BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 07/10 pp478–490
Volume 6, Number 4, December 2023
DOI: 10.26599/BDMA.2022.9020037

AI-Based Hybrid Models for Predicting Loan Risk in the Banking Sector

Vikas Kumar*, Shaiku Shahida Saheb, Preeti, Atif Ghayas, Sunil Kumari, Jai Kishan Chandel, Saroj Kumar Pandey, and Santosh Kumar

Abstract: Every real-world scenario is now digitally replicated in order to reduce paperwork and human labor costs. Machine Learning (ML) models are also being used to make predictions in these applications. Accurate forecasting requires knowledge of these machine learning models and their distinguishing features. The datasets we use as input for each of these different types of ML models, yielding different results. The choice of an ML model for a dataset is critical. A loan risk model is used to show how ML models for a dataset can be linked together. The purpose of this study is to look into how we could use machine learning to quantify or forecast mortgage credit risk. This phrase refers to the process of evaluating massive amounts of data in order to derive useful information for making decisions in a variety of fields. If credit risk is considered, a method based on an examination of what caused and how mortgage credit risk affected credit defaults during the still-current economic crisis of 2021 will be tried. Various approaches to credit risk calculation will be examined, ranging from the most basic to the most complex. In addition, we will conduct a case study on a sample of mortgage loans and compare the results of three different analytical approaches, logistic regression, decision tree, and gradient boost to see which one produced the most commercially useful insights.

Key words: Artificial Intelligence (AI); Machine Learning (ML); loan prediction; Support Vector Machine (SVM); Random Forest (RF); accuracy

• Vikas Kumar is with the Humanities and Management Department, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar 144027, India. E-mail: vikashkumar1061@gmail.com.
• Shaiku Shahida Saheb is with the Mittal School of Business, Lovely Professional University, Phagwara 144402, India. E-mail: sahebsabi@gmail.com.
• Preeti is with the Department of Commerce & Business Administration, Kanya Maha Vidyalaya (KMV), Jalandhar 144004, India. E-mail: preetibhuker@gmail.com.
• Atif Ghayas is with the School of Management, Gitam (to be deemed university), Bangalore 561203, India. E-mail: atifghayas91@gmail.com.
• Sunil Kumari is with the Government College for Women, Indra Gandhi University, Ateli 123021, India. E-mail: raosunil77@gmail.com.
• Jai Kishan Chandel is with the Institute of Management Studies, Kurukshetra University, Kurukshetra 136119, India. E-mail: jkims2015@gmail.com.
• Saroj Kumar Pandey is with Department of Computer Engineering and Applications, GLA University, Mathura 281406, India. E-mail: patrasujitk@gmail.com.
• Santosh Kumar is with the Department of Management, Jaipuriya Institute of Management, Jaipur 302033, India. E-mail: talksant@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2022-05-17; revised: 2022-09-01; accepted: 2022-10-07

1 Introduction

1.1 Machine learning

Machine Learning (ML) is a system that combines algorithms and statistical models. Computer systems use these algorithms and statistical models to perform classification and prediction tasks automatically. It is a function of artificial intelligence. Rather than explicit instructions, these systems rely on patterns and inference. When applied to a sample dataset, prediction algorithms are made up of mathematics functions that produce a mathematical model. This dataset is known as the training dataset. Machine learning algorithms are used in a variety of applications. Machine learning has a long history of development, but recent advances in data storage and computing power have made it ubiquitous in many industries and applications, including technology and business, where "more and more machine learning tools are being adopted to manage credit risk", according to a recent Bank of Spain report.

Machine learning is the process by which computers analyse data and learn from it in order to make predictions with new data. In other words, the machine is "trained" by using large amounts of data and algorithms to find and learn patterns or trends in order to make predictions. The outcomes of machine learning methods can be difficult to interpret at times. And the effectiveness of machine learning is dependent on the starting data. The resulting system will be ineffective if the data model is incorrect or there is insufficient relevant information. Also, from the standpoint of model risk management, it is critical to understand the data so that the output of any model can be explained. As a result, selecting relevant and understandable information becomes the primary challenge in producing a successful sample.

On the other hand, we face the challenge of obtaining sensitive data that are critical to train the model. Companies that collect this information to build their own systems guard it jealously, making it difficult to build a large database on which to work on an effective predictive model.

1.2 Types of machine learning

There are three broad classifications of machine learning:

Supervised learning: Its task is to make the model learn from labelled training data. The term supervised refers to a set of samples where the output labels are already known.

Examples of supervised learning:

(1) Spam mail classification;

(2) Picture detection/ image detection;

(3) Classification of various groups on the basis of properties;

(4) I am predicting the price of bungalows based on location.

Unsupervised learning: This method is used to discover hidden structures in large amounts of data. We do not know the correct answer ahead of time. However, we look for meaningful design and information in unlabeled data without the guidance of a known outcome variable or reward function. Clustering is a data analysis technique for organizing massive amounts of data into structured clusters.

Reinforcement learning: Reinforcement learning is about taking suitable actions to maximize reward in a scenario. For example, in a chess game after the opponent chance we get a state of game and based on this state it gives us the best option to win the game. This learning based on the states and for each state it describes the best option.

The main goal here is to create an agent that can improve system performance through interaction with the environment. A reward signal is used to convey information about the environment. The agent employs reinforcement learning to learn a series of actions that, through trial and error, can maximize the reward. A chess engine is a well-known example of reinforcement learning. Here, based on the sequence of moves, the engine (agent) decides on subsequent activities by analyzing the state of the board (environment). The reward is agreed upon following the match's win or loss condition.

(1) Preprocessing: This step entails transforming the data as needed. All redundant information, as well as null values, are removed. This allows the ML algorithms to run faster.

(2) Choosing and training a predictive model: Different machine learning models have been developed for various tasks. Cross-validation techniques are used to determine which model will work best, and the training dataset is further divided into training and validation sets.

(3) Evaluating models and predicting unseen data instances: In this step, we estimate the performance for generalization error using the test dataset.

1.3 Big Data and machine learning

Once the common link or origin of Artificial Intelligence (AI) and what we know as statistical learning has been established, it is necessary to make a distinction and comparison between what we define as Big Data and machine learning. It is clear that both refer to the handling of large amounts of data, and both for storage and for analysis and application in everyday problems. The amount of data we encompass today is unmanageable. These authors in numerous studies provided a series of surprising data about the information that we are capable of handling.

(1) 90% of all data have been created in the last two years.

(2) In 2010, the total number of mobile phones was 5 billion.

(3) Thirty billion pieces of content are shared on Facebook in a month.

