Nov 2024

Customer Segmentation Report

Subhas Mukherjee
Avantika Roy
Priyam Pal
Jeet Nandi
Manu Bhaskar

1. **Collecting and Preparing Data**
    - Loading Data: We started by loading four CSV files: orders, products, order_products, and aisle.
        - Orders Data: Contains information about each order.
        - Products Data: Lists details about each product.
        - Order-Products Data: Shows which products were in each order.
        - Aisle Data: Groups products by aisle (category).
2. **Merging the Data Together**
    - Combining Datasets: Merged these files to create one complete dataset.
        - Merging Orders and Order-Products: Linked them by order_id so each order shows the products in it.
        - Adding Products Data: Merged with the products list on product_id to include product details.
3. **Sampling Data for Easier Analysis**
    - Stratified Sampling: Selected a sample that represents all customer types, using 99% of the data to keep it accurate.
        - Sample Size: We kept the sample large to cover most customer behaviors.
4. **Adding Aisle Names**
    - Mapping Aisle IDs to Names: Converted the aisle data into a dictionary (list of pairs) to make aisle names easier to read.
        - Aisle Mapping: Replaced aisle_id with the actual aisle name for better understanding.

.

5. Cross-Tabulating Users and Aisles

- **Creating a User-Aisle Table: Created a table to see how often each user buys from each aisle.**
  - **Row Normalization: Adjusted each row to compare user data more easily.**
  - **User-Aisle Matrix: Shows each user’s shopping habits across different aisles.**

.

6. Reducing Dimensions with PCA (Principal Component Analysis)

- **Simplifying Data: Used PCA to reduce the number of variables and keep the most important information.**
- **Normalization: Standardized values so they are all on a similar scale.**
- **Choosing Key Components: Selected components that capture the most important data patterns.**

7. Finding Clusters with K-Means

- K-Means Clustering: Used K-means to group customers into segments based on similar buying patterns.
  - Elbow Method: Chose the best number of clusters by looking at a graph of within-cluster variation.
  - Optimal Cluster Number: Found the right number of clusters that balanced detail with simplicity.

8. Visualizing Clusters

- Plotting Clusters: Made a scatter plot of clusters using the first two principal components to show group differences.
  - Understanding Cluster Differences: Noted how different clusters stand out, helping us understand each group’s buying habits.

9. Finding Popular Products in Each Cluster

- Top Products per Cluster: Listed the most popular products for each customer group.
  - Popular Product Analysis: Showed common products in each cluster to identify unique preferences.
  - Bundle Suggestions: Suggested product bundles based on popular items within each group.

10. Final Insights and Recommendations

- Customer Insights: Summarized key buying behaviors and patterns for each group.
- Purchase Patterns: Highlighted trends like when customers shop most and which products they often buy together.
- Product Bundling Ideas: Recommended bundles tailored to each group to boost sales and satisfaction