A Comparative study analysis of Loan Approval Prediction Algorithms

[1]Dr.E.ChandraBlessie, [2]R.Rekha
[1]Associate Professor, [2]Ph.D., Research Scholar,
[1,2] Department of MCA,
[1,2]Nehru College of Management, Coimbatore, Tamilnadu, India

Abstract: Recent revolutions in the technology such as Big Data, data availability and computing power, most banks or lending financial institutions are in need of improving their business by revising or innovating business models. Credit risk predictions, monitoring, model reliability and effective loan processing are considered as the important functions. The main features for issuing the loan are Loan amount, costumer's history and many other factors. The main challenging factor is to assess whether the applicant would be eligible to receive the loan or not. This work presents a comprehensive analysis of various prediction algorithms that have been carried out in financial sectors for various operations. From the detailed analysis, it is observed that the classification played a vital role in data mining tasks and various decision tree algorithms have been used in the bank sectors.

Index Terms **– Loan approval, Big Data, Finance and Prediction algorithms.**

I. INTRODUCTION

Information and Communication Technology has intervened in all parts of the world and among the various technologies, it is found that data mining has occupied a great position in the financial segments such as prediction of payment default, marketing, credit analysis, advertisement, cash management and forecasting operations. The major risks that are encountered by the banking industry is credit risks which describes the risk of loss and loan defaulters. To overcome the risk, data mining techniques could be applied. Knowledge extraction or knowledge mining from huge data storage is referred as data mining [1]. Different data mining tasks could be depicted in the below Fig 1.

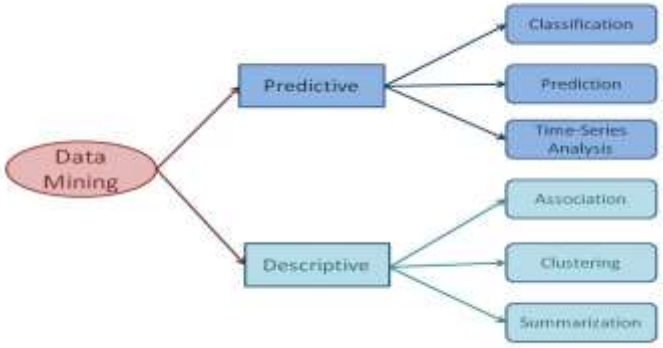

Figure 1. Data mining tasks

There are two kinds of functions involved in data mining and they are

1. Classification and Prediction.
2. Descriptive.

Descriptive functions are described as

1. Mining of clusters
2. Mining of Associations
3. Mining of correlations
4. Mining of frequent patterns
5. Class/Concept Description

1.1 Classification and Prediction

It is the process of identifying a model that explains about the concepts or data classes. The objective of this process is to use the model so that predicting the class is done. Training data sets is used by this model. The model could be denoted in various forms such as

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae

- Neural Networks

Several prediction and estimations have come up in all sectors of the market. A financial system is defined as the group of procedures that tracks the financial activities. It is a system which facilitates the lenders and borrowers to exchange funds. Moreover, this system covers the exchange of money between investors, lenders and borrowers, and financial transactions. Banks plays a vital role in the nation's economy. It also realizes that retaining the customers and preventing fraud must be the strategy tool for healthy competition. The most important operations in the bank sector is the distribution of the loans since the bank's assets are raised due to the profit earned from the loans that are distributed by the banks. In spite of many advertisements regarding the sanctioning of loan availed as a marketing strategy by the bank, approving the loan to a particular person is a tedious process. A detailed fundamental analysis is required in order to analyze the data provided by the entity. Many risk factors are related to the bank and as well as to the persons who got loans from the bank. A comprehensive analysis of risk needs a better understanding. Risk in bank loans could be categorized as liquidity risk, interest rate risk and credit risk. To overcome these risks, various data mining algorithms could be applied.

II. RELATED WORKS

Shorouq Fathi Eletter and Saad Ghaleb Yaseen [3] developed a model which made use of artificial neural network in order to evaluate the credit applications so that the officials could make decisions in loan issuing factor in the Jordanian commercial banks. This model is based on the multi-layer feed-forward network with back propagation learning algorithm. Evaluation of the proposed model is done by taking into account of various representative cases of loan applications.

