Matthew Blessing
*Student ID**: 000507424***
OFM3 - Task 3, D212 - Data Mining II

OFM 3 - Performance Assessment Task 3: Association Rules and Lift Analysis

Part I: Research Question

A1.
For this analysis, we will explore if it is possible to take existing customer transaction data and accurately determine which items are frequently bought together.

A2.
The goal of our analysis will be to utilize market basket analysis to determine item combinations that are frequently together by customers and determine the items' rules of association, along with their support, confidence, and lift values.

Part II: Market Basket Justification

B1.
Market basket analysis works by examining transaction data to determine items that are frequently purchased together, referred to as itemsets. These itemsets are used to identify rules of association, which suggest that when certain items are purchased, customers may also purchase other items. For instance, transaction data may show that if a customer buys pencils and pens, they are likely to also buy paper. We can define this rule as follows: {pencil, pen} => {paper}, which can be read as "if pencil and pen, then paper", where 'pencil' and 'pen' are known as the antecedents and 'paper' is known as the consequent (Introduction to Market Basket Analysis - What Is Market Basket Analysis?, n.d.).

B2.
Within our dataset, transactions are listed by rows, with the items in each transaction being listed in columns. The first transaction in our dataset is a rather large transaction containing the following 20 items:

- 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 Lightning cable
- FEIYOLD Blue light Blocking Glasses

B3.
Market basket analysis works on the assumption that when items that are purchased together, they may have the potential to complement each other, meaning that the purchase of one item will often lead to the purchase of the complementary item as well (McColl, 2018).

Part III: Data Preparation and Analysis

C1.
We begin preparation for our data by first importing our standard Python libraries:

```
#imports common libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import statsmodels.api as sm
sns.set_style('darkgrid')

#sets the jupyter notebook window to take up 90% width of the browser window
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:90% !important; }</style>"))
```

We then read in our market basket dataset into a data frame:

```
#imports our market basket data as a pandas data frame
df = pd.read_csv('teleco_market_basket.csv')
```

We utilize Pandas .info() function to get an overview of our variables, observation counts, and data types:

```
df.info()
```

We then check for null values:

```
#sums the null values per column
df.isnull().sum()
```

We then check the top rows of our data:

```
#checks the top rows of our data
df.head()
```

Since we will be utilizing a machine learning algorithm, we need to replace null values with 0s:

```
#replaces any null values with a 0
df.fillna(0, inplace=True)
```

We then double check to ensure there are no remaining null values:

```
#checks for the sum of any null values
df.isnull().sum()
```

We then need to convert the column names to enable us to easily create lists for our market basket analysis. We do this by converting the string column names to numbers using a for loop across all 20 columns and 15,002 rows and export this list, df_list, to a new data frame, df2:

```
#creates a list and creates a for loop that iterates over all values in our data frame to change the column names from strings to numbers, and then exports to a new data frame
df_list = []
for i in range(0, 15002):
 df_list.append([str(df.values[i, x]) for x in range(0, 20)])
df2 = pd.DataFrame(df_list)
```

We then check the top rows of our new data frame:

```
#checks the top rows of our new data frame
df2.head()
```

We then export our cleaned and prepared dataset:

```
#exports the new dataframe to a .csv
df2.to_csv('D212_Task2_Dataset.csv')
```

Please see the attached 'D212_Task2_Dataset.csv'.

C2.

Now that our data is prepared, we can begin our market basket analysis.

