Project 4 - West Nile Virus Prediction

Si Hao, Way Keat, Jervin

Problem Statement

At the Disease And Treatment Agency, we are working with data from Chicago's mosquito surveillance and control system to see if we can learn anything about the recent epidemic of the West Nile Virus (WNV) in the city.

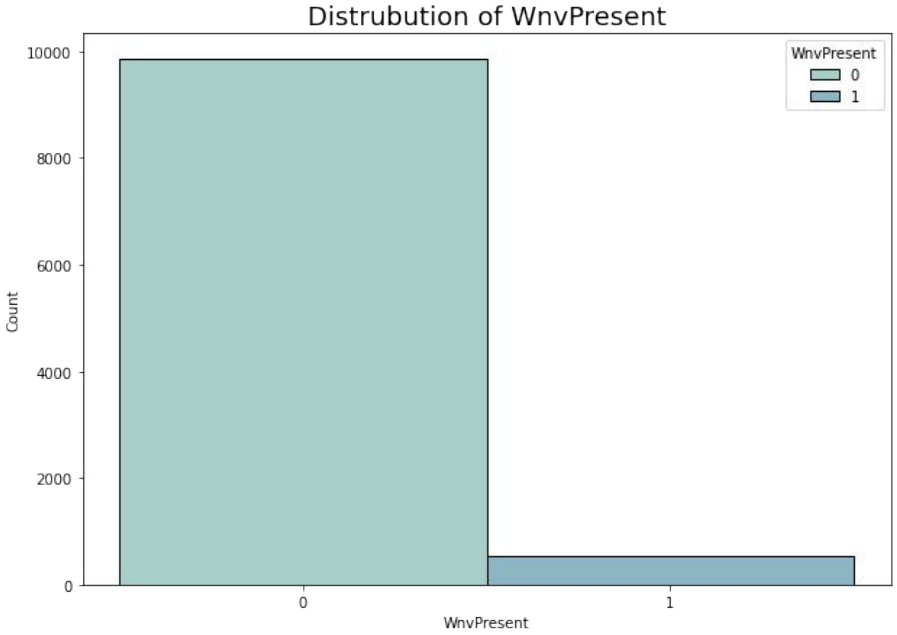
We're tasked to make predictions about the presence of WNV in various sections of the city so officials can decide if, where and when to spray pesticides.

Objective:

1. Predict the presence of WNV based on data for a specific site at a specific time.
2. Conduct a cost-benefit analysis - when to spray the pesticides and where to spray the pesticides

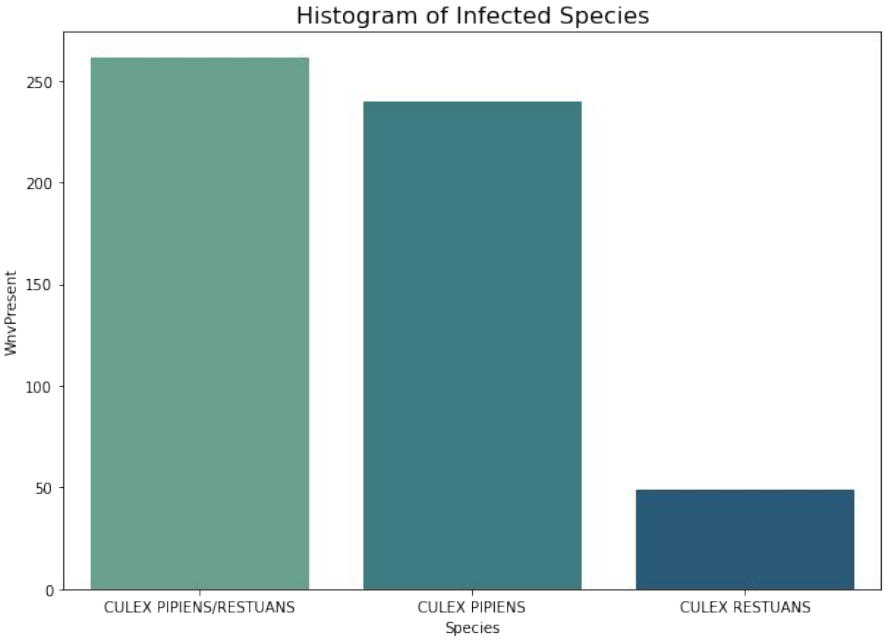
Exploratory Data Analysis

- WnvPresent is our target variable
- Present: 5.3%
- Not Present: 94.7%
- Imbalanced classification problem
- Might affect model's accuracy if not correctly addressed
- Primary metric: AUC



