Project 4

West Nile Virus

Ivan Cho
Andrea Wong
Carina Rebellon

Problem Statement

Our team aims to build a classification model to **predict the presence of West Nile Virus** in Chicago **supporting the Chicago Department of Public Health in its prevention efforts and control activities** while also educating the general public and health care providers, enabling the Chicago Department of Public Health **to effectively plan and manage their resources in preventing West Nile Virus.**

How far can mosquitoes fly?

How far can mosquitoes fly?

Mosquitoes are **short-distance flyers**. They can only travel **100-200 feet** at a time looking for water containers for breeding. They live their whole life within this short range. Reaching speeds of up to 1 to 1.5 miles per hour, they are **one of the slowest flying insects** around, despite their small body weight.

When is mosquito season?

When is mosquito season?

Breeding season is usually **July through September, while peak West Nile Virus season is usually not until late August** through early September or even October in some areas. Temperatures need to be around freezing before they will start to die off for the winter.

Who do you think mosquitoes have a preference for ?

Who do you think mosquitoes have a preference for ?

“Mosquitoes are attracted to the carbon dioxide, lactic acid and octenol found in our breath and sweat, and they also sense the heat and humidity that surrounds our bodies. They may also have a **preference for beer drinkers**.”

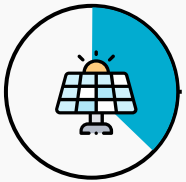
Full Moon makes mosquitoes more active.

Werewolves, zombies… and mosquitoes? Yep, the full moon can be blamed for many strange occurrences both real and imagined. Scientists have never actually seen a real-life zombie, but they do know mosquitoes exist— and that **a full moon can increase mosquito activity 500 percent!**

Datasets

Weather

Data detailing the weather conditions from 2007 to 2014 during the months of the test
E.g. temperature, dew point, total precipitation



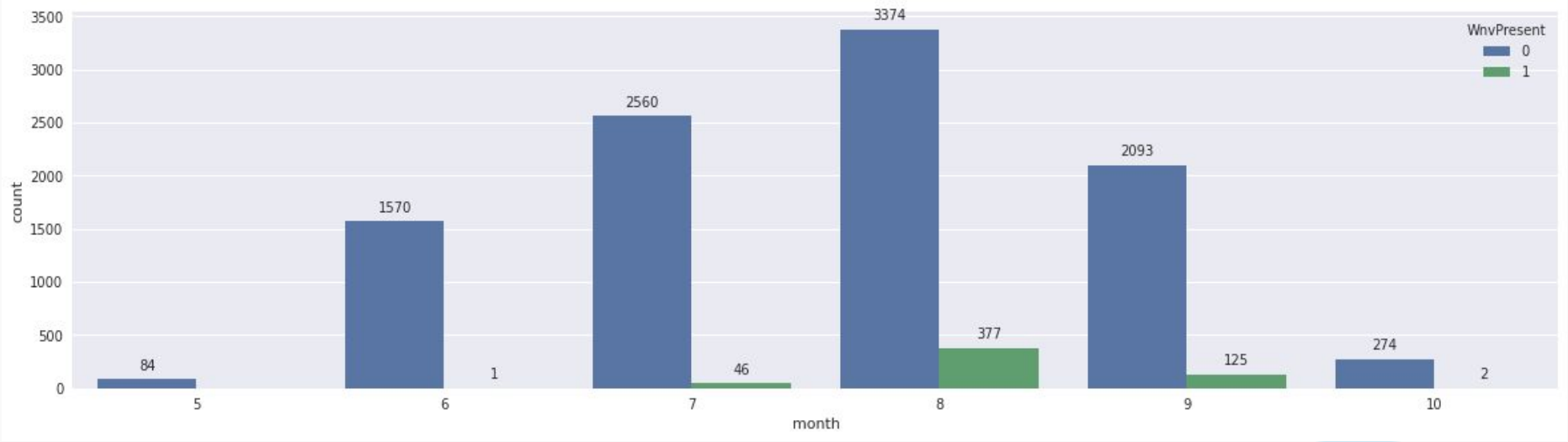
Spray

Data on Chicago's spray effort in 2011 & 2013 including the date, time, latitude and longitude of spray locations

