ISSN: 0976-3104
REGULAR ISSUE
Mittal et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

PREDICTION OF CREDIT RISK EVALUATION USING NAIVE BAYES, ARTIFICIAL NEURAL NETWORK AND SUPPORT VECTOR MACHINE

Lohit Mittal*, Tarang Gupta, Arun Kumar Sangaiah
School of Computing Science and Engineering, VIT University, Vellore, TN-110003, INDIA

ABSTRACT

An accurate prediction of credit risk evaluation is very useful for the banking and financial industry in minimizing the risk in lending credit to the customer and decreasing the chances of making wrong decision. As the increase in customers and emergence of different trade, it is difficult for bank management to analyze individual physically hence data mining algorithms are implemented in order to reduce the work effort by the bank management. This study attempted to implement three data mining model and compared their performances in predicting the risk in giving credit to the customer. In this study neural network based on back propagation, naive Bayes algorithm and support vector machine were implemented. Eight technical indicators were used as input for the above models. Data preprocessing were done to increase the performance of their prediction. A comparative analysis of the models was carried out and experimental results showed that the performance of Support Vector Machine (92%) was higher than the other two. Naïve Bayes (87%) performance was found little better than that of Artificial Neural Network (85%). Findings from this study can help in improvement of the existing system.
.

Received on: 5th-May-2016
Revised on: 22nd -June-2016
Accepted on: 29th -June-2016
Published on: 28th -July-2016

KEY WORDS

Support Vector Machine; Artificial Neural Network; Binary Classification; Prediction; Credit

***Corresponding author: Email:** lohitmittal22@gmail.com, **Tel:** +91-9159065928

INTRODUCTION

The credit provided by the bank is one of the important features of the banking industry. It is the main source of income for the banks. A credit risk is defined as a failure in returning the loan by the borrower to the lender. Due to increase in globalization and opening of various business opportunities, many small scales or large scale firms require credit to invest in their business or sometimes credit is required in situations like purchasing new house, vehicle or any other necessary expensive stuff. Sometimes a loan is needed for education. As the foreign study rates are very high, students tend to apply for educational loans. In the recent years, banks are facing crises in financial sector as the risk in lending credit to the borrowers is increasing at alarming rates.

As borrower credit risks can not only be determined by the assets the borrower posses during the application of loan, there are several other factors also responsible which can determine the credit risk. As some of the bank handle customers overseas and exponential increase in borrowers it is very difficult for the bank management to analyze the borrowers. Hence banking industry requires a credit risk evaluation system to filter the borrowers' base upon some of their characteristics and with precise accuracy to reject their loan application at preliminary stage to reduce the risk in lending credit .In this study the factors taken under consideration are age, years in service, total income (Per annum), number of dependent, tenor, cash flow, collateral and credit history. A credit score is a formulated expression used to analysis of a person credit file. Banks and lenders uses credit score to determine potential risk possess by creditor while taking credit or loan from lender. The study done is to analysis, compare and study different data mining algorithms involved in classification of the above mentioned factors to determine whether there is a credit risk or not. In order to achieve this goal we have undertaken an assessment, we have chosen three methodologies.

THE IIOAB JOURNAL
ISSN: 0976-3104

Artificial Neural Network (ANN)

The study on ANN dates back to 1970s by Frank Rosenblatt(1958) [1] who worked on first perceptron algorithms and the outcome of the study was used to develop smart automated software and systems. ANN has been found to work smartly and give promising results [2-7]. Neural network received a boost in its processing after the publication of machine learning by Marvin Minsky and Seymour Papert [8]. Odom and Sharda (1990) [9] were the first to apply NNs in the credit risk evaluation. The early network was based on Hebb network which aims at updating the input vector then perceptron comes into picture which increase the accuracy and was base for many models. After perceptron neural backpropogation comes into picture which was developed by David E.Rumelhart and James McClelland [10]. Backpropogation has the feature to update its weights by keeping the history. It is known as neural processing.But in late 2000 a deep learning created more interest. The first study on credit risk evaluation was done by Angelini et al. 2008 [11] in which the banking management could calculate capital requirement using key risk drivers. An application of ANN by Mohsen et al. 2013 [12] was presented for calculating the variables required for credit risk evaluation.

