Predictive Analysis of

West Nile Virus in Chicago

To Inform Annual Deployment of Pesticides

Team of Arti, Bryan, Jefferson, Nandhini

Agenda

□ Background

□ The Data

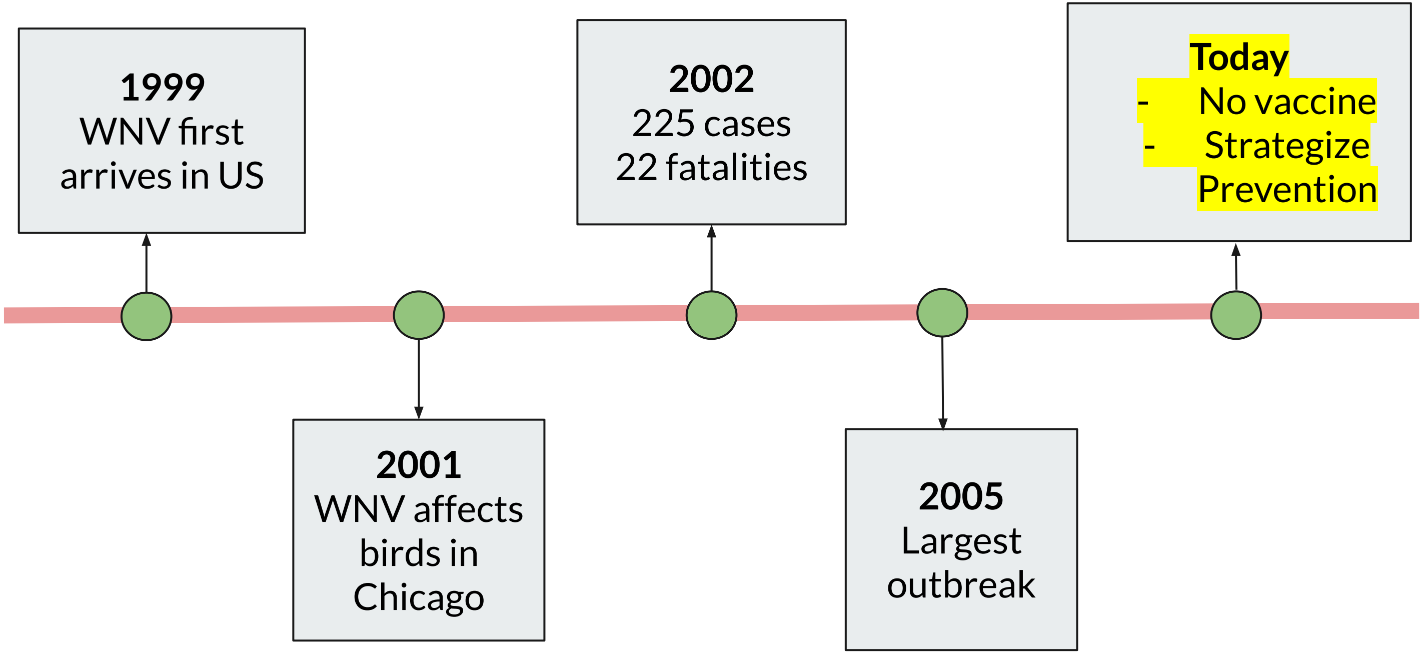
□ Modelling

□ Conclusion

Background

- Problem statement
- Analysis of costs
- Data Science Problem

Background and Problem Statement



1999
WNV first arrives in US

2001
WNV affects birds in Chicago

2002
225 cases
22 fatalities

2005
Largest outbreak

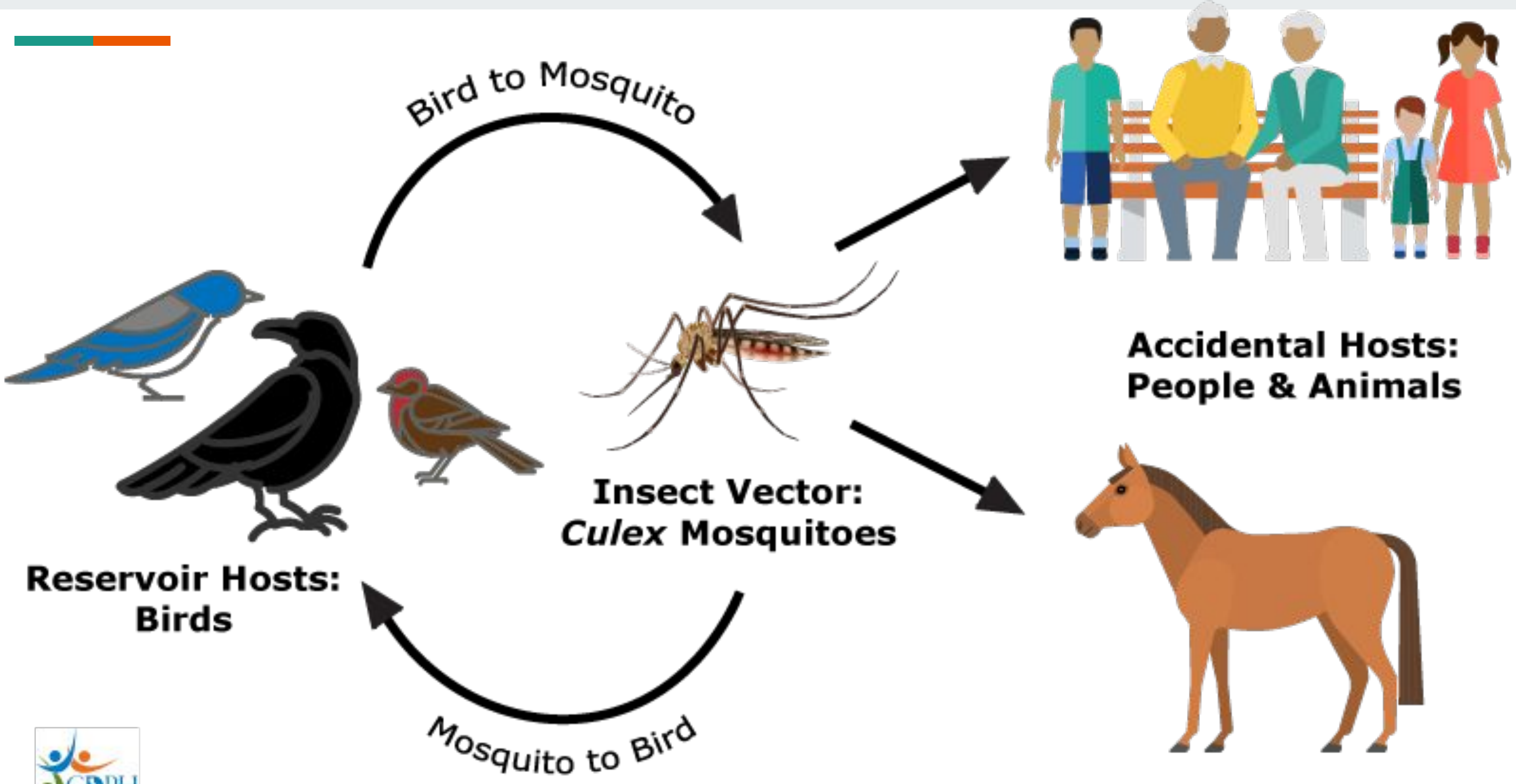
Today

- No vaccine
- Strategize Prevention

Objectives

- Utilize available data to predict the probability of WNV-carrying mosquitoes in traps over May to end October of each year.
- Determine spraying plan -- time periods and areas to deploy pesticide sprays cost-effectively.

Transmission Cycle



Bird to Mosquito

Mosquito to Bird

Reservoir Hosts:
Birds

Insect Vector:
Culex Mosquitoes

Accidental Hosts:
People & Animals

Background on WNV-related costs:

	Pesticide Spraying	**Human Treatment**
COSTS	USD 10000 for repeated spraying over 40000 m^2 (10 acres) area from May to September (source)	USD 8-9,000 / patient on average (source)

- 1 pesticide spray area may potentially reduce mosquitoes reduce several human infections
