Predictive Analytics on

West Nile Virus in Chicago

Project team:
Asyraf, Ben D, Jack, Sahaj

Agenda

1

Background

What did we do and why?

2

Data analysis

What did we find and how did we do it?

3

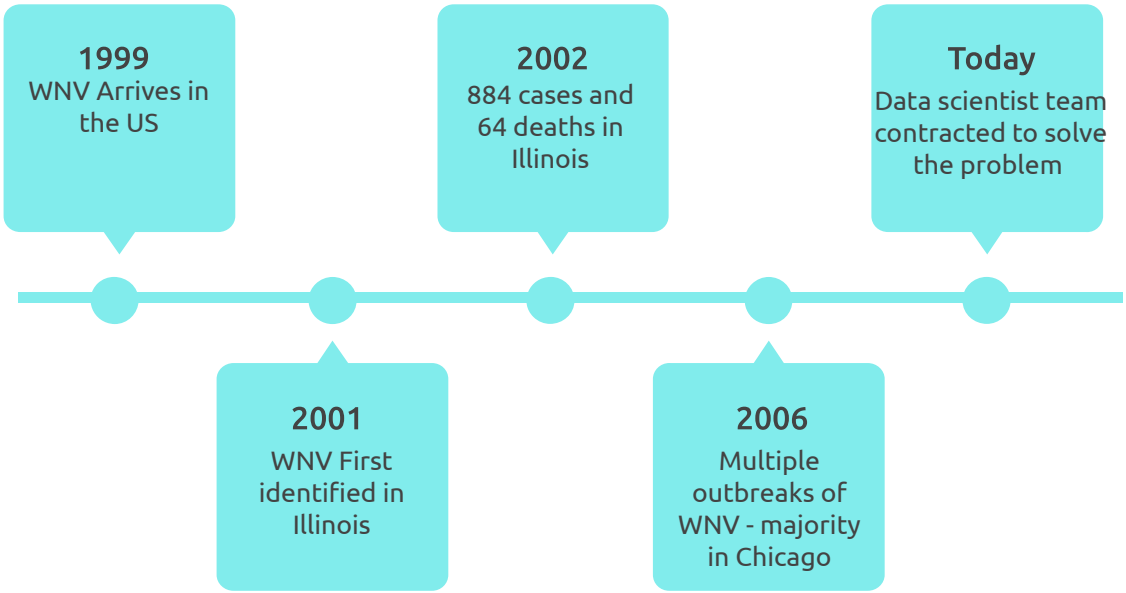
Conclusions

What should be done next?

Context

As a result of the recent outbreak of West Nile Virus in Chicago, we have been contracted to help understand the problem and propose a way forward.

An illustrative timeline of the West Nile Virus



52,000*

Cases of WNV in the US till date

Source: https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

Virus Facts

- West Nile virus can cause a fatal neurological disease in humans.
- However, approximately 80% of people who are infected will not show any symptoms.
- Birds are the natural hosts of West Nile virus.
- West Nile virus is mainly transmitted to people through the bites of infected mosquitoes.
- The virus can cause severe disease and death in horses.
- Vaccines are available for use in horses but not yet available for people.

Source: https://www.who.int/news-room/fact-sheets/detail/west-nile-virus

Overview of the datasets for this project

Weather Conditions

- 2007-2014
- Temperature highs + lows

Spraying Efforts

- 2011 and 2013
- Date and location of spray

Mosquito Trap Surveillance

- 2007-2014
- Location + number of mosquitos carrying WNV

Some data was cleaned and missing data was imputed based on reasonable inference

What are the key issues we need to understand?

SPREAD

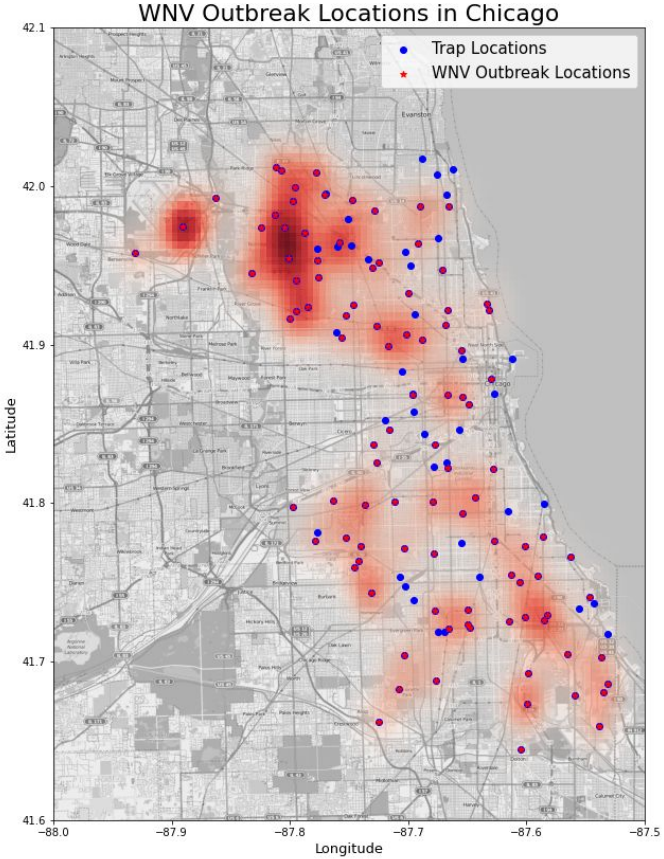
What are the contributing factors leading to spread of the virus?
Which mosquito species carry the virus?
How does it differ over the months and years?
Which parts of the city is it more concentrated in?

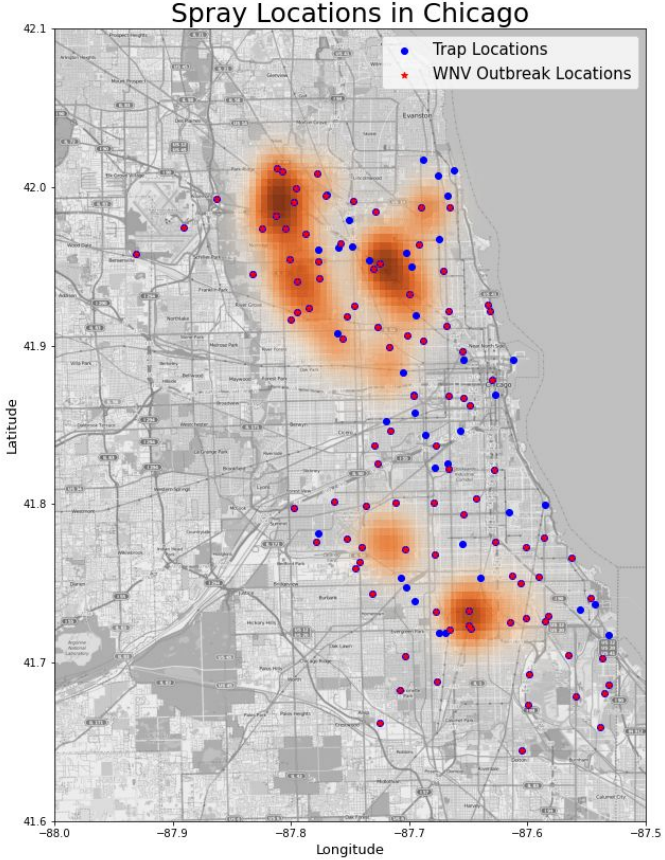
CONTROL

How can we reduce the incidence of the virus in Chicago?
What are the best strategies for controlling the virus spread?
What are the various trade-offs that need to be made and why?