Support Vector Machine (SVM)

Study using Support Vector Machine has been proven to be promising and efficient in determining credit risk evaluation when compare to other algorithms such as neural network, particle swarm optimization (PSO), and other machine learning classification algorithms. Several SVM based studies for credit risk evaluation problems have been proposed by Ahn et al., Wu et al. Zhang et al and others [13]. The results have been promising and their main work was to compare with genetic algorithms. Others important approaches include combining SVM and fuzzy logic. This fuzzy logic work was attemted by Hao et al [14] who used fuzzy sets whereas Huang et al implemented least square SVM. Xuchuan et al who used particle swarm optimization for optimal parameter selection for SVM. A comparative research by these authors proved that SVM classifier has the potential to replace other algorithms in complexity, speed and reliability.

Naïve Bayesian

The study Naïve Bayesian has been based on Appling Bayes theorem for assuming between the features. When compare to other algorithms such as neural network, particle swarm optimization (PSO), and other machine learning classification algorithms, Naïve Bayesian is proven to be promising and efficient in determining credit risk evaluationIt has more advanced methods as compared to Support Vector Machine. Naïve Bayesian has used in medical diagnosis.Russell and Norvig was the first to study about Naïve Bayesian and they have mentioned in their first book. Rish, Irina in 2001who worked on an empirical study of the Naïve Bayesian. In 2003 Rennie, J.; Shih, L.; Teevan, J.; Karger, D. worked on tackling the poor assumption of Naïve Bayesian classifiers [15-20].

MATERIALS AND METHODS

Research Data

This section describe about the data set on which the research has been carries out and the attributes used for prediction. The data set was collection from Resort Savings and Loan Plc, Lagos Island, Lagos, Nigeria. A total of 200 records are taken and accessed. As there were some missing values, the dataset required some preprocessing. So replacing the missing value with the global mean of the same attribute method was adopted. The data set was again preprocessed as it needed to be converted to binary data set where the output can be either 1 or 2. This 1 or 2 represent the label rather than their usual value, where 1 represents credit risk and 2 represents credit no risk. The output given by is in range of 0 to 1. As given by , the output is marked 2 if value is greater than 0.75 and 1 if the output is lower than 0.75, this result in a binary data set as shown in **Table-2.**

There are eight predictor attributes in the data set that are age, years in service, total income (Per annum), number of dependent, tenor, cash flow, collateral and credit history. To increase the accuracy of the output the eight data sets are converted to three categories as shown in **Table-1**. The classification is a necessary preprocessing as to increase the accuracy of all the three algorithms. Many financial officials approved those eight to be the important factors to be responsible for prediction of credit risk evaluation process.

Here Low is assigned class label 1, Moderate is assigned class label 2 and High is assigned class label 3.Hence the preprocessing of data set is done. Further there is normalization of data is done for artificial neural network as the output of neural network is in the range of 0 to 1, hence we need to normalize the value between 0-1 of the output variable.0.2 is chosen as value corresponding to class risky and 0.8 is chosen for class non risky.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table :1.input variable numerical value range

CODE	Variable Description	Low	Medium	High
C1	AGE(years)	45-60	33-45	18-33
C2	LENGTH OF SERVICE	18-30	10-18	1-10
C3	TOTAL INCOME	1-1.2x10^7	1.2x10^7-2.2x10^7	>2.2x10^7
C4	NO. OF DEPENDENT	>5	3-4	1-2
C5	TENOR	240-360	120-240	1-120
C6	CASH FLOW	1	2	3
C7	COLLATERAL	1	-	2
C8	CREDIT HISTORY	1	-	2

Table: 2. Conversion into binary data set

S.no	Variable Description	Class	Credit
1	>=0.75	1	Not-Risky
2	<=0.75	2	Risky

Table: 3. Summary statistics for the selected indicators

	Max	min	Medium	S.D
AGE(years)	57	27	36.6	6.928
LENGTH OF SERVICE	31	2	9.235	6.65
TOTAL INCOME	9076832	56789	1544842	1252815.63
NO. OF DEPENDENT	6	1	3.145	1.369
TENOR	360	45	224.7	83.6
CASH FLOW	3	1	2.13	0.829
COLLATERAL	2	1	1.75	0.434
CREDIT HISTORY	2	1	1.9	0.280

The given data set was normalized and converted. The data was stored in a CSV (comma delimiter) and was stored in mysql database using PHP Myadmin for naïve Bayes and artificial neural network. The data set is derived using java JDBC driver and loaded for further functions. The data set have been divided into set of four on the bases of age. 60 % of each data set is taken for training and the rest is for the holdout. The weights are calculated using training and tested on the holdout datasets.

PREDICTION MODEL

Bayesian Network

The naive Bayesian classification gives the class label of a data which is their in the table, the values of the entity and their attribute are predicted to be conditionally independent of one another. Bayesian classification is the statistical classifiers and every new data which we get has belongs to a class. The method used in Bayesian classifier are joint conditional probability distributions, which allows class conditional independencies to be work between subsets of variables and also it have a graphical model of relationships, by which various interpretation data is performed. The two categories of belief network are first one is a directed acyclic graph and second one is a set of conditional probability tables. Each and every node in directed acyclic graph shows a variable and it is a random variable which is either in the form of discrete or in the form of continues. The attribute can be real which are given in the data or they can be invisible variables and are believed to form a relationship. In this directed acyclic graph each and every arc represents dependence probability. An arc is made from a point C to a point D, then C is a parent D, and D is a descendant of C. Every variable in the graph is independent of its non-descendants, gives its parents.
Bayes theorem formula is given by,

THE IIOAB JOURNAL
ISSN: 0976-3104

COL 1	COL 2	COL 3	COL 4	COL 5	COL 6	COL 7	COL 8	COL 9
2	1	2	1	3	2	1	1	1
2	2	1	3	3	1	1	1	1
1	1	3	2	3	2	2	1	2
3	2	1	2	2	1	1	1	1
1	1	1	1	3	1	1	1	1
2	2	1	3	2	1	1	1	1
3	3	3	3	3	3	2	2	2
2	2	2	2	3	1	1	1	1
2	1	2	3	2	2	1	1	2
2	2	2	2	2	2	1	1	1
2	1	2	2	3	2	1	1	2
1	1	1	2	2	1	1	1	1
3	2	3	3	3	3	2	2	2
2	1	3	2	3	2	2	1	2

Fig:1.Sample data after conversion

Let assumed that there is a sample called A, the probability of all the possible events h, P(h|A) follows the Bayes theorem stated mathematically as the following equation

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayesian classification has been develop which is most optimal one. When the network data and its topology is given in th data the various variables used in the sample is known then training the network is candid. It is used to find the continuous probability table (CBT) entries. This is analog to the way of finding the computing probability which is their in naive Bayesian classification.
In Bayesian network, Naïve Bayesian classifier is also used in which we assume that attributes are conditionally independent.

$$p(C_K \mid x_1 \ldots\ldots x_n) = \frac{1}{Z} P(C_K) \prod_{i=1}^{n} p(x_i \mid C_K)$$

This is naïve hypothesis. Naive Bayesian classifier is very effective in case of cost; it reduces the calculation cost It is best for the problems where there is strong relation between the variables and the data given.

Support Vector Machine

SVM was developed in COLT-92 by Boser, Guyon, Vapnik[10] and having several application like bioinformatics, text handwriting recognition etc. Support vector machines (SVMs) are a way of classification of both linear and nonlinear data, that is, an SVM is an algorithm that works as follows.

