Journal of
Business and Strategic Management
(JBSM)

Segmenting Customers for Marketing Success: A Study Using Retail Data

Journal of Business and Strategic Management
ISSN 2520-0402 (Online)
Vol. 5, Issue No. 1, pp. 58 - 67, 2020
www.carijournals

Segmenting Customers for Marketing Success: A Study Using Retail Data

Saurabh Kumar

Data Scientist, Facebook Inc., Menlo Park, CA, USA

Accepted: 5th Dec 2020 Received in Revised Form: 16th Dec 2020 Published: 26th Dec 2020

Abstract

Purpose: Customer segmentation is a critical aspect of marketing strategy, allowing businesses to tailor their offerings to distinct groups of consumers based on behavioral, demographic, and transactional attributes. This paper explores data-driven customer segmentation techniques.

Methodology: This paper uses a publicly available dataset from an online retail store. The dataset contains over 540,000 transactions, providing a rich source of information to segment customers based on their purchasing behavior. Recency, Frequency, and Monetary (RFM) analysis and clustering techniques have been applied, such as K-means to categorize customers into distinct segments. These segments are then analyzed to uncover key insights about customer behavior, including product preferences, purchasing frequency, and spending patterns.

Findings: The results demonstrate the power of data-driven segmentation in improving targeted marketing efforts, boosting customer retention, and optimizing resource allocation.

Unique Contribution to Theory, Practice and Policy: This study provides a framework for retailers and marketers to enhance their customer segmentation strategies, ultimately improving business outcomes through personalized marketing campaigns.

Keywords: *Customer Segmentation, Retail Data, RFM analysis, Clustering, Marketing Strategy*

1. Introduction

Customer segmentation is a vital component of modern marketing strategies, enabling businesses to tailor their efforts to specific consumer groups, thereby maximizing effectiveness and efficiency. In an increasingly data-driven world, companies collect vast amounts of transaction data that can be leveraged to uncover actionable insights. However, many organizations struggle to efficiently analyze this data to create targeted marketing campaigns, improve customer retention, and optimize resource allocation. This research addresses these challenges by applying advanced segmentation techniques to a publicly available retail dataset, demonstrating the potential for improved marketing strategies.

The primary objective of this study is to utilize data mining techniques, such as Recency, Frequency, and Monetary (RFM) analysis and clustering algorithms, to categorize customers into distinct groups. RFM analysis is particularly effective in identifying high-value customers by examining their purchasing behavior in terms of how recently they purchased (Recency), how often they purchase (Frequency), and how much they spend (Monetary) [1]. Clustering algorithms such as K-means provide an additional layer of granularity by grouping customers with similar purchasing patterns [5]. By leveraging these methods, marketers can better understand customer behavior and create more personalized marketing campaigns, leading to higher conversion rates and increased customer loyalty.

This research aims to solve several key problems faced by marketers today: identifying the most valuable customer segments, reducing customer churn, and increasing return on investment (ROI) for marketing initiatives. Moreover, it explores how businesses can use open-source retail data to replicate similar segmentation processes in their own operations [2]. The findings from this study will provide a framework for businesses to enhance their customer segmentation processes and ultimately drive more effective, data-driven marketing strategies. By integrating machine learning techniques, the study contributes to the growing field of marketing analytics [6][15].

2. Literature Review

Customer segmentation has long been a key strategy in marketing, helping businesses understand and group their customers based on common characteristics or behaviors. Early methods of segmentation relied on simple demographic factors like age, income, or location. However, as data collection has become more sophisticated, marketers have shifted towards more detailed and behavior-based segmentation techniques [1].

One popular method for behavioral segmentation is RFM analysis (Recency, Frequency, Monetary), which segments customers based on their purchasing history. RFM helps identify high-value customers by analyzing how recently they made a purchase, how frequently they buy, and how much they spend [2]. Studies have shown that RFM analysis is an effective way for businesses to focus their marketing efforts on the most valuable customer groups [1].

In addition to RFM, cluster analysis is widely used in customer segmentation. Clustering algorithms, such as K-means, group customers based on their behavior and similarities. For example, Jain [5] reviewed the effectiveness of clustering algorithms in organizing data into meaningful customer segments. K-means, in particular, is a widely used algorithm due to its simplicity and efficiency. This technique allows businesses to identify customer groups with similar purchasing habits, which can then be targeted with personalized marketing strategies [5][6].

Other approaches, such as the use of decision trees (like CHAID) and logistic regression, are also common for segmenting customers. McCarty and Hastak [3] compared different segmentation methods, showing how RFM, CHAID, and logistic regression can all be useful depending on the dataset and business objectives.

