West Nile Virus Prediction

Ying Ma
08/13/2018

West Niles Virus

Facts

- Symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death
- No human vaccine
- The best method to reduce the risk is avoiding bites by infected mosquitoes

Prevention

- The Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program
- Every week from late spring through the fall, mosquitos in traps across the city are tested for the virus.

Data-driven practical insights

Data

Traps

Lab test results of the presence of WNV in the traps

Train: 2007, 2009, 2011, 2013
Test: 2008, 2010, 2012, 2014

- Species
- Trap location
- Time: year and month/week of the year

Weather

Daily weather data from NOAA

Available from beginning of May to the end of Oct, 2007 - 2014

- Temperature
- Precipitation
- Pressure
- Wind

Spray

GIS data for aerial spray done in Cook county

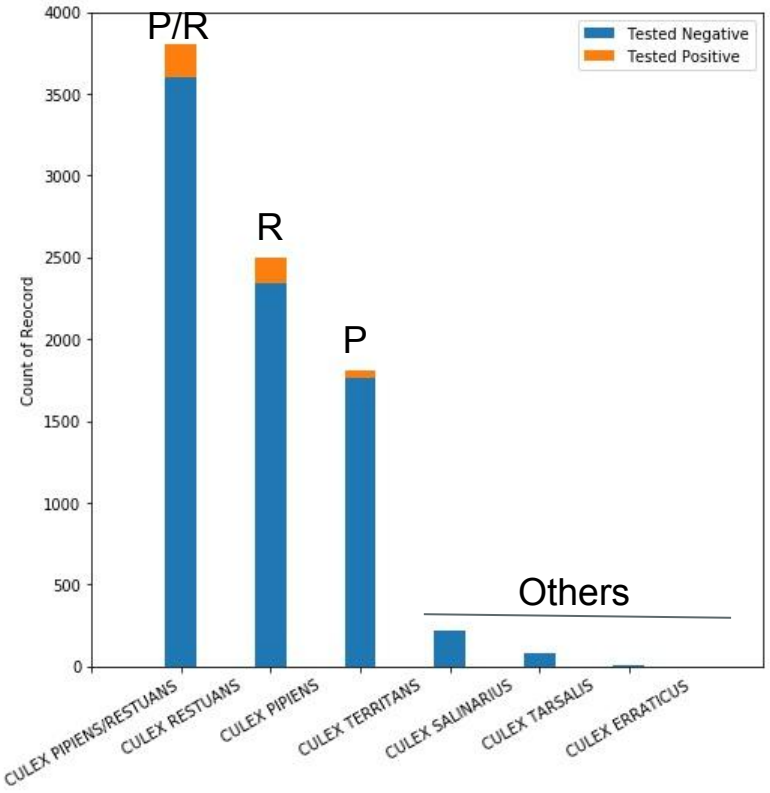
Only available for 2011 and 2013

- Date and Time
- Spray Location

Features

Traps

- Year
- Week of the year
- Species (one hot encoding)
  - Pipiens
  - Restuans
  - P/R
  - Others
- Latitude/Longitude
- Trap_id (one hot encoding)

