iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36323

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Implementation of XGBoost Ensemble Learning Model for Detecting Money Laundering

Abarna Ramprakash[1], Dr. K. Anusudha[2]
[1]*Dept. Of Electronics and Communication Engineering, Pondicherry University, Pondicherry, India*
[2]*Dept. Of Electronics and Communication Engineering Pondicherry University*

Abstract: Money laundering is the illegal process of concealing the origins of money obtained illegally by passing it through a complex sequence of banking transfers. Currently banks use rule based systems to identify the suspicious transactions which could be used for money laundering. However these systems generate a large number of false positives which leads the banks to spend a huge amount of money and time in investigating the false positives. Hence, in this paper, the monitoring of transactions is to be done using XGBoost machine learning algorithm in order to reduce the number of false positives and to increase the probability of identifying true positives.
Keywords: Money laundering, XGBoost, Machine Learning, False positives.

I. INTRODUCTION

Money laundering is the process of changing large amounts of money obtained from crimes, such as drug trafficking, into origination from a legitimate source. This is recognized as a crime by law makers around the world. Therefore, financial institutions are subject to regulations by government and regulators such as the US Office of the Comptroller of the Currency (OCC), according to which they are liable to investigate any kind of money laundering activity committed by their customers. For this purpose, banks use many rule based systems which generate alerts whenever a suspicious transaction is encountered. These alerts are manually investigated by the bank and a Suspicious Activity Report (SAR) is raised and reported to regulators when the activity is deemed suspicious. The major challenge faced by the banks is the huge number of false positives (i.e legal transactions flagged as suspicious). This leads to the banks spending a lot of time and money in investigating these false positives. Machine learning presents a quick, accurate and more cost efficient alternative compared to the existing rule based systems used by the banks. In this paper, analysis has been conducted on synthetic data using the XGBoost classification method for identifying money laundering transactions.
The paper consists of the following sections: Section-2 contains the literature survey based on detection of Anti-money laundering (AML). Section 3 includes the experimental setup and methodologies. Section 4 contains the details of the block diagram/algorithm for the chosen methodology. Section 5 contains the analysis of the results and Section 6 is the conclusion.

II. LITERATURE SURVEY

The available literature on the different methods used for detecting money laundering is limited. The existing literature on AML methods can be divided into two broad categories:

A. Unsupervised Learning
B. Supervised Learning

Unsupervised learning is a type of algorithm which learns patterns from unlabeled data. In this case, an unsupervised learning algorithm would not have any information on whether the particular transaction is related to money laundering or not. The supervised learning algorithms use labeled data to differentiate between money laundering and legitimate transactions. However, this is a problem in AML as a financial institution rarely finds out whether a particular customer is actually guilty of committing money laundering. However, this issue is resolved by modelling behavior that is suspicious instead of actual money laundering. In this paper, the focus is on unsupervised learning techniques. The existing literature shows that machine learning techniques like Naive Bayes, Decision trees, Support vector machines etc., have been used to detect money laundering.

III. SOFTWARE

Python is an interpreted high-level general-purpose programming language. Python is easily interpretable and contains a number of built-in packages. Python is It is widely used in the field of machine learning and has also been used in this paper.

©IJRASET: All Rights are Reserved

IV. PREVIOUS WORKS

In the existing work,Naive Bayes classifier was used to detect money laundering. Naive Bayes classifiers is a family of simple probabilistic classifiers based on applying Bayes theorem. It is called “naive” because its core assumption of conditional independence (i.e., all input features are independent of one another) rarely holds true in the real world. Classifiers are given different instances (object/data) to classify into predefined classes(groups), assuming there is no interdependency of features(class conditional independence). In the Bayesian analysis , the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes rule.

A. Drawbacks of Existing Work

The false positive rate for the Naïve Bayes model is 98% which is very high. A model which has very high false positive rate could lead the financial institution to spend a huge amount of money in investigating false positive cases.

V. PROPOSED METHODOLOGY

In this paper, Extreme Gradient Boosting (XGBoost) algorithm has been chosen as the preferred methodology. XGB is an ensemble based machine learning technique which produces a prediction model in the form of an ensemble of weak production models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

1) *Advantages:* The combination of individual models should produce an outcome that outperforms that of any individual member, as in every subsequent iteration the model tries to correct the error made by the model in the previous iteration. They are also relatively quick to build allowing for multiple iterations of the modelling process to refine the final model. The model also has an inbuilt cross-validation method at each iteration which takes away the need to explicitly perform cross-validation.
2) *Disadvantages:* Boosting can be more sensitive to outliers compared to other methods, thereby leading to over-fitting. It is also less easily interpretable than some other methods with some additional analysis required to understand the influence of different features.

