Journal of Theoretical and Applied Information Technology
15th July 2023. Vol.101. No 13
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPLEMENTATION OF MACHINE LEARNING TECHNIQUES TO DETECT FRAUDULENT CREDIT CARD TRANSACTIONS ON A DESIGNED DATASET

NAZERKE BAISHOLAN[1], MUSSA TURDALYULY[2], SERGIY GNATYUK[3], KARLYGASH BAISHOLANOVA[1], KAZILA KUBAYEV[1]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan

[2] Satbayev University, Almaty, Kazakhstan

[3] National Aviation University, Kyiv, Ukraine

E-mail: [1] baisholan@gmail.com, [2] m.turdalyuly@gmail.com, [3] s.gnatyuk@nau.edu.ua,
[4] baisholanova.k@gmail.com, [5] kubayev@gmail.com

ABSTRACT

The rise in technology, particularly the increase in online shopping, has made it easier for cybercriminals to obtain and exploit stolen payment card information. Traditional fraud detection systems are finding it increasingly challenging to keep up with the rapid pace of technological advancement, leading to a surge in payment card fraud. Hence, it is essential for companies to continually update their fraud detection methods to keep up with the latest tactics employed by fraudsters. Machine learning algorithms have the ability to analyze large datasets and quickly identify anomalies or deviations from normal behaviour, making them a highly effective tool for payment card fraud detection. By detecting fraud early, organizations can minimize their financial losses and prevent further damage.

In this study, we generated a credit card fraud dataset that comprises three types of fraud cases. The dataset is imbalanced, with a ratio of fraudulent transactions at 0.004, making it close to real-world data. To handle the imbalance in the dataset related to credit card fraud detection, we employed popular machine learning models such as Random Forest, Decision Tree, Logistic Regression, and XGBoost. The results showed that XGBoost and Random Forest outperformed the other models on both the training and test sets. However, the Decision Tree algorithm with unlimited depth had the highest average accuracy on the training set and the lowest average accuracy on the test set, indicating that this algorithm should be avoided due to overfitting.

In conclusion, our study highlights the significance of using machine learning algorithms for payment card fraud detection. The results demonstrate that XGBoost and Random Forest are the most effective models for detecting credit card fraud in imbalanced datasets. By employing these models, organizations can improve their fraud detection capabilities and minimize the financial impact of payment card fraud.

Keywords: *Fraud Detection, Anomaly Detection, Transaction Fraud Dataset, Imbalanced Dataset, Random Forest, Decision Tree, Logistic Regression.*

1. INTRODUCTION

Payment card fraud has become a significant issue for consumers and companies, resulting in billions of dollars in losses annually [1]. In 2020, fraudulent losses in the United States amounted to $10.24 billion, a rise from the $9.62 billion reported in 2019. The total amount of fraud losses for all countries except the United States in 2020 was $18.34 billion. The global figure of $28.58 billion for card payment fraud losses in 2020 is likely an underestimate as related costs cannot be precisely quantified. The expenses incurred by issuers, merchants, and acquirers for activities such as investigating fraud, managing call centres, and maintaining operations tend to increase every year. These additional costs, which are not included in the $28.58 billion figure, mean that the actual losses experienced by the card industry are higher than the reported amount. Global fraud losses in the payment card industry are expected to reach $49.32 billion in 2030 when the total volume on all payment cards is anticipated to reach $79.140 trillion [2].

As technology advances, the methods used by fraudsters become increasingly sophisticated, making it challenging for traditional fraud detection systems to keep up. Hence, there is a pressing need for more effective fraud detection methods that can quickly identify and prevent fraudulent activities.

To address the challenge posed by payment card fraud, organizations and financial institutions are increasingly adopting machine learning techniques to enhance the detection and prevention of fraudulent activities. By analyzing historical data, machine learning algorithms can learn to recognize and anticipate fraudulent behaviour, leading to more accurate and faster fraud detection compared to traditional methods [3]. Additionally, these algorithms can identify intricate fraud patterns and adjust their models continuously to reflect changes in the data, enabling real-time protection against fraud. Consequently, the use of machine learning for financial data analysis is gaining popularity and is projected to expand in the future [4]. In this study, we focus on credit card fraud detection using machine learning algorithms. We created a credit card fraud dataset that contains three types of fraud cases and used popular machine learning models such as Random Forest, Decision Tree, Logistic Regression, and XGBoost to detect fraud in the dataset. We evaluated the performance of these models and compared their results to determine which one is the most effective in detecting credit card fraud.