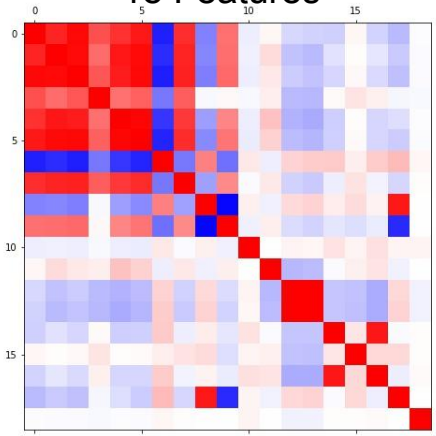


Features

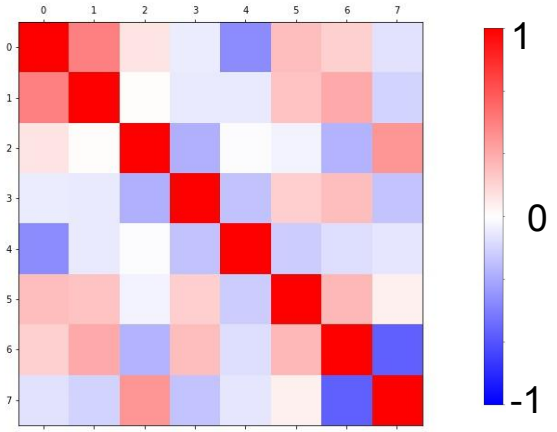
Weather

- Average values from the same week
- Data from Station 1 is kept
- Features relative to temperature, humidity, precipitation, pressure, wind are kept
- Highly correlated features revealed by correlation matrix

18 Features



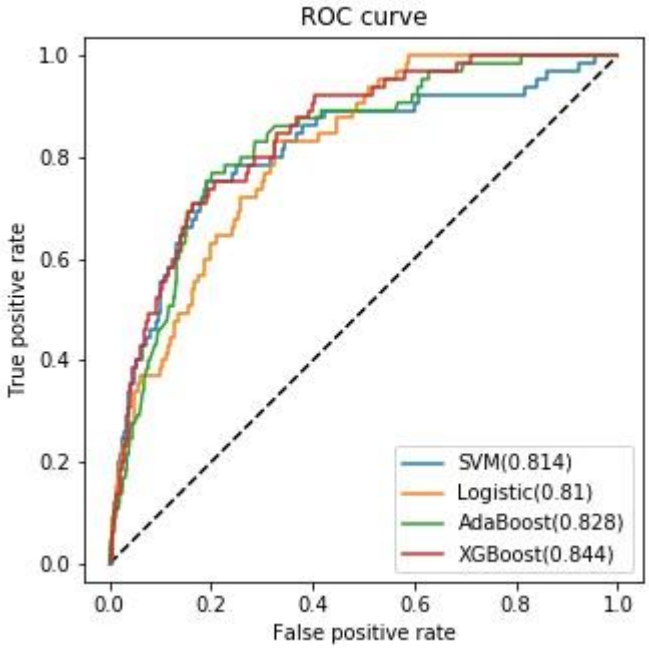
8 Features



- Temperature
  - T_diff = Tmax - Tmin
  - T_avg
  - T_depart
- Relative Humidity
  - Calculate from WetBulb and DewPoint

- Total percipitation
- Atmospheric Pressure
  - Station Pressure
  - Sea Level Pressure (drop)
- Wind
  - Resultant wind speed
  - Direction (cos)
  - AvgSpeed (drop)

Model



XGBoost

- ~~Interpretability~~
- Good performance
- Deals well with non-linear decision boundaries
- Works well with categorical and ordinal data
- Not strongly affected by class imbalances

Distance Encoding

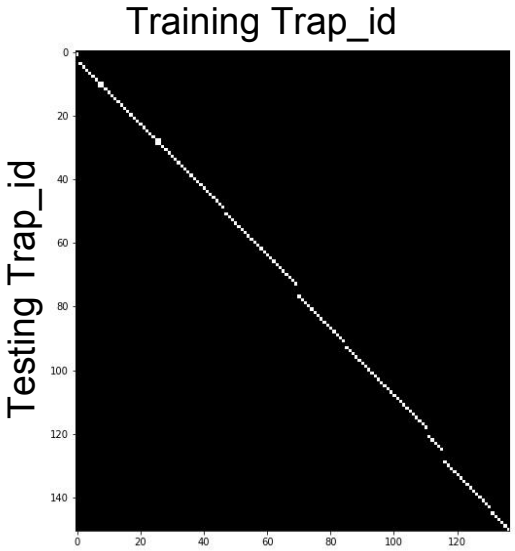
Problems with one hot encoding of trap location:

- Some traps/satellite traps in the test set that didn't appear in the training set.
- Trap location migrated over the years.

