Asian Journal of Research in Computer Science

11(3): 9-22, 2021; Article no.AJRCOS.72569
ISSN: 2581-8260

Emerging Approach for Detection of Financial Frauds Using Machine Learning

Upasana Mukherjee[1], Vandana Thakkar[1], Shawni Dutta[1], Utsab Mukherjee[1] and Samir Kumar Bandyopadhyay[2*]

[1]Department of Computer Science, The Bhawanipur Education Society College, India.
[2]The Bhawanipur Education Society College, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2021/v11i330263
Editor(s):
(1) Dr. Hasibun Naher, BRAC University, Bangladesh.
Reviewers:
(1) Shamitha S Kotekani, Visvesvaraya Technological University (VTU), India.
(2) Maad M. Mijwil, Baghdad College of Economics Sciences University, Iraq.
Complete Peer review History: https://www.sdiarticle4.com/review-history/72569

Original Research Article

Received 15 June 2021
Accepted 21 August 2021
Published 27 August 2021

ABSTRACT

The growth of regularly generated data from many financial activities has significant implications for every corner of financial modelling. This study has investigated the utilization of these continuous growing data by a means of an automated process. The automated process can be developed by using Machine learning based techniques that analyze the data and gain experience from the underlying data. Different important domains of financial fields such as Credit card fraud detection, bankruptcy detection, loan default prediction, investment prediction, marketing and many more can be modelled by implementing machine learning methods. Among several machine learning based techniques, the use of parametric and non-parametric based methods are approached by this research. Two parametric models namely Logistic Regression, Gaussian Naive Bayes models and two non-parametric methods such as Random Forest, Decision Tree are implemented in this paper. All the mentioned models are developed and implemented in the field of Credit card fraud detection, bankruptcy detection, loan default prediction. In each of the aforementioned cases, the comparative study among the classification techniques is drawn and the best model is identified. The performance of each classifier on each considered domain is evaluated by various performance metrics such as accuracy, F1-score and mean squared error. In the credit card fraud detection model the decision tree classifier performs the best with an accuracy of 99.1% and, in the loan default prediction and bankruptcy detection model, the random forest classifier gives the best accuracy of 97% and 96.84% respectively.

**Corresponding author: E-mail: 1954samir@gmail.com;*

Keywords: Financial analytics; parametric and non-parametric; credit card fraud detection; bankruptcy detection; loan default prediction.

1. INTRODUCTION

Since the past few centuries, the main aim of human being in business and finance is analysis of capital, investment, profit, etc. Finance analytics provide contradictory perceptions on the financial data with respect to a given business, this provides the basic understanding that can simplify strategic verdicts and actions to improve the overall performance of the business. The aim of financial analytics is to provide better business strategies through more factual data that are reliable instead of relying on intuition. Traditionally, the chief officers followed historical data and trends and based on that future predictions were made [1]. It is needless to mention that companies are switching their focus from traditional data analysis into modern technologies like advanced data analytics coupled with machine learning and automation. According to the experts, it is considered that predictive analysis is one of the most significant elements in the field of finance. A primary section of the predictive analysis is the capacity to compare past data with new data so that the best assessment can be made with respect to any company [2]; this improves forecasting and decision making. The data can be a company's macro-economic data, industry trends, fuel price, etc. It is obvious that in order to get a better result by a decision-making process, machine learning must be incorporated. Human beings are making different applications using traditional programming since centuries, with the advancement of technology, traditional programs are being replaced by machine learning algorithms. The machine learning algorithms are designed in such a way that it can essentially addresses the question of building computers to progress automatically through the past experiences. Machine Learning is regarded as the subset of artificial intelligence where the computer can predict based on past data. It is the fastest growing technology in the field of business, healthcare, education, weather forecast, etc. The predictive modelling of the algorithm can build or reform the prevailing algorithm to learn new data and extract pattern from the existing ones. Machine Learning uses different data mining techniques to extract data so that it can foresee future results [3]. In simpler words, machine learning is nothing but an algorithm that acts like human beings. Just the way a human being can analyse, predicts any situation based on the experience, machine learning also does the same. It is further classified into supervised and unsupervised learnings. This research is carried out using different supervised schemes and they are compared.

The greatest threat to the society is money fraud. Yet, this can be detected by gathering data, analysing them. In the late 20th century, to increase the productivity of work, most of the private as well as public companies have shifted partially of fully to e-commerce. This marks the beginning of the era of cyber money fraud. The cyber money fraud is proved to be the never-ending anarchy in the society. This may include credit card scam, loan fraud or many other types of frauds. However, the problem of money fraud can be detected and controlled using machine learning techniques. The nature of scams constantly deviates and this makes the prediction of all the money scams as a data-mining problem. At first, a vast data must be gathered so that the computer can classify which one is fraud and which transaction is not. Then the model segments the data into training and testing sets. The principle is to feed the model by the testing data so that the probability of fraud is predicted.

In order to avoid the undesired losses, different well-known money frauds like credit card fraud, loan related frauds and bankruptcy frauds are highly important to be focussed [4]. It is needless to say that the most common scam in the field of finance is credit card fraud. It occurs when a card is stolen, or someone's personal information is gathered to perform CNP transactions. This affects more than 10.7 million people globally every year [5]. In the financial industry, to detect and to control credit card scams, machine learning techniques are becoming more dominant since the past decade. Besides this, the loan related scams also create a menace [6]. A loan fraud is nothing but offering loan schemes with suspiciously favourable circumstances without any company details. The machine learning based models are highly potent to save billions of dollars in scams and with the advancement of technology, such frauds can be identified easily with the help of AI by filtering out loan defaulters in the past with significant accuracy [6, 7]. Apart from the scams using credit card or loan frauds, one of the most significant scams in the financial domain is the

bankruptcy fraud. Here if the company is unable to repay the debt to the creditors then the stakeholders of a business would suffer loss. So, the bankruptcy estimation can also be done by using different machine learning tools [8].

