WGU D212 PA 3

Andrew Shrestha

Part I: Research Question

A1. Question: By obtaining and analyzing a list of customer transactions consisting of real purchases, is it possible to identify items that are frequently bought together in combination with the purpose of creating effective promotional marketing? This question will be answered using a Market Basket Analysis Technique.

A2. Objectives and Goals: The goal of this research analysis is to add a greater insight into stakeholder decision making with the aim in increasing customer loyalty to the company. With our results identifying items that are frequently bought together in combination, stakeholders will be able to more effectively utilize discounts and promotional marketing decisions to reduce customer incentives to churn.

Part II: Method Justification

B1. Explanation of Market Basket Analysis

Market basket analysis is a form of frequent itemset mining in which various associations and correlations between items within either transactional or relational databases are identified. In the case of our scenario, we will be identifying the buying habits of our customers and finding the associations between the telecom products/services they choose to purchase **(Amruta Kadlaskar, 2021).**

It is expected that there will be several different products that customers will purchase in combination that have either a direct or indirect relationship. For example, we would imagine that any of the HP Ink products would be likely to be purchased along with a USB printer cable since this would make sense as these two products would have direct usage if a customer had an HP printer.

B2. Example of Transactions in the Dataset

For this example, I have chosen Transaction on line 35 of the dataset. This will consist of the following:

- Logitech M510 Wireless Mouse
- HP 61 ink
- Falcon Dust Off Compressed Gas
- HyperX Cloud Stinger Headset
- HP 925 ink
- Premium Nylon USB Cable

B3. Summary of One Assumption for Market Basket Analysis

The underlying assumption that we are basing this analysis on is that fact that the joint occurrence of either two or more product/services in most baskets imply that they are complements in purchase and thus the selection of one will lead to the selection of other **(Wagner A. Kamakura 2012).**

Part III: Data Preparation

C1. Transformation of the dataset to be compatible with Market Basket Analysis Technique

```
# Standard imports
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline


# Loading Dataset to begin Transformation
teleco =
pd.read_csv(r"C:\Users\andre\OneDrive\Desktop\teleco_market_basket.csv")


teleco.columns


Index(['Item01', 'Item02', 'Item03', 'Item04', 'Item05', 'Item06', 'Item07
',
       'Item08', 'Item09', 'Item10', 'Item11', 'Item12', 'Item13', 'Item14
',
       'Item15', 'Item16', 'Item17', 'Item18', 'Item19', 'Item20'],
      dtype='object')


teleco.shape


(15002, 20)


teleco.describe()


teleco.head()
```

```
# Check to see if there are any missing values in the dataset before analysis
data_nulls = teleco.isnull().sum()
print(data_nulls)
```

```
Item01     7501
Item02     9255
Item03    10613
Item04    11657
Item05    12473
Item06    13138
Item07    13633
Item08    14021
Item09    14348
Item10    14607
Item11    14746
Item12    14848
Item13    14915
Item14    14955
Item15    14977
Item16    14994
Item17    14998
Item18    14998
Item19    14999
Item20    15001
dtype: int64
```

```
# Drop the records that are found to have missing values in order to increase
accuracy of our dataset
teleco.dropna(how='all', inplace=True)

teleco.head()
```

```
# Replace the NaN values with 0 of the remaining records
teleco.fillna(0, inplace=True)
```

```
teleco.head()
```

```
# Transform the dataset into a list format for our data analysis
teleco_list = []
for i in range(0, 7501):
    teleco_list.append([str(teleco.values[i, j]) for j in range(0, 20)])
teleco_cleaned = pd.DataFrame(teleco_list)


teleco_cleaned.head()


# Extract our cleaned and prepared dataset
teleco_cleaned.to_csv('prep_teleco_market_basket.csv')
```

C2. Code used to Generate Association Rules for Apriori Algorithm

```
# Utilize the apriori algorithm which identifies items and expands them in
our frequent item set
!pip install apyori
from apyori import apriori

# Train algorithm
rule_list = apriori(teleco_list, min_support = 0.003, min_confidence = 0.3,
min_lift = 3, min_length = 2)


rule_list = list(rule_list)
results = pd.DataFrame(rule_list)


# Observe results list
results


# Create a separate dataframe for support
support = results.support
```

```
#Create our four lists containing Left Hand Side, Right Hand Side,
Confidence, and Lift
first_values = []
second_values = []
third_values = []
fourth_values = []


#Create a loop that repeats over our list
for i in range(results.shape[0]):
    single_list = results['ordered_statistics'][i][0]
    first_values.append(list(single_list[0]))
    second_values.append(list(single_list[1]))
    third_values.append(single_list[2])
    fourth_values.append(single_list[3])



#Convert our list into a proper dataframe
lhs = pd.DataFrame(first_values)
rhs = pd.DataFrame(second_values)
confidence = pd.DataFrame(third_values, columns=['confidence'])
lift = pd.DataFrame(fourth_values, columns=['lift'])


#Combine the newly created list dataframe
results_final = pd.concat([lhs, rhs, support, confidence, lift], axis=1)
results_final.fillna(value=' ', inplace=True)


results_final
```

C3. Values for Support, Lift, and Confidence

```
# Setting the column names
results_final.columns = ['lhs', 1, 2, 'rhs', 1, 2, 'support', 'confidence',
'lift']
results_final_1 = results_final[['lhs', 'rhs', 'support', 'confidence',
'lift']]
results_final_1


results = list(rule_list)
for i in results:
    print('\n')
    print(i)
    print('**********')
```

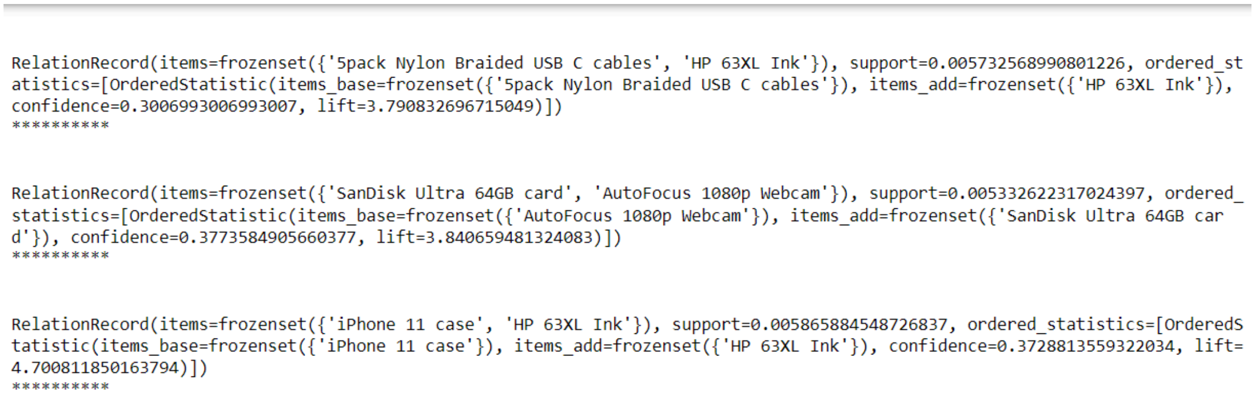
C4. Top 3 Rules that are generated by Apriori Algorithm

I. Rule 1: Customers who purchase the 5 pack Nylon Braided USB C Cables then also go on to include the HP 63XL Ink in their basket
    - This rule states that out of all the customers who decided to purchase the Nylon Braided USB C Cables, 30% of them then went on to purchase the HP 63XL Ink in their basket as well. This is supported by the confidence level we obtain of 0.3007 or roughly translating to 30%.
    - This rule states that only 0.57% of the total number of transactions collected have both items together in a purchase. This is supported by the support level of 0.0057.
    - This rule states that when a customer makes the decision to purchase the Nylon Braided USB C Cable, they are increasingly likely to purchase the HP 63XL Ink by a factor of just below 3.8 times as compared to if the Nylon USB Cable pack were not purchased at all. This can be supported by the Lift level of 3.7908 which in this case will translate to just below 3.8 times.

II. Rule 2: Customers who purchase the Auto Focus Webcam 1080p then also go on to include the SanDisk Ultra 64GB Card in their basket
    - This rule states that out of all the customers who decided to purchase the Auto Focus Webcam 1080p, 37.7% of them then went on to purchase the SanDisk Ultra 64GB Card in their basket as well. This is supported by the confidence level of 0.3774 or roughly translating to 37.7%.
    - This rule states that only 0.53% of the total number of transactions collected have both items together in a purchase. This is supported by the support level we obtain of 0.0053.
    - This rule states that when a customer makes the decision to purchase the Auto Focus Webcam 1080p, they are increasingly likely to purchase the SanDisk Ultra 64GB Card by a factor of just over 3.8 times as compared to if the Webcam were not purchased at all. This can be supported by the Lift level of 3.8407 which in this case will translate to just above 3.8 times.

