 doi:10.20944/preprints202108.0028.v1

Emerging Approach for Detection of Financial Frauds Using Machine Learning

1. Vandana Thakkar, Final Year Student, Department of Computer Science, The Bhawanipur Education Society College
2. Upasana Mukherjee, Final Year Student, Department of Computer Science, The Bhawanipur Education Society College
3. Shawni Dutta, Lecturer, Department of Computer Science, The Bhawanipur Education Society College
4. Prof. Samir Kumar Bandyopadhyay, Academic Advisor, The Bhawanipur Education Society College

Abstract

The growth of regularly generated data from many financial activities has significant implications for every corner of financial modeling. 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 other financial models can be modeled by implementing machine learning models. 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, recall, precision, 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.

Keywords: Financial Analytics, Parametric and Non-parametric, Credit card fraud detection, bankruptcy detection, loan default prediction

1. Introduction

Finance analytics provide 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. The goal of financial analytics is to shape the strategy for business through reliable, factual insight rather than intuition. The chief financial officers traditionally relied on historical data and trends to forecast future performance [1]. However, they are changing their focus as they increasingly tap into technologies, such as advanced data analytics, machine learning and automation. Many experts consider predictive analytics an essential element in the digital transformation of finance. A key part of this is the ability to examine historical and new data to assess what's relevant to a specific company, be it macroeconomic data, industry trends or petroleum prices, to improve forecasting and decision-making [2]. To support the decision making process, use of machine learning is often relevant in practice.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

Machine learning is a subfield of computer science, but is often also referred to as predictive analytics, or predictive modeling. Its goal and usage is to build new and/or leverage existing algorithms to learn from data, in order to build generalizable models that give accurate predictions, or to find patterns, particularly with new and unseen similar data. Machine learning includes the study of an algorithm that can automatically extract the data. Machine learning utilizes data mining techniques and another learning algorithm to construct models of what is happening behind certain information so that it can predict future results [3].

Fraud detection process using machine learning starts with gathering and segmenting the data. Then, the machine learning model is fed with training sets to predict the probability of fraud. To avoid unwanted losses amongst numerous well-known frauds, credit card fraud, loan related frauds are necessary to be focused [4]. Credit card fraud which occurs when a card is stolen or someone's personal information is hacked to perform so-called card-not-present (CNP) transactions. It is the most common form of identity theft, affecting more than 10.7 million annually [5]. In the financial industry, Credit card fraud detection using Machine Learning is becoming the prevalent method of fraud prevention. And in case of Loan related frauds, it happens if a person contacts and offers a loan scheme with suspiciously favourable conditions and asks for the bank details or for payment upfront, without having any proper company information or even using an international contact number [6].

This technology has the potential to help save financial institutions billions of dollars in fraud losses over the coming years. Such frauds can easily be handled by AI using previous loan application records to filter out loan defaulters [7]. Bankruptcy is another state of insolvency wherein the company or the person is not able to repay the creditors the debt amount. Bankruptcy prediction using different machine learning models is of importance to the various stakeholders of the company as well as the society on the whole [8].

In this current study, three significant financial based tasks such as credit card detection, bankruptcy detection and loan defaulter identification are carried out based on parametric [9] and non-parametric machine learning models [9]. In our implementation, among numerous parametric models, Logistic regression [10], Naive Bayes [11] models are used whereas, decision tree [12] and random forest [13] models are used from non-parametric machine learning categories. The performance retrieved from these predictive models are compared and the best model is identified for each considered financial domain. The performance of these models are evaluated by various evolutionary metrics [14] such as accuracy, recall, precision, F1-score and mean squared error(MSE).

The major contribution of this paper 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.

The rest of this paper is arranged as follows. Section 2 gives the summaries of the related work. Section 3 provides the background that is the description of the methods and some related terms which are used. Section 4 describes the three datasets. Section 5 describes the methodology and the algorithm of the three

 doi:10.20944/preprints202108.0028.v1

experiments. Section 6 gives the results of the experiments. Section 7 provides the conclusion of this paper.

2. Related Work

Fraud detection is generally viewed as a data mining classification problem. It involves monitoring the behavior of users in order to estimate, detect, or avoid undesirable behavior. Many researches have been conducted based on data mining in the field of financial and banking sector. This paper mainly focuses on four 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

In this study a comparative analysis of credit card fraud detection is done using Naive Bayes, KNN and Logistic Regression techniques of machine learning. They have used a highly skewed dataset(source: ULB Machine Learning Group (European cardholders containing 284,807 transactions)). The comparative study is done based on accuracy, sensitivity, precision, specificity, Matthews's correlation coefficient metrics and balanced classification rate. It is observed that a hybrid of under-sampling and oversampling is carried out on the highly unbalanced dataset to achieve two sets of distribution, that is, 10:90 and 34:64 for analysis. It was observed from their analysis that KNN with accuracy 97.92% is better than Naive Bayes with accuracy 97.69% and Logistic Regression with accuracy 54.86%.They have used the python language for the implementation. Also, the expected future areas of research could be in examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data. And, the effects of other sampling approaches can be also investigated [15].

