West Nile Virus Prediction

Yu Pei
Graduate group in Biostatistics
University of California, Davis
whpei@ucdavis.edu

Abstract—**In this report, we analyzed the mosquitos surveillance data provided by Chicago Department of Public Health(CDPH) to identify West Nile virus(WNV) hotspots. A more accurate method of predicting outbreaks of West Nile virus in mosquitos will help the City of Chicago and CPHD more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus. Data provided includes weather, location of traps and spraying efforts. In the modeling efforts, we extracted several weather, spatial and temporal features based on exploratory analysis and builded several prediction models, including logistic regression, gradient boosting machine and random forest. We were able to get 0.750 area under the ROC curve(AUC), currently rank in the top 20% among all participants.**

I. Introduction

A. Background

West Nile virus is most commonly spread to humans through infected mosquitos. Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death. In 2002, the first human cases of West Nile virus were reported in Chicago. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today. Every week from late spring through the fall, mosquitos in traps across the city are tested for the virus. The results include the number of mosquitos, the mosquitos species, and whether or not West Nile virus is present in the cohort. Identifying outbreak hotspots will help the City of Chicago and CPHD more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

The organization of the report is as follows, we first present some data exploratory results and discuss the features we used. Then we state the models we tried with the corresponding results. Finally we discuss the results we got and point out some future works.

B. Data Exploratory Analysis and Feature Engineering

The training set consists of data from year 2007, 2009, 2011, and 2013, while test set contain data for year 2008, 2010, 2012, and 2014. There are four parts for the provided data:

1) Main dataset: each entry represents one time point for a given trap location and a given mosquito species. There are total of 7 species present, about 150 unique locations. In the training set we have number of mosquito counts and the true outcome(whether the virus is present in that trap). When the number of mosquitos exceed 50 in a trap, they are split into another record (another row in the dataset), such that the number of mosquitos are capped at 50. For test dataset, we don't have number of mosquito information and the true label.
2) Spraying efforts: The City of Chicago started spraying pesticide from 2010 to reduce the number of mosquitos in the area, and therefore might eliminate the appearance of West Nile virus. We are provided with spraying data for year 2011 and 2013. Currently the models didn't use this information.
3) Weather: Daily weather measurements from two weather stations in Chicago, obtained from National Centers for Environmental Information(NOAA). Relevant information includes temperature, humidity, wind speed etc.
4) Map data: map data for the city of Chicago, primarily for use in visualizations.

It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet, we first matched weather information to each data point based on location and date.

Second, we can see from the heatmap (Fig. 1) that there is a strong correlation between the number of mosquitos and WNV presence. We can also observe a clustering effect for the presence of WNV, indicating spatial closeness could be a strong predictor as well.

