West Nile Virus Prediction Milestone Report

Capstone Project of Springboard Data Science Intensive Workshop

Mentor Alan Si

By Ruiye Ni

March 2016

1. Overview

Background

West Nile (WN) virus is a member of flavivirus, and the mosquito is the primary vector. Though infection of WN virus is subclinical in most cases, it would still pose a big threat to human body health if an infection leads to clinical fever and fatal meningoencephalitis (Campbell, Marfin et al. 2002).

Since the first detection of WN virus in New York City in 1999, the virus has been keeping expanding from the eastern parts to the western parts of the USA. Given the potential big threat of WN virus to human health and its ongoing spread, how to effectively prevent the infection of WN virus is particularly relevant for the public health. One way to cope with the future eruption of WN virus is to predict the virus presence in particular areas, and corresponding preventive measures can therefore be taken in the area having high probability of WN virus detection.

In this report, I used dataset released by Chicago Department of Public Health to explore the potential relationship between spatial/temporal information and detection of WN virus in Chicago city (Kaggle 2015). I further built predictive models to forecast future presence of WN virus in a test dataset which is also from Chicago city. Data analysis was implemented in Python2 Jupyter Notebook.

Conclusions

- WN virus detection fluctuated across years in Chicago city.
- WN virus is most likely to be detected in August across years.
- WN virus is more common in species of mosquitos in CULEX PIPIENS and CULEX RESTUANS than other types.

Practical application

The predictive model developed in the project has the potential to help Chicago Department of Public Heath to take timely actions in the areas with predicted high probability of detection of WN virus and effectively prevent the human infection of WN virus in the future.

Limitations

I cannot fully explain why there is such a big difference from year to year based upon the current analysis. Further analysis can be done to explore the impact of weather conditions and spraying effort on WN virus detection given the weather and spray dataset. Other factors, such as the distribution of sensitive human population can be explored if there is such a relevant dataset.

2. Data Model and Database

Datasets

The datasets are composed of four csv files： train, test, weather and spray.

Train dataset has 12 variables as shown in **Table 1** and has 10506 samples, covering data from 2007, 2009, 2011 and 2013. There are eleven predictors variables column 1 ~ column 11 (temporal and spatial information) and one target variable "WnvPresent", indicating whether WN virus was detected at a specific location or not.

Table 1 Train Dataset

Column	Variable Name	Data Type
1	Date	object
2	Address	object
3	Species	object
4	Block	int64
5	Street	object
6	Trap	object
7	AddressNumberAndStreet	object
8	Latitude	float64
9	Longitude	float64
10	AddressAccuracy	int64
11	NumMosquitos	int64
12	WnvPresent	int64

Test dataset has 11 variables as shown in **Table 2** with 116293 test samples, covering data from 2008, 2010, 2012 and 2014. Test dataset's structure is very similar to train dataset, except that it doesn't have target variable "WnvPresent" and one predictor "NumMosquitos".

Table 2 Test Dataset

Column	Variable Name	Data Type
1	ID	int64
2	Date	object
3	Address	object
4	Species	object
5	Block	int64
6	Street	object
7	Trap	object
8	AddressNumberAndStreet	object
9	Latitude	float64
10	Longitude	float64
11	AddressAccuracy	int64

Weather dataset has 22 variables displayed in **Table 3**, which includes various weather metrics from 2007 to 2014.

Table 3 Weather Dataset

Column	Variable Name	Data Type
1	Station	object
2	Date	object
3	Tmax	float64

4	Tmin	int64
5	Tavg	object
6	Depart	object
7	DewPoint	int64
8	WetBulb	object
9	Heat	object
10	Cool	object
11	Sunrise	object
12	Sunset	object
13	CodeSum	object
14	Depth	object
15	Water1	object
16	SnowFall	object
17	PrecipTotal	object
18	StnPressure	object
19	SeaLevel	object
20	ResultSpeed	float64
21	ResultDir	int64
22	AvgSpeed	object

The last spray dataset has 4 variables as shown in **Table 4** and provides the GIS data of spraying efforts in 2011 and 2013.