The three significant financial scams – credit card fraud, bankruptcy and loan defaulters are identified. This research has been carried out using parametric methods [9] like logistic regression [10], Naïve Bayes [11] as well as the non-parametric methods [9] like decision tree [12] and random forest [13]. We have retrieved the performance from these models and the best model is identified for each financial field. The figures of merits to evaluate the performance of each model are accuracy, precision, F1-score and mean squared error(MSE) [14].

The major contribution of this project can be summarized as follows-

1. Model financial analytics using machine learning techniques.
2. Prediction of prominent financial domains such as credit card detection, bankruptcy detection and loan defaulter identification using machine learning based methods.
3. All these financial tasks are assessed by means of parametric and non-parametric models such as Random Forest, Logistic Regression, Gaussian naive bayes and decision tree.
4. A comparative study is being conducted for each of the aforementioned financial domains by applying the employed models.
5. Based on the comparative analysis, the most efficient model is picked up for each task.

Some literature review has been done in section 2; the background of the research is provided in the section 3; section 4 explains the used datasets. In section 5, we have described the methodology and the algorithm used. The results are written in the section 6, finally the future scopes are written in section 7 that concludes the work.

2. RELATED WORK

In general, Fraud detection is viewed as a data mining classification problem. It comprises monitoring the behaviour of users in order to estimate, detect, or avoid undesirable behaviour. Many researches have been conducted based on data mining in the field of financial and banking sector. This paper mainly focuses on three main fraud occasions in real-world transactions, that is, a study on Credit card fraud detection, Loan default prediction, and Bankruptcy detection.

2.1 Credit Card Fraud Detection

Awoyemi, John O., et al. [15] have made prediction of credit card scam using Naïve Bayes, KNN along with logistic regression techniques of machine learning using python language and have made a comparative analysis of the result based on accuracy, sensitivity, precision, specificity, Matthews's correlation coefficient metrics and balanced classification rate in each case. The authors have deployed a highly skewed dataset. It is noted that in the research carried out by Awoyemi, John O., et al.. [15] ; that on an unbalanced dataset, a fusion of under-sampling and over-sampling is done to get two sets of distribution having 10:90 and 34:64 for analysis. The accuracy of KNN is found to be 97.92% and in NB it is 97.69% but logistic regression shows significantly low efficiency of 54.86%.

Ong Shu Yee, and et al. [16] have implemented a framework that can detect credit card fraud by using an amalgamation of machine learning and data mining technologies, the research is carried out using supervised learning. Five Bayesian network classifiers, viz., K2, Tree Augmented Naïve Bayes (TAN), and Naïve Bayes, logistics and J48 classifiers have been used. To monitor the performance parameters the researchers have used WEKA tools and obtained more than 95% accuracy for all the classifiers after using preprocessing of the data.

Maes S. et al. [17] have implemented a framework that is capable of detecting credit card scam by involving Artificial Neural Networks and Bayesian Belief Networks as the machine learning techniques based on the dataset of Serge Waterschoot at Europay International (EPI). It was observed that both the techniques give promising result in the above dataset.

Raj, S. and et al. [18] have made a survey work on different credit card scam detections using machine learning. The authors have reviewed :- i) A fusion approach using Dempster–Shafer theory and Bayesian learning ii) BLAST-SSAHA Hybridization iii)Hidden Markov Model iv) Fuzzy Darwinian Detection of Credit Card Fraud v)

Bayesian and Neural Networks. The authors have noticed that the accuracy of Fuzzy Darwinian(best), Dempster and Bayesian theory is high in terms of true positive or false positive and the processing speed for BLAH-FDS and ANN is very fast in detecting credit card scams.

Khare N and et al. [19] in their work tested the behaviour of decision tree, random forest, SVM and logistic regression on a skewed dataset of credit card scam hence a collative comparison is made and have observed that random forest classifier works most accurately with an accuracy of 98.6%.

Thennakoon, Anuruddha, et al. [20] aimed to detect real-time credit card scams by deploying machine-learning models coupled with API module to decide whether a particular transaction is honest or not. The problem was carried out by some supervised models like Logistic Regression, Gaussian Mixture Models, Naïve Bayes, K-Nearest Neighbour and Support Vector Machine; they have obtained four fraud patterns viz risky mcc, unauthorized web address, ISO response code, dealing above 100 USD. The accuracy levels for each model are 74%, 83%, 72% and 91% respectively; with mentioning the future research scope as location based scam detection.

Dhankhad S. and et al. [21] have made a framework where comparative investigation have been made in credit card scams using real-world dataset by applying some supervised learning processes like Logistic Regression (LR), Decision Tree Method, Random Forest Method, Naive Bayes, K-Nearest Neighbourhood (KNN), Gradient Boosted Tree Classifier (GBT) and XGBoost Classifier (XGB) methods of machine learning besides this, ensemble methods or the meta-classifiers excels the result by coupling multiple learning classifiers. Besides accuracy, the authors have also included some other figures of merits to justify the result; they are F1-score, precision, recall, G-mean, TP, FP and specificity.

