West Nile Virus Prediction & Prevention

DSI19 Project 4

Members: Nemo, Dennis, Adrian

Introduction

Problem Statement

To effectively make use the budget on hand, maximising mosquito control. The state would like my team to come out with a cost-benefit solution for the pesticide deployment with the data provided.

Objective

- Predict when and where different species of mosquitoes will test positive for WNV.
- Predict where to spray pesticides.
- Predicting features that will affect the number of mosquitoes
- Cost-benefit analysis.

EDA

Overall methodology:

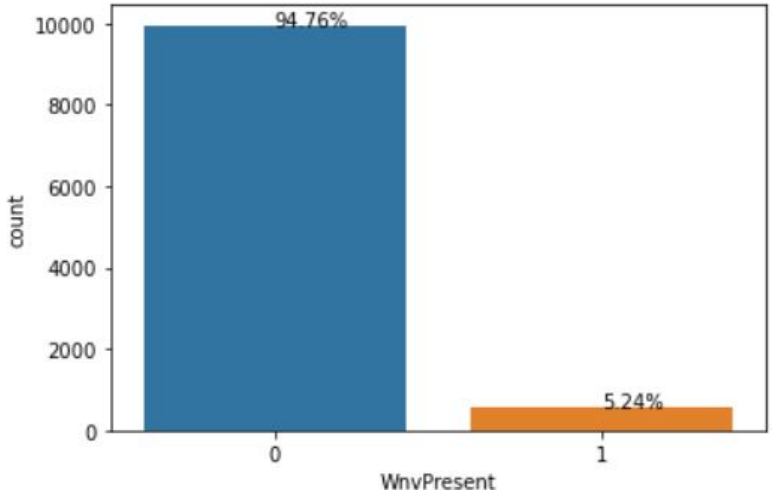
First, explore weather and train csvs together after dropping heavily missing columns: ‘Depth’, ‘Water1’, ‘SnowFall’, and ‘Heat’

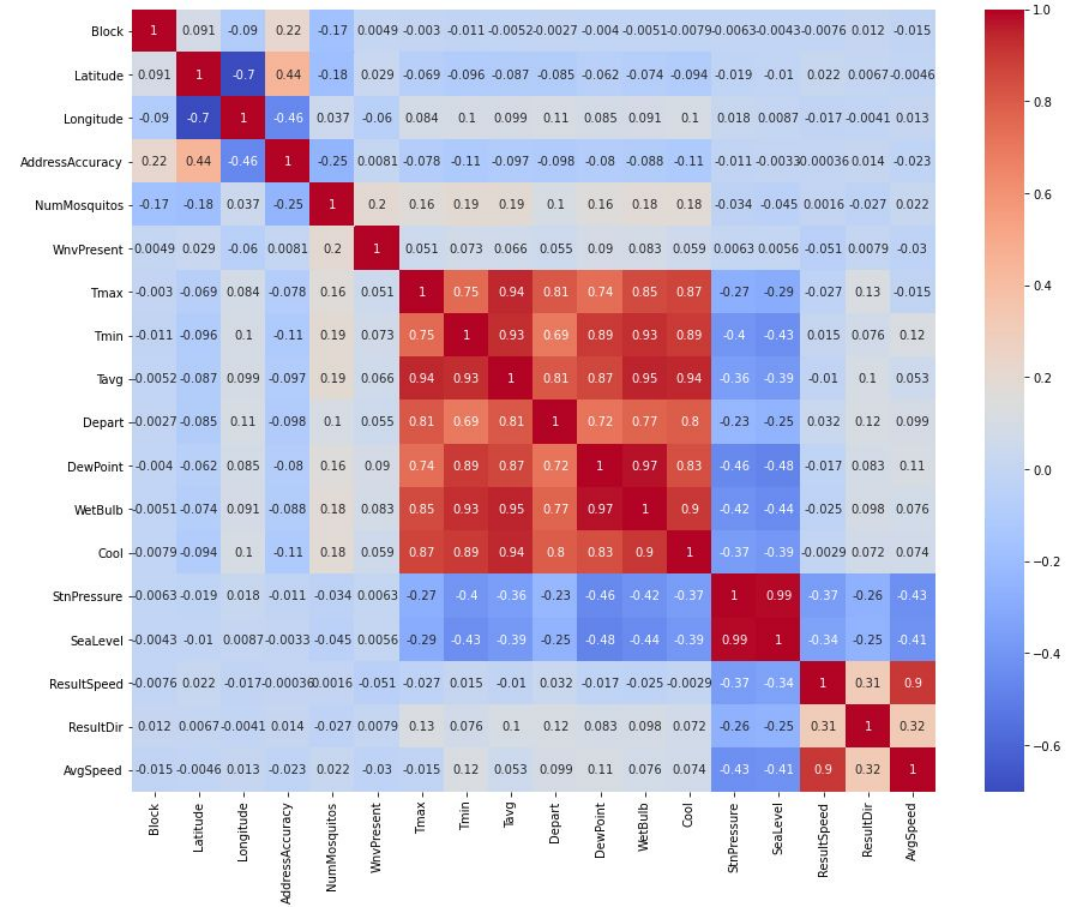
Then, with a fresh set, clean weather data before combining with train and test data

Imbalanced classes! Many 0s, very few 1s

In hindsight, we should resample the data

Count of WnvPresent mosquitos in dependent variable



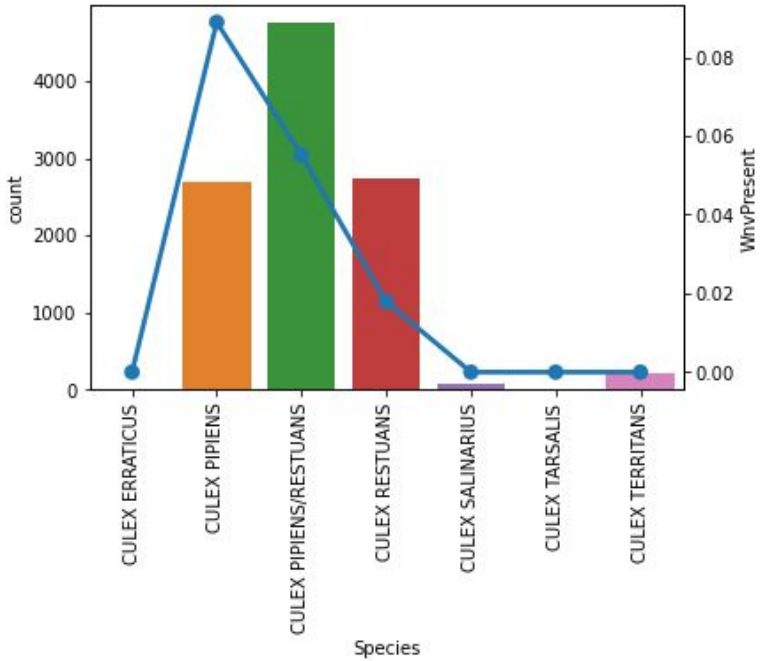


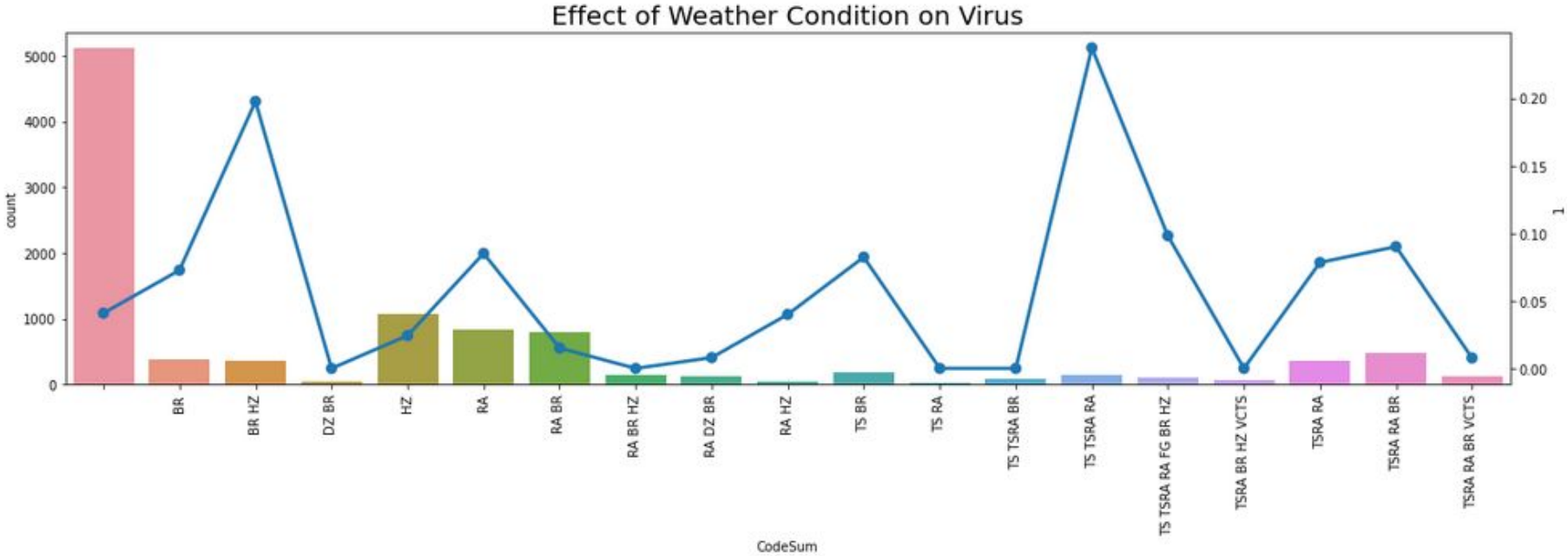
- WnvPresent (6th row) is not correlated with any other independent variable, even no. of mosquitos
- Big red square in the middle: Temperature is correlated with each other
- StnPressure and Sealevel are heavily correlated with each other, but negatively correlated with temperature and wind speed
- Direction of the wind is slightly correlated with the wind speed

- Only 3 mosquito breeds carry the virus:
  - Culex Pipiens
  - Culex Restuans
  - Culex Pipiens/Restuans crossbreeds

- Culex Pipiens has the highest infection rate. This is important information

Count of the number of infected mosquitos among different species



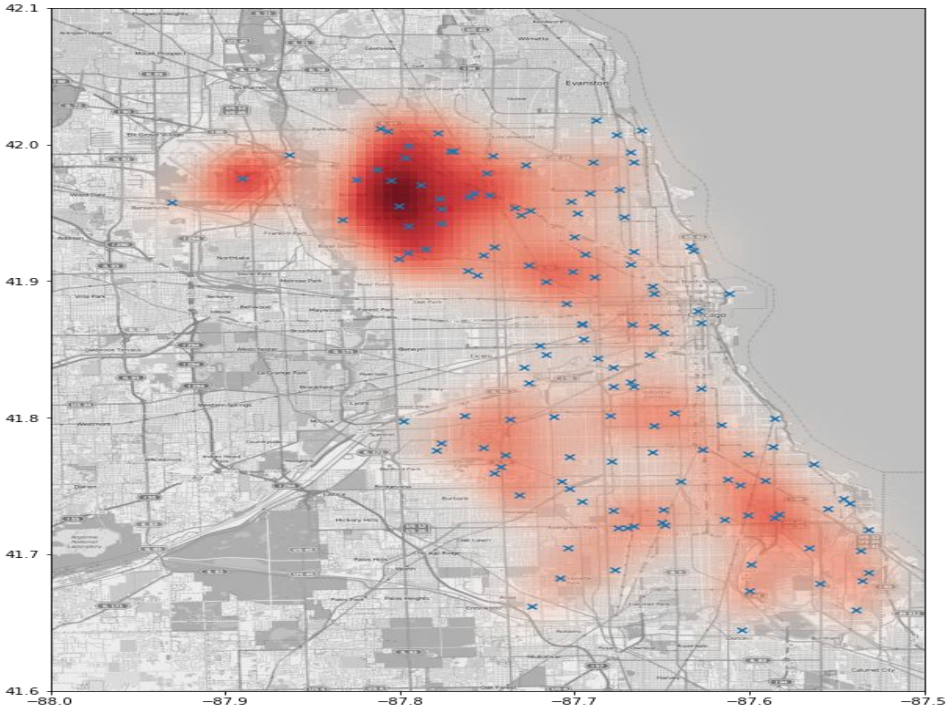


- Thunderstorms, rain and thunderstorms, and rain have the highest number of infected mosquitos caught.

- Mist and haze has the second highest.

- North side has many infected mosquitos: A good choice for spraying

Number of traps and infected mosquitos caught



Models we used

Logistic Regression

K-Nearest Neighbors

Random Forest

Support Vector Machines

AdaBoost

Most models: Low bias (95%+), very high variance (20-30% difference)

AdaBoost: Higher bias (73%), very low variance (1% difference)

Adaboost as production model

Kaggle score: 64%

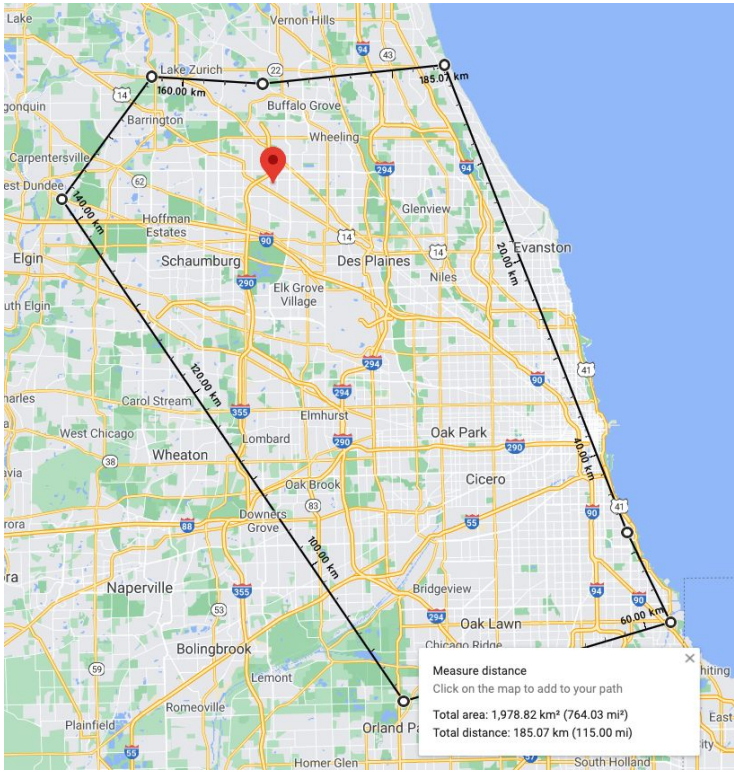
...**But very low probabilities guessed (0.1 & 0.2%)**

Model guessed majority class!

Cost Benefit Analysis: Cost

Cost of spraying:

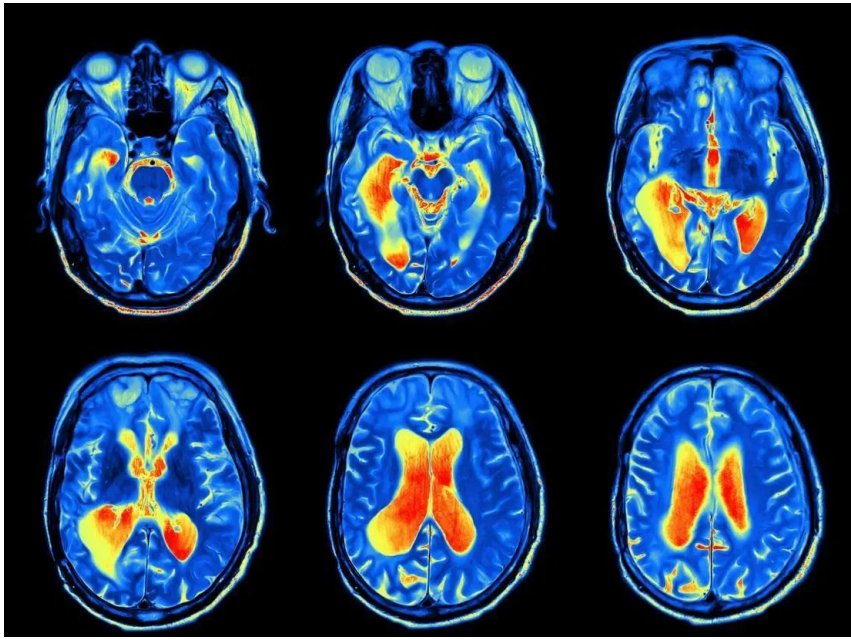
- Spray lasts for 1 year
- total area is 1978.82 km2
- cost of 200 dollar per acre,
- Total cost of spray is 489k USD



Cost Benefit Analysis: Benefits

WNV's Consequence on Health

- Mild symptoms - 1 in 5 people who are infected develop a fever
- Serious symptoms - 1 in 150 people's central nervous system or brain are damaged



Cost Benefit Analysis: Benefits

Economic Consequence

- GDP per capita of an average person in Illinois is around in 61713 USD in 2019.
- 225 were infected with WNV in 2019 in the city of chicago

Formulas

benefit per year due to virus = regaining the economic damage caused by virus

cost per year due to virus = hospitalised + permanently loss of ability to work

Different Classes

- true positive: correctly detects that the virus is present (thus, spray)
- true negative: correctly detects that the virus is not present (thus, do not spray)
- false positive: incorrectly detects that the virus is present (thus, spray but incorrectly so)
- false negative: incorrectly detects that the virus is not present (thus, do not spray but incorrectly so)

Benefits vs Cost

Benefit of True Positive = 2777496.42

Benefit of True Negative = 0

Cost of False Positive = -488977.071

Cost of False Negative = 0

TN = 694/2627

FP = 1795/2627

FN = 16/2627

TP = 122/2627

Expected Value = -311404

Implications of Negative Expected Value

Explanation:

A negative value indicates that the project is expected to generate greater disbenefits than actual benefits; meaning that on a net basis, the project would make conditions worse rather than better.

Possible Reasons:

- Spraying the whole area rather than just clusters of it was too aggressive
- Cost of Spray is too high as the data taken was not from an industrial spray website, rather, it was for homes
- Possible benefits of spraying is understated

Solutions & Recommendations

1. Awareness Campaigns
   - Identify and report breeding areas to authority with cash rewards
   - Set up pop up booths to educate residence

2. Only Spray Areas most affected by the virus