WEST NILE VIRUS PREDICTION

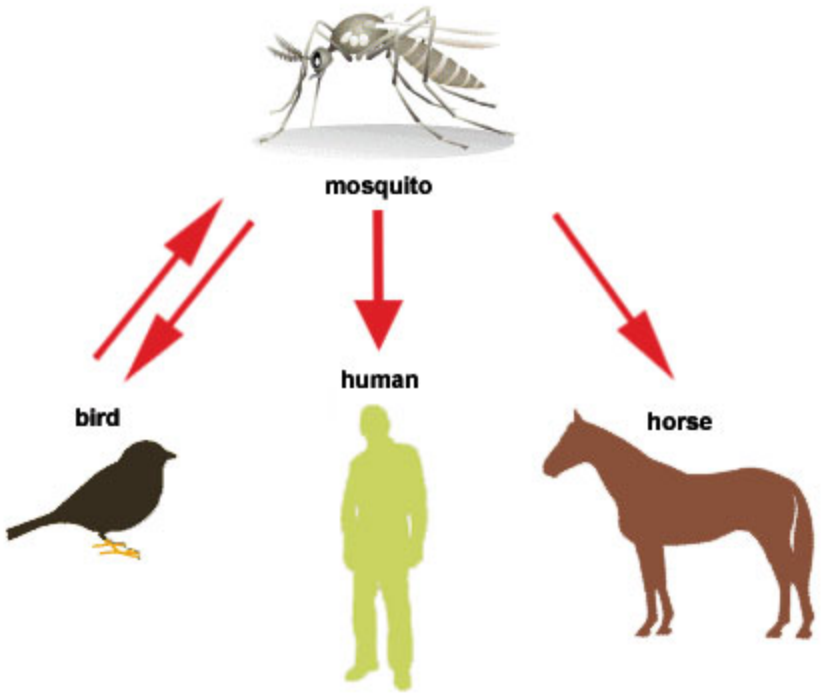
Capstone Project @ Springboard

Mentor: Alan Si

By Ruiye Ni

Mar.2016

West Nile Virus



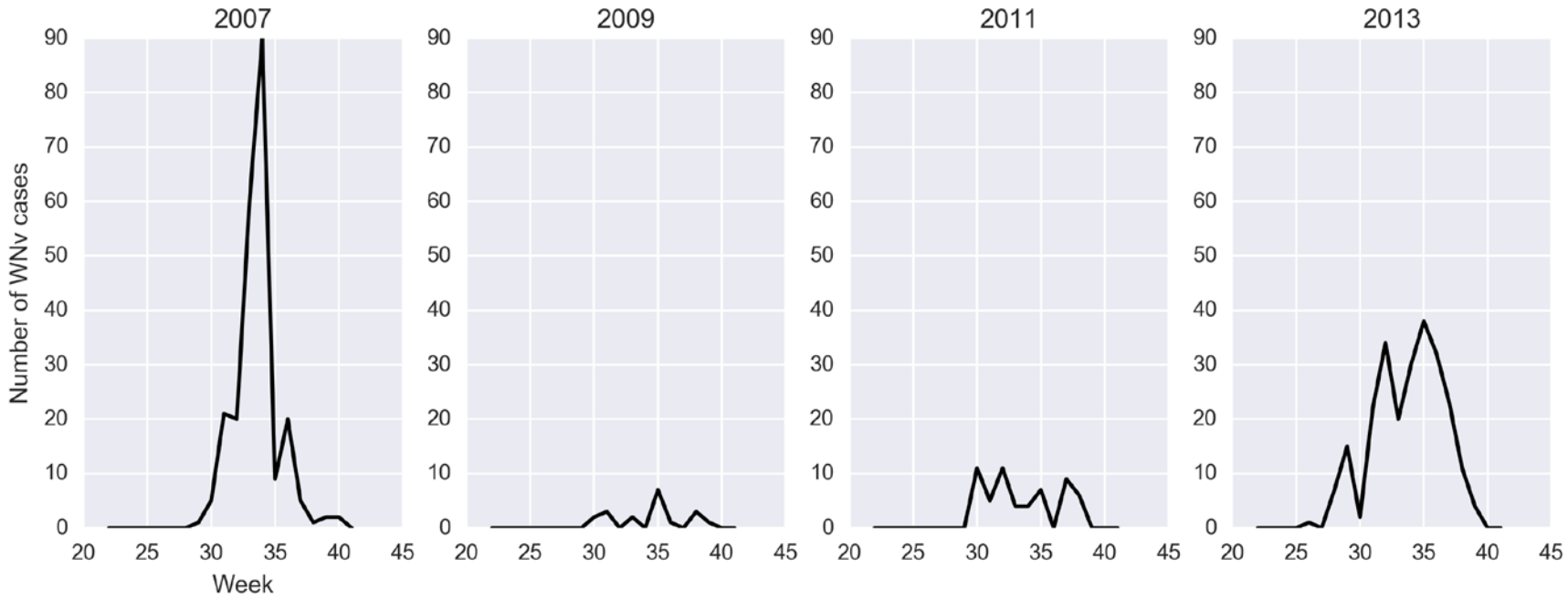
http://www.ci.berkeley.ca.us/Health_Human_Services/Environmental_Health/West_Nile_Virus.aspx

