Introduction

Up until the big data revolution, the approach to money laundering detection has been to look at evidence provided from audit trails and the wider business context. Such investigative techniques mostly focused on detecting suspicious patterns from the data available to identify laundered activities. The emergence of technology (such as machine learning and algorithms) and the digital economy has changed the manner in which data are mined for further analysis. Money laundering in major industrial centres is becoming more complex and the sheer volume of the data is far beyond what investigators can realistically comprehend and analysed. For example, a recent media report coming out of Canada noted that money laundering activities in Vancouver, British Columbia is estimated to be $1 billion a year and is disrupting the real estate and financial sectors (Meissner, 2019). Data analysts, despite how adept they are, simply cannot comprehend the aggregate data available from money laundering activities using traditional approaches. This is where data mining techniques can be useful. Data mining techniques can allow data analysts to analyze the data and make informed decisions on whether a particular transaction or a series of particular transactions are likely to have been laundered, and hence, needed to be red flagged by anti-money laundering (AML) compliance teams for further investigation.

The traditional evidence-based approach, which focuses on collating data from internal controls, external audits, and whistle blowing hotlines to detect money laundering activities simply does not have the tools to address the complexity of these transactions. A more efficient method is to employ data mining (such as statistical methods, artificial intelligence, pattern
recognition, neural networks etc.) and various analytical approaches to discover knowledge and patterns in the dataset. In this sense, data mining can be seen as the assemblage of large masses of data that can then be analyzed to make informed decisions on a phenomenon, in this case money laundering transactions (Written et al., 2017).

As a corollary in understanding data mining to detect money laundering activities, the pre-processing of the data is of critical importance in making inference from the data. Context is important. To get the most from the data, data analysts need to understand the wider context of the sector in which there are suspicions of money laundering activities (Lokanan, 2018).

Considerable thought must go into the data mining preparation and subsequent processes for inferences to be drawn from the data. While, there are various methods for detecting suspicious money laundering activities, this paper will focus on using statistical techniques to mine and analyze suspicious transactions.

The Data Mining Process

No discussion on data mining to detect money laundering activities is complete without addressing the general data mining process. Handling and processing large volume of transaction data is a very complex phenomenon. The large volume of data from various contexts and sectors requires precision in the data mining process. Due to the diversity of contexts from where the data is collected, the data mining process needs to be a very interactive and iterative process (Sethi, 2001), and is particularly so when collating data on complex financial transactions. Given the challenging and temporal nature of transactional database and the complexity of the data entry involved when coding the transactions (Lokanan, 2019), the mining of the data is pivotal to discover un-hypothesized relationship in the data (Romney and Steinbart, 2015, p. 110). Data mining process involves understanding the domain to collate large volume of unstructured data that allows data analysts to gain insight into the business processes of the institutions through
which laundered funds flow through. Through the discovery from large patterns of data, insights are gained and patterns are discovered. These patterns are then interpreted and disseminated for further analysis. Figure 1 below show the different stages of the data mining process.

**Figure 1.** Stages in the Data Mining Process

*Understanding the Data*

Understanding the data has a lot to do with understanding the domain. Attempting to mine data without an understanding of the business domain can lead to weaknesses in the analytical outputs. To obtain a clear understanding of the business domain, it is first important to identify the business goals and objectives, and then assess the current situation of the business. It is important here to focus on the contextual factors of the business, namely, challenges, resources constraints, and any other contingency issues that can affect the mining of the data. Given the context of the business, this knowledge is then employed to create data mining goals that can be supported by analytic techniques to achieve the business objectives of the company (Kennedy and Moss, 2015). Here the objective is to mine data in a way that will assist data scientists to analyze and make sense of the transactions and in so doing, detect laundered activities.
Data Selection

Understanding the sources of the data is important for data selection. Data analysts deciding to use the data cannot analyze all the data available for suspicious transactions. The analyst is better served by selecting a database and a subset of data records to be used for data mining (Romney and Steinbart, 2015, pp. 110-112). This is where the importance of securing a very good understanding of the sectors and institutions (e.g., banks, securities firms, real estate agencies etc.) helps, because the analysts can then target certain fields and employed their resources to capture useful data for further analysis. At this stage in the data mining process, the properties of the data are carefully examined to ensure goodness of fit. The experienced analysts can do this through careful queries and visualization to ensure completeness of the data (i.e., missing values and outliers).

