Predicting the Outbreak of West Nile Virus in Chicago

Richard Hsu
July 30, 2016

Outline

- Problem
- Data
- Model
- Results
- Conclusion

Problem

- First cases of West Nile Virus were reported in Chicago in 2002
- Predict geographically where there is a high chance that the virus is present

Data

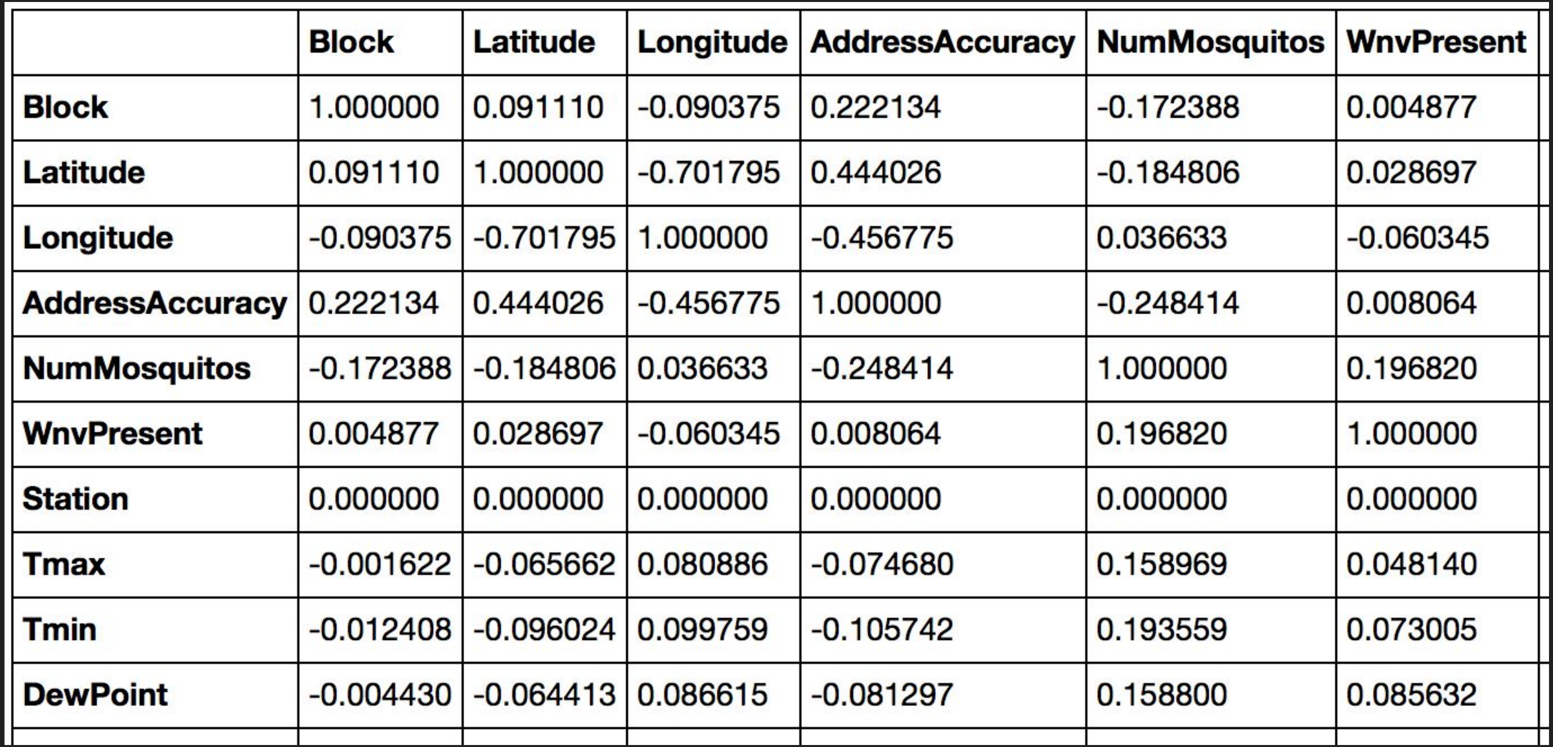
- Mosquito traps were set up throughout the city and each year, from spring to late summer, health workers tested the mosquitos in the traps and determined the species as well as whether or not the virus was present (WnvPresent)
- Weather data was also available from NOAA from 2007 to 2014
- Independent Variables (Weather)
  - Tmax
  - Tmin
  - Tavg
  - PrecipTotal
  - DewPoint
  - WetBulb
  - Heat
- Independent Variables (Main)
  - Address
  - Species
  - Block
  - Latitude
  - Longitude
  - Trap
  - AddressAccuracy

