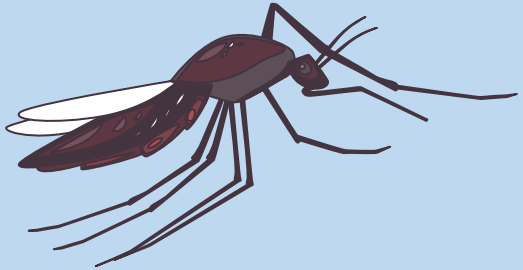
Project 4: West Nile Virus Prediction



Dillon, Shaalini, Vee Vian & ZheQin

Contents

Introduction

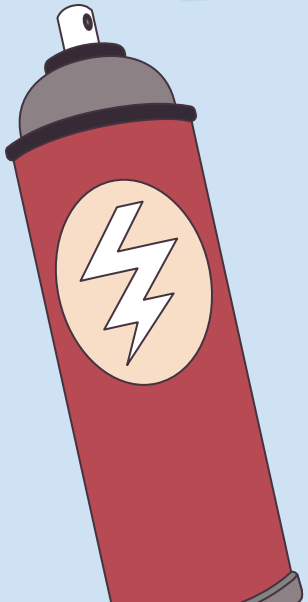
Due to the recent epidemic of **West Nile Virus (WNV)** in **Chicago**, we've had the Department of Public Health set up a surveillance and control system. Through data collection, we will learn about mosquito population to **derive an effective plan to deploy pesticides** throughout the city.

Primary stakeholders: CDC, Centers for Disease Control and Prevention
Secondary stakeholders: Government of Chicago.

Problem Statement

As Data Scientists from the **DATA-SCIENCE**, *Disease And Treatment Agency, division of Societal Cures in Epidemiology and New Creative Engineering*, our task is to:

1) Build a **model(LR/RF)** and **make predictions** to determine the *period and location of pesticide spraying* in Chicago
2) Conduct a **cost-benefit analysis** for cost of spraying vs economic/social cost and **provide feasible recommendations to reduce WNV infection rate**



West Nile virus

Transmission

Spread to people by the *bite of an infected mosquito* (feed on infected birds).

Symptoms

- No symptoms in most people.
- Febrile illness (fever) in some people - 1 / 5 people
- Serious symptoms in a few people - 1 / 150 people

Treatment

No vaccine or specific medicines are available yet

Data Cleaning

Exploratory Data Analysis (EDA)

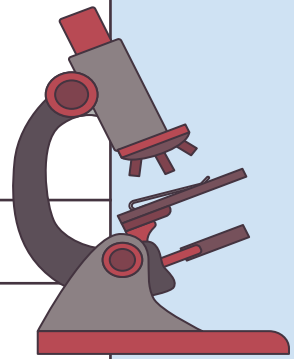
Pre-processing & Feature Engineering

Overview of Datasets

Train and Test Dataset	Main datasets where public health workers set up mosquito traps across the city to capture mosquitoes and test for the presence of West Nile virus.
Weather Dataset	Contains information about the weather condition of Chicago from 2 different Weather Stations
Spray Dataset	Contains details of pesticide spraying in Chicago such as location, date and time of spraying

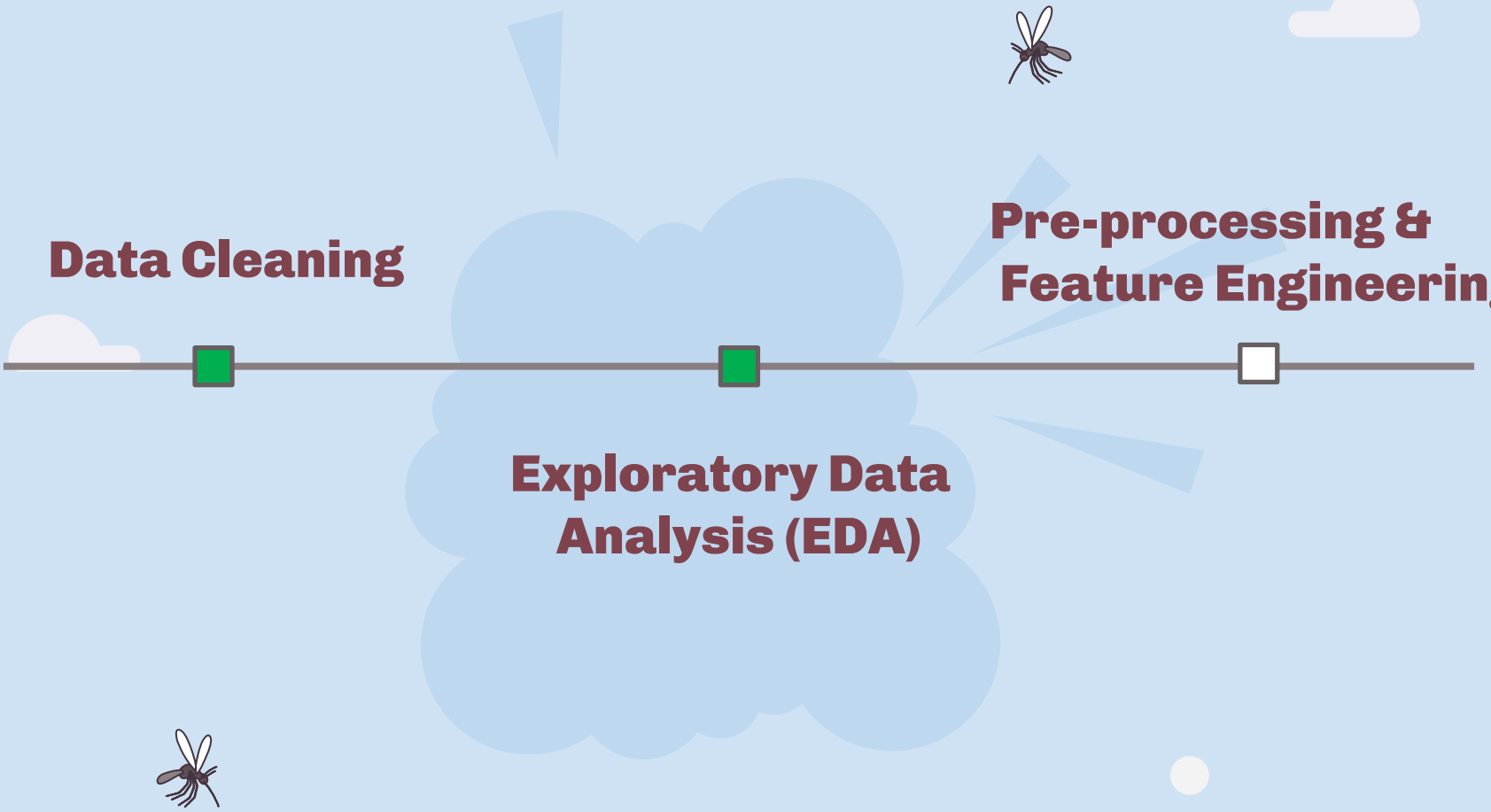
Data Cleaning

Lowercase	All column names to lowercase
Duplicates/Missing values	Drop rows with missing values and duplicates
Removal of Columns	Spray dataset - time Weather dataset - snowfall, depth, water1
Creating new features	Train and Test dataset *nearest station* to identify which weather station is nearest to the coordinates of the trap Weather dataset *trange* for temperature range between tmax and tmin Train, Test and Weather dataset *year, month, day*
Reshape	Traps with > 50 mosquitos combined for the same location
Conversion of Dtype	Date to Datetime type



