

# *Data Mining Based Classifiers for Credit Risk Analysis*

Armend Salihu

*South East European University, North Macedonia*

*ar.salihu@gmail.com*

Visar Shehu

*South East European University, North Macedonia*

*v.shehu@seeu.edu.mk*

In order to pay back the principal borrowed from the depositary bank, the interest collected by principal creditors will be collected. In this paper, we have presented the main classifiers which are used in credit evaluation. From the research, we have noticed that there are some classifiers who find application in the credit allocation decision. Credit risk assessment is becoming a critical field of financial risk management. Many approaches are used for the credit risk evaluation of client data sets. The evaluation of credit risk data sets leads to an option of cancelling the loan or refusing the request of the borrower, which requires a detailed examination of the data set or of the customer's data. This paper discusses various automatic methods of credit risk analysis used for the estimation of credit risk. The data mining method was defined with the emphasis on different algorithms, such as neural network, and as the most widely employed approach for credit risk analysis.

*Key Words:* banking loan analysis, classifiers, credit risk analysis, machine learning, data mining

*JEL Classification:* E51, G21, G32, H81

*<https://doi.org/10.26493/1854-6935.18.147-167>*

## **Introduction**

One of the most important and crucial procedures for banks is the loan allocation mechanism for approving requests for loans. This vital process consists of the compilation, review, and final credit evaluation of the different factors used to determine credit applications from sources such as credit application forms, interbank data exchange, credit office information, and the relevant in-bank information (Brown and Zehnder 2007).

In the conventional assessment approach, banks accept or decline applications from commercial and/or retail customers, typically discretionary judgments of credit allocation specialists. The 'credit scoring

technique' or more generally referred to as the 'Scorecard' is also an approach for evaluating requests for loans. The scorecards would estimate the probability that the borrower will repay the credit he/she requested and address the issue of whether or not the loan fails at any time. An application scorecard is generated in terms of credit risk based on previous statistics, by classifying it as either good or bad. Applications for past loans are analyzed to classify characteristics that have a substantial effect on the discrimination of good and bad credit risk applications (Anderson 2007). The advantages of using credit allocation scorecards can be summarized as better decisions, quicker and easier, more logical loan choices, observable risk management and decreased transaction costs.

Data mining is a major area of research, aiming at the collection and analysis of large amounts of data with machine learning. Data mining is currently popular in the banking sector, as there are successful analytical techniques for identifying knowledge related to credit scores. The primary goal is to exploit a huge amount of information available in databases to make key decisions (Hamid and Ahmed 2016; Sudhakar and Reddy 2014).

Decision making is an important element in the selection of lenders with a lot of information and experience. Identifying less risky borrowers is key to good credit scores. The financial sector is not only unpredictable, but it is also very competitive (Altman 1968). The possibility of failure to provide the desired results is significant for lenders, and lenders must take adequate measures to remedy the situation and mitigate the risks in order to obtain favorable debt payment (Akkoc 2012). The selection procedure for lenders can contribute to the identification of good lenders who have the willingness and capacity to repay debt during the fixed payback period. Any of the credit scores approaches currently in operation are criticized for failing to address the lender's ability to repay the debt at reduced risk within a specified time (Chen et al. 2012).

The debtor is obliged to pay the principal and additional interest if a loan has been issued by a bank or a financial institution. The most significant factor is the amount borrowed and the interest rate of the bank. Loans are typically secured or not secured. A loan secured requires a loan obligation for an estate, such as a car, home, etc. If the borrower makes an error or fails to repay the loan, the lender is entitled to the estate. The unsecured credit option is not preferred, nor is it common. If the creditor fails to repay the unsecured loan, the lender has no right to withdraw anything.

When a customer loan is accepted, the bank first reviews the customer profile and documents. Each bank has a credit score for each customer, represented as numbers based on the borrower's credit file (Chopde et al. 2012). Credit scoring is a compilation of a variety of functions, mechanisms, and requests which must be considered in full (Hájek 2011). Accordingly, credit funding decisions in such situations are challenging and complex (Malhotra and Malhotra 2003). Systems are therefore important to describe such complex scenarios. The findings show that the implementation of certain decision-making parameters will make it easier to answer these questions (Louzada, Ara, and Fernandes 2016). Some models such as furrowed logic, deep neural networks, and decision tree models specifically consider and identify the underlying connections, such as operational risks. Additional techniques such as fuzzy sets or computational models are increasingly being used in credit score, alongside the multi-criterion approaches. These instruments were included in the decision-making of the lender to deal with the financial risk aspects. The ability of these instruments to resolve ambiguity makes it possible for the lender to address the issues in a manner not protected by traditional approaches (Zurada 2010). This method is referred to as a loan assessment that takes time but is typically a binary decision resulting in approval or refusal. The credit process fails under the circumstances that are at the core of the bank. The two main reasons for the need for an expert support system are the absence of accurate methods of measuring the lack of public credit risk system and credit risk in many banks (Sudhakar and Reddy 2014).

There are many risks associated with bank loans, for the bank and for those who get loans. The risk assessment of bank loans should be understood as the nature of the risk (Sudhakar and Reddy 2014). Credit risk is the risk that the loan will not be repaid on time or at all; liquidity risk is the risk that too many deposits will be withdrawn too soon, leaving the bank short of immediate cash; and interest rate risk, the risk that interest rates on bank loans will be too low to earn the bank significant funds (Hamid and Ahmed 2016).

