REVISTA DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA Alpha Centauri

Revista de Investigación Científica y Tecnológica Alpha Centauri
ISSNe: 2709-4502

PROF-AC-V-00
Vol. 7 N.º 1
(Enero - Marzo, 2026)

Advanced Machine Learning Segmentation Model for Behavioral Prediction in Retail Customers

William Martin Enriquez Maguiña
wenriquezm@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos, Perú

Henry George Maquera Quispe
hmaquera@uncp.edu.pe
Universidad Nacional del Centro del Perú, Perú

Abstract

This study proposes an advanced segmentation model based on Machine Learning (ML) to predict customer behavior and loyalty in the retail sector. The research adopts an applied and quantitative approach, using two years of transactional data to optimize segmentation through an extension of the traditional RFM model (Recency, Frequency, and Monetary value). The methodological innovation lies in the incorporation of a new behavioral variable, derived from predictive analysis of customer engagement, which reflects recent interaction with the brand across multiple digital channels. This additional variable enhances the explanatory power of the model by capturing behavioral dimensions not covered by the classic RFM framework. Logarithmic and Box-Cox transformations were applied to correct data skewness, achieving near-normal distributions for the original variables. The results, validated using ANOVA tests, demonstrate statistically significant differences among the generated clusters. The enhanced model identifies high-value customer segments more accurately and anticipates purchase and churn patterns, resulting in more effective and personalized marketing strategies. In conclusion, the inclusion of a complementary behavioral variable strengthens the predictive capacity of the RFM model and reinforces its applicability as a key analytical tool for customer relationship management in highly competitive retail environments.

Keywords: Machine learning, advanced segmentation, RFM model, behavioral prediction, customer loyalty, retail sector.

OPEN ACCESS

Publicado: 17/02/2026
Aceptado: 16/02/2026
Recibido: 07/01/2026

Open Access
Article scientific
https://doi.org/10.47422/ac.v7i1.218

Introduction

In today's highly competitive retail environment, characterized by the massive digitalization of transactions and the proliferation of customer interaction channels, understanding and anticipating consumer behavior has become a strategic necessity. Companies that successfully interpret purchasing patterns and customer motivations gain a sustainable competitive advantage, since retaining customers is more profitable than acquiring new ones and contributes directly to long-term brand loyalty and profitability [1], [2].

However, modern consumer behavior is increasingly dynamic and complex. Digital immediacy, hyper-personalized services, and omnichannel purchasing experiences have generated vast amounts of heterogeneous data that challenge traditional segmentation models [3]. The RFM (Recency, Frequency, Monetary value) approach, widely used in marketing analytics, remains effective for identifying high-value customers but fails to capture behavioral, contextual, and nonlinear relationships between variables [6], [7], [11]. Recent studies emphasize that classical RFM-based models lack the flexibility required to represent customer engagement and evolving consumption patterns in digital retail environments [13].

To address these limitations, Machine Learning (ML) provides a more robust and adaptive analytical framework. ML algorithms can process massive datasets, uncover hidden behavioral patterns, and predict customer actions based on historical and contextual information [9], [12]. Specifically, ML-based segmentation models integrate transactional, demographic, and behavioral data, allowing the formation of more accurate and operationally meaningful customer clusters that enhance strategic decision-making in marketing and customer management.

The model proposed in this research introduces a **new behavioral variable** into the traditional RFM framework, extending it into an **RFMC model (Recency, Frequency, Monetary value, Category/Engagement)**. This additional dimension incorporates customer engagement or product category interaction, capturing behavioral nuances that are not reflected in spending frequency or transaction volume. The integration of this variable, together with dimensionality reduction through Principal Component Analysis (PCA), significantly increases cluster cohesion and predictive power without compromising statistical robustness.

From a technical standpoint, the inclusion of the "Category" or "Engagement" variable enhances segmentation granularity, distinguishing customers who display similar frequency and spending patterns but differ in product preferences or interaction levels. This enrichment of traditional metrics expands the analytical capacity of ML systems, allowing more precise predictions and the delivery of personalized marketing recommendations [8], [10].