The objective of this paper [16] is to detect credit card fraud by using a combination of machine learning and data mining. They have used five Bayesian network classifiers, namely, K2, Tree Augmented Naïve Bayes (TAN), and Naïve Bayes, logistics and J48 classifiers. In this implementation, WEKA was used to measure the performances of the classifiers. Their study evaluated the performances of the classifiers using True Positive Rate, False Positive Rate, Precision, Recall, F-Measure and accuracy. After they did the preprocessing of the dataset using normalization and Principal Component Analysis, all the classifiers achieved more than 95.0% accuracy compared to results attained before preprocessing the dataset [16].

Maes S. et al. [17] have discussed automated credit card fraud detection by using machine learning techniques. They have used Artificial Neural Networks and Bayesian Belief Networks as machine learning techniques. The source of their dataset is Serge Waterschoot at Europay International(EPI). After applying these two techniques they have concluded that good results can be achieved by both techniques. They have compared the results achieved by ANN and BBN for a false positive rate of 10% and 15% respectively. BBN has a shorter training time and has achieved better results but the fraud detection process is faster with ANN. The future work with ANN using Pruning algorithms could improve the performance of backdrop and with BBN a structure learning method can be implemented based on dependency analysis and the results are compared with the STAGE algorithm [17].

 doi:10.20944/preprints202108.0028.v1

A comparative study of various methods used in credit card fraud detection is carried out in [18]. The methods are:- 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. Results show that the fraud detection systems such as Fuzzy Darwinian(best), Dempster and Bayesian theory have very high accuracy in terms of TP and FP. At the same time, the processing speed is fast enough to enable online detection of credit card fraud in the case of BLAH-FDS and ANN.

Another research [19] checks the performance of the Decision tree, Random Forest, SVM and logistic regression on highly skewed credit card fraud data and then does a collative comparison to evaluate which model is the best. They have used a highly skewed dataset(source: ULB Machine Learning Group (European cardholders containing 284,807 transactions)). These techniques are applied to the raw and preprocessed data. They have evaluated the performances of the techniques on the principle of accuracy, sensitivity, specificity, precision. Their results indicate that the optimal accuracy for logistic regression, decision tree, Random Forest and SVM classifiers are 97.7%, 95.5% and 98.6%, 97.5% respectively. So they have concluded that the Random Forest Classifier gives the best accuracy(for the given dataset).

The aim of the paper [20] is to detect real-time credit card fraud detection using machine learning models. Here, the authors take the use of predictive analysis done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or not. They have used Logistic Regression (LR), Gaussian Mixture Models (GMMs), Naive Bayes, Risk-Based Enable (RBE) models of machine learning. Four machine learning algorithms that are Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Logistic Regression were prioritized in the analysis. LR, NB, LR and SVM are the machine learning models that captured the four fraud patterns (Risky MCC, Unknown web address, ISO Response Code, Transaction above 100$) with the highest accuracy rates. Also, the accuracy rates that the models indicated are 74%, 83%, 72% and 91% respectively. They have used the dataset from a financial institution(according to a confidential disclosure agreement). The future research could be to focus on location-based fraud detection.

Dhankhad S. et al. [21] have done a comparative analysis on credit card fraudulent transaction detection using a real-world dataset applying various supervised machine learning algorithms. Here they have used 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. Also, Ensemble learning (also known as meta-classifier) helps to improve the results by combining multiple machine learning classifiers to improve the predictive outcomes. Accuracy is one important method to compare the performance of classification models is Accuracy but the other factors such as F1-Score, Precision, TPR, FPR, Recall, G-mean and Specificity are also taken into analysis. By comparing we get the result that all the proposed models were superior in overall performance but the overall results show that the most promising for predicting fraudulent transactions is the stacking classifier which is using LR as meta classifier in the dataset and it is followed by the random forest and XGB classifier. The future work will be conducted using the voting classifier and checking the performance with other ML learning methods, increase the size of training and testing dataset, also all the

 doi:10.20944/preprints202108.0028.v1

machine learning algorithms can be used to find out the features importance [21].

