Data Cleaning and Profiling

Task 3

Nathan Hefner

A. Describe the purpose of your report by doing the following:

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

The question I am proposing is, what items are typically bought together?

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

Our goal will be to increase sales by recommending items to shoppers that are frequently sold together. In turn, this will improve customer satisfaction and result in a higher likelihood of them returning as customers.

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

1. Explain how the market basket technique analyzes the provided dataset, including expected outcomes.

Market Basket Analysis is a data mining technique that specializes in identifying relationships between products frequently purchased together by customers. The algorithm works by going through a series of analysis steps. First, the dataset must be cleaned and prepared by ensuring it has transaction IDs, customer IDs, and a list of items purchased in each transaction. Next, the transactions are encoded into a binary form with 1s acting as the product being present is true and 0s acting as the product being present is false. We then need to calculate support, which is the ratio of transactions containing the itemset to the total number of transactions. The last step before applying the apriori algorithm is to set a minimum support threshold to filter out insignificant values. The apriori algorithm's first step is initialization, which identifies individual itemsets in the data that meet the minimum support threshold. It then generates candidate itemsets along with deleting any that do not meet the minimum support threshold. Finally, after identifying frequent itemsets, the algorithm creates association rules and calculates their confidence and lift values to determine their strength. Confidence is the chance of a product being bought when a different product is purchased, and lift is the probability percentage compared across all purchases. The expected outcome would be finding hidden relationships between groups of 2 or more items, the commonality being they are commonly purchased together.

2. Provide one example of a transaction in the dataset.

537468	LUNCH BAG RED RETROSPOT	10	12/7/2010 10:36	$1.65	$16.50	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	RETROSPOT HEART HOT WATER BOTTLE	6	12/7/2010 10:36	$4.95	$29.70	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	BIRD HOUSE HOT WATER BOTTLE	12	12/7/2010 10:36	$2.55	$30.60	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	SCOTTIE DOG HOT WATER BOTTLE	3	12/7/2010 10:36	$4.95	$14.85	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	GREY HEART HOT WATER BOTTLE	4	12/7/2010 10:36	$3.75	$15.00	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	HOT WATER BOTTLE TEA AND SYMPATHY	4	12/7/2010 10:36	$3.95	$15.80	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	FELTCRAFT DOLL MARIA	6	12/7/2010 10:36	$2.95	$17.70	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	LUNCH BAG SPACEBOY DESIGN	10	12/7/2010 10:36	$1.65	$16.50	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	LUNCH BAG SUKI DESIGN	10	12/7/2010 10:36	$1.65	$16.50	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	10 COLOUR SPACEBOY PEN	24	12/7/2010 10:36	$0.85	$20.40	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	MAGIC DRAWING SLATE CIRCUS PARADE	24	12/7/2010 10:36	$0.42	$10.08	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	CLOTHES PEGS RETROSPOT PACK 24	12	12/7/2010 10:36	$1.49	$17.88	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	DELUXE SEWING KIT	9	12/7/2010 10:36	$5.95	$53.55	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	POPPYS PLAYHOUSE BEDROOM	6	12/7/2010 10:36	$2.10	$12.60	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	POPPYS PLAYHOUSE BATHROOM	6	12/7/2010 10:36	$2.10	$12.60	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	POPPYS PLAYHOUSE KITCHEN	6	12/7/2010 10:36	$2.10	$12.60	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	FELTCRAFT PRINCESS CHARLOTTE DOLL	4	12/7/2010 10:36	$3.75	$15.00	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied
537468	FELTCRAFT PRINCESS LOLA DOLL	4	12/7/2010 10:36	$3.75	$15.00	United States	No	Medium	Southeast	Consumer	No	Credit Card	Very Satisfied

This example from the dataset shows how the dataset is made up of Order ID, which is used for every item purchased, along with product name, quantity, invoice date, unit price, total price, country, whether a discount was used, order priority, region, the market segment to which the customer belongs, whether it has expedited shipping or not, payment method, and finally customer satisfaction, this one being very satisfied.

3. Summarize one assumption of market basket analysis.

One assumption of market basket analysis is the independence of transactions, which means each transaction is treated as an independent event. This means the analysis does not consider external factors like seasonal purchases, trends, or customer demographics. Therefore, the accuracy of predictions may be limited and incomplete. However, we can also assume that by discovering which products were purchased together, we can predict how customers will behave in the future.

C. Prepare the dataset for further analysis by doing the following:

1. Wrangle (i.e., transform) data by doing the following:

a. Select x number of categorical variables, choosing at least two ordinal variables and at least two nominal variables.

My two ordinal variables will be Order Priority (high, medium, and low, as the ranking) and customer order satisfaction (Very Satisfied, Satisfied, Dissatisfied, Very Dissatisfied, Prefer to not respond).

