A DATA-DRIVEN APPROACH AGAINST WEST NILE VIRUS IN CHICAGO

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

DSI-28
10 Jun 2022

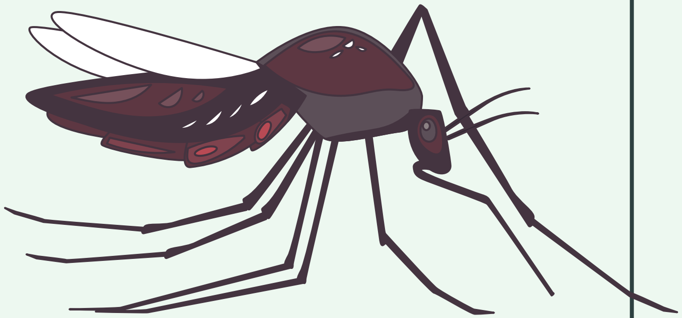
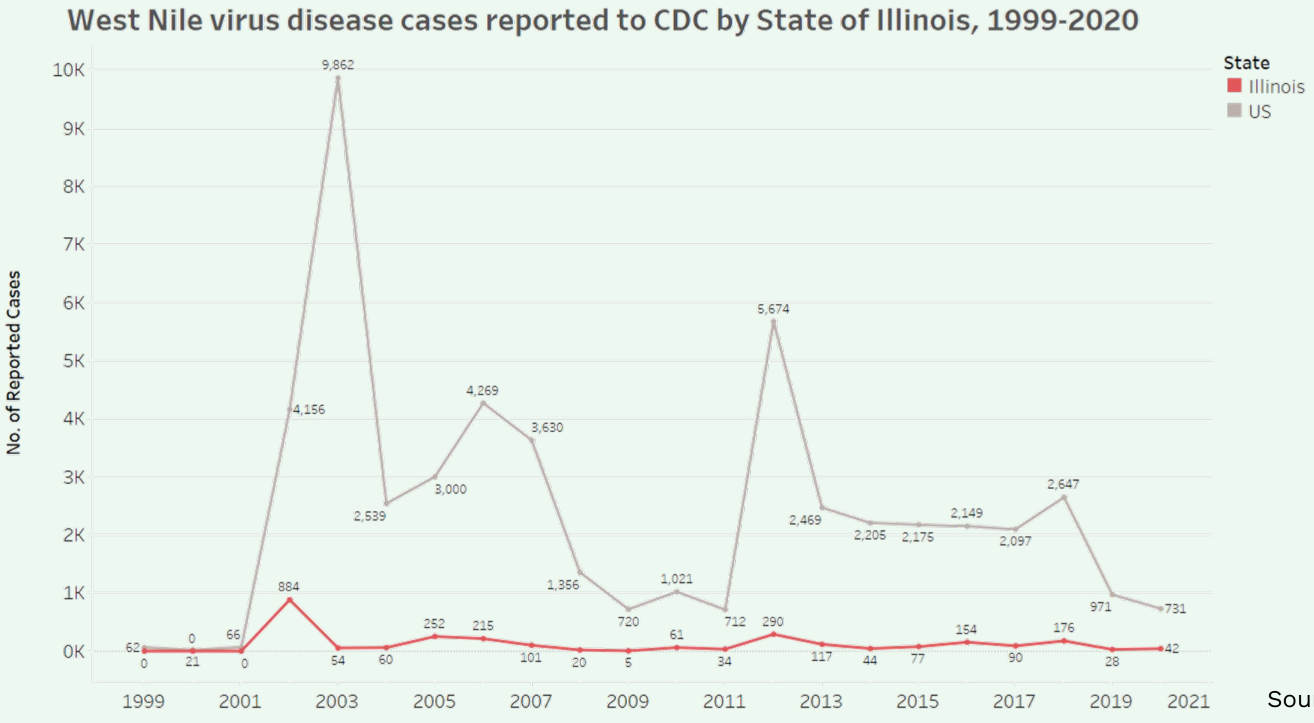


TABLE OF CONTENTS

01.

BACKGROUND

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



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 take proactively reduce no. of mosquitos in residential 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

02.

Data Cleaning & EDA

Scope of Data

1. **Train Set**
   - Contains data from 2007 - 2013 (2 years interval)
   - No missing values
2. **Test Set**
   - Data from 2008 - 2014 (2 years interval)
   - No missing values
3. **Spray Set (Will not use)**
   - Contains data from 2011 and 2013
   - Missing values in spray timing
4. **Weather Set**
   - Contains data from 2007-2014 (8 years)
   - Missing values in many columns

