International Journal of Engineering & Scientific Research
Vol. 6 Issue 2 , February 2018,
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://esrjournal.com, Email: esrjeditor@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Comparative analysis of classification based data mining algorithms for credit risk analysis

Mrs.Yogita Bhapkar[*], Yashwantrao Mohite College of Arts, Science and Commerce, Pune

Dr.Ajit More[**],MCA Director, BVDU IMED, Pune

Abstract: Data mining is known as Knowledge Discovery in Database. It refers to mining knowledge from large amounts of data. Data mining is the procedure to haul out patterns from different types of datasets. Data mining is purely based on the concepts of artificial intelligence and statistical theory. Data mining has performed extraordinary commitment in learning revelation in variety of application areas. It is popular research area also. Data mining has variety of applications in the fields like Science, Telecommunication ,Information Technology, Biology, Medical science, Banking ,Marketing and many more. This research paper provides focus on data mining application in banking sector. This research paper provides the study of loan applicants by using data mining classification method. Now a days, There are numerous dangers identified with bank loans, for the individuals who get the loans. Number of transactions in banking industries are rapidly growing and huge volumes of databases are available .These databases represent the customers behavior and the risks related with loan .Data Mining is one of the most encouraging and essential area of research with the aim of extracting important information from remarkable amount of database. In this paper we are presenting a comparative study of models developed by using data mining technique classification for classifying loan applications. The model has been built using data from bank. Here performance of Naïve Bayes,J48 and Bagging algorithms are measured. Data mining tool used in this study is WEKA. At the end we have discussed results and performances of algorithms are compared on the basis of accuracy and model development time.

Keywords : Data mining, KDD, classification, Naïve Bayes,J48,Bagging .

1. Introduction:

Database management system gives the contribution for massive gathering of all kinds of datasets. Now a days we are handling tremendous data with variety of transactional databases from business, scientific domain, document reports , military applications, government organizations, banking , finance and many more sectors but for the purpose of decision making it is not just enough to retrieve the data. We search for the patterns from the database by taking efforts but that requires variety of approaches to be applied may be for many years.

Traditional method to obtain knowledge from the database for all the domains are depends on manual analysis and interpretation. The success of traditional data analysis depends on efforts and capabilities of data analyst to read the data, analyze data for interesting findings but it requires to apply intelligence and scanning of database for multiple times .But this approach is time consuming and requires more manpower also it is expensive. As data size is increasing hence sometimes it is unfeasible to use traditional approaches for all the domains.

It has been observe that information about financial business organizations such as financial details of customers is the most valuable assets. Hence we found that there is need for tools and techniques that are capable to store, manage and analyze large volumes of data. However to store, breaking down, such immense measure of information gives numerous obstructions and difficulties means we are drowning in data, and yet ravenous for knowledge.

By considering this necessity implementing the applications using data mining techniques has been emerged. Then We found that the need of data mining has become increasingly growing in all the domains including banking industries. This study covers a complete process of data collection, data preparation, data preprocessing, applying data mining techniques and finally knowledge discovery from the bank database.

2.Knowledge Discovery Process :

The main purpose of using data mining method in this study is to extricate data from bank database and change it into an understandable structure for the future utilize. This involves bank database collection, analysis of data aspects, data preprocessing, model development, finding interestingness metrics, study different accuracy measures, finding time as well as memory complexity and finally data visualization. Using data mining techniques we have predicted the future trends and behaviors of bank customers, which allows banking industries to make loan perditions and take knowledge base decisions. This study also moves data mining beyond the traditional methods for credit risk analysis which are too time consuming and have less accurate results. KDD process contains steps from raw data collection to new knowledge discovery.

