International Journal of Management, Accounting and Economics
Vol. 3, No. 8, August, 2016
ISSN 2383-2126 (Online)

www.ijmae.com

Classification of Bank Customers by Data Mining: a Case Study of Mellat Bank branches in Shiraz

Dariush Farid
Associate Professor, Faculty of Economics, Management and Accounting, University of Yazd, Yazd, Iran

Hojjatollah Sadeghi
Assistant Professor, Faculty of Economics, Management and Accounting, University of Yazd, Yazd, Iran

Elahe Hajigol
Industrial PhD, University of Yazd, Yazd, Iran

Nadiya Zarmehr Parirooy[1]
Master of Business Administration Financial trends Yazd University, Yazd, Iran

Abstract

This research predicts through studying significant factors in customer relationship management and applying data mining in bank. Financial institutions and other firms in competitive market need to follow proper understanding of customer behavior. Customers' data are analyzed to identify specific opportunities and investment, to classify and predict the behaviors; further, data are eventually used for decision-making. Therefore, data mining as knowledge exploring (discovery) approach plays a significant role through a variety of algorithms. This study classifies bank customers by using decision tree algorithm. Three decision tree models including ID3, C4.5, and CART were applied for classifying and finally for prediction. Results of simple sampling method and k-fold cross validation show that forecast accuracy of C4.5 decision tree using simple sampling was higher than other models. Thus, predicting customers' behavior through C4.5 decision tree was considered the ideal prediction for bank.

Keywords: Validation, data mining, decision tree, customer relationship management.

Cite this article: Farid, D., Sadeghi, H., Hajigol, E., & Parirooy, N. Z. (2016). Classification of Bank Customers by Data Mining: a Case Study of Mellat Bank branches in Shiraz. *International Journal of Management, Accounting and Economics, 3(8)*, 534-543.

[1] *Corresponding author's email: nadiya.zarmehr@yahoo.com*

Introduction

For decades, financial institutes followed the approaches focused on production and transaction. With the growth of technology and developing competitive factors, enterprises were increasingly required to establish and maintain effective customer relationship. In competitive market with other banks and financial institutions, the banks will have to seek for proper knowledge of their customers. Purpose of customer identification is to distinguish and detect more high-value customers; to maintain and attract valuable customers. IT growth and delivering electronic banking services, developing marketing activities and customer relationship management increase the possibility of turning high-value customers away. The point is highlighted when it is found out that the cost of attracting new customer is almost five times the cost of maintaining old customers (Norouzi, 2009).

Theoretical basics

Customer relationship management is a strategic system for collecting customers' business needs and behaviors so that it establishes more strong relationships. Eventually, strong customer relationship is the key to business success (Nguyen, 2007). Customer relationship management as a process helps us to collect customer hidden data, sale, effectiveness of marketing activities, fast customer responding, as well as market tendency (Hicks, 2006).

Customer relationship management (CRM) processes

- Awareness and knowledge discovery
- Market planning
- Customer interaction
- Feedback analysis

In knowledge discovery phase, customers' data are analyzed to identify specific opportunities and investment and to classify and predict the behaviors; finally, the data are used for decision-making. In market planning, customers' tendencies as well as distribution channels aid the organization to enable strategic communication development activities and customer knowledge orientation (Toy Yeen, 1998).

CRM functional features

According to studies conducted on customer relationship management, functional features are as follows (Pepard, 2000):

- Focus on strengthening closer and deeper customer relationship
- Greater benefit of current (existing) customers to new customers
- Analyzing customer data for decision-making

- Effective customer relationships based on data transformed into information
- Person-to-person marketing and database marketing

CRM requirements

Considering customer significance as any firm's critical basics, the requirements of applying CRM in a firm, regarding to firm activities' development and complexity, are described as follows (Smith, 2006):

- Improve services
- Customer satisfaction
- Reduce costs
- Person-to-person relationship (even with millions of customers)

Data mining notion

Regarding highly-varied clients, customers, markets, variety and complexity of services and business environments, as well as the need for accessing to proper information for accurate and timely decision making, it is critically necessary for firms to find and classify functional and effective information of huge data; in other word, it is an expertise and art (skill). Indeed, data mining responds this need of firms and institutions. The higher the data and the more complicated the relationships, the harder the access to information contained in data; therefore, the role of data mining as knowledge discovery approach will be more evident.

Regional and global competitive markets require institutions creating competitive advantages in production, service delivery, increased satisfaction and higher customer attraction so that it may achieve a better place (in customer perspective) and develop the firm (Shahrabi, 2007).

Data mining stages

Data mining stages are described in the following (Deshpande, 2010):

1. Data selection: data related to analysis and decision making are separated from other data.
2. Information pre-processing: data processing, cleaning, and integration.
3. Data conversion: selected data are properly converted for data mining.
4. Data mining: potentially useful patterns are extracted through intelligent methods; decision is adopted for these methods.

