# A supply chain analytics approach to product assortment optimization

*Rajesh Kumar (rajesh.kumar7@hp.com), Vishwanathan Rajagopalan, Jayesh Baldania, Nidhi Sagar, Santanu Sinha, Priyanka Dahiya, Jyotirmay Nag, Sudip Chatterjee*
*Global Analytics, Hewlett Packard, Bangalore, India*

## Abstract

Retailers and CPG manufacturers struggle to create a store level product assortment for their thousands of stores, which is critical to achieving revenue and margin targets. Our solution to this complex problem of choosing right product mix and new product introductions uses a combination of store clustering; product profiling and attributes based assortment optimization. Our solution built on a big data platform of Vertica /R is easily able to scale in the voluminous point-of-sale data environment and when applied in the context of a major CPG manufacturer has helped to identify new revenue opportunities of up to sixteen percent.

**Keywords:** Assortment optimization, Big-data, Store clustering

### Introduction

Assortment planning is a problem that retail chains and CPG manufacturers have dealt with from time immemorial. Assortment planning can be defined as determining the set of products that needs to be carried at each store at each point in time. It is done for a fixed period of time like six months to a year and also involves questions like how much of the buying budget and store space should be allocated to each product category. Amongst these assortment related problems, we tackle the following in our paper:

**Finding a manageable level for assortment planning**

Most of the retailers have thousands of stores and thousands of SKUs (Stock keeping units). Typically one would like to do an assortment at the individual store level but due to the number of store-SKU combinations that can exist, this becomes a daunting task. In this context one of the important questions is what is the right level for assortment planning?

**Planning assortment of existing products**

There is an opportunity to improve the assortment of existing products by introducing top products in stores where they are currently not selling and reviewing the non-performing ones. In this context the following questions emerge:

- How do we compare the performance of various products taking into account multiple performance criteria?

- How can we improve sales from current products that we sell?
- Which top products are not positioned in the stores?
- Which are the non-performing products?

**Planning for new product introductions**

One of the challenges in assortment planning is the decision on which new products to introduce in the market. In this context the following questions emerge:

- What are the gaps in the product attribute mix that we offer to the customers?
- Which new products can we strategically introduce in the market?
- How to optimize the mix of product attributes and their stocking levels in the stores?
- What is the impact on revenue by offering more/less of an attribute?

In this paper we try to answer these business questions.

## Literature review

There are multiple papers available in the area of retail assortment planning. However we refer to some of the closely related papers that we have looked at while coming up with our approach. Fisher and Vaidyanathan (2007) look at the assortment planning problem from the point of view of attributes that customers want from the products that they buy. They also look at demand substitution between different attributes that the customers want. Saure and Zeevi (2013) look at the problem from a view of customer utility maximization and propose a tradeoff between the exploration that a customer does when he/she visits the store and exploitation in terms of the revenue maximization for the retailer. Hubner and Kunh (2012) look at the assortment problem from the perspective of space elasticity of demand for the shelf space allocated to different products and the demand substitution between various product categories.

In contrast, we look at the assortment planning problem from three angles: First, at what level should the assortment planning problem be handled? e.g. store level, store cluster level, geography level etc. Second, we look at how product assortments can be improved from the current list of SKUs. Thirdly, we also look at which new SKUs can be introduced in the stores that will help increase the revenues. We use a combination of clustering, scoring and optimization techniques in our approach. We also use point-of-sale (POS) data in our analysis which means massaging millions of lines of records and hence we had to use a big data setup of Vertica/ R to manage this huge amount of data which is also something new in our approach. Vertica is a big data management platform and R is an open source statistical analysis package.

## Methodology

Our approach comprises of fours steps to solve the problem:

**Data massaging**

Our data source as mentioned above was point-of-sale data for 18 weeks of one of the largest retailers in UK. This data was for products related to a big CPG manufacturer. The category of products that was chosen for analysis was baby care, specifically Diapers (Nappies and Pants). This POS data was cleaned and massaged to segment all products based on the attributes that customers look for while buying them. These attributes included: Baby size – this is dependent on the weight of the baby in kilo grams, Diaper type – Nappy (taped version) or pants (pull on version), Quality of diaper which is dependent of price point and finally Pack Size of diapers which varies from 19 to 104 pieces per pack. Thus we had the following values for these attributes in increasing order:

- Baby sizes: Min, Mid, Max, Large, Extra Large
- Diaper Type: Nappies, Pants
- Quality: Low, Medium, Premium
- Pack sizes: Small, Medium, and Big

The summary of sales based on product attributes is shown in Table 1.

