Evaluating Simulated Transaction Data for the Classification and Prediction of Money Laundering Activities

Ajay D Nair
Micro Data Analysis
Dalarna University
Borlänge, Sweden
h24ajana@du.se

Navib Bajracharya
Micro Data Analysis
Dalarna University
Borlänge, Sweden
v25navba@du.se

Abstract— **Money Laundering is an inherent evil in today's modern society. There is a need to be more vigilant towards financial transactions by banks and other financial institutions. While prior research has addressed detection methods, limited work has been done comparing different data balancing techniques using the SAML-D dataset. This project aims to classify and predict money laundering activity using the simulated SAML-D financial dataset from Kaggle. We evaluate the performance of two ensemble learning models—XGBoost and LightGBM—under three conditions: oversampling, under sampling, and without any balancing the dataset. Our models achieved high prediction accuracy; oversampling and under sampling does increase accuracy. Machine learning Models do classify and predict suspicious transactions. XGBoost is the better performing model whereas oversampling improves accuracy better than undersampling.**

Keywords- SAML-D; money laundering; machine learning; classification; Light GBM; XG Boost (key words)

I. INTRODUCTION

Fintech's and online banking have tremendously grown over the past years and with this comes the risk of increased financial crime[1].More and more fraudsters use loopholes in newly made financial technologies to launder their money derived from illicit activities and preventing money laundering as a financial crime has been getting more difficult. Money laundering as discussed by Altman [2] is the process wherein illegal money derived from various illicit activities like sale of drugs, stolen money, human trafficking, wildlife trade is obscured in complex financial transactions with the legal financial system to use it legally thereafter. Criminals often use money mules to this job when the money to be laundered is significantly in less amounts. Money mules are vulnerable people attracted through online parttime job spams or even at times forced into doing the job in return for a percentage of the withdrawn funds sent to their accounts , from there these funds are used by criminals in activities which they intended to further obscuring funds like buying assets[3]. Ngai [4] highlights data mining playing an important role in financial crime, uncovering patterns and hidden truth in large financial datasets which can be then used in decision making. This research identifies data mining techniques like classification, prediction, visualization. Also stating that research on data mining for money laundering has been few mainly due to the rare availability of real-world data which is due to data privacy. Synthetic financial data is an emerging area to help address challenges of data privacy and unavailability of real world data[2]. Our project uses such synthetic simulated dataset called the SAML-D dataset from Kaggle. This dataset was developed to allow researchers test and evaluate their models improving anti money laundering monitoring[5].While there are studies using classifiers to predict suspicious transactions with SAML-D dataset they necessarily do not compare the prediction results with or without balancing data.Cross validation done in our project also is something which past studies have not done. Our project aims to answer questions related to machine learning classification models and their prediction and the questions are as follows:

- Can machine learning models accurately classify transactions as suspicious based on transaction features?
- How does the performance of different machine learning models (e.g., XGBoost vs LightGBM) compare in detecting money laundering within imbalanced data?
- Does oversampling or under sampling the data improve the F1 score and recall without significantly sacrificing precision in classification?

Following the research question, we design hypothesis to be researched and proven:

- Hypothesis(H1): Alternative (H_1): ML models can accurately classify fraudulent transactions (AUC-ROC ≥ 0.85).
- Null Hypothesis(H0): Null (H_0): Model performance is no better than random guessing (AUC-ROC = 0.5).

Two ML classifiers, XG BOOST and Light GBM as known to handle imbalanced datasets were used. Both models used unbalanced and balanced data separately for the purpose of comparing results. The F1 score and AUC ROC score were obtained from each model. After analysis we conclude that ML models do classify financial transactions with great accuracy and can handle imbalanced data. However, on balancing the data, results improve. Risk profiling was done, and the models predicted transactions and flagged it with a

risk profile. On proving the hypothesis we reconfirm classifiers can work very well with fraud detection and balancing the data before training can give better results.Our work is an addition to the already existing researches undertaken in the field of data mining in identifying financial fraud[2].

