Development of Bank's Customer Segmentation Model based on RFM+B Approach

Article · January 2021

DOI: 10.24507/icicelb.12.01.17

2 authors:

Uus Firdaus
Binus University

Ditdit Nugeraha Utama
Binus University

ICIC Express Letters
Part B: Applications
Volume **12**, Number **1**, January **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **17–26**

DEVELOPMENT OF BANK'S CUSTOMER SEGMENTATION MODEL BASED ON RFM+B APPROACH

Uus Firdaus and Ditdit Nugeraha Utama

Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia 11480
uus.firdaus@binus.ac.id; ditdit.utama@binus.edu

Received April 2020; accepted July 2020

Abstract. *Analysis of recency-frequency-monetary (RFM) is an analytical method that focuses on customer behavior. Fundamentally, R shows the last transaction, F is the number of transactions, and M represents a total amount of expenses. It has often been applied and provides an effective analysis for decision makers to promote their product strategies. However, this is considered not able to accommodate the segmentation needs of banking customers; thus customer balance should be involved in the analysis process theoretically. The customer balance (B) is potentially able to be functioned for the customer segmentation process and fruitful in marketing strategies. The developed model is called the recency-frequency-monetary-balance (RFM+B) model. It is a segmentation model of bank's customer considering four aspects: recency, frequency, monetary, and balance, where it is developed by using main method K-Means. The constructed model is applied successfully in one bank's 65 thousand customers coming from 147 thousand transaction data in period of the first half of 2017: cash payments, cash deposits, overbooking, and transactions through ATMs. The result shows that clusters 0 until 3 are dominantly filled by customers with high R (with average 113.17 times), high B (with average 3,487,790,000 Indonesian rupiah), high F (with 315.73 times), and high M (with average 5,000,000,000 Indonesian rupiah) respectively.*

Keywords: Customer segmentation, K-Means, Bank's customers, Segmentation model, RFM+B model

1. **Introduction.** In banking companies, there is a third party fund (TPF) which is an important source of funding and is often referred to as core liability. TPF in Indonesian banks basically consists of savings, time deposits, and current accounts. Savings and current accounts are designed to be withdrawn at any time by the owner through bank channels [1]. If the banks did not care about management of TPF, they are going to lose a strength in generating profits and reduce the loan ultimately [2]. In addition, the big number of data is a bank's high potency when they are managed appropriately [3]. Also, a well-data-process is going to produce fruitful information to improve bank's ability in making profit practically [4].

For one bank, publicizing the right type of product to definite customers is compulsory. The wrong strategy can trigger the bank's marketing performance declines. Here, customer segmentation is going to be a logical and scientific reason for banks to promote convinced types of product to accurate customer. In addition, the good customer segmentation model is a verified standard way operated by banks for customer segmentation to amplify effectiveness of marketing operation. Bank requires a standard model that is able to make customer segmentation objectively to increase customer loyalty, retain existing customers, and attract new customers [5].

In other studies, there is another valuable attribute that is able to be considered for customer segmentation, i.e., balance (B). Attribute B for customers is very important, as

DOI: 10.24507/icicelb.12.01.17

it is going to affect in grouping customers and in determining the strategies which will be carried out by the marketing department [6].

Analysis of recency-frequency-monetary (RFM) is an analytical method that focuses on customer behavior, and classifies customers using the variables recency (R), frequency (F), and monetary (M). R shows the length of time since the last transaction. F is the number of transactions in a period. M exhibits the total amount of expenses in a period [7]. Analysis of customers using the RFM-model can also be combined with other grouping methods in the clustering algorithm in data mining [8].

[5] did a research related to the segmentation of consumer consumption behavior. Various customers were segmented based on their consumption behavior to ensure objective assessment standards based on the RFM attribute model [5,7-11], and the attributes become very effective in the customer segmentation [5]. [12] analyzed the telecommunications industry in China, and they operated a segmentation strategy based on customer value and extracted customer data characteristics, and then classified customers based on customer value which was derived from historical value, current value, long-term value, loyalty, and credit. [13] has discussed classifying user behavior. They extracted features that can reflect their behavior as input for the C-SVM model that groups user behavior into different classes. The cluster behavior method functioned by extracting features from the behavior and then using it as an input model.

