Credit Evaluation Model of Loan Proposals for Indian Banks

Seema U. Purohit, Venkatesh Mahadevan, and Anjali N. Kulkarni

***Abstract*—The failure and success of the Banking Industry depends largely on industry's ability to properly evaluate credit risk. Credit Evaluation of any potential credit application has remained a challenge for Banks all over the world till today. This paper checks the applicability of one of the new integrated model on a sample data taken from Indian Banks. The integrated model is a combination model based on the techniques of Logistic Regression, Multilayer Perceptron Model, Radial Basis Neural Network, Support Vector Machine and Decision tree (C4.5) and compares the effectiveness of these techniques for credit approval process.**

***Index Terms*—Credit evaluation, decision process, models.**

I. Introduction

A. Credit Evaluation

When it is required to obtain credit scoring, one has to undergo a process of evaluation before the credit score is sanctioned. This process is called as credit evaluation, which may take time, but concludes in either an approval or a rejection.

Credit approval for financial sanctions is a vital component of any evaluation process as it is related to the economy of a country. Before a potential debtor wants to obtain credit, he must be evaluated on certain areas. There are five C's involved in credit evaluation. As discussed in [1] they are: character, credit report, capacity, cash flow, and collateral.

The character of a person applying for a credit is a big factor to the decision for credit approval. A person with a sound financial objective is likely to be granted a credit approval quickly and are possibly than an individual who is in bad shape, not just on the financial facet, but also on other aspect

Credit history is another important factor considered by lenders in their decision to grant and approve credit applications. The credit report is a record of an individual's past borrowing and reimbursing transactions. It also includes information about late payments and bankruptcy.

A credit report can be tarnished. A credit score can be at its low. Under these circumstances it is unlikely for you to earn the confidence of the lender for a credit approval. However, if your cash flow is good, there is a possibility of getting the credit approval.

Lenders may also have to check the liquidity of an individual. This can be done by checking the bank statements of an individual borrower. In the case of businesses, lenders may have to obtain a copy of the audited financial statements. The financial statements of businesses and bank statements can be utilized to show the capacity of a borrower to settle and repay a line of credit. The capacity of the borrower to pay a credit is determined during credit evaluation and approval. Collateral is a common term in credit. A lender seeks for security whenever the borrower defaults the credit payment. If no collateral is present as security for a credit, it is likely that the lender will give the borrower a high-interest rate credit. Credit evaluation is a process taken by the lender with the participation of the credit applicant. If you want to undergo this process, it is important to make substantial preparation so you are more likely to obtain a credit quickly and less expensively.

B. Decision Process for Credit Evaluation

Credit managers rely heavily upon external data sources to guide them in the credit decision process. To approve or reject a credit request is a delicate task. A credit manager must evaluate the risk associated with extending credit and declining an applicant based on numerous factors [2].

The need for sufficient and reliable information is the foundation of a successful credit decision. A credit manager may call on references, run background checks, pull a credit report, verify bank accounts or ask questions of the applicant to validate the information on the credit application. Credit managers are challenged with the task of obtaining readily available information to support their decision while sending a timely response to the applicant. A major obstacle in achieving this task is the turnaround time associated with checking references. The process varies from business to business and may include a background check, a verification of a bank deposit or credit references with existing suppliers.

C. Motivation for Designing Different Credit Functions Using Different Models

Some businesses require written requests, while othersmay offer to do a phone interview at their convenience. The credit function is the heart of banking, under the ever changing market conditions. The lack of general credit review system in many banks and the lack of precise methods for measuring credit risk are two important reasons why an expert support system is necessary [2]. Such a system can be implemented by using advantages of Radial Basis Neural Network techniques, Multilayer perceptron Model, Regression techniques, SVM and Decision trees. Decision trees and Logistic Regression are well-established traditional

Manuscript received May 21, 2012; revised June 25, 2012.

