

**UNIVERSIDADE FEDERAL DO AMAZONAS – UFAM
INSTITUTO DE CIÊNCIAS EXATAS E TECNOLOGIA – ICET
CURSO DE ENGENHARIA DE PRODUÇÃO**

VITOR LOPES DE MATOS

**AVALIAÇÃO DO USO DE INTELIGÊNCIA COMPUTACIONAL PARA PREVISÃO
DE VENDAS**

ITACOATIARA

2025

VITOR LOPES DE MATOS

**AVALIAÇÃO DO USO DE INTELIGÊNCIA COMPUTACIONAL PARA PREVISÃO
DE VENDAS**

Trabalho de Conclusão de Curso apresentado ao
Curso de Engenharia de Produção da Universidade
Federal do Amazonas (UFAM), como requisito para
obtenção do título de Engenheiro de Produção.

Orientador: Profa. Dra. Iracyanne Retto Uhlmann

ITACOATIARA

2025

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

M433a Matos, Vitor Lopes de
Avaliação do uso de inteligência computacional para previsão de vendas / Vitor Lopes de Matos. - 2025.
79 f. : il., color. ; 31 cm.

Orientador(a): Iracyanne Retto Uhlmann.
Trabalho de Conclusão de Curso (graduação) - Universidade Federal do Amazonas, Instituto de Ciências Exatas e Tecnologia de Itacoatiara, Curso de Engenharia de Produção, Itacoatiara, 2025.

1. Previsão de vendas. 2. inteligência computacional. 3. Aprendizado de máquina. 4. Séries temporais. 5. Varejo. I. Uhlmann, Iracyanne Retto. II. Universidade Federal do Amazonas. Instituto de Ciências Exatas e Tecnologia de Itacoatiara. Curso de Engenharia de Produção. III. Título

VITOR LOPES DE MATOS

**AVALIAÇÃO DO USO DE INTELIGÊNCIA COMPUTACIONAL PARA PREVISÃO
DE VENDAS**

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Produção da Universidade Federal do Amazonas (UFAM) como requisito para obtenção do grau de Engenheiro de Produção.

Este trabalho foi defendido e aprovado pela banca em 10/12/2025.

BANCA EXAMINADORA

Prof.^a Dr.^a Iracyanne Retto Uhlmann - UFAM
Orientadora

Prof. Dr. Carlos Alberto Oliveira de Freitas - UFAM
Avaliador

Prof. Dr. Hidelbrando Ferreira Rodrigues - UFAM
Avaliador

Dedico este trabalho à minha mãe, Ângela, e ao meu pai, Jair, que foram minha base, minha força e meu porto seguro em toda esta caminhada. Cada conquista desta jornada carrega o apoio, o amor e os valores que vocês me ensinaram.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, por me fortalecer e iluminar em cada etapa desta caminhada, dando-me sabedoria e perseverança para continuar, mesmo nos momentos mais difíceis.

Agradeço à minha mãe Ângela, e ao meu pai Jair, por todo o apoio, amor e dedicação. Eles foram essenciais nesta jornada e sempre acreditaram no meu potencial. À minha irmã Karina e ao meu irmão Kaique, que estiveram presentes com incentivo e companheirismo ao longo de todo o processo.

Aos meus amigos, que estiveram comigo nos momentos de risadas, conversas, diversões e, também, nos momentos de dificuldade. A ajuda e a presença de vocês tornaram essa trajetória mais leve e marcante, e sou profundamente grato por cada incentivo e por cada momento compartilhado.

À minha orientadora, professora Iracyanne, pela orientação paciente, pela confiança depositada e por todo o aprendizado proporcionado durante o desenvolvimento desta pesquisa e nas disciplinas ministradas. Ao professor Joel, pelo incentivo que ultrapassou a sala de aula e contribuiu também para meu crescimento pessoal, e a todos os professores que eu tive a honra de estudar, deixo os meus sinceros agradecimentos.

A todos vocês, registro minha profunda gratidão.

“Consagre ao Senhor tudo que você faz, e os seus planos serão bem-sucedidos.”

Provérbios 16:3.

RESUMO

A crescente digitalização dos processos no varejo ampliou a disponibilidade de dados operacionais e reforçou a necessidade de previsões de vendas mais precisas para apoiar decisões estratégicas e otimizar planejamento, estoques e desempenho financeiro. Esse cenário evidencia a oportunidade de aplicar técnicas de inteligência computacional capazes de lidar com padrões complexos, sazonalidade e variáveis exógenas, superando limitações de métodos estatísticos tradicionais. O objetivo deste estudo foi avaliar modelos de inteligência computacional aplicados à previsão de vendas, comparando-os a um modelo estatístico de referência. A metodologia envolveu coleta de dados reais de vendas e fluxo de clientes via SAP e API Digifort, pré-processamento, imputação de zeros artificiais por GLM de Poisson, criação de variáveis sazonais e contextuais, construção de diferentes versões de bases de treino e aplicação dos modelos Floresta Aleatória, XGBoost e SARIMAX, avaliados por *holdout* temporal 80/20 e validação cruzada *walk-forward*. Os resultados mostraram que os modelos de inteligência computacional apresentaram menor erro preditivo e maior estabilidade temporal, com destaque para a Floresta Aleatória, enquanto o SARIMAX teve desempenho inferior diante da maior complexidade das variáveis explicativas. Conclui-se que as técnicas de inteligência computacional se mostram mais eficazes e robustas para previsão de vendas no varejo físico, oferecendo suporte valioso à tomada de decisão gerencial.

Palavras-chave: previsão de vendas; inteligência computacional; aprendizado de máquina; séries temporais; varejo.

ABSTRACT

The growing digitalization of retail processes has expanded the availability of operational data and reinforced the need for more accurate sales forecasting to support strategic decisions and optimize planning, inventory management, and financial performance. This scenario highlights the opportunity to apply computational intelligence techniques capable of handling complex patterns, seasonality, and exogenous variables, overcoming the limitations of traditional statistical methods. This study aimed to evaluate computational intelligence models applied to sales forecasting and compare them with a statistical reference model. The methodology involved collecting real sales and customer flow data via SAP and the Digifort API, performing preprocessing procedures, imputing artificial zeros using Poisson GLM, creating seasonal and contextual variables, building different versions of training datasets, and applying the Random Forest, XGBoost, and SARIMAX models, assessed through an 80/20 temporal holdout and *walk-forward* cross-validation. The results showed that computational intelligence models presented lower predictive error and greater temporal stability, with Random Forest standing out, while SARIMAX performed worse when faced with higher variable complexity. The study concludes that computational intelligence techniques are more effective and robust for sales forecasting in physical retail, offering valuable support for managerial decision-making.

Keywords: sales forecasting; computational intelligence; machine learning; time series; retail.

LISTA DE ILUSTRAÇÕES

Figura 1 - Procedimento Metodológico.....	27
Figura 2 - Fluxograma da metodologia aplicada para substituição dos valores zerados	37
Figura 3 - Comparação entre a série original e a série ajustada, com destaque para os valores imputados em substituição aos zeros artificiais	38
Figura 4 - Dispersão dos dados originais	40
Figura 5 - Dispersão dos dados tratados	40
Figura 6 - Boxplot das vendas por dia da semana	42
Figura 7 - Boxplot da contagem de clientes por dia da semana.....	43
Figura 8 - Boxplot das vendas por mês.....	44
Figura 9 - Boxplot da contagem de clientes por mês	44
Figura 10 - Resíduos de vendas com marcação de <i>outliers</i> e janelas de campanha	45
Figura 11 - Comportamento das vendas reais e previstas pela Floresta Aleatória na base extra-expandida	54
Figura 12 - Relação entre valores reais e previstos pelo modelo XGBoost base extra- expandida.....	54
Figura 13 - Relação entre valores reais e previstos no teste pelo modelo SARIMAX na base extra-expandida	59
Figura 14 - Esquema da validação cruzada temporal <i>walk-forward</i>	60
Figura 15 - RMSE dos modelos antes e depois da validação cruzada	65
Figura 16 - R ² dos modelos antes e depois da validação cruzada	66

LISTA DE QUADROS

Quadro 1 - Enquadramento metodológico	30
Quadro 2 - Resumo das fontes e períodos de cobertura	32
Quadro 3 - Estrutura final da base integrada.....	34
Quadro 4 - Quantis por dia da semana q10 e q99	36
Quadro 5 - Estatísticas descritivas antes e depois do tratamento da variável de contagem de clientes e para variável de vendas.....	39
Quadro 6 - Comparação entre as versões da base de treino.....	42
Quadro 7 - Cenário A - métricas por base e modelo	46
Quadro 8 - Cenário B - métricas por base e modelo	46
Quadro 9 - <i>Holdout</i> temporal 80/20 para o Floresta Aleatória	49
Quadro 10 - <i>Holdout</i> temporal 80/20 para o XGBoost.....	49
Quadro 11 - <i>Holdout</i> temporal 80/20 para o SARIMAX	50
Quadro 12 - Comparação com a linha de base <i>lag-7</i>	50
Quadro 13- Hiperparâmetros utilizados.....	51
Quadro 14 - XGBoost na base extra-expandida: antes e depois da regularização ...	52
Quadro 15 - Comparação entre XGBoost regularizado e Floresta Aleatória.....	55
Quadro 16 - Hiperparâmetros escolhidos por base (validação temporal nos 10% finais do treino).....	56
Quadro 17 - Desempenho no período de avaliação 80/20 com RMSE e ganho em relação ao <i>lag-7</i>	57
Quadro 18 - Configurações do SARIMAX por base	58
Quadro 19 - Desempenho do SARIMAX por base (80/20).....	58
Quadro 20 - Resultados da validação cruzada temporal sem definição de janelas ..	61
Quadro 21 - Resultados médios da validação cruzada temporal com janelas de 15 dias.....	62
Quadro 22 - Valores de MAE nos testes 80/20 e na validação cruzada com janelas de 15 dias.....	67

