European Online Journal of Natural and Social Sciences 2014;
Vol.3, No.3 pp. 564-571
ISSN 1805-3602

Credit Ranking of Bank Customers (An Integrated Model of RFM, FAHP and K-means)

Roohollah Mohammadi[1], Bijan Bidabad[2], Tahereh Nourasteh[3], Mahshid Sherafati[4*]
[1] Novin Pajoohan Research Institute, Tehran, Iran; [2] Bank Melli Iran; [3] Export Development Bank of Iran; [4] MBA Department, Management Faculty, Multimedia University, Malaysia
*E-mail: mahshidsherafati@yahoo.com

Received for publication: 19 March 2014.
Accepted for publication: 22 July 2014.

Abstract

In this paper, with the aim to rank customers in terms of credit, three patterns namely Hsieh (RFM), FAHP, and K-means were integrated. The main effective factors on ranking customers including transactions, repayment and RFM (Recency, Frequency and Monetary) variables were defined. For classifying the legal customers of Export Development Bank of Iran in terms of credit, 5 variables were extracted from the bank's database and normalized accordingly. The weight of each variable was calculated through interviewing bank experts using FAHP. Using the values of the variables and K-means algorithm, the optimal clusters of customers were determined. Finally, bank customers were ranked in 5 credit clusters and the value of each cluster was estimated.

According to the findings, recency, repayment behavior, transaction, frequency, and monetary variables had maximum effects on customers' ranks, respectively. Therefore, 54% of the customers fell in the third cluster (with cluster value of 0.95) and the fifth cluster (with cluster value of 0.76) composed of good and very good customers. Credit risk of the two clusters (especially the third one) was at least. 32% of the customers positioned in the second cluster (with cluster value of 0.59) including the average customers in terms of credit. 14% of the customers fell in the fourth and first clusters with cluster values of 0.42 and 0.26 including highest risky customers.

Keywords: Credit Ranking, Behavioral Ranking, Hsieh, FAHP, RFM, K-means

Introduction

Banks, finance and credit institutions employ credit applicants' records to rank customers in terms of credit and evaluate the probability of refund of loans fault. Credit and behavioral ranking models are the most common statistical models used in bank's customer ranking (Thomas, 2000). In this regard, these models classify customers into different clusters (Lancher et al., 1995) through cluster multivariate statistics (Morrison, 1990; Hand, 1981; Johnson & Wichern, 1998). K-means is used for ranking new customers in terms of credits using some determinants like age, marital status, income etc. (Chen & Huang, 2003). Banks use ranking method through cluster analysis for predicting future purchasing behavior of customers or their credit status (Setiono et al., 1998).

Credit and behavioral scoring models are used extensively in different banks (Thomas, 2000). By credit and behavioral scoring models customers are classified into different groups. Generally, classification methods are applied to bank databases to identify a new applicant for granting him a credit. On the other hand, behavioral scoring tries to guess the present and future behavior of customers.

Literature Review

RFM

RFM (Recency, Frequency, and Monetary) method was introduced by Hughes as a method for customer evaluation in such a way that customers are differentiated using recency (how recent the last purchase is), frequency (frequency of purchases) and money value of the purchased goods and services (Hughes, 1994). Moreover, R stands for the time span from the last purchase till present time, F for the frequency of purchases in a specific period and M for the nominal purchased value in a defined period (Wang, 2010).

It has been shown in many studies that the higher R and F, the more will be the probability of new transactions with the customer, and the higher M, the more will be the probability of customer return (Wu and Lin, 2005). Moreover, some studies assert that RFM are quite efficient in ranking customers (Newell, 1997). RFM variables have also been used for selecting direct marketing method as expanded RFM model by adding two variables: the time of the first purchase and the probability of abandonment (Yeh et al., 2009). Jonkera used this model to classify customers to find optimal marketing strategy (Jonkera et al., 2004). Some researchers used this model to estimate customer life time value (LTV) (Ramzi and Ghanbari 2009; Sohrabi and Khanlari, 2007; Liu and Shih, 2005). However, this model is extensively used for ranking customers (Hsie, 2004) and (Ghazanfari et al., 2010).

