

A DATA-DRIVEN APPROACH AGAINST WEST NILE VIRUS IN CHICAGO

By: Amira, Joseph, Joshua,
Nelson, Zhi Hong

DSI-28
10 Jun 2022

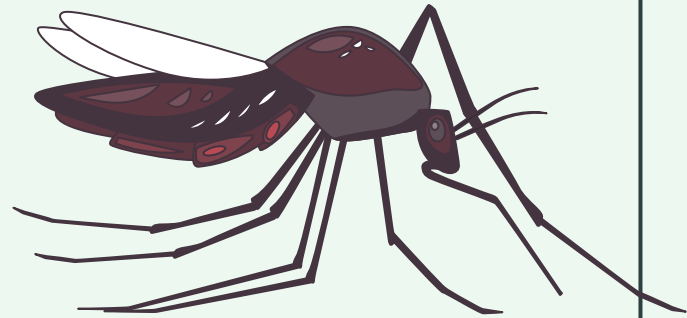


TABLE OF CONTENTS

01

BACKGROUND

02

**DATA CLEANING &
EDA**

03

**FEATURE
ENGINEERING**

04

**MODELLING &
EVALUATION**

05

**COST-BENEFIT
ANALYSIS**

06

**CONCLUSIONS &
NEXT STEPS**

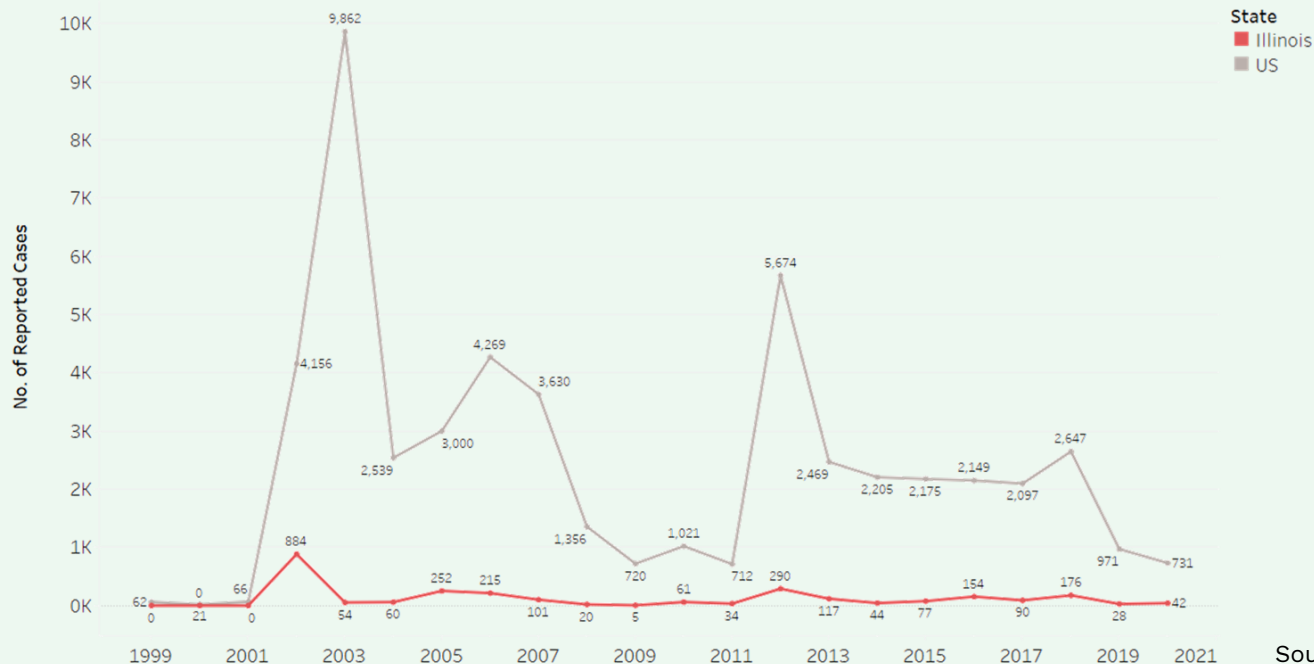
01.

