CS4035 - Cyber Data Analytics
Assignment 1: Credit Card Fraud Detection

By : Hendra Hadhil Choiri (4468457)
Prahesa Kusuma Setia (4501543)

Introduction

Credit card fraud detection is a challenging task. Given a large labeled transaction data, which only a tiny amount of transactions is labeled as fraudulent. The features are extracted based on some observations and analyses. Then, the data are resampled to handle the imbalance data. After that, the classifier will be trained by using these data to be able to detect the fraud.

Data Observation

First task to be done is understanding the data. In the given data, there are 17 attributes. Some of important attributes to be considered are: currencycode (there are 5 currencies in the data: MXN, AUD, NZD, GBP, SEK), amount, creationdate, issuercountrycode, shoppercountrycode, card_id, and simple_journal. Simple_journal is used as the output label of fraud detection, there are 3 cases: Chargeback (fraud), Settled (normal), and Refused (rejected). We cannot make sure whether refused transaction tends to be fraud or normal, so we remove these transactions (even though these data can be exploited by using semi-supervised learning).

The data are then analyzed by doing below observations (by using Excel). The values to be compared are mainly the average amounts and number of transactions in the case of each fraud and normal transactions.

Observation based on currency. It is not fair to compare the transaction amounts of different currency. So, we averaged the amounts per currency, then compare it for the case of fraud and normal transactions. The visualization is shown in Figure 1. It can be observed that the average amount of fraud is always higher than average amount of normal transaction (which means also higher than average amount of overall amount average).

Observation based on transaction timestamp. Next analysis is beyond the transaction date. The daily number of transactions is shown in Figure 2. It is easily recognized that on 8 October 2015, the number of transactions is very high, maybe there is shopping event. However, overall there is no high corellation between number of fraud to number of daily transactions. If we check the daily average transaction amount, we can see in Figure 3 that fraudulent transaction amount is usually higher than average amount of transactions in that day.

We are also interested in what time the transactions are usually done. Figure 4 shows the number of transactions in hourly basis. We can observe that generally, at midnight to the morning, the number of transactions is less than in the afternoon. For fraudulent transactions, we can check that the probability that the fraud happens in the morning is slightly higher than in the afternoon.

Feature Extraction

Based on the observation and some considerations, some attributes are extracted from the data. Thase attributes will be used as features for machine learning. Feature extractions and machine learning are done by using Python (source code attached).

average_amount_daily : average of amount in the same date
relative_amount : amount - average_amount_daily with respect to threshold set
transaction_hour : hour part of creationdate
is_same_currency_shopper : binary value, true if the currency matches with the country
is_same_issuer_shopper : binary value, true if issuer and shopper are in the same country

Handling of Imbalance Data & Machine Learning

In the given data, there are 345 samples of fraudulent transaction. This number is very small compared to 236691, the number of normal transactions. To handle this imbalance data, we used SMOTE + Tomek.
For the machine learning, we consider various classifiers to be trained: 1-NN, 3-NN, and RandomForest. We also use 10-fold Cross Validation to measure the performance of our classifiers.

The metrics we used are precision, recall, accuracy, and F score. The result can be shown in Table 1. We observe that Random Forest perform the best when we increase the ratio of fraud data to 3%.

Table 1. Classification performance of various classifiers and data imbalance resampling

Classsifier	% Fraud	Precision	Recall	Accuracy	F1
1-NN	**0.14% (original)**	0.9986	0.9602	0.9589	0.9761
1-NN	**1.50%**	0.9943	0.9084	0.9047	0.9408
1-NN	**3%**	0.9916	0.8997	0.8955	0.9341
3-NN	**0.14% (original)**	0.9986	0.9699	0.9686	0.9819
3-NN	**1.50%**	0.9923	0.9190	0.9132	0.9467
3-NN	**3%**	0.9887	0.9084	0.9012	0.9378
RandomForest	**0.14% (original)**	0.9986	0.9608	0.9553	0.9713
RandomForest	**1.50%**	0.9962	0.9111	0.9036	0.9351
RandomForest	**3%**	0.9962	0.9010	0.8955	0.9414

White-box Explanation

From the data observation, it can be concluded that if a transaction is fraudulent, the amount is more likely to be above average amount of that day. Then, by using 1-NN classifier, whenever we find a transaction that is predicted as fraud, we can find which a sample of fraud similar to that transaction.

