Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

DEVELOPING A MODEL FOR VALIDATION AND PREDICTION OF BANK CUSTOMER CREDIT USING INFORMATION TECHNOLOGY (CASE STUDY OF DEY BANK)

M. H. Yazdani

Department of Humanities, West Tehran branch, Islamic Azad University, Iran

Published online: 15 February 2017

ABSTRACT

Credit risk is the most important risk of banks. The main approaches of the bank to reduce credit risk are correct validation using the final status and the validation model parameters. High fuel of bank reserves and lost or outstanding facilities of banks indicate the lack of appropriate validation models in the banking network. The weakness of the previous models is due to the choice of inappropriate decision parameters, technical weakness of the model and lack of access to desired data. In this paper, in order to establish a communication between the final status and the parameters of facilities granted, data mining technique with the help of machine learning and neural networks have been used. A database of facilities granted by Dey Bank was created and a model with data mining approach was prepared. This model has good accuracy is able to validate real customer. According to the analysis, interest rate parameter is more important in determining a customer validation. The model has higher accuracy and comprehensiveness compared to the similar cases, due to the database size, type of data mining and learning algorithms applied.

Keywords: validation, credit risk, Dey Bank

Author Correspondence, e-mail: mflay1366@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i1s.693

1. INTRODUCTION

Bank's board is responsible for periodic (at least annual) approval and review of important credit risk solutions and policies. The solutions should indicate to what extent the bank can tolerate the risks and what the expected level of profitability against different credit risks is.

Senior management of the bank is responsible for implementing credit risk solutions approved by the Board of Directors. In addition,

the development of policies and procedures to determine, evaluate, monitor and control the credit risk is in charge of senior management. Such policies and procedures must specify the credit risk in all of its activities, both at the individual and portfolio levels [1].

Banks should identify and manage the credit risk in all of their products and activities. Banks must ensure that the risk of new products and activities, before offering or performing, has been evaluated with the help of appropriate risk management procedures and adequate controls and already approved by the Board of Directors or other appropriate committees. Granting of loans can be as much profitable as it may involve the bank in a variety of risks. As the banks concerned the overall profitability, they should also evaluate the relationship between risk and return for any credit. When considering the possibility of granting loans to the relevant conditions, it is necessary for the banks to evaluate the risk against expected return by regulating the terms of pricing and non-pricing (for example: documentation, contracts, etc.) at the fullest possible level. When speculating the risks, banks shall consider all possible negative scenarios and their potential impacts on the financial condition of borrowers or parties. A common problem among banks is that they do not tend to price a credit or a set of credits appropriately and thus, do not consider sufficient compensations to deal with potential risks [2 and 3].

One of the main projects of the Risk Department of Dey Bank is designing a validation model. The aim of this project is to design a mechanism to measure and manage the credit risk of Dey Bank. Such mechanism includes credit risk assessments at both individual facility and portfolio of facilities. This model aligns the decision of credit risk management and credit department to identify and allocate facilities to reliable customers. At the same time, it supports portfolio management decisions for the optimal utilization of diversification effects. Such a model has the following features [4-9]:

This model provides a context for the unity of lending decisions among different branches of the bank. Thus, a context is created for evaluating the performance of credit department of various branches based on a single criteria and it will be possible to design performance-based reward systems in branches.

Taking very conservative policies against credit decisions reduces the credit risk, but this risk mitigation is over at the expense of reducing bank interest income. Such policies, especially in competitive environments, lead to the elimination of financial institution from the competition scene. On the other hand, unplanned granting the credit also increasingly raises the credit risk

and thus, the credit losses of banks. So, granting credit to the applicants requires a compromise between risk and returns.

Such a model can provide the possibility of increasing the quality of bank loan portfolios. So, using this model is expected to reduce outstanding facilities of the bank and as a result, predict incoming cash flows arising from the repayment of interest and principal of facilities with more certainty. Thus, credit risk management mechanism not only reduces the costs of outstanding facilities and increases the profitability, but also helps to fix the issues of liquidity risk.

Inattention to credit risk makes the banks to increase the reserves to cover the credit risk. The major consequence of increased reserves is reduction in return on investment and thus reduce reduced profitability. Credit risk management model causes resource estimation to be non-experimental and non-conservative by providing reliable estimates. Such a situation is likely to reduce the amount of reserves and increase the accuracy of estimations.

