

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Avaliação do impacto de ações promocionais nas vendas no contexto do retalho alimentar

Maria Francisca Azevedo Paupério



Mestrado Integrado em Engenharia Informática e Computação

Orientador: Vera Lucia Miguéis Oliveira e Silva

Co-orientador: Liliana Bernardino

23 de Julho de 2020

Avaliação do impacto de ações promocionais nas vendas no contexto do retalho alimentar

Maria Francisca Azevedo Paupério

Mestrado Integrado em Engenharia Informática e Computação

Resumo

O setor do retalho alimentar tem vindo a sofrer uma elevada pressão sobre a atividade promocional. As empresas neste setor disputam os mesmos consumidores e recorrem constantemente a ações promocionais para se destacarem e se manterem competitivas no mercado.

Grande parte do orçamento de marketing das empresas retalhistas é alocado em promoções. Contudo, Portugal é o quarto país europeu onde há uma maior atividade promocional e a eficácia das promoções tem vindo a diminuir. Como o investimento nas promoções é elevado, torna-se fundamental garantir uma estratégia promocional eficiente e adaptada à realidade extremamente promocional que se vive nos dias de hoje.

De facto, a eficácia das promoções passa por compreender os efeitos que resultam da ação promocional. No entanto, há uma tendência em avaliar, unicamente, os efeitos incrementais das vendas ou do número de transações do produto promovido, desprezando os efeitos que influenciam as vendas de outros produtos dentro do mesmo ecossistema. Assim sendo, partindo do pressuposto de que existem relações diretas produto a produto, procurou-se apurar e quantificar essas relações, de forma a identificar quais os produtos com maior impacto absoluto, positivo ou negativo, nas vendas de outros produtos. Este impacto surge associado ao efeito *Halo*, quando é positivo, ou à canibalização quando é negativo.

Tendo em consideração que estas relações só se tornam visíveis quando num par de produtos um deles está em promoção e o valor base das vendas é um dado indispensável para avaliar qualquer efeito de uma promoção, começa-se por propôr um modelo de previsão, através de uma rede neuronal *LSTM*, para calcular este valor. Seguidamente, percebendo o impacto da promoção no produto promovido aplica-se um modelo de regressão linear para extrair a força da relação.

Concluindo, esta dissertação contribuiu para a criação de uma ferramenta de apoio à decisão, que dará suporte ao negócio, aquando da definição da estratégia promocional, visando aumentar a eficiência da atividade promocional no retalho alimentar: evitando a promoção de produtos com alta taxa de substituíbilidade e favorecendo a promoção de produtos com alta taxa de complementaridade.

Palavras-chave: Canibalização, Efeito *Halo*, Valor Base das Vendas, Estratégia Promocional, Retalho Alimentar, Regressão Linear, *LSTM*

Abstract

There has been an increasing pressure on the promotional activity for the food retail industry. Companies from this field of work share customers and rely on promotional activity to stand out and remain competitive in the market.

Promotions and sales represent a big portion of the marketing budget in retail companies. Moreover, Portugal is the fourth European country where promotional activity is more intense, and its effectiveness is decreasing. As investment in promotions is considerable, so is of the highest importance to assure an efficient promotional strategy that is suitable for today's reality.

In fact, measuring promotions effectiveness goes by understanding the results from these promotional activities. However, there is the practice of evaluating, exclusively, the incremental effects that has on sales, or on the number of sold items of the discounted product, not considering the impact that it might have on the other products' sales. Assuming there are relations product by product, the goal was to identify and quantify the impact of those relations, in order to identify the most impactful ones, positively or negatively, when selling products. When there is a positive impact on the product sales it is called Halo effect, on the other hand, when there is a negative effect it is called switching.

These relations are only clear when in a couple of products, one is at promotion and the sales baseline are necessary to evaluate the promotion impact. To estimate this baseline it is used a predictive model with a LSTM neuronal network. Afterwards, it is used a linear regression model to extract the relation power.

In conclusion, this master thesis dissertation contributed to the creation of a decision supporting tool to assist the business when defining the promotional strategy. Its goal is to improve the effect of the promotional activity in the food retail industry: avoiding promotions on products with high replaceability rate and favoring the ones with high complementarity rate.

Keywords: Switching, Halo Effect, Baseline, Promotional Strategy, Food Retailing, Linear Regression, LSTM

Agradecimentos

Esta dissertação é o culminar de um percurso académico desafiante e com alguns obstáculos, muitos deles colocados por mim, que insistia em testar os meus limites, procurando ser melhor. Foram quatro meses que replicaram e resumiram os últimos cinco anos e que, mais uma vez, acredito que me fizeram crescer.

Tive a sorte de desenvolver este projeto numa área que, até então, tinha sido pouco explorada por mim e que me suscitava muita curiosidade - a mineração de dados. Para além disso, tive ainda mais sorte por ter sido acompanhada e orientada pela Professora Vera Miguéis, que foi incansável ao longo de todo o processo e a quem deixo o meu maior agradecimento. Pelos desafios que me lançou, pela orientação interessada, pelo conhecimento partilhado, pelas críticas minuciosas e, acima de tudo, pelo apoio e motivação, o meu muito obrigada.

Gostava, também, de agradecer ao Continente pela oportunidade que me deram de poder desenvolver este projeto final num ambiente empresarial e para um negócio que me diz tanto. Em especial, a toda a equipa na qual integrei, por tudo o que me ensinaram e pela simpatia e disponibilidade, não podendo deixar de destacar as 3 pessoas que me orientaram mais de perto: o Filipe Miranda, pela disponibilidade e apoio permanente, ao longo de todo o processo, a Ana Freitas, pelo cuidado e por todos os conselhos e conhecimento partilhado e, por fim, a Liliana Bernardino, por quem tenho uma grande admiração, pelos desafios lançados e por todo o apoio.

Não queria deixar de referir o Jason Brownlee, que é um especialista em *Machine Learning* que se dedica ajudar os outros a serem melhores nesta área, através da partilha do seu conhecimento e com quem tive o prazer de falar para trocar algumas ideias.

Um agradecimento enorme aos meus amigos, que nunca me deixaram desistir e que sempre se preocuparam comigo, com especial destaque para o Luís Fragoso, o João Bordado, o André Reis e o Padre Basílio, que foram acompanhando o desenvolvimento da tese e me motivaram e ajudaram a fazer um trabalho melhor.

Por último, um agradecimento sem fim aos meus pais e irmãos, Pedro, Beatriz e Sofia, por acreditarem em mim e nas minhas capacidades e pelo apoio incondicional, mesmo quando as minhas inseguranças os levam ao limite.

Francisca Paupério

*“All models are wrong,
but some are useful.”*

George Box

Conteúdo

1	Introdução	1
1.1	Contexto	1
1.2	Motivação e Objetivos	2
1.3	Estrutura da Dissertação	3
2	Enquadramento Teórico	5
2.1	O Retalho Alimentar em Portugal	5
2.2	Ações Promocionais	6
2.3	Avaliação das Promoções	8
2.3.1	Conceitos Importantes	8
2.3.2	Retalhistas vs. Fabricantes	10
2.3.3	Efeitos das Promoções nas Vendas	10
2.3.4	Teorias da Decomposição dos Efeitos Promocionais	12
2.3.5	Metodologias Relacionadas	13
2.4	Mineração de Dados no Retalho Alimentar	16
2.4.1	Processo de Mineração de Dados	17
2.4.2	Métodos de Aprendizagem Supervisionados	19
3	Metodologia e Estruturas de Dados	21
3.1	Visão Geral	21
3.2	Modelo Principal	22
3.2.1	Entendimento do Problema	23
3.2.2	Compreensão dos Dados	24
3.2.3	Preparação dos Dados - Modelo Secundário	25
3.2.4	Modelação	32
3.2.5	Avaliação	33
3.2.6	Desenvolvimento	33
4	Resultados	35
4.1	Caso de Estudo	35
4.1.1	Afinação de Parâmetros do Modelo <i>LSTM</i>	35
4.1.2	Apuramento da Relação entre um Par de Produtos	42
4.1.3	Outros Resultados	46
5	Conclusões, Limitações e Trabalho Futuro	49
5.1	Conclusões	49
5.2	Limitações e Trabalho Futuro	50
	Referências	53

Lista de Figuras

1.1	Logótipo do Continente.	2
2.1	Tipos de promoções no setor do retalho.	6
2.2	Objetivos das promoções retalhistas.	7
2.3	Avaliação do impacto da promoção nas vendas do produto promovido.	9
2.4	Síntese dos efeitos das promoções para um <i>bump</i> positivo num produto promovido.	12
2.5	"Repartição Geral da Promoção".	14
2.6	Estrutura do <i>CRISP-DM</i>	18
2.7	O módulo de repetição numa <i>LSTM</i> contém quatro camadas de interação.	20
3.1	Processo <i>CRISP-DM</i> adaptado ao contexto do problema.	22
3.2	Tabela de dados inicial com duas variáveis por dia.	28
3.3	Tabela de dados final, considerando um intervalo de 7 dias.	28
4.1	Vendas da Batata Frita Ondulada Marca Continente (2016 a 2019).	36
4.2	Vendas da Batata Frita Ondulada Marca Continente em janeiro de 2016.	36
4.3	Diagrama de caixas para afinação do tamanho da <i>lag</i>	38
4.4	Diagrama de caixas para afinação do número de <i>epochs</i>	38
4.5	Diagrama de caixas para afinação do tamanho do lote.	39
4.6	Diagrama de caixas para afinação do número de neurónios.	39
4.7	Curva de aprendizagem do modelo para os conjuntos de dados de treino e de validação.	40
4.8	Vendas reais versus vendas previstas a partir dos dados de validação e de teste no ano 2019 (batata frita marca Continente).	41
4.9	Vendas reais versus valor base das vendas para o ano 2019 (batata frita marca Continente).	41
4.10	Vendas reais versus vendas previstas a partir dos dados de validação e de teste no ano 2019 (batata frita marca <i>Ruffles</i>).	42
4.11	Vendas reais versus valor base das vendas para o ano 2019 (batata frita marca <i>Ruffles</i>).	42
4.12	Gráfico da Regressão Linear do par (A em promoção, B sem promoção).	43
4.13	Gráfico da Regressão Linear do par (A em promoção, B sem promoção) após remover o primeiro <i>outlier</i>	44
4.14	Gráfico da Regressão Linear do par (A em promoção, B sem promoção) após remover o segundo <i>outlier</i>	44
4.15	Gráfico da Regressão Linear do par (A sem promoção, B em promoção).	45
4.16	Gráfico da Regressão Linear do par (A sem promoção, B em promoção) após remover <i>outliers</i>	46
4.17	Gráfico da Regressão Linear do par (C em promoção, D sem promoção).	47

4.18	Gráfico da Regressão Linear do par (C sem promoção, D em promoção).	48
------	---	----

Lista de Tabelas

3.1	Estrutura comercial.	23
3.2	Estrutura orientada ao cliente - <i>Virtual Product Hierarchy (VPH)</i>	24
3.3	Estrutura da tabela base agrupada ao produto, ao dia e à loja.	26
4.1	Valores considerados para cada parâmetro.	37
4.2	Tabela de resultados referente à regressão linear da Figura 4.12.	43
4.3	Tabela de resultados referente à regressão linear da Figura 4.13.	44
4.4	Tabela de resultados referente à regressão linear da Figura 4.14.	45
4.5	Tabela de resultados referente à regressão linear da Figura 4.15.	45
4.6	Tabela de resultados referente à regressão linear da Figura 4.16.	46
4.7	Tabela de resultados referente à regressão linear da Figura 4.17.	47
4.8	Tabela de resultados referente à regressão linear da Figura 4.18.	48

Abreviaturas

BOGO	<i>Buy-One-Get-One</i>
CNT	Continente
DC	Departamento Comercial
POS	<i>Point Of Sale</i>
SKU	<i>Stock Keeping Unit</i>
VPH	<i>Virtual Product Hierarchy</i>

Capítulo 1

Introdução

1.1 Contexto

O setor do retalho alimentar tem vindo a sofrer uma elevada pressão sobre a atividade promocional. As empresas neste setor investem uma parte significativa dos seus orçamentos de marketing em promoções, com o intuito de se destacarem dos seus concorrentes. As promoções são dos instrumentos de marketing mais eficazes para estimular o aumento das vendas a curto prazo e mais atrativos para gestores orientados a objetivos, que procuram retornos rápidos para os seus investimentos [VHN08, MMR20].

Após uma promoção, rapidamente se calcula o aumento das vendas durante o período promocional, em relação às vendas expectáveis para o mesmo período sem promoção, a fim de se avaliar o resultado do evento promocional. Embora seja simples inferir se o resultado foi positivo ou negativo para o produto que estava em promoção, torna-se bastante desafiante interpretar a performance real da promoção e calcular o seu retorno da perspectiva de um retalhista. De facto, para um retalhista a visão altera quando se considera o produto como substituto de outros produtos na mesma loja.

De uma forma geral, quando se analisa o efeito promocional de uma determinada oferta, tende-se a focar no efeito incremental das vendas, no número de transações ou até num aumento de clientes. Muitas vezes, o efeito da oscilação das vendas entre produtos - canibalização - acaba por ser menosprezado.

Como o investimento nas promoções é elevado, é fundamental garantir que os efeitos da atividade promocional sejam medidos corretamente; é essencial calcular o impacto das promoções entre produtos da mesma categoria ou de categorias diferentes, que possam ser percebidos como substitutos.

Hoje em dia, as empresas retalhistas geram, capturam e armazenam grandes quantidades de dados, diariamente, com ênfase em dados transacionais. Com o crescimento do volume e fluxo de dados (*Big Data*), passou a ser possível extrair informação dos dados transacionais para auxiliar os gestores nos processos de tomada de decisão e na definição de estratégias promocionais mais assertivas. Dado o aumento da quantidade de dados disponíveis, emergiu a oportunidade para

as ferramentas e técnicas analíticas avançadas evoluírem. Esta evolução despoletou progressos significativos na modelação e análise de dados no contexto do retalho alimentar, com foco nas promoções, dada a sua relevância.

As empresas de retalho alimentar em Portugal não são exceção. A empresa em análise neste projeto, o Continente, está a par da evolução das técnicas e ferramentas no âmbito da análise avançada e está a investir cada vez mais em equipas com habilidades nessa área.



Figura 1.1: Logótipo do Continente.

O Continente foi a primeira cadeia de hipermercados em Portugal. Abriu a primeira loja em 1985 e tem, neste momento, 302 lojas espalhadas por todo o país.

Em 2007, lançou um cartão de cliente (cartão Continente), através do qual tem vindo a oferecer descontos e outras regalias, com o intuito de fidelizar os seus clientes. Em 2017, o cartão atingiu 3,7 milhões de contas ativas.

A quantidade de informação que o Continente consegue extrair dos pontos de venda tem vindo a crescer e o cartão torna possível associar transações a pessoas. Para se manter numa posição de liderança no setor do retalho alimentar em Portugal, o Continente procura tirar cada vez maior proveito da informação e das ferramentas a que tem acesso.

1.2 Motivação e Objetivos

Em 1999, Inman e Winer [IW99] alegaram que quase 60% das compras de supermercado não são planeadas e são consequência de decisões tomadas dentro das lojas. Na sequência desta referência, em 2004, um estudo [DNV04] investigou as promoções como premissa para uma maior despesa em loja. Este estudo não só constatou que as promoções aumentavam o consumo numas categorias de produto mais do que em outras, como defendia que fomentavam o aumento global do valor da cesta em todas as categorias (não só nas categorias dos produtos promovidos). Estas observações sustentam a extrema importância das promoções como ferramenta indispensável para o sucesso de uma empresa de retalho.

Para avaliar o impacto financeiro das promoções nas vendas nos supermercados, os gestores devem estimar que parte da nova procura provém da canibalização de outros produtos comparando com outras fontes do aumento das vendas [MMR20]. A maior parte dos estudos têm seguido uma abordagem holística para medir o aumento de uma promoção, no qual o desempenho não é

analisado, apenas, como uma promoção individual, mas como uma promoção que pode impactar as vendas de outros itens dentro da mesma categoria [Pes18]. Ailawadi et al., 2006 [AHCT06], para além de estudar os efeitos dentro da categoria, também defende que pode existir um efeito positivo, externo à categoria, que se reflete num aumento do total das vendas da loja - chamando a este efeito *Halo*.

Para o Continente, as promoções são também um foco importante na estratégia do negócio e o seu estudo aprofundado é inevitável. No entanto, suspeita-se que os efeitos das promoções não se fazem sentir unicamente dentro da mesma categoria. Assim sendo, surgiu a necessidade de aprofundar a análise aos efeitos promocionais, procurando relações produto a produto.

Com base nas evidências bibliográficas sobre a existência de um efeito promocional positivo e outro negativo, considerou-se relevante aferir os efeitos ao nível do produto, assumindo que os efeitos são complementares.

A canibalização é vista como efeito de substituíbilidade, que faz diminuir as vendas de produtos não promovidos em prol de um aumento das vendas do produto promovido. Por outro lado, o efeito *Halo* pode ser visto como um efeito de complementaridade, que faz aumentar as vendas de produtos sem atividade promocional.

Desta forma, surge a necessidade de procurar uma relação para cada par de produtos (A, B). Quando a promoção de A se traduz num aumento de vendas em ambos os produtos, estes são complementares. Por outro lado, caso implique uma diminuição nas vendas de B, os produtos são substitutos.

O desenvolvimento deste projeto resulta numa ferramenta de apoio à definição da estratégia promocional, que permite evitar promover produtos com uma alta taxa de substituíbilidade (canibalização) e favorecer promoções com altas taxas de complementaridade (efeito *Halo*).

