Market Basket Analysis

Phase 5

—

2021506011

Madras Institute of Technology

B.Tech IT

Understanding Of The Problem

Objective: The primary aim of the analysis is to uncover hidden patterns and associations between products. This involves understanding the relationships between items that customers frequently purchase together.

Problem: The retailer wants to target customers with suggestions on the itemset that a customer is most likely to purchase. I was given a dataset containing data of a retailer; the transaction data provides data around all the transactions that have happened over a period of time.

Retailers will use results to grow in his industry and provide for customer suggestions on itemset, we will be able to increase customer engagement and improve customer experience and identify customer behavior. I will solve this problem by using Association Rules, a type of unsupervised learning technique that checks for the dependency of one data item on another data item.

Methodology: The analysis will utilize association analysis techniques, with a specific focus on the Apriori algorithm. The Apriori algorithm is a widely used method in MBA, allowing the identification of frequent itemsets and association rules in transactional data.

Business Goal: The ultimate goal of this analysis is to gain insights into customer purchasing behavior. By understanding what products are often bought together, the retail business can make informed decisions to enhance customer experience and increase sales.

Cross-selling Opportunities: One of the key objectives is to identify potential cross-selling opportunities. Cross-selling involves offering customers complementary products based on their purchase history, thereby increasing sales and customer satisfaction.

Actionable Recommendations: The project's output will be actionable recommendations. These recommendations will be derived from the patterns and insights discovered during the analysis. These suggestions could include optimizing product placement, creating targeted marketing campaigns, or improving customer service strategies.

Design Thinking

To tackle this problem effectively, we will follow a systematic approach, encompassing the following key stages:

1. Empathize:

Understanding the retailer's perspective and goals was paramount. By empathizing with their need to enhance customer engagement, the team gained insights into the significance of analyzing transaction data. Recognizing that the key to improving customer experience lies in understanding customer preferences and behavior, the team proceeded with a focus on uncovering item associations within the data.

2. Define:

Defining the problem was crucial. The challenge was articulated: discover item associations in transaction data to provide personalized item suggestions. Formulating the problem as a market basket analysis task, with the Apriori algorithm as the chosen technique, provided a clear direction for the analysis. This step ensured that the team had a well-defined problem statement to address.

3. Ideate:

The ideation phase involved exploring a range of techniques and libraries for market basket analysis. Several options were considered, but the Apriori algorithm stood out due to its efficiency in handling large datasets and its ability to generate meaningful association rules. By selecting this technique, the team laid the foundation for a robust solution that could effectively identify item associations from the transaction data.

4. Prototype:

Prototyping involved implementing the chosen Apriori algorithm using the arules and arulesViz libraries in R. This phase required extensive data preprocessing, where the dataset was cleaned, missing values were handled, and irrelevant items were filtered out. Relevant features like 'Year', 'Month', and 'Total price' were extracted, providing a richer context for the analysis. The data was transformed into a suitable format for market basket analysis, and the Apriori algorithm was applied with a defined minimum support threshold. This step resulted in the generation of association rules, capturing valuable insights into item relationships.

5. Test:

The testing phase was critical for evaluating the effectiveness of the discovered association rules. The team thoroughly analyzed the rules, considering metrics such as support, confidence, and lift. Visualization techniques, including scatterplots, were employed to gain a visual understanding of the associations. Business implications of these rules were meticulously assessed, identifying top items, high-

confidence associations, and strategic insights. These insights provided actionable recommendations for the retailer, paving the way for personalized marketing strategies and improved customer engagement.

Dataset Description

File Name:Assignment-1_Data.xlsx

Dataset Name:retaildata

BillNo: This attribute represents a 6-digit number assigned to each transaction. It serves as a unique identifier for individual transactions, allowing the retailer to track and manage sales effectively.

Itemname: This attribute contains the names of the products purchased in each transaction. It provides details about the items bought by customers, enabling the retailer to analyze popular products and customer preferences.

Quantity: Quantity indicates the number of units of each product purchased in a transaction. This numeric attribute helps in understanding the volume of items sold, aiding in inventory management and demand forecasting.

Date: Date signifies the day and time when each transaction occurred. This temporal information is crucial for analyzing sales trends over time, identifying seasonal patterns, and understanding customer behavior based on different times of the year.

Price: Price denotes the cost of each product. It is a numeric attribute representing the monetary value of items sold. Price information is essential for calculating revenue, profit margins, and average transaction value. It also allows the retailer to analyze customer spending patterns.

CustomerID: CustomerID is a 5-digit number assigned to each customer. It serves as a unique identifier for individual customers, enabling the retailer to

track customer behavior, preferences, and purchase history. Customer segmentation and personalized marketing strategies can be derived from this attribute.

Country: Country contains the name of the country where each customer resides. This attribute provides geographical information about the customers, allowing the retailer to analyze international sales, regional preferences, and target specific markets with tailored marketing campaigns

Phases of Development:

1. Data Source

The first step in any data analysis project is selecting an appropriate dataset. For this Market Basket Analysis, we will utilize the dataset available at the following link: [Market Basket Analysis Dataset. This dataset contains transactional data, including lists of purchased products.

The transaction data can be loaded from the provided Excel file using appropriate tools or libraries in R, such as readxl or openxlsx. The structure of the data can be explored to understand the attributes and their formats.This step is crucial for understanding the raw data and preparing for preprocessing..

2.Data Processing

Before conducting any analysis, we must ensure that the data is clean, structured, and in a suitable format for association analysis. Key tasks at this stage will include:

- Data Cleaning:

Columns with missing or irrelevant values, such as 'POSTAGE' or 'Adjust bad debt', were identified, and these rows were removed or handled appropriately. For instance, rows with negative or zero quantity were removed, and items that do not contribute to meaningful analysis can be filtered out. Missing values in essential columns like 'CustomerID' were addressed by either removing or imputing them based on the specific context of the dataset.

Data Transformation:

.'Year' and 'Month' was extracted from the 'Date' column to enable analysis based on temporal patterns.

'Total price' was calculated by multiplying 'Quantity' and 'Price', providing a comprehensive view of transaction values.

Grouping and Aggregating Data:

The preprocessed data was grouped based on relevant features such as 'Year', 'Month', and 'Country'. Aggregating data at this level is essential for meaningful market basket analysis. The grouped data was converted into a suitable format for the Apriori algorithm, where each row represents a transaction with items listed.

3. Association Analysis

Utilizing the Apriori Algorithm:

The Apriori algorithm was applied to the preprocessed data, and a minimum support threshold was set to filter out infrequent itemsets. This threshold determined the minimum occurrence of an itemset to be considered significant. Different support thresholds were experimented with to find the balance between generating meaningful rules and computational efficiency.

Generating Association Rules:

Association rules were generated based on metrics such as confidence and lift. Confidence measured the likelihood of one item being bought when another was bought, while lift measured the likelihood of items being bought together compared to being bought independently. The generated rules were evaluated to identify interesting and actionable patterns, focusing on rules with high confidence and lift, indicating strong associations between items.

Association Rules And Business Implications:

1.Top Association Rule

Rule: The rule states that there is a 5% likelihood that a customer who buys 'JUMBO BAG RED RETROSPOT' will also buy 'JUMBO BAG PINK POLKADOT'.

Implication: To capitalize on this association, the retailer should promote 'JUMBO BAG PINK POLKADOT' specifically to customers purchasing 'JUMBO BAG RED RETROSPOT'. By doing so, they can increase the sales of both items. This targeted marketing approach can be applied online or in-store, ensuring that customers are aware of the complementary product, thereby encouraging additional purchases.

2. High Confidence Rule:

Rule: This rule highlights that 82% of customers who bought "PINK REGENCY TEACUP AND SAUCER" also bought "GREEN REGENCY TEACUP AND SAUCER".

Implication: To enhance customer satisfaction and potentially increase sales, the retailer could bundle these two items together. Alternatively, they could offer discounts on one item when the other is purchased. This bundling strategy ensures that customers are incentivized to buy both items together, leading to increased revenue. Moreover, this approach enhances customer experience by offering them related items in a convenient package.

3. High Lift Rule:

Rule: Customers are 15.5 times more likely to buy "GREEN REGENCY TEACUP AND SAUCER" if they buy "PINK REGENCY TEACUP AND SAUCER".

Implication: To leverage this strong association, the retailer can create special promotions targeted at customers who buy "PINK REGENCY TEACUP AND SAUCER". These promotions can include discounts, exclusive offers, or bundled deals that feature "GREEN REGENCY TEACUP AND SAUCER". By doing so, the retailer can significantly boost sales of "GREEN REGENCY TEACUP AND SAUCER" among customers who have already shown interest in the pink variant. This focused marketing approach maximizes the likelihood of cross-selling between these items.

4. Strategic Insight:

The strategic insight emphasizes the importance of identifying and promoting items with the best combinations, such as different colors of "REGENCY TEACUP AND SAUCER". By offering various color options, the retailer can cater to diverse customer preferences, ensuring a wider appeal. These combinations can be strategically promoted through marketing campaigns, featured displays, or online recommendations. This approach not only maximizes cross-selling opportunities but also enhances customer satisfaction by providing them with a range of choices tailored to their tastes.

Conclusion

By following this structured approach, we will be able to uncover valuable insights into customer behavior and provide the retail business with concrete strategies to optimize their operations and increase profitability. This iterative process of data analysis and insights generation will enable data-driven decision-making and continuous improvement in the retail business's performance.