West Nile Virus Analysis

Felicia | Simran | Sid
Team #4

Background

West Nile Virus (WNV)

- Leading cause of mosquito-borne disease in continental United States
- 20% infected people develop severe symptoms
- Can only spread from mosquito → human

Chicago

- First human cases of WNV reported in 2002
- Established a comprehensive surveillance and control program by Chicago Department of Public Health (CDPH) by 2004
- Test mosquitos in traps across the city every week (late spring through the fall)

Problem Statement

Due to the recent epidemic of West Nile Virus in the Windy City, the data science team at Disease And Treatment Agency was tasked to derive an effective plan to deploy pesticides throughout the city.

Using various location, weather conditions and time lags, we will be analysing classification techniques to obtain the best model that can predict the presence of WNV across Chicago.

Workflow

Data Exploration

Collect and explore data from Kaggle

1

2

Data Cleaning & Analysis

Perform cleaning and visualizations

Modelling

Preprocess and build models. GridSearch for hyperparameter tuning

3

4

Evaluation

Results & discussion

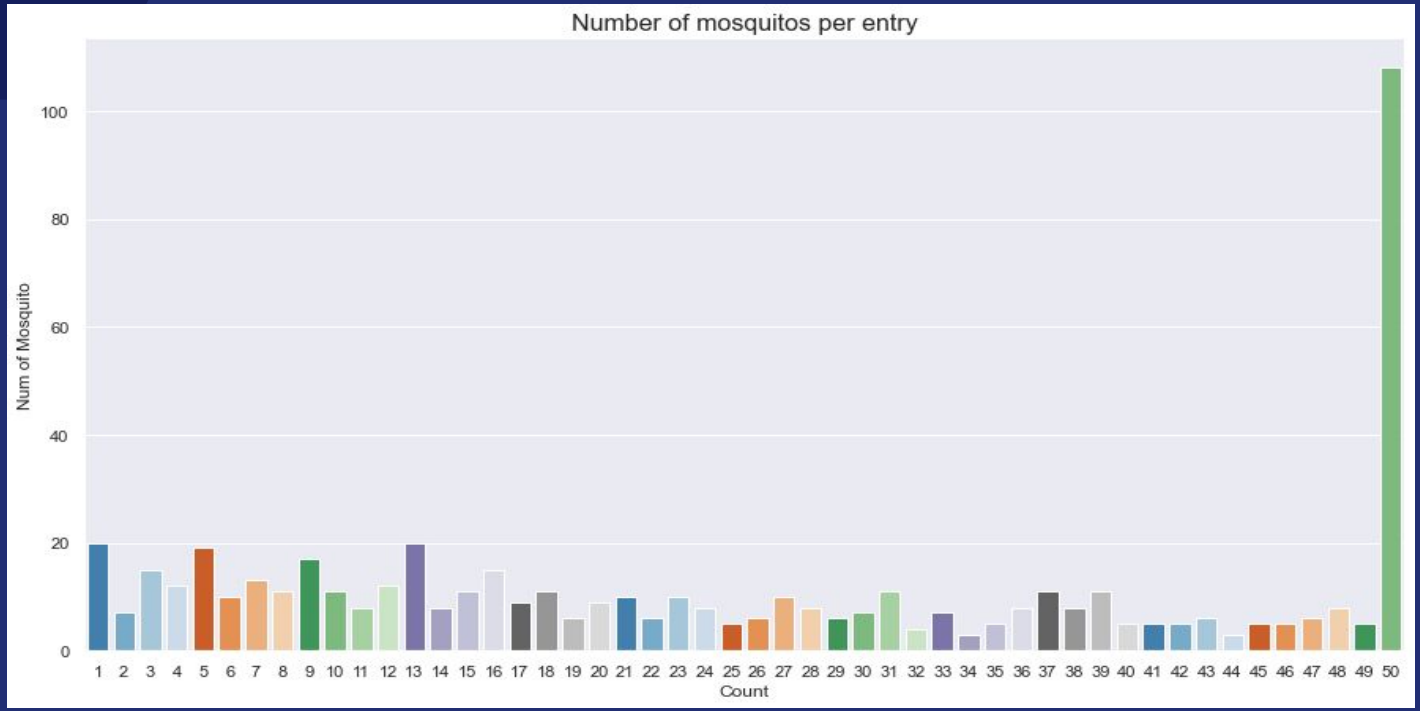
Conclusion

Final thoughts & remarks

5

Data Exploration

Train Data Quality



