Financial Crime Detection: A Machine Learning Approach to Anomalous Financial Transaction Detection using SWIFT synthetic dataset

Sagar Pathak
Computer Science
University of Memphis
spathak1@memphis.edu

Bidhya Shrestha
Computer Science
University of Memphis
bshrstha@memphis.edu

ABSTRACT

UPDATED—December 1, 2022. A financial crime is an activity committed by an individual or group of individuals that involves illegal conversion of the property's ownership from one to another. It includes different types of financial fraud, money laundering, bribery and corruption, insider trading, terrorist financing, and cybercrime. Such crimes can adversely affect the social-security systems and destabilize a country's economy, governance, politics, and society. With the rise of artificial intelligence and big data, new possibilities have arisen in using advanced machine learning models to detect financial fraud. We experimented the dataset with different machine learning classification models named as Logistic Regression, Random Forest, Decision Tree, and XGBoost. The experiment showed that the XGBoost performs best for the fraudulent vs non-fraudulent data classification in balanced dataset with hyperparameter tuned with 0.953 AUPRC (Area Under Precision Recall Curve). The XGBoost wins over all other machine learning models used in this project on imbalanced as well as balanced dataset.

Author Keywords

Financial Crime; Money Laundering; XGBoost; AUPRC; Random Forest; Decision Tree; Logistic Regression

INTRODUCTION

No system is immune to the threat or challenges, and the more entities involved, the more it's prone to vulnerabilities within the system. The primary goal of the financial system is very straightforward, which is to facilitate the transfer of resources from savers to those who need funds [5]. The more it looks straightforward, the more it is intricate in nature, consisting of several parties with their own federation, policies, rules, and regulations. Financial crime is one of the most significant threats or challenges that financial institutions pose nowadays; however, with technological advancement in big data and artificial intelligence, it can be framed in a data science problem and devise solutions using machine learning techniques and algorithms.

Moreover, it is well known that with the unparalleled growth of mobile banking and digital payments, billions of people can access financial services at their fingertips. These systems are advantageous in time savings and providing ease of use, scalability, higher speed, and lower transaction costs to the business, financial service providers and individual customers. However, financial crimes are also taking a new bigger shape, and it is very likely that criminals can bypass the traditional fraud and financial crime detection systems built under static rules and threshold conventions. According to the UN, approximately 800 to 2000 billion dollars money laundered annually, which is in contrast two to five percent of global GDP [2]. So the scale of this problem is huge and also, to come up with the solution is very challenging because of high volume of financial transactions made each seconds as well as the sensitivity and the privacy of the transaction data. It requires to have a system that can efficiently identify and categorize as fraud and non-fraud activity through the nature of the transactions made. Thus, this project tries to surface the financial crime committed through digital channels and aims to solve financial fraud problem by creating a solution to classify the transactions committed over the network into fraudulent or non-fraudulent, which in result helps to build an illegal financial activity detection system that could be effective and advantageous for financial institutions.

RELATED WORK

Gottschalk, P. (2010) [8] classified financial crimes into four main categories: corruption, fraud, theft, and manipulation. These categories are subdivided into several subclasses on which money laundering comes under the manipulation category. The purpose of laundering is to make the transaction appear as if the proceeds were acquired legally, as well as disguises its illegal origins [3]. The research by Zeinab Rouhollahi (2021) [12] on financial crime detection using artificial intelligence mainly addressed money laundering detection using data collected through Australian banks. It utilized the Snorkel Model [11] to auto-label the transactions with the help of different labeling functions created in coordination with banking experts. This research used five different classification methods: logistic regression, Nearest Neighbours, Naive Bayes, Neural Network, and Random Forest as the classifier to categorize money-laundering transactions. The model accuracy, precision, recall, and F1 score is evaluated for each

models where Neural Network (92.4% accuracy) performed the best as comparison to other classification methods.

