Optimizing Customer Value: RFM Based Segmentation in Online Retail

Deepti Gururaj Baragi

April 23, 2024

Course: CS 695 Capstone Course

Instructor: Prof. Nasir Eisty, Ph.D.

1 Introduction

In today's online retail world, understanding customers' value is crucial for success. Businesses utilize various methods like surveys and market research to learn about customer needs and tailor their offerings and marketing strategies accordingly. This personalized approach not only boosts customer satisfaction but also drives growth in the competitive online retail market. These methods serve as invaluable tools, providing insights that guide strategic decision-making and enhance overall business performance.

However, amidst the ever-evolving digital marketplace, simply understanding customer value at surface level is no longer sufficient. Businesses must delve deeper, into the intricate layers of customer data to uncover the nuances that drive purchasing decisions and foster long-term loyalty. This deeper understanding enables businesses to tailor their products, services, and marketing strategies with precision, ensuring alignment with customer needs and desires.

2 Problem Statement

This project aims to address the complexity arising from the overwhelming volume of customer data available in the digital realm. Businesses require sophisticated analytical methods to sift through this data and extract actionable insights. Thus, there is a need for tools that can navigate through this data maze and provide meaningful insights into customer behavior and preferences. This will be accomplished by leveraging data analytics methods, particularly RFM analysis, to gain a deeper understanding of customer value in online retail.

3 Dataset Description

The dataset used in this project is a transnational dataset containing all the transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company primarily sells unique and all-occasion gifts.

3.1 Variables Description

- **InvoiceNo:** Invoice number, a six-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'C', it indicates a cancellation.
- **StockCode:** Product code, a five-digit integral number uniquely assigned to each distinct product.
- **Description:** Product name.
- **Quantity:** The quantities of each product per transaction, numeric.
- **InvoiceDate:** Invoice Date and time, indicating the day and time when each transaction was generated.
- **UnitPrice:** Unit price, product price per unit in sterling, numeric.
- **CustomerID:** Customer number, a five-digit integral number uniquely assigned to each customer.
- **Country:** Country name, indicating the name of the country where each customer resides.

4 Methodology

The project utilized a multi-step approach to analyze customer behavior and value in the online retail sector.

4.1 Data Cleaning and Wrangling

The initial phase of the project involved data cleaning to ensure the integrity and quality of the dataset. The dataset, sourced from the UCI Machine Learning Repository, underwent a meticulous cleaning process. This included checking for missing data and removing duplicate records to eliminate any inconsistencies.

Upon initial loading and inspection of the dataset, several challenges were identified, necessitating data wrangling to prepare it for analysis. Some of the key challenges encountered and the corresponding actions taken include:

- **Missing Values in 'Description' and 'CustomerID' Columns:** Attempted to link each InvoiceNo to a single CustomerID to fill missing values, but no linkage was found. As missing CustomerID values could not be filled, observations with NaN values in the CustomerID column were dropped. Dropping NaN values in the CustomerID column also removed all missing Description rows.
- **Transactions from the last month of 2011 only had data for 9 days** - they were also removed.

4.2 Exploratory Data Analysis (EDA)

Following data cleaning, EDA was conducted to gain insights into the structure and distribution of the data. Various analyses were performed to understand the characteristics of the dataset and identify patterns and trends. Some of the key aspects explored during EDA include:

- **Top 10 Countries by Percentage of Customers:** Analyzed the distribution of customers across different countries to identify the top markets for the online retail business.
- **Number of Invoices per Year:** Examined the trend in the number of invoices generated each year to understand the overall business activity and growth over time.
- **Products Contributing to Maximum Price Value:** Identified the products that contributed the most to the total revenue generated by the online retail business.
- **Top Selling Products: Distribution of Sales for the Top 20 Products:** Investigated the distribution of sales across the top 20 products to identify the best-selling items and understand their contribution to overall sales.

4.3 Cohort Analysis

Cohort analysis is a powerful tool for understanding customer behavior over time, particularly in terms of retention and loyalty. Some of the key analyses performed during cohort analysis include:

- **Monthly Acquisition Cohorts:** Grouped customers based on the month of their first purchase to analyze their behavior over time.
- **Cohort Retention and Purchase Behavior:** Calculated retention rates and examined purchase behavior (e.g., frequency, monetary value) within each cohort to understand customer retention and loyalty trends.

4.4 Data Modeling

4.4.1 Recency, Frequency, Monetary (RFM) Analysis

RFM analysis is a method used to analyze customer behavior based on three key metrics: Recency, Frequency, and Monetary value. Each metric provides valuable insights into different aspects of customer engagement and spending behavior. The RFM analysis process involves:

- **Recency (R):** Evaluates the time elapsed since the customer's last purchase, indicating how recently a customer has interacted with the business.
- **Frequency (F):** Refers to the total number of purchases made by a customer within a specific period, indicating how often a customer engages with the business.
- **Monetary (M):** Represents the total amount spent by a customer on purchases, offering insights into the customer's spending behavior and contribution to business revenue.

