Predicting West Nile Virus in Chicago

Team A

Dora, HJ, Kenny

Agenda

- Background
- Problem Statement
- EDA
- Feature Engineering
- Modeling
- Cost Benefit Analysis
- Conclusion

Background

Who are We?

DATA SCIENCE, consultants engaged by Department of Public Health (DOPH)

What happened?

Epidemic of West Nile Virus in Chicago

What are we doing?

DOPH set up a surveillance and control system.

We collect data of the mosquito population, species, WNV, weather conditions

Problem Statement

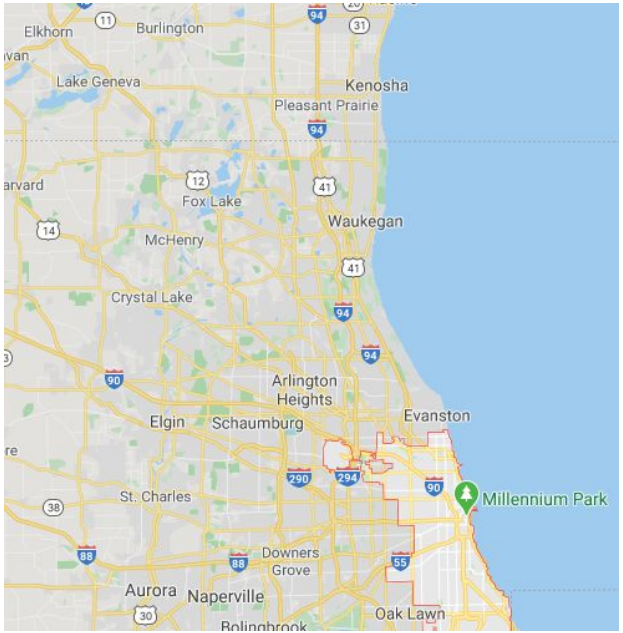
Should we perform aerial spraying?

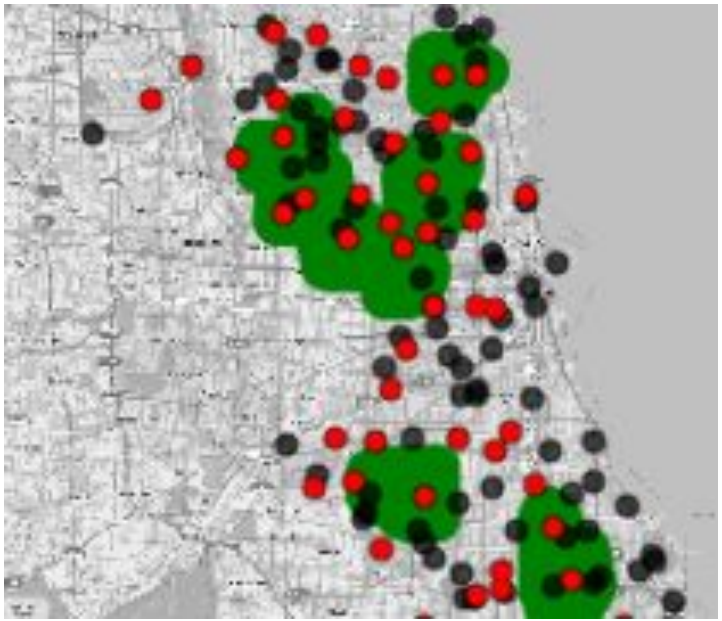
WHEN and WHERE to conduct aerial spraying?

