Modeling customer churn: Differentiated models per customer segment

Ana Catarina Costa Carvalho

Master's Dissertation

Supervisor: Prof. Vera Miguéis

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Integrated Master in Industrial Engineering and Management

June 2020

Abstract

Retaining customers has been considered one of the most critical challenges among those included in Customer Relationship Management and it is particularly relevant in saturated markets, such as the grocery retail sector.

In this regard, this dissertation aims to leverage transactional data collected through a retailer's loyalty card to create churn prediction models focused not only in detecting the customers more likely to churn but also in identifying the factors that best explain it. The goal in doing so is to collect relevant business insights to develop effective and proactive customer retention strategies.

One-to-one marketing strategies for customer retention, although effective, are difficult to implement from an operational point of view. To ensure that campaigns are customized and still operable a two-stage approach that comprises both supervised and unsupervised data mining techniques was developed. Firstly, customers were divided into logical groups through hierarchical clustering, according to their past behavior at the retailer and for each of them differentiated churn models were developed resorting to Regularized logistic regressions and Decision trees. It was concluded that there are no major differences in terms of predictive performance among the alternative classification techniques, however when using Regularized logistic regression more actionable and relevant insights to increase customer retention were identified.

Current state of the art classification algorithms are not properly aligned with commercial goals, in the sense that, models assume that misclassification errors carry the same cost, which is not suitable for churn modeling since failing to identify a profitable customer implies a higher cost than for an unprofitable customer. Motivated by the importance of align classification models with commercial goals, in a later phase of this study we incorporated financial costs throughout Bayes minimum risk, which classifies customers after the training phase of the algorithm considering different costs per customer.

The results highlighted the importance of using real financial costs, as the application of the Bayes minimum risk model led to cost saving compared with the use of a non cost-sensitive method, such as Regularized logistic regression.

Resumo

A retenção de clientes é considerado um dos desafios mais críticos na gestão do relacionamento com os consumidores, sendo particularmente relevante em mercados saturados, como é o caso do mercado retalhista alimentar.

Neste sentido esta dissertação pretende, tendo por base os dados transacionais do cartão de fidelização de um retalhista, desenvolver modelos de previsão focados na identificação de clientes com maior probabilidade de abandono, assim como na identificação dos fatores que melhor explicam o mesmo. A informação recolhida permitará ao retalhista em questão desenvolver estratégias de retenção de clientes mais eficazes e proativas.

Estratégias de marketing customizadas para cada um dos clientes apesar de eficazes são díficeis de implementar do ponto de vista operacional. Para garantir campanhas customizadas mantendo a sua operacionabilidade foi adotada uma abordagem repartida em duas fases que compreende técnicas de aprendizagem supervisionada e não supervisionada.

Numa fase inicial os clientes foram agrupados, em função do seu comportamento no retalhista, e posteriormente foram desenvolvidos modelos de previsão de abandono diferenciados para cada um dos clusters, recorrendo a modelos de Regressão logística regularizada e a Árvores de decisão. Conclui-se que não existem diferenças consideráveis no desempenho preditivo entre ambos, contudo as Regressões logísticas regularizadas neste contexto forneceram informações mais acionáveis e, por esse motivo, mais relevantes para aumentar a retenção dos clientes.

Os algoritmos de classificação mais utilizados pela comunidade académica e científica não estão devidamente alinhados com os objetivos comerciais, na medida em que, os modelos assumem que todos os erros de classificação acarretam o mesmo custo para a empresa, o que não se enquadra com a realidade, falhar na classificação de um cliente rentável tem um impacto diferente do que num cliente menos rentável. Motivados pela importância de alinhar os modelos com os objetivos comerciais, numa fase posterior do estudo procurou-se incorporar custos financeiros das campanhas de retenção através do uso de algoritmos de classificação que incorporam custos diferentes para cada cliente, nomeadamente o *Bayes minimum risk*.

Os resultados evidenciam a importânica da incorporação de custos reais na classificação dos clientes. A aplicação do algoritmo *Bayes minimum risk* conduziu a uma poupança nos custos de campanhas de retenção quando comparado com um modelo independente dos custos financeiros, neste estudo a Regressão logística regularizada.

Acknowledgements

My first thanks goes to Prof. Vera Miguéis for her guidance, which was critical for the completion of this dissertation. I am grateful for her total availability, support, insights and critical look during the development of this dissertation. Thank you for showing me that in everything we do we must apply a dose of passion, but also a lot effort and work.

I would like to thank Sonae MC for giving me the opportunity to develop this project and to be part of such a professional and dedicated team.

Thank you, Patrícia Castro, Filipe Miranda, Margarida Costa and Francisca Paupério for all the support provided. In particular, I owe my gratitude to Liliana Bernardino, Ana Freitas and Francisco Barbosa for giving me the opportunity to work alongside them and for always being open to help me. Their support and advice was invaluable.

Finally, yet important I would like to acknowledge the strong support of my family and friends. To Rosa and João, my parents, for showing me that dedication and humility are the perfect combination to achieve our goals. To Joana and Diogo for the support throughout the whole path, for always encouraging me to overcome the challenges and to be happy. I also thank Tiago for his support, help, comprehension and love. I am also grateful to my dearest friends for all the encouragement received.

“When we confront facts and fears, we achieve real power and unleash our capacity for change.”

Margaret Hefferman

Contents

Acronyms and Symbols

ALBA	Active Learning Based Approach
AUC	Area Under the Curve
B	Behavioral Information
BC	Behavioral Change Detection Variables
BMR	Bayes minimum risk
CLV	Customer Lifetime Value
CRISP-DM	Cross Industry Standard Process for Data Mining
CRM	Customer Relationship Management
CSDT	Cost Sensitive Decision Trees
CSLR	Cost Sensitive Logistic Regression
DT	Decision Tree
ELP	Every Day Low Price
FN	False Negative
FP	False Positive
KDD	Knowledge Discovery in Databases
KITH	Kids in the House
KNN	K-Nearest Neighbor
LR	Logistic Regression
MLE	Maximum Likelihood Estimate
NN	Neural Network
NV	No Value
PB	Promo Busters
PS	Price Sensitivity
RFM	Recency, Frequency and Monetary
RLR	Regularized Logistic Regression
ROC	Receiver Operating Characteristic
RR	Regularized Regression
SD	Socio-Demographic Information
SOW	Share of Wallet
SVM	Support Vector Machine
T	Transactional
TN	True Negative
TP	True Positive

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents and contextualizes the project in question, as well as, the company and some of the challenges that retailers currently face.

1.1 Project Motivation and Objectives

Companies nowadays devote more attention and effort to managing churn as part of their customer relationship management program. They are aware of the importance of maintaining long-lasting relationships with customers, as a way to ensure their long term sustainability (Rygielski et al., 2002).

In saturated markets, as the food retail, it becomes even more crucial to invest in customer retention strategies and for Sonae MC, especially in the scope of its loyalty program, managing customer relationship is a central business issue.

In this sense, the present project arises with the objective of developing differentiated churn models for Sonae MC per customer segments. This study consists in grouping customers according to their behavior at Sonae MC and developing for each group different churn prediction models that, besides predicting customers most likely to leave the Portuguese retailer, should also provide insights about the customer churn path. The identification of customers and their churn behavior will allow more effective and targeted retention actions, instead of resorting only to undifferentiated discounts for all customers predicted as churners. This project supports Sonae MC long term strategy, not only by increasing customer retention rates through a customer-oriented strategy but also endeavoring, at a later stage, to assess whether the inclusion of real costs in the classification of customers as churners or not impact the financial results of retention campaigns.

1.2 Big data Analytics in Retail

Advances in information technology have enabled the collection, storage and analysis of huge amounts of data. The volume, variety and velocity of this data are increasing exponentially, and as a result, big data has become a buzzword. It is estimated that the world's data volume doubles every 18 months (Gaurav Pant, 2014).

Analytics has been used in a variety of business domains and the major goal in doing so is to transform data into knowledge and extract business meaningful insights to support decisions that most impact companies.

In order to keep competitive, in a data-driven marketplace, companies experience a huge pressure to use data to make effective business decisions (Zdravevski et al., 2020). Zdravevski et al. (2020) points out that "the success of companies hugely depends on how well they can analyze the available data and extract meaningful knowledge".

Cebr (2012) highlighted that the industries that can most benefit from big data analytics are manufacturing, central government, healthcare, telecom, banking and retail industries. Leaders in retail industry are more aware of the importance of analytics in business and a study from Manyika et al. (2011) concluded that a retailer embracing data to its full potential can improve its operating margin by more than 60%.

Nowadays, retailers are more focused than ever on making customer-centric decisions and to do so, they resort to customer data that can be extracted from multiple sources. Data used by retailers can be obtained through points of sale, online transactions platforms, mobile apps, social media, enterprise systems, public sources (e.g. Census) and from syndicated sources (Gaurav Pant, 2014). This data has been applied to different functional areas, such as marketing, operations and customer intelligence (Intel, 2016). Some of the most common examples of applications of analytics in retail are price optimization, demand forecasting, customer segmentation and campaign optimization.

1.3 Characterization of the food retail sector in Portugal

> "73% of Portugal retail customers buys in more than one supermarket banner per week, with 21% of the inquiries saying that they shop in three or four". (The Consumer Goods Forum, 2018)

The food retail market in Portugal is a highly competitive and promotional market, where current players compete fiercely for customer acquisition and loyalty. In addition, consumers in the retail sector are now more demanding. Customers are highly informed about the different products and services available, expect more customization of the service and have access to increasingly attractive promotional campaigns.

The two main players in the food retail market in Portugal are Sonae MC and Pingo Doce, whose market shares[1] are 21.9% and 20.8% respectively. These two clearly stand out from the other food retailers, with the third, Auchan having 9.5% of share. The recent entry, in 2019, of a new player, the Spanish retail market leader called Mercadona, which is considered to have the

[1] Sonae MC data referes to June 2018, Pingo Doce and Auchan data referes to December 2017. Source: Sonae Mc Presentation

potential to shake up the current reality of the retail sector in Portugal, makes even more crucial for already established retailers to find new ways of differentiating and innovating (The Consumer Goods Forum, 2018).

A study carried out by Accenture (2015) reported that the retail sector has demonstrated an ability to reinvent itself over the years, namely by creating differentiating experiences for consumers. In order to improve consumer experience, retailers have invested in an omni-channel reality, by adapting sales channels and points of contact and have focused on creating more integrated and efficient operations. Retailers also have demonstrated an effort to leverage the power of information, that enables them to support their decisions and respond proactively to consumer behavior.

1.4 Sonae MC and its loyalty program

Sonae MC has a chain of food-based stores, i.e. hypermarkets, large supermarkets and small supermarkets, which differ between them in terms of sales area, products offered and target customers. Additionally, Sonae MC sells online at Continente Online.

Cartão Continente was launched on January 23, 2007 with the purpose of increasing customer loyalty to the brand and enabling the company to study the behavior of its costumers on an individual basis. It currently has around 4 million active loyalty accounts and approximately 85% of the total transactions are done by customers using loyalty cards. Cartão Continente also includes 19 permanent partners[2] and some occasional partners, which intends to cover the customer's spending ecosystem, such as food, transportation, health and fashion.

The loyalty card is the main source of knowledge about consumer behavior and has allowed the company to improve the management of the range of products offered, support extensions and adaptations of stores, guide product innovation and guide customer strategy and loyalty programs, in which churn management is included.

1.5 Dissertation Structure

The present dissertation is divided into six chapters. The first chapter presents the motivation and contextualizes the project and some of the major topics to understand the reality of retail sector in Portugal and more specifically the Sonae MC and its loyalty program. The Second chapter comprises a literature review on the main topics including customer retention, churn management and churn modeling using non cost-sensitive and cost-sensitive algorithms. Chapter three describes how the company currently handles churn, details the project objectives and the strategy defined to achieve them. Chapter four covers in detail the methodology adopted in all phases of

[2] Internal Partners: Continente Stores, Meu Super, Note!, Mo, Wells's, Zu, Bagga and Zippy; External Partners: Galp, Ibersol group (includes brand as: Burger King, SOL, KFC, Pizza Hut, Ò Kilo, Roulotte, Pans & Company, Miit and Pasta Caffé)

the project. On Chapter five the results obtained resorting to data from Cartão Continente are provided and some suggestions about how the results could entail the customer retention program are provided. Chapter six presents conclusions of the work done and suggests future work that could be done in these topics.

Chapter 2

Literature Review

2.1 Customer Retention

Companies are facing increasingly competitive pressure which has prompted them to implement new strategies, particularly in the way they establish the relationship with customers. The concepts of mass production and product-orientation that emerged with the Industrial Revolution have been supplanted by a customer-oriented strategy, where companies seek to know their customers better to meet their needs and expectations (Rygielski et al., 2002).

With the increased focus on the customer there are business approaches that take on greater importance such as Customer Relationship Management (CRM). It comprehends a series of processes and systems that endeavor to create a deeper understanding of a customer and influence customer behavior, through meaningful communications, to build long term and profitable relationships with specific customers (Ling and Yen, 2001).

CRM has developed some analytical tools and tasks to extract more knowledge from the typically huge set of information regarding each customer that companies have, some given by the customer during registration and other gained from customers activity at the company, to be more efficient in the different dimensions where it performs (Lazarov and Capota, 2007).

The four dimensions of CRM proposed by Kracklauer et al. (2004) are:

1. Customer Identification
2. Customer Attraction
3. Customer Retention
4. Customer Development

In food retail sector, Customer Retention has been considered the most challenging and crucial of all dimensions (Miguéis et al., 2012). Besides being a good indicator of customer satisfaction, customer retention has become an important managerial issue. It is particularly relevant in saturated markets, as retail market, once the growth of the number of new customers is decreasing and companies strive to keep existing customers (Shaon and Rahman, 2015).

This dimension can have a huge impact on companies, namely because of the economic value of customer retention, summarized by Van den Poel and Lariviere (2004) as follows:

- Acquiring a new customer is five to six times more expensive than retaining an existing one;
- Customer retention reduces the need to pursue new customers who may be more higher risk, allowing greater focus to fulfil the demands of the existing ones;
- The positive word of mouth from customers is a good way to get new ones;
- People tend to share negative experiences more easily than positive service experiences, which can have a major impact on the image of companies to other consumers;
- Long-term customers are less sensitive to competitors' marketing activities;
- When there is a long-term relationship, there is a database related to customer that allows to meet, in a less costly manner, their demands;
- Long-term customers tend to buy more.

Among the four dimensions, customer retention is the most common one for which data mining is used to support decision making. Most of the research in this area is related to loyalty programs and one-to-one marketing (Ngai et al., 2009).

One-to-one marketing differs from traditional marketing as it is customer orientated, focuses on understanding and addressing the different needs of individual customers, to increase customer satisfaction and consequently customer retention (Amorim, 2013).

To increase customer retention, companies can also resort to churn management. The realization of its importance has led companies to invest in understanding more about churn, what drives it and how to manage it.

2.2 Churn Management

Glady et al. (2009) defines churn as a "marketing-related term characterizing a consumer who is going from one company to another".

According to Lazarov and Capota (2007) there are three types of churn: deliberate, incidental and non-voluntary. The first one relates to customers that decide to quit the contract and to switch to another provider. Some of the reasons that could be pointed out for this are unsatisfaction, no rewards for customer loyalty and not having competitive price plans. Incidental churn means that the customer leaves the company without the aim of switching to another competitor, and it may be motivated by changes in the circumstances that prevent the customer from further requiring the service, such as change in financial circumstances or change of the geographical location of the customer where the company is not present in. Finally, the non-voluntary churn is due to the decision of the company to discontinue the contract with the customer and it could be due to abuse or non-payment of the service.

Most of the churn management solutions are focused in reducing deliberate churn. In spite of being the hardest to predict, it is also the most interesting for companies. There is no need to predict non-voluntary churn as the initiative to withdraw the relationship comes from the company and incidental churn only explains a small percentage of company's churn (Hadden et al., 2007).

According to Lejeune (2001), “churn management consists of developing techniques that enable firms to keep their profitable customers and it aims to increase customer loyalty”.

There are two basic approaches to manage churn, which are the untargeted and the targeted approach. The untargeted approach does not address specific customers and relies on superior products along with mass advertising to boost brand loyalty and ultimately customer retention. The targeted approach includes reactive and proactive churn management. Company actions that attempt to prevent churn from happening are the result of proactive churn management and involve identification of customers who are more likely to churn and the development of actions that should be taken to prevent it, like special programs or promotions. Reactive churn management means that the company takes actions that attempt to prevent churn only after the customer had contacted the company showing the desire to cancel their relationship. It includes incentives to stay, for example a rebate (Tamaddoni Jahromi, 2009).

The reactive churn management should not be preferred over proactive churn management, because companies should continuously search for potential sources of customer dissatisfaction in advance to build loyalty in their customers and the likelihood to prevent the customer from churning is higher when using proactive actions to prevent it (Asaari and Karia (2000);Valtola (2019)).

The prevention of churn takes on even greater emphasis in retail, as this is the most familiar sector with fluctuations in consumer behavior, as the customer makes the decision where to buy often, which is even more aggravated by the lack of need to communicate to retailer the intention to leave. Having that in mind, retailers must constantly reconvince customers that loyalty is a good investment (Mattison, 2006).

According to Mattison (2006) there are two crucial aspects that justify the effort required from retailers to retain customers:

1. Customers may, in many cases, have access to the same products and on similar conditions at different retailers, as the majority of retailers’ suppliers are the same;
2. Retailers are forced to compete for consumer attention on a daily basis, as customers can easily purchase products from other retailers.

Retailers and the scientific community are more aware of the economic value of customer retention (see Chapter 2.1) and the need to increase customers loyalty, which has led them to invest in the application of data mining techniques to find out what can be done to increase customer retention.

2.3 Churn Prediction

Advances in technology have revolutionized customer churn management. Nowadays the use of data mining models for churn prediction is very common and widely proliferated, not only across the academic community, but also across the professional one.

2.3.1 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) refers to the whole process of discovering useful knowledge from data. Companies from different areas such as marketing, finance (especially investment), fraud detection, manufacturing and telecommunications resort to various KDD techniques to support their decisions (Fayyad et al., 1996).

Churn Management is one of the areas where KDD and data mining techniques can play a central role in alleviating churn issues (Lejeune, 2001).

KDD process comprises several steps, see Figure 2.1, which include data selection, preprocessing, transformation, data mining and interpretation/evaluation of the results obtained. It is worth to mention that Fayyad et al. (1996) distinguishes KDD from data mining. Data mining, according to the same author, is a step in the KDD process that includes data analysis and discovery algorithms to extract patterns from the data. The other steps in the KDD process are also essential to ensure the extraction of useful information from the data, once the blind application of data mining techniques can lead to the identification of meaningless or invalid patterns. However, in industry, media, and some research fields, the term data mining is often used to refer to the whole knowledge discovery process (Han et al., 2011a).

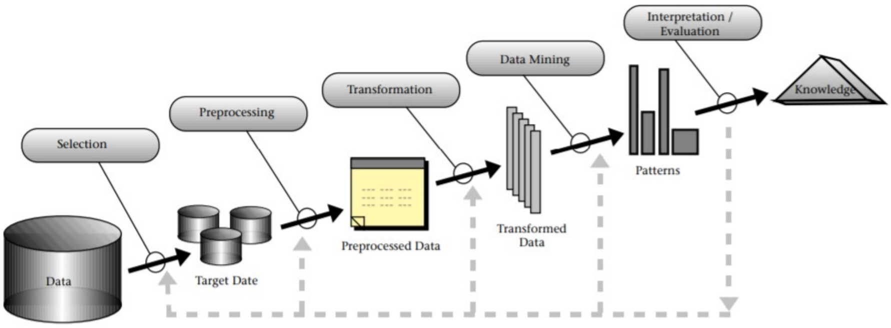


Figure 2.1: Steps of the KDD Process. Source: Fayyad et al. (1996)

Adapted for churn prediction purposes, data mining techniques for churn identification are methods that use historical data to find patterns which can point out possible churners (Lazarov and Capota, 2007).

2.3.2 Churn Concept Definition

One of the first steps when modeling churn is the definition of the periods under study and the definition of the attrition criterion that defines switchers.

The main periods recognized in churn prediction are the observation period, latency and the period for which the target variable is defined, known as target period, see Figure 2.2 (Veloso, 2013).

The observation period is the period in which historical customer behavior data is gathered to be used to train the prediction model. As the volume of data considered during this period

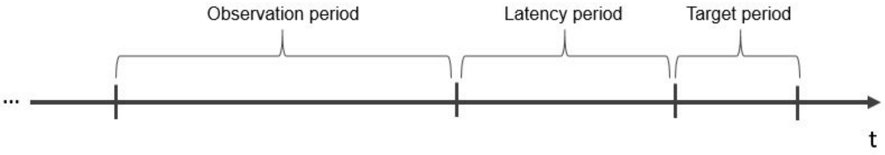


Figure 2.2: Periods considered for churn prediction

increases, the accuracy of the model is expected to increase as well, while the computational performance is expected to decrease.

The latency period is used mainly for logistics purposes. It is the period between the input variable window and the beginning of the target period, during which retention actions are planned to retain customers that were identified as potential churners. The duration of this period highly depends on the company's capabilities to allocate retention efforts.

Finally, the target period is the period for which customers are classified as being churners or not.

In addition to defining the periods most appropriate to the reality of the company in question, there is also the need to define the concept of churn. As in retail there is no need for a formal communication of abandonment of the organization, churn is said to occur if a customer's transactions do not happen during a certain time, defined by the company according to rules that can be determined resorting to business knowledge (Shetty et al., 2019).

The definition of the churn concept can have a major impact on the applicability and relevance of the churn model in a company. So, to define when a customer is going to be considered a churner, the modeler must be aware of the business rules and the application scenario of the model (Yan et al., 2004).

Some strategies to define churn can be divided into static, dynamic and partial definitions (Veloso, 2013). In the static definition, a fixed period is established, and a customer is considered a churner if he/she makes no purchase during a defined period. It is a widely used strategy in managing churn in retail. For example, Bernardino (2012) classifies loyal customers of a food-based retail company as churners if they do not make purchases for a month.

In the case of the dynamic definition, a different concept of churn is applied for each client. Two different approaches to this concept were applied by Chiang et al. (2003) to predict churn in a network banking service. In the first approach, a customer is considered a churner when the period between transactions is longer than the average interval between transactions in the past. In the second, churn occurs when the period between transactions is longer than the longest time interval without transactions.

There is also the concept of partial churn, which is said to occur when a client reduces the volume spent in one company to a competitor, not having completely ceased the relationship with it. A study from Miguéis et al. (2012) performed a dynamic evaluation based on the expenses of the customer whose goal was to determine if there was a three-month period in which the customer made less than 40% of the purchases of the previous quarter. The detection of partial churn can also contribute to churn management since partial churn can lead to the total abandonment of the relationship between the company and the client (Buckinx and Van den Poel, 2005).

