D212 Advanced Data Acquisition

Performance Assessment: Task 3 – Market Basket Analysis

Jill Rivera

Western Governors University

Table of Contents

[I. Research Purpose 3](#_Toc116152544)

[A. Research Question 3](#_Toc116152545)

[B. Research Goal 3](#_Toc116152546)

[II. Analysis Justification 3](#_Toc116152547)

[A. Market Basket Analysis 3](#_Toc116152548)

[B. Example 5](#_Toc116152549)

[C. Assumption. 5](#_Toc116152550)

[III. Data Preparation and Analysis 5](#_Toc116152551)

[A. Data Cleaning. 6](#_Toc116152552)

[B. Apriori Algorithm 11](#_Toc116152553)

[C. Association Rules. 13](#_Toc116152554)

[D. Top 3 Rules 15](#_Toc116152555)

[IV. Data Summary and Implications 16](#_Toc116152556)

[A. Association Rule Significance 16](#_Toc116152557)

[B. Practical Significance 17](#_Toc116152558)

[C. Recommendations 17](#_Toc116152559)

[V. Works Cited 17](#_Toc116152560)

[Figure 1: Association Rule Diagram (Li, 2017) 4](#_Toc116649852)

[Figure 2: Example Transaction 5](#_Toc116649853)

[Figure 3: Importing Python Packages 6](#_Toc116649854)

[Figure 4: Loading the CSV File with Pandas 6](#_Toc116649855)

[Figure 5: Data Frame Structure 7](#_Toc116649856)

[Figure 6: First Five Rows of Raw Data 8](#_Toc116649857)

[Figure 7: Data Types 8](#_Toc116649858)

[Figure 8: Loop to Convert Data Frame into Lists 9](#_Toc116649859)

[Figure 9: List of Columns After Transformation 9](#_Toc116649860)

[Figure 10: Clean Data Frame Structure 10](#_Toc116649861)

[Figure 11: Clean Data Frame Example 10](#_Toc116649862)

[Figure 12: Exporting Clean Data 11](#_Toc116649863)

[Figure 13: Defining Output Variables 11](#_Toc116649864)

[Figure 14: Apriori Algorithm 4/1000 Lift 12](#_Toc116649865)

[Figure 15: Apriori Algorithm 3/1000 Lift 13](#_Toc116649866)

[Figure 16: Defining Items 13](#_Toc116649867)

[Figure 17: Support Parameters 14](#_Toc116649868)

[Figure 18: Confidence Parameters 14](#_Toc116649869)

[Figure 19: Lift Parameters 15](#_Toc116649870)

[Figure 20: Pruning 15](#_Toc116649871)

[Figure 21: Top Three Rules 16](#_Toc116649872)

[Figure 22: Top Association Rule 17](#_Toc116649873)

# **Research Purpose**: A telecommunications company is attempting to understand their customer base in a more detailed manner. The company has a data set containing customer transactions for products that they offer to customers and would like to figure out which products the customers’ favor. A market basket analysis on the data set will be conducted to produce this information.

## Research Question: The analysis will answer the question of: What group of products are the customers most likely to purchase?

## Research Goal: The goal of this analysis is to find a group of products that the customers want to buy. This is a great way to evaluate the products the company offers and determine their value for the business.

# **Analysis Justification**: In data analysis, finding patterns in data is common theme. Pattern detection can enable the evaluator to see trends in the data and give recommendations for changing or strengthening that pattern. This analysis will use the market basket analysis method to predict patterns in the telecommunication component data gathered.

## Market Basket Analysis: This analysis method searches for product combinations that are frequently purchased together. This method benefits from using larger data sets so that more transactions can be compared and produce more accurate results. Market basket analysis uses association rules to calculate the relationship between products. The product is a rule, which consists of an antecedent and a consequent. These are just lists of items that are purchased by a customer, with the antecedent occurring first and the consequent being a co-occurrence in the transactions. An itemset or set of items in the antecedent and consequent, is produced for every rule.