5. Interpretation and evaluation: interesting patterns, at this stage, representing knowledge are identified based on adopted measures; the discovered knowledge is provided to the user.

Decision tree algorithm

Decision tree in data mining is a model used for showing the classifiers and regressions. It consists of some nodes and branches. In the classifying decision tree, leaves indicate classes. Decision is made according to one or more particular traits in other nodes (non-leaf nodes).

Decision tree is a popular technique in data mining as it is simple and understandable. In other word, decision tree describes all contents alone independent of an expert to interpret the output. In fact, this is a graphic method and may be simpler than other classifying methods due to this interpretation. However, large numbers of nodes may make decision tree graphic representation difficult (Ismaeili, 2012).

Data mining functions

- Data mining and banking
- Data mining and stores
- Data mining and telecommunications
- Data mining and bioinformatics
- Etc.

Literature review

Yang (2013) used apriori algorithm to improve bank customers' segmentation. Experimental results showed that this algorithm may effectively overcome conventional algorithm, which results in precise customer segmentation, more reasonable results, effective improving of decision making as well as higher benefits for bank (Yang, 2013).

Bhapkar, in 2014, conducted a study using a series of data including positive and negative cases of loan depositors. Considering all banks, customers' demographic features such as age, sex, occupation, annual income, total assets, and other income, etc., were collected and the customers were classified. According to this study, it concluded that data mining techniques are efficient for customers' analysis including banks. As a result, the customers were classified in to four: secure, highly secure, risk taking, highly risk taking (Bhapkar, 2014).

Wang et al (2014) introduced a research focused on hierarchy-based customers fuzzy clustering by optimizing logistic network. The results revealed the proposed approach better for solving customer cluster problem comparing the three other dominant algorithms (Wang, 2014).

Oscar et al (2015) merged the two approaches of clustering and data envelopment analysis (DEA) and through which identified management clusters in bank branches and reviewed productivity performance. The considered variables were divided into input and output. Inputs included three variables; while, outputs were four. Thus, 78 branches were placed in 4 clusters. Finally, productivity performance was investigated. The result of this research was effective for designing a bank network since it aids in making decision of whether to found or roll up a branch (Oscar, 2015).

Kim et al (2015), in a research, studied the effect of customers earning performance on the terms of precious and non-precious loans through using a volume sample of 3725 loan facilities within 1995-2011. There were three main hypotheses and the results obtained using regression. The results of this study show that customer earning performance significantly contributes in loan contracts (Kim, 2015).

Conceptual model

Regarding the significance of customer classification in banking system and conducted studies, data sources are thoroughly studied by mathematical and statistical methods and the proposed data mining approach is described. In general, the dataset for customers' classification in banking system is introduced; in addition to data cleaning and preprocessing, data will be prepared for evaluation; and finally, the proposed method is thoroughly investigated.

The dataset is studied here to apply the suggested method and to analyze the results. Modeling is conducted according to the information collected through Mellat Bank in Shiraz and the customers are classified. 10 variables of the initial 20 variables were selected interviewing 9 banking experts and frequent reviewing. Of the excluded variables is customer "credit rating", for instance; banking system evaluates this rating based on existing data. This was of the most important variables; however, due to bank security system, giving information of credit rating was ignored. Totally, this dataset offers information such as customer income, current bank debt, education level, sex, marital status, age, and the years of interacting with bank.

The main purpose of preprocessing is to determine data quality. There are four main operations in preprocessing: data cleaning, data integrating, data reduction, Transformation and Discretization. This research cleans unused data and removes the samples lacking adequate information. Therefore, information volume is reduced by data preprocessing and efficient data are provided; as a result, classification is imposed by better speed and higher accuracy. Regarding the nature of required data, some features are excluded in order to have less processing in more proper time for identifying authentic customers.

Research method

Data preparation

Since data are provided to the researcher in paper files, some unrelated data are also seen. Therefore, in some cases, empty features were excluded by data mining according

to experts. Then, data were prepared in excel files and transformed into acceptable formats of data mining tools like rapidminer.

Sampling of data sources

In data mining studies, in general, and classification researches, in particular, main data are divided in to two parts as follows:

- Train data: are used for training classification algorithm such as K-nearest neighbor algorithm, decision tree, and neural network, etc. It is necessary to separate about 70-80% of research all data sets as train data. Hence, bank customers' classification model merely use train data for learning. It is worth to notify that test data are absent in train data.

- Test data: around 20-30% of data are attributed to test data. These data are used for evaluating bank customers' classification algorithm accuracy.

Therefore, to train and evaluate the proposed model of customers' classification, 70% of data were applied as train data and 30% as test data.