(4) The US Library of Congress houses some 235 terabytes of information.

However, the popularity that the term Big Data has gained in recent years leads us to a vague and complex definition. And even more so in relation to the different data analysis systems that have been used recently, among which the aforementioned machine learning, blockchain, and data mining, among others, stand out. Despite the above, Abdou et al.'s definitions reveal a common point about their definition[1]. Big Data is best defined as a set of data. These data are used on a large scale to extract new insights or create new forms of value. On the other hand, the data suppose the confluence of a multitude of technological trends that have burst into society with force, generating greater mobility, an increase in social networks and geolocation, an increase in broadband, a reduction in connection costs, and computing in the cloud. Other authors delved deeper into the characteristics of the information stored or contained in Big Data.

1.4 Opportunities to create credit risk models

The main advantage of using machine learning in credit risk management over a simpler logistic model is its superior performance, particularly in terms of discrimination. Using a broader set of variables to predict defaults improves accuracy significantly, regardless of the model used.

This allows us to highlight two distinct opportunities: efficiency and automation. Because these models are based on algorithms, they are more secure than human-based analysis. Furthermore, machine learning implies less time and resources in credit risk evaluation due to process automation, as well as better results (efficiency) because it is based on real case histories. As a result, future clients will be able to receive more personalized attention based on their risk of non-payment. For example, if a company has a low default risk, it can be offered better terms than others with a higher default risk.

Although there are still fields to be explored in the field of artificial intelligence, we have seen the obvious benefits that the use of machine learning can bring to credit risk management in terms of predicting the risk of non-payment of each client with great precision and automatically.

1.5 Motivation

This work was approached in two ways. The first involved the research method, by which information was sought about the history of artificial intelligence and machine learning, to understand how it has evolved to the present day. Credit risk and what happened in 2021 were investigated, studying the causes that produced the recession, as well as the consequences derived from the large number of defaults in mortgage loans. At the same time, the models for calculating credit risk and some techniques such as decision trees or gradient boosting that are used in this case study were also investigated, without entering into the mathematical development. On the other hand, the methodology chosen for the last section of the work was the case study. This research method is defined as empirical research that investigates a contemporary phenomenon within its real-life context; when the limits between the phenomenon and the context are not evident; and in which multiple data sources are used. The main contribution of this method is that it allows to give answers to contemporary situations and that it can facilitate the understanding of complex realities such as machine learning. A set of mortgage loans dated before the recession was analyzed. The dataset was analyzed using the Python programming language.

2 Literature Review

In this section we have reviewed the exiting literature and work done by various researcher.

Abdou et al.[1] presented a framework of finding the

prediction in loan risk. The framework accuracy and precision have been checked with the historical data of loan. Accuracy and precision values of all the models are calculated. This prototype model can be used to sanction the loan request of the customers or not. The model which has high accuracy and high precision is selected for verification and test data are passed to it to identify the loan risk. Models such as K-Nearest Neighbors (KNN), logistic regression, Naïve Bayesian Classifier, and decision tree are applied to the dataset. Considering rate of accuracy, the decision tree model works best for it. This section consists of the discussion about the exiting research work done by the various researchers, which were used as a help for identifying the problem statement and the issues that are pertaining in the field of sentiment analysis.

Bülbül et al.[2] proposed an innovative and efficient method for financial risk analysis, which includes a mathematical explanation of the high accuracy SA algorithm. Many works with various methods can be found on SA, but Abdou et al.[1] presented a process with an accuracy of 86.8%, which is very high. Celik and Karatepe[3] proposed an algorithm HC4.5 that performs financial risk analysis efficiently. The work, however, did not include sentimental analysis of data containing jargons and slangs.

Chen et al.[4] presented a model which focuses primarily on individual loan derived from a combination of the influence of neighbors' opinions and personal experiences. Various models of field-dependent cognition have previously been presented, but no significant work on field-independent awareness has been found. In general, Chen et al.[4] provided a model called the Cognitive Scale (CS) model to determine opinion dynamics based on a combination of field-independent and field-dependent cognition. Data were gathered by 1000 agents in an opinion evolution system. Despite the fact that it is a very theoretical approach, it can be validated through numerical simulations. The field dependent measures are represented by Eqs. (1) and (2).

$$x_i(t+1) = c \cdot [\omega \cdot x_i(t) + (1-\omega) \cdot x_i(t-1)] + (1-c) \cdot \frac{1}{\sum_{j \epsilon_{|N i} i N_i} x_j(t)} \quad (1)$$

$$N_i(t) = \{j \in V || x_i(t) - x_j(t)| \leqslant \in_i\} \quad (2)$$

Cozarenco and Szafarz[5] presented a paper which aims to extract feature of finical risk analysis dataset from objective sentences using a subjective sentence classifier. Previous papers on subjectivity detection have been published, but none of them have been able to fully comprehend accurate sentences. Cozarenco and Szafarz[5] have developed an algorithm for classifying subjectivity and extracting features from it. It is a completely novel concept that can be applied to further research in the field of sentiment analysis. However, a good algorithm for objective data classification is still required.

Cuestas et al.[6] proposed a paper that provides algorithms for first determining polarization in risk and then reducing its impact. Many people have conducted polarization research and discussed filter bubbles and echo chambers. Nonetheless, significant work in the field of optimizing the effect of polarization is lacking. An algorithm has been developed to moderate the opposing viewpoints that are the source of polarization. The solution approach, however, is impractical for large datasets, and the problem is NP-hard.

Silva et al.[7] discussed a very accurate financial risk analysis during credit card used by a person. Previously, sentimental analysis was performed on various datasets, but none of them had the accuracy that Silva et al.[7] have derived from their approach. Every day, a retweet network was built for SA. Data from the 2017 Chilean election were obtained from Twitter. Additional research can be conducted to extract useful data from social networking sites.

Depren and Kartal[8] addressed the problem of finding significant reviews from the reviews available online by customers. Distinguishing helpful reviews from the crowd is difficult, but it can benefit both the customer and the manufacturer. Dima and Vasilache[9] proposed and tested a system for finding useful critiques on a three-class classification problem.

Enjolras and Madies[10] emphasized the importance of sarcasm in sentiment analysis "Who cares about witty tweets? The effect of sarcasm on sentiment analysis is being investigated." The irony is sometimes difficult for humans to understand, and their analysis is a bit tricky, so it is generally ignored in sentiment analysis. Enjolras and Madies[10] provided a set of rules for detecting and analyzing sarcasm.

Gandhi et al.[11], based on Structure and Dynamic Dictionary, proposed a method for detecting negative news from online news articles. Because the news article may be violent and unpleasant, automatic news classification is required[11].

