West Nile Virus Prediction-Kaggle Competition

Rafel Taye

DAT7-General Assembly

1. INTRODUCTION

West Nile virus is most commonly spread to humans through infected mosquitos. People who are infected with the virus develop range of symptoms ranging from persistent fever, to neurological illnesses that can result in morbidity. In 2002, the city of Chicago experienced the first human case of West Nile virus. By 2004 the city and Chicago Department of Public Health (CDPH) implemented a surveillance and control program to combat this virus that is still in effect today. A more accurate method of predicting the West Nile virus in mosquitos will greatly assist the city of Chicago and CPHD combat an outbreak by efficiently and effectively allocate their resources. I intend to provide a more accurate prediction of when and where different species of mosquitos will test positive for West Nile virus given weather, location, testing, and spraying data from the city of Chicago.

Every week from late late-May to early-October, mosquitos throughout the city are trapped and tested for the virus. Every week from Monday through Wednesday, these traps collect mosquitos, and the mosquitos are tested for the presence of West Nile virus before the end of the week. The test results include the number of mosquitos, the mosquito species, and whether or not West Nile virus is present in the collected population. I will be analyzing weather data and GIS data and predicting whether or not West Nile virus is present, for a given time, location, and species. The results of these tests influence how the city allocates its resources in combating this virus by spraying airborne pesticides.

2. DATA

Data provided are from the city of Chicago are:

2.1 Main dataset: The test results obtained are organized in such a way that when the number of mosquitos exceeds 50, they are split into another record (another row in the dataset), such that the number of mosquitos is capped at 50. Each row cannot have more than 50 mosquitos for a given location. The location of a trap is described by the block number and street name and is mapped into Longitude and Latitude as a feature in the dataset. These are derived locations. For example, Block=79, and Street= "W FOSTER AVE" has an approximate address of "7900 W FOSTER AVE, Chicago, IL", which translates to (41.974089, -87.824812).

Some traps are "satellite traps". These are traps that are set up near (usually within 6 blocks) an established trap to enhance surveillance efforts. Satellite traps are postfixed with letters. For example, T220A is a satellite trap to T220. Please note that not all the locations are tested at all times. Also, records exist only when a particular species of mosquitos is found at a certain trap at a certain time. In the test set, it is being asked for all combinations/permutations of possible predictions and is only scoring the observed ones.

2.2 Spray Data: The City of Chicago also performs regular spraying to eliminate mosquitos. We are given the GIS data for their spray efforts in 2011 and 2013. Spraying can reduce the number of mosquitos in the area, and therefore might eliminate the appearance of West Nile virus. The diagram below illustrates the different mosquito traps located through out the city (blue dots) and the airborne pesticide spray locations (red shade) on the city map.

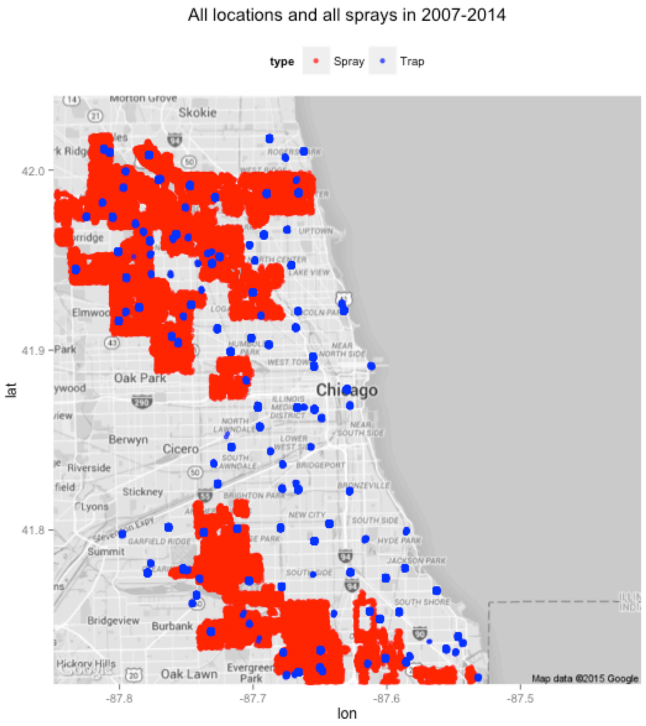

Figure 1-City of Chicago image taken from Kaggle

3.1 Weather Data: It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet. We are provided with the dataset from National Oceanic and Atmospheric Administration (NOAA) of the weather conditions of 2007 to 2014, during the months of the tests. The weather information provided is only for two locations through out the city, the northern and southern part of the city. This presents a challenge when trying to combine the weather data with the main data set. I have to calculate the shortest distance between a mosquito trap location and the two weather locations, and assign the weather information of station to which the trap location is the closest. I plan to use a library called "Geopy" to calculate geodesic distance between two points using the Vincenty distance. The calculation is based on the assumption that the figure of the earth is a spheroid, and hence more accurate than methods such as Euclidean which assumes a straight distance. The two weather stations are listed below

