Credit Risk Prediction in Digital Banking Using Hybrid Machine Learning Techniques

Ramin Farrokhi[1*], Milad Sarvestani[2]

[1]Universal Technology Company, Chief Financial Officer

[2]Universal Technology Company, Assistant Financial Manager

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.

Keywords

Artificial Intelligence (AI)Machine Learning (ML)loan predictionSupport Vector Machine (SVM)Random Forest (RF)accuracy.

Introduction

As the financial market competes sharply nowadays, the traditional banking profit is largely reduced. This causes banks to focus on consumer banking in order to make higher interest profits, i.e. consumer loans. However, the quality of issuing consumer loans is bank dependent and the audit process is forced to be as simple as possible. Consequently, the potential risk gradually arises.

With the rapid growth in credit industry and the management of large loan portfolios, credit rating (or credit scoring) models have been extensively used for the credit admission evaluation. The credit rating models are developed to classify loan customers as either a good credit group (accepted) or a bad credit group (rejected) with their related characteristics such as age, income and martial status or based on the data of the previous accepted and rejected applicants [1]. The benefits of considering credit scoring include reducing the cost of credit analysis, enabling faster decisions, insuring credit collections, and diminishing possible risks [24]. Even if a slight improvement in credit scoring accuracy might reduce large credit risks and translate into significant future savings.

The traditional approach to predict the consumers' credit risk is based on some statistical methods, such as logistic regression. However, related studies have shown that machine learning techniques or data mining techniques, such as neural networks, decision trees, etc., are superior to traditional (statistical) methods [2], [9], [22]. That is, using machine learning techniques can provide higher predication accuracy.

In machine learning, the hybridization approach has been an active research area to improve the classification/prediction performance over single learning approaches [8], [11], [13], [17], [19]. In general, it is based on combining two different machine learning techniques. For example, a hybrid classification model can be composed of one unsupervised learner (or cluster) to pre-process the training data and one supervised learner (or classifier) to learn the clustering result or vice versa [18].

Therefore, to develop a hybrid learning credit model, there are four different ways to combine the two machine learning techniques. They are: (1) combining two classification techniques, (2) combining two clustering techniques, (3) one clustering technique combined with one classification technique, and (4) one classification technique combined with one clustering technique.

In literature, related work developing credit rating models based on hybrid machine learning techniques only compare with some chosen single learning based models as the baselines (see Section 2.4). That is, none of the existing studies compares different hybrid models to identify which hybrid approach can perform the best for credit rating in terms of high prediction accuracy and low error rates.

Therefore, the aim of this paper is to examine the prediction performance of these four types of hybrid learning models for credit rating in addition to the single classification and clustering techniques as the baseline models. Moreover, the profit made by these models is also compared based on a chosen bank as the case. Four well-known classification techniques, which are decision trees, Bayes classification, logistic regression, and neural networks, and two clustering techniques, which are *K*-means and expectation maximization are used to develop the hybrid models. Therefore, the contribution of this paper is to find out which combination method and techniques for the hybrid learning model can perform the best as well as provide maximum profits for credit rating. Credit risk management is one of the most critical functions of financial institutions, particularly in the banking sector, where lending activities constitute a major source of income and exposure to risk. Credit risk refers to the possibility that a borrower will fail to meet contractual obligations, resulting in financial losses for lenders. Effective credit risk assessment is therefore essential for ensuring financial stability, safeguarding capital adequacy, and maintaining long-term profitability (OECD, 1999; Williamson, 1999).

Traditionally, credit risk evaluation has relied on statistical and econometric models such as logistic regression, linear discriminant analysis, and basic decision tree techniques. While these methods have been widely adopted due to their simplicity and interpretability, they are constrained by assumptions of linearity and normality. As financial markets become more complex and data-driven, these traditional approaches often fail to capture non-linear patterns, complex interactions, and high-dimensional relationships inherent in modern financial data (Rezaee & Mimmier, 2003; Messier, 2000).

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has introduced powerful alternatives for financial risk modeling. Algorithms such as Support Vector Machines (SVM), Random Forests (RF), and Decision Trees (DT) offer enhanced flexibility and robustness in modeling complex, non-linear relationships. These techniques have demonstrated strong performance in predicting financial distress, improving audit quality, and assessing institutional risks, particularly when dealing with large and heterogeneous datasets (Apak & Ganji, 2025; Ganji, 2024).

More recently, deep learning models—especially Convolutional Neural Networks (CNNs)—have gained attention for their superior feature extraction capabilities. Although CNNs are traditionally associated with image and signal processing, their ability to automatically learn hierarchical representations makes them suitable for structured financial data as well. However, deep learning models often suffer from limited interpretability, which remains a concern in regulated financial environments (Ganji & Ganji, 2025).

To address this limitation, hybrid modeling approaches that integrate deep learning with classical machine learning algorithms have emerged as a promising solution. By combining the feature extraction power of CNNs with the classification strengths of algorithms such as SVM, RF, and DT, hybrid models can achieve higher predictive accuracy while maintaining interpretability through feature importance and explainable AI techniques. Prior studies have shown that such hybrid frameworks outperform standalone models in risk assessment, fraud detection, and financial decision-making contexts (Mehmet & Ganji, 2021; Ganji, 2025).

Furthermore, global economic disruptions—such as the COVID-19 pandemic—have underscored the need for adaptive and resilient credit risk assessment systems. Financial uncertainty, increased default rates, and volatile borrower behavior necessitate more advanced predictive tools capable of responding to rapidly changing conditions. Machine learning and bio-inspired algorithms have proven effective in addressing these challenges by enabling proactive risk identification and improved decision-making (Ganji, 2024; Ganji, 2025).

Against this background, the present study proposes a hybrid credit risk prediction model that integrates Convolutional Neural Networks with Support Vector Machines, Random Forests, and Decision Trees. The primary objective is to enhance the accuracy, robustness, and practical applicability of credit risk assessment models in banking environments. By leveraging both deep learning and traditional machine learning techniques, this research aims to contribute to the growing body of literature on intelligent financial risk management and provide actionable insights for financial institutions.

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).

(1)

$$xi(t+1) = c \cdot [\omega \cdot xi(t) + (1-\omega) \cdot xi(t-1)] +$$

$$(1-c) \cdot 1 \sum j \epsilon \mid NiiNixj(t)$$

(2)

$$Ni(t) = \{j \in V || xi(t) - xj(t)| \leqslant \in i\}$$

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. 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 ($xi,\ i\ =\ 1, 2, \ldots,\ k$). To do this, establishing that the event to be predicted is binary, the probability of each event is given by Eq. (3).

(3)

$$Pi = 11 + e - (\beta 2 + \beta 1x1 + \cdots + \beta kxk)$$

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 Finally, the leaf nodes represent the output, that is, the classification offered by the decision tree.The way in which the model is classified is very intuitive, as shown in 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={(x1, y1), (x2, y2),…,(xn,yn)} where xi∈Rdy,yi∈{0,1}, the separating hyper plane has the form wTx+b=0, where w∈F and b∈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.

(4)

$$minw, b, k \parallel w \parallel 22 + C\sum xni = 1ki,$$

$$yi(wT\phi(xi) + b) \geqslant 1 - \xi i, ki \geqslant 0, \forall i = 1,2, \dots, n; \forall i = 1,2, \dots, n$$

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→0), the case of perfectly separable examples (ξ→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→0), all examples will be allowed to be misclassified ($\xi \rightarrow 0$).

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 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:

(5)

$$c *= argmaxc\{P(C = c)\prod i = 1nP(Xi = xi \mid C = c)\}$$

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.

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, 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 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 low probability to anomalous events.

Therefore, as seen 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, 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.

$$IDFi = log(nDFi).$$

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

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 .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.

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

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

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.

Table 1Numerical 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

Parameter	Time (s)	Amount
25%	54 201.500	5.600
50%	84 692	22.00
75%	139 320.000	25 691.160

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 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.

Table 2Accuracy 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 3Comparison 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

References :

Mehmet, H., & Ganji, F. (2021). Detecting fraud in insurance companies and solutions to fight it using coverage data in the COVID-19 pandemic. PalArch's Journal of Archaeology of Egypt / Egyptology, 18(15), 392–407.

https://scholar.google.com/citations?view_op=view_citation&hl=tr&user=_RyCeTEAAAAJ&citation_for_view=_RyCeTEAAAAJ:Y0pCki6q_DkC

Ayboğa, M. H., & Ganii, F. (2022). The Covid 19 Crisis and The Future of Bitcoin in E-Commerce. Journal of Organizational Behavior Research, 7(2), 203-213. https://doi.org/10.51847/hta7Jg55of

GANJI, F. (2024). LEVERAGING BIO-INSPIRED ALGORITHMS TO ENHANCE EFFICIENCY IN COVID-19 VACCINE DISTRIBUTION. TMP Universal Journal of Research and Review Archives, 3(4). https://doi.org/10.69557/ujrra.v3i4.103.

