U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

PROMOTIONAL SALES FORECASTING IN A RETAIL SETTING

Ana Margarida de Figueiredo Pereira

Dissertation
Master in Data Analytics

Supervised by
Jorge Valente

2023

Acknowledgements

This dissertation represents the culmination of my academic career and of my journey at Faculdade de Economia do Porto as I complete my master's degree in Data Analytics following my bachelor's degree in Economics. This was a hard but gratifying path.

First, I would like to thank LTPlabs for the opportunity to develop this dissertation in a business environment. It allowed me to develop my skills in a real environment in a highly specialized organization. Everyone at LTP was verry important over these past months as they fostered a great and supportive environment. However, a special acknowledgement goes to the Data Science team that I had the pleasure to join for this project. Working with Bruno Batista, Rui Azevedo and Klaus Garrido I was able to greatly improve my skills and learn immensely, their direction was fundamental in achieving all my objectives with this dissertation. Also, my supervisor Diogo Miranda played a verry important role in the revision and guiding of the focus of my work.

Second, a very special acknowledgement goes to my professor Jorge Valente, that from the beginning was available to promptly answer all my questions and concerns. He always guided me in the best way to achieve and strive for the best out of this project.

This was a long path surrounded by the best people, so to my friends that shared any part of it with me I also owe a thank you. Especially Inês, form the bachelor's degree, to the master's, to being roommates, the long study hours finally paid off!

Last, but not least, I have to thank my family for always supporting me, believing in my abilities and pushing me to always give my best.

Abstract

Having accurate demand predictions is essential to optimize retail operations such as stock replenishment, inventory management and ultimately meet customer needs. The demand forecasting task is challenging due to the various factors that affect demand - the economic scenario, market competition and consumer preferences to name a few. Promotional actions introduce additional layers of complexity to this task as they encompass many levels of product interactions and effects.

This dissertation focuses on promotional sales forecasting applied in a major Portuguese retailer scenario. It aims at improving the current forecasting solution by actively accounting for promotion cannibalization and stockpiling effects in the prediction process. To achieve this a great effort was put into engineering specific features that can capture these effects at various levels. To predict promotional sales two main approaches are explored, one using regression techniques and another using machine learning algorithms.

For the regression-based methods, in order to account for heterogeneity between store types and for homogeneity between similar products nine variations of the regression model are estimated. For the machine learning portion, two tree based algorithms are considered, specifically Random Forest and Gradient Boost Regressor.

The multiple regression methods proved that including promotional effects' related variables has a positive impact in prediction accuracy. Furthermore, this approach yielded better results than expected, outperforming the machine learning techniques in many cases.

Resumo

Ter previsões da procura precisas é essencial para otimizar as operações de retalho, tais como a reposição de stocks, a gestão de inventários e, em última instância, satisfazer as necessidades dos clientes. A tarefa de previsão da procura é um desafio devido aos vários fatores que afetam a procura, tais como o cenário económico, a concorrência no mercado e as preferências dos consumidores, para citar apenas alguns. As ações promocionais introduzem uma adicional complexidade a esta tarefa, uma vez que abrangem vários níveis de interações de produtos e diferentes efeitos.

Esta dissertação centra-se na previsão de vendas promocionais aplicada num cenário de um grande retalhista português. O seu objetivo é melhorar a atual solução de previsão através da consideração ativa de efeitos de canibalização e de armazenamento no processo de previsão. Para atingir este objetivo, foi feito um esforço no cálculo de variáveis específicas capazes de capturar estes efeitos a vários níveis. Para prever as vendas promocionais, são exploradas duas abordagens, uma utilizando técnicas de regressão e outra utilizando algoritmos de aprendizagem automática.

Para os métodos baseados na regressão, de modo a ter em conta a heterogeneidade entre os tipos de lojas e a homogeneidade entre produtos semelhantes, são estimadas nove variações do modelo de regressão. Relativamente à parte de aprendizagem automática, são considerados dois algoritmos baseados em árvores, especificamente *Random Forest* e *Gradient Boost Regressor*.

Os diferentes métodos de regressão provam que a inclusão de variáveis relacionadas com os efeitos promocionais tem um impacto positivo na precisão das previsões. Além disso, esta abordagem produziu resultados melhores do que o esperado, superando as técnicas de aprendizagem automática em muitos casos.

Table of Contents

1.Introduction **1**
2.Literature Review **5**
2.1. Univariate Forecasting Methods 5
2.2. Linear Causal Methods 7
2.3. Nonlinear and Machine Learning Methods 9
2.4. Promotions 11
2.5. Practical Applications 12
3.Project Description **17**
3.1. Current Methodology 17
3.1.1.Description of the current methodology 18
3.1.2.Shortcomings of the current methodology 19
3.2. Proposed Approach 20
3.3. Available Data 21
4.Methodology **25**
4.1. Preprocessing 25
4.1.1.Baseline 26
4.1.2.Cleaning 27
4.1.3.Classification 27
4.2. Input Variables 28
4.2.1.Product attributes 29
4.2.2.Promotional effects variables 33
4.3. Regression Approach 39
4.4. Machine Learning Approach 41
4.5. Training and Test Splits 44
4.6. Evaluation Metrics 45
4.7. Used Packages 46
5.Results **47**
5.1. Regression Results 47
5.2. Machine Learning Results 51
5.3. Overall Comparison 52

6.Conclusion **56**
6.1. Limitations 57
6.2. Future Work 58
References **59**
Appendixes ..... **i**
Appendix A i
Appendix B ii
Appendix C ix
Appendix D xxii

List of Tables

Table 3.1 - Position of each category related to relevant factors 23
Table 3.2 - Percentage of weeks of the period with the average number of products on promotion 24
Table 3.4 - Number of unique products in each category and Percentage of products present per store type 24
Table 4.1 - Number of varieties in each product attribute 31
Table 4.2 - Number of clusters and single product clusters per category 32
Table 4.3 - Description of the available variables 37
Table 4.4 - Discarded features per category 41
Table 4.5 - Overview of the hyperparameters 44
Table 5.1 - Regression results for the category Beers and Ciders 48
Table 5.2 - Machine Learning results per category 51
Table 5.3 - Results for the regression, machine learning and RDF per category 52
Table B.1 - Clusters considering 3 attributes for Beers and Ciders ii
Table B.2 - Clusters considering 2 attributes for Beers and Ciders iii
Table B.3 - Clusters considering 3 attributes for Liquid Fats iv
Table B.4 - Clusters considering 2 attributes for Liquid Fats v
Table B.5 - Clusters considering 3 attributes for Eggs v
Table B.6 - Clusters considering 2 attributes for Eggs vi
Table B.7 - Clusters considering 3 attributes for Sweets and Chocolate vii
Table B.8 - Clusters considering 2 attributes for Sweets and Chocolate viii
Table C.1 - Variance Inflation Factor for category Beers and Ciders store type L x
Table C.2 - Variance Inflation Factor for category Beers and Ciders store type M xi
Table C.3 - Variance Inflation Factor for category Beers and Ciders store type S xii
Table C.4 - Variance Inflation Factor for category Liquid Fats store type L xiii
Table C.5 - Variance Inflation Factor for category Liquid Fats store type M xiv
Table C.6 - Variance Inflation Factor for category Liquid Fats store type S xv
Table C.7 - Variance Inflation Factor for category Eggs store type L xvi
Table C.8 - Variance Inflation Factor for category Eggs store type M xvii
Table C.9 - Variance Inflation Factor for category Eggs store type S xviii

Table C.10 - Variance Inflation Factor for category Candy and Chocolate store type L xix
Table C.11 - Variance Inflation Factor for category Candy and Chocolate store type M xx
Table C.12 - Variance Inflation Factor for category Candy and Chocolate store type S xxi
Table D.1 - Regression results for the category Liquid Fats xxii
Table D.2 - Regression results for the category Candy and Chocolate xxii
Table D.3 - Regression results for the category Eggs xxiii

List of Figures

Figure 5.1 - Results for store type L in category Liquid Fats 49
Figure 5.2 - Results for store type M in category Candy and Chocolate 49
Figure 5.3 - Results for store type S in category Eggs 50
Figure A.1 - Detailed description of the promotional information i

Acronyms

ADL – Autoregressive Distributed Lag

ANN – Artificial Neural Networks

ARIMA – Autoregressive Integrated Moving Average

GBM – Gradient Boosting Machines

GBR – Gradient Boosting Regressor

LASSO – Least Absolute Shrinkage and Selection Operator

MAE – Mean Absolute Error

MAPE – Mean Average Percentage Error

ML – Machine Learning

MPE – Mean Percentage Error

MSE – Mean Squared Error

PCA – Principal Component Analysis

RDF – Retail Demand Forecasting

RF – Random Forest

SKU – Stock Keeping Unit

SVR – Support Vector Regression

TS-N-CV – Time-Series Nested Cross-Validation

VIF – Variance Inflation Factor

WMAPE – Weighted Mean Absolute Percentage Error

WMPE – Weighted Mean Percentage Error

1. Introduction

Grocery retail is recognized as a highly profitable and fiercely competitive sector. In 2020, the retail industry in Portugal constituted 37% of the sales volume of the entire commerce sector, with food or food-related retail accounting for 48% of commercial establishments. The number of retail establishments saw a 1.4% increase in 2020, emphasizing the ongoing nature of competition. The retail industry's commercial margin reached 275.6 thousand euros per retailer company in 2020, reflecting its profitability. Supermarkets, which make up 68.7% of the retail landscape, played a significant role in driving sales volume (INE, 2020).

In addition to its profitability and competitiveness, it is crucial to highlight the significant role that marketing tactics and promotions play in the current food retail scenario (Ailawadi et al., 2006; Fildes et al., 2022; Hewage & Perera, 2022). Portuguese consumers exhibit a strong inclination towards price promotions. This phenomenon has led to a setting where retailers must rely heavily on discounts to capture the attention of customers. As a result, nearly half of all grocery purchases now consist of discounted and promoted products. The proliferation of mega promotions in Portuguese supermarkets has become a prevailing trend, with 50% discounts becoming increasingly commonplace (Mendes & Pereira, 2021).

In the intensely competitive and highly lucrative grocery business, retailers face increasing complexity, intensified competition and higher consumer expectations. To succeed in this environment, accurate demand forecasting and efficient stock replenishment processes are vital in delivering the right products to the right place, optimizing operations and meeting customer needs.

Having accurate demand forecasts plays a critical role in the retailer's decision-making process across many levels. To determine the best forecasting method it is essential to understand beforehand the scope and environment for which the forecasts are needed. Furthermore, the prediction granularity relates to the time horizon, forecasting gap, product hierarchy level and retail chain position. These different aspects should be considered depending on the decision level. Strategic decisions, such as expansion of the sales channel or stores, require high level aggregated predictions over a longer time horizon. While tactical decisions, related to

communication and advertisement planning, or product line breadth and depth, might need more tailored predictions for a more near future.

The present project focuses on assisting operational decision-making, specifically on ordering and stock planning. The focus lies on product forecasting at the store and store type level for a short time horizon, which is crucial to minimize stock-outs.

Forecasting, in general, but especially at the product store weekly level, is challenging due to the many factors that can impact demand. There are external factors, such as special events, or the time of the week, month or year, or the socio-economic conditions, as well as retailer related factors as communication and promotional actions. Additionally, there are further complications introduced by product interactions such as cannibalization, substitution and complementarity effects (Abolghasemi et al., 2020; Fildes et al., 2022; Ma et al., 2016; van Heerde et al., 2000).

Given all the factors stated, sales data tends to exhibit fluctuations, seasonal patterns and occasional stock-outs. Surmounting these challenges is crucial to generate accurate predictions that can, in the end, prevent revenue losses or poor customer service.

The presence of promotions is one of the many factors that contribute to the complexity of demand forecasting, as has been mentioned. Promotional efforts encompass price cuts and marketing tactics, introducing additional challenges into the forecasting process.

Different types of promotions, such as percentage-based price reductions or bundle deals, coupled with diverse advertising and communication methods, contribute to the intricate landscape. Moreover, the way promoted items are displayed in stores can vary, ranging from dedicated placement at the store front to regular location with changed price tags.

Promotional type, feature advertising and display are among the most influential factors affecting the success of promotional efforts, as they can significantly impact consumer perception and behavior. Additionally, these tactics enhance and amplify the existing interaction effects between products. Specifically, common responses to promotions, as identified by (Gupta, 1988), include brand switching, purchase acceleration, and holdout, leading to substitution or cannibalization effects and stockpiling or pre- and post-promotional dips.

The demand predicting task becomes more intricate when these various promotional aspects, interaction effects, and competitive dynamics are taken into account. However, research by (Ma et al., 2016) and (Srinivasan et al., 2005) demonstrates that incorporating these factors can substantially improve forecasting accuracy.

Due to the wide range of characteristics and intricate interaction effects, forecasting product sales using formal model-based approaches can be an elaborate task. Consequently, many retailers rely on a combination of simple statistical approaches and judgmental adjustments.

This dissertation arises in the context of an ongoing internal project being developed at an analytical driven consultancy company for a Portuguese food retailer. As previously emphasized previously, having accurate demand predictions is pivotal and the current methodology applied by the retailer uses somewhat simpler models that do not include key promotional factors. To this extent, the primary objective of the project is to develop a promotional forecasting methodology that can surpass the results that are currently being obtained. In the end, getting better demand predictions will have managerial relevance by enhancing the replenishment planning processes. Moreover, a significant effort was put into capturing various product interactions and promotional effects, highlighting their substantial impact on the forecasting results.

The developed work can be characterized by five main stages, the first of which focused on understanding the methodology that is currently applied by the retailer to tackle the problem and on studying its limitations and opportunities for improvements. The second stage was to collect the necessary data to take on this project. The third stage consisted of data exploration which entailed firstly understanding the data available and secondly exploring how it could be leveraged and manipulated according to the goals for the forecasting methods. This included some data cleaning, processing and feature engineering.

The fourth stage can be looked at as the development of a first model and this process can be divided into different sub steps. First, a baseline was estimated. This is outside the scope of this dissertation, but the baseline will be mentioned and used as part of the available data. This is followed by the implementations of simple univariate models to serve as benchmarks. Lastly,

the focus is on the development of linear regression and machine learning methods including promotion and product interaction features.

The final stage entailed analyzing the first results obtained, making comparisons with the existing methodology results and with the benchmarks. From this point on the process involved improving the built models, exploring additional features and evaluating and comparing results.

The remainder of this dissertation is structured in 6 chapters, as follows. Chapter 2 gives some theoretical background and literature review related to the most relevant topics presented in this project. It includes an overview of significant and widely applied forecasting techniques, both for promotional and non-promotional scenarios. It also gives some theoretical background on specific techniques that will be mentioned later on. Chapter 3 describes in greater detail the problem being addressed, the retailer's context and the current approach that is followed. Chapter 4 comprehensively describes the methodology developed and adopted. Chapter 5 depicts and discusses the results of the implementation of the proposed methodology. Chapter 6 finally summarizes the main conclusions that were drawn, acknowledges the limitations of this dissertation of this dissertation and presents some possibilities for future work.

2. Literature Review

Forecasting involves analyzing historical data and projecting future situations and considerable efforts have been dedicated to the development and improvement of various models and techniques. Throughout the years there has been a transition from relying solely on expert judgment and intuition to data-driven decision making and utilizing more complex modeling.

In the present chapter, some of the most relevant forecasting techniques will be presented, starting with simpler benchmark techniques, proceeding to linear and econometric models and ending with sophisticated and more intricate machine learning methods.

2.1. Univariate Forecasting Methods

Univariate forecasting methods focus on using past sales historical data as input to obtain predictions. Traditional time series techniques such as simple moving averages, exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) are commonly employed examples of this type of approach. These are frequently utilized as benchmarks for performance evaluation, especially the base-times-lift and exponential smoothing approaches, as seen in the studies by (Ali et al., 2009; Hewage & Perera, 2022; Huang et al., 2014).

However, univariate techniques should be primarily adopted for situations where predictions are necessary at a higher level of aggregation or where products have minimal promotional activity and conditions are more stable (Ali et al., 2009; Fildes et al., 2022; Martins et al., 2022).

Exponential smoothing is a simple univariate method that relies on the previous sales pattern. The method assigns weights to past observations that decrease exponentially over time. This translates the assumption that future sales will be closer to more recent past sales than more distant past sales.

In Simple Exponential Smoothing sales predictions are a combination of the most recent observation and the previous forecast. More specifically:

$$FS_t = \alpha . S_t + (1 - \alpha) . FS_{t-1}$$

Where:

FS_t represents the forecasted sales in period t

S_t represents the most recent observed sales

FS_{t-1} represents the previous forecast

α represents the smoothing factor

In this most simple type of exponential smoothing, the alfa value controls the rate at which the importance of past information decreases. The default value for the factor is 0.2 (Ali et al., 2009).

This is a widely used method due to its simplicity, efficacy and interpretability. Moreover, nowadays it is widely available: not only it is integrated in many commercialized prediction software but it is also easily computed using free software such as *R*. (Ali et al., 2009; Hewage & Perera, 2022).