1.3 Estrutura da Dissertação

Para além da introdução, este documento contém quatro capítulos adicionais.

No capítulo 2, é feito um enquadramento teórico onde são definidos alguns conceitos triviais no contexto do retalho alimentar. Adicionalmente, é apresentada uma revisão da literatura existente sobre a decomposição das fontes da nova procura decorrente das promoções. Por fim, são descritas as principais pesquisas empíricas que investigam os efeitos do aumento de vendas, com foco na canibalização e no efeito *Halo*, os quais fornecem uma base para explicar a metodologia aplicada nesta dissertação.

No capítulo 3, é exposta a metodologia desenhada para dar resposta às necessidades e objetivos anteriormente evidenciados. Seguindo uma estrutura previamente definida, o capítulo relata detalhadamente todo o processo de implementação da metodologia.

No capítulo 4, são apresentados e analisados os resultados da metodologia aplicada a dois exemplos práticos, a fim de demonstrar as informações que se podem retirar do modelo desenvolvido.

Para finalizar, o capítulo 5 salienta as conclusões finais e destaca as limitações encontradas e potenciais trabalhos futuros para a continuação deste projeto.

Capítulo 2

Enquadramento Teórico

O setor do retalho alimentar é um setor altamente competitivo e promocional. No entanto, Portugal destaca-se por ser um dos países europeus onde há uma maior atividade promocional [Por19].

Este capítulo começa por fazer um enquadramento no setor do retalho alimentar em Portugal, explicando, também, a visão e o comportamento do consumidor perante ações promocionais. De seguida, é feita uma revisão teórica dos principais conceitos relacionados com as promoções, concentrando-se em mais detalhe no estudo dos efeitos das promoções nas vendas, na ótica dos retalhistas. Por fim, são apresentadas algumas metodologias que serviram de ponto de partida para o desenvolvimento da metodologia desta dissertação.

2.1 O Retalho Alimentar em Portugal

O panorama no mercado nacional revela a importância central do fenómeno promocional, que se torna notória quando comparamos a realidade portuguesa com a europeia. Em 2019, estudos [Por19] revelaram que quase metade (46%) das vendas em Portugal no ano 2018 foram realizadas em promoção. Na verdade, no ano 2018, foram registadas vendas em promoção na ordem dos 3,5 mil milhões de euros, posicionando Portugal como o quarto país europeu onde a promoção adquire maior peso. No entanto, mais de metade das vendas em promoção correspondem a vendas não incrementais, ou seja, vendas que seriam feitas de qualquer forma (com ou sem promoção).

Relativamente ao comportamento do consumidor português, quase metade (contrapondo com 1/3 dos consumidores europeus) assume que, mesmo que não mude de loja em função das promoções, procuram-nas ativamente quando fazem as suas compras. Para além disso, um terço dos consumidores admite escolher a marca de acordo com as promoções em vigor. No entanto, apesar da importância atribuída pelo consumidor à existência de promoções, tornam-se cada vez mais

preponderantes outros fatores que influenciam a decisão de compra: 70% dos consumidores procuram alimentação saudável, 2/3 mostram-se dispostos a pagar mais por maior qualidade e metade admitem pagar mais para poupar tempo [Por19].

Entre 2017 e 2018, mais de metade (54%) das promoções nas categorias dos bens de grande consumo diminuíram a sua eficiência. Com efeito, a consultora Nielsen [Por19] prevê que, mantendo-se a estratégia atual, a eficiência não vai aumentar em Portugal, sendo necessário encontrar e adaptar novas estratégias. Adicionalmente, a consultora afirma ainda que é necessário explorar outras oportunidades de trabalhar as promoções.

2.2 Ações Promocionais

As promoções são conhecidas como um evento de marketing focado na ação, cujo objetivo é estimular o cliente com o propósito de influenciar o seu comportamento [BN90].

Também conhecido como *call-to-action*, é um evento que aciona uma ação. Geralmente, uma promoção é acompanhada por algum tipo de comunicação, pois é temporária e a comunicação ajuda a aumentar a visibilidade e a alcançar mais clientes [BN93].

As promoções são projetadas para diferentes propósitos e para diferentes audiências, sendo possível distinguir-se diferentes tipos de promoções. Na Figura 2.1 estão esquematizados os três principais tipos de promoções, que se distinguem pela entidade que promove e pelo público-alvo. Posteriormente, é dada uma explicação para cada tipo de promoção.

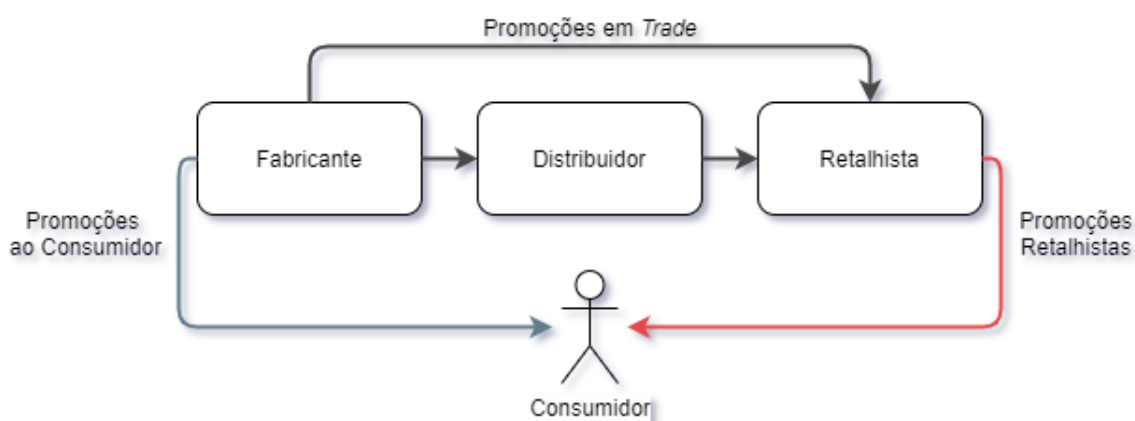


Figura 2.1: Tipos de promoções no setor do retalho.

- **Fabricante → Consumidor:** oferecidas diretamente pelo fabricante ao consumidor, para estimular compras num intervalo curto de tempo, após a receção da promoção;
- **Fabricante → Distribuidor/Retalhista:** oferecidas no canal de distribuição ("em *trade*"), com o intuito de estimular os retalhistas a oferecer promoções aos consumidores, ou as distribuidoras a oferecer promoções aos retalhistas;

- **Retailista → Consumidor:** oferecidas pelos retalhistas aos consumidores para estimular o aumento das vendas de determinado produto, categoria ou loja.

Tendo em consideração que este projeto é desenvolvido no domínio de uma empresa retalhista, o foco restringe-se, de agora em diante, às promoções retalhistas.

Para além do momento para o qual é planeada, os principais elementos de uma ação promocional retalhista são os objetivos, os diferentes tipos de descontos de preço, o meio no qual o desconto é oferecido e o meio no qual é comunicado [BN93].

Antes de definir o desconto ou o meio em que este será oferecido e comunicado, o retalhista tem de definir os objetivos que pretende atingir com a promoção. Na Figura 2.2 são enumerados alguns objetivos das promoções [OBM15].



Figura 2.2: Objetivos das promoções retalhistas.

Partindo de um ou mais objetivos bem definidos, escolhe-se o tipo de desconto de preço. Existem inúmeros tipos de desconto [BN93], que vão surgindo da inovação dos próprios retalhistas:

- **Corte de Preço:** redução temporária do preço de um produto;
- **Cupão:** emissão de um cupão de desconto para determinado produto;
- **BOGO:** promoções "Leve 3 Pague 2"; inclui produtos com oferta de produtos complementares;
- **Sorteio:** o cliente entra num concurso no qual tem a oportunidade de ganhar um prémio;
- **Amostra/Teste Grátis:** oferta de amostras do produto para dar a conhecer e incentivar a compra;
- **N-por:** o retalhista oferece um desconto no preço se o cliente comprar N produtos;
- **Cartão de Desconto:** promoção dirigida aos clientes fidelizados;

- **Rebate:** o cliente recebe um aviso de um desconto de que pode usufruir, perante o comprovativo da receção do aviso;
- **Pacote (Pack) de Promoção:** desconto para produtos complementares vendidos em conjunto.

O meio em que o desconto é oferecido está muito relacionado com os objetivos que se querem atingir com a ação promocional. Enquanto que alguns são oferecidos em loja, para que qualquer cliente que visite a loja possa usufruir dele, outros são emitidos e enviados diretamente para grupos específicos de clientes. Por exemplo, para atrair novos consumidores para determinado produto, a técnica de envio de cupões de desconto para clientes fidelizados (cuja informação para entrar em contacto está na base de dados do retalhista) é a mais adequada. Por outro lado, quando se faz o lançamento de um produto novo, é habitual acompanhar o lançamento com a oferta de amostras do produto em loja, para dar a conhecer.

Por fim, é muito importante comunicar os descontos de preços para que os consumidores tenham conhecimento. Normalmente, os retalhistas recorrem à comunicação em loja (*displays*), folhetos semanais e publicidade em diferentes meios, como televisão, rádio, *outdoors* e comunicação digital. A definição do meio em que o desconto é comunicado está relacionado com a dimensão que se quer dar à promoção. Recorre-se à comunicação em etiquetas nas prateleiras de supermercado para descontos pequenos e mais habituais e à comunicação em meios externos à loja, como publicidade em televisão, para promoções de maior dimensão, às quais se pretende dar muita visibilidade.

2.3 Avaliação das Promoções

Como foi referido anteriormente, no enquadramento do retalho alimentar em Portugal, os estudos da Nielsen detetaram uma diminuição da eficiência promocional e afirmam ser necessário explorar outras oportunidades de trabalhar as promoções. No entanto, para explorar novas oportunidades, é necessário extrair os resultados das promoções que têm vindo a ser feitas, bem como, saber avaliá-las corretamente.

Uma promoção é vista como um evento que provoca uma alteração no decorrer natural das vendas de um supermercado. Para se apurar uma alteração, tem que haver um valor esperado, isto é, um valor base a partir do qual se pesa a alteração.

2.3.1 Conceitos Importantes

Para um melhor entendimento dos métodos de avaliação das promoções, é necessário definir alguns conceitos neste contexto:

- **Valor Base (*Baseline*):** O valor base é uma estimativa do valor das vendas que teria sido registado caso não tivesse ocorrido uma promoção. O cálculo do valor base é visto como

um cálculo preditivo das vendas, num determinado período de tempo, para o qual não são considerados fatores externos que possam afetar a sua performance natural.

- **Impacto nas Vendas (*Bump*):** O impacto nas vendas remete para a diferença entre as vendas reais e o valor base do produto promovido. Este impacto pode ser positivo, quando há um incremento nas vendas, ou negativo, quando as vendas ficam aquém do expectável.
- **Sazonalidade:** A sazonalidade é um fator externo que influencia as vendas e deteta-se através da presença de variações que ocorrem em intervalos regulares inferiores a um ano.
- **Tendência:** A tendência caracteriza o padrão de evolução das vendas e remete para uma subida ou descida constante da média das vendas. Normalmente, avalia-se de ano para ano.
- **Rotatividade:** A rotatividade de um produto é um indicador que revela o período em que o produto se mantém na prateleira do supermercado antes de ser substituído; é um dos indicadores mais importantes dentro da gestão de *stocks* e serve para alinhar o setor de compras com o de vendas.

Na Figura 2.3, está representado o método de avaliação do impacto de uma ação promocional nas vendas do produto promovido, recorrendo aos conceitos de *bump* e *baseline*.

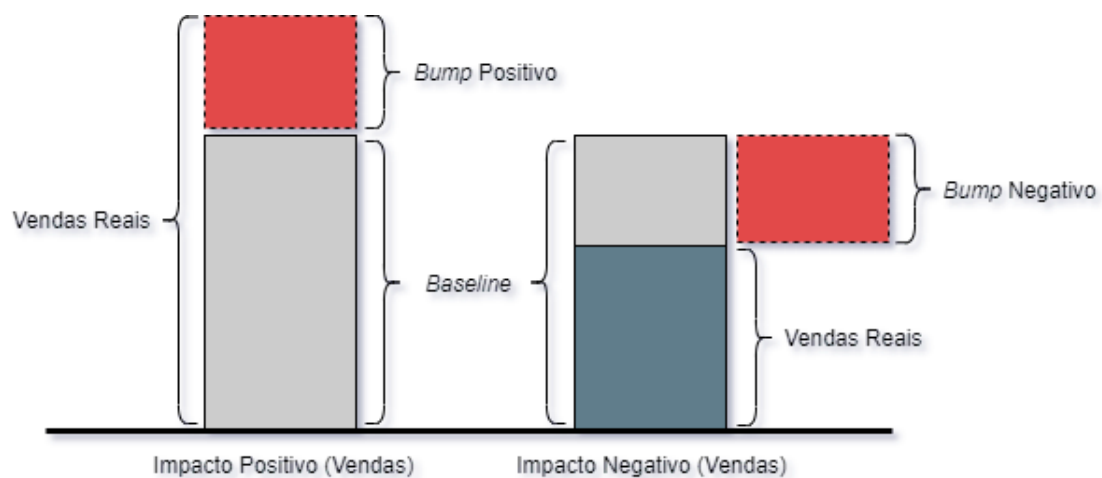


Figura 2.3: Avaliação do impacto da promoção nas vendas do produto promovido.

Ao longo das últimas décadas, têm surgido inúmeros estudos no âmbito da avaliação das promoções. As primeiras abordagens estavam relacionadas com a perda da lealdade às marcas em prol das promoções. Os investigadores começaram por avaliar o impacto das promoções do ponto de vista dos fabricantes [AHCT06].

Embora seja simples inferir se o resultado foi positivo ou negativo para o produto que estava em promoção, torna-se bastante desafiante interpretar a performance real da promoção e calcular

o seu retorno da perspectiva de um retalhista, pois um *bump* positivo nem sempre se reflete num impacto global positivo.

Entretanto, para tirar proveito de uma maior disponibilidade de dados recolhidos diariamente pelos pontos de venda (*POS*), muitos investigadores orientaram a sua investigação para a avaliação das promoções na ótica dos retalhistas [BN93].

2.3.2 Retalhistas vs. Fabricantes

No contexto do retalho alimentar, identificam-se duas entidades que negociam regularmente quais os produtos que entram para a gama das lojas, os seus preços e o valor e momento em que acontecem as atividades promocionais associadas aos produtos. Estas decisões não afetam as duas entidades da mesma forma.

Quando se avalia o impacto de uma promoção, por um lado existe a visão dos fabricantes e fornecedores que procuram um incremento nas vendas de um determinado produto; por outro lado, existe a visão do retalhista que procura um aumento no global das vendas de uma categoria ou loja. Para os fabricantes, se o incremento representar um declínio nas vendas dos produtos das marcas concorrentes, a promoção é mais benéfica. Por oposição, para o retalhista se o incremento for proveniente de uma troca entre marcas ou produtos, dentro da mesma loja, a receita será igual ou menor devido a margens inferiores nos produtos promovidos [BN93].

Assim sendo, é necessário ter uma compreensão completa dos efeitos das promoções para projetar estratégias promocionais eficazes do ponto de vista dos retalhistas, especialmente, quando o *bump* do produto promovido é positivo.

2.3.3 Efeitos das Promoções nas Vendas

A identificação e medição dos efeitos das promoções nas vendas, tem sido o grande desafio dos retalhistas [AG⁺14]. A avaliação do impacto da promoção apenas nas vendas do produto promovido não é suficiente da perspectiva de um retalhista. No entanto, é proposto que este incremento seja modelado, fazendo corresponder o incremento em determinado período de tempo t a um comportamento do consumidor que o justifique [VHN17].

Assim sendo, começa-se por distinguir os efeitos das promoções pelo período em que é detetado: os efeitos a curto-prazo, que se fazem sentir durante o evento promocional e os efeitos a longo-prazo, que só se fazem sentir depois do evento promocional terminar [GNA10].

De seguida, são enumerados os principais efeitos das promoções, descrevendo o comportamento do consumidor que o despoleta:

- **Efeitos a curto-prazo**

- **Maior Taxa de Consumo:** Engloba um aumento de consumo em virtude da promoção. O cliente compra por impulso mais quantidade e consome mais, por consequência. Este comportamento é o mais vantajoso para os retalhistas e vem refletido num

bump positivo do produto promovido, pois está associado a um efeito de expansão/-crescimento da categoria.

- **Canibalização:** Entre marcas, categorias ou lojas. O Cliente deixa de comprar o que está acostumado para comprar o que está em promoção. Por exemplo, em proveito da promoção, o cliente pode trocar uma garrafa de água por um refrigerante (canibalização entre categorias), uma Coca-Cola por uma Pepsi (canibalização entre marcas), comprar numa loja diferente para aproveitar a promoção (canibalização entre lojas) e, por fim, trocar uma referência por outra da mesma marca. A canibalização não é vantajosa para o retalhista, sendo a canibalização entre referências da mesma marca igualmente não proveitosa para o fabricante [VHLW04].
- **Novos Consumidores:** Abrange os clientes que não compravam o produto e que passam a comprar em proveito da promoção. O surgimento de novos clientes pode estar relacionado com uma mudança de hábitos dos clientes ou com os efeitos da canibalização anteriormente referidos. Caso não esteja associado ao efeito da canibalização, este efeito é vantajoso para o retalhista e vem refletido num *bump* positivo do produto promovido, pois está associado a um efeito de expansão/crescimento da categoria.