Table 4 Spray Dataset

Column	Variable Name	Data Type
1	Date	object
2	Time	object
3	Latitude	float64
4	Longitude	float64

3. Data Exploratory Analysis

Yearly WN virus detection

To explore the train dataset, I started with looking into the temporal pattern of number of WN virus across years in train dataset in **Figure1**. WN virus was most active in 2007 with a peak between week30 and week 35. There was an acute decrease in number of WN virus cases in 2009 and level was relatively stable in 2011. A rebound of WN virus incidences can be observed in 2013. Therefore, a general trend was followed in year 2007, 2009, 2011 and 2013 that the WN virus presence was centralized around week 35. Yearly unique dynamics of virus occurrence, however, also clearly exists. For example, 2007 and 2013 have much more virus detection than 2009 and 2011.

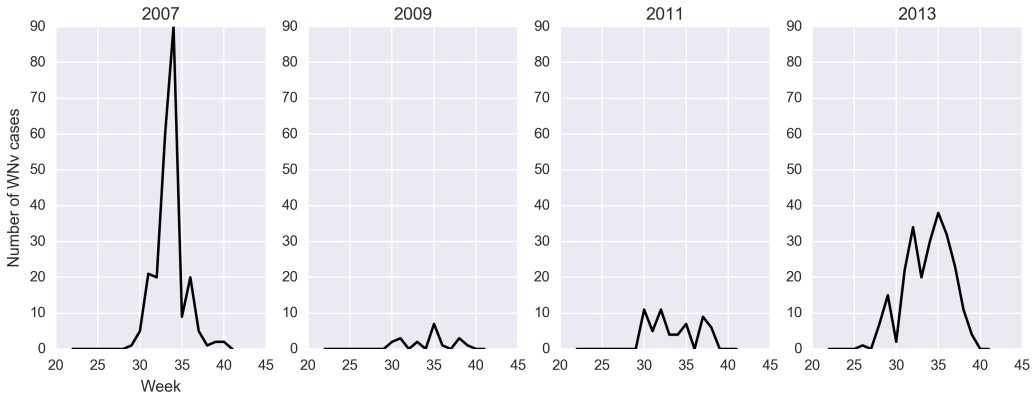

Figure 1 Number of WN virus cases for each year by week in train dataset

The timing trend of WN virus detection is more obvious in **Figure2**. Here, number of WN virus cases for each year is summarized by month. In agreement with Figure1, detection of WN virus reached a peak in August for each year. The observation makes sense since the mosquito is a primary vector of WN virus, and hot and humid weather is favored by mosquito population to multiply.

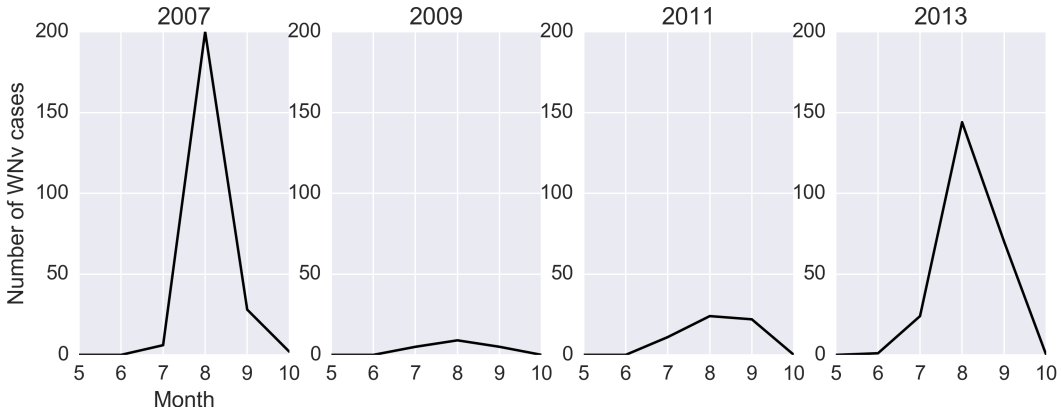

Figure 2 Number of WN virus cases for each year by week in train dataset

Spatial spread of WN virus detection

To further explore the relationship between WN virus detection and the location where the virus was detected, I associated the number of mosquito captured in trap with its location in **Figure 3**. Bigger size of dots in the figure indicates more number of mosquitos. By comparing 2007/2013 with 2009/2011, we can see that 2007 and 2013 have more areas covered by bigger green dots than 2009/2011, and the scattering of green dots are more dense as well.

