A Novel GBT-Based Approach for Cross-Channel Fraud Detection on Real-World Banking Transactions

Uğur Dolu, Emre Sefer

A Novel GBT-based Approach for Cross-Channel Fraud Detection on Real-World Banking Transactions

Uğur Dolu[1,2][0000−0001−9711−4204] and Emre Sefer[2][0000−0002−9186−0270]

[1] Yapi Kredi Technology, Istanbul,Turkey ugur.dolu@ykteknoloji.com.tr
[2] Ozyegin University, Istanbul, Turkey
{ugur.dolu,emre.sefer}@{ozu,ozyegin}.edu.tr

Abstract. The most recent research on hundreds of financial institutions uncovered that only 26% of them have a team assigned to detect cross-channel fraud. Due to the developing technologies, various fraud techniques have emerged and increased in digital environments. Fraud directly affects customer satisfaction. For instance, only in the UK, the total loss of fraud transactions was £1.26 billion in 2020. In this paper, we come up with a Gradient Boosting Tree (GBT)-based approach to efficiently detect cross-channel frauds. As part of our proposed approach, we also figured out a solution to generate training sets from imbalanced data, which also suffers from concept drift problems due to changing customer behaviors. We boost the performance of our GBT model by integrating additional demographic, economic, and behavioral features as a part of feature engineering. We evaluate the performance of our cross-channel fraud detection method on a real banking dataset which is highly imbalanced in terms of frauds which is another challenge in the fraud detection problem. We use our trained model to score real-time cross-channel transactions by a leading private bank in Turkey. As a result, our approach can catch almost 75% of total fraud loss in a month with a low false-positive rate.

Keywords: Gradient Boosting Tree · Cross Channel Fraud · Concept Drift · Imbalanced Data.

1 Introduction

Frauds are an inevitable loss for banks and financial institutions. Banks lose reputation and money as a result of fraudulent transactions since customers give importance to reliability when choosing the bank to keep their money. The number of fraudulent transactions and the amount lost due to fraud tend to increase each year. When fraudulent transactions which are taking place all over the world are analyzed, the financial cost of fraud in 2021 is £4.37 trillion [9] which was £3.89 trillion [8] in 2019. The method used by fraudsters also change over time. The most common techniques are phishing, social engineering, verbal persuasion, and computer viruses like Trojan. The total loss of fraud transactions

in the UK alone in 2020 is £1.26 billion. Among these fraudulent transactions, 38% of them are performed by remote banking, 45% of them by credit card, whereas 16% of them are performed by social engineering.

For the remote banking, the total amount of fraud loss was £197.3m in 2020. However, when the first six months of 2020 and 2021 are compared, the amount of loss advanced from £79.7m to £133.4m which marks an increase of 67%. Remote banking has three subcategories which are internet banking, mobile banking, and telephone banking. The number of fraudulent transactions in telephone banking has been decreasing over the years. From 2020 to the first six months of 2021, the number of transactions was decreased by 50%. In 2020 the total financial cost of social engineering fraud methods was £479m. However, when the first six months of 2020 are analyzed, the amount is observed as £207.8m, and when the first six months of 2021 are examined, 71% of the increase is detected, which brings the number up to £355.3m [18, 19].

Fraudsters prefer to use remote banking channels with social engineering techniques to persuade victim customers. According to the statistics taken from our real dataset from a private bank in Turkey, more than 40 million transactions take place in each month. However, only around 800 transactions are labeled as fraud. The ratio of fraud transactions is approximately 0.002. Besides that, most academic works used synthetic or static datasets. This kind of work is not applicable to realistic scenarios. In this case, the main dilemma is that fraud methods and customer behaviors are evolving over time, but the model is trained with only synthetic or static data [12].

The concept drift and highly imbalanced data problems are dominant in the detection of channel frauds and need to be targeted with caution. The transactions from accounts and channels form a highly imbalanced dataset, including very few fraud transactions and many more legitimate ones. To solve this problem, various methods were employed, namely undersampling [20] and oversampling [3]. However, these techniques are still problematic because the underlying dataset is exceedingly imbalanced, and instances of the dataset individually carry important information (such as transactions belonging to the same account or customer).

