JMIBR
Journal of Modern Issues in Business Research

OPEN ACCESS

2021, Vol. 1, No. 1

Applying Supervised Machine Learning Algorithms for Fraud Detection in Anti-Money Laundering

Omri Raiter

Abstract

As international money transfers become more automated, it becomes easier for criminals to transfer money across borders in a fraction of a second, while it also becomes easier for regulators to inspect and monitor international money mobility and identify unusual patterns of money movement. Machine learning algorithms may be a useful addition to the current money laundering detection issues. This research empirically tested four machine learning algorithms (Logistic regression, SVM, Random Forest, and ANN) using a synthetic dataset that closely matches regular transaction behavior. After observing the performance of different algorithms, it can be stated that the Random Forest technique, when compared to the other techniques, provides the best accuracy. The least accurate approach was the Artificial Neural Network (ANN).

Keywords: AML, ANN, Logistic Regression, Random Forest, SVM

Corresponding author: Contactme@OmriRaiter.com

Introduction

Money laundering dates back to the early twentieth century which is regarded a gangster era in American history (Strafer, 1989). Gambling, prostitution, and alcohol sales were all increasing at the time, and people were earning a huge amount of cash that required to be laundered so the authorities would not know where their money came from (Unger and Van der Linde, 2013). As a result, a means for concealing the sources of gangsters' finances had to be devised, as they could not just deposit the monies in the bank (Muller, 2007) . The fact that this money could not simply

be put into a bank would generate a slew of concerns, since the bank, and eventually the government, would want to know where the money came from, and the gangster would be unable to offer a plausible explanation. Furthermore, the monies could not just be spent on a variety of high-end things, as this would create suspicion (Sullivan, 2015). Furthermore, the monies were frequently in little dollar notes and low-value coins, adding to the gangsters' difficulties in holding enormous sums of money.

Money laundering may be broken into three parts, which may be referred to as placement, layering, and integration. Not all money laundering transactions, however, go through all three rounds. Even still, classifying distinct phases in what might be a difficult process is important. The placement phase is the initial step in the procedure. The money or other monies produced from unlawful acts are physically relocated to a site or into a less suspect shape to law enforcement authorities and extra favorable to the criminal throughout the placement phase, according to the Board of Governors of the Federal Reserve System (Bosworth-Davies, 2007). The proceeds are subsequently invested in (non-traditional) financial institutions or the retail economy. This insertion can be accomplished by transporting money to another nation, enlisting the help of security brokers or bank employees, purchasing assets with cash, and a variety of other means (Saeed, Mubarik and Zulfiqar, 2021) (Teichmann, 2017) (Raiter, 2021). Layering is the second phase, which entails intricate financial transactions in order to establish sophisticated channels that obfuscate and mask the money from its illicit source, making it harder for law enforcement authorities to track down where the money came from. Circulating money through several financial accounts, transferring money across many distinct businesses, or reselling assets purchased during the placement phase - locally or internationally - are all examples of layering strategies (Schroeder, 2001) (Tong, 2021). Integration is the third and final phase (Sharman, 2008). Normal financial procedures are used to turn the unlawfully acquired funds into ostensibly legitimate business money (Compin, 2008). The money seems legitimate if the documentation for the company's revenues or spending is clever. Dealing with properties and making fraudulent loans are also used in this operation.

Machine learning technologies can be a helpful contribution to the existing challenges of money laundering detection (Guevara, Garcia-Bedoya and Granados, 2020) (Labib, Rizka and Shokry, 2020) (Paula *et al.*, 2016). Fresh approaches can provide new insights, show previously unknown

trends, and help detect questionable transactions more precisely (Villalobos and Silva, 2017). Machine learning technologies, on the other hand, have limitations that may severely impact the identification of financial activities. Due to the intricate structures and patterns of money laundering activities, detecting them remains a difficult task. Because machine learning methods have prospective benefits and limitations in the detection of money laundering, this study examines and identifies these strengths and drawbacks, as well as whether and how machine learning can accompany the rule-based method of analyzing financial network typologies (Canhoto, 2021).

Once the detection mechanism - or a component of the detection process - has been refined, upgraded, or replaced, machine learning approaches may be effective. Machine learning approaches might partly replace staff screening and analysis duties once machine learning models perform as well as humans (Alkhalili, Qutqut and Almasalha, 2021) .

Methodology

a) Machine learning techniques

i) Support Vector Machine (SVM)

SVM classification's main goal is to successfully divide classes based on a set of criteria. We can illustrate this with a simple two-class problem, as shown in Figure 1. To the left is the simplest scenario, in which the classes are linearly separable. The SVM algorithm distinguishes classes by maximizing the distance between their nearest components (Deris, Zain and Sallehuddin, 2011) (Ding, Qi and Tan, 2011). Support vectors are the closest members of each class on either side of the line.

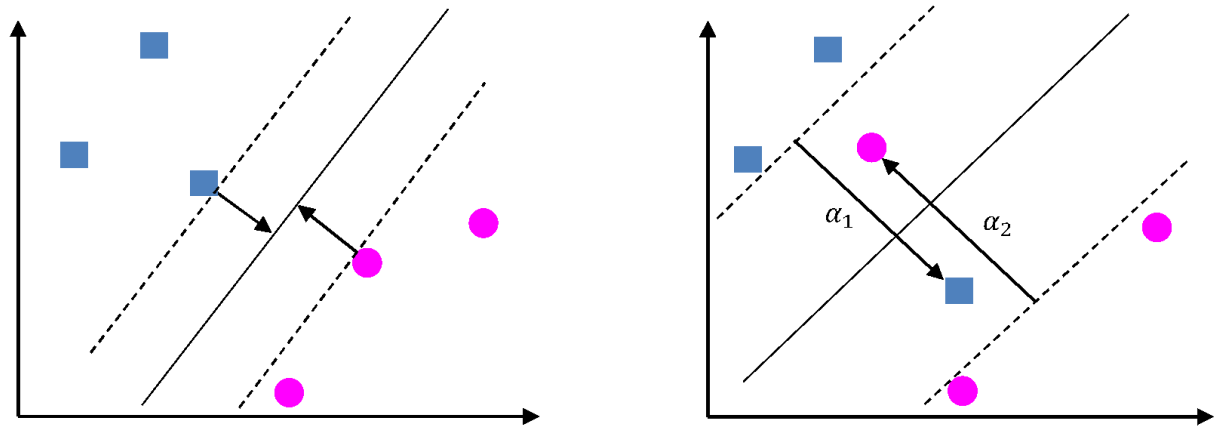


Figure 1 Simple 2-class support vector machine

ii) Logistic Regression

Logistic regression is a supervised classification technique that calculates the likelihood of a binary dependent variable being predicted from the dataset's independent variable (Healy, 2006). Logistic regression is similar to linear regression in that it produces a straight line, while linear regression produces a curve. Based on the usage of one or more predictors or independent variables, logistic regression generates logistic curves that depict the values between zero and one (Hosmer Jr, Lemeshow and Sturdivant, 2013) .

There are many different forms of logistic regression models, including binary, multiple, and binomial logistic models (Menard, 2002). The binary logistic regression model is used to predict the likelihood of a binary response based on one or more factors.

$$p = \frac{e^{\alpha+\beta_n X}}{1 + e^{\alpha+\beta_n X}}$$

Above equation represents the logistic regression in mathematical form.

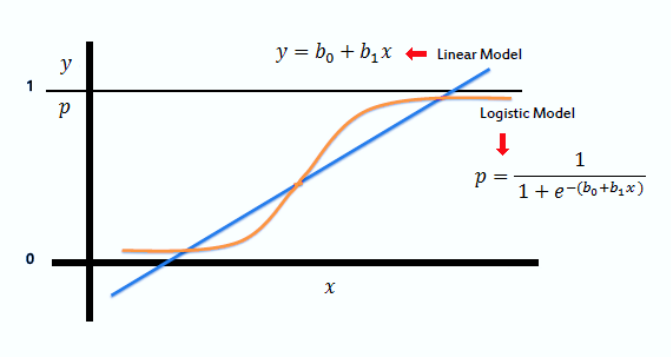


Figure 2 : Logistic Curve

The contrast between linear regression and logistic regression is seen in this graph, where logistic regression depicts a curve and linear regression depicts a straight line (Wright, 1995).

iii) Random Forest

Random Forest is a classification and regression technique. In other words, it's a set of decision tree classifiers. The benefit of random forest over choice trees is that it corrects the behavior of overfitting to the training set (Schonlau and Zou, 2020). A random portion of the training set is sampled to train each individual tree, and then a decision tree is constructed, with each node splitting on a feature chosen at random from the whole feature set (Shaik and Srinivasan, 2019) (Couronné, Probst and Boulesteix, 2018). Random forest training is incredibly quick, even for big data sets with numerous characteristics and data instances, since each tree is trained autonomously of the others. It has been discovered that the Random Forest approach delivers a decent approximation of the generalization error and is resistant to overfitting (Biau and Scornet, 2016). Random Forest may be used to rank the relevance of variables in a regression or classification issue in a natural fashion.

