Enhancing loan fraud detection process in the banking sector using data mining techniques

Fahd Sabry Esmail[1], Fahad Kamal Alsheref[2], Amal Elsayed Aboutabl[3]
[1]Department of Business Information Systems, Faculty of Commerce and Business Administration, Helwan University, Helwan, Egypt
[2]Department of Information Systems, Faculty of Computing and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
[3]Department of Computer Science, Faculty of Computing and Artificial Intelligence, Helwan University, Helwan, Egypt

Article Info

Article history:

Received Jun 20, 2023
Revised Jul 5, 2023
Accepted Jul 30, 2023

Keywords:

Bank loan
Data mining
Fit statistics
Fraud detection
Loan fraud

ABSTRACT

Ongoing loan fraud is a source of concern for financial institutions, as it has a direct financial impact and also scares off customers. This pattern, which can be traced to the development of modern technology, the introduction of novel ideas, and the quickening pace of international connections, makes the detection of fraud an expensive endeavour. This article proposes a novel framework for enhancing the fraud detection of loan banking using data mining algorithms. The framework extracts a number of predictive analysis techniques for identifying loan fraud. Several methods employing a wide range of pipeline architectures have been tried in order to select the optimal champion model. Autotuning has also been used to find the best possible setting for the model's hyperparameters. The results of the evaluation show that autoencoder with gradient boosting outperformed the other classification algorithms with an accuracy of 98.62%. The proposed framework has the potential to significantly improve the fraud detection process of loan banking, which can ultimately lead to better faster fraud detects rates by combining data mining techniques with dimensionality reduction strategies in the feature space.

Corresponding Author:

Fahd Sabry Esmail
Department of Business Information Systems, Faculty of Commerce and Business Administration
Helwan University
Helwan, Cairo governorate, Egypt
Email: fahd.Sabry21@commerce.helwan.edu.eg

1. INTRODUCTION

Recently, fraud has evolved in both organisations and the banking sector. This is because of the developments in information technology, whose escalating waves have caused a great deal of chaos in a variety of enterprises. Financial crime and fraud are not legitimate ways for corporations and organisations to conduct daily operations. Fraudulent activities are always growing, and both their cost and clients' expectations are rising along with them. Financial losses have been incurred as a result of fraud, which also raises the cost of investigations and legal actions, erodes client trust, and damages brand reputation. It really is the corporate world's biggest nemesis.

The number of loan applications has increased recently since so many people rely on them for various reasons, [1], [2]. According to research, defaulters who refuse to repay the money they have taken do not allow individuals to get loans from banks or through other channels. This one action denied potential recipients the chance to get a loan [3], [4]. Credit can result from banks' inability to handle creditors' debts efficiently [5], [6]. A credit crisis occurs when reckless and inappropriate lending continues over an extended period of time, resulting in losses for banks and lending organisations [5]. The appraisal of loan defaulters and risk has received attention recently since credit risk in the banking sector is one of the most significant challenges [7], [8].

Sudhamathy and Venkateswaran [9] offers a framework that may be used to estimate a bank loan applicant's likelihood of default. The metrics generated by the prediction demonstrate the great accuracy and precision of the built-in model. To improve the accuracy of fraud detection, Carcillo *et al.* [10] created a hybrid technique that incorporated supervised and unsupervised methods. Unsupervised anomaly ratings computed at different levels of granularity and using an actual, labelled credit card fraud identification dataset are reviewed and appraised. The efficiency of the combination, which also improves identification precision, is supported by experimental findings.

Worryingly, the loan default rate causes banks to lose money to borrowers [11]. Where the banks face challenges with loan defaults and fraud, for addressing these issues requires a thorough analysis of historical data. This analysis should identify new and relevant classifications and predictions to aid in loan facilitation and decision-making. Credit risk assessment involves both structured and unstructured management decisions. Structured decisions rely on known loan-granting processes and computational tools, while non-structured decisions rely on managers' intuition and experience, often involving subjective elements. To improve the accuracy of loan facilitation decisions and minimize risks, banks need to develop more sophisticated and objective analytical tools and techniques, such as data mining, machine learning, and other advanced analytics methods. In order to forecast fraud in bank loan administration and prevent loan defaults, past loan data can be used through data mining, which can reveal hidden patterns that would not have been discovered through manual examination by a credit officer. Traditional and statistical approaches have limited accuracy potential in this direction, as a person cannot effectively assess a credit history given the volume and variety of data. Therefore, method based on cases, analogies, and statistics have been applied.

The study was motivated by the lack of a comprehensive framework capable of handling different types of data, including classified and unclassified data. To address this gap, the research team aimed to develop an integrated framework that could incorporate various feature selection techniques, mining, and machine learning algorithms. The objective of this approach was to provide financial decision-makers with a powerful tool for identifying and preventing loan fraud, which is a significant issue in the financial industry. Ultimately, the study aimed to advance the goal of promoting integrity and transparency in the financial sector by contributing to the development of a more effective framework for fraud detection and prevention. The article also includes novel findings that either corroborate or conflict with the present study's findings and other pertinent literature in this specific domain.

2. RELATED WORK

In this section, various sequential models and data mining techniques for fraud detection are reviewed. Numerous credit applications and their transaction histories are examined. The challenge of binary classification mostly arises while categorising various credit-related transactions since they might either be legal transactions or fraudulent transactions.

In predictive analytics, data mining is a crucial step. The extraction of information or specifics from a vast amount of data is known as "data mining". Although data mining is a component of knowledge discovery in databases (KDD). Data mining functions try to figure out the many patterns that come up in data mining jobs. To create models from datasets, data mining techniques are used, and the datasets represent a collection of details. Data mining algorithms learn from datasets, or they learn to anticipate the crucial consequence of a certain input [12]. This type of knowledge acquisition has no impact on the workstations' ability to hold onto data, but it does change how they operate so that future improvements may be made [13].

