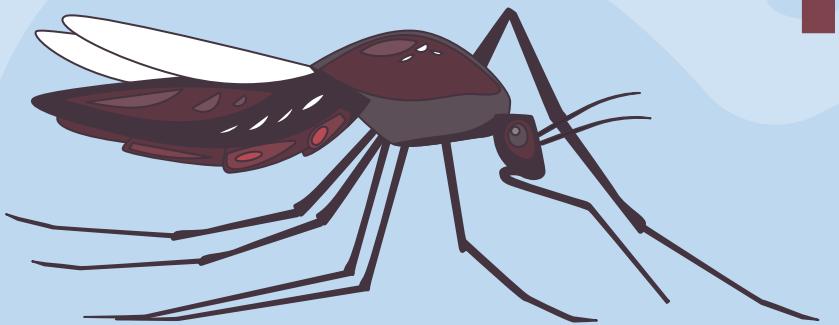
WNV Prediction in Chicago



Emma
Junhao
Benjamin
Gabriel

Overview

- Problem Statement
- Interesting Features
- Model Selection
- Feature Importance
- Cost Benefit Analysis
- Conclusion/Recommendations



Problem Statement

- We are a team of data scientists at the Disease And Treatment Agency. In the last few years, there have been several episodes of the West Nile Virus (WNV) outbreak in Chicago. The aim of our project is to leverage on weather, location and time data to predict West Nile Virus occurrences in Chicago.
- By **understanding these features and whether they can predict the virus**, we will then evaluate **whether spraying areas with pesticides will be an effective tool vis-a-vis medical and economic costs**

Interesting Features

Relative Humidity

Humidity - ambiguous relationship with WNV
Instead of using DewPoint x Tavg interaction term - worked out RH=100 x (e / es)

Lagged Weather Features

Culex: egg to adult ~7 to 10 days.
Lagged weather features by 14 days - factor in time needed to pick up the virus

Risk Classification

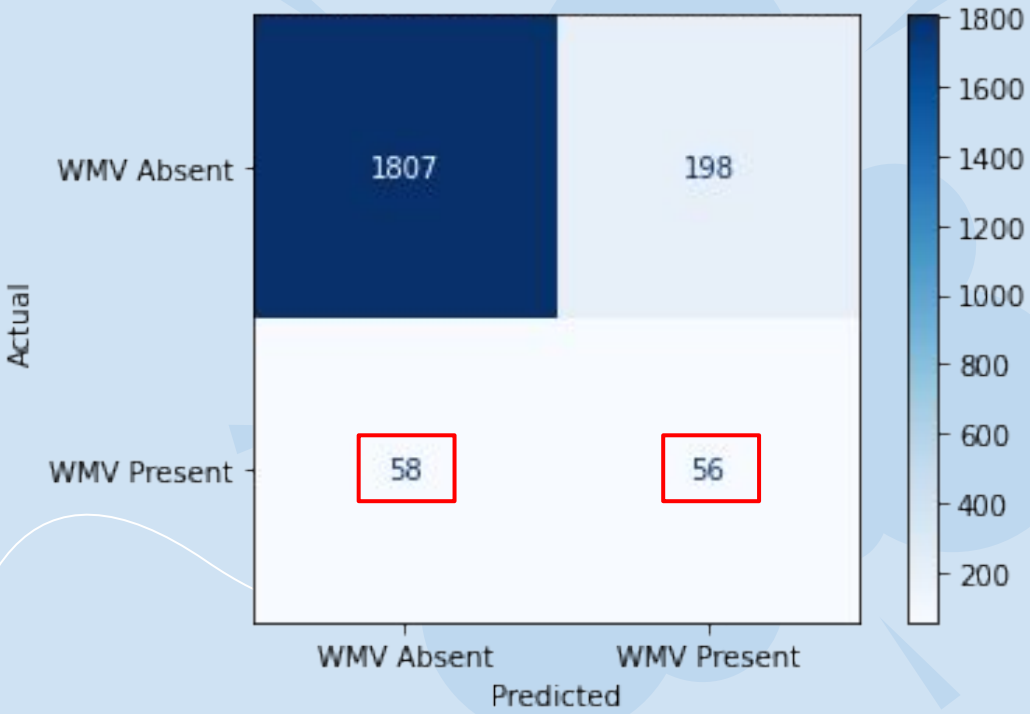
136 unique trap addresses
5 categories based on WNV occurrences:
Low, Very Low, Medium, High, Very High

Which model is suitable?

