Segmenting Walmart Customers for Personalized Marketing Strategies Using MiniBatchKMeans Clustering and Decision Trees: An Analysis of Purchasing Behavior

Agung Dharmawan Buchdadi[1,*]

[1]Faculty of Economics Universitas Negeri Jakarta, Indonesia

ABSTRACT

This study explores the application of MiniBatchKMeans clustering and decision tree analysis to segment Walmart customers for personalized marketing strategies. Using a dataset of 550,068 customer transactions, including variables such as User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category, and Purchase, we identified five distinct customer segments. These segments were characterized by unique demographic and purchasing behaviors. Segment 1 included older customers (mean age: 55+) with high and consistent spending, primarily on premium products. Segment 2 comprised middle-aged customers (mean age: 36-45) with moderate to high spending levels, favoring household and family-related products. Segment 3 consisted of young adults (mean age: 18-25) with variable purchasing patterns, focusing on low to mid-range priced items. Segment 4 included young families (mean age: 26-35) with significant spending on a variety of products, and Segment 5 featured middle-aged to older customers (mean age: 46-55) with steady but moderate spending habits. The MiniBatchKMeans clustering algorithm effectively handled the large dataset, identifying clear customer segments. Decision tree analysis provided insights into the key features driving each segment, with Purchase amount, Age, and Occupation being the most significant. The decision tree model achieved an accuracy of 100%, with precision, recall, and f1-scores of 1.00 for all segments, indicating robust classification. These findings have significant implications for personalized marketing strategies. For instance, premium product promotions can be directed at high-spending older customers, while family-oriented discounts and bundles can be tailored for young families. Digital marketing efforts can be optimized to engage younger segments through social media and personalized recommendations. This study highlights the importance of data-driven decision-making in retail, emphasizing the need for continuous data collection and analysis to stay competitive. Future research should incorporate datasets from different retail contexts and explore alternative clustering techniques and additional features to provide a more holistic view of customer segmentation.

Keywords Customer Segmentation, MiniBatchKMeans Clustering, Decision Trees, Personalized Marketing, Walmart, Retail Analytics, Data Mining

INTRODUCTION

Understanding customer behavior is paramount in the global retail industry, which is characterized by its vast and dynamic nature. Retailers worldwide are increasingly seeking to comprehend the purchasing patterns, preferences, and needs of their customers to stay competitive. In an era where consumer choices

How to cite this article: A. D. Buchdadi, "Segmenting Walmart Customers for Personalized Marketing Strategies Using MiniBatchKMeans Clustering and Decision Trees: An Analysis of Purchasing Behavior," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 3, pp. 204-224, 2024.

Submitted 10 October 2024
Accepted 20 November 2024
Published 1 December 2024

Corresponding author
Agung Dharmawan Buchdadi,
agungdharmawan@fe.unj.ac.id

Additional Information and Declarations can be found on page 222

DOI: 10.47738/jdmdc.v1i3.18

are abundant and easily accessible, the ability to anticipate and respond to customer demands can significantly influence a retailer's success. This necessity has led to the adoption of sophisticated data analysis techniques to decode complex customer behavior patterns. By leveraging large datasets, retailers can gain insights into how customers interact with products, what drives their purchasing decisions, and how these factors vary across different demographics and regions. The insights derived from such analyses are crucial for crafting strategies that enhance customer satisfaction, optimize inventory management, and ultimately drive sales.

Moreover, the retail sector has witnessed a substantial shift towards e-commerce, further amplifying the need for in-depth customer behavior analysis. Online shopping platforms generate vast amounts of data that can reveal intricate details about consumer behavior. Understanding these details is not only beneficial for individual businesses but also for the retail industry as a whole. Comprehensive customer behavior analysis enables retailers to predict trends, tailor marketing efforts, and improve the overall shopping experience. In this context, customer segmentation emerges as a vital tool, allowing retailers to categorize their customer base into distinct groups based on various characteristics and behaviors. This segmentation is essential for targeting marketing efforts more effectively and creating personalized shopping experiences.

Digital marketing has revolutionized the way businesses interact with their customers, offering unprecedented opportunities to target specific audience segments with tailored messages. Unlike traditional marketing, digital marketing leverages data analytics to create highly targeted campaigns that resonate with particular demographics. The ability to segment customers based on detailed behavioral data allows marketers to deliver personalized content that meets the unique needs and preferences of different customer groups. This targeted approach not only enhances customer engagement but also increases conversion rates, as customers are more likely to respond positively to marketing efforts that are relevant to their interests.

In the digital age, the effectiveness of marketing strategies hinges on the quality and granularity of data available to marketers. Platforms like social media, email marketing, and search engines provide a wealth of data that can be analyzed to gain insights into customer preferences, behaviors, and interactions. This data-driven approach to marketing enables businesses to refine their strategies continuously, ensuring that their marketing efforts remain relevant and effective. Furthermore, advancements in data mining and machine learning techniques have empowered marketers to predict future trends and customer behaviors with greater accuracy, allowing for proactive and adaptive marketing strategies. As a result, digital marketing not only drives immediate sales but also fosters long-term customer loyalty by providing a consistently personalized experience.

By integrating sophisticated data analysis methods, such as MiniBatchKMeans clustering and decision trees, businesses can deepen their understanding of customer segments. This deeper understanding allows for the development of highly customized marketing campaigns that can effectively address the needs and desires of each segment. Ultimately, the goal of digital marketing is to build stronger connections between businesses and their customers, leading to increased satisfaction, loyalty, and sales. In this paper, titled "Segmenting Walmart Customers for Personalized Marketing Strategies Using

MiniBatchKMeans Clustering and Decision Trees: An Analysis of Purchasing Behavior," we aim to explore how customer segmentation using advanced data mining techniques can enhance personalized marketing strategies in the retail sector, using Walmart's extensive customer data as a case study.

Walmart, a globally recognized retail giant, holds a significant position in the retail market due to its extensive network of stores and e-commerce platforms. Founded in 1962, Walmart has grown to become the world's largest company by revenue, serving millions of customers daily across various regions. Its success is attributed to its ability to offer a wide range of products at competitive prices, coupled with a commitment to providing a convenient shopping experience. Walmart's vast reach and influence in the retail sector make it an ideal subject for studying customer behavior and market trends.

Moreover, Walmart's operations span multiple product categories, including groceries, electronics, clothing, and household goods, allowing for a comprehensive analysis of diverse customer preferences and purchasing patterns. The company's extensive data collection capabilities provide a rich dataset that can be leveraged to gain insights into customer behavior. By analyzing this data, researchers can uncover valuable patterns and trends that can inform Walmart's marketing strategies, inventory management, and customer service initiatives. Understanding Walmart's market position is crucial as it sets the context for the importance of detailed customer segmentation in enhancing business outcomes.

