West Nile Virus Prediction

City of Chicago

Jonathan Jacobs
Springboard Data Science Career Track

OUTLINE

Problem Statement

Data: Sourcing and Cleaning

Exploratory Data Analysis

Machine Learning Analysis

Conclusions

Problem Statement

PROBLEM STATEMENT

40,000

Catch basins are treated with larvicide each year by the Chicago Department of Public Health

83

Traps are monitored and maintained

PROBLEM STATEMENT

6

Yet there are still cases of WNV each year. In the summer of 2020 6 cases were reported.

PROBLEM STATEMENT

Predicting the presence of WNV carrying mosquitos can help **drive infections down to zero within two years.**

Data: Sourcing and Cleaning

DATA: SOURCING

WEATHER

MOSQUITO AND SPRAY DATA

CDPH
Chicago Department of Public Health

Data for this project comes from the 2015 kaggle competition sponsored by the Robert Wood Johnson Foundation.
Data for the competition is provided by the Chicago Department of Public Health.

DATA: CLEANING

WEATHER

1472 unique dates recorded

Two weather stations

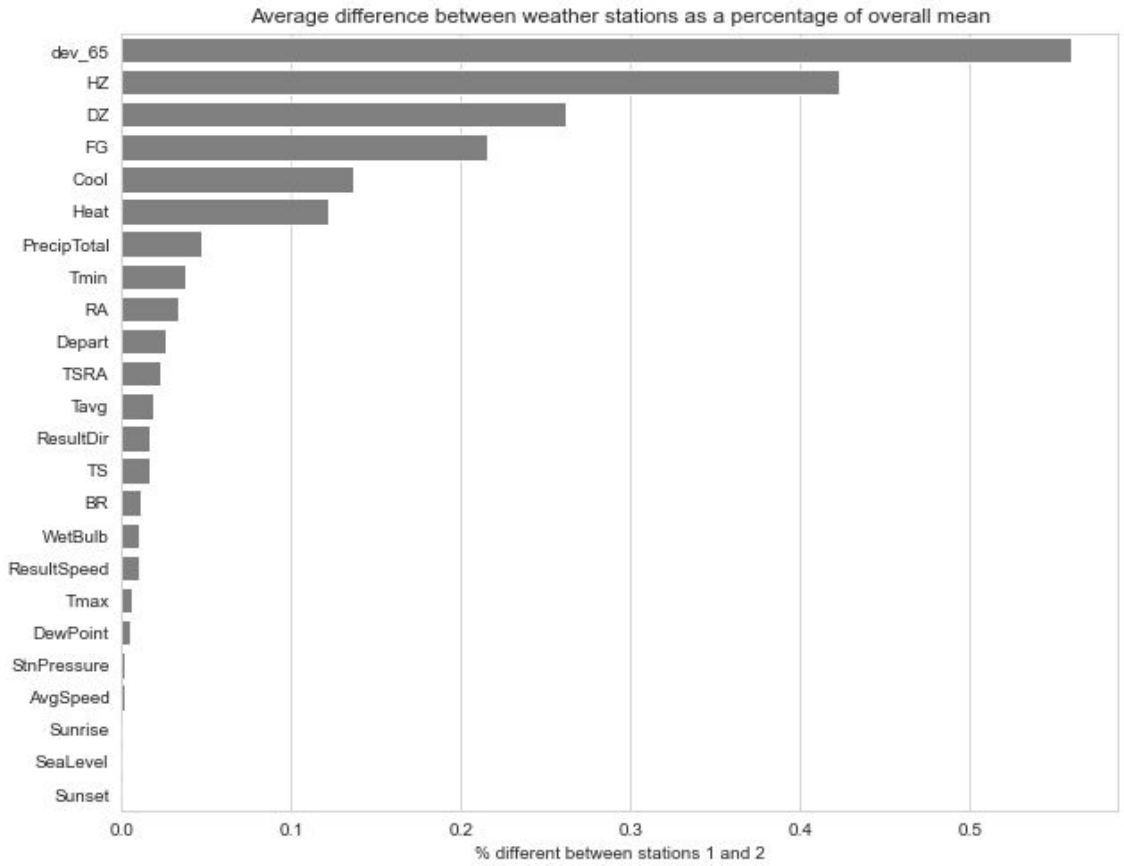
- 20 features recorded at station 1
- 19 features recorded at station 2

Few missing entries, with no structural or serious data issues found

DATA: CLEANING

HANDLING TWO WEATHER STATIONS

The two weather stations are averaged. This chart shows the average difference between stations 1 and 2 as a percentage of the overall mean. Although statistically different, the features are not practically different.