Sudhakar *et al.* [14] proposed a practical prediction model for identifying reliable customers who have requested bank loans. To forecast the characteristics important for believability, decision trees (DT) are used. Client loan requests may or may not be approved using this prototype method. To forecast the state of loans, the model presented in Hamid and Ahmed [15] was constructed using information from the banking industry. J48, Bayes Net, and naive Bayes are the three classification algorithms used in this model. Weka is used to put the model into use and verify it. Based on its accuracy, the best algorithm, J48, was chosen [16]. Implements an enhanced risk prediction clustering multi-dimensional algorithm to identify unsuitable loan applicants. To prevent duplication, the association rule is integrated into this work's usage of risk assessments at the primary and secondary levels. Zamani and Mogaddam [17] utilized a DT model as a classifier and implemented a genetic approach for selecting features. Bekhet and Eletter [18] conducted a study that led to the creation of two data mining models designed to assist Jordanian banks in determining the optimal credit amount to offer borrowers for credit scoring purposes. In terms of accuracy rate, it is demonstrated that the regression model outperforms the radial function model. Blanco *et al.* [19] developed a number of multilayer-based credit scoring models. According to the research, it outperforms models that have used logistic regression (LR) methods. The outcomes show that the neural network model performs better than the other three methods. Another approach was proposed by Harris [20] credit scoring models built using a variety of default definitions and based on support-vector machines are compared. The study found that the performance of the broad-definition models outperformed that of the restricted-definition models. Using methods like bayes classification, the bagging method, DT, random forests (RFs), boosting, and other methods, financial data analysis is carried out in Desai and Kulkarni [21]. All of these methods-multilayer perceptrons, LR, DT, support vector machines (SVM), and neural networks (NN) are incorporated in this model. The usefulness of using the aforementioned methods to score credit is investigated. Based on the analysis's findings, the performance is exceptional. Islam and Habib [22] proposed using a model to forecast potential business areas in retail banking. The study also used client records from a retail bank's business division and records from both rural and urban areas of Bangladesh. The primary transactional determinants of clients were broken down using these records, and a design for the retail bank's expected subdivisions was predicted. The DT and artificial neural networks (ANN) are the two categories of data mining that were selected. The genetic algorithm (GA) and principal component analysis (PCA) were then used as feature selection methods. The approaches were evaluated using the German and Sudanese credit datasets. In most cases, ANN outperforms DT, according to the classification's results. As a feature selection method, it was also discovered that GA is superior to PCA. Compared to the Sudanese data set, the German data set had an accuracy of 80.67%. Final thoughts: it was demonstrated that ANN performed better than DT and its hybrid models, GA-DT and PCA-DT. However, when it comes to taking into account changes in client behaviour and the ability of fraudsters to create new fraud patterns, the process becomes more challenging. Models for identifying fraud can help in this case by finding anomalies with the use of unsupervised learning approaches.

3. THE PROPOSED METHOD

3.1. Steps of the proposed framework

As shown in Figure 1 of the proposed framework, several data mining techniques have been implemented, such as artificial NN, bayesian networks (BN), RFs, DT, gradient boosting, autoencoder with gradient boosting, PCA with gradient boosting, and ensemble models. Also, a number of performance indicators, including cumulative lift, lift, accuracy, receiver operating characteristic (ROC) separation, and F1 score, have been used to identify the most effective data mining algorithms. By employing preprocessing, modification, modeling, and evaluation phases, the issue of detecting fraudulent entities was resolved.

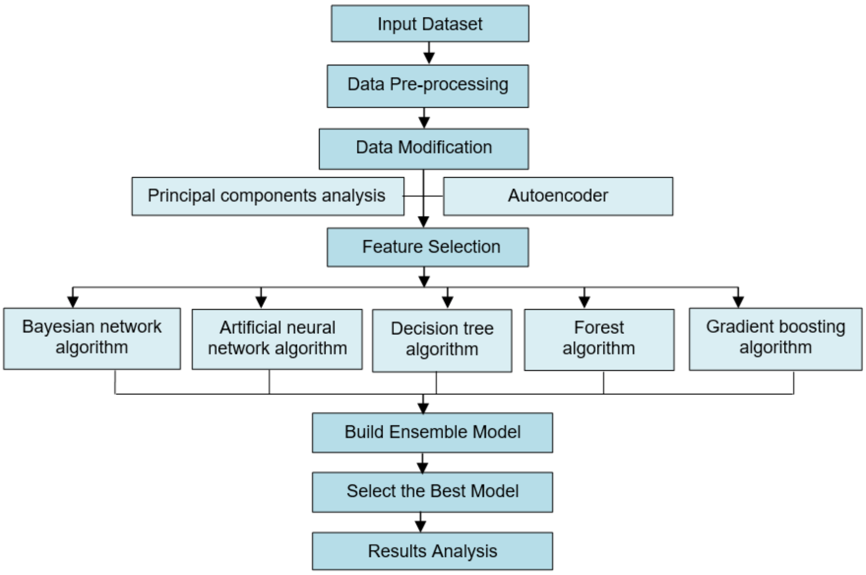


Figure 1. Steps of the proposed framework

In the preprocessing phase, the data was cleansed and enhanced to meet the necessary quality standards, followed by the modification step, where techniques such as PCA and autoencoder were employed to decrease the dimensionality of the initial feature space. Subsequently, models were developed and their effectiveness was evaluated, leading to the discovery of the most efficient model and a set of fundamental features for detecting fraudulent entities. The k-nearest neighbours method was used to fill in any gaps in the data, statistical analysis was used to weed out any outliers, and the support vector machine method was used to remove any abnormalities.

3.2. Description of the dataset

The original source of the dataset is the kaggle repository for data scientists and people interested in machine learning. Data from transactions made possible by credit loans is included in the dataset. To compare accuracy, five primary data mining techniques are probably employed. The dataset contains 100,000 unique customer records pertaining to transactions. All of the dataset's attributes are accurate and integer, and it contains a multivariate feature. The continuous (numerical) variable's input were an autoencoder and PCA. A total of 10 distinct input characteristics are used to train and evaluate the model, as shown in Table 1. To obtain two distinct groups of distribution, a hybrid oversampling and undersampling strategy is utilized to preprocess an imbalanced dataset. Statistical analysis system (SAS) enterprise miner is the experimental configuration used to detect fraud in credit loans.

