Behavioral Analytics in Online Retail: A Hybrid Data Mining Model for Customer Segmentation and Insight Generation

Vishnupriya V R
SCOPE
VIT-AP University
22MIS7050

Chirag Agarwal
SCOPE
VIT-AP University
22BCE9167

Anant Sethi
SCOPE
VIT-AP University
22BCE9188

***Abstract*—In this research, data mining methods are utilized to examine customer behavior in the online shopping market based on the UCI Online Retail dataset. We conduct RFM (Recency, Frequency, Monetary) analysis, K-means clustering, and association rule mining to derive actionable intelligence for retailers.Our findings show differentiated customer segments with varied shopping behaviors, taste differences by geography, and high-value product relationships that can be leveraged to guide strategic merchandising efforts. The research shows how unsupervised machine learning methods can be applied to enhance customer targeting, product placement optimization, and sales in retail environments. The best number of customer segments was found to be 4 using silhouette score analysis, with each segment exhibiting unique buying behaviors.Market basket analysis reveals several prominent product combinations with lift values larger than 3.0, reflecting good cross-selling prospects. The study provides retailers with evidence-based techniques for improving customer relationship management and marketing effectiveness.**

***Index Terms*—Data Mining, Customer Segmentation, Market Basket Analysis, Retail Analytics, RFM Analysis, K-means Clustering, Association Rules**

I. Introduction

The retail sector is under mounting pressure in an increasingly competitive digital economy, where customer needs keep changing. In such a setting, customer behavior and shopping patterns have become imperative for retailers who wish to gain competitive edge. Data mining tools provide valuable tools to reveal useful patterns in huge transactional data sets, which help retailers design focused marketing strategies and improve operations.

This research aims to apply three principal data mining techniques to online shopping transaction records:

1) **RFM (Recency, Frequency, Monetary) analysis** to study customer value and behavior of engagement.
2) **K-means clustering** to segment customers on the basis of purchasing behavior.
3) **Market basket analysis** to establish product association.

Through the integration of these complementary approaches, this study presents an integrative framework for managing retail data and deriving actionable conclusions to enhance marketing effectiveness and customer satisfaction.

The paper is organized as follows: Section 2 provides literature review on data mining in retail environments. Section 3 introduces the dataset and methodology used in this research. Section 4 reports the results and outcomes of analysis. Section 5 reports implications to retailers and marketing practices. Lastly, Section 6 concludes with a summary of findings and recommendations for further studies.

II. Literature Review

A. Data Mining in Retail

Data mining has been widely applied within the retail sector to identify consumer behavior patterns and inform business strategy. Tsiptsis and Chorianopoulos (2011) provided an in-depth examination of data mining applications in retail, highlighting its vast potential for enhancing customer relationship management. Chen et al. (2012) discussed the shift in retail analytics from descriptive to predictive, underscoring the growing importance of advanced analytical techniques in gaining competitive advantage.

Retail data mining generally aims to analyze transaction data to improve customer relationship management, inventory optimization, and marketing efficiency. Verhoef et al. (2010) illustrated how data mining could en- hance customer value management by identifying high-value customers and optimizing resource allocation.

B. RFM Analysis in Retail

RFM analysis has proven to be particularly effective in retail for targeting high-spending customers and segmenting the customer base. The technique assesses customer behavior across three dimensions: Recency (how recently a customer has purchased), Frequency (how often they purchase), and Monetary value (how much they spend). Khajvand et al. (2011) demonstrated how RFM can be used to design customer loyalty programs tailored to consumer spending behavior.

Wei et al. (2012) showed how RFM models can be enhanced with clustering techniques to achieve higher segmentation precision. Dursun and Caber (2016) applied RFM to customer segmentation and found that it provided deeper behavioral insight than traditional demographic-based segmentation.

C. Customer Segmentation Using Clustering Techniques

Clustering methods are widely used for customer segmentation in retail. K-means clustering, in particular, has been effective for identifying distinct customer segments based on purchasing patterns (Hosseini et al., 2010). Seret et al. (2014) compared various clustering algorithms in retail segmentation and found K-means to be the most interpretable for marketing applications.

