Predictive Modeling For West Nile Virus

Chicago Department of Public Health

Problem for Analysis

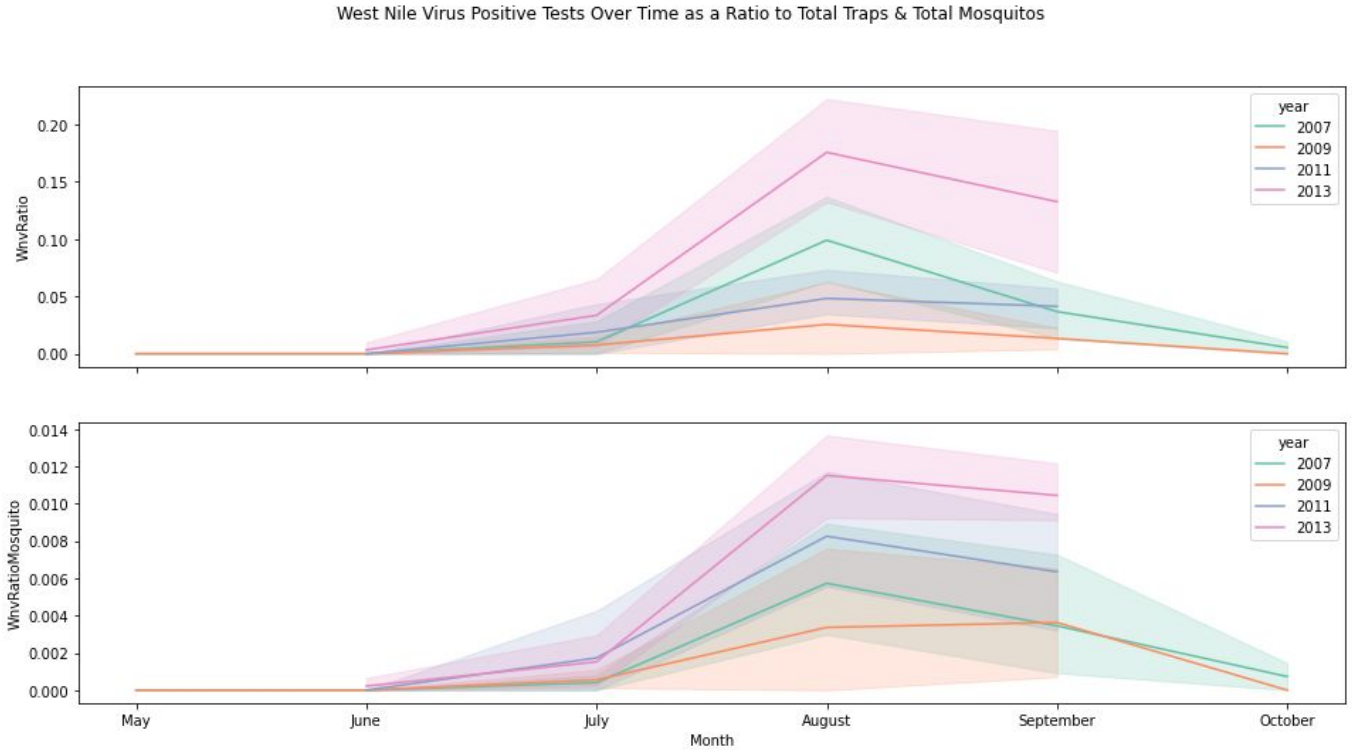
The City of Chicago has a comprehensive surveillance and control program to control the spread of West Nile Virus.