The rest of the paper is organized as follows. Section II provides a literature review on credit card fraud detection and the use of machine learning algorithms. Section III describes the methodology used in this study, including the dataset creation process and the machine learning models employed. Section IV presents the experimental results and compares the performance of the different models. Finally, Section V concludes the paper and provides suggestions for future research.

2. LITERATURE REVIEW

Manuscripts must Machine learning is a field of study that involves training algorithms on data, enabling software applications to carry out tasks and make predictions. This process allows the software to find information and intuitively provide outcomes [5]. ML involves teaching computers to perform functions by training them on data, enabling them to make predictions find information, and complete tasks without being explicitly programmed. It is a crucial aspect of AI that allows computers to learn and adapt to a new report [6, 7]. Machine learning is closely related to other fields, such as statistics, data mining, and pattern recognition, which work together to enable software applications to make predictions and perform tasks based on learned patterns from data.

Detection of fraudulent transactions is challenging due to the small number of fraudulent transactions relative to the total number of transactions. It can cause difficulties in detecting fraud in real-time or even after it has occurred [8].

Ileberi et al. [9] used the genetic algorithm (GA) to identify the most relevant features and combined it with several machine learning classifiers such as Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), and Artificial Neural Network (ANN). The results indicated that the combination of GA and Random Forest (GA-RF) had the best performance, with an accuracy rate of 99.98%. Meanwhile, the GA-DT combination also achieved good results, with an accuracy of 99.92%.

The three machine learning algorithms, Logistic Regression, Naive Bayes, and K-Nearest Neighbor, were compared and analyzed for their performance in fraud prediction. The evaluation was based on accuracy, sensitivity, specificity, precision, F-measure, and AUC. Logistic Regression was the best algorithm for fraud prediction compared to Naive Bayes and K-Nearest Neighbor. The results showed an improvement when under-sampling techniques were applied to the data before building the prediction model [10].

Chang et al. [11] attempted to identify an efficient way to identify fraudulent activities in an imbalanced dataset. They used four supervised machine learning algorithms, performed feature engineering and analysis, and evaluated their performance. The results revealed that utilizing the NearMiss undersampling method to combine the models improved their accuracy compared to the determining metrics AUROC and precision.

Another study [12] compares the effectiveness of machine learning and deep learning algorithms in detecting credit card fraud. The European credit card benchmark dataset was used for empirical analysis. The study starts by applying a machine learning algorithm, and then three convolutional neural networks (CNN) based models are used to enhance the fraud detection performance. The results showed that the proposed model outperformed the existing machine learning and deep learning algorithms regarding the accuracy, precision, f1-score, and AUC. Efforts were also made to balance the data and minimize the false negative rate. The results suggest that the proposed approach can be applied effectively for real-world credit card fraud detection.

The study [13] presents a machine learning approach that utilizes a Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) with an attention mechanism to improve credit card fraud detection performance. This method is compared to traditional classifiers such as Naive Bayes, Support Vector Machine (SVM), and Artificial Neural Network (ANN). The results indicate that the LSTM-RNN with the attention mechanism produces a high level of accuracy.

The authors in the paper [14] propose a fraud detection method that combines machine learning (ML) classifiers with a voting ensemble learning approach. They use dimensionality reduction techniques such as PCA, LDA, Autoencoder, and SMOTE to increase fraud detection accuracy by utilizing the strengths of multiple ML classifiers. The results show that using PCA's voting ensemble learning technique outperforms other ML classifiers, achieving 100.0% accuracy, 97.3% precision, 73.5% recall, and 83.7% f1-score.

Compared with other studies [15-19], we used Area Under the Receiver Operating Characteristic Curve (AUC ROC), Average precision (AP) and Card precision (CP) metrics.

While the Matthews Correlation Coefficient (MCC) and AUC ROC take into account specificity and sensitivity, the AP and CP provide a more comprehensive and balanced evaluation by considering all possible decision thresholds. The AP and CP measures are consistently effective in assessing ranking quality, unlike other metrics, and specificity, which rely on specific decision thresholds and are less informative for evaluating detection algorithms.

Average precision and Card precision are better suited for fraud detection than other metrics [20].

3. RESEARCH METHODOLOGY

The proposed methodology uses the simulated data generated for our research and will rely on a supervised learning approach. The development of our fraud detection system will involve three primary stages:

1) Establishing the training and testing datasets. A portion of the available transactions will be designated as the training dataset, which will be used to build the prediction model. The rest of the transactions will serve as the testing dataset and will be used to evaluate the prediction model's performance.