In this paper, the XGBoost model is built using decision trees as the base models. Some of the features of XGBoost include parallelization and tree pruning. Parallelization refers to the process of building trees sequentially using parallelized implementation, which results in faster computation. The below figure shows how XGBoost uses parallelization at a branch level.

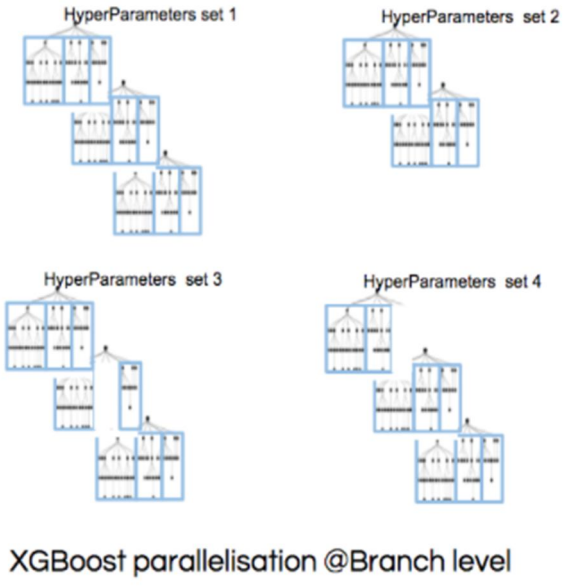


Fig. 1: XGBoost parallelization

Tree pruning refers to the process by which portions of the decision tree are removed in order to make the tree more efficient. XGBoost uses “max_depth parameter to perform this pruning.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

VI. BLOCK DIAGRAM/ALGORITHM

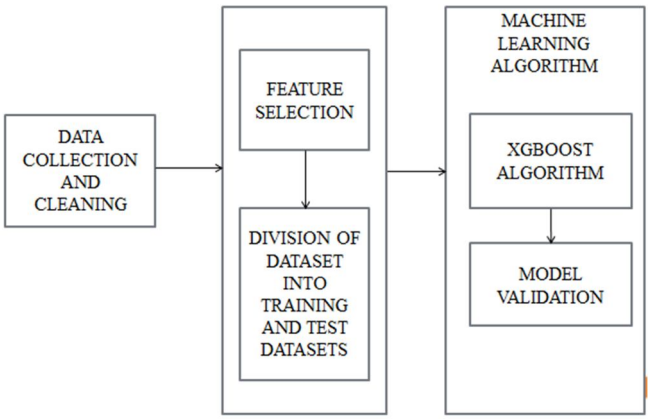


Figure 2. Block Diagram of XGBoost Machine Learning algorithm

1) STEP 1:Data acquisition and Data cleaning

Synthetic dataset was used for the purpose of this paper. The data was checked for missing values and cleaned using Python code.

2) STEP 2: Feature selection

The necessary features for the model which affect the transactions are selected in this step. The selected features are shown in Table 1.

Table 1: Features selected

Feature	Description
Type	Type of transactions like payment in/out, cash, debit, transfer
Amount	Amount transacted by the customer
Old balance of originator	Old balance of the customer who initiates/originates the transaction
New balance of originator	New balance of the customer who initiates/originates the transaction
Old balance of Beneficiary	Old balance of the customer who receives the transaction amount
New balance of Beneficiary	New balance of the customer who receives the transaction amount

3) Step 3: Division of dataset

The entire dataset was divided into two subsets as Training dataset (70%) and Test dataset (30%).

4) Step 4: Model Implementation

The model was designed with the help of XGBoost classifier.

5) Step 5: Hyperparameter tuning

The hyperparameters used in the model were tuned using RandomizedSearch CV technique. Some of he hyperparameters tuned using this technique are as follows:

a) Learning Rate - This makes the model more robust by shrinking the weights on each step.
b) Max_depth - This is the maximum depth of the decision tree.
c) Min_child_weight - Defines the minimum sum of weights of all observations required in a child.
d) Gamma - A node will be split only when the split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split in the node.

©IJRASET: All Rights are Reserved

VII. RESULTS

The following metrics were evaluated to determine the performance of the XGBoost model.