In ranking candidates in both good and bad grades, the banking system tests the reliability of the data sets. Applicants in the good groups are more likely to repay the cashback to the bank. Applicants in the bad class seem unable to return cash to the bank and are thus defaulting on loans. Various forms of credit risk management approaches are used to minimize the defaulter rate of credit data (Bask et al. 2011). Even with a slight

improvement in the accuracy of the loan assessment, major losses may often be minimized. The benefits of a reliable credit risk dataset are that the cost of credit ratings is minimized, excellent decision-making takes less time, and the risk of borrowing is avoided. Since credit risk assessment plays a key role in the banking sector and is very important and a major challenge facing banks, the correct classification of credit information in order to avoid economic losses plays a significant role. The decreased default rate in the non-reliable credit risk data set gives inspiration to this sector (Pandey et al. 2013).

In the following, in the second chapter are presented algorithms for credit risk analysis, divided in categories: Classification Algorithms, Grouping Algorithms, Neural Networks and The combinations of Classifiers. In the third part is presented the utilization of data mining algorithms for credit risk analysis also divided in categories like the second chapter. In the fourth part are presented Discussion, Implication, and Conclusion.

### **Algorithms for Credit Risk Analysis**

For the better and more reliably analyzed credit risk, different kinds of methods are used for the evaluation of credit datasets.

#### **CLASSIFICATION ALGORITHMS**

*Bayesian Classifier.* The Bayesian network is referred to as the belief network. Bayesian is an acyclic direct graphic or Directed Acyclic Graph statistical model (Pandey et al. 2013). Each graph node displays a random variable in which the edges reflect the respective variable's functions. This is based upon the presupposition that the decision problem is formulated in probability terms, and that all applicable probability values are generated by the decision theory.

The Bayesian model is one of the most common test methods used in data mining. Coined by Thomas Bayes (1763) and complemented by Pierre-Simon Laplace (1812) in a paper published two years after his death:

$$P(h|d) = \frac{p(D|h)P(h)}{P(D)}, \quad (1)$$

where  $P(h)$  is the prior probability of hypothesis  $h$ -prior,  $P(D)$  is the prior probability of training data  $D$ -evidence,  $P(D|h)$  is the probability of  $D$  given  $h$ -likelihood, and  $P(h|D)$  the is the probability of  $h$  given  $D$ -posterior probability.

*Naïve-Bayes Classifier.* It is a simple probabilistic classification based on Bayes theorem. This classification is called naive because it assumes that one class is independent (Pandey et al. 2013; Huang, Chen, and Hsu 2004). Another important classification variable is necessary for the Naïve Bayes Classifier, a small number of data to calculate parameters such as mean and variance.

There are several types of Naïve Bayes Classifiers:

- *Multinomial Naïve Bayes* is primarily used for issues with document classification, i.e. whether a document is a sport, political, technical etc. category. The features/predictors used by the classifier are the frequency of the terms present in the text.
- *Bernoulli Naïve Bayes* is similar to the naive multinomial bayes, but Boolean variables are the predictors. The parameters used to predict the class variable only take value Yes or No whether a word is present in the text or not.
- *Gaussian Naïve Bayes* When the predictors are continuous and non-discrete, it is concluded that the values from the Gaussian distribution are being sampled.

*Bayesian Belief Network.* The chain rule states that a Bayesian Network represents the general distribution of all variables represented by the Directed Acyclic Graph. For each node of the network, it is possible to measure the margins and conditions of probabilities.

A Bayesian Network consists of stochastic nodes that are either unknown, unknown or latent variables, which can be observed. In the first case, the distribution of the node is indicated by the distribution of the observation error or the distribution of data, while in the other case pre-dispositions for the nodes are defined. Each node has a subset of the other nodes that affect the respective distribution. The Bayesian networks have their relationships to decide the nodes and the network is a directed acyclic graph. A node relies only on its nodes of the 'parent.' Bayesian networks therefore have a local property in Markov. Because of its parent, each node is dependent on other nodes.

*Decision Tree.* The Decision Tree is a model for mapping the comment on a branch element to achieve a target in leaves. This is one of the supervised techniques best monitored. This learning method marks each internal node or non-leaf node with an input function. A class or probabilistic class distribution is available to each leaf node in the tree (Bask et al. 2011;

Curran and Orr 2011). The branches between the nodes indicate potential values in these characters, while the terminal nodes indicate the ultimate value of the factors to be observed (Wang, Wang, and Lai 2005).

*Binary Decision Trees.* Non-linear multi-stage classifier are binary decision trees. This classification scheme functions until the right class is identified by sequentially rejecting classes. The feature area is divided into regions that fit the various groups. When large numbers of classes are provided this classification system will be highly useful. The tree is searched sequentially and a decision is taken on the form  $x_i \leq \alpha$ , with  $x_i$  being a function of the axis, and  $\alpha$  being a threshold value, in each node. This processing scheme is an important part of many tree-based vector-quantification algorithms. The processing scheme is a key component of the measurement period

*Fuzzy Decision Tree.* Fuzzy decision tree is one of the most common data mining and machine learning algorithms. The decision tree is called a tree structure classifier, where each node is either a leaf indicating the target value attribute (class) or a decision which specifies for each potential result one branch and sub-tree (Wang, Yeung, and Tsang 2001).

*Support Vector Machine (svm).* The support vector machine is another supervised learning method testing data with a corresponding learning algorithm for classification and regression. svm constructs a high- or infinite-dimensional plane or set of hyper-planes that can be used for classification, regression and other tasks (Chorowski, Wang, and Zurada 2014). svm was first introduced by Vapnik in 1995 for machine learning and proven its usefulness in several fields. A good distance from the closest training information point (functional margin) accomplishes a good separation (Danenas and Grasva 2015). The greater the margin, the lower the classification mistake (Huang, Chen, and Hsu 2004). svm is useful in text and hypertexts, as its implementation can substantially reduce inductive and transductive requirements for labeled training instances (Hearst et al. 2008). The svm can also be used to classify pictures. Many experimental results indicate that svm is more reliable than conventional query refinement (Hui and Yang 2011). svm can be used to identify handwritten characters.

