Previsão de Vendas de um Marketplace Digital através de Técnicas de Machine Learning

Ricardo Moura Vieira Batista

Dissertação de Mestrado

Orientador na FEUP: Prof. José Luís Cabral de Moura Borges

Mestrado em Engenharia e Gestão Industrial

2022-07-11

Resumo

O Dott é uma *start-up* tecnológica que conecta compradores e vendedores através do seu *marketplace* digital. Com o crescimento do consumo online, e com a pandemia da Covid-19 a atuar como um acelerador, a quantidade de vendas tornou-se cada vez mais imprevisível, fazendo com que a variabilidade seja um dos fatores mais importantes a ter em conta.

Este projeto foi proposto pelo Dott com a intenção de ajudar na tomada de decisão em diversas áreas. Neste momento, o Dott não possui qualquer tipo de processo de previsão de vendas. O objetivo do projeto passa por construir um modelo de *machine learning* que utilize informação armazenada na base de dados do Dott para gerar previsões da quantidade vendida de cada produto para a semana seguinte, permitindo à empresa obter conhecimento antecipado sobre as suas vendas.

Começou-se por enquadrar o problema e obter os dados necessários à construção do modelo. Após a obtenção dos dados históricos relativos às vendas de cada produto, seguiu-se a análise dos mesmos. Por fim, procedeu-se ao pré-processamento dos dados de forma a ficarem disponíveis para serem imputados nos diferentes algoritmos.

Para a fase de treino dos modelos de ML foram definidas duas abordagens distintas. A primeira abordagem ("9 variáveis") implicou a utilização de dados de entrada dos modelos com apenas 9 variáveis independentes. A possibilidade de diminuir o *overfitting* e o custo computacional serviu como motivação para o desenvolvimento desta abordagem. A segunda abordagem ("60 variáveis") implicou a utilização dos dados de entrada com 60 variáveis independentes. Para a análise das diferentes abordagens foram gerados cenários diferentes, considerando vários algoritmos de ML, nomeadamente *Linear Regression*, *Decision Tree*, *Random Forest*, *LightGBM*, *XGBoost* e *K-Nearest Neighbor*.

Ambas as abordagens de treino foram testadas com observações desconhecidas, assegurando que os modelos foram avaliados em condições próximas da realidade. Os resultados demonstraram que o melhor modelo da abordagem "60 variáveis"retornou os melhores resultados. Esta conclusão validou a abordagem "60 variáveis".

No entanto, quando aplicado o método de avaliação aos 25 produtos mais vendidos, foi demonstrado que as previsões geradas não eram suficientemente precisas.

Abstract

Dott is a tech startup that connects buyers and sellers through its digital marketplace. With the growth of online consumption, and with the Covid-19 pandemic acting as an accelerator, the amount of sales has become increasingly unpredictable, making variability one of the most important factors to take into account.

This project was proposed by Dott with the intention of helping decision making in several areas. At the moment, Dott does not have any kind of sales forecasting process. The project's goal is to build a machine learning model that uses information stored in Dott's database to generate forecasts of the quantity of each product sold for the following week, allowing the company to obtain advance knowledge about its sales.

The initial step was to frame the problem and obtain the data needed to build the model. After obtaining the historical data on the sales of each product, a briefly analysis of the data was performed. This was followed by a data preprocessing step so that it would be available to be imputed in the different algorithms.

For the training phase of the ML models two distinct approaches were defined. The first approach ("9 variáveis") implied the use of model input data with only 9 independent variables. The possibility of decreasing overfitting and computational cost served as motivation for the development of this approach. The second approach ("60 variáveis") implied the use of the input data with 60 independent variables. For the analysis of the different approaches, different scenarios were generated, considering several ML algorithms,namely Linear Regression, Decision Tree, Random Forest, LightGBM, XGBoost and K-Nearest Neighbor.

Both training approaches were tested with unseen observations, ensuring that the models were evaluated under conditions close to reality. The results showed that the best model of the "60 variáveis"approach returned the best results. This conclusion validated the "60 variáveis"training approach.

However, when the evaluation method was applied to the 25 best-selling products, it was shown that the forecasts generated were not sufficiently accurate.

Agradecimentos

Para começar, gostaria de agradecer ao meu orientador, Professor José Luís Borges, pela disponibilidade e conhecimento fornecido.

Ao Miguel Serrano e André Clemêncio, os meus mentores do projeto dentro da empresa, pela orientação, apoio e constante encorajamento. Forneceram-me as ferramentas necessárias para me orientar sempre na direção correta. A ambos, um grande obrigado por partilharem comigo esta experiência e por me acompanharem durante estes meses.

À Dott por ser uma empresa que me permitiu crescer e desafiar-me a mim próprio todos os dias. Uma palavra especial para todos que me proporcionaram excelentes momentos.

A toda a capacidade do Mestrado em Engenharia e Gestão Industrial da FEUP. O sucesso e a qualidade deste curso estão sem duvida ligados aos professores apaixonados pelo ensino e que preparam os seus alunos para os maiores desafios.

À minha família, especialmente à minha mãe, Maria Celeste Moreira Moura, ao meu pai, Valdemar Rodrigo Vieira Batista e à minha irmã, Romana Moura Vieira Batista por todo o apoio, principalmente em momentos mais difíceis, pela confiança depositada nas minhas capacidades durante a minha jornada pessoal e académica e por me permitirem a oportunidade de seguir os meus sonhos. Às minhas empregadas, que considero fazerem parte da minha família, Mia e Linda, por desde sempre estarem presentes na minha vida.

A todos os meus amigos que acompanharam o meu percurso na Faculdade de Engenharia da Universidade do Porto, Alice Sousa, Catarina Silva, Dalila Ribeiro, Daniel Carvalho, Daniel Figueiredo, Diogo Costa, Duarte Araújo, Eduardo Monteiro, Hugo Braga, João Gonçalves, Luís Meireles, Marta Pereira, Pedro Valente, Raquel Machado, Rodrigo Castro, Rodrigo Santos, Sofia Magalhães e Tiago Tavares por tornarem este caminho inesquecível.

Aos meus amigos do secundário, Alice Sousa, Ana Luísa, Barbara Viegas, Carolina Faria, Catarina Silva, Gabriela Cardoso, Gonçalo Gomes, Inês Cardoso, Isabel Graça, João Silva, Luís Moutinho, Mário Pereira, Paulo Vieira e Ricardo Martins pela companhia desde muito novo e por todos os momentos extraordinários partilhados.

"Não há impossíveis e a prova disso foi o Benfica ter levado 7x0 do Celta de Vigo."

Jorge Nuno Pinto da Costa

Conteúdo

Acronyms and Symbols

PME	Pequenas Médias Empresas
ML	*Machine Learning*
GMV	*Gross Merchandise Value*
B2C	*Business-to-Consumer*
ICT	*Information and Communications Technology*
ARIMA	*Autoregressive Integrated Moving Average*
MAE	*Mean Absolute Error*
MSE	*Mean Squared Error*
RMSE	*Root Mean Squared Error*
MAPE	*Mean Absolute Percentage Error*
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
RF	*Random Forest*
LR	*Linear Regression*
DT	*Decision Tree*
LGBM	*LightGBM*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O consumo online registou grandes taxas de crescimento nos últimos três anos. A adesão ao comércio digital, que vinha a aumentar a um ritmo constante ao longo dos anos, sofreu um aumento significativo devido à pandemia. A passagem dos consumidores de retalho tradicional para compras online trouxe uma grande mudança no paradigma de muitas empresas, que passaram o seu foco para o serviço online. Deste modo, empresas que não possuíam uma plataforma preparada para o comércio online, optaram por integrar os seus produtos em *marketplaces* digitais de forma a ganhar mais visibilidade. Porém, o aumento de vendas inesperado fez com que as empresas enfrentassem diversos problemas, entre eles, a rotura de stock dos produtos mais comercializados. Além disso, a qualidade do serviço ao cliente tornou-se, ainda mais, num ponto fulcral para as empresas como forma de diferenciação e vantagem competitiva no que diz respeito ao comércio online. Assim, a existência de um modelo eficaz de previsão de vendas tornou-se, mais do que nunca, num dos pontos mais importantes para qualquer empresa da atualidade.

Esta dissertação realizada na equipa de Dados e Análises do Dott, centra-se no desenvolvimento de um algoritmo de *Machine Learning* que permita à empresa ter um modelo de previsão de vendas. Ao longo deste capítulo, na secção 1.1, é dada uma visão geral do comércio eletrónico e da importância de um modelo de previsão de vendas no negócio online juntamente com as razões e motivações para a criação deste projeto. De seguida, na secção 1.2, é feita uma breve apresentação da empresa, bem como do departamento onde ocorreu o desenvolvimento do projeto. Posteriormente, na secção 1.3, são explicados os objetivos do projeto, seguidos da secção 1.4, onde a metodologia aplicada é descrita. Finalmente, a estrutura desta dissertação é delineada na secção 1.5.

1.1 Motivação

Entre 2019 e 2020, as vendas de retalho através do comércio eletrónico em todo o mundo passaram dos 3,351 biliões de dólares para os 4,248 biliões de dólares, registando, assim, um aumento de 27,76%. Em 2021, este valor ascendeu aos 4,938 biliões de dólares, sendo que as

previsões apontam a que as receitas do comércio eletrónico de retalho atinjam, em 2022, os 5,542 biliões de dólares (Chevalier, 2022).

Desde o início da pandemia da Covid-19, os consumidores em todo o mundo têm estado fortemente dependentes do comércio eletrónico, desde bens essenciais a presentes de férias. Juntamente com as encomendas domésticas generalizadas e as preocupações com o vírus, a pandemia acelerou a adesão do comércio eletrónico pelos consumidores e empresas (Mathradas, 2020).

A indústria de retalho não se manteve indiferente à pandemia. Os *marketplaces* digitais também acompanharam a taxa de crescimento de consumidores durante a pandemia da Covid-19. Na verdade, este crescimento foi superior nestes mercados visto que muitos vendedores procuraram *marketplaces* digitais onde pudessem vender os seus produtos para colmatar a falta de uma plataforma própria preparada para o comércio online. Os *marketplaces* digitais, devido à sua robusta estrutura de comércio eletrónico, aumentaram as suas receitas. A Amazon, considerado o maior *marketplace* digital do mundo, registou um aumento do valor bruto de mercadorias (GMV) de 22% em 2021 relativamente a 2020 (Weise, 2021). Alibaba, outro grande *marketplace* digital, registou um aumento do GMV de 2020 para 2021 em 14% (Ma, 2022). Este aumento de vendas nos *marketplaces* deveu-se não só à pandemia, como também ao facto de mais produtos terem sido disponibilizados devido à entrada de novos vendedores.

A previsão de vendas pode ajudar as empresas a aumentar a qualidade da tomada de decisões e a colmatar possíveis problemas. Exemplificando, no mundo atual, a situação mais desfavorável que uma organização enfrenta, quando se trata de venda online, é a ocorrência de ruturas de stock. Os consumidores ficam desanimados quando se apercebem que um artigo desejado se encontra fora de stock. Estes eventos estão entre as experiências mais frustrantes para os consumidores online e criam desilusão e frustração também para os retalhistas. Além da perda de receitas, podem levar a oportunidades perdidas para atrair compradores, e a potenciais danos para a marca de um retalhista (Keenan, 2021). Neste caso, a previsão de vendas não só ajuda os retalhistas eletrónicos a fornecer aos clientes os produtos que desejam, como também ajuda a mantê-los à frente de qualquer concorrência potencial. Os retalhistas eletrónicos podem aumentar o inventário de certos produtos antes que os níveis de stock se tornem demasiados baixos durante uma venda. Isto mantém os clientes satisfeitos, ao mesmo tempo que impede de passarem para produtos semelhantes de marcas rivais.

A previsão de vendas por parte de um *marketplace* digital pode servir como a prestação de um serviço com valor acrescentado para os vendedores e uma vantagem competitiva em relação a outros *marketplaces* para atrair mais vendedores para a sua plataforma.

O contexto atual contribuiu para o aumento da importância das empresas possuírem um modelo de previsão de vendas eficaz. Assim, esta janela de oportunidade serviu como principal motivação para a realização deste projeto.

1.2 Contexto Empresarial

O Dott é uma *start-up* criada a partir de uma *joint venture* formada pelos CTT e pela Sonae, dois dos maiores intervenientes portugueses nas cadeias logísticas e nos mercados retalhistas, respetivamente. A ideia original veio dos CTT que inovaram e passaram por uma grande transformação digital. Ao unir forças com a Sonae, o Dott conseguiu fornecer um nível específico de conhecimentos de marketing online aos seus parceiros (D'Orey, 2021).

A empresa lançou a sua plataforma web no dia 1 de Maio de 2019. Com apenas um ano de funcionamento, o Dott reuniu mais de 700 vendedores e 2,5 milhões de produtos à disposição dos seus clientes distribuídos por 17 categorias diferentes, tais como: Moda, Saúde e Cuidados, Desporto entre outros. No final de 2021, com apenas dois anos e meio de existência, o Dott apresentava no seu *marketplace* duas mil empresas e comercializava seis milhões de produtos diferentes. O Dott é o primeiro mercado generalista em Portugal e rege-se pelo lema de mudar a forma como os portugueses compram e vendem online, sendo um *marketplace* puro que liga os compradores e vendedores através da sua plataforma. Uma das grandes virtudes da empresa é a facilidade da entrada de marcas portuguesas no seu *marketplace*. O objetivo passa por atrair pequenas e médias empresas (PME) no mundo do comércio eletrónico para adotarem este canal de venda, fazendo com que o investimento inicial destas empresas seja nulo. Desta forma, o Dott pretende ajudar estas PME portuguesas, disponibilizando uma forma barata de entrar no mundo digital, visto que é o Dott que lida com toda a cadeia logística, pagamentos, faturação, manutenção de catálogos e web-site, prestando assistência através dos seus gestores de conta e equipa de apoio ao cliente.

A empresa necessita de uma ferramenta que disponibilize informação sobre a quantidade de vendas futuras de cada produto para poder tomar decisões em várias áreas. Por exemplo, o Dott tem em vista colmatar problemas prejudiciais à empresa como a ocorrência de eventos de rutura de stock que resultam em perda de receitas e clientes. Assim, esta ferramenta pode ajudar a perceber a quantidade necessária de cada produto para notificar o seu vendedor, servindo como um serviço com valor acrescentado para os vendedores presentes no *marketplace* Dott. Neste sentido, a empresa procura um modelo de previsão de vendas.

O departamento de *Data and Analytics* é essencial no que diz respeito ao sucesso da organização. A equipa de analítica dedica-se à recolha de todos os grandes dados do Dott, e baseia-se nesses dados para conceber futuros modelos e estratégias de negócio. Esta equipa trabalha desde a inteligência artificial ao marketing e dita como as políticas devem assegurar a melhor produtividade e melhor valor da marca, processando milhares de dados. Por isso, este projeto surgiu e foi desenvolvido pelo departamento de dados e análises que é o departamento responsável pelo desenvolvimento de qualquer tipo de algoritmo de *machine learning* no Dott.

1.3 Objetivos do Projeto

Os objetivos deste projeto foram definidos de forma a acrescentar valor à empresa e aos seus parceiros com base no aumento do nível de serviço. O principal objetivo deste projeto foi o desenvolvimento de um modelo ML de previsão de vendas semanal de modo a prever as unidades vendidas de cada produto presente no *marketplace* Dott.

A construção deste modelo tem como objetivo propulsionar o crescimento da empresa. As previsões ajudam a identificar sinais de alerta precoces e, consequentemente, atuar sobre eles. Assim, o algoritmo pretende possibilitar uma previsão de vendas contínua, permitindo à organização responder com menos falhas ao ritmo atual do mercado.

