Project 4

Group 4
4 of us
4 man
4 months of GA

Methodology

1. Background
2. Data Analysis
3. Modelling
4. Business Case Recommendations

Context

As a result of the outbreak of West Nile Virus in the city of Chicago, our Data Scientist Team has been contracted to help to understand the problem.

1. How can we predict potential outbreak areas of the West Nile Virus? - SPREAD
2. What is the best strategy for controlling the spread moving forward? - CONTROL

Illustrative Timeline of the West Nile Virus

1999
WNV arrives in the USA

2001
WNV first outbreaks identified in Illinois

2002
884 total cases identified in Illinois and 66 deaths

2006
Multiple outbreaks observed of WNV

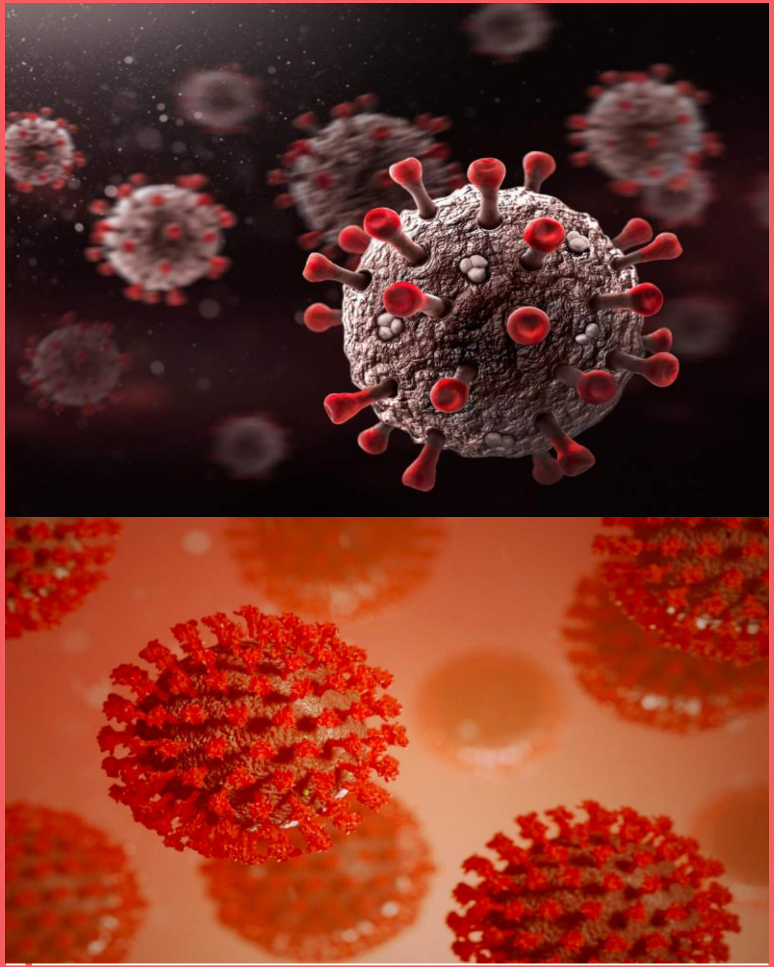
Today
Our team contracted to put forward solution to WNV

52,231

Number of deaths due to WNV since 1999

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.
- West Nile virus is mainly transmitted to people through the bites of infected mosquitoes.



Datasets Overview

Weather

- Date Range: 2007-2014
- Temperatures Max + Min
- Various meteorological factors

Spraying

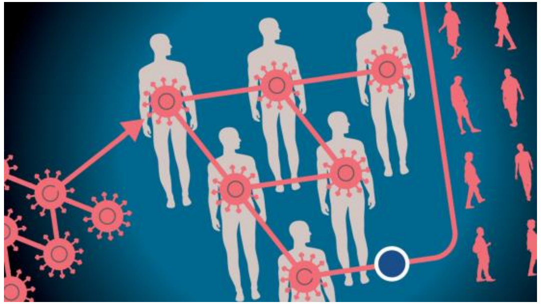
- Date Range: 2011-2013
- Latitude and longitude measurements

Train/Test

- Date Range: 2007-2013
- Location and Number of Mosquitoes
- Species of Mosquitoes

Some Key Issues To Understand - SPREAD

- **What** are the contributing factors/features which are leading to an increase in WNV?
- **Which** mosquitoes are responsible for spreading the virus?
- **How** does WNV infection rates differ over the months?
- **Where** are the major concentrations of WNV?



Some Key Issues To Understand - CONTROL

- **What** are the pros and cons of potential solutions to resolving WNV?

