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
- Has cost US $800 million since 1999

Chicago

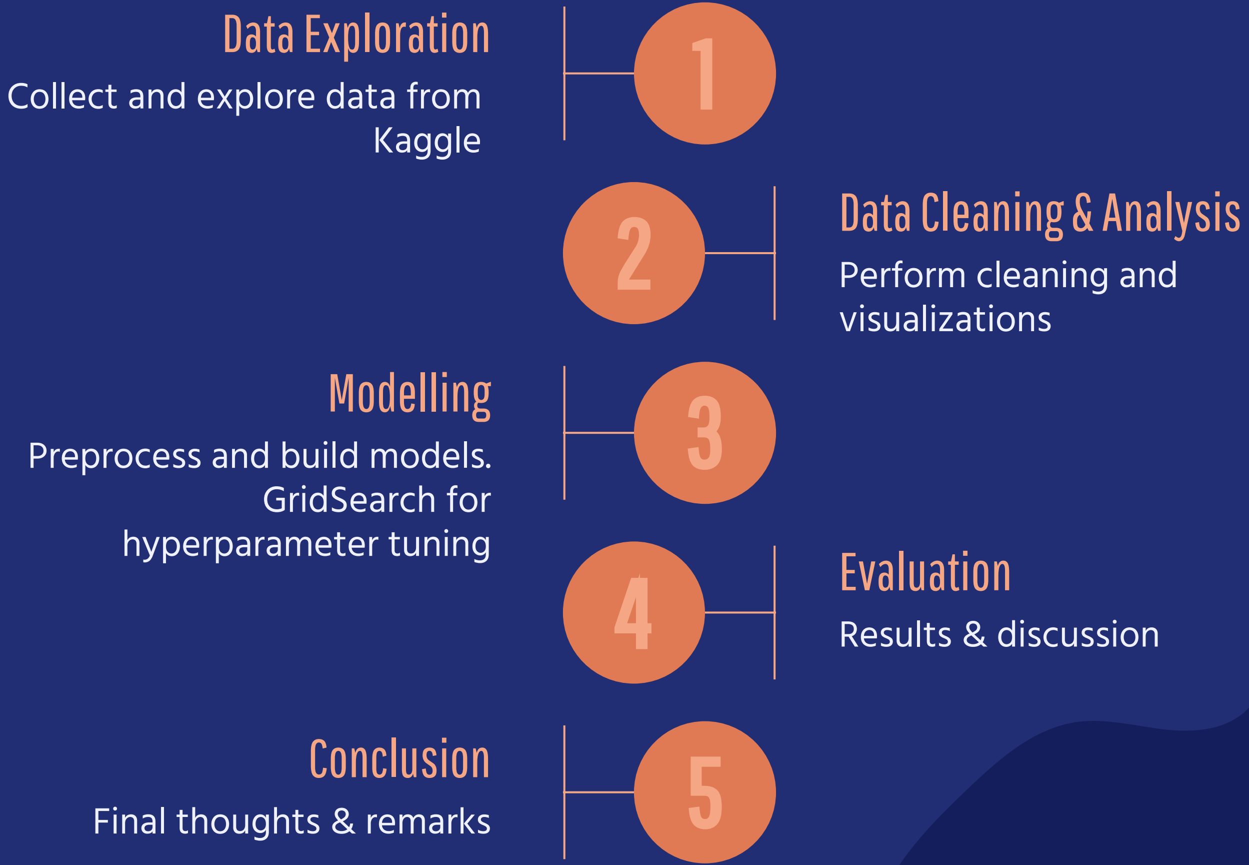
- First human cases of WNV reported in 2002
- Established a comprehensive surveillance and control program by Chicago Department of Public Health (CDPH) by 2004
- Test mosquitos in traps across the city every week (late spring through the fall)