Furthermore, [11] marked customers with three different scores for the variables R, F, M. The assessment was done by calculating scores for each instance, where the score of 5 in each parameter was the highest and one was the lowest score. [14] has explained the results of customer segmentation by finding the value and type of customer used as a marketing strategy to increase company profits. RFM to identify customers combined elbow methods and sum of squared errors (SSE) for the K-Means algorithm. [15] has examined customer behavior with the RFM model in 2 steps. First, the results of RFM modeling are followed by the K-Means algorithm, and next step was to analyze the characteristics of the cluster using silhouette and connectivity measurements. There was also a segmentation of internet banking users by getting RFM scores and comparing several methods of grouping from transactions [16]. Customer segmentation can also be based on customer characteristics in the environment such as geographic, demographic, and behaviors that are in a broad geographical scope such as countries, regions, cities [17], and customer characteristics based on job profile, current balance in accounts, age, gender, balance level [18].

The proposed model is considered to accommodate the needs of segmentation for bank customers, the study includes customer balances in the analysis process. B is a savings balance held by a customer until the end of the period within the scope of the study. The model is carried out to segment customers based on transaction behavior and the balance they have, so that it can be performed to identify potential customers, evaluate and recommend product marketing strategies and customer growth strategies [6]. We proposed an RFM+B model combining four imperative attributes mentioned, where the main method used is K-Means [8,12]. This paper consists of the following arrangement. Theories and methods in research are found in Section 2. In Section 3 research methodology, framework and stage are discussed. Section 4 talks about the proposed model, and the conclusions are given in Section 5.

2. Theories and Methods.

2.1. RFM+B model.

The existing, the RFM analytic model separates important customers with three attributes: recency, frequency, and nominal purchase (monetary). Segmentation for consumer consumption behavior exploits their consumption behavior. Through the RFM attribute to ensure that assessment standards are set objectively based on the RFM attribute of the model. The present value of R is greater if the interval is smaller, the value of F will be greater if the number of transactions is greater, while M value is greater if the amount of money used is also in large quantities [10]. B is a savings balance held by a customer at the end of the data period.

R depicts the presence of the last purchase refers to the interval between the time at which the last consumption behavior occurred and the current time. The shorter interval will affect greater R. F is about the number of customer deals. F is defined as the number of purchases made by a customer in each period. The higher F value indicates that the customer is loyal to the company. M is about how much customers money spend. M is defined as the amount of money spent by a customer during a certain period. The higher M means more income customers provide to the company [19]. B itself is a balance held by a savings customer at the end of the data period within the scope of the study. It is utilized to segment savings customers for helping bank management in setting product marketing strategies [6].

2.2. **K-Means clustering.** Clustering is a grouping process which is classified as an unsupervised classification. Clustering can be interpreted as the process of grouping or classifying objects based on information obtained from data that can describe the relationships between objects. It can be used to find distribution patterns in a dataset that functions for the data analysis process [17].

K-Means is a simple clustering algorithm that has the ability to bend large amounts of data, and partition the dataset into clusters of k [20]. K-Means algorithm has a high level of accuracy, is effective and requires a relatively fast execution time because it is linear [21]. The process of modifying the number of k is done to obtain clusters whose members have a high degree of similarity. Algorithm is still possible to be improved, such as overcoming weaknesses of previous algorithms in image segmentation [22].

Determining the number of clusters (k) is an initial step in the K-Means algorithm in order to get optimal grouping. To determine the best number of clusters you can use the elbow method. The elbow method will provide the best cluster value taken from SSE value. The significant SSE value decrease makes the elbow form to determine the optimal number of clusters. High SSE value means high error, and it is poor cluster quality [21]. Thus, good quality which has a minimum SSE is value reached by reducing each pixel to the center of the nearest cluster to form a well segmented image [22].

3. **Research Methodology.**

3.1. **Research framework.** This study produced a model for classifying savings customers based on the value of customer attributes. These attributes are generated based on an analysis of customer behavior related to transactions that have been carried out and the current balance. This RFM model retains information about the time of the last purchase, the total number of transactions carried out, and the total nominal money spent [10]. As a development of this model, in the process of analysis, the attribute of the customer's current balance is considered as one of the important attributes.

