Which are more impactful, profitable customers or profitable products? An empirical study

Huong Nguyen
Stockholm School of Economics

Acknowledgements:
I am grateful to my supervisors, Sara Rosengren, Rickard Sandberg and Emelie Fröberg, for the their advices and insightful comments.

Cite as:
Nguyen Huong (2020), Which are more impactful, profitable customers or profitable products? An empirical study. *Proceedings of the European Marketing Academy*, 49th, (64094)

Paper from the 49th Annual EMAC Conference, Budapest, May 26-29, 2020.

Which are more impactful, profitable customers or profitable products?

An empirical study

-

Abstract

As discussed in the literature, shopping behaviors and cross-buying are two of the key drivers in customer lifetime value. This study examines how customer segments and the product categories impact the retailer's profitability (gross margin) in multiple regression models. Results show that the best customers, identified by Recency-Frequency-Monetary clustering, are potentially more impactful comparing with the products frequently bought together, defined by Market Basket Analysis. The findings can be leveraged to identify customers with most potential and implemented with corresponding strategies. Retailers are suggested to focus on their customer relationship management system and their best customers, especially households with high frequency. The method can be flexibly replicated with other point-of-sales data sets to uncover local patterns.

Keywords: *retailer profitability, RFM, market basket analysis*

Track: *Retailing & Omni-Channel Management*

1. Introduction

A.G. Lafley Chairman–CEO of Procter and Gamble Co. (P&G) once said "*Twenty percent of the brands and products account for 80 percent of sales*". As a result, in 2014, P&G decided to cut more than half of their brand portfolio and pay more attention to those 20 percent of customers who generate 80 percent of the sales (Schrage, 2014). Latest reports reveal the 65 brands that remain in P&G's mix account for 85 percent of the firm's topline and 95 percent of its profit (Lash, 2019).

This example shows how powerful it is to identify and leverage *the 20 percent*. Many retailers have access to the same type of data as P&G, but only a few utilize it to optimize their profitability (Kumar et al., 2017). This research is designed to help the retailers answer the question: *"Do profitable customers or profitable products have stronger impact on my gross margin?"*. To answer, I first segment the customers based on their Recency-Frequency-Monetary (RFM) variables, identify the potential product categories that are frequently bought together by market basket analysis (MBA) and finally, measure the impact of customers and categories on gross margin using multiple regression models.

2. Literature review

From both marketing literature and practice, an increasing number of metrics have been developed to measure retailer profitability (Petersen et al., 2009), among which customer life time value (CLV) and cross-buying/up-buying metrics are most relevant to the research question.

2.1. Customer value, RFM model and profitability

Inspired by the 80/20 rule, *"only certain types of customers are worth attracting and nurturing"* (Duboff, 1992, p.10), customer value has long been an appealing topic in marketing. One of the most well-known behavior-based methods is RFM analysis, which extracts customer profiles using three factors: the time since their last purchase (recency), the frequency with which they make purchases (frequency) and the amount they typically spend (monetary value). Each of the RFM variables has been shown to be a key driver in computing future customer profitability (Fader et al., 2005). RFM is also a tool for behavior-based segmentation, as based on the principle

that customers who have recently purchased, who have the highest frequency and who spent the most, are predicted to repeat the same behavior and be the most profitable (Yang, 2004). Despite limitations, RFM analysis has been considered an indicator to measure customer value, loyalty and is one of the most widely used techniques to identify the best customer-segment(s) (Hughes, 1996). It is especially favored by practitioners.

2.2. Cross-selling, MBA and profitability

Research shows that customers who cross-buy are more profitable than those who do not (Kumar et al., 2008). Therefore, cross-buying and up-buying is a common strategy to increase profit from current customers. Following the principle that purchase intention between categories is not independent (Sarantopoulos et al., 2016), MBA, or multicategory choice models uncover the pattern of interactions in purchase incidence decisions across product categories (Dippold & Hruschka, 2013). In other words, MBA analyses the buying habits of customers by finding associations between the items they place in shopping baskets. These results of *frequent itemsets*, or rules, help retailers to eventually increase the basket size and hence, profitability (Moodley et al., 2019).