Hypothesis

- According to theory, hot and dry conditions favor the presence of the West Nile Virus vs. cold and wet.
- Therefore, temperature and dewpoint will play an important role in terms of feature importance in deciding whether the virus is present.

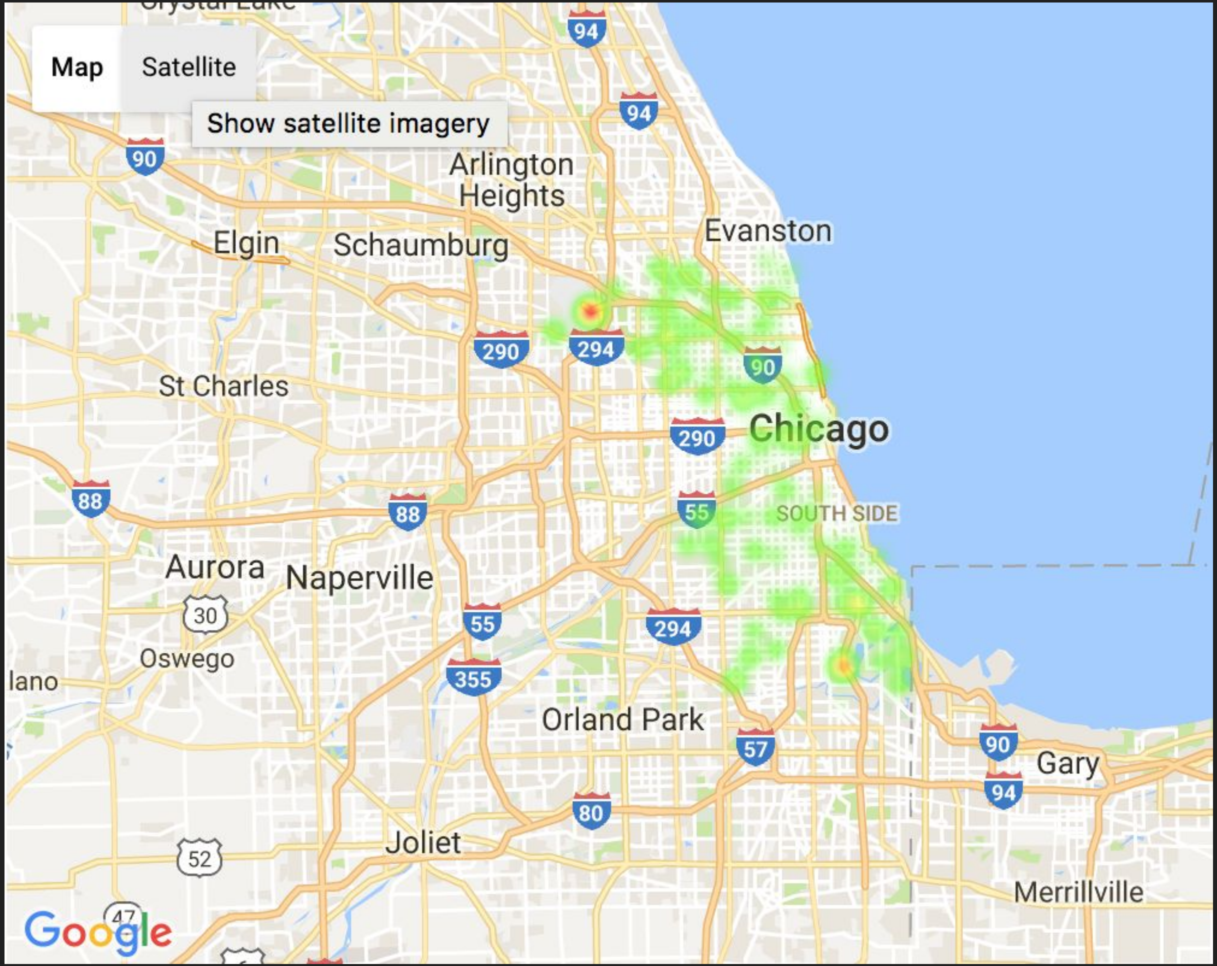
Data Processing

- Weather data needed to be combined with the main dataset and an inner join was carried out
- None of the data columns had missing values