Summary of the Datasets

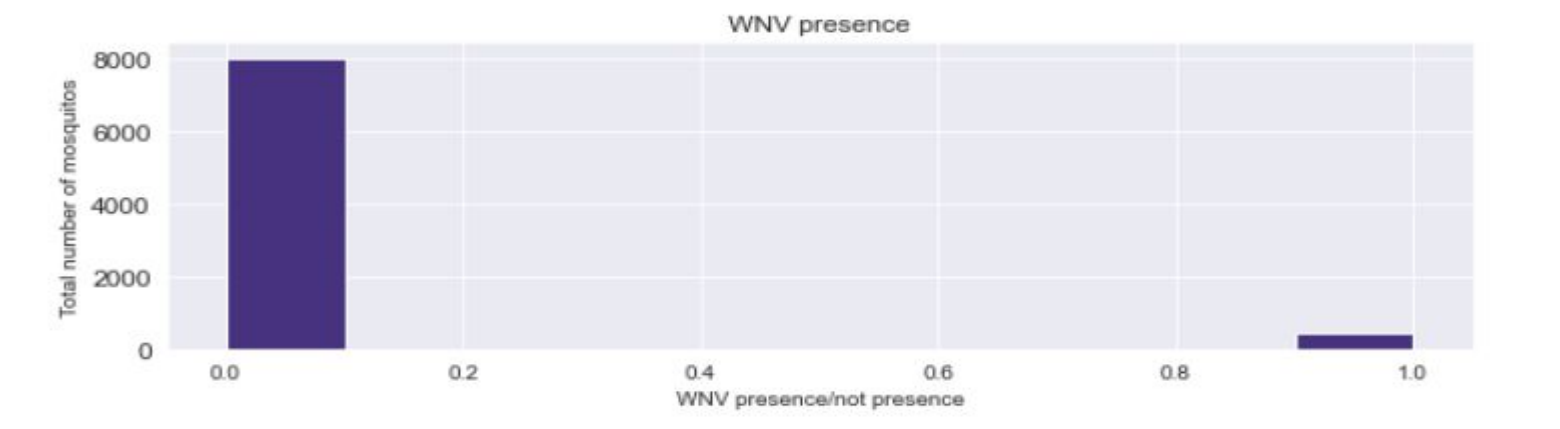
	Period								Rows		Columns	
Dataset	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Bef.**	**Aft.**	**Bef.**	**Aft.**
Train	✡		✡		✡		✡		10,506	8,475	12	16
Test		✡		✡		✡		✡	116,293	116,293	11	15
Spray					✡		✡		14,835	14,294	4	3
Weather	✡	✡	✡	✡	✡	✡	✡	✡	2,944	2,918	22	22


Data Cleaning
Exploratory Data Analysis (EDA)
Pre-processing & Feature Engineering

Exploratory Data Analysis (EDA)

Train Dataset

WNV Present, in %

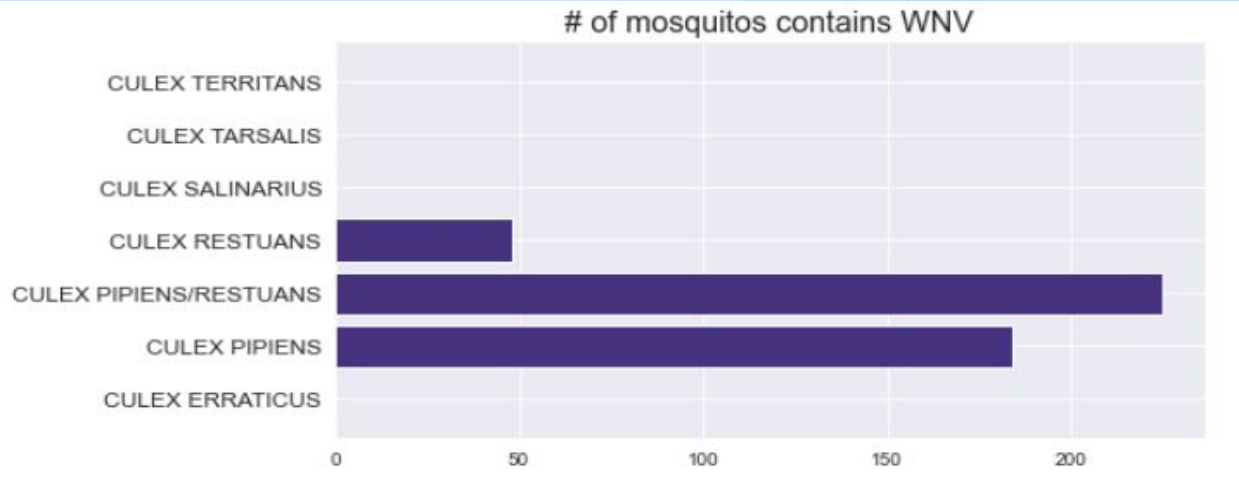


- 95% train data with no WNV present, while **only 5% with WNV present**
- Highly **unbalanced dataset**

Exploratory Data Analysis (EDA)

Train Dataset

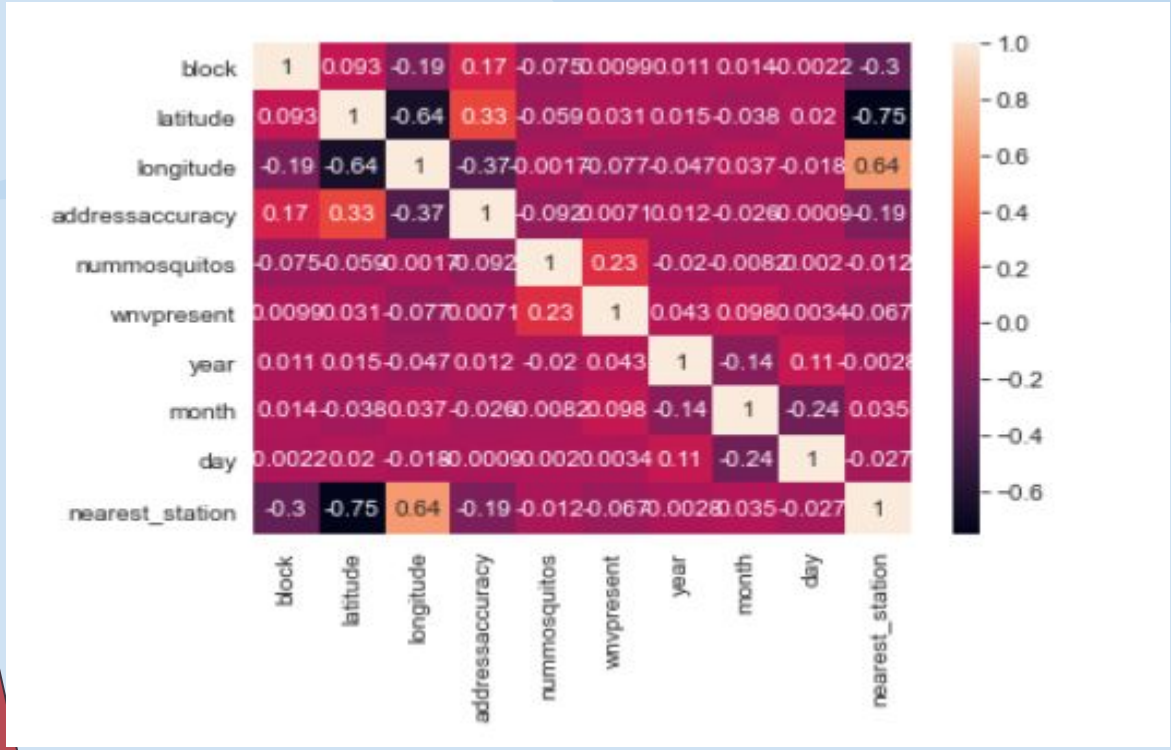
WNV Present, in # by species



- Top 3 species made up **> 96%** of the sample of the species sampled
- They are the **only species** detected with WNV presents