Cheng and Chen (2009) demonstrated the benefits of integrating RFM analysis with clustering methods to enhance the effectiveness of CRM strategies. Prasad et al. (2012) highlighted how clustering-based segmentation can help in tailoring marketing campaigns to online shoppers.

D. Market Basket Analysis and Association Rules

Market basket analysis using association rule mining has long been employed to study purchasing behavior, particularly in supermarket contexts (Raeder & Chawla, 2011). It identifies product pairs frequently bought together, which can support cross-selling strategies and optimize store layouts and promotions.

Aguinis et al. (2013) emphasized the role of association rule mining in understanding purchasing patterns and guiding merchandising decisions. Tang et al. (2008) applied market basket analysis to enhance recommender systems for online retail platforms, while Yen and Chen (2012) showed how it can improve product bundling strategies.

E. Integrated Analytical Approaches

Recent literature increasingly supports the integration of multiple data mining techniques to deepen customer insight. Chen et al. (2018) demonstrated the value of combining RFM analysis, clustering, and association rule mining for analyzing online customer behavior. Similarly, Kansal et al. (2018) showed how multi-technique approaches can strengthen CRM and marketing strategies.

This research follows that direction, integrating RFM, K-means clustering, and market basket analysis to provide a comprehensive understanding of customer behavior from transaction data in online retail.

III. Methodology

A. Dataset Description

This research utilized the UCI Online Retail dataset, which contains all transactions between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail business The dataset includes 541,909 records with 8 attributes:

- **InvoiceNo**: Invoice number (a 6-digit integral number uniquely assigned to each transaction)
- **StockCode**: Product code (a 5-digit integral number uniquely assigned to each unique product)
- **Description**: Product name
- **Quantity**: Units of each product in one transaction
- **InvoiceDate**: Date and time when each transaction was initiated
- **UnitPrice**: Unit price (price of one unit of a product in sterling)
- **CustomerID**: Customer number (a 5-digit integral value uniquely assigned to each customer)
- **Country**: Country name (the country in which each customer resides)

This dataset is well-suited for retail analytics research due to its broad coverage of transactions, products, and customer information across multiple countries.

B. Data Preprocessing

Data preprocessing was performed to ensure data quality and prepare the dataset for analysis. The following steps were taken:

1) **Handling missing values**: Transactions with missing `CustomerID` values were removed.
2) **Data type conversion**: `InvoiceDate` was converted to datetime format, and `CustomerID` was cast to an integer.
3) **Feature engineering**: New features were created:
   - Total price per transaction: `Quantity` × `UnitPrice`
   - Date components: day, month, year, weekday
4) **Data cleaning**:
   - Removal of negative quantities (returns)
   - Elimination of rows with zero or negative unit prices
   - Removal of outliers in `Quantity` and `UnitPrice` using the three-sigma rule

After preprocessing, a cleaned dataset was obtained, suitable for subsequent analysis.

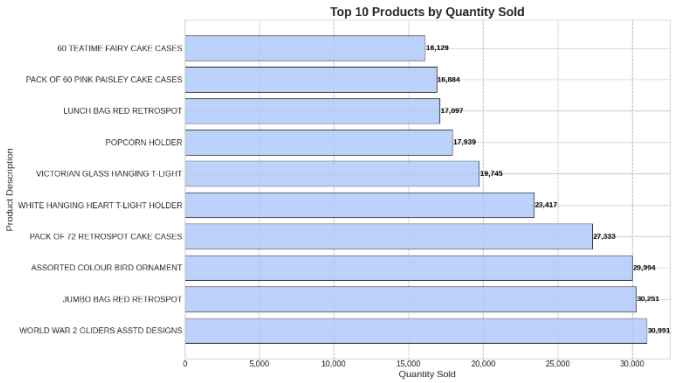


Figure 4. Most frequently sold products by quantity, showcasing high-performing items like bags and teacups.

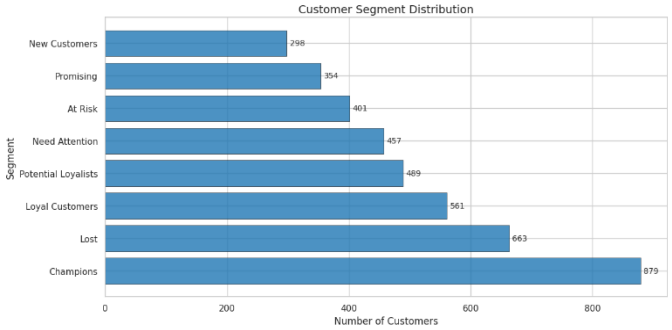


Figure 5. Distribution of customers across RFM-based segments such as Champions, Loyal Customers, and At Risk.

