HackerRank Email Opening Prediction

Vinh Dang

September 4, 2016

Data Preprocessing

The problem is defined as, given an email with several pre-calculated features, we need to predict whether or not this email will be opened by users.

The first task is obviously reading datasets

```
train = read.csv("training_dataset.csv")
test = read.csv("test_dataset.csv")
```

Let's take a look into two datasets

```
str(train)
```

```
## 'data.frame':    486048 obs. of  54 variables:
##  $ user_id                             : Factor w/ 30538 levels "//17xrIotw4mNNpre+QPI1IXTDM9B/Gb4a9
##  $ mail_id                             : Factor w/ 164 levels "/jUknxtF81t/2czkMnheze2WsqJrqNajin0ZC
##  $ mail_category                       : Factor w/ 19 levels "","mail_category_1",..: 12 2 2 2 14 16
##  $ mail_type                           : Factor w/ 5 levels "","mail_type_1",..: 2 2 2 2 2 2 2 2 3 3
##  $ sent_time                           : int  1463497837 1461357640 1463499639 1463182983 1461855019
##  $ open_time                           : int  1463540868 NA NA NA NA NA 1460231830 NA 1462890129 NA
##  $ click_time                          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ unsubscribe_time                    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ last_online                         : int  1459520208 1461210367 1463411072 1462767962 1461248422
##  $ hacker_created_at                   : int  1432533023 1432184291 1433045937 1432184291 1432998058
##  $ hacker_timezone                     : int  18000 -25200 18000 -25200 18000 18000 3600 18000 18000
##  $ clicked                             : Factor w/ 2 levels "false","true": 1 1 1 1 1 1 1 1 1 1 ...
##  $ contest_login_count                 : int  1 3 3 3 5 2 1 1 53 1 ...
##  $ contest_login_count_1_days          : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ contest_login_count_30_days         : int  0 1 0 0 0 0 0 0 5 0 ...
##  $ contest_login_count_365_days        : int  1 3 3 3 5 2 1 1 53 1 ...
##  $ contest_login_count_7_days          : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ contest_participation_count         : int  1 3 3 3 13 3 3 2 91 2 ...
##  $ contest_participation_count_1_days  : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ contest_participation_count_30_days : int  0 1 0 0 0 0 0 1 7 1 ...
##  $ contest_participation_count_365_days: int  1 3 3 3 13 3 3 2 91 2 ...
##  $ contest_participation_count_7_days  : int  0 0 0 0 0 0 0 1 1 1 ...
##  $ forum_comments_count                : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ forum_count                         : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ forum_expert_count                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ forum_questions_count               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hacker_confirmation                 : Factor w/ 2 levels "false","true": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ipn_count                           : int  17 12 46 15 107 27 20 9 106 8 ...
##  $ ipn_count_1_days                    : int  0 0 0 2 0 0 0 1 0 0 ...
##  $ ipn_count_30_days                   : int  3 2 1 3 4 2 0 6 4 0 ...
##  $ ipn_count_365_days                  : int  17 12 46 15 107 27 20 9 106 8 ...
##  $ ipn_count_7_days                   : int  0 0 1 3 0 0 0 6 2 0 ...
##  $ ipn_read                           : int  0 1 0 1 11 11 0 0 28 0 ...
##  $ ipn_read_1_days                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ipn_read_30_days                   : int  0 1 0 0 1 0 0 0 2 0 ...
##  $ ipn_read_365_days                  : int  0 1 0 1 11 11 0 0 28 0 ...
##  $ ipn_read_7_days                    : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ opened                             : Factor w/ 2 levels "false","true": 2 1 1 1 1 1 2 1 2 1 ...
##  $ submissions_count                  : int  13 99 16 101 60 101 20 14 394 3 ...
##  $ submissions_count_1_days           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_30_days          : int  0 46 3 9 1 0 12 14 19 0 ...
##  $ submissions_count_365_days         : int  13 99 16 101 60 101 20 14 394 3 ...
##  $ submissions_count_7_days           : int  0 4 0 0 0 0 12 14 0 0 ...
##  $ submissions_count_contest          : int  0 16 0 16 17 13 0 0 265 0 ...
##  $ submissions_count_contest_1_days   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_contest_30_days  : int  0 3 0 0 1 0 0 0 19 0 ...
##  $ submissions_count_contest_365_days : int  0 16 0 16 17 13 0 0 265 0 ...
##  $ submissions_count_contest_7_days   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_master           : int  13 83 16 85 43 88 20 14 129 3 ...
##  $ submissions_count_master_1_days    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_master_30_days   : int  0 43 3 9 0 0 12 14 0 0 ...
##  $ submissions_count_master_365_days  : int  13 83 16 85 43 88 20 14 129 3 ...
##  $ submissions_count_master_7_days    : int  0 4 0 0 0 0 12 14 0 0 ...
##  $ unsubscribed                       : Factor w/ 2 levels "false","true": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
str(test)