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



1

Data Exploration

Collect and explore data from Kaggle

2

Data Cleaning & Analysis

Perform cleaning and visualizations

3

Modelling

Preprocess and build models. GridSearch for hyperparameter tuning

4

Evaluation

Results & discussion

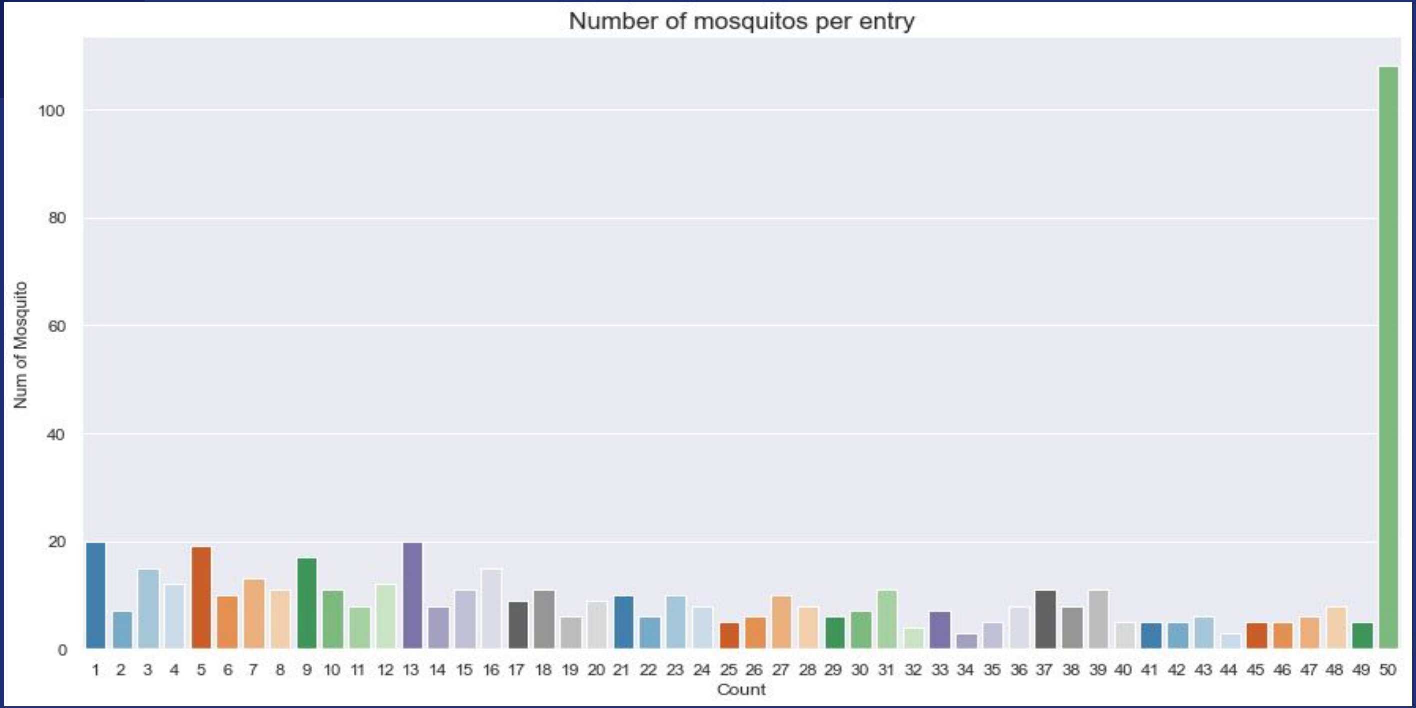
5

Conclusion

Final thoughts & remarks

Data Exploration

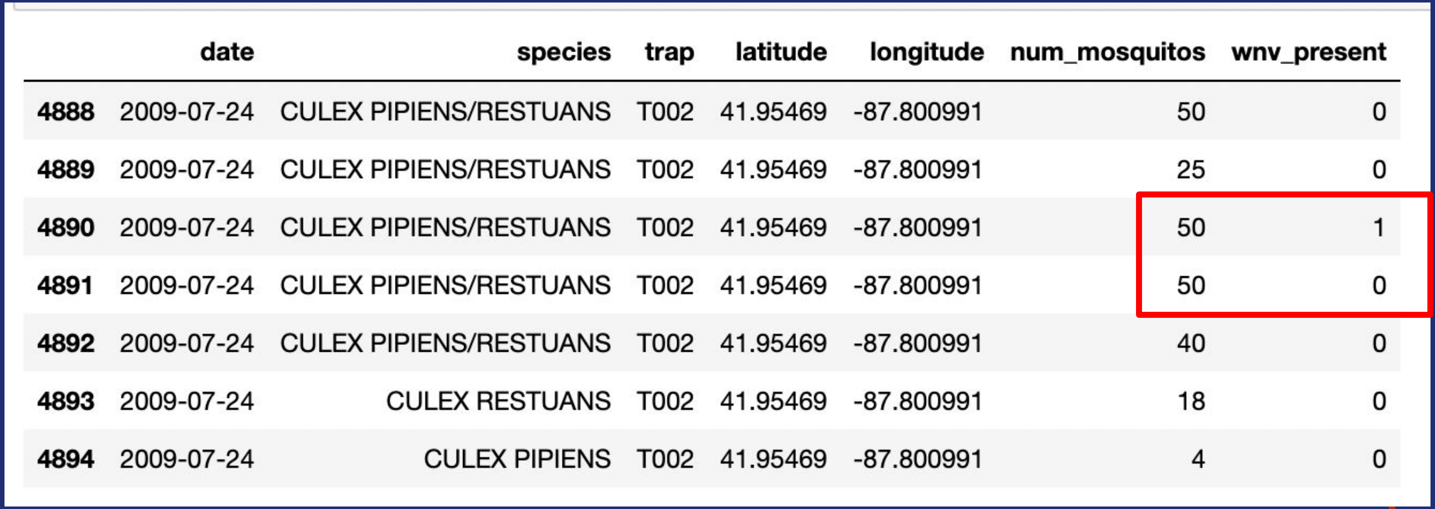
Train Data Quality



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)

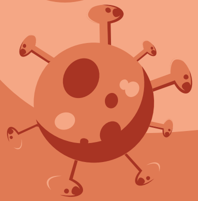
Daylight
⟶ Convert sunrise and sunset into daytime in mins

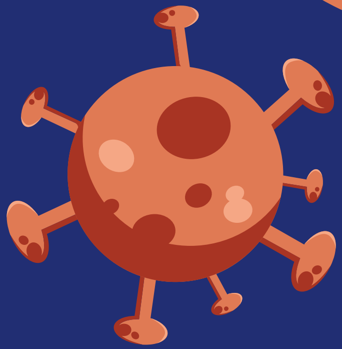
Relative Humidity
⟶ Moisture content in the atmosphere, at constant temperature and pressure.