From the research framework in Figure 1, there are problems in marketing products. It is affected because banks do not have a model for marketing strategies and verified standards in segmenting savings customers based on transactions and customer balances. Big bank's data are exploited as research material, including data of savings and transaction in specified period. RFM+B model was developed by adding RFM attributes/variables with the customer's B. B is a savings balance held by a savings customer at the end of the data period in the scope of the study. This B is used to help segment savings customers, so that it can help company in product marketing strategies. The results are normalized for the grouping stage by the K-Means clustering method, and the elbow principle is used in determining the optimal number of clusters and evaluated with SSE. The position of the

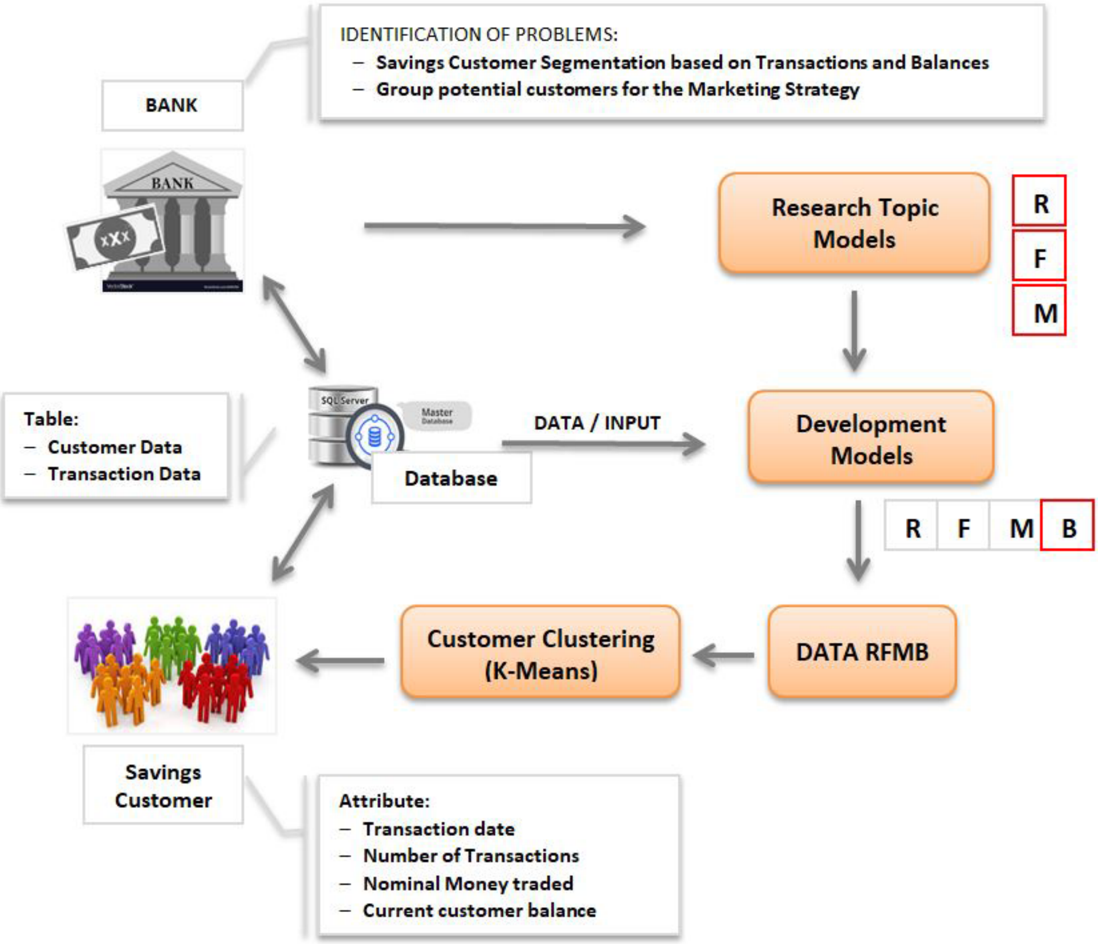

FIGURE 1. Research framework

research to be carried out lies in the development of model and the customer segmentation model.

3.2. **Research stages.** The cross-industry standard process (CRISP) is a method benefited in this study to solve problems and refers to the six stages of CRISP, which includes the process of understanding business, understanding data, data preparation, modeling, evaluation, and the final process is development [19]. Such six stages are simplified into four simple stages of the research that are configured in Figure 2. They consist of bank's data identifying, method searching, model developing, and model evaluating.

