Enhancing Customer Segmentation for Personalized Marketing in E-commerce with Deep Learning Techniques

Dedeepya Yarra
ID: 11594115
Department of Information Science,
University of North Texas

Haritha Kolli
ID: 11584536
Department of Information Science,
University of North Texas

Jeevan Kumar Tirumalgari
ID: 11601255
Department of Information Science,
University of North Texas

Sreedeep Uppalanchi
ID: 11547873
Department of Information Science,
University of North Texas

Sai Pooja Sanku
ID: 11553365
Department of Information Science,
University of North Texas

Abstract

E-commerce businesses that want to improve their marketing tactics have a huge obstacle when it comes to consumer segmentation. This study delves into cutting-edge analytic methods for identifying and catering to your most valuable clientele. K-means, Gaussian Mixture Models, autoencoders, and a new hybrid approach are just some of the unsupervised learning and dimensionality reduction techniques we use. An extensive database of users' internet activities is analyzed using these methods. Our study identifies real customer categories with practical consequences for targeted messaging and product suggestions. Metrics from the evaluation show that our data-driven categorization is accurate. This research lays out actionable steps for optimizing digital marketing and enhancing consumer happiness in e-commerce via the use of advanced analytics. With the help of the gleaned consumer insights, businesses may better connect with their various client segments through tailored offerings. Our findings reveal data-driven strategies for raising customer involvement and boosting sales by tailoring promotions to underlying beliefs and preferences.

Introduction

In the highly competitive world of Internet commerce, companies must use innovative techniques to attract and retain customers, boosting sales, income, and survival. Understanding basic client categories helps create focused products, experiences, and marketing tactics that meet their needs. This research uses advanced analytics to give client segmentation data for an e-commerce company. Behavior-based machine learning clustering algorithms like K-means and hierarchical approaches categorize clients. This research aims to help firms understand consumer values so they may use data to make decisions that appeal to key segments. This rigorous process allows focused segmentation to improve client experiences, pleasure, and lifetime value. In today's data-driven marketing world, customer segmentation plays a crucial role in getting familiar with one's customers and is key to the organization's success. It helps to identify customer preferences by categorizing them into groups with similarities based on their behavior, demographics, and many other factors. So, that business can improve their strategy to interact with each customer group efficiently to grow their business. As there is a multitude of customer data available, in our study, we focus on exploratory data analysis like preprocessing the raw data and then analyzing that data to draw valuable insights to add value to the firm. We focus on visualizing the insights drawn from the analysis. Primarily, we explore various clustering algorithms to categorize the customers into multiple clusters to obtain the best model with effective results for further study.

Related Work

In the e-commerce industry, customer segmentation is essential to individualized marketing techniques (Ballestar, Grau-Carles, Sainz, 2018). It includes dividing customers into groups based on commonalities in order to maximize engagement and the value they add to the organization. The importance of client segmentation in the context of the cash-back business model has been highlighted by Ballestar et al. (2018). Their study shows how segmentation strategies may be used in the real world to increase customer engagement and loyalty in e-commerce. The authors provide examples of how segmentation approaches might improve client retention and satisfaction. He and Li (2016) explore customer segmentation research and application, focusing on e-commerce websites. Their research offers insights into how

segmentation strategies might be applied practically in the world of digital technology. He and Li expose the particular potential and problems connected with client segmentation in online retail environments by concentrating on e-commerce platforms. The use of clustering algorithms for efficient client segmentation in e-commerce has been studied by Punhani, Arora, Sabitha, and Shukla (2021). Their study places a strong emphasis on the usage of cutting-edge machine learning methods to classify clients according to their actions and preferences. The study emphasizes how clustering algorithms can offer insightful data for marketing strategy optimization and enhancing overall business performance. When working with data from several categories, Wu and Chou (2011) suggest a soft-clustering strategy for customer segmentation in e-commerce. Their study focuses on the challenges of categorizing customers with varied buying habits across several product categories. The authors give examples of how soft-clustering methods can successfully manage the complexities of multi-category data in e-commerce environments. Tsai, Hu, and Lu (2015) underline the relevance of client segmentation in retail contexts. The authors want to improve marketing strategies in this competitive business by using two different clustering algorithms. This will ultimately increase client engagement and loyalty. This demonstrates how flexible client segmentation approaches are across different industries.

