Task 3: Market Basket Analysis – D212

Abhishek Aern

Western Governor University

Contents

Part I - Research Question

A1: Question for analysis

What are the frequently co-purchased items by customers in the provided telecom company dataset? Also, What are the relationships and associations between different items in customer transactions?

A2: Goal of analysis

The goal of exploring the relationships and associations between different items in customer transactions is to gain insights into customer preferences, uncover cross-selling opportunities, and identify patterns of co-occurrence. By understanding which items are frequently purchased together, we can identify potential bundling or package offerings, optimize product placement, improve recommendation systems, and enhance the overall customer experience.

By addressing the research question, the telecom company can leverage the findings to make data-driven decisions and strategic recommendations to increase customer satisfaction, drive sales, and maximize business growth.

Part II - Method Justification

B1: Explanation of Market Basket Analyzes

(Kadlaskar, 2021) Market Basket Analysis is a data mining technique used to uncover relationships and associations between items in a dataset of customer transactions from the telecom company dataset. This method identifies customer buying habits by finding associations between the different items that customers place in their shopping baskets or purchase history.

I am using the Apriori algorithm which is a popular and widely used algorithm in the field of Market Basket Analysis and association rule mining. It is designed to discover frequent itemsets and generate association rules from transactional datasets.

The association rule or also known as the "if-then" rule, is a fundamental concept in Market Basket Analysis. It represents a relationship or association between items in a dataset of customer transactions. An association rule is structured as "IF {antecedent} THEN {consequent}", indicating that if certain items (antecedent) are present in a transaction, then there is a likelihood or tendency for certain other items (consequent) to be present as well.

The antecedent and consequent parts of an association rule are sets of items. The antecedent represents the items that are present on the left-hand side of the rule, while the consequent represents the items on the right-hand side. The rule is considered significant if the presence of the antecedent items implies the presence of the consequent items with a certain level of confidence.

The generated association rules are evaluated based on various metrics such as support, confidence, and lift. Support measures the frequency of the rule in the dataset, confidence measures the conditional probability of the consequent given the antecedent, and lift indicates the strength of the association between the antecedent and the consequent. These metrics help in interpreting the significance and usefulness of the association rules.

The outcomes of Market Basket Analysis provide insights into customer behavior, purchasing patterns, and item associations. Businesses can utilize these insights to make informed decisions such as product bundling, cross-selling, targeted marketing campaigns, store layout optimization, and inventory management (Kadlaskar, 2021).

B2: Transaction Example

Here are some examples of transactions in a dataset:

```
print ("Transaction 1:" , transactions[0])
print ('**********************************************************************')
print ("Transaction 5:" , transactions[4])
print ('**********************************************************************')
print ("Transaction 30:" , transactions[29])
```

```
Transaction 1: ['Logitech M510 Wireless mouse', 'HP 63 Ink', 'HP 65 ink', 'nonda USB C to USB Adapter', '10ft iPHone Charger Ca
ble', 'HP 902XL ink', 'Creative Pebble 2.0 Speakers', 'Cleaning Gel Universal Dust Cleaner', 'Micro Center 32GB Memory card',
'YUNSONG 3pack 6ft Nylon Lightning Cable', 'TopMate C5 Laptop Cooler pad', 'Apple USB-C Charger cable', 'HyperX Cloud Stinger H
eadset', 'TONOR USB Gaming Microphone', 'Dust-Off Compressed Gas 2 pack', '3A USB Type C Cable 3 pack 6FT', 'HOVAMP iPhone char
ger', 'SanDisk Ultra 128GB card', 'FEEL2NICE 5 pack 10ft Lighning cable', 'FEIYOLD Blue light Blocking Glasses']
**********************************************************************
Transaction 5: ['Dust-Off Compressed Gas 2 pack', 'Screen Mom Screen Cleaner kit', 'Moread HDMI to VGA Adapter', 'HP 62XL Tri-C
olor ink', 'Apple USB-C Charger cable']
**********************************************************************
Transaction 30: ['Nylon Braided Lightning to USB cable', 'VIVO Dual LCD Monitor Desk mount', 'Creative Pebble 2.0 Speakers', 'D
ust-Off Compressed Gas 2 pack']
```

In the above screen print, each transaction represents a customer purchase, and the items in squared brackets [ ] represent the products bought in that transaction. These transactions serve as the input for performing Market Basket Analysis to uncover associations and patterns in customer purchasing behavior.