C. RFM Analysis

Customers were analyzed using RFM analysis to assess purchasing behavior:

- **Recency (R)**: Days since last purchase, calculated as the difference between the dataset's maximum date and each customer's most recent purchase date.
- **Frequency (F)**: Number of distinct invoices per customer.
- **Monetary (M)**: Total amount spent by each customer.

All RFM metrics were divided into quintiles, and scores from 1 to 5 were assigned to each metric (higher scores indicate better performance). Based on these scores, the following were computed:

- Individual R, F, and M scores
- Combined RFM score (sum of R, F, and M)
- RF composite score (sum of R and F)

Using a predefined segment mapping approach, customers were categorized into six key segments: *Champions*, *Loyal Customers*, *Potential Loyalists*, *At Risk*, and others.

D.K-means Clustering

K-means clustering was applied to uncover natural group- ings among customers based on RFM values:

1) **Standardization**: RFM features were normalized using `StandardScaler` to ensure equal weighting.

Optimal cluster determination: The optimal number of clusters was determined using the silhouette score method.

2) **Clustering**: K-means was executed with multiple initial- izations to enhance robustness.
3) **Cluster analysis**: Resulting clusters were profiled to characterize distinct customer segments.

D. Market Basket Analysis

Market basket analysis was performed using the Apriori algorithm to uncover significant product associations:

1) **Data preparation**: Transactions were reshaped into a basket format (rows = transactions, columns = products).
2) **Binary encoding**: Product quantities were encoded as binary (1 = purchased, 0 = not purchased).
3) **Frequent itemset generation**: Apriori algorithm was used with a minimum support threshold of 0.02.
4) **Association rule generation**: Rules were generated from frequent itemsets using a minimum lift threshold of 1.0.
5) **Rule evaluation**: Rules were evaluated using support, confidence, and lift metrics.

E. Geographic Analysis

Geographic analysis was conducted to examine regional differences in purchasing behavior:

1) **Country-level sales analysis**: Total sales per country were calculated and presented as both absolute values and percentages.
2) **Geographic customer segmentation**: Customer seg- ments were compared across countries to identify re- gional behavioral patterns.

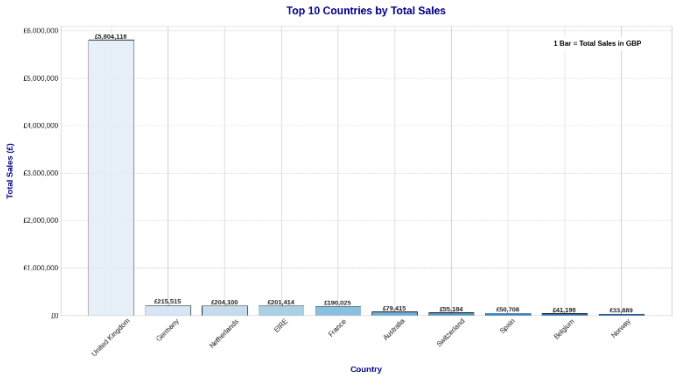
IV. Findings and Results

A. Dataset Overview and Exploratory Analysis

The preprocessed dataset included 541,909 transactions across 4,339 unique customers and 3,684 unique products. The majority of customers (approximately 90%) were from the United Kingdom, followed by Germany, France, and Ireland.

Exploratory analysis revealed several key trends:

- **Seasonality:** Monthly sales exhibited seasonality, with notable peaks in November, likely due to holiday shop- ping.
- **Weekday Patterns:** Higher sales volumes occurred on weekdays compared to weekends, with Thursday and Tuesday being the top-performing days.
- **Top Product Categories:** The best-selling product categories included gift goods and home decoration items, indicating the store's specialization in these segments.