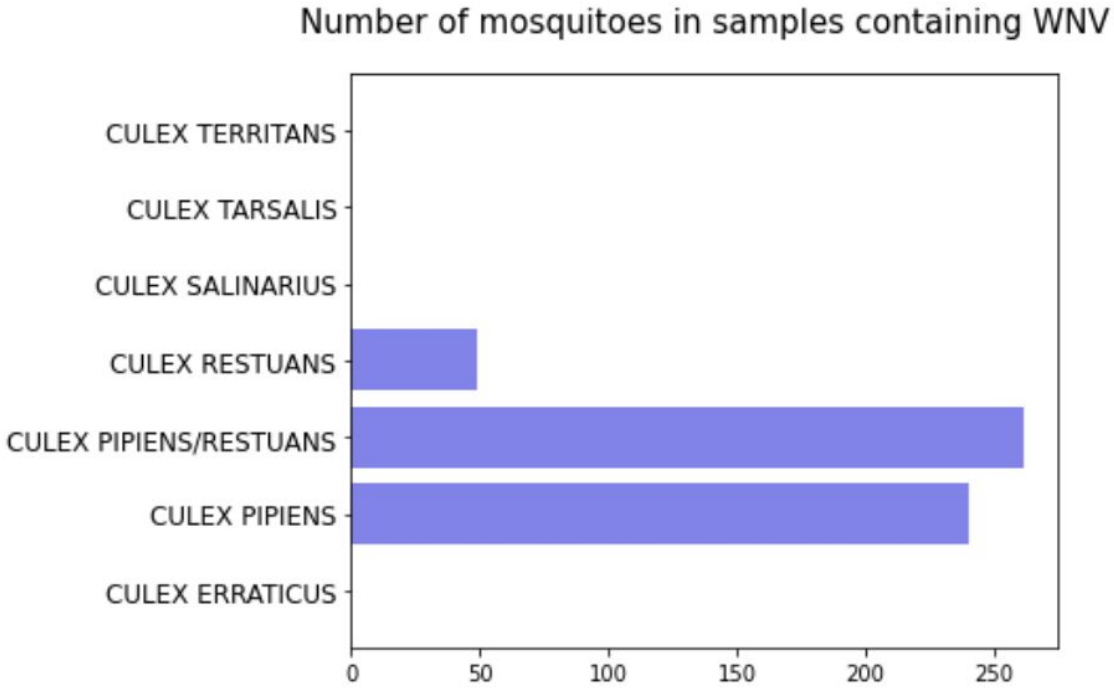
Data Analysis

Let’s see what we found

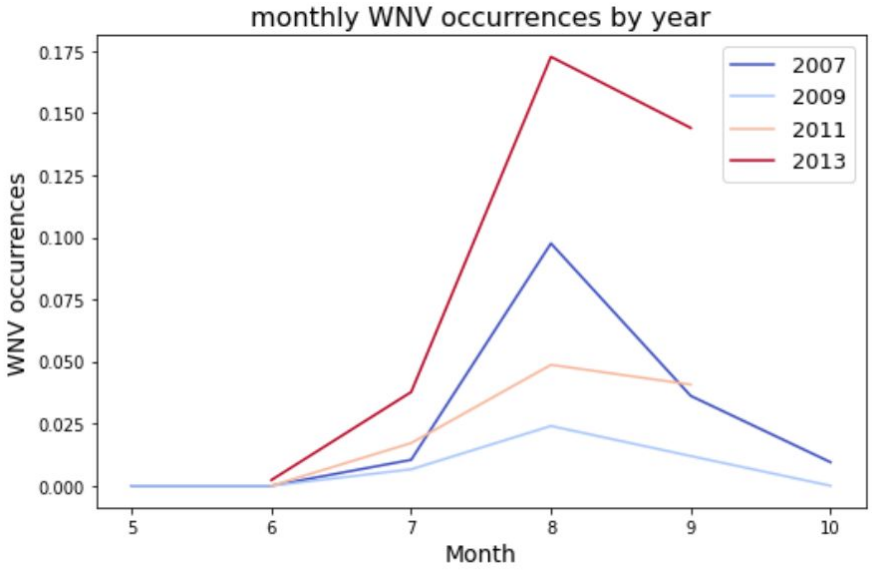

WNV Outbreak Locations in Chicago
Trap Locations
WNV Outbreak Locations
Latitude
Longitude
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5


Spray Locations in Chicago
Trap Locations
WNV Outbreak Locations
Latitude
Longitude
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

Only 2 out of 6 species sampled carry the virus



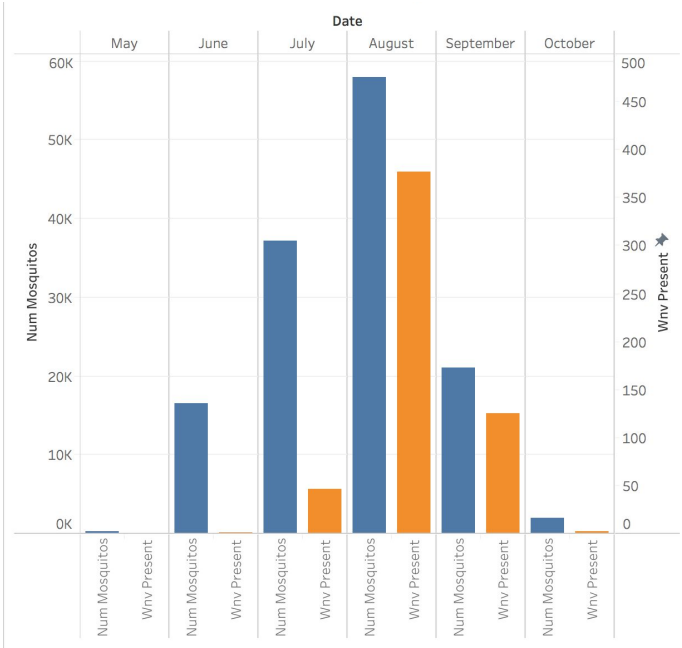
WNV seems to follow a pattern with spikes mid-year



- The presence of WNV in mosquitoes spikes in August.
- 2013 stands out with the highest number of WNV incidences
  - Simply a coincidence that Chicago went through a large heat wave* in 2013?
- 2009 has the lowest number of WNV cases - probably due to the temperature, as it ranked the 30th coldest August on record.
- Ultimately, the higher number points to an increase in temperature during the July - September period.
  - Positively correlated to temperature

*Source: https://www.chicagotribune.com/news/breaking/ct-heat-wave-chicago-met-20160719-story.html

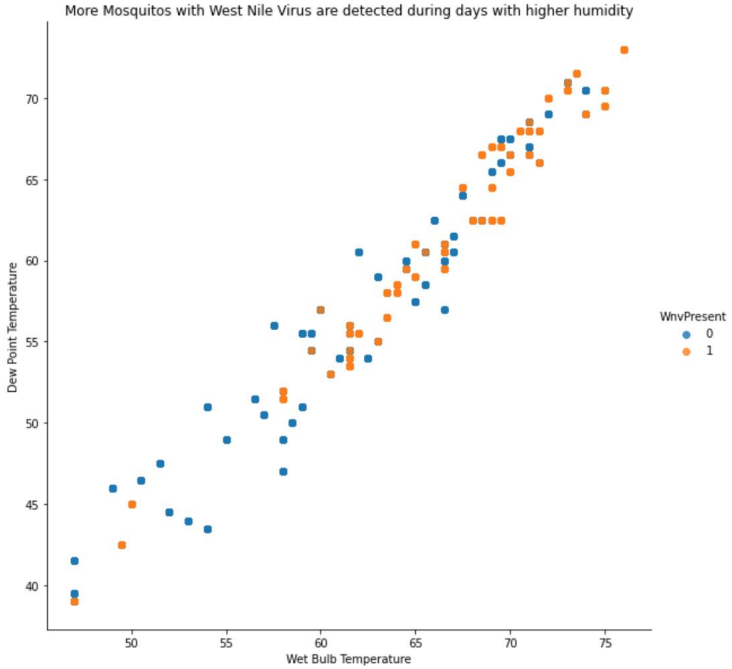
August appears to be the month with highest incidence of WNV

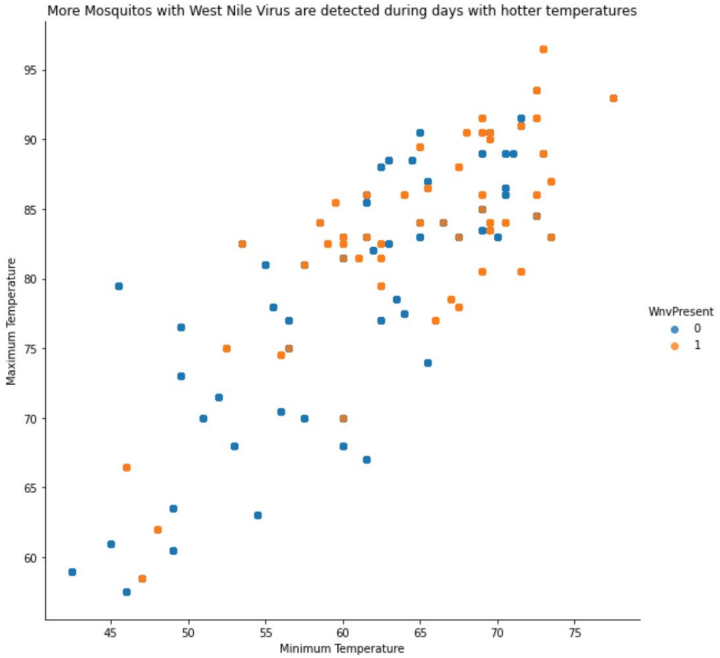


- Not surprising that the presence of WNV in mosquitoes is the highest in August.
- June and July are the peculiar months as the presence of WNV are disproportionate to the number of mosquitoes present.
  - Existence of external factors (birds migration pattern in North America*)?
  - Lag period of a month for mosquitoes to interact with virus. (incubation period)?