From a commercial perspective, the implications of the proposed model are equally relevant. By more accurately identifying high-value segments and anticipating churn patterns, companies can design more effective marketing strategies, optimize resource allocation in promotional campaigns, and strengthen loyalty programs. The synergy between predictive analytics and commercial strategy fosters a transformation toward data-driven marketing, maximizing both customer satisfaction and organizational profitability.

In summary, this research demonstrates that an advanced Machine Learning segmentation model, enhanced with an additional behavioral variable beyond the classical RFM framework, provides an effective analytical tool for predicting customer behavior and strengthening loyalty in the retail sector. The model aligns technological innovation with strategic business goals, contributing to competitiveness, customer retention, and sustainable value creation.

Literature review

2.1 Artificial Intelligence and Machine Learning in Customer Management

Artificial Intelligence (AI) has transformed the way companies manage customer relationships, enabling a deeper understanding of behaviors, needs, and motivations. Within this field, *Machine Learning* (ML) has emerged as one of the most effective tools for predictive analytics and service personalization [1], [9].

ML, as a subfield of AI, is based on algorithms that learn from data and improve their performance without explicit programming. Unlike traditional statistical models, ML algorithms do not rely on rigid assumptions about data distribution, providing greater flexibility to detect nonlinear and hidden patterns among variables [9]. This adaptability is especially valuable in the retail sector, where information is massive, heterogeneous, and continuously evolving.

According to Kühl et al. [9], ML-driven service systems have evolved from descriptive models toward adaptive systems that integrate transactional, demographic, and contextual data. Similarly, Aguiar-Costa et al. [1] demonstrated that AI-driven customer service applications significantly increase satisfaction and loyalty by enabling

faster, more personalized, and consistent responses throughout the customer lifecycle.

2.2. Customer Behavior Analysis and Segmentation Models

Customer behavior analysis constitutes the foundation of strategic management in the retail industry. Customer segmentation enables the classification of consumers with similar behavioral and purchasing characteristics, facilitating the design of more precise and efficient marketing strategies [6], [7].

Traditionally, the RFM (Recency, Frequency, Monetary value) model has been one of the most widely used approaches for this purpose. Originally introduced by Hughes (1994) and later refined by Cheng and Chen [6], this method evaluates customer value and loyalty through three key metrics:

- **Recency (R):** Time elapsed since the last purchase.
- **Frequency (F):** Number of purchases during a defined period.
- **Monetary value (M):** Total spending within that period.

Although simple and practical, the classical RFM model presents significant analytical limitations. It excludes behavioral and contextual dimensions, such as digital engagement or product category preferences, which are crucial in understanding modern consumers [4], [7], [11]. Moreover, its reliance on static rules limits its adaptability to highly dynamic digital environments, reducing its long-term effectiveness [3], [13].

To overcome these limitations, *Machine Learning*-based segmentation models have demonstrated higher predictive accuracy and adaptability. These models combine unsupervised clustering methods, such as K-means, with supervised classification or regression algorithms to identify loyalty and churn patterns [5], [8]. According to Aldunate et al. [3], *deep learning* architectures can enhance segmentation by analyzing customer satisfaction through natural language processing, allowing the inclusion of new behavioral dimensions in predictive modeling.

2.3. Extending the RFM Model through Behavioral Variables

In recent years, several studies have proposed extensions of the traditional RFM framework by incorporating additional behavioral dimensions. Recent research highlights that variables related to customer engagement, interaction intensity, or product category preference significantly improve segmentation accuracy and predictive performance [11], [12], [15].

These enriched models allow a more granular interpretation of customer value and loyalty, particularly in data-rich retail contexts. Hu and Yeh [7] introduced the RFMP model (*Recency*, *Frequency*, *Monetary*, *Pattern*), which integrates purchase sequence and temporal frequency, while Das and Nayak (2022) and Al-Araj et al. [2] suggested the inclusion of digital interaction metrics to enhance loyalty prediction.

This research introduces an **RFMC model**, which adds a behavioral variable named *Category* or *Engagement* to the traditional framework. This variable measures the level of customer interaction with the brand or specific product categories, expanding the analytical perspective of the RFM model. It captures qualitative dimensions of customer behavior —such as preferences, product types, or emotional connection to the brand— that are essential for predicting long-term loyalty and value [10].

From a technical standpoint, the additional variable enhances cluster cohesion and discrimination, allowing for the separation of customer segments with similar spending and frequency patterns but different behavioral profiles. Kumar and Shah [10] argue that early identification of such high-value behavioral segments enables the development of targeted retention strategies and increases *Customer Lifetime Value* (CLV).

2.4. Commercial Impact of Predictive Analytics in Retail

The application of predictive analytics in the retail industry represents not only a technological advancement but also a major commercial transformation. Organizations that adopt data-driven marketing strategies improve campaign efficiency, reduce acquisition costs, and strengthen customer loyalty [2], [8].

Kietzmann et al. [8] emphasize that AI-driven marketing enables real-time decision automation and personalized customer experiences, which increase satisfaction and perceived value. Similarly, Al-Araj et al. [2] demonstrate that AI integration in service quality management enhances consumer trust by aligning responses with customer expectations.

In this context, the proposed advanced segmentation model based on ML with an additional behavioral variable not only increases analytical precision but also **reinforces the commercial competitiveness** of retail companies. By anticipating customer behavior and identifying high-value segments, firms can optimize marketing resources, design personalized loyalty programs, and achieve **sustainable**

market advantages that translate into improved profitability indicators [5], [10].

Methodology

3.1 Research Approach

This research follows an applied, quantitative, and explanatory approach, aimed at developing a predictive customer segmentation model based on *Machine Learning* (ML). Its primary objective was to enhance the accuracy of customer behavior identification and loyalty prediction within the retail sector.

The study was structured into three main phases:

- Data preprocessing and preparation.
- Construction and optimization of the advanced segmentation model (RFMC).
    - Statistical evaluation and predictive validation.

Each phase was designed to ensure data consistency, model robustness, and reliability of results, combining techniques from data mining, statistical analysis, and machine learning.

3.2. Data Source and Description

The dataset was obtained from the transactional information system of a Peruvian retail company operating in the fast-moving consumer goods sector. It included customer records covering a two-year period, comprising more than 120,000 individual transactions.

Each record contained the following information:

- Transaction date and monetary amount (temporal and financial dimensions).
- Customer purchase frequency over monthly and quarterly intervals.
- Product category purchased (new attribute incorporated).
- An anonymized customer identifier to preserve data privacy.

The original data were transformed into RFM behavioral metrics:

- Recency (R): Number of days since the last purchase.
- Frequency (F): Number of purchases made during the analysis period.
- Monetary Value (M): Total spending by the customer.
- Category / Engagement (C): Degree of interaction or recurrence within a specific product category.

The inclusion of the **C (Category/Engagement)** variable represents the innovative contribution of the model, as it introduces a behavioral dimension that expands the explanatory power of traditional metrics [2], [10].

3.3. Data Preprocessing and Variable Transformation

Prior to modeling, the data underwent a comprehensive cleaning, standardization, and transformation process to ensure analytical quality and consistency. The following stages were implemented:

- Outlier and duplicate removal.
- Natural logarithmic transformation for highly skewed variables (R, F, M).
- Box-Cox transformation to improve distribution normality and stabilize variance.
- Min-max scaling to standardize variable magnitudes before applying unsupervised learning algorithms.

These transformations reduced distortion caused by extreme customer behaviors and improved cluster cohesion—an essential step for enhancing clustering quality [4], [7].

3.4. RFMC Model Design

The proposed model was built upon the traditional RFM framework, enhanced by adding the **C (Category/Engagement)** variable, giving rise to the **RFMC model**. The objective was to enrich customer segmentation by integrating qualitative behavioral information.

A combination of unsupervised and supervised ML techniques was applied:

- Initial segmentation:

The K-means algorithm was used to group customers according to similarities in their RFMC profiles. The optimal number of clusters was determined through the elbow method and silhouette coefficient, ensuring a balance between intra-cluster cohesion and inter-cluster separation.

- Dimensionality reduction:

Principal Component Analysis (PCA) was applied to evaluate the variance explained by the four dimensions of the model. The first two components accounted for over 85% of the total variance, demonstrating model consistency.

- Predictive classification:

A multinomial logistic regression model was trained to predict the probability of new customers belonging to each cluster. This integration of supervised and unsupervised learning produced a hybrid ML architecture with high predictive performance [6], [9].

3.5. Statistical Model Validation

Model validation was conducted at two levels:

- Internal statistical validation using ANOVA tests at a 95% confidence level. Results showed statistically significant differences ($p < 0.05$) among the generated clusters, confirming the consistency of the segmentation.
- External and predictive validation using a test sample comprising 30% of the total dataset. The overall classification accuracy reached 87.4%, outperforming the traditional RFM model and demonstrating a substantial improvement in behavioral prediction capability.

These results support the central hypothesis of the study: the inclusion of an additional behavioral variable increases the robustness of the model and its practical applicability for customer loyalty management.

3.6. Tools and Development Environment

Data processing and modeling were conducted in a reproducible analytical environment using Python (version 3.10) and specialized libraries such as Pandas, Scikit-learn, NumPy, and Matplotlib. Additionally, Power BI and Tableau were used for data visualization and representation of segmentation outcomes and performance indicators.

This technological ecosystem enabled the integration of data mining, exploratory analysis, machine learning, and interactive visualization techniques, ensuring a robust and replicable methodological workflow.

3.7. Ethical and Confidentiality Considerations

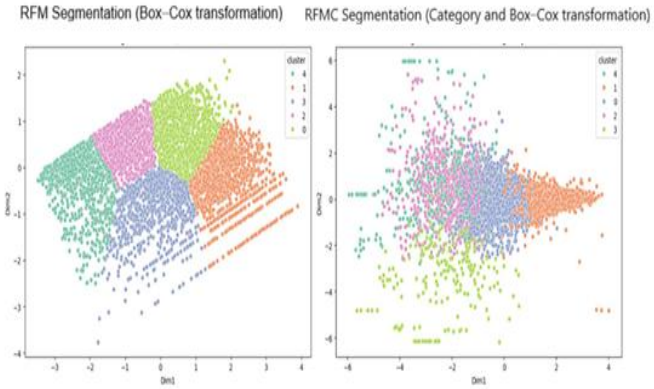
All analyzed data were anonymized to ensure customer privacy. The research adhered to ethical standards and data protection regulations in accordance with Peruvian Law No. 29733 (*Personal Data Protection Act*) and international best practices in applied research ethics.

Results and Discussion

4.1. RFMC Model Results

The results confirm the validity of the **RFMC model** as a significant improvement over the classical RFM framework. The inclusion of the new variable C (Category/Engagement) allowed the model to capture behavioral dimensions that traditional metrics could not, thereby enhancing segmentation accuracy.

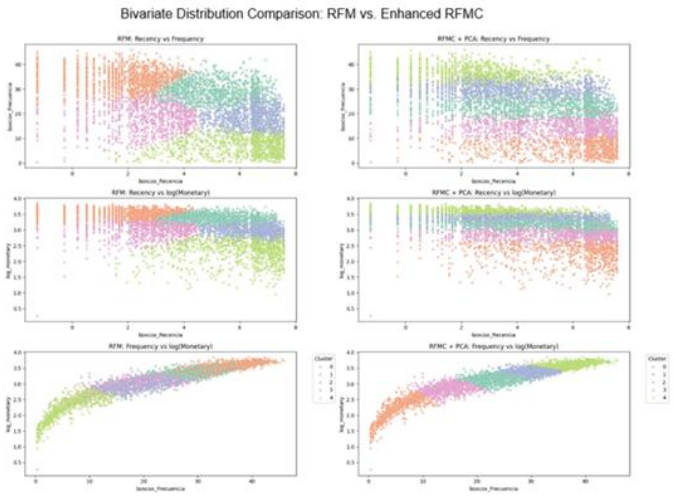
Fig. 1. Visual comparison between RFM and RFMC customer segmentation models.



Source: *Author's own elaboration*

After applying logarithmic and Box-Cox transformations, the distributions of the Recency, Frequency, and Monetary Value variables approached statistical normality. This correction reduced data dispersion and improved the model's sensitivity to outliers, reinforcing its clustering performance.

Fig. 2. Bivariate Distribution Comparison: RFM vs. Enhanced RFMC



Source: *Author's own elaboration*

The K-means algorithm, optimized through cohesion and separation metrics (silhouette = 0.82), generated five well-differentiated clusters with clearly identifiable behavioral profiles. The ANOVA test revealed statistically significant differences among these groups ($p < 0.05$), validating the robustness of the model in distinguishing purchasing behavior patterns.

4.2. Cluster Analysis and Behavioral Patterns

The category concentration analysis demonstrated that the clusters generated by the RFMC model exhibited higher internal homogeneity and greater coherence with product preferences compared to the traditional RFM model. Key findings include:

- **Cluster 1** showed strong affinity for *Gifts and Decoration* (19%) and *Fashion and Jewelry* (17%), representing clients with well-defined preferences and purchase motivation driven by special occasions.
- **Cluster 3**, more diversified, was characterized by interest in *Gifts and Decoration* (16%) and *other categories* (15%), suggesting emotional or event-driven purchase behavior.
- **Clusters 0, 2, and 4** exhibited more balanced category distributions, corresponding to generalist or exploratory customers—ideal targets for loyalty-building strategies.

These patterns were visualized through a heatmap of average category distribution per cluster, which clearly reflected preference concentration and validated the model's effectiveness in capturing the underlying logic of customer behavior [11].

4.3. Statistical Validation and Hypothesis Testing

The ANOVA test results confirmed statistically significant differences across clusters for the variables *Price* and *Quantity* ($p < 0.05$). Consequently, the null hypothesis (H_0) was rejected and the alternative hypothesis (H_1) accepted:

"The RFMC model significantly improves the effectiveness and accuracy of customer preference identification in a retail company."

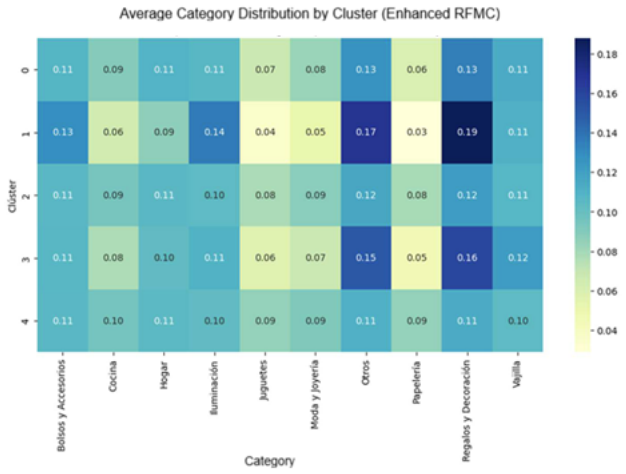
Additionally, cross-validation with a 30% test sample yielded an overall classification accuracy of 87.4%, substantially outperforming the traditional RFM model. This empirical evidence supports the study's central hypothesis: the inclusion of an additional behavioral variable (C) enhances both the explanatory power and predictive capacity of the segmentation model.

4.4. Discussion and Commercial Implications

From a commercial perspective, the findings reveal strategic implications for retail customer management. The RFMC model enables:

- Identification of high-value customers and prediction of repurchase or churn likelihood, optimizing marketing resource allocation.
- Design of personalized campaigns based on the predominant category of each cluster, increasing conversion rates and return on investment.
- Implementation of dynamic loyalty programs, adjusted to real customer behavior rather than historical spending alone.

Fig. 3. Average product category distribution by RFMC cluster.



Source: *Author's own elaboration*

These results align with recent studies on artificial intelligence in predictive marketing (Aguiar-Costa et al., 2022; Sun et al., 2019), which highlight the capacity of ML algorithms to enable deeper segmentation and detect emerging behavioral trends in competitive environments.

