West Nile Virus in Chicago

By DSI 14 Group 6

Anastasiya | Bryan | Junyuan | Roshan

Introduction

West Nile virus (WNV) is the leading cause of mosquito-borne disease in the United States

We aim to assess the cause of WNV spread in Chicago, IL and predict hotspots to conduct spraying

01

Problem Statement

02

Exploratory Data Analysis

03

Feature Engineering

04

Model Evaluation

05

Conclusion and Recommendations

01

Problem Statement

Key Matters

Weather

Will weather conditions affect the presence of mosquitos?

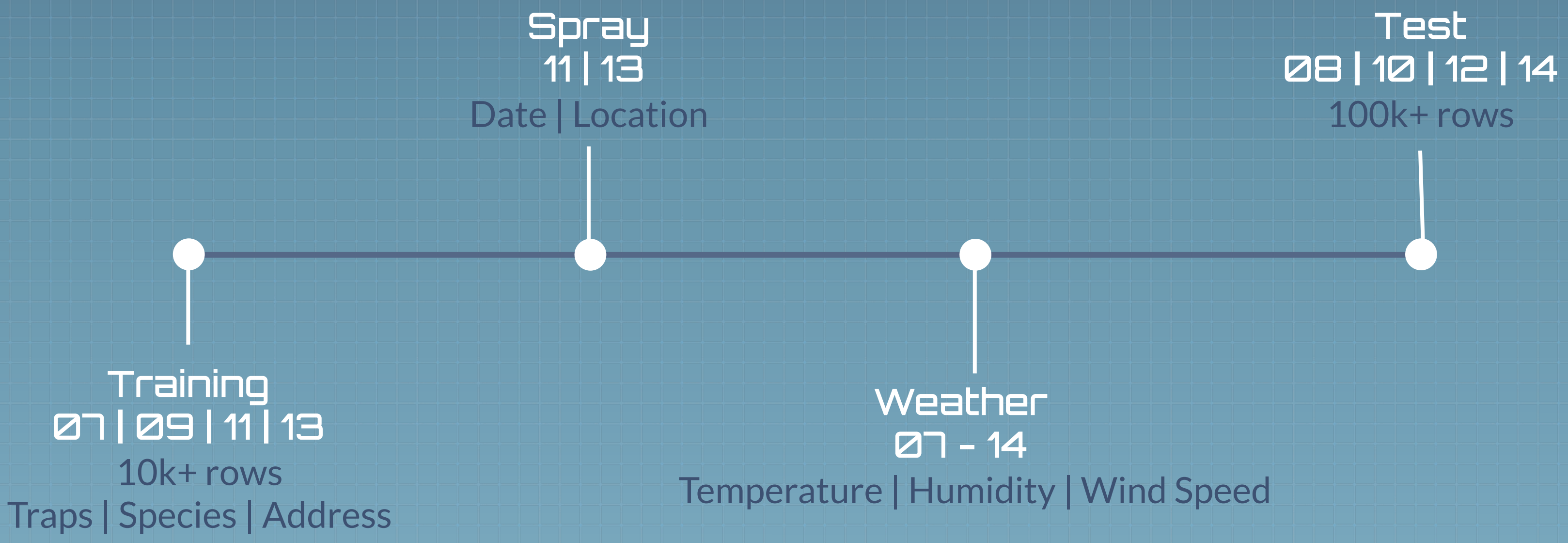
Geography

Are mosquito breeding areas affected by geolocation?

Species

Does a particular species carry the West Nile Virus?

Datasets



Training
07 | 09 | 11 | 13
10k+ rows
Traps | Species | Address

Spray
11 | 13
Date | Location

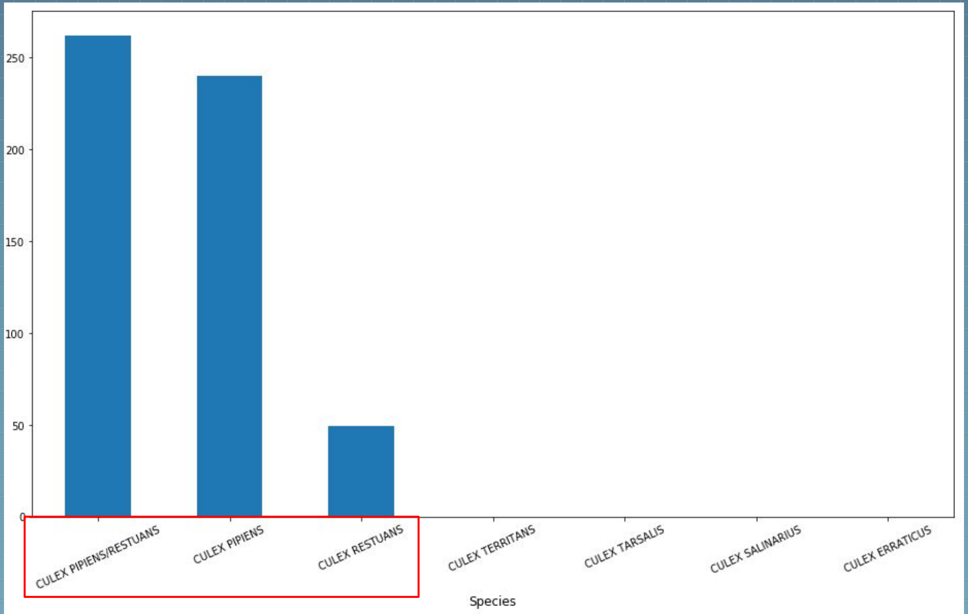
Weather
07 - 14
Temperature | Humidity | Wind Speed