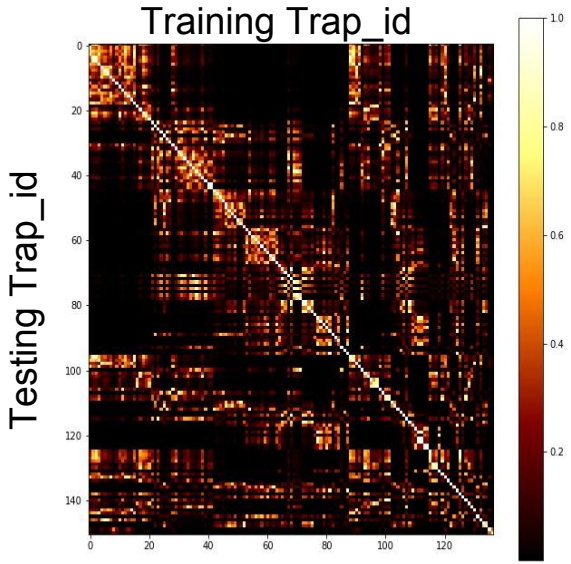
Solution: distance encoder

- Distance between two sites are calculated by Haversine formula

- Weight = 2/(1+exp(λ*distance))
  - Distance = 0 => weight =1
  - Distance ->∞ => weight->0

