Predicting the West Nile Virus in Chicago

DSI13 Project 4 Group 3
Toh Jun Kai, Tan Hueeming, Huang Shilin, Elton Yeo

Overview

- Problem Statement and Context
- Data Cleaning
- Data Visualisation
- Modelling
- Cost-Benefit Analysis
- Recommendations
- Next Steps

Problem Statement and Context

To predict the when and where the West Nile Virus will occur in mosquitos by taking into account a range of variables (e.g. location, temperature etc.).

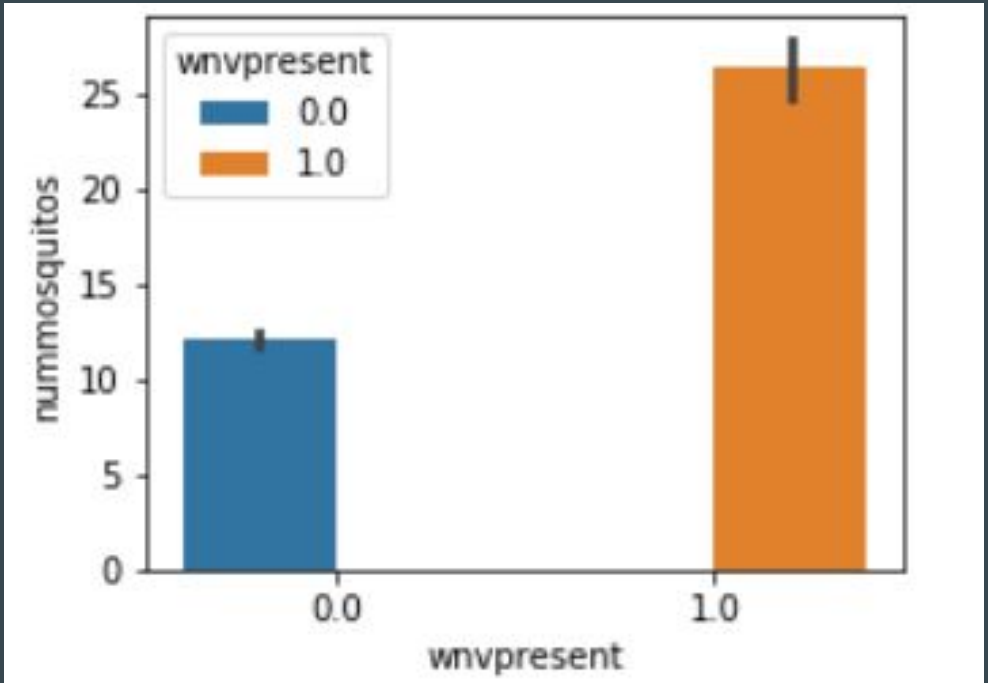
Models: Linear regression, K-nearest neighbours, Random forest

Evaluation: Receiver Operating Characteristic (ROC) Area Under Curve (AUC) score

Data Cleaning - Train and Test

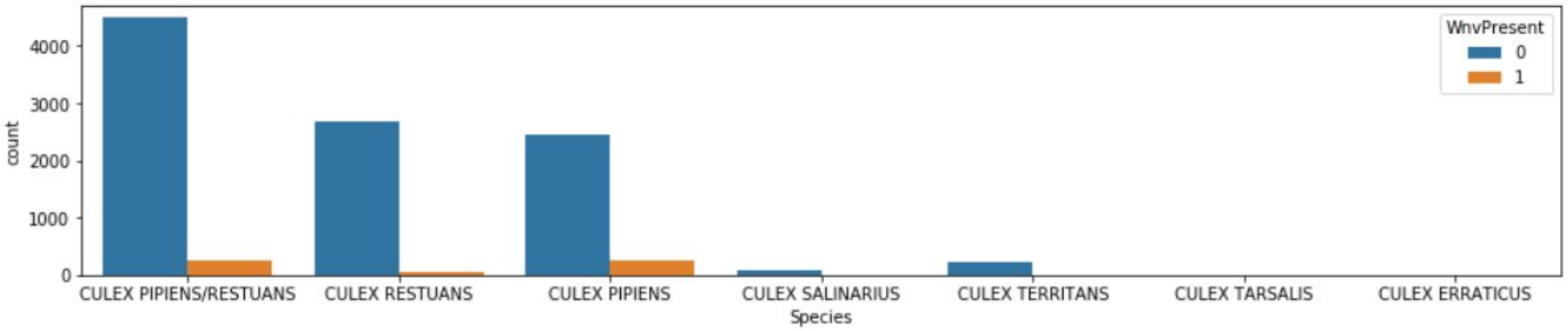
Data Visualisation - Train

The more mosquitos in the trap, the more likely WNV is present.