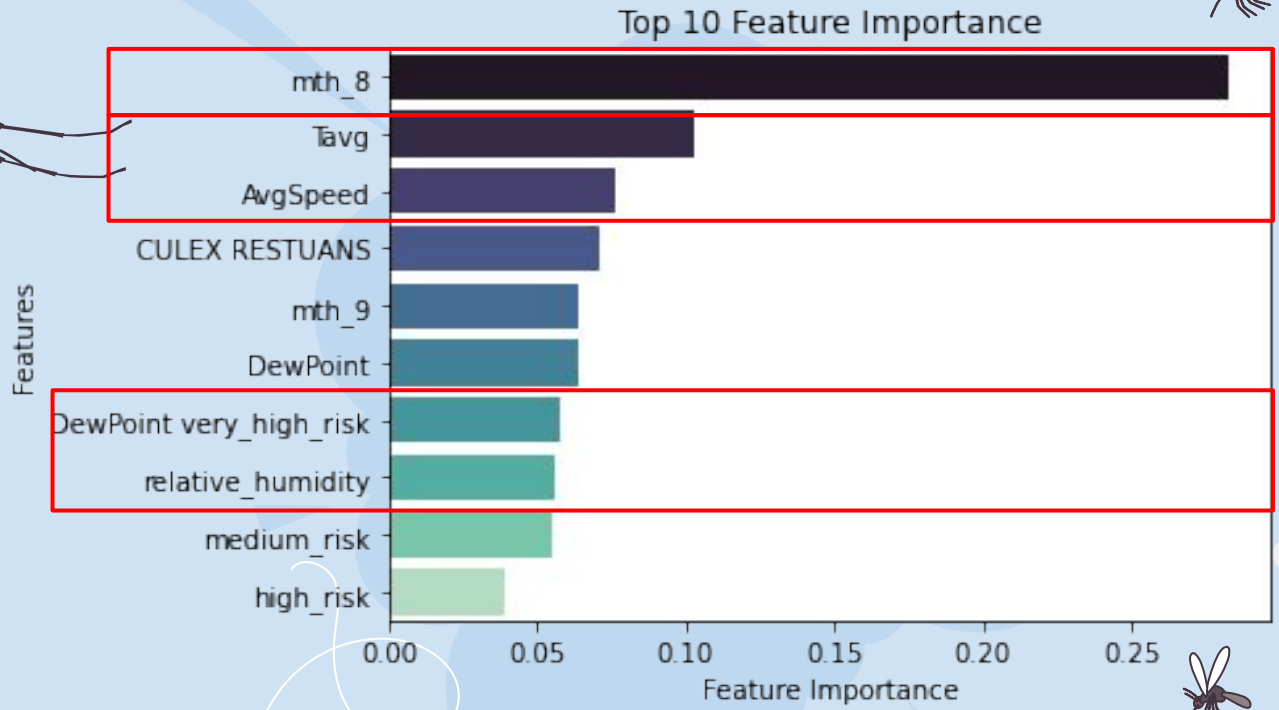
(Selected Model)

