UNIVERSIDADE CATÓLICA PORTUGUESA

Métrica para a estimativa das quotas de mercado de um *player* do setor do retalho

por

Gonçalo de Almeida Branco Monteiro Pinto

UNIVERSIDADE CATÓLICA PORTUGUESA

Métrica para a estimativa das quotas de mercado de um *player* do setor do retalho

Trabalho Final na modalidade de Relatório de Estágio
apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em Gestão
por

Gonçalo de Almeida Branco Monteiro Pinto

sob a orientação de:
Professor Leonardo Costa, UCP

Católica Porto Business School, Universidade Católica Portuguesa
abril de 2022

Agradecimentos

Agradeço à Sonae MC, pela oportunidade de estágio que permitiu o desenvolvimento deste trabalho final de mestrado. A experiência foi muito enriquecedora, para um jovem profissional como eu. Sinto-me muito grato a todos aqueles que participaram comigo nesta jornada. Desde já, um agradecimento especial a toda a equipa da direção cliente, equipa liderada pela Engenheira Telma Paulo. Integrei esta equipa durante o estágio, sob a orientação da Engenheira Mariana Esteves. Quero destacar também os meus colegas estagiários que fizeram comigo esta caminhada, José Carvalhaes e Catarina Graça. Eles foram importantes em toda a minha adaptação, por estarem sempre lá.

Agradeço ao meu orientador, Professor Leonardo Costa, bem como à coordenadora de *Analytics & Insight* da direção do Cartão Continente, Engenheira Ana Freitas, por toda a orientação que foi crítica para a elaboração deste Trabalho Final de Mestrado.

Agradeço à Universidade Católica Portuguesa no Porto e a todos os seus professores que trabalham afincadamente na formação de jovens como eu, principalmente aos professores que fazem parte do Mestrado em Gestão com especialização em *Business Analytics*, encabeçada pela Professora Maria Conceição Silva, que fizeram de tudo para que houvesse um bom ensino, apesar da situação pandémica vivida. Obrigado por nos incentivarem a busca pela excelência.

Por fim, agradeço à minha família e a todos os meus amigos, que souberam não só festejar as vitórias como também me apoiar e ajudar a enfrentar os maiores desafios.

"Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."

Clive Humby, UK Mathematician an architect of Tesco's Clubcard, 2006.

Resumo

O objetivo deste Trabalho Final de Mestrado (TFM) foi o da criação de uma métrica capaz de replicar o valor real da quota de mercado, através de dados transacionais extraídos do cartão cliente de um retalhista. O método utilizado fez uso de algoritmos de *machine learning* na identificação das top 15 categorias de loja, da análise de *cluster* na agregação dessas mesmas categorias e da análise de regressão na identificação dos fatores que afetam a quota de mercado de um determinado *cluster*. A partir deste último passo foi possível construir a métrica pretendida com base nos coeficientes das variáveis identificadas como significativas para o referido *cluster*: vendas brutas, número de transações, número de artigos disponíveis e percentagem de desconto. A métrica resulta da soma ponderada pelos coeficientes destas variáveis transacionais. Os resultados mostram que com a métrica é possível monitorizar a quota de mercado por via da sua estimação interna, sem ter de depender de dados de quotas de mercado fornecidos por fonte externa.

Palavras-chave: Quota de Mercado; Dados Transacionais; Big Data; Machine Learning; Análise *Cluster*; Análise de Regressão e Métrica

Número de palavras: 6532

Abstract

The objective of this Master Final Assignment (MFA) was to create a metric capable of replicating the real value of a retail market share, through transactional data extracted from a retailer's customer card. The method used made use of machine learning algorithms to identify the top 15 store categories, cluster analysis to aggregate these same categories and regression analysis to identify the factors that affect the market share of a given cluster. From this last step, it was possible to build the desired metric based on the coefficients of the variables identified as significant for that cluster: gross sales, number of transactions, number of items available and discount percentage. The metric results from the sum weighted by the coefficients of these transactional variables. The results show that with the metric it is possible to monitor the market share through its internal estimation, without having to rely on market share data provided by an external source.

Keywords: Retail Market Share; Transactional Data; Big Data; Machine Learning; Cluster Analysis; Regression Analysis and Metric

Number of words: 6532

Índice

Capítulo 1 - Introdução

O objetivo deste Trabalho Final de Mestrado (TFM) é o da criação de uma métrica interna capaz de replicar o valor real de uma quota de mercado, através de dados internos transacionais extraídos do cartão cliente de um retalhista, autonomizando a empresa de fontes de informação externa sobre a referida quota.

O método utilizado faz uso de algoritmos de *machine learning* na identificação das top 15 categorias de loja, da análise de *cluster* na agregação dessas mesmas categorias e da análise de regressão na identificação dos fatores que afetam a quota de mercado de um determinado *cluster*. A métrica interna pretendida resulta da soma ponderada dos coeficientes desta última regressão nas variáveis transacionais.

A transformação digital verificada nos diferentes setores de atividade abriu as portas a mudanças de mentalidade e à forma de como cada negócio é explorado. A tecnologia passou a estar ao serviço da caminhada rumo aos objetivos das empresas, permitindo as mesmas melhorarem as suas performances, como também proporcionou um maior alcance de mercado. Antes, a concorrência encontrava-se na porta ao lado. Hoje, com a transformação digital, atribui-se ao cliente maior conhecimento e facilidade de procura do mesmo bem entre milhares de competidores. Os setores de atividade encontram-se mais expostos à concorrência do que nunca. A pandemia Covid-19 foi exemplo de forte catalisador desta transformação digital. A análise de dados e extração de conhecimento através dos mesmos compõem um dos desafios mais cruciais para o sucesso de uma empresa no longo-prazo.

Existe uma quantidade enorme de dados a ser gerada através de diversas fontes, cuja posse e tratamento pertence ao horizonte dos *big players* dos múltiplos setores. Através de dados recolhidos, as empresas conseguem estudar e prever, com enorme profundidade, o comportamento dos seus clientes, agora posicionados no centro de qualquer negócio. Percebe-se a importância de cada empresa estudar o seu posicionamento no respetivo setor, de forma a obter maior capacidade de reação, resultando em vantagens competitivas face aos restantes *players*.

O conceito de *Big Data* é descrito como um conjunto de dados de grande dimensão, composto por bases de dados complexas que são difíceis de recolher, armazenar e analisar efetivamente, utilizando sistemas de gestão de informação atuais (Fan W, Bifet A 2013). *"You can't manage what you don't measure"* (McAfee A., Brynjolfsson E., 2012) Esta é uma frase proferida por diversos autores, desde o início da última década. A expressão revela a importância da análise de dados provenientes das respetivas

cadeias de valor de cada negócio. Hoje, o que é *mainstream* é a utilização de dados e de modelos, com o intuito de melhorar continuamente os processos ou atividades das empresas. E tem vindo a ocorrer uma transição de uma análise meramente descritiva, para uma análise com capacidade preditiva, resultando em mais conhecimento e contribuindo para uma melhor tomada de decisão e performance das empresas.

As principais características do conceito de *Big Data* são o volume, velocidade e variedade. Com as cadeias de valor dos respetivos negócios cada vez mais digitalizadas, as empresas passam a ter à disposição grandes quantidades de dados, em tempo real, conferindo maior agilidade aos analistas no sentido de obtenção de vantagens competitivas. (IBM, 2022). Segundo Randy Bean (2021), pelo terceiro ano consecutivo, o investimento em dados e iniciativas de inteligência artificial tornou-se universal, com 99,0% das empresas a reportarem um investimento nessa área.

Mas o investimento que as empresas realizam não pode estar somente centrado na recolha e armazenamento dos dados. O seu tratamento é crucial. É através desta etapa que os dados acrescentam valor à empresa. E por isso os especialistas em artigos de investigação referem que *data is the new oil* (os dados são o novo petróleo). A conclusão a retirar da frase é que os dados sem tratamento não apresentam valor em si mesmo. Apenas quando são devidamente tratados e/ou refinados os mesmos podem constituir uma ferramenta de apoio à tomada de decisão, informação, permitindo às empresas reagirem de outra forma às diferentes forças do mercado, tornando o processo de tomada de decisão mais eficaz (Kenway Consulting, 2020).

Os avanços tecnológicos permitiram a digitalização dos processos que compõem a cadeia de valor da indústria do retalho. Também nesta indústria, o foco dos negócios é cada vez mais o consumidor, sendo que este tem vindo a mudar as suas exigências e a requerer uma análise cada vez mais detalhada.

Na indústria do retalho, a análise de *Big Data* proporciona perspetivas interessantes acerca do comportamento do consumidor. Já existem modelos preditivos capazes de extrair tendências dos dados disponíveis, no sentido de prever vendas futuras, frequência de compra de diversos segmentos de clientes, a probabilidade de fazerem compras online, etc (Camm et al., 2014; Shmueli and Koppius, 2011). As empresas conseguem acompanhar em tempo real as mudanças nos hábitos de consumo, de modo a estarem preparadas para satisfazerem o máximo de necessidades do consumidor. Com este intuito, os *big players* do retalho de bens, como é o caso da SONAE MC, promovem a utilização do cartão fidelização, através de uma estratégia

de comunicação e campanhas promocionais personalizadas, permitindo assim fazer um *tracking* à evolução do comportamento do consumidor dentro de loja.

Os resultados de uma empresa do retalho (ou de qualquer outro setor) são, numa boa parte, uma consequência da qualidade do serviço prestado. Existem diversos estudos realizados neste setor que tentam explicar a performance através de múltiplas variáveis. A quota de mercado engloba as performances no que diz respeito às vendas de todos os players de um determinado setor. Este Trabalho Final de Mestrado serve para verificar se, através de dados transacionais provenientes de transações realizadas em loja (dito de outra forma, se conseguimos explicar uma métrica externa a partir de dados internos detidos pela entidade), é possível encontrar variáveis internas que permitam prever as flutuações no valor da quota da empresa, sem recorrer a fontes de informação externas e antecipando estas flutuações.

A métrica que é desenvolvida representa o valor esperado da quota de mercado e constitui uma forma indireta de monitorizar a performance da empresa face aos seus concorrentes. Compreendê-la e perceber como a mesma afeta a quota de mercado da empresa traz vantagens competitiva a esta última.

O projeto tem vindo a ser desenvolvido na SONAE MC, subsidiária da Sonae Holding. A Sonae é um grupo empresarial multinacional que atua em diversos setores, destacando-se o retalho, os serviços financeiros, a gestão de centros comerciais e as telecomunicações. É atualmente o maior empregador privado em Portugal. A Sonae MC é atualmente líder em Portugal no retalho alimentar, apresentando assim um conjunto de marcas: Continente (hipermercados), Continente Modelo e Continente Bom dia (supermercados de conveniência), Meu Super (supermercados em franchising), Bagga (cafetarias e restaurantes), Go Natural (supermercados e restaurantes saudáveis), Makenotes, note! (livraria/papelaria), ZU (produtos e serviços para cães e gatos), Well's (saúde, bem-estar e ótica) e Dr. Well's (clínicas medicina dentária e estética).

O projeto surgiu no contexto do estágio curricular desenvolvido no departamento Direção Cliente, encarregue de todas as atividades relacionadas com Cartão Continente, mais concretamente na equipa comercial. A gestão e estabelecimento de parcerias, extração de insights dos dados de clientes, operacionalização de estratégias de marketing e medição dos resultados de publicidade, são tarefas desenvolvidas neste departamento.

Lançado em 2007 e a aumentar cada vez mais a sua base de utilizadores, o programa de fidelização Continente conta atualmente com cerca de 4 milhões de contas ativas,

o que corresponde a 90% das famílias portuguesas. O processo de fidelização de clientes integra atualmente 19 parceiros permanentes dos setores da alimentação, combustível, saúde e moda. Os parceiros internos incluem lojas Continente: Well's, Note!, Bagga, Meu Super, Zippy, and Mo. Os parceiros externos são integrados pelos postos de abastecimento Galp e Ibersol, um grupo com diversas marcas que gere o Burger King, SOL, KFC, Ô Kilo, Roulotte, Pans & Company, Pizza Hut, Miit e Pasta Café. Para além destas parcerias, o programa de fidelização de clientes tem vindo a desenvolver diversas parcerias dentro de outros setores, tais como a cultura e literatura, serviços financeiros e transportes.

O Trabalho Final de Mestrado desenvolve-se em 5 capítulos. Após este primeiro capítulo introdutório, no capítulo 2 é apresentada uma revisão de literatura de suporte ao modelo empírico desenvolvido nesta dissertação. No capítulo 3 é explicada a seleção e recolha dos dados a usar e no capítulo 4 são apresentados o modelo empírico e a sua implementação e os resultados e a respetiva discussão. O capítulo 5 finaliza o Trabalho Final de Mestrado, relembrando ao leitor o objetivo e o método utilizado, sumariando os principais resultados obtidos, apontado as limitações do trabalho e fazendo algumas sugestões para investigações futuras.

Capítulo 2 - Revisão de Literatura

2.1 O contexto

O mundo está a registar mudanças cada vez mais rápidas e frequentes, como nunca antes visto:

- O contexto político, através de políticas macroeconómicas e não só, cria pressões sobre a sociedade (ex: o Brexit);

- Socialmente, assistem-se também a grandes mudanças nos estilos de vida e nos hábitos de consumo (geração Z a entrar no mercado de trabalho e pandemia Covid-19) e a um crescimento da desigualdade nos países da OCDE nas últimas décadas (Piketty, 2014; Gornick & Jantti, 2014);

- A transformação digital está a crescer de forma exponencial e começa a estar inserida em todas as atividades do quotidiano (*AI*, *Big Data* and *Data Analytics*, *Blockchain* e outras tecnologias digitais);

- A temática da sustentabilidade ambiental, com os problemas da perda de biodiversidade e das alterações climáticas a atingirem níveis alarmantes (Rijmenam, 2019).

Os últimos dois anos foram marcados pela pandemia Covid-19 que teve impacto em diversos setores da economia. As pessoas evitaram ou foram obrigadas a evitar, em certas alturas, uma grande exposição física em espaços públicos e passaram a estar mais tempo em casa (por vezes em teletrabalho). Este maior tempo em casa criou-lhes oportunidades de mudarem rotinas, como consumidores, e tornarem-se clientes mais bem informados (Sayyida, S., Hartini, S., Gunawan, S., & Husin, S., 2021). A informação acerca dos produtos, e a sua facilidade de obtenção, leva a que os clientes queiram pagar menos por mais e a procurarem até encontrarem a solução que melhor os satisfaz. Segundo dados da Nielsen, encontrados num artigo elaborado pela KPMG sobre o tema do futuro no retalho, acredita-se que 81% dos clientes estão a fazer pesquisa online antes de se comprometerem com a compra (KPMG, 2021).

2.2 Quota de mercado

O conceito de quota de mercado é usado para evidenciar a porção do total das vendas de mercado (em quantidade ou volume) para um determinado produto, categoria ou unidade de negócio, num dado período de tempo e numa determinada

área geográfica. Esta métrica é usada para dar uma ideia da dimensão do mercado que possui o alvo de análise. (Hayes, A. 2021)

Para efeitos deste trabalho utilizamos o conceito de quota de mercado mais específico de Weiss (1968):

$$S_i = \frac{Q_i}{Q} \quad (1)$$

Onde Q_i corresponde às vendas no mercado da empresa ou marca *i* de determinada categoria de produto e Q corresponde ao total das vendas no mercado da referida categoria:

$$Q = \sum_{i=1}^{m} Q_i \quad (2)$$

Em que *m* é o número de empresas ou marcas concorrentes.

"The economic well-being of a business firm can often be summarized in terms of its market share. Market share responds to price, advertising expenditures, retail availability, and product characteristics" (Weiss, 1968, p. 290). A análise da quota de mercado e a enumeração das principais características associadas são relevantes.

As principais características associadas são: a competitividade, o valor descritivo e preditivo e a orientação para o lucro das empresas (Nakanishi & Cooper, 2010). Primeiro, considera-se que o conceito é um indicador de competitividade, pois implica que os efeitos das ações de uma empresa sejam analisados em conjunto com as posições do mercado e as ações das empresas concorrentes. Segundo, como métrica interna a desenvolver, é considerada uma métrica tanto descritiva como preditiva. Fazer previsões precisas acerca do valor futuro da quota é um grande contributo para análise da performance do mercado, mas não dispensa informação base acerca dos restantes competidores e da influência de ações de marketing na performance da marca. Por último, a referida métrica deve ser orientada para o lucro, pois a empresa apenas está interessada no crescimento da quota se o mesmo tiver consequências positivas no seu lucro. Por vezes a expansão pode não compensar, visto que novas estratégias de comunicação com o mercado podem ser bastante dispendiosas.

O tema da quota de mercado tem surgido em diferentes estudos nos últimos anos. Sendo esta uma métrica que dá informação acerca da performance dos *players* de um determinado setor, descobrir variáveis internas que a afetem é um desafio. Por exemplo, existe um estudo que se debruçou sobre o impacto da localização e da atratividade das lojas como fatores significativos que provocam mudanças na quota de mercado, provando que esta era sensível aos mesmos (Drezner, 1994).

Outra relação importante a ter em conta é aquela que existe entre o fornecedor e o retalhista na gestão das categorias. Dupre & Gruen (2004) referem que o retalhista deve olhar para cada categoria como um centro de lucro, e que a quota de mercado das mesmas é influenciada pelo mix de marcas, SKUs (termo inglês *Stock Keeping Unit*, com tradução para "número de referência único") e estratégia de preço.

2.3 Algoritmos *machine learning* e modelos utilizados no trabalho

2.3.1 Modelos supervisionados e não supervisionados

Modelos supervisionados representam algoritmos que apresentam a capacidade de identificar padrões em dados de treino, com o objetivo de prever observações futuras (Singh, Thakur, & Sharma 2016).

O processo de aprendizagem de qualquer modelo supervisionado *machine learning* subdivide cada conjunto de dados em duas amostras: uma amostra de treino e uma amostra de teste, sendo, por norma, a divisão das observações de 70% para a primeira e 30% para a última (Nasteski, 2017).

No processo de treino, as observações da respetiva amostra são utilizadas na construção do modelo. O algoritmo perceciona o comportamento dos fatores manipuláveis, aprende e retorna um modelo com o impacto dos fatores interiorizado. As restantes observações, inseridas na amostra de teste, vão ser alvo do motor de execução do modelo criado, cujo resultado são previsões da variável dependente da amostra de teste. A qualidade do modelo teste é calculada pela percentagem de erro verificada entre o valor real e o valor previsto da amostra de teste.

O modelo de regressão linear (que vamos utilizar) é um dos exemplos mais conhecidos de modelos supervisionados, em que o *output* consiste num número real para determinados *inputs*.

Por outro lado, existem os modelos não supervisionados. O principal objetivo destes algoritmos é representarem a estrutura estatística dos padrões comuns no conjunto dos dados. Não há distinção entre variáveis, ou seja, não existe variável dependente ou variável alvo. Todas as variáveis que compõem as observações são tratadas como inputs deste modelo (Dayan, 1999). O método de *clustering*, especialmente o algoritmo *k-means clustering* (que vamos utilizar), é dos modelos não supervisionados mais utilizados e será utilizado como recurso neste estudo.

O algoritmo *k-means clustering* permite ao analista definir o número de centróides em que pretende segmentar o conjunto de observações. As observações são atribuídas

a um determinado *cluster*, dependendo da sua distância ao centro do centróide. O critério de segmentação tem por base a minimização do valor da distância euclidiana verificada entre as observações e os centróides (Mendes De Morais et al., 2012):

$$J = \sum_{j=1}^{k} \sum_{n \in S_j} ||x_n - c_j||^2 \quad (3)$$

Em cada interação das observações com o algoritmo também as dimensões dos centróides são recalculadas.

Existe um problema com este algoritmo *k-means clustering* que é o de requer a definição do número de *clusters* (k) *a priori*. Para construir o *algoritmo k-means clustering*, de modo a encontrar de forma automática a solução ótima do número de *clusters* a segmentar a distribuição dos dados, é possível utilizar o conceito de entropia. No entanto, o número de *clusters* também pode ser decidido através de outros meios, nomeadamente o método do cotovelo (explorado no Capítulo 4) e o coeficiente de silhueta.

O coeficiente de silhueta avalia a qualidade do *cluster*. Varia entre -1 e 1, sendo 1, quando os *clusters* estão perfeitamente divididos e apresentam distinção clara, e -1, quando os *clusters* estão mal segmentados. O analista consegue avaliar os valores dos coeficientes de silhueta dos *clusterings* realizados com diferentes números de k, pertencentes ao intervalo decidido pelo mesmo. A solução ótima de k acaba por estar ligada ao *clustering* que apresenta um maior coeficiente de silhueta (Bhardwaj, 2020).

2.3.2 Regressão Linear

O modelo de regressão linear parte de uma equação que estabelece relações estatísticas entre variáveis explicativas e a variável de resposta, no sentido de produzir novas observações através de estimação. No caso deste trabalho, a regressão linear que se realizou foi uma regressão linear múltipla, no sentido de examinar o impacto de diversas variáveis transacionais no valor da variável dependente que é quota de mercado.

Nas diversas áreas de estudo, como economias, tecnologias, ciências sociais, os modelos matemáticos sob a expressão

$$y_i = f(x_1, \dots, x_n, a_0, \dots, a_l) \quad (4)$$

são dados como aceite, seguidos de considerações acerca de relações entre as variáveis. Contudo, estes modelos devem ser testados (Späth, 2014). Para tal, são recolhidas observações que contam com os valores $(x_1, \dots, x_n)$ das variáveis independentes e com valores da variável independente y_i $(i = 1, \dots, m)$, resultado da

interação das variáveis explicativas. Os parâmetros $(a_0, \dots, a_l)$ são determinados através de estimativas que tornam o poder explicativo do modelo

$$y_i = f(x_1, \dots, x_n, a_0, \dots, a_l) + \varepsilon_i \qquad (i = 1, \dots, m) \qquad (5)$$

o maior possível. A interseção, a_0, é o valor esperado de y_i quando o valor de todas as variáveis independentes é igual a zero. O termo ε_i é assumido como o desvio entre o valor real da observação y_i e o valor que o modelo estima como observação y_i. Pode ser escrito sob a seguinte forma:

$$\varepsilon_i = y_i - E(y_i | a_0, \dots, a_n) \qquad (6)$$

Para melhor compreensão deste valor, coloquemos como hipótese a existência de um modelo perfeito, isto é, um modelo que apresenta uma explicação para toda a variação do valor da variável dependente. Por outras palavras, é visto como um modelo que descreve o processo de determinação do valor da variável dependente para qualquer caso da população (Berry, 1993). A existirem, são modelos bastante extensos, incluindo todas as variáveis que influenciam y_i, ainda que muitas destas apresentem um menor impacto comparativamente a outras. Contudo, muitos autores acreditam que modelos perfeitos não são totalmente determinísticos, indicando aleatoriedade intrínseca no comportamento humano, dificilmente incorporada nos modelos através de variáveis (Gujarati, 2021).

Um aspeto importante a ter em consideração é o número de observações que fazem parte da criação de um modelo. Normalmente é assumido como número de observações (m) ótimo como sendo o quadrado do número de variáveis explicativas (n): $m = n^2$ (Späth, 2014). Contudo, o valor tende a ser bastante maior ou bastante menor, o que pode causar problemas ao modelo em qualquer um dos casos.

2.3.3 Regressão *Clustering*

A distribuição complexa de dados do mundo real é frequentemente moldada pela mistura de várias distribuições de dados individualmente mais simples. O método de *clustering* é uma das ferramentas utilizadas para revelar como os dados de um *dataset* bastante complexo se encontram distribuídos.

O método de *clustering*, assim como o método de regressão linear, é uma técnica de *data mining* (extração de informação através de dados) bastante relevante. Enquanto as regressões lineares são realizadas em *datasets* com dados classificados, estudando a relação entre variáveis independentes e uma variável dependente, o processo de *clustering* encontra-se relacionado com *datasets* que não apresentam a tal variável

dependente explicada. O conceito aqui retratado – Regressão *Clustering* – tem por objetivo resolver o meio termo que existe entre a finalidade de cada método na sua individualidade: tratar dados que pertencem a *datasets* que não apresentam informação suficiente para uma boa qualidade de aprendizagem de qualquer modelo realizado sobre estes (Zhang, 2003).

Capítulo 3 - Recolha e Preparação de Dados

Neste capítulo é detalhado o procedimento de coleção e preparação de dados. O processo envolve as seguintes etapas: i) seleção de categorias; ii) extração de dados de vendas para as categorias selecionadas; iii) extração de dados transacionais e de segmentação do cliente. O procedimento foi realizado utilizando o *Structured Query Language* (SQL) para extração de dados da base de dados da SONAE MC, seguido de análise de dados realizada através da linguagem de programação *Python*.

3.1 Seleção de categorias

As lojas Continente oferecem uma variedade de produtos que vão desde a alimentação à higiene, até equipamento para casa, contando para o efeito com uma estrutura mercadológica[1] devidamente segmentada, que vai desde o nível dos artigos individuais até ao nível das Direções Comerciais (DCs). As DCs do Continente partem de uma visão global do *layout* de loja, dividindo a mesma entre a DC de Frescos (talho e peixaria, por exemplo) e a DC Alimentar (além de bens alimentares, sem os produtos frescos, também engloba outros bens do dia-a-dia, como é o caso dos produtos de higiene e limpeza do lar).

Este TFM foca-se apenas na análise das 15 maiores categorias da DC Alimentar (que não inclui os produtos frescos), com base numa métrica ponderada composta por vendas brutas e número de transações, *insights* estes de negócio analisados pela equipa de *Analytics* da Direção Cliente. No Quadro 1 abaixo encontram-se as categorias de produtos selecionadas.

Categorias	Código
Limpeza e tratamento de roupa	0601
Limpeza geral	0603
Produção de papel e consumo	0604
Cerveja	0303
Refrigerantes	0302

[1] Estrutura mercadológica: separação e organização de produtos no sistema da empresa, criado tendo em conta os critérios utilizados pelo cliente na escolha de cada produto para compra (marca, preço, tamanho, etc...). A estrutura mercadológica neste estudo assenta na seguinte forma: Direção Comercial – Unidade de Negócio – Categoria – Subcategoria – Unidade base – Artigo.

Vinhos Classif. Nac./Estrang.	1701
Gorduras líquidas	1001
Ingredientes básicos	1002
Refeições congeladas	0702
Iogurtes e sobremesas	0804
Leite e natas	0801
Bolachas	0203
Doçaria	0202
Conservas	0103
Bebidas quentes	1403

Quadro 1: Categorias de produtos selecionadas

Fonte: Sonae e Autor

3.2 Extração de dados de vendas para as categorias selecionadas

Uma vez selecionadas as categorias de produtos, foram extraídos dados relativos às transações de clientes *loyal*, ou seja, dados apenas de transações efetuadas com o cartão cliente, observadas de forma agregada a quadrissemana desde janeiro de 2020 até janeiro de 2022 (exclusive). Foram consideradas todas as insígnias onde decorreram as transações: Continente, Continente Bom Dia, Continente Modelo e Continente Online.

As variáveis transacionais selecionadas para melhor descrever a performance de cada categoria encontram-se no Quadro 2 abaixo.