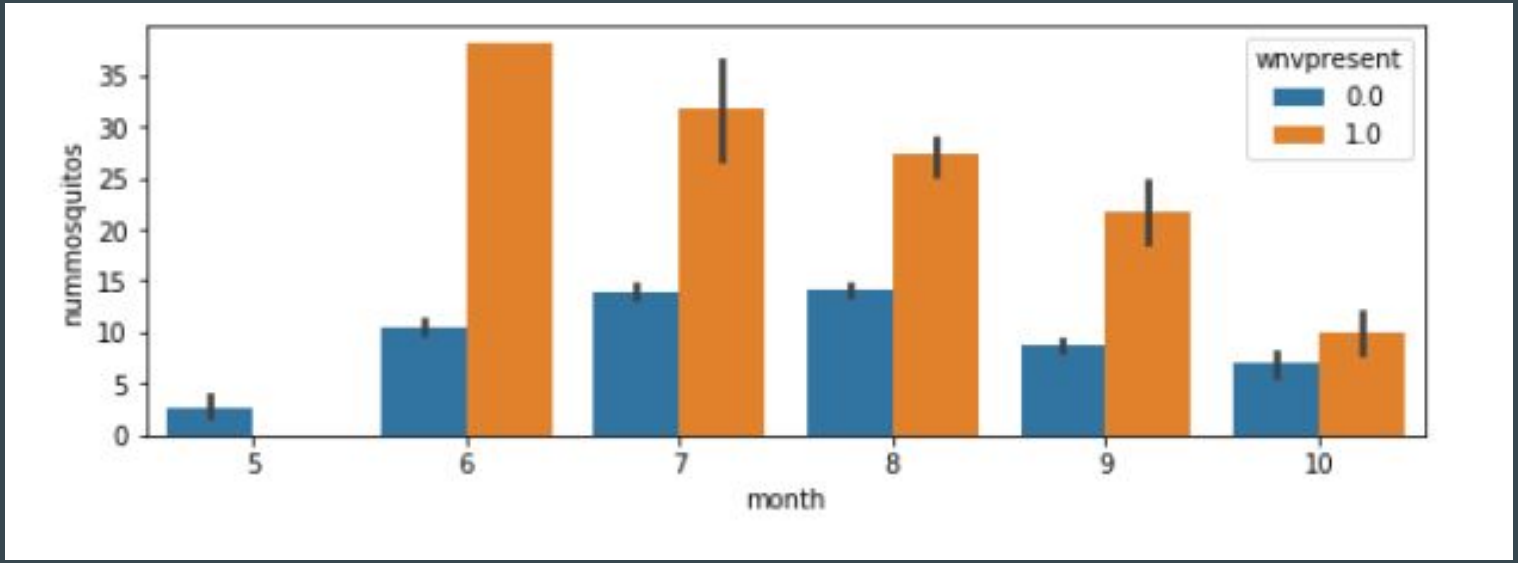
Data Visualisation - Train

WNV is only transmitted by Culex Pipiens and Culex Restuans.



