Capstone Project- Online Retail Customer Purchase Analysis

SANDHANA ROJA K

Table of Contents

1. Problem Statement

The online retail store aims to understand customer purchase patterns to enhance business strategy. This involves deriving insights from historical data and segmenting customers based on their purchasing behaviour.

2. Project Objective

The primary objective of this project is to analyse customer purchasing history and segment customers to provide actionable insights for the online retailer.

3. Data Description

The dataset `online_retail.csv` consists of 387,961 rows and 8 columns:

- **Invoice:** Invoice number

- **StockCode:** Product ID

- **Description:** Product Description

- **Quantity:** Quantity of the product

- **InvoiceDate:** Date of the invoice

- **Price:** Price of the product per unit

- **CustomerID:** Customer ID

- **Country:** Region of Purchase

4. Data Pre-processing Steps and Inspiration

Initial Data Exploration

- Loaded the dataset to examine its structure and contents.

- Performed exploratory data analysis (EDA) to understand data distribution, identify missing values, and detect outliers.

Data Cleaning

- **Handling Missing Values:** Removed or imputed missing values.

- **Correcting Data Entry Errors:** Ensured consistency in data entries.

- **Ensuring Appropriate Data Types:** Converted data types where necessary (e.g., dates).

Feature Engineering

- Created new features such as:

  - **TotalSpend:** Total spend per invoice (Quantity * Price).

  - **TotalSpendPerCustomer:** Aggregated spend per customer.

  - **PurchaseFrequency:** Number of purchases per customer.

5. Choosing the Algorithm for the Project

For customer segmentation, the **K-means clustering** algorithm was chosen due to its simplicity and effectiveness in handling large datasets.

6. Motivation and Reasons for Choosing the Algorithm

The **K-means** algorithm was selected for its:

- **Simplicity:** Easy to implement and interpret.

- **Scalability:** Efficient for large datasets.

- **Effectiveness:** Works well with continuous data, suitable for our numeric purchasing data.

7. Assumptions

- Each row in the dataset represents a unique transaction.

- The dataset is representative of the overall customer purchasing behaviour.

- No significant external factors (e.g., economic downturns) disproportionately affecting the dataset.

8. Model Evaluation and Techniques

Data Segmentation using RFM Analysis

- **Recency (R):** Time since the last purchase.

- **Frequency (F):** Total number of purchases.

- **Monetary (M):** Total spend.

Clustering with K-means

- Applied the Elbow method to determine the optimal number of clusters.

- Performed clustering and analyzed the characteristics of each cluster.

9. Inferences from the Same

- Identified distinct customer segments with varying purchasing behaviours:

  - **High-Value Customers:** Frequent buyers with high total spend.

  - **Occasional Shoppers:** Infrequent buyers with moderate spending.

  - **Low-Value Customers:** Infrequent buyers with low total spend.

- **Key Insights:**

  - High-value customers tend to purchase luxury items.

  - Promotions and targeted marketing can be designed for different segments to maximize engagement and revenue.

10. Future Possibilities of the Project

- **Enhanced Personalization**: Tailor marketing efforts based on customer segments.

- **Predictive Analytics:** Predict future purchasing behaviour and trends.

- **Loyalty Programs:** Design and implement loyalty programs to retain high-value customers.

11. Conclusion

The project successfully analyzed customer purchasing patterns and segmented customers based on their purchasing behaviour, providing valuable insights to the online retailer. These insights can inform strategic decisions to enhance customer satisfaction and business performance.

12. References

1.Data Source: Online Retail Dataset-Intellipaat

2."K-Means Clustering: A Review" by S. B. Kotsiantis and D. Kanellopoulos (2006)

3. "RFM Analysis: A Simple yet Effective Tool for Customer Segmentation" by W. J. Reinartz and V. Kumar (2000)