Universidade do Minho
Escola de Ciências

Carlos Manuel da Silva Castro

Engagement Score:
Aplicação ao Retalho Alimentar

outubro de 2018

Engagement Score: Aplicação ao Retalho Alimentar

Carlos Manuel da Silva Castro

UMinho | 2018

Universidade do Minho
Escola de Ciências

Carlos Manuel da Silva Castro

Engagement Score: Aplicação ao Retalho Alimentar

Tese de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação de
Professora Doutora Susana Margarida Ferreira Sá Faria
Doutora Liliana Andreia de Sousa Bernardino

outubro de 2018

Anexo 3

DECLARAÇÃO

Nome
Carlos Manuel da Silva Castro

Endereço eletrónico: carlosmscastro95@hotmail.com Telefone: 912 942 048 /

Número do bilhete de Identidade: 14588552 6ZY3

Título dissertação
Engagement Score: Aplicação ao Retalho Alimentar

Orientador(es):
Susana Margarida Ferreira Sá Faria

Liliana Andreia de Sousa Bernardino Ano de conclusão: 2018

Designação do Mestrado:

Mestrado em Estatística

Nos exemplares das teses de doutoramento ou de mestrado ou de outros trabalhos entregues para prestação de provas públicas nas universidades ou outros estabelecimentos de ensino, e dos quais é obrigatoriamente enviado um exemplar para depósito legal na Biblioteca Nacional e, pelo menos outro para a biblioteca da universidade respetiva, deve constar uma das seguintes declarações:

1. É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE/TRABALHO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

Universidade do Minho, 31/ 10/ 2018

Assinatura: Carlos Manuel Silva Castro

"O sofrimento é passageiro, desistir é para sempre"

Lance Armstrong

AGRADECIMENTOS

Numa fase tão importante, muitas pessoas tiveram um papel fundamental querendo portanto deixar o meu sincero agradecimento.

À professora Susana Faria pela excelente orientação e apoio constante, partilhando os seus conhecimentos, que tiveram um papel fundamental para a realização deste trabalho.

Um agradecimento à Sonae MC e à entidade gestora do Cartão Continente pela oportunidade da realização deste projeto, pela total disponibilidade de acesso aos dados e ferramentas, fazendo com que fosse possível dar o meu contributo.

À minha diretora e orientadora Liliana Bernardino, aos meus colegas de equipa, Patrícia Castro, Filipe Miranda e Camila Matos o meu muito obrigado por todo o apoio durante este processo. Por fim, queria agradecer em especial à Ana Freitas e Ana Januário, por todo o apoio, motivação e dedicação para que todo este trabalho fosse possível.

A todos os meus colegas e amigos que me acompanharam durante este percurso, e por tudo o que fizeram por mim.

Um agradecimento muito especial a uma pessoa que esteve a meu lado durante grande parte desta jornada que já dura há 5 longos anos, aquela pessoa que sabe bem o quão difícil foi, mas que foi algo que sempre ambicionei. A ti Carla Brites, o meu muito obrigado por tudo!

Por último, queria agradecer à minha família, em especial aos meus pais e ao meu irmão por todo o esforço que fizeram para que pudesse cumprir este sonho, por todo o apoio, paciência, estando sempre presentes e por nunca me terem deixado desistir.

A todos o meu muito obrigado!

Carlos Castro

"A gratidão é o único tesouro dos humildes."

William Shakespeare

RESUMO

Com o passar do tempo, é cada vez mais importante para as empresas analisarem atentamente o comportamento dos seus clientes, pois dele depende o desempenho de qualquer empresa. Mais do que conhecer, é necessário prever o seu comportamento no futuro para que sejam tomadas decisões de negócio.

O *Engagement Score* (ES) é uma métrica utilizada para medir o envolvimento de um cliente com uma marca/ empresa, que no contexto da área de retalho alimentar desta empresa é obtida a partir de uma probabilidade.

Para o cálculo desta métrica, foram inicialmente construídos *clusters* para se definir quais os clientes mais envolvidos à marca usando variáveis de negócio, que caracterizam o cliente segundo o seu comportamento de compra. Numa segunda fase, e percebendo quais os clientes mais envolvidos, foram aplicados modelos de regressão logística, que estimaram a probabilidade de o cliente estar envolvido com a empresa.

Para cada cliente, foi calculada a métrica que mede o seu envolvimento com a empresa. Todos os clientes foram divididos em cinco segmentos e posteriormente caracterizados, conhecendo assim um pouco melhor os clientes presentes em cada um deles. Aquilo que se verifica, é que os clientes mais leais, são clientes que não se resumem apenas às insígnias Continente, procurando comprar nas restantes marcas do ecossistema. São clientes que vivem sobretudo em cidades do litoral do país, em cidades com maior poder de compra e com filhos no agregado familiar.

O conhecimento do *Engagement Score* permite à empresa ter um maior conhecimento sobre quais são os seus clientes "mais valiosos", tornando-se mais eficaz nas estratégias de *marketing* aplicadas, de forma a aumentar a fidelidade dos clientes.

Os resultados obtidos desta dissertação serão aplicados na empresa de retalho e irão sustentar a tomada de decisão nas estratégias do cliente.

Palavras-chave: *Engagement Score*; modelos de regressão logística; análise de *clusters*

ABSTRACT

Over time, it is increasingly important for companies to carefully analyze the behavior of their customers, as it is central to the performance of any company. More than knowing, it is necessary to predict their behavior in the future so that key decisions can be made in favor of the company.

The Engagement Score (ES) is a metric used to measure the involvement of a customer with a brand / company, which in the context of the food retail area of this company is obtained from a probability.

For this metric's calculation, clusters were initially built to define which customers are most involved in the brand using business variables, which characterize the customer according to their buying behavior. In a second phase, and realizing which customers were most involved, logistic regression models were applied, which estimated the probability of the customer being involved with the company.

For each customer, the metric that measures their involvement with the company was calculated. All customers were divided into five segments and later characterized, thus knowing a little better the customers present in each one of them. What happens is that the most loyal customers are those who are not limited to the Continente brands, looking to buy in the remaining brands of the ecosystem. They are clients who live mainly in towns on the country's coast, in cities with greater purchasing power and with children in the household.

Knowledge of the Engagement Score enables the company to gain a better understanding of who their "most valuable" customers are, and in this way becoming more effective in applied marketing strategies in order to increase customer loyalty.

The results obtained from this dissertation will be applied in the retail company and will support the decision making in the client's strategies.

Keywords: Engagement Score; logistic regression models; cluster analysis

CONTEÚDO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CLV - *Customer Lifetime Value*
DC - Departamento Comercial
ES - *Engagement Score*
MCH - Modelo Continente Hipermercados, S.A.
MLG - Modelos Lineares Generalizados
NPS - *Net Promoter Score*
OR - *Odds Ratio*
RFM - *Recency, Frequency and Monetary Value*
SONAE - Sociedade Nacional de Estratificados
SOW - *Share of Wallet*
WOM - *Word of Mouth*

1

INTRODUÇÃO

Na atualidade, numa era da Globalização, em que os consumidores têm um maior conhecimento da oferta disponível no mercado, é cada vez mais importante para as empresas fidelizar os seus clientes, adaptando as estratégias por forma a crescerem. Por outro lado, as empresas de retalho têm vindo a crescer em número, sendo a concorrência cada vez maior, o que faz com que as empresas tentem estimular uma comunicação mais personalizada junto dos seus clientes, tentando novas estratégias e objetivos mais ambiciosos.

Desta forma, as empresas têm investido em perceber as reais necessidades dos seus clientes, para que as suas futuras decisões estejam de encontro com aquilo que os clientes precisam e procuram.

Um dos grandes investimentos das empresas, passa por criar bases de dados com o registo de todos os clientes e das suas compras, para resolver os problemas que envolvam os seus clientes e perceber as suas tendências, prever o seu comportamento e tomar decisões direcionadas por forma a satisfazer as suas necessidades.

1.1 LOCAL DE ESTÁGIO

O presente trabalho foi desenvolvido durante a realização de um estágio na SONAE no âmbito do Mestrado em Estatística, da Escola de Ciências, da Universidade do Minho.

A SONAE – Sociedade Nacional de Estratificados, fundada na Maia, a 18 de agosto de 1959 pelo banqueiro e empresário Afonso Pinto de Magalhães, natural de Arouca. Um grupo de amigos tinha conseguido convencê-lo a fabricar estratificados a partir de um desperdício bastante comum nas encostas do Douro: o engaço de uva. Em 1965, a SONAE contrata o engenheiro químico de 26 anos, Eng.º Belmiro de Azevedo, que inicia grandes reformas na produção e a empresa começa a recuperar financeiramente.

No período conturbado do 25 de Abril de 1974, Afonso Pinto Magalhães, como muitos outros empresários na altura, exílou-se no Brasil, ficando o Eng.º Belmiro de Azevedo como gestor dos destinos da SONAE.

Com a morte de Afonso Pinto de Magalhães em 1982, o Eng.º Belmiro de Azevedo travou uma longa batalha judicial com a família do antigo presidente, e conseguiu a maioria do capital, assumindo os destinos da empresa.

Nos anos 80, a SONAE diversificou-se para a área do retalho[1], e em 1985, abriu o primeiro hipermercado em Portugal, o Continente (Matosinhos), que marca o início da atividade da SONAE Distribuição, resultado da *joint-venture* entre a SONAE e a francesa Promodès (antigo grupo francês de retalhistas). Em 1991, dá-se a criação da *Sub-holding* SONAE Imobiliária, com vista à construção de Centros Comerciais de apoio aos hipermercados.

Representando mais um passo seguro na consolidação do negócio da distribuição, em 1996, a SONAE aposta no retalho especializado e surge a Worten, cadeia que rapidamente se distingue como líder em Portugal pela ampla e especializada oferta de eletrodomésticos, eletrónica de consumo e de entretenimento.

Atualmente, a SONAE é uma multinacional que gere um portefólio diversificado de negócios, estando presente em 90 países. O grupo SONAE é composto pela SONAE MC (retalho alimentar), SONAE SR (retalho não-alimentar), SONAE RP (imobiliário de retalho), SONAE FS (serviços financeiros), SONAE IM (gestão de investimentos), SONAE Sierra (centros comerciais) e por fim a NOS (telecomunicações).

A SONAE MC trabalha com um conjunto de marcas distintas que oferecem uma variedade de gama de produtos: Continente (hipermercados), Continente Modelo e Continente Bom dia (supermercados de conveniência), Meu Super (lojas de proximidade em formato *franchising*), Bom Bocado, Bagga (cafetarias e restaurantes), Go Natural (supermercados e restaurantes saudáveis), Make Notes, Note! (livraria/papelaria), Well's (saúde, bem estar e ótica), ZU (produtos e serviços para cães e gatos), Dr. Well's (clínicas medicina dentária e medicina estética) e Ibersol (restauração).

O Modelo Continente Hipermercados, S.A. lançou, a 23 de janeiro de 2007, o cartão cliente, oferecendo aos clientes descontos nos hipermercados Continente, Continente Modelo, Continente Bom Dia e Continente Online. Nos primeiros 12 dias, as adesões chegaram a um milhão, com uma cadência de 700 clientes por minuto, e no final do primeiro ano já se registavam $2,4$ milhões de cartões. Desde a sua génese, o Cartão Continente tem realizado um caminho de crescimento constante, tendo atingido em 2017, os $3,7$ milhões de contas ativas (contas que apresentam compras nos últimos 12 meses). Neste momento o cartão Continente já conta com vários parceiros, sendo 18 as marcas parceiras permanentes que exploram mais de 2000 lojas onde se pode usar o Cartão Continente, as lojas Continente, Well's, Note!, ZU, Bagga, Meu Super, Zippy, MO, os postos Galp aderentes e no Grupo Ibersol, na qual fazem parte as lojas Pasta Caffé, Miit, Pizza Hut, Pans & Company, Roulotte, Ò Kilo, KFC, SOL e Burger King.

1 Adjetivo que se utiliza no âmbito do comércio em referência à atividade que se realiza a retalhe (à unidade)

A Figura 1 ilustra o ecossistema, com os diversos parceiros envolvidos com o Cartão Continente.

Figura 1: Ecossistema do Cartão Continente

1.2 OBJETIVOS DO PROBLEMA

O *Engagement Score* consiste em classificar a fidelização de um cliente a uma determinada marca ou empresa. Existem várias definições na literatura, dependendo do contexto em que esteja inserido e a visão de negócio da empresa. No caso da SONAE, apesar de existir um programa de fidelização desde 2007, o contexto de como essa fidelização se desenvolve é considerado como não-contratual, já que o cliente não possui nenhum contrato com a empresa nem qualquer tipo de obrigação (Aghaie (2009)).

O principal objetivo deste trabalho é classificar a fidelização de cada cliente, calculando um *score* que indica o nível de envolvimento de cada cliente com a marca, sendo esta métrica obtida através de variáveis comportamentais do cliente com a finalidade de:

- Relativizar a lealdade dos clientes;
- Identificar quais os clientes com maior propensão a deixar de comprar nas lojas;
- Classificar os clientes de forma a otimizar o investimento promocional.

Os benefícios que o conhecimento desta métrica dá à empresa, passa por um maior asserto na escolha do seu público alvo para cada uma das campanhas. Permite ainda perceber se existe uma evolução do cliente com o passar do tempo, se este se encontra mais envolvido com a marca ou se por outro lado, será necessário criar estímulos para aumentar a fidelidade do cliente com a marca. Assim, a empresa está mais perto de perceber qual o comportamento futuro do cliente e agir em conformidade com as suas reais necessidades.

Na elaboração deste estudo é analisada a insígnia Continente concernente ao Modelo Continente Hipermercados, S.A. (MCH).

Uma vez que serão utilizados os dados transacionais do Cartão Continente, e existindo várias marcas associadas, foram propostas duas formas alternativas para a elaboração deste estudo. A primeira, passaria pelo cálculo do *Engagement Score* por parceiro (neste caso para o parceiro com maior impacto ao nível das vendas, transações e dados, é o Continente, o Continente Modelo e o Continente Bom Dia), a segunda alternativa seria o cálculo do *Engagement Score* para o ecossistema. Posto isto, foram analisadas as vantagens, quer a nível metodológico, quer a nível de aplicação das duas alternativas.

A nível metodológico, apresentam-se as vantagens para cada uma das alternativas e de que forma é que alguns pontos teriam de ser tidos em consideração para a construção do modelo final.

Tabela 1: Vantagens por Metodologia

Metodologia	
Por Parceiro	**Ecossistema**
Processo individual por parceiro	Processo único que agrega informação de todos os parceiros num único *score*
Potencial seleção de diferentes métricas específicas de cada negócio	Seleção *one-shot* de métricas que espelhem o comportamento dos clientes em todos os ecossistemas
Período de estabilidade adaptado a cada negócio	Potencial de sobrevalorização de clientes em faixas etárias ativas (Zippy) ou com animais de estimação (Zu) em detrimento de clientes que poderão nunca ser clientes-alvo de determinadas marcas
Entrada de novos parceiros no Cartão Continente não exige reprocessamento do *Engagement Score*	

Analisando a Tabela 1, verifica-se que a análise por parceiro permite uma abordagem que vai de encontro às características de cada negócio, desde a escolha de métricas específicas dos vários parceiros até à seleção dos períodos de estabilidade adequados a cada negócio. Como estamos perante negócios muito díspares entre si, os períodos de estabilidade variam muito

de parceiro para parceiro. Relativamente à análise do ecossistema, tem em conta o estilo de vida do cliente e as motivações para usar as diversas marcas.

Foi ainda estudada qual a aplicação que será dada aos resultados dos modelos obtidos para cada uma das diferentes abordagens.

Tabela 2: Vantagens por Aplicação

Aplicação	
Por Parceiro	**Ecossistema**
Permite a ação direta do *score* para o parceiro que se pretende acionar	Permite a ação segundo o *Engagement Score* geral com o Cartão Continente, independentemente do envolvimento com o parceiro a acionar
	Potencial para angariação/ ativação em novos parceiros

A Tabela 2 permite concluir que, por parceiro a aplicação dos resultados será mais objetiva, sabendo exatamente o nível de envolvimento do cliente com cada uma das marcas e a evolução do seu comportamento ao longo do tempo. No que diz respeito à aplicação no ecossistema, dá uma informação mais detalhada do envolvimento com o cliente com o cartão, o que permite estimar a reação e aceitação à entrada de novos parceiros, porque um cliente que seja muito envolvido com o cartão, tem maior propensão a aderir às novas marcas que entrem no ecossistema.

Em suma, após a análise das respetivas vantagens para cada uma das opções, verifica-se que por parceiro permite, estatisticamente, a obtenção de resultados com menor margem de erros, enquanto que para o ecossistema, permite um maior conhecimento do cliente e da forma como este envolve com as várias marcas. Outras das vantagens do cálculo por parceiro prende-se pelo facto de o Cartão Continente desde a sua criação ter vindo a sofrer um crescimento constante ao longo do tempo, com a entrada de novas parcerias, e desta forma não será preciso recalcular o modelo, mas apenas a criação de um novo modelo para o novo parceiro que entrar no cartão. Por último, relativamente aos períodos de estabilidade, o cálculo do *Engagement Score* para o ecossistema teria de se definir um período de estabilidade que combine as frequências naturais dos diferentes negócios, o que não permite uma correta avaliação e perceção do modelo. Com base em todos os fatores e argumentos apresentados, optou-se pela realização do *Engagement Score* para os dados do parceiro Continente (Continente, Continente Modelo e Continente Bom Dia).

1.3 ESTRUTURA DO DOCUMENTO

Esta dissertação encontra-se dividida em sete capítulos.

No Capítulo 1 é introduzido o tema do trabalho e apresentada a instituição onde foi realizado o estudo. É ainda explicada a importância do tema no mundo do retalho nos dias de hoje, a definição neste contexto e as motivações.

No Capítulo 2 faz-se uma revisão da literatura existente em relação ao *Engagement Score*, a sua importância, as diferentes definições para os vários contextos de negócio, os antecedentes que originam a uma fidelização por parte de um cliente para com uma marca, e por fim, os vários tipos de metodologias que são aplicadas na literatura e que serão usadas para o cálculo do *Engagement Score*.

A descrição das metodologias aplicadas para o cálculo do *Engagement Score* é apresentada no Capítulo 3, em particular, o modelo de regressão logística e análise de *clusters*.

No Capítulo 4 descreve-se a base de dados utilizada e apresenta-se uma análise exploratória dos dados.

No Capítulo 5 apresentam-se os obtidos nos modelos preditivos abordados no Capítulo 3.

A discussão dos resultados é apresentada no Capítulo 6.

Por último, no Capítulo 7 são apresentadas as principais conclusões retiradas do estudo, referindo-se os contributos, as suas limitações e considerações finais. Numa fase final, são apresentadas algumas recomendações para trabalhos futuros.

1.4 SOFTWARE UTILIZADO

Ao longo do estudo foram utilizados, o *SAS Enterprise Guide* para o tratamento dos dados e aplicação das metodologias, e o *R* como ferramenta de apoio na análise análise exploratória dos dados e na validação dos resultados das metodologias.

2

REVISÃO DA LITERATURA

Nos últimos anos, a fidelização dos clientes é um tema de grande interesse entre gerentes e consultores de diversas indústrias e empresas em todo o mundo, como é evidenciado pela grande quantidade de artigos, *blogs*, fóruns de discussão e seminários gerados sobre este tema (Sashi (2012)). Conceitos como a satisfação e a lealdade dos clientes, tornaram-se conceitos muito importantes na gestão moderna dos clientes (Kristensen and Eskildsen (2011)).

2.1 FIDELIZAÇÃO

Clientes fiéis são clientes que estão dispostos a comprar os produtos da mesma marca, mesmo que o mercado lhes apresente uma infinidade de opções. São clientes menos sensíveis às mudanças de preço e optam por produtos *premium* (Rishika et al. (2013)).

Nos dias de hoje, com tantas alternativas disponíveis para os clientes e com a facilidade com que a internet permite encontrar novos produtos, a fidelização dos clientes tornou-se uma das maiores preocupações para uma empresa (Cuillierier (2016)).

2.1.1 *A Importância do Conhecimento da Fidelização e a sua Evolução*

Segundo (Kotler (2000)), a estimativa de atrair um novo cliente pode custar 5 vezes mais do que manter um já existente, daí a importância que tem para as empresas fidelizar os seus clientes. Outro dos fatores de maior relevância para a fidelização de um cliente está relacionado com o facto de que a probabilidade de vender a um cliente já existente, é de 60 a 70%, enquanto que a probabilidade de vender a um novo cliente é de 5 a 20% (Saleh (2015)). Saleh (2015), afirma ainda que os clientes existentes são 50% mais propensos a experimentar novos produtos e gastar 31% a mais, do que novos clientes. O aumento das taxas de fidelização de clientes em 5% aumenta os lucros entre 25 a 95%. Um cliente fiel é mais propenso a comprar mais quando está envolvido com uma marca, do que um novo e inconstante cliente.

Os clientes dão mais atenção ao *Word of Mouth* (WOM) (boca-a-boca) porque é entendido como credível e é gerado por pessoas que não tem interesse próprio em valorizar um

produto (Brown et al. (2007)). O WOM é definido como o mais importante e efetivo canal de comunicação (Keller (2007)), impulsionando o crescimento da empresa.

A Tabela 3 ilustra algumas das diferenças de comportamento entre os clientes fidelizados e os não fidelizados a uma marca (Moutella (2002)).

Tabela 3: O impacto da fidelização

(Fonte: Moutella (2002))

Fidelização	
Clientes Fidelizados	**Clientes não Fidelizados**
Compram em múltiplos canais (telefone, loja, online, etc.)	Utilizam poucos ou um único canal
Consomem mais	Compram eventualmente
Reclamam para resolver o problema	Deixam de comprar
Recomendam a marca a 5 pessoas	Influenciam negativamente 13 pessoas
Mais barato para manter o cliente	Mais caros de se recuperar
Aumentam as vendas e reduzem as despesas	Diminuem as vendas e aumentam as despesas

As várias manifestações de envolvimento de um cliente com uma marca, passa por várias fases, desde o momento em que se adquire um produto ou serviço pela primeira vez até ao momento em que deixa de se relacionar com a empresa. As fases principais são: marketing, aquisição, retenção e por fim a sua recuperação. Este processo é denominado de ciclo de vida de um cliente como é ilustrado na Figura 2.

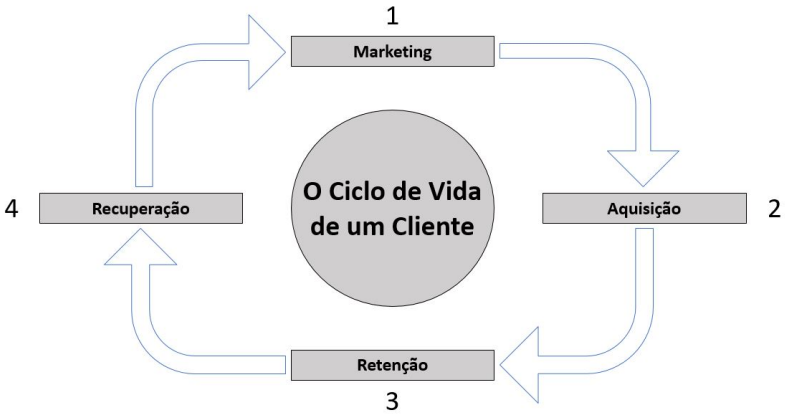

Figura 2: Ciclo de Vida de um Cliente

Numa fase inicial é fundamental pensar onde está o público-alvo, como chamar a sua atenção e incentivar o primeiro contacto com sua marca, gerando futuramente interesse pela compra dos produtos ou serviços. Uma das formas de entrar em contacto com o cliente é através

de ações de marketing (esta fase também é conhecida como a fase de consciencialização dos clientes para com a marca). Como tal, é necessário selecionar as perspetivas "corretas" para a campanha de aquisição. Dependendo do objetivo, uma perspetiva "correta" pode ser um cliente com máxima probabilidade de compra. Historicamente, a técnica mais utilizada tem sido o modelo *Recency, Frequency and Monetary Value* (RFM). Com base no pressuposto de que o cliente "certo" no futuro se parece muito mais com o cliente "certo" no passado, a abordagem tradicional de modelagem RFM cria grupos de clientes com base nas suas características de compras anteriores e atribui probabilidades ou "*scores*" a cada grupo de acordo com o seu comportamento (Bijmolt et al. (2010)).

Uma segunda fase passa pela aquisição de um cliente, que também pode ser designada como ativação, sendo essa aquisição resultado das campanhas realizadas na primeira fase do ciclo de vida de um cliente. O ponto central desta fase é estimular que o cliente experimente ou ative os produtos/serviços da marca, fazendo com que se torne um cliente regular, estimulando repetição de compras. Durante esta fase um cliente, este já se relaciona regularmente com a marca, por isso, o principal foco é torná-lo fiel, sendo esta fase responsável pelo maior gasto de tempo e de dinheiro. Afinal, a fidelidade está correlacionada com a satisfação e para o cliente ficar satisfeito, a empresa precisa de realizar algumas adequações em processos, formas de atendimento ou, simplesmente, criar ações exclusivas para que os clientes se sintam importantes. Além de tentar rentabilizar esse cliente, oferecendo produtos ou serviços de *cross sell* (oferecer a clientes existentes produtos complementares àqueles que já foram ou estão a ser adquiridos) ou *up sell* (expor ao cliente produtos *Premium* que são mais caros, fornecendo-lhe um melhor serviço/produto), vale a pena apostar em programas de fidelidade e na criação de uma área de relacionamento para atender seus clientes mais rentáveis, além de desenvolver comunicações que tragam o sentimento de exclusividade.

A retenção de clientes é a terceira fase na construção de uma estratégia de fidelização de um cliente, e refere-se à capacidade de uma empresa reter os seus clientes num determinado período de tempo. A alta retenção de clientes significa que os clientes do produto ou do negócio tendem a retornar, continuar a comprar ou, de alguma outra forma, não "desertar" para outro produto ou negócio, ou a não utiliza-lo totalmente. O objetivo dos programas de retenção de clientes é ajudar as empresas a reter o maior número possível de clientes, geralmente por meio de iniciativas de fidelidade do cliente e de fidelidade à marca. É importante lembrar que a retenção de clientes começa com o primeiro contacto de um cliente com uma empresa e continua durante toda a vida útil de um relacionamento, e os esforços de retenção bem-sucedidos levam em consideração todo esse ciclo de vida. A capacidade de uma empresa de atrair e reter novos clientes está relacionada não apenas a seus produtos ou serviços, mas também à maneira como presta serviços a seus clientes atuais, ao valor que os clientes realmente geram como resultado da utilização das soluções e à reputação que cria e em todo o mercado.

Por fim, a fase de recuperação é a última fase do ciclo de vida de um cliente que é quando, eventualmente, a empresa perde o cliente. Após a perda do cliente, a empresa precisa então de estabelecer um processo de recuperação, em que tem de analisar ou decidir quais os clientes que apresentam um baixo valor para a marca, e deixa-los sair, e quais os clientes que apresentam grande valor para a marca e procurar reconquista-los, tentando sempre perceber os motivos que levam o cliente a abandonar a empresa.

Na literatura, até ao ano 2005, poucos eram os artigos que citam o envolvimento do consumidor (*consumer engagement*), envolvimento de clientes (*customer engagement*) ou envolvimento com uma marca (*brand engagement*). Desde então, estes termos têm sido cada vez mais usados. Apesar da notória crescente popularidade do termo *engagement*, vários autores tentaram definir, sendo que são várias e variadas as suas definições. Segundo (Gonring (2008)), a fidelidade do cliente é um conceito que evoluiu ao longo do tempo e difícil de ser definido.

2.1.2 *Engagement Score*

A definição do *Engagement Score* é muito ambígua chegando mesmo a ser, em algumas situações, contraditória. Caruana (2003), afirma que o conceito de fidelidade evoluiu em largura e em profundidade ao longo dos anos, sendo que a largura reflete-se em múltiplos focos que podem incluir fidelidade às marcas, produtos, vendedores, lojas e serviços entre outros. A primeira investigação sobre este tema foi escrito por (Copeland (1923)). Segundo (Homburg and Giering (2001)), neste estudo e nas investigações realizadas na primeira metade do século XX, a fidelização do consumidor visa exclusivamente o comportamento do cliente, onde a fidelização se traduzia apenas na repetição de compra de um produto ou serviço (Brown (1952); Churchill (1942)).

Na segunda metade do século XX, vários foram os estudos realizados para tentar identificar e/ou prever o comportamento dos consumidores, como por exemplo:

- Medir a lealdade através da probabilidade de comprar ou recomprar um produto (Lipstein (1959); Kuehn (1962)).
- Identificação dos fatores de fidelização a uma loja com vista a aumentar o volume de vendas (Tate (1961)). Estes estudos, foram na década de 70 desenvolvidos de forma a explicar a fidelização a determinadas lojas (Fry et al. (1973); Lessig (1973); Bellenger et al. (1976)).

Até este momento apenas o comportamento do cliente era considerado como um fator de fidelização. No entanto, (Day (1969)), no seu artigo "*A two-dimensional concept of brand loyalty*", critica a forma de pensar, afirmando que não só o comportamento mas também a atitude dos consumidores deveriam ser tidos em conta para a fidelização.

Engel and Blackwell (1982) definiram a lealdade à marca como "a resposta preferencial, atitudinal e comportamental para uma ou mais marcas em uma categoria de produtos expressada por um período de tempo por um consumidor".

De acordo com (Assael (1992)), a lealdade à marca é "uma atitude favorável em relação a uma marca, resultando em uma compra consistente da marca ao longo do tempo". Esta definição foi também apoiada por (Keller (1993)).

Kotler et al. (1999), referem ainda que a fidelização do consumidor mede a intenção dos consumidores voltarem a fazer compras na empresa, bem como a sua vontade em estabelecer atividades de parceria com a mesma.

Money (2004), chama a atenção para a importância da comunicação ao nível da publicidade e das referências positivas WOM no processo de decisão do consumidor. O mesmo autor afirma que as referências positivas "são 7 vezes mais eficazes do que a publicidade impressa e 4 vezes mais eficazes do que o pessoal de vendas a levar os clientes a mudarem de marca" conduzindo, de acordo com (Raymond and Tanner Jr (1994)), a $61,4\%$ das novas vendas diretas. Contudo, Swan and Oliver (1989) afirmam que nem todos os clientes com uma atitude positiva perante o produto o recomendam a outros.

Assim, (Dick and Basu (1994)) foram os primeiros a defender a análise conjunta das dimensões comportamental e atitudinal, afirmando que esta abordagem agregada permite uma melhor avaliação da fidelização, dos seus antecedentes e consequentes. Estes autores propuseram um modelo integrado destas relações para os consumidores finais baseando-se em aspetos cognitivos, afetivos e conotativos. Com base neste modelo, estes autores sugerem que a fidelização não é um conceito dicotómico em que se divide entre clientes fieis ou não fieis a uma marca, mas por outro lado, existe também duas posições intermédias que são a fidelização latente e a fidelização espúria, como se pode observar na Tabela 4.