DATA: CLEANING

MOSQUITO DATA

10506 unique observations across 83 traps

- 5 features
- Target feature: WNV Present

No missing entries

DATA: CLEANING

SPRAY DATA

14835 unique spray observations

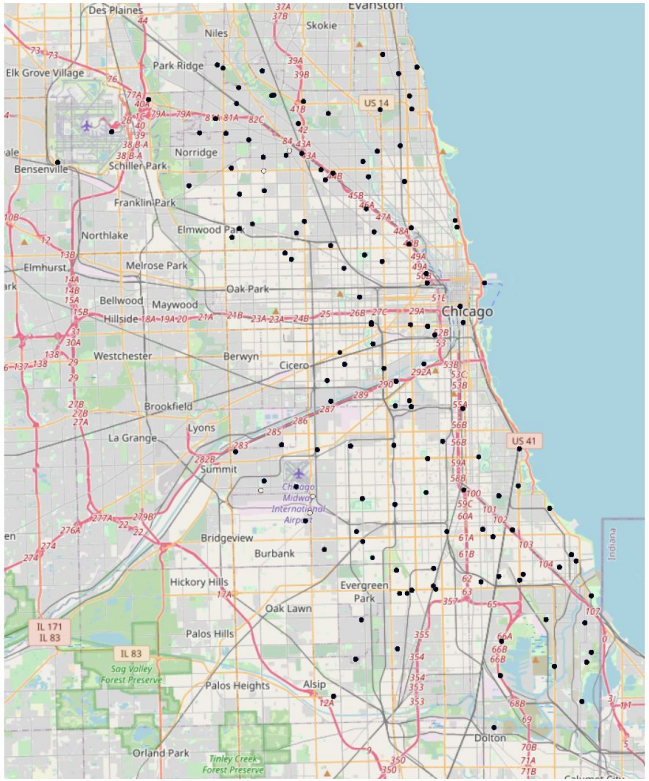
- Date, time, and location of sprays

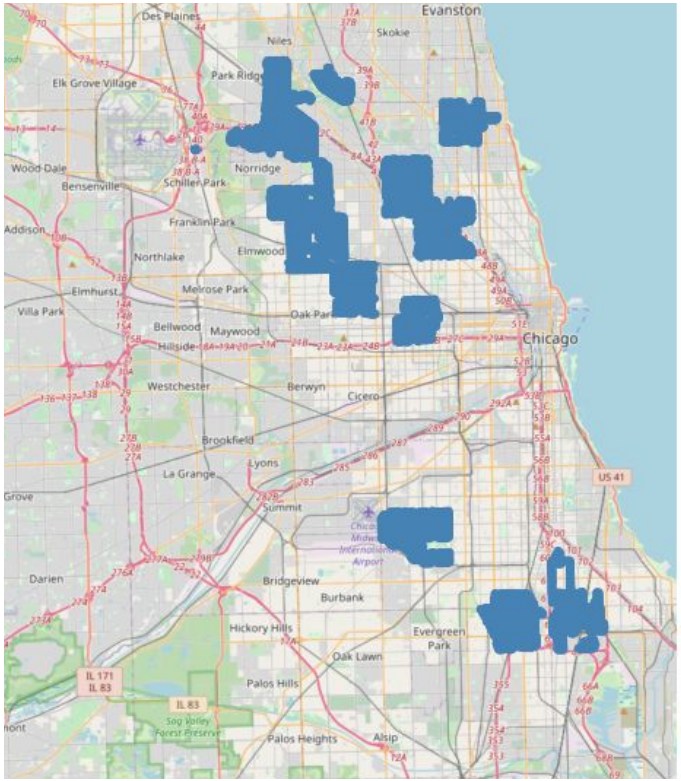
Few missing entries

UNITING SPRAY AND MOSQUITO DATA

Left: Trap locations

Right: Spray locations

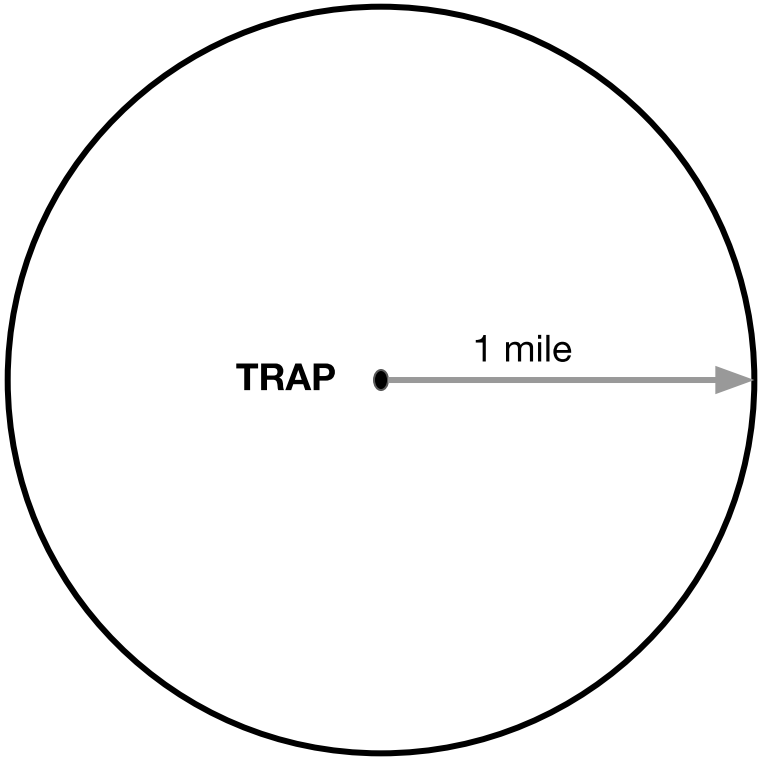




UNITING SPRAY AND MOSQUITO DATA

Spray data was incorporated into the Mosquito trap data set through distance and time proximity. A ‘1’ is recorded if a spray occurred within one mile of a trap on a particular day.

One mile is chosen because it is the furthest distance a mosquito of the correct species can travel in one day.