This model is in accordance with the Basel Committee Treaties on credit risk measurement and management. Thus, by setting up such a model, Dey Bank can be among the higher ranked banks by complying the Basel Committee's Credit Risk Standards and also will benefit from more facilities (including the enjoyment of facilities with low interest rates) in its foreign relations with other banks, financial institutions and etc.

2. RESEARCH METHODOLOGY

Given that the present article is one of the subjects of Dey Bank Department of Studies, the data was extracted from the customer database by applying appropriate filters and then, all the study population was determined. So, no sampling method (e.g. questionnaires, etc.) has not been used in this article, because the whole population is available.

The main assets of a bank with regard to its nature and activities are the loans paid to the applicants. The assets are faced with various risks called credit risk, in general. Any anomalies in the facility system, from the credit macro policies to granting the loans to customer, increases the credit risk and threaten its nature and consequently the interest of investors. Its will provide disturbances in the economy. A control system for monitoring this risk will have great influence on the bank's activities and its development among the competitors. The use of this system can inform us about the outcome of policies and loan processes. The reflection of results to the Board of Directors and analyzing them will help the senior management to reform and fix the bugs. Among the most important things in a control

system of lending is the loan applicant evaluation that is called validation. Research hypotheses include:

- There is a significant relationship between the interest rate of Dey Bank's various facilities and their risk level. It means that if the interest rate is low, then probability of default will be low too.
- Increasing the number of installments for the loan granted by Dey Bank, the risk level will be increased.
- There is a relationship between the age of the borrower and repayment of loans.
- There is a relationship between the guarantor and the probability of default. It means that, if the applicant has guarantor, then the likelihood of repayment will be higher.

In data mining, four main operations are performed that include:

1. Predictive modeling
2. Segmentation of databases
3. Link analysis
4. Identifying the deviation

One or more of the above operations are used in the implementation of various applications of data mining. For example, segmentation and link analysis are commonly used for retail applications, while each of the four operations can be used for fraud detection. Moreover, a sequence of operations can be used for a particular purpose. For example, to identify customers, the database is first segmented and then the predictor modeling is applied in the parts.

Techniques, methods and data mining algorithms are methods for implementation of data mining operations. Although each operation has its own strengths and weaknesses, various data mining tools choose the operations on the basis of certain criteria. These criteria include [10]:

Fitting with the type of input data

Transparency of data mining output

Resistance to errors in data values

The accuracy of output

Ability to work with large amounts of data

Techniques dependent on each of the four operations are shown in Table 1.

Table 1. Data mining operations and techniques

Operation	Data Mining Techniques
Predictive modeling	Classification, prediction of value
Database segmentation	Statistical clustering
Link analysis	Discovery of dependence, discovery of sequential patterns, discovery of similar time sequence
Deviation detection	Statistics, visualization of model

The model proposed in this paper is based on data mining of the database of facilities granted to real customers. For this purpose, Dey Bank database has been used. The database of the Bank facilities has been accessible. All the information that users register in the system have been obtained by applying appropriate filters.

Tips on preparing a database for data mining are as follows:

- It is 5 years from the establishment of Dey Bank. After applying the appropriate filters, the total number of facilities the bank has granted since its establishment is 37562 items including all types of facilities for legal and natural customers with current, deferred, liquidated, doubtful and ready for release accounts.
- Given that liquidated and doubtful accounts are the purpose of our modeling, we should examine those facilities whose final status is determined. After applying the appropriate screening, doubtful and liquidated facilities are separated from ready for release (delivery time to the customer), finally 6524 facilities were extracted.
- In this study, only the real customers are addressed of which 2552 items were extracted by applying the filter.

The filters are set as follows:

- Date: 20 March 2010 to 20 December 2015
- Branch: All branches
- Type of applicant: Natural
- Currency: Rial
- Facility Status: liquidated, doubtful
- Main type of loans: beautiful loan, reward, installment sale, civic participation, partnership

The following variables were extracted from the Dey Bank database:

Administration code, gender, age, number of installments, the main type of facilities, description of guarantee types 1 to 8, the economic sector, the amount of loan, interest rate, guarantor.