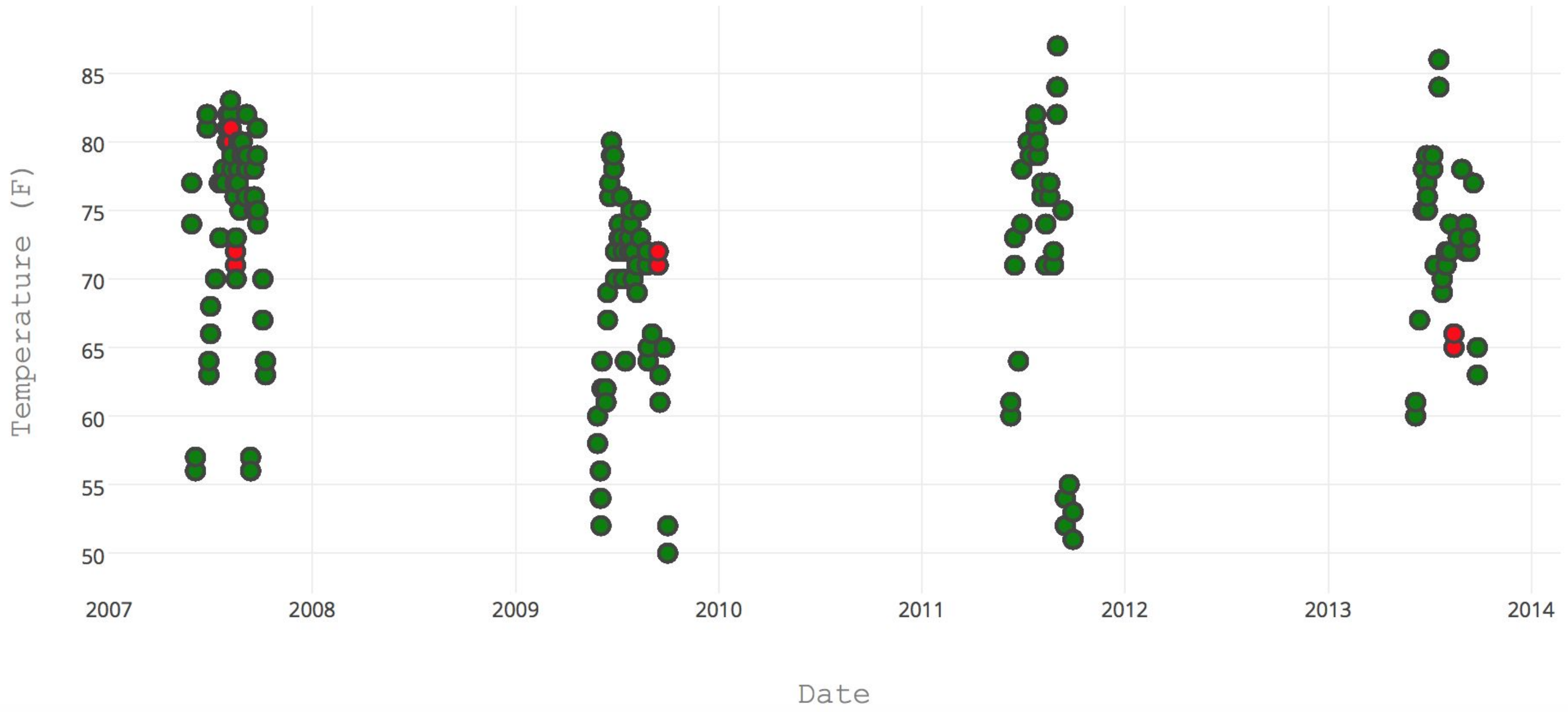


	Block	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent
Block	1.000000	0.091110	-0.090375	0.222134	-0.172388	0.004877
Latitude	0.091110	1.000000	-0.701795	0.444026	-0.184806	0.028697
Longitude	-0.090375	-0.701795	1.000000	-0.456775	0.036633	-0.060345
AddressAccuracy	0.222134	0.444026	-0.456775	1.000000	-0.248414	0.008064
NumMosquitos	-0.172388	-0.184806	0.036633	-0.248414	1.000000	0.196820
