Prediction of West Nile Virus in Chicago

Group 1
Kun Won, Lloyd, Stephanie

Executive Summary

We were hired by the Chicago Department of Public Health (CDPH) to use data to gain insights on on the occurrence of West Nile Virus in Chicago. Key insights are:

What are the key factors leading to an outbreak?

1. Seasonality, as driven by weather conditions
2. Presence of particular mosquito species
3. Location

Modelling based on these factors was effective at predicting outbreaks

Is it cost-effective to spray the city to prevent outbreaks?

Yes, economic benefits outweigh the costs based on our cost benefit analysis.

Recommendation

Chicago should build a predictive model using lagged weather, trap and location data to determine which locations it should spray, in order to better combat outbreaks

Table of contents

Outbreak of the West Nile Virus

Leading cause of mosquito-borne disease in the continental United States

Spread to people by the bite of an infected mosquito

About 1 in 150 infected people develop a serious, and sometimes fatal illness

No vaccines or medications to treat the West Nile Virus in people

Prevention & Eradication

Hired by the Chicago Department of Public Health (CDPH) to use data to gain insights on

1. **What are the key factors leading to an outbreak?**
2. **Is it cost-effective to spray the city to prevent outbreaks?**

Allow the CDPH to more efficiently and effectively allocate resources to combat outbreaks

Table of contents

Methodology

01 **Data Cleaning**

02 **EDA**

03 **Feature Engineering**

04 **Modelling**

05 **Inference**

Datasets Used

Train

Test

- Trap data (Location, Species, Date)
- Train data from 2007, 2009, 2011, 2013
- Test data from 2008, 2010, 2012, 2014
- Train data includes number of mosquitoes and presence of WNV

Weather

- Weather data from NOAA collected from 2 stations in Chicago
- Weather features matched to trap dataset by distance to nearest station

Spray

- Limited data on sprays (location, date) from 2011 and 2013
- Excluded from model due to incomplete data

Features Engineered

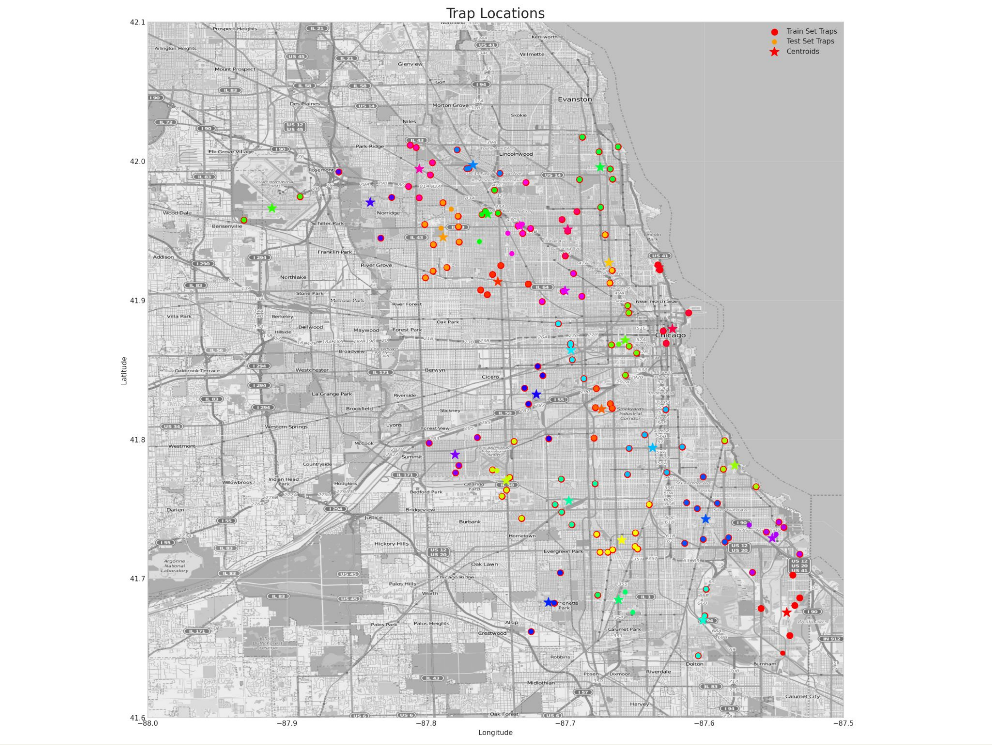
Totalling 101 explanatory variables

Clusters of Trap Locations
Using KMeans to account for geographical variations e.g. terrain, neighborhood sanitation

Relative Humidity
Approximated using dew point and average temperature

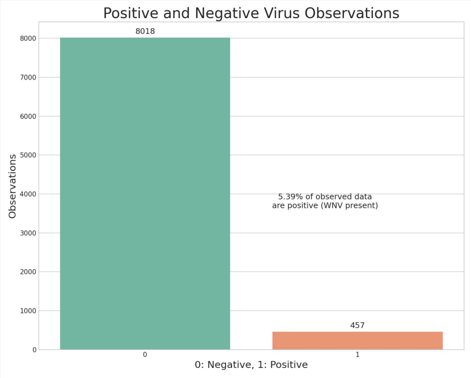
Daylight Hours
Computed from sunrise and sunset timings

Moving Averages
For Temperature, Precipitation, Humidity, Wind Speed, Daylight Hours



