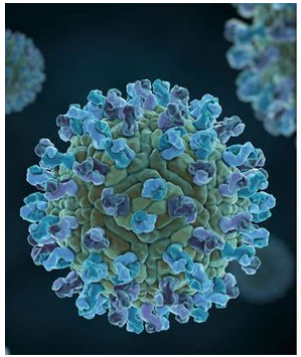
# DSI 19 Project 4
# West Nile Virus

Stephen, Zawanah, Jordan

# Problem Statement

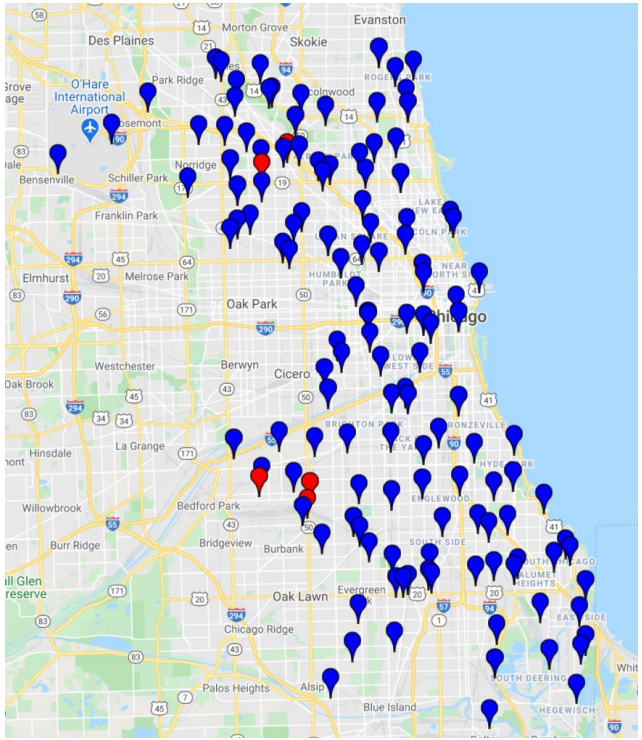
We are tasked by the Chicago's Department of Public Health to set up a surveillance and control system to combat against the West Nile Virus outbreak in Chicago. Using past history dataset of Chicago's weather, pesticide spray and mosquito traps, we are to analyse and give the most optimal conditions to combat against the virus.



# Overview of Data

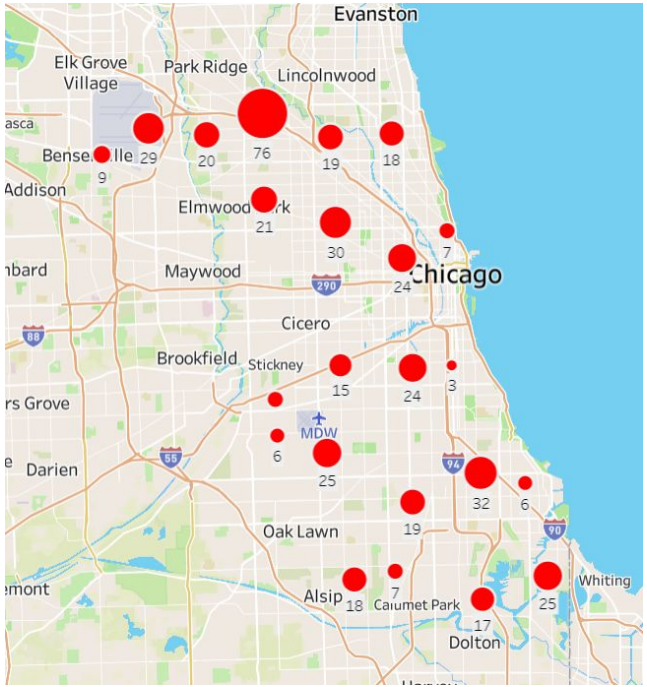
| Dataset | Description |
| --- | --- |
| Weather | Date: 2007 - 2014<br>Rows: 2944<br>Features: 22 |
| Spray | Date: 2011 - 2013<br>Rows: 14835<br>Features: 4 |
| Train | Date: 2007 - 2013<br>Rows: 10506<br>Features: 12 |
| Test | Date: 2008 - 2013<br>Rows: 116923<br>Features: 11 |

# Visuals - Train(Clean) Dataset




The pins on the map contains the location of the mosquito traps in the train dataset. Blue pins are mosquito traps without the virus while red pins contain traces of the virus.

# Visuals - Train Dataset




The red circles are locations with a high number of mosquitoes caught, the bigger the circle the more mosquitoes caught in the area.

