2025. Volume 9, Number 1 *Pages 95-103*

UOT: 004.8:519.245:336.71:658.8
DOI: https://doi.org/10.30546/09090.2025.210.010

IMPROVED RFMT-BASED CUSTOMER SEGMENTATION IN BANKING USING FEATURE ENGINEERING AND CLUSTER EVALUATION

MUSA RAHIMOV

Azerbaijan Technical University

musa.rahimov@student.aztu.edu.az

ARTICLE INFO

Article history:
Received:2025-04-06
Received in revised form:2025-04-07
Accepted:2025-04-16
Available online

Keywords:
Customer segmentation;
RFMT model;
Feature engineering;
Banking sector;
Cluster analysis
2010 Mathematics Subject Classifications: 62H30, 68T05, 91B38, 91C20

ABSTRACT

Customer segmentation is a sensitive approach in the banking sector in terms of creating individual marketing strategies and effective customer relationship management. In this study, based on customer transaction data, the RFMT (Recency, Frequency, Monetary and Tenure) model was used and an improved segmentation methodology was proposed by applying feature engineering. Thanks to inverse calculation of transactions, the Recency indicator was determined by inverse calculation of transactions, unlike the existing literature. Frequency and Monetary were determined as the sum of different purchase frequencies and the sum of log transformed financial indicators respectively. The customer's duration and the average of the full payment percentage were utilized as the Tenure indicator. The optimality of the number of clusters was calculated by the Silhouette Score (0.5320), Davies-Bouldin Index (0.7958) and Calinski-Harabasz Index (8963.6659) and it was observed that 3 clusters were the most suitable choice among all. The suggested method contains both scientific innovation and practical application value in terms of more accurate analysis of customer behavior and the development of personalized services.

1. INTRODUCTION

Customer segmentation is of great importance in the banking sector in terms of developing individual marketing tactics and effectively managing customer relationships. To do so, the widely used RFM (Recency, Frequency, Monetary) model is recognized as one of the main utilities for evaluating customer behavior and increasing their loyalty (Kumar & Reinartz, 2018). The RFM model lays a foundation for segmentation based on key indicators such as the customer's recent purchase history (Recency), purchase frequency (Frequency), and spending level (Monetary). In this study, a new dimension — Tenure, which reflects the customer's relationship with the bank — was added to the RFM model, and the RFMT model was proposed (Fuster et al., 2022). At the same time, unlike the existing literature, different feature engineering approaches were applied for each indicator in this model. Thus, the Recency indicator is determined by the inverse of purchases and cash transactions, the Frequency indicator is determined by the sum of different frequencies, the Monetary indicator is determined by financial transactions with log transformation applied, and the Tenure is determined by the average of the service period and the full payment percentage.

The K-means algorithm was used to group customers, and the effectiveness of the model was evaluated by the Silhouette Score (0.5320), the Davies-Bouldin Index (0.7958) and the Calinski-

Harabasz Index (8963.6659) (Lloyd, 1982; MacQueen, 1967). The proposed RFMT approach makes significant contributions both in increasing the accuracy of modeling from a scientific point of view and in helping banks make more targeted marketing decisions from a practical point of view (Giudici et al., 2020).

2. RELATED WORK

The RFM (Recency, Frequency, Monetary) model has been one of the widely used methods in the field of customer segmentation for a long time. This model segments customers based on three main factors that determine their behavior: the last time the customer made a purchase (Recency), the frequency of purchases (Frequency), and the amount spent (Monetary) (Kumar & Reinartz, 2018). The RFM model is used to assess customer satisfaction and increase loyalty, as this model allows for a better understanding of customer value.

The RFMT (Recency, Frequency, Monetary, Tenure) model is an extended version of the RFM model. This model includes an additional factor that reflects the length of time the customer has been with the bank (Tenure). The tenure factor helps with understanding customer behavior in the long term. (Alharthi et al,2021) note that the RFMT model provides more accurate segmentation of customers and improves the assessment of customer value in the banking sector.

The k-Means method for cluster analysis is one of the most widely used algorithms in the field of customer segmentation. This method is distinguished by its simplicity and effectiveness. The algorithm divides customers into segments by grouping similarities in the data (MacQueen, 1967). Recent studies have shown that when the k-Means algorithm is applied in conjunction with the RFM and RFMT models, better results can be achieved in customer segmentation (Zhao & Zhang, 2020).

