Credit Card Fraud Using Adversarial Attacks

Hafya Ullah, Aysha Thahsin Zahir Ismail, Lakshmi Babu Saheer(✉), and Mahdi Maktabdar Oghaz

Anglia Ruskin University, Cambridge, UK
{az303,hu142}@student.aru.ac.uk,
{lakshmi.babu-saheer,mahdi.maktabdar}@aru.ac.uk

Abstract. Banks lose billions to fraudulent activities every year, affecting their revenue and customers. The most common type of financial fraud is Credit Card Fraud. The key challenge in designing a model for credit card fraud detection is its maintenance. It is pivotal to note that fraudsters are constantly improving their tactics to bypass fraud detection checks. Several fraud detection methods for identifying fraudulent credit card transactions have been developed. However, in order to further improve on the existing strategies, this paper investigates the domain of adversarial attacks for credit card fraud. The goal of this work is to show that adversarial attacks can be implemented on tabular data and investigate if machine learning approaches can get affected by such attacks. We evaluate the performance of adversarial samples generated by the LowProfool algorithm in deceiving the classifier.

Keywords: Adversarial attacks on tabular data · Financial fraud detection · Machine learning · Lowprofool algorithm

1 Introduction

The 2021 Nilson report[1] stated that financial industries will experience fraud loss amounting to 408.50 billion dollars over the next decade. In most situations, credit card data is leaked due to phishing of financial websites where the user is unaware of the data leak. Machine learning classification algorithms are considered to be state-of-the-art techniques to identify legitimate and fraudulent transactions with greater precision and accuracy. The user spending pattern is obtained from the available transaction data that can be analyzed by machine learning classification algorithms to identify actual transactions made by the customer [1]. It is also critical to consider that the fraudsters are persistent and consistently upgrade their techniques and sophisticated activities with the aim of bypassing the fraud detection systems. This is widely referred to as concept drift [2]. This study aims to raise awareness of fraud detection in the

[1] https://nilsonreport.com/mention/1515/1link/.

M. Bramer and F. Stahl (Eds.): SGAI-AI 2022, LNAI 13652, pp. 327–332, 2022.
https://doi.org/10.1007/978-3-031-21441-7_26

financial sector by testing the robustness of machine learning algorithms against unforeseen adversarial attacks. The effectiveness of machine learning algorithms is studied on an imbalanced credit card transaction dataset by applying a novel Synthetic Minority Oversampling Technique (SMOTE) with majority undersampling. Furthermore, this paper evaluates the success rate of adversarial examples in deceiving the classifier from making the right prediction.

2 Related Work

Khatri *et al.* [5] analyzed performance metrics outside the accuracy of the algorithm when dealing with imbalanced datasets and adopting sampling approaches for obtaining satisfactory results. Wang *et al.* [6] suggests outlier detection techniques can be a good workaround to address imbalanced datasets issues in fraud detection studies. A recent study by Tanouz *et al.* [1] implements different machine learning algorithms including decision tree, Naive Bayes, Random Forest, and Logistic Regression for fraud detection in credit cards. Undersampling and oversampling techniques were used in preprocessing stage to improve the performance of their algorithm. Many recent studies [7–9] suggest numerous possibilities of machine learning approaches in developing fraud detection mechanisms. Since adversarial attacks came into light [3], it has become a subject of major importance in the machine learning domain. While it has been used mostly for image recognition tasks using Deep Neural Networks(DNN), a recent paper by Carlini and Wagner [10] discussed the use of adversarial machine learning in audio recognition. Ballet *et al.* [12] coined the idea of implementing adversarial attacks on a tabular domain during the time when adversarial approaches were popular in testing the robustness of image classifiers. This study initiated the research on the impact of unobservable adversarial attacks on organized tabular data.

Ghamizi *et al.* [15] studied failures (false negative) of the state-of-the-art techniques fraud detection techniques to generate unobservable adversarial samples. This research addressed the usefulness of "Random Forest Attack" and "Gradient-Based attack" and concluded that these state-of-the-art approaches were ineffective to generate relevant adversarial samples for any chosen domain. The main differentiating factor of this research from the existing literature is generating adversarial samples on a highly imbalanced financial transaction dataset by incorporating suitable data preprocessing strategies. The novelty of this study also includes the use of Synthetic Minority Oversampling Technique (SMOTE) with majority undersampling to overcome data imbalance issues. It is important to look at the imbalanced dataset as most of the real-world datasets in this domain are highly imbalanced with very few real samples of positive fraudulent cases of credit card usage. The adversaries are generated using the LowProfool algorithm instead of conventional adversarial generation techniques. LowProfool is implemented using the Adversarial Robustness Toolbox(ART), thereby generating adversaries constrained to the chosen domain.

