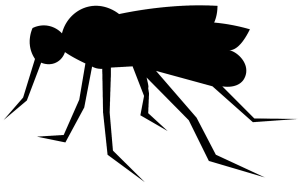
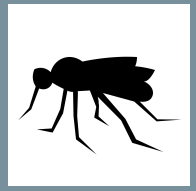
West Nile Virus Prediction

Springboard - Data Science 2022

Eric Kwon
w/ mentorship from Rahul Sagrolikar





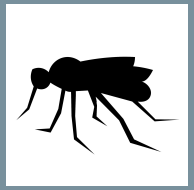
West Nile Virus in Chicago

The West Nile Virus is mosquito-transmitted and can have disastrous consequences to those infected

A **data-driven** approach is necessary to help the city combat the spread of the virus

We use a clas

- A method of
- An **extension of study spaces** for student collaboration



Prediction for Prevention

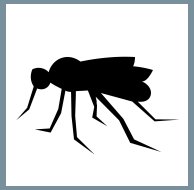
Frame as an **ML binary classification** problem:

(0) for negative
(1) for positive

Use publicly available weather data, mosquito spray data, and mosquito population data from traps to train multiple ML algorithms and assess performance by means of **AUC score**

K-Nearest Neighbor
Random Forest
Logistic Regression
Gradient Boosting
Ada Boosting

Define High Impact features with a **Shaply Analysis**



Train.csv

10506 rows x 12 columns

Data collected from mosquito traps.

- Address and location of traps
- Mosquito count
- Presence of WNV

Spray.csv

14835 rows x 4 columns

Data which records the occurrence of preventative sprays done by the city

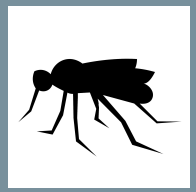
- Address and location of spray
- Date and time of spray

Weather.csv

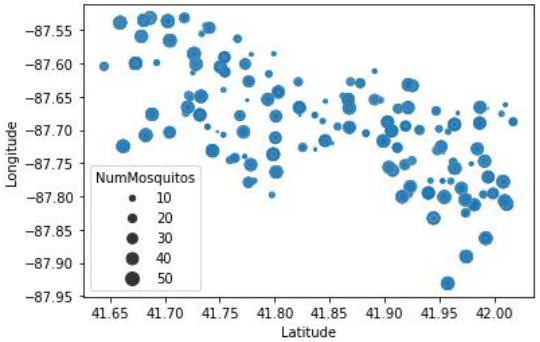
2944 rows x 22 columns

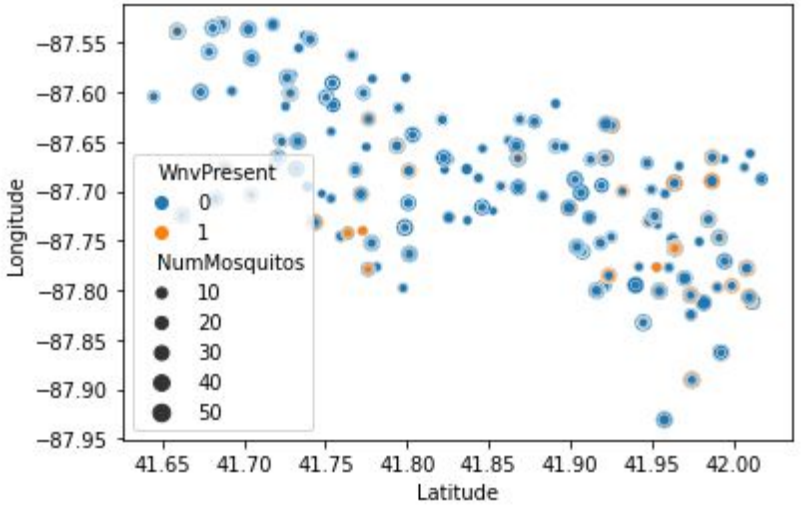
Data which contains the weather data for a given day. There are two stations recorded. We keep assess the data from station 1 only

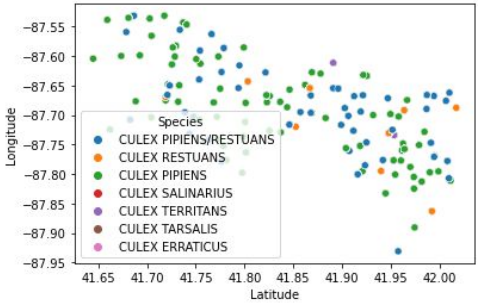
- Date
- Weather Codes
- Temperatures
- Wind Speeds
- Dew Point
- Wet Bulb
- etc.