1. It uses a nonlinear mapping to convert the given training dataset into a poly-dimensional.
2. In the same dimension, it identify for the linear optimal separating hyperplane (i.e., a “decision boundary” separating the tuples of one class from another). Having an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.
3. The SVM finds this hyperplane using support vectors (“essential” training tuples) and margins (defined by the support vectors).The task of this type of algorithm is detect complex pattern in data by various data mining approaches like clustering, classifying, ranking, cleaning etc.
4. SVM is a particular instance of kernel machine has large class of learning algorithms. The class of possible patterns are defined by various classes of kernel methods implicitly by introducing a notion of similarity between data and this methods exploits the information about the inner products between the data items.
5. There are two cases in SVM which are case when the data is linearly separable and the other is when the data is non-linearly separable.
6. Although the training time of even the fastest SVMs is slow, but they are highly accurate, have the power to solve complex nonlinear decision boundaries. They are much less prone to over fitting than other methods.
7. SVMs can be used for numeric prediction as well as classification. They have been applied to a number of areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests.

THE IIOAB JOURNAL
ISSN: 0976-3104

The various properties used by the SVM algorithm are as follows:-

1. Duality is the first property of Support Vector Machines in which SVM is represented as Linear Learning Machines in a dual fashion and the data appear only within dot products. But there are many limitation in the first feature of SVM so to minimize this limitation, one theorem was introduced called as Mercer's theorem which solve the symmetric positive definite.
   Linear learning Machines the second feature of SVM that uses dual representation concept which operates in a kernel induced space as a linear function. Various other algorithms like clustering, PCA can also be used by which dual representation often possible (in optimization problems, by Representer's theorem).The various generalization problems which can be solved in SVM are:

a. The effect of dimensionality in which it is very easy to over fit in high dimensional space.
b. The SVM problem of finding one hyperplane that separates the data: many such hyperplanes exist)

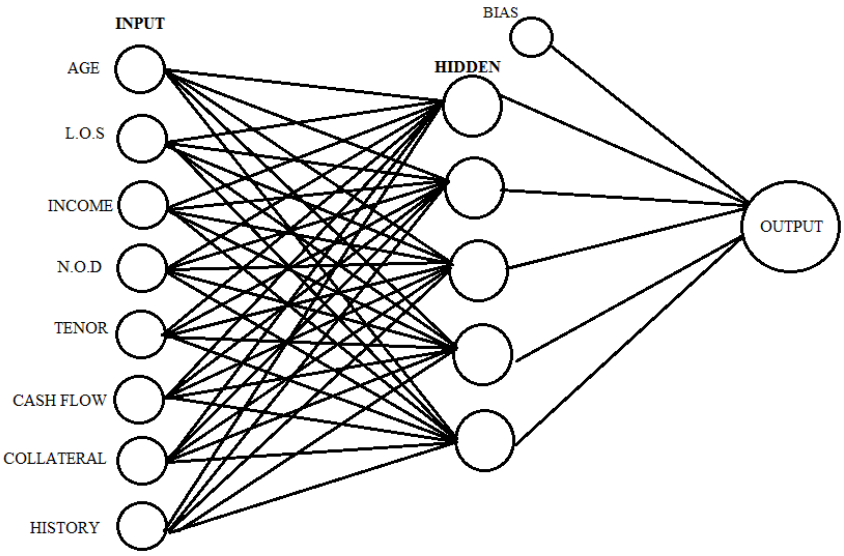

Fig: 2. Architechture of ANN

……..