B3: Assumptions of MBA

Here are some of the assumptions of Market Basket Analysis (Hua) -

Market Basket Analysis assumes that there is a relationship or association between items in a transaction. It assumes that the purchase of one item may influence the purchase of another item.

The analysis assumes that items are purchased independently of each other, meaning that the occurrence of one item does not affect the occurrence of another item. This assumption allows for the use of probabilistic measures like support and confidence in determining the strength of associations.

Part III - Data Preparation and Analysis

C1: Transforming the dataset

To transform the data for market basket analysis I am performing the following steps (WGU, Resources, n.d.) –

- The provided dataset has some rows with all null values so dropping all of them using the .dropna method. The axis=0 parameter specifies that rows should be dropped, and how='all' indicates that a row will only be dropped if all its values are NaN.

- Converting the dataset into a transaction format involves creating a list of lists, where each inner list represents a transaction and contains the items present in that transaction.

- Encoding the transaction data using the TransactionEncoder class from mlxtend.preprocessing. This will transform the transaction data into a one-hot encoded format.

After all these steps the dataset will be ready for market basket analysis. The output CSV file is provided with the submission.

C2: Code Execution

I am using the apriori algorithm (Jihargifari, 2020) to find the frequently bought items in the dataset.

```
In [19]: frequent_itemsets = apriori(df_encoded, min_support=0.05, use_colnames=True)
```

```
In [20]: frequent_itemsets
```

Out[20]:

	support	itemsets
0	0.050527	(10ft iPHone Charger Cable 2 Pack)
1	0.068391	(Anker USB C to HDMI Adapter)
2	0.087188	(Apple Lightning to Digital AV Adapter)
3	0.179709	(Apple Pencil)
4	0.132116	(Apple USB-C Charger cable)
5	0.062525	(Cat8 Ethernet Cable)
6	0.238368	(Dust-Off Compressed Gas 2 pack)
7	0.065858	(FEIYOLD Blue light Blocking Glasses)
8	0.059992	(Falcon Dust Off Compressed Gas)
9	0.163845	(HP 61 ink)
10	0.058526	(HP 62XL Tri-Color ink)
11	0.079323	(HP 63XL Ink)
12	0.071457	(Logitech M510 Wireless mouse)
13	0.095321	(Nylon Braided Lightning to USB cable)
14	0.051060	(Premium Nylon USB Cable)
15	0.052393	(SAMSUNG EVO 32GB card)
16	0.063325	(SanDisk Ultra 128GB card)
17	0.098254	(SanDisk Ultra 64GB card)
18	0.129583	(Screen Mom Screen Cleaner kit)
19	0.095054	(Stylus Pen for iPad)
20	0.081056	(Syntech USB C to USB Adapter)
21	0.076523	(TopMate C5 Laptop Cooler pad)
22	0.170911	(USB 2.0 Printer cable)
23	0.080389	(USB Type C to USB-A Charger cable)
24	0.174110	(VIVO Dual LCD Monitor Desk mount)
25	0.050927	(Apple Pencil, Dust-Off Compressed Gas 2 pack)
26	0.052660	(HP 61 ink, Dust-Off Compressed Gas 2 pack)
27	0.059725	(VIVO Dual LCD Monitor Desk mount, Dust-Off Co...

```
In [22]: # print number of items
         print("There are",frequent_itemsets.shape[0], "sets of items." )

There are 28 sets of items.
```

The first parameter df_encoded represents the preprocessed binary dataset. Each row represents a transaction, and each column represents an item.

The min_support parameter sets the minimum support threshold, specifying the minimum proportion of transactions an itemset should appear in to be considered frequent. I am using min_support=0.05, which means that an itemset should appear in at least 5% of the transactions to be considered frequent. It is a way to filter out infrequent or less significant itemsets and focus on those that occur frequently.

The use_colnames=True means that the resulting itemsets will have the item names (column names) instead of integer-coded item representation. It enhances the readability and understanding of the output.

Based on this there are 28 sets of items found.

C3: Association Rules Table

The association_rules function was used to generate the association rules from the frequent itemsets. The metric parameter specifies the metric to evaluate the rules (e.g., confidence, lift), and the min_threshold parameter sets the minimum threshold for the metric.