Another study [22] revealed the latent patterns of credit card fraudulent transactions and also avoided the model over-fitting by using a convolutional neural network (CNN) to reduce the feature redundancy effectively. Here, a novel trading feature called trading entropy is proposed to identify more complex fraud patterns and the features are transformed into a feature matrix to fit the CNN model in order to apply CNN to credit card fraud detection. The fraud detection framework consists of two parts, training and prediction parts where the training part mainly includes four modules. The four modules include, feature engineering, sampling methods, feature transformation and a CNN-based training procedure. The prediction part can judge whether it is fraudulent or not immediately when a transaction comes. Also, the detection procedure consists of feature extraction, feature transformation and the classification module. To evaluate the performance of models, the F1 score is analyzed. The results implied that the proposed method performs better than other state-of-art methods. They have used real credit card transaction data from a commercial bank which contains over 260 million transactions of credit cards in a year [22].

Halvaiee and Akbari [23] have suggested credit card fraud detection using Artificial Immune Systems (AIS), and also introduced a new model called AIS-based Fraud Detection Model (AFDM). Here, an immune system inspired algorithm (AIRS) is used and it will be improved for fraud detection. It increases the accuracy up to 25%, reduces the cost up to 85%, and decreases system response time up to 40% compared to the base algorithm. The different methods that have been used for fraud detection are Bayesian algorithm, Neural network, Markov model, account signature, Artificial Immune Systems. They suggest using cloud computing i.e. implementing a fraud detection system on a cloud-based file system, namely Hadoop, which makes data parallelization possible in large datasets. The Hadoop Distributed File System is used for storing transaction records and MapReduce API for processing those records. The four parameters which are used for evaluating fraud detection methods are True negative(TN), False negative (FN), True positive (TP), False positive and the method which offers minimum FP and FN, and maximum TP and TN is considered as the best method. Having all these changes together makes the best results. AFDM has improved detection rate up to 23%, decreased cost up to 85%, and training time up to 40%. Also higher than 50%detection rate can be achieved while having a very low false-positive rate (less than 2%), which is considerable. Then, implementing the parallel model only in a test environment shows a fair decrease in training time, which is expected to be better in a real cloud computing system. They have used the dataset from transactions of a Brazilian bank. The future improvements which can be done are weighting dataset fields in distance function, using artificial immune networks for fraud detection, using distance function based on dataset properties and cloud computing misuse detection using AIRS [23].

To overcome the strong class imbalance, the inclusion of labelled and unlabelled samples is approached by this study [24]. This study has also focused on the enhancement of the ability to process a large number of transactions with card frauds related dataset. For this purpose, different supervised machine learning algorithms such as Decision Tree, Naive Bayes Classification, Least Squares Regression, Logistic Regression and SVM have been employed and applied to detect fraudulent transactions in real-time datasets. The performance of Logistic Regression, K-Nearest Neighbour, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

Naive Bayes methods are analysed on highly skewed credit card fraud data. A model of deep Auto-encoder and restricted Boltzmann machine (RBM) that can construct normal transactions to find anomalies from normal patterns and also a hybrid method is developed with a combination of Adaboost and Majority Voting methods as supervised learning methods may fail at detecting certain cases of fraud detection. A feedback mechanism is also used to solve the problem of concept drift (high imbalance of data). Also, the Matthews Correlation Coefficient (MCC) was the better parameter to deal with imbalance dataset but by applying the Synthetic Minority Over-Sampling Technique (SMOTE), it is tried to balance the dataset, then it was observed that the classifiers were performing better than before. The other way of handling the imbalanced dataset is to use one-class classifiers like one-class Support Vector Machine (SVM). It was finally observed that the algorithms that gave better results are Logistic regression, Decision tree and Random forest [24].

2.2 Loan defaulter detection

Zhu L. et al. [25] have built a loan default prediction model by using real-world user loan data from Lending Club. They have used the Synthetic Minority Over-sampling Technique method for dealing with the imbalanced dataset and for preprocessing(such as data cleaning and dimensionality reduction) of that data. Random forest classifiers are used for predicting the default loan and to predict its performance they have used accuracy, AUC, F1-Score and recall. Then they just compared the classifier with other classifiers namely Decision tree, Logistic regression and SVM. They have seen that the best accuracy(98%) is given by the random forest classifier [25].

The aim of the paper [26] is to apply machine learning techniques to predict loan default prediction in a large imbalanced data set. They have used Logistic regression, KNN, tree-based classifier, classification and regression tree(CART), and Random Forest in their model. The authors have evaluated the performance of each and every technique using the metric Area Under the Curve (AUC), F1 score, Recall, Precision, Accuracy using the ROC Curve. As a result, they have seen that Random forest(86% highest accuracy) and KNN gives high accuracy and CART models give the lowest accuracy [26].