In this study, we have experimented with different equalization techniques; transaction-based, account-based, and customer-based to alleviate the imbalanced data problem. Furthermore, most of the traditional machine learning algorithms may easily be overwhelmed by the majority class in imbalanced data, leading to higher misclassification rates on the minority classes. To overcome this issue, we introduce boosting models such as gradient boosting trees (GBTs). We employ GBT in our study, in which the key idea is to ensemble weak decision trees, is a commonly used machine learning method for binary classification on the imbalanced data. In summary, the contributions of this paper are as follows:

- Our approach supports multiple transaction channels and is capable of detecting frauds between cross-channels.
- Our approach employs feature engineering, concatenating transaction details with customers' demographic and financial information.

– Our approach does not need any additional maintenance, the auto train algorithm will generate the training set from historical data, and will retrain the model. Our method will be applicable in realistic scenarios since: 1- An automated training mechanism helps to adapt to new behaviors of account holders and fraudsters, 2- Real data has all fraudsters and customer habits so our approach will not struggle with aging, 3- We generate training sets according to sampling method, which balances out the data and improves the performance of the model.

This paper is organized as follows: Section 2 discusses the related work. Section 3 gives the methodology for generating the training set, feature engineering, and the remaining details of our study. The performance of feature engineering techniques of GBT models for extremely imbalanced data is discussed in Section 4 in detail, which is followed by the conclusion section.

2 Related Work

The cross-channel fraud detection systems have to deal with two main issues: binary classification on an imbalanced dataset and handling the concept drift. In this section, in addition to providing a brief overview of the aforementioned problems, we also discuss some of the existing applied approaches to fraud detection for other types of transactions, including credit card.

There are many commercial solutions developed to protect financial organizations and their customers from fraudsters, and many academic works have been developed in this area. FICO is one of the market leaders in this area, with many kinds of software solutions. Most of the products try to catch fraud transactions according to the outputs of the rule-based algorithms. The rule-based algorithms are effective, although they are incapable of capturing the dynamically-changing fraudsters' strategies over time. To provide sustainability, rule-based systems are required to provide a periodic maintenance and detailed analysis to modify the rules and their conditions, which requires a significant amount of human effort. Additionally, the most recent research shows that only 26% of all financial institutions have a team allocated to detect cross-channel frauds [17].

In the literature, only a few number of studies have focused on the concept drift problem in fraud detection task. For example, [5] proposed a method based on a sliding window and an ensemble of classifiers to overcome this problem in the credit card fraud detection. In another study, a fraud detection system based on a concept drift management approach has been presented to retain new concepts on the transaction streams using the cardholder profiles [11]. However, both of these studies have been evaluated on synthetic datasets due to the unavailability of the real benchmark datasets. In this study, we used an approach for the generation of an automatic training set, which will be explained in detail later, to adapt to the drifts on the data as much as possible. The conventional fraud detection systems are developed using the previously known fraud transactions and cannot easily adapt to concept drift.

Moreover, another frequently encountered problem in real-world tasks is learning from imbalanced data for binary- or multi-class classification, where the models are trained on a dataset with an imbalanced distribution of classes [21]. Imbalanced Data, in real-world applications suffer from class imbalance problem. A predictive model trained on imbalanced data tends to correctly predict the majority class samples and to misclassify the samples from minority classes. Therefore, several studies have proposed different approaches to select a subset of training data principally for binary classification in fraud detection. For example, [14] demonstrated how the class distribution in the training set affects the performance of credit card fraud detection. In their experiments, training sets with varying the number of fraudulent transaction distributions from 10% to 90% for each month were utilized to generate a meta-learning model. They obtained the most appropriate performance on the reduction of prediction loss and training time with a $50\% - 50\%$ class distribution. [2] showed how the performance improves when the majority class is undersampled on the training set. On the other hand, some studies have employed account-based undersampling methods for the construction of the training set [10]. In this work, we also compared transaction- and account-level undersampling strategies by equalizing the counts of fraudulent and legitimate transactions while making sure that all the transactions of both types cover the same time interval in the training sets. Similarly, we use undersampling as a solution to the imbalance problem. However, instead of randomly undersampling data, we generate a training set by equalizing fraudulent and legitimate credit card counts, and transactions of each credit card cover the equal time range [1]. We applied this solution to our problem, instead of credit card count, we instead tried to equalize fraudster and legitimate account counts. Additionally, this structure was tested with different ratios between a swindler and reliable account counts.