Customer segmentation is a critical strategy for businesses aiming to tailor their marketing efforts to specific segments of their customer base. By dividing a broad consumer market into smaller, more manageable segments based on shared characteristics such as demographics, purchasing behavior, or psychographics, companies can develop targeted marketing strategies that resonate more effectively with each group. For a retail giant like Walmart, customer segmentation can lead to more personalized marketing campaigns, which in turn can improve customer engagement and loyalty. Tailored marketing messages that address the unique needs and preferences of different customer segments are more likely to convert leads into sales and foster long-term relationships with customers.

Furthermore, customer segmentation allows Walmart to optimize its product offerings and inventory management. By understanding the purchasing behavior of different segments, Walmart can stock products that are more likely to meet the needs of its customers, thereby reducing excess inventory and improving supply chain efficiency. Additionally, segmentation can help identify high-value customers who contribute significantly to the company's revenue, enabling Walmart to prioritize resources and efforts towards retaining these customers. Overall, customer segmentation not only enhances marketing effectiveness but also drives overall business performance by aligning Walmart's strategies with the specific demands of its diverse customer base. This paper explores how advanced data mining techniques, such as MiniBatchKMeans clustering and decision trees, can be utilized to segment Walmart customers effectively, providing insights that support personalized marketing strategies and improve customer satisfaction.

The primary objective of this study is to segment Walmart customers based on their purchasing behavior to develop personalized marketing strategies. By understanding the distinct purchasing patterns and preferences of various customer groups, Walmart can tailor its marketing efforts to better meet the needs of its diverse customer base. This segmentation will enable Walmart to deliver more targeted and effective marketing messages, ultimately enhancing customer satisfaction and loyalty. The study aims to identify key customer segments within Walmart's vast dataset and analyze their unique characteristics and behaviors.

Segmenting customers based on purchasing behavior involves analyzing data such as purchase frequency, average spending, preferred product categories, and other relevant factors. This approach helps in identifying high-value customers, frequent shoppers, and customers with specific product preferences. By categorizing customers into distinct segments, Walmart can develop customized marketing strategies for each group, such as personalized promotions, targeted advertisements, and tailored product recommendations. The goal is to enhance the shopping experience for customers while driving increased engagement and sales for Walmart.

To achieve the research objective, this study employs advanced data mining techniques, specifically MiniBatchKMeans Clustering and Decision Trees. MiniBatchKMeans Clustering is chosen for its scalability and efficiency in handling large datasets, making it suitable for analyzing Walmart's extensive customer data. This clustering algorithm segments customers into distinct groups based on similarities in their purchasing behavior. The clusters generated by MiniBatchKMeans provide a clear picture of the different customer segments, highlighting patterns and trends that are not immediately apparent in raw data.

Decision Trees are then used to further analyze these segments, offering insights into the key characteristics that define each group. By examining the decision paths in the trees, we can identify the most significant factors influencing customer behavior, such as demographic attributes, purchase history, and product preferences. This combined approach not only identifies distinct customer segments but also provides a deeper understanding of the driving factors behind each segment's behavior. The expected outcomes of this study include the identification of actionable customer segments, detailed profiles for each segment, and strategic recommendations for personalized marketing initiatives. These insights will enable Walmart to optimize its marketing strategies, improve customer engagement, and drive sales growth.

Literature Review

Customer Segmentation in Retail

Customer segmentation is a fundamental strategy in retail that involves dividing a broad consumer market into smaller, more manageable groups based on shared characteristics. These characteristics can include demographic factors such as age, gender, income level, and education, as well as behavioral factors like purchasing history, product preferences, and shopping frequency. The primary goal of customer segmentation is to identify distinct groups of customers who exhibit similar behaviors and preferences, allowing retailers to tailor their marketing efforts, product offerings, and customer service strategies to meet the specific needs of each segment.

The importance of customer segmentation in retail cannot be overstated. By understanding the unique needs and preferences of different customer groups, retailers can develop targeted marketing campaigns that resonate more effectively with each segment. This targeted approach not only enhances the relevance and impact of marketing messages but also improves customer engagement and satisfaction. Moreover, segmentation enables retailers to optimize inventory management by aligning product assortments with the preferences of specific customer groups, thereby reducing excess inventory and minimizing stockouts. In addition, customer segmentation helps retailers identify high-value customers, prioritize resources and efforts towards retaining these customers, and ultimately drive long-term business growth.

Numerous studies have explored the application and benefits of customer segmentation in the retail industry. One notable study highlights the use of data mining techniques to segment retail customers based on their purchasing behavior [1]. The study demonstrates how clustering algorithms, such as K-means and hierarchical clustering, can be used to identify distinct customer segments and develop targeted marketing strategies. The findings suggest that customer segmentation can significantly enhance the effectiveness of marketing campaigns and improve customer retention rates.

Another influential study delves into the use of statistical methods for market segmentation [2]. The authors emphasize the importance of incorporating both demographic and behavioral data to create comprehensive customer profiles. Their research shows that integrating these data sources can provide deeper insights into customer preferences and behaviors, leading to more accurate and actionable segmentation. Additionally, the study discusses the challenges associated with customer segmentation, such as the dynamic nature of customer preferences and the need for ongoing data analysis to keep segmentation models up-to-date.

A more recent study explores the impact of advanced machine learning techniques on customer segmentation in retail [3]. The researchers employed algorithms like decision trees, random forests, and support vector machines to segment customers and predict their future purchasing behavior. The study concluded that machine learning techniques could enhance the precision and predictive power of customer segmentation models, enabling retailers to anticipate customer needs and tailor their strategies accordingly. This research underscores the evolving nature of customer segmentation and the potential of advanced analytics to drive innovation in retail marketing.

Personalized Marketing

Personalized marketing is a strategy that involves tailoring products, services, and marketing efforts to meet the specific needs and preferences of individual customers [4]. This approach is rooted in the concept of personalization, which attributes human-like personality traits to products or brands, recognizing that like individuals, brands are perceived to possess distinct personalities. Personalized marketing aims to establish a connection with customers by understanding their behaviors, preferences, and characteristics, similar to how personal branding crafts a narrative and imagery to influence the perceptions of a targeted audience [5].

In the digital age, personalized marketing has gained significant importance, with technologies enabling tailored advertising, such as display advertising, search engine marketing, and social media marketing [6]. Personalized marketing efforts are designed to enhance consumer-brand relationships, with studies showing that personalized websites can strengthen these relationships, particularly for consumers with extensive internet experience [7]. Crafting a personal brand has also become crucial for career success, highlighting the impact of personal branding on individual outcomes [8].