K-means algorithm

Clustering method is used to categorize objects into clusters of similar objects (Han & Kamber, 2001). K-means is a clustering algorithm (Forgy, 1965) based on the mean value of the objects within the clusters. That is the objects should be partitioned in a way that the mean values of certain variables or attributes of the objects within the clusters have nearest distances to the means of the variables in the cluster, MacQueen (1967). In this study K-means is used to build clusters of customers through RFM variables. To compute K-means the following steps are to be taken. First, all m objects are portioned into K initial clusters - K initial seed centroids values could also be specified. Then, those objects whose their Euclidean distances are nearest to the mean of cluster are assigned to that cluster. Centroid is recalculated again for the clusters with new item or lost item. The process is repeated until there would be no more reassignment (Cheng & Chen, 2009).

Fuzzy numbers

An ordinary number $\ddot{a}$ can be shown as fuzzy through membership function as generalized ordinary numbers:

$$\mu_{\ddot{a}}(x) = \begin{cases} 1 & ;if\ x = a \\ 0 & ;if\ x \neq a \end{cases} \quad (1)$$

That is real number has been transformed into fuzzy number. Triangular fuzzy numbers as simplest one are defined as M on R when the membership function ($\mu_{\ddot{a}}(x): R \to [0\ ,1]$) is defined as (Jafari, Bidabad, Mohammadi, 2010):

$$\mu_{\ddot{a}}(x) = \begin{cases} \dfrac{x}{m-l} - \dfrac{l}{m-l'} & x \in [l\ ,m] \\ \dfrac{x}{m-u} - \dfrac{u}{m-u'} & x \in [m\ ,u] \\ 0 & otherwise \end{cases} \quad (2)$$

Thus, triangular fuzzy numbers is shown by parameters of “l” (smallest possible value), “m” (most promising value) and “u” (largest possible value) to express a fuzzy event as (l, m, u)

(Ertugrul and Karakasoglu, 2009). For the two positive triangular fuzzy numbers of (l_1, m_1, u_1) and (l_2, m_2, u_2) we have the following two operations we use more in this study:

$$(l_1, m_1, u_1).(l_2, m_2, u_2) = (l_1.l_2, m_1.m_2, u_1u_2) \quad (3)$$

$$(l_1, m_1, u_1)^{-1} \approx \left(\frac{1}{u_1}, \frac{1}{m_1}, \frac{1}{l_1}\right) \quad (4)$$

FAHP

As stated by Saaty (1980) AHP compares qualitative or quantitative variables through pairwise comparisons. That is a hierarchical decision tree is formed, and its indices and options are determined. Then, pairwise comparisons of each factor with rival factor are to be done to determine the weight of factors.

If exact values are to be assumed for various options of a decision-maker the process is called traditional AHP (Wang and Chen 2007). This process causes some problems when the objects are discordant and pairwise comparisons cannot be done in uncertainty state (Deng, 1999). In order to remove this shortage, in Fuzzy AHP (FAHP) the distance judgment is replaced by spot judgment (Kahraman et al., 2003).

Let X be a set of objects {x_1, ... , x_n}, and G a set of goals {g_1, ... , g_n}. In Extent FAHP (Chang, 1996), for each object Extent Analysis is done for each goal one by one and the m Extent Analysis values are obtained for each object with the signs of M^1_{gi}, …, M^m_{gi} for all i=1,…,n where M^j_{gi} (j=1,…,m) all are triangular fuzzy numbers. The Extent Analysis of Chang (1996) is performed through the following steps:

Step 1. Building fuzzy synthetic extent values for the object i as:

$$S_i = \sum_{j=1}^{m} M_{gi}^{j} \otimes gi \left[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{gi}^{j}\right]^{-1} \quad (5)$$

The symbol $\otimes$ is for extended multiplication In order to obtain $\sum_{j=1}^{m} M_{gi}^{j}$, the Fuzzy addition operation of $\sum_{j=1}^{m} M_{gi}^{j}$will be:

$$\sum_{j=1}^{m} M_{gi}^{j} = \left(\sum_{j=1}^{m} l_j, \sum_{j=1}^{m} m_j, \sum_{j=1}^{m} u_j\right) \quad (6)$$