LISTA DE ABREVIATURAS E SIGLAS

ABEPRO	Associação Brasileira de Engenharia de Produção
API	<i>Application Programming Interface</i> (Interface de Programação de Aplicações)
ARIMA	<i>AutoRegressive Integrated Moving Average</i> (Autorregressivo Integrado de Médias Móveis)
CSV	<i>Comma-Separated Values</i> (Valores Separados por Vírgula)
ENR	<i>Elastic Net Regression</i> (Regressão com Rede Elástica)
GLM	<i>Generalized Linear Model</i> (Modelo Linear Generalizado)
KNN	<i>K-Nearest Neighbors</i> (K-vizinhos mais próximos)
LSTM	<i>Long Short-Term Memory</i> (Memória de Longo e Curto Prazo)
MAE	<i>Mean Absolute Error</i> (Erro Absoluto Médio)
MAPE	<i>Mean Absolute Percentage Error</i> (Erro Percentual Absoluto Médio)
ML	<i>Machine Learning</i> (Aprendizado de Máquina)
R ²	Coefficiente de Determinação
RMSE	<i>Root Mean Squared Error</i> (Raiz do Erro Quadrático Médio)
SAITS	<i>Self-Attention-based Imputation for Time Series</i> (Imputação de Séries Temporais Baseada em Autoatenção)
SAP	<i>Systems Applications and Products in Data Processing</i> (Sistemas, Aplicações e Produtos em Processamento de Dados)
SARIMAX	<i>Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors</i> (Modelo Autorregressivo Integrado de Médias Móveis Sazonal com Regressão Exógena)
SVR	<i>Support Vector Regression</i> (Regressor por Vetores de Suporte)
XGBoost	<i>Extreme Gradient Boosting</i> (Impulsioneamento de Gradiente Extremo)

SUMÁRIO

1	INTRODUÇÃO	14
1.1	SITUAÇÃO PROBLEMA.....	15
1.2	OBJETIVOS.....	17
1.2.1	Objetivo geral	17
1.2.2	Objetivos específicos	17
1.3	JUSTIFICATIVA.....	17
2	REVISÃO DE LITERATURA	19
2.1	DESAFIO DO VAREJO E O PAPEL DA PREVISÃO DE VENDAS	19
2.2	ESTUDOS SOBRE USO DE INTELIGÊNCIA COMPUTACIONAL APLICADA À PREVISÃO DE DEMANDA.....	20
2.3	MODELOS E TÉCNICAS DE PREVISÃO ESTUDADOS NESTA PESQUISA.....	21
2.3.1	Modelo preditivo de previsão Floresta Aleatória	22
2.3.2	Modelo preditivo de previsão Impulsioneamento de Gradiente Extremo (XGBoost)	23
2.3.3	Modelo preditivo de previsão ARIMA	23
2.3.4	Técnica de validação cruzada para avaliação estatística	24
2.4	ABORDAGEM UTILIZADA PARA TRATAMENTO DE DADOS DE CONTAGEM ZERADOS	24
2.5	VARIÁVEIS EXÓGENAS, EFEITOS DE CALENDÁRIO, PROMOÇÕES E <i>OUTLIERS</i> EM SÉRIES TEMPORAIS	25
3	METODOLOGIA	27
3.1	PROCEDIMENTO METODOLÓGICO	27
3.1.1	Fase 1: Definição do problema de pesquisa	27
3.1.2	Fase 2: Aplicação dos modelos preditivos e técnica estatística	28
3.1.3	Fase 3: Análise dos modelos	30
3.2	ENQUADRAMENTO METODOLÓGICO	30
4	APLICAÇÃO DOS MODELOS E TÉCNICA ESTATÍSTICA	31
4.1	OBJETO DE ESTUDO DESTA PESQUISA.....	31
4.2	COLETA DOS DADOS REAIS DA LOJA.....	31
4.3	PRÉ-PROCESSAMENTO DOS DADOS	33

4.4	ENGENHARIA DE ATRIBUTOS.....	35
4.4.1	Criação de variáveis sazonais e contextuais	35
4.4.2	Tratamento dos zeros sequenciais	35
4.4.3	Preparação da base para aplicação dos modelos	41
4.4.4	Análise de <i>outliers</i>	45
4.5	APLICAÇÃO DOS MODELOS ARIMA, FLORESTA ALEATÓRIA E XGBOOST	47
4.5.1	Aplicação inicial e diagnóstico de sobreajuste.....	47
4.5.2	Regularização do modelo XGBoost na base extra-expandida.....	51
4.5.3	Aplicação do XGBoost regularizado e ajuste simétrico da Floresta Aleatória.....	54
4.5.4	Otimização do SARIMAX como referência comparativa	57
4.6	APLICAÇÃO DA TÉCNICA DE VALIDAÇÃO CRUZADA NOS MODELOS	59
5	ANÁLISE DOS MODELOS	64
5.1	AVALIAÇÃO DOS MODELOS	64
5.2	COMPARAÇÃO DOS MODELOS APLICADOS	67
5.3	DISCUSSÃO DOS RESULTADOS	69
6	CONSIDERAÇÕES FINAIS.....	72
	REFERÊNCIAS.....	74

1 INTRODUÇÃO

Nos últimos anos, a crescente digitalização dos processos empresariais ampliou significativamente a disponibilidade de dados operacionais e comerciais (Kádárová; Lachvajderová; Sukopová, 2023). Esse cenário abriu espaço para soluções de previsão cada vez mais precisas, capazes de auxiliar gestores na tomada de decisões estratégicas, especialmente em ambientes competitivos onde a oscilação da demanda pode impactar diretamente o planejamento de compras, a gestão de estoques e o desempenho financeiro (Makridakis; Hyndman; Petropoulos, 2020).

Sob essa perspectiva, a inteligência computacional surge como um conjunto de técnicas robustas e flexíveis que permite transformar dados em previsões mais confiáveis, mesmo quando a série apresenta comportamentos complexos, sazonalidade ou efeitos externos (Lim; Zohren, 2021). Segundo Xu *et al.* (2025), essa abordagem reúne técnicas como redes neurais e algoritmos evolutivos, que se destacam pela capacidade de aprender a partir de informações passadas e adaptar esse aprendizado a novos cenários, oferecendo soluções eficazes para problemas não lineares.

Modelos baseados em métodos estatísticos e de aprendizado de máquina têm ganhado espaço justamente pela capacidade de lidar com diferentes tipos de padrões e relacionamentos entre variáveis (Montero-Manso; Hyndman, 2021). De acordo com Punia e Shankar (2022), ao utilizar algoritmos que aprendem com o comportamento passado das vendas, torna-se possível antecipar tendências, identificar períodos de maior ou menor movimento e apoiar decisões que reduzam incertezas no processo produtivo e comercial. Além disso, a combinação entre métricas de desempenho, validação temporal e análise comparativa entre diferentes abordagens contribui para garantir previsões não apenas precisas, mas também estáveis e aplicáveis ao dia a dia da empresa.

Diante desse cenário, esta pesquisa foi intitulada “Avaliação do Uso de Inteligência Computacional para Previsão de Vendas”, alinhando-se ao que a Associação Brasileira de Engenharia de Produção (ABEPRO, 2025) identifica como parte da área de Pesquisa Operacional, especificamente na subárea de Análise de Demanda e Inteligência Computacional.

Nesse contexto, este estudo aplicou técnicas de inteligência computacional para prever vendas diárias em um contexto real, considerando diferentes conjuntos de variáveis e métodos de modelagem. A pesquisa buscou avaliar o desempenho de modelos distintos, analisar sua capacidade de generalização e identificar qual abordagem se mostra mais adequada para apoiar o planejamento operacional. Com esses experimentos realizados, pretende-se contribuir com uma solução prática e acessível, capaz de auxiliar empresas no uso de seus dados e na construção de previsões mais consistentes para suporte à tomada de decisão.

Além desta introdução, este trabalho está organizado da seguinte forma: na seção 2 é apresentado a revisão de literatura que sustenta o estudo, a seção 3 descreve a metodologia adotada na condução da pesquisa, na seção 4, é detalhado o processo de aplicação dos modelos, a seção 5 reúne a análise dos modelos, na seção 6 é evidenciado as considerações finais. Por fim, são apresentadas as referências.

1.1 SITUAÇÃO PROBLEMA

Gerenciar vendas de forma eficiente em lojas de varejo é um desafio constante, especialmente quando se trata de prever a demanda com precisão, que por sua vez é fator essencial para manter o equilíbrio operacional e financeiro (Muthukalyani, 2023). Conforme apontado por Oyewole *et al.* (2024), quando as previsões falham, toda a cadeia de suprimentos pode ser afetada, gerando rupturas de estoque, desperdício de recursos e impactos negativos no atendimento ao cliente.

Em muitos estabelecimentos, decisões estratégicas como vendas, reposição de produtos, escalas de funcionários e ações promocionais ainda são tomadas com base em dados históricos simples ou percepções empíricas, o que compromete a assertividade do planejamento (Noseda, 2021). A previsão de vendas desempenha um papel fundamental no equilíbrio entre oferta e demanda, ajudando a evitar tanto a escassez quanto o excesso de mercadorias. Tissot, Vidor e Chiwiacowsky (2022), ressaltam que o uso de métodos estatísticos, como a suavização exponencial e o modelo AutoRegressivo Integrado de Médias Móveis, do inglês *AutoRegressive Integrated Moving Average* (ARIMA), proporcionam estimativas mais precisas, especialmente quando complementadas por dados relevantes ao contexto analisado.