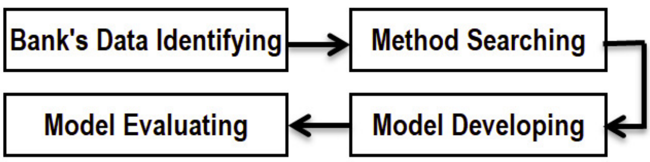

FIGURE 2. Simple research stages

The first stage is a crucial stage. It is utilized to process the data selection. The selection of variables/attributes operated is prepared to simplify the analysis process. Customer data is integrated with transaction data using the join table method in the structure query language (sql) using the Navicat premium version 11.2.7 application. It is developed by including the balance attribute in accordance with the objectives expected in this study. Transaction data and customer data cannot be directly used as input in segmenting, but requires further processing and evaluation of its attributes based on predetermined parameters as shown in Table 1, and the results of calculations for each customer can be seen in Table 1.

Table 1. Recency, frequency, monetary, balance parameters

Attributes	Value 1	Value 2	Value 3	Value 4
Recency	> 50	31-50	16-30	0-15
Frequency	≤ 6x	7-24	25-72	> 72
Monetary	$\leq 600{,}000$	600,001-6,000,000	6,000,001-24,000,000	$> 24{,}000{,}000$
Balance	$\leq 1{,}000{,}000$	1,000,001-10,000,000	10,000,001-100,000,000	$> 100{,}000{,}000$

Each customer is defined with R, F, M, B values using Python 3.7.3 programming language running on Jupyter notebook server version 5.7.8 as the processing interface. The results can be seen in Table 2, showing each customer gets a value of R, F, M, B which represents the value of loyalty. To get a score from each customer, a sum of R, F, M and B is added, so the formula for the Total Score is shown in Equation (1), and the total score for each customer can be seen in Table 3.

$$\textit{Total Score} = R + F + M + B \tag{1}$$

Table 2. Result examples for recency, frequency, monetary, balance calculation

Account	LastTrDate	Rec	Freq	Mon	Balance	R	F	M	B	RFMB
101056050	22-Jun-2017	9	72	112,083,706	9,157,730	4	3	4	2	4342
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
101102732	23-Jun-2017	8	74	217,936,797	12,987,100	4	4	4	3	4443

Table 3. Each customer's score

Account	R	F	M	B	RFMB	Total Score
101056050	4	3	4	2	4342	13
⋮	⋮	⋮	⋮	⋮	⋮	⋮
101102732	4	4	4	3	4443	16

Data from each attribute is searched for average value to get information on which customer groups and which attributes are above the average to get potential customers. Average data for each attribute based on the score obtained each can be seen in Table 4. The next step is to process the grouping using the K-Means algorithm grouping method. The stage begins by initializing a random cluster centroid vector, calculating the distance between the vector data with each centroid cluster using Equation (2), where Z_p is data point at position p, and M_j is a centroid from the data to cluster j. Then, recalculate it with Equation (3) and repeat until it meets the criteria, where n_j is a number of data point in cluster j. Repeat these steps until they stop and fulfill criteria. The satisfying criterion is amount iteration or changes in centroid position in successive iterations.

$$d(Z_p, M_j) = \sqrt{\sum_{k=1}^{d} (Z_p, k - M_j, k)} \tag{2}$$

$$M_j = \frac{1}{n_j} \left(\sum Z_p \right) \tag{3}$$

TABLE 4. Examples of recency, frequency, monetary, balance average based on total score

Total Score	Recency	Frequency	Monetary	Balance
4	113.35	1.57	235,937.41	120,375.83
5	102.48	2.84	1,750,359.49	479,599.24
⋮	⋮	⋮	⋮	⋮
12	9.81	32.19	70,547,451.28	24,978,170.58
16	5.67	135.33	2,137,014,500.00	207,613,333.33

The data is the result of segmentation of the RFM+B model which was previously normalized. At this stage, the data composition of segmented customers is based on the values in the RFM+B attribute that have been analyzed based on transaction behavior and nominal balances owned and obtained an accuracy level from the results of the grouping. Then the data is validated by the business unit. Then the data will be used as a reference in determining the company's strategy, specifically in product marketing.

4. **Proposed Model.** In this paper, we proposed the RFM+B model. It is used for segmenting customers based on customer transactions and balances. It is considered able to accommodate the needs of banking customer segmentation in this study which includes the customer's balance in the analysis process [6]. The proposed method for developing the method is expected to contribute to research related to the topic taken. The analytical model developed is the RFM model that has been used in previous studies in terms of customer/customer segmentation. The model was developed by adding a balance variable to the analysis process. Balance in the analysis process is to produce customer segmentation, as a reference for business units in evaluating marketing strategies and customer growth strategies according to targets.