The credit card fraud is another financial crime where the fraudster uses the stolen credit card details to secure goods or services in the name of the cardholder. Itoo and Singh (2021) [7] used three classification techniques, namely logistic regression, Naïve Bayes and K-nearest neighbor (KNN) to detect fraudulent credit card transactions. They used Kaggle credit card dataset [10] that contains transactions made by credit cards in September 2013 by European cardholders and found the logistic regression model to perform best with an accuracy of 96%. Primarily, the classifier was trained on the original dataset which was highly imbalanced with large negative classes of datapoints resulting poor recall score. Thus, they decided to subset the dataset into three categories where the dataset A had radio 50:50, dataset B had the ratio 34:66 and dataset C had the ratio 25:75 of fraud by non-fraud datapoints. And they achieved 96% accuracy using logistic regression model on dataset C where in contrast other models got slightly low accuracy.

There are several variable factors and attributes attached to a single financial transaction. So it is burdensome to optimize the model just by tweaking parameters. Thus, Ileberi et. al. (2022) [9] used the genetic algorithm (GA) for feature selection for credit card fraud detection using machine learning. The genetic algorithm is a type of evolutionary computation and the adaptive heuristic search algorithm inspired by Charles Darwin's theory of evolution in nature. GA is often used to solve several optimization tasks with reduced computational overhead. The research by Ileberi used GA to generate five optimal features from the dataset that originally had many feature vectors. This optimized the running time of the training process and power consumption of the system and improved the model's overall accuracy. They proposed detection model using Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network, and Native Bayes. The results of this model were assessed using the metrics such as accuracy, precision, and ROC curve and it is shown that the decision tree classifier performed well with an accuracy of 89.91% when compared to the other classifiers.

Compared with the recapitulation of these three related works, it is observed that result and the accuracy varies by the choice of different models, process and as well as the training data. The work by Zeniab Rouhollahi tried to detect money laundering activity and achieved accuracy using Neural Network. Team of Itoo and Singh worked on European credit card transaction dataset for the detection of credit card fraud and they found logistic regression model worked well into their setup, whereas Illeberi got succeed on getting higher accuracy using Decision Tree model as compared to other classification models for the financial fraud. Thus, it confers that there is no obvious or direct model that will work for all problem scenarios. So, in this paper we also trained model using different classification methods and performed the evaluation and assessment of the model.

CLASSIFICATION METHOD

In this paper we applied different classification methods on SWIFT dataset for financial crime detection. The goal of the classification problem is to classify or categorize a new data point into some defined classes such as fraud or non-fraud. There are several state-of-the-art classification algorithms which has their own pros and cons and furthermore, the usecase differs based on the nature of the process and the dataset used. These are briefly described in the below subsections. Logistic regression uses the below linear model that represents the plane or hyperplane that separates the instance of one class from another.

Logistic Regression

Logistic Regression is the algorithm that learns the linear model and can be used for classification and probability estimation. This can be very effective when the problem is linearly separable or there are a lot of relevant features. And also, when we want the efficient runtime algorithm which is easier to learn and apply into various structured datasets.

$$LinearModel : w_0 + \sum_i^n w_i * x_i$$

Random Forest

Random Forest is the supervised machine learning algorithm that combines multiple decision trees (forest) with all the random variable selection and averages out together to build a final model. This generally gives the much higher prediction accuracy but as a tradeoff, we cannot look the bottom leaf of those trees which creates difficulty in model interoperability. This is useful for regression and classification tasks and can handle large datasets efficiently.

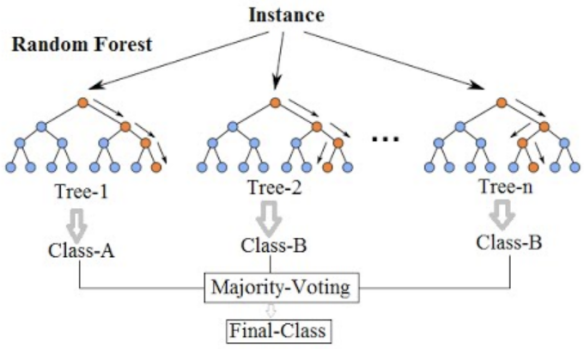

Figure 1. Illustration of Random Forest displaying n-trees.

XGBoost (Extreme Gradient Boosting)

XGBoost also known as Extreme Gradient Boosting machine learning library based on decision tree which is gradient-boosted. Gradient boosting is an algorithm where the learning happens by optimizing the loss function. This uses two type of base estimators, the average type model and the decision tree with full depth. The XGBoost is a supervised machine learning model which builds upon decision trees, ensemble learning, and gradient boosting. The basic idea of boosting is

to combine weak models with other weak models and collectively generate the strong model. The boosting helps to reduce bias and under-fitting.

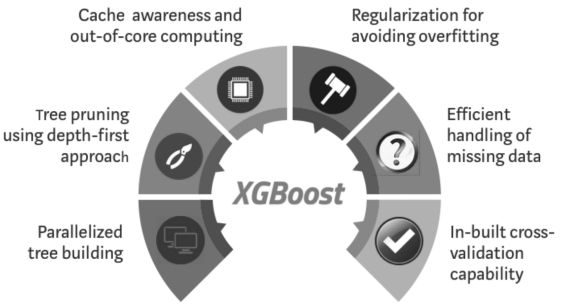

Figure 2. Features of XGBoost algorithm.

Decision Tree

Decision Tree is also a supervised machine learning model which is used to solve regression and classification problems. It helps to predict or categorize the class based on how previous set of questions were answered. Some of the key terms of decision trees are: root node, decision node, leaf node, branch, pruning of nodes, splitting of node etc. There are two types of decision trees: Categorical variable decision tree and continuous variable decision tree. Decision trees can handle high-dimensional data and requires less data cleaning than other data modeling techniques.

SWIFT SYNTHETIC DATASET

This project utilizes the dataset[4] created by SWIFT (Society for Worldwide Interbank Financial Telecommunication) which represents the transaction made through SWIFT network by 50 financial institutions connected into the network. This is not a real data but mimics the properties and behaviour of real transactions. Each row in the dataset represents a payment transaction from one sending bank to receiving bank. The dataset contains about 4 million rows in total. This dataset contains the following fields:

- **MessageId** - Globally unique identifier within this dataset for individual transactions
- **UETR** — A 36-character string enabling traceability of all individual transactions associated with a single end-to-end transaction
- **TransactionReference** - Unique identifier for an individual transaction
- **Timestamp** - Time at which the individual transaction was initiated
- **Sender** - Institution (bank) initiating/sending the individual transaction
- **Receiver** - Institution (bank) receiving the individual transaction
- **OrderingAccount** - Account identifier for the originating ordering entity (individual or organization) for end-to-end transaction,
- **OrderingName** - Name for the originating ordering entity
- **OrderingStreet** - Street address for the originating ordering entity
- **OrderingCountryCityZip** - Remaining address details for the originating ordering entity
- **BeneficiaryAccount** - Account identifier for the final beneficiary entity (individual or organization) for end-to-end transaction
- **BeneficiaryName** - Name for the final beneficiary entity
- **BeneficiaryStreet** - Street address for the final beneficiary entity
- **BeneficiaryCountryCityZip** - Remaining address details for the final beneficiary entity
- **SettlementDate** - Date the individual transaction was settled
- **SettlementCurrency** - Currency used for transaction
- **SettlementAmount** - Value of the transaction net of fees/transfer charges/forex
- **InstructedCurrency** - Currency of the individual transaction as instructed to be paid by the Sender
- **InstructedAmount** - Value of the individual transaction as instructed to be paid by the Sender
- **Label** - Boolean indicator of whether the transaction is anomalous or not. This is the target variable for the prediction task.

The dimensionality of the training dataset is 4691725x19 (rowsxcolumns). The number of non-fraudulent transactions are 4686825 and fraudulent transactions are 4900 in training set. Also, the dimensionality of training dataset is 705108x19 where this dataset consists of 704347 non-fraudulent and 761 fraudulent transactions.

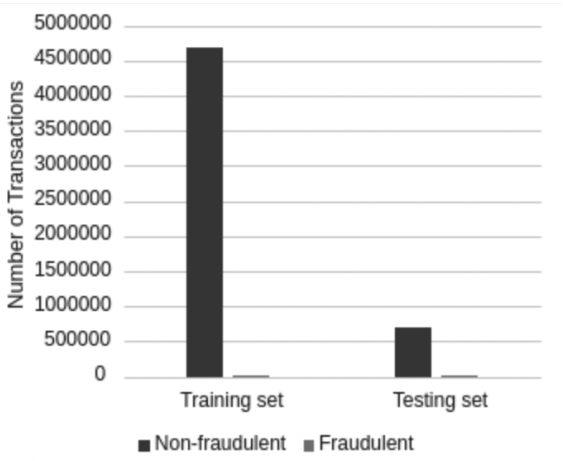

Figure 3. Comparison of fraudulent vs non-fraudulent transations in training and testing dataset

Analysing the size and dimentionality of the dataset and the records of fraudulent vs non-fraudulent transaction (as shown in Figure 3), we found that it is heavily imbalanced. Thus

we decided to run the experiment in both imbalanced and balanced dataset. We used two techniques to create balanced dataset which is described in later sections.

DATA PROCESSING

The data preprocessing is done for creating balanced dataset from the imbalanced SWIFT synthetic dataset[4]. Imbalanced datasets are those where the class distribution is heavily skewed, for example 1:100 or 1:1000 examples in the minority class to the majority class, in our case fraud vs non-fraud. We observed that our training dataset is skewed in the ratio of 1:956 in fraudulent class to the non-fradulent class. We used two techniques: Random Undersampling Method and Synthetic Minority Oversampling Technique for handling imbalanced data.

Random Undersampling (RUS)

Random Undersampling[13] is one of resampling approaches to tackle issues with imbalance data by removing instances randomly from the majority class. The Random Under Sampling method sampled the majority class of non-fraudulent transaction to 49000 and 7610 in training and testing set respectively. In this method, the data are randomly selected from the non-fraudulent transactions and then removed from the training and testing set based on sampling strategy.

Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) [6] is a preprocessing technique used to address a class imbalance in a dataset. It does the over sampling of the minority classes without overfitting by generating new synthetic examples close to the other points in feature space. We over sample minority class i.e. fraudulent class data in our project to 49000 and 7610 in training and testing sets respectively. In this method, the data from the minority class is duplicated to synthesize new data in the training and testing set. This balances the class distribution but does not provide any additional information to the model.

Data cleaning

The two columns in the dataset named OrderingName and Beneficiary Name had null values. These columns have categorical features so we handled the missing values in those columns by replacing it with the categorical feature having highest frequency i.e. mode. We replaced null values with "SYMPHORICARPOS ORBICULATUS" and "ACER SACCHARUM" in OrderingName and BeneficiaryName respectively.

Data Normalization

Data Normalization is a process of managing the data so that it looks similar across fields and appears organized in proper way. We normalized the training data and testing data since the ranges of the values were different. We transformed the columns of the dataset to the same scale without distorting the differences in the ranges.

Feature Selection

The SWIFT dataset consisted of 19 columns and to assist further on the fraud detection, we created and added eight extra feature column utilizing the existing one. Thus, we have 27 columns as a whole in the dataset. And for feature selection we dropped the categorical columns and only kept those columns which had higher impact on the target variable. We used heatmap to find out the correlation between the numerical columns.

EMPIRICAL EXPERIMENTS

The experimental was conducted on Lenovo IdeaPad 5 (16GB RAM) with Intel® Iris® Xe graphics and Macbook Pro M1 (16GB RAM) with 16-core GPU. We analyzed the performance of the chosen four models on SWIFT dataset using the Area Under Precision Recall Curve (AUPRC) as performance metrics as our dataset is highly imbalanced.

We started the experiment with the original imbalanced dataset and the result was very poor. The AUPRC for Logistic Regression was 0.003, this was extremely low because we found that the model was not able to predict any data as fraudulent due to the highly imbalances on the dataset. As a result the F1 score was 0 and such as the AUPRC. Also, none of the model got more than 0.60 AUPRC.

AUPRC on imbalanced data

Model	*AUPRC*
Logistic Regression	0.003
Random Forest	0.500
Decision Tree	0.340
XGBoost	0.598

Table 1. Experiment result for different models on imbalanced SWIFT dataset.

Then we created the balanced dataset using the technique mentioned in the data procession section and trained and tested the models on it. This time, the AUPRC of Random Forest and XGBoost increased sharply to 0.93 and 0.950 respectively as shown in the table.

AUPRC on balanced data

Model	*AUPRC*
Logistic Regression	0.678
Random Forest	0.920
Decision Tree	0.794
XGBoost	0.950

Table 2. Experiment result for different models on balanced SWIFT dataset.

Also, we tuned the hyperparameters of the models using the GridSearchCV [1] method. Hyperparameters are the top-level parameters that control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The values of hyperparameters are set before training the model and will remain the same even when the training ends.

The result is displayed in the following bar chart:

AUPRC on balanced data with hyperparameter tuning	
Model	*AUPRC*
Logistic Regression	0.683
Random Forest	0.944
Decision Tree	0.937
XGBoost	0.953

Table 3. Experiment result for different models on balanced SWIFT dataset with hyperparameter tuning.

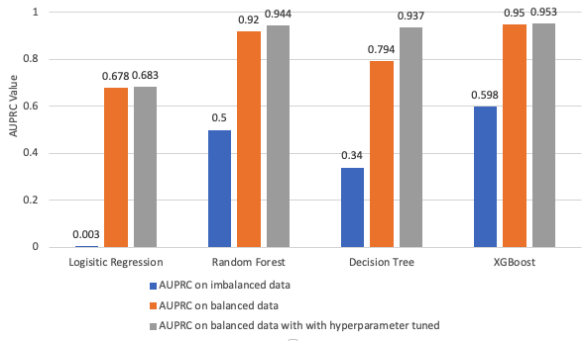

Figure 4. Performance analysis of Machine Learning models for financial crime detection

As we can see on the bar chart above, the XGBoost performed well with 0.953 AUPRC as compared to other machine learning models.

DISCUSSION

The experiment showed that the XGBoost is superior for financial fraud detection while using SWIFT synthetic dataset. The XGBoost minimises a regularised objective function that incorporates a loss function and the penalty term for model complexity. It uses a gradient descent algorithm to minimise the loss. Initially, the classification models such as Random Forest, Decision Tree, Logistic Regression and XGBoost were not performing well and the overall AUPRC was under 0.6. The Logistic Regression was giving extremely low, 0.03 AUPRC since it was not able to predict any data as the fraudulent due to high data imbalance on the dataset. Thus, we analyzed the dataset and build the balanced dataset using Random Undersampling (RUS) [13] and SMOTE [6].

The Random Under Sampling method sampled the majority class of non-fraudulent transaction to 49000 and 7610 in training and testing set respectively. Also, we over sampled the minority class i.e. fraudulent class data in our project to 49000 and 7610 in training and testing sets respectively using Synthetic Minority Over-sampling Technique (SMOTE). We ran the experiment again, and it was a drastic improvement. The result was 0.678 AUPRC using Logistic Regression, 0.92 AUPRC using Random Forest, 0.794 AUPRC using Decision Tree and 0.95 using XGBoost.

However, we wanted run the experiment by tuning the hyperparameter. HyperParameter Tuning is the way of finding out the right combination of hyperparameters that optimizes the machine learning model performance. There are different types of hyperparameter tuning methods like Random Search, Grid Search, Bayesian Optimization etc. However, we have used Grid Search Hyperparameter tuning method for our experiment.