Train - Test

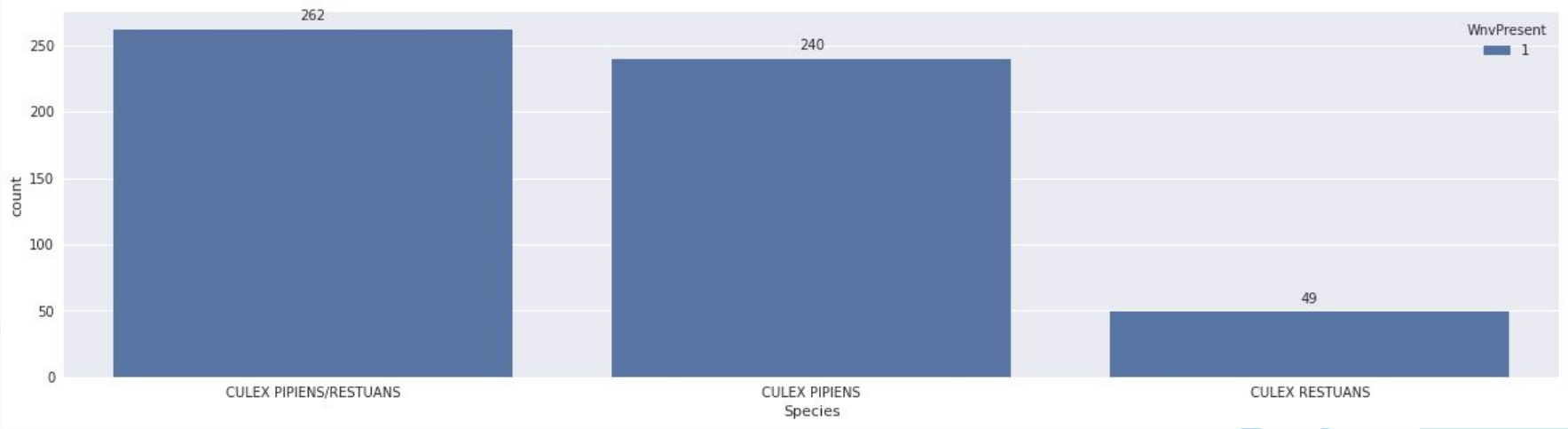
Contains data from 2007 to 2013. But only 2011 and 2013 data such as traps, species captured and etc which are used for machine learning.

Presence of West Nile Virus in traps - Month



Counts of traps with West Nile Virus present was present in July, August & September

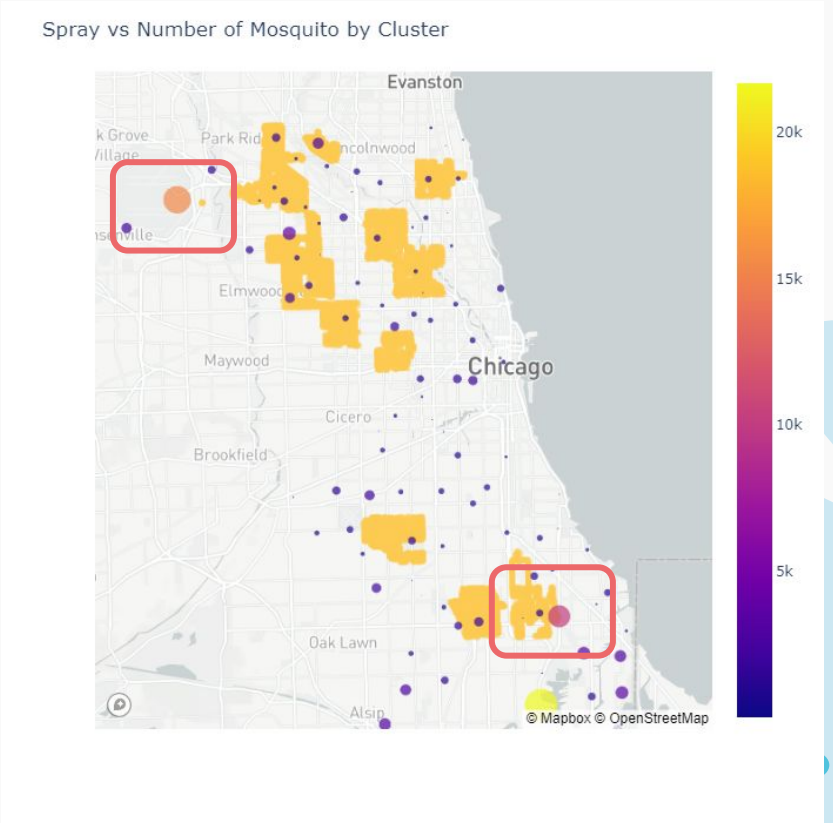
Counts of traps with mosquitoes carrying West Nile Virus