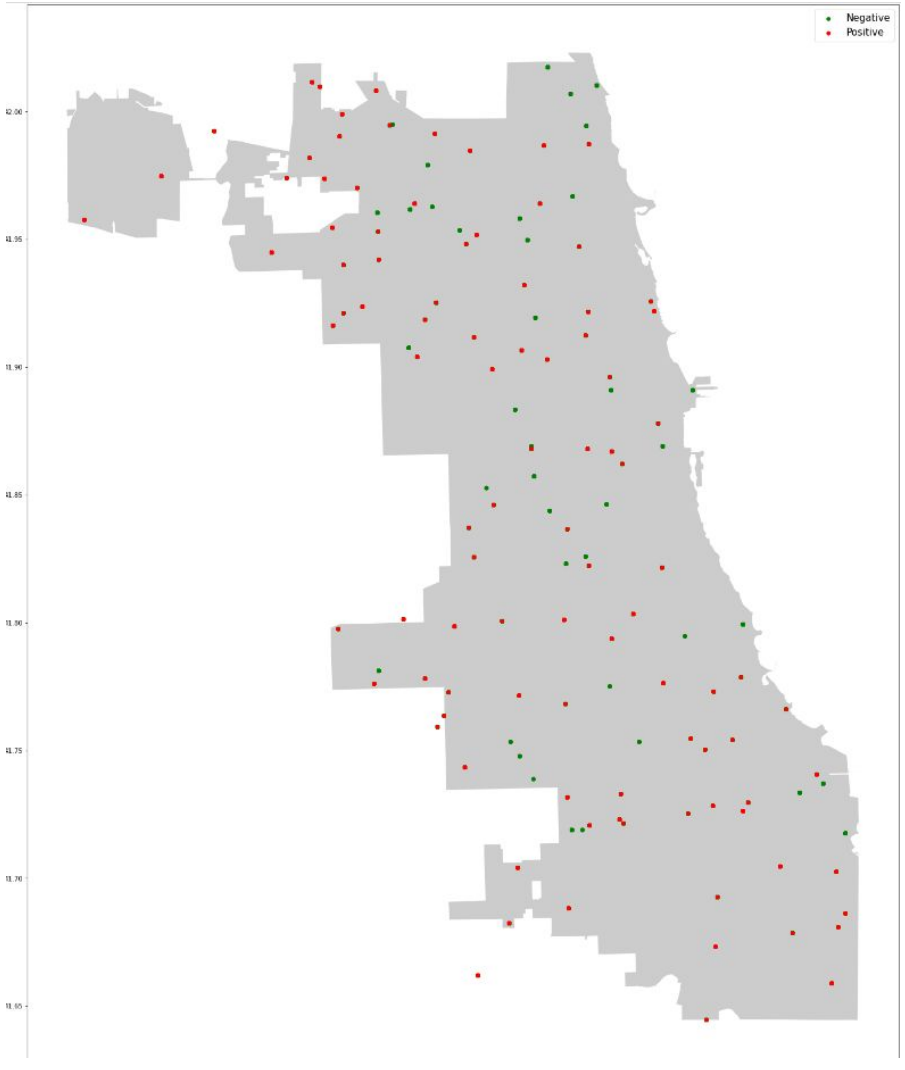
Problems to solve:

1. Build a model that can accurately predict the presence of the virus given information about the mosquito trap location, the mosquito species, and the weather conditions of the day.
2. Which features make our model more likely to predict a positive test on the virus?

Exploratory Data Analysis Part 1: Virus Trends Over Time

The Peak Month for West Nile Virus is in August. The Virus appeared to have dipped between 2007 to 2009 but is now back on the rise since, increasing from 2009 to 2011 and then again from 2011 to 2013.



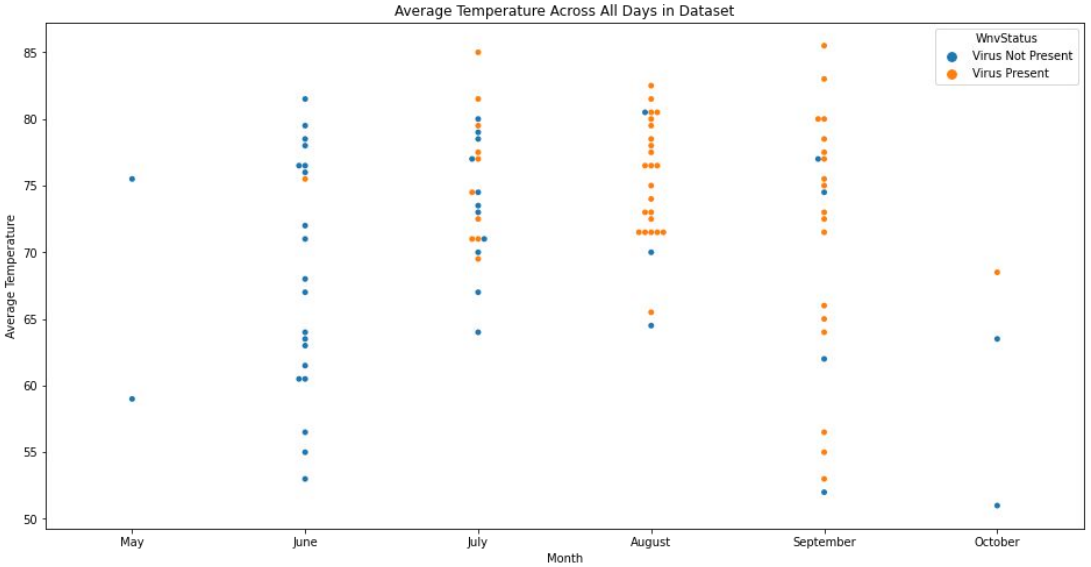


Exploratory Data Analysis Part 2: Virus Presence By Location

Each of the red dots on this map represent a location where at least one of the traps had a positive test.

Note that Northwest Chicago appears to be a hotspot.

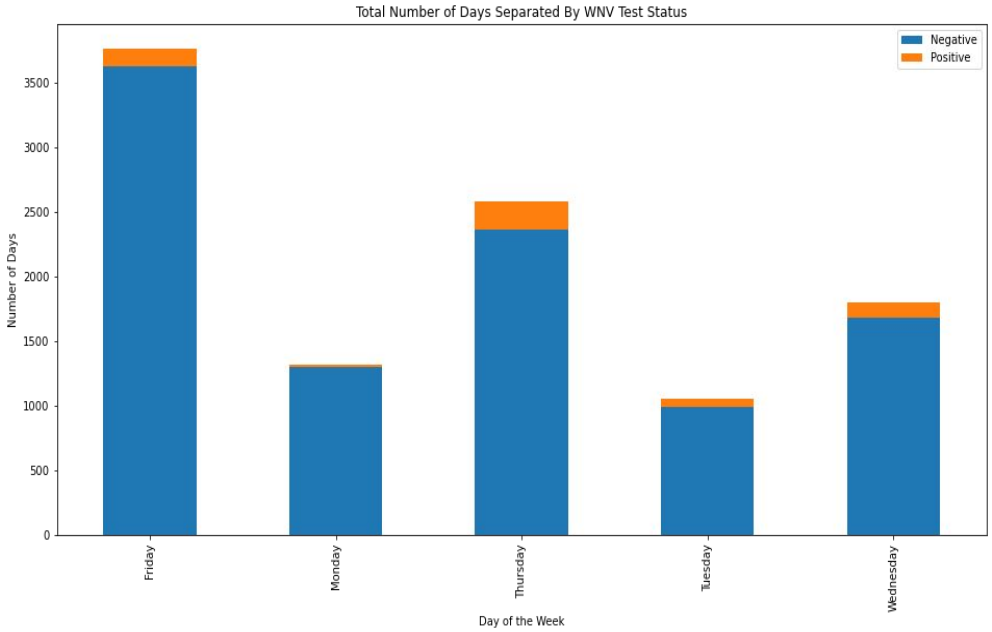
Exploratory Data Analysis Part 3: Virus Correlation with Weather Conditions



- The virus tends to show up when temperatures rise.
- This is a likely explanation for why the virus peaks in August.

Feature Engineering: Creating more explanatory variables

- Our next step before we start building the model is feature engineering. Here, we're looking to make the most out of our existing data by creating new features that will ultimately help our model make better predictions.
- Our final dataset added the following features:
  a. Day of the Week (pictured on right)
  b. Month
  c. Weather conditions
  d. Days past since a previous weather condition last occurred
  e. Municipality