Test
08 | 10 | 12 | 14
100k+ rows

02

Exploratory Data Analysis

Counts of WNV present against Species

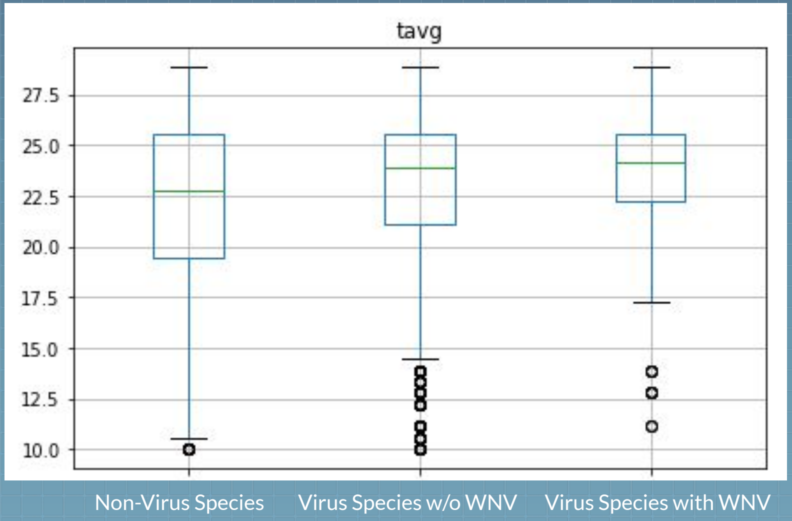


PIPIENS and RESTUAN species carry the WNV

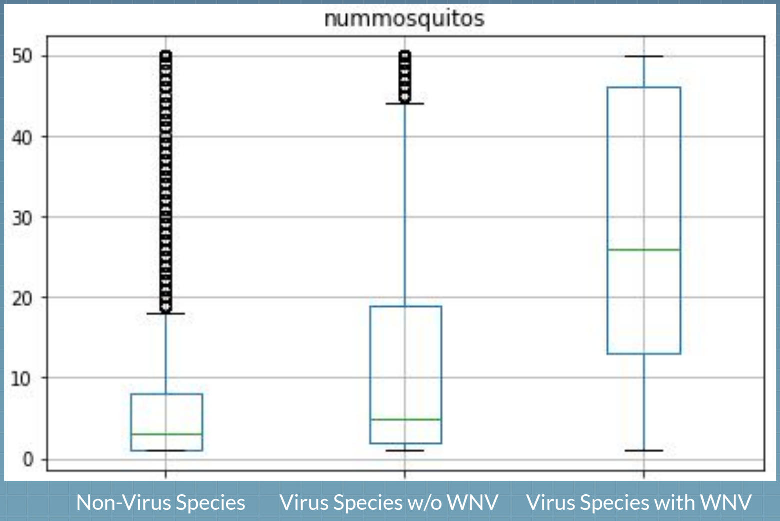
Based on this information,
a sub category is being created