Train data

Number of Mosquitos and Presence of Virus

→ Each row is capped at 50
→ Can be conflicting to train the model

	date	species	trap	latitude	longitude	num_mosquitos	wnv_present
4888	2009-07-24	CULEX PIPIENS/RESTUANS	T002	41.95469	-87.800991	50	0
4889	2009-07-24	CULEX PIPIENS/RESTUANS	T002	41.95469	-87.800991	25	0
4890	2009-07-24	CULEX PIPIENS/RESTUANS	T002	41.95469	-87.800991	50	1
4891	2009-07-24	CULEX PIPIENS/RESTUANS	T002	41.95469	-87.800991	50	0
4892	2009-07-24	CULEX PIPIENS/RESTUANS	T002	41.95469	-87.800991	40	0
4893	2009-07-24	CULEX RESTUANS	T002	41.95469	-87.800991	18	0
4894	2009-07-24	CULEX PIPIENS	T002	41.95469	-87.800991	4	0

Weather Data

Although there are no null values, they are represented differently as stated in the documentation.

'M' = missing values (for e.g. in Tavg column)	' ' = moderate (for CodeSum column)
'-' = missing values (for e.g. in Sunrise column)	'T' = trace values (for e.g. PrecipTotal column)

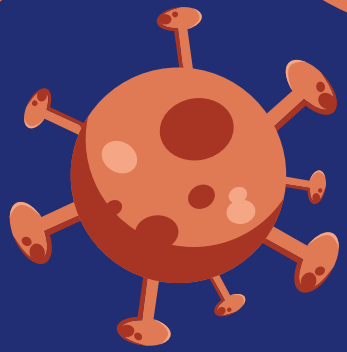
Daylight
→ Convert sunrise and sunset into daytime in mins

Relative Humidity
→ Moisture content in the atmosphere, at constant temperature and pressure.

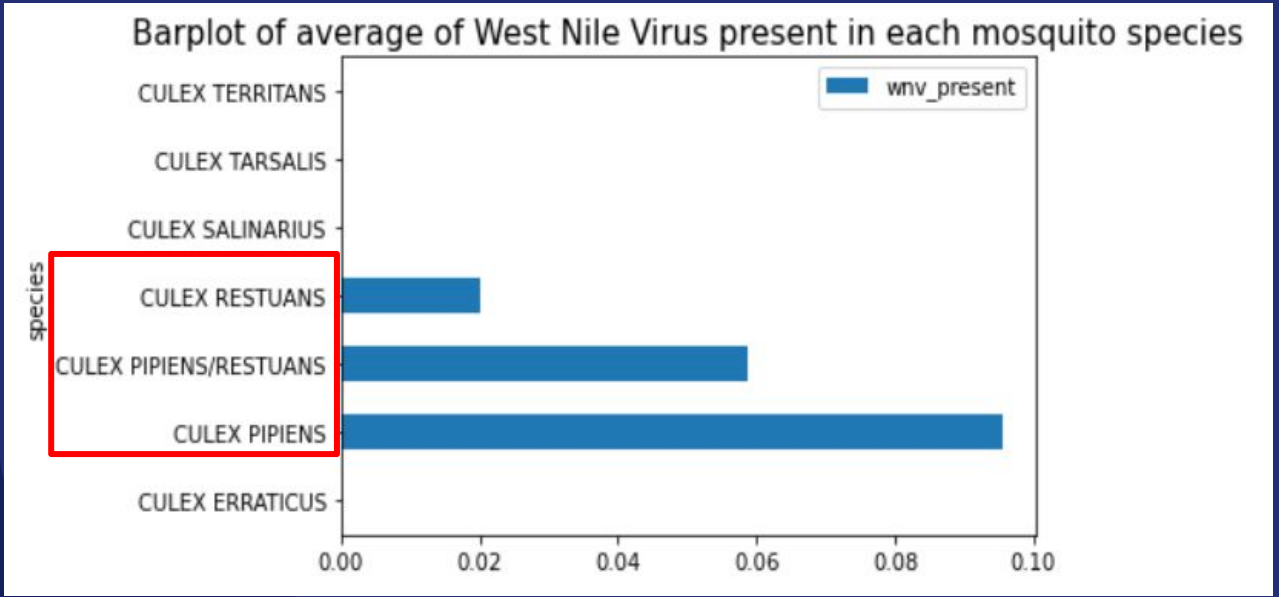
Average of Stations
→ Both stations are close to each other → Merge station data