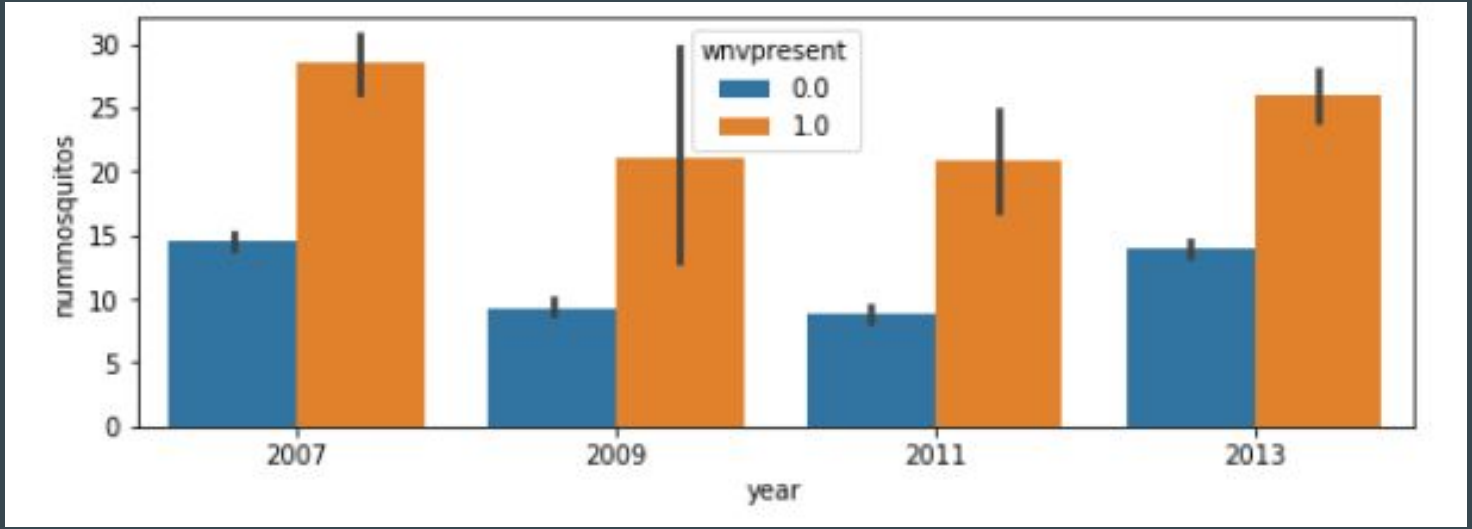
Data Visualisation - Train

Across all years in the train dataset, the number of mosquitoes peaked in June.



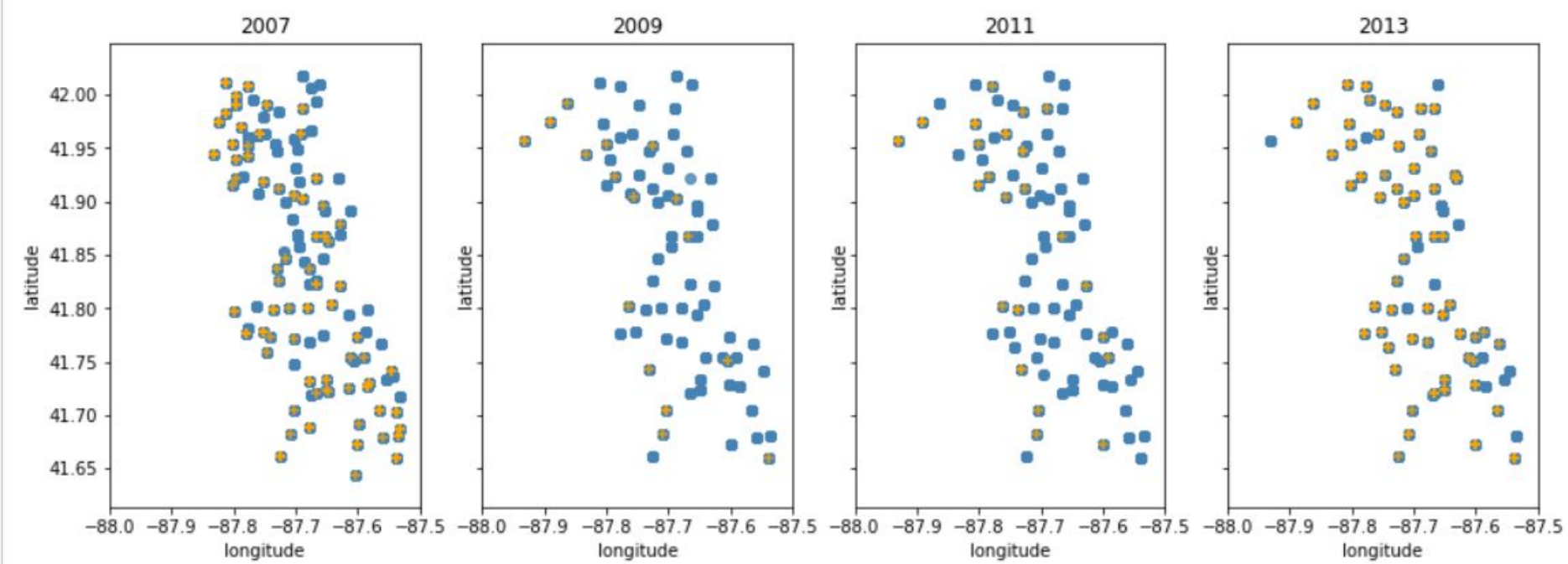
Data Visualisation - Train

Across all years in the train dataset, 2007 had the highest number of mosquitos.



