National University of Computer & Emerging Sciences
Karachi Campus

Machine Learning
Date of submission: 13 December 2023

“Customer Segmentation Using Machine Learning”

21K-4703 ALI RAZA
21K-3100 MUHAMMAD SAMEED
21K-4736 SYED SAADULLAH HUSSAINI

(BAI-5A)

Supervisor: Mehak Mazhar

Introduction

This project leverages cutting-edge machine learning techniques to meticulously segment customers, aiming to categorize them based on diverse characteristics. The ultimate goal is to refine marketing strategies and elevate the overall customer service experience.

Methodology

The approach employed in this project encompasses a series of meticulous steps:

Importing Libraries

The initiation involved importing vital Python libraries for both data analysis and machine learning. This suite of libraries, including but not limited to Pandas, NumPy, and Scikit-learn, facilitates seamless data manipulation, comprehensive analysis, and efficient model implementation.

Data Loading and Preprocessing

The customer data, featuring 2240 entries across 26 columns, was loaded into a Pandas DataFrame. Preliminary exploration uncovered missing values in the 'Income' column, promptly addressed by replacing them with the mean income. This ensures the dataset's readiness for subsequent in-depth analysis.

Feature Engineering

Removing Irrelevant Features

Two irrelevant columns, 'Z_CostContact' and 'Z_Revenue', were dropped from the dataset to enhance the efficiency of the machine learning models.

Education and Marital Status Transformation

The 'Education' column underwent transformation, classifying customers into broader groups ('PostGrad' and 'UnderGrad'). Similarly, the 'Marital_Status' column was simplified into 'Not Single' and 'Single' categories for enhanced segmentation.

Creating a Product DataFrame

A new DataFrame, 'Products_DF,' was crafted to systematically organize and scrutinize customer spending across various product categories.

Creating Additional Features

Several supplementary features were engineered to offer more context for customer segmentation, including the total number of kids, total expenses, total accepted campaigns, and the overall number of purchases.

Date Transformation and Feature Creation

The 'Dt_Customer' column was transformed into datetime format, and additional features, such as 'day_engaged,' were introduced to comprehend the duration of customer engagement with the company.

Dropping Redundant Columns

Redundant columns, identified during segmentation analysis, were prudently removed from the dataset.

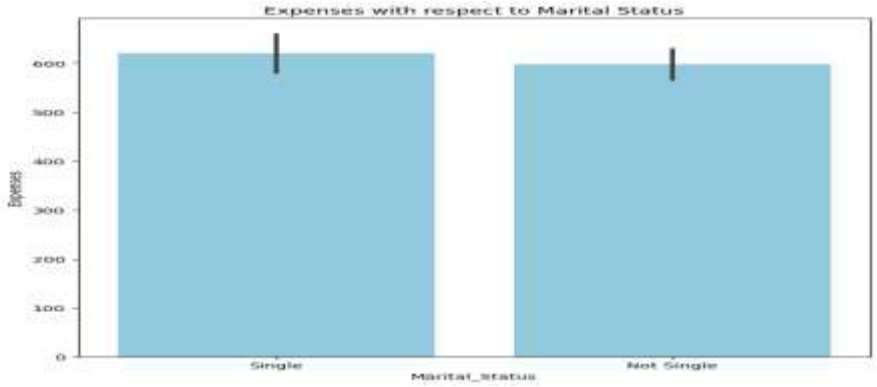
Age Calculation and Final Column Removal

Customer age was accurately calculated based on the 'Year_Birth' column, subsequently eliminating the latter from the dataset.

Recency, Frequency, Monetary (RFM) Analysis

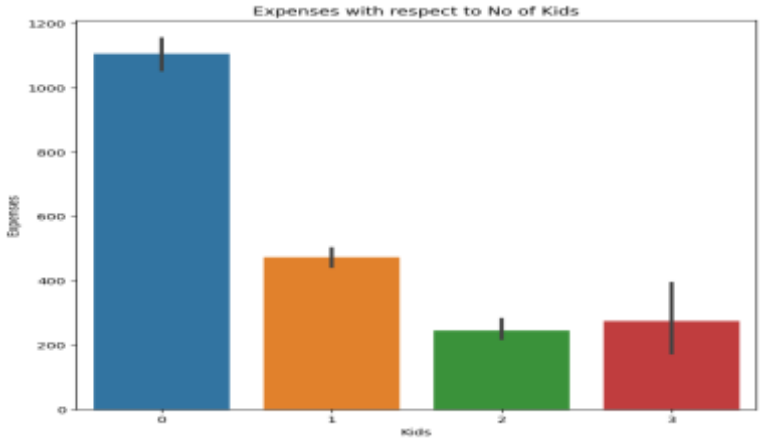
Before the preliminary customer segmentation through K-Means clustering and an exploration of purchasing patterns via the Apriori algorithm, an RFM analysis needs to be conducted. This quantitative technique scores each customer on recency, frequency, and monetary aspects, unveiling segments likely to respond favorably to engagement campaigns, thus optimizing marketing endeavors.

Recency Analysis



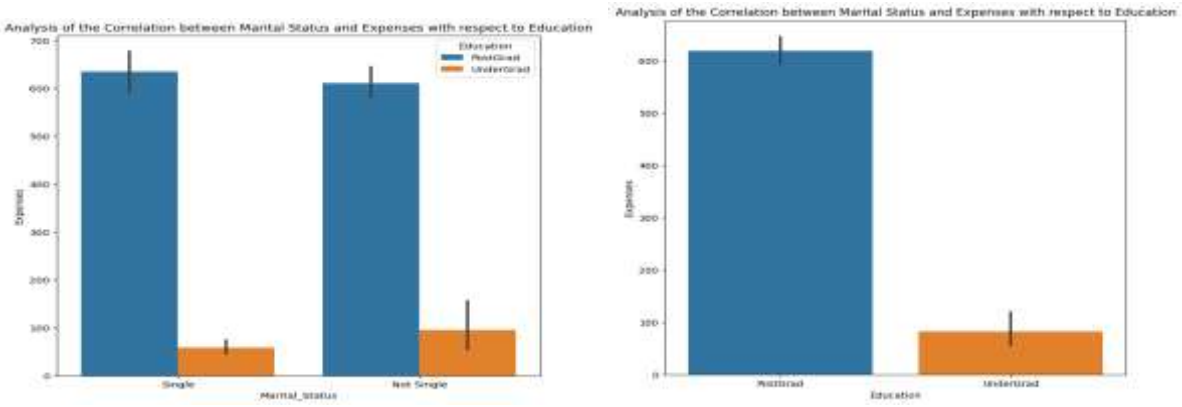
The first chart examines the recency part of the RFM analysis. Specifically, customers identified as 'Single' show a bit more frequent recent purchases compared to those who are 'Not Single.' This indicates a potential inclination to respond more to new offers and increased involvement with the brand.

Frequency Analysis



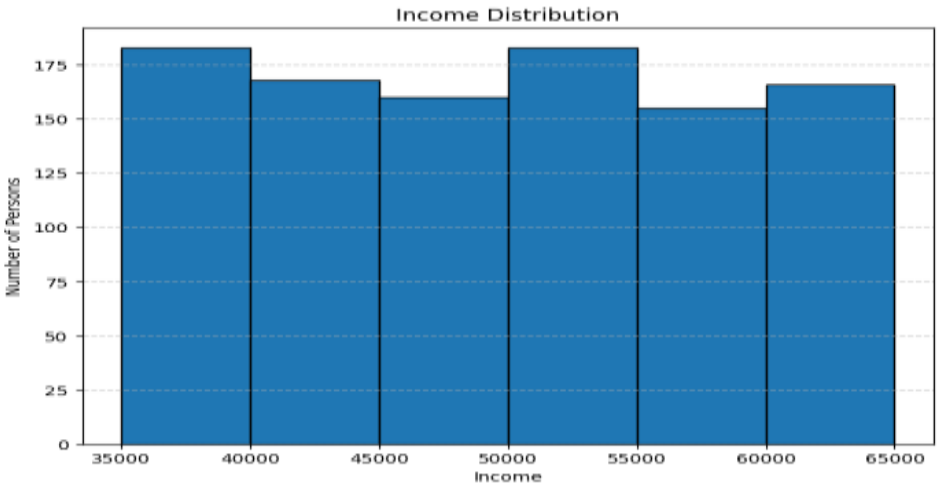
The second graph delves into the frequency of purchases concerning the number of kids. Evidently, customers with no kids exhibit the highest frequency of purchases, gradually declining as the number of kids increases. This could signify that customers with larger families might have less disposable income or time for frequent purchases.