Based on these metrics, customers are segmented into different groups using RFM scores. Common RFM segments include:

- Top Tier Customers
- High-Spending Customers
- New High-Spending Customers
- Active Loyal Customers
- Recent Customers

- At-Risk Customers
- Lost Best Customers
- Inactive Low-Spending Customers

RFM scores ranging from 1 to 4 are assigned to customers for each metric, with 4 being the highest and 1 being the lowest. These scores are then combined to create an overall RFM score, enabling businesses to categorize customers effectively for targeted marketing efforts.

4.4.2 K-means Clustering

K-means clustering was utilized to segment customers based on their RFM scores and other relevant features. The key steps involved in K-means clustering include:

1. **Determine the Ideal Number of Clusters for Segmentation:** The optimal number of clusters is determined using techniques such as the Silhouette Score. This analysis indicates the number of clusters where the mean cluster distance flattens out after a significant decrease, suggesting the ideal number of segments for effective customer segmentation.
2. **Segment Customers with K-means Clustering:** Once the optimal number of clusters is determined, the K-means algorithm is applied to segment customers into distinct groups based on their RFM scores and other relevant features.

5 Results

5.1 Exploratory Data Analysis

The data analysis phase of the project yielded valuable insights into various aspects of the online retail business. Key findings from the data analysis include:

- **Top 10 Countries by Percentage of Customers:**

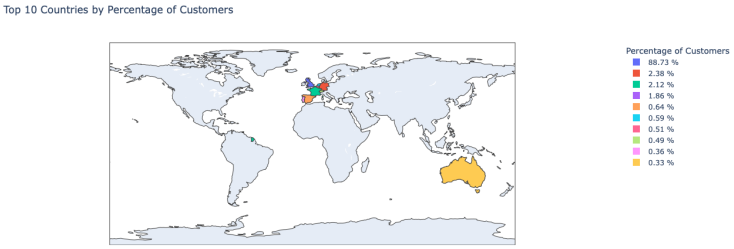

Figure 1: Distribution of Customers across Different Countries

The image above illustrates the distribution of customers across different countries. The top three countries with the highest percentage of customers are the United Kingdom (88.73%), Germany (2.38%), and France (2.12%).

- **Invoices per year:**

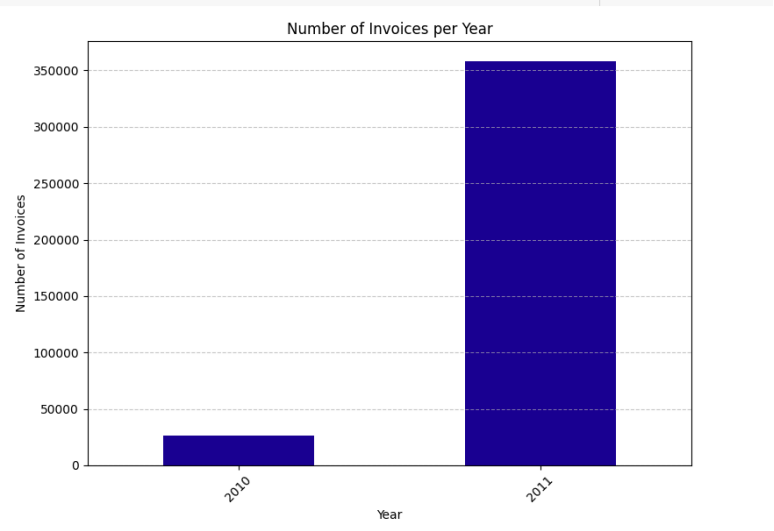

Figure 2: Number of Invoices per Year

The image above depicts the number of invoices generated per year. It provides insights into the overall business activity and growth over time, indicating the trend in transaction volume across different years.

- **Top products contributing to Maximum Price Value:**

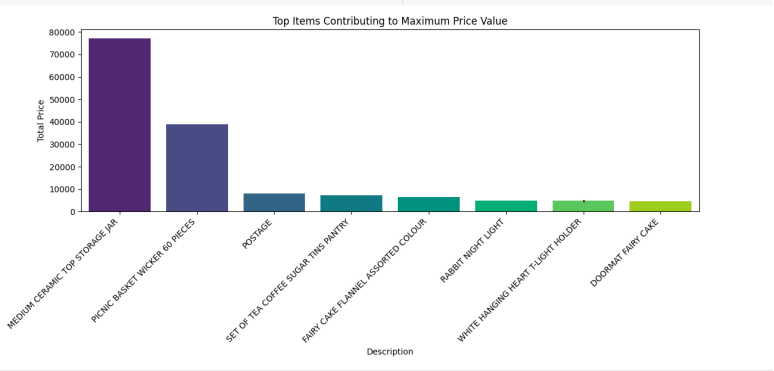

Figure 3: Top Products Contributing to Maximum Price Value

The image above illustrates the products contributing the most to the total revenue generated. The top three products contributing to maximum price value are Medium Ceramic Top Storage Jar, Picnic Basket Wicker 60 Pieces, and Postage.

