D212 Data Mining II - Market Basket Analysis

David Harvell
Master of Science, Data Analytics
October 2021

A-1. Propose one question relevant to a real-world organizational situation that you will answer using market basket analysis.

Can we identify products that are commonly purchased together for marketing and product layout purposes?

A-2. Define one goal of the data analysis.

Discover at least 3 pairs of products that are commonly purchased together. If we can do this, the information can be used for marketing and store layouts - ultimately helping to increase sales.

B-1. Explain how market basket analyzes the selected dataset.

We will first encode the data in a manner that will make it easy for analysis by pivoting the products to the columns and using one hot encoding.

After that, we will use the Apriori algorithm to prune results and examine the top associations. Apriori will first look at the frequency of single items, use that to limit sets, and then can continue limited larger sets based on the frequency of earlier sets. This allows us to create a workable set of associations to investigate. (GeeksforGeeks, 2020)

Finally, we will compute metrics like confidence, lift, and Zhang's rule. This will allow us to report the "best" associations.

B-2. Provide one example of transactions in the dataset.

> We will begin by reviewing the dataset for anomalies.

In [1]:
```
import pandas as pd
df = pd.read_csv('teleco_market_basket.csv')
df.head(10)
```

Out[1]:

	Item01	Item02	Item03	Item04	Item05	Item06	Item07	Item08	Item09	Item10
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Logitech M510 Wireless mouse	HP 63 Ink	HP 65 ink	nonda USB C to USB Adapter	10ft iPHone Charger Cable	HP 902XL ink	Creative Pebble 2.0 Speakers	Cleaning Gel Universal Dust Cleaner	Micro Center 32GB Memory card	YUNSONG 3pack 6ft Nylon Lightning Cable
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Apple Lightning to Digital AV Adapter	TP-Link AC1750 Smart WiFi Router	Apple Pencil	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	UNEN Mfi Certified 5-pack Lightning Cable	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	Cat8 Ethernet Cable	HP 65 ink	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	Dust-Off Compressed Gas 2 pack	Screen Mom Screen Cleaner kit	Moread HDMI to VGA Adapter	HP 62XL Tri-Color ink	Apple USB-C Charger cable	NaN	NaN	NaN	NaN	NaN

Some blank entries are appearing in the dataset. We will check the underlying data.

teleco_market_basket.csv

```
Item01,Item02,Item03,Item04,Item05,Item06,Item07,Item08,Item09,Item10,Item11,Item12,Item13,Item14,Item15,Item16,Item17,Item18,Item19,Item20
,,,,,,,,,,,,,,,,,,,,
Logitech M510 Wireless mouse,HP 63 Ink,HP 65 ink,nonda USB C to USB Adapter,10ft iPHone Charger Cable,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 Headset,TONOR USB Gaming Microphone,Dust-Off Compressed Gas 2 pack,3A USB Type C Cable 3 pack 6FT,HOVAMP iPhone charger,SanDisk Ultra 128GB card,FEEL2NICE 5 pack 10ft Lighning cable,FEIYOLD Blue light Blocking Glasses
,,,,,,,,,,,,,,,,,,,,
Apple Lightning to Digital AV Adapter,TP-Link AC1750 Smart WiFi Router,Apple Pencil,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,
UNEN Mfi Certified 5-pack Lightning Cable,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,
```

In [2]:
```
df.Item01.isna().sum()
```

Out[2]: 7501

In [3]:
```
len(df)
```

Out[3]: 15002

It appears as though there are blank entries between each valid entry. We will clean by removing all records where there is no Item 01 in the transaction.

In [4]:
```
df = df[df.Item01.notna()]
df.head()
```

Out[4]:

	Item01	Item02	Item03	Item04	Item05	Item06	Item07	Item08	Item09	Item10
1	Logitech M510 Wireless mouse	HP 63 Ink	HP 65 ink	nonda USB C to USB Adapter	10ft iPHone Charger Cable	HP 902XL ink	Creative Pebble 2.0 Speakers	Cleaning Gel Universal Dust Cleaner	Micro Center 32GB Memory card	YUNSONG 3pack 6ft Nylon Lightning Cable
3	Apple Lightning to Digital AV Adapter	TP-Link AC1750 Smart WiFi Router	Apple Pencil	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	UNEN Mfi Certified 5-pack Lightning Cable	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	Cat8 Ethernet Cable	HP 65 ink	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	Dust-Off Compressed Gas 2 pack	Screen Mom Screen Cleaner kit	Moread HDMI to VGA Adapter	HP 62XL Tri-Color ink	Apple USB-C Charger cable	NaN	NaN	NaN	NaN	NaN

This shows the first 5 valid transactions in the dataset.