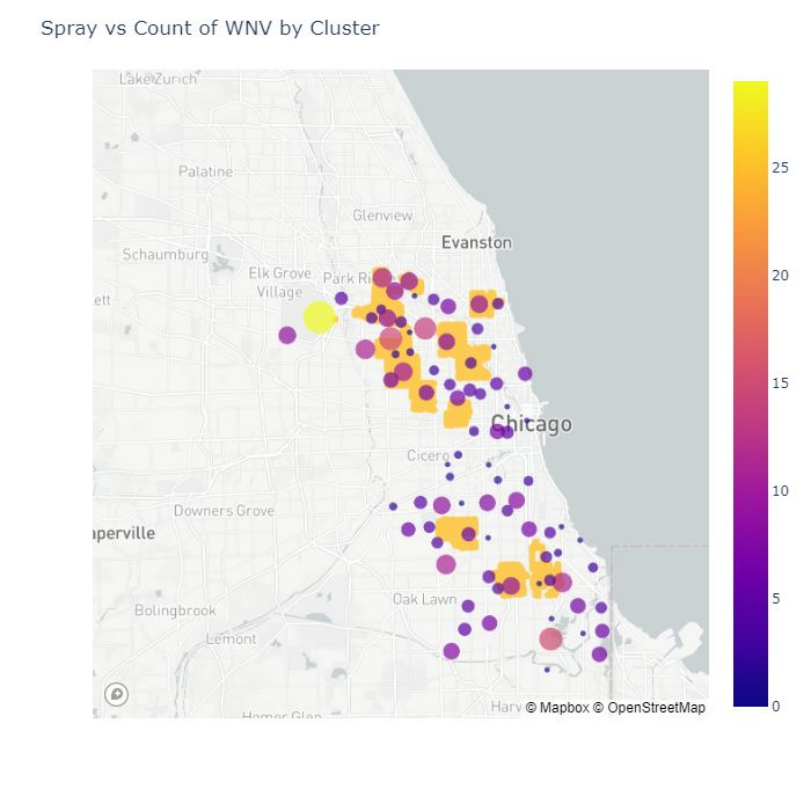
Out of the 7 species of mosquitoes in our dataset, presence of west nile virus were primarily in two species - Culex Pipiens and Culex Restuans

How has spraying affected mosquito counts in the area

In the areas that have been sprayed, the sizes of the mosquito counts are much smaller. There are two big clusters observed where there is no sprays in those areas.



Does spraying reduce counts of WNV?



While spraying may have controlled the number of mosquitoes, it did not seem to have significant impact on the number of virus present.

Feature Engineering Decisions

Cluster

Clusters were created for weather station and breeding clusters

Weather Elements

Relative Humidity

Total Sunlight

Rolling Window

Rolling window for Tavg, dew point, Precip Total and relative humidity

Lat/ Long Features

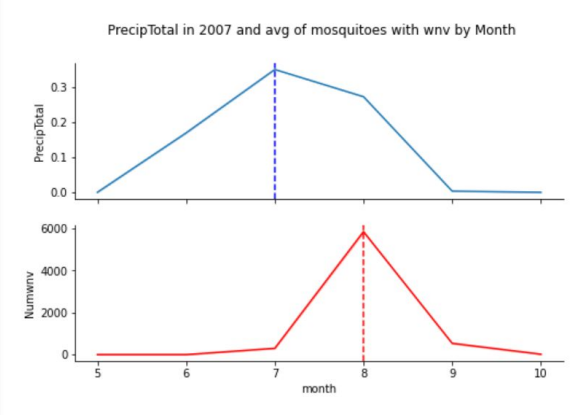
Haversine formula to calculate distance between two points

Date/ Time Features

Presence of WNV by date of the Year

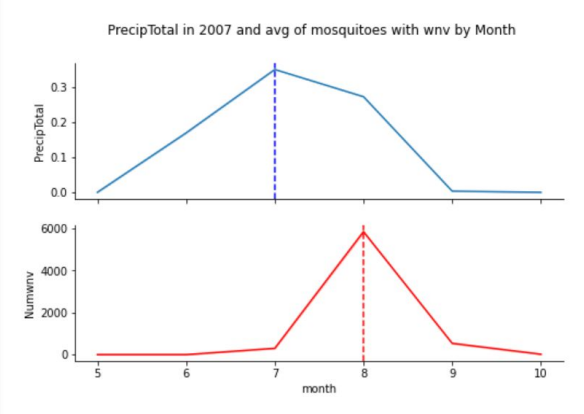
# of Days from the Day of Maximum Presence of WNV

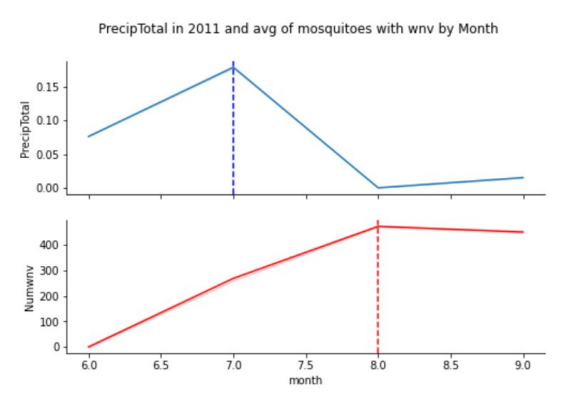
Precip Total leading to presence of Wnv

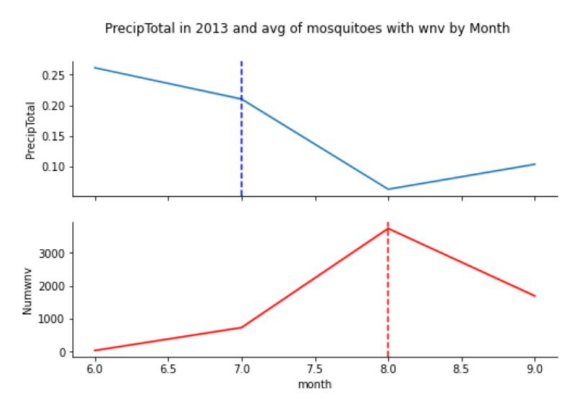


When PrecipTotal reaches peak levels for the year, around 30 days after the heavy rain, presence of WNV will increase as well. One of the reasons could be mosquitoes breeding in stagnant water

Precip Total leading to presence of Wnv

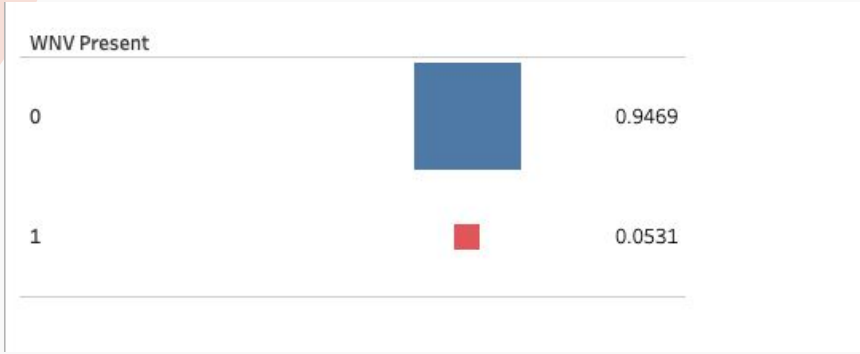






The same was observed in 2011 and 2013 as well.

Distribution of data



Imbalance data is observed and addressed using oversampling technique - SMOTE (Synthetic Minority Oversampling Technique)

Baseline Model acting as Benchmark

Logistic Regression	
Train Score	0.95
Test Score	0.95
ROC-AUC Score	0.50

Model Evaluation / Metrics

Our team will be evaluating our models by optimising these two areas: **ROC-AUC score** and **Recall Score**

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Model	Test ROC AUC	Kaggle ROC AUC	F1	Precision	Recall	Accuracy
ExtraTreesClassifier	**0.7760**	0.6350	0.2534	0.1503	**0.8070**	0.7482
RandomForestClassifier	0.7604	0.6249	0.2587	0.1565	0.8070	0.7738
SGDClassifier	0.7358	0.6881	0.2587	0.1490	0.6929	0.7742
LogisticRegression	0.7251	0.6689	0.2369	0.1436	0.6754	0.7696
RidgeClassifier	0.7153	**0.7068**	0.2200	0.1310	0.6842	0.7431
AdaBoostClassifier	0.6082	0.5551	0.2403	0.2153	0.2719	0.9089
GradientBoostingClassifier	0.5718	0.5878	0.2111	0.2878	0.1667	0.9340

**All models fitted on SMOTE-transformed training datasets.*

Model Evaluation / Metrics

Our team will be evaluating our models by optimising these two areas: **ROC-AUC score** and **Recall Score**

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Model	Test ROC AUC	Kaggle ROC AUC	F1	Precision	Recall	Accuracy
ExtraTreesClassifier	**0.7760**	0.6350	0.2534	0.1503	**0.8070**	0.7482
RandomForestClassifier	0.7604	0.6249	0.2587	0.1565	0.8070	0.7738
SGDClassifier	0.7358	0.6881	0.2587	0.1490	0.6929	0.7742
LogisticRegression	0.7251	0.6689	0.2369	0.1436	0.6754	0.7696
RidgeClassifier	0.7153	**0.7068**	0.2200	0.1310	0.6842	0.7431
AdaBoostClassifier	0.6082	0.5551	0.2403	0.2153	0.2719	0.9089
GradientBoostingClassifier	0.5718	0.5878	0.2111	0.2878	0.1667	0.9340

**All models fitted on SMOTE-transformed training datasets.*

Model Evaluation / Metrics

Our team will be evaluating our models by optimising these two areas: **ROC-AUC score** and **Recall Score**