Correlation:

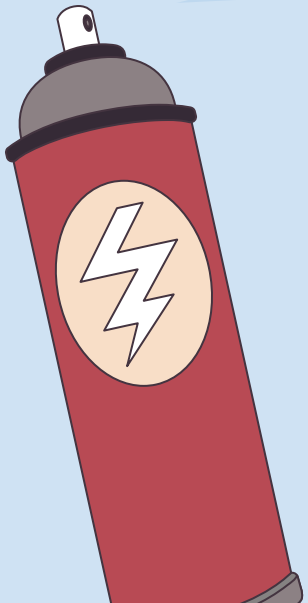
Train Dataset



Correlation:

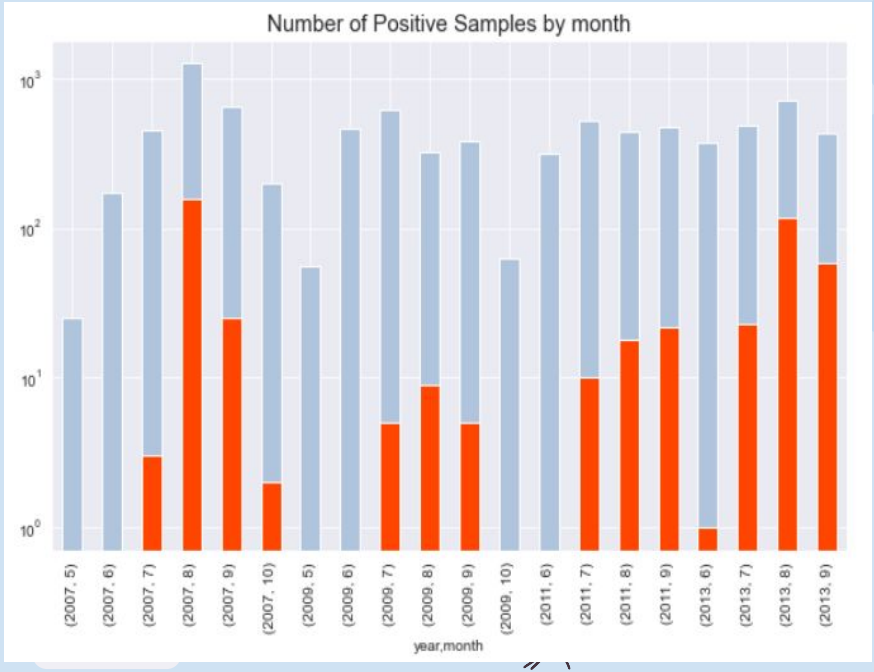
Train Dataset

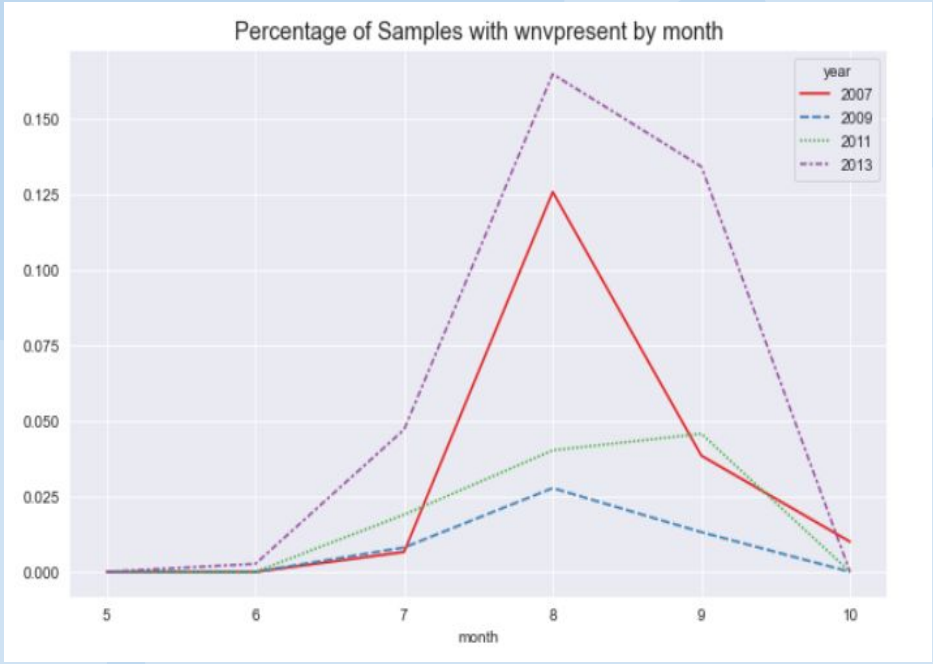
1	wnvpresent	1.000000
2	nummosquitos	0.233532
3	month	0.097948
4	year	0.043038
5	latitude	0.030862
6	block	0.009859
7	addressaccuracy	0.007057
8	day	0.003400
9	nearest_station	-0.066947
10	longitude	-0.076732



Exploratory Data Analysis (EDA)

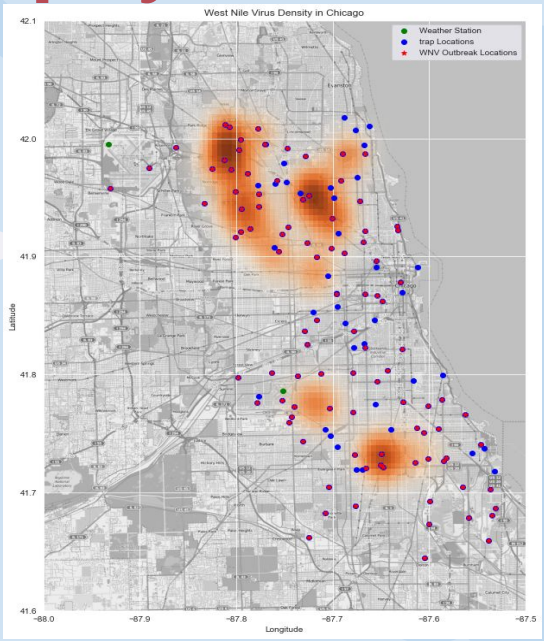
Train Dataset



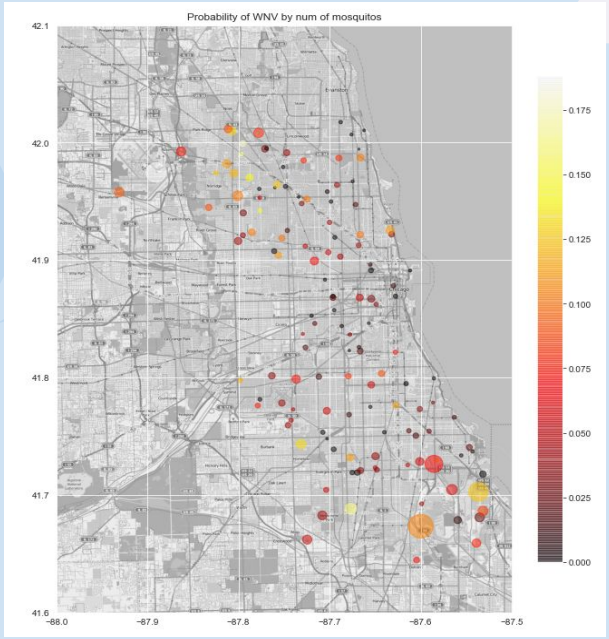


Exploratory Data Analysis (EDA)

