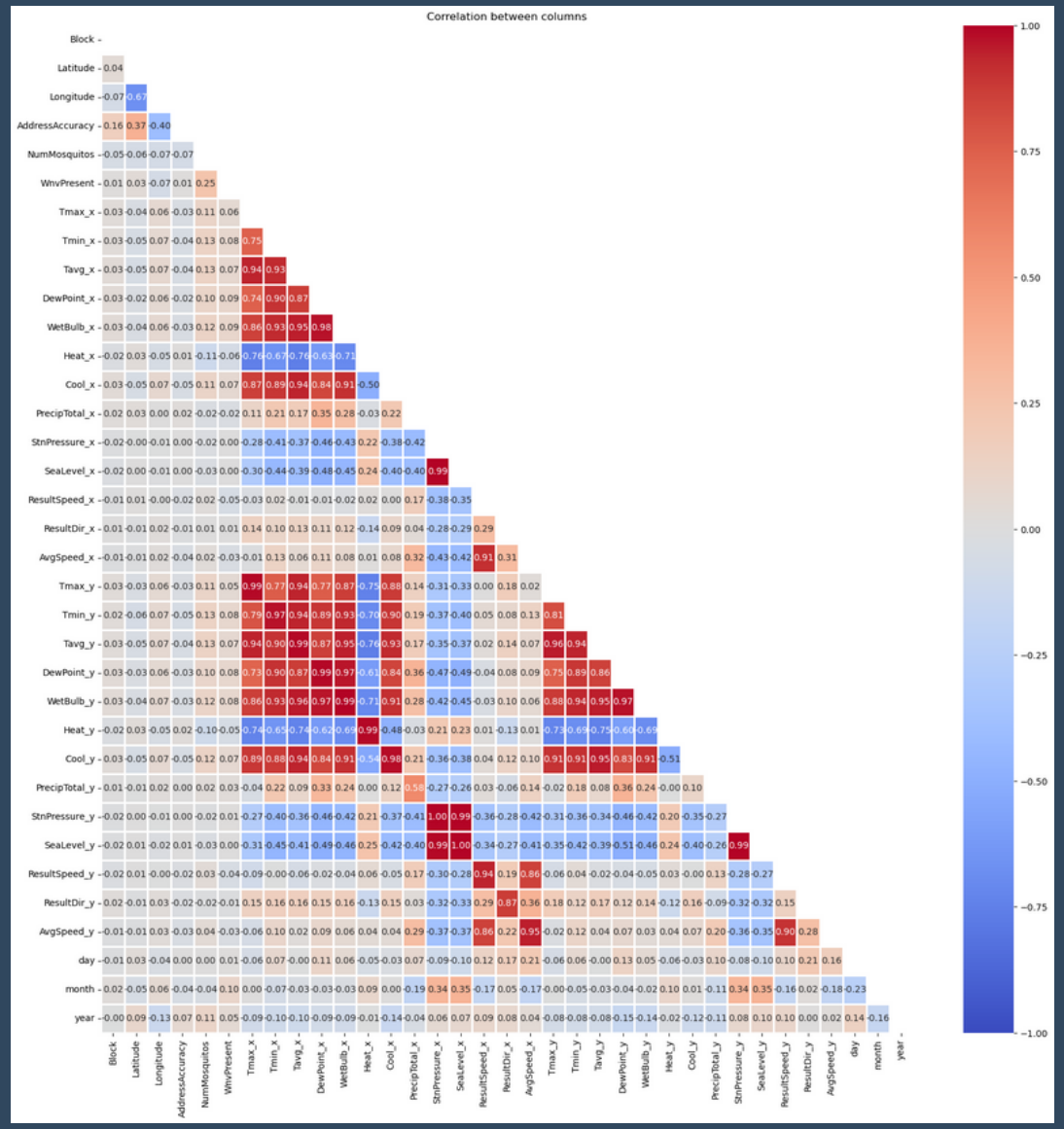
- Train data - date, location, present of virus
- Test data - date, location to be used for prediction
- Weather data - temperature, sunrise, wetbulb
- Pesticides sprayed data - date, location when sprayed