Fu, Kang, et al. [22] revealed how credit card scams can be detected by using CNN based models in a much effective way for a huge data of 260 million annual transactions. The trading entropy identifies complicated scam patterns by fitting the features into the feature matrix. During the training phase, the dataset goes through feature engineering, followed by sampling then feature transformation and finally CNN based training methods. The prediction section can decide whether it is a fake transaction or not; the F1 score is also used to validate the accuracy of the system.

The use of Artificial Immune Systems makes it possible to detect whether a credit card transaction is fraudulent or not. Manoel Fernando Alonso and et al. [23] in their project implemented a model that can detect credit card fraud using AIS. The AIS inspired algorithm improves the accuracy significantly till 85% by a significant reduction in run-time by 40%. The authors have used Bayesian algorithm, Neural network, Markov model, account signature, coupled to the Artificial Immune Systems. Use of cloud-based file system to implement credit card scam recognition makes it possible by Hadoop for large datasets.

In order to overcome the strong imbalance in class, Dornadula and et al. [24] have included labled and unlabelled samples in the research. The research focusses mainly on the enhancement of thee skill to process a colossal transaction with scams. Just like the previous cases supervised learning tools were deployed here. A deep auto encoder model and restricted Boltzmann machine makes it possible to detect glitches in normal transactions and a hybrid method was developed by combining Adaboost coupled with majority voting methods. Besides this a feedback path is also used to eradicate the issue of data imbalance.

2.2 Loan Defaulter Detection

Zhu L. et al. [25] have made a loan evasion estimation model by via real-world consumer loan data from Lending Club. The authors have used synthetic minority over-sampling technique method in order to perform dataset preprocessing like cleaning, and dimensional reduction. The research was carried out using random forest classifiers to predict loan defaulter; the performance of the framework was tested using some parameters like accuracy, F1-Score, Recall, and AUC. Besides this a comparative analysis between RF and some other classifiers viz Decision Tree, SVM, Logistic Regression was also provided by the authors hence proved that RF shows best accuracy even under skewed datasets.

Tiwari in his work [26] applied machine learning techniques and predicted loan evasion in a large dataset that was technically imbalanced and

have proven random forest gives best accuracy (86%) however CART gives a very low accuracy.

Zhou and Wang in their work [27] have customised the random forest technique by allocating weights in the decision trees in the forest and implemented this to estimate loan evasion problem. They've used dataset from Kaggle. They also compared their proposed model with existing classifiers using R language.

Hamid and et al. [28] used three algorithms viz J48, BavesNet, and Naïve Bayes to implement a model to predict and classify the claims of loans that are introduced to the clients by analysing the behaviour of customers as well as previous history of their repay. The research has been carried out using Weka platform to implement Naïve Bayes, Neural Networks, Decision Tree. The J48 algorithm shows best results with respect to the dataset used.

Arutjothi, G., and et al. [29] aimed to create a credit score of the customers by using status of loan. Hence the system was able to filter the defaulters. Min-Max normalization embedded with K-Nearest Neighbour is used to perform the prediction process by sampling the dataset from lending club randomly.

Loan assessment framework was developed by Li, Sheng-Tun, and et al. [30] using SVM in order to identify the worthy applicants of loan. The model was developed by accompanying cross-validation and paired t-test in order to compare the predicted performance.

Kou, et al. [31] aimed to implement a model that can evaluate bank loan defaulters based on multiple criteria decision-making models, on a data provided by the Chinese. Primarily, the work focuses on feature selection by independent component analysis and principle component analysis, followed by oversampling is done by SMOTE technique; then TOPSIS (Technique for order preference by similarity to ideal solution), MCDM (multiple criteria decision-making models) methods were used to evaluate the models.

In this paper [32] numerous supervised machine learning procedures have been implemented to estimate loan status forecast models that are dependent on a dataset from one Ugandan financial institution. The Random Forest method from Alternating Decision Trees (ADTs), Forest by Penalizing Attributes (Forest PA), Hoeffding Tree (VFDT), C4.5 algorithm, Logistic Model Trees (LMT), Random Tree (RT) and Random Forest were used. This resulted in the fact that the ensemble classifiers gave promising results.

2.3 Bankruptcy Prediction

Nagaraj and Sridhar [33] have proposed a predictive model for bankruptcy detection by using machine learning which acts as a decision support tool based on the dataset from UCI Machine Learning Repository getting almost 99% accuracy. It was observed that SVM based model had best efficiency as compared to Logistic Regression, Rotation Forest, Naive Bayes and Neural Network.

Wang, Nanxi, et al. [34] proposed a framework to predict bankruptcy using Support Vector Machine, with dropout and Autoencoder based on the Qualitative Bankruptcy Dataset collected from UCI. It has been proven that neural networks with added layers work with highest accuracy.

Sun, Lili [35] used Naïve Bayes model to make a bankruptcy prediction tool, the irrelevant tools were removed based on the derived correlation. A 10-fold validation has been done.

Hardinata L. et al. [36] have presented an execution of Jordan Recurrent Neural Networks to classify and predict Bankruptcy in Corporate Sectors. The feedback interaction in Jordan Recurrent Neural Networks helped the network to improve the efficiency. They have taken the dataset from University of California at Irvine. In the best performance the average accuracy of their system is 81.3785% where the number of neurons in the hidden layer is 5 [36].

In this paper [37] metaheuristic algorithm artificial bee colony (ABC), an ANN model called ABCNN has been used to create a hybrid model which can be applied in corporate bankruptcy prediction (CBP), or referred to as financial distress prediction. The authors have obtained the performance of the model in terms of some figure of merits like accuracy, type I, type II errors and AUC scores were used as well. The authors have compared the hybrid model with two other models to evaluate the efficiency. The first model was equipped with multiple discriminant analysis whereas ANN was used to train the second model. The result shows that ANN is more accurate than MDA by 10% roughly with an accuracy of almost 91%.

