Modelling Report

Camilo Bonilla

Mar 26, 2023

This document presents the results of a series of analyses that were carried out on a data set containing information about sales, marketing expenditures and several other variables related to the performance of a consumer electronics store between 2015 and 2016. The goal was to explore the relationship between the amount of money spent on marketing campaigns and the revenue from sales. Accompanying scripts can be found at: https://github.com/cabonillah/GroupM_Test.

1 Data extraction and transformation

i SQL Query

All the relevant information was extracted using the SQLite standard. Given that some of the raw tables contained redundant information, the set of tables that were used was comprised of "Sales", "Secondfile" and "SpecialSale". The final result included variables from the above-mentioned tables and it was aggregated by product category and month.

ii Final adjustments in R

Due to problems associated to later stages in the modelling process, it was determined that further transformations were necessary:

- Data were further aggregated on a quarterly (rather than monthly) basis to avoid problems related to dimensionality and rank deficiency.
- "online_mkt" and "sem" variables were omitted to avoid multicollinearity and because it was impossible to distinguish them from the "digital" variable based on the available information.
- "content_mkt" variable was omitted because content marketing is not related to a particular medium. Rather, it is an approach to marketing that might or might not include mass media.

2 Data exploration

A few plots were generated to check the hypothesis that expenditure on advertising cause an increase in sales. In fact, a relationship can be established between these two variables: Figure 2.1 shows that revenue increases with expenditure on advertising, although this relationship has diminishing returns. Figure 2.2 partially confirms this fact for each quarter and each product category combination. Finally, Figure 2.3 shows that the lowest levels of expenditure are more densely grouped toward the lower tail of the revenue distribution.

Figure 2.1: Revenue vs. Expenditure Level

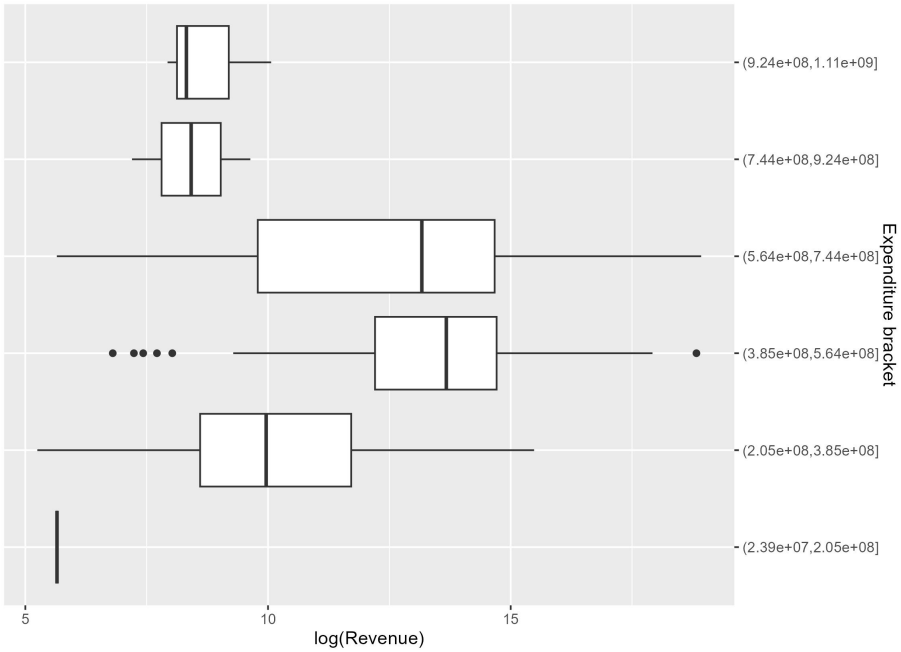


Figure 2.2: Revenue vs. Expenditure by Quarter

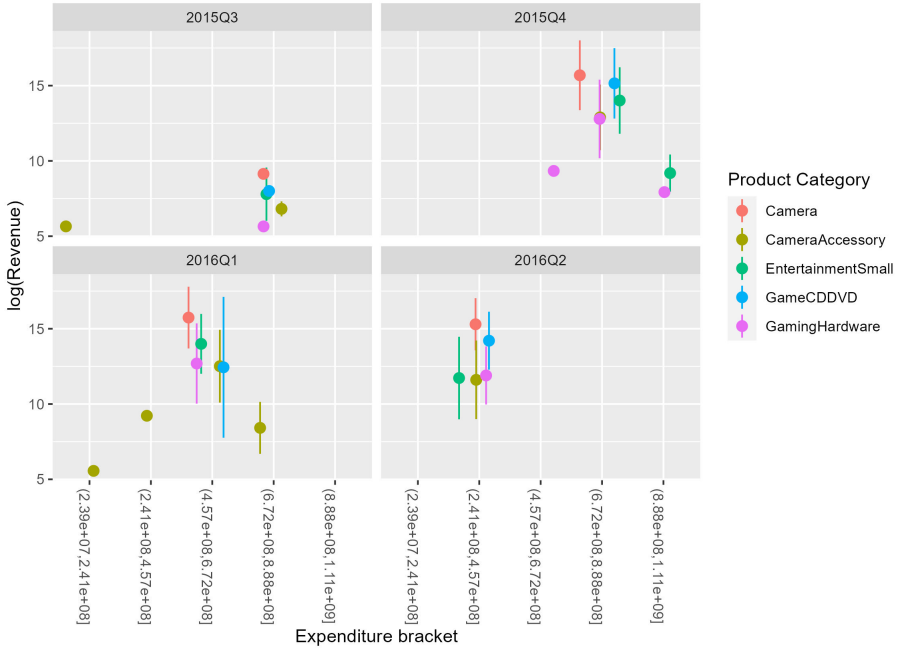
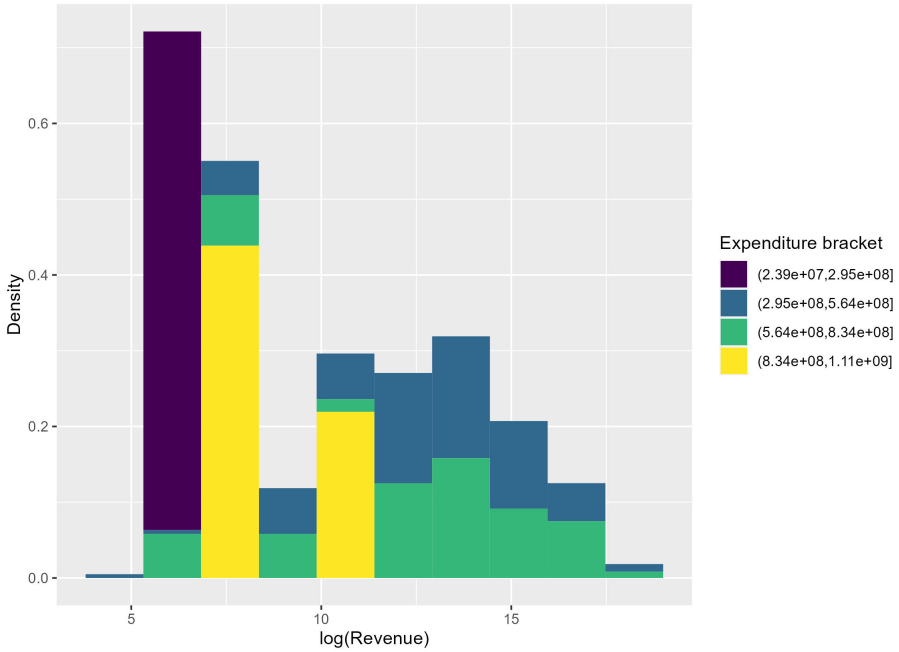


Figure 2.3: Distribution of Revenue by Level of Expenditure



i Exploring the relationship with linear models

To study the relationship between the covariates and the response variable, two models were fit: a two-way fixed effects model and a mixed efects model. The idea behind the first model is to compute different intercepts for each product type and each time period in order to generate less biased estimators. On the other hand, the mixed effects models accounts for disturbances around the parameters of interest, so that variances of the estimators are better approximated.

The following tables show the results of these two estimations. Table 1 and Table 2 show the 95% confidence intervals for the estimates. In the case of the mixed effects models, all confidence intervals were computed using bootstraping.

Table 1: Two-Way Fixed Effects Model

	2.5 %	97.5 %
log(tv + 0.5)	0.013	8.291
log(digital + 0.5)	-0.888	4.207
log(sponsorship + 0.5)	-3.984	2.032
log(affiliates + 0.5)	-35.174	-3.154
log(radio + 0.5)	-1.410	14.512
log(other + 0.5)	-12.669	1.297
sales_days	0.279	0.534
avg_sla	-0.433	-0.120
avg_proc_sla	0.035	0.207
nps	-1.992	1.296

Table 2: Mixed Effects Model

	2.5 %	97.5 %
.sig01	0.239	2.209
.sig02	0	3.994
.sigma	1.579	1.933
(Intercept)	22.244	479.600
log(tv + 0.5)	1.685	8.499
log(digital + 0.5)	0.472	4.489
log(sponsorship + 0.5)	-4.003	0.943
log(affiliates + 0.5)	-33.963	-5.862
log(radio + 0.5)	4.942	14.341
log(other + 0.5)	-12.521	-4.249
sales_days	0.276	0.473
avg_sla	-0.405	-0.103
avg_proc_sla	0.023	0.180
nps	-0.071	1.127

The estimations tend to be close to each other, and tend to be closer towards the upper bounds of the confidence intervals. This has to do with the way each model accounts for cluster-specific variation. However, both models provide evidence that there is a statistically significant relationship between some categories of expenditure and revenue.

ii Exploring the relationship with a non-linear model

In order to gain more perspective on the relationship of interest, a machine learning model was calculated with the data, using an XGBoost algorithm.

The hyperparameters of the model were found using 5-fold cross validation and a bayesian search algorithm across the search space. When the best model (MAE=0.406) was computed, it was examined to find its most influential variables. They are shown in Table 3. The results indicate a modest importance of the expenditure variables in predicting the response variable.

Table 3: XGBoost model

Variable	Importance
avg_sla	0.3789
avg_proc_sla	0.3274
sales_days	0.1172
radio	0.0682
sponsorship	0.0642
digital	0.0244
nps	0.0084
affiliates	0.0040
other	0.0037
tv	0.0032

iii Conclusion

Three models were presented in this report in order to better understand the relationship between the level of expenditure on advertising and the amount of revenue from sales.

Although some evidence was found to support this relationship (e.g. spending on TV and radio), it was not always strong and it was not consistently robust across different specifications. The use of a non-linear model at the end confirmed the idea that there are other important factors affecting sales and that spending on advertisement does not have a strong predictive power.

Perhaps the almost nonexistent variability in expenditures is a reason behind its lacking predictive power. Nonetheless, most of the other variables in the data set have this same issue, so this points to a need of more data coming from a larger time window (the current data set barely spans a year). This would provide the company with the necessary information to overcome the existing problem associated to rank deficiency in the data.