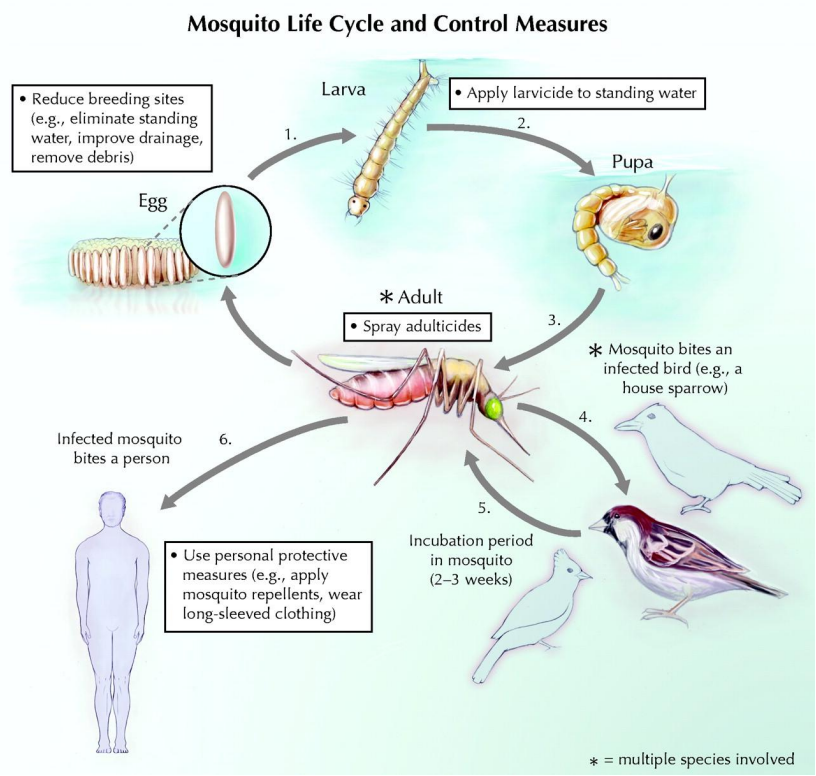


One Hot Encoding



Distance Encoding

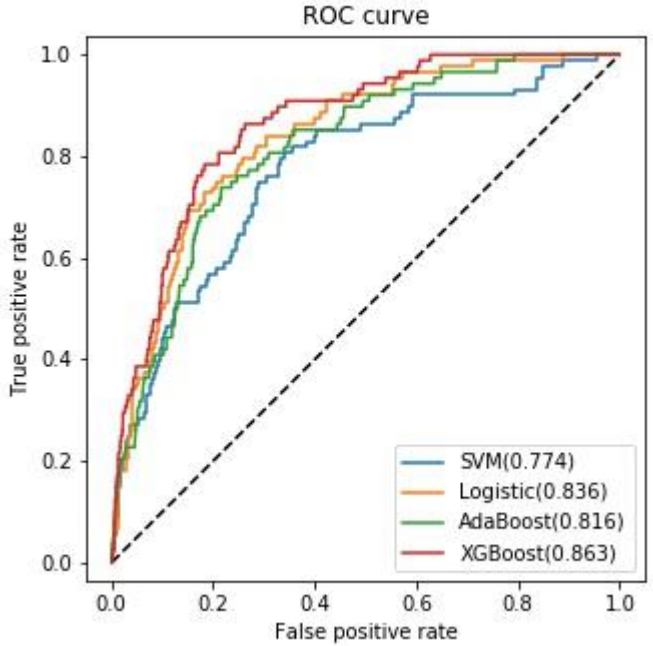
Delayed Weather Feature



- Different weather parameter may affect the mosquito and the WNV life cycle differently
- Average lifespan of adult female mosquitoes: 4-6 weeks
- Incorporated weekly weather features up to 4 weeks prior to the test date

Results with improved features

AUC of ROC of hold out test set

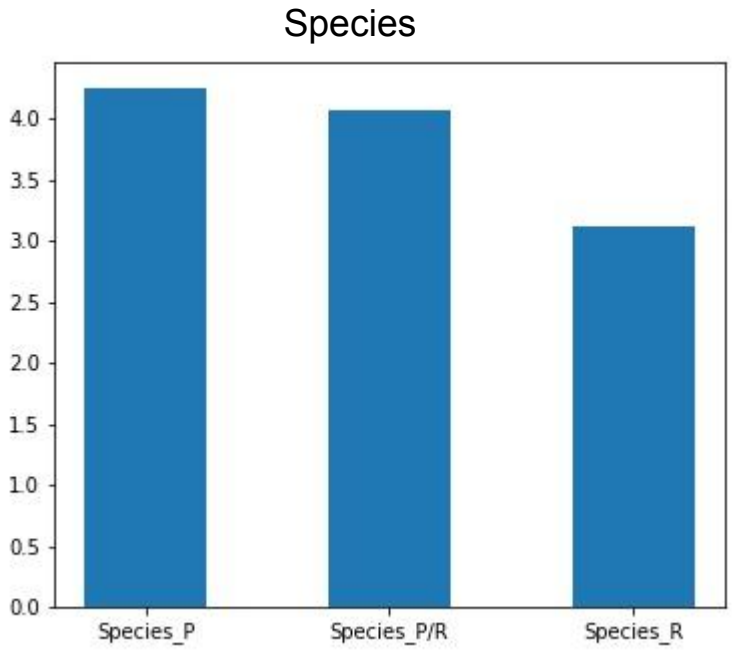


AUC of ROC of Kaggle test set

XGB	0.76686	0.78143
+ distance encoder	0.77362	0.79148
+ delayed weather	0.77896	0.79988