Variáveis	Descrição
num_trx_tot	Total de transações
vb_tot	Vendas brutas
vl_tot	Vendas líquidas
qtd_tot	Quantidades totais
num_prod_tot	Número de gamas de artigos total
perc_dscnt	Percentagem de desconto (em %)
perc_prod_mp	Percentagem de produtos marca própria (em %)

Quadro 2: Variáveis transacionais selecionadas

Fonte: Sonae e Autor

3.3 Dados de Clientes

Para uma análise mais profunda de cada categoria, considerou-se importante ter em conta variáveis relativas ao perfil ou "ADN" dos clientes responsáveis pelas transações. A atribuição de um determinado cliente a segmentos de clientes é um processo dinâmico. À medida que os hábitos e comportamentos de compra dos clientes mudam, os segmentos considerados de clientes acompanham a referida mudança. Por este motivo, existe a necessidade de atualizar o segmento de sensibilidade ao preço a que cada consumidor pertence. É um processo que decorre mensalmente, podendo o mesmo cliente pertencer a segmentos diferentes em dois períodos de tempo consecutivos. Portanto, é importante ter este detalhe bem presente no modelo, visto que, para cada quadrissemana, os valores de segmentação têm que corresponder ao seu valor mais recente. O Quadro 3 ilustra a variável de segmentação selecionada – percentagem de clientes muito sensíveis ao preço, isto é, que pertencem aos dois níveis mais baixos da escala proveniente de um modelo utilizado pela empresa.

Variáveis	Descrição
perc_cli_ss	Percentagem de clientes sensíveis ao preço (em %)

Quadro 3: Variável de segmentação selecionada

Fonte: Sonae e Autor

3.4 Variáveis transacionais consideradas e código da extração dos dados

O Quadro 4 descreve as variáveis transacionais consideradas e a Figura 1 ilustra o código de extração de dados em MySQL utilizado.

	Variáveis	Descrição
Transações	num_trx_tot	Total de transações
	vb_tot	Vendas brutas
	vl_tot	Vendas líquidas
	qtd_tot	Quantidades totais
	num_prod_tot	Número de gamas de artigos total
	perc_dscnt	Percentagem de desconto (em %)

Dimensão do produto	perc_prod_mp	Percentagem de produtos marca própria (em %)
Segmentação de cliente	perc_cli_ss	Percentagem de clientes sensíveis ao preço (em %)

Quadro 4: Descrição das variáveis transacionais consideradas

Fonte: Sonae e Autor

```
ex.sql > {} SELECT
select
                    CAT_CD_EXT,
                    year as ano,
                    count(distinct(t.transaction_id)) as num_trx_tot,
                    sum(t.gross_sls_amt) as vb_tot,
                    sum(t.net_sls_amt) as vl_tot,
                    sum(t.qty) as qtd_tot,
                    count(distinct t.product_key) as num_prod_tot,
                    sum(prod_dscnt_issued_amt+trans_dscnt_ratt_amt+direct_dscnt_amt)/sum(gross_sls_amt+direct_dscnt_amt) as perc_dscnt,
                    count(distinct case when s.segment_cd in (1,2) then t.id_cliente else null end) / count(distinct t.id_cliente) as perc_cli_ss,
                    sum(case when brand_type_cd like 'MP' then t.gross_sls_amt else null end) / sum(t.gross_sls_amt) as perc_prod_MP

      from
            (select t .*,p.CAT_CD_EXT, p.brand_type_cd
               from labmarketing.cic_sample_trx t
               inner join (SELECT product_key, CAT_CD_EXT, brand_type_cd FROM  expl.expl_dimm_product_asis where CAT_CD_EXT in ('0202','0103','1403'
               where time_key between 20200101 and 20211231
                        and loc_brand_cd in (143,303,302,888)
            ) t
                left join (select * from labmarketing.cic_segment_pss s where yearmonth in (select max(yearmonth) from labmarketing.cic_segment_pss
                where yearmonth <= 202101)) s on t.id_cliente=s.id_cliente

  inner join (select time_key, week, month, year from expl.expl_dim_time where time_level='DAY') re on re.time_key= t.time_key
     group by CAT_CD_EXT, year'''

```

Figura 1: Código MySQL de extração dos dados

Fonte: Autor

Capítulo 4 - Modelo empírico

4.1 A métrica pretendida

Este capítulo descreve o modelo empírico deste TFM e discute os resultados obtidos com o mesmo.

O objetivo último do modelo empírico é o de identificar variáveis transacionais internas significativas que afetam a quota de mercado e construir uma métrica, com ´base nas referidas variáveis, de forma a identificar flutuações nas quotas de mercado dos produtos em questão.

Sendo a quota de mercado um valor influenciado pela atividade de outras empresas dentro do setor que não a Sonae MC, entende-se como algo importante para a empresa criar a referida métrica, de forma a produzir estimativas do valor da quota de mercado exclusivamente a partir de variáveis transacionais internas, sem necessidade de recorrer a informação externa.

4.2 A matriz de correlações

A matriz de correlações foi utilizada numa fase inicial de construção da regressão, de modo a estudar o comportamento das variáveis independentes ou explicativas escolhidas. Esta matriz é vista como critério de escolha das variáveis independentes a incorporar no modelo, uma vez que é importante evitar variáveis independentes que exibam uma forte colinearidade.

A Figura 2 mostra os coeficientes de correlação linear de Pearson entre as variáveis independentes potenciais a considerar. O resultado da matriz de correlações mostra uma grande correlação entre as seguintes variáveis: vendas brutas e vendas líquidas (0,98); e quantidade total e número de transações (0,84). A escolha de variáveis a incorporar na regressão deixou por isso de parte as variáveis "vendas líquidas" e "quantidade total".

num_trx	vb_tot	vl_tot	qtd_tot	num_prod_tot	perc_dscnt	perc_cli_ss	perc_prod_mp	
1,00	0,48	0,54	0,84	0,06	-0,73	-0,48	0,51	num_trx
	1,00	0,98	0,42	0,22	-0,06	-0,17	-0,08	vb_tot
		1,00	0,54	0,18	-0,09	-0,17	-0,03	vl_tot
			1,00	-0,09	-0,65	-0,31	0,39	qtd_tot
				1,00	0,09	-0,02	-0,30	num_prod_tot
					1,00	0,36	-0,61	perc_dscnt
						1,00	-0,24	perc_cli_ss
							1,00	perc_prod_mp

Figura 2: Coeficientes de correlação linear de Pearson

Fonte: Autor

4.3 O método do cotovelo

As quotas de mercado disponibilizadas pela fonte externa para este estudo, encontravam-se agregadas em quadrissemanas dos anos 2020 e 2021. Por esse motivo, existiam observações insuficientes para correr uma regressão linear para cada categoria de produto.

A forma que se encontrou para superar a limitação encontrada foi a agregação de categorias de produtos através de um algoritmo de *k-means clustering*, de forma a contar com um maior número de observações na regressão ou regressões a correr. E o passo seguinte nesta análise foi o de encontrar o número ótimo de *clusters* a considerar, pelo método do cotovelo.

O método *k-means clustering* calcula o desvio de cada observação ao centro do respetivo *cluster* a que pertence. O objetivo é que essa distância seja o menor possível. A solução ótima parte da observação da relação entre comportamento da variabilidade das observações dentro de cada *cluster*, medida pelo WCSS (*Within Cluster Sum of Squares*), face à alteração do número de *clusters*. Este processo tem a designação de método do cotovelo. Quando o aumento do número de *clusters* já não traduz uma diminuição considerável do WCSS então está-se na presença do número *k* de *clusters* ótimo. No nosso caso, a solução aponta para k=7, como número ótimo de *clusters* a considerar, sendo que todos os *clusters* contêm no mínimo uma categoria de produto. A Figura 3 ilustra o método do cotovelo utilizado que nos permitiu retirar esta conclusão.

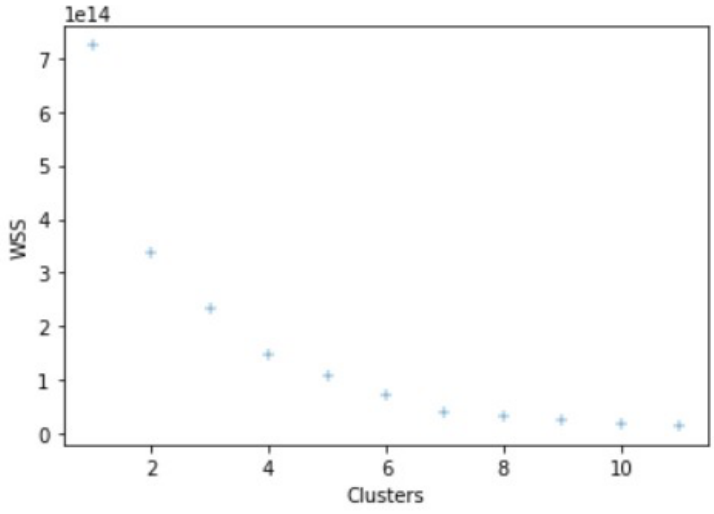

Figura 3: Método do cotovelo na determinação do número ótimo de *clusters k*
Fonte: Autor

O Quadro 5 indica as categorias de produto que integram cada um dos 7 *clusters*.

Cluster	Categorias	Código
Cluster 1	Cerveja	0303
	Refeições congeladas	0702
	Gorduras líquidas	1001
Cluster 2	Iogurtes e sobremesas	0804
Cluster 3	Ingredientes básicos	1002
	Bolachas	0203
	Conservas	0103
Cluster 4	Limpeza e tratamento de roupa	0601
	Vinhos Classif. Nac./Estrang.	1701
Cluster 5	Bebidas quentes	1403
Cluster 6	Leite e natas	0801
Cluster 7	Limpeza geral	0603
	Produção de papel e consumo	0604
	Refrigerantes	0302
	Doçaria	0202

Quadro 5: Categorias de produto por *cluster*
Fonte: Sonae e Autor

Note-se que o objetivo dos *clusters* era o de ultrapassar a limitação do número reduzido de observações por categoria de produto. Tendo em conta os resultados, o *cluster* 7 passa a ser objeto da regressão linear, pois é aquele que tem o maior número de categorias de produtos e, por isso, de observações.

4.4 A regressão linear

O modelo de regressão linear inclui o valor da quota de mercado como variável dependente e os restantes dados transacionais internos extraídos como variáveis explicativas ou independentes.

A intenção do modelo é chegar a um valor estimado da quota a partir de uma métrica que incorpore as variáveis independentes que se apresentem como significativas. A comparação entre o valor estimado que resulta da métrica e o valor observado dá uma indicação da qualidade da métrica.

O modelo da primeira regressão é o seguinte:

$$quota = \alpha + \beta 1 * vb\_tot + \beta 2 * num\_trx + \beta 3 * num\_prod\_tot + \beta 4 * perc\_dscnt + \beta 5 * perc\_prod\_mp + \beta 6 * perc\_cli\_ss + \varepsilon \qquad (7)$$

Os resultados da primeira regressão são bastante positivos. O valor calculado do *Adj. R-squared* que dá o poder explicativo do modelo é 0,825 (ver Figura 4).

A significância das variáveis independentes é frequentemente percecionada através da interpretação dos valores de *p-value* (Altman & Krzywinski, 2017). Os *p-values* resultam de testes estatísticos em que a hipótese testada é que não existe relação entre a variável dependente e a variável independente em questão (hipótese nula). Se a hipótese nula for rejeitada, a escolha recai sobre a hipótese alternativa, que afirma que a variável independente apresenta impacto significativo sobre a variável dependente (McLeod, 2019). Neste TFM foi considerado como nível de significância um alfa igual a 5%. Por esse motivo, o valor estimado do coeficiente beta, que relaciona a variável independente com a variável dependente, é considerado significativo quando o *p-value* correspondente é inferior a 0.05 (5,0%).

OLS Regression Results

Dep. Variable:	quotas	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	81.92
Date:	Tue, 05 Apr 2022	Prob (F-statistic):	9.33e-36
Time:	17:23:06	Log-Likelihood:	-185.30
No. Observations:	104	AIC:	384.6
Df Residuals:	97	BIC:	403.1
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	33.6043	4.849	6.930	0.000	23.981	43.228
num_trx_tot	-4.129e-05	1.23e-05	-3.367	0.001	-6.56e-05	-1.7e-05
vb_tot	6.788e-06	1.48e-06	4.600	0.000	3.86e-06	9.72e-06
num_prod_tot	-0.0021	0.001	-3.128	0.002	-0.003	-0.001
perc_dscnt	40.2822	2.940	13.701	0.000	34.447	46.118
perc_cli_ss	-23.3029	16.681	-1.397	0.166	-56.411	9.805
perc_prod_mp	2.2770	1.391	1.637	0.105	-0.483	5.037

Omnibus:	0.964	Durbin-Watson:	1.594
Prob(Omnibus):	0.618	Jarque-Bera (JB):	1.033
Skew:	0.218	Prob(JB):	0.597

Figura 4: Resultados da primeira regressão
Fonte: Autor

Aplicamos em seguida o processo *stepwise* para eliminar da regressão variáveis cujos coeficientes não sejam significativos. O processo consiste em descartar do modelo as variáveis cujos coeficientes não são significativos, uma de cada vez e começando pela variável cujo coeficiente é menos significativo. O processo termina quando as variáveis que restam no modelo são todas significativas (Smith, 2018).

Dito isto, a primeira variável a ser descartada ou removida do modelo de regressão é percentagem de clientes sensíveis ao preço. Excluída essa variável, realiza-se uma segunda regressão cuja equação é dada por:

$$quota = \alpha + \beta 1 * vb\_tot + \beta 2 * num\_trx + \beta 3 * num\_prod\_tot + \beta 4 * perc\_dscnt + \beta 5 * perc\_prod\_mp + u \quad (8)$$

Os resultados da segunda regressão podem ser consultados na Figura 5.

OLS Regression Results

Dep. Variable:	quotas	R-squared:	0.827
Model:	OLS	Adj. R-squared:	0.820
Method:	Least Squares	F-statistic:	118.4
Date:	Tue, 05 Apr 2022	Prob (F-statistic):	7.92e-37
Time:	17:23:34	Log-Likelihood:	-187.80
No. Observations:	104	AIC:	385.6
Df Residuals:	99	BIC:	398.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	28.9259	1.963	14.736	0.000	25.031	32.821
num_trx_tot	-4.555e-05	1.23e-05	-3.714	0.000	-6.99e-05	-2.12e-05
vb_tot	7.333e-06	1.47e-06	4.981	0.000	4.41e-06	1.03e-05
num_prod_tot	-0.0031	0.000	-6.418	0.000	-0.004	-0.002
perc_dscnt	40.7839	2.935	13.897	0.000	34.961	46.607

Omnibus:	1.735	Durbin-Watson:	1.500
Prob(Omnibus):	0.420	Jarque-Bera (JB):	1.463
Skew:	0.290	Prob(JB):	0.481

Figura 5: Resultados da segunda regressão
Fonte: Autor

O valor do *Adj. R-squared* é agora 0,820. O valor do *p-value* do coeficiente da variável "percentagem de produtos de marca própria" diminuiu após este processo de *stepwise*.

Contudo, é preciso replicar o processo de *stepwise*, pois apesar do *p-value* da referida variável ter diminuído ainda é superior a 0.05. A variável é então removida e a equação da terceira regressão é dada por:

$$quota = \alpha + \beta 1 * vb\_tot + \beta 2 * num\_trx + \beta 3 * num\_prod\_tot + \beta 4 * perc\_dscnt + v \quad (9)$$

A Figura 6 ilustra os resultados de estimação da terceira regressão. Este já pode ser considerada a regressão final, pois já só contempla variáveis independes cujos coeficientes são significativos na explicação da quota de mercado, com *p-values* inferiores a 0,05 (5%).

OLS Regression Results

Dep. Variable:	quotas	R-squared:	0.827
Model:	OLS	Adj. R-squared:	0.820
Method:	Least Squares	F-statistic:	118.4
Date:	Tue, 05 Apr 2022	Prob (F-statistic):	7.92e-37
Time:	17:23:34	Log-Likelihood:	-187.80
No. Observations:	104	AIC:	385.6
Df Residuals:	99	BIC:	398.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	28.9259	1.963	14.736	0.000	25.031	32.821
num_trx_tot	-4.555e-05	1.23e-05	-3.714	0.000	-6.99e-05	-2.12e-05
vb_tot	7.333e-06	1.47e-06	4.981	0.000	4.41e-06	1.03e-05
num_prod_tot	-0.0031	0.000	-6.418	0.000	-0.004	-0.002
perc_dscnt	40.7839	2.935	13.897	0.000	34.961	46.607

Omnibus:	1.735	Durbin-Watson:	1.500
Prob(Omnibus):	0.420	Jarque-Bera (JB):	1.463
Skew:	0.290	Prob(JB):	0.481

Figura 6: Resultados da terceira regressão

Fonte: Autor

4.5 A métrica de estimação da quota de mercado obtida

Tendo em conta os resultados da terceira regressão, a métrica de estimação da quota de mercado obtida a partir dos dados de variáveis internas é dada pela equação:

$$quota\ estimada = 28{,}9259 + 0{,}000007 * vb\_tot - 0{,}00005 * num\_trx - 0{,}0031 * num\_prod\_tot + 40{,}784 * perc\_dscnt \quad (10)$$

A Figura 7 ilustra a relação entre quota de mercado observada e quota de mercado estimada com dados quadrimestrais para categoria de produto Doçaria. A Figura 8 ilustra a métrica obtida para a mesma categoria, mas com dados semanais. Com a métrica obtida, a empresa passa a contar com um valor estimado da quota de mercado para as categorias de produtos que integram *cluster* analisado, com periocidade semanal. É expectável que esta métrica permita obter valores estimados da quota próximos dos valores reais da quota, ao nível semanal e para as categorias de produto que integram o *cluster* analisado. O facto permite antecipar e justificar mudanças no valor da quota do mercado das categorias incluídas no *cluster* analisado e atuar, sempre que necessário.

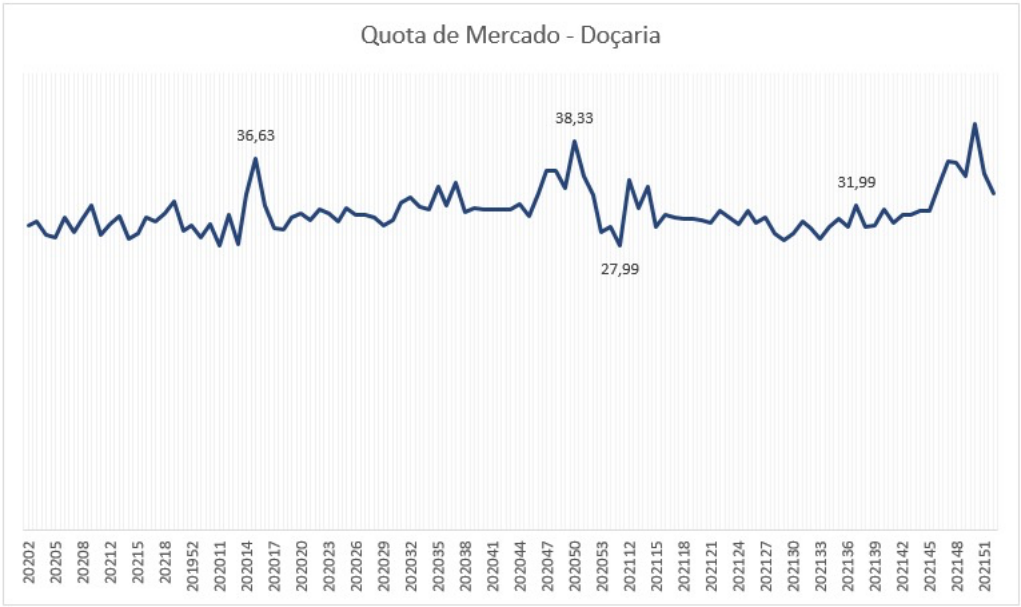

Figura 7: Quota de mercado estimada para a Doçaria (dados semanais)
Fonte: Autor

Capítulo 5 – Conclusão

O Trabalho Final de Mestrado teve como objetivo construir uma métrica baseada em variáveis transacionais internas para estimar a quota de mercado. A abordagem utilizada fez uso da matriz de correlações, da análise *k-means cluster*, do método de cotovelo e da análise de regressão.

A quota de mercado engloba o desempenho de vários *players* de um setor. Assim sendo, pode ser considerado redutor considerar só dados transacionais internos para estimar o valor da quota de mercado da empresa. Todavia, a obtenção dos dados dos competidores apresenta dificuldades relevantes e elevados custos. Por exemplo, não é fácil obter a informação sobre a atividade promocional dos outros *players*, o impacto da localização das respetivas lojas e a gama de produtos por categoria de produto oferecida pelos diversos *players*. Estes são exemplos de variáveis que podem influencia o valor da quota, mas não estão ao alcance da empresa. Mas a empresa dispõe de variáveis internas transacionais que lhe permitem uma estimativa melhor da referida quota de mercado e atuar, em função da informação obtida, sempre que necessário e em antecipação.

Os resultados mostram que a métrica obtida permita uma estimativa da quota de mercado da Doçaria que pode ser útil à empresa.