Data Cleaning Process

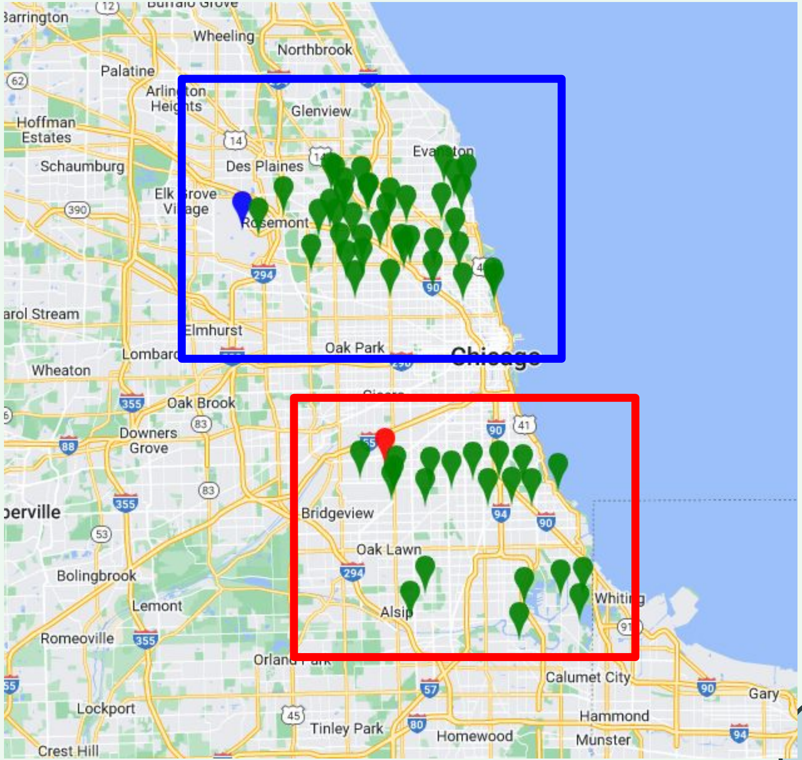
1. Dropped non-essential columns

2. Combined duplicated rows due to mosquitoes count limit

3. Assignment of stations based on each point’s lat and lon using Haversine Distance

Haversine Distance

- Station 1
- Station 2
- Traps deployed



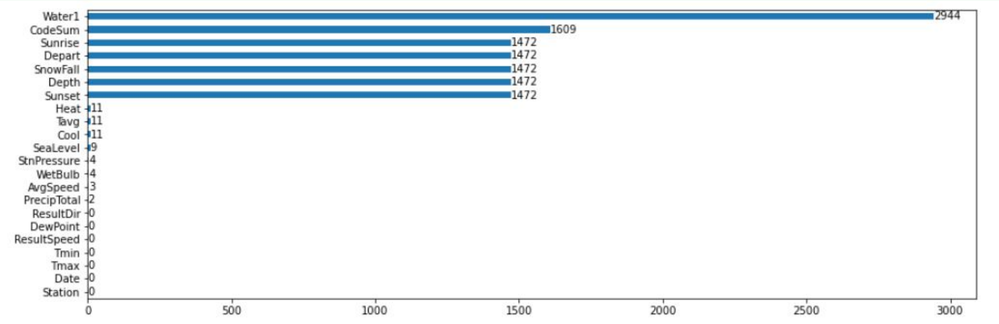
Missing Values

1. Dropped columns with > 80% data missing

2. Forward-fill sunrise and sunset

3. Filling average temp with Min and max

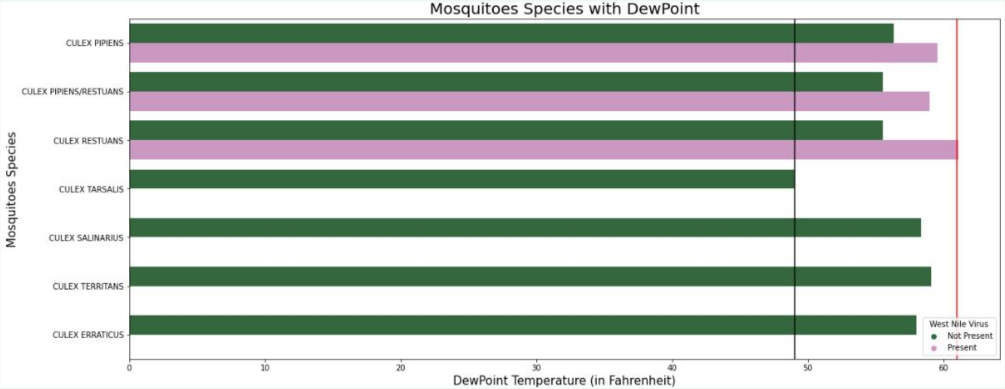
4. Mean for <0.3% data



Dew Point Temperature

1. All mosquitoes species thrives with dewpoint between 49-61 Fahrenheit (9.4-16 Celsius)

2. Could increase resources when temperatures are within ideal conditions

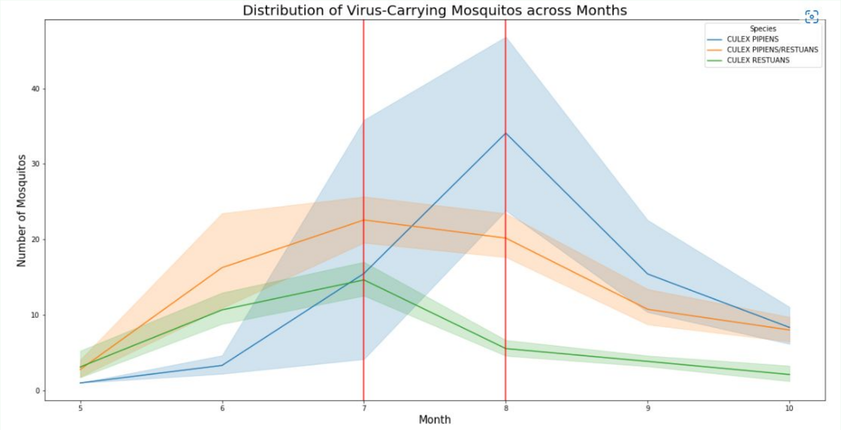


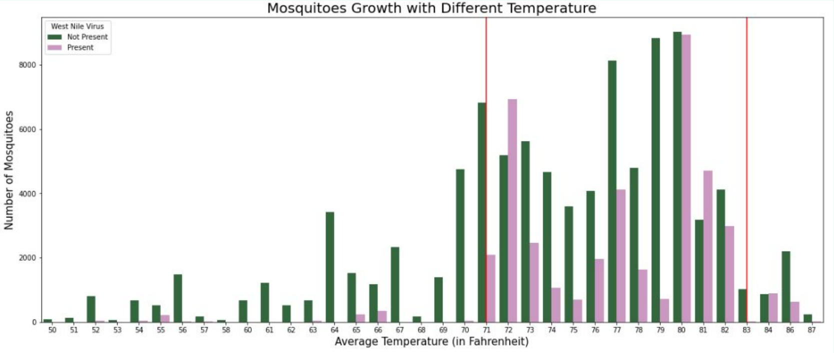
Seasonality

1. Chicago warm season starts from June to September.