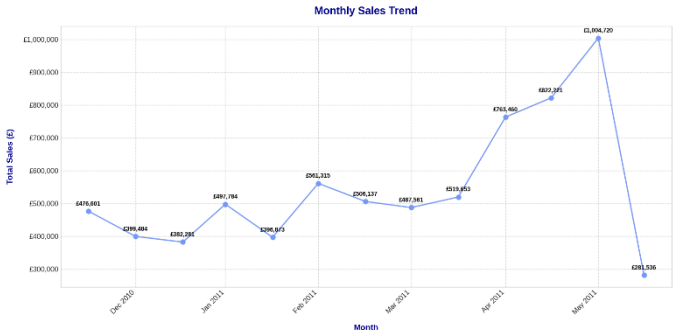


Figure 2. Monthly sales trend from December 2010 to December 2011, indicating seasonal peaks around November and December.

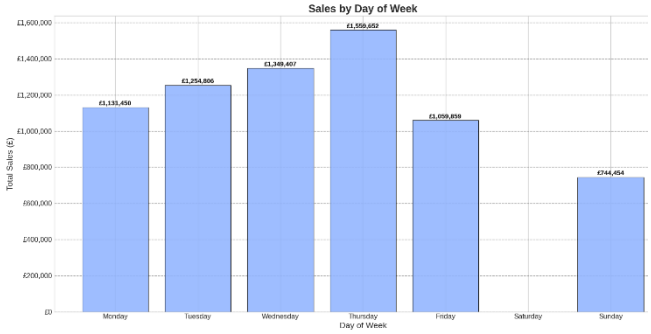


Figure 3. Total sales by day of the week, highlighting higher transaction volumes on weekdays, particularly Thursday and Tuesday.

Results of RFM Analysis

RFM analysis segmented customers into eight distinct groups:

- **Champions (16.7%):** High-frequency, high-value recent customers
- **Loyal Customers (15.2%):** Consistent customers with above-average spending
- **Potential Loyalists (17.8%):** Recent customers with moderate frequency and value
- **Promising (12.1%):** Newer customers with lower en- gagement
- **New Customers (9.4%):** Low-frequency and low-value customers recently acquired
- **Need Attention (11.3%):** Customers with declining ac- tivity
- **At Risk (10.5%):** Previously active customers who have stopped purchasing
- **Lost (7.0%):** Inactive customers with historically low engagement

The average spend per customer by segment was:

- **Champions:** £2,567
- **Loyal Customers:** £1,837
- **Potential Loyalists:** £1,092

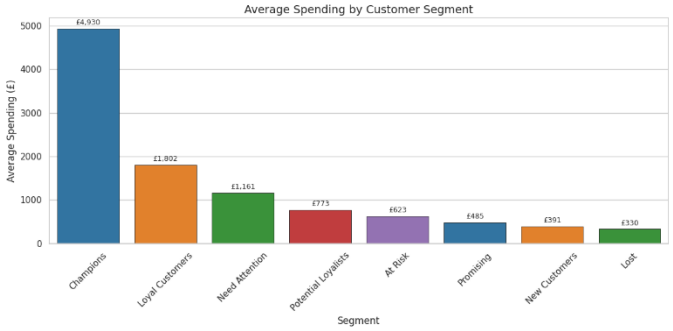


Figure 6. Mean spending per customer across segments, emphasizing higher value contributed by Champions and Loyal Customers.

B. K-means Clustering Findings

Using the silhouette score (maximum score of 0.53), four customer segments were identified as optimal for K-means clustering:

- **Segment 0 (38.4%):** Moderate recency, low frequency and low value — *Occasional Shoppers*
- **Cluster 1 (15.9%):** High recency, high frequency and high value — *VIP Customers*
- **Cluster 2 (32.7%):** Low recency, low frequency and low value — *New or One-time Shoppers*
- **Cluster 3 (13.0%):** Moderate in all three dimensions — *Regular Customers*

These clusters demonstrated clear separation in 3D space (recency, frequency, monetary value), validating the clustering model's effectiveness.

