DSI24 Project 4: West Nile Virus

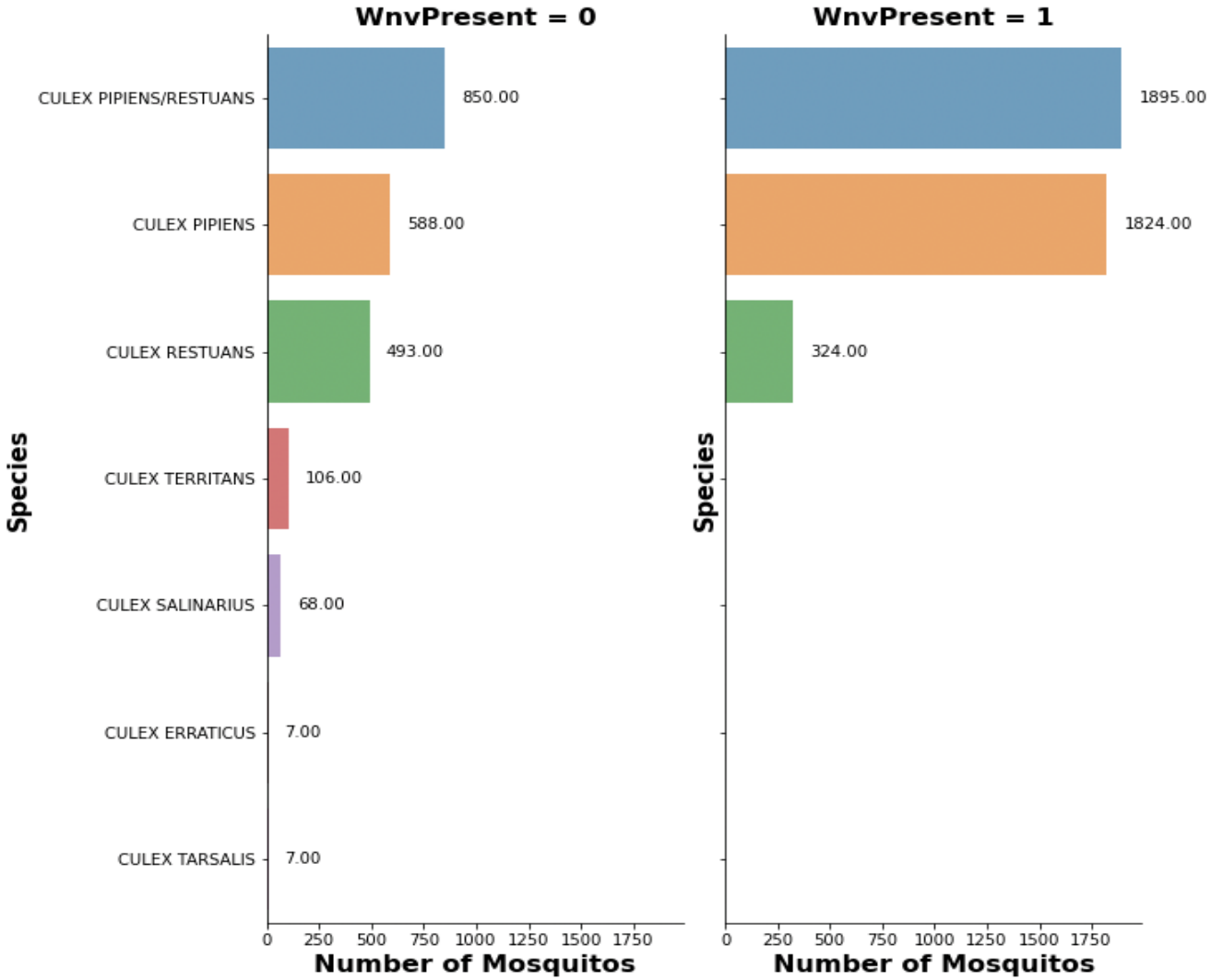
By Team CJE

Problem Statement

We are a team working for the Mayor of Chicago who is interested in predicting the area of Chicago that has high presence of WNV in order to take the appropriate mitigation measures

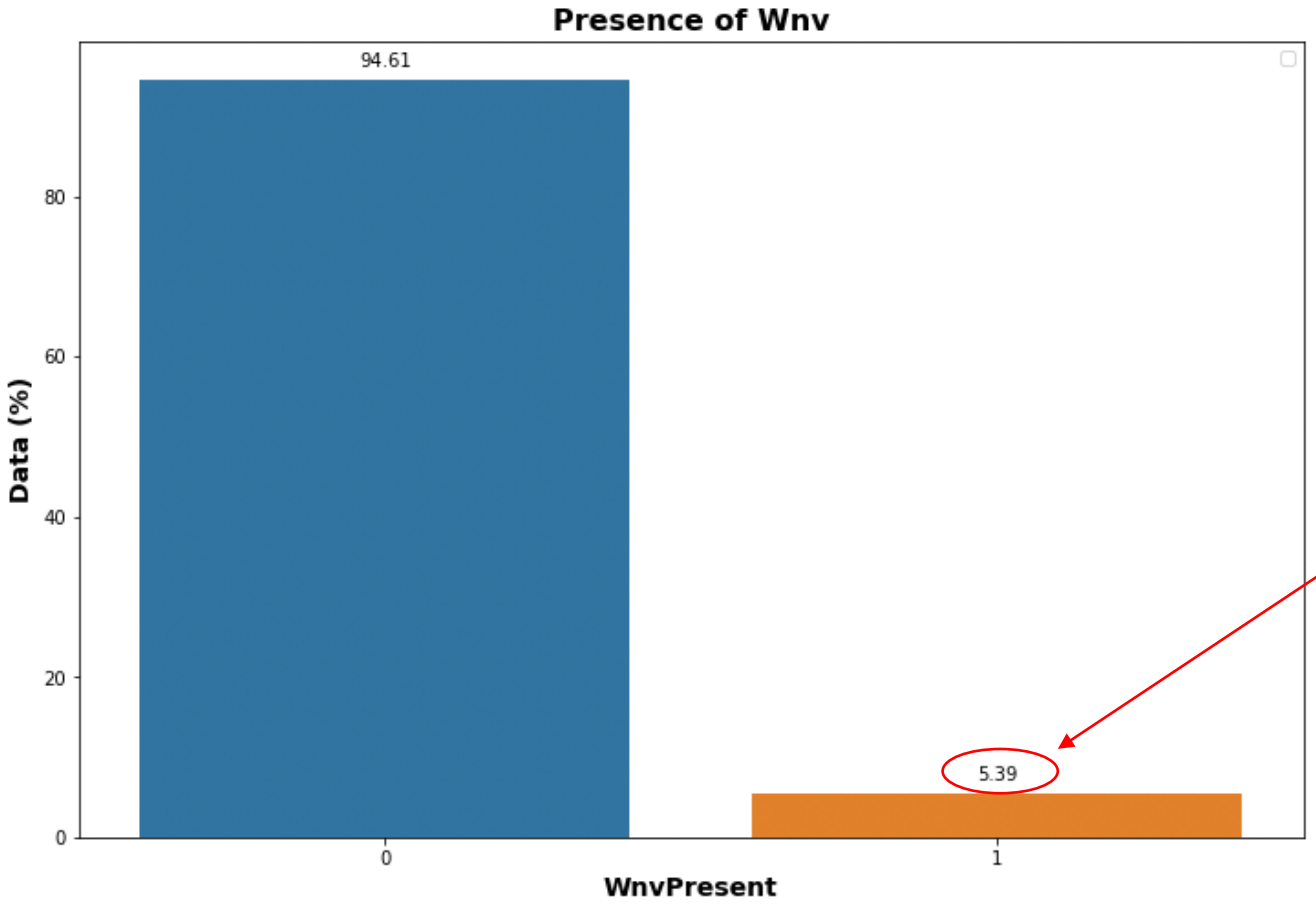
- Predict the presence of WNV given a location, species with the weather condition
- Determine if spraying help to mitigate the epidemic in a cost-effective way