II. Literature Review

A review of the recent literature shows how classifying ML models have been used to predict money laundering. Data Mining and machine learning are growing to be important factors to uncover patterns and truths in financial transactions helping uncover money laundering. Grigorescu and Amza [11]used tree-based models Random Forest and logistic regression and KNN models in classifying money laundering transactions. Particularly they used it for the convenience of explanation instead of complex models like LightGBM and XGBoost with greater accuracy.Similarly Jayan[6], used XGBoost model in classifying transactions with SAML-D dataset but did not perform cross-validation or balancing of data. As most AML datasets are imbalanced due to its nature it can still give poor recall on the minority class which is identifying transactions that are 'is laundering'[2].

In recent research focused on evaluating models for effectively classifying and predicting money laundering in financial transactions, ensemble models have demonstrated superior performance. A comparative study by Lokanan and Maddhesia[12]reinforces this finding, highlighting the advantages of ensemble tree-based algorithms such as XGBoost and LightGBM over other models. SAML-D dataset used in our project is a large dataset and has highly less transactions identified as fraud or laundering when compared to not laundering. This infact is comparable to a real dataset as fraudsters or criminals doing transactions tend to match in with regular transactions done by the public and to detect these transactions are very difficult and many go missing. To solve this class imbalance in the dataset we need some measure such as SMOTE (oversampling) or RUS (undersampling). These algorithms has been evaluated and verified to be used extensively with datasets with high class imbalance and proven to be effective [7], [8]. ZhangHao [9] also treated imbalance in data by undersampled ensemble approach and used XGBoost to classify financial transactions to detect money laundering and achieved great accuracy. In another research by Sun and Du (2025), transaction behaviour was used to profile risk and then detect and flag transactions for risk of money laundering which was also employed in a different manner by ZhangHao[9]. These studies and projects collectively affirmed ML models, particularly ensemble methods like XGBoost and LightGBM. Moreover, the application of techniques like SMOTE and RUS addresses class imbalance issues, potentially improving F1 scores and recall. Lastly, the integration of explainability methods enhances the interpretability of risk scores, aligning with the need for transparent and accountable AML systems.

III. Method Description

A. The Dataset

The dataset used in this project is the SAML-D (Synthetic Anti-Money laundering Dataset) [13]. It is a synthetic financial dataset which simulates transactions and is labelled as suspicious (potential money laundering) or not. The dataset incorporates 12 features and 28 typologies (split between 11 normal and 17 suspicious). These were selected based on existing datasets, academic literature, and interviews with AML specialists. The dataset comprises 9,504,852 transactions, of which 0.1039% are suspicious[5]. Initial exploratory data analysis also supported this as shown as in Fig 1 revealing a significant class imbalance as is with real world data[2].

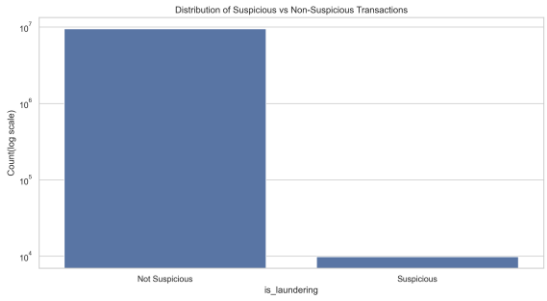

Figure 1. Showing imbalance of classes in the dataset

The dataset contains 12 features which are provided in Table 1 with their description.

Table 1: Features in the dataset with description

Features	***Description***
Time	(HH:MM: SS)
Date	(YYYY-MM-DD)
Sender Account	Account number
Receiver Account	Account number
Amount	Transaction values
Payment Currency	Currency in which payment is sent
Received Currency	Currency in which payment is received
Sender Bank Location	Indicating high risk locations
Receiver Bank Location	Indicating high risk locations
Payment Type	Like credit card, debit card, cash etc.
Is Laundering	Binary indicator for suspicious or not
Laundering Type	Money laundering types, for deeper insights

The dataset included outliers in transaction amounts, with some extremely high values affecting model stability. These were handled by applying log transformation. All amounts were converted to USD using Google exchange rates, and a 'log amount' feature was introduced to normalize skewness. After deriving necessary feature columns which are discussed in data mining method the date, time, amount, sender account, receiver account, payment currency, received currency,

sender bank location, receiver bank location and laundering type were removed. Laundering type was removed as a target variable as it is mostly used by AML investigators and is not necessarily needed in our machine learning prediction for suspicious or not. No noise or null values were observed in the dataset

Correlation analysis revealed generally weak relationships among variables; however, a strong association was observed between the currency conversion indicator and the cross-border transaction indicator. Additionally, a moderately weak relationship was identified between the receiving country's money laundering risk and the cross-border transaction indicator as shown in Fig 2. However, correlation is handled by the models effectively.