2.3.3 Data mining techniques and explanatory variables

Prediction is a supervised technique, in which historical data is used to predict values of certain attributes in unknown situations (Goldschmidt and Passos, 2005) .

Depending on the type of the target variable, prediction problems can be divided into classification or as regression problems. In classification problems, the goal is to determine a function that maps a set of records to predefined categorical data. Once the function has been defined, it can be applied to unknown data to determine its class. In the case of regression problems, the rational is the same but the target variable is numerical (Goldschmidt and Passos, 2005).

In churn analysis, as customers can be classified as being either churner or non-churners, the problem is generally conceptualized as a classification problem.

Ngai et al. (2009) states that the choices of data mining techniques should be based on the data characteristics and business requirements. When choosing between techniques to predict churn it is important to take into consideration three key aspects which are accuracy, comprehensibility, and justifiability (Verbeke et al., 2011).

Accuracy of a churn model is related to its predictive capabilities and is the percentage of instances correctly classified as churners or as non-churners. Comprehensibility and justifiability are topics related with understanding the causes/path of churning and what can be done to prevent it. According to Ahn et al. (2006), in order to be more effective in managing churn, companies should focus their efforts in knowing and understanding customer's behavioral churn path, the factors associated with churn and what can be done to prevent it.

Although comprehensibility and justifiability are recognized as crucial from a business point of view, they have been disregarded in many of the studies carried out in this area (Verbeke et al., 2011). There are several data mining techniques that can be applied to churn classification problem and there is no consensus as to which is more suitable for this purpose. To this end, in the past, preference was given to methods with higher predictive capabilities, such as Support Vector Machines (SVM) and Neural Networks (NN). In more recent studies at the top of the list of the techniques used are Decision trees (DT) and NN (KhakAbi et al., 2010).

A study from Khodabandehlou and Rahman (2017) used two-year data of one store from a grocery retailer and compared SVM, DT and Artificial Neural Network (ANN) only in terms of their predictive performance. The model with the highest accuracy in this case was the ANN. In terms of the variables included, the study also concluded that an extension of the traditional RFM (recency, frequency and monetary) may lead to better results than the use of RFM variable alone, as a better performance was accomplished when including the following variables: number of purchased items, number of returned items, the discount, the distribution time and prize added to RFM (prize is number of items given for free as a reward for buying a certain number of some of the items).

In more recent studies there has been an effort to increase their interpretability, which means that an effort is being made so that human beings can understand the causes that drive the decision and thus extract important insights for business (Molnar, 2019). Some of the studies focused on churn prediction in retail businesses that stand out in terms of interpretability are presented below.

The study carried out by Bernardino (2012) analyzed for 13 months a group of clients classified, by the grocery retailer, as being loyal and resorts to logistic regression (LR), DT and ANN to classify them as being churner or non-churners for the next month. To choose the best model, both model performance and interpretability were taken into consideration. The model chosen was DT with entropy as division criterion. The model uses around 100 variables to predict who is more likely to churn and some of the most discriminatory factors are response rate to promotional actions, variables related to the type of basket chosen, the amount spent, the number of visits in the current month and the number of visits in the last 12 months.

Also, in the study of Veloso (2013) DT led to the best results, when the quality of the model is based on its interpretability and performance. In this case DT outperformed LR, NN, Random Forest and SVM. The study used 6 months data from a grocery retailer and the factors highlighted as the most relevant for identifying churners were: purchase frequency, recency, number of visits, amount spent per visit, number of children in the household and gender. Additionally, in this study, some of the financial impacts that could be faced by the company when using the proposed model were analyzed. However, this analysis has some limitations, e.g. the cost of retention campaigns and their success rate was arbitrated and considered equal for all customers and these costs had no impact on the results of the model.

There are other articles, but in less number, such as Verbeke et al. (2011), that study how more novel data mining techniques, such as Active Learning Based Approach (ALBA), can be applied to predict churn with high accuracy without disregarding the importance of understanding the causes that underline the classification as being a churner or a non-churner. ALBA is a technique that combines SVM models, defined as black-box models, with extraction rule techniques.

2.3.4 Cost-Sensitive Classification

Most of the studies carried out on this topic use evaluation metrics that can be obtained using a confusion matrix (Table 2.1), such as accuracy, sensitivity, AUC (Area Under the Curve) and Lift for example. However, these metrics tacitly assume that misclassifications penalize the results the same way, and in real contexts misclassifying a churner as a non-churner has different implications than the other way round.

From a business point of view, and specifically in the case of churn modeling, it is more interesting to take into consideration real costs to maximize the results of retention campaigns.

According to Hadden et al. (2007), churn management efforts should not focus on the entire customer base, as customer retention costs money to organizations and not all customers are worth retaining. Despite this, most of the studies previously carried out, especially in food retail, do not take this into consideration.

Table 2.1: Confusion Matrix. Source: Bahnsen et al. (2015b)

	Actual Positive y=1	**Actual Negative** y=0
Predicted Positive c=1	True Positive (TP)	False Positive (FP)
Predicted Negative c=0	False Negative (FN)	True Negative (TN)

In cost-sensitive classification, the penalizations and the performance metrics take into consideration the financial costs that the company will face due to that classification. A study from Bahnsen et al. (2015b), analyzed the impact of including cost-sensitive analysis in classification of churn in an TV Cable provider, and concluded that using this approach can increase cost-saving of retention campaigns up to 26.4%. A saving of this magnitude can have a huge impact on the company's financial performance and should not be ignored.

Although standard cost-sensitive approaches assume constant costs for each type of class and the same costs among examples, this is not the best way to deal with churn problems since misclassifying a profitable customer as a non-churner has a greater impact that the same error with an unprofitable customer (Glady et al., 2009).

State of the art studies for cost-sensitive analysis resort to example-dependent-cost-sensitive method. Under which costs vary not only between classes but also between examples, see Table 2.2.

Table 2.2: Classification Cost Matrix. Source: Bahnsen et al. (2014)

	Actual Positive y_i=1	**Actual Negative** y_i=0
Predicted Positive c_i=1	C_{TPi}	C_{FPi}
Predicted Negative c_i=0	C_{FNi}	C_{TNi}

The total cost when using this methodology can be computed using the Equation 2.1.

$$Cost(f(S)) = \sum_{n=1}^{N} y_i(c_i \times C_{TPi} + (1-c_i) \times C_{FNi}) + (1-y_i)(c_i \times C_{FPi} + (1-c_i) \times C_{TNi}) \quad (2.1)$$

To successfully apply the aforementioned concepts in churn prediction, it is important to identify the possible implications that different combinations of the predicted class and the actual class may have for the company. These implications are schematized in Figure 2.3 and the costs that accompany such decisions are presented in the cost matrix, Table 2.3.

The relevant costs associated to retention campaigns are C_{oi}, that is the cost of the offer made to the client i to retain it, and C_a, that is the cost of contacting the customer and finally the customer lifetime value (CLV) for each client.

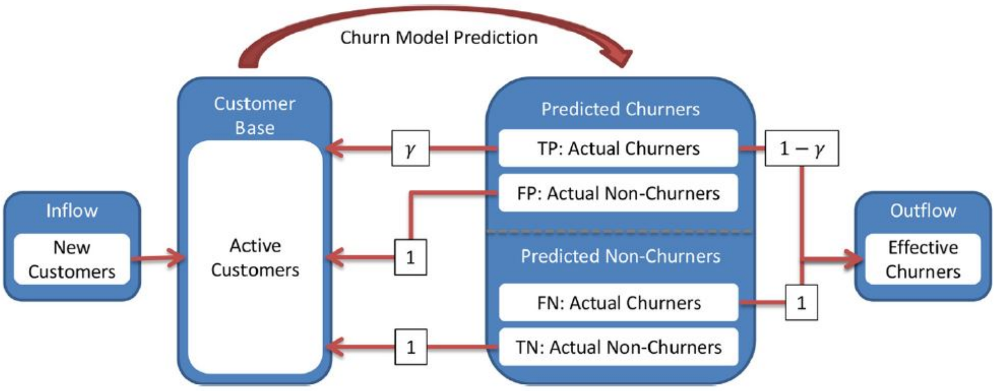


Figure 2.3: Churn campaign analysis. Source: Bahnsen et al. (2015b).

Table 2.3: Proposed churn modeling example-dependent cost matrix. Source: Bahnsen et al. (2015b)

	Actual Positive y_i=1	**Actual Negative** y_i=0
Predicted Positive c_i=1	$C_{TPi}= \gamma_i(C_{oi}+C_a)+(1-\gamma_i)(CLV_i+C_a)$	$C_{FPi}= C_{oi} + C_a$
Predicted Negative c_i=0	$C_{FNi} = CLV_i$	C_{TNi}=0

According to Pfeifer and Bang (2005), “CLV is the present value of all the future cash flows attributable to the customer relationship over the lifetime of that relationship”. CLV is important in churn prediction, especially when dealing with misclassification costs, because if the relationship with the customer is ceased, the company, in the long term, will lose the CLV of that customer.

When a customer is classified as a churner, retention campaigns are triggered and typically an offer is sent to prevent him/her from leaving the company. This offer is not necessarily accepted by all customers, who may or not accept the offer with a probability γ_i.

To compute C_{TPi} we need to consider two different costs (see Table 2.3), because the customers may or not accept the offer sent by the company. If the customer accepts and remains a customer, the cost is the cost of the offer (C_{oi}) added by the cost of contacting the customer (C_a). If the customer does not accept the offer and leaves the company the cost of sending the offer will still exist, as well as the cost of his/her CLV.

If the customer is identified as a churner and in fact was not (FP), then he/she will gladly accept the offer, and the cost will be C_{oi} plus C_a. In case a customer is classified as a non-churner but in fact is one (FN), no offer will be sent to him/her and the cost to the company will be his/her CLV. Finally, if the client is properly classified as a non-churner (TN), no offer will be sent to the client and the relationship still remains.

Returning to the Equation 2.1 and replacing the variables by the cost values (Table 2.3), the cost can be defined by the Equation 2.2 (Bahnsen et al., 2015b).

$$Cost(f(S)) = \sum_{n=1}^{N} y_i(c_i(\gamma_i(C_{oi}+C_a)+(1-\gamma_i)(CLV_i+C_a))+(1-c_i)(CLV_i))+(1-y_i)(c_i(C_{oi}+C_a)) \tag{2.2}$$

In example-dependent-cost-sensitive algorithms, the costs can be incorporated before, during or after training the classification model (Bahnsen et al., 2015b).

In the methods that incorporate costs before training, the data is transformed to take into account the costs before being given as input to the model. This transformation does not take into consideration the full cost matrix, it only considers the costs of misclassifying, which leads to less satisfactory results, which, for this reason, are not very detailed in this analysis.

The most referenced methods to incorporate costs during training are the cost-sensitive logistic regression (CSLR) (Bahnsen et al., 2014) and cost-sensitive decision trees (CSDT) (Bahnsen et al., 2015a). In CSLR the objective function is changed to take the costs presented in Table 2.3 into account. In CSDT the split is done by using a cost-based impurity measure as splitting criteria instead of using the traditional Gini index or Information gain.

A study of Bahnsen et al. (2015a) concluded that this methodology when compared to standard DT, leads to models with higher cost-saving, but generally there are no high improvements in accuracy, as the model is focused on maximizing the savings not the predictive performance. The same study also concluded that in the case of the CSDT algorithm, there was no significant improvement when using any pruning procedure, neither in savings nor in predictive performance.

Finally, the Bayes minimum risk (BMR) is a decision model that incorporates costs after the training of the model and it quantifies the trade-off of risks between the decision of classifying a customer as churner or non-churner using probabilities and the costs that accompany such decisions. In the end, each customer will be classified as belonging to the class that represents the lowest risk for the company (Bahnsen et al., 2014).

Bahnsen et al. (2014) compared how the timing of cost incorporation can impact savings. This study sought to predict whether or not a credit should be attributed to a client, and the main conclusions are that the inclusion of costs after training the model increases the savings of the company but the savings are even greater when the costs are taken into account during the training of the algorithm.

2.3.5 Final perspective on data mining techniques for churn prediction

Churn prediction is a common problem in many businesses and has received attention from academics and professionals over the years (Zdravevski et al., 2020). Considering the literature review carried out on this theme, it can be seen that in more recent studies there is a greater focus on interpretable models, rather than exclusive focus on measuring predictive capabilities, since these models can provide more interesting insights for the business.

Studies developed in food retail sector have followed this trend and, therefore, DT have become popular to predict churn. Decision trees have been highlighted for their ease of interpretation, low computational time and because they have a good performance when compared with more complex techniques (Hastie et al., 2009).

It is also noteworthy that, although there is an extensive literature on the application of data mining techniques in business contexts to predict churn, only limited research has been made concerning the financial impact of these techniques and their adaptation to maximize the value of customer retention campaigns.

Chapter 3

Sonae MC Case Study

3.1 Churn at Sonae MC

Although Sonae MC is the leader in the food retail market in Portugal, the sector in which it operates is a very competitive one, with increasingly demanding customers, which makes abandonment by customers a reality. Consequently, customer churn is a dimension of the churn relationship management that has been in the agenda of Sonae MC. The customer database created from the use of the Cartão Continente can be an important source of information to identify those customers most likely to churn and to understand what actions could be taken to increase customer retention.

3.2 Current practice of churn prediction at Sonae MC

Sonae MC already has a model that is used to manage churn. The major goal of the model is to predict which customers will do at least one visit during the period considered as well as the number of visits expected for each client in a given period. As input this model only takes into consideration three variables namely: total length of tenure, total number of transactions and the time since last transaction.

The actual model has an accuracy of about 95%. Since the model only considers three variables, it performs well in terms of computation time but does not provide information about the causes that lead to the conclusion that a specific customer has higher likelihood of churning. For Sonae MC computation time is not a critical issue, so it is more interesting to have a model that, besides performing satisfactorily in terms of computational time, also provides information that allows the company to take targeted actions for customers instead of generic actions that could be less effective.

Currently the company, as a retention strategy sends, by mail, an undifferentiated offer to customers identified as churners. This solution does not maximize the value for the company, as the offer may not be attractive to all customers and therefore the customer may leave the company anyway. Moreover, the fact that the strategy adopted is an attractive promotional offer may mean that customers who are constantly identified as churners have access to higher benefits than the others.

Another issue pointed out by the company to the current model is the fact that it constantly signals as churners those customers who are of lower value to the company and that occasionally visit it to redeem discounts from previous retention campaigns. This could be due to the fact that the current model only considers a single period to predict whether a customer is a churner or not and the period considered is the same for all customers. From a business point of view, this is not the ideal once customers have different behaviors in terms of the frequency with which they visit stores.

Another drawback of the current model is that it does not take into account the financial implications that the company will face with the retention campaigns, that are dependent on the output of the churn prediction model.

3.3 Proposed approach for churn prediction

3.3.1 Goals Definition

The objective of Sonae MC with this project is to develop a model, relying on customer behavior in the Cartão Continente ecosystem, that can predict with the desired antecedence which customers will leave the following insignias: Continente, Continente Modelo, Continente Bom Dia and Continente Online.

In addition to identifying the customers most likely to churn, the solution must conciliate data mining techniques with business knowledge, to understand the factors that best explain churn, depending on customer profile. These insights will help define which proactive retention campaigns can be most effective for customers. However, it is not feasible for a company like Sonae MC to have a true one-to-one relationship with all customers, thus, customers with similar profiles will be grouped, so that retention programs can be targeted and still operable.

Given that the model is expected to be applied in a real context, it should be aligned with business objectives, in the sense that, besides creating more value for customers, the model should also take into account real financial costs that arise from retention campaigns, in order to maximize their results.

Hence, the proposed approach seeks to reach a solution more aligned with company's financial goals, as well as enhance customer service and, consequently, customer retention rates.

3.3.2 Methodology

The methodology followed to tackle the purposed goals was CRISP-DM, see Figure 3.1.

In the beginning the objective was to gain business understanding about the current churn model, its applicability and expectations for the new model. The expectations were defined together with those who will be the end users of the model, i.e. the parties involved in the customer retention process.

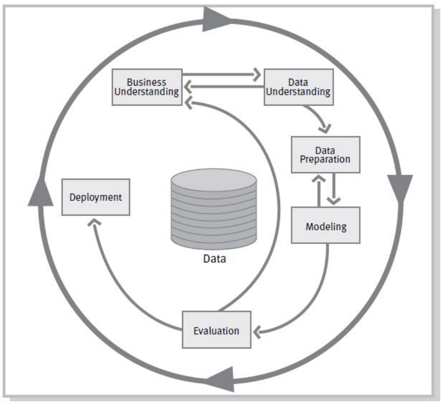


Figure 3.1: CRISP-DM reference model phases. Source: Chapman et al. (2000)

This was followed by an exploratory analysis of the customer's DNA data, transactional data and customer's data constructed by the company, to better understand what kind of information could be used in later phases.

The data preparation and modeling evolved in parallel. The modeling process can be divided into two major phases (see Figure 3.2), the creation of groups of customers with similar behaviors and also the creation of churn prediction models for each of the previously identified clusters. In the first phase, customers were grouped according to their behavior profile through hierarchical clustering techniques. The objective of this phase was to identify groups of customers with more similar behavior, so that at a later stage the creation of predictive models could be targeted. The second phase can be divided into two parts, the first one is independent of costs of retention campaigns, where explanatory models were developed for each segment using Regularized logistic regression and Decision tree. The goal in doing so, is to predict the likelihood of churning and identifying the factors that best explain churn. In the second part, as not all customers are worth retaining, a cost sensitive algorithm, i.e. Bayes minimum risk was applied. This algorithm is based on the results of the Regularized logistic regression and aims to assess the impact of using a cost-sensitive churn model in the results of retention campaigns.

Finally, the results were evaluated and suggestions were presented on how the project can entail the customer retention program by presenting proposals for actions that can be undertaken to increase customer retention in the different groups.

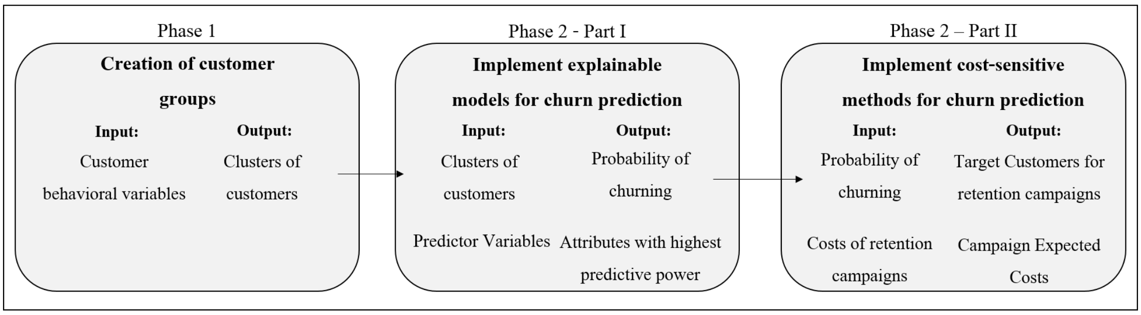

Phase 1
Creation of customer groups
Input:
Customer behavioral variables
Output:
Clusters of customers
Phase 2 - Part I
Implement explainable models for churn prediction
Input:
Clusters of customers
Predictor Variables
Output:
Probability of churning
Attributes with highest predictive power
Phase 2 – Part II
Implement cost-sensitive methods for churn prediction
Input:
Probability of churning
Costs of retention campaigns
Output:
Target Customers for retention campaigns
Campaign Expected Costs

Figure 3.2: Overall perspective of the methodology applied.

Chapter 4

Methodology

This chapter is divided into two main sections corresponding to the two main phases covered in this project. The first section describes the methodology adopted for the creation of clusters of customers and the second describes the development of the churn prediction models.

4.1 Behavioral Clusters

Tailor-made campaigns for each client would be the ideal to maximize customer retention, however, from an operational point of view, for an organization such as Sonae MC, this is not feasible.

To overcome this limitation in this study we resort to a strategy commonly used in customer relationship management, i.e. clustering. It allows to group a large number of customers that share similar characteristics. By segregating clients into logical groups it is possible to develop more effective proactive strategies for each of them (Mattison (2006), Han et al. (2011b)).

4.1.1 Data Collection

Sonae MC has a portfolio of customer DNA variables, in which each of the segmentations focuses on the analysis of the customers and their behavior at Sonae MC from different perspectives (see Appendix A). The combination of the different segmentations already created provides a holistic view of the customer and its behavior at Sonae MC. Under the assumption that customers with similar behavior at Sonae MC will also behave similarly in terms of churn, these were the variables used to create the clusters.

4.1.2 Clustering Techniques

Clustering analysis is described by Wei et al. (2003) as a “process whereby a set of objects is divided into several clusters in which each of the members is in some way similar and is different from the members of other clusters”. Clusters are better, more distinctive, the greater the similarity within a group and greater the difference between groups (Tan et al., 2016).

There are many clustering algorithms in the literature although the two most widely studied, according to partitioning criteria, are partitional and hierarchical clustering (Karypis et al., 2000).

In partitional clustering the clusters are typically found at once, having a one-level (un-nested) partition of the data and every point must belong to exactly one group. Partitioning methods use an iterative relocation technique, in which, given the number of clusters, the method creates an initial solution and then several iterations are made to minimize the intra-cluster distances and to maximize the inter-cluster distances. K-means and K-medoids algorithms are the most used to compute distance between objects in partitional clustering (Han et al., 2011b).

On the other hand, hierarchical clustering produces a nested sequence of partitions and each intermediate level is a division of the cluster from the next higher level into smaller clusters.

Hierarchical clustering can be classified as being either agglomerative or divisive, depending on how hierarchical decay is made. The agglomerative clustering starts with points as individual clusters and at each step of the methodology and the most similar or closest are aggregated in one bigger cluster, until all the points are in one single cluster. The divisive clustering starts with a cluster with all the objects and at each step, each cluster is separated into several smaller sub clusters, until only singleton clusters of individual points remain.

The output of a hierarchical clustering algorithm can be graphically displayed as a tree, called dendrogram, see Figure 4.1. By looking at the dendrogram it is possible to see which clusters are formed at each cut and the hierarchy between them, i.e. it is possible to see to which cluster a given client would be assigned if a smaller number of clusters were chosen.

In this study hierarchical clustering was chosen, as one of its main advantages is that once the dendrogram is built, it is possible to choose the number of clusters by cutting the tree at different levels to obtain different clustering solutions without having to run the clustering algorithm again.

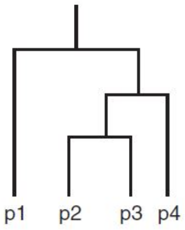


Figure 4.1: Dendrogram Example. Source: Tan et al. (2016).

In data mining applications, such as clustering, we need ways to assess how alike or unalike objects are in comparison to one another and then this values are stored in a matrix called dissimilarity matrix (Han et al., 2011b).