Inside the bracket of (5) is derived through following fuzzy addition as:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} M_{gi}^{j} = \left(\sum_{i=1}^{n} l_i, \sum_{i=1}^{n} m_i, \sum_{i=1}^{n} u_i\right) \quad (7)$$

Step 2. For the triangular fuzzy numbers $M_1 = (l_1, m_1, u_1)$ and $M_2 = (l_2, m_2, u_2)$, the degree of possibility of $M_2 \geq M_1$ will be:

$$V(M_2 \geq M_1) = \begin{cases} 1 & if\ m_2 \geq m_1 \\ 0 & if\ l_1 \geq u_2 \\ \dfrac{l_1 - u_2}{(m_2 - u_1) - (m_2 - l_1)} & otherwise \end{cases} \quad (8)$$

Values of degrees of possibilities of $V(M_1 \geq M_2)$ and $V(M_2 \geq M_1)$ are used for M_1 and M_2 comparison. The intersection of the two triangular fuzzy numbers of M_1 and M_2 is depicted by Figure 1. In this figure the value of d is highest at point D which is at the point of intersection of μ_{M1} and μ_{M2}.

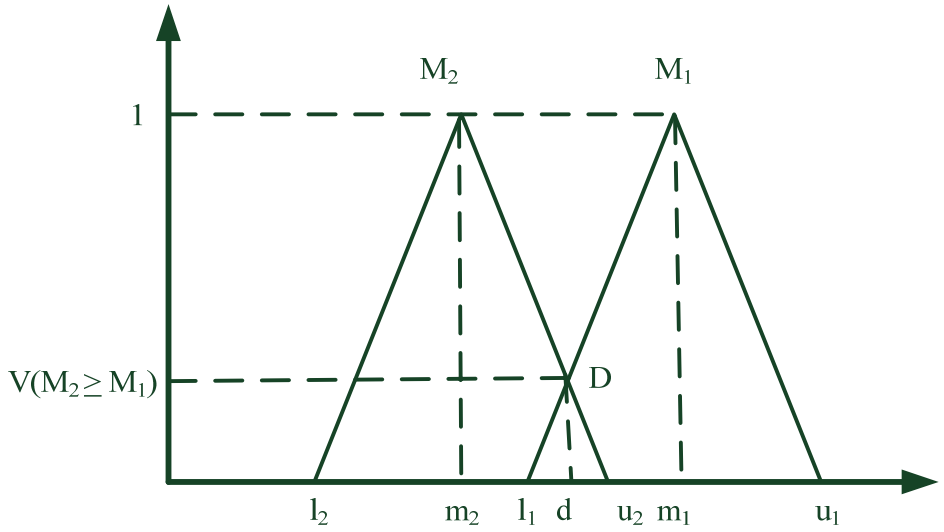

Figure 1: Possibility degree of $M_1 \geq M_2$

Figure 1. Intersection of two Fuzzy numbers ($M_1 \geq M_2$) and possibility degree of $M_1 \geq M_2$

Step 3. For a set of k convex fuzzy numbers M_i for i = 1,..., k, the degree of possibility will be greater than all k numbers if:

$$V(M \geq M_1, \ldots, M_K) = V[(M \geq M_1), \ldots, (M \geq M_K)] = \min V(M \geq M_i),\ i=1,\ldots,k \qquad (9)$$

If $d(A_i)$ were defined as minimum of all degrees of possibilities for k convex fuzzy numbers M_i for i = 1,..., k ($d(A_i) = \min V(S_i \geq S_k)$; k =1,..., n; k≠ i) then the vector of weights n elements A_i for i=1,…,n will be:

$$w' = \left(d'(A_1), \ldots, d'(A_x)\right)^T \qquad (10)$$

Step 4. The normalized vectors of non-fuzzy weight will be derived after normalization as:

$$w = \left(d(A_1), \ldots, d(A_n)\right)^T \qquad (11)$$

Methodology

RFM model was used to rank legal persons customers of Export Development Bank of Iran in terms of credit. In RFM ranking, financial perspectives are of less relevance and the main emphasis is on quality issues. In this model, the volume, frequency and recency of transactions are taken into consideration, which are not directly related to customers' yields. In addition to RFM variables, transaction data and repayment behavior were also used. 296 customers of the bank in 2009 were used as the sample and 5 variables of were selected for the model as follows:

R variable stands for Recency in RFM and shows the time between the last exportation of the customer till the time of study. In fact, R shows the last time in the respective span in which the

customer had exported some goods or services. F is used for frequency of exports in the respective time span as a RFM variable. M is the total value of exports in dollars in the respective time span as monetary variable in RFM. The transaction data is related to the number of customer's transactions in the respective time span and repayment behavior is related to the number of the delayed debts to Export Development Bank or other banks in Iran. Variables were extracted from the Export Development Bank's database and normalized accordingly. The weights of variables were calculated using FAHP and finally, a value for every customer was estimated. After then, the bank's customers were ranked using K-means.

To estimate the weight of each variable, a questionnaire based on FAHP logic distributed among 30 bank experts. In this questionnaire, variables were compared pairwise. After converting the linguistic variables into triangular fuzzy numbers, the mean of the collected data for 30 samples was calculated and the extent fuzzy data was assembled. Table 1 shows the summary result of this step.

Table 1 - Initial pairwise comparison matrix after integrating 30 data points

Variable	R	F	M	TD	RB
R	(1,1,1)	(1,2.33,3)	(3,3.67,5)	(0.33,4.11,7)	(0.14,4.05,7)
F	(0.33,0.55,1)	(1,1,1)	(0.33,1.44,3)	(0.14,1.78,5)	(0.2,1.18,3)
M	(0.2,0.28,0.33)	(0.33,1.44,3)	(1,1,1)	(0.14,1.11,3)	(0.14,1.76,5)
TD	(0.14,1.11,3)	(0.2,4.07,7)	(0.33,4.11,7)	(1,1,1)	(3,3,3)
RB	(0.14,2.45,7)	(0.33,2.78,5)	(0.2,4.73,7)	(0.33,0.33,0.33)	(1,1,1)

RB, TD, M, F, and R stand for repayment behavior, transaction, monetary, frequency and recency, respectively. After integration of the data, S_i vectors were calculated according to (5) as shown below:

$$\begin{aligned}
S_1 &= (5.47,15.16,23) \otimes \left(\frac{1}{89.66},\frac{1}{51.28},\frac{1}{16.06}\right) = (0.06,0.29,1.43)\\
S_2 &= (2,5.95,13) \otimes \left(\frac{1}{89.66},\frac{1}{51.28},\frac{1}{16.06}\right) = (0.02,0.12,0.81)\\
S_3 &= (1.81,5.59,12.33) \otimes \left(\frac{1}{89.66},\frac{1}{51.28},\frac{1}{16.06}\right) = (0.02,0.11,0.77)\\
S_4 &= (4.67,13.29,21) \otimes \left(\frac{1}{89.66},\frac{1}{51.28},\frac{1}{16.06}\right) = (0.05,0.26,1.31)\\
S_5 &= (2,11.29,20.33) \otimes \left(\frac{1}{89.66},\frac{1}{51.28},\frac{1}{16.06}\right) = (0.02,0.22,1.26)
\end{aligned} \tag{12}$$

Next, the possibility degrees (of S_i vectors) in comparison with others were calculated as described by (8). Then, d(I)s were calculated as follows:

$$\begin{aligned}
d'(I1) &= MIN(S_1 \geq S_2, S_3, S_4, S_5) = MIN(1,1,1,1) = 1\\
d'(I2) &= MIN(S_2 \geq S_1, S_3, S_4, S_5) = MIN(0.82,1,0.84,0.89) = 0.82\\
d'(I3) &= MIN(S_3 \geq S_1, S_2, S_4, S_5) = MIN(0.8,0.99,0.83,0.87) = 0.8\\
d'(I4) &= MIN(S_4 \geq S_1, S_2, S_3, S_5) = MIN(0,98,1,1,1) = 0.98\\
d'(I5) &= MIN(S_5 \geq S_1, S_2, S_3, S_4) = MIN(0.94,1,1,0.97) = 0.94
\end{aligned} \tag{13}$$

The final matrix was estimated as follows:

$W' = (1, 0.82, 0.8, 0.98, 0.94)^T$ (14)

$W = (0.22, 0.18, 0.176, 0.216, 0.21)$ (15)

Thus, based on FAHP, variables were prioritized as shown by Table 2.