Relativamente às aplicações deste algoritmo, que se encontram fora do âmbito deste projeto, o desenvolvimento do modelo de previsão de vendas tem como intuito acrescentar valor à empresa. Por um lado, uma das principais aplicações é acrescentar valor ao serviço prestado aos vendedores presentes no *marketplace* a partir do fornecimento da informação da quantidade necessária por semana, relativa a cada produto, de modo a evitar roturas de stock e, assim, evitar insatisfação por parte dos consumidores e prejudicar a imagem do vendedor. Por outro lado, permitirá à empresa atuar continuamente sobre as previsões geradas, nomeadamente, na criação de campanhas de marketing a partir da *newsletter* do Dott, destacando, estrategicamente, produtos específicos.

1.4 Método seguido no projeto

Inicialmente, o projeto passou por uma análise geral de modo a ter uma perspetiva empresarial das implicações e requisitos de um modelo de previsão de vendas. Foi, também, realizada uma análise mais profunda nas áreas relevantes para assegurar uma compreensão profunda de todos os tópicos.

Após a compreensão do problema, foi realizada uma revisão bibliográfica minuciosa de modo a perceber o estado de arte, bem como uma análise séria a conceitos técnicos abordados ao longo da dissertação, mais especificamente, os princípios de Machine Learning, com especial foco em regressão de séries temporais.

Relativamente ao desenvolvimento do modelo de previsão de vendas, numa primeira fase foram aplicados métodos de tratamento de dados ao conjunto de dados históricos de vendas obtido através da base dados do Dott. De seguida, os dados foram aplicados aos algoritmos explorados. Finalmente, diferentes cenários foram testados relativamente aos diferentes modelos de *machine learning* para determinar a combinação ótima dos fatores que melhor se adequam ao problema.

No final, foi realizada uma comparação entre todos os cenários, e retirada a conclusão. A comparação foi estabelecida através das métricas de avaliação de um modelo de regressão.

1.5 Estrutura da dissertação

Esta dissertação está estruturada em seis capítulos. No presente capítulo é introduzido o tema da dissertação e apresentada a empresa onde a mesma foi desenvolvida. Além disso, é apresentada a motivação do projeto, os seus objetivos e a metodologia utilizada. No capítulo seguinte é realizada uma revisão de todos os conceitos teóricos fundamentais à realização desta tese.

A cadeia de valor da empresa é apresentada no terceiro capítulo, bem como a sua estrutura e diferentes departamentos. Numa segunda parte é descrito o problema aquando da apresentação da situação *AS-IS*.

No quarto capítulo, a metodologia aplicada neste projeto é especificada. Neste capítulo é descrito o conjunto de dados utilizado, bem como todas as etapas de pré-processamento necessárias. Posteriormente, é apresentada uma explicação relativa a cada um dos modelos propostos e termina com a definição dos diferentes cenários a serem testados no capítulo 5.

No quinto capítulo são apresentados os resultados obtidos da aplicação dos modelos propostos e comparados os seus desempenhos. No último capítulo, é estabelecida uma visão geral dos resultados do projeto e realizado um reflexo relativo aos seus objetivos. Por fim, esta dissertação é finalizada com a sugestão de possíveis melhorias a realizar no futuro.

Capítulo 2

Revisão de literatura

A relevância científica desta tese de mestrado é demonstrada neste capítulo. Este segundo capítulo resume a literatura publicada sobre a previsão de vendas, incluindo a literatura escrita sobre a área do comércio eletrónico. Por fim, são apresentados os conhecimentos teóricos dos modelos de *machine learning* desenvolvidos neste projeto.

2.1 Comércio Eletrónico

Um relatório publicado pela Cramer-Flood (2022) estima que as vendas do comércio eletrónico de retalho irão aumentar, a nível mundial, de 4,938 biliões de dólares em 2021 para 7,391 biliões de dólares em 2025, correspondendo a 23,6% do total das vendas de retalho. Embora esta situação possa variar entre países e regiões, a ausência de infraestruturas adequadas, de socioeconomia e a falta de estratégias governamentais nacionais de ICT criaram uma barreira significativa na adesão e crescimento do comércio eletrónico no que diz respeito aos países em desenvolvimento (Lawrence and Tar, 2010). Não obstante, estes países apresentam as maiores taxas de crescimento do comércio eletrónico nos últimos anos (América Latina registou uma taxa de crescimento de 25% de 2020 para 2021 em comparação com 9,5% da Europa) (Chevalier, 2021; Department, 2021).

Contudo, ainda não foi alcançado o mundo ideal de transações seguras através na Internet, uma vez que algumas questões de privacidade do comprador, ainda não ultrapassadas, têm impedido o desenvolvimento futuro das tecnologias. Com o acesso ao mercado global através da Internet, os compradores ganham uma clara vantagem, tendo a possibilidade de comparar preços entre regiões, descobrir se preços variam em função da fragmentação das encomendas e tomar conhecimento de produtos substitutos. Devido à transparência do mercado, o cliente pode comparar facilmente os serviços de vários web-sites de comércio eletrónico. Neste caso, os concorrentes estão a um clique de distância do cliente. Se os clientes não estiverem satisfeitos com os produtos, preços ou serviços oferecidos por um determinado web-site, são capazes de mudar muito mais facilmente em relação à presença em loja física (Khan, 2016).

Por um lado, do ponto de vista do consumidor, o comércio eletrónico traz diversas vantagens como: a inexistência de filas, a facilidade de comparar preços, não haver a necessidade de deslocação e a inexistência de horários por parte das lojas (Niranjanamurthy et al., 2013).

Por outro lado, segundo Lee et al. (2019), o sucesso de uma plataforma de comércio eletrónico depende significativamente do número de vendedores e da sua qualidade. Fatores como o aumento do número de clientes, possuir um canal adicional de vendas, baixos entraves à entrada e apoio de vendas e logística podem ser considerados como algumas das vantagens para os vendedores que aderirem a um *marketplace* digital (Kawa and Wałęsiak, 2019).

2.1.1 *E-commerce* em Portugal

Entre 2015 e 2021, a população portuguesa com acesso à Internet verificou uma tendência de crescimento, sendo que em 2021 atingiu os 81% (Lone et al., 2021). Em 2020, 4,4 milhões de portugueses efetuaram pelo menos uma compra online, registando um crescimento de 46% em relação ao ano anterior (Pimenta, 2020). Embora em 2021 também tenha registado um grande crescimento, cerca de 26%, Portugal continuou abaixo da média europeia, ficando entre os últimos 10 lugares no que diz respeito à taxa de crescimento do comércio eletrónico dos países europeus (Lone et al., 2021).

2.1.2 *Marketplace* Digital

O comércio eletrónico B2C é um fenómeno empresarial importante que proliferou com o aumento das tecnologias de informação e da Internet, funcionando como uma plataforma multilateral. Este tipo de plataformas específicas reúne dois grupos distintos de clientes interdependentes, operando como intermediários, ao criar valor por conectar estes dois grupos (Eisenmann et al., 2006; Osterwalder et al., 2010). Os *e-marketplaces*, também conhecidos por *marketplaces* digitais, são plataformas multilaterais que combinam clientes e vendedores, onde a propriedade e controlo dos bens é deixada a cargo dos vendedores (Hagiu, 2007). Assim, o serviço principal dos *e-marketplaces* é proporcionar um espaço de mercado central, onde o comércio eletrónico possa ser realizado (Bru, 2002). Estes mercados funcionam como intermediários online (Sarkar et al., 1995).

2.2 Previsão de Vendas

O ambiente empresarial dinâmico e complexo do comércio eletrónico traz grandes desafios à tomada de decisões empresariais (Akter and Wamba, 2016). Como em qualquer negócio, há uma necessidade de prever e, consequentemente, fornecer a quantidade certa de produto para venda aos clientes (Xia and Wong, 2014). Desta forma, a previsão de vendas é útil para a gestão da mão-de-obra, do fluxos de caixa e dos recursos, tais como a otimização da cadeia de abastecimento dos fabricantes. A previsão de vendas traduz-se na antecipação de vendas futuras. Normalmente, os fatores que influenciaram as vendas passadas terão um peso semelhante nas vendas futuras,

fazendo com que a utilização de dados históricos relativos a vendas passadas seja uma boa base para a previsão (Gooijer and Hyndman, 2006).

Existem inúmeros modelos de previsão que podem ser utilizados para prever as vendas futuras, no entanto, a escolha do modelo certo é crucial para se conseguir uma precisão que permita um bom desempenho operacional, bem como a redução dos custos de inventário e transporte, o que por sua vez levará a uma melhoria dos lucros da empresa (Ren et al., 2017). A elevada volatilidade na procura de produtos aumenta a complexidade desta tarefa (Liu et al., 2013). Além disso, o mercado é influenciado por fatores conhecidos como variáveis explicativas, que aumentam a variabilidade das vendas. Algumas destas variáveis não são possíveis de controlar, tais como fatores económicos, situação do mercado, datas de calendário, condições meteorológicas, entre outros (Nelson and Kang, 1984). Em contrapartida, as variáveis intrínsecas à empresa e ao produto podem ser controladas. De forma a lidar com todos os parâmetros que introduzem a não-linearidade nos modelos, o algoritmo de previsão deverá incluí-los como variáveis, em vez de se limitar a considerar a série temporal de dados históricos (Du et al., 2015).

Os métodos de previsão de vendas apropriados dependem em grande parte dos dados disponíveis. A inexistência de dados ou existência de dados apenas não relevantes leva a que métodos de previsão qualitativos devam ser aplicados. No caso de existir informação numérica sobre o passado, então podem ser aplicados métodos de previsão quantitativos. Existe uma vasta gama de métodos de previsão quantitativa, muitas vezes desenvolvidos para aplicações específicas (J. Hyndman and Athanasopoulos, 2014).

2.3 Técnicas Estatísticas de Previsão

2.3.1 Série Temporal

Segundo J. Hyndman and Athanasopoulos (2014), tudo que seja observado sequencialmente ao longo do tempo pode ser considerado uma série temporal. Os métodos mais simples de séries temporais utilizam apenas informação sobre a variável a prever, e não fazem qualquer tentativa de descobrir os fatores que afetam o seu comportamento. Desta forma, extrapolam tendências e padrões sazonais, mas ignoram todas as outras informações, tais como iniciativas de marketing, atividade concorrente, mudanças nas condições económicas, entre outros.

Os modelos de *exponential smoothing* e ARIMA são as duas abordagens mais utilizadas relativamente à previsão de séries temporais, e fornecem abordagens complementares. Enquanto os modelos de *exponential smoothing* são baseados numa descrição da tendência e sazonalidade dos dados, os modelos ARIMA visam descrever as auto-correlações nos dados.

Uma série temporal pode ser decomposta em três fatores principais: o nível, que representa o valor médio da série; a tendência, que reflete a tendência descendente ou ascendente da série; e a sazonalidade, que se refere ao comportamento cíclico de uma série (Wei, 2006).

2.3.2 Vantagens e Desvantagens de Técnicas Estatísticas de Previsão

Os métodos estatísticos anteriormente descritos têm diversas vantagens associadas, tais como a rapidez, mesmo quando se trata de um grande conjunto de dados, e serem de fácil compreensão (Liu et al., 2013). Apesar destes pontos a favor e da sua popularidade mundial, a linearidade que categoriza tais modelos traz algumas limitações quando aplicados a vendas de alta complexidade e aleatoriedade inerentes aos dados históricos disponíveis. Por vezes, as informações sobre as vendas são caracterizadas por fortes padrões não lineares que não são totalmente reconhecidos pelos modelos estatísticos anteriormente referidos. A influência da procura e variações de preços de outros produtos, por exemplo, não são tidas em conta como variáveis e, além disso, esse tipo de modelos apenas aceitam dados quantitativos, que são grandes desvantagens (Loureiro et al., 2018).

Existem muitos outros métodos estatísticos que podem ser utilizados para a previsão de vendas. Contudo, nenhum deles consegue lidar com a não linearidade introduzida pelas vendas de produtos, que juntamente com o avanço das tecnologias informáticas, aumentaram o interesse em modelos de inteligência artificial para a previsão de vendas. O seu sucesso deve-se principalmente à capacidade de aprendizagem por tentativa erro, bem como à melhoria contínua ao longo do tempo (Ren et al., 2015).

2.4 Técnicas Avançadas de Previsão

Dependendo do contexto e dos dados disponíveis, os métodos de inteligência artificial são mais poderosos e versáteis do que os métodos estatísticos, devido à sua boa capacidade de lidar com a alta complexidade e não linearidade dos dados históricos disponíveis.

Os métodos de *machine learning* são computacionalmente mais exigentes do que os métodos estatísticos. Em muitos casos, a explicabilidade dos modelos nos métodos de ML pode não ser totalmente alta. No entanto, em aplicações empresariais com grandes quantidades de dados, as técnicas de ML podem ser mais adequadas devido ao grande número de variáveis associadas aos dados disponíveis (Cerqueira et al., 2019).

2.5 Princípios de *Machine Learning*

Os algoritmos de *machine learning* são cada vez mais incorporados em vários programas e sistemas para apoiar as organizações na tomada de decisões, especialmente na análise preditiva e no reconhecimento de padrões.

Um processo de *machine learning* consiste geralmente numa série de passos que formam um ciclo que é iterado até se obter um resultado desejável. O processo engloba a compreensão do domínio e o conhecimento prévio do problema, a seleção dos dados, a limpeza e pré-processamento dos dados, a aprendizagem dos modelos, a interpretação dos resultados e, finalmente, a consolidação e a aplicação do conhecimento obtido.

Os algoritmos de *machine learning* possuem grande qualidade devido à sua capacidade de aprender a partir dos dados disponíveis. Estes podem ser divididos em três modelos principais: *supervised learning*, *unsupervised learning* e *reinforcement learning*.

Os métodos de *machine learning* mais utilizados são os métodos de aprendizagem supervisionada. Os sistemas de aprendizagem supervisionada baseiam-se na observação de vários exemplos de um vetor x aleatório e o seu valor de resposta y, para depois prever y em função de um vetor x novo. Ou seja, neste método de aprendizagem, cada elemento que sirva como treino deve ser acompanhado pelo seu valor de resposta. A aprendizagem supervisionada está dividida em duas categorias: classificação e regressão. Um problema de classificação está relacionado com a tarefa de classificar uma observação num conjunto discreto de classes, enquanto um problema de regressão está associado à capacidade de prever valores de uma variável contínua para uma dada observação (Jordan and Mitchell, 2015).

A aprendizagem não supervisionada é um tipo de algoritmo que aprende padrões e tendências de similaridade a partir da análise de dados não etiquetados. Implica a observação de diferentes instâncias de um vetor aleatório x com o objetivo de aprender as suas propriedades (Jordan and Mitchell, 2015).

Por fim, o método de aprendizagem de reforço é um exemplo de *machine learning* onde a máquina é treinada para tomar decisões específicas com base no ambiente externo, maximizando um objetivo pré-definido. A máquina é treinada continuamente com base no ambiente a que está exposta e aplica os seus conhecimentos enriquecidos a problemas específicos (Jordan and Mitchell, 2015).

2.5.1 Modelos de Regressão

No contexto desta dissertação, a categoria de regressão foi explorada.

A regressão é um algoritmo de aprendizagem supervisionado utilizado para fazer previsões de valores contínuos. É uma técnica que investiga a relação entre variáveis independentes e uma variável dependente. O modelo aprende a partir de um conjunto de dados com o objetivo de prever um valor a partir de variáveis independentes. A regressão é um elemento chave da modelação preditiva, pelo que pode ser encontrada em diversas aplicações de *machine learning*.

