**West Nile Virus Prediction**

Introduction

This project is engaged with the spread of West Nile virus (WNV) in Chicago, USA.

WNV first arrived in the Western Hemisphere in 1999 (1) and during the summer and fall of 2002, an epidemic (223 cases) and epizootic of WNV infections occurred in Chicago (2). It is most commonly spread to people by the bite of an infected mosquito. Cases of WNV occur during mosquito season, which starts in the summer and continues through fall. There are no vaccines to prevent or medications to treat WNV in people. Fortunately, most people infected with WNV do not feel sick. Yet about 1 out of 150 infected people develop a serious, sometimes fatal, illness (3).

**Transmission**

Mosquitoes become infected when they feed on infected birds, which circulate the virus in their blood for a few days. During later blood meals (when mosquitoes bite), the virus may be injected into humans and animals, where it can multiply and possibly cause illness.

The virus may also be transmitted through contact with other infected animals, their blood, or other tissues (4).

**Symptoms**

Most people infected with WNV do not develop any symptoms (80%). About 20% of the infected people will develop a fever with other symptoms such as headache, body aches, joint pains, vomiting, diarrhea, or rash. Most people will recover completely, but fatigue and weakness can last for weeks or months. However about 1 in 150 people who are infected will develop a severe illness affecting the central nervous system. The severe illness symptoms are characterized with neck stiffness, stupor, disorientation, coma, tremors, convulsions, muscle weakness, vision loss, numbness and paralysis. Severe illness can occur in people of any age; however, people over 60 years of age are at greater risk. The highest risk for getting severely ill when infected with WNV is found among immunocompromised people (for example, transplant patients) or other medical conditions(3).

**Diagnosis**

When symptoms as described above appear, in area known or suspected to be infected with WNV it is recommended to see a healthcare provider and to be tested for MNV infection. The diagnosis is generally accomplished by testing of serum or cerebrospinal fluid (CSF) to detect WNV-specific IgM antibodies (3).

**Treatment**

There is no vaccine or specific antiviral treatments for WNV infection.

Clinical management is supportive (3). Patients with severe illness will require close monitoring, pain control and other treatment depending on their symptoms.

**Prevention**

In the absence of a vaccine, prevention of WNV disease depends on community-level mosquito control programs to reduce vector densities, personal protective measures to decrease exposure to infected mosquitoes (3).

**Project Objectives**

This project is aim to predict when and where different species of mosquitoes will test positive for WNV in traps found around Chicago, USA in the given years: 2007, 2009, 2011 and 2013.

**Project Overview**

WNV is a disease that can easily become epidemic. It has no cure and can be lethal in certain risk groups. The resources used for fighting WNV should focus on prevention. The prevention in this case is to eradicate mosquitoes, the disease vector.

Prevention efforts can be done on the personal level and on the community level.

* Monitoring infected mosquitoes (and birds, see below) can provide an indication to when and where the eradicating efforts should take place, in aim to control the disease spreading.

The spread of mosquitoes is highly dependent on weather conditions. Elevated temperature together with the presence of fresh or stagnant water can increase the spread of mosquitoes. For instance, in some mosquitoes’ species life cycle can be 14 days at 21° C but will take only 10 days at 27° C (5). An increase in the number of virus spreading mosquitoes may occur between 2 weeks after and up to 2 months after a hurricane. Especially in areas that did not flood but received more rainfall than usual (6). On the other hand, strong winds will drive mosquitoes away. For those reasons, much of the analysis in this project is based on weather records.

In this project, no reliable data was found related to reporting dead birds or regarding monitoring infected birds although birds are known to be carrier of WNV. In other geographical areas around the word and later years, dead or infected birds monitoring is a common practice regarding WNV monitoring.

Available records regarding spraying are integrated in the data as well as hospitalization records.

Data

The data in this project was based on 2 main sources:

1. Dataset of mosquitoes catch and presence of West Nile Virus (WNV table).

2. Dataset of weather observations from 2 local climatological data stations.

The first table was consisted of the following features:

Date, Address of the Trap, Species of mosquitoes, Block, Street name, Trap number, Full Address, Latitude, Longitude, Address Accuracy, number of Mosquitoes, number of mosquitoes caught in this trap and WnvPresent (presence of WNV).

Each observation was dedicated to mosquito species caught in a certain trap on a specific date and whether or not it was positive to the presence of WNV.