WnvPresent	0.004877	0.028697	-0.060345	0.008064	0.196820	1.000000
Station	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Tmax	-0.001622	-0.065662	0.080886	-0.074680	0.158969	0.048140
Tmin	-0.012408	-0.096024	0.099759	-0.105742	0.193559	0.073005
DewPoint	-0.004430	-0.064413	0.086615	-0.081297	0.158800	0.085632

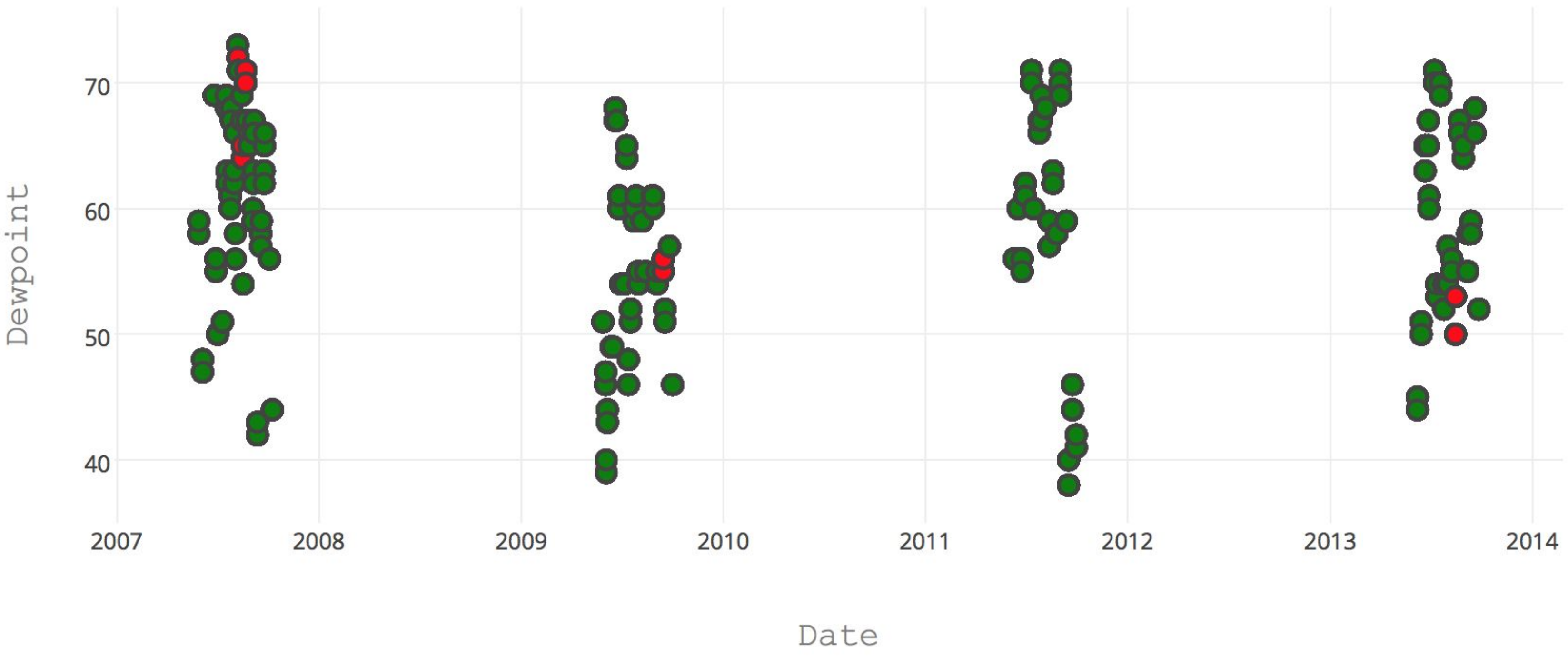
Distribution of Virus Presence From Joined Training Data Set



The Impact of Temperature on the Presence of West Nile Virus



The Impact of Dewpoint on the Presence of West Nile Virus

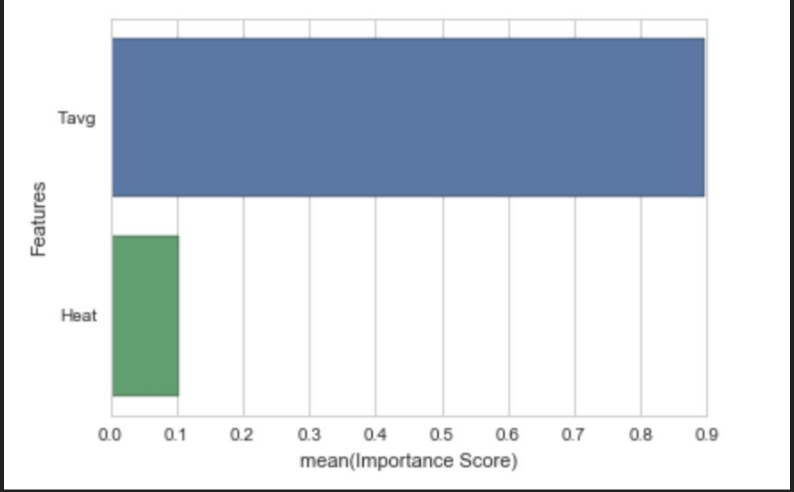


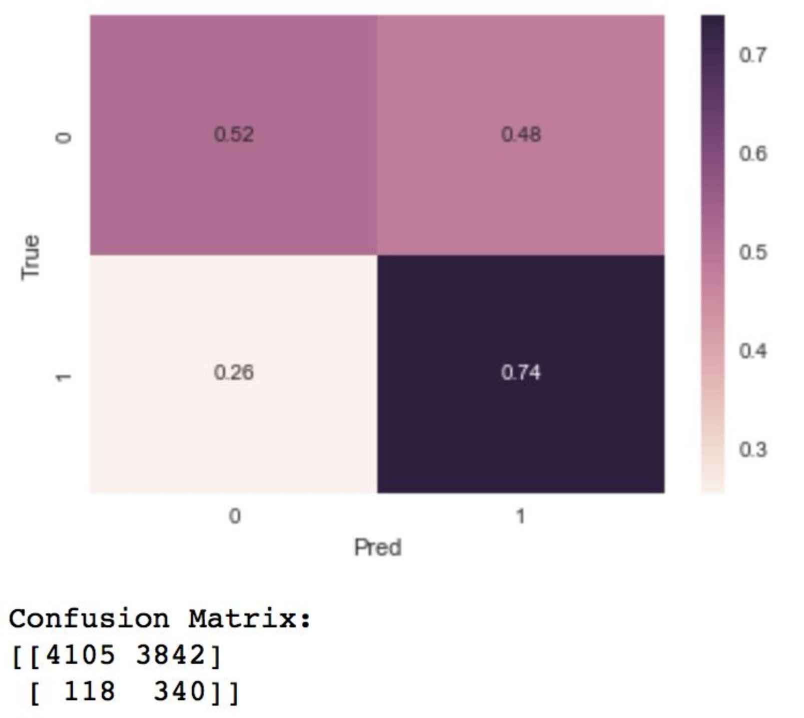
Random Forest

- Large Numbers of Features (Temperature, Dewpoint, Heat, Mosquito Species, Address, etc.)
- Classification
- Easy to overfit

Baseline

- Two Features: Tavg and Heat
- Accuracy of Training: 0.526
- Accuracy of Testing: 0.529

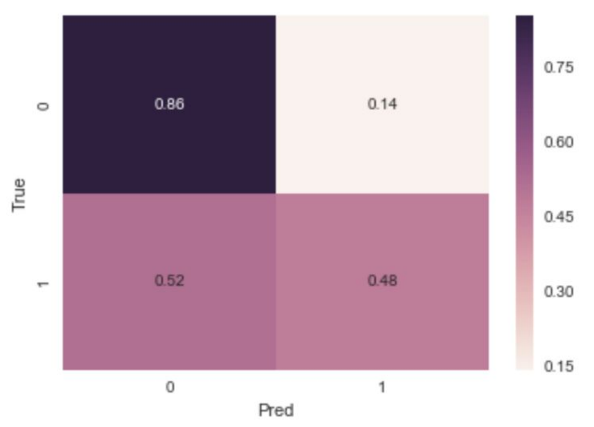




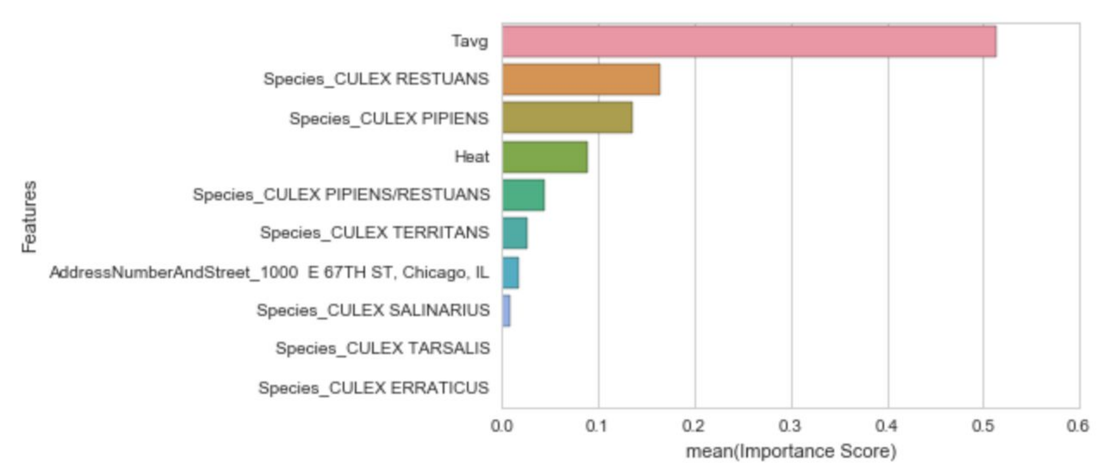
```
Confusion Matrix:
[[4105 3842]
 [ 118  340]]
```

Feature Engineering

- Features: Tavg, Heat, Address, Species
- Accuracy of Training: 0.872
- Accuracy of Testing: 0.837
- ROC AUC: 0.679

