Fraudulent or not?

Sari Vesiluoma

6 11 2019

Introduction

The goal in this project is to learn how to predict a fraudulent financial transaction. The data used here is called Synthetic Financial Datasets for Fraud Detection generated by the PaySim mobile money simulator (https://www.kaggle.com/ntnu-testimon/paysim1).As described on the web page, the dataset is a synthetic one, generated using the simulator called PaySim. It uses aggregated data from a private dataset to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behaviour.

PaySim simulates mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The synthetic dataset is scaled down 1/4 of the original dataset.

I have downloaded the dataset from the net (the link above) and I have unzipped it to the same folder where my R script and the rmd file are. Here, I am reading the data from my folder.

<<<<<<< HEAD

The dataset, here referred with a variable name fraud_or_not, has the following dimensions

```
## Parsed with column specification:
## cols(
##   step = col_double(),
##   type = col_character(),
##   amount = col_double(),
##   nameOrig = col_character(),
##   oldbalanceOrg = col_double(),
##   newbalanceOrig = col_double(),
##   nameDest = col_character(),
##   oldbalanceDest = col_double(),
##   newbalanceDest = col_double(),
##   isFraud = col_double(),
##   isFlaggedFraud = col_double()
## )
```

```
## [1] 6362620      11
```

Next I will analyse the data and split it to training and test sets. I will use different machine learning algorithms to try to predict which transaction is fraudulent and which not. In this kind of a case the speciality is that the amount of fraudulent transactions is very minor compared to the amount of non-fraudulent transactions, as we will see.

Analysis

Let's look the data first as is. Like can be seen from the summary below, there are e.g. no NA values which would need to be cleaned.

```r
summary(fraud_or_not)
```

```
##       step           type               amount           nameOrig
##  Min.   :  1.0   Length:6362620     Min.   :       0   Length:6362620
##  1st Qu.:156.0   Class :character   1st Qu.:   13390   Class :character
##  Median :239.0   Mode  :character   Median :   74872   Mode  :character
##  Mean   :243.4                      Mean   :  179862
##  3rd Qu.:335.0                      3rd Qu.:  208721
##  Max.   :743.0                      Max.   :92445517
##  oldbalanceOrg      newbalanceOrig       nameDest
##  Min.   :       0   Min.   :       0   Length:6362620
##  1st Qu.:       0   1st Qu.:       0   Class :character
##  Median :   14208   Median :       0   Mode  :character
##  Mean   :  833883   Mean   :  855114
##  3rd Qu.:  107315   3rd Qu.:  144258
##  Max.   :59585040   Max.   :49585040
##  oldbalanceDest      newbalanceDest         isFraud
##  Min.   :        0   Min.   :        0   Min.   :0.000000
##  1st Qu.:        0   1st Qu.:        0   1st Qu.:0.000000
##  Median :   132706   Median :   214661   Median :0.000000
##  Mean   :  1100702   Mean   :  1224996   Mean   :0.001291
##  3rd Qu.:   943037   3rd Qu.:  1111909   3rd Qu.:0.000000
##  Max.   :356015889   Max.   :356179279   Max.   :1.000000
##  isFlaggedFraud
##  Min.   :0.0e+00
##  1st Qu.:0.0e+00
##  Median :0.0e+00
##  Mean   :2.5e-06
##  3rd Qu.:0.0e+00
##  Max.   :1.0e+00
```

```r
str(fraud_or_not)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 6362620 obs. of  11 variables:
##  $ step          : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ type          : chr  "PAYMENT" "PAYMENT" "TRANSFER" "CASH_OUT" ...
##  $ amount        : num  9840 1864 181 181 11668 ...
##  $ nameOrig      : chr  "C1231006815" "C1666544295" "C1305486145" "C840083671" ...
##  $ oldbalanceOrg : num  170136 21249 181 181 41554 ...
##  $ newbalanceOrig: num  160296 19385 0 0 29886 ...
##  $ nameDest      : chr  "M1979787155" "M2044282225" "C553264065" "C38997010" ...
##  $ oldbalanceDest: num  0 0 0 21182 0 ...
##  $ newbalanceDest: num  0 0 0 0 0 ...
##  $ isFraud       : num  0 0 1 1 0 0 0 0 0 0 ...
##  $ isFlaggedFraud: num  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   step = col_double(),
##   ..   type = col_character(),
##   ..   amount = col_double(),
##   ..   nameOrig = col_character(),
##   ..   oldbalanceOrg = col_double(),
##   ..   newbalanceOrig = col_double(),
##   ..   nameDest = col_character(),
##   ..   oldbalanceDest = col_double(),
##   ..   newbalanceDest = col_double(),
##   ..   isFraud = col_double(),
##   ..   isFlaggedFraud = col_double()
##   .. )
```

```r
fraud_or_not %>% head()
```

```
## # A tibble: 6 x 11
##    step type  amount nameOrig oldbalanceOrg newbalanceOrig nameDest
##   <dbl> <chr>  <dbl> <chr>            <dbl>          <dbl> <chr>
## 1     1 PAYM~  9840. C123100~        170136        160296. M197978~
## 2     1 PAYM~  1864. C166654~         21249         19385. M204428~
## 3     1 TRAN~   181  C130548~           181             0  C553264~
## 4     1 CASH~   181  C840083~           181             0  C389970~
## 5     1 PAYM~ 11668. C204853~         41554         29886. M123070~
## 6     1 PAYM~  7818. C900456~         53860         46042. M573487~
## # ... with 4 more variables: oldbalanceDest <dbl>, newbalanceDest <dbl>,
## #   isFraud <dbl>, isFlaggedFraud <dbl>
```

The data has 11 columns which are:

Table 1: Explanations of the features

feature	expl
step	maps a unit of time in the real world. 1 step is 1 hour of time. Total steps 744 (30 days simulation).
type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
amount	amount of the transaction in local currency.
nameOrig	customer who started the transaction
oldbalanceOrg	initial balance before the transaction
newbalanceOrig	new balance after the transaction
nameDest	customer who is the recipient of the transaction
oldbalanceDest	initial balance recipient before the transaction
newbalanceDest	new balance recipient after the transaction
isFraud	Transactions made by the fraudulent agents inside the simulation
isFlaggedFraud	An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

Results

Conclusion