Identifying Fraudulent Activities

Yingying Xu

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(randomForest)
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
data=read.csv("Fraud_Data.csv")
ip_address=read.csv("IpAddress_to_Country.csv")

#check duplicates
nrow(data)==length(unique(data$user_id))
```

```
## [1] TRUE
```

```
head(data)
```

```
##   user_id         signup_time       purchase_time purchase_value
## 1   22058 2015-02-24 22:55:49 2015-04-18 02:47:11             34
## 2  333320 2015-06-07 20:39:50 2015-06-08 01:38:54             16
## 3    1359 2015-01-01 18:52:44 2015-01-01 18:52:45             15
## 4  150084 2015-04-28 21:13:25 2015-05-04 13:54:50             44
## 5  221365 2015-07-21 07:09:52 2015-09-09 18:40:53             39
## 6  159135 2015-05-21 06:03:03 2015-07-09 08:05:14             42
##       device_id source browser sex age ip_address class
## 1 QVPSPJUOCKZAR    SEO  Chrome   M  39  732758369     0
## 2 EOGFQPIZPYXFZ    Ads  Chrome   F  53  350311388     0
## 3 YSSKYOSJHPPLJ    SEO   Opera   M  53 2621473820     1
## 4 ATGTXKYKUDUQN    SEO  Safari   M  41 3840542444     0
## 5 NAUITBZFJKHWW    Ads  Safari   M  45  415583117     0
## 6 ALEYXFXINSXLZ    Ads  Chrome   M  18 2809315200     0
```

```
head(ip_address)
```

```
##   lower_bound_ip_address upper_bound_ip_address   country
## 1               16777216               16777471 Australia
## 2               16777472               16777727     China
## 3               16777728               16778239     China
## 4               16778240               16779263 Australia
## 5               16779264               16781311     China
## 6               16781312               16785407     Japan
```

```
summary(data)
```

```
##     user_id                    signup_time
##  Min.   :     2   2015-01-01 00:00:42:     1
##  1st Qu.:100643   2015-01-01 00:00:43:     1
##  Median :199958   2015-01-01 00:00:44:     1
##  Mean   :200171   2015-01-01 00:00:45:     1
##  3rd Qu.:300054   2015-01-01 00:00:46:     1
##  Max.   :400000   2015-01-01 00:00:47:     1
##                   (Other)            :151106
##              purchase_time    purchase_value           device_id
##  2015-06-08 09:42:04:     3   Min.   :  9.00   CQTUVBYIWWWBC:    20
##  2015-07-17 23:22:55:     3   1st Qu.: 22.00   EQYVNEGOFLAWK:    20
##  2015-09-10 09:04:53:     3   Median : 35.00   ITUMJCKWEYNDD:    20
##  2015-01-08 09:32:50:     2   Mean   : 36.94   KIPFSCNUGOLDP:    20
##  2015-01-09 14:08:40:     2   3rd Qu.: 49.00   NGQCKIADMZORL:    20
##  2015-01-12 02:56:04:     2   Max.   :154.00   ZUSVMDEZRBDTX:    20
##  (Other)            :151097                    (Other)      :150992
##     source         browser      sex            age
##  Ads   :59881   Chrome :61432   F:62819   Min.   :18.00
##  Direct:30616   FireFox:24610   M:88293   1st Qu.:27.00
##  SEO   :60615   IE     :36727             Median :33.00
##                 Opera  : 3676             Mean   :33.14
##                 Safari :24667             3rd Qu.:39.00
##                                           Max.   :76.00
##
##    ip_address            class
##  Min.   :5.209e+04   Min.   :0.00000
##  1st Qu.:1.086e+09   1st Qu.:0.00000
##  Median :2.155e+09   Median :0.00000
```

```
##  Mean   :2.152e+09   Mean   :0.09365
##  3rd Qu.:3.243e+09   3rd Qu.:0.00000
##  Max.   :4.295e+09   Max.   :1.00000
##
```

```
summary(ip_address)
```

```
##  lower_bound_ip_address upper_bound_ip_address               country
##  Min.   :1.678e+07      Min.   :1.678e+07      United States     :46868
##  1st Qu.:1.920e+09      1st Qu.:1.920e+09      Canada            : 6989
##  Median :3.231e+09      Median :3.231e+09      Russian Federation: 6739
##  Mean   :2.725e+09      Mean   :2.725e+09      Australia         : 6316
##  3rd Qu.:3.350e+09      3rd Qu.:3.350e+09      Germany           : 5999
##  Max.   :3.758e+09      Max.   :3.758e+09      United Kingdom    : 5401
##                                                (Other)           :60534
```

Merge countries to the original data set by IP

```
n=nrow(data)
data_country=rep(NA,n)
for (i in 1:n){
  tmp=as.character(ip_address[data$ip_address[i] >= ip_address$lower_bound_ip_address
                              & data$ip_address[i] <= ip_address$upper_bound_ip_address, "country"])
  if (length(tmp)==1) {data_country[i]=tmp}
}
```