We begin by importing the Apriori algorithm and train the algorithm on our created list, df_list, as well as set our thresholds for minimum support, minimum confidence, minimum lift, and minimum length:

```
In [16]: #imports the apriori algorithm
         from apyori import apriori

         #trains the apriori algorithm on our df_list and sets minimum support, confidence, lift, and length
         rule_list = apriori(df_list, min_support = .003, min_confidence = 0.3, min_lift = 3, min_length = 2)
```

We then define our rule list:

```
In [45]: #defines the rule list
         rule_list = list(rule_list)
```

We then check the number of rules the algorithm has defined:

```
In [18]: #prints the number of rules
         print(len(rule_list))

         242
```

Next, we create a data frame called 'results' from our generated 'rule_list', and displays the data frame to ensure its structure looks correct:

```
In [25]: #creates a dataframe from our rule list
         results = pd.DataFrame(rule_list)

In [26]: #displays the results dataframe
         results
```

Out[26]:

	items	support	ordered_statistics
0	(Dust-Off Compressed Gas 2 pack, 10ft iPHone C...	0.011532	[((10ft iPHone Charger Cable 2 Pack), (Dust-Of...
1	(Screen Mom Screen Cleaner kit, 10ft iPHone Ch...	0.007599	[((10ft iPHone Charger Cable 2 Pack), (Screen ...
2	(3A USB Type C Cable 3 pack 6FT, Dust-Off Comp...	0.008532	[((3A USB Type C Cable 3 pack 6FT), (Dust-Off ...
3	(3A USB Type C Cable 3 pack 6FT, VIVO Dual LCD...	0.006732	[((3A USB Type C Cable 3 pack 6FT), (VIVO Dual...
4	(Screen Mom Screen Cleaner kit, Anker 2-in-1 U...	0.004933	[((Anker 2-in-1 USB Card Reader), (Screen Mom ...
...	...	...	...
237	(VIVO Dual LCD Monitor Desk mount, SanDisk Ult...	0.004333	[((SanDisk Ultra 64GB card, Nylon Braided Ligh...
238	(VIVO Dual LCD Monitor Desk mount, Screen Mom ...	0.004133	[((Screen Mom Screen Cleaner kit, Nylon Braide...
239	(SanDisk Ultra 64GB card, SanDisk 128GB Ultra ...	0.003200	[((SanDisk Ultra 64GB card, SanDisk 128GB Ultr...
240	(Screen Mom Screen Cleaner kit, SanDisk Ultra ...	0.004866	[((Screen Mom Screen Cleaner kit, SanDisk Ultr...
241	(Stylus Pen for iPad, SanDisk Ultra 64GB card,...	0.003200	[((Stylus Pen for iPad, SanDisk Ultra 64GB car...

242 rows × 3 columns

We now have our association rules provided by the Apriori algorithm, along with support values for each rule. But we still need to add our lift and confidence values.

C3.

To obtain our lift and confidence values, we begin by defining our support values as the support values within the results data frame. We then create four empty lists, one of left hand values, one for right hand values, one of confidence value, and one for lift value.

We then use a for loop to iterate through these lists and we append the calculated values to the lists:

```
In [27]: #defines the support values fromn the support column in results
         support = results.support

         #creates four lists for lhs, rhs, confidence, and lift values
         lhs_val = []
         rhs_val = []
         conf_val = []
         lift_val = []
```

```
In [35]: #creates a for loop that iterates over the lists defined above
         for i in range(results.shape[0]):
          single_list = results['ordered_statistics'][i][0]
          lhs_val.append(list(single_list[0]))
          rhs_val.append(list(single_list[1]))
          conf_val.append(single_list[2])
          lift_val.append(single_list[3])
```

We then convert our lists containing our lhs, rhs, confidence, and lift values into data frames:

```
In [29]: #converts the lists into dataframes
         lhs = pd.DataFrame(lhs_val)
         rhs = pd.DataFrame(rhs_val)
         confidence = pd.DataFrame(conf_val, columns=['confidence'])
         lift = pd.DataFrame(lift_val, columns=['lift'])
```

We then concatenate the four data frames into a single data frame called ‘basket_results’:

```
In [30]: #concats list into dataframe
         basket_results = pd.concat([lhs, rhs, support, confidence, lift], axis=1)
```

We then fill any null values with an empty space for ease of viewing our rules:

```
In [31]: #replaces null values with a space
         basket_results.fillna(value='', inplace=True)
```

