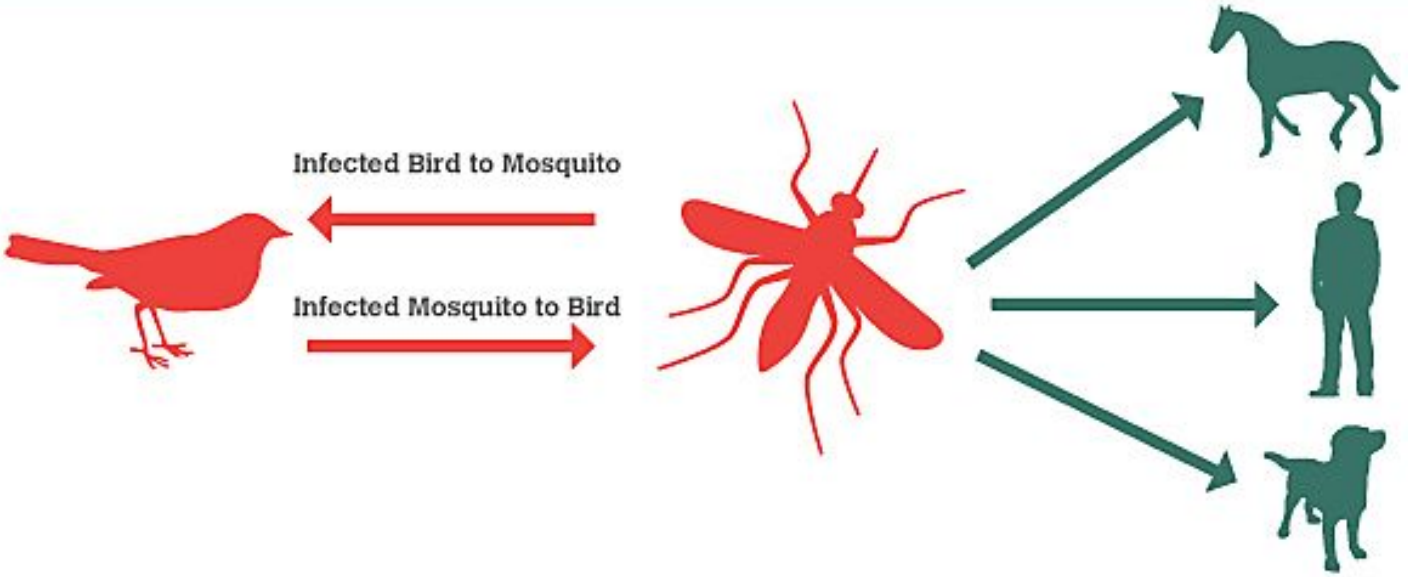
SparkBeyond: West Nile Virus

Ryan Grosso - Insight Data Science Fellow
11/13/2017

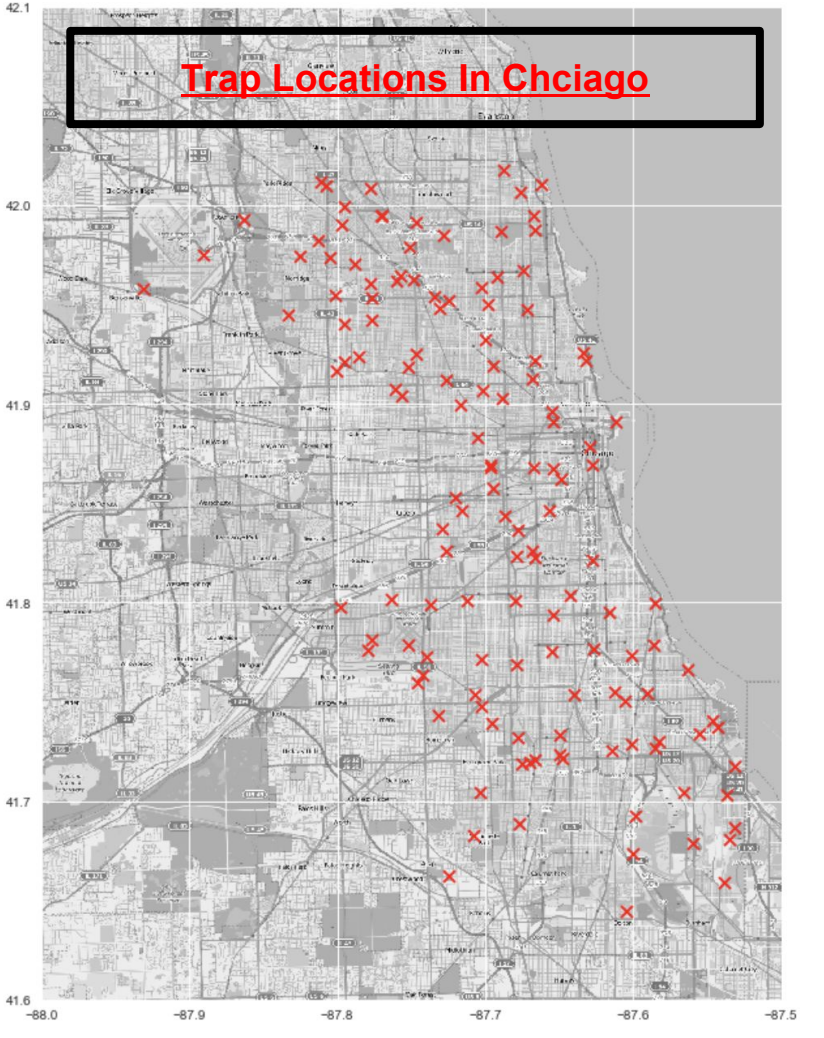
How West Nile Virus Is Transmitted



About the data

- WNV- Trap data
  - Train (2007, 2009, 2011, 2013)
  - Test (2008,2008,2010,2012,2014)
- Chicago weather
  - 2007-2014
- Spray data (2011 - 2013)

Datasets for this talk
2011 and 2013



Deciding on the goal

Build a model that can predict the likelihood that a trap will test positive for WNV.

