Customer and Market Analysis for an online retail

First name	Last Name	IIT Email
Bhavya Sree	Bindela	bbindela@hawk.iit.edu

1. Introduction

Introduce the background of your application and give me the motivations why you want to do that.

Customer and Market analysis for an online retail non-store is to explore different types of customers and their granular purchase behavior patterns using unsupervised learning techniques.

There will be different types of customers for any firm like some customers will be frequent buyers, some customers will be less frequent but buy high-value products, and so on. These are called different segments of customers.

Customer segmentation is to uncover the behavior of the customers based on their interactions with the firm and divide them into groups, so that the company can target their customers efficiently for future customer retention strategies.

Market Basket Analysis is to uncover the product purchase behavior of customers by extracting the association rules of the products. These will be helpful to understand purchase decisions made by the customers and the company can use the association rules in their future marketing strategies.

2. Data Sets

Briefly introduce your data sets, such as which application or domain the data belongs to, where did you collect it, how large it is, how many features there are, what is your target variable, and so forth

This dataset contains all the transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based and registered online retail non-store.

This dataset has been collected from UCI Machine Learning Repository.

The dataset can be found in the below link

https://archive.ics.uci.edu/ml/datasets/online+retail

Format of the dataset

Number of Instances: 541909

Number of Attributes: 8

Attribute	Description	Data Type
InvoiceNo	Invoice number	Character
StockCode	Product Code	Character
Description	Product Name	Character
Quantity	Quantity of each product per transaction	Number
InvoiceDate	Invoice Date and Time	Number
UnitPrice	Price of the Product per unit	Number
CustomerID	Customer Number	Number
Country	Country Name	Character

In this dataset, there is no target variable, as we will be doing unsupervised learning on the online retail dataset.

The whole project will be implemented in R.

3. Research Problems

List your research problems, that is, what kinds of the problems you want to solve.
You cannot simply say I want to explore the data and find the patterns
If you decide to work on a classification task, you must identify labels.
If your project is involved with multiple data mining tasks, you should clearly mention each problem and why you want to do that.
You should provide finer-grained research problems that can be solved by data analysis/mining techniques. If it is an implementation project, you should introduce the challenges in implementing or development and how you will evaluate them

1. Identify the customers into different groups based on their behavior.

   This dataset contains the data of all the transactions made in a store in a period of time. Based on the transactions made, I want to divide the customers into different groups. So that the company can target their customers efficiently for their future strategies.

   For example, the company can send some vouchers to the group of customers who have purchased high value from the store to show their gratitude and in turn to attract the major customers more.

   As this is unlabeled data and we are not sure how many groups the customers can be divided into. Therefore, we use the clustering technique for customer segmentation.

2. Identify the products that are purchased together often.

   Identifying the products that are purchased together often helps the owner to design the marketing strategies for the products.

For example, if most of the customers bought two items A and B together, the company can plan a marketing strategy like buy 2 quantities of A and get 1 quantity of B with a 30% discount. (considering A is the item with high cost and B is the item with low cost). As the customers are in need of both items, they prefer to buy 2 quantities of A rather than 1 quantity.

As this is transactional data, in order to find the association or correlation between the products, we use the association rule mining technique.

4. **Potential Solutions**

For each problem you list above, figure out feasible solutions, and introduce your plan to perform experiments

Load the data and preprocess the data.
Preprocessing includes

- Data Cleaning – Handling Missing data
- Data Transformation – Transformation of data types (if needed)
- Data Reduction – Reduce the data to meaningful format (if needed)
- Outliers Detection.

1. Identify the customers into different groups based on their behavior.

Recency, Frequency, Monetary (RFM) is a method that is mostly used for customer segmentation. In this method, we use time from the last purchase (Recency), a total number of purchases in a period of time (Frequency) and total amount spent in a period of time (Monetary) attributes.

At first, as a part of data Reduction, calculate Recency, Frequency, Monetary for every customer and form a new dataset.

On the basis of calculated Recency, Frequency, and Monetary values, customers can be segmented into different groups using Clustering techniques.

In this case, we use both K-means Clustering and Hierarchal clustering algorithms for customer segmentation.

At first, we randomly select the K-value and divide customers into K groups.

Then we calculate optimal K-value using elbow method and then redivide the customers into optimal K groups.

Optimal K-value is used both in K-Means clustering and hierarchical clustering for customer segmentation.

2. Identify the products that are purchased together often.

At first, group the products based on InvoiceNo. This gives us the list of products that are bought together.

Apply Apriori association rule mining algorithm on the dataset to find the association rules of the products i.e., to know the items that are bought together.

5. Evaluations

There could be multiple solutions for a same problem, You must figure out how to evaluate them and the details about your evaluations, for example, hold-out or N-folds evaluation?, which metrics you will use for evaluations.

There are no clear evaluations for unsupervised learning techniques. Clustering is good as long as it can serve the usage. Association rules are good as long as they have much support and we can interpret the results.

1. Identify the customers into different groups based on their behavior

   In clustering, inter-cluster distances should be maximum and intra-cluster distances should be minimum. In other words, the variance between the points inside the cluster and the total Within-Cluster-Sum of Squared Errors (WSS) should be less.

   Therefore, we use elbow method to find the optimal K values, where within-cluster-sum of squared Errors is calculated for different values of K.

   Based on the elbow method, we choose the value of K up to which WSS values diminishes more and adding another cluster won't diminish WSS much more.

   As we divide the clusters based on the optimal K-value, the clusters should be good to interpret the results.

   Furthermore, we will look into at least 10 customers of each cluster and try to figure out why they were put in a same cluster and check whether the customers of a cluster possess the same characteristics.

2. Identify the products that are purchased together often.

   We will be identifying association rules at minimum support of 0.01 and minimum confidence of 0.5. With this, we will get the rules with at least 50 percent confidence.

   The rules with high confidence will be considered as final rules and we interpret them.

6. Expected Outcomes

Introduce your expected outcomes for your project

1. Identify the customers into different groups based on their behavior

Customers will be divided into different groups such as the customers who purchased more in one group, the customers who purchased less in one group, frequent customers in one group, customers who do not visit for a long time in one group, and so on.

A plot will be drawn to show different customer clusters.

A dendrogram will be drawn to show the hierarchical relationship between the customers.

2. Identify the products that are purchased together often.

Inspect top 20 association rules of the products with high confidence.

References:
https://looker.com/blog/creating-actionable-customer-segmentation-models

---

After you finished your proposal, you should ask yourself the following questions:

1. Is my objective/goal is clear in the proposal? Am I able to decompose the high-level objectives into some practical problems which can be solved by my proposed solutions?

2. Can my solutions help me solve the proposed problems? why? are there any requirements on the data given by my solutions? Did I introduce the data? Can I used my solutions on the proposed problems?

3. Do I have a clear evaluation approach? Can I reasonably evaluate my solutions to tell that my solutions are good ones?

4. Can the reader understand every details in my proposal?

Some students just propose different techniques I taught in the class, such as classification, clustering, association rules, and put all the data mining tasks on the proposal. Well, what is your goal? are you sure these techniques can solve the problems you proposed? Please think deeper about it!! It is an examination about your understanding of the different knowledge and techniques. You should be able to figure out what techniques can be used to solve which problems, and which cannot.

5. You can choose to work on an easy project or a challenging one. Your project will be compared with others, and your grade will be affected by the degree of difficulty of your topic