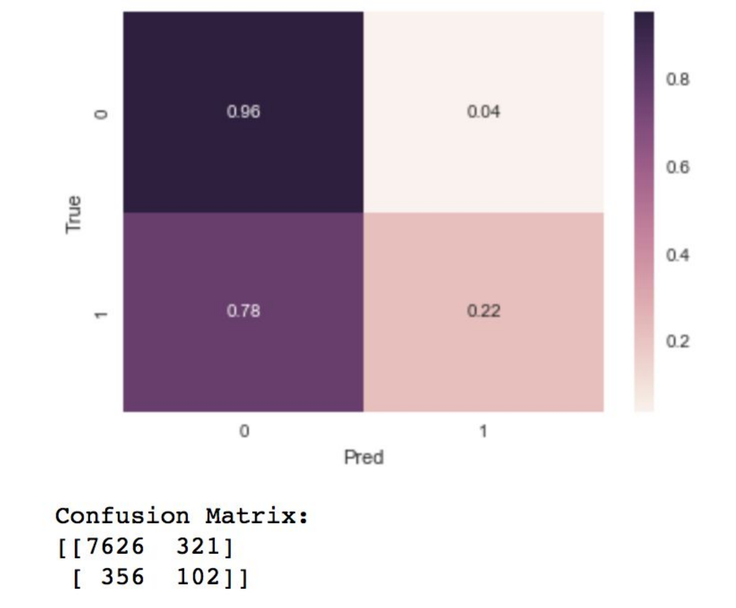


```
Confusion Matrix:
[[6815 1132]
 [ 239  219]]
```

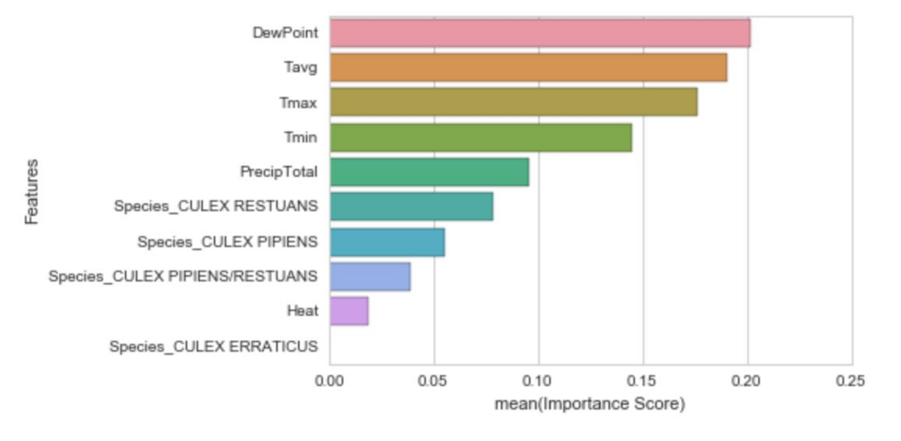


Feature Engineering Cont.

- Features: Tavg, Heat, Address, Species, Dewpoint, PrecipTotal, Tmax, Tmin
- Accuracy of Training: 0.964
- Accuracy of Testing: 0.919
- ROC AUC: 0.784