The second Table was consisted of different weather observation. Some are quantitative, like temperature, barometric pressure, precipitation and a variable of different weather descriptive phenomena such as fog or thunderstorm.

Data Processing

At first, the two tables were treated separately. In WNV table, from the date-variable: A month, month in year, and a season variables were driven. Number of mosquitoes-caught per date and per trap was united, while keeping the differentiation between Mosquitoes species.

The outcome variable – “WnvPresentB” was defined as: whether or not (1 / 0) a WNV was tested positive among the all mosquitoes-caught in a trap in a given date. The **amount** and **species** of the mosquitoes were specified for each observation and thus play a rule in the prediction models.

The weather data contain observation from 2 separate weather station, the data was divided by those station and the two dataset were treated in the exact the same manner. The variable of the descriptive weather phenomena was transformed to a binary indicator variable for each phenomenon.

The three tables were then joined based on the dates; in this stage, variables without differential distribution were removed, redundancy of weather variables (within and between the two stations) were tested by Pearson correlation test (table 1) and pairs- plot visualization (see fig.1). Variables found to be highly correlated were removed.

Table 1: a correlation matrix using Pearson correlation test of the weather variables (temperature related mainly).

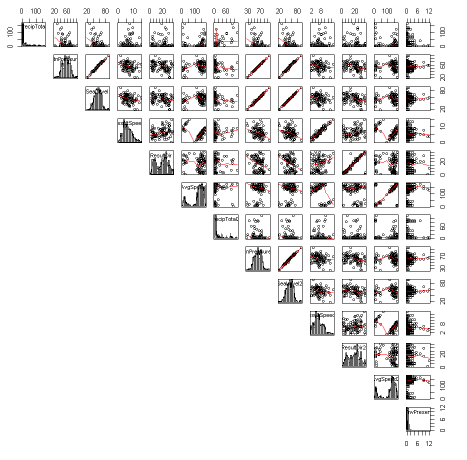


Fig 1: Pairs plot of weather variables (temperature related mainly) from station 1 & 2, variables with identical distribution can be identify (nb “ffWNV2 data exploration R”).

Data exploration

Two significant trends revealed by data exploration:

1. Two mosquito’s types (*Culex Pipiens* and *Culex Pipiens* or *Restuans*) were correlated to the presence of WNV (table 2 & fig.2).

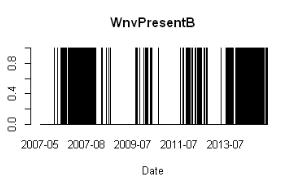
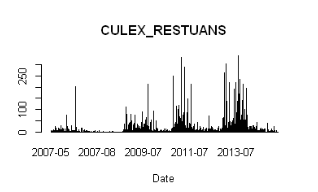
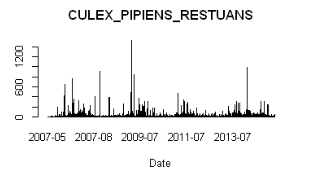


Fig. 2 Presence of mosquito types Culex Pipiens and Culex Pipiens or Restuans is visually and statistically correlated to positive tested WNV