## The components of the association rule are calculated to analyze the itemset. Those components are support, confidence, and lift. Support is how often the itemset appears in the data. Support is calculated by dividing the number of transactions with the itemset by the total number of transactions. Confidence calculates if the itemset is popular with individual or combined sales. Confidence is calculated by dividing the combined transactions by the individual transactions. Lift is the ratio between confidence and support and tells the likelihood of both items being bought together. (Li, 2017) A diagram of how the association rules of market basket analysis is shown to visualize these calculations.

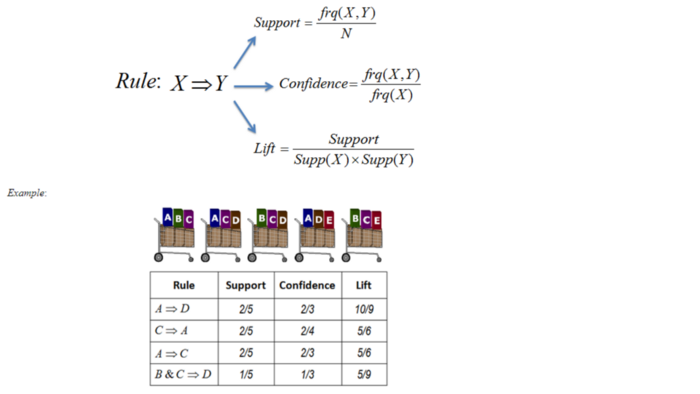


Figure : Association Rule Diagram (Li, 2017)

## Example: Using the iloc function in Python, an example transaction is displayed. This is transaction number 25 in the data set, and you can see that the customer purchased six items. The remaining item slots are listed as NaN since they are empty.

## 

Figure 2: Example Transaction

## Assumption: The biggest assumption of market basket analysis is that there will be overlap of items in the transactions. If the data set were to consist of transactions with no overlap of items, this analysis technique would not produce any results. There must be multiple transactions with the same group of products to calculate the association rules.

# **Data Preparation and Analysis**: The market basket analysis is performed in a Jupyter Notebook using the Python 3 programming language.

## Data Cleaning: The data set is manipulated and cleaned prior to performing the market basket analysis. First, the libraries needed are imported.

Figure 3: Importing Python Packages

## Using Pandas, the CSV file is read and the iloc function shows an example transaction.

## 

Figure 4: Loading the CSV File with Pandas

## The data set is evaluated using the info, format, and shape commands. This data set consists of 7501 rows of data and 20 data attributes. Table Description automatically generated

Figure 5: Data Frame Structure

## The head command shows the first five rows of the data set. Text, table Description automatically generated with medium confidence

Figure 6: First Five Rows of Raw Data

## Using the dtypes command, all attributes are shown to be object in type. Table Description automatically generated

Figure 7: Data Types

## To perform the analysis, the data needs to be set up in lists so that each transaction shows all the items purchased on the same row. A simple loop is created to accomplish this. Text Description automatically generated

Figure 8: Loop to Convert Data Frame into Lists

## There are now 120 columns or transactions in the data frame, and using the columns command they are listed. Text Description automatically generated

Figure 9: List of Columns After Transformation

## Now, the null values can be dropped using the drop command to create the final clean data set of 7501 rows and 119 columns.

Figure 10: Clean Data Frame Structure

## The info and head commands show the cleaned data frame’s first four columns. Table Description automatically generated

Figure 11: Clean Data Frame Example

## This data frame is exported into a CSV file labeled “D212Task3Clean.CSV” using Pandas and is attached to this submission.

Figure 12: Exporting Clean Data

## Apriori Algorithm: To perform the analysis, first the outputs of antecedents, consequents, support, confidence, and lift are defined. A picture containing text Description automatically generated

Figure 13: Defining Output Variables

## Using MLXtends frequent\_patterns (Raschka, 2022) package, the itemsets and rules are identified using the minimum support of 4/1000 and sorted by lift.