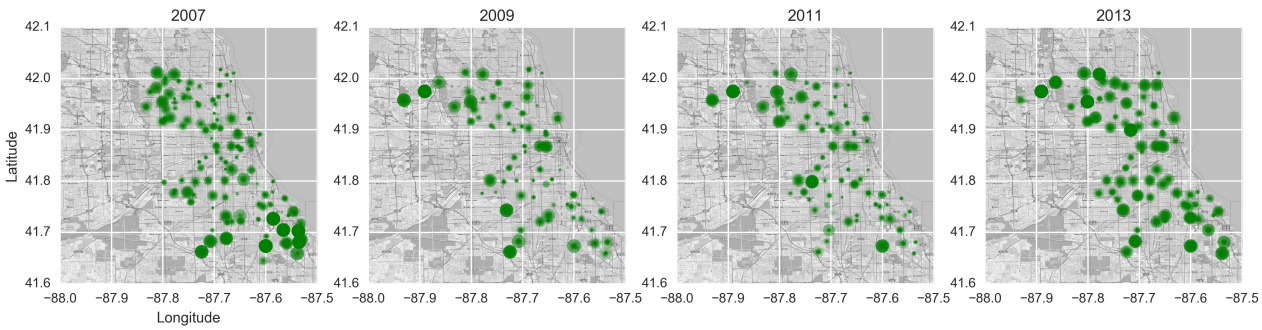

Figure 3 Map of mosquito number for each year in training dataset

Do more mosquitos indicate a higher probability of WN virus detection? I explored this question in **Figure 4**. Consistent with Figure 3, year 200/ 2013 have magenta areas covered than 2009/2011, which indicates a positive association between number of mosquitos and the detection of WN virus. Another interesting trend in Figure 4 is that 2007 and 2013 have cases of WN virus close to the Lake Michigan while the location of WN virus detection in 2009 and 2011 is further away from the Lake Michigan.

Temporal and spatial visualization of WN virus detection was inspired by Kaggle blog(Montoya 2015).

Figure 4 Map of WN virus detection for each year in training dataset

Mosquito types vs. WN virus presence

Last, I explored the relationship between the type of mosquitos and the presence of WN virus. As shown in **Figure 5**, WN virus was detected in mosquito species of CULEX PIPENS and CLUEX RESTUANS. These two species are common throughout much of North America. CLUEX RESUANTS is a native species, and CLUEX PIPENS is emigrated from Europe since 1600s.

(Helbing, Moorhead et al. 2015)

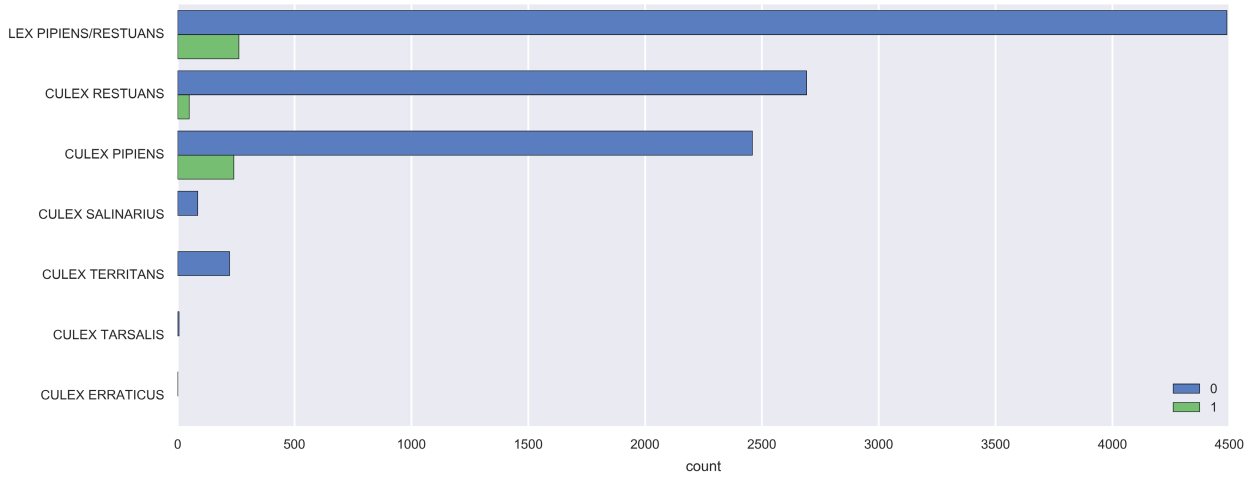

Figure 5 Map of WN virus detection for each year in training dataset

4. Predictive models

Data Cleaning & Checking

- Data Leakage
  Data leakage is the variable in the historically collected dataset that highly informative on the target variable, but is not available for future prediction. In WN virus train dataset, there is a variable named "NumMosquitos" is highly correlated with the detection of WN virus, but in practice we cannot know the number of mosquitos ahead of time. Similarly, the spray dataset will not be available for the test data by the time we make prediction. In order to build a more practical predictive model, I dropped the "NumMosquitos" variable and spray dataset.