Monetary Analysis with respect to Marital Status and Education



The analysis of customer expenses segmented by marital status and education level uncovers intriguing patterns. The third graph illustrates that single customers with a postgraduate education incur the highest expenses, potentially reflecting greater financial freedom and a penchant for premium products. Conversely, the fourth graph simplifies the comparison, focusing solely on education levels, highlighting that postgraduates, in general, incur higher expenses than undergraduates. This duality suggests that educational attainment strongly influences spending behavior, with postgraduates exhibiting a higher spending capacity, possibly linked to advanced degrees' associated higher income levels.

Income Distribution Analysis



Contrary to expectations of a bell curve in typical income distributions, the histogram portrays a uniform distribution across income brackets from Rs. 35,000 to Rs. 65,000. This suggests an equal probability of individuals falling into any income category within this range—a characteristic of a uniform distribution rather than a normal distribution.

Conclusions from RFM Analysis

The RFM analysis provides nuanced insights into customer behavior. Single, postgraduate customers emerge as a prime segment for premium offerings, while families appear receptive to value-based promotions. The income distribution analysis supports the strategy of segment-based targeting. By leveraging this segmentation, customized marketing efforts can enhance customer experience and

optimize the marketing budget.
The RFM analysis emphasizes personalized customer engagement. 'Single' customers with frequent purchases could benefit from exclusive rewards and new product launches. Families, with lower purchase frequencies, may find appeal in bulk purchase discounts. This targeted strategy aligns marketing efforts with specific segment traits, enhancing connections.

Discussion and Observations

Detailed disscussion and observations from the notebook include:

K-Means Clustering Results:

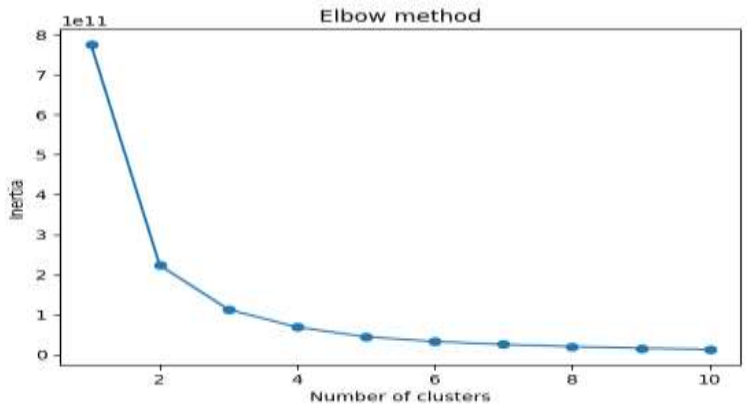
The application of K-Means clustering led to the segmentation of customers into distinct groups. The results show how these groups differ in terms of purchasing behavior, frequency, and monetary value. The cluster analysis reveals patterns and characteristics unique to each segment, providing insights into customer behavior.

K-Means Clustering Analysis

K-Means clustering is a pivotal machine learning technique used in our study to identify distinct customer groups based on their purchasing behavior, frequency, and monetary value. The process began with the selection of the optimal number of clusters using the elbow method.

Elbow Method Results

The elbow graph illustrates a clear bend at two clusters, indicating that additional clusters do not significantly contribute to explaining variance within the data. This 'elbow' suggests that two clusters are the optimal number for our segmentation (see Elbow Method graph).



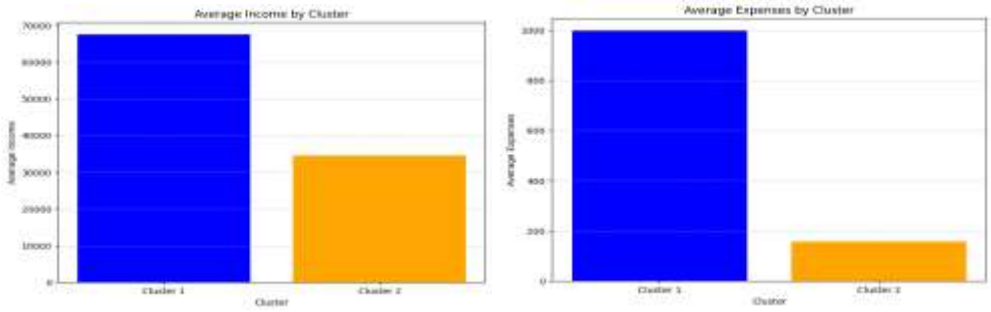
Selection and Application of K-Means Clustering

With two clusters identified, the K-Means algorithm was applied to the dataset, yielding the following results:

Cluster Profiles

1. **Cluster 1 - Highly Active Customers:**
    - Customers in this cluster exhibit high expenses, correlating with a higher average income.

- These customers are highly engaged and contribute significantly to revenue, making them ideal targets for up-selling and premium services. (see Average Income by Cluster and Average Expenses by Cluster graphs below).



2. **Cluster 2 - Least Active Customers:**
    - This cluster comprises customers with lower expenses and income.
    - While less active than Cluster 1, these customers still represent a substantial portion of the customer base.
    - Strategies focused on retention and value-oriented offerings may be effective for this segment.

Implications for Customer Segmentation

The distinct characteristics of the two clusters allow for tailored marketing strategies. Cluster 1, with higher income and expenditure, may respond well to luxury products, while Cluster 2, with lower income, may require more conservative marketing approaches.

Conclusions from K-Means Clustering

K-Means clustering effectively segments customers, enabling the development of targeted marketing and service strategies. The insights from the clusters, combined with the RFM analysis, equip us with a nuanced understanding of customer behavior.This comprehensive analysis is instrumental in shaping informed business decisions that cater to the diverse needs of our customer base.

Apriori Algorithm Analysis:

The Apriori algorithm, focusing on market basket analysis, uncovers associations and relationships between different products purchased by customers. Segmentation based on age, income, and engagement levels is extended to product categories, identifying customers as 'Least Active' or 'Highly Active' based on their purchasing activity.

Association Rule Mining

After preprocessing and applying one-hot encoding, the Apriori algorithm was employed with a minimum support threshold of 0.08 and a maximum length of 10 to mine for frequent itemsets and generate association rules with a lift metric.

Insights from Association Rules

Two key insights emerged from the association rules generated:

- **Patterns Among Least Active Customers in Fruit Purchases**
  The analysis revealed a strong association between customers who are categorized as 'Least Active' in purchasing fruits and other product segments. For instance, Customers categorized as 'Least Active' in purchasing fruits showed a strong association with being 'Least Active' in other product segments. This suggests a significant probability that customers less active in certain categories will exhibit similar behavior in fruit purchases.

- **Correlations with Highly Active Fruits Purchasers:**
  The segment of 'Highly Active' fruits purchasers shows a strong association with being 'Highly Active' in buying other product categories, including Gold, Meat, Wines, and Sweets. This suggests that customers who are highly active in purchasing fruits are very likely to be active in other high-value product categories as well, such as meats, wines, and sweets.

Conclusions from Apriori Analysis
The insights from the Apriori analysis offer practical guidance for cross-selling and promotions. Understanding linked purchasing behavior allows for strategic bundle offers, personalized recommendations, and focused marketing campaigns. For instance, marketing plans can be devised to cross-promote fruit items to active wine buyers or craft value bundles for less active segments, encouraging purchases across categories.

The Apriori analysis seamlessly enhances customer segmentation achieved through RFM and K-Means clustering. It identifies specific product associations, refining the targeting of diverse customer groups. Integrating these findings empowers our marketing strategies with data-driven precision, increasing the likelihood of effectively engaging varied customer segments.

Conclusions

Concludingly, our Machine Learning Project has successfully categorized customers through sophisticated data analytics, bolstering the impact of marketing endeavors. The strategic use of RFM analysis and K-Means clustering distinctly grouped customers, exposing actionable insights into their purchasing patterns. Furthermore, the Apriori algorithm provided valuable insights into product associations, enhancing the precision of our marketing strategies. These methods, coupled with rigorous data processing and feature engineering, have empowered us with a data-driven approach to customize customer service and marketing endeavors, fostering a business growth strategy centered around the customer.