Station 1: CHICAGO O'HARE INTERNATIONAL AIRPORT Lat: 41.995 Lon: -87.933 Elev: 662 ft. above sea level
Station 2: CHICAGO MIDWAY INTL ARPT Lat: 41.786 Lon: -87.752 Elev: 612 ft. above sea level

2.4 Map Data: The map files mapdata_copyright_openstreetmap_contributors.rds and mapdata_copyright_openstreetmap_contributors.txt are from Open Street map and are primarily provided for use in visualizations.

3. FILE DESCRIPTION

3.1 train.csv & test.csv

3.1 train.csv & test.csv - the training and test set of the main dataset. The training set consists of data from 2007, 2009, 2011, and 2013, while in the test set you are requested to predict the test results for 2008, 2010, 2012, and 2014.

- Id: the id of the record
- Date: date that the WNV test is performed
- Address: approximate address of the location of trap. This is used to send to the GeoCoder.
- Species: the species of mosquitos
- Block: block number of address
- Street: street name
- Trap: Id of the trap
- AddressNumberAndStreet: approximate address returned from GeoCoder
- Latitude, Longitude: Latitude and Longitude returned from GeoCoder
- AddressAccuracy: accuracy returned from GeoCoder
- NumMosquitos: number of mosquitoes caught in this trap
- WnvPresent: whether West Nile Virus was present in these mosquitos. 1 means WNV is present, and 0 means not present.

3.2 spray.csv - GIS data of spraying efforts in 2011 and 2013

- Date, Time: the date and time of the spray
- Latitude, Longitude: the Latitude and Longitude of the spray

3.3 weather.csv - weather data from 2007 to 2014.

3.4 **SampleSubmission.csv** - a sample submission file in the correct format

4 DATA PROCESSING & EXPLORATION

The data was provided in a neat csv file, so there was no major data pre-processing necessary. I decided to do feature engineering because I thought there other factors could be predictive of West Nile virus. Looking at the number of species present could be an indicative of West Nile virus presence.

CULEX PIPIENS/RESTUANS	4752
CULEX RESTUANS	2740
CULEX PIPIENS	2699
CULEX TERRITANS	222
CULEX SALINARIUS	86
CULEX TARSALIS	6
CULEX ERRATICUS	1

As a preliminary exploration, I plotted the Mosquito species Vs West Nile Present on a scatter plot to see if there is any correlation. Indeed, only three of the seven different mosquito species carry the West Nile virus. Therefore we know mosquito species could be a good predictor feature

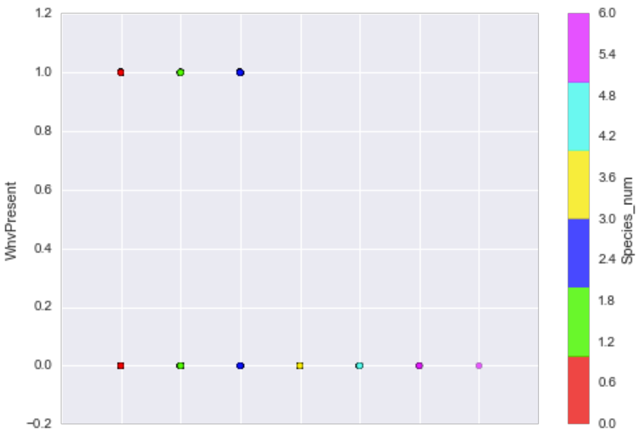

Figure 2: Mosquito species vs Virus Present

I also created a heat map of the correlation between the features for the train data set.

