Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 364–369

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

# On the benefits of machine learning classification in cashback fraud detection

Bryan Karunachandra, Nathaniel Putera, Stephen Rian Wijaya, Dewi Suryani*, Julian Wesley, Yudy Purnama

*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia*

**Abstract**

Technology development has been getting more advanced and greatly facilitated human life. One of them is machine learning automation which has been proven to be consistent for doing various computations against extensive data such as transaction data in the e-commerce area. Seeing this opportunity, we implemented the machine learning approach to detect fraudulent cashback transactions in e-commerce that are currently rife in Indonesia. The training data used to build the machine learning model were the transaction data from one of the leading e-commerce in Indonesia that had been processed. The supervised classification algorithms used were K-Nearest Neighbor (k-NN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). In the end, the best steps and methods that could be taken against fraudulent cashback activities in the future are shown in this paper.

© 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 7th International Conference on Computer Science and Computational Intelligence 2022

*Keywords:* Fraud Detection; E-Commerce; Classification; Machine Learning; Cashback Fraud

## 1. Introduction

Nowadays, Indonesia is experiencing fairly rapid economic and technological development. It can be seen from a public transport ticket or food ordering point of view, where initially people will directly come to the station or restaurant, now they can do it over the Internet or online. In addition, shopping has also begun to penetrate the online world, which is commonly called electronic commerce (e-commerce). In Indonesia, e-commerce has successfully attracted many people and become a rapidly growing industry in terms of daily transaction volume. Several e-commerce startups developing in Indonesia were even able to enter the top 10 of Southeast Asia's unicorns, namely Tokopedia, Bukalapak, etc.[1]. One of e-commerce's methods to attract a large number of customers is by offering many promotions such as cashback, discounts, free delivery, etc. However, this phenomenon is not only attracting customers'

* Corresponding author.
*E-mail address:* dewi.suryani@binus.ac.id

1877-0509 © 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 7th International Conference on Computer Science and Computational Intelligence 2022

10.1016/j.procs.2022.12.147

attention but also attracting criminals' interest to abuse it for their own benefit. One example of promotion misconduct done by the criminal is by creating multiple fraudulent accounts and transactions.

By looking at the various types of promotion that have the possibility of fraud, we conducted a survey of 70 respondents from e-commerce users. The survey shows that the most frequent frauds that occur on e-commerce platforms were from transactions with promotion type of cashback. Based on the result of the survey, this study focuses on detecting cashback fraud. The importance of fraud detection is felt not only by the e-commerce parties but also by the general public to anticipate it. On the other hands, the development of technology is getting more advanced and has greatly facilitated human life, especially machine learning automation which has been proven to be consistent for doing various computations against large data and can solve several classification tasks such as handwriting recognition[2][3], sign language recognition[4], criminal detection[5], credit card fraud detection[6], etc. These also encouraged us to conduct this research on cashback fraud detection by implementing classification using machine learning methods. In order to get the best machine learning approach for our case, we compare 3 different algorithms in this work, i.e., convolutional neural networks (CNN), long short-term memory (LSTM), and k-nearest neighbor (k-NN).

The remaining of this paper is organized as follows: In Section 2, we describe our approach relations to the other approaches, and then we present our research methodology in Section 3. Afterwards, the details of our experimental settings and the results are explained in Section 4. Last but not least we conclude our works in this paper and discussion our future work in Section 5.

## 2. Related Works

Several approaches had been widely used for detecting fraud, such as data mining, feature engineering, and the use of machine learning and deep learning classification algorithms[7][8][9]. All those algorithms and methods are commonly measured by using a confusion matrix. The confusion matrix maps four groups based on true and false value[10]. The obtained value can be divided into various metrics such as accuracy, precision, recall, etc. Based on various experiments conducted by Wickramanayake et al.[7], the random forest machine learning algorithm produces the best retrieval value to detect online fraud. Moreover, in the deep learning algorithm experiment, long short-term memory (LSTM) currently outperformed the other algorithms for detecting anomalies tasks in transaction data with sequential types, which presented in the research of Jurgovsky, J. et al.[11], and Singh, A.[12]. On the other hands, Malini, N. and Pushpa, M.[13] have taken the advantage of the k-Nearest Neighbor (k-NN) algorithm which has been optimized by combining it with various outlier detection methods. Their research resulted in the reduction of false alarm rates and the enhancement of fraud detection rates.