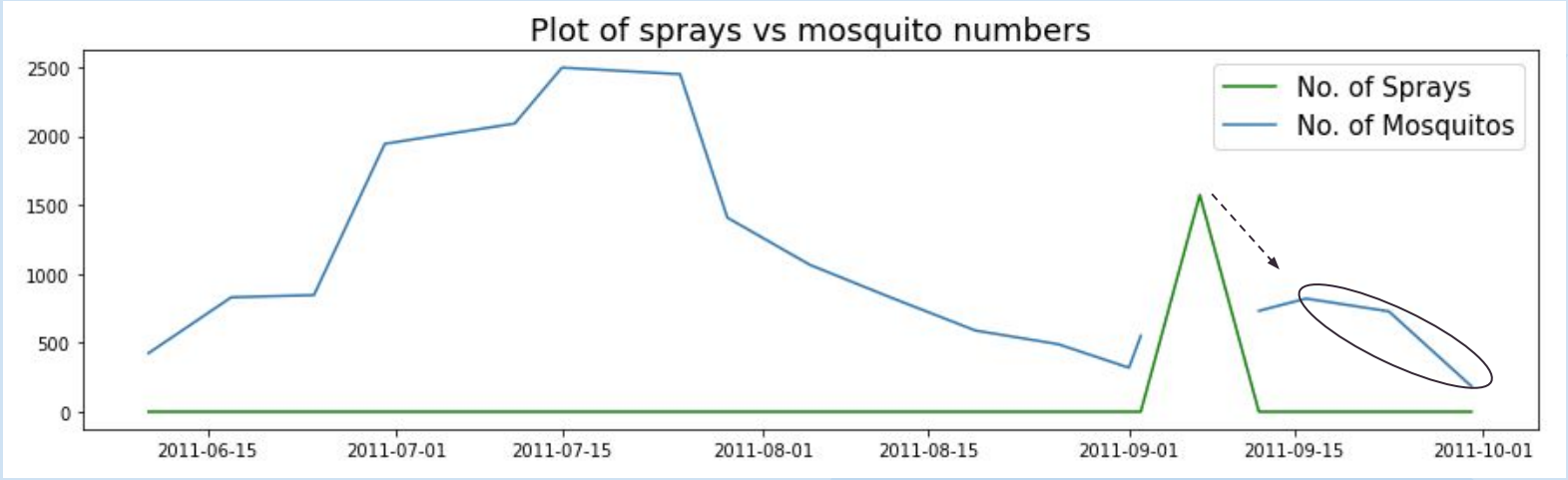
Logistic Regression	Gradient Boost	XGBoost	ADA Boost	Random Forest
81.8% Test Score	**82.8%** Test Score	**82.9%** Test Score	**83.0%** Test Score	**81.5%** Test Score
84.5% Train Score	89.4% Train Score	89.6% Train Score	89.3% Train Score	93.7% Train Score
• Baseline • No parameter tuning	• SMOTE • Best recall: 0.49 • Precision: 0.22 • Best F1: 0.3	• SMOTE • Recall: 0.48 • Precision: 0.21 • F1: 0.29	• SMOTE • Worst recall: 0.11 • Best precision: 0.25 • Worst F1: 0.15	• SMOTE • Recall: 0.38 • Most overfitted model

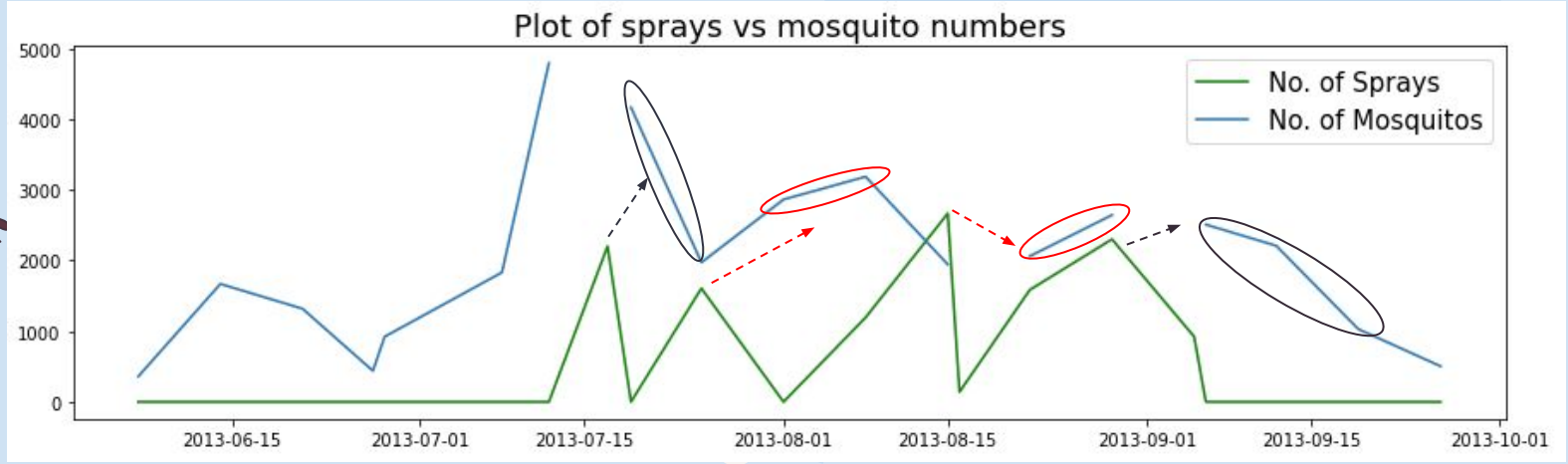
Confusion Matrix of Actual vs Predicted WMV

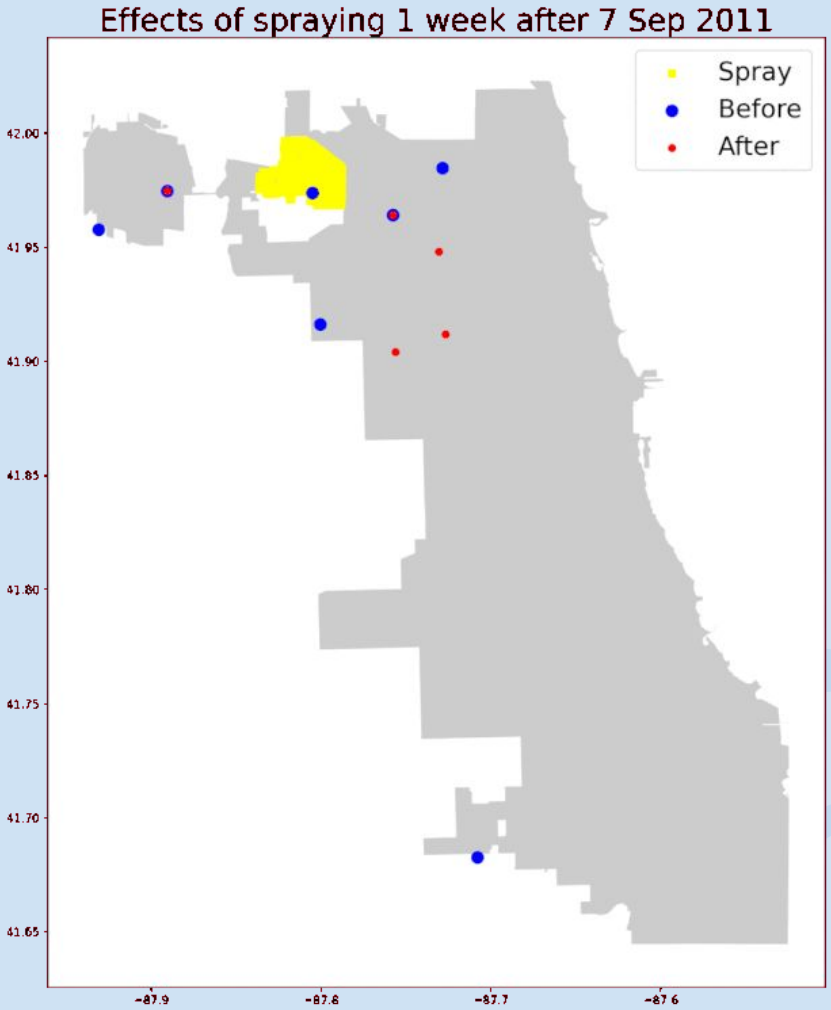


Top 10 Feature Importance




Plot of sprays vs mosquito numbers
No. of Sprays
No. of Mosquitos
2500
2000
1500
1000
500
0
2011-06-15
2011-07-01
2011-07-15
2011-08-01
2011-08-15
2011-09-01
2011-09-15
2011-10-01


Plot of sprays vs mosquito numbers
No. of Sprays
No. of Mosquitos
5000
4000
3000
2000
1000
0
2013-06-15
2013-07-01
2013-07-15
2013-08-01
2013-08-15
2013-09-01
2013-09-15
2013-10-01



Date	Effective Areas	Ineffective Areas
7 Sep 2011	1	0
17 Jul 2013	1	1
25 Jul 2013	1	0
8 Aug 2013	2	0
15 Aug 2013	1	3
22 Aug 2013	0	3
29 Aug 2013	1	2
5 Sep 2013	0	1
Total	**7**	**10**

41.18% success rate!

Cost Benefit Analysis

Taking into account:
Previous Sprays 41.9%
< Predicted Sprays 49%

Spray

Benefit

- Reduce disease in population
- Lower load on hospital services
- Overall better quality of life
- Approximate cost of $4.5 million

No Spray

Cost

- High Medical Expenses
- Loss of Workplace Productivity
- More death rates
- Approximate cost of $1 million

Conclusion

- **Gradient Boost Classifier** has a roc-auc score of **83%** with highest recall score of **49%**
- Concentrate spray efforts during month of **August**
- Consolidate weather information on **Temperature / Wind Speed / Humidity**
- **Location** a relatively strong predictor, WNV may be **transient**

Recommendations

Adult Control	Spray during (month/location/weather)
Larval Control	Larva-ciding water bodies
Public Education	Educated public on personal protection
Bird Surveillance	Monitor birds known to be carriers, data collected to enhance model