In the process of data mining for classification problems, different approaches can be used based on the number of variables and the number of categories in final class. It will be very influential on the performance of the final model. If the goal is achieving the least error in modeling the final class and the number of classes in the final class is limited, then linear classification techniques are used. Despite the low error in these models, the predictability of model relies heavily on education data. It will usually include high level of error in perdition, especially if the data space is different from the education environment. But if the ultimate goal is modeling the data set to predict the overall problem space, non-linear classification techniques are used whether the final class is low or high. Although these techniques in modeling the final class show sometimes more error than linear techniques, they do better than linear techniques in the prediction and design of predictive model.

After separation of the final variables from all variables, an appropriate database for data mining process is obtained. Weka software has been used to do data mining process. Since the present research aims to provide a predictive model, the Weka data mining algorithm with nonlinear nature is used. Neural network and decision tree are the best techniques.

The sample of data mining technique based on neural networks on the above data set is shown in Figure 1. In the neural network designed, the learning rate 0.6 and the memory rate 0.2 are applied. This leads to improving the predictability of neural network. The number of neural network epochs was 10000 and the Feed Forward & Feed Backward learning algorithm and is considered.

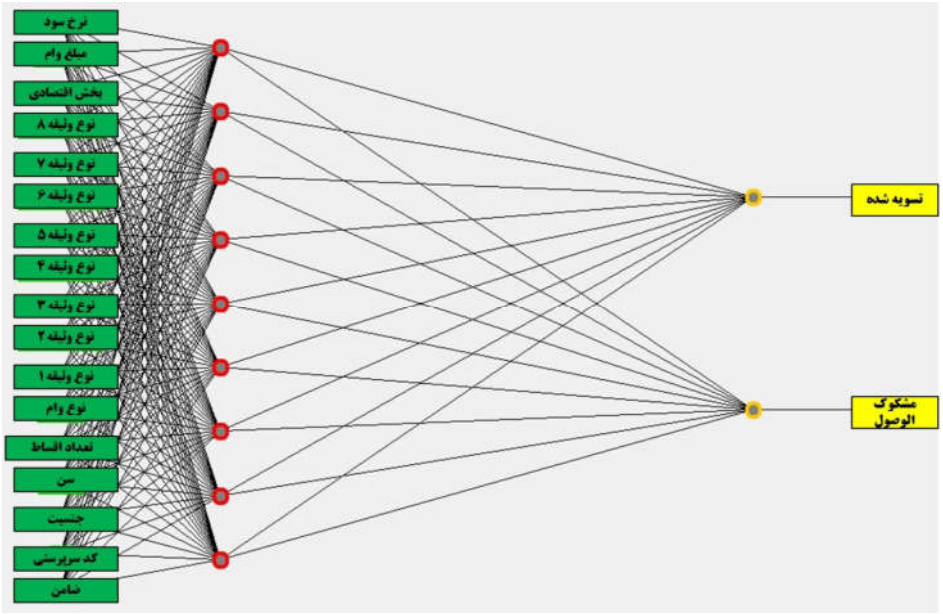

Fig.1. Neural network designed

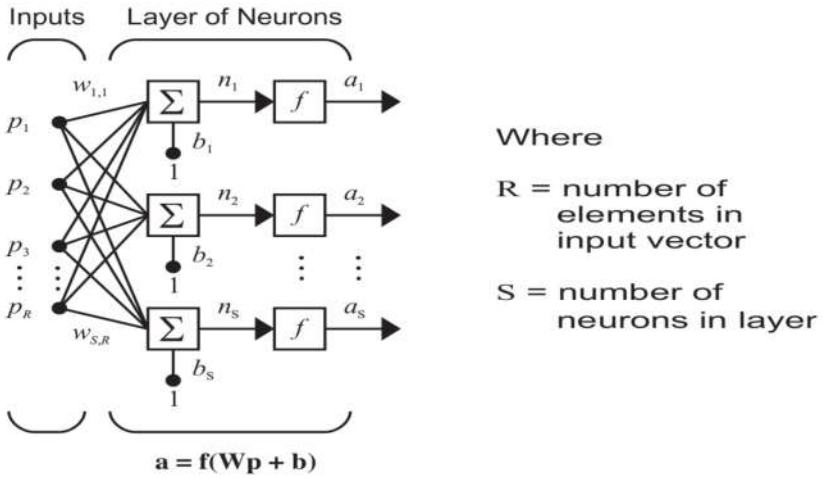

Fig.2. General Structure of the neural network used

The number of nodes in the input layer was 17 and the neural network has a hidden layer with 9 nodes. Log Sigmoid was the transfer function of between the layers, and its function is shown in Figure 3.