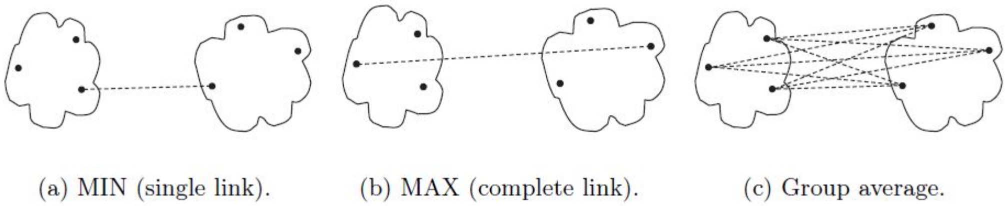
As previously mentioned, all the variables used to cluster customers are the result of previous segmentations, and all of them are categorical with the vast majority admitting more than two levels. To assess the similarity between two different clients (i, j) using the above mentioned variables, the most suitable measure and the one used in this project, is the simple matching coefficient, whose mathematical expression is presented below, in Equation 4.1.

$$d(i,j) = \frac{p-m}{p} \tag{4.1}$$

where m is the number of matches and p the total number of variables.

Once we have computed the similarity matrix between clients, the following step is to quantify the distance between clusters that should be computed after each merge, in Figure 4.2 three different ways of computing cluster proximity are represented. On the single link, distance between clusters is determined by the distance of the two closest objects, in the complete link the opposite happens, i.e. the distance between clusters is the greatest distance between any two objects. Group average computes cluster proximity by computing the average distance between all pairs of points from different clusters.

Along with these popular methods to compute cluster proximity, there is another method, called Ward's. This is the correct hierarchical analog to K-means, since the proximity between two clusters is computed by the distance between the centroids of clusters (Tan et al., 2016). The Ward's method joins clusters so that, after each step, the within-cluster variance is minimized across all clusters. This method was the one applied to create the behavioral clusters since it has a tendency to remove small clusters and to produce more similarly-sized clusters (Löster, 2017).


(a) MIN (single link).

(b) MAX (complete link).

(c) Group average.

Figure 4.2: Graph-based definitions of cluster proximity. Source: Tan et al. (2016).

4.1.2.1 Determining the optimal number of clusters

From a business point of view, a balance should be sought between the number of clusters chosen and its ability to create distinct clusters. Consider two extreme scenarios. If the entire dataset was included in the same cluster, this would maximize the compression of the data, but there is no point in doing clusters. On the other hand, if there are as much clusters as observations this will reduce the sum of within-cluster variance of each cluster to the minimum (Han et al., 2011b). However, a very high number of clusters may turn their implementation in a business context not actionable.

To assess the optimal number of clusters (K) statistical metrics and dendrograms can be useful tools. The statistic metrics more appropriated to deal with hierarchical clustering are pseudo F index and Pseudo T-Squared index. The pseudo F index is the ratio of between-cluster variation to within-cluster variation, see Equation 4.2, and peaks of F score index are indicators of greater separation between clusters (Wilkinson et al., 2012).

$$PseudoF = \frac{\frac{GSS}{K-1}}{\frac{WSS}{N-K}} \tag{4.2}$$

where N is the number of observations, K is the number of clusters at any step in the hierarchical clustering, GSS is the between-group sum of squares and WSS is the within group sum of squares.

The Pseudo-T^2 Index is a ratio of the sum of squared errors when the merging clusters remain separate to the sum of squared errors when the merging clusters are joined, see Equation 4.3, and a candidate to clustering solution is achieved when the value of Pseudo-T^2 falls or has a trough (Larson, 1993).

$$Pseudo - T^2 = \frac{B_{KL}}{\frac{W_K+W_L}{N_K+N_L-2}} \tag{4.3}$$

where K and L are the clusters merged to form a new cluster, N_k and N_L are the number of observations in clusters K and L, W_k and W_L are within cluster sum of squares of clusters K and L, and B_{KL} is the between-cluster sum of squares.

As previously mentioned, the dendrogram can also be used to determine the number of clusters as the length of the lines can be a good proxy to analyze the distance between clusters.

4.1.3 Clusters Generalization

In this study, due to the computational effort required to compute the dissimilarity matrix a sample of 10 000 clients was selected to generate the clusters. Once the clusters for this sample were formed, resorting to the previous steps presented, there was the need to generalize these results to all customers, i.e. each of the customers was assigned to one of the 5 clusters previously defined.

For this purpose, the classification algorithm K-Nearest Neighbor (KNN) was applied. It is a method based on learning by analogy, which means that, when faced with an unknown tuple, this method investigates in the search space the most similar training tuples to it and the class assigned to this object will be the majority class of the K neighbours identified as being more similar to this one (Han et al., 2011b).

In this study the K chosen was 1. The goal in doing so, is to compare customers without a cluster assigned with all the other customers with a cluster already attributed and assign to the customer, whose cluster is not known, the same cluster as the customer most similar to it. In this method there is the need to compute the “closeness” between customers and to ensure coherence between this method and the method used for the definition of clusters, the metric used to compare the proximity between customers was also the simple matching coefficient (see Equation 4.1). With this approach each client was assigned to one of the five clusters previously created.

4.2 Churn Prediction Model

Based on the assumption that customers belonging to the same cluster will behave similarly, not only from the different perspectives already studied by Sonae MC, but also in terms of churn behavior, the next step was the development of a predictive churn model for each cluster. In this section is presented the construction of the predictive models, which involved the following steps: (i) Churn Concept Definition; (ii) Data Preparation and Understanding; (iii) Selection of the non cost-sensitive models, its application and assessment, and finally (iv) Selection of the cost-sensitive model and its assessment.

4.2.1 Churn Concept Definition

As outlined in the literature review section, the definition of the criteria under which a customer is considered a churner may vary depending on the company's objectives. In the case of food retail, the definition of the concept of churn is particularly important, as churn can occur gradually and there is no need for a formal cessation of the contract.

In this project, after a brainstorm with the parties involved, it was decided to adopt the concept of total churn with a static definition, which means that a client will be considered a churner if he/she does not make any purchases at Continente, Continente Modelo, Continente Bom Dia or Continente Online during a predefined period. Hence, the next phase of the project involved the definition of the period after which a client is considered a churner for these insignias.

Although most studies consider the same period for all customers, from a business point of view that is not the ideal. For instance, the period for classifying a frequent customer as a churner should be shorter than for an infrequent one. Choosing the same period for everyone may imply, if the period is long, that retention actions are not taken at a reasonable time for frequent customers. On the other hand, if the period is short it could mean that most of the infrequent customers will be identified as churners when in fact it is part of their natural behavior. Therefore, different target periods were defined for each of the clusters.

The methodology underlying the definition of the target period involved the study of the evolution of the churn rate for each cluster when considering different target periods.

To compute the churn rate, the following formula was used:

$$ChurnRate = \frac{x-y}{x} \tag{4.4}$$

where x is the number of customers at the beginning of the period (base number of customers) and y the number of base customers who made a purchase in the target period.

To compute the churn rate it is necessary to define a base number of customers. For this analysis only customers who made at least a purchase in the month before the analysis were considered, to ensure that they were not already churners as this could bias the analysis (Veloso, 2013).

The churn rate for each target period was computed in different periods of the year to reduce the impact that eventual outliers and seasonality could have on the definition of the target period. In this analysis weekly target periods were analyzed, where the following period results in adding one more week to the preceding one.

To clarify the procedure described, please consider as an example the churn rate computed for a target period of 1 week. This will be computed at the beginning of each month, where base customers will be the customers that have performed at least a purchase in the past month and the churn rate will be the percentage of those customers who have not made a purchase in the following week. Using the same target period these computations will be repeated for all the months of the year. After having computed the different churn rates for each target period, values that distance more than 2 standard deviation from the average churn rate will be removed to exclude possible outliers and seasonal effects. Then all this procedure will be repeated for the next target period, that in this analysis is the previous one added of one week. This analysis was performed for up to the maximum target period of 16 weeks.

As the evolution of the churn rate was not enough to conclude which target period should be considered, at each point the tangent line was drawn and finally the difference of slopes between them was computed. The goal in doing so is to determine the target period from which the evolution of churn rate stabilizes and approaches to zero, which means that lengthening the target period does not mean a substantial increase in the churn rate.

Once the target periods for each cluster had been defined, the next step was to collect the data needed to develop the predictive churn model.

4.2.2 Data Collection and Preparation

The phase of data collection, understanding and preparation is one of the most critical steps in KDD process, once the quality of the inputs can strongly impact the quality of the results of the model.

One of the capabilities that a quality model must have is to be capable of generalizing for the future, based on past data, and not just memorizing what happened at a particular moment in the past (Yan et al., 2001). To this end, there are some strategies related to the extraction of data that can be used such as training the model using different time window datasets that are then gathered in a single training set (see Figure 4.3). One of the disadvantages of adopting this strategy is to increase the size of the training dataset and thus its learning time (Yan et al., 2001).

Real datasets are highly susceptible to noisy data and, in order to avoid a "garbage in garbage out" situation, a data mining analyst typically spends about 80% of the time preparing the data (Refaat, 2010).

Data preprocessing can comprise several steps such as data cleaning, data integration, data transformation and data reduction. In most of these steps, business knowledge and information collected during the process of data understanding are essential to support decisions.

In this project, the removal of inconsistent data that may result from information entry errors or equipment malfunctioning was one of the first steps of data preprocessing performed. Treatment of missing values was also performed. This step is especially important when dealing with Regularized logistic regression.

In what regards to the treatment of outliers this is the phase where knowledge of the business is of utmost importance to ensure that values that may be relevant to the model are not eliminated

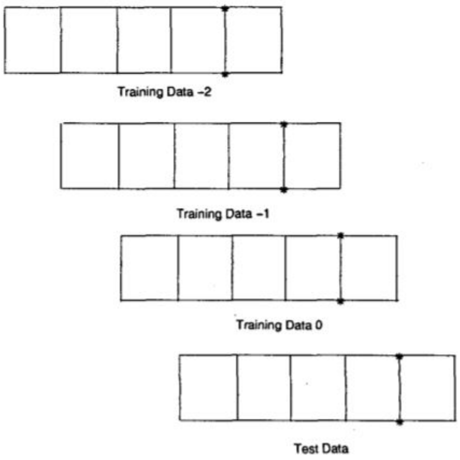


Figure 4.3: Extract training data from several shifted time windows. Source: Yan et al. (2001)

or replaced. Outliers can be very informative and their removal in an undifferentiated way can make the model poorly adapted to customers who have more extreme behaviors. To this end, a descriptive analysis of the data should be done to understand which criteria should be used to define outliers.

There are several ways to identify outliers. A common practice is computing the mean of values and their standard deviation and then values that distance from the mean more than two or three standard deviations are considered outliers.

For skewed distributions, it is more informative to also provide the two quartiles, along with the median (Han et al., 2011b). One simple rule of thumb for finding outliers is based on the quartiles of the data where box-plots are drawn, (see Figure 4.4) and metrics such as lower quartil (Q_1), upper quartil (Q_3) and interquartile range (IQR) are computed and values falling at least $1.5 \times IQR$ above the Q_3 or below the Q_1 are considered outliers.

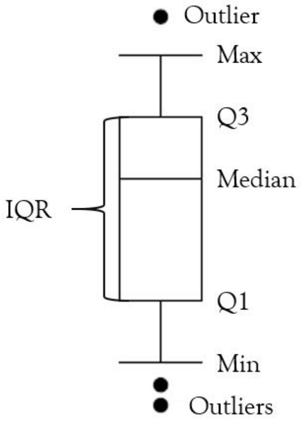


Figure 4.4: Boxplot to visualize outliers. Source:Han et al. (2011a).

Instead of using $1.5 \times IQR$ in some cases analysts use $3 \times IQR$. By doing this, only extreme values are identified as outliers. Detection of outliers can also be done resorting to clustering techniques where similar values are grouped into clusters and the remaining ones are considered as outliers and therefore disregarded.

Once the outliers are identified, the most commonly adopted methodologies to handle them are the elimination of values, bins construction and regression, in which the outliers are replaced by the result of one function.

Another aspect to take into account when treating data is that sometimes the data collected for various predictor variables may differ in their measurement units, which can have an impact on the model, since variables with higher magnitude will have more influence over other ones. To prevent this situation from happening data should be normalized into one single scale. Some of the most popular strategies to normalize data are Z-Score and Min-Max normalization.

In Z-Score normalization data is transformed in such a way that the resulting distribution has a mean of 0 and a standard deviation of 1, while Min-Max normalization performs linear transformations on the original data so that the values fall within the range of values specified (Hodeghatta and Nayak, 2016). In this study to smooth the impact of outliers' data was normalized resorting to Z-Score normalization.

After the data was processed, the dataset should be separated into at least two separate datasets, i.e. the training dataset, which will be used to build the model, and the test dataset, which will be used to evaluate the model's performance. The larger the training data, the better the classifier and the larger the test data, the more accurate the performance measures.

The following step is to select the most suitable models to predict churn in this context.

4.2.3 Model Selection

The churn problem is generally conceptualized as being a classification problem as customers can be classified as being either churners or non-churners.

As mentioned in the literature review, there are several data mining techniques that can be used in classification problems, however some stand out in terms of the interpretability of their outputs, such as Logistic Regression and Decision trees.

4.2.3.1 Regularized logistic regression

The logistic regression model is used to examine the relationship between a binary response variable and one or more categorical or numerical predictive variables. This technique allows the estimation of the probability of occurrence of a given event and to identify which independent variables contribute more to explain the response variable.

To ensure that the output of the function, i.e. the probability, is between 0 and 1, the logit function is applied. Its inverse (see Equation 4.5) relates the response variable to the input variables. The log-likelihood function of the logit transformation of Equation 4.5 is defined by the Equation 4.6.

$$\pi_i = p\left(y_i = 1|x_i\right) = \frac{exp\left(x_i^T\beta\right)}{1+exp\left(x_i^T\beta\right)} \quad ,i = 1,2,...,n \tag{4.5}$$

where $y_i \in \{0,1\}$ is the response variable for observation i, $i = 1,2,...,n$. X is the input variables and $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ the vector of unknown coefficients.

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \, log\left(\pi_i\right) + (1-y_i)\, log\left(1-\pi_i\right)\} \tag{4.6}$$

Logistic Regression to be reliably applied require certain assumptions to be met, such as none or little multicollinearity among the independent variables. In this sense, regularized methods, will be used in this study, due to their capacity of controlling the effects of overfitting and multicollinearity (Algamal and Lee, 2015).

In Regularized logistic regression (RLR), see Equation 4.7, the objective is to minimize the negative log-likelihood function, but in this case albeit with a penalty term P. The goal of using the penalty term is to constrain the size of the coefficients, such that the only way the coefficients can increase is if a comparable decrease in the negative log-likelihood function is experienced (Boehmke and Greenwell, 2019). The main benefits pointed out of using Regularized logistic regression are the fact that classification accuracy can be improved by shrinking the regression coefficients, and, as it selects the features that exhibit the strongest effects, the number of features will be reduced, which will make the model easier to interpret.

$$RLR = -\ell(\beta) + P \tag{4.7}$$

This method is a soft thresholding method which means that it inherently incorporates feature selection, by reducing the coefficients of irrelevant variables to close to zero and in some cases may even eliminate them. This approach when compared with hard thresholding feature selection, which includes the most traditional techniques of feature selection such as forward selection and backward elimination, presents several benefits as hard thresholding feature selection is computationally inefficient, does not scale well and only admits including or not a variable.

There are several approaches in regularized regression (RR) that differ between them in the way the penalty term is computed. The most common penalty parameters are Ridge, Lasso and Elastic Net.

The Ridge regression minimizes Equation 4.7, subject to a $L2$ penalty and the estimation of vector β is computed using the Equation 4.8. In this method the coefficients of irrelevant variables are pushed towards zero, as well as coefficients of correlated features are pushed towards each other. Ridge model does not perform feature selection, as it reduces coefficients without canceling any of them. As all variables are included, this model should be preferred over the other ones if the inclusion of all features is considered important.

$$\hat{\beta}_{Ridge} = argmin_{\beta} \left[-\ell(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \tag{4.8}$$

where λ is the tuning parameter, which controls the trade-off between fitting the data to the model and the effect of regularization.

The Lasso Model imposes a $L1$ penalty on the regression coefficients (see Equation 4.9). It is very similar to the Ridge model, however it performs automated feature selection, that is, in this model the coefficients may even reach the value zero. In high dimensional datasets the Lasso model can be used to identify and select the features that are more relevant, i.e. with the largest (and most consistent) signal.

$$\hat{\beta}_{Lasso} = argmin_{\beta} \left[-\ell(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{4.9}$$

Finally, there is the Elastic Net model that is in between the methods above mentioned, see Equation 4.10.

$$\hat{\beta}_{Elastic} = argmin_{\beta} \left[-\ell(\beta) + \lambda \left((1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \right] \tag{4.10}$$

where α controls the “mix” of Ridge and Lasso regularization

While Lasso method performs feature selection, the Ridge method is more systematic when dealing with correlated features. So, Elastic Net by combining both performs effective regularization via the Ridge penalty with feature selection characteristics of the Lasso penalty (Boehmke and Greenwell, 2019). It is also worthy of mention that the Elastic Net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together (Zou and Hastie, 2005).

In this study we applied RLR resorting to Elastic Net method, as it inherently performs feature selection and, from the methods presented above, is the most appropriated to deal with highly correlated variables.

To apply this method it is necessary to define the tuning parameter, λ, which controls the size of the penalty, and, in addition, there is the need to determine the α parameter in which values ranging from 0 to 1 correspond to Elastic Net, when $\alpha = 0$ corresponds to Ridge and $\alpha = 1$ to Lasso. Both these parameters must be optimized in a join manner.

4.2.3.2 Decision tree

In previous studies Decision trees have proven to outperform several methods, in terms of predictive performance, as well as, in terms of interpretability of results.

Decision tree, also known as classification tree, is defined by Aggarwal (2014) as being a technique for partitioning data into homogeneous groups, which is accomplished by dividing the training dataset into smaller subsets, until each subset contains only one class or there are no remaining attributes for further partitioning.

The output of a Decision tree is a tree-shaped structure where, in an intuitive way, it is possible to access sets of mutually exclusive rules that make possible the classification of the target variable as belonging to the positive class or not. The construction of the tree is done by choosing, based on a split criterion such as information gain or gini index, the attribute that best divides the data. This division is done recursively until the tree is complete. According to Raileanu and Stoffel (2004), it is not obvious which of the split criterion produces the best results for a given data set.

When using the information gain, the goal is to split the data in such a way that the information gain is the highest. The greater the loss of entropy, a measure of homogeneity of the dataset, the greater the information gain. The entropy is calculated after each split (see Equation 4.11) and the information gain is the measure of decrease in entropy due to the split. The condition that ensures the highest gain is the one chosen to split the data.

$$Ent(S) = \sum_{c=1}^{N} p_c \log_2(p_c) \tag{4.11}$$

where S is the training set and p_c the relative frequency of C_c in S.

Another metric to evaluate the quality of a node, in terms of discrimination between different classes, is the gini index. It measures the “purity” of a node. If there is only one class present in the data then it is maximally pure, if there is a mix of elements from different classes in the node then the node is impure (Molnar, 2019). The greater the purity of a node, the lower the value of the gini index (Equation 4.12). So, the attribute that provides the smallest gini index after the split is the one chosen to split the node.

$$Gini(S) = 1 - \sum_{c=1}^{N} p_c^2 \tag{4.12}$$

where S is the data set with N classes and p_c the relative frequency of class c in S.

Some of the advantages of using Decision Trees that can be pointed out are that they are simple to understand and interpret, can handle both numerical and categorical data and their performance is not affected by nonlinear parameters. Some of its disadvantages are related with the fact that small variations can make the model unstable and in some cases overfitting may occur, which makes the model poorly perform for untrained data. Overfitting is typically overcome by pruning the tree to remove nodes that may lead to it. In this study, to prevent overfitting it was defined the maximum depth of the tree. To sum up, when using decision tree two parameters were defined: the splitting criterion and the maximum depth of the tree.

In both data mining techniques, i.e. Regularized logistic regression and Decision Tree, there is the need to determine the best parameters and to do so a cross validation strategy known as K-fold Cross Validation was applied. With this approach, the instances in the evaluation data set are split into K mutually exclusive subsets of approximately equal size. In each learning-and-testing process the model is trained using $K-1$ subsets and is tested in the remaining one. This process is repeated K times. The goal in doing so, is to ensure that there is no overfitting when choosing the parameters.

Some authors, such as Hastie et al. (2009), encourage a 5-fold or 10-fold cross-validation. Therefore in this study, when needed, a cross-validation, a 10-fold validation was adopted. To assess the quality of the model, several metrics can be used and the ones analyzed in this study are presented below.

4.2.4 Classifier Performance Evaluation

There are several metrics that can be computed to assess the performance of a model. One of the most common to evaluate the predictive capability of a classification model is the accuracy, that measures the percentage of instances correctly classified by the model, which can be obtained through the confusion matrix (see Table 2.1), by computing the ratio presented in Equation 4.13.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.13}$$

To measure the performance of the binary prediction models, we compute the receiver operating characteristic curve (ROC). This curve, characterizes the trade-off between true positives and false alarms for every individual cut-off. The true positive rate is the percentage of elements that are churners and were correctly classified as being so. False alarms or false positive rate, are the proportion of false non-churners, computed by dividing the number of elements misclassified as churners by the number of non-churner elements. The AUC (Area Under the Curve) is higher when the ROC Curve raises abruptly, meaning that models with a higher AUC are better at discriminating between classes. A model with an AUC close to 1.0 is perfect at discriminating, while an AUC close to 0.5 has apparently poor discrimination ability (Hanley and McNeil, 1982). When comparing models by their predictive capacity AUC is preferred over accuracy due to its independence from the chosen threshold (Ling et al., 2003).

Lift is a widely used indicator in churn prediction by assessing how much better the model is when compared to a random classification. This metric allows us to analyze how much larger the proportion of customers who actually drop out on customers classified as churners when compared with the population of customers as a whole (see Equation 4.14). On some occasions, when there are budgetary constrains, lift can help companies select the customers most likely to churn and use them as target of specific campaigns (Vuk and Curk, 2006).

$$Lift = \frac{Precision}{\frac{P}{P + NP}} = \frac{\frac{TP}{TP + FP}}{\frac{TP + FN}{TP + TN + FP + FN}} \tag{4.14}$$

These metrics are some of the traditional metrics that can be used to assess the performance of the models, but all of them tacitly evaluate misclassifications and correct classifications with the same weight, although they actually have different impacts on the company.

When models are intended to be implemented in real life contexts and to support decisions that aim to maximize value for the company, such as this one, which is intended to maximize the value of Sonae MC's retention campaigns, it is important to take the financial implications of classifying a customer as a churner or not into account.