Table 2 – The final matrix for prioritization of the model variables using FAHP

Criterion	Weight
Recency	0.220
Repayment behavior	0.216
Transaction data	0.210
Frequency	0.180
Monetary	0.176

Having determined the weights of indices, in the next step, the model variables' values were normalized using the following formula:

$$y_{ci} = \frac{X_{ci} - X_{Min}}{X_{Max} - X_{Min}} \quad (16)$$

Where, X_{ci} is the main value of variables for i^{th} customer. X_{min} and X_{max} are the minimum and maximum values of each variable among all sample customers, respectively. Then, using K-means algorithm the 293 cases were clustered. At first, number of optimal clusters was estimated as shown by Table 3.

Table 3 - Results of K-means clustering to determine the number of clusters

Number of clusters	K-means
2	0.2204839
3	0.1487382
4	0.1291029
5	0.1098234
6	0.1319821

Number of optimal clusters was 5. Next, the value of each customer on the basis of the 5 variables was calculated as follows:

$$V(C_i) = W_1 \times R(C_i) + W_2 \times F(C_i) + W_3 \times M(C_i) + W_4 \times TD(C_i) + W_5 \times RB(C_i) \quad (17)$$

In which $RB(C_i)$, $TD(C_i)$, $M(C_i)$, and $R(C_i)$ are repayment behavior, transaction, monetary, frequency and recency respectively. W_i represents the weight of the variables already estimated using FAHP. All the customers were ranked into 5 clusters. The results are shown by Table 4.

According to the results, in 2009, the legal customers of the bank were divided into 5 groups of customers: the third cluster (with a cluster value of 0.95) added to the fifth cluster (with a cluster value of 0.76), totally 162 companies (i.e. 54% of customers), included as good and very good customers of the bank. Credit risk of the two groups (especially the third cluster) was at least. The second cluster included 96 companies (32% of customers with a cluster value of 0.59) were average customers of the bank. The fourth and first clusters with cluster values of 0.42 and 0.26 included

high risk bank customers (14% of customers). Payment of loans to the latter two groups, especially the first cluster indicates a high probability of faults.

Table 4 - Value of clusters

Number of clusters	Average of R	Average of F	Average of M	Average of RB	Average of TD	Value of each cluster	Cluster members
3	0.82	0.91	1	1	1	0.95	34
5	0.57	0.83	0.78	0.72	0.88	0.76	128
2	0.38	0.69	0.64	0.62	0.64	0.59	96
4	0.24	0.54	0.56	0.52	0.31	0.42	20
1	0.19	0.47	0.34	0.31	0.19	0.26	18

Conclusion

One of the problems in analyzing bank customers' data is that the information is multi-dimensional. Since behavioral and credit models include two main aspects of customers' behaviors, are amongst the most successful models in evaluating bank customers.

In this paper, credit and behavioral models were integrated and used together with K-means algorithm for ranking bank customers in terms of credit. Since in RFM ranking, financial perspectives are of least value, in this paper, the transaction data and repayment behavior were both included in investigation. The results of the suggested method, considering its simple application and multi-dimensionality of information, can help banks and finance and credit institutions to rank their customers purposefully.

References

Chang, D. Y. (1996) Applications of the extent analysis method on fuzzy AHP. European Journal of Operational Research, 95, 649–655.

Chen, M. C., Huang, S. H. (2003) Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications, 24, 433–441.

Cheng, C.H., You-Shyang Chen. (2009) Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications 36, 4176–4184.

Deng, H. (1999) Multicriteria analysis with fuzzy pair-wise comparison. International Journal of Approximate Reasoning, 21, 215–231.