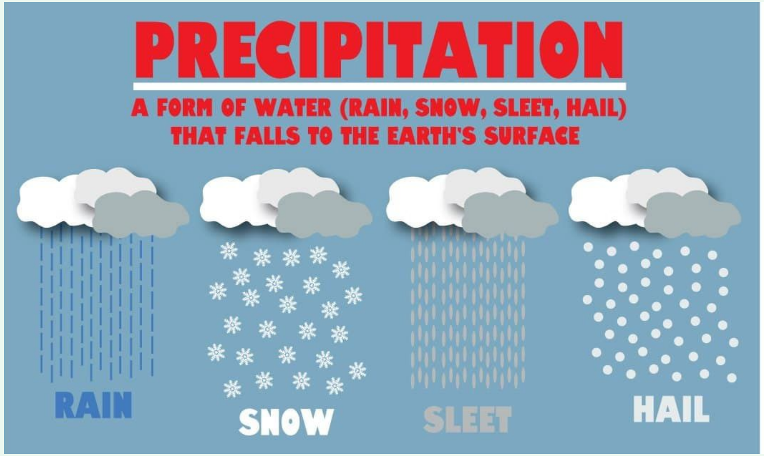
2. Mosquito growth starts to peak around July to August

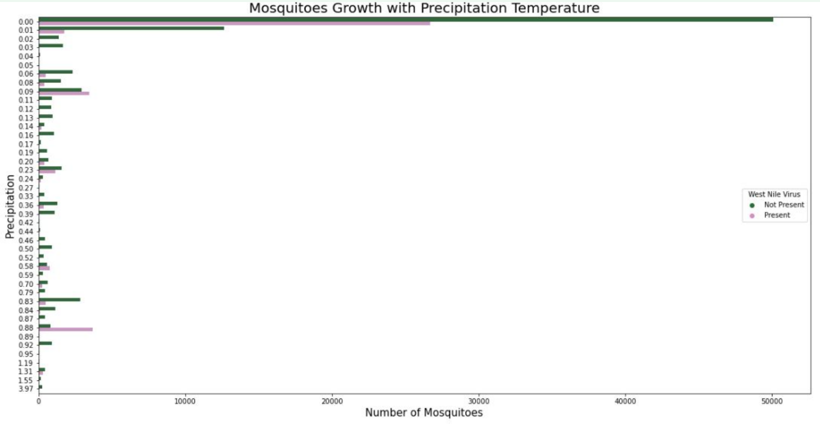
3. Spike in Mosquito breeding with temperature between 71-85 Fahrenheit (21.6 -29.4 Celsius)





Precipitations



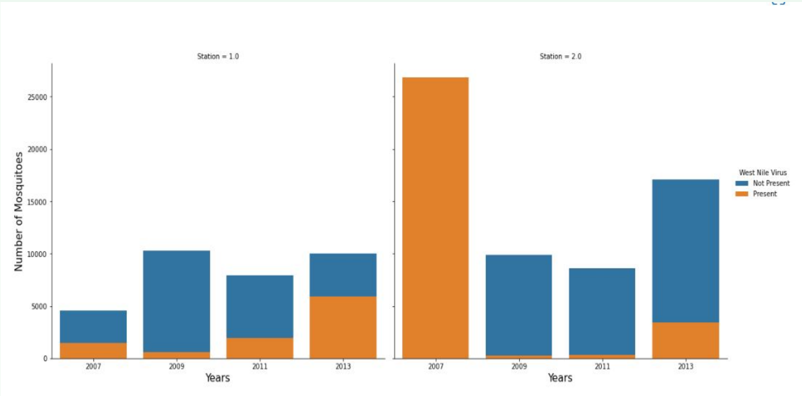


Mosquitoes Locations

1. Station 2 mosquitoes count is doubled of station 1

2. 2007 has the highest amount of mosquito caught.
   - High WNV count is due to WNV present in same trap

3. Although station 1 count is lower, but WNV is growing/double of station 2



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 farenheit_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 = farenheit_to_celcius(avg_temp)
    dew_point = farenheit_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)

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

where the temperatures in the formula are in Celsius.

