Influencers in Social Networks, a Kaggle Competition

Christopher Peters[1]

[1]Louisiana State University

EXST 7152: Advanced Topics in Statistical Modeling, 2013

Influencers in Social Networks

Goal: predict which people are influential in a social network.
Data: a pair-wise preference learning task. Each datapoint describes two individuals, A and B.

1. Dependent variable: a binary label representing a human judgement about which one of the two individuals is more influential.
2. Independent variables: 11 pre-computed, non-negative numeric features based on Twitter activity.

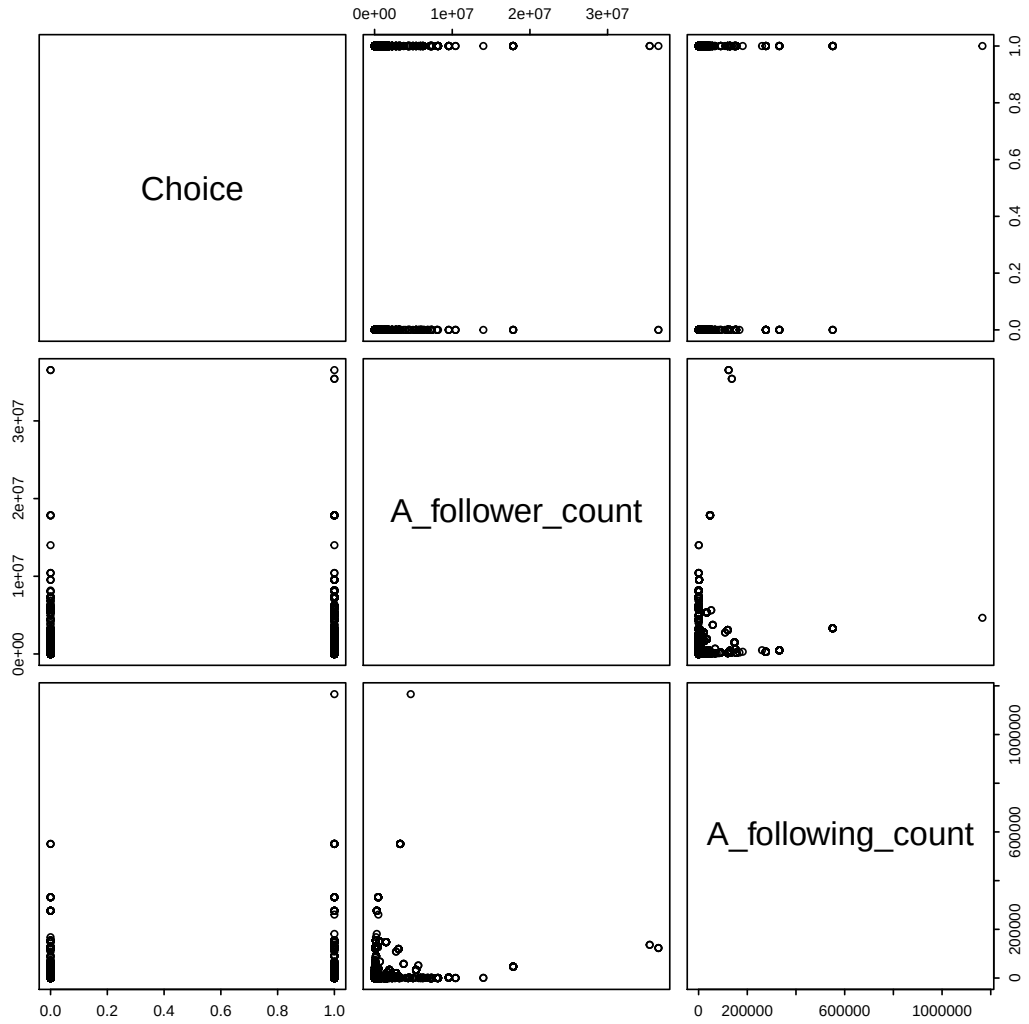
Data summary

```
setwd("C:/R_stuff/influencer/influencer")
data <- read.csv("train.csv")

summary(data[, 1:3], digits = 2)
```

```
##      Choice     A_follower_count
##  Min.   :0.00   Min.   :1.6e+01
##  1st Qu.:0.00   1st Qu.:2.7e+03
##  Median :1.00   Median :4.6e+04
##  Mean   :0.51   Mean   :6.5e+05
##  3rd Qu.:1.00   3rd Qu.:3.9e+05
##  Max.   :1.00   Max.   :3.7e+07
##  A_following_count
##  Min.   :      0
##  1st Qu.:    322
##  Median :    778
##  Mean   :  12659
##  3rd Qu.:   2838
##  Max.   :1165830
```