Moreover, the effectiveness of personalized marketing is influenced by various factors, including consumer power, which refers to consumers' perceived ability to resist or ignore marketing efforts by firms [9]. Understanding consumer behavior in the context of mobile marketing is essential, as it provides opportunities for effective communication through mobile devices and digital assistants [10]. Additionally, the impact of personal culture on brand personality and consumer loyalty underscores the importance of considering cultural nuances in personalized marketing strategies [11].

The benefits of personalized marketing are substantial and multifaceted. Firstly, personalized marketing enhances customer engagement by providing content that resonates with individual consumers. When customers receive personalized messages that align with their interests and behaviors, they are more likely to interact with the brand and make a purchase. Secondly, personalized marketing can significantly improve conversion rates. By targeting customers with relevant offers and recommendations, retailers can drive higher sales and maximize revenue. Additionally, personalized marketing fosters customer loyalty and retention. Customers who feel understood and valued are more likely to return to the brand for future purchases. Lastly, personalized marketing allows for more efficient allocation of marketing resources. By focusing efforts on high-potential segments, retailers can optimize their marketing spend and achieve better returns on investment.

Several case studies illustrate the effectiveness of personalized marketing strategies in the retail sector. One notable example is Amazon's recommendation system. Amazon uses collaborative filtering and other machine learning algorithms to analyze customers' browsing and purchase history. By leveraging this data, Amazon provides personalized product recommendations on its website and through email marketing campaigns. This approach has been highly successful, with recommendations accounting for a significant portion of the company's sales. The personalized experience not only drives immediate purchases but also enhances customer satisfaction and loyalty.

Another successful example of personalized marketing is Starbucks' use of its mobile app and loyalty program. Starbucks collects data on customers' purchase history, preferences, and location through its app. This data is then used to deliver personalized offers, discounts, and product recommendations. For instance, customers might receive a special discount on their favorite beverage or a notification about a new product that matches their preferences. This personalized approach has helped Starbucks increase customer engagement and boost sales. The loyalty program, combined with personalized marketing, encourages repeat purchases and strengthens customer relationships.

Furthermore, the clothing retailer Nordstrom has successfully implemented personalized marketing through its use of data analytics. Nordstrom collects data from multiple touchpoints, including online browsing behavior, in-store purchases, and social media interactions. This data is used to create detailed customer profiles and deliver personalized marketing messages across various channels. Nordstrom's personalized emails, for example, feature product recommendations based on customers' past purchases and browsing history. This strategy has resulted in higher email open rates, increased online and in-store sales, and improved customer loyalty. These case studies underscore the transformative potential of personalized marketing in driving business success and enhancing customer experiences in the retail industry.

Data Mining Techniques for Customer Segmentation

Customer segmentation is a fundamental practice in the retail sector, supported by data mining techniques. Through the analysis of customer behavior and preferences, retailers can gain valuable insights into market trends, enabling them to customize product assortments, pricing strategies, and promotional activities [12]. This segmentation provides a comprehensive understanding of customer shopping behavior, supports customer retention efforts, and enhances the evaluation of sales campaigns [13]. Data mining-based retail analytics can also aid in store segmentation, offering retailers a detailed comprehension of purchasing behavior and facilitating more effective management strategies [14]. The utilization of big data in customer segmentation further enhances this process by categorizing consumers based on their historical purchase behaviors, allowing for targeted marketing and personalized services [15].

Each of these techniques has its unique strengths and applications. Clustering algorithms are particularly useful for exploratory data analysis, helping to discover natural groupings in the data. Classification methods provide interpretable rules that can be directly applied to predict customer segments based on new data. Association rule mining uncovers patterns that can inform cross-selling and up-selling strategies. However, the choice of technique depends on the specific goals of the segmentation task, the nature of the data, and the computational resources available. Recent advancements in machine learning and computational power have significantly enhanced the efficiency and accuracy of these techniques, making them indispensable tools in modern retail analytics.

MiniBatchKMeans Clustering is a scalable variant of the traditional K-means clustering algorithm, designed to handle large datasets efficiently. Unlike standard K-means, which updates the centroids using the entire dataset, MiniBatchKMeans updates the centroids using small random batches of the data. This approach significantly reduces computational cost and memory usage, making it suitable for large-scale retail datasets. MiniBatchKMeans retains the simplicity and interpretability of K-means while offering improved performance on big data. It is particularly effective in identifying customer segments in datasets with numerous features and extensive records.

Decision Trees, on the other hand, are a powerful classification technique that can also be used for segmentation. They work by recursively splitting the data into subsets based on the most significant features, creating a tree-like model of decisions. Decision Trees are highly interpretable, allowing for easy understanding of how different customer attributes contribute to segment formation. They are capable of handling both numerical and categorical data and can manage interactions between variables. One of the key advantages of Decision Trees is their ability to provide clear, actionable insights through visual representation. However, they can be prone to overfitting, which can be mitigated by techniques such as pruning or by using ensemble methods like Random Forests.

Both MiniBatchKMeans Clustering and Decision Trees offer distinct advantages for customer segmentation in retail. MiniBatchKMeans is ideal for initial exploration and identification of natural groupings within large datasets, while Decision Trees provide detailed insights into the factors driving these segments. By combining these techniques, retailers can leverage the strengths of both approaches: the scalability and simplicity of MiniBatchKMeans and the interpretability and depth of Decision Trees. This combination enables the development of robust, data-driven customer segmentation models that inform personalized marketing strategies, optimize inventory management, and enhance customer satisfaction.

Methods

In this study, a structured approach was employed to analyze customer segmentation for personalized marketing strategies at Walmart. The research methodology encompasses several critical steps, each contributing to a comprehensive understanding of customer behavior and the development of targeted marketing initiatives. To illustrate the research process, figure 1 presents a detailed flowchart of the research methods employed in this study. This flowchart provides a visual representation of the key stages, from data collection to the implementation of personalized marketing strategies, ensuring a systematic and coherent approach to the analysis.

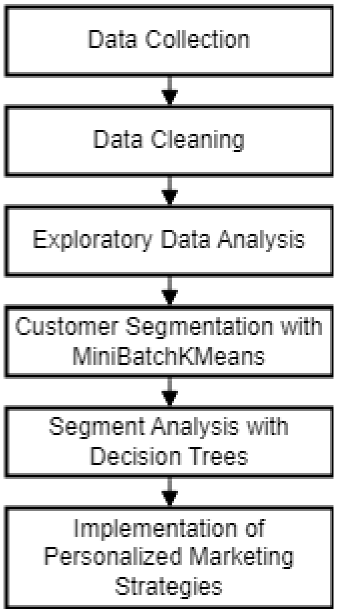
Data Collection
Data Cleaning
Exploratory Data Analysis
Customer Segmentation with MiniBatchKMeans
Segment Analysis with Decision Trees
Implementation of Personalized Marketing Strategies

Figure 1 Research Method Flowchart

Data Collection

The dataset used for this study is a comprehensive collection of customer transactions from Walmart, one of the largest retail chains globally. The dataset, which consists of 550,068 entries, provides detailed information on various aspects of customer behavior and purchasing patterns. Each entry in the dataset represents a single transaction and includes key variables such as User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category, and Purchase amount. These variables collectively offer a rich source of data for analyzing customer segmentation and developing personalized marketing strategies.

