R Project Report

Jude Al-Aitani

05/02/2021

# Credit Card Fraud Prediction

### Packages used

library("tidyr")  
library("caret")  
library("parallel")  
library("doSNOW")

### Summary

Knowing ones money is kept secure in case of any financial emergency is essential for personal safety needs and for peace of mind. Banks are accountable for providing clients with financial security and are held responsible for protecting bank accounts from any suspicious fraudulent transactions. This project is conducted to support banks in managing credit card transactions by creating a supervised machine learning model that will be able to accurately identify fraudulent attempts. Consequently, this model will help increase accuracy of online transactions and banks will be able to apply higher security measures and provide safer banking experiences.

### Introduction

Modern-day robbery is no longer about armed criminals storming banks wearing masks. Instead, criminals nowadays developed sophisticated tactics of cyber hacking bank accounts. In this era of digitalization, financial fraud is a drastically increasing problem in scope and ubiquity. Therefore, the urgency of developing fraud detection and prevention systems has become a higher priority.

### Literature Review

The development of new technologies in the internet and e-commerce sector cultivated digitalization in payment methods by providing a variety of electronic payment options including payment cards (credit, debit, pre-paid), digital and mobile wallets, bank transfers, electronic cash, contactless payment methods and cash on delivery. As a result of the rise and rapid use of credit cards for online purchases, credit card fraud has dramatically increased. Namely, credit card fraud is essentially known as the crime of deceiving and scamming individuals in their financial transactions. In fact, according to the Federal Trade Commission, credit card fraud ranked second place for the topmost common types of identity theft for two year in a row. Moreover, (Shen et al., 2007) stated that credit card transactions had a total loss of over 700 million dollars of fraud in the U.S.A. in 2004. Because of the infeasibility of manually processing the infinite numbers of online transactions occurring at the same time per second. Anti-fraud machine learning models are employed by financial institutions to counter fraudulent transactions. The goal of this project is to continuously develop models that support fraudulent detection. Fraudulent detection is defined as the act of recognizing indications of fraud where no prior suspicion or tendency to fraud exists (Jayasree & Balan, 2013). The authors emphasized the importance of identifying cybersecurity breaches to disable fraudulent activities. In other words, developing sophisticated approaches to prevent fraud attacks by implementing advanced analytical techniques to protect oneself of becoming a victim to cybercriminal theft.

### Data

The Credit Card Transactions Fraud Detection Dataset for this project was sourced from [here](https://www.kaggle.com/kartik2112/fraud-detection). The dataset in Kaggle was cleaned and merged therefore further filtering was not required.

### Methodology

The training dataset and testing datasets that were sourced from Kaggle were reviewed using Microsoft Excel. Additional information available on the Kaggle website (linked above) was also studied.

After having a thorough understanding of the data, the data was imported into R studio. In R studio, a subset of the data was chosen that contained selected columns using the following command:

trainingData <- read.csv("C:\\Users\\mizzj\\OneDrive\\Desktop\\R proj\\New folder\\fraudTrain.csv", header=TRUE)  
  
#Set up factors  
trainingData$merchant <- as.factor(trainingData$merchant)  
trainingData$category <- as.factor(trainingData$category)  
trainingData$amt <- as.factor(trainingData$amt)  
trainingData$gender <- as.factor(trainingData$gender)  
trainingData$city <- as.factor(trainingData$city)  
trainingData$state <- as.factor(trainingData$state)  
trainingData$zip <- as.factor(trainingData$zip)  
trainingData$city\_pop <- as.factor(trainingData$city\_pop)  
trainingData$job <- as.factor(trainingData$job)  
trainingData$is\_fraud <- as.factor(trainingData$is\_fraud)  
  
#Subset data to the features we wish to use  
features <- c("merchant", "category", "amt", "gender", "city", "state", "zip", "city\_pop",  
 "job", "is\_fraud")  
  
trainingData <- trainingData[, features]

Once the columns were picked and the training dataset was ready, the next step was to setup the parameters that will be used for creating and tuning the model. For this project, repeatedcv method was used in conjunction with grid search. Grid parameters were also provided.

train.control <- trainControl(method="repeatedcv", number=2, repeats = 2, search="grid")  
  
tune.grid <- expand.grid(n.trees = seq(10, 1000, by = 100)  
 , interaction.depth = 1  
 ,shrinkage = 0.1  
 , n.minobsinnode = c(5, 10, 20, 30) )

In order to speed up the model creating process, virtual machines were setup in the background with the help of parallel and doSNOW packages.

cl <- makeCluster(5, type = "SOCK")  
registerDoSNOW(cl)

After having the requisites in place, the model was then created using the following command. Once the model was created, the virtual machines that were setup for parallel computing were stopped.

caretModel <- train(is\_fraud ~ .,  
 data= trainingData,  
 method='gbm',  
 tuneGrid= tune.grid,  
 trControl= train.control)

## Iter TrainDeviance ValidDeviance StepSize Improve  
## 1 1.3773 nan 0.1000 0.0041  
## 2 1.3695 nan 0.1000 0.0035  
## 3 1.3629 nan 0.1000 0.0030  
## 4 1.3571 nan 0.1000 0.0025  
## 5 1.3509 nan 0.1000 0.0025  
## 6 1.3463 nan 0.1000 0.0017  
## 7 1.3418 nan 0.1000 0.0021  
## 8 1.3378 nan 0.1000 0.0014  
## 9 1.3332 nan 0.1000 0.0023  
## 10 1.3294 nan 0.1000 0.0023  
## 20 1.2986 nan 0.1000 0.0009  
## 40 1.2628 nan 0.1000 0.0006  
## 60 1.2349 nan 0.1000 0.0002  
## 80 1.2117 nan 0.1000 0.0001  
## 100 1.1915 nan 0.1000 0.0003  
## 120 1.1709 nan 0.1000 0.0002  
## 140 1.1533 nan 0.1000 -0.0003  
## 160 1.1353 nan 0.1000 -0.0002  
## 180 1.1180 nan 0.1000 0.0003  
## 200 1.1017 nan 0.1000 -0.0000  
## 220 1.0861 nan 0.1000 -0.0001  
## 240 1.0717 nan 0.1000 0.0001  
## 260 1.0569 nan 0.1000 0.0001  
## 280 1.0422 nan 0.1000 0.0001  
## 300 1.0289 nan 0.1000 -0.0003  
## 320 1.0157 nan 0.1000 -0.0002  
## 340 1.0028 nan 0.1000 0.0000  
## 360 0.9899 nan 0.1000 -0.0003  
## 380 0.9781 nan 0.1000 0.0000  
## 400 0.9654 nan 0.1000 -0.0001  
## 420 0.9538 nan 0.1000 -0.0001  
## 440 0.9412 nan 0.1000 -0.0002  
## 460 0.9303 nan 0.1000 -0.0001  
## 480 0.9182 nan 0.1000 0.0000  
## 500 0.9077 nan 0.1000 -0.0003  
## 520 0.8962 nan 0.1000 -0.0000  
## 540 0.8853 nan 0.1000 0.0001  
## 560 0.8747 nan 0.1000 -0.0003  
## 580 0.8649 nan 0.1000 -0.0005  
## 600 0.8549 nan 0.1000 -0.0006  
## 620 0.8452 nan 0.1000 -0.0001  
## 640 0.8351 nan 0.1000 0.0001  
## 660 0.8257 nan 0.1000 0.0001  
## 680 0.8158 nan 0.1000 0.0001  
## 700 0.8081 nan 0.1000 0.0001  
## 720 0.8001 nan 0.1000 -0.0004  
## 740 0.7915 nan 0.1000 -0.0002  
## 760 0.7826 nan 0.1000 -0.0000  
## 780 0.7740 nan 0.1000 -0.0000  
## 800 0.7659 nan 0.1000 -0.0001  
## 820 0.7579 nan 0.1000 -0.0001  
## 840 0.7498 nan 0.1000 0.0001  
## 860 0.7415 nan 0.1000 -0.0001  
## 880 0.7335 nan 0.1000 -0.0001  
## 900 0.7258 nan 0.1000 0.0002  
## 910 0.7220 nan 0.1000 0.0000

stopCluster(cl)

The model was then applied to the test data and prediction was stored in the variable preds using the following command.

testData <- read.csv("C:\\Users\\mizzj\\OneDrive\\Desktop\\R proj\\New folder\\fraudTest.csv", header=TRUE)  
preds <- predict(caretModel, data= testData)

### Results

The supervised machine learning model created by the Caret package detected the fraud and legitimate transactions with an accuracy over 90%.

confusionMatrix(preds, as.factor(testData$is\_fraud))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 911 86  
## 1 89 915  
##   
## Accuracy : 0.9125   
## 95% CI : (0.8993, 0.9246)  
## No Information Rate : 0.5002   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8251   
##   
## Mcnemar's Test P-Value : 0.8798   
##   
## Sensitivity : 0.9110   
## Specificity : 0.9141   
## Pos Pred Value : 0.9137   
## Neg Pred Value : 0.9114   
## Prevalence : 0.4998   
## Detection Rate : 0.4553   
## Detection Prevalence : 0.4983   
## Balanced Accuracy : 0.9125   
##   
## 'Positive' Class : 0   
##

### Implications

1. Machine learning models in fraud analysis is a project under continuous technological development due the unique nature of each fraudulent transaction. Despite continuous efforts of development in anti-fraud machine learning models, countering fraudulent transactions could never be eradicated.
2. The implication of the inability of processing Big data on my personal device restricted the ability to process a larger dimension of parameters required to better train the model to provide higher accuracy. Building supervised machine learning models requires devices with dynamic capabilities.

### Conclusion

1. Utilizing advanced supervised machine learning models that provide the highest accurate predictions will enable banks to stop fraud transactions before they occur that will result in eliminating the substantial financial losses associated with disputed transactions and chargebacks.
2. It is crucial for both parties involved, banks and their consumers to consider applying the higher security measures to reduce fraud. For example: using virtual credit cards, deleting credit card information after each transaction, encrypting sensitive data, conducting frequent updates on software and web applications, and only selecting highly trusted and secure e-commerce platforms and payment processors.
3. Combating fraud is an activity that requires joining all possible resources. Since regulators, banks, and investment firms are continuously involved in monitoring suspicious activities and possible money laundering cases they must detect and inform each other about suspicious activities.

### References

Jayasree (R. V. S. (2013)) Shen (2007, June) (n.d.)

n.d. <https://public.tableau.com/profile/federal.trade.commission#!/vizhome/TheBigViewAllSentinelReports/TopReports>.

Jayasree, & Balan, V. R. V. S. (2013). “A Review on Data Mining in Banking Sector. American Journal of Applied Sciences.”

Shen, Tong, A. 2007, June. “Application of Classification Models on Credit Card Fraud Detection. In 2007 International Conference on Service Systems and Service Management.”