- Non-virus species
- Virus species w/o WNV
- Virus species with WNV

Comparing other features against new subcategory

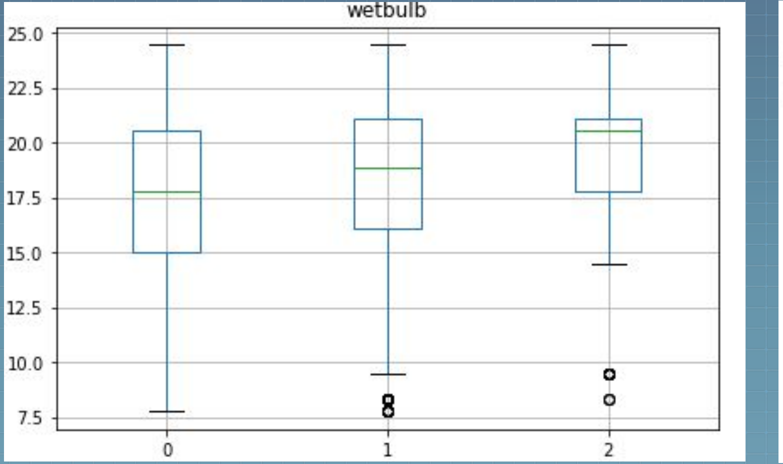


HIGH temperature, **HIGH** mosquitos with WNV

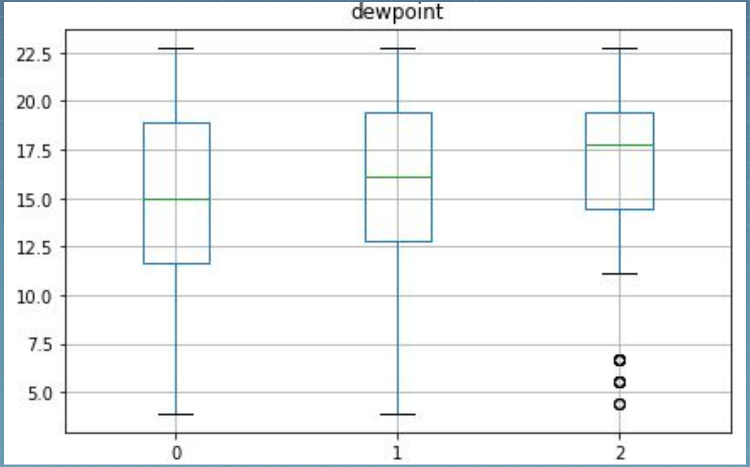


MORE mosquitos, **MORE** WNV carrying mosquitos

Comparing other features against new subcategory



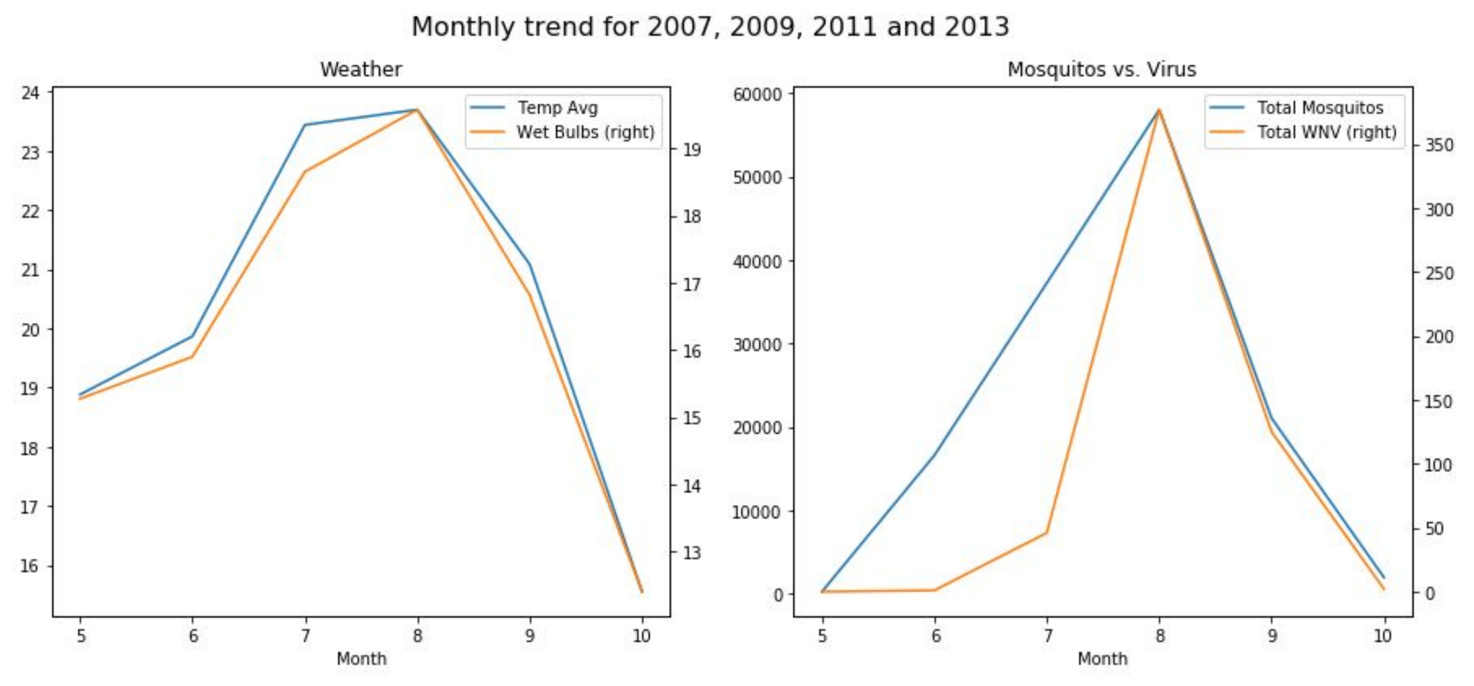
Non-Virus Species Virus Species w/o WNV Virus Species with WNV



Non-Virus Species Virus Species w/o WNV Virus Species with WNV

HIGH humidity, **HIGH** mosquitos with WNV

Monthly Trend



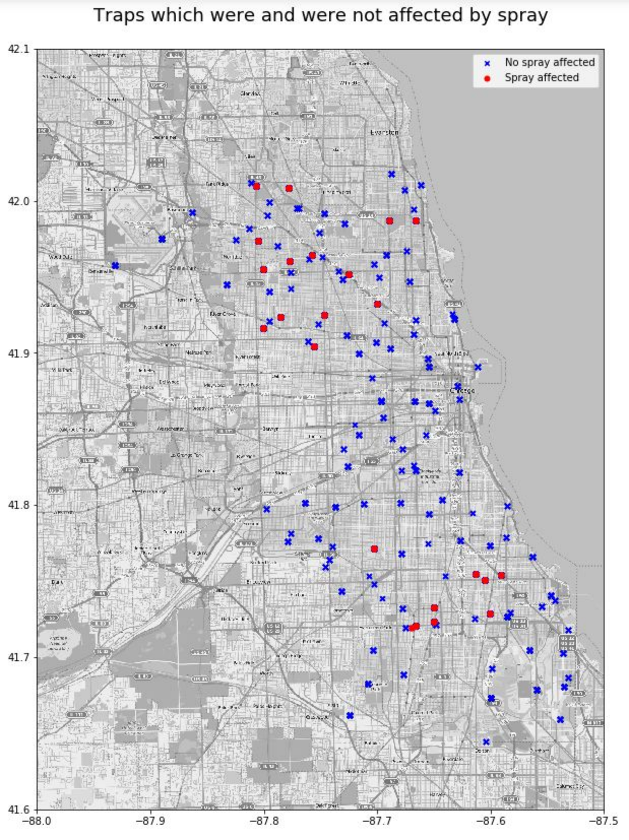
- **JULY** and **AUGUST** huge spike in temperature and mosquitos
- **WNV** develops as more mosquitos increase