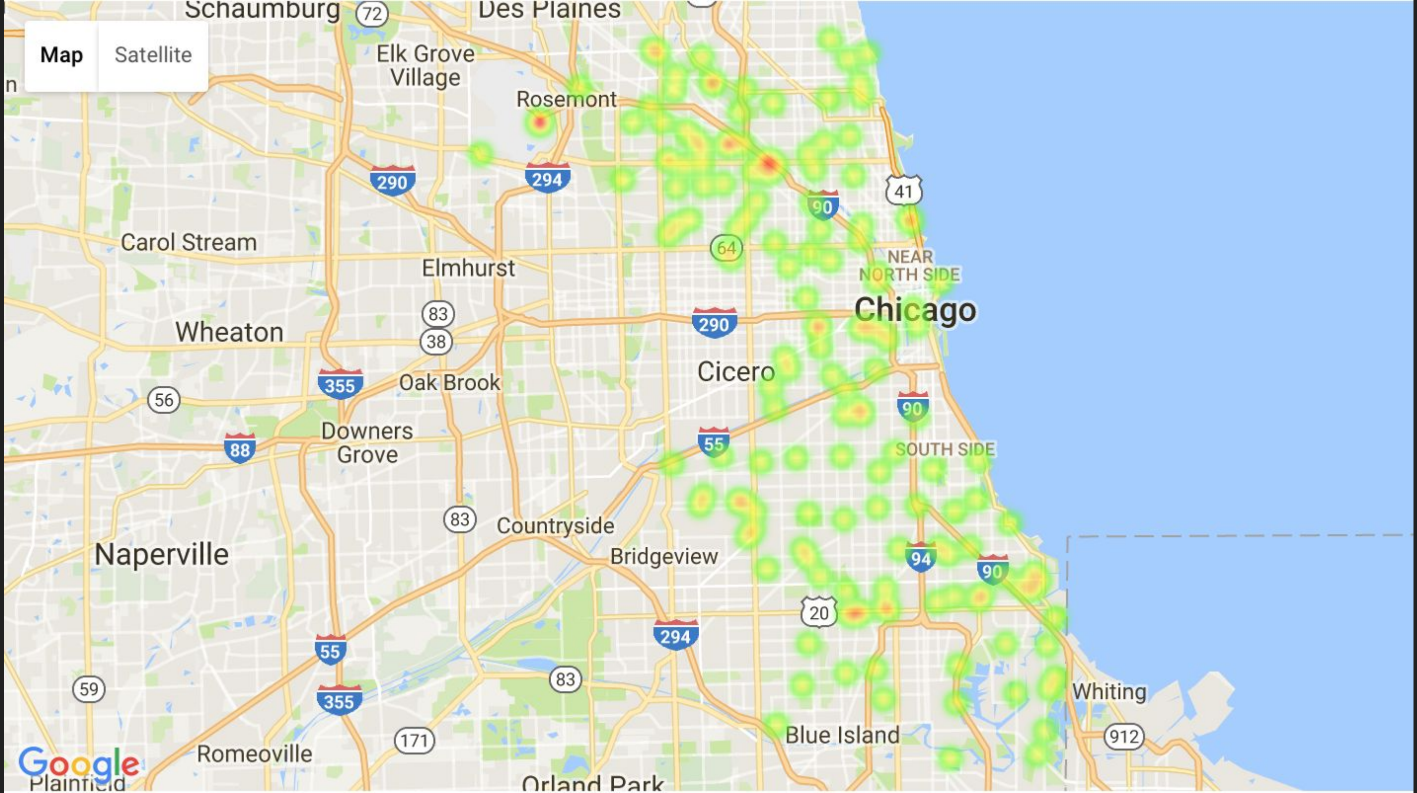
```
Confusion Matrix:
[[7626  321]
 [ 356  102]]
```



Tuning Parameters with Grid Search

- N_estimators - 200
- Max_features - 30
- Accuracy of Training: 0.966
- Accuracy of Testing: 0.914

Prediction (2008, 2010, 2012, 2014)



Conclusion

- A lot of features are involved in determining the presence of West Nile Virus
- The features tell us which are the most relevant: Dewpoint, Tavg, etc.
- The more information you feed into a model, the greater the accuracy

Next Steps

- Since more data is better for the model, we can pull in the mosquito spraying data
- We can look at the number of cases of West Nile Virus reported
- Important for public health planning and allocation of resources to the right place at the right time
- Extend this analysis to studying the Zika Virus

Acknowledgements

- Jim
- Kate and Austin
- Tim
- Everyone in the class
- GA

Citations

- Kaggle Competition: https://www.kaggle.com/c/predict-west-nile-virus/data
- GA - DS-DC-13 (Class Notes)

Any Questions?