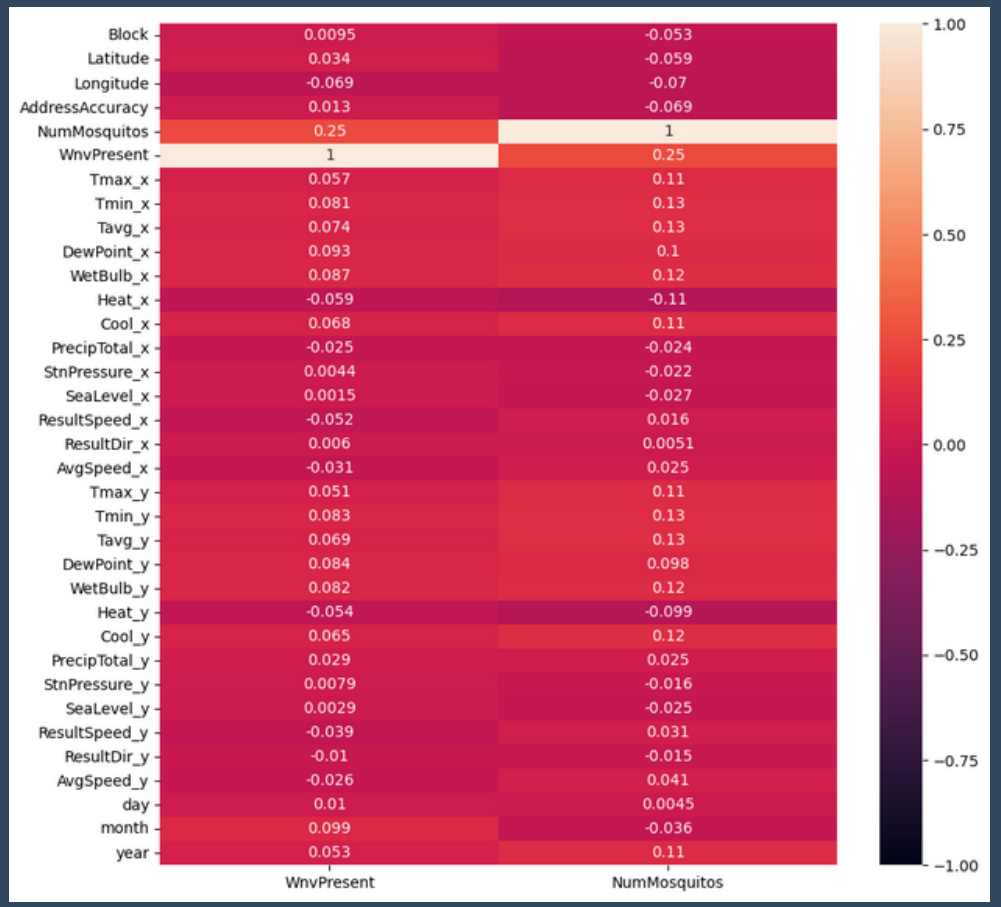
Cleaning

- 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





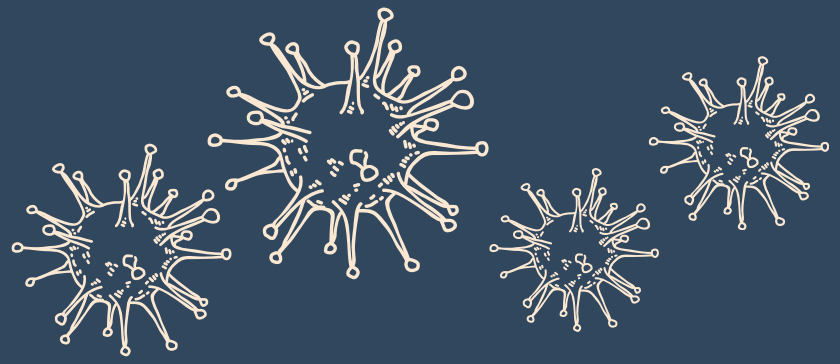
Correlation:

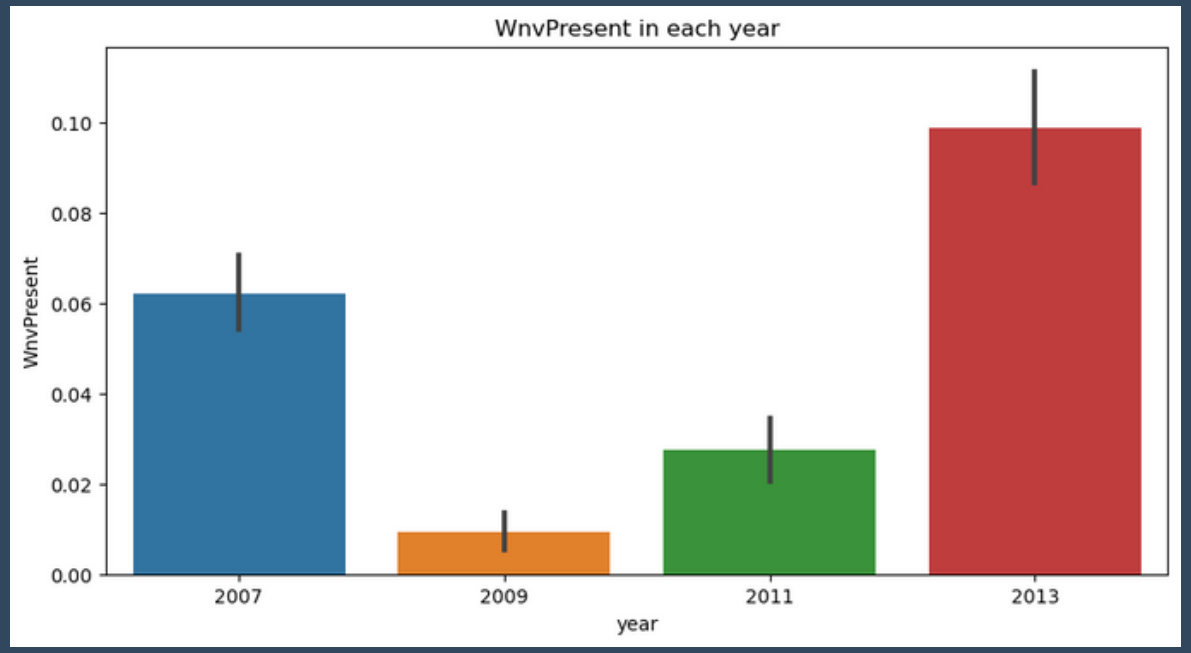
The only significant correlations (>0.5) are among weather data. The correlation toward present of virus and number of mosquito are very low. Even between those 2 are 0.25, which is the highest