4.1. **Model flowchart.** In the following flowchart can describe the sequence of processes in detail and show the relationship between one process with another process. The process sequence in the development of this segmentation model starts with data collection and ends with the segmentation process based on combined data from the supporting attributes. Master data is the result of integration between customer data and transaction data. Furthermore, the master data will be accumulated for each attribute owned by each customer. The result of the accumulation of each attribute is carried out of a process to get the value of each attribute used as a model that produces the attributes forming data segmentation, namely RFMB data. The process can be seen in Figure 3.

4.2. **Model parameters.** In the developed model, there are several parameters functioned in the analysis process. These parameters describe interrelated processes. The interrelationship of the processes that occur between these variables can be seen with the influence diagram in Figure 4. It shows that the component that becomes the output is the segmentation of savings customers. It depends and is formed by several variables derived from input variables and system component variables. For each variable there is a dependence on the data filling and the process carried out by other variables. The recency component variable depends on the input that originates from the input date of the last transaction, the frequency variable comes from the input of the total number of transactions, the monetary variable comes from the input of the total nominal transaction, while the balance variable comes from the system components namely customer balances. Output variables depend on a combination of recency, frequency, monetary, and balance variables.

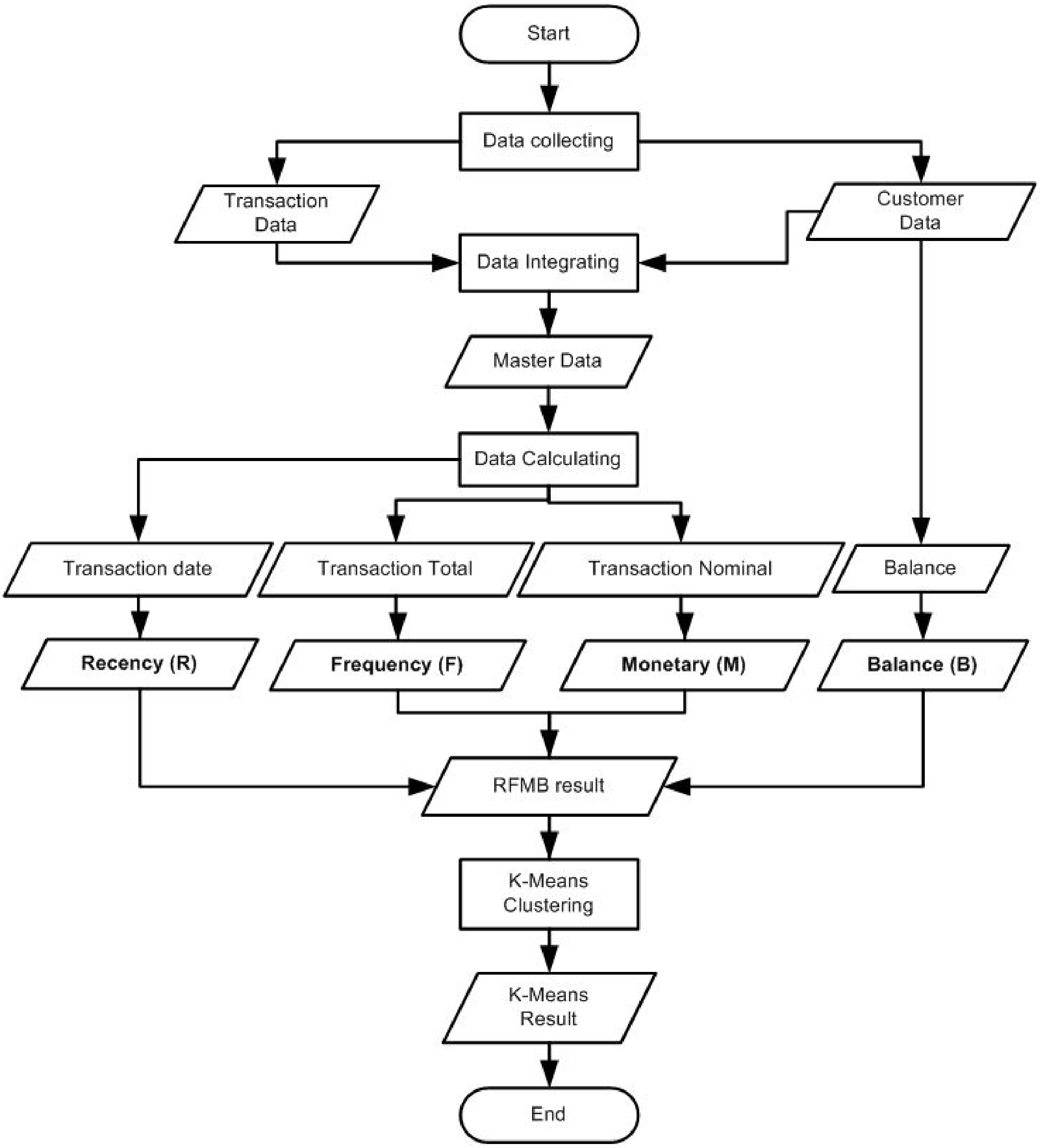

Figure 3. RFM+B model flowchart

4.3. **Model accuracy and clustering result.** Finally, RFM+B model's accuracy was measured. This process was performed by taking account of the SSE. SSE states the total sum of the square values of the distance of the data with the center of the cluster; a smaller SSE value means better clustering results [20]. The equation for this SSE can be seen in Equation (4), where d is the minimum distance calculated between the data and the cluster center point, p_i symbolizes a feature or attribute of the i-th data, and m_i represents a feature or attribute of the i-th cluster center point. Equation (4) produces a value to determine a significant difference in value. The attribute operated is the data from the segmentation of savings customers in the RFM+B model that has been obtained, and then the normalization process is carried out on these attributes to equalize the data ranges of each attribute R, F, M, and B, where the range is in between 0 and 1.

$$SSE = \sum_{i=1}^{k} d(p_i - m_i)^2 \tag{4}$$

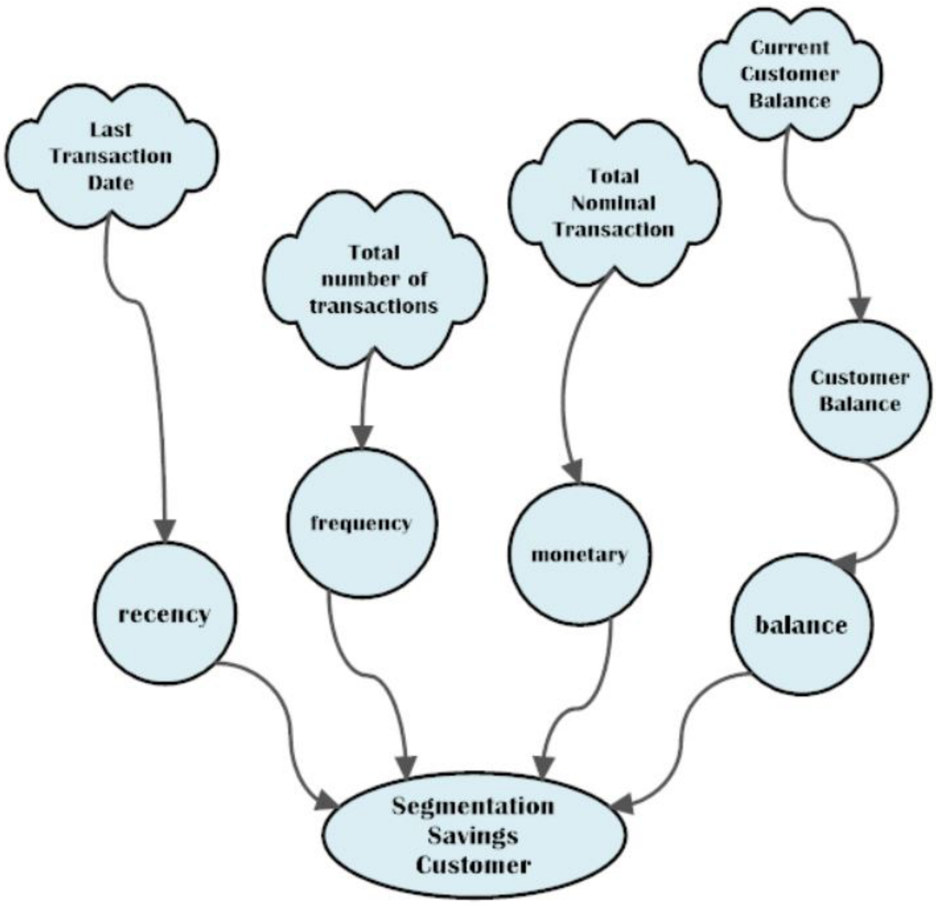

FIGURE 4. Influence diagram for model parameters

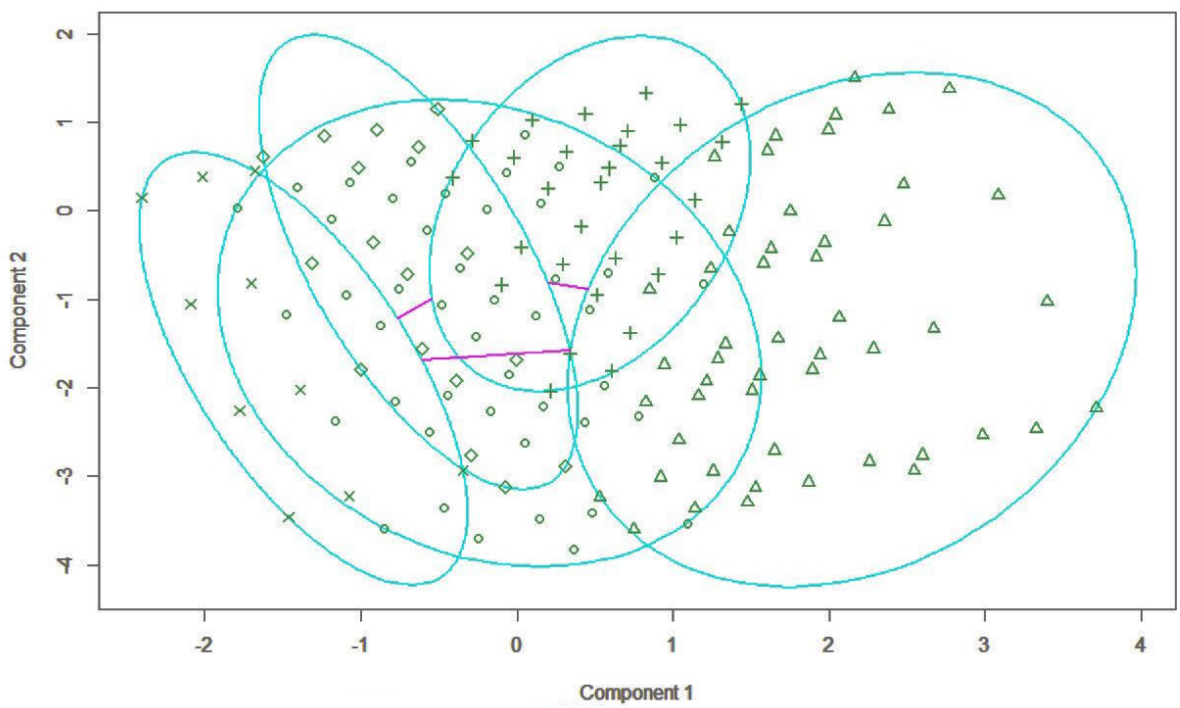

FIGURE 5. Component diagram of customer segmentation

Figure 5 presents the level of accuracy obtained from the process of grouping savings customers using data from the RFM+B model, and it is 77.58%. The level of accuracy obtained is then conveyed to the business unit to get a confirmation regarding the segmentation process, and it needs further development and re-segmentation process. With the level of accuracy obtained, business units at banking institutions can receive the results of this savings customer segmentation research. And the final results in the grouping process obtained five clusters as shown in Table 5 with average value for each attribute.

Table 5. Result of clustering

Cluster	Recency	Frequency	Monetary	Balance
0	113.17	3.58	8,421,517.58	12,534,863.29
1	62.00	11.00	244,640,902.75	3,487,790,000.00
2	9.95	315.73	2,117,917,775.39	17,245,277.77
3	92.00	2.00	5,000,000,000.00	27,241,600.00
4	18.67	17.73	36,010,389.79	15,598,664.16

5. **Conclusion and Further Works.** The RFM+B was successfully constructed. This is a customer segmentation model for bank in particular and for banking industry in general. It was developed based on K-Means clustering as a main model's method. This contribution is expected to generate company's profit and the growth of TPF in the company is also going to increase. The model as well is going to be used as a reference for marketing strategies. Based on average values that have been obtained, the most customers' R in cluster 0 is 113.17, the highest F in cluster 2 is 315.73, the highest M in cluster 3 is 5,000,000,000, and the largest customer's B in cluster 1 is 3,487,790,000. In addition, its contribution can be a reference in subsequent similar studies and become a new analytical model benefited in the banking industry. The work being possibly done in the future is a study regarding use of other or combined clustering methods that is going to make the model more optimal.