Exploratory Data Analysis

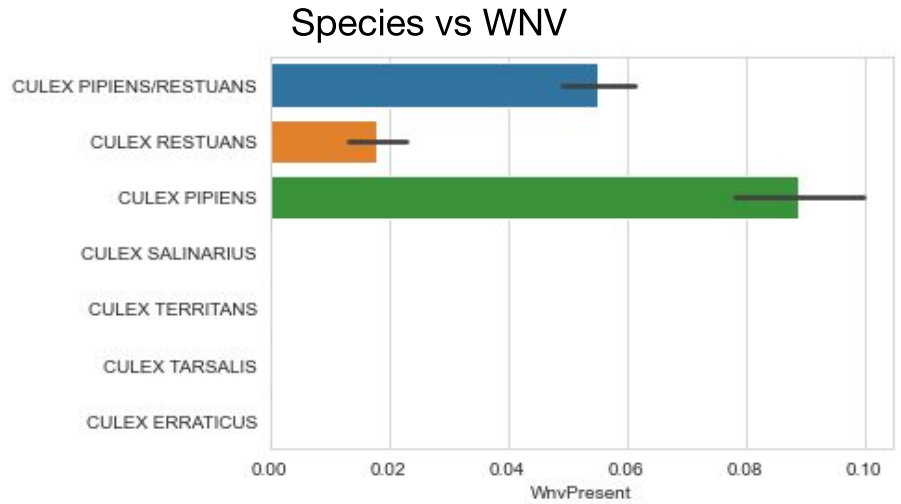
EXPLORATORY DATA ANALYSIS

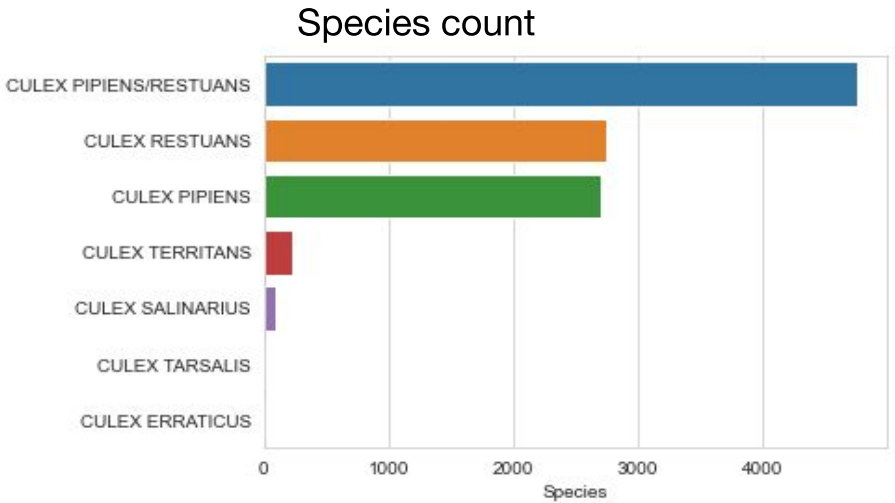
MOSQUITO SPECIES

Only two mosquito species can carry WNV, Pipiens and Restauns, yet those species are over-represented

The chart, top, shows the presence of WNV versus different mosquito species

The chat, below, shows the distribution of mosquito species





EXPLORATORY DATA ANALYSIS

LAGGED FEATURES

Features were lagged between 2 and 60 days using two strategies:

1. Uniform moving average
2. Exponentially weighted moving average

1143

Total Features

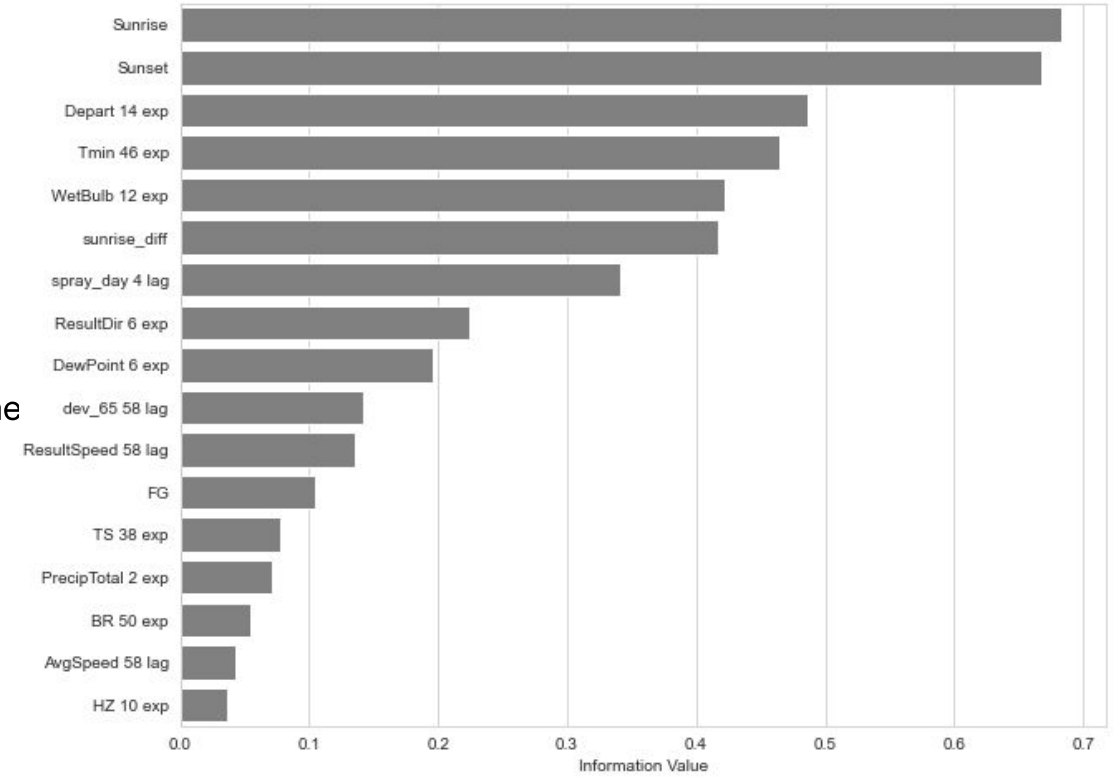
EXPLORATORY DATA ANALYSIS

INFORMATION VALUE

17 features remain from the analysis

Thresholds of information value (IV) > 0.02 and < 0.8 used to cull about 1000 features

For each feature, the lagged variable with the highest IV was kept

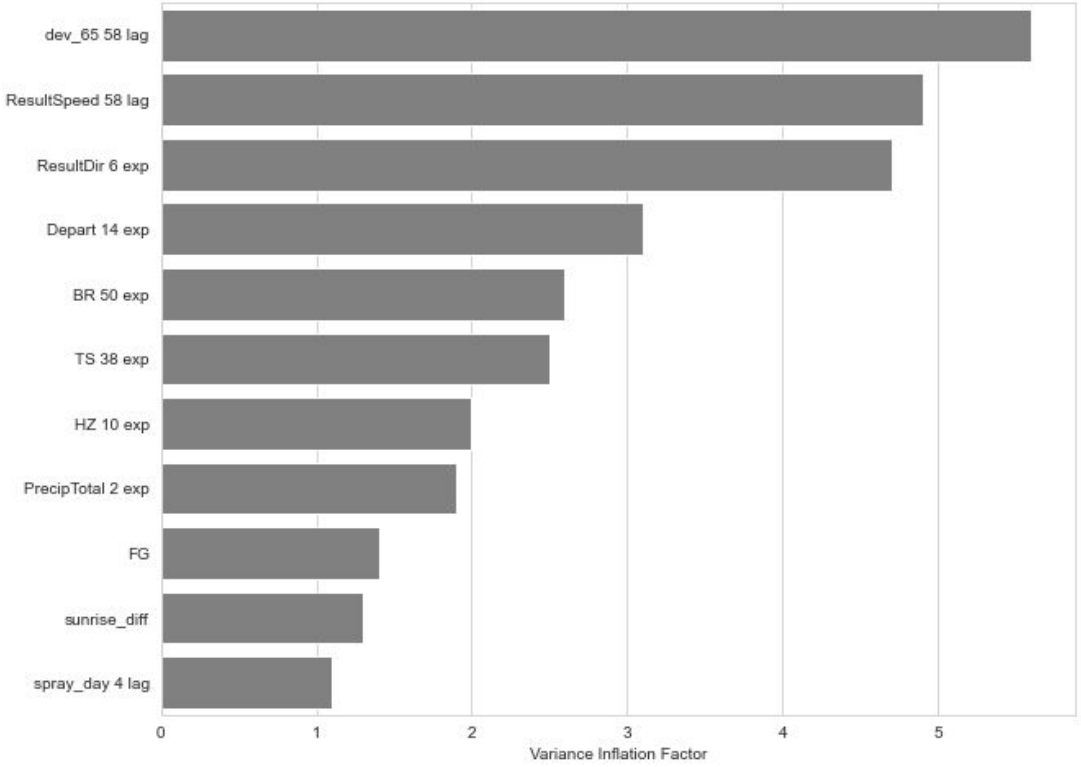


VARIANCE INFLATION FACTOR

Variance inflation factors (VIF) are a measure of multicollinearity, which contribute to modeling error.

Threshold of 6 was used to remove features one-by-one

Final VIFs are shown in the chart, right

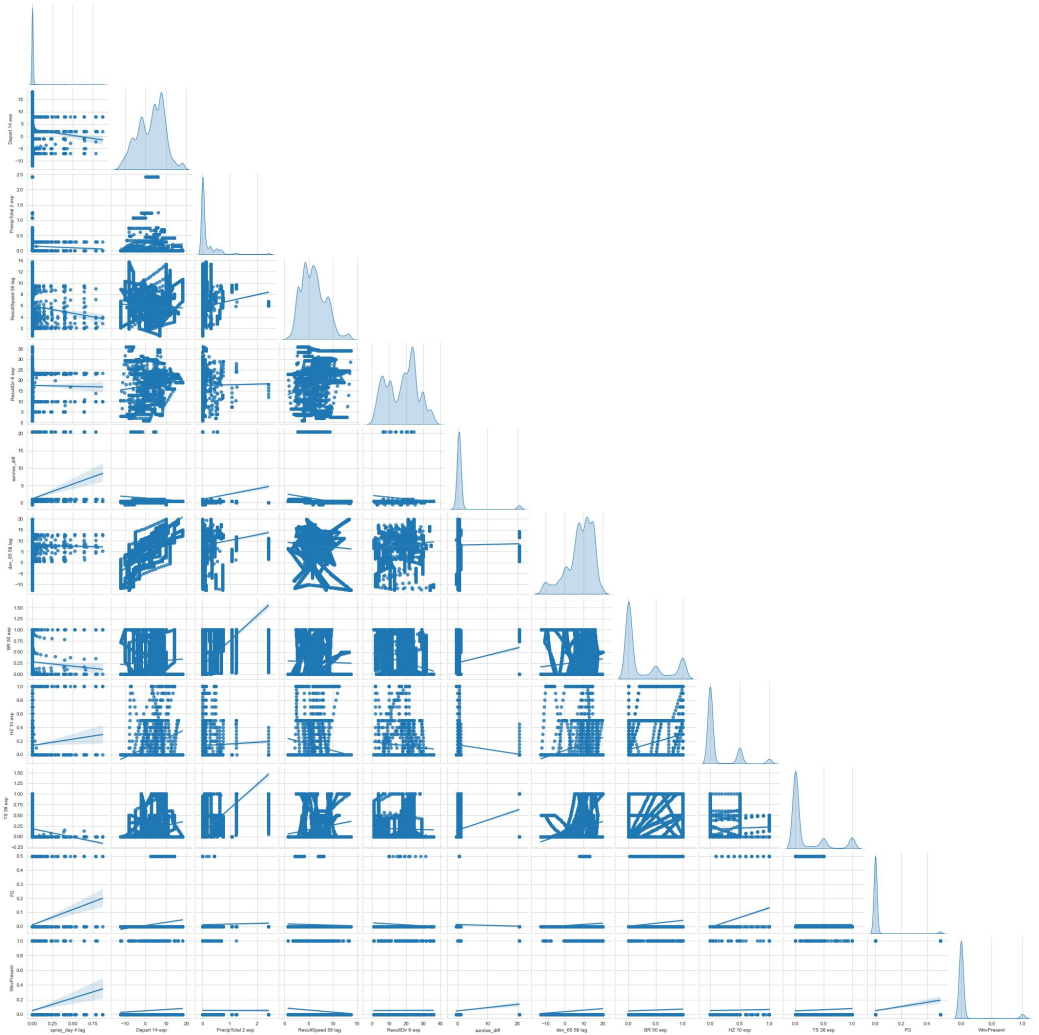


DATA ANALYSIS

FINAL PAIRPLOTS

At a high level, no clear relationships exists between features, or with the target variable (bottom row)

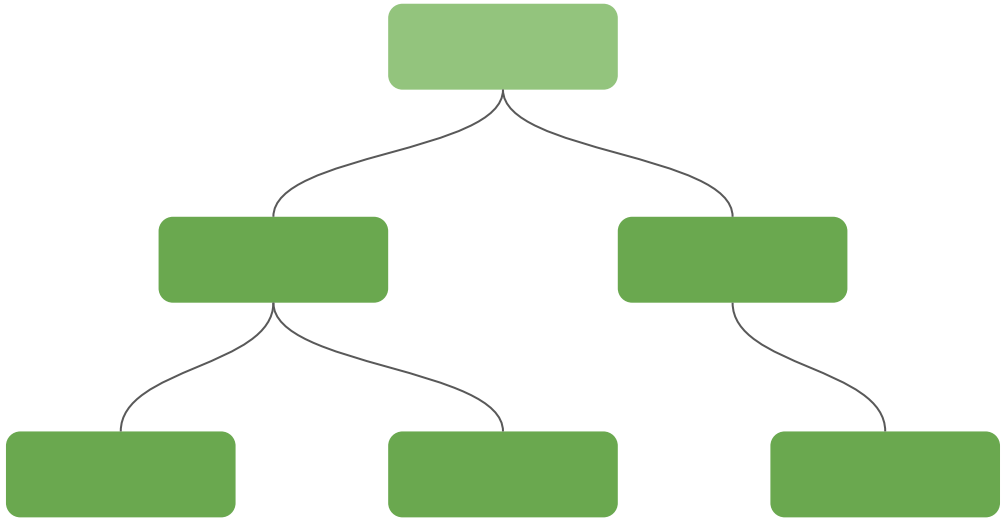
Features are now ready for modeling and prediction



Machine Learning Analysis

RANDOM FOREST CLASSIFIER

Many ‘naive’ decision makers are combined to make one more informed classification decision

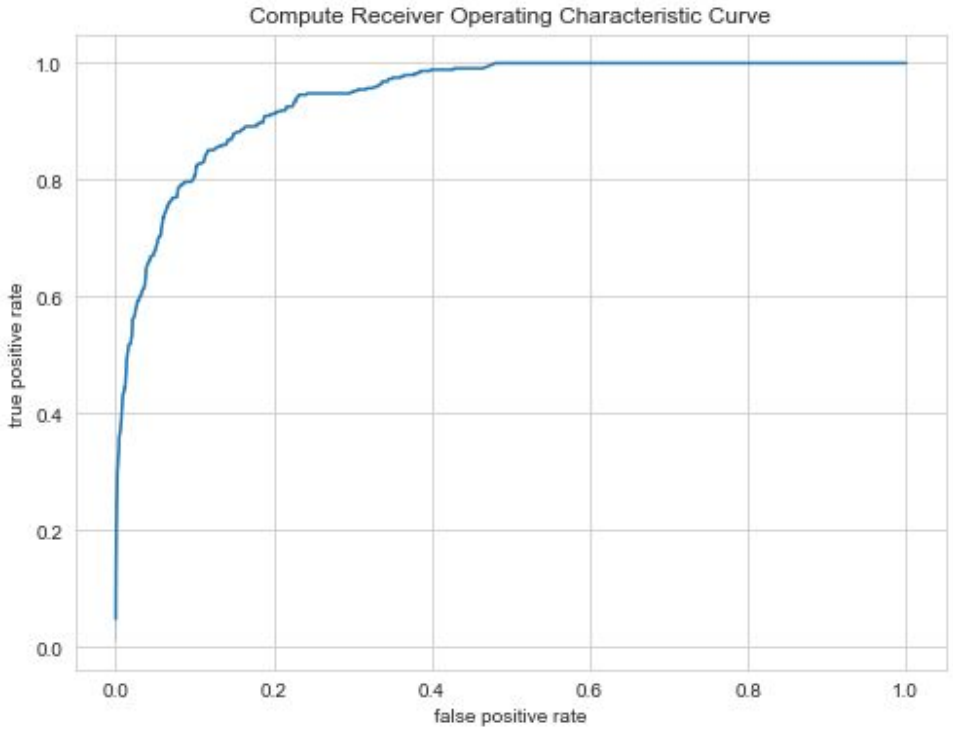


M L A N A L Y S I S

ROC AUC

Area under the receiver operating characteristic curve

- A model evaluation metric that is resistant to unbalanced classes
- The ROC is a measure of false/true positive rates as the threshold for class membership is changed
- ROC AUC of 0.8 or greater indicates good model performance



ML ANALYSIS

RANDOM FOREST CLASSIFIER

Two parameters need to be chosen:

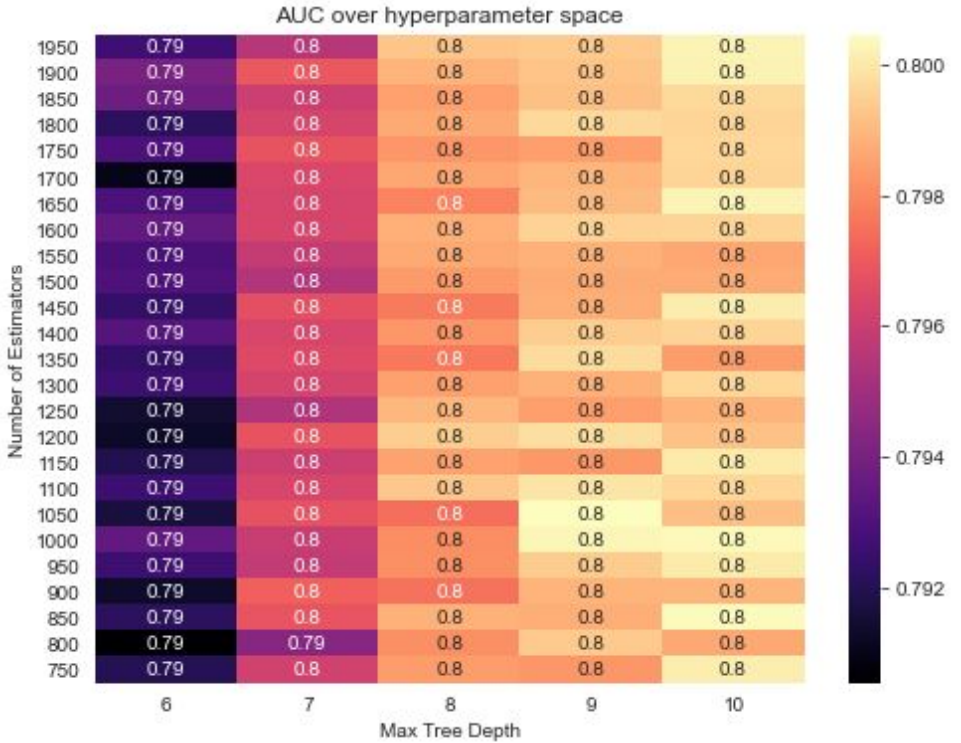
- Number of estimators
- Maximum depth of trees