DARK HOURS

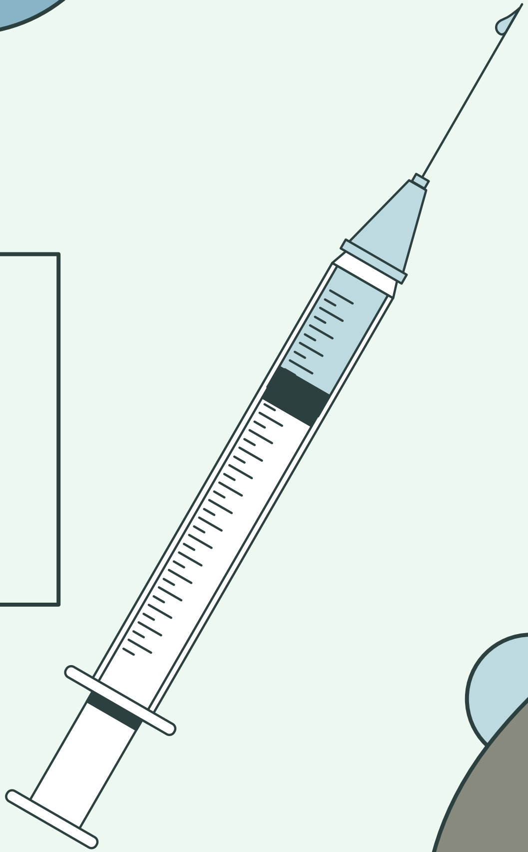
```
# Function to covert time to the equivalent float representation
def conv_time_to_float(timee):
    ## Extract the last two digits (as minutes)
    timee /= 100
    min_ = timee % 1
    ### Convert minute to decimal representation
    min_conv = min_ / .6

    ## 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.

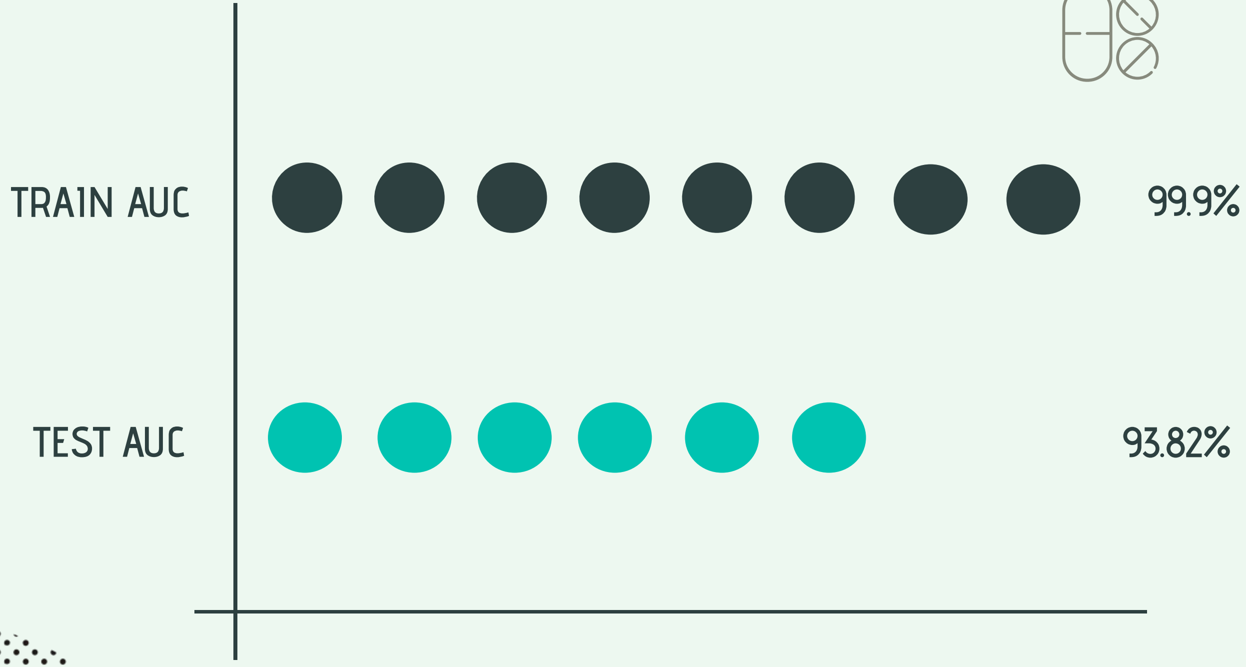
MODELLING



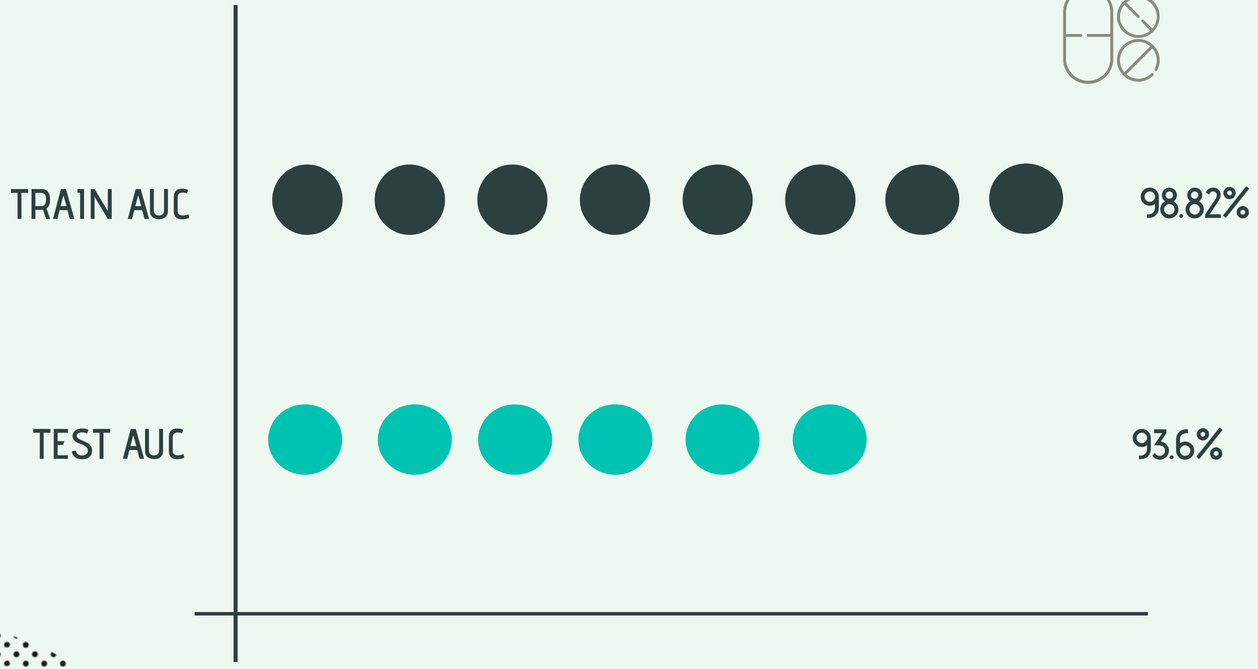
MODELLING

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