Datasets

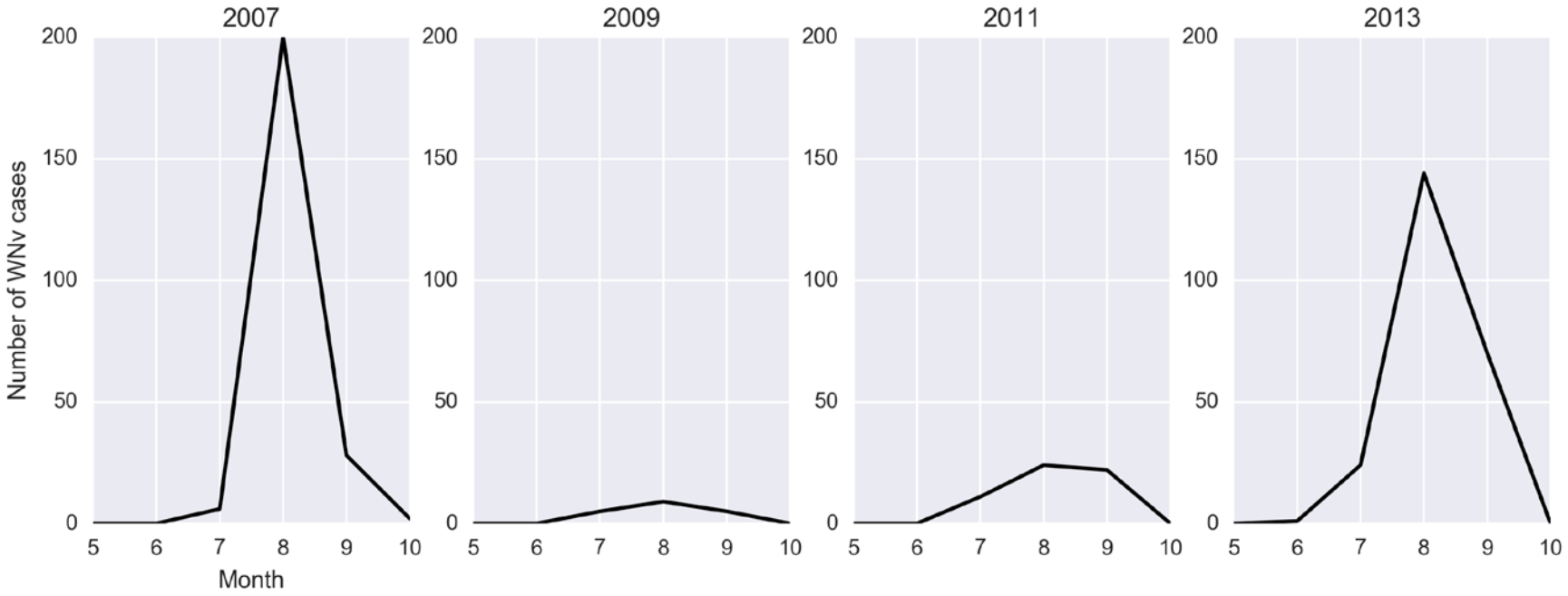
- Chicago Department of Public Health
  - Train.csv
  - Test.csv
  - Spray.csv
  - Weather.csv

Exploratory Analysis

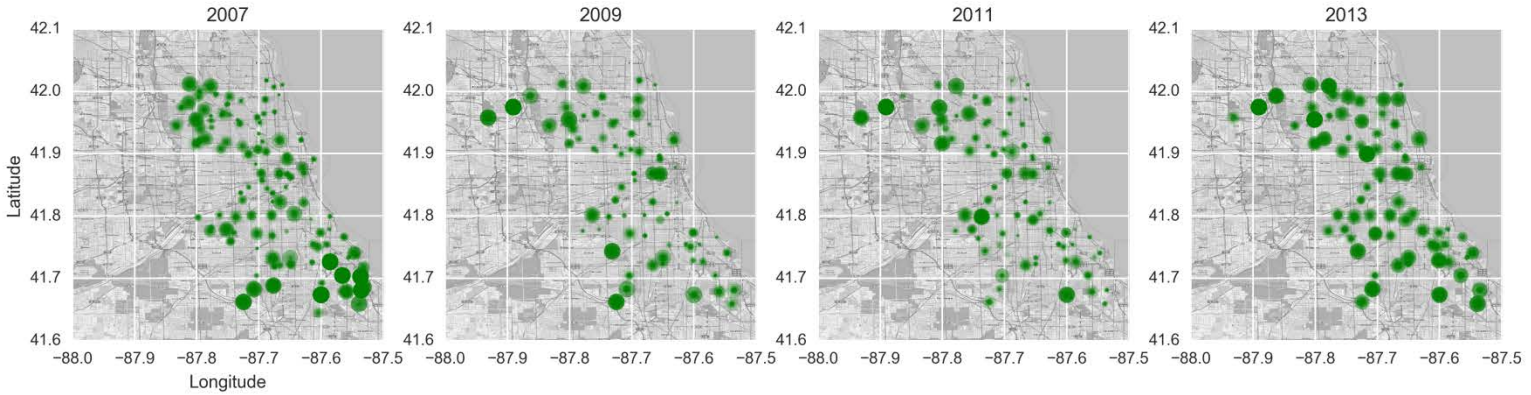
- Yearly WN virus detection by week



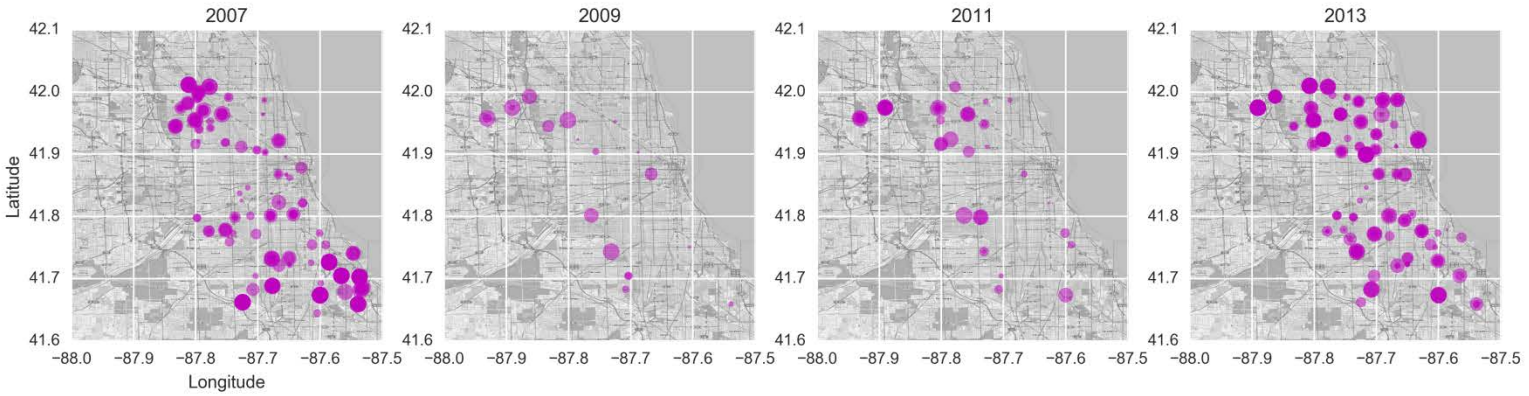
- Yearly WN virus detection by month