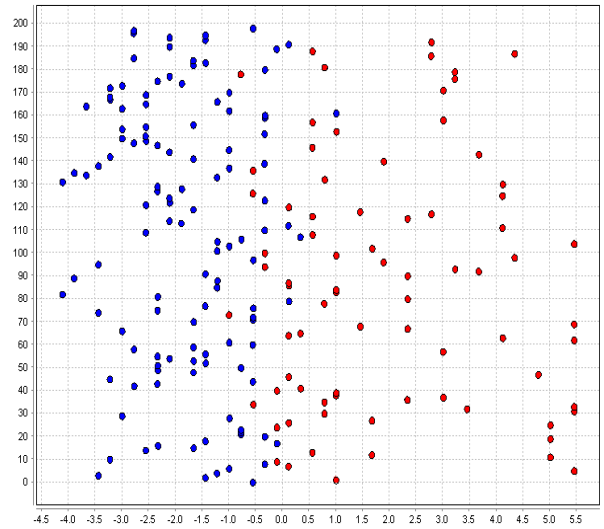

Fig: 3. SVM polynomial output

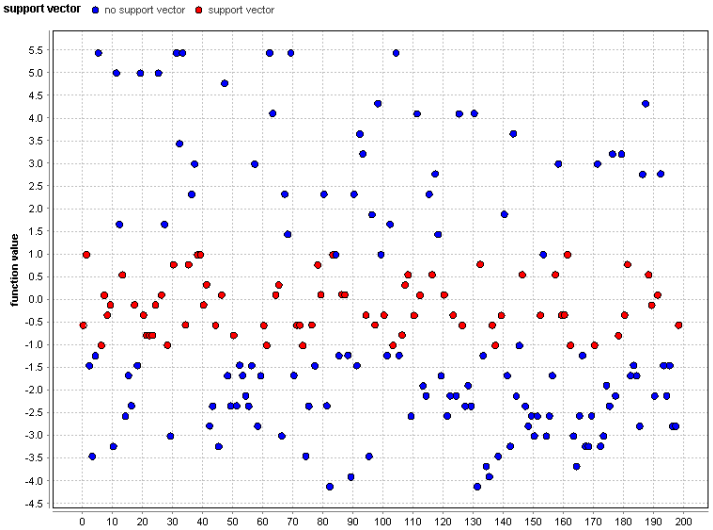

Fig: 4. SVM radial output

……..

THE IIOAB JOURNAL
ISSN: 0976-3104

Artificial Neural Network

Artificial Neural Network is an algorithm or a system used to calculation purpose, work on various methods, and to study the various problems related to biological field. Artificial Neural Networks are details processing algorithms which are built and to model the human brain. The main objective of the neural network research is to make a calculation system for designing the experts system to perform calculation faster.

Various tasks performed by ANN such as pattern-matching, classification, optimization of function, approximation, vector quantization, data clustering. ANN was introduced by biological nervous system such as brain processes information. The net input to the neuron u is given by the formula:

$$v = \sum_{j=1}^{m} w_j x_j$$

The activation function is applied to yin to compute the output and the weight represents the strength of synapse connecting the input and the output neurons. The weights may be positive or negative. The positive weight means the synapse is excitatory and the negative weight means the synapse is inhibitory.

Properties of ANN are:-

1. The speed or cycle time of execution of ANN is of few nanoseconds.
2. The processing time of ANN is very fast and it can perform several parallel operations simultaneously.
3. The size and complexity of ANN depends upon the chosen application and the network designer
4. The ANN stores the data in its contiguous memory locations and in ANN the loading may sometimes overload the memory
5. The main property of ANN is learning and learning. There are two kinds of learning in ANN
    1. Parameter learning: It modernize the weight which are linked in the network.
    2. Structure Learning: It concentrate on the topology of the network and check whether any change in the network.
6. learning can be :
    1. Supervised learning: Learning through a teacher/ supervisor
    2. Unsupervised learning: Learning in the absence of supervisor.
    3. Reinforcement learning: Learning basing upon a critic information and it is similar to supervised learning

ANN also has activation function which is used to calculate the exact output. The activation function is applied over the overall input to find the overall output of an ANN. There are many activation function like identity function, binary step function, sigmoidal function, bipolar sigmoid function, hyperbolic tangent function etc