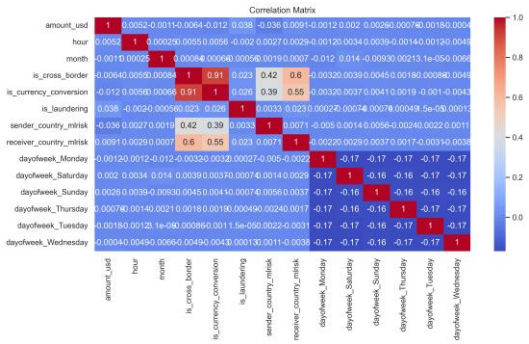

Figure 2: Correlation analysis of features

B. *Data Mining Method*

Two ensemble learning models were employed: LightGBM and XGBoost. These are gradient boosting frameworks optimized for speed and performance and are known to handle tabular, imbalanced datasets well. The dataset's imbalance was tested to be addressed by using SMOTE (Synthetic Minority Oversampling Technique), and RUS (Random Under Sampling). Oversampling increased the dataset size by nearly 200%, while undersampling reduced it to just 0.2% of its original size. The model uses engineered features such as Time-based features (year, month, hour, day of week), currency conversion indicators, cross-border transaction flags, and frequency-based features for sender and receiver accounts. Another feature 'ML risk' scores for sender and receiver countries were added based on the Basel AML Index[10]. It classifies and predicts transactions as suspicious based on these features. The model is trained on 80% of the data and tested on the remaining 20%. Under sampling and Oversampling of data was done prior to training and results of prediction were compared to one without balancing. Under sampled data being a smaller one compared to the original dataset was used to perform statistical evaluation called cross-validation. This ensured robustness of model's performance on unseen data. Evaluation was done using metrics suited for imbalanced classification: F1 score, ROC-AUC, precision, and recall. A three-tier risk scoring system classified transactions as Low (<0.3), Medium (0.3–0.7), or High Risk (>0.7) based on predicted laundering probabilities from the better-performing model. Finally, the training data was statistically evaluated with cross-validation which is discussed in Section C of Results and Analysis.

IV. Results and Analysis

A. *Model Performance Results*

Both models, LightGBM and XGBoost gave out strong results with and without balancing. The performance scores are shown in Table 2, and their respective Confusion matrices are shown in Fig 3, 4 and 5 below.

Table 2. Model performance with or without balancing

Model	***Balancing***	***Precision***	***Recall***	***F1 Score***	***AUC-ROC***
LightGBM	None	0.87	0.78	0.63	0.90
LightGBM	Under-Sampling	0.95	0.95	0.95	0.99
LightGBM	Over-Sampling	0.98	0.98	0.98	0.99
XGBoost	None	0.90	0.78	0.66	0.92
XGBoost	Under-Sampling	0.94	0.94	0.94	0.99
XGBoost	Over-Sampling	0.99	0.99	0.99	0.99

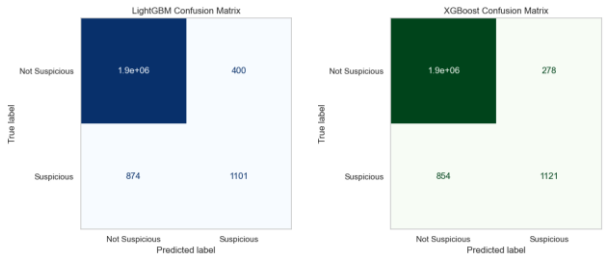

Figure 3: Confusion Matrix (No Balancing used)

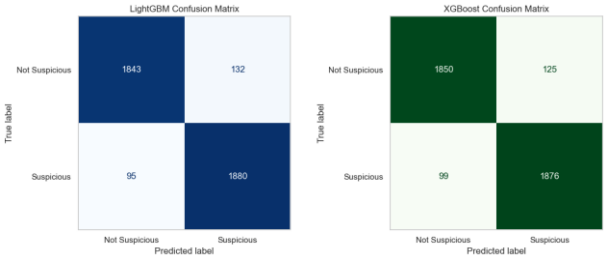

Figure 4: Confusion Matrix (Undersampling used)

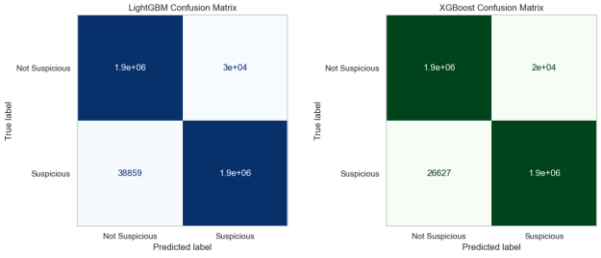

Figure 5: Confusion Matrix (Oversampling used)

The metrics which evaluate them are precision, recall, F1Score and AUC-ROC. These metrics show that both models, particularly after applying Oversampling and Undersampling referring to Table 1, were able to distinguish suspicious transactions significantly better than random chance (AUC > 0.85). This proves our hypothesis. Fig 6 shows AUC ROC curve and shows the performance of the XGBoost Model on Under Sampled Data.

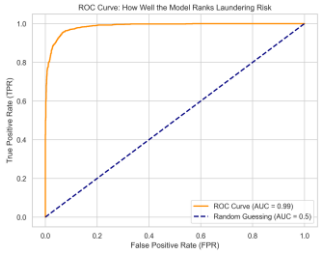

Figure 6: AUC ROC of XGBoost Model

XGBoost performed better than LightGBM under all but Undersampling by 1%. Oversampled data gives a better F1 score than undersampled data with both the models 5% more with XGBoost model and 3% more in LightGBM model.

Confusion matrix of oversampled and undersampled data in Fig 4 and 5 shows accurately predicted data, whereas models in the imbalanced dataset have more false positives and false negatives. The 5-fold cross validation applied gave the result which came out to be a mean cross validated ROC-AUC metric score of 98.70%. The risk segregation of low, medium and high was accurately established through the model's predicted probabilities for the laundering values and the division came out as follows shown in Fig 7.

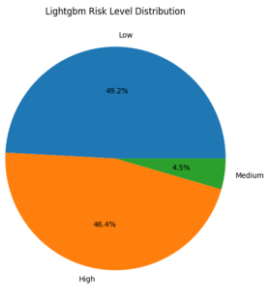

Figure 7: Risk Level Distribution

B. *Feature Importance Analysis*

The most significant predictive features identified for both models in common were receiver frequency, sender frequency and log amount. Fig 7 and 8 shows the 15 most important features used by their models respectively in prediction. These top features are the primary drivers of the model's predictions and can therefore be used to fine-tune the model for improved performance and efficiency.

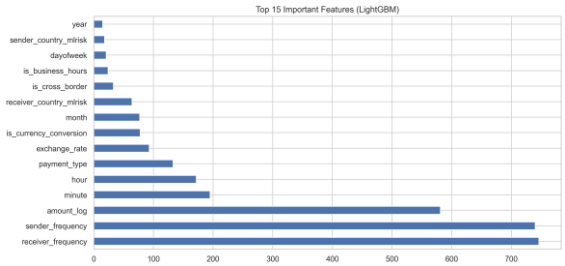

Figure 7: Features ranking for usage in prediction (LightGBM)

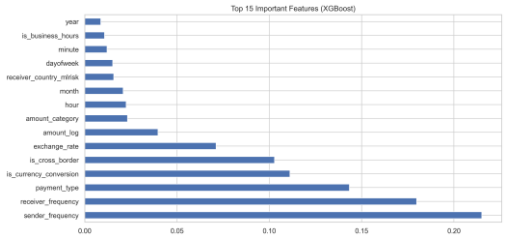

Figure 8: Features ranking for usage in prediction (XGBoost)

C. *Cross validation*

A statistical technique cross validation was applied to assess how well our model generalizes to unseen data. We used 5-fold stratified cross-validation to ensure each fold maintains the same fraud/non-fraud ratio as the original dataset. The training data was split into 5 equal folds out of which model trained on 4 folds and validated on the remaining 5th fold repeating 5 times with each fold serving as the validation set at least once. ROC-AUC score was recorded on each fold and the results were:

1st fold score: 0.9870
2nd fold score: 0.9856
3rd fold score: 0.9862
4th fold score: 0.9887
5th fold score: 0.9858
Mean CV ROC-AUC (Average of 5 folds): 0.9867

The mean score suggests that the model consistently ranked 98.67% of suspicious and non-suspicious transaction pairs correctly across all folds that explain good generalization. The model seems to have worked universally with high accuracy and reliability. Minimal variance was witnessed between the folds as performances were generalized beyond the training set.