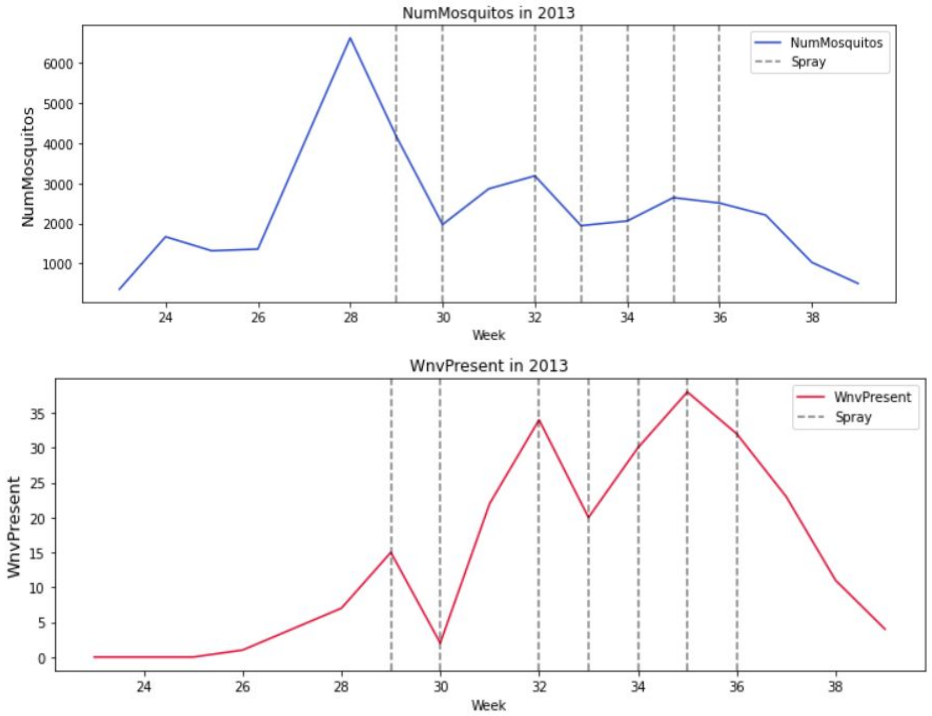
*Source: https://www.nationalgeographic.com/magazine/2018/03/bird-migration-interactive-maps/

There is a positive relationship between WNV and humidity/ temperature





Spraying seems to have a negligible effect- are there avenues for improvement?



Spraying appears to reduce number of mosquitoes but not the virus incidence.

Possible hypotheses for this -

1) Improper spray targeting
2) Incorrect timing of sprays (proactive vs reactive)

To be explored further upon model evaluation

Data Modelling

Steps and findings

Machine learning workflow

Step 1

Prepare Data

Feature engineering, train test split.
Benchmark and baseline models.

Step 2

Initial Run

Ran models on the data without any over or under sampling methods

Step 3

Balancing the data

Using SMOTE to address the class imbalance within our target variable

Step 4

Final Models

Hyperparameter tuning and identifying the top performing model based on AOC-RUC curve

Feature Engineering

Weather

1. Relative Humidity
2. Winter Temperature
3. Summer Temperature
4. CodeSum -> rain & fog/mist

Location

1. Traps (One Hot)
2. Species (Re-map)

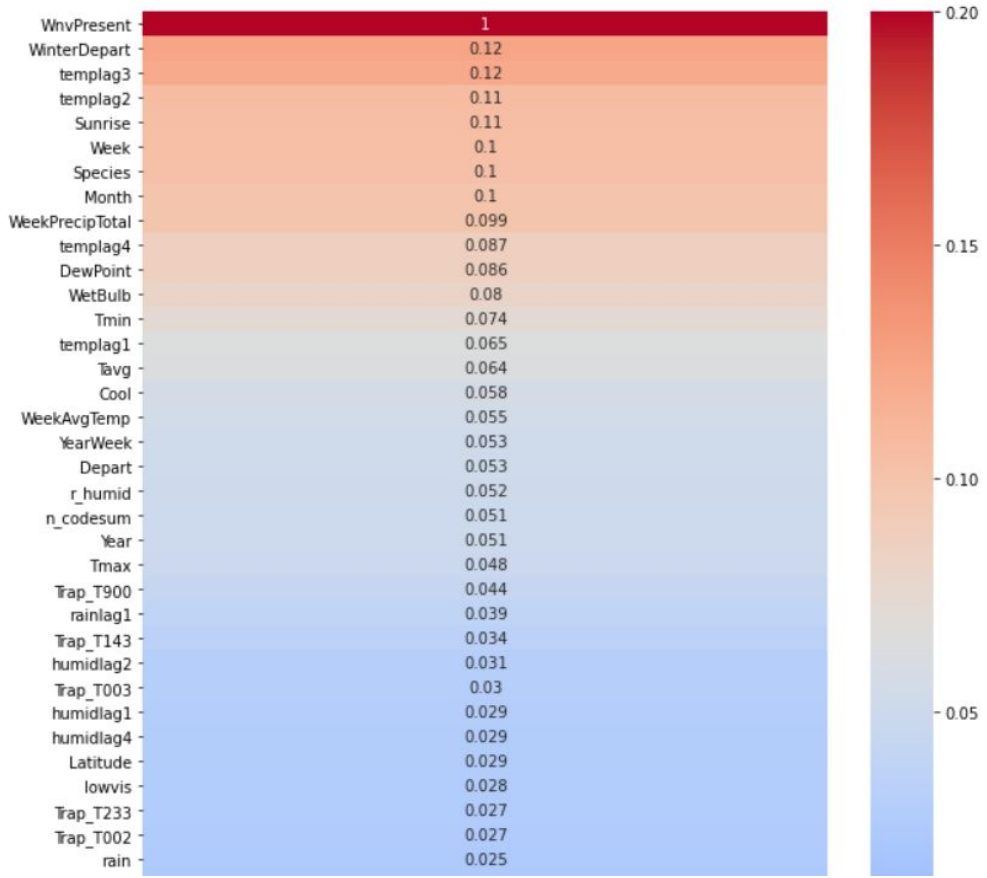
Time/Lag Features

1. Average Weekly Temperature
2. Cumulative Weekly Precipitation
3. Average Weekly Relative Humidity
4. Daily Number of CodeSum