Exploratory Data Analysis

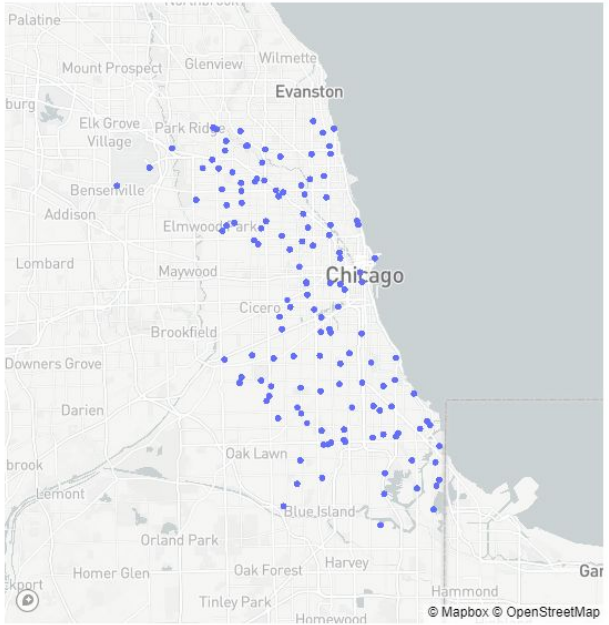
- Vast majority of mosquitos are either Culex Pipiens or Restuans
- Difficulty in differentiating them
- Assuming that both species are equally unidentifiable, Culex Pipiens has been most responsible for WNV.



