International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 2, No. 2, 2015, pp. 83-94.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Provide a method for validation of bank customers using data mining techniques

Fatemeh Sadeghi Laghareh[a] , Seyed Javad Mirabedini[b] , Ali Haroun Abadi[c]

[a] *Department of Computer , Electronic Branch, Islamic Azad University, Tehran, Iran.*

[b] *Faculty Member of Islamic Azad University Central Tehran Branch.*

[c] *Faculty Member of Islamic Azad University Central Tehran Branch.*

Abstract

Today in the banking industry loans have Fundamental role, as a large part of the assets of a bank of Loans paid to persons and companies to be formed and so with increasing the number of loan applications from people and with respect to the risk in these activities, providing a way to manage these loans is essential. If banks and financial institutions before the granting facilities, Validation of Seekers Facility Properly do And our credit rating is correct And facilities are properly paid, Certainly the cost of granting facilities will be lower Compared to when is deferred And we want to collect receivables. If validation Customers do not And monitoring and control does not exist. Resources goes to waste. One of the main reasons for the increase in bad loans due to a lack of full Validation and ranking customers in Iranian banks with the aim of transparency and competition in the banking system is. So if there is a system of true validation amount bad loans dropped is granting facilities processes faster, as well as credit risk and bank costs significantly reduced. . Many Data mining techniques to predict and validation bank customers there are, that each of them can be According to the data and desired efficiency chose. But these techniques terms of accuracy have error and this reduces performance some of them in this article using the logit regression model runs for validation customers suggested we hope to be fruitful.

Keywords: validation ، banking customer ،data mining

Introduction:

Process economy of the country towards competitive market can have a significant impact in the country's progress and Transparency information in Iranian banking system Increase. To this end, the best solution is greater accuracy in the validation of our customers. System that is in the process of customer orientation in banks can have a big impact this system needs various substrates. Process data collection and information in Iran's banking system due to lack of comprehensive bank of information from banking customers have problems That this problems in payment facilities to customers or provide long-term plans in the field of amount facility which will be paid in the future Occur And also can not considered as a guarantee for collection of receivables .

One of the main reasons for the increase in bad loans due to a lack of full validation and ranking customers in the Iranian bank Are with the aim of transparency and competition in the banking system.

The importance and necessity of research:

Advances in use of and collection of data using computer and also the public use the Web as a universal information system, data mining as a useful tool to help solve the complex issues raised.

In recent years, something important validation for financial institutions such as banks and insurance companies has become. In fact, poor management In terms of credit allocation, losses many financial institutions, especially in the United States of America and Europe have imposed and even some of them has led to the point of bankruptcy. Validation of to reduce the risks arising from granting credit will pay. This is done with a correct forecast customer behavior.

In fact, the issue of validation as a category issue can be raised the goal of which extract the model is efficient and appropriate that customers based on the properties classified in two categories: the good and the bad.

In the past, in order to make model validation, Classical methods of data mining such as linear regression and Linear audits analysis was used. But recently, artificial intelligence and machine learning tools, Attractive for validated Researchers have been

Research method

Data Mining and Knowledge Discovery in Databases

Data mining is a process in which the use from data analysis tools to look for patterns and connections between existing data that may lead to extract new information from the database cause are.

In other words, the data mining process is use smart techniques, knowledge of the collection of the data. Collection of data that are processed, known as the training set. Knowledge extracted in the form of models, patterns or rules various forms for provide knowledge are extracted. This knowledge can be a basis for future decisions, subsequent operations or changes needed in the system be placed.

The concept of data mining

Data Mining Most useful heuristic analysis scenario is where thought and harvested predetermined about the remarkable results achieved there is not. In fact, data mining search process required for find new information, valuable and unknown among a host of data.

Including data mining work can be to clustering ، Categories ،visualization of data ، deviation analysis and involvement rules cited. Validation method for the detection group or different groups in a seeking population of credit cards or loans. Based on this validation among other things classification and clustering placed And because the classification and clustering of data mining tasks is ، So we can used for validation of data mining.