Table 1. Feature descriptions for loan fraud detection

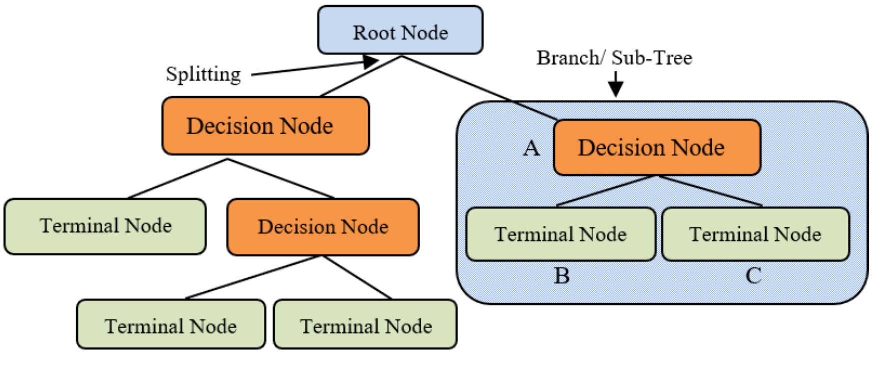
Feature	Description
Credit score	Creditworthiness
Annual income	yearly income of the borrower
Term	The term that apply to money borrowing like short term, long term
Current loan amount	Amount of the current loan
Purpose	The purpose of a loan like home improvements, debt consolidation, buy a car, buy house, educational expenses medical bills, wedding, vacation, small business, major purchase, and other
Maximum open credit	The highest credit limit
Years of credit history	Length of credit history
Current credit balance	The total amount of money you currently owe on your credit
Home ownership	Home ownership like home mortgage, own home, and rent
Number of credit problem	Numerous credit issues

4. RESEARCH METHODS

Data mining involves analyzing large datasets to identify anomalies, trends, and correlations for predicting outcomes. It is utilized to reduce risks and detect loan fraud through various approaches. In this study, five data mining algorithms were applied and analyzed to gain a deeper understanding of their efficacy and relevance in detecting fraudulent activities, resulting in valuable insights and patterns that facilitate informed decision-making.

4.1. Decision tree

The DT are the most popular method for solving classification issues in data mining. This approach employs supervised learning, where the model is trained to identify the object type represented by the data, depending on the nature of the data provided. A tree's accuracy greatly depends on the choice of strategic splits. To promote homogeneity, DT divide nodes into sub-nodes using a variety of techniques, including Gini. With relation to a target variable, a node's purity increases [16]. Based on all available characteristics, the aim is to choose the split that yields the most homogenous sub-nodes see Figure 2.



Note: A is parent node of B and C

Figure 2. Decision tree

4.2. Random forest

The RF is an aggregate classifier that combines several DTs. The purpose of using numerous trees is to properly train them such that each one contributes to structure of a model has been illustrated in Figure 3. After the tree has been built, the findings will be merged. Both classification and regression issues may be resolved with RFs [23].

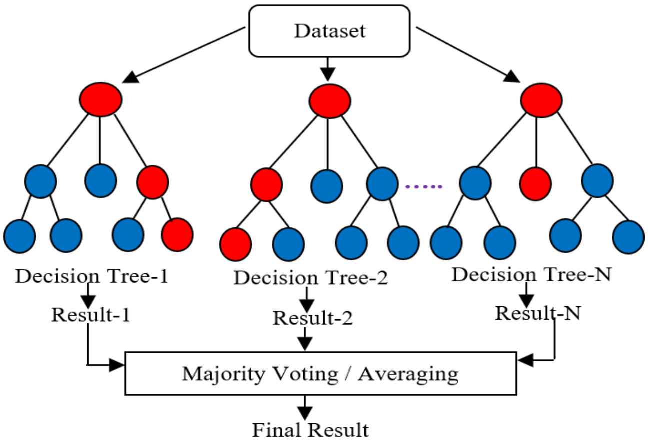


Figure 3. Random forest

4.3. Multilayer perceptron classifier

The MLP is an ANN model that utilizes a feedforward approach, where the input data is propagated to several pertinent outputs. The system comprises three levels, namely, input, output, and hidden layers, with the input layer receiving the signal for processing [22]. The backpropagation algorithm for MLP training was shown in Figure 4 for categorising indivisible datasets, the hidden layer is necessary.

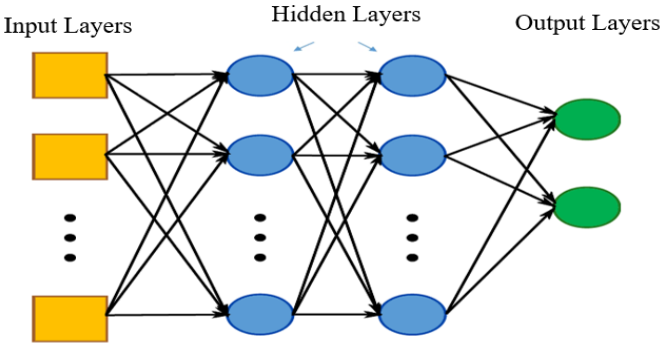


Figure 4. MLP neural network

4.4. Bayesian network

A BN is a probabilistic visual model that assesses the conditional dependency structure of a collection of random variables using the bayes theorem:

$$P\ (A|B) = \frac{P\ (B|A)\ P\ (A)}{P\ (B)} \tag{1}$$

assuming that A and B are events, P (A|B) represents the probability of A being true given that B is true, P (B|A) represents the probability of B being true given that A is true, and P (A) and P (B) denote the independent probabilities of A and B, respectively. A and B events are separate from one another. Bayesian classifiers can be taught more rapidly than other methods, but learning takes longer [8].