1. Elevated temperatures were positively correlated to WNV presence (table 2 & fig.3).

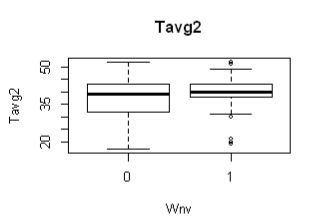
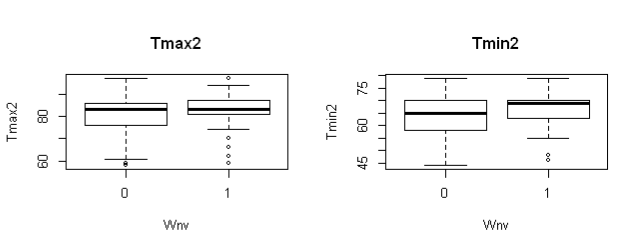
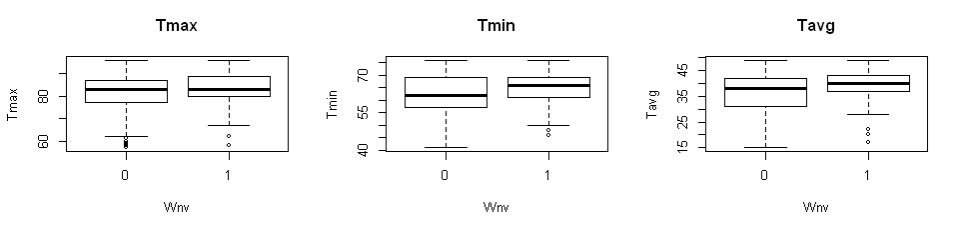
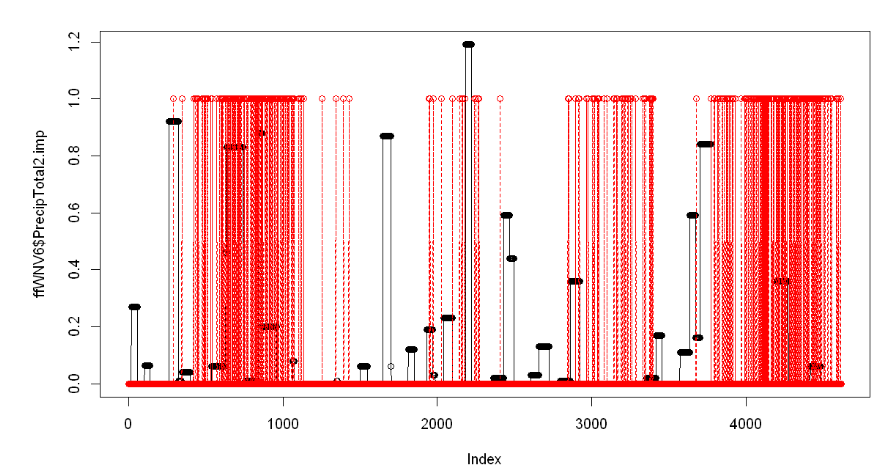


Fig. 3: Maximum, minimum and average temperatures in relation to WNV presence. The positive correlation was visible mainly in the minimum temperature (Tmin), though in all 3 temperature variables when WNV was positive the temperature was statistically significantly higher.

Table 2: “Tableone” analysis containing variables and categories in which p.v.<0.05

* Full data exploration can be found in notebook: “ffWNV2 EDA 030219”, “ffWNV5 EDA 110219 R”, “ffWNV2 data exploration R”.
* Table one and can be found in notebook: “ffWNV5 - Table1 & exploreData”.
* Visual exploratory of the data can be found in PDF files: Exploratory Data Analysis weaq, Exploratory Data Analysis wead, Exploratory Data Analysis Other, Exploratory Data Analysis Newvar, Exploratory Data Analysis MOS.

Additional interesting feature can be seen when the precipitation variable was exhibited versus the presence of WNV (fig.4). It appears that in the days post major precipitation-events, an increase in WNV presence occurred. This observation is supported by previous knowledge and is treated in the variables engineering (see below).



WNV

PrecipTotal

Fig.4: In days following precipitation events (>0.5 or >1) WNV was detected (WNV=1). The x axis is index with direct relationship to time (precipitation-black dots, WNV-red circles).

Variables engineering

1. Two variables added based on hospitalization data in Cook County. The data is originally displayed as percentage from total number of hospitalization per Quarter. In aim to adjust to the project dataset, the given data was divided to 3 and for each observation the variable stand as hospitalization per month (<http://www.idph.state.il.us/emsrpt/form-hospitalization.asp>).
2. Percentage of parasitic infection hospitalizations from total number of hospitalizations in Cook County. Taken from Cause of Hospitalization characterized as “Infection/Parasitic – Other” code 010-139 (exclude 038) by the ICD-9 codes for the principal diagnosis associated with the hospitalization that include West Nile Virus (ParasiticIn.).
3. Percentage of pneumonia/influenza hospitalizations from total number of hospitalizations in in Cook county, Taken from Cause of Hospitalization characterized as Pneumonia/Influenza code 480-4883 by the ICD-9, due to the similarity in symptoms of WNV illness to influenza (Influenza.).
4. Six additional variables were engineered based on the fact that 2 weeks (and up to 2 months) after an area received more rainfall than usual an increase may occur in the number of virus-spreading mosquitoes:
5. For each of the weather stations, two binary indicator variables were added based on the original precipitation variables (PrecipTotal & PrecipTotal2):

* pre45\_05: Events of precipitation that were higher the 0.5in were marked as 1 if occurred in the last 45 days
* pre45\_1: Events of precipitation that were higher the 1in were marked as 1 if occurred in the last 45 days

In aim to calculate those variables supplementary data was retrieved from ([https://www.ncdc.noaa.gov S1](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094846/detail) & [https://www.ncdc.noaa.gov S2](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00014819/detail)), see notebook “adding Pre”.