1. The Artificial Neural Networks

One of the most important technologies in machine learning is artificial neural networks. They are brain-inspired systems that are designed to mimic how people learn, as the neural segment of their name suggests (Micheli-Tzanakou, 2011). Neural networks are made up of input and output layers, as well as (in most instances) a hidden layer that consists of units that convert the input into something that the output layer can use (Rosa et al., 2020). They're great for finding patterns that

are much too intricate or numerous for a human programmer to cite and teach the machine to recognize.

Although neural networks (also known as perceptrons) have been present since the 1940s, they have only recently become a viable artificial intelligence application (Zou, Han and So, 2008). This is due to the backpropagation method, which enables networks to alter their hidden layers of neurons in instances when the output does not meet the creator's expectations. The development of deep learning neural networks, in which various layers of a multilayer network extract different characteristics until it can identify what it is seeking for, has also become a key challenge in ANN (Buscema, 2002) (Rosa et al., 2020).

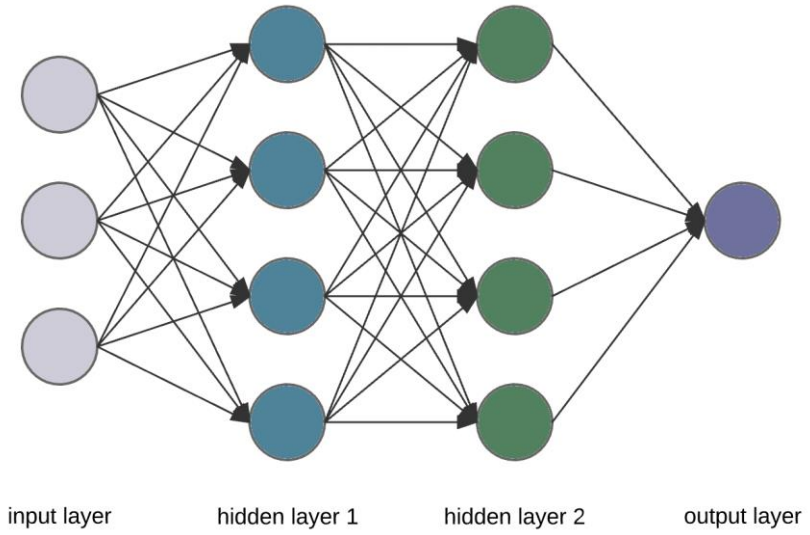


Figure 3: Typical ANN architecture

b) performance of algorithms:

Several criteria, including as F1-score, accuracy, precision, and sensitivity have been used to evaluate algorithm performance.

c) data

There are few publicly accessible statistics on financial services, particularly in the rapidly growing field of mobile money transfers. Many academics, and especially those of us working in the field of fraud detection, rely on financial information. The fundamentally private character of financial transactions contributes to the lack of publicly accessible statistics. As a solution to this

challenge, a synthetic dataset developed using the PaySim simulator was proposed in the literature (Lopez-Rojas and Barneaud, 2019). PaySim generates a synthetic dataset using aggregated data from the private data that mimics the regular functioning of transactions and injects harmful behavior to test the effectiveness of fraud detection algorithms later on (Lopez-Rojas, Elmir and Axelsson, 2016).

Features are given as follows (Patil, Framewala and Kazi, 2020):

CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER are the different types of transactions.

amount - the transaction's value in local currency.

nameOrigin is the name of the client who initiated the transaction.

oldbalanceOrg - before to the transaction, the original balance

newbalanceOrig - following the transaction, the new balance

nameDest - the customer who is the transaction's beneficiary.

oldbalanceDest - the receiver of the original balance before the transaction. It's worth noting that there's no information for clients whose names begin with the letter M. (Merchants).

newbalanceDest - following the transaction, the new balance recipient. It's worth noting that there's no information for clients whose names begin with the letter M. (Merchants).

isFraud - This represents the fraudulent agents' transactions inside the simulation. The fraudulent activity of the agents in this dataset tries to profit by seizing control of clients' accounts and attempting to empty the money by transferring to another account and then cashing out of the system.

isFlaggedFraud - The business model strives to regulate large transfers from one account to another and flags any efforts that seem to be fraudulent. An attempt to transfer more than 200.000 in a single transaction is considered unlawful in this dataset.