```

```
## 'data.frame':    207424 obs. of  48 variables:
##  $ user_id                            : Factor w/ 26877 levels "//17xrIotw4mNNpre+QPI1IXTDM9B/Gb4a9
##  $ mail_id                            : Factor w/ 57 levels "/jUknxtF81t/2czkMnheze2WsqJrqNajin0ZCG
##  $ mail_category                      : Factor w/ 15 levels "","mail_category_1",..: 2 2 9 2 10 2 2
##  $ mail_type                          : Factor w/ 2 levels "","mail_type_1": 2 2 2 2 2 2 2 2 2 2 ..
##  $ sent_time                          : int  1467708425 1466570440 1463671887 1467719224 1467723250
##  $ last_online                        : int  1467620141 1466482562 1463411072 1467632347 1467115996
##  $ hacker_created_at                  : int  1433145409 1433734262 1433045937 1432109057 1432012189
##  $ hacker_timezone                    : int  18000 18000 18000 18000 18000 18000 25200 18000 18000
##  $ contest_login_count                : int  1 3 3 2 1 2 1 3 2 11 ...
##  $ contest_login_count_1_days         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contest_login_count_30_days        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contest_login_count_365_days       : int  0 1 3 1 0 1 0 2 1 8 ...
##  $ contest_login_count_7_days         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contest_participation_count        : int  1 4 3 2 1 2 2 5 2 14 ...
##  $ contest_participation_count_1_days : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contest_participation_count_30_days : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contest_participation_count_365_days: int  0 2 3 1 0 1 1 4 1 10 ...
##  $ contest_participation_count_7_days : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ forum_comments_count               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ forum_count                        : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ forum_expert_count                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ forum_questions_count              : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ hacker_confirmation                : Factor w/ 2 levels "false","true": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ipn_count                          : int  13 22 46 16 13 16 22 43 16 50 ...
##  $ ipn_count_1_days                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ipn_count_30_days                  : int  0 0 1 0 0 0 9 3 1 0 ...
##  $ ipn_count_365_days                 : int  13 21 46 16 13 16 22 43 16 50 ...
##  $ ipn_count_7_days                  : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ ipn_read                          : int  0 0 0 0 0 0 0 1 0 2 ...
##  $ ipn_read_1_days                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ipn_read_30_days                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ipn_read_365_days                 : int  0 0 0 0 0 0 0 1 0 2 ...
##  $ ipn_read_7_days                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count                 : int  21 35 16 42 42 42 56 109 42 106 ...
##  $ submissions_count_1_days          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_30_days         : int  0 0 3 0 1 0 32 0 9 0 ...
##  $ submissions_count_365_days        : int  21 31 16 42 29 42 53 83 42 104 ...
##  $ submissions_count_7_days          : int  0 0 0 0 0 0 19 0 0 0 ...
##  $ submissions_count_contest         : int  0 7 0 41 0 41 1 1 41 78 ...
##  $ submissions_count_contest_1_days  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_contest_30_days : int  0 0 0 0 0 0 0 0 9 0 ...
##  $ submissions_count_contest_365_days : int  0 3 0 41 0 41 1 1 41 77 ...
##  $ submissions_count_contest_7_days  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_master          : int  21 28 16 1 42 1 55 108 1 28 ...
##  $ submissions_count_master_1_days   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submissions_count_master_30_days  : int  0 0 3 0 1 0 32 0 0 0 ...
##  $ submissions_count_master_365_days : int  21 28 16 1 29 1 52 82 1 27 ...
##  $ submissions_count_master_7_days   : int  0 0 0 0 0 0 19 0 0 0 ...
```

Several features appeared uniquely in train dataset. While I am not intending to use any unsupervised techniques, they should be removed.

```
train$click_time = NULL
train$clicked = NULL
train$open_time = NULL
train$unsubscribe_time = NULL
train$unsubscribed = NULL
```

What is the distribution of the train set

```
summary (train$opened)
```

```
##  false   true
## 324701 161347
```

Seems that the opened emails cover 66% of the train dataset (majority threshold). Any predictive model should do better than that.

Model Selection

The problem is a binary-classification. I validated three predictive models: logistic regression, random forest and deep feed-forward neural network (DNN).

Using 5-folds cross validation on the training dataset, we chose DNN as our final model, because it achieved the best F1-score and accuracy in compare to other models.