Validation

This method is briefly called k-fold cross validation, which is sometimes referred as displacement estimation. This evaluation method is the results of a dataset statistical analysis determining to what extent it is generalizable and independent of train data. This technique is particularly used for prediction functions in order to determine to what extent the model will be useful in practice. The method is practical where data gathering is difficult, costly and or impossible.

Data, in such validation, are partitioned into k subsets. Of these k subsets, one is applied for validation and k-1 for training. The procedure is repeated k times and all data are exactly used once for training and once for validation. Finally, average results of k-time validation are selected as final estimation. K is usually 10, which is also true for this research.

Classification using decision trees

In general, decision tree is like a flowchart. It consists of two nodes. Ending node, which is called leaf and other nodes are non-ending node testing features. Each branch or edge is a test result.

There are various algorithms for making decision tree including: ID3, C4.5, CART[1]. These algorithms apply a greedy approach; further, it follows top-down strategy and division.

[1] Classification and regression tree

ID3 decision tree analysis

The best feature in ID3 decision tree is selected based on a particular criterion and is used as tree node. Next, the related branch is considered for each value of the selected feature and each class leaf is determined. If all samples are totally classifies, the tree will stop; otherwise, it would be repeated for new characteristics as node, branch and leaf.

Information Gain is the criterion used for ID3 decision tree selecting features of various tree levels. The best feature (variable) is initially selected using this criterion. The selection is based on maximum value of applied criterion.

C4.5 decision tree analysis

C4.5 decision tree also employs a particular criterion, like ID3 decision tree, referred as Gain Ration. Information Gain criterion leads to better results in large number of variables; whereas, C4.5 decision tree (ID3 successor) benefits Gain Ratio criterion dealing with this problem. Hence, this new criterion is normalizing the pervious decision tree criterion.

CART decision tree analysis

This algorithm utilizes Gini Index criterion. Gini index provides binary division for each variable.

Research findings

Measurement accuracy of the three decision tree algorithms are initially represented in the following table.

Table 1 Overall results of decision tree measurement accuracy

	Percentage of validation	Percentage of X-validation
ID3	81.67	79.50
C4.5	83.33	82
CART	80	75.50

According to the above table and obtained results, it is concluded that maximum accuracy is assigned to C4.5 decision tree. 70% and 30% of data are train and test data, respectively using simple sampling method. Therefore, of the three algorithms, C4.5 decision tree is introduced as the best predictive model.

Results of prediction accuracy of all three trees can be assessed together since the same results obtained. According to the obtained results, prediction accuracy of all three tress was higher using simple sampling method than validation method; moreover, C4.5 decision tree of the three algorithms showed higher prediction precision. Thus, C4.5 decision tree was used to provide the model and to better predict customers. Therefore, the prediction by this algorithm is more precise comparing other algorithms. Tree's final form is illustrated as follows:

International Journal of Management, Accounting and Economics
Vol. 3, No. 8, August, 2016
ISSN 2383-2126 (Online)

www.ijmae.com

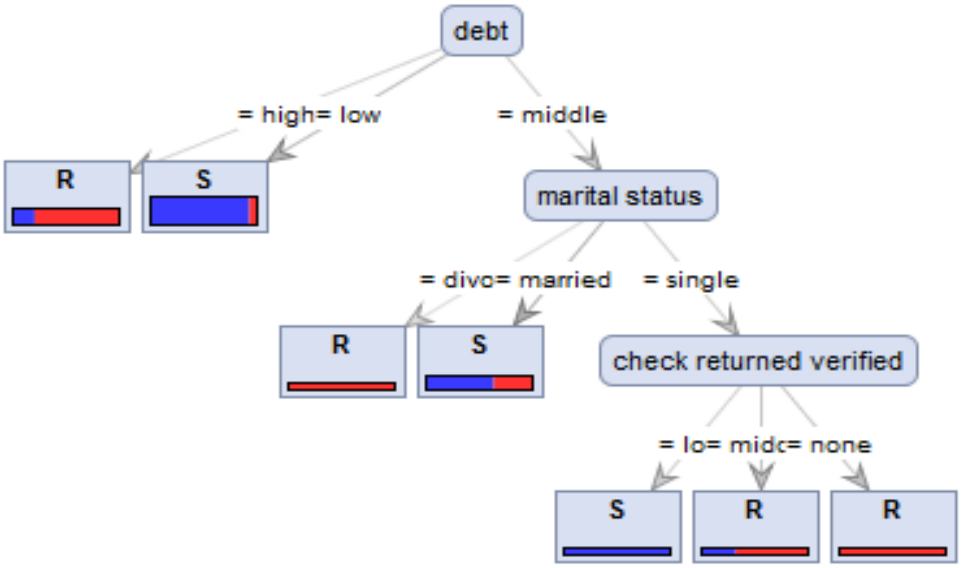

Figure 1 C4.5 decision tree

Conclusion

As mentioned earlier, this is an applied research. Therefore, research results may be used for planning and decision making of senior managers in firms and financial institutions.

