Finhacks 2018

Auto-A

12 Oktober 2018

Laporan berikut menjelaskan tentang metodologi dan proses penentuan apakah transaksi yang ditandai terbukti melakukan penipuan atau tidak berdasarkan indikator yang di dapatkan dari dataset.

Kata kunci: finansial, penipuan, fraud detection, finhacks

Ringkasan Eksekutif

Laporan berikut menjelaskan tentang metodologi dan proses penentuan apakah transaksi yang ditandai terbukti melakukan penipuan atau tidak berdasarkan indikator yang di dapatkan dari dataset. Dataset disediakan oleh panitia Finhacks 2018.

Penulis mulai mengeksplorasi data dan menganalisa hubungan antar variabel hingga menghasilkan prediksi transaksi yang terbukti melakukan penipuan dengan metode Random Forest Model. Metode random forest sendiri dipilih karena bisa melakukan regresi dan klasifikasi sekaligus.

flag_transaksi_finansial, status_transaksi, bank_pemilik_kartu dihapus karena dinilai tidak memiliki pengaruh statistic yang signifikan, sehingga bisa dibilang tidak akan banyak memengaruhi keluaran yang diinginkan.

Dataset

Dalam dataset terdapat 28 variabel, dimana variabel terakhir yaitu flag_transaksi_fraud menjadi target yang akan dicari. flag_transaksi_fraud sendiri diubah menjadi nilai kategorikal karena isinya mendeskripsikan apakah suatu transaksi yang ditandai merupakan penipuan atau tidak.

Proses

Pada lomba ini, kami akan menggunakan metode Random Forest. Random forest adalah sebuah metode machine learning yang mampu melakukan regresi dan klasifikasi sekaligus. Random Forest mampu menemukan relasi yang lebih kompleks dengan waktu yang relatif efisien.

Langkah pertama adalah mendeklarasikan semua library yang diperlukan

```
library(randomForest)
```

```
## randomForest 4.6-14

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

library(ggplot2)

##
## Attaching package: 'ggplot2'

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

Setelah itu, file “fraud_train.csv” dan “fraud_test.csv” dibaca dengan perintah ‘read.csv’. Perintah ‘summary’ digunakan untuk menunjukkan ringkasan dari dataTrain dan dataTest.

```
dataTrain <- read.csv('fraud_train.csv')
summary(dataTrain)

##        X         id_tanggal_transaksi_awal tanggal_transaksi_awal
##  Min.   :    1   Min.   :2457297           Min.   :2457303
##  1st Qu.: 3784   1st Qu.:2457404           1st Qu.:2457451
##  Median : 7475   Median :2457500           Median :2457543
##  Mean   : 7508   Mean   :2457490           Mean   :2457541
##  3rd Qu.:11265   3rd Qu.:2457581           3rd Qu.:2457632
##  Max.   :15000   Max.   :2457662           Max.   :2457754
##
##    tipe_kartu      id_merchant     nama_merchant    tipe_mesin
##  Min.   :  0.00   Min.   :    -2   Min.   :   2   Min.   :     -4
##  1st Qu.: 93.00   1st Qu.:    -2   1st Qu.:1798   1st Qu.:1130699
##  Median :103.00   Median :    -2   Median :1798   Median :1836319
##  Mean   : 85.34   Mean   : 39301   Mean   :1678   Mean   :1649037
##  3rd Qu.:111.00   3rd Qu.:    -2   3rd Qu.:1798   3rd Qu.:2419350
##  Max.   :138.00   Max.   :720990   Max.   :1859   Max.   :6928943
##
##  tipe_transaksi  nama_transaksi  nilai_transaksi      id_negara
##  Min.   : 26.0   Min.   : 1.00   Min.   :       1   Min.   : -2.00
##  1st Qu.: 26.0   1st Qu.: 9.00   1st Qu.:  200000   1st Qu.: 96.00
##  Median :156.0   Median :10.00   Median :  570000   Median : 96.00
##  Mean   :178.8   Mean   :10.73   Mean   : 1315219   Mean   : 96.06
##  3rd Qu.:301.0   3rd Qu.:11.00   3rd Qu.: 1250000   3rd Qu.: 96.00
##  Max.   :640.0   Max.   :20.00   Max.   :75000000   Max.   :216.00
##
##   nama_negara      nama_kota      lokasi_mesin  pemilik_mesin
##  Min.   : 1.00   Min.   :  1.0   Min.   :   2   Min.   :   1
##  1st Qu.: 5.00   1st Qu.:102.0   1st Qu.:1914   1st Qu.: 613
##  Median : 5.00   Median :128.0   Median :3720   Median : 613
##  Mean   : 5.02   Mean   :148.3   Mean   :3948   Mean   : 766
##  3rd Qu.: 5.00   3rd Qu.:203.0   3rd Qu.:5637   3rd Qu.: 613
##  Max.   :16.00   Max.   :293.0   Max.   :8697   Max.   :2688
##
```