No contexto do varejo físico, informações como o fluxo de clientes vêm ganhando cada vez mais espaço como base para decisões estratégicas. Esses dados, quando bem utilizados, ajudam a prever indicadores-chave de desempenho, como vendas, conversão e movimentação na loja. Panay *et al.* (2021) afirma que métricas como o número de visitantes, a taxa de conversão e o volume total de vendas são fundamentais para que gestores possam antecipar comportamentos e se planejar melhor, seja para reforçar o estoque, organizar a equipe ou ajustar o fluxo de caixa. Desta forma, modelos preditivos que oferecem não só boas estimativas, mas também explicações sobre a relevância de cada variável, têm se mostrado aliados valiosos na tomada de decisões mais assertivas e flexíveis diante de cenários imprevisíveis.

Haque, Amin e Miah (2023) complementam essa visão ao mostrar que muitas empresas ainda não aproveitam todo o potencial das ferramentas de previsão, apesar do avanço das técnicas como o modelo ARIMA, o algoritmo dos K-vizinhos mais próximos, do inglês *K-Nearest Neighbors* (KNN), e o método de Florestas Aleatórias, do inglês *Random Forest*. Seu estudo indicou que modelos mais sofisticados, como o Florestas Aleatórias, tendem a oferecer melhor desempenho preditivo em séries temporais de vendas agregadas, reforçando a importância de escolher a técnica mais adequada ao tipo de dado e ao objetivo da previsão.

Diante desse cenário, o escopo desse estudo é a avaliação do uso de inteligência computacional para previsão de vendas em uma loja física de varejo, utilizando dados de fluxo de clientes, vendas e dados temporais. A pesquisa foi desenvolvida com base em técnicas de previsão consolidadas na literatura e aplicadas sobre um conjunto real de dados fornecido por uma loja.

Dessa forma, o estudo busca responder: como diferentes modelos de inteligência computacional, comparados a um modelo estatístico clássico, se comportam na previsão de vendas em uma loja de varejo físico, a partir de dados de fluxo de clientes, data e quantidade de vendas?

1.2 OBJETIVOS

1.2.1 Objetivo geral

Avaliar modelos de inteligência computacional aplicados à previsão de vendas, comparando-os a um modelo estatístico de referência, utilizando dados históricos de fluxo de clientes, vendas e informações temporais, apoiando a tomada de decisões gerenciais.

1.2.2 Objetivos específicos

- Identificar na literatura estudos relacionados a técnicas e modelos de inteligência computacional aplicados a previsão de vendas;
- Aplicar modelos e técnicas de previsão de vendas utilizando dados de fluxo de clientes, vendas e informações temporais, testando sua eficácia no contexto do varejo;
- Analisar o desempenho dos modelos utilizados, considerando métricas de acurácia e sua utilidade para a tomada de decisões gerenciais.

1.3 JUSTIFICATIVA

No varejo atual, altamente competitivo e dinâmico, saber prever com precisão a demanda por produtos tem se mostrado fundamental para que as empresas tomem decisões mais assertivas. Quando as vendas podem ser antecipadas com maior clareza, o controle de estoque, a distribuição de recursos e até o atendimento ao cliente tendem a ser mais eficientes, refletindo positivamente nos resultados do negócio (Jiang; Ruan; Sunj, 2021).

Nos últimos anos, pesquisas vêm destacando como a análise de dados e o uso de algoritmos de aprendizado de máquina, do inglês *Machine Learning* (ML), vêm ganhando espaço no desenvolvimento de modelos preditivos mais eficazes e confiáveis no varejo (Learning; Nithinraj; Jaisachin, 2024). Ganguly e Mukherjee

(2024) observam que, ao combinar algoritmos como Floresta Aleatória, KNN e o Regressor por Votação, do inglês *Voting Regressor*, é possível alcançar previsões mais precisas, reduzindo erros e melhorando o desempenho dos modelos, especialmente quando se leva em conta fatores como a sazonalidade, o histórico de vendas e o comportamento de compra dos consumidores.

Nesse mesmo sentido, Mishra e Sinha (2025) mostram que técnicas de análise de dados aplicadas a lojas físicas ajudam a entender melhor o comportamento de compra, permitindo segmentar produtos com mais precisão e prever a demanda com base na realidade local. Os autores destacam, ainda, que o uso de algoritmos como Regressor por Vetores de Suporte, do inglês *Support Vector Regression (SVR)*, Floresta Aleatória e Impulsionamento de Gradiente Extremo, do inglês *Extreme Gradient Boosting (XGBoost)* tem gerado bons resultados na previsão de vendas de produtos com alta rotatividade, especialmente ao se analisar indicadores como Erro Absoluto Médio, do inglês *Mean Absolute Error (MAE)* e Raiz do Erro Quadrático Médio, do inglês *Root Mean Squared Error (RMSE)*.

Tony *et al.* (2021) defende que os modelos baseados em ML representam um avanço em relação às abordagens estatísticas mais tradicionais, sobretudo quando se lida com grandes volumes de dados e mudanças frequentes no padrão de consumo. Técnicas como Floresta Aleatória, segundo os autores, vêm se mostrando mais precisas do que métodos clássicos como a Regressão Linear, do inglês *Linear Regression*, o que reforça sua aplicação no varejo físico.

Além do interesse do pesquisador na área de ciência de dados, com tantas opções de técnicas para previsão de demanda, este estudo surge da crescente necessidade de melhorar a precisão dos modelos preditivos no varejo físico. Modelos mais assertivos podem transformar a maneira como as empresas antecipam suas vendas, facilitando um planejamento mais eficiente e a alocação de recursos de maneira mais estratégica.

A pesquisa também traz uma contribuição teórica significativa ao investigar a eficácia desses modelos no contexto do varejo físico, um campo que ainda é pouco explorado em comparação ao varejo digital. Ao comparar e testar essas abordagens em um cenário prático, o estudo oferece contribuições relevantes tanto para profissionais que buscam aprimorar suas estratégias de vendas, quanto para pesquisadores que desejam aprofundar seu entendimento sobre como essas técnicas preditivas podem ser aplicadas no setor.

2 REVISÃO DE LITERATURA

Nesta seção, são abordados os estudos mais atuais para entendimento das técnicas, ferramentas e aplicações feitas nesta pesquisa.

2.1 DESAFIO DO VAREJO E O PAPEL DA PREVISÃO DE VENDAS

O varejo tem passado por mudanças cada vez mais rápidas, impulsionado por consumidores com hábitos voláteis, sazonalidades difíceis de prever e um cenário econômico global instável. De acordo com Ahn *et al.* (2024), a cadeia de suprimentos tem se tornado mais complexa, exigindo uma gestão mais ágil e sustentável para lidar com as variações de demanda e evitar desperdícios. Nesse contexto, antecipar o comportamento do consumidor deixou de ser apenas um diferencial competitivo e passou a ser uma necessidade para manter a operação eficiente e competitiva.

De maneira tradicional, as previsões de vendas são feitas com base em análises simples, geralmente apoiadas em médias históricas ou na experiência dos gestores. No entanto, como apontam Jahin *et al.* (2024), esse tipo de abordagem não consegue acompanhar as rápidas mudanças provocadas por fatores externos, promoções pontuais ou novas tendências de mercado. As consequências disso são previsões imprecisas, que podem gerar tanto excesso quanto falta de produtos, comprometendo o nível de serviço e elevando os custos operacionais.

Outro ponto importante é que o desafio não está apenas em prever o volume total de vendas, mas em entender como essa demanda se distribui entre diferentes lojas ou unidades de negócio. Bi *et al.* (2022) destacam que, em muitos casos, a ausência de previsões localizadas atrapalha o planejamento logístico e financeiro, gerando desorganização nas operações do dia a dia.

Diante dessa realidade, o uso de modelos preditivos mais avançados que combinam técnicas de ML e análise de séries temporais tem se mostrado uma alternativa eficaz para reduzir incertezas e apoiar o planejamento das empresas. Modelos como Floresta Aleatória, Regressor por Votação, ARIMA e redes neurais com Memória de Longo e Curto Prazo, do inglês *Long Short-Term Memory* (LSTM) vêm

demonstrando bons resultados em estudos recentes, especialmente por sua capacidade de lidar com dados complexos e variáveis (Deng; De Oliveira, 2024).

2.2 ESTUDOS SOBRE USO DE INTELIGÊNCIA COMPUTACIONAL APLICADA À PREVISÃO DE DEMANDA

Diversos estudos acadêmicos têm investigado como modelos preditivos, incluindo técnicas de inteligência computacional, podem contribuir para a melhoria das previsões de demanda no varejo e em outros setores.

A pesquisa de Yang, Chen e Zhou (2025), desenvolveu um modelo linear semifuncional com erros autorregressivos para previsão de vendas de energia elétrica. O destaque deste estudo é a incorporação da autocorrelação dos dados, fator muitas vezes ignorado em modelos tradicionais. Ao reconhecer essa característica, os autores conseguiram obter resultados de previsão mais robustos, demonstrando que levar em consideração dependências temporais pode fazer diferença significativa.

Outro exemplo é a pesquisa de Burinskiene (2022), que estudou a previsão de vendas em redes farmacêuticas utilizando técnicas de suavização exponencial e modelos de média móvel. O estudo mostrou que ao integrar a detecção de valores atípicos, do inglês *outliers*, e trabalhar em níveis de lojas específicas, foi possível reduzir de forma significativa os erros de previsão e melhorar a disponibilidade de produtos, o que é essencial para o varejo, onde imprecisões de estoque podem resultar em perdas consideráveis.