I am using the confidence metric with min_threshold=0.2 which means that only association rules with a confidence value of 0.2 (20%) or higher will be considered significant. In other words, the rule should have a confidence level of at least 20% to be included in the results.

Below are my generated association rules printed –

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Apple Pencil)	(Dust-Off Compressed Gas 2 pack)	0.179709	0.238368	0.050927	0.283383	1.188845	0.008090	1.062815	0.193648
1	(Dust-Off Compressed Gas 2 pack)	(Apple Pencil)	0.238368	0.179709	0.050927	0.213647	1.188845	0.008090	1.043158	0.208562
2	(HP 61 ink)	(Dust-Off Compressed Gas 2 pack)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357	0.308965
3	(Dust-Off Compressed Gas 2 pack)	(HP 61 ink)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256	0.339197
4	(VIVO Dual LCD Monitor Desk mount)	(Dust-Off Compressed Gas 2 pack)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314	0.369437
5	(Dust-Off Compressed Gas 2 pack)	(VIVO Dual LCD Monitor Desk mount)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008	0.400606

C4: Top Three Rules

The top three rules can be identified based on Lift, Confidence, or Support depending on the objective of the analysis. The support matrix indicates the proportion of transactions in the dataset that contains a specific itemset. Higher support values indicate that the itemsets are popular and frequently co-purchased by customers, so I am using the Support matrix to sort the top three rules to identify frequently co-purchased itemsets to answer my research question.

Here are the top 3 rules based on the support matrix.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
4	(VIVO Dual LCD Monitor Desk mount)	(Dust-Off Compressed Gas 2 pack)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314	0.369437
5	(Dust-Off Compressed Gas 2 pack)	(VIVO Dual LCD Monitor Desk mount)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008	0.400606
2	(HP 61 ink)	(Dust-Off Compressed Gas 2 pack)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357	0.308965

Another representation -

```
Rule: VIVO Dual LCD Monitor Desk mount -> Dust-Off Compressed Gas 2 pack
Confidence: 0.34
Support: 0.06
Lift: 1.44

Rule: Dust-Off Compressed Gas 2 pack -> VIVO Dual LCD Monitor Desk mount
Confidence: 0.25
Support: 0.06
Lift: 1.44

Rule: HP 61 ink -> Dust-Off Compressed Gas 2 pack
Confidence: 0.32
Support: 0.05
Lift: 1.35
```

Here is my interpretation of the top 3 rules –

Rule 1: VIVO Dual LCD Monitor Desk mount -> Dust-Off Compressed Gas 2 pack

Confidence: 0.34

Support: 0.06
Lift: 1.44

Customers who purchased the "VIVO Dual LCD Monitor Desk mount" have a 34% probability of also purchasing the "Dust-Off Compressed Gas 2 pack". The support value indicates that this association occurs in 6% of all transactions. The lift value of 1.44 indicates that the purchase of the "VIVO Dual LCD Monitor Desk mount" is 1.44 times more likely to be accompanied by the purchase of the "Dust-Off Compressed Gas 2 pack" compared to their individual occurrence.

Rule 2: Dust-Off Compressed Gas 2 pack -> VIVO Dual LCD Monitor Desk mount

Confidence: 0.25
Support: 0.06
Lift: 1.44

Customers who purchased the "Dust-Off Compressed Gas 2 pack" have a 25% probability of also purchasing the "VIVO Dual LCD Monitor Desk mount". The support value indicates that this association occurs in 6% of all transactions. The lift value of 1.44 indicates that the purchase of the "Dust-Off Compressed Gas 2 pack" is 1.44 times more likely to be accompanied by the purchase of the "VIVO Dual LCD Monitor Desk mount" compared to their individual occurrence.

Rule 3: HP 61 ink -> Dust-Off Compressed Gas 2 pack

Confidence: 0.32
Support: 0.05
Lift: 1.35

Customers who purchased the "HP 61 ink" have a 32% probability of also purchasing the "Dust-Off Compressed Gas 2 pack". The support value indicates that this association occurs in 5% of all transactions. The lift value of 1.35 indicates that the purchase of the "HP 61 ink" is 1.35 times more likely to be accompanied by the purchase of the "Dust-Off Compressed Gas 2 pack" compared to their individual occurrence.