Zhou and Wang [27] have proposed an improved Random Forest technique for default loan predictions. They have allocated the weights to the decision trees in the forest. They made the prediction decision based on the weighted majority which was in the ensemble of all the trees in the forest. The weights were calculated by out-of-bag errors in training. They have compared their proposed technique with the original Random Forest method, SVM, KNN and C4.5 classifiers concerning the overall accuracy and balanced accuracy. They have implemented it by using the R language. The imbalanced dataset which they have used in their work is from Kaggle [27].

In this paper [28] three algorithms have been used such as J48, BayesNet and NaiveBayes algorithms to build predictive models that are used to predict as well as classify the applications of loans that introduced the customers to good or bad loan by investigating the customer's behavior and previous pay back credit.The model has been implemented by using Weka application. Then the several classification classifiers like Naive Bayes Classifier, Neural Network Classifier, Decision Tree Classifier

 doi:10.20944/preprints202108.0028.v1

are used to produce a model for predicting the class of unknown records. And, after applying the classification's data mining algorithms that is J48, BayesNet and NaiveBayes on the model, the conclusion is that the best algorithm for the loan classification is the J48 algorithm as it has high accuracy and also low mean absolute error. Also it is capable of classifying the instances correctly than the other algorithms and the confusion matrix of the three algorithms showed that the J48 algorithm is the best one. The dataset is taken from the banking sector [28].

The aim of this paper [29] is to create a credit scoring model by using the loan status as the credit scoring model is used for accurate analysis of credit data to find defaulters and valid customers. Here, the machine learning classifier based analysis model for credit data is created using the combination of Min-Max normalization and K Nearest Neighbor (K-NN) classifier and is implemented using the software package R tool. A random sampling method is used to solve the problem of an unbalanced dataset. The K-NN classifier has been used for the prediction and to predict its performance they have calculated the accuracy, Root Mean Squared Error and Correlation.Then after comparing with other classifiers it has been concluded that the K-NN based credit scoring system provides higher accuracy than other classifiers which can be effectively used by commercial loan lenders to predict the loan applicant. The dataset is taken from Lending Club [29].

A loan evaluation model using SVM is developed by [30] to identify potential applicants for consumer loans in order to meet the upcoming Basel II requirement. The development of the model is accompanied by cross-validation and paired t test in order to compare the predictive performance. However the analysis of misclassification errors in terms of Type I and Type II and their effect on selecting network parameters of SVM is conducted. The misclassification analysis that is, by adopting a SVM with RBF (radial basis function) kernel, leads to the development of a visual interactive tool in determining conservative, aggressive, or compromised loan evaluation policy by support decision makers and facilitates the development of a useful visual decision-support tool. It also shows that SVM surpasses traditional neural network models like MLP(multilayer perceptrons), MDA(multiple discriminant analysis), logistic regression analysis,etc in generalization performance and visualization via the visual tool, which helps decision makers determine appropriate loan evaluation strategies. The data is collected from a local bank in Taiwan. The future research is to enhance the SVM with rule extraction to allow one to systematically uncover the tacit knowledge embedded in the application cases with satisfactory accuracy and interpretability [30].

The aim of this paper [31] is an evaluation approach for bank loan default classification models based on multiple criteria decision making (MCDM) methods. The main concentration is on feature selection, unbalanced data, and model assessment in bank loan default prediction. A procedure is developed to address the three problems, that is, firstly, Independent component analysis (ICA) and Principal component analysis (PCA) were used to select relevant features. Secondly, the SMOTE(Synthetic Minority Oversampling TEchnique) approach was utilized to deal with the unbalanced data by creating synthetic default examples. Thirdly, Technique for order preference by similarity to ideal solution (TOPSIS), a MCDM method was utilized to rank a selection of default prediction models. The performances of classifiers were measured using accuracy, Type-I error rate, Type-II error rate, and AUC. And, the experimental results showed that K-NN has good potential in default prediction. Also, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

outcome indicated that there is no significant difference between PCA and ICA on default prediction of the specific dataset. The future research includes introducing more feature selection techniques, sampling approaches, classification algorithms, and MCDM methods to the process. And, another research direction is to resolve this disagreement and help decision-makers pick the most suitable classifier(s). The dataset was provided by a Chinese commercial bank [31].