- **Efeitos a longo-prazo**

- **Stockpiling:** O cliente compra em mais quantidade e não aumenta o consumo; compra por antecipação; opta por armazenar os produtos em casa, o que leva a uma diminuição das vendas nos períodos pré ou pós-promocionais. Os produtos perecíveis não têm as características necessárias para serem armazenados durante muito tempo, por isso, este comportamento verifica-se mais com produtos passíveis de serem congelados ou produtos com um grande prazo de validade. Este efeito não é vantajoso para o retalhista.
- **Fidelização:** O cliente, devido ao estímulo que recebeu da promoção, deixa de comprar um produto a que era fiel. Acontece quando um cliente, em proveito da promoção experimenta um produto novo, que lhe agrada mais que o que estava habituado a comprar. Este efeito é comparado a uma canibalização a longo-prazo, sem o prejuízo da margem ser inferior durante o período promocional. Este efeito pode ser positivo ou negativo para o retalhista, pois dependerá das margens dos produtos.

Tendo em consideração que os efeitos estão relacionados entre si e que só é possível avaliar se alguns efeitos são positivos ou negativos ao analisar o comportamento que o justifica, na Figura 2.4 é feita uma síntese dos efeitos das promoções que originam um *bump* positivo num produto promovido.

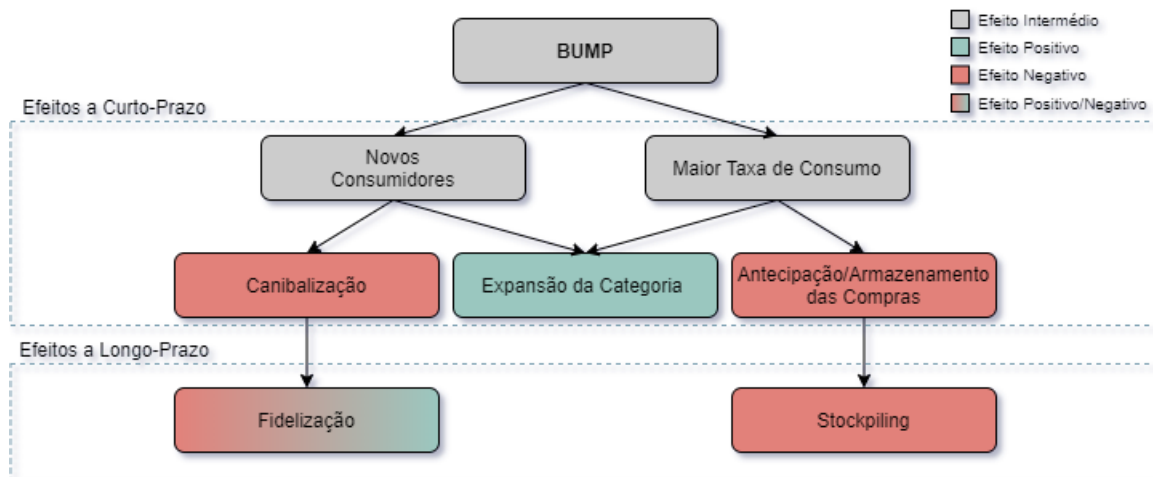


Figura 2.4: Síntese dos efeitos das promoções para um *bump* positivo num produto promovido.

Adicionalmente, um estudo de Ailawadi et al., em 2006 [AHCT06], expõe, pela primeira vez, um efeito a curto-prazo chamado efeito *Halo*, que defende que uma promoção numa categoria pode afetar as vendas em outras categorias, durante o período da promoção. Por um lado, o efeito é considerado positivo se a promoção influenciar o cliente a comprar produtos de outras categorias dentro da mesma loja, que não teria comprado caso não existisse essa promoção. Por outro lado, o efeito é negativo, se a promoção resultar numa redução das compras em outras categorias.

Na verdade, são muitos os efeitos, ao nível do comportamento dos clientes e consequentemente das vendas, que podem resultar de uma promoção. No entanto, a medição destes efeitos exige que a decomposição do incremento das vendas nos diferentes tipos de efeitos seja investigada de um modo quantitativo.

2.3.4 Teorias da Decomposição dos Efeitos Promocionais

A análise quantitativa dos efeitos promocionais despertou a curiosidade de muitos investigadores e foram desenvolvidas duas teorias muito distintas: a teoria da elasticidade, também conhecida por teoria das vendas brutas, e a teoria das vendas unitárias, também conhecida por teoria das vendas líquidas [VHLW04].

Teoria da Elasticidade

A teoria da elasticidade surgiu, em 1988, quando Gupta [Gup88] identificou os comportamentos dos consumidores acionados por uma promoção, como fundamento para a sua perspetiva inovadora da decomposição dos efeitos promocionais.

Gupta implementa uma análise de elasticidade para avaliar o impacto relativo da promoção em determinados comportamentos de compra: a escolha da marca, diretamente relacionada com um incremento proveniente da canibalização entre marcas; a aceleração da compra, relativa a uma antecipação no momento da compra; e a decisão da quantidade de produto associada ao armazenamento (*stockpiling*). Posto isto, a elasticidade das vendas é calculada através da soma

das elasticidades destas três variáveis. O modelo da elasticidade pressupõe que a escolha da marca e da quantidade dependem da decisão do momento da compra e são independentes entre si. Esta análise foi implementada considerando diversas marcas e diferentes tipos de promoção e Gupta concluiu que 84% do incremento nas vendas deriva da canibalização entre marcas, no máximo 14% deriva de uma antecipação do momento da compra e menos de 2% deriva de um aumento da quantidade para armazenamento. Olhando só a cortes de preço, Gupta obteve resultados de canibalização que se aproximam dos 98% [Gup88].

Alguns modelos estendem o modelo de Gupta, para integrar mais variáveis comportamentais [Chi91, Chi93], ou para ter em conta a heterogeneidade dos consumidores [VHN17].

Teoria das Vendas Unitárias

Em 2003, surge a primeira abordagem à decomposição dos efeitos das promoções baseado em vendas unitárias [VHGW03]. São derivadas expressões analíticas que transformam a decomposição em elasticidades numa decomposição em vendas unitárias, mostrando as diferenças entre as duas abordagens. Os autores de Van Heerde et al., 2003 [VHGW03], concluem que, sendo por definição a intenção das promoções promover um aumento das vendas, é preferível avaliar a eficácia de uma promoção através de uma análise às vendas unitárias do que às elasticidades.

Em 2004, Van Heerde et al. [VHLW04] conclui que um aumento de 100 unidades numa marca promovida, resulta de uma perda de 33 unidades para outras marcas da mesma categoria e de 32 unidades de períodos pré ou pós promocionais, refletindo-se num aumento de 35 unidades provenientes de efeitos de expansão da categoria.

Existem diferentes formas de modelar os efeitos das vendas unitárias, mas a ideia central é focar nas mudanças nas vendas reais dos produtos promovidos e de outros produtos dentro de uma categoria, em vez de estimar probabilidades e elasticidades para as decisões dos consumidores [Pes18].

2.3.5 Metodologias Relacionadas

A Fórmula para o Impacto Total de uma Promoção

A *CrossCap*, empresa que fornece uma plataforma de gestão de *marketing* usada por empresas líderes do setor do retalho alimentar a nível mundial, sugere uma fórmula para avaliar o impacto global de uma promoção, repartindo o incremento do produto promovido pelos efeitos com mais impacto para o retalhista [Cro].

$$\text{ImpactoGlobal} = \text{Bump} + \text{Halo} - \text{Canibalização} - \text{Stockpiling} \quad (2.1)$$

Para além dos efeitos considerados para a fórmula, tem também em consideração o facto da margem de um produto promovido ser, normalmente, inferior à margem do produto sem promoção. Os fatores que considera importantes para a avaliação de uma promoção estão sintetizados no gráfico da Figura 2.5.

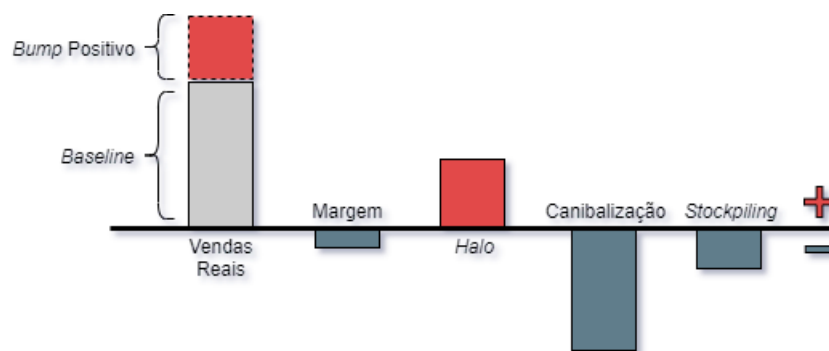


Figura 2.5: "Repartição Geral da Promoção".

Como já foi referido anteriormente, para avaliar estes efeitos é necessário calcular um valor base. A *CrossCap* afirma que todos os retalhistas com quem trabalham concordam que o incremento nas vendas é uma métrica importante para determinar o sucesso promocional no retalho, no entanto há mais de uma forma para os retalhistas calcularem este aumento. Deste modo, são sugeridas três estratégias para calcular o valor base das vendas, indicando os seus prós e contras [Cro17]:

1. Vendas antes da promoção:

- **Prós:** É considerado o método mais simples; o valor base para o período promocional é igual ao valor das vendas no período imediatamente antes da promoção, com a mesma duração do período promocional;
- **Contras:** Não considera a sazonalidade e a tendência;

2. Média de vendas de 52 semanas (eliminando períodos promocionais históricos):

- **Prós:** É uma forma de calcular o valor base para um produto por um longo período, eliminando períodos promocionais e obtendo um valor médio;
- **Contras:** Produtos novos necessitam de produtos semelhantes para auxiliar no cálculo; não capta a sazonalidade nem a tendência associada ao produto;

3. Média sazonal do ano passado (eliminando períodos promocionais históricos):

- **Prós:** Necessita de uma quantidade menor de dados; olha apenas para o período sazonal no ano anterior, resolvendo o problema da sazonalidade;
- **Contras:** Produtos novos necessitam de produtos semelhantes para auxiliar no cálculo; exige uma compreensão da sazonalidade de cada produto, que pode variar de loja para loja, dependendo da zona do país, por exemplo; no caso de haver tendência esta abordagem também pode não ser muito fiável.

Apesar de nenhuma das estratégias ser 100% fiável, a *CrossCap* aconselha a seguir o terceiro método, indicando que funcionará para a maioria dos casos [Cro17].

A Rentabilidade de uma Promoção para um Retalhista

Estudar os efeitos da promoção em termos de vendas unitárias permite medir o impacto no lucro das promoções. Existe apenas um estudo [AHCT06] que aborda este impacto, possivelmente, devido à falta de dados de custo disponíveis publicamente. Ailawadi et al. calcula o aumento bruto em três componentes (canibalização, *stockpiling*, expansão de categoria) e estimou até que ponto a promoção aumenta as vendas de outras categorias de produtos na loja - efeito *Halo* [AHCT06]. Daqui em diante, o foco será o cálculo do valor base, da canibalização e do efeito *Halo*.

Em Ailawadi et al., o cálculo do valor base utiliza uma média móvel do valor das vendas unitárias nas semanas vizinhas não promocionais. O número de semanas utilizado para o cálculo da média varia de acordo com a rotatividade e sazonalidade do produto. Utilizam um maior número de semanas anteriores (*lag*) para produtos de baixa rotatividade e baixa sazonalidade e, por oposição, um número mais pequeno de semanas anteriores (*lag*) e posteriores (*lead*) à semana em promoção para produtos sazonais.

Segundo este estudo, deteta-se um efeito de canibalização se o *bump* dos produtos em promoção numa categoria c , numa loja l , numa semana t , for superior ao aumento no total das vendas da categoria c , na loja l , na semana t .

Assim sendo, o modelo proposto para calcular a percentagem da canibalização, passa por uma regressão que relaciona as vendas totais numa categoria, com o *bump* dos produtos promovidos nessa mesma categoria:

$$VendasTotais_{clt} = \beta_{0c} + \beta_{1cl}BumpTotal_{clt} + \varepsilon_{clt} \quad (2.2)$$

Esta regressão linear é estimada separadamente para cada categoria. A variável dependente é as vendas unitárias na categoria c na loja l na semana t ; a variável independente é a soma dos *bumps* dos itens promovidos na categoria c na loja l na semana t ; β_0 é a ordenada na origem; β_{1cl} é o coeficiente de regressão para a categoria c na loja l ; e ε_{clt} é o termo de erro aleatório.

Considerando que i corresponde a cada produto promovido na categoria c :

$$BumpTotal_{clt} = \sum_{i \in c} Bump_{i clt} \quad (2.3)$$

Um modelo de coeficientes aleatórios é estimado para cada categoria, para englobar as diferenças entre lojas:

$$\beta_{1cl} = \beta_{1c} + \beta_{1cl}^* \quad (2.4)$$

onde,

$$\beta_{1cl}^* \sim N(0, \sigma_c^2)$$

Por fim:

$$\%Canibalização = 1 - (\beta_{1c} + \beta_{1cl}^*) \quad (2.5)$$

Antes de se aplicar o modelo, é removida a sazonalidade dos dados através de um fator de sazonalidade semanal. O fator de sazonalidade para a categoria c na semana t é a média das vendas por categoria na semana t dividida pela venda semanal média por categoria. Tanto as vendas totais da categoria c como o *bump* total, na equação 2.2, são divididos pelo fator de sazonalidade.

Ailawadi et al. define o efeito *Halo* pela primeira vez defendendo que uma promoção numa categoria pode influenciar as vendas em outras categorias diferentes. Por um lado, o efeito é considerado positivo se a promoção influenciar o cliente a comprar produtos de outras categorias dentro da mesma loja, que não teria comprado caso não existisse essa promoção. Por outro lado, o efeito é negativo, se a promoção resultar numa redução das compras em outras categorias.

O modelo que utiliza para calcular a percentagem deste efeito é semelhante ao da canibalização e é detetado se o aumento do *bump* dos produtos em promoção numa loja l , numa semana t , for superior ao aumento no total da loja l , na semana t , depois de ajustar as vendas de acordo com as percentagens de canibalização.

$$VendasTotaisAjustadas_{lt} = \beta_0 + \sum_{d=1}^4 \beta_{dl} BumpTotal_{dlt} + \epsilon_{lt} \quad (2.6)$$

Esta regressão linear é estimada separadamente para cada loja. A variável dependente é as vendas ajustadas na loja l na semana t ; a variável independente é a soma dos *bumps* dos itens promovidos por departamento d na loja l na semana t ; β_0 é a ordenada na origem; β_{dl} é o coeficiente de regressão para o departamento d na loja l ; e ϵ_{lt} é o termo de erro aleatório.

Considerando que i corresponde aos produtos promovidos de cada departamento comercial:

$$BumpTotal_{dlt} = \sum_{i \in d} Bump_{idt} \quad (2.7)$$

Um modelo de coeficientes aleatórios é estimado para cada departamento, para englobar as diferenças entre lojas:

$$\beta_{dl} = \beta_d + \beta_{dl}^* \quad (2.8)$$

onde,

$$\beta_{dl}^* \sim N(0, \sigma_d^2)$$

Por fim:

$$\%Halo = 1 - \beta_{dl} \quad (2.9)$$

Tal como para o modelo da canibalização, é removida a sazonalidade dos dados antes de se aplicar o modelo.

2.4 Mineração de Dados no Retalho Alimentar

Por muito tempo, os gestores confiaram na sua intuição para tomar decisões em relação às estratégias do negócio, entre elas, as estratégias promocionais. Na verdade, estudos sugerem que

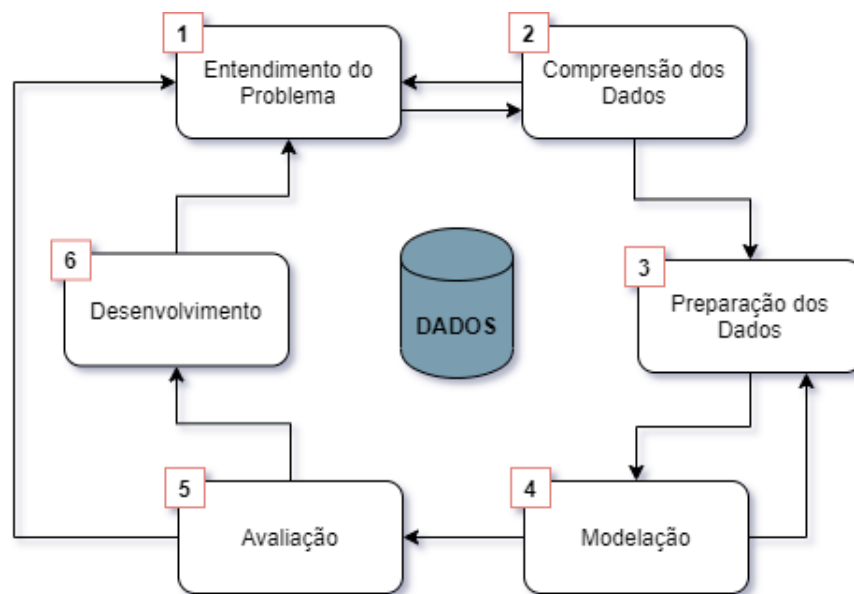
40% das principais decisões nas empresas não se baseavam em factos, mas na intuição dos gestores [DHM10]. Atualmente é evidente que as empresas são cada vez mais movidas por análises de dados [BC12].

