Estimating customer future value of different customer segments based on adapted RFM model in retail banking context

Mahboubeh Khajvanda, Mohammad Jafar Tarokhb a *

a Post graduate student, IT Group - Faculty of Industrial Engineering K. N. Toosi University of Technology, Tehran, Iran
b Associate professor, IT Group - Faculty of Industrial Engineering K. N. Toosi University of Technology, Tehran, Iran

Abstract

One of the important challenges in customer-based organizations is customer cognition, understanding difference between them, and ranking them. Customer need-based segmentation was common in past years, but recently customer value as a quantifiable parameter could be used for customer segmentation. In this regard, customer segmentation based on customer lifetime value (CLV) and estimating the value of each segment would be useful for making decision in marketing and customer relationship management (CRM) program which can be adapted with the characteristics of each segment. Customer future value as a part of customer lifetime value can be estimated based on customer segmentation. In this regard, this study provide a framework for estimating customer future value based on adapted weighted RFM analysis which is a CLV calculating model, for each segment of customer in retail banking scope.

Keywords: Customer lifetime value; Customer segmentation; Data mining; RFM analysis; Time series; ARIMA model

1. Introduction

CLV is a concept in customer relationship management scope which was defined more than 30 years ago by Kotler as “the present value of the future profit stream expected over a given time horizon of transacting with the customer” [1]. There are several model for estimating CLV in past researches which were considered by Gupta et al. [2] and Jain and Singh [3], but in retail banking environment, a model to determine CLV should satisfy some conditions that two important of them are [4]: First, in order to be easily implementable, it should focus on the valuation of homogeneous segments of customers instead of individual clients. Second, it needs to be easy to understand and parsimonious in nature to ensure its applicability in many business contexts. This specifically implies limiting data requirements to the information available in an average bank’s information system.

In this regard, in current study customer segmentation methods were exploited to group customers in to homogeneous segments and for satisfying last condition, one of the most powerful and simplest models to implement CRM and estimating customer value which is RFM model – Recency, Frequency, and Monetary value [5, 6,7] was used.

Current study provides a framework for segmenting customers in homogenous segments, calculating CLV value of different segments and finally estimating customer future value of each segment with analyzing trend of customer value in different seasons by time series method. The rest of this study is organized as follows. Section 2 describes...
the framework, and section 3 explains a case study and empirical results. Finally, section 4 draws conclusions and summarizing the contributions of this work.

2. Framework of estimating future value of different customer segments

Current study provides a framework for segmenting customers in homogenous segments, calculating CLV value of each segment and finally estimating future value of each segment with analyzing trend of customer value in different seasons by time series method. Fig. 1 represents the framework.

3. Empirical results

This study as it is shown in Fig 1 is summarized in seven steps:
Step 1: In the first step, data was collected from an Iranian private bank in retail banking scope. The data was row data of fifty hundreds customers’ transactions from the first of the spring 2008 to the end of the summer 2009 which was conclude customer ID, type of deposit, serial of customer deposit, branch code, transaction date, the amount of balance before transaction and the amount of transaction.

Step 2: The data which was received in step one, divided to six parts based on seasonal division. Then RFM parameters were extracted for each customer. Table 1 shows definition of basic RFM model parameters and adapted RFM parameters in this study.
Table 1. Definition of RFM and Adapted RFM model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Adapted parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency (R)</td>
<td>last purchase date in a particular period</td>
<td>Interval between the time of the last transaction and first day of each season</td>
</tr>
<tr>
<td>Frequency (F)</td>
<td>Number of purchases in a particular period</td>
<td>Number of days which occur a transaction during each season</td>
</tr>
<tr>
<td>Monetary (M)</td>
<td>value of purchases in a particular period</td>
<td>Daily average amount of money in all of the customer’s deposits during each season</td>
</tr>
</tbody>
</table>

Step 3: In this step, customer segmentation was done with different clustering algorithms and k-means algorithm was chosen as the best algorithm by Dunn index. This index introduces optimum number of clusters in a clustering algorithm and also the best algorithm in segmenting data. The main goal of this measure is to maximize intercluster distances (distance between different clusters), whilst minimizing intracluster distances (distance between members of a cluster) [8]. For any partition \( C = \{ C_1, C_2, ..., C_k \} \), where \( C_i \) represents the \( i^{th} \) cluster of such partition, the Dunn index, \( D \), is defined as in equation (1):

\[
D(C) = \frac{\min_{i=1, ..., k} \max_{j \neq i} \delta(C_i, C_j)}{\max_{i=1, ..., k} \Delta(C_i)}
\]

(1)

Where \( \delta(C_i, C_j) \) defines the distance between clusters \( C_i, C_j \) (intercluster distance), and \( \Delta(C_i) \) represents the intracluster distance of cluster \( C_i \) or the size of the cluster \( C_i \) and \( k \) is the total number of clusters. In this study:

\[
\delta(C_i, C_j) = d(\bar{C}_i, \bar{C}_j)
\]

(2)

\[
\Delta(C_i) = \max_{x \in C_i} [d(x, \bar{C}_i)]
\]

(3)

where \( \bar{C}_i \) and \( \bar{C}_j \) are the centroid of cluster \( C_i \) and \( C_j \). Thus large Value of \( D \) corresponds to good clusters. Therefore, the number of cluster that maximizes \( D \) is taken as the optimal number of clusters, also the algorithm which has the maximize \( D \) is the best clustering algorithm.