RESULTS AND DISCUSSION

Experimental Results

The Naïve Bayes algorithm and Artificial Neural Network with back propagation was implemented in Java Servlet Package, JDK 8.0 and the data set was uploaded to MySQL as a CSV file whereas for the Support Vector Machine Rapid Miner Tool was used .At initial stage the data set was divided into four sets. The age was chosen to determine the dividing factor as young (18-32), middle (32-35), semi-old (36-44), and old (45-60).At initial stage of ANN the best combination of the four parameters needs to be determined that is number of epoch, learning rate, number of neurons and mc. The best combination is observed and collected results are presented in **Table-4** and **Table-5**. It should be taken care that not only parameter combination but they need to be applied and checked for best accuracy. Now with fixation of three combinations (ep; mc ;n) as shown in **Table-4** , learning rate is iterated and a plot of different accuracy is shown in**[Figure-1]**as a non-linear graph is obtained. As seen from figure the best performance was obtained at lr = 0.2. The average accuracy observed is 85 percent. Next is Support Vector Machine, it is also applied on four sets and average performance is calculated. First SVM polynomial is applied and then radial SVM is applied. In polynomial SVM is iterated for combination of three parameter (d; y; c) as shown in **Table-6**. In polynomial SVM the results obtained were 87 percent whereas in case of radial basis the results obtained were around 76 percent. In radial basis two parameter combinations were iterated as shown in **Table-7**. The average accuracy of polynomial is taken for comparison. Next is Naïve Bayes algorithm. In this algorithm is applied to four sets in one process as seen in **Table-8** and then the algorithm is applied to the whole dataset as a lump sum and a confusion matrix is created , as it can be seen that the best iteration results have been taken for comparison.

THE IIOAB JOURNAL
ISSN: 0976-3104

TABLE: 4. ANN Holdout Accuracy Parameter combination (ep;mc;n)

AGE(years)	(5000;0.7; 5)		(5000;0.7; 5)		(5000;0.7; 5)	
	Training	Holdout	Training	Holdout	Training	Holdout
18-32	98	86.3	98	83.4	98	83.2
32-35	99	85.4	99	84.1	99	84.2
36-44	98	85.5	98	83.2	98	80.2
45-60	98	86.3	98	85	98	85.5
AVERAGE	98.5	85.4	98.5	83.9	98.5	83.27

Table: 5. Best three combination of ANN model

No.	Lr	Ep	Mc	n	Training	Holdout	Average
1	0.1	5000	0.7	5	98.18	86	93
2	0.1	7000	0.1	6	98.54	85	92
3	0.1	6000[7]	0.4	7	98.18	85.6	92.3

As naïve Bayes algorithm doesn't have any parameters hence there is no adjusting need to be done. The Algorithm's efficiency increases with the increase of the data set tuples. Finally all the algorithms accuracy has been compared as shown in **Table-10**. It can be seen that all the algorithms give a promising result in prediction of given problem.

Table: 6. Prediction performance (%) of polynomial SVM model

AGE(years)	(3;2.6;100)		(3;2.5; 100)		(3;3.2; 100)	
	Training	Holdout	Training	Holdout	Training	Holdout
18-32	99	92.3	99	90.5	99	90.1
32-35	98	91.6	98	91.6	98	90.3
36-44	97	92.0	97	91.0	97	92.1
45-60	99	91.8	99	90.3	99	92.2
AVERAGE	98.2	92.1	98.2	91.1	98.2	92.8