```
data$country=data_country
data[,"signup_time"]=as.POSIXct(data[,"signup_time"],tz="GMT")
data[,"purchase_time"]=as.POSIXct(data[,"purchase_time"],tz="GMT")
summary(as.factor(data$country))
```

```
##                   United States                           China
##                           58049                           12038
##                           Japan                  United Kingdom
##                            7306                            4490
##               Korea Republic of                         Germany
##                            4162                            3646
##                          France                          Canada
##                            3161                            2975
##                          Brazil                           Italy
##                            2961                            1944
##                       Australia                     Netherlands
##                            1844                            1680
##              Russian Federation                           India
##                            1616                            1310
## Taiwan; Republic of China (ROC)                          Mexico
##                            1237                            1121
##                          Sweden                           Spain
##                            1090                            1027
##                    South Africa                     Switzerland
##                             838                             785
##                          Poland                       Argentina
##                             729                             661
```

```
##                      Indonesia                         Norway
##                            649                            609
##                       Colombia                         Turkey
##                            602                            568
##                       Viet Nam                        Romania
##                            550                            525
##                        Denmark                      Hong Kong
##                            490                            471
##                        Finland                        Austria
##                            460                            435
##                        Ukraine                          Chile
##                            429                            417
##                        Belgium       Iran (ISLAMIC Republic Of)
##                            409                            389
##                          Egypt                  Czech Republic
##                            359                            349
##                       Thailand                     New Zealand
##                            291                            278
##                         Israel                    Saudi Arabia
##                            272                            264
##                      Venezuela                        Ireland
##                            251                            240
##                 European Union                         Greece
##                            238                            231
##                       Portugal                        Hungary
##                            229                            211
##                       Malaysia                      Singapore
##                            210                            208
##                       Pakistan                    Philippines
##                            186                            177
##                       Bulgaria                       Morocco
##                            166                            158
##                        Algeria                           Peru
##                            122                            119
##                        Tunisia             United Arab Emirates
##                            118                            114
##                        Ecuador                      Lithuania
##                            106                             95
##                     Seychelles                          Kenya
##                             95                             93
##                     Kazakhstan                     Costa Rica
##                             92                             90
##                         Kuwait                       Slovenia
##                             90                             87
##      Slovakia (SLOVAK Republic)                        Uruguay
##                             86                             80
##  Croatia (LOCAL Name: Hrvatska)                        Belarus
##                             79                             72
##                     Luxembourg                         Serbia
##                             72                             69
##                        Nigeria                         Latvia
##                             67                             64
##                         Panama                        Bolivia
##                             62                             53
```

```
##              Dominican Republic                          Cyprus
##                              51                              43
##                         Estonia                           Oman
##                              42                              41
##                      Bangladesh              Moldova Republic of
##                              37                              37
##                        Paraguay                         Georgia
##                              35                              32
##                       Sri Lanka           Bosnia and Herzegowina
##                              31                              30
##                     Puerto Rico                          Jordan
##                              30                              28
##                         Lebanon                     El Salvador
##                              28                              25
##                           Qatar                           Sudan
##                              25                              25
##                          Angola                        Macedonia
##                              24                              24
##             Syrian Arab Republic                       Azerbaijan
##                              24                              23
##                         Namibia                          Malta
##                              23                              22
##                         (Other)                           NA's
##                             550                           21966
```

Feature engineering

Time difference between purchase and signup

```
data$purchase_signup_diff=as.numeric(difftime(as.POSIXct(data$purchase_time,tz="GMT"),as.POSIXct(data$s
```

Device ID with different users

```
data=data %>%
  group_by(device_id) %>%
  mutate(device_id_count=n())
```

Same IP with different users

```
data=data.frame(data %>%
                group_by(ip_address) %>%
                  mutate(ip_address_count=n()))
```

Day of the week/ week of the year

```
data$signup_time_wd=format(data$signup_time,"%A")
data$purchase_time_wd=format(data$purchase_time,"%A")

data$signup_time_wy=as.numeric(format(data$signup_time,"%U"))
data$purchase_time_wy=as.numeric(format(data$purchase_time,"%U"))
```

```
data_rf=data[,-c(1:3,5)]

#replace the NA in the country var
data_rf$country[is.na(data_rf$country)]="Not_found"

#keep top 50 countries, else with "other"
country_n=length(unique(data_rf$country))
data_rf$country=
  ifelse(data_rf$country %in% names(sort(table(data_rf$country),decreasing=TRUE))[51:country_n],
                        "Other", as.character(data_rf$country)
         )

#as.factor in class
data_rf$class=as.factor(data_rf$class)

#all cahracters become factors
data_rf[sapply(data_rf,is.character)]<-lapply(data_rf[sapply(data_rf,is.character)],as.factor)

head(data_rf)
```

```
##   purchase_value source browser sex age ip_address class       country
## 1             34    SEO  Chrome   M  39  732758369     0         Japan
## 2             16    Ads  Chrome   F  53  350311388     0 United States
## 3             15    SEO   Opera   M  53 2621473820     1 United States
## 4             44    SEO  Safari   M  41 3840542444     0     Not_found
## 5             39    Ads  Safari   M  45  415583117     0 United States
## 6             42    Ads  Chrome   M  18 2809315200     0        Canada
##   purchase_signup_diff device_id_count ip_address_count signup_time_wd
## 1              4506682               1                1        Tuesday
## 2                17944               1                1         Sunday
## 3                    1              12               12       Thursday
## 4               492085               1                1        Tuesday
## 5              4361461               1                1        Tuesday
## 6              4240931               1                1       Thursday
##   purchase_time_wd signup_time_wy purchase_time_wy
## 1         Saturday              8               15
## 2           Monday             23               23
## 3         Thursday              0                0
## 4           Monday             17               18
## 5        Wednesday             29               36
## 6         Thursday             20               27
```

Train/Test Split

```
train_sample=sample(nrow(data_rf),size=nrow(data)*0.66)
train_data=data_rf[train_sample,]
```

```
test_data =data_rf[-train_sample,]
```

Random Forest

```
rf = randomForest(y=train_data$class,x=train_data[,-7],
                  ytest=test_data$class,xtest=test_data[,-7],
                  ntree=50,mtry=3,keep.forest = TRUE)
rf
```

```
##
## Call:
##  randomForest(x = train_data[, -7], y = train_data$class, xtest = test_data[,      -7], ytest = test_
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 4.42%
## Confusion matrix:
##       0    1  class.error
## 0 90364   17 0.0001880926
## 1  4392 4960 0.4696321642
##                 Test set error rate: 4.2%
## Confusion matrix:
##       0    1  class.error
## 0 46574    6 0.0001288106
## 1  2151 2648 0.4482183788
```

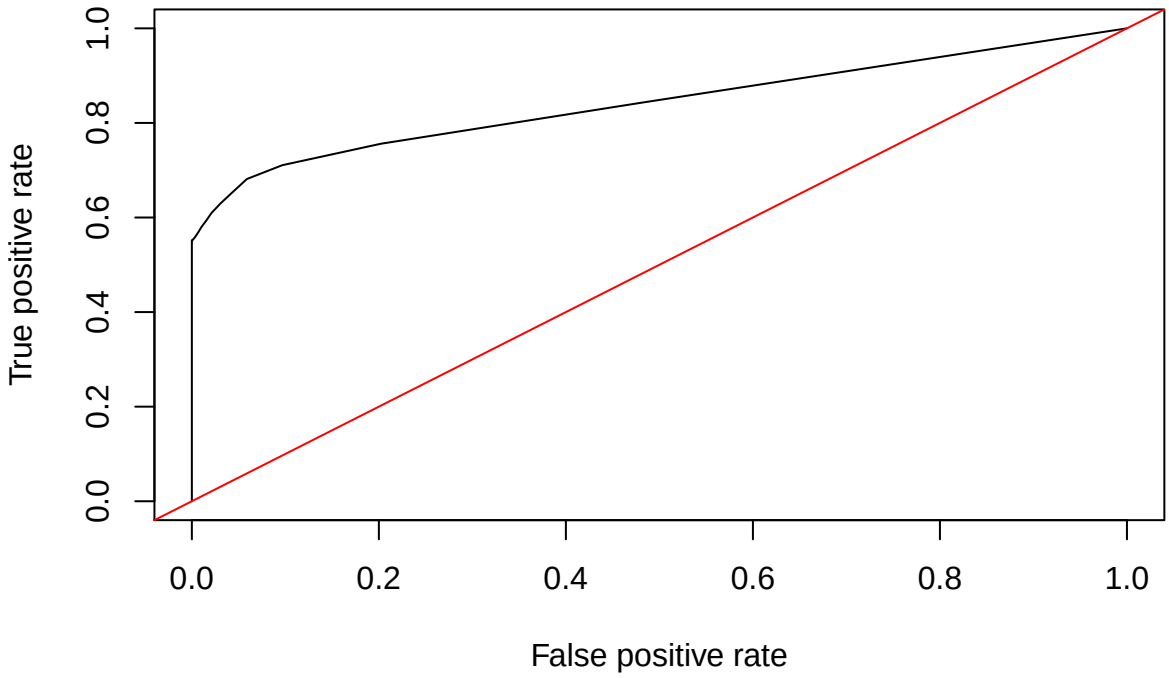
Model Prediction and Actual Values

```
rf_results=data.frame(true_values=test_data$class,
                      predictions=rf$test$votes[,2])
identical(as.numeric(as.character(rf$test$predicted)),ifelse(rf_results$predictions>0.5,1,0))
```

```
## [1] TRUE
```

```
pred=prediction(rf_results$predictions,rf_results$true_values)

# plot the ROC
perf=performance(pred,measure = 'tpr',x.measure = "fpr")
plot(perf)+abline(a=0,b=1,col='red')
```


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

```
## numeric(0)
```