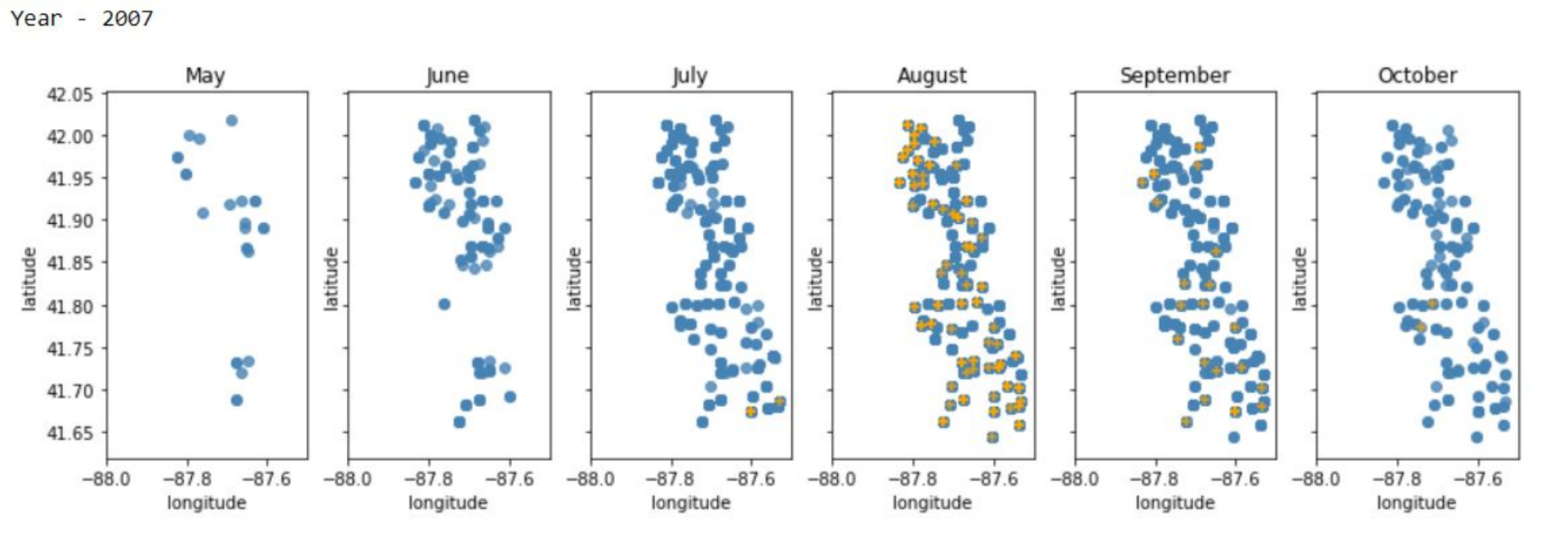
Data Visualisation - Train

Spread of WNV across the years in the train dataset.



Data Visualisation - Train

Spread of WNV across the months in 2007 only.