Mosquito surveillance

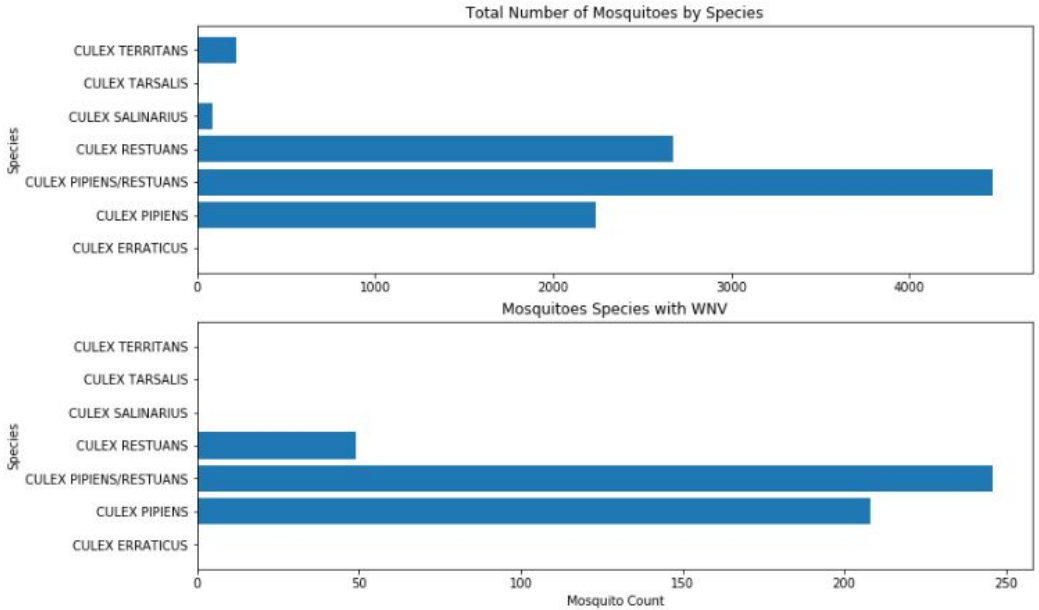
Photo: The Atlanta Journal-Constitution

Trap Locations

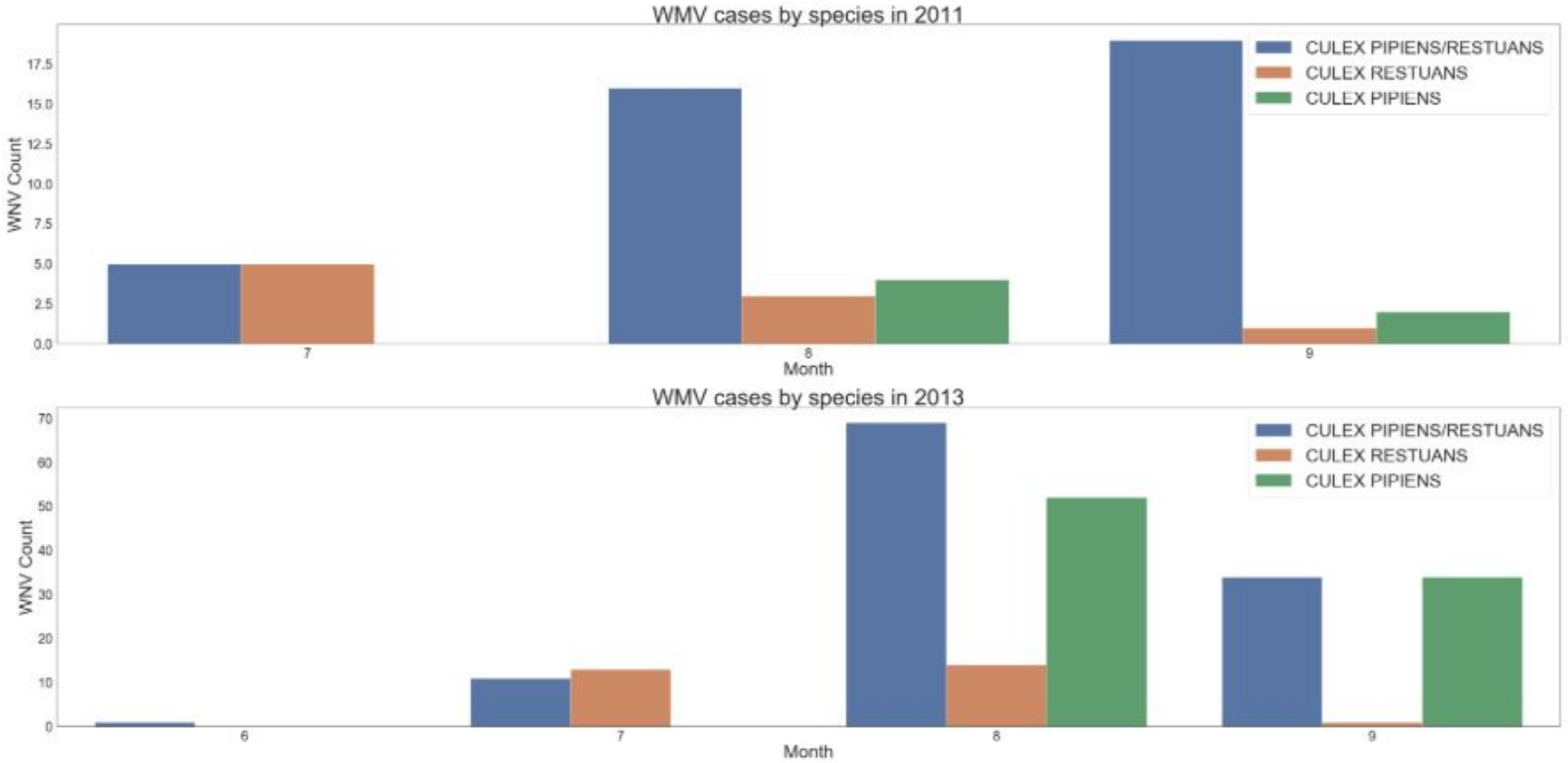




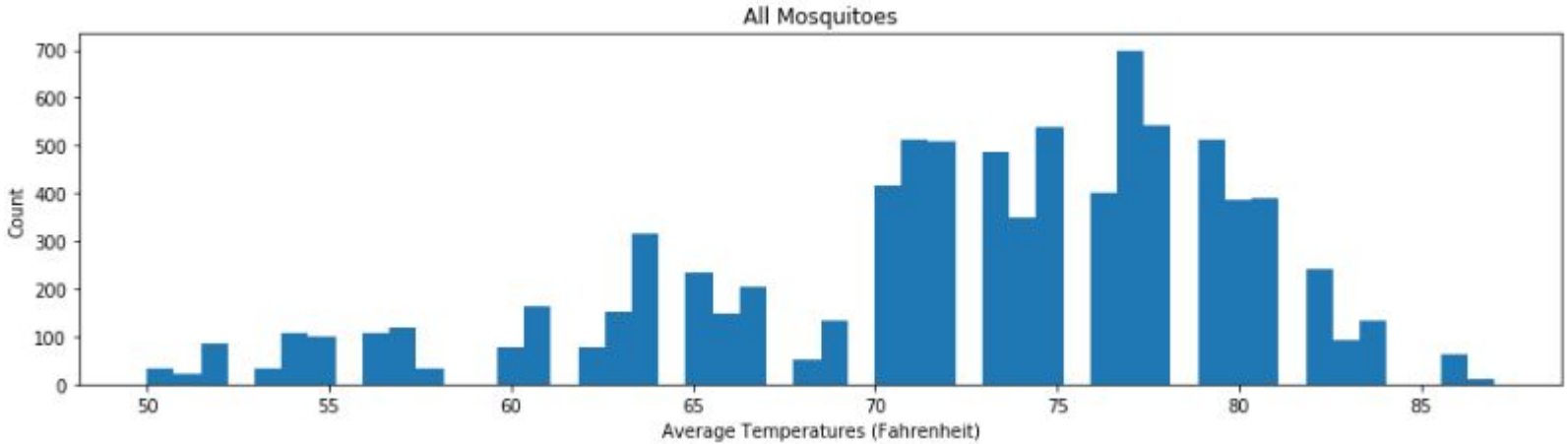
Mosquito Species and WNV



WNV species over the months

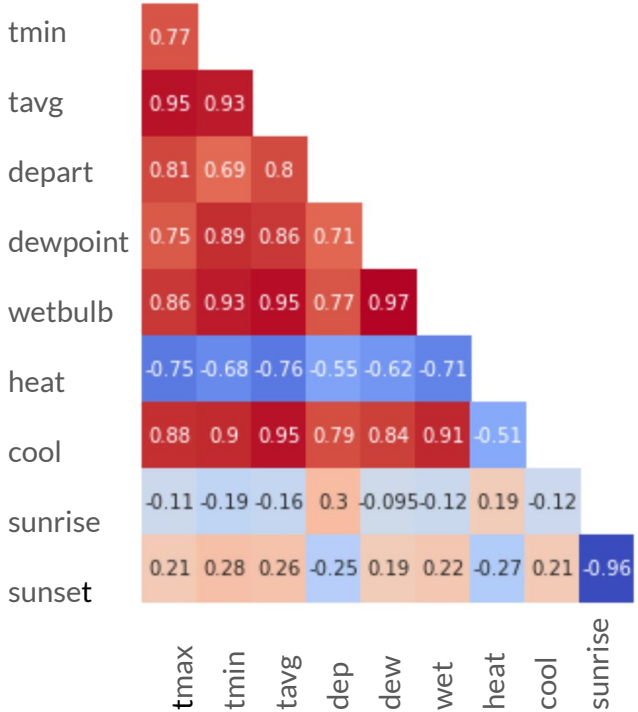


Mosquitoes Thrive in Warmer Weather



Feature Engineering - Collinear Features

- Merged train/test data with weather dataset
- Weather features highly collinear
- Dropped **tmin, tmax, depart**
  - Engineered temperature range feature
- Dropped **wetbulb**
- Dropped **heat and cool**