Furthermore, the combination of convolutional neural networks and long short-term memory (CNN-LSTM) has also been tested for a similar case with the unbalanced dataset. The architecture that was implemented is the Siamese neural network. With that architecture, the task was divided into feature extraction using the CNN model in order to find the intrinsic patterns and the transaction information recall model using the LSTM model[14,15,16,17]. The result of this model has produced a 95% recovery value and 96% precision. As well as the work of Malini, N. and Pushpa, M.[13], we also implemented the k-NN algorithm to detect fraud, however, we focused on cashback fraud detection instead of the credit card.

## 3. Research Methodology

The detail of our methodology architecture is depicted in Fig. 1.

### 3.1. Data Pre-processing

In the data preprocessing stage, the data need to be cleaned first. For some missing data whose number of rows was less than 5%, they were cleaned using the Complete Case Analysis (CCA) technique. Meanwhile, for other missing data, the filling could be done by looking at the existing missing data mechanism. Generally, the missing data mechanism falls into three categories, namely, Missing data Completely At Random (MCAR), Missing data At Random (MAR), and Missing data Not At Random (MNAR). One of the characteristics that this technique applied in

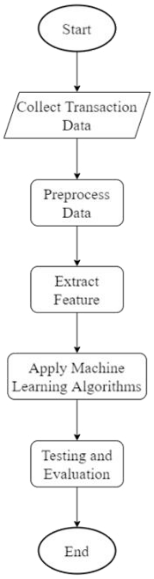

Fig. 1. Methodology of this research.

our case was the column *cashback amount* that is filled with a value of 0 (zero), assuming that the transaction was a transaction without using a cashback promotion.

### 3.2. Feature Extraction

At this phase, feature generalization, min-max normalization, and also hot encoding were performed into the data that have been cleaned. Feature generalization is the process of grouping or simplifying data with a large number of unique values. For example, the *buyer ip* column can be grouped according to five network groups, namely classes A, B, C, D, and E. Min-max normalization is used to reconstruct numeric data in a range of 0 (zero ) to 1 (one). Moreover, hot encoding is used to binarize a column with a category data type. After the data were completed going through the 3 processes above, the feature extraction will be carried out using 5 different algorithms. The first algorithm is the Pearson Correlation which looks for a correlation between two columns, where when it approaches -1 or 1 then the relationship becomes stronger, and when approaching the number 0 it indicates otherwise. The second algorithm is Chi Square which finds the degrees of freedom of a column, where the smaller the degrees of freedom, the stronger the relationship is. The third algorithm is Recursive Feature Elimination (RFE) which removes features repeatedly until it meets the specified number of features. The fourth algorithm is the Least Absolute Shrinkage and Selection Operator (LASSO), a method of reducing the freedom of a machine learning model by regularizing features. The final algorithm is Random Forrest which contains a series of decision trees that handle prediction using ”yes” or ”no” conditions and voting. In the end, all the features that passed the test by the five algorithms were accumulated and a vote was taken to determine the 15 potential features that would be used in this work.

### 3.3. Classification

This step was the main phase of this research where several experiments were carried out using 3 different algorithms, namely CNN, LSTM, and k-NN. We tried CNN and LSTM because they have been well performed on time series data and for several other reasons. The CNN used in this study is one-dimensional CNN because the data source is not an image (two-dimensional data). LSTM is used because it can solve the problem of vanishing gradient. The combined CNN-LSTM method was included based on related works, which showed that combining these models could produce better precision than each model as stand alone method. The last method we use is k-NN, a simple method commonly used in classification. The performance results of the k-NN classification are highly dependent on the quality of the input data because it is based on the closest distance between the points of its neighbors. The parameter k that represents how many closest neighbors are to a point is determined by the set learning method[14].

## 4. Experimental Settings

### *4.1. Dataset*

Dataset used in this experiment is transaction data that had been gathered from e-commerce in Indonesia. In the beginning, the dataset was divided into 6 main tables, i.e., *addresses* (141402, 13), *registers* (73536, 23), *shippings* (125067, 17), *transactions* (140727, 51), *users* (109300, 20), and *vouchers* (104827, 14). However, after the preprocessing had been conducted using the method which has been explained in Section 3, the dataset is organized into 4 types of datasets with the combinations of the 6 main tables above. The details of the dataset will be shown in Table 1. Each dataset is divided into training data, validation data, and test data. For training data and test data, they are taken from the total data with a ratio of 70%: 30% respectively, while the validation data is extracted from 20% of the training data. As an additional note, columns in brackets '[ ]' which appears in Table 1 are not included in the machine learning process as their purposes are only to support the results.

Table 1. Dataset details.

| Code | Name | Selected Columns | Shape |
|---|---|---|---|
| DS-1 | DATASET1 | cashback amount, trx amount, shipping cost, coded amount, trx_created_at, buyer_id, seller id, city, postal code, payment method 0, payment method 1, payment method 2, payment method 3, payment method 4, payment method 5, payment method 6, payment method 7, payment method 8, payment method 9, payment method 10, payment method 11, payment method 12, buyer ip A, buyer ip B, buyer ip C, label | (68831, 26) |
| DS-2 | DATASET2 | cashback amount, trx amount, shipping cost, coded_amount, payment method_0, payment method 1, payment method 2, payment method 3, payment method 4, payment method 5, payment method 6, payment method 7, payment method 8, payment method 9, payment method 10, payment method 11, payment_method_12, buyer ip A, buyer ip B, buyer ip C, label | (68831, 21) |
| DS-3 | DATASET3 | [buyer id], [seller id], [trx_id_x], trx_amount, coded_amount, percentage, payment method 0, payment method 1, payment method 2, payment method 3, _ payment method 5, payment method 6, payment method 7, payment_method_9, has bank account 1, max voucher usage, insurance cost, label | (48941, 18) |
| DS-4 | DATASET4 | [buyer id], [seller id], [trx id x], trx amount, shipping cost, coded_amount, cashback amount, payment method 1, payment_method_2, payment_method_3, payment method 5, payment method 6, payment method 8, payment method 9, payment method 10, payment method 12, insurance cost, level 3, label | (78516,19 ) |

### *4.2. Hardware Specification*

Hardware specifications used in this experiment are described in Table 2.

Table 2. Hardware specifications.

| Specification | Details |
|---|---|
| RAM | 16GB |
| Processor | Intel Core i7 Skylake |
| GPU | NVIDIA GeForce GTX 960M |

### 4.3. Hyperparameter Setup

In addition to the hardware specifications, we also need to define the initial hyperparameter to be used. The details of the initial hyperparameter used for our artificial neural network (ANN) model are described in Table 3. This initial hyperparameter was applied in CNN, LSTM, and CNN-LSTM experiments, which were called ANN experiments.

Table 3. Initial hyperparameter for ANN experiments

| Code | Learning Rate | Dropout 1 | Dropout 2 | Dropout 3 | Dropout 4 | Epoch |
|---|---|---|---|---|---|---|
| ANN-1 | 0.01 | 0.1 | 0.1 | 0.1 | 0.5 | 1000 |
| ANN-2 | 0.01 | 0.1 | 0.1 | 0.5 | - | 1000 |
| ANN-3 | 0.01 | 0.1 | 0.1 | 0.1 | - | 1000 |
| ANN-4 | 0.01 | 0.1 | 0.1 | - | - | 1000 |

Meanwhile. the details of the initial hyperparameter used for our k-NN model are described in Table 4. The value of *k* is obtained using the assembly learning method that is applied to each data set. Meanwhile, *n-jobs* determines how many cores the CPU uses when debugging. The larger the *n-jobs*, the faster the k-NN parsing process, but it consumes more power for the CPU.

Table 4. Initial hyperparameter for k-NN experiments.

| Code | n-neighbour | n-jobs |
|---|---|---|
| KNN-1 | 5 | 16 |
| KNN-2 | 73 | 16 |

## 5. Results

During this research, several preliminary experiments were carried out to determine what type of dataset to use for the main experiment. Based on the result of these preliminary experiments, we obtained DATASET3 with the code DS-3 which will be used for our main experiment because this dataset is the best dataset with a wide coverage of features. Furthermore, from the initial experiments, it was also found that many factors affect the accuracy, such as hyperparameters, the quality of the dataset, and also the architectural model used. Using the experimental settings that have been determined in Section 4, the main experiment was carried out using the DS-3 dataset, the detailed result can be seen in Table 5.

Table 5. The result of main experiments.

| Algorithm | Validation Accuracy (%) | Test Accuracy (%) | Test Time | Hyper-parameter Code |
|---|---|---|---|---|
| CNN | 63.44 | 49.39 | ±0:01:06.00 | ANN-1 |
| LSTM | 54.48 | 51.13 | ±0:01:11.00 | ANN-1 |
| CNN-LSTM | 53.58 | 52.14 | 0:00:58.995 | ANN-3 |
| k-NN | 84.36 | 83.82 | 0:00:02.031 | KNN-1 |

## 6. Conclusion and Future Works

In this study, we implemented several machine learning algorithms to detect e-commerce cashback fraud. Based on the experiments that have been carried out, it was shown that the k-NN algorithm is the best machine learning algorithm in our case with an accuracy of 83.82%. In the future, comparisons can be made using deep learning, where it is wished that the output of deep learning could produce much better results.

## Acknowledgements

This work is supported by Research and Technology Transfer Office, Bina Nusantara University as a part of Penelitian Terapan Binus entitled Fraud Detection using Data Mining Approaches with contract number: No. 025/VR.RTT/IV/2020 and contract date: 6 April 2020.

## References

[1] Team, T.A.P.. Asean unicorns on the rise. The ASEAN Post: Something Linky; 2020. Accessed: December 2020.

[2] Suryani, D., Doetsch, P., Ney, H.. On the benefits of convolutional neural network combinations in offline handwriting recognition. In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE; 2016, p. 193–198.

[3] Alkawaz, M.H., Seong, C.C., Razalli, H.. Handwriting detection and recognition improvements based on hidden markov model and deep learning. In: *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE; 2020, p. 106–110.

[4] Koller, O., Camgoz, N.C., Ney, H., Bowden, R.. Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. *IEEE transactions on pattern analysis and machine intelligence* 2019;**42**(9):2306–2320.

[5] Shirsat, S., Naik, A., Tamse, D., Yadav, J., Shetgaonkar, P., Aswale, S.. Proposed system for criminal detection and recognition on cctv data using cloud and machine learning. In: *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking(ViTECoN)*. IEEE; 2019, p. 1–6.

[6] Adepoju, O., Wosowei, J., Jaiman, H., et al. Comparative evaluation of credit card fraud detection using machine learning techniques. In:

[7] *2019 Global Conference for Advancement in Technology (GCAT)*. IEEE; 2019, p. 1–6.

[8] Wickramanayake, B., Geeganage, D.K., Ouyang, C., Xu, Y.. A survey of online card payment fraud detection using data mining-based methods. *arXiv preprint arXiv:201114024* 2020;.

[9] West, J., Bhattacharya, M.. Intelligent financial fraud detection: a comprehensive review. *Computers & security* 2016;**57**:47–66.

[10] Hassan, A.K.I., Abraham, A.. Modeling insurance fraud detection using imbalanced data classification. In: *Advances in nature and biologi-cally inspired computing*. Springer; 2016, p. 117–127.

[11] Miao, J., Zhu, W.. Precision–recall curve (prc) classification trees. *Evolutionary Intelligence* 2021;:1–25.

[12] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.E., He-Guelton, L., et al. Sequence classification for credit-card fraud detection. *Expert Systems with Applications* 2018;**100**:234–245.

[13] Singh, A.. Anomaly detection for temporal data using long short-term memory (lstm). 2017.

[14] Malini, N., Pushpa, M.. Analysis on credit card fraud identification techniques based on knn and outlier detection. In: *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. IEEE; 2017, p. 255–258.

[15] Hassanat, A.B., Abbadi, M.A., Altarawneh, G.A., Alhasanat, A.A.. Solving the problem of the k parameter in the knn classifier using an ensemble learning approach. *arXiv preprint arXiv:14090919* 2014;.