3 Methodology

3.1 Dataset

This project utilizes publicly available transactional credit card data sourced from several European card companies [15]. The data set contains online credit transactions within a 48-h time frame. The dataset is extremely imbalanced where the total number of fraudulent transactions constitutes approximately 0.17% of total transactions. The actual features in the dataset are hidden to ensure the confidentiality of individual card owners. The dataset represents features V1, V2, ...V27, V28 obtained via Principal Component Analysis (PCA) transformation. The known features in the dataset include time and amount of transactions. The class feature represents the category of each transaction; 0 represents a legit transaction and 1 denotes a fraudulent transaction. The initial data analysis shows that the dataset contains 284807 transactions with no null values however there were 1081 duplicate rows that were removed resulting in 283726 unique transactions. The dataset is extremely unbalanced with only 473 fraudulent transactions in the entire dataset. The correlation among the variables is represented using a heat map shown in Fig. 1. The features which are highly correlated include V10, V12, V14, V16, and V17. These values were capped by efficiently replacing extreme values with other close values of the variable by determining the minimum and maximum range using the mean and standard deviation. This project adopts Synthetic Minority Oversampling Technique (SMOTE) [15] which creates data samples from the minority fraud class in our dataset along with an undersampling strategy to reduce the number of data samples belonging to the legitimate class. This can eliminate the bias and noise induced by a SMOTE-only approach.

3.2 Classification Algorithms

This project investigates state-of-the-art machine learning algorithms including Logistic Regression, K-Nearest Neighbours, and Random Forest for credit card fraud detection. Figure 2 illustrates Logistic Regression's feature ranks measured using the "correlation coefficient". The results indicate that V2, V4, and V11 have positive importance scores. The K-Nearest Neighbour algorithm in this study uses Euclidean distance to measure similarities. The K value is set to 5 empirically. Random forest is usually less sensitive to changes made in training data and reduces the overfitting of the model to a greater extent [9]. Figure 3 demonstrates the feature importance plot generated by the random forest classifier model. The most important features are V17, V12, V14, V10, V11, and V3.

3.3 Adversarial Attack on Tabular Data

Adversarial data samples are generated using an adversarial algorithm whose primary goal is to create a data sample similar to the input sample by making

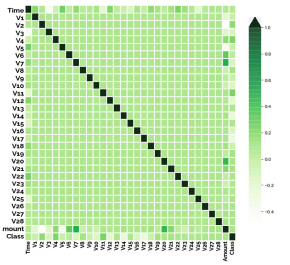

Fig. 1. Correlations

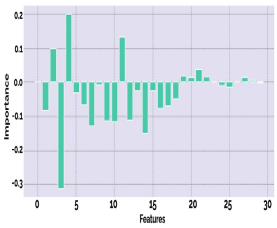

Fig. 2. LR features

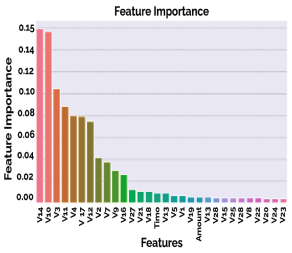

Fig. 3. RF features

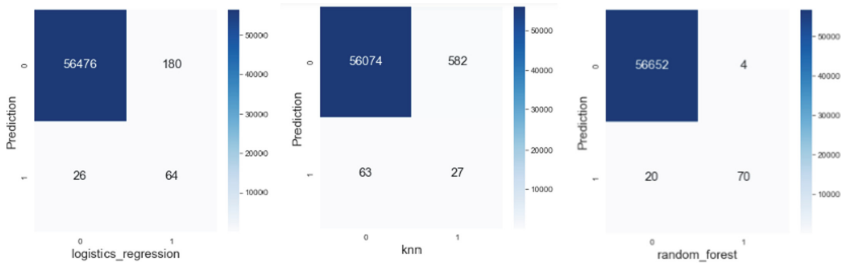

Fig. 4. Confusion matrix indicates algorithm performance across different classifiers

infinitesimal perturbations which lead to misclassification. This study employs evasion attacks [17] to generate adversarial samples on tabular data by inducing minimal changes to the input fed to the classifier thereby obtaining an incorrectly classified result. The adversarial generation by Lowprofool revolves around the computation of a feature importance vector for each input sample by computing lp-norm. The parameter p of lp-norm is set as 2, thereby computing a Euclidean distance while generating the feature importance vector [12]. The manipulated input sample are steered towards the opposite target class ensuring that manipulations are induced into less relevant features and are unobservable by the fraud detection system. The reliability of an adversarial attack is often evaluated using the metric referred to as success rate, which is the percentage of incorrect predictions.