Tabela 4: Tipos de Fidelização

(Fonte: Dick and Basu (1994))

		Comportamento	
		Forte	Fraco
Atitude	Forte	Fidelização Verdadeira	Fidelização Latente
	Fraca	Fidelização Espúria	Inexistência de Fidelização

Mais tarde, (Oliver (1997)), com base nos fatores cognitivo, afetivo e conotativo, propõe uma nova sequência para a classificação dos clientes fiéis (à marca) que se hierarquiza da seguinte forma:

- **Fidelização cognitiva**: está associada à convicção que a marca apresenta um desempenho superior quando comparado à concorrência. Esta convicção é derivada de informações disponíveis sobre o preço, qualidade e características do produto;

- **Fidelização afetiva**: traduz-se no sentimento de afeto que o cliente cria relativamente a dada marca, resultado de experiências satisfatórias da aquisição e consumo do produto;

- **Fidelização conotativa**: indica a intenção comportamental de recomprar o produto, devido às sucessivas experiências afetivas positivas;

- **Fidelização de ação**: representa o nível de fidelização mais elevado, observa-se quando a intenção de voltar a comprar a marca se torna numa realidade.

No mercado atual, caracterizado pela diversidade de ofertas concorrentes e novidades contínuas, existe uma tendência decrescente para os consumidores se tornarem leais e os atuais clientes só se tornam fiéis se conseguirem perceber que a empresa continua a ser a melhor alternativa de mercado.

Em síntese, podemos esperar que um cliente fiel seja aquele que volta a comprar à organização determinado produto ou similar, mantendo um nível de consumo frequente por um longo período de tempo, de forma preferencialmente exclusiva, não mudando de fornecedor mesmo num momento de crise, tornando-se não só cliente mas acima de tudo parceiro da organização, recomendando-a e defendendo perante outros potenciais clientes (Lovelock and Wright (2002)).

Tendo em conta que a fidelização é um dos grandes objetivos a ser alcançado pelas empresas, é importante tentar perceber de que forma é que se pode conquistar a fidelidade dos clientes, mantendo os seus clientes mais valiosos. Desta forma, é importante determinar os principais elementos que levam à fidelização, ou seja, os antecedentes da fidelização.

2.2 antecedentes da fidelização

Ao longo dos anos vários foram os conceitos apontados como sendo os principais determinantes da fidelização. Vários autores consultados tentaram explicar a fidelização, mas com os diversos estudos concluiu-se que a fidelização era resultado da combinação de diversos conceitos. Neste estudo, serão abordados como os principais antecedentes da fidelização, i) a satisfação, ii) a qualidade percebida, iii) o valor percebido, iv) o preço percebido, v) a confiança, vi) o compromisso, vii) a comunicação, viii) a imagem organizacional e ix) as barreiras à mudança de fornecedor.

As seguintes secções oferecem uma breve explicação para cada um dos vários antecedentes de fidelização do cliente.

2.2.1 *Satisfação*

A satisfação e a fidelização do cliente estão relacionadas, funcionando a satisfação como um antecedente da sua fidelização (Fornell (1992); Reichheld (1996)). Como tal, a satisfação do cliente é uma questão central para as empresas que pretendam criar uma vantagem competitiva sustentável (Patterson (1993)).

Embora exista na literatura diferenças significativas na definição de satisfação, todas partilham de três pontos em comum (Giese and Cote (2000)):

- A satisfação do cliente é baseada numa resposta emocional ou cognitiva;
- A resposta diz respeito a um foco particular (produto, experiência de consumo, etc.);
- A resposta ocorre num determinado momento (após o consumo, após a escolha, etc.).

Westbrook (1987) considera que a satisfação é uma resposta emocional global às expectativas reais de consumo. Howard and Sheth (1969) afirmam que a satisfação corresponde à perceção por parte do consumidor, de uma recompensa adequada pelo produto ou serviço. A qualidade é avaliada comparando os resultados obtidos pela experiência e as expectativas criadas em relação à marca.

Anderson et al. (1994); Fornell (1992), enumeram vários benefícios importantes para a empresa resultantes da elevada satisfação e fidelização dos clientes, dos quais se destacam:

- Redução da sensibilidade/ elasticidade ao preço (clientes satisfeitos estão mais dispostos a pagar pelos benefícios que recebem e têm maior probabilidade de tolerar aumentos de preço, aumentando as margens);
- Redução dos custos de transação no futuro (como os clientes satisfeitos tendem a comprar com maior frequência e em maior volume, e compram outros bens ou serviços oferecidos pela empresa). Se a empresa tiver uma alta taxa de retenção de clientes, não precisa de investir muito capital para a aquisição de novos clientes;
- Aumento do lucros e redução dos custos decorrentes de falhas (a empresa fornecendo consistentemente produtos e serviços que satisfaçam os clientes, reduzem os custos relacionados com o tratamento de devoluções e a gestão de reclamações);
- Redução da taxa de rotação de pessoal pois a satisfação dos clientes afeta a satisfação dos empregados que passam a querer manter-se na empresa;
- Os custos de atrair novos clientes são mais baixos para empresas que atingem um alto nível de satisfação (dado que os clientes satisfeitos são mais propensos a dar referências positivas WOM);

- Um aumento da satisfação dos clientes leva a um aumento da reputação da própria empresa.

Um cliente satisfeito está convicto que a aquisição de determinado produto ou serviço foi um bom negócio, facilitando o aumento da relação de continuidade entre o cliente e a empresa. A fidelização leva à obtenção de uma posição competitiva no mercado e a possibilidade da obtenção de um lucro superior. A satisfação do cliente é definida como a avaliação total do desempenho baseado em todas as experiências (positivas e negativas) anteriores com a empresa (Jones et al. (2000)).

2.2.2 *Qualidade Percebida*

Na literatura, a qualidade percebida é muitas vezes associada por vários autores à satisfação dos clientes. Segundo Parasuraman et al. (1988), a qualidade é reconhecida sob a forma de atitude, sujeita a uma avaliação contínua e permanente, e a satisfação é referida como uma medida específica de transação.

De acordo com Fogli (2006), a qualidade do serviço é "um julgamento ou uma atitude global em relação a um determinado serviço". Wong and Sohal (2003), tentaram avaliar o impacto das dimensões da qualidade do serviço em dois níveis de relações de retalho, ou seja, nível interpessoal (pessoa a pessoa) e nível de loja (pessoa a empresa). As suas descobertas sugeriram que existe uma associação positiva entre qualidade do serviço e lealdade do cliente.

Assim, a melhor forma de medir a qualidade é através da satisfação dos clientes. Muitos são os autores que se têm dedicado a investigar o conceito de qualidade e a sua ligação à satisfação do cliente e à sua intenção de compra.

Desta forma (Parasuraman et al. (1988)), determinaram que são cinco as dimensões utilizadas pelos clientes para avaliar a qualidade do serviço:

- **Tangibilidade** - associada às instalações, equipamentos e aparência do pessoal;
- **Fiabilidade** - capacidade de prestar o serviço de forma fiável;
- **Capacidade de resposta** - disponibilidade em ajudar os clientes e prestar um bom serviço;
- **Confiança/Segurança** - conhecimento e cortesia dos funcionários, sendo capazes de demonstrar confiança;
- **Empatia** - oferecer aos seus clientes um atendimento cuidado e uma atenção individualizada.

2.2.3 *Valor Percebido*

Com a constante evolução dos mercados, onde surgem constantemente novos produtos, é de extrema importância que as empresas tenham em conta o valor dos seus produtos e/ou serviços, uma vez que, com a constante evolução, estes tem um ciclo de vida cada vez mais curto, provocando uma alteração nos hábitos dos consumidores. Parasuraman (1997) defende que o valor percebido é um dos elementos principais na obtenção de vantagem competitiva por parte de uma organização face às demais.

Segundo Bolton and Drew (1991), o valor percebido resulta da constatação entre os benefícios que um serviço oferece e os custos, monetários ou não, relacionados com a sua utilização. Já Gassenheimer et al. (1998) defendem que o valor percebido é resultado de uma perceção de valor social e económico que se altera de acordo com o nível de relacionamento e que se baseia na satisfação percebida e na comparação entre as alternativas existentes. Outro dos autores que se pronunciou em relação a este tema foi (Zeithaml (1988)), afirmando que o valor percebido é definido como a avaliação global do consumidor sobre a utilidade de um produto, com base em perceções recebidas e pareceres emitidos sobre o produto.

Analisando as anteriores definições apresentadas, verifica-se que o conceito de valor percebido resulta de um balanço feito pelos consumidores entre benefícios e custos da oferta do produto e/ou serviço da empresa.

Desta forma, quando estiver criado valor para os consumidores, estes tendem a recomprar e também referencia-lo positivamente a outros potenciais consumidores (WOM).

2.2.4 *Preço Percebido*

A definição de preço está associada ao preço pago por um cliente pela compra de um produto e/ou serviço.

Segundo Zeithaml (1988), o preço percebido é uma variável que pretende transformar o valor percebido em lucro para a empresa, como tal, é necessário ter em conta que o preço não pode ser definido como o valor pago pelo cliente, mas um resultado de todos os fatores envolventes como os custos monetário e não monetários. Zeithaml (1988), defende também que os clientes fiéis, regra geral, não estão conscientes das pequenas variações de preço entre os fornecedores, nomeadamente em mercados em que existe pouca diferenciação de preços.

Neste tipo de mercados, a variável preço influencia o consumidor quando são feitas grandes promoções para incentivar os clientes que habitualmente fazem compras na concorrência a experimentarem os seus produtos ou serviços. Por outro lado, cria um impacto nos seus habituais clientes, já que estes acabam também eles por usufruir das promoções.

2.2.5 *Confiança*

Como em todas as relações é fundamental criar uma relação de confiança a longo prazo entre o consumidor e o fornecedor, e quanto maior for a confiança do cliente perante a organização mais ele irá demonstrar a sua fidelidade perante o fornecedor (Dwyer et al. (1987); Sirdeshmukh et al. (2002)).

Vários autores analisaram o impacto da confiança na construção de uma relação de fidelização, sublinhando o significado da confiança em explicar a lealdade (Lim and Razzaque (1997); Garbarino and Johnson (1999); Chaudhuri and Holbrook (2001)). Garbarino and Johnson (1999); Chaudhuri and Holbrook (2001) encontraram a confiança como credibilidade para afetar lealdade. Ball et al. (2004), realçam que em caso de mercados com elevada concorrência, a falta de confiança obstruí a formação de fidelidade. Zeithaml et al. (1996); Castañeda (2011); Shainesh (2012), entenderam a confiança como um marcador significativo de lealdade ao cliente.

Resumidamente, pode-se afirmar que a confiança é um antecedente da fidelização e que influência as intenções de compra.

2.2.6 *Compromisso*

O compromisso num relacionamento é um estado psicológico onde um cliente tem planos de continuar a relação com seu fornecedor existente (Moorman et al. (1992); Morgan and Hunt (1994)). Moorman et al. (1992), definiram o compromisso como "uma atitude ou desejo duradouro para uma determinada marca ou empresa. É o grau em que os clientes como membros de uma organização estão emocionalmente conectados a uma organização, marca ou produto, sustentada pelo desejo contínuo de manter a adesão". Já Wong and Sohal (2002), declaram "o compromisso aparenta ser uma das mais importantes variáveis a utilizar para compreender a força de uma relação de marketing, e é um conceito útil para medir a probabilidade da fidelização de um cliente, bem como, para prever a frequência de compra futura".

Autores como (Delgado-Ballester and Luis Munuera-Alemán (2001); Fullerton (2005); Pritchard et al. (1999); Verhoef and Langerak (2002); Wong and Sohal (2002); Zins (2001)), dividem o compromisso em duas componentes:

- **Compromisso afetivo**: traduz o sentimento de identificação pessoal e social do cliente para com o produto ou serviço. Este sentimento influência a resistência à mudança pelo seu elevado custo psicológico.

- **Compromisso de continuidade**: observa-se quando o cliente mantêm a relação com o atual fornecedor por não identificar alternativas viáveis, sendo este um fator que se baseia mais na dependência do que na dedicação.

Fullerton (2005), ao estudar o comportamento dos clientes, identificou que o compromisso afetivo relacionado de forma positiva com o WOM, com a predisposição em pagar mais e está relacionado de forma negativa com a possibilidade do cliente trocar de fornecedor. Por outro lado, o compromisso de continuidade apresentou um relacionamento inverso relativamente aos três aspetos referidos anteriormente.

2.2.7 *Comunicação*

A comunicação é percebida como uma troca de informações (formais e informais) significativas entre um cliente e um fornecedor (Anderson and Narus (1984)). Anderson and Narus (1990); Kotler (2000), definem a comunicação como um diálogo interativo entre a empresa e os seus clientes que deverá ocorrer desde a fase de pré-venda até ao pós-venda. A forma de comunicação com o consumidor tem de ser planeada desde o primeiro contacto, e deve ser feita de forma permanente para fortalecer a confiança, não devendo portanto estar a empresa longos períodos sem contactar os clientes (Moutella (2002); Claycomb and Martin (2001)). Keller (2009) defende que a comunicação deve ser vista pela forma como as empresas tentam comunicar, persuadir e relembrar os clientes, sobre os seus produtos ou serviços que fornecem, realçando que a comunicação é vista como a "voz" da empresa.

Oly Ndubisi and Kok Wah (2005), sugeriram que, no marketing de relacionamento, a comunicação envolve a prestação de informações credíveis, o cumprimento de promessas e a informação em caso de problemas relacionados à entrega.

Para Keller (2009), a comunicação é composta por oito principais modos: publicidade, promoções de vendas, eventos e experiências, relações públicas e publicidade, marketing direto, marketing interativo, *Word of Mouth* e venda pessoal.

2.2.8 *Imagem Organizacional*

A imagem organizacional é de extrema importância, pois uma boa imagem pode impulsionar as vendas através da satisfação e respetiva fidelização do consumidor.

Aaker (1996) definiu a imagem organizacional como "o resultado líquido de todas as experiências, impressões, crenças, sentimentos e conhecimento que as pessoas têm sobre uma empresa". Chun and Davies (2006), no seu estudo, concluíram que na área do retalho, uma boa imagem correlaciona-se positivamente com a satisfação do cliente, o que faz dela um preditor significativo de lealdade. Keller (2009) destaca que a imagem organizacional é definida como a perceção e a preferência do consumidor a uma marca, refletindo-se em diferentes associações mantidas na memória relativamente à marca.

A imagem organizacional é composta por duas componentes principais (Kennedy (1977)):

- **Componente funcional**: composta por características tangíveis e que são mensuráveis;

- **Componente emocional**: caracterizada pelos sentimentos e atitudes associados a uma empresa de acordo com as experiências de cada cliente em relação à empresa.

Para Nguyen and Leblanc (2001), a imagem de uma organização é descrita como a impressão estabelecida na mente do consumidor sobre a empresa, podendo estar relacionada com atributos físicos ou comportamentais da mesma, como o nome, a estrutura organizacional, a variedade de produtos e/ou serviços, a tradição, a ideologia e a sensação de qualidade. Realçam ainda que no momento da escolha por parte do consumidor, o mesmo tem em atenção à imagem que tem predefinida da empresa, que pode ser proveniente de relações passadas, ou até de experiências de outros consumidores, que passaram informações positivas da organização (WOM), ou ainda, através de informações que lhes cheguem pelos de meios de comunicação.

2.2.9 *Barreiras à Mudança de Fornecedor*

Sendo os clientes a peça fundamental das organizações, devido à sua elevada importância em termos financeiros, é pertinente identificar alguns aspetos que as organizações podem fazer para reter os clientes (Fornell (1992); Jones et al. (2000)). Citando (Jones et al. (2000)), as barreiras à mudança de fornecedor "representam qualquer fator que torna mais difícil para os consumidores mudarem de fornecedor".

Caruana (2003), analisando o que leva a uma mudança de fornecedor por parte do cliente, identificou oito motivos: falha no serviço base (26%), falha nos pontos de contacto dos serviços (21%); preço (17%); resposta a uma falha do serviço (11%); inconveniência do serviço (10%); concorrência (4%) e mudança involuntária (2%). Percebendo as razões que levam os clientes a mudar de fornecedor, a empresa deve tentar impedir que os clientes desertem, e para isso as empresas têm de perceber o que pode constituir uma barreira à mudança.

É fundamental perceber-se então quais a barreiras à mudança, para a obtenção de vantagem competitiva face à concorrência (Caruana (2003); Burnham et al. (2003)). Na literatura, são vários os fatores apresentados como barreiras à mudança de fornecedores.

Para Ping Jr (1993); Jones et al. (2000); Burnham et al. (2003); Balabanis et al. (2006); Beatty et al. (1996), as barreiras à mudança, são justificadas por:

- **Custos de mudança**, remete à perceção do consumidor em relação ao tempo, dinheiro e esforço despendido associados à troca de fornecedores. À medida que os custos percebidos aumentam, a probabilidade dos consumidores mudarem de fornecedor diminui, podendo dividir-se estes custos subdividirem-se em:
  - **Procedimentais**, baseado em riscos económicos, custos de avaliação, aprendizagem e de iniciação, envolvendo principalmente o gasto de tempo e esforço;
  - **Financeiros**, que consiste na perda de benefícios financeiros;

  - **Relacionais**, as relações interpessoais entre os clientes e os empregados das organizações aumenta a fidelização do cliente à marca;
  - **Existência de programas de fidelização**;
  - **O grau de monopólio do mercado da organização**, caso não exista uma relação de mercado de concorrência perfeita;
  - **Custos artificiais**.

- **Investimento**, medido através do tempo, energia e dinheiro gastos na relação com o fornecedor;
- **Singularidade do investimento no fornecedor**;
- **Atratividade das alternativas**, perceção que o cliente tem das alternativas existentes no mercado. Quando existem poucas alternativas viáveis, a possibilidade de um cliente desertar é reduzida, aumentando a sua retenção à marca.

Jones et al. (2000), refere ainda que as barreiras à mudança de fornecedor não devem ser consideradas de forma isolada, como uma medida explicativa da fidelização dos clientes.

Os antecedentes mencionados, e em conjunto, levam à formação de fidelização de um cliente com a marca, como está representado na Figura 3.

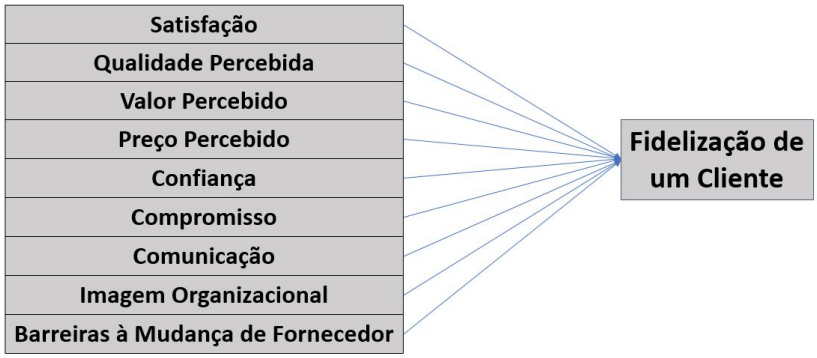

Figura 3: Formação de fidelidade de um cliente

2.3 MEDIDAS PARA O CÁLCULO DA FIDELIZAÇÃO

Várias foram as metodologias propostas ao longo dos anos para o cálculo do envolvimento de um cliente com uma marca. As medidas mais usadas pelas empresas para medir a fidelização dos clientes passa pelo cálculo do *Net Promoter Score* (NPS), *Recency, Frequency and*

Monetary Value (RFM), *Customer Lifetime Value* (CLV) e *Share of Wallet* (SOW). Todos estes modelos já existem na empresa do retalho alimentar onde esta dissertação foi realizada, contudo aquilo que se pretende é a criação de uma métrica única.

2.3.1 *Net Promoter Score (NPS)*

Com base em toda a análise e perceção da importância de fidelizar um cliente a uma marca, uma nova métrica foi criada por Reichheld (2003), para ajudar a perceber e a calcular a fidelização dos clientes a uma marca. Desde então, muitas organizações usam o *Net Promoter Score* (NPS) para medir a fidelização do cliente com a marca. O NPS consiste em avaliar os seus clientes em 3 grupos, os promotores, os passivos e os detratores (Reichheld and Covey (2006)), grupos esses que são obtidos com base na resposta a uma questão: "Qual a probabilidade de recomendar a marca a um amigo ou colega?" usando uma escala de 11 pontos, variando entre 0 (*não provável*) a 10 (*extremamente provável*). Os promotores com uma classificação de 9 ou 10 pontos, são definidos como clientes leais entusiastas que continuam a comprar na empresa e aconselham a amigos e familiares para fazerem compras na empresa. Os clientes passivos apresentam uma classificação de 7 ou 8 pontos, estão satisfeitos com a marca mas não são entusiasmados e que facilmente podem ser perdidos para outras marcas. Por fim, os detratores são clientes infelizes, que não recomendariam a marca e que apresentam valores menores ou iguais a 6 pontos.

O NPS é calculado subtraindo à percentagem de clientes promotores a percentagem de clientes detratores,

$$NPS = Percentagem \quad de \quad Promotores - Percentagem \quad de \quad Detratores \tag{2.3.1}$$

e varia entre -100 e 100.

Contudo, alguns autores não concordam com o uso desta métrica, alegando que fazer um cálculo que não se baseie em dados comportamentais é enganoso, sendo certo que existe a possibilidade de um cliente ser considerado promotor e deixar de fazer compras na marca. No estudo realizado por (Zaki et al. (2016)), verifica-se a falta de confiança no NPS como única medida de fidelidade de clientes em grandes organizações, propondo novos métodos que combinem múltiplas fontes de dados de clientes, incluindo dados demográficos, comportamentais e de atitude.

Umas das criticas feitas ao NPS por parte de Kristensen and Eskildsen (2011) prende-se com o facto da escala padrão do NPS não incluir a possibilidade de ter uma categoria "Sem Resposta", uma vez que, é uma recomendação padrão por parte da bibliografia na área da pesquisa empresarial. Os mesmo autores afirmam ainda que o NPS não é um novo número científico para os negócios e que ganhou popularidade porque é muito simples, pedagógico e fácil de calcular, mas não indica uma real informação sobre a satisfação e fidelização dos

clientes. Outro estudo realizado por (Eskildsen and Kristensen (2011)) onde foi inserida a opção das pessoas escolherem a opção "Sem Resposta", é evidente que a escala padrão das respostas para as mulheres e homens foram muito diferentes quando foram forçados a dar uma resposta. O NPS não é independente do género, ou seja, o género influencia o NPS. Este é um problema grave do NPS, em que é maior o número de homens detratores em proporção aos detratores femininos e maior o número de promotoras femininas proporcionalmente aos promotores do sexo masculino. Isso leva a um NPS artificialmente baixo para os homens o que provoca que se retirem conclusões arbitrárias. Por fim, outra das criticas feitas ao NPS é que este não é exclusivo de clientes que visitaram a loja.

2.3.2 *Recency, Frequency and Monetary Value (RFM)*

O *Recency, Frequency and Monetary Value (RFM)* é um modelo que diferencia a importância dos clientes por três variáveis (Cheng and Chen (2009)):

- Recência da última compra (R): refere-se ao intervalo de tempo entre a última compra e o presente. Quanto menor for o intervalo, menor é R.
- Frequência das compras (F): refere-se ao número de transações num período específico de tempo. Quanto maior o número de transações maior o valor de F.
- Valor monetário das compras (M): refere-se ao montante gasto no período de tempo definido *à priori*. Quanto maior for o valor gasto, maior será o valor de M.

O primeiro modelo paramétrico em que foram usadas variáveis RFM foi proposto por (Hughes (1994))

Embora o modelo RFM seja um modelo adequado para diferenciar os clientes em grandes bases de dados por três variáveis, existem estudos que apresentam diferentes opiniões em relação às três variáveis do modelo de RFM.

Para Hughes (1994) a importância dessas variáveis é configurada através de pesos, considerando que as três variáveis são de igual importância e como tal deverão ter o mesmo peso. Bob (1994) considera que os pesos dessas variáveis devem ser estabelecidos de acordo com as características da empresa.

Contudo, segundo Tichindelean (2013), os modelos de *scoring* baseados em variáveis RFM são utilizados numa fase inicial de aquisição de novos clientes, ao selecionar os clientes certos para futuras ações de marketing. Embora tenha várias vantagens, o modelo de *scoring* de RFM é considerado como tendo as seguintes limitações:

- Esta metodologia identifica o envolvimento dos clientes apenas no período de tempo atual, as expectativas em relação aos períodos futuros não podem ser obtidas;

- As variáveis usadas (RFM) são indicadores observados (efeitos observados) do envolvimento dos clientes, outros fatores não são levados em consideração;

- O modelo ignora a possibilidade de que o envolvimento dos clientes seja resultado de anteriores ações de marketing da empresa;

- O modelo não tem em conta o poder de compra dos clientes (daí a importância do uso de modelos como o *Share of Wallet* que será apresentado na secção 2.3.4).

2.3.3 *Customer Lifetime Value (CLV)*

O *Customer Lifetime Value* (CLV) é um conceito que apresenta diferentes definições na literatura, e como não existe uma definição única, o seu significado varia conforme o contexto onde está a ser desenvolvido e a visão de negócio de quem o desenvolve. Para a empresa de retalho alimentar em estudo nesta dissertação, o *Custumer Lifetime Value* é visto como o valor de vendas líquidas reportadas que o cliente gastará na empresa num dado período de tempo futuro.

As vendas líquidas reportadas são calculadas a partir da subtração dos descontos líquidos às vendas líquidas. As vendas líquidas são da subtração entre as vendas brutas e o valor do IVA praticado numa transação. O desconto líquido é o valor que o cliente acumula em cada compra sem o valor do IVA, podendo ser posteriormente utilizado numa futura transação.

As organizações passaram a contemplar os seus relacionamentos com o cliente como a soma do valor de vendas de todas as transações entre si e estes, durante o tempo em que o cliente permanece com a empresa. Os clientes passaram a ser vistos como ativos, levando à aposta por parte das organizações na sua aquisição e retenção (Singh and Jain (2013)). Pfeifer et al. (2005) defendem que o CLV é o valor de um cliente medido em função dos proveitos futuros que irá gerar no decorrer do relacionamento com a empresa e consiste essencialmente numa métrica de marketing que projeta o valor do cliente ao longo do tempo que o cliente realiza transações com a organização. Por outro lado existem definições que tem por base exatamente o lucro que o cliente dá à empresa, (Glady et al. (2009)).

Apesar das mais variadas definições de CLV, todas as abordagens compartilham a mesma essência, o valor presente dos rendimentos futuros, trazidos pelo cliente, aos quais se subtrai os custos correspondentes, no decorrer de um período de tempo, no qual o cliente realiza transações com a organização (Madeira (2014)).

Contudo, o uso do CLV também apresenta algumas desvantagens:

- O momento em que o cliente abandona é baseado em pressupostos. O tempo entre a compra e o gasto em cada compra é muito variável o que dificulta a previsão do seu abandono (Castro (2017));

- O método trabalha essencialmente com informações extraídas da base de transações de consumidores, não considerando portanto elementos que dizem respeito às motivações que levam às decisões de compra (Venkatesan and Kumar (2004));

- Os dados referentes à taxa de retenção de um serviço pode estar associado a motivações que não a efetiva fidelização dos clientes, entre elas a ausência de alternativas, a falta de informação e/ou o elevado custo da troca (Gupta et al. (2006)).

2.3.4 *Share of Wallet (SOW)*

As empresas gastam muito tempo e dinheiro com o objetivo de perceber e tentar aumentar a fidelidade do cliente, medindo e criando métricas, como a satisfação e o *Net Promoter Score*. Mas as medidas tradicionais de lealdade correlacionam-se mal com o que mais importa, o *Share of Wallet* (SOW). O SOW corresponde a uma percentagem dos gastos de um cliente numa categoria que é captada por uma determinada marca, loja ou empresa. Os clientes podem estar muito satisfeitos com uma marca e recomendá-la a outras pessoas, mas se eles gostam dos concorrentes tanto ou mais, a empresa está a perder vendas. Fazer mudanças para aumentar a satisfação não será necessariamente suficiente. Isso não significa que as métricas tradicionais não sejam valiosas, até porque é muito útil saber os clientes estão satisfeitos e se eles recomendam a marca aos seus amigos e colegas, mas essas medidas em si não dizem como é que os clientes irão dividir os seus gastos entre a empresa e a concorrência (Timothy et al. (2011)).

A primeira definição de SOW na literatura é a de Keiningham et al. (2003): "A percentagem do volume total de negócios realizado com a empresa por uma organização de clientes dentro de um período de 12 meses". Novas definições foram emergindo, como (Cooil et al. (2007)), que propõem: "O *Share of Wallet* é a percentagem de dinheiro que um cliente atribui em uma categoria atribuída a uma empresa específica". Todas essas definições, no entanto, carecem de precisão. Letaifa and Perrien (2008) propuseram uma definição mais completa de SOW, que leva em consideração o fator tempo, a continuidade do conceito ao longo do tempo e os tipos de fatores que influenciam essa continuidade. Segundo os autores, "O SOW é a proporção de ativos ou negócios investidos por uma cliente numa determinada empresa ou marca (em percentagem do total de negócios ou ativos investidos pelo cliente em uma determinada indústria), ao longo de um determinado tempo. Essa proporção pode mudar ao longo do tempo devido a fatores pessoais e/ou situacionais".

O SOW é indiscutivelmente o indicador mais importante da lealdade de um cliente, sendo resultado de experiências de clientes consistentemente positivas. Na verdade, as empresas planeiam competir principalmente com base na experiência do cliente, explicando a obsessão com a satisfação do cliente e com os níveis de NPS. Contudo, o NPS está tão fracamente correlacionada com o SOW que os clientes atribuem às marcas que usam. Não é a sua

pontuação que é importante, mas sim a classificação, isto é, a forma como os clientes veem uma marca em relação à concorrência. Acontece que, sabendo a classificação da sua marca e o número de marcas concorrentes que um cliente também usa, os gerentes conseguem prever com maior precisão o SOW que os clientes usam na empresa, recorrendo à *Wallet Allocation Rules* (Regras de Alocação da Carteira):

$$SOW = \left(1 - \frac{ranking}{número \quad de \quad marcas + 1}\right) * \left(\frac{2}{número \quad de \quad marcas}\right) \tag{2.3.2}$$

onde, o *ranking* corresponde à posição relativa que um cliente atribui a uma marca em comparação com outras marcas também utilizadas pelo cliente na mesma categoria. O número de marcas corresponde ao número total de marcas usadas pelo cliente numa categoria (Keiningham et al. (2015)).

A satisfação e a lealdade do cliente não são suficientes. Os clientes têm uma razão lógica para usarem as marcas que usam em cada uma das categorias. Como tal, a chave para aumentar o SOW, é dar aos clientes menos motivos para estes usarem a concorrência

Também o SOW, como as métricas enumeradas anteriormente, apresenta as suas desvantagens, tais como:

- É um indicador do passado, não permite uma visão futura do cliente, medindo apenas o comportamento num único ponto no tempo;
- É transacional, como tal, não mede a mentalidade do cliente;
- É insensível ao mercado, e tendo os vários mercados diferentes níveis de concorrência, o peso do SOW também deveria ser diferente para cada um dos diferentes mercados.

2.4 CONCLUSÃO

Este capítulo teve como principal objetivo perceber e definir o conceito de fidelização, bem como a sua evolução ao longo dos anos. Inicialmente, a fidelização estava focada apenas nas repetições de compra de um determinado produto ou serviço. No entanto, com o passar do tempo e com as várias evoluções que a sua definição foi sofrendo, percebeu-se que era preciso estudar os comportamentos dos clientes, os seus antecedentes e a real importância da fidelização. A lealdade de um cliente não só garante compras repetidas e publicidade positiva (WOM), como um maior valor em termos de confiança por parte do cliente, o que origina outros benefícios essenciais para o crescimento de um empresa como a preferência exclusiva e prioritária dos seus produtos ou serviços, maior SOW entre outros.

