MKT 680 Marketing Analytics

Report for Project 2: Recommender Systems

Group 7: Myra Liu, Khasim Shaik, Jacky Yang, Jingwen (Kivi) Zuo

Date: 02/28/2019

Overview

This report is about recommending 2 products from at most 5 suppliers to at least 500 selected customers who purchased more than $5,000 in 2017 from Pernalonga, a leading supermarket chain in Lunitunia. Pernalonga wants to experiment on personalized promotions funded by partnering with suppliers to grow revenue. The objectives are to target the most profitable and promotion-sensitive customers and to recommend products with the highest expected value. In addition, the metrics applied to define the target customers and products will be discussed in more details.

Business & Data Understanding

Firstly, some issues are identified to better prepare the data:

- Bags is the most frequently purchased item in the dataset, but bags should not be included to conduct the recommender system from the business understanding.
- The transaction ID is not unique for all transactions. Base on the assumption that a transaction is made when a customer visits a specific store on a given day, the unique identifier for a transaction should be a combination of customer ID, transaction ID and store ID.
- A dubious transaction identified in the dataset is that one transaction has a negative paid amount. Without further details about this transaction, it is removed.
  Product id: 357541011, cust_id: 93409897 tran_prod_paid_amt: -0.55

Secondly, the focus of analysis in the project is based on transaction level, as the metrics created measure the impact of promoted products' on transaction values. It is more valuable to find products that induce more profitable transactions for the most promotion-sensitive customers.

Thirdly, due to the fact that some products sold by the retailer do not belong to any supplier as "No Label" or belong to the retailer as "Private Label", they do not satisfy the requirement of partnering with suppliers. They are removed from the dataset.

Target Customer

The complete data includes over 29 million historical transactions for 7,920 customers in 2016 and 2017. According to the scope of the project, there are two thresholds applied first, and there are 1,897 customers left in the dataset:

- Customers who purchased over L$5,000 in 2017
- The product that was purchased at least 15 times in 2017

In order to narrow down the scope and efficiently target a certain group of customers who are most likely to purchase more when they have promoted items in the transactions, two criteria are created to assess and filter the target customer. To clarify some of the major dollar amounts that we will use below, the sales amount equals to the sum of the discounted amounts and the paid amounts in the dataset.

- Profitability measures the profits generated in a transaction in a ratio of weighted average profits over the total paid amount generated on the customer level. Without the products' cost information, the lowest discounted price of a given product will be used as a benchmark of cost to measure the same product's profits in other transactions, as it is assumed that the lowest discounted price is higher than or equal to the actual cost. The importance factor used to calculate the weighted average profits for a customer is the ratio of a single transaction's paid value over the customer's total paid value, and the aggregated profitability is the sum of the product of the ratio and the profits generated over total paid amounts in the transaction.

$$Weighted\ avg.\ profits = Sum\left(\frac{a\ transaction's\ paid\ value}{a\ customer's\ total\ paid\ value} \times a\ transaction's\ profits\right)$$

$$Profitability = \frac{Weighted\ avg.\ profits}{Total\ paid\ value}\ for\ a\ given\ customer$$

The final profitability for a customer is measured on the transactional level, as it is important to identify the profits triggered by a promoted product. The process can be explained more clearly as below:

For each customer:
    For each transaction:
        For each product:
        Calculate the profits
    Sum the profits for transactions with discounted items
Calculate the ratio of profits/total sales

- Promotion-sensitivity measures if a customer's transaction value increases with higher promotion level, and it is defined by the correlation of transactions' weighted average promotion levels and sales value for a given customer within a timeframe. The importance factor used to calculate weighted average promotion level is the same as for profitability, and the promotion level equals the transaction's total discounted amount divided by total sales amount.

$$Weighted\ avg.\ promo.\ level = Sum\left(\frac{a\ transaction's\ paid\ value}{a\ customer's\ total\ paid\ value} \times a\ transaction's\ promo.\ level\right)$$

$$= \frac{A\ customer's\ total\ discounted\ value}{A\ customer's\ total\ sales\ value}$$

The process can be explained more clearly as below:

For each customer:
    For each transaction:
        For each products:
        Calculate the discount level
    Calculate the weighted average discount level on transaction level based on the product's paid amount/transaction value
Conduct the correlation of the weighted average discount level vs. transaction value

The final target customer is filtered by a threshold of 0.5 on both profitability and promo-sensitivity. Therefore, there are 1,348 customers left as the customer base for the recommender system.

Similarity Measurement

After narrowing down to the customer base, the next step is to calculate the similarity between all pairs of customers using their baskets of promoted products purchased in 2017. The input for the cosine similarity algorithm is a matrix of binary variable on only promoted products for each customer. We used cosine similarity rather than jaccard similarity because the similarity result will be the same when the "rating input" is binary. The binary matrix is used because of lack of rating information and the similarity is based on if the customer purchased a promoted product. The columns represent the product pool that includes all promoted products bought by the target customers. The rows represent the basket for each customer in vectors. Using only promoted products is to match with incremental sales that are calculated based on promoted products.

Example below:
Matrix of customer-product (1: purchased, 0: not purchased)

prod_id	145519008	145519009	145519010	145519011	145519012	148066012	152576008	152576009	152576010	152576011	...	999996327	999996335
cust_id													
29568	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
29909	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
39856	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
289996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
329968	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0
339627	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
550000	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	...	0.0	0.0
559804	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0
709543	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0

Matrix of similarity score

```
1 cosine_sim
```

cust_id	29568	29909	39856	109693	289996	299749	329968	339627	550000	559804	...	99569634	99569937	99579555
cust_id														
29568	0.000000	0.145913	0.160532	0.136753	0.112693	0.154232	0.182402	0.160581	0.227255	0.115634	...	0.114461	0.116369	0.120433
29909	0.145913	0.000000	0.168587	0.135266	0.185779	0.171765	0.167499	0.155666	0.176300	0.158150	...	0.185171	0.144945	0.125139
39856	0.160532	0.168587	0.000000	0.085088	0.157458	0.150848	0.208132	0.185535	0.182110	0.125663	...	0.153531	0.116864	0.143413
109693	0.136753	0.135266	0.085088	0.000000	0.117892	0.134240	0.113980	0.115783	0.134252	0.137635	...	0.128745	0.121737	0.082465
289996	0.112693	0.185779	0.157458	0.117892	0.000000	0.130478	0.169963	0.145390	0.184386	0.146500	...	0.212189	0.171211	0.169136
299749	0.154232	0.171765	0.150848	0.134240	0.130478	0.000000	0.176295	0.114414	0.141316	0.134529	...	0.132755	0.082012	0.132274
329968	0.182402	0.167499	0.208132	0.113980	0.169963	0.176295	0.000000	0.141194	0.197369	0.118305	...	0.170085	0.115465	0.119932
339627	0.160581	0.155666	0.185535	0.115783	0.145390	0.114414	0.141194	0.000000	0.164919	0.140896	...	0.159484	0.131366	0.121828
550000	0.227255	0.176300	0.182110	0.134252	0.184386	0.141316	0.197369	0.164919	0.000000	0.097542	...	0.140234	0.137088	0.107481

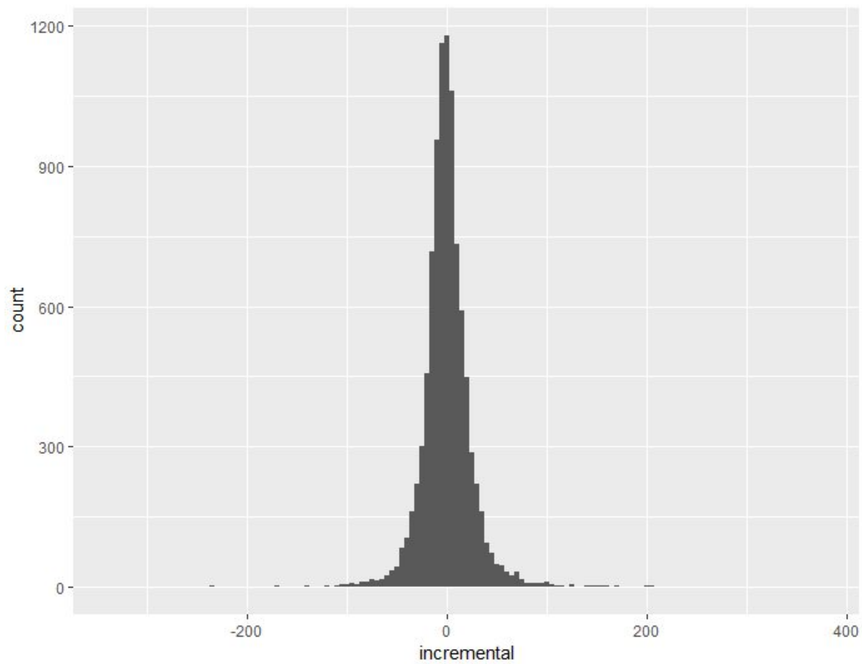
Incremental Sales

Before computing the purchase probability, incremental sales is also calculated to be implemented to the select the products that will be recommended to customers. Incremental sales measure the difference between a product's weighted average transaction value with itself on promotion and without promotion on product level.

The transactions applied to the calculation is firstly filtered by the target customer base, as the incremental sales should be specifically targeting the selected customer base. The weighted average sales for each product is calculated as below with promotion and without promotion respectively:

$$Weighted\ avg.\ trans\ value = Sum\left(\frac{a\ product's\ paid\ value}{a\ product's\ total\ paid\ value} \times a\ transaction's\ paid\ amount\right)$$

The distribution of incremental sales for products purchased by the target customer is normally distributed with a mean of 0.1 and a maximum of $374.1.



A subset of top-ranked incremental sales is shown below:

	prod_id	avg_sales_np	avg_sales	incremental
9964	999954065	51.70458	425.76583	374.06125
1316	999168161	30.65192	344.51008	313.85817
1317	999168162	35.58303	307.41766	271.83463
8093	999470815	54.24921	307.34138	253.09217
7013	999356553	90.83426	311.79582	220.96156
1318	999168163	48.96665	255.33424	206.36759
3712	999231755	70.23990	276.16969	205.92979
5398	999269716	41.12826	243.34000	202.21174
1315	999168160	34.24799	233.19399	198.94600
1201	999165814	43.38640	241.84000	198.45360
1314	999168159	50.58768	241.30214	190.71446
1312	999168027	64.63276	249.60814	184.97538
2250	999181191	52.97822	235.01000	182.03178
7176	999364679	59.24846	231.41145	172.16299
9491	999749706	65.28147	233.39000	168.10853
141	233442011	16.70000	181.30000	164.60000
10293	999995048	66.02219	227.63824	161.61605

Purchase Probability

Next, the top 20 most similar customers are identified for each target customer by the ranking of similarity scores. A product pool is formed by all products purchased by this group of customers. The products that a customer has purchased are not eliminated because promotion of purchased products could also induce a larger transaction, which will be assessed by incremental sales in next step.

Therefore, each product's probability of purchase within the customer group could be calculated by the ratio of customers who purchased out of all 21 customers. As a result, each customer has a list probability of buying a product. The list of purchase probability is embedded in the codes and implemented to calculate expected value.

Expected Value

The expected value is measured by the product of purchase probability and incremental sales, and it is used to comprehensively rank the top products that will be recommended to customers. As the recommended products have to be limited within 5 suppliers, 5 top-ranked products are kept to better narrow down the supplier selection in the next step. A subset of 5 top-ranked products for customer 29568 and 29909 is shown below.

Customer	Product	Expected Value
29568	999885829	15.97917033
29568	999378733	10.3837859
29568	999424824	7.901821439
29568	999630595	7.307987668
29568	999749460	7.085251978
29568	999259780	6.810730815
29568	999958544	6.748552405
29568	999662852	6.654033931
29568	999749463	6.451385908
29568	999274617	6.420384948
29909	999953616	15.93981523
29909	999885829	12.78333627
29909	999457945	10.5407654
29909	999364679	9.481683951
29909	999749460	8.097430832
29909	999424824	7.901821439
29909	153701005	7.078732571
29909	999958544	6.748552405
29909	999274617	6.420384948
29909	999378733	6.057208439

Supplier Selection

Due to the scope of the project, at the meantime of recommending two products to target customers, the suppliers should also be limited to at most five. A more comprehensive view is to create a table of customers, products and the product's expected value for this customer. The total expected value will help with identifying the top 5 suppliers that will generate the highest expected value in the target customer base.

prod_id	brand_desc	category_desc_eng
999259753	FERRERO ROCHER	BONBONS
999958544	FERRERO ROCHER	BONBONS
999378733	FULA	OIL
999424824	MIMOSA	CHEESE TYPE FLAMENGO
999383364	PERECÍVEIS CARNE	FRESH PORK
999749460	PERECÍVEIS CARNE	FRESH BEEF
999749463	PERECÍVEIS CARNE	FRESH BEEF
999749469	PERECÍVEIS CARNE	FRESH BEEF

The final suppliers are decided as Ferrero Rocher, Fula, Mimosa and Pereciveis Carne by purchase frequency on promotion in the target customer base. Although Serrata ranks as the third in the list, it is removed from the list because only one oil supplier should kept and Fula generates higher expected value in the target customer base.The total expected value of the four suppliers is about \$16,012 for the target customer base and an average of \$20.63 from each customer.

Final Recommended List

The complete list can be referred to the “Recommendation_list.csv” file.

prod_id	customer	brand_desc	ExpectedValue	category_desc_eng
999378733	29568	FULA	$ 10.38	OIL
999424824	29568	MIMOSA	$ 7.90	CHEESE TYPE FLAMENGO
999378733	39856	FULA	$ 10.38	OIL
999424824	39856	MIMOSA	$ 10.87	CHEESE TYPE FLAMENGO
999749460	289996	PERECÃVEIS CARNE	$ 11.13	FRESH BEEF
999749463	289996	PERECÃVEIS CARNE	$ 9.03	FRESH BEEF
999424824	329968	MIMOSA	$ 9.88	CHEESE TYPE FLAMENGO
999749460	329968	PERECÃVEIS CARNE	$ 11.13	FRESH BEEF