1. For each of the weather stations, the amount of rain events was counted 45 days backwards (RA45 and RA45\_2). Data regarding rain events in the first 45 dates of each year were missing and could not be traced, therefore this data was imputated (see below)
2. Last added variable was the “TrapClusterNum”. This variable clusters the areas of mosquito- trap stations were engineered with KNN model (see notebook “Trap-clustering by lon. lat.”). The construction of this variable was based on the assumption that mosquito prevalence will increase closer to the Lake Michigan since standing water is the preferred mosquitoes’ habitat.

Outliers determination and treatment

Weather data was tested for outliers, most of the data appear to be uniform and no outliers that need to be treated were identified (See nb: ffWNV4 R Data Cleansing).

Missing values

PrecipTotal2 was the only variable that originally had missing observations (and was not filtered out due to redundancy). Two subsequent observations were missing in this variable (out of 1470). The test for Missingness Generation Mechanism resulted with inconclusive outcome for these two missing observations. It can be assumed that the missingness was a result of measurement failure (MCAR), thus allowed imputation to PrecipTotal2 (see notebook: “PrecipTotal2 imp.”).

The engineered variables (RA45 & RA45\_2) that calculated based on events of rain were missing 352 observations resulting from their calculation method (first 45 dates of each year).

The test for Missingness Generation Mechanism resulted with inconclusive outcome for both variables. The missingness is clearly a result of the variable generation method (MAR) and thus imputation is allowed as well (See notebooks: “w1” & “w01 R Data Cleansing”)

Methodology

Feature Selection

The feature selection in this project was based on voting of five tests: univarable (when p.v. <0.05), Lasso (regression analysis), RandomForest, GradientBoost and SVM (see attached notebook “f9 Feature Selection”).

The complete dataset contained 69 features

Based on the above analysis two set of categories were made (see table 3):

1. Dataset A: Features that gained at least two votes (sum≥2), this category consists of 16 variables (all mosquito species will be used in dataset A since those are subcategories of the same variable; even though only 3 mosquito species had sum≥2)

2. Dataset B: Features that gained at least one votes (sum≥1), this category consists of 47 variables (all variables from data set A are included in data set B).

Dataset B was chosen to be the one for constructing the model.

Data Preparation

The dataset was divided into 3 sub-datasets: train constructed by 80% of the observation (n= 3693), development (dev) and test, each with 10% of the observation (n= 462 and 461 respectively); see attached notebook “f9 Data Preparation Py-TRAIN TEST”.

In the next step the “Month” and the “TrapClusterNum”, two categorical variables, were converted by “one hot encoding” in aim to better fit prediction models.

Considering the fact that the dataset was inbalance regarding the outcome (the y variable, “WnvPresentB”) and positive WNV was found only in 8% of the cases (in 385 out 4616 of the observations). Balancing the data was tested by several methods (Randomly Over Sampling Examples, Under sampling, Over sampling) the Over sampling method resulted with the best performance (AUC =0.915) and this dataset will be used for training the model (see notebook: “Data Enrichment WNV-260219”).

Table 3: feature selection voting score table: Variable name, Variable description / calcification,

Score of voting of five tests: univarable, Lasso, RandomForest, GradientBoost and SVM. Dataset: associate with dataset A or B.



Model Selection

Seven supervised classification models were initially used in this project: XGBoost, AdaBoost, Random Forest, Logistic regression, CART tree, Naive Bayes and SVM. To assess and compare the precision of the models three metrics were calculated: area under curve (AUC), accuracy, and sum of squared errors (ERR) (Table 4). The XGBoost model gained the best score with AUC of 0.89, accuracy of 0.82 and ERR of 82. CART tree model and SVM both had the height accuracy of 0.89 and the lowest ERR (50 and 53 respectively) but the AUC was much lower compared to the XGBoost model (0.69 and 0.52 respectively).

The AUC metrica was chosen to be the evaluator for model selection. AUC takes into consideration both sensitivity and specificity while accuracy (which is presented as well) does not indicate the false positive and the false negative cases.

Table 4: Seven models were assessed by: area under curve (AUC) Accuracy, and sum of squared errors (ERR). This assessment was done on the DEV dataset. XGBoost was the model with the highest AUC (see notebook:” WNV models- Python -train=over test=dev“).