Data summary



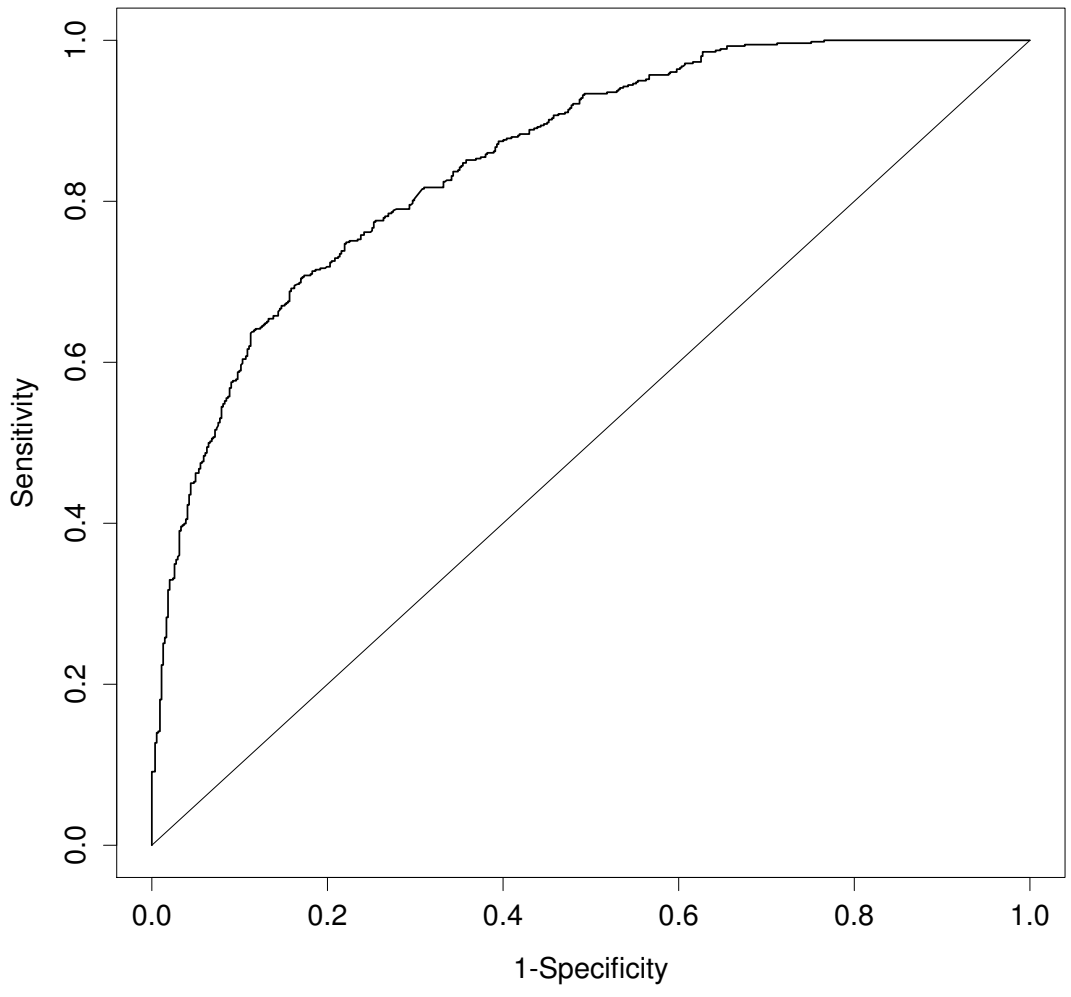
Adaboost

```
setwd("C:/R_stuff/influencer/influencer")
data <- read.csv("train.csv")
train <- sample(1:length(data$Choice), 0.8 * length(data$Choice),
    replace = FALSE)

library(gbm)

ada_1 <- gbm(Choice ~ ., distribution = "adaboost",
    data = data[train, ], n.trees = 2000, shrinkage = 0.005,
    interaction.depth = 6, train.fraction = 1, cv.folds = 4)
```

Adaboost Results: 0.889



Caret

How can we keep track of all of these moving parts better?

1. Basis
2. Loss function
3. Penalization
4. Other tuning (mtry, number of terminal nodes, etc.)

```
install.packages("caret")
library(caret)
```