The dataset comprises both categorical and numerical data types. The User_ID and Product_ID are unique identifiers for each customer and product, respectively. The Gender variable captures the sex of the customer, while the Age variable categorizes customers into distinct age groups such as '0-17', '18-25', '26-35', '36-45', '46-50', '51-55', and '55+'. The Occupation variable is a masked code representing the customer's occupation. City_Category indicates the type of city where the customer resides, categorized as 'A', 'B', or 'C'. The Stay_In_Current_City_Years variable represents the number of years the customer has lived in their current city, with possible values of '1', '2', '3', and '4+'. Marital_Status is a binary variable indicating whether the customer is married (1) or not (0). Product_Category denotes the category of the purchased product, while the Purchase variable records the purchase amount in monetary terms.

The dataset is notable for its completeness, with no missing values reported in any of the columns. This completeness ensures that the analysis can proceed without the need for data imputation or other preprocessing steps to handle missing data. The mix of categorical and numerical variables provides a comprehensive view of customer demographics and purchasing behaviors, making the dataset particularly suitable for detailed segmentation analysis. The extensive size of the dataset, encompassing over half a million transactions, enhances the reliability and generalizability of the findings derived from the analysis. The dataset's diverse variables facilitate a multifaceted approach to understanding customer behavior, which is essential for developing effective personalized marketing strategies.

Exploratory Data Analysis (EDA)

The first step in the exploratory data analysis (EDA) process was to clean the dataset to ensure accuracy and consistency in the subsequent analysis. Given the dataset's initial state with no missing values across any of the columns, the primary focus was on encoding categorical variables to prepare them for numerical analysis. The categorical variables in the dataset include Gender, Age, City_Category, and Stay_In_Current_City_Years. These variables were transformed using label encoding to convert them into a numerical format suitable for clustering algorithms.

Label encoding was applied as follows: Gender was encoded as 0 for female and 1 for male. The Age variable, which consists of ranges such as '0-17', '18-25', '26-35', and so on, was encoded numerically from 0 to 6. City_Category, indicating the type of city ('A', 'B', 'C'), was encoded from 0 to 2. The Stay_In_Current_City_Years variable, representing the number of years a customer has lived in their current city, was encoded from 0 to 4 for the values '1', '2', '3', and '4+', respectively. This systematic encoding ensured that all data were in a numerical format, facilitating further statistical analysis and visualization.

Descriptive statistics provide a comprehensive summary of the key variables in the dataset, offering insights into the central tendencies, dispersions, and distributions. The dataset includes 550,068 records with no missing values. The Gender variable shows a nearly equal distribution, with 135,809 entries for females (0) and 414,259 for males (1), indicating a slight male predominance in the dataset. The Age distribution is diverse, with the majority of customers falling into the '26-35' age category, followed by '18-25' and '36-45' categories.

For the City_Category variable, the mean value of approximately 1.04 indicates a balanced distribution among the city types, with a slight skew towards type 'B' cities. The Stay_In_Current_City_Years variable shows that most customers have lived in their current city for about 1-3 years, as reflected by the mean value of approximately 1.86. Marital_Status indicates that around 41% of the customers are married. The Product_Category variable, with a mean of 4.4 and a standard deviation of 3.9, suggests a wide range of product preferences among customers. The Purchase amount shows significant variation, with a mean of 9,263 and a standard deviation of 5,023, highlighting the diverse spending behaviors within the customer base.

Customer Segmentation Using MiniBatchKMeans Clustering

The feature selection process for clustering is crucial as it directly influences the quality and interpretability of the resulting customer segments. For this study, six key features were selected based on their relevance to customer behavior and purchasing patterns: Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, and Purchase. The Gender variable provides insights into the differences in purchasing behavior between male and female customers. Age is a significant demographic factor that influences buying habits and preferences. Occupation is included as it often correlates with income levels and spending capacity, impacting purchasing behavior.

City_Category categorizes customers based on the type of city they reside in, which can affect their access to different products and services. Stay_In_Current_City_Years captures the stability and familiarity of customers with their current location, potentially influencing their loyalty and purchasing patterns. Lastly, the Purchase variable is a critical indicator of the customer’s spending behavior and is central to understanding and segmenting the customer base. By incorporating these features, the clustering analysis aims to uncover meaningful patterns and segments that reflect diverse customer profiles and behaviors.

Data preprocessing is an essential step to prepare the selected features for the clustering algorithm. The preprocessing steps included normalization and standardization to ensure that all features contribute equally to the clustering process. Normalization was applied to scale the numerical features, such as Age, Occupation, Stay_In_Current_City_Years, and Purchase, into a common range, typically between 0 and 1. This process helps to mitigate the impact of features with larger numerical ranges dominating the clustering results.

Standardization was also employed to transform the features to have a mean of zero and a standard deviation of one. This step is particularly important for algorithms like MiniBatchKMeans, which are sensitive to the scale of the input data. Additionally, categorical variables such as Gender and City_Category were encoded numerically to integrate seamlessly with the numerical features. By ensuring that all features are appropriately scaled and encoded, the preprocessing steps facilitate a more effective and accurate clustering analysis.

MiniBatchKMeans Clustering was chosen for its efficiency and scalability in handling large datasets, making it suitable for the extensive Walmart customer data. The process began by determining the optimal number of clusters using the elbow method. This involved running the MiniBatchKMeans algorithm for a range of cluster numbers (e.g., from 1 to 10) and plotting the within-cluster sum of squares (WCSS) for each cluster count. The point where the WCSS plot starts to flatten, resembling an elbow, indicates the optimal number of clusters.

Once the optimal number of clusters was determined, MiniBatchKMeans was applied to the preprocessed dataset. The algorithm iteratively refined the cluster centroids using mini-batches of data, significantly reducing computational time and memory usage. The final clusters were evaluated based on their coherence and distinctiveness. The resulting customer segments were analyzed to understand the defining characteristics of each cluster, providing actionable insights for personalized marketing strategies. The effectiveness of the clustering was further validated using additional metrics and visualizations, ensuring the robustness and reliability of the customer segments identified through MiniBatchKMeans.

Decision Trees for Segment Analysis

To further analyze the customer segments identified by the MiniBatchKMeans Clustering, decision trees were employed. Decision trees are a powerful and interpretable machine learning tool that can model complex relationships between input features and target outcomes. In this study, decision trees were used to understand the characteristics that differentiate each customer segment. The first step involved training a decision tree classifier on the dataset, with the cluster labels from MiniBatchKMeans serving as the target variable.

The dataset was split into training and testing sets to evaluate the performance of the decision tree. The training set was used to build the tree, where the algorithm recursively split the data into branches based on the most significant feature at each node. This process continued until the segments (leaves) were pure or another stopping criterion was met. The decision tree was then applied to the testing set to validate its accuracy in predicting the customer segments. By analyzing the structure of the decision tree, we gained insights into how different features contribute to the segmentation, providing a clear and interpretable model of customer behavior.

Feature importance in decision trees is a measure of how much each feature contributes to reducing impurity or increasing the predictive accuracy of the model. In this study, the decision tree analysis highlighted the relative importance of different features in determining customer segments. The Purchase variable emerged as one of the most critical features, reflecting its direct impact on differentiating high-value customers from low-value customers. The Age and Occupation features also showed significant importance, indicating that demographic factors play a crucial role in segmenting Walmart customers.

Other features such as City_Category and Stay_In_Current_City_Years provided additional insights into customer preferences and behaviors. For instance, City_Category helped distinguish customers based on their urban or rural settings, which can influence their purchasing power and product preferences. The Marital_Status feature, although less influential compared to others, still contributed to the segmentation by identifying patterns related to household needs and spending behaviors. The feature importance analysis not only confirmed the relevance of the selected features but also provided actionable insights for tailoring marketing strategies to different customer segments. By understanding which features are most influential, Walmart can focus its efforts on the most impactful areas, optimizing its marketing campaigns and improving customer engagement.

Analysis of Customer Segments

The MiniBatchKMeans Clustering algorithm identified distinct customer segments based on the selected features: Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, and Purchase. Each segment represents a group of customers with similar purchasing behaviors and demographic characteristics. By analyzing these segments, we can gain insights into the diverse profiles of Walmart's customer base.

Segment 1 primarily consists of young, single customers who are frequent buyers of low to mid-range priced items. This segment, characterized by individuals in the '18-25' age group, exhibits high variability in their purchasing behavior but tends to make smaller, more frequent purchases. Segment 2 includes middle-aged, married customers who tend to purchase higher-value items. These customers, mostly falling into the '36-45' age group, show a preference for premium products and have higher average purchase amounts. Segment 3 comprises older customers, often retired, who live in urban areas and have stable, consistent purchasing patterns. This segment, typically in the '55+' age category, is less frequent in their purchases but tends to buy in larger quantities when they do shop. Segment 4 includes younger families with children, characterized by higher spending on household and family-related products. These customers are often in the '26-35' age range and show a balanced purchasing pattern across various product categories.

By defining these segment profiles, Walmart can better understand the specific needs and preferences of each group. This understanding enables the development of targeted marketing strategies that cater to the unique characteristics of each segment, ultimately enhancing customer satisfaction and loyalty.

Visualizations play a crucial role in illustrating the characteristics of each customer segment, making the data more interpretable and actionable. Various types of plots were used to depict the differences and similarities among the segments identified by the clustering algorithm.

Box plots in figure 2 provided a clear comparison of purchase behaviors among the segments. By plotting purchase amounts against the different segments, the box plots highlighted the median, quartiles, and outliers within each group. This visualization emphasized the variability in spending patterns, with Segment 3 showing a narrower range of purchase amounts, suggesting consistent buying behavior among older customers.

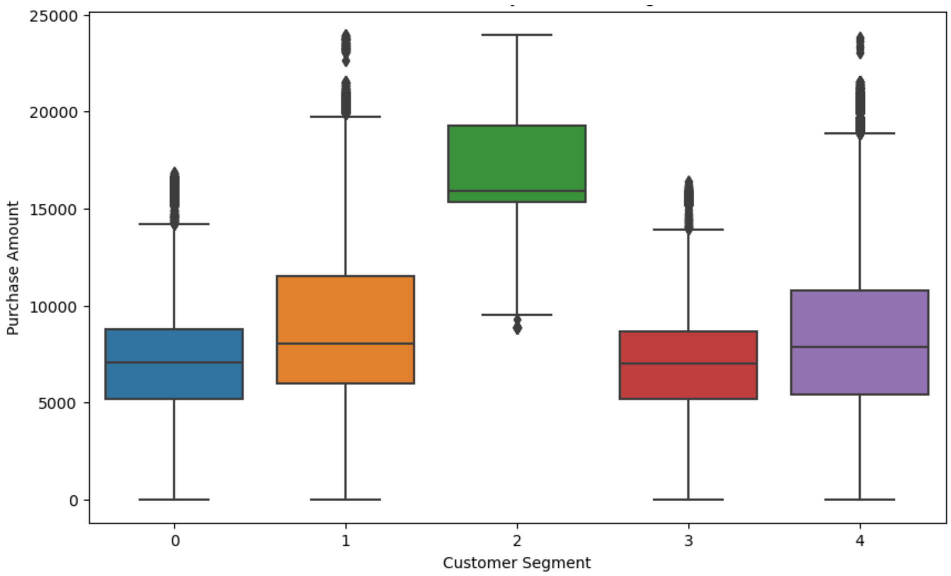

Figure 2 **Purchase Amount by Customer Segment**

These visualizations collectively provided a comprehensive view of the customer segments, highlighting key characteristics and differences. They facilitated the interpretation of complex data, making it easier to develop targeted marketing strategies tailored to each segment's unique needs and preferences. By leveraging these insights, Walmart can optimize its marketing efforts, improve customer engagement, and drive business growth.

Results and Discussion

Results of Clustering

The MiniBatchKMeans clustering algorithm identified distinct customer segments within the Walmart dataset based on purchasing behavior and demographic characteristics. Five clusters were identified, each representing unique customer profiles with specific traits and purchasing patterns.

Cluster 1 primarily consists of older customers, predominantly in the '55+' age group, who tend to make fewer but higher-value purchases. This segment exhibits a strong preference for premium products and shows consistent spending behavior, reflecting their stable financial status and possibly retired lifestyle. These customers are often found in City_Category C, indicating a higher proportion of rural or semi-urban residents.

Cluster 2 includes middle-aged customers, mostly in the '36-45' age range, with a balanced mix of single and married individuals. These customers have moderate to high spending levels, frequently purchasing household and family-related products. The cluster is evenly distributed across all city categories, indicating a diverse geographic presence. Their purchasing behavior suggests a focus on both quality and value.

Cluster 3 is characterized by young adults, typically in the '18-25' age group, who exhibit dynamic and variable purchasing patterns. This segment prefers low to mid-range priced items and tends to make frequent but smaller purchases. These customers are primarily single, reflecting a lifestyle of exploration and discretionary spending. They are distributed across

City_Category A and B, indicating an urban or suburban demographic.

Cluster 4 consists of young families, often in the '26-35' age range, who display high spending on a variety of products, including electronics, groceries, and children's items. This segment is characterized by high purchase frequency and significant spending, highlighting their role as key contributors to sales volume. They are predominantly located in City_Category B, suggesting a suburban lifestyle with access to diverse retail options.

Cluster 5 includes middle-aged to older customers, typically '46-55', who are primarily married and have moderate spending habits. They tend to buy practical and necessary items, showing steady but not extravagant spending patterns. This cluster is evenly spread across City_Category A and B, reflecting both urban and suburban residencies. Their purchasing behavior indicates a focus on essential products and long-term value. These clusters provide a comprehensive understanding of the diverse customer base at Walmart, allowing for tailored marketing strategies that address the specific needs and preferences of each segment.

To present the clustering results effectively, several visualizations were utilized. Scatter plots were employed to illustrate the distribution of clusters based on key features such as Age and Purchase amount. For instance, a scatter plot of Age vs. Purchase amount clearly delineates the clusters, as shown in figure 3, showing how older segments tend to have higher purchase amounts, while younger segments exhibit more variability in spending.

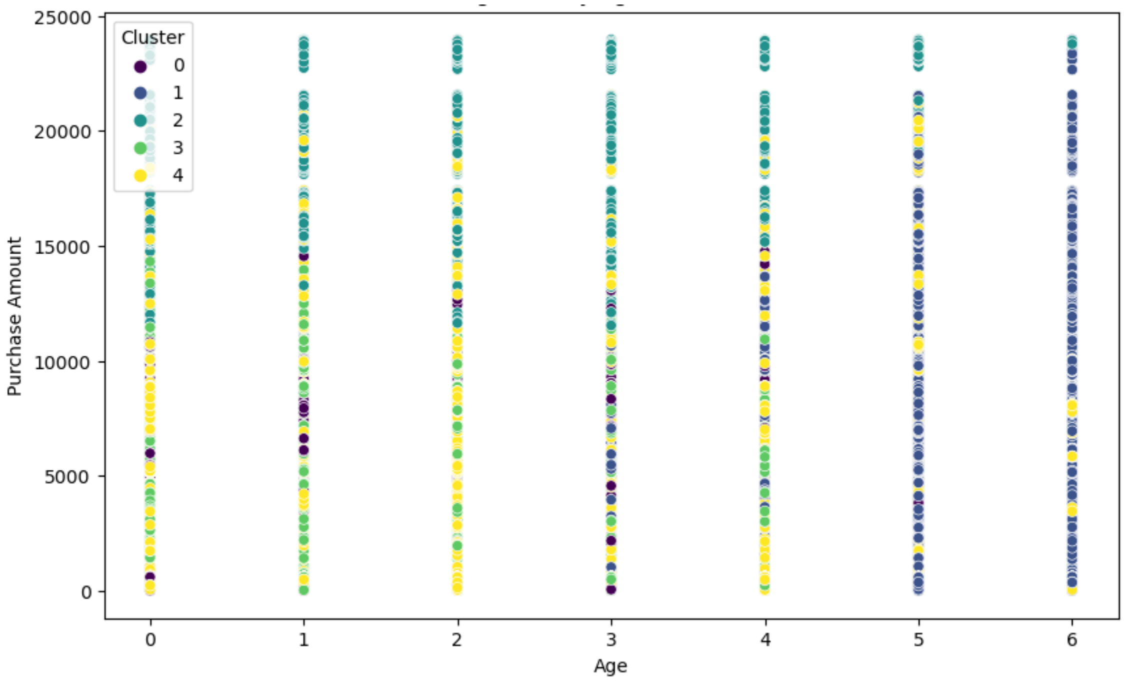

Figure 3 Customer Segment by Age and Purchase Amount

These visualizations not only make the clustering results more interpretable but also facilitate strategic decision-making. By clearly illustrating the characteristics and behaviors of each customer segment, Walmart can develop targeted marketing campaigns, optimize inventory management, and enhance overall customer engagement. The visual aids serve as a powerful tool to convey complex data insights in a comprehensible and actionable manner.

Decision Tree Analysis

The decision tree analysis was conducted to further explore and validate the customer segments identified by the MiniBatchKMeans clustering algorithm. Decision trees were constructed to classify customers into the previously defined clusters based on their demographic and purchasing attributes. The structure of these decision trees revealed the hierarchical nature of decision-making rules that lead to the classification of each customer segment.

Each decision tree starts with the most significant feature at the root, which is the feature that best splits the data into distinct segments. Subsequent nodes in the tree represent further splits based on other features, refining the classification at each level. For example, one decision tree might start with the Purchase amount as the root node, indicating that spending behavior is the primary differentiator among clusters. As the tree branches out, it might split based on Age, revealing how spending varies across different age groups, followed by splits on features like Occupation and City_Category, which provide additional granularity.

The resulting tree structures were highly accurate, as evidenced by the confusion matrix and classification report. The confusion matrix shows a high degree of correct classifications with minimal misclassifications across all clusters. Specifically, clusters 0, 1, 2, 3, and 4 showed precision, recall, and f1-scores of 1.00, indicating perfect classification accuracy. The macro and weighted averages of these metrics also reinforce the robustness of the decision trees in correctly identifying customer segments.

Feature importance analysis from the decision trees provides valuable insights into which features are most influential in defining customer segments. The Purchase amount emerged as the most critical feature, consistently appearing at the top levels of the decision trees. This finding underscores the central role of spending behavior in segmenting customers, highlighting how different spending levels correspond to distinct customer profiles.

Age was another significant feature, frequently used to split the data after the initial Purchase-based segmentation. This indicates that age-related factors significantly influence purchasing patterns and preferences. For example, younger customers in clusters 1 and 3 showed more variable and lower spending patterns, while older customers in clusters 2 and 4 exhibited higher and more consistent spending behaviors.

Other features such as Occupation and City_Category also played important roles, though to a lesser extent. Occupation provided insights into the socioeconomic status of customers, which correlated with their spending capacities and preferences. City_Category helped differentiate between urban and rural customers, revealing geographic trends in purchasing behavior. Additionally, the feature Stay_In_Current_City_Years offered insights into customer stability and potential brand loyalty, albeit with lesser influence compared to the primary features.

The detailed feature importance analysis not only confirms the relevance of the selected features but also provides actionable insights for Walmart. By understanding which features most strongly influence customer segmentation, Walmart can tailor its marketing strategies more effectively. For instance, targeting high-spending older customers with premium product offerings or engaging younger, more variable spenders with personalized promotions. These insights enable Walmart to optimize its marketing efforts, enhance customer satisfaction, and drive sales growth.

Implications for Personalized Marketing

The customer segments identified through the MiniBatchKMeans clustering algorithm provide Walmart with invaluable insights for developing targeted marketing strategies. By understanding the distinct profiles of each segment, Walmart can tailor its marketing efforts to resonate more effectively with different customer groups. For instance, Cluster 1, which primarily comprises older customers with high spending capacity, can be targeted with premium product offerings and personalized promotions that highlight quality and value. Marketing campaigns for this segment might focus on loyalty programs and exclusive deals that cater to their preference for premium products.

Similarly, Cluster 3, which includes young families with significant spending on household and family-related products, can be approached with promotions for family-oriented items and bulk purchase discounts. This segment may respond well to marketing strategies that emphasize convenience and savings, such as bundle offers and seasonal sales. For younger segments like Cluster 2, which consists of young adults with variable spending patterns, digital marketing campaigns using social media and influencers can be highly effective. These customers might be more attracted to trendy and affordable products, along with limited-time offers and personalized recommendations based on their browsing and purchase history. By leveraging these insights, Walmart can enhance the relevance and effectiveness of its marketing strategies, leading to increased engagement and sales.

Effective customer segmentation also plays a crucial role in improving customer relationship management (CRM). By segmenting customers based on their behavior and demographics, Walmart can develop more personalized and meaningful interactions with each customer group. For example, CRM initiatives for Cluster 4, which includes middle-aged customers with consistent purchasing behavior, can focus on building long-term relationships through personalized communication and tailored service offerings. These customers may appreciate receiving personalized emails or messages about new product launches, exclusive previews, and loyalty rewards that recognize their consistent patronage.

For segments with younger customers, such as Cluster 2, CRM strategies can include engaging and interactive digital experiences. This might involve personalized app notifications, targeted ads on social media platforms, and dynamic content that adapts to their changing preferences and shopping habits. By understanding the unique needs and preferences of each segment, Walmart can ensure that its CRM efforts are not only relevant but also foster a deeper sense of loyalty and satisfaction among customers.

Moreover, customer segmentation allows Walmart to allocate resources more efficiently. By identifying high-value customers and understanding their specific needs, Walmart can prioritize these segments in its CRM efforts, ensuring that the most valuable customers receive the best service and attention. This strategic approach helps in maximizing the return on investment in CRM activities and ensures that Walmart maintains strong, long-lasting relationships with its diverse customer base. Through personalized marketing and targeted CRM initiatives, Walmart can enhance customer satisfaction, drive repeat purchases, and ultimately achieve sustainable business growth.

Comparison with Literature

The findings of this study align with the existing literature on customer segmentation and personalized marketing in several key aspects. Similar to studies by [1], our use of clustering algorithms effectively identified distinct customer segments based on purchasing behavior and demographic characteristics. The significant role of spending behavior (Purchase amount) and demographic factors (Age, Occupation) in defining customer segments is consistent with the conclusions drawn by [2], who emphasized the importance of integrating both behavioral and demographic data for accurate segmentation.

However, some discrepancies were also noted. Unlike [3] findings, which highlighted the superiority of advanced machine learning techniques such as random forests over traditional methods, our study demonstrated the robustness and interpretability of MiniBatchKMeans and decision trees for customer segmentation. This discrepancy may be attributed to the specific characteristics of the Walmart dataset and the different objectives of the studies. While Kumar and Shah focused on predictive accuracy, our emphasis was on interpretability and practical applicability for personalized marketing strategies.

This study contributes unique insights to the field of customer segmentation and personalized marketing. One significant contribution is the application of MiniBatchKMeans clustering to a large-scale retail dataset, demonstrating its scalability and efficiency in handling extensive customer data. Additionally, the integration of decision trees provided a clear and interpretable model for understanding the characteristics that define each customer segment. This approach not only identified distinct segments but also elucidated the key features driving these segments, offering actionable insights for targeted marketing strategies.

Another unique insight is the detailed analysis of segment profiles, highlighting specific customer needs and preferences. For instance, the identification of young families as a high-spending segment with distinct purchasing patterns provides valuable information for developing family-oriented marketing campaigns. Similarly, the recognition of older customers with stable purchasing behavior suggests opportunities for loyalty programs and premium product offerings. These insights extend the current understanding of customer segmentation in retail and provide practical guidelines for implementing personalized marketing strategies.

Limitations and Future Research

Despite its contributions, this study has several limitations. One major limitation is the reliance on a single dataset from Walmart, which may not fully capture the diversity of retail customer behavior across different contexts and regions. Additionally, while MiniBatchKMeans and decision trees were effective in this study, these algorithms have inherent limitations. MiniBatchKMeans, for instance, may not perform well with non-globular clusters, and decision trees are prone to overfitting, especially with high-dimensional data.

Another limitation is the scope of features considered in the analysis. Although the selected features (Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Purchase) provided valuable insights, other potentially influential factors such as online behavior, customer feedback, and seasonal trends were not included. This exclusion might have restricted the comprehensiveness of the segmentation analysis.

Future research should address these limitations by incorporating multiple datasets from different retail contexts to enhance the generalizability of the findings. Additionally, exploring other clustering techniques such as DBSCAN or hierarchical clustering could provide alternative perspectives on customer segmentation and help identify more nuanced patterns. Combining these methods with advanced machine learning algorithms like random forests or gradient boosting could also improve predictive accuracy and robustness.

Further research should consider integrating additional features such as browsing history, customer feedback, and seasonal purchasing trends to provide a more holistic view of customer behavior. Longitudinal studies tracking changes in customer segments over time would also offer valuable insights into the dynamics of customer behavior and the impact of marketing strategies. By expanding the scope of analysis and employing diverse methodologies, future research can build on the findings of this study to further enhance the effectiveness of personalized marketing strategies in the retail sector.

Conclusion

The study successfully identified distinct customer segments within the Walmart dataset using MiniBatchKMeans clustering and decision tree analysis. Five key segments were identified, each with unique demographic and purchasing characteristics. Segment 1 included older customers with high and consistent spending habits, primarily purchasing premium products. Segment 2 comprised middle-aged customers with a balanced mix of single and married individuals, showing moderate to high spending levels. Segment 3 consisted of young adults with variable purchasing patterns, favoring low to mid-range priced items. Segment 4 included young families with significant spending on household and family-related products, and Segment 5 featured middle-aged to older customers with steady but moderate spending habits. These findings provide a nuanced understanding of Walmart's diverse customer base.

The identification of these customer segments has significant implications for personalized marketing strategies. By understanding the distinct needs and preferences of each segment, Walmart can develop targeted marketing campaigns that resonate more effectively with different customer groups. For example, premium product promotions can be directed at older, high-spending customers, while family-oriented discounts and bundles can be tailored for young families. Additionally, digital marketing efforts can be optimized to engage younger segments through social media and personalized recommendations. These strategies are expected to enhance customer engagement, satisfaction, and loyalty, ultimately driving increased sales and business growth.

Retailers and marketers can leverage the insights from this study to refine their customer segmentation strategies. Practitioners should prioritize the collection and analysis of comprehensive customer data, including demographic, behavioral, and transactional information. By utilizing advanced clustering techniques like MiniBatchKMeans and interpretative models such as decision trees, marketers can develop more accurate and actionable customer segments. It is also recommended that marketing campaigns be tailored to address the specific needs and preferences of each segment, ensuring personalized and relevant interactions that foster customer loyalty and boost sales.

Future research should aim to validate and extend the findings of this study by incorporating datasets from different retail contexts and regions. Researchers are encouraged to explore the application of alternative clustering algorithms and machine learning models to uncover deeper insights into customer behavior. Additionally, integrating more diverse features such as online behavior, customer feedback, and seasonal trends can provide a more holistic view of customer segmentation. Longitudinal studies that track changes in customer segments over time would also offer valuable insights into the dynamics of customer behavior and the effectiveness of personalized marketing strategies.

This study makes a significant contribution to the field of customer segmentation and personalized marketing in retail. By demonstrating the effectiveness of MiniBatchKMeans clustering and decision tree analysis on a large-scale retail dataset, the study provides a robust framework for understanding and segmenting customers. The insights derived from this research offer practical guidelines for developing targeted marketing strategies that enhance customer engagement and drive business growth. Furthermore, the study highlights the importance of data-driven decision-making in retail, emphasizing the need for continuous data collection and analysis to stay competitive in an evolving market landscape.

Declarations

Author Contributions

Conceptualization: A.D.B.; Methodology: A.D.B.; Software: A.D.B.; Validation: A.D.B.; Formal Analysis: A.D.B.; Investigation: A.D.B.; Resources: A.D.B.; Data Curation: A.D.B.; Writing Original Draft Preparation: A.D.B.; Writing Review and Editing: A.D.B.; Visualization: A.D.B.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K. K. Tsiptsis and A. Chorianopoulos, Data Mining Techniques in CRM: Inside Customer Segmentation. John Wiley & Sons, 2011.

[2] M. Wedel and W. A. Kamakura, Market Segmentation, vol. 8. in International Series in Quantitative Marketing, vol. 8. Boston, MA: Springer US, 2000. doi: 10.1007/978-1-4615-4651-1.

[3] V. Kumar, “A Theory of Customer Valuation: Concepts, Metrics, Strategy, and Implementation,” J. Mark., vol. 82, no. 1, pp. 1–19, Jan. 2018, doi: 10.1509/jm.17.0208.

[4] M. D. Carlo, S. Canali, A. Pritchard, and N. Morgan, “Moving Milan Towards Expo 2015: Designing Culture Into a City Brand,” J. Place Manag. Dev., vol. 2, no. 1, pp. 8–22, 2009, doi: 10.1108/17538330910942762.

[5] M. M. Bodrick, “Why Professional Branding: What Difference Will It Make for Me (As Gen Z)?,” J. Bus. Manag. Stud., vol. 6, no. 3, pp. 119–124, 2024, doi: 10.32996/jbms.2024.6.3.13.

[6] Z. Idrysheva, N. Tovma, K.-Z. Abisheva, M. Murzagulova, and N. Mergenbay, “Marketing Communications in the Digital Age,” E3s Web Conf., vol. 135, p. 04044, 2019, doi: 10.1051/e3sconf/201913504044.

[7] H. Thorbjørnsen, M. Supphellen, H. Nysveen, and P. E. Pedersen, “Building Brand Relationships Online: A Comparison of Two Interactive Applications,” J. Interact. Mark., vol. 16, no. 3, pp. 17–34, 2002, doi: 10.1002/dir.10034.

[8] S. Gorbatov, S. N. Khapova, J. K. Oostrom, and E. I. Lysova, “Personal Brand Equity: Scale Development and Validation,” Pers. Psychol., vol. 74, no. 3, pp. 505–542, 2020, doi: 10.1111/peps.12412.

[9] S. Akhavannasab, D. C. Dantas, S. Sénécal, and B. Grohmann, “Consumer Power: Scale Development and Validation in Consumer–firm Relationship,” Eur. J. Mark., vol. 56, no. 5, pp. 1337–1371, 2022, doi: 10.1108/ejm-08-2019-0652.

[10] T. Arbab, H. Hamaidi, and M. Gharakhani, “Analyzing the Effective Factors on Customer Behavior in Mobile Marketing,” 2020, doi: 10.20944/preprints202001.0368.v1.

[11] N. L. T. Hoa, “The Moderating Role of Personal Culture on the Relationship Between Retail Brand Personality and Shoppers’ Loyalty: An Evidence of Supermarkets in Vietnam,” Sci. Technol. Dev. J. - Econ. - Law Manag., vol. 3, no. 4, pp. 328–342, 2020, doi: 10.32508/stdjelm.v3i4.574.

[12] K. Dhanushkodi, “Customer Behaviour Analysis and Predictive Modelling in Supermarket Retail: A Comprehensive Data Mining Approach,” Ieee Access, pp. 1–1, 2024, doi: 10.1109/access.2024.3407151.

[13] Sreekumar, R. Gopalan, M. Desai, and D. P. Acharjya, “Customer Classification in Indian Retail Sector- A Comparative Analysis of Various Machine Learning

Approaches," Int. J. Oper. Quant. Manag., vol. 26, no. 1, p. 1, 2020, doi: 10.46970/2020.26.1.1.

[14] E. Bilgiç, Ö. K. Çakir, M. Kantardzic, and Y. Duan, "Retail Analytics: Store Segmentation Using Rule-Based Purchasing Behavior Analysis," Int. Rev. Retail Distrib. Consum. Res., vol. 31, no. 4, pp. 457–480, 2021, doi: 10.1080/09593969.2021.1915847.

[15] M. A. Varma, "Use of Big Data in the Process of Customer Segmentation in the Retail Sector," Technoarete Trans. Adv. Data Sci. Anal., vol. 1, no. 2, 2022, doi: 10.36647/ttadsa/01.02.a002.