Em suma, sendo este um conceito relativamente recente, tem sofrido grandes evoluções nos últimos tempos, não existe nenhuma definição que reúna total consenso na literatura, tendo várias definições dependendo do contexto no qual está inserido. Na literatura existem vários

métodos propostos sendo que a sua aplicação vai de encontro com a definição que cada negócio dá ao tema.

3

METODOLOGIA

A análise de regressão é um método conceptualmente simples para investigar relações funcionais entre variáveis. A relação é expressa na forma de uma equação ou um modelo que associa a variável resposta (ou dependente) e uma ou mais variáveis explicativas (ou independentes ou covariáveis). Denota-se a variável resposta por Y e as variáveis explicativas por $X_1, X_2, ..., X_p$ onde p é o número de variáveis explicativas. A relação entre Y e $X_1, X_2, ..., X_p$ pode ser expressa pelo modelo de regressão linear

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon, \tag{3.0.1}$$

onde $\beta = (\beta_0, ..., \beta_p)^T$ é o vetor de parâmetros de regressão e ε é o vetor dos erros aleatórios.

Durante muitos anos, os modelos de regressão linear clássicos foram muito utilizados. No entanto, estes modelos apresentam algumas restrições:

- a relação entre as variáveis (variável resposta e as variáveis explicativas) é descrita por uma função linear de parâmetros.
- condicional aos valores das variáveis explicativas, as respostas são independentes e seguem uma distribuição normal.

3.1 MODELOS LINEARES GENERALIZADOS

Embora vários modelos não lineares ou não normais tenham entretanto sido desenvolvidos para fazer face a situações que não eram adequadamente explicadas do modelo de regressão normal, Nelder and Wedderburn (1972) introduziram os Modelos Lineares Generalizados (MLG). Estes modelos são caraterizados como sendo uma extensão do modelo de regressão linear, onde a distribuição da variável resposta condicionada à variável explicativa pertencente à família exponencial e a função que relaciona o valor esperado e o vetor de variáveis explicativas pode ser qualquer função diferenciável. São vários os modelos lineares generalizados, destacando-se os modelos: Regressão Logística, Regressão Poisson e Regressão Binomial Negativa.

Neste capítulo, o modelo de Regressão Logístico será introduzido.

3.1.1 *Família Exponencial*

Como referido anteriormente, os MLG pressupõem que a variável resposta tenha uma distribuição pertencente a uma família particular, a família exponencial.

Definição (Família Exponencial):

A variável aleatória Y tem uma distribuição pertencente à família exponencial de dispersão se a sua função de densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) possa ser escrita na seguinte forma:

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{3.1.1}$$

onde θ é a forma canónica do parâmetro de localização e ϕ é um parâmetro de dispersão suposto, em geral, conhecido. $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções específicas para cada distribuição (Turkman and Silva (2000)).

Neste caso a distribuição descrita em 3.1.1 faz parte da família exponencial uniparamétrica.

Segundo McCullagh and Nelder (1989), se Y é uma variável aleatória com distribuição pertencente à família exponencial, em concordância com as condições anteriores, então:

$$\begin{aligned} &E(y) = \mu = b'(\theta), \\ &Var(Y) = \sigma^2 = b''(\theta)a(\phi) \end{aligned} \tag{3.1.2}$$

O valor médio surge como uma função de θ e a variância como um produto de duas funções que dependem do parâmetro de localização e do parâmetro de dispersão, respetivamente.

A família exponencial abrange inúmeras distribuições, por exemplo distribuições discretas tais como a distribuição Binomial e a distribuição de Poisson ou distribuições contínuas como a distribuição Normal ou distribuição Inversa Gaussiana.

3.1.2 *Descrição do Modelo Linear Generalizado*

Os modelos lineares generalizados são representados na forma matricial por

$$Y = Z\beta + \varepsilon,$$

onde Z é uma matriz de dimensão $n \times (p+1)$ de especificação do modelo (em geral a matriz de covariáveis X com um primeiro vetor unitário), associada a um vetor $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta p)^T$ de parâmetros, e ε é um vetor de erros aleatórios com distribuição que se supõe $N_n(0, \sigma^2 I)$.

Estas hipóteses implicam que $E(Y|Z) = \mu$ com $\mu = Z\beta$, ou seja, o valor esperado da variável resposta é uma função linear das covariáveis.

Os Modelos Lineares Generalizados assentam sobre três componentes fundamentais:

- **Componente aleatória** - a variável resposta Y corresponde a uma variável aleatória do qual se recolhem n observações independentes e cuja distribuição de probabilidades faz parte da família exponencial de distribuições e tem-se

$$E[y_i|x_i] = \mu_i] = b^{'}(\theta_i), \quad i = 1, ..., n$$

- **Componente estrutural ou sistemática** - consiste numa combinação linear de variáveis explicativas.

  O valor esperado μ_i está relacionado com o preditor linear

$$\eta_i = z_i^T \beta$$

  através da relação

$$\mu_i = h(\eta_i) = h(z_i^T \beta)$$

  onde h é uma função monótona e diferencial e

$$g = h^{-1}$$

  é a função de ligação que relaciona a média de y_i ao preditor linear, ou seja,

$$g(\mu_i) = \eta_i, \quad i = 1, ..., n$$

- **Função de ligação** - como o nome sugere, esta função estabelece uma ligação entre a componente aleatória e a componente sistemática, isto é, relaciona o valor esperado $E(Y) = \mu$ com o preditor linear η. No caso dos MLG, tal ligação é estabelecida através de uma função $g(\mu)$ monótona e diferenciável, designada por uma função de ligação:

$$g(\mu) = \mathbf{Z}\beta \tag{3.1.3}$$

  A função ligação é utilizada com o objetivo de efetuar uma transformação do valor esperado da variável resposta, para uma escala em que este não fique restrito. A função de ligação mais simples é obtida quando $g(\mu) = \mu$, isto é, a função identidade como é o caso do modelo de regressão linear clássico. No caso da regressão logística a função de ligação mais usual denomina-se de função *logit*, como será explicitado posteriormente.

3.2 MODELO DE REGRESSÃO LOGÍSTICA

O modelo de regressão Logística é um modelo linear generalizado, aplicado quando se pretende modelar uma variável resposta categórica, dado um conjunto de variáveis explicativas (Agresti (2002)).

Analogamente à regressão linear, os modelos dizem-se de regressão logística simples envolvem apenas uma variável explicativa e de regressão logística múltipla quando estão presentes no modelo mais do que uma variável explicativa. Para além dessas designações também se dividem estes modelos em regressão logística *dicotómica* (a variável resposta é do tipo binária) e *politómica* (quando a variável resposta possui mais do que duas categorias), podendo ser nominal ou ordinal. Designa-se de regressão logística ordinal quando existe uma ordem subjacente entre as categorias da variável resposta e regressão logística nominal quando no contexto do problema não existe esta ordem. Relativamente às variáveis explicativas, estas podem ser quantitativas ou qualitativas.

No estudo será apenas tratado os modelos de regressão logística binária.

Seja $Y \in \{0, 1\}$ a variável resposta binária, onde 1 é traduzido como o "sucesso" do evento em estudo e 0 é traduzido como o "insucesso". Seja $Y_1, ..., Y_n$ uma amostra aleatória, $X = (X_1, ..., X_p)$ um vetor das p covariáveis, e uma determinada observação $\mathbf{x}_i = (x_{1i}, ..., x_{pi})$ do objeto i, $i = 1, ..., n$, assume-se que:

$$Y|(X = \mathbf{x}_i) \sim Bin(1, \pi_i) \tag{3.2.1}$$

sendo $\pi_i = \pi(\mathbf{x}_i) = P(Y = 1|X = \mathbf{x}_i)$ a probabilidade de sucesso dado que se observou $X = \mathbf{x}_i$ (Agresti (1996)). Analogamente, a probabilidade de insucesso é definida como sendo complementar da anterior, $1 - \pi_i = P(Y = 0|X = \mathbf{x}_i)$. É possível verificar, pelas propriedades do valor esperado e da variância (equação 3.1.2), que:

$$\begin{aligned} \mu_i &= E(Y|X = \mathbf{x}_i) = \pi_i, \\ \sigma^2 &= Var(Y_i|X = \mathbf{x}_i) = \pi_i(1 - \pi_i), \end{aligned} \tag{3.2.2}$$

observando-se que o valor esperado e a variância dependem da probabilidade π_i (Rodrıguez (2007)). Isto sugere que modelos onde se assume variância constante, como o caso dos modelos lineares, não são apropriados para modelar dados binários e que o valor esperado da variável resposta varie no intervalo de $[0, 1]$.

Em problemas de regressão, pretende-se modelar $E(Y|X = \mathbf{x}_i)$, isto é, o valor esperado da variável resposta dada determinada observação $\mathbf{x}_i$ (Hosmer and Lemeshow (2000)).

Neste seguimento, e de acordo com a estrutura dos MLG descritos na Secção 3.1 torna-se necessária a aplicação de uma função de ligação $g(\mu_i)$ apropriada a este tipo de dados

que seja capaz de relacionar o valor esperado da variável resposta, π_i, com o preditor linear $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}, i = 1, ..., n$. Pretendemos então que:

$$\eta_i = g(\mu_i) = g(\pi_i), \tag{3.2.3}$$

No caso da regressão logística, a função de ligação mais comum é a função *logit* (Hosmer and Lemeshow (2000)), definida por

$$g(\pi_i) = logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right), \tag{3.2.4}$$

e obtemos

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}, \quad i = 1, ..., n \tag{3.2.5}$$

Note-se que o logaritmo faz com que o termo do lado esquerdo da equação varie entre $-\infty$ e $+\infty$ como o outro membro.

Assim, o modelo de regressão logística é definido por:

$$\begin{aligned} &Y|(X = \mathbf{x}_i) \sim Bin(1, \pi_i) \\ &logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}, \quad i = 1, ..., n \end{aligned} \tag{3.2.6}$$

Adicionalmente, resolvendo a equação 3.2.5 em ordem a π_i, obtém-se

$$\begin{aligned} log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} &\Leftrightarrow \frac{\pi_i}{1-\pi_i} = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}} \\ &\Leftrightarrow \pi_i = e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}} (1 - \pi_i) \\ &\Leftrightarrow \pi_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}}}, \quad i = 1, ...n, \end{aligned} \tag{3.2.7}$$

A regressão logística permite assim estimar a probabilidade de ocorrência de um evento $\pi_i = P(Y = 1|X = x_i)$, tendo em conta as diversas covariáveis. Para determinar o valor da variável resposta Y_i (0 ou 1) associado a um determinado objeto, é usual assumir-se um valor de corte para π_i, por exemplo $0,5$. Assim sendo, assume-se que se $\pi_i \geq 0.5 \rightarrow Y = 1$ (o objeto pertence à classe definida como o "sucesso" do evento), caso contrário, $Y = 0$. Um valor de corte ótimo poderá ser estimado através dos dados, por forma a melhorar o ajuste do modelo de regressão logística aos dados.

Por fim, é importante realçar que o cálculo de π_i não está restrito ao tipo de variáveis explicativas. Como referido no inicio desta secção, as variáveis explicativas, $X_i, ..., X_p$ podem ser de natureza quantitativa ou qualitativa. Caso as variáveis sejam explicativas, são então

transformadas nas denominadas variáveis "*dummys*" (tomando apenas dois valores possíveis 0 e 1), onde para k valores possíveis de uma variável categórica são criadas $k-1$ variáveis *dummys*, e considerando uma das categorias da variável como referência.

Nestes casos, a equação 3.2.5 pode ser reescrita como:

$$logit(\pi_i) = \beta_0 + \beta_1 x_{1i} + ... + \sum_{l=1}^{k_j-1} \beta_{jl} x_{jli} + \beta_p x_{pi} \tag{3.2.8}$$

considerando-se a j-ésima variável categórica com K categorias, x_{jli} representa cada variável *dummy* criada para o objeto i e β_{jl} os respetivos coeficientes.

3.2.1 *Estimação dos coeficientes do modelo*

Os parâmetros dos modelos lineares generalizados podem ser estimados aplicando o método de máxima verosimilhança, à semelhança dos restantes MLG (Agresti (1996)). As estimativas de máxima verosimilhança $\widehat{\beta}$, são obtidas maximizando a função de verosimilhança. Assumindo independência nas observações, esta função corresponde a considerar o produtório de uma função de probabilidade,

$$l(\beta, y) = \prod_{i=1}^{n} f(y|x_i) \tag{3.2.9}$$

O valor que maximiza $l(\beta, y)$ é o que maximiza a função obtida pelo seu logarítmo, a função de log-verosimilhança, $L(\beta, y)$, uma vez que a função logaritmo é uma função monótona. A função de log-verosimilhança é usualmente, mais adequada para calcular o máximo, dado que permite escrever a verosimilhança como uma soma de parcelas ao invés de uma multiplicação de parcelas e, consequentemente, simplificar o cálculo da derivada.

$$L(\beta, y) = log\left(\prod_{i=1}^{n} f(y|x_i)\right) = \sum_{i=1}^{n} log(f(y|x_i)) \tag{3.2.10}$$

Após a obtenção da função log-verosimilhança, pretende-se então calcular os seus máximos. O máximo da função é o zero da primeira derivada, sem haver necessidade de avaliar o sinal da segunda derivada, uma vez que se prova que esta tem sempre sinal negativo (Casella and Berger (2002)). Neste sentido, calculam-se as $p+1$ equações de verosimilhança correspondentes aos $p+1$ coeficientes do modelo (p coeficientes associados às p variáveis explicativas mais o parâmetro constante), que correspondem a derivar a equação (3.2.10) em ordem a cada β_j e igualar a zero.

Relembrando, a componente aleatória de um modelo de regressão logística para resposta binária:

$$Y|(X = x_i) \sim Bin(1, \pi_i), \tag{3.2.11}$$

e, por definição, a função de probabilidade condicionada é:

$$f(y|x_i) = \pi(\mathbf{x}_i)^{y_i}(1 - \pi(\mathbf{x}_i))^{1-y_i} \tag{3.2.12}$$

Neste caso, a contribuição para a função de probabilidade é $\pi(x_i)$ quando $y_i = 1$ e $1 - \pi(x_i)$ caso contrário. Portanto, de acordo com a equação (3.2.10), a função log-verosimilhança pode ser escrita como:

$$L(\beta, y) = \sum_{i=1}^{n} \{y_i log(\pi(\mathbf{x}_i)) + (1 - y_i) log(1 - \pi(\mathbf{x}_i))\} \tag{3.2.13}$$

As equações de verosimilhança assumem a seguinte forma:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0 \quad para \quad j = 0 \tag{3.2.14}$$

e

$$\sum_{i=1}^{n} x_i [y_i - \pi(\mathbf{x}_i)] = 0 \quad para \quad j = 1, ..., p \tag{3.2.15}$$

Em geral, não é possível resolver-se as equações (3.2.14) e (3.2.15) de forma explícita. Nesses casos, para obtermos as estimativas de $\boldsymbol{\beta}$, é necessário recorrer-se a métodos numéricos.

3.2.2 *Avaliação da qualidade do modelo*

Depois de se obter as estimativas dos parâmetros do modelo, é essencial avaliar a sua qualidade no que diz respeito, em particular, a:

- Significância estatística das variáveis do modelo;
- Qualidade de ajustamento do modelo aos dados;
- Proporção de variância explicada pelo modelo.

Significância Estatística das Variáveis

O *teste de Wald* permite avaliar individualmente a significância estatística de cada coeficiente do modelo. As hipóteses a testar são:

$$H_0 : \beta_j = 0, \quad versus \quad H_1 : \beta_j \neq 0, \quad j = 0, ...p, \tag{3.2.16}$$

A estatística de teste, e a respetiva distribuição, sob a hipótese H_0 é calculada por

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim N(0,1), \tag{3.2.17}$$

onde $\hat{\beta}_j$ é a estimativa de máxima verosimilhança de β_j e $SE(\hat{\beta}_j)$ representa o erro padrão de $\hat{\beta}_j$. A hipótese H_0 é rejeitada se o valor observado de W for superior a $z_{1-\alpha/2}$, onde $z_{1-\alpha/2}$ representa o quantil $(1-\alpha/2)$ da distribuição $N(0,1)$

Outra forma de se inferir sobre um parâmetro populacional é com base nos intervalos de confiança (IC). O intervalo de $100(1-\alpha)\%$ de confiança para um $\beta_j, j = 1, ..., p$ é dado por:

$$\left(\widehat{\beta_j} - z_{1-\alpha/2}SE(\hat{\beta}_j); \widehat{\beta_j} + z_{1-\alpha/2}SE(\hat{\beta}_j)\right), \tag{3.2.18}$$

onde $z_{1-\alpha/2}$ representa o $(1-\alpha/2)$ quantil da distribuição $N(0,1)$. Com $100(1-\alpha)\%$ de confiança, este intervalo conterá o valor do parâmetro β_j. Neste caso, um IC para β_j que não contenha o valor 0 indica que β_j é estatisticamente diferente de zero com $100(1-\alpha)\%$ de confiança e portanto a variável a ele associada é significativa no modelo.

Qualidade de Ajustamento do Modelo

Antes de analisarmos um determinado modelo, m, com funções de verosimilhança e log-verosimilhança dadas por l_m e L_m, respetivamente, é necessário considerar a existência de outros dois modelos recorrentemente utilizados em comparações (Lindsey (2000)):

- Modelo nulo: é constituído apenas pelo termo constante β_0, não contendo nenhuma variável explicativa, apresentando o menor valor da função de verosimilhança, l_0. A função de log-verosimilhança é definida por L_0.

- Modelo saturado: representa exatamente a amostra, uma vez que é estimado um parâmetro para cada observação. Este modelo é o que tem o maior valor da função de verosimilhança, l_s. A sua função de log-verosimilhança assume-se como L_s.

Um dos testes de ajustamento mais usados em MLG é denominado teste razão de verosimilhança que utiliza a estatística de teste *Deviance*, permitindo comparar a discrepância entre os valores observados e previstos.

A *Deviance*, utiliza o modelo saturado, s, para avaliar a qualidade de ajustamento do modelo, m, com base nas funções de versomilhança:

$$D = -2ln\left(\frac{l_m}{l_s}\right) = -2(L_m - L_s) \tag{3.2.19}$$

onde l_m e l_s representam as funções de verosimilhança do modelo em estudo e do modelo saturado respetivamente e L_m e L_s as funções de log-verosimilhança do modelo em estudo e do modelo saturado.

A estatística D assume sempre um valor superior ou igual a zero. Um modelo é tanto melhor quanto menor for o valor de D, sendo que $D = 0$, traduz um modelo com um ajustamento perfeito aos dados, como é o caso do modelo saturado.

Sob a hipótese nula:

H_0: O ajustamento do modelo (m) é igual ao ajustamento do modelo saturado (s)

H_1: O ajustamento do modelo (m) nãp é igual ao ajustamento do modelo saturado (s)

e tendo em conta a estatística de teste $D \sim X^2_{1-\alpha}(J-(p+1))$, onde J representa o número de padrões de covariáveis distintas, p o número de covariáveis do modelo ajustado m e χ^2 a distribuição de Qui-Quadrado. A hipótese nula é rejeitada com um nível de significância α se $D > X^2_{1-\alpha}(J-(p+1))$.

Para os modelos de regressão logística, um dos testes mais utilizados para avaliar a qualidade de ajustamento é o teste de *Hosmer e Lameshow* (Hosmer and Lemeshow (2000)).

Sob a hipótese nula:

H_0: O modelo ajusta-se aos dados

H_1: O modelo não se ajusta aos dados

Num modelo de regressão logística com p variáveis explicativas e n indivíduos, podem existir indivíduos diferentes que apresentem os mesmos valores para o conjunto das p variáveis.

Neste caso, denota-se por J o número de valores diferentes de $\mathbf{x}$, sendo $\mathbf{x} = (x_1, x_2, ..., x_p)$, e por $m_j, j = 1, 2, ..., J$, o número de indivíduos com $\mathbf{x} = \mathbf{x}_j$.

Hosmer and Lemeshow (2000) propuseram uma estatística da qualidade de ajuste de um modelo de regressão logística que pressupõe que os dados sejam agrupados em k grupos segundo as probabilidades estimadas, nomeadamente considerando os seus percentis ou decis. Para cada grupo k, denota-se por n_k o número de indivíduos e por c_k o número de valores diferentes do conjunto das p variáveis explicativas. A soma dos valores da variável resposta e a média das probabilidades estimadas para o grupo k denotam-se por o_k e $\bar{\pi}_k$, com $o_k = \sum_{j=1}^{c_k} y_i$ e $\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k}$. A estatística de *Hosmer-Lemeshow*, C, tem uma distribuição aproximada de um χ^2_{k-2} sob a hipótese do modelo ser adequado. A hipótese nula deve ser rejeitada para valores elevados da estatística de teste, a qual tem a seguinte formulação:

$$C = \sum_{k=1}^{g} \frac{o_k - n_k \bar{\pi}_k}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Destaque-se que sendo a estatística $\widehat{C}$ correspondente à estatística do qui-quadrado usual, tem limitações semelhantes. Segundo Hosmer and Lemeshow (2000), as frequências esperadas em cada grupo deverão ser pelo menos 5. A obtenção dos g grupos considerados, pressupõe algumas etapas a descrever de seguida:

- as n probabilidades de sucesso para cada indivíduo $(P(Y_i = 1))$ são ordenadas crescentemente

- as probabilidades são agora agrupadas em g grupos com o mesmo número de observações (g/n), divididos por decis. Hosmer and Lemeshow (2000) aconselham 10 grupos, sendo que cada grupo representa um decil.

- obtém-se uma tabela de contingência $g \times 2$ que consta do cruzamento entre a resposta binária e os g grupos. Esta estatística, $\widehat{C}$ corresponde à estatística usual do X^2 de Pearson que permite comparar a frequência dos valores observados com os valores esperados.

Proporção de variância explicada pelo modelo

No modelo de regressão linear clássica o coeficiente de determinação R^2 representa a proporção de variância da variável resposta explicada pelas variáveis explicativas. Um valor de R^2 próximo de 100% indica que o modelo é capaz de explicar bem a variabilidade dos dados.

Têm sido propostas algumas medidas, denominadas como *pseudo* - R^2 que visam ser uma medida correspondente de R^2 para a regressão logística. Estes *pseudo* - R^2 baseiam-se na comparação da função log-verosimilhança do modelo nula com a do modelo em estudo (Hu et al. (2006)). Uma medida muito usual é o *Cox & Snell* R^2:

$$Cox\ \&\ Snell\ R^2 = 1 - exp\left(\frac{-2(L_m - L_0)}{n}\right) \tag{3.2.20}$$

onde L_m e L_0 são as funções de log-verosimilhança do modelo em estudo e do modelo nulo, respetivamente e n é a dimensão da amostra. Este coeficiente apresenta a limitação de nunca atingir o valor 1, o que dificulta a sua interpretação. Assim, surge o *Nagelkerke's* R^2 como uma normalização do *Cox & Snell* R^2 com a vantagem de variar entre 0 e 1.

$$Nagelkerke's\ R^2 = \frac{Cox\ \&\ Snell\ R^2}{1 - exp\left(\frac{2L_0}{n}\right)} \tag{3.2.21}$$

3.2.2.1 *Métodos de seleção de variáveis*

Existem três métodos de seleção de variáveis: *backward, forward e setwpise.* Estes métodos baseiam-se nos p-valor obtidos através de testes de seleção de modelos, em que cada passo se avalia a inclusão ou exclusão de variáveis. Os métodos são caracterizados por:

- ***Backward* (seleção regressiva, eliminação sucessiva)**: - parte do modelo com todas as variáveis explicativas, isto é, modelo completo, testando-se a remoção de cada variável. As variáveis vão sendo retiradas sucessivamente do modelo até que não existam mais variáveis cuja remoção do modelo produza alterações significativas ao mesmo. Neste método, qualquer variável que seja excluída do modelo não pode voltar a ser adicionada.

- ***Forward* (seleção progressiva)**: - parte do modelo nulo, isto é, modelo sem variáveis explicativas, testa-se a adição de cada variável explicativa ao modelo. Começa-se por

incluir a variável mais importante e sucessivamente inclui-se as variáveis, até que não existam mais variáveis cuja adição ao modelo produza alterações significativas ao mesmo. Neste método, qualquer variável que seja adicionada ao modelo manter-se-á até ao modelo final.

- ***Stepwise* (*both*, passo-a-passo)**: - é uma combinação dos dois métodos anteriores, que testa em cada passo as variáveis que devem ser incluídas ou excluídas, partindo do modelo nulo. O processo é repetido até não existir nenhuma variável cuja adição ou remoção do modelo produza alterações significativas ao mesmo, isto é, até não existirem variáveis que "passem" os critérios de inclusão e exclusão.

Nestes métodos, é necessário definir à partida os critérios de inclusão e/ou exclusão de variáveis. Estes critérios são definidos através dos valores-p associados a um determinado teste à nulidade do coeficiente de cada variável.

3.2.2.2 *Interpretação do modelo*

Uma vez ajustado o modelo e após avaliar a significância dos coeficientes estimados, é agora necessário interpretar os seus valores.

Pretende-se agora apresentar a interpretação do modelo e, em particular dos seus coeficientes. No modelo de regressão logística

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.2.22}$$

cada parâmetro β_j é interpretado como uma mudança provocada em $logit(\pi_i)$ devido ao aumento de uma unidade na j-ésima variável explicativa, X_j, quando as restantes variáveis se mantêm constantes (Hosmer and Lemeshow (2000); Faraway (2006)). Para as variáveis explicativas categóricas, "uma unidade" refere-se à categoria em estudo quando comparada com a de referência.

É habitual extrair informação das *odds* e das *odds ratio*. As *odds* correspondem ao quociente entre a probabilidade de "sucesso" e a probabilidade de "insucesso", isto é:

$$odds = \frac{P(Y=1)}{1-P(Y=1)} = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi_i}{1-\pi_i} \tag{3.2.23}$$

Odds ratio (OR) é quociente entre duas *odds*, permitindo comparar as probabilidades de sucesso/insucesso para dois grupos (G_1, G_2) distintos:

$$OR = \frac{odds(G_1)}{odds(G_2)} = \frac{\frac{P(Y=1|G_1)}{1-P(Y=1|G_1)}}{\frac{P(Y=1|G_2)}{1-P(Y=1|G_2)}} \tag{3.2.24}$$

É importante frisar que os grupos em comparação (G_1, G_2), são muitas vezes complementares, nomeadamente quando se comparam grupos de indivíduos com e sem uma determinada característica. No entanto, não é limitativo, pois permitem que a comparação seja sempre efetuada com uma classe definida como referência. Recorrendo aos *odds ratio* conseguimos obter o quão mais provável é acontecer o evento num determinado grupo em comparação com o outro.

É possível ter os seguintes casos:

- Se $OR > 1$, o "sucesso" é mais provável no grupo G_1;
- Se $OR = 1$, o "sucesso" é igualmente provável nos dois grupos;
- Se $OR < 1$, o "sucesso" é mais provável no grupo G_2.

Os OR são obtidos através dos parâmetros do modelo estimado na amostra. Para interpretar as estimativas associadas aos coeficientes do modelo, é conveniente proceder à análise da natureza das variáveis explicativas.

Variável explicativa dicotómica

Considerando o vetor de variáveis explicativas $X = (X_1, ..., X_v, ...X_p)$ e X_v como uma variável dicotómica codificada por duas categorias, a e b. Temos então que:

$$\begin{aligned} log[\widehat{OR}] &= log[\widehat{OR}(X_v = a, X_v = b)] \\ &= log\left[\frac{\frac{\widehat{\pi}(X_v=a)}{1-\widehat{\pi}(X_v=a)}}{\frac{\widehat{\pi}(X_v=b)}{1-\widehat{\pi}(X_v=b)}}\right] \\ &= logit\left[X_v = a\right] - logit\left[X_v = b\right] \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + ... + \widehat{\beta}_{v-1} X_{v-1} + \widehat{\beta}_v \times a + \widehat{\beta}_{v+1} X_{v+1} + ... + \widehat{\beta}_p X_p \\ &- \widehat{\beta}_0 - \widehat{\beta}_1 X_1 - ... - \widehat{\beta}_{v-1} X_{v-1} - \widehat{\beta}_v \times b - \widehat{\beta}_{v+1} X_{v+1} - ... - \widehat{\beta}_p X_p \\ &= \widehat{\beta}_v \times (a - b) \end{aligned} \tag{3.2.25}$$

Conclui-se então que $\widehat{OR} = e^{\widehat{\beta}_v \times (a-b)}$. Note-se que a maioria das variáveis dicotómicas são codificadas em 1 e 0 nos *softwares* estatísticos. Assim, para este caso ($a = 1$ e $b = 0$) temos $\widehat{OR} = e^{\widehat{\beta}_v}$. Em variáveis de natureza dicotómica, e assumindo agora que a variável apresenta codificação 1 e 0, os $\widehat{OR}$ são interpretados como: é mais/menos provável e $e^{\widehat{\beta}_v}$ vezes a ocorrência do evento para os objetos com $X_v = 1$ em comparação com os que assumem $X_v = 0$, quando as restantes variáveis se mantêm constantes.

Variável explicativa categórica com k categorias

Quando a variável X_v é categórica com k categorias, esta é transformada em $k - 1$ *dummys* e o processo é análogo ao da variável dicotómica. Esta codificação permite a comparação

de categorias duas a duas, onde cada categoria é sempre comparada com uma categoria de referência.

Variável explicativa contínua

Assumindo-se como contínua a variável X_v, a interpretação do respetivo $\widehat{OR}$ varia de acordo com as unidades da variável. Contudo, esta abordagem de uma unidade nem sempre é a mais correta, e, como tal, consideremos que a variável X_v aumenta de c unidades, ou seja, $X = (X_1, ..., X_1 + c, ..., X_p)$. Temos então que:

$$\begin{aligned} log[\widehat{OR}(X_v + c, X_v)] &= log\left[\frac{odds(X_v + c)}{odds(X_v)}\right] \\ &= logit[X_v + c] - logit[X_v] = (X_v + c)\beta_v - \beta_v X_v \\ &= c\beta_v \end{aligned} \tag{3.2.26}$$

Conclui-se que $\widehat{OR} = e^{c\beta_v}$, traduzindo o OR para o "sucesso" por aumento de c unidades na variável X_v e considerando as restantes covariáveis constantes.

Analogamente ao que se efetua para os β_j, a significância das variáveis no modelo pode também ser obtida com recurso aos $\widehat{OR}$. O teste descrito pela equação (3.2.16) para β_j é equivalente a testar a hipótese seguinte:

$$H_0 : OR = 1 \quad versus \quad H_1 : OR \neq 1 \tag{3.2.27}$$

uma vez que se provou que $OR = e^{\beta_j}$. Está-se a testar a hipótese de a probabilidade de sucesso do evento ser igualmente provável em ambos os grupos considerados ($OR = 1$). Isto significa que, pretendemos rejeitar a hipótese nula para que uma variável seja considerada significativa.