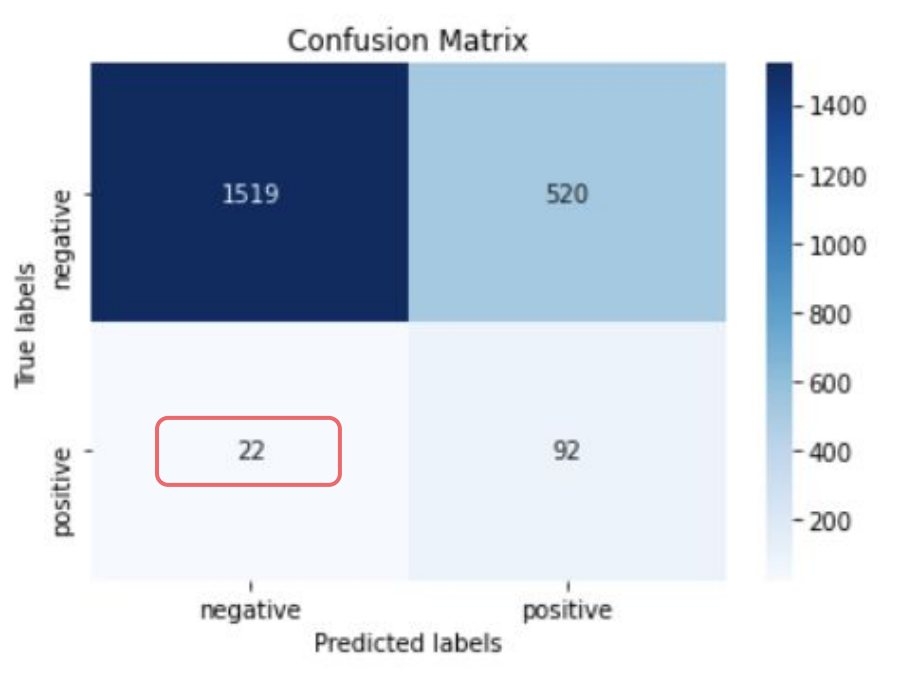
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Model	Test ROC AUC	Kaggle ROC AUC	F1	Precision	Recall	Accuracy
ExtraTreesClassifier	**0.7760**	0.6350	0.2534	0.1503	**0.8070**	0.7482
RandomForestClassifier	0.7604	0.6249	0.2587	0.1565	0.8070	0.7738
SGDClassifier	0.7358	0.6881	0.2587	0.1490	0.6929	0.7742
LogisticRegression	0.7251	0.6689	0.2369	0.1436	0.6754	0.7696
RidgeClassifier	0.7153	**0.7068**	0.2200	0.1310	0.6842	0.7431
AdaBoostClassifier	0.6082	0.5551	0.2403	0.2153	0.2719	0.9089
GradientBoostingClassifier	0.5718	0.5878	0.2111	0.2878	0.1667	0.9340

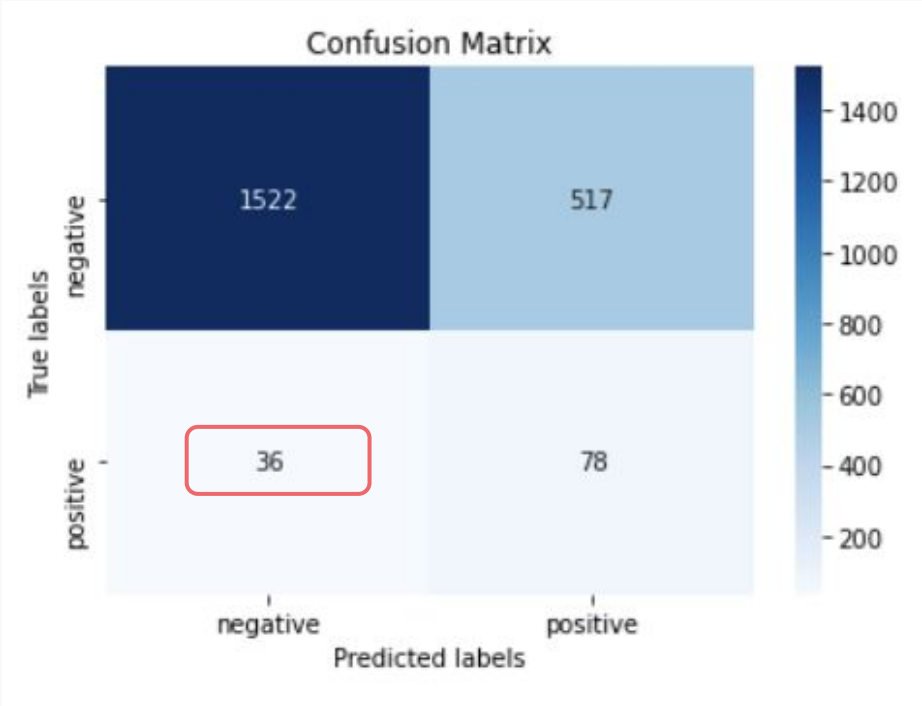
**All models fitted on SMOTE-transformed training datasets.*