In this paper [32] various supervised machine learning methods have been applied for estimating loan status prediction models based on an electronic loan applications dataset from one Ugandan financial institution. The Random Forest approach 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, resulted in the highest classification accuracy and when used as a base classifier with other classification methods such locally weighted learning, there was a slight improvement in accuracy. Also, ensemble classifiers gave promising results. Cross validation accuracy results not only show significant differences in the loan status prediction accuracies from the various prediction models, but also indicate supervised learning methods that consistently resulted in models with promising and near satisfactory loan status prediction performance. The future work is that more Ugandan loan application data sets should be developed and used for modeling loan status prediction and also, it should also be interesting to evaluate classification models developed from similar data sets from different contexts on the Ugandan loan applications data set [32].

2.3 Bankruptcy prediction

Nagaraj and Sridhar [33] have proposed a predictive model for bankruptcy detection by using machine learning.Their model behaves as a decision support tool. They have used the dataset from UCI Machine Learning Repository. RBF based SVM performed best (99% accuracy) in comparison with other methods namely Logistic Regression, Rotation Forest, Naive Bayes and Neural Network in their implementation. They have used R for their implementation [33].

In this paper [34] three methods namely Support Vector Machine, Neural Network with dropout and Autoencoder have been proposed to predict bankruptcy. They used the Qualitative Bankruptcy Dataset collected from UCI. Among the three methods neural network with added layers and with dropout performed the best. They also compared these three methods with the former methods for predicting bankruptcy which are Logistic Regression, Genetic algorithm and Inductive Learning and concluded that the proposed methods have better accuracy [34].

Using naive Bayes Bayesian network (BN) model bankruptcy prediction tool is developed. Based on the derived correlation the irrelevant variables were removed. Later on, the proposed model is developed by a means of 10-fold validation technique. This study has also considered the case of over-fitting and finally concluded the applicability of this technique in business modeling for decision making process apart from the current bankruptcy prediction domain.

Hardinata L. et al. [36] have presented an implementation of Jordan Recurrent Neural Networks for the classification and prediction of Corporate Bankruptcy. The feedback interaction in Jordan Recurrent Neural Networks helped the network to improve the efficiency. They have taken the dataset

 doi:10.20944/preprints202108.0028.v1

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 model's performance was evaluated not only in terms of prediction accuracy but also the type I and type II errors and AUC-score were used. The hybrid model is compared with the two other models in order to investigate its efficiency: the first model is multiple discriminant analysis (MDA) and the second one is an ANN trained by the most common learning algorithm, a back propagation (BPNN). The experimental results indicate that ANN models, on average, are approximately 10% more accurate in relation to MDA in different periods. The ABCNN model led to 92% accuracy, whereas BPNN's and MDA's led to 91% and 81% accuracy, respectively in one year before the bankruptcy. However, when it was applied in the period three years before the bankruptcy, it was found that ABCNN's result was 80.94% followed by BPNN with 81.05% , then MDA - with 67.22%. The results of the experiments indicate better performance of the hybrid ABCNN, followed by the traditional back-propagation BPNN, with their flexible non-linear modeling capability, over other multivariate statistical methods. The dataset is about CBP of Polish firms and it was collected from (EMIS). The future research includes applying deep neural networks model, incorporating macroeconomic variables with financial ratios, using a parallel processing architecture [37].

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). Also, a complete graphical visualization was generated to improve the understanding of the different attained performances, effectively compiling all the conducted experiments in a meaningful way. The probabilistic GP classifier used in this paper showed to be superior in comparison with both SVM and LR methods in a broad range of studied scenarios and datasets. Moreover, regarding the DIANE data, the GP proved to be less sensitive to the class balance, maintaining a comparable performance to that of the balanced dataset. For the credit risk datasets, the results were generally worse across all the models. However, in the majority of the cases, the GP proved to have higher classification performance. And, using the real-world bankruptcy data, an in-depth analysis is conducted showing that, in addition to a probabilistic interpretation, the GP can effectively improve the bankruptcy prediction performance with high accuracy when compared to the other approaches. The main dataset used in this study was extracted from the DIANE database. The future work includes trying another set of kernels, also applying the same experimental setup to more domains [38].

The main aim of this paper [39] was to investigate the accuracy of predicting bankruptcy using a three-fold cross validation scheme to compare the classification and prediction of bankrupt firms by robust logistic regression with the Bianco and Yohai (BY) estimator versus maximum likelihood (ML) logistic regression. The analysis indicates that if the BY robust logistic regression significantly changes the estimated regression coefficients from ML logistic regression, then the BY robust logistic regression method can significantly improve the classification and prediction of bankrupt firms. And at worst, the BY robust logistic regression makes no changes in the estimated regression coefficients and has the same classification and prediction results as ML logistic regression. Also, this strongly shows that BY robust

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

logistic regression should be used as a robustness check on ML logistic regression and if a difference exists, BY robust logistic regression should be used as the primary classifier of bankrupt firms. The data sample for this bankruptcy prediction study consists of U.S. corporations that filed for bankruptcy in 2008-2009 as listed in a bankruptcy research database [39].

