MKT 680

Project Report

Promotion and Marketing Activities Effectiveness Analysis with Mix Modeling for Mahou San Miguel

Students:

Meng Cheng

Feifan Gu

Sidi Liu

Under supervision of

Prof. Alvin Lim   Prof. David Sackin

Emory University

Goizueta Business School

CONTENTS

1. Introduction

Mahou San Miguel is a Spanish brewer that sells three San Miguel beer products in Pernalonga’s stores and regularly partners with Pernalonga to promote its products via weekly marketing activities. Mahou San Miguel wants to verify the effectiveness of its promotions and marketing partnership with Pernalonga and identify the key elements that drive incremental sales for continuing into 2020. Being familiar with Pernalonga’s transaction and products, our data science team has the perfect timing to demonstrate our capabilities in marketing analytics and help our client to solve this problem.

In this project, we will mainly leverage Marketing Mix Modeling (MMM) for the analysis. It is a data analytics technique that helps quantify the impact of several marketing vehicles on revenue or sales volume. Within this project, we have seven types of vehicles through which promotion could take place: paid search, radio, TV, web display, Email, flyer, and store display. We will decompose the weekly sales volumes for each San Miguel product into base(including seasonality and holidays), shelf price, discount and the seven vehicles, and justify the validity of the models that lead to the decomposition with various aspects of model diagnostic.

2. Data Understanding and Preparation

We have several datasets in this project, including transaction records, promotion history, seasonality index and holiday info of the three San Miguel beer products, which we identify with their product ids: 138936951, 138936952 and 138936953. When we discuss products in the following sections, we always refer to these ones.

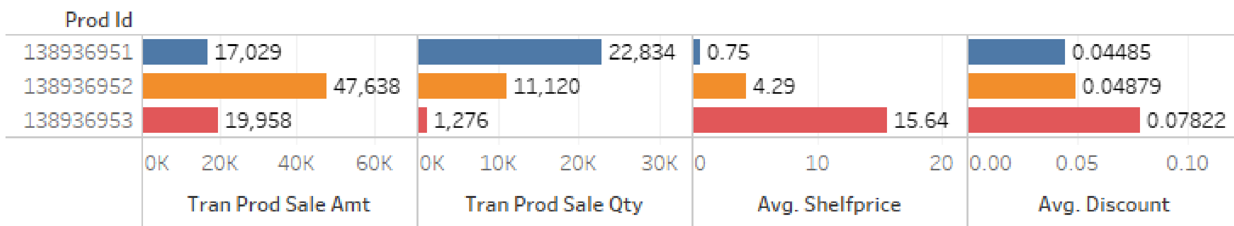

Fig1: the sum of sales amount and quantity and average shelf price and discount of the three products

As you can see from the three products' basic transaction features, product 138936951 has small profits but quick turnovers, while 138936953 is basically the opposite and 138936952 is in the middle of them and makes the most revenue.

With the understanding of their historical transaction, we had to do some data preparation where the first step is to aggregate the transaction data to a weekly level. We used every Sunday as the first day of a week and assigned each date a "yearweakid", where weekid is the week of the year as a decimal number (00-53). For example, 2016-06-28's yearweekid is 201626, since it is within the 26th week of year 2016. Also, with this calculation method, there are only 2 days in the first week of 2016, so we bind the data of the last week of 2015, which is also the only week of 2015 we have in the dataset, to 2016's first week. Then we calculated the sum of the volume, average shelf price, and average discount of each product each week, which will be future target variables and two of the predictor variables in the modeling part.

Fig2: promotion of the three products on seven vehicles through the two years

Promotions are also important variables in the models. As we stated in the introduction part, we have seven vehicles for promotion. As you can see from their weekly change plot, Email, paid search, TV, radio and web display are the same for the three products; store display is only for product 138936952 and 138936953; flyer is for the three of them but with a different schedule.

So, for the second step of data preparation, for each product sold in each week, we would like to calculate the strength of each type of promotion. For Email whose unit is circulation, paid search and web display whose unit are impressions, we would directly use the amount as their strength; for flyer and store display, whose amount is 1 or 0, which means exists or not, they would be directly used as factor variables; for TV and radio which we were provided with GRP as a unit, we would like to transform them into Reach, which would be a better measurement in terms of marketing influence. The specific calculation method is as follows.

1. We calculated the decay parameter of TV advertisements considering their 8-week-half-life: $\alpha_{TV} = 1 - 0.5^{\frac{1}{8}}$;
2. We converted GRPs to AdStock GRPs with this decay parameter: $AdStock\ GRPs\ in\ week_i = \alpha_{TV} \times GRPs\ in\ week_i + (1 - \alpha_{TV}) \times AdStock\ GRPs\ in\ week_{i-1}$;
3. We converted AdStock GRPs to 2+Reach: $Reach = 0.95 \times (1 - e^{-0.020\ AdStock\ GRPs})$;
4. We repeated the process to calculate the 2+Reach of radio advertisements, except that $\alpha_{radio} = 1 - 0.5^{\frac{1}{4}}$ since radio advertisements have a 4-week half-life, and use the Reach formula for radio: $Reach = 0.90 \times (1 - e^{-0.025\ AdStock\ GRPs})$.

With these preparations being done, we got all the quantified promotion variables. With holidays and important holidays including Xmas and New Year determined for each week if it contains any of them, and seasonality index joined by year-week id, we had the three weekly-level datasets prepared for mix modeling on each of the three products.

3. Model Building

After finalizing the datasets for those three different San Miguel Beer products that are sold in Pernalonga's stores, we started building linear models to identify the key sales vehicles. To be more specific, we tested both the additive model and multiplicative model, but ended up using the logit model. We left out the additive model and multiplicative model since the dependent variables in both cases are not unbounded. Unbounded means as price approach 0, demand/sales volume approaches infinity. In our case, even when the price of the product is small enough, the demand did not become unrealistically large. For instance, when the average shelf price of product 1 dropped to 0.69, the distribution of its demand/sales did not change much, and thus we pick the logit model. The logit model is also the most suitable model since it captures the implicit interaction between the causal values. In our case, an audience can be exposed to a lot of similar marketing vehicles, such as TV and radio, at the same time. Therefore, there are likely implicit interactions between those variables, and we need to take them into consideration when building the model. For the logit model, the dependent variable is represented as an inverse logistic transformation of the sum-product of the causal coefficients and the (transformed) values of the causal values. Independent variables are different marketing vehicles and components that influence product sales such as flyer, email, paid search, web search, TV, Radio, whether it is a holiday, etc.

For the logit model, the base form for the model is showed below:

$$\log\left(\frac{y_t}{1-y_t}\right) = \sum_i \beta_i g_i(x_{it}) + \epsilon_t$$

We will evaluate the performance and results of the model in the next section. Keep in mind that we will transform the dependent variable back to regular sales volume when measuring the performance.

4. Model Diagnostics

All the models were significant at p-values < 0.05 level. After running logit models for the three different products, we summarized the model performance based on their error in predicting the sales volume.

ProductID	RMSE	MAPE	DURBIN-WATSON STATISTIC
138936951	34.8380	0.2084	1.75
138936952	18.6935	0.1442	2.07
138936953	4.8962	0.4581	2.41

As we can see, the model we built for product 3 has the lowest RMSE but the highest MAPE. This is because the sales volume for product 3 is generally lower compared with the other two products. In addition, the model we built for product 2 seems to have the best overall performance as it has the lowest MAPE and relatively low RMSE among the three models.

Also, to test the autocorrelation of errors, we used the Durbin Watson Statistic Test (dwtest). It is a test for autocorrelation in the residuals from a statistical regression analysis. The resulting values of dwtest range between 0 to 4 with 2 being a strong indication of no autocorrelation. Values below 2 indicate the possibility of positive autocorrelation and values above 2 indicate possible negative autocorrelation. As shown above, all three products have a dwtest value close to 2, which indicates that there is likely no autocorrelation within the observation.

Next, we summarized the variables that are statistically significant by product.

ProductID	Positive Impact	Negative Impact
138936951	Email. TV	Important Holiday. Shelf Price
138936952	Email. Seasonality. Discount	Important Holiday. Shelf Price
138936953	Email. Seasonality. Discount	Important Holiday. Shelf Price. Store Display

As we can see, all three products' sales volume are positively affected by the same marketing vehicle – Email, and two of them are positively affected by seasonality and discount. This indicates that product 2 and product 3 are season-sensitive products and we can target our customers effectively by sending out marketing emails and providing discounts. Furthermore, factors such as the shelf price of a product and whether it is during an important holiday negatively affect these products' sales. This shows that these products are unlike holiday-related products (gifts, turkeys, etc.) and are less popular during holidays and our customers are sensitive to the shelf price of these products.

5. Sales Volume Decomposition

The goal of Marketing Mix Modeling is to decompose a measure of the marketing goal, which is the sales volume in this case, into its base and other components due to each marketing vehicle. We apply the logit model to calculate the DueTos for each marketing vehicle, including Discount, Email, Flyer, Paid Search, Radio, Shelf Price, TV, Web Display as well as the Base.

A bit of domain knowledge is needed to set base, for media and promotions including Discount, Email, Flyer, Paid Search, Radio, TV and Web Display, the base value is set to 0 in general. For Shelf Price, we take the average price of the two-year modeling time series as base. It should be noticed that holiday and seasonality are included in the base in our case as the base is usually computed as the sum of sales components due to causal factors that are not considered in decision making or cannot be affected or changed by the decision-maker. The formula for DueTos calculation is shown below.

$$\mathrm{DueTo}_{it} = \widehat{y_t} - \widehat{y_t}(x_{it} = x_{itBASE})$$

The raw DueTos results need to be adjusted by scaling for the purpose of debias, in other words, to ensure that the sum of adjusted DueTos is equal to the real sales volume. For debias, we adjust each DueTo with its proportion of total estimated weekly sales volume.

$$\sum_i \mathrm{DueTo}_{it} = y_t$$

The result of the sales volume decomposition for each product is shown below.

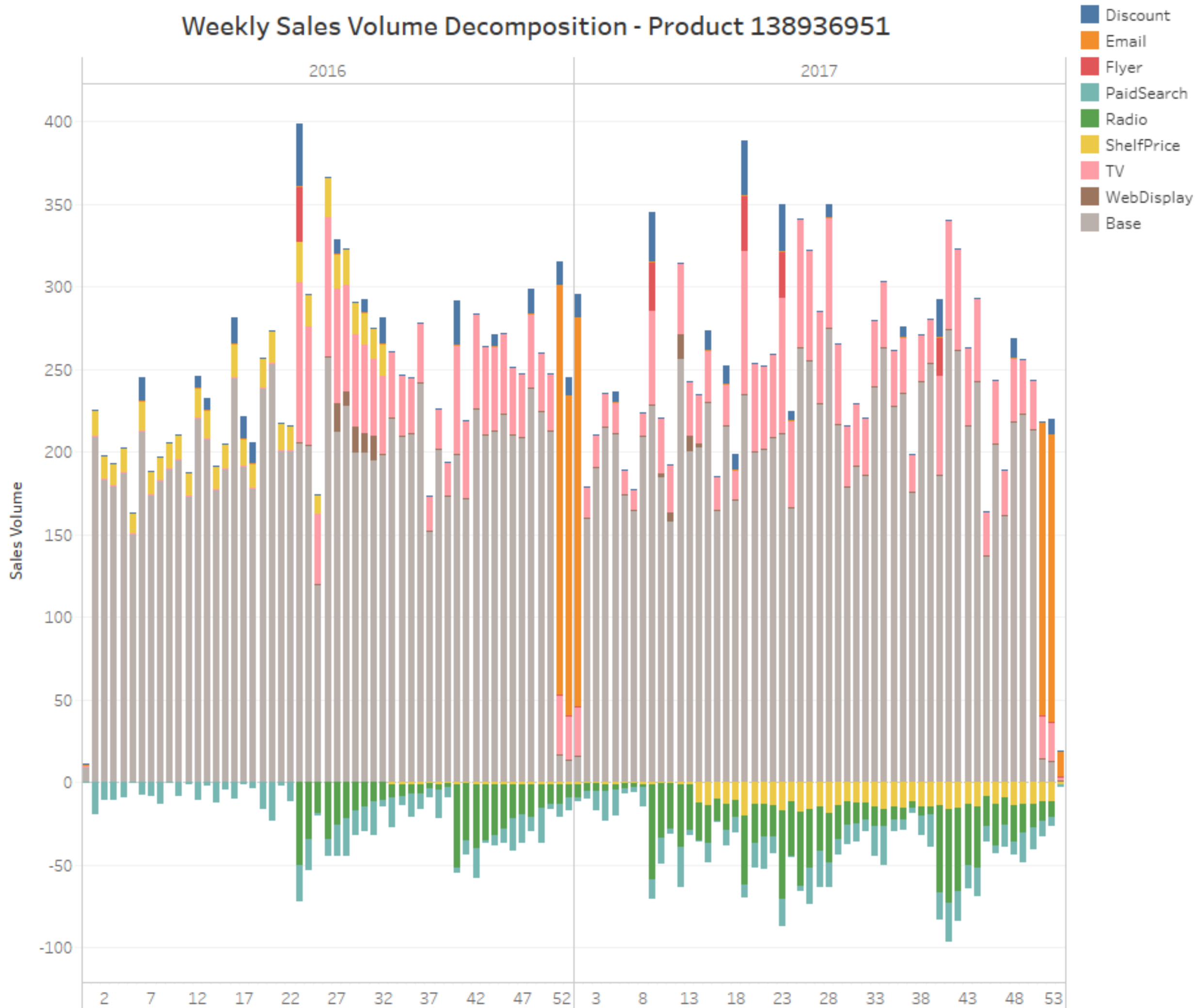

Fig: Sales Volume Decomposition Result – Product 138936961

For product 138936961, the base sales volume is relatively stable and makes up the largest proportion to the historical sales volume. Aside from base, TV contributes most while Email has a significant contribution during the time around Christmas and New Year. However, Paid Search and Radio only contribute negatively to the sales volume. Shelf Price and Discount show a relatively small impact along the two years.

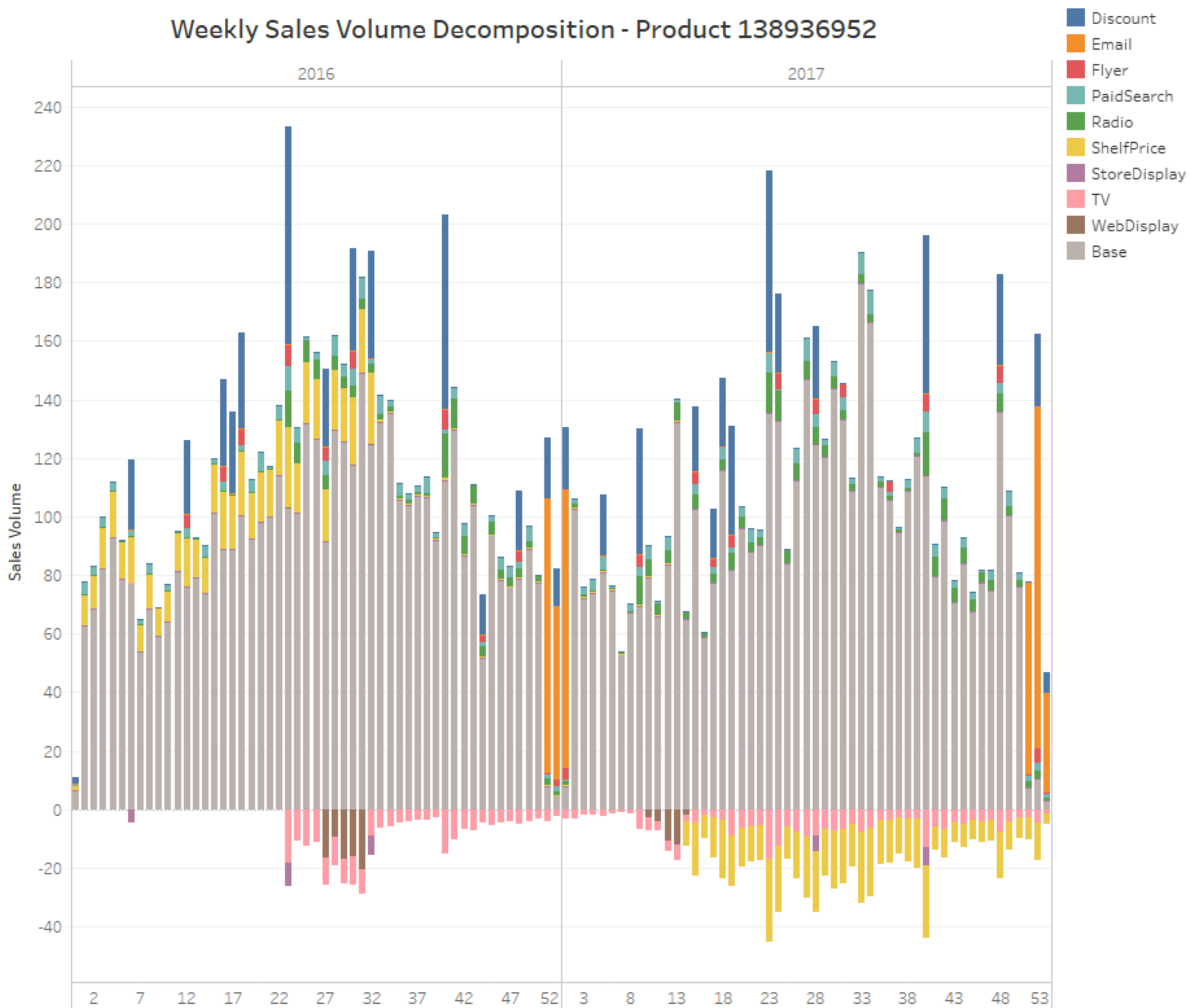

Fig: Sales Volume Decomposition Result – Product 138936962

For product 138936962, the base sales volume is smaller compared to product 138936951. Aside from base, Discount contributes the most in a positive way. Shelf Price shows an impact along the two years with a significant proportion. We can conclude that customers of product 138936962 are more price sensitive compared to customers of product 138936961. Similar to product 138936961, Email contributes a lot during the time around Christmas and New Year for product 138936962. TV, Store Display and Web Display show a negative impact on the sales volume.

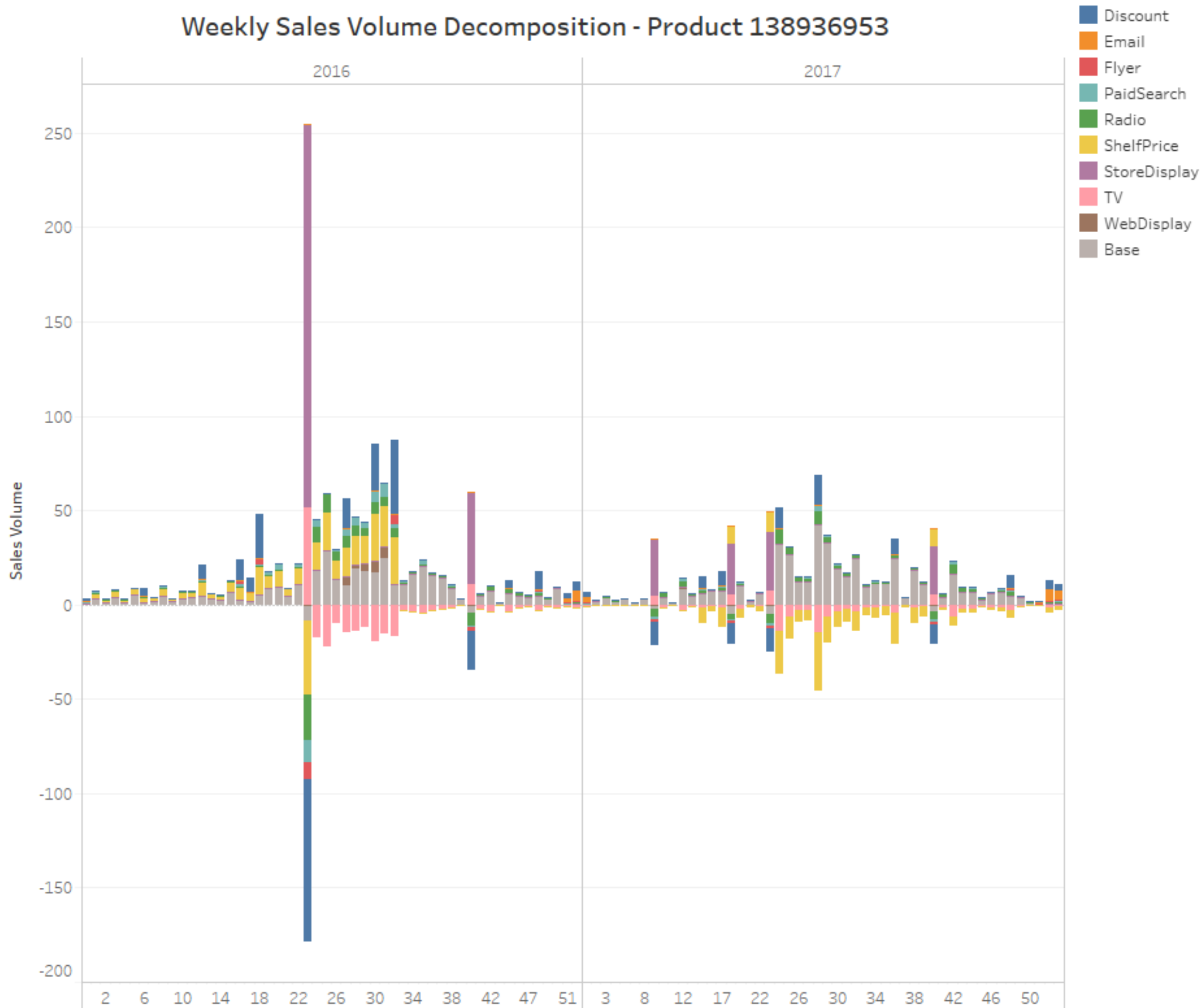

Fig: Sales Volume Decomposition Result – Product 138936963

For product 138936953, the contribution of base sales volume is relatively small, and the historical sales volume is very unstable with several peaks along the two-year period. Store Display shows the biggest positive impact on sales volume. The impact of Shelf Price and Discount are significant as well which suggests customers of product 138936963 are the most price sensitive. The contribution of TV, Radio and Flyer are mixed, and it is hard to tell whether each of the three marketing vehicles has a positive impact on the sales volume in general. Also, the historical sales volume of this product is so instable that we believe it is likely to be a seasonal product and relies heavily on promotion.

Based on the sales volume decomposition analysis above, our team recommends that San Miguel continues to utilize Discount to motivate the purchase of product 138936952 and 138936953 and use Email during the time around Christmas and New Year as it is a low cost, high impression marketing vehicle. Also, we suggest that San Miguel may reconsider the use of radio and Paid Search, especially for product 138936951, which is their strongest product and contribute the most to the total sales volume along the past two years. We also suggest that San Miguel may continue to use TV for product 138936951 but reconsider for other two products as the impact is unclear and not effective enough to justify the high cost of TV advertisement.

6. Conclusion

After completing our analysis, we were able to verify the effectiveness of its promotions and marketing partnership with Pernalonga, with the goal of identifying promotion and marketing activities that can drive significant incremental sales for continuation into 2020. Overall, different products respond differently to those marketing vehicles during the previous promotions. Moving forward, depending on the product they are trying to promote, they might want to increase and decrease spending in different marketing vehicles correspondingly to come up with the most cost-effective group of vehicles.