HALO EFFECT IN PROMOTIONS – APPLICATION TO FOOD RETAIL

Bruno Daniel Castro Borges

Dissertation

Master in Modeling, Data Analytics and Decision Support System

Supervised by

Professor Doutor João Gama (a)
Liliana Bernardino (b)

(a) University of Porto, School of Economics and Management
(b) Internship Advisor

2019

Acknowledgements

"It is difficult to understand that the most important thing isn't the objectives but the effort one dedicates to conceiving them" (Marcelo Bielsa).

"Tough times don't last always. Your hard times are there to shape you and develop your character. It causes you to become more aware of life and you develop an attitude of gratitude. Don't lose hope because it gets better" (Amaka Imani Nkosazana).

It was a tremendously difficult journey. It has involved a lot of effort and dedication. I would like to take this opportunity to thank who was part of this.

I would like to thank all my teachers during the master's degree, specially the professor João Gama for his advice.

Special thanks to the whole team where I was inducted at SonaeMC, including Liliana Bernardino, for their support and availability.

To all my loyal friends that were always there for me.

Finally, and most importantly, I would like to thank all my family specially the three most important people in my life:

Thank you, mother. Thank you, father. Thank you, my brother.

Abstract

Nowadays, promotions are an important element for retailers to remain competitive. For business development and profitability, it is crucial to make the best of promotions. Therefore, measuring the real effectiveness of promotions becomes an extremely important task.

This dissertation aims to realize which products, when promoted, have the potential to leverage the sales of other products and, thereby, increase the store profit. This leverage effect is called halo effect.

Results show that more than 200 products have the potential to increase total store sales when promoted.

These results were achieved using two different methodologies. The first used was ego network analysis using Jaccard index to evaluate differences between the ego networks with and without promotions. The second methodology aimed to compare the strength of associations between products with and without promotions using association rules. The use of these methodologies, for this purpose, is innovative and thus contributes to enrichment of the literature in this field.

This dissertation thus makes it possible to understand which products have the most halo effect allowing the company not to invest in campaigns that only brings to the store, customers who exclusively buy the promotional products(s).

Keywords: *Data Mining, Market Basket Analysis, Association Rules, Ego-Networks, Promotions, Halo-Effect*

Table of Contents

List of Figures

List of Tables

List of Formulas

1. Introduction

Grocery retailing is a highly competitive industry. Small improvements in operational decisions may change the competitive balance permitting the chain to survive and prosper. On average, a supermarket's margin is about only 1% of net sales (Ma & Fildes, 2017).

Therefore, the need to establish goals and set new strategies has become fundamental, being retailers more aware to the importance of analytics in adding value to the business (Kotler, Armstrong, Saunders, & Wong, 1999).

Retailers realized that competitive advantage will no longer achieved by only using systems for purpose of inventory management. Competitive advantage will be achieved for those retailers who are able to extract the knowledge hidden in data generated by electronic purchase registration systems. These data can be used to optimize the retailer's marketing decisions (Brijs, Goethals, Swinnen, Vanhoof, & Wets, 2001).

Alongside the importance of data analysis, retailers are also alerted to the importance of customer satisfaction and therefore, adopt a customer-oriented strategy is essential to keep the existing customers or to attract new ones.

Consumer promotions emerge as being an important component of competitive dynamics in retail markets playing a relevant role in the retailer's profit being responsible for almost a quarter of the marketing budget of consumer product companies (Raghubir, Inman, & Grande, 2004).

1.1. Problem Description

Given the widespread use of retail promotions and the amount of money spent on them, promotions planning has the potential to make a significant difference in the retailer's profits (Ma & Fildes, 2017). Therefore, measure the real effectiveness of a promotion is crucial.

The hardest part of measuring effectiveness is determining a promotion's real impact on demand. Measuring the real impact requires powerful analytical techniques to make sense of huge streams of sales and causal data. The noisy nature of retail data complicates the task (Liebson, 2003).

Typically, when a promotion is made, 4 effects occur:

- Baseline – An estimate of what would be sold if a promotion had not occurred;
- Pull-forward – Sales decrease in the promoted product after the promotion;
- Cannibalization – This effect occurs when the promoted item´s sales increase but "eats up" the sales of another item;
- Halo-effect – The effect that occurs when promoting one item leads to an increase in sales of another item.

The study of the halo-effect is the purpose of this dissertation. The intention is to study the effect that a certain promotional action on a product or a set of products has on other product(s) and consequently on overall store sales.

Understanding this effect will allow company to know which products, when promoted, have a greater power of increment in global sales of the store allowing company to desinvest in some campaigns that only brings to the store customers that exclusively buy the promoted product(s).

1.2. The Company

SONAE is a multinational that controls a very diversified portfolio of business in retail, technology, financial services, and telecommunications being present in 90 countries. SONAE is constituted by SONAE MC (food retail), SONAE SR (non-food retail), SONAE RP (real state retail), SONAE FS (financial services), SONAE IM (investment management), SONAE SIERRA (shopping centers) and NOS (telecommunications).

SONAE MC works with a variety of brands and products: (i) Continente – hypermarkets; (ii) Continente Modelo (middle size) and Continente Bom dia (small one) - convenience supermarkets; (iii) Meu Super - proximity stores in franchising format; (iv) Bagga - coffee shops and restaurants; (v) Go Natural - healthy restaurants and healthy supermarkets; (vi) Make Note and Note! - bookshop/stationary; (vii) Well´s - health, wellness and optics; (vii) ZU - products and services for pets; and (ix) Dr-Well´s - dentistry and aesthetics medicine clinics.

In 2007, Modelo Continente Hipermercados, S.A, launched the customer card offering to the customers, discounts in Continente, Continente Modelo, Continente Bom dia and Continente Online. In 2017 cartão Contiente reached 3.7 million of active accounts. At this moment, cartão Continente has several partners where cartão Continente can be used such as: Continente stores, Wells´s, Note!, ZU, Bagga, Meu Super, Zippy, MO, Galp and Ibersol group (Pastas Caffeé, Milt, Pizza Hut, Pans & Company, Roulotte, Ò Kilo, KFC, SOL and Burger King).

2. State of the art

2.1. Data mining

Data mining is a set of techniques used to extract information from large databases in order to discover patterns and relationships in the data and it is useful in several areas such as: support decision, estimation, forecasting and prediction and it is becoming crucial to the companies to make important business decisions giving them a competitive edge (Ahmed, 2004).

The exponential growth in information and amount of data leads to a situation in which retailers are struggling to get the information and mainly the right information since the traditional tools and methods to process and analyze information can´t deal with the enormous amount of information produce nowadays. To deal with this, KDD appears (Knowledge discovery in databases). KDD major application areas are marketing and fraud detection. One of the most famous KDD techniques is extraction of association rules (Brijs, Swinnen, Vanhoof, & Wets, 1999).

2.2. Association Rules

Association Rules were introduced by (Agrawal, Imielinski, & Swami, 1993) when they were searching relationships between items purchased by customers during a visit to a supermarket.

Association rules are one of the data mining techniques and allow to discover if the presence of an item or a set of items in the basket imply the presence of other distinct item or set of items (Agrawal & Ramakrishnan, 1994).

Extraction of association rules is one of the most well-known problems in Big Data analysis where algorithms are used to identify in an automatic way relationship between items in a database.

Association rules represent combinations of items that occur with a certain frequency in a database. Discovering these associations can help for instance develop marketing strategies by understand which items are often purchased together.

These rules are calculated from data and have probabilistic nature (Gama et al., 2017).

From a database that stores products purchased by customers, an association rule is an expression A→B where A and B can be sets composed by one or more items. The product A is called the antecedent of the rule and B is the consequent. This rule is used to indicate that customers who buy product A tend to also buy product B.

2.2.1. Measures of Interestingness in Association Rules

In order to evaluate the quality of a rule, interest measures are used. These measures can be divided into two classes: objectives and subjective.

The objectives identify statistically the strength of a rule while in the subjective, a specialist opinion is taken into consideration to determine the strength of a rule. (Gonçalves, 2005)

2.2.1.1. Objective Measures of Interestingness

Objective measures of interestingness are statistic indexes used to select interesting rules among many rules that can be discovered. Support, Confidence, Lift and Yules Q are some examples of objective measures (Gonçalves, 2005).

These measures are defined in terms of the frequency counts tabulated in a 2×2 contingency table (Tan, Kumar, & Srivastava, 2002).

Support

The support of a rule (A → B) is the percentage of transactions that contains A and B (Lenca, Meyer, Vaillant, & Lallich, 2008).

$$\textbf{Support (A}\rightarrow\textbf{B)} = \frac{\textbf{Frequency of A and B}}{\textbf{Total of Transactions}}$$

Formula 2.1 - Support of the Association Rule (A→B)

Support is a symmetric measure that is support (A→B) = support (B→A).

According to (Gama et al., 2017) support can be absolute or relative. The absolute support of an itemset represents the number of transactions that the itemset appears in the dataset. The relative support represents the proportion of transactions that contain the itemset and it is calculated dividing the absolute support by the number of total transactions.

Confidence

Confidence is the percentage of transactions in which B was bought given that A was bought (Jiawei, Han; Kamber, Micheline; Pei, 2011).

$$\textbf{Confidence (A} \rightarrow \textbf{B)} = \frac{\boldsymbol{support\ (A \cup B)}}{\boldsymbol{support\ (A)}}$$

Formula 2.2 - Confidence of the Association Rule (A→B)

The numerator of the rule is the number of transactions that A and B occur together, and the denominator is the number of transactions in which item A occurs.

Confidence is an asymmetric measure that is confidence (A→B) ≠ confidence (B→A).

Lift

Another measure widely used in this type of problem to assess levels of association is the coefficient of interest or lift. This metric is defined by the ratio between the confidence and the expected value of the confidence. The expected value of the confidence is the number of transactions that include the consequent divided by the total of transaction (Gama et al., 2017).

$$\textbf{Lift (A}\rightarrow\textbf{B)} = \frac{\boldsymbol{Confidence\ A \rightarrow B}}{\boldsymbol{support\ (B)}} = \frac{\boldsymbol{support\ (A \cup B)}}{\boldsymbol{support\ (A) \times support\ (B)}}$$

Formula 2.3 - Lift of the Association Rule (A→B)

In other words, lift tells us the more often B becomes when A occurs. The variation is between zero and infinite and the greater the lift, more interesting the rule is because the greater is the dependency between the items. (Brin, Motwani, Ullman, & Tsur, 1997).

A lift greater than 1 means that the occurrence of A has a positive effect on the occurrence of B. A lift smaller than 1 indicates that the occurrence of A has a negative effect on the occurrence of B. A value of 1 shows that A has no effect on B.

Lift is a symmetric measure that is lift (A→B) = lift (B → A).

Yules Q

Yule's Q is a measure of association used to indicate the strength of relationship between dichotomous variables at nominal or higher level. It is a normalize variant of the odds ratio to fall between -1 and 1 (Tan, Kumar, & Srivastava, 2004).

The antecedent and consequent are statistically independent if this measure obtains a value of zero.