Advanced analytics and data-driven strategies have also become mostly common due the increase in development of the e-commerce's industry. In e-commerce the transformational potential of big data analytics was highlighted by Akter and Wamba's (2016). Their work highlights how organizations may use the abundance of data to their advantage by gaining significant insights into their target markets, hence enhancing consumer segmentation tactics and overall business performance. One of the key components of ecommerce industry is the utilization of recommendation systems, which offer clients personalized suggestions based on their interests and behavior (Schafer et al., 2001). By using data mining techniques to examine previous interactions, these systems let companies provide specialized product recommendations. In addition to helping to boost sales, this fosters a more enjoyable retail environment for customers. Visualization is a critical component of successful consumer segmentation, according to Kamthania, Pawa, and Madhavan (2018). Their study stresses the value of visually expressing segmentation results and uses the K-mode clustering technique to categorize customers. This graphic method helps to create focused marketing plans for each group of customers and gives a greater understanding of consumer groups. Furthermore, Vanderveld et al. (2016) presented an engagement-based system to determine the client lifetime value in online commerce. The newly developed method takes into account consumer interactions and engagement levels in addition to transactional data. These indicators give firms a comprehensive understanding of the worth of their customers, enabling them to develop more precise and successful segmentation strategies.

Jha (2020) provides a thorough analysis of several segmentation approaches in a critical review of the customer segmentation techniques used in e-commerce. Jha offers helpful insights into the practical application of these strategies by analyzing their strengths and weaknesses, pointing organizations in the direction of the best course of action for their settings. Enhancing the current variants is a more favorable approach for developing a K-means-based clustering technique that is both robust and scalable (Ikotun et al.,2022) gave us insight into addressing K-means initialization problems and has taken priority over mixed data type issues.
This information will be an essential foundation for our project, which aims to assess and apply the best clustering model for customer segmentation in the e-commerce industry.

Data set

For this customer segmentation project, we have obtained a comprehensive dataset containing behavioral data, from Oct 2019 - April 2020. This dataset has been sourced from a prominent multi-category online store and comprises a multitude of events that capture interactions between users and products. Each event reflects a many-to-many relationship between products and users, providing a rich and intricate view of customer behavior over time. We are grateful to the Open CDP project for making this data accessible, and we utilized open-source customer data platform technologies to manage and analyze this valuable dataset securely and efficiently. The robustness of this dataset will serve as the cornerstone of our research, enabling us to derive meaningful insights and achieve our project's objectives effectively. The dataset consists of event data, where each row represents a specific user-product interaction event within the online store. The key attributes in the dataset include:
1. Event time: The timestamp of when the event occurred (in UTC).
2. Event type: Describing the type of event (e.g., view, cart, remove from cart, purchase) View: User viewed product, Cart: User added the product to the shopping cart, remove from the cart: User removed product from the shopping cart, Purchase: The user purchased a product.
3. Product ID: The unique identifier of the product.
4. Category ID: The category ID of the product.
5. Category code: The category taxonomy (code name) of the product, when applicable.
6. Brand: The brand name (down case) of the product, which may be missing.
7. Price: The price of the product.
8. User ID: The permanent user ID.
9. User session: A temporary user's session ID, changes with each session.
10. Event Types: The data set captures four primary event types:

Handling Missing Data: We found 5.14 percent of missing data mainly from attributes like category code, and brand. We replaced the null values with the mode, which is a reasonable strategy, especially for categorical variables because it helps

maintain the distribution of the existing data. The duplicate values have been dropped and removed from the data set.

Feature Engineering: The skewness was rectified by square rooting.Interquartile Range (IQR) has been used to drop the outliers in the 'Price' feature.
From the category code feature, we have extracted the main category and sub category indicating the product details.
We have extracted the date features like month, day, and year from the 'event time' category.
By considering the time from the 'event time', we created a new feature 'timing' indicating the user's session timing – morning, afternoon, evening, night, and midnight.

