Bradley_Holt_D212_Task3

January 8, 2022

1 Part I: Research Question

1.1 A1. Question

What item or items should be used in an upcoming promotional sale for Telco Company?

1.2 A2. Data Analysis Goal

One goal of the data analysis is to identify items that are frequently purchased together in each transaction of the market_basket dataset by creating association rules. Association rules are x_item then y_item ({antecedent} -> {consequent}) relationships created between items and/or itemsets. However, because itemsets increase exponentially with each unique item within a data set, after a certain amount of items it would be impossible to enumerate them all and create all the itemsets. Because of this, market basket analysis must use pruning to reduce itemsets before effective association rules can be created.

2 Part II: Market Basket Justification

2.1 B1: Market Basket Explaination

Market basket analysis is a popular method to find associations and correlations between items in transactional or relational detasets. This type of analysis can be used to: *** 1. Build Netflix-style recommendations engine. 2. Improve product recommendation on an e-commerce store. 3. Cross-sell products in a retain setting. 4. Improve inventory management. 5. Upsell products.

(Hull) ***

Market basket analysis is based on the use of association rules to group items into related objects. Association rules explain what items are associated with each other using a methods which make sense for the specific data set. For example, determining what items are frequently purchased together is optimal for creating association rules in a transactional data set or determining what shows or movies are frequently watched consecutively in a steaming data set. Such rules take the form of an if-then relationship between two sets of items. The first is called the antecedent and the second is called the consequent.

A problem with market basket analysis is in large data sets the number of association rules can be impossible to create. Therefore, item sets must be filtered before association rules can be created. One way to accomplish this is to use the Apriori Algorithm to reduce the number of itemsets in the data set. In short, the Aprirori Algorithm is a frequent itemset generator. It scans the list of items and determines what items occur frequently together and continues combining items into itemset based on if they are frequently occuring or not. If an item or item set is not frequently occuring it removes them from the list of itemsets to be used to create association rules.

One association rules are created, different metrics are computed through market basket analysis methods in order to be used to find and filter for the best association rules. Some of these metrics include support, lift, and confidence. Based on the requirements of the organization's needs these metrics are used to finalize association rules that can then be used to address the organization's research question.

Expected outcomes:

1. List of transactions are created.
2. Each unique item is encoded into true/false boolean values for each transaction.
3. Apriori Algorithm used to create list of frequent items purchased in each transaction.
4. Association rules are created from the frequently occuring item list to create {antecedent} -> {consequent} rules that include the support, lift, confidence and other metrics for each rule.
5. Metrics are then used to filter for the best top three association rules to address the research question.

2.2 B2: Example of One Transaction

The market basket analysis is converted to a list of transaction as created below. Row index three of the transaction list is selected to provide the example transaction:

['Apple Lightning to Digital AV Adapter', 'TP-Link AC1750 Smart WiFi Router', 'Apple Pencil', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan']

In this specific transaction 3 items were purchased: Apple Lightning to Digital AV Adapter, TP-Link AC1750 Smart WiFi Router, and Apple Pencil. The remaining 17 items are 'nan' because the original DataFrame supported up to 20 items per transaction. Other transactions can include more or less items.

```
[ ]: # Libraries used in task
     import pandas as pd
     import numpy as np
     from mlxtend.preprocessing import TransactionEncoder
     from mlxtend.frequent_patterns import apriori
     from mlxtend.frequent_patterns import association_rules
```

```
[ ]: # Load transactions from pandas.
     data = pd.read_csv("C:/Users/holtb/Data/D212_Data_Mining_II/data/
      ↪teleco_market_basket.csv")
```

```
[ ]: #Build list of transactions
     transactions = []

     for i in range(0, len(data)):
         transactions.append([str(data.values[i,j]) for j in range(0, len(data.
      ↪columns))])
```

```
print(transactions[3])
```

```
['Apple Lightning to Digital AV Adapter', 'TP-Link AC1750 Smart WiFi Router',
'Apple Pencil', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan',
'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan', 'nan']
```

2.3 B3: Summarization of One Assumption

The underlying assumption of market basket analysis is that the occurence of two or more items in the basket implies that the items are complements in the purchase. In the case of the market basket data set, it's implied that the customer purchased the consequent item or items because of the purchase of the antecedent item or items. This may not always be the case and common sense must be used when evaluating association rules.