Pasupuleti *et al.* (2024) também apresentaram avanços significativos ao aplicar algoritmos de ML para otimizar a previsão de demanda em uma cadeia de suprimentos de grande porte. O uso de regressão, agrupamento, conhecido como *clustering* e séries temporais, permitiu não apenas prever vendas com mais precisão, mas também reduzir falhas no estoque e aumentar a eficiência logística, reforçando a aplicação prática dessas técnicas em ambientes reais.

Zubair *et al.* (2024) conduziram uma análise comparativa entre Regressão Linear, Floresta Aleatória e XGBoost, utilizando dados de vendas no varejo. O modelo Floresta Aleatória obteve o melhor desempenho, sendo identificado como o mais adequado para lidar com dados de alta variabilidade e complexidade, especialmente pela capacidade de capturar interações não lineares entre variáveis.

Mahin *et al.* (2025) exploraram a aplicação do Regressor por Votação, Floresta Aleatória, KNN, e Regressão com Rede Elástica, do inglês *Elastic Net Regression* (ENR). A pesquisa demonstrou que o Regressor por Votação e o Floresta Aleatória foram capazes de reduzir o RMSE, a níveis quase nulos, além de alcançarem um Coeficiente de Determinação (R^2) quase perfeito, mostrando-os serem extremamente eficiente em cenários de alta sazonalidade e tendências flutuantes.

Bianchessi (2023) analisou métodos de previsão de demanda em uma microcervejaria no Rio Grande do Sul, comparando o modelo ARIMA, as LSTM, e o XGBoost. A conclusão foi que o ARIMA, apesar de ser uma técnica mais clássica, ainda se mostra altamente competitivo para séries temporais de vendas bem estruturadas, atingindo um Erro Percentual Absoluto Médio, do inglês *Mean Absolute Percentage Error* (MAPE), de apenas 10,72%.

Entre todas essas pesquisas, a que mais se aproxima do objetivo deste estudo é a de Mahin *et al.* (2025), que comparou vários modelos preditivos para otimizar a previsão de vendas no varejo físico. Assim como na presente pesquisa, eles utilizaram variáveis como dados históricos, sazonalidade e características específicas de cada ponto de venda, o que torna sua abordagem alinhada ao objetivo de gerar previsões mais assertivas.

2.3 MODELOS E TÉCNICAS DE PREVISÃO ESTUDADOS NESTA PESQUISA

A literatura especializada em previsão de vendas no varejo físico destaca diversos modelos preditivos que vêm sendo aplicados com sucesso em diferentes contextos. Entre essas técnicas, sobressaem-se tanto modelos tradicionais de séries temporais quanto abordagens modernas baseadas em ML.

Entre os métodos clássicos, destaca-se o modelo ARIMA, muito utilizado em séries temporais com tendência e sazonalidade. Segundo Fatima e Rahimi (2024) o ARIMA é eficaz para capturar dependências temporais lineares e servir como referência de desempenho para técnicas mais complexas, devido à sua robustez estatística e interpretabilidade. Nesta pesquisa, o ARIMA foi empregado apenas para fins comparativos, permitindo avaliar se os modelos baseados em ML, que lidam melhor com múltiplas variáveis e relações não lineares, realmente superam um método estatístico consolidado.

Além do ARIMA, foram adotados os modelos de inteligência computacional Floresta Aleatória, XGBoost e a técnica de validação cruzada para avaliação estatística. Esses modelos são amplamente reconhecidos por sua eficiência no tratamento de dados tabulares e por sua capacidade de identificar padrões. Segundo Kamble *et al.* (2024), além de apresentarem alta precisão preditiva, esses modelos se destacam pela escalabilidade e pela facilidade de adaptação a diferentes contextos de vendas. Eles também oferecem recursos que permitem interpretar a importância de cada variável, o que contribui para análises mais assertivas.

Krishna, Farheen e Kalyani (2025) apontam que ambos os modelos têm alcançado bons resultados em métricas bastante utilizadas na literatura, como o MAE, o RMSE e o R^2 . Isso reforça sua aplicação para este estudo, que utiliza dados de fluxo de clientes, datas e volume diário de vendas.

Para garantir maior confiabilidade nas previsões, foi aplicada a técnica de validação cruzada, do inglês *cross-validation*. Essa abordagem alterna os conjuntos de treino e teste ao longo de várias iterações, permitindo uma avaliação mais justa e robusta dos modelos. Conforme destacam Linder e Wolfinger (2022), essa técnica contribui para evitar o sobreajuste, do inglês *overfitting*, e melhora a capacidade dos algoritmos de generalizar seus resultados, o que é especialmente importante em cenários com alta variabilidade temporal, como o do varejo.

2.3.1 Modelo preditivo de previsão Floresta Aleatória

O modelo Floresta Aleatória foi introduzido por Leo Breiman em 2001 como uma técnica de ML baseada em conjuntos de árvores de decisão. Sua proposta visava melhorar a estabilidade e a precisão dos modelos, reduzindo o risco de sobreajuste, a partir da construção de múltiplas árvores de decisão e da combinação de seus resultados de forma agregada (Breiman, 2001).

Esse algoritmo funciona criando diversas árvores de decisão com subconjuntos aleatórios de dados e variáveis. Ao final, as previsões de cada árvore são agregadas, geralmente por média ou por votação, resultando em uma previsão mais robusta. Essa abordagem reduz a variância e melhora a estabilidade do modelo. Jo *et al.* (2025) e Avinash *et al.* (2025) confirmam a eficácia do Floresta Aleatória na modelagem de dados complexos, destacando sua habilidade de lidar com ruído, interpretar a

importância das variáveis e manter bons resultados mesmo em bases com muitas variações.

2.3.2 Modelo preditivo de previsão Impulsioneamento de Gradiente Extremo (XGBoost)

O algoritmo XGBoost foi desenvolvido por Tianqi Chen e Carlos Guestrin em 2016 e se destaca por ser uma evolução dos métodos de *boosting*, com foco na eficiência computacional e na melhoria da performance preditiva (Chen; Guestrin, 2016). Ele funciona ajustando novos modelos sobre os resíduos dos anteriores, de forma sequencial, para minimizar os erros de forma otimizada.

O XGBoost funciona construindo modelos de forma sequencial, onde cada nova árvore tenta corrigir os erros cometidos pelas anteriores. Ele se destaca por otimizar uma função de perda com regularização, o que ajuda a evitar que o modelo fique complexo demais e acabe se ajustando demais aos dados, o famoso sobreajuste. Entre seus pontos fortes estão a capacidade de lidar bem com dados esparsos, o tratamento automático de valores ausentes e o alto desempenho mesmo com grandes volumes de dados, ele oferece ótimos resultados em tarefas de previsão, com boa interpretação das variáveis e uma performance estável mesmo em cenários com alta variabilidade (Shaik *et al.*, 2025).

2.3.3 Modelo preditivo de previsão ARIMA

O modelo ARIMA é um dos métodos clássicos mais conhecidos na análise de séries temporais. Ele foi desenvolvido a partir dos estudos de Box e Jenkins na década de 1970, que consolidaram uma metodologia capaz de descrever e prever comportamentos ao longo do tempo com base em padrões históricos (Box *et al.*, 2015).

O modelo ARIMA funciona ajustando valores futuros de uma série a partir das relações entre observações passadas e erros anteriores. Ele combina três etapas principais: a autorregressiva, que usa valores defasados da própria série; a integração, que transforma dados não estacionários em uma sequência mais estável;

e a média móvel, que considera o comportamento dos resíduos de previsão. Essa estrutura permite capturar padrões lineares de tendência e dependência temporal, sendo uma das formas mais conhecidas e consolidadas de modelagem de séries temporais (Kontopoulou *et al.*, 2023).

No presente estudo, será utilizada a variante Modelo Autorregressivo Integrado de Médias Móveis Sazonal com Regressão Exógena, do inglês *Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors* (SARIMAX), que estende o modelo ARIMA tradicional ao permitir a inclusão de variáveis externas e lida melhor com séries temporais (Alharbi; Csala, 2022).

2.3.4 Técnica de validação cruzada para avaliação estatística

A validação cruzada, introduzida por Mervyn Stone em 1974, é uma técnica estatística utilizada para avaliar o desempenho de modelos preditivos a partir da subdivisão dos dados em múltiplos conjuntos de treino e teste (Stone, 1974). A abordagem mais comum, conhecida como *k-fold*, consiste em dividir o conjunto de dados em k partes, treinando o modelo k vezes com diferentes combinações de dados.

No método mais comum, o *k-fold*, os dados são divididos em k partes. O modelo é treinado em $k - 1$ delas e testado na parte restante, repetindo o processo k vezes para garantir que cada ponto de dado seja usado tanto no treino quanto na validação. Segundo Sullivan *et al.* (2025) e Hwang *et al.* (2025) essa técnica permite uma avaliação mais robusta da generalização do modelo, reduzindo o risco de sobreajuste e proporcionando maior confiabilidade aos resultados. Os autores salientam que a validação cruzada continua sendo uma ferramenta indispensável para avaliar e comparar modelos em cenários complexos e com variabilidade temporal.

2.4 ABORDAGEM UTILIZADA PARA TRATAMENTO DE DADOS DE CONTAGEM ZERADOS

Séries temporais de contagem podem apresentar períodos com valores iguais a zero devido a falhas de registro, interrupções de captura ou limitações dos sensores. Esses valores podem distorcer a estrutura da série e comprometer a qualidade das

previsões, exigindo técnicas específicas para sua identificação e correção (Kim *et al.*, 2023). A literatura apresenta diferentes abordagens para tratar séries temporais de contagem com excesso de zeros ou lacunas sequenciais.