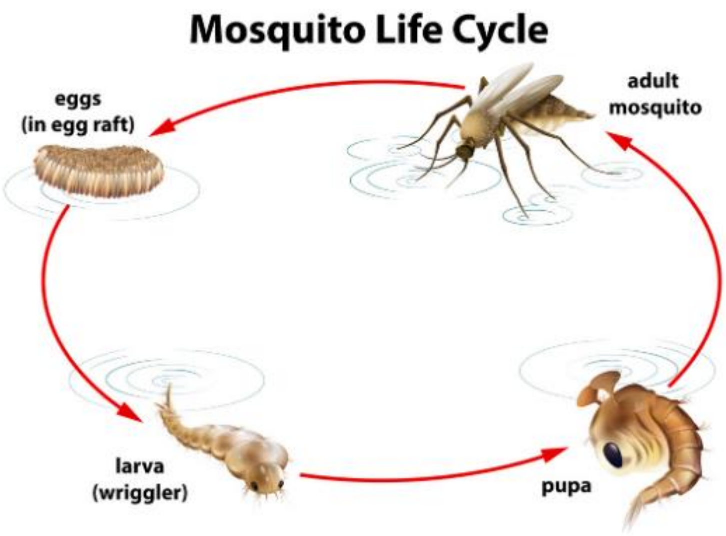
- **Which** is the best strategy we can put forward to control the WNV outbreaks from a cost-benefit point of view (hospitalisation vs spraying costs).

- **How** can we reduce the incidence rates of WNV across Chicago?

- **Where** should we be focusing the city’s solutions and resources?

The Mosquito Life Cycle

- Mosquitoes generally have a 4-stage life cycle.
- Each life cycle takes approximately 14 days, with eggs becoming larvae within 2 days
- Taking this into account, it is best to tackle the Mosquito population levels during their egg/larvae stages
- Therefore, it is important to be aware of two-week interval periods when measuring mosquito levels