Acknowledgment. We would like to thank BINUS University who has supported our works, particularly BINUS Graduate Program, Master of Computer Science.

REFERENCES

[1] K. T. Kustina et al., Branchless banking, third-party funds, and profitability evidence reference to banking sector in Indonesia, *Jour. of Adv. Research in Dynamical & Control Systems*, vol.11, no.2, pp.290-299, 2019.
[2] J. Yi, Analysis and improvement strategy for profit contribution of bank customer under big data background, *Proc. of 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Changsha, pp.338-341, 2019.
[3] S. M. Kostic et al., Data mining and modeling use case in banking industry, *Proc. of the 26th Telecommunications Forum (TELFOR)*, Belgrade, pp.1-4, 2018.
[4] K. I. Moin and Q. B. Ahmed, Use of data mining in banking, *International Journal of Engineering Research and Applications (IJERA)*, vol.2, no.2, pp.738-742, 2012.
[5] C.-H. Cheng and Y.-S. Chen, Classifying the segmentation of customer value via RFM model and RS theory, *Expert Systems with Applications*, vol.36, no.3, pp.4176-4184, 2009.
[6] U. Firdaus and D. N. Utama, Balance as one of the attributes in the customer segmentation analysis method: Systematic literature review, *Advances in Science, Technology and Engineering Systems Journal*, vol.5, no.3, pp.334-339, 2020.
[7] Y.-H. Hu and T.-W. Yeh, Discovering valuable frequent patterns based on RFM analysis without customer identification information, *Knowledge-Based Systems*, vol.61, pp.76-88, 2014.
[8] M. Tavakoli et al., Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study, *Proc. of the 15th International Conference on e-Business Engineering (ICEBE)*, Xi'an, pp.119-126, 2018.

[9] A. Dursun and M. Caber, Using datamining techniques for profiling profitable hotel customers: An application of RFM analysis, *Tourism Management Perspectives*, vol.18, pp.153-160, 2016.

[10] S. Monalisa et al., Analysis for customer lifetime value categorization with RFM model, *Procedia Computer Science*, vol.161, pp.834-840, 2019.

[11] A. J. Christy et al., RFM ranking – An effective approach to customer segmentation, *Journal of King Saud University – Computer and Information*, 2018.

[12] S. H. Han et al., Segmentation of telecom customers based on customer value by decision tree model, *Expert Systems with Applications*, vol.39, no.4, pp.3964-3973, 2012.

[13] Z. Yang and X. Su, Customer behavior clustering using SVM, *Physics Procedia*, vol.33, pp.1489-1496, 2012.

[14] Dedi et al., Customer segmentation based on RFM value using k-means algorithm, *Proc. of International Conference on Informatics and Computing (ICIC)*, Semarang, pp.1-14, 2020.

[15] F. A. Bachtiar, Customer segmentation using two-step mining method based on RFM model, *Proc. of International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, pp.10-15, 2018.

[16] M. Aryuni, E. D. Madyatmadja and E. Miranda, Customer segmentation in XYZ bank using K-Means and K-Medoids clustering, *Proc. of International Conference on Information Management and Technology (ICIMTech)*, Jakarta, pp.412-416, 2018.

[17] W. Qadadeh and S. Abdallah, Customers segmentation in the insurance company (TIC) dataset, *Proc. of INNS Conference on Big Data and Deep Learning*, Edinburgh, pp.277-290, 2018.

[18] M. A. Mawoli and D. Abdulsalam, Effective market segmentation and viability of Islamic banking in Nigeria, *Australian Journal of Business and Management Research*, vol.1, no.10, pp.1-9, 2012.

[19] M. C. Hendrawa and I. P. G. H. Suputra, Customer segmentation using RFM model, *Electronic Journal of Computer Science Udayana (Jurnal Elektronik Ilmu Komputer Udayana)*, vol.8, no.2, pp.153-161, 2019.

[20] J. Qi, Y. Yu, L. Wang and J. Liu, K*-means: An effective and efficient k-means clustering algorithm, *Proc. of IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, Atlanta, GA, pp.242-249, 2016.

[21] S. Nawrin et al., Exploreing K-means with internal validity indexes for data clustering in traffic management system, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.8, no.3, pp.264-272, 2017.

[22] Z. Khan, J. Ni, X. Fan and P. Shi, An improved K-means clustering algorithm based on an adaptive initial parameter estimation procedure form image segmentation, *International Journal of Innovative Computing, Information and Control*, vol.13, no.5, pp.1509-1525, 2017.