The clustering algorithms which are used in this study include k-means [9], Two step algorithm and x-means algorithm [10]. Two step algorithm and x-means algorithm don’t need the exact number of clusters which was defined by user while they find the optimum number of cluster in a range which defined by user. The defined range in this study for both of this algorithm was 3 to 10. Both of algorithms chose 4 clusters as optimum number. But k-means algorithm was executed 8 times and Dunn index was calculated in each time, because this algorithm need . The optimum number of cluster (K) was obtained 4(See Fig. 2).

![Fig. 2. Optimum number of clusters](image-url)
Table 2. Result of Dunn index in clustering algorithms

<table>
<thead>
<tr>
<th></th>
<th>k-means</th>
<th>Tow step</th>
<th>x-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dunn index value</td>
<td>0.427</td>
<td>0.334</td>
<td>0.254</td>
</tr>
</tbody>
</table>

Step 4: K-means algorithm is sensitive about outlier data, so anomaly detection method was used to solve this problem and then the data was clustered again in each season. For example, the result of customer segmentation in summer 2009 by k-means algorithm after anomaly detection is shown in table 3.

Table 3. Clustering customer in summer of 2009

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Percent of customer</th>
<th>Recency (R)</th>
<th>Frequency (F)</th>
<th>Monetary (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>44</td>
<td>86.5</td>
<td>9.8</td>
<td>49498711</td>
</tr>
<tr>
<td>c2</td>
<td>9</td>
<td>24.1</td>
<td>2.1</td>
<td>3921154</td>
</tr>
<tr>
<td>c3</td>
<td>11</td>
<td>91.1</td>
<td>38.3</td>
<td>118829402</td>
</tr>
<tr>
<td>c4</td>
<td>36</td>
<td>64.2</td>
<td>4.4</td>
<td>24341737</td>
</tr>
</tbody>
</table>

Step 5: For calculating CLV value of each segment, normalized form of adapted RFM parameters were needed. Thus data was normalized by Min-Max normalization method, as it is shown in equation (4) (Tan et al., 2006):

\[ v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} \] (4)

where \( \text{min}_A \) and \( \text{max}_A \) are the minimum and maximum values of an attribute, \( A \). Then min-max normalization maps a value, \( v \), of \( A \) to \( v' \) in the range of \([\text{newmin}_A, \text{newmax}_A]\).

Step 6: After normalizing RFM parameters, we calculated CLV score of each cluster based on weighted RFM model as follow in equation (5):

\[ CLV_{ci} = N_R_{ci} \times W_R + N_F_{ci} \times W_F + N_M_{ci} \times W_M \] (5)

where \( N_{R_{ci}} \) refers to normalized Recency of cluster \( ci \), \( W_R \) is the weight of Recency, \( N_{F_{ci}} \) is normalized Frequency, \( W_F \) is the weight of Frequency, \( N_{M_{ci}} \) is normalized Monetary, and \( W_M \) is the weight of monetary.

According to the assessments obtained by the AHP method based on expert people idea, the relative weights of the RFM variables are as follow: \( W_R = 0.081 \), \( W_F = 0.317 \) and \( W_M = 0.602 \). In table 4 the normal R, F and M values for each cluster’s centroid and CLV of them is shown for the summer of 2009.

Table 4. Calculating CLV value of each cluster in summer of 2009

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Percent of customer</th>
<th>Normalized R</th>
<th>Normalized F</th>
<th>Normalized M</th>
<th>CLV Value</th>
<th>CLV Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>44</td>
<td>0.929</td>
<td>0.103</td>
<td>0.015</td>
<td>0.117</td>
<td>CLV2</td>
</tr>
<tr>
<td>c2</td>
<td>9</td>
<td>0.251</td>
<td>0.012</td>
<td>0.001</td>
<td>0.025</td>
<td>CLV4</td>
</tr>
<tr>
<td>c3</td>
<td>11</td>
<td>0.979</td>
<td>0.434</td>
<td>0.036</td>
<td>0.238</td>
<td>CLV1</td>
</tr>
<tr>
<td>c4</td>
<td>36</td>
<td>0.687</td>
<td>0.039</td>
<td>0.007</td>
<td>0.072</td>
<td>CLV3</td>
</tr>
</tbody>
</table>

In this regard, customers in cluster C3 which are 11% of all customers, with the most CLV value are the best and gold customer. In the other side, customers in cluster C2 with the least CLV value are the most invaluable customers. The rank of each segment of customers also can be seen in table 4.

