Vol: 04 Issue: 04

International Journal of Data Science and Advanced Analytics (ISSN: 2563-4429)

Data Science & Machine Learning Methods for Detecting Credit Card Fraud

Daniella Maya Haddab
Weizmann Institue of Science, Rehovot, Israel.
mayahdaniella@mail.com

Corresponding author email: mayahdaniella@mail.com

Abstract- Credit Card fraud is the physical loss of a credit card or the loss of sensitive information. For detection there are many machine learning algorithms that can be used. The research demonstrates several algorithms that may be utilized for determining transactions as fraud or genuine. The data set utilized in credit Card Fraud Detection was used in the study. The SMOTE technique was utilized for oversampling because the dataset was extremely imbalanced. Furthermore, feature selection was carried out as well as the set was split into 2 components, test data as well as training data. The algorithms employed in this investigation were Logistic Regression, Naive Bayes, Random Forest, and Multilayer Perceptron. The results demonstrate that each algorithm could be utilized with high accuracy for fraud detection of credit cards. For the detection of additional irregularities, the proposed model could be used.

Keywords: Credit card; Fraud; Logistic Regression; Multilayer Perceptron; Naive Bayes; Random Forest.

1. Introduction

Right now, there are many new businesses around the world [1]. Most companies are attempting to offer best service quality for their customers. To be successful, businesses regularly process a lot of data. This information is from many energy sources and is in different formats. Additionally, this information has several of the primary key areas of the company's long-term business. Due to that, businesses have to keep that information, process it, and what's truly crucial, to help keep it protected. Without securing information, many of it may be used by other businesses, or perhaps worse, stolen. In most cases, financial info is stolen, which may harm the total individual or company.

There are many types of frauds [2]. Check fraud occurs when a person forges an inspection or perhaps pays for one with a check, realizing there's not enough cash. Internet sales is fraud, where fraudsters sell fake products or maybe counterfeit clothes, or take payment while not delivering the product. You will find a few more, like charities fraud, identity theft, credit card fraud, debt elimination, insurance fraud, as well as others. Due to increasing interest in cashless transactions, among the most typical frauds are charge card frauds. Credit card fraud describes the circumstances in which a fraudster uses a credit card for their needs, while the owner of that charge card isn't mindful of which. Fraudulent transactions conducted using credit cards acquired globally amounted to $2.9 billion in 2021 [3]. Though there are huge volumes of increased charge card transactions, the number of frauds is proportionally the same, or even have reduced because of advanced fraud detection methods. Nevertheless, fraudsters are always coming up with new means to steal info [4].

You will find 2 types of charge card frauds. It is theft of actual physical flash cards, along with various other. You are stealing sensitive info out of the card, for example card number, CVV code, card type, along with other. By stealing charge card info, a fraudster can broach a huge amount of cash or even make a great amount of purchase prior to the cardholder finds out. Due to that, businesses work with several machine learning techniques to identify what transactions are fraudulent and which aren't.

The goal of this paper is to analyze various machine learning algorithms, for example Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB) and Multilayer Perceptron (MLP), to determine which algorithm is the most suitable for credit card fraud detection.

Most of the article is organized as follows: in Section II studies which deal with chosen issues, Section III offers a brief explanation of the dataset that can be used in the test, after that the outcomes are provided in Section IV. Finally, concluding remarks are reviewed in Section V, followed by a summary of literature.

2. Literature Review

We wanted to find a method to detect and prevent fraud, because fraudulent activities lead to major losses. Some techniques have already been suggested and tested. Below we will be looking at some of them in some detail.

Known classical algorithms like GradientBoost (GB), Support Vector Machines (SVM), Decision Tree (DT), LR and RF have proved useful. In paper

[5], GB, LR, RD, SVM, along with some classifiers, were utilized, leading to a high recall on a European dataset of more than 91 %. High accuracy and recall were achieved by under-sampling the information just after balancing the set. The European dataset was also used in the paper [6] and comparisons were made between the models based on LR, RF and DT. RF was the best of the three models, with a 95.5 % accuracy, followed by DT with 94.3 % accuracy, in addition to LR with 90 % accuracy.

Techniques based on k-nearest neighbors (KNN) and outlier detection can be effective in fraud detection, according to [7] and [8]. These techniques have been proven to reduce false alarms and increase the fraud detection rate. The KNN -algorithm additionally performed well in the experiment for paper [9], where the authors evaluated and compared it with other classical algorithms.