S. U. Purohit is with Department of Mathematics, Kirti College and Department of Technology and Mgmt, NMITD, Mumbai, India
(e–mail: supurohit@gmail.com)

V. Mahadevan is with Information Systems & eBusiness, Swinburne University of Technology, Melbourne, Australia.
(e –mail: vmahadevan@swin.edu.au)

A. N. Kulkarni is with College, Panvel Navi Mumbai
(e-mail: anjali_kulkarni74@gmail.com)

statistical techniques, whereas Radial Basis Neural Networks are relatively new data mining tools that have been successfully used for classification and prediction. Multilayer perceptrons using a backpropagation algorithm are the standard algorithm for any supervised-learning pattern recognition process.

Decision trees are particularly useful for classification tasks. Like Radial Basis Neural Networks, decision trees learn from data. Using search heuristics, decision trees are able to find explicit and understandable rules-like relationships among independent and dependent variables. The purpose of the logistic regression model is to obtain a regression equation that could predict in which of two or more groups an object could be placed (i.e. whether a credit should be classified as approved or rejected).

SVM is a class of data driven and non linear methods that do not require specific assumptions on the underlying data. This feature is suitable for practical business problem where there are massive data.

The feature of different model that can be used for designing credit function can be given as follows:

A. Decision tree model

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute [3].

advantages of using decision learning tree algorithms are:

1) They generalize in a better way for unobserved instances, once examined the attribute value pair in the training data.
2) They are efficient in computation as it is proportional to the number of training instances observed.
3) The tree interpretation gives a good understanding of how to classify instances based on attributes arranged on the basis of information they provide and makes the classification process self-evident.

The operation of decision tree is based on ID3 or C4.5 algorithms. It builds tree based on the information (information gain) obtained from the training instances and then uses the same to classify the test data. ID3 algorithm generally uses nominal attributes for classification with no missing values. ID3 can even work well on datasets with missing attribute values to certain extent.

C4.5 handles both continuous and discrete attributes. While handling the data C4.5 allows missing attribute values to be marked as (?). Missing attribute values are simply not used in gain and entropy calculations. Decision tree are self explanatory and can be easily converted to set of rules so they are used in credit evaluation process.

B. Radial basis neural network model

Artificial neural networks are one of the most common data mining tools. Neural networks are particularly useful for the tasks of classification, prediction, and clustering in business applications. Neural network models are characterized by three properties: the computational property, the architecture of the network, and the learning property [4].

Computational Property: - Neural networks are made up of neurons or nodes, which are simple processing elements. Each neuron contains a summation node and often a nonlinear sigmoidal activation function of the form

$$F(n) = 1/(1+\exp(-2n)) \tag{1}$$

where $n = WP$ is the output from a summation node; λ is the steepness of the activation function; W is a weight matrix and P is an input vector. Because a single neuron has a limited capability, neurons (sometimes hundreds) are organized in layers and are interconnected between layers using connections called weights. Each weight carries a numerical value that represents the strength of connection or expresses the relative importance of each input to the neuron.

Architecture: - Radial basis neural network is most commonly used architectures used in financial applications is radial basis neural network. Radial Basis Functions are powerful techniques for interpolation in multidimensional space. They can model non linear function using a single hidden layer which removes some design decisions about number of layers. The simple linear transformation in the output layer can be optimized fully using traditional linear modeling techniques which are fast and do not suffer from the problems such as local minima. RBF networks can therefore be trained extremely quickly.

Learning: - Neural networks use a three types of learning modes supervised learning, unsupervised learning, and reinforcement learning. During supervised learning, which is the most common for the mentioned feed-forward networks, weights are initialized at small random values and training patterns are presented to the network one pattern at a time. The output produced by the training pattern is compared with the actual response provided by a teacher. The differences modify the weights of the network to make them closer to the actual output.

This process is repeated for all training patterns contained in a training set until the cumulative error between the actual outputs and the network's output is reduced to a small value. Weights are crucial to the operation of the neural network because through their repeated adjustment the neuron (or network) learns. Knowledge of the network is encoded in its weights. . The most attractive features of these networks are their ability to adapt, generalize, and learn from training patterns. Due to this feature it is used in credit evaluation process.

C. Logit Regression Model

The purpose of the logistic regression model is to obtain a regression equation that could predict in which of two or more groups an object could be placed (i.e. whether a credit should be classified as a good credit or a bad credit) [4]. The logistic regression also attempts to predict the probability that a binary or ordinal target will acquire the event of interest (e.g. credit payoff or credit default) as a function of one or more independent variables (i.e. amount of credit, borrower job category, reason of credit). The logit model is represented by the logistic response function *P(y)* of the form:

$$P(y) = 1/(1+\exp(-z)) \tag{2}$$

where $z = b_0 + \sum b_i x_i, \forall i = 1 to m.$

The function *P(y)* describes a dependent variable *y* containing two or more qualitative outcomes. *Z* is the function of *m* independent variables *x* called predictors, and *b* represents the parameters. The *x* variables can be categorical or continuous variables of any distribution. The value of *P*(*y*) that varies from 0 to 1 denotes the probability that a dependent variable *y* belongs to one of two or more groups. The principal of maximum likelihood can commonly be used to compute estimates of the *b* parameters [4].

This means that the calculations involve an iterative process of improving approximations for the estimates until no further changes can be made. Unlike radial basis neural networks, logistic regression models are designed to predict one dependent variable at a time. Logistic regression output provides statistics on each variable included in the model. Researchers then can analyze the applications with these statistics to test the usefulness of specific information. This model is easy to use and has good flexibility so it is used in credit scoring application.

D. Multilayer Perceptron Neural Network Model

A Multilayer Perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network MLP is suitable for credit evaluation process because of its classification accuracy [5].

If a multilayer perceptron has a linear activation function in all neurons, that is, a simple on-off mechanism to determine whether or not a neuron fires, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model Each neuron uses a *nonlinear* activation function which was developed to model the frequency of action potentials, or firing, of biological neurons in the brain. This function is modeled in several ways, but must always be normalizable and differentiable. The two main activation functions used in current applications are both sigmoid, and are described by

$$\begin{aligned} \phi(y_i) &= \tanh(v_i) \\ \phi(y_i) &= (1+\exp(-v_i))^{-1} \end{aligned} \tag{3}$$

in which the former function is a hyperbolic tangent, which ranges from -1 to 1, and the latter is equivalent in shape but ranges from 0 to 1. Here y_i is the output of the *i*th node (neuron) and v_i is the weighted sum of the input synapses. More specialized activation functions include radial basis functions which are used in another class of supervised neural network models.

The multilayer perceptron consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes. Each node in one layer connects with a certain weight w_{ij} to every node in the following layer.

Multilayer Perceptron most commonly seen in speech recognition, image recognition, and machine translation software, but they have also seen applications in other fields such as cyber security.

E. Support Vector Machine Model

A support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis.

Support vector machine was first proposed by Vapnik (1998). Its main idea is to minimize upper bound of the generalization error and it maps the input vector into high dimensional feature space through some nonlinear mapping. In this space, the optimal separating hyper plane, which separates the two classes of data with maximal margins, is constructed by solving constrained quadratic optimization problem whose solution has an expansion in terms of a subset of training patterns that lie closest to the boundary. It is been discussed in [6] how SVM has a best method for classification in terms of credit approval process.

Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. SMO is widely used for training support vector machines and is implemented by the popular libsvm tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers.

II. Literature Review

Data mining techniques have also been successfully applied to credit-risk assessment problems. The initial research focused on determining the usefulness of data mining tools, such as Radial Basis Neural Networks and decision trees, and examining how these tools should be applied in a credit-risk assessment context. In one of the early papers, McLeod *et al*. (1993) discussed general features of neural networks and their suitability for the credit-granting process.

Glorfeld and Hardgrave (1996) presented a comprehensive and systematic approach to developing an optimal architecture of a neural network model for evaluating the creditworthiness of commercial loan applications. The neural network developed using their architecture was capable of correctly classifying 75% of loan applicants and was superior to neural networks developed using simple heuristics. Tessmer (1997) examined credits granted to small Belgian businesses using a decision tree-based learning approach. Tessmer focused on the impact of Type I credit errors (classifying good loans as bad loans), and Type II credit errors (classifying bad loans as good loans), on the accuracy, stability and conceptual validity of the learning process. Subsequent authors built on the existing research by comparing the performance of various data mining techniques in various credit risk assessment contexts.

Desai *et al*. (1996) analyzed the usefulness of neural networks and traditional techniques, such as discriminant analysis and logistic regression, in building credit scoring models for credit unions. Desai studied data samples containing 18 variables collected from three credit unions and showed that neural networks were particularly useful in detecting bad loans, whereas logistic regression outperformed neural networks in the overall (bad and good loans) classification accuracy.

Barney *et al*. (1999) compared the performance of neural networks and regression analyses in identifying the farmers who had defaulted on their Home Administration Loans and those farmers who paid off the loans as scheduled. Using an unbalanced data, Barney found that neural networks outperform logistic regression in correctly classifying farmers into those who made timely payments and those who did not.

Jagielska *et al*. (1999) investigated credit risk classification abilities of neural networks, fuzzy logic, genetic algorithms, rule induction software, and rough sets and concluded that the genetic/fuzzy approach compared more favorably with the neurofuzzy and rough set approaches.

A. Significance and Objective of Study

Credit scoring is a very important task for lenders to evaluate the credit applications they receive from consumers as well as for insurance companies, which use scoring systems today to evaluate new policyholders and the risks these prospective customers might present to the insurer. Credit scoring systems are used to model the potential risk of credit applications, which have the advantage of being able to handle a large volume of credit applications quickly with minimal labor, thus reducing operating costs, and they may be an effective substitute for the use of judgment among inexperienced credit officers, thus helping to control bad debt losses. This study explores the performance of credit scoring models using approaches: Logistic regression, Multilayer Perceptron Model, Radial basis neural network, SVM and decision trees (C4.5).

Data mining techniques have been applied to solve classification problems in [7] for a variety of applications such as credit scoring, bankruptcy prediction, insurance underwriting, and management fraud detection. The lack of research in combining data mining techniques with domain knowledge has prompted researchers to identify the fusion of data mining and knowledge-based expert systems as an important future direction. Here by combining the advantages of data mining classification methods logistic regression, decision tree, Multilayer Perceptron Model, SVM and radial basis neural network a new integrated model will generate which will be best for a credit approval system.

Objective of the study is to find the Integrated model which can be constructed by combining the advantages of Radial Basis neural network, Multilayer Perceptron Model, Logistic regression, SVM and decision tree models (C4.5 model which works well with numeric attribute) to be applied for different types of credits for credit approval process such that there will be minimum defaulters and credit risks.

III. Problem Definition

The credit function is the heart of banking, under the ever changing market conditions. The lack of general credit review system in many banks and the lack of precise methods for measuring credit risk are two important reasons why an expert support system is necessary. It is with this spirit researchers have taken up the tasks for checking the applicability of the integrated model on the data collected from the Indian Banks.

IV. Methodology Adopted

A. Identification of Independent and Dependent Variables

The data set used in this research is divided into training and testing data sets. All training cases are set by default taking into account the banks' guidelines for personal credit approval in the banks. Data used is of 500 customer's data. The data required for the current study was collected from different banks such as SBI, IDBI, AXIS and Syndicate banks. It consists of different independent variables and one dependent variable.

Variables are the conditions or characteristics that the investigator manipulates, controls or observes. It is necessary to optimize variables by using SVM as mentioned in [8], [9], [10]. Variables are classified as dependent and independent variables. An independent variable is the condition or characteristic that affects one or more dependent variables: its size, number, length or whatever exists independently and is not affected by the other variable. A dependent variable changes as a result of changes to the independent variable.

Independent Variables

1) Age of customer
2) Sex
3) Marital status
4) Service period
5) Current account
6) Saving account
7) Payment history
8) Occupation
9) Home ownership
10) Address time

Dependent variable:

1) Credit (Approved or Not)

Using this data set, a model is built, which consists of a decision tree model (C4.5), radial basis neural network model, logistic regression model, SVM and multilayer perceptron model, to predict whether a future applicant's a credit is approved or rejected.

Using the neural network node, we can construct, train, and validate a network. The number of neurons in the hidden layer is determined experimentally from the number of observations in the data set and the number of weights in the network.

We can use the decision tree node to classify observations by segmenting the data created according to a series of simple rules. We can use the entropy gain reduction method to build the tree. The regression node fitted the logistic regression model to the data. The ensemble node combines the five models by averaging the posterior probabilities for the class target variable.

V. Credit Evaluation model

A. Requirements and Architecture

Credit evaluation system requires the following four components to work with data: computer system, WEKA software, customer, and data. Credit evaluation system can be built by combining the advantages of logistic regression, radial basis neural network, multilayer perceptron, SVM and decision tree.

B. Technology

Research methodology used is quantitative analysis. For logistic regression, radial basis neural network, multilayer perceptron, SMO and c4.5 models two data sets are used. One of them is used as training data set and other is used as test data set. The data is normalized first. A new model will be developed by combining the advantages of all these five models. Then this model will be applied to this normalized data.

C. Software Customization

All the experiments are carried out through WEKA. In these experiments 100,200,300,400,500 customer's data is used. If data is of 100 customers then training instances are 90 and testing instances are 10. If data is of 200 customers then training instances are 180 and testing instances are 20. If data is of 300 customers then training instances are 270 and testing instances are 30.If data is of 400 customers then training instances are 360 and testing instances are 40. If data is of 500 customers then training instances are 450 and testing instances are 50.

VI. Research Findings

Following tables present classification percentage of logistic regression, radial basis network, and multilayer perceptron model, SVM and decision tree according to data.

TABLE I. Results for Decision Tree (C4.5)

No. of customers	No. of customers in training set	No. of customers in testing set	Classification percentage
100	90	10	94.80
200	180	20	76.75
300	270	30	81.33
400	360	40	84.25
500	450	50	84.77

TABLE II. Results for Radial Basis Network

No. of customers	No. of customers in training set	No. of customers in testing set	Classification percentage
100	90	10	94.20
200	180	20	75.70
300	270	30	83.87
400	360	40	85.65
500	450	50	86.53

TABLE III. Results for Logistic Regression

No. of customers	No. of customers in training set	No. of customers in testing set	Classification percentage
100	90	10	88.70
200	180	20	74.45
300	270	30	77.40
400	360	40	81.58
500	450	50	82.79

TABLE IV. Results for Multilayer Perceptron Neural Network

No. of customers	No. of customers in training set	No. of customers in testing set	Classification percentage
100	90	10	86.90
200	180	20	74.70
300	270	30	78.47
400	360	40	80.47
500	450	50	80.85

TABLE V. Results for SVM (SMO)

No. of customers	No. of customers in training set	No. of customers in testing set	Classification percentage
100	90	10	88.80
200	180	20	76.50
300	270	30	83.00
400	360	40	85.43
500	450	50	85.97

VII. Conclusion

Research findings indicate that SVM, decision tree and logistic regression is the best methodology for classifying the loan applications. By analyzing the performance of these models on standard dataset it is found that in case of missing data multilayer perceptron model and logistic regression is also good. So the objective is to develop new integrated model which takes advantages of all the five models.

Acknowledgment

In this study many persons and Institution assisted us and without their help this study would not have been completed. A. N Kulkarni specially thanks Bank officials of many banks in Maharashtra, who helped by providing the data needed for the study and the management of NMITD for motivating at every stage of this study.

References

[1] M. Handzic, F. Tjandrawibawa, and J. Yeo, "How/neural networks can help loan officers to make better informed applications decisions," in *Proc. 2003 Informing Science + IT Education Conference* 2003, pp. 97-108.

[2] J. E. Boritz and D. B. Kennedy, "Effectiveness of neural network types for prediction of business failure," *Expert Systems with Applications*, vol. 9, no. 4, pp. 503-512, 1995.

