İSTANBUL UNIVERSITY PRESS

Acta Infologica
acin 2024, 8 (1), 60–70
DOI: 10.26650/acin.1475658

Research Article / Araştırma Makalesi

Diagnosis of Internal Frauds using Extreme Gradient Boosting Model Optimized with Genetic Algorithm in Retailing

Aytek Demirdelen[1], Pelin Vardarlıer[2], Yurdagül Meral[2], Tuncay Özcan[3]

[1]İstanbul Medipol University, Institute of Social Sciences, PhD Program in Business Administration, İstanbul, Türkiye
[2]İstanbul Medipol University, Faculty of Business and Management Sciences, Department of Human Resources Management, İstanbul, Türkiye
[3]İstanbul Technical University, Faculty of Business Administration, Department of Management Engineering, İstanbul, Türkiye

Corresponding author : Tuncay Özcan
E-mail : tozcan@itu.edu.tr

ABSTRACT

Fraud is one of the most vital problems that can lead to a loss of organizational reputation, assets and culture. It is beneficial for companies to anticipate possible fraud in order to protect both culture and company assets. The aim of this study is to provide a fraud detection model using classification and optimization algorithms. For this purpose, this study proposes a novel hybrid model called XGBoost-GA to enhance the prediction quality for cashier fraud detection in retailing. In the proposed model, the genetic algorithm (GA) is used to optimize the parameters of extreme gradient boosting (XGBoost) model. The proposed XGBoost-GA model is compared with XGBoost, logistic regression (LR), naive bayes (NB) and k-nearest neighbor (k-NN) algorithms. The performance comparison is presented with a case study with the actual data taken from a grocery retailer in Turkey. Numerical results showed that the proposed hybrid XGBoost-GA model produces higher accuracy, recall, precision and F-measure than other classification algorithms. In this context, the use of proposed model in fraud detection will be beneficial for companies to use their resources effectively. Classification algorithms will also accelerate organizations in terms of detecting the possible damage of fraud to company assets before it grows.

Keywords: Fraud Detection, Retailing, Machine Learning, Extreme Gradient Boosting, Genetic Algorithm

Submitted : 01.05.2024
Accepted : 14.05.2024
Published Online : 03.06.2024

1. INTRODUCTION

The short story Minority Report, written by Philip K. Dick in 1956, was brought to the big screen by Steven Spielberg in 2002 with the same name. The film reached large audiences with its detection, prediction, actors and special effects of crime and criminals and raised questions in the minds. In this movie, the crimes are seen by the oracles before they happen and can be prevented beforehand. In this context, we are depicted in a world where not only the detection of crime but also foresight is dominant in the fight against crime and the detection is thought to be made before the action. This depiction is partially possible today, thanks to the advanced computer. In this context, early detection of fraud plays a critical role. A proactive approach in early detection is beneficial in protecting companies' assets (Erol, 2016).

The retail sector, which has a high transaction volume, also takes measures with audit activities to detect abuses it faces. Loss prevention and internal audit processes, which are among the business processes of the retail industry, have also undergone digital transformation and had to be reshaped. This transformation has gained importance for companies in terms of early detection of fraud and reduction of possible damage.

The impact of digital transformation on business processes is naturally reflected in audit methodologies. Audit processes are being reshaped, data analytics, automation etc. test methods have started to be used in field applications of process controls. The world of auditing resulting from Industry 4.0 will automate existing procedures, expand audit scopes, save time and increase the quality of audit assurance (Esmeray, 2018).

In this context, artificial intelligence, machine learning and advanced data analytics applications, which are part of the change created by the elements of Industry 4.0, have enabled the application of a proactive approach in the detection of frauds. Therefore, this study aims to present a hybrid model using extreme gradient boosting (XGBoost) and genetic algorithm (GA) for cashier fraud detection in the retail sector.

The remainder of this study is as follows: In Section 2, a review of the literature studies on the fraud detection problem is provided. In Section 3, gradient boosting algorithm, genetic algorithm and the proposed XGBoost-GA model are introduced. In section 4, the performance analysis of the developed models for the cashier fraud detection are given. In Section 5, the findings and conclusions are presented.