Methodology

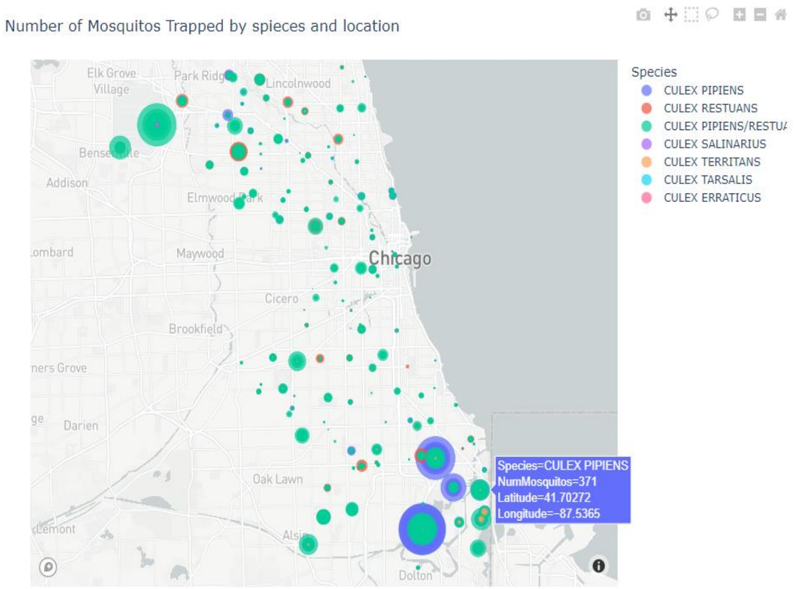
1 Background

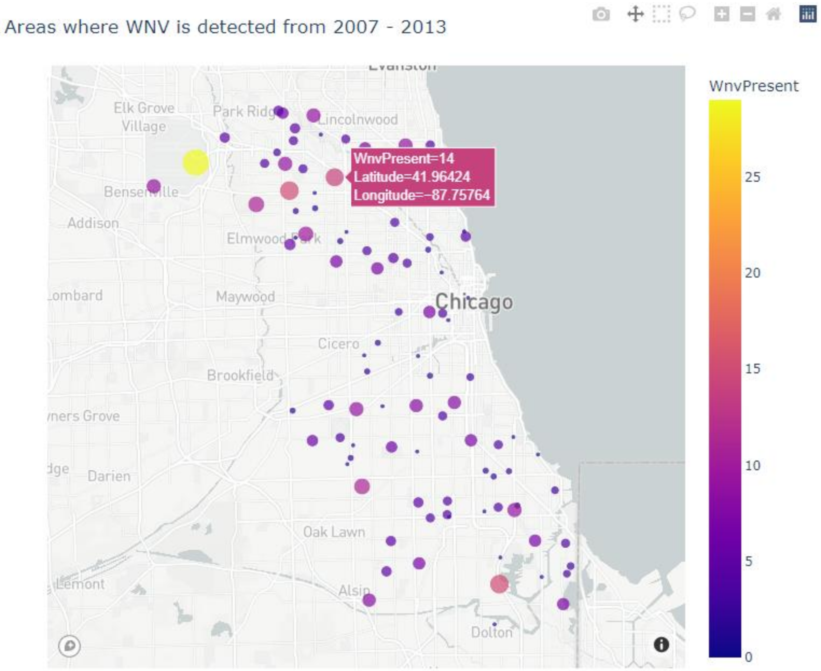
2 Data Analysis

3 Modelling

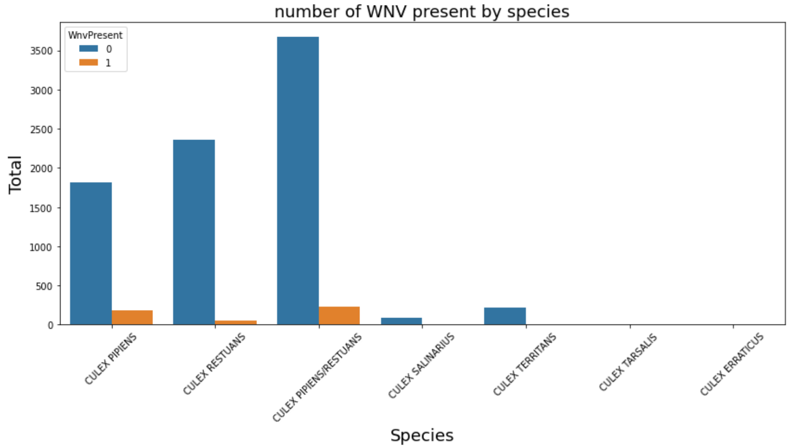
4 Business Case Recommendations

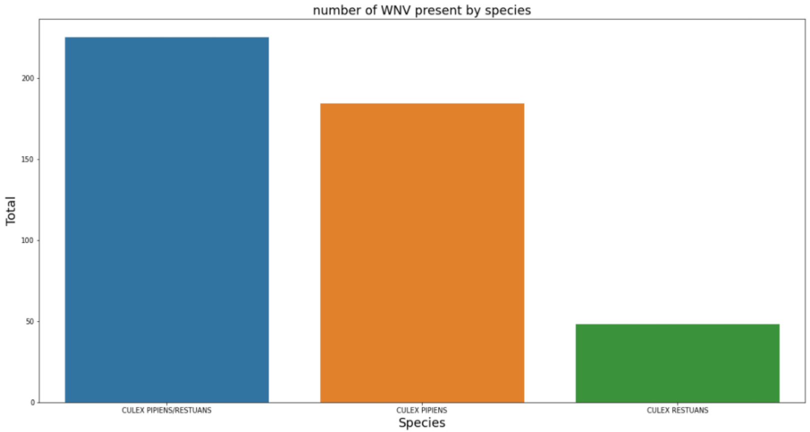
Number of Mosquitoes and Presence of WNV



