West Nile Virus Prediction

Squad 2:
Charlie, Cheng, Kulwant, Nick

Contents

- Context
- Data
- Features
- Models
- Conclusion

Context

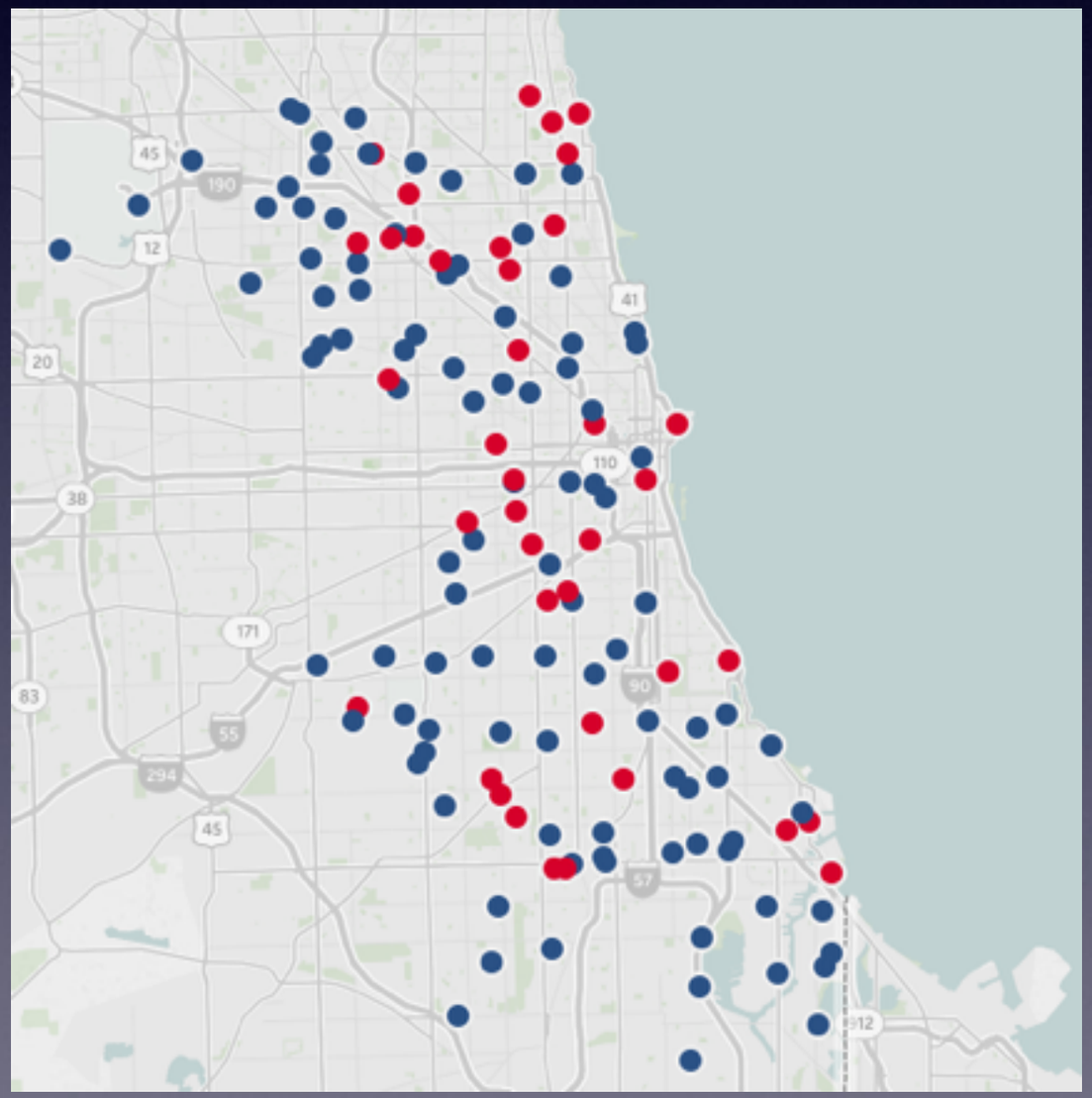
- West Nile Virus is an arbovirus
- Not particularly lethal
    - Only 1 in 150 infected people show symptoms that lead to death
- But it does scare people
- Chicago DPH would like to be able to predict where infected mosquitos are likely to be

Data

- Three datasets considered:
  - Train dataset
  - Weather dataset
  - Spray dataset

EDA

Incidence of West Nile Virus Around Chicago



Train Dataset

- Reports results from specific mosquito traps
- Organized such that when the number of mosquitos exceed 50, they are split into another record (another row in the dataset), such that the number of mosquitos per row is capped at 50.

Weather Dataset

- Two weather stations
- Lots of data
- Data frequently recorded as strings

Spray

- Latitude, longitude, dates of spraying for mosquitos
- Only two years' worth of data
- We assume one spray would have a lasting effect (3 weeks) for a wide range

Feature Selection

- Based on Random Forest feature importance

	Importance	Selected
Longitude	0.226082	X
Trap	0.213049	
Latitude	0.202767	X
Species	0.163729	X
Week	0.016426	X
Sunrise1	0.014968	X
Sunset1	0.014003	
Tmax2	0.009448	
WetBulb1	0.009019	
Sprayed	0.008520	

Models

- We threw everything but the kitchen sink at WNV
- KNN, Random Forest, SVM, Naive Bayes, Neural Network
- Train-Test Split
- Grid search to optimize ROC-AUC

Model Performances

- Random Forest: 0.61572
- Naive Bayes: 0.68945
- KNN: 0.69683
- SVM: 0.70307
- Neural Network: 0.71183

Conclusion

- Recalling that only 5% of the mosquitos captured were found to have WNV
- And that WNV is only permanently debilitating or fatal in 1 case in 150
- We found that, based on the features we selected, we had an auc of .71