Spray & Train Datasets



Spray and Train trap locations

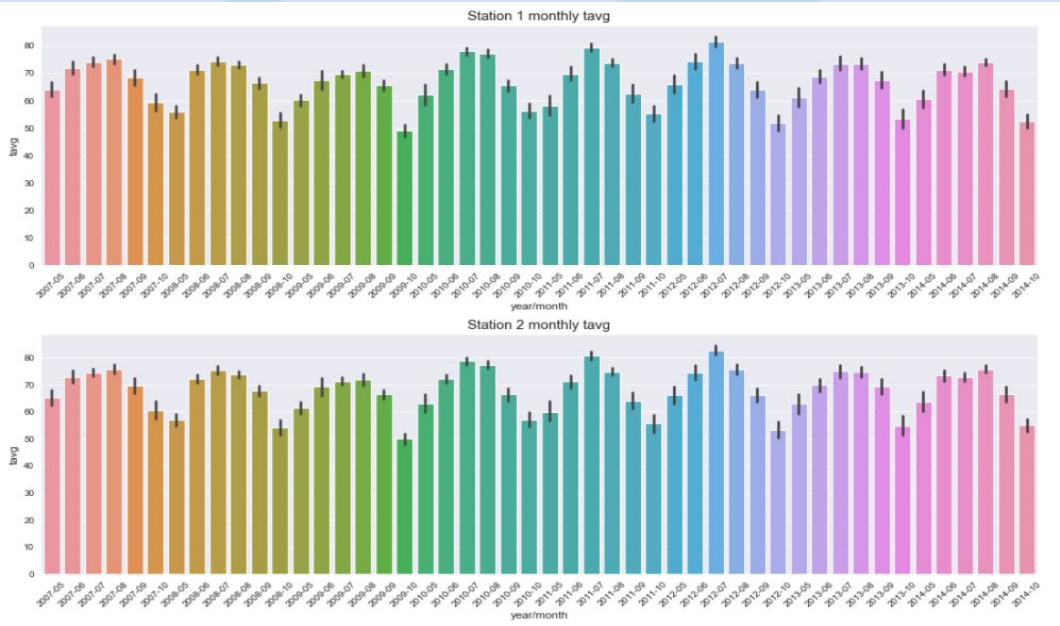


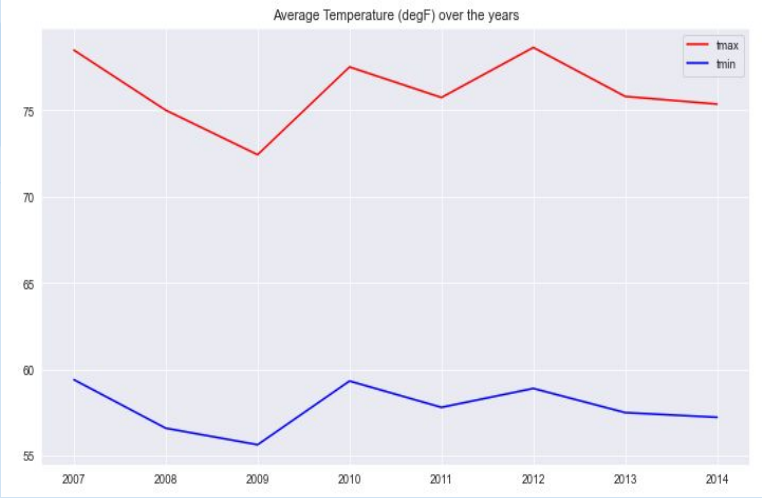
Probability of WNV by nummosquitos

- Traps are spread out across the Windy City
- Area with darker orange - region with more spray concentration
- Previous sprays did not cover most of the WNV outbreak area

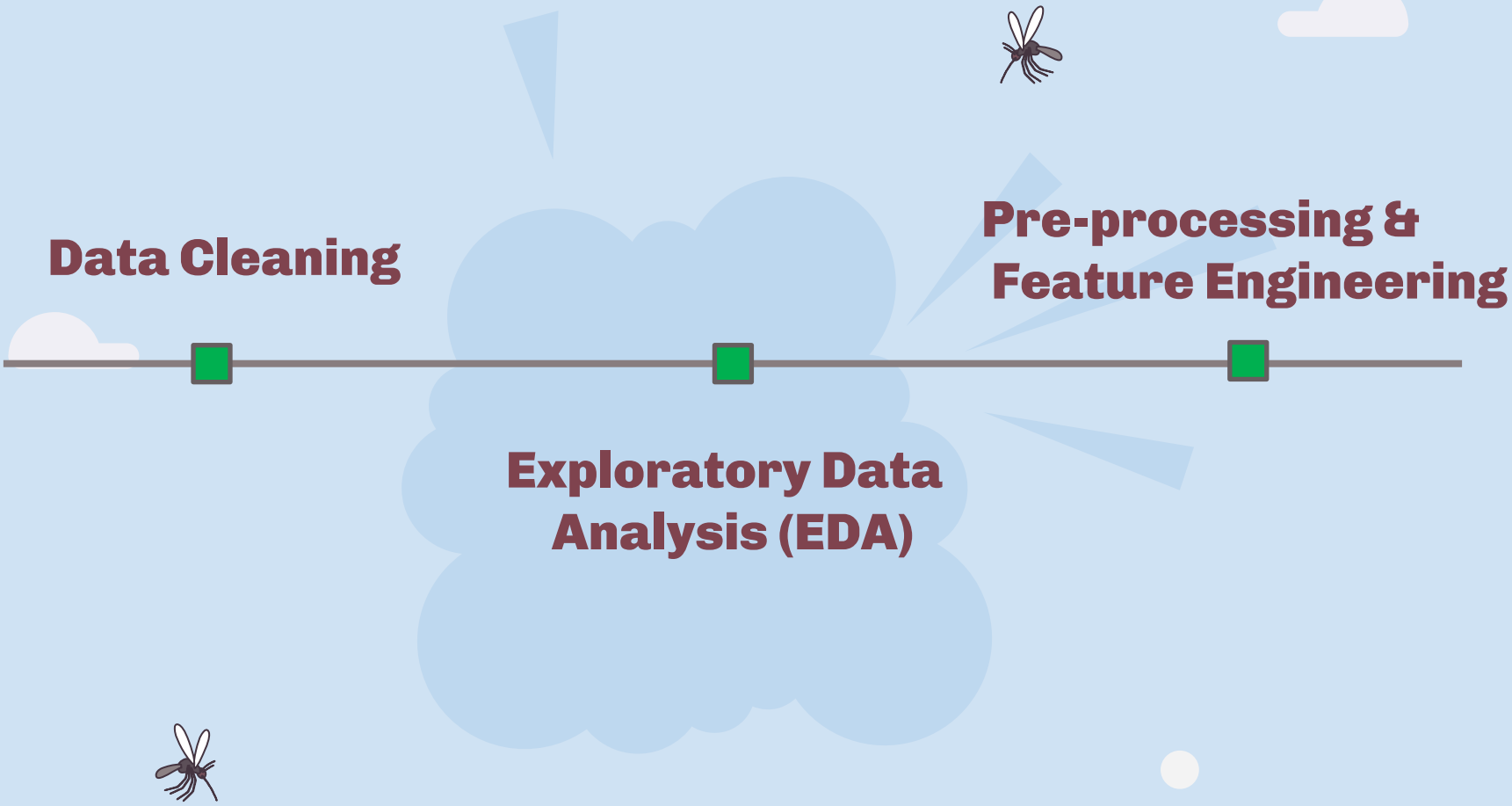
Exploratory Data Analysis (EDA)