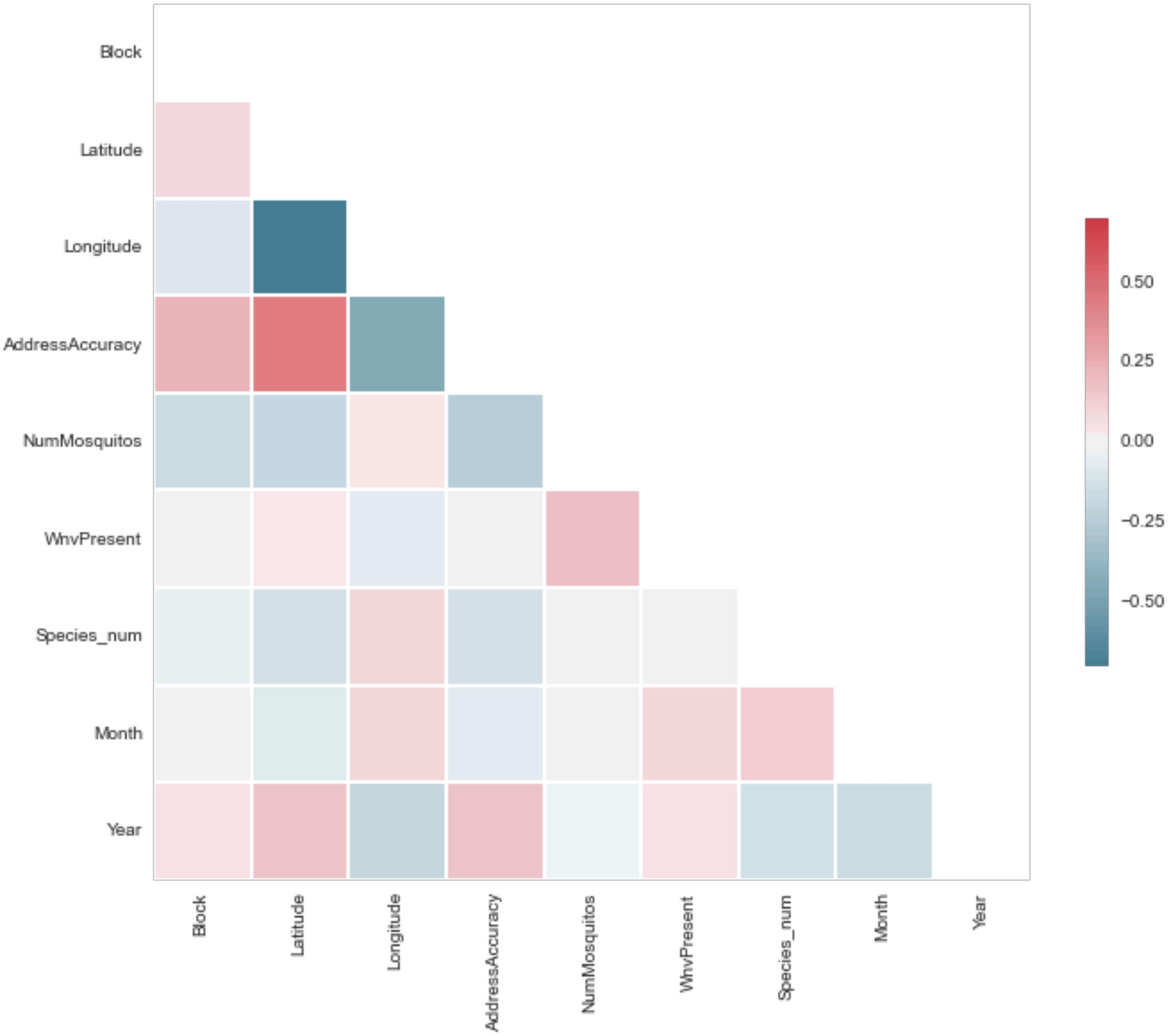

Figure 3: Heat Map of feature correlations

It is hard to tell what feature has strong correlation with presence of the West Nile virus. Further data exploration and weather data merge could be key to unlocking key features. To merge the weather data, I did some exploration and found that some of the columns are string objects and have some values marked ‘M’ for missing values and ‘T’ for trace amount. I replaced these values with zero and changed string objects to float columns to allow me to do some more visualization using correlation heat map and scatter matrix plots as follows.