2. LITERATURE REVIEW

Fraud detection is an interesting research topic for both practitioners and academics. In the literature, there are many studies on credit card, telecommunication, tax/customs and insurance fraud. These studies can be summarized as follows:

Hanagandi et al. (1996) created a scoring system for credit card fraud detection by combining a radial basis function network with a density-based clustering algorithm. In this study, artificial neural networks were used to create the model. Shen et al. (2007) investigated the effectiveness of applying classification models to credit card fraud detection problems. In this study, the performance of decision tree, artificial neural network and logistic regression algorithms in fraud detection was tested. Seyedhossein & Hashemi (2010) proposed a method based on the creation of customer profiles for credit card fraud detection. The focus of this study is on cases of fraud that are not detected at the transaction level. In the proposed method, daily amounts spent on an individual credit card account were examined by time series analysis. Bhattacharyya et al. (2011) compared the performance of logistic regression, random forest, and support vector machines for credit card fraud detection. Sahin & Duman (2011) analyzed the decision tree and support vector machines (SVM) for credit card fraud detection in their study. Perols (2011) used six popular statistical and machine learning models to detect financial statement fraud. Numerical results showed that logistic regression and support vector machines performed well according to artificial neural network, C4.5 and stacking. Mahmoudi & Duman (2015) used a linear separator called Fisher Discriminant Function (FDA) in their study to detect credit card frauds. Vlasselaer et al. (2015) proposed a new approach called APATE (Anomaly Prevention using Advanced Transaction Exploration) to detect fraudulent credit card transactions in online stores. In the proposed approach, the characteristics of incoming transactions and the time since the last shopping date, shopping frequency and shopping amount derived from customer spending history are combined. Then, using this data from the network of credit card holders and businesses, a time-dependent risk score was derived for each network object. Renjith (2018) used the support vector machines method to detect fraudulent sellers in online sales areas. Additionally, it was stated that the algorithm used would not be sufficient to make a decision for a new seller whose historical data is not available. Shukur & Kurnaz (2019) used Logistic Regression, Artificial Neural Networks and K-Nearest Neighbor methods for credit card fraud detection. In this study, numerical results showed that Logistic Regression had the best classifier performance and K-Nearest Neighbor algorithm had the worst classifier performance. Nadim et al. (2019) compared different machine learning methods for credit card fraud detection according to performance criteria such as accuracy, precision, sensitivity and specificity. As

a result of this comparison, it was revealed that logistic regression, random forest and XG-Boost algorithms gave the best results according to the accuracy rate, and random forest and XG-Boost algorithms gave the best results according to the cost criterion. Niu et al. (2019) performed the performance analysis of supervised and unsupervised learning techniques for credit card fraud detection using AUC-ROC curves. As a result of the study, among supervised learning techniques, the XG-Boost classifier was the most successful method with an accuracy rate of 98.94%, while the Decision Tree classifier was the least successful method with an accuracy rate of 95.42%. Pehlivanli et al. (2019) used support vector machine and artificial neural network methods to detect fraudulent purchases in the retail industry. In this study, different kernel functions were tested for the support vector machine and it was revealed that the support vector machine performed better. Varmedja et al. (2019) used logistic regression, Naive Bayes and random forest algorithms for the detection of credit card fraud with an original data set and compared the performance of these methods according to precision, sensitivity and accuracy values. Performance analysis showed that the most successful method among the methods used was the random forest algorithm. Walke (2019) compared the performance of supervised learning techniques and unsupervised learning techniques in solving the fraud detection problem. As a result of this study, it was observed that supervised learning techniques were more successful than unsupervised learning techniques. Askari & Hussain (2020) proposed a hybrid algorithm based on fuzzy logic and decision tree for fraud detection in online transactions. Parmar et al. (2020) used many different classification algorithms in detecting credit card fraud and tested the performance of these algorithms with the accuracy rate and F-score obtained from the confusion matrix. As a result of this analysis, it was concluded that the best results were obtained with the K-Nearest Neighbor method and the worst results were obtained with the Logistic Regression method. Roseline et al. (2022) performed pattern recognition on the card transaction database to detect credit card fraud and used machine learning algorithms to identify suspicious transactions. In this study, the class imbalance problem was addressed and machine learning algorithms such as Naive bayes, SVM, ANN and LSTM were used. Yi et al. (2023) presented a machine learning method integrated with the Egret Swarm Optimization Algorithm (ESOA), a meta-heuristic algorithm for financial fraud detection. Huang et al. (2024) proposed a machine learning-based K-means clustering method to improve the accuracy and efficiency of financial fraud detection.