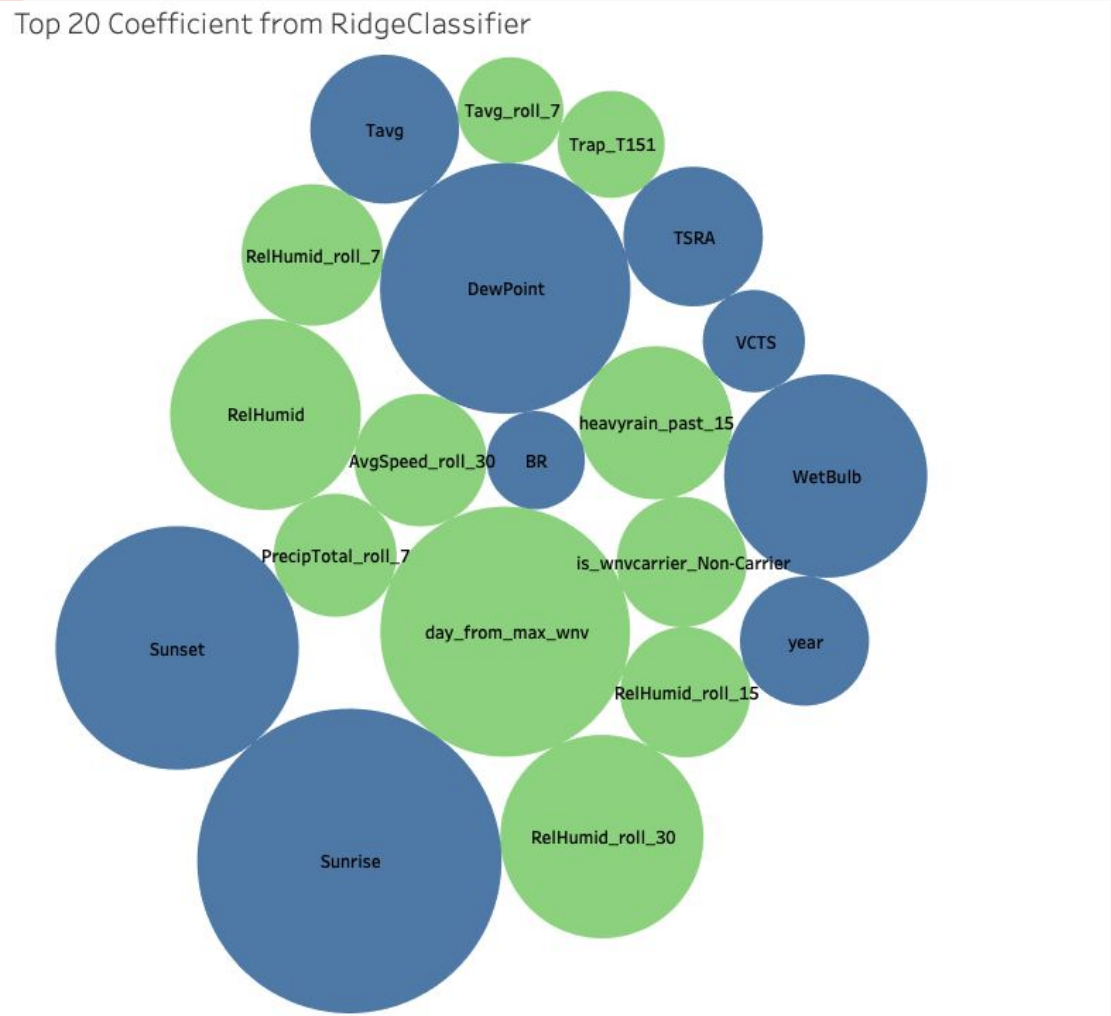
Confusion Metric

ET - Highest ROC-AUC and Recall



Ridge - Kaggle's Best Score





Some of the top features in the Ridge Classifier model include:

- Rolling average features (Tavg, Avg speed)
- Date time features (day from max wnv)
- Weather features (relative humidity)

Recommendations to Redirect resources and Strengthen Mosquito Control Program in Chicago

Continue spray efforts to reduce counts of mosquitoes as it is scientific proven to kill mosquitoes temporarily.

Remove all potential breeding areas- **remove, puncture or regularly drain all water-retaining objects**

Support the **mosquito control program**

- Improve processes in monitoring mosquito traps, investigating breeding sites and educating residents and schools
- **Initiate studies on the specific species** that are west nile virus carriers, **better understand the behavior** of these particular species to predict the mosquito larval occurrences, flight behavior

References

- https://www.renesas.com/us/en/blogs/understanding-relative-humidity-and-dew-point
- https://www.sciencedaily.com/releases/2020/09/200915105932.htm#:~:text=West%20Nile%20virus%20spreads%20most,published%20today%20in%20eLife%20shows
- https://kestrelmeters.com/blogs/news/the-science-of-mosquito-abatement#:~:text=Wind%20works%20as%20a%20natural,MPH%20wind%20gust%20is%20substantial
- https://en.wikipedia.org/wiki/Rain#:~:text=Light%20rain%20%E2%80%94%20when%20the%20precipitation,50%20mm%20(2.0%20in)%20per
- https://www.orkin.com/pests/mosquitoes/mosquito-facts
- https://www.smithsonianmag.com/science-nature/14-not-so-fun-facts-about-mosquitoes-36242998/e
- https://off.com/en/education/insects-101/5-surprising-facts-about-mosquitoes




Questions?