Antunes F. et al. [38] assumed a probabilistic point-of-view by applying three different classification models, Gaussian processes (GP) in the context of bankruptcy prediction, comparing it against the support vector machines (SVM) and the logistic regression (LR). Besides this, the authors have provided a clear graphical visualization for better understanding of different performances. This paper shows that the probabilistic GP classifier is superior than LR and SVM for a broad dataset.It was observed that in the DIANE data, GP is less sensitive to the class balance which maintains a comparable performance of the balanced dataset. The authors wished to work on the same module using other kernels on some other datasets.

The primary motive of Hauser, Richard P., and David Booth [39] was to inspect the efficiency of predicting bankruptcy using a three-fold cross validation system in order to compare the classification and prediction of bankruptcy by a robust logistic regression coupled with the Bianco and Yohai estimator with respect to maximum likelihood logistic regression. It was observed that the Bianco and Yohai logistic regression changes the estimated regression coefficients from maximum likelihood logistic regression hence the BY based robust logistic regression can be used to improve efficiency of the bankruptcy classification and prediction problems. However, in some cases the BY robust logistic regression makes no deviations in the projected regression coefficients and has the same classification and prediction results as ML logistic regression. The data was collected from US corporation in 2008 – 2009.

3. BACKGROUND

Machine learning methods are majorly considered into two categories, viz., Supervised and Unsupervised learning methods. Unsupervised learning is independent of trained data sets to predict the results, rather, it uses some direct procedures such as clustering and association in order to forecast the results. The trained datasets are defined as the input for which we get known outputs. On the other hand, the Supervised learning is a learning technique in which the machine is trained or taught using some well-defined labelled data. Followed by, the new sets of data are provided to the machine so that the supervised learning procedure investigates the training data and provides an accurate result from categorized data. Classification is defined as a supervised technique that maps the data into some predefined groups or classes. This is because the classes are determined before investigating the data. In economic analytics the data cataloguing is regarded as organizing crucial financial information. It is needless to say that there has been a colossal study that exploited the power of classification to distinguish and fight credit card scam, predict evasion of loans and bankruptcy detection. A classifier is an algorithm that automatically categorizes data into one or more of a set of classes [3,9].

Based on the learning traditions, machine learning procedures can be of two classes such as parametric models and non-parametric models. Parametric Methods employs a constant number of parameters to figure the model which is used to regulate a probability model used in Machine Learning as well. A parametric algorithm shows faster computation. However, it makes tougher assumptions of the data. If the assumption turns out to be correct, the algorithm may work. On the contrary, if the assumptions are wrong, the algorithm fail to work properly. The learning model that summarises data with a set of constraints of fixed size of a predefined mapped function that is independent of the number of training examples is called parametric model. A few illustrations of the parametric machine learning models are Logistic Regression, Linear Discriminant Analysis, Perceptron, Naive Bayes, Simple Neural Networks, and many more [40]. Besides this, the Non-Parametric Methods use a flexible number of parameters to implement the framework. It uses a flexible or a varying number of parameters that has a provision to grow while it learns from more data. It is quite obvious that a non-parametric algorithm is computationally sluggish as compared to the previous one but it makes fewer assumptions about the data. Non-parametric methods are preferred when we have a lot of data with no prior knowledge, and when we don't want to worry too much about choosing just the right features. Some popular examples of non-parametric learning algorithms are k-Nearest Neighbours, Decision Trees, Support Vector Machines, and many more [40].

Among several prevailing supervised classification methods, the following are the models that are being executed in this research. This study focusses on the following Parametric and Non-parametric models.

A. Decision Tree Classifier

B. Random Forest Classifier
C. Logistic Regression
D. Naive Bayes Classifier

3.1 Decision Tree Classifier

A Supervised learning non-parametric system with a tree-structured system that can be used for both classification and Regression problems is called decision tree. Each segment or layer is regarded as a node that represents the structures of a dataset, the decision rules are represented by the 'branches' and the outcome is represented by a 'leaf'. Beginning with a root, followed by expansion via branches or the decisions and ending into leaves gives the name 'decision tree'. It is a white box type of ML procedure that shares inner decision-making judgement, that is unavailable to the algorithms like neural networks. Besides this, the training time is faster compared to the neural network algorithm. For a given set of data, the time complexity of decision trees is dependent to the number of records and number of attributes. Therefore, the 'Decision tree' can be used to perform data-mining problems with high-dimensional data with a decent accuracy.

3.2 Random Forest Classifier

Random Forest is a supervised nonparametric machine learning algorithm. It is used for both Regression and Classification problems. Fundamentally, it follows the perception of ensemble learning. Ensemble learning is a process of cascading multiple classifiers to solve a problem which helps to improve the performance of the model. Random Forest process is a combination of multiple decision trees and a technique called Boosting and Aggregation or Bagging. Therefore, the result not only relies on a single decision tree rather it takes the forecast from every tree and then just predicts the final output based on most of each prediction. The number of trees in the forest and accuracy are proportional to each other [13].

