International Journal of Industrial Engineering and Technology.
ISSN 0974-3146 Volume 7, Number 1 (2015), pp. 39-50

http://www.irphouse.com

Evaluation and segmentation based on customer lifetime value: case study

Sahar Ghoreishi and Mohammad Jafar Tarokh

Industrial Engineering Department
K.N Toosi University of technology

Abstract

In last decades firms which have directly or indirectly contact with customer migrate from Product-oriented to be customer-oriented; hence some of products and customers are not profitable in the same way some of them bring detriment to firm. In this regard, firms should recognize loyal, profitable and potential customers which bring added value for them. In order to distinguish the profitable customers they supposed to cluster customers and study their behavior's group for the sake of having best investment on best segment. In this paper we clustered customers and fended each cluster rule base on customer value in a bank. Additionally, it's proved that participation bonds are the factor that affects clustering of RFM attribute result.

Key words Data mining, clustering algorithms, segmentation, CLV

1. Introduction

Along with the development of the financial business, the competitive mode of bank "to the product as the center" has been gradually replaced by the new business model "to the customer as the center", Customers become the most important resource for the bank, and to pay attention to the customer management, mining the customer value is development of banking core power [9]. The customer relationship management and customer value management have become the core motive of commercial banks developing [8]. In this regard, to manage the customers and their value we need some methods for evaluating in order to make a suitable decision. We could use data mining methods for customer clustering and finding patterns by classification algorithms.

2. Literature review

2.1. CRM

CRM is necessary for banks to improve the customer service in order to keep the profile for the long-run.

CRM can be realized by financial product in the process of operation. So the most important way to improve the customer service is obtaining customer's information and distinguishing valuable customer to keep and develop from the process [1]. Customer relationship management (CRM) can be defined as "managerial efforts to manage business interactions with customers by combining business processes and technologies that seek to understand a company's customer, i.e. Structuring and managing the relationships with customers". The fastest way to build a successful customer-focused business is to divide the customer base into groups or segments in order to identify customers with the greatest profit potential [2]. Commercial bank CRM's contains two aspects: the first is a sales management of customers; the second is service management of customers. This is an important characteristic to distinguish commercial bank with other industry. So called sales management of customers is Doing sales of financial products to customers and exploiting market by active and effective ways; and service management of customers is doing good service and keeping management to customers who are from sales management in order to solidify and expand foundation of customers [1]. In the commercial banks CRM refers to the promotion and application of CRM in the banking field of commercial banks and falls into three levels: communication-level CRM, operation-level CRM and support-level CRM. The communication-level CRM is the interface of CRM to interact, collect and output information with customers, including phone, fax, Internet, Email, wireless- network, traditional counter, branch and others; customers can use various ways to get contact with customer service centers, to retain the required information and service; thus, integration between advanced technology and all kinds of bank resources is effectively realized. The operation-level CRM is made up of each sub module executing fundamental functions of CRM, including sales management module, marketing management module, customer service & support module, business intelligence management module, call center management module and E-commerce management module. The support-level CRM means data warehouse, data mining technology, operating system, network communication protocol and others, which are used in CRM, and is the foundation guaranteeing the normal operation of the whole CRM system. In the whole CRM structure, support-level CRM completes processing information accumulated from communication-level and operation-level, and then produces the analytical information based on the two. This information can be transferred to each functional module and front-consumer service system of the bank through systems integration, and finally forms internal dynamic, integrated consumer's analysis management and service of the bank [4].

2.2. Data mining

Data Mining is a process of abstracting unaware, potential and useful information and knowledge from plentiful, incomplete, noisy, fuzzy and stochastic data. These information and knowledge can't be achieved relying on a simple data search. The key of data mining includes three parts: data, information and business decisions. Data is the most valuable only when mobilized or converted into useful information. Accessing to data is not the ultimate goal of data mining. In fact, the final aim of data mining is using that information to improve business decision-making efficiency and to develop more appropriate decisions [3].