4.5. Gradient boosting

The gradient boosting is also known as the statistical prediction model. Although it allows for the expansion and optimization of differential loss functions, it nevertheless acts much in the same way as earlier boosting approaches. Regression and classification algorithms frequently include gradient boosting [24]. An approach for regression that resembles boosting is gradient boosting as shown in Figure 5.

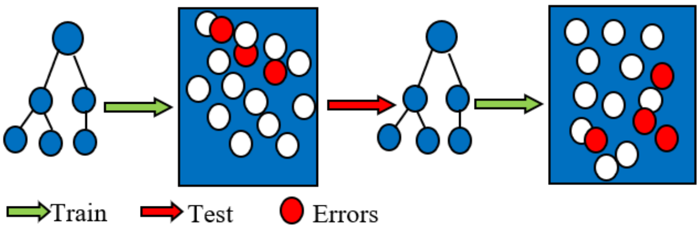


Figure 5. Gradient boosting

4.6. Feature selection and extraction

Feature selection and extraction are techniques used in machine learning to identify and select the most relevant features or attributes from a dataset. A proper study of when to employ PCA and autoencoder methods were necessary, two of the most popular dimensionality reduction methods used by machine learning researchers. Autoencoder and PCA both work for linear and non-linear surfaces, but this is the first and most noticeable difference between the two methods.

4.6.1. Autoencoder deep learning

Autoencoder is an unsupervised deep learning technique. In this simple feed-forward network, there are precisely as many inputs as outputs. The lower dimension code can reproduce the input after obtaining it from the higher dimension code. It is also known as a latent space representation since it is essentially a compressed knowledge representation. Since it is primarily used for input reconstruction, the training must be as exact as possible [25]. There are encoders and decoders in the architecture as shown in Figure 6.

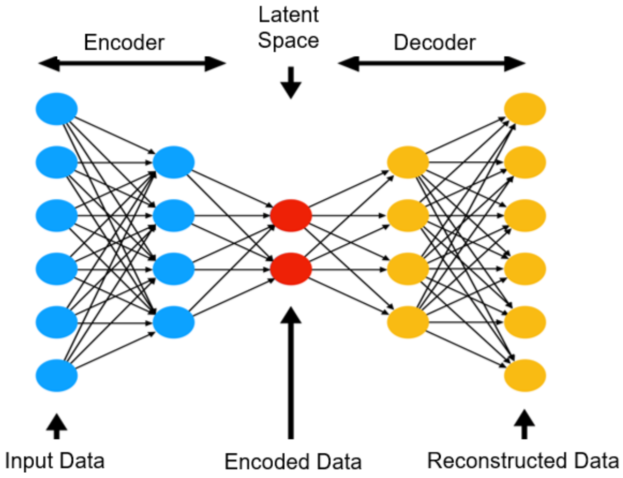


Figure 6. A typical autoencoder

4.6.2. Principal component analysis

The PCA is a technique for reducing the number of dimensions. PCA narrows down the lower order dimensions to a relatively narrow range. The features or dimensions are employed need to be kept to a minimum in order to minimize overfitting, which also aids in identifying the correlation between the variables. When lowering dimensions, it is crucial to remember that we shouldn’t sacrifice information in order to cut back on features. In other words, information must not be lost throughout the process. PCA bears this in mind and does a great job of reducing the number of dimensions and retrieving information [17]. Figure 7 depicts the idea of the PCA algorithm.

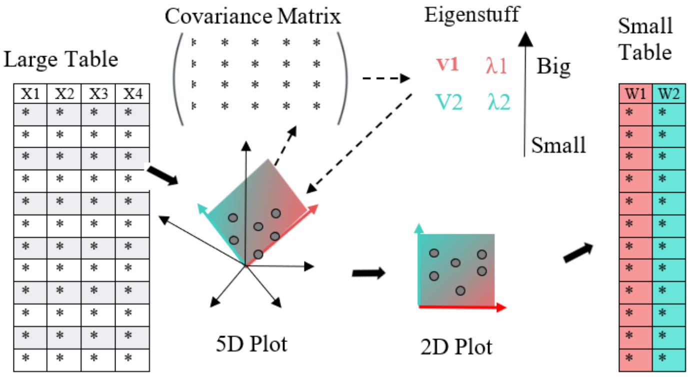


Figure 7. The idea of PCA algorithm

4.7. Ensemble learning model

Ensemble techniques are strategies for creating several models and combining them to achieve better results. Ensemble models employing majority voting predict the outcome of each test case using each model, and the ultimate output prediction is selected based on the highest number of votes received. The ensemble creates a novel model by utilizing a function of posterior probabilities (for class objectives) or projected values (for interval targets) derived from multiple models. The majority voting algorithm operates under the Algorithm 1.

Algorithm 1. Ensemble learning model for majority vote

```
Step1. Use the 5 classifiers BN, MLP, DT, GB, and RF.
Step2. Compare the output of the five classifiers.
Step3. Apply majority voting to each instance that has been validated.
Step4. Vote by majority for each instance that is valid.
```

5. RESULTS AND DISCUSION

The following methods were used to solve the classification issue: gradient boosting, DT, ANN, BN, RFs, and ensemble methods of NN and DT. The SAS enterprise miner was used to implement the proposed solution to the classification problem. The PCA and the autoencoder are two of the techniques that may be used by SAS enterprise miner's "feature extraction" node to build new features. All two building methods were applied in order to achieve the best results. The outcomes were compared after that. The hyperparameters were fitted using the validation set after training the models on the training data. Table 2 presents a performance measure for several algorithms.

Table 2. The performance measure for several algorithms

Model name	Accuracy	Precision	Recall	Area under ROC
Random forest	0.8276	0.6935484	0.056367	0.827442
Decision tree	0.8241	0.5430917	0.074345	0.752058
Bayesian network	0.8033	0.4476386	0.449064	0.759582
Neural network	0.822	0	0	0.5
Gradient boosting	0.8646	0.7229588	0.388015	0.897017
PCA with gradient boosting	0.98523	0.99497	0.99017	0.899
Autoencoder with gradient boosting	0.98616	0.99571	0.99036	0.917

Table 2 shows that PCA with gradient boosting and autoencoder with gradient boosting achieved the highest accuracy rates of 0.98523 and 0.98616, respectively, with high precision and recall rates as well. These two algorithms outperformed the other models by a significant margin. The findings of this table indicate that using a combination of techniques, such as PCA or autoencoder with gradient boosting, can significantly improve the performance of algorithms for a binary classification task of loan fraud detection.

