Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining 
A Case Study: Branches of Mellat Bank of Iran

Somayyeh Zamani, 
Department of Accounting of Payame Noor, I. R. of Iran 
Corresponding Author
Abdolkarim Mogaddam 
Department of Accounting of Payame Noor, I. R. of Iran

Abstract

In the development of trade and business throughout the world, needs to become more widespread financial dealings that led to the development of the business activities of banks and new banks were established. Banking and financial systems, credit risk management is a major problem. So check applicants’ credit facility to repay the process is important and many methods have been proposed for this work is the study of data mining techniques to identify genuine bank customer credit risk will be used. Numeric and non-numeric data is the selected data set consisting of 1000. This data set contains 20 features. 7 of the 20 features and 13 features a numerical and non-numerical attributes are nominal. Actually this is 20 character is input. There's also a feature called Class, category and class of a row of characters in the show. (Ie the output of the problem) that are included in this category are 2 classes, good and bad. In the first 300 customer data poorly (Y = 1) and 700, proper and good customer (Y = 0) exist.

Keywords: Credit Risk, Ranking, Mellat Bank.

Introduction

Banks as the main part of the financial system has been one of the main risks that they face different credit risk. A significant volume of facilities or outstanding fuel banks, indicates a lack of appropriate models for credit risk measurement and risk management systems in the banking network [1, 3]. One of the most important tools banks need to manage and control credit risk is, is "customer credit rating system". Using the data mining analysis of information related to bank customers using credit ratings can be creditworthy loan applicants and classify them according to customer accounts and bad, without judgment on the basis of smart payment systems. So banks to determine the needs of its customers, the granting of credit facilities to identify the characteristics of them. This validation by a decrease in bank risk including credit risk is. Accreditation refers to the practice where customers natural and legal validity of financial institutions and bank credit is measured according to the information received from them. And the possibility of a better understanding of the financial situation and people to repay loans received and provide more services. According to this method, the credit risk is measured and customers based on their credit risk classification and grading are [7, 11]. Credit ratings, risk management tool that uses data and statistical techniques to rank the customer pays a little Applicants facilities. Credit ratings of the different attributes of the default of its loan applicants separately examined and calculate the probability of default of the loan applicant's credit ratings at risk of deals [16, 17]. Credit ratings of credit models into two groups: good and bad credit applicants divide. Good credit group, a group that will repay its debts on time and bad credit debt group that has a high probability they will default. Due to the dynamic development of the credit industry, today the industry has an important role in the economy. Although credit demand, increased competition and the development of new channels in the new economic environment, has
created new opportunities for credit institutions. But also has new tools and methods they need. Thus, institutions such as the review, empowerment and the arrival of new technologies has led the credit management process. Credit ratings models, one of the most important and most fundamental decision-making system that most of the required information and credit institutions to provide credit management [9, 10, 13]. The purpose of credit rating models to predict the risk of non-payment by the customer’s credit or classified into two groups: good and bad credit applicants. In other words, credit ratings set of models and methods of decision-making related to their creditors in granting credit to customers to help. A lender must decide two types: first, whether or not to give credit to a new client, and the second on how to behave with customers [21, 22, 26]. One of the ways that credit institutions and banks for two reasons: the existence of a system for your customers require credit ratings. Customer credit rating system for banks and credit institutions provide the possibility of relying on clients Bank of Golestan province in terms of credit risk by using data mining.

Literature Review
such a system, based on the existing duty rates, as far as possible reduce the risk of their credit portfolios. Considering the importance of this research deals with the rank ready real c

Credit Ratings
Credit rating is an estimate of credit condition or the ability to pay debt. Financial institution follows the certain procedure, using the statistical method to lay down a number of rating standards which accesses an enterprise’s credit situation and gives an overall default risk assessment. Getting each credit attribute gives the quantification, and calculates its points and rating. According to the rating the quality of credit is decided [28], [32]. In addition, when this credit intensity changes, a financial institution can promptly make the suitable revision to the credit rank, to reflect as present the credit quality.

Data mining
The concept of data mining Using machine learning, statistical analysis, and other modeling techniques, the patterns and relationships in data can be found. The activities to discover hidden knowledge contained in data sets have been attempted by researchers in different disciplines for a long time. By the end of the 1980s, a new term, Knowledge Discovery in Databases (KDD), was coined [31, 37] and quickly adopted by artificial intelligence and machine learning practitioners to cover the overall process of extracting knowledge from databases, from setting the business goal to eventual analysis of the results. In this context, the word “data mining” was used for one step in the KDD process --- the step when the mining algorithms were applied. This interpretation was formalized in 1994 [25, 29]. Recently, as a result of the increasing attention of vendors and the popular trade press in this area, the word "data mining" has been used and has come to mean, like KDD, the overall process of extracting knowledge from databases [17, 20]. This paper adopts this recent interpretation of data mining. This interpretation emphasizes that data mining is not just a set of mining algorithms, but rather a process: A process that aims at solving a definite problem or making a decision, utilizes various mathematical and computer techniques to analyze the relevant data stored in large databases, finds a solution based on the discovered patterns in data and applies the solution to the predefined problem.