The decision function  $f(x)$  is given by:

$$f(x) = \text{sgn}(w, x + b). \quad (2)$$

To compute optimal hyper-plane the optimized problem is to be solved;

minimization  $1 \div 2\|w^2\|$ , subject to:  $y_i = ((w \cdot x_i) + b - 1) \geq 0$ . The margin of hyper-plane  $= 2 \div \|w\|$  equivalent optimization is:

$$\sum_{i=1}^l y_i \alpha_i = 0. \tag{3}$$

GROUPING ALGORITHMS

*K-Nearest Neighbor (KNN)* is the non-parametric regression and classification method. This involves a favorable and unfavorable package of preparation. It is also called the lazy algorithm. It does not use any data point for generalization (Islam, Wu, and Ahmadi 2007). This ensures that the training stage is really quick and all knowledge is stored. All training data are required during the test process. A measured distance is used to assess the size of a training data collection (Huang, Chen, and Siew 2006). For the valued entrance variable, the most common Euclidian distance is used.

Euclidean distance is calculated as:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}. \tag{4}$$

However, other distance or resemblance metrics such as Minkowski Distance, Jaccard, Simple Matching Coefficient, Cosine Similarity etc. can be used depending on the data collection.

With KNN regression, a median or average of most similar *K* cases is used to forecast. In the majority of comparable cases where KNN is used for classification, KNN may be computed as a class at the highest frequency. It is expected that each case will vote for its party and the group that has the largest number of votes.

The classifier determines the *K* points in the data nearest to  $x_o$ , given that the positive integer *K* and a test observation are carried out. Thus if *K* is 5, five observations nearest to observation  $x_o$  are identified. Usually these points are shown by  $N_o$ . The KNN classifier then calculates the conditional likelihood for class *j* as a fraction of the observation points in  $N_o$ . This is expressed mathematically by:

$$Pr(Y = j|X = x_o) = \frac{1}{K} \sum_{i \in N_o} I(y_i = j). \tag{5}$$

It is no surprise that altering *K* results vary drastically. If *K*=1 is minimally restricted, KNN models can generate low bias, but high variance. *K*=1 is not limited. As the number *K* increases, the classifier's flexibility

is reduced and the decision limit is increasingly linear. These models are low variance, but highly predictive. Both of these models don't work particularly well on a test accuracy basis, so we need to find a model with a well-balanced varying and bias.

*K-Means* is an unsupervised learning technique used when information is given unlabeled. The objective of this algorithm is to find groups with the number of  $K$  groups in the details. According to the specification, the algorithm assigns one  $K$ -group to each data point (Curran and Orr 2011). The  $K$ -cluster centroids can be used as low-marked data and labels for training information. This algorithm displays a number of functional vectors as an aggregated dataset. The quantity of seed by  $k$  selects randomly as a cluster base. Assign the cluster's closest data point.

$K$ -Means is among the most popular clustering algorithms with its enhanced version  $K$ -Medoids. Their drawbacks, however are that before the algorithm is used the number of  $K$  clusters must be known, algorithms sensitive to surface, noise, and the initial location of the centroids significantly affects the algorithms outcome. It should be noted that clusters cannot be used to predict credit rating (or any other prediction). Instead they are generally used as a pre-processing step for data mining in accordance with a supervised classification algorithm.

#### NEURAL NETWORKS

*Artificial Neural Network.* An artificial neural network is a group of neural networks linked to one weighted node (Dhaiya and Singh 2016). Each node can replicate the neuron of a creature, and the synaptic interactions between these nodes are the same. The neural network consists of three layers, the input, hidden layer and output layer, defined as a multi-layer perception (Olafsson, Li, and Wu 2008). In the MLP, the layer network, connected as an input unit layer to the hidden unit layer, is connected to the output unit layer.

*Multilayer Perceptron.* MLP consists of the input, output and one or more hidden layers between them (Chen and Huang 2003). It interconnects each layer entirely. The processing part is called nodes that function as a neuron, except for the input layer for each layer. Any node in one layer in the next layer is connected by weights with another node. Non-linear activation is available in many neuron layers. The network will learn from these layers the relation between input and output vectors. The input nodes are entered and the result nodes are released. This is referred as



front propagation only. Interestingly enough, back-propagation is a training algorithm where the values are forwarded, the error is measured and replicated back to the previous stages. In other words, forward spreading is part of the back spreading algorithm, but is performed before the node signals are propagated again.

Extreme Learning Machine (ELM): ELM is designed by Huang for widespread feedback networks with one hidden layer (Chen and Huang 2003). ELM randomly selects the hidden node parameter, which can describe a network as a linear system (Huang and Zhu 2006). ELM tends to meet minimal exercising defects, resulting in effective broader use and minimum weight requirements. ELM is easy to learn and provides excellent generalization efficiency in many real and artificial applications (Zhou et al. 2012). ELM is a modern and efficient single-layer feedback network training algorithm. Given  $N$  distinct training samples  $(x_i, t_i) \in R_n * R_m, (i = 1, 2, \dots, N)$ , the output of an SLFN with  $N$  hidden nodes can be represented by:

$$O_j = \sum_{i=1}^N \beta f_i(x_i) = \sum_{i=1}^N \beta f(x_j; a_i, b_i), j = 1, \dots, N \quad (6)$$

where,  $O_j$  is the output vector of SLFN with respect to input sample  $x_i$ .  $a_i = [a_{i1}, a_{i2}, \dots, a_{in}]^T$  and  $b_i$  are the learning parameter generated randomly of the  $j$ th hidden node that is  $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T$  is the link connecting the  $j$ th hidden node and the output node.  $f(x_j; a_i, b_i)$  is the activation function of the original ELM.

#### THE COMBINATION OF CLASSIFIERS

A group of independently qualified simple classifiers is part of the group of classifiers. The basic classifiers are chosen to create a classifier. The new cases are classified collectively by voting and the invisible ones. The vote may be weighted or not. In a mixed classification, the basic classifiers are combined to achieve greater output than the single classifier. Various aggregation methods used to improve the classification accuracy. Like regression, individual trees can be grouped and tuned as ensemble patterns. Multiple trees are grown simultaneously by using ensemble methods. Individual trees have a high variance but low bias. The variance of the ensemble system is typically significantly reduced when combining the expected values of the individual trees.

Bagging, boosting and random forests are the most aggregation methods (Pandey et al. 2013).

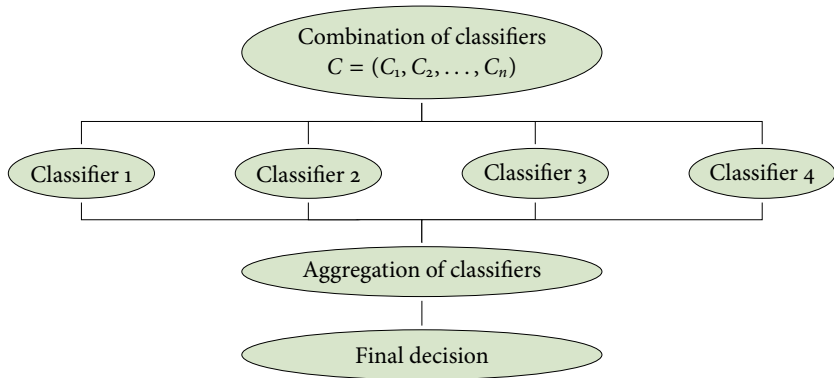


FIGURE 1 The Mathematical System of Combination of Classifiers

The mathematical system of combination of classifiers: a combination model of several classifiers is presented in figure 1.

The  $C = (C_1, C_2, \dots, C_n)$  are a sample of  $n$ -dimensional classifiers with different attributes. The classifiers can be different or same type of classifier with different attributes.

**Bagging.** The Bagging method was developed in 1996 by Breiman (Hui and Yang 2011). This is a meta-algorithm for machine learning designed to improve the reliability and accuracy of machine learning algorithms used for classification and regression statistics. It contributes to reducing overfitting. The bagging is an average type approach. The comparable classification is chosen primarily as a simple bagging classification. With bagging, by separate training sets of the same size and sampling of the training sets ‘with replacement’ can produce distinct decision-making structures (Jain and Kumar 2007).

**Boosting.** Boosting is an integration technique, trying to add new models which succeed when previous models are lacking (Pandey et al. 2013). Increasing the variance is not the goal. It is ideal for designs with high biases and low variance. Most of the stimulating algorithms are a weak classification and add to a high ranking (Huang, Chen, and Siew 2006). The boosting produces a classification ensemble through the re-evaluation and cost-function or majority vote of the training data kit.

**Random forests:** Random forests also use the construction of multiple classification trees based on bootstrapped training samples similar to bagging. However a random sequence of predictors is picked from the whole set of predictors for every split in a tree. Usually, the size of the

random subset is the square root of the number of features. The division can only be chosen from the randomly selected predictor sub-group. A new predictor sample is taken on each split, so that the majority of predictors cannot be taken into account on each split. This method decreases the association of the individual trees.

## **Utilization of Data Mining Algorithms for Credit Risk Analysis**

### **CLASSIFICATION ALGORITHMS**

The efficiency of the bank's credit portfolio is in line with the Bank's competitiveness and profitability. The efficiency of the bank's loan rises with the number of highly credible customers. The key decision support system used to test the credit ratings of the applicant is the credit rating. Therefore, the credit score can be defined as the modelling method of credit ratings for the applicant (Crook, Edelman, and Thomas 2007).

The credit scoring literature includes different meanings. The statistical methods are used to direct the choice of a loan that transforms the related variables of the loan into numeric action (Anderson 2007). Malhotra and Molhotra (2003) have identified them as an analytical model that has empirically established data from previous applications to predict applicants' reputation using the default probability.

The roots of the scoring systems can be traced to the 1930s when some postal companies started to use an evaluation system to eliminate differences between loan analyses. Because of the important responsibility that credit analysts had to provide military services during the Second World War, the management of the credit risk was a concern for businesses operating in the financial sector. Therefore, businesses needed analysts to write down the loans they used. These guidelines were used by non-experts to help companies enforce their loan decision processes and to control their expert systems (Thomas, Edelman, and Crook 2002).

The scholarly literature on credit scoring dates back to the 1940s when Durand published his work outcomes in order to detect credit factors which significantly affected a good and bad classification of loan applications based on the concept of loan scoring 'same characteristics could be used in separation of the groups within the same group' (Durand 1941).

Altman (1968) used a biased appraisal methodology by using information from companies operating in the United States – manufacturing sector. He has developed a general accuracy model in his research to forecast the company's failure.

Since banks supply a wide range of items, the number of credit applications has greatly increased. The huge demand growth in the banking sector has led to higher applications for credit scoring. Banks have begun reviewing almost all use of mortgage loans, credit cards, small business loans, and domestic loans (Emel, Oral, and Yolalan 2003).

Beaver (1966) submitted a survey to examine potential bankruptcy predictors. He showed statistically that many loan variables were primarily used to determine the risk of bankruptcy.

As the early credit rating models for approved applications took note of only the historical data, the consistency of the rejected applications could not be known. Feelders (2000) applied accepted and rejected applications to construct a mixture model and has shown that when rejected applications are included the performance of the model dramatically improves.

The banks used credit scores for residential mortgages and small business loans in the 1980s by doing a good review of credit card application methods (Thomas, Edelman, and Crook 2002). There have been so many applications that the conventional approach in which a loan specialist has reviewed the applications one by one has become economically difficult to implement. Both credit applicants' and banks' willingness in the fairest time to review applications encouraged banks to use credit scoring in evaluating credit applications (Lewis 1992).

Banks are concentrating more on conducting many observational experiments to benefit from the already efficiently applied classification methods. This helps them to distinguish between good and bad credit applications and build strategies that optimize the benefits of new and existing customers (Finlay 2010).

The credit score system uses a prior credit record to obtain a quantitative model to determine loan applications for separation of approval or refusal. Both credit rating applications in the credit rating system are continuously covered by the credit evaluation criteria. With the credit scoring system, credit decisions can be made promptly. Moreover, credit scoring maintains improved customer satisfaction due to the brief timeframe taken to complete the loan request. It can also provide a comprehensive approach for assessing success (Nisha 2017). One of the best business predictive applications is the credit score model (Bailey 2004).

There have been many quantitative methods in the literature for credit scoring. Although these approaches have problems modelling complex economic systems, they are based on fixed characteristics and statistical assumptions, several statistical methods exist for the evaluation of

credit risk. Studies comparing statistical methods and methods of machine learning for assessing credit risk have shown that machine learning techniques are better than statistical methods (Saberi et al. 2013).

The risk assessment of credit by nature is a problem of classification. Harris (2013) presents a bilateral classification technique for quantitative assessment of credit risk using Support Vector Machines (SVM). The study applied the SVM technique using both large and narrow data when widths are less than 90 days and narrows over 90 days. The comparison of the models' performance showed that the models developed using narrow data outperform with broad data. Danenas and Garsva (2012) presented another research using SVM in order to assess SVM credit risk in association with the selection of development parameter.

#### GROUPING ALGORITHMS

A decision on credit allocation requires a precise decision support system since even a slight improvement in accuracy is a huge financial saving for financial firms. Credit scoring is the most widely used method that enables lenders to make decisions about credit allocation. The goal of clusters is to group patterns based on a criterion of similarity (or dissimilarity) in groups (or clusters). Because it can be easily and quickly used, *K*-means is a famous clustering method (Saritha et al. 2010; Gholamiana, Jahanpourb, and Sadatrasoul 2013).

#### NEURAL NETWORKS

Useful research, which showed superior efficiency compared to statistic loan assessment models conducted by Sousa, Gama, and Brandao (2016) suggested a fresh dynamic structure for credit risk assessment. The technique comprises of sequential knowledge through the implementation of new information that enables the projections to be adjusted with changing information volumes.

Bekhet and Eletter (2014) have introduced artificial neural network models for Jordanian corporate banks' lending decisions. The radial function model was used in the neural network model, rather than the logistic regression. Although the neural network model is less reliable in general, the authors have found that the default alternative is more efficient when recognizing clients.

In Akkoç's work (2012) which utilizes statistical techniques and neuro-fuzzy, the adaptive neuro-fuzzy inference system (ANFIS) is presented as a model for credit scoring. The efficient use of international bank credit

card information, linear discrimination analysis, logistical regression and the artificial neural network have been evaluated against this suggested model. The overall right classification rate and the expected error cost were lower with the ANFIS technique.

#### THE COMBINATION OF CLASSIFIERS

The Bavarian Method and two-stage regression trees process (Kao, Chiu, and Chiu 2012) have developed a template that involves two techniques. The first step is to build a hierarchical Bayes model for customers that represent reimbursement choices and credit usage behavior. These predicted customer performance results are used as an inscription into the CART algorithm in the second level, as are client credit reports and demographic information.

The results of the CART are used to draw up policy guidelines in order to decide whether applicants are to grant loans, define loan limits, the annual performance rate and other levels of bank items. Based on these findings, the authors conclude that a cardholder's credit report is the most explanative of credit scoring and demographic variables are less relevant as they are less successful in credit score.

Regression is one of the well-known and robust approaches used in the literature for credit scoring and evaluation. Van-Gool et al. (2012) applied a logistic regression in the evaluation of applicants' features and the characteristics of loans and of the branches and credit officers. The applicant assessed the socio-economic attributes, the loan characteristics and experiences of credit officers in another logistics regression inquiry (Kinda and Achonu 2012). Dinh and Kleimeier (2007) have used logistic regression by analyzing credit characteristics, applicant characteristics and connections between the applicants and the bank.

Recent studies have shown that traditional techniques of statistical analysis and artificial intelligence (AI) typically apply to the choice of features that can enhance credit risk identification accuracy.

With regard to conventional statistics, current research investigates factors influencing credit risk to clients, primarily through statistical methodologies such as multiple bias analysis, multiple logic regression and the Markov chain. Multiple Discriminant Analysis was suggested for study by Pinches and Mingo (1973), and McAdams (1980) to investigate the contributing factors in the credit score. Pishbahar, Ghahremanzadeh, and Ainollahi (2015) have examined 779 farmers' details and identified key reimbursement effects by using the Nested Logit Model

(NLM). Karan, Ulucan, and Kaya (2013) analyzed loan appraisal metrics by constructing a rational regression model. Afolabi (2010) assessed some socioeconomic characteristics of 286 smallholder farmers in Nigeria through a quantitative assessment methodology. In order to investigate credit-risk factors, Karminsky and Khromova (2016) used Ordering Probit Regression, ANOVA and Survival-Duration Models. Bai, Shi, and Liu (2018) tested their credit qualifications with Fuzzy rough description technology and the  $F$  test method). For the study of customer credit scores, Zhang and Chi (2018) applied a genetic algorithm. For credit rating forecasts and yield substantially more dependable predictions, Petropoulos, Chatzis, and Xanthopoulos (2016) proposed a hidden Markov model. Hwang (2013) compared conventional statistical techniques and the result showed that the most effective models are ordered logit regression and ordered probit models. Shi, Yang, and Wang (2016) proposed a new approach by means of fluffy cluster analysis to distinguish the customer's standard. Shi, Wang, and Qi (2015) combined logistic regression and correlation analysis to extract features. A selection function also included  $R$  cluster analysis and coefficient of variation (Shi, Chen, and Wang 2016).

In the recent creation of an effective credit scored model for Artificial Intelligence (AI) technology, researchers have proposed an approach in hybrid information mining. A neural network, a support vector machine (SVM), genetic algorithm, and other techniques are investigated to test the credit scoring mechanism, Akkoc (2012) suggested a 3-stage hybrid credit scoring model of the Adaptive Neuro-Fuzzy Inference Method. Results have shown that the full value of assets, total liabilities and operating profit margins are necessary in order to ensure the credit risk of American examples (Huang, Chen, and Hsu 2004). In the credit risk assessment, Kim and Ahn (2012) investigated SVM and demonstrated the further enhancement of SVM general efficiency by selecting features. The genetic algorithm by Hájek (2011) used to choose input parameters. Hájek (2011) used genetic algorithms to select input variables. Hájek and Michalak (2013) have demonstrated that wrappers have improved the accuracy of US and European datasets better than filters.

The majority of risk characteristics depend on financial indices or private customer data that ignore macroeconomic variables. Second, most rating systems focus only on consistency in classification, but are not in a position to recognize key variables influencing customers' willingness to make repayments.

## Conclusion

Risk is an aspect of banking activities which cannot be removed entirely, but which can be minimized by use of appropriate methods. One of the key priorities of the banking system is to maintain a sustainable and sound credit system from the credit application to the loan termination. Credit risk is closely related to the assessment and management of the excellence of this process. This is the most significant type of risk for banks.

No reimbursement or reputation measurements should be performed with care, speed, accuracy and realism in the highly dynamic banking sector today. This improves credit effectiveness and customer response rates and allows us to review applications quicker and for more customers. The banks' failure to evaluate credit requests means that the use of funds is inefficient. Where a bank finances a lender believing that the loan is not risky, but the issue is that there is no reimbursement or a bank does not lend the credit to an applicant that has no reimbursement problem causes the bank to lose heavily.

The main aim of the conventional credit assessment methodology is to give credit only to those customers who pay back the loan. The credit underwriting experts review the applications. The loan of the borrower is opened if the application is approved. This method, however, creates consumer discontent, as the subjective judgment of loan underwriting experts is introduced because of incoherence in the credit decision-making system and the absence of evaluation of each applicant with equally objective variables.

Credit applications are analyzed using the decision support system as a scorecard with different classifiers in the current loan rating system approach. The strategy seeks to evaluate credit applications that develop quicker, easier, more effectively and more accurately. The pace of the method is related to the objective nature of the loan evaluation and the time taken to make decisions. Predictive intensity and quality depend on the efficacy and accuracy of the template used in the credit assessment process.

Various classifier forms are discussed in this paper and various classification types are discussed. This paper is focused on main classification algorithms.

## References

- Afolabi, J. A. 2010. 'Analysis of Loan Repayment among Small Scale Farmers in Oyo State, Nigeria.' *Journal of Social Sciences* 22 (2): 115–9.



- Akkoc, S. 2012. 'An Empirical Comparison of Conventional Techniques, Neural Networks and the Three Stage Hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) Model for Credit Scoring Analysis: The Case of Turkish Credit Card Data.' *European Journal of Operational Research* 222 (1): 168–78.
- Altman, E. I. 1968. 'Financial Ratios, Discriminant Analysis and the Prediction of Corporation Bankruptcy.' *The Journal of Finance* 23 (4): 589–609.
- Anderson, R. 2007. *The Credit Scoring Toolkit*. New York: Oxford University Press.
- Bai, C., B. Shi, and F. Liu. 2018. 'Banking Credit Worthiness: Evaluating the Complex Relationships.' *Omega* 83:26–38.
- Bailey, M. 2004. *Consumer Credit Quality: Underwriting, Scoring, Fraud Prevention and Collections*. Bristol: White Box Publishing.
- Bask, A., H. M. Ratanen, M. Tinnila, and T. Lauraeus. 2011. 'Towards E-Banking: The Evolution of Business Models in Financial Services.' *International Journal of Electronic Finance* 5 (4): 333–56.
- Bayes, T. 1763. 'An Essay towards Solving a Problem in the Doctrine of Chances.' *Philosophical Transactions of the Royal Society of London* 53:370–418.
- Beaver, W. H. 1966. 'Financial Ratios as Predictors of Failure.' *Journal of Accounting Research* 4:71–111.
- Bekhet, H. A., and S. F. K. Eletter. 2014. 'Credit Risk Assessment Model for Jordanian Commercial Banks: Neural Scoring Approach.' *Review of Development Finance* 4 (1): 20–8.
- Brown, M., and C. Zehender. 2007. 'Credit Reporting, Relationship Banking, and Loan Repayment.' *Journal of Money, Credit and Banking* 39 (8): 1883–918.
- Chen, M. C., and S. H. Huang. 2003. 'Credit scoring and rejected instances reassigning through evolutionary computation techniques.' *Expert Systems with Applications* 24:433–441.
- Chen, W, G. Xiang, Y. Liu, and K. Wang. 2012. 'Credit Risk Evaluation by Hybrid Data Mining Technique.' *Systems Engineering Procedia* 3:194–200.
- Chopde, K., P. Gosar, P. Kapadia, N. Maheshwari, and P. M. Chawan. 2012. 'A Study of Classification Based Credit Risk Analysis Algorithm.' *International Journal of Engineering and Advanced Technology* 1 (4): 142–4.
- Chorowski, J., J. Wang, and M. J. Zurada. 2014. 'Review and Comparison of SVM and ELM Based Classifiers.' *Neurocomputing* 128:506–16.
- Crook, J. N., D. B. Edelman, and L. C. Thomas. 2007. 'Recent Developments in Consumer Credit Risk Assessment.' *European Journal of Operational Research* 183 (3): 1447–65.

- Curran, K., and J. Orr. 2011. 'Integrating Geolocation into Electronic Finance Applications for Additional Security.' *International Journal of Electronic Finance* 5 (3): 272–85.
- Danenas, P., and G. Garsva. 2012. 'Credit Risk Evaluation Modeling Using Evolutionary Linear SVM Classifiers and Sliding Window Approach.' *Procedia Computer Science* 9:1324–33.
- . 2015. 'Selection of Support Vector Machine Based Classifier for Credit Risk.' *Expert Systems with Application* 42:3194–204.
- Dhaiya, S., and N. P., Singh. 2016. 'Impact of Bagging on MLP classifier.' *International Conferences on Computing For Sustainable Global Development* 7 (4): 3794–9.
- Dinh, T., and S. Kleimeier. 2007. 'Credit Scoring Model for Vietnam's Retail Banking Market.' *International Review of Financial Analysis* 5 (16): 471–95.
- Durand, D. 1941. *Risk Elements in Consumer Installment Financing*. New York: National Bureau of Economic Research.
- Emel, A., M., R. Oral, and R. Yolalan. 2003. 'A Credit Scoring Approach for the Commercial Banking Sector.' *Socio-Economic Planning Sciences* 37 (2): 103–23.
- Feelders, A. J. 2000. 'Credit Scoring and Reject Inference with Mixture Models.' *Intelligent Systems* 9 (1): 1–8.
- Finlay, S. 2009. 'Credit Scoring for Profitability Objectives.' *European Journal of Operational Research* 202 (2): 528–37.
- Gholamiana, M. R., S. Jahanpourb, and S. M. Sadatrasoul. 2013. 'A New Method for Clustering in Credit Scoring Problems.' *Journal of Mathematics and Computer Science* 6:97–106.
- Hajek, P. 2011. 'Municipal Credit Rating Modeling by Neural Networks.' *Decision Support Systems* 51 (1): 108–18.
- Hajek, P., and K. Michalak. 2013. 'Feature Selection in Corporate Credit Rating Prediction.' *Knowledge-Based Systems* 51:72–84.
- Hamid, A., and T. M. Ahmed. 2016. 'Developing Prediction Model of Loan Risk in Banks Using Data Mining.' *Machine Learning and Applications: An International Journal* 3 (1) <https://www.doi.org/10.5121/mlaij.2016.3101>.
- Harris, T. 2013. 'Quantitative Credit Risk Assessment Using Support Vector Machines: Broad versus Narrow Default Definitions.' *Expert Systems with Applications* 40 (11): 4404–13.
- Hearst, M. A., S. T. Dumais, E. Osman, and J. Platt. 2008. 'Support Vector Machines.' *IEEE Intelligent Systems* 13 (4): 18–28.
- Huang, J., H. Chen, and C. J. Hsu. 2004. 'Credit Rating Analysis with SVM and Neural Network: A Market Comparative Study.' *Decision Support System* 37 (4): 543–58.

- Huang, G. B., L. Chen, and C. K. Siew. 2006. 'Universal Approximation Using Incremental Networks with Random Hidden Computation Nodes.' *IEEE Transaction on Neural Networks* 17(4): 879–92.
- Huang, B. G., and Q. Y. Zhu. 2006. 'Extreme Learning Machine: Theory and Application.' *Neurocomputing* 70:489–501.
- Hui, X., and S. G. Yang. 2011. 'Using Clustering-Based Bagging Ensemble for Credit Scoring.' Paper presented at the IEEE 2011 International Conference on Business Management and Electronic Information (BMEI), Guangzhou, 13–15 May.
- Hwang, R. C. 2013. 'Forecasting Credit Ratings with the Varying-Coefficient Model.' *Quantitative Finance* 13 (2): 1–19.
- Islam, J. M., J. Wu, and M. Ahmadi. 2007. 'Investigating the Performance of Naïve-Bayes and K- Nearest Neighbor Classifiers.' In *2007 International Conference on Convergence Information Technology*, 1541–6. <https://www.doi.org/10.1109/ICCIT.2007.148>.
- Jain, A., and A. M. Kumar. 2007. 'Hybrid Neural Network Models for Hydrologic Time Series Forecasting.' *Applied Soft Computing* 7 (2): 585–92.
- Kao, L. J., C. C. Chiu, and F. Y. Chiu. 2012. 'Bayesian Latent Variable Model with Classification and Regression Tree Approach for Behavior and Credit Scoring.' *Knowledge-Based Systems* 36:245–52.
- Karan, M. B., A. Ulucan, and M. Kaya. 2013. 'Credit Risk Estimation Using Payment History Data: A Comparative Study of Turkish Retail Stores.' *Central European Journal of Operations Research* 21 (2): 479–94.
- Karminsky, A. M., and E. Khromova. 2016. 'Extended Modeling of Banks' Credit Ratings.' *Procedia Computer Science* 91:201–10.
- Kim, K., and H. Ahn. 2012. 'A Corporate Credit Rating Model Using Multi-Class Support Vector Machines with an Ordinal Pairwise Partitioning Approach.' *Computers & Operations Research* 39 (8): 1800–11.
- Kinda, O., and A. Achonu. 2012. 'Building a Credit Scoring Model for the Savings and Credit Mutual of the Potou Zone (MECZOP)/Senegal.' *Journal of Sustainable Development* 7 (1): 17–33.
- Laplace, P. S. 1812. *Théorie analytique des probabilités*. Paris: Courcier.
- Lewis, E. M. 1992. *An Introduction to Credit Scoring*. San Rafael, CA: Athena.
- Louzada, F., A. Ara, G. B. Fernandes. 2016. 'Classification Methods Applied to Credit Scoring: A Systematic Review and Overall Comparison.' *Surveys in Operations Research and Management Science* 21:117–34.
- Malhotra, R., and D. K. Malhotra. 2003. 'Evaluating Consumer Loans Using Neural Networks.' *Omega* 31 (2): 83–9.
- McAdams, L. 1980. 'How to Anticipate Utility Bond Rating Changes.' *The Journal of Portfolio Management* 7 (1): 56–60.
- Nisha, N. 2017. 'An Empirical Study of the Balanced Scorecard Model: Ev-

- idence from Bangladesh.' *International Journal of Information Systems in the Service Sector* 9 (1): 68–84.
- Olafsson, S., X. Li, and S. Wu. 2008. 'Operations Research and Data Mining.' *European Journal of Operational Research* 187 (3): 1429–48.
- Pandey, T. N., A. K. Jagadev, D. Choudhury, and S. Dehuri. 2013. 'Machine Learning-Based Classifiers Ensemble for Credit Risk Assessment.' *International Journal of Electronic Finance* 7 (3–4): 227–49.
- Petropoulos, A., S. P. Chatzis, and S. Xanthopoulos. 2016. 'A Novel Corporate Credit Rating System Based on Student's-t Hidden Markov Models.' *Expert Systems with Applications* 53 (1): 87–105.
- Pinches, G. E., and K. A. Mingo. 1973. 'A multivariate Analysis of Industrial Bond Ratings.' *The Journal of Finance* 28 (1): 1–18.
- Pishbahar, E., M. Ghahremanzadeh, and M. Ainollahi. 2015. 'Factors Influencing Agricultural Credits Repayment Performance among Farmers in East Azerbaijan Province of Iran.' *Journal of Agricultural Science and Technology* 17:1095–101.
- Saberi, M., M. S. Mirtalaie, F. K. Hussain, A. Azadeh, O. K. Hussain, and B. Ashjari. 2013. 'A Granular Computing-Based Approach to Credit Scoring Modeling.' *Neurocomputing* 122:100–15.
- Saritha, S. J., P. Govindarajulu, K. R. Prasad, S. Rao and C. Lakshmi. 2010. 'Clustering Methods for Credit Card using Bayesian Rules Based on K-Means Classification.' *International Journal of Advanced Computer Science and Applications* 1 (4): 92–5.
- Shi, B., N. Chen, and J. Wang. 2016. 'A Credit Rating Model of Microfinance Based on Fuzzy Cluster Analysis and Fuzzy Pattern Recognition.' *Journal of Intelligent & Fuzzy Systems* 31 (6): 3095–102.
- Shi, B., J. Wang, and J. Qi. 2015. 'A novel imbalanced data classification approach based on logistic regression and Fisher discriminant.' *Mathematical Problems in Engineering* 945359. <https://doi.org/10.1155/2015/945359>.
- Shi, B., H. Yang, and J. Wang. 2016. 'City Green Economy Evaluation: Empirical Evidence from 15 Sub-Provincial Cities in China.' *Sustainability* 8 (6): 1–39.
- Sousa, M. R., J. Gama, and E. Brandao. 2016. 'A New Dynamic Modeling Framework for Credit Risk Assessment.' *Expert Systems with Applications* 45:341–51.
- Sudhakar, M., and K. Reddy. 2014. 'Credit Evaluation Model of Loan Proposals for Banks Using Data Mining Techniques.' *International Journal of Latest Research in Science and Technology* 2 (4): 126–34.
- Thomas, L. C., D. Edelman, and J. N. Crook. 2002. *Credit Scoring and Its Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

- Van-Gool, J., W. Verbeke, P. Sercu, and B. Baesens. 2012. 'Credit Scoring for Microfinance: Is It Worth It?' *International Journal of Finance & Economics* 17 (2): 103–20.
- Wang, X., D. S. Yeung, E. C. C. Tsang. 2001. 'A Comparative Study on Heuristic Algorithms for Generating Fuzzy Decision Trees.' *IEEE Transactions on Systems, Man, and Cybernetics – part B: Cybernetics* 31 (2): 215–26.
- Wang, Y., S. Wang, and K. K. Lai. 2005. 'New Fuzzy SVM to Evaluate Credit Risk.' *IEEE Transaction on Fuzzy Systems* 13 (6): 820–31.
- Zhang, Y., and G. Chi. 2018. 'A Credit Rating Model Based on a Customer Number Bell-Shaped Distribution.' *Management Decision* 56 (5): 987–1007.
- Zhou, H., Y. Lan, Y. C. Soh, and G. B. Huang. 2012. 'Credit Risk Evaluation Using Extreme Learning Machine.' Paper presented at the *IEEE International Conferences on System, Man and Cybernetics*, Seoul, 14–17 October.
- Zurada J. 2010. 'Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions?' Paper Presented at the 43rd Hawaii International Conference on System Sciences, Kauai, HI, 5–8 January.