Customer Segmentation Project Report

Naman Manchanda

June 12, 2024

1 Introduction

Customer segmentation is a crucial task in marketing and business analytics. It involves dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, and spending habits. In this project, we aim to segment customers based on their purchasing behavior using clustering techniques.

2 Data Description

The dataset contains information about customers, including their demographics, purchasing behavior, and response to marketing campaigns. It consists of 2240 rows and 29 columns. Some of the key columns include:

- **ID**: Unique identifier for each customer
- **Year_Birth**: Year of birth of the customer
- **Education**: Level of education of the customer
- **Marital_Status**: Marital status of the customer
- **Income**: Income of the customer
- **Dt_Customer**: Date of customer registration
- **Recency**: Number of days since the last purchase
- **MntWines, MntFruits, ...**: Amount spent on different product categories
- **Response**: Response to marketing campaigns (target variable)

3 Data Preprocessing and Feature Engineering

Before clustering, we performed several preprocessing steps and feature engineering techniques, including:

- Handling missing values in the **Income** column
- Encoding categorical variables (**Education**, **Marital_Status**)
- Feature scaling to standardize the numerical features
- Date feature extraction from **Dt_Customer**

4 Principal Component Analysis (PCA)

Since the dataset has 28 columns, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the data while preserving most of its variance. PCA helped in capturing the underlying structure of the data and reducing noise.

5 Clustering Techniques

We employed two clustering techniques for customer segmentation: Gaussian Mixture Model (GMM) clustering and K-Means clustering.

5.1 Gaussian Mixture Model (GMM)

GMM is a probabilistic clustering algorithm that models the distribution of the data as a mixture of several Gaussian distributions. We initially applied GMM clustering to the PCA-transformed data.

5.2 K-Means Clustering

K-Means is a centroid-based clustering algorithm that partitions the data into K clusters based on Euclidean distance. We then modified the K-Means algorithm to use cosine similarity instead of Euclidean distance.

6 Results

After clustering using both GMM and modified K-Means, we evaluated the performance of each clustering technique using the Silhouette Score. The Silhouette Score for the modified K-Means clustering with cosine similarity was found to be 0.65, indicating good cluster separation.

7 Conclusion

In conclusion, customer segmentation is a vital task for targeted marketing strategies. By applying clustering techniques such as GMM and modified K-Means with cosine similarity, we were able to effectively segment customers based on their purchasing behavior. These segments can be used by businesses to tailor marketing campaigns and improve customer satisfaction.