A comparison was made between a few traditional algorithms and deep learning methods in paper [10], compared to the earlier papers. All methods evaluated have an accuracy of approximately 80 %. The authors of the paper [11] set the following algorithms side by side: LR, GB, RF, DT, KNN, NB, XGBoost (XGB), MLP and stacking classifier (a mix of several machine learning classifiers) using a European dataset. All algorithms achieved a high accuracy of more than 90% due to the thorough data preprocessing. A stacking classifier with an accuracy of 95 % and a recall value of 95 % was most successful. A neural network was evaluated with the European dataset in the paper [12]. The experiment consisted of a rear propagation neural network optimized with the Whale algorithm. The neural network was made up of two input layers, 20 concealed as well as 2 output layers. Their optimization algorithm obtained remarkable results on 600 sample samples: It has a 97.40 % accuracy rate and a 96.83 % recall rate. The authors of [13] and [14] used neural networks to demonstrate improvements in results when ensemble methods are used. Three data sets were used in paper [15] for comparison between Auto-encoder and Restricted Boltzmann Machine algorithms, which led to the conclusion that algorithms such as MLP are ideal for credit card fraud detection.

Many papers focus on utilizing deep neural networks to detect fraudulent transactions. These models are computationally costly, however, and work better on large datasets [16]. As we observed in some papers, this approach can yield good results, but what if it can be done with fewer resources? Our main objective is to demonstrate that with appropriate preprocessing, different machine learning algorithms can produce satisfactory results. The authors of most of these papers used the sampling technique under the motivation for utilizing a different oversampling technique.

The editors of this paper chose to evaluate the suitability of LR, RF, MLP and NB for credit card fraud detection, considering given facts. An experiment was carried out to achieve this, and the transformation of these input variables was performed to keep this information private. The three features were not transformed. The time feature displays the time between the first transaction and each additional transaction in the dataset. The feature "Amount" is the quantity of transactions made by credit card. The feature class provides the label and uses only 2 values: The value is 1 in the event of a fraudulent transaction and 0 in any event.

The dataset consists of 274,807 transactions, where 490 transactions were rip-offs, and the remainder were authentic. This dataset is extremely imbalanced, as only 0.273 % of transactions are classified as fraud.

The preprocessing of data is crucial, as the distribution ratio of classes plays a crucial role in the accuracy and precision of the model.

B. Preprocessing feature selection is a fundamental method that picks the variables that are most pertinent in the provided dataset. Choosing the right features can improve accuracy, reduce training time, and reduce over-fitting, by carefully selecting them and removing them as they don't fit. This process can be helped with visualization techniques. In this experiment, the feature selector tool [18] by Will Koehrsen was used for that purpose. It was determined by using this tool, whose features would be the most crucial. In addition, features that don't contribute to the cumulative significance of 95 % have been removed. 27 attributes were selected for further experiment after the feature selection method.

When classification categories are not approximately equally distributed, machine learning algorithms have difficulty learning. Due to the highly imbalanced data, it is essential to do some balancing so that the model can be trained efficiently. Methods for adjusting class distribution consist of frequently used methods for sampling the majority type, oversampling the minority type, or maybe a combination of these two. The Synthetic Minority Oversampling Technique (SMOTE) is a popular Oversampling method that is useful when applied to imbalanced datasets [19, 20]. To improve random oversampling, the SMOTE method was proposed.

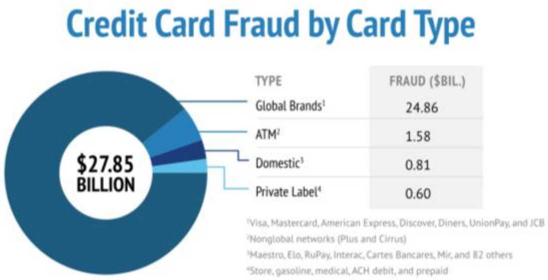

Figure 1: Credit card fraud type and value (Source: European Central Bank (2018)

3. Methods of Research

The Credit Card Fraud Detection dataset was utilized in this research and may be downloaded from Kaggle [17]. This data set contains transactions that happened in two days and was made by European cardholders in September 2013. There are 31 numeric features in the dataset. The PCA is performed because some input variables contain financial information. Figure 1. Figure 1. Before and after sampling, many machine learning algorithms depend on the size of the input. Scaling is done to bring all capabilities to the same magnitude, considering the highly variable values of amount and time. The environment for the experiment is the Windows OS, and the software program operating environment is Spyder, the scientific Python development atmosphere part of the Anaconda platform. Other used libraries include: Imblearn, pandas, numpy sklearn, matplotlib and matplotlib.