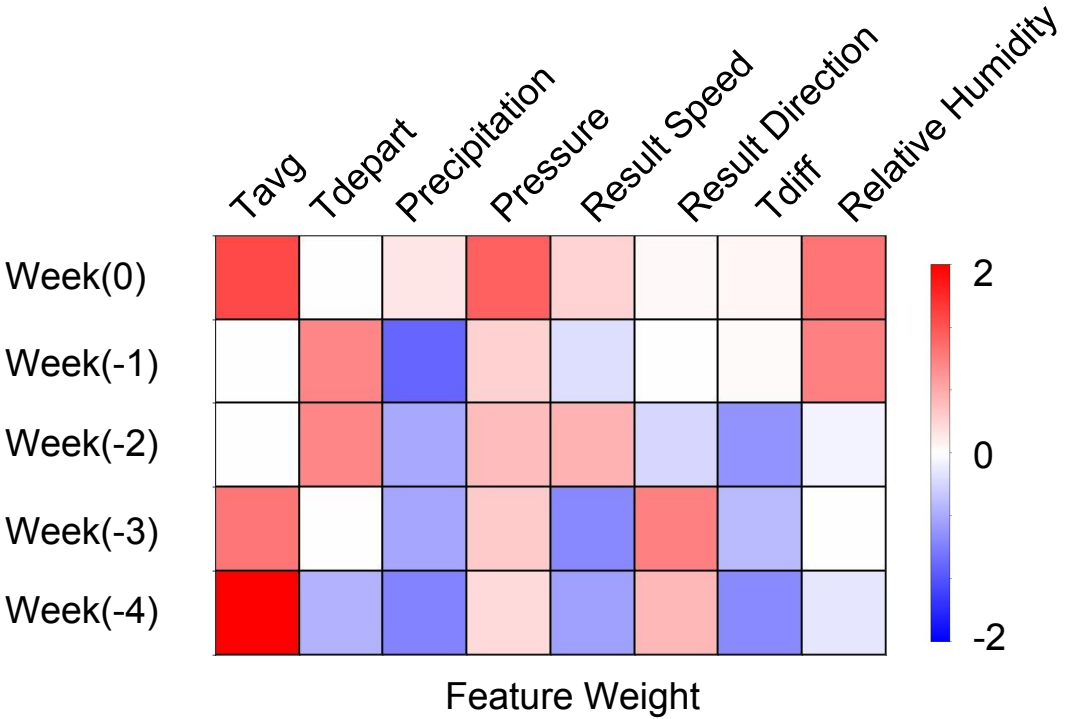
Feature interpretation is done on the coefficients of regularized logistic regression.

Feature Weight: Species



- Species P, R and P/R have high positive weight which suggests that they're the major driving factor for WNV presence
- Other species are not tested positive in the WNV test in the training set, therefore they don't have predictive value.
- However, because other species are less representative in the training set, I can't entirely eliminate the possibility of other species getting infected
- More data about other species and expert opinion is needed

Feature Weight: Weather



Insights:

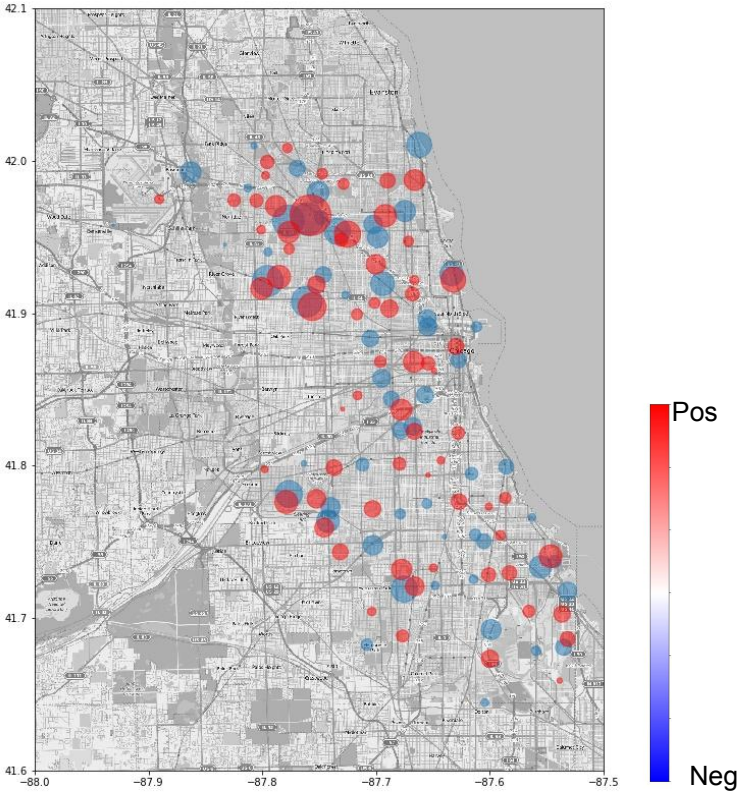
1. Weather of the week prior to the test is important to the model

2. WNV likes higher temperature, but doesn't tolerate high temperature difference.

3. WNV prefers less precipitation in longer term, however higher relative humidity facilitates the spread of WNV in shorter term.