Techniques Used in Modelling

SMOTE
Bootstrapping train set to deal with imbalanced classes



StandardScaler
Scales variables for regularisation to prevent overfitting

One-Hot Encoding
Dummified mosquito species and clusters

Feature Selection
Selected moving averages which were correlated with the target

GridSearchCV
Finding and tuning to the best hyperparameters

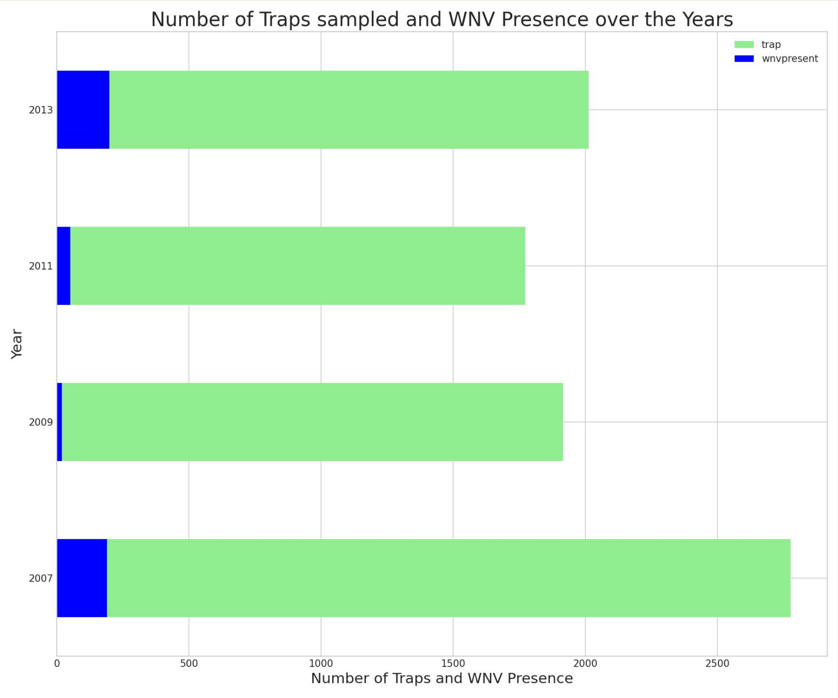
Logistic Regression, Random Forest, XGB Classifier, Gradient Boost
Classification Models used for Supervised Learning

Best Model
Logistic Regression

Model	Train AUC Score	Test AUC Score	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Dummy (Baseline Model)	0.500	0.500	0.946	0.000	1.000	0.000	0.000
XGB Classifier	0.998	0.843	0.800	0.675	0.807	0.166	0.266
Random Forest	0.983	0.836	0.829	0.570	0.844	0.172	0.264
Gradient Boost	0.961	0.829	0.785	0.711	0.790	0.161	0.263
Logistic Regression	0.901	0.839	0.756	**0.754**	0.756	0.149	0.249

Table of contents

Outbreaks were worse in 2007 and 2013. Why?