If Yule's Q is greater than 0, means that the antecedent and consequent are more likely to occur together than if they are independent.

Negative values suggest that the antecedent and consequent are less likely to occur together than if they were independent (Shaikh, McNicholas, Antonie, & Murphy, 2013).

Like lift and support, Yule's Q is a symmetric measure.

$$\textbf{Yule's Q (A}\rightarrow\textbf{B)} = \frac{P(A,B)*P(\overline{A},\overline{B}) - P(\overline{A},B)*P(A,\overline{B})}{P(A,B)*P(\overline{A},\overline{B}) + P(\overline{A},B)*P(A,\overline{B})}$$

Formula 2.4 - Yule's Q of the Association Rule (A→B)

2.2.1.2. Subjective measures of interestingness

Subjective measures depend on knowledge and interest of the user and according to (Silberschatz & Tuzhilin, 1996) there are two major reason why a pattern is interesting : (i) unexpectedness and (ii) actionability.

The first one means that if a pattern is surprising to the user, then it is interesting because contradict the expectations of the user.

The second one means that if the user can do something about the pattern and take advantage, it is certainly interesting (Silberschatz & Tuzhilin, 1995).

2.3. Ego-Networks

Ego network analysis is one of several approaches to Social Network analysis. Social Network analysis conceptual origins can be related to three different areas: (i) sociology; (ii) anthropology and (iii) role theory (Tichy, Tushman, Fombrun, & Tushman, 1979).

A social network can be defined as a set of social entities such as groups, organizations and people with a certain pattern of relationship between them. These networks are modeled using math graphs where vertices are the entities and the edges are the relations between them. (Gama et al., 2017)

Using SNA, it is possible to understand who the most important actors or vertices are, the most influential, and it is possible to detect the formation of communities.

An ego network is the network of alters (edges) that form around a particular actor or node. Each node in a whole network (social network) has an ego network (Crossley et al., 2015).

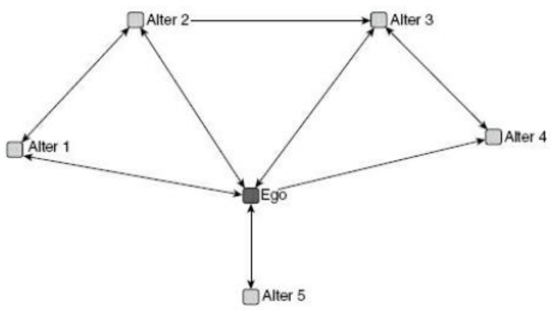

Figure 1 - Example of and Ego-Network (Crossley et al., 2015)

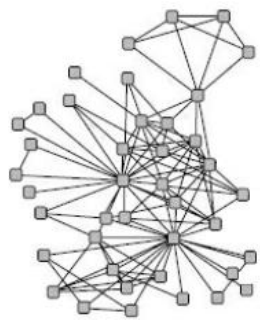

Figure 2 - Example of a Social (whole) Network (Crossley et al., 2015)

2.3.1. Comparing Two Ego-Networks

The Jaccard index compares members for two sets to see which members are shared. The range varies between 0 and 1. The higher the value, the more similar the two sets.

$$J(X,Y) = \frac{X \cap Y}{X \cup Y} = \frac{number\ of\ elements\ (alters)\ in\ both\ sets}{number\ of\ elements\ in\ either\ set}$$

Formula 2.5 - Jaccard index

2.4. Promotion Sales

Sales promotions are a fundamental strategic within the marketing value chain. They have multiple objectives such as: (i) correcting or increasing sales performance; (ii) attracting new shoppers; (iii) increasing average spend per visit; (iv) increasing purchase quantity; (v) increasing cross-category sales; (vi) seasonal demand smoothing; (vii) launching a new product or category; (viii) brand building/rewarding loyalty; (iv) tactical competitor response and (x) employee engagement (Ogden-barnes & Minahan, 2015).

A widely used promotion strategy is called the deep discount. This strategy consists in promoting some products at deep discounts trying to attract customers who maybe, will buy also non-promoted items once inside the store (Gauri, Ratchford, Pancras, & Talukdar, 2017).

Allegedly, sixty percent of the household supermarket purchases are not planned and the decision is the result of in store decisions (Drèze, Nisol, & Vilcassim, 2004).

Personalized Coupons, rebates, in-store temporary discounts, feature advertising, and in-store displays are some examples of promotions. When studying promotions it is possible to identify 3 generic effects: (i) immediate effects - the impact in the week w that promotion is realized; (ii) medium-term effects - the weeks surrounding week w and (iii) long-term effects - after the weeks corresponding to medium-effects (Harald Van Heerde & Neslin, 2017).

Nowadays, retailers are facing fierce competition hence profit margins are decreasing and expenses with advertising and promotions are increasing. To remain competitive, retailers must adopt an enterprise approach when planning, executing and analyzing their promotions. So, it is not surprising that in last years, retailers are constantly attempting to measure the effectiveness of their promotions (Liebson, 2003).

2.5. Decomposition of sales promotion effect

Promotions normally leads sales on the promoted item to increase. However, does not mean that this increase is actually beneficial. To know that it's crucial that managers acquire a method to decompose the promotions effects. (Harald Van Heerde & Leeflang, 2004)

Decomposition of sales promotion effect have been studied in the past years. There are two approaches to get a decomposition of sales-promotion effects: gross (elasticity based) and net (unit sales based). The gross decomposition explains the customer choice process detailed but, in order to assess the net losses in cross-brand sales and net growth in category sales, net decomposition is required. The main difference between these two approaches is the way that the net approach accounts for the increase in purchase incidence that benefits all brands in the category. If a brand is promoted, the category purchase incidence probability in the household models increases. Because this increase also applies to the non-promoted brands, it tends to compensate for the gross sales loss due to brand switching (Harald Van Heerde & Leeflang, 2004).

The effect of a promotion may result in an increase in primary demand or secondary demand. While primary demand refers to a demand for a whole product category, secondary demand refers to a demand for a particular brand of a product.
Secondary demand can only increase due to brand switching (cross-brand effects) and therefore retailer has no interest in changes in secondary demand.

Primary demand can increase due to multiple changes in consumer behavior: stockpiling, purchase incidence, store switching and category switching. Growth in primary demand is much more important for retailers than secondary demand.

The first major contribution was given by Sunil Gupta. In his article (Gupta, 1988) he introduces the gross approach and tries to answer two questions: (i) is the increase in sales due to consumers switching from other brands? and (ii) Are consumers advance their purchases or stockpile the product increasing the brand´s sales?

According to him, these questions can be answer by decomposing the sales bump during the promotion into sales increases due to stockpiling, brand switching and purchase time acceleration His study was a reference because it was the first considering all 3 dimensions of household´s decision: (i) purchase timing; (ii) brand choice; (iii) purchase quantity.

Gupta study focused on coffee category and concluded that of the total sales increase due to promotion, more than 84% was accounted for by brand switching, 14% by purchase time acceleration and 2% by stockpiling. For the retailers this is not a good outcome because brand switching is not benefic for them.

Using the same approach (elasticities based), (Chintagunta, 1993) formulated a utility framework for modeling the 3 components of a household´s purchase decision making process (category purchase, brand choice and purchase quantity). The study concluded that household does not make a category purchase whenever the reservation price is lower than the adjusted price of all brands in the product category.

Still elasticity based, (Bucklin, Gupta, & Siddarth, 1998) added another dimension: across segment dimension. Market-level sales elasticities were decomposed by segments and purchase behavior. They modeled the choice decision with multinomial logit, the incidence decision with nested logic and the quantity decision with poison regression. The study was made on yogurt category and five distinct response segments were revealed. They conclude that in the two largest segments, price had a great impact on choice but a small impact on quantity. Price discounts induce these households to shift brands and buy earlier but have a little effect on their stockpiling behavior. In the two smallest segments price had a smaller impact on choice but a greater impact on quantity. The last segment had a low overall level of price sensitivity.

After that, (Bell, Chiang, & Padmanabhan, 1999) proposed and elasticity based approach decomposing the total price elasticity for 173 brands across 13 different products category. They concluded that 25% of the elasticity was due to primary demand expansion and 75% to secondary demand effects (brand switching). These results are more interesting to the retailers than Gupta´s results.

While all the approaches mentioned above are elasticity based, a different approach, a net (unit sales based) approach was propped by (H Heerde, Gupta, & Wittink, 2003). The main goal of the study was to clarify if 75% of the sales promotion bump was due to brand switching or not. Several authors interpreted this 75% as follows: of a brand gains 100 units during a promotion and 75% of the sales elasticity is attributable to brand switching, other brands in category lose 75 units. But according to their study, this interpretation is wrong because the secondary demand component of the elasticity decomposition cannot be interpreted as the ratio of the loss in sales of competing brands to the benefit in sales of the promoted brand.

The procedure consisted in review and clarify the elasticity decomposition based on household data. Elasticity decomposition was transformed into unit sales effect decomposition. The study concluded that those 75% are equal to 33% when talking in net effect. So, if a promoted brand gains 100 units, the other brands together lose 33 units.

Another study was conducted by (Harald Van Heerde & Leeflang, 2004) and the objective was to answer the following questions: (i) If the promoted brand sells 100 more units, how many units do other brands loose? (ii) how many units come from other periods? (iii) how many units is due to category expansion?

Store-level scanner data was used and a decomposition of the own-brand sales effect into net cross-brand, cross-period, and category-expansion effects was made. Primary demand effects were split into cross-period and category-expansion effects.

Cross-period effects are the part of primary demand effects that represent temporal shifts in sales. Category expansion effects are the parts of primary demand effect that is not due to temporal shifts.

	Category-expansion effects (no temporal shifts)	**Cross-period effects (temporal shifts in sales)**
Purchase incidence effect	Cross-category Cross-store	Purchase time acceleration
Purchase quantity effect	Consumption increase	stockpiling

Table 1 - Splitting primary demand effects

Two additional decompositions were added: (i) cross-store and (ii) cannibalization effects, and cross-store effect was split from the category expansion effect. Cross-store effects are related to store switching, which has two forms, direct store switching and indirect store switching. The first one is related to outside store suggestions (e.g., featured price cuts) and if households use these suggestions to decide where to purchase specific products and leads sales in stores competing with the store promoting the product to decrease. The second one is when a household goes to several stores in a given week, and is influenced by inside store suggestions (e.g. displays with price promotions). (Harald Van Heerde & Leeflang, 2004)

Across 4 categories, results showed that on average if a price promoted brand gains 100 units, the net loss for other brands is 33 units (secondary demand effect is 1/3 and primary demand is 2/3).

The following table shows a summary of the studies cited above.

Author	**Year**	**Category**	**Approach**	**Conclusion**
Gupta	1988	1(Coffee)	Gross approach (elasticity based)	84% of the effect of a promotion is at secondary level (brand switching) and 16 % at primary level (14% purchase time acceleration and 2% by stockpiling).
Chintagunta	1993	1 (Yogurt)	Gross approach (elasticity based)	Household does not make a category purchase whenever the reservation price is lower than the adjusted price of all brands in the product category.
Bucklin, Gupta, & Siddarth	1998	1 (Yogurt)	Gross approach (elasticity based)	In the two largest segments price had a large impact on choice and incidence but a small impact on quantity. In the two smallest segments price had a smaller impact on choice and incidence but a greater impact on quantity. The last segment had a low overall level of price sensitivity.
Bell, Chiang & Wittink	1999	13 different products (173 brands)	Gross approach (elasticity based)	25% of the elasticity was due to primary demand expansion and 75% to secondary demand effects (brand switching).

H.Heerde, Gupta Wittink	2003	13 different products (173 brands)	Net approach (unit sales based)	If a promoted brand gains 100 units, the other brands together lose 33 units.
Harald Van Heerde & Leeflang	2004	4 (Tuna, Tissue, Shampoo, Peanut Butter	Net approach (unit sales based). Split cross- category from cross-store effect	If a promoted brand gains 100 units, the other brands together lose 33 units.

Table 2 - Summary of the related work in decomposition of promotion sales effects

2.6. Halo effect

Halo effect occurs when promoting one product has a positive effect on the sales of other products. This effect is important for retailers since their the main goal is to drive store traffic and overall sales (Liebson, 2003).

There are few studies that estimate halo rates. The first and the most important study was conducted by (Ailawadi, Harlam, César, & Trounce, 2007).

They studied a U.S drugstore chain called CSV, with over 4000 retail stores in 35 states. This was the first study to estimate the halo effect of the promotion. They defend that besides the incremental lift within the category, retailer must consider any halo effect that the promotion may have on sales of other categories in the store to determine the net unit impact of a promotion in the store. They defined store net unit impact as being:

Store net unit impact = Gross lift x (1-switching - % stockpiling + % Halo)

Formula 2.6 - Store net unit impact (Ailawadi et al., 2007).

In which, the gross lift for a promoted product in store s in week w is equal to the unit sales of that product during the promotional week minus its baseline unit sales in that week. They estimated the baseline as a moving average of the item´s unit sales on neighboring promotional weeks (Ailawadi et al., 2007).

Switching occurs if the gross lift is entirely due to switching from other products in the category. If for every unit increase in the gross lift of all promoted items in category c in store s in week w, there is a an analogous unit increase in total category units in the store on that week, there is no switching (Ailawadi et al., 2007).

The stockpiling is the percentage that is taken from future category sales in the store. Halo effect occurs if, for every unit increase in the gross lift from promoted products in a store s in a given week w, there is a alteration in total store units (Ailawadi et al., 2007).

Their results regarding to the halo effect was that for every unit of gross lift, 0.16 units of some other product is purchased elsewhere in the store.

Therefore, their study allows to understand the effect that an promotional action has on the various departments and on the global but does not allow to conclude which products are responsible for this increase.

3. Data

3.1. Data Collection

From the database where all transactions are registered, the first task was to extract and create the data table needed for the project. This was executed using SAS Enterprise Guide and Oracle SQL Developer.

3.2. Extraction of the Segmentation data

Through queries, information on the customer price sensitivity segmentation was extracted. This segment is updated monthly and it is dynamic, meaning that in a given month a certain customer can be in one level of the segment and then switch to another level, depending on the most recent purchase behavior. Only the most recent segmentation was considered - December 2018.

3.3. Data Description

In this section, the description of all variables, including segmentation data, will be presented.

The analysis focused in 4 commercial departments: (i) alimentar - DC 10; (ii) peixaria e talho - DC 11; (iii) Frutas e legumes, charcutaria, padaria e take away - DC 12 and (iv) nutrição saudável. It was considered all transactions in which loyalty card was used in 2018 at Continente, Continente Modelo, Continente Bom Dia and Continente Online. The following table summarizes all the variables.

Variable	Description
Time_Key	Transaction Date (year/month/day)
Transaction_ID	Transaction identifier
ID_Cliente	Customer identifier
VPH4_Key	Virtual Product Hierarchy type 4 identifier
VPH4_Dsc	Virtual Product Hierarchy type 4 description
Gross_sls_amt	Gross sales of the product in the transaction (€)
Desconto Cupão	Gross discount (€) on card associated to the product (e.g: 25% in Bananas)
Desconto Direto	Direct gross discount (€) associated to the product (e.g: 25% in Bananas)
Percentagem Desconto Cupão	Gross discount (%) on card associated to the product
Percentagem Desconto Direto	Direct gross discount (%) associated to the product (e.g: 25% in Bananas)
Promotion	Binary variable indicating if the item is bought on Promotion. 1-yes; 0-No
Segment_CD	Code of the customer's price sensitivity segmentation with 8 levels (1-8)
Segment_DSC	Description of the customer's price sensitivity segmentation (PSS_1 to PPS_8)
Semanas_Promo	Number of weeks in which VPH4_Key is on promotion

Table 3 - Descriptive summary of the variables

The product structure considered was VPH level 4. VPH is an acronym for Virtual Product hierarchy and it´s related with how the customer perceives the product. It´s important to distinguish between 2 types of structure: i) Market Structure - Store oriented and VPH - Customer oriented.

To provide a better understanding of these two structures, an example is given below.

Market Structure Type	Market Structure Description
Commercial Department	F&L, C&Q, Pad, Takeaway
Business Unit	Frutas e vegetais
Category	Frutas
Sub-Category	Bananas
Unit Base	Importada
SKU	Banana importada CNT

Table 4 - Market Structure

VPH_TYPE	VPH type description	VPH description
1	Category	Frutas
2	Category - Brand	Frutas da marca Continente
3	Product	Banana
4	Product - Brand	Frutas banana da marca continente

Table 5 - VPH structure

Thus, two distinct products in terms of market structure (2 different SKU's), banana importada CNT and banana CNT emb, are perceived by the customer as being the same product. Therefore, these two products will merge in one product - frutas banana da marca continente in terms of VPH4.

3.4. Data Transformation and Cleaning

In this stage, were performed tasks such as:

- Some observations that accumulated discounts and appeared as having 100% discount or more - inconsistent data - were not considered in the analysis.

- Observations having an associated discount if customer pays with caixa geral de depósitos debit card were not considered.
- Observations with an associated discount not related to the product were not considered.
- Therefore, it was considered that an item is purchased with promotion if it has a direct discount or coupon offer greater or equal to 10% and less or equal to 75%.
- A new attribute named promotion was created with values 0 or 1 indicated the absence or presence of promotions.

4. Description of the Methodologies used

In this study two main approaches were considered: (i) ego network analysis and (ii) association rules. In this chapter both approaches are described. Ego network analysis in 4.1 and association rules in 4.2.

For both approaches it was necessary to work with two different datasets, a dataset with no promotions and a dataset with promotions. To do that, the data table described in the previous section was split into 2 different datasets. The idea is as follows:

If a given transaction has no item purchased on promotion, then the transaction is allocated to the no promotion dataset. On the other hand, if the transaction has at least one item purchased on promotion, then the transaction is allocated to the promotion's dataset.

A small sample of the two datasets is provided below for better understanding.

Transaction_id	VPH4_Key	Promotion
(…)123	(…)100200	0
(…)123	(…)100300	0
(…)345	(…)100400	0

Table 6 – Sample of the dataset without promotions

Transaction_id	VPH4_Key	Promotion
(…)789	(…)300100	0
(…)789	(…)400100	1
(…)567	(…)500100	1

Table 7 – Sample of the dataset with promotions

As can be seen, in dataset with promotions, the transaction (…)789 has at least one item purchased on promotion therefore the transaction is allocated to the dataset with promotion

4.1. Ego-Network approach

The main idea behind this approach is to compare two ego networks for each item. The ego network when the item is on promotion and the ego network when it is not on promotion. Ego-net is the relation between items when they are purchased together in the same transaction represented as a network, an ego network in this case. If, for example banana is purchased with apple and orange in the same transaction and the banana ego network is being analyzed, banana would be the ego and apple and orange would be the alters. Both ego and alters are represented by a node and the relation between ego and alters is represented by an edge.

To evaluate the differences between the ego networks, the Jaccard index was calculated to check the similarity between each ego network. The closer to one the Jaccard is, more similar are the ego-nets and less halo-effect is expected to have the item in analysis because if the ego networks are similar, there are no differences when the product is promoted. On the other hand, the closer to zero more halo-effect is expected to have.

The software used to calculate this index was SAS enterprise Guide and Oracle SQL Developer.

The first stage was to calculate for each pair of items the number of times they were purchased together in the same transaction (or same basket). This was executed using self-join in SQL. Having the number of transactions of each pair of items with promotions and without promotions, the weighted Jaccard index was calculated for each product using the following formula:

$$\textbf{Weighted Jaccard (x,y)} = \frac{\sum_i \textbf{min}\,(x_i,y_i)}{\sum_i \textbf{max}\,(x_i,y_i)}$$

Formula 4.1 - Weighted Jaccard

It was considered the weighted Jaccard index instead of the conventional Jaccard index because this way, the number of transactions is taken into consideration.

For better understanding an example that illustrates what was made is presented below.

Item1	Item2	Trasactions without promotions	Transactions when item1 is promoted	Min	Max	Sum_min	Sum_max
Banana	Apple	30	15	15	30		
Banana	Orange	15	60	15	60	50	170
Banana	Pear	20	80	20	80		

Table 8 – Illustrative example of the Jaccard index application

Thus, the Jaccard index for banana would be: 50 / 170 = 0,294.

Not all associations were considered. There were some restrictions in both datasets. This process is described in 6 together with the results.

4.2. Association Rules approach

Typically, market basket analysis and associations rules are used to find out consumption patterns, to improve the layout of the store or still to promote certain items.

The approach used in this study is quite different. Here, association rules are used to evaluate the strength of the connections between items with promotions vs without promotions. Therefore, association rules were applied to both datasets to evaluate the differences considering lift and Yule's Q as association measures.

Initially, for each pair of items that occur in same basket, measures like (i) support (ii) confidence (iii) lift and (iv) Yule's Q were calculated.

The following table shows a 2x2 contingency table which was taken into consideration to calculate all metrics.

	B	$\overline{B}$	
A	n(AB)	n($A\overline{B}$)	n(A)
$\overline{A}$	n($\overline{A}$B)	n($\overline{AB}$)	n($\overline{A}$)
	n(B)	n($\overline{B}$)	N

Table 9 – 2x2 contingency table for A and B variables (Geng & Hamilton, 2006).

- n(AB), n(B), n(A) and N were calculated using Oracle SQL Developer.

- n($A\overline{B}$), n($\overline{A}$B) and n($\overline{AB}$) were calculated in SAS Enterprise Guide using the following formulas:

$$\mathbf{n(\boldsymbol{A\overline{B}}) = n(A) - n(AB)}$$

Formula 4.2 - n($A\overline{B}$) formula used based on a 2x2 contingency table

$$\mathbf{n(\boldsymbol{\overline{A}}B) = n(B) - n(AB)}$$

Formula 4.3 - n($\overline{A}$B) formula used based on a 2x2 contingency table

$$\mathbf{n(\boldsymbol{\overline{AB}}) = N - n(AB) - n(\boldsymbol{A\overline{B}}) - n(\boldsymbol{\overline{A}}B) - n(\boldsymbol{\overline{AB}})}$$

Formula 4.4 - n($\overline{AB}$) formula used based on a 2x2 contingency table

In which:

- n(AB) is the number of times that A and B occur together;
- n($A\overline{B}$) is the number of times that A occurs and B does not occur;
- n($\overline{A}$B) is the number of times that B occurs and A does not occur;
- n($\overline{AB}$), is the number of times that neither A nor B occur;
- n(A) and n(B) are the number of times A and B occurs respectively and N is the total number of transactions in the dataset.

Support, confidence, lift and Yule's Q were also calculated in SAS Enterprise Guide using the following formulas:

$$\textbf{Support (A→B) = n(AB) / N}$$

Formula 4.5 – Support formula based on a 2x2 contingency table

$$\textbf{Confidence (A→B) = n(AB) / n(A)}$$

Formula 4.6 - Confidence formula based on a 2x2 contingency table

$$\textbf{Lift (A→B) = n(AB) / (n(A) * n(B)) / N}$$

Formula 4.7- Lift formula used based on a 2x2 contingency table

$$\textbf{Yule's Q (A→B)} = n(AB)*n(\overline{AB}) - n(A\overline{B})*n(\overline{A}B) \; / \; n(AB)*n(\overline{AB}) + n(A\overline{B})*n(\overline{A}B)$$

Formula 4.8 – Yule's Q formula used based on a 2x2 contingency table

Observing the formulas above can be concluded that lift, support and Yules Q of a rule (A→B) is equal to the rule (B→A) as stated before. However, this is true only when we are talking about the dataset without promotions.

When the subject is the dataset with promotions the antecedent of the rule is am item on promotion. Therefore, a rule (A→B) corresponds to an association of a product A with product B when product A is purchased with discount and product B is purchased without discount. So, (AB) is the number of times that product A and product B are purchased together in the same transaction or basket when A has an associated discount and B does not have an associated discount. Following the same line of thinking, n(BA) is the number of times that product B and product A are purchased together in the same transaction when B has an associated discount and A does not have an associated discount, causing n(AB) different from n(BA).

Still talking about dataset with promotions, while the antecedent of the rule is always on promotion, the consequent is always without promotions and the reason that happens is because if the goal is to get an individual score of halo effect for each product, if the consequent were on promotions, the effect that would occur when antecedent is promoted could not be imputed to it because the hypothetical sales growth in the consequent would probably be due to promotions on it.

After all these calculations for both datasets the next step was to get a halo-effect score for each product. This stage is divided in two steps:

1. Get the weighted arithmetic of the lift and the weighted arithmetic of the Yule's Q for each item in both datasets using the following formula:

$$\overline{\boldsymbol{x}} = \frac{\sum_i^n \boldsymbol{wi} * \boldsymbol{xi}}{\sum_i^n \boldsymbol{wi}}$$

Formula 4.9 – Weighted arithmetic mean

The following table illustrates better what was done in point 1.

Item1	Item2	Lift	Co-Occurence
Banana	Orange	2,1	100
Banana	Apple	1,1	200
Banana	Pear	1,5	300

Table 10 – Illustrative example of the weighted arithmetic of the lift (banana)

Therefore, the weighted arithmetic mean of the lift (banana) would be:

$$\frac{2{,}1*100+1{,}1*200+1{,}5*300}{100+200+300} = 1{,}467$$

This way the co-occurrence (number of transactions between items) is used as weight. This weighted arithmetic mean was applied in both dataset for all items using lift and Yule's Q.

2. Get the variation (with and without promotions) of these two metrics in both sets getting this way a halo-effect score for each product.

Lift and Yule' s Q variation was calculated using the weighted average mean in both sets. As Yule's Q takes values between -1 and 1, the absolute variation was used.

In the case of the lift as it varies from 0 to infinite, the relative variation was used. Below, both formulas are described:

$$\textbf{Halo Effect Yule's Q} = \mathbf{A - B}$$

Formula 4.10 – Absolute Variation of the Yule's Q – with and without promotions

In which:

A = Weighted arithmetic mean Yule's Q with promotion

And

B = Weighted arithmetic mean Yule's Q without promotion

$$\textbf{Halo Effect Lift} = \frac{(\boldsymbol{A} - \boldsymbol{B})}{\boldsymbol{B}}$$

Formula 4.11 – Relative Variation of the lift – with and without promotions

In which:

A = Weighted arithmetic mean Lift with promotion

And

B = Weighted arithmetic mean Lift without promotion

There were some restrictions in both datasets, not all associations were considered as had already happened in the first methodology. This is detailed in 6 alongside the results.

5. Exploratory Analysis

In this section, an exploratory analysis of both datasets is presented with emphasis in two associations metrics: (i) lift and (ii) Yule's Q.

5.1. Dataset with no promotions

In this dataset, there are about 4 million transactions with 6298 different products being transacted. The average number of items per basket is 3,05 meaning that on average each transaction has 3 items in the basket.

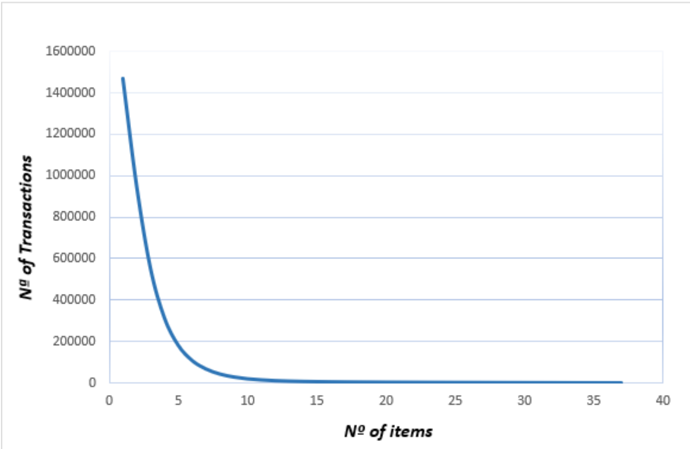

Figure 3 - Number of items vs Number of transactions – dataset without promotions

The figure 3 shows that more than 1.4 million of transactions have one item transacted, more than 900 thousand have two items transacted and more than 500 thousand have 3 items in the basket. It is possible to conclude that customers do not buy much products without promotions and they prefer to buy on promotion which is expected.

Using SAS and SQL all pairs items transacted in the same basket were obtained. Defining a minimum support and a minimum confidence it is possible to check how many rules are created. This is illustrated below, in figures 4 and 5.

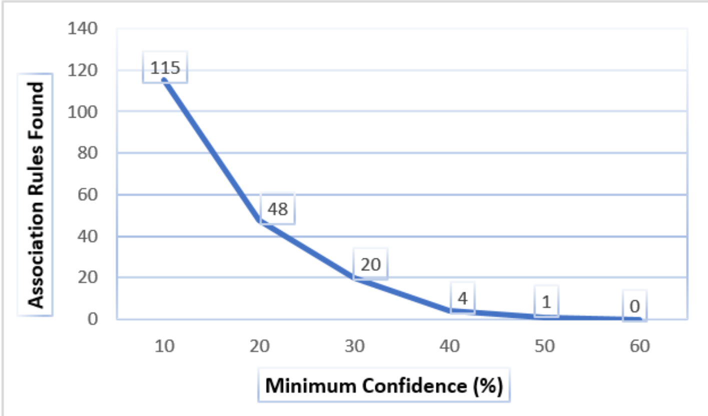

Figure 4 - Associations Found with support 0.1% and varying the confidence

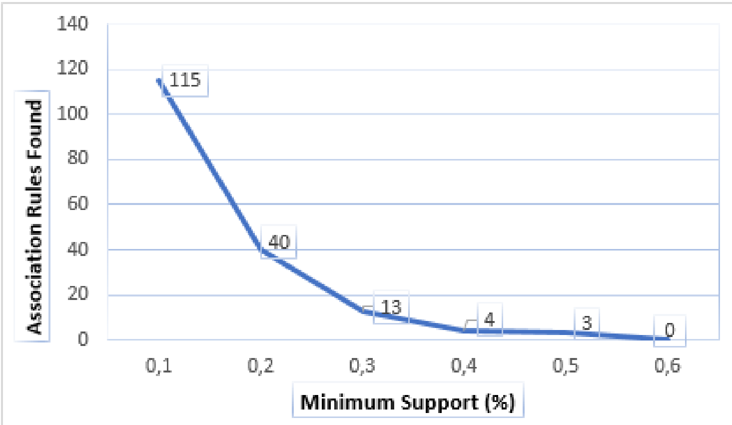

Figure 5 - Associations Found with confidence 10% and varying the support

When a minimum support and confidence are set, all associations that do not verify the conditions are not considered.

As can be seen from the figures above the greater the support and confidence, less associations are considered. With a support of 0,1% and a confidence of 10%, 115 are considered rules but this value drops to zero when the minimum confidence is 60%.

Fixing confidence in 10% and varying the support, the number of rules is relatively low and similar to the number of rules in figure 4.

The following table shows the 10 most frequent item sets in this dataset.

VPH4_key	VPH4 description	Support
11501401701	Frutas banana da marca continente	0,065
10301400101	Água lisa da marca continente	0,047
11802400225	Take away frango assado da marca continente at	0,044
11301400301	Enchidos de fiambre da marca continente	0,029
11503401201	Legumes cenoura da marca continente	0,025
10304400401	Refrigerantes com gás da marca coca cola	0,023
11602402101	Padaria cacetinho congelado da marca continente	0,021
11602403201	Padaria congelado padeira da marca continente	0,021
10301400510	Água lisa garrafão da marca continente	0,020
11602401903	Padaria congelado da marca continente	0,020

Table 11 - List of the 10 most frequent item sets - dataset with promotions

From this table, the first 3 products can be highlighted: (i) frutas banana da marca continente; (ii) água lisa da marca continente and (iii) take away frango assado da marca continente at, with a support value of 0.065, 0.047 and 0.044 respectively. These values mean that bananas da marca continente is present in 6.5 % of the transactions, água lisa da marca continente is in 4,7 % of the transactions while take away frango assado da marca continente is present in 4,4% of the transactions.

These values are not that high meaning that the products that are sold by the company have a great diversity and a shared importance.

The next step was to analyze the rules with greater lift defining a minimum support of 0.1% and a minimum confidence of 10%. The table below describes it.

Antecendent	Description	Consequent	Description	Lift
11503402401	Legumes curgete da marca continente	11503401901	Legumes abóbora da marca continente	36,394
11503402401	Legumes curgete da marca continente	11503402201	Legumes nabo da marca continente	32,517
10108400401	Conservas vegetais de feijão da marca continente	10108402106	Conservas vegetais de grão de bico da marca continente	29,357
11503401901	Legumes abóbora da marca continente	11503401401	Legumes alho da marca continente	20,741
11503402401	Legumes curgete da marca continente	11503401401	Legumes alho da marca continente	20,537
11101400447	Carnes frango da marca continente at	11101400301	Carnes suíno da marca continente	20,182
11503402301	Legumes pepino da marca continente	11503400701	Legumes tomate da marca continente	20,082
11503402201	Legumes nabo da marca continente	11503401401	Legumes alho da marca continente	18,337
11503402401	Legumes curgete da marca continente	11503400401	Legumes couve de brócolo da marca continente	18,147
11304401802	Queijos flamengo da marca terra nostra	11301400302	Enchidos de fiambre da marca nobre	17,773

Table 12 - Rules with greater lift - dataset without promotions

As stated in 2.2.1 lift is a symmetric measure meaning that lift of the rule (A→B) is equal to lift of rule (B→A). Even though confidence can be different, it was decided to describe the first rule avoiding duplicates.

Analyzing the table, it is possible to see that the products with greater lift are normally legumes da marca continente. Analyzing the lift of the rule (11503402401→11503401901) which is equal to 36,394 can be stated that when item 11503402401 is purchased, is 36,394 times more likely item 11503401901 to be in the basket than would be expected, which shows a strong relation between these two items.

The same analysis was made considering the Yule's Q. Like lift, Yule's Q is a symmetric measure and therefore the same procedure was done, describing the first rule instead of repeating the rules only switching the antecedent by the consequent. This is exposed in the table below.

Antecendent	Description	Consequent	Description	Yule's Q
11503402401	Legumes curgete da marca continente	11503401901	Legumes abóbora da marca continente	0,969
11503402401	Legumes curgete da marca continente	11503402201	Legumes nabo da marca continente	0,963
10108400401	Conservas vegetais de feijão da marca continente	10108402106	Conservas vegetais de grão de bico da marca continente	0,954
11503402401	Legumes curgete da marca continente	11503401401	Legumes alho da marca continente	0,942
11503401901	Legumes abóbora da marca continente	11503401401	Legumes alho da marca continente	0,936
11503402401	Legumes curgete da marca continente	11503401201	Legumes cenoura da marca continente	0,934
11101400447	Carnes frango da marca continente at	11101400301	Carnes suíno da marca continente	0,934
11503402301	Legumes de pepino da marca continente	11503400701	Legumes de tomate da marca continente	0,933

11304401804	Queijos de flamengo da marca gresso	11301400301	Enchidos de fiambre da marca continente	0,923
11503402201	Legumes de nabo da marca continente	11503401201	Legumes de cenoura da marca continente	0,923

Table 13 - Rules with greater Yule's Q - dataset without promotions

From the table can be seen that the top rules using lift and Yules Q are similar. This attests the strong dependence between these items. The rule (11503402401→11503401901) with a yules Q of 0,969 confirms the strong relation between these two items.

5.2. Dataset with promotions

In this dataset there are about 15 million transactions with 6654 different items transacted where 6037 of these are purchased on promotion. The average number of items by basket is 9,89 which is quite different from the average number of items by basket in dataset without promotions.

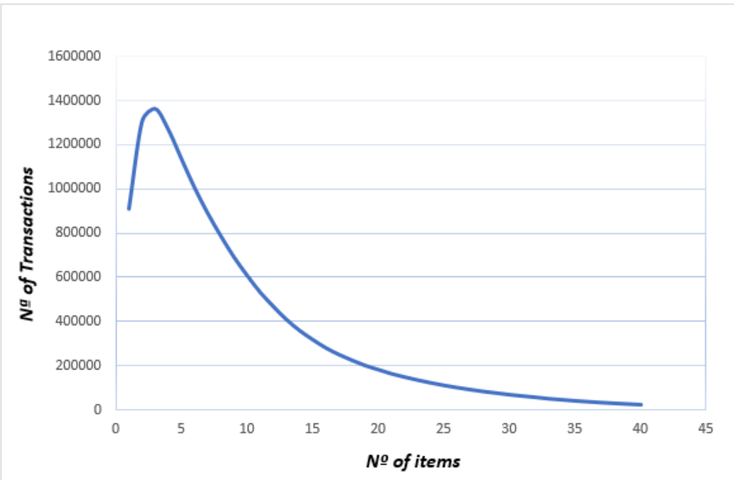

Figure 6 - Number of items vs Number of transactions – dataset with promotions

The figure 6 shows a distribution quite different from the distribution in figure 3. In this dataset, more than 1.3 million of transactions have 3 items in the basket being 3 the most usual size of the basket. The average of 10 items purchased by transaction does not mean that all items are purchased on promotion. It means that among these 10 items, at least one is purchased on promotion.

As was done in dataset without promotions, for this dataset all pair of associations were also calculated. The figure below shows what happens to the association rules found when a minimum support and minimum confidence are set.

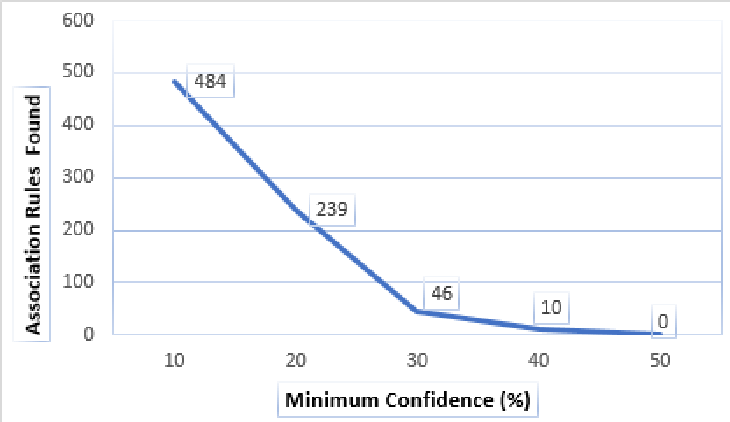

Figure 6 - Associations Found with support 0.1% and varying the confidence

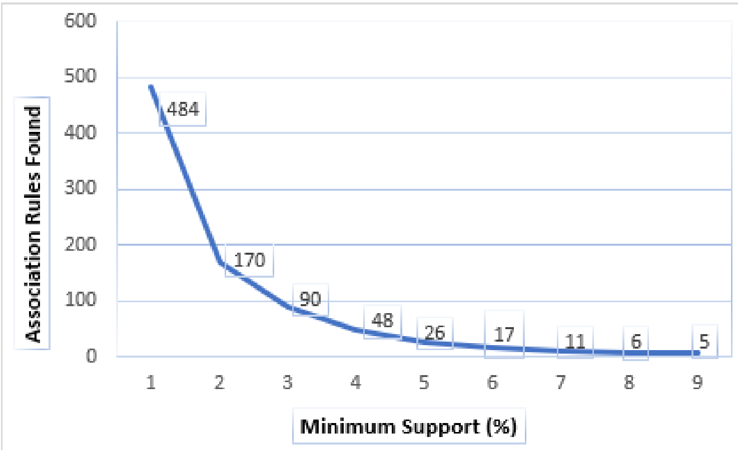

Figure 7 – Association Rules Found with confidence 10% and varying the support

With a support of 0,1% and a confidence of 10%, 484 are considered (greater than in the other dataset) but this value drops to zero when the minimum confidence is 50%.

Fixing confidence in 10% and varying the support, we get 5 rules with a support of 0.9%.

In the following table, the 10 most frequent item sets in this dataset are described.

VPH4_key	VPH4 description	Support
11101400301	Carnes de suíno da marca continente	0,101
11101400101	Carnes de bovino da marca continente	0,037
10802400202	Leite meio gordo da marca gresso	0,032
11503400701	Legumes tomate da marca continente	0,031
10802400201	Leite meio gordo da marca mimosa	0,031
11101400447	Carnes de frango da marca continente at	0,031
10603400101	Amaciador de roupa concentrado da marca comfort	0,024
10302400101	Cervejas com álcool da marca super bock	0,023
11203405201	Peixe dourada fresca da marca continente	0,022
10104400201	azeite de virgem extra da marca oliveira serra	0.022

Table 14 – List of the 10 most frequent item sets - dataset with promotions

From this table it is immediately notice that carnes de suíno da marca continente has a greater support value than the other items. This value means that this item is purchased on promotion in 10.1% of the transactions.

From the list can be highlighted the Continente's own brand having 5 products in this top. All these values are related with transactions on promotion. Again, these support values are not too high which indicates the great diversity of the products sold by the company.

Next, the top 10 rules with greater lift are presented.

Antecendent	Description	Consequent	Description	Lift
11503401901	Legumes abóbora da marca continente	11503402401	Legumes curgete da marca continente	11,786
11503402401	Legumes curgete da marca continente	11503401901	Legumes abóbora da marca continente	10,372
11503402401	Legumes curgete da marca continente	11503402201	Legumes nabo da marca continente	8,926
11301400302	Enchidos de fiambre da marca nobre	11304401802	Queijos flamengo da marca terra nostra	7,402
10112400102	Farináceos de trigo da marca branca de neve	10101400201	Açúcar branco da marca continente	6,845
11503402401	Legumes curgete da marca continente	11503401401	Legumes alho da marca continente	6,232
11503401401	Legumes alho da marca continente	11503402401	Legumes curgete da marca continente	6,139
11503403504	Legumes batata doce da marca continente	11503402401	Legumes curgete da marca continente	6,103
11503402301	Legumes pepino da marca continente	11503400701	Legumes tomate da marca continente	5,936
10107400103	Conservas peixe de atum da marca continente	10120400102	Salsichas de porco da marca continente	5,912

Table 15 - Rules with greater lift - dataset without promotions

As explained in 4.2, in this dataset, lift is not a symmetric measure meaning that the lift of the (A⟶B) is different to the lift of the rule (B⟶A).

Analyzing the table, the rules with greater lift are legumes da marca continente as was already the case in the other dataset. The average lift value is smaller comparing with dataset without promotions.

The lift of the first rule (11503401901 ⟶ 11503402401) means that when the antecedent is on promotion, it is 11,786 times more likely to buy item 11503402401 (with no promotion).

The same analysis was made considering the Yule's Q which is also not a symmetric measure in this dataset. The results are shown below.

Antecendent	Description	Consequent	Description	Yule' s Q
11503401901	Legumes abóbora da marca continente	11503402401	Legumes curgete da marca continente	0,886
11503402401	Legumes curgete da marca continente	11503401901	Legumes abóbora da marca continente	0,866
11503402401	Legumes curgete da marca continente	11503402201	Legumes nabo da marca continente	0,842
11503402201	Legumes nabo da marca continente	11503401201	Legumes cenoura da marca continente	0,813
10112400102	Farináceos de trigo da marca branca de neve	10101400201	Açúcar branco da marca continente	0,812
11301400302	Enchidos de fiambre da marca nobre	11304401802	Queijos flamengo da marca terra nostra	0,807
11503402401	Legumes curgete da marca continente	11503401401	Legumes alho da marca continente	0,794
11503402401	Legumes curgete da marca continente	11503401201	Legumes cenoura da marca continente	0,772
11503400501	Legumes couve repolho coração da marca continente	11503401201	Legumes cenoura da marca continente	0,763
11503401401	Legumes alho da marca continente	11503402401	Legumes curgete da marca continente	0,762

Table 16 - Rules with greater Yule's Q - dataset with promotions

Analyzing the table can be seen that the top rules considering lift and Yules Q are similar. This attests the strong dependence between these items.

Comparing this table with the table with no promotions, here the average Yule's Q is smaller., like happened when the analysis was focused in the lift in this dataset.

The next table shows the possible existence of halo effect in singular associations.

Antecendent	Description	Consequent	Description	Lift without promotion	Lift with promotion
11802401905	Take away refeição pato da marca cont fácil bom	11802402027	Take away acompanhamento da marca nordigal	0,910	5,710
10302400105	Cervejas com álcool da marca cristal	11501401401	Frutas melão da marca continente	0,850	5,689
11802400161	Take away assado da marca continente at	11602402015	Padaria tosta da marca continente	0,730	4,091
11802400606	Take away sopa da marca cont fácil bom	11802401828	Take away empada da marca continente at	2,568	5,886
11602400501	Padaria cereal chapata da marca continente	11304406801	Queijos vaca light barrar da marca vache qui rit	2,417	5,869

Table 17 – Example of associations that can indicate halo-effect

Analyzing the table, can be noticed that the lift of these 5 associations is greater when the antecedent is on promotion than when it is not. It means that when the antecedent is on promotion, it has a halo-effect in the products in the consequent of the rule. It is more likely to purchase the item in consequent of the rule when antecedent is promoted than when it is not.

This event is more likely to occur with complementary goods – a good´s demand increases when the price of another good decreases.

The following table shows the possible existence of cannibalization (opposite of halo effect).

Antecendent	Description	Consequent	Description	Lift without promotion	Lift with promotion
11101400447	Carnes de frango da marca continente at	11101400328	Carnes de suíno da marca montaraz	3,477	0,990
11101400328	Cafés de cápsula da marca continente	11101400328	Cafés de cápsula da marca delta q	1,235	0,383
10301400101	Águas lisa da marca continente	10301400505	Águas lisa garrafão da marca caramulo	1,008	0,656
10802400202	Leite meio gordo da marca gresso	10802400208	Leite meio gordo da marca vigor	1,193	0,563
11304401903	Queijos flamengo fatiado da marca limiano	11304401802	Queijos flamengo da marca terra nostra	1,098	0,423

Table 18 - Example of associations that can indicate no halo-effect or cannibalization

In this table all associations have a smaller lift when the antecedent is promoted than when it is not. This may be an indicator of cannibalization which is the opposite of halo-effect. This is more likely to occur in substitute goods – similar products, like can be seen in the table.

6. Results

Before presenting the results it´s important to talk about the criterion used to exclude non relevant associations. This was done take into consideration the number of transactions between each pair of items (co-occurrence). For the association between two pairs of items to be considered in the analysis there was a minimum number of transactions it had to respect. This criterion was applied in both datasets. The following two graphics illustrate that.

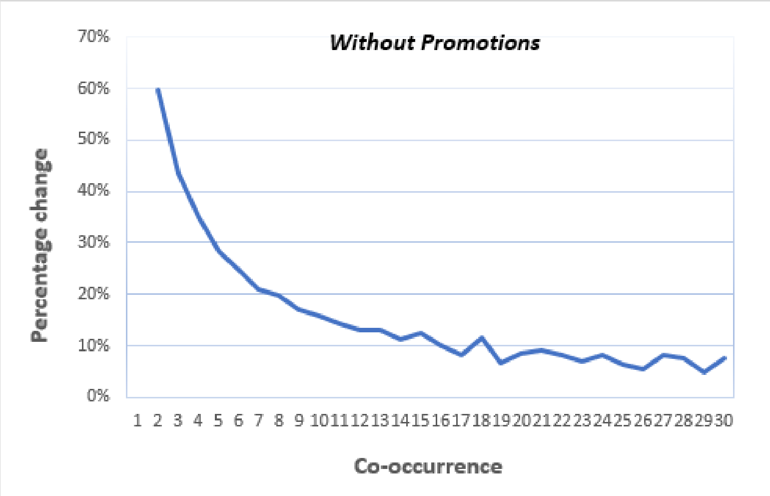

Figure 8 – Criterion used to set a minimum support in dataset without promotions

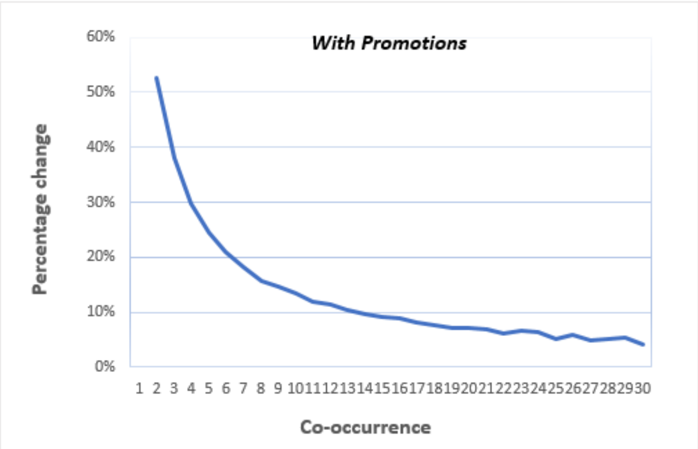

Figure 9 - Criterion used to set a minimum support in dataset with promotions

The idea behind this method is to choose a point where the line begins to stabilize and an inflexion point emerges, cleaning the irregular associations. The y axis represents the percentage change going from one point to another. The x axis represents the co-occurrence, the number of transactions between two pairs of items. As can be seen in the first graph, it drops fast in the first 6/7 points and then starts to stabilize. This analysis is a little subjective and when the line graph begins to stabilize any point can be chosen. 14 was the chosen one, where the inflexion point is perceptible.

In the second graph, it goes down quickly and starts to stabilize. In this dataset 10 was chosen as the minimum number of transactions between two items.

Setting a restriction in the number of transactions a minimum support is being establish when it comes to association rules while when it comes to ego-net it is a weight that is being set instead of support.

Hereafter, the 5 products with higher halo-effect are presented, considering the two methodologies used. It is important to say that even initially considering initially over 6000 products, after applying the minimum support restrictions for both datasets, 3751 products ended up with a score.

Item	Halo Effect Jaccard
11301400304	0
10207400630	0
10207400721	0
11301402307	0
11501400707	0

Table 19 – Items with greater halo effect using Jaccard index – first methodology

As stated in 4.1 Jaccard index measures the similarity. A value of zero means that the two ego networks are completely different and this way indicating halo-effect.

Item	Halo Effect Lift
11501400319	240,296
11304401909	183,712
11501400919	24,097
11501400220	22,097
11501400707	21,707

Table 20 – Items with greater halo effect using Lift – second methodology

Item	Halo Effect Yule's Q
10505401697	0,569
10506401007	0,547
10505402234	0,444
11604400215	0,395
10506401003	0,333

Table 21 – Items with greater halo effect using Yule's Q – second methodology

These values alone say very little to the business. This way, to interpret and validate these scores is a fundamental task trying to add value to the business. That´s exactly what next chapter is about.

7. Interpretation/Validation of the Results

Given the nature of the results, it is essential to interpret and understand what those results can represent in terms of incremental sales adding, this way, value to the business. To do that, it is important to answer the following question: What happen to the incremental sales if the halo-effect goes up? In order to answer this question 2 major steps were taken:

1- Find out which of the three metrics is better when comparing with average market ratio using a polynomial quadratic regression model;

2- Take some action trying to improve the model considering some business structures such as:

- Customer typology;
- Product typology;
- Sales Restriction, considering the most representative products in terms of gross sales (in euros);
- Number of weeks on promotion.

In relation to point one, three regression analysis were made. One for each of the metrics. The goal is to understand the relation between each of the metrics and the average basket ratio through a polynomial regression.

The average basket is defining as the average value of the basket (in euros) containing the product being analyzed.

$$\textbf{Average Basket} = \frac{\boldsymbol{Gross\ Sales\ amount}}{\boldsymbol{number\ of\ transactions}}$$

Formula 7.1 – Average Basket

Therefore, using the banana as an illustrative example, the average basket of the banana is the total value of sales (in euros) of all baskets containing bananas divided by the number of transactions (quantity) containing bananas

The ratio is just a simple arithmetic operation dividing the average basket on promotion by the average basket without promotion.

$$\textbf{Average Basket Ratio} = \frac{\boldsymbol{Average\ Baket\ with\ promotion}}{\boldsymbol{Average\ Basket\ without\ promotion}}$$

Formula 7.2 – Average Basket Ratio

Before presenting the regressions, a framework is required. Regression analysis describes the relationship between two variables. A quadratic polynomial regression is called like that because the curve describing the relationship between the variables is a second-degree polynomial. In terms of the business there was the assumption that regression analysis was no more complex than a quadratic polynomial regression allowing that way the business interpretability.

Therefore, between a linear regression and a quadratic polynomial regression the second one was chosen because the coefficient of determination was slightly better.

Coefficient of determination is used to measure the quality of the adjustment and it is represented by r^2. This coefficient varies between 0 and 1 being that 1 represents a perfect adjustment. A coefficient of 0,8 means that the variable x explains 80% of the values of the variable y.

At an initial stage, mild and extreme severe outliers were removed, in order to improve the model. But in a second analysis only severe outliers were disregarded. The reason is that by eliminating the mild outliers, some observations with positive halo effect were being eliminated which is not desirable because these observations are potentially the main source of interest.

The three graphics are presented below.

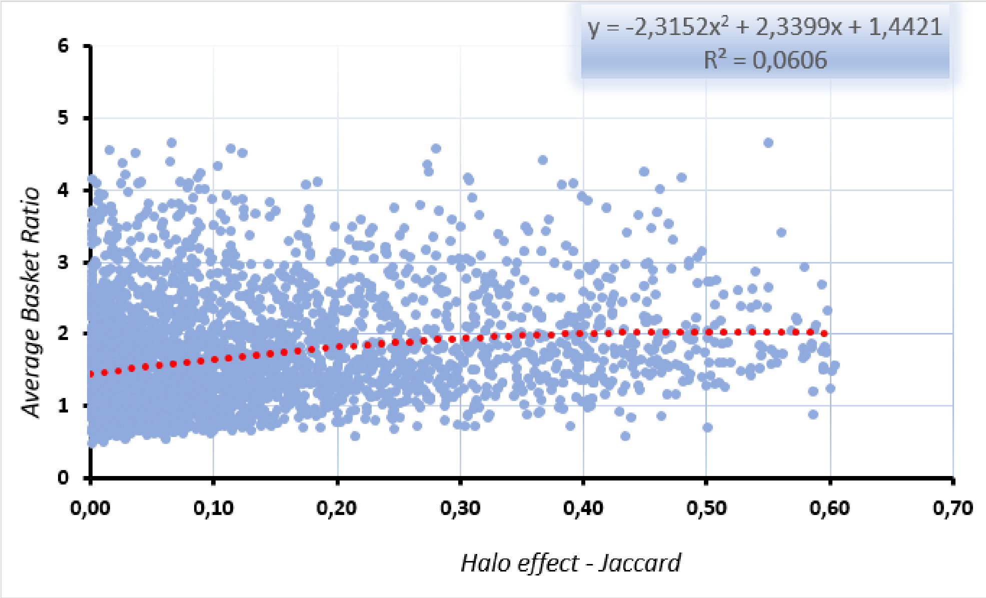

Figure 10 – Quadratic Polynomial Regression – Average Basket Ratio VS Halo Effect Jaccard

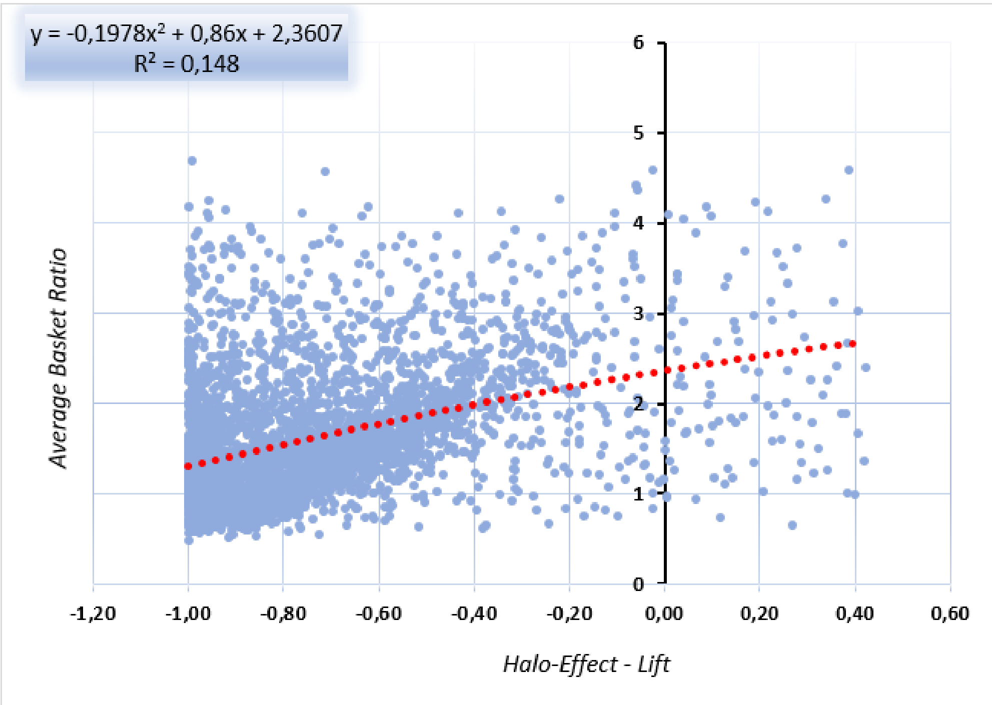

Figure 11 - Quadratic Polynomial Regression – Average Basket Ratio VS Halo Effect Lift

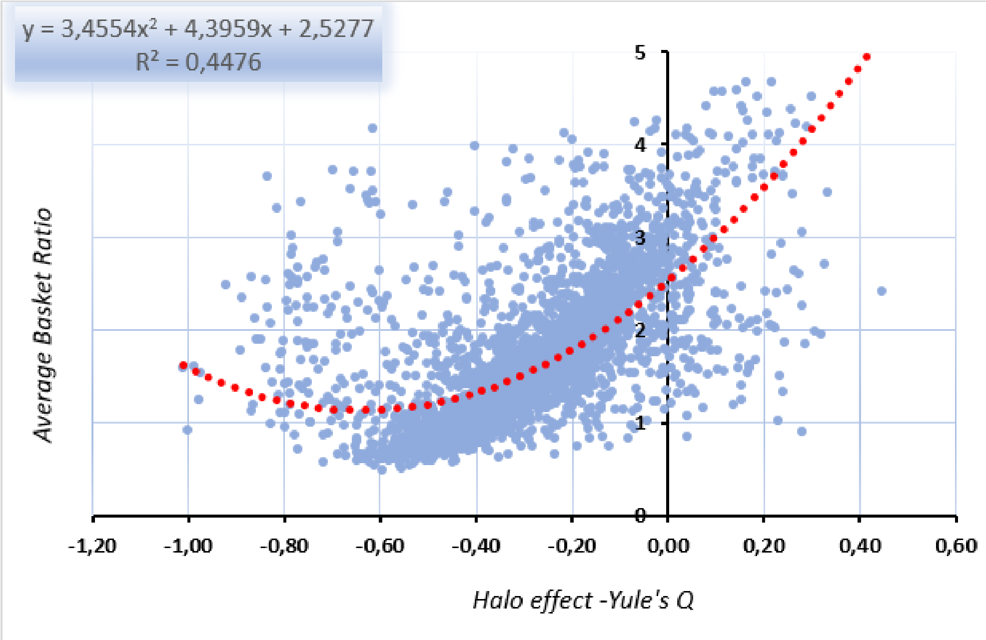

Figure 12 - Quadratic Polynomial Regression – Average Basket Ratio VS Halo Effect Yule's Q

As can be seen, in the three regressions, the r^2 is not good. However, the Yule's Q is the measure that presents a better adjustment and appears to be related with the average basket ratio with a r^2 = 0,4476. This value means that the halo-effect – Yule's Q variable explains 44,76% of the values of the average basket ratio variable. Looking at the curvature on the graph (fig 13), as halo effect increases, the average basket ratio also increases meaning that when the products are on promotion, the customer spends more money which is expected and also it is the definition of halo-effect.

Both lift (r^2 = 0,148) and Jaccard (r^2 = 0,0606) are not a good measure to translate Halo-effect into incremental sales.

Therefore, from now on, the measure used in the rest of the interpretation is the Yule's Q.

As the r^2 obtained was not good some additional approaches were considered. These approaches will be described below.

Customer Typology

Depending on their purchase behavior, customers are allocated in a 8 levels price sensitivity segmentation going from pss1 (customers that are very price sensitivity) to pss8 (price upscale).

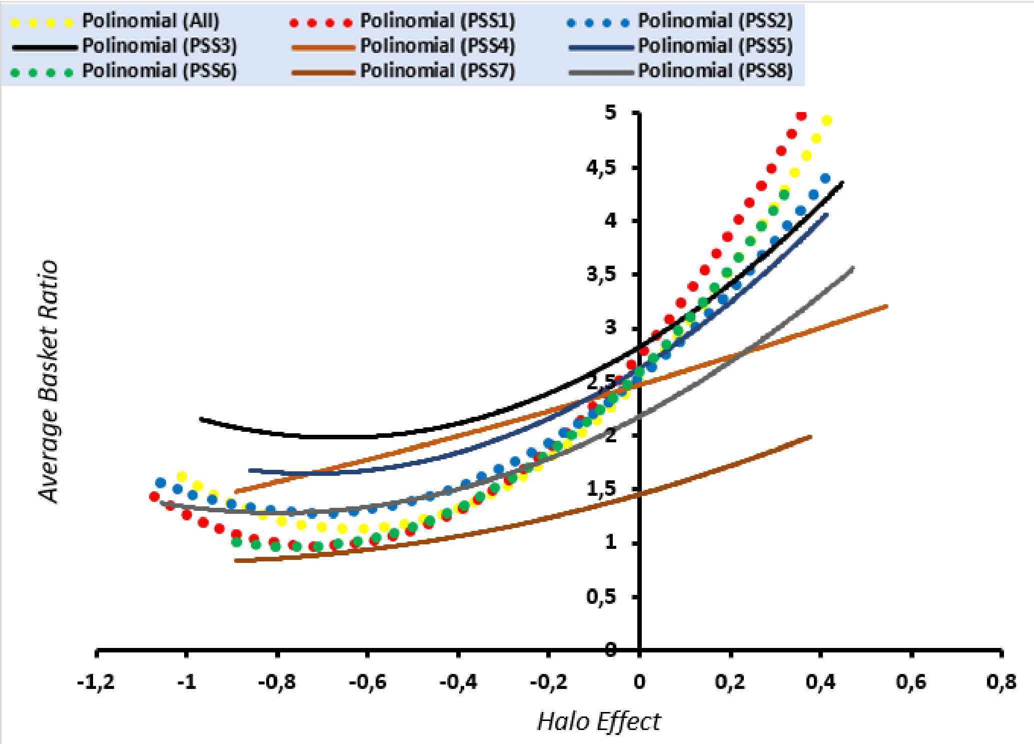

Figure 13 - Quadratic Polynomial Regression – Price Sensitivity Segmentation

Overall, the segments have a different behavior which is good and expected. However, the coefficient of determination for all segments got worse when comparing with the coefficient of determination in figure 13, representing all customers.

Segment_CD	r^2
-	0,4476 (all customers)
1	0,419
2	0,294
3	0,208
4	0,158
5	0,296

6	0,397
7	0,225
8	0,245

Table 22 – Coefficient of Determination considering price sensitivity segmentation

In view of these results, to draw some solid conclusions, the models would have to be improved in a future work.

Product Typology

In this analysis, the regression analysis was divided in 4, one for each department considered in this study.

DC	r^2
10	0,454
11	0,379
12	0,354
41	0,265

Table 23 - Coefficient of Determination by department

Again, the coefficient of determination did not improve overall, being that way, inconclusive.

Most representative products in terms of sales

In this approach were only considered products with at least 100 thousand € in sales.

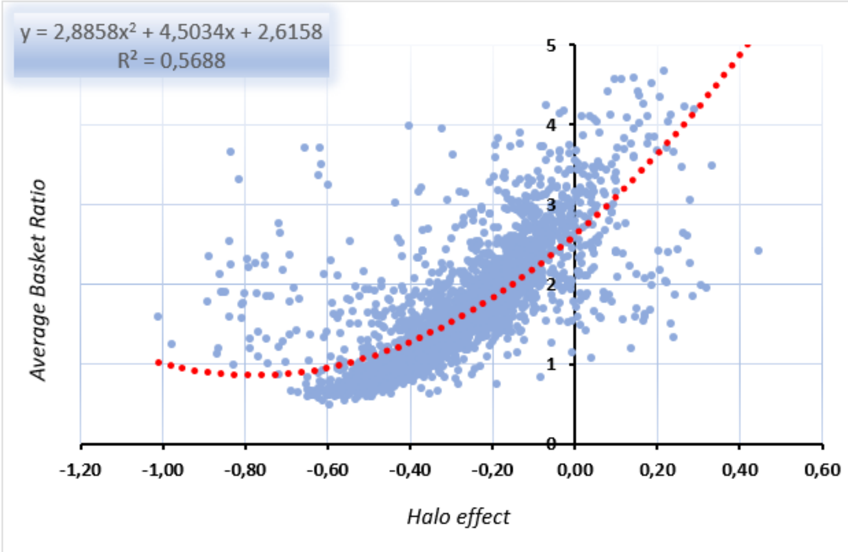

Figure 14 - Quadratic Polynomial Regression considering top selling products (€)

There was a slight improvement in the coefficient of determination. It means that the top sales products are more correlated with the average basket ratio which was, in some way, expected.

Minimum of weeks on promotion

The last analysis considered, was the one that shows more conclusive results. Here, the analysis focused on the number of weeks a certain product is promoted.

The following table shows the relation between that and the coefficient of determination between halo-effect and average basket ratio.

Minimum of weeks on promotion	r^2
10	0,479
20	0,514
30	0,571
40	0,666
50	0,754

Table 24 - Coefficient of Determination considering a minimum of weeks on promotion

As can be noticed, the coefficient is improving when products with more weeks on promotion are considered. As products with less weeks on promotion are being added, the coefficient of determination decreases. The reason is that products with more weeks on promotion are also the top products in terms of sales and those products have a better correlation with the average basket.

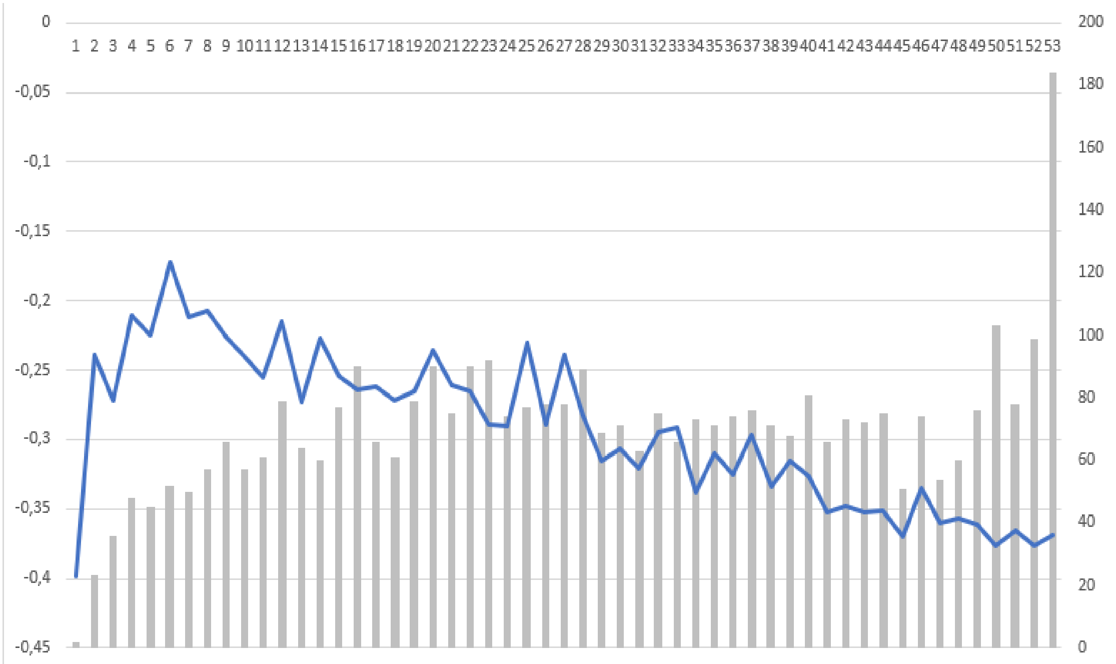

Figure 15 – Average Halo Effect VS Number of Weeks on Promotion

Another major conclusion can be drawn supported by the figure above. The vertical axis on the left that takes values between -0,45 and 0 is the average halo effect represented on the blue line. The vertical axis that takes values between 0 and 200 are the number of products that are on promotion in a given number of weeks represented on the grey bars. The horizontal axis is the number of weeks on promotion and takes values from 1 to 53 weeks.

As an example, the fourth grey bar, from left to right, means that 40 items were promoted in 4 weeks throughout the year and these 40 items have an average halo effect of -0,21.

Analyzing the graphic, it is possible to conclude that as the number of weeks on promotion increases, the average halo effect decreases. This can be explained by the fact that the products which are most often on promotion are the top selling products meaning that even without promotion they have a high demand, an example of that are the essential goods. On the other hand, products less often on promotion, can bring customers to the store trying to take advantage on that less frequently promotions and lead them to buy also other products.

In summary, considering only the products that are more often on promotion, the coefficient of determination improves but at the same time these products are associated with less average halo effect and considering only them can be not desirable for the business.

8. Conclusions

8.1. Main Conclusions

In general, the objective of this dissertation was achieved. It was possible to identify which products, when promoted, have the greatest halo effect contributing to the increase in total store revenue, although further work will be necessary to realize the extent of this increase. In a population of almost 4000 evaluated, more than 200 have halo effect. Taking this into consideration, the company can understand which campaigns are worth investing in and which ones it makes no sense to invest in, thus optimizing its promotional actions.

The approaches used, allowed to estimate a halo effect by product which contributes to extant literature since it´s the first study that estimates the halo effect at this level.

In the personal field, it was a very enriching experience. I had the opportunity to deepen my knowledge in data analysis and machine learning areas. Additionally, I had the opportunity to develop hard skills such as: SQL and SAS language.

Professionally, it was a rewarding experience since it allowed me to face challenging problems, exploring a large amount of data and understanding the dynamics of such big company, developing important skills for the future.

8.2. Limitations and Future Work

Even though the main goal was achieved, obtaining a score for almost 4000 products, there are some limitations that will be described in this section.

One of the limitations is the fact that in this study the quantities or volumes transacted were not taken into consideration. It was only used the number of transactions as indicator of halo effect. The difference in the transacted quantities may also be an indicator of halo effect. Incorporation the traded quantities can be interesting in a future work.

Another limitation is the fact that the coefficient of determination in most of the regression models considered are not good which limits the results interpretability. One possible action that could be taken by the company would considerer for example the 10 products with greater halo effect according to this study and for a given store in a given period of time analyze the sales before, during and after promotional periods of these 10 products.

Nowadays promotions are everywhere therefore, several products are promoted at the same time. It is impossible to change that. Thus, considering the methodologies used in this project, it is not possible to be 100% sure when saying that the sales on product x improved due to promotion on product y because other products are also on promotion which represents one more limitation of this project even though this limitation is softened by the fact that the same products are not always on promotion at the same time thus allowing to analyze the promotional impact.

9. References

Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining Association Rules between Sets of Items in Large Databases*.

Agrawal, R., & Ramakrishnan, S. (1994). Fast Algorithms For Mining Association Rules. *20th Int. Conf. Very Large Data Bases,VLDB, 1215, 487–499*.

Ahmed, S. R. (2004). *Applications of Data Mining in Retail Business*.

Ailawadi, K. L., Harlam, B. A., César, J., & Trounce, D. (2007). Practice Prize Report—Quantifying and Improving Promotion Effectiveness at CVS. *Marketing Science*, *26*(4), 566–575. https://doi.org/10.1287/mksc.1060.0245

Bell, D. R., Chiang, J., & Padmanabhan, V. (1999). The Decomposition of Promotional Response: An Empirical Generalization. *Marketing Science*, *18*(4), 504–526. https://doi.org/10.1287/mksc.18.4.504

Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., & Wets, G. (2001). *A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Model*. 1–13.

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (1999). Using association rules for product assortment decisions: A case study. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99*, 254–260. https://doi.org/10.1145/312129.312241

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *26*(2), 255–264. https://doi.org/10.1145/253262.253325

Bucklin, R. E., Gupta, S., & Siddarth, S. (1998). Determining Segmentation in Sales Response across Consumer Purchase Behaviors. *Source Journal of Marketing Research*, *35*(2), 189–197. https://doi.org/10.2307/3151847

Chintagunta, P. K. (1993). *Investigating Purchase Incidence , Brand Choice and Purchase Quantity Decisions of Households . Marketing Sci. 12*(2), 184–208.

Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J., & Tranmer, M. (2015).

Social Network Analysis for Ego-Nets (R. Chris, ed.). SAGE Publications Ltd.

Drèze, X., Nisol, P., & Vilcassim, N. J. (2004). *Do Promotions Increase Store Expenditures ? A Descriptive Study of Household Shopping Behavior Do Promotions Increase Store Expenditures ? A Descriptive Study of* (Vol. 2).

Gama, J., Carvalho, A. P. de L., Katti, F., Ana Carolina, L., & Oliveira, M. (2017). *Extração de Conhecimento de Dados – Data Mining* (2nd ed.). Edições Sílabo, Lda.

Gauri, D. K., Ratchford, B., Pancras, J., & Talukdar, D. (2017). An Empirical Analysis of the Impact of Promotional Discounts on Store Performance. *Journal of Retailing*, (1987), 1–21. https://doi.org/10.1016/j.jretai.2017.06.001

Gonçalves, E. C. (2005). Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. *INFOCOMP Journal of Computer Science*, *4*(1), 26–35.

Gupta, S. (1988). Impact of Sales Promotions on When, What, and How Much to Buy. *Journal of Marketing Research*, *25*(4), 342. https://doi.org/10.2307/3172945

Heerde, H, Gupta, S., & Wittink, D. (2003). Is 75 percent of the sales promotion bump due to brand switchig? {N}o, only 33 percent is. *Journal of Marketing Research*, *40*(4), 481–491.

Heerde, Harald Van, & Leeflang, P. S. H. (2004). *Decomposing the Sales Promotion Bump with Store Data*. (June 2014). https://doi.org/10.1287/mksc.1040.0061

Heerde, Harald Van, & Neslin, S. A. (2017). *Sales promotion models*. https://doi.org/10.1007/978-3-319-56941-3

Jiawei, Han; Kamber, Micheline; Pei, J. (2011). *Data Mining : Concepts and Techniques. 3rd Edition*. https://doi.org/10.1093/nar/gku1019

Kotler, P., Armstrong, G., Saunders, J., & Wong, V. (1999). Principles of Marketing. In *The British Journal of Psychiatry* (2nd ed., Vol. 111). Prentice Hall Europe.

Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, *184*(2), 610–626. https://doi.org/10.1016/j.ejor.2006.10.059

Liebson, M. (2003). *An Enterprise Approach to Maximising Promotion Effectiveness*.

Ma, S., & Fildes, R. (2017). A retail store SKU promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research*, *260*(2), 680–692. https://doi.org/10.1016/j.ejor.2016.12.032

Ogden-barnes, S., & Minahan, S. (2015). *Sales Promotion Decision Making - concepts, principles, and practice*.

Raghubir, P., Inman, J. J., & Grande, H. (2004). *The three faces of consumer promotions*.

Shaikh, M. R., McNicholas, P. D., Antonie, M. L., & Murphy, T. B. (2013). Standardizing interestingness measures for association rules. *Statistical Analysis and Data Mining*. https://doi.org/10.1002/sam.11394

Silberschatz, A., & Tuzhilin, A. (1995). On Subjective Measures of Interestingness in Knowledge Discovery. *Kdd´95*, 275–281.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/69.553165

Tan, P. N., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32–41.

Tan, P. N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*(4), 293–313. https://doi.org/10.1016/S0306-4379(03)00072-3

Tichy, N. M., Tushman, M. L., Fombrun, C., & Tushman, M. L. (1979). Social Network Analysis for Organizations. *Academy of Management Preview*, *4*, 507–519.