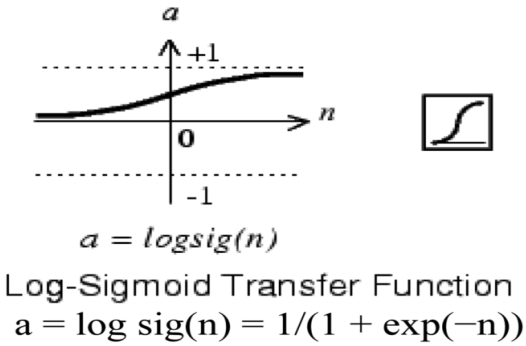

Fig.3. Log Sigmoid transfer function

The sample of the weights assigned to the neural network nodes after the learning process is shown in Figure 4.

```
Sigmoid Node 2
    Inputs    Weights
    Threshold    -3.486786291034266
    Attrib interest rate    8.828441934984665
    Attrib loan amount    3.5023041866504965
    Attrib Economic Section    2.5155700426618886
    Attrib Type of guaranty 8    1.0899643892563693
    Attrib Type of guaranty 7    -1.6017388967899795
    Attrib Type of guaranty 6    -0.8429287876309188
    Attrib Type of guaranty 5    -6.324357847615961
    Attrib Type of guaranty 4    1.848490980653218
    Attrib Type of guaranty 3    -6.03513941093107
    Attrib Type of guaranty 2    1.8605551561887765
    Attrib Type of guaranty 1    3.0626490072171206
    Attrib Loan Type    -5.0124169830048615
    Attrib Number of installments    3.7025442603456846
    Attrib Birth Date    6.780407371460648
    Attrib Gender    0.9184029048343795
    Attrib Area    -6.289092466811125
    Attrib guarantor    5.156066928192624
```

Fig.4. The weights assigned to the neural network nodes

The weights assigned to the neural network nodes in Figure 4 indicate the importance of the variables.

Interest rate: interest rate by a weight factor of 8.82 is very importance for the repayment of loans and it is expected that increased interest rates of the loans will increase the risk of default.

Birth Date (coefficient of 6.78): the applicant's date of birth is very important in validation after the interest rate. Granting the loans to the applicants aging from 35 to 45 has the least risk.

Guarantor (coefficient of 5.15): the facilities where the applicant has a guarantor have less risk of repayment.

Number of installments (coefficient of 5.15): The number of installments is important in terms of the repayment amount, because as the number of installments is higher, the installment amount will be less and thus favorable for the customer.

Type of guarantee: if the customer gives a valuable guarantee to the bank, the probability of default is reduced; because the customer must pay the installments to prevent the bank from capturing the collateral.

All other variables are less important.

There are a lot of criteria to assess the performance of different techniques. But most of the articles use the two values (pd) (False Alarm Rate) and (pf) (Hit Rate) to determine the

probability of correct diagnosis and the likelihood of false detection. These two value are extracted from equations (1) and (2).

Pd = D / (B + D) (1)

Pf = C / (A + c) (2)

Defects		Actual	
		no	yes
Prd	no	A	B
	yes	C	D

In addition to these two values, Mean Absolute Error (MRE) and Root Mean Squared Error (RMSE) can be used to show the performance of two different techniques. Equations (3) and (4) are the methods of obtaining the two values.

(3) $$MRE = \frac{|a_1 - c_1| + |a_2 - c_2| + \ldots + |a_n - c_n|}{n}$$

(4) $$RMSE = \sqrt{\frac{(a_1 - c_1)^2 + (a_2 - c_2)^2 + \ldots + (a_n - c_n)^2}{n}}$$

Where a is the actual output value and c is acceptable output value. These two equations represent technical error in determining the truth or falsity of a code.