Halteh et al.[12] proposed a new method for

introducing government policy to the public. Financial risk analysis tools can be used to determine the public's opinion on these new policies, and the government can then decide whether the system is acceptable or not[12].

Li et al.[13] proposed in their paper that social networking websites are a rich source for opinion mining, so the proposed system presents an approach to extract data from Twitter and perform linguistic analysis on their opinions, and that information is in the form of graphs and charts. This proposed project was in the works and had yielded promising preliminary results. The proposed system has limitations in that it does not focus on classifying each review as a whole, and the dataset is limited to Twitter. It should be extended to other social networking sites, ecommerce website comments, and blog posts[13].

Mall[14] explained the basic workflow of the financial risk analysis process. Natural language processing is used in the analysis. It is critical to create a system that can extract sentiments or opinions from available text and easily classify the statement. People nowadays have a habit of using casual language and slangs on social media, which makes analysis difficult. As a result, the system must identify and mine slang and conversational language. The proposed system should work on broadening the definition of "fake reviews" or "spam message"[14].

Martin-Oliver et al.[15] addressed the problem of identifying helpful financial risk analysis that will benefit both consumers and businesses. The helpful/unhelpful review threshold can be determined based on the amount of data that the user wishes to prune, and the system is tested on a three-class classification problem. The proposed method yields significant results, demonstrating that helpful reviews can be distinguished from unhelpful ones with high precision. The proposed system is based on human-observed features and implements some of them, but it does not attempt to automatically extract parts of the training corpus[15].

Pokrason[16] discussed data mining, analysis, and its challenges. The mining problems discussed were about different languages and how to approach each language based on its orientation, grouping of synonym words, the direction of opinion based on the situation, and finding spam and fake reviews. The discussion focused primarily on what mining is, the various methods for mining, and the challenges encountered during opinion mining. The solutions to these challenges, as well as the various algorithms that can be used to optimize the mining process, were not discussed in Ref. [17].

Qiu[18] discussed financial risk analysis for loan using a novel Naïve Bayesian Classifier. The data from the user review are analyzed to learn about the user's loans and to learn more about the product. The reviews are preprocessed to remove noise before the words are extracted. Using the Naïve Bayesian Classifier, the extracted terms are classified as positive or negative. Thus, the proposed system can collect useful comments or information from the Twitter website and perform loan analysis on the data effectively using a trained Naïve Bayesian Classifier[18].

Issue wise solution approach

This section contains problem statement, solution approach, and limitations associated with the papers. Some of the techniques have been discussed in terms of research gaps.

The solution is presented in a step-by-step fashion. The first step is to review the text[19], create tokens, and remove noise from the book. The second step consists of subjectivity classification, followed by the extraction step. Saha et al.[19] established two algorithms: (1) for the extraction of explicit feature-loan data, and (2) for the construction of training corpora. Techniques such as Naïve Bayesian, decision trees, Support Vector Machine (SVM), and others are used to model an implicit feature-opinion couple. The TF-IDF representation is used to determine the importance of a term across the entire corpus. Saha et al.[19] used a publicly available customer-review dataset for the experiment, and the results of the experiment are shown using tabular data. Reference [19] focused on a field that has never been adequately considered in the past.

The system proposed in Ref. [20] discovers features for classifying reviews by observing why people classify any thought as useful or not. After that, the components were used to compute Log Support Confidence (LSC). The higher the LSC of the review, the more useful it is to the customer. A threshold is calculated, and any review with LSC greater than the point is considered useful.

The hashtags are tokenism used to create tokens; Ref. [20] assisted in locating sarcasm and its scope within these tokens. The content is used to determine the sentiment analysis area. The analysis of Twitter data is carried out to investigate the effects of sarcasm on sentiment analysis.

The news article is broken down into separate sentences. The type of punishment is scrutinized. The polarity of various types of penalties is then calculated and added to determine the total contradiction of the news article.

3 Methodology

It is the suggested approach in brief in this segment. Figure 1 introduces the flow chart of the approach that can be applied on any dataset to find out the accuracy measures of an ML model.

3.1 Machine learning technique used in credit analysis

3.1.1 Logistic regression

Logistic regression is the statistical analysis to predict the probability that a certain event occurs. It is suitable for binary classification problem, and it is taken between the prediction results from 0 and 1. We performed a quantitative analysis using a linear synthetic function derived from the dependent and independent variables.

Logistic regression model trys to explain how likely an event occurs based on predictive or independent variables $(x_i, i = 1, 2, \ldots, k)$. To do this, establishing that the event to be predicted is binary, the probability of each event is given by Eq. (3).

$$P_i = \frac{1}{1 + e^{-(\beta_2 + \beta_1 x_1 + \cdots + \beta_k x_k)}} \tag{3}$$

3.1.2 Decision trees

Decision trees are used in the areas of economics and decision analysis. However, they also have a great impact on machine learning. The objective of the decision tree is to generate a model that, from a set of variables, is capable of predicting an output value (class). For a better explanation, an example of a decision tree in which the output is a categorical value is shown in Fig. 1. Analyzing the components of the example, you can see that there are 3 types of nodes. First, the rectangles refer to the characteristic that is analyzed at that depth of the tree. The intermediate nodes such as sunny, cloudy, or rainy represent the values that take the variables from which they are hanging, as shown in Fig. 2. Finally, the leaf nodes represent the output, that is, the classification offered by the decision tree.

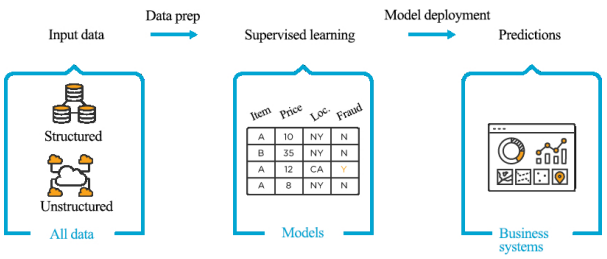


Fig. 1 Supervised learning[1].

The way in which the model is classified is very intuitive, as shown in Fig. 3. Once the model has been learned, the new instance to be classified must meet the criteria from top to bottom until it reaches a leaf node, obtaining a classification of the instance.

The technique used in this paper is to learn the categorization model. In the first phase, the earnings in information (info-gain) 1 of all the variables with respect to the class are computed and sorted. Then, based on the characteristic that has the greatest gain in information, the tree is branched until it reaches a leaf node; At the same moment, process is finished.

