Proceedings of ICFTBA 2025 Symposium: Data-Driven Decision Making in Business and Economics
DOI: 10.54254/2754-1169/2025.BL29390

E-commerce Customer Segmentation Using Machine Learning

Yuqin Yang

Business School, Hunan International Economics University, Changsha, China
BeckyYyq1997@outlook.com

Abstract. E-commerce has increased exponentially over the past few years, with opportunities for growth as well as daunting tasks for businesses selling in competitive online spaces. One of the persisting challenges is how to identify and retain valuable customers whose purchasing behavior changes quickly in response to promotions, new product launches, or social media. This study responds by developing a hybrid segmentation model that incorporates Recency-Frequency-Monetary (RFM) measures with the K-Means clustering algorithm. Using transactional-level data, it construct behavioral features and apply clustering to detect patterns not normally captured by static thresholds. Four segments are revealed through analysis: high-value loyal purchasers, mid-value customers with growth potential, disengaged segments at risk for churn, and a small premium spending segment. RFM affords interpretability, and K-Means detects latent structure that yields analytical insight. Overall, the findings provide managers with concrete recommendations for loyalty programs, reactivation campaigns, and premium services, showcasing how machine learning can complement the role of traditional metrics in e-commerce.

Keywords: E-commerce, customer, machine learning

1. Introduction

Over the past two decades, e-commerce has shifted from being a side channel for retailers to becoming one of the dominant modes of trade. More than 2.6 billion people now shop online, and global transaction values are forecast to surpass 6.5 trillion U.S. dollars by 2025. While this surge creates enormous opportunities, it also means firms face fiercer competition and significantly higher customer acquisition costs. For the majority of firms, the question now is no longer that of acquiring new customers but that of keeping current customers and cementing their loyalty in the long term.

Conventional segmentation techniques, such as those based on categorizing customers simply on age, gender, or area, tend to miss the varied and dynamic behavior of digital markets. The Recency-Frequency-Monetary (RFM) model is widely used due to its ability to reduce customers' activity to three simple measures—how recently someone made a purchase, how frequently customers make purchases, and the amount they spend. Its simplicity facilitates its use as managers can easily implement it. Yet, by using fixed cutoffs, RFM is not an immediate means of reacting to sudden change in behavior, such as the spilling over of holiday sales orders or unanticipated demand due to

DOI: 10.54254/2754-1169/2025.BL29390

viral social media campaigns. Golden moments for real-time engagement are therefore lost occasionally [1-3].

Machine learning provides a more immediate method of customer segmentation than rule-based approaches. For example, K-Means is widely utilized by retailers in situations where datasets are too big or complex to process by grouping, since the algorithm is able to identify patterns that go unnoticed in static methods. With such techniques, firms are capable of reacting more quickly to changes in behavior and identifying valuable groups more accurately. Few studies have, however, integrated machine learning with behavioral measures such as RFM. This research fills that gap by proposing a hybrid model that combines the interpretability of RFM with the analytical power of K-Means, to further improve both accuracy and strategic applicability in e-commerce settings.

2. Methodology

2.1. Model overview

An effective segmentation strategy must be meaningful to managers but also subtle enough to pick up differences in customer behavior. RFM has been an effective marketing practice for some time now as it bears a direct relation to purchase behavior with three simple measures: recency, frequency, and monetary value. Managers like it because it is simple and its categories are intuitive. Static cut-offs make RFM miss important differences from time to time and lag in order to maintain pace with rapidly changing trends in online consumer behavior.

To go beyond these limitations, the RFM approach is blended with the K-Means clustering algorithm. Unlike threshold-based segmentation, K-Means splits customers according to patterns of similarity in their purchase history. The hybrid approach inherits the simplicity of RFM but becomes more flexible and specific. At a practical level, RFM gives a benchmark for customer value, and clustering finds more subtle patterns that would be missed by rigid rules. Other algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) or hierarchical clustering are capable of identifying more complex groupings, but the current study provides the RFM–K-Means combination with a shot for a practical balance between managerially usefulness and adequate analytical complexity [4-6].

2.2. Data source and preprocessing

The empirical analysis uses a publicly available e-commerce dataset from Kaggle, comprising 541,909 transaction records. Each entry includes variables such as invoice number, product code, description, quantity, invoice date, unit price, customer ID, and country. This rich dataset captures both behavioral and monetary dimensions of customer activity, providing a robust foundation for segmentation.

Data preprocessing involved several key steps to ensure quality and reliability. Transactions without customer IDs were removed because they could not be linked to individual profiles. Negative quantities, often representing returns, were excluded to avoid distorting spending calculations. Duplicate records were eliminated, and monetary values were calculated as the product of quantity and unit price, then aggregated per customer. Because monetary values were highly skewed. z-score normalization was applied to place all variables on a comparable scale and prevent any single feature from dominating the clustering process. Also, Finally, outliers were carefully examined and handled, as they can significantly affect centroid positioning in K-Means.

DOI: 10.54254/2754-1169/2025.BL29390

2.3. Analytical procedure

Table 1 and Figure 1, Figure 2, Figure 3, Figure 4 provide a detailed comparison of the four segments in terms of recency, frequency, monetary value, and revenue contribution. Segment A consistently dominates across most metrics, while Segment C shows the lowest engagement. Segment B represents a large base with moderate activity, and Segment D, although smaller, contributes through high-value transactions. These contrasts highlight how recency and frequency alone do not fully explain revenue distribution, reinforcing the need for a hybrid framework.

Beyond these broad distinctions, the segments differ in measurable ways that highlight managerial priorities. For instance, the average recency for Segment C is nearly three times longer than that of Segment A, suggesting a much higher probability of churn. Segment B, despite accounting for the largest share of customers, generates less than half the revenue of Segment A, underscoring the importance of strategies that increase purchase frequency. Segment D, though small, achieves an average transaction size significantly above other groups, which makes it an important contributor to profit margins even with fewer orders. These contrasts demonstrate that customer value cannot be inferred from size alone, but instead requires careful analysis of both behavioral intensity and monetary contribution.

3. Results and comparative analysis

3.1. Results and segment analysis

Table 1. Average recency-frequency-monetary values by customer segment

Segment	Recency (days)	Frequency (purchases)	Monetary (USD)	Size (%)	Revenue (%)
A - Loyal Very Important Person (VIP)	10	50	$5,000	15%	60%
B - Mid-Value	30	20	$1,500	40%	25%
C - At-Risk	90	5	$300	30%	10%
D - Premium Small	60	8	$2,500	15%	5%

Figure 1, Figure 2, Figure 3, Figure 4, Figure 5 highlight how differences in recency, frequency, monetary value, revenue, and product preferences translate into distinct customer behaviors and managerial priorities.

DOI: 10.54254/2754-1169/2025.BL29390

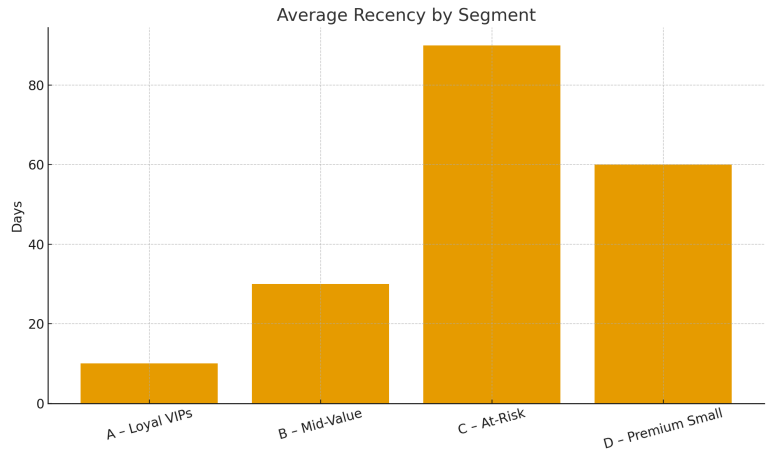

Figure 1. Average recency by segment (picture credit: original)

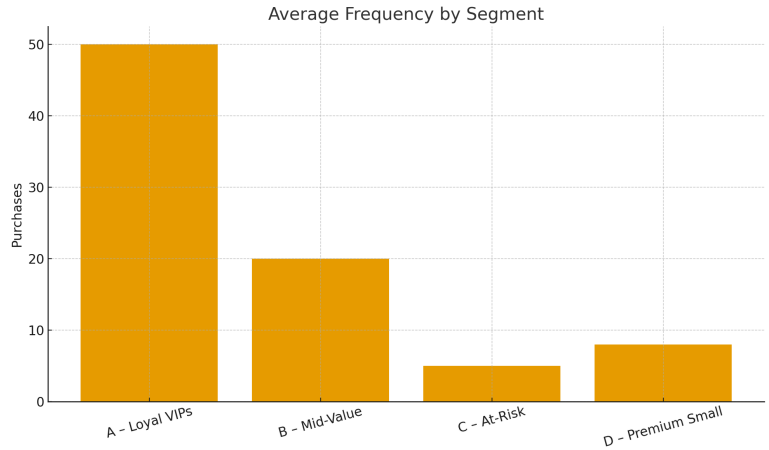

Figure 2. Average frequency by segment (picture credit: original)

Figure 1 shows that Segment A has the shortest recency, while Segment C has the longest. This contrast highlights the difference between highly engaged buyers and customers who are largely inactive.

Figure 2 indicates that Segment A records the highest purchase frequency, with Segment B following behind. Segment C shows minimal activity, while Segment D's frequency remains modest even though spending levels are high.

DOI: 10.54254/2754-1169/2025.BL29390

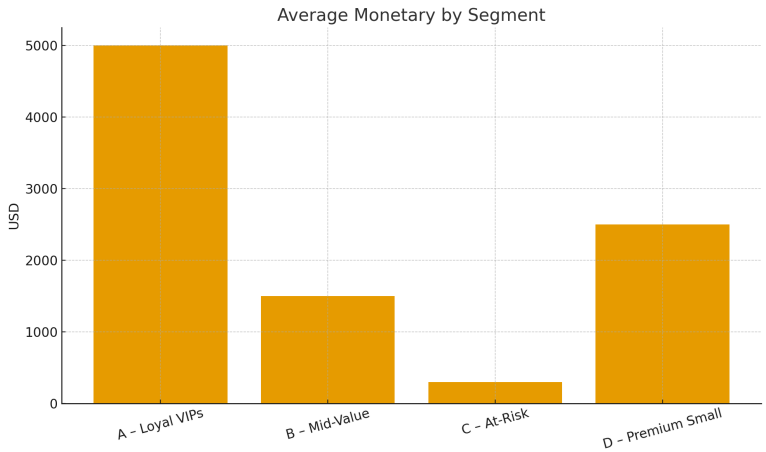

Figure 3. Average monetary value by segmenT (picture credit: original)

Figure 3 illustrates that Segment A generates the highest total spending, while Segment D stands out for high individual purchase values despite fewer transactions.

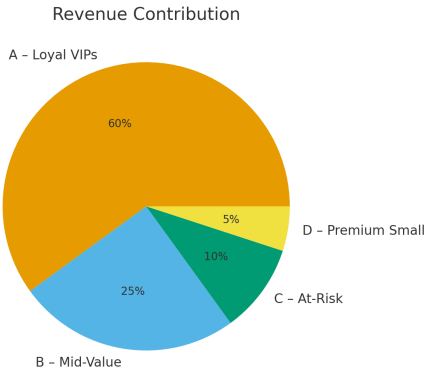

Figure 4. Revenue contribution by segment (picture credit: original)

Figure 4 shows that Segment A contributes more than half of total revenue. Segment B provides a moderate share, while Segments C and D account for smaller proportions.

DOI: 10.54254/2754-1169/2025.BL29390

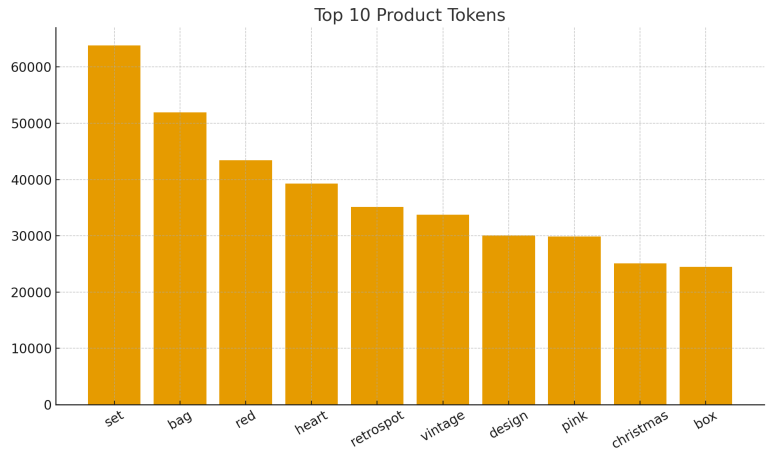

Figure 5. Top-10 product description tokens (picture credit: original)

Figure 5 expands the discussion by comparing product-level contributions across segments. For Segment A, purchases are concentrated in fast-moving consumer goods and seasonal promotions, reflecting their responsiveness to frequent campaigns. Segment B displays a more diversified product mix, with mid-range electronics and clothing as signature categories, suggesting opportunities for targeted cross-selling. Segment C's limited purchases are largely restricted to discounted or clearance products, reconfirming the threat of disengagement and the need for reactivation promotions. Segment D, although relatively small, is strongly associated with premium and limited-edition products, indicating potential to strengthen brand positioning through exclusivity.

3.2. Discussion and implications

3.2.1. Managerial implications

The findings yield several important conclusions for managers. Segment A, which generates most revenue, must continue to receive loyalty schemes and premium services. Segment B, less profitable on a per-customer basis yet having more customers than any other segment and therefore looking like an attractive target for cross-selling campaigns and promotion efforts, must receive attention. Segment C exhibits disengagement and will likely require reactivation efforts such as reminders, promotion, or seasonal discounts in order to stop loss from eroding further. Segment D, although small in customer count compared to other segments, exhibits spectacular per-customer revenue and must receive premium memberships or special editions [7].

In reality, the methods can be easily used in existing e-commerce programs. For Segment A, loyalty programs may include premium access to mega shopping events such as Singles' Day or Black Friday, and tiered reward escalation. Bundling and targeted suggestions, such as offering accessories with top-selling electronics, can tempt Segment B to spend more. Segment C is likely to be stimulated by time-sensitive coupons or reminder messages that bring them back to the platform. For Segment D, highly chosen premium clubs or high-end product drops could cement their spending inclination.

By bridging analytical insights with practical marketing tactics, businesses are in a better position to pass on segmentation learnings to measurable results, so strategic decisions are not only information-driven but also actionable [8].

DOI: 10.54254/2754-1169/2025.BL29390

3.2.2. Methodological insights

A comparison of the two techniques also highlights complementary strengths. RFM is intuitive with simple-to-explain benchmarks that managers can read at a glance. Compared to static cut-offs, K-Means uncovers subtle behavioral distinctions otherwise unexpected. Managers appreciate RFM since it is straightforward, and K-Means teases out subtle differences that RFM alone would miss. In combination, they render complex data more straightforward to distill into actionable strategy. The hybrid approach improves accuracy and provides assurance to managers that they can trust the results to make everyday decisions. Above all, it is very easy to implement using standard software, so firms without advanced analytics capabilities can utilize it. By striking a balance between simplicity and rigor, it enables companies to better implement the learnings in marketing strategy development. Most importantly, the method is not that hard to implement with regular analytics software, so even companies that are not in the situation to get fancy deep learning systems installed can make use of it.

3.2.3. Limitations

Some of the constraints should be pointed out. One, the sample is from one online shop, which restricts the generalizability of the results. Two, the RFM attributes in isolation are being analyzed and not behavioral and attitudinal traits such as browsing, product interest, or brand loyalty. Three, K-Means algorithm must be aware of the number of clusters beforehand and suffers from the problem of sensitivity to initial conditions and hence the results become less objective.

4. Conclusion

In this study, a hybrid model that joins the simplicity in readability of the RFM model with the analysis capability of the K-Means algorithm was investigated. The procedure describes how conventional behaviour markers can be enhanced with machine learning so that both implicit and explicit customer tendencies can be detected. The model gives recommendations for loyalty schemes, reactivation, and premium activities that illustrate the usefulness of the research to e-commerce managers.

Subsequent research can extend this work by including behavioral and attitude information such as browsing history, product selection, or involvement in loyalty programs. Testing more advanced methods, including deep learning, could reveal additional dynamics than clustering alone. Another area is real-time segmentation using streaming data, and firms can respond in a timely manner as consumers' behavior shifts. These measures would take customer management nearer to an adaptive and resilient shape.

References

[1] Wong, K., Chan, H., Lee, M. and Ho, J. (2024) Exploring Customer Segmentation in E-Commerce Using RFM Analysis with Hierarchical Clustering. Journal of Telecommunications and the Digital Economy, 12, 45–62.
[2] Zhao, Y. and Balagué, C. (2023) Customer Lifetime Value Prediction in E-Commerce: A Deep Learning Approach. Journal of Business Research, 154, 113–125.
[3] Haddadi, A., Rahimi, M., Shafiei, S. and Karami, A. (2025) A Hybrid Model for Improving Customer Lifetime Value. Information Sciences, 678, 119-134.
[4] Sun, Y., Li, H., and Wang, J. (2024) A Dynamic Customer Segmentation Approach by Combining LRFMS and Time-Series Clustering. Expert Systems with Applications, 241, 12-23.

DOI: 10.54254/2754-1169/2025.BL29390

[5] Zhang, R., Liu, P., and Chen, Q. (2025) Customer Segmentation Using RFM and K-Means Clustering to Support CRM in Retail Industry. Decision Support Systems, 172, 114-151.
[6] Vo, N., Pham, T., and Nguyen, L. (2025) Automatic K-Optimization and RFM-Based K-Means for Customer Segmentation. ACM Transactions on Management Information Systems, 13, 67–85.
[7] Wang, L., Zhou, Y., and Fang, M. (2025) Data-driven Strategic Customer Segmentation Considering Behavioral Dynamics. Information Sciences, 701, 120-122.
[8] Liu, J., Wang, R., and Zhang, X. (2025) Unveiling Consumer Patterns with Machine Learning: From RFM to Deep Clustering. Applied Intelligence, 63, 1123–1142.