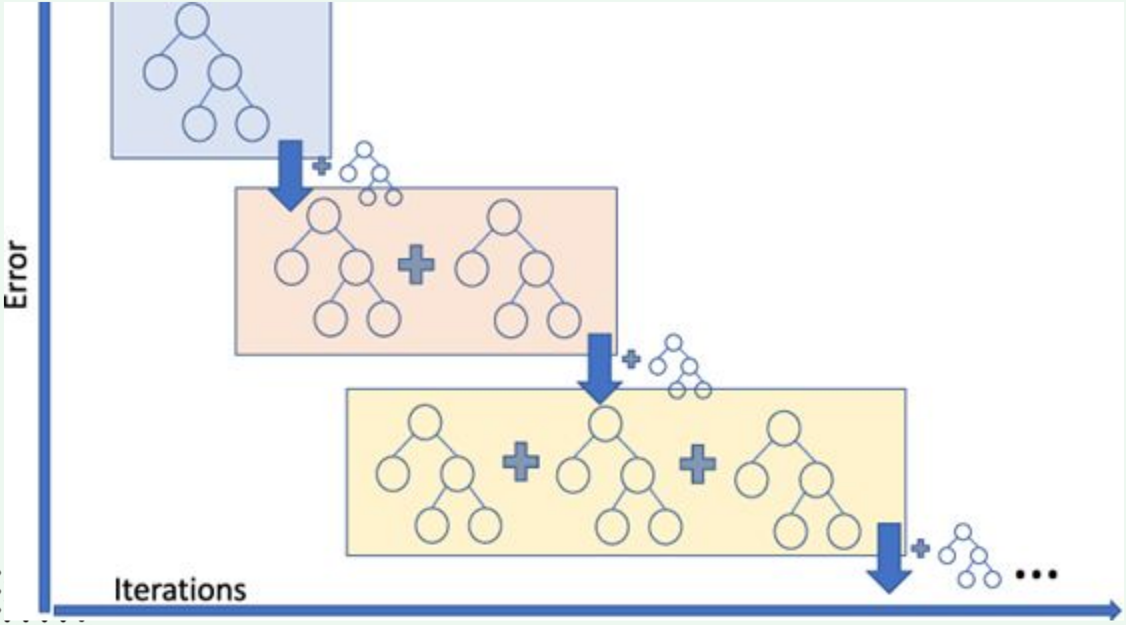
RANDOM FOREST CLASSIFIER



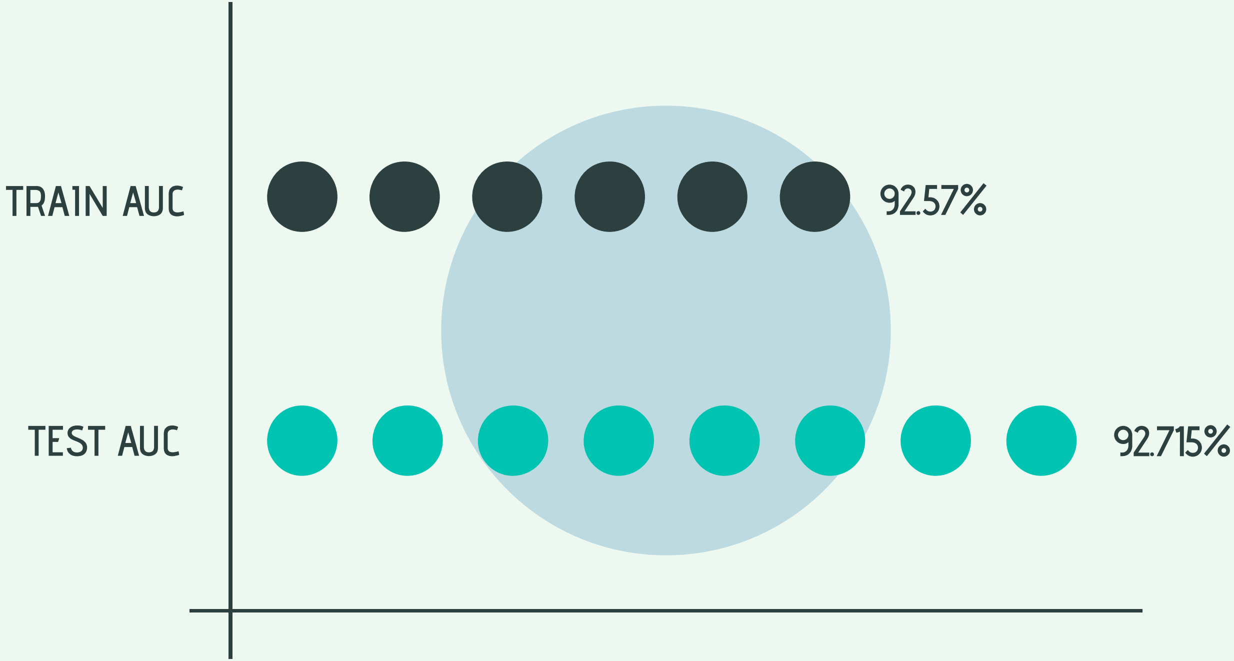
ADA BOOST CLASSIFIER



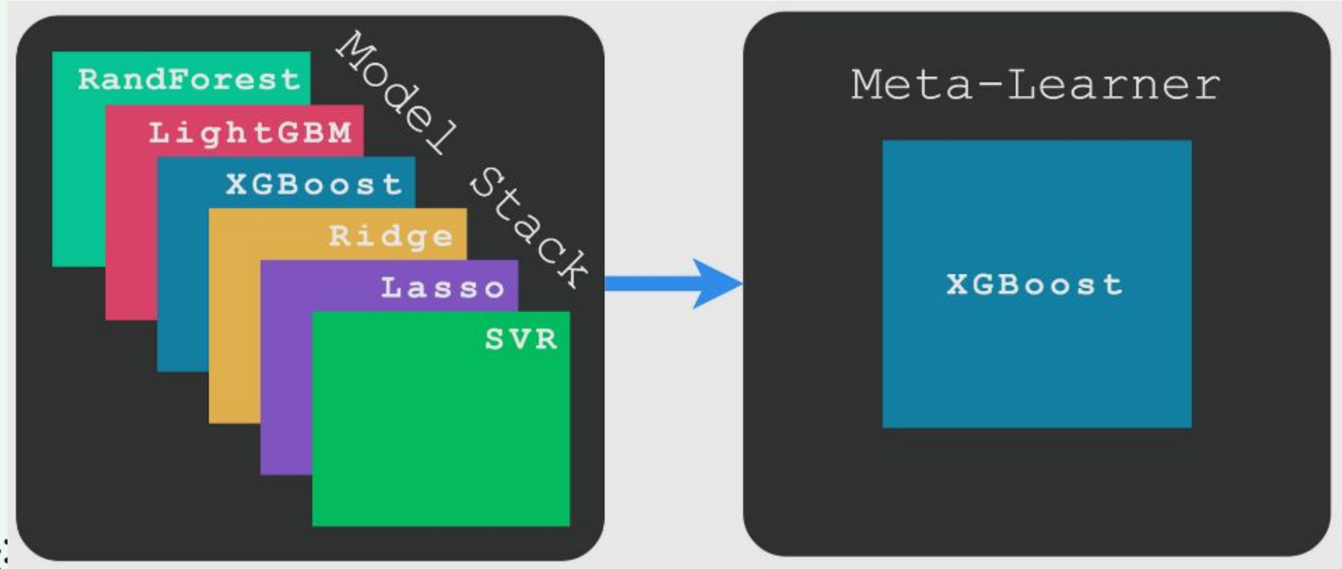
GRADIENT BOOSTING CLASSIFIER



GRADIENT BOOSTING CLASSIFIER



STACKING CLASSIFIER



STACKING CLASSIFIER

XGBRF CLASSIFIER

RANDOM FOREST CLASSIFIER

EXTRA TREE CLASSIFIER

GB CLASSIFIER

ADABOOST CLASSIFIER

XGBCLASSIFIER (GBTREE)

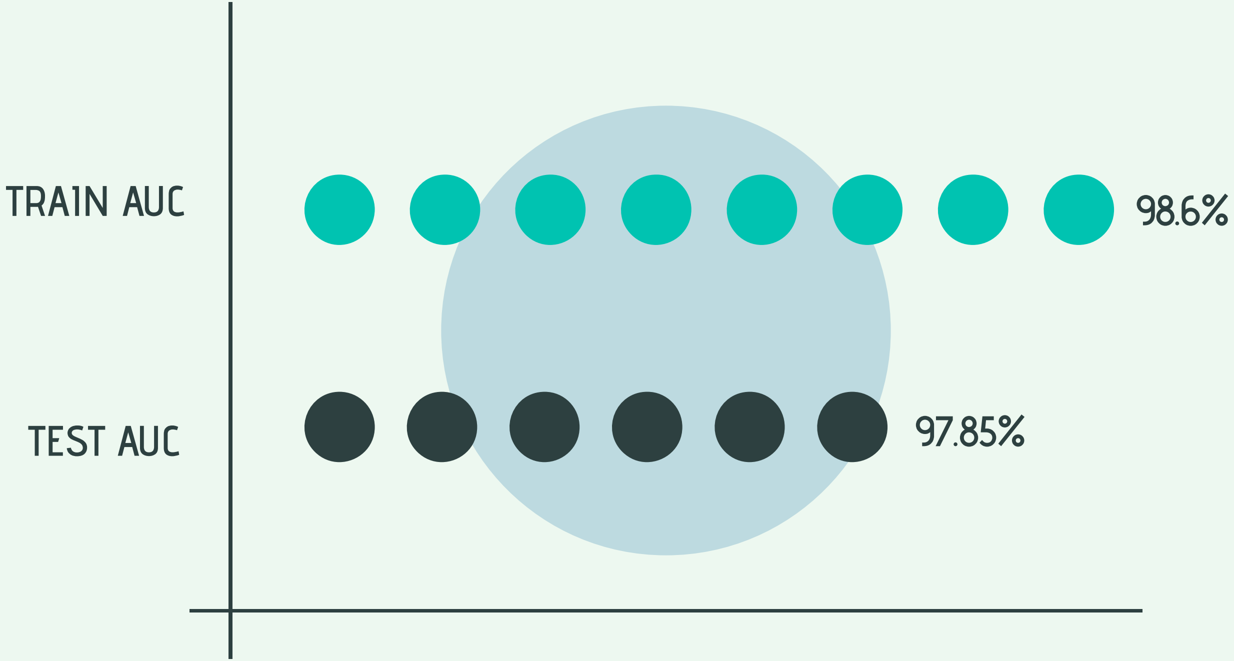
RIDGE CLASSIFIER

→

→

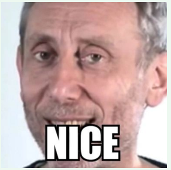