*Table 1 - Summary of diaper sales based on product attributes*

**Quality and Pack size**

Baby size and Type

| BS_NP | Low_Small | Low_Medium | Low_Big | Medium_Small | Medium_Medium | Medium_Big | Premium_Small | Premium_Medium | Premium_Big |
|---|---|---|---|---|---|---|---|---|---|
| Min_Nappies | 0 | 0 | 0 | 238597 | 187873 | 111600 | 0 | 0 | 0 |
| Mid_Nappies | 27802 | 19832 | 5228 | 60246 | 176363 | 79331 | 26072 | 57659 | 10083 |
| Max_Pants | 0 | 0 | 0 | 0 | 0 | 0 | 69762 | 32425 | 0 |
| Max_Nappies | 66504 | 85147 | 13053 | 95129 | 311295 | 155748 | 32217 | 212585 | 49819 |
| Large_Pants | 0 | 0 | 0 | 0 | 0 | 0 | 88958 | 37304 | 0 |
| Large_Nappies | 55564 | 77416 | 14847 | 175359 | 151422 | 186606 | 54677 | 245870 | 56587 |
| Extra_Large_Pants | 0 | 0 | 0 | 0 | 0 | 0 | 163557 | 30608 | 0 |
| Extra_Large_Nappies | 20285 | 0 | 37113 | 75907 | 139150 | 72085 | 0 | 0 | 0 |

**Store clustering**

One of the business problems in assortment planning is deciding the right level at which to do the planning. Ideally one would like to do this at the store level; however because of the large number of stores and SKUs, the task becomes very difficult.

To address this, our solution looks at clustering the stores first into homogeneous groups based on multiple criteria. The idea here is that alike stores should sell similar type of products. The philosophy in doing a store clustering is to create store groups which are alike from a perspective of product types that they sell, store type, store sale volumes and revenue and store location. We have written algorithms that can work with big data like POS data along with store attribute and product data to generate store clusters in the Vertica/R environment. Our algorithms use supervised clustering techniques like Classification and regression trees (CART). Our algorithm enables the understanding of why stores are grouped together. Based on a Scree plot we determine the optimal number of clusters which in our case comes to around 10-12. The Scree plot shows that the error in predicting the classes does not decrease beyond 10-12 clusters as shown in figure 1. Post the store clustering, the cluster properties are as shown in the table 2. The idea is to create similar assortments for all stores falling in the same cluster.

**Product profiling**

One of the questions in assortment optimization is how to increase revenues from the current set of SKUs that a retailer has. We resort to product profiling to help with this question. Various criteria like recency of sales of a product, frequency at which it sells, the monetary value attached to its sales etc. are used to determine how important a product is and whether it should find its place in the assortment or not. However other factors like substitution and cannibalization are looked into while determining whether a product should be taken out from the assortment. We use a Scoring methodology to create the profile of all products based on their sales characteristics like recency, frequency and monetary value etc. The score is called an RFM score and is calculated as follows:

RFM Score = Recency*Frequency*Monetary value

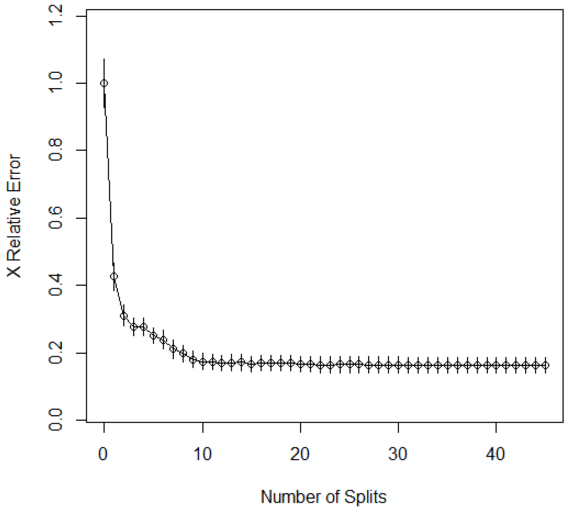


*Figure 1 - Scree plot to evaluate optimal no of clusters*

We create a pareto plot of the cumulative RFM scores of SKUs arranged in descending order of their RFM scores as shown in figure 2. We decide a cut-off on the profile scores for the products and the products which are above this cut-off are treated as 'Top products' which should be listed in all stores of the cluster. The products which are below this cut-off need to be reviewed to be taken out of the store. Interestingly we find that a large number of stores do not list the top products. For e.g. 79 of the top 101 RFM SKUs of cluster one is not listed in 30-40 stores (figure 3). The recommendation is to list these top SKUs in these stores.