Statistical Summaries Visualization: The visualizations have been made using Power BI making them interactive dashboards. The User's session of the event analysis corresponding to the weekday and the top 5 brands have been shown below.
View event type Session:

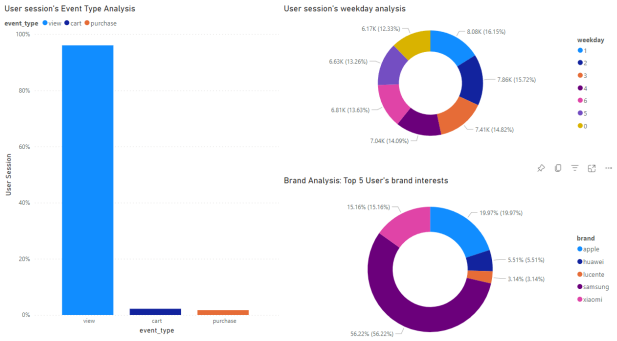

Fig. 1. User Session Analysis

The top 5 most explored main categories the customers with the top 5 user's brand interests:
Electronics:

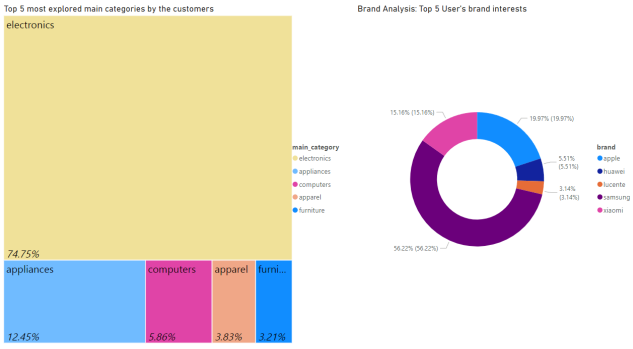

Fig. 2. Explored Categories by Brand

System Architecture

Selecting Modeling Approaches

Our study employs a multifaceted approach to customer segmentation, leveraging a suite of clustering algorithms each with its unique strengths and capabilities. The goal is to transcend traditional one-dimensional segmentation by harnessing the predictive power of machine learning and the nuanced discernment of deep learning techniques.

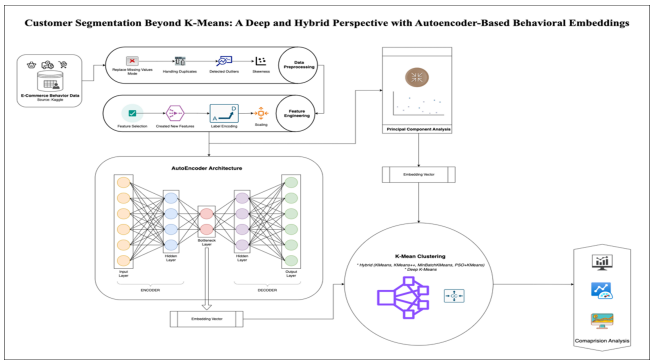

Fig. 3. A Deep and Hybrid perspective with Autoencoders based Behavioral Embeddings

K-Means Clustering: The initial phase of our analytical procedure involved utilizing the K-means algorithm, renowned for its straightforwardness and effectiveness in processing substantial datasets. K-means facilitated the partitioning of our dataset into discrete, non-overlapping segments by reducing the variance within each cluster. This stage established a preliminary clustering framework, enabling us to set a benchmark for evaluating more sophisticated clustering techniques.

Gaussian Mixture Model (GMM) To surmount the constraints of K-means, particularly its predilection for equidistant clusters, we applied the Gaussian Mixture Model (GMM). This probabilistic approach posits that the data is generated from multiple Gaussian distributions, each with unknown parameters. By employing the expectation-maximization technique, GMM adapts to clusters of various dimensions and configurations, thus providing a versatile and probabilistic approach to modeling the diverse customer data.

Deep K-Means Clustering Delving deeper into clustering sophistication, we explored Deep K-means, which marries the depth of neural network-based feature extraction with the traditional K-means clustering. This hybrid technique uses an autoencoder to distill the raw data into a more potent and condensed feature set. K-means is then applied to this transformed dataset, resulting in clusters that are more cohesive and reflective of complex customer behaviors that simpler models might miss.

Particle Swarm Optimization (PSO) with K-means We refined our clustering approach by integrating Particle Swarm Optimization (PSO) with K-means. PSO, which optimizes solutions through a process that mimics the social behavior of birds or fish, was instrumental in optimizing the initial centroid placement for K-means. This optimization is crucial as it enhances the sensitivity of K-means to its initial conditions, thereby improving segmentation quality.

Hybrid Model The culmination of our clustering analysis was the creation of an advanced hybrid model. This model synthesizes the core advantages of deep K-means, K-means++, min batch K-means, and PSO K-means. It employs

an ensemble-like majority voting mechanism for assigning cluster membership, thereby creating a robust and comprehensive segmentation solution that is greater than the sum of its parts.