```
##  waktu_transaksi  kuartal_transaksi kepemilikan_kartu  nama_channel
##  Min.   :    47   Min.   :1.000     Min.   :1.000     Min.   :1.000
##  1st Qu.:102622   1st Qu.:2.000     1st Qu.:2.000     1st Qu.:1.000
##  Median :140707   Median :3.000     Median :2.000     Median :1.000
##  Mean   :138896   Mean   :2.855     Mean   :1.932     Mean   :1.404
##  3rd Qu.:175420   3rd Qu.:3.000     3rd Qu.:2.000     3rd Qu.:1.000
##  Max.   :235914   Max.   :4.000     Max.   :2.000     Max.   :5.000
##
##    id_channel    flag_transaksi_finansial status_transaksi
##  Min.   :3.000   Mode :logical            Min.   :3
##  1st Qu.:9.000   FALSE:13125              1st Qu.:3
##  Median :9.000                            Median :3
##  Mean   :8.237                            Mean   :3
##  3rd Qu.:9.000                            3rd Qu.:3
##  Max.   :9.000                            Max.   :3
##
##  bank_pemilik_kartu rata_rata_nilai_transaksi maksimum_nilai_transaksi
##  Min.   :999        Min.   :   50000          Min.   :    38000
##  1st Qu.:999        1st Qu.:  568563          1st Qu.:  2500000
##  Median :999        Median : 1024239          Median :  6000000
##  Mean   :999        Mean   : 1364132          Mean   : 12287603
##  3rd Qu.:999        3rd Qu.: 1679778          3rd Qu.: 15000000
##  Max.   :999        Max.   :24666667          Max.   :100000000
##                     NA's   :21                NA's   :21
##  minimum_nilai_transaksi rata_rata_jumlah_transaksi flag_transaksi_fraud
##  Min.   :       1        Min.   : 1.000             Min.   :0.00000
##  1st Qu.:   25000        1st Qu.: 1.680             1st Qu.:0.00000
##  Median :   36964        Median : 2.100             Median :0.00000
##  Mean   :   76519        Mean   : 2.436             Mean   :0.06933
##  3rd Qu.:   63200        3rd Qu.: 2.790             3rd Qu.:0.00000
##  Max.   :75000000        Max.   :19.780             Max.   :1.00000
##  NA's   :21              NA's   :21
```

Data-data yang tidak diperlukan dihapus karena hanya memiliki satu nilai. Data-data yang dihapus adalah “flag_transaksi_finansial”, “status_transaksi”, dan “bank_pemilik_kartu”

```
dataTrain$flag_transaksi_finansial <- NULL
dataTrain$status_transaksi <- NULL
dataTrain$bank_pemilik_kartu <- NULL
summary(dataTrain)
```

```
##        X         id_tanggal_transaksi_awal tanggal_transaksi_awal
##  Min.   :    1   Min.   :2457297           Min.   :2457303
##  1st Qu.: 3784   1st Qu.:2457404           1st Qu.:2457451
##  Median : 7475   Median :2457500           Median :2457543
##  Mean   : 7508   Mean   :2457490           Mean   :2457541
##  3rd Qu.:11265   3rd Qu.:2457581           3rd Qu.:2457632
##  Max.   :15000   Max.   :2457662           Max.   :2457754
##
##    tipe_kartu      id_merchant     nama_merchant    tipe_mesin
```

```
##  Min.   :  0.00   Min.   :    -2   Min.   :   2   Min.   :     -4
##  1st Qu.: 93.00   1st Qu.:    -2   1st Qu.:1798   1st Qu.:1130699
##  Median :103.00   Median :    -2   Median :1798   Median :1836319
##  Mean   : 85.34   Mean   : 39301   Mean   :1678   Mean   :1649037
##  3rd Qu.:111.00   3rd Qu.:    -2   3rd Qu.:1798   3rd Qu.:2419350
##  Max.   :138.00   Max.   :720990   Max.   :1859   Max.   :6928943
##
##  tipe_transaksi  nama_transaksi  nilai_transaksi      id_negara
##  Min.   : 26.0   Min.   : 1.00   Min.   :       1   Min.   : -2.00
##  1st Qu.: 26.0   1st Qu.: 9.00   1st Qu.:  200000   1st Qu.: 96.00
##  Median :156.0   Median :10.00   Median :  570000   Median : 96.00
##  Mean   :178.8   Mean   :10.73   Mean   : 1315219   Mean   : 96.06
##  3rd Qu.:301.0   3rd Qu.:11.00   3rd Qu.: 1250000   3rd Qu.: 96.00
##  Max.   :640.0   Max.   :20.00   Max.   :75000000   Max.   :216.00
##
##   nama_negara      nama_kota      lokasi_mesin  pemilik_mesin
##  Min.   : 1.00   Min.   :  1.0   Min.   :   2   Min.   :   1
##  1st Qu.: 5.00   1st Qu.:102.0   1st Qu.:1914   1st Qu.: 613
##  Median : 5.00   Median :128.0   Median :3720   Median : 613
##  Mean   : 5.02   Mean   :148.3   Mean   :3948   Mean   : 766
##  3rd Qu.: 5.00   3rd Qu.:203.0   3rd Qu.:5637   3rd Qu.: 613
##  Max.   :16.00   Max.   :293.0   Max.   :8697   Max.   :2688
##
##  waktu_transaksi  kuartal_transaksi kepemilikan_kartu  nama_channel
##  Min.   :    47   Min.   :1.000     Min.   :1.000     Min.   :1.000
##  1st Qu.:102622   1st Qu.:2.000     1st Qu.:2.000     1st Qu.:1.000
##  Median :140707   Median :3.000     Median :2.000     Median :1.000
##  Mean   :138896   Mean   :2.855     Mean   :1.932     Mean   :1.404
##  3rd Qu.:175420   3rd Qu.:3.000     3rd Qu.:2.000     3rd Qu.:1.000
##  Max.   :235914   Max.   :4.000     Max.   :2.000     Max.   :5.000
##
##    id_channel    rata_rata_nilai_transaksi maksimum_nilai_transaksi
##  Min.   :3.000   Min.   :   50000          Min.   :    38000
##  1st Qu.:9.000   1st Qu.:  568563          1st Qu.:  2500000
##  Median :9.000   Median : 1024239          Median :  6000000
##  Mean   :8.237   Mean   : 1364132          Mean   : 12287603
##  3rd Qu.:9.000   3rd Qu.: 1679778          3rd Qu.: 15000000
##  Max.   :9.000   Max.   :24666667          Max.   :100000000
##                  NA's   :21                NA's   :21
##  minimum_nilai_transaksi rata_rata_jumlah_transaksi flag_transaksi_fraud
##  Min.   :       1        Min.   : 1.000             Min.   :0.00000
##  1st Qu.:   25000        1st Qu.: 1.680             1st Qu.:0.00000
##  Median :   36964        Median : 2.100             Median :0.00000
##  Mean   :   76519        Mean   : 2.436             Mean   :0.06933
##  3rd Qu.:   63200        3rd Qu.: 2.790             3rd Qu.:0.00000
##  Max.   :75000000        Max.   :19.780             Max.   :1.00000
##  NA's   :21              NA's   :21
```

Perintah berikut bertujuan untuk menghitung berapa jumlah NA dalam tiap kolom pada objek “dataTrain”