  - Postulate that spraying costs to treatment could effectively be in a ratio of **1:20.**
- Logically we could spray all areas with mosquitoes detected in Chicago at least once every 3 weeks, but we also don’t want to do that!

What if Model Predicts Wrongly?

	Not Spraying Infected Areas (Infected Humans!) In model: False Negatives	Spraying on Areas with Low Chance of Infection (Side Effect Cost of Pesticide Use) In model: False Positives
COSTS	● VERY EXPENSIVE! ● Economic and political costs ● Healthcare costs	● Less expensive but want to reduce as much as possible ● Environment cost ● Human Health Costs

The Data

- Exploratory Data Analysis
- Feature Engineering

What to Look up for?

Virus Growth Driver

- Mosquito Population affecting spread
- Factor affecting Mosquito Population
- Spread pattern
- Species effect
- Mosquito Season

Control Measure

- Past effort effects
- Strategies to eliminate virus
- Trade-offs

Dataset

Mosquito Trap

2007 - 2014*

- Trap location
- Number of Mosquito
- WNV presence

Weather Condition

2007 - 2014

- Temperature
- Humidity
- Geographical properties

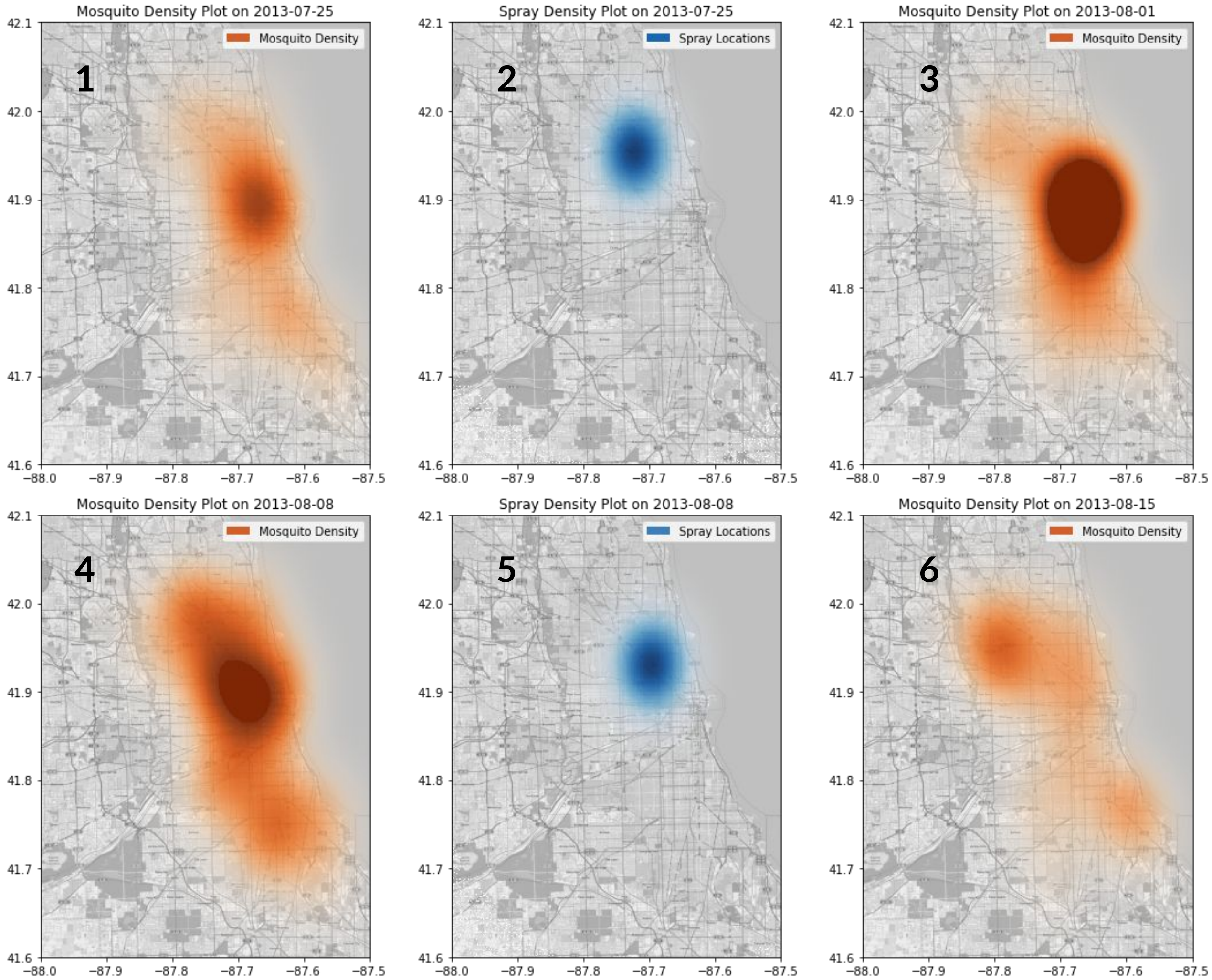
Spray Effort

2011 & 2013

- Spray location

* Only odd years contains Number of Mosquito and WNV Presence

Is spraying effective?