Thus, it is pointless for the model to evaluate misclassifications and correct classifications with the same weight and to assume that costs are equal for all customers. Not all customers compensate the effort of retaining them and there are others who, despite having a lower likelihood of churning, should receive an offer to maintain the relationship due to its value to the company.

4.2.5 Cost-Sensitive Classification

In an attempt to maximize the results of retention campaigns, real financial costs were incorporated throughout a cost sensitive classification and to ensure that costs associated with misclassification and correct classification vary, not only between classes but also across clients, an example-dependent-cost-sensitive algorithm was implemented.

In this study we implemented Bayes minimum risk, which is an after training algorithm. As mentioned in the literature review section, by using this algorithm, we are computing the expected cost of classifying each customer as a churner and as a non-churner and in the end the customer will be classified as belonging to the class that presents lower risk for the company.

In the case of binary classification, the risk of predicting the example as belonging to the negative or positive class is calculated using the Equations 4.15 and 4.16, respectively. One of the requirements of this method is that the input of the model has to be the actual probability of the event happening, i.e. p_i.

According to Bahnsen et al. (2014), the MBR model leads to better results, in the sense of higher savings, regardless of the algorithm used for estimating the probabilities. Therefore in this study, we will use as input to MBR the probabilities of churning previously computed using Regularized logistic regression.

$$R(c_i = 0|X) = C_{TNi} \times (1 - \widehat{p_i}) + C_{FNi} \times \widehat{p_i} \tag{4.15}$$

$$R(c_i = 1|X) = C_{TPi} \times \widehat{p_i} + C_{FPi} \times (1 - \widehat{p_i}) \tag{4.16}$$

where $\widehat{p_i}$ is the estimated positive probability for example i.

In order to compute the risk associated with classifying a customer as churner or non-churner we need to compute the financial impact of the different decisions, ie. false positives, false negatives, true positives and true negatives, for each customer and to do so, we used a similar approach to the one defined by Bahnsen et al. (2015b), which was analyzed in detail in the literature review section (see Figure 2.3 and Table 2.3). This approach was designed to maximize the results of retention campaigns. Nevertheless, the computations should be adjusted to the reality of the company in question, depending on the way retention campaigns are managed.

The example of food retail is a very specific one. Customers decide where to buy on a daily basis and when a customer is classified as churner in the target period, there is no point in saying that it will also be churner in the following periods. Due to this, the amount lost for the company if the client actually withdraws the relationship, CLV, will be the expenses of the client during the target period, that in this case will be the expenses of the client in the previous year distributed uniformly in the target period, which we call SV_{TarPer}.

Regarding the cost of the offer, as the target period considered often differs from the coupon delivery period, is was agreed with the parties involved, that the cost of each offer (C_o) will be the same annual value offered as nowadays but distributed uniformly in the target periods defined.

Once a class has been assigned to all customers using the aforementioned methodology, is now time to compute the costs (see Equation 2.1) and the savings associated with using MBR (see Equation 4.17).

$$Savings(f(S)) = \frac{Cost(f_l(S)) - Cost(f(S))}{Cost(f_l(S))} \tag{4.17}$$

where $Cost(f_l(S)) = min\{Cost(f_0(S)), Cost(f_1(S))\}$, f_0 refers to a classifier that predicts all the examples in S as belonging to the class c_0, while in f_1 all the examples belong to the class c_1.

In previous studies the savings were computed comparing the cost-sensitive model with the options of classifying everyone as a churner or as non-churner, see Equation 4.17. However, in this case Sonae MC has already a culture associated with offering coupons to customers based on predictive models and a total change in the dynamic would not be suitable. Given this, we adjust the computation of the savings to the reality of Sonae MC, which means that the savings of this model will be assessed by comparing the costs when using the cost-sensitive methodology with the costs when using a non-cost sensitive methodology to classify customers, such as the Regularized logistic regression .

Chapter 5

Results

This chapter presents the main results of the application of the methodology described above, using data from the Cartão Continente database. It is divided in three sections, i.e. first one concerning the construction of the behavioral clusters, the second one concerning the development and evaluation of the predictive churn model, which uses as input the results of the clustering algorithm, and finally are presented some of the managerial implications that result from this study.

5.1 Behavioral Clusters

5.1.1 Data Collection and Understanding

The collection and preprocessing of data from Cartão Continente database was conducted using SQL language and SAS software combined.

Due to computational effort required to analyze all the clients from the Sonae MC database, a sample of 10% of active clients[1], containing around 400 000 customers, was explored.

As stated in the presentation of the methodology, this study relies on variables previously constructed by Sonae MC that seek to characterize the customer and its behavior at Sonae MC from different perspectives. A summary of the variables used to create clusters is presented in Table 5.1.

It is not imperative that all customers have a segment assigned in all segmentations. In some of the segments there are requirements for the customer assignment, such as having performed a minimum number of transactions in the past months. For this analysis, it was decided not to remove these data but to assign them to the “No Value” segment, as the lack of segment is in itself an indicator of customers’ behavior.

[1] Clients with at least one transaction in Continente stores in the last year.

Table 5.1: Data used to build Cluster

Variables	Description	Levels
Customer_id	Customer Account Number	
Segm_lifestyle	Lifestyle Segmentation	SL_1; SL_2; SL_3; SL_4; SL_5; SL_6; SL_7
Segm_Value	Value Segmentation	SV_1; SV_2; SV_3; SV_4; SV_5; SV_6; SV_7
Segm_Lifestage	Lifestage Segmentation	SLT_1; SLT_2; SLT_3; SLT_4; SLT_5
Segm_Engagment	Engagement Segmentation	SE_1; SE_2; SE_3; SE_4; SE_5
Segm_SOW	Share of Wallet Segmentation	SOW_1; SOW_2; SOW_3; SOW_4
Segm_Baby	Baby Segmentation	YES; NO
Segm_Junior	Junior Segmentation	YES; NO
Segm_Cacapromo	Abusive promotional behavior Segmentation	SCP_1; SCP_2; SCP_3
Segm_Grocer	Grocer Segmentation	SG_1; SG_2
Segm_Pay	Payment Day Segmentation	SP_1; SP_2; SP_3; SP_4
Segm_PSS	Price Sensitivity Segmentation	SSS_1; SSS_2; SSS_3; SSS_4; SSS_5; SSS_6; SSS_7; SSS_8
Segm_Gender	Gender Segmentation	SG_1; SG_2: SG_3
Pref_Insignia	Customer Preferred Insignia	Continente; Continente Bom Dia; Continente Modelo
Pref_Purchasemission	Customer main pruchase Mission	PM_1; PM_2; PM_3; PM_4; PM_5; PM_6
Pref_Channel	Customer Preferred Channel	Online; Physical

5.1.2 Modeling Clusters

The modeling of behavioral clusters was done using SAS software, while its generalization was coded using R programming language.

When creating clusters, due to the computational effort required to classify all the above-mentioned clients, a sample of 10 000 clients was selected. The sample was stratified according to the lifestyle segmentation, as this is the one that most closely resembles the natural distribution of the customers.

To assess the optimal number of clusters, both statistical metrics and dendrogram were analyzed. The analysis of the statistical criteria (see Figure 5.1) shows that Pseudo F index presents a high value for K=2 and then does not present a clear peak, so priority was given to Pseudo-T^2 index, which decreases abruptly when K=5, suggesting that 5 is a good candidate for the number of clusters. The analysis of the dendrogram obtained through the clustering process (see Figure 5.2), shows that the cut K=5 stands out due to its ability to discriminate clusters. Having that in mind, customers will be segregated into 5 clusters, given that it is a reasonable number both from an analytical point of view, as previously seen, as well as from a business point of view, as is totally feasible to deal with 5 segments of clients in churn retention campaigns.

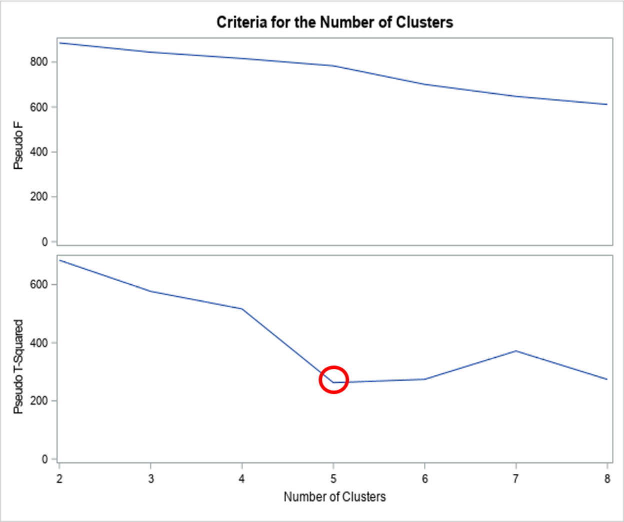


Figure 5.1: F-Score Index and Pseudo-T^2 index of behavioral clustering

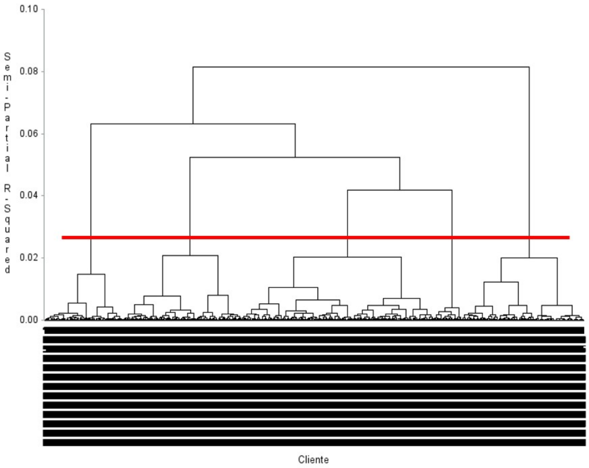


Figure 5.2: Dendrogram of behavioral clustering

5.1.3 Clusters Characterization and generalization

The 5 clusters obtained were characterized by comparing the distribution of the different customer DNA segmentations in each of the clusters with their natural distribution in the sample used.

A more detailed description of the clusters is presented below.

Premium:

This cluster stands out for including customers really involved with Continente, with high levels of loyalty and includes customers whose spending is mostly done in Continente (high share of wallet). The customer lifestyle that stands out is the “Urbanos Sofisticados”. These are customers that value quality and are willing to pay for it. In terms of lifestage segmentation, the segments highlighted are “Family Supporters” and “Family with kids”, i.e. senior consumers (more than 65 years old) with children in their household and families with kids.

In this cluster, online purchases have a high relevance and the purchase mission that stands out is “Abastecimento”, which means that the purchases of these customers are typically of high value and include a wide variety of products. In this group, the rate of “Grocers” is higher than in the others. This may be due to the large volume spent in Continente and the high frequency of stores visits.

Every Day Low Price:

In this cluster, economic customers stand out. However, this segment is not very price sensitive and prefers own brand products, which means that they are customers who typically buy own brand products and do not change their purchase pattern if others are on promotion. This cluster is essentially made up of smaller households, with “Active Adults” and families where children are not present. They are frequent customers whose share of wallet is medium to low.

Kids in the house:

The lifestyle segment more present in this cluster is “Pais Práticos”. Typically these customers have children and value practical and affordable products. In terms of lifestage segment, “Family with kids” and “Family Supporters” stand out and, as expected, the index related with having babies or juniors is high. In this cluster, the SOW is medium to low, and the most relevant purchasing missions are “Orientada”, which stands out for including non-food product usually with promotional influence, and “Suprir Faltas”, that includes some non-fresh items that fill specific purposes.

No Value:

This cluster concentrates customers that are typically weak clients, that do not meet admission conditions for segmentations. These customers visit stores only occasionally and in this segment customers are very price sensitive and the most relevant purchase mission is “Orientada”, which means that when customers come to Continente they are looking for non-food products with a promotional influence.

Promo Busters:

In this cluster customers are very price sensitive, typically visit stores occasionally and are customers with low to medium share of wallet. In this cluster the lifestage segments that stand out are “Seniors” and “Family Supporters”, who are typically customers over 65 years old. Concerning lifestyle segments, the ones with higher emphasis are “Tradicionais Frequentes”, which includes senior customers with time availability and that have a cautious relationship with money, and “Promocionais Atentos”, who are customers motivated by promotional activities and not by the retailer who promotes them.

It is possible to access clusters proximity by analyzing the dendrogram and the partition sequence during its construction, see Figure 5.3. The “No Value” group clearly differs from the others, having been separated right from the beginning, it can also be concluded that the groups closest to each other are “Every Day Low Price” and “Promo Busters”, where the economical factor seems to be the most relevant.

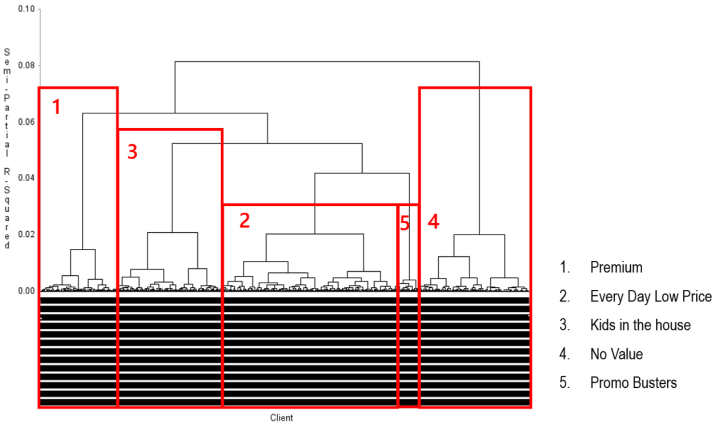


Figure 5.3: Identification of clusters in dendrogram

Looking at some relevant business metrics (see Table 5.2), it can be concluded that the cluster with the highest percentage of customers is the “Every Day Low Price” and although the customers from the “Premium” cluster only represent 14.6% of the total number of customers considered, they are the ones that most spend on Continente, representing around 44% of the sales of all the clients analyzed. They also have the highest average basket[2] and are the most frequent, having five days as the average number of days between transactions. On the other hand, “No Value” clients only account for 2.4% of sales, despite being 22.7% of the customers considered. Their average basket is the lowest and are the least frequent, having almost one month of average time between transactions.

Table 5.2: Relevant business metrics about clusters before generalization

Clusters	**Percentage of Customers**	**Percentage of Sales**	**Average Basket[a] (€)**	**Average number of days between transactions**
1. Premium	14.6	43.9	15.0	4.9
2. Every Day Low Price	35.7	30.7	9.0	8.7
3. Kids in the house	22.0	19.1	9.9	10.0
4. No Value	22.7	2.4	7.5	26.0
5. Promo Busters	5.0	3.9	9.3	11.6

[a] These values were multiplied by a constant for confidentiality issues.

As the clusters were created considering only a sample of 10 000 customers, there was the need to generalize them for the remaining customers in the original sample. For this purpose, the KNN method with K=1 was used with simple matching coefficient as proximity measure. In the

[2] Amount spent per transaction

Table 5.3, business metrics related to this cluster are presented and as expected they are really similar to those computed for the 10 000 customers.

Table 5.3: Relevant business metrics about clusters after generalization

Clusters	**Percentage of Customers**	**Percentage of Sales**	**Average Basket**[a] **(€)**	**Average number of days between transactions**
1. Premium	15.5	45.4	15.0	5.0
2. Every Day Low Price	41.4	34.1	9.0	9.2
3. Kids in the house	16.2	13.8	9.9	10.0
4. No Value	22.1	3.1	8.4	25.4
5. Promo Busters	4.80	3.6	9.3	12.3

[a] These values were multiplied by a constant for confidentiality issues.

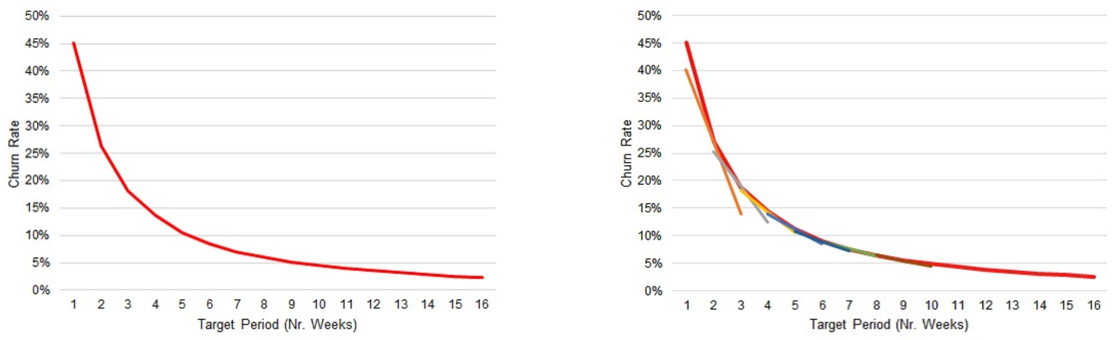
5.2 Churn Prediction Model

5.2.1 Churn Concept Definition

Once the clusters were formed, the next step was the definition of the target period for each of them. Having different periods for each cluster allows to carry out targeted retention campaigns, which are expected to enhance their effectiveness.

To identify the target period for each cluster, an analysis of the evolution of the churn rate for increasing target periods was performed. In this study weekly target periods were analyzed, where each target period results of adding one more week to the preceding one, up to the maximum target period of 16 weeks.

The graphic of the churn rate evolution in function of the target period for cluster "Every Day Low Price" is presented in the Figure 5.4a. When the target period increases, the churn rate diminishes as expected, since a customer has more time to buy before being considered a churner.



(a) Simple

(b) With tangent in each point

Figure 5.4: Evolution of churn rate dependent on target period for cluster "Every Day Low Price"

Only by the visual inspection of the graphics, it is difficult to conclude which target period to consider, so the tangent line at each point was drawn (see Figure 5.4b), and the difference between the slopes at each point was computed (see Figure 5.5). The target period chosen is the one in which the slope of the tangents stabilizes close to zero, which means that increasing the target period would not imply a substantial decrease in the churn rate. As a result, the target period for cluster "Every Day Low Price" is 7 weeks, see Figure 5.5.

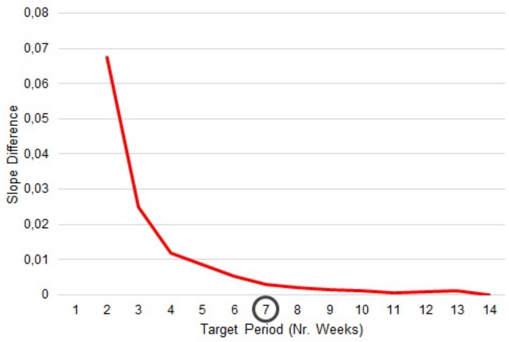


Figure 5.5: Evolution of slope difference

The same methodology was applied to all clusters (see Appendix B) and the target periods defined are presented in Table 5.4.

Table 5.4: Target periods for each cluster

Clusters	**Target Period (Weeks)**
1. Premium	4
2. Every Day Low Price	7
3. Kids in the house	7
4. No Value	9
5. Promo Busters	7

5.2.2 Data Collection and Preparation

The extraction of the data was performed combining SAS software and SQL. In the construction of the model, data from Cartão Continente databases was used.

The choice of the variables to include in the model depended both on business knowledge, which allowed the selection of the variables that are expected to be more relevant, as well as the availability of information.

Despite the model being intended to predict churn in Continente, Continente Modelo, Continente Bom dia and Continente Online, the customer behavior in the several partners of Sonae MC was also included in the analysis, as it may be relevant to predict churn in the aforementioned insignias.

In this study we included more than 600 variables that embraced several dimensions of the client and his behavior which can be divided into:

1. **Socio-Demographic Information (SD):** Data regarding the client's demographics and history in the company, such as the client's age, address region, data regarding the longevity of clients of Continente and competition index associated to client preferential store;
2. **Behavioral Information (B):** Variables that relate to customer behavior in the ecosystem, e.g. if the customer has adhered to the app Cartão Continente and which insignias were visited by the customer in the previous year;
3. **Transactional Information (T):** Transactional information of the client over different time periods, which was important even when defining the target period;
4. **Behavioral Change Detection Variables (BC):** Variables related to changes in customer behavior that were expected to be relevant for predicting churn, such as a flag if the customer's most recent value segment is lower than the most frequent class in the previous year and the ratio of sales amount from different periods.

When collecting variables, we sought to analyze them considering different time periods, namely variables regarding the year prior to the analysis (12 months before), periods equal to the target periods (4, 7 and 9 weeks) immediately before the analysis, and finally variables regarding the customer's behavior during the homologous periods of the target periods in the last year (homologous 4, 7 and 9 weeks). A description of the variables included in the model can be found in Appendix C.

For this study, data from 2018 was extracted to train the model and validate parameters and data from 2019 was used to test. Several shifted time windows were used to make the model more stable. To be easily deployed in Sonae MC's current practices, the model will be run at the beginning of each month, so for each month of the two years' period we have collected the aforementioned variables to predict churn.

As previously mentioned, the customer is considered a churner if he/she has not made any purchase during the target period, previously defined. To ensure that the training of the model is not biased by customers that were already churners, only those customers who have made at least one purchase in the previous period, which in this case will be equal to the target period, were considered.

After extracting the data considered relevant for the study, an exploratory analysis was performed using graphic displays of basic statistical descriptions to visually inspect the data. The analysis of the longevity of customer account, (see Table D.1 in Appendix D), shows that the average longevity of "Premium" customers is higher than for the other customers and it also shows that churner customers have on average a lower account longevity than non-churners. This may mean that customers with a longer lasting relationship are customers with less likelihood of abandoning Continente.

Figure D.1 in Appendix D reveals that in churner customers, the percentage of customers whose region/address is unknown is higher, which can be justified by the fact that the customers are less involved and therefore provide little information about them or by the fact that it may be a cause of churn, as the customer does not receive retention coupons sent by mail and as a consequence may have lower retention rates.

As far as the brand analysis of the products purchased is concerned, in all clusters, customers spend most of their annual expenditure on supplier branded products. The cluster that least stands out in the percentage of purchases of own branded products is the “Promo Busters”. It also seems that in terms of the proportion of amount spend per brand, the behavior seems to be similar for both churners and non-churners.

Finally, we analyzed the correlations between variables, to assess if predictor variables are independent from each other, using Pearson correlation coefficients and, as expected, there are correlated variables, such as sales amount and number of transactions in the previous year at Continente insignias. Figure D.2a presents this correlation matrix for cluster “Premium”. Besides this, there are other variables that by definition are expected to be correlated, since the same variable is being analyzed considering different time periods, Figure D.2b presents the correlation matrix for cluster “Premium” for the variable amount of sales.

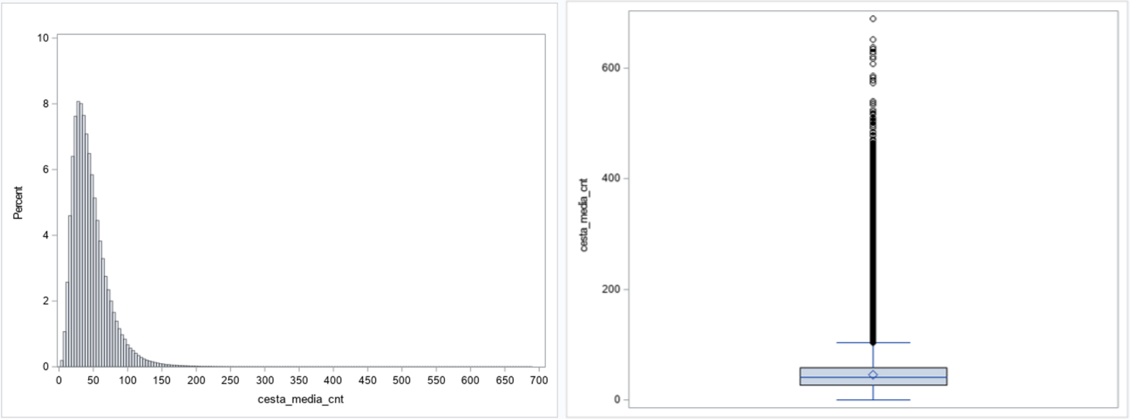
In the preprocessing phase, although the steps are the same for all clusters, they were performed separately for each cluster as each of them isolates clients with different behaviors and it is important to separate them specially when handling outliers.

The removal of inconsistent data and the treatment of missing data was performed in parallel. To treat the missing values in the case of categorical variables the value “NA” was defined to replace those values. In the case of numerical variables, substitution depended on the variables, in some cases the existence of missing values meant, for example, that the customer had no purchases in that period and these cases were replaced by zero, while in other cases such as age, the missing values were replaced by the average of the remaining observations.

As the variables considered for the treatment of outliers are skewed to right (take as example Figure 5.6), we chose to consider outliers only those observations that distance more than 3 times the interquartile range from the third quartile. It is not intended to remove all the observations that are considered outliers, because it could make the model poorly adjusted to clients with more extreme behaviors. Thus, it was decided to remove only those instances whose number of outliers was greater than 15% of the variables considered.

After the steps of the data preprocessing, it is clear, as typically happens in churn prediction studies, that the dataset is unbalanced with a higher number of non-churners than churners (see Table 5.5).

The class imbalance problem can impact the quality of the model as it could be biased towards the majority class, which would result in a poor performance of the model in predicting the positive class. To overcome this issue data must be balanced. There are multiple sampling strategies to balance dataset, such as under sampling and oversampling. In this study, due to the large volume of existent data, it was decided to follow the strategy of under sampling.



(a) Histogram

(b) Box-Plot

Figure 5.6: Distribution of the average basket of the "Premium" cluster

Table 5.5: Percentage of churners per cluster

Clusters	**Nr. of Churn Instances**	**Total Nr. of Instances**	**Percentage of Churn Instances**
1. Premium	32 844	1 271 932	2.6%
2. Every Day Low Price	275 211	2 074 576	13.3%
3. Kids in the house	112 139	1 050 748	10.7%
4. No Value	260 934	673 794	38.7%
5. Promo Busters	35 704	320 492	11.1%

5.2.3 Churn Prediction Models

5.2.3.1 Regularized logistic regression

To determine the best parameters for Elastic Net method, i.e. α and λ, a 10-fold cross validation was combined with grid-search for parameter tuning, where multiple parameter combinations were evaluated to determine the one that led to the best results. Results of parameter tuning are presented in Table 5.6.

Table 5.6: Elastic Net parameters chosen after parameter tuning

Clusters	α	λ
1. Premium	0.8	0.00131274
2. Every Day Low Price	0.9	0.00050267
3. Kids in the house	0.1	0.00620031
4. No Value	0.5	0.00084056
5. Promo Busters	1	0.00248654

For the cluster "Promo Busters" the method that has led to the best results is Lasso ($\alpha = 1$) while for "Kids in the house" the best results were obtained when the method gets closer to Ridge.

Once applied the model in unseen data, the results are the ones presented in Table 5.7, and the ROC Curves are the ones presented on Appendix E. The model with highest values for all metrics is the one built for “Premium” customers, while the one for “No Value” customers has the lowest performance. This could be due to a more random behavior and due to the fact the "No Value" customers buy with less frequency.

Table 5.7: Results for Regularized logistic regression

Clusters	AUC (%)	Accuracy (%)	Lift
1. Premium	85.75	77.88	1.59
2. Every Day Low Price	82.49	75.40	1.46
3. Kids in the house	80.59	73.45	1.42
4. No Value	72.65	66.05	1.24
5. Promo Busters	80.01	72.98	1.43

The predictive model built for the “Premium” cluster has a lift of 1.59, which means that the model is better at predicting churner than randomly generated predictions. This conclusion applies to all clusters, as they all have a lift superior to 1.

Looking at the importance of variables for each cluster (see Figure 5.7), which is determined by the magnitude of standardize coefficients, that, combined with the analysis of the signal of coefficients, makes possible to identify the variables with higher influence on the model and if they impact negatively or positively the probability of churning.

In all clusters the change of value segment to a lower one is a predictive factor of customer churn. Additionally, customers who adhere to the app Cartão Continente have a lower likelihood of churning. However, stating that the app is the cause to reduce the probability of churn is fallacious, since customers who adhere to the application are more involved with Continente and therefore less susceptible of churning. This leads us to conclude that using the application does not negatively impact the customer experience to the point of increasing the likelihood of churning.

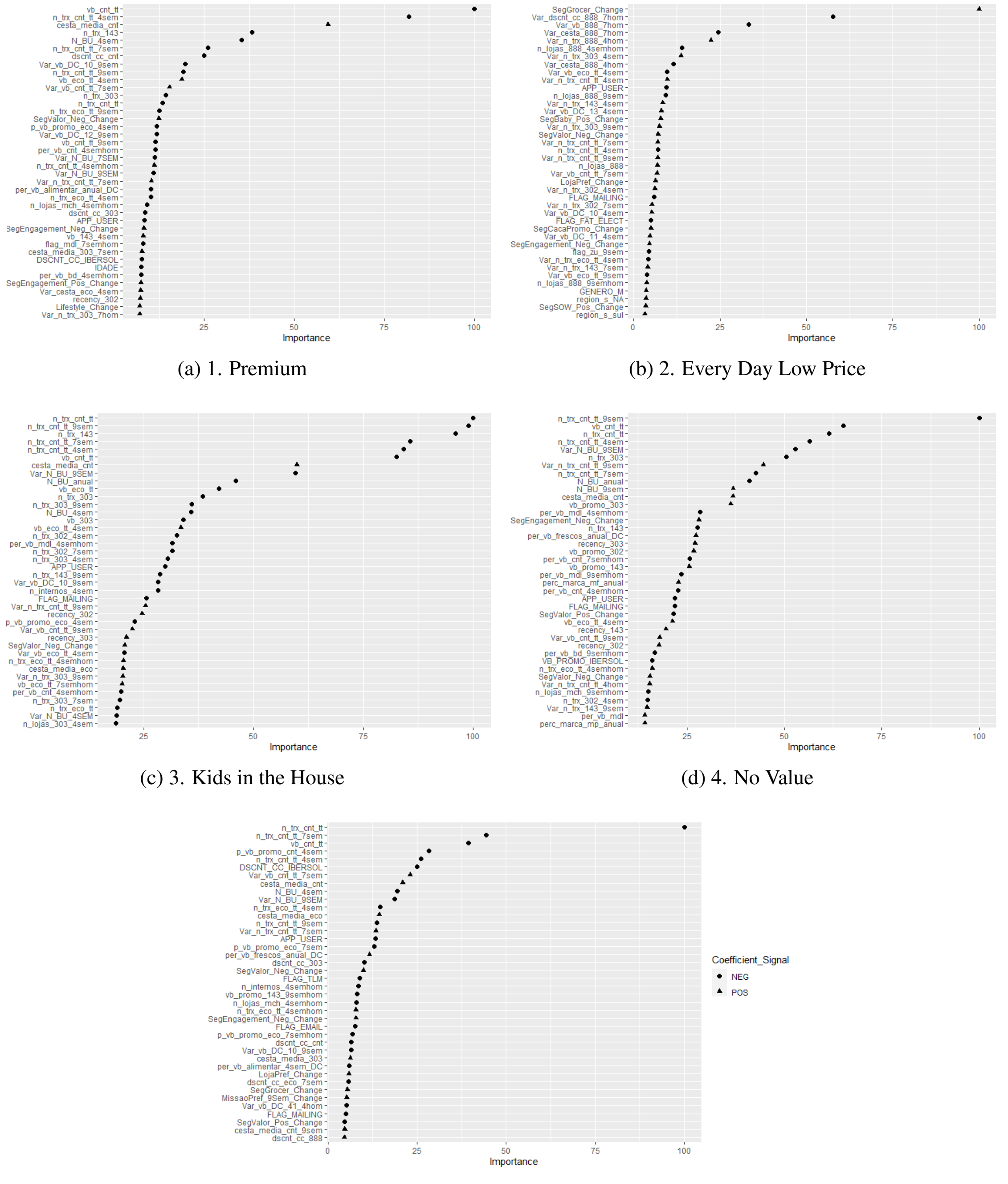
For the vast majority of customer groups, customers with the highest sales volume in the past year and with the highest frequency of transactions in Continente insignias are less likely to abandon. This is in line with what is presented in previous studies, namely Axelsson and Notstam (2017), where frequency and value spent are emphasized as important factors for reducing the likelihood of churn.

The most actionable conclusions that result of the analysis of the variables with higher influence on the prediction models, for each cluster, are presented below.

Premium:

Premium customers are more likely to abandon Continente [3] when they change their lifestyle segment, which happens when the customer changes their consumption habits. Furthermore, a decrease or increase in the engagement segment also means higher likelihood of churning. To sum up, changes in customer behavior of “Premium” customers have a positive impact on their

[3] Continente, Continente Modelo, Continente Bom Dia and Continente Online



(a) 1. Premium

(b) 2. Every Day Low Price

(c) 3. Kids in the House

(d) 4. No Value

(e) 5. Promo Busters

Figure 5.7: Importance of predictive variables per customer group: Regularized logistic regression

probability of churn. This is in accordance with DeSouza (1992) who suggested that changes in customer behavior are indicators of higher promptness to withdraw the relationship with the company.

Clients with higher average baskets are more likely to churn, by crossing this information with the fact that in this cluster the mission that most stands out is “Abastecimento”, one can conclude that some of these customers with higher baskets value do not return to Continente in the following month to do more regular purchases. Thus, an opportunity to increase retention is to promote to these customers purchases of the mission “Suprir Faltas” and/or “Suprir Faltas de Frescos” which, although typically of lower value, occur more frequently. To increase the number of purchases of these missions, more regular communications can be made to promote attractive offers on the products that typically characterize them.

When the customer buys products from more diverse business units, the likelihood of churning is lower. So, to increase the retention of “Premium” customers, it might be interesting, as these customers are already involved with Continente, to invest in a cross-selling strategy in order to encourage the customers to increase the number of business units bought.

Higher volume of sales within the ecosystem [4] in the last 4 weeks increases the probability of churning. However, when customers have a high percentage of sales with promotion in the ecosystem, they are more likely to return. This lead us to infer that this customers return to Continente to use the amount accumulated in the card with sales in Continente insignias and in internal partners[5].

Every Day Low Price:

Customers whose preferential store has changed are more likely to churn. To keep a relationship with them and consequently, increase customer retention, attractive offers should be sent to these customers.

Additionally, “ELP” customers, whose region/address is unknown are more prone to churn. This may result from the fact that, as the address is unknown, these customers do not receive the coupons sent by mail, which are the most attractive and are, therefore, more likely to churn. Or it could be case that these customers have not given all the data on purpose because they are less willing to establish a relationship with Continente, which makes them more likely to churn. Given that, with the information available it is not possible to confirm which hypothesis is true it would be interesting in the future to invest efforts in a more detailed study to assess the underlying reasons. Assuming that the former is the main cause of churn it will be interesting to encourage these customers to use the app Cartão Continente and send retention coupons through this channel.

Active[6] customers of Continente Online without online purchases in the last 9 weeks are more likely to churn. However, customers who had already visited Continente during the homologous periods considered, are less likely to churn, which leads us to conclude that customers who have already bought online and keep their purchasing using the online channel should be encouraged to increase their basket and the amount spent online as this would reduce their likelihood of churning. These incentives could take the form of special promotions like discounts coupons, for online shops, whose value is a percentage of their purchases.

[4] Continente, Continente Modelo, Continente Bom Dia, Continente Online,Meu Super, Wells, Zippy, MO, Bagga, Note and Zu

[5] Meu Super, Wells, Zippy, MO, Bagga, Note and Zu

[6] Customers with at least one transaction in the previous year.

Kids in the House:

When customers from this cluster buy products from more diverse business units or if this number has increased in more recent periods compared to last year, their likelihood of churning decreases. Higher average spending per transaction in Continente insignias and in the ecosystem increase the probability of the customer being classified as a churner. Market studies, previously carried out by Sonae MC, concluded that some families with children prefer to purchase large quantities and less frequently, due to the difficulty in reconciling the family routine with the process of shopping. It is argued by the families that the process of shopping with children is stressful and leads them, as much as possible, to reduce the number of visits to the stores. Due to the previous stated, a possible solution to increase customer retention would be to encourage them to use more practical solutions for their shopping, that Sonae MC has available, such as Continente Online, Click & Go or to more conservative ones offer coupons for products more related to purchase missions that happen more regularly, such as“Suprir Faltas” or “Suprir Frescos”.

Clients with higher recency (time since last visit) in Continente Bom Dia and Continente Modelo are more likely to churn, so when families with kids take more time without going to proximity stores, they are more likely to leave. Previous studies carried out on this theme also concluded that customers with a higher value of recency are less likely to repeat purchases (Hadden et al., 2007).

Additionally, clients from this cluster, who visited more internal partners in the last month are less likely to churn. This could be due to the fact that these customers return to Continente stores to spend the amount accumulated in Cartão Continente with purchases in these partners. Moreover, as these customers visit more partners within the ecosystem, they may be more involved with the brand. Thus, one possible strategy is to give these customers incentives to visit internal partners, which in the end will impact negatively the probability of churning at Continente insignias.

No Value:

When there is a change of value segment to a higher level, the probability of churning is lower. To induce this changes in value segment, higher average baskets and higher number of transactions should be promoted. This can be done using coupons and discounts to attract these customers that typically have low share of wallet.

Clients “No Value” that in the previous year spent a high percentage of their sales buying fresh food (vegetables, fruits, meat, fish and delicatessen) have a higher likelihood of churn in the target period, which lead us to infer that fresh food in Continente do not represent a force of attractiveness for these customers.

“No Value” customers with a high volume of sales on promotion in the insignias Continente, Continente Modelo and Continente Bom Dia are more likely to abandon. Typically, these customers are clients who regularly buy in the competitors and come to Continente on a mission “Orientada”, which means that they are focused on taking advantage of promotions on non-food products, but then they do not return on the following weeks.

Customers who purchase products from a larger number of business units are less likely to leave Continente. As these customers typically shop at competitors, there is not much knowledge of their preferences. Therefore one way to leverage the relationship with them can be by offering transaction discounts that will not only stimulate customers to buy from Continente, but will also incentive them to do purchases of more products and consequently increase their likelihood of buying products belonging to more business units.

Promo Busters:

In "Promo Busters" their churn behavior is highly driven by their promotional activity, the higher the percentage of sales on promotion in the previous 4 weeks, the less likely the customer is to leave Continente, this behavior may be due to the fact that customers return to Continente to spend the accumulated amount on the Cartão Continente, from previous campaigns. The same happens when the amount accumulated in Cartão Continente with purchases at Ibersol [7] is higher.

"Promo Busters" with higher value accumulated in card throughout the previous year also have a higher probability of returning in the following 7 weeks. The rational behind this behavior is the same as the presented above.

In this cluster when customers change value segment to a higher one their probability of churning decreases. Similarly to "ELP" customers, when the customer's preferred store changes, the probability of churning increases and for this reason, attractive offers should be sent to those customers. Visiting more partners also decreases the likelihood of churn which may indicate greater involvement with the Sonae MC brand and with its loyalty program.

"Promo Busters" whose percentage of previous year's fresh products sales is high are also more likely to churn. This demonstrates that fresh food do not represent a force of attractiveness in Continente for "Promo Busters".

5.2.3.2 Decision trees

For this analysis, due to the computational time required, a sample of 20 000 instances randomly chosen and stratified by target variable, for each cluster was analyzed. In this method the same variables were included but as the Decision trees are able to cope with data having different range of values no normalization was performed.

Using 10-fold validation with AUC as metric for assessing the predictive quality of the model, the best results for cluster "No Value" and "Promo Busters" were obtained by setting the splitting criterion as entropy and the max-depth of the tree to 4. For the remaining clusters, the best performance was achieved with the same splitting criterion and max-depth of the tree equals 5.

Thereafter, the model was tested on unseen data for all clusters and the results are the ones presented on Table 5.8 and the ROC Curves on Appendix F .

Comparisons between Regularized logistic regression and Decision trees should not be made without considering that the sample was reduced before applying the Decision trees. In this case,

[7] Includes brand as: Burger King, SOL, KFC, Pizza Hut, Ò Kilo, Roulotte, Pans & Company, Miit and Pasta Caffé.

Table 5.8: Results for Decision tree

Clusters	AUC (%)	Accuracy (%)	Lift
1. Premium	84.56	76.73	1.58
2. Every Day Low Price	82.19	75.27	1.44
3. Kids in the house	79.23	71.78	1.25
4. No Value	71.92	65.27	1.22
5. Promo Busters	79.07	73.00	1.46

Decision trees do not provide information as actionable as Regularized logistic regressions and the information gathered is less relevant for retention campaigns. This could result from the fact that when pruning using the max-depth of the trees, the solution chosen is the one that maximizes AUC but may compromise the identification of actionable knowledge.

Some of the major conclusions that can be extracted from the analysis of the Decision trees (see Appendix G) are presented below.

Premium:

Customers with less than weekly frequency are classified as churners, except those who have increased the number of business units purchased in recent weeks, have more than 31 transactions in the past year and have accumulated more than 10€ in Cartão Continente in the previous month. On the other hand, keeping everything the same, if the number of transactions in the last year is less than 32, the customer is more likely to return if he/she have purchased more perishable products (such as fruits and vegetables, delicatessen and take away) in recent weeks. This shows that the amount accumulated in Cartão Continente and the increase in sales on fresh products are considered important factors when “Premium” customers are less frequent.

Most "Premium" customers, whose weekly frequency is close to 1 or higher, will be classified as non-churners. When “Premium” customers are more frequent, the number of transactions and amount spent in the ecosystem start to have more relevance when predicting churn, since customers more involved in the ecosystem are less prone to churn.

Every Day Low Price:

Most customers with less than 22 transactions per year are classified as defected. Exceptions to this rule are new customers (longevity less than one year) who have made several transactions in the last few weeks but less than one per month in the past year or customers whose amount spent in own brand purchases increased in the last month, that have more than 3 transactions in the last 9 weeks and have obtained more than 15€ in discounts on Cartão Continente with expenses made in the ecosystem during the same period. This means that despite having a small number of transactions in the past year, these customers, in the last weeks have shown more frequent activity both in Continente insignias, as well as, in ecosystem.

When “ELP” customers have made more than 21 transactions in the previous year the classification as churner or not will depend a lot on their behavior in the ecosystem. It is expected that customers with more transactions in Continente are also more involved with the ecosystem.

Kids in the House:

Customers from “Kids in the House” cluster with high frequency during the year (more than 25 transactions) will only be considered churners if in more recent periods they have made only purchases in Continente occasionally, have accumulated small amounts of discounts in the ecosystem in the past month and if their ratio between the number of transactions in the last 7 weeks when compared to the same period of the previous year is less than 0.3. Apparently, these are customers who are substantially reducing their purchasing frequency.

Customers with a small number of transactions in the previous year will be considered churners. Exception to this rule are new customers who have made at least 8 transactions in the last year and in the last 9 weeks have made more than 40% of the total number of transactions performed in the last year or customers who had on average more than one transaction per month but less than 2, who have made on average more than 1 transactions every two weeks in the last 9 weeks, who are increasing the number of business units bough (more than 50% of the total bought in the year was bought in the last month) and also customers who have not increased the number of business units purchased but have accumulated discounts on Cartão Continente, with expenses incurred in the ecosystem in the last 4 weeks. To sum up, customers with few transactions during the year will be more likely to return in the following 7 weeks if they have increased their purchase frequency and the number of products purchased or if they have accumulated some amount of money in Cartão Continente in more recent periods.

No Value:

As expected, due to the sporadic frequency with which “No Value” customers visit Continente, this is the decision tree whose separation criterion related to the number of annual transactions is lower, i.e. only 11 transactions.

Customers whose frequency is less than one per month will be considered churners unless in the past year they have made at least 6 transactions, have bought in the last month more than 10% of the number of business units bought in the past year and have recently increased their purchasing frequency. The same applies to customers who meet the same criteria but, have not bought in the last month more than 10% of the number of business units bought in the past year and have a very high percentage of purchases on promotion in the homologous period. One hypothesis for this is that these are customers who despite coming only sporadically to Continente use transaction coupons in almost all transactions and if this behavior of coming to Continente to use coupons is prolonged in time, the customer is expected to return at least once in the next 9 weeks.

Promo Busters:

Customers who have an intermediate number of transactions in the last year (between 15 and 21) and have accumulated discounts in the ecosystem in the last 4 weeks are classified as non-churners. This demonstrates that “Promo Busters” more involved with the ecosystem are less likely to churn than those who are not. Customers whose number of transactions in Continente in the last year is higher than 21 will be classified as non-churners, with the exception of customers who have a volume of purchases in the last month of less than about 14€ and who have considerably reduced the number of transactions in the ecosystem compared to the homologous period. In resume, frequent customers will be considered churners if they have spent less in more recent periods and if they have reduced the purchasing frequency compared to previous periods.

5.2.4 Cost-Sensitive Classification

5.2.4.1 Bayes minimum risk

To apply the Bayes minimum risk it is necessary to compute the risk of classifying a customer as churn or non-churner and to do so we need to compute the costs associated with every possible combination of the predicted class and the actual class for each customer. This costs were calculated using the expressions presented in Table 2.3, with some adjustments to adapt to the reality of food retail and more specifically the reality of Sonae MC. For this analysis, only costs associated with retention campaigns for customers identified as churners were considered.

At first it is necessary to be aware of the current functioning of the retention campaigns in Sonae MC, to understand how it can be reflected in this analysis. Typically, a customer identified as churner receives an extra offer of two coupons per month and each coupon for loyal customers has a value of 15% on the amount spent on the transaction, while the remaining ones receive a discount coupon of 5€ on purchases with a minimum value of 20€.