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

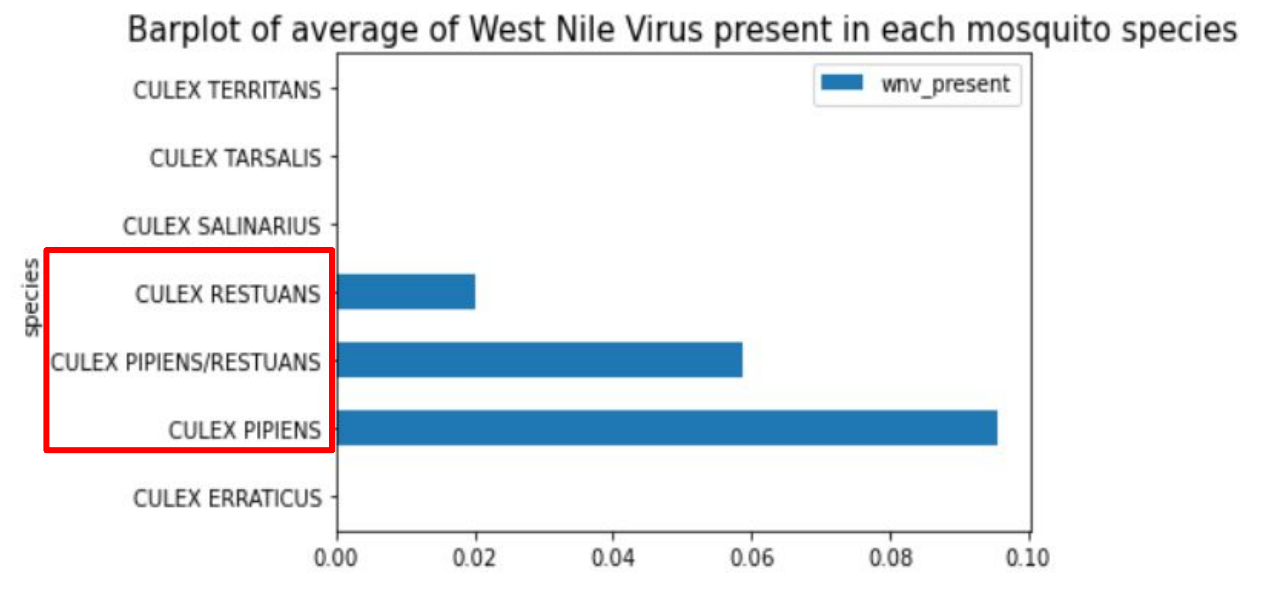
Exploratory Data Analysis

Initial investigation of data

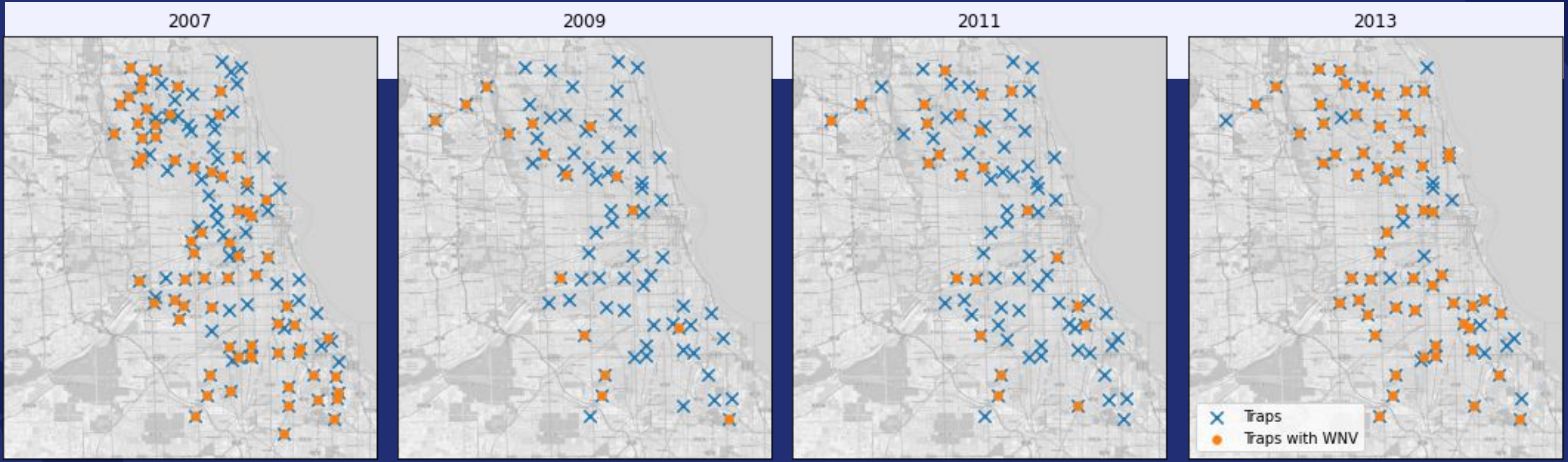




Mosquito species that spread WNV



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

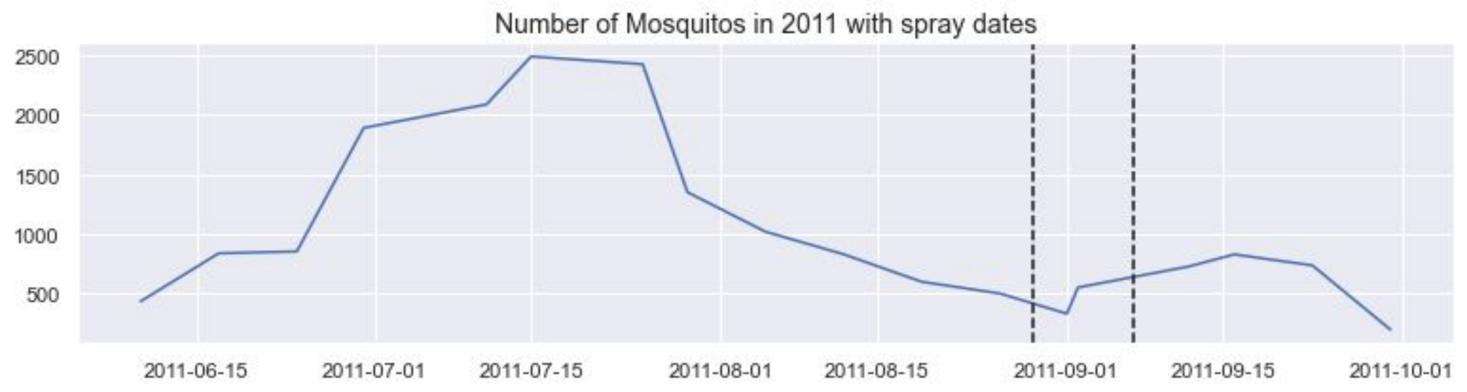