Weather Dataset





- Highest average temperature are generally in August
- However, for 2010-2012, highest average temperature were recorded in July instead


Data Cleaning
Pre-processing &
Feature Engineering
Exploratory Data
Analysis (EDA)

Pre-processing and Feature Engineering

Merge	Weather data to train & test data
Created additional features	1. heat_cool - between heat and cool columns 2. tavg, trange - between tmax and tmin
Removed Column ***Before & after Correlation plot***	Columns from Weather dataset ['tmax']; ['tmin']; ['heat']; ['cool']; ['dewpoint']; ['wetbulb']; ['heat_cool']; ['sunrise']; ['sunset'];['avgspeed']; ['sealevel']; ['codesum'] Columns from Train and Test dataset [date]',['address'],['nummosquitos'],['block'],['street]',['trap'], [year'], ['station']
Dummify	Use pd.getdummies on 'species' column
Data Imputation	On weather data, used SimpleImputer with the mean value
Data Imbalance	Set 'class_weight' to 'balanced' for both models
Train_test_split (Train Dataset)	70 : 30

Modelling

Logistic Regression

Random Forest

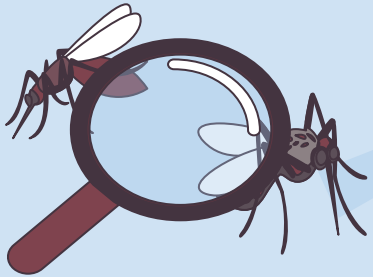
Cost Sensitive Learning for the Imbalance date

“Class weighs” set to balanced

Logistic Regression

1. **Set up parameters**
   - **C** : [0.001, 0.01, 0.1, 1, 10],
   - **penalty** : ['l1', 'l2']

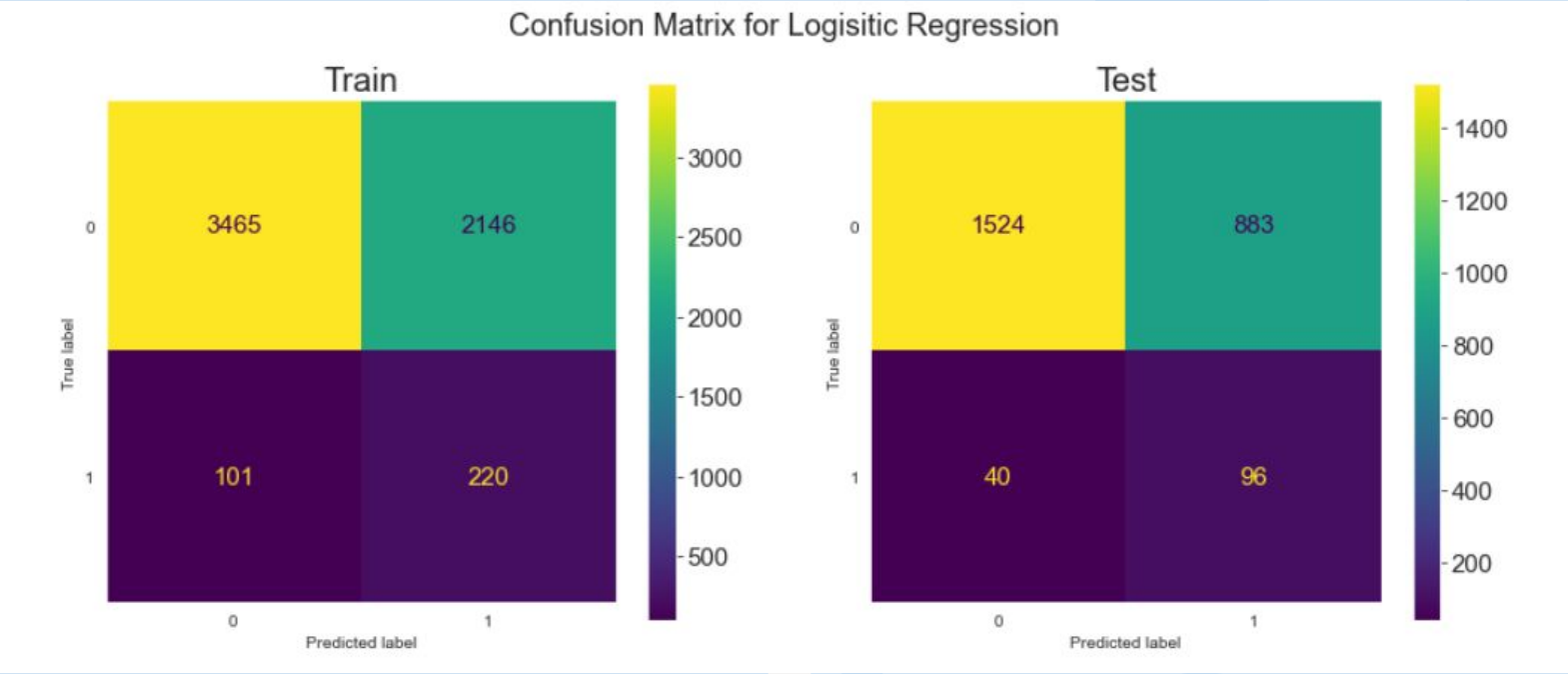
2. **GridsearchCV**
   - Parameter Pipelines (from #1)
   - **class_weight** argument to **'balanced'** - to address the imbalance data
   - **Cross-Validation** - RepeatedStratifiedKFold (n_splits=10, n_repeats=3, random_state=1)



Best score: **0.715**
Best parameters set:
C: **10**
penalty: **'l1'**

Logistic Regression

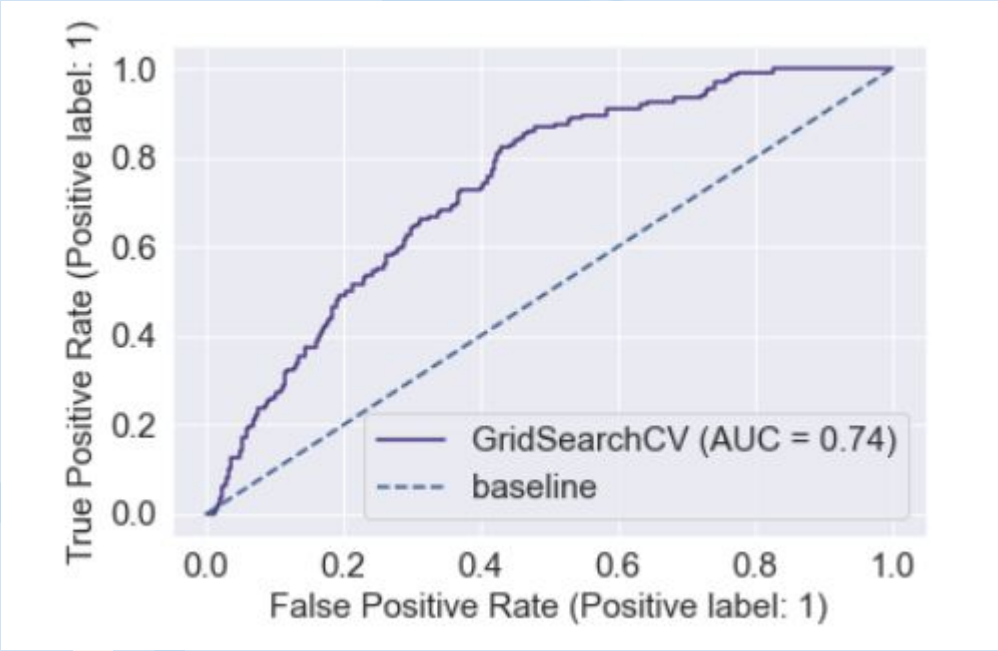
Confusion Matrix



Logistic Regression

Confusion Matrix

	Train	Test
ROC AUC	0.725	0.736
F1	0.164	0.172
Recall	0.685	0.172
Precision	0.093	0.098