EDA – Train dataset



- 3 species contribute to the WNV
- Found out that there is another type of mosquito 'UNSPECIFIED CULEX' in the Test data
- Might have some negative implication on training our model if 'UNSPECIFIED CULEX' is a positive carrier of WNV

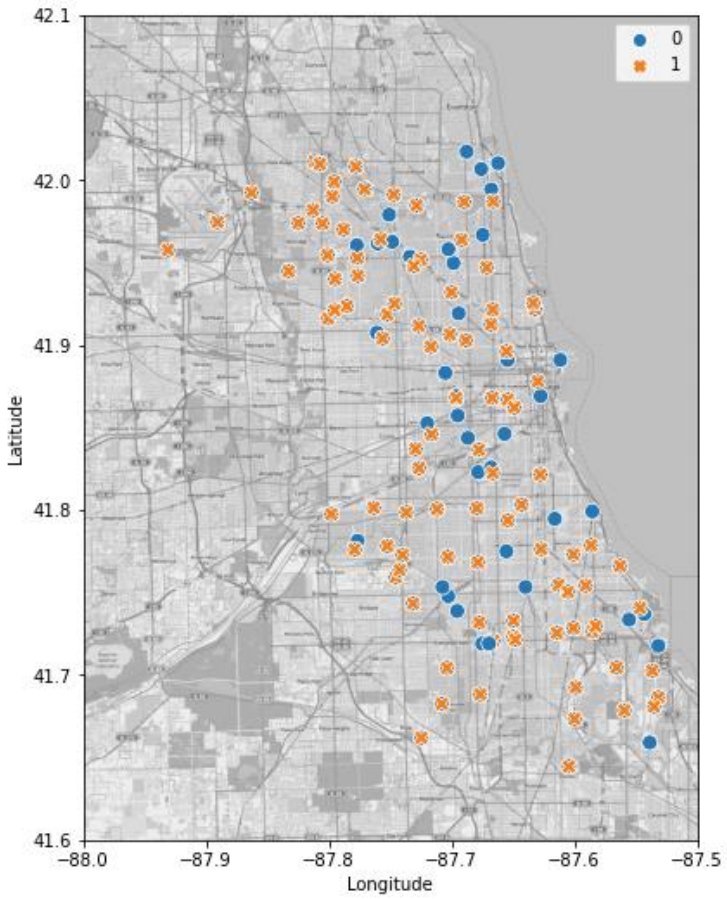
EDA – Train dataset Data Imbalance



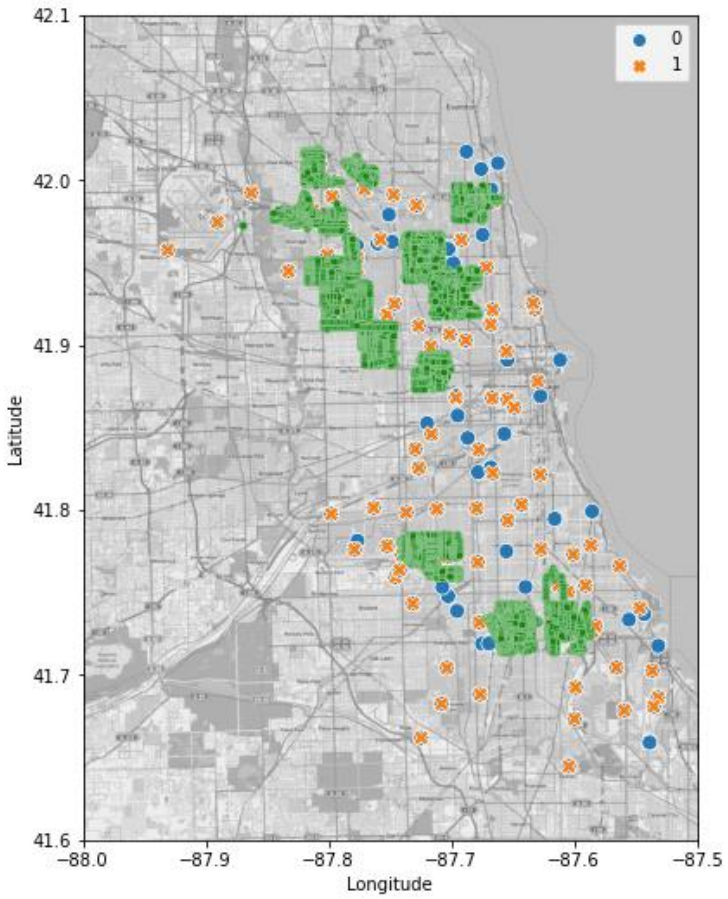
- Imbalance train dataset with >90% of the data having WNV-negative values
- Oversample the dataset using SMOTE-NC
- Mixture of continuous and categorical features