Results

We begin results reporting with exploratory analysis. Figure 4 displays the transaction types. The transaction type TRANSFER appears to have the highest average Transaction Amount value, followed by CASH IN for the highest average Old Origination Account Balance value, TRANSFER for the highest average Old Destination Account Balance value, and TRANSFER for the highest average New Destination Account Balance value. Also, it seems that none of the averages for the transaction type PAYMENT were captured by the box plots. Furthermore, these box plots suggest a large number of outliers in the data.

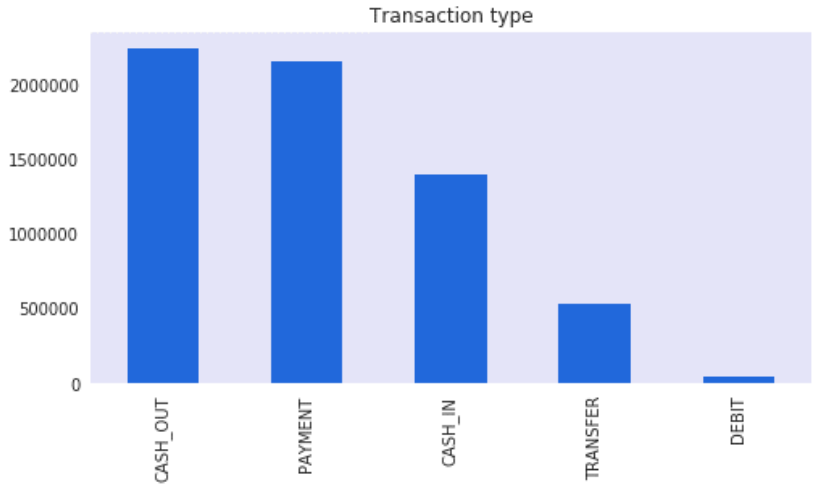


Figure 4. Transaction type

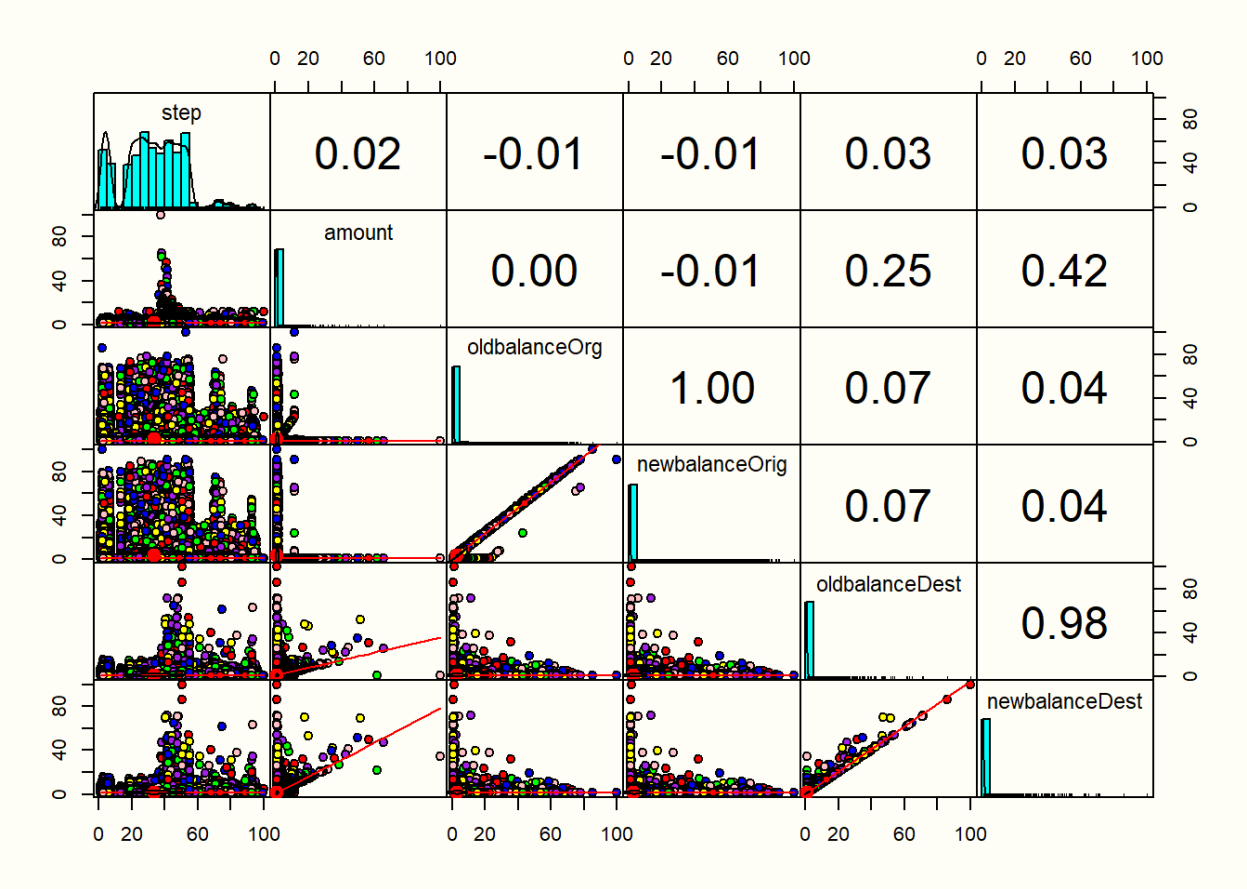
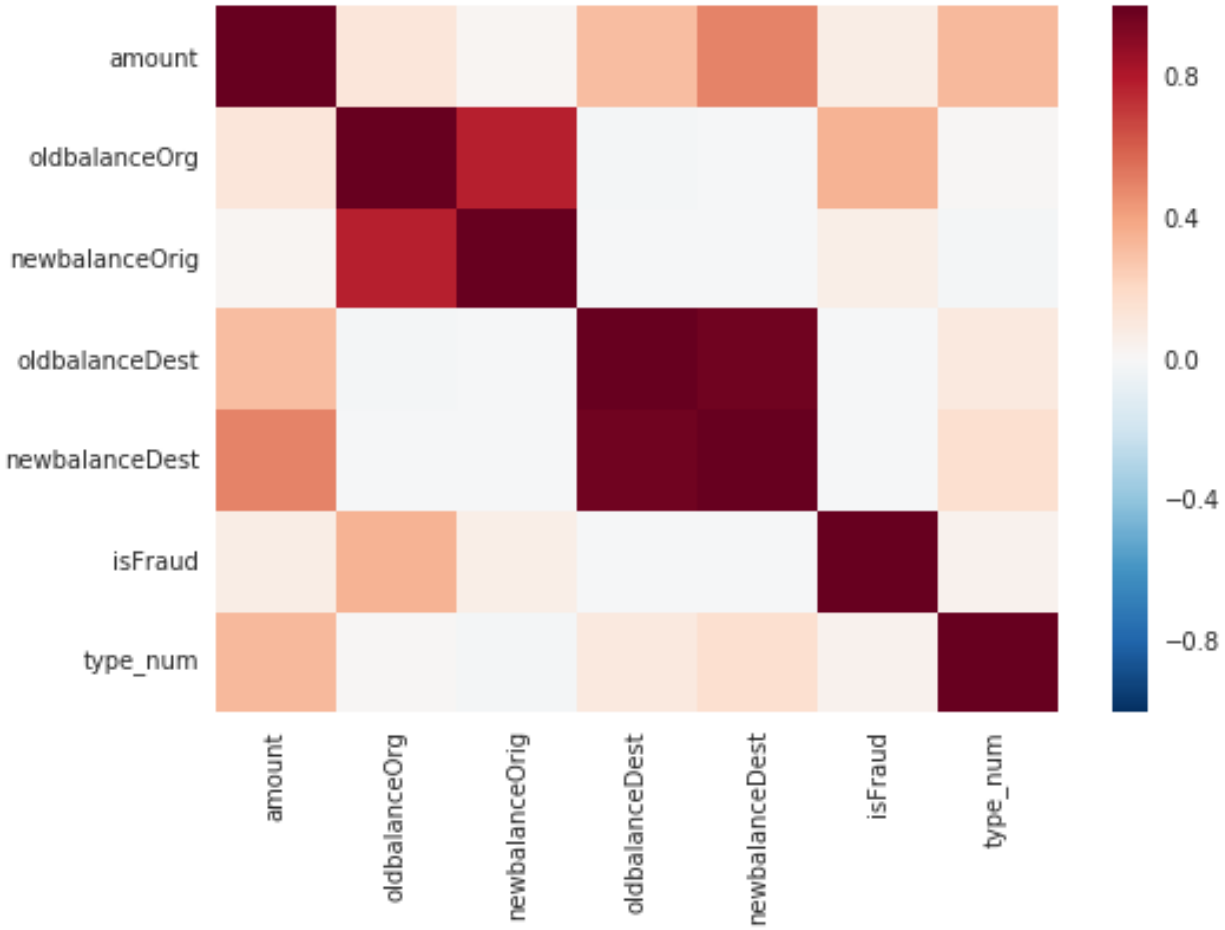