Um problema de aprendizagem supervisionado pode ser estruturado em quatro passos diferentes. O primeiro passo consiste na obtenção e pré-processamento dos dados. Este primeiro passo é de extrema importância porque a quantidade e qualidade dos dados recolhidos têm um grande impacto nos resultados do modelo. Após a recolha dos dados, é necessário conduzir etapas de pré-processamento para corrigir eventuais problemas no conjunto de dados recolhidos. Estas etapas podem ser dividas em quatro fases: a limpeza dos dados, onde o objetivo é remover dados incompletos e dados incorretos do conjunto de dados; a integração dos dados, onde o objetivo é combinar dados vindos de múltiplas fontes num conjunto único; a redução dos dados, onde o processo ajuda na redução do volume de dados e facilita análise; e a transformação dos dados, que consiste na alteração do formato ou estrutura dos dados para que possam ser integrados nos

modelos a aplicar. Ainda nesta fase deve-se proceder à análise exploratória dos dados de forma a aprofundar o conhecimento sobre os mesmos.

O segundo passo trata da seleção do modelo. Este refere-se ao processo de seleção de um modelo final de *machine learning* de entre uma coleção de candidatos. É um procedimento que pode ser utilizado para comparar modelos de diferentes tipos e modelos do mesmo tipo com hiperparâmetros diferentes. Encontrar o melhor algoritmo pode ser algo complexo quando existe um grande leque de modelos possíveis. O objetivo final é escolher um modelo cujo erro de generalização seja o mais baixo. Este erro é resultado do valor esperado da taxa de erro das previsões medido a partir dos dados não observados. O facto da medição deste erro ser executado no conjunto de dados deixados para testar o modelo, torna impraticável a utilização deste erro para selecionar o melhor modelo. Deste modo, a seleção do modelo durante a fase de treino do modelo torna-se imperativa (Murphy, 2012).

Existem duas abordagens principais para a fase de seleção do modelo. Na primeira, o conjunto de dados é dividido em três partes: um conjunto de treino, um conjunto de validação e um conjunto de teste. O conjunto de treino é utilizado para construir os modelos. O conjunto de validação é utilizado para estimar o erro de previsão para a seleção do modelo. Por fim, o conjunto de teste é utilizado para avaliar o erro geral do modelo final escolhido (Raschka, 2018).

A segunda abordagem, amplamente utilizada, introduzida por Geisser (1975), é a validação cruzada. Esta técnica é geralmente preferida, porque permite ao modelo treinar em diferentes conjuntos de treino, dando uma indicação mais precisa do seu desempenho. O primeiro passo na validação cruzada é baralhar o conjunto de dados e dividi-lo em k grupos. Para cada grupo, é considerado um conjunto de validação e as restantes k-1 dobras são os conjuntos de sub-treino utilizados para construir o modelo como representado na Figura 2.1. De seguida, o erro médio é calculado sobre todas as dobras e utilizado para o erro de teste. A seleção do modelo é realizada na montagem do modelo final, uma vez que todos os modelos serão descartados no final desta fase. De outro modo, depois da escolha do modelo mais adequado, será construído um novo modelo final com todos os dados de treino disponíveis e, posteriormente, avaliado no conjunto de dados preservados para teste não utilizados na fase de seleção do modelo (Bishop, 2006).

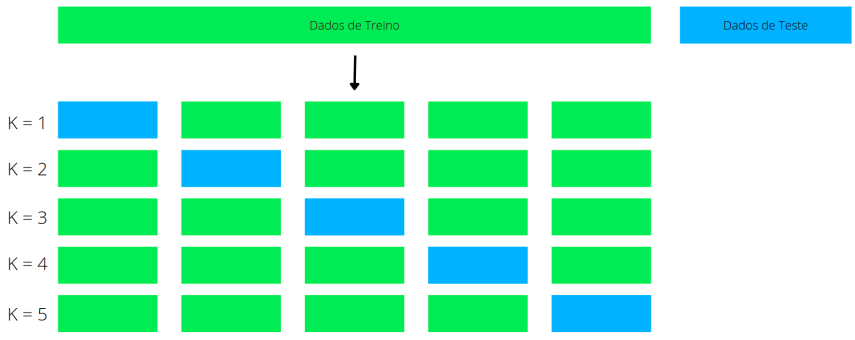

Figura 2.1: Exemplo de 5-*fold cross-validation*

O próximo passo no fluxo de trabalho de um problema de aprendizagem supervisionado é o treino do modelo. Nesta fase, a qualidade dos dados de treino e o algoritmo de treino são ambos elementos críticos. Esta fase pretende criar a melhor representação matemática da relação entre as variáveis independentes e a variável dependente, onde o algoritmo utilizado procura as melhores combinações dos pesos atribuídos às diferentes variáveis de forma a minimizar o erro de previsão.

Após concluída a fase de treino, é necessário avaliar o modelo. O objetivo desta fase é avaliar o desempenho do modelo no mundo real. Desta forma, é importante utilizar novos dados para evitar o *overfitting* do modelo em dados anteriormente utilizados para a fase de treino. A precisão das previsões apenas pode ser determinada considerando o desempenho de um modelo em novos dados que não foram utilizados na montagem do modelo. Uma vez que os dados de teste não são utilizados na determinação das previsões, fornecem uma indicação fiável da qualidade do modelo a prever com base em dados nunca antes vistos (J. Hyndman and Athanasopoulos, 2014).

2.5.2 *Linear Regression*

A *Linear Regression* é um modelo de *machine learning* supervisionado no qual o modelo encontra a relação linear entre as variáveis independente e dependente.

A regressão linear pode ser de dois tipos: simples ou múltipla. A regressão linear simples é onde apenas uma variável independente está presente e o modelo tem de encontrar a relação linear desta variável com a variável dependente. Na regressão linear múltipla há mais do que uma variável independente para o modelo encontrar a relação (Weisberg, 2005).

No contexto deste projeto, foi utilizada a regressão linear múltipla. A equação 2.1 ilustra a sua representação.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \tag{2.1}$$

onde y é o valor previsto, b_0 é o valor que y toma quando todas as variáveis são 0, b_1, b_2 e b_n são os coeficientes de regressão relativos às variáveis independentes e x_1, x_2 e x_n são as variáveis independentes.

2.5.3 *Support Vector Regression*

Support Vector Regression é um algoritmo de *machine learning* supervisionado que é utilizado para prever valores discretos. Este algoritmo segue o mesmo princípio que o *Support Vector Machine*. Num SVM, os dados de treino são representados no espaço separados por categorias. Os novos pontos são adicionados ao espaço através da previsão da categoria onde se enquadram. Relativamente ao SVR, a ideia base é encontrar a linha mais adequada que traduz os dados. Esta linha é considerada a partir do plano que possui o número máximo de pontos (Smola and Schölkopf, 2004).

A vantagem do SVR é a sua capacidade de trabalhar com dados dispersos e gerar modelos com pequena quantidade de dados. Porém, uma grande desvantagem deste algoritmo é o facto de não ser adequado com uma quantidade de dados elevados.

2.5.4 *Decision Tree*

Dada a sua capacidade de resolver problemas complexos e a sua fácil interpretabilidade, os algoritmos de *Decision Trees* são bastante utilizados.

O algoritmo constrói modelos de regressão ou classificação sob a forma de uma estrutura de árvore. As árvores de decisão são geradas com uma abordagem de cima para baixo a partir de um nó de raiz (Yang, 2019). O conjunto de dados é dividido em subconjuntos cada vez mais pequenos, ao mesmo tempo que se desenvolve incrementalmente a árvore de decisão associada. Esta árvore é construída a partir de regras que classificam os dados consoante uma série de perguntas sobre as variáveis independentes associadas às observações. Estas regras são aprendidas sequencialmente, utilizando os dados de treino um de cada vez. No fim, esta árvore é constituída por nós de decisão e nós de folhas (Figura 2.2). Um nó de decisão representa uma questão e possui dois ou mais ramos que dão origem a quantas respostas forem possíveis. Um nó de folha representa a decisão que corresponde ao valor final. Os ramos são os segmentos da árvore que estabelecem a ligação entre os diferentes tipos de nós (Quinlan, 1986).

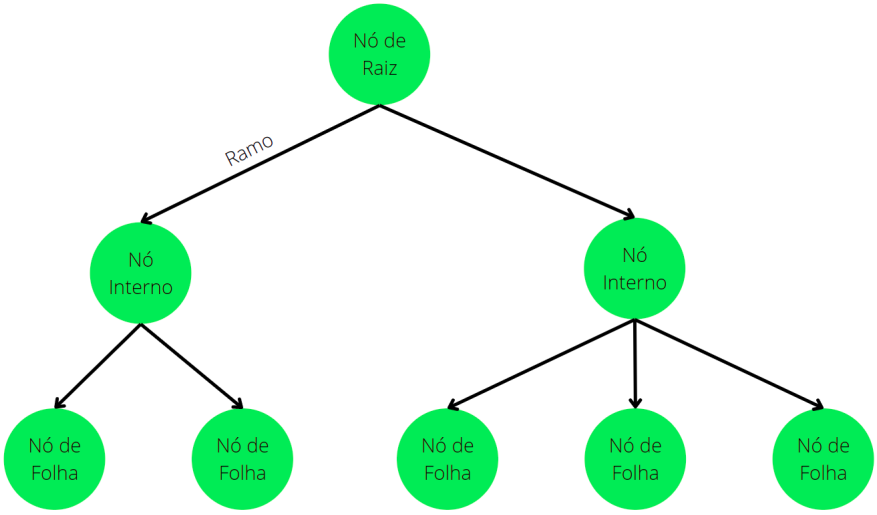

Figura 2.2: Representação de uma *Decision Tree*

2.5.5 *Random Forest*

Random Forest é um algoritmo de *machine learning* supervisionado que é amplamente utilizado em problemas de classificação e regressão. É classificado como uma técnica de *ensemble learning*, ou seja, que combina múltiplos modelos (conhecidos por *week learners*) de forma a obter previsões. Esta técnica pode ser divida em dois tipos de modelos: modelos de *bagging* e modelos de *boosting*. Na construção de um modelo de *ensemble learning* é necessário primeiro estabelecer quais os modelos base a agregar. Na maior parte das vezes é utilizado apenas um algoritmo de forma a possuir *week learners* homogéneos que são treinados de diferentes maneiras. Os modelos de *bagging* consideram *week learners* homogéneos, aprendendo-os, independentemente uns

dos outros, em paralelo e combinando-os seguindo algum tipo de processo determinista. Assim, o algoritmo *Random Forest* é considerado um modelo de *bagging* que constrói múltiplas árvores de decisão, relativas a várias subamostras do conjunto de dados, que trabalham em conjunto para formar um *ensemble*. Cada árvore retorna um valor de resposta. O *Random Forest* baseia-se no conceito fundamental que um grande número de árvores atuando em conjunto superará qualquer modelo individual do algoritmo. Desta forma, calcula a média dos resultados para melhorar a precisão preditiva. No *Random Forest* n observações são retiradas do dataset. De seguida, para cada amostra, são construídas *decision trees* individuais, que, por usa vez, irão gerar um output. O output final gerado é baseado na média dos outputs de todas as árvores de decisão.

2.5.6 *LightGBM*

LightGBM é um modelo de *ensemble learning* de *boosting*. Os algoritmos de boosting visam converter *week learners* em *strong learners*, criando modelos sequenciais de forma adaptativa (modelo base depende dos anteriores) de modo a que o modelo final tenha maior precisão. Entre os diversos algoritmos de *boosting* é ainda considerado um modelo de *gradient boosting* porque utiliza um algortimo de descida de gradiente para minimizar a função de perda ao adicionar novos modelos. O *LightGBM* utiliza algoritmos de árvores de decisão e desenvolve as árvores verticalmente, o que significa que cresce em termos de folhas e não de nível como representado na Figura 2.3.

O facto de conseguir lidar com grandes quantidades de dados e obter resultados rapidamente constituem duas das grandes vantagens deste algoritmo.

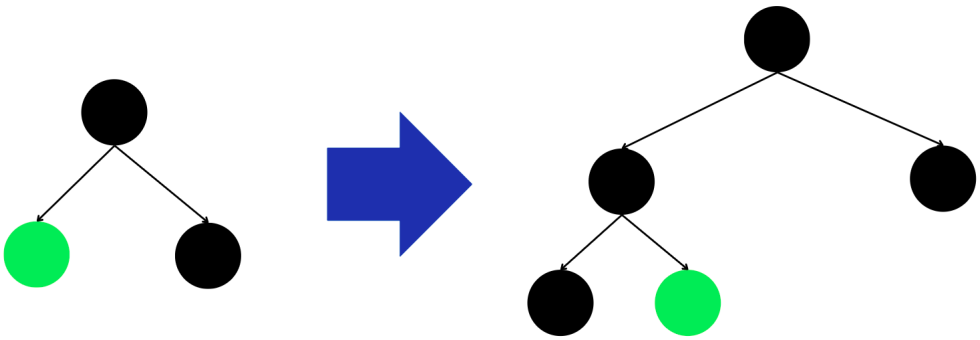

Figura 2.3: Representação da evolução em termos de folhas

2.5.7 *XGBoost*

O *XGBoost* é, tal como o anterior, um algoritmo de *gradient boosting*. Este algoritmo minimiza uma função objetivo que combina uma função de perda convexa (baseada na diferença dos

valores previstos e os valores reais) e um termo de penalização pela complexidade do modelo (as funções das árvores de decisão). O treino do modelo prossegue iterativamente, adicionando novas árvores que preveem os erros de árvores anteriores que são depois combinadas com árvores anteriores para fazer a previsão final. Ao contrário do *LightGBM*, este algoritmo desenvolve a árvore em termos de nível como representado na Figura 2.4.

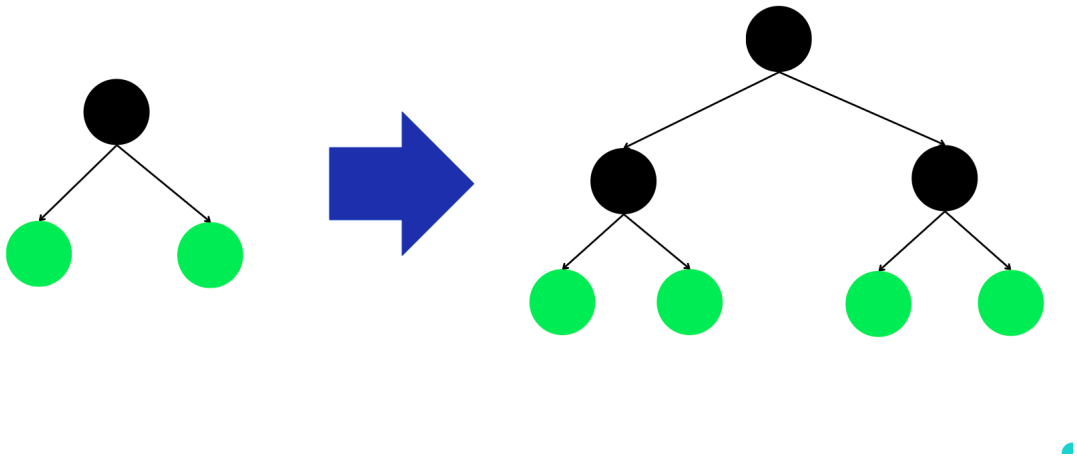

Figura 2.4: Representação da evolução em termos de nível

2.5.8 *K-Nearest Neighbor Regressor*

O algoritmo *K-Nearest Neighbor* é utilizado para prever a classe de uma amostra com base na classe dos seus registos vizinhos (Keller and Gray, 1985).

Este método é conhecido como um algoritmo de aprendizagem preguiçosa, uma vez que se concentra em armazenar todas as instâncias correspondentes ao conjunto de dados de treino num espaço n-dimensional em vez de construir um modelo interno genérico (Song et al., 2017). Logo, cada vez que é necessária uma previsão para uma nova amostra, o algoritmo tem de procurar por todo o conjunto de dados de treino as instâncias k mais próximas para devolver a classe mais semelhante como previsão, resultando num custo computacional elevado. Por outro lado, os benefícios do algoritmo incluem um procedimento de implementação simples e um nível alto de robustez em relação a dados de treino ruidosos.