Mosquito Density Plot on 2013-07-25

Spray Density Plot on 2013-07-25

Mosquito Density Plot on 2013-08-01

Mosquito Density Plot on 2013-08-08

Spray Density Plot on 2013-08-08

Mosquito Density Plot on 2013-08-15

Preview of Data Analysis

- **WNV presence** is proportional to the **mosquito population**

- **Mosquito population** is affected by **temperature**

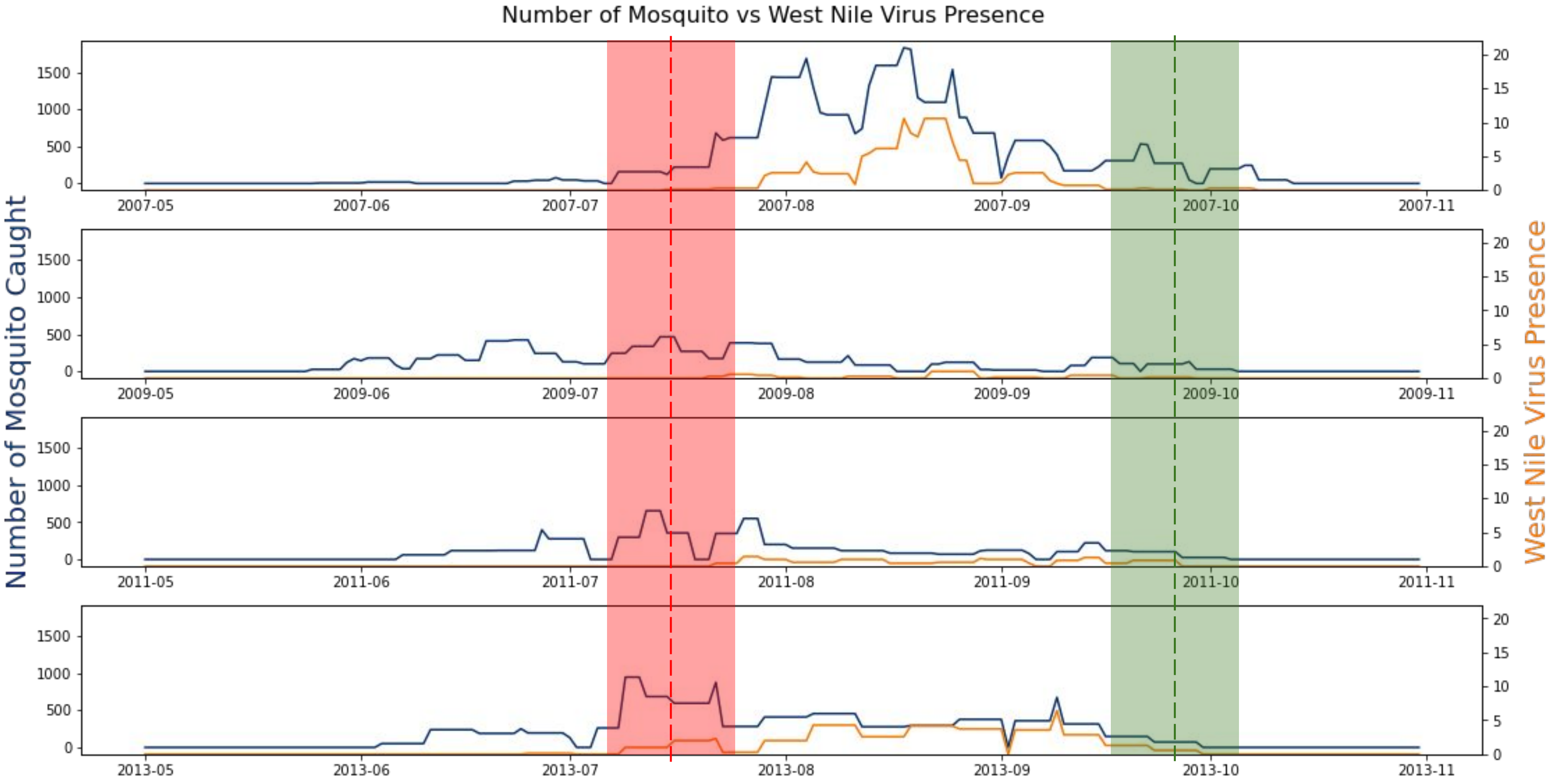
- **Two species carry all the virus**
  - They spread quite equally across Chicago

