Classification on Internet Banking Adoption Dataset Using WEKA

Nidhi Nigam Verma

Research scholar, Computer Science and Application, Dr. APJ Abdul Kalam University, Indore, India

Abstract: ***Data mining is a process of extracting usable information or knowledge from large amounts of data. For the banking industry, data mining, its importance, and its techniques are vital because it helps to extract useful information from a large amount of historical data which enable to make useful decisions. In banking, industry customers are becoming more and more demanding in terms of the level of service, responsiveness, costs, quality of products etc. Data Mining is very useful for banking sector for better acquiring and targeting new customers and helps to analyze customers and their transaction behaviours. Banks could only reach to these goals furthermore if they take initiative to invest in new technologies. In the recent era, a new technology that has achieved considerable attention, especially among banks, is Internet banking. Its large scope of applications, its advantages brings an immoderate change in a common human's life. Classification is one of the most commonly used and applied data mining techniques. It is applied where the cases of data in the volume are known. In this paper the J48 algorithm for classification is applied on internet banking adoption dataset in order to predict group membership for data elements in a dataset. The analysis here is done with the help of WEKA tool for data mining.***

Keywords: Classification, Data mining, WEKA, J48 Algorithm for classification, Internet Banking Adoption

I. INTRODUCTION

Data Mining is an analytic process designed to explore large amounts of data. The most common and widely used techniques which are used for data mining and data recovery operations includes Association rule mining, Classification algorithm, Clustering, Sequential Patterns ,Regression and Decision Trees. Classification is a most commonly used data mining technique to predict group membership for data elements in a dataset. Popular classification techniques which are used include decision trees and neural networks [27]. Classification is a data mining technique which is used mostly in order to defend from fraud and escape from risk of credit [25].In classification technique of data mining data accuracy is very important. Especially in classification method test data are utilized to anticipate the accuracy of classification. If that accuracy is acceptable then only that value can be applied to new data set.

It builds up a model from the sampled data items with class labels known and also utilizes this model for prediction of the class of objects for the population whose classes are not known yet. Each row in the database contains one or more predicting values or attributes which finds out the predicted class label of that row according to the model constructed. These models are developed using a neural network or decision tree model [26]. On the other hand prediction models employ with continuous valued functions. It is mostly used for prediction of unavailable or missing numerical data values from the sampled values of attributes. Banks use classification technique to categorize information as a target for prediction. Also techniques can be utilized to develop models that can easily predict default risk levels. Credit risk and fraud detection applications are mostly well suited to this kind of analysis. This approach often employs neural network or decision tree based classification algorithms.

Internet Banking is the most used feature by the citizens of India after the effect of demonetization. This feature is assumed to be one of the most flexible, adaptable and secure ways of transacting among the users or customers to bank. However, it depends on the trust that an individual has on the bank he/she is operating with. One of the most popular techniques of data mining used by banking industry is the classification. The main aim of this paper is to apply J48 algorithm for classification on internet banking adoption dataset with the help of WEKA tool.

.

II. METHODOLOGY

Banking customers having accounts in different bank branches located in various cities of Madhya Pradesh was the target population for this research study. Basically, there are two types of quantitative research methods experiment and survey. In this study specifically survey method is used where the data is collected by means of a questionnaire to determine the opinion of a target population of Madhya Pradesh [22].

A. Primary Data for the Research

Primary data for this has been collected using a self structured questionnaire designed purposely for this study. Appropriate secondary sources have as well been relied upon for designing a suitable comprehensive questionnaire to gain deeper insights in this field. It was specially prepared for the sample population in order to explore the most important factors for adoption of internet banking services and challenges associated with it. Questionnaire also includes questions regarding the satisfaction level of the customers using internet banking on identified factors. For this study data is collected using Google forms and through emailing of the questionnaires to users. Data for the research has been collected from 545 customers which includes both users and non-users of Internet banking. The collected data should to be analysed by using the appropriate analytical tool or technique in order to understand the various factors and reasons behind Internet banking adoption. For this research

study, WEKA tool, a data mining tool, is being used.