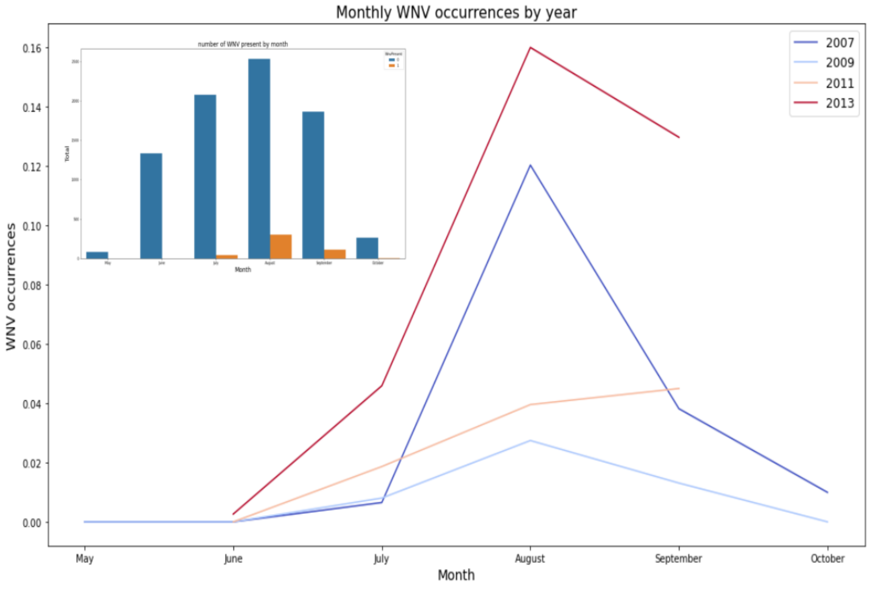


Only 2 of the mosquito species carry the virus





WNV occurrences spike during summer months

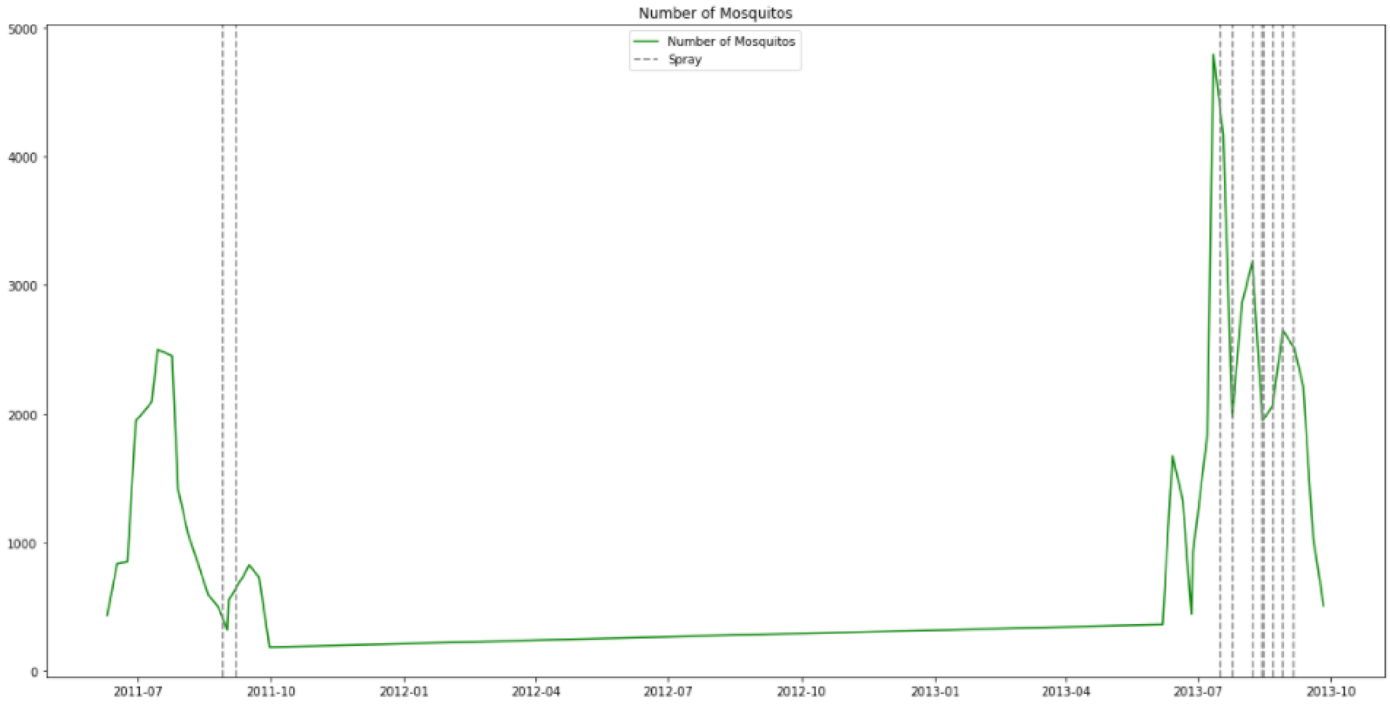


- The presence of WNV in mosquitoes species tends to peak in August

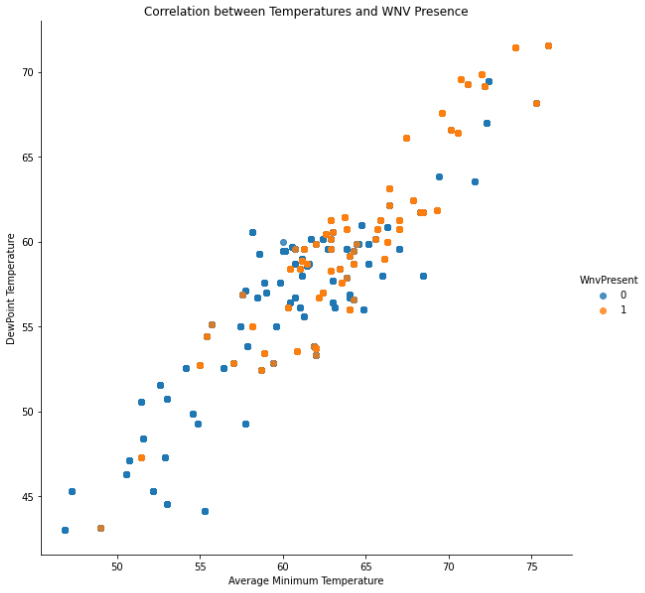
- 2013 has the highest amount of WNV recorded

- Ultimately, the months that have the highest recorded average temperatures (July-September) seem to imply a positive correlation between WNV occurrences and temperature.