In the following section, the previously mentioned algorithms used in the experiment are discussed. Experimental logistic regression is among the most widely used classification algorithms in machine learning. The logistic regression method describes the connection between predictors that may be categorical, binary and continuous. A dependent variable may be binary. Based on some predictors, we predict whether something will happen. For any set of predictors, we compute the likelihood of belonging to each group. Naive Bayes is among the supervised learning algorithms in which there's no dependence between attributes. Based on the Bayes theorem. Following algorithms are available depending on the type of distribution used: Bernoulli distribution, Multinomial distribution, Gaussian distribution. The Bernoulli distribution is used in this research to detect bribery transactions.

The random forest algorithm may be used in both regression and classification problems. It's made up of numerous decision trees. This algorithm provides better results when more trees are within the forest and prevents model over-fitting. Every decision tree in the forest provides some results. To obtain a more exact and stable prediction, these results are merged. A multilayered perception is an artificial neural network, which consists of at least three levels of nodes: input layer, concealed layer, as well as output layer. Use the activation function of each node. The activation function calculates the weighted total of the data and contributes bias. This enables us to find out what neurons should be taken out and never considered in exterior connections.

The used ANN in the experiment was comprised of four concealed layers, 50, 30, thirty, as well as fifty units in each concealed layer, with a real activation function. As shown in, deeper networks obtain better results than smaller networks [20]. Based on our experience, we started with a few layers and increased them gradually to get the results we wanted. Based on this research, the best hyper-parameters were therefore chosen. The further improvement of the network led to greater computational time, and the obtained results did not differ much from the selected architecture. With Adam, stochastic gradient-based optimizer, weight optimization was accomplished.
The train and the test set were divided in 80: The ratio was 20 and the design was updated through many epochs based on the tolerance for optimization (TOL). If the loss or rating for specific successive iterations isn't improving by more than TOL, convergence is believed to have been attained and training stops.

4. RESULTS AND DISCUSSION

To figure out which algorithm is best for the detection of fraud transactions, various criteria have been used for algorithm comparison. The most typical indicators for determining the end result of machine learning algorithms are precision, recall, and accuracy. All the metrics pointed out can be computed from a Confusion matrix. These metrics were used to guide the evaluation of the performance of a model. The models were tested on original and over-sampled data, and the results have proven that sampling is important.
Because the test set contains 20% of the entire dataset, the total number of samples is 56962.

Among the 98 fraud transactions, the LR model (table 1) achieved:

- precision: 59.82%,
- recall: 92.84%,
- accuracy: 96.46%.

Table 1: Matrix for LR

		Predicted	
		0	1
Actual	0	55124	14409
	1	9	91

NB model obtained following results (Table 2):
precision: 18.13%,
recall: 83.65%,
accuracy: 99.83%

Table 2: Matrix for NB

		Predicted	
		0	1
Actual	0	51442	621
	1	19	91

RF model obtained following results (Table 3):
precision: 97.38%,
recall: 83.63%,
accuracy: 98.96%.

Table 3: Matrix for RF

		Predicted	
		0	1
Actual	0	56591	3
	1	19	82

MLP model obtained following results (Table 4):
precision: 80.21%,
recall: 81.63%,
accuracy: 99.43%

Table 4: Matrix for MLP

		Predicted	
		0	1
Actual	0	51843	24
	1	19	81

By analyzing obtained outcomes, it's apparent that accuracy is incredibly high, though that does not imply that results are ideal. Accuracy must be taken "with a grain of salt" - desirable, it must be viewed in conjunction with various other metrics. Based on given results, it's proven that a classic algorithm, including RF, can provide results similar to an easy neural network.

Comparison of the obtained outcomes with outcomes attained in investigations on the same dataset, with classical algorithms [5] and [23], indicates that oversampling the information can boost the fraud detection rate. Like in papers [ten] and [11], it's proven that classical algorithms are often as effective as serious learning algorithms. Although papers [twelve] and [15] represent heavy learning algorithms as ideal for this type of problem, it must be decided based on the situation in which of those should be used. For instance, deep networks work more efficiently with more data and could be adapted to various domains more effortlessly compared to classical algorithms. On the flip side, when there's not a lot of info, it's possibly preferable to handle classical algorithms. These algorithms are simpler to understand and less expensive, both in the computational and financial sense [24].

5. CONCLUSION

Charge card frauds are a serious business problem. These frauds can lead to large losses, both business and private. Due to that, companies invest increasingly more cash in developing new ways and ideas that will enable them to detect and prevent frauds.

The primary objective of this paper was to compare specific machine learning algorithms for detection of fraudulent transactions. Hence, a comparison was created, and it was started that Random Forest algorithm provides the best results, i.e. best classifies if transactions are fraud or perhaps not. This was established using various metrics, like recall, precision, and accuracy. Because of this problem type, it's vital that you have recall with high value. Include choice in addition to balancing the dataset has proven to be essential in obtaining significant results.