B. WEKA tool for Data Mining

"WEKA" stands for Waikato Environment for Knowledge Analysis. Basically WEKA is named subsequent to a flightless bird of New Zealand. It is a set of various machine learning algorithms that can easily be applied on any data set directly or can be called from your Java code. WEKA basically contains various tools for data mining and data pre-processing which are clustering, classification, association rules, regression and visualization [23].It is freely available and an open source software for data mining and its applications under GNU general public license, which is developed by the university of Waikato in New Zealand.

C. Sampling Procedure

A sampling, defines the population from which our research sample is drawn. As there is hardly enough money or time to collect information from everyone in overall population, the goal becomes choosing a representative sample, sometimes called as a subset of that population. Area for this research is Madhya Pradesh, India. Madhya Pradesh is situated at the center part of India, so known as the heart of India. For this research sample which is considered are from major cities of Madhya Pradesh. Research Questionnaire was sent to them online, their opinions and concerns are collected in order to measure the adoption rate of internet banking in Madhya Pradesh.

III. EXPERIMENTAL WORK

a. Primary Data Collection

For this study data is collected using Google forms and through emailing of the questionnaires to users. Only 502 responses were totally complete and found eligible for our analysis. So for this study, a sample of 500 responses was considered.

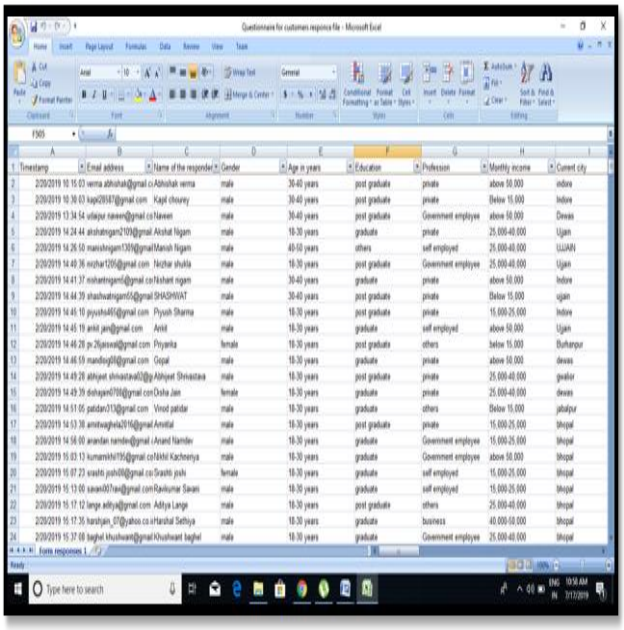

Fig1: Questionnaire for customer's response file containing 502 responses

b. Analysis of dataset using WEKA

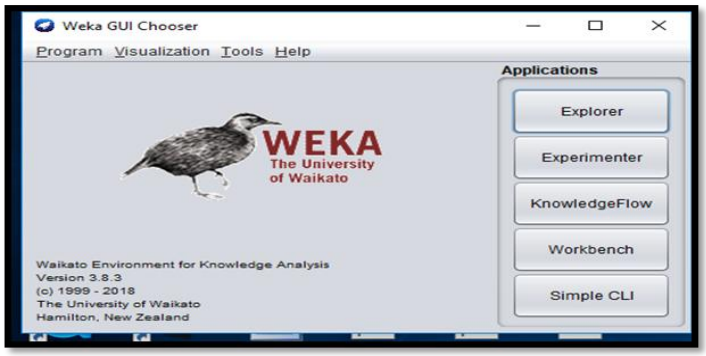

Fig2: Weka first screen while startup

c. Loading Dataset into WEKA Explorer

For this study we are having a dataset on internet banking adoption, which we have collected online through a well structured questionnaire using Google forms. We have received 502 responses which we are going to consider here for the purpose of analysis using WEKA [24]. Here we import our data file which is CSV format and loading it is converted in ARFF file format with the help of save button in WEKA explorer.

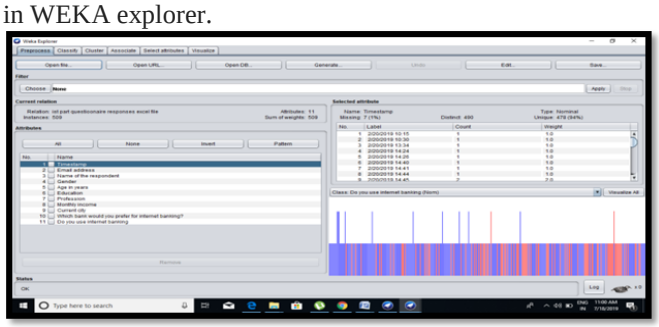

Fig3: Internet Banking Dataset in WEKA Explorer

d. Classification on internet banking dataset

For Classification, training dataset is used basically to develop a model which classifies the test data samples into known classes. Data classification is classic task in data mining field. There are so many classifiers that are being used to classify the data which includes Naive bayes, functions, lazy, Meta, misc, rules based and J48 tree etc.

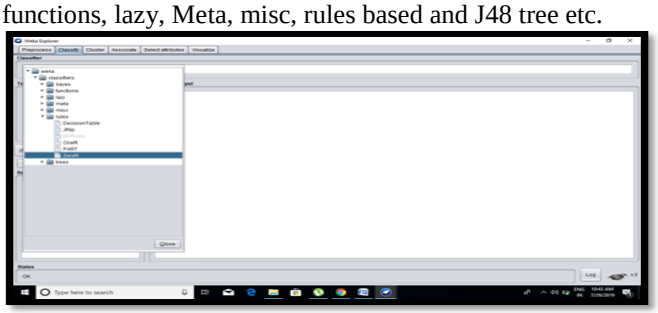

Fig 4: Classify panel in WEKA 3.8 with classifiers

Out of so many classifiers, here we are using J48 tree to classify internet banking adoption dataset.

J48 tree

This algorithm basically generates the rules for predicting the target variable. With the use of the tree classification the

decisive distribution of data set is very easily understandable. The WEKA tool also offers so many options related with tree pruning, which can be used as a powerful tool for précising. The classification is performed in a recursive manner till each and every single leaf is pure, that means the classification of the dataset must be as perfect as possible. It develops the decision tree using labeled training data set. To take the decision, the attribute with highest normalized information gain is chosen. After that the algorithm begins recursion on other level of smaller subsets. The splitting procedure continues until all the instances of a subset belong to the same class.

Basic Steps of J48 Algorithm:

- If all the instances are of the same class, the tree only represents by a leaf, hence the leaf is returned by labeling with that class name.
- The probable information is calculated for each and every attribute, provided by a test on that attribute. Then the information gain is calculated which would result from a test on the attribute.
- Finally on the basis of the current selection criterion the best attribute is found, which is then selected for branching.

e. **Analyzing results of J48 decision tree algorithm:**

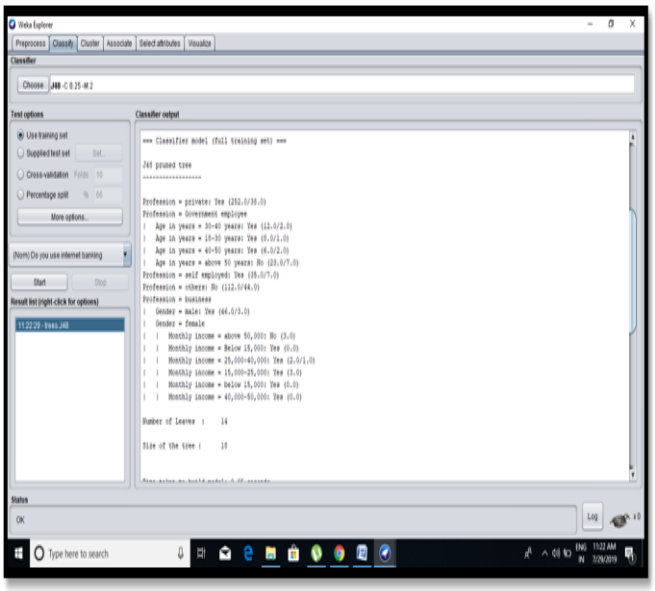

Fig 5: Classification results

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

Profession = private: Yes (252.0/38.0)

Profession = Government employee

| Age in years = 30-40 years: Yes (12.0/2.0)

| Age in years = 18-30 years: Yes (8.0/1.0)

| Age in years = 40-50 years: Yes (6.0/2.0)

| Age in years = above 50 years: No (23.0/7.0)

Profession = self employed: Yes (35.0/7.0)

Profession = others: No (112.0/44.0)

Profession = business

| Gender = male: Yes (46.0/3.0)

| Gender = female

| | Monthly income = above 50,000: No (3.0)

| | Monthly income = Below 15,000: Yes (0.0)

| | Monthly income = 25,000-40,000: Yes (2.0/1.0)

| | Monthly income = 15,000-25,000: Yes (3.0)

| | Monthly income = below 15,000: Yes (0.0)

| | Monthly income = 40,000-50,000: Yes (0.0)

Number of Leaves : 14

Size of the tree : 18

Time taken to build model: 0.05 seconds

The Classifier Model part shown in the above table is basically an evaluation on the training data and it illustrate the model as a decision tree and provides some necessary information about the tree, size of the tree, number of leaves in a tree and total time taken to build that model. Classifier model is basically a textual representation of pruned decision tree that was developed on the full training data. In the above result table you can see, the very first split is on the 'profession' attribute, and on the second level, the respective splits are on 'Private', 'Government Employee', 'Self employed', 'Others' and 'Business'.

In the tree textual form, a colon represents a class label which has been assigned to a specific leaf and followed by the total number of instances that reaches to that leaf. Number of leaves here are 14, and the size of the tree is 18. The program gives a time it took to build the model, which is 0.05 seconds.

=== Evaluation on training set ===

Time taken to test model on training data: 0.05 seconds

=== Summary ===

Correctly Classified Instances	397	79.0837 %
Incorrectly Classified Instances	105	20.9163 %
Kappa statistic	0.4789	
Mean absolute error	0.3053	
Root mean squared error	0.3907	
Relative absolute error	75.5064 %	
Root relative squared error	86.9353 %	
Total Number of Instances	502	

Evaluation on training set provides estimated values of the tree's predictive performance which is provided by WEKA's evaluation model. It results the list of statistical values which shows how exactly the classifier was capable to predict the actual class of the attribute instances under the selected test module. In summary part of the classifier output, you can see that the success rate is 79.0837% of the training data and mean absolute error is 0.3053. The reason behind the errors here are not 0 or 1 is just because not all the instances of the training set are classified correctly. By examining this part of the classifier output you can find how successful this particular model is, and other statistical results of that model along with total number of instances.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.859	0.383	0.852	0.859	0.855	0.479	0.768	0.857	Yes
0.617	0.141	0.630	0.617	0.624	0.479	0.768	0.546	No
0.791	0.315	0.790	0.791	0.790	0.479	0.768	0.770	**weighted avg**

WEKA makes you to visualize a graphical representation of the classified tree, as the result of the J48 algorithm is the Decision tree, we can see it in graphical form by selecting visualize tree option. The tree structure consists of one root node, intermediate lower level nodes and leaf nodes.

Each and every node in the decision tree carry a decision and that decision lead to final result. Decision tree partition the total input space of a data set into mutual exclusive areas, where each partitioned area having one label, an action or a value to illustrate its data points. Splitting criterion of a decision tree is basically used to determine the best attribute to split the tree of the training dataset that would reach to a particular node. Below Fig26 shows the decision tree, which is a result of using J48 for internet banking adoption data set to find whether one using internet banking and its services or not.

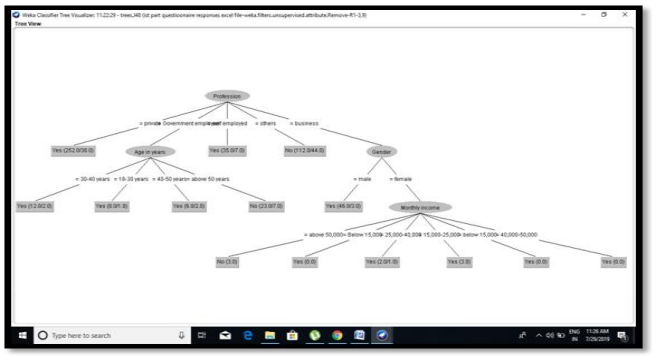

Fig 6: Tree Visualization of the result of the J48 classification algorithm

For each leaf belongs to tree, you will see total number of instances that falls into that particular leaf, and the total number of inaccurately classified instances of that leaf, which are both from the training set. In other words we can say that in each leaf the first number is basically the number of instances (total weight of instances) reaching to the leaf and the second number is the total number (total weight) of those instances which are wrongly classified.

IV. CONCLUSION

Here in this paper we discussed the results of applying J48 algorithm of classification on internet banking adoption. Through this paper we have shown how classification is useful for predication and can helps to sort out problems regarding internet banking adoption in Madhya Pradesh. J48 algorithm for classification basically generates the rules for predicting the target variable. The result of J48 algorithm is a decision tree, with the use of this tree classification the decisive distribution of data set is very easily understandable. These results will be useful for bank managers and other banking authorities, in order to enhance the internet banking adoption rate in the future.

REFERENCES

[1] Dr. Md. R. Farooqi and N. Iqbal, "Effectiveness of Data mining in Banking Industry: An empirical study", international Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017,pp. 827-830.

[2]S.Pulakkazh and R.V.S. Balan," Data Mining In Banking and its applications- A review",
Journal of Computer Science, Vol. 9 Issue 10, 2013, pp. 252-1259.

[3]M.Dash and S.K. Nayak," Data Mining Need In Profiling Internet
Banking Customer" International Journal of Computer Science and Management Research, 2012

[4]K. Rangra and Dr. K. L. Bansal," Comparative Study of Data Mining Tools ",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014,pp. 216-223.

[5]S. Raj,"A Review Of Data Mining Applications In Banking ",International Journal of Computer Science and Management Research, June 2015.

[6]V. Jayasree and R.V. Siva Balan, "A Review on Data Mining in Banking Sector", American Journal of Applied Sciences, Vol. 10 Issue 10, 2013, pp. 1160-1165,

[7]Dr.C. Rajabhushanam and S. cumaar, "An Study Of Data Mining Applications In Banking", International Journal of Pure and Applied Mathematics, Volume 116 No. 15, 2017, pp. 265-271.

[8]Mr. N.K. Dokania and Ms. N.Kaur," Comparative Study Of Various Techniques In Data Mining", International Journal Of Engineering Sciences & Research Technology, Vol. 7 Issue 5, May 2018, pp. 202-209.

[9]W. Niyagas, A Srivihok, and S Kitisin, Non-members," Clustering e-Banking Customer using Data Mining and Marketing Segmentation", ECTI Transactions On Computer And Information Technology Vol.2, No.1 May 2006,Pp. 63-69.

[10] K.I. Moin, Dr. Q.B. Ahmed, "Use of Data Mining in Banking", International Journal of Engineering Research and Applications (IJERA) ,Vol. 2, Issue 2,Mar-Apr 2012, pp.738-742.

