# A Holistic Analytics Approach for Determining Effective Promotional Product Groupings

Mehul Zawar, Siddharth Harisankar, Xuanming Hu, Rahul Raj, Vinitha Ravindran, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
mzawar@purdue.edu; sharisan@purdue.edu; hu661@purdue.edu; rajr@purdue.edu;
ravindr1@purdue.edu; lanhamm@purdue.edu

## ABSTRACT

With companies across industries continually striving to get ahead of the competition, product pricing could be the deciding factor in driving or destroying the margins of a company. Promotional grouping of products is an effective pricing strategy used across multiple industries such as retail, healthcare, and many more. Promotional product groupings or bundling can be seen everywhere, from buffets served at restaurants to the suite of products sold together by MS Office. The fact that the component products are readily available means that bundling is one of the most flexible elements of product strategy. However, some caveats come with bundling, most of which stem from inadequate planning. Bundling could lead to the cannibalization of products that are not present in bundles. Furthermore, it could lead to customers not choosing to buy the desired product because she would have to buy the other product bundled with it.

The study encapsulates the selection and creation of labels for promotional product groupings for individual SKUs of a consumer product goods company. The groupings are based on historical data of the company's incremental sales and competitors' sales data collected in the same time frame. Currently, product grouping analysis is done manually, which could be compromised by human error and an individual's unique decision framework that could be biased. A more pertinent issue faced is that the company would fail to recognize the life of a successful promotion. Failure to do so could lead to stagnant promotional groupings that would not only fail to gain traction with customers but also siphon off the already existing sales, eventually leading to the company being overtaken by its competitors and lose market share.

In order to develop recommendations for an ideal product grouping strategy, the study initially delves into the existing promotional groupings of the company and compares it with those of its competitors. Detailed competitive analysis provides an idea of the company's success with its past bundling strategy. The study uses machine learning models to identify the drivers of a successful promotion and finally uses optimization to suggest an ideal bundling strategy that would maximize revenue.

**Keywords:** promotion analysis, product grouping, bundling, optimization, binary classification

# INTRODUCTION

The pricing strategy adopted by different companies is a crucial deciding factor that impacts their revenue and margin. Selecting the right pricing tactics and applying them in the right manner can have a major impact on the bottom line. Promotional Product Grouping (or PPGs) is a technique where multiple products are grouped and sold as a single unit for one price. Grouping or bundling is a way to get customers to buy multiple products. The discounted price of such bundles attracts customers in droves. This could be advantageous to a company not only because of the increased sales but also because bundling enables consumers to discover new products that are grouped with goods that are commonly purchased. As with any strategy, improper management of bundling could hurt sales. As per a research report published by The Boston Consulting Group, 20 to 50 percent of promotions generate no noticeable lift in sales.

Another 20 to 30 percent dilute margins, where the increase in sales does not offset the cost of promotion.[1] A survey of marketing professionals from leading companies was conducted to analyze the problems faced by different companies while creating a good promotion plan. Some of the issues identified from the survey includes not having a track of their effective promotions that help build profits. Hence, the companies are hesitant to cut back the sales volume, leading them to run all the promotions year long. This not only results in monetary losses but is also driving customers toward better opportunities elsewhere. The major reason for the failure of the promotions is due to the lack of a clear goal. Companies do not consider whether the promotions are being held to improve margins, to reach out to new customers, or to improve some other metrics.

This paper looks at the promotional grouping strategy employed by a global consumer products firm (hereafter referred to as 'Company'). The research conducts a competitive analysis of similar groupings sold across the industry. This analysis could help identify the areas where the Company is finding success with its strategy and where it is surpassed by its competitors. Several parameters, such as total sales and quantity sold, are taken to generate a score, which is then compared across the various competitors to identify which grouping had the highest' success'. Building on this analysis, the study uses machine learning models to estimate the importance of various drivers behind a successful promotional bundle. The model created considers a systematic analysis of historical performance to determine whether specific promotions are meeting the Company's strategic objectives. It learns from past PPG groupings, sales patterns, and the PPGs of the competitors and will then recommend what PPG label to place at each SKU level. For example, a particular flavor of a shampoo grouped with other personal care products could prove to have traction with customers. This could be successfully adopted by the Company to boost sales. Finally, the study develops an optimization model to generate promotional groupings that can be adopted by the Company to maximize revenue.