3.1.3 Support vector machines

SVMs try to divide the space of characteristics by means of hyper planes in such a way that they act as a boundary between the classes, thus creating as many subspaces as labels. The process to determine these boundaries is based on moving the instances to the space of characteristics F and looking for the hyper plane that separates them. This change of space is done using a kernel, being able to be polynomial, spherical, and linear, among others. Formally, given a dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}_{\mathrm{dy}}, y_i \in \{0, 1\}$, the separating hyper plane has the form $w^{\mathrm{T}}x + b = 0$, where $w \in F$ and $b \in \mathbb{R}$. So that the decision boundary is not over fitted, slack variables are added. Therefore, the objective of the SVM algorithm is to perform the following minimization.

$$\min_{w,b,k} \frac{\| w \|^2}{2} + C \sum x_{i=1}^{n} k_i,$$

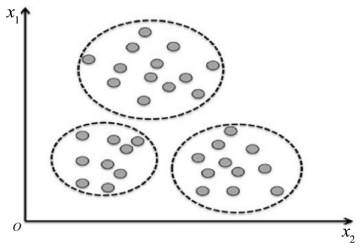


Fig. 2 Clustering in unsupervised learning[21].

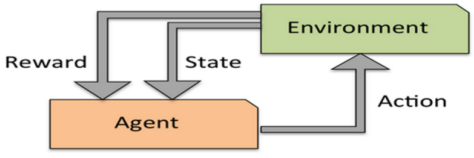


Fig. 3 Reinforcement learning[22] roadmap for building machine learning systems.

$$\begin{aligned} &y_i\left(w^{\mathrm{T}}\phi\left(x_i\right)+b\right) \geqslant 1-\xi_i, \quad \forall i=1,2,\ldots,n; \\ &k_i \geqslant 0, \qquad\qquad\qquad\quad \forall i=1,2,\ldots,n \end{aligned} \tag{4}$$

where $C > 0$ is a constant, large enough chosen by the user, which allows to control to what extent the cost term of non-separable examples influences the minimization of the norm, that is, it will allow to regulate the compromise between the degree of over-adjustment of the final classifier and the proportion of the number of non-separable examples. Thus, a very large C value will allow very small values of ξ. In the limit ($C \to 0$), the case of perfectly separable examples ($\xi \to 0$) would be considered. On the other hand, a very small value of C would allow very large values of ξ, that is, a very large number of misclassified examples would be admitted. In the limit case ($C \to 0$), all examples will be allowed to be misclassified ($\xi \to 0$).

Then, by moving the points back to the original space, the decision boundaries change.

Different forms can be taken as seen in Fig. 4. For classification, it is enough to determine to which subspace the instance belongs and assign it to the class accordingly. It should be noted that the hyper plane separator tries to maximize the distance between instances of different kinds. On the other hand, the separating hyper plane can take different forms, which is defined by the kernel function. As shown above, the decision boundaries are adjusted in such a way that they create a maximum separation between the two classes and these can be different depending on the kernel used.

When it is not linearly separable in feature space, the support vector machine loosens the conditions of linearly separable called soft margin. SVM Gaussian kernel is one of the SVMs and is strong in linear inseparable pattern and data. It is a generic kernel method used

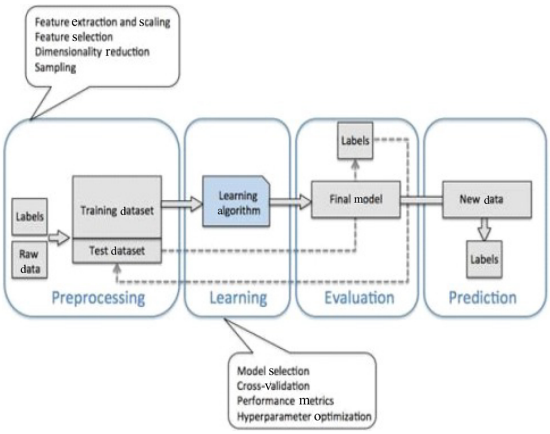


Fig. 4 Steps for building an ML system[2].

when there is no prior knowledge. Linear kernel SVM is a linear support. The training data in door vector machines are sparse, and der data (most of the items in the large-scale 0) are linearly separable. There is no need to map the feature space, and linear support vector machine that does not use a kernel method.

3.1.4 Naïve Bayesian

It belongs to the branch of probabilistic classifiers since they assign a probability of belonging to a class instead of a deterministic assignment. This algorithm is based on Bayesian' theorem, making a strong assumption about the conditional independence of the variables given the class. That is, this model assumes that the variables are conditionally independent of each other given the class. Despite this strong characteristic, the classifier obtains good results and is a good starting point. In addition, given its speed to learn, the model is widely used. Formally, the classifier assigns that class most likely to an instance following the following expression:

$$c^* = \underset{c}{\operatorname{argmax}} \left\{ P(C=c) \prod_{i=1}^{n} P(X_i = x_i \mid C=c) \right\} \tag{5}$$

On the other hand, two underlying problems of this classification model can be highlighted. First, it is showed that the decision boundaries of this model with binary variables are hyper planes. This assumes that the classifier is not capable of generating boundaries that fit the dataset perfectly and may not classify correctly.

Second, it is determined that the classifier is poorly calibrated, that is, it does not have a good Brier score.

3.1.5 K-nearest neighbors

The KNN is a deterministic classifier that is commonly described as a vague classifier (lazy) since a classification model is not generated from the training set but the assignment is done in a way independent for each of the instances.

The classification process is based on assigning the majority class among the k neighbors closest to the instance to be classified. As can be seen in Fig. 5, the classification is dynamic and is performed for each of the instances thus calculating the distances. In the example, the algorithm is based on a KNN and therefore, the assigned class is blue for that instance, as shown in Fig. 6.

3.1.6 Random forests

This random forest classification system is considered a meta-classifier since it uses classification trees as base classifiers. This model follows the bagging technique,

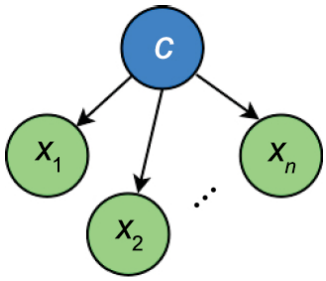


Fig. 5 Naïve Bayesian classification.

