

IMPROVING THE CREDIT SCORING MODEL OF MICROFINANCE INSTITUTIONS BY SUPPORT VECTOR MACHINE

Asuri Venkata Madhavi¹, Radhamani .G²

¹Research Scholar, Bharatiyar University, Tamil Nadu, India

²Director –Professor & Director, Department Of Computer Science, Dr.G.R Damodaran College Of Science, Coimbatore, TamilNadu, India.

Abstract

Credit risk is the most challenging risk to which all the financial institutions are exposed. Credit scoring is the main analytical technique for credit risk assessment of all financial institutions. Microfinance Institutions - one of the financial institutions who would lend money to financially weaker section people are prone to the credit risk due to their nature of service and need proactive credit risk management techniques for their long-term sustainability. Most of the Microfinance Institutions follow traditional statistical techniques for their Credit scoring. The Credit Scoring with data mining techniques in the microfinance industry is relatively a recent application. This paper explores Credit scoring of Microfinance institutions with a novel non-parametric technique called Support Vector Machine. In the proposed model datasets of a Microfinance Institution in Bangalore are compared with other traditional proposed models . The results show that Support vector machines show a higher accuracy rate of classification.

Keywords —Classification, Credit scoring, Microfinance Institutions, Support vector machine (SVM).

1. INTRODUCTION

Risk is an integral part of financial services. Managing risk is a complex task for any financial organization, and increasingly important in a world where economic events and financial systems are linked. Global financial institutions and banking regulators have emphasized risk management as an essential element of long-term success [1]. When financial institutions issue loans, there is always a risk that the borrower will repay back the amount or the loan could be a delinquent loan .Delinquent borrowers are mentioned as “Default Borrowers”. Financial Institutions always have a threat of – Default Borrowers. Financial institutions take measurements either to avoid this risk or ignore the risk .

Microfinance Institutions (MFIs) , whose main objective is to lend loans to the financially weaker communities also face credit risks that they must manage efficiently and effectively to be successful. If the MFI does not manage its credit risks well, it will likely fail to meet its social and financial objectives. When poorly managed risks begin to result in financial losses, the stake holders tend to lose confidence in the organization and funds begin to dry up. When funds dry up, the social objective of MFI to serve the poor quickly goes out of business.

To overcome this, MFI’s need to have a check on their Customers. One of the Credit Risk management techniques to find the Default Customers is “Credit Scoring”. Credit

Scoring is a technique to determine if a customer is likely to default on the financial obligation.

Specifically, credit scoring problem can be described as follows :

Let’s consider a Dataset of Customers

$$X = [(a_1, b_1), \dots (a_n, b_n), \dots (a_m, b_m)] \quad -(1)$$

Let’s consider each customer a_j which contains c attributes say $(a_j1, a_j2, a_j3, a_j4, \dots a_jc)$ then b_j denotes the type of customer ,for example Good or Bad. The objective of Credit Scoring problem is to construct a model f , for the new customer a such that the model can predict the type of the customer as mentioned in the equation (2).

$$b = f(a) \quad -(2)$$

To evaluate the credit risk and to develop Credit Scoring model as stated above most of the financial organizations use both quantitative and qualitative techniques. Techniques most commonly employed by Financial Institutions and specifically in Microfinance Institutions are discussed in the Literature Review.

The structure of the paper is as follows. Section 2 is the Literature review. The Research methodology is presented in Section 3 and Section 4 draws conclusion .

2. LITERATURE REVIEW

In general, the classification technique to analyse credit risk is applied on the dataset of the previous customers available at the financial institutions to distinguish them as good or delinquent customers, to find if there exists a relationship between the characteristics and reasons for delinquency of loans and accordingly choose an accurate classifier to implement on the new applicants. There has been a good literature survey over years that has shown that there have been different statistical and datamining classification techniques like Logistic Regression, Linear Discriminant Analysis, Naïve Bayes Classifier, Decision Trees, Neural Networks implemented by different financial institutions to discriminate customers as good or delinquent customers.

In comparison to the credit scoring models implemented by different financial institutions, the development of credit scoring models specific to Microfinance sector is very minor. Furthermore, the models which are in existence in large are based on traditional parametric statistical techniques except one or two models based on non-parametric techniques like Multi Layer Perceptron approach (MLP) [6] and Hybridized Models [8].

Among the traditional methods used by Microfinance Institutions, one of the methods used by them to assess their credit risk is CHAID (Chi-squared Automatic Interaction Detection). CHAID is a tree-structured classification method (Kass, 1980). It belongs to a family of segmentation methods known as Automatic Interaction Detection (AID). As its name suggests, the AID allows the detection of interactions between variables. Thus, the segmentation is based on the interactions.