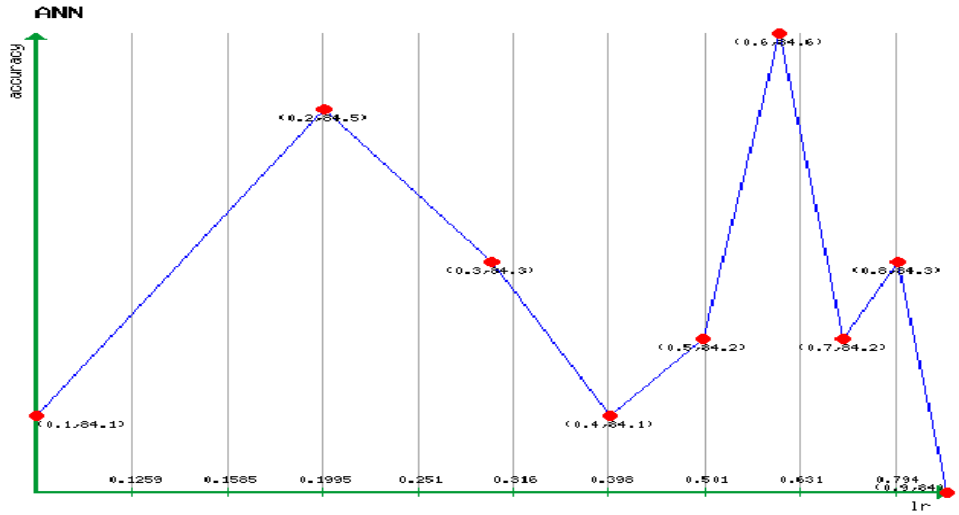

Fig: 5. accuracy vs. lr

THE IIOAB JOURNAL
ISSN: 0976-3104

Table :7 .Prediction performance (%) of radial basis SVM model

AGE(years)	(2.5;100)		(3.0; 100)		(3.1; 100)	
	Training	Holdout	Training	Holdout	Training	Holdout
18-32	99	78.2	99	75	99	77.2
32-35	98.3	76.4	98.3	76.5	98	77.1
36-44	97.8	79.0	97.8	77.3	97	78.3
45-60	98.4	77.3	99	75.2	99	76.1
AVERAGE	98.3	77.7	98.5	76	98.2	77.2

Table: 8. Naïve Bayes Accuracy

Age	Training	Holdout
18-32	100	87.2
32-35	100	89.1
36-44	100	86.4[7]
45-60	100	87.5
AVERAGE	100	87.5

Table: 9. Naïve Bayes Confusion Matrix

Confusion Matrix	risky	non risky
risky	78	5
Non risky	10	87

Table: 10. Final Comparison

Methodology	Accuracy
ANN	**83%**
SVM	**92%**
Naïve Bayesian	**87%**

The novelty of the work is to make comparison among machine learning algorithm and to determine the best which we obtained was SVM as it more efficient in classification than the other two, although the other algorithm were slightly less they can also be used to determine the credit risk evaluation. The results obtained can be base for implementation of a system for credit risk evaluation.

THE IIOAB JOURNAL
ISSN: 0976-3104

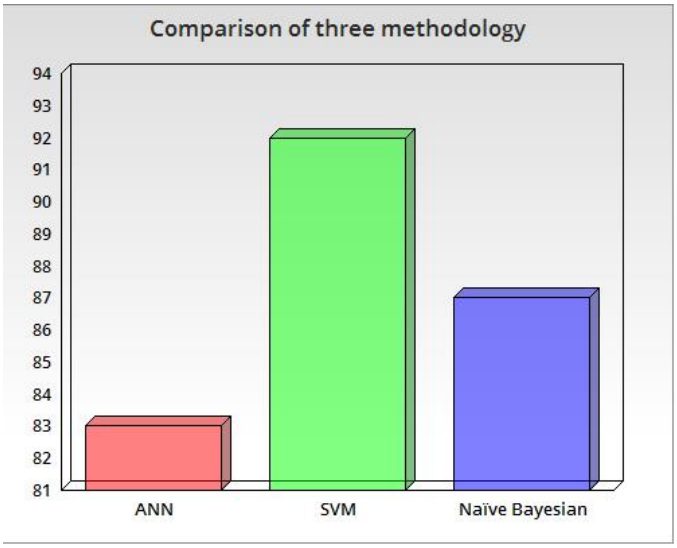

Fig: 6. Results of ANN, SVM, and Native Bayesian

CONCLUSION

Prediction of credit evaluation decision is necessary, it helps the bank management to enhance their decision making and reduce their losses by taking right decision and also filtering loan application at preliminary stage in order to reduce the work overhead. This task is difficult and requires high skills. The study attempted to predict the credit risk of an individual and compare and analyze three algorithms in doing so. Based on the experimental results obtained, some important results can be drawn such as all the three algorithms are capable of producing significant accuracy during classification. Thus we can say all the three are useful tool for prediction for other problems also. The best prediction was found by SVM (92%) followed by Naïve Bayes (87%) and last ANN(83%) **Figure-6**, however ANN performance can be increased by performing fuzzy ANN mentioned in the literature. The efficiency of the Naïve Bayes algorithm can be increased by the grouping of elements. The study aims at comparison of the above three algorithm in terms of their accuracy and the best can be used for implementation of credit risk evaluation as shown above.