- **Top selling products:**

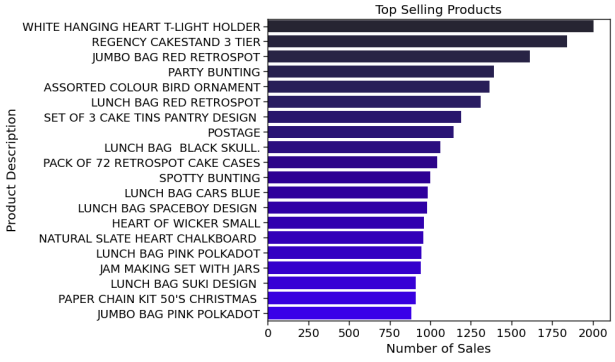

Figure 4: Top Selling Products

The image above depicts the distribution of sales for the top-selling products. The top three selling products are White Hanging Heart T-Light Holder, Regency Cakestand 3 Tier, and Jumbo Bag Red Retrospot.

5.2 Data Modeling

The data modeling phase involved analyzing customer behavior and segmenting them for targeted marketing efforts. Here are the results obtained from the data modeling process:

- **Segments and Marketing Strategies:**

	Segment	RFM	Description	Marketing
0	Top Tier Customers	[444, 443]	High-value customers who recently made frequen...	Exclusive VIP offers and early access to new r...
1	High-Spending Customers	[114, 124, 134, 144, 214, 224, 234, 244, 314, ...	Customers who consistently spend a high amount	Luxury rewards and personalized concierge service
2	New High-Spending Customers	[413, 314, 313, 414]	New customers with high spending potential	Welcome bonus and limited-time luxury package ...
3	Active Loyal Customers	[331, 341, 431, 441]	Active customers who make frequent purchases b...	Loyalty points for every purchase and monthly ...
4	Recent Customers	[422, 423, 424, 432, 433, 434, 442, 443, 444]	Customers who have made recent purchases	Flash sales and special discounts for repeat b...
5	At-Risk Customers	[244, 234, 243, 233]	Customers who were previously active but are s...	Win-back offers and personalized recovery plans
6	Lost Best Customers	[144, 134, 143, 133]	Customers who were previously loyal but have n...	Comeback deals and special reactivation bonuses
7	Inactive Low-Spending Customers	[122, 111, 121, 112, 221, 212, 211]	Customers with low spending and little recent ...	Revamp offers with irresistible discounts and ...

Figure 5: Segments and Marketing Strategies

The image illustrates the segmentation of customers based on RFM analysis scores. For instance, Top Tier Customers are those with the highest RFM scores, indicating recent purchases, high frequency, and high monetary value. On the other hand, At-Risk Customers show signs of decreasing engagement, while Lost Best Customers were previously high spenders but haven't made recent purchases.

- **RFM Segment Allocation:**

CustomerID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore	Segment
17212.0	243	1	226.85	1	1	1	111	Inactive Low-Spending Customers
16626.0	13	18	3855.44	4	4	4	444	Top Tier Customers
17148.0	42	1	124.88	3	1	1	311	Others
16419.0	103	7	945.96	2	4	3	243	At-Risk Customers
15388.0	261	1	140.54	1	1	1	111	Inactive Low-Spending Customers
17819.0	62	8	3661.07	2	4	4	244	High-Spending Customers
12726.0	19	7	2609.10	3	4	4	344	High-Spending Customers
14006.0	106	3	962.19	2	2	3	223	Others
16968.0	98	1	253.31	2	1	1	211	Inactive Low-Spending Customers
15067.0	69	3	1744.76	2	2	4	224	High-Spending Customers

Figure 6: RFM Segment Allocation

In the image, allocation of segments to each customer is based on their RFM score, with 444 representing top-tier customers and 111 representing inactive low-spending customers, among others.