Appendix

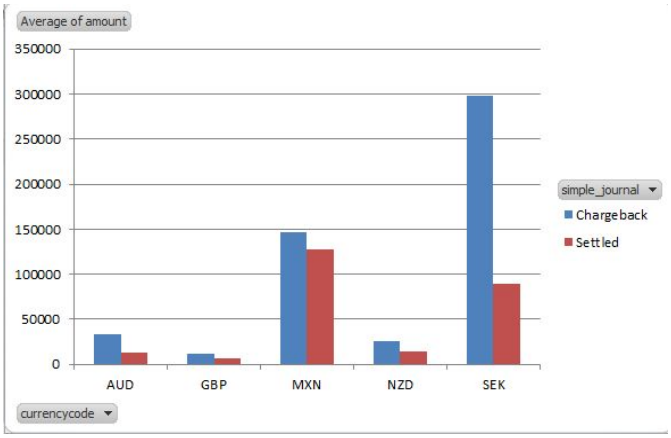

Figure 1. Average transaction amount of each currency

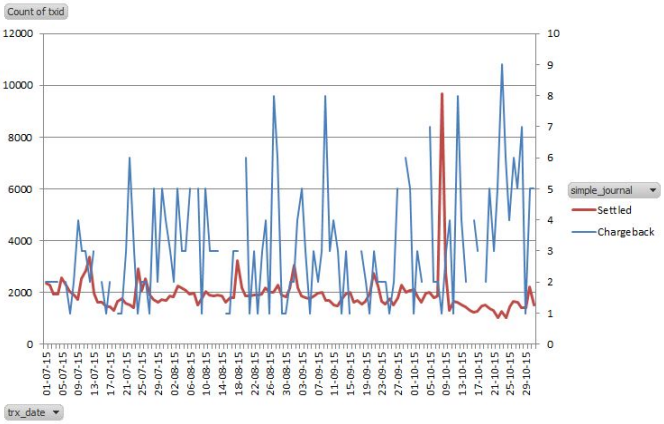

Figure 2. Daily number of transaction

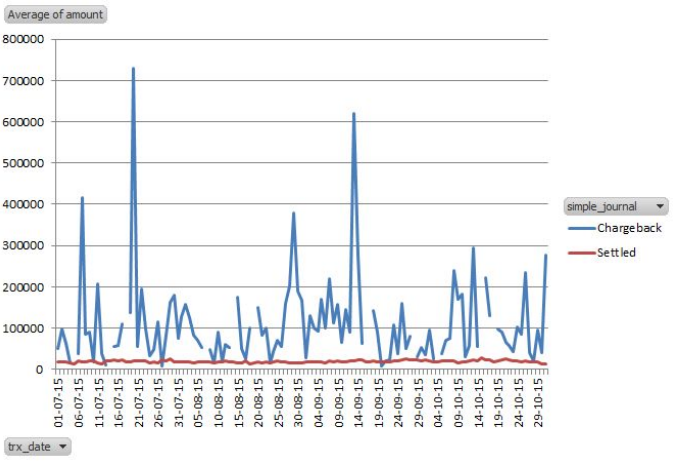

Figure 3. Daily average transaction amount

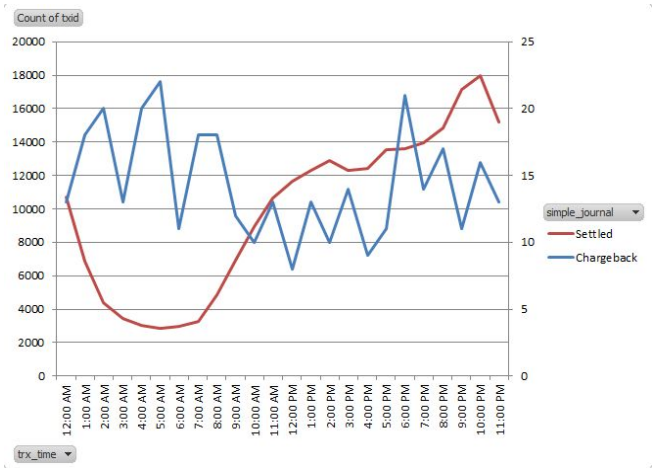

Figure 4. Number of transactions trend of each hour