Spraying vs No. of Mosquitos



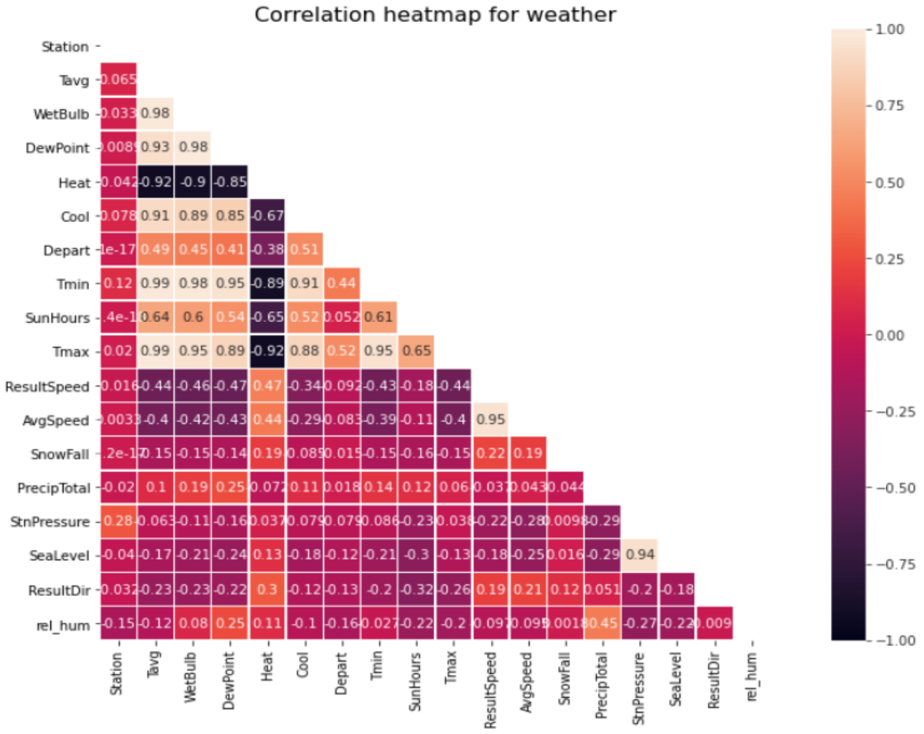
Positive Relationship between WNV and temperature



- As temperature and humidity increase, so does the possibility of WNV being present

- This supports our previous analysis that in the Summer months when temperatures are at high averages there is more WNV presence

Some weather features are highly correlated



Methodology

1. Background
2. Data Analysis
3. Modelling
4. Business Case Recommendations

Data Cleaning

S/N	Action	Remarks
1	Imputation of null values and replace non-numeric cells	1. Replace blank cells and special characters with null. Trim white spaces 2. Replace “M” with null values 3. Replace “T” with 0.005. That is the average between 0 and the smallest number 4. We make use of the other weather station to impute null values as both stations are in Chicago city and thus should have rather similar weather conditions that will not vary drastically. 5. For the rest of the null values, we impute using the next day weather condition. 6. Removed features with more than Water1 and CodeSum features as it as too many null values
2	Duplicates were removed	1. For train dataset, the maximum number of mosquitoes per trap is 50. Thus, we use groupby function to sum the total mosquitoes per trap for each location and day to remove duplicates
3	Removed features that were highly correlated	1. Removed Tavg, ResultSpeed and Wetbulb from weather dataset
4	Assigned weather station to each location and trap	1. Measure the distance between each station and trap and assign each trap to the nearer weather station

Feature Selection/Engineering

S/N	Action	Remarks
1	Created relative humidity and sunhours features	Relative humidity and seasonality are key drivers in WNV epidemiology. Link
2	Created 1 & 2 week average lagging weather conditions	Culex tarsalis, a common California (USA) mosquito, might go through its life cycle in 14 days at 70° F and take only 10 days at 80° F. Link
3	Converted species to % of WNVPresent/Total Mosquitoes	As some WNV is present only in some species and different species have different probability of having WNV present, we decided to add a feature to include the probability of a species having WNV present.
4	Mapped month to % of No. of Mosquitoes/Total Mosquitoes	From EDA, we noted some months have higher number of mosquitoes, we mapped the spread of mosquitoes for each year by the month
5	Created clusters using kmeans	As there are areas where there is higher probability of WNV present, we used location(latitude and longitude) to create 10 clusters within Chicago. This will replace all the address features.
6	Merged train and weather data set	Merged train and weather dataset based on date and weather station to include the 1 and 2 week average lagging weather conditions for each trap and location
7	Removed unnecessary features	Some features (such as dates, address,number of mosquitoes features) are not key features or features not available in test set to predict WNV present
8	Created dummy variables for traps and clusters	Traps and clusters are categorical features

Steps before modelling

Train Validation Split

Perform train validation split with 80% train and 20% validation set

Create Pipeline

- Use SMOTE to account for unbalanced dataset
- MinMax Scaler to normalized the features
- Select Classification model to fit

Grid Search

Use gridsearch with 5 cross fold validation for hyperparameter tuning to select the best set of hyperparameters for each model