When the literature studies are examined, it is seen that fraud detection is considered as a binary classification problem in many studies and machine learning-based algorithms are widely used to solve the problem. At the same time, a significant part of the literature studies focuses on the problem of credit card fraud in sectors such as banking and insurance. On the other hand, studies on fraud detection in the retail and e-commerce sector are limited. In this direction, this study aims to develop a hybrid model called XGBoost-GA to improve the classification accuracy for cashier fraud detection in retailing.

3. METHODOLOGY

In this section, extreme gradient boosting (XGBoost), genetic algorithm (GA), the proposed hybrid XGBoost-GA model and the performance metrics used to compare classification models are introduced.

3.1. EXTREME GRADIENT BOOSTING (XGBOOST)

Extreme gradient boosting (XGBoost) is a high-performance classification algorithm based on decision trees. XG-Boost classifier introduced by Chen and Guestrin (2016). This model combines weak classifiers with stronger classifiers and at each iteration.

Suppose that $D = (x_i, y_i)$ denotes a data set with n samples and m attributes. Here, x_i denotes the input data and y_i denotes the class label for the ith sample. The predicted class labels can be calculated using Equation (1):

$$\hat{y}_i = \Sigma_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

In Equation (1), $\hat{y}_i$ indicates the predicted class label for the ith sample, K is the number of trees, f_k (x_i) denotes the predicted score of the kth tree, F is denoted by the space of all regression trees.

The objective function of XGBoost is described using Equation (2) (Chen et al., 2018).

$$\begin{gathered} F_{obj}(\theta) = L(\theta) + \Omega(\theta) \\ where\ L(\theta) = l(\hat{y}_i, y_i),\ \Omega(\theta) = \alpha T + 1/2\varepsilon w^2. \end{gathered} \tag{2}$$

In Equation (2), the objective function consists of two components. The first component reflects is differentiable convex loss function, while the second term is a regularized term that penalizes complex models. Additionally, T

indicates the number of leaves in the tree, α denotes the learning rate, ε is a regularized parameter and w is the weight of the leaves.

The objective function can be rewritten using Equation (3).

$$L(\theta) = \Sigma_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)\right) + \Omega(\theta) \tag{3}$$

The optimization goal is to construct a tree structure that minimizes the objective function (prediction error) in each iteration.

3.2. GENETIC ALGORITHM

Genetic algorithm (GA) is a population-based stochastic metaheuristic algorithm. This algorithm was inspired by Darwin's theory of evolution and was first introduced by Holland (1975). Genetic algorithm tries to find good solutions for NP-Hard optimization problems in a reasonable time by using parent selection, crossover and mutation operators. In this direction, GA is widely used in solving the parameter optimization problem of classification algorithms.

The basic principle of GA is the survival of stronger individuals to reach better solutions and the creation of better individuals from these individuals using crossover and mutation operators. In this algorithm, each individual represents a candidate solution. The steps of basic GA can be summarized as follows:

Step 1: Creating the coding structure that represents individuals Depending on the decision variable of the optimization problem, binary, discrete, permutation or real value coding can be used.
Step 2: Generating the random initial population
Step 3: Calculating the fitness value of each individual
Step 4: Selection of parents to create new individuals
Step 5: Creating new individuals using crossover and mutation operators
Step 6: Repeating Steps 3-5 until stopping criterion such as maximum number of iterations, maximum duration or target objective function value is satisfied.

3.3. PROPOSED APPROACH: XGBOOST-GA HYBRID MODEL

The XGBoost algorithm, like other classification algorithms, has a large number of parameters. The optimization of these parameters significantly improves the classification accuracy compared to the default parameters. In the parameter optimization problem, the parameters of the XGBoost model are decision variables.

These variables are presented in Table 1.

Performance metrics for classification accuracy can be used as the objective function of the optimization model. The performance metrics of the classifier can be calculated using the confusion matrix, whose general form is presented in Table 2.