Dirk Van den Poel, Bart Larivi, (2004), Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research. 157, pp. 196–217.

Ertugrul, Irfan, Nilsen Karakasoglu (2009), Performance evaluation of Turkish cement firms with fuzzy analytic hierarchy process and TOPSIS methods. Expert Systems with Applications 36, pp. 702–715.

Forgy, E. (1965) Cluster analysis of multivariate data: Efficiency versus interpreability of classifications. Biomertrics, 21, 768.

Ghazanfari, M., Malik Mohammadi, S., Alizadeh S., Fathollah, M. (2010) Categorization of customers of exporting eatable fruits. Trade Researches (Pajoohesh'haye Bazargani), No. 55, summer, pp. 151-181.

Han, J., Kamber, M. (2001) Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

Hand, D. J. (1981) Discrimination and classification. New York: Wiley.

Heckerman, D. (1996) Bayesian networks for knowledge discovery. Advances in knowledge

discovery and data mining, pp. 273–305.
Hsieh, Nan-Chen (2004) An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Systems with Applications 27, pp. 623–633.
Hughes, A. M. (1994) Strategic database marketing. Chicago: Probus Publishing Company.
Hyunseok Hwang, Taesoo Jung, Euiho Suh (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26, pp. 181–188.
Jafari samimi, A., Bidabad, B., Mohammadi, R. (2010) Simulation of continuous qualitative variables in econometric models using fuzzy functions and numbers. Australian Journal of Basic and Applied Sciences, 4(10): 4780-4791.
Johnson, R. A., Wichern, D. W. (1998) Applied multivariate statistical analysis (4th ed.). Upper Saddle River, NJ: Prentice-Hall.
Jonkera, J-J., Piersmab, N. and Van den Poelc, D. (2004); Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert Systems with Applications, Vol. 27, pp.159–168.
Kahraman, C., Cebeci, U., Ulukan, Z. (2003) Multi-criteria supplier selection using fuzzy AHP. Logistics Information Management, 16(6), 382–394.
Lancher, R. C., Coats, P. K., Shanker, C. S., Fant, L. F. (1995) A neural network for classifying the financial health of a firm. European Journal of Operational Research, 85(1), 53–65.
Liu, D,. Ya-Yueh Shih (2005) Integrating AHP and data mining for product recommendation based on customer lifetime value, Information and Management 42(3), 387-400.
MacQueen, J. B. (1967) Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley symposium on mathematical statistics and probability, pp. 281–297. Berkeley: University of California Press.
Morrison, D. F. (1990) Multivariate statistical methods. New York, NY: McGraw-Hill.
Newell, F. (1997) The new rules of marketing: how to use one-to-one relationship marketing to be the leader in your industry. New York: McGraw-Hills Companies Inc.
Ramzi, J., Ghanbari A. (2009) Proposing a new model to calculate the value of customer life period. Quarterly Journal of Information Technology Management, Vol. 1, No. 2, spring and winter, pp. 35-80.
Saaty, T. L. (1980) The analytic hierarchy process. New York: McGraw- Hill.
Setiono, R., Thong, J. Y. L., Yap, C. S. (1998) Symbolic rule extraction from neural networks, an application to identifying organizations adopting IT. Information and Management, 34(2), 91–101.
Sohrabi, B. Amir Khanlari, (2007), Customer Lifetime Value (CLV) measurement based on RFM Model. Iranian Accounting and Auditing Review, Spring,Vol. 14 No. 47, pp. 7- 20.
Thomas, L. C. (2000) A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16, 149–172.
Wang, CH,. (2010) Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. Expert Systems with Applications, 37: 8395-8400.
Wang, T. C., Chen, Y. H. (2007) Applying consistent fuzzy preference relations to partnership selection. Omega, the International Journal of Management Science, 35, 384–388.
Wu, J., Lin, Z. (2005) Research on customer segmentation model by clustering. ACM International Conference Proceeding Series, p. 113.
Yeh, C., Yang, K. and Ting, T. (2009) Knowledge discovery on RFM model using Bernoulli sequence. Expert Systems with Applications, Vol. 36, pp. 5866–5871.