Some authors choose to define promotional sales as a multiple of baseline sales, also referred as baseline-times-lift approach. The approach typically involves a two-step procedure. At first, a baseline forecast or estimation is created using a simple time series method, as simple averages or exponential smoothing, keeping in mind that the data used for this step should be cleansed of demand drivers such as past promotions and weather. The baseline forecast is then adjusted for upcoming events, typically promotions, with the adjustments often estimated based on the lift effect observed in recent similar events (Cooper et al., 1999). There are also situations where the adjustments are based on professional expert opinions (Fildes et al., 2009; Lee et al., 2007).

Using analogous past promotions that are considered similar to the upcoming promotion is a common and a natural basis for forecasting and it is shown in multiple works throughout the literature (Fildes et al., 2009; Lee et al., 2007). As stated, one of the most widely used methods for the simple baseline calculation is exponential smoothing as presented in (Hewage & Perera, 2022). However, the base-lift method with simple exponential smoothing is predominantly used as a benchmark for comparison purposes (Ali et al., 2009; Huang et al., 2014; Trapero et al., 2015).

ARIMA is a time series forecasting method, developed by (Box et al., 2016), that relies on autocorrelation with past data and grounds its predictions on extrapolations from past patterns. Due to its nature and the three components that it takes into consideration, it is a widely used method for contexts that exhibit different patterns and dependencies over time.

The purpose of the autoregressive component (AR) is to capture the relation between the present and the past values of the target variable. The integrated component (I) serves to eliminate trends and patterns from the data. It works by differentiating the data to ensure stationarity, meaning that properties as the mean and variance remain constant over time. Lastly, the moving average (MA) accounts for the relation between the target variable and the residual errors from past predictions. The parameters of the ARIMA model are then estimated using statistical techniques such as maximum likelihood estimation.

Furthermore, there have been some modifications and multivariate extensions of the simple ARIMA model, of which can be pointed the ARIMAX, the SARIMA and the SARIMAX (Arunraj & Ahrens, 2015). In these models, the X component represents the inclusion of the external factors that may be relevant to capture variations and volatility in the sales' patterns (Abolghasemi et al., 2020). Additionally, the S factor accounts for seasonality elements present in the data.

2.2. Linear Causal Methods

Another line of research relates to the use of model-based systems or econometric approaches which incorporate extra factors into product demand forecasting. These methods can use multiple linear regression models or other more complex causal models. The distinctive aspect, from what was presented before, is the inclusion of exogenous inputs such as promotional, event or seasonality information, as is the case with the SARIMAX model. Some of the noted advantages of using linear causal methods are their simplicity, ease of implementation and computational efficiency, making them not only understandable and interpretable but also scalable (Fildes et al., 2022; Martins et al., 2022).

A widely known and used model is the SCAN*PRO, first developed in (Wittink et al. 1988) and its adaptations and variations present in works of many authors (Leeflang et al., 2005)

The original SCAN*PRO is a multiplicative regression model that decomposes weekly brand store sales based on own and cross brand effects of price indexes, store displays and marketing efforts. It also accounts for seasonality with weekly indicator predictors. When it was first introduced the method considered that parameters were homogeneous across stores. Throughout the years, many additions have been introduced by the author, such as store specific parameters, variations for the estimation's level of aggregation, dynamic predictors, to name some (Leeflang et al., 2005).

Furthermore, different authors present variations of the SCAN*PRO methodology as is the case with (Abolghasemi et al., 2020; Andrews et al., 2008; van Heerde et al., 2000). The extensions of the model differ by accounting for additional factors such as dynamic promotion effects and store heterogeneity in marketing mix.

An additional stream within the linear causal methods commonly applied to the retail forecasting setting, are the techniques that are many times used to tackle dimensionality concerns. These refer to methodologies such as the Least Absolute Shrinkage and Selection Operator (LASSO) regression. The goal with the mentioned approach is to reduce the number of predictors in a regression, thus including a selection element into the estimation process (Bajari et al., 2015).

In the retail forecasting context, especially when accounting for various demand effects, as promotions, and for interactions and dynamics, dimensionality is a major concern (Trapero et al., 2015). Incorporating competitive information has been proven valuable yet including own and cross product or inter and intra category aspects can highly increase the predictor variable space. Thus, justifying the popularity of regularization techniques as the LASSO (Huang et al., 2014; Ma et al., 2016).

The LASSO is a method first proposed by (Tibshirani, 1996) for the purpose of estimating linear models. The LASSO regression works by minimizing the sum of squared errors, while penalizing the absolute values of the coefficients. The nature of the method leads to automatic feature selection as the penalty given to the values of the coefficients tends to result in coefficients equal

to zero(Fildes et al., 2022; Tibshirani, 1996). In studies such as Bajari et al., 2015; Huang et al., 2014; and Ma et al., 2016 this approach is used as a way to deal with high dimensionality and feature selection.

2.3. Nonlinear and Machine Learning Methods

Nonlinear methods, in contrast to linear regression models, directly obtain nonlinear approximation functions from data, increasing the potential accuracy, although with a risk of overfitting (Fildes et al., 2022). In the grocery retail context, research has proven that nonlinear complex models can outperform linear regressions in terms of accuracy (Ali et al., 2009; Bajari et al., 2015; Hewage & Perera, 2022; Kaneko & Yada, 2016). These nonlinear methods include Gradient Boosting Regressor (GBR) Trees algorithms, such as Light Gradient Boosting Machines (GBM) and XGBoost, Random Forest (RF), Artificial Neural Networks (ANN), Support Vector Regression (SVR) and Deep Learning algorithms, to name some (Fildes et al., 2022).

Artificial Neural Networks is a supervised machine learning algorithm which is inspired by the human brain in the way it learns from experience. It is constructed by neurons that are interconnected and organized in layers. While training, the network adjusts the weights of the neurons in order to minimize prediction errors. A popular type of this algorithm is feedforward ANN, which also consists in layers of interconnected neurons, yet these are sequentially organized. Here the network receives the input data, propagates it forward through the layers and gives the output predictions. At the same time, the connections' weights are corrected according to the calculated errors in a backpropagation manner (Abolghasemi et al., 2020).

SVR is one of the common applications of Support Vector Machines that is a supervised machine learning technique. The goal of these models is to find the optimal fit of the data points in a high dimensionality feature space. SVRs aim to find a linear function that predicts the objective variable within a set distance from the respective values to stay within a defined margin

of error. Additionally, this algorithm employs a kernel function so that low dimensional data is transformed into high dimensional (Abolghasemi et al., 2020).

The simplest Regression Trees algorithm partitions the data in order to achieve the highest uniformity at each 'leaf'. To do this in a greedy manner it selects the feature and threshold that at each 'node' are able to explain most of the variation in the data. The predictions of new data points are then determined by the value or the average value of the response variable at the 'leaf' that corresponds to the features of the new instance (Breiman, 1996).

Additionally, (Breiman, 1996) presents an ensemble method known as Bagging. This is proposed with the intention of addressing pointed limitations of the Regression Trees algorithm, specifically overfitting and noise capture. The shortcomings of the simple algorithm arise from growing a single tree for the entire data set. Bagging can combat this as it consists in growing several trees from different subsamples of the data while keeping the same distribution. As a result, predictions on unseen data are estimated by averaging the results of individual trees.

Later on, (Breiman, 2001) suggests a modification in the Bagging algorithm. The author argues that the strength and the correlation of the individual trees determine the generalization error of a forest of tree regressors. Thus, to optimize the trade-off between variance and Bias it introduces randomness to the tree building process. This addition results in the Random Forest algorithm, where each node can only be split over a random subset of features. The predictive power of the algorithm increases as the randomization leads to the decrease of both the correlation between trees and the individual predictive power of each one.

The Gradient Boosting Machine algorithm is similar to the Random Forest one as both are ensembles of classification or regression trees. However, the GBM is able to reduce the randomness introduced in the tree construction by progressively rectifying the errors in the preceding trees. The inaccuracies, also referred to as pseudo-residuals, are minimized using a gradient loss function which adjusts the predicted values at each training data point. Moreover, Gradient Boosting Machine tend to grow more shallow trees and stack them, contrary to RF which build trees in a parallel manner. However, stacking multiple models can increase computational time and poses a greater risk of capturing noise which then can lead to overfitting.

Nevertheless, it is a widely used model favored for its ability to handle complex interactions and yield accurate results.

2.4. Promotions

As mentioned before, nowadays promotions represent a very significant portion of grocery retail sales (Mendes & Pereira, 2021). Promotions, its many instruments and varying impacts, have been extensively studied in marketing and econometric research.

These instruments relate to more than just the type of price cut that is applied, which can go from a direct discount of a certain percentage of the original price, to the buy X get Y free deals. They are also related to the feature advertising used, which can include television ads, flyers or personalized messages, and to the way and where products are displayed in the stores (Fields 2022).

Each of the different aspects of the above mentioned can have varying effects on consumer behavior and, in turn, varying effects on promotion effectiveness which translates into promotional sales as it is presented by (Ailawadi et al., 2006).

One noteworthy promotional effect is switching or cannibalization that translate in an increase of sales of one product at the cost of a decrease in the sales of other. Switching and cannibalization differ as cannibalization usually refers to the decline of sales followed by the introduction of a new product while switching is not necessarily caused by the introduction of a new product (Shah & Avittathur, 2007).

Additionally, many authors have attributed part of the promotional effectiveness to purchase acceleration (van Heerde et al., 2000). Increase in quantity bought and anticipation of the purchase are at the base of stockpiling. This effect represents the consumer decision to acquire larger quantities than usual of a certain product with the implication of delaying the subsequent purchase. Stockpiling can be translated into patterns of post-promotional dips seen in promotional sales data (Ailawadi et al., 2006; van Heerde et al., 2000).

(Bell et al., 1999) recognizes in their research that the sales increase caused by a promotion can be explained by primary or secondary demand effects. Primary demand effects are related to purchase incidence and purchase quantities, whereas secondary demand effects are related to brand switching. However, (Auer & Papies, 2020) in their recent revision concluded that nowadays additional factors other than brand switching, such as purchase acceleration and stockpiling, have a high impact on demand in response to a price change.

Moreover, it was noted that the majority of the competition, especially between brands, stems from similarly priced items and less from brands or products with very different price ranges. It is also noteworthy the fact that in categories of products with lower stockpiling abilities there is switching behavior, or substitution effects, and it is more common to observe a tendency to postpone purchases (Auer & Papies, 2020).

With the goal of quantifying the impact in terms of profit that promotional actions have in a retail context (Ailawadi et al., 2006) study promotional lift with a regression-based methodology. The authors propose to quantify promotional lift decomposing it into incremental lift, switching and stockpiling. Additionally, they advance an estimation of halo effects within a store, which refers to the positive effect some promotional efforts have in the perception of the entire store and therefore increasing overall sales. Finally, they examine how the net impact is affected by promotion, brand, category and store characteristics.

2.5. Practical Applications

In order to tackle identified limitations with regression-based models specifically concerning large dimensionality and multicollinearity, (Trapero et al., 2015) suggests a novel model that incorporates Principal Component Analysis (PCA). The developed approach leverages PCA to reduce the size of the explanatory variable space and multicollinearity while identifying various demand dynamics. Additionally, the use of pooled regression allows for the estimating points where the data does not have enough information on promotional weeks. Moreover, by minimizing the Schwarz Bayesian Criterion they are able to automatically identify the function

that the error term is modeled as. Finally, all of these components are integrated as a dynamic regression that obtains the most accurate predictions when compared to benchmark models as exponential smoothing. It is also stressed that the dynamic regression presented, besides yielding the lowest forecasting error, also does not require expert knowledge to be applied in real settings as is the case with more complex methodologies such as Machine Learning (ML).

(Arunraj & Ahrens, 2015) propose to develop a hybrid ARIMA model that incorporates seasonality and external features, the SARIMAX model. The goal is to accurately estimate volatile skewed sales data that varies in time. The process started by first developing the SARIMA portion of the approach followed by the use of linear regression to combine it with key factors that impact demand. However, this SARIMAX with multiple linear regression (SARIMA-MLR) was concluded to be prone to low accuracy in the context of sparse sales as it only produces mean forecasts. To overcome this, it is advanced a hybrid model including quantile regression, the SARIMA-QL method. The authors conclude that both the SARIMA-QR and SARIMA- MLR yield better results than seasonal naïve approaches and the simple SARIMA model.

Similarly, (Abolghasemi et al., 2020) find that including promotional relevant information to the estimation process has a positive impact. They present a hybrid model, composed of an ARIMA and a piecewise regression, that characterizes demand with two factors: baseline demand and promotional demand. This hybrid approach is able to accurately reflect the significant differences between non-promotional and promotional periods and even differences within the latter, as is shown in the results. Applied to a context with various levels of volatility in demand the hybrid approach outperformed methods such as exponential smoothing with covariates. Nevertheless, the authors discuss that other approaches such as SVR and Dynamic Linear Regression are also able to obtain robust predictions in the same demand environment. The note that is stressed the most is the importance of decomposition of volatile demand series.

With the growth in popularity of more sophisticated methodologies to model consumer behavior there is more motivation to study them applied to many specific problems and contexts. (Bajari et al., 2015) compares standard econometric models to machine learning ones with the goal of finding solutions that can be efficiently applied by econometricians when

forecasting demand for large sets of data. The authors compare a total of eight approaches, two benchmark models available in commercial forecasting systems and six more complex methods. The regression-based models considered are simple linear regression, conditional logit regression, some regularization techniques, such as stepwise, forward stagewise and lasso regressions. As for machine learning they considered support vector machines, and two tree based approaches, a bagging algorithm and the random forest. Analyzing the results for estimating sales data spanning for six years the main conclusion is that the linear and logit regressions are generally outperformed by the other methods. Furthermore, an important takeaway is that there are benefits in terms of increased accuracy in using one of the six developed approaches instead of the common techniques present in statistical software.

(Ali et al., 2009) develop a sophisticated approach to forecast demand combining the use of regression trees with explicit features engineered to summarize past promotional sales by leveraging marketing research literature. When testing this method to predict sales its performance was compared to that of the three main groups of forecasting techniques that are presented above. As a benchmark is considered simple exponential smoothing, from the regression-based methods is applied the stepwise regression and from the machine learning family is chosen the support vector regression. The main conclusion from this work is that while simple time series methods have a very good performance in regular non-promoted periods, when there are promotions the proposed model can substantially improve prediction accuracy. It is also concluded that data pooling, as was done with marketing research, always has a positive impact on the results.

In a similar way, (Huang et al., 2014) present a linear causal model giving emphasis to the importance of having competitive information reflected in the explanatory variables and to the feature selection process. Specifically, grocery retail sales are forecasted at the product level by first selecting the most competitive features using LASSO and stepwise selection then using a factor analysis to finalize on the relevant representative factors. This information is then incorporated in an Autoregressive Distributed Lag (ADL) model, an econometric regression-based model estimated by Ordinary Least Squares. In the end, the ADL model, that was estimated both considering only own product information and considering cross product

information, was compared to two benchmark methods, simple exponential smoothing and base-times-lift. The ADL model outperforms both benchmarks and the one considering other product related variables, the ADL-DI, improves accuracy compared to the own product model.

(Ma et al., 2016) propose to incorporate intra and inter category promotional information into promotional sales forecasting. They developed a four-step framework where the first step is to find the most influential categories, followed by building the explanatory variable space. The third step is to estimate a LASSO regression to perform feature selection. The final step is to generate sales forecasts, leveraging a rolling scheme. In this specific case, the authors use a LASSO Granger causality in order to capture category level promotional interactions. It is shown, with the improvement in forecasting accuracy, that the developed methodology, which intakes additional information compared to the benchmark, is valuable. The benchmarks estimated for this study are simple exponential smoothing and base times lift. Furthermore, many versions of the focal ADL model are presented, differing by including own product information, intra or inter category information of the top five selling products followed by intra or inter category information of all products and intra or inter category information considering the categories' principal components following the application of a PCA. In the end, the use of inter- and intra- category features resulted in a significant increase of accuracy, being the majority of it attributed to the addition of intra-category factors.

(Kong, 2015) develops a demand predicting method that makes use of product attribute information to construct similarity features between products and product clusters. These similarity features, along with display and availability effect, time and price indexes, are taken, utilizing a linear regression method, as the inputs to predict sales. From the similarity features regression coefficients is possible to understand complementarity and substitution patterns between different product features. The results showed that the addition of display, availability and cannibalization effects had a positive impact on forecasting accuracy.

In (Hewage & Perera, 2022) the goal is to highlight the presence of different periods, namely the normal, the promotional and the post-promotional periods. With this purpose they analyzed and compared multiple models within machine learning and traditional forecasting methods. Grocery sales are forecasted at the product level for all the periods using base-times-lift

combined with exponential smoothing with promotional and post-promotional effect's adjustments as a benchmark, using ARIMA as univariate approach and using LightGBM, XGB and Random Forest as machine learning. The main conclusion goes in line with the literature, that is, the use of machine learning techniques does improve forecasting accuracy when compared to univariate methods. Conversely, the authors note that the benchmark model used can have a similar performance to the more advanced methods if an extra effort is put into cleaning and preparing the data.