To evaluate the success of classification, usually a square matrix with the size of the number of classes is formed in which ij item shows how many data of i^{th} class are in j^{th} class. This matrix is called Confusion (Figure 6). The best case occurs when the matrix is a diagonal matrix, because it shows that no data of i^{th} class is in j^{th} class and all the data of i^{th} class are in their associated i^{th} class. The more similar this matrix is to the diagonal matrix, the classification tool will be more efficient.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        2013               97.7184 %
Incorrectly Classified Instances        47                2.2816 %
Kappa statistic                          0.7038
Mean absolute error                      0.0248
Root mean squared error                  0.1477
Relative absolute error                 26.9976 %
Root relative squared error             69.0382 %
Total Number of Instances             2060
Ignored Class Unknown Instances                1
```

Fig.5. the accuracy of the proposed method

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.996     0.404      0.98      0.996     0.988      0.805    OK
                 0.596     0.004      0.894     0.596     0.715      0.8      NO
Weighted Avg.    0.977     0.385      0.976     0.977     0.975      0.804
```

Fig.5. Table of the accuracy and false positive and negative alerts rate

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1954    7 |    a = OK
   40   59 |    b = NO
```

Fig.6. Confusion Matrix

3. CONCLUSION

In this paper, neural network model was used due to its superiority on linear models.

The most important reasons for not using the linear models are as follows:

1. Linear models are useful in simple models. It means that as the model become more complex and the number of variables increased, the efficiency of model will be reduced.
2. The linear models rely heavily on sample data.
3. The linear models lack machine learning.
4. The linear models are not suitable for qualitative data.

Given the foregoing, the neural network data mining algorithms were used in this article. Data Mining has incorporated the benefits from advances in the field of artificial intelligence that include both principles for pattern diagnosis and classification problems and also the communications made through the application of neural networks and decision trees. Data mining has relatively new algorithms in this area, such as neural networks and decision tree and also new approaches to older algorithms such as separating algorithms.

It is important to note that data mining is used to make these techniques accessible for commercial problems so that these techniques are available for expert users and specialized statistics.

Data mining is getting increasingly popular due to its essential helps. Many organizations are using data mining to help managing all the loans including gaining new customers, increasing interest by retaining good customers. Determining the characteristics of a good customer of our Bank can predict the future with the same characteristics. With recording the customer who buy a particular banking product, Dey Bank may draw its attention to the similar

customers who have not purchased this product. Also, recording the case for customers who have left the organization, Dey Bank may retain those customers who are likely to leave, because retaining an existing customer is less expensive than acquiring a new customer. Data mining suggests values by exploring a wide range of loan recipients.

Telecommunications and credit card companies are two main branches of using data mining to identify fraternal use of their services. Insurance and income companies are also interested in using this technology to reduce frauds. Pharmaceutical applications are another useful area that involved data mining. Data mining can be used to assess the effectiveness of surgical procedures, medical tests and treatment. The companies active in financial businesses use data mining to determine the indicators of market and industry for income efficiency diagnosis. Retailers also use data mining to decide which product is profitable in the stores. Pharmaceutical companies create large databases of chemical compounds and genetic materials in order to discover the materials which could be a good choice for making drug. In this paper, due to the advantages mentioned for data mining, first the data mining was carried out and then modeling was performed.

The most important thing to remember is about modeling that it is an iterative process. You need alternative models to find the most useful way for solving problems. What you learn from searching for an appropriate model can guide you to return and do some changes in the data and even improve the problem statement. When you decided on the kind of prediction you want to do, you have to prepare a model for making your decision. Preparing and testing the data mining model needs the data to be broken into at least two groups: one for preparing the model and the other for testing the model. The accuracy of the model will be different if you use a different preparing and testing method.

Once a model is created and validated, it can be used in two main ways. The first way is for the analyst that introduces practices based on the simple view of the model and its results. The second way is applying the models in different data sets. This model can be used to identify the records based on their classification or giving value to a score, for example the possibility of performing an action. When obtaining a complex application, data mining is often considered as critical but small part of the final project. For example, the knowledge discovered from data mining can be combined with data experts' knowledge and input transactions. In a process diagnostic system, process patterns can be combined with the discovered patterns. When the items assumed for this process are sent to the observers for evaluation, the observers may need to access the records in database which are related to the parts claimed by a generator. In general, the steps described above are necessary to perform

