**DATA MINING AND MARKETING ANALYTICS PROJECT REPORT**

**Abhishek Gundala and Shravani Surampalli**

**MSBA – BDA - 620**

**Introduction**

Understanding customer value is crucial for e-commerce success. This study focuses on analyzing purchasing behaviors and preferences to uncover what drives customer value in online shopping. The relevance of this research lies in its potential to enhance marketing strategies, customer engagement, and ultimately, business profitability. By applying data mining techniques to e-commerce transactions, we aim to provide actionable insights for marketers to better cater to their customers' needs in a dynamic digital marketplace.

**Relevance in Marketing Analytics**

This research is vital for marketing analytics as it directly informs strategies to enhance customer value in e-commerce. Insights into purchasing behaviors allow for more effective customer segmentation and targeted marketing, leading to improved customer experiences and loyalty. By anticipating market trends and consumer preferences, businesses can adjust their strategies in real-time, ensuring competitiveness and relevance in the rapidly changing digital marketplace.

**Literature Review**

The shift to e-commerce has spurred significant research into online consumer behavior. Key themes include the impact of web design on purchasing decisions, the effectiveness of personalized marketing in enhancing customer experience, and the pivotal role of customer reviews in establishing trust. Additionally, predictive analytics is recognized for its power in forecasting consumer behavior, while customer lifetime value (CLV) analysis remains crucial for strategic planning. The integration of AI and machine learning is also noted for revolutionizing customer engagement. This project builds on these insights, focusing on the analysis of purchasing behaviors to understand e-commerce customer value.

**Data, Data Sources, and Data Characteristics**

Source of Data: This study's dataset was sourced from academic contributors at Mercer University. The dataset for this analysis originates from an e-commerce platform's transaction logs. It encompasses detailed records of customer purchases, providing a rich resource for understanding consumer behavior in online shopping.

Scope of Data: The dataset comprises 107,159 transactions, each capturing key details of customer orders.

**Data Attributes:**

| Variable | Description | Data Type |
| --- | --- | --- |
| event\_time | Timestamp when the transaction occurred (UTC). | String |
| order\_id | Unique identifier for the order. | Float |
| product\_id | Unique identifier for the product. | Float |
| category\_id | Unique identifier for the product category. | Float |
| category\_code | Classification of the product (e.g., 'electronics.tablet'). | String |
| brand | Brand name of the product. | String |
| price | Price of the product at the time of purchase. | Float |
| user\_id | Unique identifier for the user making the purchase. | Float |
| column\_name | Additional column (requires further investigation). | Integer |

Data Integrity: The dataset includes a mix of numerical and categorical data. Notably, some fields, such as category code and brand, have missing values, which will be addressed in the data preprocessing phase.

Data Use: This dataset forms the analytical base for our project, allowing us to delve into e-commerce customer purchasing behaviors and preferences to derive insights into customer value.

**Clean and Preprocess the Data**

In this phase, the dataset underwent several preprocessing steps at several stages of the project to ensure data quality and prepare it for analysis. The steps included:

Handling Missing Values: Identified and imputed or removed missing values where appropriate.

Data Type Conversion: Ensured that all variables were of the correct data type, converting strings to categorical variables where necessary.

Outlier Detection: Analyzed the data for outliers and anomalies to either correct or remove these values to prevent skewing the results.

Feature Engineering: Derived new variables that could potentially enhance the model’s predictive power.

**Methodology**

**Market Basket Analysis (MBA)**: This technique was utilized to understand the purchasing behavior of customers by identifying associations and patterns between different products purchased together. MBA is instrumental in developing cross-selling strategies and enhancing product placement.

**K-Means Clustering**: K-Means was selected for its efficiency in segmenting the customer base into distinct groups based on purchasing patterns and preferences. This segmentation helps in targeting marketing efforts and personalizing the customer experience.

**RFM Analysis**: Recency, Frequency, and Monetary (RFM) analysis was conducted to value and prioritize customers based on their purchasing history. It is a key technique for identifying potential loyal customers and optimizing marketing campaigns to increase customer lifetime value.

**Empirical Results**

**RFM Analysis**: The analysis revealed a skewed customer distribution across RFM segments, with a significant majority in specific segments suggesting a pattern of engagement and purchasing behavior.

**K-Means Clustering**: Customers were segmented into four clusters based on spending patterns. The identified clusters range from high-value, high-spending customers to low-value, low-spending ones, indicating clear distinctions in customer value and engagement.

**Conclusions and Recommendations**

### ****K-Means Clusters****:

### Cluster 1: Low Average, Varied Total

* Conclusion: A diverse group, likely a mix of one-off and long-term customers making small purchases.
* Recommendation: Introduce loyalty programs and small-scale marketing campaigns. Referral incentives could expand our customer base using this group's network.

Cluster 2: Moderate Average, Lower Total

* Conclusion: Possibly newer customers with occasional, moderately priced purchases.
* Recommendation: Increase purchase frequency through targeted promotions and by highlighting mid-range products to boost their spend.

Cluster 3: High Average, High Total

* Conclusion: Loyal customers with consistent and significant investment in our products.
* Recommendation: Retain through premium service, personalized offers, and by seeking regular feedback to meet their high expectations.

Cluster 4: Very High Average, Varied Total

* Conclusion: Customers making rare but substantial purchases, likely driven by specific needs or events.
* Recommendation: Create an exclusive VIP experience and personal outreach to enhance satisfaction and encourage repeat high-value sales.

**Conclusion from RFM Analysis**

The distribution shows a common pattern in customer behavior: a large number of customers make purchases infrequently and spend, while a smaller number of customers are highly engaged with frequent purchases and higher spending. The segments with low R, F, and M scores, indicating that most customers have not made recent purchases, don't buy frequently, or don't spend much.

### Recommendations Based on RFM Analysis

* For segments with high frequency and monetary values but low recency: Re-engage these customers with targeted communication reminding them of the value of your products or offering them incentives to return.
* For segments with high recency but lower frequency and monetary values: Encourage repeat purchases and higher spending through cross-selling and upselling strategies. Personalized recommendations based on their recent purchases could be effective.
* For the large segments with lower values across R, F, and M: These customers may need a reactivation strategy or may be targeted less frequently due to their lower engagement and value. Tailored promotions or special events could be used to try to increase their engagement.
* For smaller, high-value segments: Focus on retention strategies, such as loyalty programs, exclusive offers, and personalized services. These customers are valuable and likely contribute a significant portion of revenue.

**References**

Jain, D., & Singh, S. S. (2002). Customer lifetime value research in marketing: A review and future directions. Journal of Interactive Marketing, 16(2), 34-46. <https://doi.org/10.1002/dir.10032>

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University - Computer and Information Sciences, 34(5), 1785-1792. <https://doi.org/10.1016/j.jksuci.2019.12.011>

Sohrabi, B., & Khanlari, A. (2007). Customer Lifetime Value (CLV) Measurement Based on RFM Model. Iranian Accounting & Auditing Review, 14(47), 7-20.

Shih, Y. Y., & Liu, C. Y. (2003). A method for customer lifetime value ranking — Combining the analytic hierarchy process and clustering analysis. Journal of Database Marketing & Customer Strategy Management, 11, 159-172. <https://doi.org/10.1057/palgrave.dbm.3240216>

**Appendix**

**EDA**

**Top 30 categories**

**A graph with orange and white text

Description automatically generated**

**Top 30 Brands**

**A graph with many numbers

Description automatically generated with medium confidence**

**Distribution of Price**

**A graph of a log distribution of a log distribution

Description automatically generated with medium confidence**

**Time SeriesA graph of a graph showing a number of blue lines

Description automatically generated with medium confidence**

**Cluster Analysis**

**A graph with many dots

Description automatically generatedA graph with red green and blue dots

Description automatically generated**

**RFM Analysis**

**A graph of a customer distribution

Description automatically generated**

**Market Basket Analysis**

**A screenshot of a data

Description automatically generated**

**A close up of text

Description automatically generated**