3. Background

Machine learning algorithms are majorly categorized into two types, namely, Supervised and Unsupervised learning methods. Unsupervised learning does not depend on trained data sets to predict the results, but it utilizes direct techniques such as clustering and association in order to predict the results. Trained data sets are defined as the input for which the output is known. Supervised learning is a learning process in which we teach or train the machine using data which is well leveled 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 labeled data. Classification is a supervised method that maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. In financial analytics data classification is about organizing crucial financial information. There are a large number of studies that exploited the power of classification to detect and combat credit card fraud, predict default loans and bankruptcy detection. A classifier is an algorithm that automatically categorizes data into one or more of a set of classes [3,9].

Depending on the learning assumptions, machine learning techniques can be of two categories such as parametric models and non-parametric models. Parametric Methods uses a fixed number of parameters to build the model which is used to determine a probability model that is used in Machine Learning as well. A parametric algorithm is computationally faster, but makes stronger assumptions about the data; the algorithm may work well if the assumptions turn out to be correct, but it may perform badly if the assumptions are wrong. A learning model that summarises data with a set of parameters of fixed size (predefined mapped function) (independent of the number of training examples). Some examples of parametric machine learning algorithms include Logistic Regression, Linear Discriminant Analysis, Perceptron, Naive Bayes, Simple Neural Networks, and many more. Non-Parametric Methods use a flexible number of parameters to build the model. It uses a flexible number of parameters, and the number of parameters often grows as it learns from more data. A non-parametric algorithm is computationally slower but makes fewer assumptions about the data. Non-parametric methods are good when we have a lot of data and no prior knowledge, and when we don't want to worry too much about choosing just the right features. Some examples of popular nonparametric machine learning algorithms are k-Nearest Neighbors, Decision Trees, Support Vector Machines, and many more [41].

Among numerous existing supervised classification techniques, the following are the models that are being implemented. In this study, we have used the following Parametric and Non-parametric models.

A. Random Forest Classifier
B. Decision Tree Classifier
C. Logistic Regression
D. Naive Bayes Classifier

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

3.1. Random Forest Classifier

Random Forest algorithm is a supervised machine learning algorithm. It is an example of a non-parametric model. For both Regression and Classification problems this algorithm can be used. Basically, it follows the concept of ensemble learning. Ensemble learning is a process of combining multiple classifiers to solve a problem and this helps to improve the performance of the model. Random Forest algorithm combines multiple decision trees and a technique called Boosting and Aggregation(also known as Bagging). So it is not dependent only on a single decision tree; rather it takes the prediction from every tree and then just predicts the final output bases on the majority of each prediction. The number of trees in the forest and accuracy are proportional to each other [13].

3.2. Decision Tree Classifier

A Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. It is called a decision tree because similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. A Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. Decision trees can handle high-dimensional data with good accuracy. It is also an example of a non-parametric model [12].

3.3. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning based parametric technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). It is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Also, it can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification [10].

3.4. Naive Bayes Classifier

Naive Bayes is a supervised machine learning algorithm. For the Classification problem, this algorithm can be used. It is based on the Bayes Theorem. It is an example of a parametric model. Naive Bayes is called naive because the assumption is based on the conditional independence of each pair of features, that is the occurrence of any feature is independent of the occurrence of other features. Hence it can not learn the relationship between those features. There are many types of Naive Bayes Classifiers present from those some are Gaussian Naive Bayes(follows normal distribution), Multinomial Naive Bayes(follows multinomial distribution) and Bernoulli Naive Bayes(follows multinomial distribution but here the independent variables are boolean variables). In this paper, we have used the Gaussian Naive Bayes Classifier. In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

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)".