- Identify as many traps as possible ➜ Optimize over recall

  - Use features that do not involve measuring the trap to predict WNV

  -Generate historical features to use as predictors

- Assuming the city has the best interest for the people.
- Assume that the spray chemicals get better over time.

  -This study is not going to test the effectiveness of the spray although it could.

Features

Mosquitoes take about 10-14 days to fully grow in standing water.

Generate features that are dependent on historical weather data.

Weather Features:

- Temperature : Average, Max, Min
- Precipitation : Average , Max, Min, Acuml.
- Wind : Average
- Pressure : Average

Seasonal Effects: Months

Location Effects: Where the trap is located (TrapID)

Features

Mosquitoes take about 10-14 days to fully grow in standing water.

Generate features that are dependent on historical weather data.

Weather Features:

Continuous

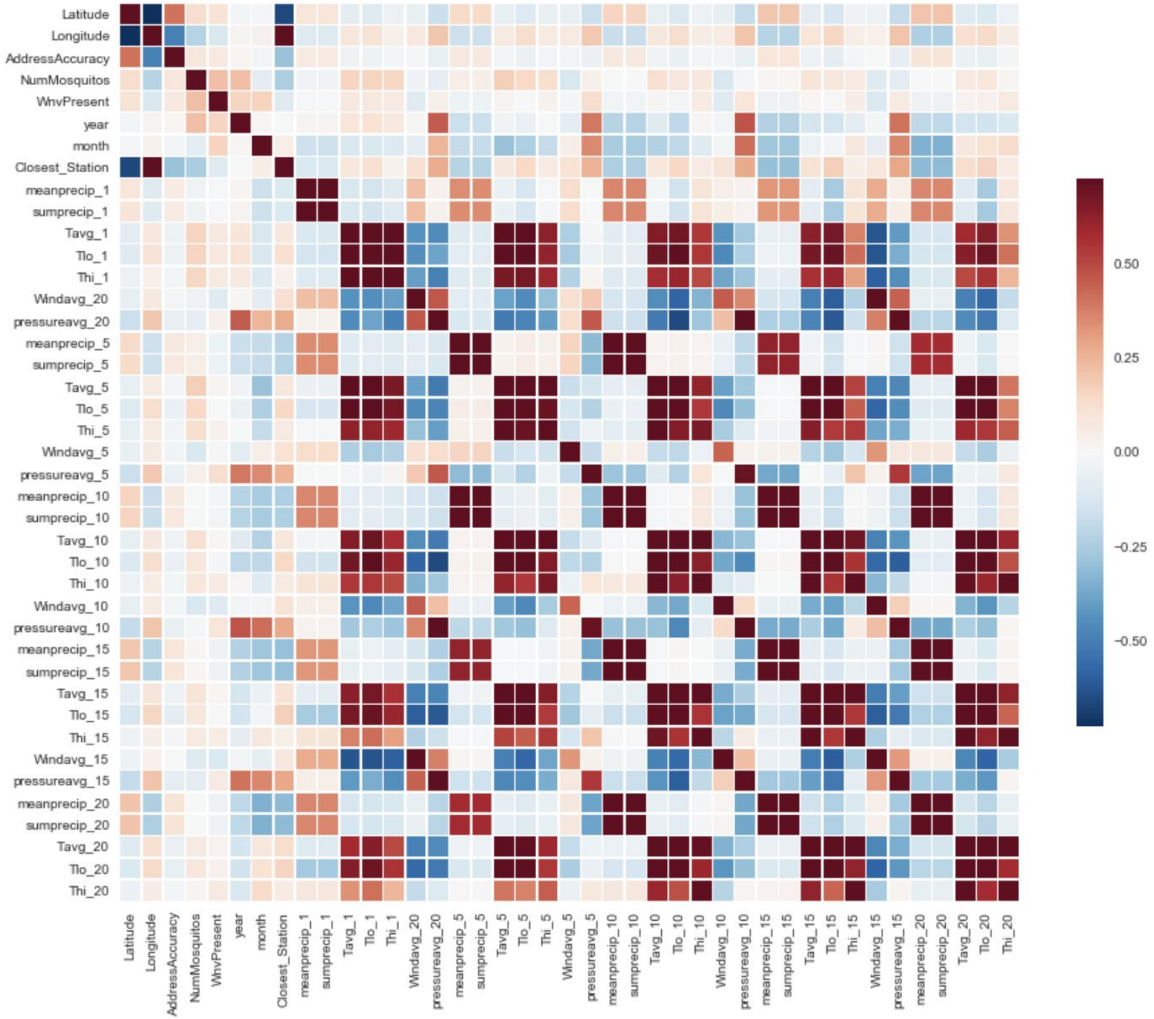
- Temperature : Average, Max, Min
- Precipitation : Average , Max, Min, Acuml.
- Wind : Average
- Pressure : Average