Table 1. The parameters of XGBoost used in parameter optimization

Parameter	Data Type	Description
alfa (a)	Real	Learning rate
gamma (g)	Integer	The minimum loss value to make a partition on a leaf node of the tree
max depth (*md*)	Integer	The maximum number of splits
min child weight (*mcw*)	Integer	The minimum sum of sample weight in a child node
max delta step (*mds*)	Real	A measure of regularization
subsample (*ss*)	Real	A parameter used to control overfitting
n estimators (*n*)	Integer	The number of trees

Table 2. General form of confusion matrix for binary classification problem

		Predicted Class	
		C_0	**~ C_0**
Actual Class	**C_0**	True Positive (TP)	False Negative (FN)
	~ C_0	False Positive (FP)	True Negative (TN)

At this point, the most important of these performance metrics and their calculation formulas are as follows.

$$Accuracy\ Rate = \frac{TP + TN}{TP + TN + FN + FP} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

In the parameter optimization problem, F-measure is used as the objective function due to the unbalanced distribution of the class label. The genetic algorithm is used to solve the parameter optimization problem. The pseudocode of the proposed hybrid XGBoost-GA model is presented in Table 3.

Table 3. The pseudo-code of the proposed XGBoost-GA hybrid method

1: **Load** the dataset
2: **Divide** data into training and testing datasets
3: **Initialize** GA parameters (max number of iteration=1000, population size=20, elit ratio=0.1, mutation probability= 0.05, crossover probability=0.8, parents portion= 0.3)
4: **Define** a,γ, *md, mcw, mds, ss* and *n* parameters of XGBoost *randomly*
5: **Define** $Fmeasure_{best}$=0
6: **Set** *i*=1
7: While (*i* < max number of iteration)
8: **Calculate** F-measure of XGBoost by using training data
9: **Set** fitness function=$MAPE_i$
10: Calculate $F\text{-measure}_i$
11: **if** ($Fmeasure_i$ >$Fmeasure_{best}$) **then**
12: **Update** $Fmeasure_{best}$= $Fmeasure_i$
13: **Update** a,γ, *md, mcw, mds, ss* and *n*
14: **end if**
15: i=i+1
16: **end while**
17: Create a NGBM(1,1) model with finalized a,γ, *md, mcw, mds, ss* and *n* parameters
18: Calculate Fmeasure value of the testing data

4. APPLICATION

In this section, the application steps and performance analysis of the proposed approach are presented with dataset taken from a retail chain in Turkey.

4.1. DATA SET

In this study, real-life data from a retail chain in Turkey is used to detect cashier fraud. This dataset consists of 10 attributes and 13520 cashier transactions. The attributes included in the data set are presented in Table 4.

Table 4. The overview of the attribute in the dataset

Attribute	Type	Distinct Value
Transaction Type	Nominal	3
City	Nominal	5
Time Period	Nominal	3
Gender	Nominal	2
Age	Numeric	31
Seniority	Numeric	87
Position	Nominal	7
Marital status	Nominal	3
Category	Nominal	5
IsFraud	Binary	2

The descriptive statistics of numeric and nominal attributes in the dataset are given in Table 5 and Table 6, respectively.

Table 5. Descriptive statistics of numeric attributes in the dataset

Attribute	Average	Std. Dev.	Minimum	Q1	Median	Q3	Maximum
Age (year)	30.559	6.766	18	26	29	35	52
Seniority (month)	51.320	43.790	4	16	44	67	226

Table 6. Descriptive statistics of nominal attributes in the dataset

Attribute	Possible Values and Percentages
Transaction Type	Price Check (60%), Cancel Line (21%), Cancel Receipt (19%)
City	Istanbul (44.5%), Ankara (6%), Izmir (6%), Adana (4.7%) and Other (38%)
Time Period	Midday (51.5%), Night (30.5%), Morning (18%),
Gender	Female (59%), Male (41%)
Position	Cashier (42%), Staff (26%) and Other (32%)
Marital status	Single (45%), Married (27%) and Uknown (22%)
Category	Food (69%), Fresh food (13%), Bazar (12%) , Textile (4%) and Electronics (2%)
IsFraud	Fraud (10%), Non-Fraud (90%)

4.2. DETECTION OF CASHIER FRAUD WITH PROPOSED APPROACH

This study aims to propose a hybrid model using extreme gradient boosting (XGBoost) optimized with the genetic algorithm (GA) for cashier fraud detection in the retail sector. Additionally, the performance of this proposed model is

evaluated by comparing it with XGBoost, logistics regression, naïve bayes, k-nearest neighbor (k-NN). The dataset is divided into 80% training data and 20% testing data for validation of the classification models.