- **City area** always have high **mosquito population** density

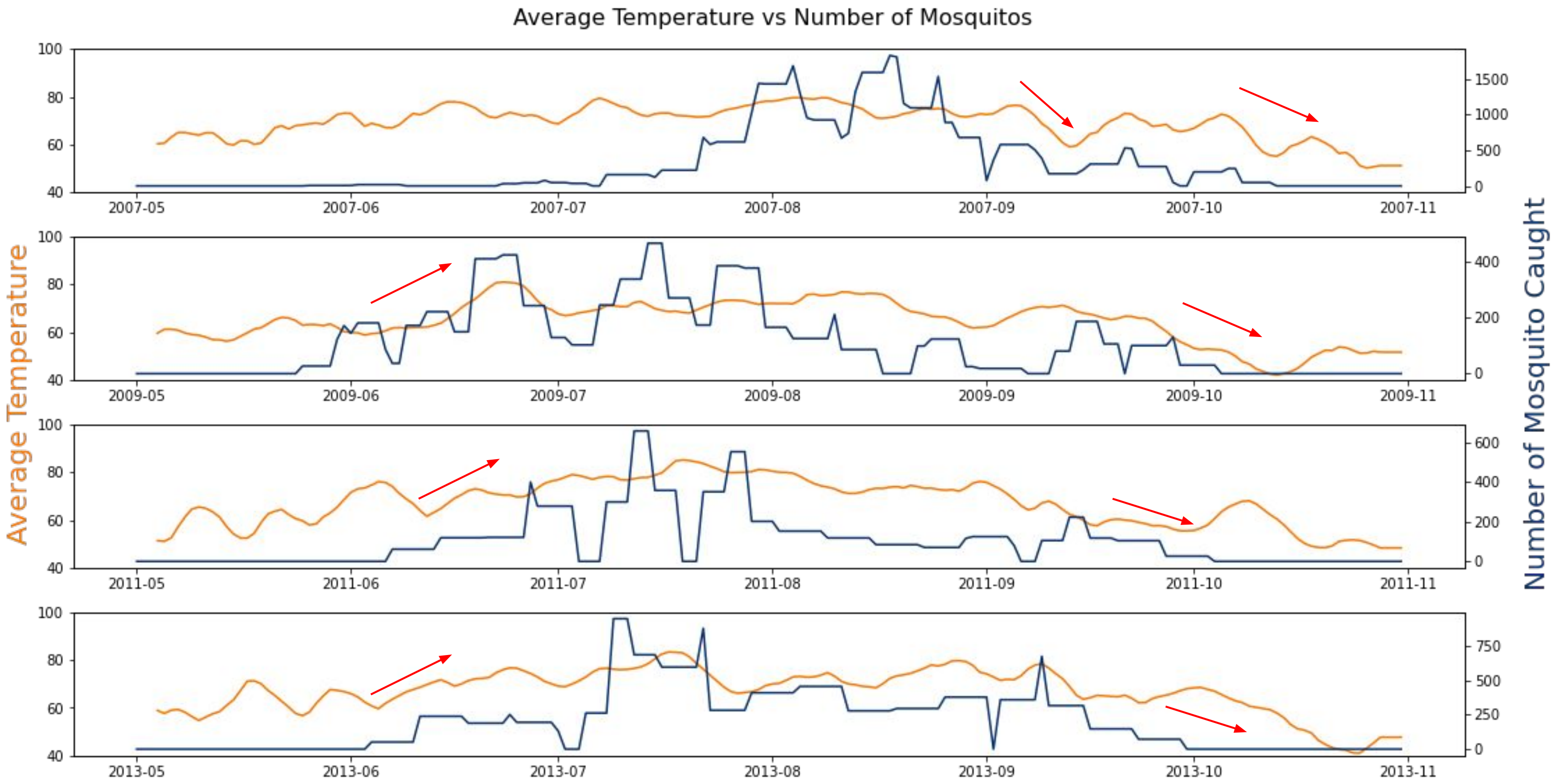
- The virus spread **sporadically**

- Trap can be **categorized** into **5 clusters** depending on **location and infection rate**