The use of RFM and RFMT models in conjunction with the k-Means algorithm is considered a powerful tool for predicting customer behavior and increasing customer loyalty in the banking sector. With the development of artificial intelligence technologies, combining these models with machine learning methods brings customer segmentation to a new level (Fuster et al., 2022).

Recent research has explored advanced feature engineering techniques to improve the accuracy of RFM-based customer segmentation models. For instance, Almeida and de Oliveira (2022) proposed an extended RFM model by incorporating additional behavioral and temporal variables, demonstrating improved clustering performance in retail banking. Similarly, Rahmani and Farrokhnia (2020) introduced a hybrid approach that combines k-means with customized features, yielding more homogeneous customer clusters. Jain and Singh (2023) compared traditional RFM clustering with machine learning-based segmentation methods and found that integrating artificial intelligence significantly enhances predictive power. These studies highlight the growing interest in optimizing customer segmentation frameworks using extended variables and intelligent algorithms.

3. METHODOLOGY

3.1 Dataset Description

The dataset was developed for customer segmentation purposes, covering data based on real customer behavior. The features in the dataset give chance to determine the usage of banking services, spending behaviors, and loyalty levels of customers. The dataset was obtained from the

Kaggle open data platform, which provides publicly available anonymized financial data for research and educational purposes. In total, this dataset contains data on 8950 active credit card users. 12 features of the dataset, which consists of 18 customer features, were used for feature engineering when building the RFMT model:

Table 1. Features selected for building the RFMT model

Feature name	About feature
PURCHASES_TRX	Number of transactions
CASH_ADVANCE_TRX	The number of cash advance transactions made by the customer.
PURCHASES_FREQUENCY	Purchase frequency
ONEOFF_PURCHASES_FREQUENCY	Frequency of single purchases
PURCHASES_INSTALLMENTS_FREQUENCY	Frequency of installment purchases.
PURCHASES	The total amount of purchases made by the customer
CASH_ADVANCE	Total amount of cash advance transactions.
PAYMENTS	Total amount paid by the customer
BALANCE	Customer account balance
MINIMUM_PAYMENTS	Minimum amount paid by the customer
TENURE	Duration of the customer's business relationship with the bank (indicated in months)
PRC_FULL_PAYMENT	Percentage of customers who paid in full

3.2 Data Preprocessing

Unfilled Values Management. NaN values were replaced with the average values of the columns to ensure completeness in the data set.

Infinite Value Provisioning. Infinite (inf) and negative infinite (-inf) values were replaced as NaN and then filled with corresponding values.

Log Transformation. A log transformation was applied to the M (Monetary) feature to reduce extreme values effect.

Normalization. Features in different units were brought to the range [0, 1] using the Min-Max Scaling method.

Quintile-Based Grouping. RFMT features were segmented with a score system from 1 to 5 and grouped for more convenient analysis.

Multidimensional Data Dimension. Normalized and transformed data for each feature ensures more accurate performance of the model.

Feature Selection. The main features to be used in modeling were determined within the RFMT model.

3.3 Feature Engineering

The feature engineering process aims to generate more appropriate features using the underlying features of the data set. In this study, the following methods were applied to create the features R (Recency), F (Frequency), M (Monetary), and T (Tenure). After normalization, each feature was segmented into five groups using quintile-based scoring from 1 to 5.

R (Recency) - Based on the time elapsed since the customer made a purchase last time. It is used to assign a higher "Recency" score to customers who make fewer transactions. The following formula was used for this value:

$$R\_Score = 1/(1 + PURCHASES\_TRX + CASH\_ADV\ ANCE\_TRX) \quad (1)$$

F (Frequency) - Indicates the customer's shopping frequency. It is used to identify customers who make more purchases:

$$F\_Score = PURCHASES\_FREQUENCY + ONEOFF\_PURCHASES\_FREQUENCY + PURCHASES\_INSTALLMENTS\_FREQUENCY \quad (2)$$

M (Monetary) - Indicates the total financial value of the customer. Used to identify customers with higher financial value. Calculated using the following formula:

$$M\_Score = log(1 + PURCHASES + CASH\_ADVANCE + PAYMENTS + BALANCE + MINIMUM\_PAYMENTS) \quad (3)$$

T (Tenure) - Reflects the duration of the customer's business relationship with the bank. It is used to identify customers who have been cooperating with the bank for a long time:

$$T\_Score = (TENURE + PRC\_FULL\_PAYMENT)/2 \quad (4)$$

The features created above were combined to create a four-digit RFMT score for customers. This score was used to segment customers:

$$RFMT\_Score = R\_Score + F\_Score + M\_Score + T\_Score \quad (5)$$

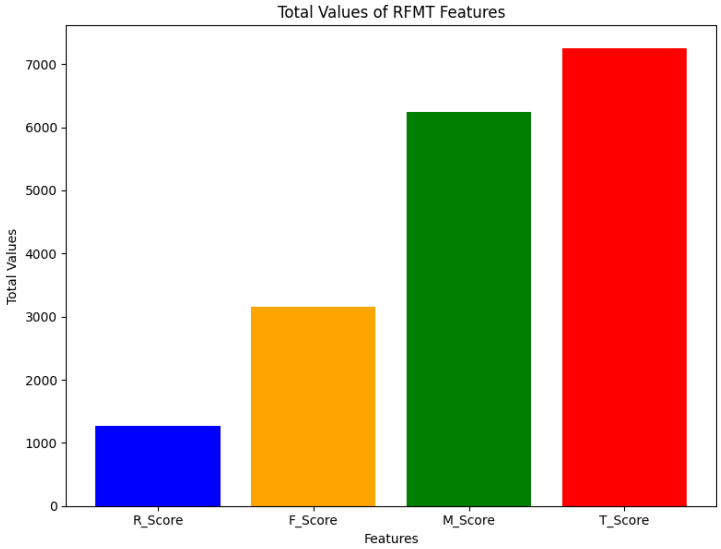

Fig. 1 Total values of the RFMT features.

Figure 1. This graph helps to understand the overall behavior of RFMT characteristics and illustrates how different features are distributed across the customer base. A high T_Score indicates that customers generally have a long-term relationship with the bank, whereas a low R_Score reflects high customer activity, as Recency was computed inversely based on transaction counts.

3.4 Model and Evaluation Metrics

The K-Means clustering algorithm is one of the most popular and widely used methods for splitting observations into a specified number of clusters. This algorithm iteratively assigns observations to clusters that are close to their mid points and recalculates the cluster centers. The K-Means method is considered an ideal choice for analyzing and segmenting customer behavior

in large data sets (Lloyd, 1982). The algorithm performs optimization by minimizing the distance to the cluster centers for each observation:

$$J = \sum_{i=1}^{k} \sum_{j \in C_i} \|x_j - \mu_i\|^2 \quad (6)$$

where x_j is the observation, μ_i is the cluster center, and C_i represents the i-th cluster.

Silhouette Score

The silhouette metric is used to assess the quality of clusters. This metric measures how close each observation is to its cluster and how far it is from other clusters. The silhouette value ranges from -1 to 1, where 1 indicates that the cluster is perfect and 0 indicates that the clusters overlap (Rousseeuw, 1987):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (7)$$

$a(i)$ – the average distance of the *i*-th observation within its own cluster

$b(i)$ – the average distance of the *i*-th observation to the nearest other cluster

Davies-Bouldin Index

The Davies-Bouldin metric is used to measure the similarity of clusters. It is estimated based on the distance between clusters and the internal variance of each cluster. A smaller Davies-Bouldin value indicates better clustering (Davies & Bouldin, 1979):

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_{j)}} \right) \quad (8)$$

σ_i– the average intra-cluster distance (dispersion) of the *i*-th cluster, $d(c_i, c_{j)}$ – the distance between the centroids of clusters *i* and *j*.

Calinski-Harabasz Index

This index is used to measure the density and distance of clusters. A higher Calinski Harabasz value indicates better clustering quality (Calinski & Harabasz, 1974):

$$CH = \frac{tr(B_{k)}}{tr(W_{k)}} \times \frac{n - k}{k - 1} \quad (9)$$

$tr(B_{k)}$– the trace of the between-cluster dispersion matrix, $tr(W_{k)}$ – the trace of the within-cluster dispersion matrix, n – the total number of samples, k the number of clusters.

Elbow Method

The Elbow method is one of the most widely used methods for selecting the optimal number of clusters. In this method, the distortion value (inertia) decreases as the number of clusters increases. The "elbow" point on the graph determines the optimal number of clusters (Thorndike, 1953):

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (10)$$

k the number of clusters, $x \in C_i$ observations belonging to the i-th cluster, μ_i the centroid of the i-th cluster, $\|x - \mu_i\|^2$ the squared distance between the observation and the cluster centroid.

4. RESULTS AND DISCUSSION

In this section, the results obtained were studied. With the purpose to better understand the behavior of customers, cluster analysis was applied using R, F, M and T indicators. Thus, cluster analysis was performed for different numbers of clusters from 2 to 10 and metrics such as Silhouette Score, Davies-Bouldin Index and Calinski-Harabasz Index were calculated for each cluster number, as shown in Table 2.