In the parameter optimization model, F-measure is used as the fitness function of the GA due to the unbalanced distribution of the class label, as can be seen in Table 6. The classification models and the parameter optimization problem are programmed using the Python language and related packages. The developed codes are run on a PC with Intel ® Core™ i5-7200U CPU at 2.71 GHz, 8GB RAM, and Windows 10 Pro.

The parameters of genetic algorithm for the optimization model are as follows: the maximum iteration number is 1000, the population size is 20, elitism ratio is 0.1, the probability of crossover is 0.8 and the probability of mutation is 0.05. The default values are used for the other variables such as crossover and selection function. 10 independent replications are carried out using this parameter set. In the genetic algorithm, the F-measure converges very fast to a stationary point, as can be seen in Fig. 1. Fig. 1 shows the value of fitness function according to the number of iterations.

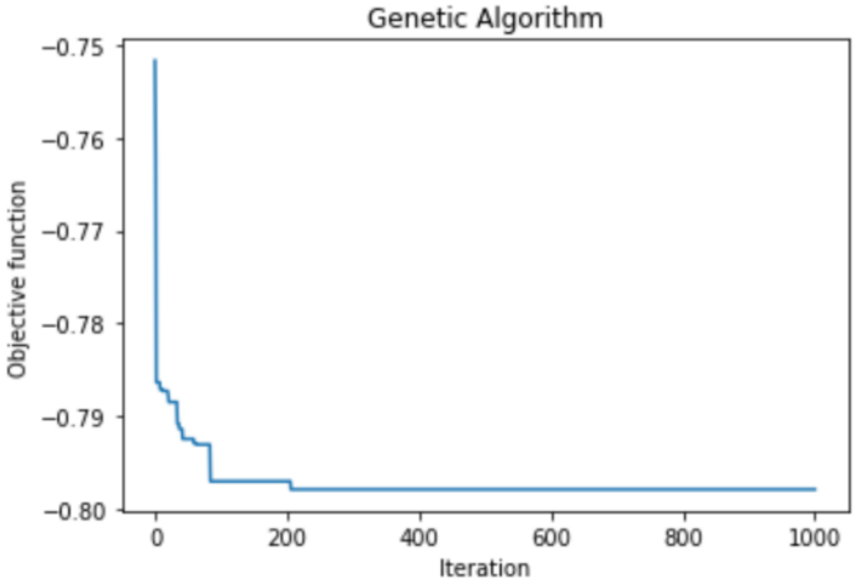

Figure 1. Parameter optimization process with genetic algorithm

In the proposed XGBoost-GA model, the F-measure is found to be 79.8% for the training data. The parameters of the XGBoost model are alfa=0.9857, gamma=1, max depth=10, min child weight =3, max delta step=1, subsample=0.8869 and the number of trees=95.

The performance analysis of the classification models for training data are given in Table 7 and Fig. 2.

Table 7. Performance analysis of the classification models for training data

Model	F-measure	Accuracy	Precision	Recall
XGBoost-GA	0.798	0.962	0.817	0.780
XGBoost	0.752	0.948	0.826	0.691
Logistics Regression	0.090	0.907	0.658	0.048
Naive Bayes	0.393	0.916	0.628	0.286
k-NN	0.718	0.949	0.772	0.672