Species Mapping

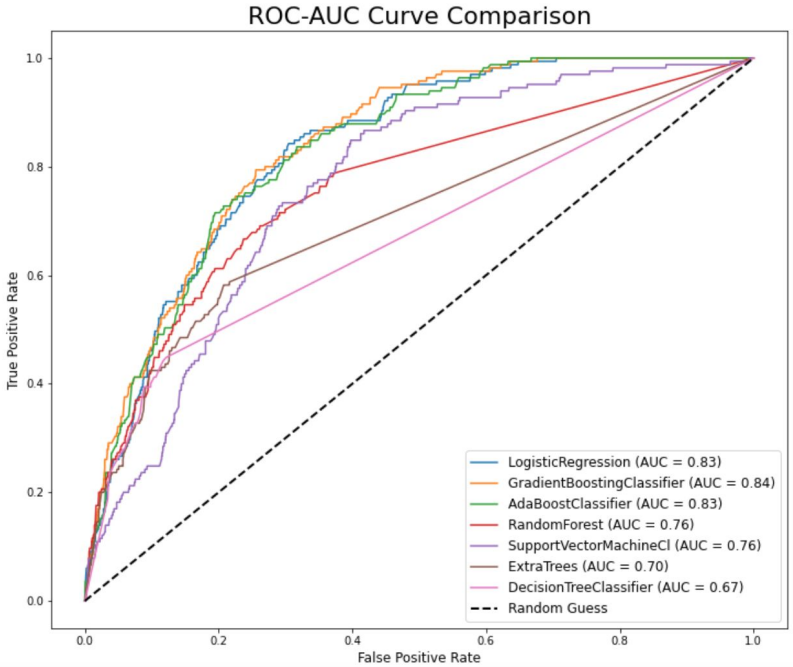
Species	NumMosquitos	WnvPresent	
CULEX ERRATICUS	7	0	
CULEX PIPIENS	44671	240	**0.005**
CULEX PIPIENS/RESTUANS	66268	262	**0.004**
CULEX RESTUANS	23431	49	**0.002**
CULEX SALINARIUS	145	0	
CULEX TARSALIS	7	0	
CULEX TERRITANS	510	0	

```
train['Species'] = train['Species'].map({'CULEX PIPIENS/RESTUANS': 2, 'CULEX PIPIENS': 2, 'CULEX RESTUANS': 1}) \
                                .fillna(0)
```



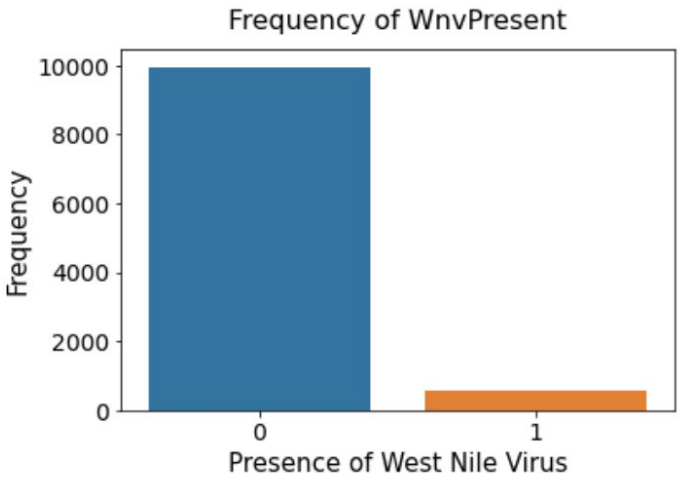
Feature Correlation with WNV Present

[INITIAL RUN] How do the models stack up?



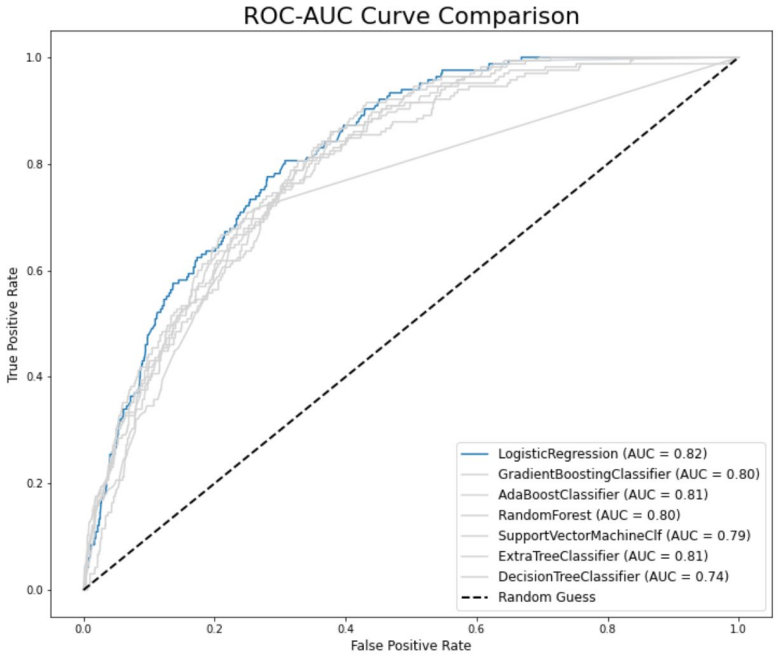
- Our non-boosting tree classifiers (Decision Tree, RF, ET) performed pretty badly here.
- In comparison, our Logistic Regression and Boosting models seem to be performing better in terms of AUC.
- This might be due to our classes being **poorly separated**, or due to our **imbalanced classes**.

Why we decided to use SMOTE to address the class imbalance



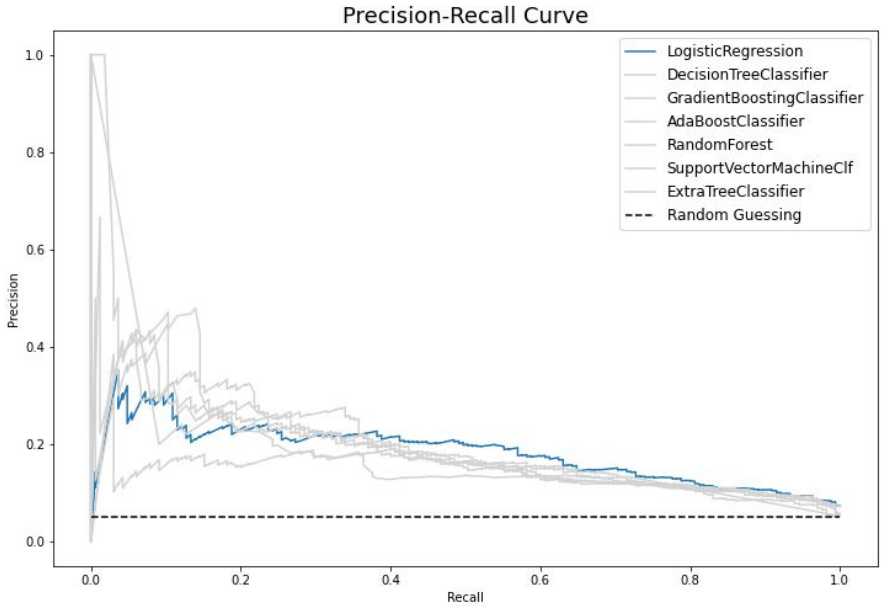
- SMOTE is a commonly used oversampling method that attempts to balance class distribution.
- Another option that we considered was class weights.
- However, we opted for SMOTE as some models like our Gradient Boosting classifier can't use class weights.

[FINAL RUN] How do the models stack up?



- Our Logistic Regression model seems to be a clear winner here in terms of AUC score.
- SMOTE also seems to have helped most of our models improve their AUC score.
- Our decision tree classifier did the worst out of all models.

Trade-off between precision and recall - how do we decide the final model?

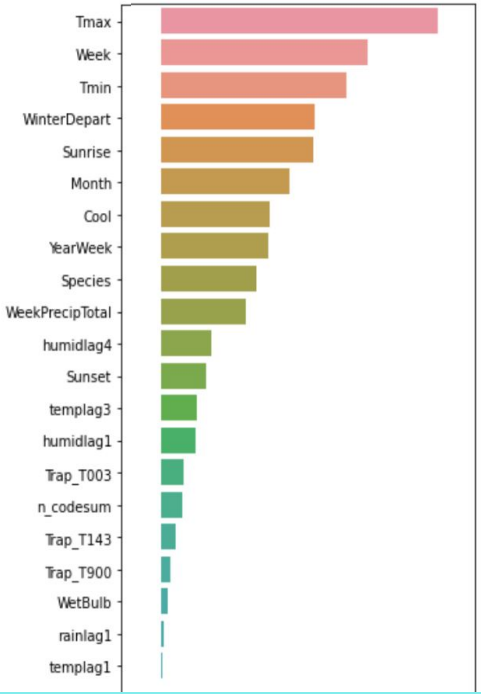


- The PR-AUC curve helps show our model's ability to to classify most or all of our positive class.
- Our logistic regression model is able to outperform the other two models in terms of recall at most thresholds.
- We decided to prioritize recall over precision, as ignoring WNV could lead to human death.

