DYNAMIC CUSTOMER SEGMENTATION FOR ENHANCED MARKETING STRATEGY

Data 602 Introduction to Data Analysis and Machine Learning
Shashankrajgupta Gunta, XY43338
Premanth Alahari, MO99570
Rohith Eerabattini, HE39576

Project Summary

The aim of this Project is to do customer segmentation for an e-commerce company according to unsupervised learning concepts. The customer segmentation process is used to create small groups out of the whole customer base that have the same characteristics, purchase behavior, or preferences. By correctly developing groups of customers, their business can achieve the information and data needed to customize marketing strategies for particular categories properly, and therefore, customer satisfaction will be improved, and the profitability of their business will be increased. To use machine learning methods for the division of online retail customers into meaningful groups according to their purchasing behavior and characteristics. This division will assist the business in learning its customer base, discovering patterns and trends, and creating personalized marketing campaigns for each segment so that it can be targeted effectively.

Problem Statement

The task is to identify major customer segments using a transnational dataset containing all transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail business. The dataset includes various attributes such as Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country.

Dataset

The project focuses on a dataset, which contains a great deal of information about the online customers of retailers. This includes customer demographics, purchase history, frequency of purchases, the monetary part that a single purchase accounts for, and many other elements that serve as the basis of customer segmentation.
Please refer to this link for the Dataset used for this project:
https://archive.ics.uci.edu/dataset/352/online+retail

Data Description

- **Invoice No**: Invoice number. Nominal is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- **Stock Code:** Product (item) code. Nominal is a 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product (item) name. Nominal.
- **Quantity:** The quantities of each product (item) per transaction. Numeric.

- **Invoice Date:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
- **Unit Price:** Unit price. Numeric, Product price per unit in sterling.
- **Customer ID:** Customer number. Nominal is a 5-digit integral number uniquely assigned to each customer.
- **Country:** Country name. Nominal is the name of the country where each customer resides.

Data Cleaning and Exploration

The initial steps entailed data cleaning to tackle the missing values, which were mostly in the customer ID and the description columns. The null customer IDs were replaced with identification numbers that were based on invoice numbers in order to achieve data integrity. The findings of the exploratory data analysis indicated that there were many interesting things that could be learned about the transaction patterns, customer distribution across countries, and other significant statistics such as the average order size and the customer count.

RFM Analysis

RFM (Recency, Frequency, Monetary) technique is the most common customer segmentation tool applied in marketing and analytics. It helps businesses understand and categorize their customers based on three key factors: The following are the aspects that are taken into consideration when measuring the effectiveness of a product launch: The time since they bought a product (Recency), How often they buy products (Frequency), and how much they spend (Monetary value). The RFM (Recency, Frequency, Monetary) analysis, which is one of the methods of collecting data on customer engagement levels and purchasing habits, has frequency distribution in its essence.

Observations

- Data cleaning and exploration revealed missing values in the Customer ID and Description columns, which were subsequently dropped.
- Further exploration uncovered insights such as the highest-selling products and anomalies like negative quantities and zero-unit prices.
- RFM (Recency, Frequency, Monetary) analysis was identified as a crucial technique for customer segmentation, providing valuable insights into customer behavior and preferences.

Clustering Techniques

1. **KMeans Clustering:** The Elbow Method and Silhouette Score were used to find the best number of clusters using Recency and Monetary variables. Besides, the DBSCAN Algorithm was also used for Frequency and Monetary variables for validation.
2. **Hierarchical Clustering:** Studied hierarchical clustering as another way of customer segmentation and used the dendrogram analysis to find the best number of clusters.
3. **DBSCAN Clustering Algorithm:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Algorithm is used as an alternative ML Algorithm technique for

clustering validation. It is based on the density of data points and chooses the clusters as the regions of high density separated by the regions of low density.

Elbow Method

The Elbow method is employed to decide the best number of clusters in a data set. Through the plot of the explained variation against the number of clusters, the “elbow” point, where the rate of decrease sharply changes, is the optimal number of clusters.

Cross-verifying with Elbow Visualizer

The process of confirming the results of the Elbow Method is of great importance for robustness. Visualization helps in finding the elbow point more precisely and verifies the selected number of clusters.

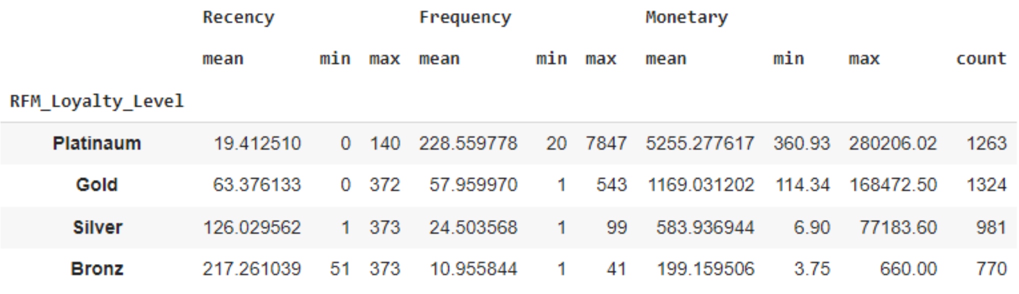
Silhouette Score

The Silhouette Score is a tool to check the quality of the clustering by evaluating the compactness and the separation of the clusters. A higher Silhouette Score means better clustering. It is especially helpful for assessing the validity of the chosen number of clusters.