Feature Engineering in real-time transactions, customer behaviors are crucial, and their habits can evolve over time. To minimize the error that arises from behaviors [6], the features such as "familiar device", "is account whitelisted or blacklisted", and "is device used for another account before" are integrated to the main transaction dataset. Additionally, to detect more accurate behavioral data, we have also included historical maximum and minimum credit scores, customer age, and financial age [15]. Furthermore, combining transactions with commercial or individual customer-type flags has improved the model's prediction performance in a positive way.

Another cross-channel fraud detection framework [12] proposed a graph analysis extracted from realized transactions in real-time. Their analysis examines the shortest paths between transactions and strongly-connected components in the transaction graph to detect fraudulent transactions. Another work applies a recurrent neural network [13] for cross-channel fraud detection. In the literature, there are not many examples of GBTs for the detection of cross-channel frauds. In this paper, we come up with a GBT model which is capable of generating a prediction score for each transaction that is from internet banking, mobile

banking, ATM, and telephone banking. We mainly employ GBTs as they have outperformed the other boosting algorithms.

3 Methodology

In this section, we first give an overview of our study by defining its components. Then, we give details of our generating training set algorithm, feature engineering, encoding mechanism, and training algorithm, which is GBT.

3.1 Overview of the Study

Two main phases are included namely, offline algorithm training, and real-time fraud detection on incoming transactions. During the offline training phase, initially, a training set is composed by the mechanism, which is explained in Section 3.2. After building a training set, a GBT model is trained. In the real-time detection phase, transactions and their metadata are fed to apply pre-processing steps, i.e., data sanitization and handling the missing values. Afterwards, it is tried to match the receiver customer identity from the receiver account number if the customer identity is not known. Subsequently, the pipeline will extract demographic, economic, and behavioral features for the sender and receiver and merge them with the main data. The encoder will handle the categorical string values and convert them to the encoded integer to feed the trained GBT model. Finally, the model generates a score for each transaction to classify them as fraud or not. The real-time detection pipeline is illustrated in Figure 1.

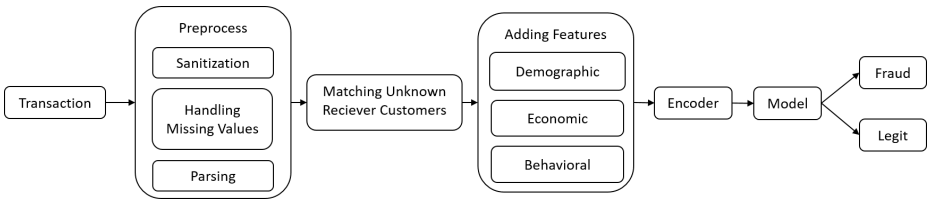

Fig. 1. A flow of a transaction in the study.

3.2 Generating Training Sets

We experimented two types of data balancing mechanisms, which are transaction-level and account-level. In transaction-level balancing, we pick all the fraudulent transactions within a time frame and randomly select the same number of transactions from the legitimate ones. The fraudulent and legitimate transactions are belong to the same customers who has been a sender or receiver of a fraudulent transaction at least once during the time frame of the training set. For the

account-level balancing, we employ the technique used by [21]. If at least one transaction is identified as a fraudulent transaction in a given time range, we get all transactions of the receiver and sender of the fraudulent transaction and label them as fraudulent accounts. If any customer has no fraudulent transaction, then we denote the sender and receiver customers as legitimate accounts. We equalize the number of fraudulent and legitimate accounts and create a training set from their transactions. Even though, it still results in a high-class imbalance, it preserves the patterns while decreasing a lot of initial imbalance [21].

