Date: September 14, 2022
To: Mitch Cochran, MA, MHS, CISM, CGCIO
From: Thao Hau Nguyen, BA 61 NEU Vietnam

Subject: Market Basket Analysis applied Association rules in Python

In this report, I will describe the dataset related to information about items from 5617 receipts, which will help retailers to uncover associations among these types of goods and further increase sales.

Firstly, description of the dataset about the shopping basket is shown below with the following attributes:

- ReceiptsID: This is the distinct information about each customer's invoice which is recorded immediately after a consumption.
- juices: A yes/no column indicating whether the customer purchased juices in their cart.
- paper_goods: A yes/no column indicating whether the customer purchased paper goods such as paper napkins, towels, and toilet tissue in their cart.
- frozen_goods: A yes/no column indicating whether the customer purchased frozen goods in their cart.

The same explanations about the choices of customers whether they purchased with the rest of columns: snack_foods, canned_goods, beer_wine_spirits, dairy, breads, produce, desserts and meats.

Secondly, I applied Association rules in Python to do shopping basket analysis. The main objective of analysis is to find which products are most frequently purchased together. In the data preparation, to reduce the data set to only those attributes related to our question, I deleted the ReceiptID column and made sure that all products that customers bought or did not buy during shopping would now be represented by values 1 and 0.

```
df = pd.read_csv('/Users/nguyenthaohau/Desktop/National Economic University/NEU 2022-2023 HK01/Data Mining/Exercise chapter 5/Chapter05Exercise.csv')
del df['ReceiptID']
df.head()
```

✓ 0.3s Python

	juices	paper_goods	frozen_foods	snack_foods	canned goods	beer_wine_spirits	dairy	breads	produce	desserts	meats
0	1	0	1	0	0	0	0	0	1	0	1
1	1	1	0	0	0	0	1	0	0	1	0
2	1	1	1	0	1	1	1	1	1	1	1
3	0	1	1	0	0	0	0	0	1	1	1
4	0	0	0	1	0	1	0	0	0	0	0

The next step is to create the Apriori Model. For this, I set a min_support value with a threshold value of 15% and printed them on the screen as well. Then, I chose the 55% minimum confidence value. In other words, when product X is purchased, we can say that the purchase of product Y is 55% or more:

```
rules[ (rules['confidence'] >= 0.55) ]
```

✓ 0.4s Python

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(snack_foods)	(frozen_foods)	0.336122	0.338615	0.188713	0.561441	1.658051	0.074897	1.508086
1	(frozen_foods)	(snack_foods)	0.338615	0.336122	0.188713	0.557308	1.658051	0.074897	1.499638
2	(beer_wine_spirits)	(frozen_foods)	0.334698	0.338615	0.186220	0.556383	1.643114	0.072887	1.490892
3	(snack_foods)	(beer_wine_spirits)	0.336122	0.334698	0.187823	0.558792	1.669541	0.075323	1.507911
4	(beer_wine_spirits)	(snack_foods)	0.334698	0.336122	0.187823	0.561170	1.669541	0.075323	1.512836
5	(snack_foods, frozen_foods)	(beer_wine_spirits)	0.188713	0.334698	0.155955	0.826415	2.469135	0.092793	3.832717
6	(snack_foods, beer_wine_spirits)	(frozen_foods)	0.187823	0.338615	0.155955	0.830332	2.452142	0.092356	3.898108
7	(frozen_foods, beer_wine_spirits)	(snack_foods)	0.186220	0.336122	0.155955	0.837476	2.491580	0.093362	4.084799

In looking at the rules (where metric is confidence with the min equals 55%), some combinations such as the snack foods and frozen foods, or beverage products (beer, wine or spirits) and snack or frozen foods are purchased together. Snack foods, frozen foods and beverage products are purchased together in a manner that is significantly higher than the overall probability would suggest (with the confidence are all over 82%). With the min confidence is 55%, the purchases among snack foods, frozen foods and spirit drinks are related with more than over 15% of the observations in the data set which supported them. However, their confidence percentages did very as the premises and conclusions were reversed.

Besides, I also saw that the support percentage of carts which includes above 3 items is smaller than basket with 2 items. The reason explained for this can be purchase with more items requires more cash and need to consider some other external factors from the customers. But this is not affected the association among 3 items. From my opinion, the basket with 2 or 3 items all supports the sales of goods and increase the revenue for the retailer.

Because I set metric as confidence, I then check the lift column to reduce the drawbacks of the first metric. The lift of 7 above associations is greater than 1, which means that item Y is *likely* to be bought if item X is bought. For example, in the first index value, people who buy snack foods is likely to buy frozen foods because the lift is greater than 1. Moreover, looking at the leverage, consumers who buy snack foods will likely consume 7.5% more than frozen foods users who don't buy snacks.

Therefore, there are numerous ways to increase the profit if snacks, frozen and beverage products are purchased more frequently. For examples:

- Combining products will increase cross-selling.
- When specific things are kept together, the shop layout might be modified to increase sales.
- It is possible to enhance sales of products that customers do not purchase such as meats, canned goods or dairy products by engaging in promotional activities, which are an advertising campaign.
- If the client purchases 3 best-seller items, a combined discount can be provided.

It’s all for my analysis. Thank you for reading!

Thao Hau Nguyen