Capstone Project - Online Retail Customer Segmentation

Problem Statement

An online retail store wants to analyze customer purchasing behavior to improve marketing and retention. Using historical transaction data, the goal is to uncover useful customer insights and segment customers effectively.

Project Objective

The objective of this project is to segment customers using RFM (Recency, Frequency, Monetary) analysis, identify purchasing trends, and provide actionable insights that can guide loyalty programs, marketing strategies, and sales growth.

Data Description

The dataset contains 541,909 rows and 8 columns, representing 1 year of transaction data from a UK-based online retailer. Features include InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country.

Data Pre-processing Steps and Inspiration

1. Removed rows with missing CustomerID
2. Removed rows with Quantity <= 0 or UnitPrice <= 0
3. Converted InvoiceDate to datetime format
4. Added TotalPrice column = Quantity * UnitPrice
5. Filtered valid transactions for analysis

Choosing the Algorithm for the Project

KMeans Clustering was used on RFM variables after standard scaling to segment customers into behaviorally similar groups.

Motivation and Reasons For Choosing the Algorithm

KMeans is an effective unsupervised learning method for discovering hidden patterns in customer transaction data. It's scalable, easy to interpret, and well-suited for numerical RFM data.

Assumptions

1. RFM metrics effectively capture purchasing behavior
2. The latest invoice date reflects recency correctly
3. 4 clusters are optimal for customer segmentation based on trial and silhouette insights

Model Evaluation and Techniques

1. Boxplots were generated for each RFM metric by cluster
2. Cluster centers and summary statistics were calculated
3. Visual inspection and statistical comparison across clusters helped evaluate segmentation quality

Inferences from the Same

Clusters revealed groups like high-value loyal customers, infrequent but big spenders, recent one-time buyers, and dormant users. These segments can be targeted differently in marketing strategies.

Future Possibilities of the Project

1. Automate monthly segmentation refresh
2. Add features like product categories or return rate
3. Integrate segmentation into CRM for targeted offers

Conclusion

The RFM-based segmentation helped identify key customer groups. These insights can guide promotions, improve retention, and enhance customer satisfaction. The project is scalable and can be extended further with product-level personalization.

References

1. UCI Machine Learning Repository

2. Streamlit Documentation

3. scikit-learn Documentation

4. Online Retail Dataset - Kaggle