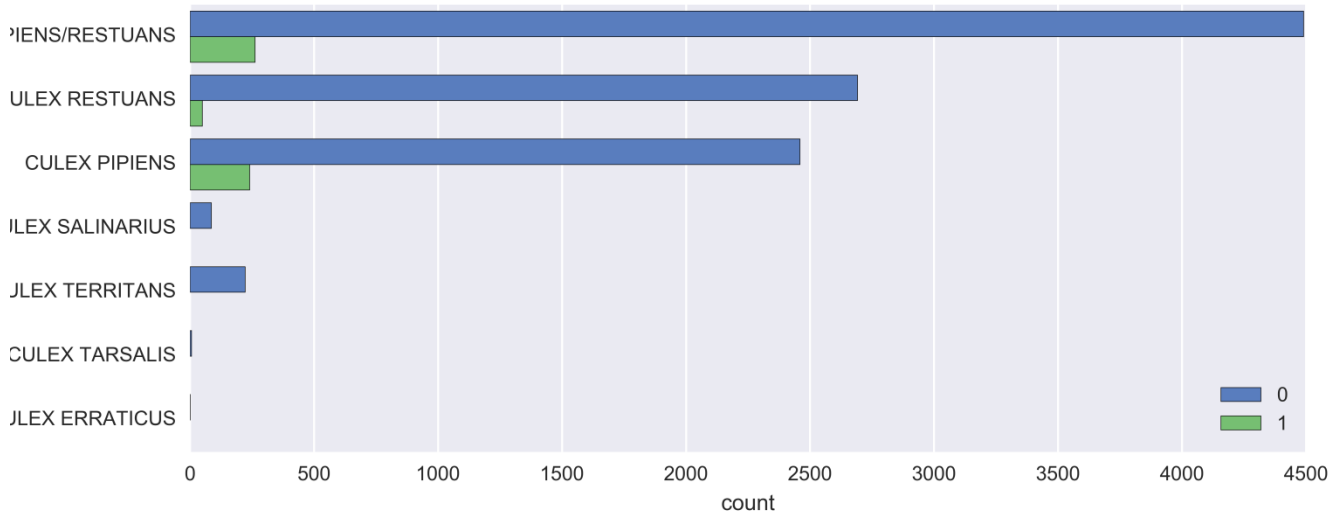
- Number of mosquitos @ locations



- Detection of WN virus @ locations



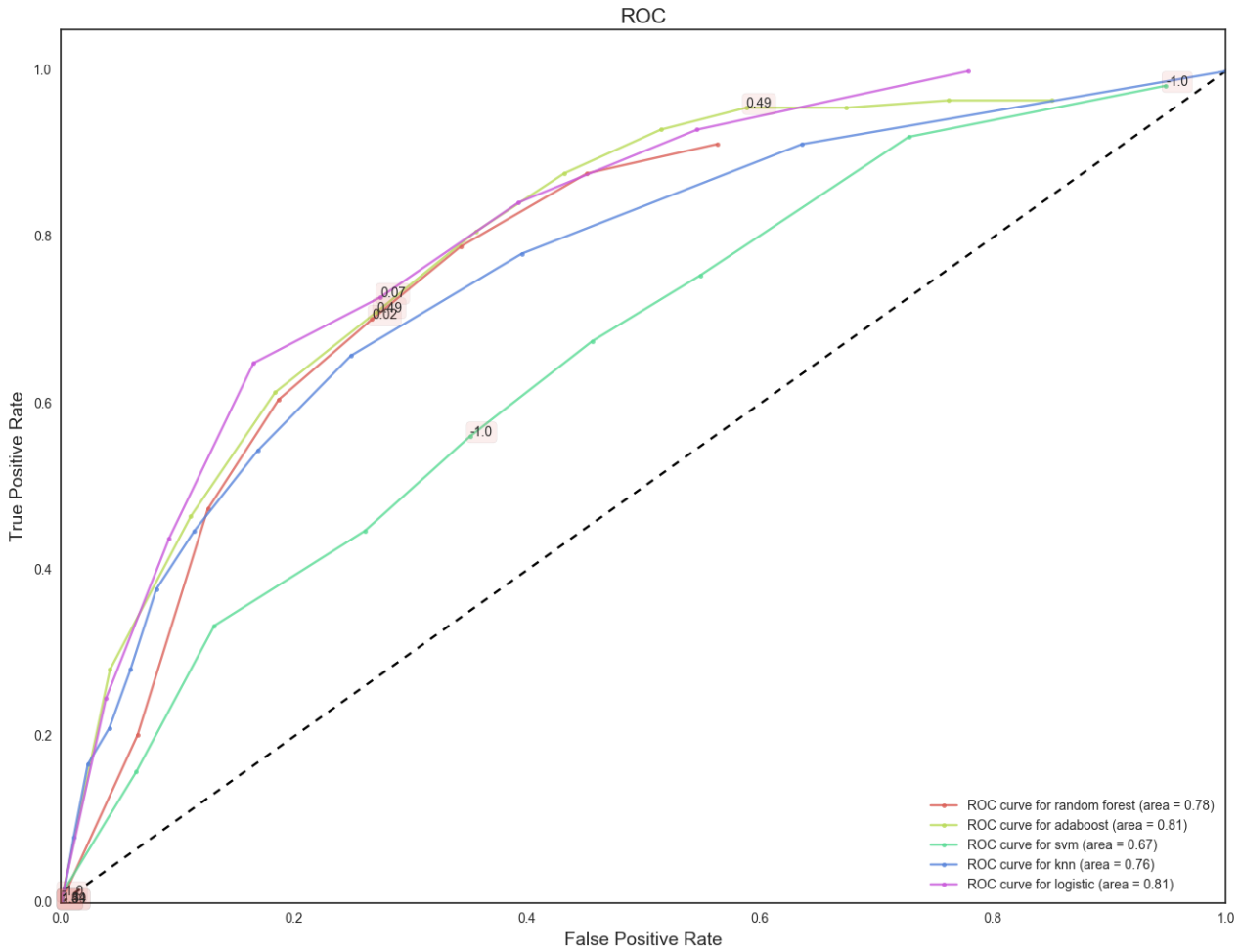
- Mosquito species vs. WN virus detection