Techniques and data mining algorithms

Data mining algorithms, specify the types of problems that can be modeling and solved. Data mining functions generally placed into two categories
1. Supervised learning algorithms (supervised)
2. Learning algorithms without supervision

Table 1, data mining algorithms for supervised functions described.

algorithms	**functions**	**Explanation**
Decision trees	Classification	Decision trees, predictive information in the form of humanitarian law understandable Their extracted. These rules in the form of expressions if - then – in otherwise they are; They explain the decisions which led to is predicted.
Generalized linear models (GLM)	Classification and Regression	GLM, logistic regression for classification Binary purposes and linear regression for continuous objectives implementation does. GLM classification from confidence boundaries for predicted probabilities support. GLM regression to predict from the boundaries of trust support.
Minimum Description Length (MDL)	Significance	MDL is a theoretical model of the information. MDL assumes that simplest, most compact presentation of information, the best and the most probable explanation for the data.
Simple Bayes (NB)	Classification	NB, predictions by using the Bayes theorem does that probability of a forecast of the evidence underlying does conclusion, as in the data been observed.
Support vector machine (SVM)	Classification and Regression	Separate versions of SVM, from core different functions that use different types of data collection guide. Core linear and Gaussian (nonlinear) are supported. SVM classifier to separate the target classes with the widest margin possible tries. SVM regression tries to find a continuous function so that the maximum number of data points in tube with size Epsilon placed around it.

Table 2 data mining algorithms for describes the functions without supervision.

algorithms	**functions**	**Explanation**
APRIORI	Relationship	APRIORI, market basket analysis done by the finding items at the same time (set frequent items) in a set does. APRIORI the rules with support greater than the

		minimum support and trust more than the minimum specified trust specifies. For example, find items that are purchased together and their relationship specifies
K- averages	Clustering	K- Average, a clustering algorithm is based-distance data to number from the predetermined clusters are partitions. Each cluster has a center (center of gravity) is. Items (people in a population) who are in a cluster near are center. For example, demographic information to clusters segment is And the possibility that a person To a given cluster Belongs Ranking Is
non-negative matrix factorization (NMF)	Feature Extraction	NMF new Characteristics by using linear combinations from main characters to produce. Coefficients of linear compounds, non-negative. During the applying the model, a model NMF, the original data To a new set of features discovered by does mapping model. For example, according To demographic data about a set from customers, Characteristics for the general characteristics of your customers Category
single-class support vector machine	Anomaly detection	SVM single - class profile of a class are created, and when applied, flag cases that are somewhat different from the profile used. This diagnosis provides rare cases that are not necessarily related to one another. As an example, According to demographic information about a set of customers, customer transaction behavior that is different from the normal identify.

Credit risk

granting facilities one of the activities of major banks Constitute that gain or loss arising from facilities important role in the profitability of banks. For banks this is a natural thing that most people who refer to them to achieve the goals and resolve their needs to be able to provide that to reliance on limited resources at their disposal, they are not. Credit risk, is risky upon which the borrower is able or willing to repay the principal and interest of loans received with respect to the conditions specified in the contract is not. Even in cases of delay in repayment of the facilities and loans, banks with problems and dilemmas such as providing needed liquidity facing is. As a result, one of the major risk that banks encounter with it is credit risk. Banks in supply request clients facilities, Should In addition Observance instructions and Policies credit and programs of the bank, Action to investigate And estimating the needs of the applicant with regard to the risks of not returning are resources.

Credit Limit: order of credit limit or ceiling, the maximum amount that the bank is willing as financing or commitment in the form of specific tools at the same time at the disposal a customer has put or the commitment for him. Limit facilities for a maximum period one year or a production period is established and owner of Limit facilities could gradually or all from the relevant facilities Ceiling specified Used And for each paid again "from new facilities Use. Limit facility maturities examined Undergraduate located And in terms of bank resources, conditions time, the performance applicant, Expert report of the situation the applicant, the amount of approved credit limit, Increased or decreased or will become void. So credit limit isnot never constant and by changing the conditions of the bank and the customer's situation changes. (Directive No. 3 Bank Sepah)

Validation of bank customers
Validation refers to a process in which each seeking facilities ، a quantity or points allocated that represents an estimate of the future performance he In repayment of the facilities is requested. Validation from important issues in the field of financial decisions. Generally, financial decisions ،
A large number of quantitative and qualitative factors, each of which have special features included. For example, in the evaluation process Loan applicants ، The main point of concern is credit analysts ، recognize applicants They credibility Can be guaranteed Applicants who are high risk and should probably demands to be rejected. Validation one of the applications is forecast ،that in this field we need to predict risk borrowers ، Instead of looking for the most profitable customers are and we avoid risky customers. An important part of this analysis to past information of clients is. Credit and behavioral scoring ، technique that are helps organizations, about loan or loan Request Seekers ، decisions. Information Resources This type of decision from details of request and request and Also information related to of the funding agencies are obtained and a final decision for rejecting or accepting the demand is. .

History of credit risk assessment (altman،1968)

By starting late fifties, more Companies wanting to improve credit ratings systems have and so need to have updated information. "altman" as one of the pioneers of discuss validation Many attempts to find a significant relationship between the variables accounting and the probability of inability to pay a debt of this company in the future was this relationship became known as the Z-Score.

Effective index in validation of bank applicants

Credit risk management right for customers in the business market, getting Inquiry are listed below from relevant principles is required. (Circular No. 3 Bank Sepah)

1. Inquiry checking back
2. Getting credit code
3. Getting a tracking number debts and obligations
4. Inquiry debts and obligations of the applicant
5. Inquiry blocked account
6. Inquiry creditworthy
7. Inquiry authentication systems consulting company Iran credit rating
8. Inquiry from beneficiaries of the system (derived from the banking laws)

The overall process of data mining

to explore knowledge and finding models in the collection of data, or the rules that govern them, there are various methods. Each of these methods, the data in a specific way Processed Placed And In terms of applications, Desired information does extracted from them. Seven stages of knowledge exploration include:

Data cleansing ،Data integration, selected data ، data conversion ،data mining, evaluate the pattern ، knowledge representation.

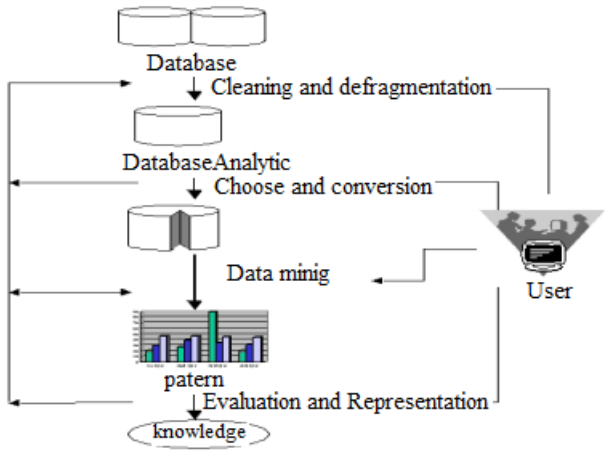


Figure 1: The main steps of the process of knowledge discovery from database

In other words, to use from data mining techniques, in order to build a predictive model able to، visualize report meaningful information to the user is simple.

Table 3shows the flow of data mining methods is our system.

Understanding the problem	• of knowledge about the banking sector • challenges and the most important concerns of Bank
Understanding the data	• Collect data about the characteristics needed • Customer Data
Filtering data	• Filtering data, prepare Features • Choice of data collection, check the lost data, and preparing it for extracted
Modeling System	• modeling system for customer retention, prevent and detect fraud • Find the most effective They feature
Evaluation System	• Check data from new customers with design model and • Verification Results
The analysis results	• Comparison with other models • Performance Analysis

History Research

Researches And numerous applications in the field of validation to identify good customers and bad banks have been that due to an error and a lot of time gradually gave replaced by the parametric and nonparametric methods (thomas, 2000) ، parametric methods such as Probit ، logit discriminant analysis and logistic regression ،from beginning of emergence validation and Behavior Survey Were used

Then nonparametric and methods data mining like decision trees, neural networks and expert systems ، were used. (Sabzevari et2007) The usefulness of technology decision tree based on genetic algorithm Another method validation approach using genetic algorithms in making decision tree. This method was proposed in 2008 (hung et.al, 2008) . Hsieh Wang by using collective classification method and neural networks Algorithms, and support vector algorithm and also The use of class-based classification ، classification system provided bank customers. Sastrstyk and et all , by used from artificial intelligence to classify customers applying for loans Paid. They have used a genetic algorithm to select the input feature. Also SOM algorithm for clustering samples have used neural networks to modeling and knowledge discovery. The results of using the proposed method of data validation customers in Germany and Australia showed precision high this method logic was.

Table 4shows some validation methods studies outside the country .

Researcher /date	Model validation	Row
(Susac, Sarlija, & Bensic, n.d.)[22]	**Logistic regression, decision trees, neural networks**	1
(Lee, Chiu, Lu, & Chen, 2002)[14]	**Composition analysis A distinction and back propagation algorithms In neural networks**	2
(Kim & Sohn, 2004)[12]	**Patterns of wrong classification**	3
(Lee & Chen, 2005)[13]	**The combination of artificial neural networkmodels and methods MARS**	4
(Xu, Zhou, & Wang, 2008)[26]	**Ranking analysis of links using SVM**	5
(Abdou & Pointon, 2008)[3]	**Neural Networks and General Techniques**	6
(Nanni & Lumini, 2009)[17]	**Class composition Replaced by a classifier**	7

decision tree

One of most the important classification tree method used in data mining is becoming beloved forecasting models (Burez and Van den Poel, 2007). These trees are widely used in medicine, computer science, business, engineering, insurance, botany, psychology etc. are used.

The flexibility of this method will become an attractive option. This issue means suggested avoiding the use of traditional methods is not really When the Theory and distributional assumptions hard traditional methods. Be established, possibly classification methods is preferable Decision Tree A graph, such as tree To represent the relationships between a set of variables One of the methods of knowledge discovery in data mining, which is generally used for classification and decision tree algorithm is a decision tree can produce intelligible rules, and even In a large tree or the complex, a way can be easily passed and this makes the relatively is easy to interpret classification or prediction. Different algorithms for building decision trees is introduced, that famous method which can be C5, C4.5, ID3, OC, CART, CHAID, SERCH, AID said.

logistic regression: Logistic regression or logit regression, a type of regression analysis is used to predict the result of a definite dependent variable based on one or more variables predict. To replace data fit in a straight line, logistic regression from logistic curve uses. (Formula 1) Multivariate logistic curve is as follows. (Aghabeighi et al.) Formula 1

$$p = \frac{e^{c_0+c_1x_1}}{1+e^{c_0+c_1x_1}}$$

Performance logarithmic function can be used to obtain a logistics function. (Formula 2)

$$log_e = \frac{p}{1-p} = c_0+c_1x_{10}$$

Logistic regression is simple and easy implementation and good performance in a wide range of issues offers. Logistic regression is kind of regression model where the dependent variable is a qualitative variable at two levels. In other words, in this case the response variable y determine the credit status and is divided into two categories.

First group: "creditworthy customers" means a group of bank customers compared to the timely settlement of its obligations before the expiry time for payment of installments are acted.

The second group of "bad customer accounts," means group that to timely repayment obligations facilities received failed. Therefore the variable y, defined that if y = 1 is customer-i-good account if

y = 0 is a customer of i non-creditworthy is in these case y_i has a binomial distribution is as follows: (Formula 3)

$$\begin{cases} P_r(Y_i = 1) = P_i \\ P_r(Y_i = 0) = 1 - P_i \end{cases}$$

$$0 \leq P_i \leq 1$$

And here pi is the probability that creditworthy clients. In logistic regression model from logit concept are used for dependent value. In other words, from logarithm of success chance (the ratio of probability of success to probability of failure) Logit is used; therefore in state of one variable, logistic regression model as one of two the following formula. (Formula 4)

Therefore logistic regression model is generally defined as follows (Formula 5)

$$\log it\,(P_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots . + \beta_k X_{ik}$$

According to the logit definition following ، pi value (probability of a person's credit- worthiness) is equal to (Formula 6)

$$P_i = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}}}$$

Mr. Mohammad Jalili and et al in the paper, "Validation of real customers in the banking system" Validation of real customers in the banking system paid. The research population is consists of real customers in the banking system are that in two categories creditworthy (low credit risk) and bad accounts (credit risk above) have been checked, Work process is this form that first to identify variables affecting customers credit status paid , After than classification variables model and estimates people to creditworthy and non-creditworthy in each class, then customers according to the instructions of the banks into two group creditworthy and non-creditworthy separated . Finally, check the model accuracy in determining the customer's credit status with credit status them has detected by the bank, Finally, to measure that results show Logit model with 68 percent accurately from predicted enjoys a good.

Logit model There are different statistical methods to estimate credit rating models. Accordingly, Models to determine the credit rating to two groups of-parametric models (linear probability models, logit models - probit, models based on discriminant analysis, neural networks) and nonparametric (mathematical programming, classification tree models, nearest neighbor, the process hierarchy analytical, expert systems divide). The general form of the proposed model is as follows:

Y = F (X1 , X2 , X3 ,…Xn) Equation (1)

Where y is the response variable and determining status of Applicant Credibility from discrete characteristics is given, because of the credit risk into two categories, creditworthy customers (zero) and bad accounts (one) divide. Now consider the following model:

Equation (2)

$$P_i = E\,(Y = 1 \mid X_i) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}}$$

That e is the base of natural logarithms. For convenience we rewrite discussed above equation as follows.
Equation (3)

$$P_i = \frac{1}{1+e^{-z_i}}$$

That is Zi = B1 + B2Xi. The above equation suggests that what is known as logistic cumulative distribution function. If Pi probability of timely payment of loan is expressed by equation (3) ، In this case, (1-Pi) the possibility of unpaying loan on time will be as follows.

$$1 - P_i = \frac{1}{1+e^{z_i}} \qquad \textbf{Equation (4)}$$

$$\frac{Pi}{1-Pi} = \frac{1+e^{z_i}}{1+e^{-z_i}} = e^{z_i} \qquad \textbf{Equation (5)}$$

If from Equation (5) take the natural logarithm ، interestingly result below is obtained:

$$L_i = \ln\left(\frac{Pi}{1-Pi}\right) = Z_i = \beta_1 + \beta_2 X_i \qquad \textbf{Equation (6)}$$

Odds ratio Logarithm, relative to linear parameters and values of after the estimate the coefficients from maximum likelihood method is obtained. Therefore with estimate the coefficients, Obtained model is that each of the coefficients, value variables (logarithm of the odds in favor of a failure to repayment) per unit change in the independent variable indicates by calculating the probability of failure in repayment from the equation (2) can be used to classify customers.
Model estimation

Table 4 coefficient estimate of logic function and final effects model

In this part of the research, 519 customer data after from processing in software sata to design an optimal model for measuring credit risk, to enter the model is ready And them to, trial and error method entered to the model and finally, after testing various combinations from types of variables mentioned and omitted variables, That the error rate is more than 05, had significant effect in the model, ***Had*** not And no significant effect on heart, In optimal model mentioned variables In Table 3 as a variable effect was. Therefore, general form of logit function or logarithmic probability failure in repayment to probability of failure in repayment, as follows:

Meaningful coefficients:
To test Meaningful of coefficients of the variables, Logit regression function From Parent statistical is used. It is noteworthy that the null hypothesis means that the variable of interest any effect on the

dependent variable have not. As shown in Table 4 Significance level for all coefficients obtained is less than 0.05 then these coefficients are meaningful.