Grid Search is an exhaustive process of hyperparameter tuning in which the model is trained with each and every possible combination of hyperparameters. The model performance of each combination is recorded and finally, the best model with the best set of hyperparameters is returned. As Grid Search tries out every combination, it took a long time to tune the hyperparameters of the models, especially for XGBoost.

The Decision Tree AUPRC improved a lot when the hyperparameters are tuned. However, overall the XGBoost performed best with 0.953 AUPRC and then Random Forest with 0.942 AUPRC followed by Decision tree with 0.935 AUPRC and Logistic Regression with 0.683 AUPRC.

CONCLUSION

We analyzed the performance of Logistic Regression, Random Forest, Decision Tree, and XGBoost algorithm to classify fraudulent vs non-fradulent bank transactions. The dataset was highly imbalanced so we balanced it through sampling. We performed hyperparameter tuning, using GridSearchCV[1]. The four machine learning classification models are evaluated based on AUPRC value, which ranges from 0 to 1 (0 means poor and 1 means best). From the experiments, we concluded that XGBoost performs the best for detecting fraudulent transactions in highly imbalanced and balanced data, while Logistic Regression performs the lowest. GridSearchCV has improved the performance of Decision Tree immensely, but still XGBoost got the highest AUPRC.

REFERENCES

[1] 2022a. GridSearchCV for Beginners. It is somewhat common knowledge in the... | by Scott Okamura | Towards Data Science. https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee. (11 2022). (Accessed on 12/01/2022).

[2] 2022. Money Laundering. Website. (2022). Retrieved from https://www.unodc.org/unodc/en/money-laundering/overview.html.

[3] 2022. Money Laundering - Financial Action Task Force (FATF). https://www.fatf-gafi.org/faq/moneylaundering/. (November 2022). (Accessed on 11/20/2022).

[4] 2022b. SWIFT Synthetic Dataset: PETs Challenge. https://www.drivendata.org/competitions/98/nist-federated-learning-1/page/524/. (11 2022). (Accessed on 12/01/2022).

[5] Anjan V. Thakor Arnoud W. A. Boot. 2015. Financial System Architecture. Blog. (03 April 2015). Retrieved August 22, 2014 from https://academic.oup.com/rfs/article/10/3/693/1635559.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (jun 2002), 321–357. DOI: http://dx.doi.org/10.1613/jair.953

[7] Satwinder Singh Fayaz Itoo, Meenakshi. 2021. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. (August 2021). DOI: http://dx.doi.org/10.1007/s41870-020-00430-y

[8] Petter Gottschalk. 2010. Categories of financial crime. *Journal of Financial Crime* 17 (10 2010), 441–458. DOI: http://dx.doi.org/10.1108/13590791011082797

[9] Emmanuel Ileberi, Yanxia Sun, and Zenghui Wang. 2022. A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data* 9 (02 2022). DOI: http://dx.doi.org/10.1186/s40537-022-00573-8

[10] Kaggle. 2021. Kaggle-Credit Card Fraud Dataset Dataset. (June 2021). https://paperswithcode.com/dataset/kaggle-credit-card-fraud-dataset (Accessed on 11/20/2022).

[11] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré . 2017. Snorkel. *Proceedings of the VLDB Endowment* 11, 3 (nov 2017), 269–282. DOI: http://dx.doi.org/10.14778/3157794.3157797

[12] Zeinab Rouhollahi. 2021. Towards Artificial Intelligence Enabled Financial Crime Detection. (2021). DOI: http://dx.doi.org/10.48550/ARXIV.2105.10866

[13] Mulyana Saripuddin, Azizah Suliman, Sera Syarmila Sameon, and Bo Norregaard Jorgensen. 2022. Random Undersampling on Imbalance Time Series Data for Anomaly Detection. (2022). DOI: http://dx.doi.org/10.1145/3490725.3490748