Conclusions

Figuring out what to do next and what the trade-offs are

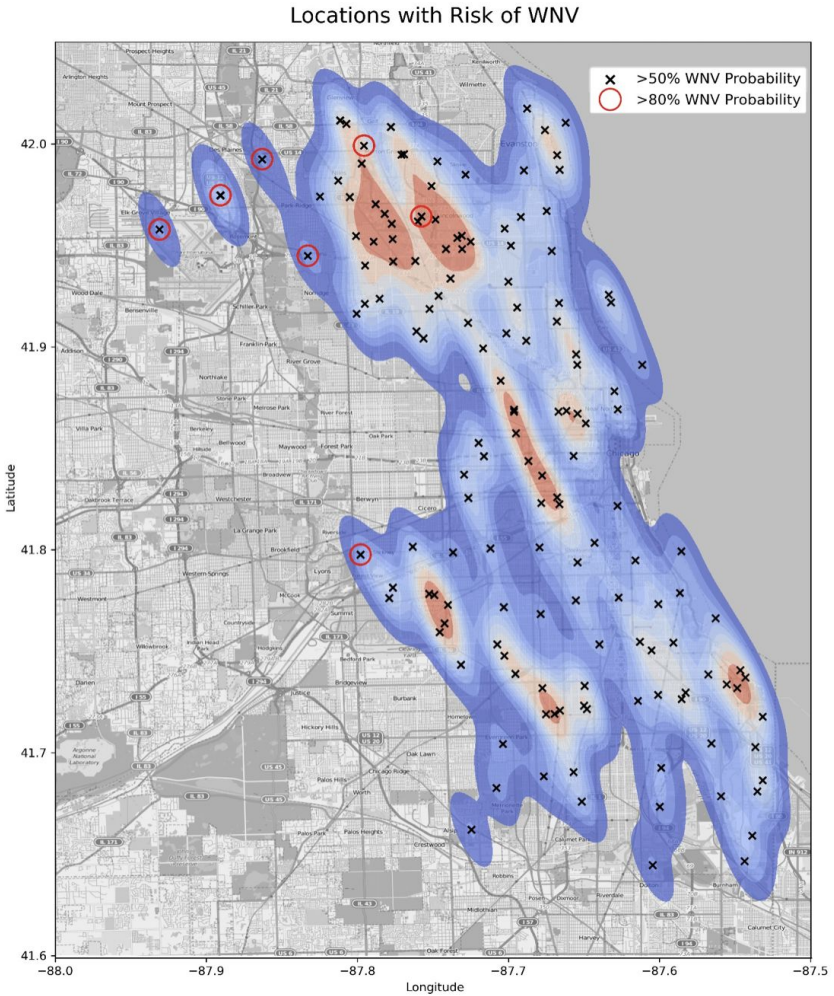
We can infer that weather conditions are the main predictor of WNV being present



Changing trends in global climates have affected patterns of infectious disease transmission*. Additionally, changes in weather — increased rainfall, humidity and heat waves have impacted patterns of insect activity. These changes have also created environments that better suit virus transmission.

> "[Some examples are] diseases, such as West Nile, Zika, [and] malaria. We should have every reason to expect that climate is going to make places [that used to be] less suitable to these diseases more suitable [to them] and vice versa."
>
> – Dr. Aaron Bernstein

*Sources:
https://idpjournal.biomedcentral.com/articles/10.1186/s40249-019-0565-1
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4342965/



Locations with 80% probability of WNV

- Elk Grove Village (7,500 acres)
- Des Plains (9,000 acres)
- Norridge (1,100 acres)
- Lincolnwood (1,700 acres)
- Stickney (1,200 acres)
- Forest View (900 acres)
- Morton Grove (3,100 acres)

Around 24,500 acres of area in Chicago identified as high risk, housing an approximate population of 148,500 people.

Source:Wikipedia

Let’s not forget the Human Cost

With a fatality rate of 5%, WNV is no small matter. Let’s look beyond just the numbers.

WLS-TV

West Nile Virus Illinois: Chicago resident becomes 1st 2020 death related to mosquito-borne illness, IDPH says

A Chicago resident is the first person to die this year from the mosquito-borne West Nile Virus, according to the Illinois Department of Public ...

Oct 9, 2020

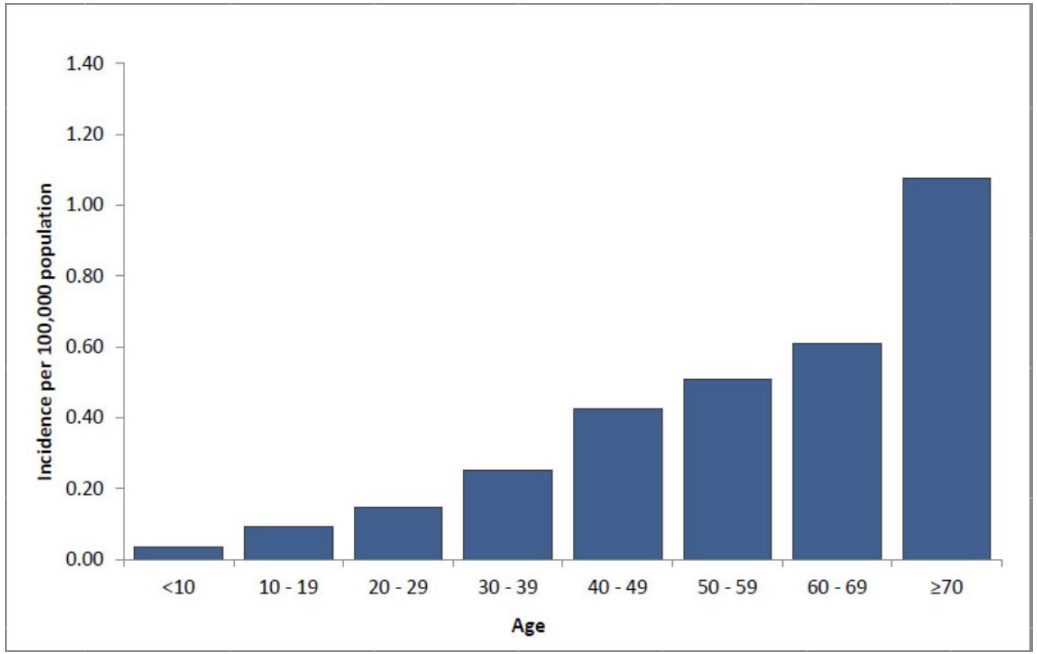
Chicago Tribune

Darien woman confirmed to have first human case of West Nile virus in state this year

Darien woman confirmed to have first human case of West Nile virus in state this year ... Chicago area and around the state have been confirmed to be West ... human cases of the virus in Illinois in 2019, resulting in one

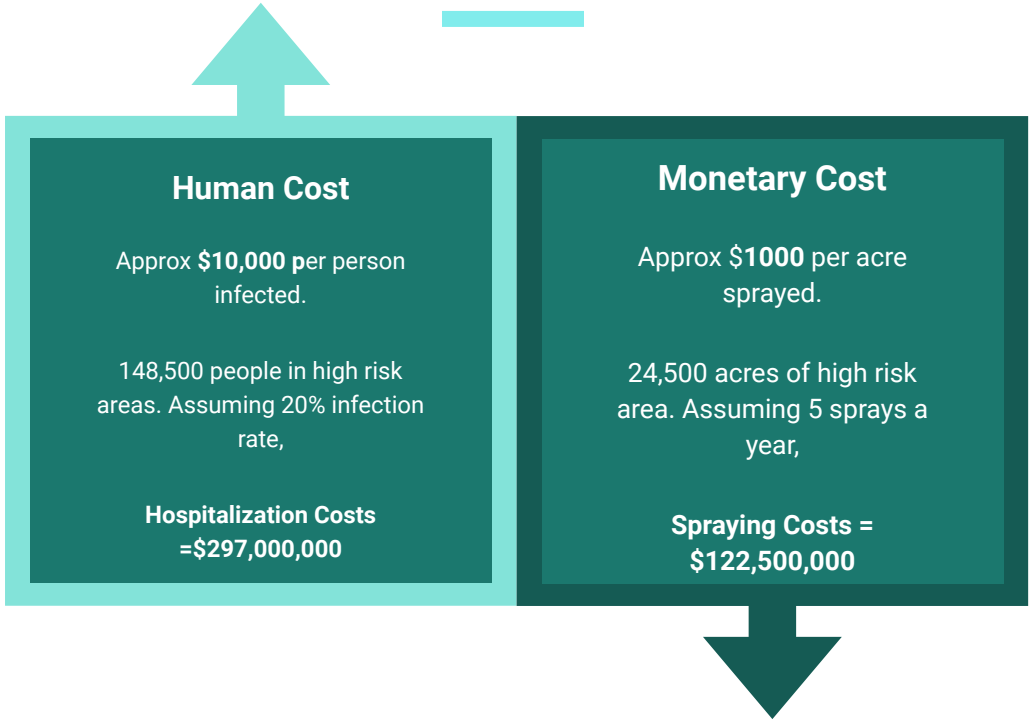
Sep 9, 2020

It disproportionately affects the elderly



Source: https://www.cdc.gov/westnile/index.html

The human cost is 2x more than the monetary cost



*Sources:
https://www.callnorthwest.com/2020/05/how-much-does-a-mosquito-treatment-cost/
https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-019-4596-9