2) Building the prediction model. Using the training dataset, we will create a model that can accurately predict if a transaction is genuine or fraudulent. For this task, we will utilize the sklearn library in Python, which offers convenient functions to train prediction models.

3) Evaluating the prediction model's performance. The accuracy of the prediction model will be evaluated using the test dataset, which contains new data.

3.1 Dataset Description

The created dataset consists of 1,746,520 transactions from 01.01.2022 to 31.03.2022. It includes unique 10,000 client IDs and 20,000 POS terminal IDs. All transactions were generated based on client properties, such as frequency, spending amount, and available terminals. We linked each client profile to specific POS terminals within a 5 km radius of their last location. Transactions were classified as legitimate or fraudulent using three different fraud cases:

Case 1: Any POS terminal transactions with an amount greater than 300 are considered fraud. This case provides a clear fraud pattern that should be easily detected by a basic fraud detector. This serves as a way to test the implementation of the fraud detection methodology.

Case 2: This case represents criminal use of the POS terminals. A list of three POS terminals was randomly generated on a daily basis. All transactions at these POS terminals were flagged as fraudulent within 20 days. To detect this type of fraud, we added features that monitor the number of fraudulent operations on the POS terminal.

Case 3: This case imitates a card-not-present (CNP) fraudulent transaction where the customer's credentials have been leaked. Each day a list of 5 clients is randomly selected. Over the next ten days, the amounts of 1/4 of their transactions are multiplied by four and labeled as fraud. To detect this fraud, we added features that track a client's spending behavior.

The data set for the proposed study is highly unbalanced, with a disproportionate number of data points belonging to the majority class (non-fraud) compared to the minority class (fraud).

Of the 1,746,520 transactions, only 8,613 were found to be fraudulent. The rate of fraudulent transactions is 0.4%. This underscores the importance of correctly identifying fraud data points, which are a minority.

3.2 Proposed Classification Methods

The proposed approach involves using four ML classifiers for fraud detection. These classifiers are Logistic Regression, Random Forest, and eXtreme Gradient Boosting. These algorithms were selected based on their common use in literature for similar problems and their proven effectiveness in various applications. The following section will briefly explain the basic mechanics of each of these algorithms.

3.2.1 Logistic Regression

Logistic regression (LR) is a commonly used machine learning algorithm due to its brief analysis and straightforward processing of class features. In LR, the goal is to model the relationship between the predictor variables and the likelihood of a binary outcome, such as fraud or non-fraud. It can connect different factors, particularly those with significant influence, and adapt to various aspects based on the predictor variables and the outcome.

LR employs values greater than 1 and less than 0 to deal with anomalies in the dataset. Additionally, it is not limited to only classifying and predicting binary outcomes but also multinomial outcomes and uses the sigmoid function to calculate the values of the parameters' coefficients [21].

3.2.2 Random Forest

Random Forest (RF) is a popular machine learning algorithm used in fraud detection. It is an ensemble method that combines multiple decision trees to make a prediction. In the RF algorithm, a large number of trees are trained on bootstrapped samples of the training data, and the forecast is made by aggregating the results from all the trees [22]. This helps reduce the variance in the model and improves its generalization ability. RF can handle high dimensional data and complex relationships between features, making it a suitable choice for fraud detection tasks [23]. The algorithm can also identify important features that contribute to the prediction and is relatively robust to overfitting compared to other machine learning models.

3.2.3 Decision Tree

A Decision Tree (DT) is a tree-based model used for classification and regression [24]. The model represents data as a tree-like structure where each internal node represents a feature, and the connections represent the outcomes of the feature. Leaf nodes represent class labels. The tree is constructed by dividing the dataset into subsets based on the results of a feature value test. This process is repeated until the subsets have the same result as the target attribute or until further splits do not improve predictions. DTs are suitable for exploratory knowledge discovery applications because they do not require domain knowledge or parameter configuration and can handle high-dimensional data. As a result, DT classifiers are often accurate in their classifications, and the induction of categorization information through DTs is a typical inductive approach.

3.2.4 eXtreme Gradient Boosting

Random eXtreme Gradient Boosting (XGBoost) is a machine learning algorithm used for various applications, including fraud detection [25]. It is an optimized version of Gradient Boosting decision trees, which is an ensemble learning algorithm that creates a set of decision trees from the training data and combines them to make predictions.

In the case of fraud detection, XGBoost can be trained on a dataset of transactions to identify patterns and relationships between various features that are indicative of fraudulent behaviour. XGBoost also can handle large datasets and complex relationships between features, making it a suitable choice for fraud detection applications [26]. Additionally, XGBoost can handle missing values and noisy data, which is a common issue in fraud detection datasets.

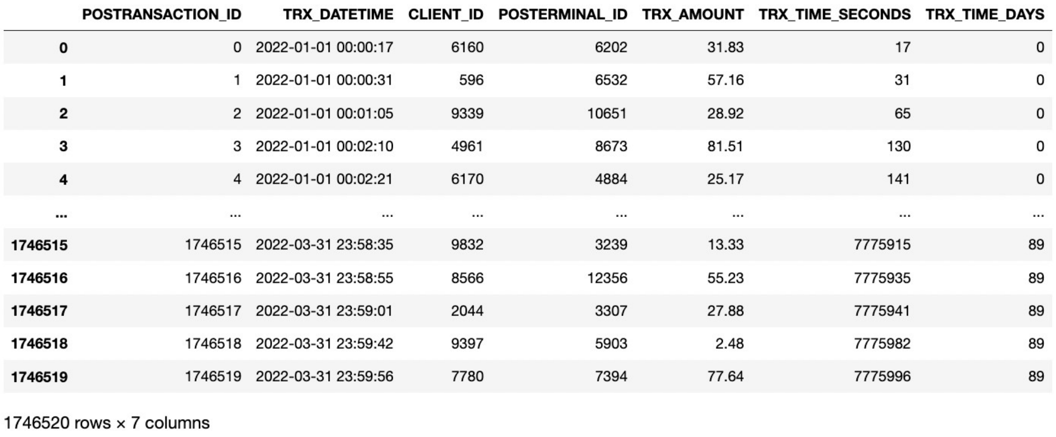
4. RESULTS

The research was conducted in Apple M1 Pro, GPU, 16 Cores. Both training and testing algorithms are implemented in Jupyter Notebook 6.4.12. We used Python as the programming language due to its ease of use, interpreted, high-level, object-oriented, scripting nature, and its popularity in the field of machine learning. To analyze and visualize our data, we used some of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Python libraries and packages, such as Numpy, Pandas, Sklearn, Matplotlib, and Seaborn.

Using a "generate_dataset" function, we created our own dataset with the following features: 10,000 clients, 20,000 POS terminals, and 90 days of transactions from 01.01.2022 to 31.03.2022. The radius is set to 5km. It took us less than 3 minutes to generate 1,746,520 POS transactions (Fig. 1). However, this number is lower compared to the actual fraud detection systems in the real world, where millions of transactions might need to be processed on a daily basis. But this number will suffice for the purpose of our study, particularly to maintain reasonable execution times. The dataset was stored in the format of the time series databases [27]. Compared to other studies, our dataset includes transactions over a period of 90 days, therefore the number of total transactions was increased and the ratio between of fraudulent and genuine transactions was changed



	POSTRANSACTION_ID	TRX_DATETIME	CLIENT_ID	POSTERMINAL_ID	TRX_AMOUNT	TRX_TIME_SECONDS	TRX_TIME_DAYS
0	0	2022-01-01 00:00:17	6160	6202	31.83	17	0
1	1	2022-01-01 00:00:31	596	6532	57.16	31	0
2	2	2022-01-01 00:01:05	9339	10651	28.92	65	0
3	3	2022-01-01 00:02:10	4961	8673	81.51	130	0
4	4	2022-01-01 00:02:21	6170	4884	25.17	141	0
...	...	...	...	...	...	...	...
1746515	1746515	2022-03-31 23:58:35	9832	3239	13.33	7775915	89
1746516	1746516	2022-03-31 23:58:55	8566	12356	55.23	7775935	89
1746517	1746517	2022-03-31 23:59:01	2044	3307	27.88	7775941	89
1746518	1746518	2022-03-31 23:59:42	9397	5903	2.48	7775982	89
1746519	1746519	2022-03-31 23:59:56	7780	7394	77.64	7775996	89

1746520 rows × 7 columns

Figure 1: Generated 1,746,520 POS transactions

The distribution of transaction amounts has a majority of its values for small amounts (Fig. 2). The distribution of transaction times follows a Gaussian distribution centred around noon on a daily basis (Fig. 3).

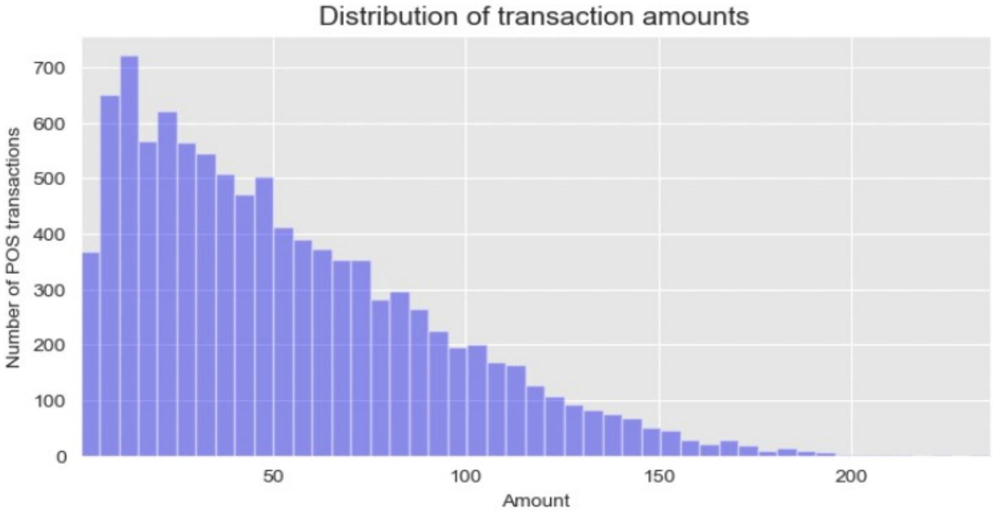


Figure 2: Distribution graph of transaction amounts

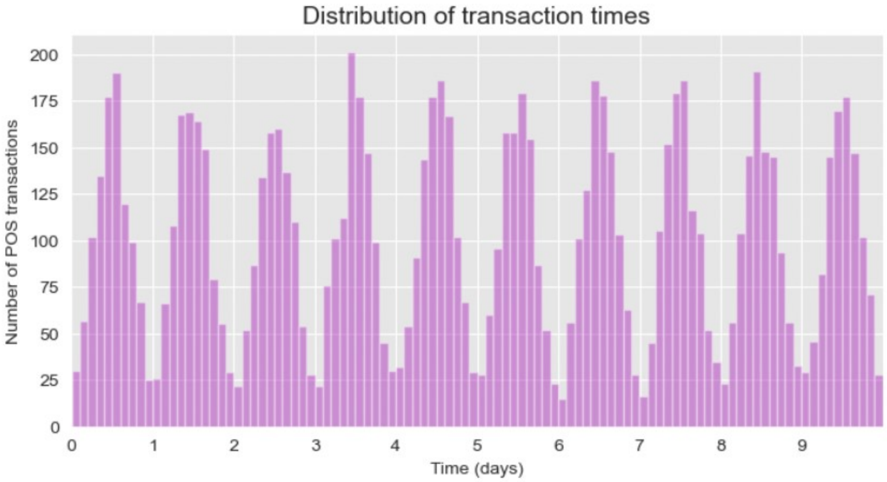


Figure 3: Distribution graph of transaction times

In our study, the transactions from 21.02.2022 to 15.02.2022 were used as the training set, while the transactions from 8.03.2022 to 14.03.2022 were used as the test set. The goal of using the training set is to train a prediction model, while the purpose of using the test set is to evaluate the prediction model's performance on new data. This one week of transaction data is sufficient to train the first prediction model and assess its performance. Further evaluation will be carried out by using more considerable periods for training and testing to see how it affects the performance results. “Fig. 4” shows that the daily number of transactions is the same during the training and testing periods. The average number of frauds during the training period was about 85 per day, while during the testing period, there were jumps of up to 100 per day.

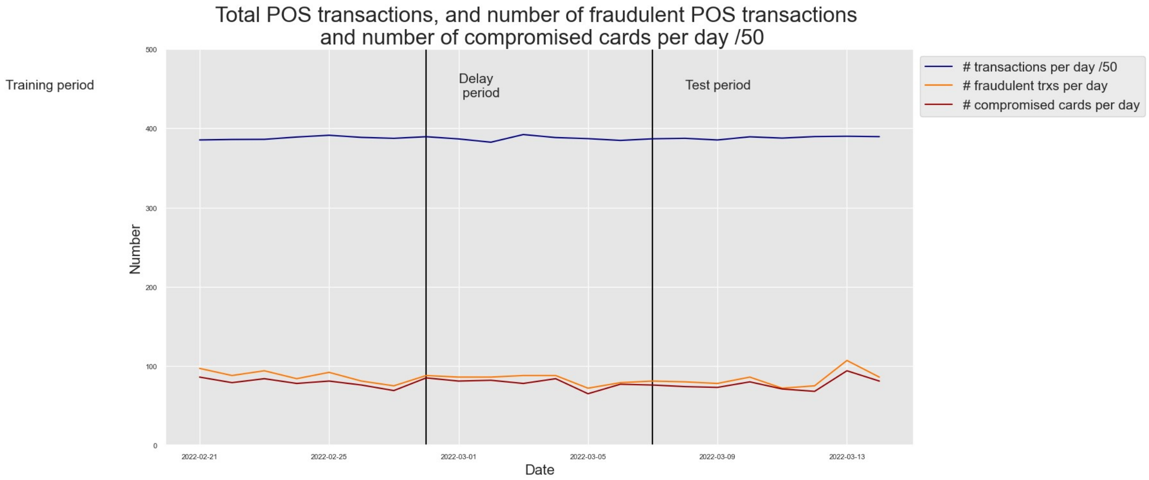


Figure 4: The Total POS Transactions, Number Of Fraudulent POS Transactions, And Number Of Compromised Cards Per Day /50

While extracting from the transaction dataset for the training set, we found 135,892 transactions, including 600 that are fraudulent. The test set contains 126,251 transactions, of which 506 were fraudulent. It means that the ratio of fraudulent transactions is 0.004.

Next, we will train some standard classifiers. We will start with the Decision Tree model. “Fig.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5" shows an example of training a decision tree algorithm with a maximum depth of 2.

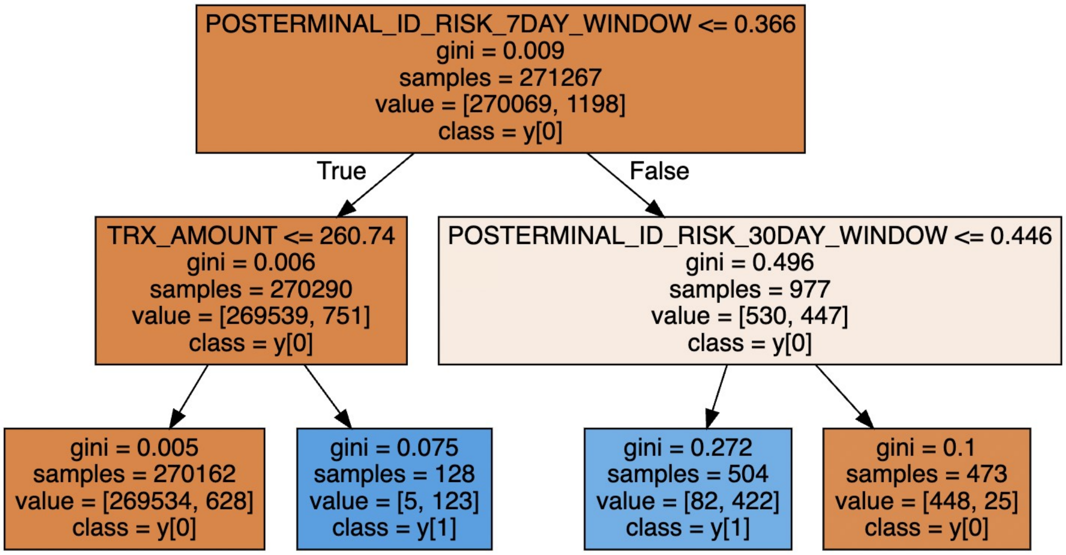
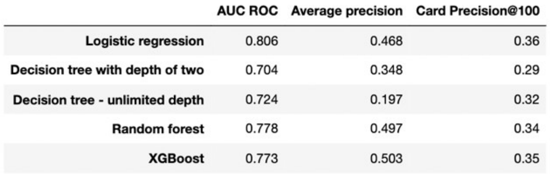


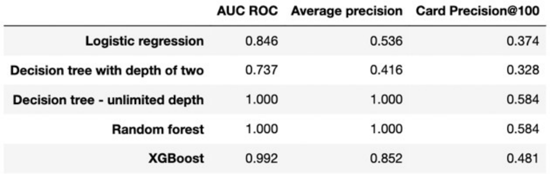
Figure 5: A Decision Tree Model That Allows To Establish The Probability Of Fraud

After training four other prediction models (DT with unlimited depth, LR, RF, and XGBoost), we finally evaluated the prediction efficiency of our chosen models on the test (Fig. 6) and training (Fig. 7) set, as well as their execution time (Fig.8).



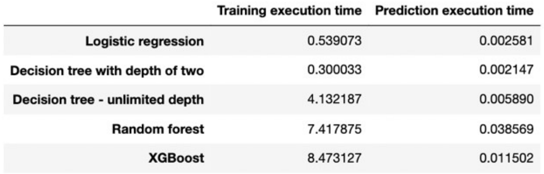
	AUC ROC	Average precision	Card Precision@100
Logistic regression	0.806	0.468	0.36
Decision tree with depth of two	0.704	0.348	0.29
Decision tree - unlimited depth	0.724	0.197	0.32
Random forest	0.778	0.497	0.34
XGBoost	0.773	0.503	0.35

Figure 6: Performances On The Test Set



	AUC ROC	Average precision	Card Precision@100
Logistic regression	0.846	0.536	0.374
Decision tree with depth of two	0.737	0.416	0.328
Decision tree - unlimited depth	1.000	1.000	0.584
Random forest	1.000	1.000	0.584
XGBoost	0.992	0.852	0.481

Figure 7: Performances On The Training Set



	Training execution time	Prediction execution time
Logistic regression	0.539073	0.002581
Decision tree with depth of two	0.300033	0.002147
Decision tree - unlimited depth	4.132187	0.005890
Random forest	7.417875	0.038569
XGBoost	8.473127	0.011502

Figure 8: Execution Times

5. CONCLUSION

The predictive models have effectively captured fraud patterns from the training data. The XGBoost shows better stable results (in terms of Average Precision) compared to other models.

Comparing the classifiers' performance may vary depending on the metric used. For instance, a DT with a depth of two may have a lower AUC ROC than a DT with unlimited depth but a higher average precision.

Some classifiers (DT with unlimited depth and RF) exhibit perfect performance on the training set (AUC ROC and Average Precision of 1) but lower results on the test set. The DT with unlimited depth, for example, has the lowest Average Precision on the test set and is an example of overfitting, which should be avoided and will be addressed in our future research.

The training time for ensembles of models (RF and XGBoost) is significantly longer compared to single models (LR and DTs), which is to be expected.

A limitation of research in this area related to the quality of the datasets. A limited number of datasets are currently available on Kaggle [28]. Due to privacy and security concerns, the dataset with credit card transactions may not fully represent the complexities and variations of real-life fraud scenarios.

The limitations of having a limited dataset and restricted access to real-life fraud scenarios pose challenges in accurately modeling and assessing the effectiveness of fraud detection algorithms.

To overcome these limitations, in further research we will explore collaborations with financial organizations that can provide access to more extensive and diverse datasets, enabling a more accurate evaluation of fraud detection algorithms.

Extensive training and a proposal of our prediction model will be discussed in a later study.

REFERENCES:

[1] KPMG. (2022) *A triple threat across the Americas: KPMG 2022 Fraud Outlook*. Accessed: Feb. 02, 2023. [Online]. Available: https://kpmg.com/xx/en/home/insights/2022/01/kpmg-fraud-outlook-survey.html

[2] HSN Consultants, Inc. (Dec. 7, 2021). *The Nilson Report 2021, Issue 1209*. [Online]. Available: https://nilsonreport.com

[3] A.E. Abdou, K. Wael, R. Ismail and S. Abdel-Badeeh. "Machine Learning Techniques for Credit Card Fraud Detection," *Future Computing and Informatics Journal*, vol. 4: Iss. 2, Article 5, 2019.

[4] F. Rundo, F. Trenta, A.L. di Stallo and S. Battiato. "Machine learning for quantitative finance applications: A survey", *Applied Sciences*, vol. 9(24): 5574, 2019.

[5] Y. Abakarim, M. Lahby and A. Attioui. ''An efficient real time model for credit card fraud detection based on deep learning,'' in *Proc. 12th Int. Conf. Intell. Systems: Theories Appl.*, pp. 1–7, Oct. 2018.

[6] Priyadharshini. (2023). Machine Learning: What it is and Why it Matters. Accessed: Feb. 04, 2023. [Online]. Available: https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article

[7] L. Tagliaferri. (2022) *An Introduction to Machine Learning*. Accessed: Feb. 04, 2023. [Online]. Available: https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning

[8] P. Sharma, S. Banerjee, D. Tiwari and J.C. Patni. "Machine learning model for credit card fraud detection - A comparative analysis", *International Arab Journal of Information Technology*, vol. 18 (6), pp. 789 – 796, 2021.

[9] E. Ileberi, Y. Sun and Z. Wang. "A machine learning based credit card fraud detection using the GA algorithm for feature selection", *Journal of Big Data*, vol. 9 (1), art. no. 24, 2022.

[10] F. Itoo, S. Meenakshi and Singh. "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection", *International Journal of Information Technology* (Singapore), vol. 13(4), pp. 1503–1511, 2021.

[11] V. Chang, L.M. Thao Doan, A. Di Stefano, Zh. Sun and G. Fortino. "Digital payment fraud detection methods in digital ages and Industry 4.0", *Computers and Electrical Engineering*, vol. 100, 107734, 2022.

[12] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed. "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," in IEEE Access, vol. 10, pp. 39700-39715, 2022.

[13] J. Femila Roseline, G.B.S.R. Naidu, V. Samuthira Pandi, S. Alamelu alias Rajasree and Dr.N. Mageswari. "Autonomous credit card fraud detection using machine learning approach", *Computers and Electrical Engineering,* vol. 102, 108132, 2022.

[14] M.R. Baker, Z.N. Mahmood and E.H. Shaker. "Ensemble learning with supervised machine learning models to predict credit card fraud transactions", *Revue d'Intelligence Artificielle*, vol. 36 (4), pp. 509-518, 2022.

[15] T.E. Mathew. "An Ensemble Machine Learning Model for Classification of Credit Card Fradulent Transactions", *Journal of Theoretical and Applied Information Technology,* vol.101(9), pp. 3530 – 3546, 2023.

[16] M. Moreira, C. Junior, D. Silva, M. Junior, I. Costa, C. Gomes, M. Santos, "Exploratory analysis and implementation of machine learning techniques for predictive assessment

of fraud in banking systems", *Procedia Computer Science*, vol. 214, pp. 117-124, 2022.

[17] J.K. Afriyie, K. Tawiah, W.A. Pels, S. Addai-Henne, H.A. Dwamena, E.O. Owiredu, S.A. Ayeh, J. Eshun,
"A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions", *Decision Analytics Journal*, vol. 6, 100163, 2023.

[18] J. Roseline, G. Naidu, V. Pandi, S. Rajasree, N. Mageswari, "Autonomous credit card fraud detection using machine learning approach″, *Computers and Electrical Engineering*, vol. 102, 108132, 2022.

[19] H. Fanai and H. Abbasimehr, "A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection",
Expert Systems with Applications, vol. 217, 119562, 2023.

[20] G. Fan and M. Zhu, "Detection of rare items with TARGET", *Statistics and Its Interface,* vol. 4, pp. 11–17, 2011.

[21] Y. Jain, T. Namrata, D. Shripriya and S. Jain. "A comparative analysis of various credit card fraud detection techniques," *Int. J. Recent Technol. Eng*., vol. 7, 402–403, 2019.

[22] N. Shirodkar, P. Mandrekar, R.S. Mandrekar, R. Sakhalkar, K.C. Kumar and S. Aswale. "Credit card fraud detection techniques–A survey", *In Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering* (ic-ETITE), Shiroda, India, pp. 1–7, 13–15 May 2020.

[23] K. Seeja and M. Zareapoor. "Fraudminer: A novel credit card fraud detection model based on frequent itemset mining", *Scientific World Journal,* pp. 1–10, 2014.

[24] M. Xu, P. Watanachaturaporn, P.K. Varshney and M.K. Arora. "Decision tree regression for soft classification of remote sensing data", *Remote Sensing of Environment*, vol. 97(3), pp. 322–336, 2005.

[25] N. Dhieb, H. Ghazzai, H. Besbes and Y. Massoud. "A secure AI-driven architecture for automated insurance systems: Fraud detection and risk measurement". *IEEE Access*, vol. 8, pp. 58546–58558, 2020.

[26] C.V. Priscilla and D.P. Prabha. "Influence of optimizing xgboost to handle class imbalance in credit card fraud detection". In: *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology*, ICSSIT, pp. 1309–1315, 2020.

[27] S. Gnatyuk, R. Berdibayev, I. Azarov, N. Baisholan and I. Lozova. "Modern Types of Databases for SIEM System Development", *CEUR Workshop Proceedings*, vol. 3187, pp. 127-138, 2021.

[28] Kaggle, Accessed: March 24, 2023. [Online].
Available: https://www.kaggle.com/