Zeno Talent

Customer Segmentation using Advanced Clustering

September 19, 2025.

Overview Report

Problem statement

Businesses often struggle to understand their customers' diverse needs and behaviors. Treating all customers as a single group leads to ineffective marketing, poor targeting, and a loss of potential revenue.

The goal of this project is to perform **customer segmentation** using advanced clustering techniques. By grouping customers with similar characteristics and behaviors, businesses can:

- Identify high-value customer groups.
- Personalize marketing strategies.
- Improve customer experience and retention.
- Optimize resources for better decision-making.

About the Dataset

The dataset (`Test.csv`) provides demographic and behavioral attributes of customers. It does **not include annual income**, but it contains other useful information for segmentation.

Features

- **ID** → Unique identifier for each customer.
- **Gender** → Male/Female.
- **Ever_Married** → Whether the customer has been married.
- **Age** → Customer's age.

- **Graduated** → Whether the customer has a graduation degree.
- **Profession** → Type of profession (Engineer, Executive, Marketing, etc.).
- **Work_Experience** → Years of work experience.
- **Spending_Score** → Categorical measure of spending behavior (*Low, Average, High*).
- **Family_Size** → Number of members in the customer's family.
- **Var_1** → Encoded category label (*Cat_1 … Cat_6*), representing hidden customer classes provided by the dataset creator.

Key Notes

- Clustering will rely on Age, Work Experience, Family Size, Profession, and Spending Score.
- Var_1 (Cat_1 … Cat_6) is a **categorical grouping variable** that can be used later to compare how well the clusters align with predefined groups.

Methodology

The project follows a systematic approach:

Step 1: Data Preprocessing

- Handle missing values (e.g., missing Profession, Work_Experience).
- Encode categorical variables:
  - Gender → 0/1
  - Spending_Score → Low = 0, Average = 1, High = 2
  - Profession & Var_1 → Label encoding or one-hot encoding.
- Normalize numerical variables (Age, Work_Experience, Family_Size) for fair comparison.

Step 2: Exploratory Data Analysis (EDA)

- Analyze demographic distributions (age groups, gender split, marital status).
- Study profession-wise customer distribution.
- Check relationship between Spending_Score and Age/Family_Size.
- Visualize category distribution in **Var_1 (Cat_1 … Cat_6)**.

Step 3: Feature Selection

Select key attributes that influence customer behavior and segmentation:

- **Demographic**: Age, Gender, Family_Size, Ever_Married.
- **Professional**: Profession, Work_Experience, Education.
- **Behavioral**: Spending_Score.

Step 4: Clustering Techniques

Apply and compare different clustering algorithms:

1. **K-Means Clustering** → Partition customers into fixed k groups.
2. **Hierarchical Clustering** → Create dendrograms to identify nested group structures.
3. **DBSCAN (Density-Based Spatial Clustering)** → Detect irregular clusters and outliers.
4. **Gaussian Mixture Models (GMM)** → Capture overlapping groups with probabilistic membership.

Step 5: Cluster Evaluation

Evaluate and compare cluster performance using:

- **Silhouette Score** – Measures cluster cohesion and separation.
- **Davies–Bouldin Index** – Lower values indicate better clustering.
- **Calinski-Harabasz Index** – Higher values indicate well-defined clusters.
- Compare clusters with `Var_1` to see if segmentation aligns with predefined categories.

Step 6: Results & Insights

- Identify distinct customer groups (e.g., *Young Professionals with Low Spending*, *Large Families with Average Spending*, *Experienced Executives with High Spending*).
- Provide business strategies for each group, such as:
  - Loyalty programs for high spenders.
  - Budget offers for low spenders.
  - Special campaigns for young unmarried customers.