Hoje em dia, as empresas retalhistas geram, capturam e armazenam grandes quantidades de dados diariamente. As informações, os padrões e as relações disfarçadas nestes dados provaram ser muito valiosas, pois podem ser usadas para fazer previsões e auxiliar em tomadas de decisão importantes para a definição da estratégia do negócio. De um modo geral, embora seja reconhecida a importância e o valor destes dados, a extração da informação relevante é extremamente desafiante.

A mineração de dados é o processo de descobrir informações úteis em grandes repositórios de dados, identificando padrões e regras de associação que, de outra forma, permaneceriam desconhecidos [TSK13]. Por conseguinte, no setor do retalho alimentar, as técnicas de mineração de dados podem contribuir para a criação de segmentações de clientes, através do reconhecimento do seu perfil, marketing direcionado, através do envio de promoções a segmentos de clientes, entre outros.

2.4.1 Processo de Mineração de Dados

O *CRISP-DM*, *Cross-Industry Standard Process for Data Mining Projects*, estabelece diretrizes claras para auxiliar a extração de informação em problemas de análise de dados. De acordo com o *CRISP-DM*, o processo de mineração de dados é um processo cíclico e consiste em seis etapas distintas de transformação, responsáveis pela conversão de dados brutos em dados relevantes. É um processo padrão que pretende tornar os projetos de mineração de dados menos dispendiosos, mais confiáveis, mais fáceis de gerir, mais rápidos e passíveis de serem repetidos [WH00]. O processo é adaptável e os ciclos podem repetir-se dependendo do resultado de cada etapa. A estrutura do *CRISP-DM* é ilustrada na Figura 2.6 e serve como ferramenta conceptual orientando o processo de análise de dados ao longo desta dissertação. Posteriormente, é descrita cada etapa do processo.

Figura 2.6: Estrutura do *CRISP-DM*.

- **Entendimento do Problema:** A primeira etapa passa por compreender de forma clara o problema, perceber o enquadramento do projeto e assimilar os objetivos e requisitos da perspectiva do negócio.
- **Compreensão dos Dados:** Começa pela extração dos dados, seguida de uma análise para identificar as características relevantes para o projeto. Nesta etapa, é importante começar a detetar os possíveis problemas na qualidade dos dados.
- **Preparação dos Dados:** É nesta etapa que se constrói a estrutura de dados final, pronta para ser modelada. Esta etapa pode envolver tarefas como limpeza, agrupamento ou formatação de dados.
- **Modelação:** Técnicas e métodos de modelação de dados são aplicados na tentativa de encontrar padrões nos dados.
- **Avaliação:** Após a aplicação de vários modelos, os resultados devem ser avaliados de acordo com os objetivos do negócio. No final desta etapa, deve ser tomada uma decisão sobre o uso dos resultados alcançados.
- **Desenvolvimento:** Por fim, o modelo criado é reorganizado de forma a poder ser utilizado facilmente pelo negócio.

No desenvolvimento deste projeto, a etapa da preparação de dados exigiu a aplicação de um método de aprendizagem supervisionada, para completar a informação necessária para o modelo principal. Deste modo, no próximo tópico, é feita uma abordagem a estes métodos orientada ao modelo que se ajusta ao problema em questão.

2.4.2 Métodos de Aprendizagem Supervisionados

Os algoritmos de mineração de dados podem ser classificados como algoritmos supervisionados ou não supervisionados. Algoritmos supervisionados, por sua vez, são classificados em problemas de regressão e classificação. Na aprendizagem supervisionada, o algoritmo tem o objetivo de aprender uma função que mapeia uma entrada para uma saída, com base em pares de entrada-saída. Ao analisar os dados de treino rotulados, é inferida uma função que possa ser utilizada para mapear novos exemplos. Num problema de regressão procuram-se resultados contínuos, fruto de uma função contínua; por outro lado, num problema de classificação os resultados são discretos, isto é, mapeiam-se variáveis de entrada em categorias distintas.

2.4.2.1 Redes Neurais

As redes neurais são um modelo computacional inspirado no funcionamento do cérebro humano e são compostos por um conjunto de neurónios organizados por camadas:

- **Camada de Entrada:** responsável por carregar os dados iniciais no sistema, para serem processados pelas camadas que se seguem.
- **Camada(s) Escondida(s)/Oculta(s):** camada na qual os neurónios recebem um conjunto de entradas ponderadas e retornam uma saída através de uma função de ativação; um modelo pode ser constituído por várias camadas escondidas.
- **Camada de Saída:** responsável por retornar os resultados finais para o programa.

A forma como os neurónios e as camadas estão conectadas definem a arquitetura da rede neuronal. Por um lado, existem as redes neurais *feedforward*, nas quais neurónios presentes na mesma camada não estão conectados entre si; neste caso, a informação percorre a rede numa só direção, desde a camada de entrada para a de saída. Por outro lado, existem as redes neurais recorrentes (*RNN*), nas quais a informação que sai de cada neurónio alimenta neurónios da mesma camada e de outras camadas, podendo, inclusivamente, servir de entrada para o próprio neurónio; a informação percorre a rede em duas direções (entrada $\leftrightarrow$ saída).

No entanto, as redes neurais *feedforward* não são capazes de guardar informação para manter um pensamento persistente. Em alternativa, as *RNN* permitem que as informações persistam, destacando-se por terem uma memória interna que lhes permite usar informação passada para ajudar a aumentar o desempenho do modelo. Após produzir um resultado, ele é copiado e enviado de volta para a rede recorrente, servindo como entrada para a etapa seguinte.

As *RNN* padrão têm uma limitação na quantidade de informação que são capazes de armazenar. Todavia, existe uma versão modificada das redes neurais recorrentes, que por natureza relembra a informação por grandes períodos de tempo - a rede neuronal *Long Short-Term Memory*. É um modelo reconhecido por alcançar resultados impressionantes e ao qual se deve grande parte do sucesso associado às *RNN* [GBC16].

Tal como as *RNN* padrão, as *LSTM* são representadas por uma cadeia interligada de módulos de rede neuronal repetidos, mas cada módulo, em vez de ter uma única camada, tem quatro que interagem de uma maneira especial, representadas a amarelo na Figura 2.7.

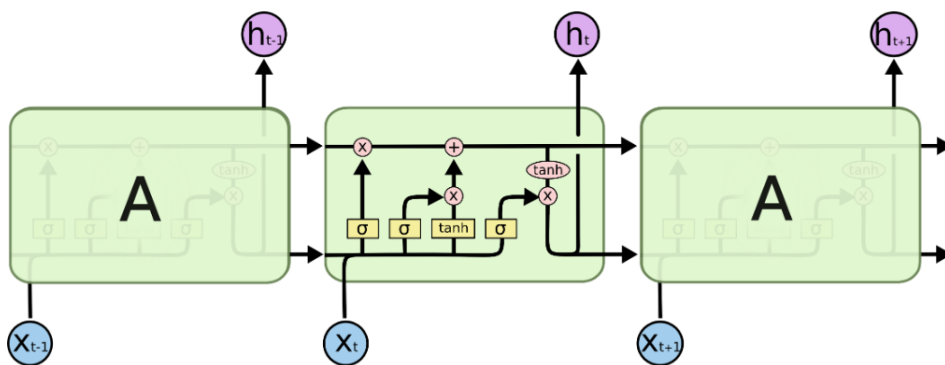


Figura 2.7: O módulo de repetição numa *LSTM* contém quatro camadas de interação.

O elemento chave para a *LSTM* é o estado da célula, que consegue manter a informação na memória durante longos períodos de tempo; é a linha horizontal que percorre a parte superior do diagrama na Figura 2.7. Este modelo tem a capacidade de remover ou adicionar informações ao estado da célula, através de três estruturas chamadas portões que o protegem e controlam: portão de esquecimento, portão de entrada e portão de saída.

2.4.2.2 Séries temporais e previsões com o modelo *LSTM*

As redes neurais recorrentes são muito utilizadas para problemas com dados sequenciais não estruturados, como processamento de linguagem natural e reconhecimento de fala, mas as suas características são interessantes para qualquer problema com carácter de sequência.

As *LSTM* são projetadas explicitamente para evitar o problema de dependência a longo prazo. Lembrar informações por longos períodos de tempo é o comportamento padrão destes modelos e não algo que eles procuram aprender. As *LSTM* destacam-se para problemas de classificação, processamento e previsão de séries temporais [GBC16], para além de ser capaz de modelar problemas com múltiplas variáveis de entrada.

Por último, este modelo também apresenta mecanismos para capturar características de ciclos e sazonalidades das séries.

Capítulo 3

Metodologia e Estruturas de Dados

Medir a eficácia das promoções é considerada uma tarefa desafiante, dada a natureza ruidosa dos dados captados diariamente no setor do retalho alimentar.

Neste capítulo é descrita a solução implementada e as etapas necessárias para resolver o problema em questão. Tendo em conta o âmbito e os objetivos do projeto, a abordagem do problema foi estruturada de acordo com a arquitetura processual *CRISP-DM*. Adicionalmente, foi acrescentada uma secção no início do capítulo que expõe as diretrizes que conduziram à metodologia adotada.

3.1 Visão Geral

São inúmeros os efeitos que resultam de uma promoção, no entanto, de acordo com estudos teóricos, só a canibalização e o efeito *Halo* evidenciam a existência de relações entre produtos. Com o intuito de compreender e concretizar estas relações, foi levado a cabo um estudo que visa a avaliação do impacto das ações promocionais nas vendas do produto promovido e nos produtos que o rodeiam, com base no cálculo destes dois efeitos.

Partindo do estudo de Ailawadi et al. [AHCT06], onde são desenvolvidos dois modelos semelhantes para o cálculo da canibalização dentro de uma categoria e do efeito *Halo* no total das vendas de uma loja, construiu-se um modelo para calcular estes efeitos ao nível do produto, numa relação direta num par de produtos.

Durante o estudo dos métodos de avaliação das promoções, que foram revistos no Capítulo 2, foi detetada a necessidade de calcular um valor base para as vendas, como premissa obrigatória para o cálculo dos efeitos das promoções. Na verdade, existem muitas estratégias para calcular o valor base das vendas, porém grande parte não é capaz de captar fatores como a sazonalidade e a tendência (estratégias 1 e 2 da *CrossCap*); ou exige uma compreensão da sazonalidade de cada produto (estratégia 3 da *CrossCap*); ou necessita de semanas vizinhas não promocionais (estratégia de Ailawadi). Tendo em consideração o ambiente extremamente promocional que se vive no setor

do retalho alimentar em Portugal e considerando essencial atentar à sazonalidade e à tendência associadas aos produtos, torna-se complicado aplicar qualquer uma das estratégias mencionadas anteriormente. Deste modo, houve a necessidade de idealizar um modelo novo para o cálculo do valor base.

Em síntese, ambos os modelos foram estruturados de acordo com a arquitetura processual *CRISP-DM* e o modelo do cálculo do valor base surge na etapa da preparação dos dados para o modelo principal das relações produto a produto. Na Figura 3.1 está esquematizado o processo adaptado ao contexto do problema.

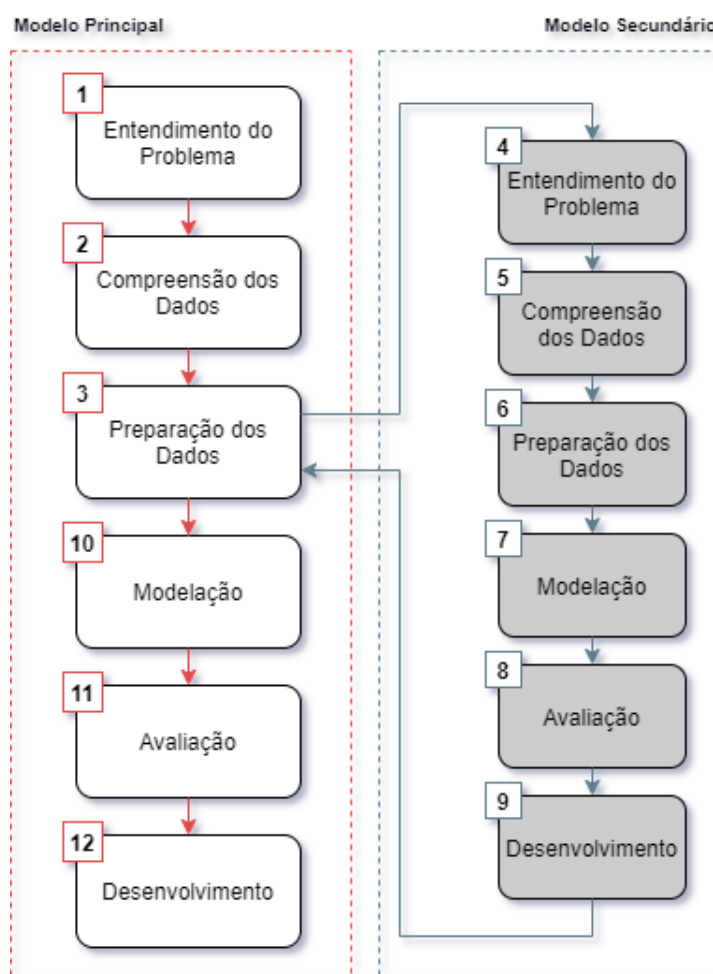


Figura 3.1: Processo *CRISP-DM* adaptado ao contexto do problema.

Seguidamente, percorrendo o esquema da Figura 3.1, é explicada cada etapa do processo, identificando os pressupostos que fundamentam os modelos implementados.

3.2 Modelo Principal

O modelo principal é o modelo que pretende dar resposta aos objetivos do negócio, nomeadamente perceber de que forma se relacionam os produtos.

3.2.1 Entendimento do Problema

O retalho alimentar tem sofrido uma elevada pressão sobre a atividade promocional. Assim sendo, este projeto surge com os seguintes objetivos:

- Apurar e calcular os efeitos da atividade promocional entre produtos dentro da mesma categoria ou de diferentes categorias;
- Identificar quais os produtos que sendo alvo de atividade promocional ou destaque originam a canibalização ou o efeito *Halo* em outros produtos;

Deste modo, é importante começar por compreender as dimensões do negócio, a forma como os produtos são geridos e estão organizados e a estratégia promocional que lhes é aplicada.

3.2.1.1 Enquadramento no Negócio

O Continente opera em três tipos de loja física e uma *online*. As lojas físicas podem ser de grande, média ou pequena dimensão e já são 302 espalhadas por todo o país. Uma loja Continente de grande dimensão tem entre 45 000 e 50 000 referências de produtos disponíveis para venda, em simultâneo.

3.2.1.2 Estruturas de Produtos

Para ser capaz de gerir esta enorme quantidade de produtos, o Continente organiza as suas equipas de acordo com uma estrutura comercial, pela qual, também, organiza os produtos em loja. No entanto, essa estrutura não está totalmente alinhada com a forma como o cliente percebe os produtos. Assim sendo, de um ponto de vista analítico, surgiu a necessidade de criar uma nova estrutura orientada ao cliente, chamada Hierarquia Virtual do Produto (*VPH*), que também veio a facilitar a definição e comunicação das promoções.

As tabelas 3.1 e 3.2 especificam os níveis em que cada uma das estruturas se subdivide, através de um exemplo.

Níveis	Exemplo
<i>SKU</i>	Sumo Laranja CNT 750ml
Unidade Base	Sumos Fruta
Sub-categoria	Transformados
Categoria	Especialidades
Unidade de Negócio	Frutas e Legumes
Departamento Comercial	Frutas e Legumes, Charcutaria e Queijos, Padaria e <i>Take Away</i>

Tabela 3.1: Estrutura comercial.

Níveis	Exemplo
1 - Categoria	Sumos
2 - Categoria-Marca	Sumos da marca Continente
3 - Produto	Sumos Laranja Natural
4 - Produto-Marca	Sumos Laranja Natural da marca Continente

Tabela 3.2: Estrutura orientada ao cliente - *Virtual Product Hierarchy* (VPH).

Na estrutura VPH, mesmo o nível mais granular, produto-marca (nível 4), é constituído por várias referências. Por exemplo, o VPH nível 4 “Sumos Laranja Natural da marca Continente” é composto por sumos de diferentes volumes, porém na visão do cliente será sempre um sumo e não uma especialidade dentro da categoria frutas e legumes, como a estrutura comercial o sugere.

3.2.1.3 Estratégia Promocional

A estratégia promocional do Continente está assente em dois tipos de promoção: os descontos diretos (ou cortes de preço) e os descontos em cartão.

Do ponto de vista da análise das promoções, a grande diferença entre os dois tipos de promoção é o público atingido por cada uma. Enquanto o desconto direto é uma promoção global, definida para todas as lojas com o produto, o desconto em cartão é uma promoção direcionada, enviada para um grupo de clientes específico - dentro do grupo de clientes que possuem o cartão e estão fidelizados.

Após compreender as necessidades do negócio e a forma como estão organizados os produtos e as promoções, partiu-se para a extração e compreensão dos dados disponibilizados para a análise.