Exploratory Data Analysis

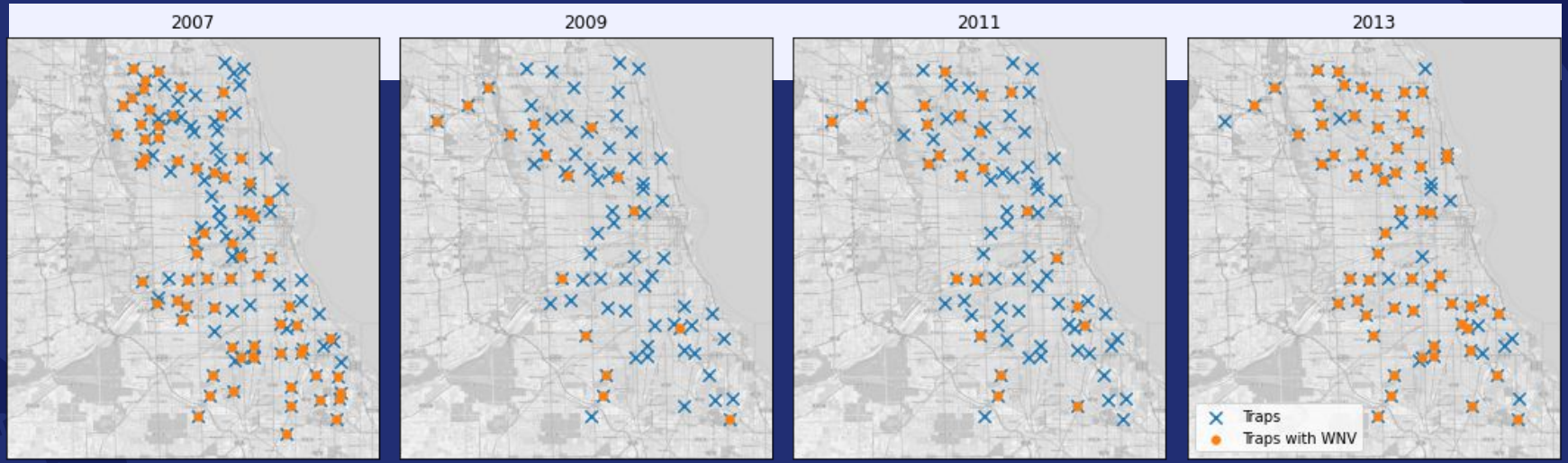
Initial investigation of data



Mosquito species that spread WNV



YoY WNV presence



Feature highlights from EDA

Selected features

1. Location (longitude and latitude)
2. Average temperature (lag 28)
3. Daylight
4. Week
5. Year
6. Species (one hot encoded)
7. Relative humidity (lag5)
8. Precipitation (lag 14)

Rejected features

1. **Number of mosquitoes**
2. **Trap**
3. Snow fall
4. Water1
5. Sunrise-sunset

Rolling features

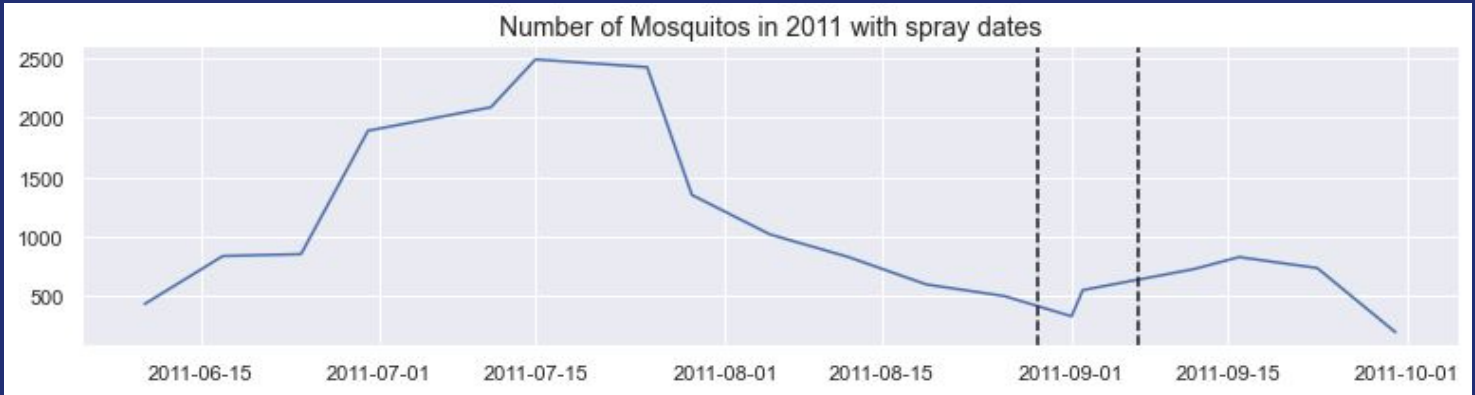
We will use rolling average for the following features:

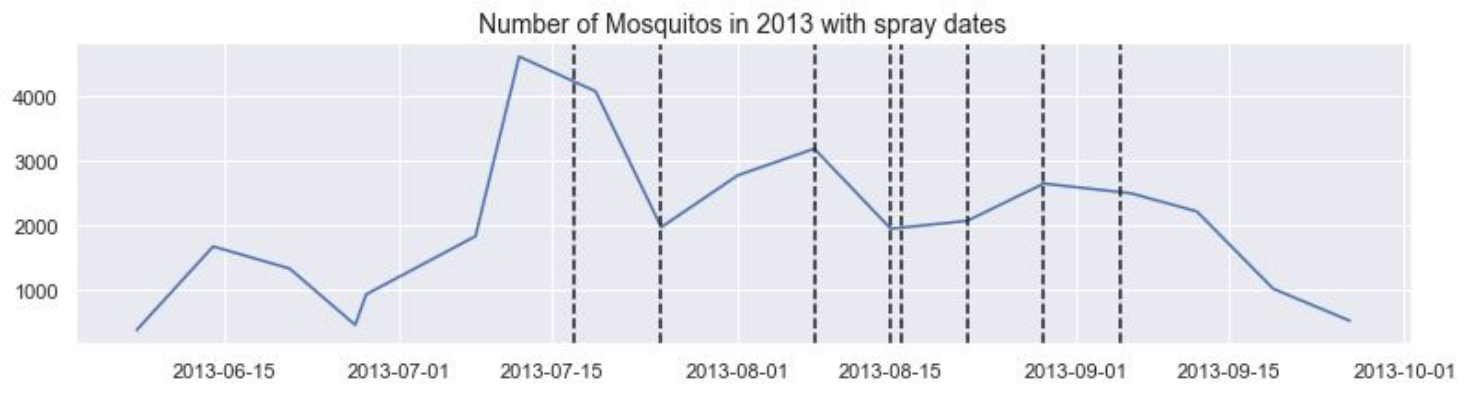
- 'tavg'
- 'precip_total'
- 'r_humid'

E.g. Breeding cycles are significantly affected by small changes in temperature.

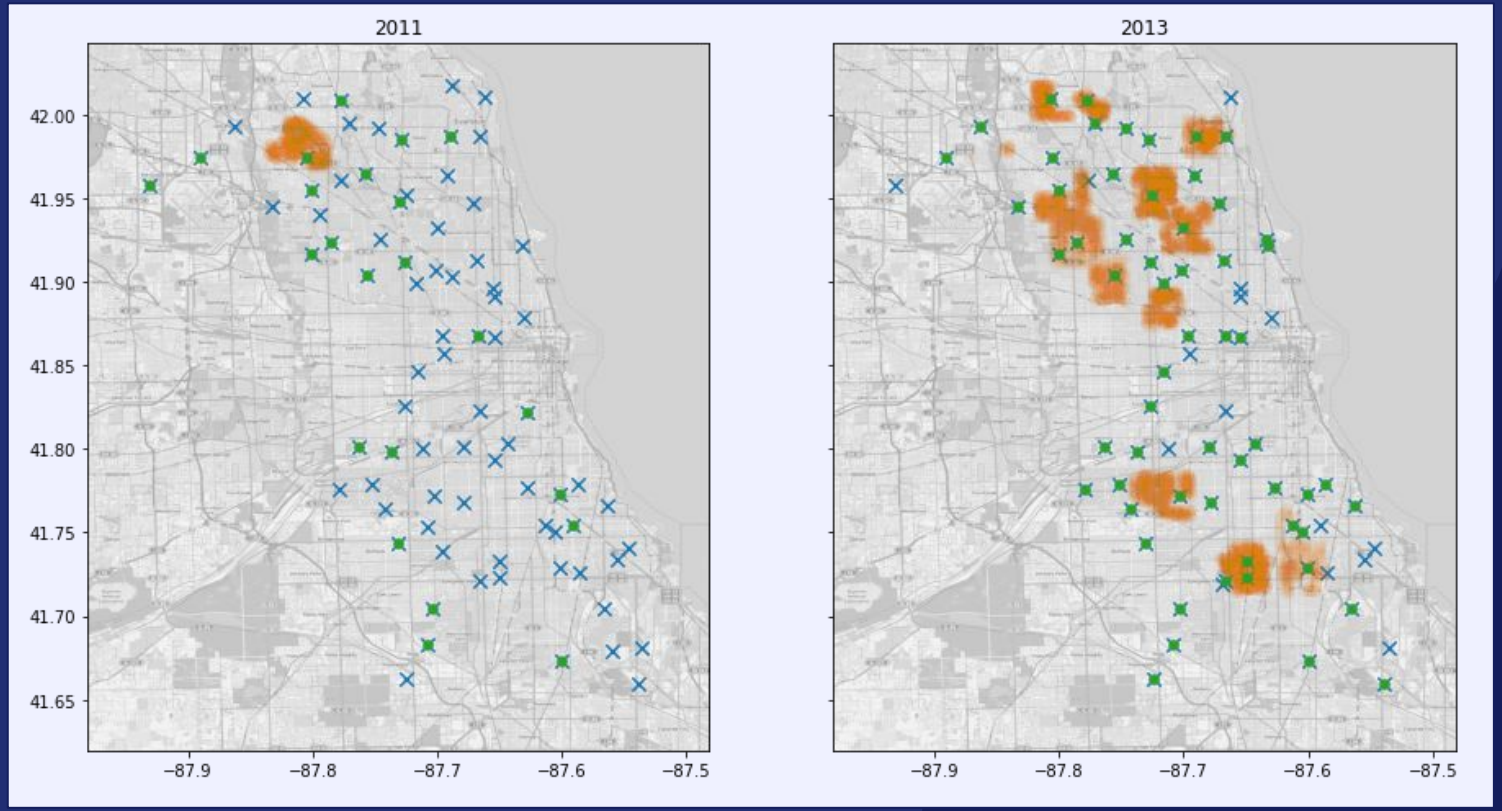
Mosquito Species	Temperature	Lifecycle (days)
CULEX	70° F	14
CULEX	80° F	10