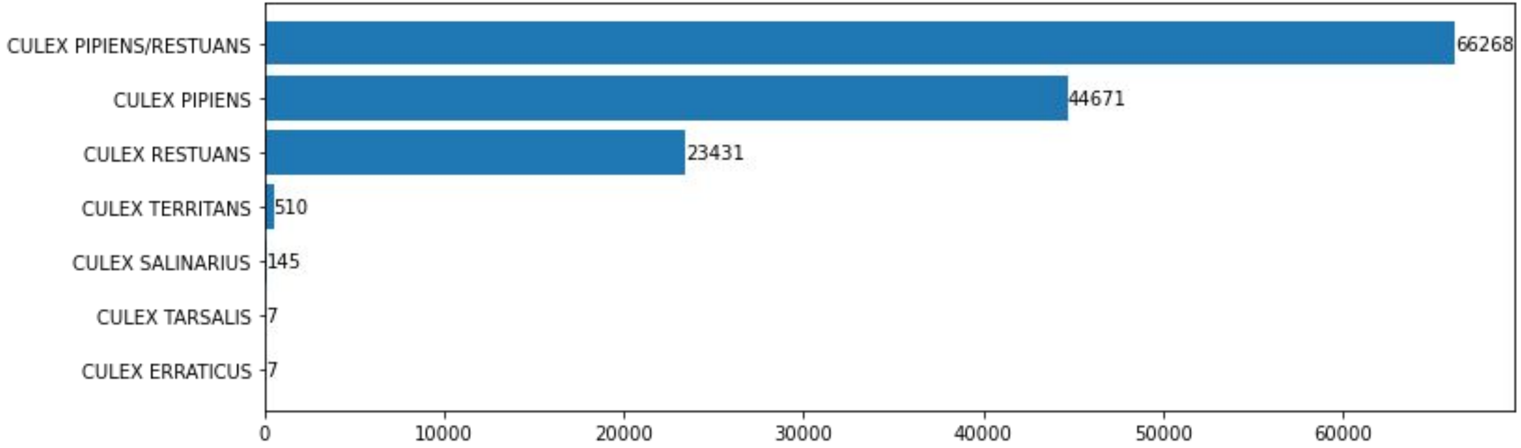
WNV - Number of Mosquitos



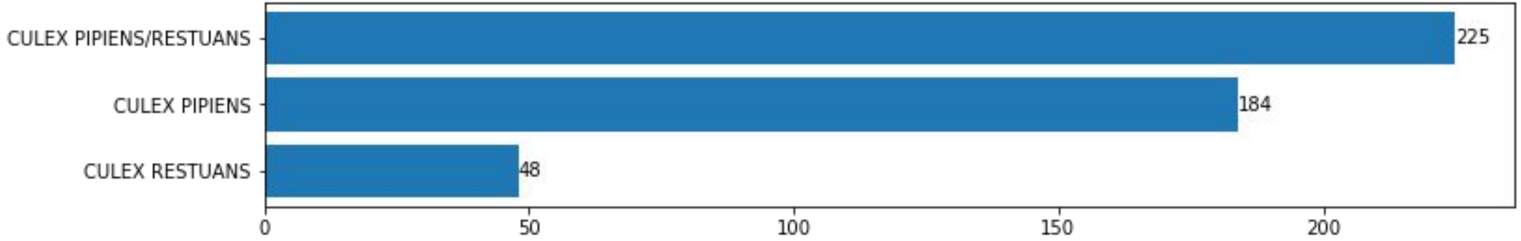
Number of Mosquitos - Temperature



Two Species Carry all the Virus

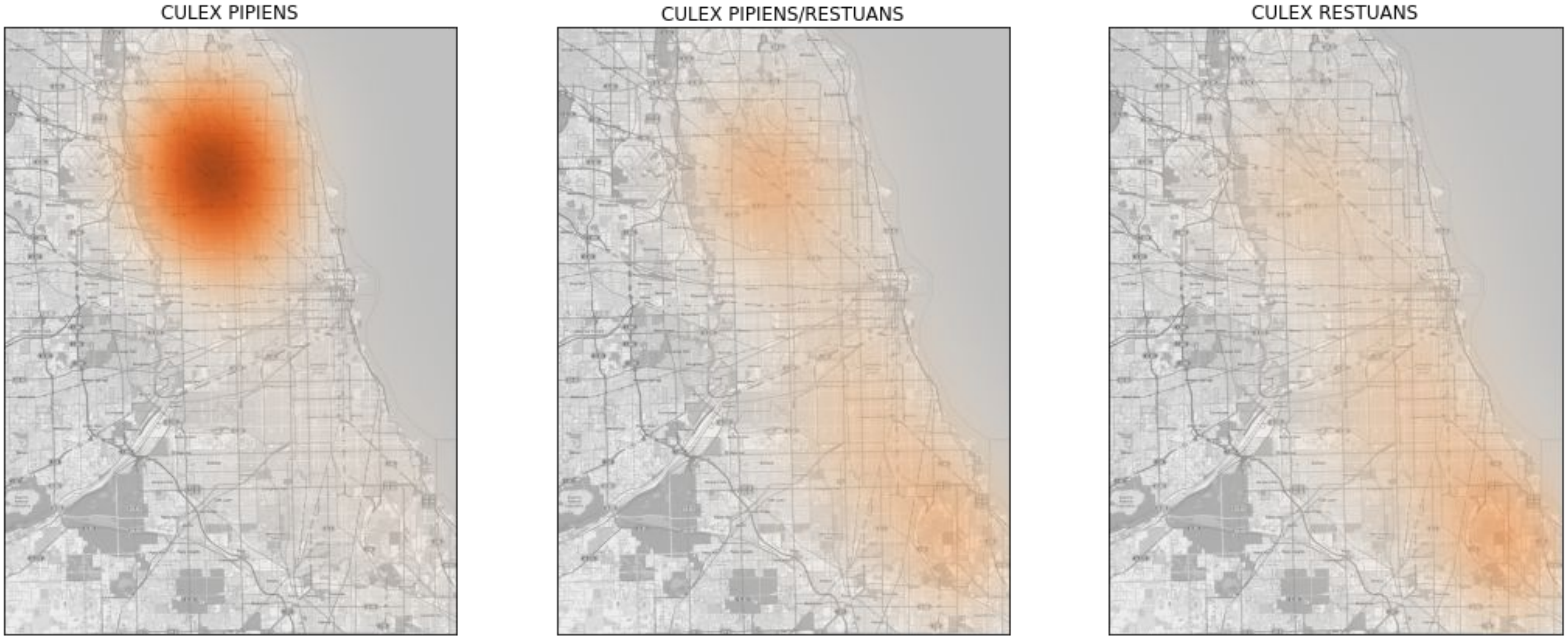


Number of Mosquito Caught



Number of Trap with West Nile Virus per Species

How are the Carrier Species Distributed?

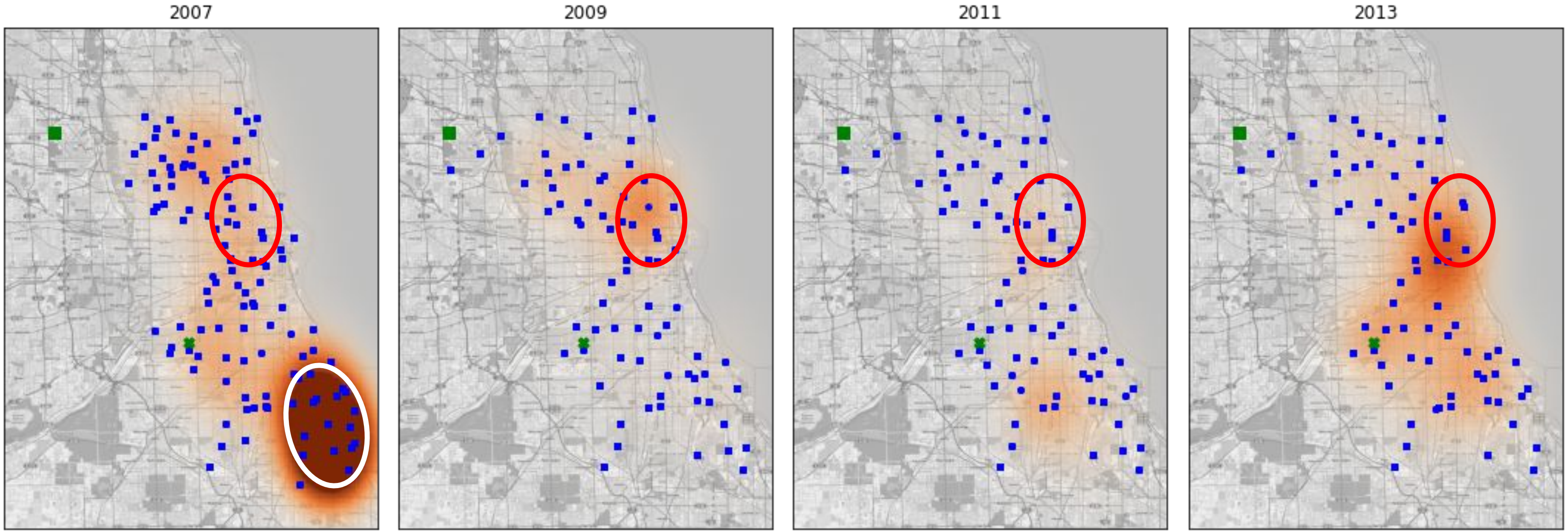


CULEX PIPIENS

CULEX PIPIENS/RESTUANS

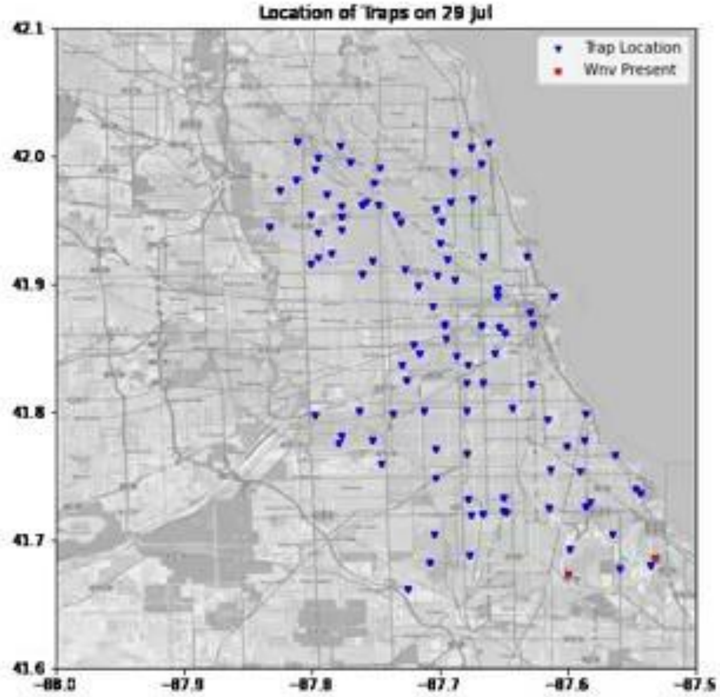
CULEX RESTUANS

Does Mosquito Appear on the Same Location Yearly ?