```
colSums(is.na(dataTrain))
```

```
##                          X  id_tanggal_transaksi_awal
##                          0                          0
##     tanggal_transaksi_awal                 tipe_kartu
##                          0                          0
##                id_merchant              nama_merchant
##                          0                          0
##                 tipe_mesin             tipe_transaksi
##                          0                          0
##             nama_transaksi            nilai_transaksi
##                          0                          0
##                  id_negara               nama_negara
##                          0                          0
##                  nama_kota               lokasi_mesin
##                          0                          0
##              pemilik_mesin            waktu_transaksi
##                          0                          0
##          kuartal_transaksi          kepemilikan_kartu
##                          0                          0
##               nama_channel                 id_channel
##                          0                          0
##  rata_rata_nilai_transaksi   maksimum_nilai_transaksi
##                         21                         21
##    minimum_nilai_transaksi rata_rata_jumlah_transaksi
##                         21                         21
##       flag_transaksi_fraud
##                          0
```

Untuk menghilangkan nilai NA yang ada pada dataset tersebut, digunakan perintah sebagai berikut :

```
dataTrain <- dataTrain[complete.cases(dataTrain),]
```

Dalam “dataTrain”, semua variabel direpresentasikan dalam bentuk angka atau integer sehingga sistem menganggap semuanya numerical. Namun pada kenyataannya, terdapat beberapa variabel yang merupakan kategorikal, yaitu “flag_transaksi_fraud”.

```
dataTrain$flag_transaksi_fraud <-
factor(as.character(dataTrain$flag_transaksi_fraud))
```

Menggunakan cara yang sama dengan “dataTrain”, proses pengolahan data diulang kembali untuk mengolah “dataTest”

```
dataTest <- read.csv('fraud_test.csv')
summary(dataTest)
```

```
##        X         id_tanggal_transaksi_awal tanggal_transaksi_awal
##  Min.   :   18   Min.   :2457297           Min.   :2457302
##  1st Qu.: 3784   1st Qu.:2457407           1st Qu.:2457456
##  Median : 7355   Median :2457507           Median :2457554
```