3 Part III: Data Preparation and Analysis

3.1 C1: Dataset Transformation and Copy

[ ]:
```
# Instantiate transaction encoder
encoder = TransactionEncoder().fit(transactions)

# Encode list of transactions to list of True/False booleans
onehot = encoder.transform(transactions)

# Transform true/false list to pandas DataFrame and relabel columns to items␣
↪while dropping nan values
onehot = pd.DataFrame(onehot, columns= encoder.columns_).drop('nan', axis =1)
onehot.head()
```

[ ]:
```
   10ft iPHone Charger Cable  10ft iPHone Charger Cable 2 Pack  \
0                      False                             False
1                       True                             False
2                      False                             False
3                      False                             False
4                      False                             False

   3 pack Nylon Braided Lightning Cable  3A USB Type C Cable 3 pack 6FT  \
0                                 False                           False
1                                 False                            True
2                                 False                           False
3                                 False                           False
4                                 False                           False

   5pack Nylon Braided USB C cables  ARRIS SURFboard SB8200 Cable Modem  \
0                             False                               False
1                             False                               False
2                             False                               False
```

```
3                             False                                    False
4                             False                                    False

   Anker 2-in-1 USB Card Reader  Anker 4-port USB hub  \
0                         False                 False
1                         False                 False
2                         False                 False
3                         False                 False
4                         False                 False

   Anker USB C to HDMI Adapter  Apple Lightning to Digital AV Adapter  …  \
0                        False                                  False  …
1                        False                                  False  …
2                        False                                  False  …
3                        False                                   True  …
4                        False                                  False  …

   hP 65 Tri-color ink  iFixit Pro Tech Toolkit  iPhone 11 case  \
0                False                    False           False
1                False                    False           False
2                False                    False           False
3                False                    False           False
4                False                    False           False

   iPhone 12 Charger cable  iPhone 12 Pro case  iPhone 12 case  \
0                    False               False           False
1                    False               False           False
2                    False               False           False
3                    False               False           False
4                    False               False           False

   iPhone Charger Cable Anker 6ft  iPhone SE case  nonda USB C to USB Adapter  \
0                           False           False                       False
1                           False           False                        True
2                           False           False                       False
3                           False           False                       False
4                           False           False                       False

   seenda Wireless mouse
0                  False
1                  False
2                  False
3                  False
4                  False

[5 rows x 119 columns]
```

```
[ ]: # Exporting onehot encoded DateFrame to CSV
onehot.to_excel('C:/Users/holtb/Data/D212_Data_Mining_II/data/
 ↪market_basket_onehot.xlsx')
```

3.2 C2: Generating Association Rules with Apriori Algorithm

```
[ ]: # Compute frequent itemsets using the Apriori algorithm
frequent_itemsets = apriori(onehot,
                            min_support = 0.025,
                            max_len = 2,
                            use_colnames = True)
frequent_itemsets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28 entries, 0 to 27
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   support   28 non-null     float64
 1   itemsets  28 non-null     object
dtypes: float64(1), object(1)
memory usage: 576.0+ bytes
```

```
[ ]: # Print a preview of the frequent itemsets
frequent_itemsets.sort_values('support', ascending=False).head(10)
```

```
[ ]:      support                                   itemsets
 6   0.119184         (Dust-Off Compressed Gas 2 pack)
 3   0.089855                           (Apple Pencil)
 24  0.087055       (VIVO Dual LCD Monitor Desk mount)
 22  0.085455                  (USB 2.0 Printer cable)
 9   0.081922                               (HP 61 ink)
 4   0.066058                (Apple USB-C Charger cable)
 18  0.064791          (Screen Mom Screen Cleaner kit)
 17  0.049127                 (SanDisk Ultra 64GB card)
 13  0.047660  (Nylon Braided Lightning to USB cable)
 19  0.047527                      (Stylus Pen for iPad)
```

3.3 C3: Association Rules Table

```
[ ]: # Compute all frequent association rules
rules = association_rules(frequent_itemsets,
                          metric = "support",
                          min_threshold = 0.0)
rules.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
```

```
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   antecedents         6 non-null      object
 1   consequents         6 non-null      object
 2   antecedent support  6 non-null      float64
 3   consequent support  6 non-null      float64
 4   support             6 non-null      float64
 5   confidence          6 non-null      float64
 6   lift                6 non-null      float64
 7   leverage            6 non-null      float64
 8   conviction          6 non-null      float64
dtypes: float64(7), object(2)
memory usage: 560.0+ bytes
```

[ ]:
```
# Print frequent association rules
rules.head(6)
```

[ ]:
```
                         antecedents                           consequents  \
0                     (Apple Pencil)    (Dust-Off Compressed Gas 2 pack)
1   (Dust-Off Compressed Gas 2 pack)                        (Apple Pencil)
2   (Dust-Off Compressed Gas 2 pack)                          (HP 61 ink)
3                        (HP 61 ink)    (Dust-Off Compressed Gas 2 pack)
4   (Dust-Off Compressed Gas 2 pack)  (VIVO Dual LCD Monitor Desk mount)
5 (VIVO Dual LCD Monitor Desk mount)    (Dust-Off Compressed Gas 2 pack)

   antecedent support  consequent support   support  confidence      lift  \
0            0.089855            0.119184  0.025463    0.283383  2.377689
1            0.119184            0.089855  0.025463    0.213647  2.377689
2            0.119184            0.081922  0.026330    0.220917  2.696664
3            0.081922            0.119184  0.026330    0.321400  2.696664
4            0.119184            0.087055  0.029863    0.250559  2.878170
5            0.087055            0.119184  0.029863    0.343032  2.878170

   leverage  conviction
0  0.014754    1.229130
1  0.014754    1.157425
2  0.016566    1.178408
3  0.016566    1.297989
4  0.019487    1.218168
5  0.019487    1.340729
```