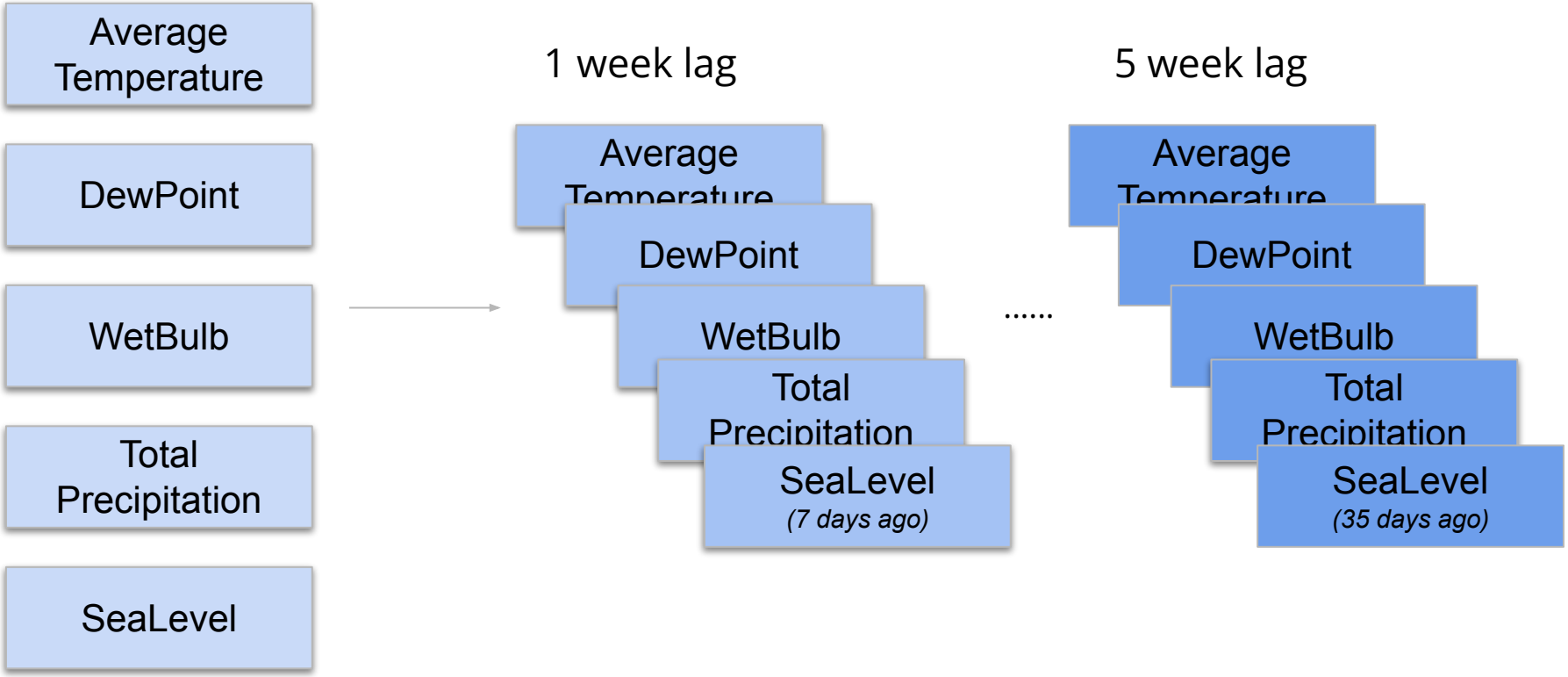
Exploratory Data Analysis

- Location of traps are uniformly distributed
- Lacking information on trap checking schedule
- Some bias might exist because traps might not be check at fixed intervals