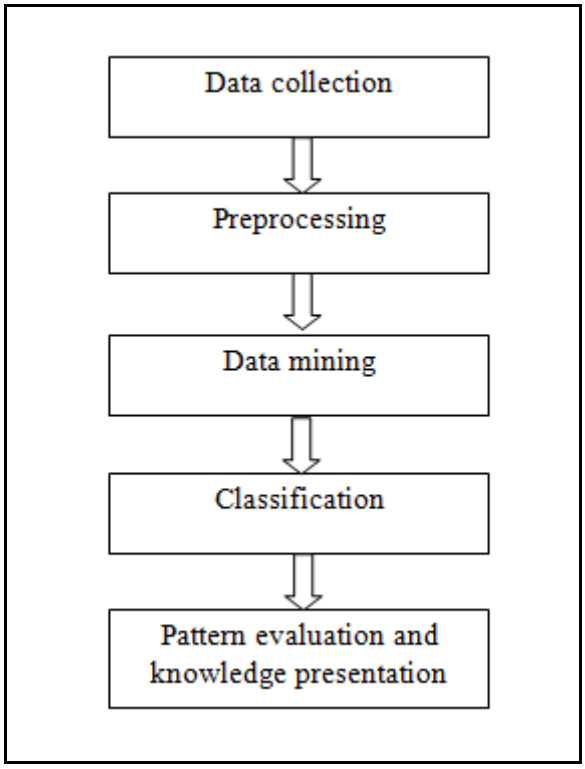

Figure 1 KDD steps

We have performed following iterative steps during this study:

- **Data collection:**

In order to develop bank credit risk analysis model, database of home loan applications is to be created. Database has to be enough and adequate also contains all the variations. Data collection was performed using questionnaire method . For this study we have collected 200 samples of customers applying for home loan from three different banks using questionnaire prepared.

- **Database preprocessing:**

Data preprocessing operations are completed to enhance the quality of credit risk database and to evacuate misrepresentation of data. We need to examine data from bank database for providing loan to the customers so as to improve the quality and reduce distortions. In the data preprocessing stage we have first cleaned our dataset. At this point noisy and irrelevant data present into original dataset are removed,missing values are replaced with real values. Duplicate values present in the dataset are also removed. We identified and corrected the inaccurate, incorrect, irrelevant data and then modify or replace it with the correct values. We have performed data preprocessing on credit risk database using WEKA tool preprocessing algorithms.

- **Data mining:**

At this stage we have applied some statistical techniques as well as machine learning algorithms on bank dataset to extract variety of patterns and to discover knowledge from loan database which are potentially useful and that is apply for decision making.

- **Classification:**

The goal of home loan applications classification is to assign a class label value to test sample. Classification is comprehensively grouped into two kinds: supervised classification and unsupervised classification. In this study we

have used supervised classification methods to develop appropriate model for credit risk analysis for home loan database. Supervised learning is suitable for our study because we known the target values of sample dataset.

- **Pattern evaluation and knowledge representation:**

Knowledge discovery from credit risk database involves discovery of interesting patterns representing knowledge from the bank dataset. KDD involves assessment and interpretation of the patterns retrieved from bank dataset. After performing experiments on credit database we have discovered different loan patterns.

3. Credit risk analysis :

Data mining methods are utilize for credit risk analysis and credit risk management in finance industry. Bank officers must know the clients they are managing are reliable or not. Banks give loan to their clients by checking the details from loan application, different subtle elements like age, salary, loaning rate, reimbursement period, property sold, demography, loan amount, installments, credit card advances and record as a consumer of the borrower. customers with bank for longer periods, with high pay are probably going to get credits effectively. Despite the fact that, there are chances for credit defaulters. Data mining systems recognizes clients who reimburse advances inside a given time constrain from the individuals who don't. Client's unwavering quality can be confirmed by data mining strategies, it examinations the capacity of the client to pay. Using data mining techniques it become easy to decide whether a customer is risky or safe for getting home loan .

4.Database development:

To develop database for credit risk analysis questionnaire was designed. Data is collected from three different banks of type private ,cooperative and nationalize bank. Database of 200 samples with 12 features was ready to carry out the experiments. This database contain last column as a target attribute or class label, it has four class values namely 'Safe', 'Risk', 'More Safe', More Risk'. Complete database description is given below:

No	Attribute name	Description	Data type
1	Age	Customer age	Numeric
2	Gender	Male/Female	Nominal
3	Occupation	Service or business	Nominal
4	Net income	Customers Net income	Numeric
5	Other income	Other income (if any)	Numeric
6	Present lone	Customers current loan(if any)	Numeric
7	Avg. expenses	Customers average expenses	Numeric
8	Loan required	Amount of loan required	Numeric
9	Repayment period	Loan repayment period(in year)	Numeric
10	House cost	Amount of house cost	Numeric
11	Total assets	Customers total assets	Numeric
12	Bank	Type of bank	Nominal
13	Result	Class labels-Safe ,Risk ,More Safe ,More Risk	Nominal

Table 1 Database Description

5.Exploring database using WEKA tool :

First we load our data set on WEKA tool and then we performed series of operations using WEKA preprocessing filters. While as we can also perform all the operations from the command line, for that purpose we use WEKA GUI interface. Following are the visualizations of credit risk database using WEKA tool

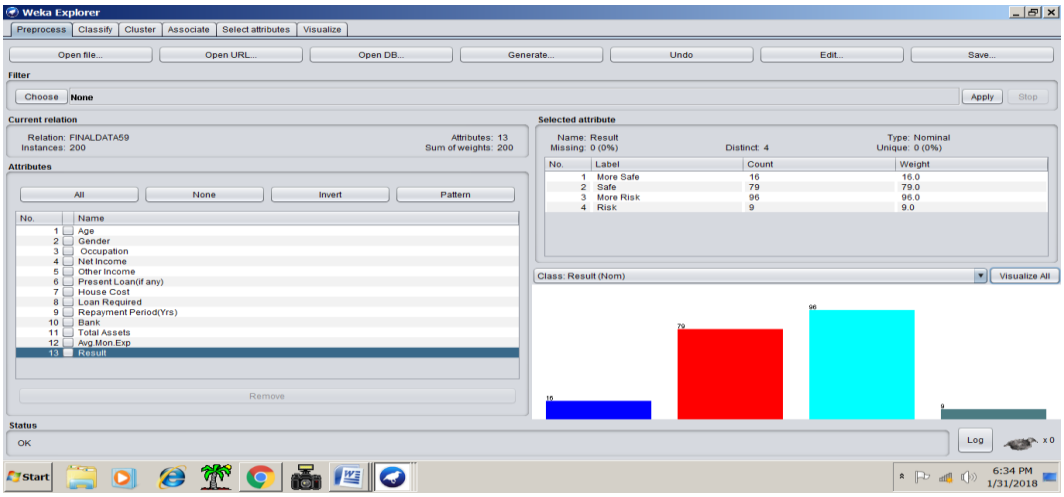

Figure 2 WEKA Database visualization(a)

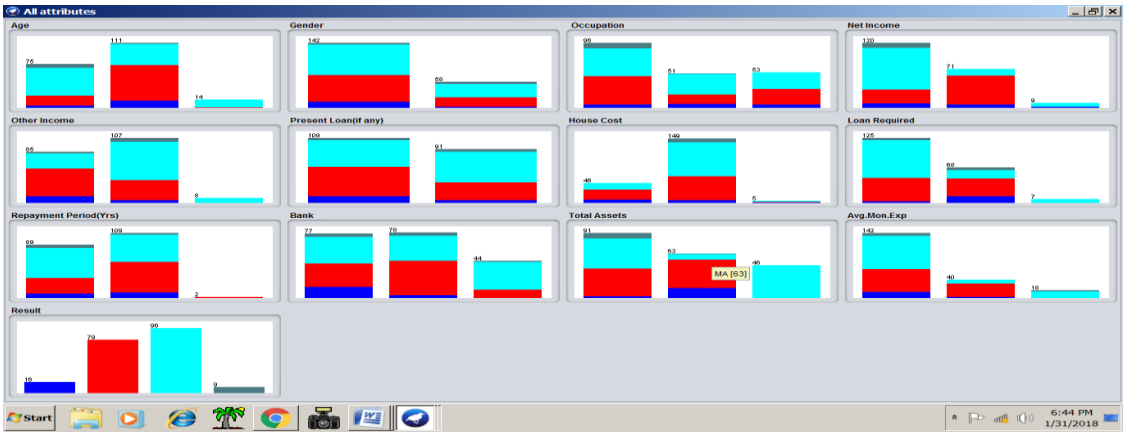

Figure 2 WEKA Database visualization(b)

6. Methodology:

In this research paper we have implemented data mining algorithms for credit risk analysis model development. We have used Naïve Bayes ,J48 and Bootstrap aggregation methods. Classification results of these methods are discussed in the next section.

- **Classification using Naïve Bayes:**

Naïve Bayes is a simple and probabilistic classifier , based on Bayes theorem. It is machine learning classification algorithm. This assumes that the contributions of all attributes within dataset are independent here each attribute contributes equally for the classification. Naïve Bayes is based on Bayes rule which depends on conditional

probability .Using independent attributes conditional probabilities are calculated. Naïve Bayes do the classification by combining the impact of the different attributes on which predictions are measured. This approach is called Naïve because it assumes independence between the various attribute values within dataset.

- **Classification using decision tree:**

Decision tree is a common approach used for classification , tree representation is easy for implementation and also easily understood as compared to other classification algorithms.C4.5 algorithm is one of the popular and most effective decision tree classification algorithm. J48 algorithm is an extension to C4.5 algorithm. Dataset is the input to J48 and decision tree generated is the output. Decision tree consists of root node, intermediate node and leaf node. Nodes are responsible for decision making .Decision tree generates rules for the prediction of the target variable. Using this technique, we have constructed decision tree to model the classification of home loan customers.

- **Classification using Bootstrap aggregation or Bagging:**

Bagging is a simple and powerful classification base on ensemble learning . An ensemble learning is a method which combines the predictions derived from multiple machine learning algorithms .These predictions are more accurate than any individual model's prediction. It is a machine learning ensemble meta classifier which is designed to improve the accuracy and stability of machine learning algorithms .Bagging is an approach of model averaging. Bagging includes training of many classifiers on different subsets of the training data set and results are drawn using the majority voting on the results of all classifiers. Bagging reduces the variance associated with prediction, hence accuracy of prediction using bagging is improved.

7.Classifiers performance evaluation :

Performance of classification algorithm is generally examined by measuring classification accuracy. Traditionally algorithm evaluation is done by using space and time overhead but this is secondary approach. To figure out which is a superior approach is relies upon problem interpretation. Usually accuracy is computed by percentage of tuples put into right class. Reality that the cost is related with a wrong assignments of tuples to the class. We have studied accuracy measures of classifiers like TP Rate, FP Rate, Precision, Recall, F-Measure, correctly classified and incorrectly classified tuples. In this paper we have used cross validation fold testing method where number of folds taken are 10. WEKA tool provides a toolbox of machine learning algorithms. It provides excellent GUI .Recently WEKA tool is recognized as a landmark system for data mining in machine learning

In the next section we have discussed the results obtained after performing the tests on credit database using WEKA tool.

8. Results and analysis:

- **Correctly and incorrectly classified instances by Naïve Bayes algorithm :**

Classifier algorithm	**Correctly predicted**	**Incorrectly predicted**
Naïve Bayes	76%	24%

- **Detailed accuracy of Naïve Bayes algorithm for all four classes:**

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.250	0.011	0.667	0.250	0.364	0.904	More Safe
0.810	0.174	0.753	0.810	0.780	0.897	Safe
0.875	0.212	0.792	0.875	0.832	0.918	More Risk
0.000	0.016	0.000	0.000	0.027	0.857	Risk
0.760	0.172	0.731	0.760	0.737	0.906	Weighted Avg.

- **Correctly and incorrectly classified instances by J48 algorithm :**

Classifier algorithm	Correctly predicted	Incorrectly predicted
J48	75%	25%

- **Detailed accuracy of J48 algorithm for all four classes :**

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.125	0.043	0.200	0.125	0.154	0.746	More Safe
0.823	0.198	0.730	0.823	0.774	0.881	Safe
0.865	0.154	0.838	0.865	0.851	0.928	More Risk
0.000	0.010	0.000	0.000	0.000	0.809	Risk
0.759	0.156	0.707	0.750	0.727	0.890	Weighted Avg.

- **Correctly and incorrectly classified instances by Bagging algorithm :**

Classifier algorithm	Correctly predicted	Incorrectly predicted

Bagging	85.84%	14.15%

- **Detailed accuracy of Bagging algorithm for all four classes:**

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.625	0.010	0.870	0.625	0.727	0.948	More Safe
0.919	0.141	0.850	0.919	0.883	0.922	Safe
0.888	0.081	0.867	0.888	0.877	0.950	More Risk
0.462	0.003	0.857	0.462	0.600	0.923	Risk
0.858	0.102	0.859	0.858	0.854	0.936	Weighted Avg.

Conclusion:

The work presented in this paper has addressed the problem of credit risk analysis. This is the problem of classification of customer's applications applying for home loan to the bank. An approach of classification mainly depends on the nature of the data and features of customers. Since decision of providing home loan for any financial originations plays important role in their economy, home loan application classification process needs to be much efficient and accurate. Present work addressed this problem by a classification approach of data mining that classifies customers into 'Safe', 'More Safe', 'Risk', 'More Risk' these four classes. From the experiments performed on credit database using WEKA tool , we developed models using Naïve Bayes,J48 and Bagging algorithms .We then compared the accuracies and we observed that accuracy of Boot strap ie Bagging is more as compared with Naïve Bayes and J48 .With this we conclude that the classifier developed using Bagging algorithm is best suitable for the credit risk database used in this study.

References:

1. M. Sudhakar, C.V.K. Reddy; 2016;Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique; International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5(3), pp. 705-718, 2016.
2. J. H. Aboobyda, M.A. Tarig ;2016;Developing Prediction Model Of Loan Risk In Banks Using Data Mining Machine Learning and Applications; An International Journal (MLAIJ), vol. 3(1), pp. 1–9, 2016.
3. K. Kavitha; 2016;Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques; International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016.
4. Z. Somayyeh, M. Abdolkarim; 2015;Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran; Jurnal of UMP Social Sciences and Technology Management, vol. 3(2), pp. 307–316, 2015.
5. A.B. Hussain, F.K.E. Shorouq;2014; Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach; Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014.
6. T. Harris; 2013 ;Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions Expert Systems with Applications; vol. 40, pp. 4404–4413, 2013.
7. A. Abhijit, P.M. Chawan; 2013 ;Study of Data Mining Techniques used for Financial Data Analysis; International Journal of Engineering Science and Innovative Technology, vol. 2(3), pp. 503-509, 2013.
8. D. Adnan, D. Dzenana ; 2013;Data Mining Techniques for Credit Risk Assessment Task; in Proceedings of the 4th International Conference on Applied Informatics and Computing Theory (AICT 13), pp. 105-110,2013.
9. G. Francesca; 2012 ;A Discrete-Time Hazard Model for Loans: Some Evidence from Italian Banking System; American Journal of Applied Sciences, 9(9), pp. 1337–1346, 2012.
10. P. Seema, K. Anjali ; 2011;Credit Evaluation Model of Loan Proposals for Indian Banks, World Congress on Information and Communication Technologies; IEEE, pp. 868–873, 2011.

11. E.N. Hamid, N. Ahmad; 2011;A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining; International Journal of Information and Education Technology, vol. 1(2), pp. 150-155, 2011.
12. K. Abbas, Y. Niloofar ;2011;A Proposed Classification of Data Mining Techniques in Credit Scoring; International Conference on Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, p. 416-424,2011.
13. B. Twala ;2010;Multiple classifier application to credit risk assessment; Expert Systems with Applications, vol. 37(4), pp. 3326–3336,2010.
14. N.C. Hsieh, L.P. Hung; 2010;A data driven ensemble classifier for credit scoring analysis; Expert Systems with Applications, vol. 37, pp. 534–545, 2010.
15. Z. Defu, Z. Xiyue, C.H.L. Stephen, Z. Jiemin; 2010 ;Vertical bagging decision trees model for credit scoring; Expert Systems with Applications, vol. 37, pp. 7838-7843, 2010.
16. Chitra. K ,Subashini.B;2013;Data Mining Techniques and its Applications in Banking Sector;International Journal of Emerging Technology and Advanced Engineering,Vol3, Issue 8,2013.
17. Jayasree.V , Vijayalakshmi.R , Balan.S ;2013;A review on data mining in banking sector; American Journal of Applied Sciences, pp.1160- 1165, 2013
18. Bhambri.V;2012;Implementation of data mining in banking sector- a feasibility study; IJRIM, Vol2, Issue 9,2012.
19. Gupta.G ,Aggarwal. H; 2012;Improving Customer Relationship Management Using Data Mining;International Journal of Machine Learning and Computing,Vol.2, pp.874-877,2012.
20. Prasad. U. D ,Madhavi. S; 2012;prediction of churn behavior of bank customers using data mining tools; Business Intelligence Journal, Vol.5, pp.96-101, 2012.