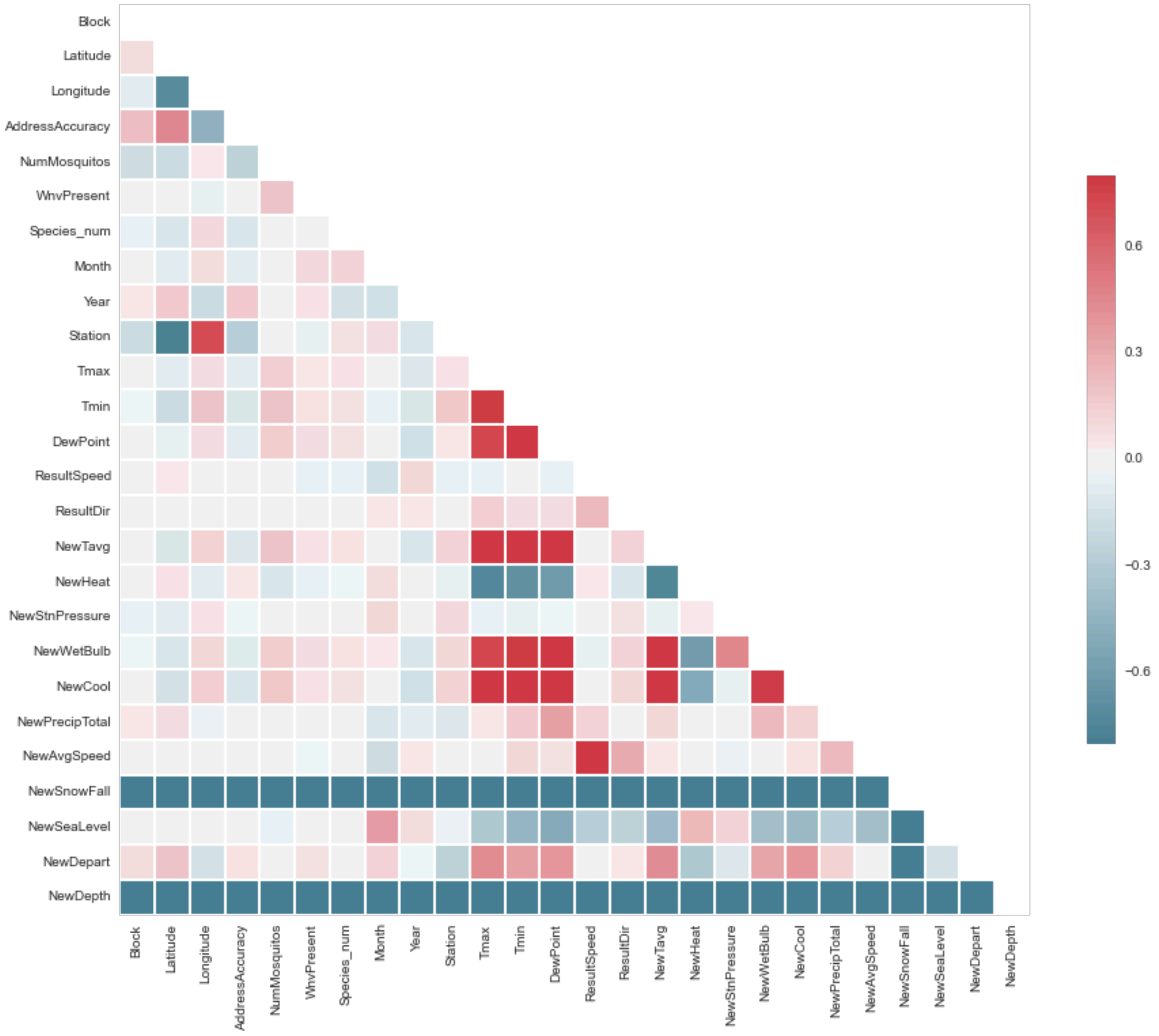

Figure 4- Correlation heat map of combined weather and train data sets

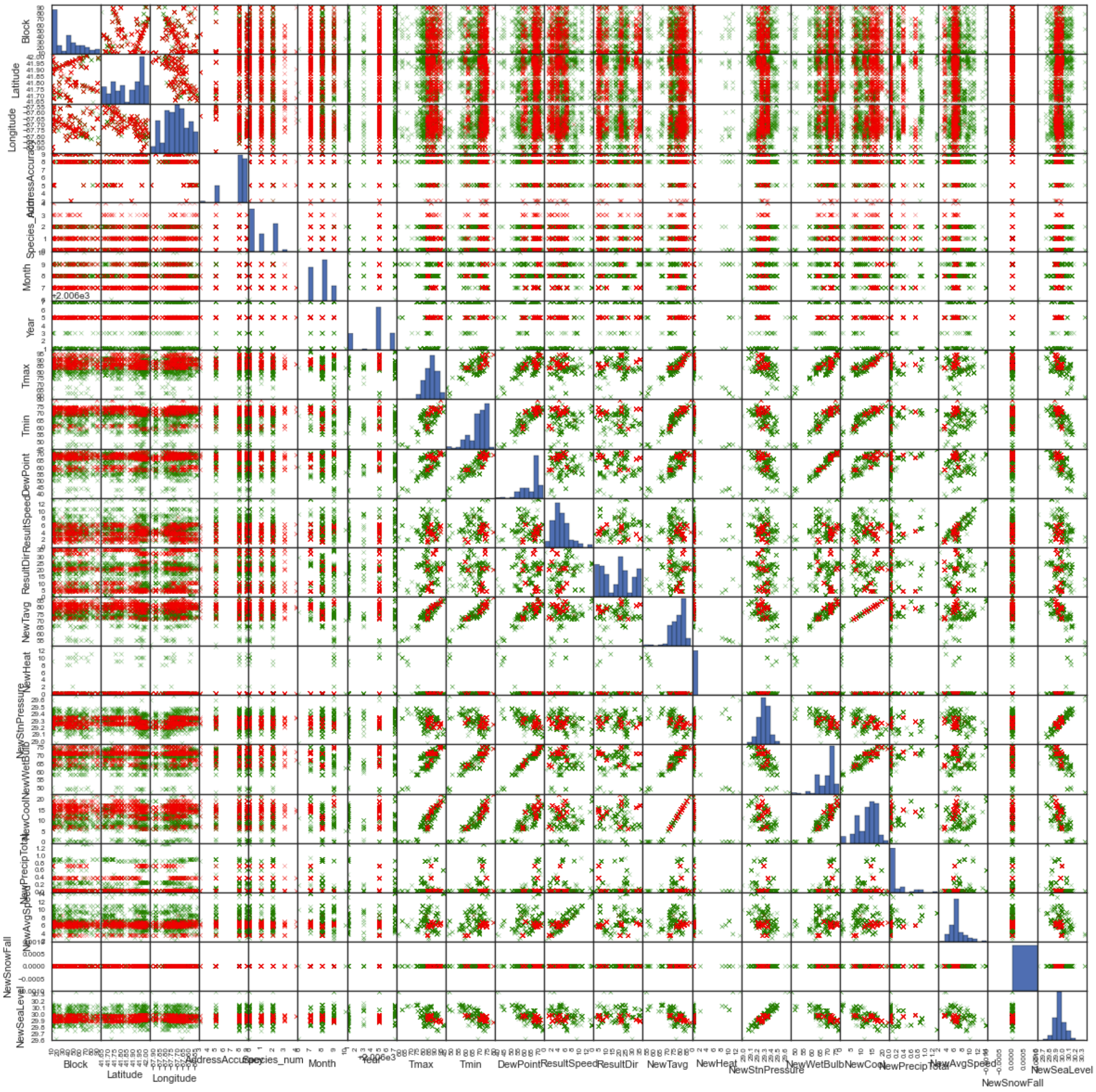

Figure 5- Scatter Matrix Plot of features vs WnvPresent

Using the above two visualizations I selected the following features as my baseline for the selected modeling techniques.

Month,Tmax, Tmin, DewPoint, NewTavg, NewWetBulb, NewCool

5. **MODELING AND VALIDATION:** The following three modeling techniques were tested with the selected baseline features

5..1. **Logistic Regression:** this modeling technique was relatively easy to use than the other ones. With minimal tuning, the ROC-AUC score when I was cross validating came out to be around 0.8. When submitting to the kaggle website I received a score of 0.56 which is significantly lower that my cross validation score. However, I was able to beat the baseline score of 0.5 on kaggle leader board.

5..2. **Decision Tree:** I initially thought this would provide me with the better score. Indeed when I was cross validating my model using the training data, I was getting an ROC-AUC score of close to 0.98. This was a significant improvement compared to the Logistic Regression Model. However, I received a significantly lower score (around 0.4) when submitting to the kaggle completion.

5..3. **Random Forest Classifier:** Random Forest classifier was my next try. This model is known to be robust and minimizes errors in predicting. This model performed better than the Decision Tree model as expected, but performed worse than the baseline score received by using Logistic Regression Model.

Challenges

- Cleanly merging new data frame with the train data set
- Engineering features column names not standard with the other ones, which gave me a hard time visualizing and modeling
- Selecting the best predictor features
- Calculating distance to nearest weather station and assigning all weather attributes to a mosquito trap location
- Parameter tuning for the different modeling techniques

6. CONCLUSIONS & KEY LEARNINGS

I learned a great deal when doing this project. Performing a through data exploratory and having a deep understanding of what the data is trying to communicate is a big part of the battle. Spending a lot of time on feature engineering and selection should be the next priority. These two stages will allow models to learn the predictors better hence providing a better machine learning performance. A special attention should be paid to significant class imbalance as this might throw off your models ability to predict accurately, this should be fixed by balancing the data set to have roughly the same amount of class. I will continue this project in the coming few weeks refining and tuning my modeling techniques and paying special attention to selecting the best predictor features and hopefully come close to the leading score of 0.86.