BACKGROUND



State of Illinois has been effectively tackling WNV over the past two decades

West Nile virus disease cases reported to CDC by State of Illinois, 1999-2020



Source: CDC

Existing surveillance and control strategy is important to curb the prevalence of WNV in Chicago



Treating catch basins with larvicide to limit mosquito breeding and reduce adult mosquito population density



Setting mosquito traps across the city to detect WNV



Testing dead birds found in the city to detect WNV



Routine spraying of adulticides



Educating the public to proactively reduce no. of mosquitos in their own areas and take personal precaution to avoid bites

We propose to improve the control strategy with a targeted approach for adulticide sprays

1

Machine Learning Model to accurately predict WNV occurrence across Chicago

Evaluation Criteria:

- High ROC AUC score above 85% on validation set
- High ROC AUC score above 70% on truly unseen data

2

Cost Benefit Analysis of adulticide spraying

- To quantify spraying costs and external costs of WNV
- To weight the benefits of spraying against its costs

The background features a light teal color with several abstract elements: a large blue shape with a black dot pattern in the top left, a solid blue circle in the top right, a grey shape with a black dot pattern in the bottom left, and a large grey shape in the bottom right. A thin black line runs vertically on the left and horizontally across the top and bottom. A small grey circle is positioned to the left of the main text box, and a small blue circle is to its right.

02.

**Data Cleaning &
EDA**

Scope of Data

1. **Train Set**

- Contains data from 2007 - 2013 (2 years interval)
- No missing values

1. **Test Set**

- Data from 2008 - 2014 (2 years interval)
- No missing values

1. **Spray Set (Will not use)**

- Contains data from 2011 and 2013
- Missing values in spray timing

1. **Weather Set**

- Contains data from 2007-2014 (8 years)
- Missing values in many columns

Data Cleaning Process

1. Dropped non-essential columns
1. Combined duplicated rows due to mosquitoes count limit
1. Assignment of stations based on each point's lat and lon using Haversine Distance