Este método de *machine learning* pode ser adaptado para um problema de regressão, onde a variável de resposta é prevista através da interpolação local dos alvos associados aos vizinhos mais próximos no conjunto de dados de treino.

2.5.9 Métricas de Avaliação

Um erro de previsão é a diferença entre um valor observado e a sua previsão. O erro de previsão é calculado da seguinte forma:

$$e_t = (y_t - \hat{y}_t) \tag{2.2}$$

onde, e_t é o valor do erro de previsão, y_t é o valor real no instante t e $\hat{y}_t$ é o valor previsto pelo modelo no instante t.

Os erros de previsão podem ser divididos em três categorias diferentes: erros dependentes da escala, erros de percentagem e erros de escala. No seguimento do interesse deste projeto foram apenas explorados os erros dependentes da escala e erros de percentagem.

1. Erros dependentes da escala: Os erros de previsão estão na mesma escala que o dados. As medidas de previsão baseadas apenas em escalas são, portanto, dependentes da escala e não podem ser utilizadas para fazer comparações entre séries que envolvem unidades diferentes. As três medidas dependentes de escala mais utilizadas são baseadas nos erros absolutos ou nos erros quadráticos.

Mean Absolute Error é a média dos erros absolutos, como demonstrado na equação 2.3 onde n é o número de observações.

$$MAE = \frac{1}{n}\sum_{n=0}^{n}|e_t| \tag{2.3}$$

Mean Squared Error é a média dos erros quadráticos, como demonstrado na equação 2.4.

$$MSE = \frac{1}{n}\sum_{n=0}^{n}e_t^2 \tag{2.4}$$

Root Mean Squared Error é a raiz quadrada da média dos erros quadráticos, como demonstrado na equação 2.5

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=0}^{n}e_t^2} \tag{2.5}$$

2. Erros Percentuais: O erro percentual é dado por

$$p_t = \frac{e_t}{y_t} * 100 \tag{2.6}$$

onde y_t é o valor da observação e p_t é o valor do erro percentual. Os erros percentuais têm a vantagem de não conterem unidades, e por isso são frequentemente utilizados para comparar os desempenhos previstos entre conjuntos de dados. O erro percentual mais utilizado é:

Mean Absolute Percentage Error é o erro percentual absoluto médio, como pode ser observado na equação 2.7

$$MAPE = \frac{1}{n}\sum_{n=0}^{n}|p_t| \tag{2.7}$$

Capítulo 3

Descrição do Problema

De forma a descrever o problema em análise, este capítulo está dividido em duas partes. Inicialmente, é apresentada a cadeia de valor da empresa e a estrutura do Dott, descrevendo todos os seus departamentos. Por fim, é apresentada a situação *AS-IS* da empresa na qual é descrita o problema que deu origem a este projeto.

3.1 Cadeia de Valor da Empresa

O Dott é o primeiro grande mercado generalista a operar em Portugal e, de acordo com D'Orey (2021), o CEO do Dott, é o maior shopping de Portugal no que diz respeito a compras online. O seu objetivo é mudar a forma como os portugueses compram e vendem online e ajudá-los a digitalizar as suas marcas, abrindo novos canais de venda aos seus negócios. Apesar de não existirem apenas vendedores portugueses no seu mercado, o Dott pretende ser um mercado de português para português, oferecendo uma gama diversificada de produtos e proporcionando aos seus clientes a melhor experiência na compra online.

Contudo, para atingir os objetivos anteriormente mencionados, o Dott ainda enfrenta desafios que precisam de ser ultrapassados - ganhar a confiança dos compradores e vendedores portugueses. Portanto, o seu objetivo é aumentar a quantidade de portugueses que estão atualmente a comprar online, remodelar o comportamento do consumidor online, uma vez que a maioria dos compradores está a comprar de países estrangeiros como a China, e, finalmente, continuar a aumentar o número de vendedores presentes no mercado, porque estes são ainda uma pequena percentagem das empresas que estão atualmente a vender online. Além de proporcionar acesso ao mercado a todas as marcas, independentemente da sua dimensão ou nacionalidade e assegurar a competitividade e flexibilidade, por exemplo, não cobrando taxas de entrada, permitindo aos vendedores definir preços e gerir o stock, o Dott pretende conceder aos compradores a melhor experiência de compra online a fim de alcançar o sucesso entre os seus clientes.

O modelo de negócio do Dott é constituído por diferentes tipos de pontos chaves. O primeiro ponto está relacionado com os principais parceiros do Dott: os CTT (responsáveis por toda a logística), a Sonae (principal investidor), a Izbeerg (responsável por disponibilizar um serviço

que permite a criação do *marketplace* digital) e a Zendesk (responsável pelo fornecimento de um software que ajuda a empresa a oferecer uma experiência única e a melhorar o serviço ao consumidor).

O segundo ponto é relativo às principais atividades da empresa: o desenvolvimento, preservação e melhoria da plataforma, a estratégia e marketing e o serviço do consumidor. Os recursos fundamentais para a empresa compõem o terceiro ponto chave. Fazem parte destes recursos as parcerias com os vendedores, os recursos humanos e o equipamento tecnológico.

No que concerne a estrutura de custos da empresa, que constitui o quarto ponto chave da cadeia de valor do Dott, esta está dividida em: custos de emprego, custos de marketing e custos operacionais. Adicionalmente, o Dott compromete-se com as seguintes propostas de valor: ser o maior shopping online com produtos diversificados, disponibilizar um serviço de entrega excecional onde o consumidor pode escolher a janela de tempo da entrega, funcionar como um canal de vendas online para pequenas e médias empresas portuguesas, fornecer uma experiência de compra intuitiva e fácil e disponibilizar preços competitivos.

Relativamente ao sexto ponto chave, o Dott destaca as relações com o consumidor, tais como: manter uma relação contínua com o vendedor e providenciar aos clientes um serviço de excelência, pagamentos seguros e garantias nos produtos. Outro dos pontos fulcrais no que diz respeito à cadeia de valor da empresa são os canais, constituídos por: o web-site online e uma aplicação para telemóvel.

O oitavo ponto chave diz respeito à segmentação de indivíduos. Por um lado, a vertente dos compradores classificados em três categorias diferentes: os *Milenials*, *Gen X* e *BabyBoomers*. Por outro, a vertente dos vendedores: as pequenas e medias empresas com produtos inovadores e as marcas internacionais.

Por fim, o último ponto chave do modelo de negócio do Dott são as fontes de receitas que são compostas pela comissão de vendedores e a publicidade na plataforma.

3.2 Estrutura Organizacional

O Dott possui uma estrutura organizada e divisional. Esta estrutura encontra-se dividida em nove equipas: Informação e Tecnologia, Dados e Análise, Produto, Operações e Estratégia, Parcerias, Apoio a Clientes e Vendedores, Recursos Humanos, Finanças e Crescimento.

A equipa de Informação e Tecnologia é responsável pela manutenção e desenvolvimento dos produtos do Dott, ou seja, o mercado online, onde o cliente pode selecionar e comprar diferentes produtos de diferentes lojas, bem como o centro de revenda, uma plataforma online onde os vendedores podem atualizar e manter as suas encomendas. Além disso, é da sua responsabilidade dar apoio tecnológico às outras equipas.

O principal objetivo da equipa de Dados e Análise é recolher, reunir e analisar dados dos diferentes sistemas com os quais a empresa trabalha. Basicamente, os seus membros certificam-se de que as *pipelines* de dados estão a funcionar corretamente e que as equipas de dados têm acesso

ao hardware de que necessitam, preparando os dados para consumo. Além disso, são responsáveis pelo desenvolvimento de algumas funcionalidades como o motor de busca, recomendação de produtos e outros algoritmos baseados em *Machine Learning*.

A função da equipa de Produto é articular o valor comercial ao produto, para que compreendam a intenção por detrás do novo produto ou lançamento do produto, permitindo-lhes gerir e dar prioridade a ideias e características. Trabalham nas especificações técnicas juntamente com outras equipas - Operações, Vendas e Apoio ao Cliente - e asseguram que as equipas têm toda a informação de que necessitam para melhorar e completar o produto.

A equipa de Operações e Estratégia tem como objetivo otimizar o desempenho e as relações com os parceiros. De facto, esta equipa acompanha e analisa os processos atuais, identificando os mais importantes e formulando, analisando e implementando oportunidades de melhoria, trabalhando em estreita colaboração com parceiros, tais como vendedores e logística de terceiros. Além disso, são responsáveis pela tomada de decisões estratégicas para satisfazer as expectativas dos clientes e aumentar os níveis de satisfação.

Parcerias é responsável pela procura de novos parceiros, que partilham a mentalidade do Dott de inovação e melhoria contínua. estes parceiros integrarão o Dott através de APIs, por exemplo, para dar a ambos a oportunidade de crescerem enquanto trabalham em conjunto.

Para manter o ritmo de crescimento do Dott, a empresa está em constante procura de pessoas talentosas que partilhem o seu espírito, ou seja, permitindo a cada empresa alcançar coisas maiores através do comércio eletrónico. A equipa de recursos humanos é também responsável por assegurar que cada trabalhador esteja satisfeito e disponha dos recursos necessários para contribuir para a visão e crescimento da empresa.

A equipa financeira é responsável por acompanhar todas as finanças da empresa e assegurar que todas as transações seguem determinados procedimentos. Além disso, como parte da estratégia da empresa, é também da sua responsabilidade estabelecer indicadores chave de desempenho para ter uma perceção melhor dos custos esperados para o futuro.

A equipa de Apoio ao Cliente e ao Vendedor é responsável pela gestão da comunicação com clientes, vendedores e parceiros logísticos, concentrando-se em garantir a melhor experiência de compra. Esta equipa é a primeira linha de resposta do Dott, visando assegurar níveis de serviço de excelência a todos os canais de comunicação disponíveis desde a pré-venda até aos momentos pós-venda.

A equipa de Crescimento encontra-se dividida em dois lados. O lado do vendedor é responsável pela aquisição de novas marcas competitivas para o mercado e pela sua incorporação para garantir que sabem operar no software do Dott e seguir os processos estabelecidos e ajudá-los com quaisquer dificuldades que enfrentem. A parte do comprador é responsável por tudo relacionado com a aquisição e retenção de clientes Dott, desde o planeamento de uma campanha à definição da estratégia para fazer com que um cliente regresse ao mercado no futuro para realizar outra compra.

3.3 Situação Atual e Problema

Dada a dimensão atingida pelo Dott ao fim de dois anos e meio torna-se imperativo que todo o planeamento e departamentos estejam ao mais alto nível de forma a contribuir para o sucesso e melhoria contínua da empresa. No cenário atual, não existe qualquer método de previsão futura de vendas. Este facto aliado ao fator de dimensão do Dott, ou seja, sendo um *marketplace* com mais de 200 mil utilizadores, mais de 2 mil vendedores e cerca de seis milhões de produtos diferentes, constitui uma desvantagem competitiva em relação aos restantes *marketplaces* digitais.

A previsão de vendas é uma parte essencial das operações de venda para qualquer empresa. Sem uma ideia geral de como poderão ser as vendas futuras é difícil alocar os recursos da melhor forma. O objetivo da construção de um algoritmo de previsão de vendas é fornecer informações que sirvam para tomar decisões comerciais intencionais e com impacto positivo. A previsão de vendas permite identificar potencias problemas antecipadamente e tomar medidas de forma a mitigá-los. A abordagem dos problemas futuros no presente terá um impacto significativo na forma como são tratados.

Relativamente ao Dott, o problema atual advém da variabilidade e imprevisibilidade da quantidade de vendas semanais acentuada pela pandemia. O Dott pretende o desenvolvimento de um modelo de previsão de vendas com objetivo de funcionar como uma ferramenta de ajuda para diversas áreas, tanto na tomada de decisão como na prevenção de possíveis problemas. Como por exemplo, a ausência de stock suficiente para satisfazer as vendas de uma semana resultará na perda de receitas e na insatisfação de clientes que serão perdidos. Neste caso, o conhecimento da quantidade a ser vendida de um produto poderá ajudar a mitigar este problema se o vendedor for notificado atempadamente. Por outro lado, o modelo pode funcionar como uma ferramenta de suporte na decisão dos produtos a destacar na *newsletter* Dott e a promover em campanhas de marketing.

Capítulo 4

Metodologia

Neste capítulo, é definida a metodologia utilizada para construir o modelo de *machine learning* de previsão de forma a resolver o problema mencionado no Capítulo 3.

Inicialmente, é concedida uma descrição inicial do conjunto de dados recolhido, e são detalhadas as etapas de pré-processamento necessárias para preparar os dados para aplicar os modelos. De seguida, os algoritmos são apresentados com mais detalhe, tendo por base uma forma padrão inicial para os diferentes modelos específicos. Além disso, este capítulo explica ainda os processos de avaliação dos modelos. Finalmente, esta secção termina com uma visão geral dos diferentes cenários a serem comparados relativamente ao seu desempenho no que diz respeito à previsão.

4.1 Pré-Processamento de Dados

4.1.1 Caracterização dos Dados

Antes de se proceder ao pré-processamento dos dados, é importante ter um conhecimento geral dos mesmos. Os dados utilizados para este projeto foram obtidos a partir de uma *query* de SQL que extraiu os dados desejados da base de dados do Dott (BigQuery). O conjunto de dados extraído continha 5.824.377 instâncias e 114 variáveis, que reflete a informação de vendas dos produtos presentes no *marketplace* do Dott desde 28 de Junho de 2021 até 31 de Dezembro de 2021.

As variáveis advêm do conjunto de dados retirado a partir de fontes internas do Dott que correspondem a dados históricos relativos a cada produto. As variáveis ano de venda e número da semana de venda foram classificadas como categóricas ordinais, enquanto a variável de identificação do produto foi classificada como categórica nominal. Todas as restantes variáveis, excluindo a quantidade de vendas de cada produto, foram consideradas como quantitativas contínuas. Por fim, a variável em estudo, número de vendas, foi classificada como quantitativa discreta.

O conjunto de dados disponível inclui as seguintes variáveis:

year_nr: corresponde ao ano da instância observada.

week_nr: corresponde ao número da semana da instância observada.

product_variation_id: corresponde ao código de identificação de cada produto.

A descrição das restantes variáveis pode ser consultada na Tabela A.1 e na Tabela A.2 do Anexo A.

A variável **nr_sales** é a variável dependente e corresponde ao número de vendas. Depois de uma análise superficial, constatou-se que o conjunto de dados inicial possuía um total de 305.598 produtos diferentes e que 294.856 correspondiam a um total de vendas nulas no conjunto de todas as semanas como pode ser observado na Figura 4.1. Estes produtos, que correspondem a 95,58% (5.566.703 observações) do conjunto de dados original, não foram incluídos nos algoritmos aplicados, visto que traria um custo computacional muito superior e iria enviesar o modelo de previsão para vendas nulas.

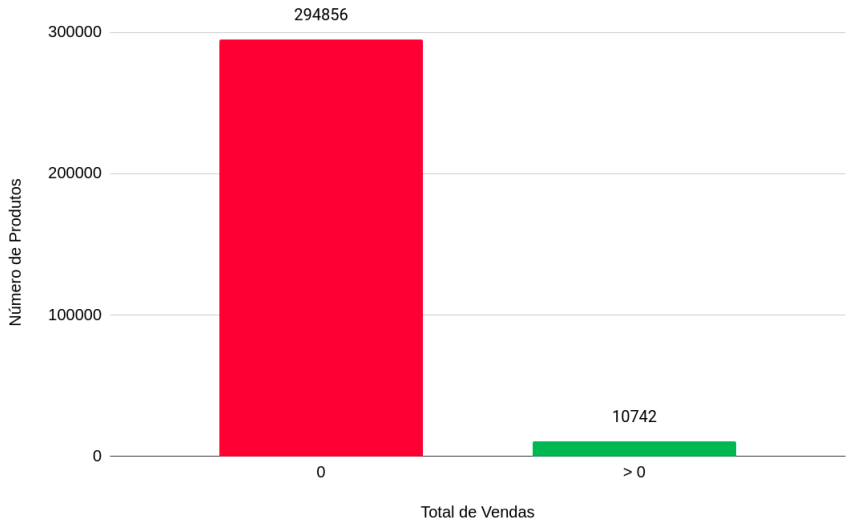

Figura 4.1: Frequência do total de vendas de cada produto

Relativamente aos 10.742 produtos, que refletem as vendas da empresa no período dos dados obtidos (semana 26 à semana 52), estes correspondem a 4,42% dos dados originais, reduzindo o conjunto de dados para 257.674 observações.

4.1.2 Preparação dos Dados

Limpeza de dados

O conjunto de dados continha produtos que na primeira semana (26) não tinham observações, uma vez que ainda não existiam no mercado do Dott, bem como produtos não tinham instâncias relativas à última semana (52). Estes últimos correspondiam a produtos que deixaram de ser comercializados no mercado do Dott e, por isso, não fazia sentido serem alvos de análise. Assim, foram excluídos 1.773 produtos do conjunto de dados, explicando assim a eliminação de 27.780 instâncias (10,78%). Relativamente aos produtos que não eram comercializados na primeira semana do conjunto de dados, permaneceram no modelo visto que continuam a ser vendidos, justificando a previsão das suas vendas.

Adicionalmente, o conjunto de dados apresentava valores negativos em algumas instâncias relativas às variáveis relacionadas com descontos. Estas observações foram consideradas como

valores em falta visto que refletiam erros de cálculo. A percentagem de *missing values* referente a cada variável ilustrada da Figura 4.2 demonstra a baixa existência destas ocorrências. Assim, e devido à existência de observações com mais do que um valor em falta, estas instâncias foram eliminadas. De notar que foi necessário eliminar todas as observações correspondentes a produtos que continham *missing values*, uma vez que não faria sentido ter observações descontinuadas no tempo, resultando na eliminação de 12.254 observações (6,32%).

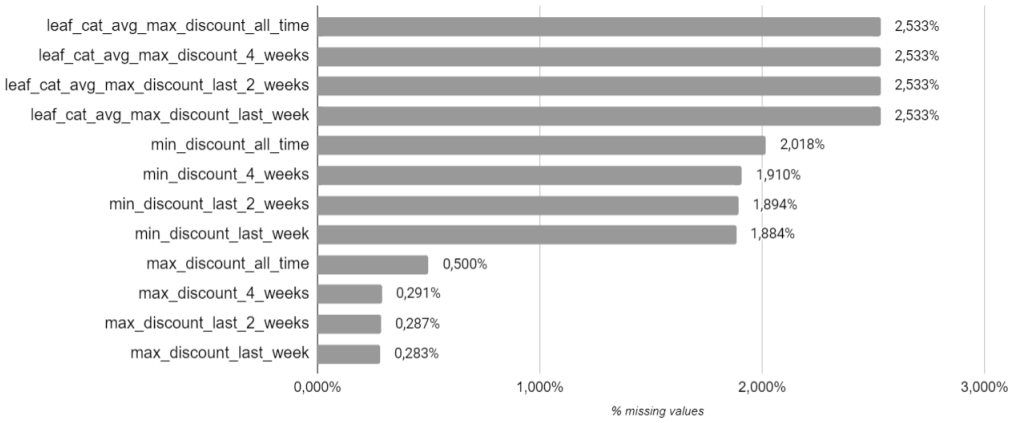

Figura 4.2: Frequência de *missing values* por variável

4.2 Análise Exploratória

A análise exploratória de dados é uma forma de compreender os dados antes de qualquer modelação ser posta em prática e de recolher o máximo de conhecimento sobre os dados. Esta análise refere-se ao processo crítico de realizar investigações iniciais sobre dados para descobrir padrões, detetar anomalias e verificar hipóteses a partir de testes estatísticos e representações gráficas.

4.2.1 Frequência dos Dados

Após uma análise prévia, foi verificado que nas 215.359 instâncias restantes, existem apenas 70 valores de vendas únicos e que 35 desses valores ocorrem apenas uma única vez. As observações com vendas nulas dominam o conjunto de dados correspondendo a 199.448 instâncias (92,61%) como pode ser observado na Tabela 4.1. É de notar que a não eliminação destes dados pode diminuir a robustez do modelo e a sua capacidade de previsão futura em produtos com quantidades de vendas significativas.

Isto representa um possível problema para o modelo de *machine learning*, visto que, para além do modelo ficar enviesado para a previsão de vendas nulas, também todas as vendas que apenas são observadas uma vez terão de ser incluídas ou no conjunto de treino ou no conjunto de teste. Assim, verifica-se um problema, no qual o modelo terá dificuldade em prever quantidades

Tabela 4.1: Frequência de vendas semanais

Quantidade Vendida	Frequência
0	199.448
1	10.842
2	2.595
>2 & <10	2.122
>9 & <100	332
>99 & <1.000	20

vendidas se estas observações forem incluídas no conjunto para teste. Será, ainda, difícil analisar o desempenho do modelo para prever observações únicas que forem incluídas no conjunto de treino.

Além disso, o conjunto de dados tinha dados relativos apenas ao ano de 2021, portanto a variável referente ao ano acabou por ser retirada. A mesma lógica foi aplicada à variável relativa às semanas, que apenas possuía valores entre 26 e 52 o que fez com que essa variável perdesse importância para o modelo visto que não era possível retirar o fator da sazonalidade devido à ausência de dados das mesmas semanas em anos diferentes.

Por fim, a variável que identifica o produto acabou por ser excluída do modelo visto que existiam cerca de 9 mil produtos diferentes, o que implicaria proceder ao método de *one hot encode* desta variável e adicionar cerca de 9 mil colunas ao conjunto de dados, o que não foi possível devido às restrições tecnológicas e ao custo computacional que isso adicionava aos modelos. Além disso, foi considerada a hipótese de adicionar a categoria de cada produto, porém as categorias dos produtos do *marketplace* Dott estavam mal organizadas e, por isso, não adicionavam informação que fosse relevante.

4.2.2 Dados não-balanceados

Um conjunto de dados que possui uma distribuição desigual da variável de resposta reflete um problema de dados não-balanceados. A presença de bastantes valores de um determinado intervalo pode subestimar intervalos iguais com muito menos valores. A generalidade dos algoritmos de ML é construída com base em problemas que assumem distribuições iguais dos dados.

Através da análise da Tabela 4.1 foi facilmente verificado que as vendas nulas dominam o conjunto de dados, e que o intervalo de valores entre 0 e 9 vendas semanais possui bastantes mais observações do que qualquer outro intervalo. Portanto, foi concluído que a distribuição do número de vendas semanais possui grande assimetria à direita. Além disso, a média de vendas semanais é cerca de 0,167, o que pode levar ao problema em que o modelo prevê melhor os valores em torno da média, mas é bastante ineficiente para prever os valores menos observados.

No entanto, estas observações não foram eliminadas, visto que esta assimetria no conjunto de dados corresponde a um retrato da realidade da distribuição de vendas do *marketplace* Dott.

4.3 Transformação de variáveis

A codificação de dados e *scaling* das variáveis são os últimos passos necessários no que diz respeito ao pré-processamento.

4.3.1 Codificação de Dados

Por um lado, os dados numéricos envolvem variáveis que são compostas por números inteiros ou decimais. Por outro lado, os dados categóricos consistem em variáveis que podem conter valores nominais ou ordinais. Uma variável ordinal tem uma ordem inerente às suas categorias, enquanto que as variáveis nominais não. A maior parte dos modelos de ML exige que as variáveis de entrada sejam numéricas, o que significa que a existência de qualquer tipo de variável categórica tem de ser codificadas antes de serem usadas. Neste projeto existem três variáveis categóricas, porém não foi necessário qualquer tipo de codificação, visto que foi optado por deixar estas três variáveis de fora do modelo pelas razões acima referidas.

4.3.2 *Feature Scaling*

Feature Scaling é uma das etapas mais importantes do pré-processamento de dados. Tem como finalidade evitar que o algoritmo utilizado fique enviesado para as variáveis com maior ordem de grandeza.

Existem diferentes técnicas de *Feature Scaling*, tais como a padronização (*Z-Score Normalization*) e a normalização. Neste projeto, foi utilizado o método de padronização dos dados, também conhecido por normalização *Z-Score*, devido à sua robustez em relação a *outliers*.

A Equação 4.1, onde z representa o valor após a normalização, x representa o valor observado, μ representa a média das observações da variável e σ representa o desvio padrão da variável, demonstra que a padronização utiliza a média e o desvio padrão para calcular o novo valor. Esta padronização tem como objetivo colocar as variáveis com média igual a 0 e um desvio padrão a igual a 1.

$$z = \frac{(x - \mu)}{\sigma} \tag{4.1}$$

O conjunto de dados para teste desempenha o papel de dados nunca antes vistos, pelo que não é suposto serem acessíveis durante a fase de treino. A utilização de qualquer informação proveniente do conjunto de teste antes ou durante o treino do modelo revela um potencial enviesamento na avaliação do seu desempenho. Por este motivo, a padronização dos dados foi aplicada após a separação dos conjuntos de treino e teste, utilizando apenas o conjunto de treino.

4.4 Importância e Seleção das Variáveis

A importância das variáveis refere-se a técnicas que atribuem uma pontuação às variáveis independentes com base na sua utilidade na previsão das vendas. A seleção das variáveis é o

processo que seleciona as variáveis que mais contribuem para a variável dependente. Logo, a seleção das variáveis é feita através do método da importância das variáveis. Esta técnica revela diversas vantagens na construção de modelos de *machine learning*, tais como: reduzir o *overfitting*, melhorar o desempenho do modelo e diminuir o tempo de treino.

Existem 3 técnicas de seleção de variáveis: *filter*, *wrapper* e *embedded*. A técnica *embedded*, que aprende quais as variáveis que contribuem melhor para o desempenho do modelo enquanto este está a ser criado, foi aplicada neste projeto. Mais especificamente, o método de *XGBoost* foi utilizado. Neste método, a importância fornece uma pontuação que indica o quão útil cada variável foi na construção das *boosted decision trees* dentro do modelo.

A pontuação é atribuída explicitamente para cada atributo no conjunto de dados, permitindo que os atributos sejam classificados e comparados entre si. Esta importância é calculada para uma única árvore de decisão baseada na quantidade que cada ponto de divisão das variáveis melhora a medida de desempenho do modelo, ponderada pelo número de observações pelas quais o nó é responsável. As importâncias das variáveis são então calculadas a partir da média do seu valor em todas as árvores de decisão dentro do modelo. Por fim, a soma total das importâncias de todas as variáveis é, obrigatoriamente, um.

Assim, com intuito de diminuir a dimensão do conjunto de dados, foram construídos dois conjuntos de dados para, posteriormente, servirem como diferentes cenários para comparação do desempenho dos modelos consoante o seu input. A primeira seleção das variáveis corresponde a uma explicação de 80% da variável dependente, reduzindo o conjunto de dados de 110 variáveis para 9. No segundo caso, foram escolhidas as variáveis que explicavam 100% das vendas semanais, segundo o método *XGBoost*, o que resultou na seleção de 60 variáveis. Na Tabela B.1 e na Tabela B.2 do Anexo B podem ser consultadas as importâncias de cada variável individual ordenadas por ordem decrescente. Para efeitos de referência, o cenário constituído pelo conjunto de dados com apenas 9 variáveis será referido como "Cenário 9 variáveis", e o outro cenário como "Cenário 60 variáveis". As primeiras 9 variáveis constituem o conjunto de dados do "Cenário 9 variáveis"e as primeiras 60 variáveis constituem o conjunto de dados do "Cenário 60 variáveis"

4.5 Métodos de Previsão

4.5.1 Estrutura do Modelo

Após o tratamento de dados, todos os algoritmos, explicados no Capítulo 2, possuem um procedimento geral em comum: recebem as variáveis independentes que constituem o conjunto de dados.

O resultado final do modelo será uma quantidade vendida prevista para cada produto relativa a uma semana.

O grande objetivo de *machine learning* é disponibilizar um modelo que consiga generalizar na presença de dados novos. O conjunto de dados é dividido num conjunto de treino, onde o

algoritmo aprende os valores de vendas já existentes e um conjunto de teste que avalia o desempenho do algoritmo. Neste projeto, o conjunto de dados guardados para treino corresponde a 80% do conjunto de dados inicial, enquanto 20% do conjunto de dados inicial foram reservados para a fase de teste do modelo. Esta partição foi executada consoante a ordem temporal, logo todos os dados foram ordenados por ordem crescente de semanas, fazendo com que os dados de teste correspondessem às semanas seguintes dos dados de treino. Desta forma, a variável relativa às semanas serviu apenas para ordenar os dados por ordem temporal, visto que não foi considerada como variável para ser imputada no modelo como já tinha sido referido.

4.5.2 Seleção do Modelo

Antes da fase de treino e avaliação do modelo final, é necessário escolher o modelo que melhor se adapta ao problema. Esta fase consiste na escolha do modelo final entre os vários candidatos considerados para o problema de previsão. Esta etapa pode ser usada para comparar diferentes tipos de modelos, bem como modelos do mesmo tipo com diversos hiperparâmetros. Os hiperparâmetros correspondem aos parâmetros relativos a cada algoritmo e são definidos antes da adaptação do modelo aos dados de treino.

Time series cross-validation

Para a fase de seleção do modelo, a técnica *cross-validation* foi escolhida em vez de uma simples divisão em conjunto de treino e conjunto de validação. A validação cruzada é utilizada para ter uma estimativa mais credível da capacidade de um modelo de ML para fazer previsões sobre dados novos. Esta técnica é amplamente preferida, pois permite ao modelo treinar em diferentes conjuntos de treino, dando uma indicação mais precisa do desempenho do modelo. A simples divisão em conjunto de treino e conjunto de validação está dependente de como os dados são divididos, sendo menos representativo da realidade.

Tanto a seleção do algoritmo como a otimização dos seus hiperparâmetros são aspetos importantes da fase de seleção do modelo. A seleção de algoritmos é utilizada para avaliar o desempenho de um conjunto de algoritmos de ML e o *tunning* dos hiperparâmetros é utilizado para encontrar o melhor conjunto de hiperparâmetros para esse algoritmo ML. Uma versão mais sofisticada dos conjuntos de treino/validação é a validação cruzada de séries temporais. Neste procedimento, há uma série de conjuntos de validação, cada um deles consistindo em observações sequenciais. O conjunto de treino correspondente consiste apenas em observações que ocorreram antes da observação que forma o conjunto de validação. Assim, nenhuma observação futura pode ser utilizada na construção da previsão. Uma vez que não é possível obter uma previsão fiável com base num pequeno conjunto de treino, as primeiras observações não foram consideradas como conjuntos de validação. A Figura 4.3 ilustra o método utilizado para a fase da seleção do modelo, onde as observações verdes representam os conjuntos de treino e as observações vermelhas formam os conjuntos de validação.

No contexto deste projeto, o primeiro conjunto de validação foi inicializado na semana 43. No final, é relatada uma precisão relativa a cada uma das iterações realizadas. A precisão da previsão do modelo é calculada através do cálculo da média sobre os conjuntos de validação.

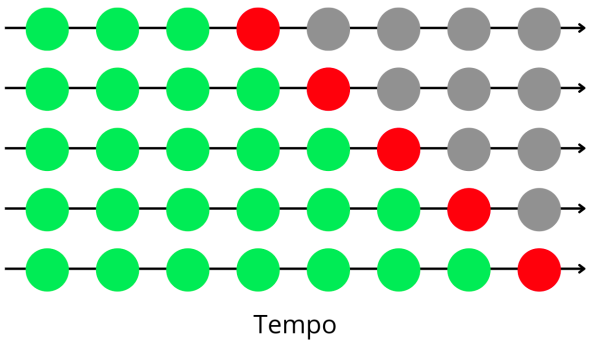

Figura 4.3: *Time series cross-validation*

O conjunto de teste foi utilizado nesta fase, de modo a assegurar que o modelo continuava desconhecido dos dados utilizados numa fase posterior para reportar o desempenho do modelo final.

É importante referir que esta fase não pretende aplicar o modelo final, uma vez que todos os modelos são descartados no final. Em vez disso, após a escolha do modelo mais adequado, um novo modelo final será aplicado a todos os dados de treino disponíveis e subsequentemente avaliado no conjunto de teste preservado.

Os algoritmos de *machine learning* testados foram *K-Nearest Neighbor* (KNN), *Linear Regression*, *Decision Tree*, *Random Forest*, *LightGBM* e *XGBoost*. O algoritmo *Support Vector Regression* não foi utilizado devido à dificuldade de ser usado com uma quantidade de dados elevada. Estes modelos foram escolhidos devido ao seu grande poder no que diz respeito a problemas de regressão.

Finalmente, apesar do desempenho do modelo ter sido o critério prioritário relativamente à escolha do modelo, foram ainda considerados outros dois fatores importantes tais como o custo computacional e o grau de interpretabilidade. Este último fator, revela-se fulcral na decisão do modelo, pois quão mais complexo for o modelo, melhor ele consegue captar os padrões nos dados. No entanto, à medida que a complexidade do modelo aumenta, torna-se cada vez mais difícil compreender todas as variações e determinar o peso que uma variável tinha na variável de resposta. Assim, a atribuição de maior importância aos fatores fica totalmente dependente do objetivo do projeto.

No contexto deste projeto, os fatores considerados para a seleção do algoritmo foram os seguintes, em ordem decrescente de importância: desempenho do modelo, custo computacional do modelo e interpretabilidade do modelo.

Embora a interpretabilidade do modelo tenha sido considerado o último critério no processo de tomada de decisão, a sua importância deve ser realçada. Em certos algoritmos de ML é crucial ser capaz de explicar porque é que um modelo executa da forma como o faz.

4.6 Método de avaliação

As abordagens acima referidas foram avaliadas a partir do mesmo método, calculando uma métrica de erro relativamente aos valores previstos pelo modelo relativamente aos dados reservados para teste. Assim, foi preservado o principal objetivo da avaliação que é avaliar o comportamento de um modelo quando colocado num ambiente desconhecido.

As métricas de avaliação referem-se a métricas numéricas que permitem comparar o desempenho do modelo final escolhido. Neste projeto foram consideradas 2 métricas de avaliação: MAE, RMSE. A métrica de avaliação percentual MAPE acabou por não ser utilizado devido aos seus problemas no que diz respeito a valores observados nulos.

4.7 Definição de Cenários

Foram criados diferentes cenários com objetivo de determinar a combinação ótima de fatores que levam ao modelo mais adequado para a previsão de vendas.

O principal objetivo era comparar as duas abordagens de treino propostas e validar qual das hipóteses resultava num desempenho mais elevado. As duas metodologias possíveis de treino dos algoritmos corresponderam à diferença no conjunto de dados imputados. No primeiro caso, foram escolhidas apenas 9 variáveis independentes. No segundo caso, foram escolhidas 60 variáveis independentes. Para cada abordagem, foram avaliados diferentes modelos de *machine learning*.

4.8 Ferramentas de programação

Este projeto exigiu duas linguagens de programação diferentes, cada uma com um objetivo distinto e específico. Inicialmente, foi necessário utilizar linguagem SQL de forma a obter os dados necessários para o algoritmo de previsão de vendas. Estes dados foram retirados da base de dados do Dott (BigQuery).

No que diz respeito ao desenvolvimento dos modelos de previsão de vendas, foi utilizada a linguagem de programação *Python*. Além de ser uma das linguagens de programação mais utilizadas em todo o mundo devido à facilidade da sua utilização e ao grande leque de bibliotecas de pacotes existentes, a escolha desta linguagem deveu-se, também, ao facto de assegurar a consistência no resto dos algoritmos do Dott.

Os algoritmos construídos tiveram por base 3 principais bibliotecas fornecidas em Python. Pandas e NumPy foram duas dessas bibliotecas, sendo a primeira uma biblioteca que se demonstrou essencial para a manipulação de dados, principalmente pelo facto de fornecer estruturas de dados e ferramentas de alto desempenho e fáceis de usar que facilitam a codificação. Em relação ao NumPy, foi uma biblioteca essencial, que permitiu o manuseamento de vários objetos de matrizes n dimensionais, funções avançadas e álgebra linear.

Por fim, o pacote scikit-learn foi essencial para o desenvolvimento deste projeto. Esta biblioteca simples e consistente, fornece implementações eficientes de uma grande variedade de algoritmos comuns, essenciais à construção dos modelos de *machine learning* deste projeto.

Todo o projeto foi desenvolvido numa máquina com CPU Intel® Core™ i7-8565U @ 1.80GHz processador com 8GB de memória RAM.

Capítulo 5

Resultados

Os resultados obtidos da aplicação dos algoritmos e metodologia definidos nos capítulos anteriores são apresentados neste capítulo onde são avaliados de forma a determinar o modelo mais adequado para prever a quantidade de vendas semanais.

5.1 Análise dos Cenários

Os resultados de desempenho relativos às diferentes iterações da aplicação da técnica *time series cross-validation* dos diversos algoritmos utilizados nos 2 cenários podem ser consultados no Anexo C.

5.1.1 Cenário 9 Variáveis

Neste cenário foram utilizadas as variáveis independentes que correspondiam a uma explicação de um total de 80% da variável dependente. Desta forma, neste cenário, apenas 9 variáveis independentes foram inseridas nos diferentes algoritmos.

O modelo com melhor desempenho foi escolhido na fase da seleção do modelo. A Tabela 5.1 resume o desempenho de cada modelo, obtido a partir da média das diferentes iterações efetuadas para cada algoritmo.

Tabela 5.1: Resultados das métricas utilizadas para a fase de seleção dos modelos com base no conjunto de dados com 9 variáveis

Algoritmo	MAE	RMSE
K-Nearest Neighbor	0.225	1.990
Linear Regression	0.286	2.116
Decision Tree	0.283	2.682
Random Forest	0.231	**1.957**
LightGBM	0.187	2.013
XGBoost	**0.184**	2.152

A partir dos resultados ilustrados na Tabela C.1 do Anexo C foi constatado que todos os algoritmos apresentam piores desempenhos na primeira e última iteração. Estas iterações correspondem

às semanas 37 e 47. Esta diminuição no desempenho de todos os modelos pode ser explicada por uma maior variabilidade na quantidade de vendas nestas semanas.

O algoritmo *Support Vector Regression* não foi testado devido aos seus custos computacionais extremamente elevados. Este modelo não se adequa a conjuntos de dados muito extensos e com grande dimensionalidade.

Para a fase de seleção de modelos, todos os algoritmos de ML foram avaliados usando as métricas de desempenho referidas na metodologia. Com base nos resultados da Tabela 5.1, foi verificado que o modelo *XGBoost* possui melhor desempenho no que diz respeito à métrica do erro médio absoluto. Esta é uma métrica que, como o próprio nome indica, calcula a média dos erros totais, logo dá uma indicação média do erro das previsões. Assim, esta métrica poderá ser enganadora no que diz respeito a valores de vendas superiores mas que se verifiquem menos vezes. Relembrando que o conjunto de dados possui cerca de 95% de valores nulos, o MAE pode estar muito enviesado pela quantidade de vendas nulas e não traduzir a realidade no que diz respeito a quantidades de vendas muito superiores a 0.

Relativamente à métrica RMSE, o modelo que apresentou melhor desempenho foi o *Random Forest*. Esta métrica foi preferida à MAE, uma vez que realça, através da sua fórmula de cálculo, previsões que fiquem mais distantes do seu valor real, e, por isso, traduz de melhor forma o verdadeiro desempenho do modelo.

Por fim, e seguindo o contexto real deste projeto, foi necessário perceber qual dos modelos generalizava melhor os produtos mais vendidos, visto que possuem uma importância muito superior relativamente a produtos que no total do conjunto de dados apenas venderam 1 unidade. Desta forma, foram escolhidos os 25 produtos mais vendidos como exemplo. Assim, foi estabelecida uma forma de avaliação do desempenho dos modelos em que as métricas de avaliação foram calculadas apenas para as previsões que pertenciam aos 25 produtos com um total de vendas mais elevado, dando assim preferência a produtos com bastantes vendas e não com muitos valores nulos. Estes resultados estão representados na Tabela 5.2 e são resultado da média das várias iterações presentes na Tabela C.2 do Anexo C

Tabela 5.2: Resultados das métricas utilizadas para a fase de seleção dos modelos com base no conjunto de dados com 9 variáveis apenas para os 25 produtos mais comercializados

Algoritmos	**MAE**	**RMSE**
K-Nearest Neighbor	11.152	30.124
Linear Regression	12.074	31.873
Decision Tree	15.092	44.028
Random Forest	**10.723**	**27.507**
LightGBM	11.495	30.656
XGBoost	11.918	33.594

Neste caso, o modelo *Random Forest* foi o melhor modelo em ambas as métricas de avaliação.

Após a seleção do melhor modelo de ML, o desempenho global do modelo foi avaliado a partir dos dados preservados para teste. Neste caso o algoritmo *Random Forest* foi escolhido por ser o modelo com melhor desempenho na fase de seleção.

Na fase de avaliação do modelo final, os resultados obtidos presentes na Tabela 5.3 demonstraram pobres resultados de previsão. Os resultados das duas métricas que poderiam à primeira vista parecer resultados prometedores relativamente ao modelo aplicado, depois quando comparados com a média de vendas (0,206) do conjunto de dados previstos, resultaram na dificuldade em perceber a qualidade do modelo, devido à grande quantidade de vendas nulas que enviesou a média de vendas e as métricas de avaliação.

Deste modo, foram medidas, novamente, as métricas para os 25 produtos com total de vendas superior de forma a perceber a verdadeira qualidade do modelo. Com o resultado do erro médio absoluto igual a 23,252, foi constatado que o modelo não possui grande capacidade preditiva visto que a média de vendas semanais destes 25 produtos é de 19,181, ou seja, o erro é superior à própria média em cerca de 4 unidades, o que significa que por cada observação o modelo pode falhar em mais do que a média de vendas.

Tabela 5.3: Resultados da avaliação do modelo final do cenário 9 variáveis

Produtos	**MAE**	**RMSE**
Todos	0.307	3.835
Top 25	23.252	69.424

5.1.2 Cenário 60 Variáveis

Neste cenário foram utilizadas as variáveis independentes que correspondiam a uma explicação de um total de 100% da variável dependente. Desta forma, neste cenário, foram consideradas 60 variáveis do conjunto de dados.

À semelhança do que foi feito na abordagem anterior, durante a fase de seleção do modelo foi selecionado o modelo com melhor desempenho. A Tabela 5.4 resume o desempenho de cada modelo, obtido a partir da média das diferentes iterações efetuadas para cada algoritmo.

Tabela 5.4: Resultados das métricas utilizadas para a fase de seleção dos modelos com base no conjunto de dados com 60 variáveis

Algoritmos	**MAE**	**RMSE**
K-Nearest Neighbor	0.210	1.929
Linear Regression	0.274	2.170
Decision Tree	0.312	3.032
Random Forest	0.214	1.988
LightGBM	0.188	**1.913**
XGBoost	**0.187**	2.402

A partir dos resultados ilustrados na Tabela C.3 do Anexo C foi constatado que, tal como no cenário anterior, todos os algoritmos apresentam desempenhos piores na primeira e última iteração.

Relativamente à métrica RMSE, o modelo que apresentou melhor desempenho foi o *LightGBM*. Por fim, foi estabelecida, mais uma vez, a forma de avaliação relativa aos 25 produtos mais comercializados. Os resultados estão representados na Tabela 5.5 e são resultado da média das várias iterações presentes na Tabela C.4 do Anexo C

Tabela 5.5: Resultados das métricas utilizadas para a fase de seleção dos modelos com base no conjunto de dados com 9 variáveis apenas para os 25 produtos mais comercializados

Algoritmos	**MAE**	**RMSE**
K-Nearest Neighbor	10.973	29.629
Linear Regression	12.249	32.887
Decision Tree	16.434	50.841
Random Forest	11.010	29.074
LightGBM	**10.824**	**28.845**
XGBoost	13.118	35.508

Neste caso, o modelo *LightGBM* foi o melhor modelo em ambas as métricas de avaliação.

Após a seleção do melhor modelo de ML, o desempenho global do modelo foi avaliado a partir dos dados preservados para teste. Neste caso o algoritmo *LightGBM* foi escolhido por ser o modelo com melhor desempenho na fase de seleção.

Na fase de avaliação do modelo final, os resultados obtidos presentes na Tabela 5.6 demonstraram, novamente, resultados muito à quem. Os resultados das duas métricas que poderiam à primeira vista parecer resultados prometedores relativamente ao modelo aplicado, depois quando comparados com a média de vendas (0,206) do conjunto de dados previstos, resultaram na dificuldade em perceber a qualidade do modelo, devido à grande quantidade de vendas nulas que enviesou a média de vendas e as métricas de avaliação. Assim, foram medidas, novamente, as métricas para os 25 produtos com total de vendas superior de forma a perceber a verdadeira qualidade do modelo. Com o resultado do erro médio absoluto igual a 21,047, foi constatado que o modelo não possui grande capacidade preditiva visto que a média de vendas semanais destes 25 produtos é de 19,181, ou seja, o erro é superior à própria média em cerca de 2 unidades, o que significa que por cada observação o modelo pode falhar em mais do que a média de vendas.

Tabela 5.6: Resultados da avaliação do modelo final do cenário 60 variáveis

Produtos	**MAE**	**RMSE**
Todos	0.238	3.380
Top 25	21.047	59.472

5.2 Comparação dos Cenários

Relativamente aos dois cenários, e comparando relativamente às previsões de todos os produtos, através da métrica RMSE foi verificado que os algoritmos *Linear Regression*, *Decision Tree*, *Random Forest*, e *XGBoost* tiveram melhor desempenho quando imputados pelo conjunto

de dados com apenas 9 variáveis independentes. Enquanto, os algoritmos *K-Nearest Neighbor* e *LightGBM* apresentaram melhor desempenho quando imputados pelo conjunto de dados com 60 variáveis independentes. Quando comparado apenas para os 25 produtos mais comercializados, a mesma regra foi demonstrada.

No que diz respeito aos modelos finais, ficou demonstrado para ambas as métricas aplicadas, e ambos os casos do tipo de produtos estudados, que o modelo *LightGBM* proveniente do cenário 60 variáveis apresentou melhor desempenho que o modelo *Random Forest* proveniente do cenário 9 variáveis quando submetidos aos dados de teste preservados para avaliação.

Embora exista um modelo ótimo para cada cenário e seja possível comparar estes modelos finais, os resultados obtidos para ambos não demonstraram bons resultados numa perspetiva do negócio da empresa.

5.3 Oportunidades de Melhoria

No âmbito da discussão dos resultados e da procura de melhorias do modelo preditivo, neste subcapítulo são mencionadas diversas opções a ser exploradas com intuito de obter melhores resultados.

Primeiramente, o *tunning* dos hiperparâmetros dos modelos utilizados pode ser uma forma de melhorar o processo de previsão de vendas. Um hiperparâmetro é um parâmetro cujo valor é utilizado para controlar o processo de aprendizagem do algoritmo. A otimização dos hiperparâmetros trata da escolha de um conjunto de hiperparâmetros ótimos para um algoritmo de *machine learning*.