EDA – Spray vs WnvPresent

Location of WNV

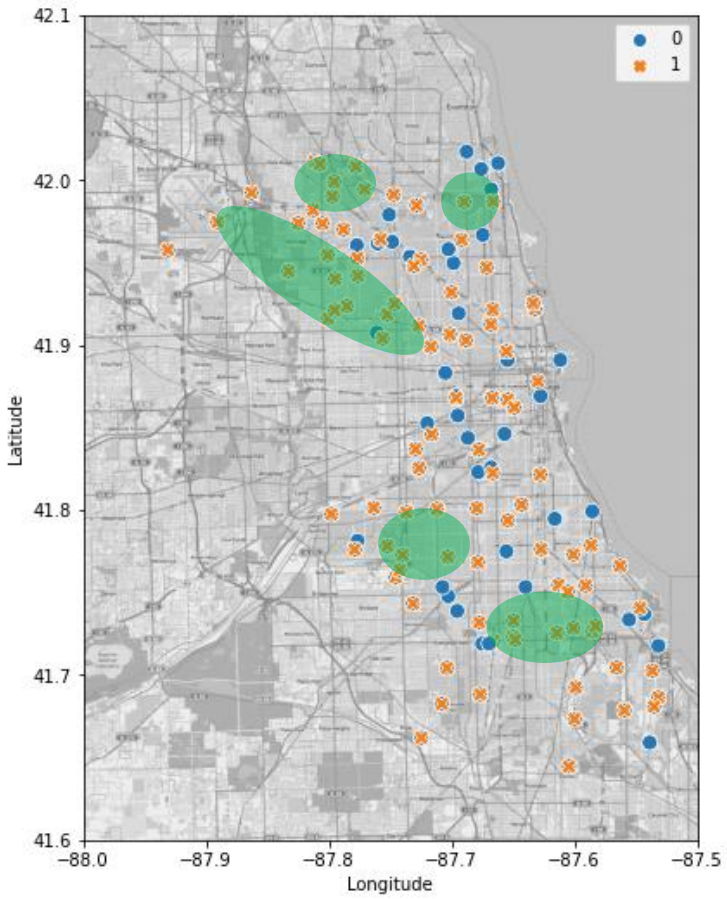


Location of Spraying

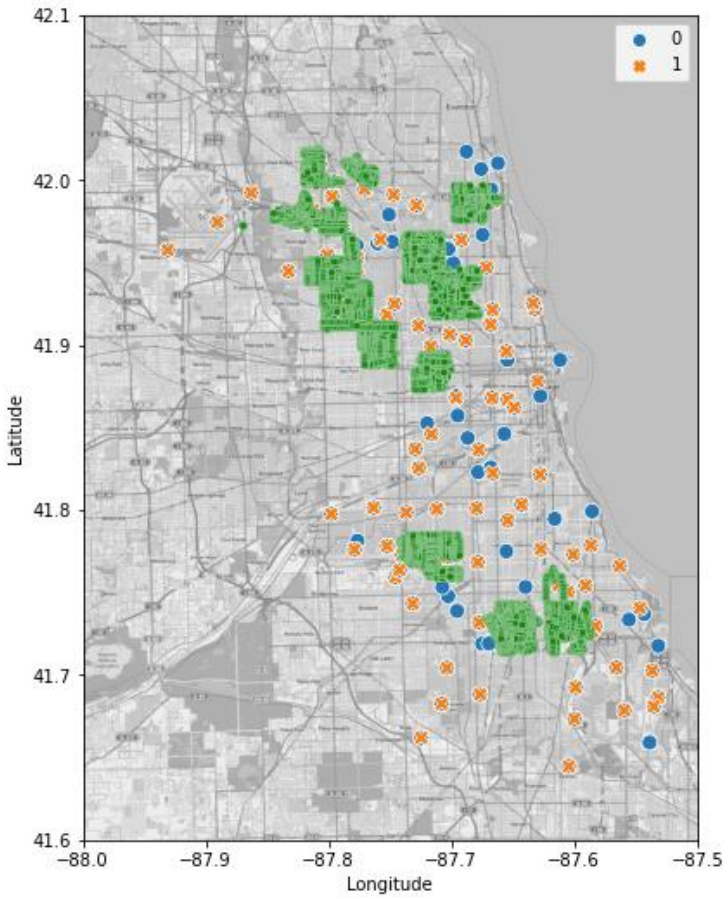


EDA – Spray vs WnvPresent

Location of WNV

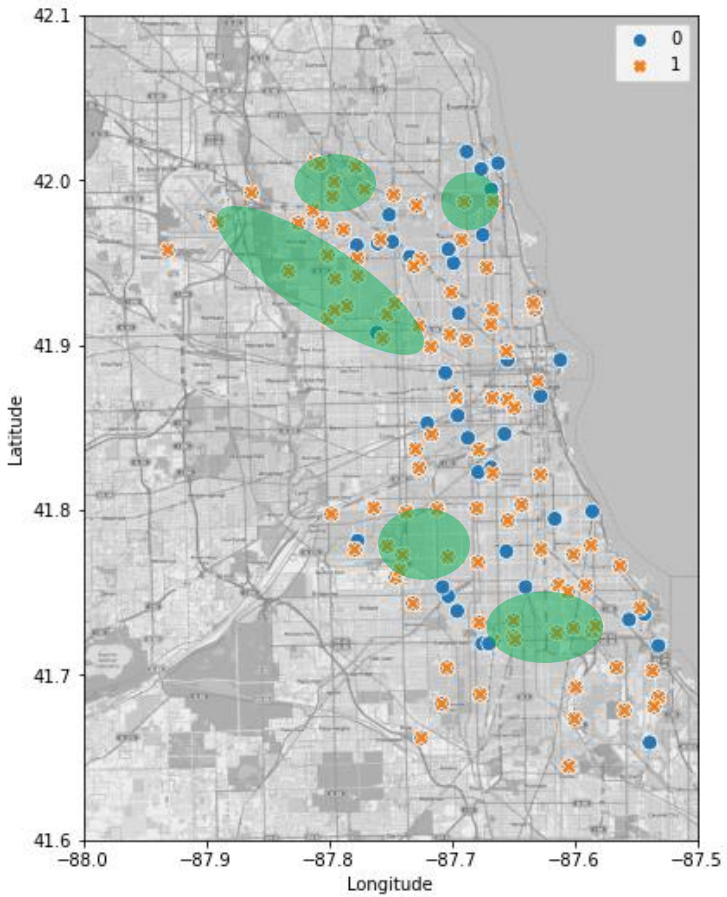


Location of Spraying

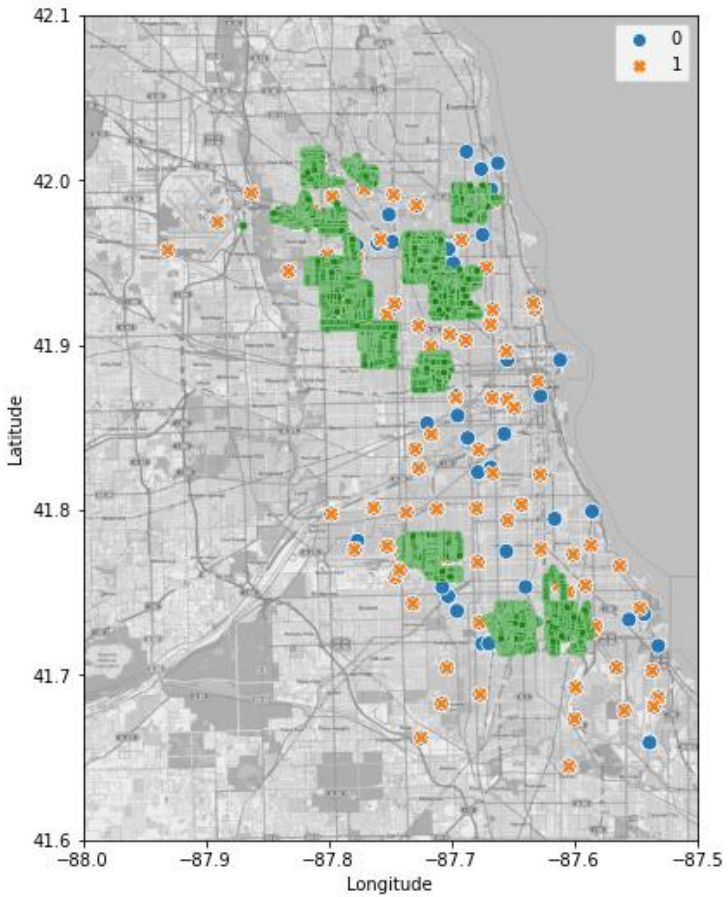


EDA – Spray vs WnvPresent

Location of Wnv



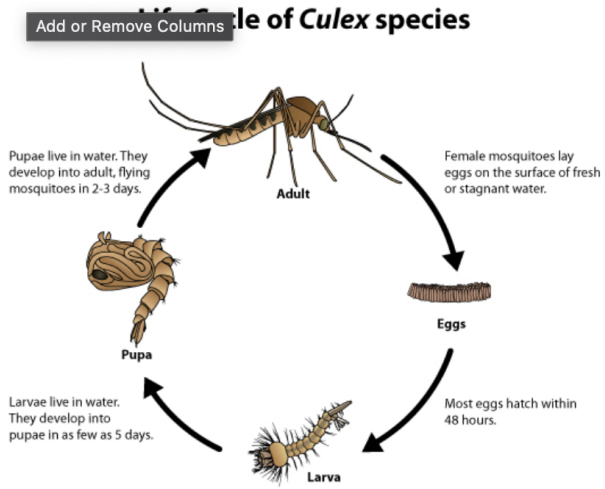
Location of Spraying



SPRAYING DOES NOT HELP!!!!

Modeling Assumptions

- Species: 3 out of 7 are historically WNV-positive
- Trap ID: Not all traps are equal (e.g. T115)
- Longitude and latitude: Areas towards the east seem to be less prone to WNV
- Weather-related features: Temperature, Dew Point, Wet Bulb, Total Precipitation
  - Our research shows that mosquitoes have a life cycle of 7 days
  - Any effect on mosquito-breeding due to changes in weather might be seen only a few days later
  - Weather variables lagged by up to 7 days
- Month and week of the year: August seems to be most prone to WNV, followed by September



Results

BASELINE ACCURACY: 0.95

BASELINE ROC_AUC: 0.50

Step 1: Test six baseline models. Pick the best two models for intensive hyperparameter tuning.

Step 2: Intensive grid search of hyperparameters seems to work slighter better for Logistic model but worse for XGBoost