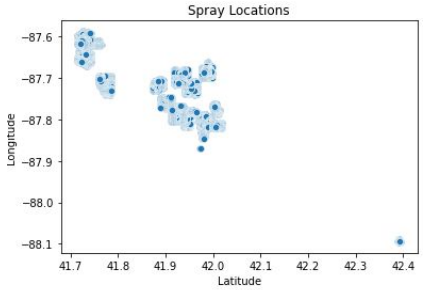


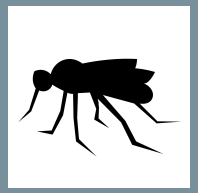
Seaborn Plots to visualize relationships











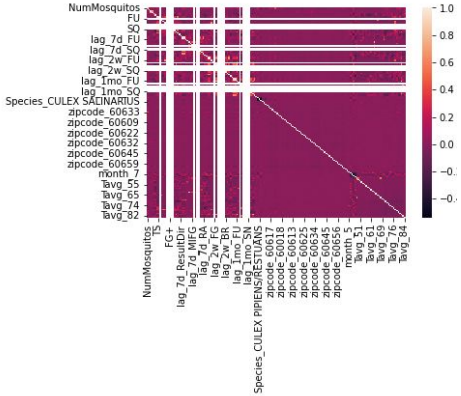
Train.csv | Spray.csv | Weather.csv

Data Cleaning and Feature Engineering:

- Search for missing or null values and drop or impute per engineering judgement
  - *-convert dates to datetime object and extract the month as a new categorical feature*
  - *-station 1 weather values only,*
  - *-drop columns: ['Depth', 'Water1', 'SnowFall', 'SeaLevel']*
  - *-impute Tavg by averaging Tmax & Tmin*
  - *-drop any row containing 'M' or 'T' which is defined as missing in PDF weather guide*
- Aggregate location information by categorizing latitude/longitude to zipcodes
  - *-Geopy API*
- Categorize weather code values as categorical columns and normalize times
  - *-apply and lambda functions with sets*
- Categorize the occurrence of sprays by tallying up the number of sprays at a specific zipcode within a 2-week or 1-month window
  - *-apply and lambda functions*
- Create new features based on lagged weather data per 7-days, 2-weeks, 1-month
  - *-pandas merge method for SQL-like join statements*
- One-hot encode categorical features ['Species', 'zipcode', 'month', 'Tavg']

Feature Selection

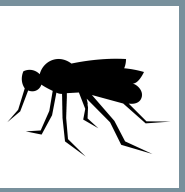
Correlation Matrix

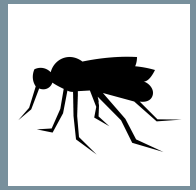


Keep features above a 0.01 correlation threshold (arbitrary)

<u>Eliminate redundant features:</u>

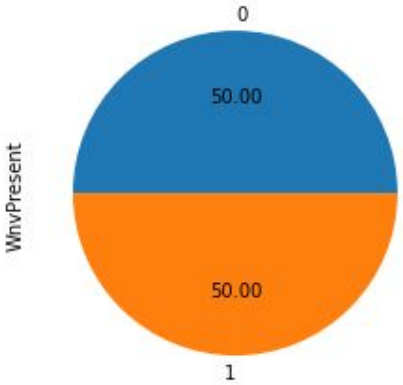
Check for **multi-collinearity** using statsmodels variance_inflation_factor "VIF" and prune features so as to limit VIF < 10.0





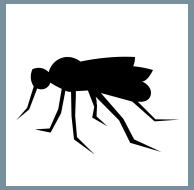
Prepare the Data

Random Under Sampler to focus analysis on presence of WNV which only affects a minority of the dataset



Test-train-split with a test size of 25%

Random Under Sampler to focus analysis on presence of WNV which only affects a minority of the dataset



Using pipeline for grid search or random CV search

KNN

Results:

```
From Confusion Matrix:
True Negatives: 1515
False Positives: 898
False Negatives: 20
True Positives: 109
```

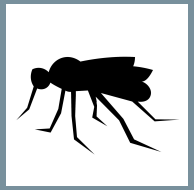
	No WNV	WNV Present	accuracy	macro avg	weighted avg
precision	0.986971	0.108242	0.638867	0.547606	0.942377
recall	0.627849	0.844961	0.638867	0.736405	0.638867
f1-score	0.767477	0.191901	0.638867	0.479689	0.738268
support	2413.000000	129.000000	0.638867	2542.000000	2542.000000

Random Forest

Results:

```
Fitting 10 folds for each of 10 candidates, totalling 100 fits
From Confusion Matrix:
True Negatives: 1782
False Positives: 631
False Negatives: 26
True Positives: 103
```

	No WNV	WNV Present	accuracy	macro avg	weighted avg
precision	0.985619	0.140327	0.741542	0.562973	0.942723
recall	0.738500	0.798450	0.741542	0.768475	0.741542
f1-score	0.844350	0.238702	0.741542	0.541526	0.813615
support	2413.000000	129.000000	0.741542	2542.000000	2542.000000



Logistic Regression

From Confusion Matrix:
True Negatives: 1828
False Positives: 585
False Negatives: 34
True Positives: 95

	No WNV	WNV Present	accuracy	macro avg	weighted avg
precision	0.981740	0.139706	0.756491	0.560723	0.939009
recall	0.757563	0.736434	0.756491	0.746999	0.756491
f1-score	0.855205	0.234858	0.756491	0.545031	0.823724
support	2413.000000	129.000000	0.756491	2542.000000	2542.000000

Gradient Boosting

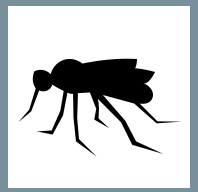
From Confusion Matrix:
True Negatives: 1805
False Positives: 608
False Negatives: 27
True Positives: 102

	No WNV	WNV Present	accuracy	macro avg	weighted avg
precision	0.985262	0.143662	0.750197	0.564462	0.942553
recall	0.748031	0.790698	0.750197	0.769365	0.750197
f1-score	0.850412	0.243147	0.750197	0.546779	0.819595
support	2413.000000	129.000000	0.750197	2542.000000	2542.000000

Ada Boosting

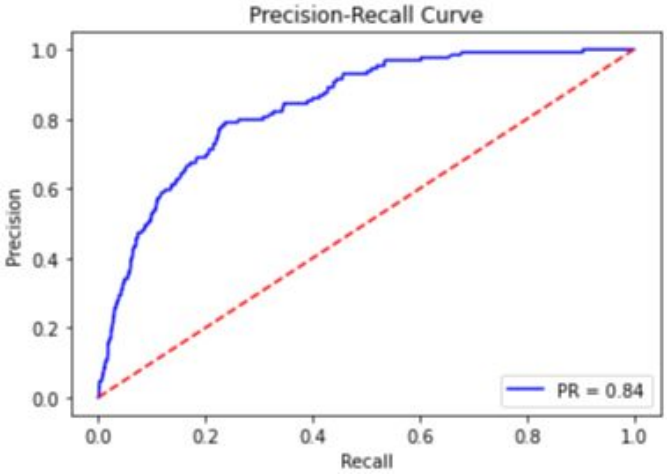
Fitting 5 folds for each of 10 candidates, totalling 50 fits
From Confusion Matrix:
True Negatives: 1783
False Positives: 630
False Negatives: 28
True Positives: 101

	No WNV	WNV Present	accuracy	macro avg	weighted avg
precision	0.984539	0.138167	0.741149	0.561353	0.941588
recall	0.738914	0.782946	0.741149	0.760930	0.741149
f1-score	0.844223	0.234884	0.741149	0.539554	0.813301
support	2413.000000	129.000000	0.741149	2542.000000	2542.000000

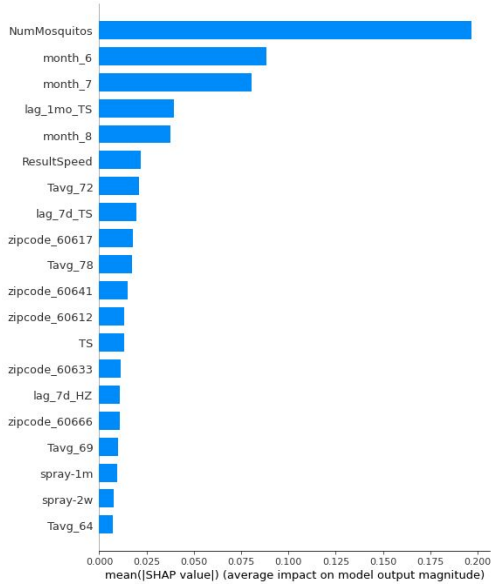


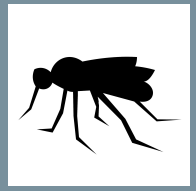
Gradient Boosting is determined to be the most effective model using f-1 score on WNV-present as a gauge

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$



Shap Values





We see from our highest performing classifier model, features related to the population of mosquitos are the greatest determining factors in the prediction of the presence of the WNV.

We observer lower than expected impact from the spray-2w and spray-1mo features. After plotting the spray data with trap locations, we see that the spray locations are highly concentrated and do not cover as large an area as the trap locations.

From the classifier model and the resultant Shaply analysis, it is hence reasonable to assume that distributing mosquito spray locations across a wider area should result in greatly reducing the presence of the West Nile Virus. This is therefore the main recommendation for the City of Chicago