Anti Money Laundering (AML) Transaction Detection Model

Transaction Dataset and Virtual Data Generator created by:

Money laundering is a multi-billion dollar issue. Detection of laundering is very difficult. Most automated algorithms have a high false positive rate: legitimate transactions incorrectly flagged as laundering. The converse is also a major problem -- false negatives, i.e. undetected laundering transactions. Naturally, criminals work hard to cover their tracks.

Access to real financial transaction data is highly restricted -- for both proprietary and privacy reasons. Even when access is possible, it is problematic to provide a correct tag (laundering or legitimate) to each transaction -- as noted above. This *synthetic transaction data* from IBM avoids these problems.

The data provided here is based on a virtual world inhabited by individuals, companies, and banks. Individuals interact with other individuals and companies. Likewise, companies interact with other companies and with individuals. These interactions can take many forms, e.g. purchase of consumer goods and services, purchase orders for industrial supplies, payment of salaries, repayment of loans, and more. These financial transactions are generally conducted via banks, i.e. the payer and receiver both have accounts, with accounts taking multiple forms from checking to credit cards to bitcoin.

Some (small) fraction of the individuals and companies in the generator model engage in criminal behavior -- such as smuggling, illegal gambling, extortion, and more. Criminals obtain funds from these illicit activities, and then try to hide the source of these illicit funds via a series of financial transactions. Such financial transactions to hide illicit funds constitute laundering. Thus, the data available here is labelled and can be used for training and testing AML (Anti Money Laundering) models and for other purposes.

The data generator that created the data here not only models illicit activity, but also tracks funds derived from illicit activity through arbitrarily many transactions -- thus creating the ability to label laundering transactions many steps removed from their illicit source. With this foundation, it is straightforward for the generator to label individual transactions as laundering or legitimate.

Note that this IBM generator models the entire money laundering cycle:

- Placement: Sources like smuggling of illicit funds.
- Layering: Mixing the illicit funds into the financial system.
- Integration: Spending illicit funds.

As another capability possible only with synthetic data, note that a real bank or other institution typically has access to only a portion of the transactions involved in laundering: the transactions involving that bank. Transactions happening at other banks or between other banks are not seen. Thus, models built on real transactions from one institution can have only a limited view of the world.

By contrast these synthetic transactions contain an entire financial ecosystem. Thus it may be possible to create laundering detection models that understand the broad sweep of transactions across institutions, but apply those models to make inferences only about transactions at a particular bank.