Table 2. Results of calculated metrics for the k-Means model

Number of Clusters	Silhouette Score	DAVIES-BOULDIN INDEX	Calinski-Harabasz Index
2	0.4344	1.1015	2631.6389
3	0.5320	0.7958	8963.6659
4	0.4877	0.8370	8551.6570
5	0.3821	0.8444	6468.1547
6	0.3925	0.9032	6257.1984
7	0.4682	0.9079	7016.9371
8	0.4567	0.9596	6373.1973
9	0.4625	0.9578	6181.9320

The highest value of the "Silhouette Score" performance metric used to assess the quality of cluster separation was acquired when the number of clusters was 3. This shows the pros of internal consistency and separation of the clusters. The lowest value for the "Davies-Bouldin Index" indicator was also observed for 3 clusters, which indicates that the inter-cluster difference is greater, and the internal variation is less. The highest result for the "Calinski-Harabasz Index" metric was obtained for 3 clusters, as 8963.67, which is a quite powerful result. Based on all these metrics, it was determined that the optimal number of clusters is 3; because in this case the Silhouette Score is maximal, the Davies-Bouldin Index is minimal, and the Calinski-Harabasz Index is high. Additionally, it can be noted that as the number of clusters increases above 5, both the Silhouette Score and the Calinski-Harabasz Index tend to decrease, while the Davies-Bouldin Index increases. These results indicate that selecting more than 5 clusters is not advisable in terms of marketing strategies.

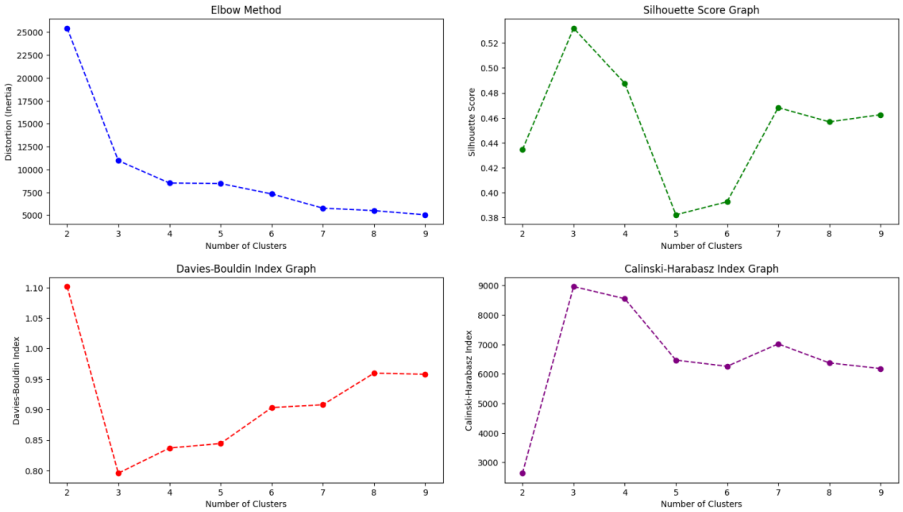

Fig. 2 Evaluation of Optimal Cluster Number Using Different Clustering Metrics

Figure 2 illustrates the results of four different clustering evaluation metrics—Elbow Method (Inertia), Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index—for cluster counts ranging from 2 to 9. According to the graphs, the optimal number of clusters is determined to be 3, as it yields the best balance across all metrics.

The Elbow plot presented in Figure 1 confirms that the optimal number of clusters is k=3. At this point, it is determined that the model is balanced, neither information loss is reduced, nor the complexity increased.

The fact that the optimal number of clusters obtained for all performance metrics is 3 and the results are good is an indicator of the effectiveness of the selected model and feature engineering.

Table 3. Statistical Analysis of Customer Behavior Across Clusters

Cluster	Number of Customers	Average Recency	Average Frequency	Average Monetary	Average Time Interval
0	4485	1.522185	1.272464	3.945819	4.940691
1	881	1.419977	1.891033	3.807037	1.938706
2	3584	1.001116	3.921875	4.055246	4.928292

According to the Table 3, it can be said that an analysis was conducted for almost every cluster and it was determined that cluster 2 mainly represents frequent, high-spending, and long-term customer behavior. The lowest Recency indicator in this cluster indicates high customer activity. At the same time, the highest values of Frequency and Monetary indicators prove that these customers use bank services regularly and efficiently. The high Time indicator also indicates that these customers have a long-term and stable relationship. Therefore, cluster 2 constitutes the main loyal and profitable customer base of the bank.