Likewise, the work of (van Heerde et al., 2000) focuses on estimating pre and post promotion dips. Following the definition of stockpiling effect, that is an increase in purchased quantity together with anticipation of purchase timing, experts expected to clearly identify a dip in sales following a promotional action, which was not always true. The authors study this seemingly paradoxical situation presenting a time series approach and an econometric one that is a variation of the SCAN*PRO model. The time series method uses an ARIMA model and both a transfer function and an intervention model in order to establish the relation between the predictors and the target variable. The econometric one includes variables that account for dynamic price effects, that capture lead and lagged price indexes. The main findings of the work are that there are significant dynamic promotional effects, post-promotional dips do exist and can be found using the proposed approach with the expected magnitude and that accounting for purchase acceleration and display extension effects does improve forecasting accuracy.

3. Project Description

The initial section of the chapter discusses the retailer's existing promotional demand forecasting methodology, that relies on the Oracle Retail Demand Forecasting (RDF) software. Additionally, the chapter highlights certain limitations associated with this approach, that will be further explored to identify areas where improvements can be made.

The subsequent portion of the chapter focuses on outlining the main objectives of the project and on providing a concise overview of the approach chosen to address the appointed challenges and limitations. The primary goal is not only to surpass the results achieved through the existing methodology, but it also aims at diving deeper into the analysis of promotional effects that currently are not being explicitly addressed. Of particular interest are the examination of cannibalization and substitution effects and the pre and post promotional dips in sales, which can significantly impact the overall demand. In the end, the project proposes a Forecasting Support System for each category to predict promotional sales, at the product-store-week granularity level, that is able to capture within-category interaction effects. It is important to stress that a more comprehensive explanation of the methodology employed will be presented in the following chapter.

Moreover, this chapter provides a comprehensive description of the available data leveraged for the development of this master thesis. The dataset encompasses sales data spanning from November 2021 to October 2022, ensuring a robust foundation for the analysis and evaluation of promotional forecasting models. By carefully selecting and preprocessing the data, the proposed approach aims to provide a comprehensive understanding of the promotional context, ultimately yielding results that align with the retailer's specific concerns and objectives.

3.1. Current Methodology

This section first describes the methodology that is currently being used in the retailer universe. That is a combination of the use of a commercial forecasting system and an external selection model developed in a previous project at LTPlabs. In the section are then advanced some of the

limitations of the approach. These limitations have been, in part, identified by the client themselves, while the remaining have been presented by the team responsible for the previous project.

3.1.1. Description of the current methodology

At the moment the company uses a combination of software predictions – generated by Oracle Retail Demand Forecast[1] - and an external model to produce their promotional demand forecasts. The process begins with generating a sales baseline. This involves processing the regular sales historic data by removing outliers, stock-out situations and promotional situations. Then three different predictions are generated, after which all of them, along with some other key features, are fed into a machine learning model. This last step will indicate which of the three predictions is best to apply to each specific situation.

The promotional forecasting methods applied by RDF are:

1. Promotional lift forecasting. Here the process starts by using their software methodology to predict regular sales. After this a promotional lift factor is applied resulting in promotional sales predictions.
2. Enhanced baseline. This is the second lift method, where, this time, the lift factor is applied to the average regular sales. As the method is based on recent sales history data it is ideal for products that have fluctuating and volatile regular demand predictions.
3. Average promotional sale. The last approach considers only promotional week's data and computes the average sale of the product. For products that are very heavily promoted and for which it is difficult to calculate promotional lifts this approach is best.

The mentioned lift is calculated as the ratio between promotional sales and baseline. Looking at the past fifty-two weeks' information this ratio is calculated at the product level and at an aggregated level, such as the subcategory. The final lift is a result of the proportion of 80% from the product level and 20% from the aggregated level. The difference between the two first

[1] This information was in part collected from Oracle's public documentation and additional information specific to this particular case was shared by the client. Nevertheless, some specificities are confidential information (Oracle, 2010).

methods lies in the baseline sales to which the lift factor is applied to. In the first case the baseline is predicted for the period in question using the Oracle's forecasting system methodology for regular demand prediction. For the second case the baseline sales are estimated as the average regular sales of the previous fifty-two weeks.

Regarding the method selection model, when choosing the most appropriate approach it considers the trade-off between accuracy and balance between over- and under-forecast, that is referred to as the Bias. It evaluates the prediction results from the three methods, while also taking into account additional relevant factors. The features considered focus on different estimations of baseline and promotional sales relative to different sale's aggregations, on some promotional information, as the number of days that each promotional action lasts for and the relative price cuts, and on seasonality indexes. In the end the model chooses the approach that can yield the most accurate result pondered by the Bias.

3.1.2. Shortcomings of the current methodology

In the context of this retailer, one of the major drawbacks with the current methodology is the lack of use of contextual and promotional specific variables, besides the discount percentage. This hinders a deep understanding of the promotional demand landscape.

A second noteworthy concern is the persistent tendency to under-forecast, which means that the predicted promotional sales fall short of the actual promotional sales. This potentially results in more frequent stock-out situations, loss of revenue, and poorer customer satisfaction. Nonetheless this tendency can be considered an inherent characteristic of the context and available data, and should, therefore, be considered when analyzing and comparing results.

There have been appointed by the retailer four main causes for these issues:

- Low reactivity: the available predicting methodologies are somewhat stable. This can be attributed to the vast history of data on which the methods are based on and on the absence of promotional effect's features. Therefore, these methods are not always the best at reacting to variations in purchasing patterns.

- Virtual promotions: there are products that are on promotion almost every week, which affects the preprocessing steps of estimating baseline sales considering non-promotional weeks, and therefore the promotional prediction results.
- Best approach choosing model: recently it has been observed that the method selection model has been favoring the average promotional sale approach, which may not always be the best one. For instance, the average method would not be the one expected to give the most accurate prevision for products with very seasonal demands.
- Existence of outliers: there are atypical sales for specific situations that are not correctly being treated as outliers that then have significant prediction errors, thus possibly clouding the analysis of the evaluation metrics. This issue is more prevalent especially on slow moving products.

3.2. Proposed Approach

Following the brief presentation of the process that the retailer undergoes to generate sales predictions and its main limitations, this section now presents the main goals and scope of the project.

As stated before, the main goal is to develop a forecasting system focusing on promotional periods that has the ability to provide better results, in terms of accuracy, than the ones currently being obtained. As a way to reach this main goal, a focal point of the project is the construction and incorporation of features that translate the promotional landscape. Namely, being able to have a deeper understanding of cannibalization and pre and post promotional dips.

First, it is necessary to note what is considered, for the sake of this project, as better results. As one of the major concerns for the client is stock-out occurrences, it is deemed that it is preferable to have over-forecast situations rather than under-forecast. Over-forecasting refers to the situation where the predicted sales exceed the actual demand, which can allow for better inventory management when stock-outs are the major concern. Conversely, under-forecasting refers to the situation where the predicted sales fall short from the actual demand, increasing the

risk of having insufficient stock to meet the demand. This, as presented in the previous section, is one of the main issues with the current methods.

Secondly it is noteworthy that there is a preference for accuracy rather than interpretability. Even though this trade-off is kept in mind during the development of the project, this fact grants the opportunity to leverage more complex and powerful yet black-box machine learning techniques.

Nonetheless, this dissertation does not aim to capture cross-category effects, but rather capture within-category ones explicitly through feature engineering or model design. Hence, it proposes a Forecasting Support System for each category to forecast sales at the Product-Store-Week granularity level.

The first approach explored is a regression method that can be described as an adaptation of the model presented by (Kong, 2015) following insights taken from the SCAN*PRO model developed by (Wittink et al., 1998) and its variation proposed by (Batista, 2016). However, it is important to note that a regression approach is not expected to yield better results than the methodology currently used by the retailer. As suggested in research, in scenarios with high promotional activity and variability in promotional sales, machine learning techniques are expected to provide more accurate predictions. Thus, the primary purpose of the regression model is to serve as a proof of concept for some of the feature engineering steps.

The second methodology proposed focuses on machine learning techniques, namely Random Forest and Gradient Boosting Regressors. These methods will include the variables constructed and tested in the first step and are expected to be serious competitors to the current methodology employed by the retailer.

3.3. Available Data

The data set utilized for this project and dissertation provides a comprehensive overview of the retailer's sales data. It spans from November 2021 to October 2022 and encompasses five key aspects.

- Store hierarchical structure: provides information about three store formats within the business. Whitin the three formats there are differences in the number of stores, locations, size of the stores and available store assortment.
- Product hierarchical structure: comprised of three relevant levels namely category, subcategory and unit base.
- Stock information: presented by a daily, product-level flag variable that indicates the occurrence of a stock-out.
- End-of-day sales volume: presented regarding both regular and promotional sales volume.
- Daily promotional information: provided at the product-store level which includes details such as the last day of the week when the promotion occurred, the respective duration, promotion type, and location, communication vehicle and discount information including type and percentage. This information is schematized in more detail in the Appendix A.

Given the vast amount of data available, a focused approach was adopted to manage the analysis in a more effective way, that will still allow the methodology to be expanded later. For this reason, four categories were chosen that are believed to provide a representative sample of varying levels of promotional intensity and for exploring different promotional characteristics within each category.

Additionally, due to the limited information regarding the exact day when a promotion occurs an assumption was made related to the promotional information to be considered. Only promotional actions that lasted for exactly 7 days were taken into account. While this assumption could be seen as a limitation by not capturing shorter duration campaigns, the overall impact on the analysis is minimal given that promotions lasting seven days represent over 90% of the data.

Regarding the stock-out flag variable it is important to acknowledge that it can be the cause for some distortions. On the one hand, there are instances when the stock-out flag is activated although there are still registered sales for the day. This indicates that the stock rupture occurred during the course of the day. Seeing as these registered sales may not be an accurate

representation of demand it was decided that only weeks with no stock-out flag for any of the seven days would be considered. On the other hand, it is possible that in some situations the stock level is very low but because it does not reach zero the flag is not activated. Again, these situations can be inaccurate depictions of true demand, however, due to the lack of store assortment information no particular action was taken in this case.

One of the objectives of the dissertation is to effectively model promotional effects, besides accurately predicting promotional sales. For this reason, it was important to select a subset of categories that would allow to test the developed methodology in different promotional and general demand landscapes. This could ultimately be pivotal to the later employment of the approach to the entirety of the retailer universe. The four categories selected are: Beers and Ciders, Liquid Fats, Eggs and Candy and Chocolate.

Therefore, the four focal categories represent both heavily promoted products, and ones with low promotional activity. Related to the promotional effects aspect, there was an effort to include representation of the two main effects mentioned, cannibalization or substitution and stockpiling, which was done for various levels of each. The diversity of products within the categories was also accounted for, thus the consideration of categories with a very homogenous set of products and ones with a wider variety of products. Lastly, the categories chosen also depict various levels of seasonality. Table 3.1 below summarizes the positioning of each category related to the mentioned aspects.

Table 3.1 - Position of each category related to relevant factors

Category	Heavily Promoted	Substitution	Stockpiling	Product Homogeneity	Seasonality
Beers and Ciders	●	●	◑	◑	◑
Liquid Fats	●	●	●	◕	○
Eggs	○	◔	○	●	○
Candy and Chocolate	○	○	○	○	●
Level:	●High	◕Medium High	◑Medium	◔Medium Low	○Low

Promotional intensity for the purpose of choosing the relevant categories to consider was defined following the work of (Huang et al., 2014). In their work a category was labeled as heavily promoted if for more than 25% of the weeks in the period the average number of existing unique products was on promotion. Although, due to the highly discount dependent retail landscape observed in Portugal, the value of 25% was raised to 30%. Table 3.2 presents the percentage of weeks in the considered period that had the average number of products on promotion for each category.

Table 3.2 - Percentage of weeks of the period with the average number of products on promotion

	Percentage of weeks
Beers and ciders	40%
Liquid fats	45%
Eggs	13%
Candy and chocolates	10%

Furthermore, the number of distinct products belonging to each of the selected categories varies considerably from category to category and their distribution across the store types is not homogenous. This information is described in Table 3.3 below.

Table 3.3 - Number of unique products in each category and Percentage of products present per store type

	Nº of unique products	Store Type		
		L	M	S
Beers and ciders	321	90%	88%	75%
Liquid fats	185	89%	86%	81%
Eggs	63	65%	92%	68%
Candy and chocolates	1942	89%	92%	71%

4. Methodology

The present chapter focuses on presenting the methodology related to the proposed approach to tackle promotional forecasting. It begins by covering the pre-processing steps taken, outlining the baseline estimation and the cleaning process, in order to have data that includes only promotional sales corrected from outliers. It also describes the ABC classification procedure that relies on the promotional sales lift.

Additionally, it describes the treatment and computation of the input variables. These include the analysis of product data to obtain relevant product attributes later used to cluster products based on attribute similarity. Furthermore, this chapter presents the construction of specific variables aimed at capturing product interactions and promotional dynamics as cannibalization and stockpiling.

This chapter then delves deeper into the techniques applied that can be categorized into two main groups: regression and machine learning. Within the machine learning techniques were developed the Random Forest and Gradient Boosting Regressor algorithms. Moreover, the hyperparameter tuning techniques are detailed, which incorporates a Time-Series Nested Cross-Validation (TS-N-CV). Finally, are introduced the relevant evaluation metrics to be considered in the following chapter.

4.1. Preprocessing

As explained in chapter 3, the data considered for the dissertation is aggregated at the Stock Keeping Unit (SKU)-store-week level. The data concerns solely fully promotional weeks, or fully non-promotional weeks and in both cases during the considered weeks there were no stock-out flags activated on any of the days. The baseline feature described next had been previously calculated at the daily level thus it had to be aggregated at the week granularity so that it could be considered. At this aggregation stage instances of weeks with both promotional and non-promotional days and with stock-outs were disregarded. The remaining sections focus on the

necessary cleaning process that was applied to the focal target variable and an additional key predictor and on the SKU classification step.

4.1.1. Baseline

To analyze the impact of promotions, the first step was to construct a baseline which represents the demand without any promotional influences at the SKU-store-week level. The estimation of this input resulted from a previous project within the company developed for the same retailer client. Constructing this baseline involves processing the available sales data and making necessary corrections. Therefore, the key aspects are addressing stock-out and promotional flags. Whenever there is a stock-out flag it is necessary to address the fact that the sales that might have happened are an under-representation of true demand, which the baseline needs to reflect. The construction of an appropriate baseline for days with promotions poses a greater challenge. The objective is to capture the demand, unaffected by promotional up-lift, cannibalization, pre and post promotional dips, or any other promotional effects. While the detailed methodology for constructing this baseline is beyond the scope of this dissertation, it is crucial to provide a brief overview, as it serves as a key input for the developed models.

Four main steps went in the estimation of the baseline. The first one was to calculate the share of sales that each day of the week represents in relation to total weekly sales. This calculation is based on information of weeks with complete non-promotional and no stock-out data. The following two steps focused on the weeks that had at least one non-promotional day and the goal was to estimate the daily baseline for the promotional days. The process starts by constructing a corrected weekly sum of the regular sales with the goal of having a first estimate of the weekly baseline sales for weeks that had both promotional and non-promotional sales. This is then multiplied the daily sales' share of each day of the week in order to obtain the final estimate of the daily baseline sales. Lastly, the focus was on the full promotional weeks. For these cases the baseline was calculated considering only full non-promotional weeks by running a 5-week exponential moving average, except for the occasions of very frequently promoted products, to which a 52-week exponential moving average was applied.

4.1.2. Cleaning

A key task that must be performed before proceeding to the feature engineering, modeling and prediction processes is ensuring that the data that will be used is reliable. One of the crucial preliminary steps was cleaning the data and removing some possible outliers. Related to this aspect, two variables were prioritized due to their relevance to the focal problem, the discount percentage and the promotional sales amount.

Firstly, the discount percentage was limited to the valid range that had been instilled by the retailer. Therefore, all instances below 5% and above 95% were identified as errors and thus disregarded. Nevertheless, this correction eliminated no more than 4% of the observations which represented, at most, a reduction of 8.5% in terms of promotional sales volume.

Secondly, some adjustments were made to the sales volume features, which proved to be highly impactful, though expected given their importance. Thus, to the baseline sales was imposed a lower limit of 1 unit, and to the promotional sales amount the lower limit of the median baseline level, that was calculated for each SKU-store pair. Following this step on average 74% of all observations are retained which represents around 84% of total promotional sales.

Both these steps were crucial as they contributed to the reduction of the noise introduced in the model, thereby contributing to prediction accuracy. Addressing the possible outliers and errors grants a more solid foundation for the next steps, which is proved by analyzing the computed variables dependent on these two. Additionally, at first the methodology was tested without these preprocessing steps, which yielded worst prediction results.