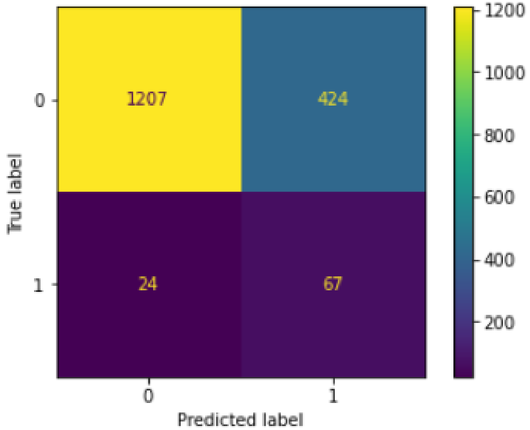
Model 1: Logistic Regression Test Results

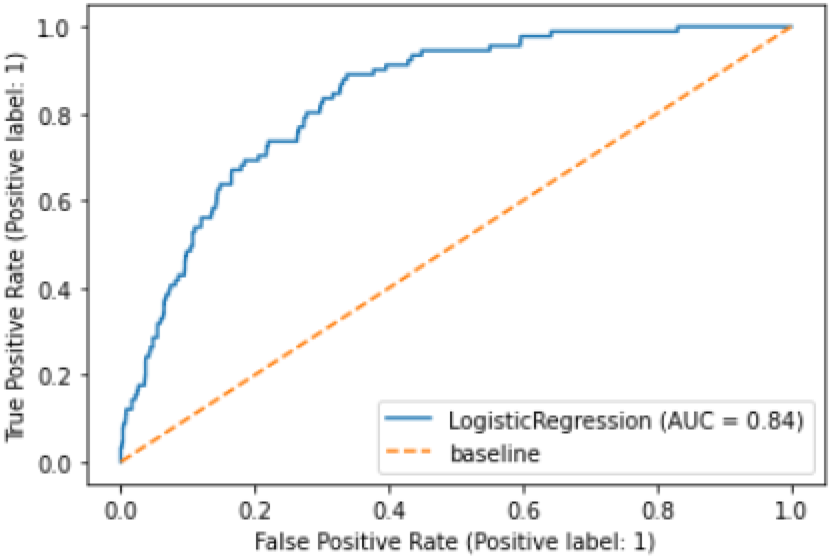
```
Classification Report
              precision    recall  f1-score   support

           0       0.98      0.74      0.84      1631
           1       0.14      0.74      0.23        91

    accuracy                           0.74      1722
   macro avg       0.56      0.74      0.54      1722
weighted avg       0.94      0.74      0.81      1722

Confusion Matrix
<sklearn.metrics._plot.confusion_matrix.ConfusionMatri
```





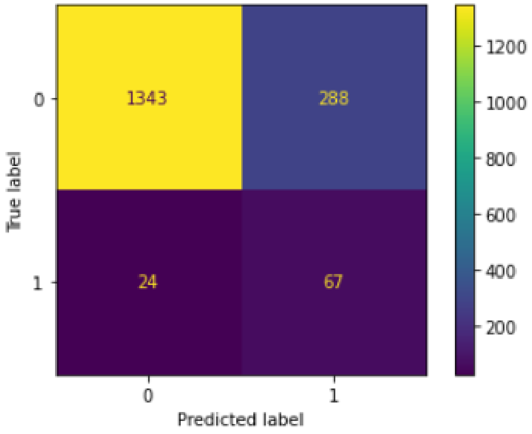
Model 2: Random Forest Test Results

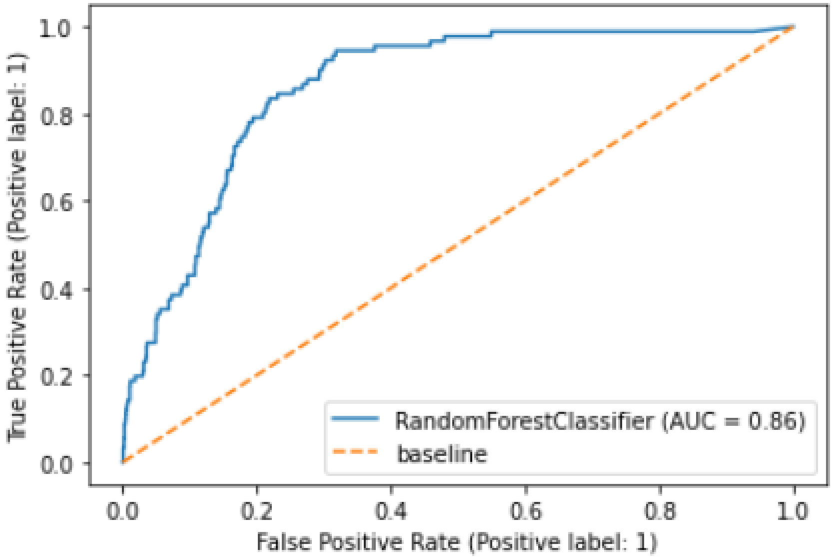
```
Classification Report
              precision    recall  f1-score   support

           0       0.98      0.82      0.90      1631
           1       0.19      0.74      0.30        91

    accuracy                           0.82      1722
   macro avg       0.59      0.78      0.60      1722
weighted avg       0.94      0.82      0.86      1722

Confusion Matrix
<sklearn.metrics._plot.confusion_matrix.ConfusionMatri
```