3.3 Logistic Regression

One of the most widely used supervised parametric models is the Logistic regression that is primarily used to predict the categorical dependent variable using a given set of independent variables showing some categorically discrete values like 1-0 or true-false, etc. but instead of providing the exact value as 0 and 1, it gives the probabilistic values which lie amongst 0 and 1. In Logistic regression, rather than fitting a regression line, fitting an "S" shaped logistic function is preferred, which forecasts two maximum values (0 or 1). It is an important machine learning process since it has the capability to deliver probabilities and categorize new data using continuous and distinct datasets. Similarly, it can be used to organize the observations using different kinds of data and can effortlessly regulate the most active variables used for the classification [10].

3.4 Naive Bayes Classifier

Naive Bayes is a supervised parametric machine learning algorithm based on the Bayes Theorem mainly used to serve classification problems. Naive Bayes is called so since the guesses are based on the conditional independence of each pair of features, that is the existence of any feature is independent of the occurrence of further features. Hence it cannot acquire the relationship amongst those features. There are many sorts of Naive Bayes Classifiers present from those some are Gaussian Naive Bayes which follows normal distribution, Multinomial Naive Bayes that follows multinomial distribution and Bernoulli Naive Bayes which follows multinomial distribution nevertheless the independent variables are Boolean variables. In this paper, we have used the Gaussian Naive Bayes Classifier. In Gaussian Naive Bayes, continuous values related with each feature are presumed to be distributed according to a Normal distribution [11].

3.5 Performance Evaluation Metrics

There are various metrics that we have used to evaluate the performance of ML algorithms, classification as well as regression algorithms.

3.5.1 Confusion matrix

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model [14]. It is used for Classification problems where the output can be of two or more types of classes. It is a table with two dimensions viz. "Actual" and "Predicted" and both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)".

3.5.2 Classification accuracy

It may be defined as the number of correct predictions made as a ratio of all predictions made. It can be easily calculated by confusion matrix with the help of the formula

$$\text{Accuracy}=\frac{TP+TN}{(TP+FP+FN+TN)}$$

3.5.3 Recall

It is defined as the number of positives returned by our ML model. It can be easily calculated by confusion matrix with the help of the formula

$$\text{Recall}=\frac{TP}{(TP+FN)}$$

3.5.4 Precision

Precision can be defined as the number of correct documents returned by our ML model. It can be easily calculated by confusion matrix with the help of the formula

$$\text{Precision}=\frac{TP}{(TP+FP)}$$

3.5.5 F1-score

F1 score gives the harmonic mean of precision and recall. Mathematically, it is the weighted average of precision and recall. The best value of F1 would be 1 and the worst would be 0. It can be calculated with the formula

$$\text{F1-score}=2*\frac{(Precision*Recall)}{(Precision+Recall)}$$

3.5.6 Mean squared error

The MSE(Mean Squared Error) is calculated as the mean or average of the squared differences between predicted and actual values.

$$\text{MSE}=\frac{1}{n}\sum_{i=0}^{n} \quad (Yi-Yipred)^2$$

Here, Yi is the i'th actual value and Yipred is the i'th predicted value. The difference between these two values is squared, which has the effect of removing the sign, resulting in a positive error value.

4. DATASET DESCRIPTION

4.1 Data Description of Credit Card Fraud Detection Model

From Kaggle the dataset of the credit card fraud detection model is collected [41]. The dataset contains 284807 rows and 31 columns. For each data, there are 28 transactions (Attributes are V1, V2, V3, ..., V28). The target variable is Class and Time, Amount, V1 to V28 are the independent variables. Fig. 1 shows the data type of the attributes. In this dataset, there are 492 fraud cases and 284315 valid cases of credit card transactions. This count is shown in Fig. 2.

```
Time      float64
V1        float64
V2        float64
V3        float64
V4        float64
V5        float64
V6        float64
V7        float64
V8        float64
V9        float64
V10       float64
V11       float64
V12       float64
V13       float64
V14       float64
V15       float64
V16       float64
V17       float64
V18       float64
V19       float64
V20       float64
V21       float64
V22       float64
V23       float64
V24       float64
V25       float64
V26       float64
V27       float64
V28       float64
Amount    float64
Class       int64
dtype: object
```

Fig. 1. Attribute description of credit card transaction dataset

4.2 Data Description of Default Loan Prediction Model

This dataset is collected from Kaggle. The dataset contains 10000 records [42]. Here the dependent attribute is "Defaulted?" and the independent attributes are Employed, Bank Balance, Annual Salary. Fig. 3 shows the data type of the attributes. Here the number of defaulted cases are 333 and rest 9667 cases are not defaulted which are shown in Fig. 4.

4.3 Data Description of Bankruptcy Detection Model

For this model, the dataset is taken from Kaggle [43]. It contains 6819 rows and 96 columns. The target variable is "Bankrupt?" and the rest 95 attributes are the independent variable. Fig. 5 shows the data type of the attributes. This dataset contains 220 bankrupt and 6599 not bankrupt cases which is shown in the Fig. 6.