3.4 C4: Top Three Rules

1. {Apple Pencil} -> {Dust-Off Compressed Gas 2 pack}
2. {HP 61 ink} -> {Dust-Off Compressed Gas 2 pack}
3. {VIVO Dual LCD Monitor Desk mount} -> {Dust-Off Compressed Gas 2 pack}

```
[ ]: # Compute top 3 association rules filtering by confidence
rules = association_rules(frequent_itemsets,
                            metric = "confidence",
                            min_threshold = 0.27)
rules.head()
```

```
[ ]:                    antecedents                        consequents  \
0                   (Apple Pencil)  (Dust-Off Compressed Gas 2 pack)
1                     (HP 61 ink)  (Dust-Off Compressed Gas 2 pack)
2  (VIVO Dual LCD Monitor Desk mount)  (Dust-Off Compressed Gas 2 pack)

   antecedent support  consequent support   support  confidence      lift  \
0            0.089855            0.119184  0.025463    0.283383  2.377689
1            0.081922            0.119184  0.026330    0.321400  2.696664
2            0.087055            0.119184  0.029863    0.343032  2.878170

   leverage  conviction
0  0.014754    1.229130
1  0.016566    1.297989
2  0.019487    1.340729
```

4 Part IV: Data Summary and Implications

4.1 D1: Summarization of support, lift, and confidence

- Support: The support metric is the popularity of an item/itemset in the transcation. Mathematically, the support of item A is the ratio of transactions involving A to the total number of transactions. The results of the analysis indicates the most popular association rule is {VIVO Dual LCD Monitor Desk mount} -> {Dust-Off Compressed Gas 2 pack} occuring in approximately 3% of the transactions. Specific item supports can be located in the antecendent support and consequent support columns.
- Lift: Lift can be defined as the increase in the sale of A when you sell B. Additionally, lift greater than 1 provides evidence that that specific association rule did not occur in the rule list by chance. The results of the analysis indicates that the sale of the antecendent increases the sale of consequents by over 2 times and the rules were not created by chance.
- Confidence: Confidence is the likelihood that a customer bought both the antecendent item(s) (A) and consequent item(s) (B). Mathmatically, it is created by dividing the number of transactions involving both A and B by the number of transactions involving B. The higher the confidence, the stronger the association rule is. The results of the analysis indicate that it is 30% likely that if a transaction includes the antecedent, then the consequent will also be included.

4.2 D2: Practical Significant of Findings

The identification of the most popular combinations of specific items is the underlying significance of these findings. The results of the top three indicates the Dust-off Compressed Gas 2 pack is purchased in combination with the purchase of the Apple Pencil, HP 61 ink, and VIVO Dual LCD Monitor Desk mount about 3% of the time.

4.3 D3: Recommended Course of Action

The top three results of the association rules indicate that the Dust-Off Compressed Gas 2 pack is the most likely consequent of Apple Pencil, HP 61 ink, and VIVO Dual LCD Monitor, however the metrics may not be significantly strong enough to improve the purchase of the compressed gas. Additionally, Dust-Off Compress Gas 2 pack has the highest support of single items of all items (11%). It may not make sense to use the compressed gas as a consequent promotional item as it already is the most popular item. But because it's popularity, the item could potentially be used to promote purchases of less popular items. The association rules should be explored again using the Dust-Off Compressed Gas as the antecedent to find potential items to promote along with it.

5 Part V: Attachments

5.1 E. Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1b806205-516f-403b-9974-ae17010e018d

5.2 F. Third Party Code

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2. (Publisher link).

Hull, I. “Market Basket Analysis in Python” [MOOC]. Datacamp. https://app.datacamp.com/learn/courses/market-basket-analysis-in-python

Python Software Foundation. Python Language Reference, version 3.7. Available at http://www.python.org

Sebastian Raschka, “MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack”; The Journal of Open Source Software. Volume 3, (2018). DOI: 10.21105/joss.00638. (The Open Journal)

5.3 G. In-text citations

Hull, I. “Market Basket Analysis in Python” [MOOC]. Datacamp. https://app.datacamp.com/learn/courses/market-basket-analysis-in-python