Some researches have examined the relationship between product assortment and sales at stock-keeping-unit (SKU) or category levels (Boatwright & Nunes, 2001; Sloot et al., 2006). Looking from the *frequent itemsets* level is believed to better enhance profitability analysis because (1) it reflects the association of product categories and (2) financial contribution of *frequent itemsets* is interestingly higher than the categories or SKUs in isolation. Marketing activities on one category can be expected to influence the purchasing decision of the other(s), hence increasing profitability with optimal effort (Manchanda et al., 1999).

3. Method and data

A point-of-sales data set has been obtained from a Northen European grocery retailer. It includes all transactions by loyal members from 2014 to 2016 with the transaction date, customer ID, receipt ID, product ID, price, quantity and gross margin reported. Household-level is believed to reflect grocery purchase behavior better, hence “customer” in this report refers to “household”.

3.1. *Customer segmentation with RFM and K-means clustering*

RFM variables have been transformed as follows: *Recency* - Number of days from the last purchase to the day of analysis; *Frequency* - Number of visits throughout the observed period; *Monetary* - Average gross margin generated from each basket. These variables will then be used as inputs for *K-means* clustering process. To find the optimal number of clusters, gap statistics method was applied, $k = 10$ was chosen. The result is summarized in Table 1.

Table 1: 10 clusters, RFM values and margin contribution

Cluster	Recency (days)	Frequency (times)	Monetary (EUR)	Margin/customer (EUR)	Total margin (EUR)	Cluster (%)	Margin (%)
1	1	1089	5	5,369	434,922	1.46	4.15
2	812	25	6	149	13,416	1.62	0.13
3	27	49	6	301	358,426	21.42	3.42
4	9	161	7	1,157	1,292,565	20.11	12.33
5	4	550	7	3,793	2,048,237	9.72	19.54
6	30	72	24	1,710	321,429	3.38	3.07
7	5	269	8	2,171	2,337,777	19.39	22.30
8	4	747	6	4,342	1,085,416	4.50	10.35
9	5	393	7	2,897	2,529,262	15.72	24.13
10	355	59	7	411	61,274	2.68	0.58

Recency = Number of days since the last purchase; Frequency = Number of visits; Monetary = Gross margin per basket; Margin/customer = Monetary * Frequency; Total margin = Monetary * Frequency * Number of households who belong to that cluster; Cluster (%) = Number of households who belong to that cluster/Total number of households; Margin (%) = Total margin of that cluster / Total margin of all clusters

3.2. *Product categorization with MBA*

As suggested by both researchers (Bell et al., 2011) and practitioners (ECR Europe, 2011), basket-level is utilized in this step because every single visit is believed to carry valuable insights into the shopper needs. Regarding product taxonomy, category-level is chosen for two reasons: (1) shoppers' needs are often expressed at category level rather than SKU level (Griva et al., 2018) and (2) working at category level potentially enables more easily generalizable and replicable results. Applying *apriori* algorithm, the output is the set of association rules. In MBA, the performance of the rules is measured by *support, confidence* and *lift*. *Support* addresses how frequentlyt two or more categories, for example *Milk* and *Eggs*, are simultaneously contained in

one basket. *Confidence* is a conditional probability that determines how frequently $Eggs$ appear in a basket given that it contains $Milk$. *Lift* indicates the importance of the rule, measured by the ratio of two probabilities, that is the probability of $Milk$ and $Eggs$ co-occurring to the expected probability if the two groups of attributes were independent (Tan et al., 2002). To ensure the strong association, five typical $ItemSet(s)$ has been identified based on the *100-highest-Lift* rules (Table 2). One basket may contain none, one or several $ItemSet(s)$.