My two nominal variables will be Payment Method (PayPal or Credit Card) and Region (Northeast or Southeast).

b. Perform the appropriate encoding method (ordinal, label encoding, one-hot encoding) for each variable selected in part C1a.

We will use ordinal encoding for the two ordinal variables; therefore, we will assign numbers to the ranked values. We use ordinal encoding to assign numbers to the ranked values, and this will transform the data into a numeric format to be used for future analysis, as the data cannot be in categorical form for the Apriori Algorithm.

For Customer Order Satisfaction, we will use the following ranking (Very Satisfied=4, Satisfied = 3, Dissatisfied = 2, Very Dissatisfied = 1, Prefer to not respond = 0)

Below is my code for encoding customer satisfaction.

```
satisfaction_order = [['Prefer not to answer', 'Very Dissatisfied', 'Dissatisfied', 'Satisfied', 'Very Satisfied']]

encoder = OrdinalEncoder(categories=satisfaction_order)
df['CustomerOrderSatisfaction_Encoded'] = encoder.fit_transform(df[['CustomerOrderSatisfaction']])
```

For Order Priority, we will also be using ordinal encoding. This will be the following rankings (High = 3, Medium = 2, Low = 1).

```
priority_order = [['Low', 'Medium', 'High']]

encoder = OrdinalEncoder(categories=priority_order)
df['OrderPriority_Encoded'] = encoder.fit_transform(df[['OrderPriority']])
```

For the two nominal variables, we will use the One-Hot Encoding method to assign binary numbers to each value. We do this because there are no logical rankings for nominal variables, but we still need to assign a number to the value to prepare them for the Apriori Algorithm.

```
df_encoded = pd.get_dummies(df, columns=['PaymentMethod'])

one_hot_cols = [col for col in df_encoded.columns if 'PaymentMethod' in col]
df_encoded[one_hot_cols] = df_encoded[one_hot_cols].astype(int)
```

Here is the code for encoding PaymentMethod with One-Hot Encoding method.

```
df_encoded = pd.get_dummies(df, columns=['Region'])

one_hot_cols = [col for col in df_encoded.columns if 'Region' in col]
df_encoded[one_hot_cols] = df_encoded[one_hot_cols].astype(int)
```

This is the code for encoding Regions.

c. Transactionalize the data for market basket analysis.

The transactional data must be encoded into the correct format for analysis. Each transaction is represented in binary numbers, with 1 being True and 0 being False.

```
transactional_data = df.groupby("OrderID")["ProductName"].apply(list).tolist()
```

Here is my code to encode the dataset into a transactional dataset.

3. Execute the code used to generate association rules with the Apriori algorithm. Provide a screenshot that demonstrates that the code is error-free.

Below is the code I used to run the Apriori Algorithm on the dataset.

```
te = TransactionEncoder()
te_ary = te.fit(transactional_data).transform(transactional_data)
df_apriori = pd.DataFrame(te_ary, columns=te.columns_)

frequent_products = apriori(df_apriori, min_support=0.01, use_colnames=True)

rules = association_rules(frequent_products, metric="lift", min_threshold=1)

print(rules.head())
```

```
                         antecedents                         consequents  \
0  (CHARLOTTE BAG DOLLY GIRL DESIGN)               ( DOLLY GIRL BEAKER)
1               ( DOLLY GIRL BEAKER)  (CHARLOTTE BAG DOLLY GIRL DESIGN)
2        (DOLLY GIRL CHILDRENS BOWL)               ( DOLLY GIRL BEAKER)
3               ( DOLLY GIRL BEAKER)        (DOLLY GIRL CHILDRENS BOWL)
4         (DOLLY GIRL CHILDRENS CUP)               ( DOLLY GIRL BEAKER)

   antecedent support  consequent support   support  confidence       lift  \
0            0.058957            0.020408  0.011338    0.192308   9.423077
1            0.020408            0.058957  0.011338    0.555556   9.423077
2            0.040816            0.020408  0.015873    0.388889  19.055556
3            0.020408            0.040816  0.015873    0.777778  19.055556
4            0.036281            0.020408  0.013605    0.375000  18.375000

   representativity  leverage  conviction  zhangs_metric   jaccard  certainty  \
0               1.0  0.010135    1.212828       0.949880  0.166667   0.175481
1               1.0  0.010135    2.117347       0.912500  0.166667   0.527711
2               1.0  0.015040    1.602968       0.987842  0.350000   0.376157
3               1.0  0.015040    4.316327       0.967262  0.350000   0.768322
4               1.0  0.012865    1.567347       0.981176  0.315789   0.361979

   kulczynski
0    0.373932
1    0.373932
2    0.583333
3    0.583333
4    0.520833
```

4. Provide values for the support, lift, and confidence of the association rules table. Include a screenshot of the values.