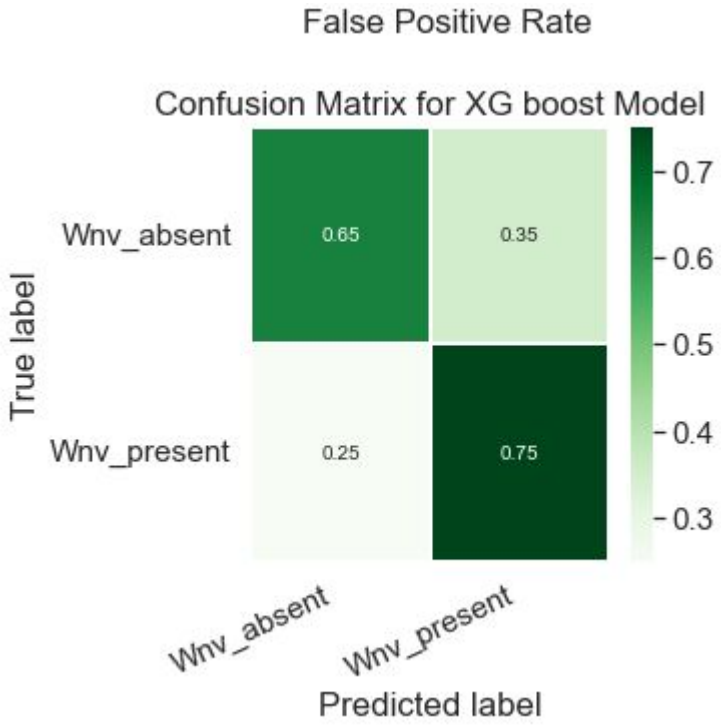


Feature Engineering Part 2: Optimizing features

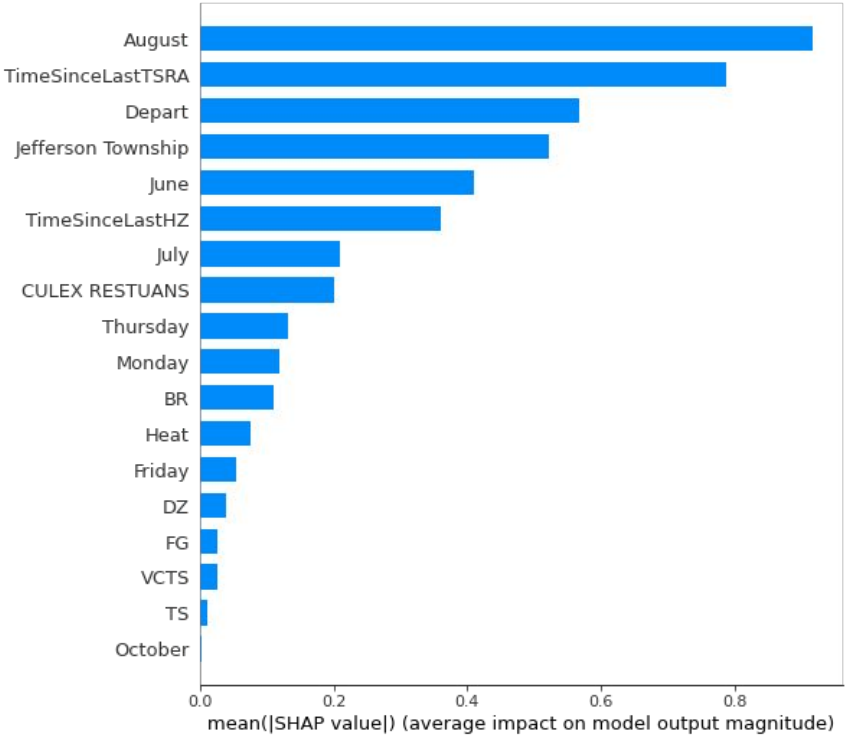
- After adding all of our new features, our dataset had over 70 explanatory variables.
- The next step in our feature process was to cut down these features to only the ones which had a high “information value” meaning that they had more value in their ability to predict the virus. We also remove any features which had high collinearity (i.e. correlation with other features).
- At the end of this process, we were left with **18** variables entering our model building stage.