CONFLICT OF INTEREST

The authors declare no competing interest.

ACKNOWLEDGEMENT

We are thankful to VIT University for providing necessary resources for successfully implementation of this system

FINANCIAL DISCLOSURE

None declared.

REFERENCES

[1] Rosenblatt F. [1958] The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain". *Psychological Review* 65 (6): 386–408.doi:10.1037/h0042519. PMID 13602029.

[2] Sangaiah AK, Thangavelu AK., Gao XZ, Anbazhagan N, Durai MS. [2015a]An ANFIS approach for evaluation of team-level service climate in GSD projects using Taguchi genetic learning algorithm. *Applied Soft Computing* 30: 628-635.

[3] Sangaiah AK., Gao XZ, Ramachandran M, Zheng X.[2015b] A fuzzy DEMATEL approach based on intuitionistic fuzzy information for evaluating knowledge transfer effectiveness in GSD projects. International *Journal of Innovative Computing and Applications* 6(3-4):203-215.

[4] Huang Z, Chen H, Hsu CJ, Chen WH, Wu S. [2004] Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems* 37(4): 543-558.

[5] Huang W, Lai K., Nakamori Y, Wang S. [2004] Forecasting foreign exchange rates with artificial neural networks: A review. *International Journal of Information Technology*, 3: 145-165. DOI: 10.1.1.121.8174

THE IIOAB JOURNAL
ISSN: 0976-3104

[6] Jang JS, Sun CT. [1993] Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks* 4 (1): 156-158
[7] Hawley DD, Johnson JD, Raina D. [1990] Artificial neural systems: A new tool for financial decision-making. *Fin Anal J* 46: 63-72. DOI: 10.2307/2F4479380 Huang
[8] Minsky M, S Papert. [1969] An Introduction to Computational Geometry. MIT Press.ISBN 0-262-63022-2
[9] Odom M, Sharda R. [1990] A neural network model for bankruptcy prediction. Proceedings of the International Joint Conference on Neural networks, 163-168.
[10] Rumelhart DE, James McClelland. [1986] Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge: MIT Press.
[11] Angelini E. di Tollo, G Roli A. [2008]. A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, 48 (4): 733-755
[12] Mohsen N, Mojtaba A. [2013] Measuring credit risk of bank customers using artificial neural network. *Journal of Management Research* 5(2):17. Mulhim
[13] Ahn H, Lee K, Kim K.[2006] Global Optimization of Support Vector Machines Using Genetic Algorithms for Bankruptcy. *Lecture Notes in Computer Science* 4234(5):420-429. Springer.
[14] Wu C, Tzeng G, Goo Y, Fang W.[2007] A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications* 32: 397-408.
[15] Zhang D, Chen Q, Wei L.[2007] Building Behavior Scoring Model Using Genetic Algorithm and Support Vector Machines. *Lecture Notes in Computer Science* 4488:482-485.
[16] Hao PY, Lin MS, Tsai, LB.[2008] A New Support Vector Machine with Fuzzy Hyper-Plane and Its Application to Evaluate Credit Risk. *2008 Eighth International Conference on Intelligent Systems Design and Applications*, 83-88.
[17] Russell Stuart, Norvig Peter. [1995] Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.
[18] Rish Irina. [2001] An empirical study of the naive Bayes classifier (PDF). IJCAI Workshop on Empirical Methods in AI.
[19] Zhang, Harry. The Optimality of Naive Bayes (PDF). FLAIRS2004 conference.
[20] Rennie J, Shih L,Teevan J, Karger D. [2003] Tackling the poor assumptions of Naive Bayes classifiers (PDF). ICML.

ABOUT AUTHORS

Lohit Mittal *is currently studying B.tech (computer science and technology) at VIT University, Vellore.*
Tarang Gupta *is currently studying B.tech (computer science and technology) at VIT University, Vellore.*
Dr. Arun Kumar Sangaiah *is a Associate Professor at VIT university and PhD in computer science.*