

WEST NILE VIRUS PREDICTION

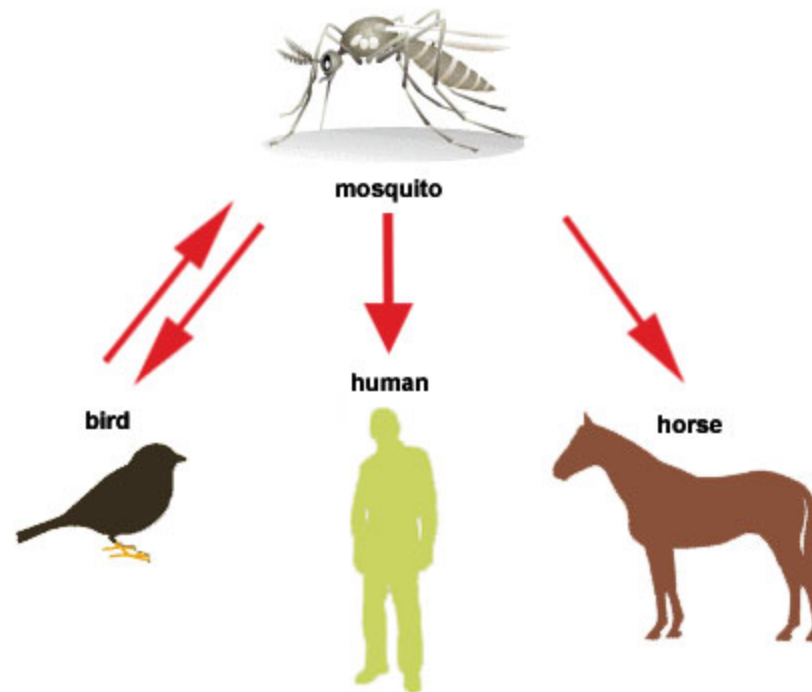
Capstone Project @ Springboard

Mentor: Alan Si

By Ruiye Ni

Mar.2016

West Nile Virus