Besides that, as each customer receives two coupons each month, there is the possibility of rebating none, one or the two of them. It was assumed that the probability of a given customer rebating a coupon is equal to the proportion of rebates, which that customer carried out in the year preceding the analysis for each one of the hypothesis previously presented. For customers who did not receive any coupons at the moment of the analysis, it was assumed that the probability of these coupons being rebated is equal to the average rebate rate of the cluster to which they belong.

Due to the aforementioned, the formula to compute the cost of a true positive and the cost of a false positive is presented in Equations 5.1 and 5.2, respectively.

$$C_{TP} = C_a + P_0 \times SV_{TarPer} + \textstyle\sum_{n=1}^{N} P_n \times \frac{n \times C_o \times 12 \times TarPer}{52} \quad ,n = 1,2 \tag{5.1}$$

where SV_{TP} is the sales value during the target period, C_a is the cost of sending the offer, P_n is the probability of rebating n coupons, C_o is the cost of the offer, which is 15% of their average basket for loyal customers and 5€ for the remaining ones and the $TarPer$ is the target period, in weeks.

$$C_{FP} = C_a + \textstyle\sum_{n=1}^{N} P_n \times \frac{n \times C_o \times 12 \times TarPer}{52} \quad ,n = 1,2 \tag{5.2}$$

where C_a is the cost of sending the offer, P_n is the probability of rebating n coupons, C_o is the cost of the offer which is 15% of their average basket for loyal customers and 5€ for the remaining ones and $TarPer$ is the target period, in weeks.

The dispatch of the offer (C_a) has a cost of 6 cents per offer. Once all the costs for the different customers have been calculated and knowing the probability of them being churners (computed through Regularized logistic regression) it is possible to compute the risk associated with classifying a customer as target for a retention campaign using the method Bayes minimum risk. In the end each client will be assigned to the class that presents lowest risk for the company.

In Table 5.9 we compared the costs of retention campaigns when using the cost sensitive method, i.e. Bayes Minimum risk, with the costs of retention campaigns when sending coupons to customers predicted as churners whose class was defined using regularized logistic regression without considering the associated costs. This table presents the total savings, the savings per customers and the accuracy of the model. It should be noted that the values shown are related to the target period of the cluster to which customers belong to. Furthermore, the risk of losing the customer considers costs that, if they occur, do not imply effective outflows of money, such as the cost of losing the customer's sales. Thus, in this case the savings do not represent real savings of money but also consider the prevention of potential losses.

Table 5.9: Comparison savings between using a non cost-sensitive method and a cost-sensitive method

Clusters	Savings (%)	Savings per customer (€)	Accuracy (%)
1. Premium	25.00	13.20	52.94
2. Every Day Low Price	10.75	3.09	46.34
3. Kids in the house	11.60	3.51	46.12
4. No Value	5.16	0.71	46.00
5. Promo Busters	14.60	3.83	46.24

The implementation of the cost-sensitive model generates savings in the retention campaigns of more than 5% for all clusters, however the cluster with the highest saving is the "Premium". The "Premium" customers are typically the ones who spend more money in Continente, being therefore the most valuable and with the greatest impact if lost. In this sense, the risk of losing them is the highest one, so the model will be more prone to predict them as churners and to send them retention offers even if their probability of churning is not very high. As the cost of the offer is typically much smaller than the cost of losing these customers, it is expected that the model will result in higher savings for customers with higher value.

Finally, in line with what was concluded in the study Bahnsen et al. (2015b), cost sensitive-models do not lead to higher predictive performance. In this case the accuracy decreases to values around 50% for each cluster, which reveals a poor predictive performance.

5.3 Managerial Implications

The results of this study bring many business insights that can be used by the retailer to improve their retention campaigns and consequently increase customer retention rate.

The separation of customers into different groups has made possible, not only to identify which factors are the most important to predict churn in each one, but also to define different actions depending on the cluster. For example, although both “Premium” and “No Value” customers with more business units purchased are less likely to churn, the retention strategies adopted for each of them should be different. In the former the customers are much more involved with Continente and therefore a cross-selling strategy to increase the number of business units bought makes more sense. On the other hand, for “No Value” customers, about whom there is not much information regarding their preferences and behavior, because they are typically customers of other competitors, a transaction discount coupon may be more effective.

Furthermore, changes in customer behavior in different customer groups have a completely different impact on the probability of churning. For example, a change of the customers belonging to the “Premium” segment to a higher engagement segment is not desirable. It is an indicator that a loyal customer is changing his behavior, which is reflected in an increase of his likelihood to churn. A change to a higher value segment observed for a customer belonging to the "No Value" segment is a positive sign. This is an indicator of bigger baskets, higher frequency and less probability of churn.

Another example of an insight that may have an impact on the management of the relationship with customers is the fact that options should be explored so that more convenient solutions can be offered and/or promoted to customers “Kids in the house”, such as Continente Online, Click & Go or to offer attractive options in more proximity stores for them to visit more regularly.

Fresh products have proven not to be a source of attractiveness to the company for some clients. This insight can have implications in several areas of the company and an effort should be made to understand what can be done to increase their attractiveness.

Finally, it turns out that the cost-sensitive model provides savings up to 25.0% compared to a non-cost-sensitive model. According to this model, the most valuable customers, who are typically the most loyal, are also the ones most worth retaining and a considerable part of them are classified as target for retention campaigns.

However, it is not correct to say that this model should be preferred over the other as it has a much lower predictive performance. In these cases it is up to the company, depending on its objectives, to decide which metric it wants to be guided by.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

As discussed in the literature review, companies are becoming more focused in developing customized and effective retention campaign. In this dissertation, we proposed a two-stage approach for predicting churn, in a food retail company, focused both on predictive performance and understanding the main drivers for it.

In the first phase of this project, 5 customer groups were created, each cluster grouping similar customers according to their behavior, from distinct perspectives, in the retailer in question. The methodology adopted can be replicated to group similar customers according to categorical variables independently of the size of the dataset chosen.

Secondly, differentiated churn prediction models were developed for each customer group previously identified. In this study, we have proposed to use both Regularized logistic regression and Decision trees to model churn prediction and it was concluded that there are no major differences in their predictive performance. Although the first one proved to be more enlightening about what retention actions could be undertaken. This algorithm has led to models with area under the curve ranging from 72.65% to 85.75%, depending on the cluster.

One can say that the segregation of clients into logical groups at an earlier stage of this study and the subsequent development of differentiated model per each of them proved to be a good solution to extract information about the main churn drivers for each cluster. These drivers were found to be different, thus allowing the definition of targeted retention strategies for each one. Even in cases when the main churn drivers were the same, depending on the cluster and the characteristics of its customers, different retention actions were defined.

Finally, we demonstrated the importance of using an example-dependent-cost-sensitive method, i.e. Bayes minimum risk, that incorporates real financial costs to maximize the results of retention campaigns. In this study the cost-sensitive method provided cost saving of retention campaigns up to 25%.

6.2 Limitations and Future Work

Other studies, such as Ferreira (2019), have shown the importance of complaint analysis as an opportunity to correct and solve consumer problems in order to maintain satisfaction and therefore retention rates. This study did not include information on customer complaints due to lack of data related to it. It would be interesting in the future to consider including other variables to examine the effects of service quality and satisfaction on customer retention rate.

Another limitation of this study has to do with the incorporation of costs from retention campaigns. Firstly, this approach only considers the costs associated with retention campaigns for customers identified as churners. Secondly, to compute the expected losses if the customer actually abandons the company we considered the average expenses of the customer during that period and to compute the probability of rebating coupons we assumed as proxy the percentage of coupons rebated by the customer in the previous year. To tackle these limitations, a model to predict the expected amount spent by the customer during that period could be developed, as well as, a model to predict the likelihood of rebating coupons. As the model to predict the probability of rebating could be dependent on the value of the offer sent to the customer, it would be interesting to vary the cost of retention offers and to investigate the impact on the company results.

In order to fully assess the benefits for the company of adopting the methodology proposed by this study, it would have been interesting to meet with the parties involved, to define new customer retention strategies resorting to insights gathered using explainable models, quantify these strategies and finally, include them in the cost-sensitive method instead of assuming similar costs to the existing ones.

It would also be interesting to consider applying the present methodology to predict partial churn, which can also be an important source of knowledge and insights for the company, especially if the goal is to detect changes in customer spending.

In this study we explored how explainable data mining models can be successfully applied to predict churn but in a future study one could extend the present one by comparing the performance of black-box models with rule extraction, such as ALBA, with the models used.

Finally, as we used an after training algorithm, Bayes minimum risk, in future studies it would be interesting to consider the implementation of a cost-sensitive algorithm during the training phase, such as cost-sensitive Decision trees, to assess whether there are any improvement in savings.

Bibliography

Accenture (2015). Retalho: Onde está e quem é o consumidor?

Aggarwal, C. C. (2014). *Data classification: algorithms and applications*, pp. 620–621. CRC press.

Ahn, J.-H., S.-P. Han, and Y.-S. Lee (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry. *Telecommunications policy 30*(10-11), 552–568.

Algamal, Z. Y. and M. H. Lee (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in biology and medicine 67*, 136–139.

Amorim, J. F. C. d. (2013). *Continente online: starting a one to one marketing program.* Ph. D. thesis.

Asaari, M. H. A. H. and N. Karia (2000). Churn management towards customer satisfaction: A case of cellular operators in malaysia. In *Conference Proceeding, The International Conference on E-Commerce: Emerging Trends in Electronic Commerce (ETEC2000)*, Volume 21, pp. 5.

Axelsson, R. and A. Notstam (2017). Identify churn: A study in how transaction data can be used to identify churn for merchants.

Bahnsen, A. C., D. Aouada, and B. Ottersten (2014). Example-dependent cost-sensitive logistic regression for credit scoring. In *2014 13th International conference on machine learning and applications*, pp. 263–269. IEEE.

Bahnsen, A. C., D. Aouada, and B. Ottersten (2015a). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications 42*(19), 6609–6619.

Bahnsen, A. C., D. Aouada, and B. Ottersten (2015b). A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics 2*(1), 5.

Bernardino, L. A. d. S. (2012). Previsão de churn no retalho alimentar. pp. 91–136.

Boehmke, B. and B. Greenwell (2019). *Hands-On Machine Learning with R*, pp. 121–139. Chapman and Hall/CRC.

Buckinx, W. and D. Van den Poel (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European journal of operational research 164*(1), 252–255.

Cebr (2012). Data equity—unlocking the value of big data. *A report for SAS*, 1–9.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc 9*, 13.

Chiang, D.-A., Y.-F. Wang, S.-L. Lee, and C.-J. Lin (2003). Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications 25*(3), 293–302.

DeSouza, G. (1992). Designing a customer retention plan. *Journal of Business Strategy 13*(2), 24–28.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). From data mining to knowledge discovery in databases. *AI magazine 17*(3), 37–37.

Ferreira, M. I. (2019). O paradigma da retenção de consumidores: Otimização da estratégia anti-churn num retalhista português.

Gaurav Pant, Sahir Anand, A. K. N. A. (2014). State of the industry research series : 3rd annual analytics in retail study. pp. 12–14.

Glady, N., B. Baesens, and C. Croux (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research 197*(1), 402–403.

Goldschmidt, R. and E. Passos (2005). *Data mining: um guia prático*, pp. 2592–2602. Gulf Professional Publishing.

Hadden, J., A. Tiwari, R. Roy, and D. Ruta (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research 34*(10), 2902–2917.

Han, J., M. Kamber, and J. Pei (2011a). Data mining concepts and techniques third edition. *Morgan Kaufmann*, 8.

Han, J., M. Kamber, and J. Pei (2011b). Data mining concepts and techniques third edition. *Morgan Kaufmann*, 443–471.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*(1), 29–36.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.*

Hodeghatta, U. R. and U. Nayak (2016). *Business Analytics Using R-A Practical Approach*, pp. 117. Springer.

Intel (2016). Retail data-driven decision-making. pp. 1–2.

Karypis, M. S. G., V. Kumar, and M. Steinbach (2000). A comparison of document clustering techniques. In *TextMining Workshop at KDD2000 (May 2000)*, pp. 1–20.

KhakAbi, S., M. R. Gholamian, and M. Namvar (2010). Data mining applications in customer churn management. In *2010 International Conference on Intelligent Systems, Modelling and Simulation*, pp. 220–225. IEEE.

Khodabandehlou, S. and M. Z. Rahman (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, 65–89.

Kracklauer, A. H., D. Q. Mills, and D. Seifert (2004). Customer management as the origin of collaborative customer relationship management. In *Collaborative Customer Relationship Management*, pp. 3–6. Springer.

Larson, R. B. (1993). Food consumption, regionality, and sales promotion evaluation. *Unpublished Ph. D. thesis, Department of Agricultural Economics, Purdue University*.

Lazarov, V. and M. Capota (2007). Churn prediction. *Bus. Anal. Course. TUM Comput. Sci 33*, 34.

Lejeune, M. A. (2001). Measuring the impact of data mining on churn management. *Internet Research 11*(5), 375–387.

Ling, C. X., J. Huang, H. Zhang, et al. (2003). Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, Volume 3, pp. 519–524.

Ling, R. and D. C. Yen (2001). Customer relationship management: An analysis framework and implementation strategies. *Journal of computer information systems 41*(3), 82–97.

Löster, T. (2017). Comparison of results of selected clustering methods on real data set. In *11th International Days of Statistics and Economics: Conference Proceedings*, pp. 889.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers (2011). Big data: The next frontier for innovation, competition, and productivity. pp. 2.

Mattison, R. (2006). *The telco churn management handbook*, pp. 79–82. Lulu.

Miguéis, V. L., D. Van den Poel, A. S. Camanho, and J. F. e Cunha (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert systems with applications 39*(12), 11250–11256.

Molnar, C. (2019). *Interpretable machine learning*, pp. 1–24. Lulu.

Ngai, E. W., L. Xiu, and D. C. Chau (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications 36*(2), 2592–2602.

Pfeifer, P. E. and H. Bang (2005). Non-parametric estimation of mean customer lifetime value. *Journal of Interactive Marketing 19*(4), 48.

Raileanu, L. E. and K. Stoffel (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence 41*(1), 77.

Refaat, M. (2010). *Data preparation for data mining using SAS*, pp. 2. Elsevier.

Rygielski, C., J.-C. Wang, and D. C. Yen (2002). Data mining techniques for customer relationship management. *Technology in society 24*(4), 483–502.

Shaon, S. K. I. and H. Rahman (2015). A theoretical review of crm effects on customer satisfaction and loyalty. *Central European Business Review 4*(1), 23.

Shetty, P. P., C. Varsha, V. D. Vadone, S. Sarode, and D. Pradeep Kumar (2019). Customers churn prediction with rfm model and building a recommendation system using semi-supervised learning in retail sector. *International Journal of Recent Technology and Engineering 8*(1), 3353–3358.

Tamaddoni Jahromi, A. (2009). *Predicting customer churn in telecommunications service providers*. Ph. D. thesis.

Tan, P.-N., M. Steinbach, and V. Kumar (2016). *Introduction to data mining*. Pearson Education India.

The Consumer Goods Forum (2018). Portugal: A Retail Snapshot. `https://www.theconsumergoodsforum.com/portugal-retail-snapshot/`. Accessed March 10, 2010.

Valtola, V. (2019). A study of customer retention. pp. 20–25.

Van den Poel, D. and B. Lariviere (2004). Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research 157*(1), 196–217.

Veloso, F. J. M. (2013). *Um modelo para previsão de churn na área do retalho*. Ph. D. thesis.

Verbeke, W., D. Martens, C. Mues, and B. Baesens (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications 38*(3), 2354–2364.

Vuk, M. and T. Curk (2006). Roc curve, lift chart and calibration plot. *Metodoloski zvezki 3*(1), 89.

Wei, C.-P., Y.-H. Lee, and C.-M. Hsu (2003). Empirical comparison of fast partitioning-based clustering algorithms for large data sets. *Expert Systems with applications 24*(4), 351–363.

Wilkinson, L., L. Engelman, J. Corter, and M. Coward (2012). Systat 8.0 statistics. *A report for SAS*, 1–44.

Yan, L., D. J. Miller, M. C. Mozer, and R. Wolniewicz (2001). Improving prediction of customer behavior in nonstationary environments. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, Volume 3, pp. 2260. IEEE.

Yan, L., R. H. Wolniewicz, and R. Dodier (2004). Predicting customer behavior in telecommunications. *IEEE Intelligent Systems 19*(2), 50–54.

Zdravevski, E., P. Lameski, C. Apanowicz, and D. Ślzak (2020). From big data to business analytics: The case study of churn prediction. *Applied Soft Computing*, 1–3.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

Appendix A

Customer ADN Variables previously constructed by Sonae MC

Over the years Sonae MC has carried out several projects to profile its customers, and for this reason there are already several segmentations models of its customers.

The client profile built through the existing segmentations has been used in several models and process across the company. The models previously developed by Sonae MC that were used to support this dissertation are briefly described below.

LifeStyle Segmentation

LifeStyle segmentation is based on the study of transactions and makes possible to understand customer's motivations and choices: why they choose a store or certain products.

There are seven groups. The “Urbanos Sofisticados” and “Saudáveis Exigentes” are groups more focused on quality, then “Os pais práticos”, “Os Generalistas Disciplinados” and “Os Tradicionais Frequentes” are more focused on family and then “Económicos Focados” and “Promotionais Atentos” drive their options based on price and promotions.

Value Segmentation

This segmentation was intended to determine the value of the customer to the Continente by studying customer behaviour, taking into account customer loyalty and spending. Variables such as frequency of purchase, product diversity, recency and average customer spending were used to define the segments.

This segmentation also has 7 segments, Loyal Small, Loyal Medium, Loyal Large, Frequent Small, Frequent Medium, Occasional Small and Occasional Medium.

Price Sensitivity Segmentation

Price Sensitivity (PS) segmentation classifies customers according to the importance of price change for each one. There are 8 segments: Price Upscale, No PS, Neutral PS, Low PS, Medium PS, High-Medium PS, High PS and Very High PS. In “Price Upscale" and in the “No PS” are included customers who typically choose products with high prices and have low promotional activity. The “Neutral PS” differs from "No PS” because it includes customers who give preference to Continente brands and the “No PS” does not show such preference. At the other extreme are the customers with “High PS” and “Very high PS” who are characterized by high promotional activity.

Share of Wallet Segmentation

The purpose of share of wallet (SOW) segmentation is to determine what portion of the customer’s total expenditure is spent in the Continente in relation to the total customer’s expenditure on the different retailers. Clients are classified as having very low, low, medium or high SOW. For example, a customer with very high SOW is a customer who spends most of his available money to groceries in the Continente.

LifeStage Segmentation

By combining socio-demographic information and analytical inference, a segmentation about life stage of customers was constructed. This segmentation can be shown into a matrix in which one axis is divided by the age of the customers, whether is under or over 65 years old, and the second whether they have dependents or not in their aggregate.

There are five segments (Figure A.1): Active Adults, Senior, Family with kids, Family with young adults and Family Supporters.

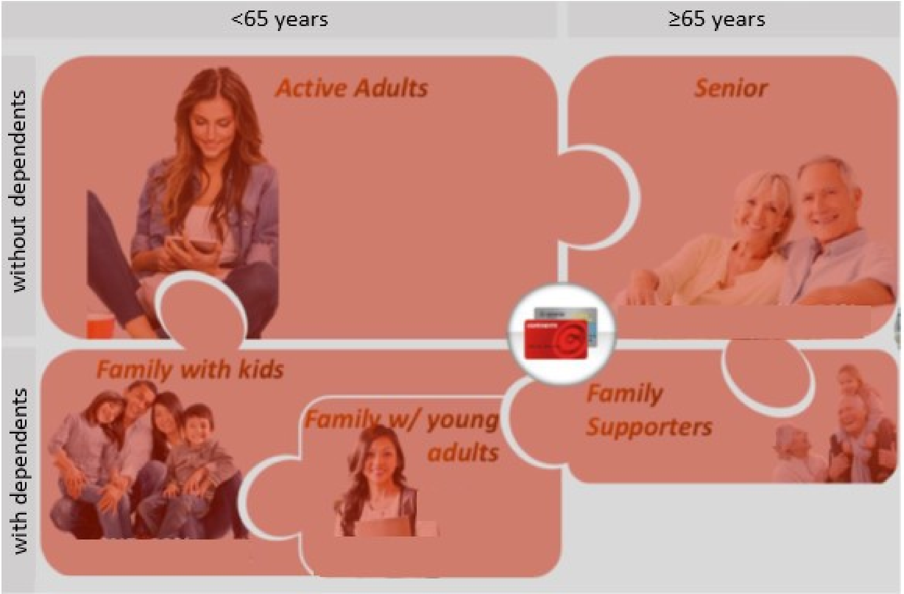


Figure A.1: LifeStage Segmentation. Source: Adapted from Sonae MC

Baby & Junior Segmentation

Through the identification of characteristic products and the analysis of customer transactions, segmentations have been created to identify whether customers are more or less likely to have babies and/or juniors in their household. There are four segments according to the likelihood which are high, medium, low or no value. So, in total, there are 8 segments: high baby, medium baby, low baby, no value for baby, high junior, medium junior, low junior and no value for junior. It can be the case that a single client belongs to low junior and high baby meaning that there is a high likelihood of having a baby and a low likelihood of having a junior in their household.

Grocers Segmentation

Grocer Segmentation allows the detection of customers with a commercial and/or abusive promotional behaviour. Consumers can be classified as promotional high, promotional medium and grocer.

Engagement Segmentation

Through clustering techniques and logistic regression models, customers were classified according to their involvement, i.e. their loyalty to the Continente. Five segments were created: Very Low, Low, Medium, High and Very High.

Payment Day Segmentation

The payment day segmentation identifies the period of the customer's preferred month of purchase. There are four different segments.

Purchase Mission Segmentation

Through this segmentation each transaction can be classified as belonging to one of six possible groups. The construction of the six segments was based on the creation of important core product groups and the identification if they were or not present on the transactions. The possible purchasing missions are "Abastecimento", "Suprir Faltas", "Suprir Falta de Frescos", "Orientada", "Consumo imediato" e "Cafetaria".