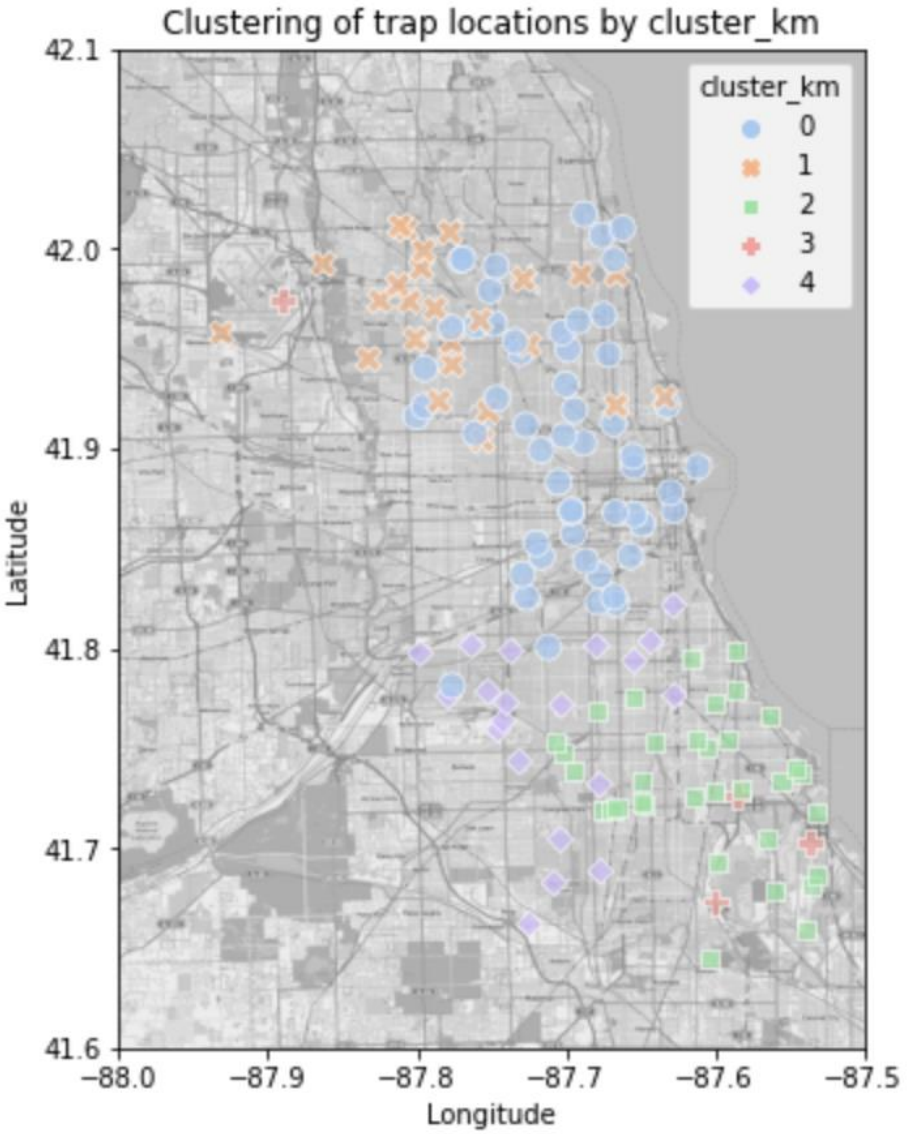
Step 3: Drop `month` which has 95% correlation to `week`. But results worsen.

Steps 4 and 5: Play around number of lagged days for weather variables. Jump in Kaggle scores. Lagged up to 3 days work best.

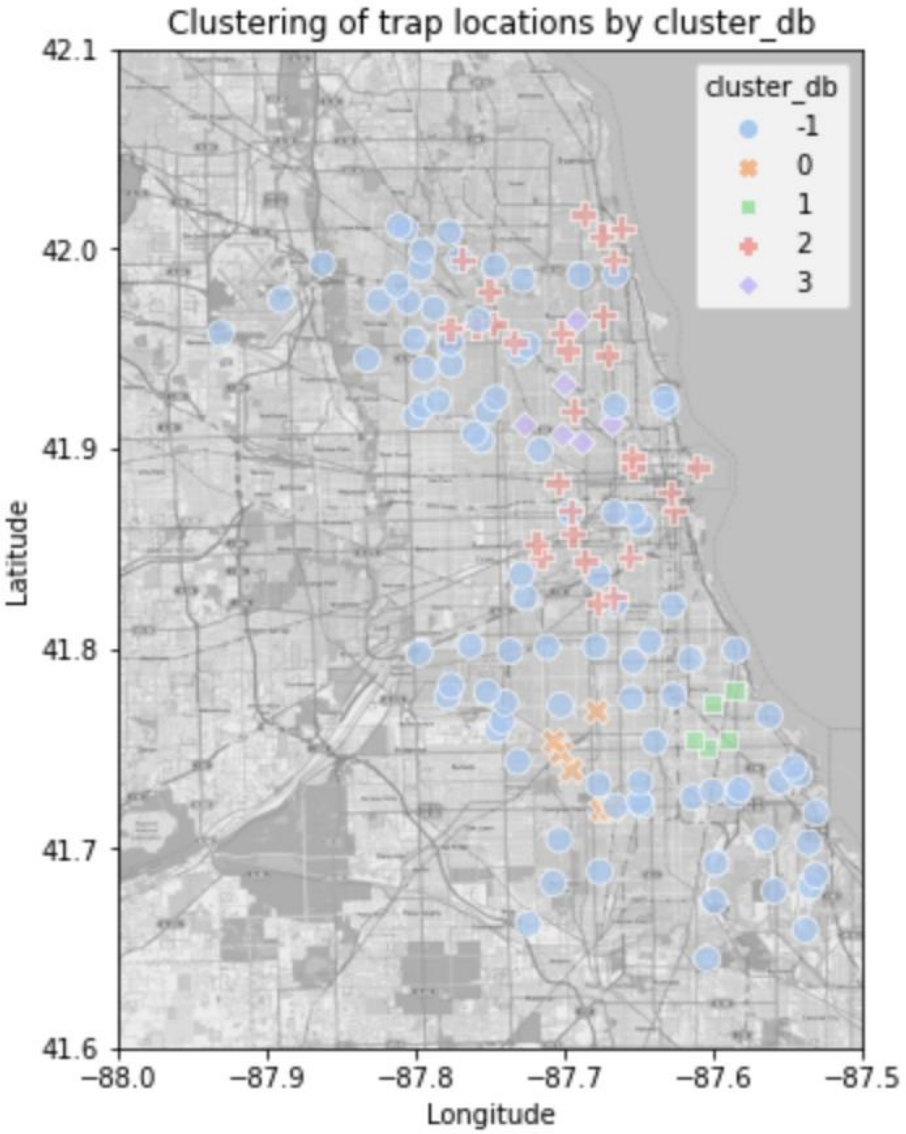
Steps 6 and 7: Drop GPS coordinates, create 'mosquito-infection clusters'. DBSCAN works better than KMeans.