4 Experimental Results and Observation

Model performance is tested with three different classifiers. Table 1 shows how novel SMOTE with undersampling had improved the fraud class performance for Logistic Regression. The overall system performance is shown in Table 2. Figure 4 illustrates the confusion matrix generated for Logistic Regression, K-nearest neighbors, and Random Forest algorithms. It is observed that Random Forest

exhibits better results compared to K-nearest neighbors and Logistic Regression. The Random Forest algorithm correctly predicts 56652 legal transactions and 70 illegal transactions. The second-best performer is the Logistic Regression detecting 64 illegal transactions. The LowProFool algorithm was able to mislead the classifier with a high fooling rate. The results of the original prediction were compared with the prediction made using the adversarial samples. Table 3 shows the performance of the Logistic Regression algorithm on adversarial samples for the fraud class. Result demonstrates that many data samples that were initially classified as illegal transactions (1) are converted to legitimate transactions (0) with the help of the lowprofool approach.

Table 1. Impact of SMOTE sampling on model performance

Class	Precision	Recall	F-score
1 (fraud)	0.15	0.63	0.24
0 (legal)	1.00	0.99	1.00
1(fraud) - SMOTE	0.84	0.58	0.68
0 (legal) - SMOTE	0.99	0.99	1.00

Table 2. Model performance metrics

Metrics	LR	KNN	RF
Precision	0.9980	0.9972	0.9994
Recall	0.9932	0.9856	0.9995
F1-score	0.9954	0.9912	0.9994
Accuracy	0.9932	0.9856	0.9995
AUC	0.8246	0.6211	0.8832

Table 3. Performance metrics with adversarial samples for class 1 (fraud)

Model	Precision	Recall	F1score
Real samples	0.15	0.63	0.24
Lowprofool adversarial samples	0.0098	0.0017	0.0039

5 Conclusion

This study aims to investigate the robustness of machine learning algorithms against unforeseen adversarial attacks in credit cards. It evaluates the performance of Logistic Regression, k-nearest neighbors, and random forest algorithms on a credit card transaction dataset to identify illegal transactions followed by generating unobservable adversarial samples by making infinitesimal changes to input and eluding the classifier with a high success rate. Primary results can be utilized to evaluate the robustness of classification algorithms and the emerging need for suitable defensive techniques in financial fraud detection models.

References

1. Sailusha, R., Gnaneswar, V., Ramesh, R., Rao, G.: Credit card fraud detection using machine learning. In: 2020 4th International Conference on Intelligent Computing And Control Systems (ICICCS), pp. 1264–1270 (2020)
2. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Mach. Learn. **23**, 69–101 (1996)
3. Abdallah, A., Maarof, M., Zainal, A.: Fraud detection system: a survey. J. Netw. Comput.Appl. **68**, 90–113 (2016)
4. Awoyemi, J., Adetunmbi, A., Oluwadare, S.: Credit card fraud detection using machine learning techniques: A comparative analysis. In: 2017 International Conference on Computing Networking And Informatics (ICCNI), pp. 1–9 (2017)
5. Khatri, S., Arora, A., Agrawal, A.: Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence), pp. 680–683 (2020)
6. Wang, H., Bah, M., Hammad, M.: Progress in outlier detection techniques: a survey. IEEE Access. **7**, 107964–108000 (2019)
7. Papernot, N., et al.: Technical report on the cleverhans v2. 1.0 adversarial examples library. ArXiv Preprint ArXiv:1610.00768 (2016)
8. Azhan, M., Meraj, S.: Credit card fraud detection using machine learning and deep learning techniques. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 514–518 (2020)
9. Gupta, A., Raghav, A., Srivastava, S.: Comparative study of machine learning algorithms for Portuguese bank data. In: 2021 International Conference on Computing, Communication, And Intelligent Systems (ICCCIS), pp. 401–406 (2021)
10. Cheng, M., Yi, J., Chen, P., Zhang, H., Hsieh, C.: Seq2sick: evaluating the robustness of sequence-to-sequence models with adversarial examples. Proc. AAAI Conf. Artif. Intell. **34**, 3601–3608 (2020)
11. Huang, S., Papernot, N., Goodfellow, I., Duan, Y., Abbeel, P: Adversarial attacks on neural network policies. ArXiv Preprint ArXiv:1702.02284 (2017)
12. Ballet, V., Renard, X., Aigrain, J., Laugel, T., Frossard, P., Detyniecki, M.: Imperceptible adversarial attacks on tabular data. ArXiv Preprint ArXiv:1911.03274 (2019)
13. Cartella, F., Anunciacao, O., Funabiki, Y., Yamaguchi, D., Akishita, T., Elshocht, O.: Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. ArXiv Preprint ArXiv:2101.08030 (2021)
14. Ghamizi, S., Cordy, M., Gubri, M., Papadakis, M., Boystov, A., Le Traon, Y., Goujon, A.: Search-based adversarial testing and improvement of constrained credit scoring systems. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 1089–1100 (2020)
15. Dal Pozzolo, A., Caelen, O., Johnson, R., Bontempi, G.: Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence, pp. 159–166 (2015)
16. Fernández, A., Garcia, S., Herrera, F., Chawla, N.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)
17. Nicolae, M., et al.: Adversarial Robustness Toolbox v1. 0.0. ArXiv Preprint ArXiv:1807.01069 (2018)