For testing, we equalize fraudster and legitimate accounts with different ratios. Because higher sampling of legitimate accounts will increase the diversity of legitimate transactions in the training set, and as a result, it is believed to improve model performance on detecting non-fraud transactions.

The generation of a training set is automated in our study so that it can create new training sets and train new models by itself as time passes. Since the performance of a model decreases over time due to concept drift caused by a change in the behavior of account holders and/or fraudsters, this automation mechanism helps to adapt to new behaviors and prevents a dramatic decrease in prediction performance.

3.3 Feature Engineering

The metadata of any cross-channel transaction contains numerical, categorical, and textual information. All related data is embedded in a feature vector, and this vector is fed to the model to get the prediction for the transaction. The textual fields may contain more than one different value. These fields are parsed to extract features. For instance, a feature, which indicates belonging to a blacklist or white-list of a customer, is fed the model with the value "1" and "0". The parsed fields refer to "1" for blacklist and "0" for white-list. Another example is a field containing "AZ" which means "A" for a device like ATM, and Z for login duration, which indicates more than 29 minutes. The Model uses numerical fields as they are, but they are encoded as categorical fields. More detail for encoding is given in Section 3.4. In addition, demographic, economic, and behavioral features for the receiver and sender of a transaction are added. For instance, features like customers' age and financial age are some of the features which belong to demographic information. Credit scores, client types, and historical payment habits are also included to estimate the fraud probability of a transaction. They are called as behavioral features. After we determine the preliminary model features, the GBT model is trained with behavioral, economic, and demographic features. We also studied the feature importance.

3.4 Encoding

As mentioned above, transaction data has text and numeric values. The model can be fed directly by numeric values, although text values need to be encoded. As an encoding method, we preferred a label encoder. A dictionary, which contains a mapping for text and its corresponding encoded value, is created from the

history of each encoded features of the dataset. The encoded list is composed in ascending order of feature values. For the real-time encoding, the value -9999 is added as a response to all unseen values. This provides continuity for a real-time scoring mechanism.

4 Experiments

Our proposed approach is focused on detecting fraudulent transactions on cross-channel operations by using GBT models. Various decision tree algorithms are tested during the research period of our study. However, XGBoost [4] beats Random Forest, and also Decision Tree Classifier. The following section describes the experiments and gives detailed results of our study.

4.1 Experimental Setup

The experiments are performed on a real dataset of a private bank in Turkey. The dataset is divided into two mainframes which are the training set and the test set. The data of the test set includes 6 months of channel transactions from May 2021 to October 2021. During the test set period, out of more than 260 million transactions, around 4000 of them were a fraud. Month by month detailed explanation is given in Table 1.

Table 1. The legitimate and fraudulent transaction counts of test sets used in experiments.

Test Set Name	Trx CNT	Fraud CNT	Ratio
202105	40013025	644	0.0016%
202106	40867466	670	0.0016%
202107	42121304	690	0.0016%
202108	44368519	951	0.0021%
202109	46337966	826	0.0018%
202110	48144671	863	0.0018%

Table 2. The legitimate and fraudulent transaction counts of training sets used in experiments.

SetID	Trx CNT	Fraud CNT	Ratio	Sender Cust. CNT	Fraud Sender Cust. CNT
A	1634587	10394	0.6359%	20475	10237
B	13339287	10394	0.0779%	204619	10237
C	15532174	11357	0.0731%	223668	11191

For composing the training set, three different strategies are examined. Let's say, the first one is called A, which has an equal number of fraudulent and

legitimate accounts in the given time frame. The time range of set A is 16 months and contains data from December 2019 and up to May 2021. Another set, which is B, contains 19 times more legitimate transactions than fraudulent transactions for the same time range as A. The set C is generated with the same algorithm as the B but the time range is from March 2020 to August 2021. Comparison between set A and B indicates the importance of the undersampling method of data. The only difference between sets B and C is the time range difference. The detailed explanation of training sets is given in Table 2.