[3] M. Bensic, N. Sarlija, and M. Zekic-Susac, "Modeling small-business credit scoring by using logistic regression, neural networks and decision trees," *international Journal of Intelligent Systems In Accounting, Finance And Management*, vol. 13, no. 3, pp.133-150, July2005.

[4] J. Zurada and M, Zurada., "How Secure Are "Good loans": validating loan-granting decisions and predicting default rates on consumer loans," *The review of business information systems*, vol. 6, no.3, pp. 65-83, 2002.

[5] H. G. Nguyen, "Using neural network in predicting corporate failure," *Journal of Social Sciences,* vol.1, no. 4, pp.199-202, 2005.

[6] C. L. Huang, M. C. Chen, and C. J. Wang, "Credit scoring with data mining approach based on support vector machines," *Expert systems with applications*, vol. 33, no.4, pp. 847-856, November 2007.

[7] D. J. Hand and W. E. Henley, "Statistical classification method in consumer scoring: a review," *Journal of the royal statistical society, series A (Statistics in Society)*, vol. 160, no. 3, pp.523-541, 1997.

[8] Y. W. Chen and C. J. Lin, "Combining SVMs with various feature selection strategies," in Feature extraction, foundations and applications, New York: Springer, 2005, pp.319-328.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157-1182, 2003.

[10] Y. Q. Wang, "Building credit scoring systems based on support-based support vector machine," in *Proc. Fourth international conference on natural computation October 2008*, pp. 323-327.

Dr. S. Purohit, born in Ahmadabad, India on 8th August 1962. She has received M.Sc. degree in Applied Mathematics and Ph. D. in Computational Mathematics in 1990 from University of Mumbai and working as a Professor in Kirti M. Doongursee College and Director, NMITD, Mumbai. Her specializations include Research methodology and research based teaching and learning techniques. Currently she is working on problems in Artificial Intelligence, Optimization, Simulation and Mathematical Modeling and published and presented her work in International Conferences and referred Journals including IEEE explorer. She is a Research Guide for doctoral students in the subject of Computer Science, Information Technology and Management. She is an educational consultant and is an active member contributing to curriculum designs. Dr. Purohit is a Senior Member of IACSIT, Singapore, MIR LABS, USA,IEEE, WIE, and CSI, India.

V. Mahadevan has received a B.E. in Electrical and Electronics Engineering in 1989 form Institution of Engineers, India, M.E. in Computer Systems Engineering in 1997 form University of Auckland, New Zealand Ph.D. in E – business Technology in 2007 form University of Sydney, Australia and has an extensive research experience in the different phases of the usability design lifecycle, from user research, requirement specifications and conceptual design through to usability attributes testing of a Tele-collaboration (TC) business system. He has joined the Faculty of Higher Education, Lily dale as Lecturer in Information Systems (IS) and e-Business. He has presented his research findings at several national and international conferences and workshops. He is Founder Chair of Forum for Telecommunications Networking Simulations (using OPNET, OMNeT++ etc) at UTS. Cross-institutional Coordinator for Telecommunications Networking Simulations for the Technical University of Budapest (Hungary) at UTS. He worked as a Director and senior member of IACSIT Sydney Chapter and Chennai (India) Chapter. He is an Editorial Board Member, Australian Journal of Information Systems (JIS) and ICTACT Journal of Communication Technology (ICTACTJCT). Dr. Mahadevan is a Member of IEEE (SIG - Society on Social Implications of Technology), Association of Information Systems (SIG - E-Business), International Association for Development of Information Society (IADIS), International Consortium for Electronic Business (ICEB), Human Capital Institute (HCI), Sydney Networks Interested Research Forum (Sydney), Institution of Engineers (India).

A. Kulkarni, born in May 1974 in India. She received her B.Sc. degree in Computer Science from Amravati University, India, in 1994. She received her M.Sc. degree in Computer Science from Amravati University, India, in 1996. She completed her M.Phil. (Master of Philosophy) in the subject of Computer Science from Yashwantrao Chavan Open University, India in the year 2008. Her research interest areas' include Artificial Intelligence and data-mining techniques. She is working as Assistant Professor in the CKT College, Panvel Navi Mumbai.