Journal of Operational and Strategic Analytics
https://www.acadlore.com/journals/JOSA

Optimizing Decision-Making Through Customer-Centric Market Basket Analysis

MD Jiabul Hoque[1,2*], Md. Saiful Islam[2], Syed Abrar Mohtasim[1]

[1] Department of Computer and Communication Engineering, International Islamic University Chittagong, 4318 Chattogram, Bangladesh
[2] Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering & Technology, 4349 Chattogram, Bangladesh

* Correspondence: MD Jiabul Hoque (jiabul.hoque@iiuc.ac.bd)

Received: 01-26-2024 **Revised:** 03-22-2024 **Accepted:** 04-12-2024

Citation: M. D. J. Hoque, M. S. Islam, and S. A. Mohtasim, "Optimizing decision-making through customer-centric Market Basket Analysis," *J. Oper. Strateg Anal.*, vol. 2, no. 2, pp. 72–83, 2024. https://doi.org/10.56578/josa020201.

Abstract: In the realm of understanding consumer purchasing behaviors and refining decision-making across diverse sectors, Market Basket Analysis (MBA) emerges as a pivotal technique. Traditional algorithms, such as Apriori and Frequent Pattern Growth (FP-Growth), face challenges with computational efficiency, particularly under low minimal support settings, which precipitates an excess of weak association rules. This study introduces an innovative approach, termed Customer-Centric (CC)-MBA, which enhances the identification of robust association rules through the integration of consumer segmentation. By employing Recency, Frequency, and Monetary (RFM) analysis coupled with K-means clustering, customers are categorized based on their purchasing patterns, focusing on segments of substantial value. This targeted approach yields association rules that are not only more relevant but also more actionable compared to those derived from conventional MBA methodologies. The superiority of CC-MBA is demonstrated through its ability to discern more significant association rules, as evidenced by enhanced metrics of support and confidence. Additionally, the effectiveness of CC-MBA is further evaluated using lift and conviction metrics, which respectively measure the observed co-occurrence ratio to that expected by chance and the strength of association rules beyond random occurrences. The application of CC-MBA not only streamlines the analytical process by reducing computational demands but also provides more nuanced insights by prioritizing high-value customer segments. The practical implications of these findings are manifold; businesses can leverage this refined understanding to improve product positioning, devise targeted promotions, and tailor marketing strategies, thereby augmenting consumer satisfaction and facilitating revenue growth.

Keywords: Customer segmentation; Market Basket Analysis (MBA); Recency, Frequency, and Monetary (RFM); K-means clustering; Apriori; Frequent Pattern Growth (FP-Growth); Association rules; Customer behavior analysis

1 Introduction

The ever-changing landscape of modern commerce demands a deep understanding of customer behavior and preferences. MBA, a powerful data mining technique, empowers businesses to extract valuable insights from customer purchase patterns [1]. By identifying associations between products frequently purchased together, companies can enhance decision-making processes and drive sales growth [2]. This research introduces a novel approach to MBA that integrates RFM analysis with K-means clustering to segment customers for more effective analysis. The core innovation of this study lies in prioritizing customer segmentation over traditional product reduction techniques through the incorporation of RFM and K-means-based customer segmentation. This allows us to uncover hidden patterns within specific customer groups, leading to more targeted and actionable insights compared to traditional MBA approaches.

This research addresses the limitations of traditional MBA methods that rely on algorithms like Apriori and FP-Growth. These methods require setting a minimum support threshold to identify frequently purchased items. However, a low threshold can lead to an overwhelming number of weak association rules and high computational demands [3, 4]. This study proposes a paradigm shift by focusing on customer segmentation instead. By segmenting customers based on RFM analysis and K-means clustering (e.g., high-value customers, churn-risk customers), the

Knowledge Discovery in Databases (KDD) process can be leveraged to selectively analyze transactions within these valuable customer segments [5]. This not only reduces computational overhead but also yields stronger association rules due to the focus on relevant buying behaviors.

This research aims to investigate the effectiveness of customer segmentation using RFM analysis and K-means clustering in improving MBA. It supports more informed decision-making and optimizes sales in the process. By subjecting the customer segmentation process to a rigorous examination, this study aims to evaluate its impact on the quality of insights produced and the efficacy of the analytical procedure, explicitly focusing on the strength of association rules generated. The findings of this inquiry provide significant knowledge for organizations seeking to utilize MBA to enhance their sales performance. The core of this research is to promote a CC methodology in MBA. This emphasizes the criticality of discerning high-impact buying trends among discrete customer segments. Empirical analysis is employed to demonstrate how this approach generates practical insights that can enhance the organization's overall performance.

The rest of this study covers critical elements. After analyzing relevant literature on MBA, consumer segmentation, and data mining, the research design and framework are outlined. Section 4 covers experiment design, analysis, and a detailed discussion of their implications. Finally, Section 5 summarizes the main findings of this study, while Section 6 discusses further research.

2 Literature Review

The growing importance placed on CC initiatives highlights the value of utilizing customer data to cultivate relationships and stimulate corporate expansion [6]. MBA and customer segmentation are prominent approaches to extracting insights from customer purchase behaviours [7]. However, the current body of literature reveals a lack of agreement regarding the most effective approaches and limitations related to the ability to apply findings to a broader population [8–15].

The effectiveness of client segmentation in CRM strategies and sales forecasting is emphasized in studies conducted by Shim et al. [8] and Kasem et al. [9], respectively. Shim et al. [8] examined a specific retail environment (shopping centres), limiting their findings' applicability. In contrast, Kasem et al. [9] emphasized the need for additional investigation into the scalability of their client profiling system. Dwivedi and Singh [10] proposed a complete segmentation strategy incorporating RFM, LTV, K-means, and neural networks. However, the authors expressed concerns about integrating several methods in dynamic situations.

Multiple studies propose novel approaches to overcome existing constraints. Ahmed et al. [11] introduced the product reduction technique for MBA, while concerns were raised over its scalability for bigger datasets. The model suggested by Xiahou and Harada [12] for predicting online sales is limited in its ability to react to changing customer behaviour due to its lack of integration with real-time data streams. These inquiries highlight the need for a methodology that harmonizes effectiveness with the flexibility to adapt to evolving customer preferences.

Although Tang et al. [13] demonstrated the usefulness of the Apriori algorithm, its effectiveness in different market sizes and suitability in online environments still need to be determined. Hamdad and Benatchba [14] demonstrated association rule mining for market trend detection. However, they highlight concerns about the applicability and scalability of this method when applied to more extensive datasets. Martinez et al. [15] effectively employed the FP-Growth algorithm on a dataset about home products while acknowledging the constraints associated with its implementation inside small and medium-sized enterprises (SMEs). In general, the literature indicates a need for updated MBA that are more scalable and adaptable.

The reviewed literature highlights the potential effectiveness of sophisticated data mining approaches, such as FP-Growth, while also recognizing the computing challenges they may involve. Besides, it underscores the significant importance of data quality in assessing the efficacy of client segmentation tactics. The CC-MBA methodology proposed in this study aims to address these difficulties by:

- Integrating RFM analysis with K-means clustering as a strategy, thereby enhancing the accuracy of consumer segmentation in comparison to traditional approaches.
- Employing the KDD methodology to concentrate on transactions within essential customer segments, thereby decreasing the computational burden and generating more resilient association rules.

CC-MBA aims to enhance decision-making processes and corporate performance by prioritizing customer segmentation, providing a more adaptive and scalable method for MBA.

3 Methodology

3.1 Research Framework

The preliminary stage of this research involves thorough pre-processing of customer information, including operations such as removing duplicate entries and correcting any missing or inaccurate values. Afterward, RFM values are calculated for each customer using the data that is currently accessible. Subsequently, the RFM ratings are employed with the K-means clustering methodology to partition customers proficiently. The main goal is to

identify high-value customers who have a significant impact on business outcomes. After identifying these high-value customers, MBA procedures are utilized. As part of the segmentation process, association rules are derived from the transactional data of VIP customers. Figure 1 presents a visual representation of the proposed research framework.

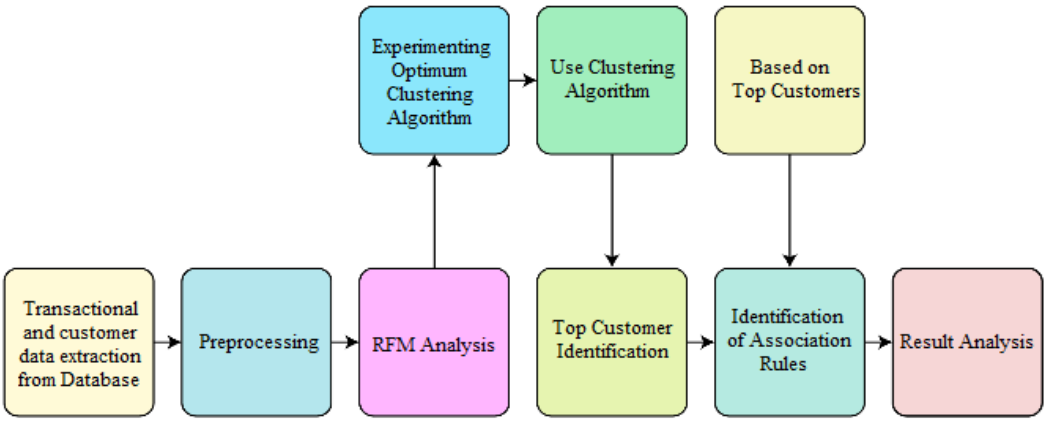


Figure 1. Proposed research framework

3.2 Customer Segmentation

This study emphasizes customer segmentation, specifically targeting individual customers and employing data mining techniques to categorize them into separate segments. Identifying and including relevant characteristic variables is crucial to providing a solid basis for predictive modelling [16]. In order to achieve this objective, descriptive statistics are calculated to capture essential characteristics of the time series data about each customer. These statistics include sums, means, medians, and standard deviations. Moreover, the establishment of a target or dependent variable is considered essential for the aim of predictive modelling. Once the target variable has been accurately defined, its values are calculated for all customers and smoothly incorporated into the pre-existing data tables.

The optimal cluster is established by experimenting with different clustering methods, specifically K-means, hierarchical clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The assessment of the performance of each algorithm is carried out by employing the silhouette score, with higher scores denoting superior performance. The grocery dataset yielded silhouette scores of 0.348 for K-means, 0.225 for DBSCAN, and 0.292 for hierarchical clustering. Similarly, the silhouette scores in the retail dataset were recorded as K-means (0.36), hierarchical clustering (0.27), and DBSCAN (0.11). The experimental findings demonstrate that K-means outperforms other clustering algorithms in RFM segmentation, supporting its choice as the most suitable method for this study. The silhouette score is succinctly summarized in Table 1.