Factors leading to an Outbreak

Number of Mosquitos

Mosquito Species

Location

Seasons

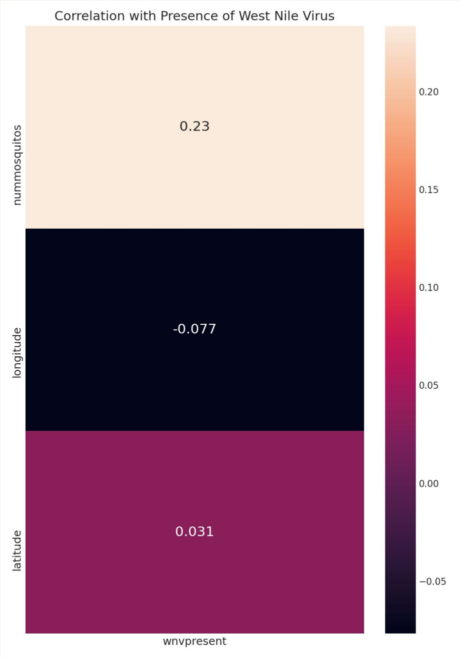
Precipitation

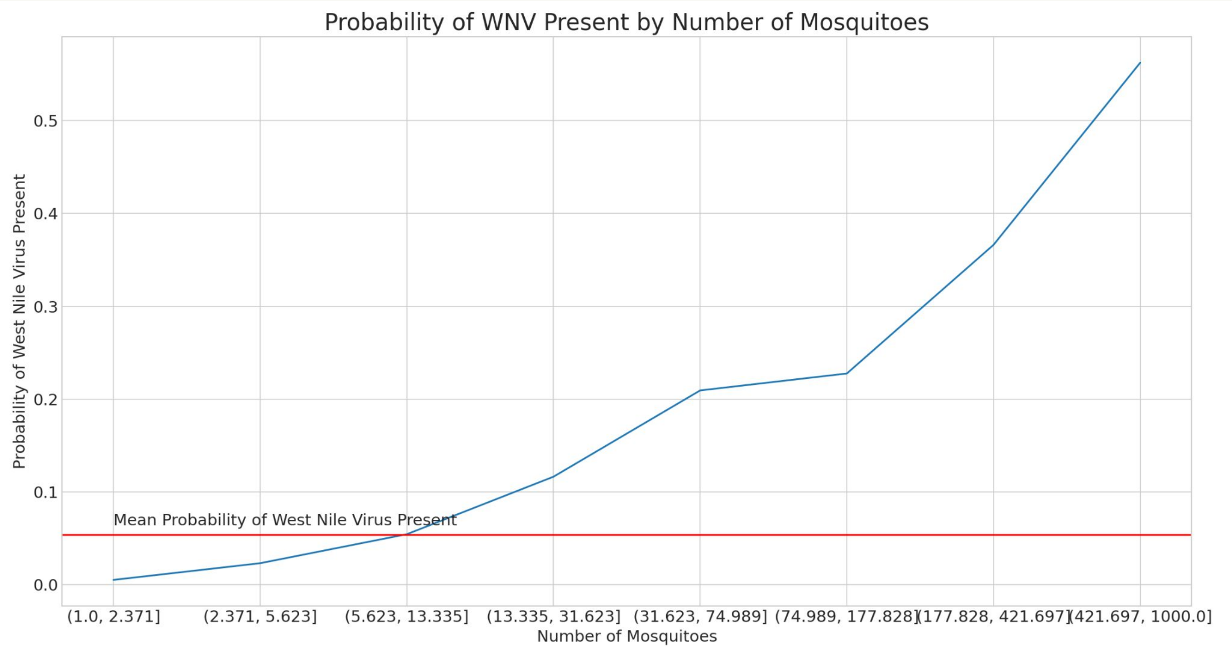
Humidity

Daylight Hours

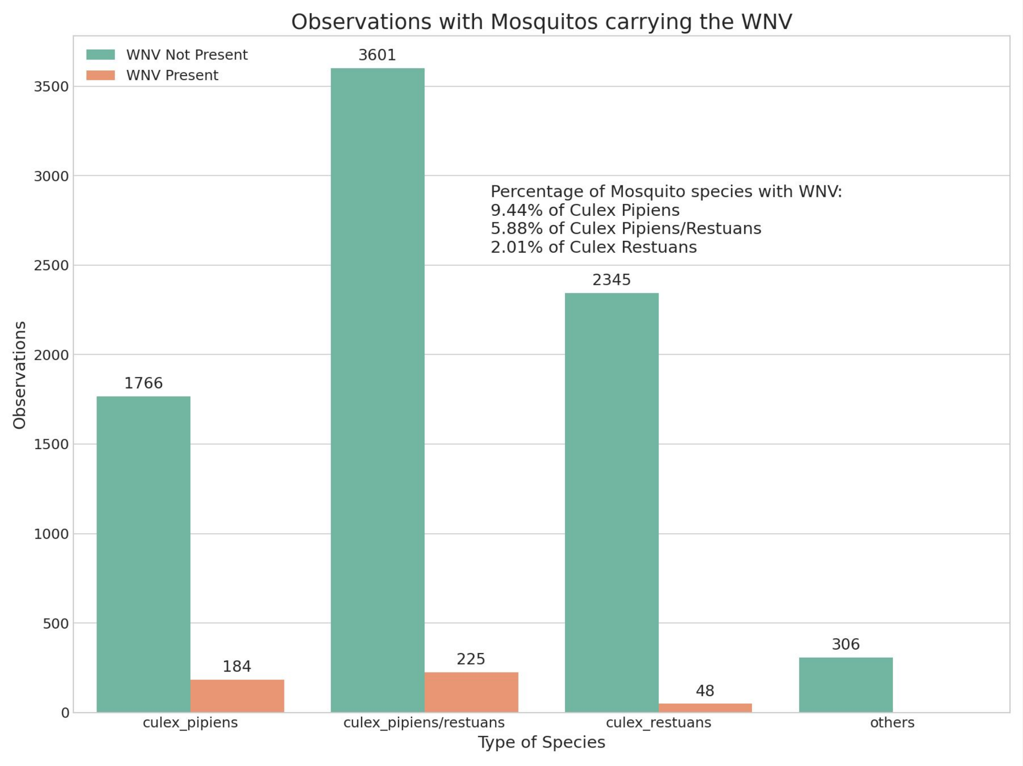
Temperature

Higher number of mosquitoes
=
Higher probability of West Nile Virus





2 out of 6
Species of mosquitoes exhibits the WNV



With WNV Presence:
Culex Pipens (9.44%)
Culex Restuans (2.01%)