Data Preparation

Data preparation is the most time consuming stage of the data mining process (Sethi, 2001). This stage involves cleaning and pre-processing the data for further empirical analysis. Cleaning is vital to avoid duplication (otherwise refer to as data corruption). Data corruption is a recurring issue when dealing with money laundering data. For example, to avoid being detected of money laundering transaction, a customer may spell his name incorrectly (Marc instead of Mark) so that the algorithm will flag him as a different person (Cao and Do, 2012). Another issue in data preparation is missing field. Due to the monotonous nature of data entry, the individual coding and entering the data may forget to fill in the desired information. While addressing missing values are a non-trivial task, it can skew the analysis if not detected in the data preparation stage (Kennedy and Moss, 2015; Lokanan, 2019). The presence of outliers in a database is due to the incorrect entries in the data field (Rahman and Islam, 2014; Sethi, 2001).
The presence of outliers in the database is difficult to detect, and more often than not, the analysts will have to undergo a data discovery step to get rid of outliers.

Part of data preparation involves pre-processing the data for meaningful analysis. Data pre-processing involves the integration of raw unstructured data from different sources into an understandable format (García et al., 2105). Raw data are incomplete and contains many errors. Because raw data is unstructured, they are very noisy and lack format. Pre-processing the raw data is a proven methodological technique to address such issues and is necessary because certain field may contain data that may not be applicable to the substantive issues at the data cleaning stage (Rahman and Islam, 2014). For example, analysts examining banks’ data for patterns in suspicions transactions may become quickly overwhelmed if they have to investigate each and every transaction. Such voluminous and unstructured data may not generate useful trends and patterns. Grouping the transactions in clusters (of $10,000, $15,000, and $20,000) may generate and prove to be more useful when conducting further analysis. The clusters built at this stage are important to detect patterns and trends in the data discovery stage.

Data Discovery

In the data discovery stage, the analyst is mostly concerned with discovering patterns and trends in the data. The data discovery stage consists of several approaches, namely association, classification, clustering, sequence analysis, and visualization (Sethi, 2001). Each one of these approaches can be use to mine data for further analysis using statistical techniques, neural networks, and machine learning. The advent of R programming and Python have made it easier for experts to analyze data mined from any one of these approaches. Figure 2 present a diagrammatic illustration of the various methodologies use to mine data.
Figure 2. Data mining methodologies

The association approach seeks to establish relationship between items in large database (Kohavi, 2000). At the very basic level, the associate approach can employ machine learning to look for patterns in a database. Algorithms can also be use to scan the database and detect association between items in the data. The association approach is very useful in detecting suspicious transactions. For example, the association approach can be use to detect the time and dates of when individuals of a certain demographic group visits banks to make their transactions. This approach to data mining supports algorithmic and machine learning techniques because they can sort out association between items from different subfields.

Data classification involves the categorization of data into classes for further analysis. Classification is particularly useful when there are multiple terabytes of data. This poses serious challenges for data analysts to analyze the data and make sense of them. To address these issues, classification rules are usually put in place to allow for optimal use of the data. For example, data can be classified to mean highly sensitive or moderately sensitive, especially when dealing with money laundering transactions. Data analysts working with sensitive data may want to ensure that they are properly secured to avoid the data from being infiltrated and compromised.
Sensitive data may be used later as evidence in criminal proceedings and cannot be tempered with. This type of classification can generate training dataset with predictor rows and columns with attributes that classification algorithms can then use to detect the types of relationship that exists (see Lokanan, 2019). Table 1 provides a diagrammatic illustration of how classification can be done to facilitate decision trees and algorithmic classifiers. Note that Table 1 is structured so that the algorithm will be given a set of data points to make a prediction on suspicious money laundering transactions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Gender</th>
<th>Country</th>
<th>Suspicious Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Doe</td>
<td>45</td>
<td>Male</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>Jane Doe</td>
<td>39</td>
<td>Female</td>
<td>B</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1. Dataset for predictive analysis

