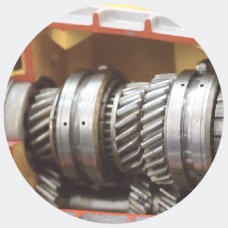
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA E GESTÃO INDUSTRIAL






MODELAÇÃO PREDITIVA BASEADA EM ALGORITMOS DE IA PARA PREVISÃO DE VENDAS NO RETALHO

PEDRO ALEXANDRE LOPES SILVA
julho de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

MODELAÇÃO PREDITIVA BASEADA EM ALGORITMOS DE IA PARA PREVISÃO DE VENDAS NO RETALHO

Forecasting de vendas de novos artigos com base em dados históricos

Pedro Alexandre Lopes Silva

Dissertação para obtenção do Grau de Mestre em Engenharia e Gestão Industrial

Orientador: Prof. Dr. Carlos Ferreira

Co-orientador: Prof. Dr. Manuel Pereira Lopes

Co-orientador: Prof. Dr. Bruno Ferraz de Sousa

Porto, Julho 2025

P.PORTO

isep

RESUMO

O lançamento de artigos sem histórico de vendas aumenta o risco operacional no retalho. Esta investigação propõe uma estrutura preditiva que estima a procura inicial de novos produtos a partir de **3 278 907 transações reais (2022-2024)** cobrindo **14 422 SKUs, 120 lojas e dois segmentos de cliente**. Seguindo o ciclo **CRISP-DM**, procedeu-se à limpeza dos dados, análise exploratória e engenharia de variáveis, antes de comparar quatro algoritmos — **XGBoost, LightGBM, LSTM e Transformer** — em cenários global e por família de produtos, avaliados com MAE, RMSE, MAPE e R^2.

Os resultados revelam dois patamares distintos: os modelos de árvores de gradiente (XGBoost ≈ LightGBM) registam erros médios substancialmente menores e R^2 positivos, ao passo que as redes neuronais sequenciais (LSTM, Transformer) apresentam elevada variabilidade e R^2 negativos em várias famílias. O **XGBoost treinado globalmente** demonstra o menor RMSE ponderado e o melhor equilíbrio viés-variância, sendo recomendado como motor de previsão único para toda a gama de artigos. Esta solução simplifica a operação, mantém precisão elevada e foi integrada num **protótipo de *dashboard web*** para validação em contexto real.

Conclui-se que um modelo único, alimentado pela diversidade de SKUs e lojas, generaliza padrões de procura com eficácia, oferecendo uma ferramenta prática para apoiar decisões ao nível do portefólio e de planeamento comercial e operacional.

Palavras-chave: previsão de vendas; novos produtos no retalho; XGBoost; machine-learning; inteligência artificial.

página propositadamente em branco

ABSTRACT

Introducing retail items with no sales history entails significant demand uncertainty. This study develops a predictive framework that leverages **3,278,907 real transactions (2022–2024)** spanning **14,422 SKUs, 120 stores and two customer segments**. Adhering to the **CRISP-DM** methodology, the workflow comprises data cleansing, exploratory analysis, feature engineering and the assessment of four algorithms — **XGBoost, LightGBM, LSTM and Transformer**— under global and family-specific settings using MAE, RMSE, MAPE and R^2.

Findings disclose **two performance tiers**: gradient-boosted trees (XGBoost ≈ LightGBM) achieve markedly lower errors and positive R^2, whereas sequence models (LSTM, Transformer) struggle with variance and often yield negative R^2. The **globally trained XGBoost** secures the lowest weighted RMSE and the best bias-variance trade-off and is thus recommended as a single forecasting engine for the entire product range. A **web-based dashboard prototype** demonstrates real-time deployment and business applicability.

The study concludes that a unified model, trained across heterogeneous SKUs and outlets, can generalize demand patterns effectively, delivering high accuracy with reduced operational complexity for assortment and inventory planning.

Keywords: sales forecasting; new products; XGBoost; retail; machine learning; artificial intelligence

página propositadamente em branco

AGRADECIMENTOS

Chegado ao fim deste percurso, não posso deixar de endereçar algumas palavras especiais a quem foi fundamental para que eu aqui chegasse.

Em primeiro lugar, deixo o meu profundo agradecimento aos meus orientadores, Prof. Dr. Carlos Ferreira e Prof. Dr. Manuel Pereira Lopes, bem como ao Prof. Bruno Sousa, colaborador da empresa que forneceu os dados e para a qual o projeto foi desenvolvido, cuja intervenção foi decisiva para articular o trabalho entre a academia e o meio empresarial. Uma menção especial ao Prof. Pereira Lopes, que me abriu as portas desta notável instituição e me desafiou a sair da zona de conforto, abraçando uma área tão fulcral quanto exigente. Fê-lo com o propósito de expandir os meus conhecimentos e ajudar-me a tornar-me um profissional mais completo — objetivo que, hoje, sinto ter alcançado.

À minha companheira de vida, Mariana, a quem nada parece impossível. A gestão, a dois, das responsabilidades pessoais, profissionais e académicas revelou-se um desafio mais profundo do que qualquer unidade curricular. Ainda assim, foste sempre o pilar essencial desta jornada: impulsionaste-me quando o receio falava mais alto e, nos momentos em que seria mais fácil desistir, não me deixaste baixar os braços. Entre aulas noturnas, fins de semana entregues a avaliações, compromissos adiados e o desgaste físico e emocional, estiveste sempre presente, com as palavras certas, fazendo-me acreditar que chegaria aqui. Cheguei — como repetes há mais de uma década. O meu agradecimento é eterno e o meu amor incalculável.

Ao amigo de sempre, Telmo. Desde cedo foste um exemplo: enquanto filho, mais tarde irmão e hoje enquanto marido da Vanessa; noutros espectros da vida, como estudante e praticante de artes marciais – sempre foste o melhor em tudo. És daquelas raras pessoas a quem ninguém encontra falhas, e tudo te sai com uma naturalidade irrepreensível, sem nunca perder esse traço humilde que te carateriza acima de todas as outras tuas características. Não guardo memórias que não te incluam; és, verdadeiramente, um *«brother from another mother».* Nas horas em que me faltava inspiração — até quando já não conseguia manter-me de olhos abertos — tu estavas lá, a empurrar-me para a frente. Chego aqui muito por tua causa. Serei sempre grato por tudo o que fazes e representas. És enorme!

Aos colegas de batalha, em especial ao grupo MALIBU — António Silva, Marta Silva, Ana Silva e Beatriz Meireles —, com quem enfrentei o enorme desafio da unidade ENGES e, acima de tudo, com quem partilhei momentos de alegria, confraternização e aprendizagem que tanto contribuíram para este resultado. À Carolina Domingos, que por circunstâncias alheias não concluiu este percurso, mas que, desde o PG-GCA até ao MEGI, partilhou comigo desafios e angústias que superámos juntos. Desejo-vos a todos o maior sucesso, pessoal e profissional.

Por último, à minha Família — pais, irmãos e avós —, presença tantas vezes omnipresente, que me dotou das ferramentas pessoais, mentais e inspiradoras necessárias para, ao fim de mais de 10 anos como trabalhador-estudante, continuar este caminho que já se faz motivo de orgulho.

A todos vós, que são parte de mim muito além deste percurso, o meu muito obrigado.

página propositadamente em branco

ÍNDICE

ÍNDICE DE FIGURAS

página propositadamente em branco

ÍNDICE DE TABELAS

página propositadamente em branco

LISTAS DE SIGLAS E SÍMBOLOS

Lista de Siglas

ARIMA	*Auto Regressive Integrated Moving Average*
DL	*Deep Learning*
EDA	*Exploratory Data Analysis*
IA	Inteligência Artificial
ISEP	Instituto Superior de Engenharia do Porto
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
SKU	*Stock Keeping Unit*

1. INTRODUÇÃO

A presente dissertação tem como foco o desenvolvimento de modelos preditivos para apoiar o lançamento de novos produtos no setor do retalho, utilizando dados históricos e técnicas avançadas de análise. Com base num **caso real de uma empresa portuguesa** desse setor, pretende-se compreender de que forma é possível **antecipar o desempenho comercial de artigos recém-introduzidos**, minimizando os riscos associados à falta de histórico.

Nos pontos seguintes, apresenta-se o enquadramento do problema, os objetivos da investigação e a metodologia adotada.

1.1. Problema, enquadramento e pertinência

No contexto atual de rápida expansão no setor do retalho, a introdução de novos produtos assume um papel cada vez mais relevante para o crescimento sustentável das empresas. A organização em análise encontra-se numa fase de forte crescimento, tanto ao nível da sua estrutura como da diversidade do seu portefólio de produtos, nomeadamente no número de *Stock Keeping Units (SKUs)* ativos. Esta evolução, embora positiva, intensifica os desafios associados à previsão da procura, especialmente no momento do lançamento de novos artigos, para os quais não existe histórico direto.

As estimativas iniciais de vendas continuam a ser, na maioria das vezes, baseadas em julgamentos subjetivos, o que conduz frequentemente a decisões de encomenda desajustadas — quer por excesso, implicando custos de stock, quer por defeito, resultando em ruturas e perda de oportunidades de venda. Este problema torna-se crítico num contexto em que a agilidade e a precisão na gestão da cadeia de abastecimento são determinantes para a competitividade.

Neste sentido, a presente investigação procura responder à necessidade de desenvolver abordagens sistemáticas e baseadas em dados que permitam:

- Identificar padrões de comportamento em artigos historicamente semelhantes (com base em atributos como família ou subfamília);
- Estimar o desempenho de vendas de novos produtos desde o momento do seu lançamento;
- E ajustar previsões após os primeiros sinais de venda, melhorando a capacidade de planeamento para as semanas seguintes.

Com a evolução das técnicas de *machine learning* e *deep learning*, surgem novas oportunidades para construir modelos preditivos robustos. Entre os modelos a considerar destacam-se o XGBoost e o LightGBM, amplamente utilizados em contextos com dados estruturados, bem como redes LSTM (*Long Short-Term Memory*) e Transformer, que oferecem capacidade para captar dependências temporais mais complexas nos dados.

1.2. Questão e objetivos de investigação

A investigação parte das seguintes questões principais:

Q1: Que modelos preditivos podem ser utilizados para estimar com eficácia o desempenho de novos produtos no retalho, desde o seu lançamento, a partir de dados históricos de artigos similares?

Q2: Modelos globais têm melhores resultados que modelos específicos por família de artigos?

Estas questões desdobram-se em objetivos específicos, sendo estes:

- Analisar dados históricos de vendas por SKU e loja, explorando agrupamentos por família, subfamília e outras variáveis explicativas (como loja e preço, p.e.);
- Identificar padrões de comportamento entre artigos semelhantes;
- Desenvolver e comparar modelos de previsão de vendas utilizando XGBoost, LightGBM, LSTM e Transformer;
- Avaliar o desempenho dos modelos na previsão de vendas no momento do lançamento (sem histórico do produto);
- Propor uma abordagem que permita melhorar a tomada de decisão na introdução de novos produtos, contribuindo para a redução de ruturas ou excesso de stock.

1.3. Opções metodológicas

A metodologia adotada segue formalmente o modelo CRISP-DM[1], um referencial amplamente reconhecido para o desenvolvimento estruturado de projetos de ciência de dados. O estudo assenta numa abordagem quantitativa, recorrendo a técnicas de análise exploratória e modelação preditiva com base em dados reais fornecidos pela empresa, nomeadamente registos de vendas, atributos dos artigos (família, subfamília, preço, tipo de cliente) e variáveis temporais (semana, sazonalidade, eventos).

As etapas metodológicas foram organizadas de acordo com as fases definidas pelo CRISP-DM, da seguinte forma:

- **Preparação e exploração dos dados**: inclui a limpeza, agregação, criação de variáveis derivadas e análise exploratória, com o objetivo de compreender a estrutura dos dados, identificar padrões e anomalias e assegurar a sua adequação à modelação (etapas de **compreensão dos dados** e **preparação dos dados**);
- **Modelação preditiva**: desenvolvimento de quatro tipos de modelos — XGBoost, LightGBM, LSTM e Transformer — com o objetivo de prever a quantidade de vendas aquando do lançamento de novos produtos, utilizando atributos disponíveis e padrões históricos (fase de **modelação**);

[1] CRISP-DM - *Cross-Industry Standard Process for Data Mining*, metodologia em seis fases originalmente proposta por Wirth & Hipp (2000)

- **Validação e comparação**: avaliação do desempenho dos modelos com recurso a métricas como MAE, RMSE e R^2, analisando-se também a sua aplicabilidade em diferentes cenários de negócio (fase de **avaliação**);
- **Discussão e implicações práticas**: interpretação crítica dos resultados e recomendações para a integração dos modelos nos processos de decisão da empresa (fase de **implementação**).

Este enquadramento assegura consistência metodológica ao longo do estudo, garantindo que cada etapa contribui de forma estruturada para a resolução do problema de negócio identificado.

1.4. Estrutura da dissertação

Ao longo do documento, cada capítulo aprofunda uma fase distinta do projeto. O capítulo 2 desenvolve uma revisão da literatura orientada pelas principais fases do CRISP-DM, oferecendo o enquadramento teórico necessário. O **capítulo 3** dedica-se à análise exploratória dos dados, enquanto o **capítulo 4** descreve em detalhe o processo metodológico de preparação, modelação e avaliação dos modelos preditivos. O c**apítulo 5** compara os resultados obtidos com diferentes algoritmos e estratégias de segmentação. O **capítulo 6** documenta a aplicação prática do modelo num protótipo de *dashboard*. Por fim, o **capítulo 7** apresenta as conclusões finais, as limitações do estudo e recomendações para futuras investigações.

2. REVISÃO BIBLIOGRÁFICA

Esta revisão bibliográfica foi estruturada para oferecer uma visão progressiva dos principais contributos académicos no domínio da previsão de vendas no retalho cosmético, com especial enfoque nos desafios trazidos pela transformação digital, pela adoção de inteligência artificial e pela escassez de dados nos lançamentos de novos produtos.

A organização das subsecções acompanha o percurso lógico de um projeto típico de ciência de dados, inspirado nas etapas centrais do modelo CRISP-DM. Partindo de uma contextualização estratégica, evolui-se para a análise das características dos dados disponíveis, passando pelas práticas de preparação e exploração, até à comparação de abordagens preditivas relevantes. Esta estrutura visa proporcionar uma leitura fluida, construindo uma ponte natural entre os fundamentos teóricos e as opções metodológicas adotadas no estudo.

Desta forma, o capítulo não se limita à apresentação da literatura relevante, mas estabelece um fio condutor que antecipa os principais desafios e opções metodológicas tratadas nos capítulos seguintes.

2.1. Compreensão do negócio

2.1.1. Transformação digital e impacto estratégico

A fase de compreensão do negócio requer analisar o contexto estratégico em que o projeto se insere. No caso do retalho cosmético, esse contexto é marcado por uma acelerada transformação digital do setor, que redefine processos e modelos de negócio. A pandemia de COVID-19, em particular, atuou como catalisador desta transformação, forçando os retalhistas a adotarem rapidamente tecnologias digitais e a reforçarem o comércio eletrónico – uma mudança súbita que tende a perdurar (Nanda et al., 2021). Tecnologias como *Internet of Things* (IoT), *blockchain*, *Big Data* e Inteligência Artificial (IA) tornaram-se a base de soluções inovadoras no retalho, enquanto ferramentas de realidade virtual/aumentada e *chatbots* emergem para aprimorar a experiência do cliente. Este panorama digital não apenas melhora a capacidade de resposta às necessidades dos consumidores em tempo real, como também tem sido fundamental para a resiliência e competitividade das empresas num mercado em rápida mudança.

A **Inteligência Artificial (IA)** desempenha um papel central na transformação digital do retalho, incluindo o segmento cosmético. As indústrias de beleza e cuidados pessoais atravessam uma rápida digitalização, em que soluções de IA e *machine learning* impulsionam o envolvimento do consumidor e a eficiência operacional (O'Higgins & Fatorachian, 2025). Estas inovações inserem-se na visão da Indústria 5.0, integrando colaboração homem-máquina para personalizar experiências de cliente e agilizar a tomada de decisões.

Do ponto de vista estratégico, a IA oferece múltiplos benefícios ao retalho. Os retalhistas utilizam-na para aumentar vendas (por via de recomendações personalizadas e campanhas direcionadas) e para otimizar operações em toda a cadeia de valor (Guha et al., 2021). Estudos indicam que a IA pode melhorar a eficiência da cadeia de abastecimento, otimizar operações em loja e tornar pagamentos e atendimento mais eficazes. De facto, a IA já demonstra utilidade em todas as áreas do negócio do retalho, desde o marketing até à gestão logística. Por exemplo, Shankar (2018)

destaca que algoritmos inteligentes contribuem para rentabilizar a estratégia omnicanal, refinando recomendações ao cliente, melhorando a gestão do relacionamento (CRM), e otimizando níveis de inventário e logística para evitar ruturas de stock. No setor cosmético – altamente influenciado por tendências de moda, sazonalidade e preferências voláteis do consumidor – estas capacidades de previsão e personalização suportadas por IA conferem uma vantagem competitiva significativa. O resultado é uma experiência de compra integrada (física e digital) mais atraente e decisões de negócio mais informadas, alinhadas com as expectativas de consumidores cada vez mais digitais.

Com a transformação digital, dados em tempo real tornaram-se um ativo estratégico no retalho. A proliferação de sensores IoT, aplicações móveis e plataformas de comércio eletrónico faz com que os retalhistas disponham de fluxos contínuos de dados atualizados ao minuto sobre vendas, stocks e comportamento do cliente. A capacidade de monitorizar indicadores em tempo real permite detetar rapidamente mudanças na procura e responder de forma ágil – seja ajustando previsões de vendas, alterando preços dinamicamente ou reabastecendo produtos críticos. Estudos recentes evidenciam que a análise de dados e o processamento em tempo real têm um impacto significativo no desempenho dos retalhistas, ao passo que a mera disponibilidade de dados históricos, isoladamente, já não é suficiente (Kameswari et al., 2024). Em outras palavras, a vantagem competitiva reside na velocidade com que as empresas transformam dados correntes em *insights* acionáveis.

No contexto cosmético, as preferências dos consumidores podem mudar em questão de horas, impulsionadas por picos de atenção nas redes sociais e pelas campanhas de influenciadores digitais (Singh et al., 2024). A disponibilização de dados em tempo real — provenientes de pontos de venda, plataformas de comércio eletrónico e monitorização social — permite aos retalhistas detetarem de imediato essas variações e reagirem com campanhas direcionadas ou realocação ágil de stock. Estudos recentes sobre *demand sensing*[2] demonstram que a integração de dados de vendas instantâneos com algoritmos de IA melhora a precisão das previsões de curto prazo e reduz erros de inventário, mas também expõe limitações ligadas à qualidade dos fluxos de dados e à integração com sistemas legados (Douaioui et al., 2024). Assim, apesar dos desafios técnicos — nomeadamente assegurar fiabilidade dos dados em tempo real e compatibilidade entre plataformas antigas e novas —, a evidência aponta que os retalhistas que dominam estas capacidades obtêm vantagem competitiva, minimizando ruturas e otimizando o serviço ao cliente (Douaioui et al., 2024).

A transformação digital tem igualmente amplificado a tomada de decisão baseada em dados (*data-driven*) na gestão da cadeia de abastecimento. Decisões que anteriormente se fundamentavam na experiência ou em históricos limitados são hoje sustentadas por análises avançadas de grandes volumes de dados, muitas vezes integradas em painéis de *Business Intelligence*. Essa mudança de paradigma reflete-se na adoção de sistemas preditivos de previsão de vendas e de ferramentas de apoio à decisão que simulam cenários logísticos. Por exemplo, algoritmos de previsão com IA conseguem analisar padrões de vendas passadas, dados meteorológicos, eventos promocionais e

[2] *Demand sensing* é um método de previsão que combina dados de vendas em tempo real com sinais exógenos (por exemplo, redes sociais, meteorologia) e algoritmos de *machine learning* para gerar *forecasts* de curto prazo muito mais reativos do que os modelos baseados apenas em séries temporais históricas. A sua adoção no setor do retalho evidenciou reduções significativas de ruturas de stock e de inventário (Folinas & Rabi, 2012); (Babai et al., 2022).

até menções em redes sociais para antecipar a procura futura de produtos cosméticos com uma precisão superior à dos métodos tradicionais. Consequentemente, os gestores podem ajustar os níveis de stock e planos de aprovisionamento de forma proativa, reduzindo tanto ruturas como excedentes de inventário. Segundo estimativas reportadas no setor, a incorporação de técnicas de análise avançada pode diminuir erros de previsão e desvios de stock de 20 a 50%, aumentando a eficiência da cadeia logística e a disponibilidade de produtos (Amar et al., 2022).

Além da previsão, a visibilidade *end-to-end* proporcionada por dados integrados fortalece a gestão da cadeia de abastecimento. *Dashboards* atualizados em tempo real permitem acompanhar níveis de stock ao longo de armazéns e lojas, status de entregas de fornecedores e até indicadores de desempenho de transporte. Com isto, as empresas conseguem identificar estrangulamentos ou atrasos e tomar decisões informadas rapidamente, como redirecionar envios ou ativar fornecedores alternativos. O uso estratégico de dados na cadeia de abastecimento também viabiliza modelos inovadores, como reposição contínua e logística preditiva, em que o envio de produtos é acionado automaticamente conforme as vendas ocorrem e atingem certos limiares (Aktas & Meng, 2017). No entanto, essa abordagem *data-driven* exige colaboração interfuncional e investimento em sistemas de informação robustos, capazes de gerir e analisar informação de múltiplas fontes.

Do ponto de vista geográfico, as tendências acima descritas manifestam-se tanto globalmente quanto a nível regional. Na Europa, a digitalização das empresas de retalho – medida pela adoção de novas tecnologias como computação em nuvem, *Big Data* e IoT – tem uma influência comprovada no desempenho económico e na sustentabilidade organizacional (Bocean & Vărzaru, 2023). Em mercados europeus suficientemente desenvolvidos, estas tecnologias exercem efeitos notórios na eficiência e inovação do setor, impulsionando novos modelos de negócio centrados em dados. Portugal, inserido neste contexto, tem também assistido a um crescente investimento na transformação digital do retalho, inclusive no segmento cosmético, embora enfrente desafios como a escassez de talento especializado e a necessidade de atualização contínua de infraestruturas tecnológicas. Iniciativas nacionais e parcerias internacionais têm emergido para promover a adoção da IA e de soluções de análise em tempo real no comércio a retalho, reconhecendo que a capacidade de gerir a informação de forma inteligente é hoje indissociável da vantagem competitiva e da resiliência empresarial (Bocean & Vărzaru, 2023).

Em suma, a transformação digital redefine a estratégia no retalho cosmético ao colocar os dados e a inteligência artificial no cerne da tomada de decisão. A disponibilidade de dados em tempo real e as ferramentas avançadas de análise permitem decisões mais ágeis e precisas, desde o planeamento de stock até ao marketing personalizado. Este impacto estratégico manifesta-se numa cadeia de abastecimento mais integrada e responsiva, capaz de alinhar-se estreitamente com a procura e de melhorar o nível de serviço ao cliente. Na fase de Compreensão do Negócio de um projeto de previsão de vendas (como preconiza o CRISP-DM), torna-se crucial entender este panorama: organizações de retalho que abraçam a transformação digital e exploram plenamente os seus dados tendem a definir objetivos de negócio ambiciosos, alicerçados em análises preditivas e eficiências operacionais. Assim, a previsão de vendas no retalho cosmético deve ser encarada não apenas como um exercício técnico, mas como parte integrante de uma estratégia digital mais ampla, que potencia a competitividade e a inovação num setor em rápida evolução.

2.1.2. Lançamento de novos artigos no retalho

Nos últimos anos, os ciclos de vida dos produtos têm-se tornado progressivamente mais curtos no setor do retalho (Van Steenbergen & Mes, 2020). Nesse contexto de constante renovação, o lançamento de novos artigos constitui um vetor crítico de competitividade. O setor cosmético, em particular, destaca-se pela intensidade da inovação e pela volatilidade das preferências dos consumidores (Zineb et al., 2025), impondo introduções frequentes de itens inéditos para conquistar ou manter quota de mercado.

Cada novo lançamento exige, todavia, estimativas fiáveis de procura mesmo na ausência de histórico de vendas. Van Steenbergen & Mes (2020) sublinham que, apesar da escassez de dados, as previsões pré-lançamento são essenciais para orientar decisões operacionais, nomeadamente o planeamento da capacidade produtiva, o aprovisionamento de matérias-primas e a gestão de inventário. Estimativas imprecisas podem conduzir tanto a ruturas de stock como a excedentes, afetando negativamente a rentabilidade do negócio e a satisfação do cliente (Van Steenbergen & Mes, 2020). Consequentemente, as empresas recorrem cada vez mais a métodos quantitativos e analíticos para enfrentar a incerteza inerente ao lançamento de novos artigos. As previsões iniciais não devem ser apenas pontuais; devem contemplar a variabilidade possível da procura para antecipar riscos e definir margens de segurança adequadas. Para tal, empregam-se técnicas que vão desde analogias com produtos semelhantes até algoritmos de *machine learning,* sempre com o intuito de calibrar o lançamento nos planos estratégico e operacional do negócio.

No retalho contemporâneo, as previsões funcionam em múltiplos níveis hierárquicos. Por um lado, previsões agregadas (total de vendas esperado) dão suporte a decisões estratégicas acerca do portefólio/sortido de artigos e de abastecimento em larga escala. Por outro, previsões altamente granulares (por *SKU* e por loja) orientam o reabastecimento local e a alocação de stock em cada ponto de venda. Muitas cadeias de retalho integram atualmente sistemas analíticos que agregam históricos de vendas, padrões sazonais e dados de comportamento do consumidor num único *framework* de planeamento da procura (Basavaraju & Fatahi Valilai, 2025). Basavaraju & Fatahi Valilai (2025) demonstram que tal arquitetura analítica maximiza a capacidade de satisfazer a procura ao combinar informação proveniente de diversas fontes, evidenciando a necessidade de flexibilidade na granularidade dos *forecasts*: enquanto a previsão ao nível *SKU*/loja permite personalizar inventário e promoções locais, previsões a níveis superiores orientam decisões estratégicas relativas ao sortido de artigos a disponibilizar — ou seja, o conjunto total de artigos considerados — bem como à cobertura geográfica.

Iniciativas recentes ilustram estas tendências. No setor da moda, por exemplo, Sousa et al. (2025) propõem um método preditivo em duas etapas para novas coleções: primeiro, reconstruir distribuições de procura a partir de vendas históricas (corrigindo efeitos de rotura); depois, aplicar algoritmos avançados—*Random Forest*, redes neuronais, entre outros—para estimar a procura dos produtos inéditos. Os resultados indicam que técnicas como o algoritmo de *Expectation-Maximization* (para tratar dados censurados) conjugado com *Random Forest* proporcionam as previsões mais precisas, contribuindo para decisões de produção melhor alinhadas às condições de mercado (Sousa et al., 2025). Estes estudos reforçam como modelos analíticos sofisticados, aplicados à granularidade adequada, podem transformar dados limitados em planos operacionais robustos.

Em síntese, o lançamento de novos artigos no retalho encerra desafios estratégicos e operacionais indissociáveis, sempre sob a égide da previsão da procura. Esta secção estabeleceu o enquadramento conceptual do tema, salientando a importância do *forecast* na ponte entre a estratégia de inovação de produto e a execução quotidiana das operações. Nas subseções seguintes, aprofunda-se esse enquadramento.

2.2. Compreensão dos dados

2.2.1. Granularidade SKU/Loja

A relação entre o produto (SKU) e o ponto de venda (loja) representa um eixo fundamental na gestão operacional e comercial no setor do retalho. Esta relação traduz-se na capacidade de compreender, ao mais alto nível de granularidade, o desempenho de cada referência comercial em cada localização, permitindo decisões mais ajustadas nas áreas de previsão de procura, gestão de inventário, definição de preços e campanhas promocionais.

A granularidade SKU/loja oferece uma visão operacional rica, particularmente relevante num contexto onde a diferenciação por perfil de loja ou cliente é crítica para a competitividade (Ying et al., 2021).

A granularidade operacional refere-se à capacidade de análise detalhada ao nível de produtos específicos em pontos de venda distintos. Como demonstram Ying et al. (2021), o uso de tecnologias de *Big Data* tem possibilitado a integração de vendas históricas, *feedback* de clientes e variáveis operacionais para gerar perfis de comportamento por loja. Esta abordagem permite identificar padrões de consumo localizados que seriam invisíveis em análises agregadas, como preferências regionais de produtos ou a eficácia de campanhas específicas numa loja em particular.

Segundo Hossain et al. (2023), a capacidade analítica sustentada com base em dados ao nível *SKU*/loja contribui significativamente para a eficácia de mercado, ao permitir uma segmentação mais fina e decisões logísticas otimizadas. Esta ideia é reforçada por Vorhies & Morgan (2005), que afirmam que a eficácia de mercado está diretamente ligada à capacidade da empresa em ajustar a sua proposta de valor com base em informação granular, específica e contextualizada.

A compreensão detalhada da performance ao nível *SKU*/loja é também crucial para a gestão do ciclo de vida dos produtos. Heidenreich et al. (2022, cit. em Tseng et al. 2022) salientam que os consumidores atuais tendem a rejeitar inovações que não se ajustem às suas necessidades específicas, sendo necessária uma análise localizada e informada para orientar os lançamentos de produtos e estratégias de substituição.

A granularidade *SKU*/loja é, por fim, essencial na construção de modelos preditivos eficazes. Thivakaran & Ramesh (2022) demonstram que algoritmos como o *XGBoost* só alcançam o seu potencial quando alimentados com dados organizados a este nível de detalhe. A capacidade de antecipar padrões de procura por loja e produto permite minimizar ruturas de stock, evitar excessos e melhorar a alocação de recursos, contribuindo para uma operação mais eficiente e sustentável.

Além disso, esta granularidade permite personalizar a experiência do cliente. De acordo com Belarbi et al. (2016, cit. em Ying et al. 2021), a análise *SKU*/loja permite ajustar não só o portefólio disponível, mas também a comunicação e o serviço prestado, aumentando a relevância da proposta de valor para cada perfil de consumidor.

2.2.2. Caraterização de produtos com dados históricos limitados

A caracterização e previsão de desempenho de produtos com **dados históricos escassos** é um desafio recorrente no retalho, particularmente evidente em segmentos como cosméticos e moda, onde novos artigos são lançados frequentemente e os ciclos de vida tendem a ser curtos (Van

Steenbergen & Mes, 2020). Nestas situações, não há dados de vendas passadas suficientes para usar métodos tradicionais de extrapolação temporal. Ainda assim, decisões cruciais de negócio (compras, gestão de stock, etc.) dependem de previsões iniciais fiáveis, pelo que é necessário adotar abordagens especializadas que contornem a falta de histórico (Van Steenbergen & Mes, 2020). Estudos recentes enfatizam a importância de técnicas analíticas inovadoras para colmatar esta lacuna, notando inclusive a escassez de pesquisas focadas no setor de cosméticos (Zineb et al., 2025).

Historicamente, muitas empresas recorreram à **analogia com produtos similares e julgamento especializado** para prever a procura de novos SKU, dadas as poucas informações disponíveis (Van Steenbergen & Mes, 2020). Por exemplo, um novo creme facial poderia ter sua procura estimada tomando como referência as vendas iniciais de um creme parecido lançado no passado. Embora simples, esta estratégia de "produto análogo" é bastante difundida na prática (Van Steenbergen & Mes, 2020). No entanto, carece de rigor estatístico e não quantifica a incerteza associada. Para melhorar esse ponto, a literatura recente tem formalizado métodos de analogia mais robustos. Uma linha de investigação combina **múltiplos produtos comparáveis**: Guo et al. (2025) propõem um *ensemble Bayesiano*[3] de perfis de ciclo de vida, onde o padrão de vendas de um novo item é previsto a partir de um conjunto de curvas de produtos similares, ponderadas segundo um modelo *Bayesiano*. Esta abordagem permite estimar não apenas um ponto de previsão, mas também uma distribuição de possíveis trajetórias de vendas, aumentando a robustez das estimativas. Os resultados demonstraram capacidade de **prever todo o perfil de vida de novos produtos**, capturando a incerteza e melhorando a precisão em categorias de ciclo de vida curto, como cosméticos e *fast fashion*. Em suma, métodos baseados em analogias – quando bem fundamentados – continuam centrais, mas agora incorporados em *frameworks* estatísticos que oferecem estimativas fiáveis e intervalos de confiança para novos lançamentos.

Outra classe de metodologias explora **algoritmos de aprendizagem automática** para compensar a falta de histórico, aproveitando dados de produtos relacionados e atributos dos artigos. Por exemplo, Van Steenbergen & Mes (2020) desenvolveram o modelo ***Demand Forest***, que combina **clusterização *k-means*** de séries de produtos similares com ***Random Forest*** para transferência de padrões e ***Random Forest* de regressão quantílica** para estimar a distribuição da procura. Essencialmente, os produtos já existentes são agrupados segundo perfis de vendas semelhantes, e esses grupos alimentam um modelo de *Random Forest* que integra também **características dos produtos** (categoria, preço, etc.) do novo item, gerando assim uma previsão calibrada para o lançamento (Van Steenbergen & Mes, 2020). Esta abordagem híbrida mostrou ganhos expressivos de precisão e reduções de até 15% nos stocks em falta ou em excesso, comparada a métodos *benchmark* tradicionais. De modo semelhante, técnicas de **aprendizagem profunda** têm vindo a ser aplicadas para extrair sinais úteis de dados auxiliares. Ekambaram et al. (2020), por exemplo, propuseram um modelo de rede neural com mecanismo de atenção que integra **dados multi-modais**– combinando imagens do produto, descrições textuais e dados de vendas de artigos análogos – para prever as vendas de novos artigos de moda (Anitha S. & Neelakandan R., 2025). O uso de atributos visuais e de contexto permite ao modelo aprender padrões de sucesso de produtos anteriores e aplicá-los no contexto de um novo SKU mesmo com histórico praticamente nulo. De forma geral, **técnicas de *Machine Learning*, incluindo *deep learning* e *ensembles***, têm

[3] O adjetivo *Bayesiano* designa qualquer método estatístico que se baseia no teorema de Bayes, o qual atualiza uma crença (a *priori*) à luz de novos dados para produzir uma distribuição a *posteriori*.

demonstrado potencial para mitigar o problema de dados limitados, aumentando a precisão das previsões ao explorar grandes bases de dados de produtos e até informações em tempo real (Anitha S. & Neelakandan R., 2025). Estas abordagens conseguem capturar efeitos de tendências de mercado e preferências dos consumidores (por exemplo, a influência de redes sociais ou pesquisas *online*) que seriam impossíveis de modelar pelos métodos tradicionais univariados.

No contexto específico do **retalho de cosméticos**, começam a emergir aplicações destas técnicas de IA para melhorar as previsões. Souza et al. (2020) avaliaram modelos baseados em **sistemas de inferência *fuzzy*** para prever vendas de uma linha de vernizes para unhas, comparando-os com o método intuitivo usado pela empresa. Os modelos *fuzzy* incorporam conhecimento heurístico e regras linguísticas, sendo úteis em cenários com poucos dados numéricos. Neste estudo, as ferramentas de lógica *fuzzy* superaram as previsões informais dos gestores, especialmente para produtos em **fase de crescimento acentuado** (séries com forte assimetria à esquerda e sem tendência central clara) e para horizontes de longo prazo (Souza et al., 2020). A capacidade dos modelos *fuzzy* de lidar com não-linearidades e variabilidade alta ajudou a reduzir vieses humanos e erros de previsão nesses casos (Souza et al., 2020). Esses resultados sugerem que técnicas estatísticas não convencionais, como este modelo ou mesmo modelos híbridos, podem trazer ganhos quando os dados históricos são escassos e os padrões de venda são atípicos, como ocorre com frequência em lançamentos "estrela" de cosmética.

Por fim, destaca-se o recurso a ***transfer learning* e modelos hierárquicos** para novos produtos. Nessa abordagem, aproveita-se informação de séries relacionadas (por exemplo, de uma categoria de produto ou de outras lojas) para **"transferir" conhecimento** para a previsão do item com poucos dados. Lei et al. (2023) desenvolveram uma estrutura de *transfer learning* que integra os dados agregados ao nível da categoria nas previsões ao nível do SKU. Em termos práticos, as vendas totais da categoria (mais estáveis e abundantes) atuam como uma forma de **regularização** ao modelar a procura de cada produto individual, evitando sobreajuste quando há pouco histórico específico (Lei et al., 2023). Testada em milhares de SKUs de uma grande retalhista online, esta técnica de ***pooling* de dados** alcançou melhorias de mais de **9% na acurácia preditiva** face a modelos que tratam cada série isoladamente (Lei et al., 2023). A utilização combinada de informações de alto nível (categoria) e baixo nível (SKU) revelou, portanto, ser uma via eficaz para reforçar previsões em cenários de dados esparsos. Estratégias análogas foram exploradas por outros autores: por exemplo, Karb et al. (2020) demonstraram que ao **pré-treinar redes neuronais** em produtos existentes e depois ajustá-las a um produto novo, obtém-se ganhos significativos na precisão das previsões iniciais em retalho alimentar. Em todos estes casos, o princípio é comum – **partilhar padrões** de séries com histórico rico para iluminar o comportamento esperado de uma série nova ou com poucos registos.

Em síntese, a literatura recente propõe diversas **metodologias inovadoras** para caracterizar e prever o desempenho de produtos com histórico limitado no retalho. Entre elas incluem-se: (a) métodos de **analogia** aprimorados, que utilizam dados de produtos comparáveis de forma estruturada (às vezes combinando múltiplos análogos); (b) algoritmos de **aprendizagem automática** e ***deep learning*** que integram atributos do produto, dados de mercado e técnicas de agrupamento para extrapolar padrões; (c) abordagens de ***transfer learning*** e **modelos hierárquicos**, que agregam informação de alto nível ou de produtos similares para reforçar previsões de séries escassas; e (d) técnicas **estatísticas robustas** (como sistemas *fuzzy* ou *ensembles Bayesianos*) capazes de capturar a incerteza e a variabilidade extrema típicas dos lançamentos. Todas estas estratégias visam contornar a ausência de histórico direto, **tirando partido de dados**

auxiliares – seja o histórico de outros produtos, características intrínsecas do novo artigo, ou conhecimentos a *priori* do domínio – de modo a melhorar a exatidão da previsão de procura em nível granular (SKU/loja) (Anitha S. & Neelakandan R., 2025). Os estudos académicos dos últimos anos reportam resultados promissores, com melhorias consistentes na precisão preditiva e na gestão de stocks quando comparados aos métodos empíricos tradicionais (Lei et al., 2023).

Deste modo, alinha-se a investigação atual com as necessidades práticas do retalho moderno: **prever a procura de novos produtos com confiança**, mesmo quando os dados históricos são limitados ou inexistentes, minimizando riscos no lançamento de artigos e contribuindo para decisões mais informadas em toda a cadeia de abastecimento.

2.3. Preparação dos dados

2.3.1. Análise Exploratória de Dados (EDA)

A análise exploratória de dados (*Exploratory Data Analysis, EDA*) constitui uma etapa fundamental nos processos de ciência de dados, particularmente em ambientes de elevada complexidade informacional como o retalho. A sua função é dupla: permite, por um lado, compreender a estrutura interna dos dados, identificando padrões, *outliers*, sazonalidades ou relações entre variáveis, e por outro, fundamentar decisões metodológicas subsequentes, nomeadamente na seleção de atributos e na escolha de algoritmos preditivos. De acordo com Zikopoulos et al., (2012, cit. em Ying et al., 2021), o crescimento exponencial do volume de dados disponíveis nas organizações — estruturados e não estruturados — exige abordagens exploratórias sistemáticas que substituam a modelação baseada em hipóteses a *priori*.

Em contextos de grande granularidade, como o retalho multiloja com milhares de unidades de manutenção de stock (*SKUs*), a EDA adquire um papel estratégico. A aplicação sistemática de técnicas exploratórias permite não só a limpeza e transformação dos dados, mas também a deteção de comportamentos sazonais, impactos de campanhas promocionais e padrões de vendas por localização ou tipo de loja. Thivakaran & Ramesh (2022) demonstram que, num estudo com dados de retalho, o recurso a métodos de *EDA* foi determinante para a identificação das variáveis mais relevantes na previsão de vendas, como o tipo de loja, localização geográfica e categoria do produto, o que se traduziu em ganhos significativos de desempenho ao aplicar algoritmos como *Random Forest* e *XGBoost*.

Para além da sua relevância técnica, a *EDA* é também reconhecida como um pilar para a construção de capacidades analíticas organizacionais. Segundo Hossain et al. (2023), empresas que integram sistematicamente práticas de *EDA* conseguem antecipar tendências de mercado, adaptar a sua estratégia comercial com maior rapidez e aumentar a sua competitividade. Este enquadramento posiciona a análise exploratória não apenas como um passo preparatório, mas como um ativo estratégico capaz de sustentar decisões informadas e resilientes.

A evolução da *EDA* acompanha a transformação digital das organizações. Tal como referem Sapountzi & Psannis (2018, cit. em Tseng et al. (2022)), o desafio central da análise baseada em *Big Data* não reside apenas na recolha de informação, mas na sua interpretação útil e operacional. A diversidade e complexidade das fontes de dados atuais — que incluem registos transacionais, sensores, redes sociais e dados de localização — exige métodos analíticos capazes de sintetizar padrões ocultos e traduzir informação em ações concretas. Neste sentido, a *EDA* contribui para a agilidade organizacional.

Tseng et al. (2022) defendem que empresas com estruturas analíticas ágeis e orientadas por dados são mais capazes de responder a variações do mercado, antecipar riscos e acelerar o ciclo de decisão. A análise exploratória, ao promover uma leitura crítica dos dados e ao evidenciar relações não triviais, atua como um mecanismo de aprendizagem contínua e adaptação estratégica, especialmente relevante no lançamento de novos produtos e na gestão de campanhas comerciais.

A análise exploratória contribui ainda para mitigar riscos associados à tomada de decisão em ambientes incertos. Baum et al. (2019, cit. em Tseng et al. (2022)) mostram que decisões apoiadas por dados — e precedidas por uma exploração sistemática dos mesmos — tendem a apresentar maiores probabilidades de sucesso em contextos de inovação no retalho. Assim, a *EDA* não é

apenas um conjunto de técnicas estatísticas, mas uma componente essencial da cultura analítica das organizações modernas.

A estrutura analítica adotada neste projeto assume a *EDA* como etapa preliminar essencial para a construção de modelos robustos de previsão, permitindo uma adaptação metodológica progressiva às particularidades dos dados de cada *SKU*.

2.4. Modelação

2.4.1. Visão geral dos modelos preditivos

A previsão de vendas no retalho desempenha um papel crucial no planeamento comercial, permitindo alinhar inventários e recursos com a procura esperada. Nos últimos anos, verificou-se um crescente interesse em abordagens de previsão baseadas em dados, dada a maior disponibilidade de dados de vendas e ferramentas analíticas avançadas. Diversas técnicas têm sido aplicadas à previsão de séries temporais de vendas, podendo-se categorizá-las em três grandes grupos: 1) métodos estatísticos tradicionais; 2) métodos de *Machine Learning* (ML, ou aprendizagem automática) e 3) métodos de *Deep Learning* (DL, ou aprendizagem profunda).

Na presente secção, analisam-se modelos representativos de cada abordagem, pelo que se discutem os seus fundamentos, aplicações ao contexto do retalho e evidências bibliográficas sobre as razões da sua aplicabilidade.

Nas subsecções seguintes são detalhadas as principais abordagens de modelos preditivos aplicadas à previsão de vendas no setor do retalho, classificadas em métodos estatísticos, de ML e de DL. Serão explicados os modelos *ARIMA* e *Prophet* como exemplos de técnicas estatísticas tradicionais, seguidos da apresentação de técnicas de ML com destaque para o *XGBoost e LightGBM*, e por fim técnicas de DL com ênfase em redes *LSTM* e no modelo *Transformer*.

Em cada subsecção, justifica-se bibliograficamente a relevância de cada método para o problema em análise.

2.4.2. Métodos estatísticos tradicionais

A previsão de séries temporais evoluiu significativamente ao longo do século XX, sendo inicialmente dominada por modelos estatísticos clássicos. Um dos marcos fundacionais foi a metodologia *Box-Jenkins,* proposta por Box e Jenkins no final da década de 1960, que sistematizou a aplicação dos modelos ARIMA (*AutoRegressive Integrated Moving Average*). Estes modelos combinam três componentes principais: a auto regressão (AR), a integração (I) — associada à diferenciação necessária para tornar a série estacionária — e a média móvel (MA), que modela os erros passados da série (Wilson, 2016).

Os modelos ARIMA são definidos pelos parâmetros p, d, q, respetivamente associados à ordem autorregressiva, à ordem de integração e à ordem da média móvel.

$$\textbf{ARIMA}(p,d,q): \quad \left(1-\sum_{i=1}^{p}\phi_i B^i\right)(1-B)^d y_t = \left(1+\sum_{j=1}^{q}\theta_j B^j\right)\varepsilon_t$$

A metodologia envolve tipicamente três fases iterativas: identificação da estrutura do modelo, estimação dos parâmetros e validação através da análise dos resíduos. Apesar de exigirem conhecimento estatístico para uma correta implementação, os ARIMA mantêm-se, ainda hoje, como referência em diversos domínios da previsão, nomeadamente no retalho (Hyndman & Athanasopoulos, 2021).

No setor do retalho, têm sido aplicadas variantes do ARIMA, como o SARIMA (que incorpora sazonalidade) e o ARIMAX (com inclusão de variáveis exógenas), para capturar padrões recorrentes

em vendas, como picos sazonais associados a festividades (Aras et al., 2017). No entanto, estes modelos assumem uma relação linear e exigem que cada série seja tratada de forma univariada, o que limita a escalabilidade quando se pretende prever milhares de séries em simultâneo (Makridakis, Spiliotis, et al., 2022).

Paralelamente aos ARIMA, os métodos de alisamento exponencial — como *Holt-Winters* — foram desenvolvidos para capturar tendências e sazonalidades com atualização recursiva. Estes métodos deram origem à família ETS (Error-Trend-Seasonal), cuja formulação em espaço de estados permite uma decomposição estruturada da série temporal em componentes latentes (Hyndman & Athanasopoulos, 2021).

Apesar da sua solidez teórica e ampla utilização, tanto os modelos ARIMA como os ETS apresentam limitações face às exigências do retalho contemporâneo, caracterizado por grande volume de dados, múltiplas variáveis explicativas e necessidade de previsões automáticas em larga escala. Neste contexto, métodos estatísticos univariados tornam-se impraticáveis, quer pela carga computacional, quer pela necessidade de intervenção humana na afinação dos modelos (Hyndman & Athanasopoulos, 2021).

Para superar estas limitações, surgiram abordagens estatísticas mais automatizadas e flexíveis, entre as quais se destaca o modelo *Prophet*, desenvolvido por Taylor & Letham (2018). O *Prophet* baseia-se numa decomposição aditiva da série temporal em componentes de tendência, sazonalidade e feriados. Este modelo foi concebido para ser utilizado por analistas de negócio sem formação estatística avançada, oferecendo maior grau de automatização e facilidade na inclusão de eventos especiais, como promoções ou feriados nacionais.

Ao contrário dos ARIMA, o *Prophet* permite a modelação de tendências com pontos de inflexão (*changepoints*), múltiplas sazonalidades em simultâneo (como padrões semanais e anuais) e integração direta de eventos externos através de variáveis *dummy*[4]. A sua robustez a valores em falta e a *outliers*, bem como a sua capacidade de gerar intervalos de previsão de forma automática, tornam-no uma opção prática para contextos empresariais (Taylor & Letham, 2018). Contudo, diversos estudos empíricos reportam que, apesar da facilidade de uso, o *Prophet* pode apresentar desempenho inferior ao de modelos clássicos afinados manualmente, como o ARIMA, ou a modelos baseados em aprendizagem automática e redes neuronais, especialmente em séries com padrões altamente não lineares (Menculini et al., 2021).

Ainda assim, tanto o ARIMA como o Prophet continuam a ser modelos amplamente utilizados como *baseline* ou referência comparativa na literatura científica. Estudos como os de Aras et al. (2017) ou Makridakis, Spiliotis, et al. (2022) demonstram que estas abordagens continuam a ser competitivas, sobretudo em séries com padrões lineares, dados limpos e com histórico consistente. No entanto, para tarefas mais exigentes — como a previsão de milhares de *SKUs* com variações não lineares, dependências temporais complexas e múltiplas variáveis explicativas — a escalabilidade e a flexibilidade tornam-se cruciais.

Por esse motivo, no presente estudo, os modelos ARIMA e *Prophet* foram considerados apenas como enquadramento teórico e metodológico, mas não foram implementados. Esta decisão deve-

[4] **Variável *dummy:*** Indicador binário (0/1) que assinala a presença de um evento externo (feriado, promoção, etc.); ao ser incluída como regressor, desloca a linha de base da série apenas nos períodos em que vale 1, permitindo ao modelo capturar o impacto específico desse evento.

se à sua menor capacidade de generalização em cenários multivariados e à dificuldade de escalar a sua aplicação para grandes volumes de séries temporais. A implementação prática foi assim orientada para modelos de *machine learning* e *deep learning*, que serão discutidos nas secções seguintes.

2.4.3. Técnicas de *Machine Learning*

A aprendizagem automática (do inglês *Machine Learning*, ML) consiste no desenvolvimento de algoritmos capazes de aprender padrões a partir de dados históricos para fazer previsões ou decisões sem serem explicitamente programados para cada tarefa. Em previsão de vendas no retalho, as técnicas de ML oferecem uma abordagem flexível que vai além dos modelos estatísticos tradicionais, pois conseguem capturar relações não lineares e interações complexas entre múltiplos fatores (por exemplo, promoção, preço, sazonalidade), muitas vezes com melhor desempenho preditivo.

Historicamente, o campo de ML consolidou-se nas últimas décadas do séc. XX, abrangendo métodos diversos como árvores de decisão, máquinas de vetores de suporte e *ensembles* (conjuntos de modelos). Em contrapartida, o *Deep Learning* (DL) emergiu de avanços em redes neuronais multi-camada, ganhando proeminência apenas no início do séc. XXI graças à maior disponibilidade de dados e poder computacional (LeCun et al., 2015)

Apesar de o DL ser hoje considerado parte do ML, por norma diferencia-se o ML "tradicional", isto é, modelos menos profundos que geralmente requerem engenharia manual de características, do DL, que se definem como modelos de redes profundas com múltiplas camadas que aprendem representações de alto nível diretamente dos dados (Schmidhuber, 2015).

No contexto de previsão de séries temporais de vendas, modelos de ML tipicamente envolvem transformar o problema temporal em um conjunto de variáveis preditoras (atributos) extraídas do histórico de vendas (*lags*[5], médias móveis, indicadores sazonais, feriados, etc.) para então aplicar algoritmos de regressão supervisionada. Entre estes algoritmos, os métodos de *ensemble* têm se destacado por combinarem múltiplos estimadores de forma a melhorar a precisão e robustez. Um exemplo notável é o *gradient boosting*, técnica introduzida por Friedman (2001) que constrói uma sequência de árvores de decisão de forma aditiva, cada nova árvore corrigindo os erros residuais do conjunto anterior.

O algoritmo *XGBoost* (*eXtreme Gradient Boosting*) proposto por Chen & Guestrin (2016) tornou-se uma implementação otimizada e amplamente adotada do *gradient boosting*, alcançando resultados de ponta em diversas tarefas preditivas. O *XGBoost* introduziu melhorias importantes, como regularização explícita (L1 e L2) para evitar sobreajuste, uso eficiente de esparsidade (valores omissos e variáveis categóricas) e otimizações de hardware que permitem treino paralelo e escalável mesmo em bases de grande dimensão (Chen & Guestrin, 2016). Essas características fazem do *XGBoost* um modelo especialmente atrativo para dados de retalho, que muitas vezes incluem muitas lojas e produtos com dados heterogéneos e ruidosos (Andrade & Cunha, 2023).

[5] ***Lag*****:** variável que representa o valor de uma série temporal em momentos anteriores (por exemplo, vendas no dia anterior ou na semana passada), utilizada como preditor para capturar dependências temporais e padrões recorrentes.

Acrescenta-se ao prescrito que estudos recentes demonstram a eficácia do *XGBoost* em cenários de previsão de vendas no retalho. Andrade & Cunha (2023) aplicaram o *XGBoost* para prever vendas diárias desagregadas por produto e loja, reportando desempenho superior a modelos tradicionais utilizados no setor (como modelos base-lineares do tipo *Base-Lift*), com ganhos significativos de precisão. De modo semelhante, os resultados da competição M5 (Makridakis, Petropoulos, et al., 2022), focada em previsões de vendas de retalho, indicaram que métodos baseados em árvores de decisão *gradient boosting* (e.g., *XGBoost* ou o similar *LightGBM*) estiveram presentes nas soluções de melhor desempenho, muitas vezes combinados com engenharia de atributos especializada.

Estas evidências sugerem que modelos de ML como o *XGBoost* conseguem capturar de forma eficaz padrões de vendas complexos (incluindo tendências e efeitos de promoções) quando alimentados com atributos relevantes, apresentando boa capacidade preditiva mesmo com dados relativamente escassos ou em cenários de curto prazo (Loureiro et al., 2018). Ademais, o *XGBoost* tende a ser mais rápido de treinar e a fornecer explicações mais diretas (por exemplo, através da importância das variáveis) em comparação com redes neuronais profundas, o que pode ser vantajoso em ambientes de retalho onde interpretabilidade e tempo de resposta são fatores a considerar.

Destaca-se ainda o *LightGBM* (*Light Gradient Boosting Machine*), proposto por Ke et al. (2017), que se diferencia do *XGBoost* pelo uso de histogramas discretos na divisão dos dados e por adotar uma estratégia de crescimento por folhas (*leaf-wise*) em vez do crescimento por nível (*level-wise*) – comparação gráfica visível na Figura 1. Esta abordagem permite ganhos substanciais em velocidade de processamento, sendo frequentemente várias vezes mais rápida que outras implementações baseadas em árvores, mantendo uma precisão comparável (Ke et al., 2017; Ileri, 2025).

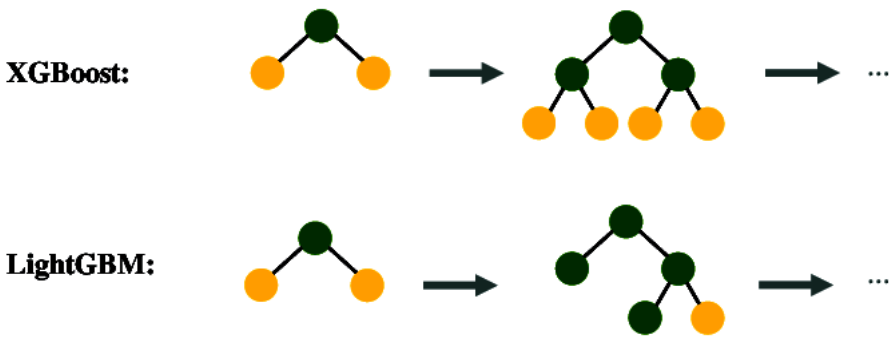


Figura 1 – Estratégia de crescimento por nível (level-wise) vs. crescimento folhas (leaf-wise) - XGBoost vs. LightGBM

O *LightGBM* tem sido amplamente adotado em contextos de previsão no retalho, não só pela sua rapidez, mas também pela flexibilidade no controlo da complexidade do modelo, através de parâmetros como o número de folhas (*num_leaves*) e a taxa de amostragem (*subsample*) (Ileri, 2025). A evidência empírica sugere que, quando devidamente afinado, o *LightGBM* atinge desempenhos equiparáveis aos do *XGBoost*, com a vantagem adicional de tempos de computação significativamente mais reduzidos em cenários com grandes volumes de dados (Ke et al., 2017; Kizgin et al., 2025).

De acordo com Theodoridis & Tsadiras (2024), modelos de *boosting* como o *XGBoost* e o *LightGBM* constituem alternativas altamente promissoras em contextos multivariados de previsão de procura, nomeadamente pela sua eficácia comprovada em diversos domínios aplicacionais.

Em estudos recentes focados em previsão de vendas em contextos de instabilidade ou crise, observou-se que estes modelos apresentam desempenhos superiores em categorias de produtos com volumes médios de venda (Kizgin et al., 2025)

Em suma, as técnicas de *Machine Learning* clássicas, oferecem modelos ágeis e de alto desempenho para previsão de vendas ao nível do retalho.

2.4.4. Técnicas de *Deep Learning*

O termo *Deep Learning* (DL) refere-se a modelos de aprendizagem profunda por redes neuronais artificiais com múltiplas camadas hierárquicas de neurónios, capazes de extrair automaticamente características de alto nível dos dados brutos (LeCun et al., 2015). A principal diferença estrutural em relação aos métodos de ML tradicionais é que, no DL, as representações relevantes (padrões temporais, sazonais, etc.) são aprendidas internamente pelas várias camadas da rede durante o treino, em vez de serem pré-definidas manualmente. Historicamente, as redes neuronais surgiram em meados do séc. XX (com o *perceptron* de 1958 como precursor), mas a "aprendizagem profunda" somente ganhou tração por volta de 2010 com sucessos em reconhecimento de imagem e voz, impulsionados por melhorias no algoritmo de retro propagação, arquiteturas mais profundas e aceleração por *GPUs* (Schmidhuber, 2015).

Acresce ainda que as redes profundas permanecem, em grande parte, *black boxes*: os mecanismos internos de decisão são opacos, o que pode dificultar a adoção em cenários de negócio regulados. A literatura de *Explainable AI*[6] alerta para a necessidade de modelos intrinsecamente interpretáveis ou de explicações fiáveis para justificar previsões em aplicações críticas (Rudin, 2019).

Definir a topologia de uma rede (n.º de camadas, neurónios por camada, funções de ativação) e calibrar hiperparâmetros-chave (taxa de aprendizagem, *batch size*, regularização, etc.) constitui um problema de otimização num espaço hiperdimensional, não linear e multimodal. Na prática, recorrem-se a estratégias de pesquisa empírica — por exemplo, *random search*, otimização *bayesiana* ou algoritmos evolutivos — que reduzem o esforço, mas não garantem otimização global e podem implicar custos computacionais elevados. No caso de séries temporais, esses métodos exigem validação temporal (*walk-forward* ou *cross-validation* bloqueada) para evitar fuga de informação e *overfitting* (Bergstra et al., 2012; Snoek et al., 2012; Hewamalage et al., 2021)

No contexto de previsão de séries temporais, o DL trouxe a promessa de modelar relacionamentos complexos de longa duração que extrapolam as capacidades dos modelos lineares e de ML superficial. Em particular, *recurrent neural network* (*RNNs*) foram adaptadas para modelar dados sequenciais, incorporando dependências temporais de forma dinâmica.

No entanto, as primeiras *RNNs* sofriam do problema do gradiente de desaparecimento ao aprender padrões de longo prazo. Uma inovação crucial para lidar com dependências de longo alcance em sequências foi a arquitetura *LSTM (Long Short-Term Memory),* proposta por Hochreiter & Schmidhuber (1997), conforme se verifica na Figura 2.

[6] ***Explainable Artificial Intelligence*** **(XAI)** refere-se a métodos que tornam compreensíveis as decisões de modelos de aprendizagem automática, seja porque o próprio modelo é transparente (p. ex., árvore de decisão rasa) ou porque se aplica uma técnica pós-hoc como SHAP ou LIME para justificar cada previsão (Rudin, 2019).

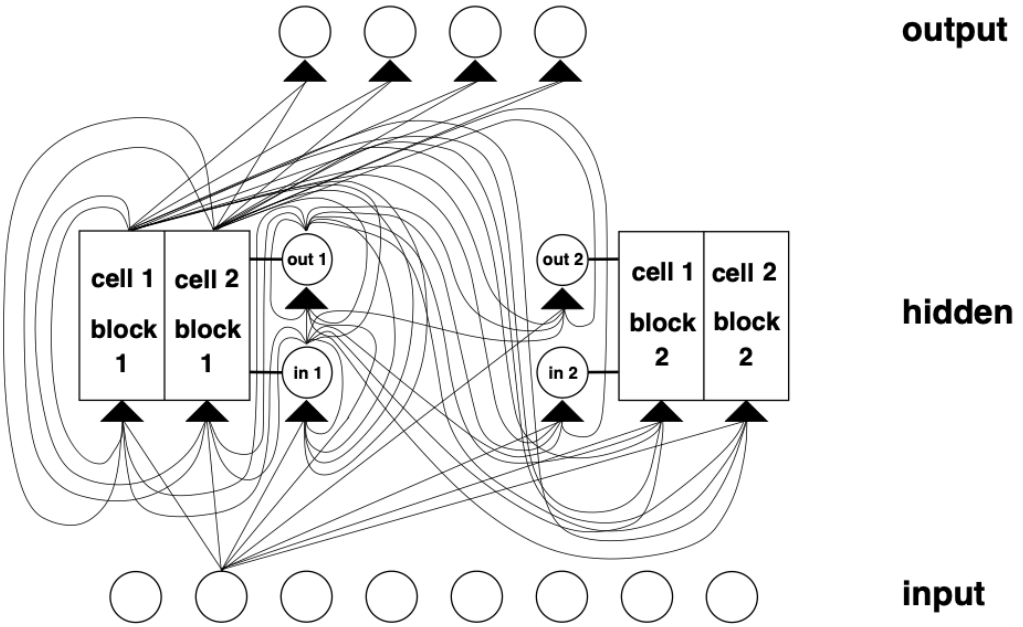


Figura 2 – Arquitetura de rede LSTM com dois blocos de memória de tamanho 2. Adaptado de Long Short-Term Memory, de S. Hochreiter e J. Schmidhuber (1997)

As redes *LSTM* introduzem células de memória com portas de entrada, saída e esquecimento, permitindo armazenar e relembrar informações ao longo de intervalos de tempo extensos de forma controlada. Assim, a *LSTM* consegue aprender padrões temporais de diferentes escalas – por exemplo, capturar tanto flutuações diárias quanto tendências sazonais mensais nas vendas – mitigando o problema do gradiente de desaparecimento. Com o advento de grandes bases de dados temporais e maior poder computacional nos anos 2010, as *LSTM* tornaram-se uma técnica padrão para previsão de séries temporais complexas em diversas áreas. Na previsão de vendas no retalho, vários estudos demonstraram a superioridade ou pelo menos competitividade de redes *LSTM* face a métodos tradicionais em cenários com dinâmicas não lineares e interações difíceis de modelar explicitamente (Hewamalage et al., 2021).

Por exemplo, o método vencedor da competição M4 em 2018 combinou exponenciação de suavização com redes do tipo *LSTM*, alcançando resultados notavelmente melhores do que abordagens puramente estatísticas (Smyl, 2020). Em aplicações específicas de retalho, redes *LSTM* têm mostrado vantagem sobretudo quando existem padrões de curta e média duração nos dados: um estudo comparativo reportou que a *LSTM* superou modelos ARIMA (SARIMA) para produtos com procura relativamente estável, enquanto para séries altamente sazonais o método estatístico mostrou melhor desempenho (Falatouri et al., 2022). Ainda assim, abordagens baseadas em LSTM vêm sendo empregues com sucesso para prever a procura em cadeias de lojas, integrando fatores como histórico de vendas, eventos promocionais e até informações climatéricas (e.g., chuva, temperatura) como entradas adicionais da rede (Loureiro et al., 2018). Uma vantagem adicional é que modelos LSTM podem ser treinados de forma global em múltiplas séries de produtos/lojas simultaneamente, aprendendo padrões compartilhados entre séries (Hewamalage et al., 2021), o que é relevante no retalho com catálogos extensos. Apesar dos avanços com *LSTM*, as RNNs apresentam algumas limitações, principalmente quando se trata de capturar dependências muito longas ou relações complexas entre posições distantes na sequência.

Com vista a contornar essas limitações, surgiu uma nova classe de modelos de DL baseada em mecanismos de atenção, culminando na arquitetura Transformer (Vaswani et al., 2017). Os Transformers abandonam a recorrência das RNNs e, em vez disso, utilizam mecanismos de autoatenção que permitem ao modelo pesar a relevância de cada elemento da sequência em relação aos outros, independentemente da sua distância temporal. Essa característica confere duas vantagens principais: (1) a capacidade de capturar relações de longo alcance na série de forma direta (uma venda atual pode ser relacionada a um padrão ocorrido muitos passos atrás, detetando-se essa influência via atenção), e (2) paralelização eficiente do treino, já que todos os passos da sequência são processados simultaneamente (Vaswani et al., 2017).

Originalmente desenvolvidos para tradução e outras tarefas de NLP, os Transformers rapidamente demonstraram desempenho superior a RNNs tradicionais em diversas tarefas sequenciais. A aplicação de Transformers em séries temporais de previsão de vendas é mais recente, mas já mostrou resultados promissores: por exemplo, Lim et al. (2021) propuseram o *Temporal Fusion* Transformer (TFT), que integra um *backbone* LSTM com mecanismos de atenção multi-cabeça focados em variáveis exógenas, alcançando estado-da-arte em vários conjuntos de dados de previsão multivariada e fornecendo interpretabilidade sobre o impacto de cada fator. Outros derivados puramente baseados em Transformer, como o *Informer* (Zhou et al., 2021) e o *Autoformer* (Wu et al., 2021), introduziram variantes de auto-atenção mais eficientes para lidar com sequências longas de séries temporais, reportando melhorias significativas em *benchmarks* de previsão de longo prazo (como energia e tráfego) face a modelos LSTM.

No domínio específico do retalho, o uso de Transformers começa a evidenciar vantagens à medida que a disponibilidade de uma grande quantidade de dados (por exemplo, históricos longos, dados de múltiplas lojas e informação contextual variada) permite explorar todo o potencial desses modelos. D. Oliveira & Ramos (2024) avaliaram extensivamente arquiteturas baseadas em Transformer (incluindo versões *vanilla, Informer, Autoformer* e TFT) num conjunto de dados público de vendas diárias de retalho (competição M5), constatando que os modelos de Transformer superaram significativamente os *baselines* tradicionais (ARIMA, ETS e ingénuo sazonal). Em particular, obtiveram reduções de 26–29% no erro absoluto médio escalado (MASE) e diminuições de até 34% nas métricas baseadas em quantis comparativamente a métodos sazonais ingénuos, evidenciando a capacidade superior dos Transformers em capturar padrões complexos de procura. Esses ganhos de precisão vêm acompanhados de um custo computacional mais elevado e maior necessidade de dados para treino, porém reforçam o potencial do Transformer para melhorar a previsão de vendas quando se dispõe de histórico abundante e variável.

Adicionalmente, Sun & Li (2024) desenvolveram um Transformer adaptado a dados tabulares de vendas empresariais, demonstrando desempenho melhor que modelos de ML convencionais (redução de MAE/MSE e $R^2 \approx 0{,}95$), o que confirma que a arquitetura de atenção, mesmo em cenários de dados relativamente agregados e de baixa dimensionalidade, consegue capturar padrões subtis e estocásticos melhor do que abordagens anteriores.

2.5. Avaliação e Implementação

2.5.1. Comparação crítica dos modelos e seleção final

Considerando o panorama descrito, o presente trabalho inclui como modelos candidatos as técnicas XGBoost e LightGBM (representando a abordagem de *machine learning*) e as arquiteturas LSTM e Transformer (representando as redes neuronais profundas). Esta seleção visa equilibrar modelos com forte desempenho em dados estruturados com redes capazes de capturar relações temporais complexas.

A escolha do XGBoost no presente estudo justifica-se por aliar a flexibilidade na incorporação de múltiplos fatores explicativos (*features* exógenas de negócio, calendários promocionais, etc.) à comprovada eficácia preditiva em dados tabulares deste domínio (Chen & Guestrin, 2016; Andrade & Cunha, 2023). Espera-se que sirva como um forte referencial de comparação para as abordagens de aprendizagem profunda, capturando os padrões essenciais dos dados de vendas com simplicidade estrutural e baixo risco de sobreajuste.

A inclusão do LightGBM como modelo adicional justifica-se pelos ganhos computacionais e pela sua competitividade em termos de desempenho preditivo, conforme evidenciado na literatura (Ke et al., 2017; Ileri, 2025; Kizgin et al., 2025). A granularidade SKU/loja adotada neste estudo é particularmente adequada a estas abordagens, na *medida* em que permite capturar padrões locais e globais de comportamento de procura — aspeto já salientado como essencial para a eficácia operacional no retalho (Loureiro et al., 2018)

A LSTM foi escolhida por ser amadurecida e comprovada em tarefas de previsão de séries temporais, conseguindo modelar de forma eficaz autocorrelações e atrasos temporais presentes nas vendas de retalho (evidenciado por Hewamalage et al., 2021). A sua estrutura de memória interna torna-a apta a aprender efeitos sazonais (por exemplo, picos semanais ou mensais) e impactos prolongados de eventos (promoções, campanhas) ao longo do tempo.

Por outro lado, o Transformer foi selecionado por representar o estado da arte atual em modelação sequencial, oferecendo a capacidade de capturar dependências complexas de longa duração e interações entre múltiplas variáveis de entrada com uma expressividade dificilmente atingível por outros modelos. Como demonstrado na literatura recente (J. M. Oliveira & Ramos, 2024); Lim et al., 2021), os Transformers tendem a superar modelos recorrentes clássicos quando há volume de dados suficiente, podendo alavancar informação global de forma mais eficaz. Em suma, a combinação de LSTM e Transformer nesta revisão permite cobrir duas gerações de redes neuronais – uma consolidada e outra emergente – fornecendo uma fundamentação teórica sólida e comparativa para a componente de *Deep Learning* do estudo.

2.5.2. Implicações práticas e plano de implantação

A Tabela 1 resume as características principais de cada modelo, as suas vantagens e desvantagens no contexto da previsão de vendas em retalho.

Tabela 1 – Análise comparativa dos modelos de previsão

Modelo	**Abordagem**	**Vantagens**	**Desvantagens**
XGBoost	*Machine Learning*	Captura não linearidades e interações; escalável; incorpora múltiplas variáveis (Chen & Guestrin, 2016; Makridakis et al., 2022).	Requer engenharia de *features*; complexidade na afinação de hiperparâmetros.
LightGBM	*Machine Learning*	Tempo de treino muito rápido e uso eficiente da memória (vantajoso em conjuntos de dados volumosos). Eficaz com grande número de variáveis numéricas e categóricas. Permite interpretação das importâncias das variáveis. (Ke et al., 2017; Ileri et al., 2025).	Não capta automaticamente padrões temporais, exigindo engenharia prévia de *features*. Número elevado de hiperparâmetros pode dificultar a afinação. Risco de *overfitting* em séries ruidosas
LSTM	*Deep Learning (RNN)*	Modela dependências temporais longas; aprende representações diretamente dos dados (Punia et al., 2020).	Necessita grande volume de dados; custo computacional elevado; menor interpretabilidade.
Transformer	*Deep Learning (atenção)*	Capta padrões de longo alcance; eficiente para grandes volumes e múltiplas variáveis (Oliveira & Ramos, 2024).	Elevada complexidade; maior exigência computacional e *tuning* rigoroso.

3. ANÁLISE EXPLORATÓRIA

A análise exploratória de dados representa uma etapa essencial nas fases iniciais do processo CRISP-DM, em particular nas etapas de **compreensão do negócio** e **compreensão dos dados**. Antes de se avançar para qualquer abordagem preditiva ou inferencial, torna-se imprescindível proceder a uma avaliação rigorosa das principais características dos dados, de modo a identificar padrões, detetar anomalias, reconhecer tendências, testar hipóteses preliminares e aprofundar o conhecimento sobre o contexto do negócio e da empresa em análise. Importa ainda referir que, por **motivos de confidencialidade**, informações concretas como nomenclaturas e categorias dos artigos, identificação das lojas, bem como famílias e subfamílias, serão apresentadas de forma codificada ao longo deste documento.

Posto isto, neste capítulo, a análise está dividida em duas grandes partes, iniciando a 1) análise dos conjuntos de dados, evoluindo para a 2) análise exploratória. O primeiro ponto incide sobre a avaliação das características estruturais dos dados, como formatos, tipos de variáveis, valores ausentes, duplicações e preparação para integração em sistemas de bases de dados. Já o segundo ponto infere-se na investigação das distribuições, relações entre variáveis e possíveis *insights* relevantes que suportam as fases seguintes da investigação.

A realização destas etapas visa assegurar que a posterior modelação se baseie em dados sólidos, consistentes e devidamente compreendidos.

3.1. Análise do conjunto de dados

Durante a análise realizada, foi inicialmente disponibilizado um conjunto de dados reais da empresa em análise, composto por dois ficheiros *parquet*: um que descreve as transações de vendas e outro que carateriza os produtos transacionados. O ficheiro das vendas é composto por 3.278.907 registos e 6 variáveis, e o ficheiro dos artigos, constituído por 22.839 registos e 4 variáveis. Refira-se que os dados, embora reais, foram sujeitos a anonimização, pelo que se encontram transversalmente codificados (SKU, Loja, Família e Subfamília). Igualmente, os preços de venda foram também multiplicados por uma constante.

Foram analisados **22.839 artigos únicos**, dos quais 14.422 apresentam registos de transações de venda. Identificaram-se **14 famílias** distintas, embora apenas 13 evidenciem vendas efetivas. De forma semelhante, contabilizaram-se **467 subfamílias** únicas, das quais 393 apresentam transações registadas.

O conjunto de dados inclui **120 lojas com registos de vendas**, bem como dois tipos de clientes, identificados pelos códigos 1 (clientes profissionais) e 2 (público geral). Neste contexto, o tipo 1 corresponde a 1.771.315 transações (57% do total), enquanto o tipo 2 representa 1.353.817 transações (43%).

O período coberto pelo conjunto de dados estende-se de **janeiro de 2022 a dezembro de 2024**, proporcionando um horizonte temporal de aproximadamente três anos. Esta abrangência permite observar fenómenos sazonais, tendências de crescimento e oscilações de curto e médio prazo. Todos os artigos têm, no máximo, **365 dias de vendas.**

Relativamente à qualidade dos dados partilhados, no conjunto de dados de vendas, foi necessário proceder a uma limpeza dos dados para garantir a consistência e a fiabilidade das análises. Este

processo incluiu a remoção de registos com valores nulos na variável price_per_item (preço por artigo), bem como de registos com valores inválidos — especificamente, casos em que o preço por artigo era inferior ou igual a zero e casos em que a quantidade vendida (qty) era inferior ou igual a zero. Como resultado deste processo de limpeza, foram eliminados 153.775 registos, o que reduziu o volume total do conjunto de dados de vendas para 3.125.132 registos.

Em relação ao conjunto de dados dos artigos, a análise revelou que não existiam valores ausentes nem registos duplicados, mantendo-se assim a totalidade dos dados disponíveis para análise. Após o processo de limpeza, os dados foram organizados e carregados numa base de dados relacional em formato SQLite, permitindo uma estruturação eficiente e potenciando a agilidade das consultas e análises subsequentes. Este procedimento assegurou que os dados a utilizar no projeto se encontrassem em conformidade com os princípios de integridade, consistência e fiabilidade.

A Figura 3 ilustra o modelo relacional construído para representar as vendas ao nível SKU / loja. O esquema materializa-se em cinco tabelas (*Family, Subfamily, Article, Store* e *Sale*).

A lógica de dados segue o paradigma estrela simplificado, em que a tabela *Sale* concentra as transações e as restantes funcionam como dimensões de referência.

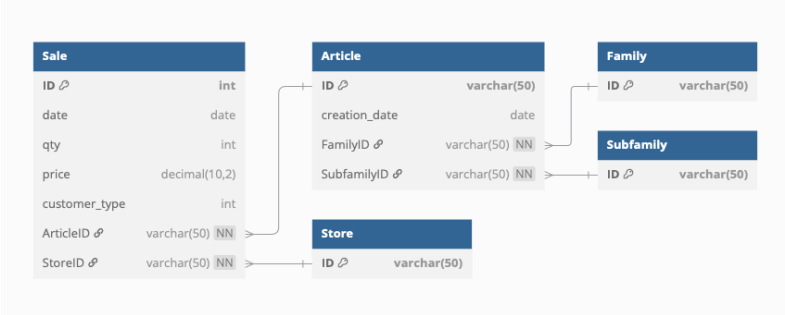


Figura 3 – Modelo relacional da base de dados

3.2. Análise exploratória

A análise exploratória desenvolvida procurou aprofundar o conhecimento dos dados disponíveis, possibilitando uma melhor compreensão dos padrões de comportamento presentes nas vendas e nos artigos comercializados. Reforça-se que os dados fornecidos apresentam o primeiro ano de vendas dos diferentes produtos.

Inicialmente, foi realizada uma análise da distribuição dos artigos pelas respetivas subfamílias, com o objetivo de identificar a concentração do portefólio em determinados segmentos. Em paralelo, analisou-se a distribuição acumulada do volume de vendas por loja, de forma a compreender a representatividade relativa das diferentes unidades comerciais.

Procedeu-se também à avaliação da distribuição dos registos por tipo de cliente, explorando as diferentes categorias existentes e a sua relevância no volume de vendas total. Foi igualmente efetuada uma análise agregada das vendas por família de artigos, permitindo destacar quais os grupos de produtos mais relevantes para o negócio. Por fim, analisou-se o valor médio por unidade vendida, comparando a variação desse indicador em função do tipo de cliente.

As análises desenvolvidas nesta fase têm como objetivo apoiar a seleção e construção das variáveis mais relevantes para as etapas posteriores do estudo, nomeadamente a modelação preditiva e a

segmentação de padrões de comportamento. Nos pontos seguintes, exploram-se as conclusões mais relevantes dentro dos elementos estudados.

3.2.1. Análise de vendas – visão geral

Antes de aprofundar a análise segmentada por variáveis específicas, foi realizada uma avaliação geral da evolução das vendas totais por dia, com recurso a todas as transações disponíveis, entre **janeiro de 2022 a dezembro de 2024,** apresentada na Figura 4.

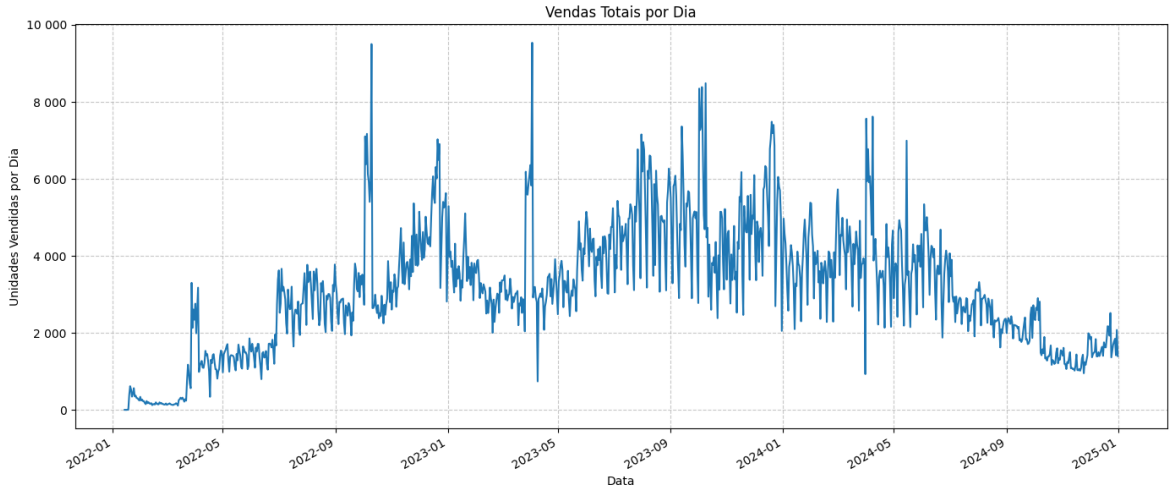


Figura 4 – Visão global das vendas

Relativamente aos padrões sazonais, observam-se picos claros de vendas durante os meses de novembro e dezembro, coincidindo com períodos promocionais como a *Black Friday* e as campanhas de Natal, que tradicionalmente impulsionam o consumo. De igual modo, identificam-se aumentos no volume de vendas durante os meses de verão (especialmente entre junho e agosto), sugerindo uma possível associação a campanhas sazonais ou alterações no comportamento de compra típicas deste período do ano.

Por outro lado, nota-se uma considerável volatilidade no número diário de unidades vendidas, com oscilações acentuadas, indicando a influência de fatores promocionais, campanhas específicas ou variações externas de mercado.

3.2.2. Análise das vendas por família – volume e quantidade

Com o objetivo de compreender a distribuição das vendas no portefólio de produtos, foi realizada uma análise do volume de vendas e do número de unidades vendidas por família de artigos.

Para esta análise, foram considerados dois indicadores distintos: (i) o número de registos de vendas por família, contabilizando o número total de transações registadas, e (ii) o total de unidades vendidas por família, considerando a soma das quantidades vendidas (qty) associadas a cada artigo. A junção das bases de vendas e de artigos permitiu realizar esta análise por agrupamento da variável *family.*

Os resultados, apresentados na Figura 5 e Figura 6, mostram que a diferença entre o volume de vendas e o número de unidades vendidas por família não é significativamente expressiva, o que indica que o preço médio por unidade entre as famílias mais vendidas é relativamente homogéneo.

Esta consistência reduz a possibilidade de distorções provocadas por variações de preço entre as diferentes famílias, permitindo uma comparação mais direta do desempenho em termos de quantidade e valor.

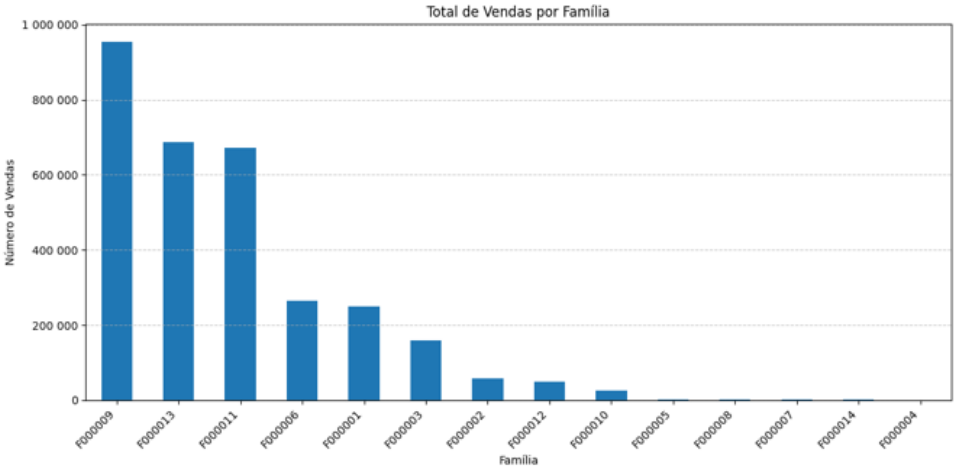


Figura 5 – Número de registos de vendas por família de artigos

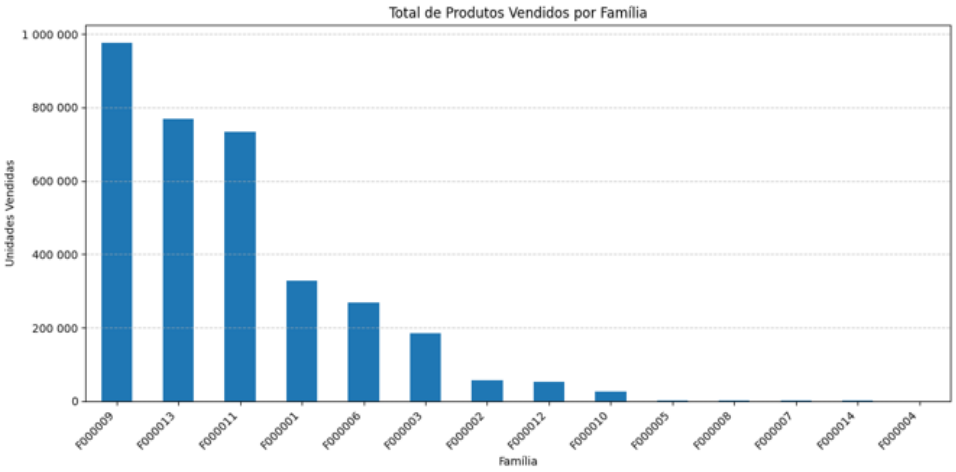


Figura 6 – Total de unidades vendidas por família

Adicionalmente, observa-se que uma reduzida quantidade de famílias concentra uma parte substancial do total de vendas, tanto em número de unidades como em valor monetário. As três principais famílias (F00009, F00013 e F00011) destacam-se de forma evidente, representando isoladamente uma parcela significativa do total. Este fenómeno é indicativo de uma forte concentração de vendas em determinados grupos de produtos, o que poderá ter implicações importantes para decisões de gestão de portefólio e estratégias lançamento de novos artigos.

3.2.3. Análise das subfamílias – artigos e vendas

Foi realizada uma análise da distribuição dos artigos e das vendas por subfamília, com o objetivo de compreender a concentração e a dispersão do portefólio de produtos.

Relativamente à distribuição de artigos, observa-se na Figura 7 que as subfamílias com um maior número de vendas concentram uma parte significativa dos artigos disponíveis. No entanto, mesmo

considerando as 50 maiores subfamílias, a cobertura em termos de total de artigos não atinge os 100%, o que evidencia uma dispersão relevante no catálogo. Adicionalmente, a curva cumulativa cresce de forma relativamente suave após as primeiras posições, sugerindo uma distribuição menos concentrada do que a típica regra de Pareto (80/20).

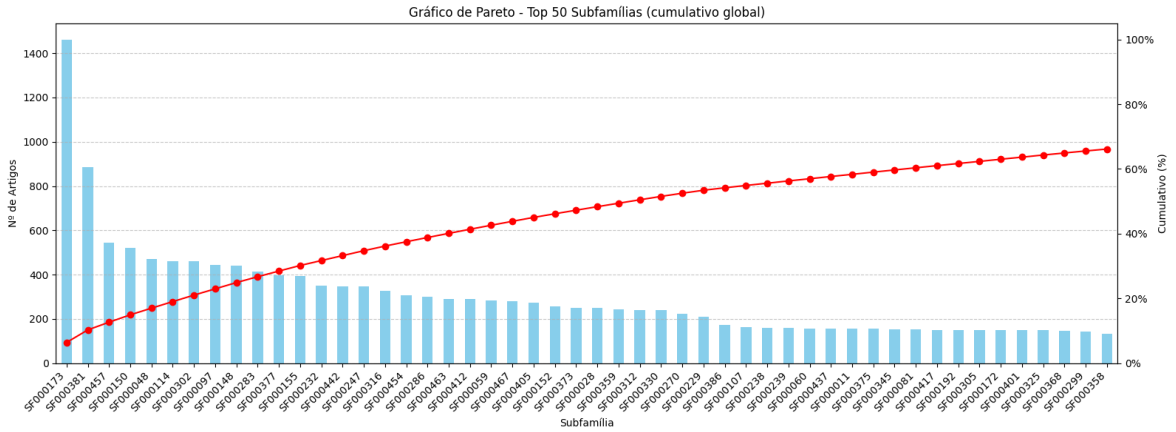


Figura 7 – Gráfico de Pareto - Subfamílias - Artigos

Em contraste, a análise da distribuição de unidades vendidas revela uma concentração muito mais acentuada, conforme se verifica na Figura 8. Um pequeno grupo de subfamílias é responsável por uma percentagem muito significativa das vendas totais, com uma curva cumulativa que sobe mais rapidamente do que na análise de artigos, indicando um comportamento típico de alta concentração.

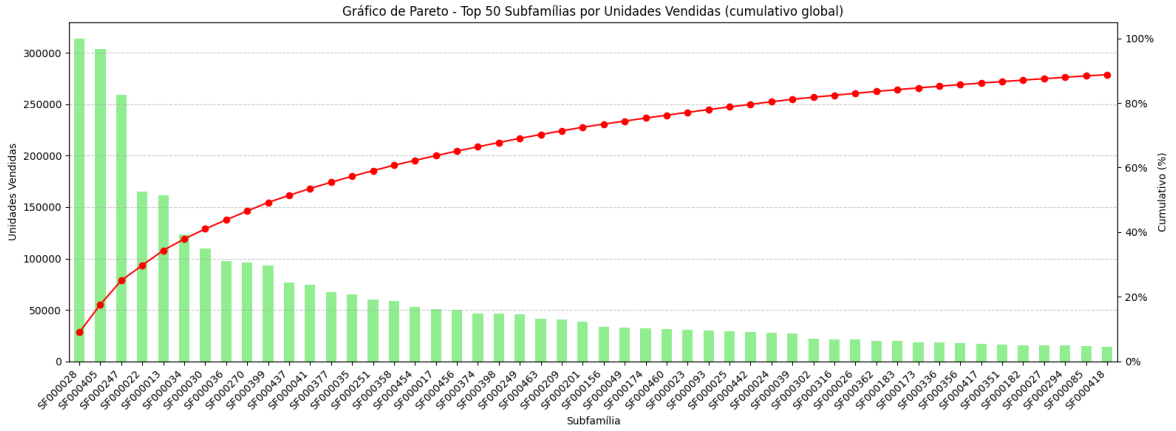


Figura 8 – Gráfico de Pareto - Subfamílias - Vendas

Adicionalmente, identificou-se a existência de subfamílias para as quais não foram registadas vendas (74 das 467 subfamílias), o que reforça a perceção de dispersão no catálogo e aponta para potenciais oportunidades de otimização e gestão do portefólio.

3.2.4. Distribuição dos preços – visão ao artigo

A análise da distribuição dos preços dos artigos teve como objetivo caracterizar a estrutura de preços do portefólio disponível, focando-se no intervalo entre os percentis 10.º e 90.º para eliminar os valores extremos e obter uma representação mais clara do comportamento central.

Os resultados, conforme verificáveis na Figura 9, revelam uma distribuição assimétrica à direita, com uma concentração muito elevada de artigos nos escalões de preço mais baixos. A maioria dos

artigos situa-se abaixo dos 20 euros, com um pico claro de frequência entre os 5 e os 10 euros. Esta configuração indica que o portefólio está fortemente orientado para artigos de gama baixa, o que poderá refletir uma estratégia de acessibilidade e volume.

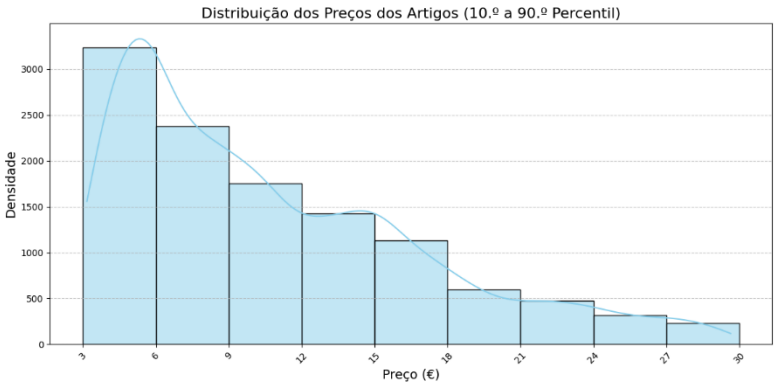


Figura 9 – Distribuição dos preços dos artigos - visão global portefólio

À medida que o preço aumenta, observa-se uma redução progressiva e acentuada na frequência de artigos, confirmando a existência de uma longa cauda direita. A curva de densidade sobreposta reforça esta assimetria, evidenciando que os artigos mais caros são significativamente menos representativos no conjunto analisado.

Este padrão sugere que os produtos de preço elevado constituem uma fração residual da oferta da empresa, estando provavelmente direcionados para segmentos de mercado muito específicos ou produtos com maior valor agregado.

Para esta análise, foi considerado o preço médio por artigo, calculado com base na variável **price_per_item** agregada por **article**. Esta abordagem permitiu suavizar flutuações pontuais de preço e obter uma visão mais robusta da estrutura de preços do portefólio.

3.2.5. Relação entre lojas e vendas

A análise da relação entre o número de unidades vendidas e o valor total faturado por loja revelou uma forte correlação positiva. O coeficiente de correlação calculado foi de 0,879, indicando que, de uma forma geral, as lojas que vendem mais unidades tendem também a faturar mais.

No entanto, o facto de a correlação não ser perfeita demonstra que existem variações significativas nos preços médios praticados entre as diferentes lojas. Isto é, conforme se verifica na Figura 10, onde se contempla o **valor total vendido (€)** no eixo do Y e o **total de unidades vendidas** no eixo do X, por loja, algumas destas conseguem gerar um volume de faturação elevado mesmo com um número relativamente menor de unidades vendidas, evidenciando a importância da composição do portefólio de produtos e do seu posicionamento de preços.

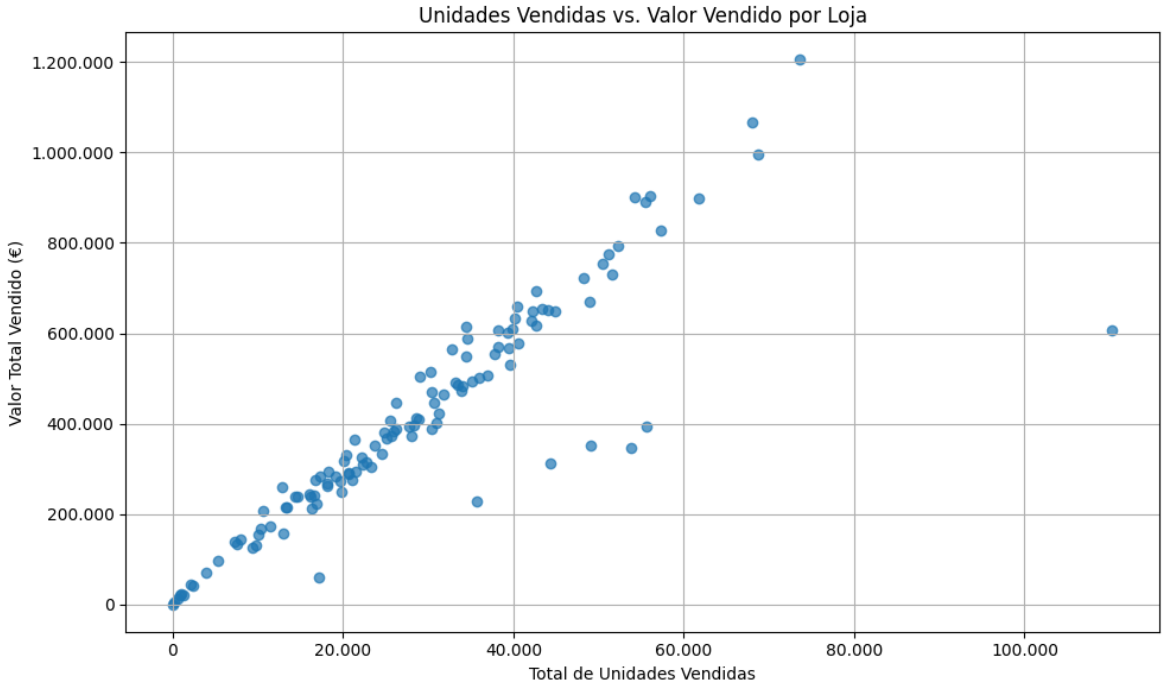


Figura 10 – Comparação das unidades vendidas com valor de vendas - visão por loja

Por outro lado, a distribuição das vendas por loja, presente na Figura 11, revela ainda uma concentração relevante no topo do ranking, com as lojas mais destacadas a representar uma percentagem significativa do valor global faturado. Apesar disso, a curva acumulativa apresenta uma progressão relativamente suave, sugerindo que as lojas intermédias têm um papel relevante na composição da faturação global. Este fenómeno é confirmado pela observação de que, mesmo considerando as 30 lojas com maior volume de vendas, não se atinge a totalidade do valor vendido, o que reforça a importância de uma rede de lojas equilibrada e diversificada.

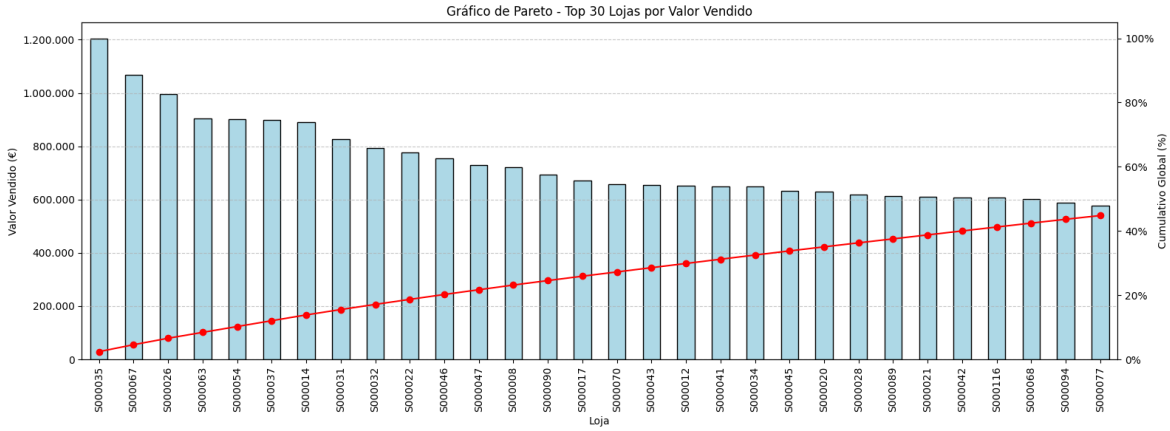


Figura 11 – Gráfico de Pareto - TOP 30 lojas com maior volume de vendas

3.2.6. Vendas por tipo de cliente

Foi realizada uma análise comparativa entre os diferentes tipos de clientes, com foco na evolução do volume de unidades vendidas, no valor total faturado e no preço médio por unidade.

Observa-se na Figura 12 que o Tipo 1 (clientes profissionais) apresenta sistematicamente um volume de unidades vendidas superior ao Tipo 2 (público geral) ao longo de todo o período

analisado. Esta tendência é consistente, embora se verifiquem variações sazonais que afetam ambos os tipos de clientes, em especial picos no mês de novembro, refletindo provavelmente campanhas promocionais e o aumento de consumo típico da época festiva.

Importa salientar que no mês de julho-2024 foi identificada uma inversão temporária deste padrão, com o Tipo 2 a registar um volume de unidades vendidas superior ao do Tipo 1, sugerindo uma possível alteração sazonal ou uma campanha dirigida especificamente a esse segmento – que se estendeu até dezembro do mesmo.

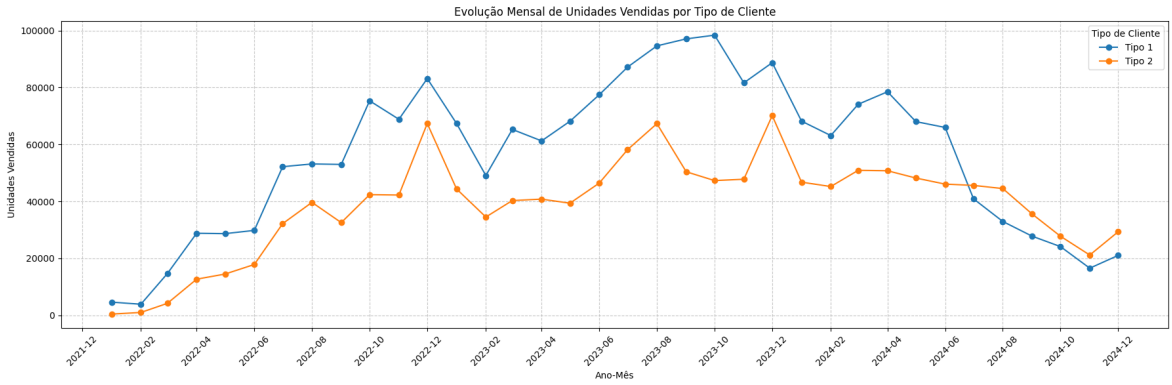


Figura 12 – Evolução mensal de vendas por tipo de cliente

No entanto, a análise do valor total vendido, presente na Figura 13, mostra que, apesar de venderem menos unidades, os clientes do Tipo 2 conseguem gerar um volume de faturação muito semelhante ao dos clientes do Tipo 1. Esta dinâmica é explicada pelo valor médio por unidade vendida, onde se verifica que o Tipo 2 apresenta preços médios significativamente superiores.

Estes resultados indicam que o Tipo 2 está associado a vendas de produtos de maior valor agregado, enquanto o Tipo 1 opera predominantemente com artigos de preço mais reduzido, mas em maiores quantidades. Esta distinção entre perfis de clientes poderá ter implicações relevantes para a definição de estratégias de implementação de novos artigos.

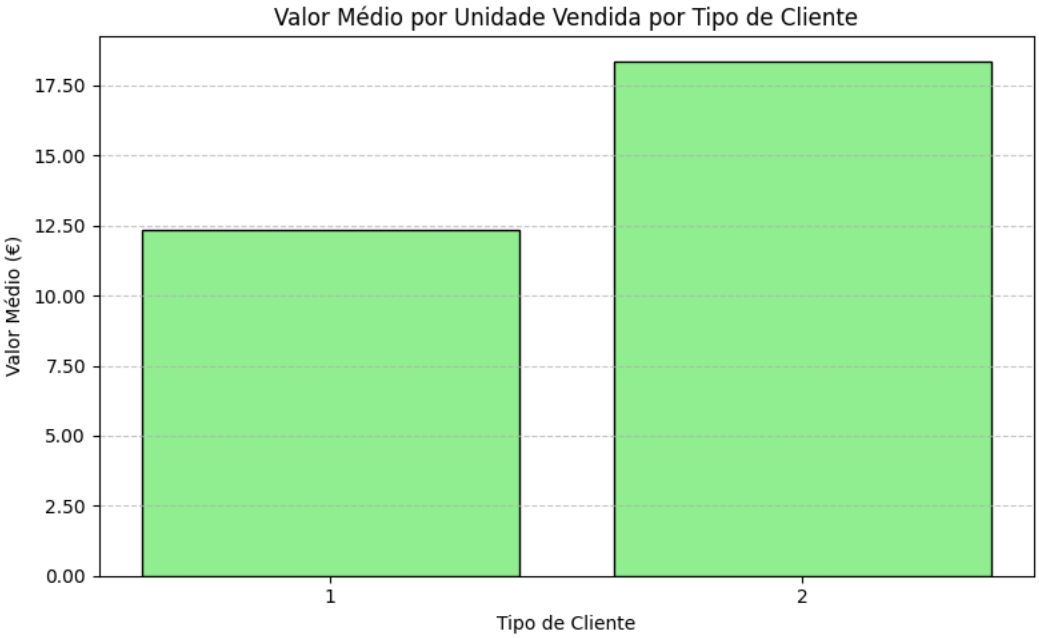


Figura 13 – Valor médio por unidade vendida e tipo de cliente

3.2.7. Distribuição da frequência de vendas por dia

Foi analisada a dispersão do número diário de registos de venda ao longo do último ano, espelhado na Figura 14. Verifica-se que a maior parte dos dias (cerca de 60 %) concentra-se no intervalo de 2 000 a 4 000 registos, com um pico (moda) em torno de 3 200 unidades. Contudo, a longa cauda direita revela um subconjunto reduzido de dias — aproximadamente 5% — que ultrapassam os 6 000 registos e chegam a extremos próximos de 8 000. Estes picos, normalmente associados a campanhas promocionais ou eventos sazonais, puxam a média para cima e tornam-na pouco representativa do dia "típico".

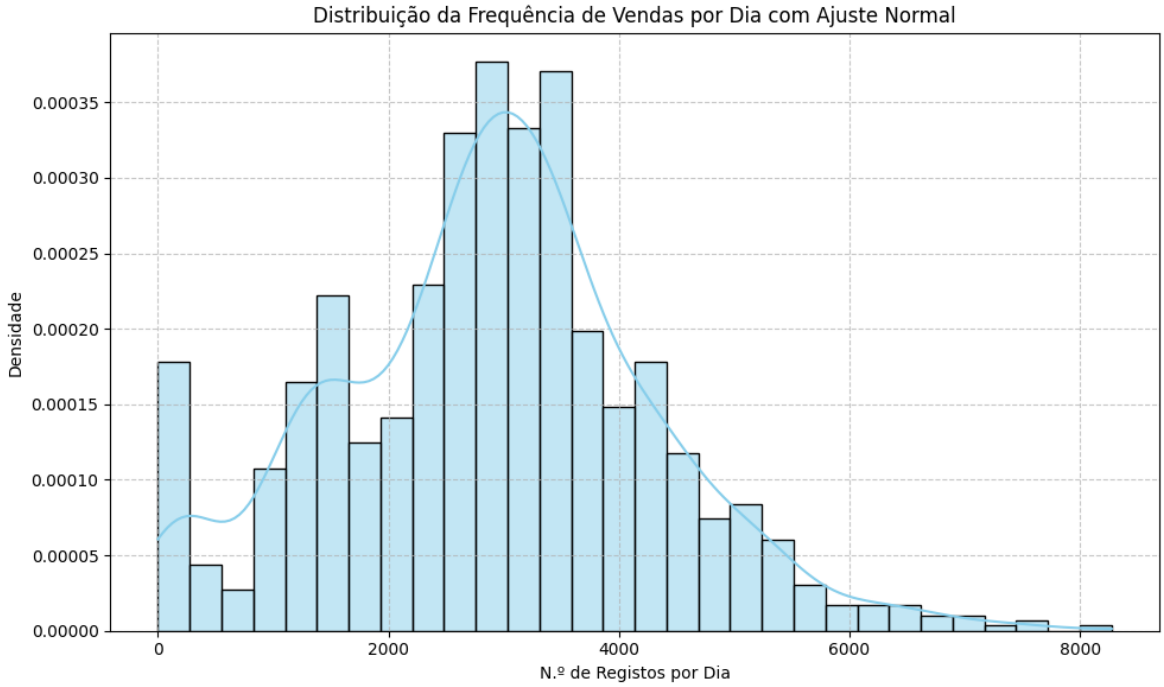


Figura 14 – Distribuição da frequência de vendas por dia

3.2.8. Distribuição da duração de venda dos artigos

O gráfico seguinte (Figura 15) verifica-se quanto tempo cada artigo permaneceu a vender entre a sua primeira e a sua última transação dentro do período em estudo — a duração efetiva do ciclo de venda de cada SKU, considerando apenas os casos em que esse intervalo é superior a zero. A curva de densidade apresenta-se agora como claramente multimodal: primeiro surge uma ligeira concentração de artigos que se esgotam muito cedo, entre 30 e 60 dias após o lançamento; em seguida aparece um pico intermédio bem visível na faixa dos 180-190 dias, sugerindo um grupo relevante de produtos com ciclo de venda semestral; por fim destaca-se o pico principal em torno dos 330-350 dias, que corresponde aos itens que permanecem ativos praticamente todo o ano. Entre estes núcleos observa-se um vale relativamente plano, no qual as vendas se diluem ao longo de quatro a dez meses.

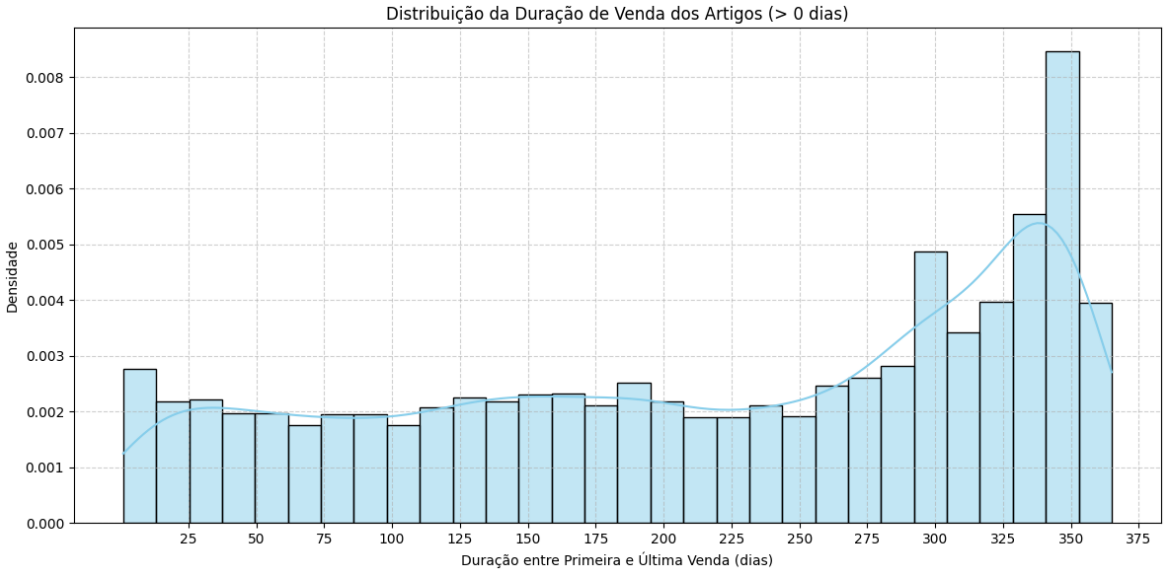


Figura 15 – Distribuição da duração de venda dos artigos

3.3. Síntese e potenciais implicações metodológicas

A análise exploratória evidenciou três fatores que deverão orientar o raciocínio na etapa seguinte. Primeiro, os ciclos sazonais — observáveis tanto à escala anual como semanal — confirmam que a procura varia de forma sistemática ao longo do calendário, tornando indispensável considerar a vertente temporal. Segundo, a forte dispersão e assimetria dos preços, associadas a variações no volume transacionado, indicam que a sensibilidade ao preço condiciona significativamente o comportamento de compra. Por fim, a acentuada heterogeneidade entre famílias de produto, lojas e segmentos de cliente demonstra que o contexto categórico influencia diretamente o desempenho de cada artigo.

Estes resultados constituem, assim, o enquadramento conceptual que sustenta a definição, no capítulo seguinte, das variáveis capazes de traduzir estes três vetores de variação e, desse modo, suportar modelos previsionais mais robustos.

4. METODOLOGIA

Este capítulo dedica-se, em exclusivo, à previsão das vendas semanais de cada produto, **integrando as etapas de preparação dos dados e modelação do processo CRISP-DM**. Partindo do histórico de transações, descreve-se a transformação desses registos em variáveis explicativas e as diferentes estratégias de modelação aplicadas para gerar previsões fiáveis que suportem a decisão comercial (Figura 16).

Nesta secção são também apresentadas as escolhas relativas à seleção e pré-processamento dos dados, à engenharia de características, à divisão dos conjuntos de treino/validação/teste e às estratégias de modelação empregues. Em particular, discutem-se as limitações de amostragem pelos grupos (família, subfamília), as variáveis derivadas usadas (temporal e de preço), a estrutura comum de divisão de dados e a abordagem de modelação global versus por família.

O código-fonte utilizado neste estudo encontra-se disponível, na íntegra, em repositório online[7], o que permite não só a replicação integral dos resultados, como também facilita análises críticas independentes e futuras extensões por outros investigadores.

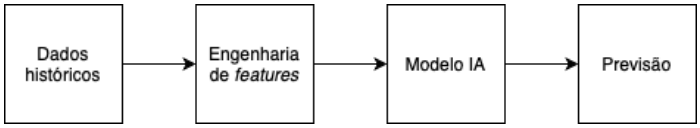


Figura 16 – Fluxograma simplificado da metodologia de previsão de vendas semanais

4.1. Seleção de dados e agrupamento de artigos

Numa fase inicial, o conjunto de dados de vendas foi filtrado para manter apenas os registos úteis à modelação. Excluíram-se registos inválidos (preço por artigo não positivo ou quantidades não positivas) e artigos sem qualquer venda no período estudado. Consequentemente, a amostra final inclui apenas SKUs com histórico de vendas. Esta decisão introduz um potencial viés de amostragem: artigos sem vendas (zero vendas) ficam completamente fora da análise, pelo que os padrões aprendidos não contemplam estes casos. É importante contextualizar que este viés é inerente à estratégia adotada – quer-se focar nos comportamentos de artigos efetivamente vendidos –, mas limita a generalização do modelo para produtos que não têm histórico de vendas. Reconhece-se esta limitação metodológica, salientando que os modelos resultantes são otimizados para prever a procura de artigos já com registos de venda.

Para análise e modelação, cada artigo (SKU) foi considerado dentro do tuplo (família, subfamília) a que pertence. Ou seja, as variáveis e padrões de venda são agregados ao nível de SKU, mas sempre associados às categorias de família e subfamília correspondentes. Esta abordagem permite agrupar artigos semelhantes com base na sua classificação hierárquica: assume-se que produtos da mesma subfamília compartilham atributos e tendências de venda. **Agregaram-se os dados ao nível da semana e calculou-se o total de vendas recentes por SKU**. Em seguida, treinaram-se modelos de previsão com base na suposição de que os SKUs pertencentes à mesma subfamília partilham padrões de comportamento, formando assim grupos coerentes. Esta abordagem permite prever a

[7] Disponível em: https://github.com/pedroalsilva/retail-forecast

quantidade vendida por semana, assegurando que cada instância de análise (SKU por semana) incorpora não apenas as características individuais do produto, mas também informações contextuais das respetivas categorias (família e subfamília).

4.2. Variáveis *input*

Foram construídas variáveis derivadas (Tabela 2) para capturar tanto tendências de vendas históricas como outras características relevantes (preço e idade do produto). Em termos temporais, utilizaram-se dois recortes de histórico de vendas: um de curto prazo (últimos 7 dias) e outro de médio prazo (últimas 4 semanas). Especificamente, para cada exemplo construiu-se a variável de quantidade vendida na última semana (*lag* de 1 semana) e a média móvel das vendas das últimas 4 semanas. Esta dualidade de recortes permite ao modelo identificar sinais iniciais de saída de produto (venda recente) bem como tendências mais estáveis a médio prazo. Por exemplo:

- Vendas **últimos 7 dias:** permite captar surto precoce de procura, revelador de lançamento bem-sucedido ou promoção temporária.
- Vendas **últimas 4 semanas**: suaviza variações pontuais, identificando comportamentos mais sustentados ao longo de um mês.

Tabela 2 – Variáveis input

Nome da variável	**Função**	**Tipo**	**Designação**
family[8]	Input	Categórico	Família do artigo
subfamily	Input	Categórico	Subfamília do artigo
price_per_item	Input	Contínuo	Mediana dos preços de venda para a semana seguinte.
week_since_creation	Input	Contínuo	Diferença entre semana da venda e semana de criação do artigo.
qty_lag_1	Input	Contínuo	Quantidades vendidas na semana anterior.
qty_rolling_mean_4	Input	Contínuo	Média das quantidades vendidas nas 4 semanas anteriores.
Qty	Target	Contínuo	Quantidades vendidas do artigo por semana.

Quanto à variável do preço, foi adotada dupla representação: calculou-se a média histórica de preço por artigo e também a mediana histórica do preço por artigo. Esta escolha justifica-se pela natureza assimétrica dos preços: a média reflete o valor médio incluindo todos os registos, mas pode ser influenciada por valores extremos (promoções ou valores muito altos); a mediana oferece uma medida robusta à presença de *outliers*. Incluir ambas permite ao modelo comparar abordagens e escolher internamente a mais informativa. Os impactos são, por exemplo, a maior estabilidade da mediana em cenários com muitos descontos ou flutuações bruscas de preço.

[8] *Feature* apenas utilizada no modelo global.

Além disso, criou-se uma variável categórica de faixa de preço para cada artigo, determinada por *K-means*. Aplicou-se o algoritmo *K-means* aos valores históricos de preço (num *feature engineering* prévio) para segmentar os SKUs em três *clusters* automáticos, interpretados como **Preço Baixo**, **Médio** e **Alto**. Isto resulta numa característica categórica adicional ("faixa de preço") que complementa o valor numérico do preço por item. A motivação é que modelos de previsão muitas vezes capturam melhor efeitos não lineares de preço quando as observações são agrupadas em categorias de faixa. Em termos práticos, definiram-se as seguintes faixas (exemplos ilustrativos):

- **Baixa**: artigos de preço baixo (*cluster* 1)
- **Média**: artigos de preço intermédio (*cluster* 2)
- **Alta**: artigos de preço elevado (*cluster* 3)

Por fim, introduziu-se a variável **week_since_creation**, que mede o tempo (em semanas) desde o lançamento do artigo até a semana analisada. Esta variável de maturidade (idade do produto) foi agregada por semanas (em vez de dias) dado que os dados históricos não tinham granularidade diária suficiente para contagem precisa. Deste modo, **week_since_creation** informa o modelo sobre a fase de ciclo de vida do produto – por exemplo, se é um item recém-introduzido (poucas semanas) ou um produto estabelecido (mais semanas no mercado) –, o que pode influenciar fortemente a procura no momento do lançamento.

4.3. Divisão treino/validação/teste

Para garantir comparabilidade entre modelos, utilizou-se um único procedimento de divisão temporal aplicado a todos. O conjunto completo (todas as famílias) foi dividido por famílias de forma não aleatória (sem *shuffle*), preservando a ordem cronológica. Concretamente, para cada família reservou-se 20% dos registos mais recentes como teste, e 80% restantes como treino e validação. Do conjunto de dados utilizados para treino e validação, retiveram-se 10% adicionais (8% do total original) para validação, resultando em 72% treino, 8% validação e 20% teste. Na Figura 17 é possível ver a separação temporal e a alocação às partições correspondentes.

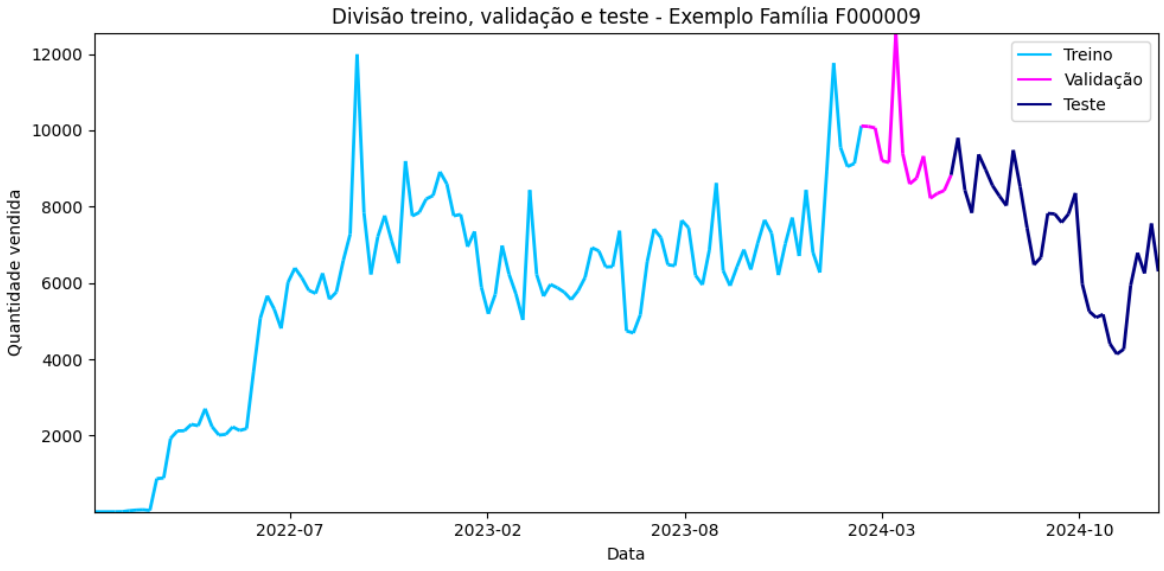


Figura 17 – Demonstração da divisão treino, validação e teste para a família **F000009**

As partições obtidas por família foram então unificadas para criar conjuntos globais usados pelos modelos gerais. Isto é, os *subsets* **X_train, X_val** e **X_test globais** foram formados concatenando as divisões de cada família, conforme ser verifica no código 1, abaixo.

```
for family in weekly_df['family'].unique():
    # Ordenação cronológica dos registos da família corrente
    df_fam = (
        weekly_df[weekly_df['family'] == family]
        .sort_values(by=['year', 'week_of_year'])
    )

    # Variáveis explicativas (X) vs. variável-alvo (y)
    X_fam = df_fam[features]
    y_fam = df_fam['qty']

    # 1.º corte: reserva 20 % mais recente para teste
    X_trainval_fam, X_test_fam, y_trainval_fam, y_test_fam = train_test_split(
        X_fam,
        y_fam,
        test_size=0.2,
        shuffle=False   # mantém ordem temporal
    )

    # 2.º corte: extrai 10 % da parcela de treino para validação (~8 % do total)
    X_train_fam, X_val_fam, y_train_fam, y_val_fam = train_test_split(
        X_trainval_fam,
        y_trainval_fam,
        test_size=0.1,
        shuffle=False
    )

    # Armazenamento estruturado – facilita reprodutibilidade e avaliação
    splits["per_family"][family] = {
        "X_train": X_train_fam,
        "X_val": X_val_fam,
        "X_test": X_test_fam,
        "y_train": y_train_fam,
        "y_val": y_val_fam,
        "y_test": y_test_fam
    }
```

Código 1 –Criação de subconjuntos temporais por família

Importa, nesta fase, reforçar o cuidado tomado para evitar qualquer fuga de informação entre treino e teste. A divisão dos dados foi realizada estritamente por ordem cronológica: os últimos 20 % das observações de cada família foram reservados para teste, garantindo que o modelo nunca contacta, durante o treino, com registos do futuro. Qualquer transformação incluída no *pipeline* — por exemplo, codificação de variáveis categóricas ou cálculo de estatísticas agregadas — é ajustada exclusivamente no conjunto de treino e depois aplicada, sem recálculo, aos conjuntos de validação e de teste. Além disso, a amostra de validação mantém a mesma ordem temporal, eliminando o risco de *look-ahead bias*. Este rigor metodológico assegura que os resultados apresentados refletem o desempenho em dados genuinamente não vistos, condição indispensável para confiar nas previsões quando o sistema estiver em produção.

4.4. Balanceamento de dados

Dado o desequilíbrio natural entre famílias (algumas muito mais populosas do que outras), avaliou-se estratégias de balanceamento para evitar que o modelo global fosse dominado pelos padrões da família maioritária. Optou-se por não aplicar o SMOTE tradicional, pois este algoritmo foi originalmente concebido para problemas de classificação e requer um alvo nominal; quando a variável-alvo é contínua, como no presente caso (quantidade vendida), a literatura recente recomenda variações específicas para regressão—por exemplo, SmoteR, SMOGN ou WSMOTER—exatamente porque o SMOTE puro não é adequado nestes cenários (Avelino et al., 2024;Camacho & Bacao, 2024).

Em alternativa, testou-se a **LN-SMOTE** (*Local Neighbourhood-SMOTE*), uma variante do SMOTE que restringe a geração de exemplos sintéticos ao micro-contexto de cada instância minoritária, reduzindo a probabilidade de sobre generalizar regiões de maioria (Maciejewski & Stefanowski, 2011).

Implementou-se a LN-SMOTE apenas ao **conjunto de treino** dos modelos globais de LightGBM e XGBoost. As tabelas presentes no anexo 1 resumem o efeito:

- **LightGBM** – a média ponderada de MAE caiu de **4.90 → 4.70** (-4 %); o RMSE desceu ligeiramente de **14.68 → 14.56**; o MAPE melhorou de **0.98 → 0.81**; e o R^2 subiu de **0.82 → 0.83**. O ganho é modesto e concentra-se sobretudo na redução do erro relativo. *Exemplo ilustrativo*: na família **F000005** o MAE baixou de 3.79 para 2.21 e o MAPE de 1.46 para 0.73, mas em **F000001** as melhorias foram marginais.
- **XGBoost** – apresentou variação quase nula ou ligeiramente negativa: MAE **4.35 → 4.39** (+1 %); RMSE **14.48 → 14.75**; R^2 **0.84 → 0.83**. O MAPE melhorou simbolicamente (0.57 → 0.56). *Exemplo*: em **F000006** o MAPE baixou de 0.49 para 0.44, mas em **F000013** a maioria das métricas manteve-se estável ou piorou ligeiramente.

Em síntese, a LN-SMOTE gerou melhorias pontuais no LightGBM e resultados neutros ou ligeiramente negativos no XGBoost. Dado o custo adicional de processamento e a ausência de ganhos consistentes nas métricas-chave, **optou-se por manter as conclusões baseadas no conjunto original, sem *oversampling* sintético**.

4.5. Modelos de previsão e calibração

Foram implementados quatro algoritmos principais de previsão de vendas: XGBoost, LightGBM, LSTM (redes recorrentes) e Transformer (redes de atenção). Cada modelo foi treinado e avaliado em dois contextos distintos: (i) global, usando o conjunto de dados total combinado, e (ii) por família, treinando um modelo separado apenas com os dados daquela família específica. Em termos práticos, isto significa que foram construídos 1 modelo global e até 13 modelos individuais (para cada uma das 13 famílias de artigos com registos de vendas), para cada técnica. O propósito dessa abordagem dupla é avaliar se a abordagem através de um modelo global, que aproveita toda a diversidade de dados, é mais vantajosa quando comparada com modelos individualizados, que capturam padrões por família de artigos

Cada modelo de previsão foi configurado com um conjunto específico de hiperparâmetros ajustados empiricamente, conforme se verifica na Tabela 3. Refira-se, para todos os efeitos, que o desempenho de qualquer algoritmo depende dos parâmetros e hiperparâmetros.

Neste seguimento, no caso do **XGBoost,** utilizou-se n_estimators = 1000, learning_rate = 0.05 e max_depth = 6, combinação que concilia um conjunto grande de árvores com uma taxa de aprendizagem baixa, reforçando a acurácia geral. Os parâmetros de subamostragem (subsample = 0.8) e de amostragem de colunas (colsample_bytree = 0.8) introduzem regularização estocástica, reduzindo correlações entre árvores (Chen & Guestrin, 2016).

De forma análoga, o **LightGBM** adotou n_estimators = 1000, learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8 e num_leaves = 31, seguindo recomendações clássicas para GBDT (Ke et al., 2017). O treino foi monitorizado por *early stopping* (paciência = 50).

No **LSTM**, a arquitetura sequencial empregou duas camadas recorrentes e duas camadas densas finais. A primeira LSTM possui 128 unidades (return_sequences=True) e a segunda 64 unidades, número que permite modelar dependências temporais sem exagerar na complexidade (Hochreiter & Schmidhuber, 1997). Em seguida, duas camadas dense (64 unidades com ReLU e 1 unidade de saída) realizam a previsão final. A função de perda escolhida foi MSE (erro quadrático médio) e o otimizador Adam (Kingma & Ba, 2014), combinados para acelerar a convergência.

Definiu-se treino por até 300 ciclos com batch size 32 e EarlyStopping com paciência 10 para interromper caso não haja melhoria, estratégia comum em redes neuronais para evitar sobreajuste. O batch size de 32 é um valor padrão que equilibra estabilidade do gradiente e eficiência computacional.

Para o **Transformer**, construiu-se um bloco de atenção simples: entrada normalizada, atenção multi-cabeça com 2 cabeças (num_heads=2, key_dim=8), dropout de 0.1 e GlobalAveragePooling, seguida de uma Dense de saída. Esse modelo minimalista segue o padrão original (Vaswani et al., 2017), mas com dimensão reduzida devido ao tamanho limitado dos dados. A perda adotada foi MAE (erro absoluto médio) ao invés de MSE, visando maior robustez a *outliers*. Utilizou-se novamente Adam e early stopping (paciência = 10) com treino máximo de 60 ciclos e batch size 32. O dropout de 10% introduz regularização adicional (Srivastava et al., 2014), ajudando a evitar sobreajuste num modelo de atenção relativamente pequeno.

Com vista a explorar ganhos adicionais, foi aplicado um **Grid Search** tanto ao **LightGBM** como ao **XGBoost**, nos modelos globais, usando o espaço de procura apresentado no **Código 2**:

```
param_grid = {
    'learning_rate': [0.01, 0.05],
    'max_depth': [-1, 4, 8],
    'num_leaves': [15, 31, 40],
    'n_estimators': [300, 500, 1000],
    'subsample': [0.8],
    'colsample_bytree': [0.8]
}
```

Código 2 – Espaço de hiperparâmetros testado no Grid Search para LightGBM e XGBoost

Com vista a explorar ganhos adicionais, realizou-se um Grid Search simultâneo para LightGBM e XGBoost usando o espaço de procura apresentado no **Código 2**. Para o LightGBM o melhor conjunto foi learning_rate = 0.05, max_depth = 8, num_leaves = 31, n_estimators = 300, subsample = 0.8, colsample_bytree = 0.8, com **RMSE = 13.00** em validação cruzada; contudo, no *hold-out* de teste o efeito foi praticamente nulo: a média pesada do **MAE** passou de **4.90 → 4.88 (-0,4 %)**, o **MAPE** de **0.98 → 0.95 (-3 %)**, enquanto **MSE** e **RMSE** variaram marginalmente (438.22 → 439.34 e 14.68 → 14.71) e o **R^2** permaneceu em 0.82. Os benefícios concentraram-se em famílias residuais, como a F000014 (0,2 % da amostra, RMSE -20 %), sendo impercetíveis nas famílias de maior peso (F000009, F000013).

Para o **XGBoost**, o Grid Search seleccionou learning_rate = 0.01, max_depth = 8, n_estimators = 300, subsample = 0.8, colsample_bytree = 0.8, alcançando **RMSE = 14.14** em validação cruzada. No conjunto de teste, os efeitos foram marginais: o **MAE ponderado** passou de **4.35 → 4.33** (-0,5 %), o **RMSE** de **14.48 → 14.44** (-0,3 %) e o **R^2** permaneceu em **0.84**; em contrapartida, o **MSE** subiu ligeiramente (433.39 → 444.22, +2,5 %) e o **MAPE** agravou-se (0.57 → 0.71). A variação continua heterogénea entre famílias: a **F000009** melhorou (MAE -2,0 %), enquanto a mais representativa, **F000013**, registou degradação (+9,0 % em MAE).

Refira-se que todos os valores acima expostos podem ser encontrados, em formato tabela, no anexo 2. Para mais, o fundamento, definição e interpretação das diferentes métricas, encontram-se em detalhe na secção 5.1.

Em síntese, os ajustes finos produziram melhorias modestas que não compensam o acréscimo de complexidade e tempo de treino. Optou-se, portanto, por manter as configurações de referência, consideradas já adequadas ao equilíbrio entre desempenho preditivo e custo computacional.

Tabela 3 – Hiperparâmetros utilizados - visão por modelo

Modelo	Hiperparâmetro	Valor	Função / Justificação
XGBoost	n_estimators	**1000**	Número elevado de árvores em conjugação com *learning_rate* baixo melhora a robustez do *ensemble* (Chen & Guestrin, 2016).
	learning_rate	**0,05**	Taxa de aprendizagem pequena permite correções finas em cada iteração, reduzindo *overfitting*.
	max_depth	**6**	Limita a profundidade da árvore, controlando a complexidade do modelo e prevenindo *overfitting*, enquanto preserva capacidade para captar interações não-lineares.
	subsample	**0,80**	Amostragem estocástica de instâncias em cada iteração (alias bagging_fraction); diminui correlação entre árvores, fornece regularização adicional e acelera o treino (Ke et al., 2017)
	colsample_bytree	**0,80**	Amostragem de atributos por árvore (alias *feature_fraction*); reduz o risco de sobreajuste e o custo computacional, promovendo diversidade no *ensemble*.
LightGBM	n_estimators	**1000**	Mantida paridade de iterações para comparação direta com XGBoost (Ke et al., 2017).
	learning_rate	**0,05**	Taxa reduzida → maior estabilidade; típica em GBDT.
	num_leaves	**31**	Valor ≈ 2^max_depth; controla a complexidade da árvore folha-wise.
	max_depth	**6**	Limita a profundidade da árvore, controlando a complexidade do modelo e prevenindo *overfitting*, enquanto preserva capacidade para captar interações não-lineares.
	colsample_bytree	**0,80**	Amostragem de atributos por árvore (alias feature_fraction); reduz o risco de sobreajuste e o custo computacional, promovendo diversidade no *ensemble*.
	subsample	**0,80**	Amostragem estocástica de instâncias em cada iteração (alias bagging_fraction); diminui correlação entre árvores, fornece regularização adicional e acelera o treino (Ke et al., 2017)
	objective	**regression**	Função objetiva para problemas contínuos.
	metric	**rmse**	Métrica principal monitorizada em validação.
	Early Stopping	**paciência = 50**	Interrompe quando a métrica de validação estagna.

LSTM	Arquitetura	128 → 64 → Dense 64 ReLU → Dense 1	Duas camadas LSTM (128 e 64 unidades) capturam dependências temporais; duas *Dense* refinam a saída (Hochreiter & Schmidhuber, 1997).
	Otimizador	**Adam** (lr = 0,001)	Otimização adaptativa robusta (Kingma & Ba, 2014).
	Função perda	**MSE**	Penaliza mais fortemente erros grandes – comum em regressão.
	batch_size	**32**	Compromisso entre ruído de gradiente e desempenho computacional.
	epochs	**300** (máx.)	Permite convergência, limitada por *EarlyStopping* (patience = 10).
Transformer	num_heads	**2**	Duas cabeças de atenção suficientes para capturar relações sazonais num *dataset* moderado (Vaswani et al., 2017).
	key_dim	**8**	Dimensão interna da atenção; reduzida para limitar parâmetros.
	feed-forward units	**64**	Camada densa interna após a atenção; ativa com ReLU.
	dropout	**0,10**	Regulariza ligações; previne *overfitting* (Srivastava et al., 2014).
	Otimizador	**Adam** (lr = 0,001)	Tal como no LSTM.
	Função perda	**MAE**	Mais robusta a *outliers*, adequada a séries com picos ocasionais.
	batch_size	**32**	Consistente com LSTM.
	epochs	**60** (máx.)	Arquitetura mais pesada → menos ciclos *(epochs)*; *EarlyStopping* (patience = 10).

4.6. Visão geral do fluxo de trabalho

A Figura 18 seguinte sintetiza o fluxo de trabalho adotado ao longo deste capítulo, desde a importação dos ficheiros de entrada até à modelação preditiva. Estão representadas as principais etapas do *pipeline* — incluindo a preparação, análise exploratória, pré-processamento, e divisão dos dados — bem como os modelos de previsão utilizados (LightGBM, XGBoost, LSTM e Transformer).

Esta organização modular permite garantir a reprodutibilidade do processo e facilita o ajuste ou substituição de componentes específicos sem comprometer a integridade do sistema global.

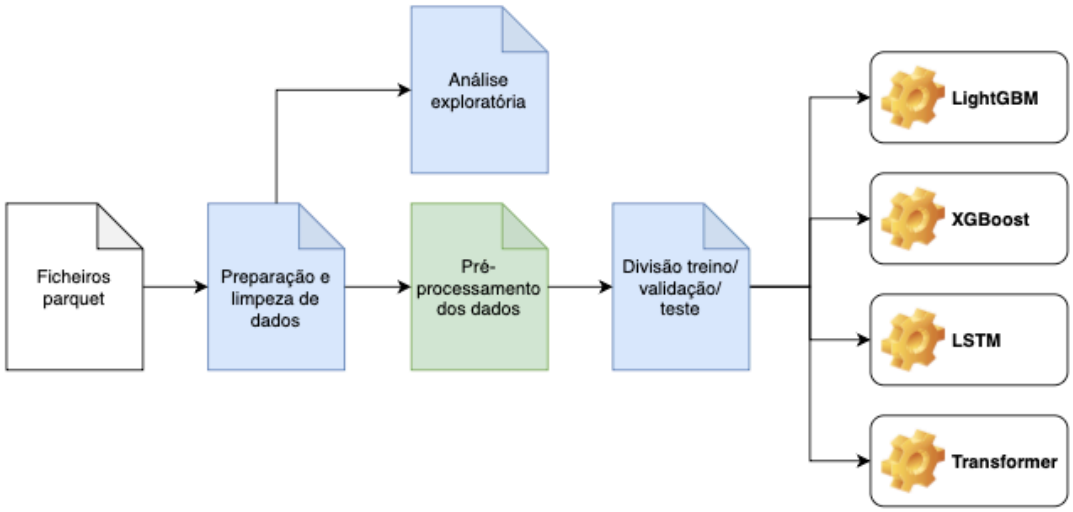


Figura 18 – Fluxo de trabalho do processo de análise e modelação preditiva

Finalizada a fase de preparação e modelação, passa-se agora à avaliação comparativa dos modelos desenvolvidos. No capítulo seguinte, analisa-se o desempenho preditivo de cada abordagem com base em métricas de erro consolidadas, explorando também a sua adequação a diferentes cenários de negócio

5. APRESENTAÇÃO E DISCUSSÃO DE RESULTADOS

Este capítulo insere-se na fase final do processo CRISP-DM, correspondendo às etapas de **avaliação** e **implementação**. Nele procede-se a uma análise crítica do desempenho dos quatro modelos testados – XGBoost, LightGBM, LSTM e Transformer – na previsão da procura de novos produtos em retalho. A avaliação assenta em seis métricas amplamente usadas em regressão (MAE, MSE, RMSE, MAPE, R^2) e considera dois cenários: global (modelo único treinado em todas as famílias) e segmentado por família (modelos independentes para cada uma das 13 famílias). O objetivo é identificar o(s) modelo(s) mais adequados a uma aplicação transversal e discutir se a avaliação deve privilegiar a perspetiva global ou a desagregação por família.

5.1. Métricas de avaliação

As métricas de avaliação quantificam a discrepância entre valores previstos e observados. O **MAE** (Erro Médio Absoluto) corresponde à média aritmética dos valores absolutos dos erros de previsão, ou seja, das diferenças |previsto – observado|, tendo a mesma unidade da variável-alvo (Hodson, 2022).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

O **MSE** (Erro Quadrático Médio) é a média dos quadrados dos erros, atribuindo maior peso aos erros mais extremos; o **RMSE** (Raiz do Erro Quadrático Médio) é simplesmente a raiz quadrada do MSE, devolvendo novamente a unidade original da previsão (Hodson, 2022). A vantagem do RMSE é manter a escala dos dados enquanto penaliza desproporcionalmente erros grandes.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

O **MAPE** (Erro Absoluto Percentual Médio) expressa o erro em termos relativos: calcula-se a média dos valores absolutos dos erros em percentagem dos valores reais. (Montaño Moreno et al., 2013). Esta métrica é adimensional e facilita comparações percentuais, mas pode ser instável quando os valores reais se aproximam de zero.

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

O **R^2** (Coeficiente de Determinação) indica a proporção da variância total dos dados observados explicada pelo modelo. Em regressão linear, R^2 varia entre 0 e 1 (quanto mais próximo de 1, melhor o ajuste) e é definido por $R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$,onde SSE é o somatório dos quadrados dos resíduos e SST o somatório total dos quadrados. Note-se que R^2 pode ser negativo em modelos que apresentam desempenho inferior à média dos dados, indicando ajuste não-positivo.

Em todas as métricas de avaliação apresentadas, foram utilizadas as seguintes definições para as variáveis:

- y_i: valor real (observado) da variável dependente no instante i;

- $\hat{y}_i$: valor previsto (estimado) pelo modelo no instante i;
- $\bar{y}$: média dos valores reais ao longo do período considerado;
- n: número total de observações no conjunto avaliado.

Refira-se que estas definições se mantêm constantes em todas as métricas de avaliação apresentadas ao longo do capítulo.

Finalmente, o número de observações do conjunto de teste utilizado em cada avaliação foi transversal ao longo de todos os testes, dado o explorado e justificado no capítulo **Balanceamento de dados.**

5.2. Comparação de desempenho dos modelos

Nesta secção apresentam-se os resultados métricos dos quatro modelos (XGBoost, LightGBM, LSTM e Transformer), avaliados tanto no contexto global (modelo treinado com os dados de todas as famílias) como no contexto por família (modelo treinado individualmente para cada família). Os critérios analisados incluem MAE (Erro Médio Absoluto), MSE (Erro Quadrático Médio), RMSE (Raiz do Erro Quadrático Médio), MAPE (Erro Percentual Absoluto Médio) e R^2 (Coeficiente de Determinação), além do número de amostras de teste. Note-se que o número de amostras de teste varia entre 731 e 3086 por família, refletindo a dimensão dos dados disponíveis para cada caso e representando o número de registos de vendas acumulados por semana.

A análise das tabelas de métricas (Tabela 4 a Tabela 7) obtidas revelam os resultados das métricas obtidas para cada um dos modelos implementados. A última linha de cada tabela resume o desempenho global através de uma média ponderada (peso = número de amostras de teste da família). No caso específico da Tabela 8, esta apresenta uma comparação dos desempenhos globais em cada um dos modelos aplicados. Esta abordagem evita que famílias com poucos registos distorçam a visão geral e, ao mesmo tempo, impede que séries muito volumosas dominem métricas simples de média aritmética.

5.2.1. Avaliação do modelo global

A leitura da linha **média pesada** presente na Tabela 4 a Tabela 8 deixa claro que os dois modelos baseados em árvores continuam a liderar ao nível dos resultados apresentados:

- **XGBoost** apresenta os menores valores agregados de erro (MAE, MSE e RMSE) e o MAPE mais baixo; além disso, mantém o melhor equilíbrio entre viés e variância, com R^2 positivo na média.
- **LightGBM** surge logo a seguir, com erros ligeiramente superiores, mas R^2 igualmente elevado.
- As redes neuronais (**Transformer** e, sobretudo, **LSTM**) ficam bastante atrás: os seus erros médios são múltiplos dos obtidos pelos modelos de árvore e a média ponderada de R^2 é negativa, indicando incapacidade de explicar a variabilidade global dos dados.

Em síntese, observam-se **dois patamares distintos** — um de alto desempenho (XGBoost ≈ LightGBM) e outro de desempenho modesto (Transformer ≈ LSTM).

5.2.2. Avaliação do modelo por família

Quando analisada cada família individualmente, confirma-se a supremacia dos modelos de árvore:

- **Consistência:** XGBoost e LightGBM obtêm, na maioria das famílias, MAE e RMSE inferiores aos das redes neuronais, mantendo R^2 positivo ou, pelo menos, próximo de zero mesmo em séries mais ruidosas.
- **Robustez em famílias pequenas:** Em agrupamentos com poucas amostras (por ex., F000008 ou F000014) ambos preservam erros muito baixos, enquanto LSTM e Transformer oscilam de forma acentuada.
- **Casos adversos:** Em famílias particularmente irregulares (ver F000005), todos os modelos sofrem alguma degradação, mas a diferença relativa continua a favorecer as árvores.
- **Benefício limitado de modelos dedicados:** Treinar um modelo distinto por família traz ganhos residuais (ou mesmo nulos) face ao modelo global, sobretudo porque o bom ajuste médio das árvores já cobre bem os padrões comuns; só em séries extremamente específicas poderia justificar-se a manutenção de modelos separados.

Tabela 4 – **LSTM** - tabela comparativa

	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	Família	Global	Família	Global	Família	Global	Família	Global	Família		
F000001	18.79	20.03	5209.81	5134.97	72.18	71.66	2.52	3.78	0.03	0.05	3086	9.8%
F000002	18.49	17.18	1833.93	1968.35	42.82	44.37	4.37	2.98	0.07	0.00	731	2.3%
F000003	12.28	19.03	414.59	687.04	20.36	26.21	3.06	5.33	-0.06	-0.76	2556	8.1%
F000005	29.99	31.48	1279.61	1044.75	35.77	32.32	9.14	16.42	-0.19	0.03	14	0.0%
F000006	12.43	12.88	917.62	1014.98	30.29	31.86	1.68	1.69	0.10	0.01	2930	9.3%
F000007	14.89	10.97	1482.83	1438.88	38.51	37.93	2.80	1.20	-0.10	-0.06	34	0.1%
F000008	15.56	2.33	278.14	12.93	16.68	3.60	10.22	0.66	-41.48	-0.97	37	0.1%
F000009	22.24	22.29	2077.14	2161.55	45.58	46.49	2.22	2.38	0.18	0.15	8003	25.3%
F000010	7.93	3.89	97.43	23.94	9.87	4.89	3.85	1.72	-3.16	-0.02	574	1.8%
F000011	11.73	13.05	483.74	544.02	21.99	23.32	2.08	2.16	0.33	0.25	3592	11.4%
F000012	21.33	19.83	1071.95	1249.64	32.74	35.35	4.81	3.06	0.11	-0.04	364	1.2%
F000013	7.35	6.40	423.89	418.15	20.59	20.45	2.25	1.35	-0.19	-0.18	9590	30.4%
F000014	1.58	1.56	4.31	3.67	2.08	1.91	0.73	0.85	-0.19	-0.02	64	0.2%
Média pesada	14.06	14.50	1397.23	1450.87	**33.45**	34.32	2.38	2.36	-0.07	-0.05		

Tabela 5 – **Transformer** - tabela comparativa

	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	Família	Global	Família	Global	Família	Global	Família	Global	Família		
F000001	20.75	21.42	5659.11	5726.40	75.23	75.67	1.52	1.51	-0.05	-0.06	3086	9.8%
F000002	16.84	18.24	2190.40	2311.67	46.80	48.08	0.86	0.88	-0.11	-0.17	731	2.3%
F000003	8.79	8.84	429.15	443.81	20.72	21.07	1.13	0.89	-0.10	-0.14	2556	8.1%
F000005	26.26	190.23	1657.31	39919.66	40.71	199.80	2.27	116.54	-0.54	-36.07	14	0.0%
F000006	14.58	13.48	1168.68	1097.30	34.19	33.13	1.31	1.18	-0.14	-0.07	2930	9.3%
F000007	12.13	52.78	1474.34	4002.43	38.40	63.26	1.38	17.80	-0.09	-1.96	34	0.1%
F000008	1.86	2.11	5.76	6.81	2.40	2.61	1.21	1.05	0.12	-0.04	37	0.1%
F000009	23.37	22.93	2691.90	2624.63	51.88	51.23	1.30	1.33	-0.06	-0.03	8003	25.3%
F000010	3.22	5.68	26.15	48.26	5.11	6.95	0.83	2.69	-0.12	-1.06	574	1.8%
F000011	12.91	13.88	757.31	827.15	27.52	28.76	1.31	1.11	-0.05	-0.15	3592	11.4%
F000012	19.67	20.46	1352.88	1466.42	36.78	38.29	2.26	1.60	-0.12	-0.22	364	1.2%
F000013	5.66	5.51	352.18	352.94	18.77	18.79	1.29	0.96	0.01	0.01	9590	30.4%
F000014	1.45	1.44	4.11	4.35	2.03	2.09	0.59	0.54	-0.14	-0.21	64	0.2%
Média pesada	13.91	14.03	1640.81	1657.28	**35.85**	35.98	1.30	1.25	-0.05	-0.09		

Tabela 6 – **LightGBM** - tabela comparativa

	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	Família	Global	Família	Global	Família	Global	Família	Global	Família		
F000001	7.73	8.33	3352.72	3470.22	57.90	58.91	1.17	1.34	0.38	0.36	3086	9.8%
F000002	7.41	7.45	686.64	741.71	26.20	27.23	0.98	0.61	0.65	0.62	731	2.3%
F000003	4.19	3.50	138.38	154.52	11.76	12.43	0.92	0.53	0.64	0.60	2556	8.1%
F000005	3.79	19.43	26.85	450.42	5.18	21.22	1.46	12.64	0.98	0.58	14	0.0%
F000006	4.25	3.93	84.15	94.27	9.17	9.71	0.91	0.54	0.92	0.91	2930	9.3%
F000007	11.34	9.22	1062.04	1366.12	32.59	36.96	1.25	0.44	0.22	-0.01	34	0.1%
F000008	2.68	1.40	9.57	4.07	3.09	2.02	2.05	0.58	-0.46	0.38	37	0.1%
F000009	6.51	6.12	212.09	208.19	14.56	14.43	0.78	0.62	0.92	0.92	8003	25.3%
F000010	2.15	1.44	6.57	4.07	2.56	2.02	1.06	0.56	0.72	0.83	574	1.8%
F000011	4.71	5.83	77.75	70.87	8.82	8.42	0.84	1.47	0.89	0.90	3592	11.4%
F000012	8.12	10.74	241.67	314.86	15.55	17.74	1.41	2.96	0.80	0.74	364	1.2%
F000013	2.96	2.66	29.45	30.74	5.43	5.54	1.14	0.86	0.92	0.91	9590	30.4%
F000014	2.30	1.07	6.50	1.84	2.55	1.36	1.51	0.52	-0.80	0.49	64	0.2%
Média pesada	4.90	4.83	438.22	453.14	**14.68**	14.89	0.98	0.88	0.82	0.82		

Tabela 7 – **XGBoost** - tabela comparativa

	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	Família	Global	Família	Global	Família	Global	Família	Global	Família		
F000001	6.96	7.15	3314.97	3431.03	57.58	58.58	0.89	0.64	0.38	0.36	3086	9.8%
F000002	6.93	7.57	779.59	1020.45	27.92	31.94	0.56	0.50	0.61	0.48	731	2.3%
F000003	4.21	3.92	107.00	199.29	10.34	14.12	0.79	0.57	0.73	0.49	2556	8.1%
F000005	2.34	11.24	18.95	165.07	4.35	12.85	0.32	4.54	0.98	0.85	14	0.0%
F000006	3.57	4.14	51.78	119.10	7.20	10.91	0.49	0.44	0.95	0.88	2930	9.3%
F000007	9.42	9.44	700.48	1380.60	26.47	37.16	0.82	0.50	0.48	-0.02	34	0.1%
F000008	1.09	1.31	2.60	3.86	1.61	1.96	0.66	0.61	0.60	0.41	37	0.1%
F000009	5.97	6.65	220.09	274.51	14.84	16.57	0.41	0.42	0.91	0.89	8003	25.3%
F000010	1.31	1.57	3.28	5.27	1.81	2.29	0.44	0.48	0.86	0.78	574	1.8%
F000011	4.28	4.92	84.85	83.40	9.21	9.13	0.44	0.55	0.88	0.88	3592	11.4%
F000012	7.26	8.64	232.76	351.66	15.26	18.75	0.95	0.74	0.81	0.71	364	1.2%
F000013	2.34	3.06	29.43	44.99	5.43	6.71	0.60	0.87	0.92	0.87	9590	30.4%
F000014	1.00	1.38	1.67	3.52	1.29	1.88	0.48	0.69	0.54	0.02	64	0.2%
Média pesada	4.35	4.90	433.39	484.59	**14.48**	16.21	0.57	0.62	0.84	0.78		

Tabela 8 – Desempenho médio ponderado dos modelos utilizados - abordagem global vs. por família de artigos

	MAE		**MSE**		**RMSE**		**MAPE**		**R2**	
Modelo	Global	Família	Global	Família	Global	Família	Global	Família	Global	Família
LightGBM	4.90	4.83	438.22	453.14	14.68	14.89	0.98	0.88	0.82	0.82
XGBoost	**4.35**	4.90	**433.39**	484.59	**14.48**	16.21	**0.57**	0.62	**0.84**	0.78
LSTM	14.06	14.50	1397.23	1450.87	33.45	34.32	2.38	2.36	-0.07	-0.05
Transformer	13.91	14.03	1640.81	1657.28	35.85	35.98	1.30	1.25	-0.05	-0.09

5.3. Análise qualitativa das previsões de vendas semanais

No gráfico de vendas do **LightGBM** (Figura 19) observa-se que modelo acompanha o perfil global da série, mas comete dois tipos de desvio. Nas primeiras semanas (março-abril 2024) sobrestima a procura: prevê 15 unidades quando as vendas reais se mantêm em ≈5. No pico de início de abril (15 unidades) já ocorre o inverso – prevê-se apenas ≈10. Entre junho e meados de setembro a linha prevista desliza suavemente de 8 para 7 unidades, sobrepondo-se razoavelmente aos valores reais, que oscilam entre 6 e 8. O grande pico de outubro (≈21) é captado quase por inteiro (≈20), mas o LightGBM demora a acompanhar a queda subsequente, prevendo ainda 14–10 unidades enquanto as vendas reais já recuam para 9–6. Em síntese, o modelo reproduz bem a sazonalidade e deteta os picos, mas tende a **sobrestimar o arranque da série e a cauda descendente** do principal pico.

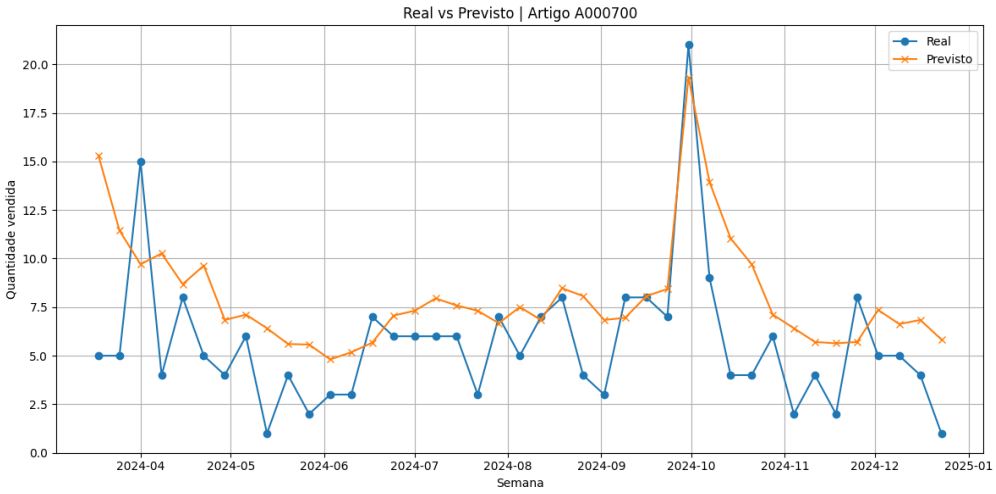


Figura 19 – previsão vendas semanais - artigo A000700 - **LightGBM**

Já o XGBoost (Figura 20) apresenta ajustes ligeiramente mais próximos dos valores reais. No arranque a sobrestimação é menor (8 previstas vs. 5 reais). O pico de abril (15) é subestimado por 3 unidades, e o de outubro (21) por 6 unidades (prevê 15). Por outro lado, a descida pós-pico é seguida com maior rapidez; a previsão cai para 9–7 quando as vendas baixam para 9–6, reduzindo o erro observado no LightGBM. De forma geral, a linha laranja mantém-se próxima à azul durante o verão e início de outono, sinal de **boa aderência na zona média da distribuição**, apesar de falhar a amplitude exata dos picos.

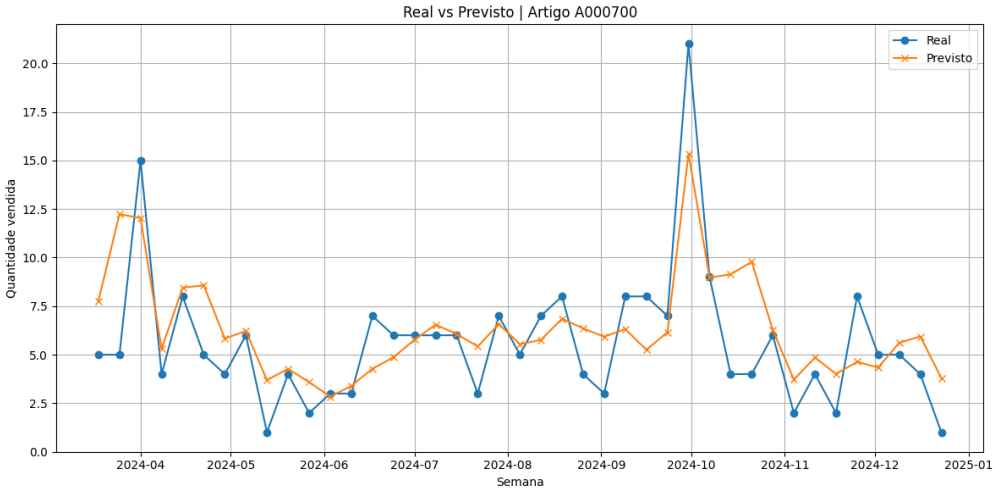


Figura 20 – previsão vendas semanais - artigo A000700 - **XGBoost**

O modelo **LSTM** (Figura 21) suaviza acentuadamente a série. Desde março até finais de maio mantém-se em 8–9 unidades, superando persistentemente vendas reais que raramente passam de 5; no pico de abril prevê apenas 9 contra 15. Após junho fixa-se num patamar quase plano de 3–4 unidades, o que leva a subestimações sucessivas dos pequenos picos do verão e a um erro grave no pico de outubro (21 reais vs. 7 previstos). O LSTM **eleva o nível de base e "corta" sistematicamente os máximos**, mostrando dificuldade em captar variações de amplitude.

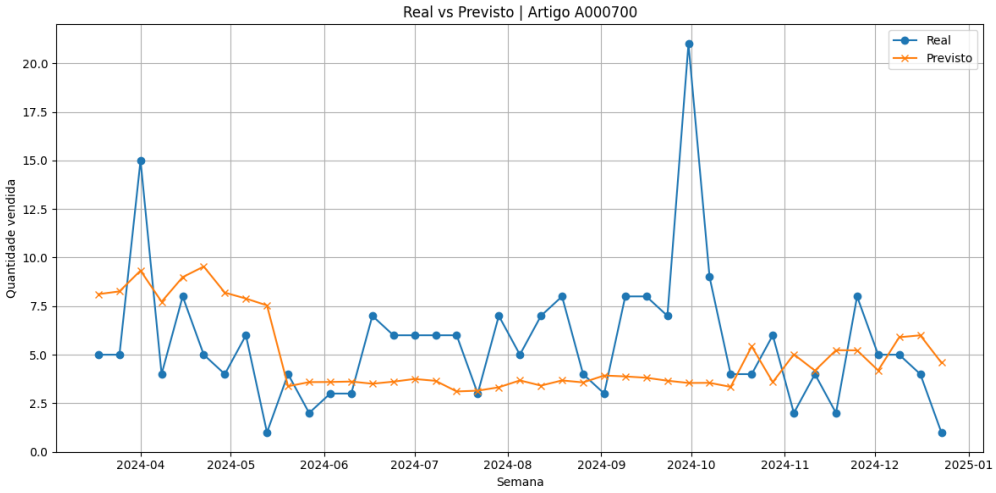


Figura 21 – previsão vendas semanais - artigo A000700 - **LSTM**

Já o modelo **Transformer** (Figura 22) apresenta o comportamento mais extremo. Durante março-maio prevê 30–26 unidades, ou seja, **quase seis vezes** o real. Em junho ocorre um corte abrupto: as previsões colapsam para ≈7 e permanecem virtualmente horizontais até ao fim do ano, ignorando tanto os pequenos picos intermédios como o pico principal de outubro (21 reais vs. 7 previstos). O modelo, portanto, **exagera a procura inicial e depois mantém um patamar constante**, falhando quer a tendência descendente do primeiro trimestre quer as oscilações sazonais seguintes.

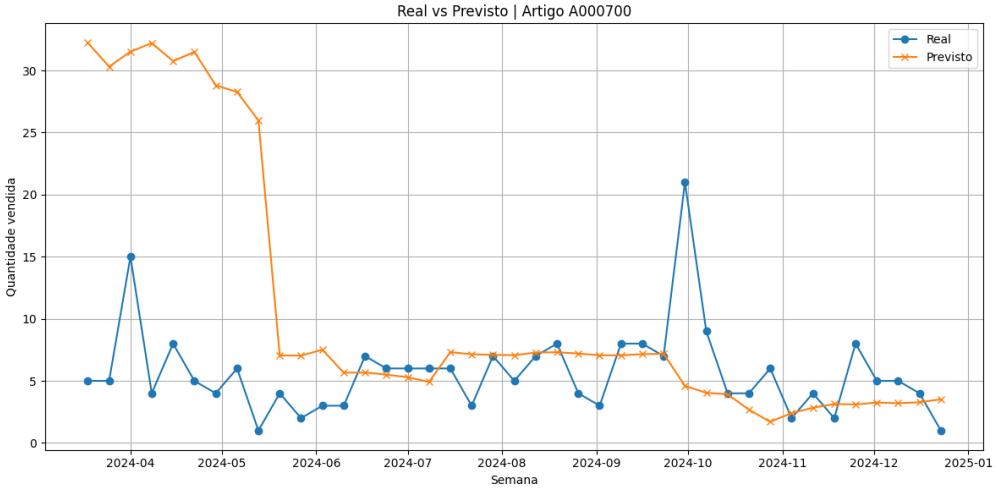


Figura 22 – previsão vendas semanais - artigo A000700 – **Transformer**

Em conjunto, os dois algoritmos de *gradient boosting* continuam a oferecer o ajuste mais fiel; o LightGBM captura melhor o pico de outubro, ao passo que o XGBoost reage mais depressa à sua descida. Já as redes neuronais (LSTM e Transformer) mantêm a tendência de amortecer a série, com especial fragilidade do Transformer na fase inicial e nos picos de maior amplitude.

5.4. Modelo recomendado para estimar o desempenho de novos produtos a partir de dados históricos de artigos similares

Com base nos indicadores apresentados no decorrer da presente secção, o **XGBoost** é o modelo preditivo mais adequado para aplicação transversal (um único modelo para todas as famílias de artigos). Este combina o menor erro médio absoluto e quadrático ponderados, o RMSE mais baixo e o melhor MAPE, além de manter R^2 consistentemente positivos. O **LightGBM** apresenta desempenho muito próximo e permanece como alternativa praticamente equivalente, mas já não supera o XGBoost em nenhuma métrica agregada.

Em contraste, os modelos de redes neurais sequenciais (**LSTM** e **Transformer**) mostraram-se menos eficazes: os seus erros médios foram várias vezes superiores e o R^2 ficou negativo em muitas famílias, sinalizando dificuldade em captar a variabilidade das séries temporais de vendas.

Adicionalmente, a linha **média pesada** de cada tabela resume o desempenho real do sistema, porque pondera cada métrica pelo número de amostras de teste disponíveis em cada família. Dessa forma, evita-se que famílias pouco representativas distorçam a avaliação e confirma-se a liderança do XGBoost, seguida de perto pelo LightGBM, enquanto as redes neuronais ocupam posições inferiores.

Considerando que a empresa parceira do projeto — responsável pelo fornecimento dos dados — adotou o **RMSE como métrica de avaliação prioritária**, e visto que o XGBoost detém agora o menor RMSE ponderado global, recomenda-se a sua adoção como modelo principal para estimar o desempenho de novos produtos. Este modelo global aproveita similaridades entre séries históricas de artigos distintos, permitindo prever vendas de lançamentos com histórico inexistente ou limitado. Além disso, oferece:

- **Robustez** a outliers e variações abruptas típicas do retalho;
- **Eficiência** de treino e inferência, facilitando atualizações frequentes;
- **Baixa complexidade operacional**, pois dispensa a gestão de múltiplos modelos por família.

Em suma, a escolha do **XGBoost – modelo global** maximiza valor preditivo com custo de complexidade mínimo, alinhando-se às exigências técnicas e operacionais do negócio. No capítulo seguinte, demonstra-se a integração deste modelo num protótipo de plataforma com interface gráfica, evidenciando o seu uso prático em contexto comercial.

O presente capítulo apresentou uma avaliação comparativa dos modelos preditivos, recorrendo a diferentes métricas de desempenho e estratégias de segmentação, tanto ao nível global como por família. A análise permitiu identificar pontos fortes e limitações de cada abordagem, contribuindo para uma compreensão mais sólida da sua aplicabilidade ao contexto em estudo. No capítulo seguinte, será demonstrada a operacionalização desses resultados através da sua integração num protótipo de *dashboard*, concebido para apoiar a tomada de decisão comercial de forma prática e acessível.

6. APLICAÇÃO *WEB – DASHBOARD*

No presente capítulo, será apresentado um componente complementar desenvolvido no âmbito desta dissertação — ainda que não estivesse inicialmente previsto nos objetivos. Trata-se de um ***dashboard web* interativo**, concebido com o intuito de facilitar a exploração dos resultados pela empresa fornecedora dos dados, permitindo a sua utilização autónoma e eventual integração no ambiente real da organização.

6.1. Interface e funcionalidades

Este capítulo apresenta a implementação prática dos modelos preditivos através de uma plataforma *web* (*frontend*), ilustrando um protótipo funcional aplicável em contexto corporativo.

A arquitetura proposta, apresentada na Figura 23, assenta em três camadas: **dados**, **aplicação** e **apresentação**. Na **camada de dados**, foi criada uma base de dados (SQLite) a partir dos ficheiros *Parquet* partilhados, com vista à futura integração nos sistemas da organização. Após a definição das tabelas e formatos, a informação é consumida pela **camada de aplicação**, onde um serviço desenvolvido em *Flask* gere a lógica de previsão com base nos modelos previamente treinados. Assim que o utilizador submete uma solicitação, o modelo adequado é acionado e devolve as previsões esperadas.

A **camada de apresentação** é composta por uma aplicação web em formato *dashboard*, que visa acrescentar valor em dois eixos principais: 1) Apoio à gestão e tomada de decisão, com métricas descritivas e operacionais ao nível de família, subfamília, artigo e loja; 2) Apoio comercial, com previsões de vendas que sustentam a negociação e inclusão de novos artigos nas lojas do retalho.

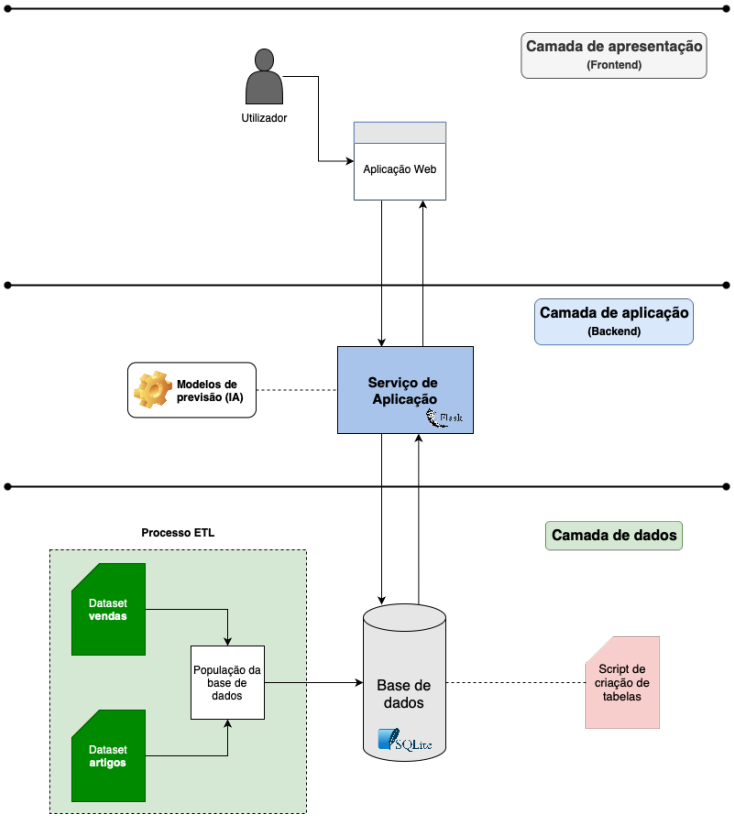


Figura 23 – arquitetura *WEB*

A Figura 24 apresenta a página inicial do *dashboard*, onde são destacados indicadores de performance global da empresa ao nível de vendas, acompanhados por uma tabela com os movimentos mais recentes, visível na secção inferior da página.

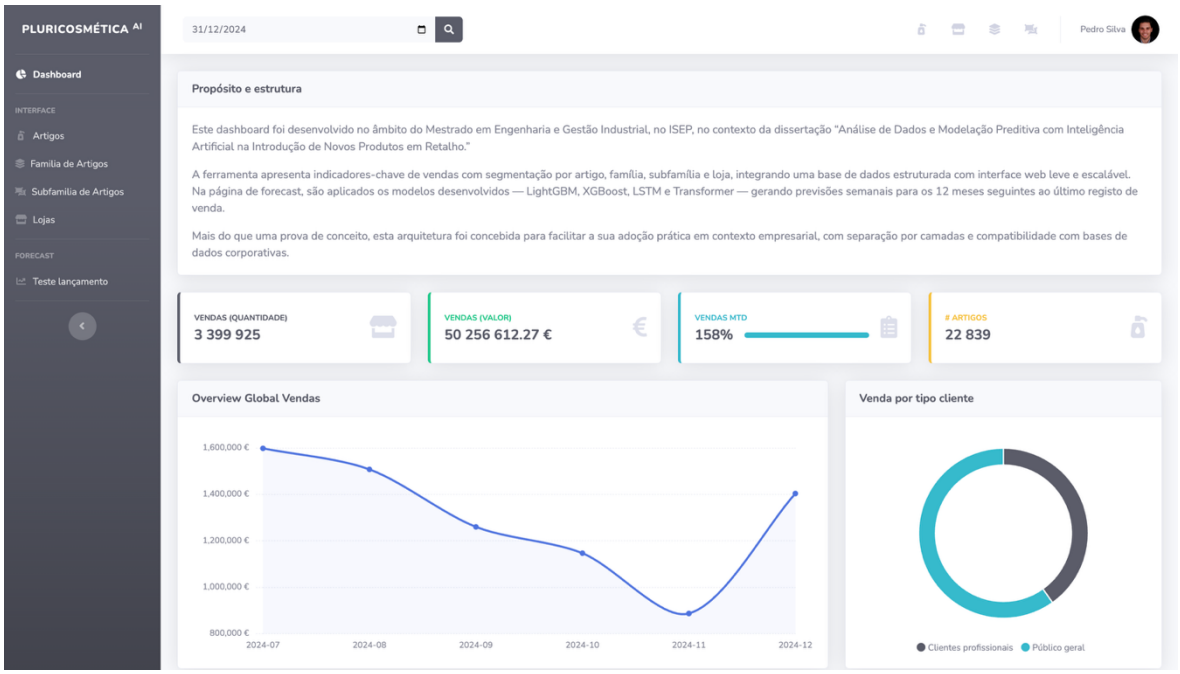


Figura 24 – Página inicial – *Dashboard*

A Figura 25 apresenta uma visão agregada e estruturada ao nível do artigo, com informação consolidada para consulta e análise. Páginas descritivas equivalentes foram desenvolvidas também para os níveis de família, subfamília e loja, mantendo uma disposição semelhante e permitindo a navegação por diferentes níveis hierárquicos da organização do portefólio.

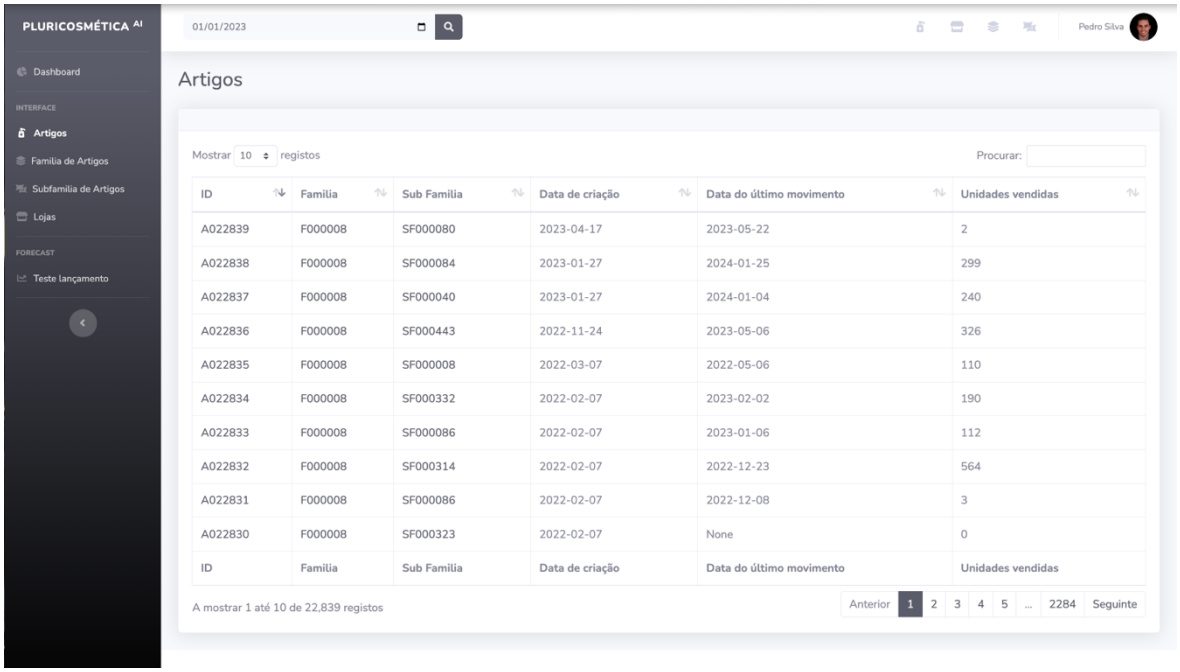


Figura 25 – Página descritiva - Artigos – *Dashboard*

As Figura 26 a Figura 28 correspondem às páginas de detalhe, que seguem uma estrutura comum: indicadores de desempenho no topo e tabelas com métricas específicas na parte inferior. O conteúdo apresentado adapta-se ao nível de análise — artigo, família/subfamília ou loja — permitindo uma exploração mais granular e orientada à tomada de decisão. Em cada uma destas páginas é possível exportar um ficheiro PDF com os resultados visíveis, facilitando a partilha com terceiros. Adicionalmente, na página de detalhe do artigo, encontra-se disponível uma ligação direta para o módulo de previsão, permitindo aplicar os modelos ao artigo em análise de forma imediata.

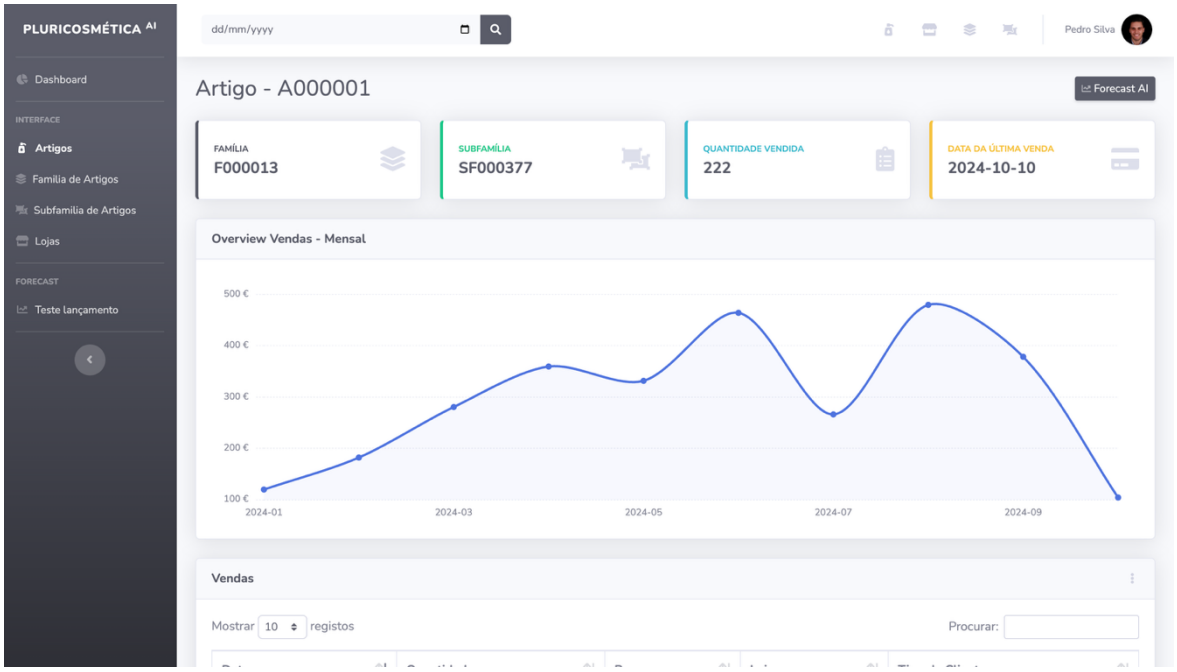


Figura 26 – Detalhe ao nível do artigo - *Dashboard*

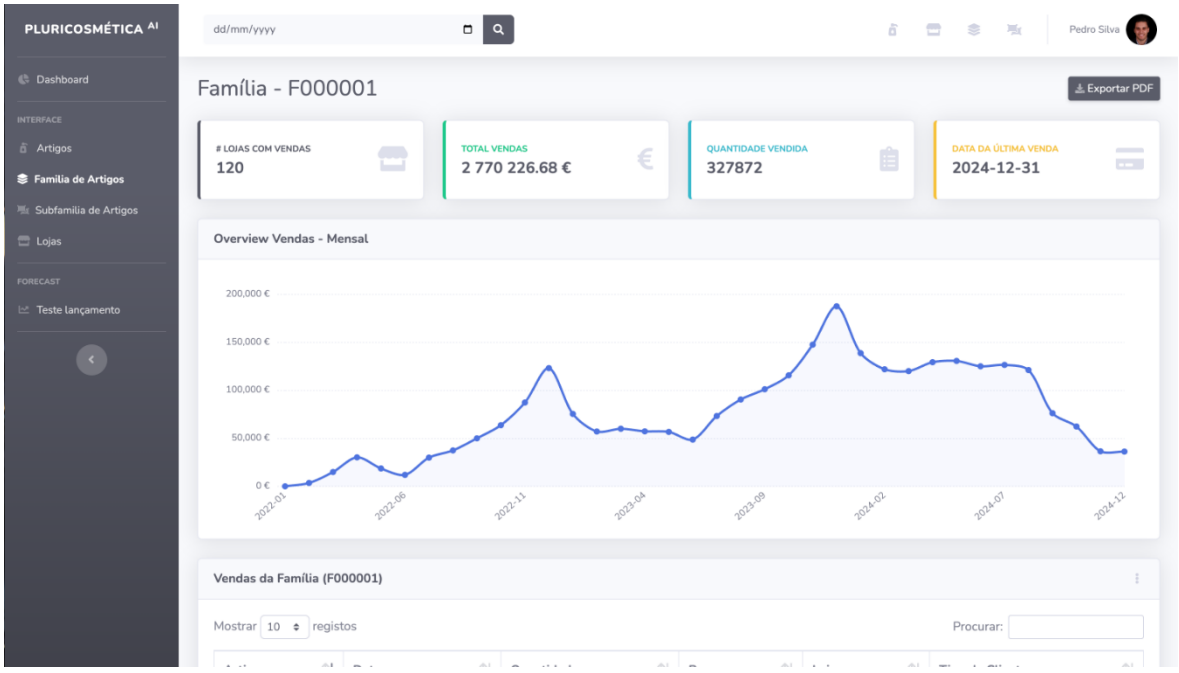


Figura 27 – Detalhe ao nível da Família - *Dashboard*

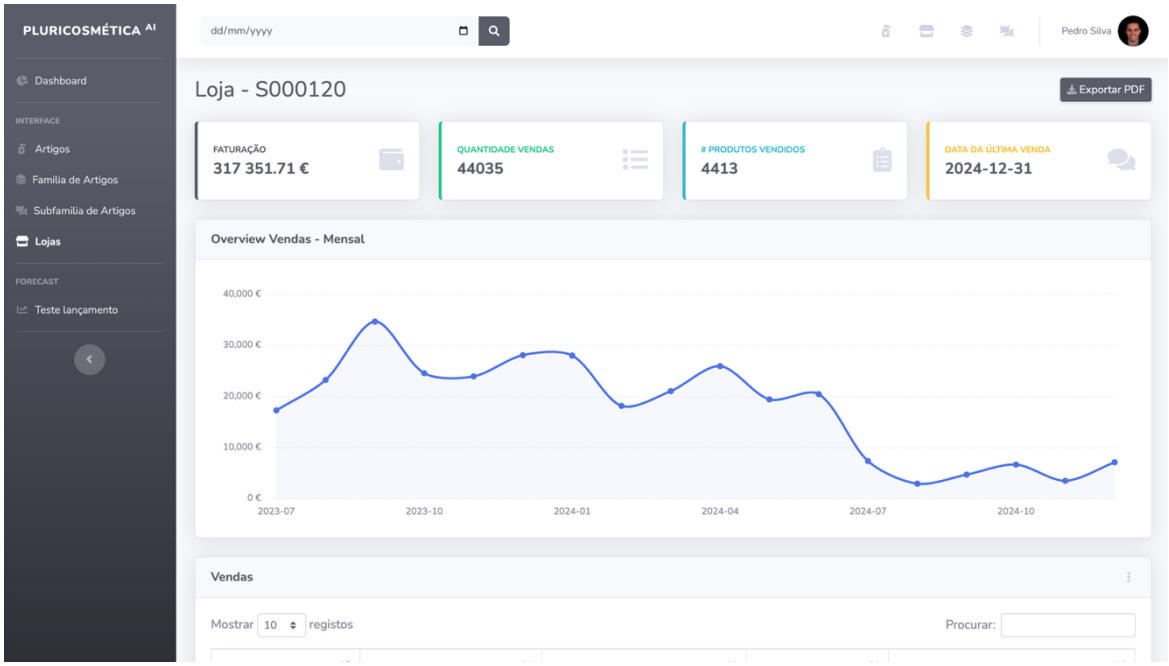


Figura 28 – Detalhe ao nível da loja – *Dashboard*

A página de *forecast* (Figura 29) permite aplicar os modelos desenvolvidos à previsão de vendas futuras, em dois cenários distintos:

1. **Artigo novo (sem histórico)**, mas semelhante a outro já existente: o utilizador preenche um formulário com as características relevantes e submete o pedido.
2. **Artigo já existente**: basta selecionar o ID do artigo, sendo a previsão gerada automaticamente.

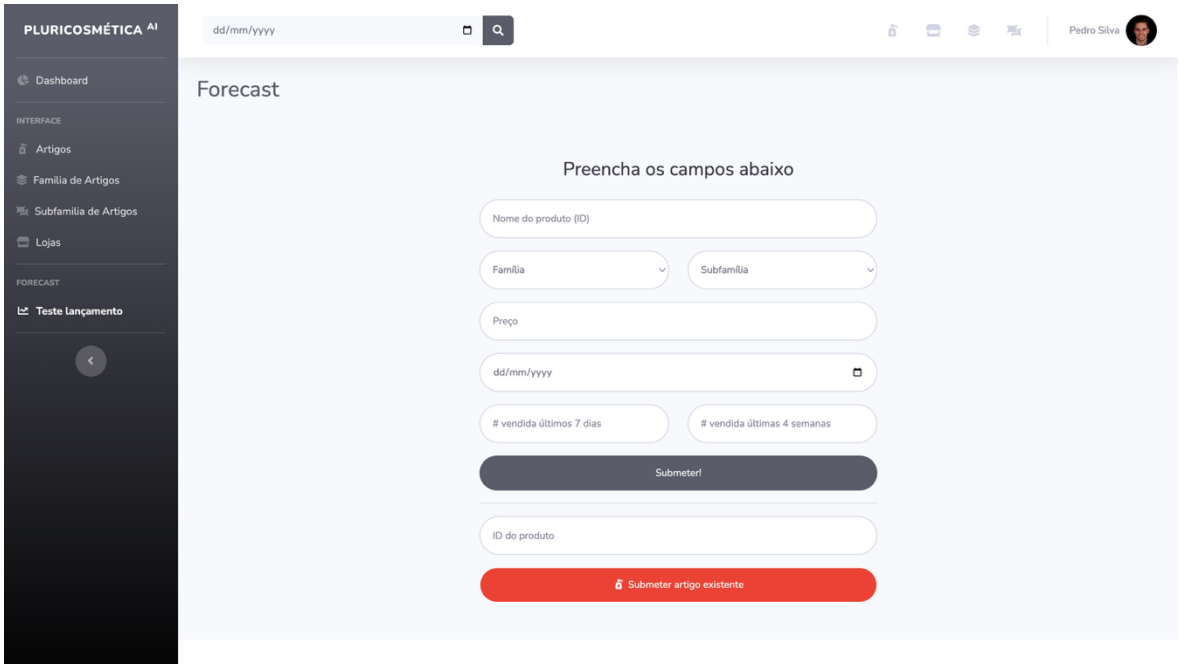


Figura 29 – *Forecast* - Página de submissão

Após a submissão, o sistema aplica todos os modelos desenvolvidos (LightGBM, XGBoost, LSTM e Transformer), em ambos os contextos — **global** e **por família**. O *output* (Figura 30) inclui dois gráficos (um para cada abordagem), com previsões semanais para os 12 meses seguintes à última venda registada. São ainda apresentados os identificadores (Família, Subfamília) e indicadores adicionais de contexto, como a quantidade vendida nos últimos 7 e 30 dias.

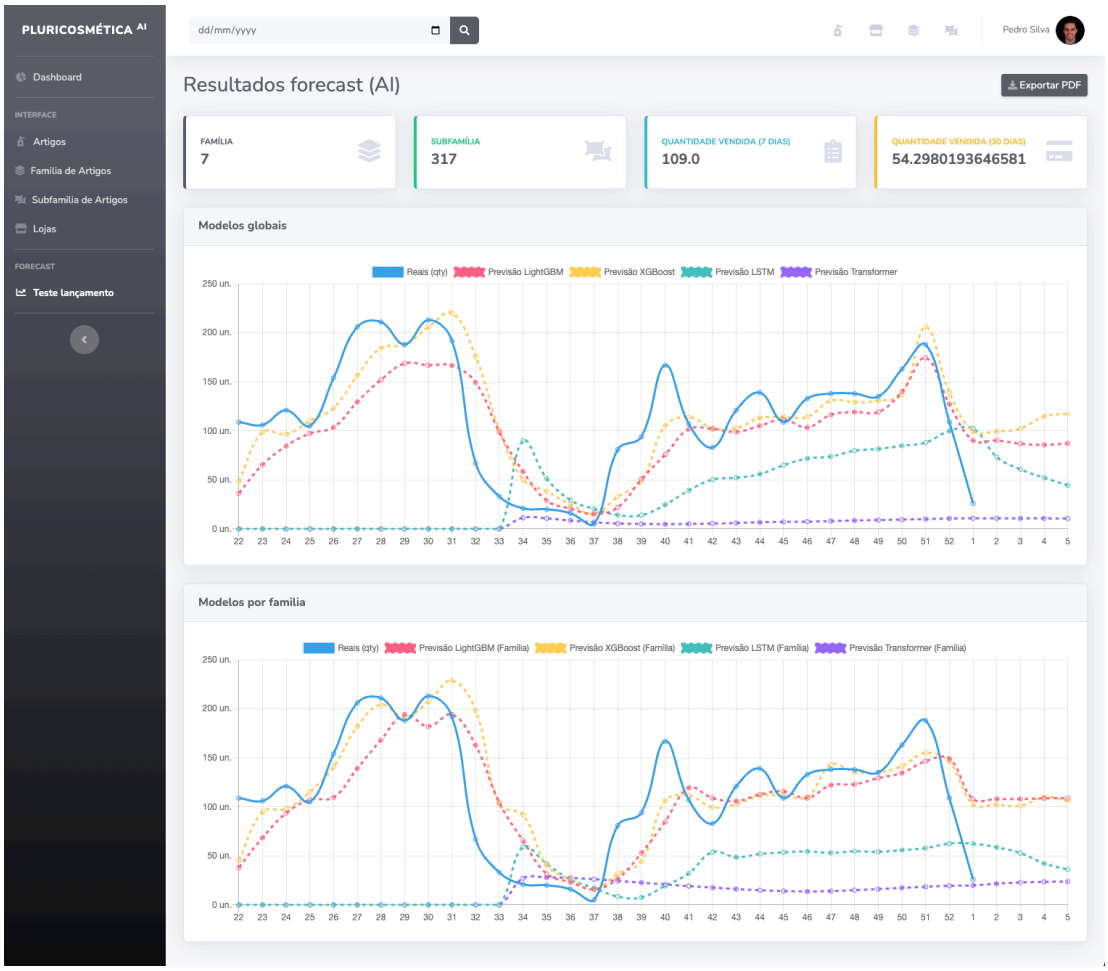


Figura 30 – Forecast - Resultados - Dashboard

Este *dashboard* demonstra como a previsão de vendas de novos artigos pode ser integrada em ferramentas operacionais e comerciais, promovendo decisões mais informadas e alinhadas com os padrões históricos observados.

Mais do que uma demonstração técnica, a estrutura desenvolvida foi pensada com **portabilidade e aplicabilidade real** em mente. A criação de uma **base de dados estruturada (SQLite)** com origem nos ficheiros de vendas e artigos em formato *Parquet* permite uma **integração ágil com sistemas corporativos existentes**, bastando, por exemplo, substituir o motor de base de dados (e.g., SQLite por PostgreSQL ou SQL Server) e configurar as ligações apropriadas.

A separação por **camadas (dados, aplicação, apresentação)** reforça essa escalabilidade: a camada de dados pode ser facilmente substituída ou estendida com novas fontes; o **serviço de aplicação (*Flask*)** pode ser colocado num servidor interno da organização ou adaptado para APIs RESTful mais robustas; e a **camada de apresentação (*dashboard*)** pode ser disponibilizada num ambiente web seguro, acessível às equipas de gestão, comerciais ou de planeamento. Além disso, a modularidade do *pipeline* de dados e modelação (Figura 18) permite que cada etapa – desde o pré-processamento

até à previsão com diferentes algoritmos – seja executada de forma independente ou programada regularmente, com possibilidade de automatização via scripts agendados (e.g., *cron jobs* ou serviços internos).

Assim, a estrutura apresentada não só prova a viabilidade técnica da abordagem proposta, como oferece um **caminho direto para adoção real em contexto empresarial**, com ganhos potenciais em previsibilidade de novos lançamentos, gestão de stock e alinhamento comercial com fornecedores e lojas.

6.2. Validação com utilizador-chave da empresa

A avaliação do *dashboard* foi conduzida exclusivamente pelo colaborador que assegurou a ligação entre o ISEP (Instituto Superior de Engenharia do Porto) e a empresa. Este profissional acompanhou todo o desenvolvimento, forneceu requisitos e testou iterações intermédias, pelo que o seu *feedback* é representativo das necessidades do negócio.

6.2.1. Objetivo

O processo de validação teve como objetivo verificar de forma rápida e objetiva se o *dashboard*:

1. **É útil** para apoiar as decisões quotidianas da área;
2. **É fácil de usar** sem formação adicional;
3. É capaz de reunir **aceitação global** (satisfação e intenção de recomendação).

6.2.2. Instrumento de recolha

Tendo por base os objetivos supra, foi aplicado um pequeno questionário de 5 (cinco) questões em escala *Likert* de 1 (discordo totalmente) a 5 (concordo totalmente). Os itens foram extraídos de instrumentos validados, conforme se verifica na Tabela 9:

Tabela 9 – Questionário aplicado ao utilizador-chave: afirmações e fontes

#	Questão	Fonte teórica
1	"As funcionalidades do *dashboard* satisfazem as necessidades da minha função."	UMUX-Lite[9] – utilidade percecionada (Lewis et al., 2013)
2	"Considero o *dashboard* fácil de usar."	UMUX-Lite / TAM[10] – facilidade de utilização (Davis, 1989)
3	"Consigo executar as minhas tarefas de forma eficiente utilizando o *dashboard*."	ISO 9241-11[11] – eficiência

[9] **UMUX-Lite** – *Usability Metric for User Experience – Lite*: versão reduzida de dois itens para avaliação rápida de usabilidade, focando utilidade e facilidade de uso.
[10] **TAM** – *Technology Acceptance Model*: modelo teórico que explica a aceitação de tecnologia com base na utilidade percecionada e facilidade de uso
[11] **ISO 9241-11** – Norma internacional que define usabilidade como a combinação de eficácia, eficiência e satisfação num contexto de uso.

4	"Estou globalmente satisfeito com o *dashboard*."	SUS[12] – satisfação (Brooke, 1996)
5	"Recomendaria este *dashboard* a outros colegas da organização."	Métrica de recomendação (NPS[13])

6.2.3. Procedimento

O processo decorreu em quatro momentos sequenciais:

1. **Demonstração breve** do *dashboard* (≈ 5 min).
2. **Exploração livre** pelo utilizador em cenários reais (≈ 10 min).
3. **Preenchimento do questionário** em ambiente online na plataforma *MS Forms* (≈ 2 min).
4. **Entrevista informal** para comentários qualitativos e sugestões (≈ 10 min).

A duração total foi de aproximadamente **30 minutos**.

6.2.4. Limitações

A avaliação envolveu apenas uma pessoa, impossibilitando inferências estatísticas profundas. Contudo, segundo Brooke (1996) e Lewis et al. (2013), instrumentos curtos como SUS ou UMUX-Lite preservam boa sensibilidade mesmo neste contexto, sendo adequados para avaliações exploratórias de usabilidade em contextos com amostras reduzidas.

6.2.5. Resultados

Depois de aplicado o questionário, seguiram-se os resultados. Quatro das cinco afirmações obtiveram a pontuação máxima (5). Apenas a facilidade de utilização recebeu 4 (“Concordo parcialmente”). O tom geral do feedback é, por isso, claramente positivo.

Com quatro respostas 5 e uma resposta 4, a média aritmética é 4,8/5. Considerando apenas os dois itens do UMUX-Lite (utilidade e facilidade) obtém-se 4,5/5. Convertendo esta pontuação para a escala 1–7 e aplicando a regressão de Lewis et al. (2013) para equivalência na métrica SUS, obtém-se cerca de **80/100**, valor geralmente enquadrado no nível *bom*.

A pontuação máxima atribuída em utilidade, eficiência, satisfação e intenção de recomendação confirma que o *dashboard* cumpre eficazmente a sua função de apoio às decisões operacionais do negócio. A única dimensão com margem de melhoria foi a facilidade de utilização.

O utilizador referiu, em particular, que o formulário de submissão de previsões não torna evidente se deve preencher a secção relativa a “Artigo existente” ou a de “Novo artigo”. Neste seguimento, recomenda-se rever o formulário para que fique logo claro que existem dois percursos distintos —

[12] **SUS** – *System Usability Scale*: escala de 10 itens desenvolvida por Brooke (1996), amplamente utilizada na avaliação da usabilidade de sistemas interativos.
[13] **NPS** – *Net Promoter Score*: métrica de marketing que avalia a intenção de recomendação de um produto ou serviço, sendo um indicador indireto de satisfação e lealdade.

“Artigo existente” ou “Novo artigo” — acompanhados de uma instrução sucinta em cada um; a secção não selecionada deve permanecer oculta ou desativada, evitando quaisquer dúvidas.

Relativamente à exportação de dados, além do PDF já disponível, deve acrescentar-se a possibilidade de descarregar os resultados diretamente em formato Excel.

7. CONCLUSÕES FINAIS

Os objetivos centrais da investigação passaram por identificar modelos capazes de antecipar a procura de novos produtos a partir de artigos historicamente semelhantes (Q1) e avaliar se um modelo único, treinado em todas as famílias, supera abordagens especializadas por família (Q2).

Primeiro, a comparação exaustiva de quatro algoritmos—XGBoost, LightGBM, LSTM e Transformer—mostrou a existência de dois patamares de desempenho bem distintos. As árvores de gradiente (XGBoost ≈ LightGBM) obtiveram, de forma consistente, os menores MAE, RMSE e MAPE e mantiveram R^2 positivos, ao passo que as redes neuronais LSTM e Transformer registaram erros várias vezes superiores e R^2 negativos em diversas famílias. Assim, responde-se à **Q1: os modelos baseados em *gradient boosting* são a opção mais eficaz para prever vendas iniciais** de novos artigos no retalho, fornecendo uma relação robusta entre precisão e simplicidade operacional.

Relativamente à **Q2**, o estudo revelou que **treinar um único modelo global não só simplifica a operação como garante desempenho igual ou superior ao de modelos específicos por família**. A média ponderada de erros comprova a liderança do XGBoost global, enquanto ganhos marginais obtidos com modelos dedicados não justificam a complexidade acrescida de manter dezenas de instâncias independentes. Recomendou-se, por isso, a adoção de **um único XGBoost global** como motor de previsão transversal, aproveitando a diversidade das séries históricas para generalizar

7.1. Limitações da investigação

Embora o conjunto de dados partilhado inclua a **data de criação de cada artigo** e algumas variáveis temporais derivadas (e.g., *week_since_creation*), o número de atributos continua reduzido a identificadores hierárquicos (família, subfamília) e a métricas transacionais básicas (quantidade vendida, preço). A falta de descritores de marketing—marca, formato, cor, posicionamento de preço relativo ou sinalização de promoções—limita o universo de *features* disponíveis, restringindo simultaneamente a explicabilidade dos modelos e a captura de fatores latentes de procura.

Cada SKU possui **no máximo 365 dias de vendas**, o que impede observar ciclos sazonais de vários anos e avaliar o desempenho dos modelos em fases de maturidade ou declínio. Além disso, não foram incorporados fatores externos (promoções específicas, ruturas de stock, meteorologia ou indicadores macroeconómicos), pelo que choques de procura exógenos permanecem não modelados.

Uma dificuldade adicional surgiu da **inconsistência taxonómica**: certas subfamílias apareciam associadas a mais de uma família. Esta ambiguidade foi mitigada ao tratar cada artigo pelo par/tupla **(família, subfamília)**, criando um rótulo único que conserva a informação hierárquica sem forçar uma pertença exclusiva a uma só família. Embora esta solução evite duplicações, continua a depender de classificações internas da empresa que podem evoluir ao longo do tempo.

7.2. Recomendações para próximas investigações

Como continuação natural deste trabalho, sugere-se o desenvolvimento de modelos de previsão **SKU–loja**, explorando diferenças micro geográficas nos perfis de cliente, poder de compra e

padrões sazonais regionais. Estratégias possíveis incluem *transfer learning* a partir do modelo XGBoost global para *clusters* de lojas, ou ainda modelos hierárquicos que permitam partilhar informação entre pontos de venda com histórico escasso.

Paralelamente, importa enriquecer o conjunto de atributos com variáveis de promoção, calendários comerciais, condições meteorológicas e indicadores de procura online, de forma a aumentar o poder explicativo e reduzir o erro residual. À medida que se acumulem dados para além de um ano, tornar-se-á viável testar os algoritmos em ciclos de vida completos e incorporar mecanismos de *demand sensing*. A aplicação sistemática de técnicas de Explainable AI (como SHAP) permitirá, por sua vez, reforçar a interpretabilidade dos modelos e facilitar a sua adoção pelas equipas de negócio.

Adicionalmente, recomenda-se a exploração de **modelos híbridos**, que conjuguem as abordagens já desenvolvidas com técnicas estatísticas complementares, capazes de isolar componentes como tendência, sazonalidade e eventos exógenos antes da previsão principal. Esta estratégia poderá passar, por exemplo, pela decomposição da série temporal para remoção de padrões regulares, aplicando-se depois o modelo de *machine learning* à componente residual — o que potenciará a deteção de relações não-lineares subtis.

A empresa envolvida no projeto manifestou já interesse em realizar uma **prova de conceito com dados reais de produção**, comparando o desempenho desta abordagem híbrida com o *pipeline* atualmente em uso. Tal ensaio fornecerá evidência empírica sobre os ganhos de precisão e robustez operacional que poderão ser alcançados.

Em síntese, esta investigação confirma que um modelo XGBoost global, treinado com dados agregados de todas as famílias, é suficiente para fornecer previsões iniciais fiáveis, com baixa complexidade operacional. No entanto, ampliar o leque de atributos dos artigos, desenvolver previsões ao nível loja e validar os modelos em contexto real abrirá espaço para ganhos adicionais de precisão e decisões de portefólio mais localizadas.

REFERÊNCIAS BIBLIOGRÁFICAS

Aktas, E., & Meng, Y. (2017). An Exploration of Big Data Practices in Retail Sector. *Logistics*, *1*(2), 12. https://doi.org/10.3390/logistics1020012

Amar, J., Rahimi, S., Surak, Z., & von Bismarck, N. (2022). *AI-driven operations forecasting in data-light environments*. https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments

Andrade, L. A. C. G., & Cunha, C. B. (2023). Disaggregated retail forecasting: A gradient boosting approach. *Applied Soft Computing*, *141*, 110283. https://doi.org/10.1016/j.asoc.2023.110283

Anitha S., & Neelakandan R. (2025). Demand Forecasting New Fashion Products: A Review Paper. *Journal of Forecasting*, *44*(2), 270–280. https://doi.org/10.1002/for.3192

Aras, S., Deveci Kocakoç, İ., & Polat, C. (2017). Comparative study on retail sales forecasting between single and combination methods. *Journal of Business Economics and Management*, *18*(5), 803–832. https://doi.org/10.3846/16111699.2017.1367324

Avelino, J. G., Cavalcanti, G. D. C., & Cruz, R. M. O. (2024). Resampling strategies for imbalanced regression: a survey and empirical analysis. *Artificial Intelligence Review*, *57*(4), 82. https://doi.org/10.1007/s10462-024-10724-3

Babai, M. Z., Boylan, J. E., & Rostami-Tabar, B. (2022). Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. *International Journal of Production Research*, *60*(1), 324–348. https://doi.org/10.1080/00207543.2021.2005268

Basavaraju, K., & Fatahi Valilai, O. (2025). Developing a demand planning strategy for joint forecasting and employing analytical tool in an empirical case study. *Discover Applied Sciences*, *7*(4). https://doi.org/10.1007/s42452-025-06740-9

Baum, D., Spann, M., Füller, J., & Thürridl, C. (2019). The impact of social media campaigns on the success of new product introductions. *Journal of Retailing and Consumer Services*, *50*, 289–297. https://doi.org/10.1016/j.jretconser.2018.07.003

Belarbi, H., Hassan, U., Bennis, H., Haj, E., Mohammed, T., Tajmouati, A., El, M., & Tirari, H. (2016). *Predictive Analysis of Big Data in Retail Industry Predictive Analysis of Big Data in Retail Industry Literature Review*. https://www.researchgate.net/publication/311900279

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. In *Journal of Machine Learning Research* (Vol. 13). http://scikit-learn.sourceforge.net.

Bocean, C. G., & Vărzaru, A. A. (2023). EU countries' digital transformation, economic performance, and sustainability analysis. *Humanities and Social Sciences Communications*, *10*(1), 875. https://doi.org/10.1057/s41599-023-02415-1

Brooke, J. (1996). SUS: A "Quick and Dirty" Usability Scale. In *Usability Evaluation In Industry* (pp. 207–212). CRC Press. https://doi.org/10.1201/9781498710411-35

Camacho, L., & Bacao, F. (2024). WSMOTER: a novel approach for imbalanced regression. *Applied Intelligence*, *54*(19), 8789–8799. https://doi.org/10.1007/s10489-024-05608-6

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine Learning and Deep Learning Models for Demand Forecasting in Supply Chain Management: A Critical Review. *Applied System Innovation*, *7*(5), 93. https://doi.org/10.3390/asi7050093

Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S. S. K., Dwivedi, S., & Raykar, V. (2020). Attention based Multi-Modal New Product Sales Time-series Forecasting. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3110–3118. https://doi.org/10.1145/3394486.3403362

Falatouri, T., Darbanian, F., Brandtner, P., & Udokwu, C. (2022). Predictive Analytics for Demand Forecasting - A Comparison of SARIMA and LSTM in Retail SCM. *Procedia Computer Science*, *200*, 993–1003. https://doi.org/10.1016/j.procs.2022.01.298

Folinas, D., & Rabi, S. (2012). Estimating benefits of Demand Sensing for consumer goods organisations. *Journal of Database Marketing & Customer Strategy Management*, *19*(4), 245–261. https://doi.org/10.1057/dbm.2012.22

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Guha, A., Grewal, D., Kopalle, P. K., Haenlein, M., Schneider, M. J., Jung, H., Moustafa, R., Hegde, D. R., & Hawkins, G. (2021). How artificial intelligence will affect the future of retailing. *Journal of Retailing*, *97*(1), 28–41. https://doi.org/10.1016/j.jretai.2021.01.005

Guo, X., Lichtendahl, K. C., & Grushka-Cockayne, Y. (2025). Bayesian Ensembles of Exponentially Smoothed Life-Cycle Forecasts. *Manufacturing & Service Operations Management*, *27*(1), 230–248. https://doi.org/10.1287/msom.2022.0359

Heidenreich, S., Killmer, J. F., & Millemann, J. A. (2022). If at first you don't adopt - Investigating determinants of new product leapfrogging behavior. *Technological Forecasting and Social Change*, *176*. https://doi.org/10.1016/j.techfore.2021.121437

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, *37*(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Hossain, M. A., Akter, S., Yanamandram, V., & Wamba, S. F. (2023). Data-driven market effectiveness: The role of a sustained customer analytics capability in business operations. *Technological Forecasting and Social Change*, *194*. https://doi.org/10.1016/j.techfore.2023.122745

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles & Practice*.

Ileri, K. (2025). Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks. *International Journal of Machine Learning and Cybernetics*. https://doi.org/10.1007/s13042-025-02654-5

Kameswari, J., Ramesh, P., Bhavikatti, V., Omnamasivaya, B., Chaitanya, G., Bastray, T., Hiremath, S., & Gondesi, G. S. (2024). Analyzing the role of big data and its effects on the retail industry. *Web Intelligence*, *22*(1), 45–63. https://doi.org/10.3233/WEB-230027

Karb, T., Kühl, N., Hirt, R., & Glivici-Cotru, V. (2020). A Network-based Transfer Learning Approach to Improve Sales Forecasting of New Products. *Proceedings of the 28th European Conference on Information Systems (ECIS)*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. https://github.com/Microsoft/LightGBM.

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*.

Kizgin, K. T., Alp, S., Aydin, N., & Yu, H. (2025). Machine learning-based sales forecasting during crises: Evidence from a Turkish women's clothing retailer. *Science Progress*, *108*(1). https://doi.org/10.1177/00368504241307719

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lei, D., Qi, Y., Liu, S., Geng, D., Zhang, J., Hu, H., & Shen, Z.-J. M. (2023). Pooling and Boosting for Demand Prediction in Retail: A Transfer Learning Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4490516

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099–2102. https://doi.org/10.1145/2470654.2481287

Lim, B., Ark, S. "O., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Loureiro, A. L. D., Miguéis, V. L., & da Silva, L. F. M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81–93. https://doi.org/10.1016/j.dss.2018.08.010

Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 104–111. https://doi.org/10.1109/CIDM.2011.5949434

Makridakis, S., Petropoulos, F., & Spiliotis, E. (2022). The M5 competition: Conclusions. *International Journal of Forecasting*, *38*(4), 1576–1582. https://doi.org/10.1016/j.ijforecast.2022.04.006

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices. *Forecasting*, *3*(3), 644–662. https://doi.org/10.3390/forecast3030040

Montaño Moreno, J., Palmer Pol, A., Sesé Abad, A., & Cajal Blasco, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*, *4*(25), 500–506. https://doi.org/10.7334/psicothema2013.23

Nanda, A., Xu, Y., & Zhang, F. (2021). How would the COVID-19 pandemic reshape retail real estate and high streets through acceleration of E-commerce and digitalization? *Journal of Urban Management*, *10*(2), 110–124. https://doi.org/10.1016/j.jum.2021.04.001

O'Higgins, B., & Fatorachian, H. (2025). Consumer trust in artificial intelligence in the UK and Ireland's personal care and cosmetics sector. *Cogent Business & Management*, *12*(1). https://doi.org/10.1080/23311975.2025.2469765

Oliveira, D., & Ramos, P. (2024). Retail demand forecasting with deep learning: Comparing LSTM and Transformer models. *Journal of Business Analytics*, *6*(1), 49–64. https://doi.org/10.1080/2573234X.2024.1234567

Oliveira, J. M., & Ramos, P. (2024). Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. *Mathematics*, *12*(17), 2728. https://doi.org/10.3390/math12172728

Punia, S., Kaur, A., & Aggarwal, A. (2020). Sales forecasting for retail chains using deep learning models. *International Journal of Data Science and Analytics*, *10*(3), 215–228. https://doi.org/10.1007/s41060-020-00221-1

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sapountzi, A., & Psannis, K. E. (2018). Social networking data analysis tools & challenges. *Future Generation Computer Systems*, *86*, 893–913. https://doi.org/10.1016/j.future.2016.10.019

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Shankar, V. (2018). How Artificial Intelligence (AI) is Reshaping Retailing. *Journal of Retailing*, *94*(4), vi–xi. https://doi.org/10.1016/S0022-4359(18)30076-9

Singh, R., Telukdarie, A., & Mongwe, R. (2024). Digital Social Media Influencers' Impact on Beauty and Personal Care Purchases in South Africa. *Platforms*, *2*(4), 193–210. https://doi.org/10.3390/platforms2040013

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian Optimization of Machine Learning Algorithms*.

Sousa, M. S., Loureiro, A. L. D., & Miguéis, V. L. (2025). Predicting demand for new products in fashion retailing using censored data. *Expert Systems with Applications*, *259*. https://doi.org/10.1016/j.eswa.2024.125313

Souza, R. F., Wanke, P., & Correa, H. (2020). Demand forecasting in the beauty industry using fuzzy inference systems. *Journal of Modelling in Management*, *15*(4), 1389–1417. https://doi.org/10.1108/JM2-03-2019-0050

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In *Journal of Machine Learning Research* (Vol. 15).

Sun, Y., & Li, T. (2024). A transformer-based framework for enterprise sales forecasting. *PeerJ Computer Science*, *10*, e2503. https://doi.org/10.7717/peerj-cs.2503

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Theodoridis, G., & Tsadiras, A. (2024). Retail Demand Forecasting: A Multivariate Approach and Comparison of Boosting and Deep Learning Methods. *International Journal on Artificial Intelligence Tools*, *33*(04). https://doi.org/10.1142/S0218213024500015

Thivakaran, T. K., & Ramesh, M. (2022). Exploratory Data analysis and sales forecasting of bigmart dataset using supervised and ANN algorithms. *Measurement: Sensors*, *23*. https://doi.org/10.1016/j.measen.2022.100388

Tseng, H. T., Aghaali, N., & Hajli, D. N. (2022). Customer agility and big data analytics in new product context. *Technological Forecasting and Social Change*, *180*. https://doi.org/10.1016/j.techfore.2022.121690

Van Steenbergen, R. M., & Mes, M. R. K. (2020). Forecasting demand profiles of new products. *Decision Support Systems*, *139*. https://doi.org/10.1016/j.dss.2020.113401

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Vorhies, D. W., & Morgan, N. A. (2005). Benchmarking marketing capabilities for sustainable competitive advantage. In *Journal of Marketing* (Vol. 69, Issue 1, pp. 80–94). https://doi.org/10.1509/jmkg.69.1.80.55505

Wilson, G. T. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, *37*(5), 709–711. https://doi.org/10.1111/jtsa.12194

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *ArXiv Preprint ArXiv:2106.13008*. https://doi.org/10.48550/arXiv.2106.13008

Ying, S., Sindakis, S., Aggarwal, S., Chen, C., & Su, J. (2021). Managing big data in the retail industry of Singapore: Examining the impact on customer satisfaction and organizational performance. *European Management Journal*, *39*(3), 390–400. https://doi.org/10.1016/j.emj.2020.04.001

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

Zikopoulos, P. C., Eaton, C., Deutsch, T., Lapis, G., & deRoos, D. (2012). *Understanding Big Data*. http://it.toolbox.com/blogs/db2luw.

Zineb, N., Rachid, B., & Fatine, E. (2025). *Enhancing Cosmetic Supply Chain Efficiency Through Demand Forecasting Using Machine Learning* (pp. 212–231). https://doi.org/10.1007/978-3-031-75923-9_13

ANEXOS

Anexo 1 – Comparação modelo global vs. *Grid search*

Comparação modelo global vs. *Grid Search* - *LightGBM*

LightGBM	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	g. search	Global	g. search	Global	g. search	Global	g. search	Global	g. search		
F000001	7.73	7.66	3352.72	3363.36	57.90	57.99	1.17	1.09	0.38	0.38	3086	9.8%
F000002	7.41	7.33	686.64	660.02	26.20	25.69	0.98	0.94	0.65	0.67	731	2.3%
F000003	4.19	4.20	138.38	139.41	11.76	11.81	0.92	0.89	0.64	0.64	2556	8.1%
F000005	3.79	3.62	26.85	28.11	5.18	5.30	1.46	1.18	0.98	0.97	14	0.0%
F000006	4.25	4.33	84.15	87.09	9.17	9.33	0.91	0.92	0.92	0.92	2930	9.3%
F000007	11.34	11.41	1062.04	1116.54	32.59	33.41	1.25	1.09	0.22	0.18	34	0.1%
F000008	2.68	2.92	9.57	11.74	3.09	3.43	2.05	2.18	-0.46	-0.79	37	0.1%
F000009	6.51	6.52	212.09	212.72	14.56	14.58	0.78	0.77	0.92	0.92	8003	25.3%
F000010	2.15	1.93	6.57	5.60	2.56	2.37	1.06	0.92	0.72	0.76	574	1.8%
F000011	4.71	4.59	77.75	76.70	8.82	8.76	0.84	0.77	0.89	0.89	3592	11.4%
F000012	8.12	8.12	241.67	249.96	15.55	15.81	1.41	1.38	0.80	0.79	364	1.2%
F000013	2.96	2.95	29.45	29.99	5.43	5.48	1.14	1.11	0.92	0.92	9590	30.4%
F000014	2.30	1.82	6.50	4.14	2.55	2.04	1.51	1.19	-0.80	-0.15	64	0.2%
Média pesada	4.90	4.88	438.22	439.34	14.68	14.71	0.98	0.95	0.82	0.82		

Comparação modelo global vs. *Grid Search - XGBoost*

XGBoost	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	g. search	Global	g. search	Global	g. search	Global	g. search	Global	g. search		
F000001	6.96	6.96	3314.97	3489.49	57.58	59.07	0.89	0.90	0.38	0.35	3086	9.8%
F000002	6.93	7.00	779.59	709.18	27.92	26.63	0.56	0.67	0.61	0.64	731	2.3%
F000003	4.21	3.48	107.00	82.66	10.34	9.09	0.79	0.69	0.73	0.79	2556	8.1%
F000005	2.34	2.17	18.95	16.23	4.35	4.03	0.32	0.65	0.98	0.98	14	0.0%
F000006	3.57	3.68	51.78	67.22	7.20	8.20	0.49	0.63	0.95	0.93	2930	9.3%
F000007	9.42	9.34	700.48	830.82	26.47	28.82	0.82	0.80	0.48	0.39	34	0.1%
F000008	1.09	1.52	2.60	4.37	1.61	2.09	0.66	1.11	0.60	0.33	37	0.1%
F000009	5.97	5.85	220.09	208.10	14.84	14.43	0.41	0.54	0.91	0.92	8003	25.3%
F000010	1.31	1.50	3.28	3.66	1.81	1.91	0.44	0.64	0.86	0.84	574	1.8%
F000011	4.28	4.17	84.85	68.65	9.21	8.29	0.44	0.58	0.88	0.90	3592	11.4%
F000012	7.26	7.69	232.76	243.26	15.26	15.60	0.95	1.15	0.81	0.80	364	1.2%
F000013	2.34	2.55	29.43	31.22	5.43	5.59	0.60	0.86	0.92	0.91	9590	30.4%
F000014	1.00	1.46	1.67	2.68	1.29	1.64	0.48	0.88	0.54	0.26	64	0.2%
Média pesada	4.35	4.33	433.39	444.22	14.48	14.44	0.57	0.71	0.84	0.84		

Anexo 2 – Comparação modelo global vs. *LN Smote*

Comparação modelo global vs. *LN Smote* – *LightGBM*

LightGBM	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	LN Smote	Global	LN Smote	Global	LN Smote	Global	LN Smote	Global	LN Smote		
F000001	7.73	7.48	3352.72	3455.45	57.90	58.78	1.17	0.95	0.38	0.36	3086	9.8%
F000002	7.41	7.35	686.64	750.96	26.20	27.40	0.98	0.82	0.65	0.62	731	2.3%
F000003	4.19	3.91	138.38	122.14	11.76	11.05	0.92	0.73	0.64	0.69	2556	8.1%
F000005	3.79	2.21	26.85	19.01	5.18	4.36	1.46	0.73	0.98	0.98	14	0.0%
F000006	4.25	4.05	84.15	77.74	9.17	8.82	0.91	0.71	0.92	0.92	2930	9.3%
F000007	11.34	10.36	1062.04	891.57	32.59	29.86	1.25	0.85	0.22	0.34	34	0.1%
F000008	2.68	1.61	9.57	4.15	3.09	2.04	2.05	1.11	-0.46	0.37	37	0.1%
F000009	6.51	6.31	212.09	210.13	14.56	14.50	0.78	0.62	0.92	0.92	8003	25.3%
F000010	2.15	1.67	6.57	4.79	2.56	2.19	1.06	0.70	0.72	0.80	574	1.8%
F000011	4.71	4.49	77.75	66.32	8.82	8.14	0.84	0.69	0.89	0.91	3592	11.4%
F000012	8.12	8.03	241.67	250.38	15.55	15.82	1.41	1.19	0.80	0.79	364	1.2%
F000013	2.96	2.82	29.45	28.06	5.43	5.30	1.14	1.00	0.92	0.92	9590	30.4%
F000014	2.30	1.20	6.50	2.08	2.55	1.44	1.51	0.65	-0.80	0.42	64	0.2%
Média pesada	4.90	4.70	438.22	445.49	14.68	14.56	0.98	0.81	0.82	0.83		

Comparação modelo global vs. *LN Smote – XGBoost*

XGBoost	MAE		MSE		RMSE		MAPE		R2		Amostra (#)	%
Família	Global	LN Smote	Global	LN Smote	Global	LN Smote	Global	LN Smote	Global	LN Smote		
F000001	6.96	7.15	3314.97	3276.70	57.58	57.24	0.89	0.79	0.38	0.39	3086	9.8%
F000002	6.93	6.95	779.59	712.48	27.92	26.69	0.56	0.66	0.61	0.64	731	2.3%
F000003	4.21	3.42	107.00	104.36	10.34	10.22	0.79	0.55	0.73	0.73	2556	8.1%
F000005	2.34	3.60	18.95	39.99	4.35	6.32	0.32	0.33	0.98	0.96	14	0.0%
F000006	3.57	3.53	51.78	50.55	7.20	7.11	0.49	0.44	0.95	0.95	2930	9.3%
F000007	9.42	7.04	700.48	314.88	26.47	17.74	0.82	0.62	0.48	0.77	34	0.1%
F000008	1.09	1.26	2.60	2.81	1.61	1.68	0.66	0.69	0.60	0.57	37	0.1%
F000009	5.97	6.12	220.09	223.05	14.84	14.93	0.41	0.41	0.91	0.91	8003	25.3%
F000010	1.31	1.31	3.28	3.41	1.81	1.85	0.44	0.41	0.86	0.85	574	1.8%
F000011	4.28	4.27	84.85	83.17	9.21	9.12	0.44	0.47	0.88	0.88	3592	11.4%
F000012	7.26	7.06	232.76	212.03	15.26	14.56	0.95	0.86	0.81	0.82	364	1.2%
F000013	2.34	2.54	29.43	43.40	5.43	6.59	0.60	0.67	0.92	0.88	9590	30.4%
F000014	1.00	1.20	1.67	2.12	1.29	1.46	0.48	0.64	0.54	0.41	64	0.2%
Média pesada	4.35	4.39	433.39	431.93	14.48	14.75	0.57	0.56	0.84	0.83		