An Analysis and Exploration on the Effectiveness of Marketing Mix Activities

Yuping Qi

Breno Albuquerque

Jacqueline Huang

Date: 10/30/2019

Executive Summary

This report is a study on the marketing activity data of a cosmetic company, and we aim to examine the marketing effectiveness and propose an allocation method. We were provided with a dataset and designated to use a focal model to show the relationship between different marketing channels and sales.

In order to construct the focal model, first we needed to select variables that provided sufficient information and showed robust significance. We used backward stepwise regression to drop variables and meanwhile monitored the quality of the model with the existing variables. The final focal model was comprised of “Catalogs_Winback”, “Catalogs_Newcust”, “Portals”, and lagged variable, and it showed that the investment on catalogues for winning back customers and portal advertisements generated most sales.

Nevertheless, the focal model was far from ideal according to the R-square record. To increase the model fit, we could incorporate different diminishing effect mechanism, timing effects, and synergy effects. We explored these elements in the Extension part of the report in detail and came up with a different model. The new model displayed a higher goodness of fit and recommends a different combination of advertisement methods, which emphasizes on “Search”, “Newsletter”, “Catalogs_Winback”, and “Catalogs_Newcust” now.

Introduction

A German cosmetics firm launched a product 4 years ago and needed to determine the effectiveness of its advertising tools. It provided data on advertising spends on multiple platforms and the final sale performance. This report will analyze the effectiveness of advertising on sales using marketing mix model approach. Additionally, there are preliminary recommendations for management in the later section.

Problem Formulation

The company has multifold advertising methods but struggles to see how each method impacts on the sale. Specifically, it has problems in determining whether one advertising approach is effective, how much one method contributes to the sale relative to each other, and how to allocate the limited budget across its advertising channels in an optimal way

By looking at the dataset provided, we can see that the company has difficulty updating its database. For instance, the firm has not invested in propagating through the social media for a while but this method still lingers in the company dataset. Another example would be the striking number of missing and repeated values. It's clear that we can clean up this trove to benefit the data analysis and further assist in the decision-making of the firm.

Data Description

The data has three components: different advertising spends, the monthly sales, and the time variable month. The advertising spending is composed of online and offline parts. The online advertising includes search, social media, newsletter, retargeting, portal, and banner. The offline is sub-categorized into mailing and catalogue for new, existing and win-back customers. These categories of data are overlapping and influencing each other. A study of these relationships can either serve as a variable selection tool or reveal deeper insights.

We started selecting variables by examining the multicollinearity relationship. The study reveals that “ADV_offline” “ADV_Total” & “ADV_online” increase the variance of the regression coefficients [1], which translates into the redundancy of the information. A deeper look into these three variables shows that the three are the sum of other variable amounts. Since we want to analyze the effect of individual, we dropped these three variables as a result. Next we created a heatmap of correlation matrix to further study the relationship between variables. According to the heatmap, most offline methods are closely correlated with each other and the same goes for the online methods. We used this finding of correlation for future reference.

Then we conducted other exploration analysis. The summary data (Table a) indicates that “SocialMedia” and “Banner” are missing 75% of the values, so we dropped them as they become less significant under this fact. We have also observed some abnormalities in the data. The “Catalogs_ExsitCus” goes extremely high right before being set to zero every 12 month (Chart 2), and “Portals” contains repeated values for the first 12 months. Therefore, we suspect there could be some mishandling of the data collection.

Model Development

After dropping variables with multicollinearity fallacy and significant missing values, we adopted a backward stepwise regression approach [2] to keep filtering the variables. As our focal model was predetermined to use square root, we first tested the AIC score of the square-root model with all the variables left. Next we used R to apply the lm() function to sale and all variables. Each variable would have a P value and the one with the highest P value got dropped. Then we tested the AIC again with the new set of the variables left. The process iterated until the AIC score ceased dropping. This regression approach eliminated “Mailing”, “Search”, “retargeting”, “newsletter”, and “Catalogs_ExistCust” and left “Catalogs_Winback”, “Catalogs_Newcust”, and “Portals” for our final dynamic focal model:

$$Y_t = 0.271Y_{t-1} + 55.225 * \sqrt{Catalogs\_Winback_t} - 26.545 * \sqrt{Catalogs\_NewCust_t} + 763.107 * \sqrt{Portals_t} + 1804.544 + \epsilon_t$$

Where Yt stands for sales, Yt-1 is lagged sales, et means error.