Factors leading to an Outbreak

Number of Mosquitos

Mosquito Species

Location

Seasons

Precipitation

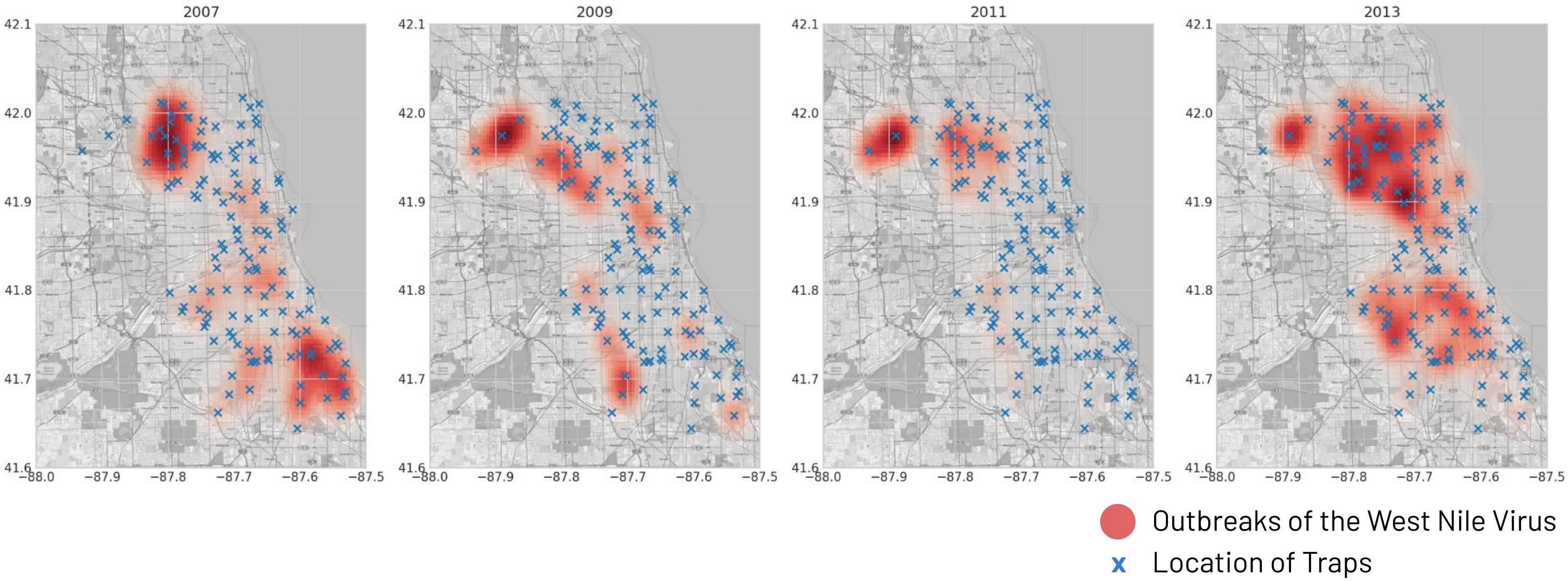
Humidity

Daylight Hours

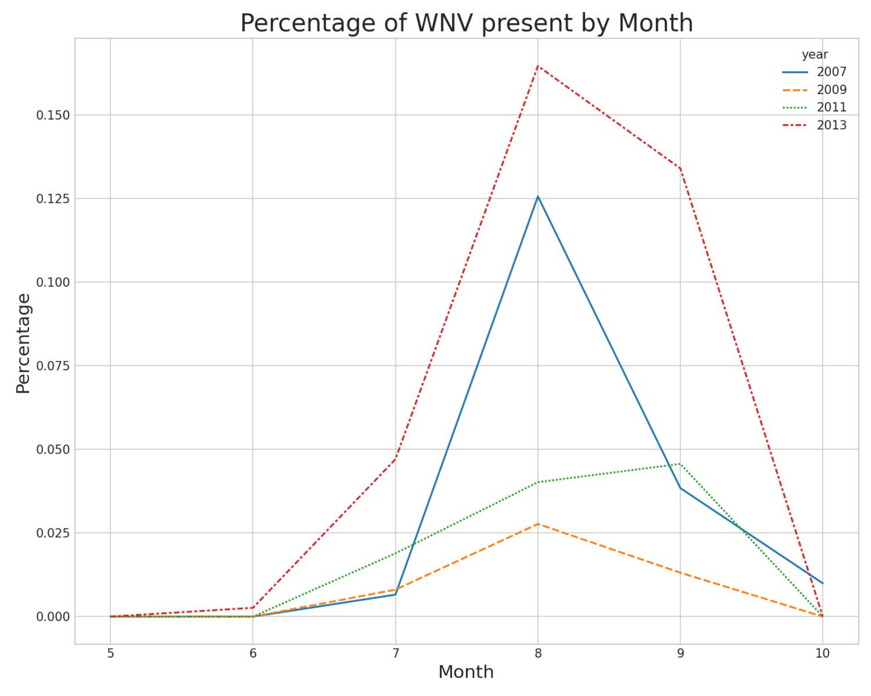
Temperature

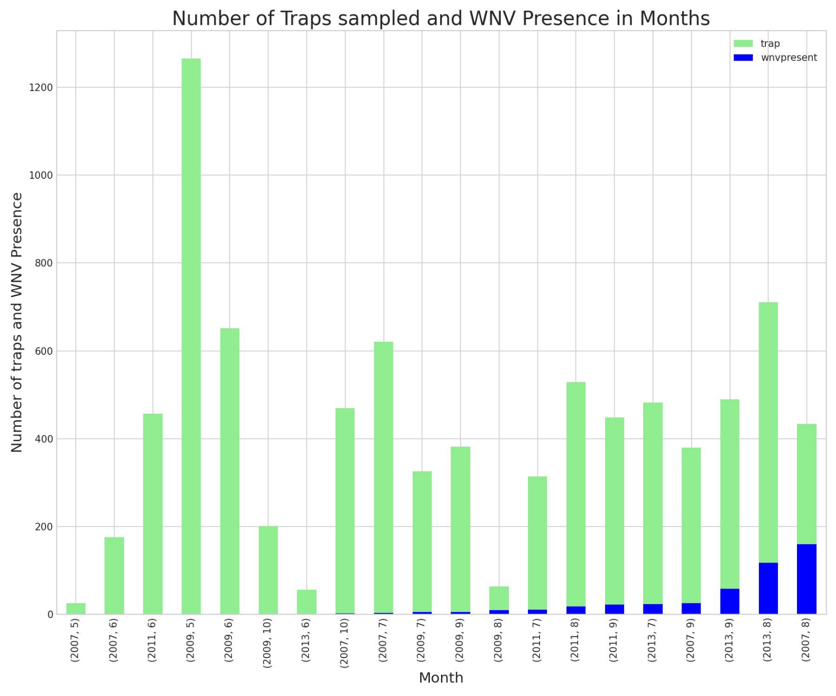
Geographical Outbreak Clusters
Around specific Neighbourhoods