Figure 5. Scatterplots among the features.

Figure 5 and 6 shows the scatterplots and correlation heatmap. There seem to be two pairs of variables with correlation coefficients of larger than 0.5 in both directions. NewbalanceOrig and oldbalanceOrg, as well as newbalanceDest and oldbalanceDest, are the two variables. It's also supported by their p-values of less than 0.05, which, given the significant cut-off point of 0.05, imply the presence of strong internal correlations.



	Techniques	**F1-score**	**Accuracy**	**Sensitivity**	**Precision**
1	Logistics regression	0.02	0.97	0.007	0.81
2	Support Vector Machine	0.011	0.96	0.09	0.79
3	Random forest	0.36	0.99	0.21	0.91
4	Artificial Neural Network	0.0032	0.84	0.0016	0.70

Table 1. Performance of the different models.

The performance of four different models are reported in the table 1. With a score of 0.99, the random forest (RF) is the provided the highest accuracy score. ANN, on the other hand, delivers the least accurate results.

Conclusion

The growing manufacture, trade, and eventual consumption of narcotics has resulted in an ever-increasing flow of illegal money. The current algorithm scrutinizes the incoming transaction data in real-time to determine if the transaction is fraudulent or authentic. The flagged transactions, some of which are illegal, are logged and referred for additional investigation in order to verify ML elements. In general, AML activities are categorized into two categories. The suspected ML operations are first recognized by continuous monitoring, and then different steps to halt or intercept these ML instances are applied. To summarize this research, machine learning models might help with money laundering detection by effectively identifying money laundering transactions from routine transactions. Furthermore, our research findings suggest that Random Forest might help in real-world anti-money laundering detecting settings. Because of its accuracy and interpretability, it may be utilized in Anti-Money Laundering (AML) architecture in financial institutions.

References

Alkhalili, M., Qutqut, M. H. and Almasalha, F. (2021) 'Investigation of Applying Machine Learning for Watch-List Filtering in Anti-Money Laundering', *IEEE Access*, 9, pp. 18481–18496.

Biau, G. and Scornet, E. (2016) 'A random forest guided tour', *Test*, 25(2), pp. 197–227.

Bosworth-Davies, R. (2007) 'Money laundering–chapter three', *Journal of Money Laundering Control*.

Buscema, M. (2002) 'A brief overview and introduction to artificial neural networks', *Substance use & misuse*, 37(8–10), pp. 1093–1148.

Canhoto, A. I. (2021) 'Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective', *Journal of business research*, 131, pp. 441–452.

Compin, F. (2008) 'The role of accounting in money laundering and money dirtying', *Critical Perspectives on Accounting*, 19(5), pp. 591–602.

Couronné, R., Probst, P. and Boulesteix, A.-L. (2018) 'Random forest versus logistic regression: a large-scale benchmark experiment', *BMC bioinformatics*, 19(1), pp. 1–14.

Deris, A. M., Zain, A. M. and Sallehuddin, R. (2011) 'Overview of support vector machine in modeling machining performances', *Procedia Engineering*, 24, pp. 308–312.

Ding, S. F., Qi, B. J. and Tan, H. Y. (2011) 'An overview on theory and algorithm of support vector

machines', *Journal of University of Electronic Science and Technology of China*, 40(1), pp. 2–10.

