Maket Basket Analysis

Code

Research Question

How can our business increase customer retention? Is there something we can do that will drive customer acquisition, while at the same time, increasing customer retention? Will increasing items per transaction increase customer retention? This analysis will fit with a broader company goal of increasing our customer retention rates.

The goal of this analysis is to understand what products our customers are typically purchasing together. Understanding this will help us to market and merchandise products together. It will allow us to provide a better experience for our customers.

This analysis will help us to understand our customers as a whole, but there are limitations with our dataset. Typically, we would want to tie these purchases to customers, so that we can build a better understanding of our customers. This dataset does not have customer_ids to do this. It will still be profitable in understanding our customers and how we can adjust our merchandising strategy to increase items per transaction.

I will be using Market Basket Analysis to conduct this research.

How it Works

Market basket analysis is a technique that looks for associations between items of an itemset. Itemsets are made up of a left-hand side and a right-hand side. The metrics describe the relationship between the left side and the right side. According to Susan Li, Market Basket Analysis, "works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items people buy." (2017)

To being, I will read in the data set and view it to see what we are working with.

Hide

```
head(df,5)
```

Item01 <chr>	Item02 <chr>	Item03 <chr>
NA	*NA*	*NA*
Logitech M510 Wireless mouse	HP 63 Ink	HP 65 ink
NA	*NA*	*NA*
Apple Lightning to Digital AV Adapter	TP-Link AC1750 Smart WiFi Router	Apple Pencil
NA	*NA*	*NA*

5 rows | 1-3 of 20 columns

The data is laid out with a row per transaction and a column per item number.

The dataset has empty rows between each transaction. I will need to clean this up before I can do any analysis.

With the data now cleaned up, I will transactionalize the dataset and view the summary statistics. The summary tells us the most frequent items, the distribution of items per transaction, and the summary stats of our transactions. I will also visualize the top 15 items in the dataset.

Hide

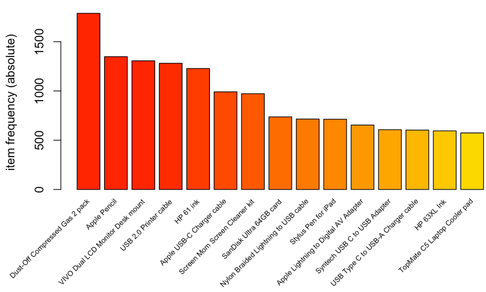
```
trans = read.transactions("transactions.csv", format="basket", sep=",")
summary(trans)
```

This bar plot shows us the top 15 items by volume.

Compressed Gas 2 pack, Apple Pencil and VIVO Dual LCD Monitor Desk mount were the top three items sold.

Hide

```
itemFrequencyPlot(trans, topN=15, type="absolute",cex.names=0.6, col = rainbow(99))
```



Sample Transaction

Below is a sample transaction. This transaction includes an Apple Lighting to Digital AV Adapter, Apple Pencil, and a TP-Link AC1750 Smart WiFi Router.

Hide

```
inspect(trans[3])
```

This code generates the association rules for our transactions.

Hide

```
#Get Rules
rules <- apriori(trans, parameter=list(supp=0.001, conf=0.8))
rules<- sort(rules, by="confidence", decreasing=TRUE)
```

The code breaks the items into the left and right sides and generates the association statistics of our market basket analysis.

Market Basket Analysis assumes that items that appear together in transactions are connected. The analysis of the statistics tells us the likelihood of the combinations appearing together again.

I am using the apriori function from the arules package. This function produces four useful statistics that helps us analyse our market baskets.

Support

Support is a calculation that tells us how often the item set is seen in our data. It is calculated by taking the number of transactions with the rule divided by the number of transactions.

Confidence

Confidence is the probability that the itemset will be seen in future transactions with these items divided by the probability of seeing just one of the items. The range for this KPI ranges from 0 to 1. The closer to 1 the more confidence you can have that these items will be purchased in combination in the future.

Lift

Lift tells us if there is an independent relationship between items on the left and right sides. This KPI is the proportion of transactions with an itemset divided by the proportion of transactions containing each item in the set. When this metric falls below 1, then the item(s) on the left side decreases the likelihood of seeing the items on the right side purchased together.

When the ratio is above 1, then the item(s) on the left side are increasing the likelihood of seeing the right side purchased together.

Count

Count is self explanatory. It is the number of times the itemset is seen in the data set.

Generating the Rules

Apriori lets us constrain the results by designating the support and confidence amounts. I have done this below by setting the support to 0.001 and the confidence to 0.8. I have also sorted the rules by confidence in descending order.

Analysis of Rules

The top three association rules are displayed below. Each have a confidence of 1.

Take the rule in the three below. Transactions containing Apple Lighting to USB Cable, FEIYOLD glasses and the 64GB memory card are also likely to contain two pack of Dust-Off Compressed air.

The support for this rule is 0.0012, meaning this set was seen .01% of all transactions. That's not very many transactions.

The model is very confident that the compressed air will be purchased again with this item combination; 100% confident.

This itemset produces a lift of 4.2 when purchased with the itemset.

Two of the three top rules produced the Dust-Off Compressed Air 2 pack on the right side of the association rules. I wonder if there was a promotion for this item? This isn't something that I will answer in this analysis, but I wanted to call it out for possible future digging.

Hide

```
#Get Rules
rules <- apriori(trans, parameter=list(supp=0.001, conf=0.8))
```

Apriori

Parameter specification:

confidence <dbl>	minval <dbl>	smax <dbl>	arem <chr>	aval <lgl>	originalSupport <lgl>	maxtime <dbl>	support <dbl>	minlen <int>
0.8	0.1	1	none	FALSE	TRUE	5	0.001	1

1 row | 1-10 of 12 columns

Algorithmic control:

filter <dbl>	tree <lgl>	heap <lgl>	memopt <lgl>	load <lgl>	sort <int>	verbose <lgl>
0.1	TRUE	TRUE	FALSE	TRUE	2	TRUE

1 row

Absolute minimum support count: 7

set item appearances …[0 item(s)] done [0.00s]. set transactions …[7640 item(s), 7502 transaction(s)] done [0.01s]. sorting and recoding items … [116 item(s)] done [0.00s]. creating transaction tree … done [0.00s]. checking subsets of size 1 2 3 4 5 6 done [0.01s]. writing … [72 rule(s)] done [0.00s]. creating S4 object … done [0.00s].

Hide

```
rules<- sort(rules, by="confidence", decreasing=TRUE)
```

Summary of The Rules

Summary of rules displays the measures from the Market Basket Analysis. The average transaction contains four items. On average, the model has a confidence level of 85%. The average lift is 4.8. This confidence number makes sense considering I set the confidence to 0.8 when I created the rule.

Hide

```
summary(rules)
```

```
set of 72 rules

rule length distribution (lhs + rhs):sizes
 3  4  5  6 
14 42 15  1 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      3       4       4       4       4       6 

summary of quality measures:
    support         confidence        coverage            lift            count       
 Min.   :0.00107   Min.   :0.80   Min.   :0.00107   Min.   : 3.4   Min.   : 8.0  
 1st Qu.:0.00107   1st Qu.:0.80   1st Qu.:0.00133   1st Qu.: 3.4   1st Qu.: 8.0  
 Median :0.00113   Median :0.83   Median :0.00133   Median : 3.8   Median : 8.5  
 Mean   :0.00126   Mean   :0.85   Mean   :0.00148   Mean   : 4.8   Mean   : 9.4  
 3rd Qu.:0.00133   3rd Qu.:0.89   3rd Qu.:0.00160   3rd Qu.: 4.9   3rd Qu.:10.0  
 Max.   :0.00253   Max.   :1.00   Max.   :0.00267   Max.   :12.7   Max.   :19.0  

mining info:
```

data <chr>	ntransactions <int>	support <dbl>	confidence <dbl>
trans	7502	0.001	0.8

1 row | 1-5 of 5 columns

Recommended Action

As I mentioned above, the two pack of compressed air were included in two of the top three rules. This is not surprising because this item was the #1 item sold. I would recommend reevaluating the dataset with this item removed to see how the basket rules change.

I wold also recommend including customerIDs to allow us to implement this analysis to specific customers for our marketing department to target in the future.

References

Li, S. (2017, September 24). A Gentle Introduction on Market Basket Analysis—Association Rules. *Towards Data Science*. https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

https://rstudio-pubs-static.s3.amazonaws.com/267119_9a033b870b9641198b19134b7e61fe56.html