Fighting Money Laundering With Statistics and Machine Learning

ABSTRACT

Money laundering is a profound global problem. Nonetheless, there is little scientific literature on statistical and machine learning methods for anti-money laundering. In this paper, we focus on anti-money laundering in banks and provide an introduction and review of the literature. We propose a unifying terminology with two central elements: (i) client risk profiling and (ii) suspicious behavior flagging.We find that client risk profiling is characterized by diagnostics, i.e., efforts to find and explain risk factors. On the other hand, suspicious behavior flagging is characterized by non-disclosed features and hand-crafted risk indices. Finally, we discuss directions for future research. One major challenge is the need for more public data sets. This may potentially be addressed by synthetic data generation. Other possible research directions include semi-supervised and deep learning, interpretability, and fairness of the results.

**EXISTING SYSTEM**

Badal-Valero et al. [37] combine Benford’s Law and four machine learning models. Benford’s Law [38] gives an empirical distribution of leading digits. The authors use it to extract features from financial statements. Specifically, they consider statements from 335 suppliers to a company on trial for money laundering. Of these, 23 suppliers have been investigated and labeled as colluders. All other (non-investigated) suppliers are treated as benevolent. The motivating idea is that any colluders, hiding in the non-investigated group, should be misclassified by the employed models. These include a logistic regression, feedforward neural network, decision tree, and random forest. Random forests [39], in particular, combine multiple decision trees. Every tree uses a random subset of features in every node split. To address class imbalance, i.e., the unequal distribution of labels, the authors investigate weighting and synthetic minority oversampling [40]. The former weighs observations during training, giving higher importance to data from the minority class. The latter balances the data before training, generating synthetic observations of the minority class. According to the authors, synthetic minority oversampling works the best. However, the conclusion is apparently based on simulated evaluation data.

González and Valásquez [41] employ a decision tree, feedforward neural network, and Bayesian network to model Chilean firms using false invoices. Bayesian networks [42], in particular, are probabilistic models that represent variable dependencies via directed acyclic graphs. The authors use data on 582,161 firms, 1,692 of which have been labeled as either fraudulent or non-fraudulent. Features include information about previous audits and taxes paid. Because most firms are unlabeled, the authors first use unsupervised learning to characterize high-risk behavior. To this end, they employ self-organizing maps [43] and neural gas [44]. Both are neural network techniques that build on competitive learning [45] rather than error correction (i.e., gradient-based optimization). While the methods do produce clusters with some behavioral patterns, they do not appear useful for false invoice detection. On the labeled training data, the feedforward neural network achieves the best performance.

Camino et al. [58] flag clients with three outlier detection techniques: an isolation forest, a one-class support vector machine, and a Gaussian mixture model. Isolation forests [59] build multiple decision trees using random feature splits. Observations isolated by comparatively few feature splits (averaged over all trees) are then considered outliers. One-class support vector machines [60] use a kernel function to map data into a reproducing Hilbert space. The method then seeks a maximum margin hyperplane that separates data points from the origin. A small number of

observations are allowed to violate the hyperplane; these are considered outliers. Finally, Gaussian mixture models [61] assume that all observations are generated by a number of Gaussian distributions. Observations in low-density regions are then considered outliers. The authors combine all three techniques into a single ensemble method. The method is tested on a data set from an AML software company. This

contains one million transactions with client-level features recording summary statistics. The authors report positive feedback from the data-supplying company; otherwise, evaluation is limited.

Sun et al. [62] apply extreme value theory [63] to flag outliers in transaction streams. The authors start by engineering two features. The first records the number of times an account has reached a balanced state, i.e., when money transferred into an account is transferred out again. The second records the number of effective fan-ins associated with an account, i.e., when money transferred into the account surpasses a given limit and the account again reaches a balanced state. Next, the Pickands–Balkema–De Haan theorem [64], [65] is invoked to model (derived) conditional feature exceedances according to a generalized Pareto distribution. The approach

allows the authors to flag transactions according to a probabilistic limit *p* (in analogy to the *p*-values used to test null hypotheses). The approach is tested on real bank data with simulated noise and outliers.

**Disadvantages**

* We find that studies on client risk profiling are characterized by diagnostics, i.e., efforts to find and explain risk factors. Specifically, unsupervised methods are used to search for new ‘‘risky’’ observations or risk factors. On the other hand, supervised methods are used with an explanatory focus.
* We also find that studies employing unsupervised methods generally use relatively large data sets. By contrast, studies employing supervised methods use small (labeled) data sets.

Proposed System

In this paper, we focus onAMLin banks and aim to provide a technical review that researchers and industry practitioners (statisticians and machine learning engineers) can use as a guide to the current literature on statistical and machine learning methods for AML in banks. Furthermore, we aim to provide a terminology that can facilitate policy discussions, and to provide guidance on open challenges within the literature. To achieve our aims, we (i) propose a unified terminology for AML in banks, (ii) review selected exemplary methods, and (iii) present recent machine learning concepts that may improve AML.

**Advantages**

* The proposed system reduced an UNSUPERVISED CLIENT RISK PROFILING problem.
* The proposed system eliminates SUPERVISED CLIENT RISK PROFILING problem.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).