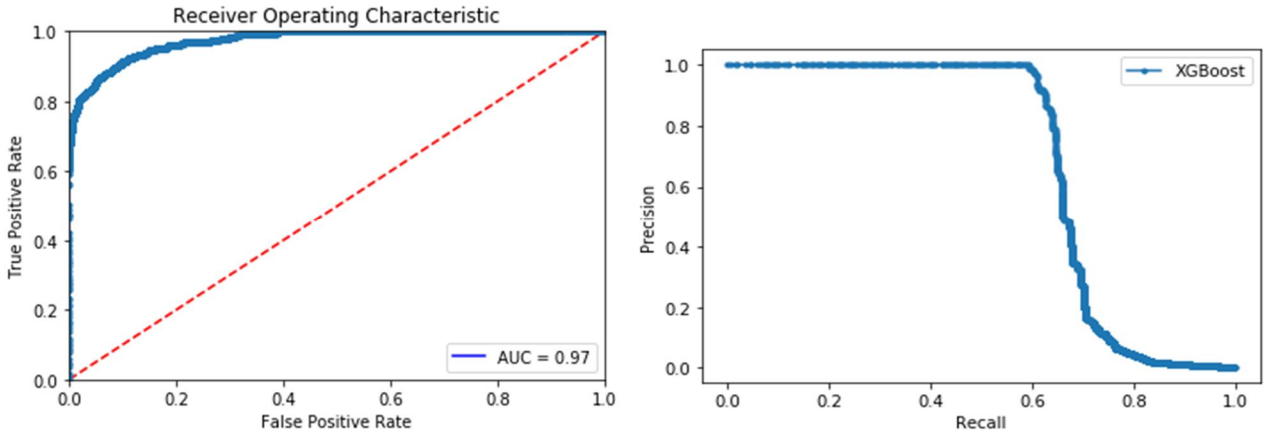


Fig. 3. ROC Curve and Precision-Recall curve

METRICS	EXISTING RESULT (NAÏVE BAYES)	PROPOSED RESULT (XGBOOST)
False positive rate	0.98	0.005
Recall or sensitivity	0.99	0.99
Precision	0.99	0.99
F-1 score	0.98	0.99
Area under ROC curve	0.63	0.97
Precision recall Area under the curve	0.08	0.68

Fig. 4. XGBoost vs Naive Bayes model

The above metrics are calculated using the following formulae:

1) Sensitivity or True Positive rate = (TP)/(TP+FN)
2) Precision = (TP)/(TP+FP)
3) Accuracy = (TP+TN)/(TP+FP+TN+FN)
4) False positive rate =(FP)/(FP+TN)=1-Specificity
5) F-1 score=2*Precision*Recall/(Precision+Recall)

VIII. CONCLUSIONS AND FUTURE SCOPE

The proposed XGBoost model shows much better performance in terms of metrics such as False positive rate, Precision-Recall Area under the curve etc., compared to the existing Naïve Bayes model.