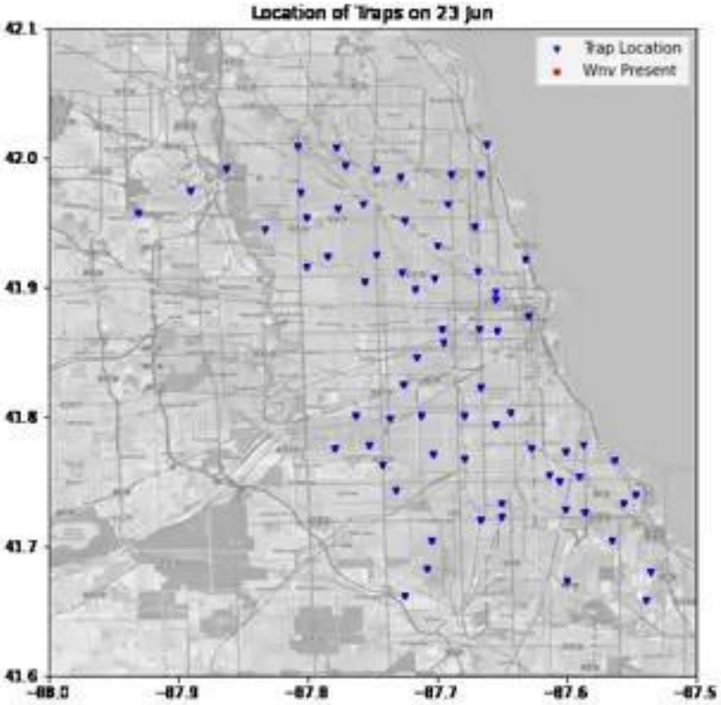


How the Virus Spread?

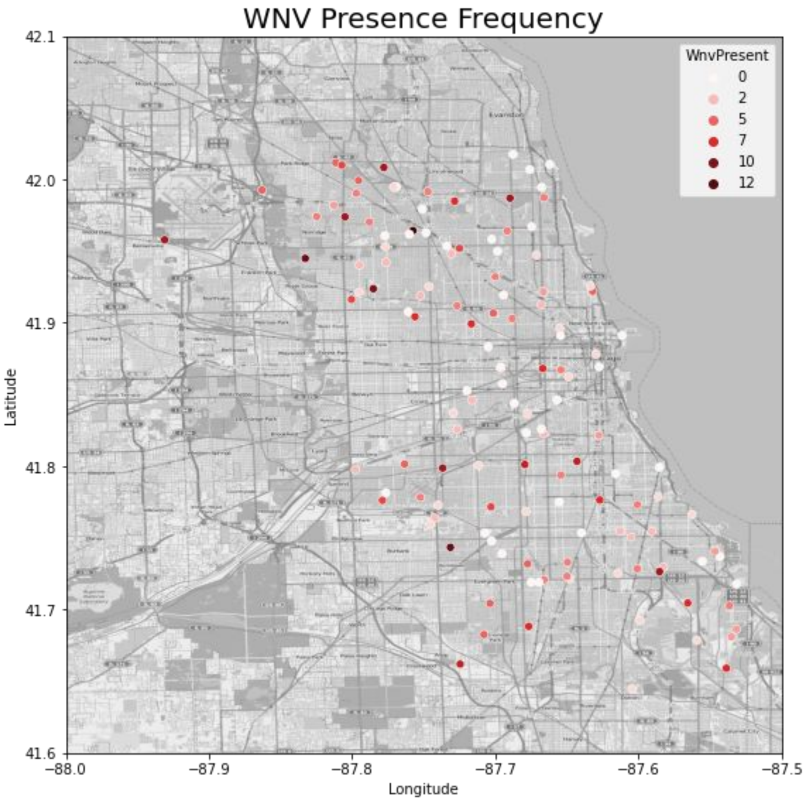
For year 2007

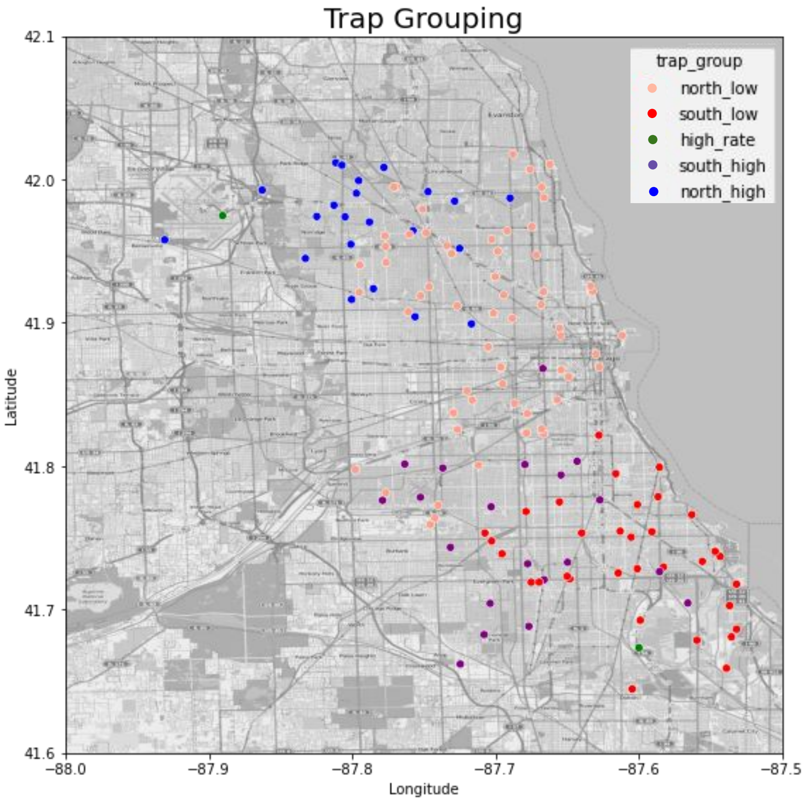


For year 2013



How the Virus Spread?





Feature Engineering

Mosquito Trap

2007 - 2014

- Trap Grouping
  - K-means, n=5
- Infection Rate
  - Per species for each week of year

Weather Condition

2007 - 2014

- Relative Humidity *
- Dark hours **
- 7-days Rolling Average

*https://parasitesandvectors.biomedcentral.com/articles/

**https://www.mosquitomagnet.com/articles/when-are-mosquitoes-most-active

The Model

- Model Performance Metric
- Imbalance Class Treatment
- Model Selection and Trade-off

Model Selection Metrics

Priority #1
False Negatives = 0!