Data mining is not a new term, which has been used for a long time, especially by statisticians [14, 28, 29]. But the idea of extracting knowledge from database has revolutionary meaning for modern enterprises. In order to make use of the data to facilitate business decision making, more and more enterprises are building their data warehouse by reorganizing their stored operational data and bringing in other external data. An ideal expectation of data mining technology is to automatically discover decision supporting knowledge from the huge data saved in the large databases. Some data mining definitions contain the term “automatic”, such as, "data mining is a set of techniques used in an automated approach to exhaustively explore and bring to the surface complex relationships in
very large datasets” [20, 29, 30]. This is likely to be misunderstood as implying that answers will magically appear when a mining tool is applied to a database. However, the overall process of data mining is far away from “automatic”, but involves so many human interventions that some authors regard the process of data mining as a combination of art and science [13, 33].

**Methodological framework**

As there are many previous works in the area of credit scoring, the literature review was based on the descriptor, “credit scoring”. Full text of articles reviewed and the ones that were not actually related to the data mining techniques are excluded. Other selection criteria are as follows [7, 11]:
- Only Science direct online journal database were used;
- Only those articles that were in published journals and used the data mining techniques are included;
- Masters and doctoral theses, conference papers, working papers and internal reports, text books are excluded from the review mainly because academics prefer journals to acquire and disseminate information.

Figure 1 shows the methodological framework of the research.

The primary databases have about 110 articles and with further investigations and refining the results 44 articles were remained and other 66 articles were eliminated because they were not related to the application of data mining techniques in credit scoring. Each of the 44 remaining articles was studied and reviewed carefully and classified in 5 tables according to their type of study [37, 38].

**Classification method**

In this section, a graphical conceptual framework shown in Figure 2 is used for classifying credit scoring and data mining techniques. The conceptual framework is designed by literature review of current researches and books in credit scoring area [1]. As shown in Figure 2, the given framework consists of two levels. The first level includes three types of credit scoring problem comprising Enterprise credit score, individual’s credit score and small and mid-sized credit score.

(i) **Individual (consumer) credit score:** The individual credit score, scores people credit using variables like applicant age, marital status, income and some other variables and can include credit bureau variables.

(ii) **Enterprise credit score:** using audited financial accounts variables and other internal or external, industrial or credit bureau variables, the enterprise score is extracted.
(ii) **Bank credit score**: For SME and especially small companies financial accounts are not reliable and it’s up to the owner to withdraw or retain cash, there are also other issues, for example small companies are affected by their partners and their bad/good financial status affects them, so monitoring the SMEs counterparts is another way of scoring them [1]. As a matter of fact, small businesses have a major share of the world economy and their share is growing, so SME scoring is a major issue which is investigated in this paper.

Although some differences can be found for scoring of export guarantees, EXIM banks and other

![Figure 2. Classification framework for intelligent techniques in credit scoring](image)

Institutions which have not the profit as their main goal, they are excluded because of their low literature [1].

The second layer, comprised from three types of solutions and variable selection, they are presented below.

- **Variable selection**: Selecting appropriate and more predictive variables is fundamental for credit scoring [7]. Variable selection is the process of selecting the best predictive subset of variables from the original set of variables in a dataset [8]. There are many different methods for selecting variables include Stepwise regression, Factor analysis, and partial least square.

- **Single classifier**: Credit scoring is a classification problem and mainly classified applicant to good or bad. There are many data mining techniques for classification including support vector machine, and decision tree.

- **Hybrid approaches**:
  
  The main idea behind the hybrid approaches is that different methods have different strengths and weaknesses. This notion makes sense when the methods can be combined in some extent. This combination covers the weaknesses of the others. There are four different hybrid methods [9].

  - **Classification + Clustering**
    
    Clustering is an unsupervised learning technique and it cannot distinguish data accurately like supervised techniques. Therefore, a classifier can be trained first, and its output is used as the input for the cluster to improve the clustering results.
    
    In the case of credit scoring, one can cluster good applicants in different groups.
- Clustering + Classification
In this approach, clustering technique is done first in order to detect and filter outlier. Then the remained data, which are not filtered, are used to train the classifier in order to probably improve the classification result.

- Classification + Classification
In this approach, the aim of the first classifier is to “pre-process” the data set for data reduction. That is, the correctly classified data by the first classifier are collected and used to train the second classifier. It is assumed that for a new testing set, the second classifier could provide better classification results than

Data description
Numeric and non-numeric data is the selected data set consisting of 1000. This data set is 20 Characters. Of these 20 properties, 7 features a numerical and non-numerical nominal 13 feature. In fact, it is 20 characters are input. There’s also a feature called Class, class category and the class of a row of characters show. (ie the output is) where this category include two classes are good and bad. The first 300 customer in bad data (Y = 1) and good customer 700 (Y=0) exist. The percentage of customers is still considered creditworthy and bad, good blue and red categories show the bad ones.