The application of Data mining technology in CRM plays an important role in improving the level and efficiency of CRM in commercial banks [4]. As the banking products have considerable homogeneity, the difference between banks lies in acquiring customers and unique business rules behind the massive business and customer information to archive a scientific decision-making. Now days database systems implemented by most banks can only realize data inputting, data querying , data collecting and other lower-level functions and are lacking in finding the connections between the data and business rules to predict future business trends [3].

2.3 Customer segmentation

Customer classification, also known as customer segmentation, divide a large number of customers in different types and the same type of customers have some similar attributes such as background information profitability, consumer preferences. Customer segmentation will enable banks to grasp the status of existing customers to provide different services for each type of customers are using different marketing methods to obtain the greatest return with minimal investment [10].

2.3.1. Segmentation Types

2.3.1.1. Value-Based Segmentation

In value-based segmentation customers are grouped according to their value. In this Regard, we supposed to segment customers based on their financial attributes.

2.3.1.2. Behavioral Segmentation

This is a very efficient and useful segmentation type. It is also widely used since it presents minimal difficulties in terms of data availability [6]. It includes customer identified information which is stored in databases.

2.3.1.3. Propensity-Based Segmentation

In propensity-based segmentation customers are grouped according to propensity scores, such as churn scores, cross-selling scores, and so on, which are estimated by respective classification (propensity) models [6].

2.3.1.4. Loyalty Segmentation

Loyalty segmentation involves the investigation of the customer's loyalty status and the identification of loyalty-based segments such as loyalty and switchers/migrates [6].

2.3.1.5. Socio-demographic and Life-Stage Segmentation

This type reveals different customer groupings based on socio-demographic and/or life-stage information such as age, income, marital status [6].

2.3.2. Clustering algorithms

2.3.2.1. Segmentation with K-MEANS

K-means is one of the most popular clustering algorithms. It starts with an initial cluster solution which is updated and adjusted until no further refinement is possible (or until the iterations exceed a specified number).

Each of the iterations refines the solution by reducing the within-cluster variation. The K in the algorithm's name comes from the fact that users should specify in advance the number of *k* clusters to be formed. The means part of the name refers to the fact that each cluster is represented by the means of its records on the clustering fields, a point referred to as the cluster central point or centroid or cluster center [6]. K-means is a common clustering algorithm that was proposed by MacQueen that based on the error squares in 1967 and the algorithm is convenient, fast and able to deal effectively with large databases though the initial value impacts on the clustering results greatly, it can easily be trapped into local optimum, and it depends on experience to determine the number of optimal classification [10]. In spite the fact that k-means is one of the most popular algorithms, it could address us into the no efficient solution. Since by different number of clusters it has different answers; however we could find efficient number of clusters with DUNN index in order to resolve the problem.

2.3.2.2. Segmentation with TWO-STEP ALGORITHM

The two-phase clustering algorithm was introduced by M.F. Jiang, S.S. Tseng and C.M. Su in. It consists of two stages. First phase is the modification of k-means algorithm by using a heuristic "if one new input pattern is far enough away from all clusters, then assign it as a new cluster center". In spite of the original k-means algorithm, where cluster centers are calculated after allocating all the objects, in modified k-means process centers are calculated after every object's allocation. Originally, in the second phase minimum spanning tree is constructed, with clusters obtained in the first phase as nodes, and then the tree is pruned by removing the longest edge. Such procedure allows effective outlier detection [10]. The Two-step algorithm involves a scalable clustering technique that can efficiently handle large data sets. It incorporates a special probabilistic distance measure that can efficiently accommodate both continuous and categorical input attributes. Another advantage of the Two-step algorithm is that it integrates an outlier handling option that minimizes the effects of noisy records which otherwise could distort the segmentation solution [6].

2.3.2.3. Segmentation with KOHONEN NETWORK/SELF-ORGANIZING MAP

Kohonen networks are special types of neural networks used for clustering. The network typically consists of two layers. The input layer includes all the clustering fields, called input neurons or units. The output layer is a two-dimensional grid map, consisting of the output neurons which will form the derived clusters. Each input neuron connects to each of the output neurons with strengths or weights. These weights (which are analogous to the cluster centers referred to in the K-means procedure) are initially set at random and are refined as the model is trained [10].

2.3.2.4. Evaluating segmentation algorithm

Comparison of different algorithms depends on business scope, constrains and type of information that is evaluated. In this regard, it doesn't make a sense to analyze different algorithm and assert an advice as a solution for all kinds of usages and businesses.

Consequently, we only review the researchers' work and study to meet the result of algorithms in specific contexts. In [7] they test that were performed to compare the algorithms that operate over several hundred of records of bank clients' data. During tests, algorithms were examined depending on a number of dimensions: attributes, efficacy in outlier detection, scalability and behavior in case of standardized and non-standardized data. While testing, the data attributes that are commonly taken into consideration in the bank customer analysis were chosen. In this research they used k-means, DBSCAN and Two-phase clustering algorithms. As the result shows Two-phase clustering algorithm detected outliers, while k-means allocated all the objects into clusters. On the contrary, DBSCAN indicated too many objects as outliers the algorithm presented the tendency to build a small amount of big clusters with many outliers, however its performance depends on the choice of parameters. In conclusion, in an environment with lots of outliers two-Phase algorithm is more suitable for detecting errors; we prefer not to use DBSCAN in this context in order to detect too many outliers since some data that are distinguished as outliers are not outliers in nature. Although, we could cluster customer by one algorithm, some scholars use two stage algorithms in order to have efficient result like using SOM -K-means algorithm for clustering for combining the advantages and disadvantages of the two algorithms to carry out application in bank customer segmentation [10].

2.4. Different analysis

2.4.1. RFM

To identify customer behavior, the well-known method called recency, frequency and monetary (RFM) model is used to represent customer behavior characteristics [9, 23]. RFM models have been used in direct marketing for more than 30 years. Given the low response rates in this industry (typically 2% or less), these models were developed to target marketing programs (e.g., Direct mail) at specific customers with the objective to improve response rates. Prior to these models, companies typically used demographic profiles of customers for targeting purposes. However, research strongly suggests that past purchases of consumers' are better predictors of their future purchase behavior than demographics [20]. The basic assumption of using the RFM model is that future patterns of consumer trading resemble the past and current patterns. The calculated RFM values are summarized to clarify customer behavior patterns. This study proposes using the following RFM variables [9]:

- Recency (R): the latest purchase amount.
- Frequency (F): the total number of purchases during a specific period.
- Monetary (M): monetary value spent during one specific period.

A large number of studies specifically in loyalty program areas considered RFM. For instance, Jonker et al. (2004) demonstrated the use of RFM value in direct-mail; they proposed optimization strategy for customer segmentation, marketing and employing stochastic dynamic programming. Also, Buckinx and Van den Poel (2005) and Fade et al. (2005) proved that RFM variables can predict accurately the CLV.

They showed how RFM variables can be used to build a CLV model that overcomes many of its limitations. They also showed that RFM is sufficient statistics for their CLV model. [5]

2.4.2. Share of Wallet

In this method the basis of calculating customer value is the ratio of sales of special product in firms to whole sales of that product to the customer in a specific period of time on the other hand, we estimate degree of customers' satisfaction. For instance, if the customer spends 500 units of money in a month and spends 300 units of that in company A. SOW of company A for this customer is 60% per month.

2.4.3. Marcov chain

This method is one of the methods that exist for calculating CLV and its calculation basis comes from Marcov chain models. In this model index of segmentation is based on customer profitability for the firms. In this regard, customers don't have equal profitability for the firms.

2.4.4. Past Customer Value

Essence of this method is based on this hypothesis that past customer behavior will show the degree of his profitability in the future. In the other word, in the past customer index we could estimate customer future value. In this method, customer value calculates with all customers' previous activity and the result is the basis of customer future value.

2.4.5. ROI

One of the methods of calculating CLV is ROI. Principle of CLV calculation is a return of the amount of money that cost for each customer. In this method it's important that how much money is cost for each customer. In this regard, customer is a tool for investing.

2.4.6. Analysis of methods

Observed on these methods that each one with different perspectives to the subject matter of CLV. Although the above methods for calculating the future value of customer generally used, but there are many problems in this methods. Though RFM, Past Customer Value, and Share-of-Wallet are commonly used for computing customer's future value, they suffer from the following drawbacks. These methods are not forward looking and do not consider whether a customer is going to be active in the future. These measures consider only the observed purchase behavior and extrapolate it to the future to arrive at the future profitability of a customer. In RFM method doesn't attention to financial aspects and it looks forward to qualities. This method sees three factors that are Recency, Frequency, Monetary Generally not correlated directly with customer profitability. Another weakness of this method is in predicting future, buying behavior and customer profitability in the future. According to the SOW method that is an index for customer value based on one sample of customer behavior it couldn't be a good measurement.

Another problem is that buying share of the company couldn't be an index for customer value; hence customers with high buying share could have less profitability for the company. Additionally, gaining information from other companies about customer buying behavior is not simple. PCV technique also fails to account for factors influencing future purchase behavior of customers. It also does not incorporate the expected cost of maintaining the customer in the future. ROI tends more towards the financial and qualitative factors are considered less. Consider the customer as a commodity or an investment tool in the current business world is not acceptable. For being able to effectively calculate CLV index, we need to know whether the customer has decided in future periods if the organization does not buy.

3. Research structure

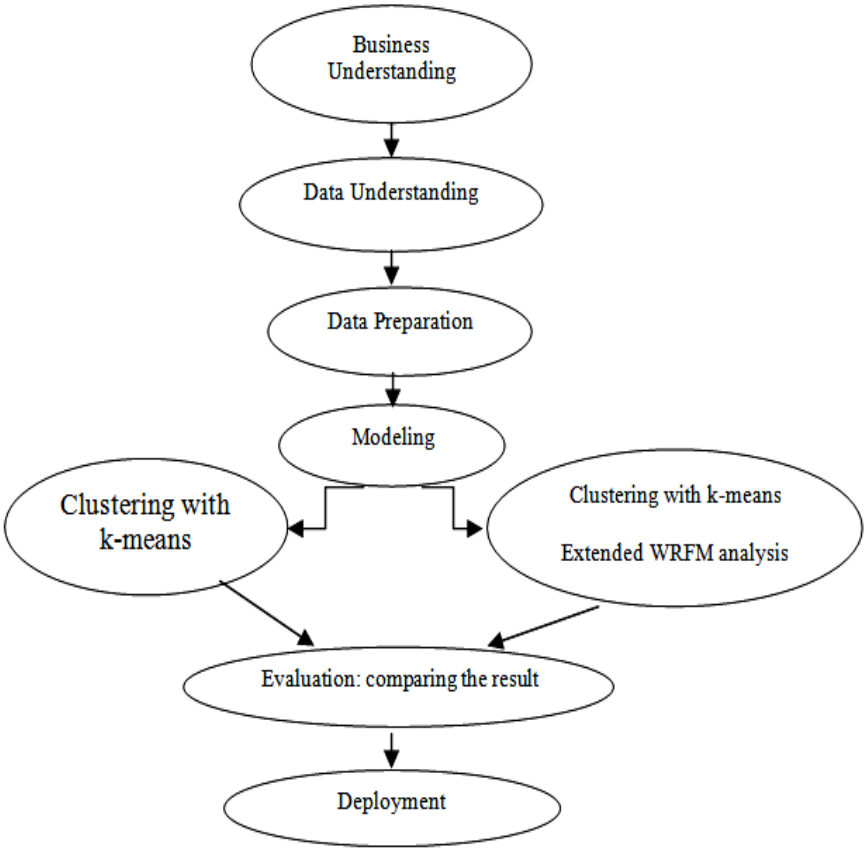

Figure 1: Research structure

Phase1: Business Understanding

In this area Business Objectives, Business Success Criteria and Data Mining Goals are considered.

The case is the retailer bank (Post bank) which covers a wide area in Iran compared with other government banks of Iran. In this regard, near two million customers' behavior traced from 22 December 2011 till 20 March 2012. The information is concluded with a customer account number, customer Identification code, transaction date, transaction money. In the banking business object is serving customers who bring more benefit to the bank. Furthermore, in retail bank, this idea of focusing on the best current customers should be seen as an on-going opportunity. To better understand the rationale behind this theory and to face the challenge of building customer loyalty, we need to break down shoppers into customer types which is one of the best ways of categorizing the customers using clustering and RFM method.

