IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 21, Issue 3. Ser. II (March. 2019), PP 24-28
www.iosrjournals.org

Using decision tree classification algorithm to design and construct the credit rating model for banking customers

FarzadShahbazi[1]
[1](*Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran)*
Corresponding Author: FarzadShahbazi

The application of classification models in credit rating of banking customers has been investigated in the present paper. Credit rating is one of the main applications of data mining in banking industry. The customers' creditworthiness can be evaluated through credit ratings. The data related to banking customers is very huge, and various classification techniques can be used to explore the hidden pattern and knowledge in data set through data mining. Several studies have been performed on the use of data mining and classification techniques in credit rating of banking customers. After preparing and preprocessing the data using C5 decision tree algorithm in this paper, the classification model has been constructed and credit rating of banking customers has been performed. A set of credit rating data has been used in this regard for teaching and testing the model. The results show that the developed model ranks banking customers with high accuracy by using decision tree making classification algorithms. The proposed classification model can also be used to credit rating of new banking customers.

Keywords:*Data mining, Credit Rating, Decision Tree*

Date of Submission: 22-02-2019 Date of acceptance: 08-03-2019

I. Introduction

As the new interdisciplinary field, knowledge discovery and data mining combine different areas, such as databases, statistics, machine learning, and other related fields, to extract information and knowledge using a large volume of data [1]. Machine learning techniques have provided the researches in different fields of engineering, science and technology with a great tool to solve many important challenges. To name a few, in chemistry, a machine learning method is proposed by J. N. Eisenberg, et al, for classification of chemical pollutants [2]. In [3], O. A. Gashteroodkhani, et al, carried out a project by using a machine learning tool i.e., support vector machine (SVM) for minimizing the electric outage time by accurately localizing faults in a power system. The method uses a metaheuristic algorithm to improve the performance of the machine learning by tuning its parameters. As mentioned in [4], metaheuristic algorithms can be used as a higher-level procedure to select a partial search algorithm that is able to find a sufficiently good solution to an optimization problem.

Another method for tuning SVM parameters are provided in [5].

Modern data mining techniques have had a significant contribution in the field of information science and can be complied with credit assessment models. Many research has been conducted on ranking the banking customers in recent years. For example, Desai et al. conducted a research in 1990s aimed at classifying customers for international loans, and achieved a list of credit predictive variables using information related to received credits [6].

In [7], Yobas et al. divide credit customers' performance into two categories of good and poor payers using fuzzy inference method and decision trees in another study. Various factors such as the main credit cards, employment status of the payer etc. have been considered in this study.

Other computational classification algorithms can be addressed inMirmozaffari et al. such as healthcare in prediction of heart disease [8], eco-efficiency of cement companies [9]and also other combinatorial optimization method such as Data Envelopment Analysis (DEA) [10]and expert systems in fuzzy [11]. Classification models of data mining can be widely used in the credit rating of banking customers. The present study aimed to propose a credit rating classification model for banking customers. This model can be used to assess the credit rating of banking customers. A credit rating data set is used in this regard for training and testing the proposed model.

The main problem of this research is how we can use the credit rating models such as the C5 decision tree to rank the credit ratings of banking customers. In this paper, the data were prepared and preprocessed. Then, the credit rating of the banking customers is performed using the C5 decision tree classification algorithm. Finally, the proposed final C5 decision tree classification model is used in practice as a high accuracy model for credit rating of banking customers.

II. Data Analysis

Data preparation and preprocessing:

The hidden patterns and knowledge in the credit rating dataset of banking customers is addressed in this paper, using the C5 decision tree classification algorithm. This dataset consists of several specifications (features or fields) and includes a large number of transactions (records). The features or fields of this dataset are described below:

Credit rating data sets for banking customers have a number of features or fields that are mentioned below. These fields include:Age, Job, marital status, education, debt, housing status, loan amount, type of contact, day, month, duration, credit status.Before describing the desired features in the dataset, one can describe the interest field in the study. The given business conducted credit rating of banking customers using data mining. In this regard, the business is looking at its creditworthy and non-creditworthy customers after receiving information from the customer and the recipient of the facility. Useful patterns and knowledge can be gathered from the generated datasets to improve and make better decisions in this business. The credit rating model can be used to determine creditworthiness of a new customer.