3.2.2 Compreensão dos Dados

O Continente tem toda a informação organizada em centenas de tabelas, mas há uma que cresce a cada dia e que é a base de qualquer análise - a tabela transacional - pois guarda toda a informação das transações efetuadas nas lojas, associando uma chave de cliente, quando a transação é feita por um cliente fidelizado. Nesta tabela, cada linha representa um produto transacionado e as variáveis mais relevantes para a análise são descritas de seguida:

- **Time Key:** Data da transação (ano/mês/dia);
- **ID da Transação:** Identificador da transação;
- **ID do Cliente:** Identificador do cliente (-1, caso não seja uma transação fidelizada);

- **SKU:** *Stock Keeping Unit*, referência do produto;
- **Quantidade:** Número de unidades ou peso (kg) do produto adquirido;
- **Vendas Brutas:** Vendas brutas do *SKU* transacionado (€);
- **Desconto em Cartão:** Desconto associado a um cupão que foi rebatido no produto (€);
- **Código da Loja:** Código único associado à loja em que foi efetuada a transação;
- **Código da Insígnia:** Código único associado ao tipo de loja (pequena, média ou grande).

Para além desta tabela, foi necessário recorrer às tabelas das estruturas dos produtos: comercial e *VPH*, para associar a cada *SKU* os vários níveis de cada estrutura que lhe correspondem; deste modo, torna-se possível filtrar ou agrupar os dados através destas variáveis, para orientar o estudo das relações para os objetivos pretendidos. Durante o projeto, foi requisito do negócio limitar a investigação a quatro departamentos comerciais e reduzir o número de *VPHs*, qualquer que fosse o nível que estivesse a ser avaliado, segundo três métricas que são explicadas na etapa da preparação dos dados.

Em relação aos descontos, que são o foco principal da análise, detetou-se à partida que o desconto direto não constava na tabela das transações e, embora o valor do desconto não fizesse parte do alcance do projeto, era importante saber se o produto estava ou não em promoção quando foi transacionado. Deste modo, recorreu-se a uma tabela com todas as promoções de desconto direto, que continha os *SKUs* em promoção por dia e por loja.

Por fim, foi nesta etapa que se definiu a granularidade da informação. Tendo em consideração que as relações entre produtos são as relações percecionadas pelos clientes, fruto das suas ações, decidiu-se que a análise seria feita agrupando os produtos ao nível 4 da estrutura *VPH*, produto-marca. De agora em diante, *VPH* nível 4 será referido somente como produto.

Após conhecer todas as tabelas necessárias e as suas dimensões, partiu-se para a preparação dos dados.

3.2.3 Preparação dos Dados - Modelo Secundário

Esta etapa começou pela criação de uma tabela base que permitisse extrair diretamente a informação necessária para o modelo principal.

Assim, relacionaram-se as tabelas das estruturas dos produtos com a tabela transacional. Para reduzir a complexidade computacional e para cumprir com os requisitos que a empresa definiu à partida, filtraram-se os departamentos comerciais e o período das transações que iria ser analisado, por fim, fez-se a redução do universo de produtos:

- **Intervalo de Tempo**

Decidiu-se que a análise das relações seria feita com base no ano 2019, mas para o cálculo do valor base foi necessário recorrer aos dados transacionais de 2016, 2017 e 2018.

- **Departamentos Comerciais**

Mantiveram-se na análise os quatro departamentos comerciais com bens de grande consumo (alimentar - DC 10, peixaria e talho - DC 11, frutas e legumes, charcutaria e queijos, padaria e *take away* - DC 12 e, por fim, nutrição saudável - DC 41).

- **Redução do Universo de VPHs**

Considerando apenas o ano 2019, com o propósito de conservar, para a análise, produtos com alguma representatividade, foram removidos produtos sazonais, regionais e de nicho, seguindo critérios definidos pela própria empresa. Neste caso, a sazonalidade foi definida por critérios que avaliam a distribuição do número de transações com referências do produto por mês; a regionalidade diretamente relacionada com o número de lojas com referências do produto à venda; e, por fim, ser ou não um produto de nicho estava relacionado com o número de clientes distintos que compraram pelo menos uma referência do produto em questão. De um total de 6073 produtos transacionados em 2019, restaram 2553, que representam 12568 *SKUs*.

Seguidamente, removeram-se os produtos transacionados com rebate de cupão de desconto em cartão e identificaram-se as promoções de desconto direto recorrendo à tabela das promoções.

O processo descrito anteriormente não permitia reunir toda a informação necessária para o modelo. Em traços gerais, o modelo necessita de receber, como variável de entrada, as vendas totais do produto agrupadas ao dia ou à semana e o incremento do produto em promoção, em relação ao valor base, também agrupado com base no mesmo período temporal. Tendo em consideração que as dinâmicas ao longo de uma semana podem diferir muito, foi decidido agrupar as vendas ao dia.

Aquando do agrupamento dos dados ao dia e ao produto (*VPH* nível 4), verificou-se que, mesmo considerando as vendas em uma só loja, nem todas as referências do produto estavam em promoção em simultâneo. Deste modo, decidiu-se calcular a percentagem das vendas em promoção em relação às vendas totais por dia e por produto (considerando todas as referências pertencentes a determinado nível 4 produto-marca de acordo com a estrutura *VPH*). Na Tabela 3.3 é ilustrada a estrutura da tabela base, agrupada ao produto, ao dia e à loja.

Produto	Código da Loja	Time Key	% Promoção	Vendas Totais	Valor Base
Banana CNT	9	7/04/2018	70%	1728,33€	?

Tabela 3.3: Estrutura da tabela base agrupada ao produto, ao dia e à loja.

Na Tabela 3.3 a coluna do valor base ainda está incompleta. Como referido anteriormente, recorreu-se a um modelo de aprendizagem supervisionada para preencher esta informação em falta, para cada produto. Seguindo as etapas do *CRISP-DM*, seguidamente, será explicado o processo de implementação do modelo secundário.

3.2.3.1 Modelo Secundário

O modelo secundário é um modelo que surge para dar resposta a uma necessidade do modelo principal - calcular as vendas base diárias para cada produto, nas semanas em que esteve em promoção.

1. Entendimento do Problema

Para avaliar os efeitos das promoções é indispensável estimar um valor base para as vendas diárias, nas semanas em que o produto esteve em promoção. O objetivo deste modelo é prever quais teriam sido as vendas totais de determinado produto, por dia, nas semanas em que esteve em promoção, durante o ano 2019, caso não estivesse em promoção.

É importante lembrar que fatores como a sazonalidade e a tendência devem ser tidos em consideração.

Para além disso, existem inúmeros fatores externos que podem influenciar as vendas do produto quer este esteja em promoção ou não, tal como uma promoção em outro produto. Teoricamente, um modelo completo seria alimentado com a informação relativa a todos os produtos dentro do mesmo ecossistema, no entanto, na prática isso não seria exequível.

2. Compreensão dos Dados

Os dados disponíveis para este modelo são os dados previamente trabalhados para o modelo principal. Deste modo, começa-se por extrair da tabela base a informação relativa ao produto para o qual se pretende calcular o valor base das vendas.

Os dados representam uma série temporal das vendas totais do produto por dia e das percentagens das vendas em promoção, desde o dia 1 de janeiro de 2016 ao dia 31 de dezembro de 2019 (um total de 1461 dias), pois os dados traduzem uma sequência.

Nesta etapa, a percentagem das vendas em promoção manifestou-se como sendo um dado muito relevante que ajudará a diferenciar as vendas reais das vendas base. As vendas base que se procuram calcular são as vendas quando a percentagem das vendas em promoção for 0.

Com base nos dados disponíveis, o modelo foi idealizado nesta etapa. Tendo acesso às percentagens das vendas em promoção, começa-se por criar um modelo de previsão de vendas sequenciais com base nas percentagens de vendas em promoção reais.

Posteriormente, a variável da percentagem irá possibilitar ao modelo que preveja o valor base das vendas na fase de desenvolvimento, adaptando os dados de previsão.

Tendo em consideração, os requisitos necessários sobre a captação da sazonalidade e da tendência e a capacidade de suportar múltiplas variáveis de entrada, o modelo selecionado foi o *LSTM*, que é uma rede neuronal reconhecida por alcançar bons resultados na previsão de séries temporais.

Por fim, sabe-se, à partida, que as lojas estão sempre fechadas no dia 25 de dezembro e desde 2017 também fecham no dia 1 de janeiro. Assim sendo, não há vendas registadas nestes dias.

3. Preparação dos Dados

Nesta etapa transforma-se o conjunto de dados para o formato que o modelo está preparado para receber.

Começa-se por garantir que os dados estão organizados de forma sequencial, para se remover todas as colunas da tabela, com a exceção das colunas da percentagem das vendas em promoção e a das vendas totais. A Figura 3.2 ilustra a estrutura dos dados. Cada linha representa um dia e tem duas variáveis.

	Percentagem (Var1)	Vendas Totais (Var2)
Data		
2016-01-01	0.00	7.62
2016-01-02	0.00	497.94
2016-01-03	0.00	492.01
2016-01-04	0.00	292.15
2016-01-05	0.49	334.20
...	...	...
2019-12-27	0.53	1051.04
2019-12-28	0.57	1547.77
2019-12-29	0.56	1508.15
2019-12-30	0.39	2265.55
2019-12-31	0.08	3194.23

Figura 3.2: Tabela de dados inicial com duas variáveis por dia.

De seguida, os dados são estruturados para suportar um problema de aprendizagem supervisionada, através da escolha do número de dias que o modelo vai considerar para prever as vendas do dia t . Posteriormente, são normalizados entre -1 e 1, pois, quando uma rede é ajustada a dados não dimensionados, é possível que valores grandes de entrada diminuam a convergência da rede e, em alguns casos, impeçam a rede de aprender efetivamente os dados do problema. A tabela final é uma tabela com o número de colunas igual ao número de variáveis multiplicado pelo número de dias mais um (número de variáveis $\times$ (número de dias + 1) = 16 variáveis), tal como ilustrado no exemplo da Figura 3.3, que considera um intervalo de 7 dias.

	Var1(t-7)	Var2(t-7)	Var1(t-6)	Var2(t-6)	Var1(t-5)	Var2(t-5)	Var1(t-4)	Var2(t-4)	Var1(t-3)	Var2(t-3)	Var1(t-2)	Var2(t-2)	Var1(t-1)	Var2(t-1)	Var1(t)	Var2(t)
7	-1.00	-0.996	-1.00	-0.736	-1.00	-0.739	-1.00	-0.845	-0.02	-0.823	0.20	-0.846	0.06	-0.824	0.24	-0.804
8	-1.00	-0.736	-1.00	-0.739	-1.00	-0.845	-0.02	-0.823	0.20	-0.846	0.06	-0.824	0.24	-0.804	0.18	-0.729
9	-1.00	-0.739	-1.00	-0.845	-0.02	-0.823	0.20	-0.846	0.06	-0.824	0.24	-0.804	0.18	-0.729	0.30	-0.805
10	-1.00	-0.845	-0.02	-0.823	0.20	-0.846	0.06	-0.824	0.24	-0.804	0.18	-0.729	0.30	-0.805	0.28	-0.804
11	-0.02	-0.823	0.20	-0.846	0.06	-0.824	0.24	-0.804	0.18	-0.729	0.30	-0.805	0.28	-0.804	-0.90	-0.853
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figura 3.3: Tabela de dados final, considerando um intervalo de 7 dias.

Nesta fase foram testados diferentes intervalos de diferentes dimensões, na procura do melhor resultado, pois este valor pode ajudar o modelo a compreender melhor a sazonalidade do produto. Estes resultados estarão espelhados num exemplo no Capítulo 4.

4. Modelação

Nesta etapa, o modelo *LSTM* é ajustado aos dados de entrada com múltiplas variáveis.

Em primeiro lugar, a tabela de dados deve ser separada numa tabela com os dados de treino, outra com os dados de validação e outra com os dados de teste. Neste caso, foram utilizados os dados de 2016, 2017 e 2018 para treinar o modelo e os dados de 2019 foram separados de forma aleatória em dois conjuntos para validar e testar o modelo, com 182 e 183 dias, respetivamente.

Adicionalmente para cada linha, tanto da tabela de treino como das tabelas de validação e teste, devem ser indicadas quais são as variáveis de entrada e quais são as variáveis de saída, a partir das quais o modelo irá mapear uma função que fará a previsão das vendas. Assim sendo, nesta situação decidiu-se que se iriam prever as vendas para um dia de cada vez (variável de saída: $Var2(t)$) e que seriam passadas ao modelo as vendas e as percentagens das vendas em promoção dos últimos dias, de acordo com o intervalo definido na etapa anterior, para além da percentagem das vendas em promoção do dia para o qual se estão a prever as vendas (variáveis de entrada: $Var1(t-x) + Var2(t-x) + \dots + Var1(t-1) + Var2(t-1) + Var1(t)$).

Para um modelo *LSTM* as variáveis de entrada têm de ser tridimensionais, indicando o número de *samples*, ou observações, que são as linhas da tabela com a sequência dos últimos dias que estão a ser considerados, o número de *timesteps*, que neste caso será 1, e, por fim, o número de variáveis por *timestep*, que serão todas as variáveis de entrada.

No que respeita à *LSTM*, a camada de entrada e de saída são definidas à partida, com o número de neurónios igual ao número de variáveis de entrada e de saída, respetivamente. Enquanto a camada de entrada é uma camada *LSTM* que recebe os dados num formato tridimensional, a camada de saída é uma *Dense Layer* na qual cada neurónio, neste caso um, recebe informações de todos os neurónios da camada anterior.

De seguida, é necessário definir o número de camadas escondidas e o número de neurónios que se coloca em cada uma. Na verdade, não há nenhum método analítico confiável para calcular o número de neurónios ou o número de camadas necessárias para uma rede neuronal adaptada a um problema de modelação preditiva. No entanto, é regra geral uma única camada escondida ser suficiente para a maioria dos casos simples e duas serem suficientes para problemas mais complexos [GBC16]. De qualquer forma, a sugestão é sempre usar a experimentação para descobrir qual configuração funciona melhor para cada problema e, após algumas experimentações, não tendo verificado melhorias significativas, optou-se por considerar uma única camada escondida. Em relação ao número de neurónios, foram experimentados vários valores e os resultados vêm refletidos no Capítulo 4.

Para além das três camadas mencionadas até então, foi utilizada uma camada de *Dropout* como técnica de regularização. Com esta camada, os dados que vão circulando no modelo são probabilisticamente excluídos das atualizações de ativação e peso durante o treino da rede, sendo necessário indicar a percentagem de informação que deve ser descartada. Mais uma vez, a sugestão é recorrer à experimentação para definir este valor. No entanto, optou-se por manter a

percentagem a 20%, pois as curvas de aprendizagem nas experimentações apresentavam um bom comportamento. De um modo geral, esta camada ajuda a reduzir o *overfitting* e a melhorar o desempenho do modelo.

```
1 model = Sequential()  
2 model.add(LSTM(neuronios, input_shape=(timesteps, variaveis)))  
3 model.add(Dropout(0.2))  
4 model.add(Dense(units=1))
```

Depois de definir a arquitetura da rede, é necessário compilá-lo, passando ao modelo uma função de perda e um otimizador.

```
1 model.compile(loss='mean_squared_error', optimizer='adam')
```

Por fim, ajustam-se os dados de treino ao modelo, definindo mais dois parâmetros: *epochs* e *batch size*. Enquanto o número de *epochs* é o número de vezes que o conjunto de dados de treino é processado na totalidade, o *batch size* define o número de *samples* a serem processadas antes dos parâmetros internos do modelo serem atualizados.

Os dados de validação também são passados ao modelo na fase de treino, para se poder acompanhar a perda durante o treino e se avaliar as curvas de aprendizagem tanto para os dados de treino como para os dados de validação.

```
1 model.fit(train_X, train_y, epochs=epochs, batch_size=batch_size,  
           validation_data=(test_X, test_y))
```

As definições das características do modelo foram orientadas por um processo de experimentação e os resultados mais relevantes estão espelhados num exemplo no Capítulo 4.

5. Avaliação

Depois de ajustar o modelo aos dados de validação, faz-se as previsões para o conjunto de dados de teste e, após inverter a normalização dos dados, calcula-se o erro do modelo.

O erro calculado foi o erro percentual absoluto médio (*MAPE*). No entanto, o erro pode ser elevado pela granularidade dos dados da previsão e pelo facto de o modelo não cobrir a dinâmica promocional dos artigos relacionados.

Assim sendo, a avaliação do modelo baseia-se numa análise cuidada da curva de aprendizagem, garantindo que não ocorreu *underfitting* nem *overfitting* e que os conjuntos de dados de treino e validação são representativos; e numa avaliação da qualidade final do modelo perante dados nunca antes vistos, garantindo a sua capacidade de generalização.

6. Desenvolvimento

Após alcançar um modelo confiável, é altura de fazer a previsão para completar a coluna em falta da tabela do modelo principal.

Convém destacar que parece contraditório fazer uma previsão passando a percentagem das vendas em promoção do próprio dia, no entanto, esta é uma previsão que visa calcular as vendas base num intervalo de tempo que já passou. Para tal, como referido na etapa da compreensão dos dados, parte-se do pressuposto que um valor base nestas condições, é uma previsão das vendas para um dia com a percentagem das vendas em promoção igual a 0.

Posto isto, recupera-se o conjunto de dados de validação e de teste, substitui-se a última variável de entrada por -1 (considerando que os dados estavam previamente normalizados) e faz-se a previsão das vendas base.

Terminando o modelo secundário, completa-se a tabela final. Posteriormente, é calculado o *bump* e removida a sazonalidade dos dados, tal como é sugerido em Ailawadi et al., 2006 [AHCT06].

