Customer Segmentation in Online Retail using the CRISP-DM Methodology

Pallavi Vangari

Abstract

This paper presents a detailed analysis of customer segmentation in an online retail dataset using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Through each phase of the methodology, the paper elucidates the process and decisions, culminating in the segmentation of customers into distinct categories.

1 Introduction

The digital transformation of commerce, especially the rise of e-commerce platforms, has given businesses access to a wealth of data regarding customer interactions and behaviors. This data, if analyzed and leveraged correctly, can yield invaluable insights for businesses, allowing them to tailor their marketing strategies, improve their sales funnel, and enhance customer satisfaction. One of the foundational techniques to achieve this is customer segmentation. This research employs the widely recognized CRISP-DM methodology to methodically segment customers using an online retail dataset.

2 CRISP-DM Methodology: An Overview

The Cross-Industry Standard Process for Data Mining (CRISP-DM) offers a structured framework for planning and executing data mining projects. Its cyclic nature allows for iterative refinement, ensuring that the model's outputs align with business objectives. The six stages of CRISP-DM are:

1. **Business Understanding:** Define the project's objectives and requirements from a business perspective. Convert this knowledge into a data mining problem definition and preliminary plan.

2. **Data Understanding:** Start with data collection, familiarize oneself with the data, identify data quality issues, and explore various statistical summaries and visual methods.

3. **Data Preparation:** All activities required to construct the final dataset, which can include table, record, and attribute selection, as well as transformation and cleaning.

4. **Modeling:** Various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.

5. **Evaluation:** Evaluate the model's ability to achieve business objectives, and seek establishment of a threshold for deploying the model.

6. **Deployment:** The process of deploying the model into a business environment.

3 Business Understanding

In the realm of e-commerce, understanding customer behaviors and patterns is paramount. With competition just a click away, retaining customers and enhancing their buying experience can make the difference. The primary objective here is to segment customers based on their purchasing behaviors, allowing for targeted marketing, personalized product recommendations, and tailored customer experiences.

4 Data Understanding

Our dataset, "online+retail+ii", is a snapshot of transactions from an online retail business. It encompasses various attributes like product codes, descriptions, quantities purchased, transaction timestamps, prices, customer IDs, and country of origin.

5 Data Preparation

5.1 Handling Missing Values

While missing data is commonplace in real-world datasets, it's crucial to address them methodically:

- **Description:** Missing product descriptions were imputed with the label "Unknown".
- **Customer ID:** Given the significance of customer IDs for segmentation, entries with missing IDs were omitted.

5.2 Addressing Outliers

The data contained negative quantities, which could denote returned items or anomalies. For the scope of this study, such entries were excluded to maintain data consistency.

6 Feature Engineering for Modeling

Effective modeling hinges on the quality and relevance of features. Derived from raw data, the following attributes were engineered to represent each customer's buying behavior:

- **Frequency:** Signifying customer engagement.
- **MonetaryValue:** Denoting the total spend, indicative of customer value.
- **Recency:** Representing engagement recency, a critical metric for retention strategies.
- **AvgQuantity:** Average items per transaction, can hint at buying patterns.
- **AvgPrice:** Average spend per item, suggesting the value segment of purchases.

7 Modeling

The K-Means clustering algorithm, a popular unsupervised machine learning technique, was employed to segment customers into clusters. The algorithm's objective is to partition customers into groups where members of each group are more similar to each other than to members of other groups.

8 Evaluation and Interpretation

Post-clustering, the segments were analyzed based on their attributes, leading to clear categorizations like Mainstream Customers, High-Value Loyal Customers,

High-Priced Item Buyers, and Inactive Low-Spenders. Such categorizations empower businesses to craft strategies tailored to each segment's unique characteristics.

9 Conclusion

Utilizing the CRISP-DM methodology, this research demonstrated a structured approach to customer segmentation in an e-commerce context. The derived segments offer actionable insights, enabling businesses to enhance their marketing and sales strategies.

10 Future Work

While the current model offers valuable insights, there's room for enhancement:

- Time-series analysis could predict future buying behaviors.
- Analyzing negative quantities could offer insights into product returns or customer dissatisfaction.
- Advanced clustering techniques or deep learning could be explored for potentially more insightful segmentations.

References

[1] Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *J. Data Warehousing*, 5(4), 13-22.