Cyber Data Analytics
Assignment 1

Ahmet Güdek Tabe Etta

May 13, 2018

Introduction

To be able to safely and trustingly perform online payments, it is important to be able to identify fraudulent transactions. But identifying fraudulent transactions in a sea of valid ones is a difficult task, mainly due to two reasons. First, all transactions look alike with little to distinguish fraudulent transactions from valid ones. And second, the number of fraudulent data points is extremely small in comparison to the entire set of transaction data. This asymmetry in class probabilities makes it diffifult to create robust and reliable classifiers, leading to high false negative rates. In this report we will explore data sampling and manipulation methods to improve the data imbalance and create classifiers building upon this. For this, we will use real transaction data with anonymised data from a bank in Mexico.

Finding patterns

Our dataset consists of the following information:

- bookingdate
- issuercountrycode
- txvariantcode
- bin
- amount
- currencycode
- shoppercountrycode
- shopperinteraction
- simple_journal

- cardverificationcodesupplied
- cvcresponsecode
- creationdate
- accountcode
- mail_id
- ip_id
- card_id

Balancing data

Building classifiers

Bonus assignment

Another way to look at this data is not on a transaction basis, but rather on a credit card or ip address. We created two separate logit classifiers relying on such information. The classifiers require the following information:

- Amount
- Average amount until current transaction
- Total amount in last 24 hours
- Time since last transaction
- The fraud/transaction ratio until current transaction

These values are calculated for every transaction of a user, where a user is defined by the card_id or ip_id columns. The data is viewed as coming in real-time, so it is first sorted by creationdate and converted to epoch timestamps for easy calculation. Next, we traverse over all transactions and calculate the desired values. Finally, we combine the outputs of the two classifiers with the classifier we created earlier by using the voting rule

$$Avg(confidence(logit_1), confidence(logit_2), confidence(logit_3)) > 0.5$$

and obtain the following results.