4.2 Evaluation Metrics

In most regression and classification problems, the error rate is the most commonly preferred success measurement metric to compare the performances of different algorithms. However, the error rate calculation may fail to provide the correct performance measure in such imbalanced datasets. In addition, the minimization of the error rate during learning may not improve Wilcoxon-Mann-Whitney statistics (or AUC scores). Therefore, AUC measures such as the area under the receiver operating characteristics curve (AUROC) is better-suited evaluation metric of binary classification with imbalanced data. Similar to the market leader as FICO, we used Account Detection Rate (ADR), Real-Time Value Detection Rate (RTVDR), and Non-Fraud Transaction Review Rate (NFTR) in our evaluation metrics [7]. The ADR indicates the rate of identified swindler accounts compared to total fraud accounts. The RTVDR refers to the rate of detected fraud amount compared to the total fraud amount. The NFTR is crucial because it directly expresses the rate of non-fraud transactions which are marked as fraud.

4.3 Results

This part includes results and charts from experiments related to generating training sets for extremely imbalanced and dynamically changing data. Furthermore, feature engineering and the combined performance of all studies are described below.

Training Set Ratio We need to create new dataset by combining fraudster and legitimate accounts with a predefined ratio, because if all the dataset is fed to a training algorithm without any adoption, the model is prone to predict all transactions to non-fraud. Our sampling method contains the ratio between legitimate and swindler accounts. First of all, the equal number of accounts from both sides, which is Set A described in Section 4.1 were selected. Set B contains 19 times more legitimate customers during the time interval, which is December 2019 up to May 2021. This optimized ratio was found according to various experiments. The detailed results for test sets are given in Figure 2. And models trained with those sets tested with the transactions which came to our banks channels during May, June, and July 2021.

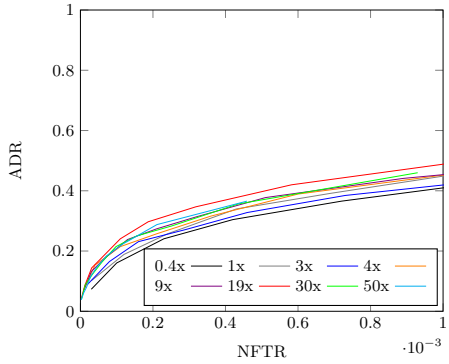

Fig. 2. Containing 19 times more transactions from legitimate customers dramatically increase detection rate.

Concept Drift As described above, customer behaviors also change over time. As a result of behavioral changes, the performance of the model will decrease slightly over time. In other words model becomes obsolete. To prevent aging of a training model as well as adopting the models performance to changing customer behaviors, training process should be renewed and model should be updated accordingly. In this way, the retrained model will compensate for those performance degradations. While the data between December 2019 and May 2021 is included in set B, the data between March 2020 and August 2021 is included in set C. Both models were tested with transactions of August and September 2021. The proof of concept drift is clearly shown in Figure 3 and 4. As the result shows clearly, if the older model used in test sets, will have less ADR in the same level of NFTR. The decrease in ADR is dramatically seen in the upcoming months September 2021.

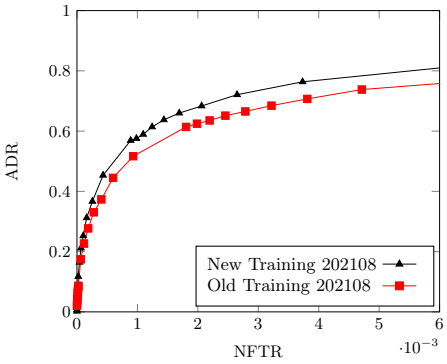

Fig. 3. A detailed result to demonstrate the Concept Drift of August 2021.