Effect of Sprays Conducted

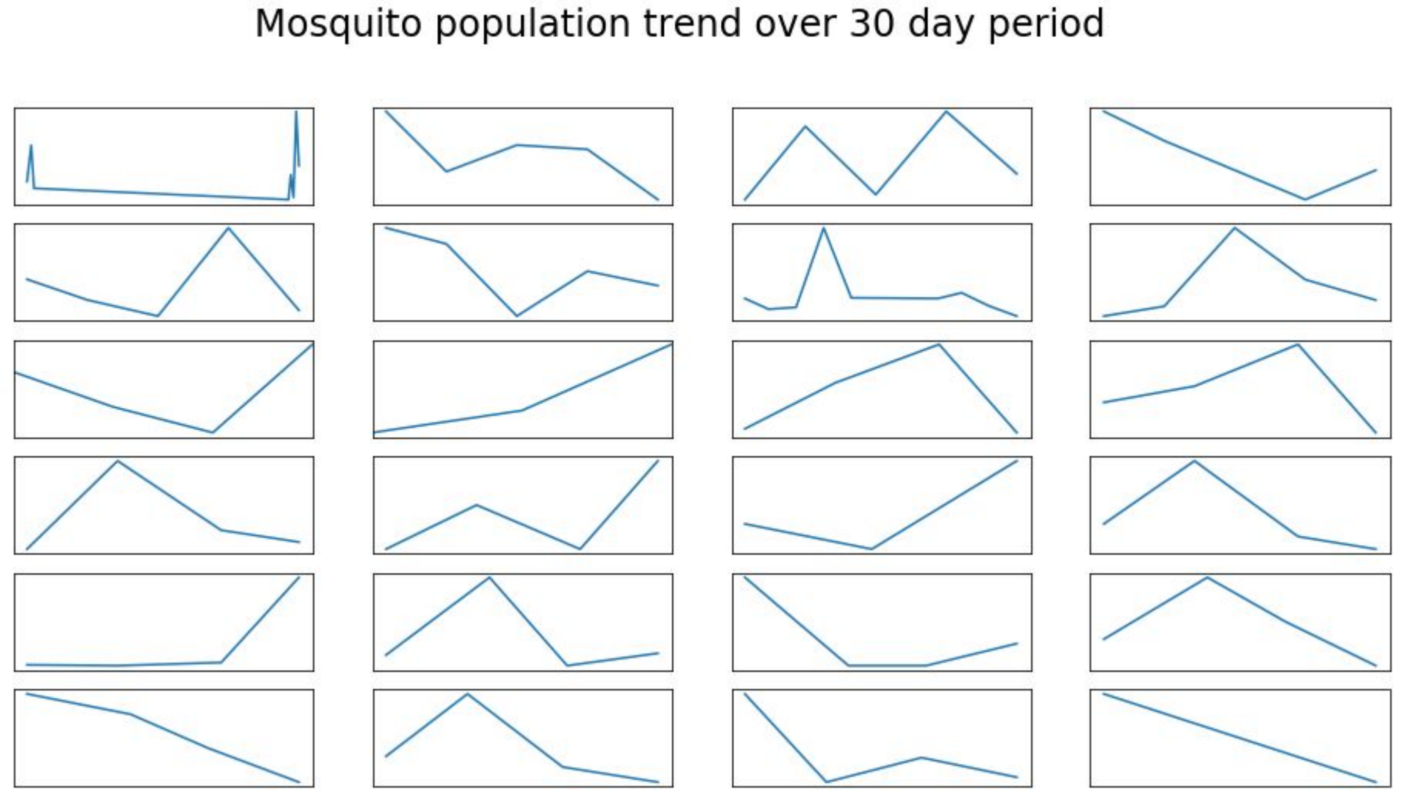
Spray was conducted in 2011 and 2013.

Number of traps affected by the spray is significantly low.

Notice the central region of the city was not affected by the spray unlike the northern and southern region.