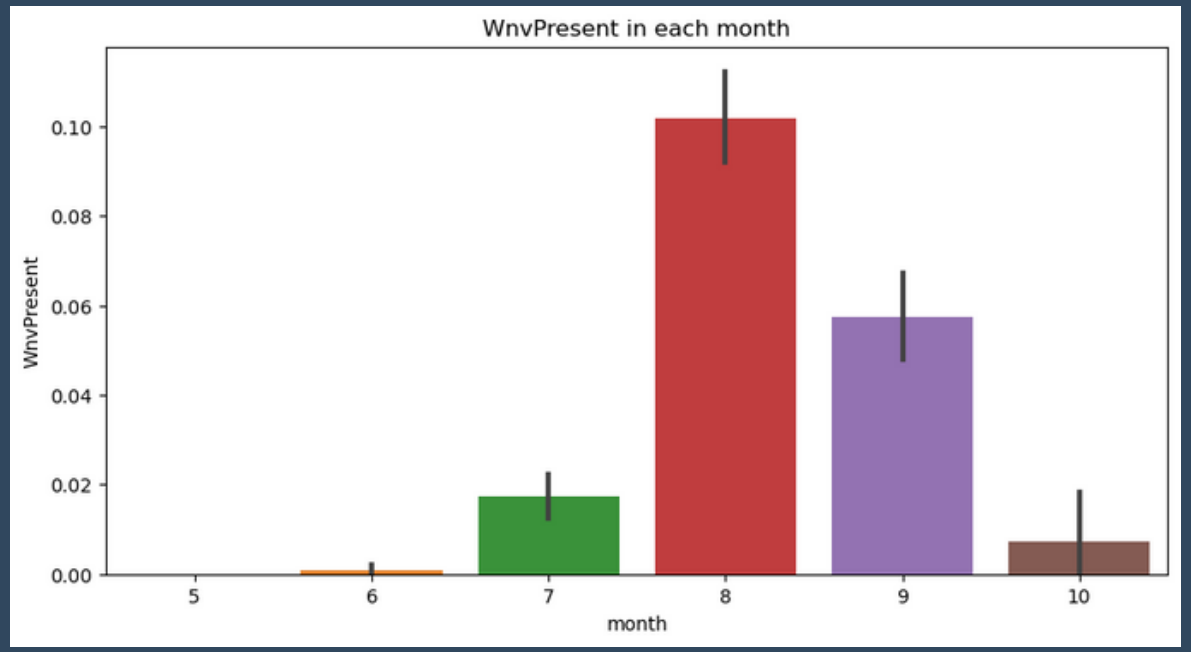
EDA..cont

Train (and test) dataset:

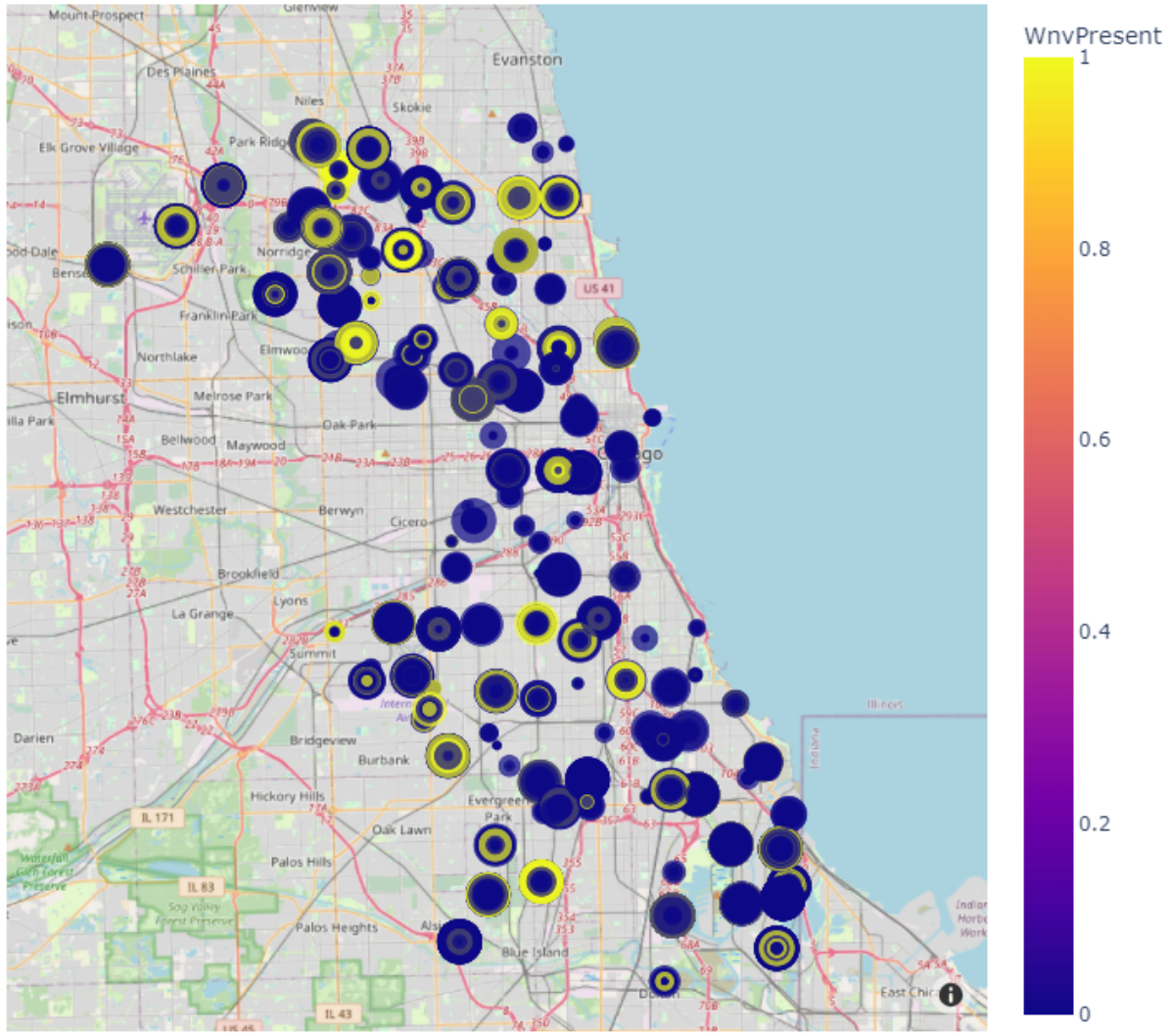
- Virus present clearly during summer
- Virus found more in 2013
- There are some location that virus found more than others, but it looks like it spread all over the city (see next page)