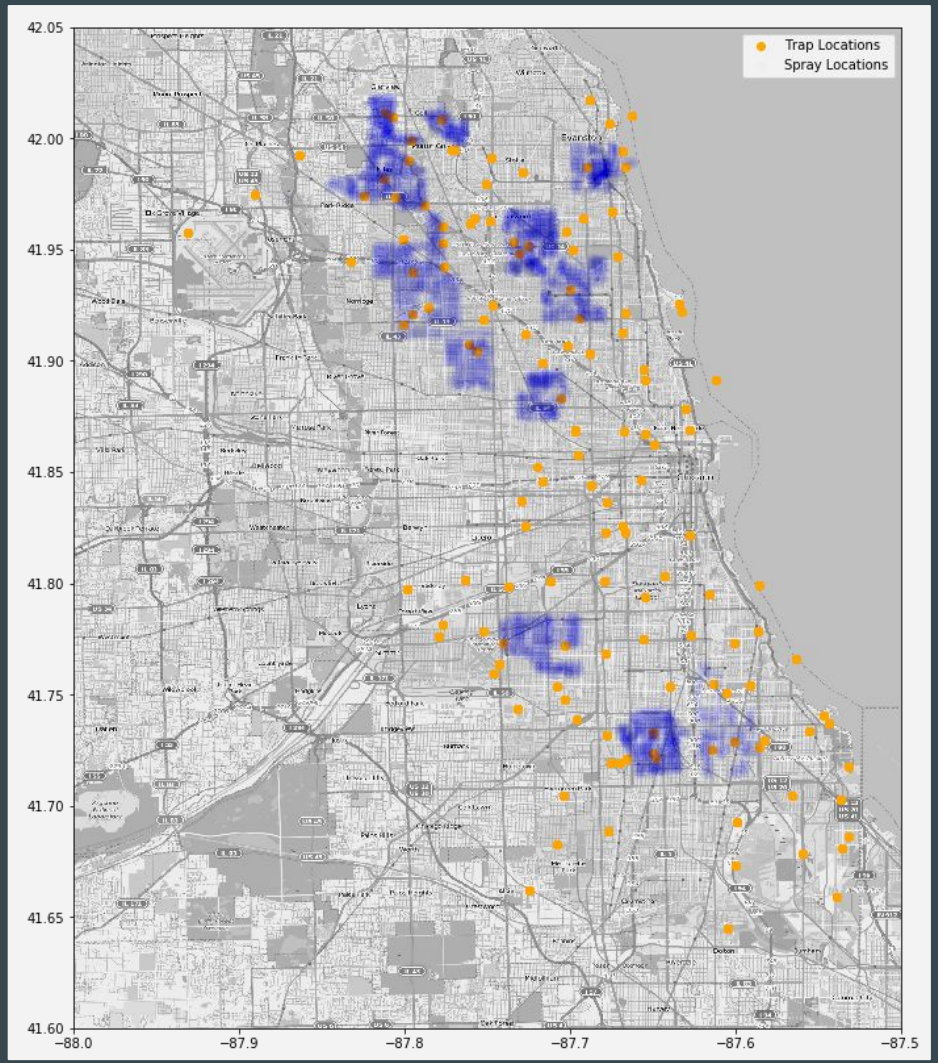


Data Cleaning - Spray

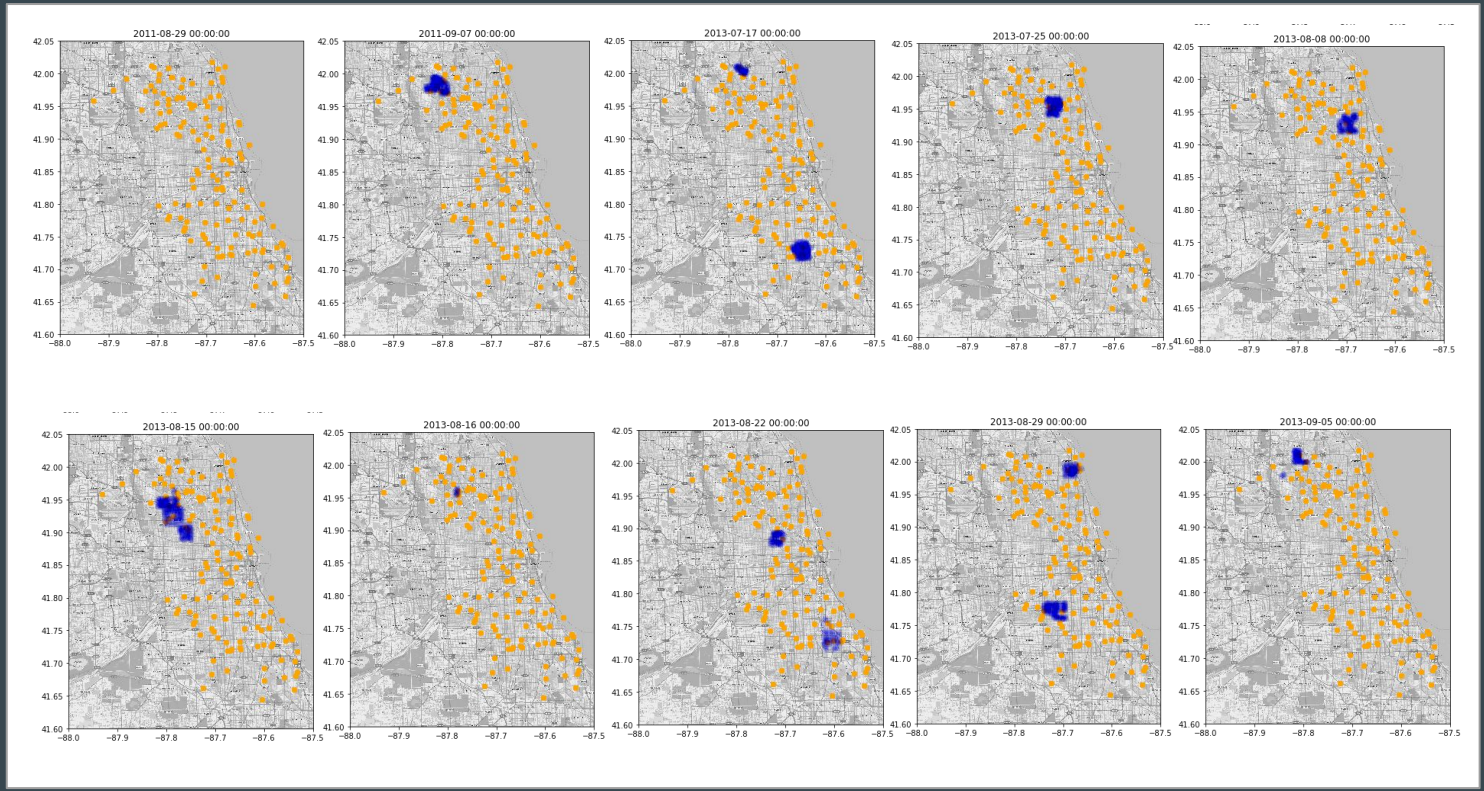
	Date	Time	Latitude	Longitude
0	2011-08-29	6:56:59 PM	42.391623	-88.089163
1	2011-08-29	6:57:08 PM	42.391348	-88.089163
2	2011-08-29	6:57:18 PM	42.391022	-88.089157
3	2011-08-29	6:57:28 PM	42.390637	-88.089158
4	2011-08-29	6:57:38 PM	42.390410	-88.088858