Amira Kamil Ibrahim Hassan and Ajith Abraham[4] proposed a loan default prediction model. This model is constructed by three different training algorithms. The prediction model is created with the supervised two-layer feed- forward network in which two attribute filtering functions were used. To train the network, back propagation based learning algorithms were used. German bank datasets were used to train the model. The author made a comparison among the models that have been resulted due to different training algorithms.

Kumar Arun, Garg Ishan, Kaur Sanmeet[5] implemented machine learning algorithms like decision trees,random forest, support vector machines, linear models, neural networks and AdaBoost for predicting the loan. The author has taken the training data set and it is given as the input to the machine learning model. Model training depends on the data set. The new applicant's data is considered as the test data set. The idea of the model is to predict whether the new applicant is eligible for loan or not.

Liyuan Liu and Jennifer Lewis Priestley[6] had done a comparative study of algorithms like regression analysis, neural networks and decision trees for predicting the commercial non financial past-due issues. The performance of algorithms has been examines and it is understood that the decision trees worked better rather than other algorithms.

Featherstone et al[7] obtained the data from the survey that has been conducted in Kansas and Indiana to investigate the lending process in agriculture sector. Certain factors such as borrower, lender, financial and non financial information were the deciding attributes for the loan approval. Tobit models were used to produce a loan approval decision model and Ordinary Least Square (OLS) models were implemented to find out the interest rates so as to inform the borrowers.

Providing loans to individuals is treated as one of the most important functionality in the financial sector. It is a monotonous task to predict the probability of occurrence of defaulters on paying the loan. It helps in decision making process a little but easier so that the right decision could be made on providing the loan based on the customer's request. Anchal Goyal , Ranpreet Kaur[8] developed an ensemble model and the models were compared in order to choose the best model.

Assesing credit risk facilitates the Basel committee on banking supervision is an important task. Angelini et al proposed a capital adequacy framework which made the banks to compute the capital requirement. Eliana Angelini et al[9] implemented artificial neural networks to carry out the model. This model deployed two neural network systems in which one is with a standard feedforward network and the other one with special purpose architecture. The model used the real world dataset based on an Italian small business. It is shown that neural networks have proven to be successful in learning and estimation of default tendency of a borrower after a detailed data analysis.

Aakash Tiwari, Aditya Prakash[11] enhanced the efficiency of J48 machine learning algorithm by examining the capability of ensemble methods. Algorithms like Bagging, Boosting and Blending increase the prejudice between sonar signals that are bounced off in the SONAR dataset. The efficiency of the classifier model is determined by the ranking and standard deviation functionalities.

Several algorithms were used to anticipate the resolution of credit applications and researchers nowadays focused their attention by integrating machine learning algorithms with the predictions. Attributes chosen for generating the algorithms might vary. There might be variation in the weights that have been applied to individual attributes. Abiola Smith et al[12] analyzed credit application data in the credit approval dataset that has been acquired from the archives of the repository of machine learning of University of California , Irvine (UCI). A detailed examination on the selection of attributes has been done. In addition, the data analysis techniques were also analyzed in order to produce the best model for the envisage of the outcome of the credit application. The authors have implemented various visualization, clustering, data reduction and classification techniques and a comparison is also done. The outcome obtained from this model could be the source of information for consumers. This model could be installed so that the automation of the credit application approval process could be performed.

Amr E. Mohamed [13] has presented a comparative study on Machine Learning techniques like Decision Tree, Artificial-Neural Network, K-Nearest-Neighbor, and Support Vector Machine for classification. From this detailed study, it is observed that each technique has been implemented in various areas and different data sets have been used for classification. Accuracy varies from method to method and from dataset to dataset. From the results, it is understood that Support Vector Machine has the largest overall accuracy with 76.3%. The performance of the algorithm depends on the nature of the dataset. Apart from accuracy, there are other metrics that might be considered for decision making.