Seasonal Effects: Months

Categorical

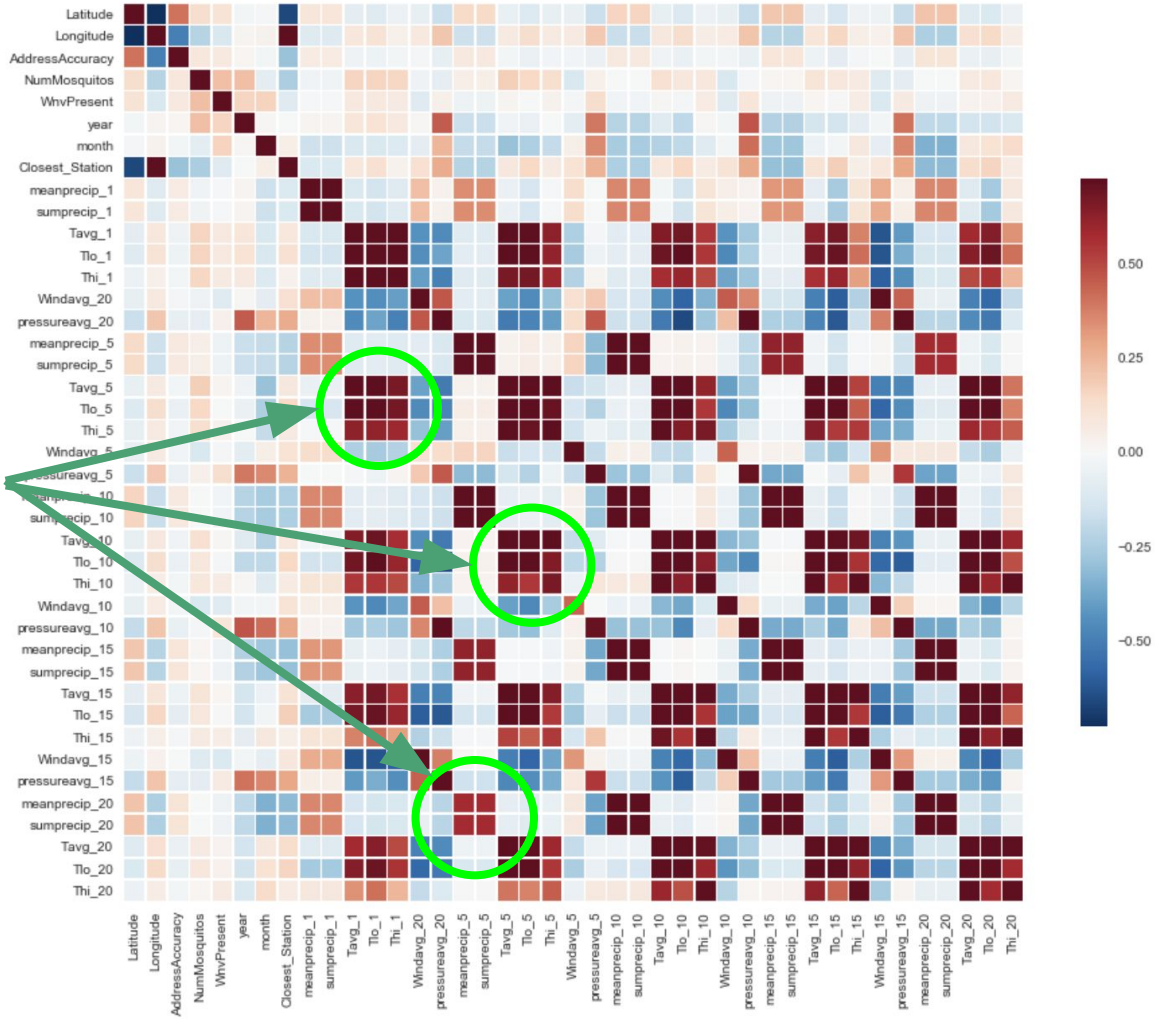
Location Effects: Where the trap is located (TrapID)

