

Capstone Project DSAI

Problem Statement 2

1. Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm. Using the `online_retail.csv` dataset, we need to find useful insights about the customer purchasing history and segment the customers based on their purchasing behavior.

2. Project Objective

The objective of this project is to analyze customer purchasing patterns to provide actionable insights that can help the online retailer enhance their marketing strategies and improve customer satisfaction. Additionally, we aim to segment the customers based on their purchasing behavior to tailor personalized marketing campaigns.

3. Data Description

The `online_retail.csv` dataset contains 5,41,909 rows and 8 columns:

- **InvoiceNo:** Invoice number
- **StockCode:** Product ID
- **Description:** Product description
- **Quantity:** Quantity of the product
- **InvoiceDate:** Date of the invoice
- **UnitPrice:** Price of the product per unit
- **CustomerID:** Customer ID
- **Country:** Region of purchase

4. Data Pre-processing Steps

- **Loading the Data:** Import the dataset using pandas; use encoding 'latin1'
- **Handling Missing Values:** Drop rows with missing `CustomerID` since these are crucial for the analysis. `Description` column also has 1454 null values which are also be dropped as the total entries is above 4 lakhs.
- **Removing Negative Values:** Entries with negative `UnitPrice` were removed, while negative `Quantity` was retained to represent product returns.
- **Handling Outliers:** Outliers were identified using the IQR method for `Quantity` and `UnitPrice`.
- **Feature Engineering:** Create a `TotalPrice` column by multiplying `Quantity` and `UnitPrice`.

- **Date Conversion:** Convert the `InvoiceDate` column to datetime format.
- **RFM Metrics Calculation:** Compute Recency, Frequency, and Monetary values for each customer.
- **Data Normalization:** Scale the RFM values to bring them on a common scale.

5. Algorithm Selection

- **RFM (Recency, Frequency, Monetary) Model:** Used for customer segmentation based on purchasing behavior.
- **K-Means Clustering:** We have chosen K-Means clustering to segment the customers based on their purchasing behavior using the RFM model. K-Means is a popular and efficient algorithm for customer segmentation due to its simplicity and speed in handling large datasets.

6. Motivation and Reasons for Choosing the Algorithm

- **RFM Model:** It provides a simple yet effective way to understand customer behavior in terms of recency, frequency, and monetary value.
- **K-Means Clustering:** It is a popular and efficient clustering algorithm that helps in identifying distinct customer segments based on the RFM model.

7. Assumptions

- Customers with missing `CustomerID` are not considered in the analysis.
- The dataset is assumed to be representative of the overall customer purchasing behavior.
- Customers with similar RFM values have similar purchasing behavior.
- The purchasing behavior can be effectively captured using Recency, Frequency, and Monetary (RFM) metrics.
- The chosen number of clusters (4) is optimal for this dataset.
- Outliers in quantity and price have been appropriately handled during preprocessing.

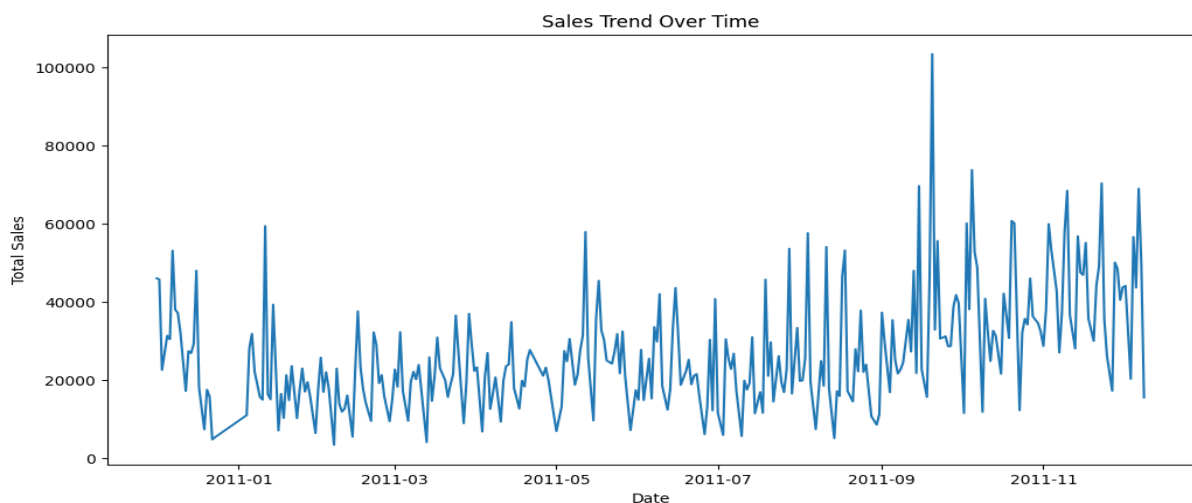
8. Model Evaluation and Techniques

- **Elbow Method:** Used to determine the optimal number of clusters for K-Means, which in this case is 4.
- **Cluster Analysis:** Evaluate the characteristics of each cluster based on mean RFM values.

Cluster	Recency	Frequency	Monetary	Characteristics
0	41.61	4.80	1478.52	Recent and moderate spenders
1	7.67	89.00	182181.98	Very recent and high spenders
2	247.95	1.81	453.49	Infrequent and low spenders
3	9.18	40.67	18441.96	Recent and high spenders

9. Inferences

- Insights on overall sales trends, popular products, and top customers.
- Distinct customer segments with specific purchasing behaviors.
- Actionable recommendations for each customer segment.



Sno.	Most Popular Products	Quantity	Sno.	Top Customers	Total Price
1	WORLD WAR 2 GLIDERS ASSTD DESIGNS	53215	1	14646	279489.02
2	JUMBO BAG RED RETROSPOT	45066	2	18102	256438.49
3	ASSORTED COLOUR BIRD ORNAMENT	35314	3	17450	187482.17

Cluster Insights (Customer Segments)

1. Cluster 0: Recent and Moderate Spenders

- Customers in this cluster have made purchases relatively recently (around 41 days ago on average).
- They purchase moderately frequently (about 5 times on average).
- Their average monetary value is moderate at approximately \$1,478.

2. **Cluster 1: Very Recent and High Spenders**

- These customers have made very recent purchases (around 8 days ago on average).
- They are very frequent buyers with an average frequency of 89 purchases.
- Their monetary value is extremely high, averaging around \$182,182.
- This cluster likely represents the most loyal and valuable customers.

3. **Cluster 2: Infrequent and Low Spenders**

- Customers in this cluster haven't made a purchase for a long time (around 248 days ago on average).
- They purchase infrequently (about 2 times on average).
- Their average monetary value is low at approximately \$453.
- This cluster may include customers who are at risk of churning or have already churned.

4. **Cluster 3: Recent and High Spenders**

- These customers have also made recent purchases (around 9 days ago on average).
- They are frequent buyers with an average frequency of 41 purchases.
- Their monetary value is high, averaging around \$18,442.
- This cluster represents valuable customers who are quite engaged with frequent purchases.

Recommendations

1. **For Cluster 0:**

- Implement targeted marketing campaigns to encourage more frequent purchases and increase the average order value.
- Offer loyalty rewards to incentivize repeat purchases.

2. **For Cluster 1:**

- Continue to engage these high-value customers with personalized offers and premium services.
- Consider exclusive access to new products or special events to maintain their loyalty.

3. **For Cluster 2:**

- Reactivate these dormant customers with re-engagement campaigns, such as special discounts or personalized reminders.
- Investigate the reasons for their inactivity and address any issues they may have encountered.

4. **For Cluster 3:**

- Provide these frequent buyers with upselling and cross-selling opportunities to maximize their spending.
- Enhance their customer experience with personalized recommendations and premium services.

10. Future Possibilities

- **Advanced Segmentation:** Explore additional clustering algorithms (e.g., DBSCAN, hierarchical clustering) to refine customer segments further.
- **Behavioral Analysis:** Incorporate more features such as product preferences, seasonal buying patterns, and feedback analysis for deeper insights.

- **Predictive Modeling:** Develop predictive models to forecast customer lifetime value (CLV) and identify potential high-value customers early.
- **Personalized Marketing:** Tailor marketing campaigns based on customer segments.
- **Customer Retention:** Implement retention strategies for high-value customers.
- **Product Recommendations:** Recommend products based on purchasing behavior.
- **Churn Prediction:** Identify and address potential churn risks.
- **Dynamic Segmentation:** Continuously update segments as new data comes in to reflect changing customer behavior.

By leveraging these insights and recommendations, the online retailer can gain valuable insights into customer purchasing patterns and effectively segment customers to improve business strategies. The retailer will also enhance their customer relationship management and ultimately drive higher sales and customer satisfaction.