U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Assortment Optimization in Online Retail

Helena Beatriz Oliveira Santos

Dissertation
Master in Data Analytics

Supervised by
Dalila Benedita Machado Martins Fontes

2019

“Without data you are another person with an opinion”

W. Edwards Deming

Acknowledgements

I would first like to thank my supervisor, Professor Dalila Fontes. She allowed this thesis to be my own, but I must acknowledge that nothing would have been possible without her help and expertise.

This dissertation was made in parallel with my professional activity and using data from my employer which I may therefore acknowledge. Moreover, I would like to thank my work colleagues Filipe, Rita B., Afonso, Rita R and my Boss Nascimento.

My parents, brother, as wells as my family in general have always been supportive of my choices and helped me through all my academic path.

I would also like to thank my friends with whom I shared my concerns during this period. Including those who join me in my concerns, while writing their own dissertations, but specially those who heard me unconditionally.

A special thanks to my basketball teammates who were the ones that probability dealt with most of my tiredness during this period.

The support of ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-031821 is acknowledge.

Abstract

The Problem proposed is to solve a real-life Assortment Problem for an Online Groceries Retailer. This comprises a problem of Assortment Planning and Optimization, where the question is one of how many unique SKUs to be kept in the assortment of a specific subcategory, rice, i. e., the products offered to customers at any given point in time.

The goal of this assortment reformulation exercise is to maximizate the revenue. Assortment affects revenues since it impacts directly the consumers' purchasing decisions, as well as, the Retailer's reputation in the longer run. In most proposed models, the revenues are estimated using a ranking-based consumer choice model and the prices of the SKUs. However, we propose an extension where the stockout levels of each SKU is also considered in the revenue estimation. Additionally, another extension is proposed at the constraints level where a prevention of eliminating all SKUs from one brand is imposed, so that the resulting assortment does not cause the Retailer to stop carrying out any of the existing brands. Additionally, our proposal presents a methodology which uses as a solo input the transactional log.

The assortment obtained from the proposed model, when analysed in the sale's period following the one used as input for the model, showed that even when reducing 33% of the original assortment 89% of sales would still be assured by the remaining SKUs. Additionally, the kept SKUs contributed to only 55% of stockout occurrences. However, these are only direct impacts. Other resulting affects cannot be quantified, such as the effect of substitutions or improvements in operational efficiency due to assortment reduction.

The proposed and studied approach is valid for the presented problem. Nevertheless, a different software could be necessary to solve more complex implementations of the model. Yet, for this particular retailer the results could be replicated for other subcategories, of similar complexity. In fact, the approach was also tested for the sugar subcategory with similarly favourable results. Finally, this approach in comparison with existing ones provides the opportunity of less blind assortment decisions. Even if the customer preferences are the main input, via past transactional logs, the business decision makers can also offer some input in the form of reduction restrictions. In our work, all brands were kept in the assortment, but in the future other business constraints can be easily added to the model.

Contents

List of Tables

List of Figures

List of Abbreviations

IIA	Indeprndence of Irrelevant Alternatives
MILP	Mixed-Integer Linear Programming
MIO	Mixed-Integer Optimization
MMNL	Mixture of MultiNomial Logits
MNL	MultiNomial Logit
NL	Nested Logit
PLD	Product Line Design
SKU	Stock Keeping Unit

1. Introduction

Product assortment, the range of products a Retailer offers its customers, is one of the main concerns in the Retail Industry, leading and intertwined with many strategic decisions. In the last decade there has been an increasing concern with the exponential growth of assortments, due to the higher costs related with keeping larger inventories and the resulting higher rates of out-of-stocks items. Therefore, retailers are interested in reducing assortments. Broniarczyk, Hoyer, and McAlister (1998) found that eliminating the lowest selling items, which accounted for up to 54% of the original assortment did not impact the sales level. Better yet, Boatwright and Nunes (2001) analysed 42 categories of an Online grocer and found an average of 11% increase in sales due to drastic cuts in the lowest selling items.

Assortment decisions can be done regarding three levels of specifies (Mantrala et al. (2009)): the variety, the depth, and the service level. The decision of variety of assortment is the definition of the number of categories to be carried out by the Retailer. Dedicated stores choose as their business strategy to carry a very limited number of categories but a very rich array of products from those categories. Contrarily, generalist stores usually carry less products from each category but have a much larger offer of categories. Therefore, the decision of variety is related to the type of Retailer and is made more on a business strategy level. At a finer level, the question of depth comprises the decision of the number of different SKUs that should be carried out within a specific subcategory of the full assortment. Finally, there is the decision also needs to be made of the service level, this means the number of units to keep in stock, at a particular store, of each of the SKUs decided to be kept in the assortment.

The Assortment Planning and Optimization Problem addressed here is specific to the Assortment Decision of depth. More specifically, we address a real-life Assortment Problem for an Online Groceries Retailer. The answer that is seek is of how many unique SKUs should be kept for a defined subcategory.

Mantrala et al. (2009) also present a high level and qualitative overview of all the factors that can affect the Assortment Decision process, which can be once more divided into three types: Consumer Perceptions and Preferences, Operational Limitations and Environmental Factors.

In the perspective of the consumers, the assortment offered can significantly impact the way the Retailer is portrayed and its reputation. Thus, sales, and the corresponding revenues, are directly affected by the assortment offered. Therefore, the decision of which products to include in the assortment is of great importance and will be the main considered factor in our work.

To start with, customers present a lot of heterogeneity in their preferences (Green and Krieger (1985)). In fact, all Retailers have various consumers and the same product is not going to be everyone's favourite. Additionally, for a specific consumer its preference will may over time (McAlister and Pessemier (1982)) and it can also be unstable, some costumers do not have a preferred product (Bettman, Luce and Payne (1998)). Furthermore, it has been proven that customers are known to seek variety in order to have flexibility of choice (Drolet (2002)). Therefore, large assortments can have advantages since they allow for very different consumer preferences and their heterogeneity, exchangeability and flexibility seeking behaviour. However, Fitzsimons et al. (1997) defend that large assortments can also be disadvantageous due to provoking an overwhelming feeling in the consumer. When consumers are faced with a big set of products to choose from, and do not have a clear preferred one, they can feel quite frustrated and give up on the purchase altogether. Other factors can contribute to a non-purchase behaviour such as the non-presence of a preferred item, for consumers with well-defined preferences, either due to out of stock situations or because the preferred product was not included in the assortment (Verhoef and Sloot (2006)). Independently, for each costumer its perception of Assortment can be distinct from the one actually offered (Broniarczyk et al. 1998).

Briefly mentioned, shelf space allocation is the most studied limitations in operations research, see Bultez and Naert (1988). Additionally, macroeconomic and environmental trends have also impact on the Assortment offered and expected by consumers (Reuters (2008)). However, these factors are not critical for the problem at hand.

In our work, a more in depth description of the problem is presented in Section 2, where the business model is also detailed to exalt the business requirements that must be accommodated by the model. On Section 3, an overview of consumer choice models and assortment optimization approaches is presented based on reviewed literature. On Section 4, the chosen methodology is shown, as well as, its extensions. The results of our approach are presented in Section 5. Finally, the main conclusions are presented in Section 6.

2. The Problem

Assortment Optimization has been presented via a general overview of its applications and factors. Now, the problem addressed is described.

The goal is to apply the most fitting of the existing approaches in the related literature to a real-life assortment decision for an Online Groceries Retailer by adapting it to the retailer's specific business needs.

The decision in hands, is one of depth. As presented in the previous chapter that entails decisions on how many, and which, SKUs of a specific subcategory should be included by the retailer in the assortment. Therefore, here and hereafter, the term Assortment is used to refers to the set of SKUs offered.

The goal is not to evaluate and reduce the whole assortment of the company since the retailer is a generalist and sells a wide range of both food and non-food products. Additionally, this is an experimental exploit for the company which wouldn't run the risk of an all-around assortment reduction, even if proven beneficial in concept. Therefore, the goal is to create a model using a specific subcategory that may, at a later stage, be replicated as for all other subcategories.

2.1 Original Assortment

Within the wide range of products offered by the company, a choice must be made regarding the subcategory to be used to conduct the first experiments. The ideal subcategories to perform an assortment reduction on would be those which are non-seasonal and non-trend dependent, this entails essential products (consumer staples) to which little to no novelty is introduced nowadays. Most basic food and beverage products fit into this description such as rice, pasta, and water. Additionally, subcategories such as these ones have scarcely taste sensitive demand and quite wide but repetitive offer.

To define the priorities in optimization and the subcategory to apply the model to, the subcategories with most SKUs, within non-taste influenced food categories, were identified and ranked decreasingly by percentage of assortment without sales, daily average, since this is a good metric for unnecessarily big assortments.

Table 1Subcategories and corresponding percentage of assortment without sales

subcategory	% of products within the assortment sold (daily average)
rice	11,8%
sugar	12,1%
pasta	14,8%
flour	17,0%
healthy basic ingredientes	21,5%
grains and seeds	22,7%

Taking into consideration the assortment analysis presented in Table 1, the subcategory chosen for the assortment reduction and the main focus of this dissertation is the rice subcategory. Additionally, the time period which will be used for taking insight into the model will be the first trimester of 2018.

The data available from the Retailer includes the transactional log of every sale of products in the rice subcategory in the first trimester of 2018. A twin data set of the immediately following semester will also be provided. The later will be used to analyse assortment obtained by the model we propose.

The format in which the transactional log is stored provides a record each time a SKU is bought. This contains the identification of the product, the brand to which the product belongs to, the date of transaction, the quantities requested and delivered, the price of the product, and a unique identifier of the customer. There are also some additional recorded details, but the ones listed are the ones that will be used as inputs to the model.

Table 2Data Entry exemplification of the transactional log

SKU	Brand	Date	RequestedQty (units)	DeliveredQty (units)	Price	Customer
X1	B1	02/01/2018	2	1	0,99	C1

2.2 Business Constraints

The business model of the Retailer also impacts the Optimization Model implemented for the problem. Therefore, an overview will be presented. This Online Retailer makes part of a bigger Retailer that invoices most of its sales in Offline Traditional Grocery Stores. Firstly, the Online Assortment may be different from the Offline one, just like differently located physical stores can have distinct Assortments. However, the supply chain resources are the same. This means that there is only one integrated team that treats all supply chain of this nation-wide Retailer as a whole. Additionally, most of the operation of the Online Retailer is located in-store. That is, the products that are sold to the online consumers come straight off the shelfs of the physical stores and the orders are prepared for shipping in the back of the store. Therefore, there is no specific warehouse dedicated to the Online Retailer. The main predictions for stock are made pertaining to the physical stores. The only contribution of the Online is a small adjustment made by the weight of its sales in each subcategory of products.

There are some exceptions of high rotating products for which the Online division has a dedicated stock in the shipping preparation area and also some centralized products, mostly non-food items of big dimensions. But since only SKUs in a food subcategory are going to be analysed here, these exceptions will not need to be considered in the model.