Data clustering is done so that similar items in the dataset are grouped into identical clusters for further analysis (Salehi et al., 2017; Torgo 2016). This approach to data mining is use in situations where training data sets of pre-classified data are not available (Sethi, 2001). Splitting the data into similar parts or clusters can make data mining more practical because the analysts can simply select one of the clusters that they are interested in to perform further analysis (De Veaux et al., 2008, p. 295). More importantly, is that if the data has been coded accurately and each cluster fairly represents the entire dataset, then the cluster approach may provide for useful analysis (p. 295). However, if the clusters are different, then it can prove very challenging to conduct useful analysis. The example in Figure 4 below shows the challenge that data analysts faced when dealing with different clusters. A closer look at Figure 3 shows that there are different types of clusters, each with its own similarities. The challenge for statistical techniques is to come up with models that will address and enhance their predictive capabilities of each clusters (see Lokanan, 2019).
Figure 3. Diagrammatic illustration of clustering in data mining

Sequence analysis is useful when conducting time series analysis (De Veaus et al., 2008, pp. 90-91). The sequence approach is particularly useful to detect patterns and trends in time series data (Dong, 2009). For example, institutions can use auto-regressive based outlier algorithm to detect outliers in large voluminous transactions (Somasundaram, 2017). Sequence analysis can also be used to detect suspicious transactions overtime and can be employed by data analysts to detect anomalies in the data. Anti-money laundering compliance agencies can also use sequencing to detect suspicious transaction by employing machine learning and neural network methods (Salehi et al., 2017).

The visualization approach to data mining has its origins in the behavioural sciences. It is base on the premise that human beings have the cognitive ability to perceive the structural form that the data will take (Kohavi, 2000; Umadevi and Divya, 2012). Data visualization is more descriptive in nature. It provides the data analysts with a set of visualization tools - statistical graphs, plots, charts to display and convey information about the data. Data scientists employ
visualization to look for complexities and outliers in the data. Tableau, Phyton, and R programming has proven to be very useful tools to assist data scientists in the visualization process in data mining. The key in using the visualization techniques to comb the dataset is to discover hidden patterns and relations in the data.

*Evaluation and Interpretation of the Data*

The evaluation and interpretation stage of the data mining process is where data analysts evaluate the quality of the data and decide whether previous stages should be revisited to ensure the validity and reliability of the data (Sethi, 2001). A thorough understanding of the domain and its context is very important at this stage to ensure that all subsequent analysis of the data will meet the business objectives. Very important also is that the data may show patterns not aligned with business objectives. It is important for data analysts to take this into consideration and identify new patterns and business objectives that they may want to relay back to the AML compliance team. Data evaluation must go through an iterative process as new objectives are identified and disseminated to the stakeholders that will be using the results of the data.

*Reporting the Data*

The knowledge of the data gathered through the data mining process needs to be presented in such a way that stakeholders who rely on the data to make material decision can benefit from them. This means that the data must be conveyed in the simplest form as possible as many of the users may not have the expertise needed to interpret the trends and patterns identified from the data. This action is important as the knowledge obtain will be used to run models and informed internal control policies designed to mitigate money laundering transactions.
Statistical Methodological and Money Laundering Detection

The statistical method of data mining and analysis has a solid theoretical foundation and came from a very mature field of quantitative science (De Veaux et al., 2008; Yasaka, 2017; Lokanan, 2018). Methodologies from the statistical tradition have been thoroughly tested over time and as such, the knowledge discovered are more likely to have external validity and generalizability than some of the more recent machine learning methods (Lokanan, 2019; Sethi, 2001). There are two types of statistical methods used to mine and subsequently analyze data: the dependence and interdependence methods. The dependence methods employ one or more independent variable to predict the effects on the dependent variable (e.g. see De Veaux et al., 2008; Lokanan, 2019; Sethi, 2001). Example of this approach includes multiple regression, logistic regression, and discriminant analysis. The interdependence methods are useful when all of the variables are independent variables (Sethi, 2001). Examples of the interdependence methods include clustering and multidimensional scaling (Sethi, 2001).