XG Boost Model Performance Overview

- Using our final 18 features, we built an XG Boost model with a recall score of 0.75. In other words, our model is able to successfully predict the presence of the West Nile Virus 75% of the time.
- Our F1 score is slightly lower at 0.695. This is because our model's precision score is slightly lower than it's recall indicating that it will occasionally mark false positives in the data.
- However in the given context, more weight is placed on a higher recall score because **the impact of a false negative is more consequential than the impact of a false positive.**

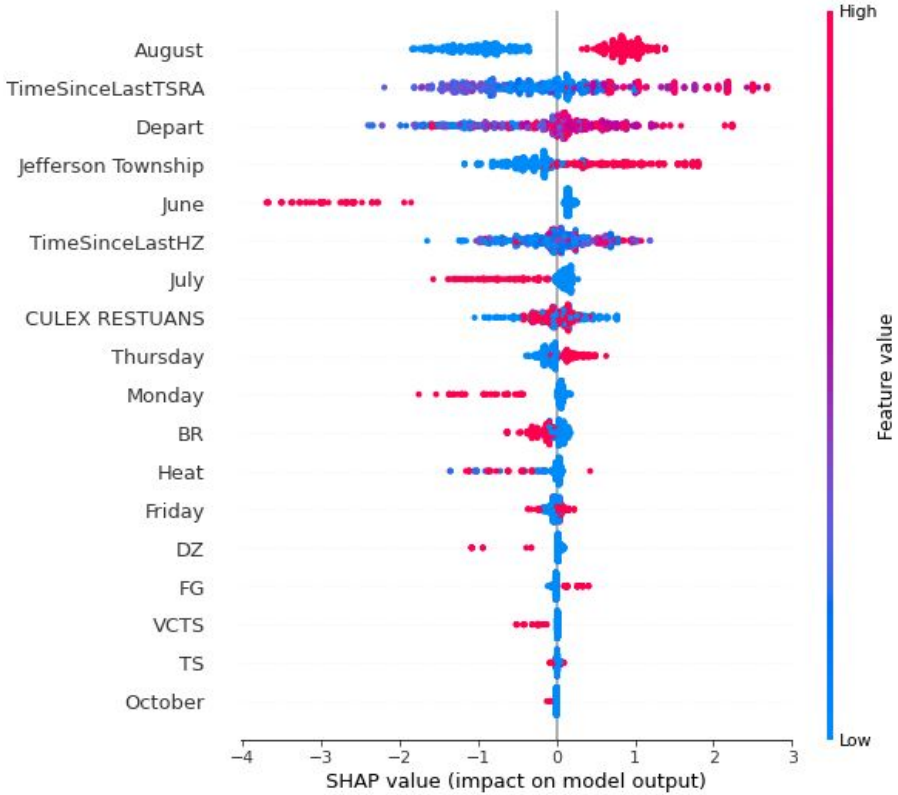


Feature Importance



- SHAP analysis on our model reveals that the most impactful features are:
  - If the data was from August
  - How much time has passed since the last time there was a rainy thunderstorm
  - Departure from normal (i.e. how much the temperature on a given day was above or below the 30-year normal)
  - If the trap is located in the Jefferson Township Municipality.

Feature Importance Part 2: Top 4 Feature Breakdown



1. <u>August:</u> If the data was from August, this made the model more likely to predict positive. This confirms our earlier observation we made in the exploratory data analysis stage.
2. <u>TimeSinceLastTSRA:</u> Generally, the longer it has been since there was last a rainy thunderstorm, the more likely the model predicts positive.
3. <u>Depart:</u> The more extreme the weather condition, the more likely the model predicts positive. This could also explain why we observed a general trend earlier that higher temperatures lead to more positive virus cases (AND it could also explain what makes August the peak month since that's when the highest temperatures occur).
4. <u>Jefferson Township:</u> Traps from Jefferson Township will make the model more likely to predict positive.

Conclusion

- The virus starts to appear in July and peaks in August before slowly disappearing by the time October starts.
- Longer periods of time without any bad/stormy weather is likely to make the virus cases go up.
- Hotter temperatures are also likely to make the virus cases go up.
- The Jefferson Township Municipality is the primary hotspot for the virus in terms of geolocation.

Recommendations

- Increase coverage of spraying in the Northwest Chicago region particularly the areas encompassed in the Jefferson Township Municipality.
- Focus the spraying schedule to be more concentrated in August.
- Add in additional out-of-schedule sprays if temperatures are noted to be particularly high for a certain day OR if at least 3 weeks have gone by without any rain or thunderstorms.