Um intervalo com $100(1 - \alpha)\%$ de confiança para os OR é então dado por:

$$\left[e^{\widehat{\beta}_j - z_{1-\alpha/2} SE(\widehat{\beta}_j)}; e^{\widehat{\beta}_j + z_{1-\alpha/2} SE(\widehat{\beta}_j)}\right] \tag{3.2.28}$$

onde $z_{1-\alpha/2}$ representa o $(1 - \alpha/2)$ quantil da distribuição $N(0, 1)$ e SE o erro padrão associado ao coeficiente. Note-se que este IC corresponde a aplicar a função exponencial à equação (3.2.18). Seguindo a analogia anterior, para uma variável explicativa X_j ser considerada significativa, o intervalo de confiança do respetivo OR não poderá conter o valor 1.

3.2.2.3 *Erro de predição*

Consiste em definir um ponto de corte, sendo 0,5 o mais comum, e desta forma criar uma variável dicotómica em que os valores ajustados são agrupados do seguinte modo:

- Se os valores ajustados forem superior a 0,5 assume o valor 1;
- Se os valores ajustados forem inferiores a 0,5 assume o valor 0.

Esta nova variável é cruzada com os valores permitindo assim calcular a proporção de casos preditos corretamente (Tabela 5).

Tabela 5: Tabela de contingência para valores observados e ajustados

		Observado	
		1	0
Predito	1	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	0	Falso Negativo (FN)	Verdadeiro Negativo (VN)

onde, o verdadeiro positivo (VP) é o número de eventos classificados corretamente como eventos. O verdadeiro negativo (VN) é o número de não-eventos classificados corretamente como não-eventos. O falso positivo (FP) é o número de não-eventos classificados incorretamente como eventos e por fim último, o falso negativo (FN) é o número de eventos classificado incorretamente como não-eventos.

Sensibilidade (S): é definida como

$$Sensibilidade = \frac{VP}{FP + VP}$$

e dá a proporção de verdadeiros positivos.

Especificidade (E): é definida como

$$Especificidade = \frac{VN}{VN + FN}$$

e fornece a proporção de verdadeiros negativos

A taxa de acerto (*accuracy*) é a percentagem de indivíduos corretamente classificados, e é dado por

$$accuracy = \frac{VP + VN}{VP + FP + FN + VN}$$

3.3 ANÁLISE DE clusters

A análise de *clusters* envolve uma série de procedimentos estatísticos que podem ser usados para classificar objetos e pessoas observando apenas as semelhanças ou dissemelhanças entre elas, sem serem definidos previamente critérios de inclusão em qualquer agrupamento. Os métodos de análise de *clusters* são procedimentos de estatística multivariada que organizam um conjunto de indivíduos para os quais é conhecida informação detalhada.

De modo sintético, o método pode ser descrito como se segue: dado um conjunto de n indivíduos para os quais existe informação sobre a forma de p variáveis, o método de análise de *clusters* procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto

possível e sempre mais semelhantes aos elementos do mesmo grupo do que os elementos dos restantes grupos (Reis (2001)).

Na análise de *clusters*, é fundamental ter particular cuidado com na seleção das variáveis de partida que vão caracterizar cada indivíduo, já que é um dos aspetos com maior influencia nos resultados de uma análise de *clusters*.

De forma resumida, a análise de *clusters* dividide-se em cinco etapas essenciais (Reis (2001)):

- A seleção de indivíduos a serem agrupados;
- A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária para o agrupamento dos indivíduos;
- A definição de uma medida de semelhança ou dissemelhança entre cada dois indivíduos;
- A escolha de um algoritmo de classificação;
- Por fim, a validação dos resultados.

3.3.1 *Definição de medidas de proximidade*

Um dos aspetos da Análise de *Clusters* é a quantificação da proximidade entre cada dois indivíduos ou objetos, que se pretende agrupar. Intuitivamente, dois objetos pertencem ao mesmo *cluster* se forem semelhantes, ou seja, quanto mais parecidos forem. Quando pertencem a *clusters* diferentes diz-se que são dissemelhantes o que reflete o grau de diferença, afastamento ou divergência entre dois objetos.

Para usar os conceitos de semelhança e dissemelhança de forma útil e eficaz é importante eliminar a sua subjetividade criando medidas concretas de proximidade.

Semelhança: Em muitas situações a medida de proximidade mais fácil de conseguir é a semelhança entre objetos. Uma definição de semelhança entre os objetos i e j, s_{ij}, deve incluir as seguintes propriedades:

1. $s_{ij} \geq 0, \forall\, i, j$;
2. $s_{ij} = s_{ij}, \forall\, i, j$;
3. s_{ij} é tanto maior quanto maior for a semelhança entre os objetos;

Vários autores definem semelhança de forma a que $0 \leq s_{ij} \leq 1$ e assumindo que $s_{ij} = 1$ se e só se $i = j$. Pode acontecer ainda que $-1 \leq s_{ij} \leq 1$, quando a semelhança depende de grandezas do tipo de correlação.

Dissemelhança: Dada uma coleção de objetos defini-se dissemelhança entre dois objetos da coleção, i e j, como a função dos objetos cujos valores d_{ij}, satisfazem as seguintes propriedades:

1. $d_{ij} \geq 0, \forall\, i, j$

2. $d_{ij} = 0, \forall\, i, j$

3. $d_{ij} = d_{ij}, \forall\, i, j$

   A propriedade 1 indica que as medidas de dissemelhança tomam valores não negativos, a propriedade 2 que a dissemelhança de um objeto a ele próprio é nula, enquanto a propriedade 3 assegura a simetria.

   Se além das propriedades anteriores se verifica ainda a propriedade triangular

4. $d_{ij} \leq d_{ik} + d_{jk}, \forall\, i, j, k$

   diz-se que a dissemelhança satisfaz as propriedades de uma semi-métrica ou semi-distância. Se a dissemelhança satisfaz também a propriedade

5. $d_{ij} = 0$ se é só se $i = j$

   diz-se que a dissemelhança é uma métrica ou uma distância.

   Muitas dissemelhanças não satisfazem a propriedade 4, mas satisfazem uma propriedade mais forte do que a propriedade triangular, a chamada propriedade ultramétrica,

6. $d_{ij} \geq max\left(d_{ik}, d_{jk}\right), \forall\, i, j, k,$

   dizendo-se assim que as dissemelhanças são ultramétricas.

Na maioria das situações práticas é suficiente que a dissemelhança satisfaça as propriedades 1, 2 e 3.

Em geral, é possível estabelecer uma relação entre as semelhanças e dissemelhanças dos mesmos objetos. A dissemelhança d_{ij} pode obter-se a partir da semelhança s_{ij}, usando uma função decrescente, como por exemplo $d_{ij} = k - s_{ij}$, onde k é uma constante adequada.

Por sua vez, dada a dissemelhança d_{ij}, pode obter-se s_{ij} usando, por exemplo, a transformação $s_{ij} = \frac{k}{k+d_{ij}}$, onde k é uma constante.

Existem várias medidas que podem ser utilizadas como medidas de distância ou dissemelhança entre os elementos de uma matriz de dados. Considere-se a matriz de dados $X = [x_{ij}], i = 1, ..., n$ e $j = 1, ..., p$. Cormack (1971) descreve uma série de medidas possíveis, entre as quais, se podem destacar como mais utilizadas:

- Distância Euclidiana: a distância entre dois casos (i e j) é a raiz quadrada do somatório dos quadrados das diferenças entre valores de i e j para todas as variáveis ($v = 1, 2, ..., p$).

$$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ij} - x_{jk}\right)^2\right]^{1/2}. \tag{3.3.1}$$

- Quadrado da Distância Euclidiana: a distância entre dois casos (i e j) é definida como o somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis ($v = 1, 2, ..., p$).

$$d_{ij}^2 = \sum_{k=1}^{p} \left(x_{ij} - x_{jk}\right)^2 . \tag{3.3.2}$$

- Distância absoluta ou *City - Block Metric*: a distância entre dois elementos (i e j) é a soma dos valores absolutos das diferenças entre os valores das variáveis ($v = 1, 2, ..., p$) para aqueles dois casos:

$$d_{ij} = \sum_{k=1}^{p} \left|x_{ij} - x_{jk}\right| . \tag{3.3.3}$$

- Distância de *Minkowsky*: definida a partir da medida anterior, pode ser considerada como a generalização da distância Euclidiana (as duas coincidem quando $r = 2$):

$$d_{ij} = \left[\sum_{k=1}^{p} \left|x_{ij} - x_{jk}\right|^r\right]^{1/r} , \; r \geq 1 \tag{3.3.4}$$

- Distância de *Mahalanobis*: também chamada distância generalizada. Esta medida, ao contrário das apresentadas anteriormente, considera a matriz de covariância $\sum$ para o cálculo das distâncias:

$$d_{ij} = \left(X_i - X_j\right)' \sum{}^{-1} \left(X_i - X_j\right). \tag{3.3.5}$$

  sendo X_i e X_j, respetivamente, os vetores de valores das variáveis para os indivíduos i e j.

- Distância de *Chebishev*: a distância entre dois casos (i e j) é o valor máximo para todas as variáveis, das diferenças entre esses dois indivíduos.

$$d_{ij} = \max_k \left|x_{ik} - x_{jk}\right| . \tag{3.3.6}$$

Para a comparação de variáveis deve-se usar semelhanças, sendo as medidas mais adequadas, em geral, medidas de correlação e associação. No caso de variáveis quantitativas duas medidas de correlação são:

- o coeficiente de separação angular

$$s_{ij} = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{\left(\sum_{k=1}^{n} x_{ki}^2 \sum_{k=1}^{n} x_{kj}^2\right)^{1/2}} = cos\, \alpha, \tag{3.3.7}$$

onde α é o ângulo entre os vetores representativos das variáveis i e j, $(x_{1i}, ..., x_{ni})'$ e $(x_{1j}, ..., x_{nj})'$;

- coeficiente de correlação linear de *Pearson*

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\left[\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2\right]^{1/2}}, \tag{3.3.8}$$

em que $\bar{x}_i$ e $\bar{x}_j$ são valores das médias para as variáveis i e j.

O valor varia entre -1 e $+1$, com o valor zero a significar não existir correlação entre os indivíduos. Este coeficiente é particularmente insensível às diferenças de escala das variáveis, uma vez que o cálculo da média de todas as variáveis para cada indivíduo impõe a estandardização prévias dessas variáveis. No entanto, é sensível às diferenças de forma de cada indivíduo e à dispersão dos valores das variáveis em torno das respetivas médias.

3.3.2 *Métodos de Análise de Clusters*

As técnicas de Análise de *Clusters* podem ser divididas em dois grandes grupos: hierárquicas e não hierárquicas. Os algoritmos não hierárquicos iniciam-se com um grupo definido de *clusters* e o processo consiste em transferir elementos entre grupos, até se otimizar determinada condição. A classificação hierárquica, a técnica mais usual e à qual se recorrerá neste estudo, forma uma hierarquia que estabelece a ligação entre um único grupo contendo todos os indivíduos em estudo e n grupos formados por um só indivíduo; em cada passo desse processo o número de *clusters* apenas aumenta ou diminui uma unidade. Se inicialmente se tem n objetos em estudo, serão necessárias $n - 1$ etapas até à conclusão do processo. Os algoritmos hierárquicos podem ser divididos em aglomerativos e em divisivos. Nos métodos aglomerativos parte-se de n grupos, cada um contendo um indivíduo, que vão sendo agrupados sucessivamente de modo a juntar todos os indivíduos num único grupo. Nos métodos divisivos o processo é inverso, ou seja, parte-se de um grupo constituído por indivíduos e por divisões sucessivas obtêm-se grupos mais pequenos até chegar a *clusters* formados por uma só unidade.

Os métodos de Análise de *Clusters* mais divulgados e mais utilizados são os aglomerativos e, isto porque, os métodos divisivos são extremamente pesados em termos de capacidade informática. É de salientar que a maioria dos algoritmos empregues nos métodos aglomerativos também podem ser utilizados em processos divisivos.

Considerando, então, os métodos hierárquicos aglomerativos, a primeira etapa destes consiste em agrupar os dois indivíduos que estiverem mais próximos, podendo para isso recorrer-se à matriz de semelhança ou dissemelhanças.

Concluída a primeira etapa, a matriz de semelhança/dissemelhança tem de ser atualizada de modo a refletir a proximidade entre o grupo recém-formado e os restantes *clusters* compostos por um só indivíduo, sendo possível assim proceder-se ao segundo passo que consiste na junção dos dois grupos mais próximos. Este processo repete-se até que todos os indivíduos estejam contidos num único *cluster*.

Após cada etapa a matriz de proximidades é atualizada e, como pelo menos um dos grupos é formado por mais do que um elemento, é necessário recorrer a métodos que permitam quantificar a proximidade entre um indivíduo e um grupo ou entre dois grupos. Definir uma distância equivale a determinar um método hierárquico aglomerativo que lhe fica associado.

3.3.2.1 *Métodos Hierárquicos aglomerativos mais comuns*

- Ligação Simples ou método do vizinho mais próximo (*Single linkage*): em que a distância entre dois grupos, r e s, é a distância entre os seus elementos mais próximos, isto é, a distância entre o par de objetos com maior proximidade sendo este par formado por um elemento de cada grupo, $d(r,s) = min\{d(i,j) : i \in r, j \in s\}$; embora cada grupo possa ser constituído por vários objetos esta medida baseia-se apenas em dois; os grupos formados exibem o efeito em cadeia, isto é, há tendência para a formação de um número reduzido de grupos com forma alongada;

- Ligação Completa ou método do vizinho mais afastado (*Complete linkage*): a distância entre dois grupos é dada pela distância dos dois elementos mais afastados, $d(r,s) = max\{d(i,j) : i \in r, j \in s\}$; este método tem tendência em desenvolver um grande número de *clusters* formados apenas por observações extremamente próximas umas das outras;

- Distância média entre *clusters (Average linkage)*: esta técnica passa pelo cálculo de uma média que envolva todos os objetos pertencentes aos grupos em questão; se os grupos r e s forem constituídos por n_r e n_s elementos respetivamente, existirão $n_r \times n_s$ pares de objetos possíveis; logo, a medida originada resulta da média entre essas medidas, ou seja, é dada por $p_{rs} = \frac{\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} p_{rs}(i,j)}{n_r} n_r \times n_s$, onde $p_{rs}(i,j)$ é a medida de proximidade entre o i-ésimo elemento do grupo r e o j-ésimo elemento do grupo s;

- Método do centróide: a distância entre dois grupos, r e s, é a distância entre os seus centróides, isto é $d(r,s) = d(\bar{x}_r, \bar{x}_s)$, onde $\bar{x}_r$ e $\bar{x}_s$ são os centróides dos grupos r e s, respetivamente, ou seja $\bar{x}_r = \frac{\sum_{i \in r} x_i}{n_r}$ e $\bar{x}_s = \frac{\sum_{i \in s} x_i}{n_s}$, e x_i é o vetor das p observações do objeto i; em cada passo do algoritmo, os grupos a aglutinar são aqueles cujos centróides estão mais próximos de acordo com a distância que foi definida; um inconveniente deste método é o facto da distância de fusão de dois grupos poder aumentar ou diminuir de passo para passo, tornando a interpretação difícil; a distância entre *clusters* pode ser qualquer medida de proximidade, como por exemplo o coeficiente de correlação ou a

distância euclidiana, mas o quadrado da distância euclidiana é a medida com maior facilidade de aplicação e clareza dos resultados que produz;

- distância mediana (*Median Linkage*): este método é semelhante ao do centróide exceto na aglutinação de dois grupos, r e s, onde os seus centróides recebem pesos iguais antes de produzirem o centróide do novo *cluster*, o novo centróide, $\bar{x}$, fica a meio dos centróides dos grupos aglutinados, $\bar{x} = (\bar{x}_r + \bar{x}_s)/2$, pretende-se evitar que o grupo com maior número de objetos absorva o grupo com menor número;

- Método de *Ward*: neste método, (Ward Jr (1963)), os *clusters* são formados de modo a minimizar a soma dos quadrados dos erros, pois o incremento da soma dos quadrados corresponde efetivamente a uma perda de informação; em cada passo do algoritmo são formados todos os pares possíveis de *clusters* e calculado o incremento da soma dos quadrados, resultante da reunião dos *clusters* de cada par; os *clusters* retidos são aqueles a que corresponde o menor incremento, ou seja, a menor perda de informação resultante da aglutinação; a distância entre dois *clusters* utilizando este método é dada por $d(r,s) = \frac{n_r n_s d_{rs}^2}{n_r + n_s}$, onde d_{rs}^2 é a distância entre os *clusters* r e s definida no método do centróide; o método de *Ward* tem tendência a formar grupos de tamanho semelhante e a encontrar soluções que podem ser ordenadas a partir dos perfis relativamente às variáveis iniciais.

Existem vários métodos aglomerativos. A sua seleção depende muito do objetivo do estudo e das propriedades dos vários métodos. Recomenda-se então a utilização de vários métodos em simultâneo comparando-se os resultados. Caso estes sejam semelhantes, é possível concluir que se obtiveram resultados com elevado grau de estabilidade e, portanto, fiáveis.

A estrutura hierárquica proveniente destes procedimentos costuma ser representada por um gráfico a duas dimensões, designado dendrograma. O dendrograma configura o esquema de uma árvore em posição invertida, com a raiz para cima e os ramos para baixo. Os nós internos representam os *clusters* e a altura dos troncos indica a distância a que se ligam, indicando as alturas pequenas que a aglutinação é feita entre *clusters* razoavelmente homogéneos.

Os métodos hierárquicos adiam a decisão para o final da análise de por onde cortar o dendrograma e, assim, obter o número de *clusters*. Um método simples e informal é a análise gráfica, onde se representa o índice de fusão contra o número de *clusters*. Se a distância entre dois *clusters* é pequena, estes devem ser agregados, se pelo contrário a distância é grande os dois *clusters* devem manter-se separados. Geralmente, a zona de cotovelo do gráfico, quando o declive da reta que une a distância entre dois *clusters* é relativamente pequeno, dá indicação do número de *clusters* a reter.

Outro critério é o critério do R^2 que representa uma medida de quão diferente cada um dos *clusters* são, em cada passo do algoritmo. O R^2 é calculado como a razão entre a soma dos

quadrados entre os *clusters* (SQC) e a soma dos quadrados totais (SQT) para cada uma das variáveis usadas na análise

$$R^2 = \frac{SQC}{SQT} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{k} n_{ij}\left(\bar{X}_{ij} - \bar{X}_i\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{k}\sum_{l=1}^{n_i}\left(\bar{X}_{ijl} - \bar{X}\right)^2}, \tag{3.3.9}$$

onde p representa o número de variáveis, k o número de grupos, n_{ij} o tamanho do grupo j na variável i, $\bar{X}_{ij}$ a média da variável i no grupo j, $\bar{X}_i$ a média da variável i e $\bar{X}$ a média da amostra global.

Sendo assim, o R^2 é uma medida da percentagem da variabilidade total que é retida em cada uma das soluções dos *clusters*. Dado que no caso de existir um único *cluster* a variabilidade entre os *clusters* é zero, e no caso de existirem tantos *clusters* quanto objetos a variabilidade é total, interessa encontrar um número mínimo de *clusters* que retenha uma percentagem significativa da variabilidade total.

4

BASE DE DADOS

Antes de se dar início à extração e construção das variáveis necessárias no cálculo do *Engagement Score*, é necessário calcular o período de estabilidade do Continente.

4.1 DETERMINAÇÃO DO PERÍODO DE ESTABILIDADE

O período de estabilidade corresponde ao tempo ao fim do qual a maioria dos clientes visita uma loja Continente. Como tal, calcula-se o número acumulado de novos clientes em cada mês, durante um período de tempo.

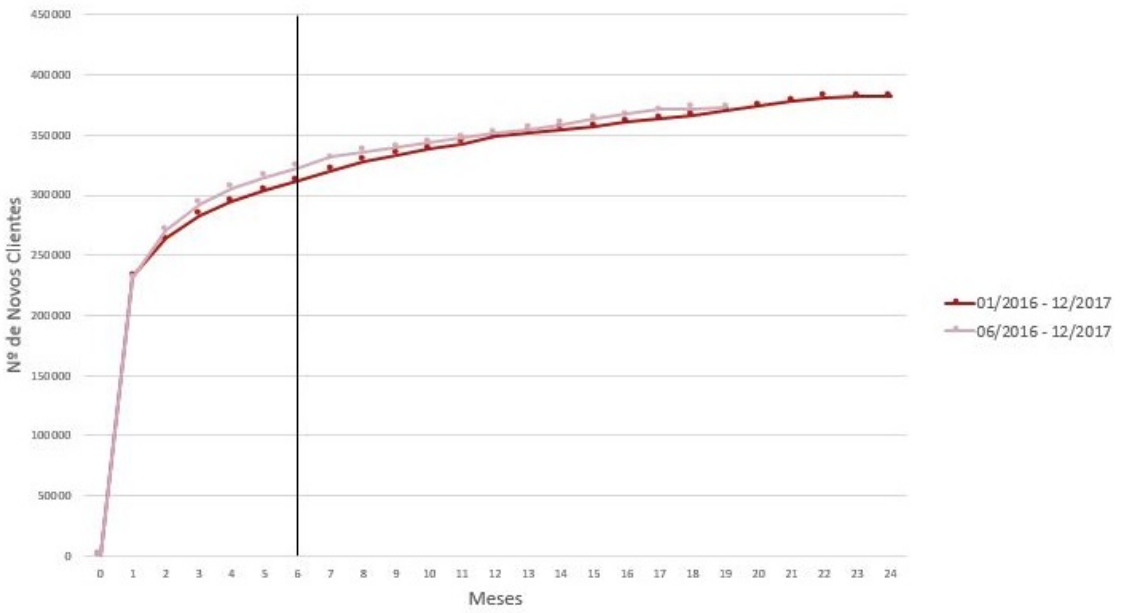

Figura 4: Número acumulado de novos clientes durante 2 anos

A Figura 4 mostra o número acumulado de novos clientes durante 2 anos, em 2 períodos de tempos distintos, por forma a detetar uma possível existência de picos (devido a fenómenos sazonais).

Para se perceber de uma forma mais concreta o período de estabilidade a ser usado, é calculado o número de novos clientes, o número de clientes mantidos e o número de clientes perdidos. Este cálculo é feito usando dois períodos de tempo distintos: se o cliente for no primeiro período de tempo e não for no segundo período de tempo em análise é considerado como cliente perdido. Se o cliente fizer compras nos dois períodos de tempo, é considerado como cliente mantido, e se o cliente só fizer compras no segundo período de tempo é considerado como um novo cliente.

Tabela 6: Clientes Novos, Mantidos e Perdidos

Jul'16 - Dez'16 / Jan'17 - Jun'17			
Novos	**Perdidos**	**Mantidos**	**Total**
254 k	326 k	2 940 k	3 520 k
7 %	9 %	84 %	100 %

Tabela 7: Clientes Novos, Mantidos e Perdidos

Jan'17 - Jun'17 / Jul'17 - Dez'17			
Novos	**Perdidos**	**Mantidos**	**Total**
335 k	226 k	2 968 k	3 528 k
10 %	6 %	84 %	100 %

As Tabelas 6 e 7 mostram para dois momentos distintos a percentagem de clientes novos, mantidos e perdidos. Através da análise destas tabelas, verifica-se que num período de análise de 6 meses abrange-se a maioria dos clientes. Com base na Tabela 6, na comparação entre os últimos 6 meses e 2016 e os primeiro 6 meses de 2017, 254000 clientes são classificados como novos clientes, o que corresponde apenas a aproximadamente 7% dos clientes.

Para se perceber se ao fim de 6 meses existe uma grande representatividade do número de clientes e quais as vendas associadas, foram analisados os resultados da Figura 4.

Tabela 8: Definição do Período de Estabilidade

Jan'17 - Jun'17	
% de Clientes	**% de Vendas**
82%	96%

A Tabela 8, mostra que ao fim dos primeiros 6 meses de 2017, tinham visitado uma loja da insígnia Continente 82% dos clientes o que representa 96% das vendas anuais. Pela análise da tabela, é claro que um período de estabilidade de 6 meses se ajusta à realidade do negócio, como tal será esse o período de tempo considerado na extração de variáveis e construção

dos modelos para o cálculo do *Engagement Score*. Num estudo é importante termos acesso à informação mais recente disponível sobre o cliente, como tal serão utilizados os últimos 6 meses de 2017.

4.2 RECOLHA DE DADOS INICIAL

A informação sociodemográfica e transacional dos clientes foi fornecida pela Entidade Gestora do Cartão Continente da SONAE MC, correspondente ao período entre 1 de julho de 2017 a 31 de dezembro de 2017 (6 meses).

Na base de dados, o cliente está identificado pela variável **id_cliente**, através de um código irreal. É importante referir que, as variáveis apresentam valores irreais de modo a garantir a confidencialidade dos valores reais da empresa.

No estudo, a base de dados *Amostra*, já estabelecida pela empresa, foi criada retirando uma amostra de 10% dos clientes da base de dados original com todos os clientes, o que corresponde a 330215 clientes. Como estamos perante uma base de dados de grandes dimensões, a amostra é representativa da população.

Usando a variável **id_cliente** da base de dados *Amostra*, é possível retirar toda a informação necessária sobre os clientes bem construir algumas novas variáveis.

1. **Extração/Construção de variáveis Comportamentais**

   Acedendo às tabelas com *Informação Transacional* foram retiradas e construídas para cada cliente, as seguintes variáveis que dizem respeito aos últimos 6 meses de 2017:

   - **id_cliente** - Número do cliente.
   - **nr_trx** - Número de transações realizadas.
   - **loc_brand_cd** - Indica em que insígnia o cliente efetuou as suas compras, codificada como: 1 - Continente, 2 - Continente Bom Dia e 3 - Modelo.
   - **vb_acu** - Vendas brutas totais do cliente no período em estudo, em euros.
   - **vl_acu** - Vendas líquidas totais do cliente no período em estudo, em euros (**vl_acu** = **vb_acu** − *valor do* **IVA**).
   - **db_cc_acu** - Desconto acumulado em cartão.
   - **db_sp_acu** - Desconto em super preço.
   - **nr_loja_visitadas** - Número de lojas diferentes que o cliente visita.
   - **compra_online** - Indica se o cliente faz compras online, codificada como: 0 - Não e 1 - Sim.
   - **cesta_média** - Valor médio gasto por transação. Esta variável é calculado pelo quociente $\left(\frac{\textbf{vb\_acu}}{\textbf{nr\_trx}}\right)$.

- **perc_dsc** - Percentagem de descontos. Esta variável é determinada pela razão $\left(\frac{\text{db\_cc\_acu}+\text{db\_sp\_acu}}{\text{vb\_acu}+\text{db\_sp\_acu}} \times 100\%\right)$.

2. **Extração da variável Loja Preferencial**

   Na tabela *Loja Preferencial* é disponibilizada, com base no histórico de compras do cliente, qual o tipo de loja preferencial. Foi retirada a informação dos últimos 6 meses de 2017 para a obtenção da variável.

   - **id_cliente** - Número do cliente.
   - **loja_preferencial** - Esta variável indica qual a preferência do cliente em relação ao tipo de loja, codificada como: 0 - online e 1 - física.

3. **Extração das Segmentação *Baby&Junior*, *Price Sensitivity*, *Share of Wallet*, Valor, Estilo de Vida, *Net Promoter Score* e *Churn***

   Existem tabelas já criadas para cada uma das segmentações, atualizadas mensalmente, que através de variáveis fornecidas pelos clientes e por pelos hábitos de compra no ecossistema do Cartão Continente, distinguem os clientes nas várias categorias existentes. Foram retiradas as segmentações de cada cliente para o último mês do período de análise, dezembro de 2017.

   - **id_cliente** - Número do cliente.
   - **segm_bj** - Categoria na segmentação *Baby&Junior*. Esta segmentação é codificada como, 1 - *Baby*, 2 - *Junior*, 3 - *Baby and Junior* e 4 - *No baby no junior*.
   - **segm_ps** - Categoria na segmentação *Price Sensitivity*. Esta segmentação é codificada como, 1 - *SEG_PS_1*, 2 - *SEG_PS_2*, 3 - *SEG_PS_3*, 4 - *SEG_PS_4*, 5 - *SEG_PS_5*, 6 - *SEG_PS_6*, 7 - *SEG_PS_7*, 8 - *SEG_PS_8* e 9 - Sem Valor.
   - **segm_sow** - Categoria na segmentação *Share of Wallet*. Esta segmentação é codificada como, 1 - *HIGH*, 2 - *MEDIUM*, 3 - *LOW*, 4 - *VERY LOW* e 5 - Sem Valor.
   - **segm_value** - Categoria na segmentação *Valor*. Esta segmentação é codificada como, 1 - *SEG_1*, 2 - *SEG_2*, 3 - *SEG_3*, 4 - *SEG_4*, 5 - *SEG_5*, 6 - *SEG_6*, 7 - *SEG_7* e 8 - Sem Valor.
   - **segm_ev** - Categoria na segmentação *Estilo de Vida*. Esta segmentação é codificada como, 1 - *SEG_EV_1*, 2 - *SEG_EV_2*, 3 - *SEG_EV_3*, 4 - *SEG_EV_4*, 5 - *SEG_EV_5*, 6 - *SEG_EV_6*, 7 - *SEG_EV_7* e 8 - Sem Valor.
   - **segm_nps** - Categoria na segmentação *Net Promoter Score*. Esta segmentação é codificada como, 1 - *SEG_NPS_1*, 2 - *SEG_NPS_2*, 3 - *SEG_NPS_3* e 4 - Sem Valor.
   - **pactive** - Indica a probabilidade de o cliente estar ativo com a marca, isto é, é o contrário da probabilidade de abandono (*Churn*). Quando o resultado do *pactive*

é próximo de 1 indica baixa probabilidade do cliente abandonar a marca estando o cliente muito ativo, por outro lado, quando o valor do *pactive* é próximo de 0, significa alta probabilidade de o cliente abandonar a marca.

- **ltv** - Variável *Lifetime Value*. Esta variável indica qual a previsão de vendas líquidas reportadas de cada cliente.

4. **Construção da Frequência Relativa de Transações por Insígnia**

   Recorrendo novamente à tabela com *Informação Transacional*, e após já terem sido extraídas variáveis desta tabela, é possível calcular a frequência relativa de transações em cada uma das Insígnias. Como tal, foram criadas três variáveis que indicam, para cada cliente, o número de transações realizadas em cada insígnia.

   - **id_cliente** - Número do cliente.
   - **nr_trx_insignia_continente** - Número de transações realizadas na insígnia Continente.
   - **nr_trx_insignia_bom_dia** - Número de transações realizadas na insígnia Continente Bom Dia.
   - **nr_trx_insignia_modelo** - Número de transações realizadas na insígnia Continente Modelo.

   O cálculo das frequências relativas das transações para cada insígnia, é obtido pelo quociente entre o número de transações realizadas em cada insígnia e o número de transações realizadas

   - **freq_continente** - Frequência relativa de transações no Continente.
   - **freq_bom_dia** - Frequência relativa de transações no Continente Bom Dia.
   - **freq_modelo** - Frequência relativa de transações no Continente Modelo.

5. **Construção da Frequência Relativa de Vendas por Departamento Comercial**

   Recorrendo à tabela *Descrição do Produto* e cruzando esta informação com a tabela *Informação Transacional* sabe-se quais os produtos adquiridos pelo cliente, o Departamento Comercial correspondente e qual o custo associado às transações. A informação retirada foi referente apenas aos últimos 6 meses de 2017. Com esta informação, construiu-se sete novas variáveis, que indicam o valor gasto por cliente para cada Departamento Comercial (DC).