Model Comparison shows Random Forest is better

	Model_name	train_accuracy	test_accuracy	accuracy_variance	train_auc	test_auc	Precision	recall	fscore
0	logistic regression	0.799	0.740	1.080	0.869	0.738	0.936	0.740	0.811
1	randomforest	0.819	0.819	1.001	0.821	0.780	0.940	0.819	0.864

Model Selected due to the following:

1. Based on the train accuracy vs test accuracy(accuracy variance), we can see that logistic regression is slightly overfitted and thus will not generalised well as compared to the RandomForestClassifier
2. As this is data classification set is highly unbalanced, we should not be using accuracy to evaluate model. Thus, we select RandomForestClassifier as it has the higher test_auc score, recall and fscore

Conclusion: We refit the whole train dataset using RandomForestClassifier to make use of the full train dataset. Subsequently, we use this model to predict WNV present for the test set for Kaggle submission.

Methodology

1 — 2 — 3 — 4

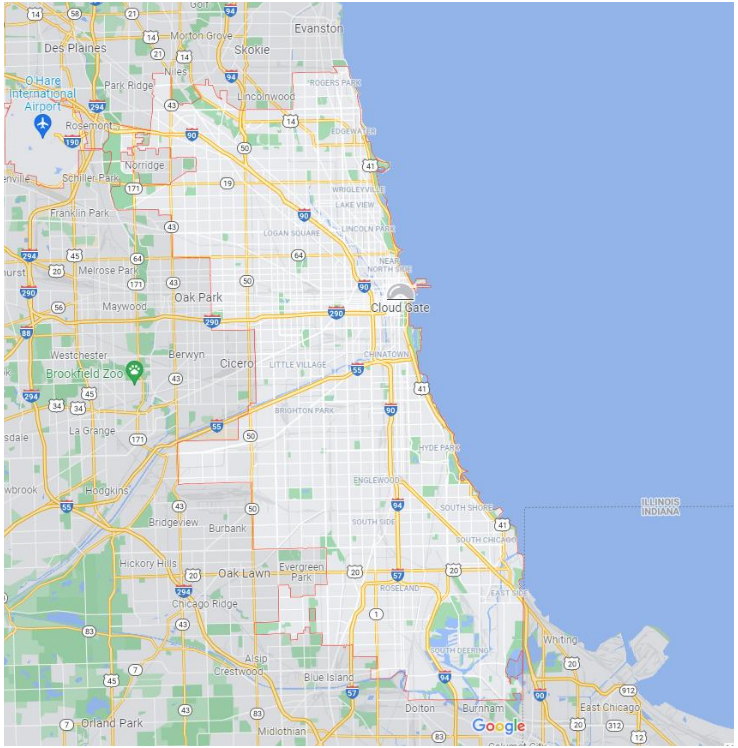
Background

Data Analysis

Modelling

Business Case Recommendations

Business Case Recommendations (Definitions)



Costs:

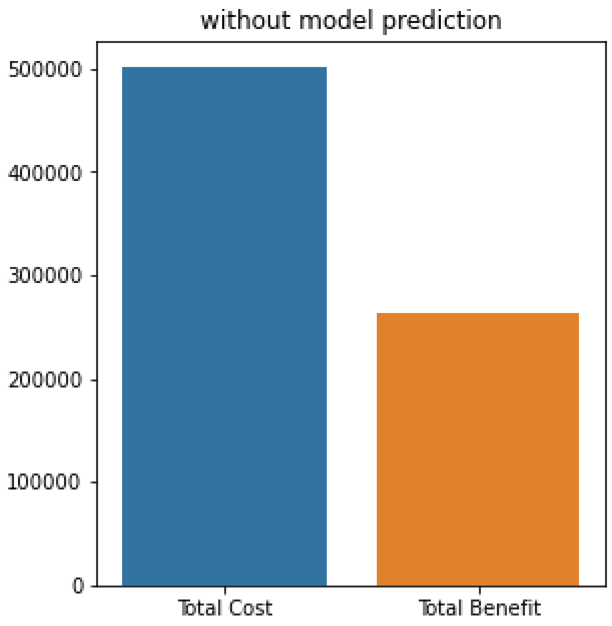
Cost of spraying of Zenivex pesticide in the city of Chicago.

Benefits:

The savings made in medical treatment costs from preventing WNV infection of persons.

