Project Report - Online Retail

Problem Statement:

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

Project Objective:

1. Using the given data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
2. Segment the customers based on their purchasing behaviour.

Data Description:

Feature Name	Description
InvoiceNo	Invoice Number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the Invoice
UnitPrice	Price of product per unit
CustomerID	Customer ID
Country	Region of Purchase

Data Pre-processing Steps:

1. Removing all the null values from Dataset.
   For K-Means Algorithm:
2. Changing datatype of CustomerID to string.
3.
   i) For monetary value we calculate the amount by multiplying quantity and unit price.
   ii) Create a new dataframe and store the total amount grouping by each CustomerID and sum of the amount.
4. Create a new dataframe and store the no of transaction using count of InvoiceNo and grouping by CustomerID.
5.
   i) For recency, we convert the Invoice Date feature to DateTime format using pandas to_datetime function.
   ii) Create a new dataframe and store the latest date of transaction, grouping by each CustomerID and maximum date (latest transaction date in dataset).
   iii) Then calculate the no of days from the last transaction in entire dataset and the last transaction date of each CustomerID, and store the data in days.
6. Finally merge the datasets created for total amount, no of days in last purchase and no of transactions.
7. Outliers analysis and removal of outliers only from total amount and no of transactions.
   Notes:
   i) Outliers have not been considered for removing from no of days since last transaction because chances of outlying data being an anomaly is not there.
   ii) For outlier removal from total amount and no of transactions, IQR has been calculated using Q1 at 0.5 quantile and Q3 at 0.95 quantile so that only too much outlying data is only removed.
8. Scaling the total amount, no of days in last purchase and no of transactions data using Standard Scaler.

   For Apriori Algorithm:
2. Remove impurities from InvoiceNo using pandas str.replace function
3. Change the InvoiceNo datatype to int
4. Sort the dataframe by InvoiceNo and reset the index
5. Create a list of lists of Items in each Invoice Number

Choosing the Algorithm for the Project and reasons for choosing the Algorithm:

Strategy 1: To understand the recency, frequency and monetary value of the customer, based on their past transaction and we leverage these features as an input to **k-means clustering** algorithm. We will use these features to identify segments of customers.

After getting segments we will do the profiling of the customer. And identify which segment is basically the done most recent transactions, most frequent transactions, and high monetary valued transactions which can be used for developing marketing campaign strategies.

Strategy 2: To understand which products are frequently brought together, based on the transactions and leveraging these transactions data as an input to **apriori (association rule)** algorithm.

These lists of can be used to develop marketing strategy and manage the online display on website or application.

Assumptions:

Nil

Model Evaluation and Techniques:

Since both the applied algorithms to derive the insights i.e. cluster segments and association rules are unsupervised learning algorithms, hence as such no direct method is applied to evaluate the model performance.
But Silhouette score and Elbow method is applied for the K-Means customer segmentation to see the optimum K-value.

Inferences:

K-Means Clustering:

The customer segments created:

Label 0: Customers with label '1' are the highest monetary value buyers and most frequent buyers.

Label 1: Customers with label '1' are the most recent buyers.

Label 2: Customers with label '2' are neither the highest monetary value transaction doers, most frequent buyers nor most recent buyers.

Association Rule:

The list of items frequently brought together produced by apriori algorithm can be used as discussed in strategy 2 of choosing the algorithm.
Also, modified list can be produced by redefining the minimum support, minimum confidence and minimum lift (hyperparameters) of the algorithm as deemed necessary for a real life project.

Future Possibilities of the Project:

Labels generated by segmenting the customers, can be used for classification in future.