XGB CLASSIFIER

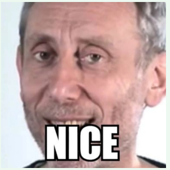
STACKING CLASSIFIER



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%
KAGGLE SCORE	67.85%	72.72%	60.91%	71.33%





Kaggle Submission

Run time = 30 secs

YOUR RECENT SUBMISSION

adaboostclassifier.csv

Submitted by Dataismybytch · 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

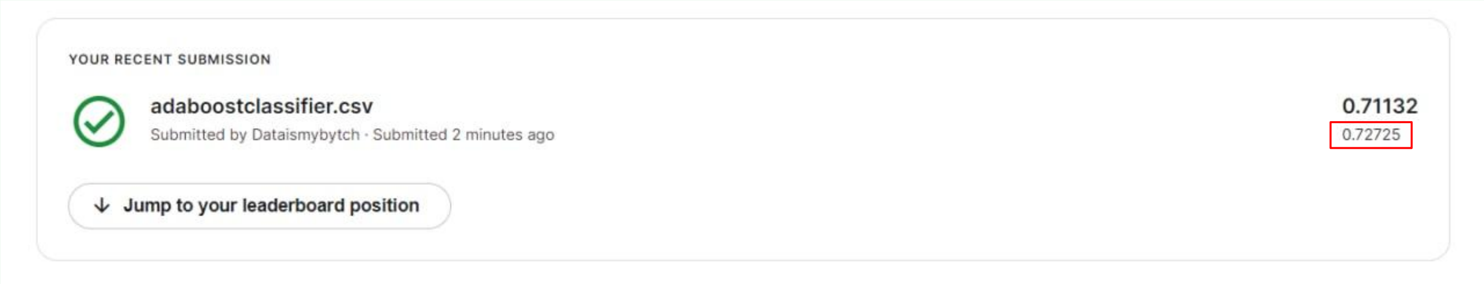
Submitted by Dataismybytch · Submitted 4 minutes ago

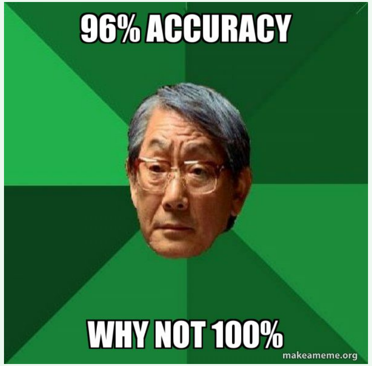
0.70923

0.71332