# Feature Engineering/ EDA

1. Distance to traps most popular with mosquitoes
    a. Grid structure
    b. Occurence of mosquitoes (WNV)
    c. Group by date and trap/location
2. Month in the year
    a. Similar weather pattern in the same month across years
3. Rolling 7/10 days of features
    a. Occurence of weather code, max temp
4. Humidity
    a. Implied by dewpoint temperature and observed temperature (Lawrence, Mark G, 2005)

# Top Features - Insight from Data

1. Species: Culex Restuans
    a. Known to be closely related to WNV
2. Short distance to traps with most mosquitoes detected with WNV
    a. T900, T115, T002, T138
3. August
4. Thunderstorm, rain, fog occured in the last 7/10 days
5. High temperature
6. High humidity

# Modelling

- **Preprocessing**
  - Select important features (chi2)
  - Adaptive Synthetic (ADASYN) oversampling technique
- **Selection**
  - ROC-AUC score
  - Recall > Precision
    - Minimising false negatives more important than minimizing false positives
  - Extent of overfit

Pipeline : Logistic Regression

| Results of Baseline for Pipeline 1 - Logistic Regression | | | | | | |
|---|---|---|---|---|---|---|
| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
| Training Set | 0.754 | 0.766 | 0.733 | 0.804 | 0.754 | [[4218 1777] [1189 4868]] |
| Validation Set | 0.72 | 0.21 | 0.12 | 0.823 | 0.769 | [[1446 577] [ 17 79]] |
| Extent of Overfit | 0.034 | 0.556 | 0.613 | -0.019 | -0.015 | |

Pipeline : K-Nearest Neighbours

| Results of Baseline for Pipeline 2 - K-Nearest Neighbours | | | | | | |
|---|---|---|---|---|---|---|
| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
| Training Set | 0.946 | 0.947 | 0.937 | 0.957 | 0.946 | [[5605 390] [ 263 5794]] |
| Validation Set | 0.902 | 0.32 | 0.233 | 0.51 | 0.715 | [[1862 161] [ 47 49]] |
| Extent of Overfit | 0.044 | 0.627 | 0.704 | 0.447 | 0.231 | |

Pipeline : Random Forest

| Results of Baseline for Pipeline 3 - Random Forest | | | | | | |
|---|---|---|---|---|---|---|
| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
| Training Set | 0.978 | 0.978 | 0.971 | 0.985 | 0.978 | [[5817 178] [ 89 5968]] |
| Validation Set | 0.908 | 0.248 | 0.198 | 0.333 | 0.635 | [[1893 130] [ 64 32]] |
| Extent of Overfit | 0.07 | 0.73 | 0.773 | 0.652 | 0.343 | |

Pipeline : Ada Boost

| Results of Baseline for Pipeline 4 - Ada Boost | | | | | | |
|---|---|---|---|---|---|---|
| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
| Training Set | 0.839 | 0.841 | 0.834 | 0.849 | 0.839 | [[4973 1022] [ 917 5140]] |
| Validation Set | 0.844 | 0.324 | 0.202 | 0.823 | 0.834 | [[1710 313] [ 17 79]] |
| Extent of Overfit | -0.005 | 0.517 | 0.632 | 0.026 | 0.005 | |

# Modelling

- **AdaBoost Classifier**
  - Least overfit model
  - Able to generalise the best among the rest of the models

Pipeline : Logistic Regression

Results of Baseline for Pipeline 1 - Logistic Regression

| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
|---|---|---|---|---|---|---|
| Training Set | 0.754 | 0.766 | 0.733 | 0.804 | 0.754 | [[4218 1777] [1189 4868]] |
| Validation Set | 0.72 | 0.21 | 0.12 | 0.823 | 0.769 | [[1446 577] [ 17 79]] |
| Extent of Overfit | 0.034 | 0.556 | 0.613 | -0.019 | -0.015 | |

Pipeline : K-Nearest Neighbours

Results of Baseline for Pipeline 2 - K-Nearest Neighbours

| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
|---|---|---|---|---|---|---|
| Training Set | 0.946 | 0.947 | 0.937 | 0.957 | 0.946 | [[5605 390] [ 263 5794]] |
| Validation Set | 0.902 | 0.32 | 0.233 | 0.51 | 0.715 | [[1862 161] [ 47 49]] |
| Extent of Overfit | 0.044 | 0.627 | 0.704 | 0.447 | 0.231 | |

Pipeline : Random Forest

Results of Baseline for Pipeline 3 - Random Forest

| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
|---|---|---|---|---|---|---|
| Training Set | 0.978 | 0.978 | 0.971 | 0.985 | 0.978 | [[5817 178] [ 89 5968]] |
| Validation Set | 0.908 | 0.248 | 0.198 | 0.333 | 0.635 | [[1893 130] [ 64 32]] |
| Extent of Overfit | 0.07 | 0.73 | 0.773 | 0.652 | 0.343 | |