The research was modeled and customers were classified according to the information collected through Mellat Bank in Shiraz. 10 variables were selected of the initial 20 variables interviewing 9 banking experts and frequent reviewing. The variables included age, education, marital status, sex, income, and annual debt, number of bank interaction; check returned non-verified, check returned verified as well as risky or safe customers. Age, number of bank interaction and debt value were categorized into specific classes.

Data volume reduced by data preprocessing; efficient data were provided; as a result, data were classified faster and more accurate. Regarding the nature of required data, some features were excluded so that less processing operations are employed for sorting and other operations; further, valid customers are discovered in more proper time. Since data provided to the researcher in paper files, some unrelated data were also seen. Therefore, in some cases, empty features were excluded by data mining according to experts' perspectives. Then, data were prepared in excel files and transformed into acceptable formats of data mining tools like rapidminer.

Grouping was conducted in two phases. The first phase is training and learning phase offering the prediction model. The second is the grouping phase, which predicts. Classification is carried out by simple sampling method and validation method. In simple sampling method, a percentage of data, 70%, are selected for training and the remaining, 30%, were applied for model evaluation and testing. The second method was k-fold cross

validation in which data were divided into 10 equal parts and each time a tenth of this value was tested and evaluated. The suggested method was using decision tree algorithms including ID3, C4.5, and CART algorithms. All three trees used the two simple sampling and validation methods leading to six results.

Therefore, the best proposed algorithm was used for model representation; further, C4.5 algorithm was finally applied for customers' better prediction. Thus, the prediction of this algorithm is more precise than other algorithms.

Limitations

Since gathering financial information of an organization or institution is impossible (follows a special law); therefore, this research faced several limitations due to studying a financial firm like bank. The limitations included failure to access bank customers' information as well as gathering few, limited number of customers caused lack of big data. On the other side, some variables that were the best features were excluded regarding bank law limitations. If more data and significant variables were provided, more reasonable and accurate results and prediction model could have been obtained.

Recommendations

- Other classification techniques in data mining may be used for evaluation.
- Customers are classified through clustering and cluster output is used as classification algorithm; further, bank customers class is identified online.
- Using support vector machine algorithm and deep learning new technique are recommended for improving and developing the current research.

References

Deshpande, S.P. & Thakare, V.M. (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems (IJDPS), 1 (1): 32-44.

Hicks, D. (2006). "Customer focus meets business agility: the business case for SOA".

Ismaeili, M. (2012). Data mining notions and techniques. Islamic Azad University, Kashan.

Kim, J-B., Song, B.Y., Zhang, Y., Earnings (2015). Performance of Major Customers and Bank Loan Contracting with Suppliers, Journal of Banking & Finance, doi: http://dx.doi.org/10.1016/j.jbankfin. 2015.06.020

Norouzi, A. (2009). Identifying and predicting customer defection rate using data mining techniques (Case study: Keshavarzi Bank). M.A. thesis, Tarbiyat Modarres University.

Oscar Herrera-Restrepo , Konstantinos Triantis , William L. Seaver , Joseph C. Paradi , Haiyan Zhu ,(2015). Bank Branch Operational Performance: A Robust Multivariate and

Clustering Approach, Expert Systems with Applications, doi: 10.1016/j.eswa.2015.12.025

Peppard, J. (2000), "Customer relationship management (CRM) in financial services", EuropeanManagement Journal, Vol. 18 No. 3, pp. 312-27

Shahrabi, J. (2007). Data mining. Gita data processing research institute and Jihad Daneshgahi, Amirkabir University, 1st edition.

Smith, R., (2006). "The state of the CRM market", Destination CRM, Viewpoint

ThuyUyen H. Nguyen, (2007), Strategies for successful CRM implementation, Information Management & Computer Security Vol. 15 No. 2, pp. 102-115

Tuuy Uyen, H. N. (1998). Strategies for Successful CRM implementation, 33-51

Wang, Y., Lao, Y., Wang, Y. (2014) "A fuzzy-based customer clustering approach with hierarchical structure for logistics network optimization", Expert Systems with Applications

Y.V. Bhapkar, A.D. More, (2014). 'Credit Risk Analysis of Bank Customers Using Data Mining Techniques', INTERNATIONAL JOURNAL OF MULTIFACETED AND MULTILINGUAL STUDIES, ISSN: 2350-0476, ISSUE-I, VOLUME-I

Yang GongXin, (2013). 'The research of improved Apriori mining algorithm in bank customer segmentation', Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering