THE 1st INTERNATIONAL CONFERENCE ON DATA SCIENCE AND OFFICIAL STATISTICS
ICDSOS
"Strengthening the Contribution of Data Science and Official Statistics to the Society in the Distruption Era"
2021

Estimating Customer Lifetime Value in the E-Commerce Industry Using Multivariate Analysis

B C Laksono[1], **I Y Wulansari**[2]

[1,2]Politeknik Statistika STIS, Jalan Otto Iskandardinata No 64C RT.1/RW 4, Bidara Cina, Jatinegara, Jakarta Timur, 13330

*Corresponding author: [1]211810211@stis.ac.id

Abstract. Companies can develop their business using big data to support decision-making. Big data in the e-commerce industry that includes size and speed of high transactions can be used to analyze customer behaviour and predict customer value. Nowadays, companies are starting to develop customer-oriented rather than product-oriented business interests. One way that can be used to determine customer value is by calculating Customer Lifetime Value (CLV). By knowing CLV at the individual level, it will be useful to help decision-makers to develop customer segmentation and resource allocation. It is important to do segmentation or customer grouping that describes customer loyalty groups. Therefore, this research aims to calculate CLV and customer segmentation using the RFM analysis method. The dimensions of forming CLV include the values of Recency, Frequency, and Monetary. In this study, concept of multivariate statistical analysis will be applied, namely K-Means Clustering and factor analysis. Segmentation is done to determine the level of customers. The higher the CLV value, more valuable customer is to maintain. In the end, the customer segmentation method built by author can be used to optimize company's strategy to get maximum profit. This method can be applied to various cases and other companies.

1. INTRODUCTION

In the era of information revolution the emerging marketing problems require in-depth research. These new emerging issues are interesting to investigate, and old issues can be analyzed better due to better data availability. In marketing research, Customer Lifetime Value (CLV) analysis is one of the new methods in developing the marketing field to determine potential customer groups. According to Gupta [1], the modern economy is accompanied by a service infrastructure, and businesses get more money by creating and maintaining long-term relationships with customers. In such an environment, marketing aims to maximize the assets of customers, which is the sum of CLV (Customer Lifetime Value) and corporate customer value.

Customer value has attracted the attention of all customer researchers because it plays a very important role at the center of all marketing activities [2]. CLV is the present value of the expected profit or loss earned by a company during a transaction with a customer. According to Khajvand [3], on a customer-centric based approach, customers are considered as an asset. The profits obtained by a company are influenced by length and number of customers and also the quality of these customers [4]. This approach is considered the best, resulting in a paradigm shift in making a company's business decisions. Of course, getting the CLV value requires supporting data. Thus, several companies such as telecommunications, retailers, banking, and many more collect customer information and transaction data to identify customer habits for a certain period. Knowing CLV at the individual level will help decision-makers develop customer segmentation and resource allocation [5&6].

One way to manage the relationship between a company and its customers is customer lifecycle, which consists of customer acquisition, customer retention, and customer development [7]. In the customer acquisition stage, company must select several customers who have good potential. Furthermore, entering the customer retention stage, the company tries to keep customers who have high values. The last stage is customer development, namely, the company's development process so that it is expected that customers will provide more benefits for the company.

It is not easy to know how companies determine good customers to acquire, retain and develop because each stage of the customer's life cycle requires high costs and resources. Therefore, it is very important to understand the customer lifetime value (CLV) in order to provide an accurate measure, that is, the present value of the expected profit or loss that a business will earn from a transaction with a customer [8].

Several models that can explain CLV include RFM, probability, econometric, computer science, and diffusion/growth models. But among them all, RFM analysis is one of the most popular and effective CLV models to be applied in market segmentation. RFM analysis is an approach to understanding customer behaviour through evaluation metrics and customer segmentation. RFM is an abbreviation for Recency, Frequency, and Monetary. RFM analysis can be used to classify customers based on the time interval of their last visit, the frequency of visits, and the amount of value that customers have issued [9]. The results of this analysis can be used for proper customer profiling. The company must maintain customers with high CLV because it will be very profitable for the company.

2. DATA AND METHODOLOGY

2.1 Data Source

The research conducted includes 4338 customers of international e-commerce in the United Kingdom in 2010-2011. The data used is secondary data as many as 541910 purchase transactions.

2.2 Recency, Monetary, Frequency (RFM) Analysis

Bult and Wansbeek first introduced the concept of RFM in 1995. The abbreviation of RFM is Recency, Frequency, and Monetary [11]. RFM is an analytical method used to segment customers into specific classes. In this study, each RFM component will be divided into three class categories, namely low, medium and high. RFM analysis relies heavily on transaction data made by customers.

a. Recency means the analysis day minus the last day the customer made a transaction. Recency indicates that the customer has recently purchased something. Customers with recent purchases are more likely to respond to new offers than customers whose purchases last longer.

b. Frequency means how often a customer makes a transaction. Frequency shows the number of purchases made by customers. If a customer makes a purchase more often, it will result in a higher positive response than a customer who buys something infrequently.

c. Money is the amount spent by a customer. Monetary value represents a purchase conversion all purchases by a customer. Customers who spend more money than they buy are more likely to respond to suggestions once than they do.

Table 1. RFM Data

No	CustomerID	Recency	Frequency	Monetary ($)
1	12346	348	1	77183.6
2	12347	25	7	615.7143
3	12348	98	4	449.31
4	12349	41	1	1757.55
5	12350	333	1	334.4
etc.				

Dataset, though. It is showing 5 of 4338 customers.

From the table above, the RFM score can be formed. The RFM score is the result of such weighting of those three components. The RFM score can help identify customers from the smallest to the largest purchasing power in the form of a specific group classification order. Therefore, the Recency, Frequency, and Monetary dimensions will be used to form the Customer Lifetime Value (CLV). RFM analysis can play a crucial role in increasing the profitability of advertising campaigns. RFM is an analytical method that divides customer segmentation into four classes: Gold, Silver, Bronze, and Non-Profit.

2.3 K-Means Clustering

Cluster analysis is a method of grouping a set of objects so that objects are classified into the same group according to certain characteristics. In classifying data, the concept of distance similarity will be used. If the two observations have a small distance, the similarity is high, so they fall into the same group. KMeans is an unsupervised machine learning algorithm designed to find clusters in the data, and number of clusters is represented by K. The concept of KMeans is to minimize variance within group and maximize variance between groups. To process the data from Kmeans clustering algorithm, the data starts from the first set of centroids randomly selected as the starting point for each group, and then an iterative calculation is performed to optimize the position of the centroid. When the centroid is stable and reaches the specified number of iterations, the process will stop. The process of grouping data into groups can be completed by calculating the shortest distance from the data to centroid point. The Minkowski distance calculation can be used to calculate distance between two data. Base on Johnson [10], the formula for calculating distance is

$$d(x_i,x_j) = \left(|x_{i1}-x_{j1}|^g + |x_{i2}-x_{j2}|^g + \cdots + |x_{ip}-x_{jp}|^g\right)^{\frac{1}{g}} \quad (1)$$

Where:
$g = 1$, to calculate Manhattan distance
$g = 2$, to calculate Euclidean distance
$g = \infty$, to calculate Chebychev distance
x_i, x_j are two pieces of data to be calculated the distance
$p =$ data dimensions
Renewal of centroid point can be done with the following formula:

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q \quad (2)$$

Where:
μ_k = K-th cluster centroid point
N_k = K-th cluster amount of data
X_q = the q-th data in the cluster

2.4 Factor Analysis

Factor analysis is a multivariate statistical analysis used to create unique variables in a multivariate data set. Base on Johnson [10], the model in factor analysis is as follows :

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2$$
$$\vdots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p$$

The matrix notation is :

$$(\boldsymbol{X} - \boldsymbol{\mu})_{(p\times 1)} = \boldsymbol{L}_{(p\times m)}\boldsymbol{F}_{(m\times 1)} + \boldsymbol{\varepsilon}_{(p\times 1)} \tag{3}$$

$\mu_i = mean\ of\ variabel\ i$
$\varepsilon_i = ith\ spesific\ factor$
$F_j = jth\ common\ factor$
$l_{ij} = loading\ of\ the\ ith\ variable\ on\ the\ jth\ factor$
$L = matrix\ of\ factor\ loading$

One of the estimation methods in factor analysis is the Principal Component method. The formula is as follows:

$$\boldsymbol{\Sigma} = \lambda_1 \boldsymbol{e_1}\boldsymbol{e'_1} + \cdots + \lambda_p \boldsymbol{e_p}\boldsymbol{e'_p} = \left[\sqrt{\lambda_1}\boldsymbol{e_1} \vdots \cdots \vdots \sqrt{\lambda_p}\boldsymbol{e_p}\right]\begin{bmatrix}\sqrt{\lambda_1}\boldsymbol{e'_1} \\ \vdots \\ \sqrt{\lambda_p}\boldsymbol{e'_p}\end{bmatrix} = \boldsymbol{LL'} + \boldsymbol{0} = \boldsymbol{LL'} \tag{4}$$

If m factors are used, then

$$\boldsymbol{\Sigma} = \left[\sqrt{\lambda_1}\boldsymbol{e_1} \vdots \cdots \vdots \sqrt{\lambda_m}\boldsymbol{e_m}\right]\begin{bmatrix}\sqrt{\lambda_1}\boldsymbol{e'_1} \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e'_m}\end{bmatrix} + \begin{pmatrix}\boldsymbol{\psi}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\psi}_p\end{pmatrix} = \boldsymbol{LL'} + \boldsymbol{\psi} \tag{5}$$

Which $\boldsymbol{\psi_i} = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2\ ; for\ i = 1,2, \dots, p$

If the estimation method uses Principal Component, the calculation of the factor score can use the following formula :

$$\hat{\boldsymbol{f}}_j = \left(\tilde{\boldsymbol{L}}'\tilde{\boldsymbol{L}}\right)^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) \tag{6}$$

When known $\tilde{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \vdots \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \vdots \cdots \vdots \sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1\right]$, then

$$\hat{\boldsymbol{f}}_j = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}}\hat{\boldsymbol{e}}_1'(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) \\ \frac{1}{\sqrt{\hat{\lambda}_2}}\hat{\boldsymbol{e}}_2'(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}}\hat{\boldsymbol{e}}_m'(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) \end{bmatrix} \tag{7}$$

For each factor scores apply,

$$\frac{1}{n}\sum_{j=1}^{n}\hat{\boldsymbol{f}}_j = 0 \tag{8}$$

and

$$\frac{1}{n-1}\sum_{j=1}^{n}\hat{\boldsymbol{f}}_j\hat{\boldsymbol{f}}_j' = \boldsymbol{I} \tag{9}$$

The notation above shows that the data produced by factor score is normal multivariate (in standardization form)

2.5 Multivariate Normal Distribution

Random variable vector $\boldsymbol{X} = [X_1, X_2, \dots, X_n]$ is assumed to have a normal multivariate (Gaussian) distribution with mean $\mu \in \mathbf{R}^{\boldsymbol{n}}$ and covariance matrix $\Sigma \in S_{++}^{n}$ [10]. Suppose it has the following probability distribution.

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{10}$$

Or can be written by $X \sim N(\mu, \Sigma)$

The following are the properties of a multivariate normal distribution:
Let X is a multivariate normal distribution, then:

1. The linear combination of the X components is normally distributed.
2. All subsets of component X have a normal (multivariate) distribution
3. The covariance value equal to zero indicates that the components are independently distributed
4. The conditional distribution of its components is (multivariate) normal

3. RESULT AND DISCUSSION

3.1 Cluster Analysis

RFM score calculation and customer segmentation are techniques that each company can develop. The results may vary according to the strategies and policies the company wants to achieve. In this study, the authors innovate to determine segmentation using one of the clustering techniques, namely K-Means Clustering. Segmentation for each component is defined and divided into 3 level categories, namely "Low", "Medium", and "High" value, so that it can be known at the outset that the value of $K = 3$. The higher the level, the more valuable customers are.

Recency, frequency, and monetary variables have been obtained in the previous sub-chapter. Those variables represent customer values. Before clustering, the data for each variable will be searched for outliers to be removed to produce an optimal cluster size. The top outlier or bottom outlier will later be included in the nearest cluster after the analysis process.

a. Recency

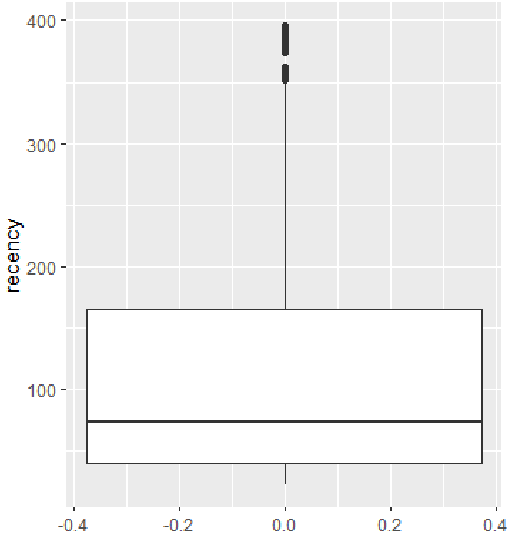

Figure 1. Boxplot of Recency

Table 2. Summary of Recency

Summary	
Min.	23,0
1st Qu.	40,0
Median	73,0
Mean	115,1
3rd Qu.	164,8
Max.	396,0

Based on the boxplot and table above, $the\ upper\ limit\ of\ outliers\ = Q3 + 1.5IQR = 164.8 + 1.5(164.8 - 40) = 352$ is obtained. So that the recency value above 352 is included in the top outlier. One hundred fifty-two customers are included in the top outliers so that later these 152 customers will be included in the cluster with the highest average. The remaining data of 4186 customers will be analyzed by cluster, and the following results are obtained:

Table 3. Clustering Result of Recency

Recency			
Cluster	Means	Level	Size
1	52,20403	High	2779
2	279,96672	Low	631
3	153,86082	Medium	776
Jumlah			4186

between_SS / total_SS = 89,9 %

Based on the table above, the level assignment is adjusted to the mean value of each cluster. Customers with lower recency values will be more feasible to maintain because their CLV values are high. Logically, customers with low recency are currently still processing transactions with the company. So it can be concluded that customers with a "Low" level are 631+152=782, customers with a "Medium" level are 776 and customers with a "High" level are 2779. The ratio between the variance compared to the total variance is 89.9%, indicating that the cluster analysis results are pretty good.

b. Frequency

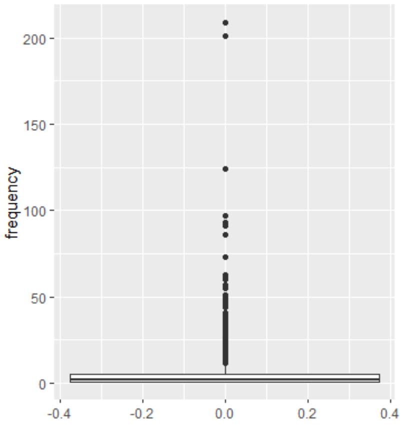

Figure 2. Boxplot of Frequency

Table 4. Summary of Frequency

Summary	
Min.	1,000
1st Qu.	1,000
Median	2,000
Mean	4,272
3rd Qu.	5,000
Max.	209,000

Based on the boxplot and table above, $\textit{the upper limit value of the outliers} = Q3 + 1.5IQR = 5 + 1.5\,(5-1) = 8$. The frequency value above eight will be considered as the top outlier. In this case, 459 customers are included in these criteria so that later these 459 customers will be included in the cluster with the highest average. The remaining data of 3879 customers will be analyzed by cluster, and the following results are obtained:

Table 5. Clustering Result of Frequency

Frequency			
Cluster	Means	Level	Size
1	3,766257	Medium	1138
2	6,820823	High	413
3	1,358677	Low	2328
	Jumlah		3879

between_SS / total_SS = 89,4 %

Based on the table above, the level assignment is adjusted to the mean value of each cluster. Customers with a higher frequency value are more feasible to maintain because the CLV value is high. So it can be concluded that customers with a "Low" level are 2328, customers with a "Medium" level are 1138 and customers with a "High" level are 413+459=872. The value of the ratio between the variance compared to the total variance of 89.4% indicates that the cluster analysis results are quite good.

c. Monetary

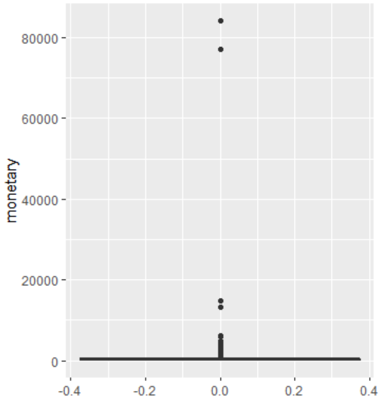

Figure 3. Boxplot of Monetary

Table 6. Summary of Monetary

Summary	
Min.	3,45
1st Qu.	178,62
Median	293,90
Mean	419,17
3rd Qu.	430,11
Max.	84236,25

Based on the boxplot and table above, the upper limit value of the outliers = Q3 + 1.5IQR = 430,11 + 1.5(430,11-178,62) = 807.345. So that the monetary value above 807.345 will be considered to be included in the upper outlier. In this case, 580 customers belong to these criteria, so that later the 580 customers will be included in the cluster with the highest average. The remaining data of 3758 customers will be analyzed by cluster, and the following results are obtained:

Tabel 7. Clustering Result of Monetary

Monetary			
Cluster	Means	Level	Size
1	352,8973	Medium	1560
2	163,5875	Low	1904
3	598,7011	High	584
	Jumlah		3758

between_SS / total_SS = 84,9 %

Based on the table above, the level assignment is adjusted to the mean value of each cluster. Customers with higher monetary values will be more feasible to maintain because their CLV values are high. So it can be concluded that there are 1904 "Low" customers, 1560 "Medium" customers, and customers with a "High" level are 584+580=1164 customers. The value of the ratio between the variance compared to the total variance of 84.9% indicates that the cluster analysis results are quite good.

After obtaining the results of clustering each RFM data, the company can use it flexibly to choose the customers they want to retain. For example, if the company intends to maintain customers by referring to high monetary values, it can select the results of the High monetary cluster. Likewise for the combination of clustering results on other dimensions, depending on company policy.

3.2 Factor Analysis

In this section, the CLV value will be calculated. The CLV value will be formed from the dimensions of recency, frequency, and monetary from the previous subchapter. The analysis used is factor analysis with the principal component estimation method. Based on factor analysis, the following results were obtained:

Table 8. Factor Analysis Result

Rotated Component Matrix[a]			
	Component		
	1	2	3
RECENCY	0,991	0,000	-0,132
FREQUENCY	-0,132	0,010	0,991
MONETARY	0,000	1,000	0,009
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.[a]			
a. Rotation converged in 4 iterations.			

Based on the table above, three factors are generated. These three factors are explained by Component 1, Component 2, and Component 3. It can be seen that Component 1 is associated with the recency dimension because it has the highest loading factor of 0.991. Component 2 has the highest loading factor on the monetary dimension of 1,000, and Component 3 has the highest loading factor on the frequency dimension of 0.991.

Table 9. The variance of each component

Total Variance Explained			
Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	1,000	33,334	33,334
2	1,000	33,333	66,667
3	1,000	33,333	100,000
Extraction Method: Principal Component Analysis.			

Based on the table above, we can see that each component has an equal contribution seen from the proportion of variance produced, which is around 33.333. So that in the formation of the CLV value, these dimensions have the same contribution. The role of the factor score of each component will be used to generate a value that will be used to form the CLV. Thus, the CLV will be created from the following formula:

$$CLV = \frac{(-Factor\ Score\ of\ Recency + Factor\ Score\ of\ Monetary + Factor\ Score\ of\ Frequency)}{3} \quad (11)$$

Table 10. Calculating Factor Score and CLV

CustomerID	**Factor Score**			**CLV**	**Customer Class**
	Recency (Comp 1)	Monetary (Comp 2)	Frequency (Comp 3)		
12346	2.25762	42.73529	-0.54851	13.30972	Gold
12347	-0.87656	0.10749	0.24012	0.408057	Silver
12348	-0.18	0.01742	-0.05971	0.045903	Silver
12349	-0.81963	0.75051	-0.54493	0.341737	Silver
12350	2.17981	-0.04676	-0.13905	-0.78854	Non Profit
etc.					

Dataset, though. Showing 5 of 4338 customers.

Based on the nature of factor score estimated by the principal component method, theoretically, each value will be normally distributed with a mean of zero and a variance of 1. CLV is formed from a linear combination of the factor score components so that the CLV will be normally distributed as well. The division of customer class is based on the following categorization

- Gold : $CLV > 0{,}67449$
- Silver : $0 < CLV \leq 0{,}67449$
- Bronze : $-0{,}67449 < CLV \leq 0$
- Non-Profit : $CLV \leq -0.67449$

The boundary values are obtained from the z score obtained by the inverse of the probability.