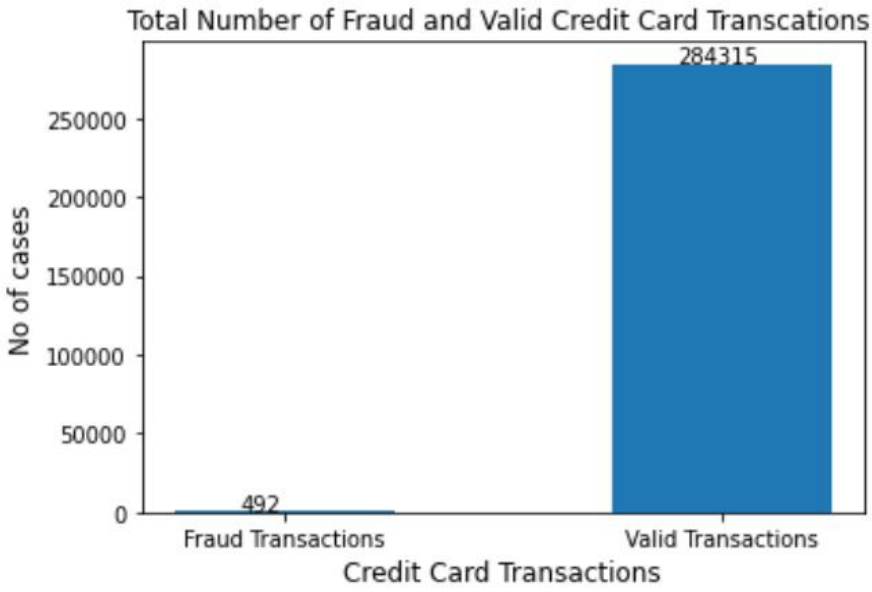

Fig. 2. Data distribution of target attribute in credit card transaction dataset

```
Index             int64
Employed          int64
Bank Balance    float64
Annual Salary   float64
Defaulted?        int64
dtype: object
```

Fig. 3. Attribute description of loan defaulter dataset

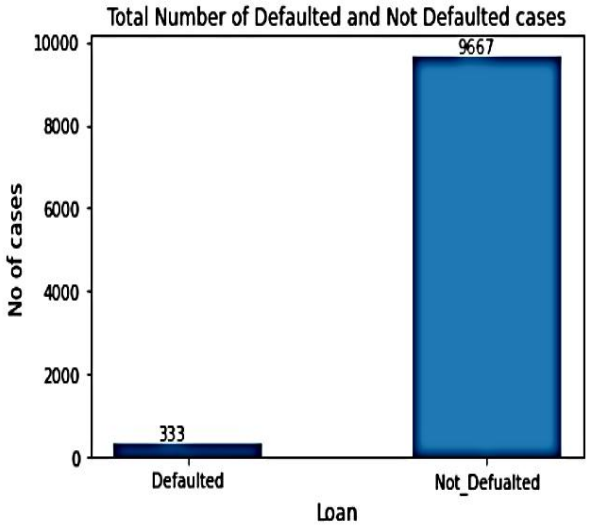

Fig. 4. Data distribution of target attribute in loan default prediction dataset

```
Bankrupt?                                                  int64
 ROA(C) before interest and depreciation before interest  float64
 ROA(A) before interest and % after tax                   float64
 ROA(B) before interest and depreciation after tax        float64
 Operating Gross Margin                                   float64
                                                           ...
 Liability to Equity                                      float64
 Degree of Financial Leverage (DFL)                       float64
 Interest Coverage Ratio (Interest expense to EBIT)       float64
 Net Income Flag                                            int64
 Equity to Liability                                      float64
Length: 96, dtype: object
```

Fig. 5. Attribute description of bankruptcy dataset

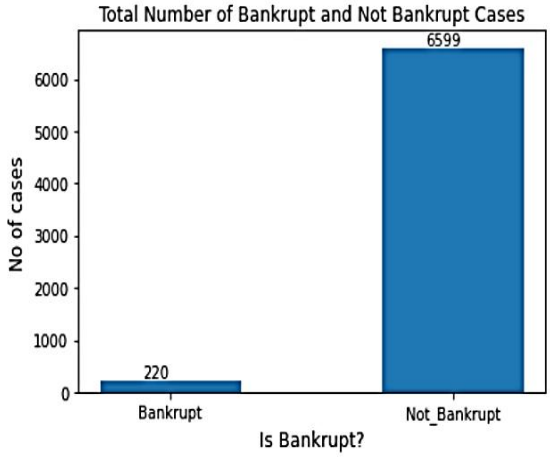

Fig. 6. Data distribution of target attribute in bankruptcy detection dataset

5. METHODOLOGY

Financial analytics provides differing perspectives on the financial data of a given business, giving insights that can facilitate strategic decisions and actions that improve the overall performance of the business. Supervised learning is a learning process in which we teach or train the machine using data which is well labelled implies that some data is already marked with the correct responses. After that, the machine is provided with the new sets of data so that the supervised learning algorithm analyzes the training data and gives an accurate result from labelled data. The datasets of credit card fraud detection, bankruptcy detection and default loan prediction models are pre-processed then splitted into 8:2 ratio that is 80 % of the data is used for tracing purpose and the rest The datasets of credit card fraud detention, bankruptcy detection and default loan prediction models are preprocessed then splitted into 8:2 ratio that is 80 % of the data is used for tracing purpose and the rest 20% data is used for testing purpose. Scaling of numeric attributes and encoding of categorical data into numeric forms were carried out in pre-processing steps. For scaling purpose that is for transforming the attributes ranging from 0 to 1, we have used MinMax Scaler. Additionally, the attributes those were not contributing to the classification problems (index, ID, etc.) were dropped. The classifiers which are implemented from sklearn library [44] are Random Forest Classifier, Decision Tree, Logistic Regression and Gaussian Naive Bayes. Each employed models were run by modifying different hyper-parameters. The best hyper-parameters for each model is illustrated here. In case of bankruptcy prediction, the decision tree was implemented with the entropy criterion, the random forest classifier was built using 150 number of estimators. In the case of credit card fraud detection and loan prediction the decision tree we have used the criterion as gini and splitter as best and, in random forest 100 no of estimators is used and, in case of logistic regression we have set the max_iter parameter as 100. After

applying all the classifiers we have evaluated the accuracy, F1-score and MSE.

