Customer Segmentation Using the KDD Methodology: A Data-Driven Approach

Soumya Bharathi Vetukuri

2024

Abstract

Customer segmentation is a crucial aspect of modern business intelligence, allowing companies to understand and target their customer base effectively. In this paper, we apply the Knowledge Discovery in Databases (KDD) methodology to a customer segmentation dataset. The KDD methodology provides a structured approach to uncovering valuable insights from raw data. We describe each phase of the KDD process, from data selection and preprocessing to data transformation, data mining, and evaluation. Through the application of clustering algorithms, we identify meaningful customer segments that align with business objectives, leading to actionable insights for personalized marketing and customer retention strategies.

1 Introduction

Customer segmentation is a key tool used by businesses to identify distinct groups of customers, which helps in personalizing marketing strategies, improving customer retention, and optimizing business operations. The Knowledge Discovery in Databases (KDD) methodology provides a systematic framework for extracting meaningful patterns from data. In this study, we apply the KDD process to segment customers based on demographic and behavioral attributes.

2 KDD Methodology

The KDD process consists of five primary phases:

1. Data Selection
2. Data Preprocessing
3. Data Transformation
4. Data Mining
5. Interpretation and Evaluation.

2.1 Data Selection

We utilized a customer segmentation dataset consisting of 8,068 records with demographic and behavioral features. The features include Age, Gender, Work Experience, Family Size, and Spending Score. The goal was to segment the customers into meaningful clusters that businesses could use to tailor their marketing and retention strategies.

2.2 Data Preprocessing

The first challenge in any data analysis project is cleaning and preparing the data. The dataset had some missing values in columns like Ever Married, Graduated, Profession and Family Size. We handled missing values using imputation techniques, such as filling missing categorical data with the most frequent category and filling numerical data with the column mean.

Next, we applied **one-hot encoding** to categorical features like Gender, Profession and spending score to convert them into a numerical format suitable for machine learning algorithms. Additionally, we normalized numerical features like Age and Family Size using StandardScaler to ensure that all features were on the same scale.

2.3 Data Transformation

Once the data was clean, we moved on to this phase. Feature engineering played a crucial role in enhancing the quality of our data. We created new features which represented interactions between key variables like age, work experience, and family size.

These new features allowed us to capture the combined effects of multiple variables, potentially revealing deeper patterns in the data that might not have been visible through the original features alone.

2.4 Data Mining

With the data prepared and transformed, we applied clustering algorithms to identify distinct customer segments. We applied several clustering algorithms to the dataset, including K-Means, DBSCAN, and Gaussian Mixture Models (GMM). K-Means clustering with $k = 4$ was selected based on the elbow method and provided well-defined clusters. We also explored DBSCAN for identifying clusters of varying densities and GMM for probabilistic clustering.

2.4.1 K-Means Clustering

We started by applying K-Means clustering, a popular algorithm that partitions customers into a predefined number of clusters based on similarity. After testing different values for `k` using the **Elbow Method**, we settled on `k=4` clusters, which provided a good balance between simplicity and accuracy.

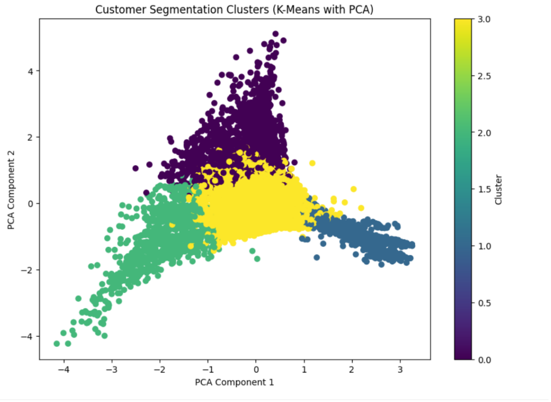

Figure 1: K-Means clustering results visualized using PCA.

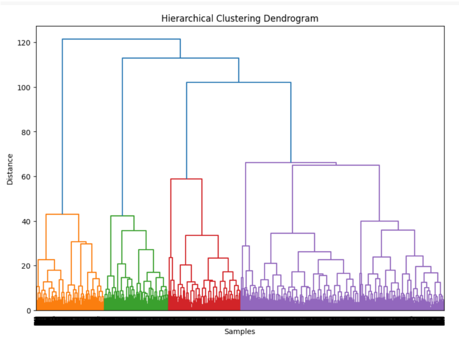

Figure 2: Agglomerative Clustering

We also explored **Spectral Clustering**, **Agglomerative Clustering**, **DBSCAN**, which identifies clusters based on density, and **GMM**, which assumes that the data is generated from a mixture of Gaussian distributions. Each algorithm produced slightly different clusters, highlighting the importance of trying multiple approaches when performing data mining.

2.4.2 Supervised Learning for Segmentation:

If you have a labeled dataset with known segments (such as `Segmentation` in your data), you might want to use supervised learning models to classify customers into segments. Some options include:

1. **Random Forest Classifier**: Random Forests are useful for classification tasks with categorical target variables. They are also robust to over fitting and can handle non-linear relationships.

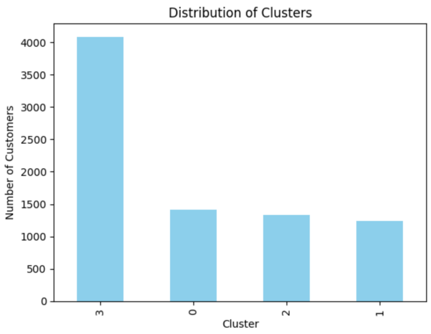

Figure 3: Distribution of Clusters

2. **Gradient Boosting Classifier**: Gradient Boosting is a powerful technique that builds an ensemble of weak learners (typically decision trees). It can capture complex patterns in the data and often outperforms other classifiers.

2.5 Results and Evaluation

The resulting clusters provided actionable insights for the business:

- **Cluster 1**: Young professionals with high spending scores, ideal for tech product marketing.
- **Cluster 2**: Older families with moderate spending, suitable for family-oriented promotions.
- **Cluster 3**: Retired customers with lower spending habits, indicating potential for cost-saving product offerings.
- **Cluster 4**: High-income frequent shoppers, candidates for personalized VIP programs.

The silhouette score was used to evaluate the quality of the clustering, and the clusters were further analyzed to identify patterns in customer behavior.

3 Conclusion

In this study, we successfully applied the KDD methodology to segment customers into actionable groups. The insights derived from these segments can help businesses tailor their marketing strategies, improve customer retention, and enhance overall customer satisfaction. Future work includes incorporating additional data sources and applying predictive models to anticipate customer behavior.

3.0.1 References

https://www.kaggle.com/datasets/vetrirah/customer