5.1. Cumulative lift

To evaluate the cumulative lift, the partitions are arranged in descending order based on the forecasted probability of the target event P_Credit_Status fraud, which indicates the expected likelihood of the event

"Fraud" for the target name credit status. The data is divided into 20 quantiles (demideciles), each comprising 5% of the total data, and the count of events is obtained across all quantiles. Figure 8 showcases the cumulative lift value of various algorithms in the train and validation partitions. The cumulative lift for a specific quantile is defined as the proportion of events between each quantile up to and including the current one, relative to the number of events that would occur randomly or uniformly. It is also known as the ratio of the cumulative response to the baseline response percentage. The first two quantiles, which constitute the top 10% of the data and 10% of the occurrences at random, comprise the cumulative lift at depth 10. Consequently, the calculations of cumulative lift demonstrate that selecting instances in quantiles is significantly more likely than choosing them randomly.

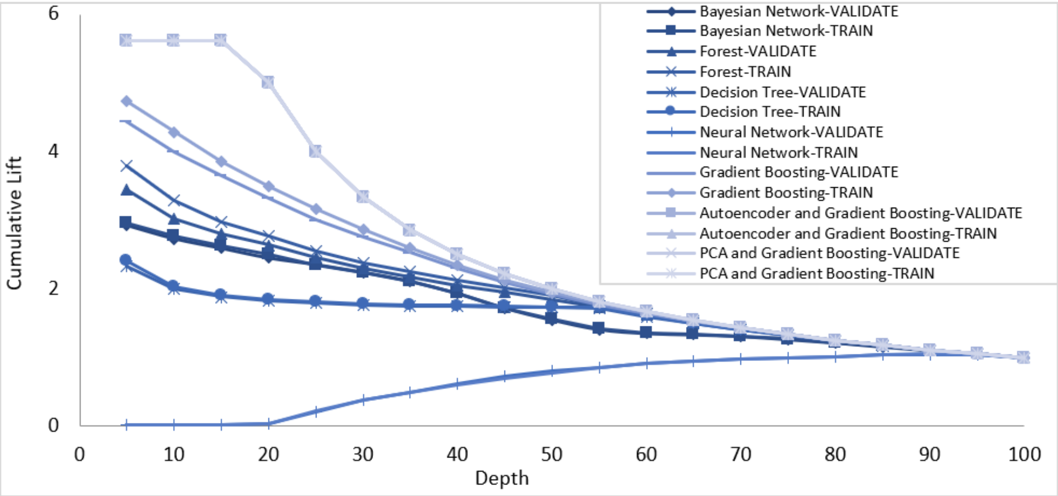


Figure 8. Cumulative lift values for the employed algorithms

5.2. Lift measure

The lift measure is obtained by sorting the partitions in decreasing order based on the expected likelihood of the target event, P credit status fraud, which signifies the probability of credit status fraud occurrence. The total number of occurrences in each of the 20 quantiles, or deciles, is determined, with each quantile representing 5% of the total data. Lift is the ratio of events in a particular quantile to events that would happen randomly or uniformly, and it represents the proportion of responses that deviate from the baseline response rate. Each of the 20 quantiles is expected to include 5% of the events. Therefore, the lift measures reveal the extent to which the potential instances of an event is distinct in each quantile compared to random instances. Figure 9 illustrates the lift measure values of different methods in the train and validation partitions.

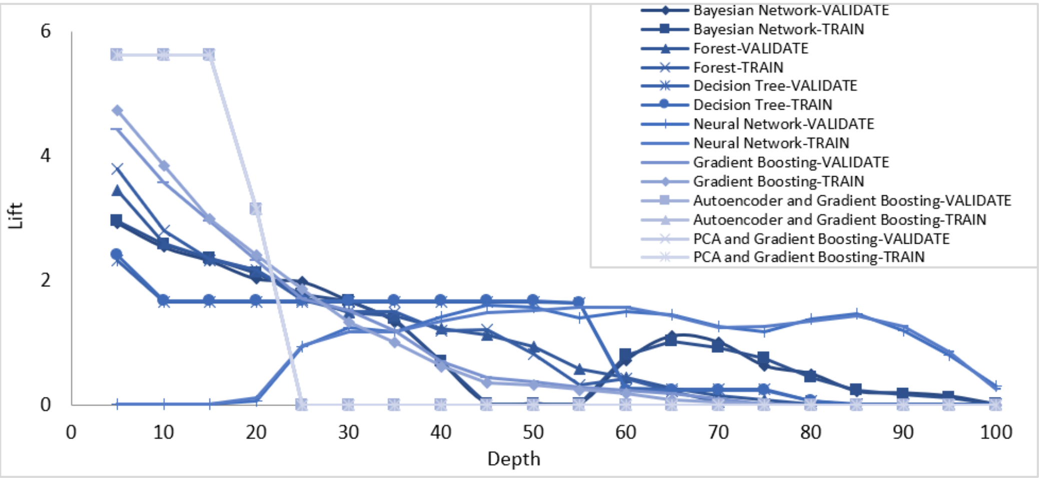


Figure 9. Lift value for the employed algorithms

5.3. Sensitivity assessment

The ROC curve, which is based on the confusion matrix, is a graph of sensitivity vs. specificity. Various cut-off numbers are used to compute these values. To aid in selecting the best data counting cut-off, the Kolmogorov-Smirnov (KS) reference line is set at the value of 1-specificity, which is the point of greatest divergence between 1-specificity and sensitivity in the validation partition. Figure 10 displays the sensitivity measure values of various algorithms in the training and validation partitions. Figure 10 illustrates the sensitivity measure values for various algorithms in both the train and validation partitions. The KS statistic measures the difference between the actual distribution functions of two models or between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. When the KS value is less than 0.05, the lack of fit’s importance becomes more evident.