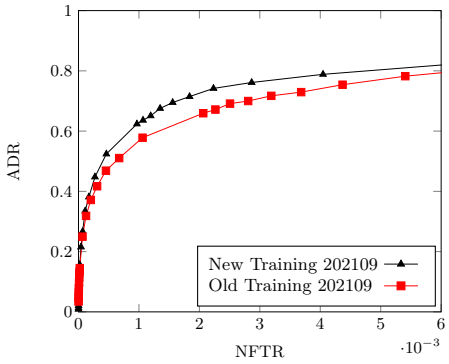

Fig. 4. A detailed result to demonstrate the Concept Drift of September 2021.

Demographic, Economic And Behavioral Features The benefit of demographic, economic, and behavioral features are helpful to characterize customer habits. As an economic feature, we added customers' previous credit scores from the credit bureau of Turkey. Historical credit scores are included as new features up to 12 months from the transaction date. This is a good indicator of senders' and receivers' personal economic situations. The detailed results and proof can be found in Figure 5. Added features are explained in Table 3.

Table 3. Demographic, and Economic features.

Feature Name	Description	Category
Account Branch	Branch Location of Account	Demographic
Client Type	Client Type (Commercial/Individual)	Demographic
Customer Age	Customer's Age	Demographic
Financial Age	Customer's Financial Age	Economic
Max & Min Credit Scores	Monthly historical max & min credit scores	Economic

Feature Importance To optimize features of the model we analyzed integrated feature importance methods of XGBoost. Highly ranked features are chosen according to [4]. This methods contains different approaches for feature importance analysis including cover, gain, total cover, total gain, and weight. All of them are blended and created a finalized optimized feature list for the model. Adding demographic features to the raw attributes of transaction gives us a data set with 172 features. We sorted them by descending order for each feature importance method, and take the union of the first 70 features in each list. The resulted feature list contains 96 features. The performance comparison between models can be inspected in Figure 6.

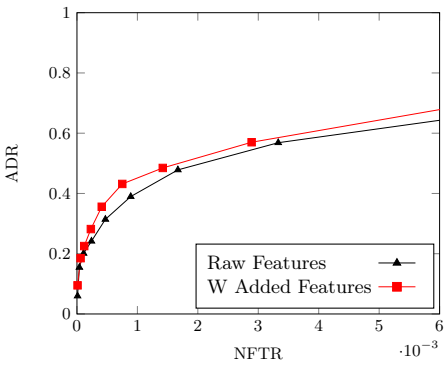

Fig. 5. An in-depth analysis of demographic, economic and behavioral features.

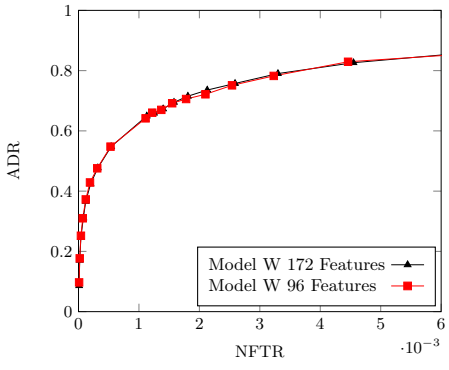

Fig. 6. Exactly the same test performance is investigated both 96 and 172 features for May 2021.

To sum up, when we combine all techniques above, the overall fraud loss amount detection performance of the model with a low false-positive rate is shown in Figure 7.

