Ruslan Maystrenko

Customer Segmentation and Market Basket Analysis

In the business landscape, understanding customer behavior and preferences is crucial for driving growth and profitability. This report presents an in-depth focus on customer segmentation and market basket mix analysis using the Online Retail dataset. Customer segmentation involves categorizing customers into distinct groups based on shared characteristics, enabling businesses to tailor marketing strategies and enhance customer experiences. Market basket analysis, conversely, identifies patterns in product purchases, revealing associations that inform product bundling and cross-selling opportunities. These analytical approaches empower businesses to optimize resource allocation, increase customer retention, and boost revenue.

Customer segmentation provides a foundation for personalized marketing by identifying high value customers, understanding their needs, and targeting them with relevant offers. This can lead to improved customer satisfaction, loyalty, and lifetime value. Market basket analysis complements this by uncovering purchasing trends, allowing businesses to strategically position products and enhance sales through informed decision making. Together, these methods offer a comprehensive view of customer and product dynamics, critical for maintaining a competitive edge.

Approaches and Methodology

Various techniques were considered for this analysis. For customer segmentation, options included hierarchical clustering, DBSCAN, and K-means clustering. K-means clustering was selected due to its efficiency with large datasets and its ability to produce clear, actionable segments. The Elbow Method and Silhouette Scores confirmed three clusters as the optimal balance of compactness and interpretability, avoiding overfitting while ensuring practical business insights.

For market basket analysis, both the Apriori and FP-growth algorithms were assessed for their effectiveness in identifying frequent itemsets and generating association rules. Although the performance difference between the two was minimal, FP-growth was chosen for its superior efficiency in processing large transactional datasets. This selection

enabled a robust analysis, suited to handle the dataset's volume, ensuring optimal performance in a retail context.

Dataset Overview

The Online Retail dataset, sourced from the UCI Machine Learning Repository, comprises 541,909 transactions from an e-commerce store, spanning December 2010 to December 2011. It includes original attributes such as:

- *InvoiceNo*, *StockCode*, *Description*, *Quantity*, *InvoiceDate*, *UnitPrice*, *CustomerID*, and *Country*

Feature engineering enriched the dataset with additional columns including:

- *Revenue* (UnitPrice multiplied by Quantity), *Year*, *Month*, *Day*, *Hour*, *Minute*, *DayOfWeek*, *IsWeekend* (indicating weekend purchases), and *IsBusiness* (identifying customers with annual revenue exceeding $30,000)

Data Wrangling

The data preparation process cleaned the dataset of inefficiencies to prepare for segmentation and association rules. Transactions with negative quantities or unit prices, indicative of returns or credits, were removed. Null values in the CustomerID field, totaling 132,220, were excluded to avoid skewing customer based analyses. The Description column was standardized by removing periods, converting to uppercase, and stripping excess spaces. Non-product transactions, such as postage and manual entries, were filtered out, reducing the dataset to 396,273 purchase transactions for analysis.

Exploratory Data Analysis Insights

Exploratory Data Analysis (EDA) provided key insights into customer and product trends:

- The United Kingdom dominates both revenue and customer count, contributing significantly more than other countries, with notable revenue spikes in October and November, likely tied to holiday shopping
- Business customers, identified by annual spending above $30,000, exhibit bulk purchasing behavior, particularly in December and January, possibly driven by seasonal demand or tax benefits

- Non-business customers show more evenly distributed spending patterns
- Product analysis highlighted "PAPER CRAFT, LITTLE BIRDIE" as a top performer in revenue and quantity sold, underscoring its popularity
- Conversely, "REGENCY CAKESTAND 3 TIER" generates high revenue with fewer units, indicating a premium price point, while "WORLD WAR 2 GLIDERS ASSTD DESIGNS" reflects high volume at a lower price

Customer Segmentation

Using K-means clustering on standardized RFM (Recency, Frequency, Monetary) metrics, three distinct customer segments emerged:

<u>Mid-Value, Active Customers (Segment 0)</u>

- *Count*: 3,226 customers
- *Characteristics*:
    - Moderate recency average of 41.42 days
    - Frequency average of 4.65 transactions
    - Monetary value average of $1,839.10
    - These customers are active and engaged, forming a stable base
- *Marketing Strategies*: Promote higher value products via targeted email discounts, implement loyalty programs to boost frequency, and leverage holiday campaigns with popular items to maximize sales.

<u>Low-Value, Inactive Customers (Segment 1)</u>

- *Count*: 1,084 customers
- *Characteristics*:
    - High recency average of 247.34 days
    - Low frequency average of 1.57 transactions
    - Low monetary value average of $625.97
    - Indicates disengagement

Marketing Strategies: Deploy personalized reactivation offers via email, provide low cost incentives through retargeting ads, and collect feedback with small incentives to address churn.

High-Value, Loyal Customers (Segment 2)

- *Count*: 25 customers
- *Characteristics*:
    - Low recency average of 6.12 days
    - High frequency average of 67.88 transactions
    - High monetary value average of $85,941.88
    - This small, loyal group drives disproportionate revenue
- *Marketing Strategies*: Offer premium loyalty programs with exclusive benefits, provide personalized recommendations via email, and ensure dedicated support for seasonal bulk orders.

Market Basket Mix Analysis

FP-growth analysis identified frequent itemsets and association rules within each segment, revealing purchasing patterns:

Segments 0 and 1 (Mid-Value and Inactive)

- Strong associations exist among "REGENCY TEACUP AND SAUCER" variants (such as Pink, Green, Roses), with high confidence (from 0.75 to 0.88) and lift (from 18.81 to 25.01)
- This unified trend suggests coordinated teacup set purchases, ideal for broad cross-selling campaigns targeting these segments

Segment 2 (High-Value)

- Unique niche rules include associations like "HERB MARKER ROSEMARY" to "HERB MARKER PARSLEY" (confidence 0.94, lift 94.28)
- These reflect curated, high value purchasing, offering opportunities for targeted bundling and premium promotions for niche specified products

Conclusion

This analysis underscores the power of data driven insights in enhancing business strategies. Customer segmentation identified three actionable groups, enabling tailored marketing to boost engagement, reactivate dormant customers, and retain high-value loyalists. Market basket analysis revealed segment specific product affinities, supporting

cross-selling for mid-value and inactive customers and niche targeting for high-value clients. These findings provide a roadmap for personalized campaigns, optimized product placement, and improved retention, driving revenue growth and competitive advantage.

Next steps include launching segment specific campaigns, monitoring key performance indicators (revenue, retention rate, order value), and exploring real-time analytics for ongoing optimization.