FEATURE Correlations

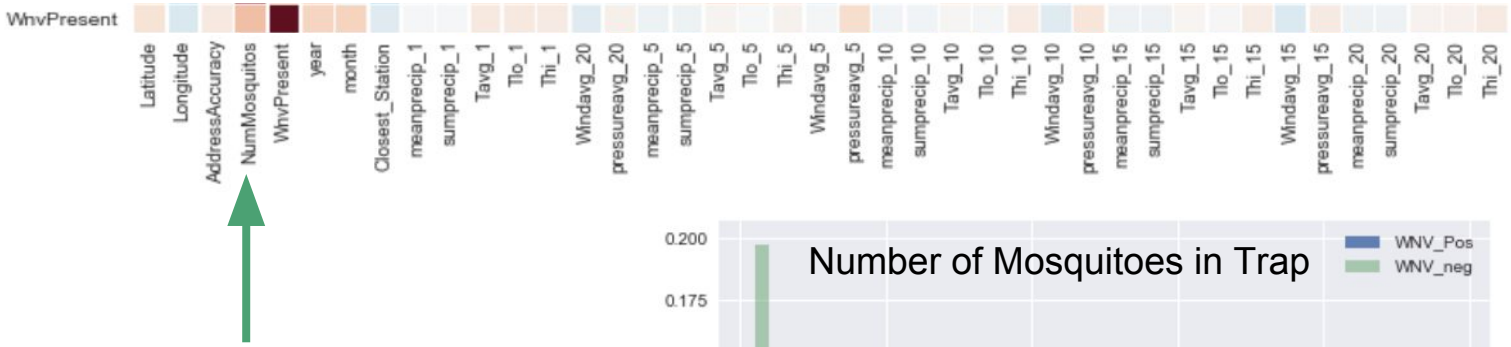


Intrinsic self-correlations

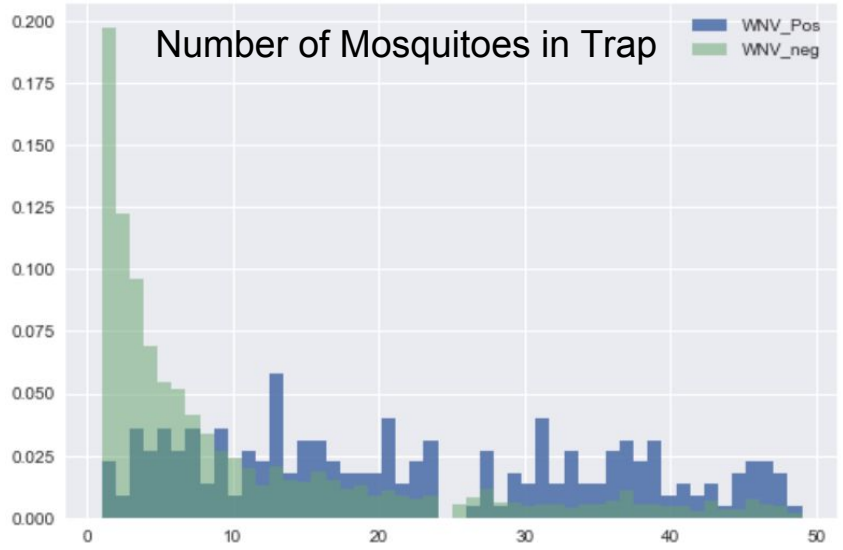
*Coming from weather data



Feature Correlation



Will not be using any data from the traps regarding number of mosquitoes, as this would require one to look at the trap



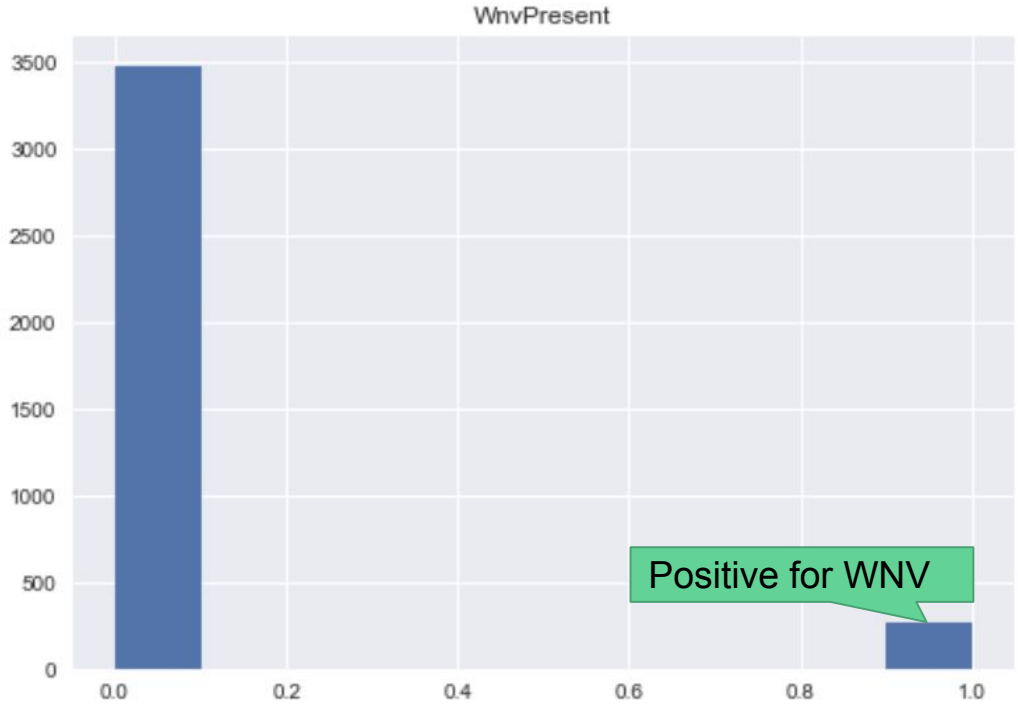
Out of the Box: Random Forest Classifier

RF-Classifiers:

- Great for categorical features.
- Robust to overfitting.
- Handles correlated features.

Class imbalance:

- UnderSampling
- OverSamping
- SMOTE



*Synthetic Minority Over-sampling Technique

RF-Classifier: undersampling

Not bad for out of the box.

- Scoring is done on accuracy

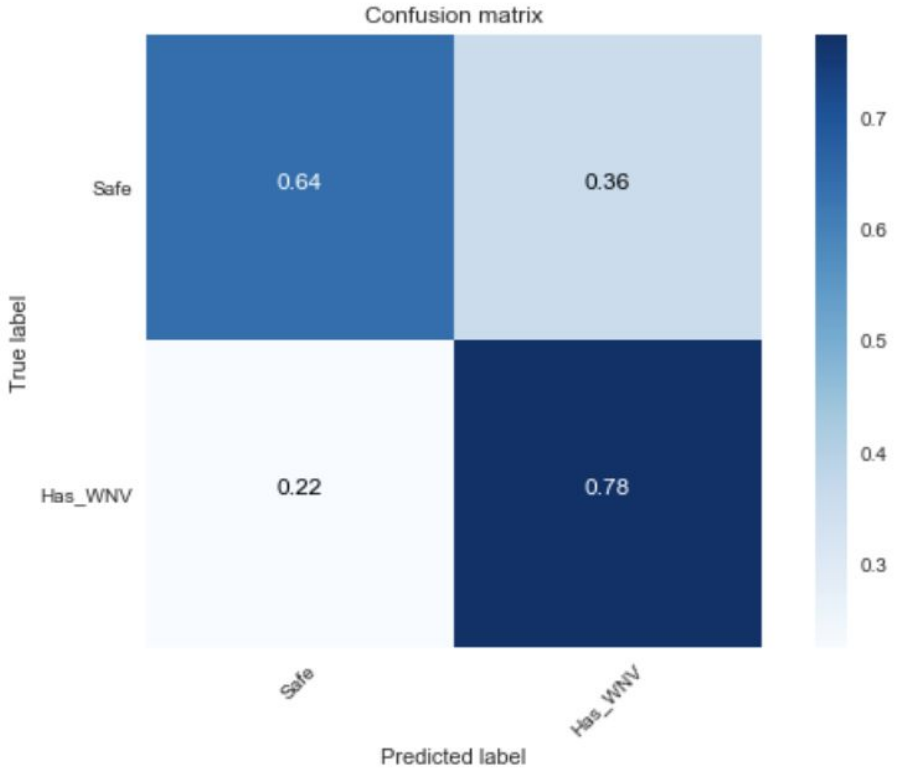
Undersampling datasets:

- N=522/class

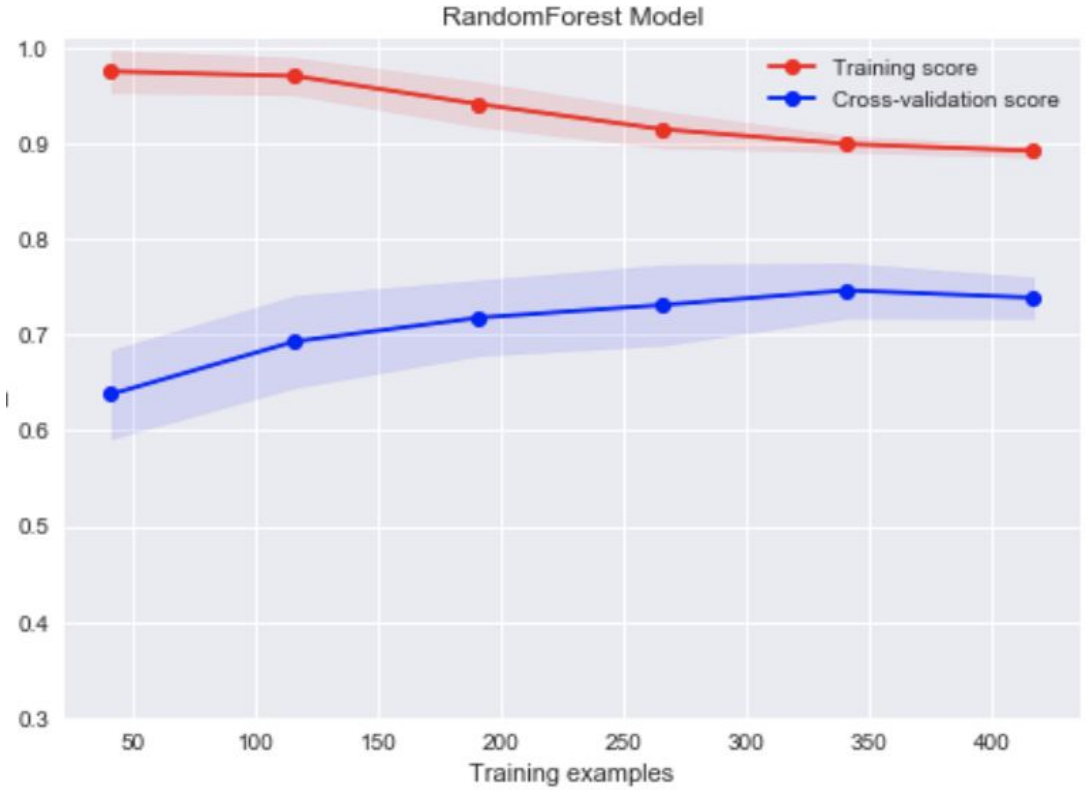
Train:Test ➔ 80:20

Question:

Why isn't the model learning?



RF-Classifier: Learning Curve undersampling



The Problem:

- Learning curves not converging

The Fix:

- Reduce correlated features
- Try Oversampling

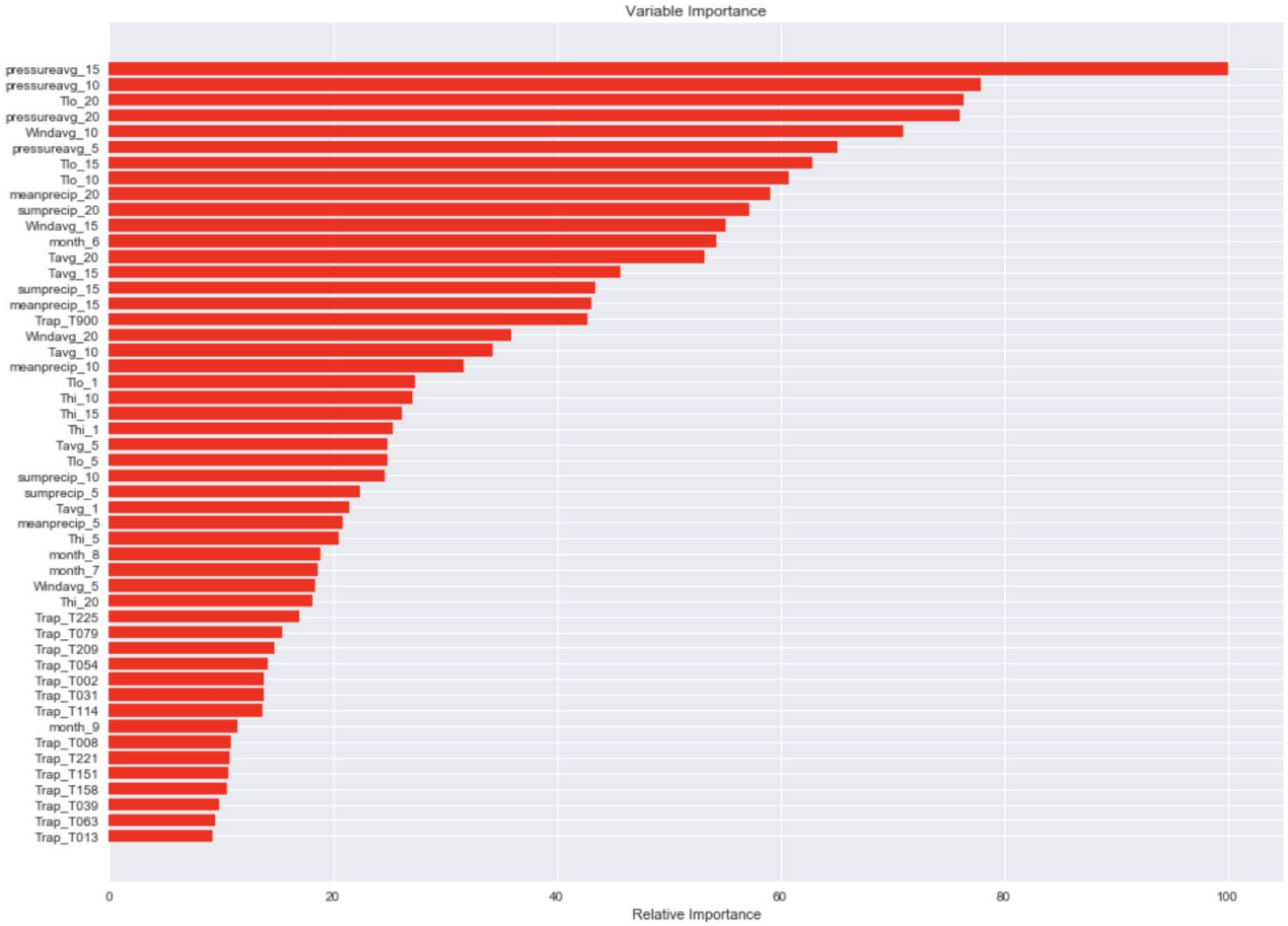
Improved Features

Chose largest importance per weather category

➜ Pressure,Wind,Temp...

➜ Keep Trapid (not very correlated)

➜ Keep seasonal features



Improved Features

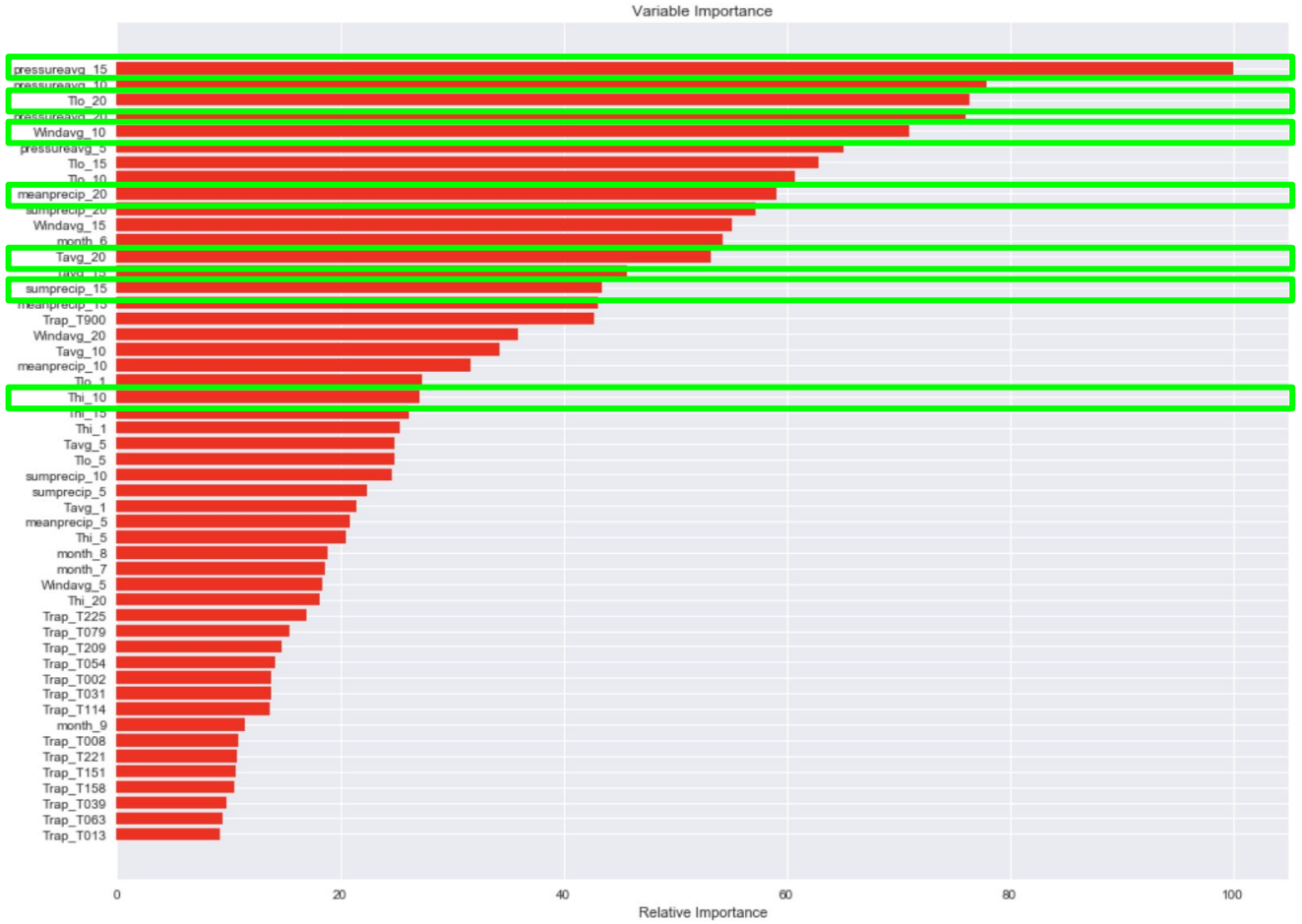
Chose largest importance per weather category