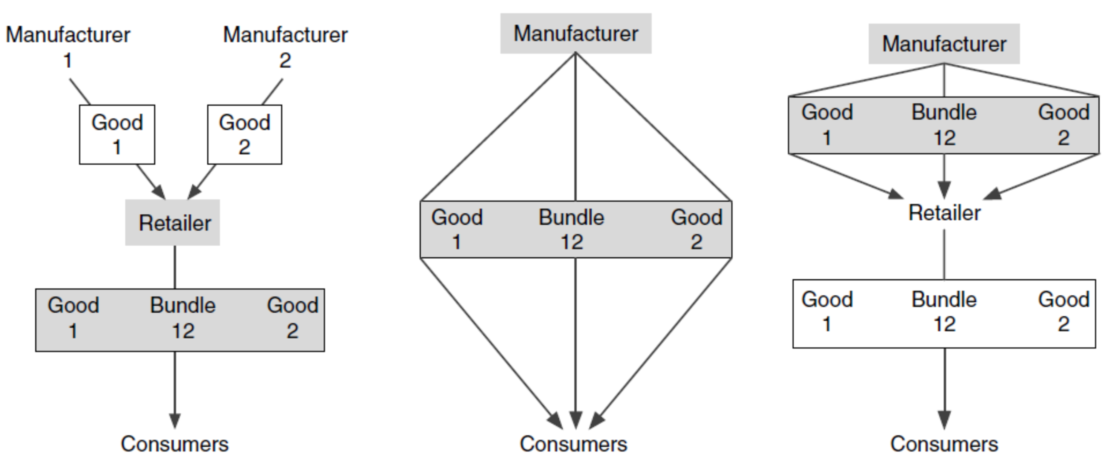
An advantage of using machine learning to study and generate promotional groupings is that it would identify trends in customer preferences. The model constantly learns from these patterns and suggests groupings that capture customer requirements of today. This is a far cry from employing a manual approach to identifying product bundles, which could fail to identify the changing trends in customer preferences. In a cut-throat industry, any lapses or delays in adapting to market requirements could lead to the Company losing significant market share.

The remainder of this paper is organized as follows: A review of the literature on various criteria and methods used for PPG selection is presented in the next section. In Section 3, the proposed methodology is presented, and the criteria formulation is discussed. In Section 4, various models are formulated and tested. Section 5 outlines the performance of our models. Section 6 concludes the paper with a discussion of the implications of this study, future research directions, and concluding remarks.

## LITERATURE REVIEW

Prior research on product bundling encompasses various types of product groupings, across various industries. Though there are fundamental differences in the approaches, a majority of the studies agree upon the fact that bundling can be viewed similar to volume discount where the volume is based on aggregate sales across products (Nalebuff 2018). Currently, bundling is approved and reviewed manually, which is time-consuming and could result in human errors. The focus is to utilize modeling to help consumer product companies identify these drivers and how their performance can be improved with respect to their competitors.

Bundling can be done at various levels. Manufacturers could group the products before passing the bundles to retailers or sell directly to consumers. Bhargava (2012) details a third type of bundling wherein retailers sell a bundle of component goods from independent manufacturers, a practice common in travel, technology, and media industries. He depicts the various types of bundling in the below diagram:




*Notes.* The shaded boxes represent the firm making the bundling decision. The $c_i$s are unit costs of component goods.

***Figure 1: Different distribution structures for product bundling***

The type of bundling is also studied by Cao et al. (2015), where they examine how bundling decisions affect wholesale price and profit.

However, this approach would not apply to the problem in hand as it would not make sense for a consumer products company to sell their products alongside a competitor's. Hence, the main point of focus would be on the third method, where the bundling decision is made by the manufacturer before passing the products on to the retailer, as depicted in the diagram. Several factors, such as marketing and economic aspects, govern the decision to bundle products. However, there are other interesting aspects which could be considered as well. Sheikhzadeh and Elahi (2013) examine the impact of heterogeneity, such as the difference in average prices, in the products to be bundled. Perhaps a more important feature would be the firm's stance on taking a risk. Some firms might prefer to experiment with novel bundling options, with the aim of increasing the expected profit, while reducing the profit variance.

While analyzing product groupings, studying the characteristics of the bundle would not yield much information as purchase patterns of the component products could vary significantly. Hence, it would be important to analyze the features of the individual products and identify characteristics that enable them to be grouped correctly.

In order to develop a model that helps companies automatically assign bundles, the study examined related papers detailing multi-classification models. Tang Fa-ming, Wang and Chen (2005) analyze different approaches of support vector machines (1 v. 1 and 1 v. rest) to develop new SVM methods based on a binary tree, which resolve unclassifiable region problems common in a conventional multiclass SVM. Their methods help reduce training time while maintaining acceptable accuracy. Fabio Aiolli and Alessandro Sperduti (2005) also introduced a novel approach of SVM Single- and Multi-Prototype SVM. This allows us to combine several vectors in a principled way to obtain large margin decision functions, and the model reduces overall problem for a series of more straightforward convex problems, with faster processing speed. Cheng and Hüllermeier (2009) conducted ML-KNN (Multi-label K-Nearest Neighbor), considering the correlations among the different labels. Thus, they are not only considering the independent variables resulting in binary classifications but also how does classification vary for different labels. Valzadegan, Jin, and Jain (2008) consider the information loss while building a multi-classification model using the Binary Tree approach. Valzadegan and his team proposed a semi-supervised boosting framework model named Multi-Class Semi-Supervised Boosting (MCSSB). This model exploits both classification confidence and similarities among examples when deciding the pseudo-labels for unlabeled examples.

While the above research methods have employed sophisticated models that improve the accuracy, a lot of the interpretability of the results is lost. As a result, they cannot accurately identify what the best grouping for a specific product is. Furthermore, multiclassification methods are hard to provide accurate results when the target variable has a large number of levels, as is the case in this study. Therefore, this research aims to leverage logistic regression to identify the key drivers of a successful grouping under each sub-category. Subsequently, the research develops a new approach to analyze data from consumer products companies and provide solutions for product bundling, namely, using optimization techniques to find out what the PPG bundle can maximize the overall revenue for the company. Rosenthal and Chaudhry (2015) used standard optimization software to analyze the relationships between different products in the vendor. While the goal of that study is different, it provides a methodology that is taken as a barometer on how to utilize optimization techniques in product bundling. Ye et al. (2017) conducted a more comparable study, where they develop a computationally efficient optimization algorithm to approximate the optimal product bundle and identifies conditions that would generate highly profitable bundling.

## DATA

The data for this study was provided by the Company. The data consists of two files – 'Transaction data' and 'Promotions data'.

Transaction data contains 553,651 observations and 21 variables over 4 years, from January 2016 to December 2019. The datasets consist of product hierarchy, description, weekly sales amount and quantity, and number of stores the product was in display for every SKU. All the products under Transaction data are from three categories: Home Care, Oral Care, and Personal Care, and 21 sub-categories, and sub-categories are crucial in later analysis.

| Category | Count of Transactions |
|---|---|
| Home Care | 90,860 |
| Oral Care | 109,766 |
| Personal Care | 353,025 |
| **Grand Total** | **553,651** |

***Table 1: Summary of transactional data by category***

Data dictionary for Transaction data is as follows:

| Particulars | Data-Type | Description |
|---|---|---|
| Category | String | Highest level of Product Hierarchy |
| Product-Category | String | Child of 'Category' |
| Sub-Category | String | Child of 'Product-Category' |
| Manufacturer | String | Company who produces/ sells products |
| Brand | String | Child of 'Product-Category' |
| Sub-Brand | String | Child of 'Brand' |
| Variant | String | Child of 'Sub-Brand' |
| Pack Type | String | Packaging Type |
| Size | String | Number of L, mls, gms, KGs, pks that a product contains per consumer unit sold |
| Segment | String | Shopper positioning of the product |
| Product Name | String | Product Name |
| Size Range | String | Size Range |
| Unit of Measure | String | Measurement unit |
| Sub-Segment | String | Child of 'Segment' |
| UPC/ SKU | Integer | Unique Product Code |
| Sales Value | Float | The Sales amount of the product for the 'Customer' by 'Week' |
| Store Count | Integer | Number of stores Product is displayed at for 'Customer' by 'Week' |
| Sales Quantity | Integer | The Sales quantity of the product for the 'Customer' by 'Week' |
| Average Sale Price | Float | Average selling price of product for a 'Customer' in a given week |
| Customer Number | Integer | Customer ID |
| Week | Timestamp | 7 day period of sales for the product within the Customer |

***Table 2: Transactional data dictionary***

Promotion data contains 559,023 records and 8 variables. Similar to Transaction data, Promotion data includes records from January 2016 to December 2019. The one variable that different is example PPG, which is the PPG bundle that been signed by Company.

| | | |
|---|---|---|
| UPC/SKU | Integer | Unique Product Code |
| Sales Value | Float | The Sales amount of the product for the 'Customer' by 'Week' |
| Store Count | Integer | Number of stores Product is displayed at for 'Customer' by 'Week' |
| Sales Quantity | Integer | The Sales quantityof the product for the 'Customer' by 'Week' |
| Average Sale Price | Float | Average selling price of product for a 'Customer' in a given week |
| Customer Number | Integer | Customer ID |
| Week | Timestamp | 7day period of sales for the product within the Customer |
| Example PPG | String | PPG bundle signed to individual product in a given week |

***Table 3: Promotional data dictionary***

## METHODOLOGY

The primary research of the paper focuses on analytically examining the promotions and identifying the significant factors that drive those promotions. Once the factors that affect the promotions are identified, the optimal product bundles, which will considerably boost the sales, are recommended in a more systematic, data-driven way that utilizes Machine Learning to bring hidden insights to life.

One of the biggest challenges the industry faces today is that companies maintain their own extensive manual list of PPGs (including competitors) in order to conduct promotions analysis. This manual approach is problematic for several reasons; the main issues are stemming from human errors and each individual's unique decision framework around what promotions label should be assigned to each SKUs.

The contributions from the analysis will provide the company with support for better decision making in its business. The success factor of a promotion must be formulated in terms of consumer preference to accomplish the objectives. Consumers tend to place a higher value on some bundles compared to other bundles due to various factors such as price and consumer utility. From a business point of view, consumer preference is identified as promotions that generate higher revenue, sell more substantial quantities, and have a longer duration of the promotion period.

The study identifies the following questions that will help the business take an analytical approach to maximize the impact of their bundles:

- What are the drivers of promotional groupings?
- How can better predictions of promotional groupings provide essential insights, decision-support, cost-savings, etc. to the business?

The study uses Python so that the codes can be easily reproduced, and the workflow can be integrated with the company's existing capabilities, solving the aforementioned problems.

**Process Flow of Methodology**

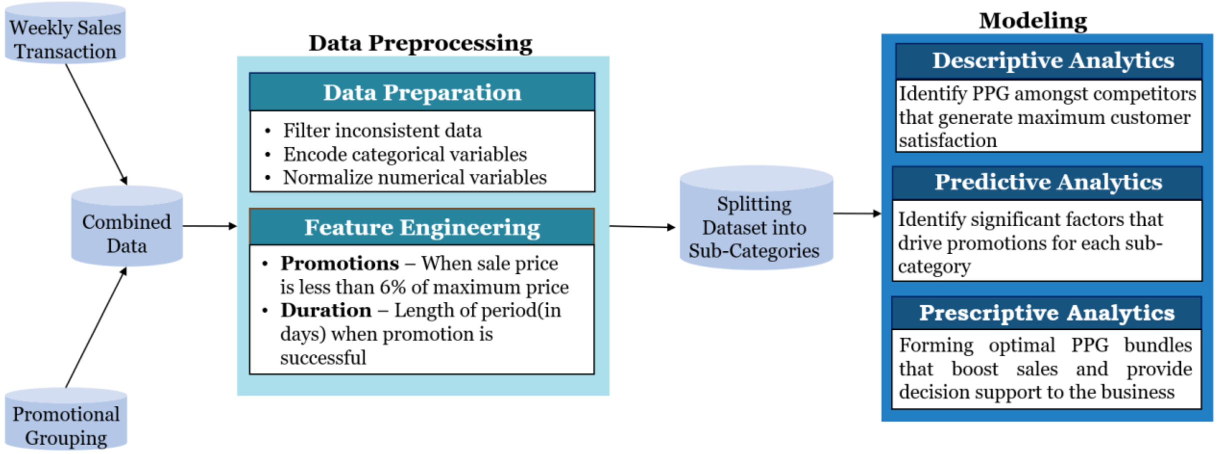



*Figure 2: Abridged methodology process flow*

**Explanatory Data Analysis (EDA)**
EDA was performed in Tableau to understand the relationship of each variable to others and obtain a thorough understanding of the data.

- Oral Care accounts for a majority of the company's revenue, sales quantity and selling price
- Compared to its competitors, the company dominated Oral Care and had an absolute advantage in the Home Care category
- Price settings for the company were lower on average

**Data Preprocessing**
The transaction and promotions data are loaded and identified for the datatypes. Cleaning and filtering were done on the datasets to remove inconsistencies. Both the datasets were merged at product, customer, and week level. Moreover, revenue, quantity, store count, and price variables were aggregated accordingly. After merging, the first data file had details of Product Information, Weekly Sales (Amount & Quantities), Number of Stores Count the product was displayed/ sold in, and the promotion group.

Data cleaning steps were performed to make the data suitable for analysis. In the Quantity column, there were some rounding-off errors in the data obtained from the system. Since quantities cannot have any decimal points, they are rounded off to the nearest 0. Furthermore, some numbers in the sales column appeared to be negative. These transactions were dropped from the dataset after receiving confirmation from the client.

Typically, one UPC code is assigned to a product. However, in the dataset, some UPC linked to more than one product. On closer inspection of the cases, this case was identified as a system issue where new names were stored in the system along with old names. Such inconsistencies have been corrected by aggregating the transactions.

**Feature Engineering**
Further information is required to obtain the solution for the research questions defined above. The dataset has details of promotion groupings at a transactional level. However, identifying promotion success is still uncertain. The first thought process was to identify for each sub-category what promotions were successful.

After conducting some detailed analysis, the study decided a promotional success period to be defined as from the starting date to the date when the store count for that bundle falls below 100. If the store count falls below 100, then it is considered that the promotion is no longer active. Thus, the duration of promotion is defined below:

$$\text{Duration: } T_{i\_below\_100} - T_{i\_start}$$

$T_{i\_below\_100}$: the first day that promotion goes below to 100 after the promotion start for product i
$T_{i\_start}$: the first day that promotion start for product i

In order to identify whether the product was sold at its base price or a discount, a Promotions variable was created. If the product was sold at a value lower than 6% of the base price, then the Promotions is 1, and is 0 otherwise.

$$\text{Promotion: } Price_{xt} \leq 0.94 * \max(Price_x)$$

max($Price_x$): Max price of Product x during the given period
$Price_{xt}$: Price of Product x sold at time t

In order to incorporate the factors and find the product bundle with the high success of the promotion, a score is calculated using the revenue, quantities, and duration values from the data. The values are standardized using a min-max scaler so that all the values are relative to each other and fall between 0 and 1. An equal weightage is given to revenue, quantities, and duration value, and the average is taken to form the score value. This score gives a value to the bundles, which has a high consumer preference:

$$\boldsymbol{Score} = Average\left(\frac{x - min(x)}{max(x) - min(x)} + \frac{y - min(y)}{max(y) - min(y)} + \frac{z - min(z)}{max(z) - min(z)}\right)$$

Where,
x: Revenue
y: Sales Quantity
z: Duration

The PPG with the highest score across sub-category and size range is called a ‘Model PPG’ which is compared with the corresponding bundles of each manufacturer at the same level. The target variable is created, which is 1 if the bundle is the same as the Model PPG, and 0 otherwise.

**Descriptive Analytics**
Once the model product bundle for each sub-category and size level is identified, it is compared with the corresponding bundles of each manufacturer at the same level. The manufacturer who has sold the maximum number of successful promotional bundles is recognized, and the significant factors driving such promotions are examined using a classification technique in the modeling phase.

**Predictive Analytics**
The study determines the important factors driving a successful promotion across each sub-category using classification algorithms. Furthermore, the model evaluates a sample promotional grouping and estimates the probability of its success.

**Prescriptive Analytics**
The next objective was to build a model that can cluster products into a new grouping that will increase sales. The optimization approach was chosen over clustering to identify groupings with business constraints. One drawback of clustering is that it does not consider business constraints while forming clusters. The non-linear optimization model in MS Excel was used to solve this question. For every category, objective function, decision variables, and constraints were defined under the same scales. This process was performed for every sub-category separately as a bundle is generally formed of products within the same sub-category. The Evolutionary method was run for all combinations to provide the solution with Global Maximum. The framework is set up interactively such that in the future, variables can be adjusted to provide new results.

## MODELING AND RESULTS

## LOGISTIC REGRESSION

The objective of this modeling is two-fold. First, to understand the factors that drive the success of the promotion across each sub-category. Second, to quantify and estimate the probability of success of a promotion using the significant factors across each sub-category.

This study employs Logistic Regression as it is one of the most popular methods to solve this binary classification problem. It uses a Sigmoid distribution, and it is easier for organizations to interpret variables and implement this algorithm for decision-making.

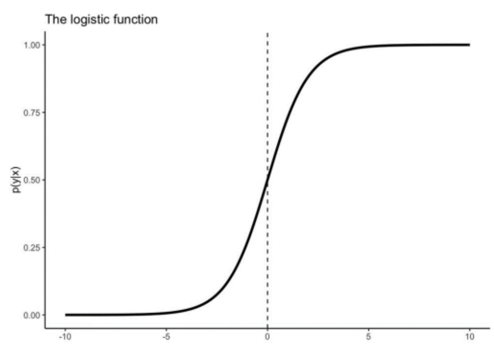



***Figure 3: Sigmoid function for Logistic Regression***

Promotion success is the target variable to be analyzed. The data is filtered for the specific manufacturer who excels in promotions of that sub-category. The independent variables used to understand the important attributes are: Variant, Pack Type, Segment, Size range, Sub-segment, and Promotions. Since all the independent variables are categorical, one-hot encoding is implemented. The data is then oversampled to balance the classes and modeled using Statsmodels in Python for Logistic Regression.

## Results

### Identifying important attributes

|  | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Promotions_1_0 | -0.9849 | 0.104 | -9.477 | 0.000 | -1.189 | -0.781 |
| Segment_EPOS_Sgel Men's | 2.0342 | 0.148 | 13.716 | 0.000 | 1.743 | 2.325 |
| Segment_EPOS_Sgel Milk/Skincare | 4.6818 | 0.366 | 12.789 | 0.000 | 3.964 | 5.399 |
| Size_Range_EPOS_Sgel 251-400ml | -14.6447 | 4.431 | -3.305 | 0.001 | -23.329 | -5.960 |
| Size_Range_EPOS_Sgel 401-500ml | 1.0792 | 0.096 | 11.282 | 0.000 | 0.892 | 1.267 |
| Size_Range_EPOS_Sgel 501-750ml | -9.2551 | 0.678 | -13.645 | 0.000 | -10.584 | -7.926 |
| Size_Range_EPOS_Sgel 751-1500ml | 11.9239 | 7.705 | 1.547 | 0.122 | -3.178 | 27.026 |
| Size_Range_EPOS_Sgel <100ml | 3.9743 | 3.115 | 1.276 | 0.202 | -2.131 | 10.080 |
| Size_Range_EPOS_Sgel >1501ml+ | 0.7517 | 0.566 | 1.329 | 0.184 | -0.357 | 1.860 |

***Table 4: Identifying important attributes for a PPG at sub-category level***

The above results depict the modeling for Shower gels. The levels Men's and Milk/Skincare under Segment, the levels 251-400 ml, 401-500 ml and 501-750 ml under Size Range, and Promotions have a significant impact on the success of a promotion as the p values are less than the 95% confidence interval. Although other levels within the variables are not significant in the model, due to the hierarchy principle, all levels are included when one of the classes is significant.

A similar approach is executed to observe the important features for other sub-categories. The results are shown as follows:

| Sub Category | Variant | Segment | Size Range | Sub Segment | Promotions |
|---|---|---|---|---|---|
| Shower Gels |  | ✔ (Men's) (Milk/Skincare) | ✔ (251-400ml) (401-500ml) (501-750ml) |  | ✔ |
| Cleaning Wipes | ✔ (Other Fragrance) | ✔ (General Purpose) |  |  | ✔ |
| Electric TB |  |  | ✔ (4pk+) | ✔ (Rfl) |  |
| Manual TB |  |  | ✔ (2pk) |  | ✔ |
| Mouthwash Bottle | ✔ (Ice) |  |  |  | ✔ |
| Sprays |  | ✔ (General Purpose) (Glass) (Kitchen) | ✔ (376-500 ml) |  | ✔ |

***Table 5: Summary of important drivers of PPGs for each sub-category***

### Estimating probability of success

To interpret the coefficients, the Logistic Regression uses the equation:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

From the above results, the probability of success of a promotion is estimated for one variable holding other factors constant.

If Men's segment is used in the promotional bundling, keeping other factors constant, there is an 88% probability of success of promotion indicating that when shower gels are bundled with Men's segment, there is an 88% chance that the promotion is successful.

Similarly, the probability is estimated for other variables across each sub-category.

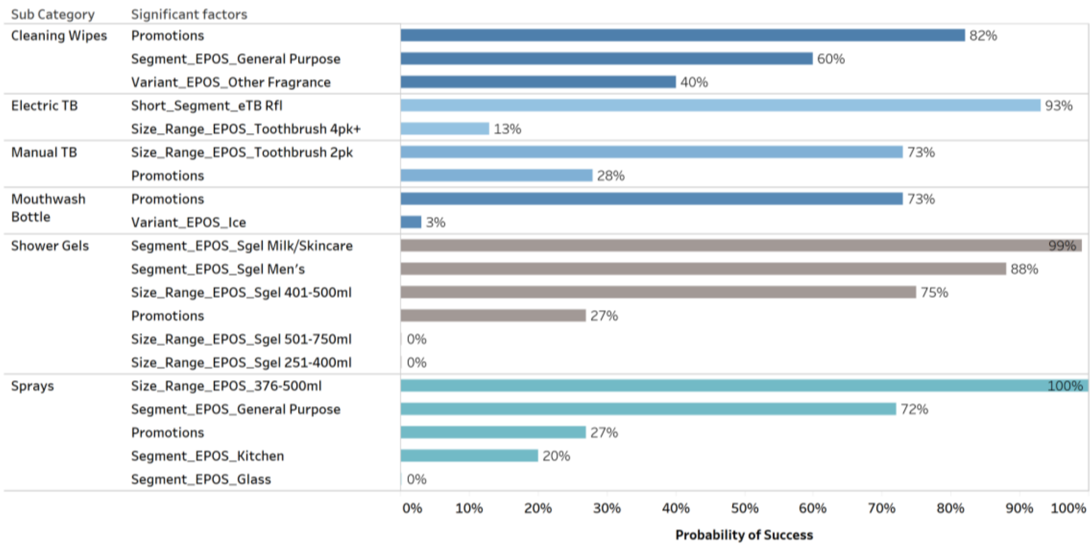



***Figure 4: Probability of success for important attributes in each sub-category***

## OPTIMIZATION

Product bundling plays a vital role in enhancing both profits and the company's growth. Bundling is based upon the idea that customers usually save more on the value of the grouped package than the individual items when purchased separately. Customers generally compare product prices and love choices before making a purchase.

In this section, the Optimization Technique was used to form new promotion bundles.

### Goal

The goal of the model is to identify new promotions groupings that can boost the Company's sales.

### Solving Method

Since the revenue function is non-linear and non-smooth, an Evolutionary Algorithm is preferred as the solving method.

Process flow for the Evolutionary Algorithm is as follows:

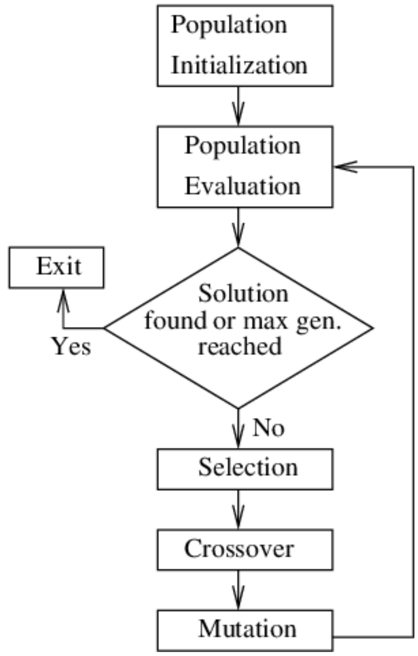



***Figure 5: Process Flow of Evolutionary Algorithm***

Evolutionary algorithms perform well-approximating solutions to all types of problems because they do not make any assumptions about the underlying fitness landscape.

## Advantages of Evolutionary Algorithm

Evolutionary algorithm optimizers are global optimization methods and scale well to higher-dimensional problems. They are robust to noisy evaluation functions, and the handling of evaluation functions that do not yield a sensible result in a given period is straightforward. The algorithms can be easily adjusted to the problem at hand.

Evolutionary algorithm calculates the global optimum, and not only the local optima, giving the most optimal results.

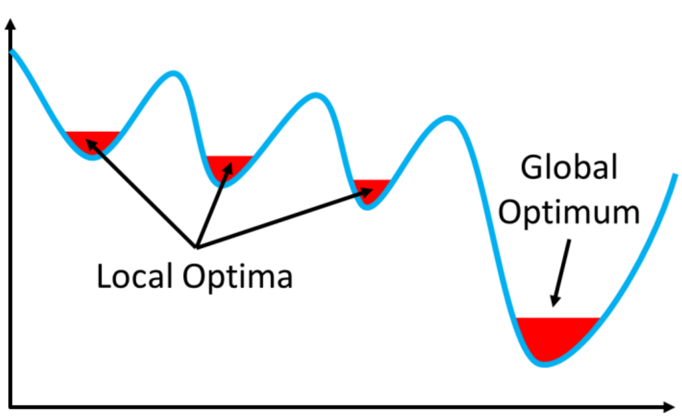



***Figure 6: Difference between Local optima and Global Maxima***

## Disadvantages of Evolutionary Algorithm

While research has been done on which algorithm is best suited for a given problem, this question has not been answered satisfactorily yet. While the standard values usually provide reasonably good performance, different configurations may give better results. Furthermore, premature convergence to a local extremum may result from adverse configuration and not yield (a point near) the global extremum.

Furthermore, as evolutionary algorithm considers all possible combinations, it is time-consuming.

## Objective Function

The objective of the Optimization Solver is to maximize revenue.

## Decision Variables

Optimization model was developed for each sub-category. Each unique product (SKU) is considered as an individual decision variable. Our aim is to find a combination of products that can be combined together to form a bundle for promotion.

## Constraints

***Number of products in a bundle:***

$$\mathbf{2 \leq n(X_i) \leq 6}$$

There will be at least 2 distinct products in a PPG and at-most 6 distinct products in a PPG.

***Decision Variables can take only Binary Non-Negative values***

$$\mathbf{X_i = \{0,1\}}$$

A product can be included in the bundle only once. This constraint ensures that no same products can be included in the bundle more than once.

***Price of the bundle***

$$\sum_{i=1}^{n} P(X_i) \leq \frac{\sum_{j=1}^{N} P(X_j)}{N} * n(X_i)$$

The price of the bundle should not be too high. The sum of the price of individual products in the bundle should be less than the average price of all the products in the bundle multiplied by the total number of products in the bundle.

***Quantity Sold***

$$\sum_{i=1}^{n} Q(X_i) * Corr(X_i, n) \geq \frac{\sum_{j=1}^{N} Q(X_j)}{N} * 2n(X_i)$$

Corr ($X_{i,n}$) refers to the Selling Correlation between two products. For example, if A & B are sold separately, quantities sold are x & y, respectively. However, if they are sold together, then the total quantity sold will be 1.1 * (x+y). The factor of 1.1 suggests that the products complement each other.

Due to the limited availability of data, the study has considered 10% as the correlation factor between any two products. A more accurate correlation factor can be computed using Market Basket Analysis if associations between products is known using the transaction data.

As per this constraint, the quantity sold for the formed PPG bundle should be more than the average sold quantities of products in the sub-category times the number of products in the bundle times 2. This will ensure that the quantities sold are higher through bundling.

***Revenue earned***

$$\sum_{i=1}^{n} R(X_i) \geq \frac{\sum_{j=1}^{N} R(X_j)}{N} * 2n(X_i)$$

Total revenue after optimization should be higher than the current cumulative revenue earned by all the products in the sub-category.

## Tool used for Optimization

Microsoft Excel Solver was used to run optimization and obtain results. The parameters used are as follows:

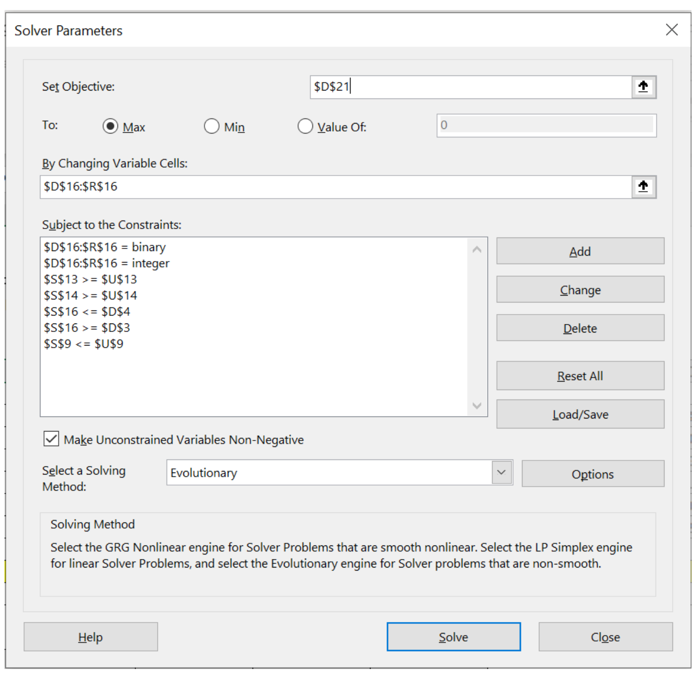



***Figure 7: Parameters used for solving Optimization Problem***

## Results

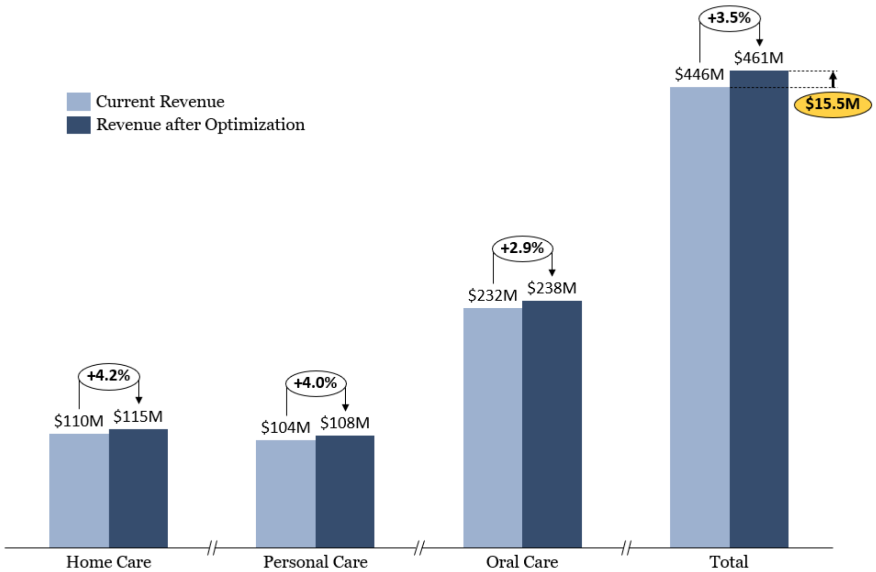



***Figure 8: Revenue generated for each category before and after Optimization***

The Company has three categories, and the chart reflects the revenue generated before and after optimization. Using optimization, the study was able to increase the revenue of the Company by 3.5% per annum, which is tantamount to a $15.5 million growth in revenue.

## CONCLUSIONS AND FUTURE SCOPE

### Manufacturing Cost of Product

The product margin can be calculated if manufacturing cost of the products are provided. This can then be used in the analysis and optimization models. This will determine successful promotion campaigns with a positive and increased margin.

### Store wise analysis

By obtaining stores wise transactions, data can be analyzed to identify demographic information and consumer preference. This data can be used to provide personalized recommendations to increase customer loyalty.

### Customer Transaction Analysis

If transactional data of customer spending is obtained, Market Basket Analysis can be used to uncover associations between items. It will help improve promotion bundle formation.

# REFERENCES


Nalebuff, B. (2004). Bundling as an entry barrier. *The Quarterly Journal of Economics*, 119(1), 159-187.

Bhargava, H. K. (2012). Retailer-Driven Product Bundling in a Distribution Channel. *Marketing Science*, *31*(6), 1014–1021. doi: 10.1287/mksc.1120.0725

Cao, Q., Geng, X., & Zhang, J. (2015). Strategic Role of Retailer Bundling in a Distribution Channel. *Journal of Retailing*, 91(1), 50–67. doi: 10.1016/j.jretai.2014.10.005

Sheikhzadeh, M., & Elahi, E. (2013). Product bundling: Impacts of product heterogeneity and risk considerations. *International Journal of Production Economics*, *144*(1), 209–222. doi: 10.1016/j.ijpe.2013.02.006

Tang, F.-ming, Wang, Z.-dong, & Chen, M.-yun. (2005). On Multiclass Classification Methods for Support Vector Machines. *Control and Decision*, *20*(7), 743–754. doi: 10.13195/j.cd.2005.07.29.tangfm.006

Aiolli, F., & Sperduti, A. (2006). Multiclass Classification with Multi-Prototype Support Vector Machines. *Journal of Machine Learning Research*, *6*, 817–850.

Cheng, W., & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, *76*(2-3), 211–225. doi: 10.1007/s10994-009-5127-5

Valizadegan, H., Jin, R., & Jain, A. K. (2008). Semi-Supervised Boosting for Multi-Class Classification. *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science*, 522–537. doi: 10.1007/978-3-540-87481-2_34

Rosenthal, E. C., Zydiak, J. L., & Chaudhry, S. S. (1995). Vendor Selection with Bundling. *Decision Sciences*, *26*(1), 35–48. doi: 10.1111/j.1540-5915.1995.tb00836.x

Ye, L., Xie, H., Wu, W., & Lui, J. C. (2017). Mining Customer Valuations to Optimize Product Bundling Strategy. 2017 IEEE International Conference on Data Mining (ICDM). doi: 10.1109/icdm.2017.65

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2010). Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints*. *International Conference on Computational Linguistics*, (23), 1272–1280.