Effect of spray on number of mosquitos

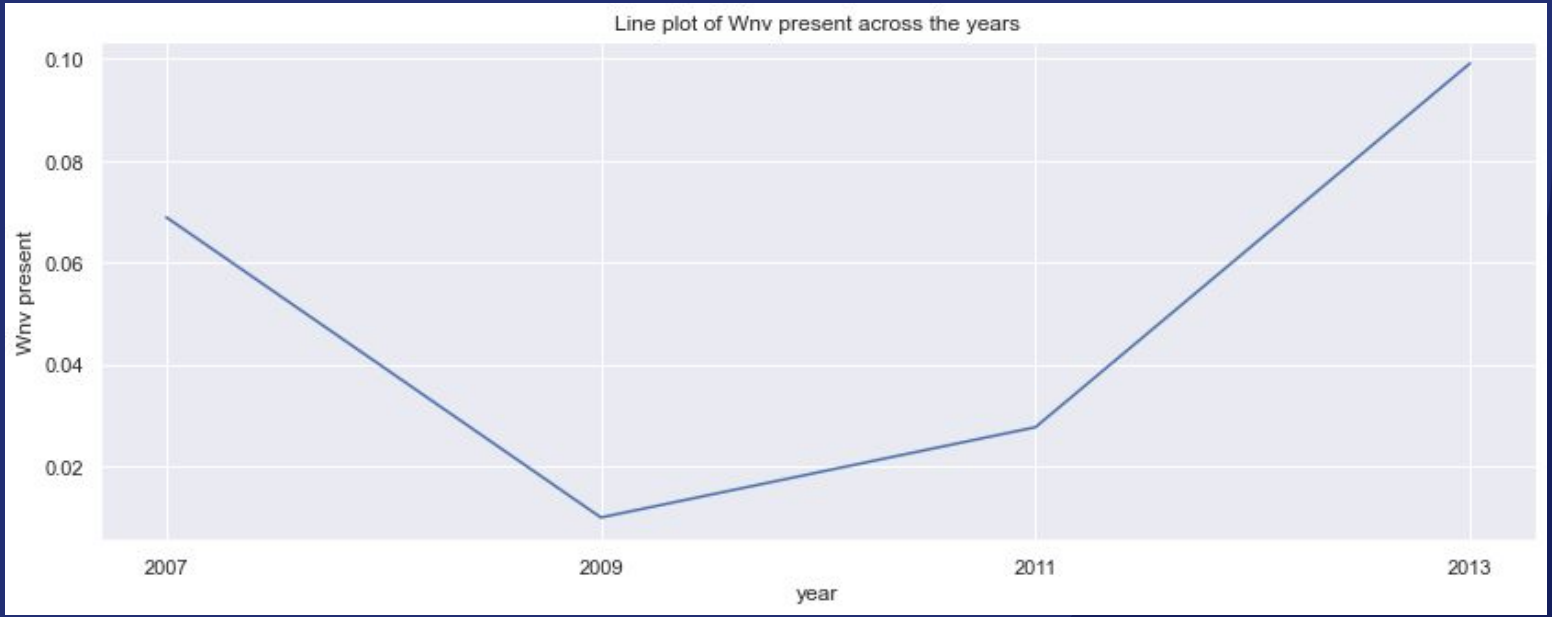




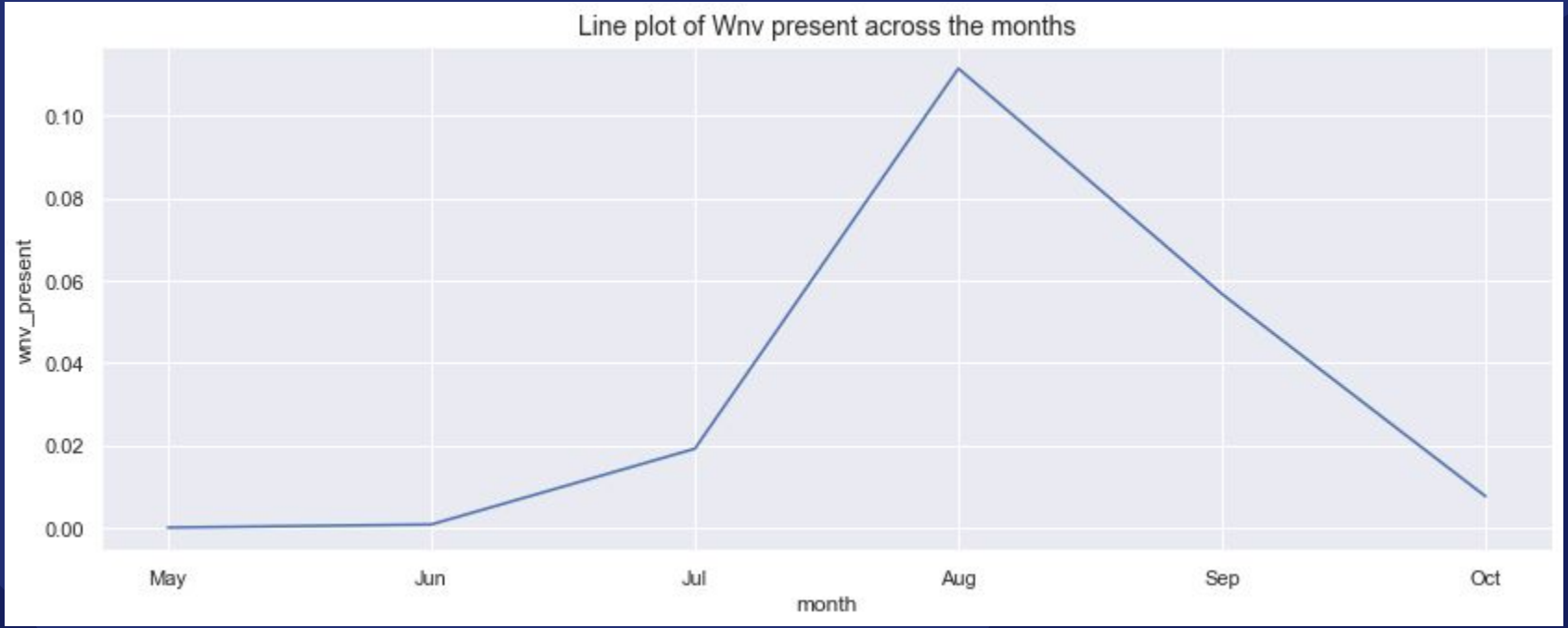
Spray locations in 2011 & 2013



Trend over the years



Seasonality in virus presence

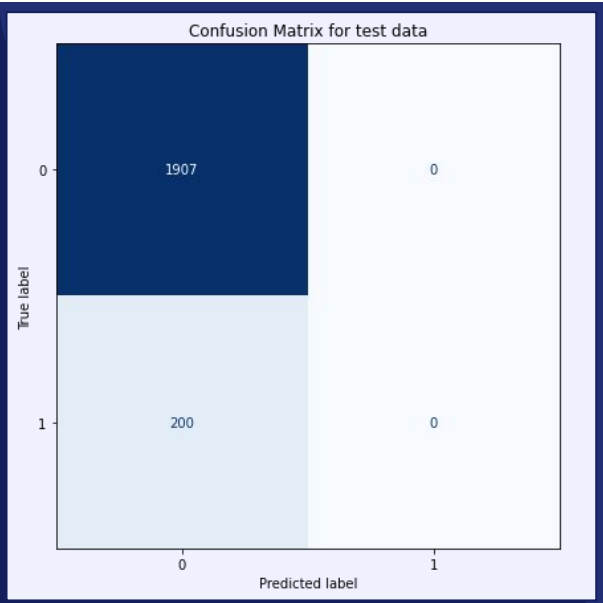


Model Evaluation

Target distribution

```
# target distribution
y_train.value_counts(normalize=True)
```

```
0    0.9595
1    0.0405
Name: wnv_present, dtype: float64
```



- 96 - 4 | highly imbalance class
- Training this results in very poor performance
- Model predicted 0 instances of the positive class correctly