Table 11. Customer Loyalty Class

Customer Class	Criteria
Non Profit	Has the lowest recency, frequency and monetary
Bronze	Customers with fairly low loyalty
Silver	Moderately loyal customer
Gold	Highly loyal customers

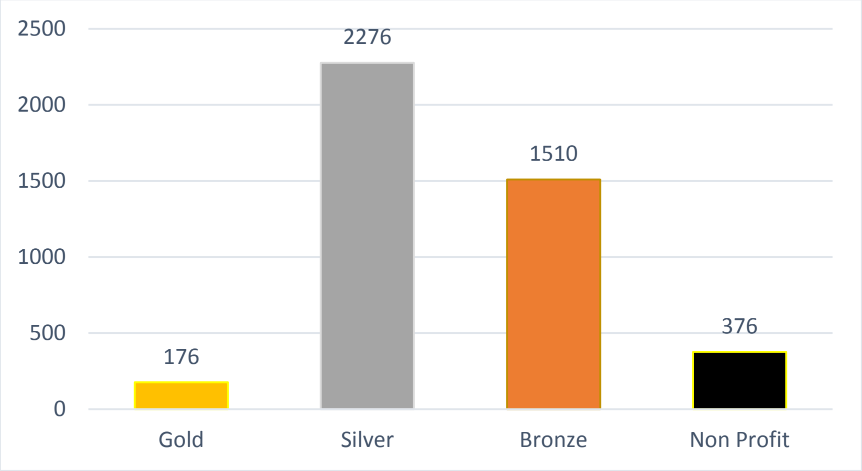

Figure 4. Number of Customers After Segmentation

Based on the results above, 176 customers in the Gold category have high value for the company. Gold customers that the company must maintain. Meanwhile, for the Non-Profit category, there are 376 customers. Non-profit customers do not need to be retained by the company.

CONCLUSION AND RECOMMENDATION

CLV is a measure that can be used to determine the value of a customer to the company quantitatively. The resulting value is an overall calculation made during the customer's transaction. The RFM model can be used to define how valuable customers are to the company. The higher the Customer Lifetime Value of a customer, then the customer deserves to be maintained. Therefore, customers with the highest loyalty class can be given optimal service so that these customers survive and provide the highest profit to the company. The customer segmentation method built by the author can be used to optimize the company's strategy to get maximum profit. This method can be applied to various cases and other companies. Suggestions for further research need to be developed regarding the method used. The customer segmentation method needs to be improved to get an accurate and robust estimation and segmentation results.

REFERENCES

[1] Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., & Sriram, S. (2006). *Modeling customer lifetime value*. Journal of service research, 9(2), 139-155.

[2] Holbrook, Morris B. (1999). *Consumer Value: A Framework for Analysis and Research*. London and New York : Routledge

[3] Khajvand, Mahboubch dkk. (2011). *Estimating Customer Lifetime Value Based On RFM Analysis of Customer Purchase Behavior: Case Study*. Procedia Computer Science 3 (2011) 57-53

[4] Reinartz, W. J., V. Kumar. (2003). *The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration*. J. Marketing 6777–99

[5] Kumar, V., & Reinartz, W. (2006). *Customer Relationship Management: A databased approach*.

[6] Lee, K., Lee, H. and Kim, S. (2007) *Factors Influencing the Adoption Behavior of Mobile Banking: A South Korean Perspective*. Journal of Internet Banking & Commerce, 12, 1-9.

[7] F. Buttle. (2009). *Customer Relationship Management Concepts and Technologies, 2nd ed.* USA : Elsevier Ltd.

[8] Berger, Nasr. (1998). *Customer Lifetime Value: Marketing Models and application*. Journal of Interaktive Marketing Volume 12, Number 1, Winter.

[9] Aggelis, Vasilis, dan Christodoulakis. (2005). *Customer Clustering using RFM Analysis*. Proceedings of the 9thWSEAS International Conference on Computers (ICCOMP).

[10] Johnson, Richard A and Wichern, Dean W. (2007). *Applied Multivariate Statistical Analysis*. New Jersey : Pearson Education, Inc.

[11] Cheng, C.H. dan Y.S. Chen. (2009). Classifying the segmentation of customer value via RFM model and RS Theory. Expert Systems with Application, Vol. 36 Issue 4 no 2, hal 216