Phase2: Data Understanding

This phase involves taking a closer look at the data available for mining consist of data collection, data description and verifying the quality of data [11]. Data collected from database of core banking system for a season due to of large volumes of transactions and bank branches.

Phase3: Data preparing

In this phase data that's collected from a database is normalized and prepared in order to modeling data with clustering and RFM method. At this level, quality of data is acceptable for transactional data in banks couldn't have outliers and errors following this further data were normalized. In the following table RFM attributes are shown and explained.

Table1: RFM attributes

RFM attributes	Original Data
Recency	The interval between the last transaction date and last date of tracing time
Frequency	Frequency of transaction during the traced time
Monetary	Monetary or transaction (negative form shows withdrawal)

Min-max normalization method is used in this phase. This method performs a linear transformation of the original data. Suppose that min_A and 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 $[newmin_A, newmax_A]$ by computing in equation (1) [11].

$$v' = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A \quad (1)$$

Phase4: Modeling

In this paper k-means method used as clustering algorithm meanwhile, it is needed to distinguish the best number of clusters that makes clustering process more efficient.

Consequently, DUNN index which introduced by J. C. Dunn in 1974 is a metric for evaluating clustering algorithms. There are many ways to define the size or diameter of a cluster. It could be the distance between the largest two points inside a cluster. Let C_i be a cluster of vectors. Let x and y be any two n dimensional feature vectors assigned to the same cluster C_i.

$$D(C) = \max_{x,y \in c_i} d(x,y) / \min \delta(C_i, C_j) \quad (2)$$

Let $\delta(C_i, C_j)$ be this inter-cluster distance metric, between clusters Ci and Cj. According to DUNN indices, the number of optimum clusters obtained 4 in both approaches. We hypothesized that Participation bonds is influence factor for evaluating customer value in this way we evaluate both RFM and extended RFM methods. In the following tables results of both methods are shown.

Table2

Cluster	Count	Recency	Frequency	Monetary
C1	1602839	16.839	5.318	-3614348.267
C2	14	9	4474.29	631386405.7
C3	2	32	3	-120000875000
C4	346765	60.465	2.084	-1705267.135

Table3

cluster	Count	Participation bonds	Recency	Frequency	Monetary
c1	1945470	FALSE	24.607	4.72	-3261784.962
c2	4129	TRUE	20.825	13.396	-2034543.097
c3	3	FALSE	38.833	1566.67	-97736975983
c4	19	FALSE	8.947	3786.263	-1136820122

As shown above, participation bonds are the factor that affects clustering of RFM attribute result. Comparing result of two approaches indicates the noticeable differences. Pearson's correlation coefficient is calculated to measure the association between new added parameter (participation bonds) and other existing parameters which are usually included in the RFM method [11]. The result shows a considerable correlation between this parameter and the RFM parameters (Recency, Frequency, and Monetary). Since the new parameter has significant impact on clustering result and strong correlation with other existing parameters, it can be concluded that this parameter should be considered as an influencing parameter in calculating customer value. For calculating weighed RFM it's needed to calculate relative weights of the RFM variables by the AHP method as follows:

First step: First of all each RFM attributes should be weighted based on expert ideas as it's shown in the table.

Table4

	M	F	R
M	1	7	9
F	1/7	1	2
R	1/9	½	1

Second step: Result matrix of weighting should be normalized as following table.

Table5

	M	F	R
M	63/79	14/17	9/12
F	9/79	2/17	2/12
R	7/79	1/17	1/12

Final step: Average of each row shows the attribute's weight that calculated as Follows: WM=0. 790333, WF=0. 132746, WR=0. 076921.
The average CLV value of each cluster can be calculated with the equation:

$$CLV_{ci} = NR_{ci} \times WR_{ci} + NF_{ci} \times WF_{ci} + NM_{ci} \times WM_{ci} \quad (3)$$

NRci refers to normal Recency of cluster ci, WRci is Weighted Recency, NFci is normal Frequency, WFci is weighted Frequency, NMci is normal Monetary, and WMci is weighted Monetary [11].Average CLV is calculated as shown in below table.