   - **id_cliente** - Número do cliente.
   - **vb_acu_novos_bazar_casa** - Vendas brutas relativas a transações no Departamento Comercial Novos Bazar e Casa.

- **vb_acu_bazar** - Vendas brutas relativas a transações no Departamento Comercial Bazar.
- **vb_acu_peixaria_talho** - Vendas brutas relativas a transações no Departamento Comercial Peixaria & Talho.
- **vb_acu_alimentar** - Vendas brutas relativas a transações no Departamento Comercial Alimentar.
- **vb_acu_textil** - Vendas brutas relativas a transações no Departamento Comercial Têxtil.
- **vb_acu_nutricao_saudavel** - Vendas brutas relativas a transações no Departamento Comercial Nutrição Saudável.
- **vb_acu_pad_taway** - Vendas brutas relativas a transações no Departamento Comercial F&L C&Q PAD TAWAY (Frutas & Legumes, Charcutaria, Padaria & *Take Away*).

O cálculo das frequências relativas das vendas para cada Departamento Comercial é obtido pelo quociente entre as vendas brutas relativas a transações de cada Departamento Comercial e as vendas brutas totais.

- **novos_bazar_casa** - Frequência Relativa de Transações no Departamento Comercial Novos Bazar e Casa.
- **bazar** - Frequência Relativa de Transações no Departamento Comercial Bazar.
- **peixaria_talho** - Frequência Relativa de Transações no Departamento Comercial Peixaria & Talho.
- **alimentar** - Frequência Relativa de Transações no Departamento Comercial Alimentar.
- **textil** - Frequência Relativa de Transações no Departamento Comercial Têxtil.
- **nutricao_saudavel** - Frequência Relativa de Transações no Departamento Comercial Nutrição Saudável.
- **pad_taway** - Frequência Relativa de Transações no Departamento Comercial F&L C&Q PAD TAWAY (Frutas & Legumes, Charcutaria, Padaria & *Take Away*).

6. **Construção da Frequência Relativa de Transações por Missão de Compra**

   Com base na tabela *Missão de Compra*, é possível atribuir qual a missão de compra a cada cliente com base nas suas transações. A variável referente às missões de compra de cada cliente, está dividia em várias categorias, *missao_1*, *missao_2*, *missao_3*, *missao_4* e *missao_5*. Foi construida a variável **nr_trx_missao** que indica o número de transações realizadas por cliente em cada missão de compra nos últimos 6 meses de 2017.

- **id_cliente** - Número do cliente.
- **nr_trx_missao_1** - Número de Transações na Missão de Compra 1.
- **nr_trx_missao_2** - Número de Transações na Missão de Compra 2.
- **nr_trx_missao_3** - Número de Transações na Missão de Compra 3.
- **nr_trx_missao_4** - Número de Transações na Missão de Compra 4.
- **nr_trx_missao_5** - Número de Transações na Missão de Compra 5.

O cálculo das frequências relativas de transações realizadas em cada missão de compra é obtido pelo quociente entre as transações relativas a cada missão de compra e as transações totais.

- **missao_1** - Frequência Relativa de Transações na Missão de Compra 1.
- **missao_2** - Frequência Relativa de Transações na Missão de Compra 2.
- **missao_2** - Frequência Relativa de Transações na Missão de Compra 3.
- **missao_4** - Frequência Relativa de Transações na Missão de Compra 4.
- **missao_5** - Frequência Relativa de Transações na Missão de Compra 5.

7. **Extração dos Dados Demográficos dos Clientes**

   Para cada cliente, foi extraída ainda informação da tabela de *Informação de cliente*, para se identificar a sua localização geográfica.

   - **id_cliente** - Número do cliente.
   - **cliente_insco** - Esta variável é codificada como 0 - Portugal Continental/Madeira e 1 - Açores.

Depois da extração e construção de todas as variáveis necessárias, procedeu-se à construção de uma tabela final que agrega toda a informação necessária para caracterizar os clientes (Tabela 9 e 10).

Tabela 9: Quadro Resumo da Variáveis Utilizadas no Estudo

Variáveis	Descrição	Tipo de variável
id_cliente	Número do cliente	
loja_preferencial	Tipo de loja preferencial (física ou online)	Binária
compra_online	Se o cliente faz compras online	Binária
nr_lojas_visitadas	Nº de Lojas diferentes que o cliente visita	Quantitativa Discreta
nr_trx	Nº de Transações	Quantitativa Discreta
vb_acu	Vendas Brutas	Quantitativa Contínua
vl_acu	Vendas Líquidas	Quantitativa Contínua
cesta_media	Valor médio gasto por transação	Quantitativa Contínua
db_cc_acu	Desconto acumulado em cartão	Quantitativa Contínua
db_sp_acu	Desconto em Super Preço	Quantitativa Contínua
segm_bj	Categoria da segmentação *Baby & Junior* a que pertence o cliente	Qualitativa Nominal
segm_ps	Categoria da segmentação *Price Sensitivity* a que pertence o cliente	Qualitativa Nominal
segm_sow	Categoria da segmentação SOW a que pertence o cliente	Qualitativa Ordinal
segm_value	Categoria da segmentação Valor a que pertence o cliente	Qualitativa Ordinal
pactive	Probabilidade de um cliente se manter ativo	Quantitativa Contínua
ltv	*Lifetime Value*	Quantitativa Contínua
segm_nps	Categoria da segmentação NPS a que pertence o cliente	Qualitativa Nominal
segm_ev	Categoria da segmentação Estilo de Vida a que pertence o cliente	Qualitativa Nominal
perc_dsc	Percentagem de Descontos	Quantitativa Contínua
freq_continente	Percentagem de transações no Continente	Quantitativa Contínua
freq_bom_dia	Percentagem de transações no Bom Dia	Quantitativa Contínua
freq_modelo	Percentagem de transações no Modelo	Quantitativa Contínua
novos_bazar_casa	Percentagem de vendas no Departamento Comercial Novos Bazar e Casa	Quantitativa Contínua
bazar	Percentagem de vendas na Departamento Comercial Bazar	Quantitativa Contínua
peixaria_talho	Percentagem de vendas na Departamento Comercial Peixaria & Talho	Quantitativa Contínua
alimentar	Percentagem de vendas na Departamento Comercial Alimentar	Quantitativa Contínua
textil	Percentagem de vendas na Departamento Comercial Têxtil	Quantitativa Contínua
nutricao_saudavel	Percentagem de vendas na Departamento Comercial Nutrição Saudável	Quantitativa Contínua
pad_taway	Percentagem de vendas na Departamento Comercial Padaria & *Take Away*	Quantitativa Contínua
missao_1	Percentagem de transações na missão de Compra 1	Quantitativa Contínua

Tabela 10: Quadro Resumo da Variáveis Utilizadas no Estudo (Cont.)

Variáveis	Descrição	Tipo de variável
missao_2	Percentagem de transações na missão de Compra 2	Quantitativa Contínua
missao_3	Percentagem de transações na missão Suprimir de Compra 3	Quantitativa Contínua
missao_4	Percentagem de transações na missão de Compra 4	Quantitativa Contínua
missao_5	Percentagem de transações na missão de Compra 5	Quantitativa Contínua
cliente insco	Indica se é um cliente dos Açores	Binária

4.3 Análise Exploratória dos Dados

A análise exploratória é uma etapa fundamental na fase inicial de um projeto, permitindo ter um conhecimento mais aprofundado sobre os dados que se está a trabalhar. Numa fase seguinte, e após esta análise das variáveis, é possível perceber se é necessário transformar as variáveis.

Na Tabela 11 apresentam-se algumas medidas estatísticas de localização e de dispersão das variáveis quantitativas.

Tabela 11: Medidas de localização e dispersão das variáveis quantitativas

Variáveis	Média	Desvio Padrão	Mínimo	Máximo	Q1	Mediana	Q3	Coeficiente de Variação	N
nr_lojas_visitadas	2,04	1,74	1	37	1	2	3	68,30	330215
cesta_media	26,31	27,22	0,09	1639,98	16,08	26,06	41,51	82,76	330215
nr_trx	16,60	24,36	1	536	5	13	27	117,45	330215
vb_acu	487,59	720,98	0,09	21586,49	125,43	358,80	830,16	118,29	330215
vl_acu	425,11	631,59	0,08	20053,73	108,42	311,44	723,60	118,86	330215
db_cc_acu	25,47	50,86	0	3268,28	0	12,42	42,64	159,76	330215
db_sp_acu	105,59	168,66	0	4765,90	22,69	73,39	178,06	127,78	330215
ltv	457,92	649,30	0	18145,18	133,86	349,03	777,51	113,43	330215
perc_dsc	16,79	9,60	0	358,82	14,80	20,77	26,82	45,77	330215

Como já referido, a base de dados é constituída por 330215 clientes, não apresentando valores em falta para nenhuma das variáveis.

Analisando a Tabela 11, relativamente ao número de lojas visitadas no período de análise, a média para o total de clientes analisado é de 2,04 lojas por cliente enquanto que o número de transações médio efetuado por cliente é de 16,60. Sobre as restantes variáveis analisadas, conclui-se que os clientes gastam em média 487,59 €, ou seja, gastam aproximadamente 81,26 €por mês nas insígnias Continente acumulando aproximadamente 25,47 €de descontos em cartão.

Verificar-se também que existem coeficientes de variação muito elevados, ou seja, existe uma dispersão muito grande dos dados, o que é explicado pelo facto dos clientes terem comportamentos de compra muito distintos entre si.

Ainda sobre a Tabela 11, é possível observar que há uma grande concentração de clientes nos valores mais baixos de cada variável. O histograma é a representação mais clara deste facto (Figura 5 e 6).

Nas Figuras 5 e 6 estão representados os histogramas das variáveis quantitativas, onde se observa uma concentração da distribuição ao lado esquerdo na maioria das variáveis, (distribuição assimétrica positiva).

Na Figura 5, as variáveis tem um comportamento semelhante, apresentando uma distribuição assimétrica positiva, ou seja, o número de clientes vai diminuindo à medida que o valor da variável vai aumentando, com exceção para a variável **pactive** que apresenta uma distribuição assimétrica negativa, isto significa que, a probabilidade de um cliente estar ativo é alta. Para o caso da variável **cesta_media**, verifica-se que existe uma grande concentração de clientes no valor mais baixo da cesta, sendo que à medida que se vai aumentando o valor da cesta média, o número de clientes vai diminuindo, o mesmo se passa com as variáveis **vl_acu**, **db_cc_acu**, **db_sp_acu**, **vb_acu** e **ltv**.

Na Figura 6, o histograma da variável **alimentar** assemelha-se a um histograma da distribuição normal, indicando que do total gasto pelos clientes na insígnia Continente, metade desse valor é gasto no Departamento Comercial Alimentar. Quando se analisa os resultados dos restantes Departamentos Comerciais, verifica-se que o valor gasto pelos clientes é inferior ao valor médio. Sobre as variáveis que se referem às missões de compra, os clientes tem um comportamento idêntico.

A Tabela 12 apresenta informação resumida sobre as variáveis binárias presentes na análise.

Tabela 12: Tabelas de Frequências das variáveis binárias

Variáveis		nº clientes	% clientes
loja_preferencial	0	3565	$1,08\%$
	1	326650	$98,92\%$
compra_online	0	324825	$98,37\%$
	1	5390	$1,63\%$
cliente_insco	0	322386	$97,63\%$
	1	7829	$2,37\%$

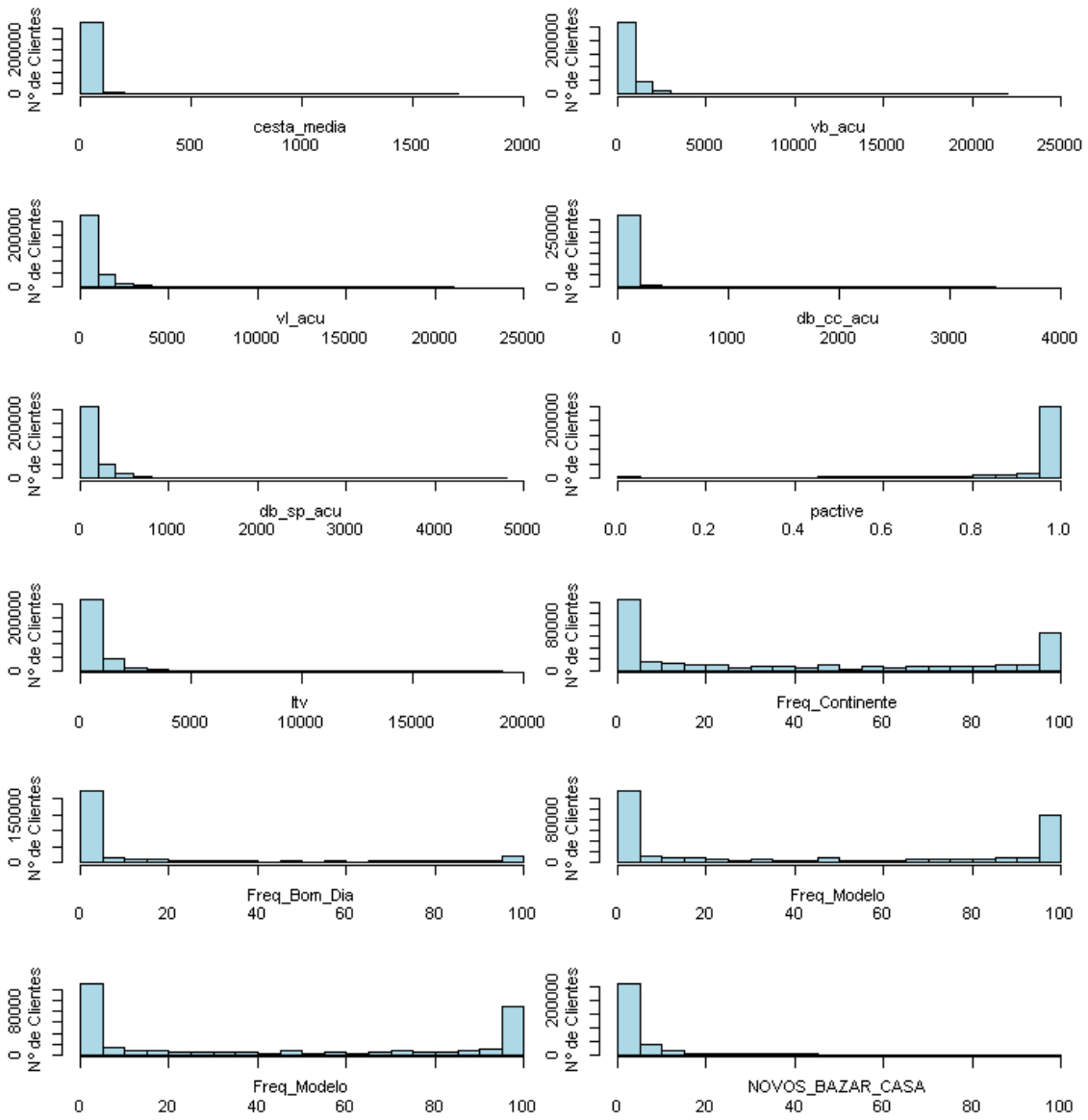

Figura 5: Histograma I - Variáveis Numéricas

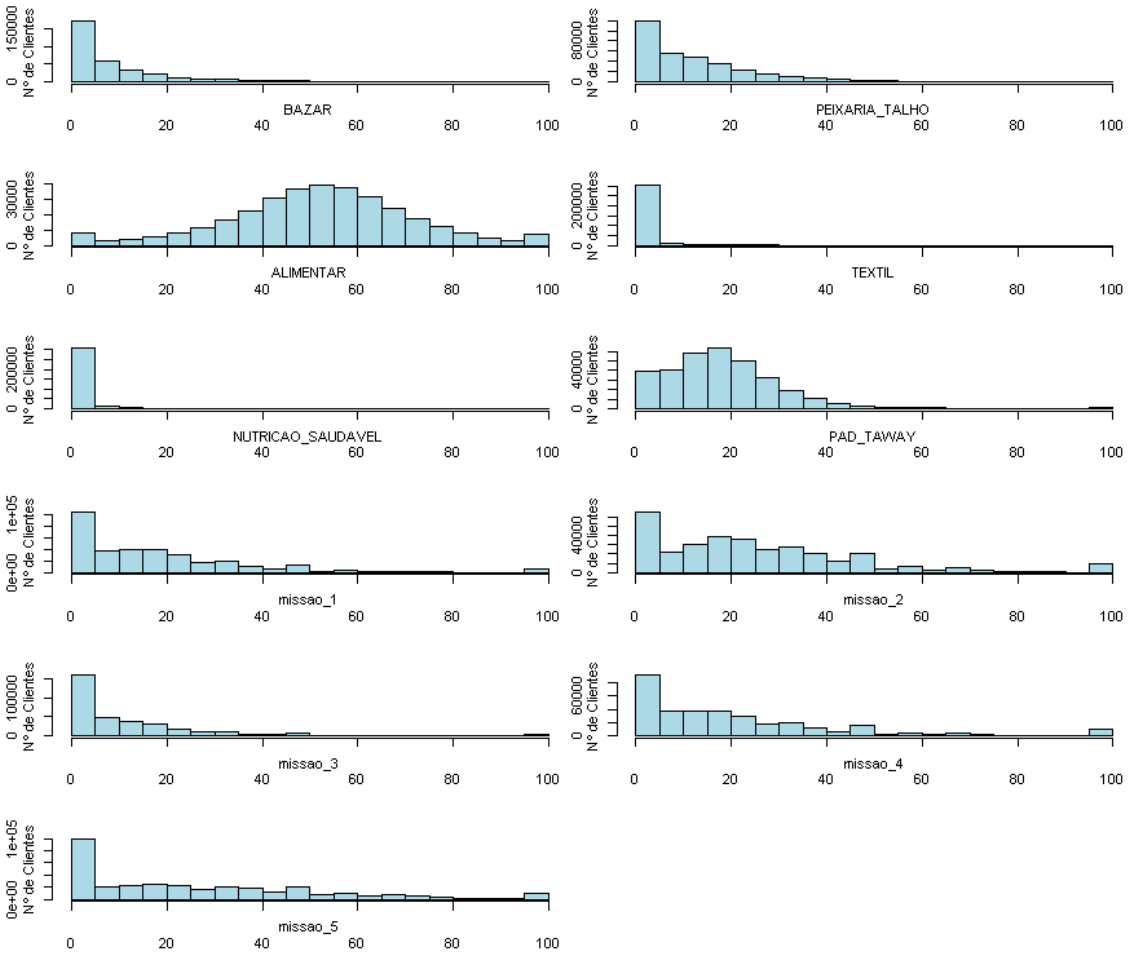

Figura 6: Histograma II - Variáveis Numéricas

Verifica-se que apenas $1,08\%$ dos clientes preferem comprar online e que apenas $1,63\%$ dos clientes realizaram compras online no período de 1 de julho a 31 de dezembro de 2017. Por fim, apenas $2,37\%$ dos clientes são dos Açores. É importante identificar os clientes desta região, uma vez que, as marcas e lojas são diferentes das restantes presentes na Madeira e em Portugal Continental, sendo necessário, por vezes analisar os seus comportamentos de forma separada.

Após uma breve análise univariada das variáveis quantitativas, é importante analisar as variáveis categóricas em estudo. A distribuição dos clientes pelos diversos segmentos analisados encontra-se representada nas Tabelas 12 a 17.

Tabela 13: Segmentação *Baby & Junior*

segm_bj	**nº clientes**	**% clientes**
Junior	45249	$13,70\%$
Baby and Junior	14214	$4,30\%$
Baby	11220	$3,40\%$
No baby no Junior	259532	$78,60\%$

Para a segmentação *Baby & Junior* (Tabela 13), aquela que apresenta uma maior percentagem de clientes é o último segmento, indicando que a maioria dos agregados familiares não tem filhos com idade bebé ou júnior.

Tabela 14: Segmentação *Price Sensitivity*

segm_ps	**nº clientes**	**% clientes**
SEG_PS_1	17887	$5,42\%$
SEG_PS_2	54377	$16,47\%$
SEG_PS_3	43022	$13,03\%$
SEG_PS_4	16207	$4,90\%$
SEG_PS_5	51271	$15,53\%$
SEG_PS_6	36765	$11,13\%$
SEG_PS_7	34736	$10,52\%$
SEG_PS_8	42769	$12,95\%$
Sem Valor	33181	$10,05\%$

A Tabela 14, referente à segmentação *Price Sensitivity* (**segm_ps**) mostra que, os clientes se distribuem mais equitativamente pelas várias categorias existentes. Contudo a categoria que apresenta mais clientes é a SEG_PS_2, que representa clientes com baixa penetração de marca própria e baixa atividade promocional, mas alto número de marcas diferentes compradas, em suma, são clientes que tendem a não ser sensíveis ao preço.

Tabela 15: Segmentação *Share of Wallet*

segm_sow	nº clientes	% clientes
High	63817	19,33%
Medium	46805	14,17%
Low	60251	18,25%
Very Low	28549	8,65%
Sem Valor	130793	39,60%

A segmentação *Share of Wallet* (Tabela 15), uma das mais importantes segmentações para se perceber o envolvimento do cliente com uma marca, tem a maioria do clientes na categoria Sem Valor, ou seja, são clientes com pouco histórico para terem uma segmentação definida. No entanto, observando a distribuição dos restantes clientes verifica-se que o segmento com mais clientes é o *High* e aquele que apresenta menor percentagem de clientes é o *Very Low*, o que representa um bom indicador. De facto, a categoria *High* corresponde a clientes que gastam a maior parte do seu orçamento disponível para compras no Continente, isto indica que são clientes com grande valor para a marca e com maior tendência a serem leais, por outro lado aqueles clientes que se encontram nas segmentações mais baixas, são clientes que fazem transações de baixo valor, e que não tem o Continente como marca preferencial.

Tabela 16: Segmentação Valor

segm_valor	nº clientes	% clientes
SEG_1	40816	12,36%
SEG_2	28281	8,56%
SEG_3	23077	6,99%
SEG_4	52928	16,03%
SEG_5	35105	10,63%
SEG_6	56071	16,98%
SEG_7	59235	17,94%
Sem Valor	34702	10,51%

A segmentação Valor (Tabela 16), baseia-se em variáveis como a recência, a frequência de compras e as vendas, ou seja, é baseada em modelos RFM que ajudam a compreender o envolvimento dos clientes. Nesta segmentação, quanto menor for a categoria da segmentação maior é o valor que o cliente representa para o Continente, isto é, um cliente presente no SEG_1 tem mais valor quando comparado com um cliente presente no SEG_7. Esta segmentação apresenta maior percentagem de clientes no segmento com mais baixo valor, contudo, a percentagem de clientes nos vários segmentos é muito semelhante.

Tabela 17: Segmentação Estilo de Vida

segm_ev	nº clientes	% clientes
SEG_EV_1	59156	17,91%
SEG_EV_2	49021	14,85%
SEG_EV_3	57506	17,41%
SEG_EV_4	43415	13,15%
SEG_EV_5	44226	13,39%
SEG_EV_6	36008	10,91%
SEG_EV_7	34457	10,43%
Sem Valor	6426	1,95%

A Segmentação Estilo de Vida (Tabela 17), é uma segmentação muito importante, pois é aquela que permite caraterizar os clientes com base nos seus comportamentos de compra e estilo de vida. O segmento com maior número de clientes presentes é o SEG_EV_1. Com base em informação fornecida pela empresa, sabe-se que este grupo é caracterizado por serem clientes com mais idade e que estão inseridos em agregados familiares de uma ou duas pessoas. São o grupo de clientes que menos usam as plataformas tecnológicas e encontram-se sobretudo no Norte e Centro do país.

Tabela 18: Segmentação *Net Promoter Score*

segm_nps	nº clientes	% clientes
SEG_NPS_1	54089	16,38%
SEG_NPS_2	122994	37,25%
SEG_NPS_3	139254	42,17%
Sem Valor	13878	4,20%

Por último, a segmentação *Net Promoter Score* (Tabela 18) é outra segmentação muito utilizada. Neste caso, verifica-se que os clientes considerados como SEG_NPS_3 são cerca de 2,6 vezes superiores aos considerados como SEG_NPS_1.

4.4 PRÉ-PROCESSAMENTO DA BASE DE DADOS

A fase de preparação de dados é fundamental em todo o processo de modelação para que os resultados sejam obtidos com a maior precisão. Nesta fase, e após a análise descritiva das variáveis, será feita uma seleção e limpeza dos dados.

Tratamento de Outliers

Fez-se um estudo de observações *outliers*, que podem ser classificados como *outliers* moderados e *outliers* severos. As observações são consideradas *outliers* moderados se estiverem entre

$Q_1 - 1,5D$ e $Q_1 - 3D$ ou entre $Q_3 + 1,5D$ e $Q_3 + 3D$, com Q_1 valor do primeiro quartil e Q_3 valor do terceiro quartil e $D = Q_3 - Q_1$ (amplitude inter-quartil). Se forem menores do que $Q_1 - 3D$ ou maiores do que $Q_3 + 3D$ são considerados *outliers* severos. Na identificação das observações *outliers* na base de dados optou-se por apenas identificar os *outliers* severos superiores de modo a evitar um corte drástico de observações e uma possível perda de informação, indo ao encontro com aquilo que são as necessidades do negócio.

Na Figura 8 estão representados os *box-plots* de algumas das variáveis numéricas da base de dados em estudo.

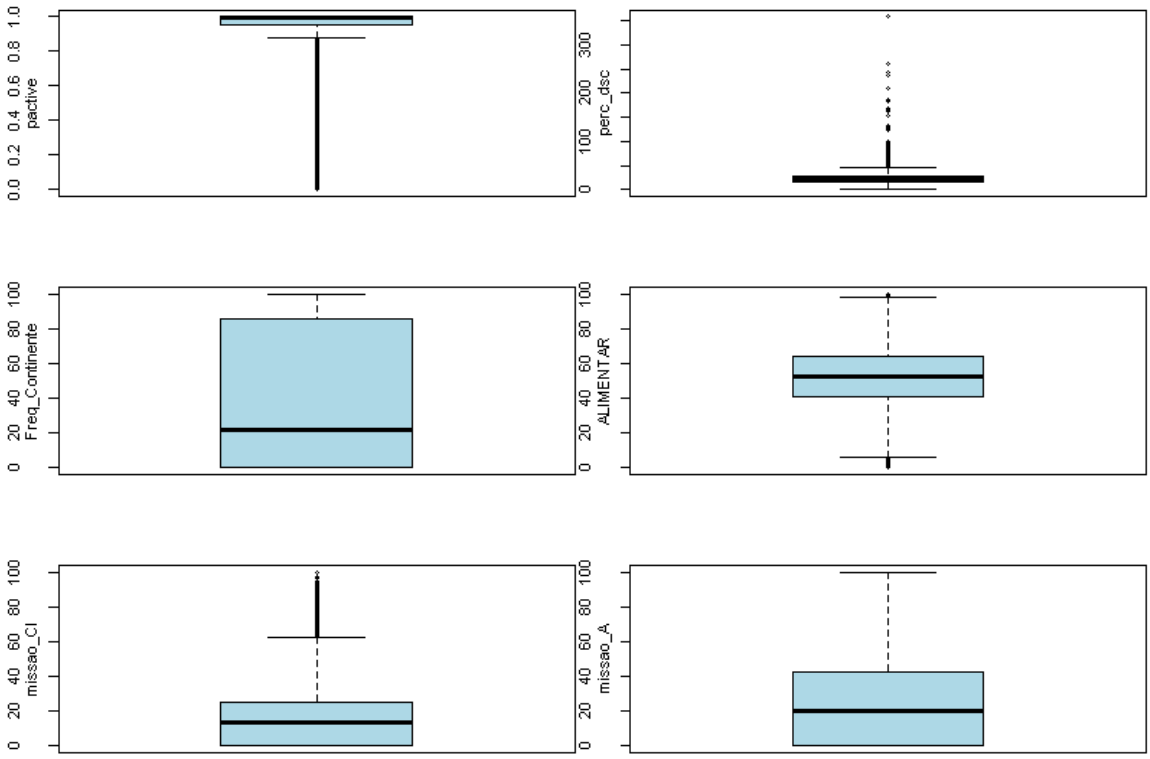

Figura 7: *Box-plot* - Algumas Variáveis Numéricas

Na Figura 9 é possível observar a existência de *outliers* severos nas variáveis **pactive** e **perc_dsc**, sendo que apenas serão retirados para a variável **perc_dsc**, pois só nesta variável é que os *outliers* severos são superiores. A Tabela 17 indica, por cada cliente, em quantas variáveis é considerado *outlier*.

Tabela 19: Distribuição dos *outliers*

Outliers	**Nº Clientes**	**% Clientes**
0	325170	$98,47\%$
1	4683	$1,42\%$
2	362	$0,11\%$

Pela análise da Tabela 17, verifica-se que $98,47\%$ dos clientes não são considerados *outliers*. As variáveis que apresentam *outliers* são o **nr_trx**, a **vb_acu** e a **perc_dsc**. Após a exclusão destes clientes a base de dados ficou reduzida a 325170 clientes.

Número mínimo de transações

Um dos tópicos mais importantes para a empresa é perceber qual o número mínimo de transações que um cliente deverá ter para ser incluído na análise. Esta escolha baseia-se no período de tempo em análise, no número de transações feitas por cada cliente e as respetivas vendas. Foi construído um gráfico que indica o valor gasto total de compras em função do número de transações, de forma a determinar-se o ponto de corte.

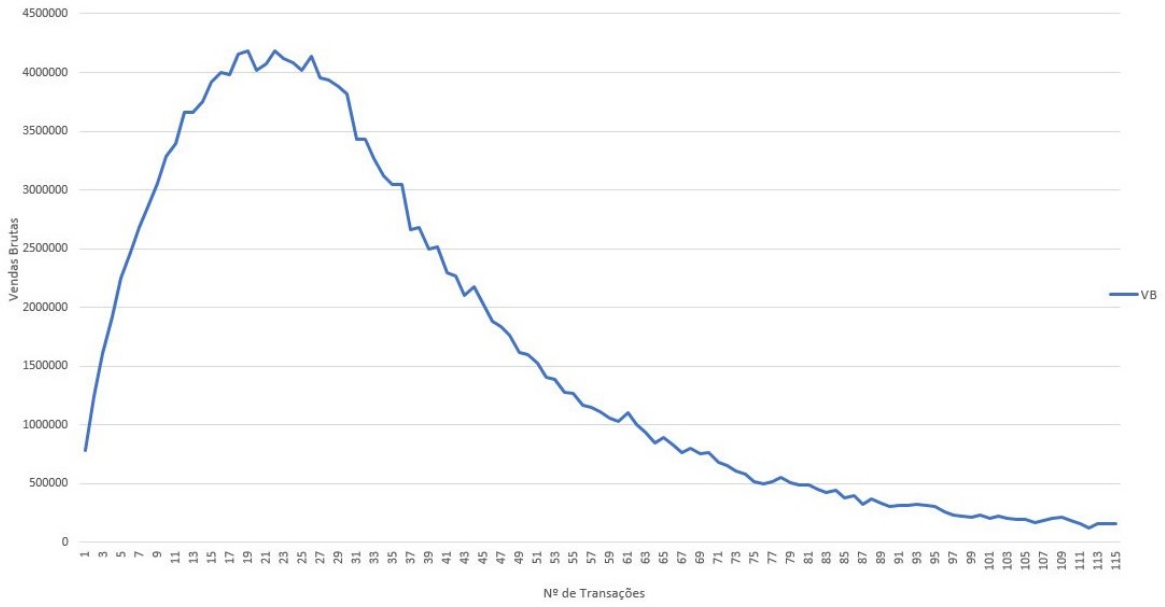

Figura 8: Distribuição das Vendas Brutas pelo número de Transações

Após a análise da Figura 8, optou-se por se remover clientes que tenham efetuado apenas uma ou duas transações, o que corresponde a 13% dos clientes da base, o equivalente a 1% das vendas brutas dos últimos 6 meses de 2017. Removeu-se este número de clientes, uma

vez que, são clientes que não fazem compras de forma recorrente nas lojas Continente e que correspondem a uma percentagem de vendas muito baixa. O número de clientes que se encontrava nestas condições eram 41928, e que foram retirados da base de dados de forma a melhorar aos resultados.