III. Rule 3: Customers who purchase the IPhone 11 Case then also go on to include the HP 63XL Ink in their basket
    - This rule states that out of all the customers who decided to purchase the IPhone 11 Case, 37.3% of them then went on to purchase the HP 63XL Ink in their basket as well. This is supported by the confidence level we obtain of 0.3729 or roughly translating to 37.3%.
    - This rule states that only 0.51% of the total number of transactions collected have both items together in a purchase. This is supported by the support level of 0.0051.
    - This rule states that when a customer makes the decision to purchase the IPhone 11 Case, they are increasingly likely to purchase the HP 63XL Ink by a factor of just over 4.7 times as compared to if the IPhone 11 Case were not purchased at all. This can be supported by the Lift level of 4.7008 which in this case will translate to just above 4.7 times.

Screenshot of values for top 3 support, lift, and confidence association rules:



```
RelationRecord(items=frozenset({'5pack Nylon Braided USB C cables', 'HP 63XL Ink'}), support=0.005732568990801226, ordered_st
atistics=[OrderedStatistic(items_base=frozenset({'5pack Nylon Braided USB C cables'}), items_add=frozenset({'HP 63XL Ink'}),
confidence=0.3006993006993007, lift=3.790832696715049)])
**********


RelationRecord(items=frozenset({'SanDisk Ultra 64GB card', 'AutoFocus 1080p Webcam'}), support=0.005332622317024397, ordered_
statistics=[OrderedStatistic(items_base=frozenset({'AutoFocus 1080p Webcam'}), items_add=frozenset({'SanDisk Ultra 64GB car
d'}), confidence=0.3773584905660377, lift=3.840659481324083)])
**********


RelationRecord(items=frozenset({'iPhone 11 case', 'HP 63XL Ink'}), support=0.005865884548726837, ordered_statistics=[OrderedS
tatistic(items_base=frozenset({'iPhone 11 case'}), items_add=frozenset({'HP 63XL Ink'}), confidence=0.3728813559322034, lift=
4.700811850163794)])
**********
```

Part IV: Analysis

D1. Summary of Support, Lift, and Confidence Results

The results that we obtained above can be summarized as follows:

A. Our support level of all three of the top 3 rules hovered in the 0.5% range. This is not as significant as we would have liked since the item groupings in each rule only occurs for around 0.5% of all the observed transactions. Thus the cases where these groupings happened were very small regardless of our lift and confidence rates.

B. Our lift ratio displayed a significant result as our highest was a factor of 4.7 times more likely between the Iphone 11 case and the HP 63XL Ink in rule 3. Our lowest factor was 3.79 times in rule 1 which again is still quite significant between the Nylon USB Cable pack and the HP 63XL Ink.

C. Our highest rate of confidence was that of 37.7% from rule 2. This doesn’t speak too well to the results as the highest rate of confidence is quite a significant portion below an optimal value of around 80-90%. Our lowest confidence was found to be at 30% from rule 1, which again is in the 30’s percentage wise. Although it is enough to not be negligent, we definitely would have liked to see a rule in which a confidence rate was much higher

D2. Significance of Findings

Unfortunately, we see from our results that there is little significance to our analysis. The reason for this is due to the fact that our confidence levels are quite a bit below 50%. This means that we are not even able to say with a high degree of certainty that the resulting item sets will be purchased. Also looking at the support levels, there are barely any item sets purchased within the same transactions as we are looking at an incredibly small percentage of 0.5%. Finally, out of the small percentage (0.5%) of customers who do decide to purchase the itemset, they are roughly around 4 times more inclined to do so through the interpretation of our lift factor.

D3. Recommended Course of Action

Due to our Results and Findings, I think that it is not recommended to take any practical decisions or measures according to this data analysis as they are not significant for leveraging discounts or bundle promotions. The only action recommended is to obtain more data so that we can come back to this again and see if we are able to find a more distinct item bundling results that will help create more incentives for purchases and subdue customer churn.

Part V: Attachments

Panapto Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=4e90d2d2-d3b8-4b99-846c-aea6003432fc

E. Third Party Codes

Chris Moffitt (2017). Introduction to Market Basket Analysis In Python.
https://pbpython.com/market-basket-analysis.html

Jihargifari (2020). How To Perform Market Basket Analysis in Python.
https://medium.com/@jihargifari/how-to-perform-market-basket-analysis-in-python-bd00b745b106

Benjamin Naibei (2021). Getting Started with Apriori Algorithm in Python.
https://www.section.io/engineering-education/apriori-algorithm-in-python/

Bashir Alam (2022). The Apriori Algorithm: What It Is And How To Use It In Python.
https://hands-on.cloud/implementation-of-apriori-algorithm-using-python/

F. Source References

Amruta Kadlaskar (2021). A Comprehensive Guide on Market Basket Analysis.
https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

Wagner A. Kamakura (2012). Sequential Market Basket Analysis.
http://wak2.web.rice.edu/bio/My%20Reprints/Sequential%20Market%20Basket%20Analysist.pdf