Ganji, F. (2024). Incorporating emotional intelligence in shark algorithms: Boosting trading success with affective AI. TMP Universal Journal of Research and Review Archives, 3(4). https://doi.org/10.69557/ujrra.v3i4.105

Ganji, F., & Ganji, F. (2025). The Role of Sports Sponsorships in Shaping Financial Strategy and Accounting Practices. International Journal of Business Management and Entrepreneurship, 4(2), 86–99. Retrieved from https://www.mbajournal.ir/index.php/IJBME/article/view/79

Ganji, F., & Ganji, F. (2025). The Impact of Financial Reporting Standards on Sports Franchise Valuation. International Journal of Business Management and Entrepreneurship, 4(1), 46–60. Retrieved from https://mbajournal.ir/index.php/IJBME/article/view/64

GANJI, F. (2024). ASSESSING ELECTRIC VEHICLE VIABILITY: A COMPARATIVE ANALYSIS OF URBAN VERSUS LONG-DISTANCE USE WITH FINANCIAL AND AUDITING INSIGHTS. TMP Universal Journal of Research and Review Archives, 3(4). https://doi.org/10.69557/ujrra.v3i4.107

GANJI, F. (2025). Exploring the Integration of Quantum Computing and Shark Algorithms in Stock Market Trading: Implications for Accounting, Finance and Auditing. International Journal of Business Management and Entrepreneurship, 4(2), 40–56. Retrieved from https://mbajournal.ir/index.php/IJBME/article/view/76

Joshua, R. (2006). A proposed corporate governance reform: Financial statements insurance. Journal of Engineering and Technology Management, 23(1-2), 130-146.

Jumaa, A.H. (2003). Institutional control and dimensions of development in the context of the practice of the internal auditing profession: The Jordanian association of certified public accountants, the fifth professional scientific conference, under the slogan of Institutional Control and Establishment Continuity. Amman, Jordan.

OECD. (1999). Principles of corporate governance, organization for economic co-operation and development publications service. Retrieved from http://www.oecd.org.

Rezaee, Z.K., & Mimmier, G. (2003). Improving corporate governance: The role of audit committee disclosures. Managerial Auditing Journal, 18(6/7), 530-537.

GANJI, F. (2025). Biomimetic Shark Algorithms: Leveraging Natural Predator Strategies for Superior Market Performance and Advanced Accounting Techniques. International Journal of Business Management and Entrepreneurship, 4(2), 57–70. Retrieved from https://mbajournal.ir/index.php/IJBME/article/view/77

APAK, . R., & GANJI, F. . (2025). Using Decision Tree Algorithms and Artificial Intelligence to Increase Audit Quality: A Data-Based Approach to Predicting Financial Risks. International Journal of Business Management and Entrepreneurship, 4(1), 87–99. Retrieved from https://mbajournal.ir/index.php/IJBME/article/view/65

GANJI, F. (2025). Investigating Neuroscience-Inspired Shark Algorithms: Mimicking Human Decision-Making in Trading Systems and Their Implications for Accounting, Risk Management, Technical Analysis, and the Stock Market. International Journal of Business Management and Entrepreneurship, 4(2), 71–85. Retrieved from https://www.mbajournal.ir/index.php/IJBME/article/view/78

Ganji, F. (2021). Knowledge and acceptance and use of technology in accounting students. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(13), 7466–7479. https://scholar.google.com/citations?view_op=view_citation&hl=tr&user=_RyCeTEAAAAJ&citation_for_view=_RyCeTEAAAAJ:4DMP91E08xMC

Saudi organization for certified public accountants, (2007). Professional conduct and ethics guide.

Williamson, Q.E. (1999). The mechanism of governance. Oxford: Oxford University Press.

Zewailf, I., & Al-Jawhar, K. (2007). The role of adherence to the elements of internal control in strengthening the pillars of institutional control. Journal of Research and Human Studies, 232-258.

Zingales, L. (1997). Corporate governance. Retrieved from http://www.theiia.org.

GANJI, F. . (2025). Advanced Accounting Techniques for Pandemic-Era Financial Reporting. International Journal of Business Management and Entrepreneurship, 4(1), 36–45. Retrieved from https://mbajournal.ir/index.php/IJBME/article/view/63

Khoury, N.S. (2003). The accounting profession between financial stumbling and institutional control in companies. United Arab Emirates: Al Bayan Newspaper.

Kuwait Stock Exchange. (2010). Public relations department. Annual bulletin.

Malhotra, N. (2003). Marketing research, (4th Edition). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Mangena, M., & Pike, R. (2005). The effect of audit committee shareholding, financial expertise and size on interim financial disclosures. Accounting and Business Research, 35(4), 327-549.

Matar, M. (2003). The role of disclosure of accounting information in promoting and activating institutional arbitration. The Jordanian Association of Certified Public Accountants, the Fifth Professional Scientific Conference. Amman: Jordan.

Matar, M., & Noor, A.N. (2007). The extent of Jordanian public shareholding companies' commitment to the principles of corporate governance: a comparative analytical study between the banking and industrial sectors. Jordanian Journal of Business Administration, 3(1), 46-71.

Messier, Jr W.F. (2000). Auditing & assurance services: A systematic approach, (second edition). New York: McGraw-Hill Companies, Inc.