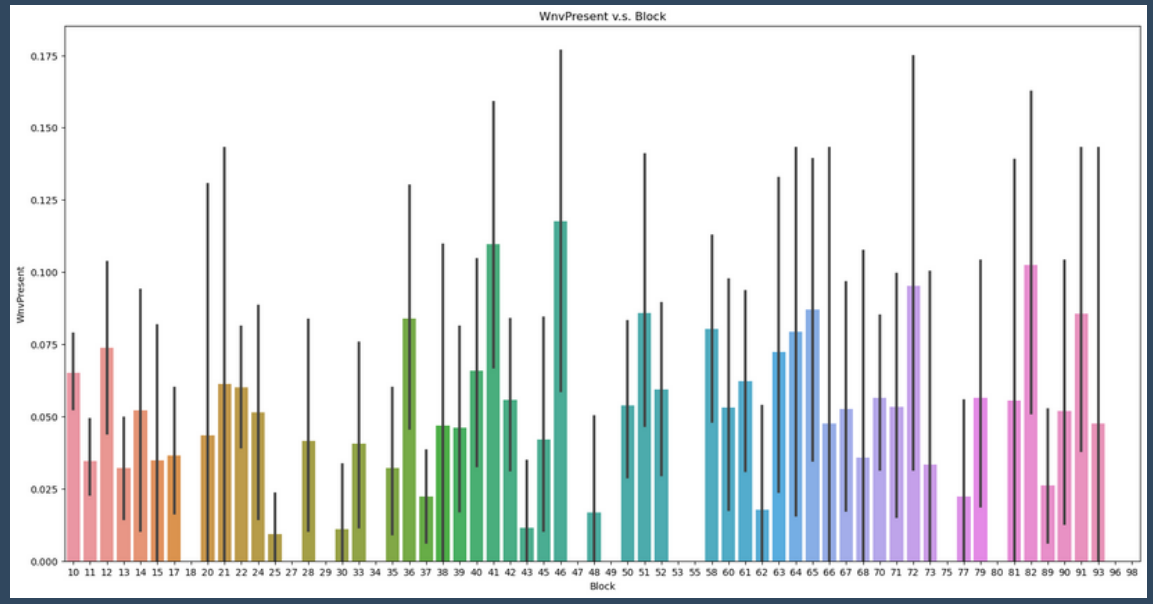






EDA..cont





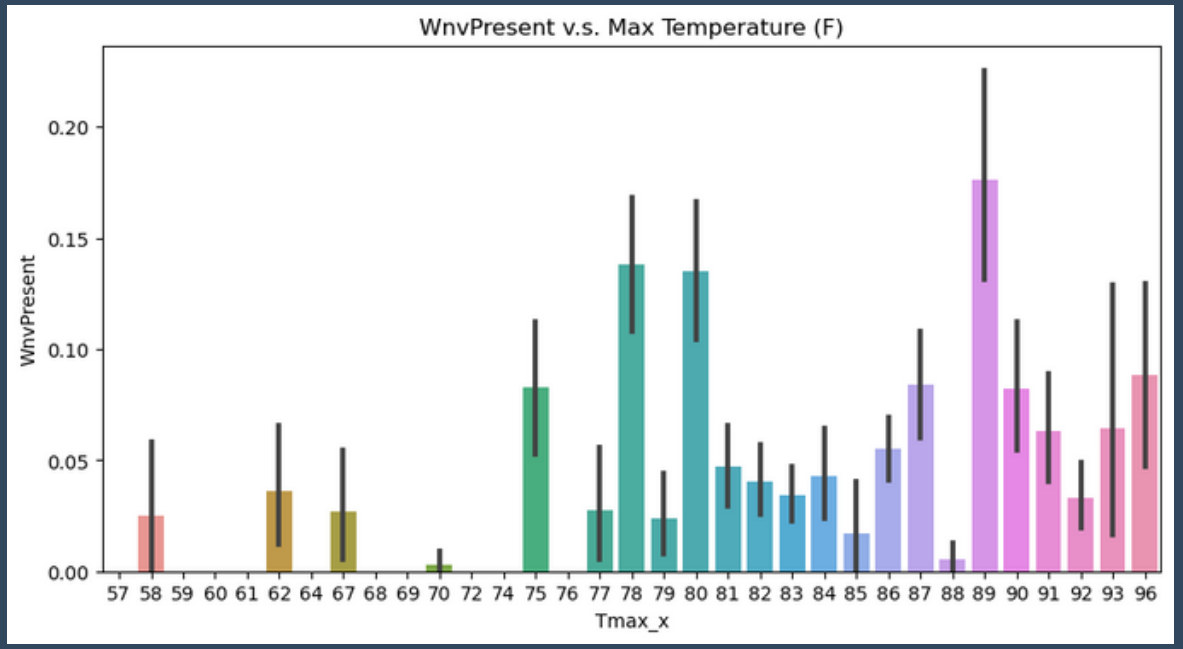
Train (and test) dataset (cont.):

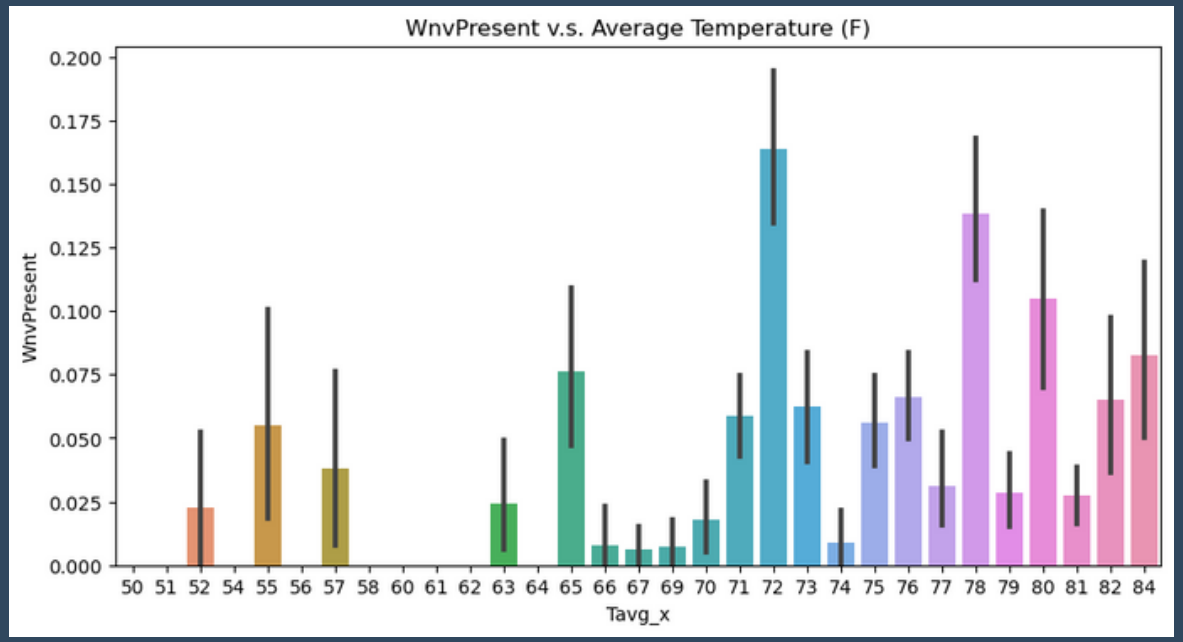
- There are some location that virus found more than others, but it looks like it spread all over the city

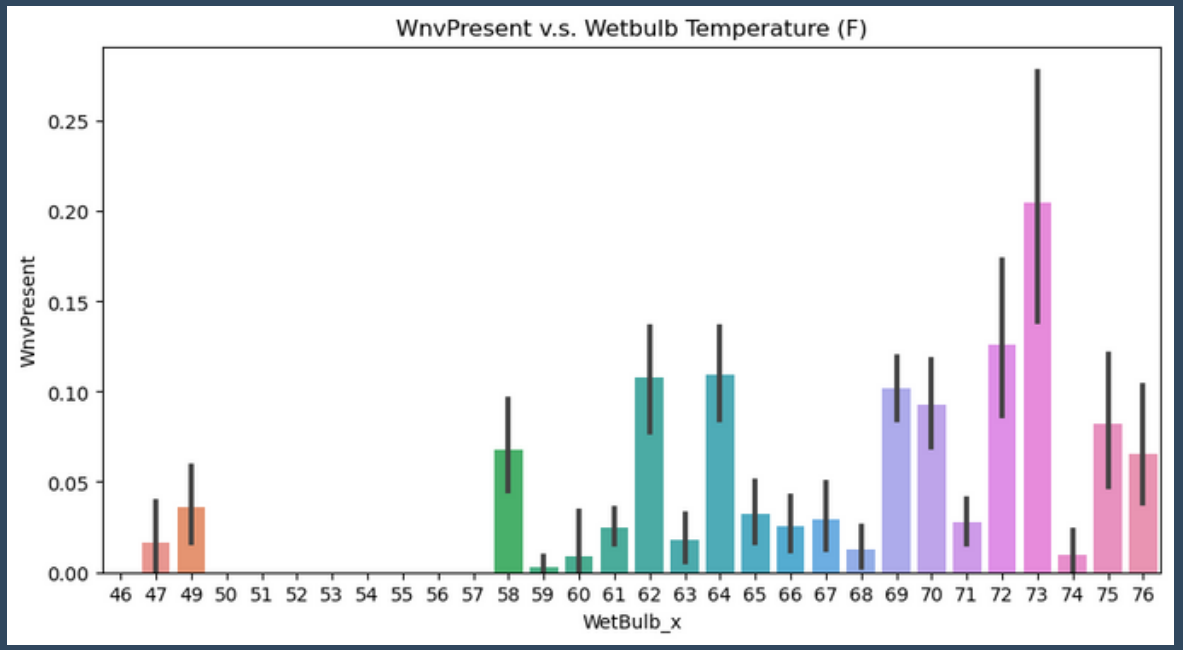
EDA..cont

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

Train data:

- Address and street address are duplicate to the lat/long and even block/trap, so consider to drop them
- Some feature with text present such spicies and traps are not number, so use hotcoded to convert them to number so the model can work with these features
- now we have 36 features (and 1 classification) to work with Model development

Modeling..cont

This is classification case, predicting the probability of label. So our group start with using 8 classification model as follow. Aiming to get the best result score to use them for futher optimization. Gradient Boosting got the highest test score, but when submitted to Kaggle XGBoost provided the best score

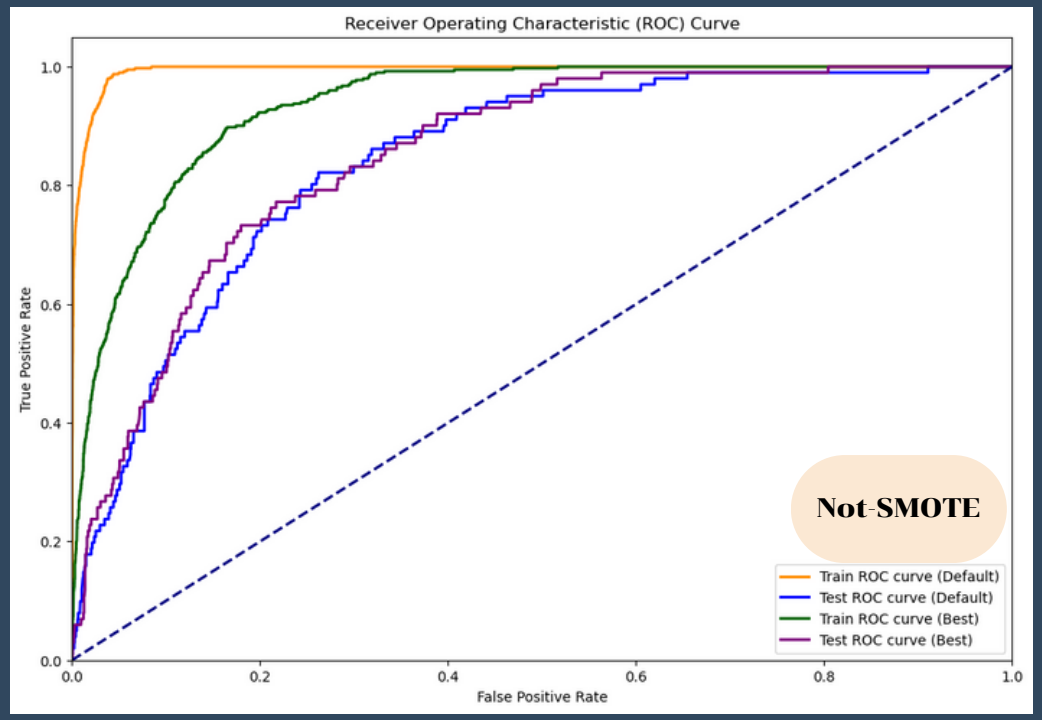
1. Logistic Regression
2. Decision Tree
3. Random Forest
4. AdaBoost
5. Gradient Boosting
6. Bagged Decision Tree
7. SVM
8. XGBoost