Cluster 1 consists of customers with an average level of activity. Although the Frequency and Monetary indicators in this cluster are at an average level, the low Time indicator indicates that the customers' relationship with the bank is short-term or non-permanent. The high Recency value compared to cluster 2 also indicates that the activity of this group is relatively low. Therefore, cluster 1 can be evaluated as premium customers with potential for development.

Cluster 0 is distinguished by its relatively high Recency and Frequency indicators, which indicate that they are less active. Also, although the Monetary value is relatively high, the frequency of purchases in this group is low. Although the high Time indicator indicates that these customers have a long-term relationship with the bank, it gives reason to say that their activity has decreased recently. This cluster consists of passive and reactivation potential customers.

Based on these results, banks can develop different marketing strategies according to customer segments. For example, for high-value and loyal customers in cluster 2, their satisfaction and commitment can be further increased through special loyalty programs, personalized services and reward campaigns. For customers in cluster 1, targeted marketing activities can be implemented to create continuous contact, send personalized offers and encourage them to be more active. Discount campaigns, personalized messaging and reactivation initiatives should be developed to reactivate passive customers in cluster 0.

Thus, the results of this analysis allow banks to develop targeted marketing strategies for different customer segments. This approach allows banks to both increase their revenues and

gain a competitive advantage. At the same time, it also makes a significant contribution to strengthening customer satisfaction and long-term relationships with the bank.

5. CONCLUSION

In this study, the RFMT (Recency, Frequency, Monetary, Tenure) model was used to analyze customer behavior and segment them. The model considered customers' purchase frequency, spending amount, last transaction history, and duration of business relationship with the bank as key indicators. The k-Means clustering algorithm based on the RFMT model was applied and customers were divided into three main clusters. The results obtained showed the behavioral characteristics of each cluster:

Cluster 2 consists of loyal customers with high purchase frequency and spending levels, high activity due to low Recency value, and a long-term relationship with the bank.

Cluster 1 includes premium customers who regularly spend a certain amount, show moderate activity, and demonstrate relatively stable behavior.

Cluster 0 represents customers with relatively high Recency and Frequency indicators, i.e., those who demonstrate less activity, are passive, and have a low purchase frequency..

Based on these results, it is proposed to develop personalized marketing strategies for different customer groups. It is recommended to introduce loyalty programs to strengthen relationships with loyal customers, develop exclusive offers for premium customers, and organize special promotional campaigns to increase the activity of passive customers. This approach will allow banks to increase customer satisfaction, as well as increase their revenues and gain a competitive advantage.

The results of the study confirm the effectiveness of the RFMT model and the k-Means algorithm in analyzing customer behavior. In addition, it is recommended to improve the models by introducing additional indicators in customer segmentation for future studies. At the same time, it may be useful to apply other clustering methods and test the models in different industries. This approach will contribute to better communication with customers and providing services that meet their needs.

REFERENCES

1. Kumar, V., & Reinartz, W. (2018). *Customer Relationship Management: Concept, Strategy, and Tools.* Springer.
2. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77, 5–47. https://doi.org/10.1111/jofi.12915
3. Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.
4. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
5. Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020). Network-based credit risk models. Quality and Reliability Engineering International, 36(1), 199–211. https://doi.org/10.1080/08982112.2019.1655159
6. Alharthi, M., Almalki, A., Alzahrani, H., & Khan, S. (2021). Enhancing customer segmentation using RFMT model in banking sector. *Journal of Financial Services Marketing*, 26(3), 125–138.
7. Zhao, L., & Zhang, W. (2020). Integration of k-means clustering with RFM model for customer segmentation. *International Journal of Data Science and Analytics, 9*(3), 155–167.
8. Almeida, T. A., & de Oliveira, M. P. V. (2022). Customer segmentation in retail banking using extended RFM and clustering techniques. *Journal of Retailing and Consumer Services, 66*, 102929.
9. Jain, A., & Singh, A. (2023). Comparative analysis of machine learning algorithms for customer segmentation using behavioral data. *Expert Systems with Applications, 215*, 119208.
10. Rahmani, M., & Farrokhnia, M. R. (2020). Feature engineering for enhancing RFM-based customer segmentation: A hybrid clustering approach. *Information Systems and e-Business Management*, 18(3), 497–514.
11. Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.
12. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.
13. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224-227.
14. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1-27.
15. Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4), 267-276.