D212 Data Mining II Performance Task 3

Sean Simmons

WDU Data Analytics

MSDA D212

February 2023

“Scenario 1

One of the most critical factors in customer relationship management that directly affects a company’s long-term profitability is understanding its customers. When a company can better understand its customer characteristics, it is better able to target products and marketing campaigns for customers, resulting in better profits for the company in the long term.

You are an analyst for a telecommunications company that wants to better understand the characteristics of its customers. You have been asked to perform a market basket analysis to analyze customer data to identify key associations of your customer purchases, ultimately allowing better business and strategic decision-making.”

Part I: Research Question

A. Describe the purpose of this data mining report by doing the following:

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

2. Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

1. One question relevant to the telecommunications organization that I will use market basket analysis to answer is, "What are the top three association rules for the customer purchasing data sorted by the lift value?" This question will help inform the organization of key metrics important to retention and will primarily use lift, but also, we will evaluate the support and confidence values to help answer this question.

2. One goal of the analysis is to use market basket analysis with the customer purchasing data to identify the top three rules of association. This goal is reasonable within the scope of the scenario because by finding these associations, we can inform the organization of their key metrics that need to be altered to improve retention and better understand their customers. The goal is also represented in the available data through the variables provided about customer purchases and transaction history. We will use support, lift, and confidence to answer this question.

Part II: Market Basket Justification

B. Explain the reasons for using market basket analysis by doing the following:

1. Explain how market basket analyzes the selected dataset. Include expected outcomes.
2. Provide **one** example of transactions in the dataset.
3. Summarize **one** assumption of market basket analysis.

1. Market basket analysis (MBA) analyzes the selected dataset by analyzing what, how, and why customers purchase items and what items are purchased together. It is an unsupervised machine learning analysis and here we are using Descriptive MBA. The expected outcome is the top three associations based on their Lift values and in a table showing their support and confidence values as well. These top three associations will be taken from the entire set of data.

   MBA is an association rule machine learning method that allows us to identify relationships between variables in large datasets. The expected associations show how two products relate to each other and we are hoping to identify which products are included in the same transactions (Brown 2019). With this analysis, the organization will be able to recommend products to customers and be informed of customers potential purchases and purchasing behavior. Lift, Support, and Confidence are explained in section D1 of this report.

2. One example of transactions in the dataset is shown below for row 3:

```
#View dataset again
telo_df.head()
```

	Item01	Item02	Item03	I
1	Logitech M510 Wireless mouse	HP 63 Ink	HP 65 ink	A
3	Apple Lightning to Digital AV Adapter	TP-Link AC1750 Smart WiFi Router	Apple Pencil	

From this image, we can see the customer in row 3 purchased an Apple Lightning to Digital AV adapter, TP-Link AC1750 Smart WiFi Router, and an apple pencil. We can see from the first purchased item that the customer owns an apple product, and then they end up buying another apple product, the apple pen. If we were to look into interactions like this, a possible outcome is that apple products are often purchased together. More transactions can be seen in the same table this screenshot was taken from in the code file provided.

3. One assumption of market basket analysis is that items purchased together, a joint occurrence, indicates these products are complements and the purchase of one item leads to the purchase of the other item and is not completely random.

Part III: Data Preparation and Analysis

C. Prepare and perform market basket analysis by doing the following:

1. Transform the dataset to make it suitable for market basket analysis. Include a copy of the cleaned dataset.

2. Execute the code used to generate association rules with the Apriori algorithm. Provide screenshots that demonstrate the error-free functionality of the code.

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

4. Identify the top **three** rules generated by the Apriori algorithm. Include a screenshot of the top rules along with their summaries.

1. I have included the prepared dataframe that is cleaned and then also encoded for market basket analysis as "Ready_to_Run.csv."

2. I have executed the code and it is included in the file "D212_Task3_Code.ipynb". Screenshots of the error-free functionality related to the apriori algorithm are as follows:

```
#Convert dataset into list format for use with Apriori algorithm, list of lists.
telo_list = []
for i in range(0, 7501):
    telo_list.append([str(telo_df.values[i, j]) for j in range(0, 20)])

#Transform the list of list into an array so we can analysis it
TE = TransactionEncoder()
array = TE.fit(telo_list).transform(telo_list)
```

```
#Create the dataframe from our array, rows are logical and columns are the purchased items. Remove the first column that is just 0
df = pd.DataFrame(array, columns = TE.columns_)
df = df.drop(['0'],axis=1)
#Extract the fully prepared dataframe
df.to_csv('Ready_to_Run.csv')
print(df.info())
print(df.shape)
df
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7501 entries, 0 to 7500
Columns: 119 entries, 10ft iPHone Charger Cable to seenda Wireless mouse
dtypes: bool(119)
memory usage: 871.8 KB
None
(7501, 119)
```

	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	...	hP 65 Tri-color ink	iFixit Pro Tech Toolkit	iPhone 11 case	iPhone 12 Charger cable	iPhone 12 Pro case	iPhone 12 case	iPhone Charger Cable Anker 6ft	iPhone SE case	nonda USB C to USB Adapter
0	True	False	False	True	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	True
1	False	False	False	False	False	False	False	False	False	True	...	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False

```
#Running our Apriori Algorithm. You can change the min_support value based on the target threshold from the organization at a later time if needed
rules = apriori(df, min_support = 0.001, use_colnames = True)
rules.head()
```

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

```
#Viewing the rules
rules_results = list(rules)
rules_results
```

```
['support', 'itemsets']
```

```
#Now let's do our association rules
as_rules = association_rules(rules, metric = 'lift', min_threshold = 1)
as_rules.head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(10ft iPHone Charger Cable 2 Pack)	(10ft iPHone Charger Cable)	0.050527	0.009065	0.001067	0.021108	2.328418	0.000608	1.012302
1	(10ft iPHone Charger Cable)	(10ft iPHone Charger Cable 2 Pack)	0.009065	0.050527	0.001067	0.117647	2.328418	0.000608	1.076070
2	(10ft iPHone Charger Cable)	(3A USB Type C Cable 3 pack 6FT)	0.009065	0.042528	0.001466	0.161765	3.803753	0.001081	1.142248
3	(3A USB Type C Cable 3 pack 6FT)	(10ft iPHone Charger Cable)	0.042528	0.009065	0.001466	0.034483	3.803753	0.001081	1.026325
4	(10ft iPHone Charger Cable)	(Apple Pencil)	0.009065	0.179709	0.002133	0.235294	1.309304	0.000504	1.072688
5	(Apple Pencil)	(10ft iPHone Charger Cable)	0.179709	0.009065	0.002133	0.011869	1.309304	0.000504	1.002838
6	(10ft iPHone Charger Cable)	(Apple USB-C Charger cable)	0.009065	0.132116	0.001866	0.205882	1.558349	0.000669	1.092891
7	(Apple USB-C Charger cable)	(10ft iPHone Charger Cable)	0.132116	0.009065	0.001866	0.014127	1.558349	0.000669	1.005134
8	(10ft iPHone Charger Cable)	(Dust-Off Compressed Gas 2 pack)	0.009065	0.238368	0.003200	0.352941	1.480655	0.001039	1.177067
9	(Dust-Off Compressed Gas 2 pack)	(10ft iPHone Charger Cable)	0.238368	0.009065	0.003200	0.013423	1.480655	0.001039	1.004417

3. The values of the association rules table are listed below and in the screenshots above:

```
#Now let's do our association rules
as_rules = association_rules(rules, metric = 'lift', min_threshold = 1)
as_rules.head(10)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(10ft iPHone Charger Cable 2 Pack)	(10ft iPHone Charger Cable)	0.050527	0.009065	0.001067	0.021108	2.328418	0.000608	1.012302
1	(10ft iPHone Charger Cable)	(10ft iPHone Charger Cable 2 Pack)	0.009065	0.050527	0.001067	0.117647	2.328418	0.000608	1.076070
2	(10ft iPHone Charger Cable)	(3A USB Type C Cable 3 pack 6FT)	0.009065	0.042528	0.001466	0.161765	3.803753	0.001081	1.142248
3	(3A USB Type C Cable 3 pack 6FT)	(10ft iPHone Charger Cable)	0.042528	0.009065	0.001466	0.034483	3.803753	0.001081	1.026325
4	(10ft iPHone Charger Cable)	(Apple Pencil)	0.009065	0.179709	0.002133	0.235294	1.309304	0.000504	1.072688
5	(Apple Pencil)	(10ft iPHone Charger Cable)	0.179709	0.009065	0.002133	0.011869	1.309304	0.000504	1.002838
6	(10ft iPHone Charger Cable)	(Apple USB-C Charger cable)	0.009065	0.132116	0.001866	0.205882	1.558349	0.000669	1.092891
7	(Apple USB-C Charger cable)	(10ft iPHone Charger Cable)	0.132116	0.009065	0.001866	0.014127	1.558349	0.000669	1.005134
8	(10ft iPHone Charger Cable)	(Dust-Off Compressed Gas 2 pack)	0.009065	0.238368	0.003200	0.352941	1.480655	0.001039	1.177067
9	(Dust-Off Compressed Gas 2 pack)	(10ft iPHone Charger Cable)	0.238368	0.009065	0.003200	0.013423	1.480655	0.001039	1.004417

4. The top three rules generated by the Apriori algorithm are included below along with the screenshot of the top rules with their summaries: Lift is our main target metric here.

Sorted by Lift:

```
#Time to Sort our Rules, we can change the lift sort by value to change the metric we are sorting by. Here it is lift because we want to see the highest lift.
sorted_rules = as_rules.sort_values(by='lift', ascending=False).head(3)
sorted_rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
24730	(HP 63XL Ink, 5pack Nylon Braided USB C cables)	(iPhone 11 case, USB 2.0 Printer cable)	0.005733	0.003066	0.001067	0.186047	60.675430	0.001049	1.224804
24735	(iPhone 11 case, USB 2.0 Printer cable)	(HP 63XL Ink, 5pack Nylon Braided USB C cables)	0.003066	0.005733	0.001067	0.347826	60.675430	0.001049	1.524543
29432	(iPhone 11 case, Dust-Off Compressed Gas 2 pack)	(Logitech M510 Wireless mouse, Apple Pencil)	0.002133	0.014131	0.001333	0.625000	44.227594	0.001303	2.628983

For Lift, you want the largest lift value, which represents the confidence to expected confidence ratio. These are the three rules with the largest lift values, indicating these items are purchased together frequently. Lift is our main metric and what I will use in the analysis in this report. I have included the next two for added analysis. These top three rules show what items are purchased together the most and it greatly informs customer purchasing behavior. The items are seen in the screenshot abovet.

Sorted by Support:

```
#Time to Sort our Rules, we can change the lift sort by value to change the metric we are sorting by. Here it is lift because we want to see the highest lift.
sorted_rules = as_rules.sort_values(by='support', ascending=False).head(3)
sorted_rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1570	(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
1571	(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
1465	(Dust-Off Compressed Gas 2 pack)	(HP 61 ink)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256

For support, you want the highest value, which represents the ratio of how popular the transaction is. With .0597 we see a roughly 6% popularity, indicating out of the 7501 transactions, this is the top 1 (and next 2 included) transactions by support value.

Sorted by Confidence:

```
#Time to Sort our Rules, we can change the lift sort by value to change the metric we are sorting by. Here it is lift because we want to see the highest lift.
sorted_rules = as_rules.sort_values(by='confidence', ascending=False).head(3)
sorted_rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
33086	(Logitech M510 Wireless mouse, FEIYOLD Blue li...	(Dust-Off Compressed Gas 2 pack)	0.001200	0.238368	0.001200	1.0	4.195190	0.000914	inf
28202	(Apple Lightning to USB cable, FEIYOLD Blue li...	(Dust-Off Compressed Gas 2 pack)	0.001200	0.238368	0.001200	1.0	4.195190	0.000914	inf
24729	(5pack Nylon Braided USB C cables, iPhone 11 c...	(HP 63XL Ink)	0.001067	0.079323	0.001067	1.0	12.606723	0.000982	inf

For confidence, you are looking for 1 or as close to 1 as possible because it is a ratio of items purchased together. These all have a value of 1 and this metric does not give us as helpful as a comparison.

Part IV: Data Summary and Implications

D. Summarize your data analysis by doing the following:

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

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

3. Recommend a course of action for the real-world organizational situation from part A1 based on your results from part D1.

1. The significance of the values is broken down and explained as follows:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
24730	(HP 63XL Ink, 5pack Nylon Braided USB C cables)	(iPhone 11 case, USB 2.0 Printer cable)	0.005733	0.003066	0.001067	0.186047	60.675430	0.001049	1.224804

a. Support: The popularity of an item in all transactions calculated through the number of transactions with it and the total number. In this analysis for the first rule in the lift sorted associations: Support is .003066, showing the frequency these two items are bought together. 3% is a relatively high number and indicates further investigation by the organization on how they can increase the total number of customers that buy these items.

b. Lift:The confidence ratio calculated using the confidence to expected confidence. In this analysis for the first rule in the lift sorted associations: 60 is the highest lift and this indicates this is a transaction that includes the second item often and the

company should investigate how to locate these items together to increase this number or try to replicate the conditions of the items in this transaction to increase the lift of other transactions.

c. Confidence: A ratio where we ask out of the transactions with one item, how many of those contain our second item of interest. In this analysis for the first rule in the lift sorted associations: 19% is how often customers purchase these items and it is almost ⅕ of the time, a relatively high amount and a good indicator to inform this organization's needs.

2. The practical significance of the findings from my analysis are that the organization now has a model to evaluate key metrics. They can see the top associations for lift (and support or confidence, as needed) to evaluate what products are doing well, why, and how they can increase the lift for all transactions. The organization now has a tool to greatly inform their customer sales and they can use it to predict where they need to place items or create bundles to increase sales and profits. This answers their need of wanting to be more informed and have a tool to provide actionable insight into customer purchasing history and behavior. This tool can also be used to predict what items might be purchased together in the future by extrapolating from the purchasing history we have.

3. The course of action I recommend from the results in D1 are for the organization to take the items in the top three rules and move them closer together and create incentives (deals, combos, price reductions) for purchasing these items together to increase sales or do the opposite to increase the profits from successfully selling the items together,

depending on the need and specific goal of this organization. As a specific example, I would recommend moving the printer cables close to the iPhone cases and think about a bundle or deal when buying these items in groups.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
24730	(HP 63XL Ink, 5pack Nylon Braided USB C cables)	(iPhone 11 case, USB 2.0 Printer cable)	0.005733	0.003066	0.001067	0.186047	60.675430	0.001049	1.224804
24735	(iPhone 11 case, USB 2.0 Printer cable)	(HP 63XL Ink, 5pack Nylon Braided USB C cables)	0.003066	0.005733	0.001067	0.347826	60.675430	0.001049	1.524543
29432	(iPhone 11 case, Dust-Off Compressed Gas 2 pack)	(Logitech M510 Wireless mouse, Apple Pencil)	0.002133	0.014131	0.001333	0.625000	44.227594	0.001303	2.628983

[ ]:

Part V: Attachments

E. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

Panopto Video Link:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b1fe4f32-cb45-408e-96d1-afb50127464c

F. Record *all* web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

References

Brown, E. D., & D.Sc. (2019, December 26). *Market Basket Analysis with Python and Pandas*. Python Data. https://pythondata.com/market-basket-analysis-with-python-and-pandas/

G. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

References

Brown, E. D., & D.Sc. (2019, December 26). *Market Basket Analysis with Python and Pandas*. Python Data. https://pythondata.com/market-basket-analysis-with-python-and-pandas/

H. Demonstrate professional communication in the content and presentation of your submission.

This aspect of the rubric is evaluated through the entirety of this report and I hope professionalism has shown continuously.