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.

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

```
## [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.

```
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
```

The data has 11 columns which are:

feature	expl
step	maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simu
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. Note that there is not information for customers that star
newbalanceDest	new balance recipient after the transaction. Note that there is not information for customers that start w
isFraud	This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the
isFlaggedFraud	The business model aims to control massive transfers from one account to another and flags illegal attem

Results

Conclusion