Table 1. Silhouette score distribution

Algorithms	Silhouette Score (Retail Dataset)	Silhouette Score (Grocery Dataset)
K-means	0.36	0.348
Hierarchical	0.27	0.292
DBSCAN	0.11	0.225

The most suitable model was carefully chosen in this study to predict the dependent variable. This is crucial in identifying the most relevant target customers in the following stages. The K-means algorithm plays a crucial role in cluster analysis and customer segmentation. The K-means algorithm is a method designed to divide a collection of objects into K clusters. The number of subgroups has been established, and the centroids are assigned randomly to observations in the dataset [17]. In order to minimize the variance within clusters, the algorithm employs an iterative process that involves two steps. Firstly, each object is assigned to the nearest centroid based on a similarity measure. Secondly, a new centroid is computed for each cluster by calculating the mean vector of all objects in the group. The iteration process continues until it achieves convergence [18]. In the usual implementation of K-means, the Euclidean distance measure is commonly used to determine the closest objects to each cluster. This is done by calculating the mean squared error of each object's feature vector with respect to the K centroid. The closest outcome is then selected [19].

The steps of the K-means algorithm are as follows:

Algorithm 1: K-means
Step 1: K clusters are defined.
Step 2: Centroids are randomly selected.
Step 3: Points are assigned to the nearest centroids.
Step 4: The centroids are updated.
Step 5: Steps 3-4 are repeated until convergence.
Step 6: End.

The sensitivity of K-means clustering to outliers in the data is well-established. To tackle this issue, Interquartile Range (IQR) statistical methods are utilized to identify outliers and eliminate them. Figure 2 graphically explains the application of K-means to RFM analysis.

The Results section may be divided into subsections. It should describe the results concisely and precisely, provide their interpretation, and draw possible conclusions from the results.

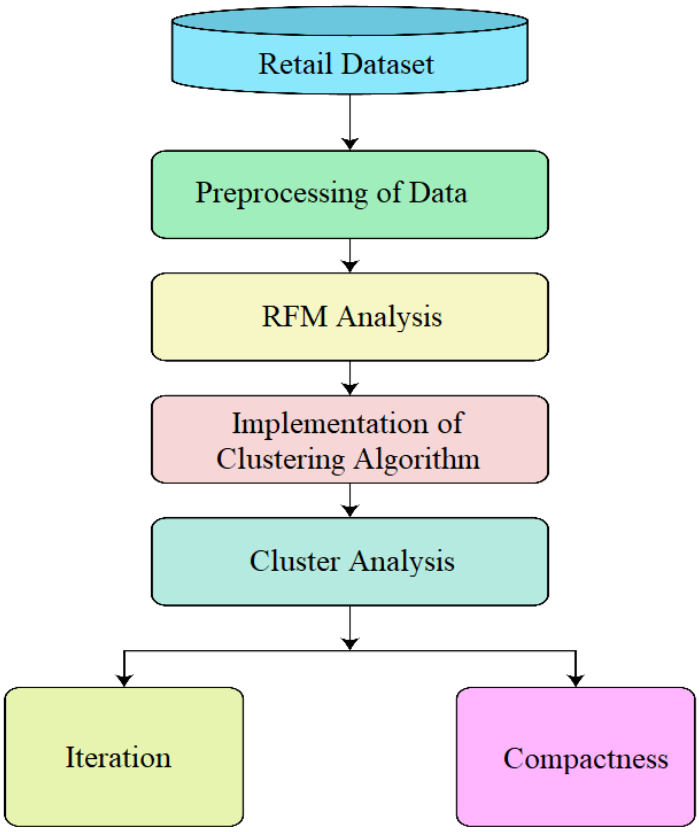


Figure 2. Process of applying K-means to RFM analysis

3.3 MBA

MBA is an extensively utilized data mining technique that is critical in marketing contexts. This study places significant emphasis on the role of MBA in simplifying the identification of helpful product correlations within transactional data about high-value customer segments [20]. The main objective of the MBA is to develop a comprehensive understanding of customer interactions through identifying patterns and associations, which improves the decision-making process [21]. Through analyzing purchase behaviours and product combinations, the MBA provides significant insights into consumer preferences and buying habits. This knowledge is instrumental in formulating data-driven strategies to cultivate customer connections and stimulate business growth.

MBA methodically examines customer buying patterns to reveal correlations among various items in their shopping carts. The primary aim is to ascertain the things commonly bought together by customers, providing substantial benefits to business owners by enhancing marketing methods [22]. It is worth mentioning that association criteria are commonly employed in MBA to identify products that are more likely to be purchased together [23].

3.4 Association Rules

Market association rules, a technique within the field of machine learning, are employed to identify the presence of co-occurring items within transaction datasets. This method uses metrics, specifically support and confidence, to assess the strength of these associations [24]. The acquisition of an item's support value is computed using the formula provided in Eq. (1). More precisely, association rules within machine learning are designed to detect co-occurring items in transactional datasets, utilizing metrics like support and confidence to evaluate the significance of these associations.

1) Support

Support quantifies the frequency with which the items mentioned in an association rule co-occur in the dataset by calculating the count of transactions containing those items ($|x|$) relative to the total number of transactions ($|D|$) [25].

$$\mathrm{Sop(x)} = \frac{|x|}{|D|} \tag{1}$$

The optimal choice of the minimal support threshold holds significant importance in the context of the Apriori algorithm. An excessively high threshold may overlook potentially valuable yet less frequent associations, whereas a shallow threshold can result in excessively feeble association rules. A minimal support criterion of 0.001% has been selected to balance capturing enough everyday objects to generate pertinent rules and preserving computational efficiency.

ii) Confidence

Confidence measures the strength of the relationship between items in a rule, indicating the percentage of transactions that include both the antecedent and subsequent terms relative to transactions that involve only the antecedent [25]. The confidence value of an item can be calculated using the formula provided in Eq. (2).

$$\mathrm{Conf(x \Rightarrow y)} = \frac{\mathrm{Sop}(x \cup y)}{\mathrm{Sop}(x)} = \frac{|x \cup y|}{|x|} \tag{2}$$

iii) Lift

Lift assesses the degree of dependence between the terms of a rule, measuring the level of certainty associated with an association rule compared to the expected confidence under the assumption of independence [25]. Mathematically, lift is defined by Eq. (3):

$$\mathrm{Lift(x \Rightarrow y)} = \frac{\mathrm{Conf}(x \Rightarrow y)}{\mathrm{Sop}(y)} \tag{3}$$

An association rule can be evaluated as follows:

If Lift $(x \rightarrow y) = 1$, then it implies that the occurrence of item " y " is independent of the event of item " x " and vice versa.

If Lift $(x \rightarrow y) > 1$, then it indicates that the occurrence of item " y " influences the probability that item " x " occurs.

If Lift $(x \rightarrow y) < 1$, it suggests that the occurrence of item " y " influences the probability that item " x " does not occur.

3.5 Apriori Algorithm

The Apriori algorithm, a fundamental component of frequent itemset mining, utilizes a level-wise search technique. The process commences by finding subsets of items that exhibit a high frequency and exceed a predetermined minimum support level within the dataset. Afterward, it expands on these standard units of one item to identify common combinations of two items (2-item sets). The process of iteration persists until all frequent k-itemsets are identified. The Apriori algorithm is notable for utilizing the anti-monotonicity property, which guarantees that any subset of a frequent itemset must maintain its frequency. Apriori employs a breadth-first search to effectively locate these common itemsets by prioritizing candidate item combinations with high frequency [26].

The Apriori algorithm utilizes two main parameters: minimum support and minimum confidence. As previously mentioned, the minimum support threshold has been set at 0.001%. Minimum confidence measures the likelihood of the consequent (item B) appearing in a transaction, given that the antecedent (item A) is already present. The optimal minimum confidence threshold is determined through a lift analysis. Lift compares the observed co-occurrence of items to the expected co-occurrence based on their individual frequencies. Rules with a lift value greater than 1 indicate a positive association between items, while a lift value less than 1 suggests a negative or random association. By analyzing the lift of association rules at different confidence levels, a threshold, which balances the number of generated rules with their actionable insights, can be identified.

4 Experimental Results and Discussion

4.1 Dataset Description

This research utilizes datasets from the reputable Kaggle platform. The grocery dataset encompasses 38,765 transactions executed by 14,963 unique customers, while the retail dataset contains 541,000 transactions from an undisclosed number of customers (due to privacy concerns). A concise summary of the content of each dataset is provided in Table 2.

Table 2. Dataset description

Dataset Name	No. of Instances	No. of Attributes
Grocery dataset	38765	3
Retail dataset	541000	8

4.2 Pre-Processing and RFM Analysis

4.2.1 On the grocery dataset

All data pre-processing involves two primary steps:

- Elimination of duplicate rows: This step ensures data cleanliness and avoids skewing the analysis.
- Standardization of data types: This step ensures consistency throughout the data for accurate processing.

Following pre-processing, this study delves into RFM analysis, a cornerstone of exploring customer behavior. Within the RFM framework, the following values can be computed:

- Recency: Quantified by the number of days since a customer's most recent purchase.
- Frequency: Determined by the number of customer transactions.
- Monetary value: Evaluated by summing the total quantity of items purchased across all customer orders.

4.2.2 On the retail dataset

In the retail dataset pre-processing, this study focuses on:

- Date and time variables: The variables were properly handled by converting them to the desired format and extracting additional temporal features (e.g., day of week).
- Data cleaning: Discrepancies or inconsistencies were addressed, including non-purchase transactions with negative unit prices and invoices containing credits.

To derive RFM values, transactional data at the customer level was aggregated by computing recency (days since the last purchase), frequency (transaction count), and monetary value (total purchase amount). Subsequently, the percentile ranks for each RFM metric were determined and mapped to a scale of 10 to facilitate comparative analysis across customer segments.

4.3 K-means Clustering

4.3.1 On the grocery dataset

K-means clustering was employed to categorize customers based on their RFM metrics. The elbow method was implemented to pinpoint the optimal number of clusters for segmentation. In this instance, five clusters were established, allowing a meaningful division of customers into distinct groups based on their RFM characteristics. Subgraph (a) of Figure 3 displays the curve of the elbow method, visualizing this process.