Traps where West Nile Virus is Detected by Year



High Seasonality
Of WNV Outbreaks exhibited





Factors leading to an Outbreak

Number of Mosquitos

Mosquito Species

Location

Seasons

Temperature

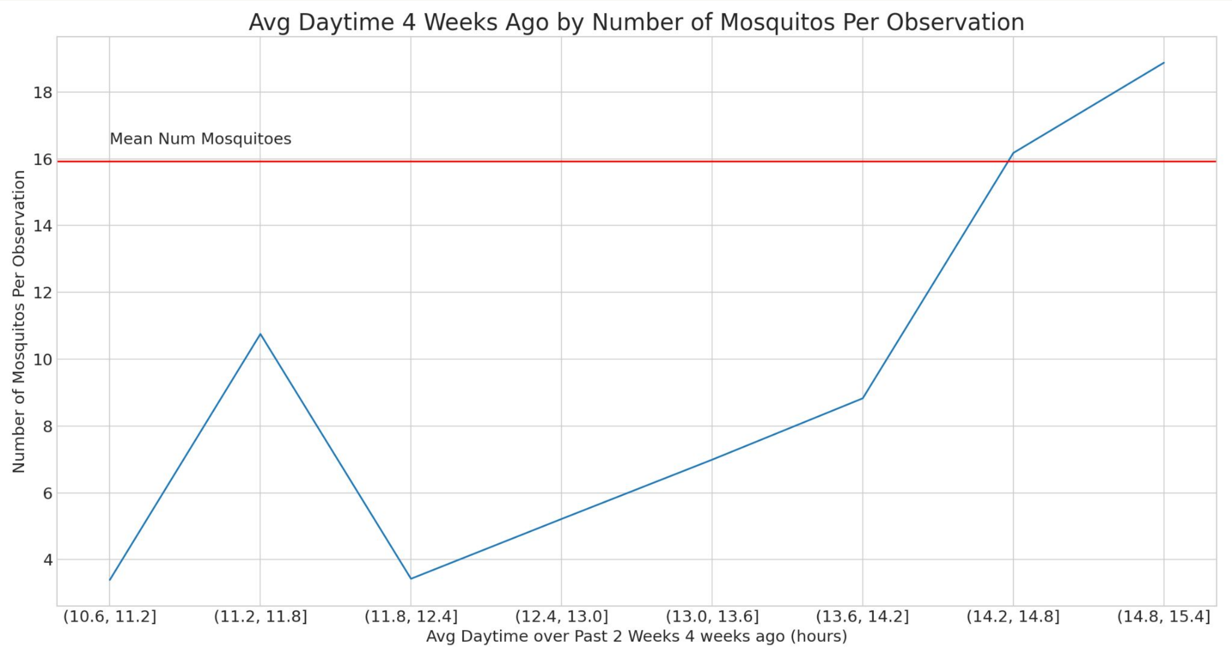
Humidity

Daylight Hours

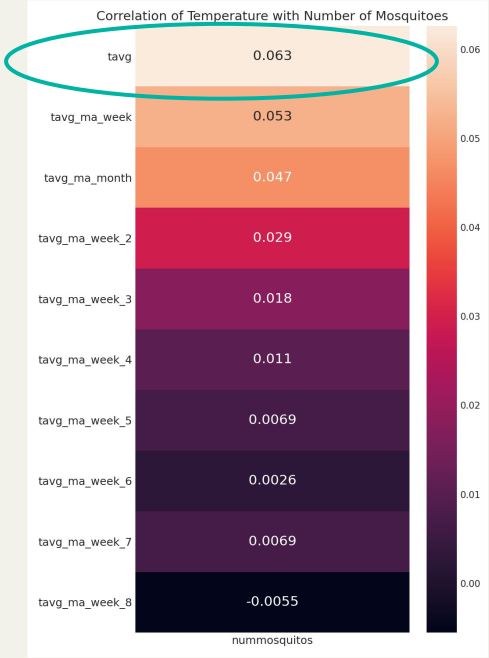
Precipitation

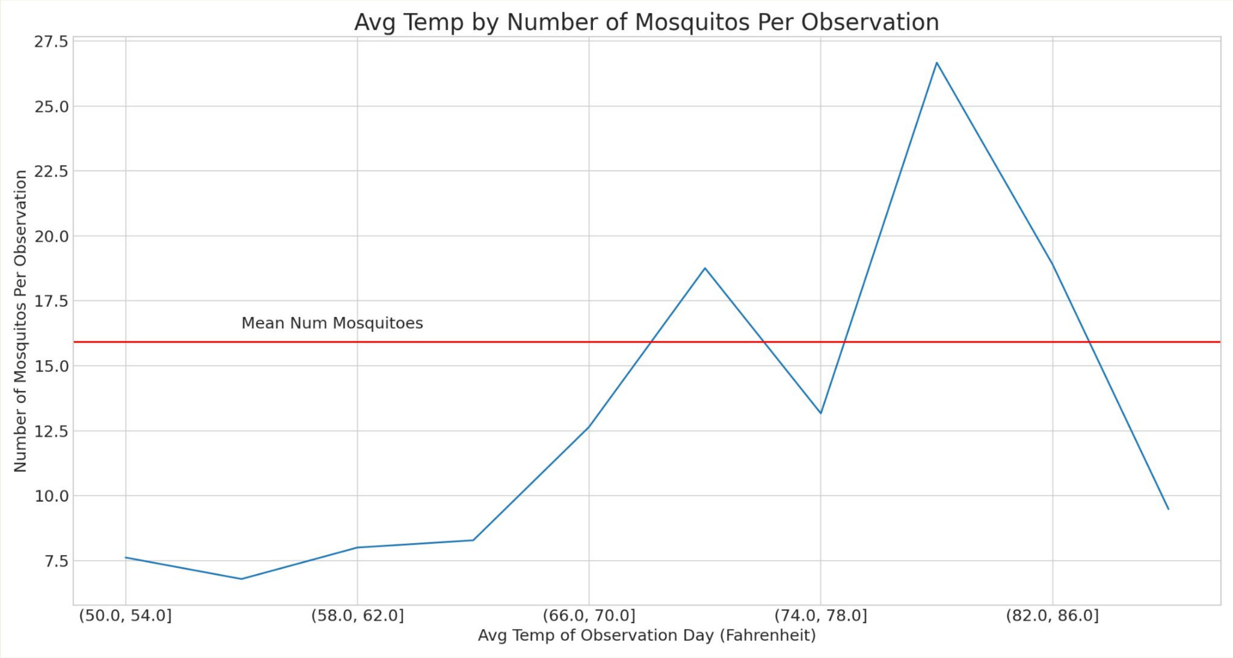
Moving Average of Daylight Hours

In relation to the no. of mosquitoes



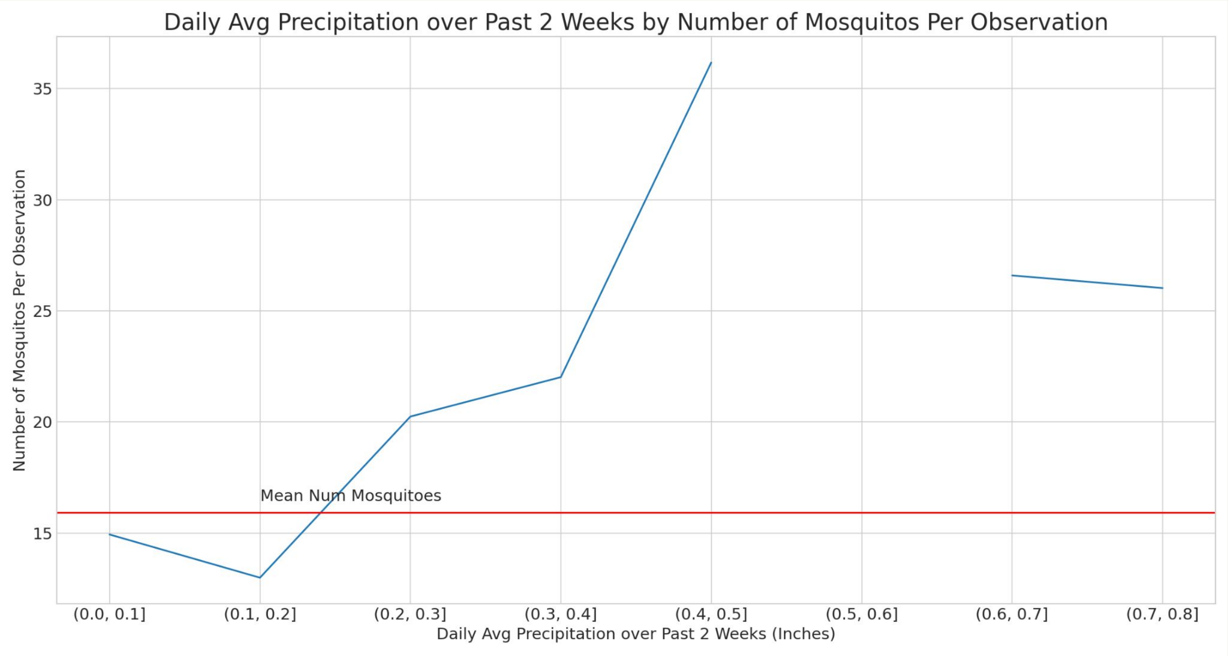
Moving Average of Temperature
In relation to the no. of mosquitoes