Além disso, a metodologia descrita neste projeto pode ser alvo de alteração na procura de um modelo melhor. No que diz respeito ao conjunto de dados, acrescentar uma variável relativa à categoria de cada produto pode oferecer informação útil aos algoritmos sobre a quantidade vendida de cada produto. Relativamente ao modelo em si, a lei de Pareto pode ser explorada, na qual o objetivo passa por realizar uma análise ABC. A curva de experiência ABC, também chamada de análise de Pareto é um método de categorização de stocks cujo objetivo é determinar quais são os produtos mais importantes de uma empresa. Neste caso, o conjunto de dados seria segmentado em 3 classes diferentes: classe A, que corresponderia a 20% dos produtos; classe B, que corresponderia a 30% dos produtos; e classe C, que corresponderia a 50% dos produtos. De seguida, cada segmento poderia ser submetido a diferentes algoritmos de forma a perceber quais os algoritmos que generalizavam melhor cada segmento.

Por fim, os algoritmos de *Deep Learning*, que não foram considerados no âmbito desta dissertação, podem ser explorados. Estes algoritmos oferecem grandes expectativas no que diz respeito à previsão de séries temporais, sendo capazes de aprender automaticamente a dependência temporal e de tratar automaticamente as estruturas temporais como tendências e sazonalidade.

Capítulo 6

Conclusão

O principal objetivo do trabalho apresentado nesta dissertação, realizado na equipa *Data and Analytics*, era desenvolver uma ferramenta de *machine learning* que permitisse ao Dott ter acesso a previsões semanais das vendas de cada produto presente no seu *marketplace*. Com o aumento do consumo online devido à pandemia Covid-19, a presença de um modelo preditivo de vendas tornou-se imperativo. No início do projeto, o Dott não possuía qualquer tipo de ferramenta para prever a quantidade de vendas. Este projeto pretendia construir um modelo de previsão eficaz para ajudar o Dott no processo de tomada de decisões e prevenção de problemas.

A metodologia escolhida consistiu no desenvolvimento de vários modelos de ML, utilizando diferentes abordagens de treino para selecionar o modelo com melhor adaptação.

O projeto começou por enquadrar o problema em questão. Após a compreensão clara do problema, os dados necessários para o desenvolvimento dos modelos foram identificados e recolhidos. Foi realizada uma análise mais aprofundada das variáveis presentes no conjunto de dados. Assim, a primeira parte do projeto consistiu no pré-processamento dos dados.

A análise do conjunto de dados revelou uma assimetria natural na distribuição dos dados relativamente à variável dependente. As vendas nulas representaram 92,61% do volume de dados, o que deu origem a uma distribuição de dados extremamente enviesada. Pelo facto desta distribuição desequilibrada ser representativa da realidade atual do Dott, não foram aplicadas nenhumas técnicas para balancear os dados.

Depois de terminada a fase de pré-processamento dos dados, o passo seguinte foi aplicar uma técnica de seleção de variáveis de forma a diminui a dimensão do conjunto de dados. Através desta técnica, foram concebidas duas abordagens de treino diferentes. A primeira abordagem implicou a utilização de dados que eram compostos por apenas as variáveis independentes que explicavam 80% da variável dependente. A perspetiva de diminuir o *overfitting* foi o propulsor por detrás desta abordagem. A segunda abordagem correspondeu ao conjunto de dados composto pelas variáveis independentes que explicavam 100% da variável de resposta.

Foram definidos diferentes cenários, considerando as variações nas abordagens de treino dos algoritmos juntamente com a experiência de diversos modelos de *machine learning*. Uma das grandes vantagens é a capacidade de integração com qualquer modelos de ML devido à estrutura

geral de cada um. Assim, o conjunto de dados não necissitou de adaptações ou ajustes dependendo do algoritmo empregado. Este projeto centrou-se na utilização dos algoritmos *K-Nearest Neighbor* (KNN), *Linear Regression*, *Decision Tree*, *Random Forest*, *LightGBM* e *XGBoost*.

Estes cenários foram comparados, terminando com a escolha do modelo com melhor desempenho. A escolha foi feita tendo em conta não só as métricas de avaliação, mas também o tempo necessário de treino. Estas métricas de avaliação foram aplicadas relativamente aos 25 produtos mais comercializados, visto que os resultados estavam extremamente enviesados devido à grande quantidade de vendas nulas. As avaliações dos modelos finais foram conduzidas utilizando observações não vistas para imitar a realidade. No final, a abordagem com o conjunto de dados constituído por 60 variáveis juntamente com o algoritmo *LightGBM* apresentou os melhores resultados de desempenho.

Esta abordagem apresentou um MAE de 0,238 e um RMSE de 3,380 em relação à previsão de vendas de todos os produtos do conjunto de dados. Porém, quando medidas as mesmas métricas em relação aos 25 produtos mais vendidos, o MAE ascendeu para 21,047 e o RMSE para 59,472. Estes resultados não foram considerados suficientemente bons uma vez que, de uma perspetiva empresarial, os produtos mais vendidos têm bastante mais influência nas decisões da empresa, logo a sua previsão teria de ser muito mais precisa. No que diz respeito a produtos que apresentam muito poucas vendas, o algoritmo apresenta um bom desempenho, porém estes produtos têm influência mínima.

Um dos principais objetivos do projeto era demonstrar que os modelos de *machine learning* podem ser uma solução viável para o desenvolvimento de uma ferramenta de previsão de vendas, o que não foi conseguido.

Bibliografia

(2002). e-marketplaces: Crafting a winning strategy. *European Management Journal*, 20.

Akter, S. and Wamba, S. F. (2016). Big data analytics in e-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Cerqueira, V., Torgo, L., and Soares, C. (2019). Machine learning vs statistical methods for time series forecasting: Size matters. *arXiv preprint arXiv:1909.13316*.

Chevalier, S. (2021). Retail e-commerce sales in latin america from 2019 to 2025.

Chevalier, S. (2022). Global retail e-commerce market size 2014-2025.

Cramer-Flood, E. (2022). Global ecommerce forecast 2022.

Department, S. R. (2021). Europe: retail e-commerce revenue forecast from 2017 to 2025.

D'Orey, G. (2021). Worten, delta e dott.pt aceleram a fundo nas vendas online e mantêm foco no digital em 2022.

Du, W., Leung, S. Y. S., and Kwong, C. K. (2015). A multiobjective optimization-based neural network model for short-term replenishment forecasting in fashion industry. *Neurocomputing*, 151.

Eisenmann, T., Parker, G., and Alstyne, M. W. V. (2006). Strategies for two-sided markets. *Harvard Business Review*, 84.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70.

Gooijer, J. G. D. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22.

Hagiu, A. (2007). Merchant or two-sided platform? *Review of Network Economics*, 6(2).

J. Hyndman, R. and Athanasopoulos, G. (2014). *Forecasting: Principles and Practice: Notes*.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349.

Kawa, A. and Wałęsiak, M. (2019). Marketplace as a key actor in e-commerce value networks. *Logforum*, 15.

Keenan, M. (2021). What are stockouts and how can i prevent them?

Keller, J. M. and Gray, M. R. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-15.

Khan, A. G. (2016). Electronic commerce: A study on benefits and challenges in an emerging economy. *Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc*, 16.

Lawrence, J. E. and Tar, U. A. (2010). Barriers to ecommerce in developing countries usman a. tar. *Information, Society and Justice*, 3.

Lee, S., Lee, S. Y., and Ryu, M. H. (2019). How much are sellers willing to pay for the features offered by their e-commerce platform? *Telecommunications Policy*, 43.

Liu, N., Ren, S., Choi, T. M., Hui, C. L., and Ng, S. F. (2013). Sales forecasting for fashion retailing service industry: A review. *Mathematical Problems in Engineering*, 2013.

Lone, S. ., Harboul, N. ., and Weltevreden, J. W. J. (2021). 2021 european e-commerce report. *Uniwersytet śląski*.

Loureiro, A., Miguéis, V., and da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, 114:81–93.

Ma, Y. (2022). Annual gross merchandise volume (gmv) transacted on alibaba's retail marketplaces in china from financial year 2014 to 2022.

Mathradas, A. (2020). Covid-19 accelerated e-commerce adoption: What does it mean for the future?

Murphy, K. P. (2012). Machine learning - a probabilistic perspective - table-of-contents. *The MIT Press*.

Nelson, C. R. and Kang, H. (1984). Pitfalls in the use of time as an explanatory variable in regression. *Journal of Business and Economic Statistics*, 2.

Niranjanamurthy, M., Kavyashree, N., and Chahar, S. J. D. (2013). Analysis of e-commerce and m-commerce : Advantages , limitations and security issues. *International Journal of Advanced Research in Computer and Communication Engineering*, 2.

Osterwalder, A., Pigneur, Y., and Clark, T. (2010). Business model generation: A handbook for visionaries, game changers, and challengers. hoboken. *Communications of the association for Information Systems, 16(1), 1.*, 75.

Pimenta, A. (2020). Ctt e-commerce report 2020.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1.

Raschka, S. (2018). Model evaluation , model selection , and algorithm selection in machine learning performance estimation : Generalization performance vs . model selection. *arXiv*.

Ren, S., Chan, H. L., and Ram, P. (2017). A comparative study on fashion demand forecasting models with multiple sources of uncertainty. *Annals of Operations Research*, 257.

Ren, S., Choi, T. M., and Liu, N. (2015). Fashion sales forecasting with a panel data-based particle-filter model. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45.

Sarkar, M. B., Butler, B., and Steinfield, C. (1995). Intermediaries and cybermediaries: a continuing role for mediating players in the electronic marketplace. *Journal of Computer-Mediated Communication*, 1.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14.

Song, Y., Liang, J., Lu, J., and Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, 251.

Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Weisberg, S. (2005). *Applied Linear Regression: Third Edition*.

Weise, K. (2021). Amazon's profit soars 220 percent as pandemic drives shopping online.

Xia, M. and Wong, W. K. (2014). A seasonal discrete grey forecasting model for fashion retailing. *Knowledge-Based Systems*, 57.

Yang, X. S. (2019). *Introduction to algorithms for data mining and machine learning*.

Anexo A

Significado das Variáveis

Tabela A.1: Significado das restantes variáveis independentes

Variáveis	Significado
add_to_cart_last_week	quantidade do produto adicionado ao cesto na semana anterior
add_to_cart_last_2_weeks	quantidade do produto adicionado ao cesto nas 2 semanas anteriores
add_to_cart_last_4_weeks	quantidade do produto adicionado ao cesto nas 4 semanas anteriores
add_to_cart_all_time	quantidade do produto adicionado ao cesto desde sempre
remove_from_cart_last_week	quantidade do produto removido do cesto na semana anterior
remove_from_cart_last_2_weeks	quantidade do produto removido do cesto nas 2 semanas anteriores
remove_from_cart_last_4_weeks	quantidade do produto removido do cesto nas 4 semanas anteriores
remove_from_cart_all_time	quantidade do produto removido do cesto desde sempre
product_page_views_last_week	número de visitas da página do produto na semana anterior
product_page_views_last_2_weeks	número de visitas da página do produto nas 2 semanas anteriores
product_page_views_last_4_weeks	número de visitas da página do produto nas 4 semanas anteriores
product_page_views_all_time	número de visitas da página do produto desde sempre
impression_views_last_week	número de vezes que a publicidade do produto apareceu na semana anterior
impression_views_last_2_weeks	número de vezes que a publicidade do produto apareceu nas 2 semanas anteriores
impression_views_last_4_weeks	número de vezes que a publicidade do produto apareceu nas 4 semanas anteriores
impression_views_all_time	número de vezes que a publicidade do produto apareceu desde sempre
avg_seconds_per_customer_last_week	média de segundos na página do produto por utilizador na semana anterior
avg_seconds_per_customer_last_2_weeks	média de segundos na página do produto por utilizador nas 2 semanas anteriores
avg_seconds_per_customer_last_4_weeks	média de segundos na página do produto por utilizador nas 4 semanas anteriores
avg_seconds_per_customer_all_time	média de segundos na página do produto por utilizador desde sempre
avg_seconds_per_customer_seasonality_last_2_weeks	média de segundos na página do produto por utilizador nas 2 semanas anteriores relativas à mesma semana no ano anterior
avg_seconds_per_customer_seasonality_last_4_weeks	média de segundos na página do produto por utilizador nas 4 semanas anteriores relativas à mesma semana no ano anterior
total_seconds_on_product_page_last_week	total de segundos na página do produto na semana anterior
total_seconds_on_product_page_last_2_weeks	total de segundos na página do produto nas 2 semanas anteriores
total_seconds_on_product_page_last_4_weeks	total de segundos na página do produto nas 4 semanas anteriores
total_seconds_on_product_page_all_time	total de segundos na página do produto desde sempre
total_seconds_on_product_page_seasonality_last_2_weeks	total de segundos na página do produto nas 2 semanas anteriores relativas à mesma semana no ano anterior
total_seconds_on_product_page_seasonality_last_4_weeks	total de segundos na página do produto nas 4 semanas anteriores relativas à mesma semana no ano anterior
nr_sales_last_week	número de vendas do produto na semana anterior
nr_sales_last_2_weeks	número de vendas do produto nas 2 semanas anteriores
nr_sales_last_4_weeks	número de vendas do produto nas 4 semanas anteriores
nr_sales_all_time	número de vendas do produto desde sempre
nr_sales_seasonality_2_weeks	número de vendas do produto das 2 semanas anteriores relativas à mesma semana no ano anterior
nr_sales_seasonality_4_weeks	número de vendas do produto das 4 semanas anteriores relativas à mesma semana no ano anterior
ratio_repurchase_last_week	rácio de recompra do produto na semana anterior
ratio_repurchase_last_2_weeks	rácio de recompra do produto nas 2 semanas anteriores
ratio_repurchase_last_4_weeks	rácio de recompra do produto nas 4 semanas anteriores
ratio_repurchase_all_time	rácio de recompra do produto desde sempre
nr_unique_customers_last_week	número de clientes únicos do produto na semana anterior
nr_unique_customers_last_2_weeks	número de clientes únicos do produto nas 2 semanas anteriores
nr_unique_customers_last_4_weeks	número de clientes únicos do produto nas 4 semanas anteriores
nr_unique_customers_all_time	número de clientes únicos do produto desde sempre
ratio_sales_in_leaf_last_week	rácio de vendas do produto relativamente à categoria mais granular a que pertence na semana anterior
ratio_sales_in_leaf_last_2_weeks	rácio de vendas do produto relativamente à categoria mais granular a que pertence nas 2 semanas anteriores
ratio_sales_in_leaf_last_4_weeks	rácio de vendas do produto relativamente à categoria mais granular a que pertence nas 4 semanas anteriores
ratio_sales_in_leaf_all_time	rácio de vendas do produto relativamente à categoria mais granular a que pertence desde sempre
ratio_sales_in_parent_cat_last_week	rácio de vendas do produto relativamente à categoria menos granular a que pertence na semana anterior
ratio_sales_in_parent_cat_last_2_weeks	rácio de vendas do produto relativamente à categoria menos granular a que pertence nas 2 semanas anteriores
ratio_sales_in_parent_cat_last_4_weeks	rácio de vendas do produto relativamente à categoria menos granular a que pertence nas 4 semanas anteriores
ratio_sales_in_parent_cat_all_time	rácio de vendas do produto relativamente à categoria menos granular a que pertence desde sempre
nr_sales_leaf_cat_last_week	número de vendas da categoria mais granular relativa ao produto na semana anterior
nr_sales_leaf_cat_last_2_weeks	número de vendas da categoria mais granular relativa ao produto nas 2 semanas anteriores
nr_sales_leaf_cat_last_4_weeks	número de vendas da categoria mais granular relativa ao produto nas 4 semanas anteriores
nr_sales_leaf_cat_all_time	número de vendas da categoria mais granular relativa ao produto desde sempre
nr_sales_parent_cat_last_week	número de vendas da categoria menos granular relativa ao produto na semana anterior
nr_sales_parent_cat_last_2_weeks	número de vendas da categoria menos granular relativa ao produto nas 2 semanas anteriores
nr_sales_parent_cat_last_4_weeks	número de vendas da categoria menos granular relativa ao produto nas 4 semanas anteriores
nr_sales_parent_cat_all_time	número de vendas da categoria menos granular relativa ao produto desde sempre
nr_reviews_last_week	número de análises ao produto por parte dos consumidores na semana anterior
nr_reviews_last_2_weeks	número de análises ao produto por parte dos consumidores nas 2 semanas anteriores
nr_reviews_last_4_weeks	número de análises ao produto por parte dos consumidores nas 4 semanas anteriores
nr_reviews_all_time	número de análises ao produto por parte dos consumidores desde sempre