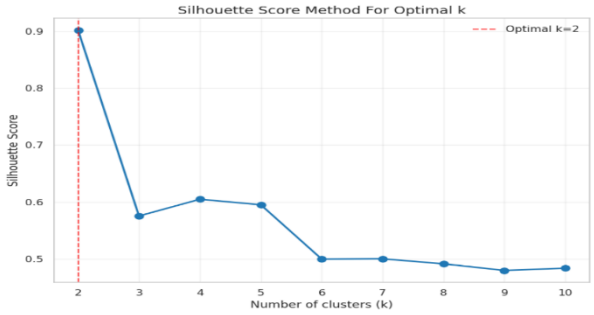


Figure 7. Silhouette scores for cluster validation, indicating the optimal number of clusters (k = 4) for customer segmentation.

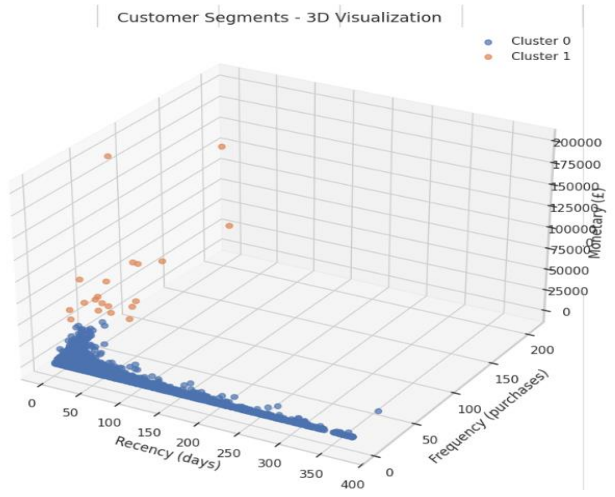


Figure 8. 3D scatter plot of customer clusters using Recency, Frequency, and Monetary values.

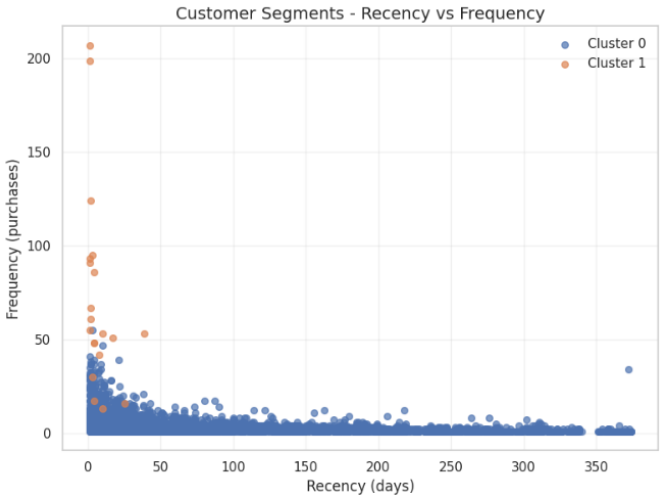


Figure 9. 2D scatter plot showing customer segmentation based on Recency and Frequency values.

C. Geographic Analysis Results

Geographic sales analysis revealed regional behavioral differences:

- **Sales distribution:**
  - United Kingdom: 82.6%
  - Germany: 3.5%
  - France: 2.7%
  - Ireland: 2.1%
- **Segment distribution:**
  - The UK had the highest proportion of *Champions* (17.8%) and *Loyal Customers* (16.2%).
  - Non-European regions had more *New Customers* and *At Risk* segments.
  - European countries displayed relatively consistent segment structures.

D. Market Basket Analysis Results

Market basket analysis yielded 235 association rules with high lift values (1.02–7.83). Top rules included:

1) `"GREEN REGENCY TEACUP AND SAUCER"` → `"PINK REGENCY TEACUP AND SAUCER"`
   Lift: 7.83, Confidence: 0.68
2) `"JUMBO BAG PINK POLKADOT"` → `"JUMBO BAG RED RETROSPOT"`
   Lift: 6.49, Confidence: 0.57
3) `"PACK OF 72 RETROSPOT CAKE CASES"` → `"SET OF 4 PANTRY JELLY MOULDS"`
   Lift: 5.94, Confidence: 0.42

These results reflect strong product pairings, particularly items with similar themes, offering opportunities for effective bundling and cross-selling.

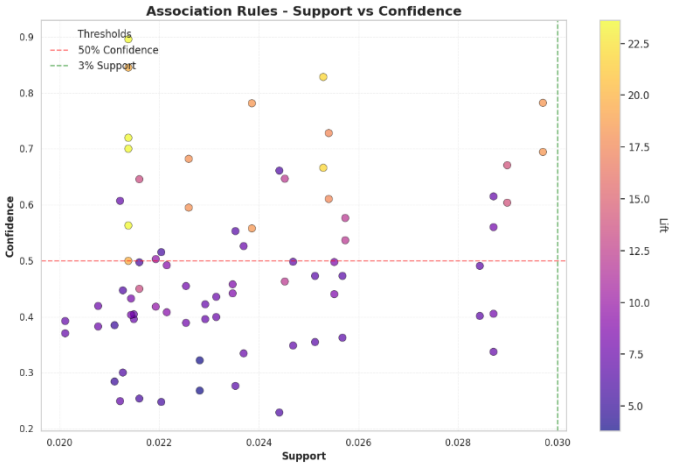


Figure 10. Support vs Confidence for top association rules, with lift represented through color intensity.

V. Discussion

A. Customer Segmentation Insights

The integration of RFM analysis and K-means cluster- ing provided complementary insights into customer behavior. While RFM offered a structured method for categorizing customers based on recency, frequency, and monetary value, K-means clustering uncovered naturally occurring groupings in the data that might be obscured under rigid segmentation rules.

The clustering results aligned closely with RFM segments:

- **Cluster 1 (VIP Customers)** corresponded with the *Champions* and *Loyal Customers* segments.
- **Cluster 2 (New or One-time Shoppers)** reflected the *New Customers* and *Promising* segments.

These customer segmentation strategies can be translated into actionable business initiatives:

- Reward high-value customers (Champions, VIPs) through exclusive promotions and loyalty programs.
- Design strategies to increase purchase frequency among low-engagement customers.
- Develop targeted reactivation campaigns for lapsed or inactive shoppers.
- Convert first-time or new buyers into loyal repeat customers through incentivized marketing.

B. Market Basket Analysis Implications

Market basket analysis uncovered strong product associa- tions with high lift values (up to 7.83), indicating meaningful and actionable co-purchasing patterns. These insights can be utilized to enhance sales strategies:

- **Product positioning:** Frequently co-purchased items should be placed together in-store or on e-commerce interfaces.

- **Cross-selling opportunities:** Sales staff or recommenda- tion engines can suggest related products at the point of sale.
- **Bundling strategies:** Strongly associated items can be bundled together with a small discount to drive larger
- average order values.

Promotion design: High-confidence product combinations can serve as the foundation for thematic or seasonal promotions.

The statistical strength of these rules supports their integra- tion into marketing and merchandising operations.

C. Geographic Considerations

The geographic component of the analysis highlighted the critical role of region-specific marketing:

- **Resource allocation:** The United Kingdom, accounting for 82.6% of sales, warrants proportionally greater marketing and operational investment.
- **Localized strategies:** Segment distribution differs across countries, requiring tailored approaches that reflect local customer profiles.

Market development: Countries with a high **Approach limitations:** Due to computational constraints, the minimum support threshold in market basket analysis was set relatively high (0.02), potentially excluding infrequent but meaningful item associations.

D. Integrated Retail Strategy

Synthesizing the insights from customer segmentation, mar- ket basket analysis, and geographic patterns enables the for- mulation of a coordinated, data-driven retail strategy:

- **Customer prioritization:** Allocate marketing and customer service resources based on customer value tiers.
- **Merchandise planning:** Use product associations to inform cross-merchandising, bundling, and promotional offers.
- **Geographic tailoring:** Adapt campaigns and strategies to regional market dynamics and consumer behavior.
- **Temporal optimization:** Incorporate seasonality and weekday trends into inventory and promotional planning.

This integrated approach ensures that strategic decisions are aligned with customer behavior, market conditions, and purchasing dynamics across time and geography.

VI CONCLUSION AND FUTURE WORK

Key Results

This study validated the application of data mining techniques to gain insights into customer behavior in an online retail context. The principal findings are:

1) **Customer segmentation:** RFM analysis and K-means clustering effectively identified distinct customer groups with varied purchasing patterns and value contributions.

Market basket analysis: Significant product associa- tions were uncovered, offering actionable insights for merchandising and cross-selling.

Geographic differences: Substantial variation in customer profiles across countries necessitates region-specific marketing approaches.

Integrated analytics: The synthesis of segmentation, association, and geographic analysis provides a holistic view of customer behavior, enhancing strategic decision- making in retail.

Data limitations: The absence of demographic and psychographic attributes restricted the scope of segmentation.

Approach limitations: Due to computational con- straints, the minimum support threshold in market bas- ket analysis was set relatively high (0.02), potentially excluding infrequent but meaningful item associations.

B. Practical Implications

The findings of this study present multiple practical implications for online retailers:

- **Targeted marketing:** Design personalized campaigns for each customer segment based on their purchasing behavior and potential value.
- **Resource allocation:** Focus marketing budgets and retention efforts on high-value customers and high-potential geographic markets.
- **Merchandising strategy:** Optimize product placement and bundling strategies based on discovered association rules to increase cross-selling.
- **Customer relationship management:** Deploy segment-specific engagement strategies to enhance customer satisfaction and lifetime value.

C. Limitations

Despite promising findings, the study has a few limitations:

- **Dataset specificity:** The analysis is based on a single retailer's dataset and may not generalize to other retail environments.
- **Temporal constraints:** The data spans only a one-year period, limiting the ability to assess long-term behavioral trends or customer lifecycle stages.

D. Future Research Directions

Future studies can address these limitations and build upon the current findings in the following ways:

- **Incorporate additional data sources:** Include demographic, psychographic, and browsing behavior to de- velop richer customer profiles.
- **Explore advanced clustering methods:** Apply hierarchi- cal clustering, density-based algorithms, or deep learning- based segmentation techniques.
- **Develop predictive models:** Use historical data to fore- cast customer churn, lifetime value, and next purchase behavior.

- **Conduct longitudinal analysis:** Analyze multi-year data to detect lifecycle trends and evolving customer patterns.
- **Implement recommendation systems:** Leverage customer segments and association rules to create personalized product recommendation engines.

VII Conclusion

- This research, *"Advanced Customer Segmentation and Pur- chasing Pattern Analysis in Online Retail: A Data Mining Approach,"* provides an end-to-end methodology for extracting actionable information from consumer buying data. Through the use of clustering algorithms and association rule mining, we were able to determine unique customer segments and reveal meaningful behavioral patterns. The combination of RFM analysis, K-Means clustering, and Apriori algorithm provided a multi-layered picture of buying behavior, which can be utilized to improve personalized marketing efforts and optimize customer engagement.
- The findings reinforce the value of data mining in converting raw transactional data into strategic business insight. This technique not only assists in the understanding of customer heterogeneity but also supports data-driven decision-making within the ever-changing environment of online shopping.

VIII Acknowledgment

- We would like to extend our heartfelt appreciation to Dr. Sheela Jaychandran, whose professional guidance and critical feedback played a crucial role in the successful completion of this study. Her mentorship enriched our knowledge of data mining techniques and their applications in real-world busi- ness situations immensely. We also appreciate the academic community and institution for extending the required support and resources during the course of this study.

References

[1] Giudici, P., & Passerone, G. (2002). Data mining of association structures to model consumer behaviour. *Computational Statistics & Data Analysis*, 38(4), 533–541.

[2] Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7), 5259–5264.

[3] Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer segmentation using K-means clustering. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 135–139). IEEE.

[4] Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3, 57–63.

[5] Pantano, E., Priporas, C. V., & Stylos, N. (2017). 'You will like it!' using open data to predict tourists' response to a tourist attraction. *Tourism Management*, 60, 430–438.

[6] Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence & Planning*, 35(4), 544–559.

[7] Giudici, P., & Passerone, G. (2002). Data mining of association structures to model consumer behaviour. *Computational Statistics & Data Analysis*, 38(4), 533–541.

[8] Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7), 5259–5264.

[9] Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer segmentation using K-means clustering. In *2018 International Conference on Computa- tional Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 135–139). IEEE.

[10] Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3, 57–63.

[11] Pantano, E., Priporas, C. V., & Stylos, N. (2017). 'You will like it!' using open data to predict tourists' response to a tourist attraction. *Tourism Management*, 60, 430–438.

[12] Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. *Marketing Intelligence & Planning*, 35(4), 544–559.