As the business model shows the resources are shared between the Online and Offline versions of the Retailer and, as already mentioned, the Offline has a significantly higher percentage of the combined Retailer's sales. Therefore, the assortment decisions made by the Online team do not necessarily provoke immediate changes regarding the shared supply chain resources since it may not fit the Offline strategy or needs.

Taking this into consideration the Optimization Problem that is going to be proposed will only take into consideration, as valuable constraints, the consumer perceptions and preferences. Since Business Constraints and Environmental Factors, such as the ones presented in the previous chapter, do not have the same impact in this Online Retailer decision.

2.3 StockOut

Taking into consideration the business model, abvove presented, there is a stockout occurrence when the online picker doesn't find in the store the needed product or in the needed quantity to fully satisfy the demand. Remember that the Retailer does not have autonomy to implement changes in the supply chain.

Therefore, the online operation must adapt to the in-store condition. For example, when it comes to fresh baked goods from the store bakery, the online retailer operational team gathers the full day aggregated demand so that first thing in the morning the order can be processed, baked and set aside for the online customers. Previously to the implementation of this method, the online picker had to grab a ticket and wait in line with the in-store customers every time bread was needed in a new order. The change aloud the picking process to be faster as well as reduce stockout since the physical customers do not affect the stock before the online operation does. However, solutions like this one cannot be applied for all product subcategories for many reasons, such as lack of man power, insufficient storing space, among others.

The Rice products are a part of the basic ingredients section and follow a multi-order picking methodology. This entails that each picker assigned to this specific zone will fulfil up to 4 orders at a time, picking up all the requested products for each order that are located at the basic ingredients section. In a day there are as many rounds as necessary to fulfil all demand, with an almost continuous flow of pickers in the aisles. During this process, if the needed product is available on the shelf it is allocated to the orders in process, otherwise there is a stockout.

The replenishment of shelf is done at the same time for each section. However, some rice SKUs have much more frequent stockout occurrences than others. Therefore, it can be concluded that stockout occurrences are SKU dependent. Many causes can be found: some products' suppliers have a larger fulfilment time period than others, some SKUs are restocked on the store's shelfs during the day, while others are not (depending manly on the in-store demand). Since these realities cannot be changed, in order to reduce overall stockout of rice products there is the need to influence a slight change in the online offer towards the SKUs with lowest levels of stockout.

3. Literature Review

Due to its relevance, not only in an academical setting, but also for real world decision making, there is a vast array of literature on the topic of Assortment Optimization. This encompasses many fields of study such as Marketing, Operations Management and Retail. The following review will only touch on the key aspects of the literature that relate to the underlying problem rather than presenting a full view of all literature. Therefore, three areas are going to be reviewed: Consumer Perceptions of Assortment, Choice Models to estimate the demand associated with each Assortment, and the Optimization Approaches used in solving similar problems.

3.1 Consumers perceptions of assortment

The consumers perceptions of Assortment differentiate from the actually offered Assortment. Therefore, the impact of Assortment reduction should be measured by the customers perceptions of the reduction since that will be the driver for the changes in sales patterns.

Broniarczyk et al. (1998) presented the three most influencing factors that affect Consumers Assortment Perceptions. These consist of the number of SKUs offered for a specific category, the shelf space allocated to that category, and the presence of a consumers most preferred SKU in that category. They studied a reduction of 25%-50% in the popcorn category and one of the most relevant findings was the indifference towards the Assortment reduction of customers whose most preferred SKU was still present, given the same shelf space allocated to the whole category.

Product dissimilarity within Assortment has also been studied by the likes of Hoch et al. (1999) and Van Herpen and Pieters (2002) with a product-based and attribute-based approach, respectively. The goal of both studies was to test the impact of uniqueness of product in the Consumers Perceptions of Assortment, in these studies the products were deconstructed into a combination of attributes. The findings lead to an agreeing conclusion showing that consumers perceive a diversified assortment if the set of offered SKUs contained diverse attributes and if the level of similarity between the products was low, this requires small redundancy between the attributes that make up the products in the assortment. The existence of very similar SKUs, even if it increases the offer, makes the Assortment seem less wide to the consumers.

Additionally, there are also studies on how the initial Assortment characteristics can impact the Consumers Perceptions in cases of Assortment reduction. Van Herpen and Pieters (2002) proved an inverse relation between the initial Assortment size and the perceptions of magnitude reduction. For a bigger initial Assortment, an equal number of SKUs cut, will seem like a smaller reduction when compared to a smaller sized initial Assortment. Additionally, Broniarczyk and Hoyer (1998) show that the sales distribution over the initial Assortment will affect the impact on sales as a consequence of assortment reduction. They showed that when the distribution of consumers through all the initially offered SKUs is close to uniform, an equivalent Assortment reduction will be perceived as bigger when compared to an initial distributed concentrated on just a few SKUs.

Out of all the Consumer Perceptions of Assortment studies that have been done, the work of Boatwright and Nunes (2001) might be the most interesting one regarding our problem, since it also involved on an Online Retailer. The purpose of their study was to analyse the sales before and after an Assortment reduction and identify the different factors that most influenced the changes in the sales. The factors in the core of the study were brand, flavour, market share and brand-size combinations offered.

Consumers are quite heterogenous, therefore, not all reacted to the reduction in the same way. Sales to some consumers decreased, due to a negative impact of the reduction, but sales to other consumers increased, as a positive effect of the decluttered Assortment. However, analysing the full set of customers as a whole, Boatwright and Nunes (2001) concluded that reducing the number of Brand-Size Combinations lead to increase in sales due to a decluttering effect. This corroborates the studies on product dissimilarity presented in this section. However, these authors found that even though small cuts in terms of number of Brands can increase sales, deep reductions in the offered brands have a negative impact on sales. The impact on sales due to reductions in flavour and market shares were found to be similar to the impact presented for reductions in the number of brands offered.

In sum, Boatwright and Nunes (2001) proved that simplifying the choice in Online Retailers via Assortment reduction can increase sales. Especially, if redundant combinations are eliminated while keeping constant the number of brands, sizes, and flavours.

3.2 Choice Models

There are many variables that can influence the Consumers Perceptions of an Assortment, consequently there is the need to model customers behaviour, for a specific set of products. On the one hand, Choice Models were used in Marketing research with the only underlying objective of studying consumer behaviour. On the other hand, in research more related to Operation Management, Choice Models are used as predictive tools to aid a larger decision. To solve our problem, the predictions of consumers choices, while deciding on products for each given assortment, will be used to calculate the expected demand associated with each assortment. Choice Models are used to estimate the conditional probability of a costumer buying a given product when a specific assortment is presented and, therefore, is one of the main factors for estimating revenues.

The array of explored Choice Models can be divided into two main distinct groups: Parametric approaches and General Choice Models, that will be reviewed in the next sections.

3.2.1 Parametric Choice Models

Out of the studied Parametric Choice Models, multinomial logit model (MNL) is the one that has been the most applied in the context of Operation Management. This model was introduced by Luce (1959) and Plackett (1975) in independent studies. Its property of being tractable when estimating parameters is one of its bigger advantages, given that these models try to incorporate cognitive complexity relying on rational choice theory. However, the MNL model does not fully capture the substitutions customers may make when their most preferred SKU is not present (Debreu (1960)). Substitution of products is, in fact, a phenomenon that wasn't even taken into consideration in some of the first decision models. Additionally, in the MNL model the assumption of independence of irrelevant alternatives (IIA) is present (Ben-Avika and Lerman (1985)), which in most cases is not the verified pattern on real sales data.

Some models were latter introduced that tried to solve these applicability issues of the MNL model such as the nested logit model (NL) and the mixture of multinomial logits model (MMNL) as thoroughly explored by Grün and Leisch (2008). These more complex models are, however, less identifiable from transactional data and are prone to overfitting. Additionally, all of these models are based on structural assumptions which can impact significantly their accuracy.

In sum, MNL is the simplest model but can lead to data underfitting, since it ignores some important consumer behaviours, like product substitution. However, the more complex models derived from the MNL, like the NL and MMNL, are prone to overfitting the data. Consequently, any of these parametric Choice Models can easily lead to suboptimal assortment decisions.

3.2.2 General Choice Models

In order to avoid the above mentioned risks associated with the implementation of Choice Models which rely on structural assumptions, one can use General Choice Models, also known as Non-Parametric Choice Models.

An example of such models is the one proposed by Rusmevichientong et al. (2006) that solves the IIA assumption problem associated with the MNL and is more stable in terms of under or overfitting the data. However, this model needs data on the consumers preferences for all possible assortments, which makes it hard or even impossible to use, since most companies do not have the required dataset for this.

The work of Blanchet et al. (2013) present a similar problem of requiring a data structure and evidence that does not exist in real transactional data. The approach proposed views SKUs as absorbing states in a Markov chain and models the consumers substitutions decisions as state transitions in the defined Markov chain. The probability of choosing a given SKU in an Assortment is given by the absorption probability of the corresponding state.

Farias et al. (2013) presented an approach that is applicable to real transactional data, out of which the preferences of customers can be observed from only the smaller set of Assortments that were offered over time. In this approach, the choice model is essentially "the probability of purchase of a particular product in N given the set of alternatives available to the customer" (Farias et al. (2013)). The probabilities are estimated for each distinct costumer type by creating a list of preferences, this list in a permutation of the products extracted from the costumer's past purchases and ordered by a descending ranking. Therefore, no assumptions are implicit in the model other than that a consumer has a preference over the offered SKUs in the assortment and that all consumers are going to purchase their most preferred SKU. The calibration of the ranks of preferences can be obtained via the conversation rates found in transactional data, even if only a subset of Assortments was previously offered to the consumers. Revenue estimation is usually the main goal in Operations Management Problems and, therefore, Farias et al. (2013) propose a formulation of a revenue function based on its choice model. They propose computing the revenue by assuming the lowest revenues possible for the Assortment based on the previously described choice model. That is, the revenue is computed using the worst-case probability distribution found in the transactional data and the underlying rankings of preferences. Additionally, Farias et al. (2013) tested their choice model and predictions against those produced by the most common choice models in literature, such as the parametric models MNL and MMNL. The findings based on both real and synthetic data have shown their model to make more accurate predictions.

Bertsimas and Mišic (2015) propose some changes to the Farias et al. (2013) model that allows to formulate the resulting Assortment Optimization Problem as a Mixed-Integer linear programming model. Bertsimas and Mišic (2015) also define a probability distribution over a set of rankings. However, instead of considering the lowest possible revenue case, they "fix a small set of rankings over the products and a probability distribution over this small set of rankings" (Bertsimas and Mišic (2015)). This can be interpreted as modelling the choice preferences of a fixed set of consumers and then applying the probabilities of a new consumer having the same preferences as ones modelled in the fixed set. Bertsimas and Mišic (2015) also show that their approach results in more accurate predictions than those of Farias et al. (2013).