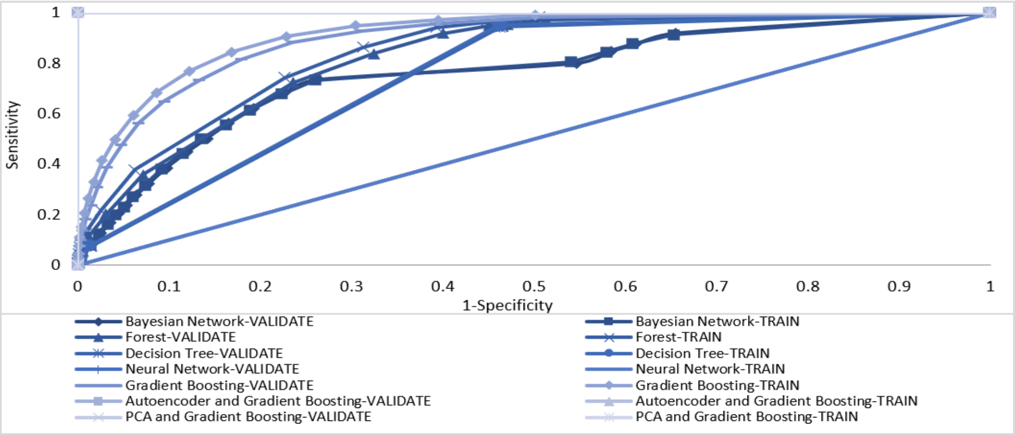


Figure 10. Sensitivity value for the employed algorithms

5.4. Accuracy

The accuracy measure, which is evaluated at various cut-off values, is the proportion of instances that are accurately classified as either events or nonevents. Cut-off levels are measured in increments of 0.05 between 0 and 1. The forecast target category is taken into account at all cut-off values. If P credit status fraud occurs, the probability that the target credit status's event “yes” happens is greater than or equal to the cut-off value. When P credit status fraud exceeds or is equal to the cut-off amount, the projected category is the actual occurrence. Otherwise, it is unimportant. When both the original classifications and the forecast categorization are true negatives or true positives, respectively, the instances is correctly sorted. If the observed sorting does not match the actual categorization, the instances has been incorrectly sorted. Figure 11 displays the various accuracy measurement results for various algorithms in the training and validation partitions. For estimating accuracy, use the (2).

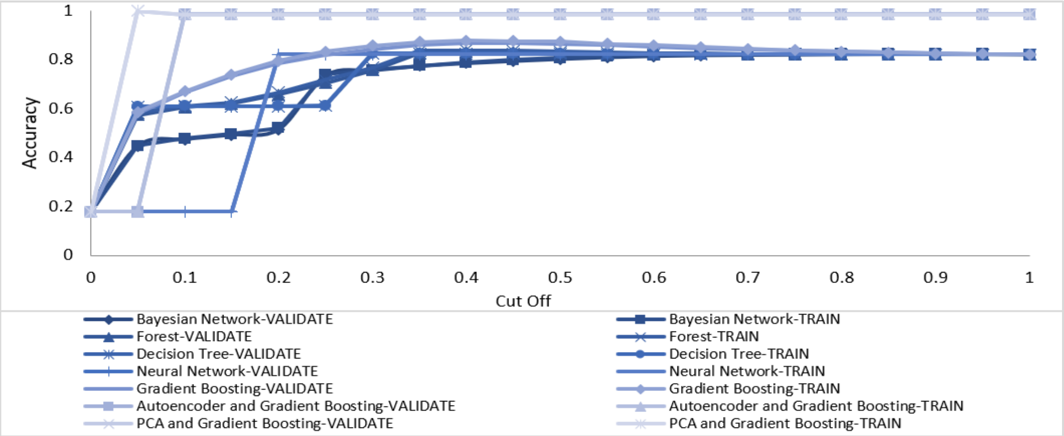


Figure 11. Accuracy value for the employed algorithms

$$Accuracy = \frac{true\ positive+true\ negative}{total\ instances} \tag{2}$$

5.5. F1 score

The F1 score is a classification criterion based on the confusion matrix assessed at various cut-off levels. It takes into account the recall and accuracy requirements (or sensitivity). In 0.05-unit increments, cut-off values vary from 0 to 1. If P credit status exists for all cut-off values, it is taken into account when categorising the prediction target. The predicted likelihood of the target fraud occurring is “yes” because it exceeds or equals the cut-off number; hence, the answer is yes. When P-credit status fraud occurs, an event is predicted to occur. Yes, it exceeds or equals the cut-off value in all other respects. For the various methods in the training and validation partitions, the F1 score measure's various values are shown in Figure 12.

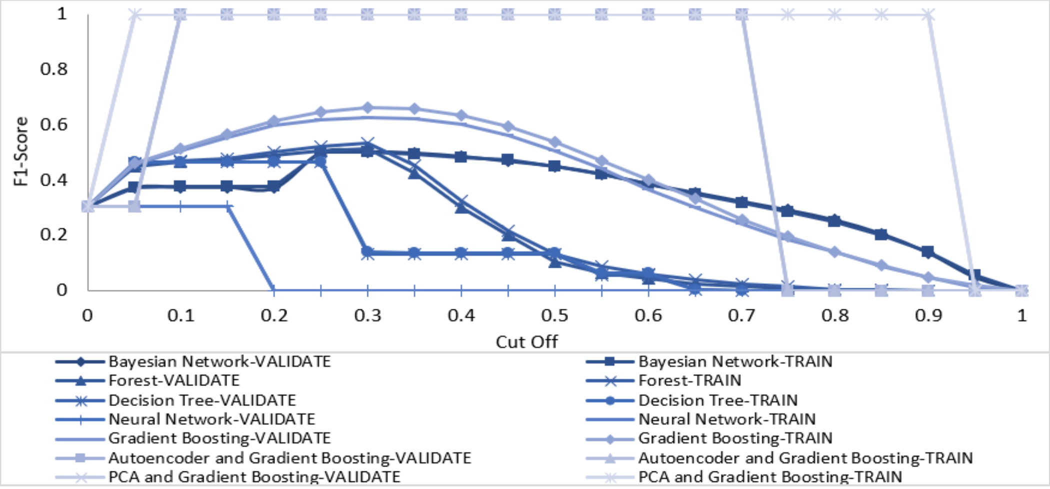


Figure 12. F1 score for the employed algorithms

5.6. Models fit statistics discussion

Many fit statistical criteria are used to choose the optimal or top model for deployment in production settings, which are shown in Tables 3-6. The average square error, misclassification rate, Gini coefficient, and KS are a few examples of these metrics. The Gini coefficient, ranging from 0 to 1, is a measure of equality, with 0 indicating absolute equivalence and 1 indicating perfect discrimination [14]. In the two-partition dataset, the neural network represents a superior model that resulted in a smaller Gini coefficient. The misclassification rate is a performance metric that indicates the proportion of incorrect predictions, without distinguishing between false positives and false negatives [15]. Better than other models, the autoencoder and gradient boosting model in the validated dataset partition has a low rate of misclassification. For average square error (ASE), while 0 indicates a flawless model, the lower the number, the better [16], [17]. The best performing model in the validation dataset partition is the PCA and gradient boosting model, which has the ideal ASE value. The KS test is utilized to identify regularly distributed samples, with the greatest difference between cumulative distributions being 1 for events and 0 for non-events. Autoencoder with gradient boosting and PCA with gradient boosting in the two-partition dataset produce a superior model outcome.

Table 3. Gini coeffcient

Model name	Train	Validate
NN	0	0
GB	0.830168523	0.794034464
Autoencoder+GB	1	1
PCA+GB	1	1
DT	0.512648002	0.504116585
Forest	0.693408266	0.654883206
Ensemble	0.526565129	0.519321977
BN	0.522007126	0.51916413

Table 4. Misclassifcation rate

Model name	Train	Validate
NN	0.470414286	0.4704
GB	0.1672	0.1827
Autoencoder+GB	0.002728571	0.002866667
PCA+GB	0.002642857	0.0029
DT	0.230357143	0.231466667
Forest	0.221457143	0.227566667
Ensemble	0.231657143	0.233033333
BN	0.464042857	0.471033333

Table 5. Average square error

Model name	Train	Validate
NN	0.115693994	0.115690301
GB	0.038851457	0.042182419
Autoencoder+GB	0.013988966	0.014053534
PCA+GB	0.002126389	0.002215149
DT	0.055028232	0.055348286
Forest	0.0498462	0.051928703
Ensemble	0.07019606	0.070354129
BN	0.103084497	0.104346093

Table 6. Kolmogorov-smirnov

Model name	Train	Validate
NN	0	0
GB	0.678390997	0.649586284
Autoencoder+GB	1	1
PCA+GB	1	1
DT	0.487690439	0.482071225
Forest	0.549830699	0.519612802
Ensemble	0.487690439	0.481818348
BN	0.47284117	0.471914669

6. CONCLUSION

Borrower repayment defaults have negatively impacted the lending industry's sustainability and led to an increase in bank loan fraud. Research indicates that current risk evaluation methods may miss important risk signals affecting repayment. By implementing a real-time transaction monitoring system, banks can prevent and discourage fraud, reducing risks and costs while safeguarding their reputation. An ensemble model integrating several data mining approaches can identify the root causes of loan fraud, with autoencoder and gradient-boosting classifier being the most accurate models. Performance criteria such as accuracy, lift, cumulative lift, and F1 score were used to identify the most effective data mining methods, while statistical metrics such as average square error, misclassification rate, Gini coefficient, and KS were used to select the optimal model. The autoencoder with gradient-boosting classifier was found to be the best model for detecting loan fraud, with the potential to significantly improve banking and fraud detection categorization techniques.

REFERENCES

[1] S. Oreski and G. Oreski, “Genetic algorithm-based heuristic for feature selection in credit risk assessment,” *Expert Systems with Applications*, vol. 41, no. 4 PART 2, pp. 2052–2064, Mar. 2014, doi: 10.1016/j.eswa.2013.09.004.

[2] C. Yannelis and G. Tracey, “Student loans and borrower outcomes,” *SSRN Electronic Journal*, vol. 14, pp. 167–186, 2022, doi: 10.2139/ssrn.4063097.

[3] P. Nascimento, “Modelling income contingent loans for higher education student financing in Brazil,” Thesis Doctor of Economics of the Federal University of Bahia (UFBA), 2018, [Online]. Available: https://repositorio.ufba.br/ri/bitstream/ri/28485/1/Tese de doutorado - Paulo Meyer Nascimento_repositorioUFBA.pdf.

[4] B. Chapman and M. Sinning, “Student loan reforms for German higher education: financing tuition fees,” *Education Economics*, vol. 22, no. 6, pp. 569–588, Nov. 2014, doi: 10.1080/09645292.2012.729327.

[5] N. Metawa, M. K. Hassan, and M. Elhoseny, “Genetic algorithm based model for optimizing bank lending decisions,” *Expert Systems with Applications*, vol. 80, pp. 75–82, Sep. 2017, doi: 10.1016/j.eswa.2017.03.021.

[6] F. S. Esmail, F. K. Alsheref, and A. E. Aboutabl, “Review of loan fraud detection process in the banking sector using data mining techniques,” *International Journal of Electrical and Computer Engineering Systems*, vol. 14, no. 2, pp. 229–239, Feb. 2023, doi: 10.32985/IJECES.14.2.12.

[7] G. L. I. Cyril and J. P. Ananth, “Deep learning based loan eligibility prediction with social border collie optimization,” *Kybernetes*, vol. 52, no. 8, pp. 2847–2867, Aug. 2022, doi: 10.1108/K-10-2021-1073.

[8] S. Akkoç, “An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: the case of Turkish credit card data,” *European Journal of Operational Research*, vol. 222, no. 1, pp. 168–178, Oct. 2012, doi: 10.1016/j.ejor.2012.04.009.