Hypothesis

The hypothesis under investigation is whether deep clustering models demonstrate enhanced efficacy compared to traditional clustering techniques across diverse datasets. This hypothesis stems from the premise that deep clustering models, leveraging advanced neural network architectures, possess the capacity to autonomously learn intricate representations of data features in an unsupervised manner. It is theorized that this inherent capability enables these models to navigate complex, high-dimensional spaces more adeptly, resulting in more refined and accurate cluster assignments compared to conventional clustering approaches. This study aims to empirically evaluate and compare the performance of deep clustering models against traditional methods, providing valuable insights into their respective strengths and limitations in clustering diverse datasets.

Results

In our analysis, we applied multiple clustering techniques to segment customers based on their interaction data from an e-commerce platform. We utilized principal component analysis (PCA) to reduce the dimensionality of the dataset, which allowed us to visualize the clusters in two-dimensional space. The following clustering algorithms were used:

K-Means Clustering: K-Means clustering partitioned the data into three clusters. As seen in the corresponding scatter plot, the clusters are fairly well-separated, indicating a good initial segmentation.

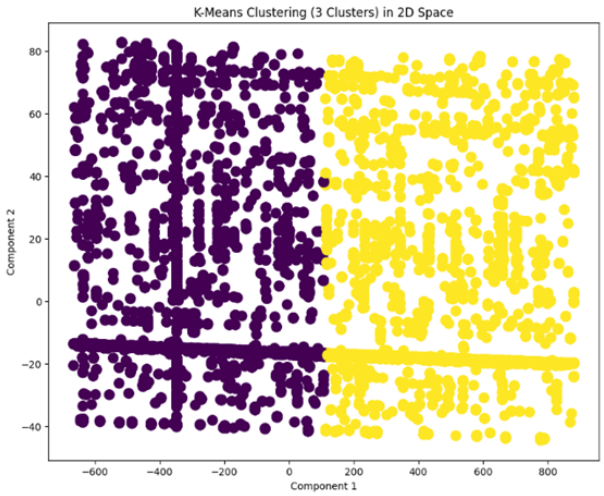

Fig. 4. K-means Clustering in 2D

Gaussian Mixture Model (GMM): Gaussian Mixture Model also identified three clusters. GMM assumes that the data is composed of a mixture of several Gaussian distributions, which can capture more complex cluster shapes than K-Means.

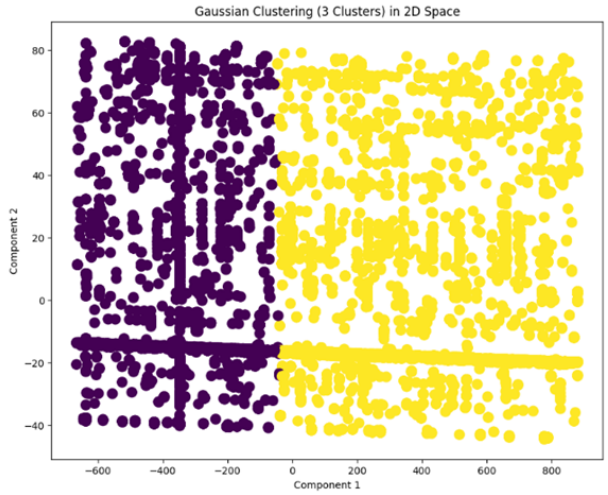

Fig. 5. Gaussian Mixture Clustering in 2D

Deep K-Means Clustering: And deep learning-based approach to K-Means clustering. The deep K-Means algorithm provided a more nuanced segmentation, identifying two distinct clusters, which suggests a different underlying pattern in the data not captured by the traditional methods.