4.1.3. Classification

As will be further explained alongside the results discussion, the wide range of promotional sales and the disparity between sales levels across different stores introduces some challenges for achieving accurate predictions. To account for this fact, a simple classification approach was used for each SKU-store pair with the goal of categorizing products into groups with similar promotional behaviors.

Initially an ABC analysis, which classifies products according to their sales volume, was considered. The basis of this analysis is that there is a set of products, the A's, that account for around 70% of the total sales volume. This group often includes fewer products, especially compared to B and C products that usually group a large number of items yet only account for 25% and 5% of the total sales volume, respectively. The premise is that the items that do not have a significant contribution to the overall volume are distinguished from the ones that can have a major impact.

Nonetheless, this technique was deemed to not have the ability to capture the behavior of each SKU when under a promotional campaign. Practically, two products with the same end of the week promotional sales amount would have the same ABC classification, however having the same volume does not imply that promotional efforts are equally as effective in both situations.

As a result, the thought process behind the classification applied is the same as the ABC analysis, but the average promotional lift was considered instead of the promotional sales volume. The first step is to calculate the average lift for each product in each store, and then sort these values in descending order. The lift factor is the result of dividing the promotional sales volume by the baseline sales. This follows the basis presented for the base-times-lift approaches presented in chapter 2. Next the cumulative sum of the average lifts is compared with the total average promotional lift of each store. Finally, the products that represent up to 70% of the total average promotional lift of the respective store are classified as A products, the ones that represent the following 25% are classified as B and the remaining ones as C.

4.2. Input Variables

For both regression and machine learning models the variables considered as promotional demand predictors are generally the same. There are 4 main sets of variables:

- Demand related variables which include both the baseline and promotional sales volume.
- Variables that classify and group products based on attributes or promotional up-lift level and that classify the time of the year.

- Variables that characterize the promotional action that was applied.
- Variables that capture promotional effects, specifically cannibalization and stockpiling effects.

The first set of variables have been described in section 4.1.1 for the baseline, while section 4.1.2 describes the cleaning process applied to the promotional sales. Similarly, from the second set of variables the ABC classification is explored in section 4.1.3, while the third set of variables was introduced in chapter 3 and is further detailed in Appendix A. The following section focuses on the products attributes, mentioned in the second set, and on the last set of variables. After the detailed descriptions of all variables and their calculation process the input variables are summarized in Table 4.3.

4.2.1. Product attributes

Following the work of (Kong, 2015) and expert's business sense, similarity between products is acknowledged as one of the main drivers for within category substitution. This fact spurred the necessity to have attributes that not only characterize each product but also empower the clustering of similar products.

Even though a product hierarchy is provided, consisting of a subcategory and unit-base, these two levels are not enough to depict a product, moreover, there were occasions with misclassifications, which made it essential to analyze and treat this data. In the end, for each category, three product attributes were retrieved from the SKU descriptions and is important to note that the brand is considered an additional key aspect in defining each product.

Beers and Ciders

- Type: related to the type of beer, being blond or stout, national or international, also related to the type of cider and related to the alcohol content.
- Size: ranges from mini to large and includes other less common sizes or formats.
- Package: distinguishes between cans and bottles.

Liquid Fats

- Type: distinguished between olive oil and vegetable oil.
- Subtype: goes into more detail regarding each type of oil, such as extra virgin olive oil, deep frying oil, and so on.
- Size: discriminates three sizes of bottles.

Eggs

- Type: includes free range, organic, cage, and so on.
- Size: specifies the sizes from small, medium or large to extra large.
- Quantity: differentiates between the amounts of eggs sold such as half a dozen, one dozen, and so on.

Chocolate and Sweets

- Type: related to the type of sweet, varies from chocolate, to chewing gum, to gummy candy.
- Format: characterizes each type of sweet with categories as bonbon or tablet for chocolate, mint or fruit flavor for chewing gum, and so on.
- Family: identifies the sort of product, encompassing children sweets, seasonal items, impulse products sold only at the checkout and regular sweets.

Table 4.1 presents the number of distinct varieties of each product attribute per category.

Table 4.1 - Number of varieties in each product attribute

	Product Attribute		Number of Varieties
Beers and Ciders	1	Type	6
	2	Size	4
	3	Package	2
Liquid Fats	1	Sub type	11
	2	Size	3
	3	Type	2
Eggs	1	Type	4
	2	Size	4
	3	Quantity	3
Chocolate and Sweets	1	Format	8
	2	Family	5
	3	Type	5

Following the extraction of product attributes, it was possible to proceed to clustering products. The goal of this clustering is to group together similar products which helps understand and study relationships and interactions of similar products versus interactions between distinct products. For clustering purposes some techniques such as K-Means and hierarchical clustering methods were considered. However, in spite of their ability to effectively pick out patterns within the data the end result did not accurately portray groups of products that were expected by expert business sense to elicit similar consumer behavior. The intricate understanding of how

consumers are anticipated to respond to specific attributes or situations is better discerned through human knowledge than solely based on sales data, at least without further information on transaction data that allows for more in-depth basket analysis.

On the contrary, seeing as the product attributes were already engineered with the purpose of aggregating products this expected behavior was kept in mind. Consequently, grouping products based on their specific attributes yielded better results than doing so with the use of clustering algorithms. The first approach was, per category, to simply group products based on the three attributes, however after analyzing what constituted each cluster it was concluded that it was better to consider grouping based on only two of the attributes. This change happened due to the lack of unique products belonging to each cluster and without it the calculation of the new features explained below would have been heavily impacted. The comparison between the numbers of clusters and single product clusters is in Table 4.2 below and the resulting clusters for each category are discriminated in Appendix B.

Table 4.2 - Number of clusters and single product clusters per category

	Using 3 attributes		Using 2 attributes	
	Nº of clusters	Nº of single product clusters	Nº of clusters	Nº of single product clusters
Beers and Ciders	29	4	12	0
Liquid Fats	27	2	6	0
Eggs	17	7	9	3
Chocolate and Sweets	32	2	11	0

4.2.2. Promotional effects variables

The phenomenon of having a sales increase at the cost of sales decrease in a different product can be observed between both promoted and non-promoted products that can belong or not to the same group or category. For this dissertation, cannibalization was only analyzed between products within the same category that are under a promotional campaign at the same time in the same store. Following the work of (Kong, 2015), two specific variables are computed to try and capture this effect.

The first feature concerns the promotional sales quota of other products that share the same characteristic as the product under analysis. The premise behind it is that the sales of a product are higher (lower) when the remaining products on promotion that share the same characteristic have a low (high) promotional sales quota.

$$Qo_{i,t,s,w} = \frac{APS_{t,s,w} - APS_i}{APS_{t,s,w}}$$

Where:

$Qo_{i,t,s,w}$ = Promotional sales quota of the other SKUs with the same variety of attribute *t* as product *i* in store s in week *w*

$APS_{t,s,w} = \sum_{\substack{i \in I \\ feature(i)=t \\ store(i)=s \\ week(i)=w}} APS_i$ = Average promotional sales of all SKUs with the same variety of attribute *t* in store *s* in week *w*

APS_i = Average promotional sales of product *i*

This feature translates that when on a promotion a product will most likely cannibalize sales of similar products if they represent a relatively small share of sales. As a predictor this was estimated with *t* being the cluster, the brand, and each product attribute, therefore five distinct variables were computed.

The rationale for the second variable is similar to the one described above but it is centered around the discount percentage. Seeing as price elasticities or price indexes could not be

estimated due to the lack of price information, the closest relation was the percentage of the price reduction applied. Here the hypothesis is that sales will increase (decrease) when the average discount percentage of the other similar products is low (high).

$$ADo_{i,t,s,w} = \frac{AD_{t,s,w} * np_{t,s,w} - D_{i,s,w}}{np_{t,s,w} - 1}$$

Where:

$ADo_{i,t,s,w}$ = Average discount percentage of the other SKUs with the same variety of attribute *t* as product *i* in store *s* in week *w*

$np_{t,s,w}$ = number of promoted SKUs with variety of attribute *t* in store *s* in weeks *w*

$AD_{t,s,w} = \frac{1}{n_{t,s,w}} \sum_{\substack{i \in I \\ feature(i)=t \\ store(i)=s \\ week(i)=w}} D_i$ = Average discount percentage of all SKUs with the same variety of attribute *t* in store *s* in week *w*

$D_{i,s,w}$ = Discount percentage of product *i* in store *s* in week *w*

Again, this predictor was calculated for the cluster, brand and product attribute level.

In order to account for the non-promoted products that are also being sold in each store each week the following variable was also calculated. It represents the percentage of similar products sold under a promotional action in each store in each week.

$$Pip_{t,s,w} = \frac{np_{t,s,w}}{n_{t,s,w}}$$

Where:

$Pip_{t,s,w}$ = Percentage of SKUs in promotion with the same variety of attribute *t* in store *s* in week *w*

$np_{t,s,w}$ = Number of promoted SKUs with variety of attribute *t* in store *s* in week *w*

$n_{t,s,w}$ = Number of SKUs with variety of attribute *t* in store *s* in week *w*

This variable was calculated for the three product attributes and the brand.

The second effect that is meant to be captured is stockpiling, although the approach described is a preliminary and surface level analysis. This is a complex effect that is very heavily intertwined with consumer behavior. It refers to the purposeful decision of purchasing larger amounts of a product, usually because it is on promotion. Stockpiling is translated by a substantial increase in sales followed by an accentuated decrease, which happens because consumers will take some time before repurchasing the same product. Additionally, there's an associated event of hold-out that occurs due to customers postponing their purchases until the product reaches a desired price. This phenomenon is represented by a lower sales volume than usual before the campaign.

To capture this effect two more features were computed. The first of which indicates the number of weeks since the product had last been on promotion on that store. This variable conveys that a recently promoted product is less likely to have a strong customer reaction while products that exhibit longer intervals between promotions are expected to have more significant sales increases.

$$wp_{i,s,w} = w - w_{-1}$$

Where:

$wp_{i,s,w}$ = Number of weeks since the last promotional action

w = Current promotional week in analysis

w_{-1} = Previous promotional week

The second variable, the percentage average promotional sales from the last promotion, was engineered with the purpose of capturing that the sales from the current promotional campaign are expected to be lower if the sales during the last promotion were above average. Conversely, if during the previous marketing effort were observed below average sale's volumes is expected that the present volume is at least average. This variable is computed according to the following formula:

$$PSp_{\ i,s,w} = \frac{PS_{i,s,w-wp}}{APS_{i,s}}$$

Where:

$PSp_{i,s,w-wp}$ = Percentage average promotional sales of product *i* in store *s* on the previous promotional week

$w - wp$ = Last promotional week

$PS_{i,s,w}$ = Promotional sales of product *i* in store *s* in week *w*

$APS_{i,s}$ = Average promotional sales of product *i* in store *s*

However, a correction is applied to both variables seeing as the stockpiling effect, as it has been defined, is only present if the last promotional campaign happened somewhat recently. Therefore, an upper limit was set on $wl_{i,s,w}$ of five weeks and for those cases $PSp_{i,s,w-wp}$ was set to zero. This follows the assumption that if the last promotion occurred more than five weeks ago then the observed purchasing patterns are regular and not influenced by the stockpiling effect. The upper bound of five was defined by the client, following an internal analysis on the occurrence of stockpiling behavior.

Table 4.3 - Description of the available variables

Variable		Description
Demand Variables		
Promotional sales	$PS_{i,s,w}$	Target feature that indicates the promotional sales volume.
Baseline sales	$RS_{i,s,w}$	Baseline sales volume.
Baseline sales lag	$RS_{i,s,w-1}$	Baseline sales volume of the previous week.
Classification Variables		
ABC classification	$L_{i,s}$	Classification feature based on the promotional uplift; products with high up-lift are classified as A, with medium up-lift are classified as B and with low up-lift are classified as C.
Cluster	C_i	Cluster to which each product belongs to based on two product attributes.
Trimester of the year	M_w	Divides the year into 4 categories one for each trimester.
Product attributes	A_1 , A_2 , A_3	Each variable represents one of the constricted product attributes
Brand	B_i	Brand of the product

Promotional information variables

Promotion type	$Pt_{i,s,w}$	Promotional tag that characterizes each promotional action.[2]
Communication vehicle	$Pv_{i,s,w}$	
Promotion location	$Pl_{i,s,w}$	
Discount type	$Dt_{i,s,w}$	
Discount percentage	$D_{i,s,w}$	Percentage of the price reduction applied.

Cannibalization variables

Promotional quota of other SKUs	$Qo_{i,t,s,w}$	Promotional quota of other products that share the same attribute t. t can be the cluster, product attribute 1, 2, or 3, or the brand.
Average discount of other SKUs	$ADo_{i,t,s,w}$	Average discount percentage of other products that share the same attribute t. t can be the cluster, product attribute 1, 2, or 3, or the brand.
Percentage of products in promotion	$Pip_{t,s,w}$	Reflects the percentage of products sharing the same variety of attribute that are promoted at the same time in each store each week.

[2] The different promotional types, locations, communication vehicles and discount types are further described in Appendix A.

Stockpiling variables

Weeks since last promotional action	$wp_{i,s,w}$	Number of weeks that have passed since the last promotional action. This variable ranges from 0 to 5.
Percentage of average sales of the last promotional action	$PSp_{i,s,w-wp}$	The percentage of the average promotional sales indicates if during the previous promotional action the product' sales volume was above or below the average promotional sales of the product.

4.3. Regression Approach

The regression model developed can be seen as an extension of the SCAN*PRO model that also builds on the one presented by (Batista, 2016). The SCAN*PRO is based on the premise that promotional sales are affected by prices, that are captured using different indexes, by the display in the store and by different marketing efforts. This can be summarized as promotional sales being influenced by price and demand related factors, and promotional related factors.

Here the rationale is that promotional sales are dependent on demand related factors, translated by the baseline and its lags, by promotional characteristics, translated by the promotional tag, and by promotional effects, translated by the cannibalization and stockpiling features. The addition of cannibalization factors takes from the work of (Kong, 2015) as well as the application of a log-linear model. The demand variables were log-transformed in order to stabilize their variance and the logarithmic version of the model can better capture the non-linear relations between variables contributing to the interpretability of this approach.

Furthermore, to account for the heterogeneity between store types one model is estimated for each store type, given that within each type the conditions for each store are assumed to be

somewhat similar. The approach also considers that product homogeneity can be better captured by also estimating one model for each ABC class.

Ultimately, there are three different versions of the regression model, that were calculated for each category-store type-class combination. The first one was estimated only considering the given demand, promotional tag information and the trimester of the year. This is presented in the formula below:

$$log(PS_{i,s,w} + 1) = \beta_0 + \beta_1.log(RS_{i,s,w} + 1) + \beta_2.log(RS_{i,s,w-1} + 1) + \beta_3.Pt_{i,s,w} + \beta_4.Pv_{i,s,w} + \beta_5.Pl_{i,s,w} + \beta_6.Dt_{i,s,w} + \beta_7.D_{i,s,w} + \beta_8.M_w + \varepsilon_{i,s,w}$$

The purpose of this initial model was to establish a base to which the results of adding the engineered features could be compared. This was helpful to evaluate the significance of the created variables and to assess if their behavior accurately portrayed the effects meant to be captured.

The next model was strengthened by incorporating all the computed variables in order to address the desired product interactions, as is presented in the following formula:

$$\begin{aligned} log(PS_{i,s,w} + 1) = {} & \beta_0 + \beta_1.log(RS_{i,s,w} + 1) + \beta_2.log(RS_{i,s,w-1} + 1) + \beta_3.Pt_{i,s,w} \\ & + \beta_4.Pv_{i,s,w} + \beta_5.Pl_{i,s,w} + \beta_6.Dt_{i,s,w} + \beta_7.D_{i,s,w} + \beta_8.M_w + \beta_9.Qo_{i,t_{C_{i'}},s,w} \\ & + \beta_{10}.Qo_{i,t_{A_1},s,w} + \beta_{11}.Qo_{i,t_{A_2},s,w} + \beta_{12}.Qo_{i,t_{A_3},s,w} + \beta_{13}.Qo_{i,t_{B_{i'}},s,w} \\ & + \beta_{14}.ADo_{i,t_{C_{i'}},s,w} + \beta_{15}.ADo_{i,t_{A_1},s,w} + \beta_{15}.ADo_{i,t_{A_2},s,w} + \beta_{16}.ADo_{i,t_{A_3},s,w} \\ & + \beta_{17}.ADo_{i,t_{B_{i'}},s,w} + \beta_{18}.Pip_{t_{A_1},s,w} + \beta_{19}.Pip_{t_{A_2},s,w} + \beta_{20}.Pip_{t_{A_3},s,w} \\ & + \beta_{21}.Pip_{t_{B_{i'}},s,w} + \beta_{22}.wp_{i,s,w} + \beta_{23}.PSp_{i,s,w-wp} + \varepsilon_{i,s,w} \end{aligned}$$

Although the inclusion of the additional predictors has a positive impact on prediction accuracy there can be an optimal combination of parameters. For this reason, a Variance Inflation Factor (VIF) analysis is conducted which addresses multicollinearity and serves as a criterion to eliminate some of the variables. It was determined that variables with VIF value, that are

discriminated in Appendix C, above a certain threshold would be disregarded. The threshold value considered was 5, following (Chan et al., 2022) and a careful analysis of the VIF results of each predictor. In the Appendix Tables C.1 to C.12 are highlighted the variables excluded for each class in each store type. Moreover, in all cases the intercept factor was excluded, as it proved to be a very high variance feature.