Below is my code for generating support, lift, and confidence values.

```
rules_sorted_by_lift = rules.sort_values(by='lift', ascending=False)
rules_sorted_by_lift.head(5)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty
28680	(PLASTERS IN TIN SPACEBOY, ALARM CLOCK BAKELIK...	(ALARM CLOCK BAKELIKE PINK, PLASTERS IN TIN CI...	0.011338	0.011338	0.011338	1.0	88.2	1.0	0.011209	inf	1.0	1.0	1.0
33189	(ALARM CLOCK BAKELIKE RED , CHILDRENS CUTLERY ...	(ALARM CLOCK BAKELIKE PINK, CHILDRENS CUTLERY ...	0.011338	0.011338	0.011338	1.0	88.2	1.0	0.011209	inf	1.0	1.0	1.0
38340	(SET6 RED SPOTTY PAPER CUPS, ALARM CLOCK BAKEL...	(SET6 RED SPOTTY PAPER PLATES, PLASTERS IN TIN...	0.011338	0.011338	0.011338	1.0	88.2	1.0	0.011209	inf	1.0	1.0	1.0
38503	(ALARM CLOCK BAKELIKE RED , ROUND SNACK BOXES ...	(ALARM CLOCK BAKELIKE PINK, CHILDRENS CUTLERY ...	0.011338	0.011338	0.011338	1.0	88.2	1.0	0.011209	inf	1.0	1.0	1.0
33828	(ALARM CLOCK BAKELIKE PINK, SET6 RED SPOTTY PA...	(PLASTERS IN TIN SPACEBOY, SET6 RED SPOTTY PAP...	0.011338	0.011338	0.011338	1.0	88.2	1.0	0.011209	inf	1.0	1.0	1.0

5. Explain the top three relevant rules generated by the Apriori algorithm. Include a screenshot of the top three relevant rules.

The first rule predicts that when the antecedents, plasters in tin spaceboy, alarm clock bakelike ivory, are purchased, the consequents, alarm clock bakelike pink, plasters in tin circus parade, alarm clock bakelike green, charlotte bag dolly girl design, will also be purchased. The support value is 0.011338, meaning that 1.11338% of all transactions include both products. The lift is 88.2, meaning buying antecedents increases the chances of purchasing the consequents by 88 times. A confidence of 1.0 means that when the antecedents are purchased, the consequents are also purchased 100% of the time.

The second rule predicts that when the antecedents, alarm clock bakelike red , childrens cutlery spaceboy, card dolly girl, are purchased, the consequents, alarm clock bakelike pink, childrens cutlery dolly girl, spaceboy birthday card, will also be purchased. The support value is 0.011338, meaning that 1.11338% of all transactions include both products. The lift is 88.2, meaning buying antecedents increases the chances of purchasing the consequents by 88 times. A confidence of 1.0 means that when the antecedents are purchased, the consequents are also purchased 100% of the time.

The third rule predicts that when the antecedents, set6 red spotty paper cups, alarm clock bakelike red, plasters in tin circus parade, are purchased, the consequents, set6 red spotty paper plates, plasters in tin spaceboy, alarm clock bakelike green, round snack boxes set of 4

woodland, will also be purchased. The support value is 0.011338, meaning that 1.11338% of all transactions include both products. The lift is 88.2, meaning buying antecedents increases the chances of purchasing the consequents by 88 times. A confidence of 1.0 means that when the antecedents are purchased, the consequents are also purchased 100% of the time.

D. Summarize your data analysis by doing the following:

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

The support value shows how many transactions there are, including both the antecedents and consequents. The lift shows how often the chances of the consequent being purchased increase after the antecedent is purchased. The confidence is a percentage of when the antecedent is bought and how often the consequent is purchased. These three values are significant when calculating and discovering frequently purchased products together.

2. Explain the practical significance of your findings from the analysis.

The practical significance of my findings appears to be a few notable things. First, parents buy toys for both children, boy and girl, on one trip. They also buy multiple toys that go together for their children, such as matching playsets or multiple toys. Our data shows that party supplies are also bought together. We can adjust our marketing strategy to increase sales of all our products by finding these trends.

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

The course of action I recommend is to ensure all these products that have a high chance of being purchased are within proximity of each other and to make sure that the customers know these products exist within the stores. For online shopping, once the customer selects one of the products, I would recommend the other before and during checkout. This will help boost sales as the customers will be recommended products they didn't realize they would also want with the product they first intended to purchase.

Sources

No sources were used except WGU Material