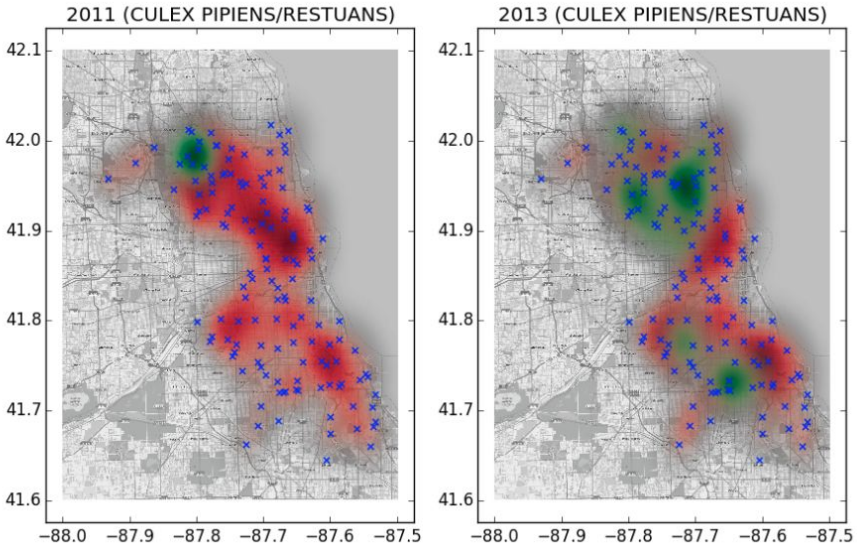
4. WNV appears to like higher atmospheric pressure.

Feature Weight: Location



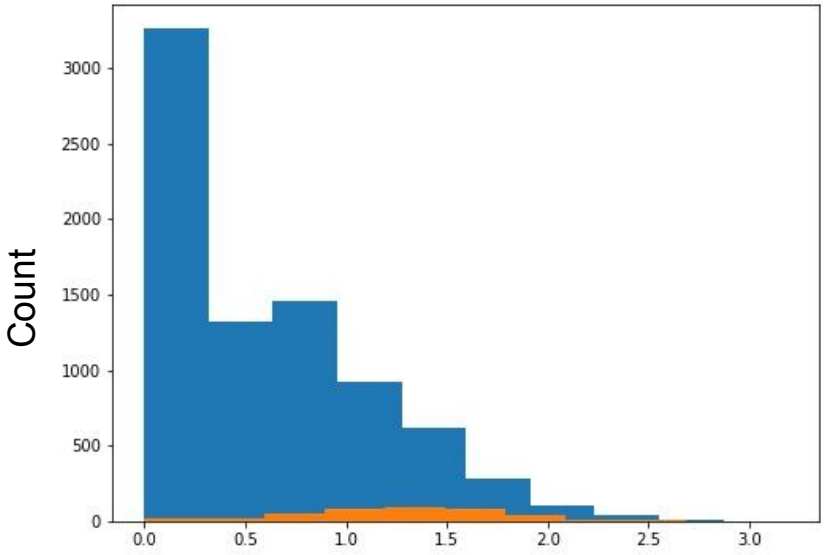
- Red: positive weight
  Blue: negative weight
- Radius of the circle stands for the weight value
- Areas at higher risk locate at top of the map
- Blue circles are more evenly distributed

Aerial Spray



Red: WNV presence
Green: Aerial spray

NumMosquitoes vs. WNV presence



log(NumMosquitoes)

Population