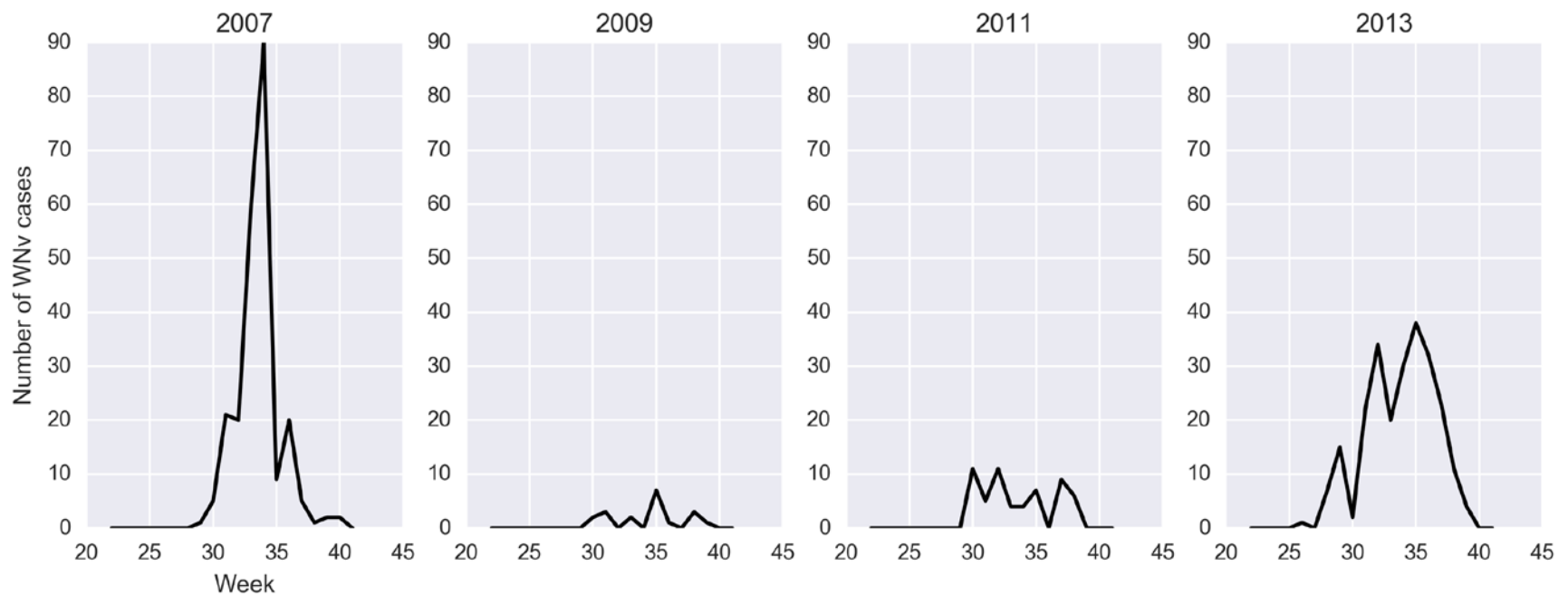


Datasets

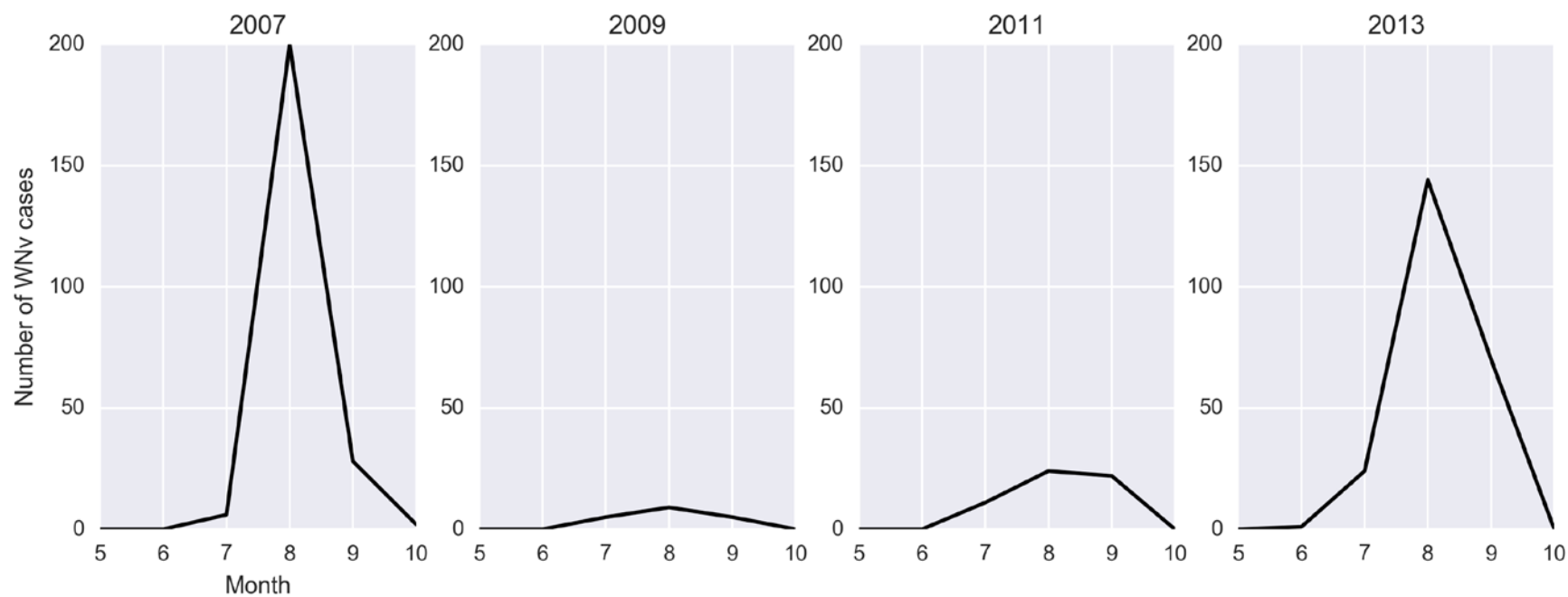
- Chicago Department of Public Health
 - Train.csv
 - Test.csv
 - Spray.csv
 - Weather.csv

Exploratory Analysis

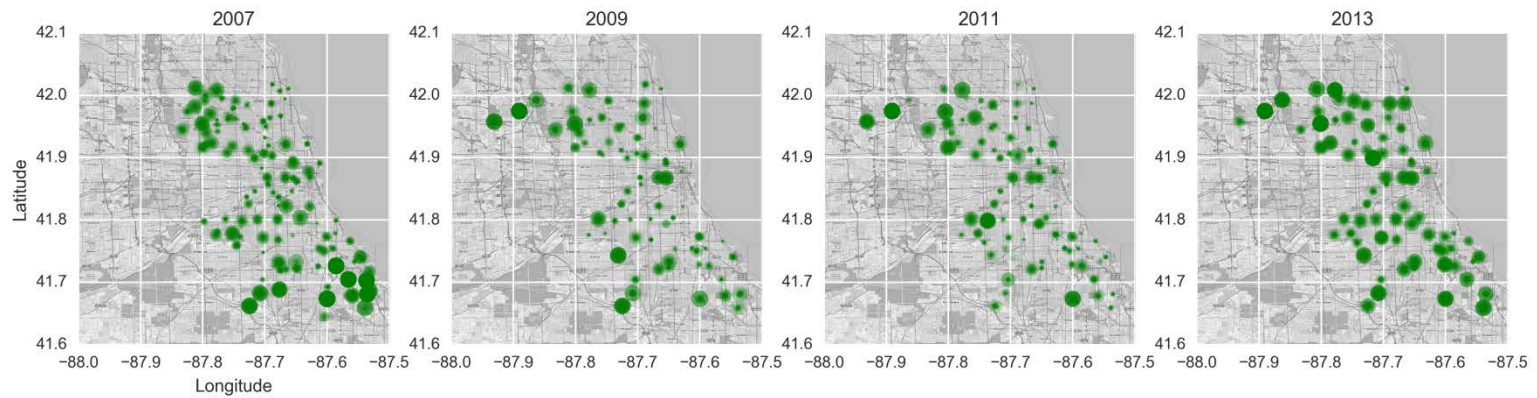
- Yearly WN virus detection by week



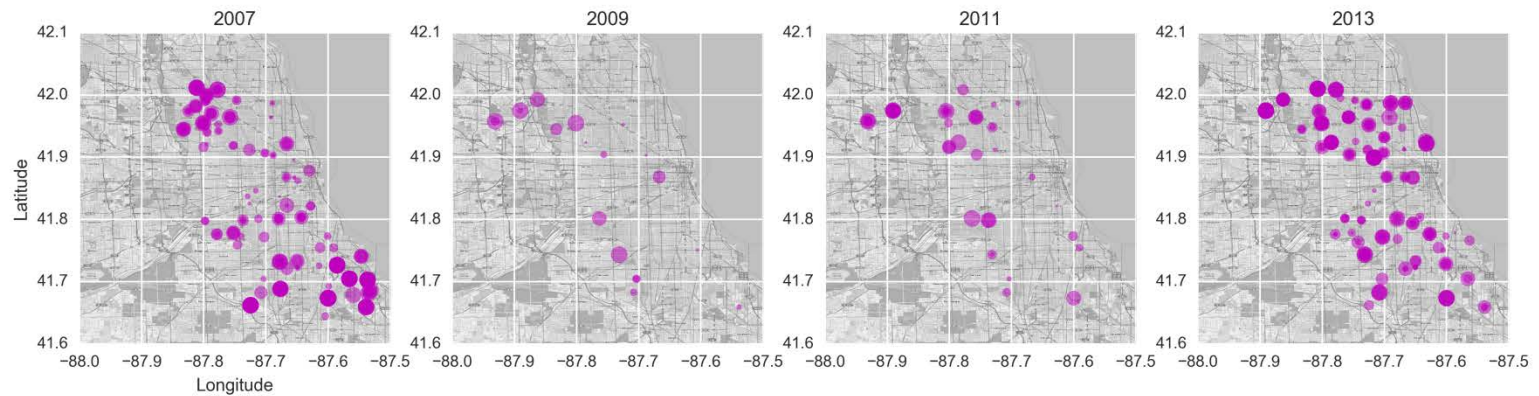
- Yearly WN virus detection by month



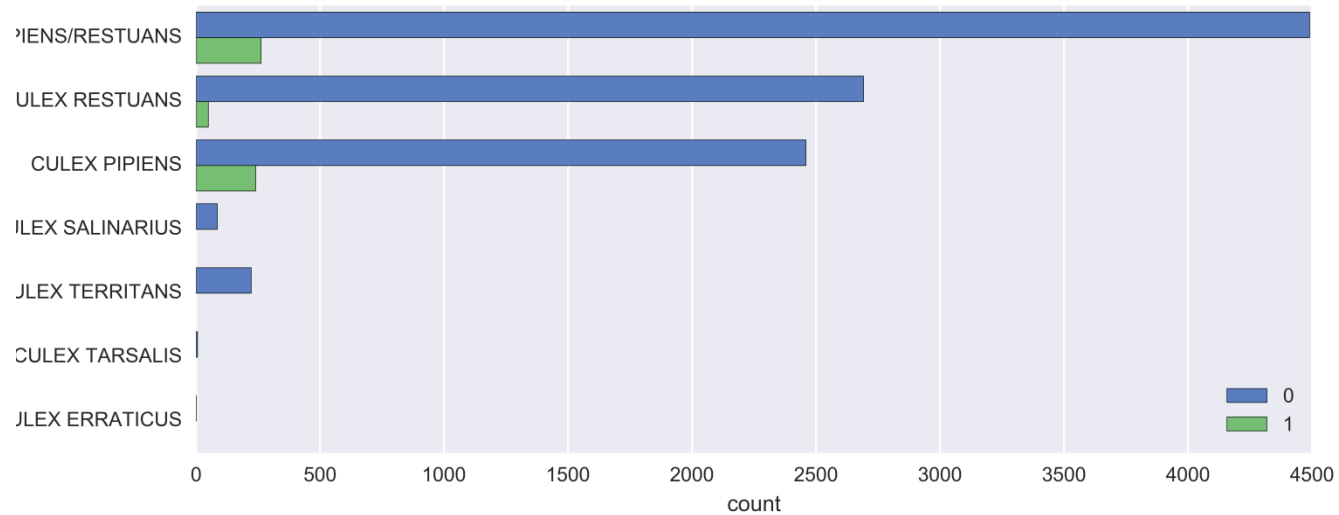
- Number of mosquitos @ locations



- Detection of WN virus @ locations



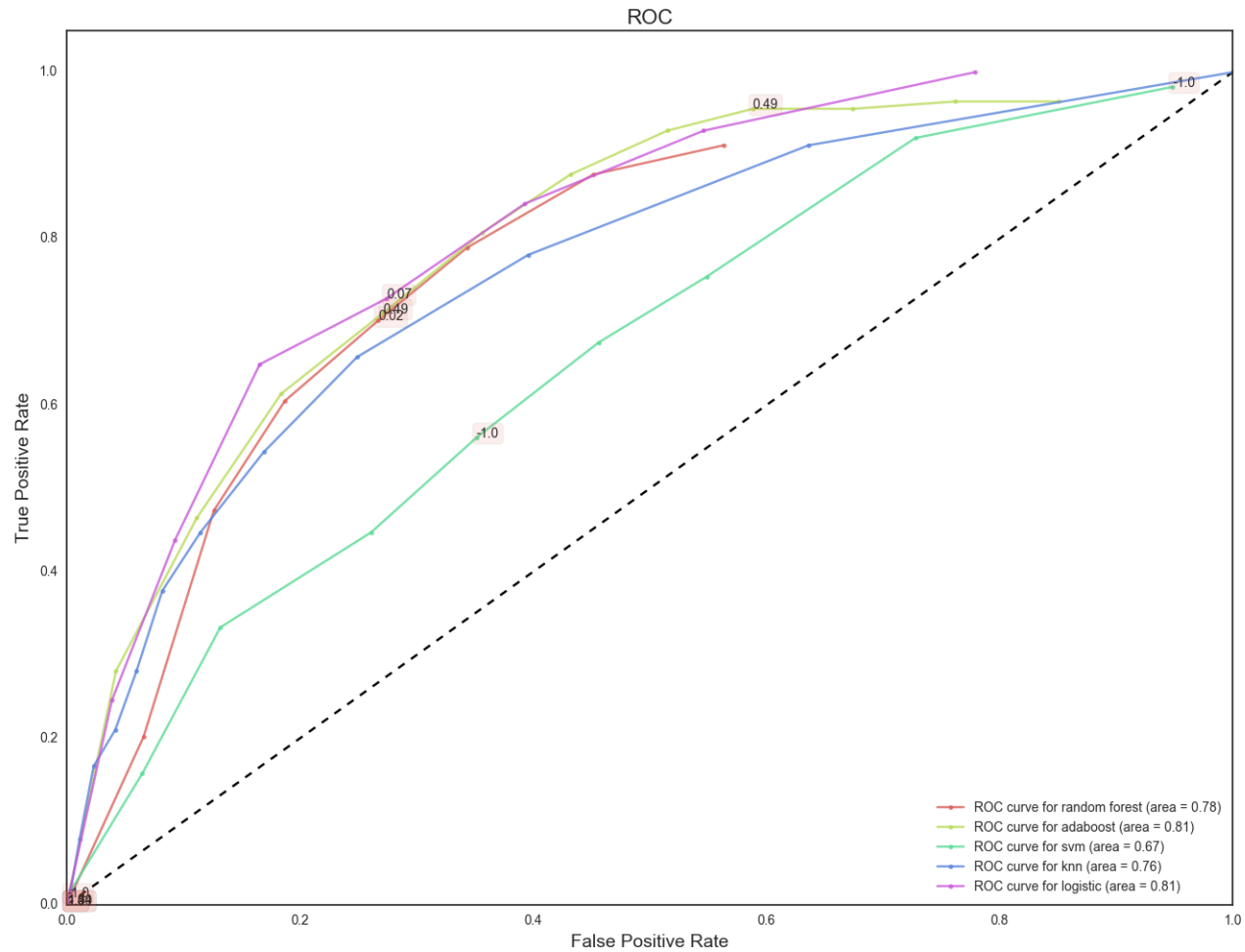
- Mosquito species vs. WNV virus detection