Random Forest Classifier

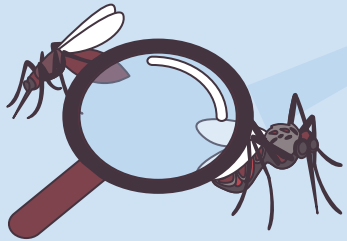
Choose Random Forest Classifier as the second model to improve the F1 score

1. Set up parameters

- **n_estimators :** [80, 100, 200, 300]
- **max_depth :** [6, 8, 10, 15, 20]
- **max_leaf_nodes :** [20, 30, 50, 70, 100, 120]

2. GridsearchCV

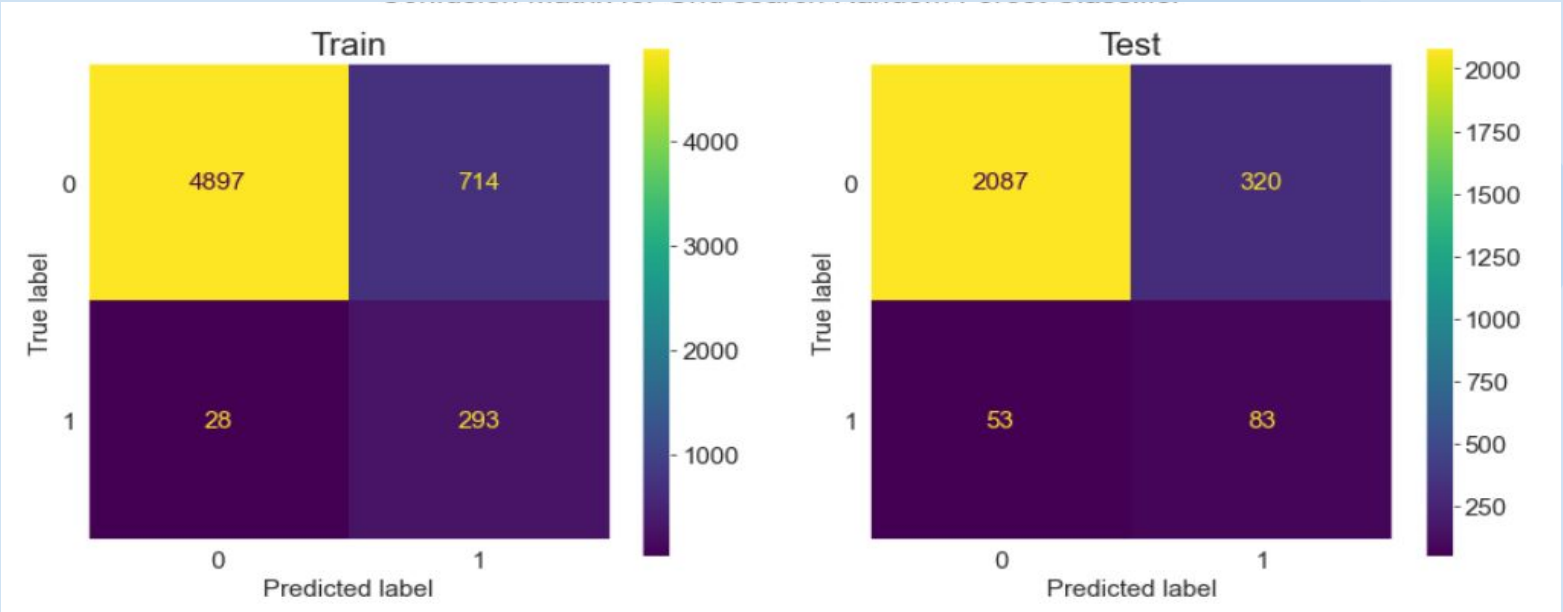
- Parameter Pipelines (from #1)
- **class_weight** argument to **'balanced'** - to address the imbalance data
- **Cross-Validation** - RepeatedStratifiedKFold (n_splits=10, n_repeats=3, random_state=1)



Best score: **0.839**
Best parameters set:
n_estimators : **80**
max_depth : **20**
max_leaf_nodes : **100**

Random Forest Classifier

Confusion Matrix



Random Forest Classifier

Scores

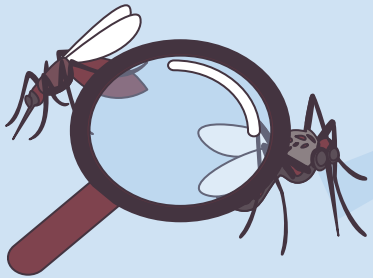
	Train	Test
ROC AUC	0.958	0.859
F1	0.441	0.308
Recall	0.913	0.610
Precision	0.291	0.206

- F1 score has been improved to ~0.44 for the train set.
- BUT, ROC AUC score for the train and test split shows sign of ***overfitting***:
  - To fix:
    - **n_estimators**
    - **max_depth**
    - **max_leaf_nodes**

Random Forest Classifier
Final

Parameters

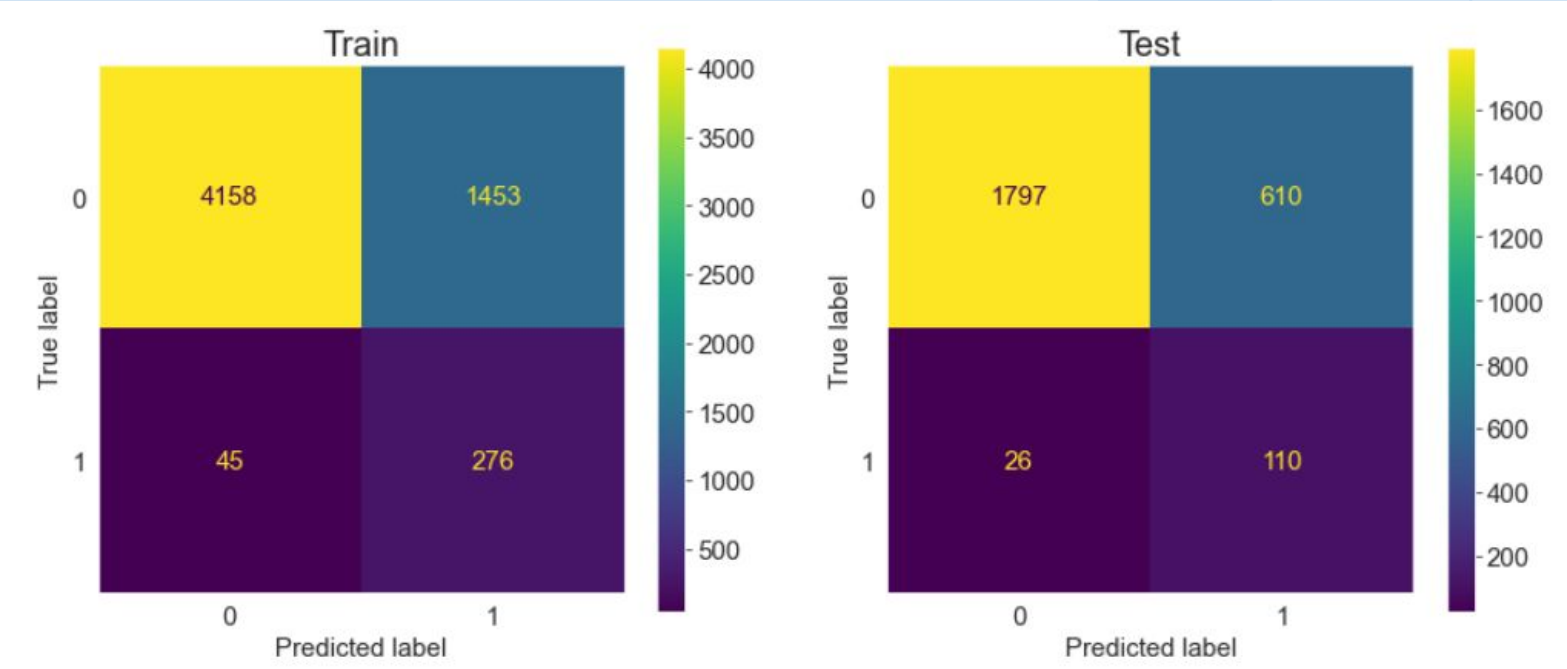
- **n_estimators :** 70 *(vs. previous: 80)*
- **max_depth :** 15 *(vs. previous: 20)*
- **max_leaf_nodes :** 20 *(vs. previous: 100)*
- **class_weight** argument to **'balanced'** - to address the imbalance data
- **Cross-Validation** - RepeatedStratifiedKFold (n_splits=10, n_repeats=3, random_state=1)