Step 7: After clustering customers in each season in previous step, in this step we want to predict future CLV value of each segment based on six recent seasons. The trends of CLV value in different seasons in each segment are shown in table 5 and Fig. 3.

Table 5. CLV value of different segments during six seasons

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CLV1</td>
<td>0.163</td>
<td>0.248</td>
<td>0.286</td>
<td>0.225</td>
<td>0.274</td>
<td>0.258</td>
</tr>
<tr>
<td>CLV2</td>
<td>0.087</td>
<td>0.120</td>
<td>0.132</td>
<td>0.113</td>
<td>0.131</td>
<td>0.126</td>
</tr>
<tr>
<td>CLV3</td>
<td>0.047</td>
<td>0.077</td>
<td>0.083</td>
<td>0.075</td>
<td>0.081</td>
<td>0.078</td>
</tr>
<tr>
<td>CLV4</td>
<td>0.015</td>
<td>0.020</td>
<td>0.022</td>
<td>0.025</td>
<td>0.025</td>
<td>0.027</td>
</tr>
</tbody>
</table>
For estimating customer future value, the multiplicative seasonal ARIMA -Auto Regressive Integrated Moving Average- which is a time series prediction method in nonstationary status and reflect the seasonal behaviour of the process, was exploited. The multiplicative seasonal ARIMA\((p,d,q)\times(P,D,Q)\_S\) model where \(p\) is the order of auto regressive process, \(d\) is the order of differencing operator, \(q\) is the order of moving average process, also \(P\) is the order of seasonal auto regressive process, \(D\) is the order of seasonal differencing operator, \(Q\) is the order of seasonal moving average process, can be represented by:

\[
\varphi(B)\varphi(B)\varphi(B)^D\varphi(B)^d x_t = \theta(B)\theta(B)\varepsilon_t
\]  

(6)

Where \(\varphi(B)\) is auto regressive process, \(\theta(B)\) is moving average process and \(\varphi(B)^d\) is d-fold differencing operator which is used to change a nonstationary time series to a stationary one. \(\varphi(B)^S\) is seasonal auto regressive process, \(\theta(B)^S\) is the seasonal moving average process and \(\varphi(B)^D\) is D-fold differencing operator. The ARIMA equation which was obtain for segments CLV1, CLV2 and CLV3 is ARIMA\((0,0,1)\times(1,1,1)\times1\), thus based on equation (6):

\[
(1-\phi_1 B)x_t = (1-\theta_1 B)(1-\theta_2 B)s_t + \theta(B)\varepsilon_t
\]  

(7)

\[
(1-\phi_2 B)(x_t - x_{t-1}) = \varepsilon_t - \theta_1 s_t - \theta_2 s_{t-1} + \theta_1 \theta_2 B^2 s_t
\]  

(8)

\[
x_t = (1+\phi_2) x_{t-1} - \phi_3 x_{t-2} + ts_t - (\theta_1 \theta_2 + \theta_2 \theta_3) s_{t-1} + \theta_2 \theta_3 s_{t-2}
\]  

(9)

The ARIMA equation which was obtain for segments CLV4 is ARIMA\((1,0,1)\times(0,0,2)\times2\), thus based on equation (6):

\[
\varphi(B)\varphi(B)^S\varphi(B)^D x_t = \theta(B)\theta(B)\varepsilon_t
\]  

(10)

\[
(1-\phi_1 B)^S(1-\phi_2 B)^D x_t = (1-\theta_1 B^2 - \theta_1 B^4)(1-\theta_3 B^4)x_t
\]  

(11)

\[
x_t = \varphi(B) x_{t-1} + \varepsilon_t - \theta_2 s_{t-1} - \theta_2 s_{t-2} + \theta_1 \theta_2 \theta_3 \varepsilon_{t-1} + \theta_2 \theta_3 \varepsilon_{t-2} + \theta_2 \theta_3 \varepsilon_{t-4} + \theta_2 \theta_3 \varepsilon_{t-6}
\]  

(12)

The outputs of software for forecasting of CLV value of each segment in next season are represented in Fig. 4. The Fig.4 show that the value of less valuable segment (CLV4) will increment in future. The CLV value of segments CLV1, CLV2 and CLV3 can be predicted just for next season because of lack of data which is the limitation of this study.
4. Conclusion

It is so important and vital for firms to know different types of customers to make decision more profitability. Knowing customers one by one in some business is not commodious, so using customer segmentation method could be useful in this situation. Customer segmentation based on their value or their lifetime value represents instead of customer need-based segmentation nowadays. Customer lifetime value includes calculating past and present value of the customers and predicting of future value of the customers. This study prepares a framework for segmenting customers based on their value for estimating future value of different segments of customer in an Iranian private bank in retail banking scope. The result of this study could be as a guideline for marketing strategies, developing and cross selling of new products for each group, and developing of private banking for the most valuable customer group.

Reference