[11]V. Bhambri, "Application of Data Mining in Banking Sector", International journal of computer science and technology, Vol. 2, Issue 2, June 2011, pp. 199-202.

[12]V. Aggelis ,"Association rules model of e-banking services ",Journal of Internet banking and Commerce, 2004 ,pp. 167-173.

[13] F. Halili and A Rustemi," Predictive Modeling: Data Mining Regression Technique Applied in a Prototype", International Journal of Computer Science and Mobile Computing, Vol. 5, Issue. 8, August 2016, pp. 207– 215.

[14] S. Jindal," Study of E-Banking Scenario in India", International Journal of Science and Research (IJSR), Volume.5 Issue.12, December 2016, pp. 680-683.

[15] S.N. Hojjati,A. Monavvarian and F.S. Roozbahani," Evaluation of Factors Influencing Internet Banking Acceptance by Internet Users in Iran", International Journal of Advanced Networking and Applications, Volume.6 Issue.6 , 2015, pp. 2503-2519.

[16] G.A. Weshah," The role of internet banking in continuous improvement areas: Quantitative evidence from Jordanian banks" International Journal of Business Performance Management, January 2013, Vol.14, No.2, pp.181- 196.

[17]R. Farooqi," Impact of Internet Banking Service Quality On Customer Satisfaction", Journal of Internet Banking and Com, April 2017, vol. 22 Issue No. 1, pp. 1-17.

[18] P. Singh and D. Kumar," Impact of Internet Banking on Customer", International Journal of Research (IJR), Vol. 1 issue.4, May 2014, pp. 394-412.

[19] S. Haq and B.M. Khan," E-Banking Challenges And Opportunities In The Indian Banking Sector", Innovative Journal Of Business And Management, vol. 2 issue.4,August 2013, pp. 56 – 59.

[20] K. Nidhi," E-Banking In India: Challenges And Opportunities "International journal of Science, technology and management, Vol. No. 05, Issue no. 08, August 2016, pp. 809-815.

[21] S. Devulapalli and S. k. Orugant, "Challenges and Opportunities of E- Banking in India", OSR Journal of Business and Management (IOSR,JBM) ,pp 56-61.

[22]Dr. G. Sharma and dr. S. malviya," Exploring Usage Pattern of Internet Banking Services Provided By Banks in Indore District of Madhya Pradesh", international journal of advance research in computer science and management studies, Volume 2, Issue 11, November 2014.

[23]S. Palaniappan, A. Mustapha, Md. C.F. Foozy, and R. Atan," Customer Profiling using ClassificationApproachfor Bank Telemarketing", Vol. 4 Issue 2, July 2016

[24]K. Kaur1 and S. Dhiman, "Review of Data Mining with Weka Tool", Vol. 4 Issue. 8, August 2016

[25]Dr. Md. R. Farooqi and N. Iqbal, "Effectiveness of Data mining in Banking Industry: An empirical study", international Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017,pp. 827-830.

[26]Dr.C. Rajabhushanam and S. cumaar, "An Study Of Data Mining Applications In Banking", International Journal of Pure and Applied Mathematics, Volume 116 No. 15, 2017, pp. 265-271.

[27]Mr. N.K. Dokania and Ms. N.Kaur," Comparative Study Of Various Techniques In Data Mining", International Journal Of Engineering Sciences & Research Technology, Vol. 7 Issue 5, May 2018, pp. 202-209.

BIOGRAPHY

Nidhi Nigam Verma, Research scholar, Computer science and application, Dr. APJ. Abdul kalam University Indore (M.P), having 8 years of experience in teaching, worked as a assistant professor, computer department in Indore.

DR Deepika Pathak , Department of computer science and application, Dr. APJ. Abdul Kalam University, Indore (M.P), Pro vice chancellor in Dr. APJ Abdul Kalam University, Indore.