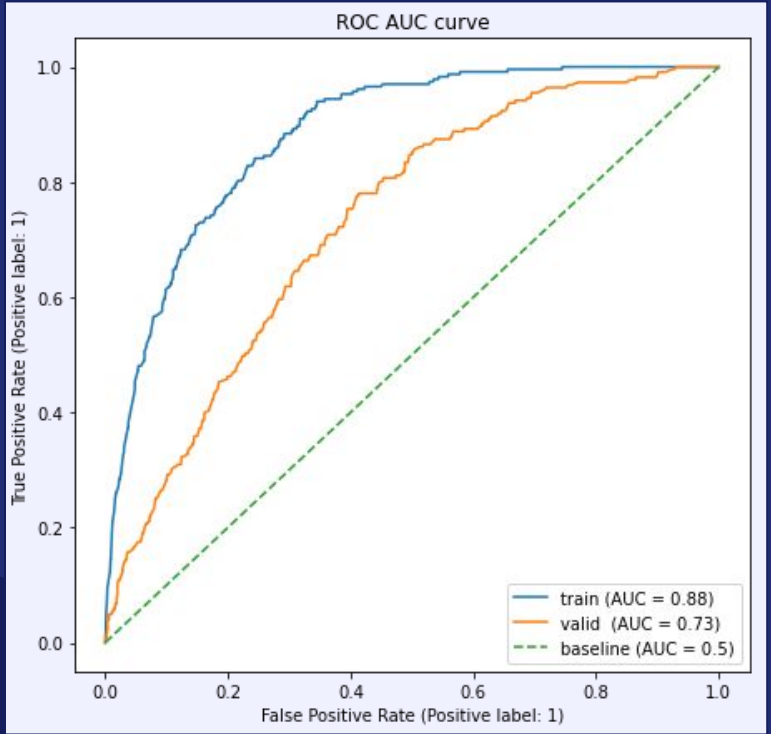
Solution?

- Oversample using SMOTE
- Synthetic Minority Oversampling TEchnique
- Post-operation distribution 50 - 50

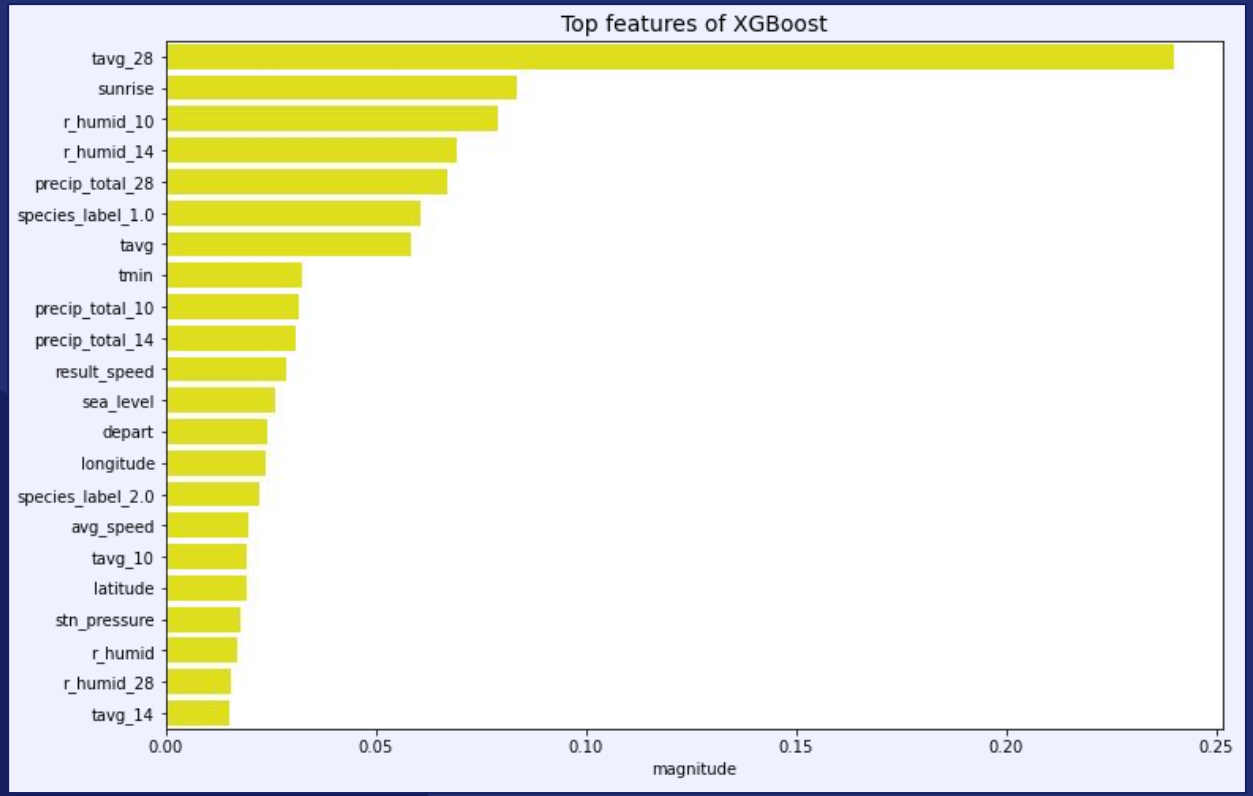
Results

	Classifier	Train Acc Score	Val Acc Score	Train ROC-AUC	Val ROC-AUC	Recall	Precision	F1-Score
0	xgb	0.741312	0.792803	0.882247	0.729746	0.408072	0.188406	0.257790
1	et	0.710290	0.865164	0.848968	0.721504	0.147982	0.179348	0.162162
2	ab	0.708256	0.779359	0.848738	0.706233	0.295964	0.141328	0.191304
3	bc	0.798779	0.429419	0.916910	0.692877	0.905830	0.124384	0.218733
4	lr	0.746228	0.783709	0.842807	0.687969	0.336323	0.158228	0.215208
5	rf	0.676894	0.850534	0.845842	0.680681	0.152466	0.152466	0.152466
6	dt	0.787930	0.827204	0.811114	0.623974	0.233184	0.163522	0.192237

ROC-AUC Curve



Feature importance



Cost Benefit Analysis

According to 'Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005'.

- 163 cases (117 WNF and 46 WNND cases)
- spray area of 477 km²

West Nile fever (WNF):

causing flu-like symptoms, mild compared to WNND

West Nile neuroinvasive disease (WNND):

severe, affecting central nervous system symptom

Summary	
Total medical, productivity, miscellaneous cost	$136,839 (WNF)
	$2,140,409 (WNND)
Total spray and labour cost	$701,790
Total economic cost	$2,979,037

Medical, productivity, miscellaneous cost per pax	$1,170 (WNF)
	$46,531 (WNND)
Spray and labour cost per km²	$1,471

On to Chicago

The average number of WNV present in the traps for train dataset is 138.

Assuming each trap with WNV present could spread to 2-3 people

Cost benefit analysis on 325 WNV cases and 112.3km^2 of spray area in Chicago.

	Sacramento County	Chicago
Population	1.36 million	2.71 million
Area	2574 km^2	606.1 km^2
Spray area	477 km^2	112.3 km^2
Total cases	163	325
Total cost of spray	$701,790	$165,193
Total medical/productivity /miscellaneous cost	$2,277,248	$4,540,535

The Annual Cost Projection for test set:

- Cost will keep increasing due to inflation rate
- Spray in areas with 12 or more WNV cases
- Currently no WNV vaccination available
- The article showed that spraying does help to an extent

Annual Cost Projection			
Year	Medical/productivity/misc cost	Spray cost	Inflation rate(%)
2005	$4,540,535	$165,193	-
2008	$4,965,529	$180,655	9.36
2010	$5,109,033	$185,876	2.89
2012	$5,328,210	$193,850	4.29
2014	$5,483,261	$199,491	2.91

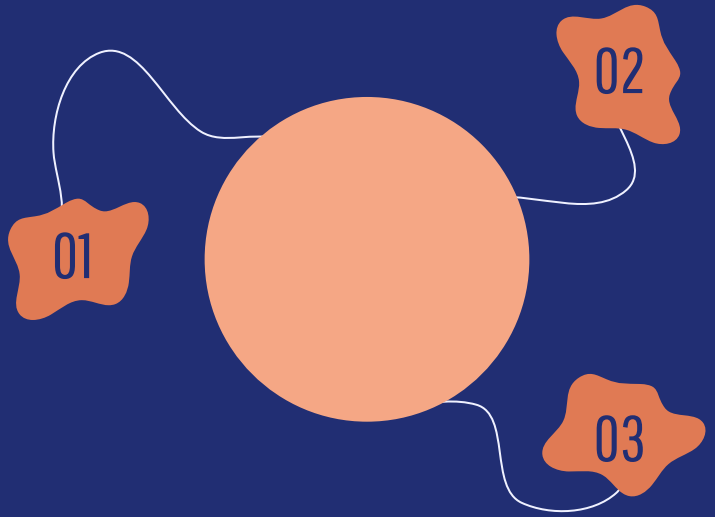
Conclusions

Evaluation

XGBoost best ROC-AUC but resource-intensive

Time of year matters

Weather 2-4 weeks prior to peak



Recommendations

Spraying 2-3 weeks prior to “virus season”

Stagnant water during rainy seasons

Improvements

Detailed feature engineering

Collect/predict number of virus-carrying mosquitoes

THANKS