Based on the rankings defined over the set of products that is identic to the Farias et al. (2013) approach, Bertsimas and Mišic (2015) propose a procedure of solving an explorer approximation problem using column generation in order to identify the probability distribution over the set of rankings. Bront et al. (2007) had already proposed a column generation algorithm, although applied to a deterministic linear programming model, with the same goal of predicting choice behaviour and its impacts on revenues.

A similar approach was proposed by Ryzin and Vulcano (2015) in which a column generation step is also presented although their approach uses a maximum likelihood probability rather than an ℓ_1 approximation. However, this alternative approach requires a prior selection of the model in order to prevent overfitting. Additionally, Bertsimas and Mišic (2015) approach is also preferable due to the smaller complexity of the model, contributing to a less likelihood of overfitting. This happens because having only a small set of rankings with a non-zero probability lead to basic feasible solutions to the bigger problem. When estimating a consumer's choice behaviour towards an assortment its relation towards every offered product must be considered. However, since Bertsimas and Mišic (2015) use a ranked list of preferences, it becomes unnecessary to analyse the products that are less favourable than the non-purchase alternative reducing the assortment to study and, consequently, the complexity of the problem.

3.3 Assortment Optimization Approaches

In addition to the reviewed literature of Choice Modelling there is also a related body of work regarding Assortment Optimization under those Choice Models. Accordingly, those approaches can also be divided into two different sections: based on Non-Parametric Choice Models and based on General Choice Models.

As mentioned previously, most Choice Models try to estimate consumer behaviour to aid Retailers in many decision-making problems other than Assortment Optimization. Therefore, a lot of authors try to use existing Choice Models to describe consumer choices and the corresponding revenues. Retailers' goal is the maximization of revenues. In Assortment Optimization there is no difference, and in fact, the goal is to maximize revenue by selecting which and how many unique SKUs should be offered to customers.

In the remaining of this section, an overview of Optimization approaches based on Choice Models is presented, although mostly restricted to the Choice Models previously introduced.

3.3.1 Parametric Choice Models

The Parametric Choice Models open possibilities of applying Assortment Optimization approaches of the type "fix-then-exploit" as characterized by Bertsimas and Mišic (2015). These approaches require a selection of a parametric model that is best thought to fit the data and following an exploration of the structure of the resulting Optimization problem.

These problems' goal is to maximize the revenue which is calculated taking into consideration estimates of consumer choice behaviour using the fixed parametric model. Some of the proposed approaches develop exact solutions. Nevertheless, some heuristics have also been proposed.

Of these approaches, the ones building upon the traditional MNL model and its variants are some of the most exploited ones. The most recent ones include the MNL model presented by Talluri and van Ryzin (2004), the NL model explored by Davis et al. (2014), and the MMNL model as studied by Rusmevichientong et al. (2004). Building on previous choice models incorporating Markov Chain, Feldman and Topaloglu (2014) propose an Assortment Optimization solution approach.

However, as previously mentioned, simple Parametric Choice Models are prone to underfitting data, while the more complex ones are prone to overfitting. Therefore, their accuracy its too volatile depending on how well the defined structure fits the data, having a direct impact on the solutions provided by these Optimization approaches.

3.3.2 Based on General Choice Models

On the other hand, Assortment Optimization approaches can also be based on General Choice Models. However, since these Choice Models are quite recent, their use for Problems is still scare, especially in terms of Assortment Optimization.

Additionally, since these approaches don't consider any preconceived distribution of consumers' behaviour, they rely on rational choice theory and, consequently, produce more complex problems. This is mostly the case of the choice model used by Farias et al. (2013), since it considers all possible ranks over the set of products, one would be required to solve some complex Mixed-Integer Optimization Problem with a large number of decision variables.

Nonetheless, the ADXOpt algorithm introduced by Jagabathula (2014) is a valid and promising approach for finding solutions to the Farias et al. (2013) Assortment Optimization Problem. However, this approach uses a local search algorithm and, therefore, cannot guarantee the solution found to be a global optimum. Although the solution found is a local optimum, it provides better Assortment Decisions than the Parametric Choice Models based approaches.

Contrarily, the approach presented by Bertsimas and Mišic (2015), where a probability function is used over a smaller subset of rankings, is formulated as a Mixed-Integer Linear Optimization (MIO) Problem much less complex than that of Jagabathula (2014).

As mentioned in the beginning of this section the body of work pertaining to Optimization of Assortment Decisions is not that vast and, therefore, Bertsimas and Mišic (2015) approach is based on a similar problem regarding the Product Line Design (PLD) context. The PLD problem solving approach that these authors base their approach on is the work of Belloni et al. (2008).

Belloni et al. (2008) approach focusses on the decision of introducing a new product that can encompass a range of distinct attributes in a market characterized by heterogenous customers and, therefore, the parallelism with the Assortment Optimization Problem is not farfetched. The real PLD problem that Belloni et al. (2008) proposed to solve was quite more complex than an average Assortment Optimization problem focused only on the SKUs for a specific subcategory. As a result, their approach, even after some sophisticated enhancements were added such as Lagrangian relaxation and valid inequalities, is computationally very demanding, having taken more than a week to solve the given PLD problem.

However, the Assortment Optimization problem proposed is fairly simpler, in comparison. Accordingly, in their computational experiments Bertsimas and Mišic (2015) found that the resulting MIO problem could be solved to optimality almost instantly even without using the referred enhancements of the original approach.

Bertsimas and Mišic (2015) MIO solution approach presents many advantages over other mentioned approaches. Firstly, the formulation allows for easy adaptation to diverse problems. Secondly, because it is highly flexible, it can easily accommodate business strategic rules as linear constraints.

3.3.3 The MIO Formulation

The Assortment Optimization Problem as in Bertsimas and Mišic (2015) will now be presented.

The probability function that describes the consumers choice behaviour, adapted from Farias et al. (2013) approach, for any given customer selecting product *i* when assortment S is offered, is formulated as:

$$P(i|S) = \textstyle\sum_{k=1}^{K} \lambda^k . \| \{i = arg_{i' \epsilon S \cup \{0\}} \min \sigma^k(i')\} , \qquad (3.1)$$

Where the probability of product *i* being purchased when the assortment *S* is offered, results of the sum of the decisions of all possible *k* customer types, weighted by the probabilities of a costumer being of each type. For each type *k,* the decision is based on the ranking based list of preferences. The subexpression:

$$\prod\{ i = arg_{i' \epsilon S \cup \{0\}} \min \sigma^k(i')\} \qquad (3.2)$$

takes the value of 1 if the SKU *i* is the most preferred SKU in the given assortment for the customers of type *k*. In fact, the SKUs are ordered from most preferred into less preferred in the vector σ^k, and, therefore, the selected SKU will be the one with the minimum index in this permutation. The subexpression 3.2 will take the value of 0 for all other products.

The total revenue of the assortment S, R(S) is calculated as follows:

$$R(S) = \sum_{K=1}^{K} r_i . P(i|S) \quad , \qquad (3.3)$$

Where r_i is the price associated with SKU *i*. By substituting P(*i*|S) in Equation 3.3 we obtain:

$$R(S) = \textstyle\sum_{i \in S} r_i \, (\sum_{k=1}^{K} \lambda^k . \prod\{ i = arg_{i' \epsilon S \cup \{0\}} \min \sigma^k(i')\})$$

Therefore, we want to find the Assortment S^* that maximizes the total revenue:

$$S^* = arg_{S \subseteq \{1,\dots,n\}} \max R(S) \qquad (3.4)$$

The complete MIO model is given in Equations (3.5) through (3.11). The decision variables are x_i and y_i^k, with $i \in$ S and $k \in$ K, where S is the given assortment and K is the set of customer groups. Both x_i and y_i^k are binary variables: x_i is set to 1 if SKU i is included in the assortment and to 0 otherwise. While y_i^k is set to 1 if it is chosen by customers of group k. Nevertheless, due to the model constraints it is sufficient to define y_i^k as non-negative since it will take the largest possible value below x_i, which is either 0 or 1.

$$\max_{x,y} \sum_{k=1}^{K} \sum_{i=1}^{n} r_i . \lambda^k . y_i^k \tag{3.5}$$

$$subjectc\ to\ \sum_{i=0}^{n} y_i^k = 1, \forall\, k \in \{1, \dots, K\}, \tag{3.6}$$

$$y_i^k \leq x_i, \forall\, k \in \{1, \dots, K\}, i \in \{1, \dots, n\}, \tag{3.7}$$

$$\sum_{j:\sigma^k(j)>\sigma^k(0)} y_j^k \leq 1 - x_i, \forall\, k \in \{1, \dots, K\}, i \in \{1, \dots, k\}, \tag{3.8}$$

$$\sum_{j:\sigma^k(j)>\sigma^k(0)} y_i^k = 0, \forall\, k \in \{1, \dots, K\}, \tag{3.9}$$

$$x_i \in \{0,1\}, \forall\, i \in \{1, \dots, K\}, \tag{3.10}$$

$$y_i^k \geq 0, \forall\, k \in \{1, \dots, k\}, i \in \{1, \dots, n\}. \tag{3.11}$$

The constraints in (3.6) to (3.11) describe the rationality in the consumers' purchase decision making. Constraint (3.6) imposes that customers of each group chose exactly one SKU, while constraints (3.7) ensures that only SKUs in the original assortment are chosen. Constraints (3.8) ensures that when a customer's type most preferred SKU is kept in the assortment ($x_i = 1$), then those customers will purchase that SKU. Constraints (3.9) ensures that the costumer's decision regarding non-preferred SKUs is not to buy. Finally, Constraints (3.10) and (3.11) force the decision variables binary nature.

The solution of this model is an optimal Assortment S* that maximizes the total revenue. The performance of Assortment S*, in the real problem, depends on the accuracy of the choice model used to estimate customers preferences.

4. Methodology

Based on the reviewed literature and the problem specificities presented formerly, the approach by Bertsimas and Mišic (2015) seems to be the most appropriate one to solve the proposed problem. With its choice modelling using ranking based preferences that can be learned from real transactional data and the resulting MIO problem that arises when optimizing the total revenue associated to the consumer choice estimated behaviour.

The present work aims to build upon the Bertsimas and Mišic (2015) model by making a few extensions based on business insights. Firstly, the revenue will not be considered as just a function of price but rather as a combination of price and stockout. Secondly, from a commercial strategy imposition, it is required that no brands be entirely removed from the assortment.

Boatwright and Nunes (2001) have also shown that, alongside the brand dimension, reductions in flavours and sizes offered also affect significantly the consumer perceptions. However, flavour and size are characteristic, that do not vary for every subcategory, and for which data is not always sufficiently detailed in the company records. Therefore, these characteristics will not be taken into consideration in our model. Thus, brand is the only product characteristic that will be identified for every SKU and added as a constraint to the problem so that no brand will be extinct as a consequence of all its products being removed from the assortment.

4.1 The Proposed Model

In the literary review the work of Bertsimas and Mišic (2015) was found to be the best approach so far to assortment optimization problems and their formulation was detailed in depth leading to the following linear programming formulation:

$$\max_{x,y} \sum_{k=1}^{K} \sum_{i=1}^{n} r_i . \lambda^k . y_i^k \qquad (4.1)$$

$$subjectc\ to\ \sum_{i=0}^{n} y_i^k = 1, \forall\ k \in \{1, \dots, K\}, \qquad (4.2)$$

$$y_i^k \leq x_i, \forall\ k \in \{1, \dots, K\}, i \in \{1, \dots, n\}, \qquad (4.3)$$

$$\sum_{j:\sigma^k(j)>\sigma^k(0)} y_j^k \leq 1 - x_i, \forall\ k\ \in \{1, \dots, K\}, i \in \{1, \dots, k\}, \qquad (4.4)$$

$$\sum_{j:\sigma^k(j)>\sigma^k(0)} y_i^k = 0, \forall\ k\ \in \{1, \dots, K\}, \qquad (4.5)$$

$$x_i \in \{0,1\}, \forall\, i \ \in \{1, \dots, K\}, \tag{4.6}$$

$$y_i^k \geq 0, \forall\, k \in \{1, \dots, k\}, i \in \{1, \dots, n\}. \tag{4.7}$$

Recall that i indexes the SKUs in the initial assortment S and k are the different customer types. The goal is to maximize the revenue through a revision of the offered Assortment. The decision variables are x_i and y_i^k where x_i represents whether the i SKU is kept in the assortment ($x_i = 1$) or whether it is removed ($x_i = 0$) and y_i^k is 1 if SKU i is the chosen one for the costumers of type k and 0 otherwise. In the objective function, r_i represents the price of SKU i and λ^k is the probability of a costumer being of type k. Lastly, σ^k is the permutation that records the preferences of customer type k. The notation $\sigma^k(i)$ represents the position of SKU i in the permutation of preferences for customer type k and $\sigma^k(0)$ indexes the position of the non-purchase alternative.

Roughly speaking the constraints entail the decision process of the costumers and the observed reality that each costumer buys its most preferred SKU in the given assortment, not excluding the non-purchase alternative. For a detailed description refer to section 3.3.3.

Let us proceed with a more in depth analysis of the preferences of the costumers:

- Constraints 4.2 indicate that any given customer type k can only have one preferred SKU.
- Constraints 4.3 declares that an SKU can only be chosen if it is present in the assortment. Therefore, if the SKU is removed from the assortment ($x_i = 0$) then y_i^k must be 0. Note that constraints 4.7 assures the non-negativity of y_i^k. However, if the SKU belongs in the assortment ($x_i = 1$) then y_i^k can be either 1 or 0 depending on whether or not the SKU is the preferred product of the customer type. In the latter case, the decision is done in order to maximize the revenues.
- Constraints 4.4 shows that if a product i is present in the assortment and is costumer type k favourite SKU then no other product will be bought by this costumer type. However, if that SKU is not in the assortment, the customer will choose the next indexed SKU in its preference's permutation. Note that such choice may be the non-purchase alternative.

- Constraints 4.5 ensures that all SKUs, for customer type *k*, indexed after the non-purchase alternative in the customer's permutation will never be purchased by the customer type *k*.

In summary, a SKU will only be bought by a specific costumer type if it is in the assortment and is the most preferred item of the costumer type under consideration. Therefore, the expression representing the buying decision of a costumer type *k* will be:

$$\sum_{i=0}^{n} x_i * Y_i, \qquad (4.8)$$

where x_i is, as before, a binary variable taking the value 1 if SKU *i* is in the assortment and 0 otherwise. While Y_i is a binary parameter set to 1 if SKU *i* is costumer type *k*'s favourite product and 0 otherwise.

However, the Retailer's revenue is the result of the buying decision of all customer types. Therefore, the total revenue will be the result of the sum of the *Yi* for all customer types, weighted by λ^k which is the probability of a costumer being of type *k*.

$$\sum_{k=1}^{k} \lambda^k * y_i^k, \qquad (4.9)$$

The goal of this optimization problem as presented by Bertsimas and Mišic (2015) is to maximize the revenue of a given assortment, for which, in their work, was used the price of the product. Adding this variable to the Expression in 4.8 and aggregating all the customer type's as shown in Expression (4.9) the resulting objective function is as follows:

$$\max_{x} \sum_{i=1}^{n} x_i * r_i \sum_{k=1}^{k} \lambda^k . y_i^k \qquad (4.10)$$

Due to operational or inventory reasons, many times, part of the demand will not be satisfied. Even if a product is present in the assortment, a costumer is of type *k* and the SKU is that costumer type's favourite, that might still not convert to the revenue of the company in case of stockout. Since the product is not stocked, it will not be delivered, and the costumer will not pay for it.

Therefore, in our estimation of revenue the stockout will also be incorporated. Thus, the revenue contributions of each SKU will be the price adjusted using information on the stockout occurrences. In result we will introduce another variable, $S_{i,}$ in the calculation of the revenue. This variable represents a percentage that affects demand. More details of its estimation and calculation are presented in the data engineering section.

Lastly, the main constraint in the present problem is the requirement of keeping at least one SKU of each of the brands present in the original assortment. Accordingly, it will be necessary a constraint for each brand, indicating which SKUs belong to which brand to assure that the sum of the decisions (x_i) for all the SKUs of one brand must be at least one. For this a new variable will be introduced z_i^b which is 1 if the SKU *i* is of brand *b* and 0 if its brand is another.

Taking into consideration these last additions to the problem the formulation that is aimed to be solved is as follows:

$$\max_{x} \sum_{i=1}^{n} x_i * \frac{r_i}{s_i} \sum_{k=1}^{k} \lambda^k . y_i^k$$

$$subjectc\ to\ \sum_{i=1}^{n} z_i^b \geq 1, \forall\, b \in \{1, \dots, b\},$$

$$\sum_{j:\sigma^k(j)>\sigma^k(0)} y_j^k \leq 1 - x_i\, , \forall\, k\ \in \{1, \dots\ , K\}, i \in \{1, \dots, k\}$$

$$y_i^k \leq x_i, \forall\, k \in \{1, \dots, K\}, i \in \{1\, , \dots\ , n\},$$

$$\sum_{i=0}^{n} y_i^k\ = 1, \forall\, k \in \{1, \dots, K\},$$

$$y_i^k \geq 0, \forall\, k \in \{1, \dots, k\}, i \in \{1, \dots\ , n\}.$$

$$x_i \in \{0,1\}, \forall\, i\ \in \{1, \dots\ , n\}$$

$$z_i^b \in \{0,1\}, \forall\, i\ \in \{1, \dots\ , n\}, b\ \in \{1, \dots\ , b\} \qquad (4.11)$$

4.2 Approach

For the model, presented in the previous Section, the problem is one of binary linear programming, a specific case of a Mixed-Integer Linear Programming Problem. There can be many distinct methods for solving this type of models, such as exact methods, cutting plane and branch and bound. However, branch and bound stands out since it builds upon the cutting plane method and opens more possibilities.

Furthermore, the main goal, even if the development and implementation have intervention of specialized professionals, is for the commercial team to have comprehension over the process and some autonomy to replicate for other subcategories. Thus, the choice of the solver approach must take into consideration several factors such as the complexity, the model characteristic and the final users.

Given the context of the problem and primary users, the software chosen to run the solver was R. This is a programming language already used in other internal process of this Online Retailer and has a global network of users active in an online community. Additionally, many of the branch and bound solvers have available callable packages in R.

4.2.1. Branch and Bound

Through the branch and bound approach, the Mixed-Integer Linear Programming problem is divided into subproblems by relaxations of the original problem and then solved by using bounds for the original problem. Its process is many times associated with a tree-like structure for which each node is a subproblem and the root is the original MILP.

Through Branch and Bound a serious of relaxations of the original MILP are solved and the obtained solutions are the bounds. Since our problem is one of maximization, here and hereafter, the solution of a subproblem will be referred as lower bounds. This process is looped until a solution, if one exists, is found using a search strategy to define the order of processing each subproblem. During the processing of each subproblem, a node is not branched if there is no possible solution or if a better solution has already been found. Otherwise, the branching occurs, here a branching method is used to define the strategy in which way the subproblem is divided into new nodes.

Accordingly, there are four key aspects to any branch and bound algorithm: The lower bounding method, the upper bounding method, the branching method and the search strategy. These elements will be briefly described and examples mentioned, although Linderoth and Ralphs (2005) is indicated for further detail.

Lower Bounding Methods

Many methods have been proposed to improve the lower bounds while solving the subproblems, such as logical pre-processing, addition of valid inequalities, reduced cost tightening and column generation.

Logical pre-processing consists of improvements to the problem's constraint system. Some of the applied routines can include: identifying infeasible instances, removing redundant constraints, tightening the bounds on variables by constraint analysis, improving the matrix of coefficients, and improving the value limitations on the constraints. Whereas in logical pre-processing the existing constraints are analysed and improved, in valid inequalities techniques new constraints are dynamically generated as a supplement to the existing ones. When valid inequalities are generated for all instances of computation with the aim to improve the bounds, the method is called branch and cut. Alternatively, cut and branch is the method in which valid inequalities are only implemented in the root node. In reduced cost tightening, the bounds to the integer variables can be tightened using the reduced cost of non-basic integer variables. As for column generation, it is considered as a dynamic generation of valid inequalities. When this routine is applied in every instance of the branch and bound process then it is referred as to branch and price.

Upper Bounding Methods

Upper bounds are found by obtaining feasible solutions to a minimization MILP, these are always the result of any branch and bound, however, some routines may be used to accelerate this process. On the one hand, to aid in acceleration one can influence the search order by preferring to compute and branch on the nodes which are closer to integer feasibility. On the other hand, the strategy might fall on the construction of a solution via a primal heuristic.

Branching Methods

Branching is the method by which the solver will partition each node into new nodes that are the subproblems.

The goal in branching is to improve as much as possible the lower bounds, however, many approaches have been created to estimate the contribution of the improvement in lower bounds. There two main groups dividing forward-looking and backwards-looking methods. Firstly, forward-looking approaches use information from current instances to evaluate candidate subproblems. Methods just like the penalty method of Driebeek and Strong Branching can be included in this cluster. As for backward-looking methods, the information obtained in previous partitions are considered along side locally computed information for the prediction of lower bounds effects through the next partition. In order to keep history of data, these methods recur to pseudocost.

Finally, using the estimated lower bounds improvements, the partitioning variable can be chosen taking into consideration the max of the sum of the improvements in all the branches, the max of the smallest improved branch or a hybrid approach.

Search Strategies

A search strategy serves to decide the order in which the subproblems will be computed. These node selection methods can be classified as: static, estimate-based, two-phase or hybrid methods.

Static methods imply that one fixed rule is applied all throughout. In a best-first approach, the first solved subproblem is always the one with the smallest lower bound. In contrast, in a depth-first search the fixed rule is of choosing the node with maximum depth in the tree. In estimate-based approaches the evaluation of the nodes is more complex. For example, in best-projection method the node is evaluated by combination of lower bound and integer feasibility, where as in the best-estimate method goes even further to additionally take into consideration the pseudocost information to rank the nodes for exploration. In a two-phase approach, the criteria to order the subproblems changes from one phase to the other. Most commonly, an initialization is made using a depth-first approach and after the first feasible solution is found, the exploration is made via a best-first strategy. Finally, in hybrid methods an interchangeable array of methods is applied given an all-round threshold to switch between strategies.

4.2.2. Non-commercial Software

There is a wide array of both commercial and non-commercial software packages to solve MILPs. Although commercial solutions have been documented to be faster than open source solvers, authors such as Meindl and Templ (2012) have tested the capability of free solvers applied to smaller problems.

In this Section, the focus will be on non-commercial alternatives, since there is no intention of investing in a paid solution for a proof of concept exercise. Table 3 presents a compilation of some of the most explored and tested open software packages in the light of the features they offer.

The choice between solvers should not only rely on the characteristics and possibilities of each one but also on the computational differences. Research has been conducted in this field putting solvers up against each other in trying to solve the same problem.

In the work of Linderoth and Ralphs (2005), SYMPHONY was highlighted as the solver that found the largest percentage of good feasible solutions and GLPK as an honourable runner up. Meindl and Templ (2012), compared both commercial and non-commercial softwares. The commercial ones were found to be the most efficient approaches, while GLPK was the best performing open source solver which was emphasised as a good free alternative for smaller complexity problems.

Among the packages available in R, the lp_solve, via the lpSolveAPI, and the GLPK are the most well documented. This presents upfront an advantage given that the tool must accommodate its less experienced users. Thus, GLPK was the select software since it was exalted on both mentioned bodies of work.

Table 3 Overview of non-commercial software

Features/Software	ABACUS	SYMPONHY	BCP	BonsaiG	CBC	GLPK	Lp_Solve	MINTO
Callable Library	X	√	X	X	√	√	√	√
Lower Bounding Methods:								
Logical Pre-processing	X	√	√	X	√	X	X	√
Additional Valid Inequalities	X	√	X	X	√	X	X	√
Reduced Cost Tightening	X	X	X	√	√	X	X	X
Column Generation	X	√	X	X	X	X	X	√
Branching Methods:								
Forward-Looking	X	X	X	X	X	√	√	√
Backward-Looking	X	√	X	√	X	X	X	√
Search Strategies:								
Static	√	X	X	X	X	√	√	√
Estimate-Based	X	X	X	X	X	√	X	√
Two-phase	√	X	√	X	√	X	X	X
Hybrid	X	√	X	√	X	X	√	√

5. Experimental Results

Given the selection of the approach to implement and its adjustments to fit the company's goals and business constraints, the next step is to gather, analyse, and treat the available data. Therefore, in this section an analysis of the original assortment will be presented as well as an overview of the required data engineering to extract costumer preferences from the transactional log. The base for this analysis is the sales of all SKUs in the rice subcategory during the first semester of 2018.

5.1 Data Analysis

As presented the proposed assortment problem has two variables contributing towards the revenue: sales and stockout. Therefore, the original assortment composed of 84 SKUs will be explored in these two dimensions. Additionally, the costumer preferences will also be studied since it also comprehends a strong role in the optimization problem.

Sales

Firstly, analysing the sales it can be seen that the demand is focused in a small part of the assortment. In Figure 1, it can be seen that 27% of the products in the assortment are responsible for more than 80% of all the sales. The figure also showcases the prevalence of a long tail in the assortment.

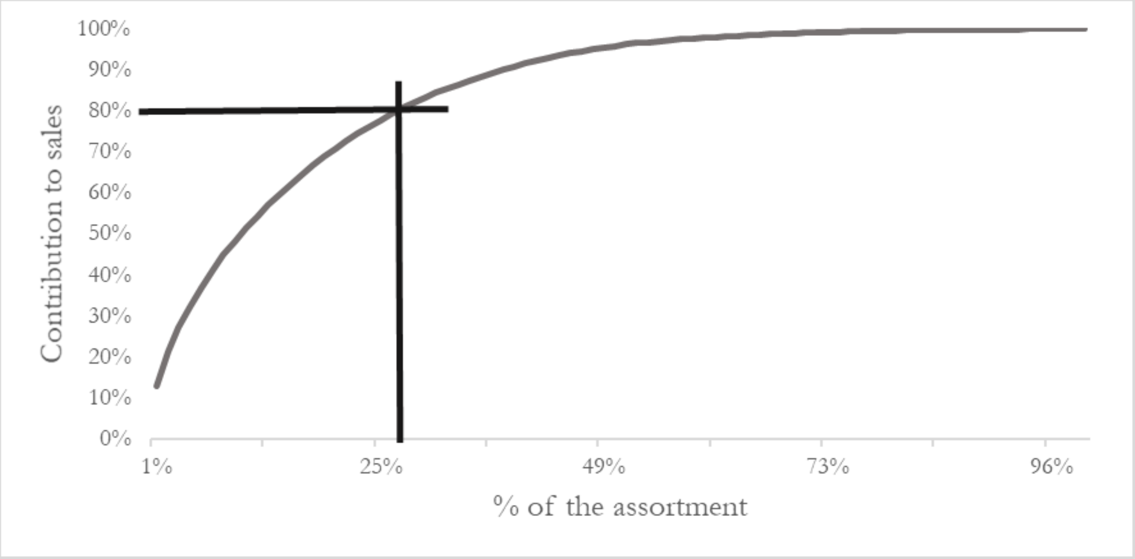

Figure 1 Contribution towards Sales

StockOut

Stockout, as mentioned, represents the part of the demand that was not satisfied, this means some of the requested quantities were not delivered to the customers. In Figure 2, it can be seen that only a quarter of the SKUs in the assortment had all its demand delivered. The SKU with the highest level of stockout only satisfied 80% of its demand.

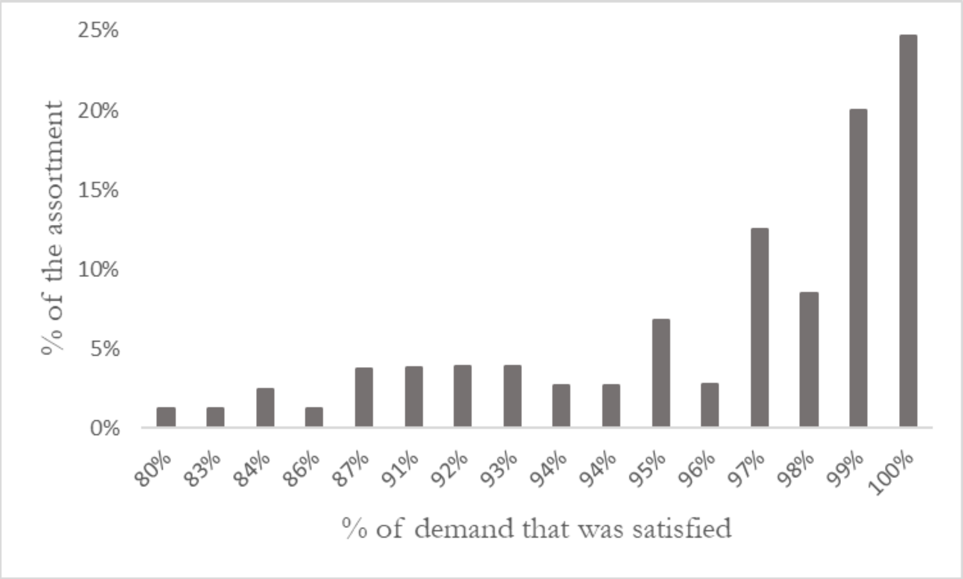

Figure 2 Distribution of satisfied demand

This unsatisfied part of demand can be evaluated in two perspectives, the value, in monetary units, that was lost and the net stockout, percentage of orders in which the stockout occurred. The value loss is a very valuable metric to the Retailer. Nevertheless, in the customer's perspective, the net stockout is more impactful, since it represents the percentage of times the customer ordered a specific SKU and it wasn't delivered. This metric doesn't take into consideration the quantities, only the number of occurrences. Consequently, a stockout of 10 units will have the same weight as a stockout of 1 unit. In fact, for a customer the most impactful incidence is the absence of the SKU, no matter the quantities missing.

Analysing the net stockout throughout the original assortment we can see that there is a high concentration of stockout occurrences in a small group of SKUs. In Figure 3, we can see that 80% of the net stockout is aggregated in 35% of the SKUs in the assortment. However, the slope of the graph is not as sharp as the one in the sales contributions.

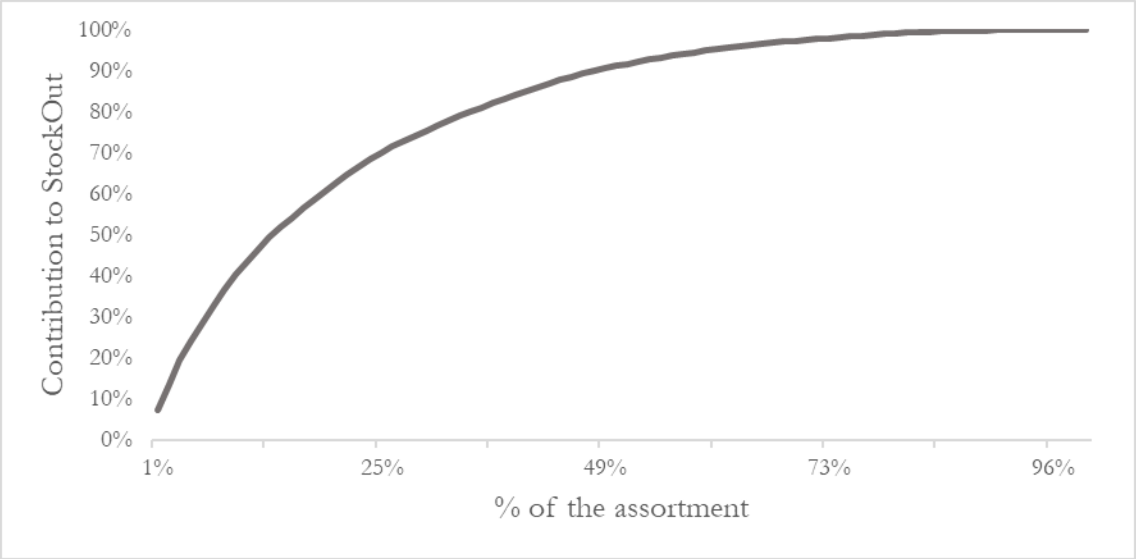

Figure 3 Contribution to Stockout

Costumers

Costumers are always in the forefront of any company's strategy and this case is not an exception. The two revenue factors are direct consequences of customer actions, the sales are a result of demand and stockout impact costumers' satisfaction and can, consequently, impact future sales. Therefore, it is also of interest to analyse how stockout has impacted customers in the past as well as the overall relation between customers and the assortment.

Firstly, in trying to assess the loyalty of costumers towards its favourite product, Figure 4 show that about half of the costumers that bought rice during the semester in analysis only bought one specific SKU. It should be recalled that the assortment is composed of SKUs with very similar characteristic and interchangeable. It can also be seen that even the costumers that bought multiple SKUs only tried out a small portion of the offered assortment. In fact, the costumer who bought more different SKUs only experienced 20% of the original assortment.

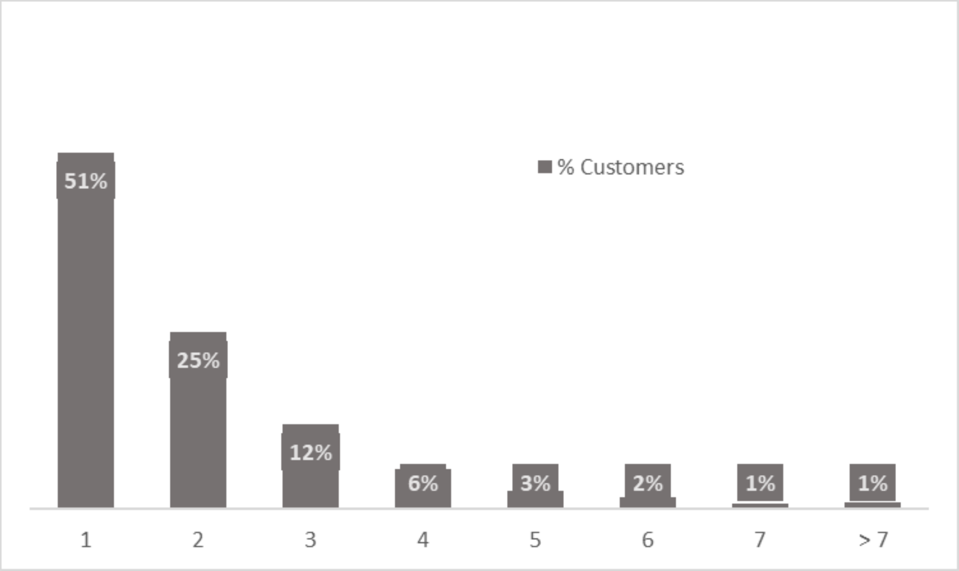

Figure 4 Distribution of Costumers Regarding the Different Number of SKUs Bought

Regarding the customers' behaviour the available discounts can also be an affecting factor, however the behaviour shown Figure 4 considers the sales for a period in which all products went through different promotional stages. Consequently, the impact of discounts is intrinsically being taken into consideration.

Secondly, in terms of the costumers' relations with stockout it can be learned from the transactional log that a high percentage of costumers when faced with a stockout stop being customers, or at least stop buying rice via the online route. Additionally, some customers came back and pursue the SKU that was unavailable, while as others change their demand to other offered SKUs.

Thus, customers of rice have well defined preferences, which is positive since the approach relies on an estimate of the consumer behaviour. Furthermore, some flexibility has also been diagnosed which indicates that it is likely that the costumers migrate demand towards a substitute SKU. In Figure 5, the rice customers, that experienced stockouts, are divided into groups depending on their behavioural response to the stockout.

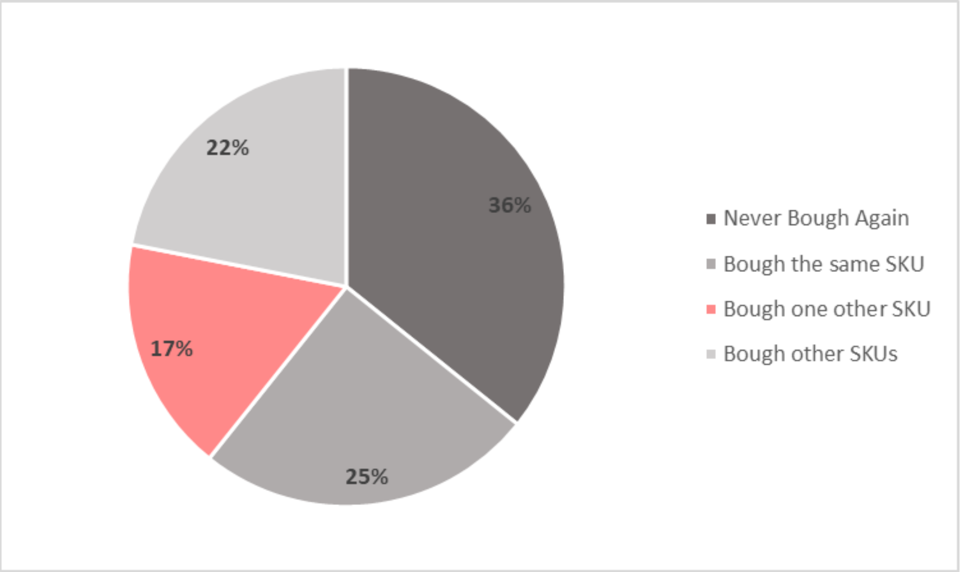

Figure 5 Costumers's Behaviour After StockOut

Full Overview

In summary, in this assortment 27% of the SKUs are accountable for 80% of the sales and 35% of the SKUs are responsible for 80% of the stockout. However, the list of top sellers differs from the list of top stocked-out SKUs.

On one hand, assortment decisions cannot be made through a simple rule, such as the approach of Broniarczyk et al. (1998), which would translate into eliminating the SKUs that simultaneously have a large contribution to stockout and a small one to sales. In fact, there is still around 20% of the products in the original assortment that are both top sellers and have high levels of strockout.

On the other hand, it is clear that the assortment does include a considerable percentage of SKUs that have small demand and, even though are also not contributing for stockout, they contribute for a higher operational complexity that can affect the efficiency and effectiveness in picking the other SKUs.

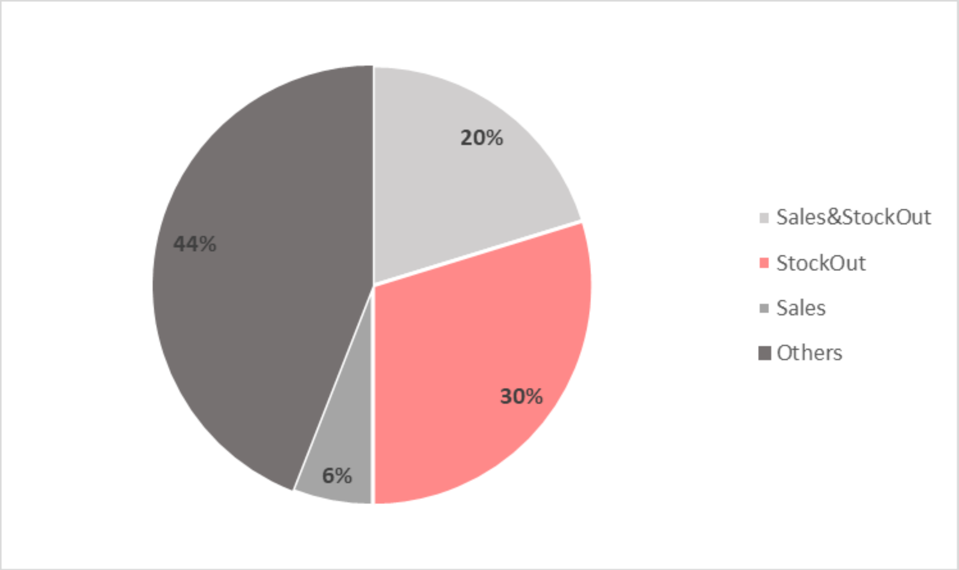

Figure 6 Assortment Overview: division of SKUs depending on their contributions towards sales and stockout

5.2 Data Engineering

The available data, as previously mentioned, is the transactional log. In this section, we present the process from which the required inputs, for the optimization model, were extracted and learnt from that data set. Viewing n as the number of products, b as the number of brands and k as the number of customer type in the original assortment. As presented in the Expression of (4.11) the problem at hand has as required inputs:

- R: a row vector of length n in which r_i records the price of SKU i;
- S: a row vector of length n in which S_i represents the penalisation of the contribution of SKU i to the revenue due to its stockout level;
- λ: a column vector of length k in which λ^k represents the probability of a costumer being of type k;
- P: a matrix of n rows by k columns where each y_i^k takes the value of 1 if the SKU i is the customer type k's most preferred product and 0 otherwise;
- H: a matrix of n rows by b columns where each z_i^b is 1 if the SKU i is of brand b and 0 otherwise.

The data engineering phase takes upon extracting these vectors and matrixes from the transactional log.

R: Price

The price is a straightforward characteristic since it is recorded in the transactional log, however, one specific SKU might have different recorded prices in the course of the semester of historical sales. The fundamental price of products is quite static nevertheless the presence of discounts affects the price effectively paid by costumers.

On the one hand, a possible approach to define the price of each SKU could be considering the last recorded price without the effect of a promotion. This value would be approximately precise since, as has already been mention, the intrinsic price of products is rather constant.

On the other hand, not taking into consideration the discounts would not be a correct quantification of the contribution of each SKU to the overall sales since the base price of products is not the value incoming from the sales. Therefore, the various recorded prices are used in the analysis and the price of the SKUs is computed via an average of all the recorded final unit prices in the time period.

Notwithstanding, an argument that price is a crucial factor in customer preferences could be used against this calculation of price. However, as it will be further explored latter in this section, the customer preferences and probabilities of being of a specific type will also be computed as a whole for the complete time period. Therefore, the possible impact of price, via discounts, in the preferences will be in still incorporated in the model.

S: Stockout

As previously mentioned, the retailer's revenue is also affected by the level of Stockout. In fact, the impact of a stockout occurrence is sometimes only associated with that particular sale but the effect it has on the costumers' satisfaction affectes future sales as well. Accordingly, stockout has a direct influence in the Retailer's revenue. The stockout must be used as a penalization of the price, since it can implicate that the customer does not actually pay for the ordered product.

Stockout occurrences are evaluated by the net stockout which is the percentage of orders where the SKU was not delivered, due to stockout, divided by the total of orders which the SKU was requested. This metric allows for a better representation of stockout without interference of the dimension of demand which varies from SKU to SKU. In fact, if one SKU is ordered more times than another, it is natural that it has more stockout occurrences. By using a percentage, it removes the demand impact, which is already considered in the consumer's choice model.

In a scenario without stockout, the revenue contribution of each SKU would be its estimated demand as the result of the consumer's choice model and its price. In our model, the variable *Si* is introduced to contribute negatively to the revenue. The SKUs with higher levels of stockout are penalised and the rest are rewarded.

Considering ns_i as the net stockout for SKU *i*, which, as previously mentioned, is the percentage of orders in which the SKU *i* had a stockout occurrence. *Si* will be computed given the following Expression:

$$S_i = \overline{ns} - ns_i \qquad (5.1)$$

λ: Probability of Customer's types

Previously to computing the probabilities of the customer types these must be defined and the rice buying customers must be labelled with the corresponding type.

Many of the reviewed papers made clusters with the customers that bought the products in the studied assortment and computed the probabilities by dividing the number of customers in each group by the total number of customers that entered the store in the time period. This presented a very effective way of contemplating the non-purchase alternative. With regards to the retailer for which the study is being made, it has a very extensive and diversified assortment. It sells almost everything from rice, the subcategory of assortment in analysis which is a reoccurring item in almost every household's shopping list, to sofas, a product seldomly sold. This implies that the retailer attracts a wide range of diverse costumers. Therefore, computing the probabilities of customer types using as denominator all the costumers of the retailer who did not buy rice would result in very small probabilities for all costumer types except the non-purchasing ones.

Therefore, customer types were derived only from the sales of rice. For each customer who bought rice during the time period under analysis a rank of preferences was made using the amount, in units, bought of each of the rice SKUs.

Note that, in many past papers on the subject, the assortments pertained to products with less frequent purchases each customer type would have typically bought only once and one SKU from the assortment and therefore that would be their most preferred one. In the present case, since a rice purchase is very frequent, some costumers have bought more than one rice SKU in the time period under analysis. Hence the need to find a strategy to identify each costumer's most preferred SKU. Accordingly, for each costumer the SKUs were ordered by the sum of ordered quantities and ranked from most bought to least bought.

Furthermore, a customer type was created for each SKU and each customer was assigned to the type *k* corresponding to its most preferred SKU. Although this process allowed all customers to be sorted into a type, some of the initial customer types ended up not having any costumers in it. As a result, those customer types were eliminated and, in the end, only 81 customer types were considered out of the 84 initially defined.

Next in order, to calculate the probability of a costumer being of one of these 81 types, the numerator used was the total number of purchases made by customers labelled with each type and the denominator was the total number of transactions of rice.

In sum, the preferences of each customer are a result of the quantities bought and the probabilities of the customer types are based on the frequency of purchase. Thus, λ^k is best described has the probability of a customer buying rice being of type *k*.

Y: Matrix of Preferences

Next in order is the computation of the matrix of preferences which will indicate for each costumer type which is the most preferred SKU.

Utilizing the customer types and respective ranks calculated as presented above, the matrix can easily be computed. For each combination of SKU and customer type the matrix takes the value of 1 when the SKU is the first ranked for that customer type and 0 otherwise.

B: Matrix of Brands

Lastly, the matrix of brands must be computed since it will be necessary for the constraints. In the original assortment the 84 SKUs correspondent to 18 different brands. The matrix of brands will be, therefore, a matrix with 18 rows and 84 columns populated by 1 if the SKU is of the corresponding brand and 0 otherwise.

5.3 Results

After the implementation of the problem presented in Section 4.1 with the data obtained from the transactional log as described in Section 5.3 in R using the Rglpk package, the solver was run recursively to gather all the output solutions. In this section, the best results obtained are showed and evaluated.

The resulting assortment keeps 67% of the original assortment. Before any other advantages are studied, the cost reduction and operational efficiencies resulting from this reduction are already positive outcomes. Furthermore, the resulting assortment will be tested among the transactional log of the semester following the period of analysis since if implemented that would be the period to which the assortment reduction would be applied.

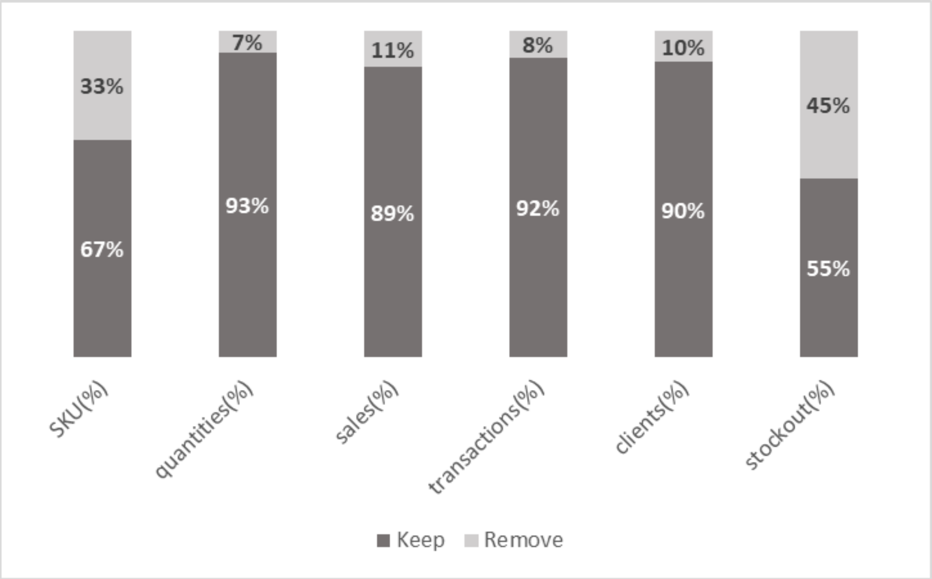

Figure 7 Analysis of the Resulting Assortment

Figure 7 shows that the SKUs which compose the resulting assortment, were responsible for 89% of the sales in the following semester, even though they only represented 67% of the assortment. Additionally, the contribution of those SKUs to the stockout occurrences was 55%, which is less expressive than its weight in the assortment.

Furthermore, the main factor in the assortment decision is the customer. As previously mentioned, assortment changes can have a great impact in costumer perceptions and their future purchase decision and behaviour. As presented by Broniarczyk et al. (1998), the highest impact is for customers whose preferred SKUs are removed. Thus, here and hereafter, affected customers are going to be considered the consumers who bought SKUs which are not included in the resulting assortment.

In Figure 8 an overview of the impact of the assortment reduction in the customers is present, based in the transactional log of the second semester of 2018. Only 10% of all the rice consumers bought at least one the SKUs that would be removed. Among these the percentage of costumers that did not buy any other SKU was 26%, which represents 4% of all the rice customers. This can be considered as the percentage of customers that would be lost due to the reduction in the assortment.

Moreover, 74% of the affected customers, additionally to buying removed SKUs, also bought SKUs that were included in the resulting assortment. This may lead to the conclusion that the removed SKUs are easily substituted by other SKUs that are kept in the assortment. Furthermore, these small impacts for the costumers are also a validation of the consumer's choice model.

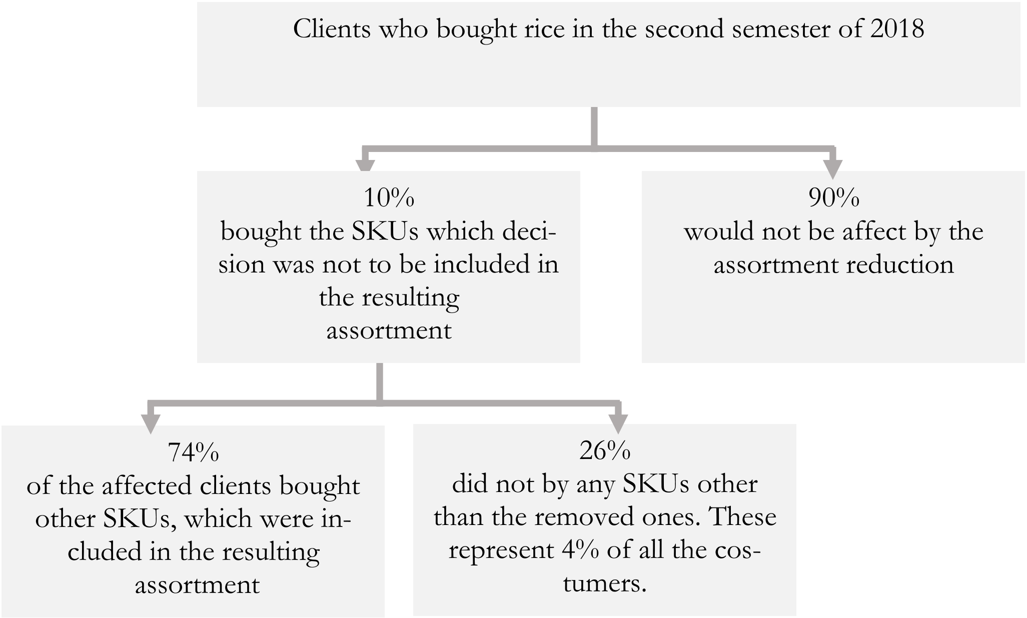
Clients who bought rice in the second semester of 2018
10%
bought the SKUs which decision was not to be included in the resulting assortment
90%
would not be affect by the assortment reduction
74%
of the affected clients bought other SKUs, which were included in the resulting assortment
26%
did not by any SKUs other than the removed ones. These represent 4% of all the costumers.

Figure 8 Assortment reduction impact on the costumers

The accuracy of the model also depends on the possibility of replicating these for a different subcategory. Therefore, the model was applied to the second biggest assortment, the sugar subcategory, as shown in Table 1. For reference, the original sugar assortment was characterised by 54 SKUs and 8 Brands.

The resulting assortment keeps 76% of the original sugar assortment, which represents a smaller reduction in comparison with the reduction in the rice subcategory. In Figure 9 the results for the sugar assortment optimization are displayed. The assortment reduction shows a positive impact in the assortment since the kept percentage of SKUs is lower than the kept percentage of demand, when evaluated in terms of quantities, price, orders and clients. Additionally, the contribution towards stockout of the kept SKUs is smaller in comparison with the weight of the kept SKUs in the original assortment.

In terms of the impact in the consumers, the percentage of affected clients was smaller than the one for clients affected by the rice assortment reduction. However, for the clients who bought, in the second semester of 2018, SKUs which would not be included in the assortment, the percentage of clients who were loyal to just one SKU is 35%. This represents an increase in 9 percentage points from the affected percentage in the rice analysis. Accordingly, it can be concluded that the sugar consumers are more sensitive to the offered assortment, which can also explain the smaller percentage of SKUs excluded from the original assortment. Once again, the accuracy of the consumers choice model translates into the results.

Therefore, it can be concluded that the proposed approach is applicable for other subcategories and could be applied for the full Retailer's assortment, but always with implementations subcategory by subcategory.

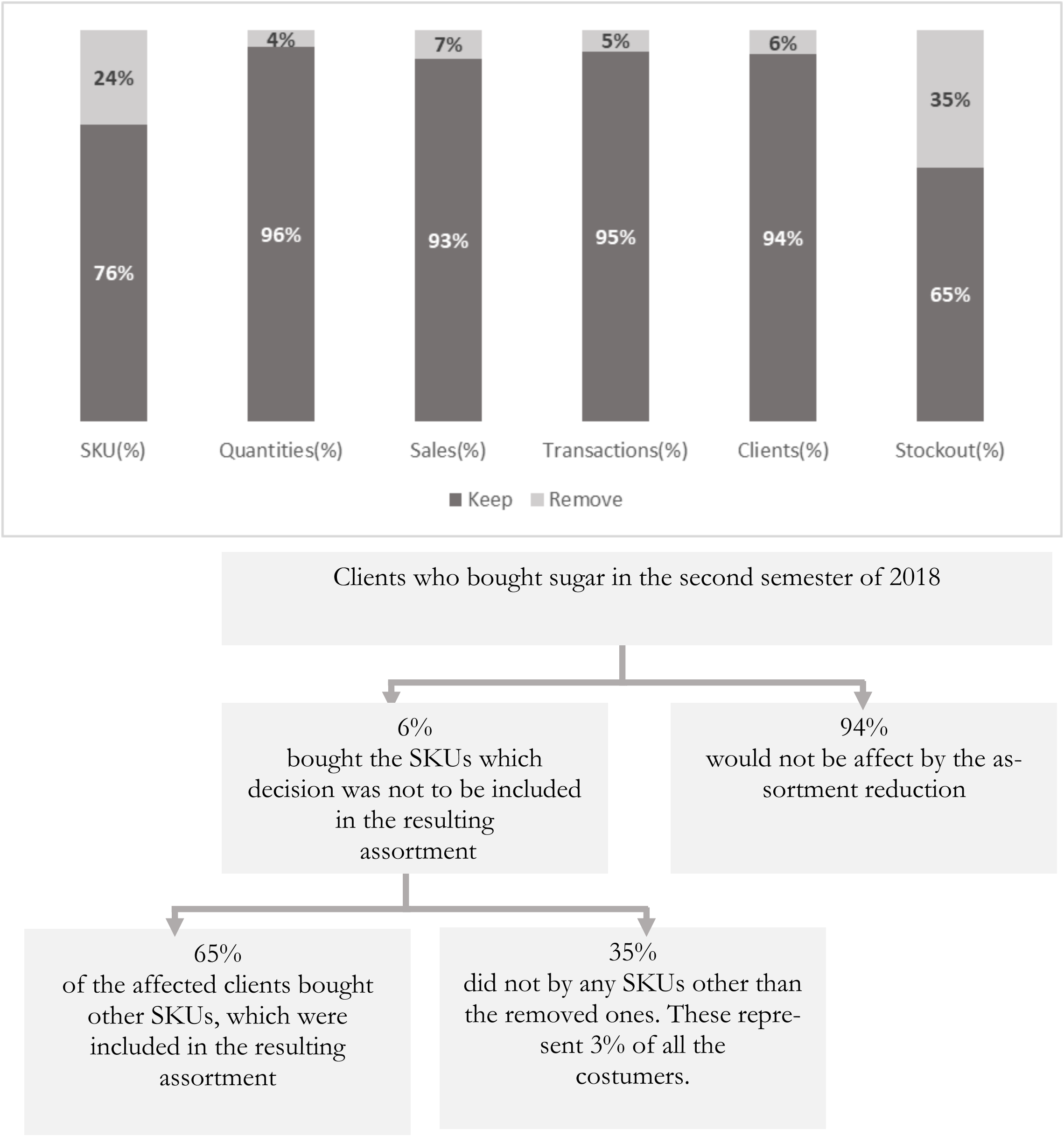

Figure 9 Results of implementing the approach on a different subcategory

6. Conclusion

Assortment Optimization is increasingly becoming a concern for many Retailers given the ever-increasing number of new products created. Therefore, there is the need to review assortments not only to include new products, but also to remove some of the current ones. This removal has two main drivers. On the one hand, it would be unsustainable for the assortment size to keep increasing due to the arrival of new products, thus the removal of older ones is needed to make spare room for them. On the other hand, the sales of same of the old products may not justify the operational complexity and costs of carrying them. In this work, an Assortment Optimization approach based on costumers' preferences was presented and applied directly from the sales transactional log.

Estimating the costumer' behaviour is a key element for the overall success of the model, the chosen approach fell, therefore, on a ranking-based approach. This approach is less prone to overfitting the data and does not require any apriori estimated data distribution. The only assumption is the rationality behind the consumer's choice, each costumer is believed to buy its most preferred SKU.

The basis for the work was the model of Bertsimas and Mišic (2015). However, this model only considered the SKUs price in estimating the revenue. Note that, the Retailer was not only focused on the prices, but also on the levels of stockouts. Therefore, we present an extension of the model by also considering stockouts in the revenue calculation.

Furthermore, the Retailer decisions are not completely autonomous, since it incorporates a larger business group. Thus, there are external conditions that must be followed. Specifically, although the number of products per brand might be reduced, the number of brands present in the assortment must remain the same. The model also had to be adapted to accommodate this constraint, which constitutes a further extension to the model.

Regarding the solver, the one implemented was Glpk and the choice was made taking into consideration computational efficiency, offered features and user interface. This may not be the most suitable software for problems of higher complexity. However, our model should always be implemented at a subcategory level because it relies on assortment comparison. In fact, the consumer choice model relies on rankings over the SKUs. In result, the proposed approach is sufficient to solve the proposed optimization problem, note also that the chosen subcategory for the test implementation represents one of the biggest subcategories in terms of assortment for the Retailer.

After solving the resulting binary linear programming model, the obtained assortment was tested, taking into consideration the costumers' behaviour as translated in the transactional log, on the following period. In this analysis, it was found that 89% of the sales would be assured by the assortment that was to be kept, which represents 67% of the original assortment. In terms of stockout the resulting assortment only represented 55% of the overall stockout occurrences.

However, the impact of the optimization in the sales does not consider that substitution effect, in fact, the sales might actually be higher because some of the costumers who bought the SKUs removed from the assortment can migrate their demand to the kept assortment. In fact, it was showed that 74% of the costumers that bought the removed products had also purchased at least one other rice SKU also kept in the assortment. Moreover, the impact of stockout reduction is actually larger, since reducing the assortment will result in operational improvements that may lead to reductions on the stockout levels experienced in the products kept in the assortment, such as faster picking time, less picking errors, among others.

The applicability of the proposed model to a different subcategory was also tested. The approach was replicated for the sugar subcategory, where very positive results were also found.

The approach proposed has the advantage of being "less blind", since in addition to customer preferences, retrieved from past transactional logs, business decision makers can also provide input or even limitations, as was the case of the impossibility of eliminating brands from the assortment.

This approach is based on historical data and, consequently, does not provide a solution regarding new products. One possibility is to introduce all new products in the assortment for the costumers to start interacting with them and after a considerable period of time, with a mature demand, use the model to rearrange the assortment. Approaches have reported accommodating new products, in general the products are deconstructed according to their characteristics such as size, brand, and flavour, and then the demand for each individual characteristic is estimated. Accordingly, Assortment decisions are made regarding product characteristics, in line with product line design problems. Such approaches require multidisciplinary decision making between product creators and assortment deciders. However, that is not the case for the Retailer, which is a distributer working as a facilitator between producers and the final consumer. Thus, making it impossible to implement such approaches in dealing with new products.

7. References

Belloni, A., Freund, R., Selove, M., & Simester, D. (2008). Optimizing product line designs: Efficient methods and comparisons. Management Science, 54(9), 1544-1552.

Ben-Akiva, M., & Lerman, S. R. (1985) Discrete Choice Analysis: Theory and Application to Travel Demand.

Bertsimas, D., & Mišic, V. V. (2015). Data-driven assortment optimization. Submitted for publication.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. Journal of consumer research, 25(3), 187-217.

Blanchet, J., Gallego, G., & Goyal, V. (2016). A markov chain approximation to choice modeling. Operations Research, 64(4), 886-905.

Boatwright, P., & Nunes, J. C. (2001). Reducing assortment: An attribute-based approach. Journal of marketing, 65(3), 50-63.

Broniarczyk, S. M., Hoyer, W. D., & McAlister, L. (1998). Consumers' perceptions of the assortment offered in a grocery category: The impact of item reduction. Journal of marketing research, 166-176.

Bront, J. J. M., Méndez-Díaz, I., & Vulcano, G. (2009). A column generation algorithm for choice-based network revenue management. Operations Research, 57(3), 769-784.Bultez, A., & Naert, P. (1988). SH. ARP: Shelf allocation for retailers' profit. Marketing science, 7(3), 211-231.

Davis, J. M., Gallego, G., & Topaloglu, H. (2014). Assortment optimization under variants of the nested logit model. Operations Research, 62(2), 250-273.

Debreu, G. (1960). Individual choice behavior-A theoretical-analysis-Luce, rd.

Drolet, A. (2002). Inherent rule variability in consumer choice: Changing rules for change's sake. Journal of Consumer Research, 29(3), 293-305.

Farias, V. F., Jagabathula, S., & Shah, D. (2013). A nonparametric approach to modeling choice with limited data. Management science, 59(2), 305-322.

Feldman, J. B., & Topaloglu, H. (2017). Revenue management under the markov chain choice model. Operations Research, 65(5), 1322-1342.

Fitzsimmons, G. J., Greenleaf, E. A., & Lehmann, D. R. (1997, July). Consumer Satisfaction with Both Product and Decision: Implications for the Supply Chain, In Consumer Satisfaction, Dissatisfaction and Complaining Behavior Biannual Conference.

Green, P. E., & Krieger, A. M. (1985). Models and heuristics for product line selection. Marketing Science, 4(1), 1-19.

Grün, B., & Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. Journal of classification, 25(2), 225-247.

Hoch, S. J., Bradlow, E. T., & Wansink, B. (1999). The variety of an assortment. Marketing Science, 18(4), 527-546.

Jagabathula, S. (2014). Assortment optimization under general choice.

Linderoth, J. T., & Ralphs, T. K. (2005). Noncommercial software for mixed-integer linear programming. In Integer Programming (pp. 269-320). CRC Press.

Makhorin, A. GLPK 4.2, 2004 Available from http://www.gnu.org/software/glpk/glpk.html

Mantrala, M. K., Levy, M., Kahn, B. E., Fox, E. J., Gaidarev, P., Dankworth, B., & Shah, D. (2009). Why is assortment planning so difficult for retailers? A framework and research agenda. Journal of Retailing, 85(1), 71-83.

McAlister, L., & Pessemier, E. (1982). Variety seeking behavior: An interdisciplinary review. Journal of Consumer research, 9(3), 311-322.

Meindl, B., & Templ, M. (2012). Analysis of commercial and free and open source solvers for linear optimization problems. Eurostat and Statistics Netherlands within the project ESSnet on common tools and harmonised methodology for SDC in the ESS, 20.

Plackett, R. L. (1975). The analysis of permutations. Applied Statistics, 193-202.

Reuters (2008), “Office Depot Introduces ‘Office Depot Green’,” April 8, available at http://www.reuters.com

Rusmevichientong, P., Van Roy, B., & Glynn, P. W. (2006). A nonparametric approach to multiproduct pricing. Operations Research, 54(1), 82-98.

Talluri, K., & Van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. Management Science, 50(1), 15-33.

Van Herpen, E., & Pieters, R. (2002). The variety of an assortment: An extension to the attribute-based approach. Marketing Science, 21(3), 331-341.

Van Ryzin, G., & Vulcano, G. (2014). A market discovery algorithm to estimate a general class of nonparametric choice models. Management Science, 61(2), 281-300.

Verhoef, P. C., & Sloot, L. M. (2006). Out-of-stock: reactions, antecedents, management solutions, and a future perspective. In Retailing in the 21st Century (pp. 239-253). Springer, Berlin, Heidelberg.