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **AUC** | **Accuracy** | **ERR** |
| XGBoost | 0.89 | 0.82 | 82 |
| AdaBoost | 0.80 | 0.81 | 86 |
| Random Forest | 0.78 | 0.75 | 114 |
| Logistic regression | 0.74 | 0.75 | 116 |
| CART tree | 0.69 | 0.89 | 50 |
| Naive Bayes | 0.67 | 0.52 | 220 |
| SVM | 0.52 | 0.89 | 53 |

Model fine-tuning

The XGBoost model had the best score and therefore was chosen for further tuning. In this project, two parameters where used for the fine-tuning:

1. **eta** - learning rate, makes the model more robust by shrinking the weights on each step

2. **max\_depth** The maximum depth of a tree, used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample (notebook: “WNV XGBoost”).

The tuning of the above mentioned parameters was partially done based on given code design by Dr. Karpati to set the best parameters (notebook: “Xgboost WNV - Hyperparameters with CV T.K”) and partially was coded based on “trial and error” (notebook: “WNV XGBoost -040319”).

Ten iterations of XGBoost model run were done manually. In each run the train, dev and test sub-dataset were used. Table 5, presents in each run the eta and max depth used and the related metrics: area under curve (AUC), accuracy, and sum of squared errors (ERR).

The AUC metrics ranged from 0.81 to 1, while the value of 1 was found in 3 iterations, only among the train data (max depth=20, 10 and 9). The Accuracy ranged from 0.82 to 1, the 1 value was found in only one iteration in train subset (max depth=20). The ERR metrics ranged widely from 0 (train subset, max depth=20, eta=1) to 1003 (train subset, max depth=2, eta=1). Surprisingly, in one of the iterations (the first one that was tested), the score of the test subset regarding AUC metrics was higher than the train score. The discrepancies and anomaly of the model metrics are addressed in the discussion.

Table 5: XGBoost tuned parameters: eta and max depth, related metrics: area under curve (AUC), Accuracy and sum of squared errors (ERR) in each of the 10 iterations (Run) of the model for the train, dev and test subsets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Run#** | **Dataset** | **max depth** | **eta** | **AUC** | **Accuracy** | **ERR** |
| 1 | TRAIN | 2 | 1 | 0.929345 | 0.851671 | 1003 |
| DEV | 2 | 1 | 0.893216 | 0.822511 | 82 |
| TEST | 2 | 1 | 0.932663 | 0.839479 | 74 |
| 2 | TRAIN | 3 | 1 | 0.952434 | 0.899586 | 679 |
| DEV | 3 | 1 | 0.868483 | 0.829004 | 79 |
| TEST | 3 | 1 | 0.907076 | 0.835141 | 76 |
| 3 | TRAIN | 3 | 0.3 | 0.933253 | 0.863502 | 923 |
| DEV | 3 | 0.3 | 0.900571 | 0.785714 | 99 |
| TEST | 3 | 0.3 | 0.924514 | 0.809111 | 88 |
| 4 | TRAIN | 4 | 1 | 0.974287 | 0.934635 | 442 |
| DEV | 4 | 1 | 0.857063 | 0.841991 | 73 |
| TEST | 4 | 1 | 0.90939 | 0.845987 | 71 |
| 5 | TRAIN | 6 | 1 | 0.998009 | 0.98181 | 123 |
| DEV | 6 | 1 | 0.852377 | 0.896104 | 48 |
| TEST | 6 | 1 | 0.881288 | 0.878525 | 56 |
| 6 | TRAIN | 6 | 0.2 | 0.972067 | 0.920586 | 537 |
| DEV | 6 | 0.2 | 0.876707 | 0.805195 | 90 |
| TEST | 6 | 0.2 | 0.910832 | 0.824295 | 81 |
| 7 | TRAIN | 10 | 1 | 1 | 0.999408 | 4 |
| DEV | 10 | 1 | 0.849305 | 0.906926 | 43 |
| TEST | 10 | 1 | 0.903689 | 0.885033 | 53 |
| 8 | TRAIN | 8 | 1 | 1 | 0.998225 | 12 |
| DEV | 8 | 1 | 0.817403 | 0.902597 | 45 |
| TEST | 8 | 1 | 0.890744 | 0.885033 | 53 |
| 9 | TRAIN | 9 | 1 | 1 | 0.998817 | 8 |
| DEV | 9 | 1 | 0.854518 | 0.896104 | 48 |
| TEST | 9 | 1 | 0.876392 | 0.893709 | 49 |
| 10 | TRAIN | 20 | 1 | 1 | 1 | 0 |
| DEV | 20 | 1 | 0.875248 | 0.917749 | 38 |
| TEST | 20 | 1 | 0.913481 | 0.904555 | 44 |

\*\*\* The discrepancies and anomaly of the model metrics are addressed in the discussion.