any data mining process. the steps of applying appropriate filters was designed according to the available population and the variables in suitable validation model.

Artificial Neural Network is a data processing system that is inspired by the human brain and assigns data processing to numerous small processors behave like an interconnected network in parallel with each other for solving a problem. In these networks, a data structure is developed using a programming knowledge that can act as neurons. This data structure is called node. After that, the network is trained by creating a network between the nodes and applying a training algorithm. In this memory or neural network, the nodes have two state: active (on or 1) and inactive (off or 0) and each edge (synapses or connections between nodes) has a weight. Edges with positive weight stimulate or activate the next inactive node, and the edges of negative weight deactivate or inhibit the next nodes (if they are enabled).

Neural networks with their remarkable ability to derive results from complex data can be used to extract patterns and detect different subfields that is very difficult to identify for human and computer. The advantages of neural networks can be as follow:

1. **Adaptive learning**: the ability to learn how the tasks are performed based on the information given to it or its initial experiences which is called network reform.
2. **Self-organization**: an artificial neural network automatically organizes and presents the data that received during training. The neurons are compatible with learning principle and the response to input changes.
3. **Real-time operators**: computations in artificial neural network can be performed in parallel using the hardware designed and fabricated to receive optimal results of artificial neural network features.
4. **Error tolerance**: sabotage in the network can decrease the performance but some of its features will continue to be, despite great difficulties.
5. **Categorization**: neural networks are able to categorize the inputs in order to get the right output.
6. **Generalization**: this feature enables the network to derive a general law from limited number of samples and then generalize the things learnt. In the absence of this feature, the system should be able to memorize infinite facts and relationships.
7. **Stability-Flexibility**: a neural network is both stable enough to maintain the information learnt and flexible enough to accept new items without losing the previous data.

Based on the evaluation results shown in Table 1 which includes the best approaches in the field of modeling and predicting the status of loans granted, three data mining approaches, expert system, scoring and rating are raised. The method proposed with data mining approach

using multilayer neural networks has obtained the best performance compared to other approaches with the highest accuracy and least error rate.

Table 1. Comparison and evaluation of different approaches with the technique proposed in this article

No.	Approach	Algorithm	Accurace	Error levle	Database size
1	Data analysis (proposed method)	Neural Networks	97/71	2/29	2071
2	Data Mining (Mirfeizi, 2012)	Probit regression	89/69	10/31	128
3	Expert System (Dahmardeh, 2012)	logistic regression	91/90	8/1	519
4	Expert System (Barzdeh, 2013)	Fuzzy Logic	78/4	21/6	102
5	Scoring and rating (Sedaqat, 2009)	Logit	81/5	18.5	200

In this paper, with data mining of the loans granted by Dey Bank, the effective parameters of real customers on the risk of default was investigated. Also, a model based on multilayer neural networks was proposed to predict the final status of facilities. It became clear after reviewing previous approaches, the data mining approach has higher accuracy and less error than other similar approaches. The parameters of interest rate and applicant's age and number of installments have significant relationship with validation. The results of this study compared with similar studies are more reliable because of the larger size of the database and the more stable data.

4. REFERENCES

1. Mansouri, Ali, (Spring 2003), Designing Mathematical Model to assign Bank credits with classic and Neural Networks approaches, thesis, Tarbiat Modarres University, Faculty of Humanities
2. Abrishami, Hamid. (2002). Econometrics. Tehran: Tehran University
3. Gujarati, D., (2004). "Foundations of econometrics" translated by Hamid Abrishami. Tehran University Publishing and Printing

4. Khosravi, Mehran; Andalib, Shahram (2014), "data mining with Weka". Arena Publishing
5. Basel Committee, (2001). Working Paper On the Internal Rating – Based ApproachTo Spesialised Lending Exposures, October.
6. Elmer, Peter J., and David M. Borowski. (1988). An Expert System and NeuralNetworks Approach To Financial Analysis,Financial Management, No 12, 66-76
7. Glantz, Morton. (2003). Managing Bank Risk, Academic Press.
8. Saunders, A. and Allen, L. (2002). Credit Risk Measurement, Second Edition, New York: John Wiley & Sons.
9. Yang, Z.R. Platt M.B. and Platt H.D. (2001). Probabilistic Neural Networks In Bankruptcy Prediction,Journal of Business Research. Feb, 67-74
10. Gordy Michael. (2001). A Risk - Factor Model Foundation for Rating. Based Bank Capital Rules.

How to cite this article:
Yazdani H M. Developing a model for validation and prediction of bank customer credit using information technology (case study of Dey bank). J. Fundam. Appl. Sci., 2017, *9(1S), 317-330.*