The chi-Square method is used in Microfinance Institutions to assess the effectiveness of credit management systems like Credit terms, Client Appraisal, Credit risk control measures and credit collection policies on Loan Performance of Microfinance Institutions. This method builds "Non-Binary Trees" based on a simple algorithm and divides the customers of microfinance into groups (nodes) of equal category and try to assess the Credit risk [2].

Apart from the traditional methods, comparison of different datamining methods like i) Decision trees ii) Clustering iii) Naïve Bayes Classifier iv) FReBE exemplar, applied on loan data of sub-prime lenders of microfinance institutions have been studied and among all the methods Decision trees appeared to be the most appropriate datamining technology [4]

A Peruvian Microfinance Institute applied Neural Networks, one of the most promising non-parametric credit scoring model based on Multilayer perceptron approach (MLP) on almost 5,500 borrowers and benchmarked its performance against other models like the traditional Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and

Logistic Regression Techniques (LR). The results revealed that Neural Network Model outperformed the other three classical techniques [6]. However, Neural networks have a problem of overfitting and opaqueness.

Along with the above stated techniques, some combined classifiers which integrate two or more classification methods have shown higher accuracy of predictability than individual methods. One such method using a hybridized and ensemble approach, based on the Particle Swarm Optimization (PSO) and Estimation of Distribution Algorithms (EDA) is used by Mexican Microfinance Institution. This ensemble method provided an ease in interpretation of the default customers for the credit officer. The disadvantage of the model is it shows a low reliability [8].

3. RESEARCH METHODOLOGY

This section proposes a model of Credit Scoring based on Support Vector Machine. Compares the sample datasets of a Micro financial Institution in Bangalore with traditional models and Support Vector Machine. The Test results show that Support Vector Machines outperforms other methods.

Support vector machine (SVM) Proposed by Vapnik, is one of the powerful data classification and function estimation tools. This is a method where the input vectors are mapped to a higher dimensional feature space and an optimal separating hyperplane in this space is constructed. The vectors (cases) that define the hyperplane are called the support vectors. SVMs can also efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

The basic idea of SVM is that the original input vector is mapped into a high or infinite dimensional feature space by a nonlinear mapping function f based on Mercer Theory.

The learning process of SVM is to solve quadratic or linear programming and its solution is optimum. SVM has been successfully applied to number of applications like biometrics, text categorization, face and fingerprint recognition etc.

To understand SVM better let's consider a training sample:

$$X = [(a_1, b_1), \dots, (a_n, b_n), \dots, (a_n, b_n)] \text{ from equation - (1)}$$

Let's consider: $\{a_k, b_k\}$, $k = 1 \dots N$ where a_k belongs to \mathbf{R}^d , in the k^{th} input pattern, d denotes the dimension of the input space and b_k is its corresponding observed result, which is a binary variable **1 or -1**.

Applying the above equation to credit scoring models, a_k denotes the attributes of applicants or creditors. b_k is the observed result of whether the debt is repaid timely or whether his or her application is approved. If the customer application

is rejected or if customer default's the debt after being approved, then $\mathbf{b}_{k=1}$,
else $\mathbf{b}_k = -1$.

That is to say that, a rule to separate the point in \mathbf{R}^d plane into two parts is generated with $\mathbf{b}_{k=1}$. The separation of groups (customers into default and non-default groups) is defined as classes. For linear separable samples of two classes, a separating hyperplane is found which maximizes the margin between itself and the nearest training points

To explain it more clearly, let's consider that for the given normal direction \mathbf{w} , there exist two extreme lines L_1, L_2 and all the lines between and paralleling with them can separate two classes exactly. SVM algorithm selects Maximum margin method to separate the two classes. Maximum margin method is to choose the line which creates maximum margin between the two lines, i.e. the distance between the stated two lines L_1, L_2 . For normalized sample, the equation of L_1, L_2 and the distance between the two lines is given as :

$$L1 = \{ \mathbf{w}^T \mathbf{f}(\mathbf{a}_k) + y \geq 1 \text{ if } \mathbf{b}_k = 1 \}$$

(or)

$$L2 = \{ \mathbf{w}^T \mathbf{f}(\mathbf{a}_k) + y \leq -1 \text{ if } \mathbf{b}_k = 1 \} - (3)$$

The distance or margin between the two boundary lines L_1 and L_2 is $2 / \|\mathbf{w}\|$. - (4)

This is because the larger the distance the generalization would be affective.

Further let us assume \mathbf{w}^*, y^* be the solution of equation (3), then $(\mathbf{w}^* \cdot \mathbf{a}) + y^* = 0$ is considered to be the line in the middle of L_1, L_2 , called optimality hyperplane, and the decision function

$f(x) = \text{sign}((\mathbf{w}^* \cdot \mathbf{a}) + y^*)$ would determine which side of the class the customers would fall (either towards Class A or Class B).

In some cases, when the training set is not linear separable, slack variables $\xi_i \geq 0$ is introduced to the i^{th} training point $(\mathbf{a}_i, \mathbf{b}_i)$, and the constraint $f_i((\mathbf{w} \cdot \mathbf{a}_i) + b) \geq 1$ relax to $f_i((\mathbf{w} \cdot \mathbf{a}_i) + \xi_i) \geq 1$. The objective of this is to maximize the margin and minimize the classification error $\sum_{i=1}^n \xi_i$. Then the QP problem is as follows:

$$\min_{\mathbf{w}; \mathbf{b}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i - (5)$$

where $C > 0$ is penalty parameter. The form of optimal hyperplane decision function is the same as the linearly separable case. For nonlinear classification problem, SVM

maps the training samples into a high-dimensional feature space via a kernel function. Once a kernel function $K(x_i, x_j)$ is selected, the hyperplane is determined by

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^n \alpha_j - (6)$$

To obtain better results, it is very important to set related parameter such as selection of kernels. The most often used kernels are linear, polynomial, radial basis function (RBF), sigmoid functions. The parameters that should be optimized are penalty parameter C and α . Thus proper parameter setting can improve SVM classification accuracy.

Hence applying the above equations the algorithm of our credit scoring model for Microfinance Institutions using SVM can be described as follows:

Step 1: Start

Step 2: Collect the Data from the Field Officers of Microfinance Institutions.

Step 3: Identify the details.

Step 4: Distinguish the applicant's details such as credit history, account balances, loan purpose, loan amount, employment status, personal information, age, housing, job etc as classification attributes.

Step 5: Use SVM to obtain the variables of importance.

Step 6: If variables of importance are listed accept them else

Step 7: Drop features with zero variable importance.

Step 8: Rebuild data set and randomly partition into a training set and a testing set.

Step 9: Choose one of the kernel functions like Radial Basis Function (RBF) Kernel Function to optimize the parameters.

Step 10: Choose penalty parameter (C) that leads to the lowest error rate.

Step 11: SVM then classifies the applicants to default and non-default classes.

Step 12: Stop

3.1 Processing of Data Sample.

To determine the performance of SVM over other datamining techniques a small sample dataset from a Microfinance Institution in Bangalore has been taken and analysed. The sample data consists of 157 instances. It consists of customer information like i) Personal Characteristics of borrowers (marital status, sex etc) ii) whether the loan is a group loan or individual loan. iii) Economic and Financial ratios of the Microfinance Institution. iv) Characteristics of current financial operation (type of interest, amount, delays in the payment) etc. 24 features of the above type are mentioned of which 16 features are selected. To perform an appropriate comparison of classification models the data is split into two subsets a) A training set of (56 instances) and b) A test sample of (101 instances) and tested on Matlab software. To check SVM ,Radial basis Kernel(RBF) is selected. The Percentage of Accurately classified data of default and non-default cases and Inaccurately classified data are monitored on five techniques like Logistic Regression(LR), Quadratic discriminant Analysis(QDA), Linear Discriminant Analysis(LDA) ,Back Propagation neural network (BP) method and Support Vector machine . The results are placed in Table-1.

Table- 1 empirical results on dataset

Model	Train sample Set (56)		Test Sample Set (101)	
	Accurate	Inaccurate	Accurate	Inaccurate
LR	77.23%	22.77%	73.15%	26.85%
QDA	71.24%	28.76%	71.50%	28.50%
LDA	77.54%	22.46%	73.45%	26.55%
BP	82.25%	17.75%	79.2%	20.8%
SVM	87.54%	12.46%	83.17%	16.83%

4. CONCLUSIONS

The empirical results of Table-1 clearly shows that SVM gives an higher accuracy rate both in test sample and trained sample which indicates clearly that SVM is a better classification technique then other traditional techniques and also to the non-parametric Back Propagation neural network technique.

This method would be best suitable for the Microfinance institutions who would deal with customers from economically weaker sections. The sustainability of the Institution depends on how they accurately differentiate the customer who would pay or not and in this regard SVM would be an appropriate method and would help to have a competitive advantage to the Microfinance Industry over other Commercial Banks.