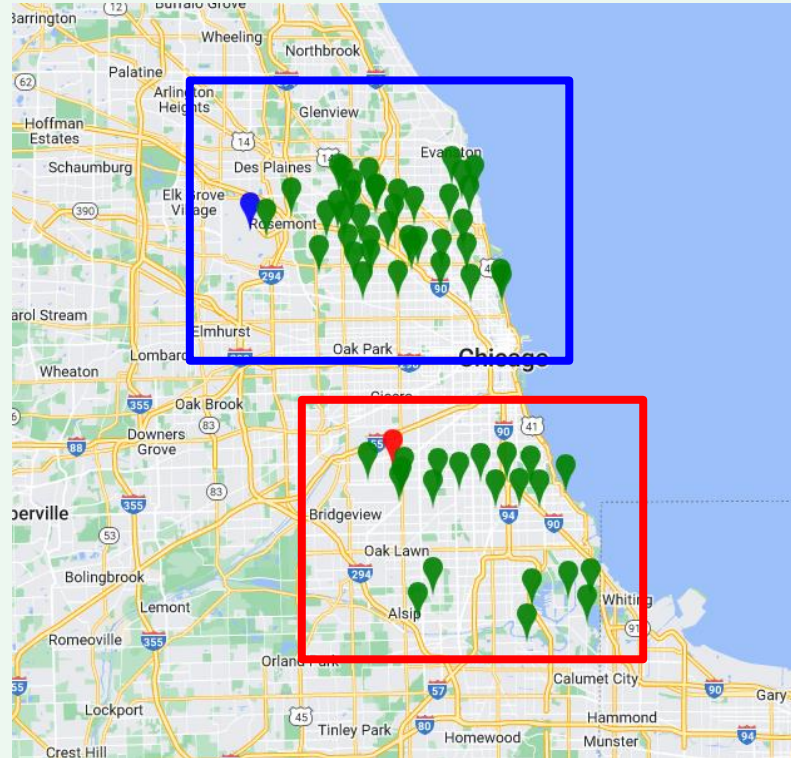
- Station 1



- Station 2

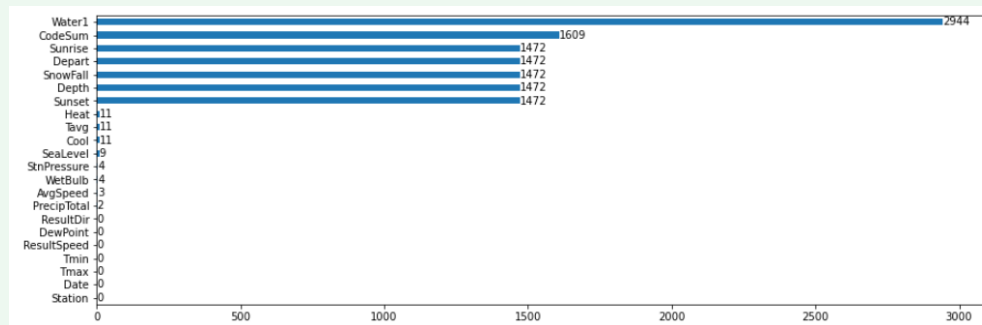


- Traps deployed



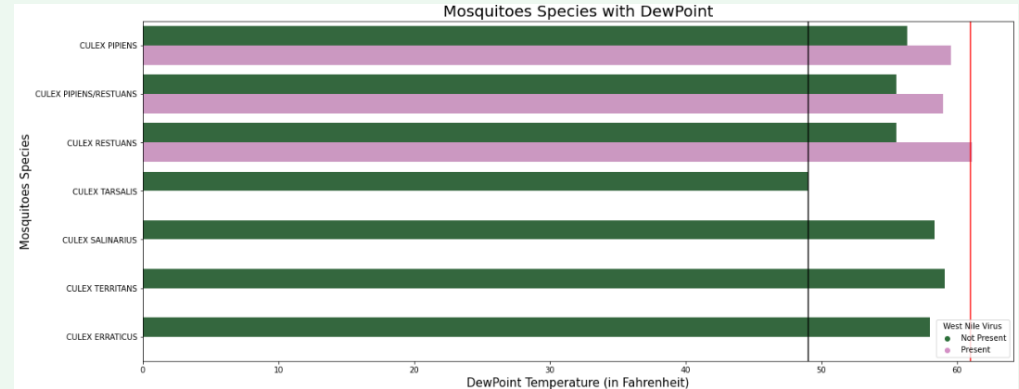
Missing Values

1. Dropped columns with $> 80\%$ data missing
1. Forward-fill sunrise and sunset
1. Filling average temp with Min and max
1. Mean for $<0.3\%$ data



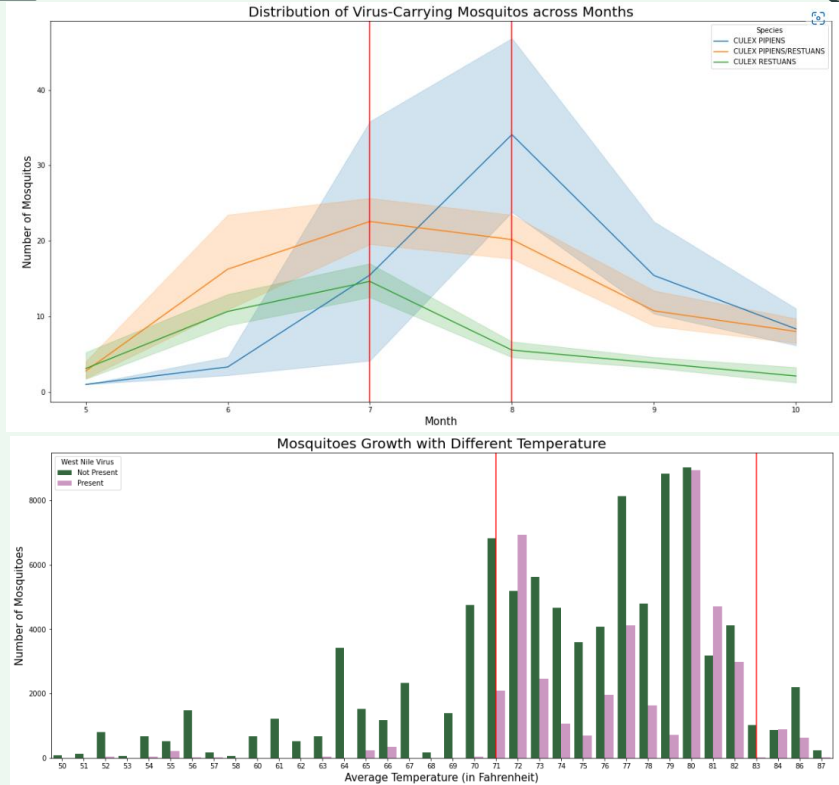
Dew Point Temperature

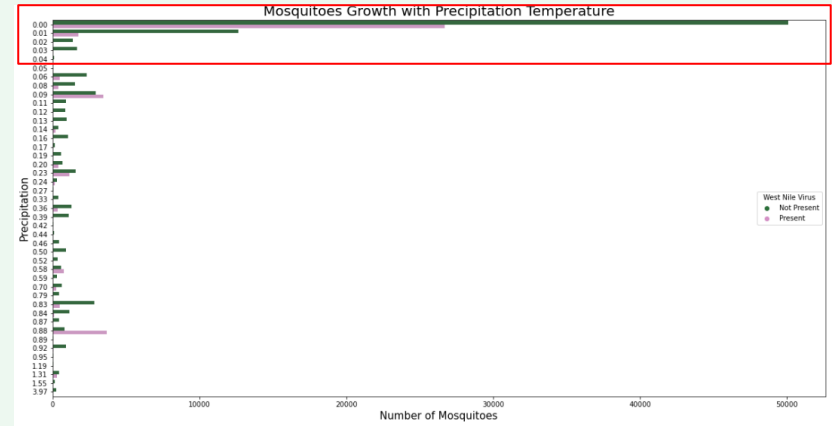
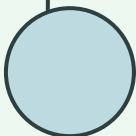
1. All mosquitoes species thrives with dewpoint between 49-61 Fahrenheit
1. Could increase resources when temperatures are within ideal conditions