Trap Locations



Preprocessing - Lag Features



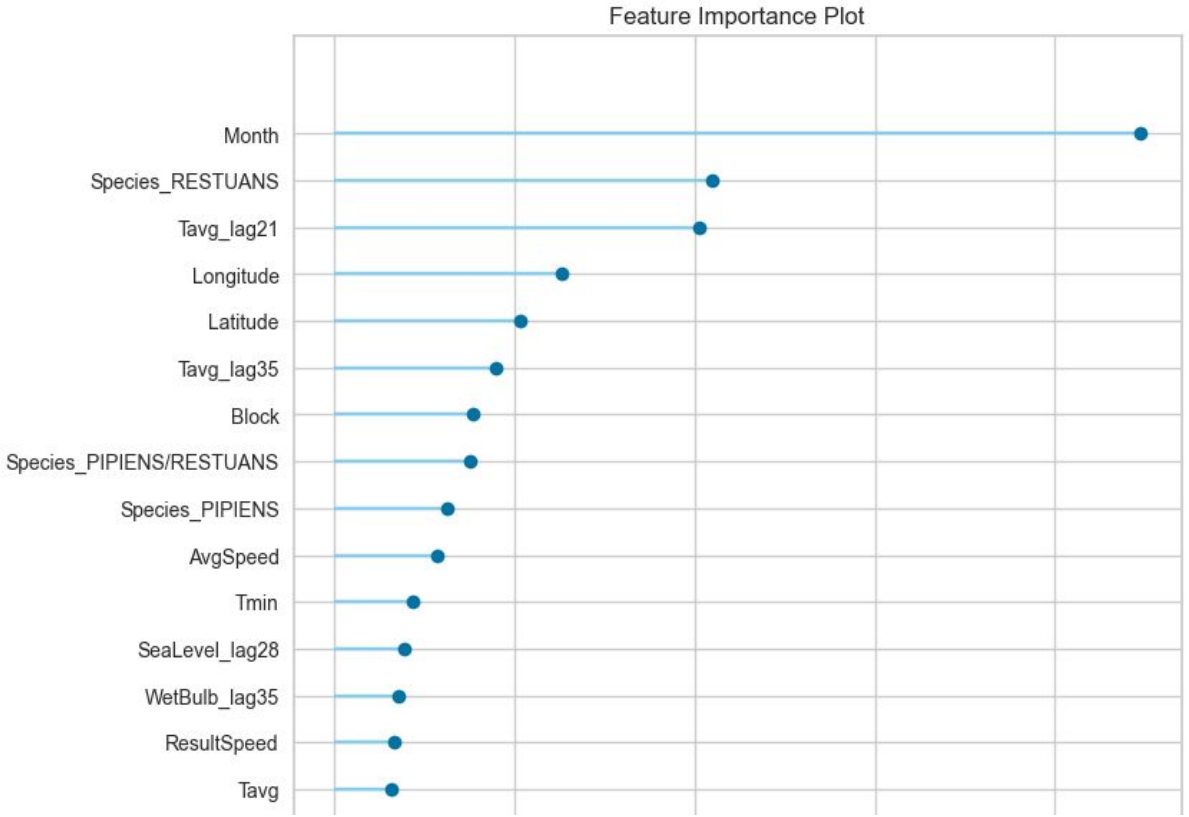
Modelling with a single model

Recap:

- Use of lagged weather indicators
- Numerical features scaled with StandardScaler.
- Categorical features one-hot encoded.
- Missing values are imputed iteratively with LightGBM's impute method
- Class imbalance is handled with SMOTE
- Evaluated using the ROC-AUC metric

	Model	Accuracy	AUC
gbc	Gradient Boosting Classifier	0.8019	0.8052
lr	Logistic Regression	0.6824	0.8036
lightgbm	Light Gradient Boosting Machine	0.8747	0.8032
catboost	CatBoost Classifier	0.8782	0.8002
ada	Ada Boost Classifier	0.7704	0.7974
xgboost	Extreme Gradient Boosting	0.8852	0.7944
lda	Linear Discriminant Analysis	0.6624	0.7909
rf	Random Forest Classifier	0.8649	0.7722
qda	Quadratic Discriminant Analysis	0.5871	0.7709
nb	Naive Bayes	0.5713	0.7310
et	Extra Trees Classifier	0.8683	0.7248
knn	K Neighbors Classifier	0.8032	0.7225
dt	Decision Tree Classifier	0.8599	0.6223
dummy	Dummy Classifier	0.9235	0.5000

Feature Importance



- Lagged weather indicators are important features
- Location - Longitude and latitude are better handled by DT-type models

Improved AUC with ensemble learning

Single model (GBC)

- Train Score = 0.8110
- Test (Kaggle) = 0.6824

Blended model (meta model)