Modelos inflado de zeros, do inglês *zero-inflated*, que combinam distribuições como Poisson ou Binomial Negativa com covariáveis adicionais, são amplamente empregados para ajustar taxas de contagem em contextos com muitos valores nulos, fornecendo estimativas mais realistas da contagem esperada (Fávero *et al.*, 2021).

Mais recentemente, surgiram técnicas de imputação supervisionada baseadas em aprendizado profundo, como o modelo Imputação de Séries Temporais Baseada em Autoatenção, do inglês *Self-Attention-based Imputation for Time Series (SAITS)*, que utiliza mecanismos de atenção para capturar dependências temporais e inter-relações entre variáveis, alcançando imputações mais precisas em séries multivariadas (Du *et al.*, 2023).

2.5 VARIÁVEIS EXÓGENAS, EFEITOS DE CALENDÁRIO, PROMOÇÕES E OUTLIERS EM SÉRIES TEMPORAIS

Séries temporais de vendas são fortemente influenciadas por fatores externos, como feriados, datas sazonais, fins de semana e ações promocionais. Esses elementos alteram o comportamento natural da demanda e, quando ignorados, podem comprometer a precisão das previsões (De Baets; Harvey, 2023).

Estudos recentes mostram que a inclusão de marcadores de calendário como variáveis explicativas, especialmente feriados e eventos sazonais, contribui para melhorar o desempenho dos modelos preditivos, pois esses efeitos explicam parte relevante da variação observada ao longo do tempo (Saputra; Kumar, 2024).

Da mesma forma, pesquisas no varejo indicam que promoções não impactam apenas o dia do evento, mas modificam a demanda ao longo de todo o ciclo, gerando efeitos de antecipação e pós-promoção. Tratar esses padrões como simples ruído pode distorcer o aprendizado dos modelos e reduzir sua capacidade de representar adequadamente a dinâmica real da série (Hewage; Perera; Bandara, 2025).

Além dos efeitos exógenos, séries temporais também podem apresentar valores atípicos, que refletem tanto erros de registro quanto variações da operação. Revisões sobre detecção de *outliers* em séries temporais destacam que muitas

anomalias são, na prática, sinais do próprio sistema, devendo ser analisadas com critério antes de qualquer limpeza automática (Blázquez-García *et al.*, 2021). Para diagnóstico, filtros robustos como o de Hampel continuam adequados, eles são baseados em mediana e desvio absoluto mediano e têm recebido aprimoramentos recentes para acelerar o processamento sem perder precisão de detecção (Roos-Hoefgeest Toribio *et al.*, 2025).

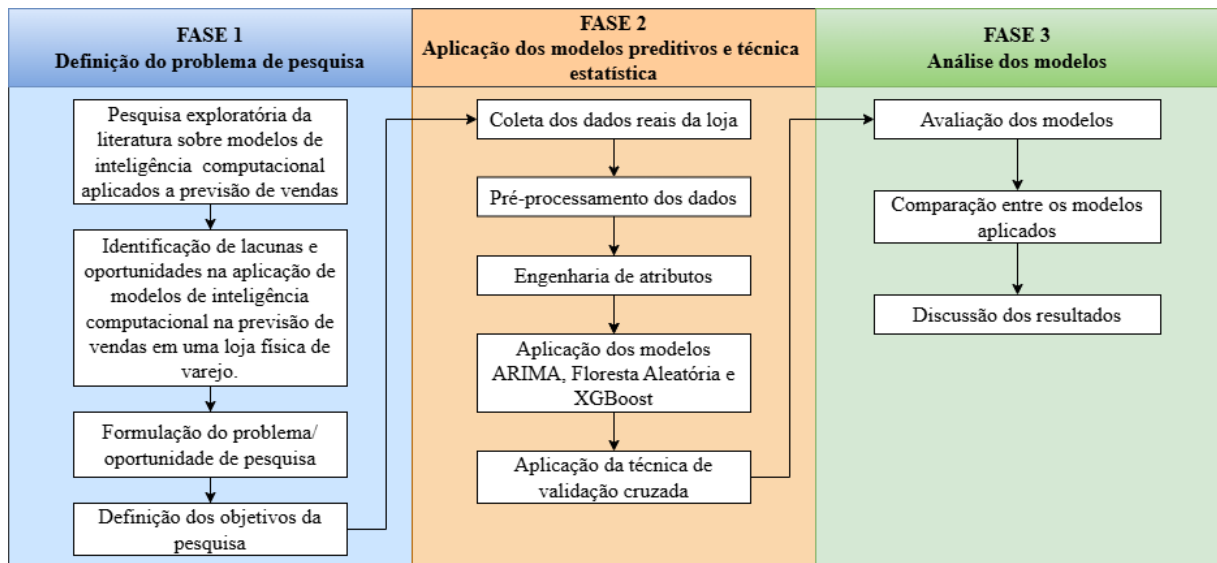
3 METODOLOGIA

Esta seção apresenta o procedimento metodológico, organizado em três fases: pesquisa exploratória da literatura, aplicação dos modelos preditivos e análise dos resultados.

3.1 PROCEDIMENTO METODOLÓGICO

A Figura 1 ilustra o procedimento metodológico seguido no desenvolvimento desta pesquisa. Ele está organizado em três fases principais, compostas por atividades interdependentes que resultam no cumprimento dos objetivos específicos definidos. As subseções descrevem cada uma dessas fases.

Figura 1 - Procedimento Metodológico



Fonte: Elaborado pelo autor (2025)

3.1.1 Fase 1: Definição do problema de pesquisa

A etapa inicial da pesquisa envolveu uma pesquisa exploratória da literatura científica. Segundo Dos Santos *et al.* (2022), esse tipo de revisão é essencial para compreender os fundamentos teóricos de um tema e orientar decisões metodológicas

ao longo do estudo. Ela permite mapear o que já foi produzido sobre o assunto, destacando tendências, limitações e oportunidades de aprofundamento.

O foco principal desta etapa foi identificar como técnicas e modelos de inteligência computacional têm sido aplicadas na previsão de vendas, com ênfase no contexto do varejo físico e outros cenários. Foram selecionados artigos que analisam o uso de algoritmos como Floresta Aleatória, XGBoost, ARIMA, entre outros. Esse levantamento ajudou a fundamentar a escolha dos modelos a serem utilizados neste estudo.

Além disso, a análise da literatura serviu de base para definir o problema e os objetivos específicos da pesquisa, e construir o referencial teórico. Isso possibilita que as decisões tomadas ao longo da pesquisa estejam alinhadas com os estudos mais atuais e relevantes da área.

3.1.2 Fase 2: Aplicação dos modelos preditivos e técnica estatística

Esta fase foi desenvolvida com base nos dados de uma loja de varejo localizada no Amazonas, a qual possui um sistema de monitoramento. Inicialmente, foi realizada a coleta de dados reais. Para Doring, Grumbach e Reusch (2024), obter dados diretamente da fonte é fundamental para garantir a representatividade e a qualidade dos resultados em pesquisas aplicadas. O conjunto de dados incluiu informações como o volume diário de vendas, fluxo de clientes e datas correspondentes.

Após a coleta, os dados passaram por processos de pré-tratamento, que envolvem etapas como limpeza, padronização e tratamento de valores ausentes. Conforme Das e Maji (2024), essas ações são importantes para preparar os dados para os modelos de previsão, evitando distorções nos resultados. Em seguida, foi realizada a engenharia de atributos, uma prática que de acordo com Ma, Jørgensen e Ma (2023), contribui para a extração de variáveis mais relevantes e informativas a partir das variáveis originais, ampliando o potencial preditivo dos algoritmos.

Posteriormente, os modelos ARIMA, Floresta Aleatória e XGBoost foram implementados com auxílio de bibliotecas em Python, utilizando a técnica de validação cruzada. Essa abordagem foi utilizada para reforçar a confiabilidade dos resultados, uma vez que permite testar os modelos em diferentes partições dos dados.

Para garantir uma avaliação consistente do desempenho dos algoritmos, foram comparados os resultados obtidos em dados de treino e em dados não vistos. Segundo Géron (2022), essa comparação é essencial para diagnosticar o sobreajuste, pois diferenças acentuadas entre as etapas indicam a chamada lacuna de generalização do inglês *generalization gap*. Essa interpretação também está alinhada com a perspectiva de Aburass e Abu Rumman (2024), que recomendam quantificar o sobreajuste observando o desempenho relativo entre treino e validação. Assim, a análise da relação entre erro de treino e erro de teste foi utilizada para verificar se os modelos estavam aprendendo padrões específicos do histórico ou se apresentavam boa capacidade de generalização.

Somado a isso, empregou-se como referência uma linha de base simples do tipo modelo ingênuo sazonal do inglês *seasonal naïve*, que utiliza o valor observado no mesmo dia da semana anterior como previsão. Esse tipo de *baseline* é recomendado para séries temporais com forte padrão semanal e funciona como patamar mínimo de comparação, permitindo avaliar se modelos mais sofisticados realmente oferecem ganhos relevantes (Hyndman; Athanasopoulos, 2021).

Por outro lado, além da validação *k-fold* tradicional, também foi utilizado o esquema de validação temporal do tipo *walk-forward*, que preserva a ordem cronológica dos dados e evita o vazamento de informação do futuro para o passado. Essa abordagem é recomendada para séries com dependência temporal e sazonalidade bem definidas, pois permite avaliar o modelo em blocos consecutivos de tempo, simulando a prática real de previsão (Cerqueira; Torgo; Mozetič, 2020).