Pipeline : Ada Boost

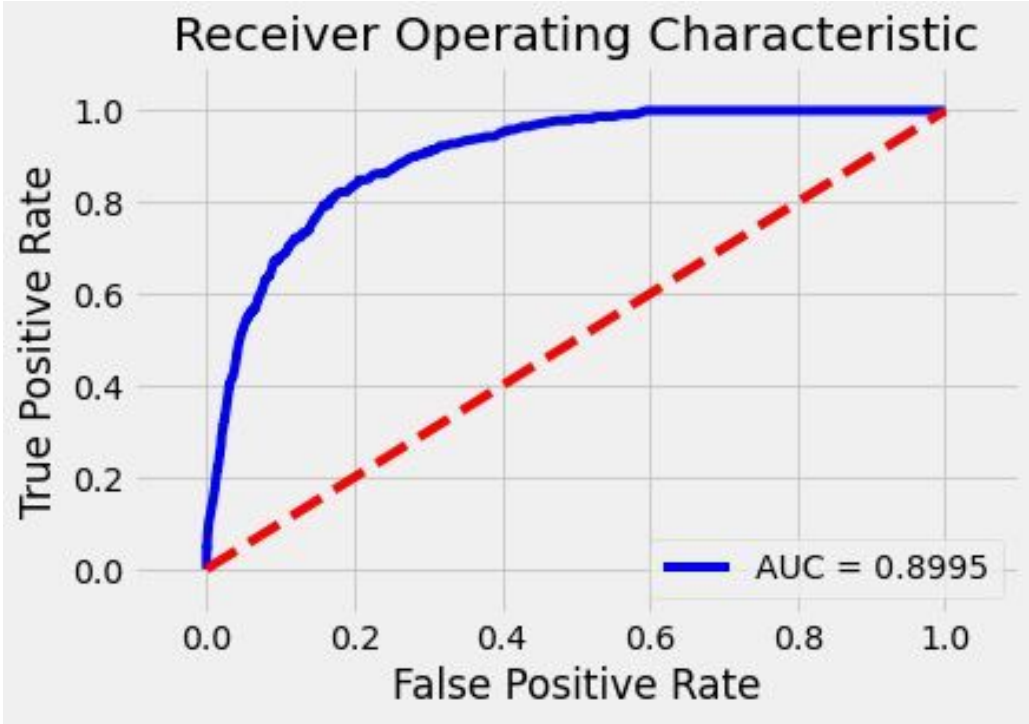
Results of Baseline for Pipeline 4 - Ada Boost

| Dataset | Accuracy Score | F1 Score | Precision Score | Recall Score | Roc-Auc Score | Confusion_Matrix |
|---|---|---|---|---|---|---|
| Training Set | 0.839 | 0.841 | 0.834 | 0.849 | 0.839 | [[4973 1022] [ 917 5140]] |
| Validation Set | 0.844 | 0.324 | 0.202 | 0.823 | 0.834 | [[1710 313] [ 17 79]] |
| Extent of Overfit | -0.005 | 0.517 | 0.632 | 0.026 | 0.005 | |

# Modelling

- **ROC-AUC curve**
- Trade-off between the true-positive rate (ie. Recall - TP/(TP+FN)) and the false positive rate

- **89%** of the time, the model is able to classify correctly the presence of WNV



# Cost and Benefit Analysis

| | False Negatives<br><br>Predicting that West Nile Virus(WNV) is not present when it actually is<br><br>---------- | False Positives<br><br>Predicting that West Nile Virus(WNV) is present when it actually is not<br><br>---------- |
|---|---|---|
| **COSTS** | • **Economic Costs**<br>  ○ Should there be a spread of WNV, may result in an outbreak, forcing businesses to close, as everyone would be asked to stay indoors<br><br>• **Healthcare Costs**<br>  ○ WNV is associated with increased healthcare resource utilization across all phases of care (ie. acute infection, continuing care, final care, up to death) | • **Aerial Spraying Costs**<br>  ○ Overtime hours<br>  ○ Plane rental hours<br>  ○ Pilot hours<br>  ○ Costs of insecticide spray |

# Conclusions and Recommendation

- Multi-pronged approach involving several stakeholders within the community

    - **Preventive**
        - Residents to keep house clean of stagnant water to avoid breeding of mosquitoes
        - Government to raise awareness to society of any clusters nearby so residents can avoid the area and prevent any potential outbreak
        - Healthcare professionals to work on a vaccine to prevent future outbreaks

    - **Detective**
        - Residents to immediately consult a doctor should he/she exhibit symptoms of WNV (eg. high fever, headache, convulsions, muscle weakness etc)
        - Be consistent in spraying efforts for areas expected to have WNV and has had a history of WNV

# Thank You

Questions