SEA Practical Application of Science

SHARE SEA EMPOWER AWARENESS

Article

Machine Learning-Driven Customer Segmentation: A Behavior-Based Approach for F&B Providers

Jacint JUHASZ [1]

Abstract: This study explores behavior-based customer segmentation by integrating Recency, Frequency, and Monetary value (RFM) analysis with the K-Means++ clustering algorithm. Using one year of invoice-level transactional data from a Romanian Food and Beverage (F&B) provider serving restaurants and coffee shops, the research aims to deliver actionable insights to enhance marketing and sales strategies. After standardizing the dataset to address scale differences, the Elbow Method was applied to determine the optimal number of clusters, resulting in five distinct customer groups: Champions, Loyal Customers, Promising, Hibernating Customers, and Lost Customers. Notably, the Champion segment, consisting of a single customer, accounts for 15% of total sales, highlighting both profitability and dependence risks. Loyal and Promising customers were identified as the most strategically valuable segments for targeted retention and growth initiatives. The clustering results were validated through visualization techniques and internal metrics, confirming the effectiveness of the segmentation. By relying exclusively on transactional data, this approach ensures GDPR compliance and offers a scalable framework for continuous monitoring and dynamic strategy adaptation. The findings provide immediate financial implications for the company, illustrating the potential of machine learning-driven behavior-based segmentation in B2B markets with frequent, recurring transactions.

Keywords: RFM Analysis, K-Means++ Algorithm, Applied Machine Learning, Business Analytics, Food and Beverage Sector, GDPR-Compliant Data Analysis,

JEL code: C38, C81, M31, L81, D22,

Citation: Juhasz, J. (2025). Machine Learning-Driven Customer Segmentation: A Behavior-Based Approach for F&B Providers. *SEA - Practical Application of Science, Issue (39),* 169-176. https://doi.org/10.70147/s39169176

Received: 4 May 2025
Revised: 12 June 2025
Published: 15 June 2025

[1] Babes-Bolyai University, Faculty of Economics and Business Administration, Cluj-Napoca, Romania,

INTRODUCTION

Importance of Customer Segmentation

In the increasingly competitive business environment, understanding customer behavior has become crucial for companies aiming to enhance their marketing effectiveness and improve customer retention. Customer segmentation, the practice of dividing a customer base into distinct groups with shared characteristics, allows companies to tailor their strategies to the specific needs of each segment, maximizing customer satisfaction and profitability. According to Kotler and Keller (2016), segmentation is fundamental for targeting efforts and personalizing marketing actions, ensuring that resources are optimally allocated. Without a clear segmentation strategy, businesses risk delivering generic experiences that fail to resonate with their diverse customer base.

Different Customer Segmentation Methods

Traditionally, customer segmentation methods have spanned a wide range of techniques, including demographic, geographic, psychographic, and behavioral segmentation. Demographic segmentation divides customers based on observable characteristics such as age, gender, income, and education level (Wedel and Kamakura, 2000). Geographic segmentation considers location-based differences, while psychographic segmentation targets customer lifestyles, values, and personalities. Behavioral segmentation, which focuses on customers' interactions with products or services, is often seen as more actionable because it is directly linked to purchasing behavior (Kotler et al., 2015).

Recent advancements in data collection and analytics have enabled the development of more sophisticated segmentation techniques. Machine learning algorithms, particularly unsupervised learning models such as k-means clustering and hierarchical clustering, have been employed to uncover hidden patterns in large customer datasets (Xu and Tian, 2015). These methods allow for more dynamic and nuanced segmentations, accommodating complex customer behavior that traditional methods might overlook.

Advantages of Behavior-Based Segmentation

Among the various segmentation techniques, behavior-based segmentation has garnered significant attention due to its direct link to customer actions and value generation. Unlike demographic or psychographic approaches, behavioral segmentation leverages actual customer data such as purchase history, frequency of transactions, and monetary value spent. This empirical grounding provides a more accurate and predictive understanding of customer value, enabling firms to prioritize high-value customers and design personalized interventions (Baesens et al., 2015).

One of the most popular models for behavior-based segmentation is the Recency, Frequency, and Monetary value (RFM) analysis. Originally developed in the context of direct marketing, RFM analysis categorizes customers based on how recently they purchased (Recency), how often they purchase (Frequency), and how much they spend (Monetary value) (Hughes, 1996). The simplicity and interpretability of RFM metrics make them particularly attractive for practical applications.

An important advantage of behavior-based segmentation lies in the availability of data. Transactional data, which underpins RFM analysis, is routinely collected by companies as part of standard business operations, making the approach cost-effective and readily implementable without the need for extensive additional data collection efforts. Moreover, behavior-based segmentation typically involves no General Data Protection Regulation (GDPR) limitations, since it relies on anonymized transactional records rather than sensitive personal data like demographics or psychographics, thereby simplifying compliance requirements (Voigt & von dem Bussche, 2017).

Furthermore, behavior-based models are highly adaptable to company-specific information. Businesses can customize the RFM model by incorporating additional firm-specific variables, such as product categories or service usage patterns, enhancing the relevance and precision of the segmentation. This flexibility ensures that the model is aligned with strategic business objectives and industry-specific dynamics.

Another critical advantage is the potential for observing changes over time. Since behavioral data is continually updated, firms can track shifts in customer purchasing patterns and respond by adapting sales and marketing strategies dynamically. This temporal sensitivity supports proactive customer relationship management and helps firms stay ahead of emerging market trends (Reinartz & Kumar, 2003).

Finally, the results of behavior-based segmentation have immediate financial implications. By identifying high-value customers and understanding their behavior, firms can directly target their most profitable segments, optimize marketing expenditures, and maximize return on investment. This direct link between segmentation results and financial performance makes behavior-based segmentation a powerful tool for achieving quick wins and long-term profitability.

In this study, we propose a behavior-based customer segmentation approach by integrating RFM analysis with machine learning clustering techniques. By doing so, we aim to provide actionable insights that businesses can leverage to enhance customer relationship management and optimize marketing strategies.

DATA AND METHODOLOGY

Data Description. The dataset utilized in this study consists of transactional records obtained from a Food & Beverage (F&B) provider operating in Romania. The company supplies products to restaurants and coffee shops across the country and makes weekly deliveries to its customers.
The transactional data collected covers a period of one year and is recorded at the invoice level. The dataset includes individual customer identifiers, transaction dates, and transaction monetary values. Importantly, the focus is on behavioral data only, ensuring compliance with GDPR by avoiding the collection of personally identifiable information (Voigt & von dem Bussche, 2017).
Key Data Fields:

- Customer ID: Unique identifier for each customer.
- Transaction Date: Date on which the transaction occurred.
- Transaction Amount: Monetary value of each transaction.

The main characteristics of the data can be seen in Figure 1.

Data Preparation. Before conducting the analysis, several preprocessing tasks were performed to ensure data quality and relevance:

- Clearing Missing and Negative Values: All records with missing fields and negative transaction values were removed to avoid distortions in the analysis.
- Pivoting Transactional Data: The transactional dataset was aggregated at the customer level to compute the three key indicators: Recency, Frequency, and Monetary Value.
- Standardizing the Dataset: Each RFM variable was standardized to eliminate the effect of differing scales among the variables, ensuring fair contribution to the clustering process.

Methodological Framework. The methodological approach can be divided into two main stages: the construction of RFM metrics and the application of a clustering algorithm. RFM (Recency, Frequency, Monetary value) analysis is a time-tested technique used to summarize customer behavior using three dimensions:

1. Recency (R): The number of days since the customer's last purchase. Customers who purchased recently are more likely to respond to new offers.

$$R_i = \text{Current Date} - \text{Last Purchase Date of Customer i}$$

2. Frequency (F): The total number of transactions made by the customer in the observation period. Frequent buyers are often more loyal.

$$F_i = \text{Count of Transactions for Customer i}$$

3. Monetary Value (M): The total monetary value of all transactions made by the customer during the period.

$$M_i = \sum \text{Transaction Amounts for Customer i}$$

The resulting RFM metrics were standardized using the z-score normalization formula:

$$Z = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the respective variable.

K-Means++ Clustering. To segment the customers based on their standardized RFM scores, the K-Means++ clustering algorithm was employed. K-Means is a partitioning method that divides the dataset into a specified number of clusters kkk, assigning each observation to the cluster with the nearest centroid.
The goal of K-Means is to minimize the Within-Cluster Sum of Squares (WCSS):

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in C_i} |x_j - \mu_i|^2$$

where μ_i is the centroid of cluster i and x_j represents the data points within cluster i.
The Elbow Method was used to identify the optimal number of clusters. In this method, WCSS is calculated for different values of ***k***, and the results are plotted to observe the point at which the rate of decrease sharply slows, forming an "elbow."
Based on the Elbow Chart (Figure 2), the optimal number of clusters was set to **5**. This choice balances model complexity with explanatory power, ensuring that the clustering solution is neither too fragmented nor too coarse.

To improve the stability and robustness of the clustering solution, K-Means++ initialization was chosen over the standard K-Means approach. K-Means++ addresses the random initialization problem by selecting initial cluster centers in a way that spreads them out, which significantly reduces the chances of poor clustering results and local minima (Arthur and Vassilvitskii, 2007). The final clustering assigned each customer to one of the five clusters based on their RFM profiles.

Cluster Profiling. After clustering, each customer segment was profiled based on the average values of Recency, Frequency, and Monetary metrics. These profiles provide actionable insights by highlighting distinctive patterns in customer behavior, such as groups of high-frequency, high-monetary customers or infrequent, low-value buyers. Descriptive statistics for each cluster were computed to assist in the interpretation and meaningful labeling of customer segments.

Validation of Clusters. The robustness of the clustering solution was assessed using the Silhouette Coefficient, which measures how well-separated the clusters are:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where a(i) is the average intra-cluster distance and b(i) is the average nearest-cluster distance. Silhouette values closer to 1 indicate better-defined and more distinct clusters.

RESULTS

Elbow Method and Cluster Determination. To determine the optimal number of clusters, the Elbow Method was applied. As shown in Figure 1, the Within-Cluster Sum of Squares (WCSS) decreased considerable until k = 5, after which the rate of decrease is more flattened, indicating diminishing returns from adding additional clusters. Therefore, the optimal number of clusters was set to 5. (see Figure 1: Elbow Chart)

Definition of Clusters. Based on the clustering results, five distinct customer groups were identified and interpreted as follows (see Figure 3):

- Champion – The best customers who require special attention due to their high value.
- Loyal Customers – Customers who generate the core revenue and require continuous care and retention efforts.
- Promising – Customers who exhibit potential for growth and are good targets for upselling and cross-selling strategies.
- Hibernating Customers – Customers with limited possibilities; often older businesses that maintain stable but unchanging demand.
- Lost Customers – Customers where reactivation efforts are challenging and retention investments may not yield a positive return on investment (ROI).

Key Insights. As can be observed, the Champion category has only one customer, but this customer generates approximately 15% of the total sales. This relationship represents a highly profitable partnership, but it also introduces strategic risks due to the company's significant dependence on a single customer. Special attention must be given to maintaining and carefully managing this collaboration to mitigate potential risks associated with dependency.

From a strategic perspective, the most attractive segments are the Loyal Customers and the Promising customers. Loyal Customers form the core revenue base and must be continuously nurtured to ensure retention. Meanwhile, Promising customers present growth opportunities, making them ideal targets for customized marketing campaigns and account management strategies aimed at increasing share of wallet.

On the other hand, Hibernating Customers represent businesses with stable but low and unchanging demands, typically older restaurants with little inclination towards expansion or upselling. While they offer stability, the upside potential is limited.

Finally, Lost Customers require special reactivation efforts if they are to be won back. However, the cost-effectiveness of targeting this group must be carefully evaluated, as the investment required for retention might not always yield a positive ROI.

Visualization of Clusters. In the graphs presenting the segments, pairs of RFM indicators (Recency-Frequency, Frequency-Monetary, Recency-Monetary) were plotted. As shown in Figures 3–5, the clustering algorithm effectively captured the differences between the groups, demonstrating clear separations across the segments.

These visualizations not only confirm the robustness of the clustering results but also serve as practical tools for designing differentiated marketing and sales strategies tailored to each customer group. Leveraging these insights can significantly enhance the company’s profitability by focusing efforts on the right customer segments with appropriate strategies.

CONCLUSIONS

This study demonstrated the value of applying RFM-based behavioral segmentation combined with the K-Means++ clustering algorithm to a real-world dataset from an F&B provider operating in Romania. The method offered several advantages:

- It leveraged available transactional data without requiring additional costly data collection.
- It avoided GDPR limitations, working only with anonymized behavioral data.
- It incorporated company-specific information, reflecting the particularities of the weekly delivery and consumption patterns of restaurants and coffee shops.
- It enabled ongoing monitoring of customer behavior, supporting dynamic adaptation of marketing and sales strategies.
- It produced results with immediate financial implications, offering clear targets for revenue growth and customer retention efforts.

By identifying key segments such as Champions, Loyal Customers, and Promising customers, and understanding their unique characteristics, the company can better allocate its resources and tailor its customer management strategies.

In summary, behavior-based customer segmentation using RFM analysis and machine learning clustering offers a powerful, practical tool for companies aiming to better understand their customers and optimize their market strategies. Future research could expand the model by integrating additional behavioral metrics, such as customer engagement or product preferences, and by exploring more advanced clustering techniques for even deeper insights.

REFERENCE LIST

[1] Arthur, D., & Vassilvitskii, S. (2007). *k-means++: The advantages of careful seeding*. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035.

[2] Baesens, B., Verstraeten, G., Van den Poel, D., Vanthienen, J., & Dedene, G. (2005). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research*, 156(2), 508–523. https://doi.org/10.1016/j.ejor.2003.11.005

[3] Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

[4] Hughes, A. M. (1996). *The complete database marketer: Second-generation strategies and techniques for tapping the power of your customer database*. McGraw-Hill.

[5] Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson Education.

[6] Kotler, P., Keller, K. L., Hoon, A. C., & Wee, C. H. (2015). *Marketing management: An Asian perspective* (6th ed.). Pearson Education.

[7] Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, 67(1), 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589

[8] Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer.

[9] Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Springer Science & Business Media.

[10] Xu, R., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

LIST OF FIGURES

	DocumentId	PartnerId	DocumentDate	Sales
count	9818.000000	9818.000000	9818	9818.000000
mean	799375.430739	6961.042575	2022-06-29 09:40:39.845182464	829.048788
min	756724.000000	4.000000	2022-01-03 00:00:00	-79375.000000
25%	776993.250000	1141.000000	2022-04-05 00:00:00	270.000000
50%	800005.000000	8191.000000	2022-06-30 00:00:00	588.000000
75%	821269.750000	10961.000000	2022-09-22 00:00:00	1027.000000
max	843594.000000	12812.000000	2022-12-30 00:00:00	24313.000000
std	25175.856289	4416.929374	NaN	1446.303041

Figure 1.
Descriptive statistics of the database

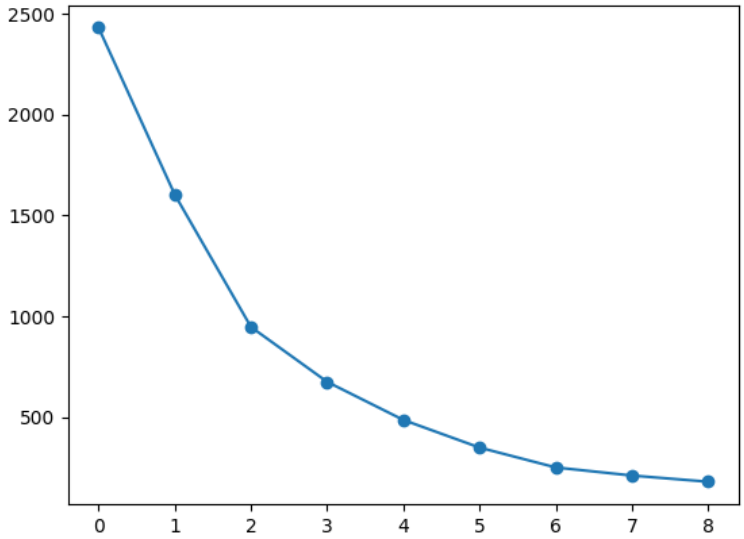

Figure 2.
Elbow-plot to determine the optimal number of clusters

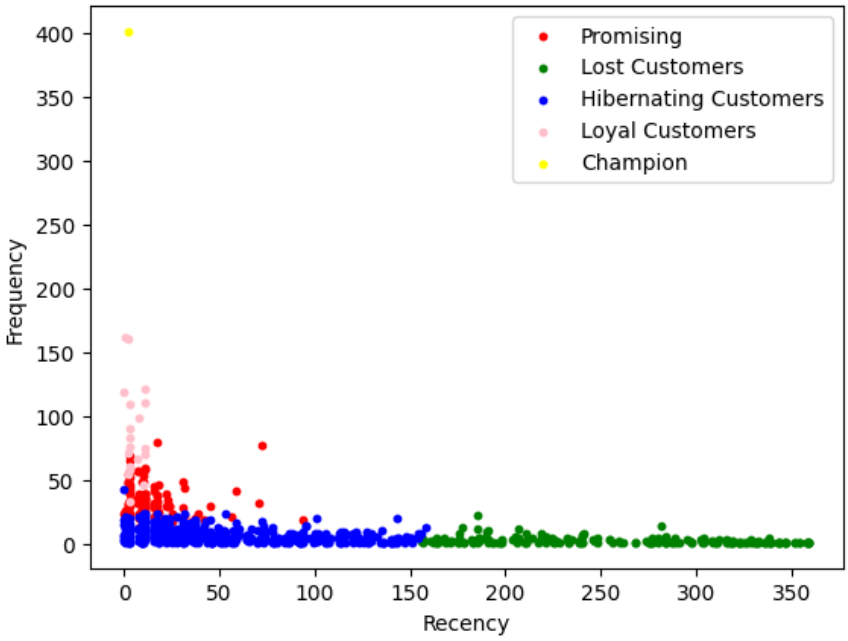

Figure 3.
Clustering results in Recency and Frequency dimensions

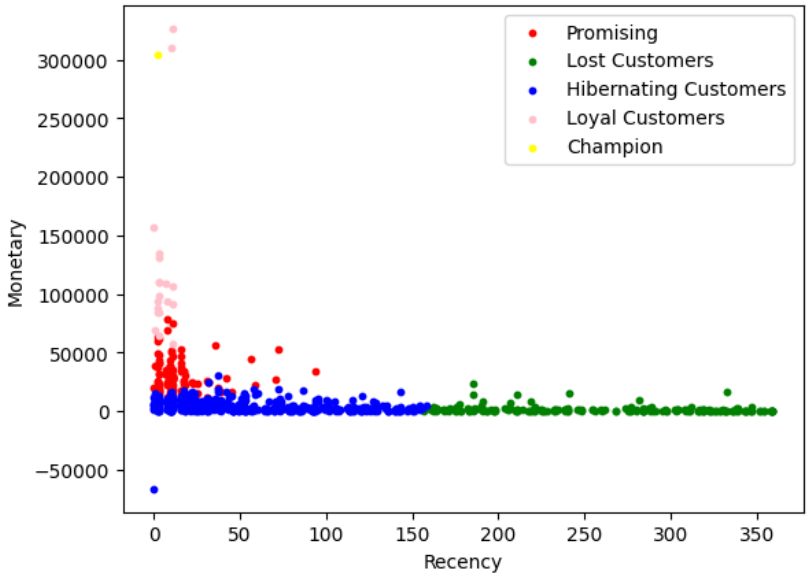

Figure 4.
Clustering results in Recency and Monetary values dimensions

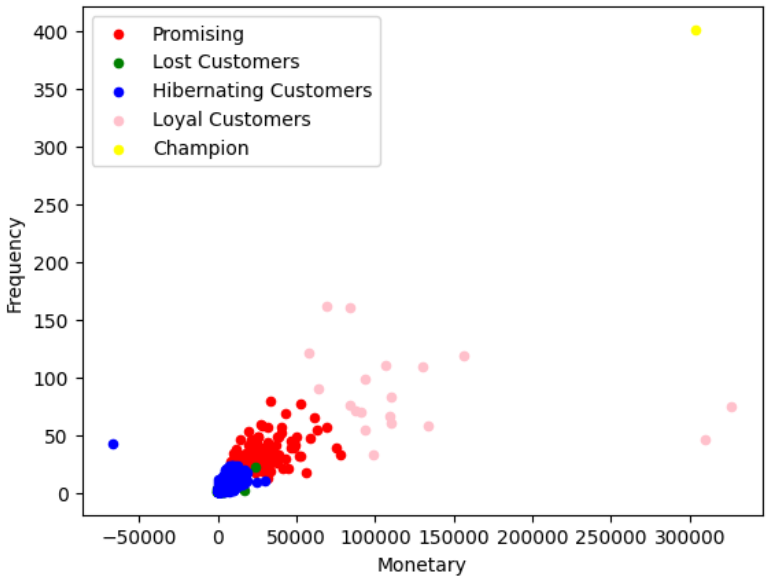

Figure 5.
Clustering results in Monetary Values and Frequency dimensions