A definição do tamanho das janelas de treino e teste é um aspecto essencial nesse tipo de validação. Hewamalage, Bergmeir e Bandara (2021) destacam que a escolha adequada da janela influencia diretamente a capacidade dos modelos em capturar dependências temporais e padrões sazonais, além de impactar sua estabilidade ao longo das divisões. Por esse motivo, diferentes tamanhos de janelas foram testados ao longo deste estudo, buscando o equilíbrio entre representatividade temporal e consistência nas métricas.

3.1.3 Fase 3: Análise dos modelos

Na terceira fase foi avaliado o desempenho dos modelos aplicados. Foram utilizadas métricas estatísticas amplamente reconhecidas na literatura, como o MAE, RMSE e R^2 . Ghafariasl, Zeinalnezhad e Ahmadishokoh (2024) afirmam que essas métricas são fundamentais para mensurar com precisão a qualidade das previsões e comparar o desempenho entre diferentes abordagens.

Com base nos resultados obtidos, foi feita uma comparação entre os modelos, o que na visão de Sekeroglu *et al.* (2022) é uma etapa fundamental para garantir a escolha da abordagem mais eficiente em contextos reais de aplicação. Essa análise levará em conta não apenas as métricas numéricas, mas também aspectos voltados a interpretação e consistência dos resultados.

Por fim, os resultados foram discutidos e organizados usando gráficos e quadros, para uma apresentação clara e objetiva dos achados. Essa etapa final serviu como base para recomendações práticas e sugestões de continuidade da pesquisa.

3.2 ENQUADRAMENTO METODOLÓGICO

As etapas deste estudo foram organizadas de modo a atender aos objetivos específicos definidos na pesquisa. Para cada objetivo, foi selecionado um método compatível com a natureza das atividades envolvidas. O Quadro 1 apresenta a síntese do enquadramento metodológico, evidenciando o alinhamento entre as fases da pesquisa, os objetivos correspondentes e os métodos adotados.

Quadro 1 - Enquadramento metodológico

FASE	OBJETIVOS	MÉTODOS
1	Identificar na literatura estudos relacionados a técnicas e modelos de inteligência computacional aplicados à previsão de vendas.	Pesquisa teórica qualitativa 1. Pesquisa exploratória da literatura.
2	Aplicar modelos e técnicas de previsão de vendas utilizando dados de fluxo de clientes, vendas e informações temporais, testando sua eficácia no contexto do varejo.	Pesquisa aplicada quantitativa 1. Aplicação de modelos preditivos e técnica estatística. 2. Estudo de caso
3	Analisar o desempenho dos modelos utilizados, considerando métricas de acurácia e sua utilidade para a tomada de decisões gerenciais.	Pesquisa aplicada quantitativa 1. Avaliação estatística dos resultados.

Fonte: Elaborado pelo autor (2025)

4 APLICAÇÃO DOS MODELOS E TÉCNICA ESTATÍSTICA

Esta seção descreve a aplicação dos modelos preditivos e da técnica estatística adotada na pesquisa, detalhando as etapas de coleta dos dados, preparação das bases e implementação dos algoritmos.

4.1 OBJETO DE ESTUDO DESTA PESQUISA

A empresa analisada nesta pesquisa é uma loja de varejo localizada no estado do Amazonas, atuando com atendimento ao público todos os dias da semana, incluindo meio período aos domingos. Trata-se de um estabelecimento de grande circulação regional.

Para apoiar suas atividades, a loja utiliza sistemas amplamente empregados no varejo, como o sistema de gestão Sistemas, Aplicações e Produtos em Processamento de Dados, do inglês *Systems Applications and Products in Data Processing* (SAP) para registro e controle de vendas, e o sistema de videomonitoramento Digifort para contabilização de fluxo de clientes e aplicações internas para consulta e integração de dados operacionais. Esses recursos fazem parte da estrutura tecnológica que sustenta o funcionamento diário da empresa.

4.2 COLETA DOS DADOS REAIS DA LOJA

A etapa de coleta dos dados constituiu-se como o ponto de partida para a aplicação dos modelos preditivos, garantindo que as informações utilizadas refletissem a realidade operacional da loja estudada. Os dados foram obtidos a partir de duas fontes distintas: o Digifort, responsável pela contagem de clientes, e o sistema de gestão SAP, de onde foram extraídas as informações referentes às vendas.

No caso da contagem de clientes, a extração foi realizada por meio de um código desenvolvido em *notebook* no Databricks, que acessou diretamente a Interface de Programação de Aplicações, do inglês *Application Programming Interface* (API) do Digifort, que é uma plataforma de videomonitoramento digital. O sistema passou a

disponibilizar registros de forma contínua a partir de 2023, porém, como sua implantação ocorreu em 2022, a empresa ainda possuía registros manuais desse ano. Para cobrir o período inicial, foi incorporada uma planilha em formato Excel, contendo a contagem de clientes entre 01/07/2022 e 31/12/2022. Assim, o conjunto final de dados abrangeu o intervalo de julho de 2022 até outubro de 2025, consolidando informações contínuas e consistentes.

Os dados de vendas foram disponibilizados por meio de uma tabela registrada na plataforma unificada de análise de dados e computação em nuvem Databricks, alimentada diretamente pelo sistema SAP da empresa. A extração também foi realizada via código em *notebook*, e os resultados foram salvos em arquivos de Valores Separados por Vírgula, do inglês *Comma-Separated Values* (CSV), possibilitando integração com os demais conjuntos e padronizando o formato de armazenamento.

O Quadro 2 apresenta um resumo das fontes utilizadas no processo de coleta, destacando as variáveis contempladas, o período de disponibilidade dos registros, os métodos empregados e o formato final adotado. Observa-se que a variável de “Contagem de clientes” foi obtida tanto pelo sistema Digifort, via API, quanto por registros manuais em planilha Excel, o que possibilitou recuperar informações desde a implantação do sistema em 2022. Já os dados de vendas diárias, provenientes do SAP, foram acessados diretamente pelo Databricks, cobrindo o mesmo período de análise.

Quadro 2 - Resumo das fontes e períodos de cobertura

Fonte	Variável	Período disponível	Método de coleta	Formato final
Digifort (API)	Contagem de clientes	01/01/23 – 31/10/25	Código em notebook Databricks	CSV
Planilha Excel	Contagem de clientes	01/07/22 - 31/12/22	Importação manual	CSV
SAP (Databricks)	Vendas diárias	2022 - 10/2025	Código em notebook Databricks	CSV

Fonte: Elaborado pelo autor (2025)

Essa integração entre diferentes fontes assegurou maior completude da base de dados, permitindo que séries temporais mais longas fossem formadas para a

posterior análise preditiva e para o treinamento dos modelos de inteligência computacional aplicados neste estudo. A escolha pelo formato CSV para todos os registros teve como objetivo simplificar a manipulação, facilitar replicações dos experimentos e manter a compatibilidade com bibliotecas analíticas do Python utilizadas nas fases seguintes.

4.3 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados representou uma etapa fundamental para garantir a qualidade, a consistência e a adequação das informações antes da aplicação dos modelos preditivos. Essa fase envolveu múltiplos procedimentos, realizados no ambiente Databricks, com o objetivo de padronizar variáveis, eliminar redundâncias e integrar diferentes fontes de dados em uma única base consolidada.

O primeiro passo consistiu na renomeação e harmonização das colunas. Essa ação foi especialmente necessária no arquivo proveniente do Excel referente à contagem de clientes, cujos rótulos não estavam uniformes em relação às demais bases. Em seguida, concentrou-se o tratamento sobre a tabela de vendas, que apresentava registros no nível transacional, em que cada linha correspondia a uma venda individual, contendo ainda informações sensíveis, como identificação do cliente e código de contrato. Por restrição da empresa, tais colunas foram descartadas, e foi implementado um procedimento de agregação por data, de modo a transformar registros diários em uma única entrada de vendas para cada dia.

Após o tratamento individual de cada fonte, foi realizada a integração das bases. Para tanto, os dados foram agrupados pelas variáveis de data e identificador da loja, formando uma estrutura única contendo as colunas padronizadas: ID da loja, Data, Nome da loja, Contagem de clientes e Vendas. Esse processo resultou em uma base consolidada, uniforme com 1150 registros e pronta para análises.

O Quadro 3 exemplifica como a base integrada ficou organizada após o processo de padronização. Cada linha corresponde a um registro diário da loja, contendo de forma consolidada o identificador e nome da unidade, a data, a contagem de clientes e o volume total de vendas. Esse formato garante consistência e facilita as análises comparativas entre fluxo de clientes e desempenho comercial.

Quadro 3 - Estrutura final da base integrada

ID da loja	Data	Nome da loja	Contagem de clientes	Vendas
1	2025-08-16	Amazonas	1624	476
1	2025-08-17	Amazonas	571	180
1	2025-08-18	Amazonas	1950	496

Fonte: Elaborado pelo autor (2025)

A etapa seguinte foi a limpeza dos dados, que envolveu a identificação e o tratamento de valores ausentes, nulos e registros duplicados. Essa tarefa teve especial relevância para os dados de 2022, uma vez que parte das informações havia sido registrada manualmente, aumentando a probabilidade de inconsistências. O processo garantiu que as séries finais utilizadas nos modelos apresentassem maior robustez e confiabilidade, minimizando distorções que poderiam comprometer os resultados preditivos.

Um problema encontrado durante essa fase foi a ocorrência de valores iguais a zero na contagem de clientes. Esses registros se devem a instabilidades do sistema Digifort, que em alguns períodos deixou de contabilizar adequadamente a entrada de pessoas na loja. Em determinados casos, foram observados intervalos de até nove dias consecutivos com valores zerados, embora houvesse registros de vendas para as mesmas datas, evidenciando que a loja operava normalmente. Esse tipo de situação requer um tratamento cuidadoso, uma vez que zeros artificiais podem induzir os modelos a interpretações equivocadas. De todos os dados, foram identificados 22 registros zerados de contagem de clientes, representando cerca de 1,9% dos dias da base de dados.

Diante deste cenário, optou-se por não realizar a substituição imediata desses valores zerados na fase de pré-processamento, mantendo-os registrados na base para preservar a integridade do conjunto original. A imputação será tratada na etapa de engenharia de atributos, momento em que serão criadas variáveis adicionais, que oferecem maior contexto para a estimativa de valores aceitáveis. Essa estratégia permite que a correção dos registros zerados seja conduzida de forma mais consistente, considerando padrões sazonais e comportamentos típicos de consumo, além de facilitar a transparência metodológica ao diferenciar a limpeza inicial da base das transformações aplicadas posteriormente.

4.4 ENGENHARIA DE ATRIBUTOS

Esta etapa apresenta o processo de engenharia de atributos realizado para enriquecer as bases de dados utilizadas nos modelos. São descritas as transformações aplicadas às variáveis originais, bem como a criação de novos atributos sazonais e contextuais capazes de melhorar o desempenho preditivo dos algoritmos.

4.4.1 Criação de variáveis sazonais e contextuais

O primeiro passo consistiu na criação da variável “Dia da semana” a partir da variável “Data”, permitindo identificar qual dia da semana cada registro estava associado. Essa informação é importante, visto que o comportamento dos consumidores varia sistematicamente entre dias úteis e finais de semana, refletindo diretamente na demanda observada em loja.

Em seguida, foram incorporadas variáveis relacionadas a feriados e datas comemorativas, reconhecidamente relevantes no varejo físico. Para tanto, foram criadas duas formas de representação desses eventos: variáveis categóricas contendo o nome do feriado “Nome do feriado” e da data comercial “Data comemorativa”, e variáveis *dummies*, que são variáveis binárias indicadoras, “Feriado” e “Comemorativa”, que assumem valores 1 quando o evento ocorre no dia e 0 caso contrário. Essa combinação permite que os modelos identifiquem não apenas a presença do evento, mas também o tipo específico de ocasião, favorecendo a captura de picos ou quedas atípicas de movimento frequentemente associados a esses períodos.

4.4.2 Tratamento dos zeros sequenciais

Superada essa etapa de criação de variáveis, passou-se ao tratamento de uma inconsistência importante, os registros zerados da variável “Contagem de clientes”, causados por falhas no sistema de monitoramento. Para lidar com esse problema, foi

inicialmente criada a coluna “Zero”, que é uma variável *dummy*, que identifica se o valor de contagem era igual a zero. Essa etapa marcou o início do pré-saneamento dos dados, permitindo separar os casos confiáveis dos registros suspeitos.

Neste estudo, considerando as características específicas da série, como a presença de sequências de zeros não reais, ocorrência de vendas registradas nos mesmos dias afetados e variáveis auxiliares já disponíveis, tratou-se o problema como um caso de excesso de zeros, como visto na Seção 2.4. Diante desse contexto, optou-se pela aplicação de um modelo de regressão de contagem, o Modelo Linear Generalizado, do inglês *Generalized Linear Model* (GLM) de Poisson com link log por se mostrar mais aderente à natureza dos dados. Diferentemente de outros estudos que utilizaram esse tipo de regressão apenas como modelo preditivo, aqui ele foi empregado como ferramenta de imputação, permitindo substituir zeros artificiais por estimativas mais confiáveis, baseadas no comportamento histórico da série.

Para assegurar consistência e evitar valores fora da faixa típica, as imputações foram combinadas com restrições baseadas nos quantis q10 e q99 de cada dia da semana, além do uso de variáveis sazonais e contextuais. Essa estratégia buscou equilibrar rigor estatístico, maior interpretação e aderência ao padrão real dos dados, reduzindo vieses sem impor suposições excessivas.

Na prática, o procedimento ocorreu em etapas sucessivas. Primeiramente, calcularam-se os quantis de referência por dia da semana, de modo a estabelecer intervalos aceitáveis de variação das contagens. Os valores obtidos são apresentados no Quadro 4, evidenciando diferenças significativas entre dias da semana, o menor fluxo típico aos domingos, e os maiores patamares observados nas sextas-feiras e nas segundas-feiras, com picos superiores a 4.000 clientes. Esses limites serviram de guia para o processo de imputação.

Quadro 4 - Quantis por dia da semana q10 e q99

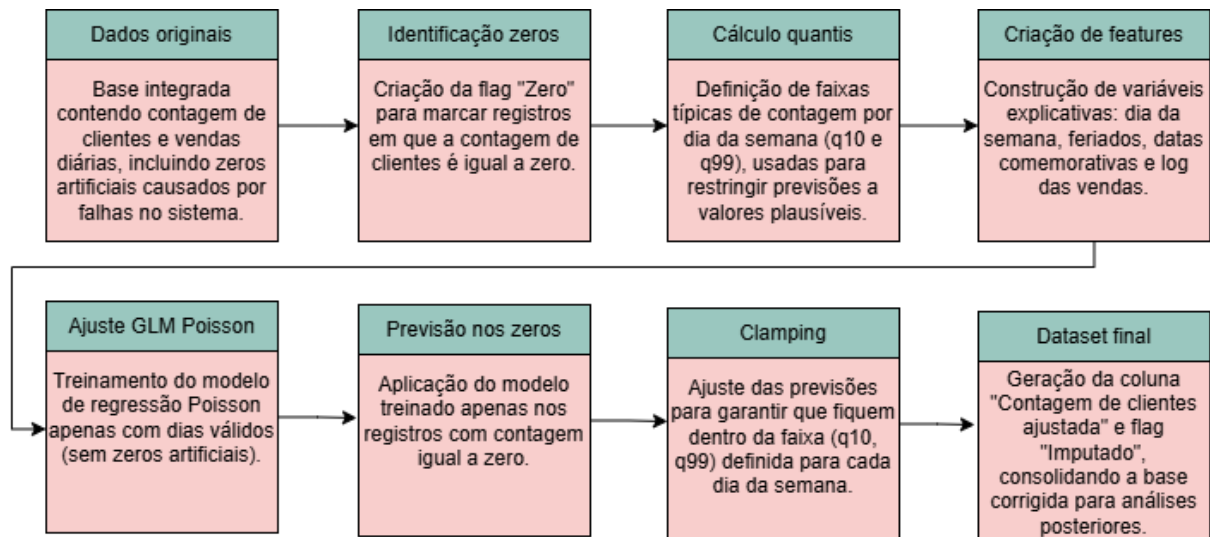
Dia da Semana	q10	q99
Domingo	452	1082
Segunda-feira	1877	4326
Terça-feira	1651	4407
Quarta-feira	1558	4116
Quinta-feira	1548	3845
Sexta-feira	1657	4566
Sábado	1580	3724

Fonte: Elaborado pelo autor (2025)

Em seguida, construiu-se um *pipeline* de *features* (esteira automática de preparo das características usadas pelo modelo), incluindo variáveis sazonais, variáveis de contexto e variáveis de escala como o logaritmo do volume de vendas. Com esse vetor explicativo, o modelo Poisson foi ajustado apenas sobre dias válidos, ou seja, com contagens maiores que zero, de forma a aprender os padrões reais de fluxo de clientes. Na sequência, o modelo foi utilizado para prever apenas os registros com contagem igual a zero.

As previsões brutas foram então ajustadas através do *clamping* (limitação por faixas), restringindo os valores ao intervalo típico observado para cada dia da semana. Essa etapa teve como objetivo evitar imputações fora da realidade histórica, assegurando consistência estatística. O resultado foi consolidado na coluna “Contagem de clientes ajustada”, acompanhada da *flag*, ou seja, do indicador “Imputado”, que indica se o valor original foi substituído. O fluxograma da Figura 2 sintetiza essa sequência metodológica, reforçando a clareza do processo adotado.