object -> datetime64

Delete duplicate entries
41.98646 (541)
41.983917 (2)



Data Visualisation - Spray



- No pattern observed
- Inconsistent reporting periods
- Spray data available for 2011 and 2013
- Not available in test set

Data Cleaning - Weather

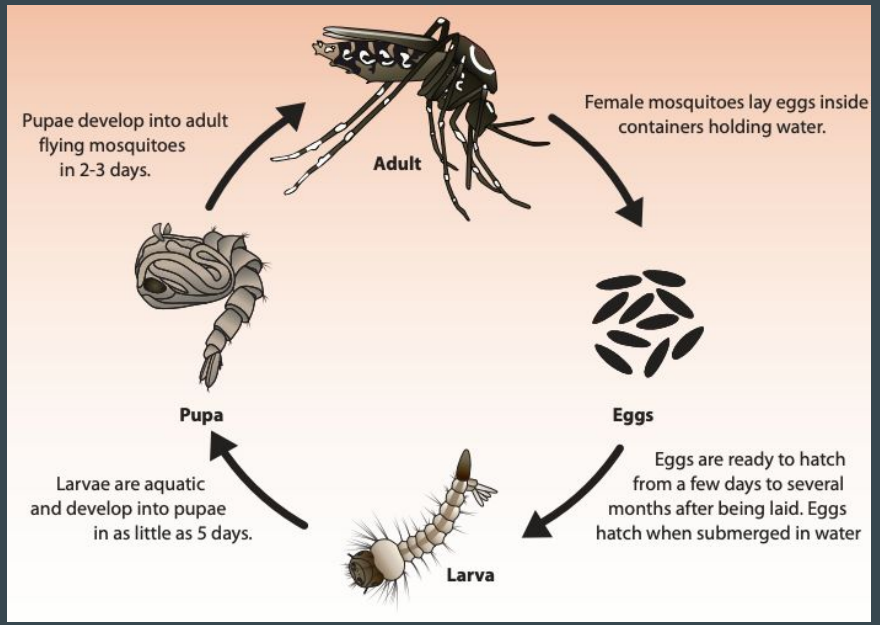
Changing ‘Date’ column data type from object to datetime.