Table 2: *ItemSets* frequently bought together

ItemSet	Margin/Sales (%)	Margin (%)	Sales (%)
Necessities - Food and Beverage	18.58	23.66	29.25
Taco Friday	25.66	4.51	4.04
Chips and Dip	20.96	1.68	1.85
Necessities - Non Food and Beverage	22.54	0.90	0.91
Kids' Foods	15.55	0.27	0.40

Margin (%) = Margin of that ItemSet/Total margin; Sales (%) = Sales of that ItemSet/Total sales
ItemSets definitions:
Necessities - Food and Beverage - denoted as "NecessitiesFnB": Milk, Bread, Butter, Yogurt, Swedish yogurt, Ham, Sausage, Hard cheese, Eggs, Juice, Banana, Mat & bakfett, Smoke & Salted meat, Creme Fraiche, Whipping cream, Carrot, Grounded coffee, Flour, Frozen chicken, Frozen fish, Flakes, Pork, Smashed Potato, Confectionery, Orange, Apple, Mushroom, Feta, Sugar, Bread butter, Bread topping (31 items)
Taco Friday - denoted as "TacoFriday": Taco shell, Tortilla, Mexican sauce, Mexican spices, Ground meat, Canned corn, Cheese, Iceberg, Tomato and Cucumber (10 items)
Chips and Dip - denoted as "ChipsnDip": Chips, Dip mix, Sour Cream and Soft drink (4 items)
Necessities - Non - Food and Beverage - denoted as "NecessitiesFnB": Shampoo, Conditioner, Toilet paper, Toothpaste (4 items)
Kids' foods - denoted as "KidsFoods": Instant Oats, Canned food for 5-7 month-old kids, Canned food for 8-11 month-old kids, Kids' deserts & snacks (4 items)

3.3. Profitability model

The impact of $Clusters$ and $ItemSets$ on $Margin$ will be examined through these models:

$$Margin_{ki} = \beta_0 + \beta_1 Cluster_k + \varepsilon_{ki} \quad (1)$$

$$Margin_{ki} = \beta_0 + \beta_2 NecessitiesFnB_{ki} + \beta_3 NecessitiesNonFnB_{ki} + \beta_4 TacoFriday_{ki} + \beta_5 ChipsnDip_{ki} + \beta_6 KidsFoods_{ki} + \varepsilon_{ki} \quad (2)$$

$$Margin_{ki} = \beta_0 + \beta_1 Cluster_k + \beta_2 NecessitiesFnB_{ki} + \beta_3 NecessitiesNonFnB_{ki} + \beta_4 TacoFriday_{ki} + \beta_5 ChipsnDip_{ki} + \beta_6 KidsFoods_{ki} + \varepsilon_{ki} \quad (3)$$

where $k = 1, \ldots, n$ and n is the number of households; $i = 1, \ldots, m$ and m is a number of months. Accordingly, $Margin_{ki}$ (measured in EUR) is the gross margin household k generates in month i;

$Cluster_k$ is dummy variable, indicating the cluster that household k belongs to; $NecessitiesFnB_{ki}$ for example is the proportion of $NecessitiesFnB$ in the monthly groceries, calculated by the number of $NecessitiesFnB$ items that household k has bought in month i divided by the total number of items that household has bought in the given month (same calculation applied for the other $ItemSets$). By construction, these are the products frequently bought together i.e. there remain other categories that have been left out in case of (1) usually being bought alone and/or (2) not usually being bought with other categories repetitively. Accordingly, the number of items from these five $ItemSets$ would not add up to the total items bought and the proportion would not add up to 100%. The result is presented in Table 3.

4. Results and discussion

From step 3.1, 5555 households have been segmented into very uneven clusters with cluster sizes ranging from 0.13% to 24.13% (Table 1). At first glance, Cluster 9, 7 and 5 are the ones that contribute the most to total margin (in the whole observed period), who have also visited the store recently and quite frequently. Accordingly, it is tempting to conclude that they are the low hanging fruits. Noticeably, Cluster 1 and 8, who have the highest margin per customer, are the ones who have the lowest Monetary values and highest Frequency. Cluster 6 has the highest Monetary value yet relatively low Frequency, hence, the low margin contribution. This implies the importance of Frequency as compared with Monetary as regard to customer profitability. On the contrary, Cluster 3 and 4, whilst accounting for more than 42% of the customer base, generates only 15% of the profit; Cluster 2 and 10 who have not visited the store in 812 and 355 days are believed to be lapsers. In short, RFM analysis shows that to optimize profitability, half the customers should be prioritized whilst the others should be given low priority.

The results of step 3.2 is five $ItemSets$ (Table 2), which includes 53 items or 3.6% of the total number of products categories (there are more than 1,500 categories sold in the observed transactions) yet generate 37% of total sales or 31% of the overall margin. As a result, I have reason to believe these are the most promising categories. $TacoFriday$ is interestingly the most profitable $ItemSet$ with highest Margin over Sales ratio. As Manchanda et al. (1999) suggested, a potential cause of cross-category purchase is the complementary nature of the products. This is believed to be a controllable factor for marketers as marketing activities on one category, e.g.,

Table 3: Impact of *Clusters* and *ItemSets* on monthly gross margin by household

	Dependent variable:		
	Monthly gross margin by household		
	(1)	(2)	(3)
Cluster 1	1,340.820***		1,339.883***
	(8.743)		(8.688)
Cluster 2	54.128***		64.048***
	(20.103)		(19.976)
Cluster 4	189.837***		190.517***
	(3.849)		(3.835)
Cluster 5	904.080***		902.506***
	(4.416)		(4.404)
Cluster 6	676.460***		665.586***
	(7.924)		(7.881)
Cluster 7	461.383***		460.904***
	(3.824)		(3.820)
Cluster 8	1,063.940***		1,063.480***
	(5.581)		(5.550)
Cluster 9	661.423***		658.892***
	(3.965)		(3.957)
Cluster 10	119.188***		123.677***
	(10.314)		(10.250)
NecessitiesFnB		−0.379***	−1.796***
		(0.137)	(0.110)
NecessitiesNonFnB		0.694	1.803**
		(0.885)	(0.703)
TacoFriday		15.388***	12.276***
		(0.404)	(0.321)
ChipsnDip		−1.057***	−1.492***
		(0.357)	(0.284)
KidsFoods		10.429***	6.772***
		(0.541)	(0.429)
Constant	178.358***	613.390***	185.398***
	(3.070)	(4.655)	(4.497)
Observations	158,179	158,179	158,179
R^2	0.369	0.011	0.377
Note:			*p<0.1; **p<0.05; ***p<0.01

Mexican sauce, is expected to influence the purchase decision of other products in the *TacoFriday* set.

With the *Clusters* and *ItemSets* set up, Table 3 measures their impact on *Margin*. Overall, the fact that $R^2 = 0.369$ of model (1) is much higher than $R^2 = 0.011$ of model (2) implies the $Clusters_k$ can better explain the variation of $Margin_{ki}$ as regard to the $ItemSets_{ki}$. Besides, when both of them are taken into model (3), the coefficients of the *Clusters* are significantly higher than the coefficients of the *ItemSets*, indicating that the best customer segments are more impactful than the best product categories. The models also show that Cluster 1, as expected, is potentially the most profitable segment, followed by Cluster 8, 5 and 9. Notice the dependent variable here is the monthly *Margin* by household, which would not reflect the cluster sizes and their contribution to margin at the store-level. Despite the low R^2, Model (2) reveals that *TacoFriday* and *KidsFoods* products are the ones that have the highest coefficients. This may come from the fact that households who buy *TacoFriday* and *KidsFoods* may have larger household size, hence, bigger grocery budgets.

$Margin_{ki}$ with interaction between $Clusters_k$ and $ItemSets_{ki}$ has been explored. Its explanatory power does not increase significantly compared with model 3, hence it is not included in this paper.

5. Concluding remarks

The results suggest that the customers holding the baskets potentially impact more on retailer profitability than the products in the baskets. Besides investing in CRM and the best-segments, it is equally essential for retailers to learn the underlying rationale behind them. In this case, frequency is seen as the key for customer profitability, meaning it might be more profitable to have customers come more often and buy small baskets (e.g. Cluster 1) than less frequent customers buying big baskets (e.g. Cluster 6). As discussed in the literature, the more often customers visit a specific store, the more they are exposed to the marketing instruments, which potentially leads to additional purchases (Kumar et al., 2008). Product categories are found not to explain the monthly *Margin* as well as the customers do, yet the *ItemSets* can still be leveraged for cross-selling and up-selling strategies. Besides, though the findings are typical for the market of study, the method of using MBA can potentially be replicated to identify the typical patterns in other markets.

Another observation is that customers who buy *KidsFoods* and *TacoFriday* products tentatively have higher monthly margin comparing to those who do not. Retailers can leverage such categories as proxies for households with kids, household size and hence their budget.
To overcome the limitations of RFM and static linear regression, for future research, profitability can be explored through CLV lens and panel analyses that promise to better capture the dynamics in customer relationship and hence enhance the predictability.

6. References

Bell, D., Corsten, D., & Knox, G. (2011). From point of purchase to path to purchase: How preshopping factors drive unplanned buying. *Journal of Marketing*, *75*(1), 31-45.

Boatwright, P., & Nunes, J. C. (2001). Reducing assortment: An attribute-based approach. *Journal of Marketing*, *65*(3), 50.

Dippold, K., & Hruschka, H. (2013). Variable selection for market basket analysis. *Computational Statistics*, *28*(2), 519–539.

Duboff, R. S. (1992). Marketing to Maximize Profitability. *Journal of Business Strategy*, *13*(6), 10–13.

ECR Europe. (2011). *The consumer and shopper journey framework.* Retrieved 2019-10-31, from https://bit.ly/2OJsVsC

Fader, P., Hardie, B., & Lee, K. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research American Marketing Association*, *XLII*, 415-430.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems With Applications*, *100*, 1–16.

Hughes, A. M. (1996). Boosting response with RFM. *Marketing Tools*, *3*(3), 4.

Kumar, V., Anand, A., & Song, H. (2017). Future of retailer profitability: An organizing framework. *Journal of Retailing*, *93*(1), 96 - 119.

Kumar, V., George, M., & Pancras, J. (2008). Cross-buying in retailing: Drivers and consequences. *Journal of Retailing*, *84*(1), 15 - 27.

Lash, E. (2019). *Wide-moat P&G's upward sales trajectory a plus, but competitive and macro pressures abound.* Retrieved 2019-10-31, from https://bit.ly/380C6N9

Manchanda, P., Ansari, A., & Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science*, *18*(2), 95–114.

Moodley, R., Chiclana, F., Caraffini, F., & Carter, J. (2019). A product-centric data mining algorithm for targeted promotions. *Journal of Retailing and Consumer Services*.

Petersen, J. A., McAlister, L., Reibstein, D. J., Winer, R. S., Kumar, V., & Atkinson, G. (2009). Choosing the right metrics to maximize profitability and shareholder value. *Journal of Retailing*, *85*(1), 95 - 111.

Sarantopoulos, P., Theotokis, A., Pramatari, K., & Doukidis, G. (2016). Shopping missions: An analytical method for the identification of shopper need states. *Journal of Business Research*, *69*(3), 1043–1052.

Schrage, M. (2014). Lafley's P&G Brand Cull and the 80/20 rule. *Harvard Business Review Digital Articles*, 2–4.

Sloot, L. M., Fok, D., & Verhoef, P. C. (2006). The short- and long-term impact of an assortment reduction on category sales. *Journal of Marketing Research*, *43*(4), 536–548.

Tan, P.-N., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 32–41). New York, NY, USA: ACM.

Yang, A. X. (2004). How to develop new approaches to RFM segmentation. *Journal of Targeting, Measurement and Analysis for Marketing*, *13*(1), 50–60.