Model No.	Classifier	CV Score (train)	ROC_AUC (train)	ROC_AUC (test)	Kaggle Score	Runtime (sec)	Remarks
1	LogisticRegression(random_state=42, solver='li...	0.989	0.994	0.840	0.659	12	NaN
2	KNeighborsClassifier()	0.916	1.000	0.794	0.593	72	NaN
3	RandomForestClassifier(random_state=42)	0.981	0.999	0.777	0.689	37	NaN
4	ExtraTreesClassifier(random_state=42)	0.957	0.973	0.831	0.708	59	NaN
5	SVC(max_iter=10000, random_state=42)	0.974	0.981	0.829	0.606	195	NaN
6	XGBClassifier(base_score=None, booster=None, c...	0.992	1.000	0.822	0.712	105	NaN
7	LogisticRegression(random_state=42)	0.989	0.994	0.841	0.662	287	Cousin of Model 1, a lot of tuning
8	XGBClassifier(base_score=None, booster=None, c...	0.991	1.000	0.817	0.7	106	Cousin of Model 6, a lot of tuning
9	LogisticRegression(random_state=42, solver='li...	0.989	0.994	0.833	0.653	40	Drop month, weather lagged 7 days
10	XGBClassifier(base_score=None, booster=None, c...	0.992	1.000	0.815	0.673	135	Drop month, weather lagged 7 days
11	LogisticRegression(random_state=42, solver='li...	0.988	0.993	0.837	0.717	19	Drop month, weather lagged 3 days
12	XGBClassifier(base_score=None, booster=None, c...	0.992	1.000	0.812	0.713	127	Drop month, weather lagged 3 days
13	LogisticRegression(random_state=42, solver='li...	0.987	0.992	0.829	0.732	10	Drop month, weather lagged 1 day
14	XGBClassifier(base_score=None, booster=None, c...	0.991	1.000	0.804	0.676	80	Drop month, weather lagged 1 day
15	LogisticRegression(random_state=42, solver='li...	0.989	0.993	0.845	0.705	24	Cousin of Model 11, with KMeans mozzie clusters
16	XGBClassifier(base_score=None, booster=None, c...	0.990	0.997	0.817	0.695	100	Cousin of Model 12, with KMeans mozzie clusters
17	LogisticRegression(random_state=42, solver='li...	0.990	0.994	0.846	0.717	23	Cousin of Model 11, with DBSCAN mozzie clusters
18	XGBClassifier(base_score=None, booster=None, c...	0.990	0.998	0.828	0.722	115	Cousin of Model 12, with DBSCAN mozzie clusters

KMeans Clustering



DBSCAN Clustering



Model Selection: Not just ROC_AUC

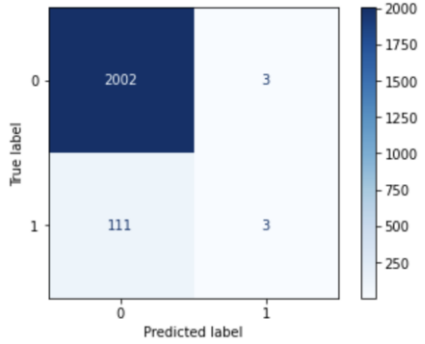
Model	Accuracy	ROC_AUC (test)
Naïve Model	0.95	0.50
Model 17 (Logistic)	0.95	0.85
Model 18 (XGBoost)	0.92	0.83

- On paper, one should pick Model 17 over Model 18, however...
- Model 17 does not seem to be useful for city officials when deciding where to focus their spraying efforts
  - 6 out of 2119 cases are predicted to be positive (99.9% negative predictions)
- Model 18 is braver, choosing 79+29 = 108 predictions to be positive (94.9% negative predictions)
- Model 18's sensitivity of 0.254 is more than 800% higher than Model 17's sensitivity of 0.026!

Selected model: XGBoost, weather variables lagged 1-3 days, DBSCAN clusters

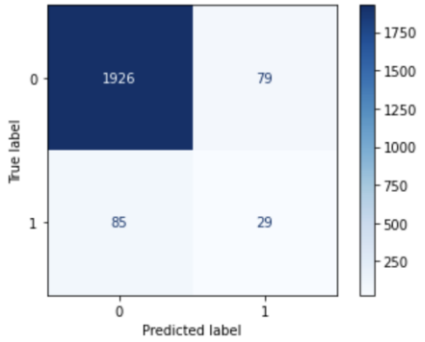
Model 17: Logistic, weather lagged 3 days, DBSCAN clusters

```
Accuracy: 0.9462010382255781
 Sensitivity: 0.026
 Specificity: 0.999
 Precision: 0.500
 Tesst ROC AUC: 0.846
```

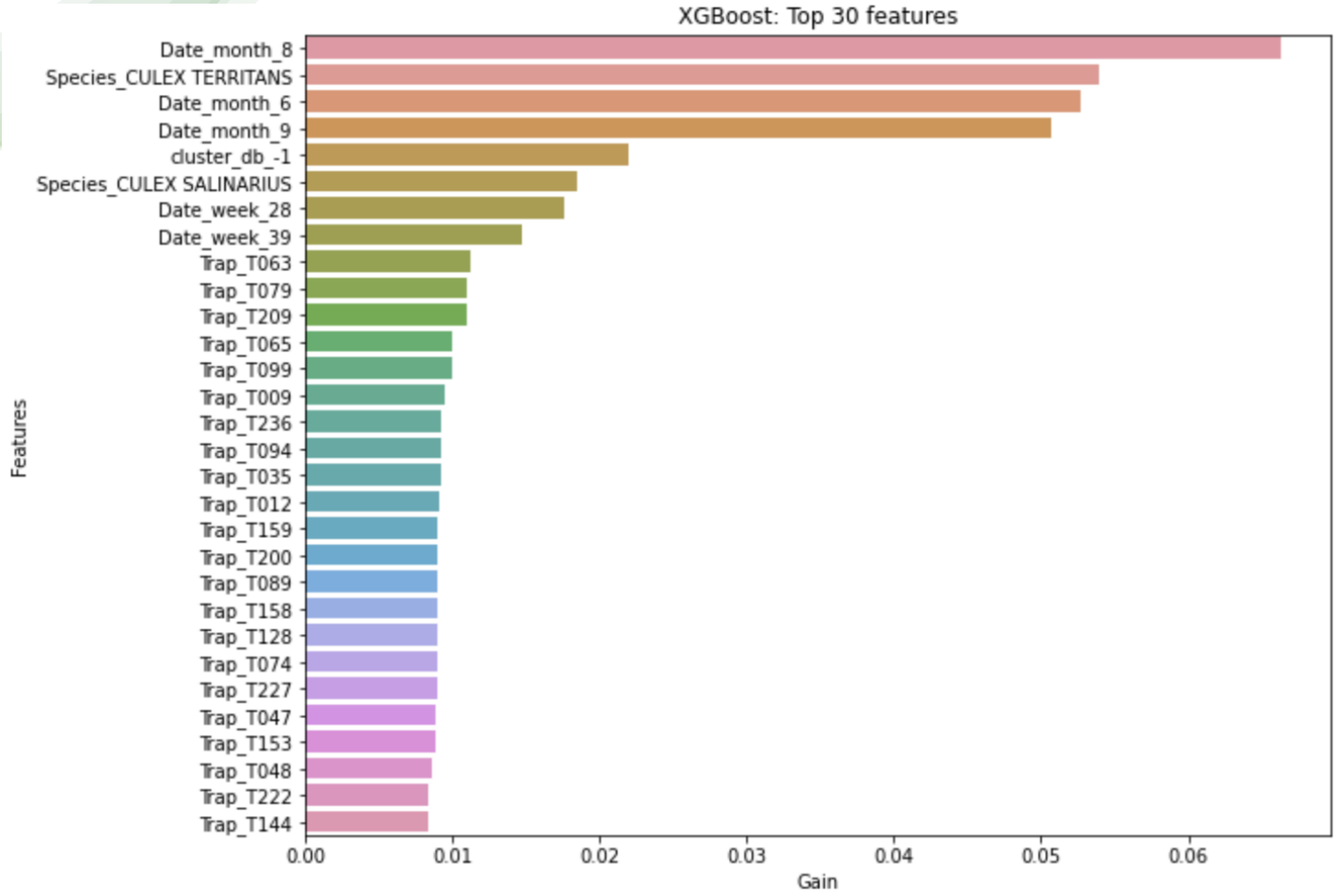


Model 18: XGBoost, weather lagged 3 days, DBSCAN clusters

```
Accuracy: 0.9226050023596036
 Sensitivity: 0.254
 Specificity: 0.961
 Precision: 0.269
 Test ROC AUC: 0.832
```



Features of Importance

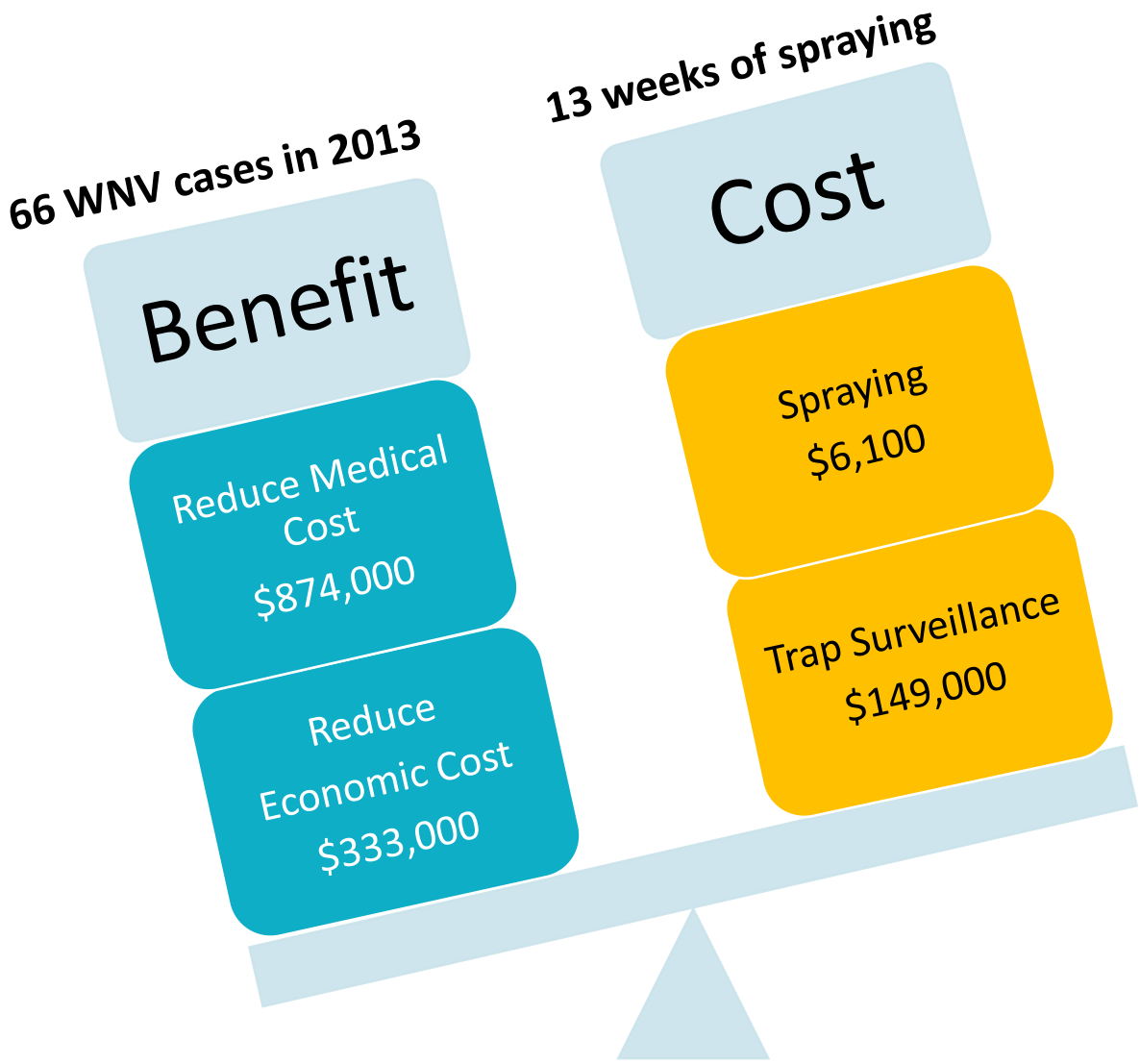


IN ORDER OF IMPORTANCE

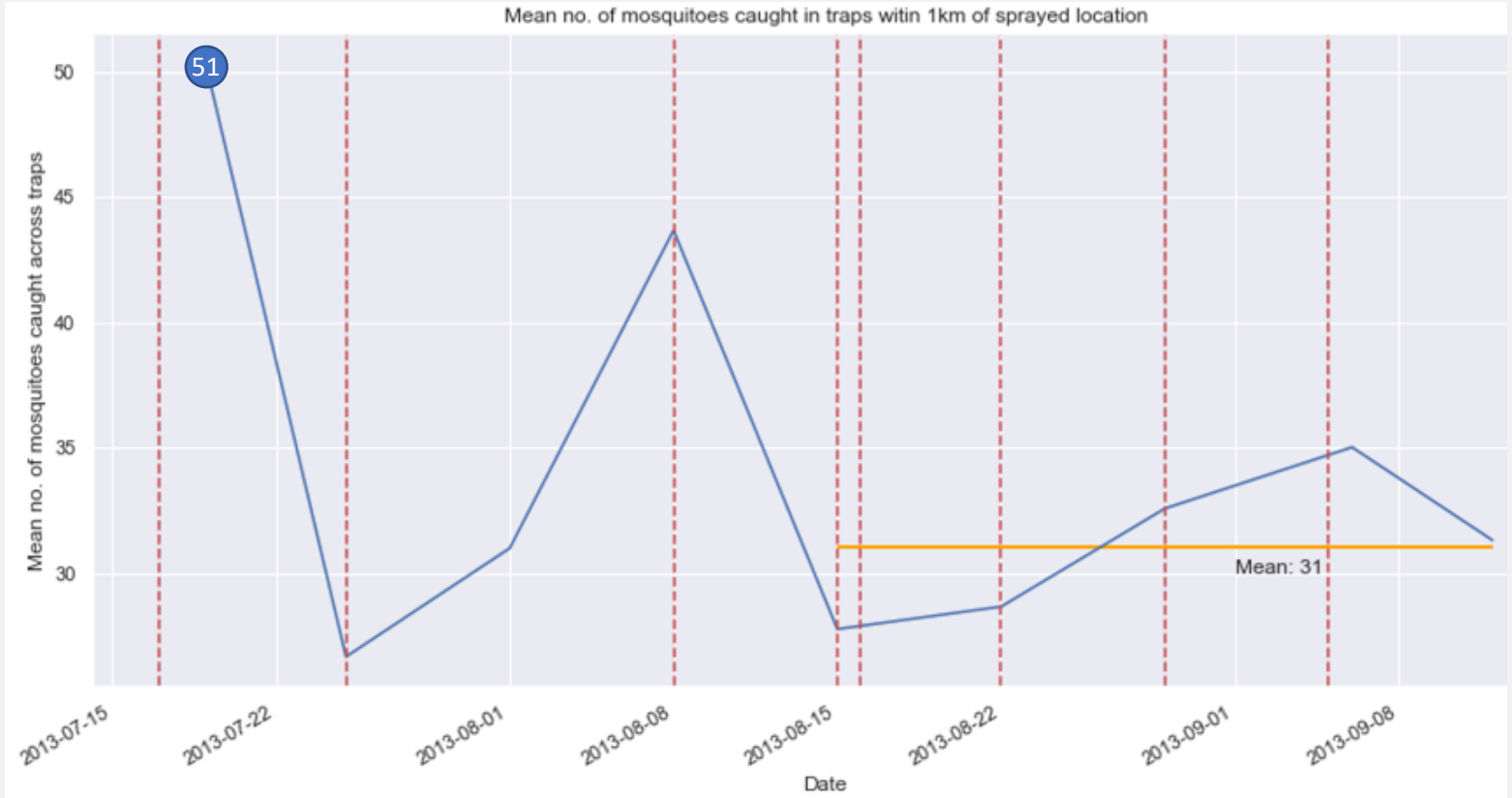
- Month: Most important predictor, August responsible for improving AUC_ROC score by 6.5%
- Species: Next most important, with Culex Territans responsible for 5.5% of improvement
- Mozzie cluster: DBSCAN cluster ‘-1’ is next on the list, 2.5% of improvement. The locations are in the Northwest and South of Chicago.
- Week: #28 and #39
- Trap: Generally responsible for <1% of the score improvement
- Weather: Interestingly, not in the Top 30 chart

Cost Benefit Analysis

- Medical cost is the key reason why we should reduce WNV cases, especially for severe cases
- Reduction of 1 case saves $18,000
- Current vector control program (VCP) needs to reduce cases by at least 58 for the benefit of the vector control program to outweigh the cost of WNV cases
- Are the current VCP capable of reducing the case load by 58? We will use mosquito numbers to help us find out!



Effects of Spraying



Legend

Date of Spraying

Mid-July

51 mosquitoes

Highest number caught between mid-July to mid-September

Mid-August onwards

31 mosquitoes

Mosquitoes number did not fluctuate that much and came down to an average of 31 mosquitoes

----> 40% decrease

Using mosquito numbers as predictor of WNV cases

66 → 40 cases

Decrease of 26 WNV cases is not sufficient to justify cost of spraying (58 required)

Recommendation & Conclusion

- The Chicago Department of Public Health should focus on all areas where the traps in the DBSCAN Cluster “-1” and traps T063, T079 and T209 are located as they are strong predictors for WNV cases

- Spraying seems to have short-term effect of reducing mosquito numbers but it does not eradicate enough mosquitoes which carries West Nile Virus to justify the cost of spraying

- There is an ‘Unspecified’ Culex species which is not part of the Train set but exist in the Test set. Further monitoring of this species is required to determine if they are carriers of the WNV, as our current model is not trained on this species