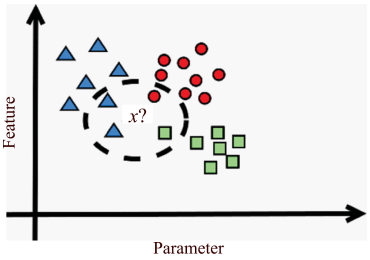


Fig. 6 KNN classification algorithm.

and uses a variety of unstable models which by themselves are very deterministic. Therefore, if trees grow enough they fit perfectly to training data, which, after averaging them, offers good results. The technique in which the base classifiers are constructed is as follows. First, a number of characteristics are selected from all the possible ones with which the classifier will be built. Similarly, a number of instances of the training set are selected. Once the model has been learned, it is tested with the instances that have not been selected in the training stage. This process is done several times until all instances have been selected as training for any of the base trees. To classify, the instances are classified by each of the base classifiers. Next, the exits of each one of them are counted to assign the label that has obtained more votes. Visually, a random forest classifier divides the space of characteristics into boxes always parallel to the axes, as can be seen above. Therefore, it has one disadvantage. In addition, this classifier can be adjusted to the data when the data are not clearly differentiated, as the case with decision trees. By combining a plurality of classifiers by decision tree when performing ensemble learning is the method to choose the attributes selected at each node of the decision tree at 201random.

3.1.7 XgBoost

Gradients are sequentially learned in ensemble learning classifiers by a plurality of decision trees, for the data each classifier is a weak point, to learn combinations of the individual classifiers as more classification ability is high. It is known as boosting.

3.2 Proposed work

This is the support vector machine but only one of the classes. To obtain the decision border, there are two aspects. In the case of the decision boundary in the feature space is considered to be a plane. Otherwise, following the theory of boundaries can be spheres in the feature space.

3.2.1 Isolation forest

The objective of this is to isolate the instances through random divisions of space. For this, first select a feature randomly. The domain of this variable is then randomly selected. This process is carried out as many times as necessary until the instance is completely isolated thus generating a tree. The logic says that it would be easier to isolate the anomalous data since they will be more peculiar data and therefore, isolating these would be done with fewer separations. For this reason, the algorithm calculates an anomaly score that measures how rare this instance is. To do this, count the number of conditions that are required to isolate this instance. This is the score with which it classifies between anomalous and common instances.

This process can be seen in Fig. 7. As can be seen, a large number of decision trees (isolation trees (tree)) have been generated. These are the ones that determine if it is normal for an instance to be isolated under N conditions. Therefore, if as in the first tree, an instance is isolated only in one condition, it is classified as atypical or anomalous.

3.2.2 Mixtures of Gaussian models

It can represent any scenario if a number of Gaussian models are used. Therefore, if the problem of the classification of rare events is understood as events that occur with a very low probability. We can think of modeling this scenario with a mixture of Gaussian models that, being learned by common cases, give a very

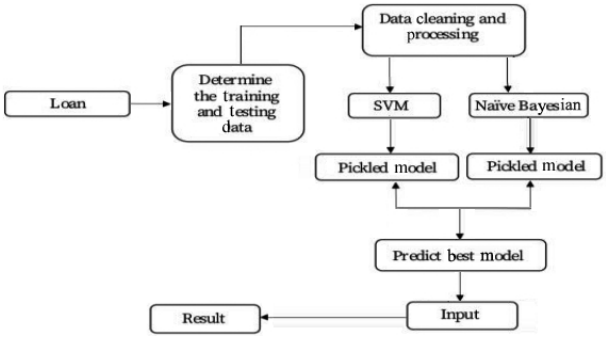


Fig. 7 Proposed work flow chart.

low probability to anomalous events.

Therefore, as seen in Fig. 6, the probability of normal events is much higher than anomalous events. However, the difficulty of adjusting the model comes in determining the amount of mixtures of Gaussian models needed to model the problem. In addition, this complexity is exacerbated as the dimension of the features increases.

3.2.3 KNN for anomaly detection

Problem belongs to the unsupervised classification; the process is more like creating Clusters 3. We can imagine that normal instances will have a greater similarity, and therefore, closeness than that are anomalous. Therefore, as can be seen in Fig. 7, two clusters are discerned in which a set of instances is clearly differentiated from the normal set.

These types of methods can be distances, such as the K-means algorithm that the centroid looks for by minimizing the quadratic Euclidean distance of the centroid to all points in the cluster. However, this method, and all those based on finding the center of a group of instances, is only capable of obtaining clusters with spherical shapes. Therefore, another approach, which is what used in this document, is the DB Scan, which is based on creating sets based on the density of nearby points. One of the characteristics by which this algorithm is used is that those points with low density are considered as atypical.

3.3 Features extraction

Vectorization is a step-in feature extraction. The concept is to obtain some specific features out of the text for the model to train on, by transforming text to numerical vectors. There are plenty of ways to achieve vectorization. For vectorization, we use two techniques. count vectorizer and TF-IDF vectorizer.

3.3.1 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a text vectorizer that converts the text into a vector. It combines 2 concepts, Document Frequency (DF) and Term Frequency (TF). Document frequency is the number of documents containing a specific term and indicates how common the term. The term frequency is the number of occurrences of a specific term in a document and indicates how important a specific term is in a document. Inverse Document Frequency (IDF) is the weight of a term, and aims to reduce the weight of a term if the term's occurrences are scattered throughout all the documents. IDF can be calculated using Algorithm 1.

Algorithm 1 Algorithm for computation of TF-IDF

Input: $\{\text{ai}_1, \text{ai}_2, \ldots, \text{ai}_m\}$, a set of word frequency
Output: FV
$\{f_{21}, f_{22}, \ldots, f_{2n}\}$, set of FF extracted feature
$S = \{s_1, s_2, \ldots, s_n\}$

$$\text{FV} \leftarrow \begin{bmatrix} f_{21} & v_{11} & v_{12} & \ldots & v_{1m} \\ f_{22} & v_{21} & v_{22} & \ldots & v_{2m} \\ f_{31} & v_{31} & v_{32} & \ldots & v_{3m} \\ f_{41} & v_{41} & v_{42} & \ldots & v_{4m} \\ f_{2n} & v_{n1} & v_{n2} & \ldots & v_{nm} \end{bmatrix}$$

$j \leftarrow 0$;
Compute the value of FV for $(j = 0; j \leqslant m; j{+}{+})$
Set S_j activated when ai_j is active to compute
Compute the value of TF-IDF value f_s of $s \in S_j$ using TF-IDF (s, ai),
$k \leftarrow 1$; $j{+}{+}$;
while $(k \leqslant n)$
  for each s in S_j
    if s is same to f_{2k} then
    $v_{ki} \leftarrow f_s$;
    end if
  end for
$K{+}{+}$;
end while
return FV

$$\text{IDF}_i = \log\left(\frac{n}{\text{DF}_i}\right).$$

And TF-IDF as mentioned earlier is $\text{TF} - \text{IDF} = \text{TF} \times \text{IDF}$.

3.3.2 Count vectorizer

Count vectorizer is a great tool provided by the sci-kit-learn library in Python. It converts the text into a vector-based on the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple texts, and we wish to convert each word in each text into vectors (for use in further text analysis). Count vectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample.

4 Result and Analysis

This section describes the results obtained as well as the methodology followed to obtain them. On the other hand, the datasets used are also detailed. In addition, the conclusions obtained from the experiments are presented. Since the datasets modified in the form of Principal Component Analysis (PCA) are used in this project,

firstly, the experiments carried out by comparison of classifier on dataset are detailed. Next, the same analysis of the actual dataset provided by the improved K-mean is described.

4.1 Dataset

As for the other dataset, it has been provided by the Kaggle (Fig. 8). These are dummy data obtained from the transaction database of bank by online user. The data corresponding to entities declare time in seconds, which describes the transaction time by the user, amount which describes the amount of transaction, and last class which describes the transaction are fraud or not in the forms of 0 and 1. With all this, it has 28 anonymized columns. It should be noted that they are modified with Principal Component Analysis (PCA) to balance the data. Furthermore, there are no missing data in dataset. However, although the other entities cannot be classified as legal, this project assumes that those entities that are not in the trout list are legal. This brings the total number of occurrences to 439, and 439 of which are trout. But it should be stressed that even businesses with "legitimate" labels could be up to no good. As describe previously in the section, this data relationship is followed by class that are

- Positive class (+): Not fraudulent (legal) (0).
- Negative class (−): Fraudulent (illegal) (1).

For both columns (anonymized and non-anonymized columns), experimentation has been carried out following the improved KMEAN validation method in the case of supervised learning algorithms which were described in Section 3.1 and the train-test method for the algorithms explained in the section.

In each of the two types of execution, the following descriptions are mentioned below:

- Comparison: Generate the ROC-AUC score of different classifiers which describes the aggregate measure of performance across all possible classifier.
- Accuracy: It is the matrix that represents the predictions of the classifier just apposed with reality.

Regarding the unsupervised classification, different merit figures are obtained given the nature of the classification. However, given that there is information on which set the instances should belong, it is possible to obtain certain merit figures described below:

Homogeneity. It represents the average of the mixture of instances of one and another class in the clusters.

Completeness. It represents that each cluster is complete with instances of a single class.

V-measure. This measure represents the harmonic mean of homogeneity and completeness.

4.2 Comparison and handling the dataset

The time is recorded in the number of seconds since the first transaction in the dataset. Therefore, we can conclude that this dataset includes all transactions recorded over the course of two days. As opposed to the distribution of the monetary value of the transactions, it is bimodal. This indicates that approximately 28 h after the first transaction there was a significant drop in the volume of transactions. While the time of the first transaction is not provided, it would be reasonable to assume that the drop-in volume occurred during the night, as shown in Fig. 9.

4.3 Train and test

This method consists of separating the dataset in two, being one part (between 70% and 85%) for training and the other for testing. It should be noted that this method does not incur biases. However, a large amount of data are needed and sometimes, this can be a problem. The graphically functioning of this technique can be seen in Fig. 10. For this dataset, a separation of 80%

Time (s)	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
74 056.0	1.386	−1.531	−0.973	−2.444	0.731	3.377	−1.535	0.796	−1.893	...	-0.367	−1.056	0.065	0.952	0.261	−0.404	0.029	0.039	102.00	0
123 276.0	−0.056	1.034	−0.457	−0.554	1.037	−0.739	0.998	−0.160	−0.075	...	-0.318	−0.757	0.173	0.516	−0.401	0.107	0.105	0.076	8.99	0
129 258.0	1.991	−0.414	−0.635	0.077	−0.246	0.233	−0.716	0.265	1.139	...	-0.135	−0.394	0.366	0.137	−0.453	−0.631	0.019	−0.042	1.18	0
127 642.0	1.995	−1.002	−2.011	−1.394	0.689	1.044	−0.435	0.222	−0.935	...	0.171	0.329	0.180	−0.944	−0.224	−0.240	−0.020	−0.072	78.00	0
138 975.0	−7.419	−16.461	−8.121	3.036	−4.551	0.453	7.062	−1.854	−1.940	...	3.556	−1.067	−4.700	−0.024	−1.065	−0.230	−1.041	0.767	5026.26	0
168 931.0	−0.393	0.932	−3.957	0.184	−0.625	0.057	4.115	−0.600	−0.665	...	0.396	1.618	0.435	0.568	−0.995	−0.475	0.449	0.006	614.84	0
101 650.0	0.061	1.532	−1.338	−0.085	0.600	0.530	−1.814	−5.457	0.774	...	-3.063	−0.489	0.396	0.095	0.446	0.161	−0.123	0.154	1.98	0
164 000.0	−0.036	0.743	0.058	0.136	1.110	0.407	0.938	0.077	−0.257	...	0.288	0.993	−0.159	0.210	−0.502	−0.661	0.258	0.240	29.00	0
124 267.0	2.124	−1.149	0.068	−0.756	−1.632	−0.547	−1.432	0.059	0.543	...	0.223	0.681	0.294	−0.028	−0.497	−0.251	0.036	−0.037	9.53	0
162 252.0	−0.807	−0.020	−0.220	−1.152	−0.493	−0.643	1.043	0.016	−1.599	...	0.251	0.877	0.214	−0.140	−0.039	−0.107	0.021	0.115	210.00	0

Fig. 8 Description of dataset used in the experiment.

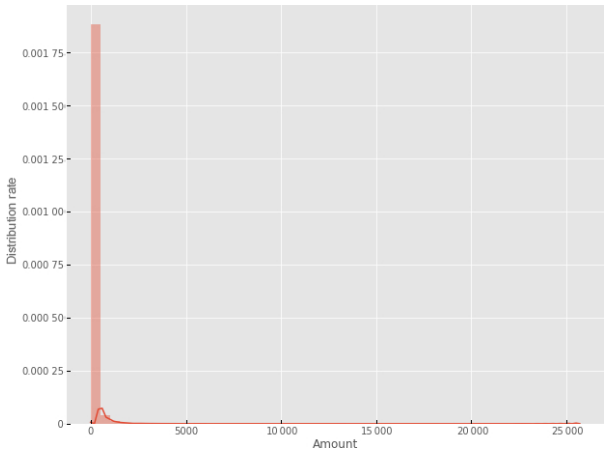


Fig. 9 Distribution of monetary value feature based on machine learning algorithm.

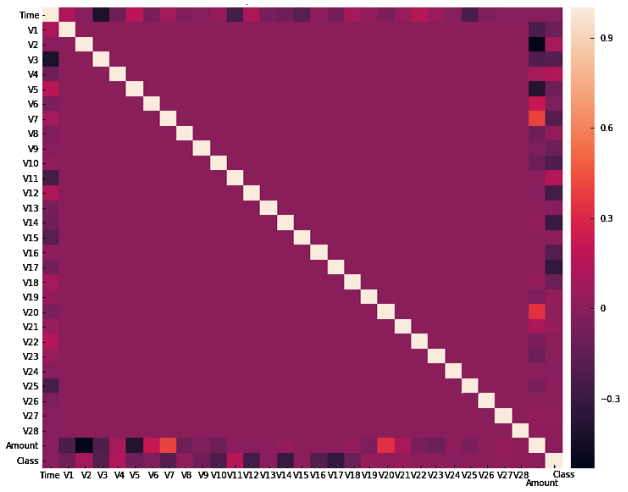


Fig. 10 Distribution of heat map of correlation of monetary value feature.

and 20% was performed for the training and test sets, respectively.

4.4 Result of comparison

In Fig. 11 we see that the cross-validation score for decision tree is the highest so we choose the decision tree model for this dataset. But as in Table 1, other algorithms have better accuracy measure, so for a large dataset, decision tree can give a wrong prediction.

5 Conclusion

Financial companies rely on credit risk prediction because it prevents them from making erroneous evaluations, which might result in squandered opportunities or financial losses. Traditional and current Artificial Intelligence (AI) technologies have been

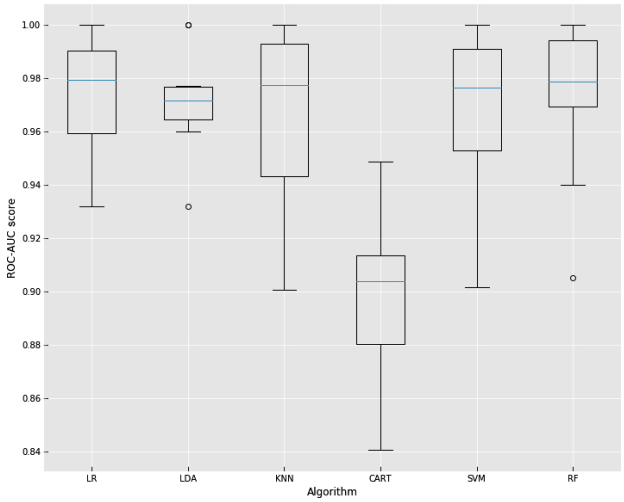


Fig. 11 Comparison of different classification algorithms.

Table 1 Numerical analysis of non-anonymized attributes.

Parameter	Time (s)	Amount
Count	284 807.000	–
Mean	94 813.860	88.350
Std	47 488.146	250.120
Min	0	0
25%	54 201.500	5.600
50%	84 692	22.00
75%	139 320.000	25 691.160

combined to create a hybrid prediction model that is more accurate than using only one methodology alone. Accuracy measurements play a crucial role to decide an ML model for a dataset. Underfitting or overfitting of model can be reduced by setting these measures shown in Table 2. A good ML model always provides accurate predictions and good decisions. Some models are very complex like neural network while some are easy to implement like regression but the difference is the accuracy measures which decides the significance of the model on a dataset shown in Table 3. 6 hybrid models were built by merging logistic regression, discriminant analysis, and decision trees with four types of neural networks: According to various performance metrics outlined in the experiment, inquiry, and statistical analysis, it is possible for the hybrid model to produce a credit risk prediction methodology that is distinct from previous methods. Using five real-world credit score datasets, the classifier was tested and verified.

References

[1] H. A. Abdou, S. Mitra, J. Fry, and A. A. Elamer, Would two-stage scoring models alleviate bank exposure to bad debt? *Expert Systems with Applications*, vol. 128, pp. 1–13,

Table 2 Accuracy measurement of different machine learning algorithms.

No.	Algorithm	Train score	Test score	Roc	Model acc	Recall	Precision	F1-score
1	Logistic regression	0.952 570	0.9547	0.955 750	0.952 000	0.710 145	0.753 846	0.731 343
2	Naïve Bayes	0.881 710	0.8793	0.925 620	0.878 700	0.782 610	0.415 380	0.542 710
3	KNN	0.969 428	0.9594	0.929 500	0.958 700	0.775 400	0.775 400	0.775 400
4	Decision tree	0.881 714	0.8794	0.925 615	0.878 667	0.782 609	0.415 385	0.542 714

Table 3 Comparison of rank of algorithms.

No.	Algorithm	Train score	Test score	Roc	Model acc	Recall	Precision	F1-score
1	Logistic regression	I	I	I	II	IV	II	II
2	Naïve Bayesian	IV	IV	III	III	II	IV	IV
3	KNN	II	II	II	I	III	I	I
4	Decision tree	III	III	IV	IV	I	III	III

2019.

[2] D. Bülbül, H. Hakenes, and C. Lambert, What influences banks' choice of credit risk management practices? Theory and evidence, *Journal of Financial Stability*, vol. 40, pp. 1–14, 2019.

[3] A. E. Celik and Y. Karatepe, Evaluating and forecasting banking crises through neural network models: An application for Turkish banking sector, *Expert Systems with Applications*, vol. 33, no. 4, pp. 809–815, 2007.

[4] J. Chen, A. L. Katchova, and C. X. Zhou, Agricultural loan delinquency prediction using machine learning methods, *International Food and Agribusiness Management Review*, vol. 24, no. 5, pp. 797–812, 2021.

[5] A. Cozarenco and A. Szafarz, The regulation of prosocial lending: Are loan ceilings effective? *Journal of Banking & Finance*, vol. 121, p. 105979, 2020.

[6] J. C. Cuestas, Y. Lucotte, and N. Reigl, Banking sector concentration, competition and financial stability: The case of the baltic countries, *Post-Communist Economies*, vol. 32, no. 2, pp. 215–249, 2020.

[7] H. D. Silva, E. J. Dockner, R. Jankowitsch, S. Pichler, and K. Ritzberger, Choice of rating technology and loan pricing in imperfect credit markets, *Journal of Risk*, vol. 17, no. 1, pp. 29–62, 2014.

[8] S. K. Depren and M. T. Kartal, Prediction on the volume of non-performing loans in Turkey using multivariate adaptive regression splines approach, *International Journal of Finance & Economics*, vol. 26, no. 4, pp. 6395–6405, 2021.