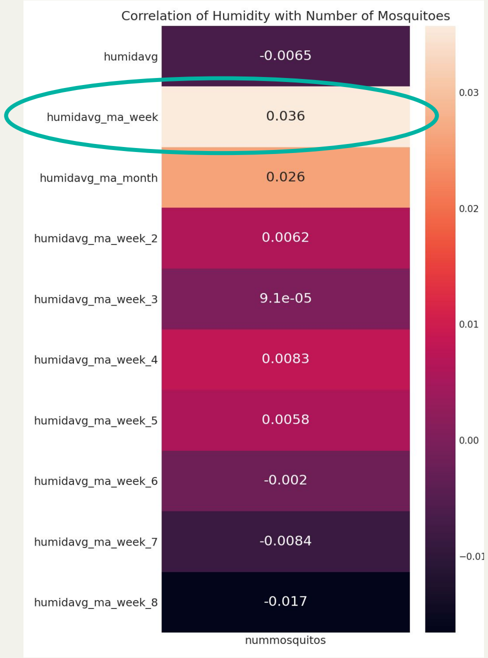
Moving Average of Precipitation

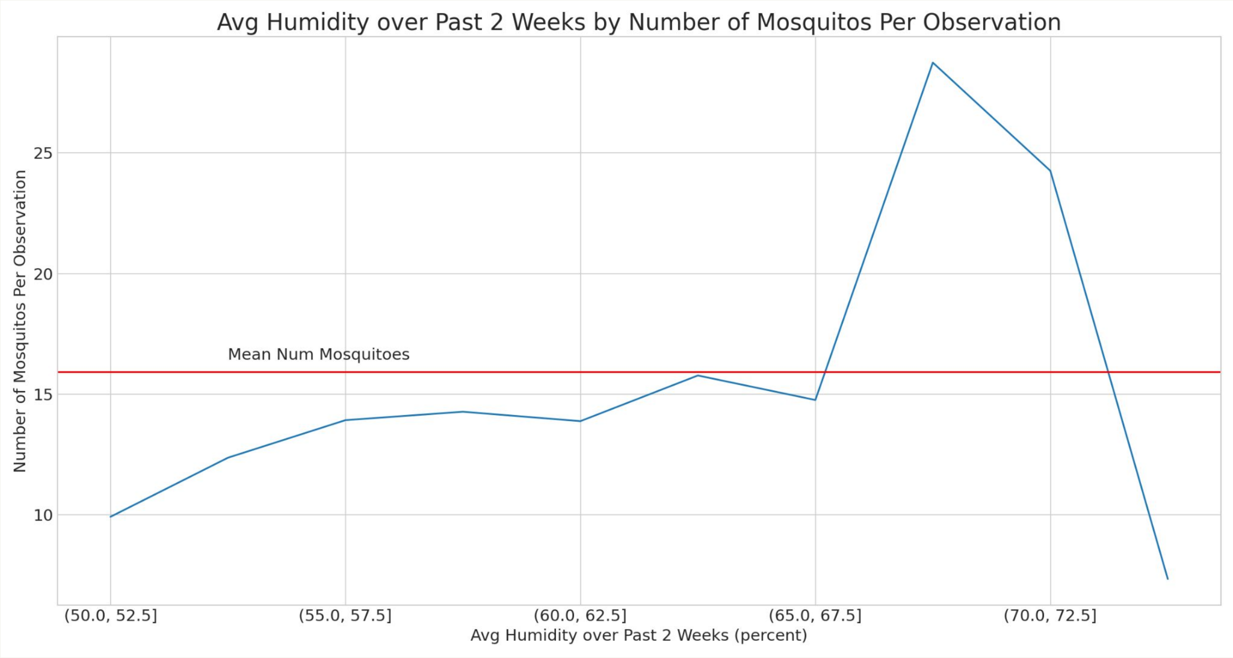
In relation to the no. of mosquitoes



Moving Average of Humidity

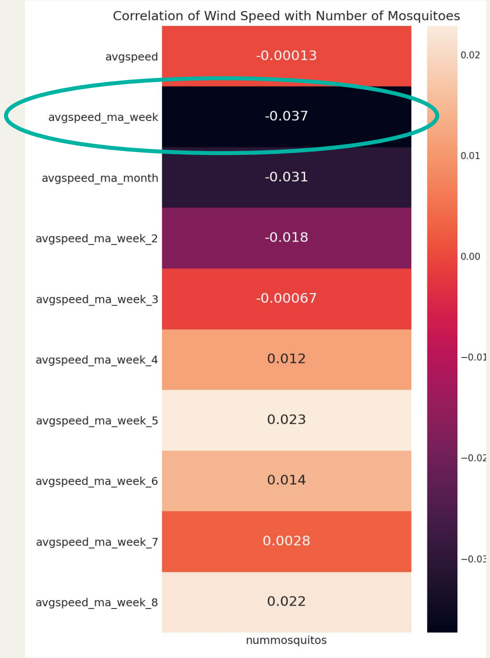
In relation to the no. of mosquitoes

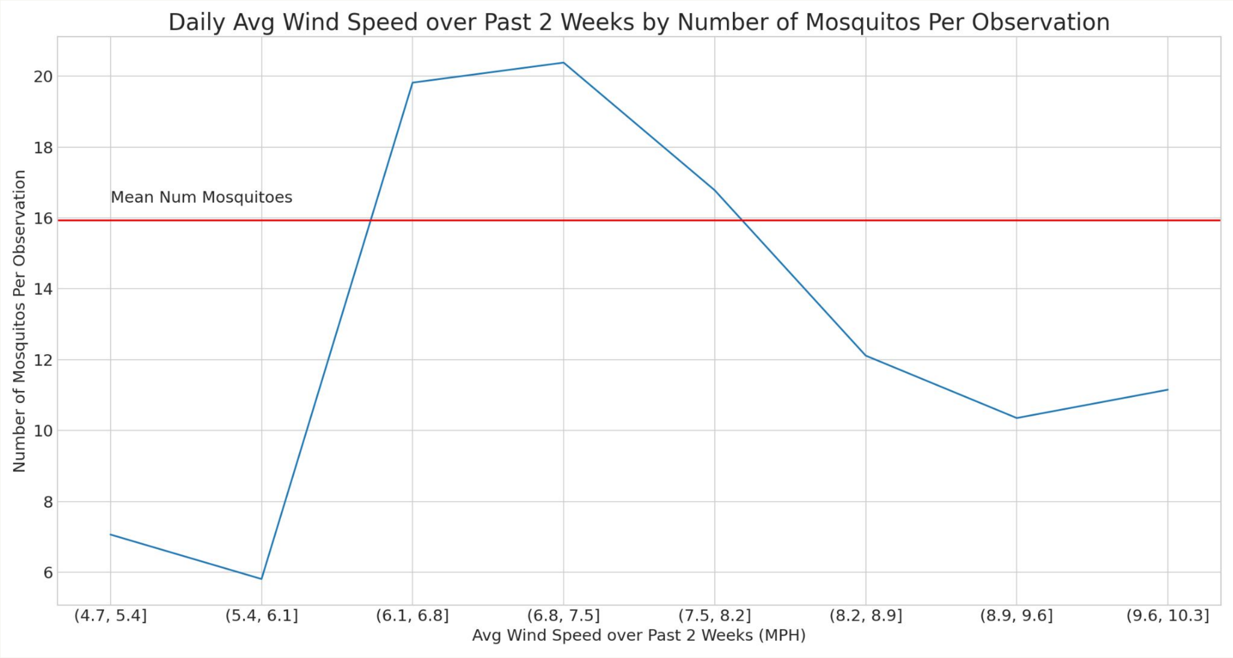




Moving Average of Wind Speed

In relation to the no. of mosquitoes