Para calcular o *bump*, como já foi referido anteriormente, foi subtraído o valor base às vendas reais, para cada produto p em cada dia t , como indica a Equação 3.1.

$$Bump_{pt} = VendasTotais_{pt} - ValorBase_{pt} \quad (3.1)$$

Relativamente à sazonalidade, é do conhecimento da empresa que de uma forma geral as vendas têm uma sazonalidade semanal, sendo o pico de vendas à sexta-feira e ao sábado; e mensal, pois verifica-se um aumento das vendas na última semana de cada mês. Ailawadi só remove a sazonalidade das semanas, pois é a grandeza base do seu modelo, enquanto que neste estudo houve necessidade de uma granularidade diária. De seguida, são enumerados os passos para remover ambas as sazonalidades:

1. Calcula-se a média dos valores base em cada semana (média semanal);
2. Calcula-se a média das médias semanais (média anual);
3. Calcula-se um fator semanal, para cada semana, dividindo a média semanal pela média anual;
4. Calcula-se um fator diário, para cada dia da semana, dividindo o valor base pela média semanal;
5. Por fim, divide-se as vendas totais e os *bumps* tanto pelo fator semanal como pelo fator diário que lhe correspondem.

Após todo este processo de transformação e limpeza dos dados, pode-se passar para a etapa da modelação, onde vão ser explicados os pressupostos que fundamentam a adaptação do modelo de Ailawadi ao contexto deste problema.

3.2.4 Modelação

Como já foi referido anteriormente, este modelo surge da adaptação do modelo de Ailawadi do cálculo da canibalização dentro de uma categoria. Recordando, o modelo relaciona semanalmente as vendas totais da categoria c (variável dependente) com os *bumps* dos produtos que estavam em promoção (variável independente). No entanto, considera também um modelo de efeitos aleatórios, para captar as diferenças entre as lojas. Como as diferenças entre lojas estão fora do âmbito do projeto, a regressão de efeitos aleatórios é substituída por uma regressão linear, tal como fez Pesonen [Pes18].

Outra diferença em relação ao modelo original, que já foi referida anteriormente, é a grandeza dos dados passados ao modelo. Enquanto que no modelo original cada ponto representa uma semana, neste modelo representa um dia. A grandeza temporal é representada por t .

O primeiro pressuposto, para utilizar o modelo de forma análoga, considera que cada par de produtos (A, B), para o qual se pretende calcular a canibalização, é uma categoria fictícia, mesmo que os produtos não pertençam à mesma categoria comercial. Assim sendo, a variável dependente do modelo, passa a ser as vendas totais dos dois produtos por dia.

De seguida, questionou-se o significado da canibalização no global da categoria, pois esta depende da troca de um produto por outro e estará relacionada com a fidelização dos clientes a ambos os produtos. Uma promoção em A pode conduzir os clientes fiéis a B a experimentar A, mas o contrário pode não acontecer. Com base no pressuposto de que a relação de A para B, não é igual à relação de B para A, definiu-se que o modelo calcularia sempre a canibalização de um produto em promoção sobre outro produto sem promoção, implicando que a variável independente seja o *bump* do produto em promoção por dia. Desta forma, a relação passa a ser composta por duas forças unidireccionais de A em promoção para B e de B em promoção para A.

Sintetizando, o modelo para calcular a relação do par (A em promoção, B sem promoção) está explicado na Equação 3.2.

$$VendasTotais_{At+Bt} = \beta_0 + \beta_1 BumpTotal_{At} + \varepsilon_t \quad (3.2)$$

Onde, β_0 é o termo de interseção no eixo y, β_1 é o coeficiente da regressão e ε_t é o erro aleatório.

Através da interpretação do modelo original, é retirado o valor da canibalização, através da subtração do coeficiente da regressão a um. No entanto, nesta fase, surge o último pressuposto, no qual o efeito *Halo* é considerado o efeito complementar da canibalização, quando a análise é feita produto a produto.

Por fim, considerando que a relação do par (A em promoção, B sem promoção) é positiva na presença de um efeito *Halo* e negativa na presença de um efeito de canibalização, a Equação 3.3 calcula a força da relação.

$$Força da Relação = \beta_1 - 1 \quad (3.3)$$

3.2.5 Avaliação

A avaliação do modelo está relacionada com a proporção da variação da variável dependente que é explicada pela variável independente, isto é, a forma como a reta consegue aproximar os dados. Para tal, recorreu-se a uma medida estatística que representa esta proporção - o R quadrado (R^2).

Para além disso, a interpretação dos resultados requer sempre o olhar crítico dos especialistas do negócio, pois como se pode calcular a relação entre qualquer par de produtos, por coincidência pode haver uma relação forte entre produtos que não faz sentido relacionar.

3.2.6 Desenvolvimento

A fase do desenvolvimento depende das relações que se pretende averiguar. De qualquer forma, para calcular uma relação para um par de produtos, começa-se por calcular o valor base das vendas de ambos os produtos e ajusta-se a regressão linear aos dados para os dias em a percentagem das vendas em promoção de A é 100% e a de B é 0% e vice-versa, obtendo as duas forças unidirecionais.

Capítulo 4

Resultados

Este capítulo concretiza os modelos descritos no capítulo anterior, através de alguns exemplos, retirados de um caso de estudo endereçado a uma loja específica. Partindo diretamente do modelo secundário, no qual é estimado o valor base das vendas de um produto por dia, é demonstrado o processo de afinação dos parâmetros do modelo para um produto e são apresentados os resultados para dois produtos distintos. De seguida, partindo dos dados de ambos os produtos e os respetivos valores base são apresentadas as forças da relação ($A \rightarrow B$ e $B \rightarrow A$), com base em modelos de regressão.

4.1 Caso de Estudo

Partindo da Tabela 3.3 do Capítulo 3, filtraram-se todos os dados relativos a uma loja Continente de grande dimensão. De um total de 2498 produtos, selecionaram-se 430 com vendas em praticamente todos os dias ao longo dos quatro anos e calculou-se o valor base para todos. No entanto, os parâmetros do modelo para calcular o valor base necessitaram de ser afinados, pois não se adaptavam de igual forma a todos os produtos, por diversas razões, entre elas as diferentes sazonalidades.

Assim sendo, de seguida é demonstrada a afinação dos parâmetros do modelo *LSTM* para um dos produtos (batatas fritas onduladas marca Continente). Adicionalmente, são apresentados os resultados do mesmo modelo para um segundo produto (batatas fritas onduladas marca *Ruffles*), após ter sido feita uma afinação de parâmetros semelhante. Por fim, será extraída a relação entre estes dois produtos.

4.1.1 Afinação de Parâmetros do Modelo *LSTM*

A etapa da afinação de parâmetros é uma etapa fulcral para o sucesso do modelo. Tendo em consideração que a previsão do valor base não é passível de ser avaliada, pelos motivos mencionados no capítulo anterior, uma afinação cuidada da configuração do modelo é a chave para aumentar

a credibilidade dos resultados.

Batatas fritas onduladas marca Continente (produto A)

Começando pela visualização dos dados das vendas do produto A, na Figura 4.1 pode-se detectar à partida um pico nas vendas no último dia de cada ano, com destaque para o pico do final do ano 2017.

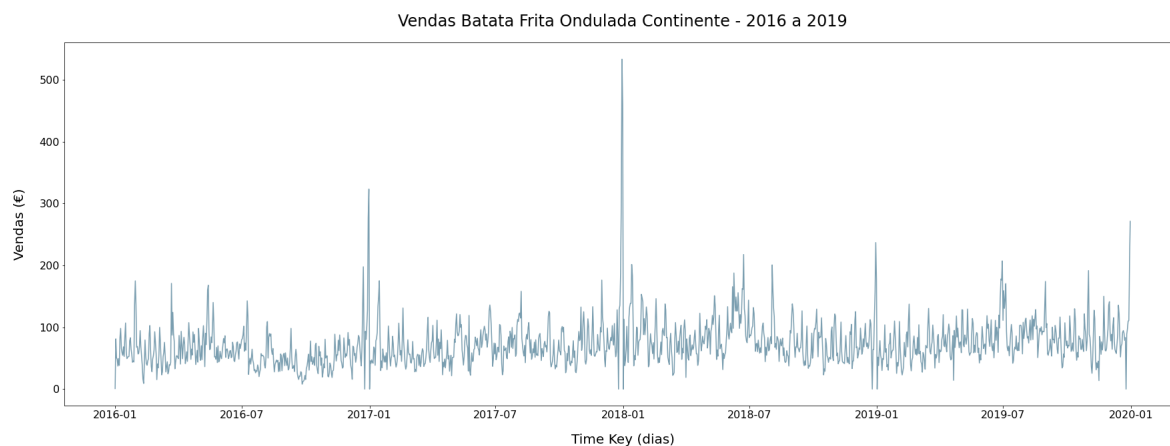


Figura 4.1: Vendas da Batata Frita Ondulada Marca Continente (2016 a 2019).

No entanto, apesar de não haver uma evidência clara de tendência, se for feita uma ampliação aos dados relativos a um mês, confirma-se a existência de sazonalidade semanal, havendo um crescimento das vendas à sexta-feira e ao sábado (ver Figura 4.2).

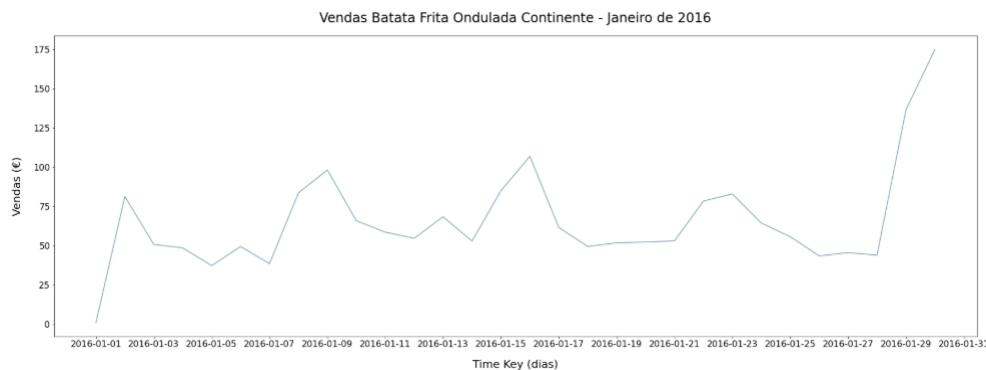


Figura 4.2: Vendas da Batata Frita Ondulada Marca Continente em janeiro de 2016.

Aquando da estruturação dos dados, foi necessário definir o número de dias que constitui cada observação (*lag*). Adicionalmente, na construção do modelo *LSTM*, foram três os parâmetros que necessitaram de ser afinados: o número de neurónios da camada escondida, o número de ciclos (*epochs*) e o tamanho do lote (*batch size*).

Para cada um destes quatro parâmetros, foram escolhidos 4 valores para a afinação do modelo. O único parâmetro que tem algum racional teórico, relacionado com o contexto do problema, por detrás dos valores que lhe foram passados é o tamanho da *lag*, pois este parâmetro ajudará o

modelo a entender a sazonalidade do produto. Assim sendo, os valores que lhe foram passados equivalem a contemplar semanas completas. Por outro lado, os valores considerados para os outros parâmetros são valores comuns em estudos de redes neurais *LSTM*. Na Tabela 4.1 estão indicados os valores que foram considerados para cada parâmetro.

<i>LAG</i>	<i>EPOCHS</i>	<i>BATCH SIZE</i>	NEURÓNIOS
7	100	128	20
14	400	256	40
21	750	512	70
28	1000	1024	100

Tabela 4.1: Valores considerados para cada parâmetro.

A afinação do modelo implica a procura da melhor combinação de valores para os parâmetros, isto é, a combinação que ajuda a minimizar o erro da previsão dos dados de validação. Tendo em consideração que as condições iniciais aleatórias para uma rede *LSTM* podem resultar em valores de saída muito diferentes cada vez que determinada configuração é treinada, o modelo foi treinado 5 vezes com cada uma das 256 combinações (4^4 , ou seja, o número de parâmetros elevado ao número de valores por parâmetro), para que houvesse, no mínimo, 5 resultados a ter em conta, aquando da escolha do valor ideal para o parâmetro. O erro utilizado para avaliar a qualidade da previsão foi o erro percentual absoluto médio (*MAPE*).

Após calcular o erro para todas as combinações, era necessário comparar as configurações através da média dos resultados de cada execução e escolher a configuração com um desempenho melhor em média. Assim sendo, fez-se uma análise, através de diagramas de caixas, representando a variação dos erros percentuais observados para cada valor que cada parâmetro tomou.

Ao longo da análise, o número de combinações que avançava para a afinação do parâmetro seguinte ia diminuindo, pois sempre que se escolhia o valor para determinado parâmetro, retiravam-se os restantes valores da tabela, para convergir para a combinação ideal.

Em primeiro lugar, fez-se a afinação do tamanho da *lag* e o resultado está representado na Figura 4.3.

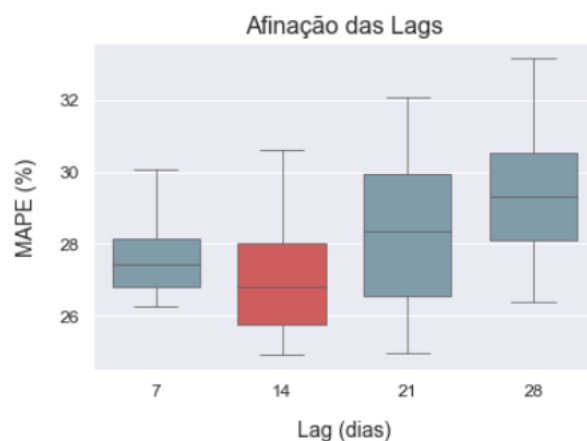


Figura 4.3: Diagrama de caixas para afinação do tamanho da *lag*.

De um modo geral, ajustar uma rede neural é encontrar o equilíbrio entre o desempenho e a variação média desse desempenho, sendo o resultado ideal um erro médio baixo com reduzida variabilidade.

A linha no centro de cada caixa mostra a mediana e a caixa mostra os percentis 25 e 75. Os traços abaixo e acima da caixa representam o erro mínimo e o erro máximo, respetivamente. Deste modo, o diagrama sugere, sinalizado a vermelho, que o valor mais promissor é o 14, isto significa que passar ao modelo a informação dos últimos 14 dias, em média, apresentou melhores resultados. Apesar de não ser o valor com a menor variação nem com o valor máximo mais baixo, é o que apresenta o valor mínimo e a mediana mais baixos.

De seguida, com a *lag* de 14 dias fez-se a mesma análise para o número de *epochs*, resultando no diagrama da Figura 4.4.

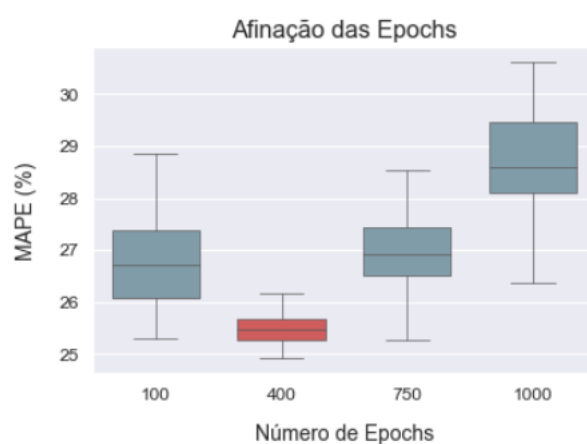


Figura 4.4: Diagrama de caixas para afinação do número de *epochs*.

Este foi sem dúvida o parâmetro mais fácil de definir, pois o valor 400 não deixa margem para dúvidas. Com a variação, a média e os extremos substancialmente mais baixos do que os restantes

valores, 400 demonstra ser o valor mais promissor para este parâmetro.

Partindo para o tamanho do lote, na Figura 4.5, pode-se constatar que, perante os valores definidos anteriormente para o tamanho da *lag* e para o número de *epochs*, o tamanho do lote não tem uma grande influência na variação do resultado, pois a escala no eixo do y é significativamente inferior.

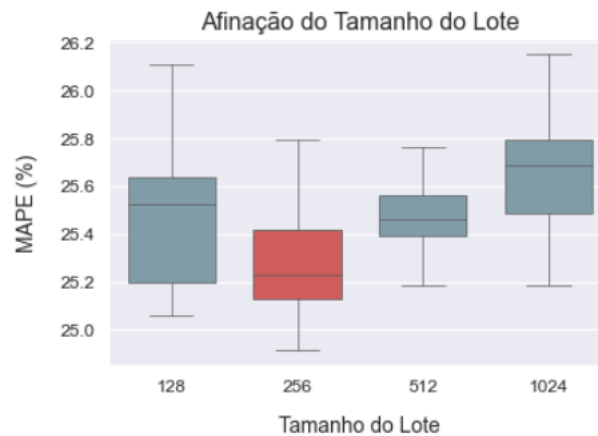


Figura 4.5: Diagrama de caixas para afinação do tamanho do lote.

De qualquer forma, o valor que apresenta uma variação mais promissora é o 256, tendo sido selecionado para a configuração ideal.

Por fim, no que respeita ao número de neurónios da camada escondida. O valor selecionado foi 20 neurónios, pois apesar do valor médio ser praticamente igual para 100 neurónios, mais de 50% dos valores concentram-se num intervalo inferior (ver Figura 4.6).

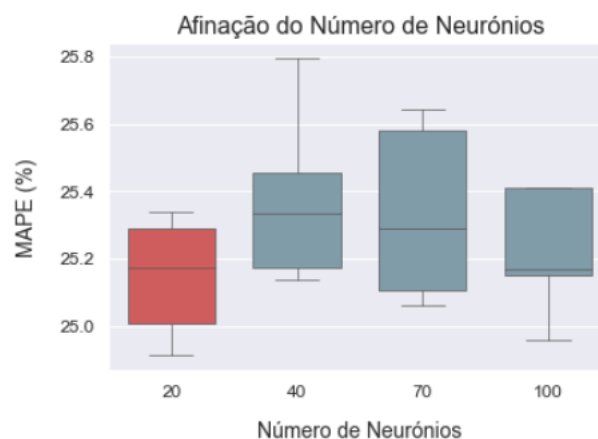


Figura 4.6: Diagrama de caixas para afinação do número de neurónios.