The description of the set of features or data fields of the credit rating of banking customers is described below.

The set of transactions in this data set equals 3146 transactions. The number of features is also 12. The present study aimed to propose a model that uses classification techniques in data mining to explore hidden patterns and knowledge in a credit rating dataset of banking customers. To this end, data were collected, prepared and pre-processed and then, the C5 decision tree classification algorithm is used on the dataset.

The data preprocessing is one of the most important activities in data mining. The set of operations performed here for preprocessing data are as follows:

Removing some transactions with missing values, converting the values of some fields such as age, financial balance, and day from decimal to integer.

The set of credit ratings features of banking customers are described in the table below:

Table no 1:*The set of credit ratings features of banking*

Feature name	Feature Type
Age	Numeric
Job	Nominal
marital status	Binary
Education	Nominal
Balance	Numeric
Housing	Binary
Default	Numeric
Type of contact	Nominal
Day	Nominal
Month	Nominal
duration	Numeric
Credit status	Nominal

III. Developing the proposed model

The model, proposed in this article, is developed after preparing and pre-processing the data. The proposed model in this paper, which uses the C5 decision tree technique, is presented in the figure below.

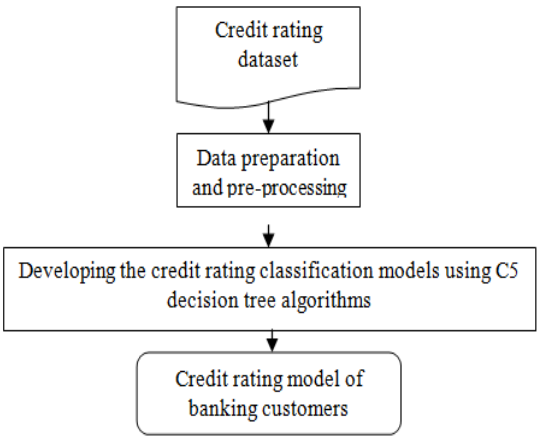

Fig.1. Proposed model

The first step in developing the proposed model is to collect credit rating data sets. Then, data were prepared, pre-processed and purified. After pre-processing the data, the C5 Decision Tree classification technique has been used. Various models can be used to build a classification model. In the present paper, the C5 classification model is used to build a credit rating model for banking customers.

IV. Decision tree algorithm

A survey on decision tree can be find in [12]. Most algorithms developed for learning decision trees are derived from a basic algorithm that uses an up-down greedy search in the decision trees space. This method is indicated by the ID3 algorithm and its fuller version namely C4.5.

ID3 decision tree algorithm: The tree algorithm make the decision as up-down procedure, and begins by asking that which attribute should be tested at the root of the tree. The algorithm then evaluates the decision based on each sample attribute using a type of statistical test to determine the most suitable attribute for classifying training examples, to answer this question. Next, the best attribute is selected and used as a test in the tree root. A corresponding node is created for each possible value and the training examples are partitioned between these nodes based on the test feature values. All the process mentioned above is repeated using the training examples attributed to each node to select the best attribute for the test in that tree node. It offers a greedy search method for an acceptable decision tree, which never returns to reassessed previous alternatives. This algorithm has some problems in training the samples non-valued attributes and it is also non-incremental and inexpensive.The C4.5 algorithm is the next generation of the ID3 algorithm which uses a kind of dimension pruning rule. It is also able to use discrete attributes, non-valued attributes, and noisy data. This algorithm selects the best attribute using the irregularity criterion and is able to apply attributes with very large amounts due to the use of the GainRatio factor. Even if there is no error in training data, pruning will be done, which will make the tree more general and less dependent on its training set.