O trabalho teve várias limitações, nomeadamente no que refere aos dados disponíveis das quotas de mercado observadas serem quadrimestrais e a quota de mercado estimada pretendida ser semanal. O facto obrigou à agregação de categorias de produtos em *clusters*, para aumentar o número de observações da regressão que serviu para construir a métrica com base numa regressão que teve de utilizar dados quadrimestrais.

A métrica construída deriva de dados internos, o que torna a sua monitorização mais acessível e intuitiva. Olhando para os resultados do modelo final recolhido na primeira fase deste projeto, é possível é aplicar a métrica obtida com dados quadrimestrais (e/ou de 4 meses acumulados) a dados semanais, mas o facto não deixa de ser uma limitação à métrica obtida. As variâncias das observações não têm de ser a mesma. Idealmente, a métrica deveria ter sido construída com dados semanais, mas isso não foi possível.

A métrica final obtida apenas incidiu nas 4 seguintes categorias de produtos: limpeza geral, produção de papel e consumo, refrigerantes e Doçaria. Por isso apenas

se aplica a estas quatro categorias. Nada impede, em investigações futuras, aplicar a mesma análise para chegar a métricas do mesmo tipo mais adequadas a outras categorias de produto.

Estando o modelo de análise concluído e conseguindo replicar o mesmo às restantes categorias de produtos, a empresa passa a contar com uma nova métrica de avaliação da sua performance e uma maior capacidade reativa e de antecipação a flutuações das suas quotas de mercado. Isto é algo passível de ser explorado em investigações futuras.

Por fim, uma frequência maior de observações disponíveis de quotas reais observadas de mercado permitiria evitar ter de agregar produtos em *clusters* e construir métricas mais adequadas a cada categoria de produto. Isto é, no que toca ao tema da criação da métrica interna como réplica da quota do mercado, cada categoria de produto passaria então a contar com uma métrica individual, com maior capacidade de explicar o comportamento da respetiva quota de mercado.

O tópico da monitorização foi abordado, mas não com a necessária profundidade. Fica em aberto a possibilidade de criação de algoritmos de deteção de mudança sobre a métrica criada. Para desvios anormais, estes modelos criam alertas para a empresa, no sentido de promover maior capacidade de resposta face a alterações da quota de mercado.

Bibliografia

McAfee A., Brynjolfsson E. (2012, outubro). Big Data: The Management Revolution. Harvard Business Review. https://hbr.org/2012/10/big-data-the-management-revolution

Bean, R. (2021, fevereiro). Why Is It So Hard to Become a Data-Driven Company? Harvard Business Review. https://hbr.org/2021/02/why-is-it-so-hard-to-become-a-data-driven-company

Piketty, T. (2014). Capital in the Twenty-First Century. Gornick, J., and Jantti, M. (2014) Income Inequality: Economic Disparities and the Middle Class in Affluent Countries.

IBM. (2022). Big Data Analytics | IBM. [online] Available at: https://www.ibm.com/analytics/big-data-analytics [Accessed 5 Apr. 2022].

Kenway Consulting. (2020). Is Data Really 'The New Oil'? [online] Available at: https://www.kenwayconsulting.com/blog/data-is-the-new-oil/ [Accessed 5 Apr. 2022].

Hayes, A. (2021). Everything You Need to Know About Market Share. Investopedia. https://www.investopedia.com/terms/m/marketshare.asp

Rijmenam, Mark (2019). Why Our Fast-Changing World Requires Us to Change How We Collaborate. [online] The Digital Speaker - Future Tech Strategist. Available at: https://www.thedigitalspeaker.com/fast-changing-world-requires-change-collaborate/ [Accessed 4 Apr. 2022].

Sayyida, S., Hartini, S., Gunawan, S., & Husin, S. (2021). The Impact of the Covid-19 Pandemic on Retail ConsumerBehavior. Aptisi Transactions on Management (ATM), 5(1), 79-88

KPMG. (2021). Transitioning from ‘retail’ to ‘consumer commerce’. [online] Available at: https://home.kpmg/xx/en/home/insights/2021/01/transitioning-from-retail-to-consumer-commerce.html [Accessed 4 Apr. 2022].

Drezner, T. (1994). Optimal continuous location of a retail facility, facility attractiveness, and market share: An interactive model. Journal of Retailing, [online] 70(1), pp.49–64. Available at: https://www.sciencedirect.com/science/article/abs/pii/0022435994900280 [Accessed 4 Apr. 2022].

Dupre, K., & Gruen, T. W. (2004). The use of category management practices to obtain a sustainable competitive advantage in the fastmoving-consumer-goods industry. Journal of Business and Industrial Marketing, 19(7), 444–459. doi:10.1108/08858620410564391.

Weiss, D.L. (1968). Determinants of Market Share. Journal of Marketing Research, [online] 5(3), pp.290–295. Available at: https://journals.sagepub.com/doi/abs/10.1177/002224376800500307 [Accessed 4 Apr. 2022].

Nakanishi, M. & Cooper, L. (2010). Market-Share Analysis: Evaluating Competitive Marketing Effectiveness. Kluwer Academic Publishers, pp. 10-11 [Accessed 4 Apr. 2022]

Zhang, B. (2003) "Regression clustering," *Third IEEE International Conference on Data Mining*, pp. 451-458, doi: 10.1109/ICDM.2003.1250952 [Accessed 5 Apr. 2022]..

Späth, H. (2014). *Mathematical algorithms for linear regression*. Academic Press.

Berry, W. D. (1993). Understanding regression assumptions (Vol. 92). Sage.

Gujarati, D. N. (2021). Essentials of econometrics. SAGE Publications.

Mcleod, S. (2019). P-values and statistical significance. [online] Simplypsychology.org. Available at: https://www.simplypsychology.org/p-value.html [Accessed 6 Apr. 2022].

Altman, N., Krzywinski, M. (2017) P values and the search for significance. Nat Methods 14, 3–4. https://doi.org/10.1038/nmeth.4120

Smith, G. (2018) Step away from stepwise. J Big Data 5, 32. https://doi.org/10.1186/s40537-018-0143-6

Davenport T, Dyche J (2013). Big data in Big Companies (White paper). May 2013 (pp. 1–31). USA. Retrieved from http://swqlab.fov.unimb.si/moodle/pluginfile.php/9774/course/section/534/Big-Data-in-BigCompanies.pdf

Fan W, Bifet A (2013). Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, 14(2), 1–5. Retrieved from http://dl.acm.org/citation.cfm?id=2481246

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). Ieee.

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

Dayan, P. (1999). Unsupervised Learning. [online] Available at: https://web.math.princeton.edu/~sswang/developmental-diaschisis-references/dun99b.pdf.

Mendes De Morais, F., Jorge, D., Fernandes, M., & Santos. (2012). Clustering de Dados Biomédicos com Algoritmos baseados em Critérios Entrópicos. https://recipp.ipp.pt/bitstream/10400.22/4464/1/DM_FredericoMorais_2012_MECIM.pdf

Bhardwaj, A. (2020). Silhouette Coefficient - Towards Data Science. [online] Medium. Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c [Accessed 6 Apr. 2022].