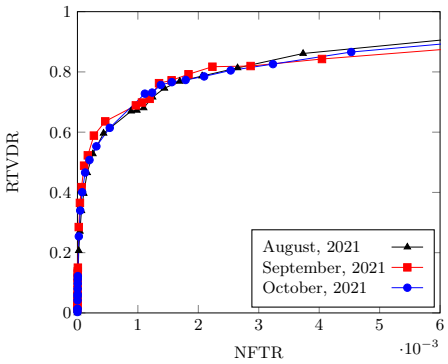

Fig. 7. The model's overall fraud amount detection performance for August, September, and October 2021

5 Conclusion

As a consequence of the emerging technologies, the variety of transaction channels has increased, but only 26% of financial institutions are tracking the frauds in their transaction channels [16]. Our proposed model tracks and scores the monetary transactions to indicate the possibility of fraudulent transactions from the cross channels in real-time. An AI-powered software solution is provided for the outcome of rarely occurred fraudulent transactions, which properly handles the imbalanced class and concept drift problems. Our model detects cross-channel frauds quite accurately over real banking datasets. It can catch almost 75% of total fraud loss with a low false-positive ratio and overcome aging with continuous training. The proposed learning strategy and trained algorithm will be integrated to live system, and our solution will become a good candidate compared to conventional rule-based solutions which require constant maintenance and analysis.

References

1. Bayram, B., Köroğlu, B., Gönen, M.: Improving fraud detection and concept drift adaptation in credit card transactions using incremental gradient boosting trees. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 545–550. IEEE (2020)
2. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C.: Data mining for credit card fraud: A comparative study. Decision Support Systems **50**(3), 602–613 (2011)

3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
5. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection and concept-drift adaptation with delayed supervised information. In: 2015 international joint conference on Neural networks (IJCNN). pp. 1–8. IEEE (2015)
6. Dheepa, V., Dhanapal, R.: Behavior based credit card fraud detection using support vector machines. ICTACT Journal on Soft computing **2**(4), 391–397 (2012)
7. FICO: Reducing fraud losses with enhanced cnp models on the fico® falcon® platform (2018)
8. Gee, J., Button, M.: The financial cost of fraud 2019: The latest data from around the world (2019)
9. Gee, J., Button, M.: The financial cost of fraud 2021: The latest data from around the world (2021)
10. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.E., He-Guelton, L., Caelen, O.: Sequence classification for credit-card fraud detection. Expert Systems with Applications **100**, 234–245 (2018)
11. Malekian, D., Hashemi, M.R.: An adaptive profile based fraud detection framework for handling concept drift. In: 2013 10th International ISC Conference on Information Security and Cryptology (ISCISC). pp. 1–6. IEEE (2013)
12. Molloy, I., Chari, S., Finkler, U., Wiggerman, M., Jonker, C., Habeck, T., Park, Y., Jordens, F., van Schaik, R.: Graph analytics for real-time scoring of cross-channel transactional fraud. In: International Conference on Financial Cryptography and Data Security. pp. 22–40. Springer (2016)
13. Patel, Y., et al.: Cross channel fraud detection framework in financial services using recurrent neural networks. Ph.D. thesis, London Metropolitan University (2019)
14. Philip, K., Chan, S.: Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In: Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining. pp. 164–168 (1998)
15. Sinayobye, J.O., Kiwanuka, F., Kaawaase Kyanda, S.: A state-of-the-art review of machine learning techniques for fraud detection research. In: 2018 IEEE/ACM Symposium on Software Engineering in Africa (SEiA). pp. 11–19 (2018)
16. Urban, M.: New survey reveals top fraud threats and vulnerabilities. FICO (Jan 11, 2011 [Online]), https://www.fico.com/blogs/new-survey-reveals-top-fraud-threats-and-vulnerabilities
17. Urban, M.: Why managing cross-channel fraud is a must. FICO (Feb 22, 2011 [Online]), https://www.fico.com/blogs/why-managing-cross-channel-fraud-must
18. Worobec, K.: 2021 fraud - the facts 2021 (2021)
19. Worobec, K.: 2021 half year fraud update (2021)
20. Yen, S.J., Lee, Y.S.: Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: Intelligent Control and Automation, pp. 731–740. Springer (2006)
21. Yeşilkanat, A., Bayram, B., Köroğlu, B., Arslan, S.: An adaptive approach on credit card fraud detection using transaction aggregation and word embeddings. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 3–14. Springer (2020)