Impact of spraying on mosquito populations in a 30 day period



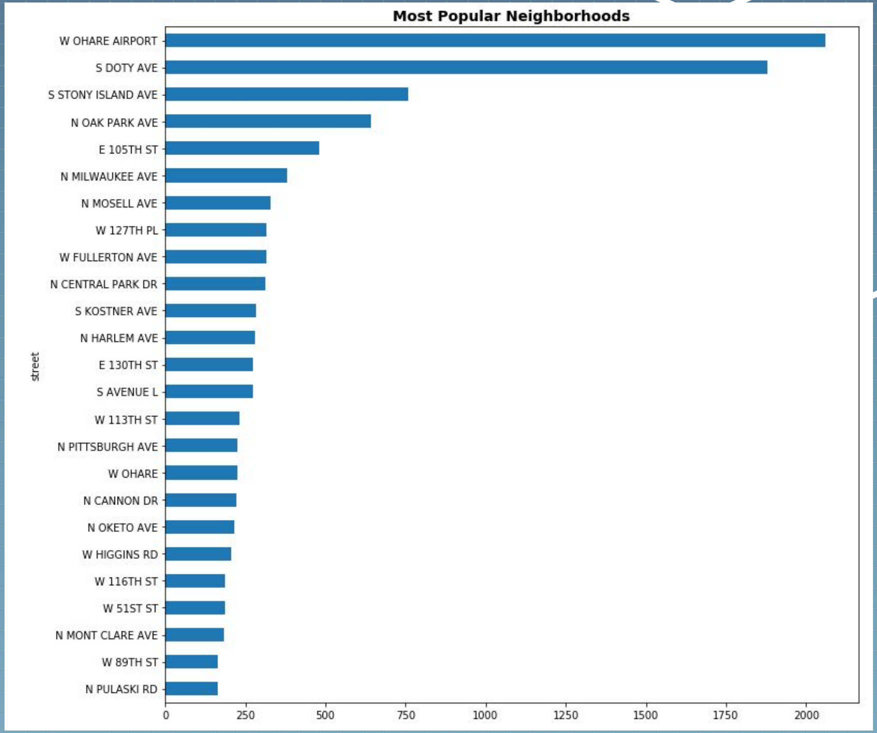
Each plot represent a trap which is affected by spraying activity.

The first point on each plot is Day 0 when spraying was conducted and the trap is within the region of influence.

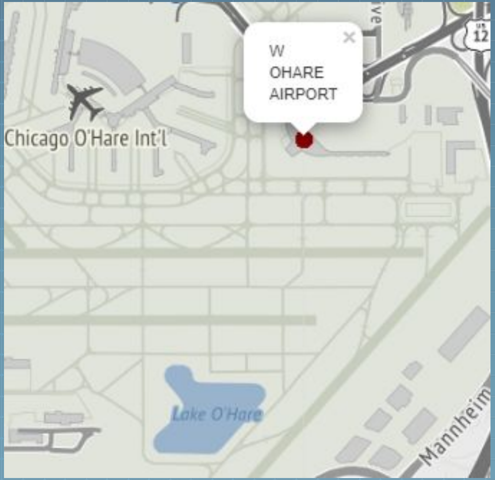
Neighborhoods

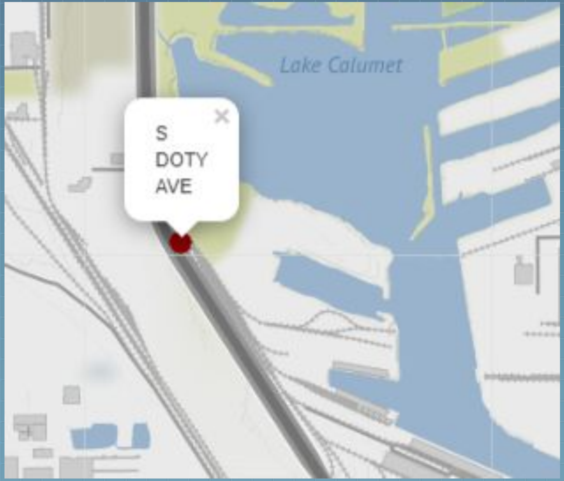
Top 5 Neighborhoods with high number of West Nile Virus mosquitos present

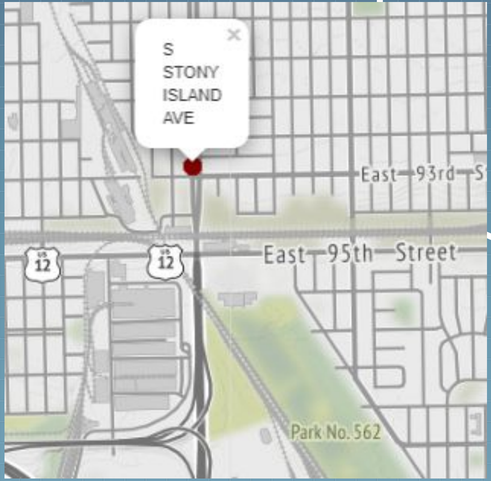
- W O’hare Airport
- S Doty Avenue
- S Stony Island Avenue
- N Oak Park Ave
- E 105th Street



Neighborhood - Surroundings







With the present of parks and lakes (stale water), we see an increase in the number of West Nile Virus mosquitoes present.

03

Feature Engineering

Weather Station

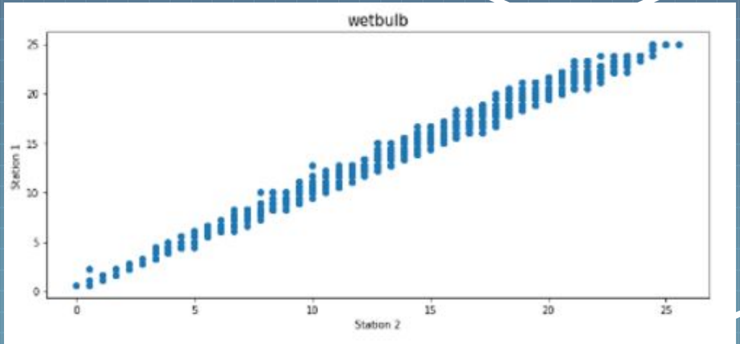
Station 1 - Chicago O'hare International Airport