Gender Segmentation

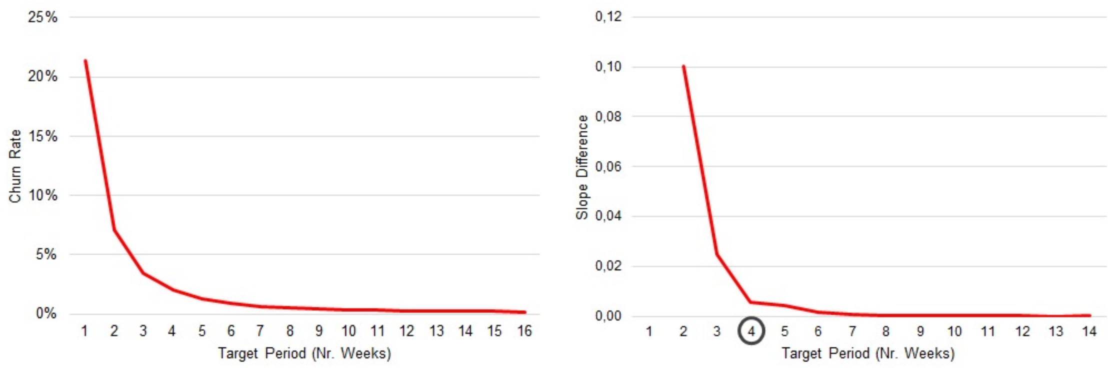
Although customers generally indicate their gender at the time of card creation, over time customers typically do not update their data and their aggregate may change. In this sense and based on the customer's transactions, the gender segmentation is performed, which at the limit can have both genders since a card is considered to be representative of a household and not just of an individual.

Preferential Store Segmentation

There is also a segmentation called preferential store segmentation that identifies the store where the customer makes the greatest amount of purchases, with greater frequency and takes also into account the rebates of discounts that the customer made in different stores. This segmentation allows not only to know the preferential store of the customer, but also the preferential insignia and the type of contact he prefers, whether physical or online commerce.

Appendix B

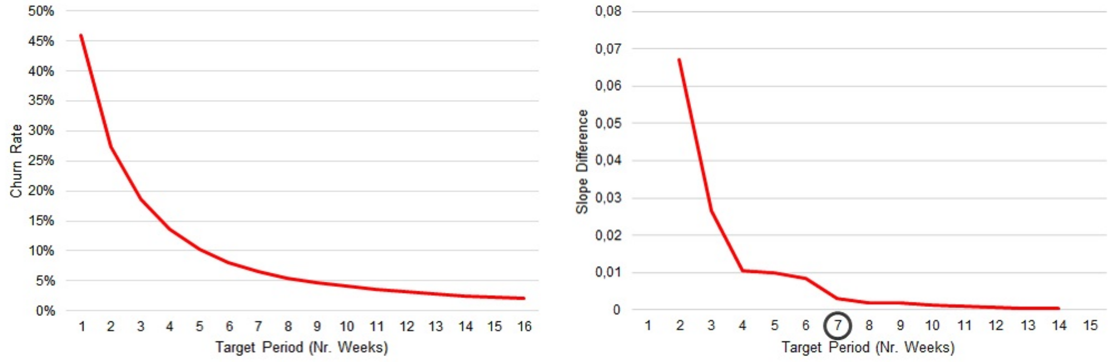
Target Period Definition



(a) Churn Rate Evolution

(b) Evolution of Slope Difference

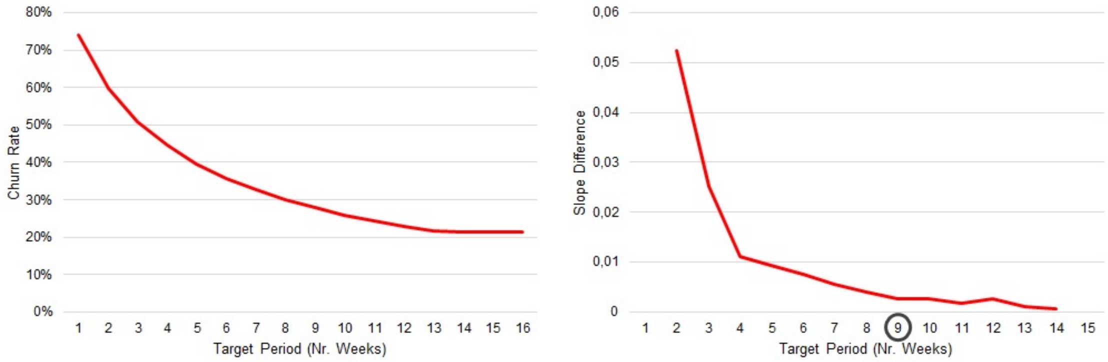
Figure B.1: Finding the target period for cluster "Premium"



(a) Churn Rate Evolution

(b) Evolution of Slope Difference

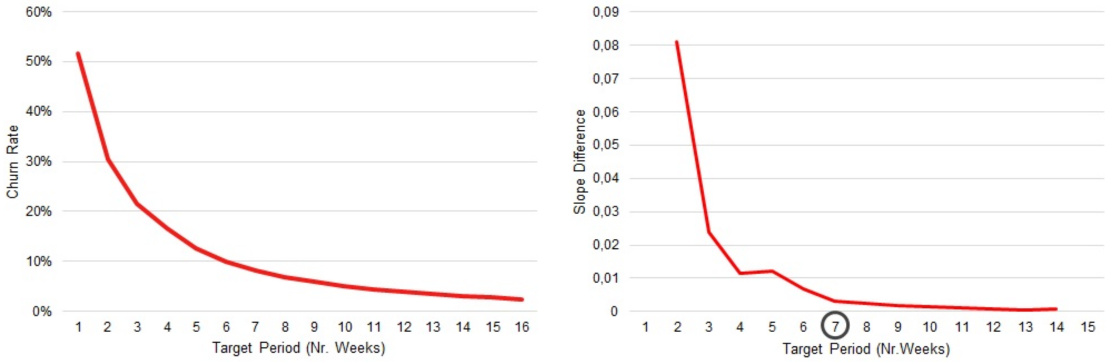
Figure B.2: Finding the target period for cluster "Kids in the house"



(a) Churn Rate Evolution

(b) Evolution of Slope Difference

Figure B.3: Finding the target period for cluster "No Value"



(a) Churn Rate Evolution

(b) Evolution of Slope Difference

Figure B.4: Finding the target period for cluster "Promo Busters"

Appendix C

Variables included in churn models

The following tables present, in more detail, the variables that were used to build the predictive models.

These variables are divided into several categories: Auxiliar (Aux), Socio-Demographic Information (SD), Behavioural Information (B), Transactional Information (T) and Behavioural Change Detection Variables (BC).

Some of these variables were collected for different time periods. These ones are identified with an “X” in the columns “annual”, “sem”or/and “semhom”. Variables with an “X” in column “annual” are variables regarding the year prior to the analysis (12 months before), column “sem” refers to periods equal to the target periods (4, 7 and 9 weeks) immediately before of the analysis, and finally the column “semhom” represents the homologous periods of the target periods in the last year (homologous 4, 7 and 9 weeks).

In this study we included variables related to changes in customer behavior. Thus, the variables with the columns “Var_sem” and “Var_hom” signalized, are variables for which the ratio of the value of recent periods (4, 7 and 9 weeks) with the annual value was computed or between its value in more recent periods with the value during the homologous period, respectively.

Finally, in the column insignia are presented the insignias for which the variable in question was extracted.

Table C.1: Predictive Variables - Part I

Variable	**Description**	**annual**	**sem**	**semhom**	**Var_sem**	**Var_hom**	**Insignia**	**Category**	**Variable Type**
Churn	Target Variable						InsCnt[a]	Aux	Categorical
ID_CLIENTE	Customer Account Number							Aux	-
region_s	Customer Address Region							SD	Categorical
GENERO	Customer Gender							SD	Categorical
FAMILY_MEMBERS	Number of family members							SD	Numeric
APP_USER	If customer adhered to app Cartão Continente							B	Categorical
FLAG_MAILING	If customer is eligible to receive Mailing							B	Categorical
FLAG_TLM	If customer phone number is known							B	Categorical
FLAG_EMAIL	If customer email address is known							B	Categorical
FLAG_FAT_ELECT	If customer adhered to electronic invoice							B	Categorical
IDADE	Customer Age							SD	Numeric
longevidade_cc_year	Longevity of customer account							SD	Numeric
CompIndex	Competition Index computed resorting to Huff formula							SD	Numeric
dist_Cnt	Driving distance from the customer's address postal code to the preferential store's postal code							SD	Numeric
dist_Conc	Driving distance from the customer preferential store to the competitor store with the highest competition index							SD	Numeric
InsigniaConc	Insignia of the store with the highest competition index. Eg. PingoDoce, Auchan, etc.							SD	Categorical
lojaconc_area	Area from the concorrential store with the highest competition index							SD	Numeric
X[b]_Change	If the customer's most segment is different from the most frequent one in the previous year							BC	Categorical
MissaoPref_Change	If the customer's most frequent Preferential Mission during this period is different than the most frequent one in the previous year		X					BC	Categorical
X[c]_Neg_Change	If the customer's most recent segment is smaller than the most frequent one in the previous year							BC	Categorical
X[c]_Pos_Change	If the customer's most recent segment is higher than the most frequent one in the previous year							BC	Categorical

[a] Includes Continente, Continente Modelo, Continente Bom Dia and Continente Online

[b] PrefStore, Lifestyle, Lifestage, SegPSS, SegCacaPromo, SegGrocer, SegPayDay

[c] SegValor, SegEngagement, SegSOW, SegBaby, SegJunior

Table C.2: Predictive Variables - Part II

Variable	Description	annual	sem	semhom	Var_sem	Var_hom	Insignia	Category	Variable Type
vbdev	Amount of sales returned	X					Cnt[a], Mdl[b], Bd[c], Onl[d], Eco[e], InsCnt	T	Numeric
perc_dev	Percentage of sales returned from the total sales amount	X					Eco, InsCnt	T	Numeric
perc_marca_mx	Amount spent on "Marca Exclusiva" over the total amount spent	X	X		X		InsCnt	T	Numeric
perc_marca_pp	Amount spent on "Primeiro Preço" over the total amount spent	X	X		X		InsCnt	T	Numeric
perc_marca_mf	Amount spent on "Marca Fornecedor" over the total amount spent	X	X		X		InsCnt	T	Numeric
perc_marca_mp	Amount spent on "Marca Própria" over the total amount spent	X	X		X		InsCnt	T	Numeric
VB	Amount of sales	X					Ibersol, Galp	T	Numeric
N_TRX	Number of transactions	X					Ibersol, Galp	T	Numeric
VB_PROMO	Amount of sales with a promotion associated	X					Ibersol, Galp	T	Numeric
DSCNT_CC	Accumulated discount on card	X					Ibersol, Galp	T	Numeric
N_LOJAS	Number of Ibersol stores visited	X					Ibersol	T	Numeric
cesta_media	Average basket	X					Ibersol, Galp	T	Numeric
p_vb_promo	Proportion of sales	X					Ibersol, Galp	T	Numeric
vb	Amount of sales	X	X	X	X	X	Cnt, Bd, Mdl, Onl, Eco, InsCnt	T	Numeric
n_trx	Number of transactions	X	X	X	X	X	Cnt, Bd, Mdl, Onl, Eco, InsCnt	T	Numeric
vb_promo	Total volume of sales with promotion	X	X	X	X	X	Cnt, Bd, Mdl, Onl, Eco, InsCnt	T	Numeric
n_lojas	Number of stores visited	X	X	X			Cnt, Bd, Mdl, Onl	T	Numeric
dscnt_cc	Accumulated amount in Cartão Continente	X	X	X	X	X	Cnt, Bd, Mdl, Onl, Eco, InsCnt	T	Numeric
cesta_media	Average basket during the period analyzed	X	X	X	X	X	Cnt, Bd, Mdl, Onl, Eco, InsCnt	T	Numeric

[a] Continente

[b] Continente Modelo

[c] Continente Bom Dia

[d] Continente Online

[e] Continente, Continente Modelo, Continente Bom Dia, Continente Online, Meu Super, Wells, Zippy, Mo, Bagga, Note and Zu

Table C.3: Predictive Variables - Part III

Variable	**Description**	**annual**	**sem**	**semhom**	**Var_sem**	**Var_hom**	**Insignia**	**Category**	**Variable Type**
p_vb_promo	Proportion of sales in promotion	X	X	X			Eco, InsCnt	T	Numeric
per_vb	Promotion of sales on promotion in total customer sales	X	X	X	X	X	Cnt, mdl, bd	T	Numeric
n_lojas_mch	Number of stores visited	X	X	X			InsCnt	T	Numeric
n_internos	Number of internal partners visited	X	X	X			Internal Partners	T	Numeric
flag	If the customer made at least one purchase	X	X	X			Cnt, Mdl, Bd, Msp[a], wells, zippy, mo, bagga, note, zu	B	Numeric
perc_vb_frescos	Percentage of fresh products bought	X	X	X	X	X	InsCnt	T	Numeric
perc_vb_alimentar	Percentage of food products bought	X	X	X	X	X	InsCnt	T	Numeric
perc_VO	Percentage of online sales	X	X		X		InsCnt	T	Numeric
Recency	Time since last visit						Cnt, Mdl, Bd, Onl	T	Numeric
N_BU	Number of different business units bougth	X	X		X		InsCnt	T	Numeric
vb_DC	Amount of sales in several departments	X	X	X	X	X	InsCnt	T	Numeric
lojaPref_Change	If the preferential store has changed in the previous two months							T	Categorical

[a] Meu Super

Appendix D

Support graphs for preprocessing

Table D.1: Account Longevity per Cluster (Years)

Cluster	Class	Average	Standard Deviation	Minimum	Maximum
1. Premium	Non Churn	9.43	3.53	0.00	13.00
	Churn	8.00	4.02	0.00	12.90
2. Every Day Low Price	Non Churn	8.32	4.06	0.00	13.00
	Churn	7.39	4.18	0.00	12.90
3. Kids in the house	Non Churn	8.25	3.94	0.00	13.00
	Churn	7.61	4.02	0.00	12.90
4. No Value	Non Churn	7.39	4.29	0.00	12.90
	Churn	7.02	4.27	0.00	12.90
5. Promo Busters	Non Churn	8.56	3.82	0.00	12.90
	Churn	7.84	3.99	0.00	12.90

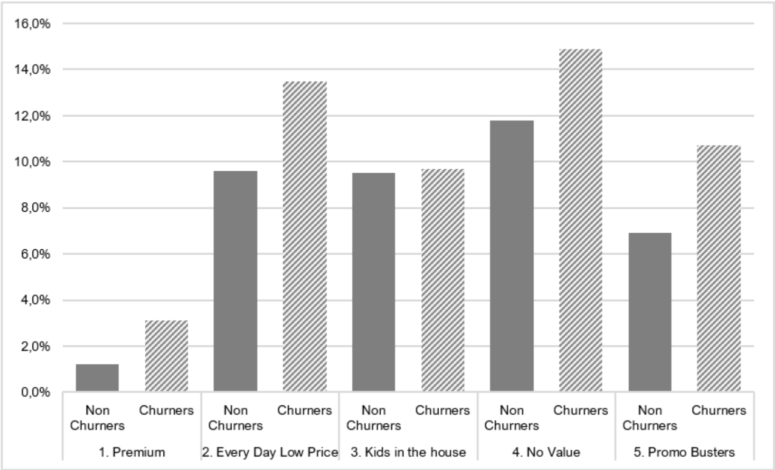
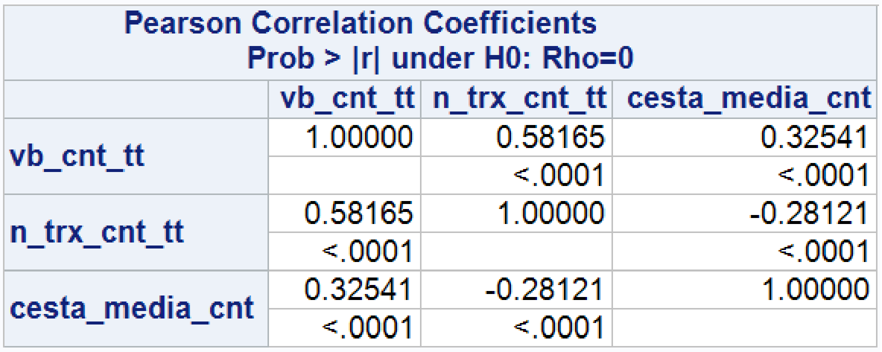


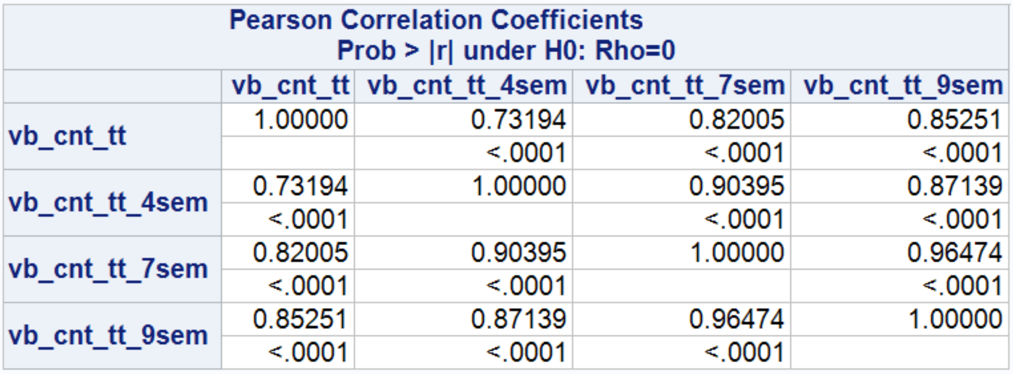
Figure D.1: Distribution of the percentage of missing values for the region according to class and cluster.



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0

	vb_cnt_tt	n_trx_cnt_tt	cesta_media_cnt
vb_cnt_tt	1.00000	0.58165	0.32541
		<.0001	<.0001
n_trx_cnt_tt	0.58165	1.00000	-0.28121
	<.0001		<.0001
cesta_media_cnt	0.32541	-0.28121	1.00000
	<.0001	<.0001	

(a) Variables: sales amount, number of transactions and average basket in the previous year



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0

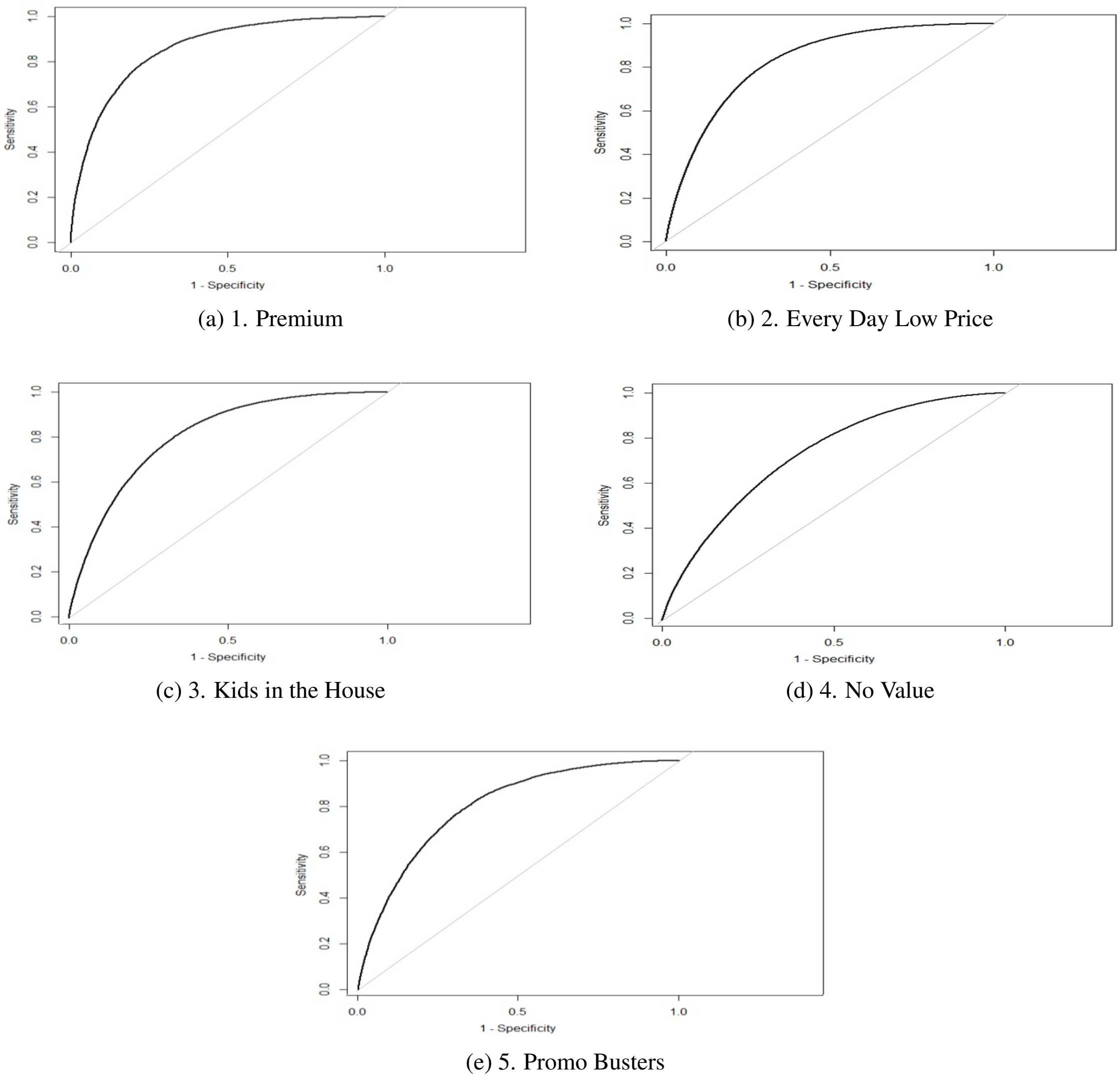
	vb_cnt_tt	vb_cnt_tt_4sem	vb_cnt_tt_7sem	vb_cnt_tt_9sem
vb_cnt_tt	1.00000	0.73194	0.82005	0.85251
		<.0001	<.0001	<.0001
vb_cnt_tt_4sem	0.73194	1.00000	0.90395	0.87139
	<.0001		<.0001	<.0001
vb_cnt_tt_7sem	0.82005	0.90395	1.00000	0.96474
	<.0001	<.0001		<.0001
vb_cnt_tt_9sem	0.85251	0.87139	0.96474	1.00000
	<.0001	<.0001	<.0001	

(b) Variables related to sales amount for different periods

Figure D.2: Pearson correlation coefficient for variables for cluster "Premium"