4.4. Machine Learning Approach

In contexts with exogenous and endogenous variables and where there are many interactions at play, machine learning methods are promising alternatives to the more common linear methods. These techniques have the capacity to directly derive nonlinear variable relations and capture hidden patterns from the data.

Tree-based methods, as the ones considered, assume independence between observations, which is not true for time-series data. In order to account for this nature of the current problem, the baseline sales as well as its lagged version of one and two weeks are included in both models as features. The remaining variables included are the ones presented in Table 4.3 and the store type information. Additionally, before the model estimation step, highly correlated features are discarded. This collinearity issue is addressed according to Pearson correlation measure, with a threshold of 99%. The eliminated variables are presented in Table 4.4 below.

Table 4.4 - Discarded features per category

	Discarded Features
Beers and Ciders	Attribute_3_can
Liquid Fats	Attribute_3_vegetable oil
Eggs	Atribute_3_2; Attribute_1_free range; Attribute_1_bio
Candy and Chocolate	Attribute_1_chocolate; Attribute_1_bonbon; Attribute_1_tablete; Attribute_1 _sugar

One of the explored techniques is the Random Forest algorithm, which is based on regression trees techniques presented by (Breiman et al., 1984). This method builds regression trees by splitting each 'node' over a random subset of variables. Predictions are then obtained by averaging each tree's results.

As every 'leaf' is constructed by a sequence of splits that depend on multiple features, tree-based algorithms tend to excel at capturing interactions between different features. Nevertheless, growing trees without any restriction can lead to overfitting and to the incorporation of noise. For this reason, it is necessary to define hyperparameters that impose stopping criteria. These relate to:

a) The maximum 'depth' or the number of splits that the tree is allowed to be partitioned in.
b) The minimum number of observations in each leaf.
c) The degree of homogeneity or 'purity' at which the algorithm should stop slitting the data.

The second technique applied is the Gradient Boosting Regressor which leverages the Gradient Boosting algorithm to make predictions of continuous outcomes in a regression task.

The conventional method works by subsequentially adding models that correct inaccuracies resulting from the previous ones. Each model is created by fitting weaker learners, such as regression trees, to the errors of the previous models. As this process is performed iteratively, it gradually leads to the decrease of the overall error of the ensemble. This error is minimized through gradient descent, resulting from the adjustments to the parameters of the week learners.

The GBR applied can be seen as an optimized version of the conventional model as it can reduce computational type and lower memory demand without jeopardizing accuracy. This is an histogram-based regressor that discretizes the continuous feature space into bins. Splits are then made based on these integer bins, leading to a substantial reduction of the number of splits. However, there is a trade-off as prediction precision could be improved if decision nodes were split based on continuous instead of discrete variables (Pedregosa et al., 2011).

Regardless, the stacking nature of these methods makes them more prone to long running times and overfitting. To combat the possible drawbacks, it is necessary to employ proper regularization and tuning techniques.

As previously mentioned, hyperparameter tuning is performed with the purpose of overcoming some of the shortcomings of the presented ML techniques.

Some of the hyperparameters are shared by the RF and GBR due to their tree-based nature. Specifically, these concern the number of trees the algorithms are allowed to grow and the number of randomly selected variables that each tree can fit and the tree depth. Conversely, the learning rate is unique to the Gradient Boosting and the minimum impurity decrease is exclusive to the Random Forest.

The learning rate acts as a regularization parameter as it controls the contribution of each tree. On the one hand, a low learning rate would require a dense number of trees resulting in lower variance. On the other hand, a high learning rate would reduce the training time at the expense of higher over-forecasting incidence.

The minimum impurity decrease determines the stopping criterion for tree growth as node impurity measures the heterogeneity within a node. Setting this parameter ensures that the only splits considered valid are the ones that result in a significant decrease in node impurity. While setting a higher threshold will lead to fewer more generalized splits, a lower threshold value allows for more splits.

During training the hyperparameters, summarized in Table 4.5, are tuned through a grid search that explores different combinations in order to minimize the generalized error and avoid overfitting. The number of trees or estimators can be either 100 or 200, the maximum depth can range from 3 to 11 with a step of 2. The possible values for the minimum impurity decrease are 0.2 and 0.5 and for the learning rate are 0.01 and 0.1.

Table 4.5 - Overview of the hyperparameters

Hyperparameters	Grid	Method
Number of trees	{100, 200}	RF and GBR
Maximum depth	{3,5,7,9,11}	RF and GBR
Minimum impurity decrease	{0.2, 0.5}	RF
Learning rate	{0.01, 0.1}	GBR

The optimal choice of parameters is carried out with a 3-fold Time-Series Nested Cross-Validation on the training set. Whitin each fold of the outer loop, an inner loop with three folds is used to run the grid search and tune the hyperparameters. Each inner fold provides a Mean Absolute Error (MAE) score, which is used so that the set of selected hyperparameters is the one that grants the inner fold with the lowest MAE. This set is then passed to the outer fold of the final cross-validation. The process is repeated three times, one for each outer loop which contributes to mitigate the risk of overfitting.

4.5. Training and Test Splits

For the regression approach the training period spans for 46 weeks from the one starting on the 15th of November of 2021 until the one starting on the 3rd of October of 2022, and the testing period is the following 3 weeks of October 2022. For the ML methods the training period is further split into three TS-N-CV which are then each split into three other inner TS-CV for hyperparameter tuning. Ultimately the time-series cross-validation methodology preserves the structure of the data allowing for better comparison of results of the different methodologies adopted.

4.6. Evaluation Metrics

A model's accuracy is key, and it can be measured by comparing the predicted results with the actual sales.

Some of the most commonly used metrics, and the ones used in the present dissertation, will be described next. In the following section n represents the total number of cases, $PRED_i$ and ACT_i represent the predicted and actual values, respectively, for case i.

The Mean Absolute Error (MAE) is the average absolute forecast error. This is computed by dividing the sum of the individual forecast errors, in absolute value, by the number of observations. More specifically:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|PRED_i - ACT_i|$$

Mean Squared Error (MSE) which is computed by averaging the squared differences between predicted sales and actual sales.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(PRED_i - ACT_i)^2$$

This metric is usually more adequately applied to situations where small frequent errors are preferred as opposed to larger sporadic ones.

Regarding the Mean Average Percentage Error (MAPE) it represents the average absolute error expressed as a percentage of the actual sales.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|PRED_i - ACT_i|}{ACT_i} \times 100$$

However, this metric has some limitations, one of them is that it penalizes more errors in lower actual values. This happened because a relatively smaller absolute error can represent a high

percentage of a low sales value. Additionally, if the actual value is very close to zero the MAPE value will be very high, even if the error is small.

The Weighted Mean Absolute Percentage Error (WMAPE) is a preferred measure related to the above mentioned, as it considers the sum of the absolute errors and the sum of the actual sales values. Therefore, it eliminated the concern of having zero or close to zero actual values.

$$WMAPE = \frac{\sum_{i=1}^{n} |PRED_i - ACT_i|}{\sum_{i=1}^{n} ACT_i} \times 100$$

Finally, the Mean Percentage Error (MPE), is a similar metric to the MAPE, but considers the actual error instead of the absolute error as a percentage of the actual sales.

$$MPE = \frac{1}{n} \sum_{i=1}^{n} \frac{(PRED_i - ACT_i)}{ACT_i} \times 100$$

Similarly, there is the Weighted Mean Percentage Error (WMPE) that mitigates the limitation of having close to zero sales.

$$WMPE = \frac{\sum_{i=1}^{n} (PRED_i - ACT_i)}{\sum_{i=1}^{n} ACT_i} \times 100$$

4.7. Used Packages

For the development of the methodology SQL was used to perform some of the preprocessing steps, as aggregating the baseline, cleaning outliers and computing the product attributes. For the remaining steps was used Python, specifically the statsmodels package for the regression approach. Regarding the machine learning portion, was mainly used an internally developed package that ultimately leverages the scikit-learn machine learning Python package.

5. Results

This chapter describes the results obtained for the different approaches that were outlined in the methodology section. First, each approach is analyzed individually, starting with regression and followed by machine learning. Then, both techniques are compared with each other and with the benchmark, which is the retailer's currently employed forecasting system.

Moreover, the results are evaluated following multiple metrics so that their possible limitations are mitigated. From the available evaluation metrics, the WMAPE considers the total absolute prediction error in relation to the total promotional sales. Conversely, the WMPE takes into account the total errors instead of the absolute ones, which can mislead interpretations as negative forecasting errors can compensate for positive ones and vice-versa. Due to this fact, through the present section, the WMAPE is used as the main score for establishing comparisons.

Lastly, it is crucial to reinforce that for this client's specific problem, the costs of inventory storage are lower than the costs of having stock ruptures. This translates into the fact that over-forecasting is preferred to under-forecasting. Over- or under-forecasting is illustrated by the scores of the WMPE, negative values correspond to the situations where the predicted values are more often below the actual sales and positive values correspond to the situation where the actual values are usually lower than the predicted ones. Thus, positive values for the WMPE score are preferred to negative ones, keeping a somewhat low absolute value in mind.

5.1. Regression Results

Regarding the regression approach, as explained in the previous chapter, for each one of the four categories a model was estimated for each class of products (A, B and C) in each store type. Additionally, in order to access the added value of the constructed promotional and interaction variables, three variations of the same model were estimated. The first variation includes only the provided promotional information, while the second encompasses all the available variables, including the constructed ones. Lastly, it was estimated a version containing the best selection of predictors according to the variance inflation factor analysis.

Given all the variations of models, three sets of results would be presented for each of the three store types, for each of the ABC class, for each of the four categories, making the analysis intricate and comparisons less clear. For ease of interpretation the results of each category for each variation of the regression model are aggregated at the store type level. This is done by calculating the weighted average of each class ABC, considering the proportion of promotional sales that each one contributes to the overall promotional sales. The results for each store type are multiplied by the corresponding class percentage of promotional sales in order to obtain the results presented below in Table 5.1.

Table 5.1 - Regression results for the category Beers and Ciders

Store type	Model	MPE	WMPE	MAPE	WMAPE
	Original	47.8%	-9.48%	77.4%	48.78%
L	All	30.4%	-8.43%	62.7%	45.63%
	Best selection	29.1%	-8.15%	62.0%	45.38%
	Original	35.4%	-9.52%	62.8%	45.13%
M	All	22.1%	-13.67%	53.4%	41.84%
	Best selection	19.9%	-13.49%	51.9%	41.61%
	Original	72.5%	16.06%	92.8%	61.83%
S	All	52.2%	7.78%	74.0%	52.78%
	Best selection	51.8%	6.71%	73.7%	52.22%

The results presented in the table above concern the Beers and Ciders category, in Appendix D are the equivalent results for the remaining categories. In the diagrams below are presented the results for the store type L, M and S in the categories Liquid Fats, Candy and Chocolate and Eggs respectively.

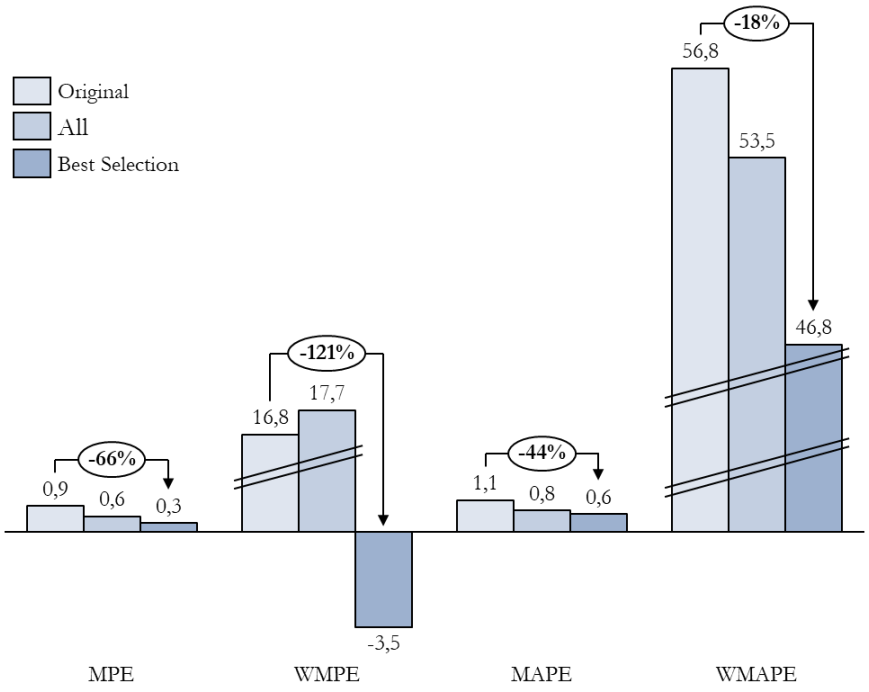

Figure 5.1 - Results for store type L in category Liquid Fats

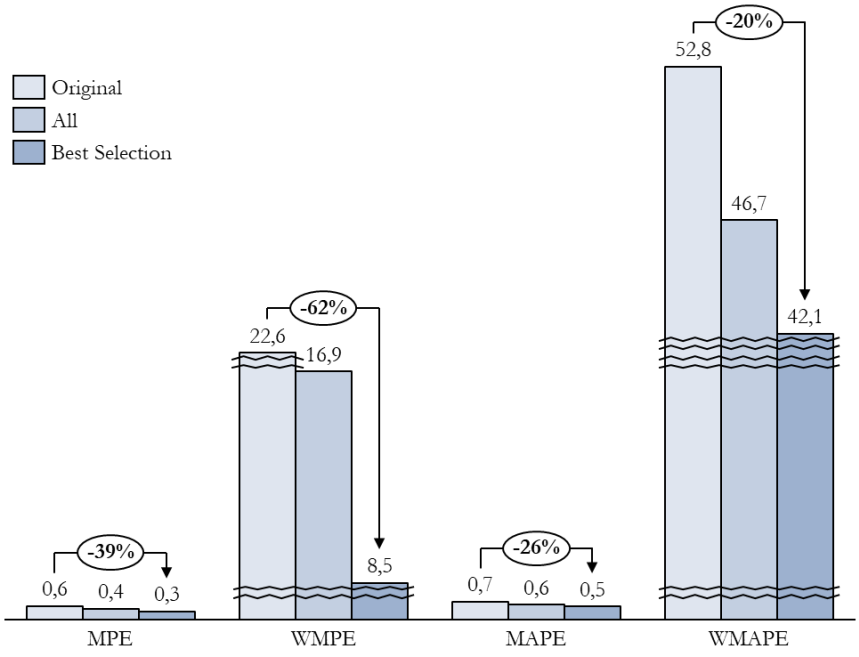

Figure 5.2 - Results for store type M in category Candy and Chocolate

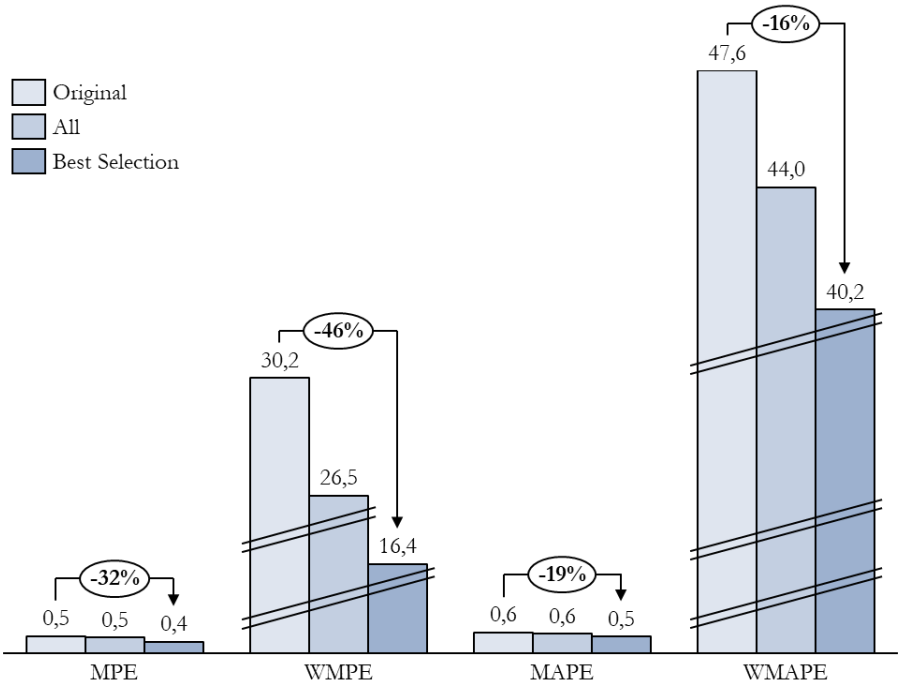

Figure 5.3 - Results for store type S in category Eggs

In what is presented it is seen that the results in general do improve with the inclusion of the computed promotional variables when compared to the original model that considers only the promotional tag information. The largest decrease in WMAPE, of 16% (13pp.) can be observed in the Liquid Fats category, class S. Moreover, in terms of WMPE, the biggest decrease is observed in the Candy and Chocolate class L, of 107%, from 7.5% to -0,5%.

From the use of the original predictors to the use of the best combination, the biggest improvements are in the Liquid Fats 29% (22.9pp) for the WMAPE. The decrease in the WMPE is highest in percentage in the Sweets and Chocolate category class L, of 139%. However, the improvements are best in the same class in the Eggs category. This is the case due to the final value in Eggs being positive (1,5%), as opposed to the negative value for the Candy and Chocolate (-3%).

Lastly, the results obtained also demonstrate that the use of the variance inflation factor, as a method to select the optimal combination of predictors, shows improvements in terms of forecasting accuracy.

5.2. Machine Learning Results

Regarding the machine learning approaches, the data used is the same for both models and it undergoes the same preprocessing steps, which ensures that both approaches are comparable. In Table 5.2 below are presented the test results. However, there is a major difference, the fact that in the machine learning approach only one model was estimated for each category. While in the regression method the store-type and ABC class are used to filter the contexts that are modeled at a time. In the ML case, both variables are taken as inputs in the single model per category that is estimated.

Table 5.2 - Machine Learning results per category

	Model	MPE	WMPE	MAPE	WMAPE
Beers and Ciders	GBR	39.2%	15.7%	51.1%	31.7%
	RF	44.6%	18.2%	55.7%	32.2%
Liquid Fats	GBR	165.6%	66.8%	174.2%	83.1%
	RF	196.3%	78.8%	202.3%	94.5%
Eggs	GBR	-0.4%	-2,7%	6.5%	5.2%
	RF	2%	-2.4%	4.4%	4.5%
Candy and Chocolate	GBR	82.3%	32.2%	103.4%	75.4%
	RF	66.1%	26.8%	82.3%	60.3%

The results show that the Random Forest has better results in general when compared to the Gradient Boosting Regressor for the less promoted categories. Yet for the Beers and Ciders and for the Liquid Fats categories the GBR has a better performance.

The category with best results over all is the Eggs. It is hypothesized that this is the case due to higher homogeneity of the category in terms of both sales values and products. Additionally, it can be seen by the overall positive values for the Weighted Mean Percentage Error that there is a general trend to over-forecast.

5.3. Overall Comparison

Throughout the dissertation has been mentioned the goal of improving forecasting accuracy. This goal was kept in mind during the development of the methodology. For this reason, the training and testing sets have been the same as well as the pre-processing steps. This ensures that the results from the different methodologies can be compared between each other and the results of the forecasting system applied by the retailer. Furthermore, the aggregation method of the results of the regression approach was done according to the proportion of promotional sales of each class and to the average of the three store types.

In the table below are presented the results comparing the regression approach that uses the best selection of predictors, the best machine learning algorithm and the RDF system.

Table 5.3 - Results for the regression, machine learning and RDF per category

	Approach	MPE	WMPE	MAPE	WMAPE
Beers and Ciders	Regression	34.2%	-4.71%	62.8%	46.49%

	ML	39.2%	15.7%	51.1%	31.7%
	RDF	14.6%	-0,4%	41.4%	32.84%
	Regression	65.9%	15.17%	87.6%	53.21%
Liquid Fats	ML	165.6%	66.8%	174.2%	83.1%
	RDF	16.2%	19.82%	50%	44.32%
	Regression	38.8%	13.70%	54.7%	40.45%
Eggs	ML	2%	-2.4%	4.4%	4.5%
	RDF	-3.6%	-12.47%	32.2%	29.57%
	Regression	34.6%	4.76%	54.9%	41.97%
Candy and Chocolate	ML	66.1%	26.8%	82.3%	60.3%
	RDF	-11.5%	-24.42%	45.7%	42.01%

The first conclusion is that the predictive performance of the forecasting system used by the retailer is not, generally, outperformed by the methods presented. There are, nonetheless, exceptions that are noteworthy. In the table above it can be seen that the machine learning approach has the best performance of the three in the Eggs category. Additionally, in the Beers and Ciders category the ML method is able to get better results than RDF in terms of WMAPE. Finally, in the Candy and Chocolate category the regression approach is very competitive with RDF, this can be further verified in Appendix D, where it can be seen that the results are better than the average benchmark for the S store format of this category.

Referring to the Weighted Mean Percentage Error scores, the proposed approaches present better prediction solutions than the forecasting system. In the Liquid Fats category, the WMPE of the regression approach is lower than that of RDF. For Eggs, the ML approach has a better overall WMPE score, nonetheless the positive value for the regression approach is preferred to the negative one of RDF, even though the aforementioned is slightly lower in absolute value. In the case of the Candy and Chocolate category the regression results are the best ones and, as the

remaining two are somewhat close in absolute values, the positive value for ML is preferred to the negative for RDF.

Additionally, the gap, in terms of performance scores, between the regression model and the retailer's benchmark tends to be shorter than that of the machine learning method. This difference is more prominent in the Liquid Fats category. The gap between the regression model and RDF is 8.9pp. and 4.7pp. and the gap between ML and RDF is 38.8pp. and 47pp., in terms of WMAPE and WMPE respectively.

A pattern across both methods is the poor performance in the Liquid Fats category. This fact is emphasized by the high WMPE observed for the machine learning method. Upon further examination of the characteristics of the specific category one of the possible explanations advanced is related to the available baseline. The Liquid Fats is, as the Beers and Ciders one, a heavily promoted category. However, in the Liquid Fats one the same SKUs are being promoted for many consecutive weeks. For the Beer's case, the consumer perception is that the products are in promotion for many consecutive weeks, however, in reality, the consecutive promotions happen in similar products but different SKUs, as the size or package can vary. This is relevant since having many consecutive promotions in the same SKU, implies that there are not many non-promotional weeks available from which data can be collected to accurately estimate baseline sales. Ultimately, the estimated baseline for the Liquid Fats category is less stable thus less reliable, yielding worst predictions.

Moreover, there are considerable differences, especially in terms of WMPE, between the performance of the regression and the ML approach. Overall, the prediction errors tend to be more centered for the regression approach while in for ML there is a higher propensity to over-forecast. A potential reason for this is the aggregation level at which the models are estimated. In both approaches the results are sales predictions for SKU-store-week combinations, however, in the regression approach, a total of nine models are estimated for each category. Conversely, only one machine learning model is estimated for each category. The regression models are estimating sales of products present in the same store type and that have a similar level of promotional uplift. The ML model must return a solution that can fit all situations, which makes it more complex to accurately encapsulate the discontinuities between store types and classes.

6. Conclusion

In an already uncertain context, that is grocery retail demand, introducing promotional dynamics heightens this uncertainty. Promotional actions do not impact the single promoted item, but rather a large set of other available products that may or may not be on promotion and of products that may or may not be similar to each other. Moreover, the impacts of a price reduction are not just felt during the targeted period. On the one hand some can keep influencing sales for a later period after the campaign. On the other hand, consumers can start anticipating these actions, therefore there can be repercussions also in the earlier periods. Having accurate demand predictions in any situation is crucial for inventory management and replenishment planning activities. It can contribute to the reduction of additional storage costs, to reducing food waste and it can also prevent stock rupture occurrences, which can have other consequences from a consumer satisfaction perspective.

In order to predict promotional sales two different methodologies were applied, a regression model and a machine learning approach. For both cases the first step was to derive product attributes from the available descriptions in order to characterize products and group them according to their similarities. A second step was to classify products according to their average promotional up-lift into three groups A, B and C. Then, a key step for the proposed methodology was to calculate different variables that are able to capture some of the relevant promotional effects, such as cannibalization and stockpiling.

Lastly, was the estimation of nine regression models for each one of the selected categories and of two ML algorithms, specifically Random Forest and Gradient Boosting Regressor. For the regression-based approach a model was fit for each store type and within each type a model was estimated for each ABC class, thus allowing to better capture intra-class and intra-type homogeneity and better distinguish inter-class and inter-type heterogeneity. This methodology first served as a proof of concept that the introduction of relevant promotional effects' predictors contributed to increase prediction accuracy. This was verified by fitting different versions of the regression equation, which demonstrated that the version including the optimal combination of the estimated predictors yielded better results. Concurrently, this approach performed better than what was first expected. The general results proved to be competitive

with the ones obtained by the benchmark method currently applied by the retailer, and even beating its performance in some specific situations.

As for the machine learning approach, only one model was estimated for each category and both the ABC class and store type used as predictors, for each of the two chosen algorithms. Overall, the RF performed better in the two more heavily promoted categories and the GBR outperformed it in the categories with less promotional activity. Contrary to what was expected from the literature review the machine learning methods do not always get better prediction results when compared to the regression or the benchmark methods. In two of the categories the ML does compete with the results obtained by RDF, however, compared to the regression results, the performance of ML is more unstable and less reliable.

Ultimately, the methodology developed was relevant to identify different data patterns in the presence of promotional effects as well as understand how to integrate those effects in the estimation process. Furthermore, it provided a deeper insight into factors that can negatively influence prediction accuracy and on possible solutions to mitigate noise introduction in the forecasting process.

6.1. Limitations

The first limitation faced during the dissertation is the data quality and availability. On the one hand, only fifty weeks of data were available which limits seasonality analysis, as there is not enough information available to translate full seasonal patterns. On the other hand, some crucial indicators for sales and promotional analysis were not provided, particularly related to price or price index and to store assortment. Regarding quality were also noted a few problems with the given variables, namely, there were price cuts registered to be above 100%, and problems with promotional and regular sales correctly being assigned as such.

Secondly, a major concern with the provided data is the estimated baselines sales. The error of this estimated baseline impacts the prediction error of the promotional sales as this is a critical

predictor in both approaches. This issue is more prevalent in the heavily promoted categories where there are less fully non-promotional weeks to accurately compute a baseline.

The last limitation is related to the established comparison between the proposed approaches and the in-use RDF forecasting system. To attest the performance of the developed methodology more benchmark models such as simple exponential smoothing as done in literature could have been estimated. The comparison with the RDF forecasting system can be unfair at times due to the fact that the retailer uses those predictions for their replenishment plan. Therefore, there is no guarantee that the observed sales are a true representation of demand, or if it was capped due to the predictions made by RDF.

6.2. Future Work

Addressing some of the limitations mentioned above there could be more effort put into leveraging data from different sources. Furthermore, it is important to give greater emphasis to the data cleaning and outlier removing process. Also related to the quality of the input data it is important to have accurate and reliable baseline sales, which would contribute to more consistent promotional forecasting results.

Related to the regression technique the approach to obtain the optimal selection of predictors can be further explored. Instead of relying solely on the variance inflation factor regression approaches such as LASSO or Elastic Net, regularizations can be applied. In regard to the machine learning approaches more values of the hyperparameters can be tested. Moreover, the ML model could be further enhanced with a similar approach to the one used for the regression model, in which multiple models were trained for different subsets of store types and classes.

Finally, an important future step is to apply the developed methodology to more categories and expand the application to the entire retailer universe.

References

Abolghasemi, M., Beh, E., Tarr, G., & Gerlach, R. (2020). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Computers and Industrial Engineering*, *142*. https://doi.org/10.1016/j.cie.2020.106380

Ailawadi, K. L., Harlam, B. A., & César, J. (2006). Promotion Profitability for a Retailer: The Role of Promotion Brand Category and Store Characteristics. *Journal of Marketing Research*, *43*, 518–535.

Ali, Ö. G., Sayin, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications*, *36*(10), 12340–12348. https://doi.org/10.1016/j.eswa.2009.04.052

Andrews, R. L., Currim, I. S., Leeflang, P., & Lim, J. (2008). Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? *International Journal of Research in Marketing*, *25*(1), 22–33. https://doi.org/10.1016/j.ijresmar.2007.10.001

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321–335. https://doi.org/10.1016/j.ijpe.2015.09.039

Auer, J., & Papies, D. (2020). Cross-price elasticities and their determinants: a meta-analysis and new empirical generalizations. *Journal of the Academy of Marketing Science*, *48*(3), 584–605. https://doi.org/10.1007/s11747-019-00642-0

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481–485. https://doi.org/10.1257/aer.p20151021

Batista, B. (2016). *Análise Quantitativa de Encomendas e Efeitos Promocionais em Produtos Perecíveis*.

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

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). Time Series Analysis: Forecasting and Control. *Journal of Time Series Analysis*, *37*(5), 709–711. https://doi.org/10.1111/jtsa.12194

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees* (1st Edition). Routledge. https://doi.org/10.1201/9781315139470

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, *10*(8), 1283. https://doi.org/10.3390/math10081283

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). PromoCast™: A New Forecasting Method for Promotion Planning. *Marketing Science*, *18*(3), 301–316. https://doi.org/10.1287/mksc.18.3.301

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

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

Hewage, H. C., & Perera, H. N. (2022). *Retail Sales Forecasting in the Presence of Promotional Periods* (pp. 101–110). https://doi.org/10.1007/978-3-030-92604-5_10

Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research*, *237*(2), 738–748. https://doi.org/10.1016/j.ejor.2014.02.022

INE. (2020). *Estatísticas do Comércio - 2020*. www.ine.pt

Kaneko, Y., & Yada, K. (2016). *A Deep Learning Approach for the Prediction of Retail Store Sales; A Deep Learning Approach for the Prediction of Retail Store Sales*. https://doi.org/10.1109/ICDMW.2016.154

Kong, E. L. (2015). *Cannibalization Effects of Products in Zara's Stores and Demand Forecasting Signature redacted Signature of Author Signature redacted*. MIT.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, *23*(3), 377–390. https://doi.org/10.1016/j.ijforecast.2007.02.006

Leeflang, P. S. H., van Heerde, H. J., & Wittink, D. R. (2005). How Promotions Work: SCAN*PRO-Based Evolutionary Model Building. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.321003

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, *249*(1), 245–257. https://doi.org/10.1016/j.ejor.2015.08.029

Martins, E., Verardi Galegale, N., & Paula Souza -CPS, C. (2022). *Retail Sales Forecasting Information Systems: Comparison Between Traditional Methods and Machine Learning Algorithms*.

Mendes, G., & Pereira, S. (2021). Promoções, saldos e black friday: como nasce um desconto. *POD Pensar - Ideias Para Consumir*.

Oracle. (2010). *Oracle Retail Demand Forecasting - user guide*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ragharan Srinivasan, S., Ramakrishnan, S., & Grasman, S. E. (2005). Incorporating cannibalization models into demand forecasting. *Marketing Intelligence & Planning*, *23*(5), 470–485. https://doi.org/10.1108/02634500510612645

Shah, J., & Avittathur, B. (2007). The retailer multi-item inventory problem with demand cannibalization and substitution. *International Journal of Production Economics*, *106*(1), 104–114. https://doi.org/10.1016/j.ijpe.2006.04.004

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Trapero, J. R., Kourentzes, N., & Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*, *66*(2), 299–307. https://doi.org/10.1057/jors.2013.174

van Heerde, H., Leeflang, P. S. H., & Wittink, D. R. (2000). The estimation of pre-and postpromotion dips with store-level scanner data. *Journal of Marketing Research*, *37*(3), 383–395. https://www.researchgate.net/publication/280051284

Wittink, D. R., Addona, M. J., Hawkes, W. J., & Porter, J. C. (1998). *SCAN* PRO: The estimation, validation and use of promotional effects based on scanner data*. Cornell University.

Appendixes

Appendix A

In this appendix the chart presents in more detail the promotional information available, including the type, location, communication vehicle and discount information. Additionally, it contains the acronyms regarding the promotional information that are used throughout the dissertation.

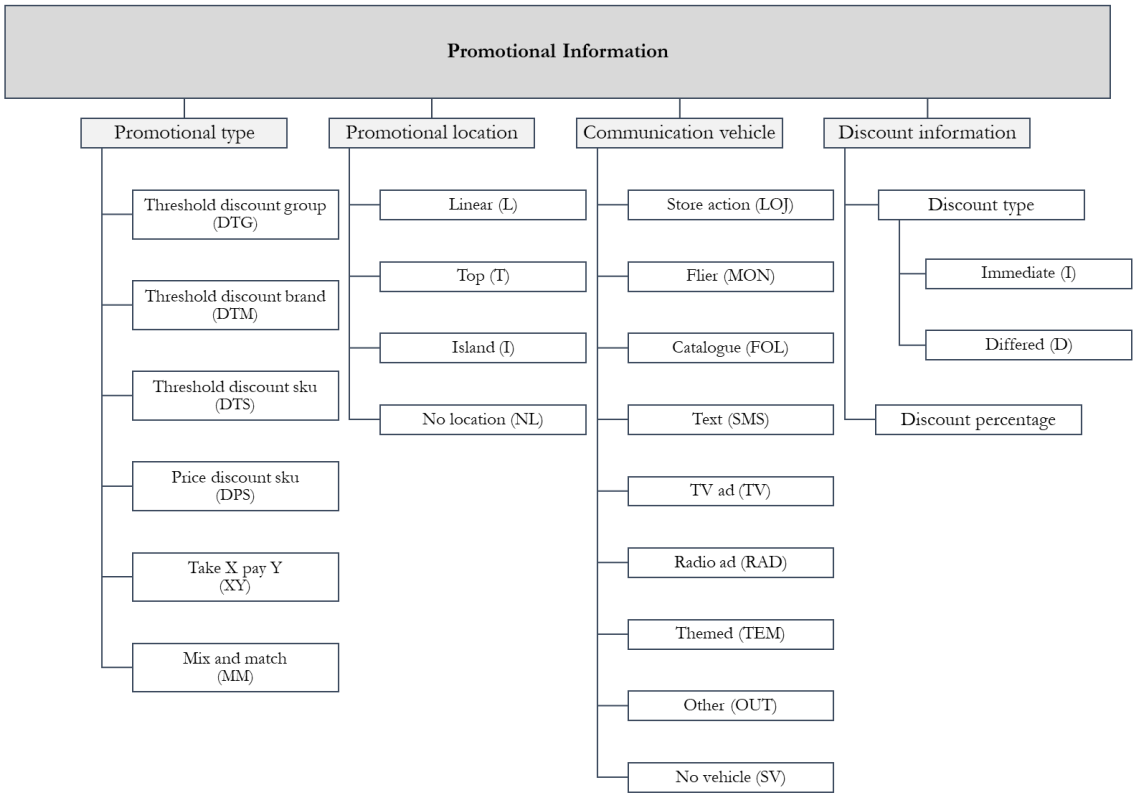

Figure A.1 - Detailed description of the promotional information

Appendix B

Appendix B includes the tables that describe what type of products encompass each cluster in each category and the respective number of products in each cluster.

Table B.1 - Clusters considering 3 attributes for Beers and Ciders

Package	Type	Size	Cluster Nº	Nº of Products
Bottle	Beer - Special and Stout	Big	1	3
		Mini	2	7
		Regular	3	84
		Other	4	7
	Beer - International	Big	5	13
		Mini	6	21
		Regular	7	43
		Other	8	3
	Beer - Non-Alcoholic	Big	9	1
		Mini	10	5
		Regular	11	9
	Other Beverages	Mini	12	3
		Regular	13	3
	Cider	Mini	14	8
		Regular	15	9
	Beer - National	Other	16	31
		Big	17	17
		Regular	18	7
Can	Beer - Special and Stout	Big	19	1
		Regular	20	3
	Beer - International	Big	21	1
		Regular	22	8
		Other	23	1
	Beer - Non-Alcoholic	Regular	24	5
	Other Beverages	Mini	25	6
		Regular	26	7
	Cider	Big	27	6
	Beer - National	Big	28	3
		Regular	29	6

Table B.2 - Clusters considering 2 attributes for Beers and Ciders

Package	Type	Cluster Nº	Nº of Products
Bottle	Beer - Special and Stout	1	101
	Beer - International	2	80
	Beer - Non-Alcoholic	3	15
	Other Beverages	4	6
	Cider	5	17
	Beer - National	6	55
Can	Beer - Special and Stout	7	4
	Beer - International	8	10
	Beer - Non-Alcoholic	9	5
	Other Beverages	10	13
	Cider	11	6
	Beer - National	12	9

Table B.3 - Clusters considering 3 attributes for Liquid Fats

Type	Subtype	Size	Cluster Nº	Nº of products
Olive Oil	Regular	1	1	2
		2	2	9
		3	3	6
	D.O.P	1	4	8
		2	5	8
		3	6	4
	Virgin	2	7	8
		3	8	2
	Extra Virgin	1	9	19
		2	10	24
		3	11	8
	Bio	1	12	4
		2	13	2
		3	14	1
	Premium	1	15	15
		2	16	6
		3	17	6
Vegetable Oil	Vegetable Oil	1	18	2
		2	19	13
		3	20	8
	Peanut	2	21	2
	Sunflower	2	22	4
		3	23	2
	Bio	1	24	11
		2	25	2
	Other	1	26	8
		2	27	1

Table B.4 - Clusters considering 2 attributes for Liquid Fats

Type	Size	Cluster Nº	Nº of products
Olive Oil	1	1	48
	2	2	57
	3	3	27
Vegetable Oil	1	4	21
	2	5	22
	3	6	10

Table B.5 - Clusters considering 3 attributes for Eggs

Type	Quantity	Size	Cluster Nº	Nº of products
Bio	1	Class M/L	1	2
		Other	2	3
Other	1	Class M/L	3	17
		Class S	4	4
		Class XL	5	3
		Other	6	1
	2	Class M/L	7	1
	3	Class M/L	8	1
		Other	9	7
Free Range	1	Class M/L	10	5
		Class XL	11	1
	3	Class M/L	12	1
Cage	1	Class M/L	13	6
		Class XL	14	3
	2	Class M/L	15	6
		Class S	16	1
	3	Class M/L	17	1

Table B.6 - Clusters considering 2 attributes for Eggs

Type	Quantity	Cluster Nº	Nº of products
Bio	1	1	5
Other	1	2	25
	2	3	1
	3	4	8
Free Range	1	5	6
	3	6	1
Cage	1	7	9
	2	8	7
	3	9	1

Table B.7 - Clusters considering 3 attributes for Sweets and Chocolate

Family	Type	Format	Cluster Nº	Nº of products
Sugar Free / Vegan	Chocolate	Bonbon	1	1
		Snack	2	20
		Tablet	3	20
	Gummy	Fruit	4	14
		Mint	5	5
	Chewing Gum	Fruit	6	1
		Mint	7	2
Impulse	Chocolate	Bonbon	8	5
		Snack	9	55
		Tablet	10	10
	Gummy	Fruit	11	50
		Mint	12	12
	Other	Snack	13	11
	Chewing Gum	Fruit	14	43
		Mint	15	35
Children	Chocolates	Bonbon	16	4
		Snack	17	40
	Chewing Gum	Fruit	18	7
Regular	Chocolates	Bonbon	19	316
		Snack	20	113
		Tablet	21	282
	Gummy	Fruit	22	261
		Mint	23	24
	Other	Other	24	3
	Chewing Gum	Fruit	25	26
		Mint	26	19
Seasonal	Almond	Sugar	27	45
		Chocolate	28	202
	Chocolate	Bonbon	29	22
		Snack	30	272
	Gummy	Fruit	31	9
	Other	Other	32	13

Table B.8 - Clusters considering 2 attributes for Sweets and Chocolate

Type	Format	Cluster Nº	Nº of products
Almond	Sugar	1	45
	Chocolate	2	202
Chocolates	Bonbon	3	348
	Snack	4	500
	Tablet	5	312
Gummy	Fruit	6	334
	Mint	7	41
Other	Other	8	16
	Snack	9	11
Chewing Gum	Fruit	10	77
	Mint	11	56

Appendix C

In this appendix are included the tables that, per category, include the Variance Inflation Factor values for each class ABC and for each store type. In each table are highlighted the cases where the VIF value is above the threshold of 5. This indicates that that predictor variable is excluded from the optimal feature selection equation of the regression model.

Table C.1 - Variance Inflation Factor for category Beers and Ciders store type L

Variables	Class A	Class B	Class C
Intercept	**611.056**	**1414.823**	**4891.769**
Promotional quota of other SKUs with the same cluster	**6.027**	**5.981**	**7.351**
Promotional quota of other SKUs with the same attribute 3	3.608	2.034	1.935
Promotional quota of other SKUs with the same attribute 1	3.756	2.152	3.028
Promotional quota of other SKUs with the same attribute 2	2.626	1.912	1.937
Promotional quota of other SKUs with the same brand	2.228	3.289	**6.099**
Discount percentage	2.271	2.208	1.500
Average discount of other SKUs with the same cluster	**5.739**	**5.663**	**6.461**
Average discount of other SKUs with the same attribute 1	4.627	4.464	4.638
Average discount of other SKUs with the same attribute 2	1.977	2.049	1.887
Average discount of other SKUs with the same attribute 3	1.856	2.073	1.885
Average discount of other SKUs with the same brand	2.538	3.813	6.641
Percentage of products in promotion within attribute 1	1.504	1.799	1.780
Percentage of products in promotion within attribute 2	1.732	1.657	1.719
Percentage of products in promotion within attribute 3	2.461	2.338	2.257
Percentage of products in promotion within brand	1.806	1.675	1.603
Weeks since last promotional action	1.113	1.123	1.108
Percentage of average sales of the last promotional action	1.132	1.069	1.027
Trimester 2	1.916	1.951	2.516
Trimester 3	2.585	2.536	2.727
Trimester 4	1.632	1.599	1.822
Promotion type DTM	1.959	1.592	1.997
Promotion type DTS	3.143	2.844	3.053
Communication vehicle LOJ	1.071	1.084	1.152
Communication vehicle MON	1.019	1.038	1.043
Communication vehicle OUT	1.472	1.435	1.567
Communication vehicle SV	1.008	1.015	1.196
Communication vehicle TEM	1.411	1.536	1.762

Table C.2 - Variance Inflation Factor for category Beers and Ciders store type M

Variables	Class A	Class B	Class C
Intercept	**7123.938**	**4452.061**	**2665.069**
Promotional quota of other SKUs with the same cluster	**6.125**	**6.920**	**8.590**
Promotional quota of other SKUs with the same attribute 3	3.546	2.541	2.374
Promotional quota of other SKUs with the same attribute 1	4.845	3.827	**5.205**
Promotional quota of other SKUs with the same attribute 2	3.395	2.693	2.713
Promotional quota of other SKUs with the same brand	2.652	3.974	**6.655**
Discount percentage	2.170	2.314	1.430
Average discount of other SKUs with the same cluster	**5.342**	**6.595**	**7.131**
Average discount of other SKUs with the same attribute 1	**4.960**	**5.854**	**5.242**
Average discount of other SKUs with the same attribute 2	2.578	2.581	2.595
Average discount of other SKUs with the same attribute 3	1.969	2.132	2.205
Average discount of other SKUs with the same brand	2.637	4.716	**7.227**
Percentage of products in promotion within attribute 1	2.103	2.471	1.969
Percentage of products in promotion within attribute 2	1.626	1.621	1.549
Percentage of products in promotion within attribute 3	2.097	1.972	1.871
Percentage of products in promotion within brand	2.505	1.979	1.851
Weeks since last promotional action	1.103	1.107	1.080
Percentage of average sales of the last promotional action	1.129	1.102	1.020
Trimester 2	2.042	2.112	2.811
Trimester 3	2.812	2.803	3.193
Trimester 4	1.640	1.597	2.038
Promotion type DTM	2.020	1.773	2.006
Promotion type DTS	3.307	3.125	2.970
Communication vehicle LOJ	1.189	1.146	1.211
Communication vehicle MON	1.020	1.040	1.044
Communication vehicle OUT	1.383	1.362	1.559
Communication vehicle SV	1.253	1.225	1.292
Communication vehicle TEM	1.400	1.417	1.630
Promotion Location L	1.199	1.692	1.508
Promotion Location T	1.168	1.588	1.382

Table C.3 - Variance Inflation Factor for category Beers and Ciders store type S

Variables	Class A	Class B	Class C
Intercept	**238.694**	**344.947**	**475.103**
Promotional quota of other SKUs with the same cluster	**5.733**	**7.137**	**7.277**
Promotional quota of other SKUs with the same attribute 3	3.175	2.260	2.337
Promotional quota of other SKUs with the same attribute 1	**5.036**	**4.945**	**5.333**
Promotional quota of other SKUs with the same attribute 2	3.165	2.496	2.499
Promotional quota of other SKUs with the same brand	2.890	3.524	**8.951**
Discount percentage	2.207	2.219	1.651
Average discount of other SKUs with the same cluster	**5.756**	**6.190**	**6.256**
Average discount of other SKUs with the same attribute 1	**4.786**	**5.926**	**5.263**
Average discount of other SKUs with the same attribute 2	2.226	2.070	1.908
Average discount of other SKUs with the same attribute 3	1.584	2.004	1.780
Average discount of other SKUs with the same brand	2.216	3.724	**9.266**
Percentage of products in promotion within attribute 1	1.918	2.185	2.089
Percentage of products in promotion within attribute 2	1.607	1.967	1.624
Percentage of products in promotion within attribute 3	1.922	2.300	1.988
Percentage of products in promotion within brand	2.226	1.870	1.815
Weeks since last promotional action	1.106	1.102	1.083
Percentage of average sales of the last promotional action	1.085	1.085	1.029
Trimester 2	1.709	1.931	2.347
Trimester 3	2.318	2.631	3.127
Trimester 4	1.513	1.465	1.887
Promotion type DTM	1.890	1.645	2.049
Promotion type DTS	2.963	2.799	3.017
Communication vehicle LOJ	1.337	1.416	1.586
Communication vehicle MON	1.041	1.106	1.070
Communication vehicle OUT	1.694	1.503	1.439
Communication vehicle SV	1.034	1.049	1.556
Communication vehicle TEM	1.209	1.262	1.139

Table C.4 - Variance Inflation Factor for category Liquid Fats store type L

Variables	Class A	Class B	Class C
Intercept	**198.336**	**535.545**	**552.108**
Promotional quota of other SKUs with the same cluster	**6.749**	3.465	4.019
Promotional quota of other SKUs with the same attribute 3	**5.178**	1.523	1.368
Promotional quota of other SKUs with the same attribute 1	4.199	3.792	4.880
Promotional quota of other SKUs with the same attribute 2	4.438	2.738	2.717
Promotional quota of other SKUs with the same brand	2.923	2.510	**5.709**
Discount percentage	2.135	1.911	2.256
Average discount of other SKUs with the same cluster	**5.609**	**5.505**	**4.740**
Average discount of other SKUs with the same attribute 1	4.050	4.531	**5.405**
Average discount of other SKUs with the same attribute 2	2.446	3.353	2.585
Average discount of other SKUs with the same attribute 3	**5.684**	3.750	3.204
Average discount of other SKUs with the same brand	3.037	3.041	**6.862**
Percentage of products in promotion within attribute 1	2.312	2.204	2.225
Percentage of products in promotion within attribute 2	3.223	2.912	2.406
Percentage of products in promotion within attribute 3	4.194	4.324	3.287
Percentage of products in promotion within brand	1.564	1.484	1.406
Weeks since last promotional action	1.248	1.259	1.193
Percentage of average sales of the last promotional action	1.313	1.244	1.050
Trimester 2	1.737	1.781	1.977
Trimester 3	1.735	1.683	1.880
Trimester 4	2.189	2.092	1.935
Promotion type DTM	2.803	2.956	8.123
Promotion type DTS	3.305	3.295	8.291
Communication vehicle LOJ	1.198	1.319	1.279
Communication vehicle MON	1.053	1.067	1.078
Communication vehicle OUT	1.297	1.253	1.261
Communication vehicle SV	1.339	1.408	2.371
Communication vehicle TEM	1.205	1.348	1.247
Promotion location T	1.114	1.177	2.014

Table C.5 - Variance Inflation Factor for category Liquid Fats store type M

Variables	Class A	Class B	Class C
Intercept	**8720.548**	**18189.23**	**4883.350**
Promotional quota of other SKUs with the same cluster	**9.946**	**6.943**	**7.484**
Promotional quota of other SKUs with the same attribute 3	**6.214**	2.676	2.002
Promotional quota of other SKUs with the same attribute 1	4.836	4.323	**5.706**
Promotional quota of other SKUs with the same attribute 2	**6.058**	3.498	**5.485**
Promotional quota of other SKUs with the same brand	2.833	2.439	4.760
Discount percentage	2.192	1.915	2.108
Average discount of other SKUs with the same cluster	**7.677**	**8.541**	**7.852**
Average discount of other SKUs with the same attribute 1	4.133	4.886	6.175
Average discount of other SKUs with the same attribute 2	2.435	4.705	5.945
Average discount of other SKUs with the same attribute 3	**7.306**	**5.390**	3.461
Average discount of other SKUs with the same brand	2.882	2.645	**5.185**
Percentage of products in promotion within attribute 1	1.973	1.747	1.427
Percentage of products in promotion within attribute 2	2.821	3.319	2.818
Percentage of products in promotion within attribute 3	3.286	3.765	2.553
Percentage of products in promotion within brand	1.604	1.395	1.493
Weeks since last promotional action	1.274	1.220	1.117
Percentage of average sales of the last promotional action	1.228	1.137	1.023
Trimester 2	1.772	1.700	1.812
Trimester 3	1.671	1.597	1.804
Trimester 4	1.911	1.756	1.717
Promotion type DTM	2.287	3.381	3.137
Promotion type DTS	2.611	3.778	3.559
Communication vehicle LOJ	1.177	1.220	1.302
Communication vehicle MON	1.058	1.055	1.108
Communication vehicle OUT	1.372	1.116	1.155
Communication vehicle SV	1.206	1.405	1.418
Communication vehicle TEM	1.152	1.397	1.225
Communication vehicle TV	1.009	1.009	1.022
Promotion location L	**14.368**	**62.298**	**21.459**
Promotion location T	**14.199**	**61.851**	**21.056**