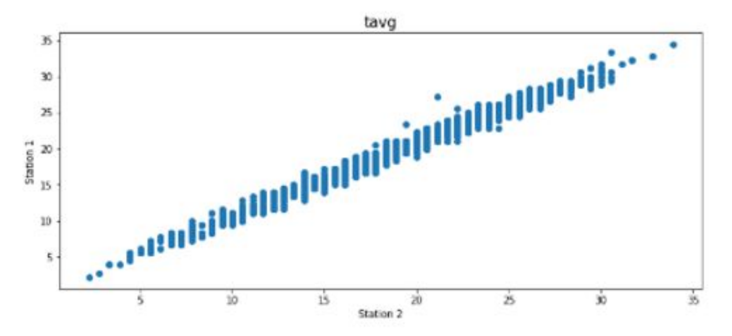
(Lat: 41.995, Lon: -87.933)

Station 2 - Chicago Midway International Airport

(Lat: 41.786, Lon: -87.752)

Since Station 1 and 2 datasets are very similar, will drop Station 2 as it has a lot more missing values than Station 1 (Depth, Snowfall, Departure, Sunrise and Sunset)





Humidity and Avg Temperature

Humidity

According to an article by the National Centre for Biotechnology Information, Humidity also tends to have a positive correlation with the population of West Nile Virus mosquitos.

Hence, we have added the humidity column to the weather data set by using difference in values between Temperature Average and Dewpoint.

Average Temperature

External research shows that mosquitoes breeding activity is closely related to weather conditions two weeks prior.

Hence, we also have added the add_tavg_14day column indicating the 14 day delay for weather data.

Species and Spray

Species

Based on the EDA above, we know the Cullex Pipiens, Cullex Restuans or Cullex Pipiens|Restuans mosquitoes are responsible for carrying the West Nile virus.
Created a iswnvspecies column to identify if the mosquitoes caught in the trap do belong to the West Nile Species group.

Spray

The spray used by city officials will have an lasting impact of 30 days.
In order to evaluate the impact on the mosquitoes population when the spray has been conducted we also set up the is_spray column with a 30 day range.

04

Model Selection and Validation

Baseline Model

0.948

Each data entry has a 5.2% chance of being positive with West Nile Virus.

- This also indicates the imbalanced nature of our dataset.
- The number suggests a high chance of getting False negatives.
- To combat this, we will perform upsampling on our train dataset.

Resampling

- Undersampling Majority Class
- Oversampling Minority Class
- Over-sampling followed by under-sampling

Modeling

(LogisticRegression, KNeighborsClassifier, SVC, DecisionTreeClassifier, RandomForestClassifier)

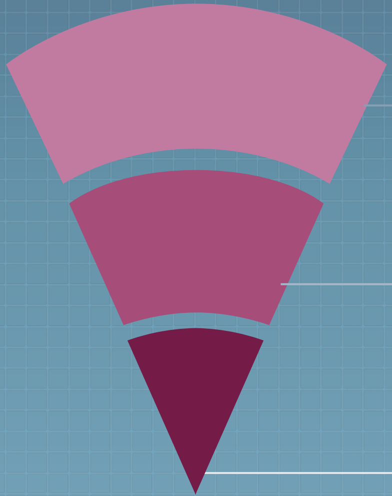
Evaluation

Train / Test accuracy, ROC scores

Resampling Metrics

Models\Resampling	undersampling	oversampling	over-undersampling
LogisticRegression	train 0.72 / test 0.61	train 0.71 / test 0.64	train 0.70 / test 0.63
	ROC score 0.753	ROC score 0.755	ROC score 0.756
KNeighborsClassifier	train 0.81 / test 0.72	train 0.92 / test 0.91	train 0.91 / test 0.91
	ROC score 0.775	ROC score 0.911	ROC score 0.906
	False Negatives: 38	False Negatives: 49	False Negatives: 55
SVC	train 0.79 / test 0.70	train 0.84 / test 0.78	train 0.84 / test 0.
DecisionTreeClassifier	train 0.95 / test 0.65	train 0.96 / test 0.93	train 0.96 / test 0.93
	ROC score 0.684	ROC score 0.961	ROC score 0.959
RandomForestClassifier	train 0.95 / test 0.68	train 0.96 / test 0.93	train 0.96 / test 0.93
	ROC score 0.782	ROC score 0.958	ROC score 0.956

Final model



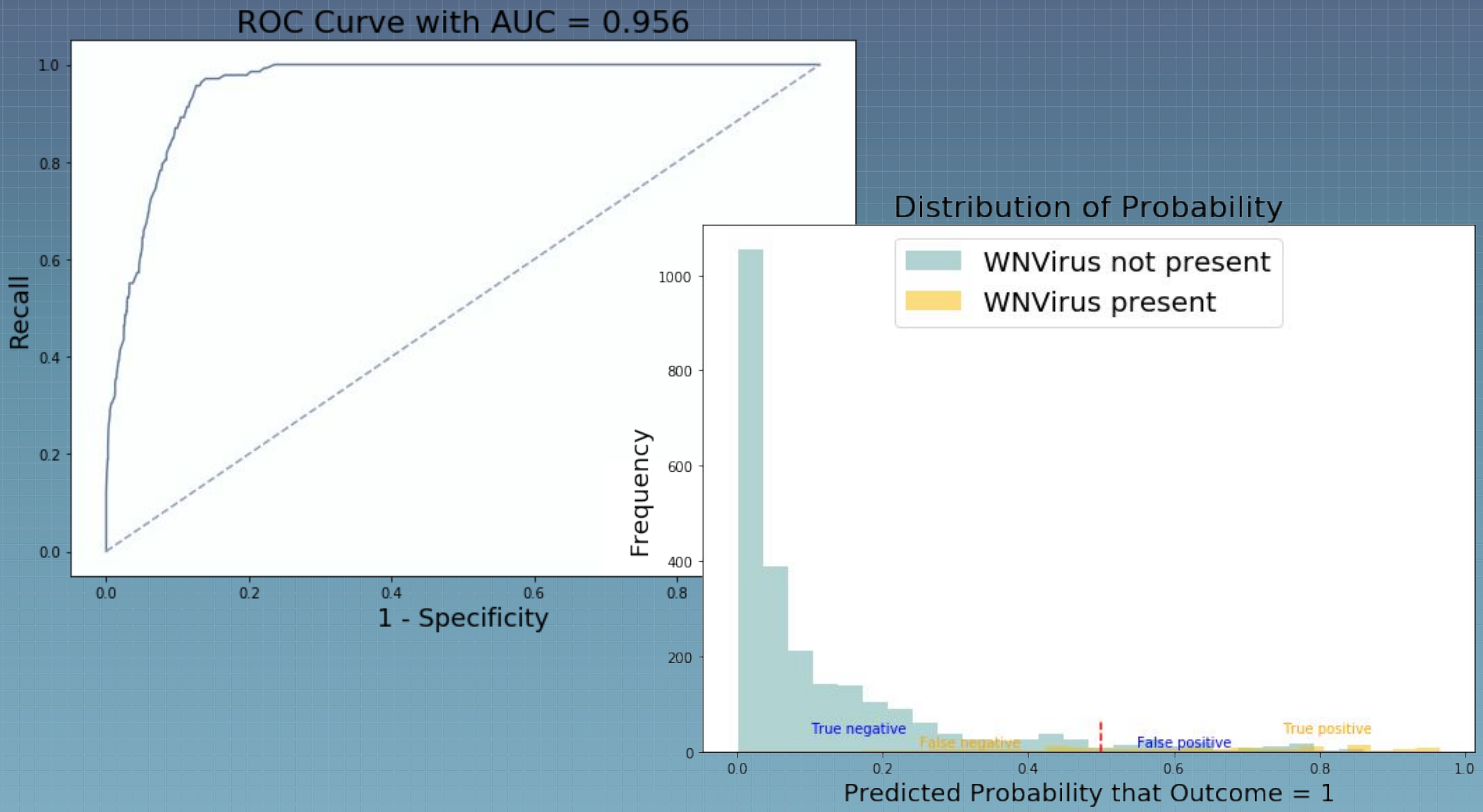
Simple Models with GridSearch:
LogisticRegression, KNeighborsClassifier,
SVC, DecisionTreeClassifier, RandomForestClassifier

Ensemble Models with GridSearch:
GradientBoostingClassifier, BaggingClassifier,
AdaBoostClassifier

VotingClassifier:
Simple + Ensemble

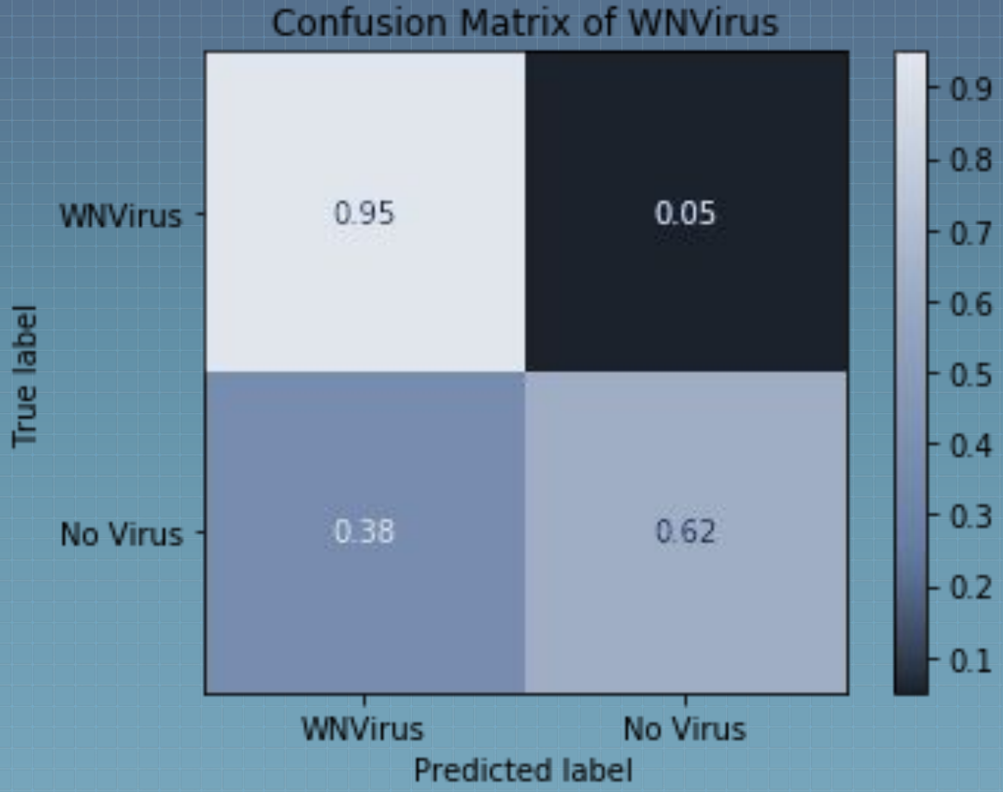
Voting Classifier metrics

Train accuracy	0.957
Test accuracy	0.942
True Negatives	2395
False Positives	94
False Negatives	58
True Positives	80
Precision	0.46
Recall	0.58
F1	0.513