Temperature Change

1. Mosquito growth starts to peak around July to August
1. Spike in Mosquito breeding with temperature between 71-85 Fahrenheit (21.6 - 29.4 Celsius)





Mosquitoes Locations

1. Station 2 mosquitoes count is doubled of station 1
1. Although station 1 count is lower, but WNV is growing/double of station 2



Note: High WNV count in 2007(station 2) is due to WNV present in same trap



03.

FEATURE ENGINEERING

LAGGED WEATHER FEATURES

```
#build lagged weather features

days = [1, 3, 5, 8, 12]

def buildLaggedFeatures(df, lag):
    new_dict={}
    for col_name in df:
        new_dict[col_name]=df[col_name]
        # create lagged Series
        for l in lag:
            if col_name!='Date' and col_name!='Station':
                new_dict['%s_Lag%d' %(col_name,l)]=df[col_name].shift(l)
    res=pd.DataFrame(new_dict,index=df.index)
    return res
```

**1, 3, 5, 8, and 12 days
lag in weather features**

To account for the life cycle of a mosquito, and for the delay required for WNV to be detected

RELATIVE HUMIDITY

```
#Function to calculate relative humidity
def fahrenheit_to_celcius(x):
    c = ((x - 32) * 5.0)/9.0
    return c

def relative_humidity(avg_temp, dew_point):
    a = 17.27
    b = 237.7
    avg_temp = fahrenheit_to_celcius(avg_temp)
    dew_point = fahrenheit_to_celcius(dew_point)
    Td_b = dew_point / b
    aT_bT = a*avg_temp / (b+avg_temp)
    ln_rh = Td_b*(a-aT_bT) - aT_bT / (Td_b + 1)
    return np.exp(ln_rh)
```

Measure of water vapour content in the air

A mosquito's survival rate reduces as relative humidity falls (3% survival rate at sub 10% Relative Humidity/during dry season)

$$T_{\text{dewpoint}} = \frac{b \left(\frac{aT}{b+T} + \ln RH \right)}{a - \left(\frac{aT}{b+T} + \ln RH \right)} \quad \text{where} \quad \begin{array}{l} a = 17.27 \\ b = 237.7 \\ RH = 0 \rightarrow 1 \end{array}$$

where the temperatures in the formula are in Celsius.

DARK HOURS

```
# Function to covert time to the equivalent float representation
def conv_time_to_float(timee):
    ## Extract the last two digits (as minutes)
    timee /= 100
    min_ = timee % 1
    ### Convert minute to decimal representation
    min_conv = min_ / .6

    ## Extract the first two digits (as hours)
    hour_ = round(timee - min_,0)

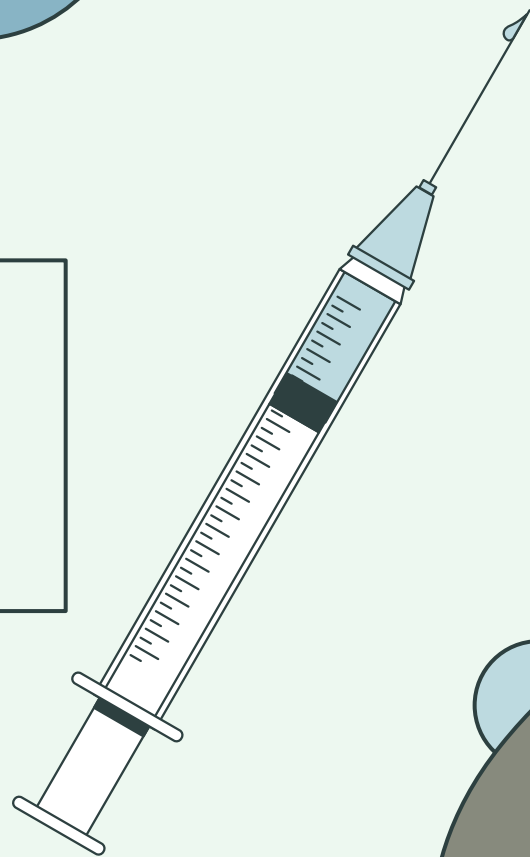
    ## Return float representation of the time
    return hour_ + min_conv
```

Hours in a day where it is night time

Mosquitoes avoids daylights to prevent dehydration from sun exposure, and are most active during the night

04.