- Merge different CSV files
  Since usually we can have weather forecast information, I can still use weather dataset to add more predictors to build models. I merged train/test dataset with weather dataset based upon the "Date" column.

- Replace missing values
  No missing values were detected in either train or test dataset, but there were missing values in the weather dataset. Therefore I don't need to worry about discarding data samples. I simply dropped columns with missing values.

- Transform categorical data
  To cope with categorical data in the train/test dataset, I used one hot encoding to convert categorical variables into binary variables.

Baselines

As a sanity check, I predicted all train dataset as WN virus negative and the results turned out to have an accuracy of 0.94, which means the train dataset is highly imbalanced and the majority cases in train dataset are WN virus negative. Therefore, we can already have very high prediction accuracy without any useful analysis. In order to build a model that is practically useful, we need to evaluate the model performance by precision and recall instead of overall accuracy, and ROC curve can serve this purpose very well.

ROC Curve

I applied multiple predictive models to see which one performs the best. The models I tried including logistic regression, random forest, decision tree, SVM, Adaboost. Five-fold cross validation was implemented to pick the best parameter for each model. Based upon the ROC curve in **Figure 6**, Adaboost and logistic regression models generally performed better than other models.

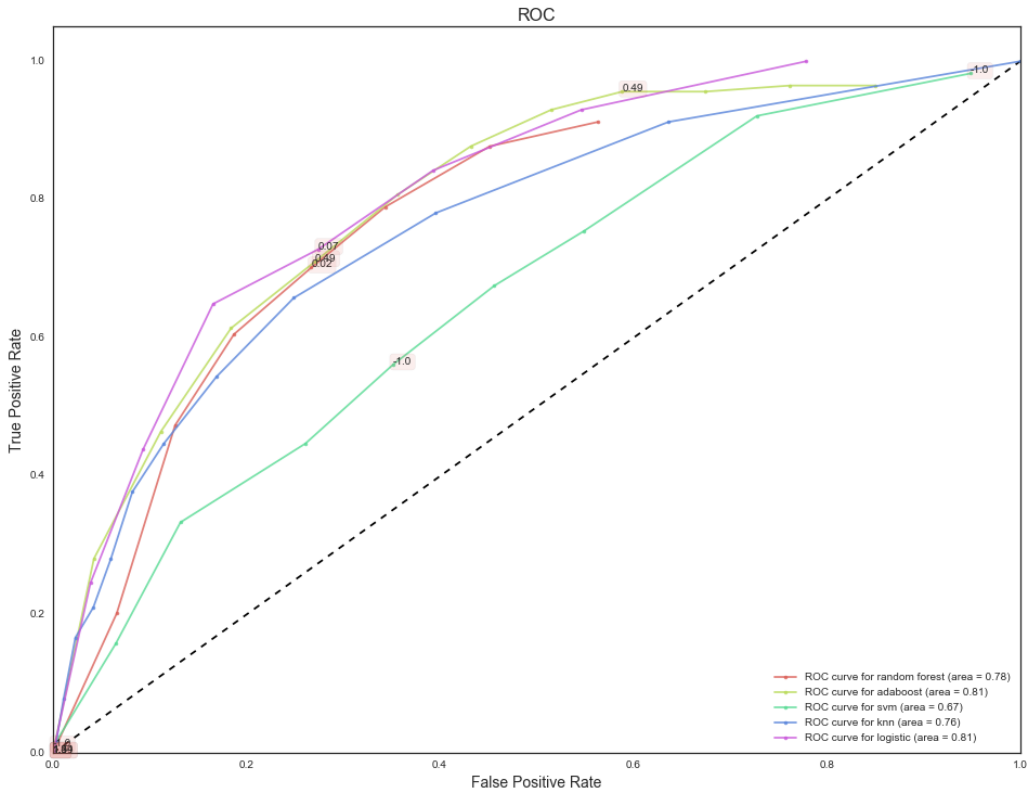

Figure 6 ROC curves of five different models

Timing

Another aspect I looked into to compare different models was the timing each model took for training. As shown in **Figure 7**, random forest and SVM are models that take the longest time to train. In practice, parallel computing is easy to be implemented for random forest model, whereas SVM with nonlinear kernel cannot be used for parallel computing. Therefore, considering both accuracy and timing, logistic regression and Adaboost seem to be reasonable choices for practical use.

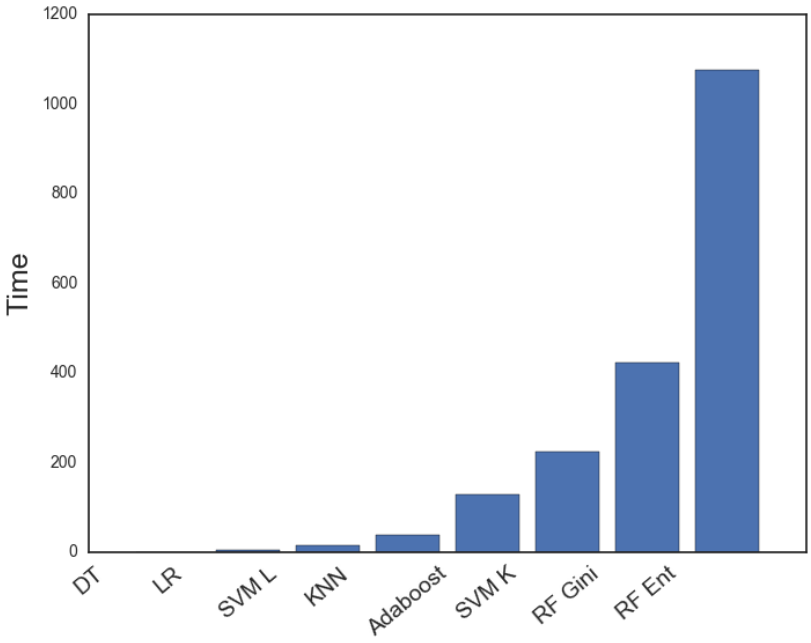

Figure 7 ROC curves of five different models

Prediction

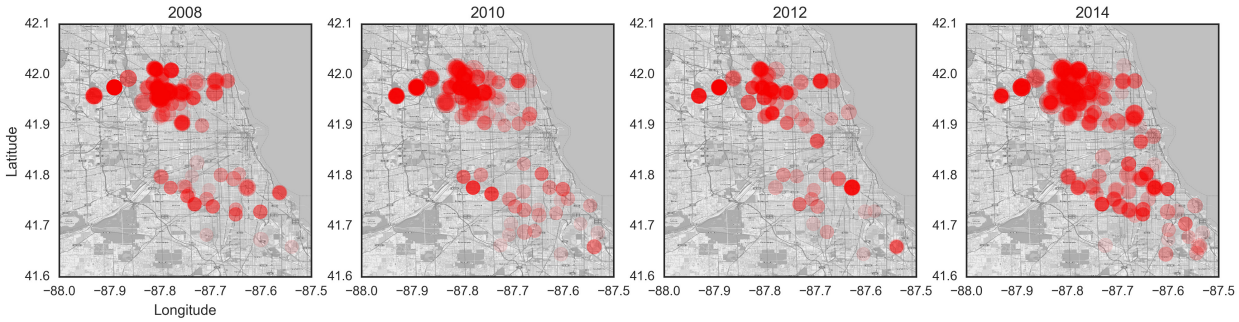

Figure 8 Map of WN virus detection prediction for each year in test dataset

Based upon the model I built in previous sections, I predicted the WN virus detection in **Figure 8**. Areas with higher than 20% probability of detection of virus are shown on the map, and the larger the marker size is, the higher the probability is.

The codes to develop the model were partly reference from a college admission prediction project (David Wihl 2015).

Reference

Campbell, G. L., A. A. Marfin, et al. (2002). "West Nile virus." The Lancet. Infectious diseases **2**(9): 519-529.

David Wihl, M. H., Lauren (2015). "Predicting college admissions."

Helbing, C. M., D. L. Moorhead, et al. (2015). "Population Dynamics of Culex restuans and Culex pipiens (Diptera: Culicidae) Related to Climatic Factors in Northwest Ohio." Environmental entomology **44**(4): 1022-1028.

Kaggle. (2015). "Predict West Nile virus in mosquitos across the city of Chicago." from https://www.kaggle.com/c/predict-west-nile-virus.

Montoya, A. (2015). "Visualizing West Nile Virus." from http://blog.kaggle.com/2015/07/14/visualizing-west-nile-virus/.