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

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

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

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

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

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

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

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

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. Figure 1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

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 Figure 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
```

Figure 1: Attribute description of credit card transaction dataset

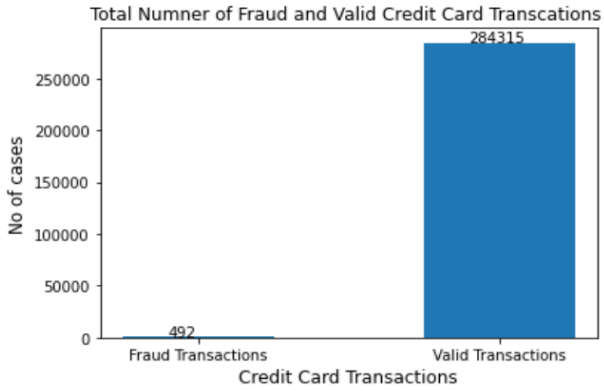

Figure 2: Data distribution of target attribute in credit card transaction dataset

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. Figure 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 figure 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

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

Figure3: Attribute description of loan defaulter dataset

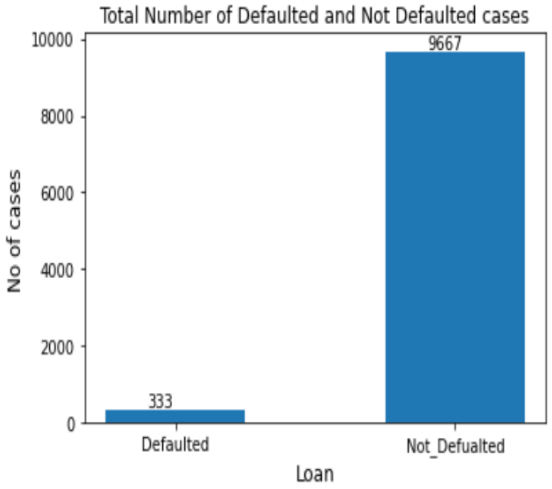

Figure 4: Data distribution of target attribute in loan default prediction dataset

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. Figure 5 shows the data type of the attributes. This dataset contains 220 bankrupt and 6599 not bankrupt cases which is shown in the figure 6.