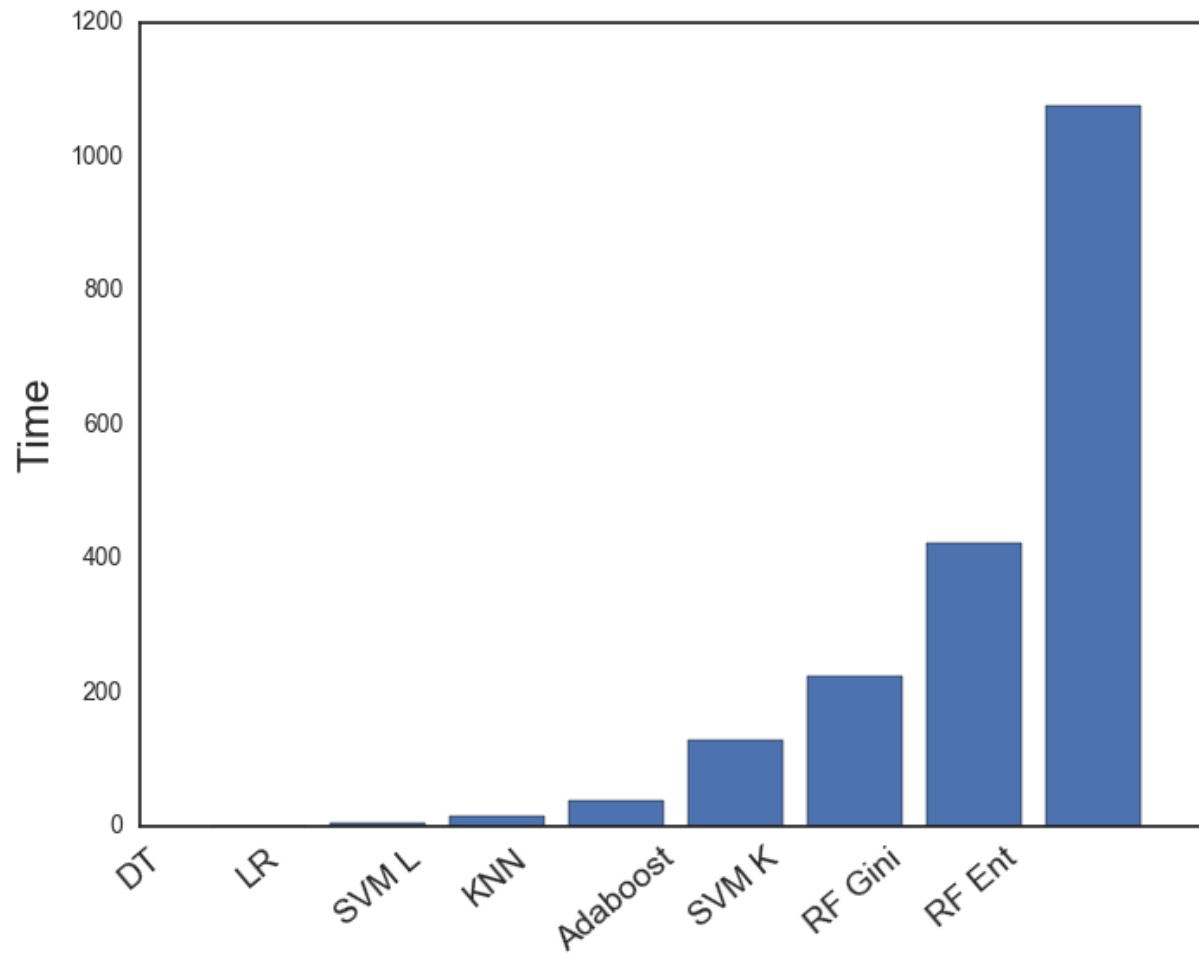
Predictive Model

- Data cleaning and wrangling
 - Replace missing value with -1
 - Convert categorical variable into dummy variables
- Data leakage
 - Drop predictor in the train dataset highly associated with target variable which is impossible to know in practice
- Imbalanced dataset
 - Use precision and recall metric to evaluate model performance instead of use accuracy
- Compare models
 - Logistic Regression

ROC Curve



Timing



Prediction