V. CONCLUSION

Our work here confirms that classification models are good in predicting money laundering when datasets are properly balanced using under or oversampling. Their accuracy hence increases after balancing. They can predict fraud better than guessing work and should be used in companies dealing with financial risk. Our work can be carried forward and compared to unsupervised models,

however accuracy with them will require time and supportive data.

VI. Discussion

Our study validates the effectiveness of ensemble learning models in detecting money laundering activities using synthetic financial data. The convenience of oversampling (SMOTE) over undersampling can imply that making synthetic minority samples preserves transaction patterns better than reducing majority samples. The feature importance plots from the models show that behavioral patterns such as sender or receiver frequency matter more than transaction amount to create effective outcomes. For future study, the models can be used on real-world transaction data to capture dynamic laundering techniques.

References

[1] E. A. Akartuna, S. D. Johnson, and A. Thornton, "Preventing the money laundering and terrorist financing risks of emerging technologies: An international policy Delphi study," Technol. Forecast. Soc. Change, vol. 179, p. 121632, Jun. 2022, doi: 10.1016/j.techfore.2022.121632.

[2] E. Altman, J. Blanusa, L. Niederhäusern, B. Egressy, A. Anghel, and K. Atasu, Realistic Synthetic Financial Transactions for Anti-Money Laundering Models. 2023.

[3] B. Nikkel, "Fintech forensics: Criminal investigation and digital evidence in financial technologies," Forensic Sci. Int. Digit. Investig., vol. 33, p. 200908, Jun. 2020, doi: 10.1016/j.fsidi.2020.200908.

[4] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decis. Support Syst., vol. 50, no. 3, pp. 559–569, Feb. 2011, doi: 10.1016/j.dss.2010.08.006.

[5] B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, H. Dogan, and G. Aksu, "Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset," in 2023 IEEE International Conference on e-Business Engineering (ICEBE), Nov. 2023, pp. 47–54. doi: 10.1109/ICEBE59045.2023.00028.

[6] "GitHub - JayanGupta/Antimoney_Laundering: Machine learning-based fraud detection using XGBoost on the SAML-D dataset with EDA, preprocessing, and model evaluation to identify suspicious transactions." Accessed: May 24, 2025. [Online]. Available: https://github.com/JayanGupta/Antimoney_Laundering

[7] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in 2021 Sixth International Conference on Informatics and Computing (ICIC), Nov. 2021, pp. 1–8. doi: 10.1109/ICIC54025.2021.9632912.

[8] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," Information, vol. 14, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/info14010054.

[9] "GitHub - ZhangHaoXuan21/SAML-D-AML-Synthetic-Dataset-Project." Accessed: May 24, 2025. [Online]. Available: https://github.com/ZhangHaoXuan21/SAML-D-AML-Synthetic-Dataset-Project

[10] "Basel AML Index - Assessing money laundering risks around the world," Basel AML Index. Accessed: May 27, 2025. [Online]. Available: https://index.baselgovernance.org/

[11] "Explainable Feature Engineering for Multi-class Money LaunderingClassification. Petre-Cornel GRIGORESCU, Antoaneta AMZA Bucharest University of Economic Studies, Romania /DOI: 10.24818/issn14531305/29.1.2025.06

[12] M. Lokanan and V. K. Maddhesia, *Predicting Suspicious Money Laundering Transactions using Machine Learning Algorithms*. 2023. doi: 10.21203/rs.3.rs-2530874/v1.

[13] "Anti Money Laundering Transaction Data (SAML-D)." Accessed: Jun. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/berkanoztas/synthetic-transaction-monitoring-dataset-aml