Correlação das Variáveis

É importante analisar a correlação entre as variáveis explicativas. A correlação pode ser calculada usando três coeficientes: Coeficiente de *Pearson*, coeficiente de *Spearman* e coeficiente de *Phi*. Como se pretende calcular a correlação entre variáveis quantitativas, o Coeficiente de *Pearson* é o mais adequado. Para variáveis ordinais deve utilizar-se o Coeficiente de *Spearman* e no caso das variáveis nominais deve utilizar-se o Coeficiente de *Phi* (Maroco (2007)). O coeficiente de correlação entre todas as variáveis foi calculado e está representado, nas Figuras 9 e 10.

	nr_lojas_visitadas	cesta_media	nr_trx	vb_acu	vl_acu	db_cc_acu	db_sp_acu	segm_bj	segm_ps	segm_sow	segm_value	pactive	ltv	segm_ev	segm_nps	perc_dsc
nr_lojas_visitadas	1.00	-0.03	0.35	0.32	0.32	0.27	0.30	-0.14	-0.16	-0.23	-0.33	0.22	0.31	-0.26	-0.06	0.10
cesta_media	-0.03	1.00	-0.11	0.38	0.38	0.34	0.30	-0.08	0.09	-0.16	-0.30	0.01	0.36	-0.18	0.01	0.01
nr_trx	0.35	-0.11	1.00	0.65	0.65	0.41	0.57	-0.13	-0.16	-0.31	-0.53	0.24	0.61	-0.20	-0.02	0.03
vb_acu	0.32	0.38	0.65	1.00	1.00	0.74	0.85	-0.19	-0.09	-0.40	-0.70	0.24	0.97	-0.31	-0.02	0.05
vl_acu	0.32	0.38	0.65	1.00	1.00	0.74	0.85	-0.19	-0.09	-0.40	-0.70	0.24	0.97	-0.30	-0.02	0.05
db_cc_acu	0.27	0.34	0.41	0.74	0.74	1.00	0.69	-0.16	-0.19	-0.34	-0.55	0.20	0.72	-0.21	-0.03	0.26
db_sp_acu	0.30	0.30	0.57	0.85	0.85	0.69	1.00	-0.18	-0.22	-0.38	-0.63	0.22	0.82	-0.24	-0.03	0.30
segm_bj	-0.14	-0.08	-0.13	-0.19	-0.19	-0.16	-0.18	1.00	0.07	0.12	0.19	-0.09	-0.19	0.09	0.03	-0.05
segm_ps	-0.16	0.09	-0.16	-0.09	-0.09	-0.19	-0.22	0.07	1.00	0.15	0.14	-0.14	-0.08	0.04	0.12	-0.35
segm_sow	-0.23	-0.16	-0.31	-0.40	-0.40	-0.34	-0.38	0.12	0.15	1.00	0.53	-0.39	-0.42	0.25	0.06	-0.12
segm_value	-0.33	-0.30	-0.53	-0.70	-0.70	-0.55	-0.63	0.19	0.14	0.53	1.00	-0.47	-0.72	0.32	0.03	-0.10
pactive	0.22	0.01	0.24	0.24	0.24	0.20	0.22	-0.09	-0.14	-0.39	-0.47	1.00	0.34	-0.16	-0.05	0.11
ltv	0.31	0.36	0.61	0.97	0.97	0.72	0.82	-0.19	-0.08	-0.42	-0.72	0.34	1.00	-0.30	-0.01	0.05
segm_ev	-0.26	-0.18	-0.20	-0.31	-0.30	-0.21	-0.24	0.09	0.04	0.25	0.32	-0.16	-0.30	1.00	0.07	-0.02
segm_nps	-0.06	0.01	-0.02	-0.02	-0.02	-0.03	-0.03	0.03	0.12	0.06	0.03	-0.05	-0.01	0.07	1.00	-0.06
perc_dsc	0.10	0.01	0.03	0.05	0.05	0.26	0.30	-0.05	-0.35	-0.12	-0.10	0.11	0.05	-0.02	-0.06	1.00
freq_continente	0.02	0.16	-0.06	0.05	0.05	0.10	0.04	-0.03	0.00	-0.07	-0.04	0.00	0.05	-0.09	-0.02	0.02
freq_bom_dia	0.10	-0.16	0.08	-0.05	-0.05	-0.07	-0.06	0.05	0.02	-0.02	0.04	0.00	-0.05	0.01	0.02	-0.06
freq_modelo	-0.09	-0.03	0.00	-0.01	-0.01	-0.04	0.00	-0.01	-0.01	0.08	0.01	0.00	-0.01	0.07	0.00	0.03
novos_bazar_casa	-0.01	0.00	-0.07	-0.06	-0.07	-0.07	-0.07	0.02	0.05	0.07	0.08	-0.04	-0.07	0.02	-0.01	-0.09
bazar	0.00	-0.02	-0.08	-0.08	-0.09	-0.03	-0.10	-0.09	0.04	0.09	0.10	-0.01	-0.08	0.06	-0.03	-0.08
peixaria_talho	-0.05	0.08	0.04	0.07	0.07	0.08	0.08	0.11	-0.08	-0.06	-0.08	0.04	0.06	0.13	0.02	0.13
alimentar	0.02	0.03	0.00	0.03	0.03	0.01	0.08	-0.07	-0.03	-0.06	-0.05	0.02	0.03	-0.08	-0.01	0.17
textile	0.01	0.01	-0.04	-0.03	-0.04	-0.01	-0.04	-0.08	0.02	0.04	0.04	0.00	-0.03	0.03	0.00	-0.06
nutricao_saudavel	0.05	0.01	0.01	0.03	0.03	0.02	-0.01	0.00	0.06	-0.03	-0.02	0.02	0.03	-0.12	-0.01	-0.07
pad_taway	0.02	-0.11	0.11	0.03	0.04	-0.02	-0.02	0.11	0.02	-0.02	-0.02	-0.02	0.04	-0.09	0.03	-0.19

Figura 9: Correlação de *Pearson* para as Variáveis em estudo

	freq_continente	freq_bom_dia	freq_modelo	novos_bazar_casa	bazar	peixaria_talho	alimentar	textile	nutricao_saudavel	pad_taway
nr_lojas_visitadas	0.02	0.10	-0.09	-0.01	0.00	-0.05	0.02	0.01	0.05	0.02
cesta_media	0.16	-0.16	-0.03	0.00	-0.02	0.08	0.03	0.01	0.01	-0.11
nr_trx	-0.06	0.08	0.00	-0.07	-0.08	0.04	0.00	-0.04	0.01	0.11
vb_acu	0.05	-0.05	-0.01	-0.06	-0.08	0.07	0.03	-0.03	0.03	0.03
vl_acu	0.05	-0.05	-0.01	-0.07	-0.09	0.07	0.03	-0.04	0.03	0.04
db_cc_acu	0.10	-0.07	-0.04	-0.07	-0.03	0.08	0.01	-0.01	0.02	-0.02
db_sp_acu	0.04	-0.06	0.00	-0.07	-0.10	0.08	0.08	-0.04	-0.01	-0.02
segm_bj	-0.03	0.05	-0.01	0.02	-0.09	0.11	-0.07	-0.08	0.00	0.11
segm_ps	0.00	0.02	-0.01	0.05	0.04	-0.08	-0.03	0.02	0.06	0.02
segm_sow	-0.07	-0.02	0.08	0.07	0.09	-0.06	-0.06	0.04	-0.03	-0.02
segm_value	-0.04	0.04	0.01	0.08	0.10	-0.08	-0.05	0.04	-0.02	-0.02
pactive	0.00	0.00	0.00	-0.04	-0.01	0.04	0.02	0.00	0.02	-0.02
ltv	0.05	-0.05	-0.01	-0.07	-0.08	0.06	0.03	-0.03	0.03	0.04
segm_ev	-0.09	0.01	0.07	0.02	0.06	0.13	-0.08	0.03	-0.12	-0.09
segm_nps	-0.02	0.02	0.00	-0.01	-0.03	0.02	-0.01	0.00	-0.01	0.03
perc_dsc	0.02	-0.06	0.03	-0.09	-0.08	0.13	0.17	-0.06	-0.07	-0.19
freq_continente	1.00	-0.29	-0.73	0.06	0.11	-0.10	-0.08	0.19	0.05	-0.04
freq_bom_dia	-0.29	1.00	-0.44	-0.05	-0.12	0.05	0.01	-0.05	0.01	0.11
freq_modelo	-0.73	-0.44	1.00	-0.02	-0.02	0.06	0.07	-0.14	-0.05	-0.04
novos_bazar_casa	0.06	-0.05	-0.02	1.00	-0.01	-0.13	-0.23	0.02	-0.02	-0.13
bazar	0.11	-0.12	-0.02	-0.01	1.00	-0.23	-0.42	0.02	-0.05	-0.26
peixaria_talho	-0.10	0.05	0.06	-0.13	-0.23	1.00	-0.38	-0.09	-0.09	-0.07
alimentar	-0.08	0.01	0.07	-0.23	-0.42	-0.38	1.00	-0.19	-0.08	-0.32
textile	0.19	-0.05	-0.14	0.02	0.02	-0.09	-0.19	1.00	-0.02	-0.10
nutricao_saudavel	0.05	0.01	-0.05	-0.02	-0.05	-0.09	-0.08	-0.02	1.00	0.00
pad_taway	-0.04	0.11	-0.04	-0.13	-0.26	-0.07	-0.32	-0.10	0.00	1.00

Figura 10: Correlação de *Pearson* para as Variáveis em estudo (Cont.)

As matrizes de correlações estão em tons de verde, amarelo e vermelho para ser mais fácil a identificação das correlações mais fortes. Duas variáveis dizem-se positivamente correlacionadas quando o coeficiente é próximo de 1 (verde) e negativamente correlacionadas se próximo de -1 (vermelho). Não há correlação linear quando o valor do coeficiente é aproximadamente 0 (amarelo).

Um dos pressupostos do modelo linear é a não-correlação entre variáveis explicativas. Assim sendo, é necessário selecionar as variáveis mais importantes e que não apresentem correlação elevada, eliminando as variáveis fortemente correlacionadas, correlação superior a $\pm 0,80$ (valor indicado pela empresa). Removeu-se as variáveis correlacionadas entre si com menor importância para o estudo, que são elas as **vl_acu**, **db_sp_acu** e **ltv**.

Desta forma, a base de dados final que será utilizada na modelação contém 283242 clientes e 31 variáveis.

5

RESULTADOS

Neste capítulo são descritos os passos desenvolvidos na construção do modelo de regressão logística, bem como a análise desse modelo.

5.1 ANÁLISE DE CLUSTERS

Numa primeira fase, e de forma a não ter de se restringir a regras de negócio, foram criados *clusters* de clientes por forma a agregar-se os clientes em termos de fidelização e identificar os clientes altamente "*engaged*", isto é, clientes muito envolvidos/fiéis ao Continente, para posteriormente utilizar estes clientes na construção do modelo de regressão logística.

Assim sendo, foram realizadas duas análises de *clusters*: a primeira análise permitiu construir *clusters* das variáveis com o objetivo de reduzir o número de variáveis e a segunda análise permitiu construir *clusters* de clientes com base nas variáveis determinadas na fase inicial, com objetivo de criar grupos de clientes com o mesmo envolvimento com o Continente.

Na Tabela 20, apresentam-se os *clusters* de variáveis. Na construção dos *clusters* de variáveis, as variáveis são selecionadas e vão formando *clusters*, enquanto o algoritmo convergir e não encontrar mais nenhum critério para a divisão das variáveis.

Esta metodologia consiste em criar *clusters*, em que as variáveis dentro de cada *cluster* (em geral, duas variáveis) apresentam uma forte correlação entre elas e uma fraca correlação com as variáveis de outros *clusters*. Por vezes, algumas variáveis que serão fundamentais para o desenvolvimento do modelo, podem não ser "selecionadas" por esta metodologia, podendo ser mantidas, escolhendo-se mais variáveis para um dado *cluster*, como é o caso do *cluster* 3 onde foram selecionadas três variáveis.

Tabela 20: Seleção de Variáveis

		R^2		$1\text{-}R^2$
Cluster	**Variáveis**	**Próprio *Cluster***	***Cluster* Seguinte**	***Ratio***
Cluster 1	**nr_lojas_visitadas**	0,19	0,03	0,83
	nr_trx	0,59	0,08	0,44
	vb_acu	0,85	0,23	0,19
	db_cc_acu	0,62	0,14	0,45
	segm_value	0,69	0,26	0,42
Cluster 2	**novos_bazar_casa**	0,18	0,02	0,84
	bazar	0,56	0,07	0,47
	alimentar	0,49	0,15	0,61
	textil	0,14	0,03	0,89
	missao_4	0,65	0,06	0,37
Cluster 3	**freq_continente**	0,82	0,03	0,19
	freq_modelo	0,84	0,03	0,17
	cliente_insco	0,15	0,02	0,87
Cluster 4	**cesta_media**	0,68	0,10	0,35
	missao_2	0,49	0,07	0,55
	missao_5	0,79	0,12	0,24
Cluster 5	**segm_ps**	0,68	0,01	0,33
	segm_nps	0,06	0,00	0,94
	perc_dsc	0,66	0,03	0,35
Cluster 6	**loja_preferencial**	0,75	0,01	0,25
	compra_online	0,75	0,01	0,25
Cluster 7	**segm_bj**	0,24	0,04	0,78
	freq_bom_dia	0,19	0,02	0,83
	pad_taway	0,52	0,02	0,50
	missao_1	0,47	0,05	0,56
Cluster 8	**peixaria_talho**	0,77	0,03	0,24
	missao_3	0,77	0,02	0,24
Cluster 9	**segm_sow**	0,65	0,19	0,43
	pactive	0,65	0,06	0,38
Cluster 10	**segm_ev**	0,57	0,08	0,47
	nutricao_saudavel	0,57	0,01	0,44

Para a obtenção dos *clusters* de variáveis foi utilizada uma função disponível no SAS que agrupa as variáveis pelos *clusters*, ocorrendo este processo em dois momentos distintos. Numa primeira fase, é utilizado o método do centróide, calculando a distância entre os *clusters*. Em cada iteração, os componentes do *cluster* são calculados e cada variável é incluída no *cluster* com o qual apresenta maior correlação. Na segunda fase do processo, cada variável é testada de forma a confirmar-se se esta pertencer a outro *cluster*, se aumenta o valor da variância explicada. Se uma variável for colocada noutro *cluster* durante esta fase, os componentes dos dois *clusters* envolvidos são recalculados antes da próxima variável ser testada.

A coluna intitulada por "Próprio *Cluster*" indica a correlação da variável com o seu próprio componente de *cluster*. Esse valor deve ser maior que a correlação com qualquer outro *cluster*, sendo que, quanto maior a correlação melhor.

A coluna intitulada de '*Cluster* seguinte" indica o valor da segunda correlação mais alta da variável com um componente de *cluster*. Este valor é baixo se os *clusters* estiverem bem separados. Já a coluna "$1 - R^2$" indica um bom ajustamento quanto menor for esse valor.

Com base nos resultados obtidos, e tendo em conta a importância de cada uma das variáveis para o estudo, foram selecionadas as seguintes variáveis com base nos valores apresentados na Tabela 20:

- ***Cluster*** 1 - **vb_acu**, **db_cc_acu**, **segm_value**;
- ***Cluster*** 2 - **bazar**, **missao_4**;
- ***Cluster*** 3 - **freq_continente**, **freq_modelo**;
- ***Cluster*** 4 - **cesta_media**, **missao_5**;
- ***Cluster*** 5 - **segm_ps**, **perc_dsc**;
- ***Cluster*** 6 - **loja_preferencial**, **compra_online**;
- ***Cluster*** 7 - **pad_taway**, **missao_1**;
- ***Cluster*** 8 - **peixaria_talho**, **missao_3**;
- ***Cluster*** 9 - **segm_sow**, **pactive**;
- ***Cluster*** 10 - **segm_ev**, **nutricao_saudavel**;

Este passo permite reduzir as variáveis em estudo, aumentando a velocidade de processamento sem perder qualidade os resultados, já que todas as variáveis selecionadas são representativas dos *clusters* em que estão inseridas.

5.1.1 *Determinação dos Clientes mais envolvidos com a marca*

Após a seleção das variáveis a utilizar, numa segunda fase, determinam-se quais os clientes mais envolvidos com o Continente, com recurso a uma nova análise de *clusters*, com base nas variáveis selecionadas anteriormente. Na Figura 11 é possível analisar a divisão dos clientes em 14 *clusters* obtidos, aplicando o método de *Ward*.

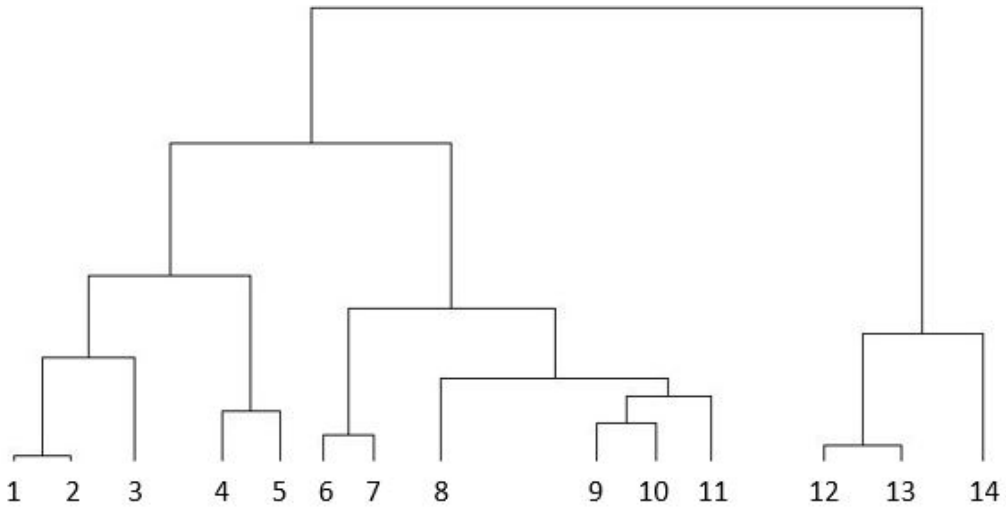

Figura 11: *Clusters* de clientes

De seguida, procede-se à sua caraterização de forma a determinar-se qual(ais) o(s) *Cluster* que agrupa(m) os clientes mais envolvidos com o Continente, com base no peso de cada variável em cada um dos *clusters*.

5.1.1.1 *Caracterização dos Clusters*

Os resultados desta análise são apresentados em cinco tabelas (Tabela 21 a 25). Para facilitar a interpretação dos resultados, as variáveis estão divididas em três grupos, aquelas que contribuem de forma muito positiva (amarelo) no *cluster*, as variáveis que estão numa posição intermédia (laranja) e por fim, as variáveis que pouco contribuem no *cluster* (vermelho). No caso das Tabelas 23, 24 e 25, referentes às segmentações em estudo, as variáveis assumem uma cor amarela se o valor for superior a 100, ou seja, se a variável se destaca pela positiva no *cluster*, caso contrário, assume uma cor vermelha.

A Tabela 21 contém informação transacional dos clientes presentes em cada *cluster*. Os valores[1] correspondem à média de cada variável em cada um dos grupos, destacando-se os

1 Os valores apresentados nas tabelas não correspondem aos valores reais de forma a manter-se a confidencialidade dos dados.

clientes presentes nos *clusters* 1, 4, 7 e 8. Os clientes presentes neste grupo são aqueles que apresentam maior percentagem de vendas face ao número de clientes, são os clientes que com maior cesta média e frequência, e são também clientes com alta probabilidade de continuarem ligados à empresa.

Na tabela 22 é analisada a informação proveniente de variáveis construídas, relativas à frequência de compra quer nas lojas Continente, quer nas lojas Continente Modelo, bem como a frequência com que cada cliente visita cada uma das direções comerciais e ainda as suas motivações de compra. Os resultados são obtidos pelas médias de cada uma das variáveis, onde se destacam os clientes dos *clusters* 1, 4, 7 e 8. São clientes que apresentam maior frequência no Continente e que têm como principal missão de compra a missão 5.

As Tabelas 23, 24 e 25 fazem referência às quatro segmentações em estudo. As Tabelas 23 a 25, indicam o peso de cada categoria em cada um dos *clusters*. Para isso, calculou-se a percentagem de clientes em cada categoria dos vários segmentos sobre a percentagem total de clientes por categoria. Dentro das segmentações, cada categoria que apresente um valor superior a 100, significa que tem uma influência positiva dentro do *cluster*, caso contrário, conclui-se que tem um impacto negativo. Desta forma calculou-se o Índex que resulta da seguinte fórmula,

$$\left(\frac{perc\_segm}{perc\_segm\_total} \times 100\right), \tag{5.1.1}$$

onde **perc_segm** indica a percentagem de clientes em cada categoria para cada um dos segmentos e **perc_segm_total** a percentagem de clientes em cada categoria para cada um dos segmentos por *cluster*. Estas tabelas estão divididas em escalas de duas cores, amarelo para as que contribuem de forma positiva e vermelho para as restantes.

A Tabela 23, mostra em qual dos segmentos da segmentação *Price Sensitivity* é que cada *cluster* se destaca. No caso da Tabela 24 é analisada a segmentação estilo de vida, e qual o impacto que esta segmentação apresenta em cada *cluster*. Nestas duas segmentações, os segmentos não são ordinais, como tal, servem apenas para caraterizar os clientes não se podendo afirmar que um cliente é mais "valioso" só por se encontrar num determinado segmento.

Por fim, para a Tabela 25, é analisada a segmentação valor e a segmentação SOW. Para ambas as segmentações, os *clusters* que mais se destacam nos segmentos de maior valor são os clientes presentes nos *clusters* 1, 4, 6, 7 e 8.

Com base na análise feita às Tabelas 21 a 25, verifica-se que os *clusters* que contêm os clientes que melhor caracterizam o comportamento de um cliente *"engaged"* com o Continente são os *clusters* 1, 4, 7 e 8, uma vez que, são os clientes que apresentam melhores resultados nas várias variáveis e segmentações analisadas.

Tabela 21: Caracterização dos *Clusters* através de variáveis transacionais

Cluster	**%clientes**	**%vb**	**cesta_media**	**freq.**	**vb_acu**	**db_cc_acu**	**pactive**	**perc_dsc**
1	8%	11%	33,54	22,48	761,22	28,83	0,79	14,36
2	13%	5%	20,54	9,69	209,49	10,15	0,77	17,65
3	9%	4%	25,10	9,33	238,23	6,97	0,71	13,67
4	10%	21%	38,27	29,02	1121,07	69,23	0,8	18,97
5	9%	3%	16,46	10,61	187,18	12,48	0,77	18,67
6	12%	15%	25,84	24,19	636,42	31,47	0,80	17,98
7	7%	17%	39,16	26,42	1212,19	67,67	0,80	17,44
8	6%	10%	30,95	24,75	817,75	44,03	0,79	17,79
9	1%	1%	20,62	11,84	270,22	12,71	0,75	13,50
10	5%	2%	17,99	9,27	178,40	10,63	0,76	19,13
11	5%	1%	18,87	6,88	146,74	7,52	0,75	15,03
12	3%	1%	24,28	11,12	274,98	11,22	0,10	16,65
13	6%	3%	23,18	11,15	265,25	13,88	0,77	17,72
14	5%	5%	23,11	20,69	472,06	32,47	0,80	19,76

Tabela 22: Caracterização dos *Clusters* através de variáveis construídas

Cluster	freq_continente	freq_modelo	bazar	peixaria_talho	nutricao_saudavel	pad_taway	missao_1	missao_3	missao_4	missao_5
1	36,06	29,56	7,10	8,82	1,59	16,36	11,67	7,83	14,04	29,17
2	31,16	34,19	5,28	6,90	0,61	15,77	17,26	5,51	15,63	16,09
3	33,96	31,81	8,95	7,53	1,03	15,18	13,74	7,65	19,83	17,25
4	33,30	35,99	6,07	12,27	0,67	14,38	11,42	7,47	10,86	34,77
5	24,21	39,98	6,44	9,78	0,49	14,82	16,70	8,46	15,04	14,41
6	25,42	42,67	6,78	11,07	0,61	14,72	14,07	8,11	13,61	24,96
7	41,78	25,01	6,14	9,78	1,67	16,22	10,87	7,96	11,64	34,33
8	34,86	32,54	7,03	8,93	1,34	15,96	12,33	8,14	13,58	27,95
9	38,85	26,89	5,82	4,20	18,44	13,47	11,46	7,82	14,95	14,18
10	18,13	41,57	3,04	32,34	0,38	12,94	11,40	32,55	10,54	6,83
11	39,82	32,70	34,76	3,13	0,34	6,83	8,90	3,37	46,08	5,05
12	28,99	35,83	5,64	10,01	0,61	13,65	16,97	8,62	13,81	20,48
13	35,96	28,30	6,99	6,99	1,34	16,10	13,89	7,88	16,45	19,33
14	25,02	40,89	5,90	14,63	0,42	14,29	14,74	10,74	11,48	22,39

Tabela 23: Caracterização dos *Clusters* na Segmentação *Price Sensitivity*

Cluster	**Segmentação Price Sensitivity**								
	SEG_PS_1	***SEG_PS_2***	***SEG_PS_3***	***SEG_PS_8***	***SEG_PS_7***	***SEG_PS_5***	***SEG_PS_4***	***Sem Valor***	***SEG_PS_6***
1	2,11	580,93	0	0	0	0	0	0	0
2	252,33	0	71,08	251,68	68,1	55,84	266,5	245,01	0
3	0,15	581,55	0	0	0	0	0	0	0
4	47,23	0,17	271,73	38,95	34,11	255,94	38,59	44,83	8,87
5	0	0	18,04	0	336,25	0,29	0,89	27,63	511,71
6	108,74	0	200,57	114,26	80,01	172,81	125,96	108,68	0
7	44,52	0	206,17	62,54	24,19	310,88	19,48	39,38	4,91
8	104,9	0	174,63	121,96	56,48	225,7	68,69	70,45	9,68
9	211,86	283,94	19,56	124,59	33,3	31,39	114,76	85,07	12,04
10	121,24	2,35	52,45	129,62	225,85	25,99	241,9	213,51	146,41
11	228,16	12,28	19,64	191,97	189,76	18,42	248,13	328,69	40,74
12	169,32	11	77,48	153,43	146,67	71,93	158,48	183,05	103,56
13	223,68	0	88,4	237,9	125,72	74,57	154,76	109,96	35,64
14	0	0	0,05	0	156,88	0	0	0	701,84

Tabela 24: Caracterização dos *Clusters* na Segmentação Estilo de Vida

Cluster	**Segmentação Estilo de Vida**							
	SEG_EV_2	**SEG_EV_4**	**SEG_EV_6**	**SEG_EV_5**	**SEG_EV_1**	**SEG_EV_7**	**SEG_EV_3**	**Sem Valor**
1	264,39	178,65	85,11	42,11	50,48	12,92	38,02	0,78
2	0	0	159,4	164,59	111,08	75,14	204,68	273,47
3	192,02	94,38	96,44	71,61	61,8	22,72	150,01	16,54
4	0,02	1,26	218,38	149,75	195,89	101,27	33,1	51,09
5	0	4,45	45,28	155,71	123,17	329,48	97,79	21,62
6	0	0	160,17	167,36	164,84	168,43	56,43	85,49
7	227,63	425,94	18,22	0	0,03	0	0,4	0
8	303,82	340,03	20,87	0	0	0	0	0
9	261,53	49,18	42,7	30,13	53,39	21,07	217,39	67,81
10	0,83	20,54	30,59	58,69	213,29	78,66	256,45	215,14
11	10,17	20	145,88	115,29	39,7	71,68	326,65	582,69
12	56,63	65,21	101,64	111,86	113,15	128,08	132,13	142,72
13	348,18	307,91	0	0	0	0	0	0
14	0	23,9	71,59	130,68	174,61	280	46,05	2,39

Tabela 25: Caracterização dos *Clusters* na Segmentação Valor e SOW

Cluster	Segmentação Valor								Segmentação SOW				
	SEG_1	**SEG_2**	**SEG_3**	**SEG_4**	**SEG_5**	**SEG_6**	**SEG_7**	**Sem Valor**	***HIGH***	***MEDIUM***	***LOW***	***VERY LOW***	***SEM VALOR***
1	219,79	168,7	178,28	207,99	4,29	7,55	0	0	191,88	164,67	30,22	9,25	73,71
2	0	0	0	0,03	218,52	203,66	206,42	89,42	11,83	65,17	165,29	206,67	108,75
3	0	0	3,55	15,03	156,71	229,52	166,58	235,97	8,18	65,95	158,04	154,44	127,02
4	302,12	313,03	125,28	97,16	0,09	6,42	0,02	0	431,89	32,31	0,18	0,09	6,03
5	0	0	0,36	1,84	269,54	153,57	216,82	84,73	2,8	57,78	209,89	172,73	99,68
6	120,62	145,58	281,6	262,27	0	0,16	0	0	3,8	211,85	61,84	23,48	147,21
7	311,68	310,6	120,73	93,24	0,28	6,16	0	0	447,03	24,69	0	0	0,3
8	180,65	177,47	204,78	230,07	0,04	1,7	0	0	0	202,4	47,27	7,1	166,4
9	15,53	12,45	55,4	71,35	138,47	150,61	183,83	126,2	26,83	85,46	149,25	200,55	100,54
10	0,11	0,15	18,92	18,96	181,63	181,42	217,64	117,27	7,01	58,09	165,83	219,02	111,93
11	0,41	0,07	5,18	21,78	75,71	217,51	248,59	165,07	2,54	13,97	133,2	348,9	124,62
12	7,63	18,92	33,75	15,95	21,46	50,29	75,39	1393,45	2,72	4,99	13,28	17,55	275,79
13	0	0	0	0,06	260,23	216,24	165,3	85,23	22,53	91,03	183,5	111,23	101,27
14	76,38	73,42	288,08	318,17	9,8	5,85	0	0	96,45	263,62	64,61	15,94	64,82

Na Tabela 26 é possível ver-se, de uma forma mais resumida, quais as variáveis que mais se destacam em cada um dos *clusters*.

Tabela 26: Caracterização dos *Clusters*

Clusters	**Variáveis que caracterizam os *Clusters* mais *Engaged***
1	Segm_ps - SEG_PS_2; Segm_ev - SEG_EV_2, SEG_EV_4; Segm_value - SEG_1, SEG_2, SEG_3, SEG_4; Segm_sow - *High*, *Medium*
4	%vb, perc_dsc, freq_modelo, missao_5; Segm_ps - SEG_PS_5, SEG_PS_3; Segm_ev - SEG_EV_6, SEG_EV_5, SEG_EV_1, SEG_EV_7; Segm_value - SEG_1, SEG_2, SEG_3; Segm_sow - *High*
7	cesta_media, Frequência, vb_acu, db_cc_acu, freq_continente; Segm_ps - SEG_PS_5, SEG_PS_3; Segm_ev - SEG_EV_2, SEG_EV_4; Segm_value - SEG_1, SEG_2, SEG_3; Segm_sow - *High*
8	Segm_ps - SEG_PS_1, SEG_PS_3, SEG_PS_8, SEG_PS_5; Segm_ev - SEG_EV_2, SEG_EV_4; Segm_value - SEG_1, SEG_2, SEG_3, SEG_4; Segm_sow - *Medium*, *No Value*

Após a escolha do grupo de clientes, e analisando as variáveis que se destacam nos diversos *clusters* foi criada uma variável, denominada por **cli_engaged** que classifica como 1 os clientes definidos envolvidos com a empresa e 0 os restantes, obtendo-se desta forma a variável resposta a ser utilizada na modelação.

5.2 Regressão Logística

Os clientes nos *clusters* de maior envolvimento com o Continente foram utilizados para a modelação, utilizando técnicas de regressão logística, resultando a probabilidade entre 0 e 1 para cada cliente.

Considere-se a variável dicotómica, cli_engaged, definida da seguinte forma:

$$cli\_engaged = \begin{cases} 1, & \textit{se o cliente está envolvido} \\ 0, & \textit{se o cliente não está envolvido} \end{cases} \tag{5.2.1}$$

Após a preparação da amostra, é crucial esta ser dividida aleatoriamente em duas amostras independentes:

- amostra de treino

- amostra de teste

A amostra de treino é utilizada para construir o modelo, baseada em toda a informação das observações. A amostra de teste permite-nos posteriormente avaliar a precisão do modelo preditivo escolhido. A amostra de treino foi obtida retirando-se aleatoriamente 70% das observações da amostra total, sendo os restantes 30% utilizados para a amostra de teste.

Após ajustar o modelo à amostra de treino, este foi aplicado à amostra de teste por forma a estimar a probabilidade de envolvimento para cada cliente na amostra de treino.

5.2.1 *Interpretação dos Resultados*

Nas Tabelas 27 e 28, apresentam-se as estimativas dos coeficientes do modelo, os respetivos erros padrão, o valor da estatística de teste do teste de *Wald* e o correspondente p-valor.

Tabela 27: Coeficientes do modelo

	Estimativa	Erro Padrão	Teste de Wald	p-valor (Teste de Wald)
	0,2848	0,2499	1,1400	0,2545
vb_acu	0,0012	<2e-16	44,4890	<2e-16
db_cc_acu	0,0059	0,0003	19,2170	<2e-16
segm_value	-1,2880	0,0083	-155,4830	<2e-16
bazar	-0,0052	0,0011	-4,6760	2,93e-06
missao_4	-0,0042	0,0008	-5,2990	1,16e-07
freq_continente	0,0030	0,0003	8,7710	<2e-16
freq_modelo	0,0019	0,0003	6,3320	2,42e-10
cesta_media	-0,0052	0,0005	-9,5760	<2e-16
missao_5	0,0099	0,0005	18,2650	<2e-16
perc_dsc	0,0099	0,0015	6,5450	5,96e-11
pad_taway	0,0028	0,0011	2,6410	0,0083
segm_sow	-1,0350	0,0069	-149,0630	<2e-16
pactive	2,9580	0,1159	25,5090	<2e-16
segm_ps (Ref. Sem valor)	-	-	-	-
(SEG_PS_3)	-0,2580	0,0546	-4,7300	2,25e-06
(SEG_PS_8)	-0,4236	0,0560	-7,5680	3,80e-14
(SEG_PS_6)	-4,3740	0,0757	-57,7980	<2e-16
(SEG_PS_1)	-0,4575	0,0654	-6,9950	2,66e-12
(SEG_PS_7)	-0,7709	0,0605	-12,7420	<2e-16
(SEG_PS_5)	-0,2898	0,0544	-5,3280	9,91e-08
(SEG_PS_4)	-0,2754	0,0720	-3,8260	0,0001
(SEG_PS_2)	1,5110	0,0534	28,3130	<2e-16

Tabela 28: Coeficientes do modelo (Cont.)

	Estimativa	Erro Padrão	Teste de Wald	p-valor (Teste de Wald)
segm_ev (Ref. Sem valor)	-	-	-	-
(SEG_EV_2)	4,4620	0,2190	20,3790	<2e-16
(SEG_EV_4)	4,6850	0,2191	21,3800	<2e-16
(SEG_EV_6)	1,3980	0,2186	6,3980	1,58e-10
(SEG_EV_5)	0,7189	0,2186	3,2890	0,0010
(SEG_EV_1)	0,7160	0,2182	3,2820	0,0010
(SEG_EV_7)	-0,0927	0,2194	-0,4220	0,6727
(SEG_EV_3)	1,4350	0,2203	6,5130	7,36e-11

Testou-se os três métodos de seleção de variáveis (*backward*, *forward* e *stepwise*), sendo que o modelo obtido pelos três métodos é o mesmo,

$$\begin{aligned} logit\,(p_i) = {} & \beta_0 + \beta_1 vb\_acu + \beta_2 db\_cc\_acu - \beta_3 segm\_value - \beta_4 bazar \\ & - \beta_5 missao\_4 + \beta_6 freq\_continente + \beta_7 freq\_modelo - \beta_8 cesta\_media \\ & - \beta_9 missao\_5 + \beta_{10} perc\_dsc + \beta_{11} pad\_taway - \beta_{12} segm\_sow + \beta_{13} pactive \\ & - \beta_{14} SEG\_PS\_3 - \beta_{15} SEG\_PS\_8 - \beta_{16} SEG\_PS\_6 - \beta_{17} SEG\_PS\_1 \\ & - \beta_{18} SEG\_PS\_7 - \beta_{19} SEG\_PS\_5 - \beta_{20} SEG\_PS\_4 + \beta_{21} SEG\_PS\_2 \\ & + \beta_{22} SEG\_EV\_2 + \beta_{23} SEG\_EV\_4 + \beta_{24} SEG\_EV\_6 + \beta_{25} SEG\_EV\_5 \\ & + \beta_{26} SEG\_EV\_1 - \beta_{27} SEG\_EV\_7 + \beta_{28} SEG\_EV\_3, \end{aligned}$$

onde p_i é a probabilidade do cliente estar envolvido com o Continente.

De forma a analisar-se os resultados, e de perceber o impacto das variáveis no modelo, foram analisados os valores da razão de riscos, como mostra os resultados da Tabelas 29 e 30.

Tabela 29: Razão de riscos (*Odds Ratio*)

	Razão de Chances	Limite Inferior	Limite Superior	p-valor
vb_acu	1,001	1,001	1,001	<2e-16
db_cc_acu	1,006	1,005	1,007	<2e-16
segm_value	0,276	0,271	0,280	<2e-16
bazar	0,995	0,993	0,997	<2e-16
missao_4	0,996	0,994	0,997	<2e-16
freq_continente	1,003	1,002	1,004	<2e-16
freq_modelo	1,002	1,001	1,003	<2e-16
cesta_media	0,995	0,994	0,996	<2e-16
missao_5	1,010	1,009	1,011	<2e-16
perc_dsc	1,010	1,007	1,013	<2e-16

Tabela 30: Razão de riscos (*Odds Ratio*) (Cont.)

	Razão de Chances	Limite Inferior	Limite Superior	p-valor
pad_taway	1,003	1,001	1,005	0.008
segm_sow	0,355	0,350	0,360	<2e-16
pactive	19,254	15,340	24,167	<2e-16
segm_ps (Ref. Sem valor)	-	-	-	-
(SEG_PS_3)	0,773	0,694	0,860	<2e-16
(SEG_PS_8)	0,655	0,587	0,731	<2e-16
(SEG_PS_6)	0,013	0,011	0,015	<2e-16
(SEG_PS_1)	0,633	0,557	0,719	<2e-16
(SEG_PS_7)	0,463	0,411	0,521	<2e-16
(SEG_PS_5)	0,748	0,673	0,833	<2e-16
(SEG_PS_4)	0,759	0,659	0,874	<2e-16
(SEG_PS_2)	4,533	4,083	5,033	<2e-16
segm_ev (Ref. Sem Valor)	-	-	-	-
(SEG_EV_2)	86,655	56,419	133,096	<2e-16
(SEG_EV_4)	108,339	70,510	166,465	<2e-16
(SEG_EV_6)	4,049	2,638	6,214	<2e-16
(SEG_EV_5)	2,052	1,337	3,150	0,001
(SEG_EV_1)	2,046	1,334	3,138	0,001
(SEG_EV_7)	0,911	0,593	1,401	0,673
(SEG_EV_3)	4,198	2,726	6,464	<2e-16

Do modelo obtido, pode-se concluir que:

- a cada aumento unitário nas vendas brutas, aumenta a chance de envolvimento do cliente em 0,1%;
- a cada aumento unitário no valor dos descontos acumulados em cartão, aumenta a chance de envolvimento do cliente em 0,6%;
- pelo aumento do valor do segmento valor do cliente, diminui a chance de ser envolvido em 72,4%;
- por cada aumento da percentagem de vendas no departamento comercial bazar, diminui a chance de ser envolvido em 0,5%;
- por cada aumento unitário da frequência relativa de transações na missão de compra 4, diminui a chance do cliente estar envolvido em 0,4%;
- por cada aumento unitário da frequência relativa de transações no Continente, aumenta a chance do cliente estar envolvido em 0,3%;
- por cada aumento unitário da frequência relativa de transações no Continente Modelo, aumenta a chance do cliente estar envolvido em 0,2%;

- por cada aumento unitário da cesta média, diminui a chance de estar envolvido em 0,5%;
- por cada aumento unitário da frequência relativa de transações na missão de compra 5, aumenta a chance de estar envolvido em 1%;
- por cada aumento unitário da percentagem de descontos, aumenta a chance de envolvimento em 1%;
- por cada aumento unitário nas vendas brutas relativas no departamento comercial peixaria & talho, aumenta a chance em 0,3%;
- pelo aumento do valor do segmento SOW, diminui a chance de estar envolvido em 64,5%;
- por cada aumento unitário da probabilidade do cliente estar ativo com a marca, aumenta em 19,254 vezes;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_3 é 22,7% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_8 é 34,5% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_6 é 98,7% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_1 é 36,7% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_7 é 53,3% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_5 é 25,2% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;
- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_4 é 24,1% inferior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_PS_2 é 5,533 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_EV_2 é 86,655 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_EV_4 é 108,339 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_EV_6 é 4,049 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_EV_5 é 2,052 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_EV_1 é 2,046 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação;

- a chance de estar envolvido com a empresa um cliente do estilo de vida SEG_EV_3 é 4,198 vezes superior que a chance de estar envolvido um cliente que não pertença a esta segmentação.

5.2.2 *Teste de Hosmer Lemeshow*

Inicialmente foi avaliada a qualidade de ajustamento do modelo, aplicando-se o método de *Hosmer Lemeshow*, obtendo-se um p-valor < 2,2e-16, rejeitando-se a hipótese de que o modelo se ajusta bem aos dados. Estes resultados podem ser influenciados pelo facto de estarmos perante dados de grandes dimensões, sendo essa uma das limitações da análise do p-valor.

Teste de *McFadden*

Foi também calculado o teste de *McFadden* que com um resultado de 0,7465 ou seja, o modelo explica 74,65% da variabilidade da variável resposta.

5.2.3 *Erro de Predição*

Na Tabela 31 apresentam-se os valores observados e os valores ajustados do modelo, quando se definiu um ponto de corte de 0,5.

Tabela 31: Resultado para valores observados e valores ajustados do modelo

Valores observados	**Valores ajustados**	**% CLIENTES**
0	0	66%
0	1	3%
1	0	4%
1	1	27%

Da Tabela 31, é possível analisar a percentagem de clientes bem classificados. Apenas 4% dos clientes que são classificados como estando envolvidos com o Continente são classificados pelo modelo como clientes não envolvidos. Por outro lado, são apenas 3% dos clientes que são classificados como não estando envolvidos com o Continente e que o modelo classifica como sendo clientes envolvidos com o Continente. Como a obtenção destes resultados é possível obter-se a taxa de acerto do modelo (*accuracy*), a especificidade e sensibilidade. Verifica-se que 92,7% dos indivíduos estão bem classificados, sendo a sensibilidade e a especificidade 87% e 95,3% respetivamente.

6

DISCUSSÃO E ANÁLISE DE RESULTADOS

Após a estimação do modelo de regressão logística, estimou-se para cada cliente, qual a probabilidade de estar envolvido com a marca.

A métrica para medir o envolvimento de cada cliente foi calculado usando a probabilidade do cliente envolvido com a marca $\times$ 100%. Os clientes foram divididos em cinco segmentos, tendo em cada um deles a mesma percentagem de clientes. Os segmentos foram caracterizados por *Very Low*, *Low*, *Medium*, *High* e *Very High*.

Tabela 32: Segmentos de clientes

Segmento	***Intervalo***	***%clientes***
Very Low	$[0\%, 33\%]$	20%
Low	$]33\%, 45\%]$	20%
Medium	$]45\%, 49\%]$	20%
High	$]49\%, 65\%]$	20%
Very High	$]65\%, 100\%]$	20%

Pela análise da Tabela 32, o segmento *Very Low*, é caraterizado por clientes que tem uma probabilidade de envolvimento com o Continente entre o 0% e os 33%. Quanto mais elevado for a categoria dos segmentos, maior é a probabilidade dos clientes estarem envolvidos com o Continente, estando no segmento *Very High* apenas clientes com uma métrica de envolvimento superior a 65%.

Após a divisão dos clientes pelos cinco segmentos, é importante perceber as principais características dos clientes, e o que os distingue, e dessa forma estando mais perto de tomar ações mais direcionadas para aumentar a fidelidade dos clientes para com a marca. Como tal, neste capítulo os segmentos serão analisados e caracterizados.

6.1 CARACTERIZAÇÃO DOS SEGMENTOS

Depois de criados os segmentos é fundamental conhecer melhor cada um deles.

	VERY LOW	LOW	MEDIUM	HIGH	VERY HIGH
% de VB e Clientes	6% 20%	9% 20%	13% 20%	28% 20%	44% 20%
SOW	*Very Low*[<6%]		*Medium*[25%<*SOW*<57%]		*High*[>57%]
Missões de compra	Missao_4 Missao_3	Missao_4 Missao_2	Missao_2 Missao_1	Missao_5 Missao_1	Missao_5

Figura 12: Quadro resumo dos segmentos

A Figura 12, permite retirar as primeiras conclusões sobre os cinco segmentos criados. Embora os segmentos apresentem a mesma percentagem de clientes (20%), apresentam grandes diferenças nas percentagens de vendas. No segmento *Very Low* temos 20% de clientes o que corresponde a 6% das vendas, enquanto no segmento *Very High*, os mesmos 20% de clientes corresponde a 44% das vendas totais do Continente. Relativamente à segmentação SOW, quanto mais alto o nível de envolvimento do cliente maior o valor nesta segmentação. Outra das métricas existente, são as missões de compra, que se baseia no registo de compras e que para os segmentos com menor envolvimento, essas compras tem como único propósito, suprir faltas que possam acontecer, ou são compras orientadas em que o cliente vai apenas com o objetivo de comprar uma lista de artigos. Nos segmentos com maior afinidade, usam as lojas Continente para abastecimento, fazendo a maioria das suas compras.

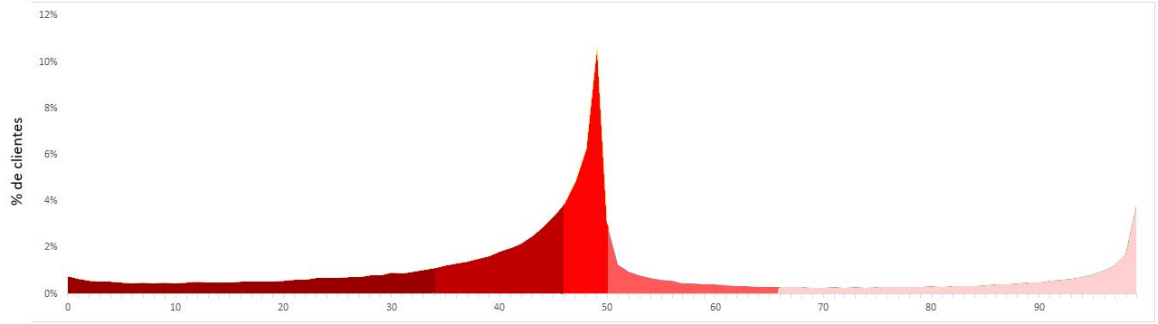

Figura 13: Distribuição dos clientes por percentagem de envolvimento

A Figura 13, mostra a distribuição de clientes por percentagem de envolvimento, com as diferentes cores a serem representativas dos cinco segmentos criados. É também visível pela figura, que há um maior número de clientes com a probabilidade de envolvimento entre os 0,4 e os 0,5, e que são mais os clientes com probabilidades de envolvimento mais próximas de 1 do que os clientes com probabilidade próximas de 0.

Aprofundando um pouco mais os resultados das Figura 14 e 15, e percebendo qual o impacto que cada grupo de clientes tem nas lojas Continente, Continente Bom Dia e Continente Modelo, foram analisados cada um dos segmentos em maior detalhe.

6.1.1 *Segmento Very Low*

Para os vários segmentos é fundamental perceber a recorrência com que estes clientes vão a Lojas Continente e qual o valor gasto em média por transação.

Figura 14: *Drivers* de compra

A Figura 14 indica que os clientes visitam as lojas em média 0,75 vezes por semana com uma cesta média de 23€.

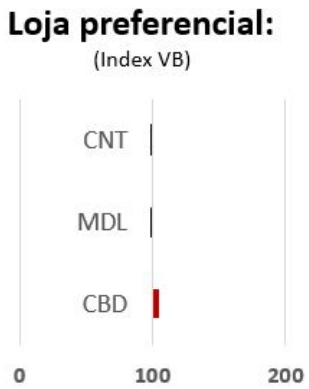

Figura 15: Loja preferencial

Segundo indica a Figura 15, são clientes que visitam preferencialmente lojas Continente Bom Dia.

Depois de uma primeira análise relativa ao comportamento do cliente em loja, é importante conhecer com maior detalhe os clientes.

Figura 16: Segmento Estilo de Vida

A Figura 16, indica que o tipo de clientes que se destaca neste segmento, são os clientes do SEG_EV_3, que são caracterizados por serem clientes de todas as idades, inseridos em agregados familiares de 1 ou 2 elementos. São clientes que estão espalhados um pouco por todo o país e para os quais as plataformas tecnológicas do Continente não são relevantes. São clientes que dão grande importância a promoções.

Figura 17: Segmento idade

Pela Figura 17, temos que os clientes deste segmento tem idade superior a 55 anos e que não têm filhos de idade bebé ou júnior no seu agregado familiar.

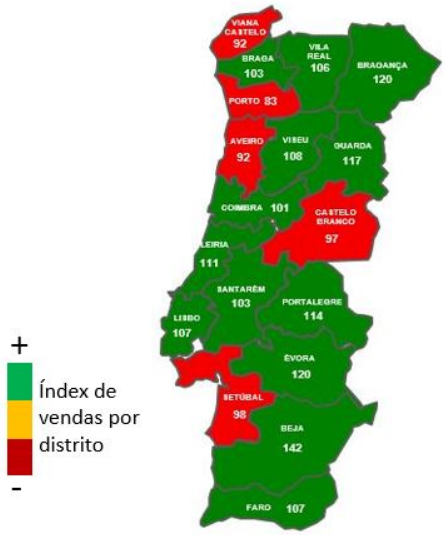

Figura 18: Distribuição geográfica

São clientes que estão espalhados um pouco por todo o país como mostra a Figura 18.

É possível ainda ver o comportamento dos clientes nas restantes marcas do ecossistema, indicando quais as marcas que os clientes mais se destacam, quer pela positiva, quer pela negativa.

Figura 19: Principais marcas

Com base na Figura 19, verifica-se que os clientes deste segmento, são clientes que no ecossistema visitam mais marcas como o Meu Super, mas que por outro lado frequentam pouco a BAGGA e a Galp. Os clientes presentes neste segmento destacam-se por serem clientes que frequentam muito pouco as restantes marcas do ecossistema.

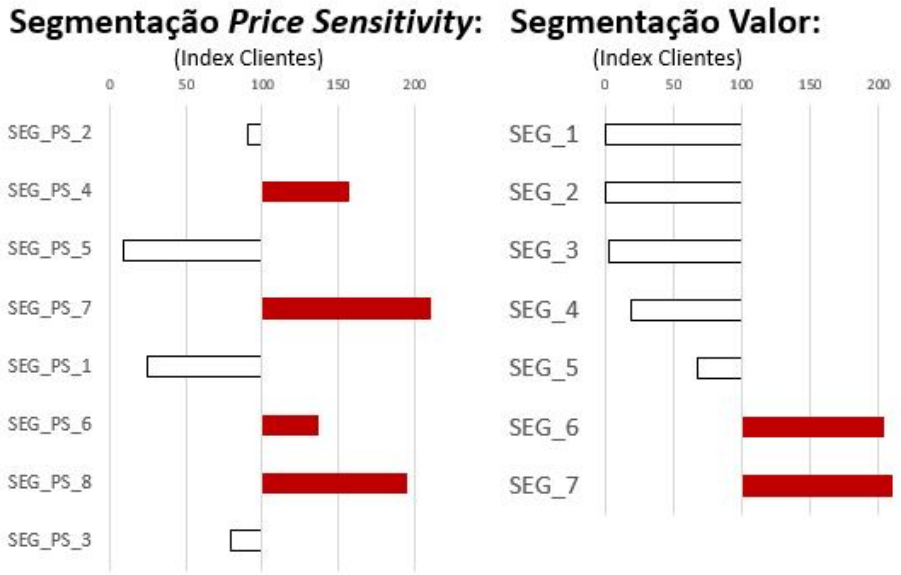

Figura 20: Segmentação *Price Sensitivity* e Valor

A Figura 20, contém informação relativa às Segmentações *Price Sensitivity* e Valor. Relativamente à Segmentação *Price Sensitivity* destacam-se os clientes nas categorias SEG_PS_4, isto é, clientes pouco frequentes e com baixa sensibilidade ao preço. Tendem a comprar produtos de marca própria em categorias específicas e produtos de marca de fornecedores em outras categorias. Destacam-se também na categoria SEG_PS_7 que é caracterizado por clientes com alta penetração de marca própria, baixa atividade promocional e baixa variedade de marcas. São também caracterizados por serem clientes SEG_PS_6 com alta penetração de marca própria, baixa variedade, mas alta atividade promocional com destaque para as marcas de fornecedores. Por último, tem também relevância na categoria SEG_PS_8, o que indica que são clientes com penetração média de marca própria e sem elasticidade, embora tenham uma variedade média, têm uma atividade promocional muito alta. Estes clientes tendem a ser sensíveis a preços / promoções. Relativamente à Segmentação Valor, destacam-se nas duas categorias mais baixas, ou seja, com menor valor.

6.1.2 ***Segmento Low***

Figura 21: *Drivers* de compra

Pela Figura 21, conclui-se que os clientes tem uma frequência semanal de 1,1 vezes o que corresponde a uma cesta média de 23€.

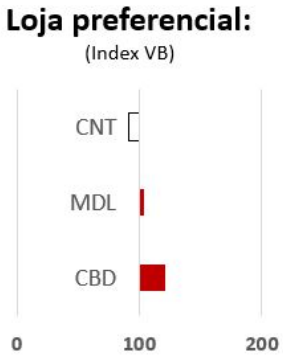

Figura 22: Loja preferencial

A Figura 22 indica que os clientes visitam preferencialmente lojas Continente Bom Dia, sendo também representativos em lojas Continente Modelo.

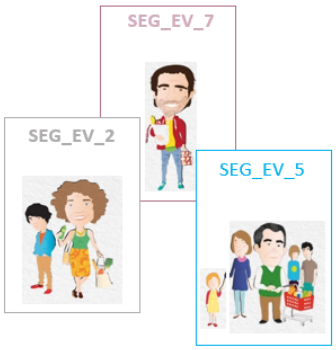

Figura 23: Segmento Estilo de Vida

Passando agora a uma caracterização mais específica de cada grupo de clientes, a Figura 23, indica que o tipo de clientes que se destaca neste segmento, são os clientes do SEG_EV_7, que são caracterizados por serem jovens inseridos em grandes agregados familiares. Encontram-se um pouco por todo o país, sobretudo no interior e norte do país. São também eles clientes que não se destacam pelo uso das plataformas tecnológicas do Continente, dando grande importância ao preço. Destacam-se também no SEG_EV_2, sendo clientes tendencialmente a partir dos 45 anos, inseridos em agregados familiares pequenos de uma ou duas pessoas. Encontram-se mais no distritos do Porto e Lisboa, e são os clientes que mais usam o Continente Online e também usam regularmente a *APP* Continente no telemóvel. São clientes que se focam essencialmente na variedade e qualidade dos frescos. Por último, destaca-se também o SEG_EV_5, que são clientes de todas as idades, mas com maior probabilidade de terem menos 65 anos, inseridos em agregados familiares médios ou grandes. Encontram-se mais nos distritos de Leiria, Santarém e no norte do país, não se destacando pelo uso das plataformas tecnológicas do Continente. As suas principais motivações de compra estão baseadas na variedade e promoção.

Figura 24: Segmento idade

Na Figura 24 verifica-se que os clientes deste segmento, tem idade abaixo dos 25 anos ou superior a 65 anos e que não contam com filhos de idade bebé ou júnior no seu agregado familiar.

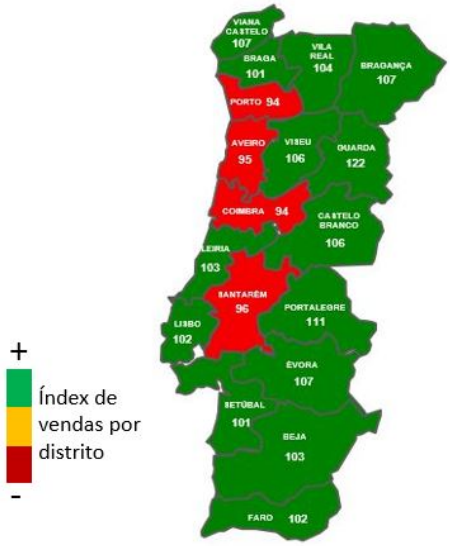

Figura 25: Distribuição geográfica

São clientes com representatividade visível em todo o país como mostra a Figura 25.

Figura 26: Principais marcas

Sobre as marcas do ecossistema que os clientes mais visitam, destacam-se as marcas Meu Super e o KFC, mas que por outro lado visitam menos marcas como a BAGGA e a Galp, como mostra a Figura 26.

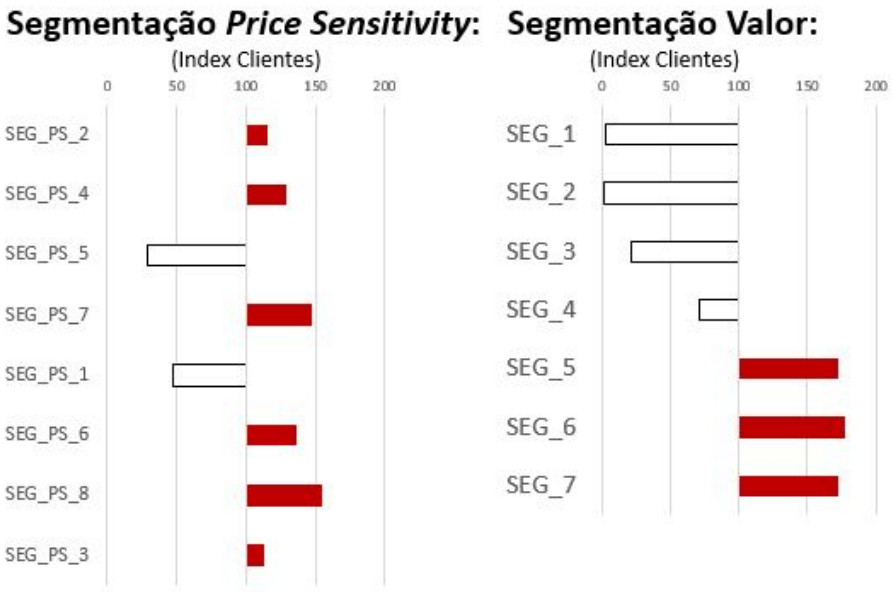

Figura 27: Segmentação *Price Sensitivity* e Valor

A Figura 27, contém informação relativa às Segmentações *Price Sensitivity* e Valor. Relativamente à Segmentação *Price Sensitivity* destacam-se os clientes nas categorias SEG_PS_2, que representa clientes com baixa penetração de marca própria e baixa atividade promocional, mas alto número de marcas diferentes compradas, se suma, são clientes que tendem a não ser sensíveis ao preço. Destacam-se também na categoria SEG_PS_3, que é caracterizado por clientes sensíveis ao preço, que têm uma grande variedade de marcas e uma penetração média de marca própria, sendo muito sensíveis ao preço. Estes clientes tal como no segmento *Very Low*, também se destacam nas categorias SEG_PS_4, SEG_PS_7, SEG_PS_6 e SEG_PS_8. No que diz respeito à Segmentação Valor, destacam-se nas três categorias com menor valor.

6.1.3 *Segmento Medium*

Figura 28: *Drivers* de compra

A Figura 28 indica que os clientes visitam as lojas em média 1,7 vezes por semana com uma cesta média de 22€.

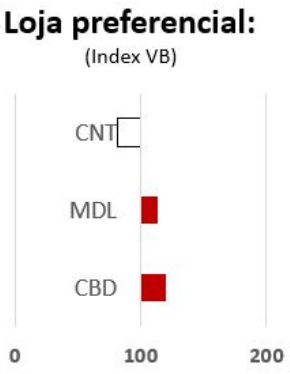

Figura 29: Loja preferencial

Segundo indica a Figura 29, são clientes que visitam preferencialmente as lojas Continente Bom Dia e Continente Modelo.

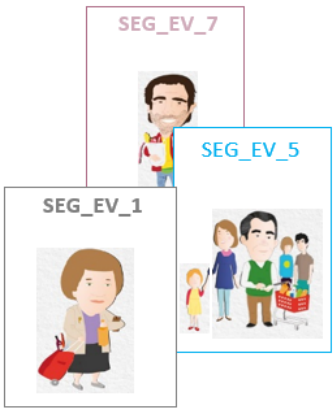

Figura 30: Segmento Estilo de Vida

A Figura 30, indica que o tipo de clientes que se destaca neste segmento, são os clientes do SEG_EV_1, que são caracterizados por serem clientes com idade e que estão inseridos em agregados familiares de uma ou duas pessoas. São o grupo de clientes que menos usam as plataformas tecnológicas e encontram-se sobretudo no Norte e Centro do País. Também se destacam os segmentos SEG_EV_7 e SEG_EV_5, que já foram explicados anteriormente.

Figura 31: Segmento idade

Pela Figura 31, temos que os clientes deste segmento tem idade compreendida entre os 18 e os 35 anos, com bebés no agregado familiar.

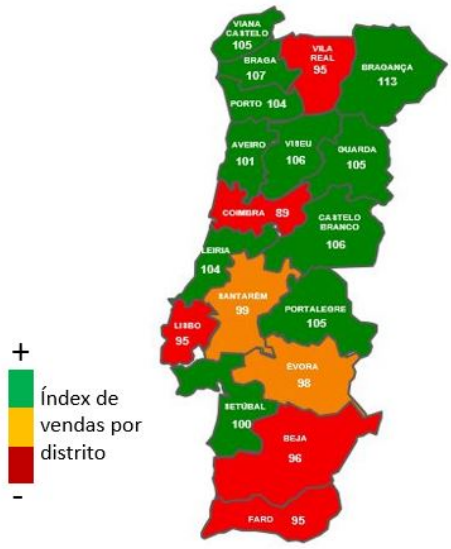

Figura 32: Distribuição geográfica

São clientes com maior representatividade nas zonas norte e centro do país como mostra a Figura 32.

Figura 33: Principais marcas

Na Figura 33, verifica-se que os clientes deste segmento, são clientes que no ecossistema visitam mais marcas como o Burguer King e o KFC, mas que por outro lado se destacam pela negativa em marcas como a NOTE! e a Zippy.

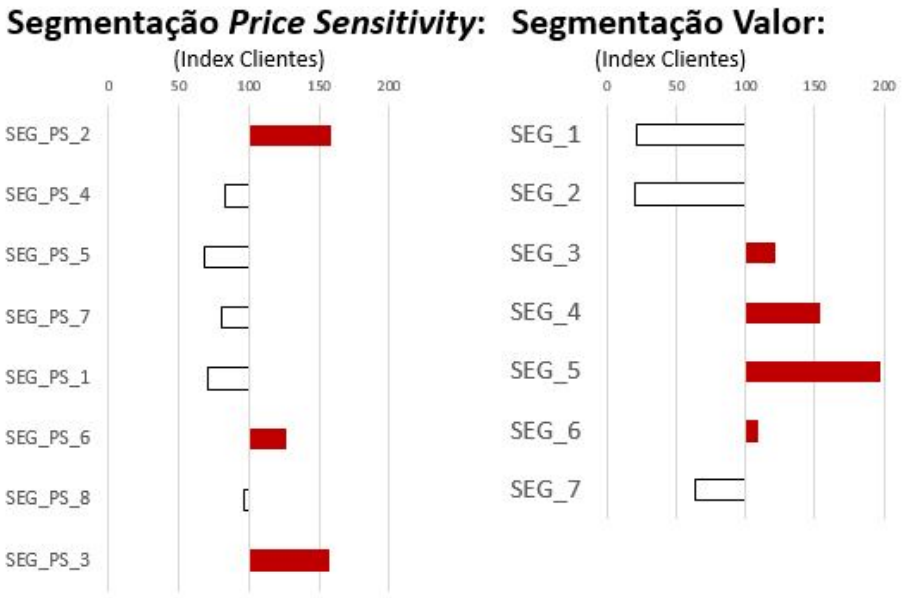

Figura 34: Segmentação *Price Sensitivity* e Valor

A Figura 34, contém informação relativa às Segmentações *Price Sensitivity* e Valor. Relativamente à Segmentação *Price Sensitivity* destacam-se os clientes nas categorias SEG_PS_2, SEG_PS_6 e SEG_PS_3. No caso da Segmentação Valor, destacam-se nas categorias intermédias, com maior peso no SEG_5, mas com representatividade de clientes com elevado valor.

6.1.4 *Segmento High*

Figura 35: *Drivers* de compra

A Figura 35 indica que os clientes visitam as lojas em média 2,6 vezes por semana com uma cesta média de 30€.

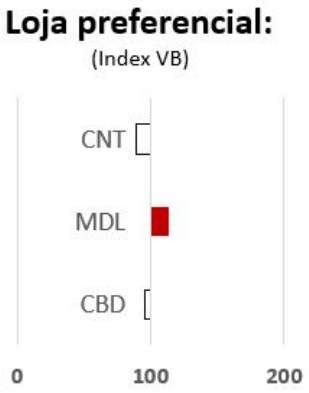

Figura 36: Loja preferencial

Segundo indica a Figura 36, são clientes que visitam preferencialmente lojas Continente Modelo.

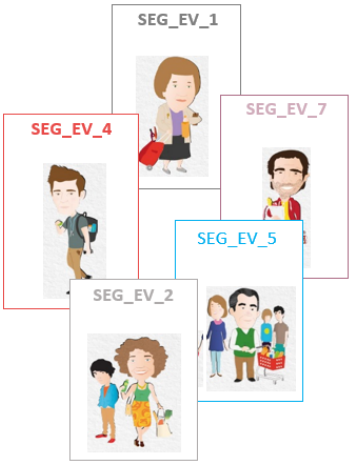

Figura 37: Segmento Estilo de Vida

A Figura 37, indica que o tipo de clientes que se destaca neste segmento, são os clientes do SEG_EV_4, que tendencialmente são pessoas mais novas entre os 25 e os 45 anos, inseridos em agregados familiares pequenos (uma ou duas pessoas). Encontram-se na sua maioria na Grande Lisboa e Alentejo. São clientes com elevada apetência tecnológica e demonstram-no na relevância que atribuem ao Continente Online e à *APP* do Continente no telemóvel. Em relação ás motivações de compra, são clientes que afirmam que não prescindem de qualidade na alimentação, tendo o preço um peso secundário na hora da escolha dos produtos. Também se destacam os clientes do SEG_EV_1, SEG_EV_2, SEG_EV_5 e SEG_EV_7.

Figura 38: Segmento idade

Pela Figura 38, temos que os clientes deste segmento tem idade entre os 45 e os 65, com filhos bebés e juniores no agregado familiar.

Figura 39: Distribuição geográfica

São clientes que se situam nos distritos de Braga, Porto, Aveiro e Setúbal como mostra a Figura 39.

Figura 40: Principais marcas

Com base na Figura 40, verifica-se que os clientes deste segmento, são clientes que no ecossistema visitam mais marcas como a Galp e a BAGGA, mas que por outro lado visitam menos marcas como a Meu Super e a Zippy.

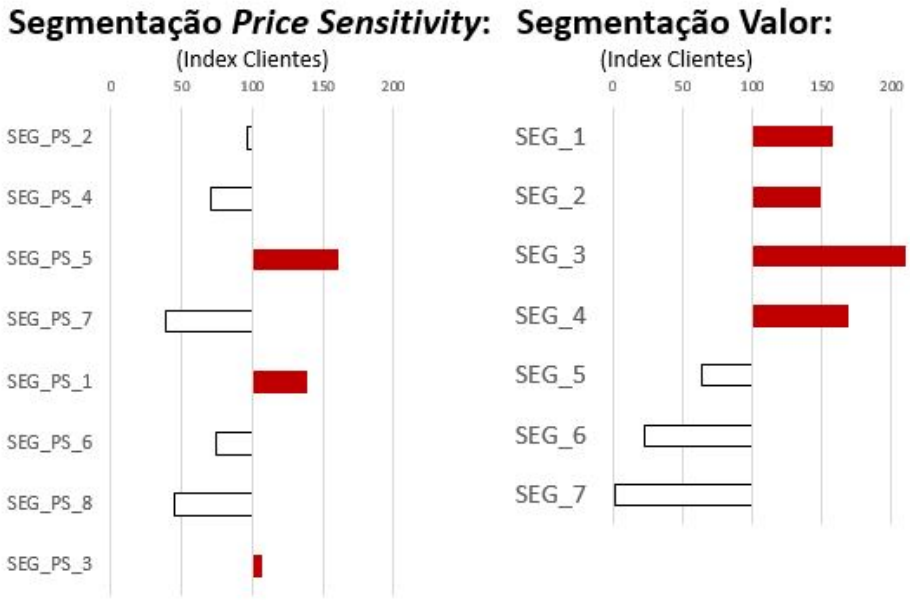

Figura 41: Segmentação *Price Sensitivity* e Valor

A Figura 41, contém informação relativa às segmentações *Price Sensitivity* e Valor. Relativamente à Segmentação *Price Sensitivity* destacam-se os clientes nas categorias SEG_PS_5, SEG_PS_1 e SEG_PS_3. Na segmentação Valor, destacam-se nas quatro categorias mais altas, com maior destaque para no SEG_3.

6.1.5 *Segmento Very High*

Figura 42: *Drivers* de compra

A Figura 42 indica que os clientes visitam as lojas em média 3,1 vezes por semana com uma cesta média de 39€.

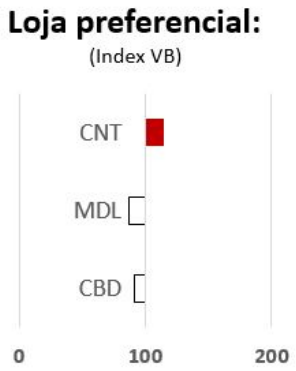

Figura 43: Loja preferencial

Segundo indica a Figura 43, são clientes que visitam preferencialmente lojas Continente.

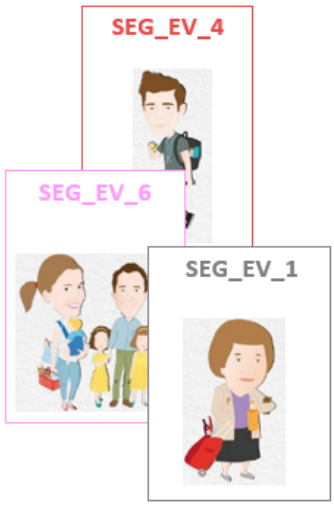

Figura 44: Segmento Estilo de Vida

A Figura 44, indica que se destaca o segmento SEG_EV_6, que é geralmente caracterizado por clientes até aos 46 anos, com agregados familiares médios ou grandes. Encontram-se mais no norte do país e em Beja, usando a *APP* do Continente no telemóvel com frequência. São clientes em que as suas motivações de compra se prendem na conveniência, variedade e promoção. Os segmentos SEG_EV_4 e SEG_EV_1, também se destacam neste segmento.

Figura 45: Segmento idade

Pela Figura 45, temos que os clientes deste segmento tem idade entre os 25 e os 55 anos, tendo filhos de idade bebé e júnior no seu agregado familiar.

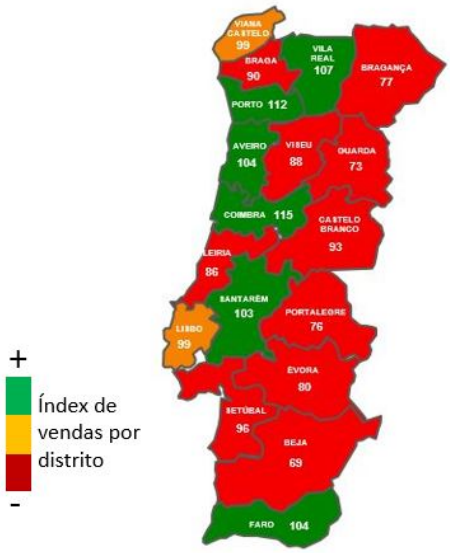

Figura 46: Distribuição geográfica

São clientes que se situam mais na zona do litoral do país com maior destaque para os distritos do Coimbra, Porto e Vila Real, como mostra a Figura 46.

Figura 47: Principais marcas

Com base na Figura 47, verifica-se que os clientes deste segmento, são clientes que no ecossistema visitam mais marcas como a Galp e a NOTE!, mas que por outro lado e atendendo ao seu estilo de vida e localização geográfica, destacam-se pela negativa em marcas como o Meu Super e o KFC.

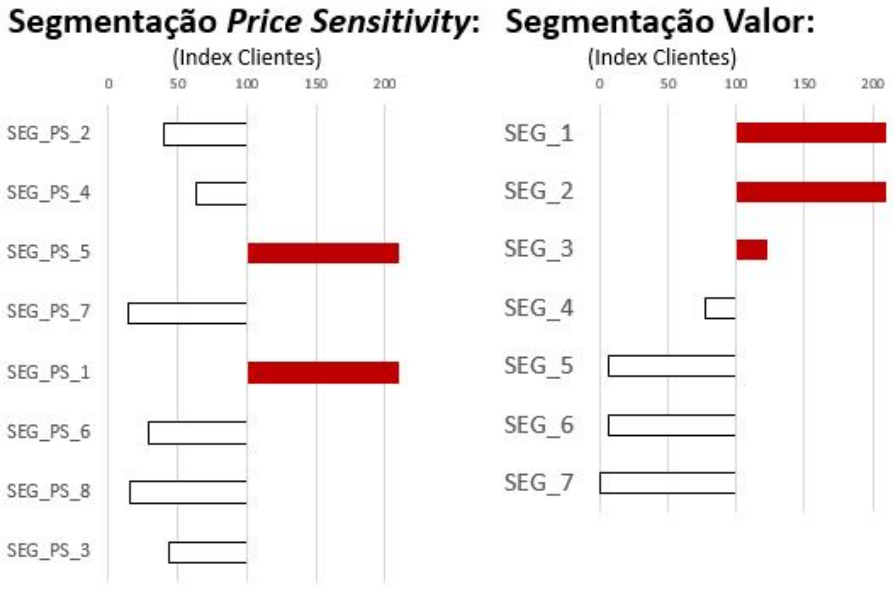

Figura 48: Segmentação *Price Sensitivity* e Valor

A Figura 48, contém informação relativa às Segmentações *Price Sensitivity* e Valor. Relativamente à Segmentação *Price Sensitivity* destacam-se os clientes nas categorias SEG_PS_5 e SEG_PS_1. Por fim, para a segmentação Valor, destacam-se nas três categorias mais elevadas, com maior destaque para as duas de maior valor.

A caracterização dos clientes pelos segmentos permite conhecer um pouco melhor cada um dos segmentos criados, validando os resultados obtidos anteriormente. Após a análise em maior detalhe, percebe-se que à medida que os clientes se situam em segmentos de maior fidelização, melhores são os resultados para o Continente, desde as segmentações, motivações de compra e as vendas associadas.

7

CONCLUSÃO

Este trabalho teve como principal objetivo calcular o nível de envolvimento de cada cliente com o Continente. Apesar das várias métricas já criadas na empresa, o principal objetivo passava pela criação de uma métrica única, que permita perceber o grau de envolvimento do cliente sem ter de se cruzar informação.

O modelo final, foi selecionado tendo por base um modelo de regressão logística desenvolvido. A escolha do modelo foi justificada pelos indicadores de desempenho e interpretabilidade dos resultados.

Do modelo final selecionado pode verificar-se que as variáveis mais importantes para a fidelização de um cliente são as segmentações SOW, *Price Sensitivity* e Valor, tendo a variável que corresponde ao valor gasto pelo cliente, um impacto pequeno. O modelo está apto e pronto a ser utilizado pela empresa.

Os resultados obtidos na modelação, na determinação dos *scores*, permitiram perceber com maior detalhe qual o tipo de clientes que a empresa tem, a sua capacidade de fidelizar clientes num mercado com tanta concorrência e desta forma repensar possíveis estratégias de *marketing*.

7.1 LIMITAÇÕES

O acesso aos dados e respetivo tratamento correram da forma pretendida, o que desencadeou que todos os processos que se seguiram terem os resultados pretendidos.

Contudo, uma das limitações deste estudo, passa pela realização apenas de modelos de regressão logística, que apesar de terem sido obtidos bons resultados, não permite comparar os valores com os resultados de outros modelos.

Destaque-se que neste estudo, as amostras foram obtidas aleatoriamente, não se tendo em conta a representatividade dos *clusters* em cada amostra. Assim, esta poderá ser uma limitação do estudo no que concerne à construção de modelos preditivos.

7.2 TRABALHO FUTURO

Após a realização deste projeto e analisando os bons resultados obtidos, poderá ser do interesse dos diversos parceiros do Cartão Continente, alargar o modelo criado às restantes marcas. Este será um possível trabalho futuro a ser realizado, adaptando as variáveis consoante o negócio que esteja a ser trabalhado.

Uma vez que o projeto foi desenvolvido a partir de uma análise *cluster*, e que os resultados são influenciados pelas variáveis consideradas inicialmente, uma nova abordagem passaria pelo cálculo da análise fatorial.

Aplicação de outros modelos de regressão poderá ainda ser um trabalho futuro a desenvolver.

BIBLIOGRAFIA

Aaker, D. A. (1996). *Building strong brands*. The Free Press: New York.

Aghaie, A. (2009). Measuring and predicting customer lifetime value in customer loyalty analysis: A knowledge management perspective (a case study on an e-retailer). *International Journal of Industrial Engineering & Production Research*, 20(1):21–30.

Agresti, A. (1996). *An introduction to categorical data analysis*, volume 135. Wiley New York.

Agresti, A. (2002). *Logistic regression*. Wiley Online Library.

Anderson, E. W., Fornell, C., and Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from sweden. *The Journal of Marketing*, 58(3):53–66.

Anderson, J. C. and Narus, J. A. (1984). A model of the distributor's perspective of distributor-manufacturer working relationships. *The Journal of Marketing*, 48(4):62–74.

Anderson, J. C. and Narus, J. A. (1990). A model of distributor firm and manufacturer firm working partnerships. *The Journal of Marketing*, 54(1):42–58.

Assael, H. (1992). *Consumer Behavior and Marketing Action*. PWS-KENT Pub.

Balabanis, G., Reynolds, N., and Simintiras, A. (2006). Bases of e-store loyalty: Perceived switching barriers and satisfaction. *Journal of Business Research*, 59(2):214–224.

Ball, D., Simões Coelho, P., and Machás, A. (2004). The role of communication and trust in explaining customer loyalty: An extension to the ecsi model. *European Journal of Marketing*, 38(9/10):1272–1293.

Beatty, S. E., Mayer, M., Coleman, J. E., Reynolds, K. E., and Lee, J. (1996). Customer-sales associate retail relationships. *Journal of Retailing*, 72(3):223–247.

Bellenger, D. N., Steinberg, E., and Stanton, W. W. (1976). Congruence of store image and self image-as it relates to store loyalty. *Journal of Retailing*, 52(1):17–32.

Bijmolt, T. H., Leeflang, P. S., Block, F., Eisenbeiss, M., Hardie, B. G., Lemmens, A., and Saffert, P. (2010). Analytics for customer engagement. *Journal of Service Research*, 13(3):341–356.

Bob, S. (1994). Successful direct marketing methods. lincolnwood. *ed: NTC Business Books*.

Bolton, R. N. and Drew, J. H. (1991). A multistage model of customers' assessments of service quality and value. *Journal of Consumer Research*, 17(4):375–384.

Brown, G. H. (1952). Brand loyalty-fact or fiction? *Advertising Age*, 23:53–55.

Brown, J., Broderick, A. J., and Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, 21(3):2–20.

Burnham, T. A., Frels, J. K., and Mahajan, V. (2003). Consumer switching costs: A typology, antecedents, and consequences. *Journal of the Academy of Marketing Science*, 31(2):109–126.

Caruana, A. (2003). The impact of switching costs on customer loyalty: A study among corporate customers of mobile telephony. *Journal of Targeting, Measurement and Analysis for Marketing*, 12(3):256–268.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Castañeda, J. A. (2011). Relationship between customer satisfaction and loyalty on the internet. *Journal of Business and Psychology*, 26(3):371–383.

Castro, P. M. M. (2017). Determinação do customer lifetime value aplicação ao retalho alimentar.

Chaudhuri, A. and Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. *Journal of Marketing*, 65(2):81–93.

Cheng, C.-H. and Chen, Y.-S. (2009). Classifying the segmentation of customer value via rfm model and rs theory. *Expert Systems with Applications*, 36(3):4176–4184.

Chun, R. and Davies, G. (2006). The influence of corporate character on customers and employees: Exploring similarities and differences. *Journal of the Academy of Marketing Science*, 34(2):138–146.

Churchill, H. (1942). How to measure brand loyalty. *Advertising and Selling*, 35(24):11–16.

Claycomb, C. and Martin, C. L. (2001). Building customer relationships: an inventory of service providers' objectives and practices. *Marketing Intelligence & Planning*, 19(6):385–399.

Cooil, B., Keiningham, T. L., Aksoy, L., and Hsu, M. (2007). A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. *Journal of Marketing*, 71(1):67–83.

Copeland, M. T. (1923). Relation of consumers' buying habits to marketing methods. *Harvard business review*, 1(3):282–289.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, 134(3):321–367.

Cuillierier, A. (2016). Customer engagement through social media.

Day, G. S. (1969). A two-dimensional concept of brand loyalty. *Journal of Advertising Research*, 9(3):29–35.

Delgado-Ballester, E. and Luis Munuera-Alemán, J. (2001). Brand trust in the context of consumer loyalty. *European Journal of Marketing*, 35(11/12):1238–1258.

Dick, A. S. and Basu, K. (1994). Customer loyalty: toward an integrated conceptual framework. *Journal of the Academy of Marketing Science*, 22(2):99–113.

Dwyer, F. R., Schurr, P. H., and Oh, S. (1987). Developing buyer-seller relationships. *The Journal of Marketing*, 51(2):11–27.

Engel, J. and Blackwell, R. (1982). *Consumer Behavior*. Dryden Press series in marketing. Dryden Press.

Eskildsen, J. K. and Kristensen, K. (2011). The gender bias of the net promoter score. In *Quality and Reliability (ICQR), 2011 IEEE International Conference on*, pages 254–258. IEEE.

Faraway, J. J. (2006). *Extending the linear model with R*. CHAPMAN & HALL/CRC.

Fogli, L. (2006). *Customer service delivery*, volume 4. San Francisco: Jossey-Bass.

Fornell, C. (1992). A national customer satisfaction barometer: The swedish experience. *The Journal of Marketing*, 56(1):6–21.

Fry, J. N., Shaw, D. C., Von Lanzenauer, C. H., and Dipchand, C. R. (1973). Customer loyalty to banks: a longitudinal study. *The Journal of Business*, 46(4):517–525.

Fullerton, G. (2005). The service quality–loyalty relationship in retail services: does commitment matter? *Journal of Retailing and Consumer Services*, 12(2):99–111.

Garbarino, E. and Johnson, M. S. (1999). The different roles of satisfaction, trust, and commitment in customer relationships. *The Journal of Marketing*, 63(2):70–87.

Gassenheimer, J. B., Houston, F. S., and Davis, J. C. (1998). The role of economic value, social value, and perceptions of fairness in interorganizational relationship retention decisions. *Journal of the Academy of Marketing Science*, 26(4):322–337.

Giese, J. L. and Cote, J. A. (2000). Defining consumer satisfaction. *Academy of Marketing Science Review*, 2000:1.

Glady, N., Baesens, B., and Croux, C. (2009). A modified pareto/nbd approach for predicting customer lifetime value. *Expert Systems with Applications*, 36(2):2062–2071.

Gonring, M. P. (2008). Customer loyalty and employee engagement: an alignment for value. *Journal of Business Strategy*, 29(4):29–40.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., and Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research*, 9(2):139–155.

Homburg, C. and Giering, A. (2001). Personal characteristics as moderators of the relationship between customer satisfaction and loyalty—an empirical analysis. *Psychology & Marketing*, 18(1):43–66.

Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley.

Howard, J. A. and Sheth, J. N. (1969). The theory of buyer behavior. Technical report.

Hu, B., Shao, J., and Palta, M. (2006). Pseudo-r 2 in logistic regression model. *Statistica Sinica*, 16(3):847–860.

Hughes, A. (1994). Strategic database marketing: The masterplan for starting and managinga profitable, customer-based marketing program. *Irwin Professional*.

Jones, M. A., Mothersbaugh, D. L., and Beatty, S. E. (2000). Switching barriers and repurchase intentions in services. *Journal of Retailing*, 76(2):259–274.

Keiningham, T. L., Aksoy, L., Williams, L., and Buoye, A. J. (2015). *The wallet allocation rule: Winning the battle for share*. John Wiley & Sons.

Keiningham, T. L., Perkins-Munn, T., and Evans, H. (2003). The impact of customer satisfaction on share-of-wallet in a business-to-business environment. *Journal of Service Research*, 6(1):37–50.

Keller, E. (2007). Unleashing the power of word of mouth: Creating brand advocacy to drive growth. *Journal of Advertising Research*, 47(4):448–452.

Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. *The Journal of Marketing*, 57(1):1–22.

Keller, K. L. (2009). Building strong brands in a modern marketing communications environment. *Journal of Marketing communications*, 15(2-3):139–155.

Kennedy, S. H. (1977). Nurturing corporate images. *European Journal of Marketing*, 11(3):119–164.

Kotler, P. (2000). *Marketing Management: The Millennium Edition*.

Kotler, P., Bowen, J. T., and Makens, J. C. (1999). *Marketing for Hospitality and Tourism, 5/e*. Pearson Education India.

Kristensen, K. and Eskildsen, J. (2011). The validity of the net promoter score as a business performance measure. In *Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2011 International Conference on*, pages 970–974. IEEE.

Kuehn, A. A. (1962). Consumer brand choice as a learning process. *Journal of Advertising Research*, 2(4):10–17.

Lessig, V. P. (1973). Consumer store images and store loyalties. *The Journal of Marketing*, 37(4):72–74.

Letaifa, S. B. and Perrien, J. (2008). A new conceptual framework for greater success with integration of e-crm. *Emergent Strategies for E-Business Processes, Services and Implications: Advancing Corporate Frameworks: Advancing Corporate Frameworks*.

Lim, K. S. and Razzaque, M. A. (1997). Brand loyalty and situational effects: An interactionist perspective. *Journal of International Consumer Marketing*, 9(4):95–115.

Lindsey, J. K. (2000). *Applying generalized linear models*. Springer Science & Business Media.

Lipstein, B. (1959). The dynamics of brand loyalty and brand switching. In *Proceedings of the fifth annual conference of the advertising research foundation*, pages 101–108. Advertising Research Foundation New York.

Lovelock, C. H. and Wright, L. (2002). *Principles of service marketing and management*. Prentice Hall.

Madeira, M. J. d. C. (2014). Customer lifetime value. Master's thesis, FEUC.

Maroco, J. (2007). *Análise estatística com utilização do SPSS, Ed. Sílabo*.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Taylor & Francis.

Money, R. B. (2004). Word-of-mouth promotion and switching behavior in japanese and american business-to-business service clients. *Journal of Business Research*, 57(3):297–305.

Moorman, C., Zaltman, G., and Deshpande, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organizations. *Journal of Marketing Research*, 29(3):314.

Morgan, R. M. and Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *The Journal of Marketing*, 58(3):20–38.

Moutella, C. (2002). Fidelização de clientes como diferencial competitivo.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nguyen, N. and Leblanc, G. (2001). Corporate image and corporate reputation in customers' retention decisions in services. *Journal of Retailing and Consumer Services*, 8(4):227–236.

Oliver, R. (1997). *Satisfaction: A Behavioral Perspective on the Consumer*. Marketing Series. McGraw Hill.

Oly Ndubisi, N. and Kok Wah, C. (2005). Factorial and discriminant analyses of the underpinnings of relationship marketing and customer satisfaction. *International Journal of bank marketing*, 23(7):542–557.

Parasuraman, A. (1997). Reflections on gaining competitive advantage through customer value. *Journal of the Academy of Marketing Science*, 25(2):154.

Parasuraman, A., Zeithaml, V. A., and Berry, L. L. (1988). Servqual: A multiple-item scale for measuring consumer perc. *Journal of Retailing*, 64(1):12.

Patterson, P. G. (1993). Expectations and product performance as determinants of satisfaction for a high-involvement purchase. *Psychology & Marketing*, 10(5):449–465.

Pfeifer, P. E., Haskins, M. E., and Conroy, R. M. (2005). Customer lifetime value, customer profitability, and the treatment of acquisition spending. *Journal of Managerial Issues*, 17(1):11–25.

Ping Jr, R. A. (1993). The effects of satisfaction and structural constraints on retailer exiting, voice, loyalty, opportunism, and neglect. *Journal of Retailing*, 69(3):320–352.

Pritchard, M. P., Havitz, M. E., and Howard, D. R. (1999). Analyzing the commitment-loyalty link in service contexts. *Journal of the Academy of Marketing Science*, 27(3):333–348.

Raymond, M. A. and Tanner Jr, J. F. (1994). Selling and sales management in action: Maintaining customer relationships in direct sales: Stimulating repeat purchase behavior. *Journal of Personal Selling & Sales Management*, 14(4):67–76.

Reichheld, F. F. (1996). *The loyalty effect*, volume 1. Harvard business school press Boston, MA.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard business review*, 81(12):46–55.

Reichheld, F. F. and Covey, S. R. (2006). *The ultimate question: Driving good profits and true growth*. Harvard Business School Press Boston, MA.

Reis, E. (2001). Estatística multivariada aplicada. lisboa: Ed. sílabo. Technical report, ISBN 972-618-247-6.

Rishika, R., Kumar, A., Janakiraman, R., and Bezawada, R. (2013). The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation. *Information systems research*, 24(1):108–127.

Rodrıguez, G. (2007). Lecture notes on generalized linear models. *Princeton University*.

Saleh, K. (2015). Customer acquisition vs.retention costs – statistics and trends. acesso em 2018-02-19.

Sashi, C. (2012). Customer engagement, buyer-seller relationships, and social media. *Management Decision*, 50(2):253–272.

Shainesh, G. (2012). Effects of trustworthiness and trust on loyalty intentions: Validating a parsimonious model in banking. *International Journal of Bank Marketing*, 30(4):267–279.

Singh, S. S. and Jain, D. (2013). Measuring customer lifetime value: models and analysis.

Sirdeshmukh, D., Singh, J., and Sabol, B. (2002). Consumer trust, value, and loyalty in relational exchanges. *Journal of Marketing*, 66(1):15–37.

Swan, J. E. and Oliver, R. L. (1989). Postpurchase communications by consumers. *Journal of Retailing*, 65(4):516–534.

Tate, R. S. (1961). The supermarket battle for store loyalty. *The Journal of Marketing*, pages 8–13.

Tichindelean, M. (2013). Models used for measuring customer engagement. *Expert Journal of Marketing*, 1(1):38–49.

Timothy, L., Lerzan, A., and Alexander Buoye, B. C. (2011). Customer loyalty isn't enough grow your share of wallet. *Harvard Business Review*, 10:29–31.

Turkman, M. A. A. and Silva, G. L. (2000). Modelos lineares generalizados-da teoria à prática. In *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*.

Venkatesan, R. and Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, 68(4):106–125.

Verhoef, P. C. and Langerak, F. (2002). Eleven misconceptions about customer relationship management. *Business Strategy Review*, 13(4):70–76.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Westbrook, R. A. (1987). Product/consumption-based affective responses and postpurchase processes. *Journal of Marketing Research*, 24(3):258–270.

Wong, A. and Sohal, A. (2002). An examination of the relationship between trust, commitment and relationship quality. *International Journal of Retail & Distribution Management*, 30(1):34–50.

Wong, A. and Sohal, A. (2003). Service quality and customer loyalty perspectives on two levels of retail relationships. *Journal of services marketing*, 17(5):495–513.

Zaki, M., Kandeil, D., Neely, A., and McColl-Kennedy, J. R. (2016). The fallacy of the net promoter score: Customer loyalty predictive model. *Cambridge Service Alliance, University of Cambridge*.

Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. *The Journal of Marketing*, 52(3):2–22.

Zeithaml, V. A., Berry, L. L., and Parasuraman, A. (1996). The behavioral consequences of service quality. *The Journal of Marketing*, 60(2):31–46.

Zins, A. H. (2001). Relative attitudes and commitment in customer loyalty models: Some experiences in the commercial airline industry. *International Journal of Service Industry Management*, 12(3):269–294.