- **K-means clustering:**

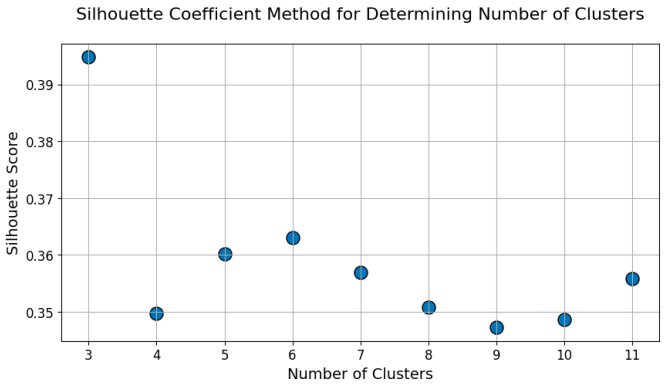

Figure 7: K-means Clustering

This image illustrates the optimal number of clusters determined by the silhouette method, indicating that 4 clusters are to be chosen for effective segmentation.

- **Assigning cluster to segments:**

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	172.0	2.0	4761.0	1952
1	21.0	2.0	4878.0	1032
2	13.0	32.0	24909.0	173
3	31.0	8.0	6707.0	1174

Figure 8: Assigning Cluster to Segments

Once the optimal number of clusters is determined, the K-means algorithm is applied to segment customers into distinct groups. These segments are then aligned with the predefined RFM segments to assign clusters to segments effectively. Based on the characteristics from the image, we can make assumptions about the segments as follows:

- **Cluster 0: Low Spending, Active Customers** - These customers have made moderate-value purchases relatively recently but with low frequency. They may be occasional shoppers or customers who haven't made purchases recently.
- **Cluster 1: Low Spending, Active Customers** - Similar to Cluster 0, these customers have made recent, low-frequency purchases but with slightly higher monetary value compared to Cluster 0.
- **Cluster 2: High-Value, Engaged Customers** - This small group consists of highly engaged and high-value customers. They have made very recent, frequent, and high-value purchases. They are likely loyal and high-spending customers.
- **Cluster 3: Moderate Spending, Engaged Customers** - Customers in this cluster have made moderately recent purchases with moderate frequency and monetary value. They are relatively engaged and valuable customers but not as much as those in Cluster 2.
- **Data Visualization: Tableau Dashboard**

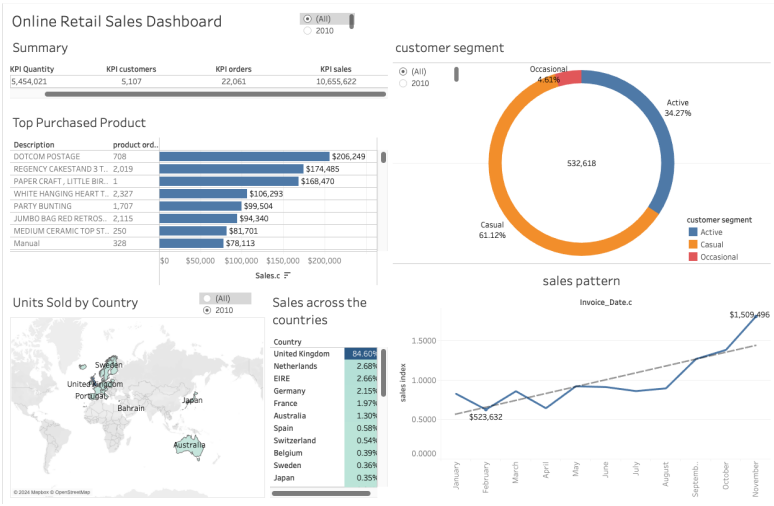

Figure 9: Tableau Dashboard

The Tableau dashboard above provides a comprehensive overview of key performance indicators (KPIs) and insights derived from the analysis. It includes visualizations for various aspects such as customer segments, quantity, products, sales, top purchased products, sales across countries, units sold by country, and sales patterns.

6 Conclusion

Understanding customer value is vital in today's online retail landscape. Traditional methods like surveys and market research provide valuable insights, but with the vast amount of customer data available, businesses must adopt sophisticated analytical methods to extract actionable insights. This project addressed this challenge by leveraging RFM analysis and K-means clustering to segment customers and tailor marketing strategies accordingly.

Cohort analysis revealed trends in customer retention and loyalty over time, enabling businesses to develop targeted strategies for customer engagement. RFM analysis provided insights into different aspects of customer behavior, allowing businesses to categorize customers effectively for personalized marketing efforts. K-means clustering further refined customer segmentation, enabling businesses to identify distinct customer groups and tailor marketing strategies to their specific needs.

7 Future Scope

Moving forward, there are several avenues for future research and development in this area. Firstly, incorporating additional variables such as demographic data and browsing behavior could enhance the accuracy of customer segmentation and predictive models. Additionally, exploring advanced machine learning techniques beyond K-means clustering could provide deeper insights into customer behavior patterns.