Meaningful and goodness fit of the model:
In logit regression of the indicators listed in Table 5 are used to evaluate the goodness on value of model:

Table 5 reviews meaning and goodness to value

Final impact analysis:
Interpretation of coefficient values logit model is complex, because the estimated coefficients confirm obtained a two-point model is that cannot be interpreted as the final effect on the dependent variable. The results related to calculation effects in the table (4) is shown. Final effect of 0.673, for a variable value loans received from banks. This means that with each unit increase in cost loans. For example, less than five million toman to the cost reach between five and ten million toman, the probability of failure in the repayment of installments 67.3% increased. This means that with increasing the amount of loans, the possibility of failure in repayment of also increase. So greatest effect is compared to other variables. After the cost loans, the applicant Wife working, return cheque status, and the duration of repayment have the greatest impact upon the possibility of failure. Therefore Bank Sepah in making decisions about the customers should pay more attention to these variables.

The predictive power of model in the threshold 0.5
Table 6 the model in predictive power threshold 0.5

In Table 6, which is called the classification table at threshold 0.5 ، predicted values for the dependent variable y According to whether higher or lower than the threshold are. In front actual values showed, they are categorized.
Proportion of observations y = 1 which have correctly predicted the "degree of sensitivity" model called the 85.5% and the ratio of observations y = 0 which have correctly predicted to "degree of discretion" model called which equal to 91.90 percent.
Also according to the table (5), error of type 1 Means credit risk in the Merit model 14.47 percent (number one minus degrees of sensitivity) and error of type 1 , Means business risk equal with 8.1 percent (number one minus degrees of recognition) are.
About 10 percent of the total sample size was not used as test data in the Fitted model After the estimate the coefficients independent model variables, to evaluate the efficiency and predictive power of the Fitted model is used.

Table 7: Evaluation of the predictive power of the model using data Control

As can be seen, Data control models, to estimate the credit risk have the high efficiency of 83 percent. This shows the use of this model can be allocated 83% of loans to real customers and the model is capable in the estimating the credit risk.

Discussion and conclusion
Mobilization and allocation of resources for investment in economic activities carried out by the financial markets that bank credit market part of the markets. Regarding the role of banks in granting credit to

customers, reviewing and evaluating credit risk or probability of failure in the repayment of credit facilities from customers is essential. With importance of this issue in our country granting credit facilities to customers based on a consistent and regular process is done, Therefore having one effective risk models not only facilitate decision in the field of granting credit but also in reducing transaction costs leads to efficient allocation of resources to different parts of the economy. In this study, one logit model for credit risk management and with use of customer information Sepah Bank and with the use of credit scoring model is used. The results indicate that the logit model to satisfactorily able to identify credit risk of customers. The final effects of the variables showed that the possibility of failure on status loan amount, the applicant wife job, status of check back, period of repayment of the loan, marital status and housing status, highly depended than others are variables. As a result, Bank Sepah for deciding on applications for using the facility must more consider variables

Proposal for future research

Considering the extent of affected areas on credit risk of customers, investigate other factors that may affect the probability of failure to repayment of loans and assess its impact could is of topics for future research. In this use of other models such as Neural Networks and combination of neural networks and fuzzy can be placed on the agenda of other researchers and compared with logit designed model.

References:

Altman E. I, (1968) , "Financial ratios discriminate analysis and the prediction of corporate Bankruptcy", The Journal of finance 23.

SHAMSFARD, M. 2011. Challenges and open problems in Persian text processing. Proceedings of LTC, 11.

SHAPLEY, L. S. 1952. A value for n-person games. DTIC Document.

SYED, A., ASLAM, M. & MARTINEZ-ENRIQUEZ, A. 2010. Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits. In: SIDOROV, G., HERNÁNDEZ AGUIRRE, A. & REYES GARCÍA, C. (eds.) Advances in Artificial Intelligence. Springer Berlin Heidelberg.

VURAL, A. G., CAMBAZOGLU, B. B., SENKUL, P. & TOKGOZ, Z. O. 2013. A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish. In: GELENBE, E. & LENT, R. (eds.) Computer and Information Sciences III. Springer London.

ZACHARY, W. W. 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research, 452-473.

ZHOU, S., YUE, K., FANG, Q., ZHU, Y. & LIU, W. An efficient algorithm for influence maximization under linear threshold model. Control and Decision Conference (2014 CCDC), The 26th Chinese, 2014. IEEE, 5352-5357.