Train Score = 0.8147

Test (Kaggle) = 0.7683

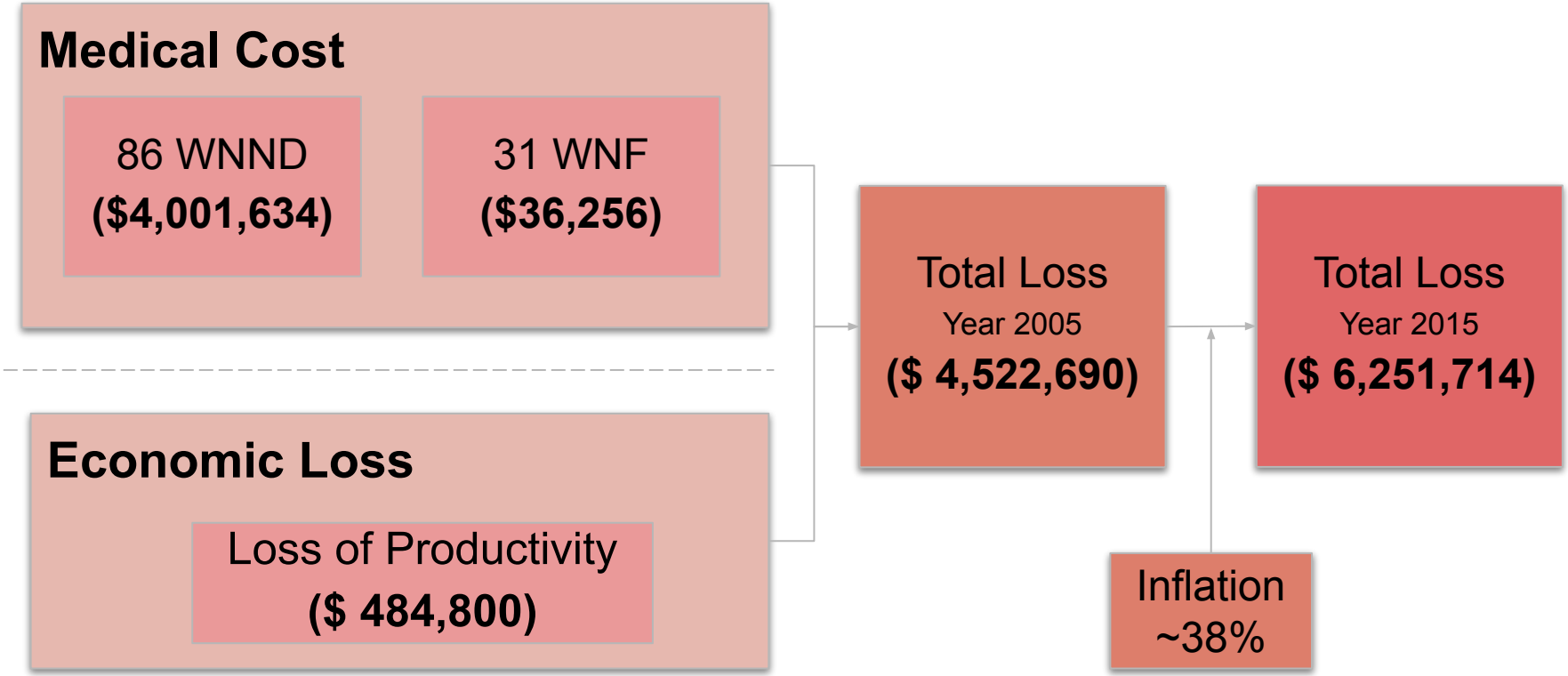
Voting Classifer

"combine conceptually different machine learning classifiers... to predict the class labels. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses."

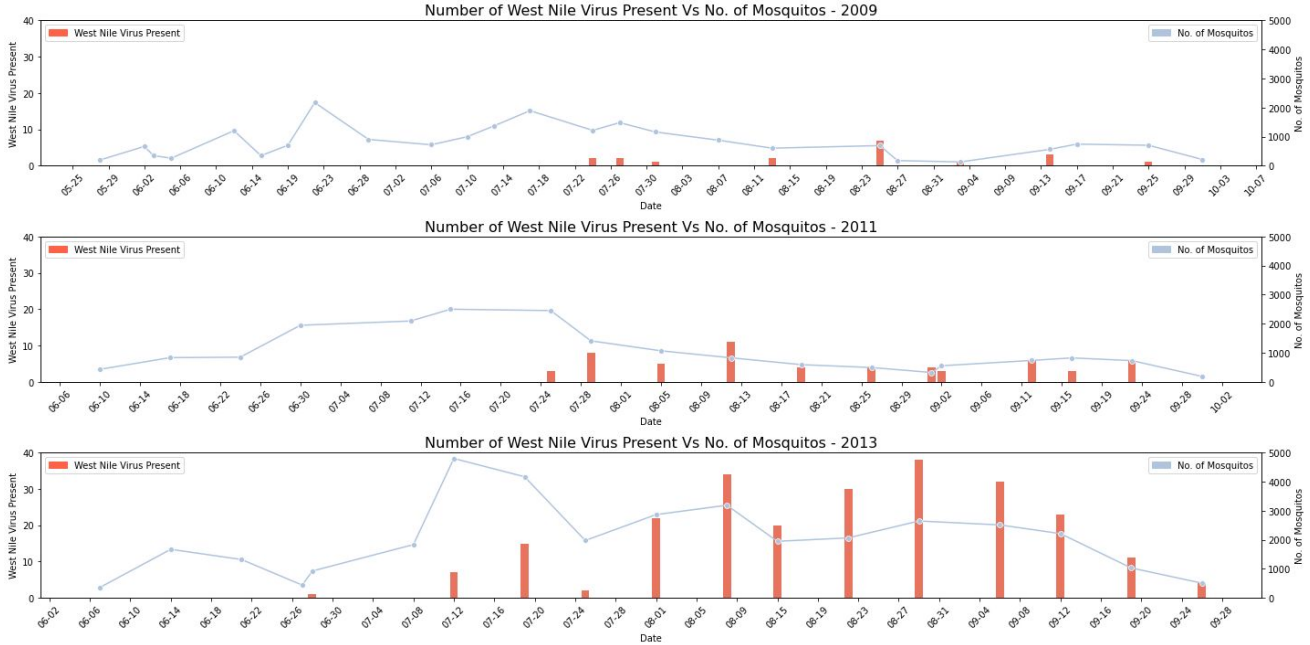
Blending of the following models:

1. Gradient Boosting Classifier
2. Light GBM
3. Logistic Regression

Cost-Benefit analysis - Cost Evaluation

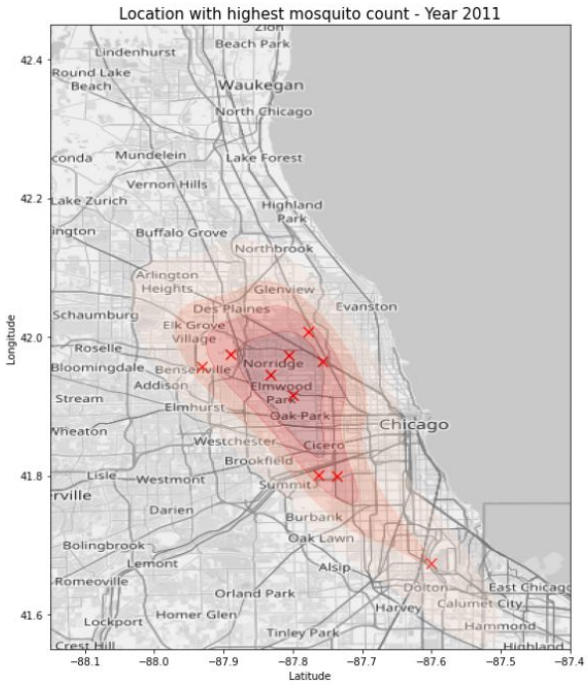


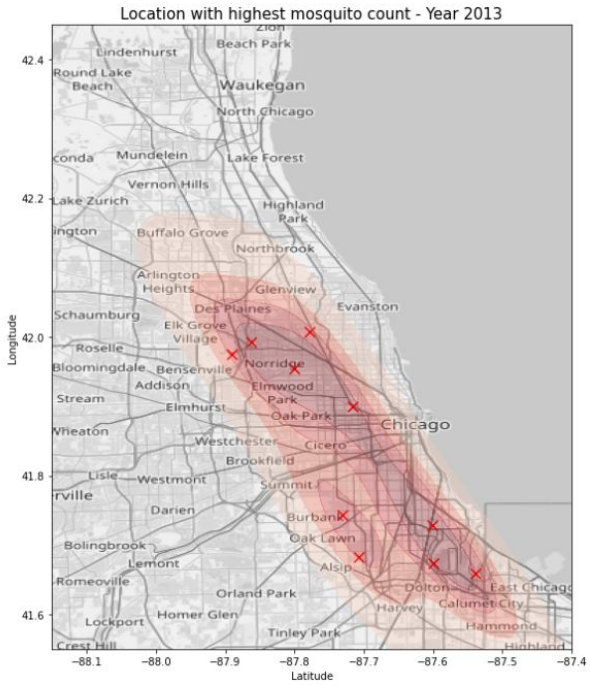
Cost-Evaluation: West Nile Virus Count



- West Nile Virus increases with the numbers of mosquitoes
- Target location with dense population of mosquitoes