Figure 14: Apriori Algorithm 4/1000 Lift

## This process is re-run using the parameters of 3/1000 for the minimum support and the minimum confidence of 3/10. Graphical user interface, text, application Description automatically generated

Figure 15: Apriori Algorithm 3/1000 Lift

## In reviewing the results, the items of “Dust-off Compressed Gas 2 Pack” and “VIVO Dual LCD Monitor Desk Mount” are identified the most, so these items are defined.

## Running the itemset on these two items shows that they are bought together 448 times in this data frame.

## Text Description automatically generated

Figure 16: Defining Items

## Association Rules: Running the Apriori algorithm again with the minimum support and confidence of 1/100 and sorted by support is run. It shows that there is a support value of 0.06 in this data set, which is a very low score. Table Description automatically generated

Figure 17: Support Parameters

## These same parameters are run again but are sorted by confidence score. It shows a confidence score of 0.51, which is a mid-range value since the range is between 0 and 1. Table Description automatically generated

Figure 18: Confidence Parameters

## The algorithm is run again and sorted by lift score. This shows a stronger lift value of 3.29 with the combination of the 128GB SanDisk card and the 64GB SanDisk card. Graphical user interface, text, application Description automatically generated

Figure 19: Lift Parameters

## Top 3 Rules: To identify the top three rules for the analysis, first the data is pruned to count the number of times the itemset occurs in the data frame to weed out infrequent itemsets. There are now 432 rules to analyze. Graphical user interface, text, application Description automatically generated

Figure 20: Pruning

## Running the analysis and sorting for lift, the top three rules are summarized. The best rule produced has a lift value of 6.116, which is a strong indication that the itemset would be purchased together. This itemset includes the “Dust-Off Compressed Gas 2 pack”, “Ankler 2-in-1 USB Card Reader”, and “FEIYOLD Blue light Blocking Glasses.”Graphical user interface, text Description automatically generated

Figure 21: Top Three Rules

# **Data Summary and Implications**: The final market basket analysis, rule significance, and recommendations are discussed.

## Association Rule Significance: The top rule that the analysis produced is shown in the figure. From this, the support for this rule is a value of 0.004 can be derived. This is calculated by dividing the number of transactions with “Dust-Off Compressed Gas 2 pack”, “Ankler 2-in-1 USB Card Reader”, and “FEIYOLD Blue light Blocking Glasses” by the total number of transactions.

## The confidence value of this itemset is calculated at 0.403, which equates to 40.3 percent of the transactions that had “Dust-Off Compressed Gas 2 pack” and “Ankler 2-in-1 USB Card Reader” as the antecedent also had the “FEIYOLD Blue light Blocking Glasses” as the consequent.

## The lift value of this rule is strong with a score of 6.116. This states that a customer is six times more likely to purchase the “FEIYOLD Blue light Blocking Glasses” if they are also buying the “Dust-Off Compressed Gas 2 pack” and “Ankler 2-in-1 USB Card Reader” together.



Figure : Top Association Rule

## Practical Significance: Market basket analysis is typically performed on very large data sets; however, this data set was only 7501 transactions. This made the calculations limited and caused a low support value to be calculated for even the top rules. Since the confidence level was calculated at 40.3 percent for the top rule, the rule should still be considered a good itemset combination and strong association.

## Recommendations: The product placement for “FEIYOLD Blue light Blocking Glasses” should be next to the computer accessories. Specifically, the blue light glasses should be placed next to the “Dust-Off Compressed Gas 2 pack” and “Ankler 2-in-1 USB Card Reader.” That same display should be next to the selection of memory cards. This will increase the likelihood of these items being purchased together and increasing sales. This will also improve customer service for the ease of finding what they need in the store.

# Works Cited

Li, S. (2017, September 24). *A Gentle Introduction on Market Basket Analysis*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

Raschka, S. (2022, October 12). *mlxtend: apriori*. Retrieved from GitHub: http://rasbt.github.io/mlxtend/api\_subpackages/mlxtend.frequent\_patterns/