We will use rolling average for the following features:

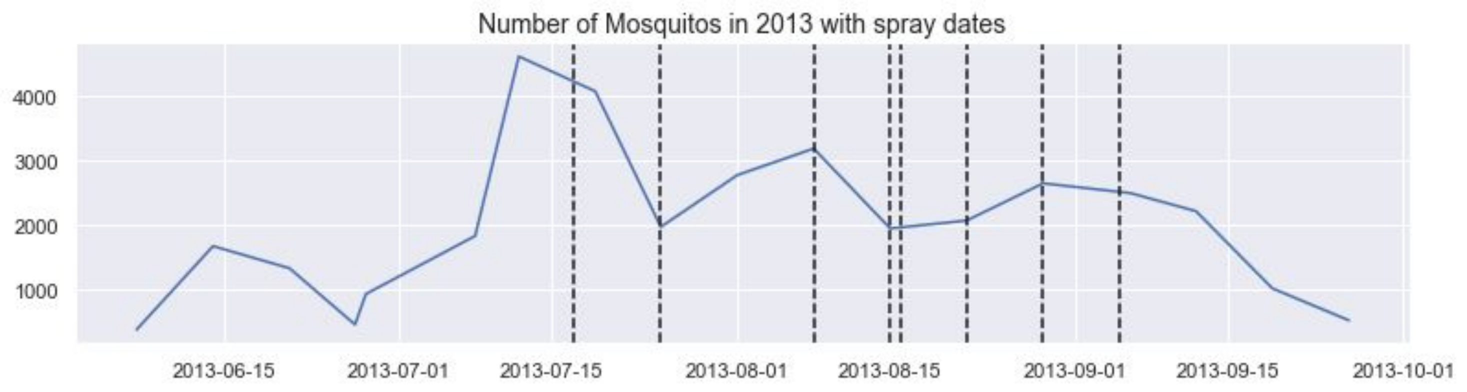
- 'tavg'
- 'precip_total'
- 'r_humid'

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

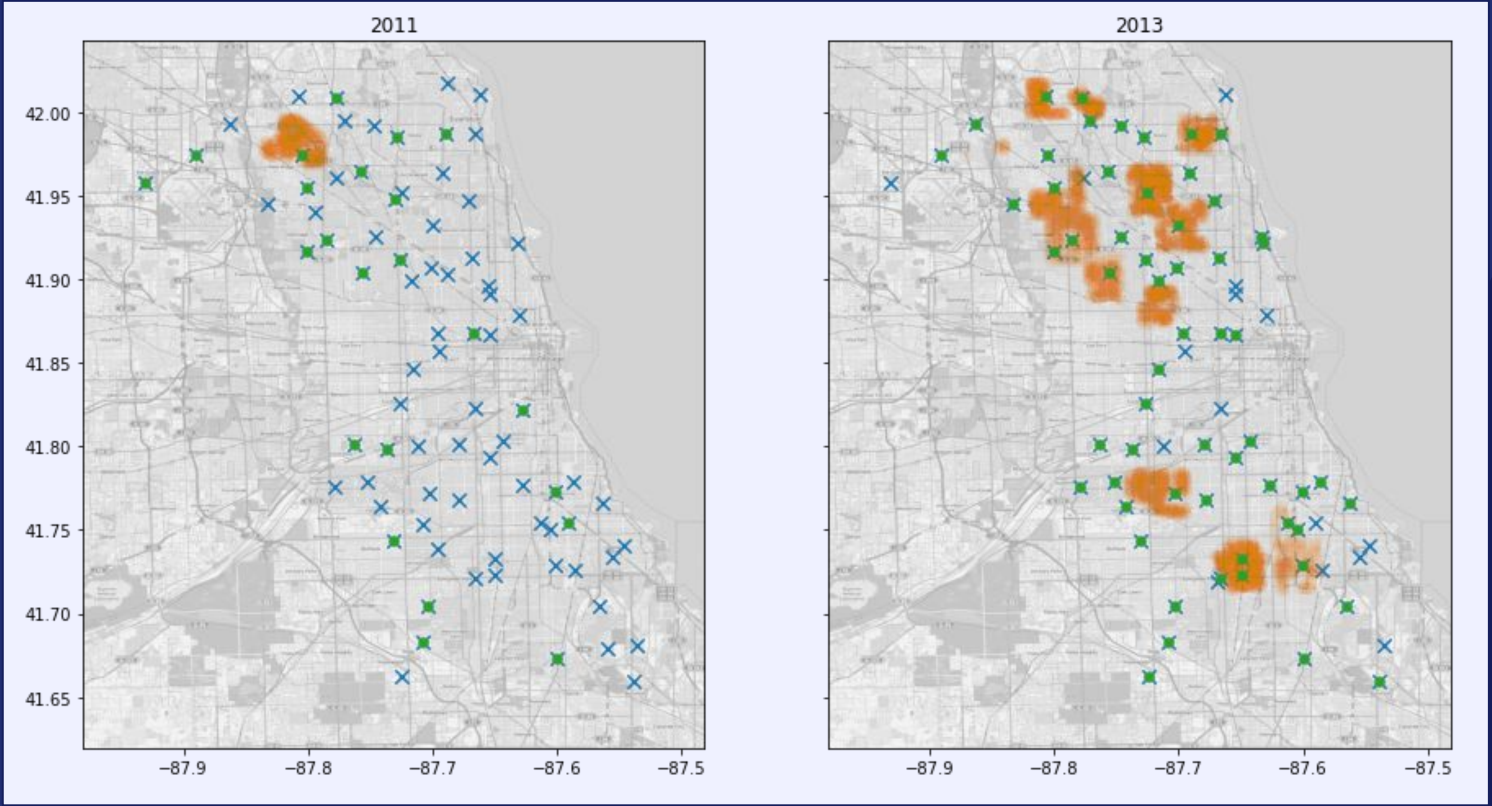
Mosquito Species	Temperature	Lifecycle (days)
CULEX	70° F	14
CULEX	80° F	10