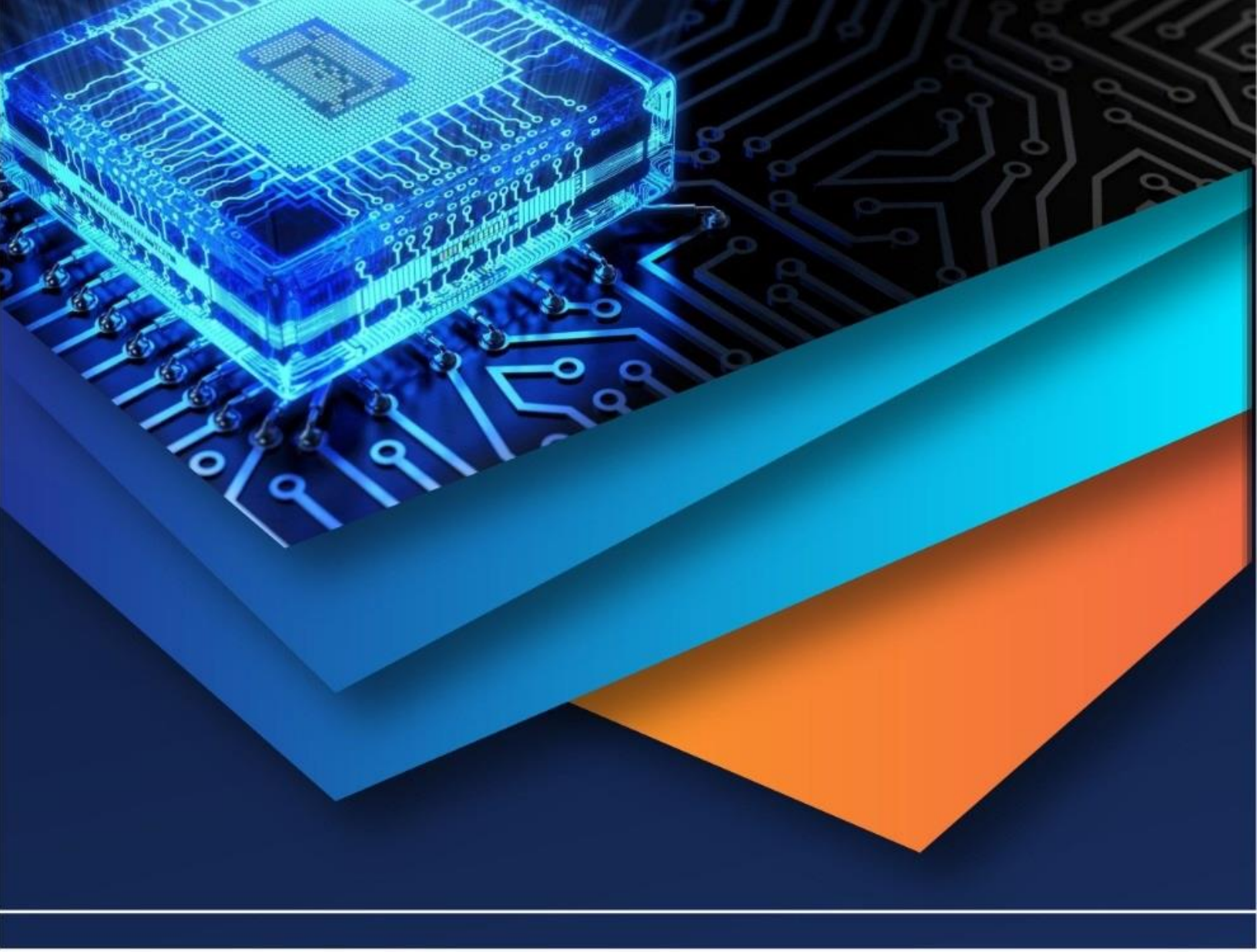
The XGBoost model has very low false positive rate(~2%) compared to the Naive Bayes model. Therefore, this model will help in reducing false positives thereby saving a lot of time and money for the bank.

©IJRASET: All Rights are Reserved

REFERENCES

[1] A.Kumar, S.Das, V.Tyagi, "Anti Money Laundering detection using Naive Bayes Classifier", 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Oct 2-4, 2020
[2] Z. Chen, L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam, "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review," Knowl.Inf. Syst., vol. 57, no. 2, pp. 245–285, 2018, doi: 10.1007/s10115-017-1144-z.
[3] C. H. Tai and T. J. Kan, "Identifying Money Laundering Accounts," Proc. 2019 Int. Conf. Syst. Sci. Eng. ICSSE 2019, pp. 379–382, 2019, doi: 10.1109/ICSSE.2019.8823264.
[4] S. Gao, D. Xu, H. Wang, and Y. Wang, "Intelligent anti-money laundering system," 2006 IEEE Int. Conf. Serv. Oper. Logist. Informatics, doi:10.1109/SOLI.2006.235721
[5] S. N. Wang and J. G. Yang, "A money laundering risk evaluation method based on decision tree," Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007, August, pp. 283–286, 2007, doi: 10.1109/ICMLC.2007.4370155
[6] Y. Jin and Z. Qu, "Research on Anti-Money Laundering Hierarchical Model," Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2018-Novem, pp. 406–411, 2019, doi: 10.1109/ICSESS.2018.8663895.
[7] Alexandre, C. and Balsa, J. (2015), "Client profiling for an anti-money laundering system", arXiv preprint arXiv:1510.00878.
[8] Colladon, A.F. and Remondi, E. (2017), "Using social network analysis to prevent money laundering", Expert Systems with Applications, Vol. 67, pp. 4
[9] Savage, D. Wang, Q. Chou, P. Zhang, X. and Yu, X. (2016), "Detection of money laundering groups using supervised learning in networks", arXiv preprint arXiv:1608.00708.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)