*Table 2- Store cluster details and cluster properties*

| Cluster # | Cluster Distribution (# of Stores) | % of XYZ Turnover | Avg. Turnover(£) | Cluster Characteristics |
|---|---|---|---|---|
| 1 | 293 | 27.89 | 27784 | Relatively Bigger **pack size** and **Store channel** as Express, High Street, Northern Ireland, Superstore, XYZ Direct, XYZ Homeplus and High **Turnover** |
| 2 | 271 | 55.94 | 60245 | Relatively Bigger **pack size**, **channel** as Dotcom, London, Extra and Very High **Turnover** |
| 3 | 170 | 8.67 | 14879 | Smaller **Pack size** & **High Turnover** |
| 4 | 170 | 2.78 | 4778 | Primarily smaller **pack size** and medium quality and Medium **Turnover** |
| 5 | 210 | 1.23 | 1704 | Primarily Smaller **pack size** , Primarily medium **quality** & Low **Turnover** |
| 6 | 158 | 0.07 | 125 | Smaller **pack size,** Medium **quality** and store **channel** as London, High Street and Low Turnover |
| 7 | 288 | 1.03 | 1040 | Smaller **pack size**, Medium **quality** and Store **channel** Express and NI and Very Low Turnover |
| 8 | 158 | 0.27 | 497 | Smaller/Medium **pack size**, Medium **quality**, Bigger **baby sizes**, Store **channel** Express and NI and Very Low **Turnover** |
| 9 | 44 | 0.07 | 443 | Medium **quality**, Primarily Bigger **baby sizes**, Only small **pack sizes** apart from small, Store **channel** Express and NI and Very Low Turnover |
| 10 | 30 | 0.05 | 437 | Smaller **pack size**, Medium **quality**, Primarily Bigger **baby sizes**, Store **channel** Express and NI, Left of Reading and Low Turnover |
| 11 | 224 | 0.69 | 904 | Smaller **pack size**, Medium **quality**, Primarily Bigger **baby sizes**, Store **channel** Express and NI, **Left** of Aberdeen and Low **Turnover** |
| 12 | 507 | 1.33 | 763 | Smaller **pack size**, Medium **quality**, Primarily Bigger **baby sizes**, Store **channel** Express and NI, **Right** of Aberdeen and Low **Turnover** |
| | | 100 | 11568 | |

3 Clusters account for 93% of Turnover

Source : Turnover of Nappies and Pants for All XYZ Stores for 18 weeks ending august 2013

**Attribute based strategic assortment optimization**

One of the questions in assortment optimization is which new SKUs to introduce in the stores. Our product attribute based approach analytically finds patterns in the way customers choose product attributes and is similar to the approach used by Fisher and Vaidyanathan (2007). The theme in this step is that customers don't buy products but they buy a bundle of attributes. For e.g. for diaper products, customers might be looking at baby size, diaper type (taped or pant), quality of the diaper and pack size while making their purchasing decision. We estimate the demand for each attribute based on historical customer preferences for these attributes. Then we look at the attribute combinations where either we may have products but have not put them in the stores or we may not have products at all. In such cases we first estimate the total demand for the attribute combination and then try to compute the fraction of this demand for the white spaces where we don't have products in such a way which minimizes the sum of errors across the entire grid. We resort to constrained optimization to solve this problem. The solution recommends volume changes for both existing products and new products. For e.g. the cells in Table 1 which have zero sales are the potential candidates for introducing new products. There are two scenarios here. Scenario 1, where the retailer has existing SKUs for the attribute combination but has deliberately not kept them in stores and Scenario 2 where the retailer needs to create new SKUs for the attribute combination. Figure 4 shows the sales volumes proposed for