Table C.6 - Variance Inflation Factor for category Liquid Fats store type S

Variables	Class A	Class B	Class C
Intercept	**90.580**	**429.074**	**635.003**
Promotional quota of other SKUs with the same cluster	**13.624**	**11.283**	**10.354**
Promotional quota of other SKUs with the same attribute 3	**7.974**	4.941	4.067
Promotional quota of other SKUs with the same attribute 1	**6.215**	**5.029**	**7.339**
Promotional quota of other SKUs with the same attribute 2	**7.898**	**4.960**	**6.670**
Promotional quota of other SKUs with the same brand	2.940	2.377	**5.161**
Discount percentage	1.795	1.992	3.382
Average discount of other SKUs with the same cluster	**10.072**	**12.953**	**12.291**
Average discount of other SKUs with the same attribute 1	**5.089**	**4.811**	**7.503**
Average discount of other SKUs with the same attribute 2	2.874	**5.049**	**8.571**
Average discount of other SKUs with the same attribute 3	**7.893**	**6.504**	4.311
Average discount of other SKUs with the same brand	2.668	2.523	4.642
Percentage of products in promotion within attribute 1	1.671	1.671	1.673
Percentage of products in promotion within attribute 2	2.727	3.637	2.468
Percentage of products in promotion within attribute 3	2.707	3.813	2.309
Percentage of products in promotion within brand	2.105	1.634	2.014
Weeks since last promotional action	1.314	1.280	1.117
Percentage of average sales of the last promotional action	1.229	1.087	1.024
Trimester 2	1.861	1.651	1.890
Trimester 3	1.812	1.698	2.119
Trimester 4	1.802	1.547	1.607
Promotion type DTM	4.028	**7.765**	**15.471**
Promotion type DTS	3.760	**9.301**	**17.426**
Communication vehicle LOJ	1.263	1.306	1.382
Communication vehicle MON	1.066	1.057	1.184
Communication vehicle OUT	1.276	1.126	1.319
Communication vehicle SV	1.406	2.919	3.804
Communication vehicle TEM	1.237	1.204	1.058
Promotion location T	1.114	1.129	1.216

Table C.7 - Variance Inflation Factor for category Eggs store type L

Variables	Class A	Class B	Class C
Intercept	**82.769**	**87.407**	**118.579**
Promotional quota of other SKUs with the same cluster	**45.579**	**28.538**	**25.222**
Promotional quota of other SKUs with the same attribute 3	**14.287**	**13.041**	**23.613**
Promotional quota of other SKUs with the same attribute 1	**38.294**	**18.428**	**20.311**
Promotional quota of other SKUs with the same attribute 2	**10.128**	**14.677**	**23.219**
Promotional quota of other SKUs with the same brand	**7.746**	**8.999**	**12.817**
Discount percentage	2.388	2.369	4.086
Average discount of other SKUs with the same cluster	**32.155**	**31.859**	**42.064**
Average discount of other SKUs with the same attribute 1	**26.054**	**22.400**	**29.366**
Average discount of other SKUs with the same attribute 2	**7.609**	**12.118**	**17.951**
Average discount of other SKUs with the same attribute 3	**9.848**	**11.176**	**24.103**
Average discount of other SKUs with the same brand	**6.124**	**8.650**	**13.708**
Percentage of products in promotion within attribute 1	1.578	2.442	3.582
Percentage of products in promotion within attribute 2	**6.265**	2.540	3.247
Percentage of products in promotion within attribute 3	4.110	2.286	2.204
Percentage of products in promotion within brand	3.244	2.790	4.635
Weeks since last promotional action	1.648	1.808	3.580
Percentage of average sales of the last promotional action	1.148	1.099	1.100
Trimester 2	1.953	1.883	2.947
Trimester 3	2.697	2.616	3.710
Trimester 4	3.851	4.039	**6.289**
Promotion type DTM	1.131	1.292	**5.019**
Promotion type DTS	1.596	2.080	**7.641**
Communication vehicle LOJ	1.603	2.078	1.807
Communication vehicle MON	1.275	1.109	1.059
Communication vehicle OUT	1.478	2.177	2.991
Communication vehicle SV	1.334	1.303	1.741
Communication vehicle TEM	1.175	1.404	2.126

Table C.8 - Variance Inflation Factor for category Eggs store type M

Variables	Class A	Class B	Class C
Intercept	**74.388**	**58.133**	**67.638**
Promotional quota of other SKUs with the same cluster	**52.757**	**21.653**	**23.666**
Promotional quota of other SKUs with the same attribute 3	**18.069**	**23.771**	**30.317**
Promotional quota of other SKUs with the same attribute 1	**48.422**	**20.291**	**26.793**
Promotional quota of other SKUs with the same attribute 2	**15.225**	**16.532**	**22.387**
Promotional quota of other SKUs with the same brand	**7.858**	**10.637**	**11.498**
Discount percentage	2.283	2.155	3.744
Average discount of other SKUs with the same cluster	**46.368**	**27.846**	**33.228**
Average discount of other SKUs with the same attribute 1	**41.186**	**25.093**	**33.371**
Average discount of other SKUs with the same attribute 2	**10.857**	**13.764**	**17.774**
Average discount of other SKUs with the same attribute 3	**12.171**	**19.754**	**25.979**
Average discount of other SKUs with the same brand	**7.873**	**11.743**	**14.656**
Percentage of products in promotion within attribute 1	1.542	2.347	2.784
Percentage of products in promotion within attribute 2	**6.436**	4.476	3.753
Percentage of products in promotion within attribute 3	3.993	1.900	2.032
Percentage of products in promotion within brand	4.284	3.175	3.586
Weeks since last promotional action	1.661	1.835	1.859
Percentage of average sales of the last promotional action	1.126	1.038	1.030
Trimester 2	1.822	1.820	1.884
Trimester 3	2.660	2.341	3.380
Trimester 4	3.859	4.088	**6.358**
Promotion type DTG	NA	1.082	1.107
Promotion type DTM	1.058	1.456	**5.251**
Promotion type DTS	1.522	3.162	**8.664**
Communication vehicle LOJ	1.546	1.773	1.614
Communication vehicle MON	1.190	1.225	1.081
Communication vehicle OUT	1.393	2.343	2.583
Communication vehicle SV	1.163	1.191	1.450
Communication vehicle TEM	1.144	1.536	2.201

Table C.9 - Variance Inflation Factor for category Eggs store type S

Variables	Class A	Class B	Class C
Intercept	**64.104**	**53.576**	**73.530**
Promotional quota of other SKUs with the same cluster	**51.682**	**23.399**	**18.101**
Promotional quota of other SKUs with the same attribute 3	**25.456**	**23.764**	**20.677**
Promotional quota of other SKUs with the same attribute 1	**48.186**	**23.095**	**20.574**
Promotional quota of other SKUs with the same attribute 2	**25.645**	**21.402**	**24.300**
Promotional quota of other SKUs with the same brand	**6.220**	**9.745**	**13.690**
Discount percentage	2.029	1.964	2.916
Average discount of other SKUs with the same cluster	**33.530**	**20.758**	**25.026**
Average discount of other SKUs with the same attribute 1	**30.743**	**18.879**	**23.869**
Average discount of other SKUs with the same attribute 2	**15.878**	**14.277**	**18.284**
Average discount of other SKUs with the same attribute 3	**18.550**	**16.919**	**17.901**
Average discount of other SKUs with the same brand	**6.997**	**10.232**	**14.259**
Percentage of products in promotion within attribute 1	1.446	1.453	1.688
Percentage of products in promotion within attribute 2	**6.802**	**5.223**	4.261
Percentage of products in promotion within attribute 3	3.618	2.460	2.278
Percentage of products in promotion within brand	3.067	2.371	3.593
Weeks since last promotional action	1.727	1.637	1.538
Percentage of average sales of the last promotional action	1.134	1.044	1.042
Trimester 2	1.711	1.967	1.826
Trimester 3	2.181	2.445	3.002
Trimester 4	3.968	4.121	**5.201**
Promotion type DTG	1.156	1.601	4.086
Promotion type DTM	**5.138**	1.308	NA
Promotion type DTS	4.671	2.589	**6.964**
Communication vehicle LOJ	1.428	1.861	1.916
Communication vehicle MON	1.213	1.173	1.041
Communication vehicle OUT	1.600	2.110	1.967
Communication vehicle SV	1.189	1.223	1.148
Communication vehicle TEM	1.099	NA	NA

Table C.10 - Variance Inflation Factor for category Candy and Chocolate store type L

Variables	Class A	Class B	Class C
Intercept	**187.725**	**348.635**	**306.085**
Promotional quota of other SKUs with the same cluster	**14.705**	**11.530**	**9.334**
Promotional quota of other SKUs with the same attribute 3	**7.118**	4.003	3.624
Promotional quota of other SKUs with the same attribute 1	**7.020**	4.760	3.752
Promotional quota of other SKUs with the same attribute 2	2.411	1.722	1.436
Promotional quota of other SKUs with the same brand	3.871	2.254	2.474
Discount percentage	2.206	2.465	2.821
Average discount of other SKUs with the same cluster	**7.735**	**6.705**	**7.564**
Average discount of other SKUs with the same attribute 1	3.926	3.114	2.998
Average discount of other SKUs with the same attribute 2	2.471	1.881	1.418
Average discount of other SKUs with the same attribute 3	4.836	3.825	4.242
Average discount of other SKUs with the same brand	4.164	3.328	3.823
Percentage of products in promotion within attribute 1	2.550	3.728	2.679
Percentage of products in promotion within attribute 2	1.466	2.264	1.802
Percentage of products in promotion within attribute 3	2.134	3.651	2.179
Percentage of products in promotion within brand	1.421	1.424	1.356
Weeks since last promotional action	1.127	1.189	1.138
Percentage of average sales of the last promotional action	1.010	1.005	1.005
Trimester 2	2.034	1.832	2.048
Trimester 3	1.919	1.907	2.122
Trimester 4	1.421	1.431	1.436
Promotion type DTG	1.019	NA	1.020
Promotion type DTM	**15.850**	**7.426**	**11.041**
Promotion type DTS	**15.927**	**7.527**	**11.221**
Communication vehicle LOJ	1.238	1.246	1.270
Communication vehicle MON	1.103	1.120	1.175
Communication vehicle OUT	1.123	1.068	1.103
Communication vehicle SV	1.311	1.199	1.295
Communication vehicle TEM	1.020	1.130	1.150
Promotion location T	1.552	1.026	1.018

Table C.11 - Variance Inflation Factor for category Candy and Chocolate store type M

Variables	Class A	Class B	Class C
Intercept	**223.915**	**2346.644**	**925.807**
Promotional quota of other SKUs with the same cluster	**17.525**	**12.780**	**12.984**
Promotional quota of other SKUs with the same attribute 3	4.000	**4.854**	**4.884**
Promotional quota of other SKUs with the same attribute 1	**10.846**	**6.006**	**6.392**
Promotional quota of other SKUs with the same attribute 2	2.074	2.091	1.940
Promotional quota of other SKUs with the same brand	2.811	2.811	2.966
Discount percentage	1.748	2.298	2.409
Average discount of other SKUs with the same cluster	**7.468**	**6.344**	**7.483**
Average discount of other SKUs with the same attribute 1	**5.304**	3.559	3.755
Average discount of other SKUs with the same attribute 2	2.286	2.237	1.948
Average discount of other SKUs with the same attribute 3	2.725	3.322	3.948
Average discount of other SKUs with the same brand	3.097	3.461	3.767
Percentage of products in promotion within attribute 1	2.101	2.466	2.141
Percentage of products in promotion within attribute 2	1.442	1.527	1.440
Percentage of products in promotion within attribute 3	2.588	2.094	1.680
Percentage of products in promotion within brand	1.372	1.358	1.283
Weeks since last promotional action	1.394	1.180	1.118
Percentage of average sales of the last promotional action	1.016	1.005	1.005
Trimester 2	1.738	1.938	1.947
Trimester 3	1.757	1.864	1.914
Trimester 4	1.698	1.342	1.411
Promotion type DTM	**7.322**	**10.111**	**11.820**
Promotion type DTS	**7.361**	**10.120**	**11.968**
Communication vehicle LOJ	1.120	1.228	1.209
Communication vehicle MON	1.104	1.120	1.101
Communication vehicle OUT	1.052	1.062	1.083
Communication vehicle SV	1.207	1.213	1.177
Communication vehicle TEM	1.163	1.252	1.140
Promotion location L	NA	1.433	1.376
Promotion location T	1.065	1.371	1.215

Table C.12 - Variance Inflation Factor for category Candy and Chocolate store type S

Variables	Class A	Class B	Class C
Intercept	**179.940**	**211.640**	**187.725**
Promotional quota of other SKUs with the same cluster	**25.914**	**19.425**	**14.705**
Promotional quota of other SKUs with the same attribute 3	4.792	**7.329**	**7.118**
Promotional quota of other SKUs with the same attribute 1	**18.205**	**9.775**	**7.020**
Promotional quota of other SKUs with the same attribute 2	2.338	2.499	2.411
Promotional quota of other SKUs with the same brand	3.437	3.884	3.871
Discount percentage	1.560	1.891	2.206
Average discount of other SKUs with the same cluster	**9.833**	**8.561**	**7.735**
Average discount of other SKUs with the same attribute 1	**7.427**	4.607	3.926
Average discount of other SKUs with the same attribute 2	2.233	2.465	2.471
Average discount of other SKUs with the same attribute 3	2.893	4.517	4.836
Average discount of other SKUs with the same brand	3.407	3.762	4.164
Percentage of products in promotion within attribute 1	2.543	2.928	2.550
Percentage of products in promotion within attribute 2	1.639	1.703	1.466
Percentage of products in promotion within attribute 3	3.186	2.639	2.134
Percentage of products in promotion within brand	1.476	1.496	1.421
Weeks since last promotional action	1.337	1.176	1.127
Percentage of average sales of the last promotional action	1.009	1.005	1.010
Trimester 2	1.884	2.058	2.034
Trimester 3	1.936	2.028	1.919
Trimester 4	1.535	1.349	1.421
Promotion type DTM	**10.262**	**21.616**	**1.019**
Promotion type DTS	**10.266**	**21.319**	**15.850**
Communication vehicle LOJ	1.175	1.225	**15.927**
Communication vehicle MON	1.150	1.099	1.238
Communication vehicle OUT	1.059	1.075	1.103
Communication vehicle SV	1.561	1.465	1.123
Communication vehicle TEM	1.026	1.039	1.311
Promotion location T	1.024	NA	1.020

Appendix D

This appendix includes the prediction results from the regression approach for all the estimated models for each store type.

Table D.1 - Regression results for the category Liquid Fats

Store type	Model	MPE	WMPE	MAPE	WMAPE
	Original	92.4%	16.78%	113.6%	56.77%
L	All	55.2%	17.67%	77.9%	53.52%
	Best selection	31.3%	-3.48%	63.8%	46.76%
	Original	143.4%	38.95%	159.4%	72.37%
M	All	102.4%	36.81%	118.9%	68.44%
	Best selection	85.1%	22.42%	102.8%	56.77%
	Original	139.9%	52.90%	1.52.3%	79.02%
S	All	91.9%	40.37%	1.05.5%	65.99%
	Best selection	81.4%	26.57%	0.96.1%	56.09%

Table D.2 - Regression results for the category Candy and Chocolate

Store type	Model	MPE	WMPE	MAPE	WMAPE
	Original	58.8%	7.50%	80.5%	51.69%
L	All	42.8%	-0.49%	66.5%	47.17%
	Best selection	37.7%	-2.95%	61.4%	44.53%
	Original	55.5%	22.62%	73%0	52.78%
M	All	44.9%	16.94%	62.1%	46.65%
	Best selection	34%0	8.51%	53.7%	42.12%
	Original	50.2%	21.72%	65.4%	48.85%
S	All	41.4%	16.02%	56.4%	42.81%
	Best selection	32.1%	8.71%	49.7%	39.27%

Table D.3 - Regression results for the category Eggs

Store type	Model	MPE	WMPE	MAPE	WMAPE
	Original	47.5%	13.44%	61.6%	42.02%
L	All	40.6%	8.20%	56%	39.18%
	Best selection	31.3%	1.53%	49.6%	37.23%
	Original	48.5%	22.55%	64.7%	46.03%
M	All	58.3%	31.07%	67.1%	46.45%
	Best selection	049%	23.23%	62.4%	43.94%
	Original	53.3%	30.15%	64.2%	47.58%
S	All	48.3%	26.51%	60.3%	43.97%
	Best selection	36%	16.35%	52%	40.19%