↓ Jump to your leaderboard position

Model Selection





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 adulticide across the entire of Chicago annually?

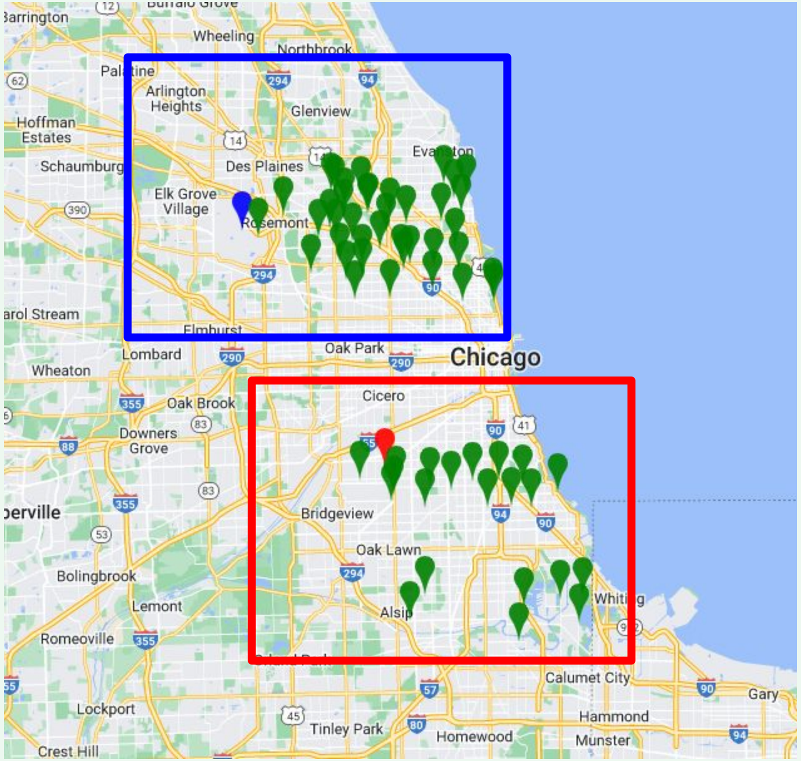
- Land Area of Chicago: ~ 145,745 acres (606 km2)
- Cost of Zenivex per acre: $0.67
- Cost of spraying Chicago (fortnightly) in a year: **~ $2.5 million**

Station locations

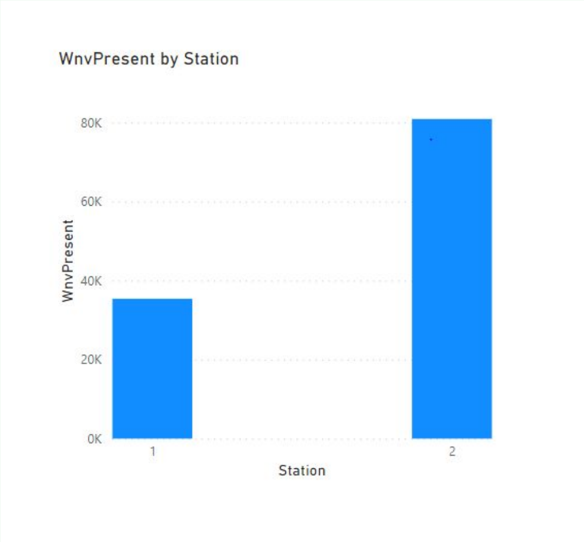
Surrounding area of station 1

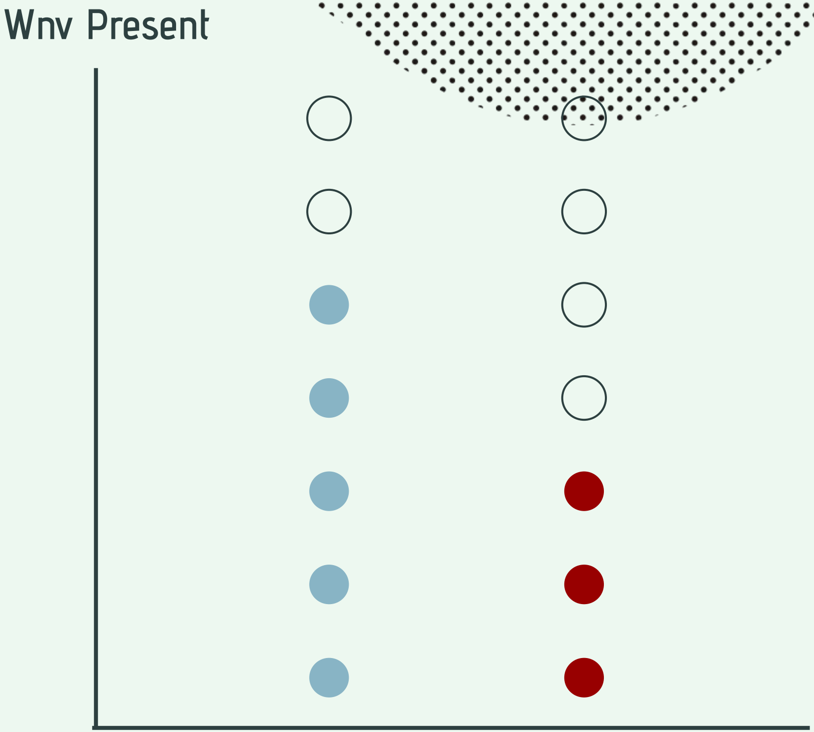
Surrounding area of station 2

Traps deployed



WNvPresent count by Stations





Station 2

80,000 counts

Station 1

37,000 counts

How much will it cost to spray Zenivex adulticide 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.
^ Mean

Comparison of Direct/Indirect Costs

$1.36 million (direct costs) **vs** $3.06 million (indirect costs)

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 adulticide outweigh its costs

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

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