Credit Card Fraud Detection

The Problem

The aim of this project is to use computer simulation for fraud detection in mobile money financial transactions. Fraud prevention is becoming a critical driver for the financial services industry. In the recent years, due to an increase in use of new technologies such as cloud and mobile financial services, the fraud problem has been intensified. Therefore achieving an accurate and less intrusive fraud detection system is crucial and banks and financial service institutions are increasingly investing in algorithms and data analytics technology to spot and combat fraud.

Clients

The primary client for this project would be banks and financial services which provide mobile money transactions. However, financial fraud is a problem which affects the finance industry, government, corporate sectors, and ordinary consumers, and therefore identifying and preventing fraud can be beneficial for many different clients.

Current Dataset

Currently, there is a lack of public research into the detection of fraud. One important reason is shortage of transaction data due to confidentiality issues. To overcome this problem, a synthetic dataset generated using the simulator called PaySim is used in this project. PaySim uses aggregated data from the private dataset to generate a synthetic dataset that resembles the normal operation of transactions and injects malicious behaviour to later evaluate the performance of fraud detection methods. The data is described in details in the following PhD thesis: http://urn.kb.se/resolve?urn=urn:nbn:se:bth-12932

Description:

The Paysim synthetic dataset of mobile money transactions available on Kaggle is used for this project. The transaction data is presented in different steps, each step representing an hour of simulation. The raw data consists of 11 columns and about 6362620 rows. The description of each column is as follow:

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 simulation).
type: CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
amount: Amount of the transaction in local currency.
nameOrig: Customer who started the transaction.
oldbalanceOrig: 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 start with M (Merchants).
newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
isFraud: This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.
isFlaggedFraud: The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

Transaction types are defined based on the following reference: http://bth.diva-portal.org/smash/get/diva2:955852/FULLTEXT06.pdf
CASH-IN is the process of increasing the balance of account by paying in cash to a merchant.
CASH-OUT is the opposite process of CASH-IN, it means to withdraw cash from a merchant which decreases the balance of the account.
DEBIT is similar process than CASH-OUT and involves sending the money from the mobile money service to a bank account.
PAYMENT is the process of paying for goods or services to merchants which decreases the balance of the account and increases the balance of the receiver.

TRANSFER is the process of sending money to another user of the service through the mobile money platform.

Data Wrangling

Other Potential Datasets

Kaggle's credit card fraud detection data:

The datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, the original features and more background information about the data is not provided.

German credit fraud dataset:

This dataset classifies people described by a set of attributes as good or bad credit risks. It contains categorical and integer data with 20 features and 1000 instances.

Initial Findings

The main technical challenge in predicting fraud is the highly imbalanced distribution between legitimate and fraudulent classes in 6 million rows of data. Another deficiency of this data stems from the possible discrepancies in its description and some redundant column values. The goal of this project is to solve these issues by a detailed data exploration and wrangling followed by choosing a suitable machine-learning algorithm to deal with the skew. Supervised classification algorithms will be used to predict fraudulent transactions.

Notes for Further Exploration