**Market Basket Analysis of Churn Dataset**

Kelseyann Wright

Student ID: 00565814

College of Information Technology, Western Governors University

Data Mining II - D212

Kesselly Kamara

January 15, 2024

**Market Basket Analysis of Churn Dataset**

**A1. PROPOSAL OF QUESTION**

The proposed question is, “Which telecom items are associated with each other in customer transactions?”. In other words, “which items are frequently purchased together”?

**A2. DEFINED GOAL**

The goal of this analysis is to use market basket analysis to identify items that are associated with each other in customer transactions, or frequently purchased together, to optimize item sales.

**B1. EXPLANATION OF MARKET BASKET**

Market basket analysis (MBA) is a popular machine learning method. MBA identifies useful association rules to discover relationships between different transactions based on the co-occurrence of data values such as items, products, or services (Lim, 2022). For example, MBA could be used to identify grocery items that are frequently purchased together, such as cereal and milk, and use that information to optimize shelf placement or sale promotions.

The measures of support, confidence, and lift are used to discover association rules (Lim, 2022). Support is the probability of an event, such as the purchase of a specific item or multiple items together and is measured as the proportion of transactions in which that event occurs (Lim, 2022). Confidence is the probability of a consequent given an antecedent, or the probability of event B given event A (Lim, 2022). For example, it could be the probability that a customer purchases milk given that they purchase cereal. It is calculated as *P(A and B) / P(A),* where *P* is probability calculated as the proportion of transactions meeting the specified condition. Lift is similar to confidence, except lift controls for how popular both items are (Lim, 2022). Lift is calculated as *P(A and B) /[P(A) \* P(B)].* A lift value greater than 1 indicates that there is an association between the items A and B, with higher values indicating stronger associations (Lim, 2022). These three measures are used in the Apriori algorithm during MBA to efficiently identify associations in large transactional datasets and can be used to assess the strength of those associations. The Apriori algorithm identifies relationships between items, extending to larger and larger item sets only if a specified minimum support threshold is met (Lim, 2022).

The expected outcome of the market basket analysis is a set of rules that indicate associations between items sold to customers. These rules identify items that are often purchased together in a single transaction and can suggest a causal relationship between the purchases.

**B2. TRANSACTION EXAMPLE**

A transaction in the dataset is the purchase of items by a customer. Each row of the initial dataset represents a transaction, with each column containing an item purchased. A screenshot of one of the transactions is provided below.

A screen shot of a computer

Description automatically generated

**B3. MARKET BASKET ASSUMPTION**

One assumption of market basket analysis is that there are associations between product purchases. In other words, certain products are purchased together in multiple transactions. This assumption is essential in order for the Apriori algorithm to establish rules.

**C1. TRANSFORMING THE DATA SET**

The transformed dataset is attached (teleco\_market\_basket\_clean\_T3.csv). The steps used to transform the dataset were:

1. Remove rows consisting of only “nan” values.
2. Convert data into a list of lists (a list of transactions, with each transaction containing a list of items purchased).
3. Use the “TransactionEncoder” function from the mlxtend library (Raschka 2018) to encode the list of transactions into a Boolean NumPy array.
4. Return the array to a data frame, with each row representing a transaction and each column representing a specific item purchased. Data values are “true” and “false” Boolean values, indicating if the item indicated in each column was purchased in the transaction.
5. Remove the column of “nan” values.
6. Export to CSV (teleco\_market\_basket\_clean\_T3.csv).

**C2. CODE EXECUTION**

The code used to generate association rules with the Apriori algorithm is attached (D212\_T3.py). A screenshot of the executed code is provided below.

A screen shot of a computer

Description automatically generated

**C3. ASSOCIATION RULES TABLE**

The complete association rules table is attached (ruletable\_T3.csv). A preview of the first few rows of the table is provided below.

A screenshot of a computer

Description automatically generated

**C4. TOP THREE RULES**

A black screen with white text

Description automatically generatedA screenshot of the top three rules (according to lift) is provided below. The rules can be interpreted as, if [antecedent] is purchased, then [consequent] is purchased. For example, the top rule indicates that if a customer purchases the VIVO Dual LCD Monitor Desk mount, they are likely to also purchase the Dust-Off Compressed Gas 2 pack.

**D1. SIGNIFICANCE OF SUPPORT, LIFT, AND CONFIDENCE SUMMARY**

The top three rules were identified according to lift. Lift is important because it considers the commonality of the purchases of the antecedent and consequent individually, helping determine if the co-occurrence of item purchases is due to actual item association or just item popularity. Looking at the top three rules from section C4, you can see all values of “lift” are greater than 1, indicating that there is a true association between the items represented in those rules. Confidence and support also provide important information. Confidence tells us the probability that the consequent will be purchased given the antecedent. For the number one rule listed in section C4, there is a 34% confidence that the dust-off compressed gas two pack will be purchased if the VIVO dual LCD monitor mount is purchased. For the second and third rules, the confidence is 25% and 32%, respectively. Support is the probability of an event. The rules table includes the support of the antecedent and consequent individually, as well as the support of events occurring together. Support is important because it provides an idea of the frequency the transaction(s) and therefore popularity of items. For example, for the number one rule in section C4, the VIVO dual LCD monitor mount is purchased in 17% of transactions, while the dust-off compressed gas two pack is purchased in 24% of transactions. The support of the overall rule based on both the antecedent and consequent is only 6%. These statistics are important for companies to consider when basing decisions off the analytical results. Companies may establish thresholds for the lift, support, and confidence values necessary to evoke actions based on the rules identified in the market basket analysis.

**D2. PRACTICAL SIGNIFICANCE OF FINDINGS**

The analysis is significant for many data-driven business decisions. Identifying item association rules can be used to optimize product placement on physical store shelves and in e-commerce lists. In addition, the rules can be used to create targeted promotions and sales to increase sales.

**D3. COURSE OF ACTION**

Based on the item associations identified, I would recommend that associated items (according to the rules in C4, be placed next to each other on store shelves. If an e-commerce site is available, when the antecedent products are added to the customer cart, I would implement a pop-up that suggests the purchase of the associated consequent item(s). I would recommend piloting this approach with the top three rules identified, then assess the success of these changes before implementing for more rules that may have weaker item associations.

**E. PANOPTO VIDEO OF CODE AND E1. PANOPTO VIDEO OF PROGRAMS**

A Panopto video providing full documentation of the code and description of programs used is attached. The link is also provided here: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f8604355-77aa-4152-a3a1-b0fb01179253>

**F. SOURCES OF THIRD-PARTY CODE**

No sources of third-party code were used.

**G. SOURCES**

Raschka, S. (2018) MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw 3(24).

Lim, Y. (2022). Data Mining: Market Basket Analysis with Apriori algorithm. Towards Data Science. https://towardsdatascience.com/data-mining-market-basket-analysis-with-apriori-algorithm-970ff256a92c

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.