We then check our basket_results data frame:

In [32]:
```
#displays our results
basket_results
```

Out[32]:

	0	1	0	1	support	confidence	lift
0	10ft iPhone Charger Cable 2 Pack		Dust-Off Compressed Gas 2 pack		0.011532	0.456464	3.829910
1	10ft iPhone Charger Cable 2 Pack		Screen Mom Screen Cleaner kit		0.007599	0.300792	4.642464
2	3A USB Type C Cable 3 pack 6FT		Dust-Off Compressed Gas 2 pack		0.008532	0.401254	3.366673
3	3A USB Type C Cable 3 pack 6FT		VIVO Dual LCD Monitor Desk mount		0.006732	0.316614	3.636945
4	Anker 2-in-1 USB Card Reader		Screen Mom Screen Cleaner kit		0.004933	0.334842	5.167998
...	...	...	...	...	...	...	...
237	SanDisk Ultra 64GB card	Nylon Braided Lightning to USB cable	0	VIVO Dual LCD Monitor Desk mount	0.004333	0.511811	5.879165
238	Screen Mom Screen Cleaner kit	Nylon Braided Lightning to USB cable	0	VIVO Dual LCD Monitor Desk mount	0.004133	0.350282	4.023689
239	SanDisk Ultra 64GB card	SanDisk 128GB Ultra microSDXC card	0	VIVO Dual LCD Monitor Desk mount	0.003200	0.400000	4.594793
240	Screen Mom Screen Cleaner kit	SanDisk Ultra 64GB card	0	VIVO Dual LCD Monitor Desk mount	0.004866	0.442424	5.082120
241	Stylus Pen for iPad	SanDisk Ultra 64GB card	0	VIVO Dual LCD Monitor Desk mount	0.003200	0.440367	5.058488

242 rows × 7 columns

And lastly, we set our column names and define the left hand and right hand sides and create a new data frame called 'basket_results2':

In [33]:
```
#sets column names to LHS, RHS, Support, Confidence & Lift and creates a new data frame, basket_results2
basket_results.columns = ['LHS', 1, 'RHS', 1, 'Support', 'Confidence', 'Lift']
basket_results2 = basket_results[['LHS', 'RHS', 'Support', 'Confidence', 'Lift']]
basket_results2
```

Out[33]:

	LHS	RHS	Support	Confidence	Lift
0	10ft iPhone Charger Cable 2 Pack	Dust-Off Compressed Gas 2 pack	0.011532	0.456464	3.829910
1	10ft iPhone Charger Cable 2 Pack	Screen Mom Screen Cleaner kit	0.007599	0.300792	4.642464
2	3A USB Type C Cable 3 pack 6FT	Dust-Off Compressed Gas 2 pack	0.008532	0.401254	3.366673
3	3A USB Type C Cable 3 pack 6FT	VIVO Dual LCD Monitor Desk mount	0.006732	0.316614	3.636945
4	Anker 2-in-1 USB Card Reader	Screen Mom Screen Cleaner kit	0.004933	0.334842	5.167998
...	...	...	...	...	...
237	SanDisk Ultra 64GB card	0	0.004333	0.511811	5.879165
238	Screen Mom Screen Cleaner kit	0	0.004133	0.350282	4.023689
239	SanDisk Ultra 64GB card	0	0.003200	0.400000	4.594793
240	Screen Mom Screen Cleaner kit	0	0.004866	0.442424	5.082120
241	Stylus Pen for iPad	0	0.003200	0.440367	5.058488

242 rows × 5 columns

We now have our association rules, along with their respective support, confidence, and lift values.

