­

**Data Mining II, Task III: Churn Data**

**Association Rules & Lift Analysis**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. Keiona Middleton

September 25, 2022

**Data Mining II, Task III: Churn Data – Association Rules & Lift Analysis**

In this paper, I will use the provided data set containing customer purchase data from a fictional telecommunications company. I will use market basket analysis to determine relationships between items purchased together to gain a better understanding of our customers. The goal is to determine which items are popular together to inform business decisions.

# Part I: Research Question

## A1. Proposal of Question

The key question I would like to answer is “which items are frequently purchased together” so that we might better understand our customers and their purchase behaviors.

## A2. Defined Goal

I will perform a Market Basket Analysis (MBA) to determine which items are frequently purchased together. The resulting association rules can offer insights into product placement and sales/marketing strategy.

# Part II: Market Basket Justification

## B1. Explanation of Market Basket

Market Basket Analysis (MBA) is a relatively simple concept with a broadly understood application – shopping at the grocery store (however, the concept can apply to other scenarios). Knowing which products are frequently purchased together can help inform business decisions such as discounting items, offering bundles, or product placement. (Jabeen, 2018) These relationships are called **association rules** and essentially follow the structure of conditional probability – *If* ***A*** *then* ***B***, or *antecedent implies consequent*. By offering incentives or simply placing related products near each other, customers may make an additional purchase (or “impulse buy”) who may not have originally intended to purchase that item. (Smartbridge, 2022)

The *apriori* function in R can efficiently mine the association rules for a given data set and provide measurements of **support, confidence,** and **lift**. These will be defined in greater detail in a later section, but in short, support is how frequent the relationship occurs overall, confidence is a measure of accuracy, and lift is a measurement of likelihood (i.e.: for the statement “customers who bought cereal are twice as likely to also buy milk,” lift = 2). The algorithm searches a data set for the relationships and complies a report of association rules and their metrics. An analysis of the metrics can inform business decisions based on the corresponding rules.

## B2. Transaction Example

The screenshot below is from the cleaned data set that has been submitted with this paper and contains the first 10 purchases. Each row represents a single transaction, with the first row containing column headers representing a count of items purchased. Some rows have many items and do not fit neatly into an image (for example, Row 2 contains the first transaction and consists of 20 items). Rows 3-5 and 7-11 have three or fewer items and are easily captured in the image below.

Graphical user interface, text, application

Description automatically generated

***Figure 1: Transaction Examples***

## B3. Market Basket Assumption

Market basket analysis assumes the data being analyzed contains independent transactions (meaning no erroneous duplicated transactions) and assumes there are trends to be found in customer purchases.

# Part III: Data Preparation and Analysis

## C1. Transforming the Dataset

The data set contains blank rows between transactions, so these rows were dropped from the data set. Please see the submitted csv file containing the data set after this step was completed.

## C2. Code Execution

Please see the attached PDF containing the code execution report.

## C3. Association Rules Table

There are 18 total rules found from the transaction data (Figure 2). The individual rules and their corresponding metrics can be seen in Figure 3.

Text

Description automatically generated

***Figure 2: Apriori Execution***

Text

Description automatically generated

***Figure 3: Association Rules Table***

## C4. Top Three Rules

The top three rules, ordered by lift, are provided in the figure below. The combination with the highest lift is when a customer purchases Dust-off Compressed Gas 2 Pack and SanDisk Ultra 64GB Card, they are 2.395 times as likely to also purchase a VIVO Dual LCD Monitor Desk Mount.

Text

Description automatically generated

***Figure 4: Top Three Rules with Metrics***

# Part IV: Data Summary and Implications

## D1. Significance of Support, Lift, and Confidence Summary

Support, lift, and confidence were introduced briefly in an earlier section and referenced throughout. Now that the *apriori()* output is available, context can be provided to these values based on the rules.

Support for a rule is defined as the proportion of transactions that contain the value(s) in the *lhs* column (A) and the value(s) in the *rhs* column (B). The formula in terms of probability can be written as

(Larose & Larose, 2019)

In terms of the first rule from Figure 4 above, this can be interpreted as:

“Customers buy Dust-off Compressed Gas 2 pack and SanDisk Ultra 64GB card together with VIVO Dual LCD Monitor Desk mount at 1.7% of all purchases.”

Confidence is the conditional probability that a customer will buy *rhs* (B) given they have “already purchased” *lhs* (A) (not meaning a separate transaction, but that they have the item(s) in their basket and will purchase it/them). The formula for calculating confidence is

(Larose & Larose, 2019)

In terms of the first rule, confidence can be interpreted as “Customers buy Dust-off Compressed Gas 2 pack and SanDisk Ultra 64GB card also buy VIVO Dual LCD Monitor Desk mount 41.69% of the time.”

Lift is a bit more complicated, but it can be summarized as the probability the rule (*confidence*) vs the probability of the second purchase (*consequent*) happening at random chance (called “prior proportion of the consequent”). The formula reads

(Larose & Larose, 2019)

The first rule has a lift value of 2.395 and, in context, means that “Customers buy Dust-off Compressed Gas 2 pack and SanDisk Ultra 64GB card together are 2.395 times as likely to also by VIVO Dual LCD Monitor Desk as other customers.”

While all metrics are important, generally ***lift*** is the best place to start when analyzing the results. When lift is greater than 1, it implies that the antecedent increases the chances the consequent will occur. When it is less than 1, the antecedent implies the customers are less likely to purchase the consequent. When lift is near or exactly 1, there is no relationship between the antecedent and the consequent. (Smartbridge, 2022) When analyzed in context with support and confidence, we can gain a more detailed view of customer purchase behavior.

## D2. Practical Significance of Findings

The significance of the metrics for the first rule are outlined in the previous section. For the remaining Top Three Rules, the context is as follows:

Customers who purchase an Apple Pencil and SanDisk Ultra 6 4GB Card

* buy them together with Dust-off Compressed Gas 2 pack at 1.01% of all purchases,
* also buy Dust-off Compressed Gas 2 pack 50.67% of the time,
* are 2.126 times as likely to buy Dust-off Compressed Gas 2 pack.

Customers who purchase a SanDisk Ultra 6 4GB Card and Screen Mom Screen Cleaner kit

* buy them together with Dust-off Compressed Gas 2 pack at 1.11% of all purchases,
* also buy Dust-off Compressed Gas 2 pack 50.3% of the time,
* are 2.112 times as likely to buy Dust-off Compressed Gas 2 pack.

## D3. Course of Action

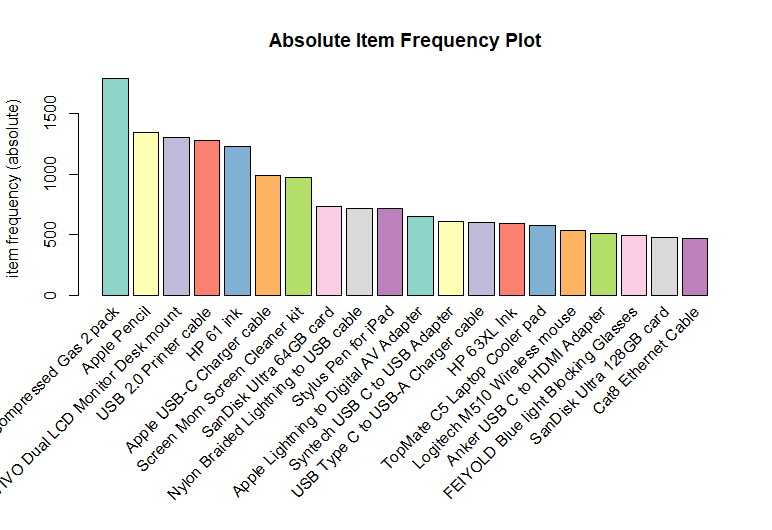
The Top Three Rules all contain the item “Dust-off Compressed Gas 2 pack” with only the first using it as the antecedent part of the rule. All remaining 17 rules have Dust-off Compressed Gas 2 pack as the consequent in the relationship. This overwhelming presence of the Dust-off Compressed Gas 2 pack implies it is a keystone product for the store – one that is vital to its success. Some actions would be to ensure we always have a sock of these items as they are the most frequent item purchased by our customers, to consider offering bundles or discounts for purchases involving the product, and/or to ensure the product is in a high-visibility or high-traffic location in the store. (Smartbridge, 2022)

While the support values are low, this is likely due to the large quantity of individual items offered for sale. This means there are many possible antecedents while the consequent tends to be consistently the Dust-off Compressed Gas 2 pack. The summary of the transaction object (*tr*) (Figure 5) and the Item Frequency Plot (Figure 6) shows that this is the most sold item by over 400 items sold.

Text

Description automatically generated

***Figure 5: Summary of Transaction Object***



***Figure 6: Absolute Item Frequency Plot***

# Part V: Attachments

## E. Panapto Recording

Please see provided video link to Panapto presentation.

## E. Web Sources

No third-party code was used in this script.

## F. Sources

**References**

Jabeen, H. (2018, August 21). *R market basket analysis using apriori examples*. DataCamp. Retrieved September 25, 2022, from https://www.datacamp.com/tutorial/market-basket-analysis-r

Larose, C. D., & Larose, D. T. (2019). *Data Science using python and R*. Wiley.

Smartbridge. (2022, July 5). *Market basket analysis 101: Anticipating customer behavior*. Smartbridge. Retrieved September 27, 2022, from https://smartbridge.com/market-basket-analysis-101/