[9] G. Sudhamathy and C. J. Venkateswaran, “Analytics using R for predicting credit defaulters,” in *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*, 2017, pp. 66–71, doi: 10.1109/ICACA.2016.7887925.

[10] F. Carcillo, Y. A. Le-Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, “Combining unsupervised and supervised learning in credit card fraud detection,” *Information Sciences*, vol. 557, pp. 317–331, May 2021, doi: 10.1016/j.ins.2019.05.042.

[11] A. Byanjankar, M. Heikkila, and J. Mezei, “Predicting credit risk in peer-to-peer lending: a neural network approach,” in *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence*, 2015, pp. 719–725, doi: 10.1109/SSCI.2015.109.

[12] I. Imran, U. Zaman, M. Waqar, and A. Zaman, “Using machine learning algorithms for housing price prediction: the case of Islamabad housing data,” *Soft Computing and Machine Intelligence*, vol. 1, no. 1, pp. 11–23, 2021, doi: 10.22995/scmi.2021.1.1.03.

[13] I. H. Witten, E. Frank, and J. Geller, “Data mining: practical machine learning tools and techniques with java implementations,” *SIGMOD Record*, vol. 31, no. 1, pp. 76–77, Mar. 2002, doi: 10.1145/507338.507355.

[14] M. Sudhakar, C. V. K. Reddy, and A. Pradesh, “Two step credit risk assesment model for retail bank loan applications using decision tree data mining technique,” *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, vol. 5, no. 3, pp. 705–718, 2016.

[15] A. J. Hamid and T. M. Ahmed, “Developing prediction model of loan risk in banks using data mining,” *Machine Learning and Applications: An International Journal*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.5121/mlaij.2016.3101.

[16] K. Khadse, “Clustering loan applicants based on risk percentage using k-means clustering techniques,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6, no. 2, pp. 162–166, 2016.

[17] S. Zamani and A. Mogaddam, “Natural customer ranking of banks in terms of credit risk by using data mining: a case study: branches of mellat Bank of Iran,” *Jurnal UMP Social Sciences and Technology Management*, vol. 3, no. 2, 2015.

[18] H. A. Bekhet and S. F. K. Eletter, “Credit risk assessment model for Jordanian commercial banks: neural scoring approach,” *Review of Development Finance*, vol. 4, no. 1, pp. 20–28, Jan. 2014, doi: 10.1016/j.rdf.2014.03.002.

[19] A. Blanco, R. Pino-Mejías, J. Lara, and S. Rayo, “Credit scoring models for the microfinance industry using neural networks: Evidence from Peru,” *Expert Systems with Applications*, vol. 40, no. 1, pp. 356–364, Jan. 2013, doi: 10.1016/j.eswa.2012.07.051.

[20] T. Harris, “Quantitative credit risk assessment using support vector machines: broad versus narrow default definitions,” *Expert Systems with Applications*, vol. 40, no. 11, pp. 4404–4413, Sep. 2013, doi: 10.1016/j.eswa.2013.01.044.

[21] D. B. Desai and R. V Kulkarni, “A review : application of data mining tools in CRM for selected banks,” *International Journal of Computer Science and Information Technologies*, vol. 4, no. 2, pp. 199–201, 2013.

[22] M. R. Islam and M. A. Habib, “A data mining approach to predict prospective business sectors for lending in retail banking using decision tree,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2. pp. 13–22, 2015, doi: 10.5121/ijdkp.2015.5202.

[23] M. S. Thomas and J. Mathew, “Supervised machine learning model for automating continuous internal audit workflow,” in *2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings*, 2022, pp. 1200–1206, doi: 10.1109/ICOEI53556.2022.9776888.

[24] N. D. Moroke and K. Makatjane, “Predictive modelling for financial fraud detection using data analytics: a gradient-boosting decision tree,” in *Applications of Machine Learning and Deep Learning for Privacy and Cybersecurity*, 2022, pp. 25–45, doi: 10.4018/978-1-7998-9430-8.ch002.

[25] A. Abhaya and B. K. Patra, “An efficient method for autoencoder based outlier detection,” *Expert Systems with Applications*, vol. 213, p. 118904, Mar. 2023, doi: 10.1016/j.eswa.2022.118904.

BIOGRAPHIES OF AUTHORS

Fahd Sabry Esmail is a Ph.D. student in the Business Information system Program at the Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt. He received the B.Sc. degree in Management Information System from the Modern Academy for Computer Science and Management Technology, Cairo, Egypt, in 2008, M.Sc. degree in Information Systems from the Faculty of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt in 2016. He is currently a Lecturer Assistant with Information Systems of Department, Modern Academy for Computer Science and Management Technology, Cairo, Egypt. He can be contacted at email: Fahd.Sabry21@commerce.helwan.edu.eg.

Assoc. Prof. Dr. Fahad Kamal ALsheref was born in Sohag, Egypt, in 1983. He received the B.Sc. degree from the Faculty of Computers and Information, Assuit University, Assuit, Egypt, in 2005, M.Sc. and Ph.D. degrees from the Faculty of Computers and Information, Helwan University, Cairo, Egypt in 2011 and 2012. He is currently an associate professor with Information systems Department, Faculty of Computers and Information, Beni-Suef University, Beni Suef, Egypt. He has authored or co-authored over 16 researches publications in peer-reviewed reputed journals. His research interests include social informatics, health information systems, machine learning, and data mining. Dr. Fahad has set a reviewer for a number of international journals. He can be contacted at email: drfahad@fcis.bsu.edu.eg.

Prof. Dr. Amal Elsayed Aboutabl is currently a Professor at the Department of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from the American University in Cairo and both of her M.Sc. and Ph.D. in Computer Science from Cairo University. She worked for IBM and ICL in Egypt for seven years. She was also a Fulbright Scholar at the Department of Computer Science, University of Virginia, USA. Her current research interests include software engineering and natural language processing. She can be contacted at email: amal.aboutabl@fci.helwan.edu.eg.