The intention of Mohammad Aizat bin Basir and Faudziah binti Ahmad's [14] work is to determine the optimal set of attributes and the classification accuracy could be progressed by assuming ensemble rule classifiers method. There are two segments involved in the research process. One phase is to determine the optimal set of attributes and ensemble classifiers method for the classification task. In this approach, 6 datasets were used and the metrics used for measurement were accuracy, rules generation and the number of selected attributes. From the experiments, it is observed that the ensemble rule classifier methods have shown a consistent improvement in classification accuracy on the selected dataset.

Customer's behavior is analyzed for predicting the credibility of loan repayment. Sivasree and Rekha[15] proposed a model which collected the behavior of the customer for the prediction. The outcome of the model is to whether to approve or reject the customer request. The model uses the decision tree data mining technique and using this model, the appropriate attributes were generated for assisting the decision.

Bhumika Gupta et al [16] presented a study on different decision tree algorithms that were implemented in data mining. The efficacy of these algorithms could be evaluated based on the metrics such as accuracy and the attributes used for classification. In addition to it, the time taken for the decision to be finalised is also considered as the metric. In this study, it is found that both CART and C4.5 performed better than ID3 when handling the missing values. The drawback in ID3 is that it is not able to handle the missing or noisy values in the dataset. In spite of this drawback, it is identified that it produced fast results. This work could give a brief idea about the attribute selection measure that could be implemented by several decision algorithms. The attribute selection measure used in the CART algorithm is GINI Index, in ID3 it is information gain and in C4.5 it is gain ratio. Implementation of algorithms depends on the applications and its usage.

Classification is the important function in the data mining. The aim of the classification is to gain knowledge of a classifier from the group of instances with class labels and to assign a class label to the test instance properly. Classification accuracy is measured from the performance of the classifier. Among the various classification algorithms, C4.5 and Naïve Bayes(NB) are considered as important algorithms in solving the classification problems. Liangxiao Jiang and Chaoqun[10] proposed a simple, effective and efficient algorithm based on NB and C4.5. Both algorithms have been evaluated individually at the training time and the class-membership probabilities are averaged based on the classification accuracies on the training data. It is also understood that it performs well and efficient than NBTree.

Data classification plays an imperative role in prediction. Data items that are seen to be similar must be clustered and thus, it aids the decision process to be more capable when considering the large voluminous datasets in bank. During the prediction of loan, the major factor that has to be considered is the risk assessment since the increase and decrease the credit limits in a bank is used to evaluate the risk. The major factor is to identify the good and bad loan applicant's status. Multidimensional risk prediction clustering algorithm proposed by Sudhakar M and Dr. C. V. K Reddy [17] has been deployed in order to identify the status of the loan applicants. Risk assessment is performed in both primary and secondary levels. Association rule is implemented in order to avoid redundancy. Furthermore, the risk percentage is used to identify whether the loan has been sanctioned to the customer or not. From the results, it is shown that the proposed method consumed less time and the accuracy is better when compared with the existing methods.

Somayyeh Zamani and Abdolkarim Mogaddam[18] have done a research on the prediction of marketing operations and a model has been developed which is used to choose the target customers. The classifier used in this model is the decision tree model. Feature selection is done as the first step and this uses the genetic algorithm. Usually among the datasets, two-thirds of customers' data were considered as the training set and one third of the data set as the test set. Genetic algorithm which has been used to choose the features on the classification of customers was a good approach. The model reduces marketing costs, reduce the level of customer satisfaction and the customers relationship is obtained. The advantages of the proposed model are equality attributes and calculation of the decision tree. When a new sample is considered, recalculation of data becomes inefficient since the decision to split the point is improbable. Hence, the proposed system allowed the user to identify a minimum for features before the data calculation takes place.

Hussain AliBekhet and Shorouq Fathi KamelEletter[19] focused their attention in the data mining tools to support credit decisions. The work proposed by the authors implemented two credit scoring models which used data mining techniques in order to support loan decisions for the Jordanian commercial banks. The evaluation of loan application facilitates the credit decision, control the tasks that is carried out in the loan office, time analysis and cost analysis. Credit scoring models are built with the help of the accepted and rejected loan applications from various Jordanian commercial banks. Logistic regression model performed better than the radial basis function model in terms of accuracy. But the radial basis function operated well in determining the default customers.