Table6

Cluster	Count	participation bonds	Recency	frequency	Monetary	CLV
c1	1945470	FALSE	0.187	0	0.898	0.724102
c2	4129	TRUE	0.144	0.001	0.898	0.720928
c3	3	FALSE	0.27	0	0.000	0.03584
c4	19	FALSE	0.011	0.403	0.890	0.757739

To shed the light of this study, we need to categorize the R, F, and M parameters in three categories (Low, Medium and High). These categories were determined by the expert's idea of the bank in the following Table5.

Table7

Cluster	R	F	M	CLV rank
C1	Medium	Low	High	2
C2	Medium	Medium	High	3
C3	High	Low	Low	4
C4	Low	High	High	1

Comparing the result of RFM parameters' values of each cluster with categorized values that showed above demonstrate each cluster have different category of attribute that explained before. Furthermore, customers who have participation bonds are in third level of CLV it demonstrates that these kinds of customers don't bring significant benefit to the bank.

Phase5 and 6: Evaluation and Deployment

In order to meet the business objects in retail bank it's important to distinguish the best customers who bring significant benefit to the bank. By study the customers, we able to categorizing them and understand we should give better services to which customers with what kind of RFM attributes.

4. Conclusion

The purpose of this paper is showing the participation bond's effect on customers' value and evaluating the customer value based on k-means algorithm and RFM method. Although, participation bond has an effect on customer segmentation it doesn't show that customer who have participation bond are valued customers. Additionally, the RFM attributes of first ranked customers are as follows: Recency= Low, Frequency=High, Monetary=High. This result could help managers to plan for bank visions and strategies.

Reference:

1. G.Lao and S.Han, (2008), "Frame Analysis of Customer Relationship Management in Commercial Bank Based on CAS". Wireless Communications, Networking and Mobile Computing, 4th International Conference, p.1-4.
2. Z.Bošnjak and O.Grljevi, (2011),"Credit Users Segmentation for improved Customer Relationship Management in Banking". 6th IEEE International Symposium on Applied Computational Intelligence and Informatics, p.379-384.

3. Z.Li Ping and S . Qi Liang, (2010), " Data Mining Application in Banking - Customer Relationship Management". International Conference on Computer Application and System Modeling, p.124-126.
4. B.Fang and S.Ma, (2009)," Data Mining Technology and its Application". First International Workshop on Database Technology and Applications In CRM of Commercial Banks, p.243-246.
5. B.Sohrabi and A.Khanlari, (2007). "Customer Lifetime Value (CLV) Measurement Based on RFM Mode"l. I ranian Accounting & Auditing Review, 14 (47), p.7-20.
6. Konstantinos Tsiptsis, Antonios Chorianopoulos, (2009). "Data Mining Techniques in CRM Inside Customer Segmentation". John Wiley.
7. D.Zakrzewska, and J. Murlewski, (2005). "Clustering algorithms for bank customer segmentation". The 5th International Conference on Intelligent Systems Design and Applications, p.197 – 202.
8. H.Su-li, (2010). "The customer segmentation of commercial banks based on unascertained clustering". Logistics Systems and Intelligent Management, 2010 International Conference, 1 p.297 - 300.
9. P. Yanyan , (2011). "Evaluation and Classification of Commercial Bank Customer Value". Business Intelligence and Financial Engineering (BIFE), Fourth International Conference, p.682 - 686.
10. Z.Li Ping, and S. Qi Liang, (2010). " Data Mining Application in Banking - Customer Relationship Management". International Conference of Computer Application and System Modeling, p.124-126.
11. M.Khajvand, S. Ashoori, and S. Alizadeh, (2011). "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study". Procedia Computer Science, 3(1), p.57–63.
12. Z.Tabaei, and M. Fathian, (2011). "Developing W-RFM model for customer value: An electronic retailing case study". Data Mining and Intelligent Information Technology Applications, 3rd International Conference, p.304 - 307.