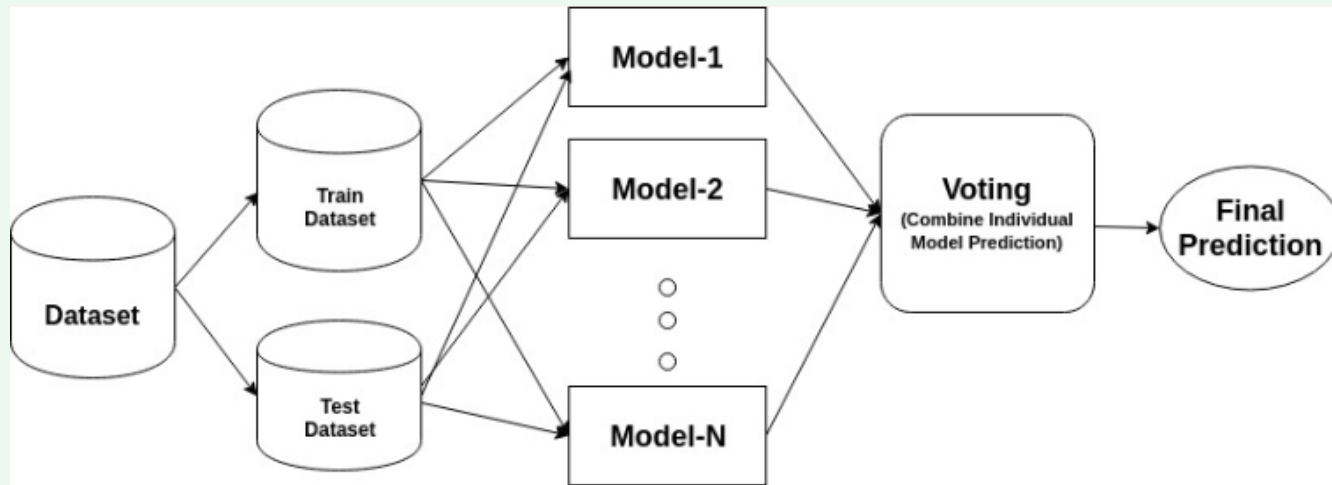
MODELLING



MODELLING

- Random Forest Classifier (Baseline)
- Adaptive (ADA) Boost Classifier
- Gradient Boost Classifier
- Stacking Classifier

RANDOM FOREST CLASSIFIER



RANDOM FOREST CLASSIFIER

TRAIN AUC

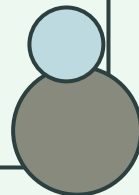


99.9%

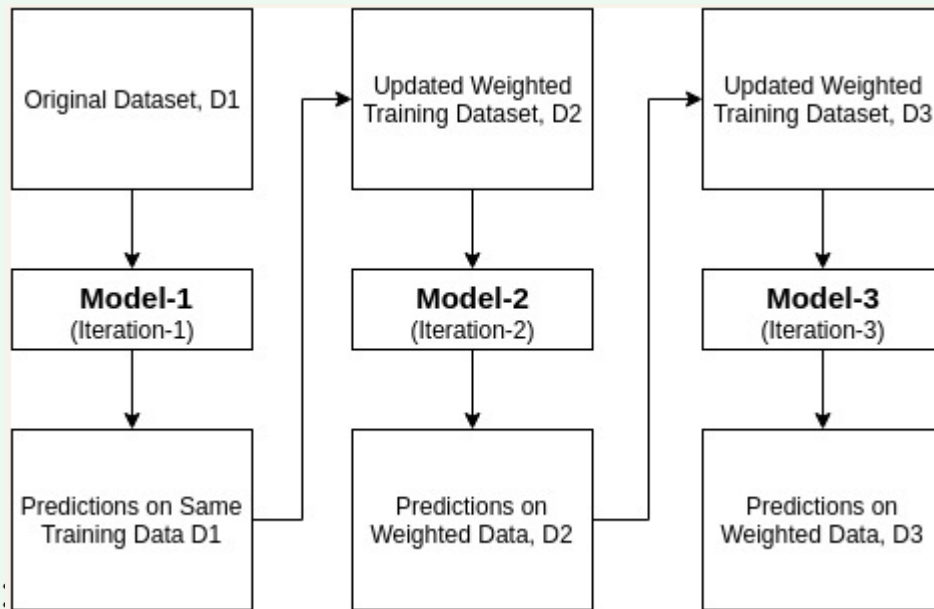
TEST AUC



93.82%



ADA BOOST CLASSIFIER



It's okay to make mistakes.