In this relatively complex algorithm, pruning is based on the binomial distribution and in the form of a recursive to tree leaves. By stopping the pruning of a branch, it does not continue upward. To avoid the presence of leaves with a test sample, no further separation is performed on branches that have already fallen into two elements. Pruning is only done when the predicted number of errors does not increase. Given the irregularities of each of them, this algorithm select an attribute for each item that has data. After choosing the best attribute, items with non-valued attribute items are assigned in the part of the data that is provided with the values of attribute, and the algorithm continues.

V. The proposed model implementation

The implementation of the proposed model is discussed in this section. After the preparation and pre-processing of data, Clementine software is used to implement the proposed model. This software is one of the applications for implementation of data mining algorithms.

The Clementine software is used to implement the proposed model. The C5 decision tree algorithm is used to implement the proposed model. The training and testing dataset is used in this regard to construct the proposed classification model.

The C5 decision tree model is constructed below for credit rating of banking customers through which, it is possible to classify and rating the new banking customers' credit. Then, the creditworthiness is considered as a target feature.

The results obtained from the implementation of the C5 decision tree algorithm are as follows.

At first, important variables or features are identified including "duration".

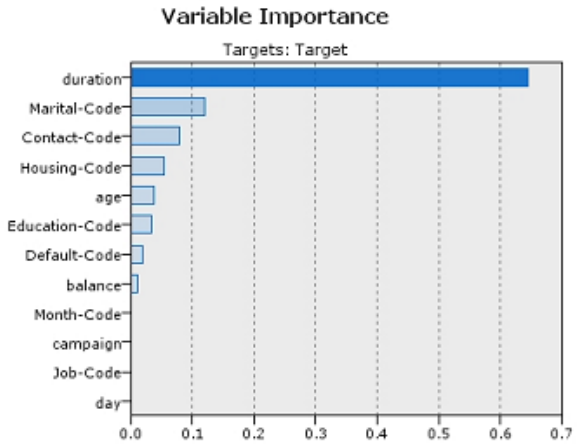

Fig.2. Credit rating of banking customers

Also, a part of the decision tree derived from the implementation of the C5 algorithm is as follows.