In the plot, higher values indicate better model performance

TOP ROC AUC: 0.816

N ESTIMATORS: 1050

MAX DEPTH: 9



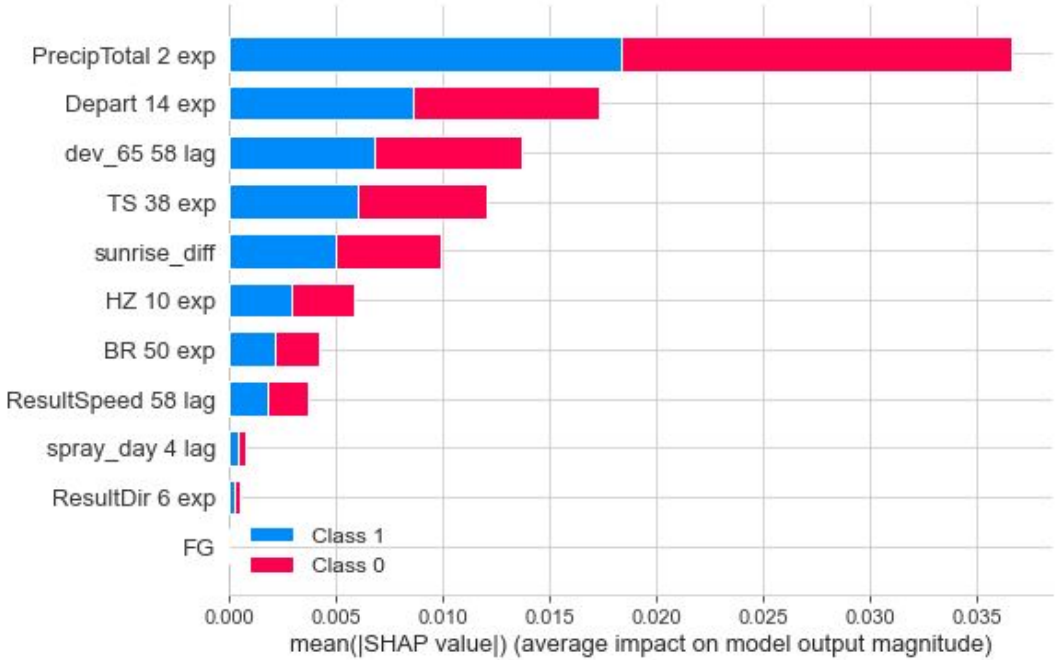
ML ANALYSIS

SHAP ANALYSIS

SHAP values quantify decision making importance for individual decisions

Precipitation and heat features are the 4 most important features

Spray data was the 3rd least important feautre

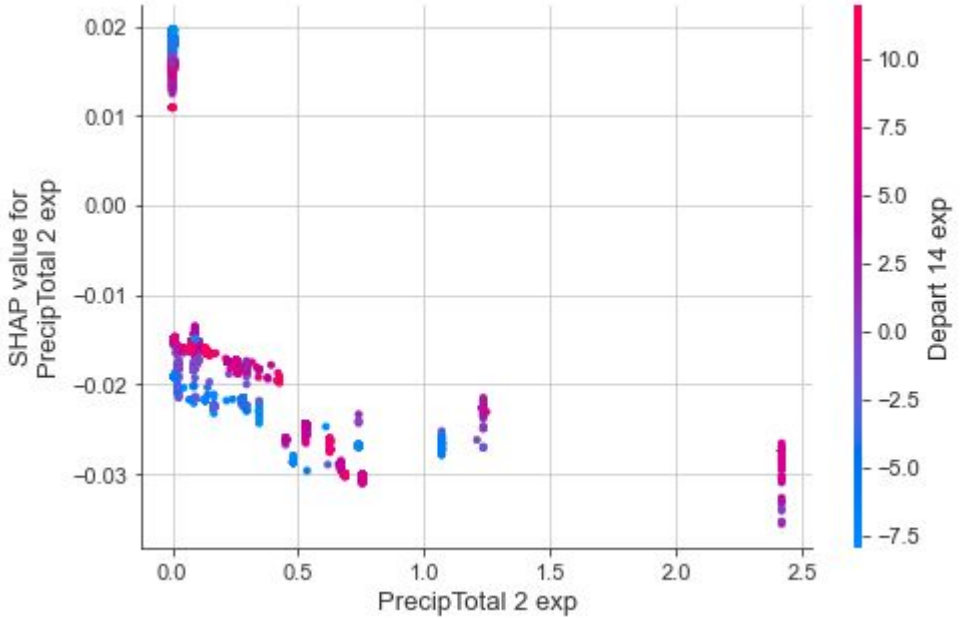


PRECIPITATION

PrecipTotal 2 exp is a 2-span exponentially weighted moving average of the daily precipitation

Heavy rainfall is associated with fewer mosquito trappings and WNV presence in short time periods

Consistent with research findings



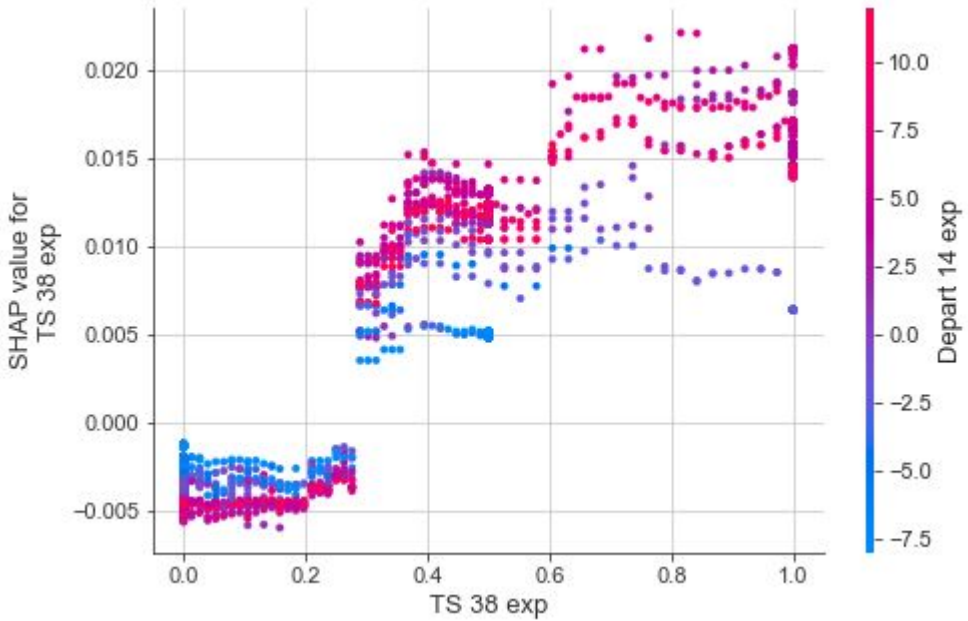
ML ANALYSIS

THUNDERSTORMS

TS 38 exp is a 38-span exponentially weighted average of thunderstorm occurrences