- Dropped **sunset**
  - Engineered daylight hours feature



Feature Engineering - Low Correlation

- Dropped features with low correlation to presence of WNV
  - All CODESUM features
    - Key weather information, i.e., temperature, precipitation, wind, already captured
  - UNSPECIFIED CULEX species feature
    - No correlation with presence of WNV
    - CULEX PIPIENS, CULEX RESTUANS are carriers of WNV

Feature Engineering - New Features

- **Relative humidity** feature
  - Mosquitos breed better in humid areas
  - Calculated by applying a formula using average temperature and dewpoint features
- Added **time lag** element to weather features to investigate the effect of previous weather patterns on WNV
  - Mosquito larvae take 10 to 14 days to develop to adult mosquito
  - Weather features with time lag of 7, 10 and 14 days added

Classification Models

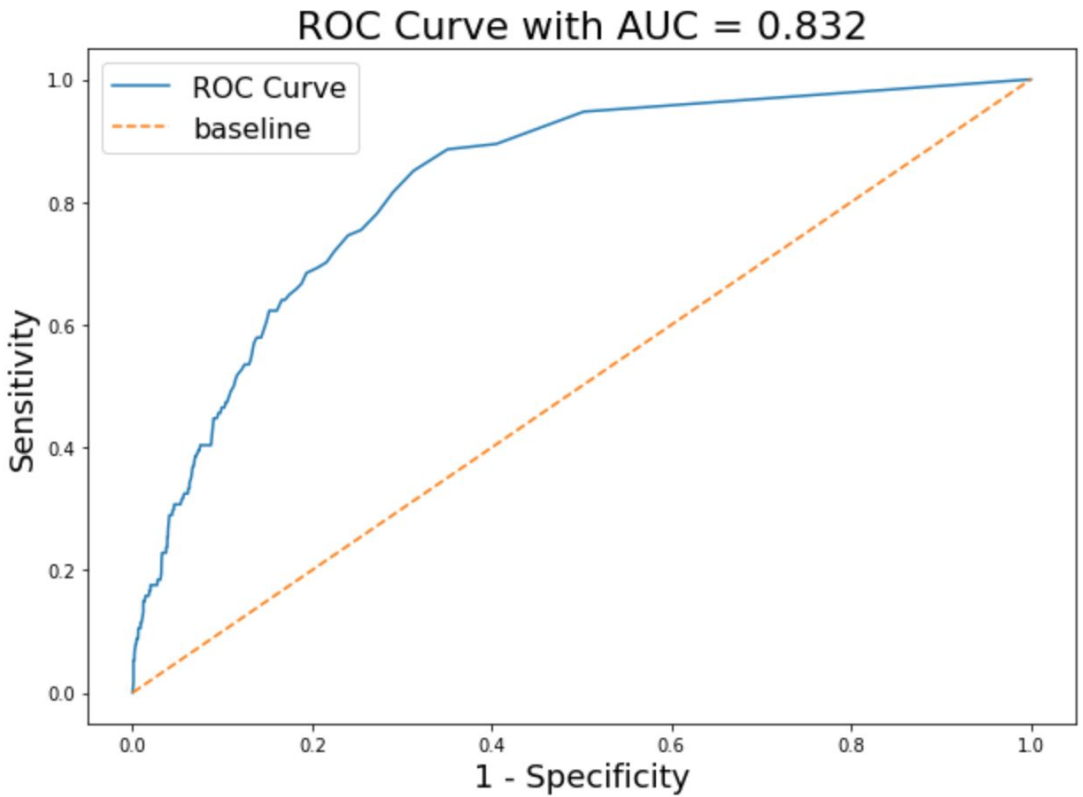
- Issue of unbalanced classes
  - SMOTE
- Classification model to predict if WNV is present
  - Logistic Regression
  - K Nearest Neighbors
  - Random Forest
  - Gradient Boosting
  - XGBoost

```
0.0    0.946077
1.0    0.053923
Name: wnv_pres,
```