Incorrectly predicting no WNV

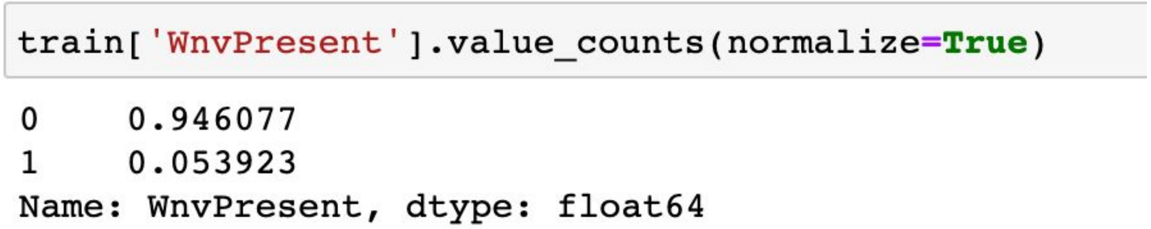
Misidentified areas may allow outbreak of WNV with increase in **medical cost, economy cost and human lives**

Priority #2
False Positives -- Minimise

Incorrectly predicting the presence of WNV

Additional spraying cost and side effects

Baseline Model



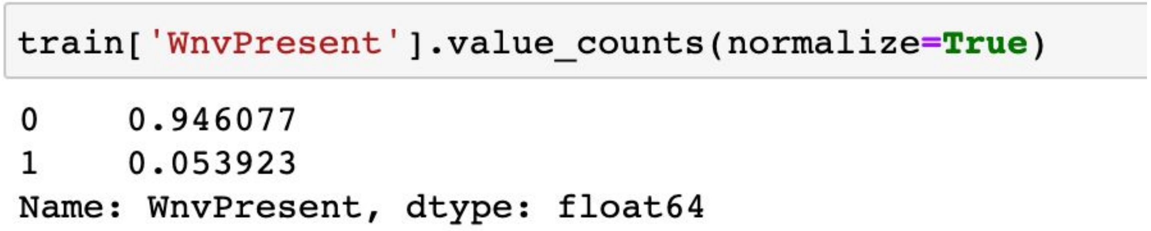
```
train['WnvPresent'].value_counts(normalize=True)
```

```
0    0.946077
1    0.053923
Name: WnvPresent, dtype: float64
```

- **Assuming all traps predicted to have WNV infected mosquitoes all the time**
    - Results in having 0 FNs, but a huge number of FPs.

- Safe, but requires a **lot of pesticide spraying!**

Imbalance Class Treatment

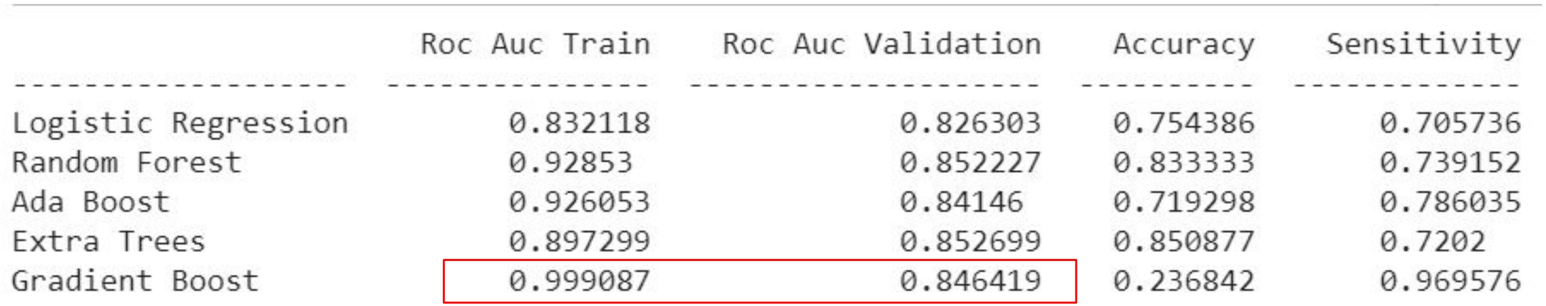


```
train['WnvPresent'].value_counts(normalize=True)
```

```
0     0.946077
1     0.053923
Name: WnvPresent, dtype: float64
```

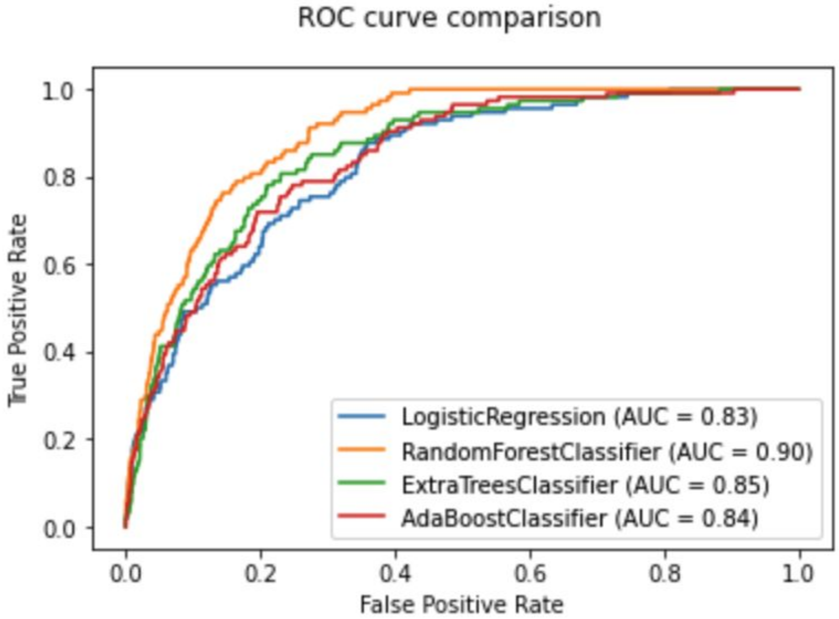
- Dataset : Imbalanced Classes
- Problem : Poor performance on the minority class
- Solution : Data augmentation of using - SMOTE