Among all XGBoost model iterations found in table 5, the best two were:

* run # 6 max depth=6, eta=0.2

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AUC** | **Accuracy** | **ERR** |
| TRAIN | 0.972067 | 0.920586 | 537 |
| DEV | 0.876707 | 0.805195 | 90 |
| TEST | 0.910832 | 0.824295 | 81 |
| ERR ratio= 6 (train ERR/ DEV ERR) | | | |

* run # 5 max depth=6, eta=1

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AUC** | **Accuracy** | **ERR** |
| TRAIN | 0.998009 | 0.98181 | 123 |
| DEV | 0.852377 | 0.896104 | 48 |
| TEST | 0.881288 | 0.878525 | 56 |
| ERR ratio= 2.6 (train ERR/ DEV ERR) | | | |

Run# 6 had the highest AUC of the DEV and the test subsets. Run# 5 had the smallest ERR ratio between the train and the DEV subset. The DEV ERR for itself was not the lowest but among the ERR below 50 it had the highest AUC.

Runs # 7-10 were excluded due to overfitting of the train subset.

Runs # 1-3 were excluded due to ERR that was 8.6 – 12.2 fold higher in the train subset compare to the DEV set.

Discussion

This project was aimed to predict when and where different species of mosquitos will test positive for WNV in traps found around Chicago, USA in the given years: 2007, 2009, 2011 and 2013.

The project was based on data taken from the Kaggle competition “West Nile Virus Prediction” 7.

In the original competition the prediction was done on both the mosquitos’ species and presence of WNV caught on each day in each trap. In this project, on the other hand, the prediction was restricted only to presence of the WNV, given the number of mosquitoes captured from each species on each day in each trap.

During the model selection phase and the model tuning phase two issues associated with anomalies of the model metrica and overfitting were recognized:

1. Anomaly was noticed in one case of model run where the AUC metrics score of the test subset was higher than the train score.
2. In three out of ten model iterations the AUC was equal to 1 and in one iteration both the AUC and Accuracy were equal to 1 and the ERR was 0.

Those issues can be explaining by the following fact:

1. Small and unbalance data set: the data originally had only 4616 observation in which the positive outcome was found in only 8% of the observations and thus an over sampling approach that was used resulted with 6762 observation for the train data set.
2. Small dev and test data sets (n= 462 and 461 respectively).
3. Processing of the data and shifting it to an outcome of prediction of WNV solely
4. Overfitting of the model

Interestingly, the highest score in this Kaggle competition was 0.88 and although the scoring method in Kaggle was not necessarily the same metrics uses in this project, there were both in the same range.

Considering all facts mentioned above, we can conclude that XGBoost model was the best model for this project, yet for further use of this model prediction an additional validation proses is needed.

Bibliography

1. Nash, D, Mostashari, F, Fine, A, et al. The outbreak of West Nile virus infection in the New York City area in 1999. N Engl J Med 2001; 344:1807–1814.

2. Watson JT, Jones RC, Gibbs K, Paul W. Dead crow reports and location of human West Nile virus cases, Chicago, 2002. *Emerg Infect Dis*. 2004;10(5):938-40).

3. Centers for Disease Control and Prevention (CDC) West Nile virus (<https://www.cdc.gov/westnile/>).

4. World Health Organization (WHO) <https://www.who.int/news-room/fact-sheets/detail/west-nile-virus>

5. The American Mosquito Control Association (AMCA)  [(https://www.mosquito.org/page/lifecycle)](%20(https://www.mosquito.org/page/lifecycle))

6. Centers for Disease Control and Prevention (CDC) [Zika Virus](https://www.cdc.gov/zika/index.html) (<https://www.cdc.gov/zika/vector/mosquitoes-and-hurricanes.html)>

7. Kaggle competition “West Nile Virus Prediction”[(https://www.kaggle.com/c/predict-west-nile-virus/data)](https://www.kaggle.com/c/predict-west-nile-virus/data).