Cost-Evaluation: Coverage

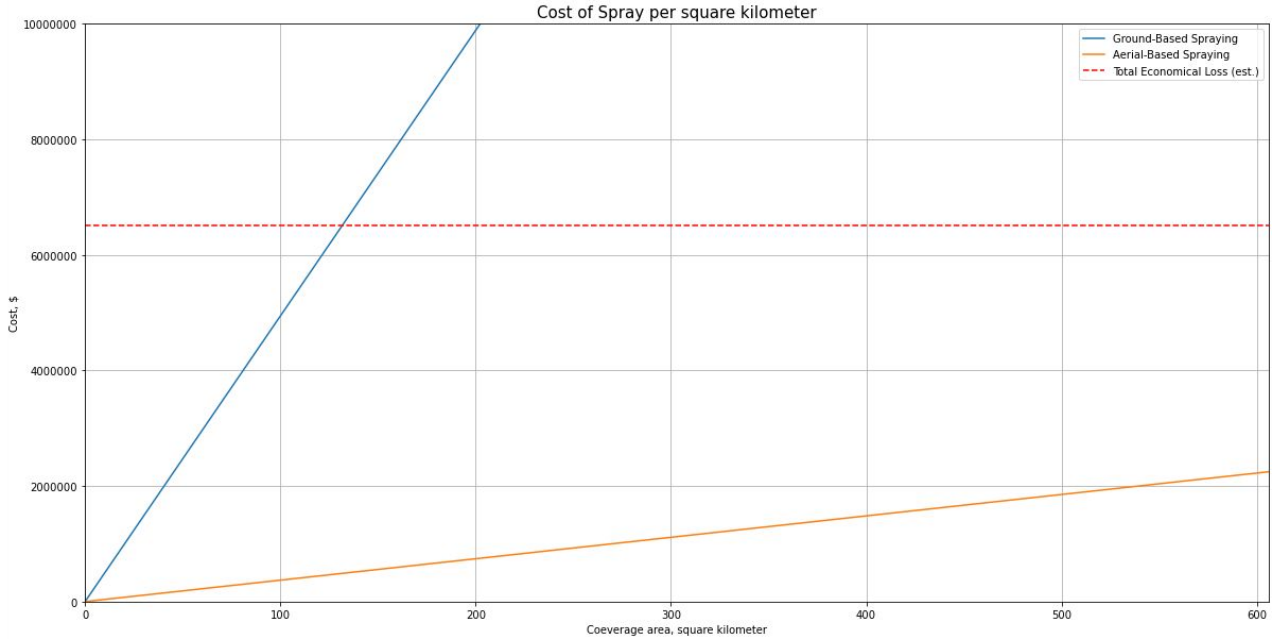




- Mosquitoes population centered around 2 regions

- Approx. 300km^2

Cost-Benefit analysis - Cost Evaluation

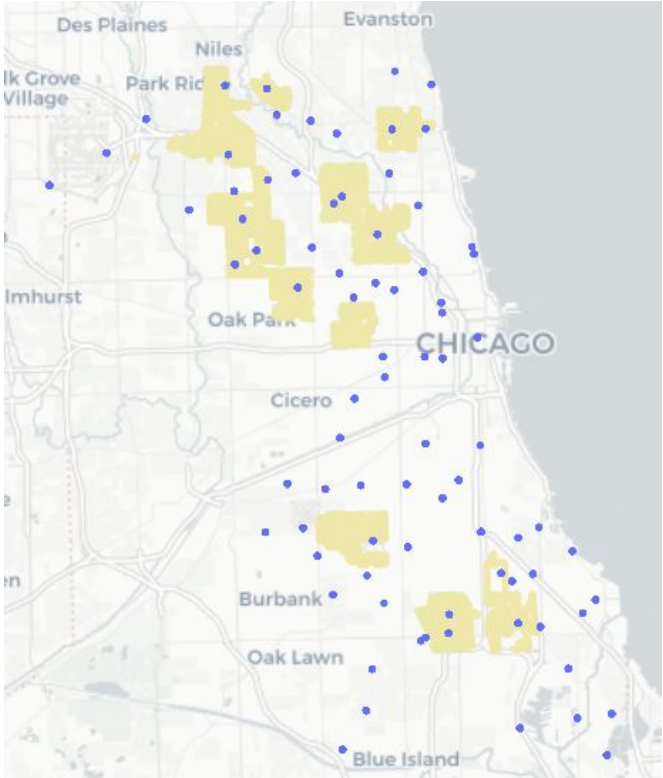


- Ground-based spray are significantly more expensive compared to aerial spray

- Recommended to use aerial spray or a combination of both

Cost-Benefit analysis - Benefit Evaluation

- Zenivex - when sprayed, the the treatment drifts with air currents, effect is strongest at the epicenter of the spray
- Across 2011/13 there were only 10 days (spray events) when spraying occurred in the city.
- Data is limited as not all trap sites are affected by the spraying.