Abhijit A. Sawant and P. M. Chawan [20] focused on the data mining techniques used for financial data analysis. The authors have studied about loan default risk analysis, Type of scoring and different data mining techniques like Bayes classification, Decision Tree, Boosting, Bagging, Random forest algorithm and other techniques. The authors selected decision trees due to its features like easy discrimination and understanding. An account would be provided to say about the acceptance and rejection of the loan for an applicant. The efficiency of the decision trees is caused by the boosting. Assessment of risk enables the bank to increase the profit and decrease in the interest rate.

A novel credit-scoring model, called vertical bagging decision trees model (abbreviated to VBDTM), has been proposed by Defu Zhang et al[21]. This model worked on the grouping of classifiers based on the predictive attributes. Classifiers are trained with sample subsets in the traditional bagging method. The same set of features were used by the classifiers

and all classifiers have the same set of attributes whereas the model consider all train samples and a part of the attributes. The model is tested on two credit databases and from the results it is understood that the performance of the proposed method outperforms in the prediction accuracy.

A comprehensive analysis of various algorithms that have been deployed is presented in the below table 2.1.

Table 2.1: Comparison of findings of various algorithms

Sl.No	Author & Title	Findings
1	Maher Alaraj, Maysam Abbod, and Ziad Hunaiti[22] " Evaluating Consumer Loans Using Neural Networks Ensembles"	Two new ensemble methods for classification of credit loan customers based on neural networks were proposed. sensitivity of these classifiers with respect to each attributes
2	Defu Zhang et al[21] , "Vertical bagging decision trees model for credit scoring"	Vertical Bagging Decision Trees Model (VBDTM) is implemented. Accuracy is monitored but classifiers use the same set of attributes for classification
3	Peter Martey Addo et al[23]," Credit Risk Analysis Using Machine and Deep Learning Models"	Binary classifiers based on machine and deep learning models on real data in predicting loan default probability has been implemented. It is observed that the tree-based models are more stable than the models based on multilayer artificial neural networks.
4	Bhumika Gupta et al[16]," Analysis of Various Decision Tree Algorithms for Classification in Data Mining"	A study on different decision tree algorithms was done. Attribute selection measure used in in the CART algorithm is GINI Index, in ID3 it is information gain and in C4.5 it is gain ratio
5	Sivasree and Rekha[15] , "Loan Credibility Prediction System Based on Decision Tree Algorithm"	Prediction of attributes for credibility is done. Ranking is performed for all attributes and attribute with high ranking is chosen to be the root node and the attributes with other ranking measure is considered as the other nodes at the below levels.
6	Aditi Kacheria et al[24], Sanctioning Prediction System	A model has been proposed to help the bankers in predicting the credible customers. Algorithm used in this approach is Naïve Bayes algorithm. Before the classification process, the quality of the data has to be improved with the help of K-NN and Binning algorithms. This leads to the improvement in the classification accuracy.

III RESEARCH METHODOLOGY

Several benefits of decision tree are

1. All possible attributes could be chosen and identifying the alternative methods in reaching the conclusion could be traced out in making the decision process a transparent.
2. Definite values could be assigned to problem, decisions and outcomes of each decision could be done.
3. A detailed analysis of the consequences of each possible decision could be implemented. This makes the task of bankers to reach the definite conclusion within a short period of time and without any negotiations in decision.

Decision tree for prediction could be done based on the attribute selection and the decision tree could be implemented as shown in the below Fig 2.