➜ Pressure,Wind,Temp...

➜ Keep Trapid (not very correlated)

➜ Keep seasonal features

-



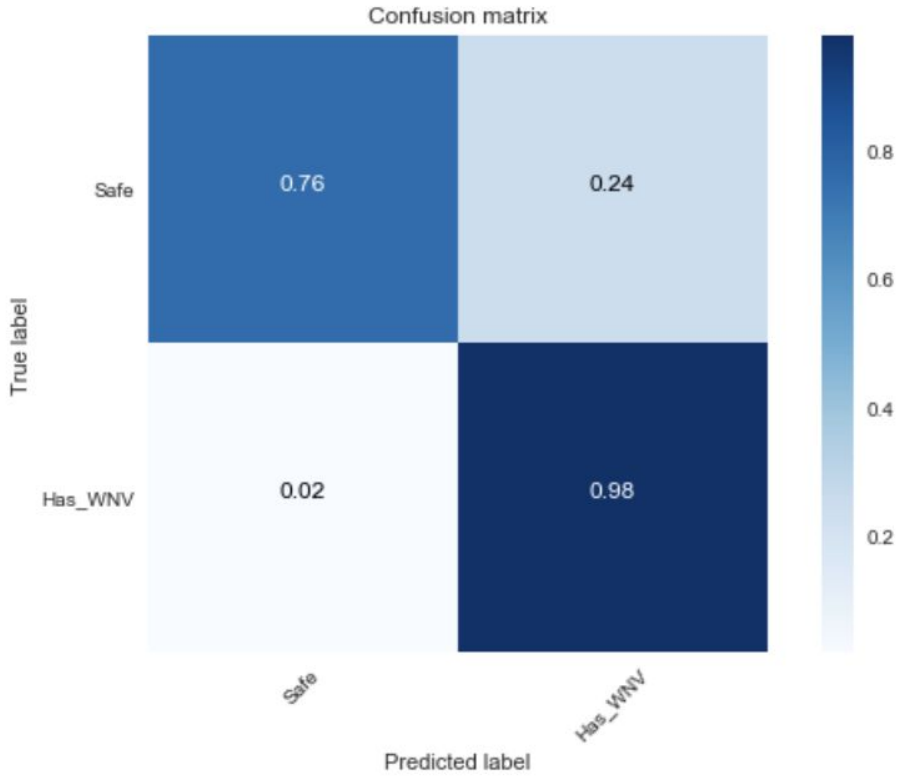
RF-Classifier: Oversampling with Enhanced Features

Recall has been improved.