Business Case Recommendations (w/o model)

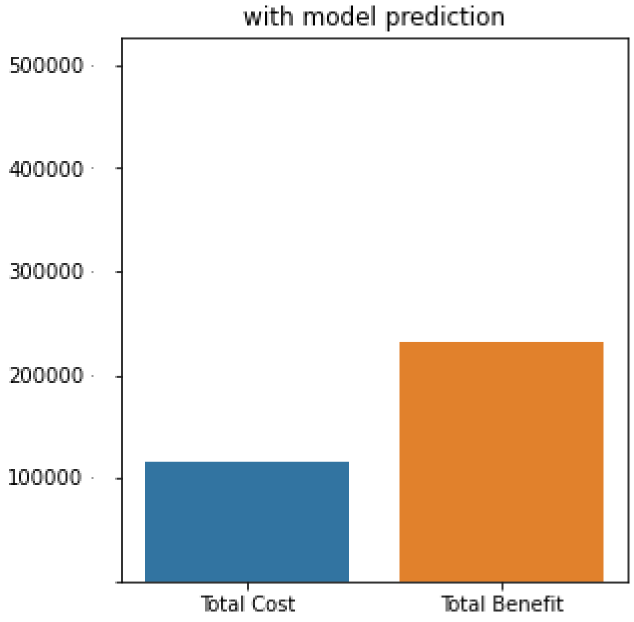
	Calculations	Amount (USD)
Costs	Chicago_area = 149800 acres Cost_pesticide_unit = 0.67 /per acre. total_costs_nopred = Chicago_area * Cost_pesticide_unit * 5 - 5 months for the summer months when mosquitoes are active.	501830.0
Benefits	For 117 cases average cost is $136,839 [1] average_cost = round(136839/117,2) = 1169.56 annual_cost_treatment = round(average_cost *225,2) = 263151.0 - 225 is the worst case in 2002 when there is no spraying done in Chicago.	263151.0
Total Benefit	Benefits - Costs	238679.0



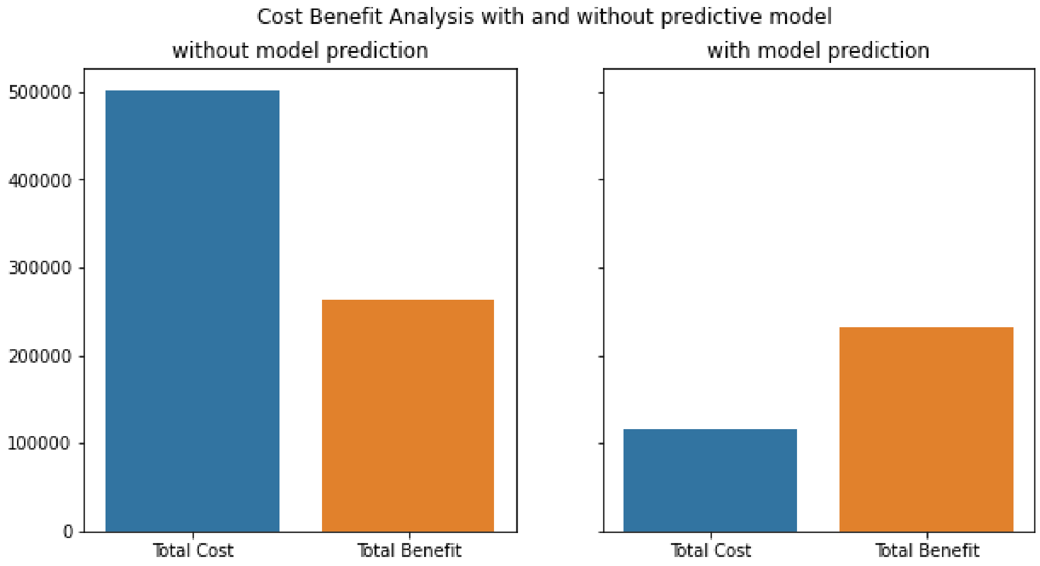
[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

Business Case Recommendations (w model)

	Calculations	Amount (USD)
Costs	percentage_area = WNV Present/ Total Rows = 23% total_costs_pred_spray = percentage_area * total_costs_nopred	115420.9
Benefits	total_benefit_pred = annual_cost_treatment * recall - Recall is 87% on used data.	231572.88
Total Benefit	Benefits - Costs	116151.98



Business Case Recommendations (Summary)



The total saving is estimated to be around **US$354830.98 with the prediction model.**
Using the predictive model, the total benefit outweighs the costs, and makes it worthwhile to conduct the spray exercise.

Appendix

Appendix - Project Management

S/N	Name	Links
1	Trello	https://trello.com/b/bnei7fte/project-space
2	Google Co-lab	https://colab.research.google.com/drive/1DqYGpvpRRNNQPLh0-kXvD7RBpym8bC8d?usp=sharing#scrollTo=ca1c6d86

Kaggle Submission

Name	Submitted	Wait time	Execution time	Score
submission.csv	just now	1 seconds	1 seconds	0.60781

Complete

Submission and Description	Private Score	Public Score	Use for Final Score
submission.csv a few seconds ago by Jeryll Chan add submission details	0.60786	0.60781	☐