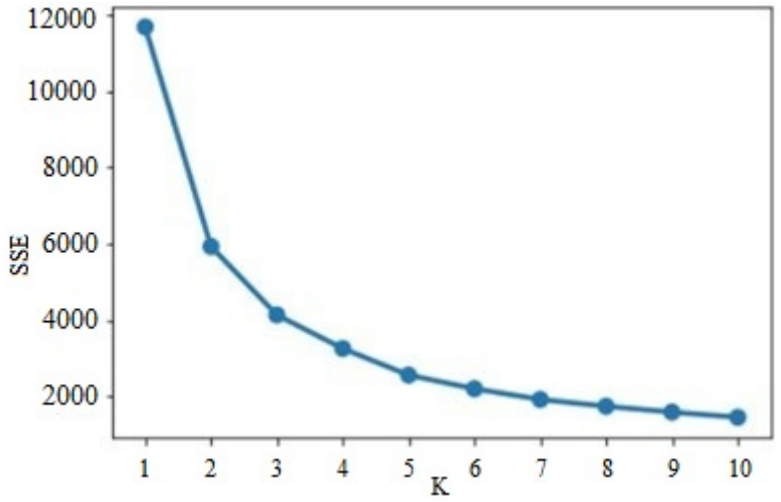
Clusters 2 and 4 were identified as key customer segments for MBA:

- Cluster 2: Characterized by high financial value and frequency with low recurrence, representing valuable customers.
- Cluster 4: Exhibited lower monetary value but moderate frequency and recurrence, suggesting a segment of loyal customers.

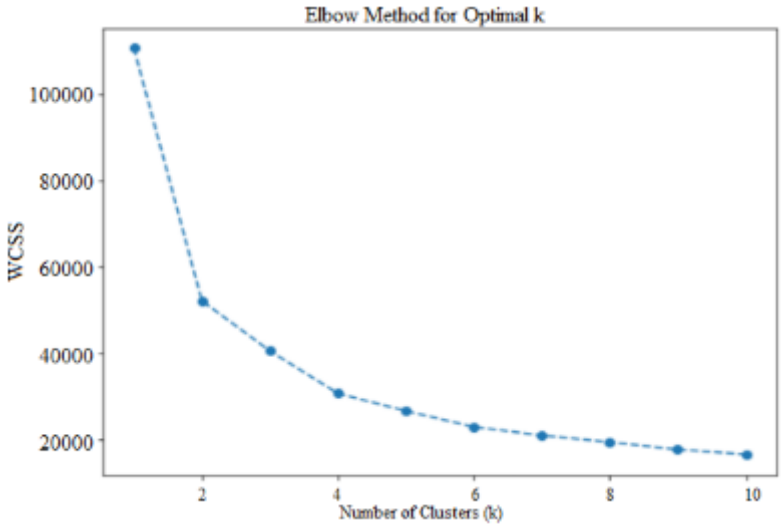
Table 3 provides a detailed view of these clusters by presenting the mean RFM values and the count of customers within each set.

Table 3. RFM values and customer count in each cluster of the grocery dataset

Clusters	Recency	Frequency	Monetary Value	Customer Counts
0	325.72	3.70	9.30	660
1	481.98	1.60	3.72	497
2	94.82	6.47	17.30	888
3	132.90	2.10	5.06	730
4	86.31	3.96	10.43	1123



(a)



(b)

Figure 3. Elbow method in K-means: (a) On the grocery dataset; (b) On the retail dataset

4.3.2 On the retail dataset

Various clustering algorithms were experimented with to identify the optimal approach for the retail dataset. K-means emerged as the front-runner, achieving the highest silhouette score, indicating a well-defined separation between clusters. The elbow method was again leveraged to determine the optimal number of clusters for segmentation, resulting in four distinct customer groups (subgraph (b) of Figure 3). Among these clusters, Cluster 1 emerged as particularly noteworthy, exhibiting high $\text{Mean}_{\text{Frequency}}$ and $\text{Mean}_{\text{Monetary}}$ values alongside relatively low $\text{Mean}_{\text{Recency}}$. This profile suggests a cohort of consistently valuable customers who make frequent purchases and contribute significantly to revenue generation. As a result, Cluster 1 was selected for further analysis in this study, particularly for MBA. This group represented a segment of customers with substantial purchasing power and a propensity for frequent transactions, making them ideal targets for targeted marketing initiatives and customer retention strategies. Table 4 presents the mean RFM values for each group, along with the count of customers within each group in the retail dataset.

A notable disparity in results was observed between the clustered and original data frames, with the clustered data frame consistently exhibiting higher support and confidence values for various rules. This discrepancy underscores the effectiveness of clustering in improving the efficacy of MBA, as evidenced by the strengthened associations evident in the clustered data frame results. Table 5 summarizes the support and confidence values for selecting unique rules in both sets, highlighting the percentage difference in support and confidence between the two groups. These findings further emphasize the value of clustering to uncover meaningful associations within customer purchasing behavior.

Table 4. RFM values and customer count in each cluster of the retail dataset

Clusters	Recency	Frequency	Monetary Value	Customer Counts
0	200	15	284	1220
1	16	225	5189	1261
2	112	70	1439	1008
3	24	29	672	850

4.4 MBA

4.4.1 On the grocery dataset

MBA was performed on Clusters 2 and 4, identified through K-means clustering. A minimum support threshold of 0.05 was applied to the clustered data frame, which comprises 10,192 transactions from the 2011 customer subset. The top 10 association rules, sorted by confidence, are presented in Table 6.

Table 5. Support and confidence values for the snippet of unique rules in the grocery data set experiment

Rule	Support (Clustered DF)	Confidence (Clustered DF)	Support (Original DF)	Confidence (Original DF)	Support (Differences)	Confidence (Differences)
(yogurt, sausage) ⇒ (whole milk)	0.2%	27.5%	0.1%	25.6%	21.08%	7.63%
(soda, yogurt)⇒ (whole milk)	0.1%	22.2%	0.1%	18.0%	28.07%	19.48%
(semi-finished bread) ⇒ (whole milk)	0.2%	20.4%	0.2%	17.6%	16.79%	15.82%
(rolls/buns, yogurt) ⇒ (whole milk)	0.2%	20.2%	0.1%	17.1%	18.34%	18.56 %

Table 6. Top 10 rules of the clustered data frame of the grocery data set

Rank	Antecedents	Consequents	Support	Confidence
1	(yogurt, sausage)	(whole milk)	0.2%	27.5 %
2	(finished products)	(whole milk)	0.1%	23.4%
3	(soda, yogurt)	(whole milk)	0.1%	22.2%
4	(dish cleaner)	(whole milk)	0.1%	20.8%
5	(semi-finished bread)	(whole milk)	0.2%	20.4%
6	(rolls/buns, yogurt)	(whole milk)	0.2%	20.2%
7	(whole milk, sausage)	(yogurt)	0.2%	19.8%
8	(rolls/buns, sausage)	(whole milk)	0.1%	19.7%
9	(soda, sausage)	(whole milk)	0.1%	19.4%
10	(detergent)	(whole milk)	0.2%	18.8%

4.4.2 On the grocery dataset

The MBA in the retail data set yielded significant insights into customer behavior and product relationships. Upon conducting MBA on the retail dataset, using a minimum threshold lift of 0.05 and sorting the values based on confidence, a notable disparity was identified in the association rules generated between the original and clustered data frames. While the top association rules in the original data frame provided valuable insights, an analysis of the clustered data frame revealed distinct patterns of product cooccurrences among specific customer segments. When analyzing the clustered data frame, five new association rules emerged that were not present in the original data frame. These rules, characterized by high confidence and lift values, indicated unique purchasing behaviors and preferences within the identified customer segment. An example of a distinct rule could be: If customers purchase *(green regency teacup and saucer)*, they are highly likely to purchase *(roses regency teacup and saucer)* with a confidence of 0.81 and a lift of 5.03. This distinct rule underscores the importance of a CC-MBA to identify purchasing patterns and tailor marketing strategies accordingly. Table 7 summarizes the support and confidence values for the top 10 association rules generated from the clustered data frames.

Table 7. Top 10 rules of the clustered data frame of the retail data set

Rank	Antecedents	Consequents	Support	Confidence
1	*(green regency teacup and saucer)*	*(roses regency teacup and saucer)*	0.2%	27.5%
2	*(gardeners kneeling pad cup of tea)*	*(gardeners kneeling pad keep calm)*	0.1%	23.4%
3	*(alarm clock bakelike green)*	*(alarm clock bakelike red)*	0.1%	22.2%
4	*(roses regency teacup and saucer)*	*(green regency teacup and saucer)*	0.1%	20.8%
5	*(jumbo bag pink polkadot)*	*(jumbo bag red retrospot)*	0.2%	20.4%
6	*(gardeners kneeling pad keep calm)*	*(gardeners kneeling pad cup of tea)*	0.2%	20.2%
7	*(alarm clock bakelike red)*	*(alarm clock bakelike green)*	0.2%	19.8%
8	*(lunch bag pink polkadot)*	*(lunch bag red retrospot)*	0.1%	19.7%
9	*(jumbo storage bag suki)*	*(jumbo bag red retrospot)*	0.1%	19.4%
10	*(lunch bag woodland)*	*(lunch bag red retrospot)*	0.2%	18.8%

4.5 Discussion

This investigation explores the domain of MBA with a concentration on CC methodologies, aiming to unravel the intricate patterns of product co-occurrences within transactional data. By prioritizing customer segmentation over traditional product reduction techniques, this study introduces an innovative methodology that not only enhances computational efficiency but also enhances the quality and relevance of the generated association rules.

In the examination of the grocery dataset, a notable increase in support and confidence values was observed when employing a CC approach to MBA. Through the segmentation of customers according to their RFM values, this study revealed previously obscured robust links between products. The top association rules derived from the clustered data frame exhibited higher support and confidence, indicating stronger and more reliable associations within specific customer segments. For example, rules like (yogurt, sausage) $\Rightarrow$ (whole milk) demonstrated markedly higher confidence and support values, emphasizing the effectiveness of CC analysis in elucidating purchasing patterns.

Similarly, the exploration of the retail data set demonstrated the potency of CC-MBA in revealing distinct purchasing patterns between different customer segments. By employing clustering techniques to segment customers based on their RFM metrics, this study discovered unique associations between products that were not apparent in the original data frame. The emergence of new association rules, characterized by high confidence and lift values, underscored the importance of customer segmentation in capturing nuanced customer behaviors and preferences. These insights are invaluable for retailers looking to customize their marketing strategies and product offerings to specific customer segments, thereby enhancing customer satisfaction and driving sales.

It is vital to recognize that the data set used in this study may not fully encompass the intricacy and diversity of real-world retail environments. Although the proposed approach yielded promising results in terms of support and confidence values, it is conceivable that with larger and more diverse datasets containing thousands of products across various categories, each customer segment may exhibit a completely distinct set of association rules. Therefore, future research seeks to authenticate the findings of this study on more extensive data sets to ensure the robustness and generalizability of the proposed approach.

5 Conclusions

This research introduces CC-MBA, a novel approach that integrates RFM analysis with K-means clustering, to enhance the efficiency and effectiveness of association rule mining. The systematic methodology employed in this study, encompassing comprehensive data pre-processing, customer segmentation using K-means clustering, and targeted MBA with the Apriori algorithm, demonstrates the potential of CC-MBA to reveal intricate patterns of customer purchasing behavior.

The findings of this study suggest that CC-MBA can lead to the identification of more significant association rules with increased support and confidence values compared to traditional approaches that don't incorporate customer segmentation. This highlights the value of CC techniques in understanding and forecasting purchasing behavior. For instance, in the retail dataset, CC-MBA uncovered distinct association rules within specific customer segments, including rules not present in the original data frame. These findings suggest that CC-MBA can be a valuable tool for retailers to gain deeper customer insights and tailor marketing strategies accordingly.

In addition to the present investigation, CC-MBA exhibits the potential for transforming decision-making methodologies in diverse sectors. Retailers can utilize CC-MBA to effectively identify consumer segments characterized by high purchase frequency or value, allowing them to implement personalized marketing. Moreover, it can reveal concealed product preferences among distinct client segments, providing valuable insights for developing product placement and bundling tactics. Furthermore, CC-MBA can be employed by subscription services to identify consumers at risk and perform specific interventions to prevent client turnover. Comprehending client buying

behaviours allows subscription services to customize suggestions and items, thus improving customer contentment.

Although CC-MBA offers significant benefits, firms may need help implementing it. These factors encompass the availability of data, which requires access to extensive customer data, and expenditures on infrastructure for data collection and management. In addition, firms may need to develop internal data analysis capabilities or engage external professionals to utilize CC-MBA properly. Furthermore, incorporating CC-MBA into current marketing and customer relationship management (CRM) systems may necessitate a high level of technical expertise and possible adjustments.

Further research is required to investigate these practical ramifications in greater depth and provide optimal strategies for applying CC-MBA. By acknowledging these constraints and investigating the pragmatic uses of CC-MBA, a more profound comprehension of its revolutionary capacity for many sectors can be cultivated.

6 Future Works

While this study has shed light on the potential of customer segmentation in MBA, it is essential to acknowledge its limitations and explore avenues for future research.

Future research endeavours should emphasize the acquisition of more comprehensive and varied retail datasets, which should span a broader spectrum of product categories, a larger client demographic, and maybe additional customer characteristics. Examining association rules across diverse datasets could lead to a more comprehensive knowledge of the usefulness of CC-MBA in real-world circumstances.

Furthermore, investigating alternate algorithms alongside the Apriori and FP-Growth algorithms employed in this work may uncover more complex relationships within the data. Algorithms designed explicitly for managing extensive and intricate datasets, such as Eclat (for data with a high number of dimensions) or Parallel FP-Growth (for distributed processing), present potential opportunities for further exploration. Furthermore, investigating rule-induction algorithms that can find both frequent item sets and classification rules has the potential to yield more profound insights into the segmentation of client behaviour.

The customer segmentation criteria substantially impact the outcomes of CC-MBA. Although this study's primary focus was on RFM measurements, future research should investigate alternative methodologies.

- Customer Lifetime Value (CLV) is a segmentation strategy that considers many factors, such as purchase frequency, value, and customer retention, to assess customer value comprehensively.
- Demographic segmentation integrates demographic data with RFM indicators to identify segments characterized by specific purchasing behaviours, such as age, region, or income.
- The analysis of purchase data segmented by marketing platforms, such as email marketing versus social media advertising, can provide valuable insights into the effectiveness of marketing campaigns.

The study of longitudinal data offers significant insights into the dynamic nature of purchasing patterns, enabling firms to understand market trends comprehensively and proactively predict future client demands. The systematic tracking of shifts in customer preferences and behaviours over an extended period achieves this. The longitudinal study of CC-MBA demonstrates its proficiency in recognizing the dynamic nature of product interactions within distinct consumer categories over an extended period.

The investigation of deep learning methodologies offers a captivating prospect for augmenting the complexity of CC-MBA. Advanced deep learning models can reveal intricate patterns and non-linear connections across products, potentially outperforming conventional techniques. Integrating deep learning with CC-MBA can reveal intricate insights into customer behaviour, enhancing the effectiveness of targeted marketing campaigns and product suggestions.

By addressing these future research directions, the CC-MBA can be further refined and strengthened, thereby ultimately enhancing its real-world applicability and value for businesses across various industries.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] K. Kafkas, Z. N. Perdahçı, and M. N. Aydın, "Discovering customer purchase patterns in product communities: An empirical study on co-purchase behavior in an online marketplace," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 7, pp. 2965–2980, 2021. https://doi.org/10.3390/jtaer16070162

[2] V. Kumar and W. Reinartz, *Customer Relationship Management*. Berlin/Heidelberg, Germany: Springer, 2018.

[3] A. H. L. Chen and S. Gunawan, "Enhancing retail transactions: A data-driven recommendation using modified RFM analysis and association rules mining," *Appl. Sci.*, vol. 13, no. 18, p. 10057, 2023. https://doi.org/10.3390/app131810057

[4] S. Tuominen, H. Reijonen, G. Nagy, A. Buratti, and T. Laukkanen, "Customer-centric strategy driving innovativeness and business growth in international markets," *Int. Mark. Rev.*, vol. 40, no. 3, pp. 479–496, 2023. https://doi.org/10.1108/IMR-09-2020-0215

[5] R. Q. Liu, Y. C. Lee, and H. L. Mu, "Customer classification and market basket analysis using K-means clustering and association rules: Evidence from distribution big data of korean retailing company," *Knowl. Manag. Res.*, vol. 19, no. 4, pp. 59–76, 2018. https://doi.org/10.15813/kmr.2018.19.4.004

[6] Y. Liu and Y. Guan, "FP-growth algorithm for application in research of market basket analysis," in *2008 IEEE International Conference on Computational Cybernetics*, Stara Lesna, Slovakia, 2008, pp. 269–272. https://doi.org/10.1109/ICCCYB.2008.4721419

[7] M. Hossain, A. S. Sattar, and M. K. Paul, "Market basket analysis using apriori and FP growth algorithm," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, Stara Lesna, Slovakia, 2019, pp. 1–6. https://doi.org/10.1109/ICCIT48885.2019.9038197

[8] B. Shim, K. Choi, and Y. Suh, "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7736–7742, 2012. https://doi.org/10.1016/j.eswa.2012.01.080

[9] M. S. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4995–5005, 2024. https://doi.org/10.1007/s00521-023-09339-6

[10] S. Dwivedi and A. Singh, "A study of customer segmentation based on RFM analysis and K-means," in *International Conference on Innovative Computing and Communication*. Singapore: Springer, 2023. https://doi.org/10.1007/978-981-99-4071-4_27

[11] S. Ahmed, J. Karmoker, R. Mojumder, M. M. Rahman, M. G. R. Alam, and M. T. Reza, "Hyperautomation in super shop using machine learning," *Eng. Proc.*, vol. 39, no. 1, p. 63, 2023. https://doi.org/10.3390/engproc2023039063

[12] X. Xiahou and Y. Harada, "B2C E-commerce customer churn prediction based on K-means and SVM," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 2, pp. 458–475, 2022. https://doi.org/10.3390/jtaer17020024

[13] K. T. Tang, Y. Sun, P. H. Lee, and Q. Huang, "Apply apriori algorithm in supermarket layout research," in *2020 International Conference on Modern Education and Information Management (ICMEIM)*, Dalian, China, 2020, pp. 521–524. https://doi.org/10.1109/ICMEIM51375.2020.00122

[14] L. Hamdad and K. Benatchba, "Association rules mining," *SN COMPUT. SCI.*, vol. 2, p. 449, 2021. https://doi.org/10.1007/s42979-021-00819-x

[15] M. Martinez, B. Escobar, G. D. Maria-Elena, and D. P. Pinto-Roa, "Market basket analysis with association rules in the retail sector using Orange. Case Study: Appliances Sales Company," *CLEI Electron. J.*, vol. 24, no. 2, Jul 2021. https://doi.org/10.19153/cleiej.24.2.12

[16] D. Chen, K. Guo, and G. Ubakanma, "Predicting customer profitability over time based on RFM time series," *Int. J. Bus. Forecast. Mark. Intell.*, vol. 2, no. 1, pp. 1–18, 2015. https://doi.org/10.1504/IJBFMI.2015.075325

[17] E. Yıldız, C. Güngör Şen, and E. E. Işık, "A hyper-personalized product recommendation system focused on customer segmentation: An application in the fashion retail industry," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, pp. 571–596, 2023.

[18] R. Sann, P. C. Lai, and S. Y. Liaw, "Understanding customers' insights using attribution theory: A text mining and rule-based machine learning two-step multifaceted method," *Appl. Sci.*, vol. 13, no. 5, p. 3073, 2023. https://doi.org/10.3390/app13053073

[19] Z. Li, X. Li, R. Tang, and L. Zhang, "Apriori algorithm for the data mining of global cyberspace security issues for human participatory based on association rules," *Front. Psychol.*, vol. 11, p. 582480, 2021. https://doi.org/10.3389/fpsyg.2020.582480

[20] M. A. Bin Ahmadon, S. Yamaguchi, A. K. Mahamad, and S. Saon, "Refining preference-based recommendation with associative rules and process mining using correlation distance," *Big Data Cogn. Comput.*, vol. 7, no. 1, p. 34, 2023. https://doi.org/10.3390/bdcc7010034

[21] J. Zhang, P. Lin, and A. Simeone, "Information mining of customers preferences for product specifications determination using big sales data," *Procedia CIRP*, vol. 109, pp. 101–106, 2022. https://doi.org/10.1016/j.procir.2022.05.221

[22] Y. D. Seo, Y. G. Kim, E. Lee, and H. Kim, "Group recommender system based on genre preference focusing on reducing the clustering cost," *Expert Syst. Appl.*, vol. 183, p. 115396, 2021. https://doi.org/10.1016/j.eswa.2021.115396

[23] E. Acar, G. Sarıyer, V. Jain, and B. Ramtiyal, "Discovering hidden associations among environmental disclosure themes using data mining approaches," *Sustainability*, vol. 15, no. 14, p. 11406, 2023. https://doi.org/10.3390/su151411406

[24] Z. Wen, W. Lin, and H. Liu, "Machine-learning-based approach for anonymous online customer purchase

intentions using clickstream data," *Systems*, vol. 11, no. 5, p. 255, 2023. https://doi.org/10.3390/systems11050255

[25] C. Zhang, J. Qiu, Y. Yang, and J. Zhao, "Residential customers-oriented customized electricity retail pricing design," *Int. J. Electr. Power Energy Syst.*, vol. 146, p. 108766, 2023. https://doi.org/10.1016/j.ijepes.2022.108766

[26] F. M. Talaat, A. Aljadani, B. Alharthi, M. A. Farsi, M. Badawy, and M. Elhosseini, "A mathematical model for customer segmentation leveraging deep learning, explainable AI, and RFM analysis in targeted marketing," *Mathematics*, vol. 11, no. 18, p. 3930, 2023. https://doi.org/10.3390/math11183930