CV Mean ROC AUC score: **0.826**
CV Std ROC AUC score: **0.033**

Random Forest Classifier

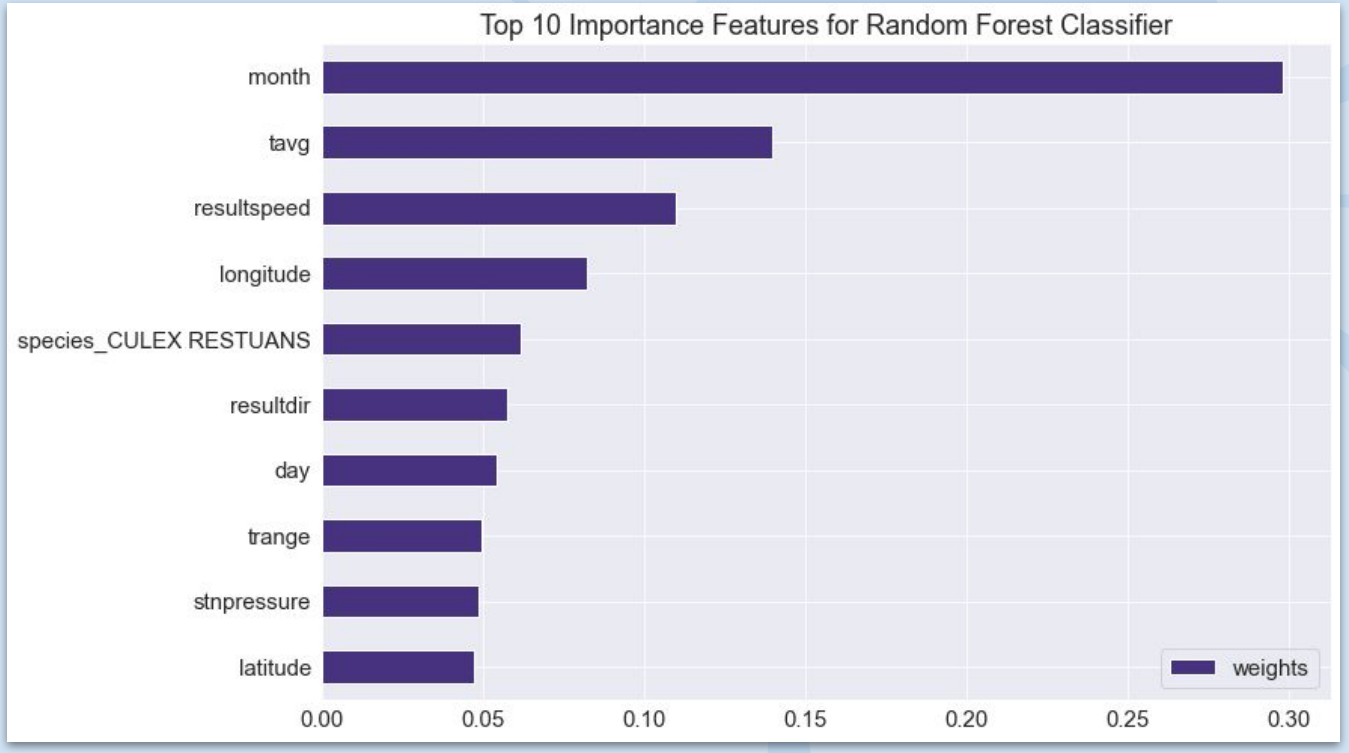
Confusion Matrix - Final



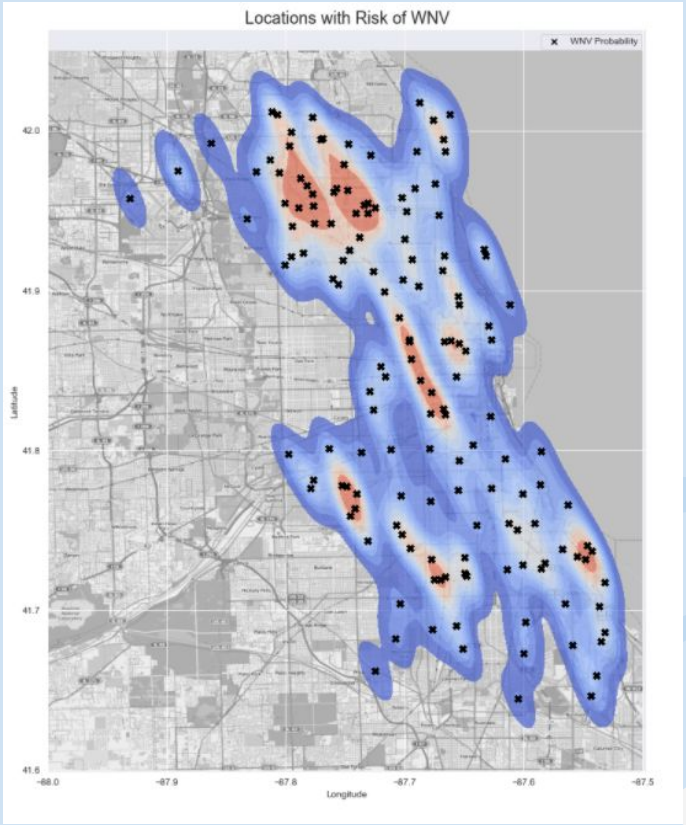
Model Results

Model	Logistic Regression		Final Random Forest	
Scores	**Train**	**Test**	**Train**	**Test**
ROC AUC	**0.725**	**0.736**	**0.876**	**0.852**
F1	**0.164**	**0.172**	**0.269**	**0.257**
Recall	0.685	0.706	0.860	0.809
Precision	0.093	0.098	0.160	0.153

Top 10 Important Features



Cost Benefit Analysis



The potential of West Nile virus present is widely spread across Chicago, with the north side of the city being the worst.