TEST ROC-AUC SCORE: 78%

RECALL: 81%

Extra Trees (Extremely Randomized Trees) the ensemble learning algorithms. It constructs the set of decision trees. During tree construction the decision rule is randomly selected. This algorithm is very similar to Random Forest except random selection of split values.

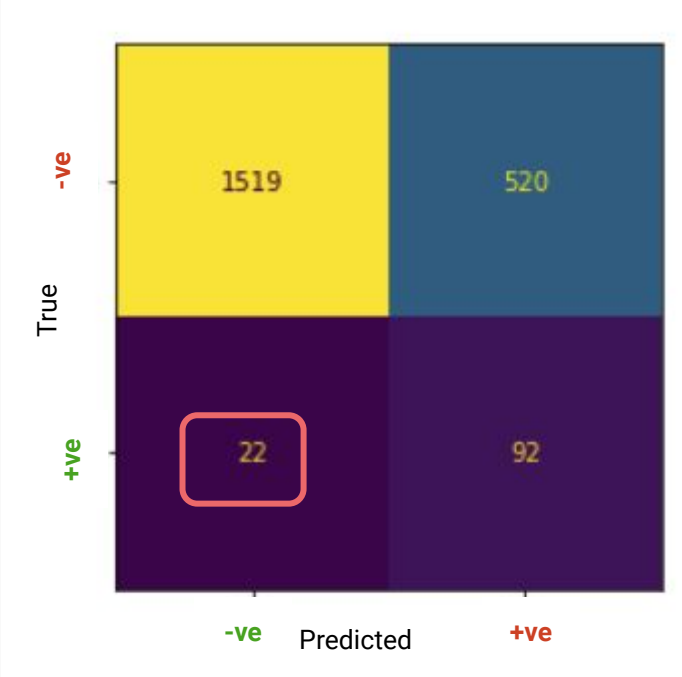
Best Params

The maximum depth of the tree: **6**

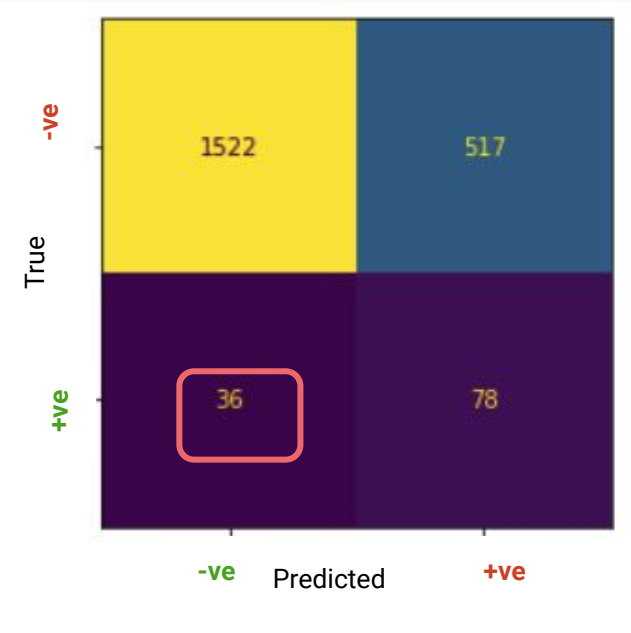
The number of trees in the forest: **75**

Confusion Metric

ET - Highest ROC-AUC and Recall



Ridge - Kaggle's Best Score



Fun Facts

How Far Can They Fly?

Answer: Mosquitoes are short-distance flyers. They can only travel 100-200 feet at a time looking for water containers for breeding. They live their whole life within this short range. Reaching speeds of up to 1 to 1.5 miles per hour, they are one of the slowest flying insects around, despite their small body weight.

When Is Mosquito Season?

Answer: Breeding season is usually July through September, while peak West Nile Virus season is usually not until late August through early September or even October in some areas. Temperatures need to be around freezing before they will start to die off for the winter.

https://www.orkin.com/pests/mosquitoes/mosquito-facts

"Mosquitoes are attracted to the carbon dioxide, lactic acid and octenol found in our breath and sweat, and they also sense the heat and humidity that surrounds our bodies. They may also have a **preference for beer drinkers**."

https://www.smithsonianmag.com/science-nature/14-not-so-fun-facts-about-mosquitoes-36242998/e

A Full Moon Makes Mosquitoes More Active

Werewolves, zombies… and mosquitoes? Yep, the full moon can be blamed for many strange occurrences both real and imagined. Scientists have never actually seen a real-life zombie, but they do know mosquitoes exist— and that a full moon can increase mosquito activity 500 percent!

https://off.com/en/education/insects-101/5-surprising-facts-about-mosquitoes

Kaggle Score

Submission and Description	Private Score	Public Score	Use for Final Score
SGDClassif.csv 2 minutes ago by Carina Rebellon add submission details	0.65518	0.68816	☐
ExtraTrees.csv 3 minutes ago by Carina Rebellon add submission details	0.60679	0.63503	☐
RidgeClass.csv 3 minutes ago by Carina Rebellon add submission details	0.67204	0.70684	☐
LogisticRe.csv 4 minutes ago by Carina Rebellon add submission details	0.64876	0.66890	☐
GradientBo.csv 4 minutes ago by Carina Rebellon add submission details	0.56272	0.58781	☐
RandomFore.csv 5 minutes ago by Carina Rebellon add submission details	0.60627	0.62485	☐

Modelling results (SelectKBest)...

Model	Test ROC AUC	KAGGLE ROC AUC	F1	Precision	Recall	Accuracy
ExtraTreesClassifier	**0.7358**	0.6357	0.2305	0.1370	**0.7281**	0.7427
SGDClassifier	0.7151	0.6777	0.2229	0.1334	0.6754	0.7506
RidgeClassifier	0.7114	**0.7013**	0.2181	0.1301	0.6754	0.7436
RandomForestClassifer	0.7251	0.6330	0.2263	0.1349	0.7018	0.7459
LogisticRegression	0.7181	0.6819	0.2303	0.1392	0.6667	0.7641
AdaBoostClassifer	0.6290	0.6555	0.2687	0.2338	0.3158	0.9090
GradientBoostingClassifier	0.6105	0.5928	0.2749	0.2990	0.2544	0.9289

West Nile Virus Transmission

01

An infected mosquito bites a bird

02

Bird transmit the disease while flying to another location

03

Mosquitoes become infected when they feed on infected birds

04

Infected mosquito feeds on humans who also becomes infected. The virus can also infect other mammals including horses

METHODOLOGY

DATA CLEANING

FEATURE ENGINEERING

CONCLUSION & RECOMMENDATIONS

EXPLORATORY DATA ANALYSIS (EDA)

MODELLING

Objectives

01

PROBLEM STATEMENT

02

EXPLORATORY DATA ANALYSIS (EDA)

03

FEATURE ENGINEERING

04

MODELLING RESULTS ANALYSIS

05

CONCLUSIONS & RECOMMENDATIONS

Beginnings of West Nile Virus

West Nile Virus was first identified in NYC in the summer and is leading cause of mosquito-borne disease in the US

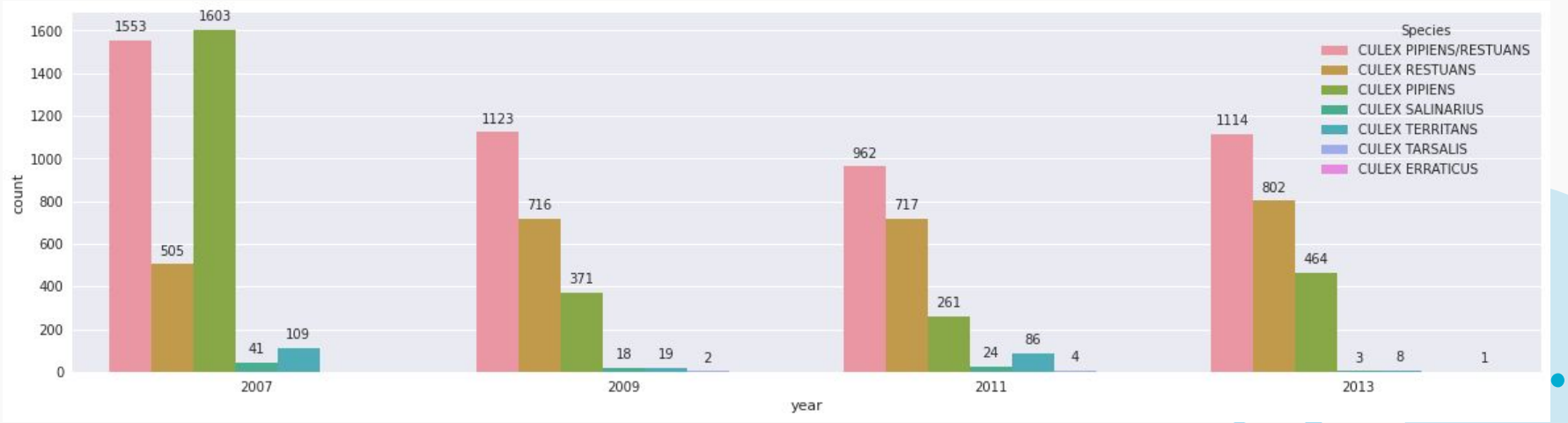
1999

WNV quickly adapted to the local populations of Culex vector mosquitoes and avian populations, rapidly spreading throughout United States

Chicago suburbs

Suitability of the environment for mosquito breeding and transmission to key avian species, especially the American robin

Counts of species captured in traps over the years



2007 saw Culex Pipiens taking up 42% of traps sampled with mosquitoes caught. However in 2009, 2011 and 2013, Culex Pipiens/Restuans represented the bulk of the traps sampled