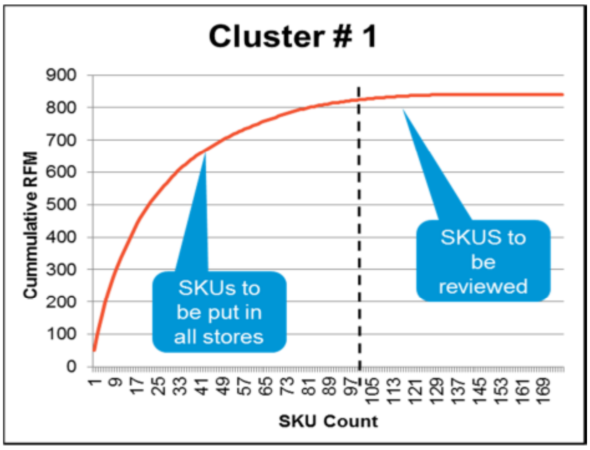


*Figure 2 - Product profiling of SKUs in cluster 1*

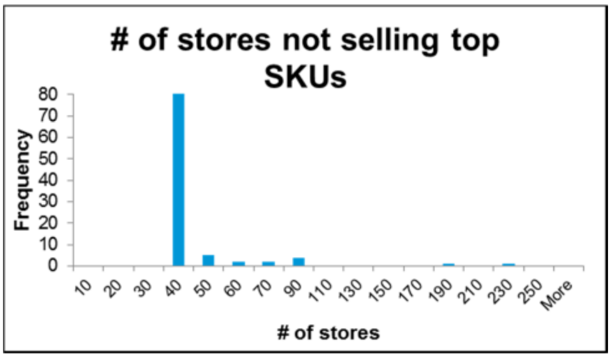


*Figure 3- Seventy nine top SKUs not selling across 30-40 stores of cluster 1*

both of these scenarios based on the above optimization model. We clearly see a lift in the volumes for both low and premium product segments.

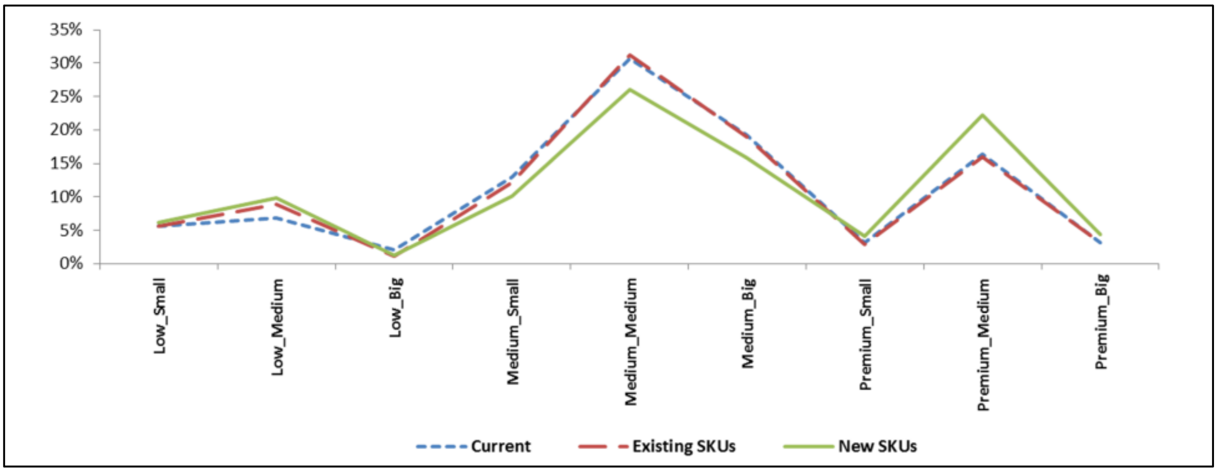


*Fig.4 Optimized sales volumes for cluster 1 stores post introduction of SKUs*

**Results**

Our assessment finds significant increase in revenue with optimization of current products and introduction of new products in all the cluster stores. We find that there is a potential to increase revenues by 10.3% by just introducing the top SKUs based on the RFM approach in the stores. We also find an additional revenue increase opportunity of 16.5% by introducing new SKUs based on the product attributes optimization in the stores.

**Future scope**

In the clustering approach, the clusters would become more meaningful and robust if we were to add other data types like demographic data or loyalty card information. We currently lacked this data but would try to make the analysis more robust in future by adding these data as well. The RFM model does not take into account the effect of cannibalization between product attributes which can be taken care of in future analysis and research. Also, this approach has been currently applied on just one retailer in UK and for just one product category. The models and the analysis methodology will be replicated for other retailers in UK along with other product categories apart from baby care.

**Disclaimer**

"This document is published only for information purpose. The document is provided without any representation and in no event shall HP be liable with the use of the information made available in the document. The information has been provided on an "as is" basis without warranty of any kind. The document is made available for information purpose only; further copying, reproduction or redistribution of the document is expressly prohibited. In no event shall HP be liable for lost profits or any direct, indirect, special, punitive, incidental or consequential

damages arising out of or in usage of the document. HP is registered trademark of Hewlett Packard and its subsidiary".

## References


Fisher M.L., Vaidyanathan R. 2007. An algorithm and demand estimation procedure for retail assortment optimization. *White paper.*

Hubner A.H., Kuhn H. 2012. Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management. *Omega* **40** (2): 199-209.

Saure D., Zeevi A. 2013. Optimal dynamic assortment planning with demand Learning. *Manufacturing and Services operations management* **15** (3): 387-404.