We have chosen a deliberate focus on sensitivity

False Positives

Incorrectly predicting presence of WNV
Pros: Increased sensitivity would result in curtailing the virus before it spread
Cons: Extra spraying costs

False Negatives

Incorrectly predicting the lack of WNV
Pros: Reduced spraying costs
Cons: Increased chance of an outbreak, leading to potential snowball effects on hospitalization and the economy.

"There’s a better way to do it - find it"

—Thomas Edison

How can we make spraying more cost-efficient?

Automation with drones*

Target larval growth area with drone tech - improving spraying productivity and speed

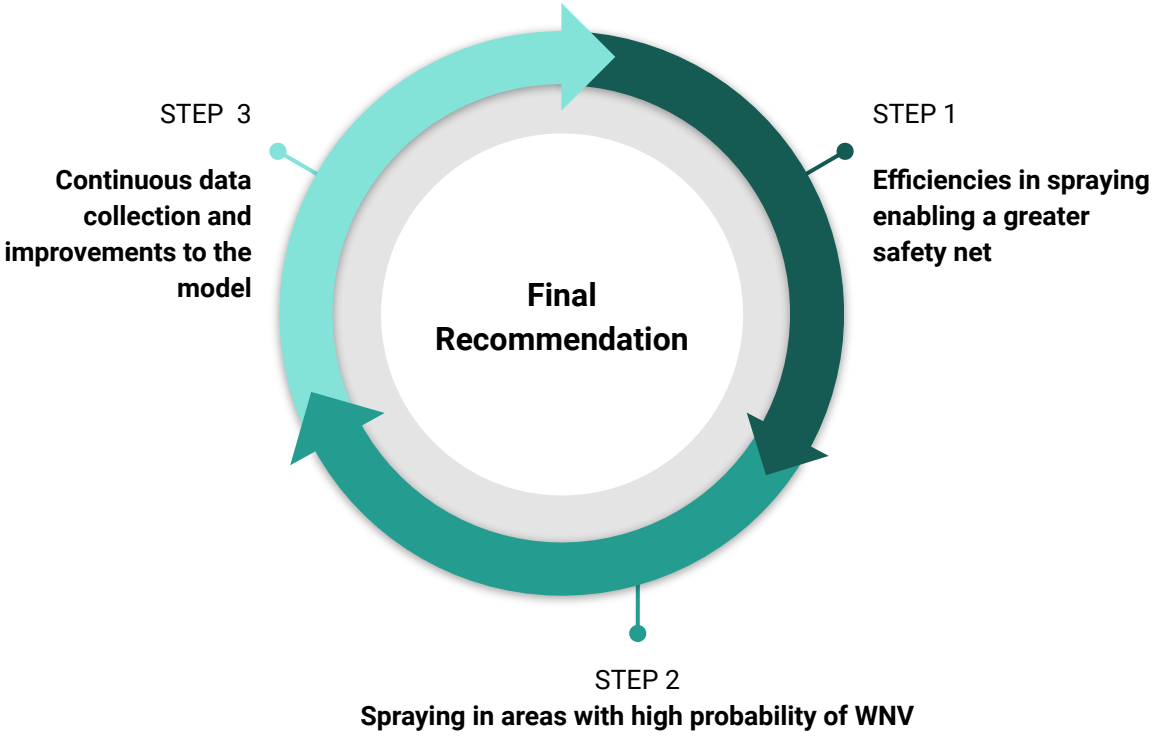
Follow best practices

Reduce spraying target areas based on guidelines from tropical countries like Singapore

Overall efficiency and cost reduction

By automation and more efficient spraying, we can reduce the cost of spraying, enabling the city of Chicago to save human lives and prevent the West Nile Virus.

*Sources:
https://www.nea.gov.sg/our-services/pest-control/mosquito-control/mosquito-control-in-condominium-estates
https://dronelife.com/2020/02/07/using-drones-to-combat-the-deadliest-animal-on-the-planet-the-mosquito/


STEP 3
Continuous data collection and improvements to the model
STEP 1
Efficiencies in spraying enabling a greater safety net
Final Recommendation
STEP 2
Spraying in areas with high probability of WNV

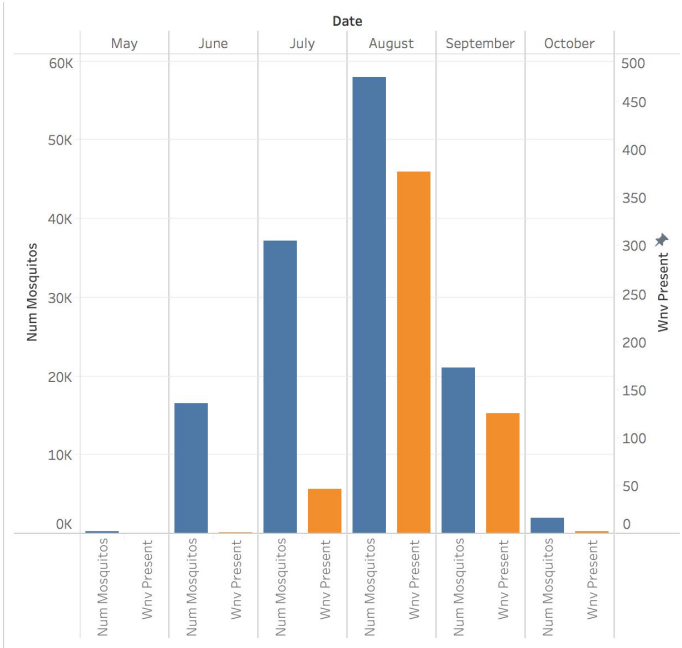
Thank you!

Does anyone have any questions?

Project team:
Asyraf, Ben D, Jack, Sahaj

Appendix

August appears to be the month with highest incidence of WNV



- Not surprising that the presence of WNV in mosquitoes is the highest in August.
- June and July are the peculiar months as the presence of WNV are disproportionate to the number of mosquitoes present.
  - Probably influenced by the birds migration pattern in North America*

*Source: https://www.nationalgeographic.com/magazine/2018/03/bird-migration-interactive-maps/

Data Cleaning

Weather:

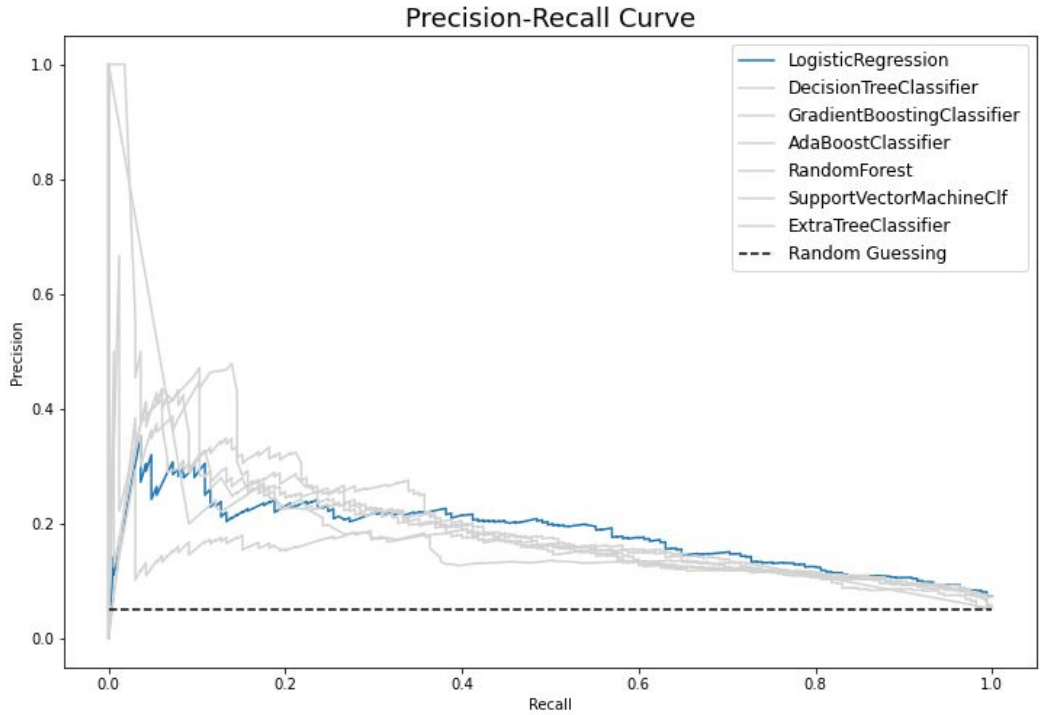
- Imputed missing values for daily average temperature `Tavg` with average of `Tmax` and `Tmin`
- Calculated missing values for `Cool` and `Warm` with `Tavg`
- Calculated missing `Depart` from Station 2 with 30 year normal temperature based on Station 1 readings and Station 2 `Tavg`.
- Imputed missing values for `WetBulb`, `PrecipTotal`, `StnPressure`, `SeaLevel`, `AvgSpeed` using readings of Station with non-missing value.
- Imputed 'T' or trace values as 0.01.
- Imputed `Sunrise` and `Sunset` for Station 2 with Station 1 values
- Split and recombined `CodeSum` to create proper spacing between different codes
- Created new feature counting number of exceptional weather phenomena based on `CodeSum`
- Created more interpretable features like `rain` and `lowvis` based on `CodeSum`
- Changed `Date` from string object to `datetime64`
- Dropped `Water1`, `Depth`, `SnowFall` due to high missing values (>99.5% 'M' or 0)
- Transformed all features into float values
- Merged Station 1 and Station 2 by averaging values of each station
- Added Year, Month, Week and Day of Week features

Train / Test:

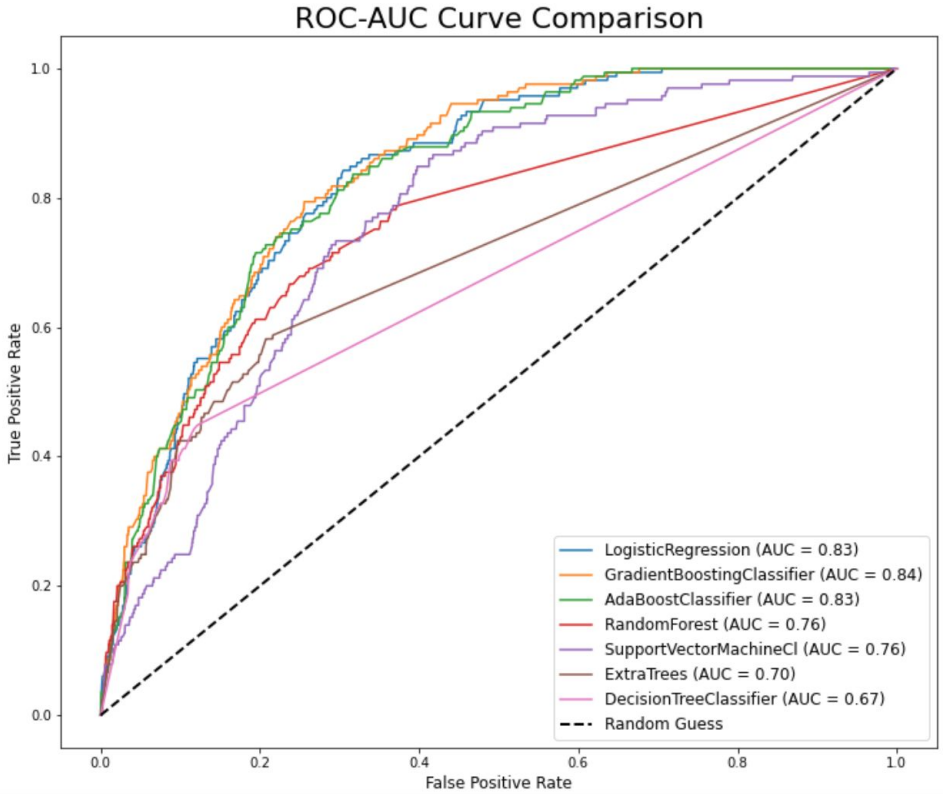
- Added Year, Month, Week and Day of Week features

Spray:

- Dropped duplicates

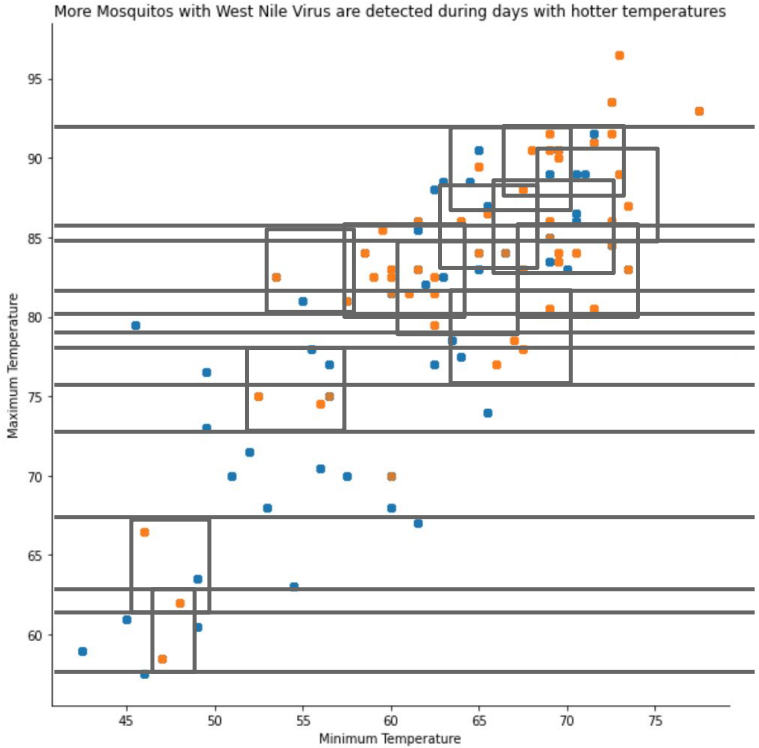


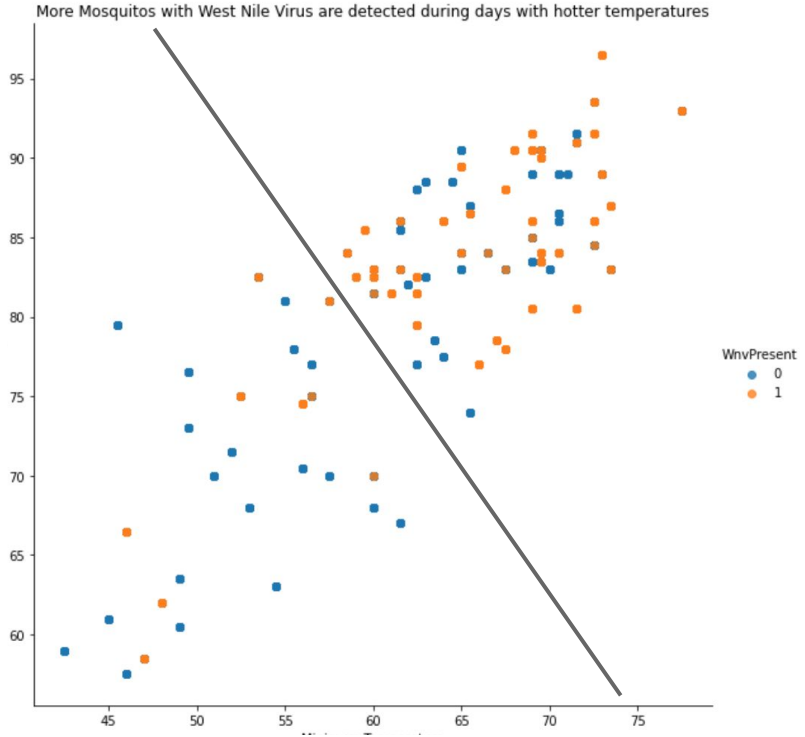
Full Precision-Recall Curve

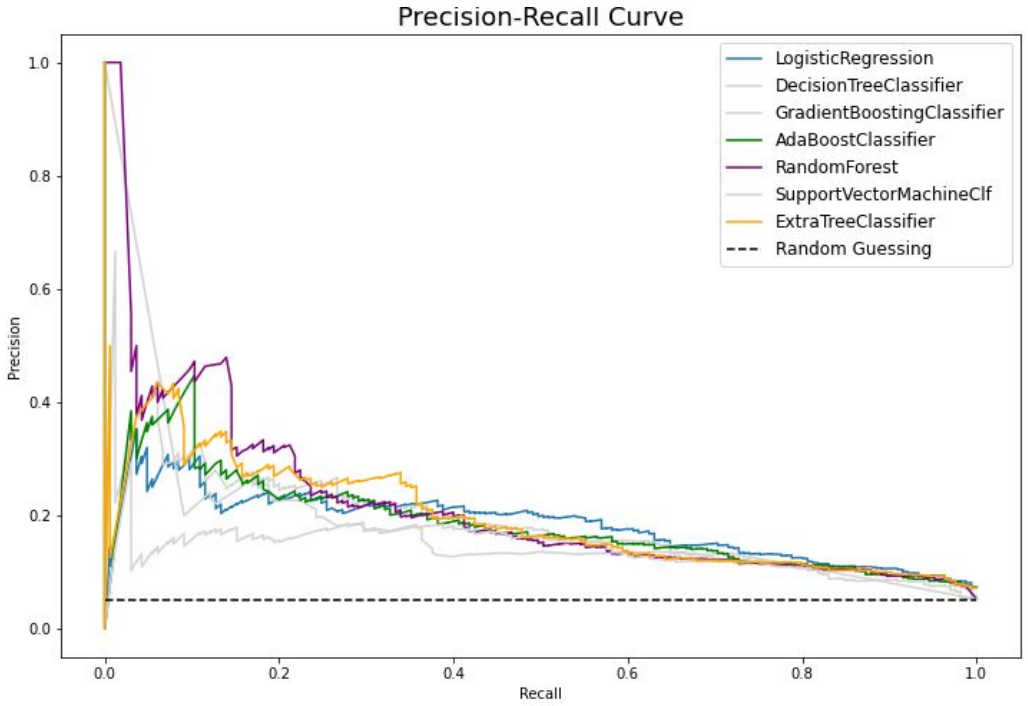


model	train_auc	test_auc	precision	specificity	recall	f_score
gb	0.906145	0.840764	0.368421	0.995983	0.042424	0.076087
lr	0.852268	0.831590	0.636364	0.998661	0.042424	0.079545
ada	0.867175	0.830862	0.666667	0.998996	0.036364	0.068966
svc	0.854914	0.763826	0.000000	1.000000	0.000000	0.000000
rf	0.986407	0.761421	0.309091	0.974556	0.206061	0.247273
et	0.988624	0.704181	0.284483	0.972213	0.200000	0.234875
dt	0.988624	0.669090	0.260000	0.962839	0.236364	0.247619

Overfitting of Tree Model vs Logistic Regression







Full Precision-Recall Curve

Model Coefficients

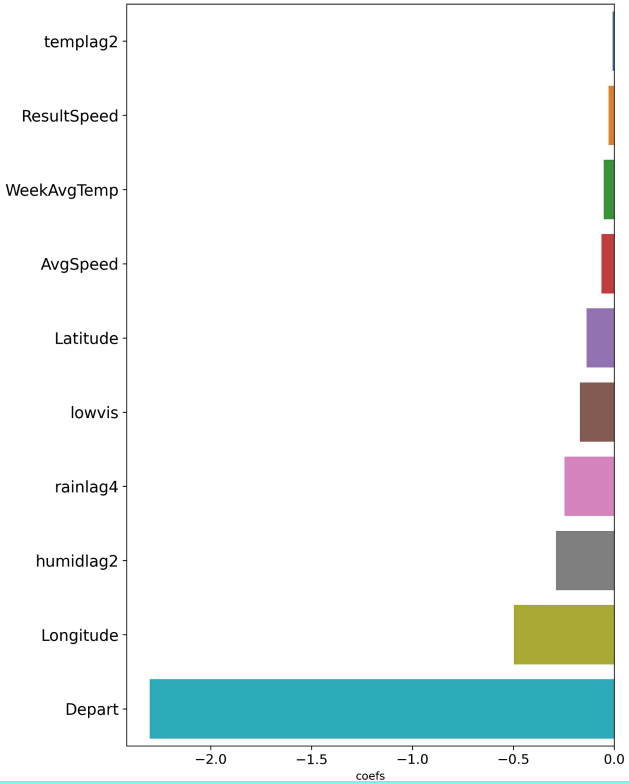
In [13]:
```
# These features are statistically significant
final_coefs[final_coefs['p_values'] < 0.05].sort_values(ascending=False, by='coefs')
```

Out[13]:

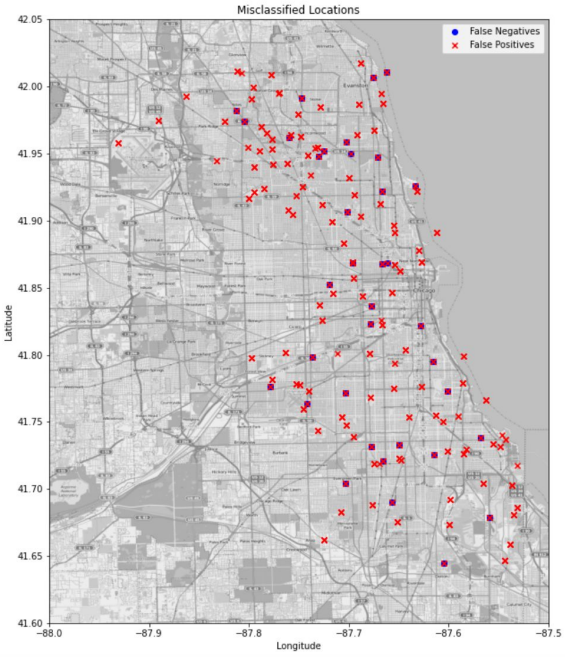
	coefs	sm_coefs	p_values	odds
Tmax	1.385	1.391	0.002	3.994
Tmin	0.926	0.936	0.023	2.524
WinterDepart	0.765	0.760	0.000	2.150
Sunrise	0.761	0.761	0.010	2.140
Month	0.643	0.651	0.046	1.903
YearWeek	0.539	0.537	0.000	1.714
Species	0.476	0.475	0.000	1.610
WeekPrecipTotal	0.426	0.425	0.000	1.531
humidlag4	0.250	0.246	0.007	1.284
Trap_T003	0.113	0.113	0.002	1.120
Trap_T143	0.070	0.070	0.017	1.073
rainlag4	-0.246	-0.243	0.035	0.782
humidlag2	-0.289	-0.287	0.009	0.749
Longitude	-0.498	-0.497	0.000	0.608
Depart	-2.302	-2.291	0.000	0.100

A one unit increase in Tmax leads to a 4x increase in probability of WNV

Negative Model Coefs



Misclassification analysis



- Imperative to minimise False Negatives as much as possible due to the human costs.
- False Negatives occur when mosquitoes breeding conditions were not ideal
    - Implies existence of external factors
- Potential hypothesis - the American Robin* is able to carry the WNV when most birds tend to die shortly after being infected.
    - The absence of such a feature in our model may have distorted the prediction - more mosquitoes in the area does not necessarily means more WNV.

*Source: https://news.wisc.edu/do-chicagos-suburbs-hold-the-key-to-understanding-west-nile-virus/