Dependence Methods

Multiple Regressions

The multiple regression method is employed where there are more than one independent (or predictor) variable and one dependent variable (De Veaux et al., 2001, Ch. 30). The objective of the multiple regression models is to find the best fit for the predictor variables with the dependent variable. A multiple regression model is represented with the following equation:

\[ Y = a + b_1X_1 + b_2X_2 + b_3X_3 \]

Where:

\( Y \) is the Dependent variable
\( A \) is the Constant or \( Y \)-intercept
\( b_1, b_2, b_3 \) are the Slopes of \( X_1, X_2, \) and \( X_3 \)
The major task as it applies to the detection of laundered activities lies in identifying the independent variables and the predictor variable. This decision is based on the outcome that needs to be predicted and the activities that affect this outcome in the context of AML compliance. A stepwise regression is usually employed to select the independent variables that statically contribute to the model to predict the dependent outcome.

**Logistic Regression**

In many money laundering cases, there may be customers who are suspected of laundering funds and customers who processed legitimate transactions. In these types of cases, a logistic regression can be used because the dependent variable can be coded to be 1 for suspicious transaction, and 0 for non-suspicious (legitimate) transactions (Yasaka, 2017). Rather than predicting the outcome (dependent) variable, logistic regression estimates the probability ($p$) that the outcome variable will be given the context and situation (De Veaux et al., 2001, pp. 690-692). The actual probability of the dependent variable determines the stage of the given situation. So for example, if the probability is greater than 0.50 then the prediction is that suspicious transactions are taking place and if it is less than 0.50, then there is very little suspicious transaction. The logistic regression equation is represented as follows:

$$Y_1 = B_0 + B_1 X_1 + E_1$$

Where:

- $Y_1$ = Dependent variable
- $B_0$ = Population Y-intercept
- $B_1$ = Population slope coefficient
- $X_1$ = Independent variable
- $E_1$ = Random error term
**Linear Discriminant Analysis**

Linear Discriminant Analysis (LDA) is a commonly used technique for classification problems in data mining (Hoque et al., 2015, p. 1). In such problem, the dependent variable is categorical, while the predictor variables are in metric form. The basic idea is to present the dataset into different classes to avoid overfitting the data and to estimate the probability of the phenomenon under investigation (Raschka, 2015). The basic LDA is represented in the following form:

\[ Z = W_1X_1 + W_2X_2 + \ldots + W_kX_k \]

Where \( X_1, X_2, \ldots, X_k \) are the independent variables, \( Z \) is the discriminant score, and \( W_1, W_2, \) and \( W_3 \) are the weights. Once the model is operationalized, the discriminant score \( (Z) \) is used to predict the relationship between the dependent variable and the independent variables (Mahmoudi & Duman, 2015). In other words, the data analysts will assign the \( N^{th} \) data record for class A (i.e., legitimate transaction) when the discriminant score is greater than or equal to the expected score otherwise it is assigned to class B (suspicious transaction).

**Interdependence Method**

**Clustering**

Clustering in data mining is an approach where each object in a particular group is similar to objects in its own group than the other groups (De Veaux et al., 2008). For example, the data can be mined in clusters of predefined classes of suspicious transactions and no suspicious transaction (i.e., legitimate transaction). Classification in clusters can be the basic premise of the clustering approach in detecting suspicious transaction. The true test of this approach is how it measures similarity and the kind of method that was employed to perform the grouping (Sethi, 2001). The most commonly used approach to measure similarity is the Euclidean
distance. Given two phenomena, (a) suspicious transaction and (b) legitimate transaction, the Euclidean distance between them is defined as the length of the line joining them. There are two types of clustering methods: (1) hierarchical, and (2) partitional.