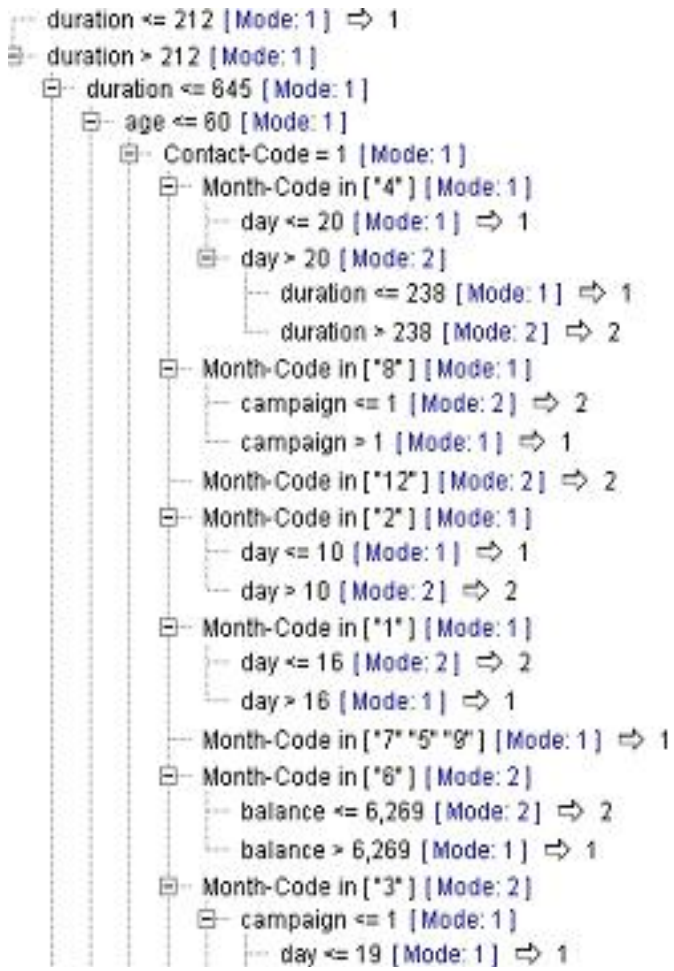

Fig.3. Decision tree derived from the implementation of the C5 algorithm

For example, this decision tree indicates that if "duration" is between 212 and 645 and less than 60, the contact code is 1, the month code is 4 and the day is less than 20, then the customer is in the class 1.Next, the decision tree is evaluated.

According to the following figure, the decision tree classification accuracy is 89.38%. That is, the model has classified customers by 89.38% of accuracy which means that 2928 of 3146 people are correctly classified in this classification.

'Partition'	1_Training		2_Testing	
Correct	2,928	93.07%	1,229	89.38%
Wrong	218	6.93%	146	10.62%
Total	3,146		1,375	

Coincidence Matrix for $C-Target (rows show actuals)

'Partition' = 1_Training	1	2
1	2,738	46
2	172	190

'Partition' = 2_Testing	1	2
1	1,173	43
2	103	56

Fig.4. The decision tree classification accuracy

V. Conclusion

Data mining techniques can be useful in providing new knowledge and patterns on banking customers' credit ratings. A model, appropriate for credit rating of banking customers has been proposed in this paper that can be used for credit rating of the new customer. The C5 Decision Tree Algorithm was used is the proposed model to construct decision tree for credit rating of banking customers. The developed decision tree ranked the banking customers credit with a high accuracy of 89.38%.

References

[1]. M. Kantardzic, Data Mining: Cocepts, Modeles, Methods, Algorithms, IEEE press, 2003.
[2]. J. N, Eisenberg, T. E., McKone. "Decision tree method for the classification of chemical pollutants: Incorporation of across-chemical variability and within-chemical uncertainty." Environmental science & technology. 1;32(21):3396-404, Nov 1998.
[3]. O. A. Gashteroodkhani, M. Majidi, M. Etezadi-Amoli, A. F. Nematollahi, B. Vahidi, "A hybrid SVM-TT transform-based method for fault location in hybrid transmission lines with underground cables" Electric Power Systems Research, vol. 170, pp. 205-214, 2019.
[4]. O. A. Gashteroodkhani and B. Vahidi, "Application of Imperialistic Competitive Algorithm to Fault Section Estimation Problem in Power Systems," in The International Conference in New Research of Electrical Engineering and Computer Science, Iran, Sep 2015.
[5]. V. Cherkassky,Y. Ma. Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks. 17(1):113-26, Jan 2004.
[6]. V.S Desai., J. N. Crook & G. A. Overstreet. "A comparison of neural networks and linear scoring models in the credit union environment", European Journal of Operational Research, 95(1), 24–37, 1996.
[7]. M.B Yobas, J. N Crook, P. Ross P."credit scoring using neural and evolutionary techniques", 1997.
[8]. M. Mirmozaffari, A. Alinezhad, and A. Gilanpour, "Data Mining Classification Algorithms for Heart Disease Prediction," Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE), ISSN 2349-1469 EISSN 2349-1477, Vol.4, Issue1, Jan 2017.
[9]. M. Mirmozaffari, "Eco-Efficiency Evaluation in Two-Stage Network Structure: Case Study: Cement Companies". Iranian Journal of Optimization (IJO). Dec. 16, 2018.
[10]. M. Mirmozaffari and A. Alinezhad, "Ranking of heart hospitals using cross-efficiency and two-stage DEA," 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, 2017, pp. 217-222.
[11]. M. Mirmozaffari, "Developing an Expert System for Diagnosing Liver Diseases", *EJERS*, vol. 4, no. 3, pp. 1-5, Mar. 2019.
[12]. S. R. Safavian, D. Landgrebe A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics;21(3):660-74. May 1991.