Mistakes are our teachers - they help us to learn

John Bradshaw

ADA BOOST CLASSIFIER

TRAIN AUC

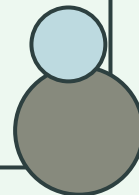


98.82%

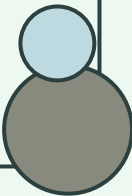
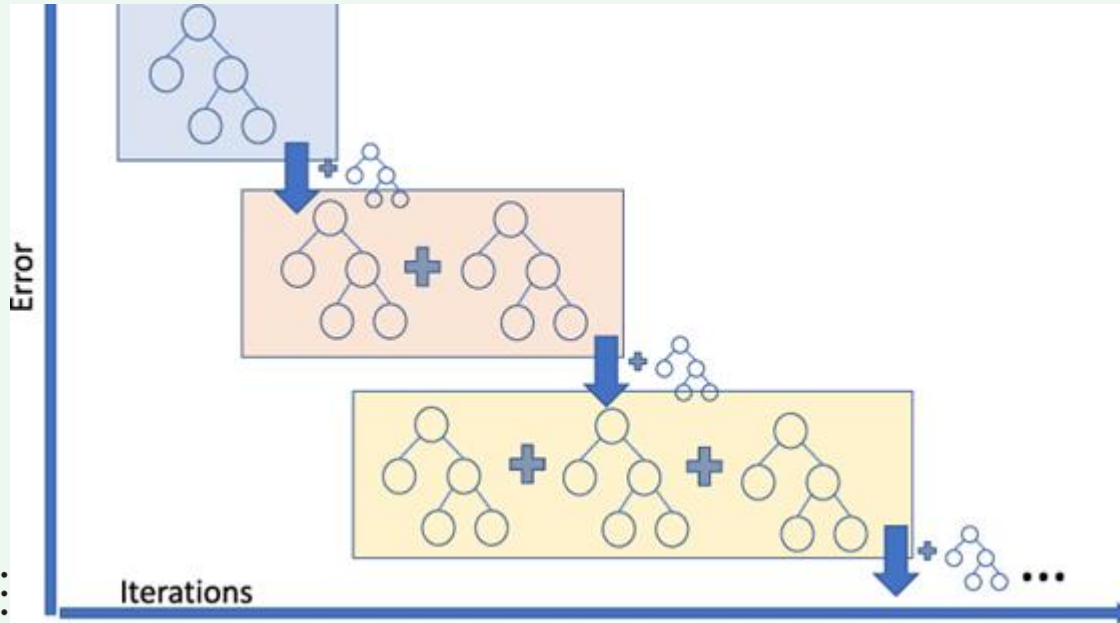
TEST AUC



93.6%



GRADIENT BOOSTING CLASSIFIER



GRADIENT BOOSTING CLASSIFIER

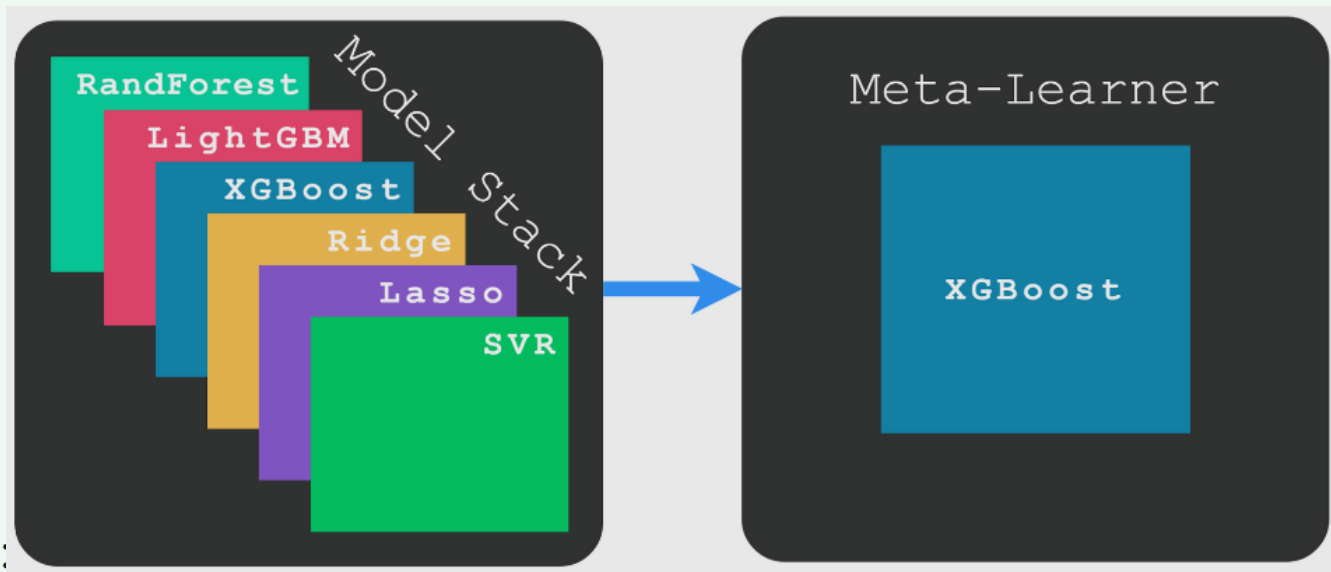
TRAIN AUC



TEST AUC



STACKING CLASSIFIER



STACKING CLASSIFIER

XGBRF CLASSIFIER

RANDOM FOREST CLASSIFIER

EXTRA TREE CLASSIFIER

GB CLASSIFIER

ADABOOST CLASSIFIER

XGBCLASSIFIER (GBTREE)

RIDGE CLASSIFIER



XGB CLASSIFIER



STACKING CLASSIFIER

TRAIN AUC

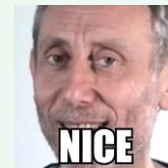
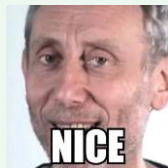


TEST AUC



MODEL EVALUATION

	RANDOM FOREST	ADABOOST CLASSIFIER	GRADIENT BOOSTING	STACKING CLASSIFIER
TRAIN	99.99%	98.82%	92.57%	98.6%
TEST	93.82%	93.6%	92.715%	97.85%
UNSEEN	67.85%	72.72%	60.91%	71.33%



MODEL SELECTION

Run time = 30 secs

YOUR RECENT SUBMISSION



adaboostclassifier.csv

Submitted by Dataismybytych · Submitted 2 minutes ago

0.71132

0.72725

↓ Jump to your leaderboard position

Run time = Let's not talk about it (30min)

YOUR RECENT SUBMISSION



ensemble.csv

Submitted by Dataismybytych · Submitted 4 minutes ago

0.70923

0.71332

↓ Jump to your leaderboard position

MODEL SELECTION

YOUR RECENT SUBMISSION



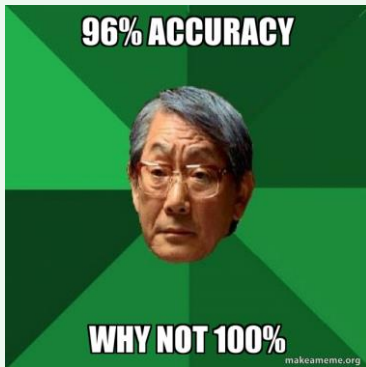
adaboostclassifier.csv

Submitted by Dataismybytch · Submitted 2 minutes ago

0.71132

0.72725

↓ Jump to your leaderboard position





05.

Cost-Benefit Analysis



Costs




Direct costs

- Procurement of adulticides (275-gal Zenivex e20)
- Healthcare costs

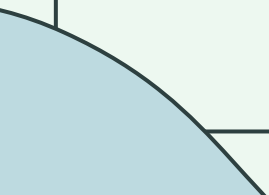


Indirect costs

- Productivity loss (severe symptomatic patients)



How much will it cost to spray Zenivex aduaticide across the entire of Chicago annually?

- Land Area of Chicago: ~ 145,745 acres (606 km²)
 - Cost of Zenivex per acre: \$0.67
 - Cost of spraying Chicago (fortnightly) in a year: ~ **\$2.5 million**
- 