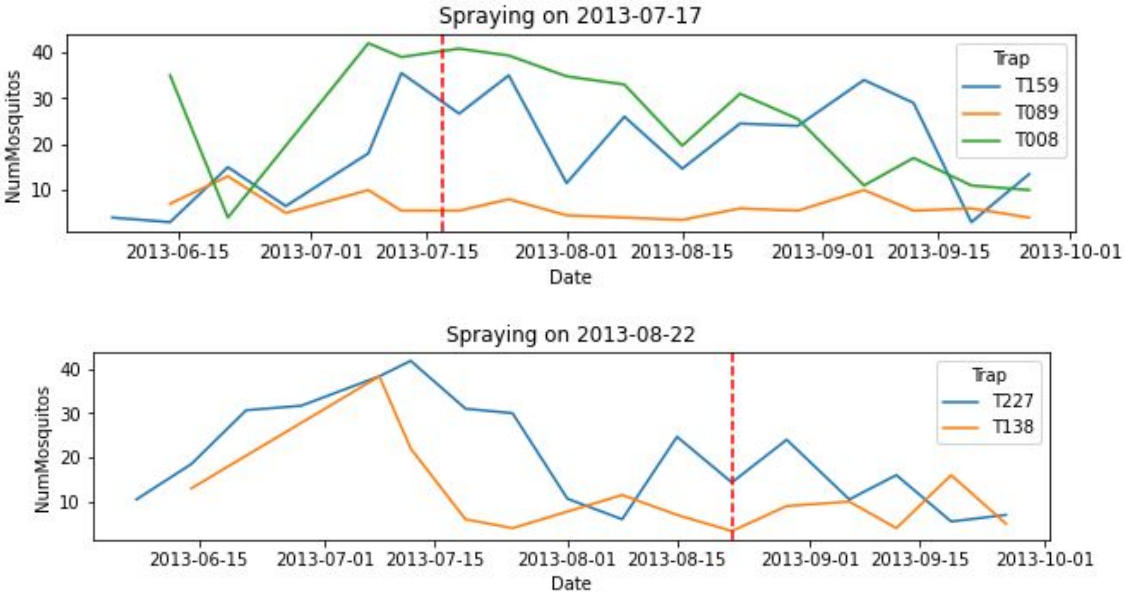


Benefit Evaluation - Analyzing each spray event

Only the traps that are close to a spray site are analyzed.

For each spray event, the traps are identified and the Number of Mosquitos in the traps are plotted over time.

The spray event is shown on the chart as a vertical red line.



Two examples of plots for each spray event - all 10 events were visualized in this manner

Benefit Evaluation - An ambiguous result

One event is inconclusive. **2011-08-29 spray event** had no traps in proximity

Two events suggest some efficacy. **2013-07-25 and 2013-08-15 spray events** caused the number of mosquitos in traps to decrease

The **remaining 7 spray events** saw the number of mosquitos in affected traps actually increase in the following weeks after spraying.

<u>Benefits of spraying are ambiguous</u>

- Causality is difficult to determine as the testing schedule is irregular
- Very few traps were in the spray zones and trap data was collected at irregular intervals
- Traps with no mosquitos, essentially a sign that spraying was successful, are not recorded in the data.
- **Not possible to determine if spraying has any benefit for controlling the mosquito population**

Conclusion

1. Achieved a score of 0.768 by blending several models together.

2. West Nile Virus has a significant economical impact.

3. Aerial spray enables better spray coverage at a significant lower price.

4. Crucial need for a stringent monitoring regime with regular interval testing - and null results (zero mosquitos) must be recorded.