Guevara, J., Garcia-Bedoya, O. and Granados, O. (2020) 'Machine Learning Methodologies Against Money Laundering in Non-Banking Correspondents', in *International Conference on Applied Informatics*. Springer, pp. 72–88.

Healy, L. M. (2006) 'Logistic regression: An overview', *Eastern Michighan College of Technology*.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013) *Applied logistic regression*. John Wiley & Sons.

Labib, N. M., Rizka, M. A. and Shokry, A. E. M. (2020) 'Survey of machine learning approaches of anti-money laundering techniques to counter terrorism finance', in *Internet of Things—Applications and Future*. Springer, Singapore, pp. 73–87.

Lopez-Rojas, E. A. and Barneaud, C. (2019) 'Advantages of the PaySim Simulator for Improving Financial Fraud Controls', in *Intelligent Computing-Proceedings of the Computing Conference*. Springer, pp. 727–736.

Lopez-Rojas, E., Elmir, A. and Axelsson, S. (2016) 'PaySim: A financial mobile money simulator for fraud detection', in *28th European Modeling and Simulation Symposium, EMSS, Larnaca*. Dime University of Genoa, pp. 249–255.

Menard, S. (2002) *Applied logistic regression analysis*. Sage.

Micheli-Tzanakou, E. (2011) 'Artificial neural networks: an overview', *Network: Computation in Neural Systems*, 22(1–4), pp. 208–230.

Muller, W. H. (2007) 'Anti-money laundering–a short history', *Anti-Money Laundering: International Law and Practice*, p. 1.

Patil, A., Framewala, A. and Kazi, F. (2020) 'Explainability of smote based oversampling for imbalanced dataset problems', in *2020 3rd International Conference on Information and Computer Technologies (ICICT)*. IEEE, pp. 41–45.

Paula, E. L. *et al.* (2016) 'Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering', in *2016 15th ieee international conference on machine learning and applications (icmla)*. IEEE, pp. 954–960.

Raiter, O. (2021) 'Macro-Economic and Bank-Specific Determinants of Credit Risk in Commercial Banks', *Empirical Quests for Management Essences*, 1(1 SE-Research Articles), pp. 36–50. Available at: https://researchberg.com/index.php/eqme/article/view/28.

Rosa, J. P. S. *et al.* (2020) 'Overview of artificial neural networks', in *Using Artificial Neural Networks for Analog Integrated Circuit Design Automation*. Springer, pp. 21–44.

Saeed, S., Mubarik, F. and Zulfiqar, S. (2021) 'Money Laundering: A Thought-Provoking Crime', in *Money Laundering and Terrorism Financing in Global Financial Systems*. IGI Global, pp. 1–29.

Schonlau, M. and Zou, R. Y. (2020) 'The random forest algorithm for statistical learning', *The Stata Journal*, 20(1), pp. 3–29.

Schroeder, W. R. (2001) 'Money laundering: A global threat and the international community's response', *FBI L. Enforcement Bull.*, 70, p. 1.

Shaik, A. B. and Srinivasan, S. (2019) 'A brief survey on random forest ensembles in classification model', in *International Conference on Innovative Computing and Communications*. Springer, pp. 253–260.

Sharman, J. C. (2008) 'Power and discourse in policy diffusion: Anti-money laundering in developing states', *International Studies Quarterly*, 52(3), pp. 635–656.

Strafer, G. R. (1989) 'Money laundering: The crime of the'90s', *Am. Crim. L. Rev.*, 27, p. 149.

Sullivan, K. (2015) 'What is money laundering?', in *Anti–Money Laundering in a Nutshell*. Springer, pp. 1–13.

Teichmann, F. M. J. (2017) 'Twelve methods of money laundering', *Journal of money laundering control*.

Tong, A. (2021) 'Comparison of the fin-tech evergreen fund in China and USA', *Available at SSRN 3904647*.

Unger, B. and Van der Linde, D. (2013) *Research handbook on money laundering*. Edward Elgar Publishing.

Villalobos, M. A. and Silva, E. (2017) 'A statistical and machine learning model to detect money laundering: an application', *Actuarial Sci. Dept. Anahuac Univ., Tech. Rep*.

Wright, R. E. (1995) 'Logistic regression.'

Zou, J., Han, Y. and So, S.-S. (2008) 'Overview of artificial neural networks', *Artificial Neural Networks*, pp. 14–22.