Gabriela Masak

D599 – Data Preparation and Exploration

Task 3: Market Basket Analysis

9/21/2024

Market Basket Analysis

**Part 1: Research Question**

In recent years, consumers have noticed the rising costs of almost every consumer good as inflation continues to permeate every corner of the economy. This has strained the wallets of consumers, whom have reacted by reducing expenditures that are viewed as less necessary in order to spend their budgets more effectively. Companies such as retail stores benefit from addressing their marketing tactics and strategies to realign with evolving purchasing habits. Techniques such as Market Basket Analysis can uncover relationships between purchasing patterns, leading to effective marketing strategies that can augment business and indirectly grow purchases in related categories or items.

Allias Megastore can use Market Basket Analysis to analyze product sales, customer behavior, geographic trends, and supply chain dynamics. These analyses can not only direct the company towards areas of business that can be improved, but can drive strategic decision-making with future sales, which will ultimately improve retail business performance. As inflation causes changes to consumer behavior, it is essential to grow and evolve with consumer habits to maintain and increase profitability. Specifically, the company can focus on generating product recommendations based on purchase history or item currently in the customer’s basket.

**Part 2: Market Basket Justification**

Market Basket Analysis is a data mining technique applied to transactional data. Allais Megastore has provided a dataset that includes order identification numbers, customer identification numbers, product names, quantities purchased, invoice dates, unit prices, total costs, countries, discounts, order pricing categorization, region, expedited shipping details, payment method, and subscription to customer rewards membership. By transforming this data, transactionalizing it, and applying a Apriori algorithm, the subsequent calculated values can uncover the association rules within the dataset. The Apriori algorithm leverages the assumption that customers purchasing a specific item will be more likely to purchase another item or set of items. For example, the transaction numbered 580263 includes several different items of the same “Danish Rose,” design category.

**Part 3: Data Preparation and Analysis**

Table 1: Categorical Variables and Label Encoding

|  |  |  |
| --- | --- | --- |
| Variable | Type | Encoding Method |
| Order Priority | Ordinal | Ordinal Encoding: 1-High, 2-Medium, 3-Low |
| Invoice Date | Ordinal | Ordinal Encoding: 1- Earliest … N-Most Recent |
| Region | Nominal | Label Encoding: 1- Northeast, 2-Southeast |
| Segment | Nominal | Label Encoding: 1- Corporate, 2-Consumer |
| Discount Applied | Nominal | One-Hot Encoding: 0-No, 1-Yes |
| Product Name | Nominal | Transaction (One-Hot) Encoding: Item Present (0-No, 1-Yes) |

Categorical variables can be further categorized as either ordinal or nominal variables, depending on whether the category can or cannot be organized by rank. Order priority can be ranked by low, medium and high, in the order of increasing priority. Invoice dates can be ranked oldest to newest. Region and segment, however, cannot be ranked in any sort of order and their organization is solely defined by their respective grouping. These distinctions can be used to wrangle data using the appropriate encoding method.

Three methods are available for encoding these variables: ordinal, label encoding, one-hot encoding. While ordinal encoding is most appropriate for ordinal variables due to maintaining rank during the transformation, nominal variables could be either label encoded or one-hot encoded. For region and segment variables, label encoding allowed for their categories to be grouped and defined by substituted numerical values. For the discount applied variable, the binary yes-or-no string data could be numerically substituted using one-hot encoding, however, only three transactions held any data for discounts, therefore the variable was deleted.

Data can further be manipulated and modified. The ability to transactionalize for market basket analysis by adding a column for each type of item and adding either a zero or one to the respective if the item was present or not within the order. Python and R make these various encoding activities instantaneous with functions to replace, merge, and sort data. A cleaned dataset can be found in attached file D599TransactionCleaned.csv.

The Apriori algorithm is applied to transaction dataset using the apriori and association\_rules methods in Python. The error-free results are displayed in Figure 2 below, with Figure 1 demonstrating that the code is error-free.

A screenshot of a computer program

Description automatically generated

Figure 1: Screenshot of Error-Free Python Code

A screenshot of a computer

Description automatically generated

Figure 2: Screenshot of Code Output (Support, Lift, Confidence)

While the Apriori algorithm generated several rules, the three most relevant are seen in Figure 3 and Table 2.

A screenshot of a computer program

Description automatically generated

Figure 3: Screenshot of Code and Results of Top Three Rules

Table 2: Table of Top Three Rules

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| antecedents | consequents | support | confidence | lift |
| frozenset({'SET20 RED RETROSPOT PAPER NAPKINS ', 'SET6 RED SPOTTY PAPER CUPS'}) | frozenset({'SET6 RED SPOTTY PAPER PLATES'}) | 0.088435 | 0.975 | 8.5995 |
| frozenset({'SET20 RED RETROSPOT PAPER NAPKINS ', 'SET6 RED SPOTTY PAPER PLATES'}) | frozenset({'SET6 RED SPOTTY PAPER CUPS'}) | 0.088435 | 0.975 | 7.9625 |
| frozenset({'SET6 RED SPOTTY PAPER PLATES'}) | frozenset({'SET6 RED SPOTTY PAPER CUPS'}) | 0.108844 | 0.96 | 7.84 |

Table 2 organizes the top three association rules, sorted by confidence and lift. The antecedents column highlights the presence of an item or group of item that is the primary purchased, with the consequents column highlighting the secondary purchased item or group of items. The relationship between the probability of purchasing these items or sets of items is displayed in rows support, confidence, and lift.

**Part 4: Data Summary and Implications**

Market Basket Analysis focuses on calculating several relevant factors linked to the statistics of consumer behavior. Support is the percentage of total transactions represented by the specific combinations of items. Lift represents the increased probability of the purchase of the secondary items given the purchase of the primary items. Confidence measures the likelihood that the secondary products are also purchased with the primary products. Combined these calculations can reveal key information about the buying patterns of consumers and used to generate recommendations that are more likely to convince the buyer to purchase more items while shopping, thereby increasing revenue and sales per order. Specifically, Megastore should use this data to focus on marketing their paper product division, as customers are likely to buy several more paper products when they buy one or two.

A Panopto audiovisual summary of the Python code, its functionality, and analysis can be found at the following link:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=2d62fe53-2293-4645-a8ad-b1fd015e4a20