Effect of spray on number of mosquitos

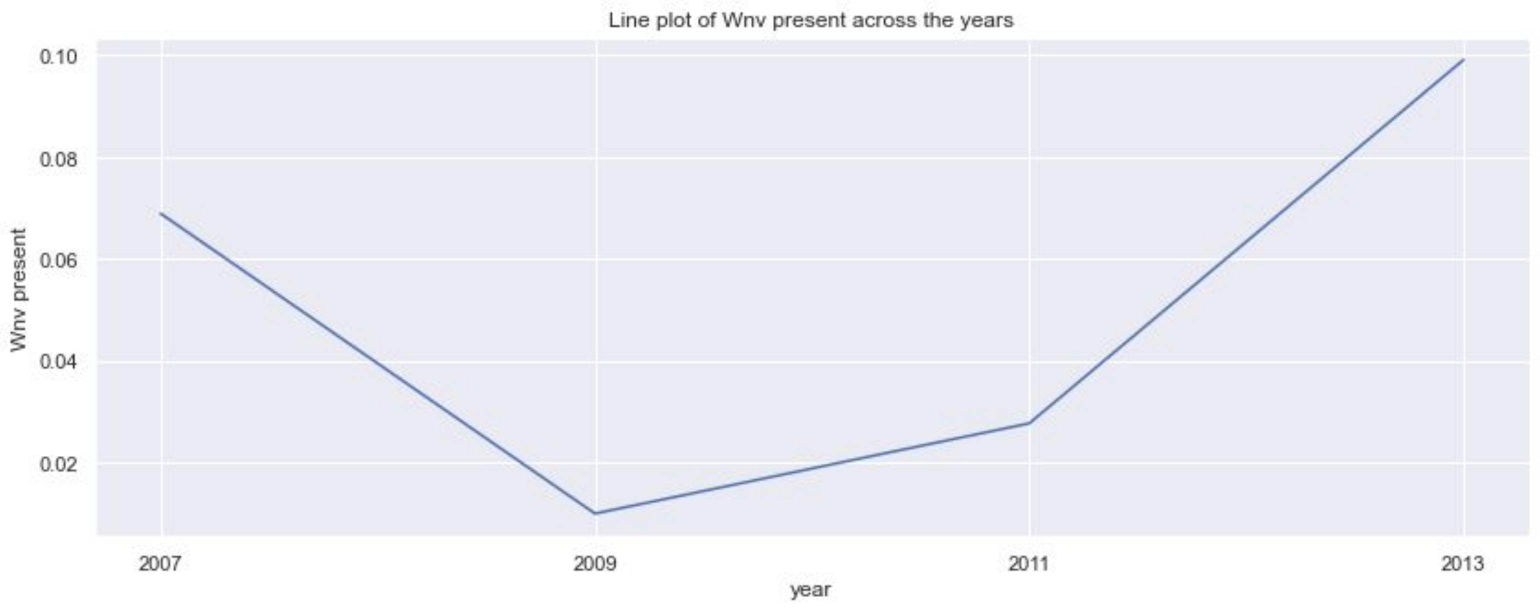




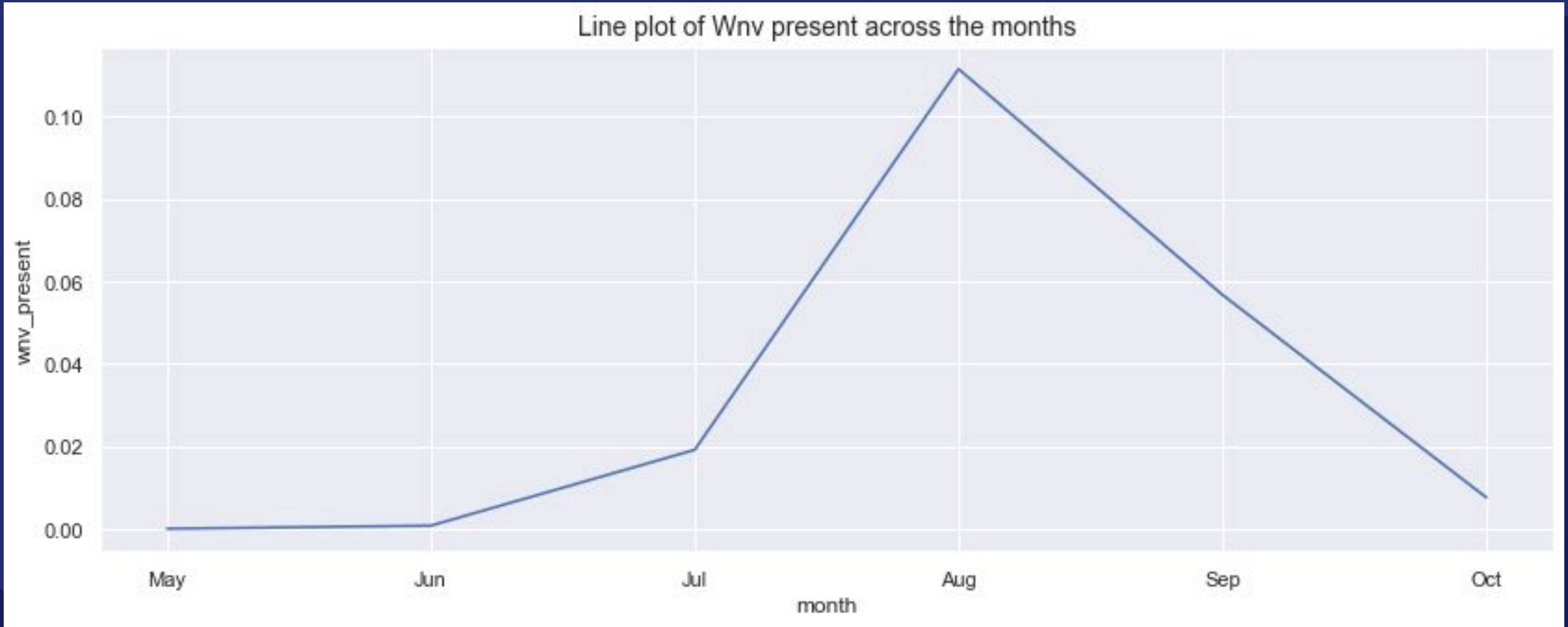
Spray locations in 2011 & 2013



Trend over the years

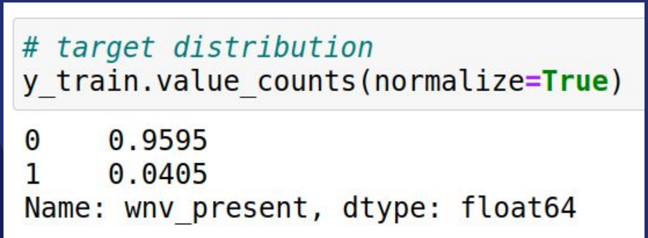


Seasonality in virus presence



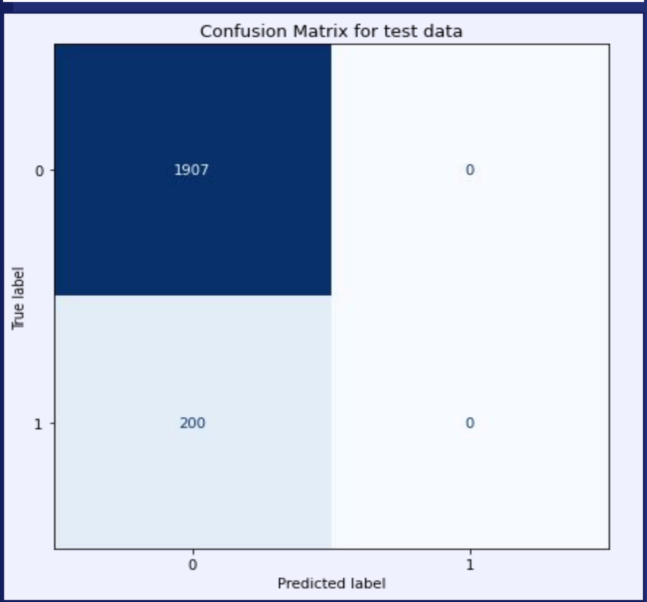
Model Evaluation

Baseline Model



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