Station locations



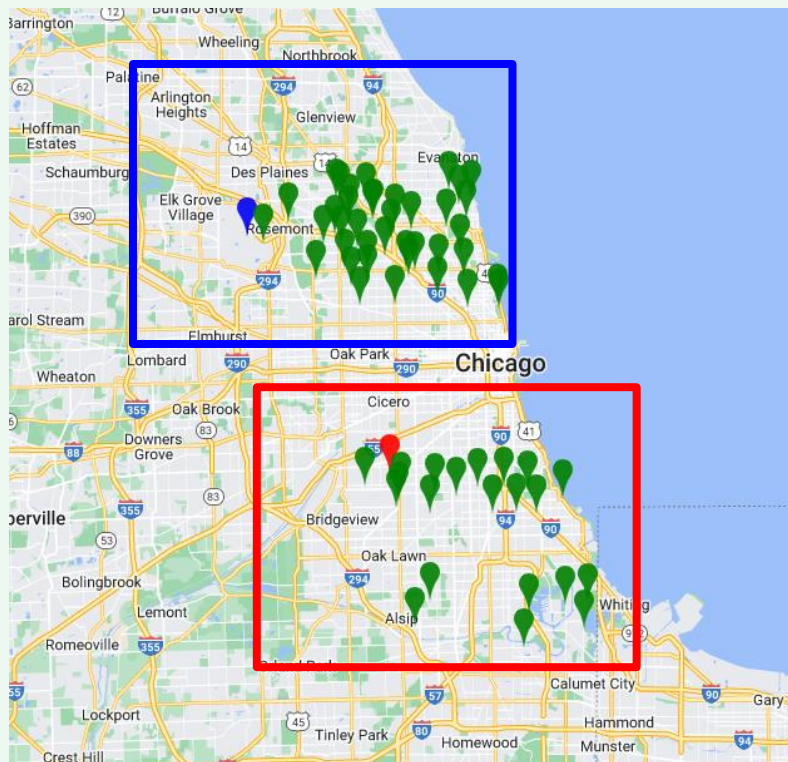
Surrounding area of station 1



Surrounding area of station 2



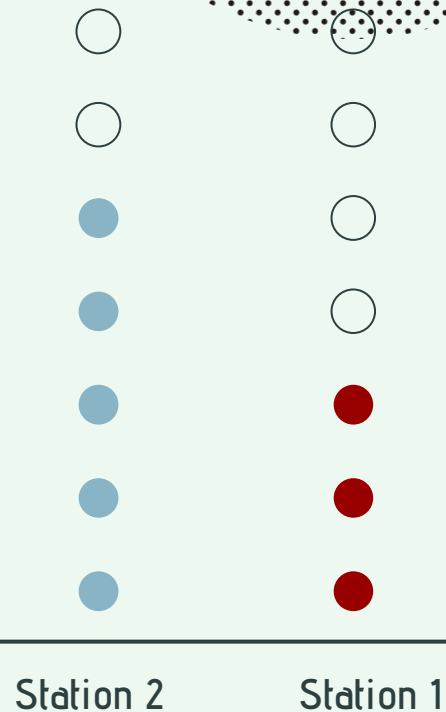
Traps deployed



WNvPresent count by Stations

**The WNV is more prevalent
In Station 2 than Station 1**

Wnv Present



How much will it cost to spray Zenivex adjuvant only at areas surrounding each station?

- Our model predicted Station 2 to capture the presence of WNV at a higher incidence.
- Surrounding Land Area of Station 1 & 2: 77 839 acres (315km²)
- Cost of Zenivex per acre: \$0.67
- Cost of spraying targeted trap locations (fortnightly): **~\$1.36 million**

High medical costs and productivity loss due to severe WNV cases

	Per WNV case#	60^ WNV cases / year
Hospitalization Cost+ per WNV case	\$40,000	\$2,400,000
Productivity Loss* per WNV case	\$11,000	\$660,000
Total Costs	\$ 51,000	\$3,060,000

* Productivity loss refers to those incurred by both patients/caretakers.

+ Hospitalization cost refers to inpatient costs, outpatient costs and long-term medical costs.

WNV cases here refer to patients who have developed neuroinvasive symptoms from WNV and require medical attention.

^ Chicago Dept of Health summary report of the 2012 season

Comparison of Direct/Indirect Costs

\$3.06 million (indirect costs) - \$1.36 million (direct costs) =
\$ **1.7million**

Yes - this project is financially feasible.



CONCLUSION

1

Best Model: ADABOOST Classifier

- High AUC ROC score on test set: 0.96
- High AUC ROC score on Kaggle: 0.72
- Prediction can be further enhanced with more recent datasets

2

Benefits of spraying Zenivex aduicide outweigh its costs

- Minimise high medical costs and productivity loss to the community
- Cost-benefit analysis can be further improved with more data points e.g. impact of neighbourhood types (residential or industrial) and presence of known water basins/ponds/drains where mosquito breeding is more likely to occur; impact of larvicide

NEXT STEPS



Collaborate with meteorologists and researchers to further investigate the impact of weather and seasons on occurrence of WNV



Improve data collection on adulticide sprays conducted in the city to accurately track the effectiveness of sprays in curbing the spread of WNV



Engage with research scientists to further investigate the virus-carrying mosquito species (Culex Pipiens & Culex Restuans) to find other methods to effectively inhibit their growth and spread



THANK YOU

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Storyset**

References

- https://www.chicago.gov/content/dam/city/depts/cdph/statistics_and_reports/CDInfo_2013_JULY_WNV.pdf
- <https://www.cdc.gov/westnile/statsmaps/cumMapsData.html#one>
- <https://dph.illinois.gov/topics-services/diseases-and-conditions/west-nile-virus.html>