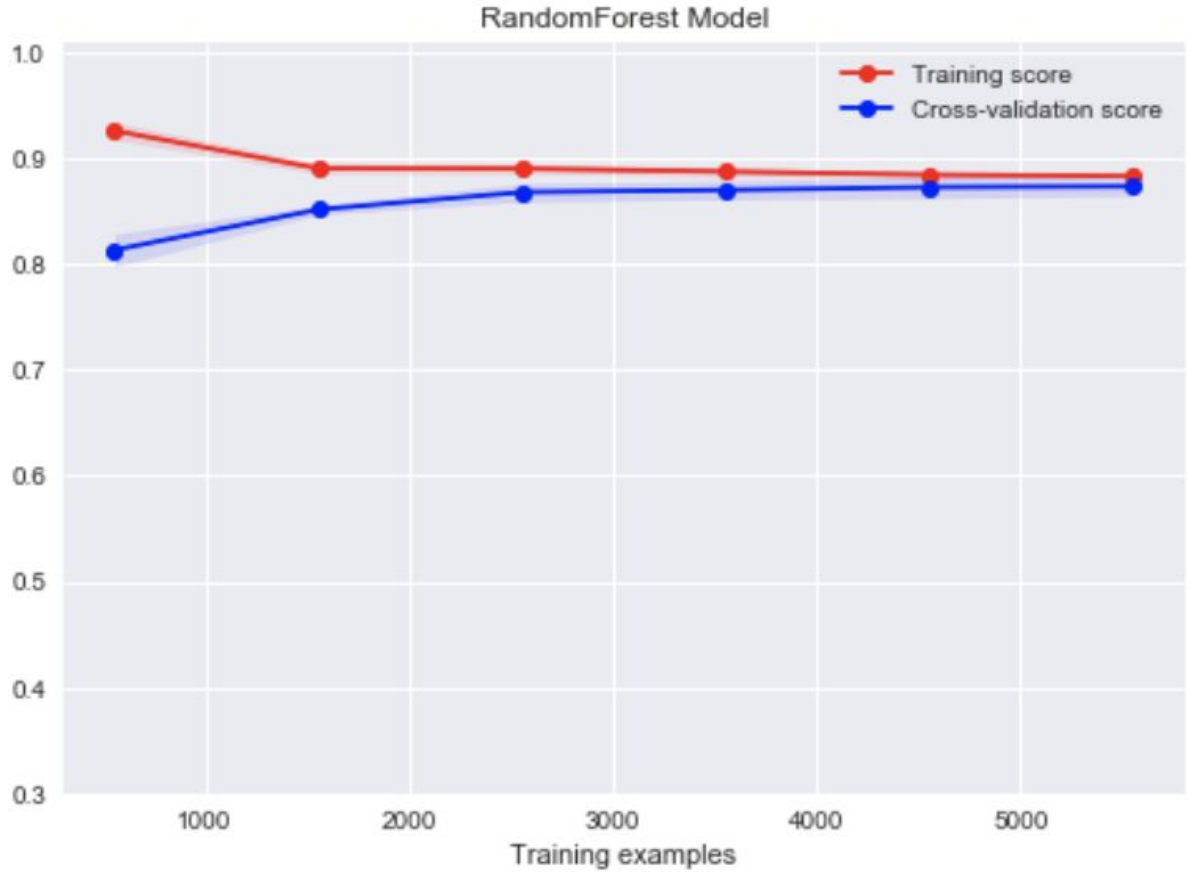
Oversampling datasets:

- N=3471/class

Train:Test ➔ 80:20



RF-Classifier: Learning Curve Oversampling



Learning curves look much better.

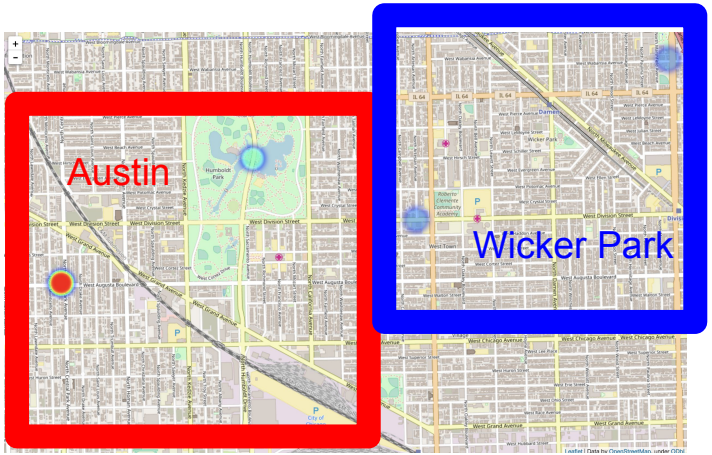
The model can now accurately capture nearly all the traps that have WNV!

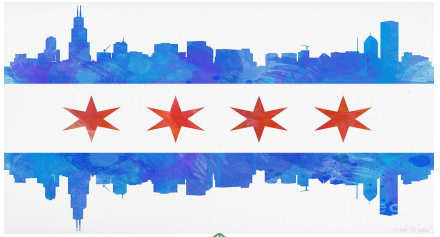
Insights and Beyond

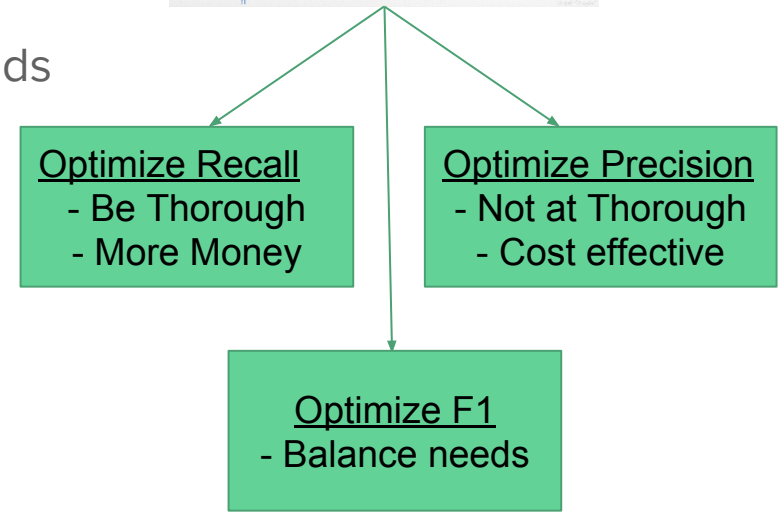
Weather is a powerful first pass predictor for WNV.

Seasonal effects also play a role.

Other useful data could come from neighborhoods







Thanks

INSIGHT

Check me out!
www.RyanGrosso.com

Fermilab

Spares

Top Feature Distributions