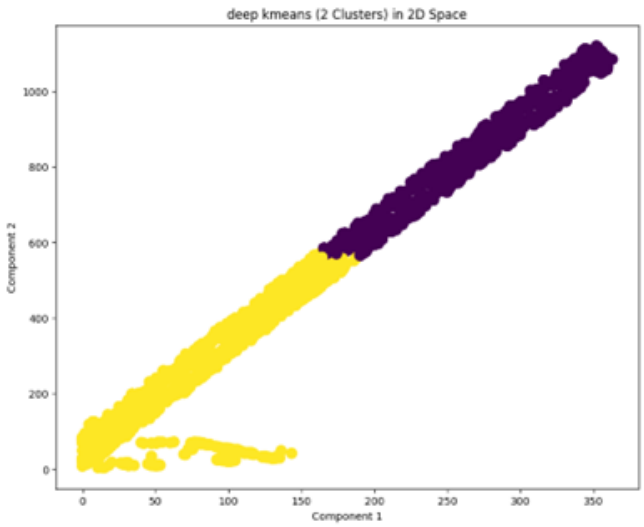

Fig. 6. Deep K-Means Clustering in 2D

Particle Swarm Optimization (PSO) with K-means: We refined our clustering approach by integrating Particle Swarm Optimization (PSO) with K-means. PSO, which optimizes solutions through a process that mimics the social behavior of birds or fish, was instrumental in optimizing the initial centroid placement for K-means. This optimization is crucial as it enhances the sensitivity of K-means to its initial conditions, thereby improving segmentation quality.

Hybrid Model: The culmination of our clustering analysis was the creation of an advanced hybrid model. This model synthesizes the core advantages of deep K-means, K-means++, min batch K-means, and PSO K-means. It employs

an ensemble-like majority voting mechanism for assigning cluster membership, thereby creating a robust and comprehensive segmentation solution that is greater than the sum of its parts.

Evaluation Techniques:

To evaluate the efficacy of our clustering techniques, we used the Silhouette Score and the Davies-Bouldin Index. The Silhouette Score helped determine the fit of objects within their own cluster compared to others, providing a clear measure of internal consistency and external separation. The Davies-Bouldin Index offered a complementary evaluation, focusing on the tightness and distinctness of the clusters, with lower scores indicating superior clustering.

The methodologies we've employed aim to distill the complex e-commerce data into actionable insights, carving out distinct customer segments that can inform more personalized marketing initiatives. The result is a nuanced view of consumer behavior that supports a targeted and efficient approach to e-commerce strategy.

For each clustering technique, we plotted the segmented groups in a two-dimensional space reduced by PCA. The plots illustrate how each method partitions the dataset, with distinct color-coding for each cluster:

The traditional K-Means algorithm resulted in three clusters with noticeable overlaps, particularly in the dense central region of the plot. This suggests some ambiguity in the boundaries that define each cluster.

The Gaussian Mixture Model yielded a similar three-cluster division, with a slightly different distribution that suggests a probabilistic overlap between the segments.

The deep K-Means method, however, showed a clear delineation between two clusters. This model's ability to reduce dimensionality and extract non-linear features likely contributed to its distinctive segmentation.

To quantitatively assess the performance of each clustering technique, we calculated silhouette scores, which measure how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters. The bar chart comparison of silhouette scores demonstrates that the deep K-Means and the Hybrid Model achieved the highest scores, suggesting they were most effective at segmenting the dataset with clear, distinct clusters

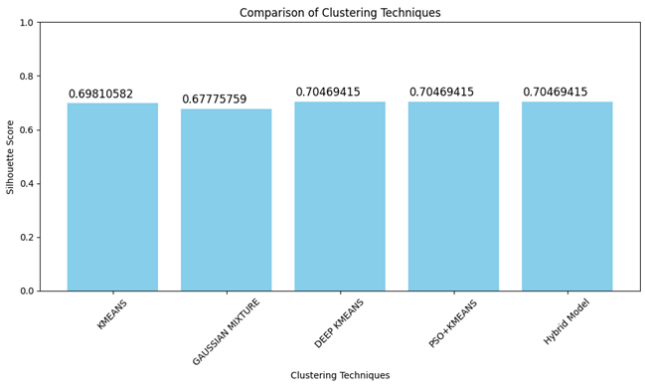

Fig. 7. Comparison of Clustering Techniques

The data analysis findings suggest that deep learning techniques, specifically the deep K-Means approach, can offer a more sophisticated customer segmentation that potentially unveils more intricate patterns within the customer behavior data. These advanced models can help e-commerce platforms to better understand their customer base, leading to more personalized marketing strategies and improved customer service.

Discussion

The weekly trends in session cart pricing further support the hypothesis that Cluster 1 customers are high spenders, particularly on weekends, indicating a possible correlation between leisure time and shopping activity. In contrast, Cluster 0's spending does not exhibit significant fluctuations throughout the week, suggesting a more consistent purchasing behavior.

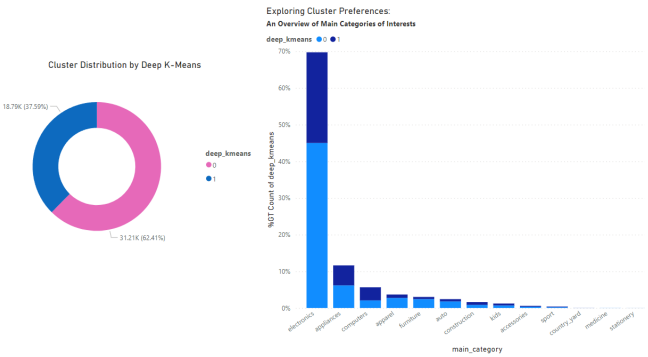

Fig. 8. Shows the Cluster Distribution along with their top product category preferences

Interpretation:

- Cluster Dominance: The larger size of Cluster '1' suggests that it represents the dominant segment within the dataset. This implies that a majority of the data points share a set of characteristics or behaviors that are distinct from those in the smaller Cluster '0'. We can say that Cluster '1' represent the most common customer profile.
- Category Preferences:The significant preference for 'electronics' in the bar chart indicates that this category is particularly important to the clusters. 'Electronics' is the most popular or sought-after category among the customer base. It might also suggest that 'electronics' has a wide appeal across different types of customers.

Targeted Strategies:

- Understanding that 'electronics' is a key category for both clusters, businesses can develop targeted marketing campaigns, tailor their product offerings, or create customized promotions to cater to this interest.
- The lower bars for categories like 'medicine', 'gourmet', and 'stationery' suggest these are niche areas. Businesses might need to either look for opportunities to grow these categories or prioritize them lower in strategic planning.

Average Price by Category: The average price by main category analysis indicated that Cluster 1 has a higher average spending across most categories compared to Cluster 0, with the most significant divergence in the 'stationery' category, where Cluster 1's average price was substantially higher. This suggests that Cluster 1 contains customers who may be less price-sensitive and potentially more quality-oriented, willing to spend more on certain products.

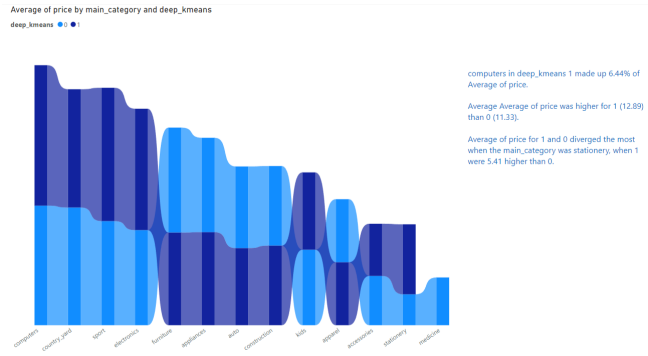

Fig. 9. Ribbon chart representing the average cart prices for the user sessions across clusters

Price Dynamics: We also examined price dynamics within each cluster, finding that Cluster 1 customers typically engage in higher-value transactions throughout the day, with peak average prices in the early morning. In contrast, Cluster 0 shows more consistent spending habits across different times of the day. This temporal pattern of spending could inform time-specific marketing strategies, such as flash sales or timed promotions to maximize revenue from each segment.

Fig. 10. Shows the Price Dynamics and Weekday user cart values across the clusters highlighting the top customer-interested category

Weekly Trends: The weekly trends in session cart pricing further support the hypothesis that Cluster 1 customers are high spenders, particularly on weekends, indicating a possible correlation between leisure time and shopping activity. In contrast, Cluster 0's spending does not exhibit significant fluctuations throughout the week, suggesting a more consistent purchasing behavior.

During the midweek, cluster 1, who are status buyers showed much interest in purchasing the electronics. Whereas, cluster 0 customers typically showed interest in buying the electronics during the weekends, when the companies must have come up with special offers and sale campaigns.

Fig. 11. Shows the purchase patterns of cluster 1 with their most interesting category trends.

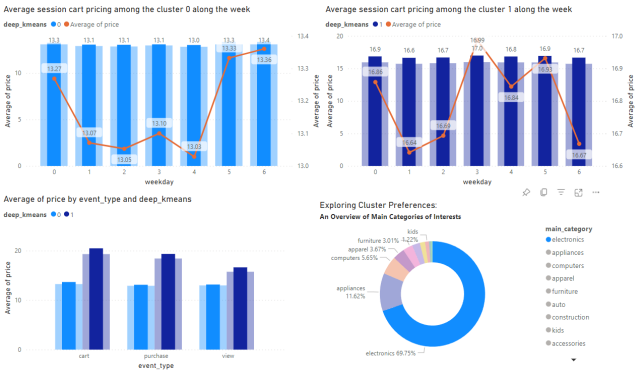

Fig. 12. Shows the electronic purchase patterns across the two clusters.

Insights into Price Variations by Event Type: The analysis of price variations by event type—such as adding to cart, purchase, and view—shows that Cluster 1 customers tend to add higher-priced items to their carts and proceed to purchase them, whereas Cluster 0 customers exhibit more price-conscious behavior, with lower average prices for viewed and added items.

When we did a deep analysis of the cluster 1 price dynamics across the various timings of the day regarding their user sessions, the cluster 1 customers exhibited higher purchase patterns during the night. Stats show that 80% of customers in cluster 1 were interested in purchasing electronics with peak purchases during the night.

Strategic Recommendations: Based on these insights, we propose the following strategies:

- Dynamic Pricing and Promotions: Implement dynamic pricing strategies for Cluster 1, taking advantage of their willingness to pay more, especially during early morning hours and weekends. For Cluster 0, consider value deals and discounts to stimulate purchasing.

- Personalized Engagement: Tailor marketing communications to the observed behavior patterns, with Cluster 1 receiving premium product promotions and Cluster 0 being targeted with cost-effective options.

- Inventory Management: Align stock levels with the high-demand categories and price points favored by each cluster, ensuring product availability that matches their spending habits.

- Temporal Marketing Tactics: Leverage the insights from time-based spending patterns to time marketing campaigns, such as early morning email blasts for Cluster 1 and consistent engagement throughout the week for Cluster 0.

- Customer Retention Strategies: Develop loyalty programs for Cluster 1 to reward their higher spending and create incentives for Cluster 0 to increase their transaction values.

Conclusion

Improving Customer Segmentation for Personalized Marketing in E-commerce," emphasizes the crucial relevance of consumer segmentation in the ever-changing e-commerce industry. We tackled the project issue by emphasizing the need to use data-driven insights to improve marketing strategy and consumer interaction activities. The study used a seven-month dataset from the Open CDP project, which was subjected to extensive preprocessing to ensure data quality and privacy compliance. The study attempted to discover unique customer segments based on behavior, preferences, and demographics by using a range of clustering algorithms and cluster selection approaches. A rigorous evaluation system, comprising metrics such as Inertia and Silhouette Score, was used to analyze segmentation quality, increasing the project's trustworthiness, and emphasizing the dataset's ethical use. E-commerce client segmentation improves personalization, increasing satisfaction, conversions, and enhanced profitability. Marketing strategies are optimized by using modern clustering and evaluation algorithms, which provide efficient resource allocation and exact audience targeting. This initiative establishes best practices in data science, fosters open data sharing, and acts as a significant resource for scholars working with real-world datasets.

References

1.Ballestar, M. T., Grau-Carles, P., Sainz, J. (2018). Customer segmentation in e-commerce: Applications to the cashback business model. Journal of Business Research, 88, 407-414.

2.He, X., Li, C. (2016, December). The research and application of customer segmentation on e-commerce websites. In 2016 6th International Conference on Digital Home (ICDH) (pp. 203-208). IEEE.

3.Punhani, R., Arora, V. S., Sabitha, S., Shukla, V. K. (2021, March). Application of clustering algorithm for effective customer segmentation in E-commerce. In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 149-154). IEEE.

4.Wu, R. S., Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications, 10(3), 331-341.

4.Tsai, C. F., Hu, Y. H., Lu, Y. H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. Expert Systems, 32(1), 65-76.

5.Akter, S., Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26, 173-194.

6.Schafer, J. B., Konstan, J. A., Riedl, J. (2001). E-commerce recommendation applications. Data mining and knowledge discovery, 5, 115-153.

7.Kamthania, D., Pawa, A., Madhavan, S. S. (2018). Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business. Journal of computing and information technology, 26(1), 57-68.

8.Vanderveld, A., Pandey, A., Han, A., Parekh, R. (2016, August). An engagement-based customer lifetime value system for e-commerce. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 293-302).

9.Jha, L. K. (2020, November). A Critical Review: Customer Segmentation Technique on E-Commerce. In e-Conference on Data Science and Intelligent Computing (p. 29).

10.Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., Heming, J. (2022). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences.