Customer Segmentation Using SEMMA: A Case Study on the Mall Customers Dataset

Mohibkhan Pathan

November 17, 2024

Abstract

Customer segmentation is an essential process for businesses to identify distinct customer groups and tailor marketing strategies. This paper demonstrates the application of the SEMMA methodology—Sample, Explore, Modify, Model, and Assess—on the Mall Customers dataset to perform customer segmentation. Using K-Means clustering, we identified five distinct customer groups based on their age, income, and spending behavior, providing actionable insights for targeted marketing strategies.

1 Introduction

Customer segmentation is a crucial tool in modern business intelligence, enabling organizations to categorize their customer base into distinct groups. This study employs the SEMMA methodology to perform customer segmentation on the Mall Customers dataset, a simple and widely-used dataset for clustering tasks. The primary objective is to identify meaningful clusters of customers based on their age, income, and spending patterns, enabling businesses to develop targeted strategies to improve customer engag...

2 SEMMA Methodology

The SEMMA methodology provides a structured approach to data mining and consists of five phases:

- **Sample**: Select a representative dataset for analysis.
- **Explore**: Perform exploratory data analysis (EDA) to identify patterns, relationships, and anomalies.
- **Modify**: Prepare and preprocess the data to ensure readiness for modeling.
- **Model**: Build machine learning models to uncover patterns and relationships.
- **Assess**: Evaluate the performance and utility of the models.

3 Dataset Description

The Mall Customers dataset consists of 200 rows and 5 columns:

- **CustomerID**: Unique identifier for each customer.
- **Gender**: Gender of the customer (Male/Female).
- **Age**: Customer's age.
- **Annual Income (k$)**: Annual income of the customer in thousands of dollars.
- **Spending Score (1–100)**: A score assigned by the mall based on customer spending patterns and behavior.

4 Phases of SEMMA

4.1 Sample Phase

The Mall Customers dataset was chosen for its simplicity and relevance to customer segmentation tasks. It contains no missing values, and all features were deemed relevant for clustering.

4.2 Explore Phase

Exploratory data analysis revealed several insights:

- Gender distribution was nearly even, with slightly more females than males.
- Age distribution showed a concentration in the 20–40 age group.
- Scatter plots of annual income and spending score revealed diverse spending behaviors across income levels.

4.3 Modify Phase

Preprocessing steps included:

- Encoding the `Gender` column as 0 (Male) and 1 (Female).
- Scaling numerical features (`Age`, `Annual Income`, and `Spending Score`) using standardization to ensure uniform scaling.

The dataset was prepared for clustering with no missing values or inconsistencies.

4.4 Model Phase

K-Means clustering was used to segment customers. The optimal number of clusters was determined to be 5 using the Elbow Method.

The clustering results revealed distinct customer segments based on their income and spending behavior.

4.5 Assess Phase

The clusters were evaluated based on their characteristics:

- **Cluster 0**: High spenders with moderate income.
- **Cluster 1**: High-income customers with low spending scores.
- **Cluster 2**: Low-income customers with low spending scores.
- **Cluster 3**: Younger customers with high spending scores.
- **Cluster 4**: Older customers with high income but low spending.

5 Key Findings and Insights

The segmentation revealed actionable insights for businesses:

- High-income, low-spending customers (Cluster 1) require targeted promotions to increase engagement.
- Younger, high-spending customers (Cluster 3) represent a loyal and valuable customer base.
- Low-income, low-spending customers (Cluster 2) may require affordability-based campaigns to improve retention.

6 Conclusion

Using the SEMMA methodology, we successfully segmented customers into five distinct groups, providing insights into their behavior and preferences. These insights can be leveraged by businesses to create targeted marketing strategies and improve overall customer satisfaction.

7 References

- Kaggle: Mall Customers Dataset. `https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python`
- SEMMA Methodology. SAS Documentation. `https://support.sas.com/resources/papers/proceedings13/069-2013.pdf`