Additional study must focus on various machine learning algorithms, like genetic algorithms, and various kinds of stacked classifiers, alongside comprehensive characteristic choice being much better results.

References

[1] A. Mishra, C. Ghorpade, “Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques” 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE.

[2] A. Pumsirirat, L. Yan, “Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine”, 2018 International journal of advanced computer science and applications, 9(1), pp. 18-25

[3] Akter Jahan, S., & Sazu, M. H. (2022). Rise of mobile banking: a phoenix moment for the financial industry. Management & Datascience, 6(2).

[4] Akter, J. S., & Haque, S. M. (2022). Innovation Management: Is Big Data Necessarily Better Data?. Management of Sustainable Development, 14(2), 27-33.

[5] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai, S. Pan, “Credit card fraud detection based on whale algorithm optimized BP neural network”, 2018 13th International Conference on Computer Science & Education (ICCSE) pp. 1-4. IEEE.

[6] Deeplearningbook.org. (2019). Deep Learning. [online] Available at: https://www.deeplearningbook.org/ [Accessed 11 Jan. 2019].

[7] European Central Bank (2018). Fifth report on card fraud, September 2018. [online]. Available at: https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport2018 09.en.html#toc1 [Accessed 21 Jan. 2019].

[8] F. Ghobadi, M. Rohani, “Cost Sensitive Modeling of Credit Card Fraud using Neural Network strategy”, 2016 Signal Processing and Intelligent Systems (ICSPIS), International Conference of pp. 1-5. IEEE.

[9] Soh, W. W., & Yusuf, R. (2019). Predicting Credit Card Fraud on a Imbalanced Data. International Journal of Data Science and Advanced Analytics (ISSN 2563-4429), 1(1), 12-17.

[10] Global Facts (2019). Topic: Startups worldwide. [online] Available at: https://www.statista.com/topics/4733/startups-worldwide/ [Accessed 10 Jan. 2019].

[11] International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC) ional conference on computation of power, energy, Information and Communication (ICCPEIC) pp. 191-

[12] Isenberg, D. T., Sazu, M. H., & Jahan, S. A. (2022). How Banks Can Leverage Credit Risk Evaluation to Improve Financial Performance. CECCAR Business Review, 3(9), 62-72.

[13] J. O. Awoyemi, A. O. Adentumbi, S. A. Oluwadare, “Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis”, Computing Networking and Informatics (ICCNI), 2017 International Conference on pp. 1-9. IEEE.

[14] J. Wang, M. Xu, H. Wang and J. Zhang, "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding", Signal Processing, 2006 8th International Conference on (Vol. 3). IEEE. 2006 8th international Conference on Signal Processing, Beijing, 2006

[15] JAHAN, S. A., & Candidate, M. B. A. Factors Affecting The Adoption Of Financial Technology Among The Banking Customers In Emerging Economies. Financial Studies, 39.

[16] Kaggle.com. (2019). Credit Card Fraud Detection. [online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud [Accessed 10 Jan. 2019]. Github (2019). Feature selector. [online] Available at: https://github.com/WillKoehrsen/feature-selector [Accessed 18 Jan. 2019].

[17] Legal Dictionary (2019). Fraud - Definition, Meaning, Types, Examples of fraudulent activity. [online] Available at: https://legaldictionary.net/fraud/ [Accessed 15 Jan. 2019].

[18] Mrs. C. Navamani, M. Phil, S. Krishnan, “Credit Card Nearest Neighbor Based Outlier Detection Techniques”

[19] N. Kalaiselvi, S. Rajalakshmi, J. Padmavathi, “Credit card fraud detection using learning to rank approach”, 2018 Internat2018

[20] N. Malini, Dr. M. Pushpa, “Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection“, Advances in Electrical, Electronics, Information, Communication and Bio- Informatics (AEEICB), 2017 Third International Conference on pp. 255-

[21] S. Dhankhad, B. Far, E. A. Mohammed, “Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study”, 2018 IEEE International Conference on Information Reuse and Integration (IRI) pp. 122-125. IEEE.

[22] S. V. S. S. Lakshmi, S. D. Kavilla “Machine Learning For Credit Card Fraud Detection System”, unpublished

[23] Sazu, M. H., & Jahan, S. A. (2022). Impact of blockchain-enabled analytics as a tool to revolutionize the banking industry. Data Science in Finance and Economics, 2(3), 275-293.

[24] Z. Kazemi, H. Zarrabi, “Using deep networks for fraud detection in the credit card transactions”, Knowledge-Based Engineering and Innovation (KBEI), 2017 IEEE 4th International Conference on pp. 630-633. IEEE.