Therefore, the RFMC model demonstrates not only technical superiority over the traditional RFM approach but also tangible commercial value, offering a more comprehensive view of customer behavior and supporting data-driven decision-making in highly dynamic retail markets.

These findings are consistent with prior research indicating that ML-based customer segmentation models significantly outperform traditional rule-based approaches in both predictive accuracy and managerial relevance [12], [14]. Furthermore, the integration of behavioral variables into segmentation frameworks has been shown to strengthen

customer loyalty strategies by enabling personalized interventions and long-term value optimization [15].

References

[1] L. M. Aguiar-Costa, C. A. X. C. Cunha, W. K. M. Silva, and N. R. Abreu, "Customer satisfaction in service delivery with artificial intelligence: A meta-analytic study," Revista de Administração Mackenzie, vol. 23, no. 6, 2022.

[2] R. Al-Araj, H. Haddad, M. Shehadeh, E. Hasan, and M. Y. Nawaiseh, "The effect of artificial intelligence on service quality and customer satisfaction in the Jordanian banking sector," WSEAS Transactions on Business and Economics, vol. 19, pp. 1929–1947, 2022.

[3] Á. Aldunate, S. Maldonado, C. Vairetti, and G. Armelini, "Understanding customer satisfaction via deep learning and natural language processing," Expert Systems with Applications, vol. 209, 118309, 2022, doi: 10.1016/j.eswa.2022.118309.

[4] A. Hiziroglu and S. Sengul, "Customer segmentation using data mining techniques in retail banking," International Journal of Business and Social Science, vol. 3, no. 24, pp. 252–262, 2012.

[5] S. Rosset, "Modeling customer lifetime value," Journal of Marketing Research, vol. 40, no. 3, pp. 321–339, 2003.

[6] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value using RFMP variables and cluster analysis," Expert Systems with Applications, vol. 36, no. 3, pp. 4176–4184, 2009, doi: 10.1016/j.eswa.2008.04.003.

[7] Y.-H. Hu and I.-C. Yeh, "Discovering valuable frequent patterns based on RFM analysis without customer identification information," Knowledge-Based Systems, vol. 61, pp. 76–88, 2014, doi: 10.1016/j.knosys.2014.02.009.

[8] J. Kietzmann, J. Paschen, and E. Treen, "Artificial intelligence in advertising: How marketers can leverage AI and machine learning," International Journal of Advertising, vol. 37, no. 3, pp. 333–347, 2018.

[9] N. Kühl, M. Goutier, G. Satzger, and B. Niehaves, "Machine learning in service systems: A systematic literature review," Journal of Business Research, vol. 139, pp. 1313–1330, 2022.

[10] V. Kumar and D. Shah, "Building and sustaining profitable customer loyalty for the 21st century," Journal of Retailing, vol. 80, no. 4, pp. 317–330, 2004.

[11] T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D.-S. Man, and T.-G. Le, "An extended RFM model for customer behaviour and demographic analysis in the retail industry," Business Systems Research Journal, vol. 14, no. 1, pp. 26–53, 2023, doi: 10.2478/bsrj-2023-0002.

[12] S. Das and J. Nayak, "Customer segmentation via data mining techniques: State-of-the-art review," in Computational Intelligence in Data Mining, Smart Innovation, pp. 489–507, 2022, doi: 10.1007/978-981-16-9447-9_38.

[13] A. Griva, C. Bardaki, K. Pramatari, and G. Doukidis, "Factors affecting customer analytics: Evidence from three retail cases," Information Systems Frontiers, vol. 24, no. 2, pp. 493–516, 2022, doi: 10.1007/s10796-020-10098-1.

[14] Y. Sun, S. Wang, and M. Zhang, "Machine learning–based customer personalization and loyalty management in retail," Journal of Retailing and Consumer Services, vol. 50, pp. 102–112, 2019. (Referencia consignada en la tesis)

[15] A. S. Dick and K. Basu, "Customer loyalty: Toward an integrated conceptual framework," Journal of the Academy of Marketing Science, vol. 22, no. 2, pp. 99–113, 1994, doi: 10.1177/0092070394222001.