Conclusion

This project was successful through tracking of the RFM analysis, where customers were segmented into four groups depending on their previous purchasing activities, such as recency, frequency, and monetary purchase scores. Following that, we used two machine learning algorithms, namely, K-means Clustering, to refine segmentation that was done by using RFM data jointly. The analysis revealed two primary customer segments. The segmentation analysis pointed out that we had two primary customer groups:

Initially, our analysis commenced with clustering based on RFM (Recency, Frequency, Monetary) analysis, resulting in the segmentation of customers into four distinct clusters.

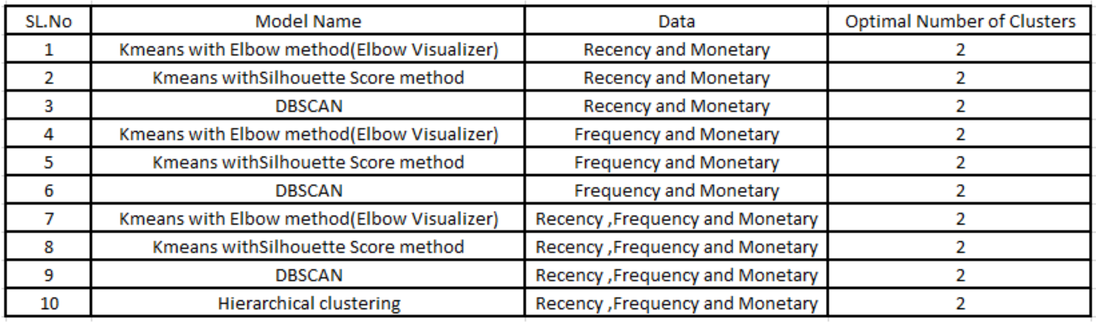


	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
RFM_Loyalty_Level										
Platinaum	19.412510	0	140	228.559778	20	7847	5255.277617	360.93	280206.02	1263
Gold	63.376133	0	372	57.959970	1	543	1169.031202	114.34	168472.50	1324
Silver	126.029562	1	373	24.503568	1	99	583.936944	6.90	77183.60	981
Bronz	217.261039	51	373	10.955844	1	41	199.159506	3.75	660.00	770

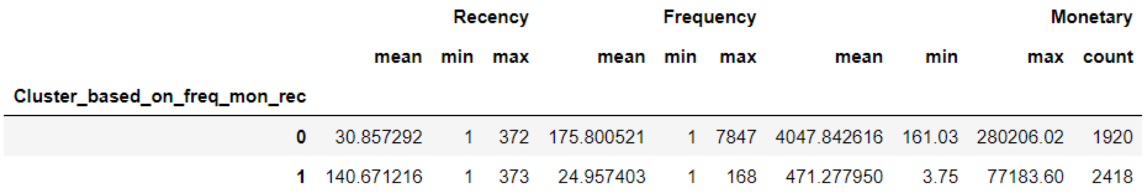
This table summarizes the average recency, frequency, and monetary values for customers segmented into different loyalty levels. Here's a breakdown:

- **Platinum:** Customers in the Platinum segment have an average recency of 19.41 days, make an average of 228.56 purchases, and spend an average of £5255.28.
- **Gold:** Gold-level customers have an average recency of 63.38 days, make an average of 57.96 purchases, and spend an average of £1169.03.
- **Silver:** Silver-level customers have an average recency of 126.03 days, make an average of 24.50 purchases, and spend an average of £583.94.
- **Bronze:** Bronze-level customers have an average recency of 217.26 days, make an average of 10.96 purchases, and spend an average of £199.16.

Subsequently, we proceeded to implement machine learning algorithms to refine customer segmentation further.



SL.No	Model Name	Data	Optimal Number of Clusters
1	Kmeans with Elbow method(Elbow Visualizer)	Recency and Monetary	2
2	Kmeans withSilhouette Score method	Recency and Monetary	2
3	DBSCAN	Recency and Monetary	2
4	Kmeans with Elbow method(Elbow Visualizer)	Frequency and Monetary	2
5	Kmeans withSilhouette Score method	Frequency and Monetary	2
6	DBSCAN	Frequency and Monetary	2
7	Kmeans with Elbow method(Elbow Visualizer)	Recency ,Frequency and Monetary	2
8	Kmeans withSilhouette Score method	Recency ,Frequency and Monetary	2
9	DBSCAN	Recency ,Frequency and Monetary	2
10	Hierarchical clustering	Recency ,Frequency and Monetary	2



	Recency			Frequency			Monetary			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster_based_on_freq_mon_rec										
0	30.857292	1	372	175.800521	1	7847	4047.842616	161.03	280206.02	1920
1	140.671216	1	373	24.957403	1	168	471.277950	3.75	77183.60	2418

1. **Cluster 0:** Cluster 0 represents customers with a low recency rate but high frequency and monetary values. These customers are loyal patrons who frequently make purchases and contribute significantly to revenue generation. Tailored marketing strategies aimed at enhancing their experience could further boost their engagement and loyalty.

2. **Cluster 1:** This segment comprises customers with a high recency rate but low frequency and monetary values. Despite their recent transactions, these customers show minimal engagement with the business. Understanding their needs and preferences could potentially re-engage them and increase their contribution to revenue.