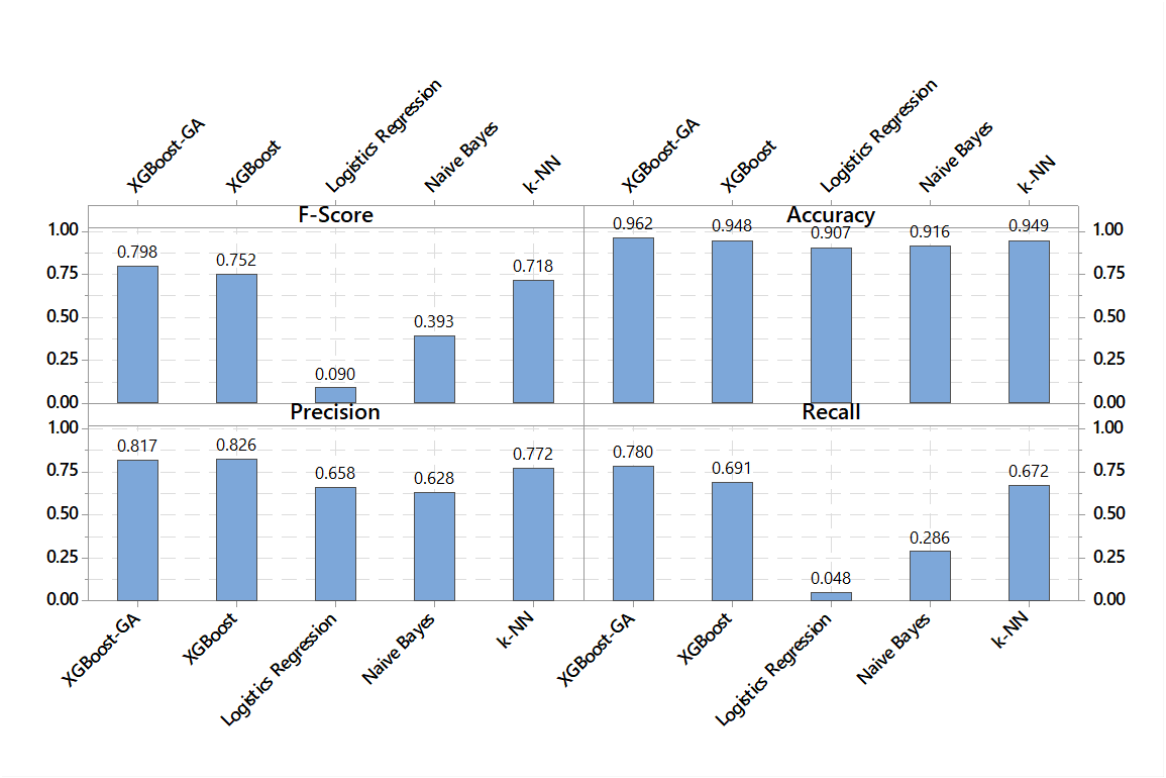

Figure 2. Performance metrics of the classification models for the testing data set

The performance analysis of the classification models for testing data are presented in Table 8 and Fig. 3.

Table 8. Performance analysis of the classification models for testing data

Model	F-measure	Accuracy	Precision	Recall
XGBoost-GA	0.733	0.959	0.718	0.749
XGBoost	0.712	0.951	0.780	0.655
Logistics Regression	0.050	0.901	0.389	0.027
Naive Bayes	0.381	0.912	0.608	0.278
k-NN	0.644	0.938	0.727	0.578

According to the results in Table 7, the proposed hybrid XGBoost-GA model has the maximum F-measure value of 79.8% whereas LR model has the lowest F-Measure of 5%. The numerical results also indicate that parameter optimization improves the classification accuracy of XGBoost model.

As can be seen from Table 8 and Fig. 3, the proposed XGBoost-GA model has a F-measure of 73.3% for validation data. The proposed hybrid XGBoost-GA model produces higher accuracy, recall, precision and F1-score than other classification algorithms such as XGBoost, logistics regression, naïve bayes, k-NN.

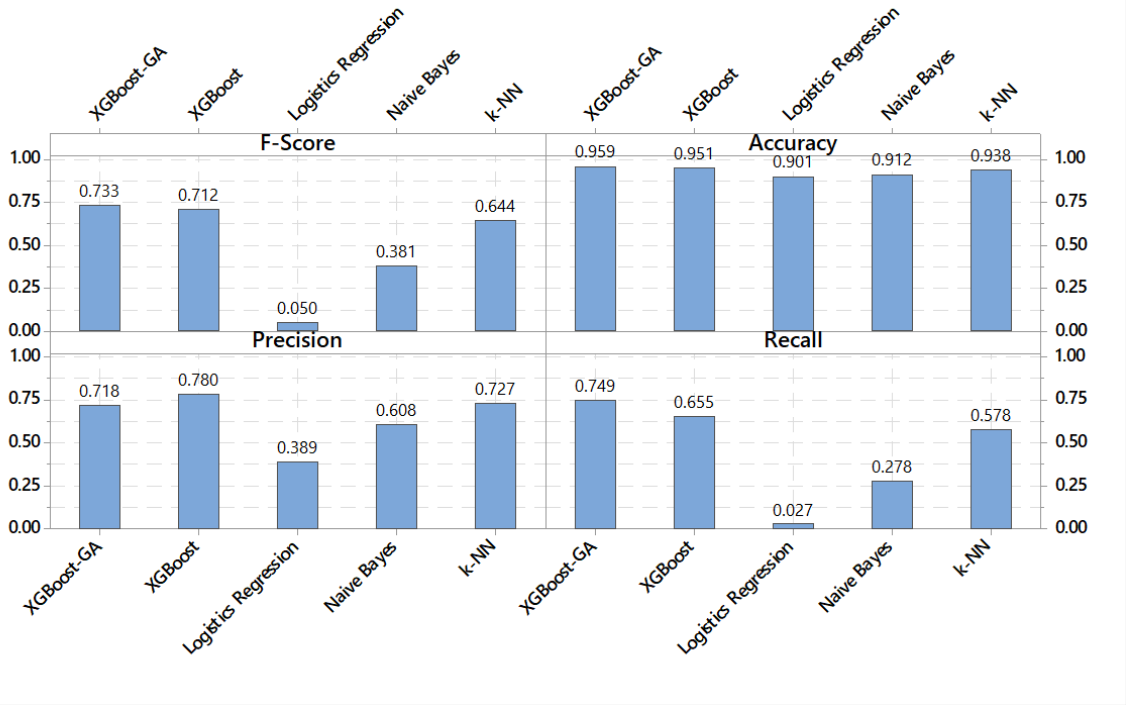

Figure 3. Performance measures of the classification models for the testing data set

5. CONCLUSIONS

With the digitalizing world, efforts to prevent and detect the increase in fraudulent activities have accelerated. While fraud detection performs digitally in sectors such as finance, banking and insurance, it is observed that applications in the retail sector have only just begun and analytical approaches are used to a limited extent in solving the problem. At the same time, while there are many studies in the literature on fraud detection for the financial sectors, there are very few studies for the retail sector.

Failure to detect and prevent fraudulent activities causes, businesses to experience significant financial losses. Gee and Button (2019) stated that financial losses resulting from fraud cases around the world are more than 80% of the UK's Gross Domestic Product. In another study published by ACFE (Association of Certified Fraud Examiners), 91 cases of fraud in the retail sector were examined and the median value of financial losses resulting from these cases was determined to be $85000.

One of the sources of fraud in the retail industry is store personnel. Accordingly, in this study, the cashier fraud detection problem is addressed with real-life data taken from a retail chain. To solve this problem, a hybrid approach is developed using extreme gradient boosting (XGBoost) and genetic algorithm (GA). In this approach, genetic algorithm is used to optimize the parameters of the XGBoost algorithm. The performance of the developed approach is compared with basic classification algorithms such as default XGBoost, logistic regression, naive bayes and k-nearest neighbor. Numerical results showed that the proposed approach has better performance than other classification algorithms for training and testing data. Also, with its high accuracy rate and F-measure, the proposed approach offers an effective solution for detecting cashier fraud in retail.

In future studies, the performance can be increased by adding new attributes to the proposed model. Additionally, approaches such as SMOTE can be used to solve the class imbalance problem. Different metaheuristic algorithms or Bayesian optimization can be used to solve the parameter optimization problem.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- A.D., P.V., Y.M., T.Ö.; Data Acquisition- A.D., T.Ö.; Data Analysis/Interpretation- A.D., T.Ö.; Drafting Manuscript- A.D., P.V., Y.M., T.Ö.; Critical Revision of Manuscript- A.D., P.V., Y.M., T.Ö.; Final Approval and Accountability- A.D., P.V., Y.M., T.Ö.; Supervision- T.Ö.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

ORCID IDs of the authors

Aytek Demirdelen 0000-0002-6005-4604
Pelin Vardarlıer 0000-0002-5101-6841
Yurdagül Meral 0000-0001-9244-1994
Tuncay Özcan 0000-0002-9520-2494

REFERENCES

Askari, S. M. S., & Hussain, M. A. (2020). IFDTC4. 5: Intuitionistic fuzzy logic based decision tree for E-transactional fraud detection. *Journal of Information Security and Applications, 52*, 102469.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018, January). XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud. In *2018 IEEE international conference on big data and smart computing (bigcomp)* (pp. 251-256). IEEE.

Erol, S. (2016). *Hile denetiminde proaktif yaklaşımlar* (Master's thesis, Sosyal Bilimler Enstitüsü).

ESMERAY, A. (2018). BİLİŞİM TEKNOLOJİSİNDEKİ GELİŞMELERİN MUHASEBE DENETİMİNE KATKISI. *Muhasebe Bilim Dünyası Dergisi, 20*, 294-309.

Gee, J., & Button, M. (2019). The financial cost of fraud 2019: The latest data from around the world.

Hanagandi, V., Dhar, A., & Buescher, K. (1996, March). Density-based clustering and radial basis function modeling to generate credit card fraud scores. In *IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr)* (pp. 247-251). IEEE.

Holland, J. H. (1975). Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.

Huang, Z., Zheng, H., Li, C., & Che, C. (2024). Application of Machine Learning-Based K-Means Clustering for Financial Fraud Detection. *Academic Journal of Science and Technology, 10*(1), 33-39.

Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by modified Fisher discriminant analysis. *Expert Systems with Applications, 42*(5), 2510-2516.

Nadim, A. H., Sayem, I. M., Mutsuddy, A., & Chowdhury, M. S. (2019, December). Analysis of machine learning techniques for credit card fraud detection. In *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)* (pp. 42-47). IEEE.

Niu, X., Wang, L., & Yang, X. (2019). A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv preprint arXiv:1904.10604*.

Parmar, J., Patel, A., & Savsani, M. (2020). Credit card fraud detection framework-a machine learning perspective. *International Journal of Scientific Research in Science and Technology, 7*(6), 431-435.

Pehlivanli, D., Eken, S., & AYAN, E. B. (2019). Detection of fraud risks in retailing sector using MLP and SVM techniques. *Turkish Journal of Electrical Engineering and Computer Sciences, 27*(5), 3633-3647.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19-50.

Renjith, S. (2018). Detection of fraudulent sellers in online marketplaces using support vector machine approach. *arXiv preprint arXiv:1805.00464*.

Roseline, J. F., Naidu, G. B. S. R., Pandi, V. S., alias Rajasree, S. A., & Mageswari, N. (2022). Autonomous credit card fraud detection using machine learning approach. *Computers and Electrical Engineering, 102*, 108132.

Sahin, Y., & Duman, E. (2011, March). Detecting credit card fraud by decision trees and support vector machines. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp. 1-6).

Seyedhossein, L., & Hashemi, M. R. (2010, December). Mining information from credit card time series for timelier fraud detection. In *2010 5th International Symposium on Telecommunications* (pp. 619-624). IEEE.

Shen, A., Tong, R., & Deng, Y. (2007, June). Application of classification models on credit card fraud detection. In *2007 International conference on service systems and service management* (pp. 1-4). IEEE.

Shukur, H. A., & Kurnaz, S. (2019). Credit card fraud detection using machine learning methodology. *International Journal of Computer Science and Mobile Computing, 8*(3), 257-260.

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision support systems, 75*, 38-48.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1-5). IEEE.

Walke, A. (2019). Comparison of supervised and unsupervised fraud detection. In *Advances in Data Science, Cyber Security and IT Applications: First International Conference on Computing, ICC 2019, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part I 1* (pp. 8-14). Springer International Publishing.

Yi, Z., Cao, X., Pu, X., Wu, Y., Chen, Z., Khan, A. T., ... & Li, S. (2023). Fraud detection in capital markets: A novel machine learning approach. *Expert Systems with Applications, 231*, 120760.

How cite this article

Demirdelen, A., Vardarlıer, P., Meral, Y., & Özcan, T. (2024). Diagnosis of internal frauds using extreme gradient boosting model optimized with genetic algorithm in retailing. *Acta Infologica, 8*(1), 60-70. https://doi.org/10.26650/acin.1475658