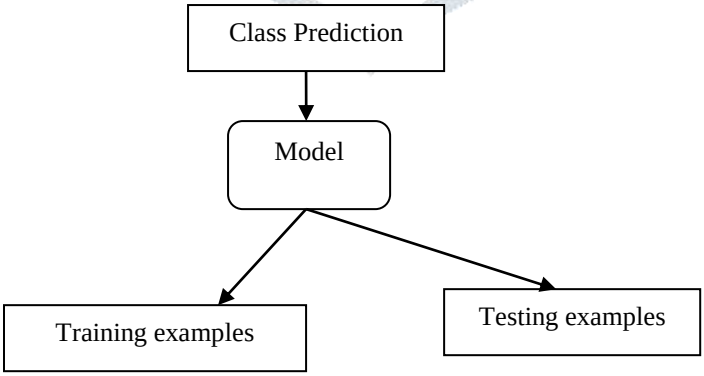

Figure 2. Decision tree model

IV. RESULTS AND DISCUSSION

A comparison of metrics on application of various algorithms have been described in the below table 4.1.

Table 4.1: A comparison of metrics

S.No	Techniques	Accuracy for Test dataset[12]	Accuracy for Training data set[21]	H	Gini
1	Logistic Regression	85.6	79.43	0.27	0.49
2	K-Nearest Neighbors	85.6	80.56	0.32	0.60
3	Linear Discriminant Analysis	76.5	70.26	0.3	0.60
4	Neural Networks	84.8	79.86	0.30	0.60
5	Classification and Regression Trees	87.2	80.62	0.26	0.52

The below graph shows (Fig.3) a comparison of existing research works of various algorithms that has been implemented for the prediction. The accuracy of the test data set is more than the training data set.

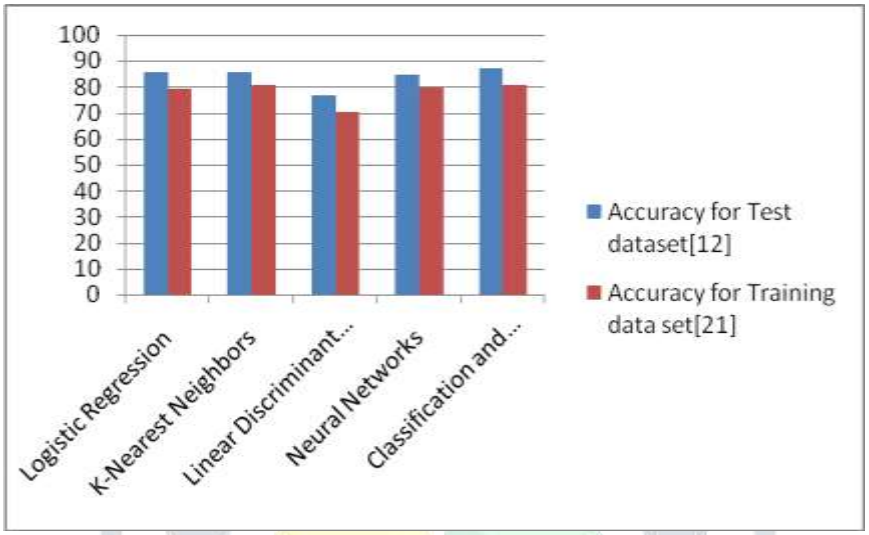

Figure 3. Comparison of accuracy of training data set vs test dataset

From the above comprehensive analysis, it is observed that the decision tree plays an imperative role in the decision making process of financial sectors. The rise of Big Data and data science approaches, such as machine Learning and deep learning models have significant role in the financial sectors. Hence, a thorough and comprehensive analysis of various prediction algorithms has been done and in addition the performance of the prediction algorithms is also monitored. It is important for the bank officials and policy makers to decide on the issuing of the loan. Hence, various prediction algorithms for predicting various operations in the financial sectors have been studied in detail and it is understood that the decision tree algorithms outperforms in terms of accuracy. The accuracy of the system could be made better by the use of decision trees and it could be compared against various classification algorithms in the future.

REFERENCES

[1]. Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI Magazine, 13(3):57.

[2]. Strahan, Philip E. "Borrower risk and the price and nonprice terms of bank loans." FRB of New York Staff Report 90 (1999).

[3]. Shorouq Fathi Eletter and Saad Ghaleb Yaseen, "Applying Neural Networks for Loan Decisions in the Jordanian Commercial Banking System", IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.1, January 2010.

[4]. Amira Kamil Ibrahim Hassan and Ajith Abraham Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks, 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE).

[5]. Kumar Arun, Garg Ishan, Kaur Sanmeet, "Loan Approval Prediction based on Machine Learning Approach", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 79-81.

[6].Liyuan Liu and Jennifer Lewis Priestley , "A Comparison of Machine Learning Algorithms for Prediction of Past Due Service in Commercial Credit", GREY LITERATURE FROM PHD CANDIDATES, DigitalCommons@Kennesa State University.

[7]. Featherstone, A.M., C.A. Wilson, T.L. Kasten and J.D. Jones. —Factors Affecting the Agricultural Loan Decision-Making Process. Agricultural Finance Review. 67(2007): 13-33

[8]. Anchal Goyal , Ranpreet Kaur , "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA) – Volume 3 Issue 1, Jan-Feb 2016
[9]. Eliana Angelini, Giacomo di Tollo,Andrea Roli, "A Neural Network Approach for Credit Risk Evaluation", 2006 Kluwer Academic Publishers.
[10]. Liangxiao Jiang and Chaoqun L, "Scaling Up the Accuracy of Decision-Tree Classifiers: A Naive-Bayes Combination" JOURNAL OF COMPUTERS, VOL. 6, NO. 7, JULY 2011.
[11]. Aakash Tiwari, Aditya Prakash, Improving classification of J48 algorithm using bagging ,boosting and blending ensemble methods on SONAR dataset using WEKA", International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume-2, Issue-9, September 2014.
[12]. Abiola Smith , Brendan Maher , and Deepesh Khaneja, "CREDIT APPROVAL ANALYSIS",
[13]. Amr E. Mohamed," Comparative Study of Four Supervised Machine Learning Techniques for Classification", International Journal of Applied Science and Technology Vol. 7, No. 2, June 2017.
[14]. Mohammad Aizat bin Basir and Faudziah binti Ahmad, "ATTRIBUTE REDUCTION-BASED ENSEMBLE RULE CLASSIFIERS METHOD FOR DATASET CLASSIFICATION" Computer Science & Information Technology (CS & IT).
[15]. Sivasree M S, Rekha Sunny T, "Loan Credibility Prediction System Based on Decision Tree Algorithm", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS090708 www.ijert.org (This work is licensed under a Creative Commons Attribution 4.0 International License.) Vol. 4 Issue 09, September-2015.
[16]. Bhumika Gupta, Aditya Rawat,Akshay Jain, Arpit Arora,Naresh Dhami," Analysis of Various Decision Tree Algorithms for Classification in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 163 – No 8, April 2017.
[17]. Dr. K. Kavitha," Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 2, February 2016.
[18]. Somayyeh Zamani and Abdolkarim Mogaddam, "Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran", Journal UMP Social Sciences and Technology Management, Vol. 3, Issue. 2,2015
[19]. Hussain AliBekhet and Shorouq Fathi KamelEletter , " Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", Review of Development Finance Volume 4, Issue 1, January–March 2014, Pages 20-28.
[20]. Abhijit A. Sawant and P. M. Chawan ,"Study of Data Mining Techniques used for Financial Data Analysis", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 3, May 2013.
[21]. Defu Zhang, Xiyue Zhou ,Stephen C.H Leung , Jiemin Zheng, "Vertical bagging decision trees model for credit scoring", Expert Systems with Applications Volume 37, Issue 12, December 2010, Pages 7838-7843.
[22]. Maher Alaraj, Maysam Abbod, and Ziad Hunaiti," Evaluating Consumer Loans Using Neural Networks Ensembles", International Conference on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014) Jan. 8-9, 2014 Dubai (UAE).
[23]. Peter Martey Addo 1,2,*, Dominique Guegan 2,3,4 and Bertrand Hassani, Credit Risk Analysis Using Machine and Deep Learning Models",article 2018.
[24]. Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta," Loan Sanctioning Prediction System", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-6 Issue-4, September 2016.