Cost of Spraying:

- Area size of Chicago: **~150k acres**
- Spray Used: **Zenivex E4** (active ingredient Etofenprox) [1]
- Cost of Zenivex E4 per acre: **USD 0.92/acre**[2]
- Cost of spraying(whole Chicago): **USD 138k**

[1] https://www.chicago.gov/content/dam/city/depts/cdph/Mosquito-Borne-Diseases/Zenivex.pdf

[2] http://www.centralmosquitocontrol.com/-/media/files/centralmosquitocontrol-na/us/resources-lit%20files/2015%20zenivex%20pricing%20brochure.pdf

Cost Benefit Analysis

Benefit of Spraying:

- **6 cases** of West Nile virus infection detected in Chicago in **2020**[1]
- Mean medical costs and productivity costs for 6 cases : **USD 197k** [2]
- Calculation:

1) Mean acute medical care costs to be avoided (for 6 cases): 6/10,000 X USD 252,115,100 = USD 151,26
2) Mean acute lost productivity to be avoided (for 6 cases): 6/10,000 X USD 22,081,260 = USD 13,249
3) Mean long-term medical care to be avoided (for 6 cases): 6/10,000 X USD 27,570,280 = USD 16,542
4) Mean long-term lost productivity to be avoided (for 6 cases): 6/10,000 X USD 26,866,800 = USD 16,120

Cost of USD 138k vs. Benefit of USD 197k

Spray OR Not Spray

[1] https://www.chicago.gov/city/en/depts/cdph/provdrs/healthy_communities/news/2020/september/first-human-cases-of-west-nile-virus-in-chicago-for-2020.html)
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3945683/

Conclusion and Recommendations

Insecticide spraying has not proven to be significant in reducing infection rates and requires more data and more campaigns for us to optimise its impact. In the meantime, we should also focus on other courses of action based on our finding

Solutions:

1. **Targets -** around Top Traps and Top WNV Addresses which are high-occurrence areas
2. **Intensify spraying cluster -** in June/July leading up to August/September
3. **Concurrent campaigns -** targeted at mosquito breeding and transmission prevention best practices

Limitations & Improvements

Our solutions are a good starting point, but other major factors should be borne in mind, including:

1. **COVID-19 -** risk prioritization needs to be adjusted in light of the pandemic
2. **Surveillance -** Applies AML to procedurally track spraying clusters vs WNV clusters in the event that there are any gaps in spray coverage.
3. **Concurrent Demographic Segmentation -** specific modelling and measures could be trained on critical population features, e.g.: age groups

We hoped to improve the model by :

1. To **collect more data** to have a more balance dataset
2. To better understand the **impact of environment to the number of mosquitoes** by having lesser missing data

Thank You

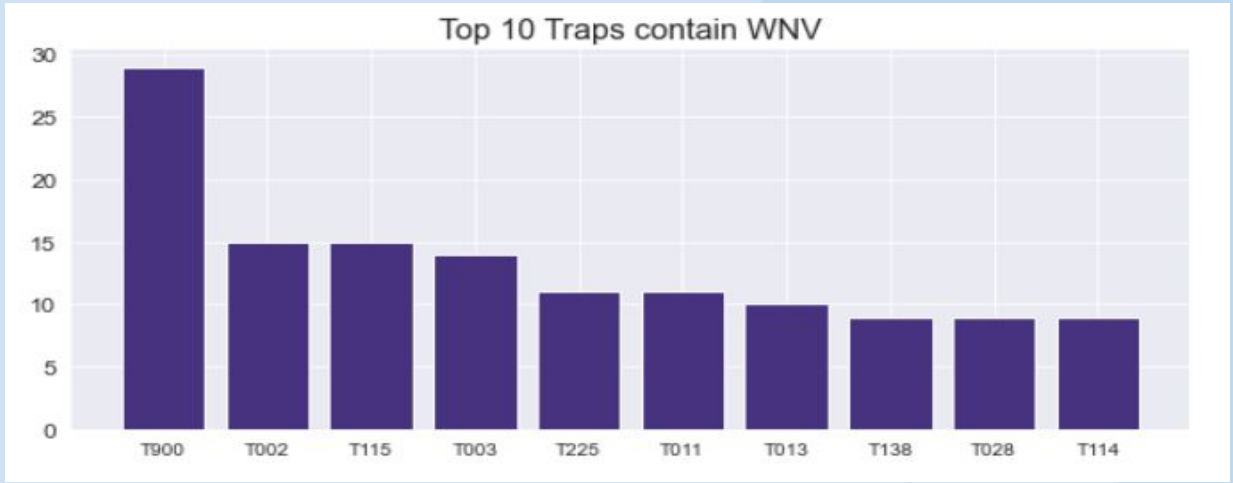
Q & A

Appendix

Supplementary Slides

Exploratory Data Analysis (EDA)

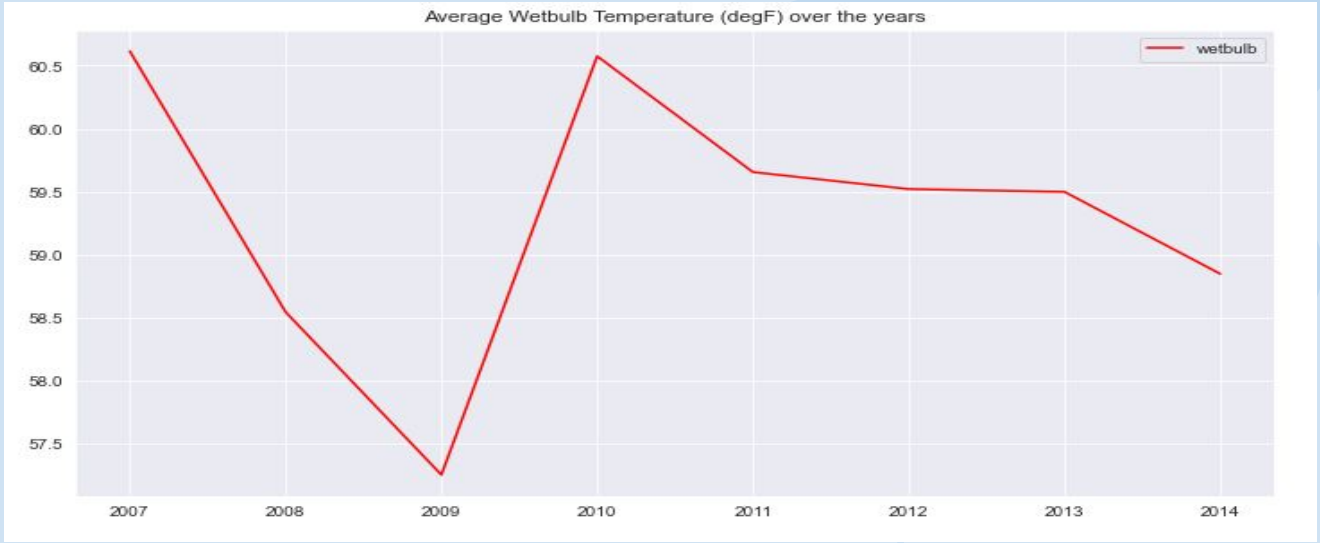
Train Dataset



- **136 traps** in the train data and trap no. **T900**(at Ohare airport) has the **most sampled** data.

Exploratory Data Analysis (EDA)

Weather Dataset



- 2009 has the lowest average wetbulb temperature
- 2007 and 2010 have the highest average wetbulb temperature