B-3. Summarize one assumption of market basket analysis.

A key assumption of market basket analysis is the existence of relationships/associations between products. We use varying metrics to measure these relationships and determine likely recurring sets of items that are purchased together. (Kamakura, 2012)

C-1. Transform the dataset to make it suitable for market basket analysis.

We have already removed the empty records. Next we will use one hot encoding to move products to the columns. This will prepare us for Apriori and Assocation Rules in the following steps.

In [5]:
```
# Transform the current dataframe into a list of lists that contain the ite
transactions = []

for index, row in df.iterrows():
    products = []
    for col in row:
        if not pd.isna(col):
            products.append(col)
    transactions.append(products)

# Encode the new lists using one hot encoding
from mlxtend.preprocessing import TransactionEncoder

encoder = TransactionEncoder().fit(transactions)
onehot = encoder.transform(transactions)
onehot = pd.DataFrame(onehot, columns = encoder.columns_)
onehot.head()
```

Out[5]:

	10ft iPHone Charger Cable	10ft iPHone Charger Cable 2 Pack	3 pack Nylon Braided Lightning Cable	3A USB Type C Cable 3 pack 6FT	5pack Nylon Braided USB C cables	ARRIS SURFboard SB8200 Cable Modem	Anker 2-in-1 USB Card Reader	Anker 4-port USB hub	Anker USB C to HDMI Adapter	Apple Lightning to Digital AV Adapter	...
0	True	False	False	True	False	False	False	False	False	False	...
1	False	False	False	False	False	False	False	False	False	True	...
2	False	False	False	False	False	False	False	False	False	False	...
3	False	False	False	False	False	False	False	False	False	False	...
4	False	False	False	False	False	False	False	False	False	False	...

5 rows × 119 columns

In [6]:
```
# Save the encoded data for upload with findings
onehot.to_csv('market_basket_encoded.csv', index = False)
```

C-2. Execute the code used to generate association rules with the Apriori algorithm.

Run Apriori for pairs of antecedents and consequents.

In [7]:
```
len(onehot)
```

Out[7]: 7501

In [8]:
```
from mlxtend.frequent_patterns import apriori, association_rules
frequent_itemsets = apriori(onehot, min_support = 0.005, max_len = 2, use_c

len(frequent_itemsets)
```

Out[8]: 552

In [9]:
```
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x:
frequent_itemsets.head()
```

Out[9]:

	support	itemsets	length
0	0.009065	(10ft iPHone Charger Cable)	1
1	0.050527	(10ft iPHone Charger Cable 2 Pack)	1
2	0.005199	(3 pack Nylon Braided Lightning Cable)	1
3	0.042528	(3A USB Type C Cable 3 pack 6FT)	1
4	0.019064	(5pack Nylon Braided USB C cables)	1

Limit to the best rules and sort by confidence.

In [10]:
```
rules = association_rules(frequent_itemsets, metric="confidence", min_thres
rules = rules.sort_values(by = 'confidence', ascending = False)
rules.describe()
```

Out[10]:

	antecedent support	consequent support	support	confidence	lift	leverage	conviction
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	0.032796	0.232527	0.014022	0.436910	1.895332	0.006287	1.366972
std	0.028483	0.019375	0.011836	0.031215	0.252447	0.005023	0.086197
min	0.010399	0.174110	0.005066	0.401254	1.683336	0.002364	1.272045
25%	0.014131	0.238368	0.005999	0.413951	1.736601	0.003062	1.299629
50%	0.018531	0.238368	0.007732	0.419028	1.757904	0.003614	1.310962
75%	0.046527	0.238368	0.020064	0.463275	1.988233	0.008973	1.447858
max	0.098254	0.238368	0.040928	0.487179	2.546642	0.017507	1.485182

C-3. Provide values for the support, lift, and confidence of the association rules table.

All relevant metrics for the remaining rules.

In [11]:
```
rules
```

Out[11]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage
8	(SanDisk Extreme 256GB card)	(Dust-Off Compressed Gas 2 pack)	0.010399	0.238368	0.005066	0.487179	2.043811	0.002587
6	(DisplayPort ot HDMI adapter)	(Dust-Off Compressed Gas 2 pack)	0.011998	0.238368	0.005733	0.477778	2.004369	0.002873
2	(Apple Lightning to USB cable)	(Dust-Off Compressed Gas 2 pack)	0.015598	0.238368	0.007332	0.470085	1.972098	0.003614
0	(10ft iPHone Charger Cable 2 Pack)	(Dust-Off Compressed Gas 2 pack)	0.050527	0.238368	0.023064	0.456464	1.914955	0.011020
4	(AutoFocus 1080p Webcam)	(VIVO Dual LCD Monitor Desk mount)	0.014131	0.174110	0.006266	0.443396	2.546642	0.003805
7	(FEIYOLD Blue light Blocking Glasses)	(Dust-Off Compressed Gas 2 pack)	0.065858	0.238368	0.027596	0.419028	1.757904	0.011898
5	(Brother Genuine High Yield Toner Cartridge)	(Dust-Off Compressed Gas 2 pack)	0.018531	0.238368	0.007732	0.417266	1.750511	0.003315
10	(SanDisk Ultra 64GB card)	(Dust-Off Compressed Gas 2 pack)	0.098254	0.238368	0.040928	0.416554	1.747522	0.017507
9	(SanDisk Extreme Pro 128GB card)	(Dust-Off Compressed Gas 2 pack)	0.018797	0.238368	0.007732	0.411348	1.725681	0.003252
3	(AutoFocus 1080p Webcam)	(Dust-Off Compressed Gas 2 pack)	0.014131	0.238368	0.005733	0.405660	1.701822	0.002364
1	(3A USB Type C Cable 3 pack 6FT)	(Dust-Off Compressed Gas 2 pack)	0.042528	0.238368	0.017064	0.401254	1.683336	0.006927

C-4. Identify the top three rules generated by the Apriori algorithm.

The top 3 rules.

In [12]:
```
rules.head(3)
```

Out[12]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	c
8	(SanDisk Extreme 256GB card)	(Dust-Off Compressed Gas 2 pack)	0.010399	0.238368	0.005066	0.487179	2.043811	0.002587	
6	(DisplayPort ot HDMI adapter)	(Dust-Off Compressed Gas 2 pack)	0.011998	0.238368	0.005733	0.477778	2.004369	0.002873	
2	(Apple Lightning to USB cable)	(Dust-Off Compressed Gas 2 pack)	0.015598	0.238368	0.007332	0.470085	1.972098	0.003614	

D-1. Summarize the significance of support, lift, and confidence from the results of the analysis.

Support
Support is the proportation of all equations that contain the assocation. This is the simplest metric. A value of one would represent the item (or combination) being in every transaction. The support isn't enough to indicate a strong relationship between items, because items with high popularity will seem to be related to the other items being purchased. (Sivek, 2020)

Confidence
Confidence is yet another proportion, but it limits the denominator to records that have the antecedent. This makes the metric more relevant when looking at relationships. It gives us the probability that the consequent will be purchased when purchasing the antecedent. The relationship will be weak when close to 0 and strongest at a value of 1. (Sivek, 2020)

Lift
Lift is another metric that assists in determining strength of relationships. It is the support of both items divided by the product of the supports for the individual items. This denominator mimics the two items being independently assigned to transactions. A lift greater than 1 indicates a probability that the items have a relationship that is not random; a lift equal to 1 represents no correlation; and finally, a lift less than 1 indicates the items are possible substitutions for one another. (Sivek, 2020)

Results of Analysis
Our top three assocations **do not have a strong support**. This is not surprising, based on the large and varied amount of items.

The assocations have **confidence between .47 and .48**. This indicates that the associations are purchased together around 50% of the time.

The **lift values are all around 2**, indicating that the assocations could be bundled together. While the bundles should work according to this analysis, all three assocations have the same consequent. Because of this, I would suggest a separate course of action (see below in D-2).

D-2. Discuss the practical significance of the findings from the analysis.

The results appear to suggest that compressed gas is one of the most common consequents. Because of the wide array of antecedents, I would believe that gas is picked up as a impulse item rather than directly tied to the antecedents.

D-3. Recommend a course of action.

I would recommend that displays of compressed gas are included near the registers, so we can try to capitalize on the impulse purchases when possible.

Code Resources

Raschka, S. (n.d.). Apriori - mlxtend. Mlxtend. Retrieved November 10, 2021, from http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/ (http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/)

Text Resources

GeeksforGeeks. (2020, April 4). Apriori Algorithm. Retrieved November 10, 2021, from https://www.geeksforgeeks.org/apriori-algorithm/ (https://www.geeksforgeeks.org/apriori-algorithm/)

Kamakura, W. A. (2012). Sequential market basket analysis. Marketing Letters, 23(3), 505–516. https://doi.org/10.1007/s11002-012-9181-6 (https://doi.org/10.1007/s11002-012-9181-6)

Sivek, S. C., PhD. (2020, November 17). Market Basket Analysis 101: Key Concepts - Towards Data Science. Medium. Retrieved November 10, 2021, from https://towardsdatascience.com/market-basket-analysis-101-key-concepts-1ddc6876cd00 (https://towardsdatascience.com/market-basket-analysis-101-key-concepts-1ddc6876cd00)