Attribute description for Dataset
In the Table Name attributes (attributes) you see:
Details of each feature can be seen in the following charts:
In any form, if desired feature in nominal (non-numeric) data extracted included several modes that feature is allocated, the number and frequency of each class to show it. For example, 1 features 4 different modes can be seen (Figure 1) that each case is determined. In the bottom of the frequency of the property class (group) shows. Good classes and bad classes red blue shows.
Development and evaluation of models using the algorithm J48

This algorithm is the same as C4.5 algorithms for building decision trees is classified.

To build the model algorithm parameters must be set. Here are the default software has been used to build the model:

1. **Test Option (Using Training Set)**

Tree model output and built-in file names (Trees-J48-Training set) and (.model Trees-J48-Training set) is stored. In summary, the results are as follows:

![Tree Model](image)

*At the root of the tree was struck by the branch as checking status.*

**Number of Leaves: 6**

**Size of the tree: 9**

**Rules are stored in files:**

Using the built-evaluated learning tree that summarized the results as follows:

**Evaluation on training set Summary**

<table>
<thead>
<tr>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>735</td>
<td>265</td>
<td>0.3587</td>
<td>0.3529</td>
<td>0.4201</td>
<td>83.9903 %</td>
</tr>
</tbody>
</table>

![Evaluation Results](image)
Root relative squared error 91.6636 %
Total Number of Instances 1000

2. Test Option (10-Fold Cross Validation)
Here is also the root of the directory tree is based on the characteristics checking status.
Tree made in the image below you can see:
Number of Leaves: 6

Size of the tree: 9

The results of tree:
Stratified cross-validation ===
Summary ===

Correctly Classified Instances 711 71.1 %
Incorrectly Classified Instances 289 28.9 %
Kappa statistic 0.259
Mean absolute error 0.3604
Root mean squared error 0.4291
Relative absolute error 85.7792 %
Root relative squared error 93.6282 %
Total Number of Instances 1000

Conclusion
In this study, a model for rating the bank’s target customers using data mining is proposed. Banks and other financial institutions provide a variety of services for customers. These organizations, with challenges such as increasing competition, the constant increase in marketing costs as well as having a direct relationship with customers face. To deal with these problems, these organizations want the choice of customers who are most likely to buy new products and services and establish a direct contact with them. The purpose of this research was to improve the point that we can predict whether or not a customer is suitable for marketing operations? To achieve this goal, we have developed a model of classifying data mining to select target customers. Using this model, banks can reduce their marketing costs.

In this study, a decision tree model was used as the classifier. Fashion modeling included several stages in the Pre-election Data feature of the genetic algorithm was used.

In order to select the algorithms and methods for the implementation of each stage, a literature review of each step thoroughly reviewed and assessed. After a review of various methods in each stage, the best and most convenient method to the innovation and
improvement will be realized. For the purposes of the research, algorithms and methods of feature selection process with genetic algorithms and data mining software was WEKA.

The first step was to feature selection using genetic algorithm was performed. The training and testing data sets to evaluate the accuracy of separation. Two-thirds (2/3) of customers were determined empirically for the training set and one third (1/3) were also set for the test. The validity of the test series was 88%, which is a good improvement over previous models. It was shown that the use of genetic algorithm to select features on the classification of customers is a good influence. As a result, banks and other financial institutions model to predict, to identify a subset of customers for appropriate marketing will help.

Using this model, not only can significantly reduce marketing costs, but also can reduce the level of customer satisfaction and improve relationships with customers is achieved. The benefits of the proposed model are as follows:

Equality attributes: When two or more characters are very similar, large samples to decide between them, with high reliability is required. It’s useless, because in this case, what attribute is selected, it creates slight differences in data collection time and a lot of time will be more samples. Calculation of the decision tree, a significant proportion of time spent to calculate the information. Re-calculate the data for each new sample is inefficient, because the possibility of the decision to split the point is unlikely.

Thus, the proposed system allows the user to specify a minimum for features, before the data is calculated. This is effectively the overall time taken to calculate the loss of information and learning system is faster.

References
5. Chen, W., Ma, C., and Ma, L., 2009, 'Mining the customer credit using hybrid support vector machine technique.' Expert Systems with Applications 36(4), 761-766.
26. Mangasarian O.L., (1965) Linear and nonlinear separation of patterns by linear programming,
37. Mohtashami T, Salami HO.2006. Discriminative factors of civil customers with low risk from customers with high risk on the commercial banks, case study: Agree bank of Iran, college of economic and Agree development, Tehran University