M Ihtesham Akram Awan (01-135212-035)

Syed Muhammad Ahmed Us S (01-135211-081)

Degree: BS(IT)7A

Report: Retail Business Customer Segmentation

Bachelor of Science in Computer Science

Department of Computer Science

Bahria University, Islamabad

Table of Contents

1. Executive Summary

In order to extract useful insights from transaction data, this study examines consumer segmentation for an online retailer. The study divides consumers into discrete groups using K-Means clustering and RFM (Recency, Frequency, Monetary) analysis, which helps the marketing team run focused ads. Understanding consumer spending trends and identifying high-value and high-risk customer segments are the main goals of the insights.

2. Introduction

A key element of business intelligence is customer segmentation, which helps organizations customize marketing tactics to meet the needs of certain clientele. This project uses data analytics to segment the market, increase sales, and retain customers.

3. Objectives/Goals

This project's main objectives are to:

- Calculate the number of products sold each month.
- To ascertain the monthly expenditures of clients.
- Dividing up the client base for more effective targeting.
- To lower marketing risks by choosing the appropriate target market for particular initiatives.
- To improve marketing effectiveness by taking into account the distinct qualities of each segment.

4. Knowledge of Business

Retail companies benefit greatly from knowing what their customers buy. The main business questions this analysis aims to address are:

• What is the monthly sales volume of products?

• How much does the typical consumer spend each month?

• How can tailored marketing strategies be fueled by client segmentation?

5. Methodology

5.1 Information Gathering

Dataset: UCI Machine Learning Library online retail transaction data time frame: December 2010– December 2011.

• **Dimensions:** 8 columns and 541,909 rows.

• **Important fields:**

o Invoice No: A transaction's unique identification.

o Product code, or stock code.

Description: Name of the product.

o Quantity: The total number of goods in each transaction.

o InvoiceDate: The transaction's date and time.

o Unit Price: The cost of each item.
o CustomerID: A special client identification number.
o Country: The nation of the customer.

5.2 Preprocessing Data

• Null Values: CustomerIDs that were missing from 25% of the data were eliminated.
• Negative Values: Transactions that represented returns and had negative quantities or prices were not included.
'United Kingdom' clients accounted for 90% of the dataset, which was a geographic limitation.

5.3 Analysis of Exploratory Data (EDA)

• Monthly Sales Analysis: o the biggest product sales (13.41% of total transactions) were registered in November.
Sales peaked at 13.41% of total sales in November as well.
• Visualization: To spot patterns and anomalies in consumer spending patterns, graphs and plots were employed.

5.4 RFM Analysis Modeling

Days since the last purchase is the recency (R) metric.
• Frequency (F): The quantity of transactions for each client.
• Monetary (M): The sum of each customer's purchases.
• Quantile Scores: o Using RFM measurements, customers received scores ranging from 1 (best) to 4 (worst).

RFM Segments:

Best Customers: R=1, F=1, M=1 (Score: 111)–409 customers are the RFM Segments.
• Loyal clients: 571 clients who make frequent purchases (Score: F=1).
• Big Spenders: 273 clients with high monetary worth (Score: M=1).
• Nearly Lost: 21 consumers with declining engagement (Score: 134).
• Lost clients: 168 clients with no recent purchases (Score: 344).
• Cheapest clients: 343 clients with little expenditure and involvement (Score: 444).

K-Means Clustering

Cluster analysis using K-Means clustering:
o To group customers into clusters, K-Means was used.
o To lessen skewness, log adjustments were applied.
• **The Elbow Method:** was used to determine the optimal clusters (K), with optimal K = 4.
• **Evaluation:** o The Davies-Bouldin Score was used to assess clusters; for K=4, the lowest

score was 1.065.

Cluster Interpretation:

• Cluster 0: 29%: Loyal Customers (high spenders, regular purchases).
Almost Lost (recent but infrequent consumers) makes up 20% of Cluster 1.
• Cluster 2: 30% - Lost poor-Cost Customers (poor spending and engagement).
• Cluster 3: 21% of the best customers (high spending and engagement).

6. Findings and Segment Analysis

• Best Customers (21%): This group has the highest value, makes frequent purchases, and generates substantial income.
• Loyal Customers (29%): Regular purchasers who are still valuable even if they haven't made any recent purchases.
• Nearly Lost (20%): Purchased recently but infrequently; at risk of churn.
• Most Cheap Customers (30%): This group is low-value and spends little.

7. Business Recommendations

• Top Clients: o Use cross-selling and upselling techniques.
o Provide exclusive discounts and loyalty prizes.
• Loyal Customers: o Use tailored email marketing to concentrate on retention.
o Use targeted promos to boost engagement.
• Nearly Lost: Create reactivation efforts to revive them.
Offer incentives and temporary deals.
• Lost Cheap Customers: o Make an effort to reactivate but keep marketing spending to a minimum.
Concentrate on promotions that are affordable.

8. Challenges and limitation

• Big Data Volume: Python was required for analysis because the dataset was larger than Excel could handle.
• Missing Values: The usable dataset was diminished by significant null entries in the customer data.
• Gaps in Demographic Data: Segmentation granularity was hampered by incomplete demographic data.

9. In conclusion

Customer segmentation has given the company important insights into consumer buying patterns, enabling it to successfully target marketing campaigns. Future marketing efforts

are guided by the analysis's identification of high-value consumers and possible churn threats through the use of RFM and K-Means clustering.

10. References

• Online Retail Dataset, UCI Machine Learning Repository.
• The documentation for Scikit-learn.
• Documentation for Seaborn, Matplotlib, Numpy, and Pandas.