REFERENCES

- [1]. Bald, Joachim. January 2000." A Risk Management Framework For Microfinance Institutions "Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH. Postfach 5180, 65726 Eschborn, Germany. Internet: <http://www.gtz.de>.
- [2]. G.D.Gyamfi, "Assessing The Effectiveness Of Credit Risk Management Techniques Of MicroFinance Firms in ACCRA", Journal Of Science and technology ,Vol.32,pp.96-103,2012.
- [3]. Haron.O.Moti,Justo simiyu Masinde,Nebat Galo Mugenda ,Mary Nelima Sindani, "Effectiveness Of Credit Management System on Loan Performance: Empirical Evidence From Microfinance sector in Kenya", International Journal Of Business ,Humanities and technology , Vol.2,No 6 ,October - 2012.
- [4]. Jia Wu,Sunil Vadera,Karl Dayson,Diane Burr ridge,Ian Clough "A Comparison of Datamining Methods in Microfinance.", IEEE Conferences,2010.
- [5]. Jozef zurada,K.Niki Kunene , "Comparisons of the performance of Computational Intelligence Methods for Loan Granting decisions" IEEE conferences ,2011.
- [6]. Antonio Blanco,Rafael Pino-Mejias,Juan Lara,Salvador Rayo, "Credit Scoring models for MFI using Neural Networks : Evidence from Peru",Expert Systems with applications,Vol.40 ,pp 356-364, 2013.
- [7]. Xiang Hui,Yang Sheng Gang , "Credit Scoring model based On Selective Neural network Ensemble",IEEE conference papers.
- [8]. Arturo Elias,Alejandro Padilla,Felipe Padilla , "Credit Scoring for Microfinance Institutions in Mexico an ensemble and Hybridized approach", International Conference on Technology and Management ,Vol.21, 2012.
- [9]. Defu Zhang,Hongyi Huang,qingshan chen,Yi Jiang, "A Comparison Study Of Credit Scoring Models", IEEE conference – 2007.
- [10]. V. Vapnik, *The Nature of Statistical Learning Theory*. New York:Springer, 1995.
- [11]. Yongqiao Wang,Shouyang Wang and K.K.Lai," A new Fuzzy Support Vector machine To Evaluate credit Risk", IEEE Transactions on Fuzzy Systems, Vol 13 No 6,2005.
- [12]. Wei Xu,Shenghu Zhou,Donmei Duan,Yanhui Chen," A Support Vector machine Based Method For Credit Risk Assessment",IEEE International Conference on E-Business Engineering.
- [13]. Cheng-Lung Huang a, Mu-Chen Chen b, Chieh-Jen Wang, " Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications 33 , 847–856
- [14]. Rashmi Malhotra,D.K Malhotra, "Differentiating Between Good Credits and Bad Credits sing Neuro –Fuzzy Systems ",European Journal of Operations Research ,136 .pp 190-211.
- [15]. Zan Huang,, Hsinchun Chena, Chia-Jung Hsua, Wun-Hwa Chenb, Soushan Wu," Credit rating analysis with support vector machines and neuralnetworks: a market comparative study", Decision Support Systems 37 , pp-543– 558

- [16]. Tian –Shyug Lee,Chih-Chou Chiu,Yu-Chao Chou,Chi-Jie Lu, “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines”,*Computational Statistics and data Analysis* ,50,pp-1113-1130.
- [17]. David Martens,Bart Baesens,Tony Van Gestel,Jan Vanthienen,“Comprehensible credit scoring models using rule extraction from support vector machines”,*European Journal Of Operations Research* 183, pp 1466-1476.
- [18]. Ligang Zhou, Kin Keung Lai,”Multi-Agent Ensemble Models Based on Weighted Least Square SVM for Credit Risk Assessment”,*IEEE conference papers* .
- [19]. Eliana angelini,Giacomo di Tollo,Andrea Roli,“A neural network approach for Credit Risk Evaluation.” ,*The Quaterly review Of Economics and Finance* ,48,pp 733-755.
- [20]. Adel Lahsasna, Raja Noor Ainon, and Teh Ying Wah,” Credit Scoring Models Using Soft Computing Methods: A Survey”, *The International Arab Journal of Information Technology*, Vol. 7, No. 2, April 2010
- [21]. Baesens B., Setiono R., Mues C., and Vanthienen J., “Using Neural Network Rule Extraction and Decision Tables for Credit Risk Evaluation,” *Computer Journal of Management Science*, vol.49, no. 3, pp. 312-329, 2003.
- [22]. Weimin Chen *, Chaoqun Ma, Lin Ma,” Mining the customer credit using hybrid support vector machine technique”, *Expert Systems with Applications* 36 , 7611–7616.