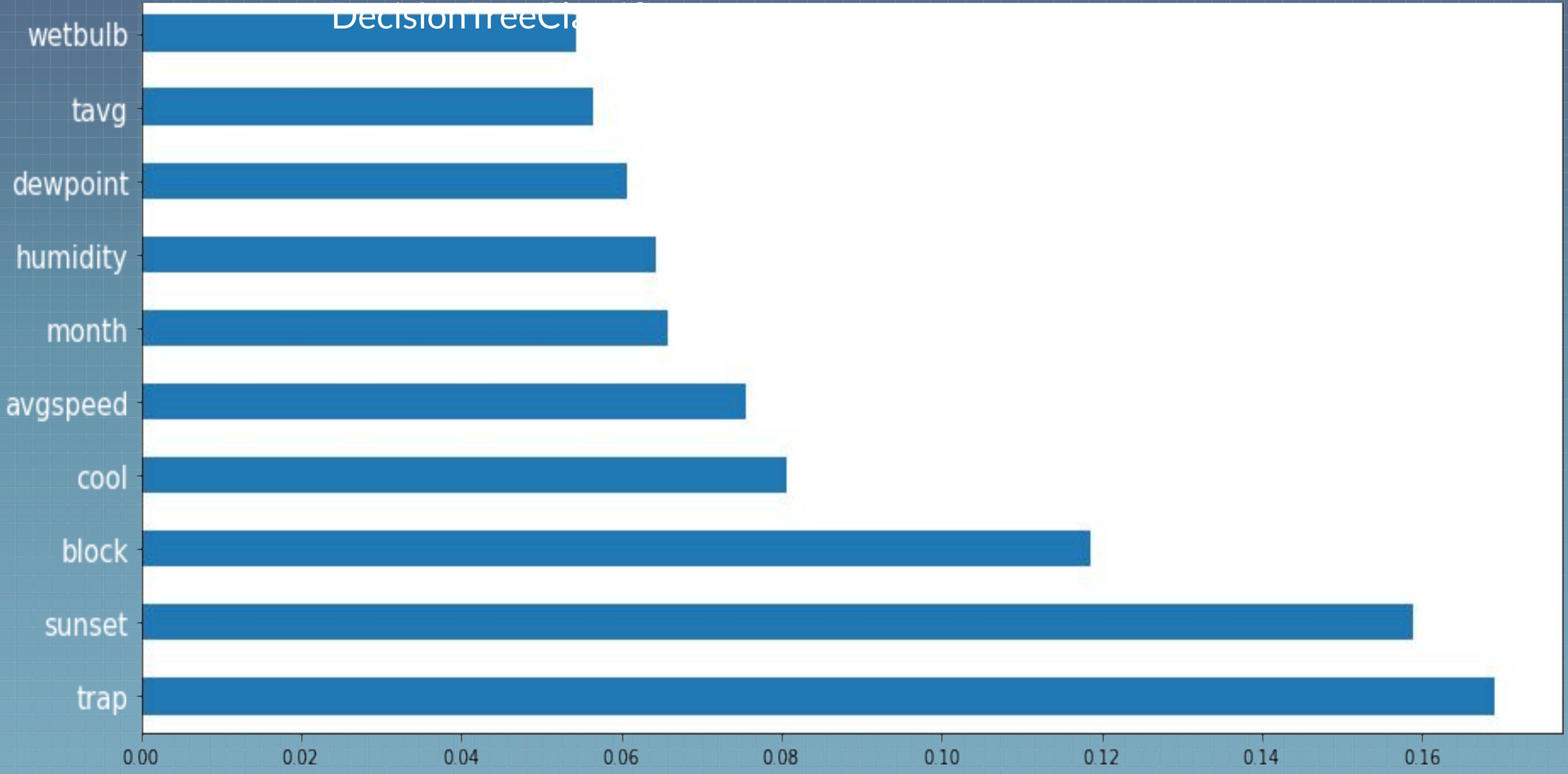
ROC Curve with AUC = 0.956

Distribution of Probability



Kaggle submission scores	
Private Score	**Public Score**
0.65931	0.65337

10 Most important features of RandomForest and DecisionTreeCla



05

Conclusions and Recommendations

Conclusion

- Out of the various approaches, the Voting Classifier comprising of a number of selected models was found to perform the best.
- The main predictors are:
  a. Location
  b. Time
  c. Weather

Recommendations

- Resolve data disparity in our data
    a. A lot less training data compared to test data.
    b. Test data lacking nummosquito column
    c. No traps in test data impacted by spraying activity

Recommendations

- With location being a strong predictor, additional geospatial information, i.e. human density or city zoning information, can be helpful.

- More in depth study on the spraying methodologies and techniques.

Thanks!

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