Add dates, breakdown by Year, Month, Week and Day

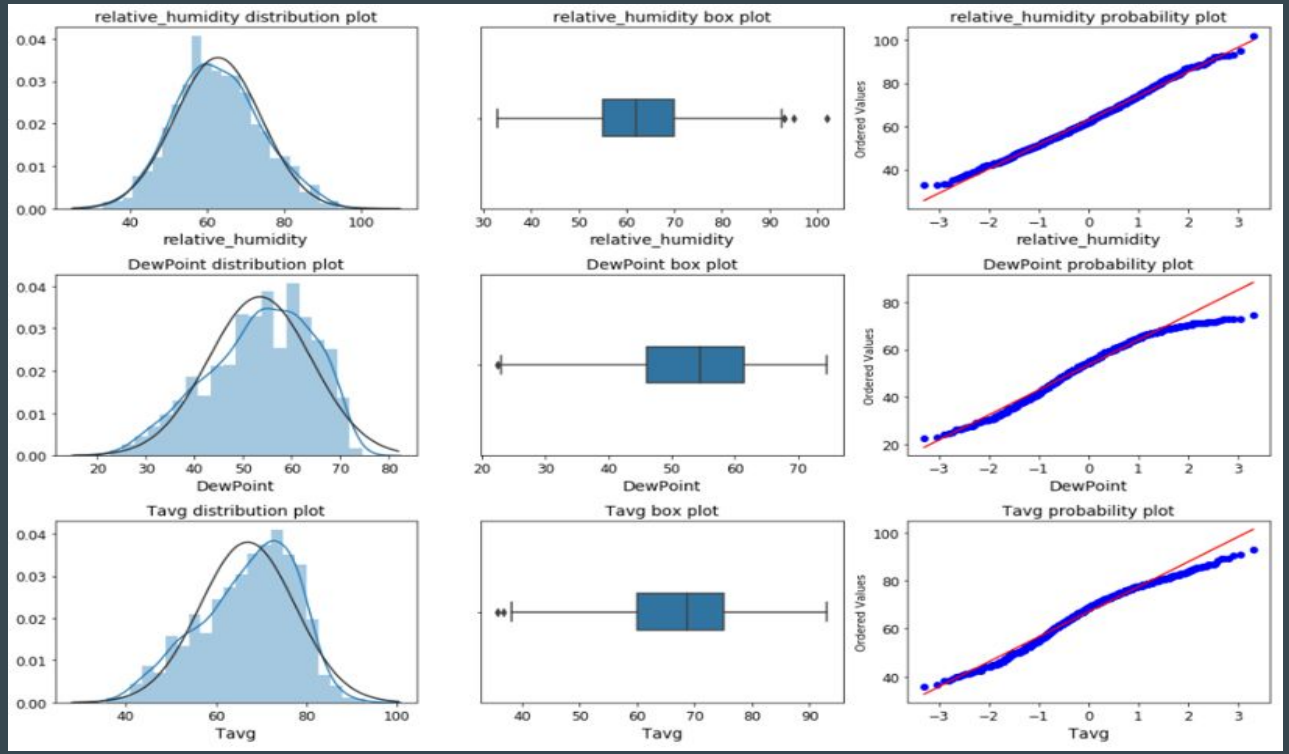
Fill values in column PrecipTotal labelled T and M with zero.

Weather Feature Engineering

1. Add daily temperature range. Tmax minus Tmin
2. Add relative humidity
3. Add number of days since it last rained
4. Introduce 14 days lagged weather feature



Data Visualisation - Weather



Baseline - AUC ROC

- ~~Accuracy~~
- AUC ROC - Area Under The ROC (Receiver operating characteristic) Curve
- AUC ROC > 0.5

Modelling

	Logistic Regression	K-Nearest Neighbors	Random Forest
Accuracy	0.69	0.93	0.66
Recall	0.75	0.07	0.87
Precision	0.12	0.18	0.12
ROC AUC	0.78	0.71	0.83

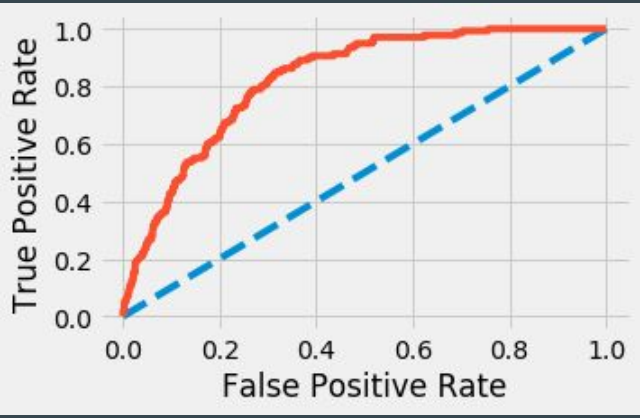
* Certain hyperparameters for all models are optimized using RandomizedSearchCV, Pipeline, and StandardScaler

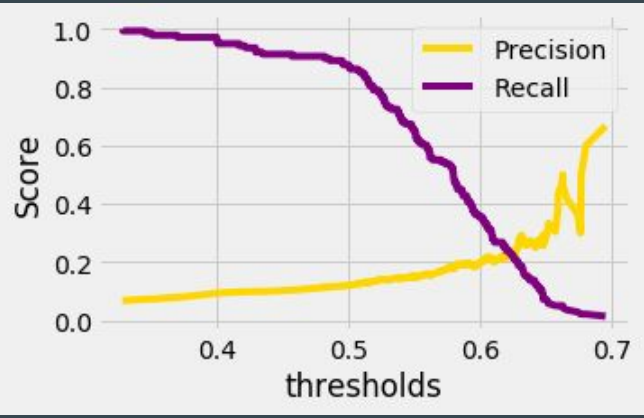
Consequences of model error

False Positive	False Negative
Inconvenience to citizens due to vector control measures e.g. spraying, avoiding certain areas during certain times 	Higher risk of contracting the virus without any mitigating factors. 