```
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
```

Figure 5: Attribute description of bankruptcy dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

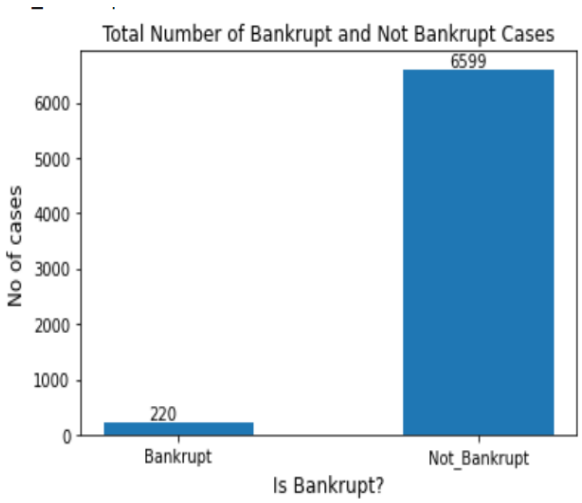

Figure 6: Data distribution of target attribute in bankruptcy detection dataset

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 labeled 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 labeled data.

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. The classifiers which are implemented from sklearn library are Random Forest Classifier, Decision Tree, Logistic Regression and Gaussian Naive Bayes. In the case of the decision tree we have used the criterion as gini and splitter as best for all the three models of financial analytics. After applying all the classifiers we have evaluated the accuracy, recall, precision, 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.

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

Step6. Evaluate accuracy, recall, precision, F1-score and MSE with the help of confusion matrix for each classifier and compare them.

6. Experimental Results

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 Table 1,2,3 for credit card fraud detection, bank loan defaulter prediction, bankruptcy prediction respectively.

Table 1 implies that the Decision Tree model can be used to build predictive tools with highest accuracy of 99.9%. However, other models have also exhibited benchmark efficiencies around 99%. This discussion reveals that Credit card fraudulent systems can be developed by using the Decision Tree model.

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 which is the highest compared to other classification techniques.

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%.

Name of Classifier	Accuray	Recall	Precision	F1-score	MSE
Random Forest	0.999052	0.554455	0.861538	0.674699	0.000948
Decision Tree	0.999192	0.782178	0.766990	0.774510	0.000808
Logistic Regression	0.999070	0.564356	0.863636	0.682635	0.000930
Naive Bayes(Gaussian)	0.992609	0.613861	0.139640	0.227523	0.007391

Table 1: Credit card fraud prediction performance

Name of Classifier	Accuracy	Recall	Precision	F1-score	MSE
Random Forest	0.970500	0.378378	0.682927	0.486957	0.029500
Decision Tree	0.951000	0.324324	0.333333	0.328767	0.049000
Logistic Regression	0.969000	0.189189	0.875000	0.311111	0.031000
Naive Bayes(Gaussian)	0.964500	0.243243	0.545455	0.336449	0.035500

Table 2: Loan defaulter prediction performance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 August 2021 doi:10.20944/preprints202108.0028.v1

Name of Classifier	Accuracy	Recall	Precision	F1-score	MSE
Random Forest	0.968475	0.108696	0.714286	0.188679	0.031525
Decision Tree	0.967009	0.152174	0.538462	0.237288	0.032991
Logistic Regression	0.944282	0.260870	0.222222	0.240000	0.055718
Naive Bayes(Gaussian)	0.388563	0.891304	0.047126	0.089520	0.611437

Table 3: Bankruptcy prediction performance

7. Conclusion

This study has provided comprehensive modeling 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 Decision Tree model with promising accuracy of 99.91%. 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. 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.

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.* 1.1 (2019): 1-20.
4. Sinayobye, Janvier Omar, Fred Kiwanuka, and 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*. (2020, February 17). Identity Guard. https://www.identityguard.com/news/basics-of-credit-card-theft

6. Awoyemi, John O., Adebayo O. Adetunmbi, and 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* 16.8 (2019): 3483-3488.
8. Wang, Nanxi. "Bankruptcy prediction using machine learning." *Journal of Mathematical Finance* 7.04 (2017): 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* 3.2 (2013): 334-337.
13. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.
14. De Sa, Christopher, et al. "High-accuracy low-precision training." *arXiv preprint arXiv:1803.03383* (2018).
15. Awoyemi, John O., Adebayo O. Adetunmbi, and 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, and Nurul Hashimah Ahamed Hassain Malim. "Credit card fraud detection using machine learning as data mining technique." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10.1-4 (2018): 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, and A. Annie Portia. "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* 118.20 (2018): 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, and 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, and 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, and S. Geetha. "Credit card fraud detection using machine learning algorithms." *Procedia computer science* 165 (2019): 631-641.
25. Zhu, Lin, et al. "A study on predicting loan default based on the random forest algorithm." *Procedia Computer Science* 162 (2019): 503-513.
26. Tiwari, Abhishek Kumar. "Machine learning application in loan default prediction." *Machine Learning* 4.5 (2018).
27. Zhou, Lifeng, and Hong Wang. "Loan default prediction on large imbalanced data using random forests." *TELKOMNIKA Indonesian Journal of Electrical Engineering* 10.6 (2012): 1519-1525.

28. Hamid, Aboobyda Jafar, and Tarig Mohammed Ahmed. "Developing prediction model of loan risk in banks using data mining." *Machine Learning and Applications: An International Journal (MLAIJ) Vol* 3.1 (2016).
29. Arutjothi, G., and C. Senthamarai. "Prediction of loan status in commercial bank using machine learning classifier." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2017.
30. Li, Sheng-Tun, Weissor Shiue, and Meng-Huah Huang. "The evaluation of consumer loans using support vector machines." *Expert Systems with Applications* 30.4 (2006): 772-782.
31. Kou, Gang, Yi Peng, and Chen Lu. "MCDM approach to evaluating bank loan default models." *Technological and Economic Development of Economy* 20.2 (2014): 292-311.
32. Nabende, Peter, and 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, and Amulyashree Sridhar. "A predictive system for detection of bankruptcy using machine learning techniques." *arXiv preprint arXiv:1502.03601* (2015).
34. Wang, Nanxi. "Bankruptcy prediction using machine learning." *Journal of Mathematical Finance* 7.04 (2017): 908.
35. Sun, Lili, and Prakash P. Shenoy. "Using Bayesian networks for bankruptcy prediction: Some methodological issues." *European Journal of Operational Research* 180.2 (2007): 738-753.
36. Hardinata, Lingga, and Budi Warsito. "Bankruptcy prediction based on financial ratios using Jordan Recurrent Neural Networks: a case study in Polish companies." *Journal of Physics: Conference Series*. Vol. 1025. No. 1. IOP Publishing, 2018.
37. Marso, Said, and Mohamed EL Merouani. "Bankruptcy Prediction using Hybrid Neural Networks with Artificial Bee Colony." *Engineering Letters* 28.4 (2020).
38. Antunes, Francisco, Bernardete Ribeiro, and Francisco Pereira. "Probabilistic modeling and visualization for bankruptcy prediction." *Applied Soft Computing* 60 (2017): 831-843.
39. Hauser, Richard P., and David Booth. "Predicting bankruptcy with robust logistic regression." *Journal of Data Science* 9.4 (2011): 565-584.
40. Khadse, Vijay M., Parikshit Narendra Mahalle, and 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)* 11.3 (2020): 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/mlg-ulb/creditcardfraud/
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* 12 (2011): 2825-2830.