Model Comparison



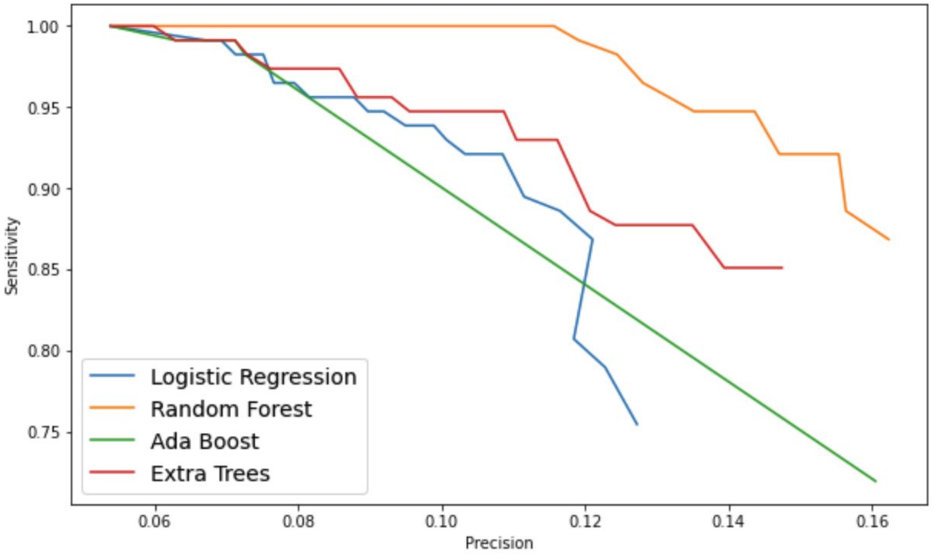
```
                     Roc Auc Train    Roc Auc Validation    Accuracy    Sensitivity
-------------------  ---------------  --------------------  ----------  -------------
Logistic Regression         0.832118              0.826303    0.754386       0.705736
Random Forest               0.92853               0.852227    0.833333       0.739152
Ada Boost                   0.926053              0.84146     0.719298       0.786035
Extra Trees                 0.897299              0.852699    0.850877       0.7202
Gradient Boost              0.999087              0.846419    0.236842       0.969576
```

Model Selection



- Visualizes the tradeoff between TPR and FPR
    - TPR - fraction of missed detection
    - FPR - fraction of false alerts based on model predictions
- Higher the AUC, better the model in classifying the data
- Random Forest performs better in terms of AUC

Selecting Best Predictive Model



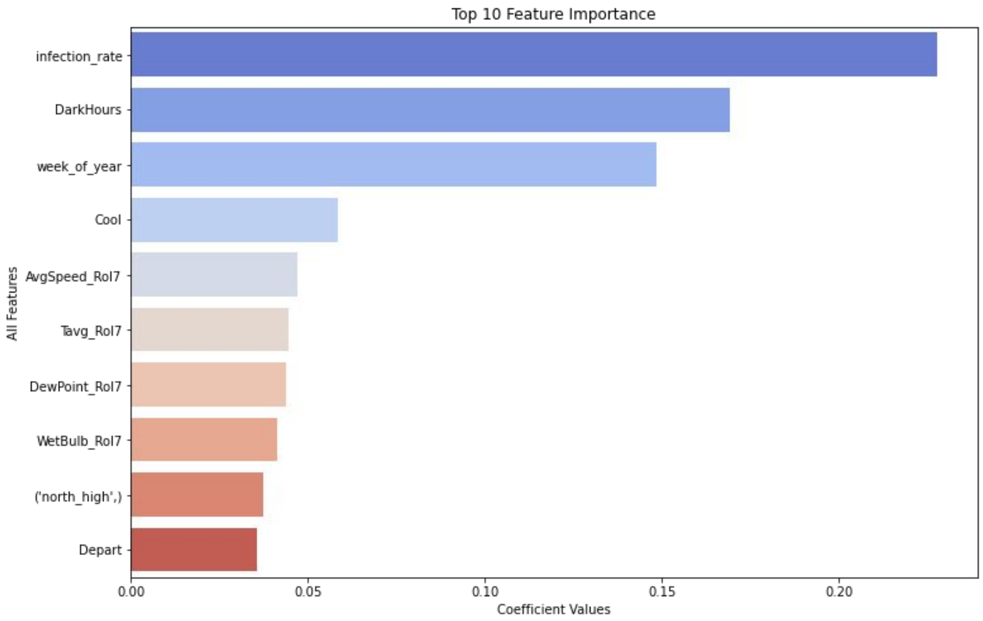
Sensitivity- Precision plot helps us to identify which model is able to classify all of our positive class well

Random forest has outperformed other models in terms of Precision for most thresholds.

We decided to prioritise Sensitivity over precision as ignoring FN(WNV) might result in outbreak.

Feature Importance

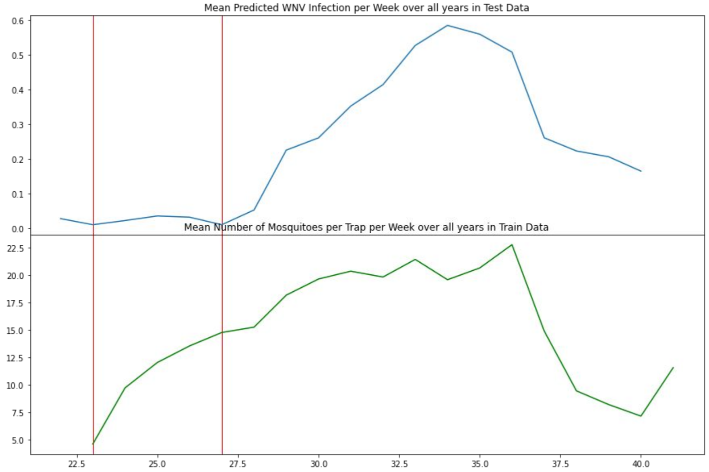
- Infection rate consistently the most important
- Length of daily dark hours important
- Mixed importance for smoothed weather-based features
- Model did not find Trap-based features important



Recommendations and Conclusion

Recommendations: Usage of Predictive Model in distribution of budget for pesticide sprays

For a given week, highest probability trap areas will be handled with priority as far as budget can go.



- Starting from Week 27 when infected mosquito pools are expected to start showing up -- not when mosquitos show up.
- Model predicted highest probability Traps have their area sprayed first.

Recommendations: Improving Model

- Number of Mosquitoes in the mosquito season are likely affected by weather conditions in Winter and Spring.
- Attempt to classify trap areas by:
  - Geographic data: Distance to large drains, ponds, etc.
  - Demographic data: Types of surrounding homes, income level, etc.

Conclusion

- Spraying is much less costly than human treatment
- Model allows for prioritized deployment of pesticide sprays to optimize spending.

Thank you!