**Submission Model and Cost Benefit Analysis**

Kaggle West Nile Virus Prediction Project

Fei Cai • Rich Enik • Andy Gonzalez • Joe Klein

Introduction

We aim to predict the presence of West Nile Virus infected mosquitoes in traps positioned across Chicago’s Cook county using weather data provided by the National Oceanic and Atmospheric Administration (NOAA) and fumigation efforts data from the City of Chicago. Using the predictions of our model, we will estimate a cost-benefit analysis that juxtaposes spraying for vector mosquito species based on model predictions with the added cost of a West Nile Virus outbreak.

Method

1. Cleaning data and feature selection
2. Model building
3. Cost Benefit Analysis
4. Cleaning data and feature selection

We began by parsing over the different data sets provided as training data from the City of Chicago and the weather and spray data as mentioned above. After performing a thorough Exploratory Data Analysis we decided to simply the variables used by eliminating those that measure the same/similar features:

- **Location of the traps**, having 7 different columns measuring some aspect of it, was reduced to a single variable containing a complete street address, and fed as dummy variables in the training data.

- Similarly, the column describing the **species of the mosquitoes** was used in the model as dummy variables for the different values of species.

- **Temperature** also had several columns measuring the same variable. We opted to use average temperature for the entry date.

- Day length had both **sunrise** and **sunset** as features in the model.

- **Rain** was quantified as total precipitation.

- **Wind** was represented as average wind speed for the entry date.

- and, **Humidity** was expressed as relative humidity, or the percentage of water vapor needed to reach saturation.

1. Model Building

Considering the imbalance of the classes/responses between mosquito traps testing West Nile Virus positive and testing negative, we decided to use the ensemble method of Random Forest. To further enhance the predictive power of our model, we normalized our variables and applied the Extreme Gradient Boosting method (XGBoost), which generalizes the model by allowing optimization of an arbitrary differentiable loss function.

To find the best parameters with which to tune the model, we used a Gridsearch Cross Validation where we tested for improved ROC-AUC scores by trying alternative changes in the maximum depth of the decision trees, weight of child nodes, learning rates, and number of estimators used.

1. Cost Benefit Analysis
2. Determine Cost
3. Calculate Benefits
4. Compare Alternative
5. Report and Plan Action
6. Determine cost:

To determine the cost of a disease you need a robust effort in which you combine medical/hospital data with survey data of those who contracted the disease. This will give us an idea of medical and abatement costs, as well as, loss of productivity cost that comes as a result of missed work and other behavioral changes.

The data does not exist for the Chicago area, so to address the lack of data we will be looking at similar analyses of costs associated with West Nile Virus outbreaks as they affected Louisiana in 2002, Sacramento, CA and Colorado in 2003 We particularly leaned on the work of Zohrabian A, Meltzer MI, Ratard R, Billah K, Molinari NA, and Roy K, in their analysis titled “West Nile virus economic impact, Louisiana, 2002” (Emerg Infect Dis. 2004;10:1736–44). Their work was emulated because they had the biggest sample size and easier to convert methods. The formulas were adjusted for inflation and to correspond with the patient breakdown as it happened during the 2012 West Nile Virus outbreak in Illinois. This outbreak was chosen as a basis on comparison because it offered a more extreme cost scenario than the alternatives. The cost analysis also takes into account the difference in cost of living between Illinois, and the basis of the formulas used, Louisiana.

Based on the decisions explained above, the cost breakdown is as follows:

**Summary of costs 2012 West Nile virus outbreak in Illinois**

| **Cost category** | **Cost ($ millions, adjusted for 2012)** | **% of total epidemic cost** |
| --- | --- | --- |
| Medical costs |  |  |
| Acute care | 5.17 |  |
| Inpatient rehabilitation facility | 0.55 |  |
| Inpatient treatment subtotal | 5.72 | 26.5 |
| Outpatient hospital care | 0.04 |  |
| Visits to medical doctors | 0.19 |  |
| Outpatient rehabilitation facilities, equipment | 0.35 |  |
| Outpatient treatment subtotal | 0.58 | 3 |
| **Total medical costs** | **6.30** | 29 |
| Nonmedical costs |  |  |
| Mortality | 7.77 |  |
| Morbidity | 1.45 |  |
| Productivity losses subtotal | 9.22 | 42.5 |
| Nursing home | 0.07 |  |
| Transportation, miscellaneous | 0.13 |  |
| **Total nonmedical costs** | **18.64** | 43.5 |
| **Total cost of illness** | ***24.94*** | ***73*** |
| Mosquito control | 5.2 |  |
| **Total public agency cost** | ***5.2*** | ***27*** |
| **Total cost of epidemic** | ***30.14*** | ***100*** |

d) Report and Plan Action:

Given weather and location data, we are able to predict the incidence of West Nile Virus in Chicago with a 76.3% accuracy, according to results provided by Kaggle.com in the competition’s website. Although exact location of where the virus will appear cannot be predicted with our model, we can suggest the coordination of mosquito abatement efforts and public awareness campaigns according to our work by using weather and trap test history. This will reduce the possibilities of an outbreak and save over $18 million, when compared to years with low incidence of the virus.

These recommendations are particularly pertinent in protecting the most vulnerable to this virus: Our elderly and children. Research into patient history of the disease shows that people over 65 years of age who contracted the disease had an increase chance of developing the more serious West Nile Neuroinvasive Disease and more likely to die from it. Programs designed to protect these communities when our model predicts appearance of West Nile, will further maximize the cost saving potential of our work.

Considering the potential human life and economic cost, 2012 efforts of $5.2 million for mosquito control falls short. However, the City of Chicago already has a multiyear contract that funds up to $135 million in mosquito abatement services. While that seemingly disproportionate number cover a larger area than the scope of this project, we are confident that the use of our model would greatly aid in keeping the use of those funds to a minimum.