C4.
Our top three association rules that have been generated by the Apriori algorithm on this dataset are as follows:

	LHS	RHS	Support	Confidence	Lift
0	10ft iPHone Charger Cable 2 Pack	Dust-Off Compressed Gas 2 pack	0.011532	0.456464	3.829910
1	10ft iPHone Charger Cable 2 Pack	Screen Mom Screen Cleaner kit	0.007599	0.300792	4.642464
2	3A USB Type C Cable 3 pack 6FT	Dust-Off Compressed Gas 2 pack	0.008532	0.401254	3.366673

- {10ft iPHone Charger Cable 2 Pack} => {Dust-Off Compressed Gas 2 pack}
    - Support: 0.011532
    - Confidence: 0.456464
    - Lift: 3.829910

- {10ft iPHone Charger Cable 2 Pack} => {Screen Mom Screen Cleaner kit}
    - Support: 0.007599
    - Confidence: 0.300792
    - Lift: 4.642464

- {3A USB Type C Cable 3 pack 6FT} => {Dust-Off Compressed Gas 2 pack}
    - Support: 0.008532
    - Confidence: 0.401254
    - Lift: 3.66673

Part IV: Data Summary and Implications

D1.
Before we continue, let us define what support, confidence, and lift actually signify:

- **Support** - The percentage of transactions that contain the items in the itemset, the higher the support value, the more frequently the itemset occurs in the dataset
- **Confidence -** The probability that the transaction containing items on the left hand side of the rule will also contain the items from the right hand side of the rule. The higher the confidence value, the higher the probability that the right hand side items will also be purchased (McColl, 2018).
- **Lift -** The probability that all of the items within a rule will be purchased together. A lift value of greater than 1 suggests that the purchase of the antecedent increases the likelihood that the consequence will also be purchased. A lift value of less than 1 suggests that the purchase of the antecedent reduces the likelihood that the consequence will be purchased, which could potentially indicate the items are substitutions for each other. A lift value equal to 1 means that purchasing the antecedent makes no difference on the purchase of the consequence (Market Basket Analysis 101: Anticipating Customer Behavior, 2021).

As we can see from our top association rules, we do not see any high confidence association rules. Our first rule only has a confidence value of 46%, and the support value suggests that only 1% of transactions contain the items defined by the association rule. This would mean that on the 1% of transactions where someone purchasing the antecedent, the 10ft iPHone Charger Cable 2 Pack, someone will also purchase the consequent, Dust-Off Compressed Gas 2 pack, only 46% of the time. That is less than a random chance, and should not necessarily be seen as significant. However, the lift values on our top association rules are far greater than 1, meaning that there is a significant increase to the likelihood that someone will purchase the consequent in all of our top three rules.

D2.

There does not appear to be any significant findings with our market basket analysis that we can specifically point to as being useful from a practical standpoint. With such low confidence and support values, it is hard to point to any association rule as insightful. While the lift values for the top association rules are significant, with such a small and limited product selection, it would appear that most of the items are purchased together by chance, rather than by being complementary products to one another.

D3.

A recommendation to the stakeholders within the organization may be to examine the products carried for sale within the company and reevaluate what products are carried. Looking at the list of items purchased by customers does not show a lot of complementary items. For instance, we see several customers have purchased printer ink and printer cables, but have not purchased any printers. If this is due to the organization not selling printers, that is a missed opportunity, as customers who purchase printers will likely also purchase ink and printer cables.

It may be required that the organization needs to streamline and eliminate several products and instead focus on carrying complementary products, rather than a varied assortment of unrelated products in order to increase additional sales.

Part V: Attachments

E1.

Please see attached Panopto video.

References

Introduction to Market Basket Analysis - What is market basket analysis? (n.d.). DataCamp. Retrieved October 26, 2021, from https://campus.datacamp.com/courses/market-basket-analysis-in-python/introduction-to-market-basket-analysis-1

Market Basket Analysis 101: Anticipating Customer Behavior. (2021, January 12). Smartbridge. Retrieved October 26, 2021, from https://smartbridge.com/market-basket-analysis-101/

McColl, L. (2018, May 31). Market Basket Analysis: Understanding Customer Behaviour. Select Statistical Consultants. Retrieved October 26, 2021, from https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/