Part IV - Data Summary and Implications

D1: Significance of Support, Lift, and Confidence

Here is the significance of all the matrix (Sivek, 2020) -

Support :

The support values in the top 3 results indicate the proportion of transactions in the dataset that contain the specific itemsets. The values range from 0.05 to 0.06, suggesting that the itemsets occur in a relatively small percentage of transactions.
These support values indicate that the co-occurrence of the items in the association rules is not very frequent in the dataset.

Confidence:

The confidence values in the top 3 results represent the conditional probability of the consequent item(s) being purchased given the antecedent item(s) are already purchased. The values range from 0.25 to 0.34, indicating moderate to low levels of confidence.
These confidence values suggest that the likelihood of co-purchasing the consequent item(s) given the antecedent item(s) is not very high.

Lift:

The lift values in the top 3 results measure the ratio of observed support to the expected support if the antecedent and consequent were independent.
The lift values are all above 1, ranging from 1.35 to 1.44 which indicates that the association between the antecedent and consequent items in the rules is slightly stronger than what would be expected by chance alone.

Overall, the top 3 results show relatively low support, moderate to low confidence, and slightly stronger than expected association (lift) between the antecedent and consequent items (Sivek, 2020).

D2: Practical Significance of the Findings

This analysis is helping to identify the frequently co-purchased items in the telecom company dataset. This information is valuable for understanding customer behavior and preferences, as well as identifying popular item combinations.

The findings indicate that the co-purchasing patterns between the identified items (such as the "VIVO Dual LCD Monitor Desk mount," "Dust-Off Compressed Gas 2 pack," and "HP 61 ink") are not very strong or frequent. This information can guide the company in determining the placement and assortment of these items in their physical or online stores. By understanding the weak to moderate associations between certain items, the company can design marketing messages or promotions that highlight the complementary or related nature of these items. This can potentially increase customer awareness and consideration, leading to higher sales and customer satisfaction.

Knowledge of the co-purchasing patterns will also aid in inventory management. The telecom company can optimize its stock levels and supply chain operations by ensuring that frequently co-purchased items are adequately stocked together. Also, the telecom company can identify customer segments based on their preferences for certain items and tailor the marketing strategies accordingly. This can improve customer targeting, engagement, and overall conversion rates.

D3: Course of Action

Based on the analysis and results my recommended course of action for the telecom company is to Promote Bundled Offers since there is a strong association between the "VIVO Dual LCD Monitor Desk mount" and the "Dust-Off Compressed Gas 2 pack" (both in terms of high confidence and lift), the company can consider promoting bundled offers that include both of these items. This can be done through targeted marketing campaigns, on-site recommendations, or package deals. By offering these items together, the company can increase the likelihood of customers purchasing them as a bundle, leading to increased sales and customer satisfaction.

Also, another recommendation would be to enhance cross-selling Efforts as the association between the "HP 61 ink" and the "Dust-Off Compressed Gas 2 pack" suggests a potential cross-selling opportunity. The company can explore strategies to promote the purchase of these items together, such as offering discounts or incentives when customers buy both items. This can be particularly effective for customers who purchase printers or other office equipment that require ink cartridges. By highlighting the convenience and cost savings of purchasing both items together, the company can encourage customers to make additional purchases and increase their overall basket size.

I would also suggest that further analysis or investigation may be required to uncover more meaningful associations between items in customer transactions as the lift values are relatively close to 1 which indicates that the association is not particularly strong.

E. Panopto Recording

Panopto recording is provided.

F. Third-Party Code References

Raschka, S. (2014). Installing mlxtend.
http://rasbt.github.io/mlxtend/installation/

WGU, Resources. (n.d.). Market Basket Analysis webinar recording.
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=dbe89ddb-e92f-4d40-a87a-af030178abf1

G. References

Hua, S. (n.d.). Market Basket Analysis.
https://sarahtianhua.wordpress.com/contact-hua/

Jihargifari. (2020, Aug 3). How To Perform Market Basket Analysis in Python.
https://medium.com/@jihargifari/how-to-perform-market-basket-analysis-in-python-bd00b745b106

Kadlaskar, A. (2021, October 2). Market Basket Analysis: A Comprehensive Guide for Businesses.
https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

Sivek, S. C. (2020, Nov 16). Market Basket Analysis 101: Key Concepts.
https://towardsdatascience.com/market-basket-analysis-101-key-concepts-1ddc6876cd00