Customer Segmentation using Unsupervised Machine Learning

A Data-Driven Approach to Market Segmentation

Om Gaud
Alston Gonsalves

July 8, 2025

Contents

1 Introduction

Customer segmentation is the process of dividing a customer base into groups based on shared characteristics such as demographics, purchasing habits, or responses to marketing campaigns. This project explores segmentation using unsupervised learning to extract actionable business insights from raw customer data.

Objective: To identify distinct customer segments based on their demographics, purchasing behavior, and responsiveness to marketing promotions, and to visualize these segments effectively.

2 Dataset Overview

2.1 Source

A marketing campaign dataset containing customer demographic and transactional data.

2.2 Key Features

- Demographics: Age, Income, Education, Marital Status
- Household: Number of kids and teens at home
- Purchase history across different product categories
- Marketing campaign responses
- Customer enrollment date

3 Data Preprocessing

3.1 Loading and Cleaning

- Loaded data using `pandas.read_csv()`
- Inspected using `data.info()` and `data.describe()`
- Removed null values using `dropna()` to ensure data integrity
- Converted date columns (e.g., `Dt_Customer`) to datetime format

3.2 Feature Engineering

- Derived customer age: `2021 - Year_Birth`
- Computed total money spent: `Spent = Wines + Fruits + Meat + Fish + Sweets + Gold`
- Computed family size and parenting status
- Encoded categorical variables (e.g., Education, Living_With)
- Removed outliers: Age ¿ 90 and Income ¿ 600000

4 Exploratory Data Analysis

4.1 Pairplot and Correlation Heatmap

- Explored key numerical relationships using `sns.pairplot()`
- Used heatmap to identify highly correlated features

4.2 Visualizations

- Swarm and boxen plots for spending across clusters
- Countplots of promotional responses
- Joint KDE plots for Age, Family Size, Education vs. Spending
- 3D scatter plots for PCA-based visualization of clusters

5 Feature Scaling and Dimensionality Reduction

- Used `StandardScaler()` to normalize feature scales
- Reduced dimensionality using `PCA` with 3 components
- Visualized reduced data in 3D for clustering insight

6 Clustering Techniques

6.1 KMeans Elbow Method

- Applied `KElbowVisualizer` to determine optimal number of clusters
- Found 4 clusters to be ideal

6.2 Agglomerative Clustering

- Performed clustering using `AgglomerativeClustering(n_clusters=4)`
- Labeled customers by cluster and appended to the main dataset
- Visualized cluster distribution in 3D space

7 Cluster Profiling

7.1 Cluster Distribution and Insights

- Analyzed how each cluster differs in terms of:
  - Spending behavior
  - Deal acceptance

  - Promotional responsiveness
  - Income and family structure

- Created meaningful profiles for each cluster:

Example:

- Cluster 0: Budget-conscious families, high deal acceptance
- Cluster 1: High-income professionals, low promo response
- Cluster 2: Mid-income, moderate spenders
- Cluster 3: Loyal premium customers with high lifetime value

8 Business Recommendations

- Personalize marketing strategies for each segment
- Invest more in promotional campaigns for responsive clusters
- Offer loyalty programs to retain premium customers
- Provide discounts or bundles for budget-sensitive segments

9 Technologies Used

- **Python libraries:** pandas, numpy, seaborn, matplotlib, sklearn
- **Visualization:** Pairplots, heatmaps, 3D scatter, KDE jointplots
- **Dimensionality Reduction:** PCA
- **Clustering:** KMeans (Elbow method), Agglomerative Clustering
- **IDE:** VS Code / Jupyter Notebook
- **Documentation:** LaTeX

10 Conclusion

This project successfully demonstrates how unsupervised learning techniques like clustering can uncover hidden customer segments from raw behavioral and demographic data. These insights can significantly improve business decision-making and marketing personalization.

11 Future Work

- Integrate segmentation into a live dashboard or web app
- Apply supervised learning to predict customer behavior
- A/B test campaigns targeted to different clusters
- Include customer churn prediction