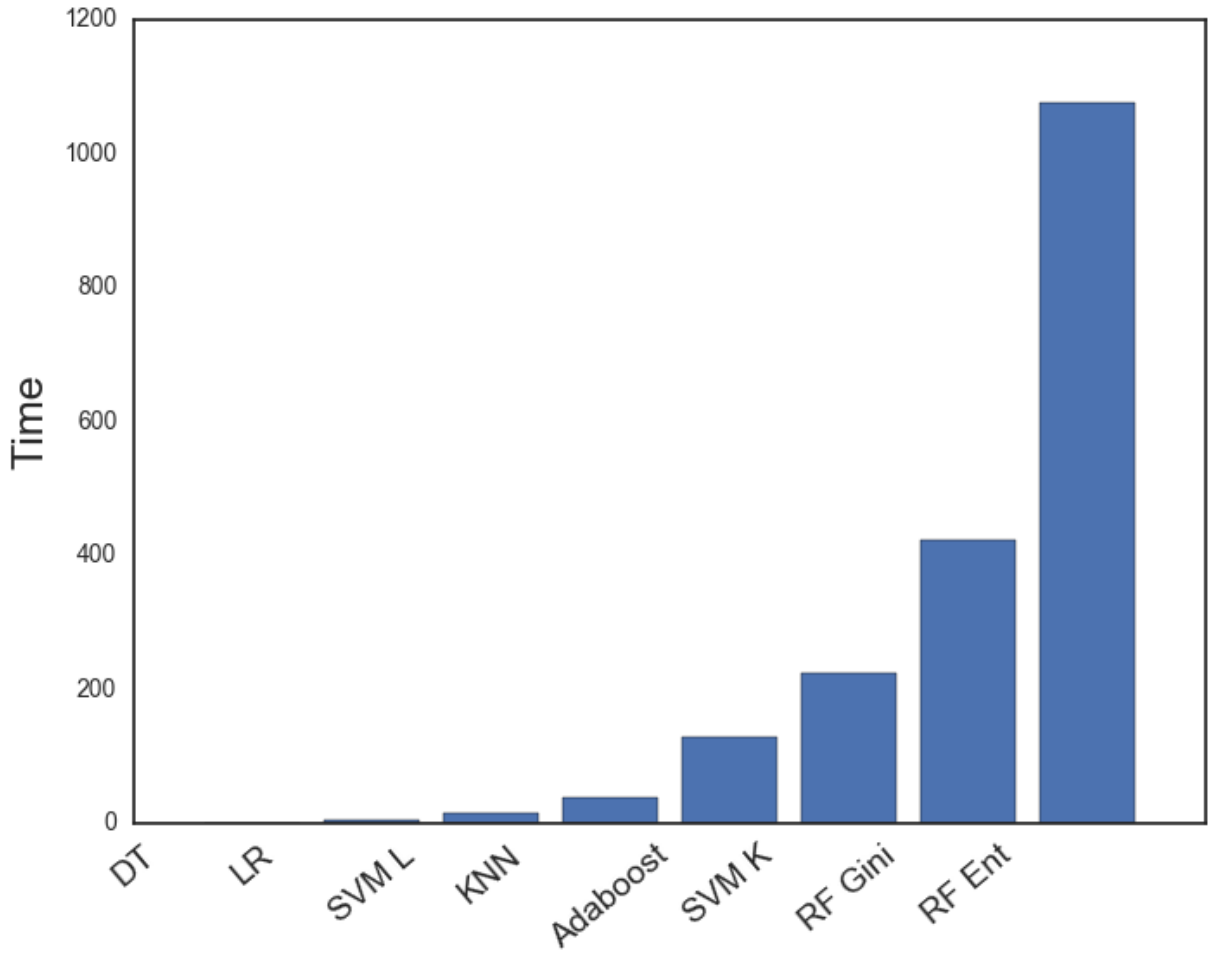
Predictive Model

- 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

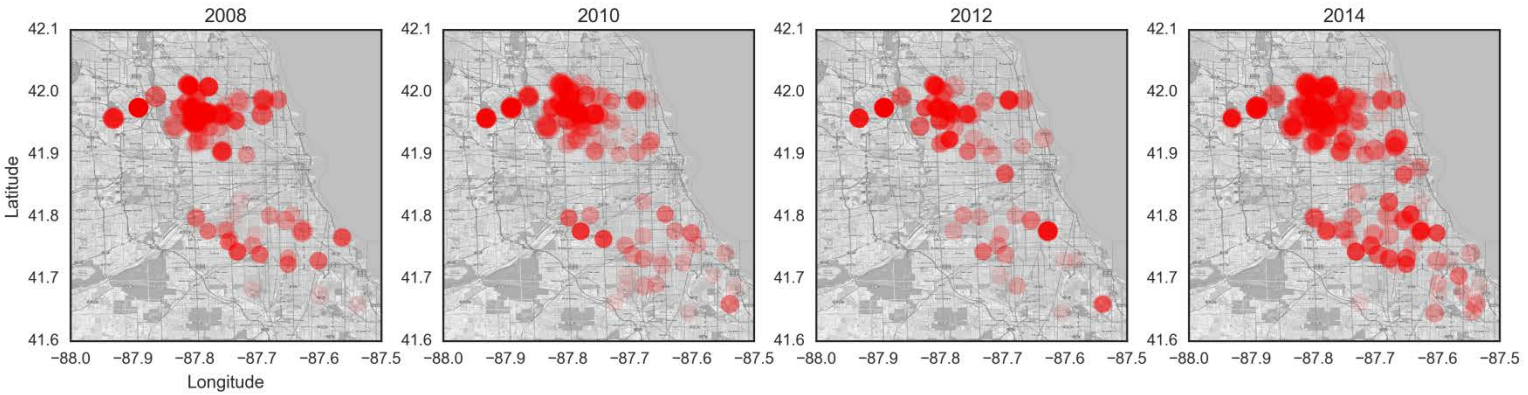
ROC Curve


ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.49
0.07
0.19
0.02
1.0
1.0
ROC curve for random forest (area = 0.78)
ROC curve for adaboost (area = 0.81)
ROC curve for svm (area = 0.67)
ROC curve for knn (area = 0.76)
ROC curve for logistic (area = 0.81)

Timing


Time
1200
1000
800
600
400
200
0
DT
LR
SVM L
KNN
Adaboost
SVM K
RF Gini
RF Ent

Prediction


2008
2010
2012
2014
Latitude
Longitude
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5