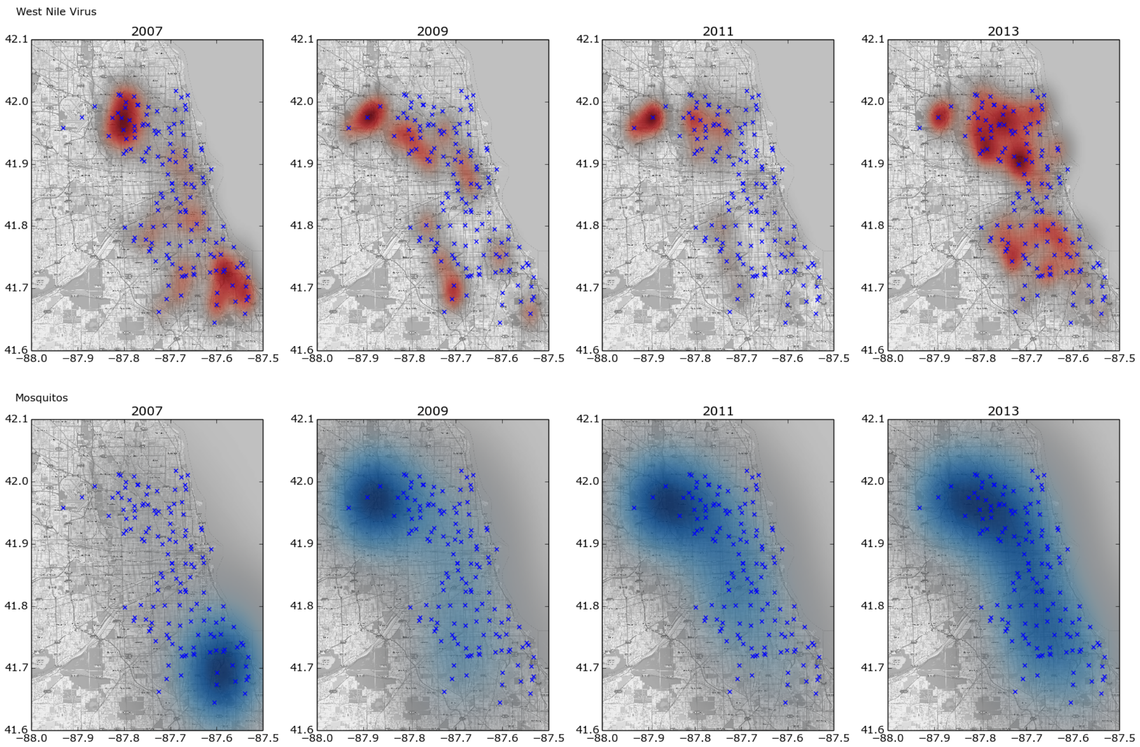

Fig. 1: Heatmaps of WNV incidences(upper) and number of mosquito counts(lower) by year. Blue dots are the location of traps(created by user Neil Summers)

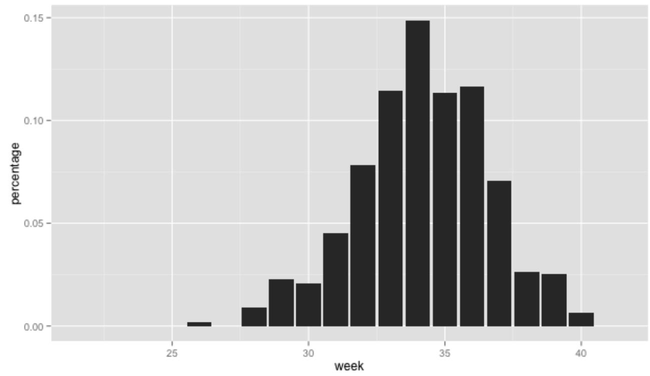

Fig. 2: Proportion of entries that has WNV present by weeks of the year

TABLE I: WNV Incidence by Species

Species	Proportion of WNV	Case counts
CULEX ERRATICUS	0.00	1
CULEX PIPIENS	0.09	2699
CULEX PIPIENS/RESTUANS	0.06	4752
CULEX RESTUANS	0.02	2740
CULEX SALINARIUS	0.00	86
CULEX TARSALIS	0.00	6
CULEX TERRITANS	0.00	222

We can see a strong time factor too. More cases happened during August and September than in the early summer as shown in Fig. 2.

Finally we can find both in literature[1] and in the training dataset (Table I) that different species of mosquitos has different ability to transmit WNV virus.

Noted that number of mosquitos in each trap is an importance predictor but is not included in the testing dataset. We imputed the variable based on time, location and species using K-nearest neighbors(KNN). And k is a hyper-parameter we can tune.

II. Methods and Results

In this section we present several classification algorithms' results. To assess the performance of a particular model and parameter combination, we performed cross validation(CV) on each available year, i.e. we train on three years data then predict the remaining years' outcome. This way we can have a general idea of how well the model will perform for test dataset.

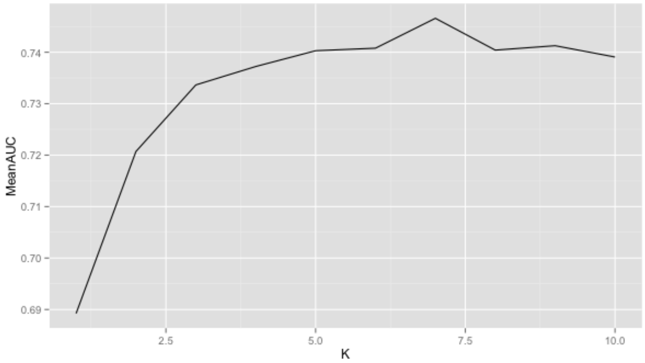

Fig. 3: The number of neighbor K effect on the mean AUC across all years

A. GLM

The first and the easiest model to try would be logistic regression. But we can see from the cross validation table II that it performs consistently worse than the other two models, except for year 2011. This is expected since some of the features has non-linear relationship with the outcome(e.g latitude and longitude). Although it doesn't perform well, it can be used to test new features' effect quickly and is very robust. Leaderboard gave 0.68 AUC, which is very close to what we got from CV results.