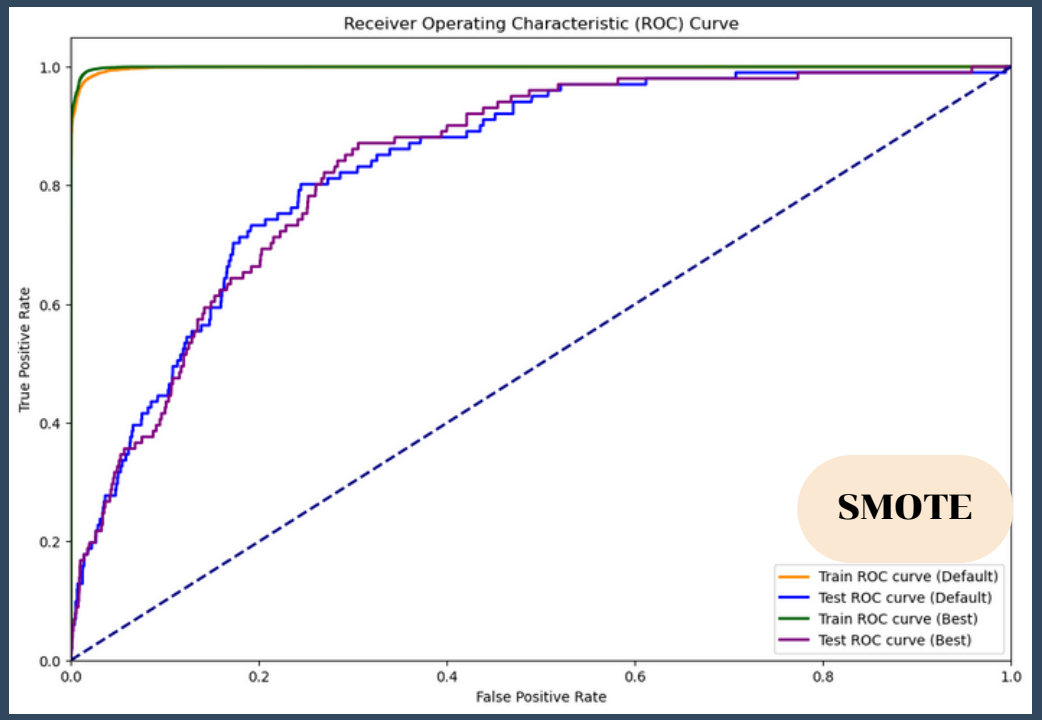
```
Evaluating Logistic Regression...
Logistic Regression - Train ROC AUC Score: 0.82
Logistic Regression - Test ROC AUC Score: 0.81

Evaluating Decision Tree...
Decision Tree - Train ROC AUC Score: 1.00
Decision Tree - Test ROC AUC Score: 0.61

Evaluating Random Forest...
Random Forest - Train ROC AUC Score: 0.99
Random Forest - Test ROC AUC Score: 0.77

Evaluating AdaBoost...
AdaBoost - Train ROC AUC Score: 0.87
AdaBoost - Test ROC AUC Score: 0.83
```

```
Evaluating Gradient Boosting...
Gradient Boosting - Train ROC AUC Score: 0.90
Gradient Boosting - Test ROC AUC Score: 0.86

Evaluating Bagged Decision Tree...
Bagged Decision Tree - Train ROC AUC Score: 0.99
Bagged Decision Tree - Test ROC AUC Score: 0.73

Evaluating SVM...
SVM - Train ROC AUC Score: 0.44
SVM - Test ROC AUC Score: 0.44

Evaluating XGBoost...
XGBoost - Train ROC AUC Score: 0.99
XGBoost - Test ROC AUC Score: 0.84
```

Modeling..cont

We selected XGBoost model to perform hyperparameter tuning, aiming to increase the prediction score. Also, since the train data has very unbalance baseline. We also recreate resample using SMOTE and build model parallel to initial train data to see which is doing better. The AUC-ROC for test split data is not 0.85, just a little bit increase from before.