During the model development process, we made several observations. First, the lagged sales is not significant during the first few iterations in the backward stepwise regression, but becomes so in the final focal model. We did not drop the lagged sales before because the focal model should contain not only the activities generating short-term sales(leading factors) but also those generating long-term sales (lagging factors) [2].

Secondly the coefficient of “Catalogs_NewCust” is negative. Even though this means that the Catalogues sent to new customers have a negative effect on the sales statistically, this observation cannot serve as a conclusion due to the limitation of the dataset as mentioned in the data description section. A likely explanation for the negative coefficient could be that it takes time for the catalogues to arrive at the customers and for the customers to actually execute the purchase and finally impact on the sales report of the company. Therefore the current sales is not positively correlated with the concurrent advertising spending on “Catalogs_NewCust”. A possible answer to this version of explanation would be to extend our analysis to consider the long-term and cross-channel effect of the current advertising activities. In this way, the coefficient of “Catalogs_NewCust” would be redressed to reflect its real impact on sales. Finally, we observed that the seasonality played an evident role in the correlation between ad activities and (lagged) sales， which we would further explore in the extension part of this report.

Results

Model fit. The R-square of the final dynamic model is 0.33 and the adjusted R is 0.26 (Table c), which are not distinctively high. However, as this model considers only the marketing activities, the R-square could increase if we factor in other influences on sale, such as the product type, the customer income level, and so on. Moreover, as all the independent variables in our model are significant enough according to our backward stepwise regression, they should be able to partially explain the predicted value, a sign of a good model fit. All the aforementioned factors gravitate towards the conclusion that our focal model has a reasonable goodness of fit despite the medium R-square score.

Sales Dynamics The lagged sales is sometimes referred to as the carryover effect and it is the main representation of the sales dynamics in our model. The lagged sales variable is not significant enough during the earlier rounds of backward stepwise regression[1] while only becomes so in the final focal model. This implies a lack of robustness for including the carryover effect in our model. In spite of that, we did not exclude this variable as the carryover effect is an important indicator in the marketing activities [5]. In the Extension, we will explore more models that might increase the robustness of sales dynamics or use other factors to take place of the carryover effect.

Marketing effectiveness. The variable selection process revealed that catalogues to win back customers and investment on portal advertisements has the strongest positive effect on the sales. The spending on catalogues sent to new customers has a statistically negatively effect on total sales, a finding that calls for further discussion. Furthermore, most current advertising spends listed on the company dataset have no significant effect on the final sales. This summary

[1] Referred to in the "Model Development" part

on marketing effectiveness is not conclusive before considering other factors such as lagged effects and seasonal effect [6].

Elasticity estimation. With the unstandardized coefficient of our focal model, we derived the elasticity of the variables “Catalogs_Winback”, “Catalogs_Newcust”, and “Portals” to be 0.072, -0.063, and 0.249 respectively. The elasticity of the portal channel is 2 times more than that of the catalogues for winning back. This is a reflection of the current popularity of online marketing strategy [7]. The negative elasticity of catalogue spendings on new customers needs further verification before removing it from the company’s advertising channels. This information has serious managerial implications by shedding light on the marketing expense distribution.

Recommendations and Managerial Implications

Even though our elasticity estimation suggests the company should focus on investing in the catalogues for new customers and portal advertisement, as data analysts, we have always been skeptical about the given data and the designated model as the status quo. In this case, we have further experimented and discovered a better model (details in Appendix 2). Instead of exploiting only portals and two catalogues proportionally to the elasticity score in the focal model, the company could focus more on advertising through "Search", "Newsletter", "Catalogs_Winback", and "Catalogs_Newcust". These channels made up a more diverse combination of online and offline advertisements than the result produced by the focal model.

In either model, there is no causality indicated in this study until we prove it with the use of a counterfactual approach. For more precise results, there are other data, such as monthly inflation, gross domestic, or even cosmetic industry-specific indexes, to narrow down the influence of media investment.

Reference

1. Multicollinearity in regression. (n.d.). Retrieved from https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/model-assumptions/multicollinearity-in-regression/.
2. https://leveleleven.com/2016/10/sales-productivity-requires-leading-lagging-sales-metrics/
3. http://ebusiness.mit.edu/research/papers/222_Brynjolfsson_Advertising_Future_Sales.pdf
4. Multiple regression analysis Efroymson, MA (1960) "." In Ralston, A. and Wilf, HS, editors, *Mathematical Methods for Digital Computers.* Wiley.
5. Assessing the Fit of Regression Models, Karen Grace-Martin, May 2005, Cornell University, Cornell Statistical Consulting Unit, StatNews #68 (https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
6. SPATIOTEMPORAL ALLOCATION OF ADVERTISING BUDGETS, Ashwin Aravindakshan, Kay Peters, Prasad A. Naik, February 1, 2012, Journal of Marketing
7. Incorporating long-term effects in determining the effectiveness of different types of online advertising, Ralph Breuer, Malte Brettel, Andreas Engelen, 16 March, 2011, Marketing Letters-A Journal of Research in Marketing
8. A Guide to Better Advertising: Online Marketing vs Tradition Offline Marketing, April 6, 2018, Lasting Trend, https://www.lastingtrend.com/blog/online-vs-offline-marketing=

Appendix (2 Parts)

Appendix Part 1: Charts

Table a : Summary

Sales (units)	Catalogs_ExistCust	Catalogs_Winback	Catalogs_NewCust	Retargeting	Newsletter
Min. :3355	Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 7.057
1st Qu.:4406	1st Qu.: 328.7	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.:16.691
Median :4690	Median : 598.0	Median : 0.00	Median : 43.63	Median : 0.00	Median :19.779
Mean :4809	Mean : 567.6	Mean : 83.42	Mean : 272.87	Mean :10.85	Mean :20.734
3rd Qu.:5195	3rd Qu.: 625.6	3rd Qu.:174.15	3rd Qu.: 487.42	3rd Qu.:18.56	3rd Qu.:25.139
Max. :6976	Max. :1298.7	Max. :438.54	Max. :1131.57	Max. :49.30	Max. :53.609
Mailings	Banner	Search	SocialMedia	Portals	
Min. : 0.00	Min. : 0.000	Min. : 38.17	Min. :0	Min. :2.544	
1st Qu.: 0.00	1st Qu.: 0.000	1st Qu.: 45.38	1st Qu.:0	1st Qu.:3.393	
Median : 0.00	Median : 0.000	Median : 66.11	Median :0	Median :4.707	
Mean :11.42	Mean : 5.179	Mean : 69.83	Mean :0	Mean :5.246	
3rd Qu.:19.24	3rd Qu.: 0.000	3rd Qu.: 88.19	3rd Qu.:0	3rd Qu.:6.867	
Max. :84.47	Max. :87.611	Max. :134.87	Max. :0	Max. :9.303	

Table b: the elasticity scores

```
> LR_elastic
                    Winback Catalog_ExistCust   Portals Newsletter
Catalog_Winback_log 0.04443779         0.0245116 0.1232468 0.09181006
>
```

Table c: the output table

```
Call:
lm(formula = sales ~ ., data = hw2_sqrt_lag)

Residuals:
     Min       1Q   Median       3Q      Max 
-1130.68  -569.93    43.59   332.50  1657.15 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)  
(Intercept)      1804.5441   798.1380   2.261   0.0299 *
Catalogs_Winback   55.2247    24.4795   2.256   0.0302 *
Catalogs_NewCust  -26.5445    13.8609  -1.915   0.0635 .
Portals           763.1066   284.4544   2.683   0.0110 *
salestm1            0.2711     0.1605   1.689   0.0999 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 684.7 on 36 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3339,	Adjusted R-squared:  0.2599 
F-statistic: 4.512 on 4 and 36 DF,  p-value: 0.004666
```

Chart 1:

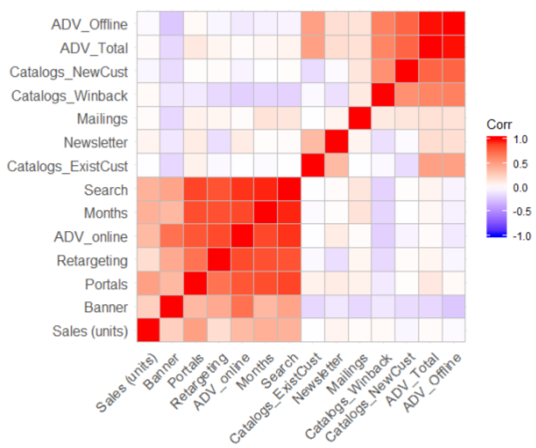
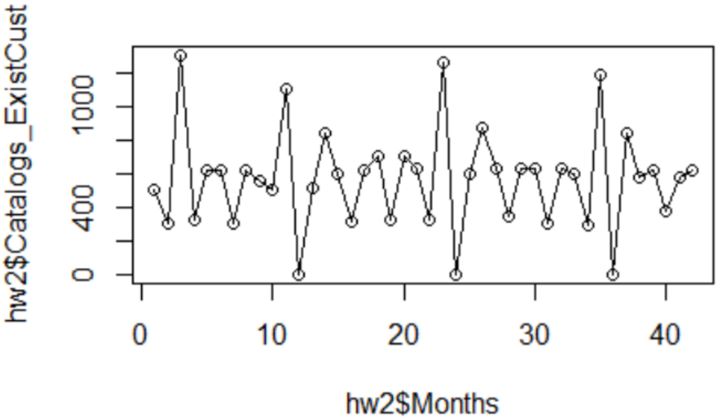


Chart 2



Appendix 2 : Model Extension

For the cosmetic sales model extension our team dedicated to asses the presence of several marketing industry expected effects. We departed from focus model approach: avoiding including covariates that would suffer from micronumerosity or cause multicollinearity. We kept the number of independent variables constrained to the recommended threshold. The most important tested specifications are shown in Tables 1 and 2. Our main findings are summarized by effect.

Seasonality

Sales exhibits a strong seasonal behavior[2]. Coefficient is always statistically significant and robust to different specifications. This means that cosmetics tend to sell more during certain periods of the year. Units sold during the current month is correlated to units sold twelve months ago. According to our extended model, this pattern explains a great amount of sales variation in the sample.

[2] Nevertheless, we found no statistical evidence the dependent variable follows a non-stationary process.

Intercept and Time Trend

Does Sales follow a (deterministic) time trend? Our analysis concluded that, despite any investments in advertising, units sold tend to grow in time, on a decreasing rate. This result may be due to earned media, product organic acceptance rate or market and economic growth during the period. We didn't control for macroeconomic or industry specific variables. Jointly with baseline revenue (intercept was always statistically significant), this effect accounts for the expected sales in the absence of any advertising expenditure.

Carryover effect (Adstock)

When controlled for seasonal effect, sales demonstrated few or no signal of carryover effects (Table 1, Column 3). It seems that after discounting for time trend and seasonality, sales of cosmetics from previous months can't explain what happens to current month sales. Thus, we decided to drop carryover term from our final model.

Diminishing returns (Adstock)

The extended model confirmed for diminishing return (saturation) effect on advertising expenditure. Log specification (for all covariates) better fits Sales when compared to the alternative - see Table 2, columns 4 and 5. We didn't try for other diminishing returns specifications.

Lagged effect

We tested if any advertising activity would take longer than a month to impact sales. For instance, we expected that investments in catalogue advertising would require more time than online advertising before exerting influence on sales. Since catalogues demands a long time to deliver, its returns may not be as immediate as of other media. Our team found that catalogue advertising actually takes more than a month in order to cause most of its impact (Table 2, Columns 6 and 7). After accounting for the proper lag, catalogue expenditure coefficient turned to positive (the expected sign).

Synergy effect

We tested for synergy in media and found robust results to interaction on Catalogue and Newsletter (Table 2, column 7). Coefficients are positive, suggesting that newsletter plays a complementary role to catalogues. The coefficients magnitudes and statistical significances were stable to several different specifications. In the final model, synergy implies that media expenditure impacts are no longer independent and crossing effects must be considered on budget distribution across media. As a consequence, defining allocation strategies based on equating media marginal impacts on sales are no longer trivial and requires more complex computations. The reason is that catalogues budget marginal effectiveness can now be expanded by just increasing newsletter expenditures. The firm should incorporate the effects of synergy in their allocation model since it would definitely improve resources allocation.

Final comments

Our final model outperforms the Focal model as it has a much lower AIC score (632.6) compared to Focal's (658.4). It is represented in column (7) of Table 2.It also managed to explain some of the negative coefficients observed when we first run the focal model with all the variables - the catalogue variables showed positive impacts when we accounted for proper lag. Beyond the estimated models presented in this Appendix, we have tested for other 67 specifications. Our team dropped variables that were not statically significant at 10%, after accounting for seasonal effects, diminishing effects and log form. As can be seen in Tables 1 and 2, in general estimated coefficients are robust (kept sign and level and showed only expected variations) to other variables inclusion or exclusion.

Not every media should be exploited when deciding where to allocate budget. Specifically, advertising on search, newsletter and catalogues to current customers and win backs should be taken into consideration.

There are other variables that are easy to collect and could have been used to increase model adherence to Sales and avoid omitted variable bias. For instance, it is very likely that unit sales are dependent on economic activity. Adding country macroeconomic data (in this case German), like monthly inflation or GD or even cosmetic industry-specific indexes would benefit the model.

It is important to remind that findings here doesn't represent causality. In order to do so it would require the use of a counterfactual approach. For instance, instrumental variables could have been tried if we have had access to good instruments.

Estimated equation

$$\begin{aligned}Sales_t = {} & 1879.6 + 264.3 \cdot \log(Month_t) + 592.7 \cdot log(Portals_t) + 80.1 \cdot log(Catalog\_Winback_t) + \\ & + 44.1 \cdot log(Catalog\_Winback_{t-1}) \cdot log(Newsletter_{t-1}) + \\ & + 38.9 \cdot log(Catalog\_ExistCust_{t-1}) \cdot log(Newsletter_{t-1}) + \epsilon_t\end{aligned}$$

where $\epsilon_t = 0.53 \cdot \epsilon_{t-12} + \mu_t$

Extended model Long-Run Elasticities:

```
> LR_elastic
                        Winback Catalog_ExistCust    Portals Newsletter
Catalog_Winback_log 0.04443779         0.0245116 0.1232468 0.09181006
> 
```

Table 1: Seasonality, Carryover, Intercept and Time Trend

	Dependent variable: $Sales_t$		
	(1)	(2)	(3)
ar1			−0.013
			(0.167)
sar1	0.724***	0.720***	0.722***
	(0.101)	(0.102)	(0.107)
intercept	4,769.804***	4,303.540***	4,304.316***
	(200.402)	(243.189)	(240.980)
$Month_t$		21.706***	21.685***
		(7.555)	(7.466)
Observations	42	42	42
Log Likelihood	−330.263	−326.465	−326.462
σ^2	320,106.300	268,247.400	267,599.100
Akaike Inf. Crit.	666.525	660.930	662.925
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2: Diminishing returns, Lagged effect and Synergy

	Dependent variable: $Sales_t$			
	(4)	(5)	(6)	(7)
sar1	0.779***	0.695***	0.554***	0.527***
	(0.102)	(0.120)	(0.154)	(0.161)
intercept	3,726.875***	4,025.695***	1,937.634***	1,879.572***
	(592.526)	(682.974)	(630.715)	(510.690)
$\sqrt{Month_t}$	127.300*			
	(72.872)			
$\sqrt{Portals_t}$	341.403			
	(232.122)			
$\sqrt{Catalog\_ExistCust_t}$	−4.476			
	(18.820)			
$\sqrt{Catalog\_Winback_t}$	−16.003			
	(12.728)			
$log(Month_t)$		258.913**	283.111*	264.359*
		(125.316)	(145.933)	(136.925)
$log(Portals_t)$		543.164*	634.561**	592.725**
		(288.646)	(305.596)	(290.005)
$log(Catalog\_ExistCust_t)$		−152.340*		
		(88.932)		
$log(Catalog\_Winback_t)$		−21.435	93.741**	80.098*
		(39.198)	(47.609)	(43.366)
$log(Catalog\_Winback_{t-1})$			125.663**	
			(48.833)	
$log(Catalog\_ExistCust_{t-1})$			81.451	
			(72.107)	
$log(Catalog\_Winback_{t-1}) * log(Newsletter_{t-1})$				44.071***
				(15.100)
$log(Catalog\_ExistCust_{t-1}) * log(Newsletter_{t-1})$				38.882**
				(17.562)
Observations	42	42	41	41
Log Likelihood	−323.090	−321.196	−311.139	−308.307
σ^2	215,468.000	212,859.700	205,280.500	180,978.400
Akaike Inf. Crit.	660.180	656.393	638.278	632.614
Note:				*p<0.1; **p<0.05; ***p<0.01