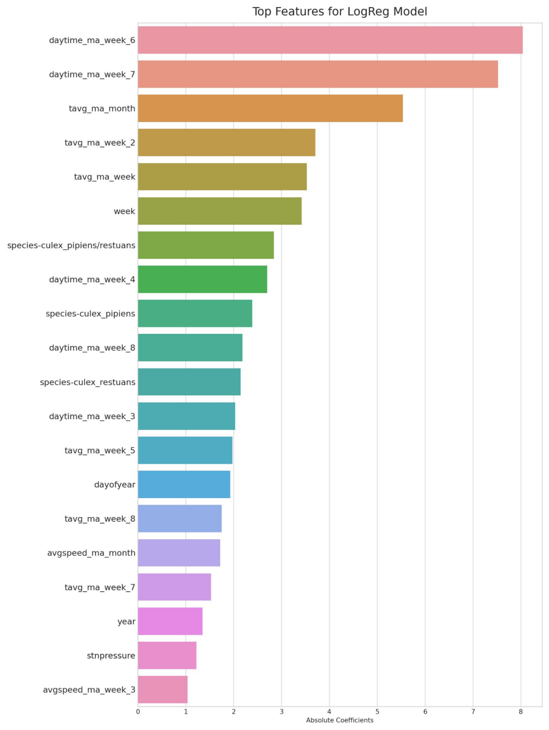


Higher Accuracy of Predictions

Led by Findings and Interpretable Model Coefficients

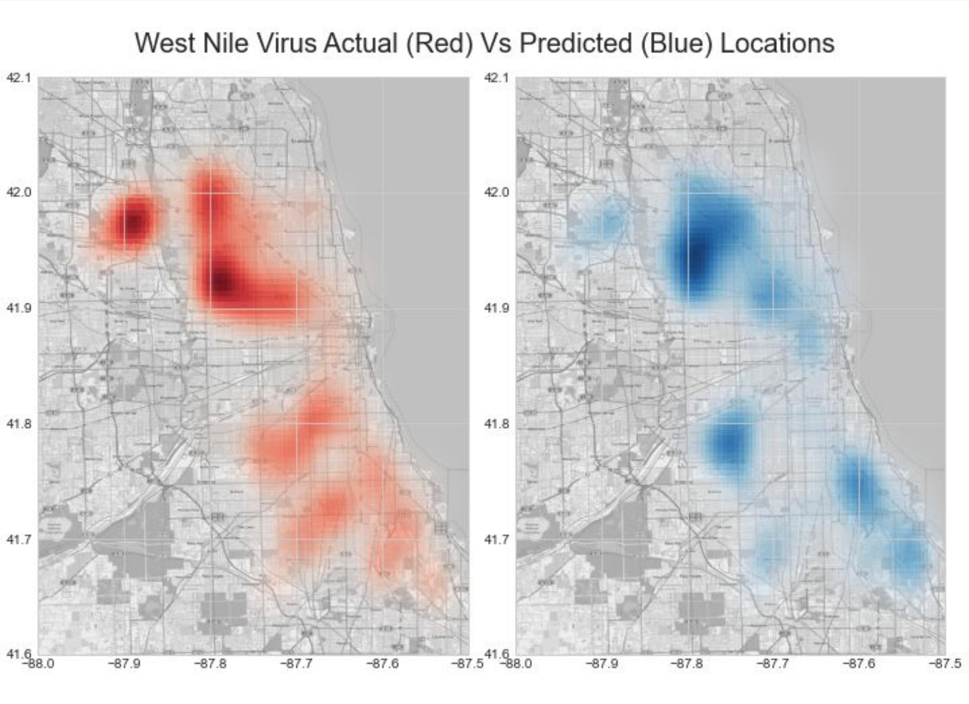
Top Features resulting in higher probabilities of the presence of WNV

1) Moving Average for Hours of Daytime
2) Moving Average for Average Temperature
3) Week of the Year
4) Species of Mosquito



