Anti Money Laundering Project Proposal

Kyle Greenberg (kkg35), Minha Kim (mk932), Christopher Chan (cmc455)

Objective/Research Question: What are strong indicative factors of money laundering?

Background: According to the IMF (International Money Fund), up to 5 percent of global GDP is estimated to be involved in money laundering (~USD 4 trillion). Not only does this include terrorism financing and drug trafficking, it can also involve government corruption and political instability. Being able to correctly identify fraudulent transactions and money laundering could alleviate such pressures on both countries and companies alike. However, money laundering detection is made significantly harder by difficulties in the acquisition of sufficient data that is correctly labeled while also covering a sufficient range of transactions. This proposal is for the analysis of a synthetic dataset generated through a simulation by IBM in order to explore potential methods of aiding money laundering detection.

Data Description: The chosen dataset is a set of 5 synthetically produced sample sets of generated data divided into the files shown in Table 1. Samples within the set consist of 11 observations: Timestamp of Transaction, sending and receiving banks, accounts, monetary values and currency of transaction payment and receipt, payment format, and whether or not it is part of a money laundering attempt. Included alongside the sample sets are 6 text files each containing a set of transactions alongside within the sample sets they correspond to, with labels indicating the money laundering strategy being used for them. Several potentially problematic aspects of the dataset are immediately noticeable: The potential for operation on such a large -scale dataset to be too demanding for our current hardware, and the absence of a *'LI-Small_Trans.csv'* file, which is referred to in both the 6 text files, and the original page detailing the dataset.

Filename	Sample Size
HI-Large_Trans.csv	(n = 179,702,229)
HI-Medium_Trans.csv	(n = 31,898,238)
HI-Small_Trans.csv	(n = 5,078,345)
LI-Large_Trans.csv	(n = 176,066,557)
LI-Medium_Trans.csv	(n = 31,251,483)

Table 1. The filename and sample size of the sets provided by IBM.

How Will This Dataset Help?

This dataset will be helpful for potentially many reasons. If the hypotheses that are formulated through using this synthetic dataset show substantial promise, they could in the future, potentially be tested on real life data. Essentially, using this data to formulate and test hypotheses could potentially make the process of training to detect money laundering easier, since initial issues and modifications that usually arise could be detected preliminarily with the synthetic dataset.