Tabela A.2: Significado das restantes variáveis independentes (continuação)

Variáveis	**Significado**
avg_rate_last_week	média do rating atribuído pelos consumidores ao produto na semana anterior
avg_rate_last_2_weeks	média do rating atribuído pelos consumidores ao produto nas 2 semanas anteriores
avg_rate_last_4_weeks	média do rating atribuído pelos consumidores ao produto nas 4 semanas anteriores
avg_rate_all_time	média do rating atribuído pelos consumidores ao produto desde sempre
nr_returns_last_week	número de devoluções do produto na semana anterior
nr_returns_last_2_weeks	número de devoluções do produto nas 2 semanas anteriores
nr_returns_last_4_weeks	número de devoluções do produto nas 4 semanas anteriores
nr_returns_all_time	número de devoluções do produto desde sempre
nr_refunds_last_week	número de devoluções do pagamento do produto na semana anterior
nr_refunds_last_2_weeks	número de devoluções do pagamento do produto nas 2 semanas anteriores
nr_refunds_last_4_weeks	número de devoluções do pagamento do produto nas 4 semanas anteriores
nr_refunds_all_time	número de devoluções do pagamento do produto desde sempre
nr_adds_to_wishlist_last_week	número de adições do produto à lista de desejos na semana anterior
nr_adds_to_wishlist_last_2_weeks	número de adições do produto à lista de desejos nas 2 semanas anteriores
nr_adds_to_wishlist_4_weeks	número de adições do produto à lista de desejos nas 4 semanas anteriores
nr_adds_to_wishlist_all_time	número de adições do produto à lista de desejos desde sempre
max_discount_last_week	desconto máximo feito do produto na semana anterior
max_discount_last_2_weeks	desconto máximo feito do produto nas 2 semanas anteriores
max_discount_4_weeks	desconto máximo feito do produto nas 4 semanas anteriores
max_discount_all_time	desconto máximo feito do produto desde sempre
min_discount_last_week	desconto mínimo feito do produto na semana anterior
min_discount_last_2_weeks	desconto mínimo feito do produto nas 2 semanas anteriores
min_discount_4_weeks	desconto mínimo feito do produto nas 4 semanas anteriores
min_discount_all_time	desconto mínimo feito do produto desde sempre
max_price_last_week	preço máximo do produto na semana anterior
max_price_last_2_weeks	preço máximo do produto nas 2 semanas anteriores
max_price_4_weeks	preço máximo do produto nas 4 semanas anteriores
max_price_all_time	preço máximo do produto desde sempre
min_price_last_week	preço mínimo do produto na semana anterior
min_price_last_2_weeks	preço mínimo do produto nas 2 semanas anteriores
min_price_4_weeks	preço mínimo do produto nas 4 semanas anteriores
min_price_all_time	preço mínimo do produto desde sempre
stddev_price_last_week	desvio padrão do preço do produto na semana anterior
stddev_price_last_2_weeks	desvio padrão do preço do produto nas 2 semanas anteriores
stddev_price_4_weeks	desvio padrão do preço do produto nas 4 semanas anteriores
stddev_price_all_time	desvio padrão do preço do produto desde sempre
nr_offers_last_week	número de ofertas do produto na semana anterior
nr_offers_last_2_weeks	número de ofertas do produto nas 2 semanas anteriores
nr_offers_4_weeks	número de ofertas do produto nas 4 semanas anteriores
nr_offers_all_time	número de ofertas do produto desde sempre
leaf_cat_avg_price_last_week	preço médio dos produtos pertencentes à categoria mais granular do produto na semana anterior
leaf_cat_avg_price_last_2_weeks	preço médio dos produtos pertencentes à categoria mais granular do produto nas 2 semanas anteriores
leaf_cat_avg_price_4_weeks	preço médio dos produtos pertencentes à categoria mais granular do produto nas 4 semanas anteriores
leaf_cat_avg_price_all_time	preço médio dos produtos pertencentes à categoria mais granular do produto desde sempre
leaf_cat_avg_max_discount_last_week	média dos descontos máximos dos produtos pertencentes à categoria mais granular do produto na semana anterior
leaf_cat_avg_max_discount_last_2_weeks	média dos descontos máximos dos produtos pertencentes à categoria mais granular do produto nas 2 semanas anteriores
leaf_cat_avg_max_discount_4_weeks	média dos descontos máximos dos produtos pertencentes à categoria mais granular do produto nas 4 semanas anteriores
leaf_cat_avg_max_discount_all_time	média dos descontos máximos dos produtos pertencentes à categoria mais granular do produto desde sempre

Anexo B

Importância das Variáveis

Tabela B.1: Importância de cada variável independente do conjunto de dados

Variáveis	Pontuação
nr_sales_last_2_weeks	0.239235
ratio_sales_in_leaf_last_2_weeks	0.113394
avg_seconds_per_customer_seasonality_last_4_weeks	0.106574
avg_seconds_per_customer_last_4_weeks	0.090999
avg_seconds_per_customer_last_2_weeks	0.079323
impression_views_last_2_weeks	0.057759
ratio_sales_in_leaf_last_week	0.057009
nr_returns_last_week	0.038437
remove_from_cart_last_4_weeks	0.025757
ratio_sales_in_leaf_all_time	0.023094
total_seconds_on_product_page_seasonality_last_...	0.020575
add_to_cart_all_time	0.015312
avg_seconds_per_customer_last_week	0.014068
nr_sales_last_week	0.011455
add_to_cart_last_2_weeks	0.008845
add_to_cart_last_4_weeks	0.007663
remove_from_cart_last_week	0.005674
ratio_sales_in_parent_cat_last_4_weeks	0.005571
ratio_sales_in_parent_cat_last_week	0.005518
nr_sales_leaf_cat_last_week	0.004882
nr_sales_leaf_cat_all_time	0.004870
nr_sales_all_time	0.004091
nr_unique_customers_last_4_weeks	0.003848
avg_seconds_per_customer_seasonality_last_2_weeks	0.003726
nr_sales_last_4_weeks	0.003672
add_to_cart_last_week	0.003422
nr_offers_all_time	0.003252
remove_from_cart_last_2_weeks	0.003233
nr_unique_customers_last_2_weeks	0.003066

Tabela B.2: Importância de cada variável independente do conjunto de dados (continuação)

Variáveis	Pontuação
total_seconds_on_product_page_all_time	0.002800
product_page_views_last_4_weeks	0.002568
nr_unique_customers_last_week	0.001986
ratio_repurchase_all_time	0.001959
nr_sales_seasonality_2_weeks	0.001885
nr_sales_leaf_cat_last_2_weeks	0.001857
nr_sales_parent_cat_last_week	0.001649
ratio_repurchase_last_week	0.001559
product_page_views_last_week	0.001557
max_price_last_week	0.001456
nr_sales_seasonality_4_weeks	0.001445
nr_refunds_last_week	0.001415
total_seconds_on_product_page_last_week	0.001355
nr_sales_parent_cat_last_4_weeks	0.001302
nr_returns_last_4_weeks	0.001152
nr_reviews_all_time	0.000999
max_price_all_time	0.000885
avg_rate_last_4_weeks	0.000778
leaf_cat_avg_max_discount_all_time	0.000664
nr_reviews_last_2_weeks	0.000605
nr_adds_to_wishlist_all_time	0.000596
nr_unique_customers_all_time	0.000595
product_page_views_last_2_weeks	0.000578
min_price_last_2_weeks	0.000576
nr_returns_last_2_weeks	0.000535
nr_sales_parent_cat_all_time	0.000486
min_price_4_weeks	0.000423
total_seconds_on_product_page_last_4_weeks	0.000325
remove_from_cart_all_time	0.000308
min_price_all_time	0.000236
ratio_sales_in_parent_cat_last_2_weeks	0.000230

Anexo C

Resultados

Tabela C.1: Resultados da seleção dos modelos com o conjunto de dados com 9 variáveis independentes

KNN_RMSE	KNN_MAE	LR_RMSE	LR_MAE	DT_RMSE	DT_MAE	RF_RMSE	RF_MAE	LGBM_RMSE	LGBM_MAE	XGB_RMSE	XGB_MAE
4.797766	0.304111	5.260228	0.308444	3.772842	0.331000	4.536543	0.298667	5.302316	0.279000	4.177080	0.263556
1.681071	0.202667	1.562512	0.313000	1.959138	0.272000	0.952599	0.208778	2.400879	0.184444	2.384091	0.168111
1.028375	0.193111	1.431821	0.321889	2.632700	0.265778	0.935949	0.203556	1.488437	0.163000	1.033387	0.150556
2.687915	0.209556	2.652839	0.273111	3.281446	0.259667	2.447834	0.205444	1.965169	0.161222	2.460578	0.158222
0.720185	0.148444	0.855050	0.202222	0.948683	0.189333	0.713053	0.153333	0.779459	0.117778	0.665582	0.111222
0.809938	0.180000	1.259056	0.226556	1.905752	0.241000	0.781381	0.179667	0.733712	0.138111	0.920869	0.140222
1.126154	0.209556	1.172604	0.261667	1.762794	0.270556	1.229273	0.220222	1.217192	0.174444	1.200000	0.168889
2.853341	0.264222	2.374588	0.301778	3.009393	0.319111	2.174882	0.247889	2.156077	0.204667	1.688688	0.186556
1.210968	0.221556	1.272487	0.280778	1.478588	0.278889	1.276628	0.231111	1.438981	0.183778	1.295634	0.175778
2.988032	0.318111	3.318450	0.370778	6.063937	0.405556	4.525422	0.358333	2.646507	0.266444	5.698947	0.314889

Tabela C.2: Resultados da seleção dos modelos com o conjunto de dados com 9 variáveis independentes para os 25 produtos mais vendidos

KNN_RMSE	KNN_MAE	LR_RMSE	LR_MAE	DT_RMSE	DT_MAE	RF_RMSE	RF_MAE	LGBM_RMSE	LGBM_MAE	XGBoost_RMSE	XGBoost_MAE
89.246400	21.960000	98.182279	23.280000	68.554212	17.840000	83.054681	20.480000	98.977573	23.600000	77.334856	19.640000
27.249193	8.592593	22.380547	8.370370	44.916796	14.259259	7.532645	4.888889	41.242732	12.814815	41.654220	11.074074
11.750380	5.928571	18.437152	8.857143	14.683567	7.035714	9.235026	4.571429	23.563743	9.321429	13.996173	6.464286
45.879501	15.642857	44.378244	17.500000	68.935011	22.178571	43.155450	15.107143	32.452163	14.000000	42.020828	14.321429
2.546126	1.655172	4.177113	2.827586	3.205599	2.275862	2.498275	1.896552	7.355505	3.758621	2.378061	1.793103
4.986939	2.608696	18.673394	6.347826	35.680953	10.173913	3.753259	2.434783	4.338904	2.652174	11.921847	4.478261
9.714680	6.000000	7.721723	5.312500	19.227259	7.750000	9.714680	5.750000	13.255895	7.218750	12.955453	6.343750
48.823071	20.076923	38.419747	15.461538	35.730831	16.000000	32.365105	13.038462	32.165438	13.307692	25.490571	11.461538
14.201959	6.217391	14.044959	7.347826	14.263057	6.565217	13.652520	6.217391	13.200790	6.434783	13.996894	6.173913
46.837151	22.843750	52.318137	25.437500	135.084349	46.843750	70.109513	32.843750	40.005859	21.843750	94.193285	37.437500

Tabela C.3: Resultados da seleção dos modelos com o conjunto de dados com 60 variáveis independentes

KNN_RMSE	KNN_MAE	LR_RMSE	LR_MAE	DT_RMSE	DT_MAE	RF_RMSE	RF_MAE	LGBM_RMSE	LGBM_MAE	XGBoost_RMSE	XGBoost_MAE
4.644100	0.292333	5.247031	0.294444	4.635635	0.349778	4.913847	0.288778	5.082279	0.278889	4.713880	0.264667
1.691778	0.198333	2.739830	0.303333	2.700062	0.317667	0.846955	0.198222	1.950128	0.177444	3.039188	0.175333
1.633027	0.186778	1.532319	0.288444	3.132181	0.286778	0.911836	0.176778	1.145183	0.159000	0.918695	0.147778
1.936377	0.181556	2.664186	0.264111	2.497198	0.273111	2.614681	0.189000	2.106762	0.162889	2.895955	0.168333
0.747812	0.145222	0.879710	0.206778	1.294304	0.246111	0.702693	0.144889	0.682398	0.118556	0.679706	0.113111
1.343503	0.178556	0.978775	0.226222	2.033743	0.283222	0.854335	0.174556	0.727095	0.141333	1.223338	0.153222
1.023502	0.200000	1.260555	0.249889	1.763582	0.294889	1.138518	0.204000	1.274406	0.178111	1.053196	0.169222
1.764055	0.221667	2.284659	0.296556	2.489355	0.316444	2.084386	0.228222	2.393904	0.212111	2.089551	0.193556
1.067916	0.206222	1.245927	0.265444	1.604715	0.294000	1.099899	0.201111	1.156527	0.175778	1.048756	0.167444
3.439024	0.287556	2.867481	0.340667	8.169632	0.461556	4.716496	0.331111	2.608852	0.272333	6.359210	0.320444

Tabela C.4: Resultados da seleção dos modelos com o conjunto de dados com 60 variáveis independentes para os 25 produtos mais vendidos

KNN_RMSE	KNN_MAE	LR_RMSE	LR_MAE	DT_RMSE	DT_MAE	RF_RMSE	RF_MAE	LGBM_RMSE	LGBM_MAE	XGBoost_RMSE	XGBoost_MAE
86.245000	21.160000	97.960400	23.120000	95.295540	23.720000	91.006154	21.800000	94.814978	22.760000	87.777674	21.000000
27.351620	9.000000	46.475003	12.518519	54.291054	14.851852	6.131280	4.259259	32.840242	10.407407	54.033255	14.333333
26.328963	10.285714	21.628354	9.142857	43.636895	12.178571	13.441673	6.607143	16.462078	8.428571	11.368817	6.035714
31.795665	12.750000	44.507223	17.607143	39.231001	11.928571	39.862711	13.750000	35.102401	14.035714	50.171136	17.571429
3.034287	2.172414	5.034365	3.689655	7.838895	3.793103	3.039964	2.206897	3.449138	2.586207	3.699953	2.379310
22.162444	6.652174	9.901691	5.000000	37.622264	12.826087	5.552125	3.695652	4.582576	3.434783	19.221817	6.608696
5.929271	4.468750	11.688670	5.937500	27.824225	9.687500	11.202678	5.750000	14.710753	6.468750	8.672874	5.218750
26.337309	11.192308	33.911083	14.884615	51.010934	19.269231	33.743945	13.269231	35.913357	13.846154	32.518634	13.076923
11.835834	5.391304	13.409925	8.086957	14.742721	8.304348	12.843540	6.260870	12.994982	6.173913	12.211897	5.826087
55.266118	22.656250	44.355806	22.500000	136.913636	47.781250	73.913801	32.500000	37.581162	20.093750	105.402502	39.125000