Appendix E

Regularized Logistic Regression: ROC Curves



(a) 1. Premium

(b) 2. Every Day Low Price

(c) 3. Kids in the House

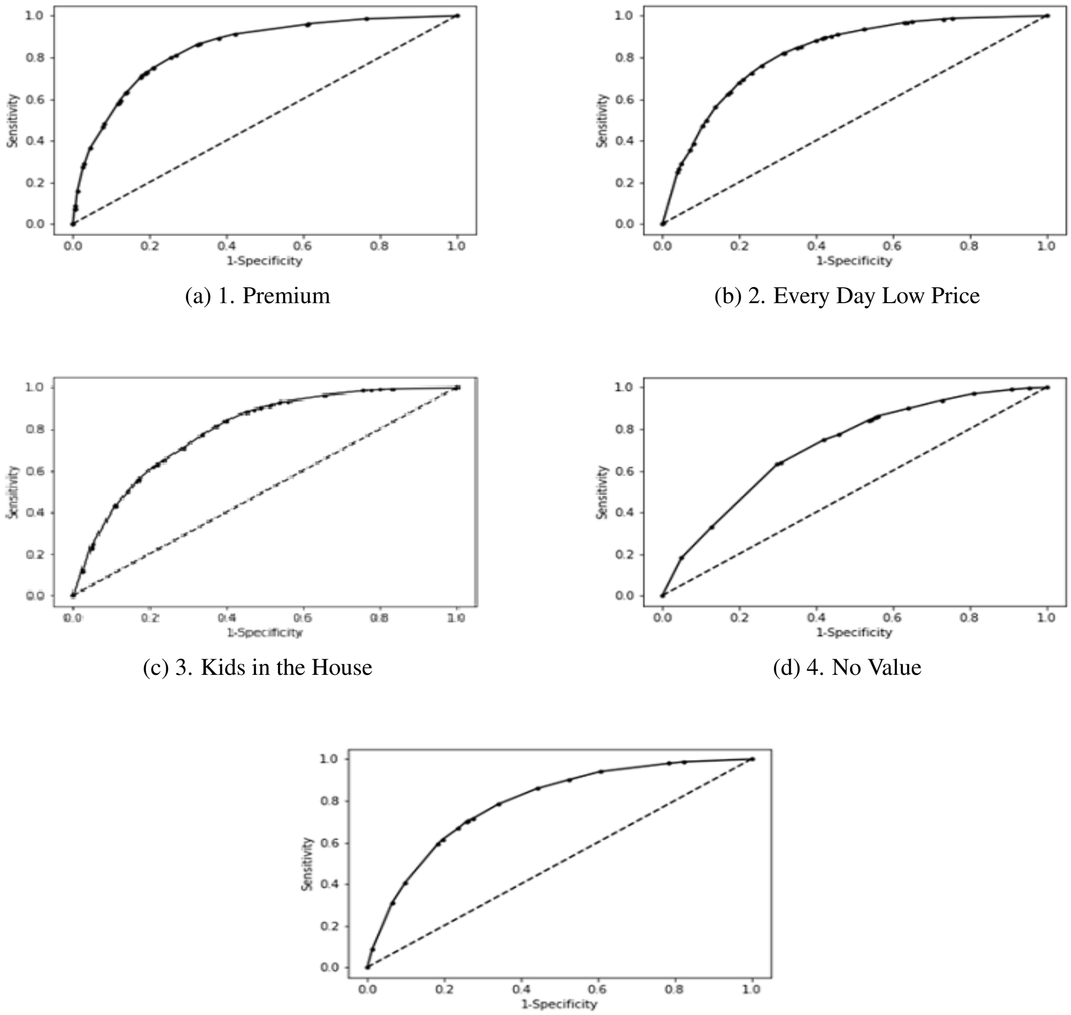
(d) 4. No Value

(e) 5. Promo Busters

Figure E.1: ROC Curve per customer group: Regularized Logistic Regression

Appendix F

Decision trees: ROC Curves



(a) 1. Premium

(b) 2. Every Day Low Price

(c) 3. Kids in the House

(d) 4. No Value

(e) 5. Promo Busters

Figure F.1: ROC Curve per customer group: Decision Tree

Appendix G

Decision trees

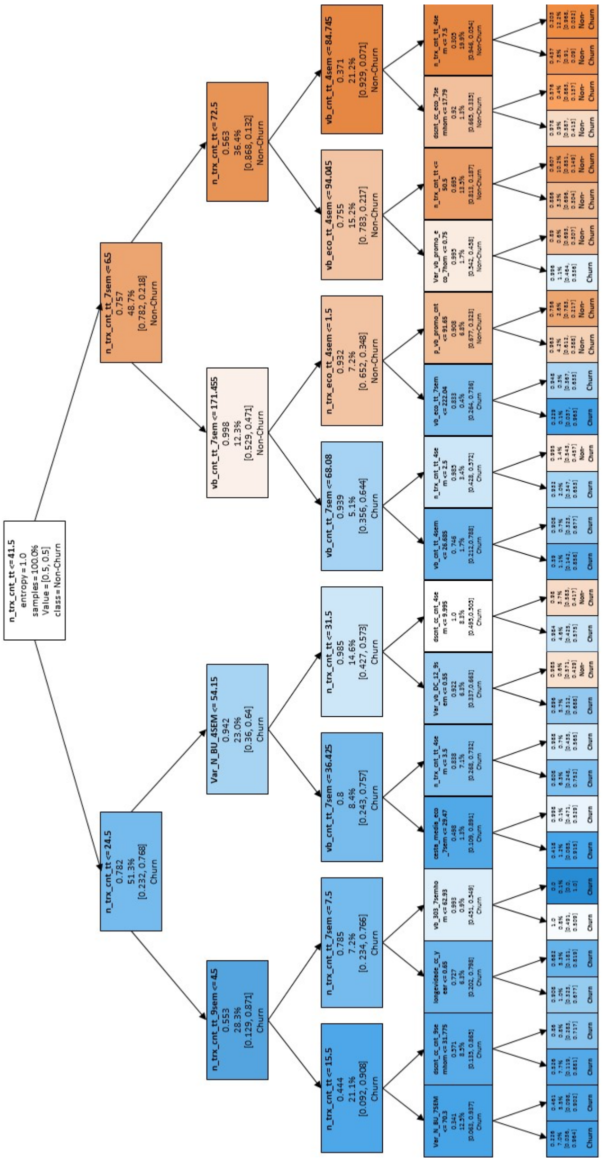


Figure G.1: Decision tree for Cluster 1. Premium

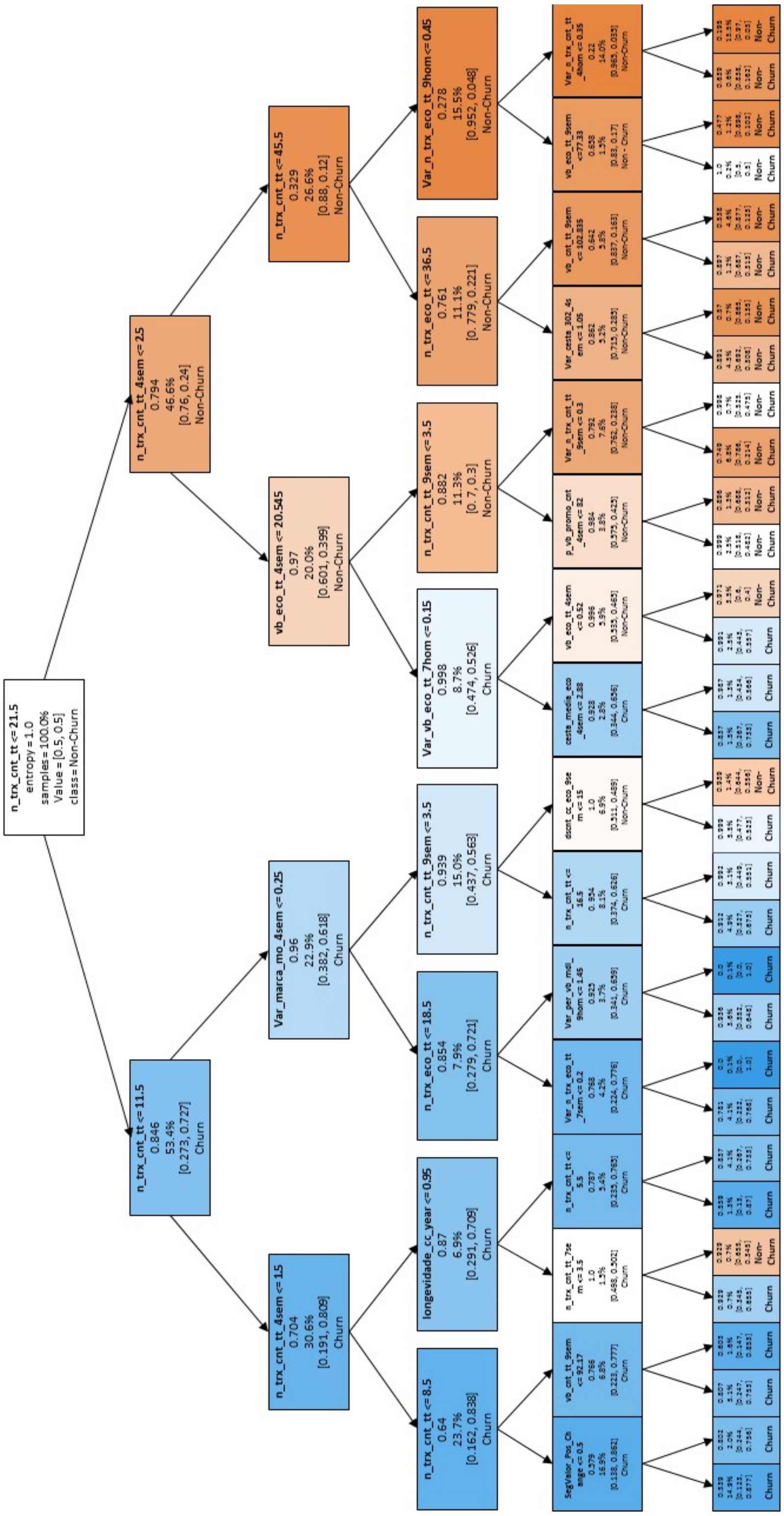


Figure G.2: Decision tree for Cluster 2. Every Day Low Price

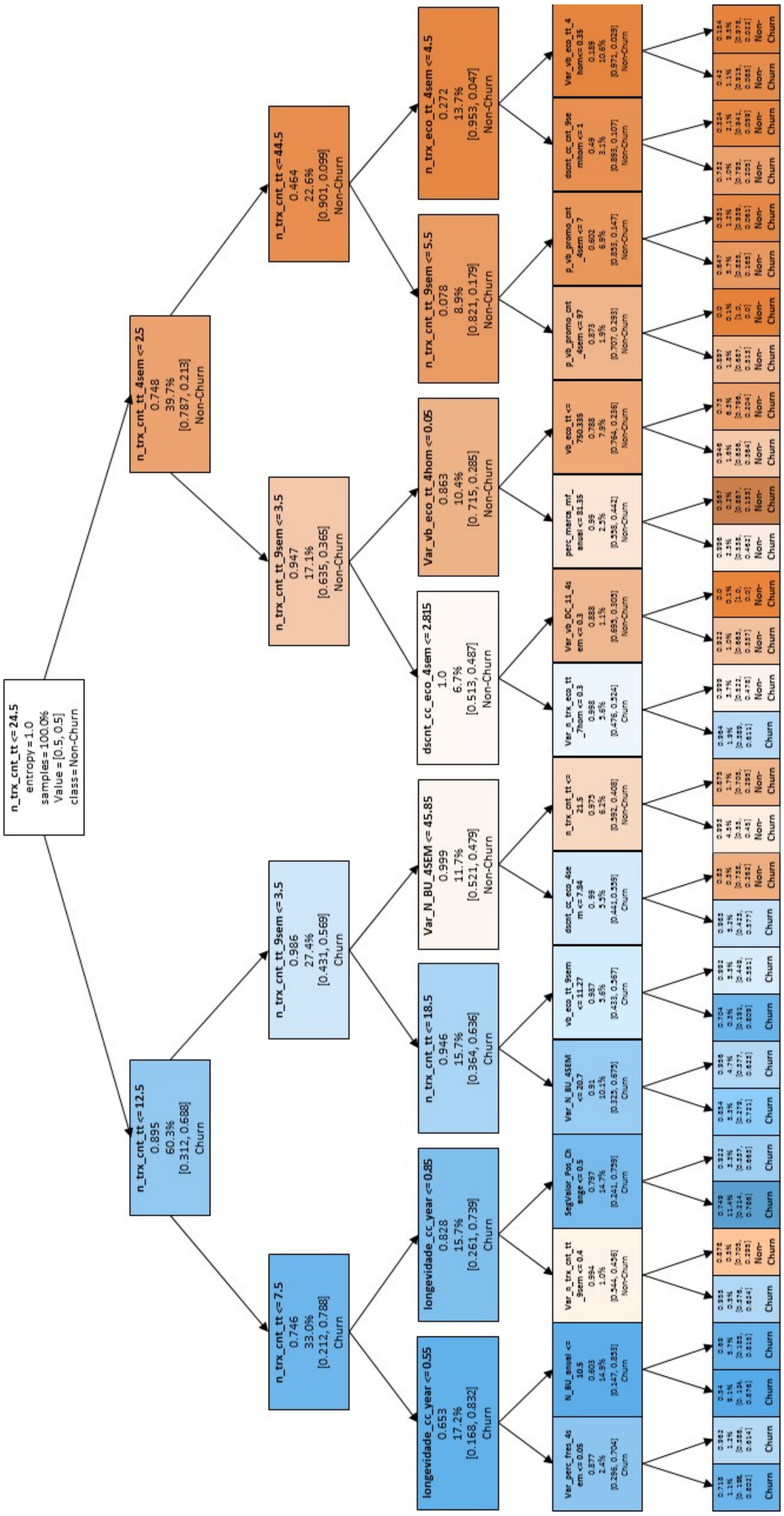


Figure G.3: Decision tree for Cluster 3. Kids in the House

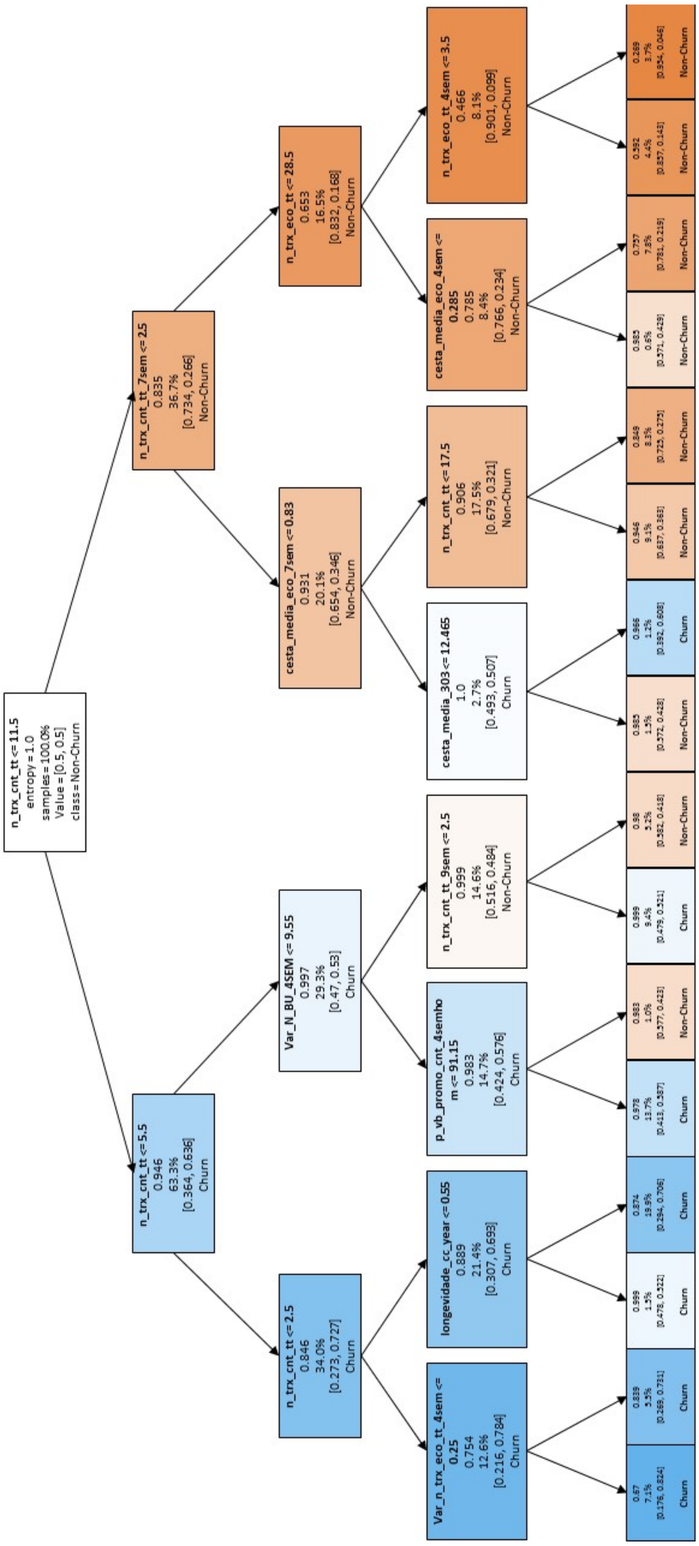

n_trx_cnt_tt <= 11.5
entropy = 1.0
samples = 100.0%
Value = [0.5, 0.5]
class = Non-Churn
n_trx_cnt_tt_7sem <= 2.5
0.835
36.7%
[0.734, 0.266]
Non-Churn
n_trx_cnt_tt <= 5.5
0.946
63.3%
[0.364, 0.636]
Churn
n_trx_eco_tt <= 28.5
0.653
16.5%
[0.832, 0.168]
Non-Churn
cesta_media_eco_7sem <= 0.83
0.931
20.1%
[0.654, 0.346]
Non-Churn
Var_N_BU_4SEM <= 9.55
0.997
29.3%
[0.47, 0.53]
Churn
n_trx_cnt_tt <= 2.5
0.846
34.0%
[0.273, 0.727]
Churn
n_trx_eco_tt_4sem <= 3.5
0.466
8.1%
[0.901, 0.099]
Non-Churn
cesta_media_eco_4sem <= 0.285
0.785
8.4%
[0.766, 0.234]
Non-Churn
n_trx_cnt_tt <= 17.5
0.906
17.5%
[0.679, 0.321]
Non-Churn
cesta_media_303 <= 12.465
1.0
2.7%
[0.493, 0.507]
Churn
n_trx_cnt_tt_9sem <= 2.5
0.999
14.6%
[0.516, 0.484]
Non-Churn
p_vb_promo_cnt_4semhom <= 91.15
0.983
14.7%
[0.424, 0.576]
Churn
longevidade_cc_year <= 0.55
0.889
21.4%
[0.307, 0.693]
Churn
Var_n_trx_eco_tt_4sem <= 0.25
0.754
12.6%
[0.216, 0.784]
Churn
0.269 3.7% [0.954, 0.046] Non-Churn
0.592 4.4% [0.857, 0.143] Non-Churn
0.757 7.8% [0.781, 0.219] Non-Churn
0.985 0.6% [0.571, 0.429] Non-Churn
0.849 8.3% [0.725, 0.275] Non-Churn
0.946 9.1% [0.637, 0.363] Non-Churn
0.966 1.2% [0.392, 0.608] Churn
0.985 1.5% [0.572, 0.428] Non-Churn
0.98 5.2% [0.582, 0.418] Non-Churn
0.999 9.4% [0.479, 0.521] Churn
0.983 1.0% [0.577, 0.423] Non-Churn
0.978 13.7% [0.413, 0.587] Churn
0.874 19.9% [0.294, 0.706] Churn
0.999 1.5% [0.478, 0.522] Churn
0.839 5.5% [0.269, 0.731] Churn
0.67 7.1% [0.176, 0.824] Churn

Figure G.4: Decision tree for Cluster 4. No Value

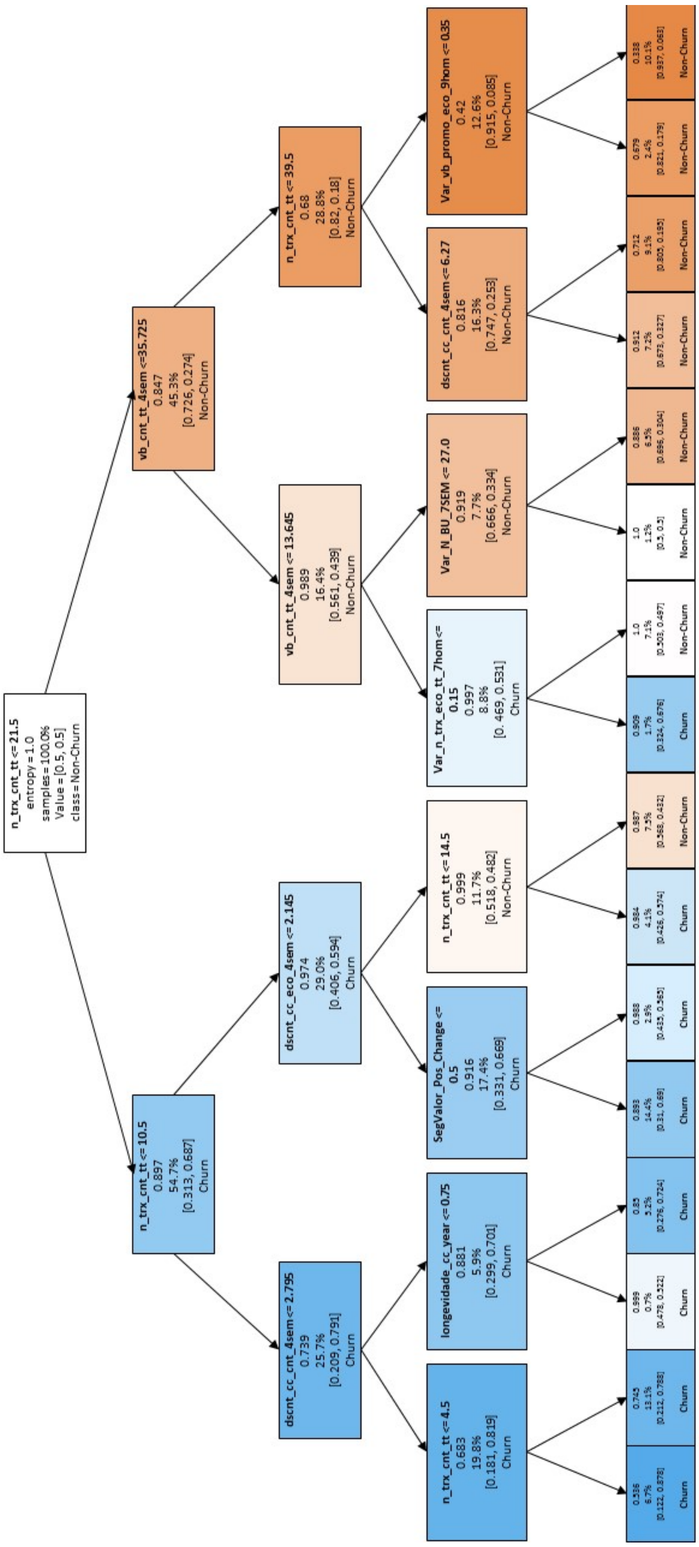


Figure G.5: Decision tree for Cluster 5. Promo Busters