Error Analysis

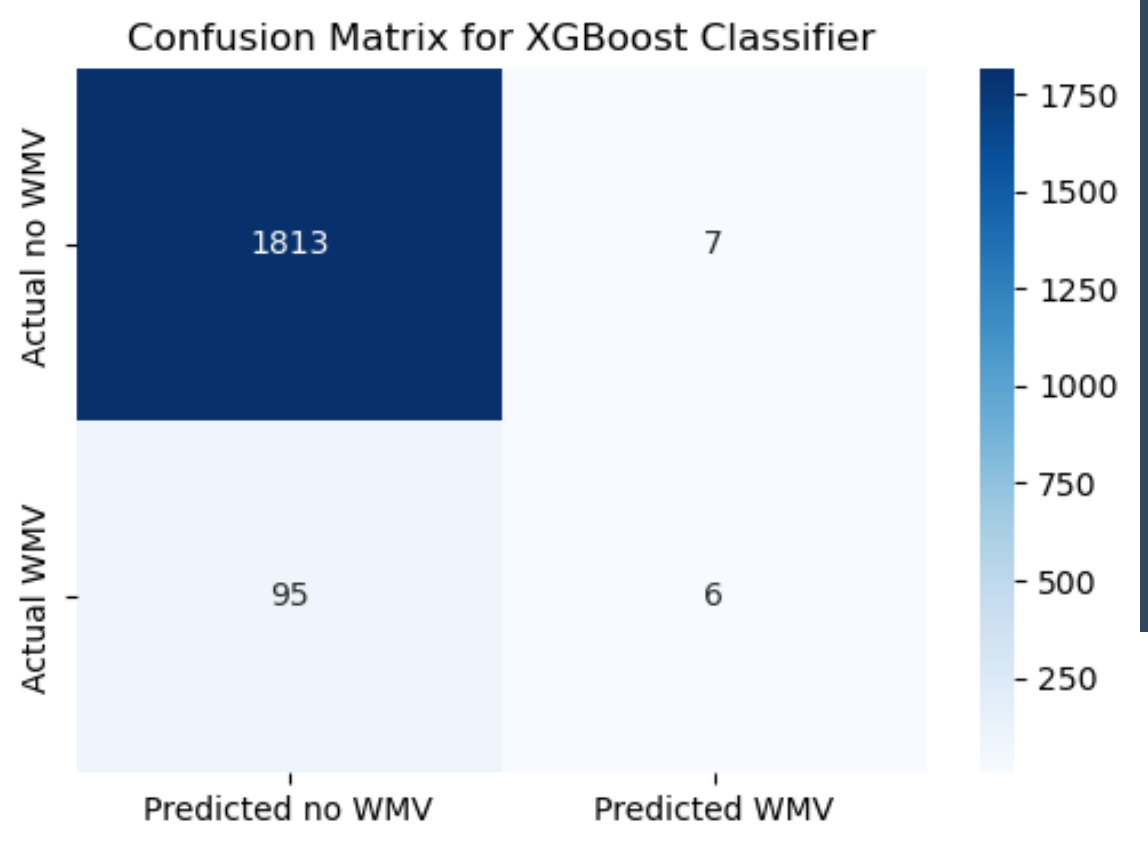
The SMOTE has a better resul. And we use both models to predict and submit to kaggle. Surprisingly non-SMOTE gave a better score on Kaggle at 0.7068 and SMOTE model is 0.6778.

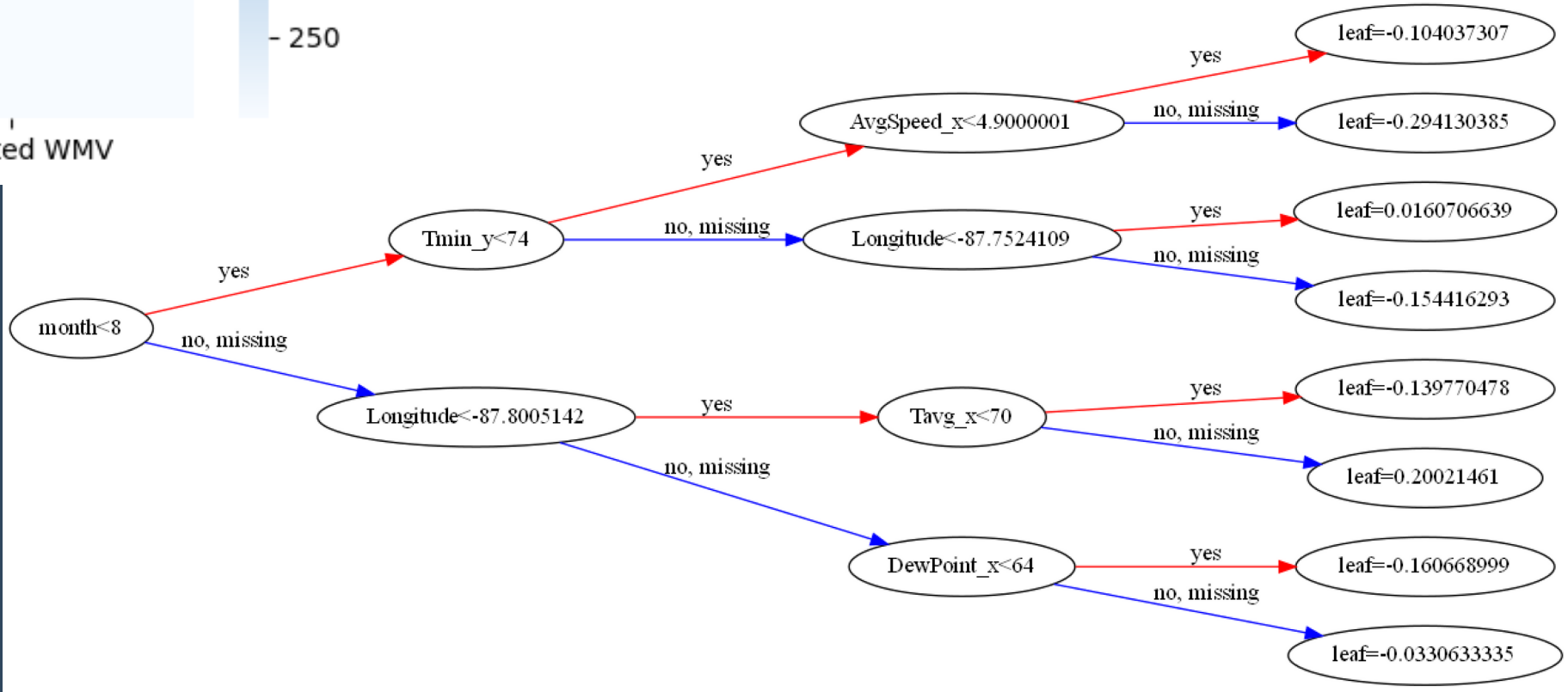
To find out what kind of prediction our model has done, we use the splited test data to check the confusion matrix and the tree

Sensitivity (TP/TP+FN) is not good at all, only 5%. Due to baseline, model is likely to predict mostly no WMV. This indicates the True Positive over Total actual positive case is very poor.

Then let's check the tree. We can see that this is align with our assumption earlier that weather might have correlation, because the tree shows in the early branch of tree.

Error Analysis..(cont.)