More rainfall over month-long periods is associated with more WNV presence

Consistent with research findings



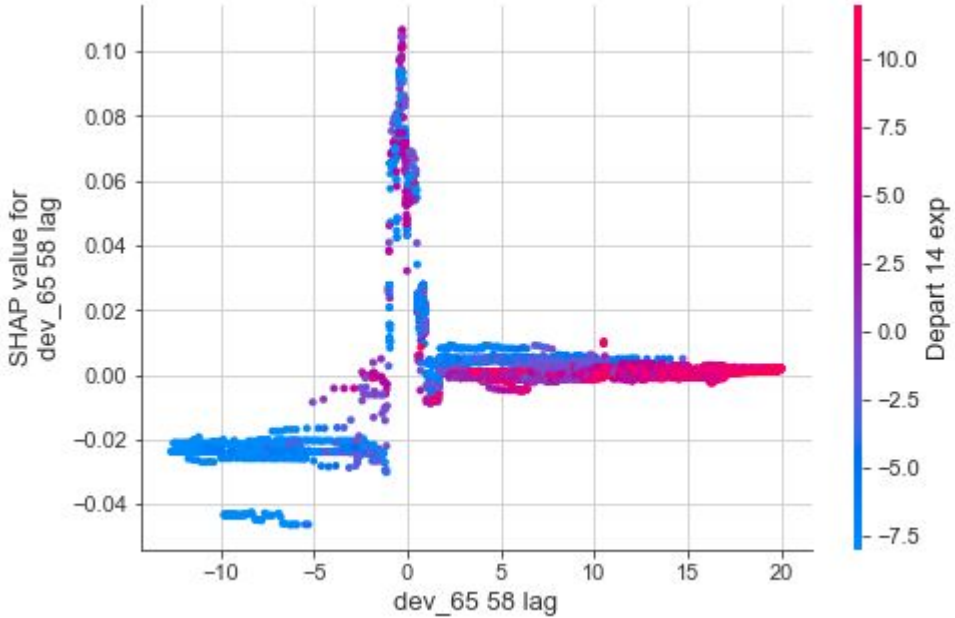
ML ANALYSIS

TEMPERATURE

Dev_65 58 lag is a 58 day trailing uniform average of deviance from 65 degrees F

Temperatures at or near 65F are strongly associated with WNV presence

Cooler temperatures are associated with less WNV presence

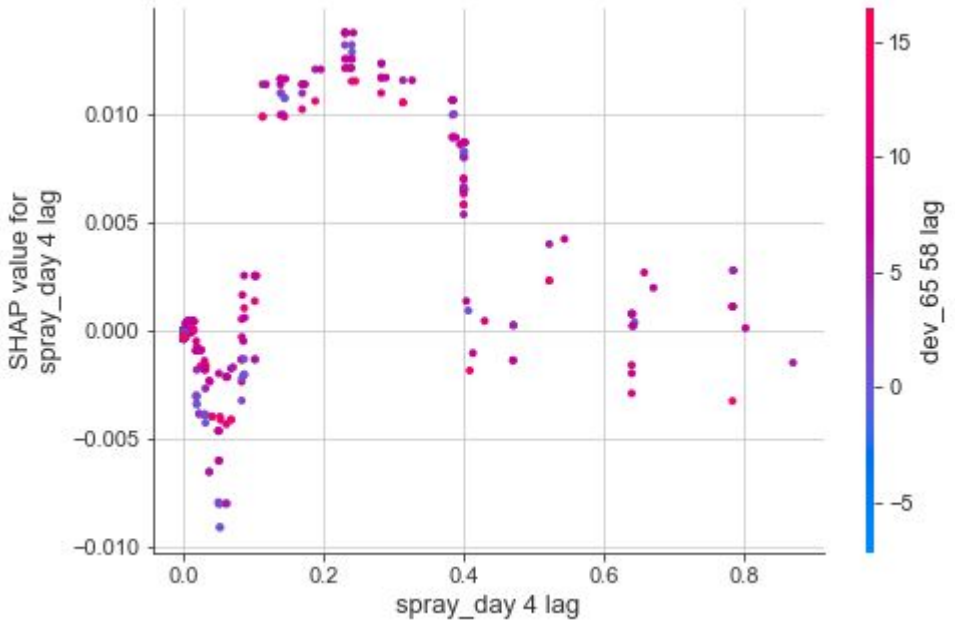


SPRAYING

WNV has a non-monotonic relationship with spray occurrences

Existing responses to mosquito populations confound the relationship we're able to see

More features quantifying human interactions with mosquito populations are needed



Conclusions

CONCLUSIONS

High monthly rates of rain are associated with more WNV

Average temperatures at or near 65F are associated with more WNV

WNV presence can be predicted reliably

Human intervention in the system is not well quantified. Modeled relationships are confounded by human intervention in mosquito populations

RECOMMENDATIONS

Field-validate the model during the 2021 mosquito season to build buy-in and consensus

Implement data-informed decision making into mosquito-control processes

FUTURE WORK

Develop metrics that better quantify human interactions with mosquito populations, as well as risk of exposure

Incorporate human factors into the prediction and management of WNV carrying mosquitoes

Continually iterate on the model in order to adapt to changing environmental circumstances