With the rise of machine learning, marketers now have more powerful tools for segmentation. Algorithms like genetic K-means and other machine learning techniques have improved the accuracy and efficiency of customer segmentation [12]. Machine learning-based models can handle large datasets and complex relationships between variables, allowing businesses to gain deeper insights into customer behavior [14].

Overall, the literature shows that combining traditional methods like RFM with advanced techniques like clustering and machine learning provides businesses with a powerful toolkit for understanding and targeting their customers more effectively [1][9]. This research builds on these methods, applying them to the Online Retail dataset to demonstrate their practical use in real-world marketing.

3. Methodology

This research uses data-driven techniques to perform customer segmentation on the Online Retail dataset from the UCI Machine Learning Repository [2]. The methodology is organized into the following steps: data preprocessing, Recency-Frequency-Monetary (RFM) analysis, and clustering using the K-means algorithm. Formulas are included to detail key calculations used throughout the analysis**.**

3.1 Data Preprocessing

Data cleaning is a crucial first step to ensure accuracy in analysis. The dataset is inspected for missing values, particularly in the CustomerID field, and transactions with missing IDs are removed. Additionally, transactions with negative quantities, which indicate product returns, are filtered out. The InvoiceDate field is converted to a datetime format, and monetary values are calculated by multiplying the unit price by the quantity purchased. To prepare the data, several preprocessing steps were applied:

$$\text{Monetary Value} = \text{Quantity} \times \text{Unit Price}$$

This step ensures the dataset is ready for segmentation based on customer purchase behavior [3]

3.2 RFM Analysis

Three key metrics for each customer has been calculated: Recency, Frequency, and Monetary Value. These metrics provide insight into customer purchasing behavior and are the foundation of segmentation.

Recency (R): The number of days since a customer's last purchase. It is calculated as the difference between the most recent transaction date and a reference date (e.g., the last date in the dataset).

$$Recency = Reference\ Date - Last\ Purchase\ Date$$

Frequency (F): The total number of transactions a customer has made.

$$Frequency = \sum(Transactions\ by\ Customer)$$

Monetary Value (M): The total amount spent by the customer across all transactions

$$Monetary\ Value = \sum(\ Purchase\ Amount)$$

Once these metrics are calculated, customers are ranked based on their Recency, Frequency, and Monetary scores, with higher ranks indicating more valuable customers [1].

3.3 K-Means Clustering:

After calculating the RFM metrics, **K-means clustering** is applied to group customers into segments. K-means is an unsupervised machine learning algorithm that minimizes the variance within each cluster. The optimal number of clusters is determined using the **elbow method**, which identifies the point where adding more clusters results in diminishing returns in terms of variance reduction [5].

Clustering Objective Function (to minimize the variance within clusters):

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

Where:

- k is the number of clusters.
- C_i represents each cluster.
- μ_i is the centroid of cluster C_i.
- x represents each data point (customer) in the cluster.

This formula helps determine how well the clustering has partitioned the data. Customers are grouped based on their RFM scores, with the goal of minimizing intra-cluster distance while maximizing inter-cluster separation.

3.4 Data Normalization

In cases where the values of recency, frequency, and monetary variables are on different scales, it's essential to normalize or standardize the data before applying K-means clustering. This ensures that no variable dominates the clustering process due to its magnitude.

$$X_{norm} = \frac{X - min(X)}{max(X) - min(X)}$$

Where X represents the values of Recency, Frequency, or Monetary, and X_{norm} is the normalized value.

This step ensures that all variables are treated equally in the clustering algorithm.

3.5 Cluster Evaluation

The clusters generated by the K-means algorithm are then evaluated to ensure they represent distinct customer segments. Each cluster is analyzed based on its average recency, frequency, and monetary values. This evaluation ensures that the segmentation is meaningful and can be used to drive actionable marketing strategies [7].

By combining RFM analysis with K-means clustering, this methodology provides a robust approach for segmenting customers based on their behavior, allowing businesses to enhance their marketing efforts, improve customer retention, and allocate resources more effectively.

3.6 Model Validation

To evaluate the quality of the K-means clustering model, it's important to measure the silhouette score or inertia to assess how well-separated the clusters are.The silhouette score measures how close each point in one cluster is to the points in neighboring clusters.

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

Where:

- $a(i)$ is the average distance between i and all other points in the same cluster.
- $b(i)$ is the minimum average distance between i and points in another cluster.

Higher silhouette scores indicate better-defined clusters.

4. Results

4.1 RFM Analysis

The RFM analysis divided customers into four segments: **Best Customers**, **Loyal Customers**, **Regular Customers**, and **At Risk Customers**, based on Recency, Frequency, and Monetary scores. The segmentation logic focused on Recency for engagement patterns, combined with Frequency and Monetary values to assess customer behavior and value.

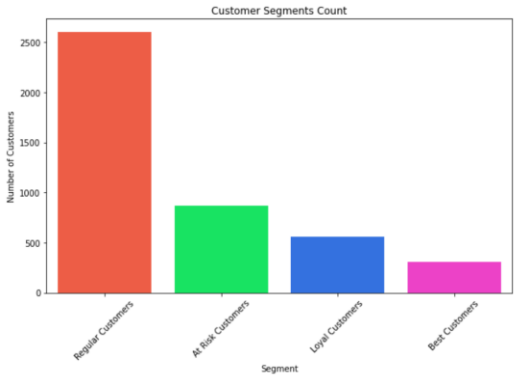

Fig.1 Customer Segment Distribution

Best Customers (306 customers) are the most valuable, with high Recency, Frequency, and Monetary scores. These customers spend the most (average $11,687) and purchase most frequently (447 transactions), making them prime candidates for VIP programs and exclusive offers.

Table 1. Segment Summary RFM Analysis

Segment	R(days)	F(#)	M($)	Customer Count
At Risk Customers	269	26	646	865
Best Customers	5	447	11687	306
Loyal Customers	7	82	1936	562
Regular Customers	63	74	1416	2605

Loyal Customers (562 customers) have made recent purchases (average Recency of 6.6 days) and buy frequently (82 transactions), though they spend less than the Best Customers ($1,936). They are highly engaged and should be targeted with retention and upsell strategies to increase their value.

Regular Customers (2,605 customers) have moderate engagement, with a Recency of 62.85 days and Frequency of 74 purchases. They spend an average of $1,416 and can be nurtured with promotions to boost their frequency and spending.

Lastly, **At Risk Customers** (865 customers) are those who haven't purchased in a long time (Recency of 268.6 days), with low Frequency and Monetary values. This segment is at risk of churn and should be re-engaged with win-back campaigns.

4.2. K-Means

After removing outliers using the Interquartile Range (IQR) method,K-Means clustering algorithm segmented the customer base into four distinct groups based on Recency, Frequency, and Monetary values. Using the Elbow Method, four clusters were identified as the optimal solution for segmenting the customer data. Each cluster represents a unique group of customers with specific characteristics. Outliers were removed

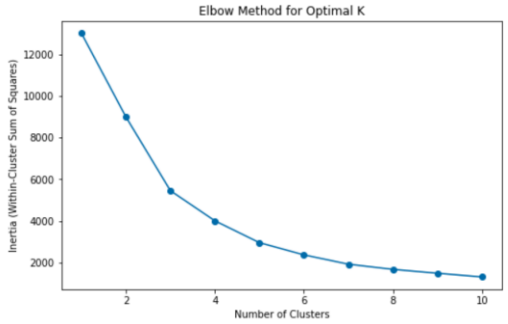

Fig.2 Elbow Method

Cluster 0, referred to as **Active Buyers**, is the largest segment with 4,072 customers. These customers exhibit moderate engagement, making 63.61 purchases on average and spending around $1,042, although their last purchase occurred approximately 97 days ago. **Cluster 1**, named **At-Risk Customers**, consists of 18 customers who, despite their high purchasing behavior (1,912 transactions on average) and significant spending ($72,103), haven't made purchases as frequently in recent times. Retention strategies focused on this small but highly valuable group are critical, as they are at risk of churn.

Table 2. Segment Summary K-Means

ClusterName	R	F	M	CustomerID
Active Buyers	41	105	2092.27617	3247
At-Risk Customers	247	28	637.285163	1081
High-Value Clients	8	827	190863.4617	6
VIP Customers	2	5807	70925.2875	4

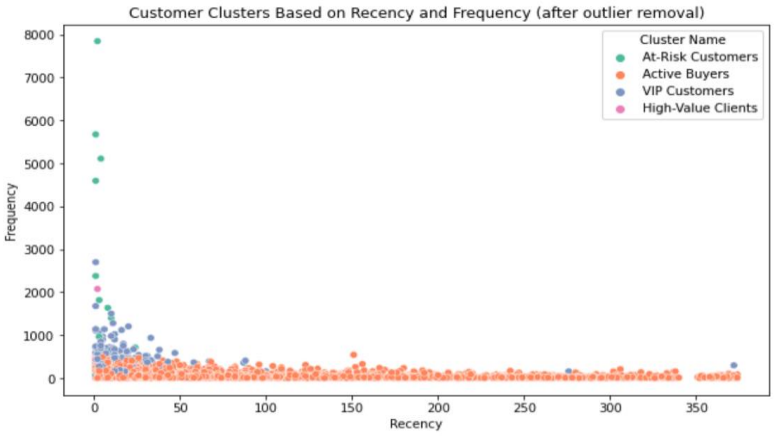

Fig 3. Frequency and Recency Scatter Plot

(**Cluster 2**, the smallest group with 4 customers, represents **High-Value Clients**. These individuals make frequent purchases (711.75 transactions) and have an exceptionally high average spend of $225,721. Maintaining their loyalty with personalized services is essential, as they provide significant revenue. **Cluster 3**, comprising 244 **VIP Customers**, is another valuable segment with frequent purchases (416 transactions) and an average spend of $10,100. These customers are highly engaged, with an average Recency of 19.68 days, and should be nurtured through loyalty programs and exclusive offers to maintain their high level of interaction. This segmentation provides the business with clear insights into how to best address and engage each customer group effectively.

5. Conclusion

The application of K-means clustering on the RFM data, after removing outliers, provided a powerful segmentation of the customer base into four distinct clusters, each representing unique patterns of customer behavior. The segmentation allows for a deeper understanding of customer engagement and spending habits, which is essential for creating personalized and effective marketing strategies. By focusing on key metrics such as Recency, Frequency, and Monetary values, this approach offers actionable insights into how to target different customer groups to maximize engagement and revenue.

The largest cluster, **Active Buyers**, represents a substantial portion of the customer base that shows consistent engagement but with room for improvement in frequency and spending. Targeting these customers with loyalty programs and special promotions could drive increased activity. In contrast, the **At-Risk Customers** and **High-Value Clients** require more specialized retention strategies, as their high transactional value and frequent purchases make them critical to the business's bottom line. Tailored retention campaigns for these high-value segments can prevent churn and enhance customer lifetime value.

The use of K-means clustering in marketing analytics allows businesses to move beyond a one-size-fits-all approach, enabling more refined targeting that can significantly improve the effectiveness of marketing campaigns. By understanding the specific needs and behaviors of each segment, marketers can deliver more relevant and timely messaging, leading to better customer satisfaction and increased loyalty. Ultimately, this method supports more efficient allocation of marketing resources, focusing efforts where they will have the greatest impact on customer engagement and business outcomes.

References:

[1]Linoff, G. S. (2011). Data mining techniques: for marketing, sales, and customer relationship management. Wiley. [PublisherLink][GoogleScholar]

[2] UCI Machine Learning Repository: Online Retail Data Set. (2015). Retrieved from https://archive.ics.uci.edu/dataset/352/online+retail.

[3] McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of business research, 60(6), 656-662.[CrossRef][PublisherLink][GoogleScholar]

[4] Saunders, J. A. (1980). Cluster analysis for market segmentation. European Journal of marketing, 14(7), 422-435.[CrossRef][PublisherLink][GoogleScholar]

[5] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.[CrossRef][PublisherLink][GoogleScholar]

[6] Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2015). Handbook of cluster analysis. CRC press.[PublisherLink][GoogleScholar]

[7] Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media.[PublisherLink][GoogleScholar]

[8] Dolnicar, S. (2003). Using cluster analysis for market segmentation-typical misconceptions, established methodological weaknesses and some recommendations for improvement.[PublisherLink][GoogleScholar]

[9] Tsiptsis, K. K., & Chorianopoulos, A. (2011). Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons.[PublisherLink][GoogleScholar]

[10] Berson, A., & Thearling, K. (1999). Building data mining applications for CRM. McGraw-Hill, Inc..[PublisherLink][GoogleScholar]

[11]Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electronic Commerce Research and Applications, 8(5), 241-251. [CrossRef][PublisherLink][GoogleScholar]

[12] Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439.[CrossRef][PublisherLink][GoogleScholar]

[13] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.[CrossRef][PublisherLink][GoogleScholar]

[14] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.[CrossRef][PublisherLink][GoogleScholar]

[15] Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of marketing, 80(6), 97-121.[CrossRef][PublisherLink][GoogleScholar]

[16] Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard business review, 96(1), 108-116.[PublisherLink][GoogleScholar]

[17] Mahajan, V., & Venkatesh, R. (2000). Marketing modeling for e-business. International Journal of Research in Marketing, 17(2-3), 215-225.[CrossRef][PublisherLink][GoogleScholar]