```
##  Mean   : 7459   Mean   :2457495           Mean   :2457547
##  3rd Qu.:11231   3rd Qu.:2457592           3rd Qu.:2457640
##  Max.   :14996   Max.   :2457662           Max.   :2457754
##
##    tipe_kartu      id_merchant     nama_merchant    tipe_mesin
##  Min.   :  0.00   Min.   :    -2   Min.   :   1   Min.   :     -4
##  1st Qu.: 93.00   1st Qu.:    -2   1st Qu.:1798   1st Qu.:1092712
##  Median :104.00   Median :    -2   Median :1798   Median :1842854
##  Mean   : 88.16   Mean   : 40161   Mean   :1666   Mean   :1646143
##  3rd Qu.:111.00   3rd Qu.:    -2   3rd Qu.:1798   3rd Qu.:2432934
##  Max.   :138.00   Max.   :699429   Max.   :1857   Max.   :6923365
##
##  tipe_transaksi nama_transaksi  nilai_transaksi      id_negara
##  Min.   : 26    Min.   : 1.00   Min.   :    9500   Min.   : 57.00
##  1st Qu.: 26    1st Qu.: 9.00   1st Qu.:  200000   1st Qu.: 96.00
##  Median :156    Median :10.00   Median :  580000   Median : 96.00
##  Mean   :181    Mean   :10.72   Mean   : 1275767   Mean   : 96.11
##  3rd Qu.:301    3rd Qu.:11.00   3rd Qu.: 1250000   3rd Qu.: 96.00
##  Max.   :640    Max.   :19.00   Max.   :64178000   Max.   :183.00
##
##   nama_negara       nama_kota      lokasi_mesin  pemilik_mesin
##  Min.   : 5.000   Min.   :  1.0   Min.   :   3   Min.   :  22.0
##  1st Qu.: 5.000   1st Qu.:108.0   1st Qu.:1935   1st Qu.: 613.0
##  Median : 5.000   Median :128.0   Median :3786   Median : 613.0
##  Mean   : 5.013   Mean   :150.3   Mean   :3965   Mean   : 779.1
##  3rd Qu.: 5.000   3rd Qu.:208.0   3rd Qu.:5648   3rd Qu.: 613.0
##  Max.   :15.000   Max.   :293.0   Max.   :8697   Max.   :2688.0
##
##  waktu_transaksi  kuartal_transaksi kepemilikan_kartu  nama_channel
##  Min.   :   253   Min.   :1.000     Min.   :1.00      Min.   :1.000
##  1st Qu.:102793   1st Qu.:2.000     1st Qu.:2.00      1st Qu.:1.000
##  Median :141534   Median :3.000     Median :2.00      Median :1.000
##  Mean   :139303   Mean   :2.861     Mean   :1.93      Mean   :1.426
##  3rd Qu.:175764   3rd Qu.:3.000     3rd Qu.:2.00      3rd Qu.:1.000
##  Max.   :235734   Max.   :4.000     Max.   :2.00      Max.   :5.000
##
##    id_channel    flag_transaksi_finansial status_transaksi
##  Min.   :3.000   Mode :logical            Min.   :3
##  1st Qu.:9.000   FALSE:1875               1st Qu.:3
##  Median :9.000                            Median :3
##  Mean   :8.183                            Mean   :3
##  3rd Qu.:9.000                            3rd Qu.:3
##  Max.   :9.000                            Max.   :3
##
##  bank_pemilik_kartu rata_rata_nilai_transaksi maksimum_nilai_transaksi
##  Min.   :999        Min.   :   89010          Min.   :   100000
##  1st Qu.:999        1st Qu.:  560535          1st Qu.:  2500000
##  Median :999        Median : 1014391          Median :  6450000
##  Mean   :999        Mean   : 1381495          Mean   : 12015116
##  3rd Qu.:999        3rd Qu.: 1665216          3rd Qu.: 15000000
```

```
##  Max.   :999        Max.   :75000000          Max.   :100000000
##                     NA's   :3                 NA's   :3
##  minimum_nilai_transaksi rata_rata_jumlah_transaksi
##  Min.   :       1        Min.   : 1.000
##  1st Qu.:   24875        1st Qu.: 1.680
##  Median :   35300        Median : 2.120
##  Mean   :  109733        Mean   : 2.431
##  3rd Qu.:   60000        3rd Qu.: 2.790
##  Max.   :75000000        Max.   :19.780
##  NA's   :3               NA's   :3
```

```
dataTest$flag_transaksi_finansial <- NULL
dataTest$status_transaksi <- NULL
dataTest$bank_pemilik_kartu <- NULL
summary(dataTest)
```

```
##        X         id_tanggal_transaksi_awal tanggal_transaksi_awal
##  Min.   :   18   Min.   :2457297           Min.   :2457302
##  1st Qu.: 3784   1st Qu.:2457407           1st Qu.:2457456
##  Median : 7355   Median :2457507           Median :2457554
##  Mean   : 7459   Mean   :2457495           Mean   :2457547
##  3rd Qu.:11231   3rd Qu.:2457592           3rd Qu.:2457640
##  Max.   :14996   Max.   :2457662           Max.   :2457754
##
##    tipe_kartu      id_merchant     nama_merchant    tipe_mesin
##  Min.   :  0.00   Min.   :    -2   Min.   :   1   Min.   :     -4
##  1st Qu.: 93.00   1st Qu.:    -2   1st Qu.:1798   1st Qu.:1092712
##  Median :104.00   Median :    -2   Median :1798   Median :1842854
##  Mean   : 88.16   Mean   : 40161   Mean   :1666   Mean   :1646143
##  3rd Qu.:111.00   3rd Qu.:    -2   3rd Qu.:1798   3rd Qu.:2432934
##  Max.   :138.00   Max.   :699429   Max.   :1857   Max.   :6923365
##
##  tipe_transaksi nama_transaksi  nilai_transaksi      id_negara
##  Min.   : 26    Min.   : 1.00   Min.   :    9500   Min.   : 57.00
##  1st Qu.: 26    1st Qu.: 9.00   1st Qu.:  200000   1st Qu.: 96.00
##  Median :156    Median :10.00   Median :  580000   Median : 96.00
##  Mean   :181    Mean   :10.72   Mean   : 1275767   Mean   : 96.11
##  3rd Qu.:301    3rd Qu.:11.00   3rd Qu.: 1250000   3rd Qu.: 96.00
##  Max.   :640    Max.   :19.00   Max.   :64178000   Max.   :183.00
##
##   nama_negara       nama_kota      lokasi_mesin  pemilik_mesin
##  Min.   : 5.000   Min.   :  1.0   Min.   :   3   Min.   :  22.0
##  1st Qu.: 5.000   1st Qu.:108.0   1st Qu.:1935   1st Qu.: 613.0
##  Median : 5.000   Median :128.0   Median :3786   Median : 613.0
##  Mean   : 5.013   Mean   :150.3   Mean   :3965   Mean   : 779.1
##  3rd Qu.: 5.000   3rd Qu.:208.0   3rd Qu.:5648   3rd Qu.: 613.0
##  Max.   :15.000   Max.   :293.0   Max.   :8697   Max.   :2688.0
##
##  waktu_transaksi  kuartal_transaksi kepemilikan_kartu  nama_channel
##  Min.   :   253   Min.   :1.000     Min.   :1.00      Min.   :1.000
```

```
##  1st Qu.:102793   1st Qu.:2.000     1st Qu.:2.00      1st Qu.:1.000
##  Median :141534   Median :3.000     Median :2.00      Median :1.000
##  Mean   :139303   Mean   :2.861     Mean   :1.93      Mean   :1.426
##  3rd Qu.:175764   3rd Qu.:3.000     3rd Qu.:2.00      3rd Qu.:1.000
##  Max.   :235734   Max.   :4.000     Max.   :2.00      Max.   :5.000
##
##    id_channel    rata_rata_nilai_transaksi maksimum_nilai_transaksi
##  Min.   :3.000   Min.   :   89010          Min.   :   100000
##  1st Qu.:9.000   1st Qu.:  560535          1st Qu.:  2500000
##  Median :9.000   Median : 1014391          Median :  6450000
##  Mean   :8.183   Mean   : 1381495          Mean   : 12015116
##  3rd Qu.:9.000   3rd Qu.: 1665216          3rd Qu.: 15000000
##  Max.   :9.000   Max.   :75000000          Max.   :100000000
##                  NA's   :3                 NA's   :3
##  minimum_nilai_transaksi rata_rata_jumlah_transaksi
##  Min.   :       1        Min.   : 1.000
##  1st Qu.:   24875        1st Qu.: 1.680
##  Median :   35300        Median : 2.120
##  Mean   :  109733        Mean   : 2.431
##  3rd Qu.:   60000        3rd Qu.: 2.790
##  Max.   :75000000        Max.   :19.780
##  NA's   :3               NA's   :3

dataTest <- dataTest[complete.cases(dataTest),]
```

Dengan menggunakan Random Forest, "dataTest" diolah sehingga menghasilkan 'flag_transaksi_fraud' yang diinginkan beserta probabiltas (certainty score).

```
set.seed(100)
modelFit <- randomForest(flag_transaksi_fraud ~., data = dataTrain)
prediction <- predict(modelFit, dataTest)
dataTest$prediction <- prediction
probability <- predict(modelFit, dataTest, type = "prob")
dataTest$probability <- ifelse(dataTest$prediction == 1, probability[,-1],
probability[,-2])
```

Perintah berikut berfungsi mengubah objek ke dalam bentuk '.csv'

```
fraud <- dataTest[,-c(2:24)]
write.csv(fraud, 'fraud_challenge.csv', row.names = FALSE)
```