Best Model: XGBoost

ROC AUC score = 0.83

Shows that the model is well capable of differentiating classes (WNV present and WNV not present)



Best Model: XGBoost

	Predicted WNV absent	Predicted WNV present
Actual WNV absent	1939	66
Actual WNV present	88	26

TN: 1939
FP: 66
FN: 88
TP: 26

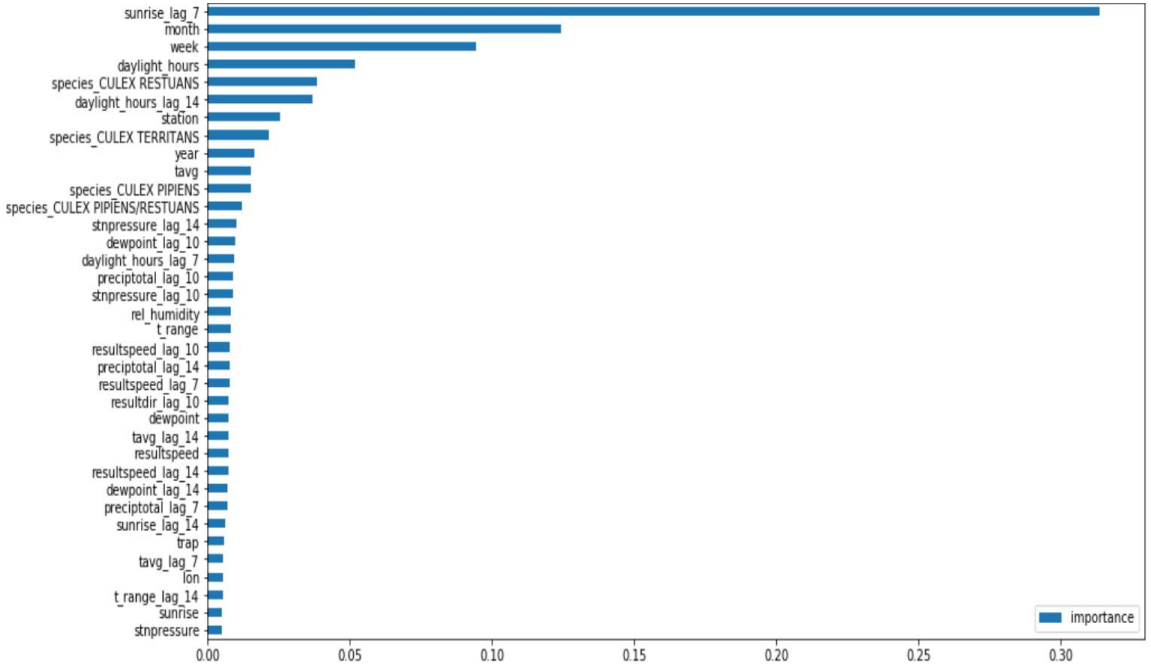
Lowest number of False Positives and False Negatives

- Low FP: Minimise wastage of resources by not spraying at areas falsely predicted to have WNV
- Low FN: Reduce human cost of WNV by accurately predicting areas where WNV is present (avoid Type II errors)