We have applied the same algorithm for the three models i.e., Credit card fraud detection model, Loan prediction model and Bankruptcy detection model.

5.1 Algorithm

Step1. Collect the dataset.
Step2. Preprocess the dataset.

a) Replace the missing values with 0 (if any).
b) Drop the irrelevant attributes (if any).
c) Transforming the attributes ranging from 0 to 1

Step3. Choose the dependent and independent attributes.
Step4. Split the dataset into two parts in 8:2 ratio for the training and testing purpose.
Step5. Build the Parametric and nonparametric Classifier models using the training dataset and then predict the test dataset.
Step6. Evaluate accuracy, F1-score and MSE with the help of confusion matrix for each classifier and compare them.

6. EXPERIMENTAL RESULTS AND DISCUSSION

All the implemented machine learning models are applied to the financial tasks that are considered in this study. Comparative analysis among the employed machine learning methods such as Random Forest, Decision Tree, Logistic Regression and Gaussian Naive Bayes model is drawn and summarized in Tables 1,2,3 for credit card fraud detection, bank loan defaulter prediction, bankruptcy prediction respectively.

Table 1 implies the result metrics of the Credit Card fraud detection. It was observed that the accuracy of Decision Tree is better than that of the peer models. However, in the field of Fraud Detection, it is necessary to take care of the false triggers. Hence, the F1-Score needs to be considered as well. The performance of the Random Forest method is superior as compared to the other employed models in terms of F1-Score. Hence, Random Forest model is regarded as the best model in this domain.

Table 2 has the resultant metrics of the classifiers for the Bank loan defaulter prediction purpose. Experimental results indicate that random forest classifier has 97% accuracy and F1-score of 0.886957 which is the highest compared to other classification techniques. This model has also shown the optimised error rate over other models.

Table 3 refers to the performance summarization of the classifiers of the bankruptcy detection model. Comparison of all the classifier performance can conclude that the Random Forest gets the highest accuracy 96.84% and F1-Score as 0.810998.

Table 1. Credit card fraud prediction performance

Name of Classifier	Accuracy	F1-score	MSE
Random Forest	0.999052	0.874699	0.000948
Decision Tree	0.999192	0.874510	0.000808
Logistic Regression	0.999070	0.782635	0.000930
Naive Bayes(Gaussian)	0.992609	0.727523	0.007391

Table 2. Loan defaulter prediction performance

Name of classifier	Accuracy	F1-score	MSE
Random Forest	0.970500	0.886957	0.029500
Decision Tree	0.951000	0.858767	0.049000
Logistic Regression	0.969000	0.861111	0.031000
Naive Bayes(Gaussian)	0.964500	0.860049	0.035500

Table 3. Bankruptcy prediction performance

Name of classifier	Accuracy	F1-score	MSE
Random Forest	0.969941	0.810998	0.030059
Decision Tree	0.967009	0.809999	0.032991
Logistic Regression	0.944282	0.790001	0.055718
Naive Bayes(Gaussian)	0.388563	0.289520	0.611437

7. CONCLUSION

This study has provided comprehensive modelling of machine learning techniques on financial applications. For this purpose, three essential financial topics such as Credit card fraud prediction, loan defaulter prediction and bankruptcy prediction is taken into consideration. In these three fields, use of well-known machine learning based classification techniques is approached. Numerous parametric and non-parametric classification techniques such as Random Forest, Decision Tree, Logistic Regression and Gaussian Naive Bayes model are implemented and their prediction performances are compared. Comparative study identifies the best possible predictive model in each of these fields. Credit card fraudulent detection systems can be developed by using the Random Forest technique. Despite of the fact Decision Tree model has a promising accuracy of 99.905%, but Random Forest model is preferred because of enhanced F1-score of 0.874699. The Random forest model has shown the best performance while predicting Bank loan defaulters as well as bankruptcy with an accuracy of 97% and 96.84% respectively and F1-Scores as 0.886957 and 0.810998 respectively. This paper has shown its contribution to the fact that data driven approaches such as machine learning techniques can perform prediction on finance based tasks. Extensive comparison presented in the study can assist to identify the most efficient predictive modelling which in turn can benefit the customers as well as organizations to facilitate the decision making process.

DISCLAIMER

The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Burdick, Doug, et al. Financial analytics from public data. Proceedings of the International Workshop on Data Science for Macro-Modeling; 2014.
2. De Prado, Marcos Lopez. Advances in financial machine learning. John Wiley & Sons; 2018.
3. Holzinger, Andreas. Introduction to MAchine Learning & Knowledge Extraction (MAKE)." Mach. Learn. Knowl. Extr. 2019;1.1:1-20.
4. Sinayobye, Janvier Omar, Fred Kiwanuka, Swaib Kaawaase Kyanda. A state-of-the-art review of machine learning techniques for fraud detection research. 2018 IEEE/ACM Symposium on Software Engineering in Africa (SEiA). IEEE; 2018.
5. The Basics of Credit Card Theft. Identity Guard; 2020, February 17. Available:https://www.identityguard.com/news/basics-of-credit-card-theft
6. Awoyemi, John O., Adebayo O. Adetunmbi, Samuel A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI). IEEE; 2017.
7. Aslam, Uzair, et al. An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience. 2019;16.8:3483-3488.
8. Wang, Nanxi. Bankruptcy prediction using machine learning. Journal of Mathematical Finance. 2017;7.04:908.
9. Alpaydin, Ethem. Introduction to machine learning. MIT Press; 2020.