Hierarchical Clustering

Hierarchical clustering is represented with a sequence of clusters that are represented in a tree like structure. There are two types of hierarchical clustering: (1) Agglomerative (bottom up) and (2) Divisive (top down). Agglomerative is very conducive to machine learning and links each data point measured by a distance to the nearest similar points, thereby creating a cluster. This is an iterative process and will continue until all the data points are grouped into their respective clusters. Figure 4 below shows an algorithmic illustration of the treelike agglomerative cluster. Note how the cluster can still be linked by the algorithmic methodology. On the other hand, divisive cluster starts off by including all the objectives in a single large cluster. The heterogeneous or dissimilar clusters are then further divided into clusters where all the data points are similar. After a series of iterative process, the data points then form their own clusters based on similar traits.
Hierarchical Clustering

**Figure 4**: Hierarchical tree like cluster

*Partitional Clustering*

Partitional clustering divides the data objects into subsets. Figure 5 presents a diagrammic illustration of partitional clustering. Note from Figure 5 that the data objects are grouped into their respective cluster only after one iteration. There are two types of partitional clusters: (1) sequential, and (2) simultaneous. Sequential clustering uses algorithm to separate the data points in particular order (Le et al., 2014). The order of the objects then influenced the cluster that they will be partitioned into. For example, all transactions that are for $10,000 will be partitioned into one cluster for further analysis. Simultaneous clustering of the data partitioned all the observations in one go for further analysis (Charrad and Ahmed, 2011; Li et al., 2018). Algorithms are employed to seek blocks of rows and columns that are interrelated to identify bi-clusters with similar behaviours (Li et al., 2018, p. 371). Simultaneous clustering is well suited for high level data analysis of suspected money laundering transaction as it is “more informative, has less parameters, is scalable and is able to effectively intertwine row and column information” (p. 371).
Multidimensional Scaling

Multidimensional Scaling (MDS) relies on projection to detect similarities from a far distance to a close distance between the data points (Liu et al., 2015). With MDS, the data analyst will obtain a two dimensional map. For this mapping to be useful, it is required that the mapping preserve the observed distance as far as possible. Again using the example between suspicious laundered transaction and legitimate transaction, an MDS analysis will allow data scientists to observe how divergent the distances are between these two phenomena from each other. Sometimes, the results from an MDS analysis between two points are easy to interpret; but, at other times, the distance becomes very nebulous, which makes it very difficult to interpret the dimensions. In such cases, it may be best to include fewer dimensions and make the configuration easier to analyzed and interpret. That said, the beauty of MDS is that it can analyze unobtrusive questions, with great accuracy. For example, data analysts can use MDS to...
investigate the similarity between transactions without the perpetrator’s knowledge of the analyst’s real intent in such evaluations.

**Conclusion**

Having described the data mining process and the statistical approach to detect money laundering activities, one may query whether it is a better approach than machine learning models. This is not an easy question to answer. With Python and R programming, machine learning is now universally used with great success. However, the statistical approach to fraud detection, in general, is a proven technique and as illustrated earlier, does have potential to be use by data analysts to detect money laundering transactions. What is known for certain is that both the statistical approach and the machine learning approaches have their own unique characteristics and may be better employed depending on the context. That said, the statistical methods have been tested across multiple dimensions and contexts and have proven to be reliable overtime. The statistical methods have also been used by both the academic community and practitioners to more generally, detect fraud in various contexts. Consequently, there are benchmarks that can be use to verify the reliability of the statistical approach to data mining, and as a consequence, the knowledge discovered using the statistical methods have external validity and have been accepted by the data mining community to be reliable.
References