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



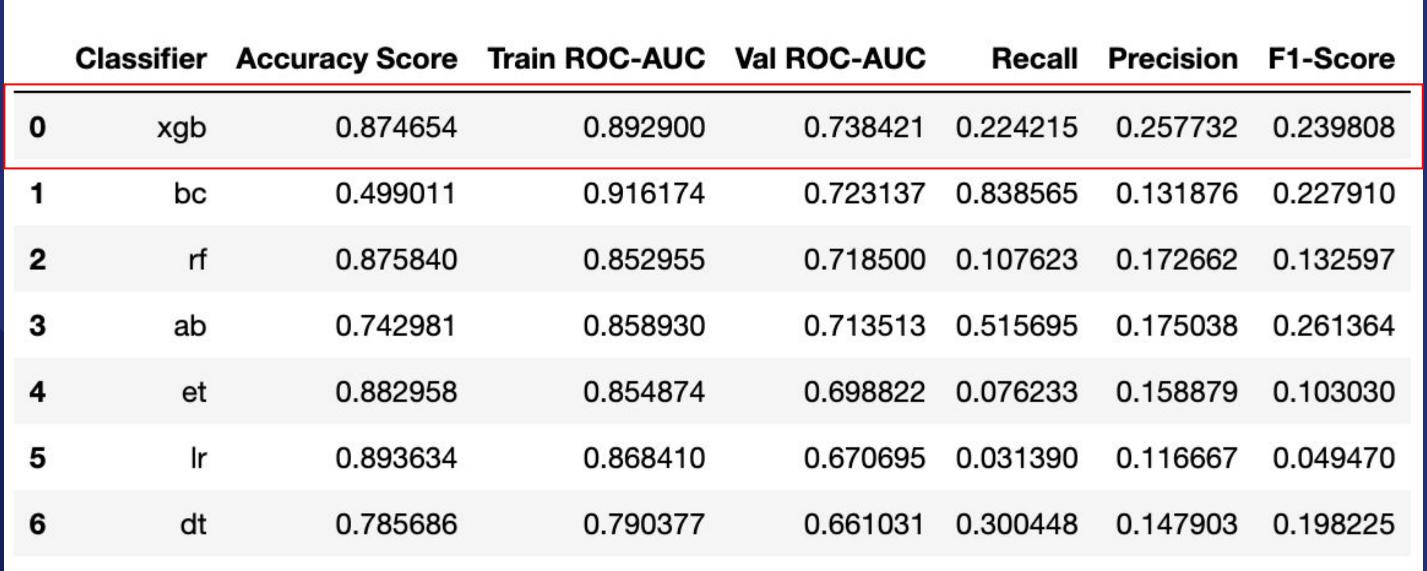
Confusion Matrix for test data

- 96 - 4 ⟶ highly imbalance class
- Training such data 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



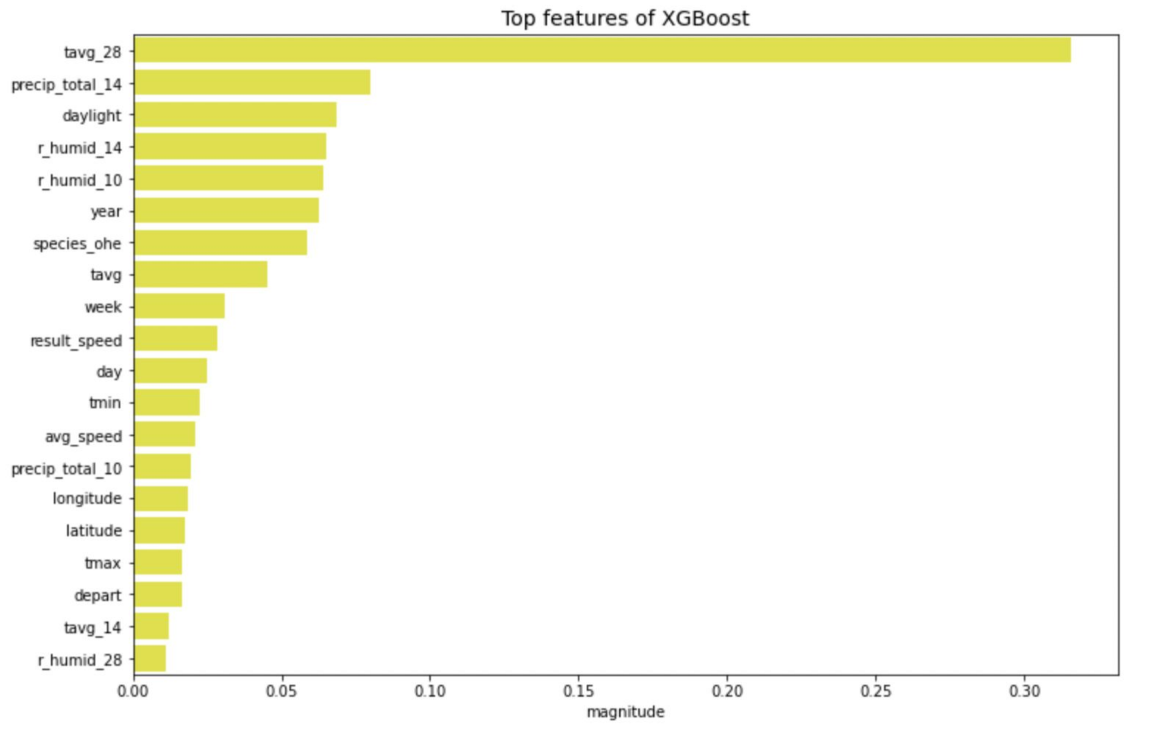
	Classifier	Accuracy Score	Train ROC-AUC	Val ROC-AUC	Recall	Precision	F1-Score
0	xgb	0.874654	0.892900	0.738421	0.224215	0.257732	0.239808
1	bc	0.499011	0.916174	0.723137	0.838565	0.131876	0.227910
2	rf	0.875840	0.852955	0.718500	0.107623	0.172662	0.132597
3	ab	0.742981	0.858930	0.713513	0.515695	0.175038	0.261364
4	et	0.882958	0.854874	0.698822	0.076233	0.158879	0.103030
5	lr	0.893634	0.868410	0.670695	0.031390	0.116667	0.049470
6	dt	0.785686	0.790377	0.661031	0.300448	0.147903	0.198225

Kaggle Score: 0.695

xgb → XGBoost
et → Extra Trees
ab → Ada Boost
bc → Bagging Classifier
lr → Logistic Regression
rf → Random Forests
dt → Decision Trees

Feature importance

- ‘Tavg_28’ was the best feature
- Expected weather features contributing to mosquito breeding in Top 20 as well
- Species of mosquitoes strongly predicts the presence of WNV as well.
- Features engineered such as daylight, r_humid also strongly predict WNV



Cost Benefit Analysis of Spraying

Cost of Spraying

Chicago Department of Health conducts seasonal spraying of Zenivex and only in affected areas (Traps)

Gross Pesticide amount per Acre: 0.87 oz
Cost: $382.78 per gallon

Brighton Park: $4530

Chicago: $39000

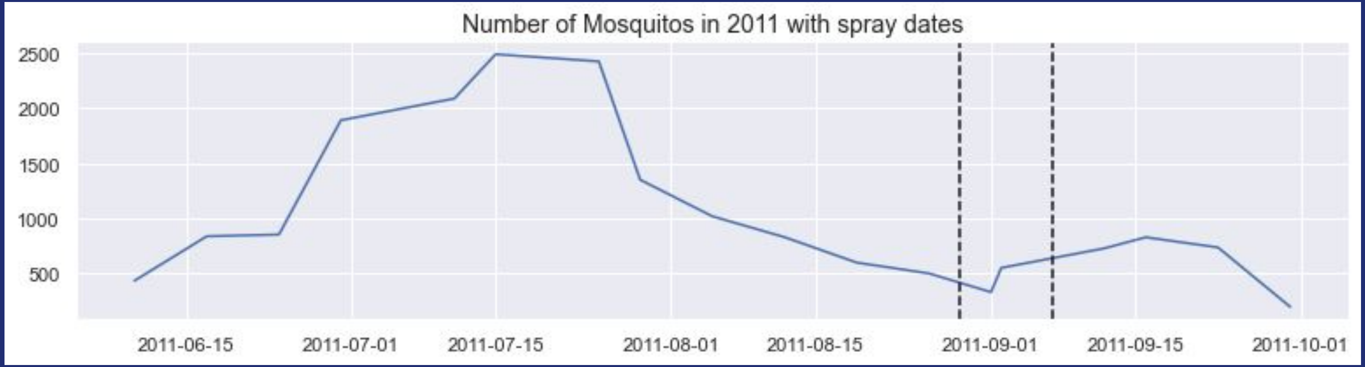
Every 14 days for 3 months: $2.34 million

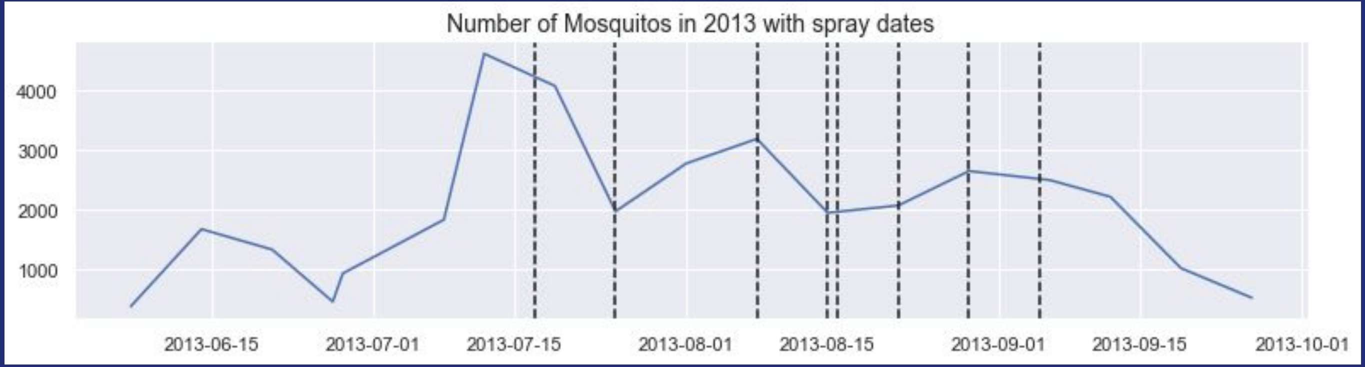
Cost of Treatment

Based on 80 patients hospitalised in 2003, research showed

- Median initial costs : $25,117
- 38 patients experiencing long term medical costs for 5 years amounting to $22,628
- Median salary: $58,247

Cost Benefit Analysis of Spraying





Spraying might not be necessary!

Key Findings

Weather

Lagged weather data: Average temperature, relative humidity, precipitation and daylight

Season

Time of the Year (month, week)

Location

Latitude and Longitude

CONCLUSION AND RECOMMENDATIONS

Government and Expert Collaboration

Government to collaborate with experts and work on the following

- Preventing breeding grounds
- More efficient spraying efforts

Campaigning

Campaign at regions where WNV is more prominent to educate public on importance of protection

Prevention is better than cure

Individual preventive measures as still as important such as wearing long sleeved clothing, wearing insect repellent, removing stagnant water

THANK YOU