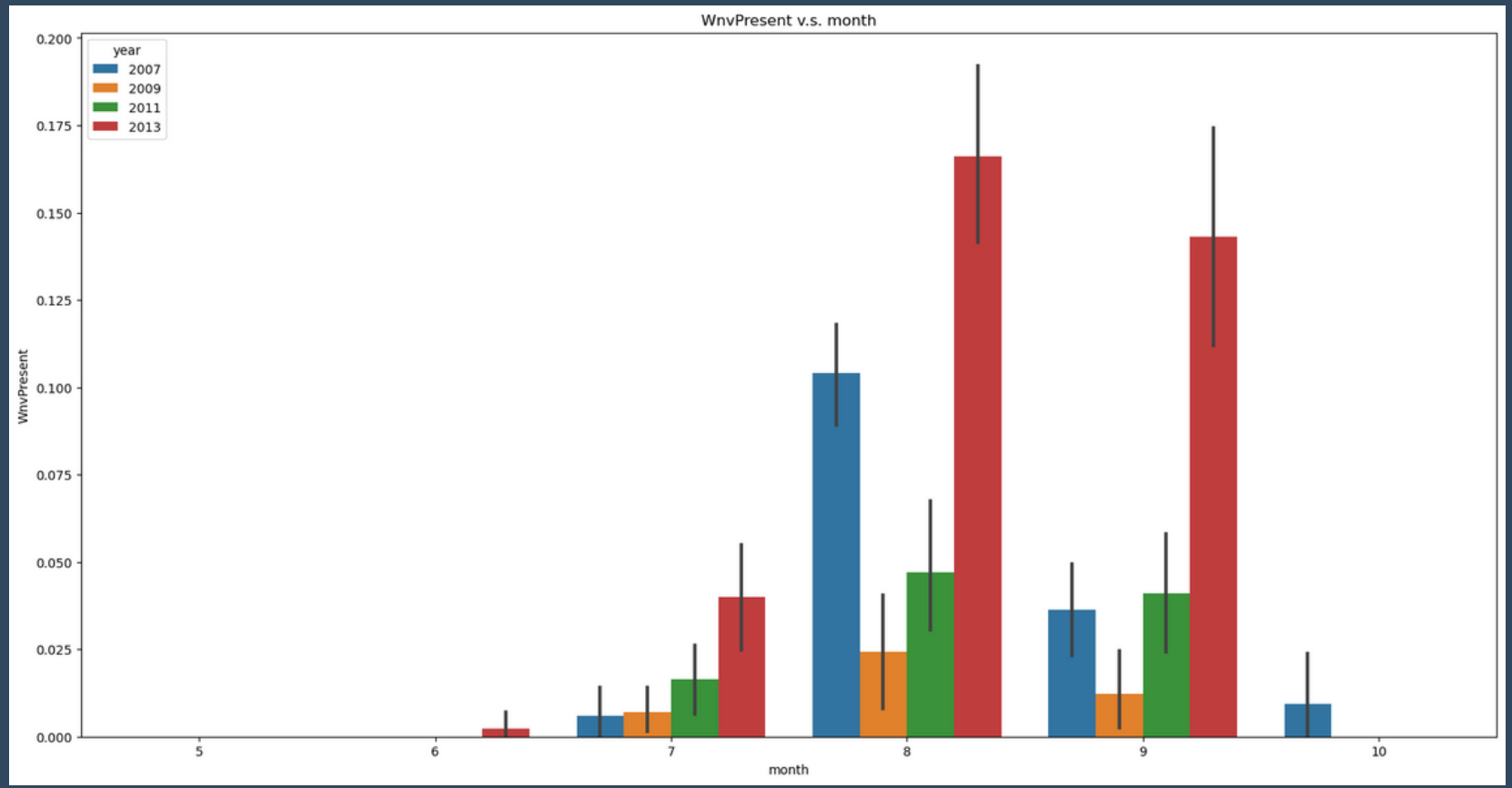


Feature Engineering

We haven't touch spray data, also from the error analysis we found that there could be a correlation between the temperature and present of virus. so let got back and check on those feature, and perform features engineering aiming to improvde the model performance.

Spray was done in 2011 and 2013, but from the data we have. 2013 has the highest virus present. This looks like pesticides spray wasn't effective, so let's leave this out as usual and focus on weather data.

We ending up created 34 more features from creating the bin for every 8 degree F of Tmax, Tmin, Tavg, and Wetbulb



Modeling after features optimization

After using new features data, the result of the model is improved. AUC-ROC of both splited train and test data of SMOTE went to 0.6. and When we made prediction and submitted to Kaggle, the result also improved. Again the non-SMOTE actually got better score than SMOTE

submission_smote2.csv Complete (after deadline) · 11h ago · smote 2	0.69897	0.72101
submission2.csv Complete (after deadline) · 11h ago · new features	0.70681	0.72142

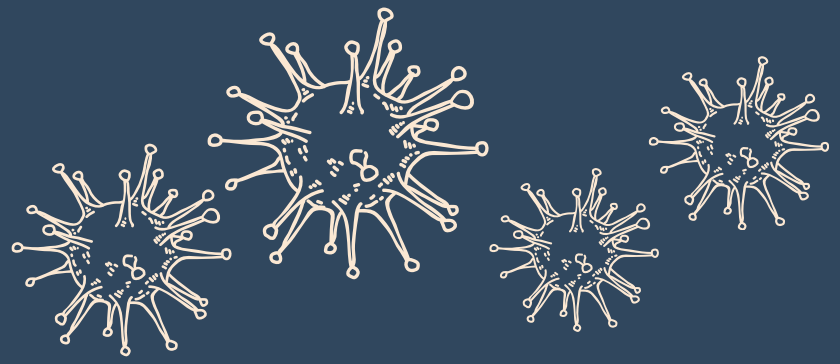
Cost-Benefit

Report Reduce Repel

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.

From our result, we found that the factor that has impact on the present of the virus is likely to be weather, or to be more presice the temperature. When we focus our feature engineer on weather factor, we got a better result. We would recommend that if the City is to plan pesticide spraying, they should concentrate to do it during summer and perhaps change the pesticide, because from spray data, it doesn't look very effective.

Limitation

We concern about the impact of pesticide spray that it not effective. But we're not sure yet, because it might be that spray was done further from Trap, so we didn't detect any effect of spray. Perhaps more data on spray would prove otherwise. Also the train data was strongly imbalance, where the test baseline shows very well balance case.

Conclusion