B. Random Forest

Random Forest[3] is a very popular ensemble method that can be used in many cases(regression and classification). The tree structure means it can have non-linear interactions between variables. The ensemble with random variables subsample can decouple the trees and stabilize the result. There are many parameters we can tune, like tree depth, number of trees to build and weighting of samples etc. Since the number of mosquitos depends on the KNN algorithm for imputation, K for number of neighbors is also a parameter. Here we show the effect of different K on mean AUC from 4 years CV result(Fig. 3), and the effect of tree numbers(Fig 4). Notice that the number of trees doesn't correlate with better performance, this is due to the randomness in the process. Based on the result, we choose $k = 7$ and number of trees to be 1800.

C. Adaboost

Adaboost[2] is also a tree ensemble method but after building a tree, it will build the next tree with re-weighted data. It is more powerful but might overfit the data. We can see from Table II that boosting tree gave a better result than random forest in CV, but uploading into leaderboard the boosting tree performed worse than random forest, So it is possible that we overfitted the data.

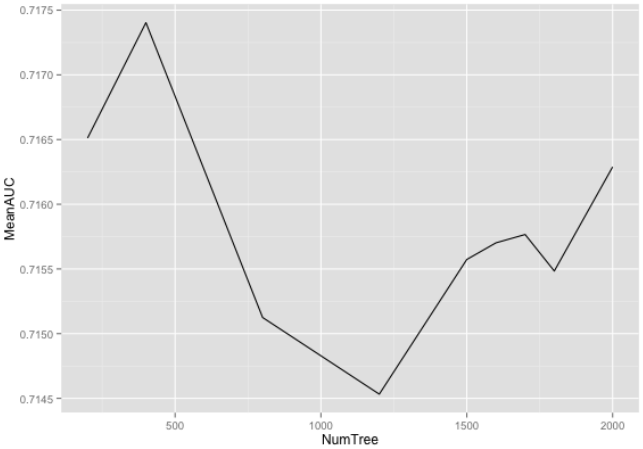

Fig. 4: The number of trees' effect on the mean AUC across all years

TABLE II: Model Cross Validation Results

Year	GLM	RandomForest	Adaboost
2007	0.5772	0.6249	0.6848
2009	0.6662	0.7682	0.8177
2011	0.7908	0.7124	0.7650
2013	0.6784	0.7568	0.7190

III. Conclusion And Future work

In this section we first list things we have tried but didn't work, then we will outline the major modeling efforts to improve the prediction accuracy. Finally I will point out future working direction that might improve the prediction accuracy.

What I have tried but didn't work:

1) The organizer padded the testing data so test data contain all the species in all the traps, but most of them are artificial records. So we tried to count how many record there are for each location and date. More than one record means the species present in the trap. we added the count as feature but that didn't help.
2) We also tried to use some time series features to capture the time dependency of adjacent test results. We calculated the moving mean, standard deviation and slope for a local linear fit. They didn't improve performance though.
3) Accounting for the spraying efforts. Since it is likely that spraying will decrease the number of mosquitos and it is an important feature in out

model, we tried to increase mosquito counts post-spraying.

What I have added in the process that improved the prediction accuracy:

1) First is noticing the interaction between variables, so instead of adding interaction manually in logistic regression, I switch to tree methods that can handle this automatically.
2) Imputing the important feature number of mosquitos helped a lot. This is confirmed by the importance index outputted by random forest.
3) The increase in imputation accuracy by adding species also gave a huge bump in AUC.

For future directions, here are some ideas we might explore in the future:

1) We can perform model diagnosis to identify where it performed poorly and try improve the model.
2) Try different imputation methods, since this is a spatial data, we might try kriging to see if it will produce better results.
3) We can refine the features we have right now, convert raw data like week and month, latitude and longitude to more useful features.

Overall this is a very rewarding experience and I have learned a lot in the process.

Acknowledgment

The authors would like to thank many fellow competitors in the forum, who have provided many helpful guidances and insights into the data.

References

[1] Michael J. Turellc, David J. Dohmc, Michael R. Sardelisac, Monica L. Oguinnc, Theodore G. Andreadisbc, and Jamie A. Blow An Update on the Potential of North American Mosquitoes (Diptera: Culicidae) to Transmit West Nile Virus Journal of Medical Entomology 42(1):57-62. 2005

[2] Y. Freund and R.E. Schapire (1997) A decision-theoretic generalization of on-line learning and an application to boosting Journal of Computer and System Sciences, 55(1):119-139.

[3] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.