Feature Importance

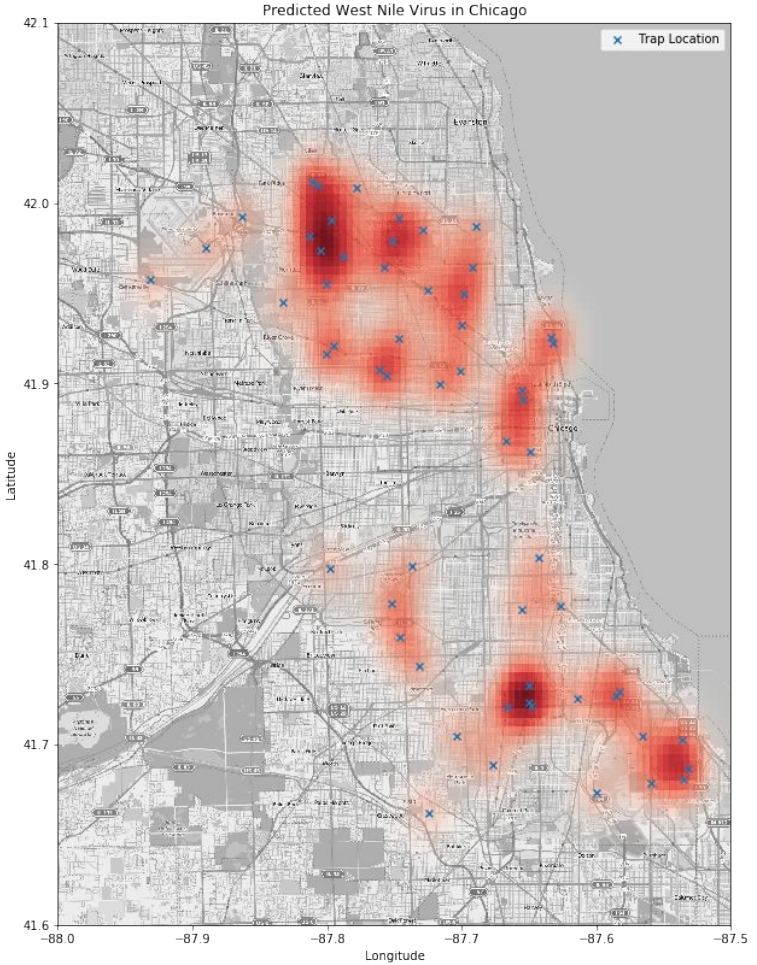
Top features affecting prediction of WNV

- Time: Week, Month
- Weather 7, 10 and 14 days ago
  - Daylight hours: sunrise
  - Temperature
  - Precipitation, humidity



Predictions Map

- Red spots are predicted areas with high concentrations of the WNV



Case Study: Sacramento (2005)

- 163 cases reported
- 2 types of WNV:
  - Cost to treat WN Fever: USD8,000
  - Cost to treat WN Neuroinvasive: USD33,000
- Cost of the WNV: USD2.3 million
  - Cost of treatment
  - Cost of productivity loss
- Cost of Vector Control: USD700,000

Sacramento County vs City of Chicago

- Sacramento:
  - Population: 1.5 million
  - Area: 2574 km^2
  - Population Density: **583** per km^2
- Chicago:
  - Population: 2.7 million
  - Area: 606 km^2
  - Population Density: **4455** per km^2

Proceed with with Aerial Spraying

- It takes 15 neuroinvasive cases to **breakeven** on the cost of aerial spraying
- The cost of the virus (USD2.3 million) vs cost of vector control (USD700,000)
- When we scale the population density, the cost decreases while the benefits per km^2 increases

Recommendations and Conclusions

- Proceed with aerial spraying
- WHEN:
  - Sunrise, where the mosquitos are most active
  - 1-2 weeks before the onset of warmer and humid months
- WHERE:
  - Focus on areas with high populations of CULEX RESTUANS and CULEX PIPIENS
  - Areas predicted to have high concentration of WNV present, according to predictions map

Next Steps

- Education is key
  - Stagnant water to be removed
  - Stay indoors outside of daylight hours in those months, when mosquitos are most active
- More data
  - Bird population data

THANK YOU