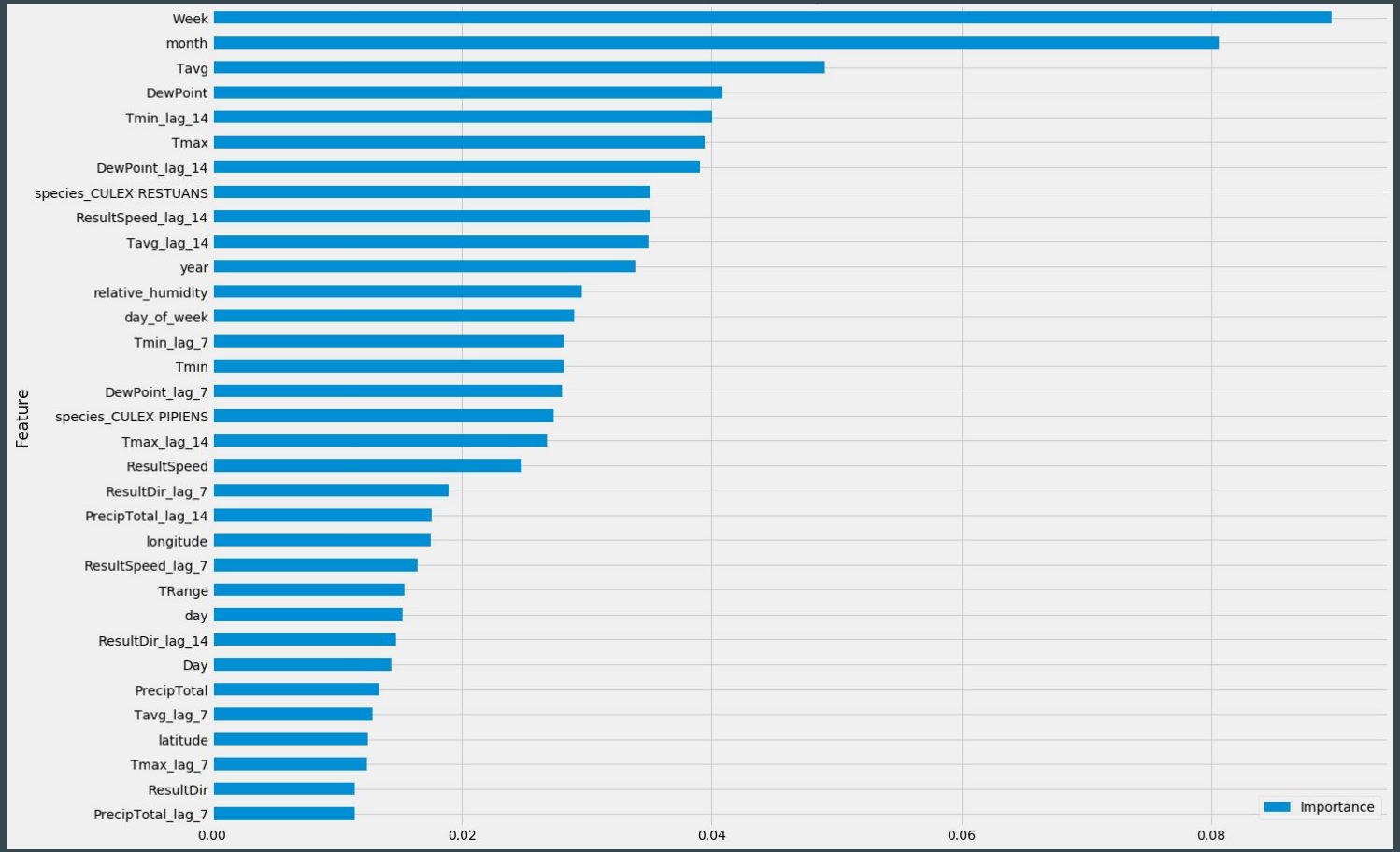
Modelling - Random Forest

- ROC AUC Score - 0.83





Features Importance



Cost-Benefit Analysis

	Spraying	**Not Spraying**
Cost	~$7 million per year Inconvenience to citizens due to vector control measures	~$17 million per year Lives lost, medical resources spent on those who contract the virus, impact on economy due to fewer tourists/workers
Benefit	Lives, and medical resources (amounting to ~$10 million) saved	Greater convenience to citizens

Recommendations

1. STRATEGICALLY FOGGING
2. CONTROL TALL GRASS AND SHRUBBERY
3. ELIMINATE /TREATING STANDING WATER
4. MOSQUITO STERILIZATION

Next Steps

Our ROC AUC score was **0.83**; however, our Kaggle score was **0.61**. Why this disparity?

It is possible that our models are almost entirely modeling noise. What can we do?

1. Restructure the data

2. Get more, cleaner data


Thank You!