Tal como no diagrama do tamanho do lote, a escala do eixo y revela uma pequena dispersão nos valores do erro, indicando que o número de neurónios, perante a escolha dos parâmetros anteriores, também já não é muito relevante.

Seguindo esta análise, a combinação de valores ideal para os parâmetros da rede neuronal foi alcançada: *lag* com 14 dias, 400 *epochs*, 256 amostras por lote e 20 neurónios na camada escondida.

Assim sendo, o modelo foi executado mais uma vez com esta combinação e, a fim de validar a configuração, analisou-se a curva de aprendizagem para os dados de treino e de validação, representada na Figura 4.7. A curva de aprendizagem é um gráfico do desempenho da aprendizagem do modelo ao longo do tempo e pode ser utilizada para diagnosticar problemas durante a aprendizagem, como o *underfitting* e o *overfitting*, bem como se os conjuntos de dados de treino e validação são representativos.

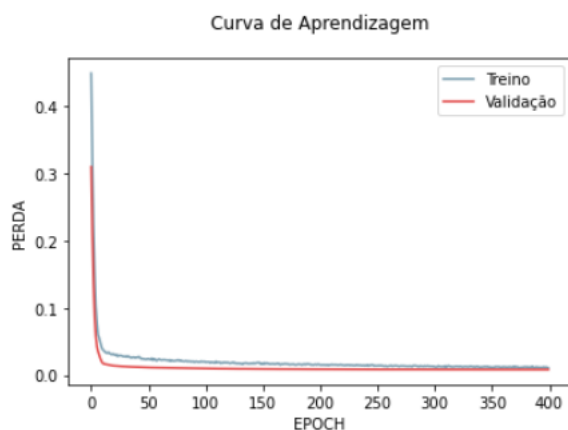


Figura 4.7: Curva de aprendizagem do modelo para os conjuntos de dados de treino e de validação.

Neste caso, o gráfico mostra um bom ajuste, tendo em conta que a perda ao longo do treino diminui para um ponto de estabilidade e a perda para o conjunto de dados de validação diminui para um ponto de estabilidade e apresenta um pequeno desfasamento, quase nulo, em relação à perda do treino.

Validada a configuração do modelo, fez-se a previsão para os dados de teste e calculou-se de novo o erro percentual em relação às vendas reais, para verificar a qualidade do modelo perante dados nunca antes vistos. O erro observado foi de 26,4% que, apesar de ser um erro considerável, é justificado pela granularidade dos dados da previsão e pelo facto das vendas reais serem influenciadas por outros produtos. De qualquer modo, o erro dos dados de teste permite assumir que o modelo é generalizável, pois não diverge muito dos erros obtidos para o conjunto de dados de validação.

Na Figura 4.8 é representado um gráfico com os valores reais, a azul, e os valores da previsão para todo o ano de 2019, a vermelho, englobando os dados do conjunto de validação e de teste. O gráfico torna visível que, de uma forma geral, a linha da previsão acompanha a flutuação da linha das vendas reais registadas.

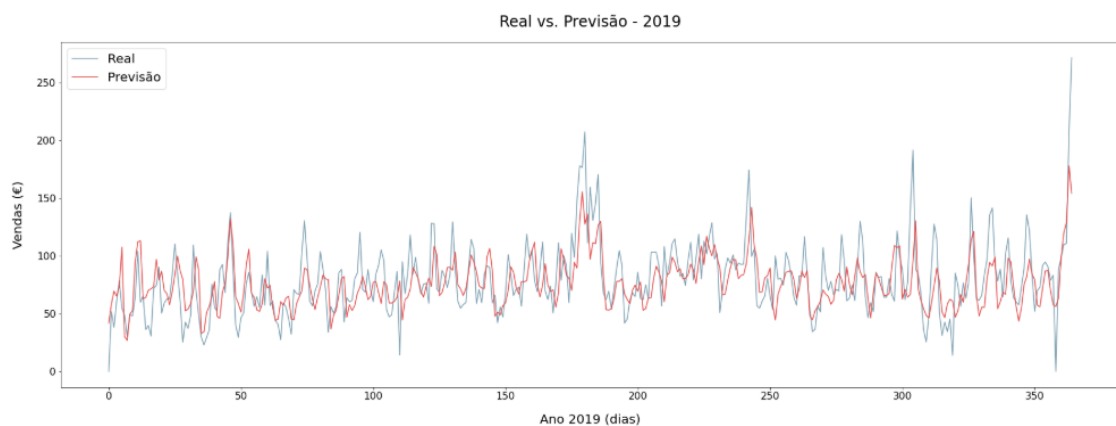


Figura 4.8: Vendas reais versus vendas previstas a partir dos dados de validação e de teste no ano 2019 (batata frita marca Continente).

No entanto, este modelo foi desenhado para prever o valor base para cada dia. Deste modo, é nesta fase que as percentagens das vendas em promoção no dia t são substituídas por 0% no conjunto de dados referentes ao ano 2019 e é feita a previsão do valor base.

Na Figura 4.9 é representado o gráfico no qual se observa o comportamento do valor base calculado em relação às vendas reais.

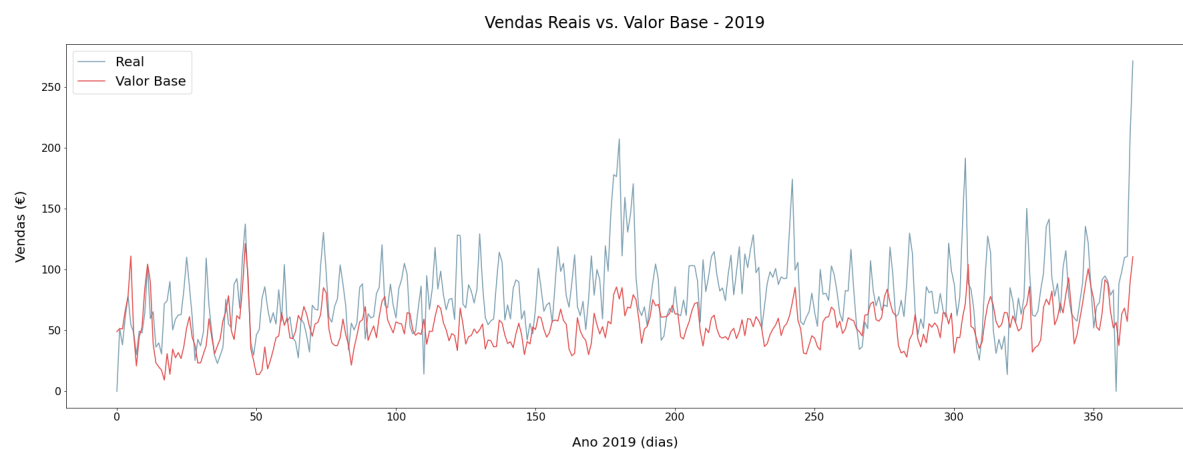


Figura 4.9: Vendas reais versus valor base das vendas para o ano 2019 (batata frita marca Continente).

Batatas fritas onduladas marca *Ruffles* (Produto B)

A mesma metodologia foi aplicada para o cálculo do valor base das batatas fritas onduladas da marca *Ruffles*. Nas Figuras 4.10 e 4.11 são representados os gráficos que resultam da previsão do conjunto de dados de validação e de teste e da previsão do valor base, respetivamente.

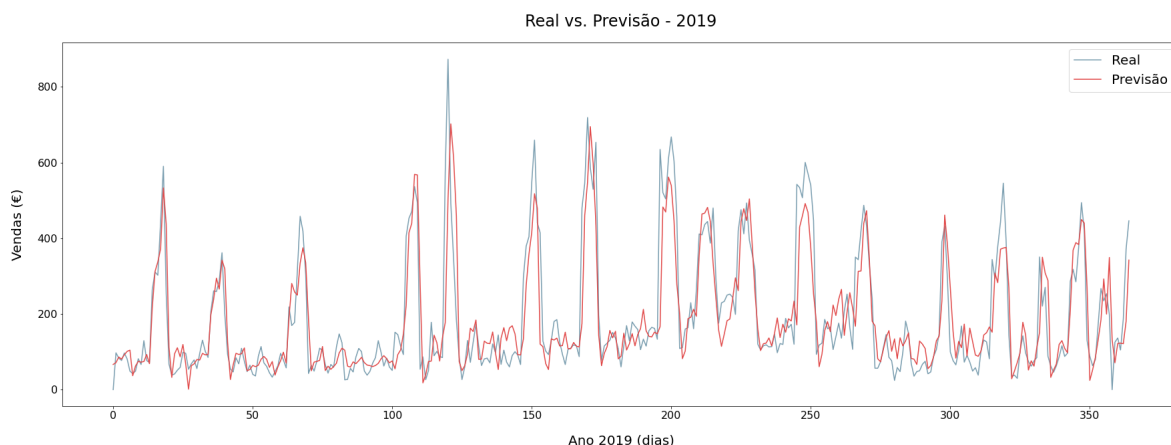


Figura 4.10: Vendas reais versus vendas previstas a partir dos dados de validação e de teste no ano 2019 (batata frita marca *Ruffles*).

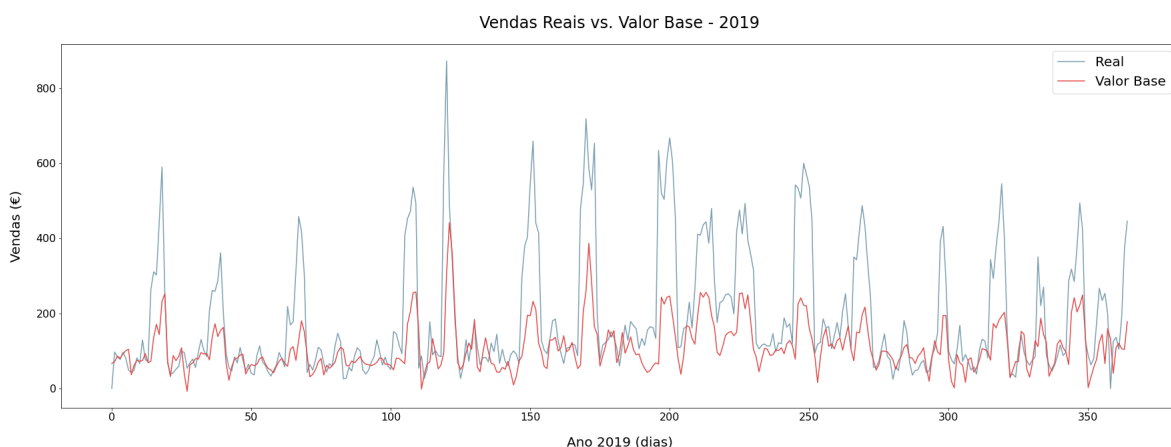


Figura 4.11: Vendas reais versus valor base das vendas para o ano 2019 (batata frita marca *Ruffles*).

Após calcular os valores base por dia para ambos os produtos, calcularam-se os *bumps* e removeu-se a sazonalidade das vendas reais e dos *bumps*, tal como foi explicado no Capítulo 3, dividindo estes valores por um fator de sazonalidade diário e outro semanal.

Terminando a fase de preparação do modelo principal, os dados alimentaram o modelo propriamente dito. De seguida, são representados os resultados da relação entre o par de produtos.

4.1.2 Apuramento da Relação entre um Par de Produtos

O pressuposto da relação entre um par de produtos ser composto por duas forças unidirecionais é agora comprovado através do exemplo prático que relaciona os dois produtos para o quais acabaram de ser ilustrados os valores base. Para tal, serão expostos os resultados do apuramento da relação para o par (A em promoção, B sem promoção) e o par (A sem promoção, B em promoção).

(A em promoção, B sem promoção)

A força desta relação é calculada passando ao modelo da regressão linear os dados dos dias do ano 2019 nos quais o produto A estava em promoção e o produto B não estava. O resultado obtido está representado na Figura 4.12, onde no eixo x está o *bump* do produto A e no eixo y estão as vendas totais de ambos os produtos nos dias que foram considerados no modelo. Na tabela 4.2 estão representados os resultados principais.

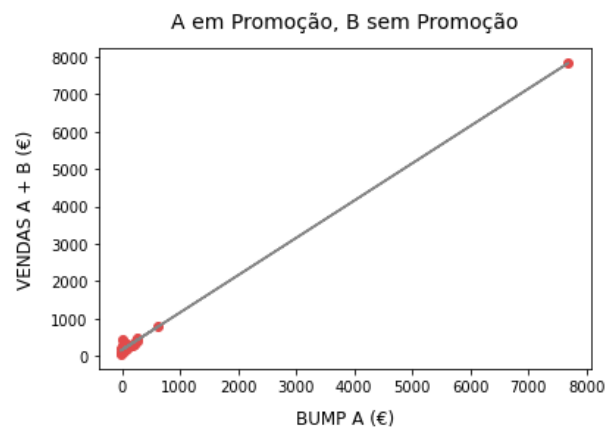


Figura 4.12: Gráfico da Regressão Linear do par (A em promoção, B sem promoção).

Interseção em $y=0$ (β_0)	166,74€
Coeficiente da Regressão (β_1)	1,00
R-quadrado (R^2)	1,00
Força da Relação (β_1-1)	0%

Tabela 4.2: Tabela de resultados referente à regressão linear da Figura 4.12.

Analisando os resultados e por observação do gráfico, verifica-se que há um ajuste perfeito da reta aos dados, refletido num valor de R^2 igual a 1, que aparentemente está a ser manipulado por um ponto com os valores de venda muito elevados. Tentou-se procurar uma justificação para esta observação, começando por identificar o dia em que ocorreu - 14 de janeiro de 2019. No entanto, não foi encontrado nenhum facto nem acontecimento que justificasse esta exceção e optou-se por remover este ponto, considerando-o um *outlier*. Os resultados, após remover este *outlier*, estão representados na Figura 4.13 e na Tabela 4.3.

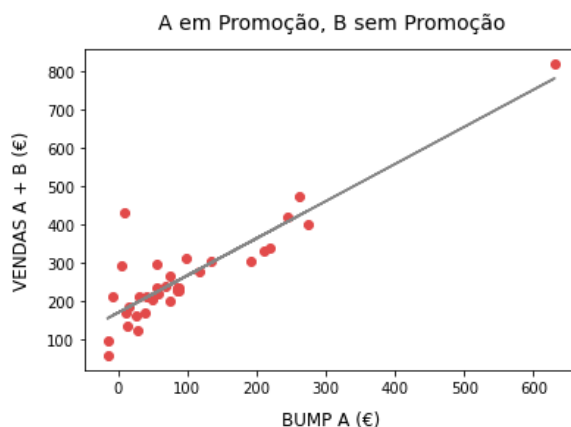


Figura 4.13: Gráfico da Regressão Linear do par (A em promoção, B sem promoção) após remover o primeiro *outlier*.

Interseção em $y=0$ (β_0)	169,17€
Coefficiente da Regressão (β_1)	0,97
R-quadrado (R^2)	0,79
Força da Relação (β_1-1)	-3%

Tabela 4.3: Tabela de resultados referente à regressão linear da Figura 4.13.

Através da análise do novo gráfico e dos novos resultados, detetou-se a existência de uma força de canibalização, mas pouco significativa, tendo em consideração que a métrica de avaliação (R^2), apesar de ter um valor considerado elevado, está claramente a ser afetado por mais um *outlier*, para o qual também não se encontrou justificação. Retirando este *outlier*, o declive da reta vai baixar e o valor absoluto da canibalização vai ser maior, no entanto os pontos vão afastar-se da reta da regressão que lhes é ajustada, baixando o valor do (R^2), como se pode verificar na Figura 4.14 e na Tabela 4.4.

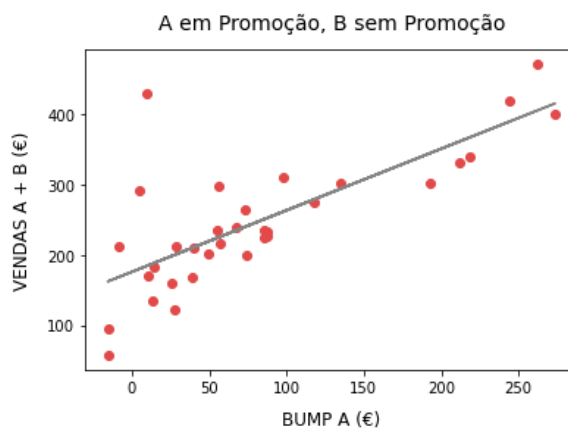


Figura 4.14: Gráfico da Regressão Linear do par (A em promoção, B sem promoção) após remover o segundo *outlier*.

Interseção em $y=0$ (β_0)	175,77€
Coefficiente da Regressão (β_1)	0,88
R-quadrado (R^2)	0,57
Força da Relação (β_1-1)	-12%

Tabela 4.4: Tabela de resultados referente à regressão linear da Figura 4.14.

Em suma, com os dados de 2019 referentes à loja selecionada, pode-se concluir que não há evidências de uma relação significativa das batatas fritas onduladas marca *Continente* sobre as batatas fritas onduladas marca *Ruffles*, analisando o período no qual as primeiras estiveram em promoção.

(A sem promoção, B em promoção)

Para calcular a força da relação no sentido contrário, foi implementado o mesmo procedimento, considerando os dados relativos aos dias do ano 2019 nos quais o produto B estava em promoção e o produto A não estava. Os primeiros resultados vêm refletidos na Figura 4.15 e na Tabela 4.5.

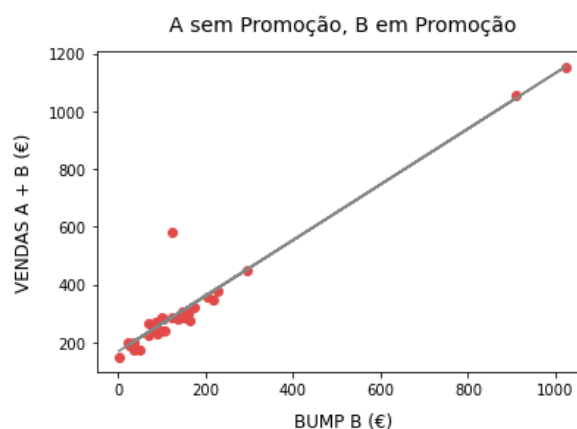


Figura 4.15: Gráfico da Regressão Linear do par (A sem promoção, B em promoção).

Interseção em $y=0$ (β_0)	167,81€
Coefficiente da Regressão (β_1)	0,96
R-quadrado (R^2)	0,93
Força da Relação (β_1-1)	-4%

Tabela 4.5: Tabela de resultados referente à regressão linear da Figura 4.15.

Neste caso, deteta-se por observação do gráfico que há 3 pontos que podem ser considerados *outliers*, não havendo nenhum evento que justifique um aumento exagerado das vendas de batatas fritas. Procedendo à remoção destes 3 pontos, obtêm-se os resultados representados na Figura 4.16 e na Tabela 4.6.

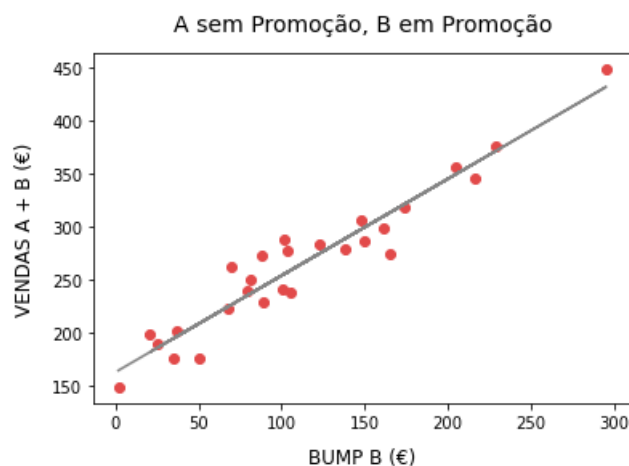


Figura 4.16: Gráfico da Regressão Linear do par (A sem promoção, B em promoção) após remover outliers.

Interseção em $y=0$ (β_0)	0,92
Coeficiente da Regressão (β_1)	162,22€
R-quadrado (R^2)	0,92
Força da Relação (β_1-1)	-8%

Tabela 4.6: Tabela de resultados referente à regressão linear da Figura 4.16.

Analisando os resultados, conclui-se que há uma relação significativa de canibalização das batatas fritas onduladas da marca *Ruffles* em promoção sobre as batatas fritas onduladas da marca *Continente*. Esta relação traduz-se numa diminuição das vendas diárias das batatas fritas da marca *Continente* num valor de 8% do incremento diário das batatas fritas da marca *Ruffles*, quando estas estão em promoção.

Em síntese, esta análise conclui que, na relação de um par de produtos, o efeito que resulta da promoção de um dos produtos não é igual nem simétrico ao efeito que resulta da promoção do outro produto. Para além disso, comprova que podem existir produtos sem relação ou com uma relação fraca. No entanto, resultados como este também são relevantes para o negócio.

4.1.3 Outros Resultados

Averiguaram-se outras relações no âmbito do caso de estudo do ano 2019 para a loja em estudo. Entre elas, foi detetada uma relação entre um par de produtos que, num sentido, apresentava um efeito de canibalização, negativo, com pouca representatividade e, no outro sentido, apresentava um efeito *Halo*, positivo e significativo. Os dois produtos são o molho de mostarda da marca *Savora* (produto C) e molho de *ketchup* da marca *Heinz* (produto D). Seguidamente, são apresentados os resultados desta relação, assumindo a afinação prévia do modelo do cálculo do valor base.

(C em promoção, D sem promoção)

Seguindo a lógica do exemplo das batatas fritas, filtraram-se os dados dos dias do ano 2019 nos quais o produto C estava em promoção e o produto D não estava e aplicou-se o modelo a estes dados. Na Figura 4.17 e na Tabela 4.7 estão refletidos os resultados.

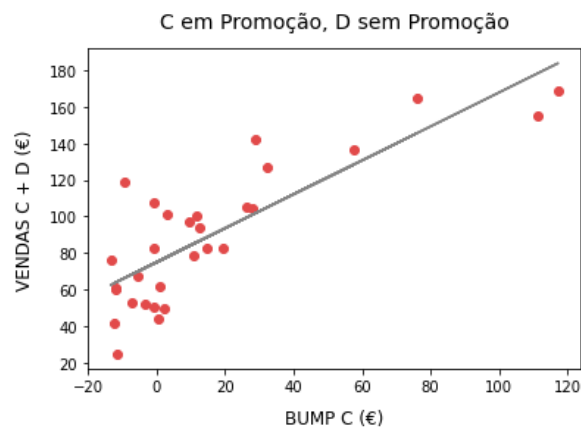


Figura 4.17: Gráfico da Regressão Linear do par (C em promoção, D sem promoção).

Interseção em $y=0$ (β_0)	74,96€
Coefficiente da Regressão (β_1)	0,93
R-quadrado (R^2)	0,67
Força da Relação (β_1-1)	-7%

Tabela 4.7: Tabela de resultados referente à regressão linear da Figura 4.17.

Através da análise dos resultados, conclui-se que a relação do molho de mostarda da marca Savora sobre o molho de *ketchup* da marca Heinz é de canibalização, no entanto o grau de confiança nesta relação é só de 67%.

(C sem promoção, D em promoção)

No outro sentido da relação, verifica-se uma relação positiva, demonstrada na Figura 4.18 e nos resultados da Tabela 4.8.

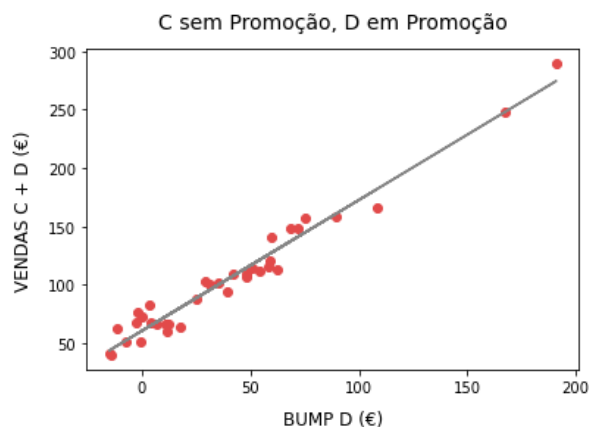


Figura 4.18: Gráfico da Regressão Linear do par (C sem promoção, D em promoção).

Interseção em $y=0$ (β_0)	60,58€
Coefficiente da Regressão (β_1)	1,12
R-quadrado (R^2)	0,96
Força da Relação (β_1-1)	12%

Tabela 4.8: Tabela de resultados referente à regressão linear da Figura 4.18.

Por oposição, a relação do molho de *ketchup* da marca *Heinz* sobre o molho de mostarda da marca *Savora* é positiva, indicando que as vendas da mostarda sobem diariamente 12% do valor do incremento diário do *ketchup*, quando este está em promoção.

Capítulo 5

Conclusões, Limitações e Trabalho Futuro

5.1 Conclusões

Tendo em consideração que o ambiente no setor do retalho alimentar em Portugal é extremamente competitivo e tem sofrido uma elevada pressão sobre a atividade promocional, surge a necessidade de garantir que os efeitos desta atividade promocional sejam avaliados. De facto, estudos comprovam que a eficácia das promoções, no mercado português, tem vindo a diminuir. Deste modo, este projeto visou apoiar a definição da estratégia promocional para uma das empresas que lidera este setor em Portugal - o Continente.

A eficácia das promoções passa por compreender os efeitos que resultam da ação promocional. No entanto, há uma tendência em avaliar, unicamente, os efeitos incrementais das vendas ou do número de transações do produto promovido, desprezando os efeitos que influenciam as vendas de outros produtos dentro do mesmo ecossistema. Assim sendo, o foco do projeto foi compreender estes efeitos entre produtos e extrair a relação entre qualquer par de produtos, quer fossem da mesma ou de diferentes categorias.

Partindo de um modelo que calcula a percentagem da canibalização dentro de uma categoria, foi desenhado um modelo semelhante para calcular a força da relação entre dois produtos, que estaria relacionada com a canibalização caso fosse negativa ou com o efeito *Halo* caso fosse positiva. Para além deste modelo, foi necessário desenvolver outro modelo que calculasse o valor base de vendas a partir do qual seria medido o incremento que resulta da ação promocional, essencial para o cálculo da relação entre produtos.

De um modo geral, a metodologia proposta comprova os pressupostos iniciais de que existem realmente pares de produtos com as vendas parcialmente relacionadas. Isto verifica-se quando há uma ação promocional num dos dois produtos, sendo esta relação não bidirecional, mas sim composta por duas forças unidirecionais. Também é importante destacar que estas forças podem ser nulas ou pouco representativas, caso não haja informação suficiente para tirar conclusões, para

além de positivas e negativas, sendo associadas aos efeitos *Halo* e de canibalização, respetivamente. Como exemplo, são apresentadas as relações entre batatas fritas Continente e batatas fritas *Ruffles* que sugerem que há uma canibalização de 8% das vendas das batatas fritas *Ruffles* sobre as do Continente, quando as primeiras estão em promoção, não se verificando uma relação direta no sentido contrário da relação; e as relações entre o molho de mostarda da Savora e o molho de *Ketchup* da *Heinz* que sugerem que, apesar de não haver evidências de uma relação significativa da mostarda sobre o *Ketchup*, há uma relação de complementaridade de 12% das vendas do *Ketchup*, quando este está em promoção.

Através da metodologia proposta é possível verificar a existência de relações entre produtos no ecossistema que se pretender, isto é, é possível aferir as relações no total das lojas, por insígnia, para um grupo de lojas ou por loja; ou com base no período de tempo que se entender, tendo em conta que quanto maior for o número de dados, maior credibilidade tem o resultado; e pode ainda ser aplicada para níveis diferentes de agrupamentos de produtos, ou seja, ir ao pormenor do *SKU* ou agrupar ao nível categoria-marca da estrutura *VPH* e perceber os efeitos entre marcas concorrentes na mesma categoria.

Do ponto de vista do negócio, estas relações são relevantes, pois permitem identificar os produtos que geram impacto negativo e os que geram impacto positivo, facilitando o processo da definição das ações promocionais. O facto das relações poderem ser extraídas para diferentes ecossistemas, torna-se ainda mais relevante, pois ajuda a personalizar as estratégias promocionais. Atualmente, com o programa de fidelização do cartão Continente, é possível dirigir as promoções para um grupo específico de pessoas. Assim sendo, saber como um grupo de clientes com comportamentos semelhantes relaciona os produtos entre si, ajudará a evitar dirigir promoções com uma alta taxa de substituíbilidade e a favorecer promoções com altas taxas de complementaridade.

Em suma, a abordagem ao problema foi validada por especialistas do negócio, que identificaram um grande potencial na metodologia proposta como ferramenta de apoio à decisão na definição da estratégia promocional da empresa.

5.2 Limitações e Trabalho Futuro

Ao longo do desenvolvimento do projeto surgiram algumas limitações que podem ser vistas como potenciais oportunidades de melhoria para um trabalho futuro, para além de outros caminhos que ficaram por explorar.

Em relação ao modelo principal, ficaram por explorar técnicas de remoção de *outliers*, tendo sido identificados e removidos por apresentarem valores extremos, quando comparados com a maioria dos outros pontos.

Após descobrir o modelo ideal para extrair as relações entre produtos, a procura de um modelo para calcular o valor base das vendas por produto foi o maior desafio, dada a quantidade de fatores externos que influenciam as vendas diariamente e dado o ambiente extremamente promocional que se vive em Portugal. Apesar do modelo proposto ser inovador, não é facilmente escalável a todos os produtos, por necessitar de uma afinação de parâmetros prévia. No entanto, ficaram por

experimental e explorar algumas técnicas para aperfeiçoar o modelo: as técnicas de *Dropout* e a adição de mais camadas, foram pouco exploradas, e as decisões foram tomadas com base em estudos semelhantes e processos de tentativa e erro; a função de perda e o algoritmo de otimização foram, também, escolhidos com base em exemplos de aplicação da rede neuronal.

Relativamente ao trabalho futuro, o modelo pode ser adaptado para englobar variáveis como o valor dos descontos, o tipo de desconto ou até o meio em que é promovido.

Para terminar, o modelo poderá ser testado para diferentes segmentos de clientes, para no futuro facilitar e otimizar o envio de cupões personalizados.

Referências

- [AG⁺14] Kusum L Ailawadi, Sunil Gupta et al. Sales promotions. *History of Marketing Science, Singapore*, pages 463–497, 2014.
- [AHCT06] Kusum L Ailawadi, Bari A Harlam, Jacques Cesar e David Trounce. Promotion profitability for a retailer: the role of promotion, brand, category, and store characteristics. *Journal of Marketing Research*, 43(4):518–535, 2006.
- [BC12] Dominic Barton e David Court. Making advanced analytics work for you. *Harvard business review*, 90(10):78–83, 2012.
- [BN90] Robert C Blattberg e Scott A Neslin. *Sales promotion: Concepts, methods, and strategies*. Prentice Hall, 1990.
- [BN93] Robert C Blattberg e Scott A Neslin. Sales promotion models. *Handbooks in Operations Research and Management Science*, 5:553–609, 1993.
- [Chi91] Jeongwen Chiang. A simultaneous approach to the whether, what and how much to buy questions. *Marketing Science*, 10(4):297–315, 1991.
- [Chi93] Pradeep K Chintagunta. Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science*, 12(2):184–208, 1993.
- [Cro] CrossCap. Guide To Analyzing The Overall Lift Of A Retail Promotion. URL: <https://www.crosscap.com/blog/guide-to-analyzing-the-overall-lift-of-a-retail-promotion>.
- [Cro17] CrossCap. 3 Strategies For Measuring Sales Lift In A Retail Promotion, 2017. URL: <https://www.crosscap.com/blog/3-strategies-for-measuring-sales-lift-in-a-retail-promotion>.
- [DHM10] Thomas H Davenport, Jeanne G Harris e Robert Morison. *Analytics at work: Smarter decisions, better results*. Harvard Business Press, 2010.
- [DNV04] Xavier Drèze, Patricia Nisol e Naufel J Vilcassim. Do promotions increase store expenditures? a descriptive study of household shopping behavior. *Quantitative Marketing and Economics*, 2(1):59–92, 2004.
- [GBC16] Ian Goodfellow, Yoshua Bengio e Aaron Courville. *Deep learning*. MIT press, 2016.
- [GNA10] K Gedenk, SA Neslin e KL Ailawadi. Sales promotion in: Retailing in the 21st century: Current and future trends, 2010.
- [Gup88] Sunil Gupta. Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing research*, 25(4):342–355, 1988.

- [IW99] J Jeffrey Inman e R Winer. Impulse buying. *The Wall Street Journal*, 15, 1999.
- [MMR20] Rod McColl, Renaud Macgilchrist e Shuddhasattwa Rafiq. Estimating cannibalizing effects of sales promotions: The impact of price cuts and store type. *Journal of Retailing and Consumer Services*, 53:101982, 2020.
- [OBM15] Steve Ogden-Barnes e Stella Minahan. *Sales promotion decision making: concepts, principles, and practice*. Business Expert Press, 2015.
- [Pes18] Aleksí Pesonen. Quantifying promotion effectiveness for a retailer. Master’s thesis, Aalto University, Finland, 2018.
- [Por19] Nielsen Portugal. 3,5 mil milhões de euros na “Selva Promocional” dos bens de grande consumo, 2019. URL: <https://www.nielsen.com/pt/pt/insights/article/2019/3-point-5-billion-euros-in-promotional-jungle-of-consumer-goods/>.
- [TSK13] Pang-Ning Tan, Michael Steinbach e Vipin Kumar. Data mining cluster analysis: basic concepts and algorithms. *Introduction to data mining*, pages 487–533, 2013.
- [VHGW03] Harald J Van Heerde, Sachin Gupta e Dick R Wittink. Is 75% of the sales promotion bump due to brand switching? no, only 33% is, 2003.
- [VHLW04] Harald J Van Heerde, Peter SH Leeflang e Dick R Wittink. Decomposing the sales promotion bump with store data. *Marketing Science*, 23(3):317–334, 2004.
- [VHN08] Harald J Van Heerde e Scott A Neslin. Sales promotion models. In *Handbook of marketing decision models*, pages 107–162. Springer, 2008.
- [VHN17] Harald J Van Heerde e Scott A Neslin. Sales promotion models. In *Handbook of marketing decision models*, pages 13–77. Springer, 2017.
- [WH00] Rüdiger Wirth e Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Springer-Verlag London, UK, 2000.