Higher Accuracy of Predictions

Led by Findings and Interpretable Model Coefficients



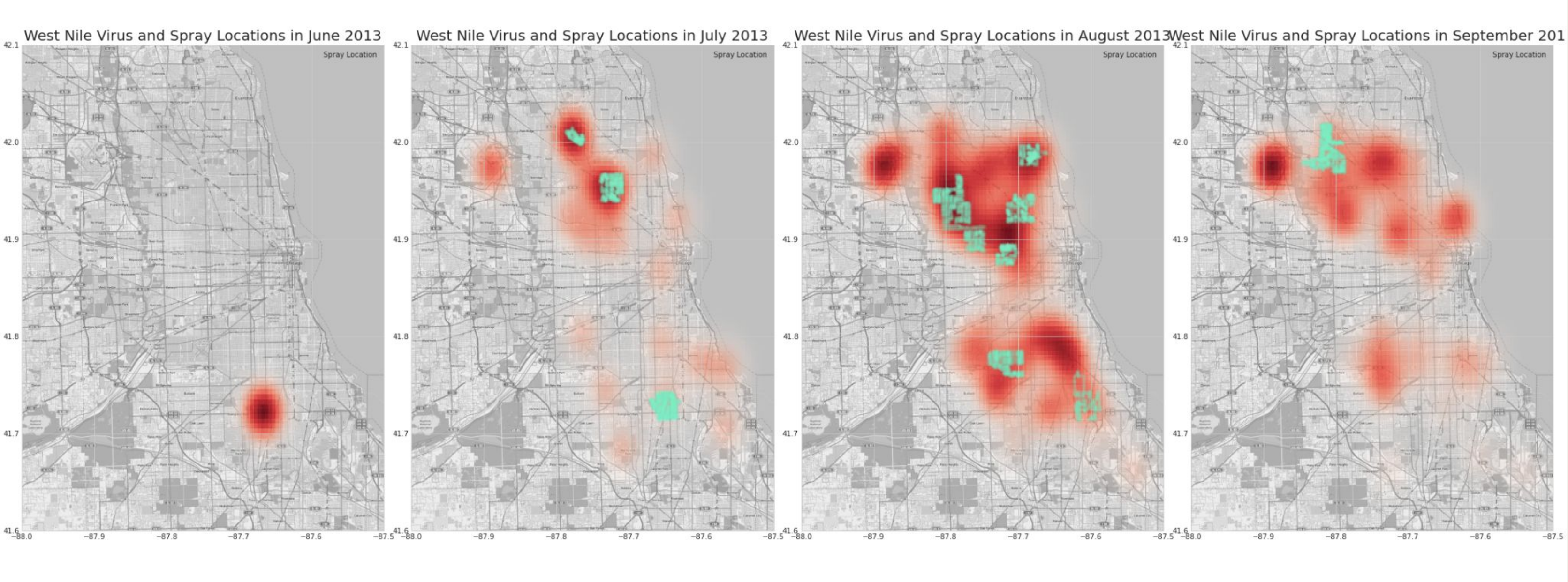
- Model is quite accurate at predicting locations of outbreaks
- Low false negative rate, at the expense of high false positive rate

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	1515 (72.2%)	490 (23.3%)
Actual Positive	**28 (1.3%)**	68 (3.2%)

Cost-Benefit Analysis

Should the Government spend more on insecticides?

- Spraying seems to be effective at controlling outbreaks within a small radius
- Spraying is reactive rather than proactive, when there is an outbreak in the area



Benefits vs Costs

Economical Cost without Spraying (Sacramento County)	
Inpatient Cost per case	$ 33,143
Outpatient Cost per case	$ 6,317
Productivity Cost per case	$ 4,775
Total Economic Cost per case (Inpatient)	$ 37,918
Total Economic Cost per case (Outpatient)	$ 11,092

Chicago Cases in 2018	
Total Cases	57
Cases Hospitalised	44
Fatalities	3

Spraying Cost Estimation		
Country	Sacramento County, California	Chicago
Area	2,574 km²	606 km²
Sprayed Area	477 km² (18%)	242 km² (~40%)
Sprayed Cost per Area	$ 1,471	$ 1,471
Total Sprayed Cost	$ 701,790	$ 356,045

Sources: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/ ; https://www.chicago.gov/content/dam/city/depts/cdph/food_env/general/West_Nile_Virus/WNV_2018databrief_FINALJan102019.pdf

Spraying is economically cost effective and should be carried out

- **Economic benefits outweigh the costs** from spraying, even if not all cases are prevented

Hospitalisation and Productivity Benefits			
	Benefits	**2018 Cases**	**Total Benefits**
Inpatient	$ 37,918	44	$1,668,392
Outpatient	$ 11,092	13	$144,196
Total			$1.8 mil

>

Spraying Cost
$ 356,045

Table of contents

Conclusions & Recommendation

What are the key factors leading to an outbreak?

1. Seasonality, as driven by weather
2. Presence of particular mosquito species
3. Location

Backtesting shows predictive models built using these variables do better than baseline

Is it cost-effective to spray the city to prevent outbreaks?

Yes, economic benefits outweigh the costs.

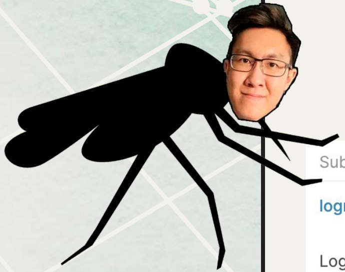
Spraying insecticides in Chicago seems to be reactive currently.

Recommendation

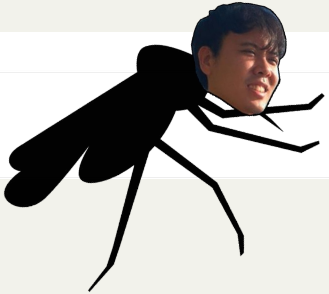
Chicago should build a predictive model using lagged weather, trap and location data to determine which locations it should spray, in order to better combat outbreaks

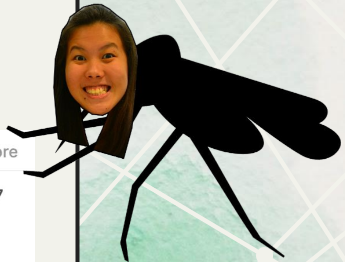
Thanks!

Do you have any questions?



Submission and Description	Private Score	Public Score
logreg.csv Logistic Regression Model	0.73270	0.73947





CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**