[9] A. M. Dima and S. Vasilache, Credit risk modeling for companies default prediction using neural networks, *Romanian Journal of Economic Forecasting*, vol. 19, no. 3, pp. 127–143, 2016.

[10] G. Enjolras and P. Madies, The role of bank analysts and scores in the prediction of financial distress: Evidence from French farms, *Economics Bulletin*, vol. 40, no. 4, pp. 2978–2993, 2020.

[11] P. Gandhi, T. Loughran, and B. McDonald, Using annual report sentiment as a proxy for financial distress in US banks, *Journal of Behavioral Finance*, vol. 20, no. 4, pp. 424–436, 2019.

[12] K. Halteh, K. Kumar, and A. Gepp, Using cutting-edge tree-based stochastic models to predict credit risk, *Risks*, vol. 6, no. 2, p. 55, 2018.

[13] Z. Y. Li, J. Crook, and G. Andreeva, Dynamic prediction of financial distress using malmquist DEA, *Expert Systems with Applications*, vol. 80, pp. 94–106, 2017.

[14] S. Mall, An empirical study on credit risk management: The case of nonbanking financial companies, *Journal of Credit Risk*, vol. 14, no. 3, pp. 49–66, 2018.

[15] A. Martin-Oliver, S. Ruano, and V. Salas-Fumás, How does bank competition affect credit risk? Evidence from loan-level data, *Economics Letters*, vol. 196, p. 109524, 2020.

[16] S. Pokrason, A healthy dose of pessimism? Influence of the Ukrainian economy on its banking sector credit ratings, *Baltic Journal of Economic Studies*, vol. 3, no. 4, pp. 216–223, 2017.

[17] Preeti and S. Roy, Application of hybrid approach in banking system: An undesirable operational performance modelling, *Global Business Review*, doi: 10.1177/09721509211026789.

[18] W. W. Qiu, Enterprise financial risk management platform based on 5G mobile communication and embedded system, *Microprocessors and Microsystems*, vol. 80, p. 103594, 2021.

[19] P. Saha, I. Bose, and A. Mahanti, A knowledge based scheme for risk assessment in loan processing by banks, *Decision Support Systems*, vol. 84, pp. 78–88, 2016.

[20] V. Srivastava, Distressed debt investments in India: What more needs to be done to strengthen regulations? *International Journal of Indian Culture and Business Management*, vol. 18, no. 3, pp. 368–380, 2019.

[21] H. A. Abdou, M. D. D. Tsafack, C. G. Ntim, and R. D. Baker, Predicting creditworthiness in retail banking with limited scoring data, *Knowledge-Based Systems*, vol. 103, pp. 89–103, 2016.

[22] A. Bhimani, M. A. Gulamhussen, and S. D. R. Lopes, Accounting and non-accounting determinants of default: An analysis of privately-held firms, *Journal of Accounting and Public Policy*, vol. 29, no. 6, pp. 517–532, 2010.

Vikas Kumar received the PhD degree in business administration with specialization in finance from Kurukshetra University, Kurukshetra, India. He is currently working as an assistant professor in the Humanities and Management Department, Dr. B. R. Ambedkar National Institute of Technology Jalandhar. Currently his job responsibilities include teaching UG and PG students, guiding PhDs and capstone projects, etc. He has been a senior research fellow at Kurukshetra University. He has qualified UGC-NET in management. He has written many research articles on the topics of behavioral finance and capital markets and valuation.

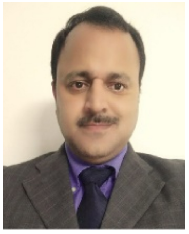
Shaiku Shahida Saheb is currently working as an assistant professor at the Mittal School of Business, Lovely Professional University in India. He was previously worked at National Institute of Technology Srinagar, J&K, India. He was a recipient of best research paper award in NIT, Srinagar. He was a former assistant professor at Ethiopian Civil Service University, Addis Ababa, Ethiopia for six years. He was nominated for best expatriate recognition in the university. He had a total of 20 years of teaching experience in India and aboard. He wrote 2 books and published 8 research papers. He has taught PhD scholars, master, and undergraduate students especially in the area of accounting and finance. He has guided one PhD student as co-advisor and supervised more than 100 master theses and projects.

Atif Ghayas received the PhD degree in business administration with specialization in finance on the topic "A study of the relationship between capital structure and profitability on the selected firm in India" from Aligarh Muslim University, Aligarh, India. He has been a senior research fellow at AMU. He is currently working as an assistant professor in the School of Management, Gitam (to be deemed university), Bangalore, India. Currently his job responsibilities include teaching UG and PG students, guiding PhDs and capstone projects, etc. He has qualified UGC-Junior Research Fellowship in management and commerce, NET in management, commerce, human resource management, and economics. He has written many research articles on the topic of capital structure and capital markets and valuation. He has won best research paper award in the international conference "ASBIC 2021".

Jai Kishan Chandel is working as an assistant professor in the Institute of Management Studies, Kurukshetra University, Kurukshetra, India.

Preeti received the PhD degree in management from Kurukshetra University, Kurukshetra, India. She works at the Department of Commerce & Business Administration, Kanya Maha Vidyalaya (KMV), India. She has been a senior research fellow at Kurukshetra University. She has qualified UGC-NET in commerce. She has written many research articles in national and international journal.

Sunil Kumari received the PhD degree in commerce from Kurukshetra University, Kurukshetra, India. She is currently working as an assistant professor in the Government College for Women, Indra Gandhi University. She has been a senior research fellow at Kurukshetra University. She has qualified UGC-NET in commerce. She has written many research articles in national and international journals.



Santosh Kumar received the PhD, MPhil, and MBA degrees from Birla Institute of Technology, Mesra. He is working as an associate professor with Jaipuria Institute of Management, Jaipur. He is a prolific writer, published 40+ research papers, articles, and chapters in international peer-reviewed journals indexed in Scopus, ABDC, etc. He has presented papers at many international conferences. His research interest is in finance and technology area.

Saroj Kumar Pandey received the BTech degree from University of Allahabad in 2008, the MTech degree in computer technology from National Institute of Technology Raipur in 2011, and the PhD degree in information technology from National Institute of Technology Raipur in 2021. He is serving as an assistant professor at the Department of Computer Engineering and Applications, GLA University, Mathura. He has published more than 20 research papers in international and national journals and conferences. He also published four book chapters in reputed publication. His areas of research include deep learning, machine learning, biomedical healthcare system, expert systems, neural networks, hybrid computing, and soft computing. He has guided various research projects in UG level.