10. Kleinbaum, David G, et al. Logistic regression. New York: Springer-Verlag; 2002.
11. Berrar, Daniel. Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands. 2018;403-412.
12. Priyam, Anuja, et al. Comparative analysis of decision tree classification algorithms. International Journal of Current Engineering and Technology. 2013;3.2: 334-337.
13. Liaw, Andy, and Matthew Wiener. Classification and regression by random Forest. R News. 2002; 2.3:18-22.
14. De Sa, Christopher, et al. High-accuracy low-precision training. 2018;arXiv preprint arXiv:1803.03383.
15. Awoyemi, John O, Adebayo O. Adetunmbi, Samuel A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI). IEEE; 2017.
16. Yee, Ong Shu, Saravanan Sagadevan, Nurul Hashimah Ahamed Hassain Malim. Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC). 2018;10.1-4:23-27.
17. Maes, Sam, et al. Credit card fraud detection using Bayesian and neural networks. Proceedings of the 1st International Naiso Congress on Neuro Fuzzy Technologies; 2002.
18. Raj S. Benson Edwin, Annie Portia A. Analysis on credit card fraud detection methods. 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET). IEEE; 2011.
19. Khare Sait. Credit card fraud detection using machine learning models and collating machine learning models. International Journal of Pure and Applied Mathematics. 2018;118.20:825-838.
20. Thennakoon, Anuruddha, et al. Real-time credit card fraud detection using machine learning. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE; 2019.
21. Dhankhad, Sahil, Emad Mohammed, Behrouz Far. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. 2018 IEEE international conference on information reuse and integration (IRI). IEEE; 2018.
22. Fu, Kang, et al. Credit card fraud detection using convolutional neural networks. International conference on neural information processing. Springer, Cham; 2016.
23. Gadi, Manoel Fernando Alonso, Xidi Wang, Alair Pereira do Lago. Credit card fraud detection with artificial immune system. International Conference on Artificial Immune Systems. Springer, Berlin, Heidelberg; 2008.
24. Dornadula, Vaishnavi Nath, Geetha S. Credit card fraud detection using machine learning algorithms. Procedia Computer Science. 2019;165 :631-641.
25. Zhu, Lin, et al. A study on predicting loan default based on the random forest algorithm. Procedia Computer Science. 2019;162:503-513.
26. Tiwari, Abhishek Kumar. Machine learning application in loan default prediction. Machine Learning. 2018;4.5.
27. Zhou, Lifeng, Hong Wang. Loan default prediction on large imbalanced data using random forests. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2012;10.6 :1519-1525.
28. Hamid, Aboobyda Jafar, Tarig Mohammed Ahmed. Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal (MLAIJ). 2016;3.1.
29. Arutjothi G, Senthamarai C. Prediction of loan status in commercial bank using machine learning classifier. International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2017.
30. Li, Sheng-Tun, Weissor Shiue, Meng-Huah Huang. The evaluation of consumer loans using support vector machines. Expert Systems with Applications. 2006;30.4:772-782.
31. Kou, Gang, Yi Peng, Chen Lu. MCDM approach to evaluating bank loan default models. Technological and Economic Development of Economy. 2014;20.2:292-311.
32. Nabende, Peter, Samuel Senfuma. A study of machine learning models for predicting loan status from Ugandan loan

applications. Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp); 2019.

33. Nagaraj, Kalyan, Amulyashree Sridhar. A predictive system for detection of bankruptcy using machine learning techniques. 2015;arXiv preprint arXiv:1502.03601.
34. Wang, Nanxi. Bankruptcy prediction using machine learning. Journal of Mathematical Finance. 2017;7.04:908.
35. Sun, Lili, Prakash P. Shenoy. Using Bayesian networks for bankruptcy prediction: Some methodological issues. European Journal of Operational Research. 2007;180.2:738-753.
36. Hardinata, Lingga, Budi Warsito. Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks: a case study in Polish companies. Journal of Physics: Conference Series. IOP Publishing. 2018;1025(1).
37. Marso, Said, Mohamed EL Merouani. Bankruptcy prediction using hybrid neural networks with artificial bee colony. Engineering Letters. 2020;28.4.
38. Antunes, Francisco, Bernardete Ribeiro, Francisco Pereira. Probabilistic modeling and visualization for bankruptcy prediction. Applied Soft Computing. 2017;60:831-843.
39. Hauser, Richard P, David Booth. Predicting bankruptcy with robust logistic regression. Journal of Data Science. 2011;9.4:565-584.
40. Khadse, Vijay M, Parikshit Narendra Mahalle, Gitanjali R. Shinde. Statistical study of machine learning algorithms using parametric and non-parametric tests: a comparative analysis and recommendations. International Journal of Ambient Computing and Intelligence (IJACI). 2020;11.3:80-105.
41. Machine Learning Group - ULB. Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine; 2016-11-03. Retrieved on Jan 25, 2021. Available:https://www.kaggle.com/mlg-ulb/creditcardfraud/
42. Kamal Das Loan Default Prediction Beginners data set for financial analytics; 2016-11-03. Retrieved on Feb 2, 2021. Available: https://www.kaggle.com/kmldas/loan-default-prediction
43. fedesoriano. Company Bankruptcy Prediction Bankruptcy data from the Taiwan Economic Journal for the years 1999–2009; 2021-01-22. Retrieved on Feb 23, 2021. Available:https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction
44. Pedregosa, Fabian, et al. Scikit-learn: Machine learning in Python. The Journal of machine Learning Research. 2011;12 :2825-2830.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/72569