Figura 2 - Fluxograma da metodologia aplicada para substituição dos valores zerados

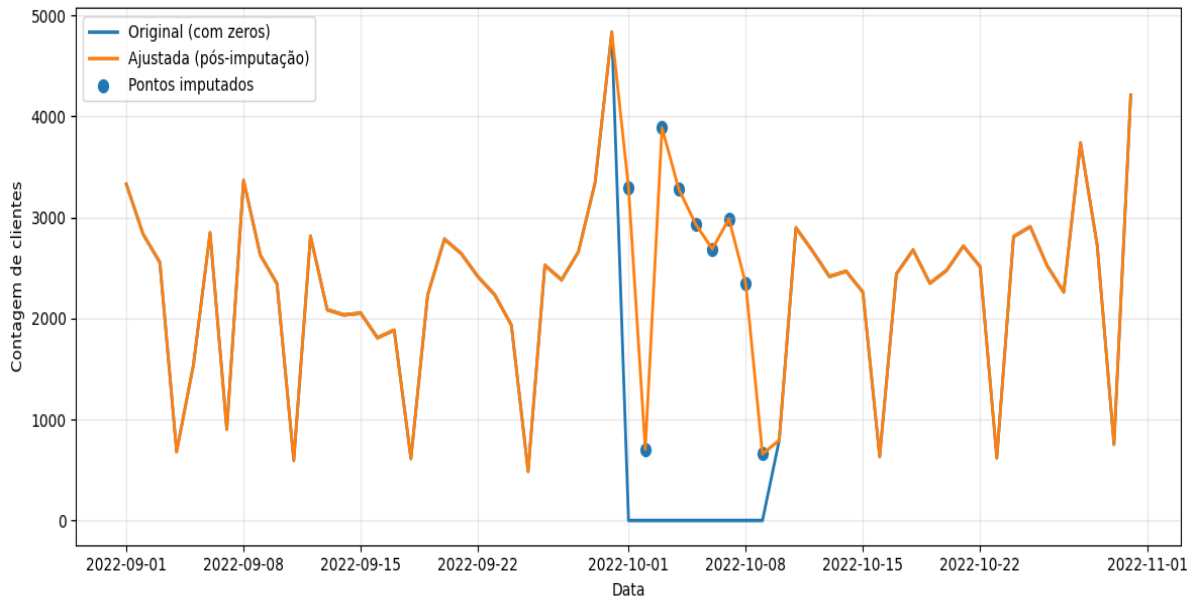


Fonte: Elaborado pelo autor (2025)

O impacto prático da imputação é evidenciado na Figura 3, a qual compara a série original com zeros, com a série ajustada. Os pontos azuis representam os valores imputados, enquanto a linha laranja mostra a série corrigida. Observa-se que o modelo conseguiu reconstruir o comportamento esperado para os dias de falha, sem gerar distorções abruptas. Ao contrário, as imputações mantiveram-se dentro da faixa

típica de cada dia da semana, preservando a sazonalidade e evitando tendências que poderiam comprometer a análise.

Figura 3 - Comparação entre a série original e a série ajustada, com destaque para os valores imputados em substituição aos zeros artificiais



Fonte: Elaborado pelo autor (2025)

Por fim, foi conduzida uma validação de simulação computacional, do inglês *in-silico* para mensurar a qualidade do processo de imputação. Nesse experimento, 10% dos dias válidos foram mascarados como zeros e submetidos ao mesmo procedimento de reconstrução. Os resultados mostraram um MAE de 149 clientes e um RMSE de 221 clientes. Considerando que as contagens típicas variam entre 1.500 e 4.500 clientes por dia, tais métricas correspondem a desvios relativos modestos de 3% a 9%, indicando que o método reproduziu de forma confiável os padrões reais da série.

Para verificar o impacto do tratamento aplicado, foram calculadas estatísticas descritivas antes e depois da substituição dos zeros. O Quadro 5 apresenta a comparação entre os dois conjuntos de dados, destacando as mudanças na variável de contagem de clientes e a estabilidade dos valores de vendas.

Quadro 5 - Estatísticas descritivas antes e depois do tratamento da variável de contagem de clientes e para variável de vendas

Estatística	Contagem de clientes (Antes)	Contagem de clientes (Depois)	Vendas
Média	2030,56	2072,61	539,95
Mediana	2076,00	2096,50	528,50
Moda	0,00	2235,00	641,00
Mínimo	0,00	237,00	59,00
Máximo	4835,00	4835,00	1429,00
Desvio Padrão	874,49	838,51	219,93
Variância	764.726,1	703.101,8	48.367,79
1º Quartil (Q1)	1635,75	1664,50	415,00
3º Quartil (Q3)	2584,25	2599,25	679,00

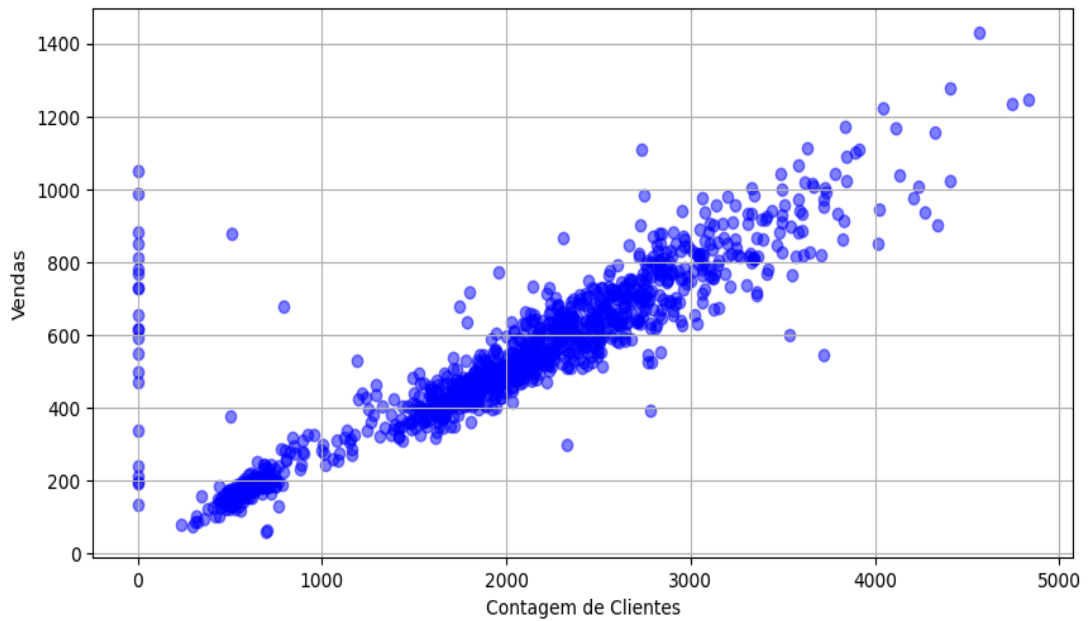
Fonte: Elaborado pelo autor (2025)

Observa-se que, após o ajuste, houve um aumento da média de 2030,56 para 2072,61 e da mediana de 2076,00 para 2096,50, refletindo a remoção de distorções causadas por registros de contagem zerados. A moda também passou de 0 para 2235, evidenciando a correção dos zeros artificiais. Além disso, o valor mínimo foi ajustado de 0 para 237, representando um patamar mais realista de fluxo de clientes.

No que se refere à dispersão dos dados, o desvio padrão caiu de 874,49 para 838,51 e a variância reduziu-se de 764.726,15 para 703.101,81. Esse decréscimo indica que os dados ficaram menos dispersos em torno da média, ou seja, a substituição dos zeros diminuiu a variabilidade extrema causada por valores artificiais, tornando a distribuição mais consistente. Por outro lado, as métricas de vendas permaneceram inalteradas, confirmando que o tratamento afetou apenas a variável de clientes, mantendo a consistência do conjunto de dados.

Além do quadro de estatísticas descritivas, as Figuras 4 e 5, de gráficos de dispersão, ajudam a enxergar com clareza o efeito do tratamento aplicado. Na Figura 4, do conjunto original, é fácil notar a concentração de pontos encostados no eixo vertical, resultado direto dos registros com contagem de clientes igual a zero. Esses valores acabavam criando uma distorção na leitura dos dados, já que não representavam a realidade do fluxo de clientes.

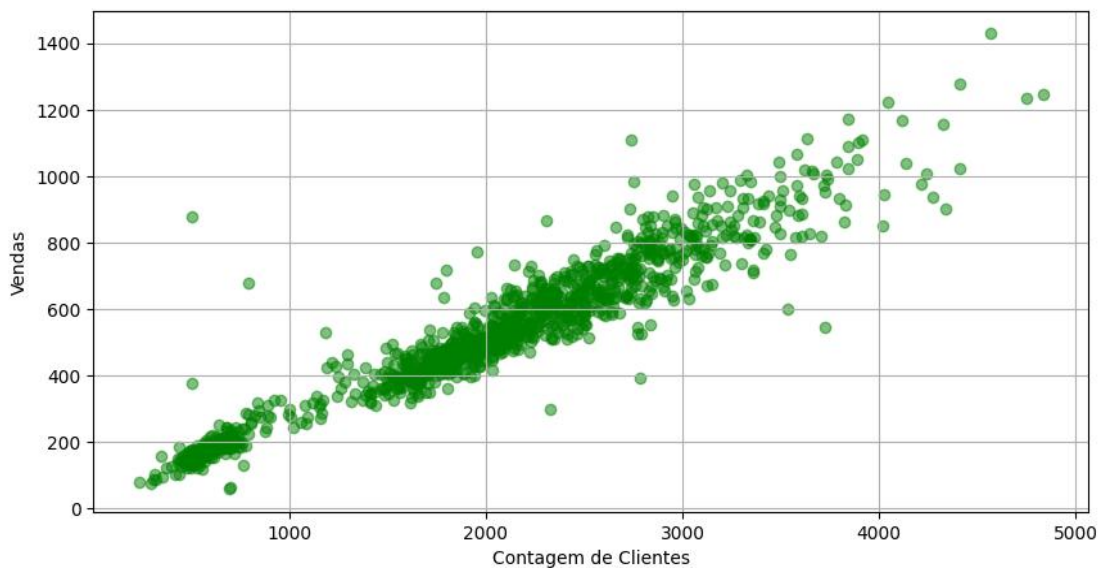
Figura 4 - Dispersão dos dados originais



Fonte: Elaborado pelo autor (2025)

Na Figura 5, após o tratamento, essa distorção desaparece. Os pontos passam a se distribuir de forma mais natural e coerente com a relação esperada entre clientes e vendas. O desenho fica mais limpo e consistente, deixando evidente a tendência, que conforme aumenta o número de clientes, também cresce o volume de vendas.

Figura 5 - Dispersão dos dados tratados



Fonte: Elaborado pelo autor (2025)

Essa comparação visual reforça o que já havia sido identificado nas estatísticas. A retirada dos zeros artificiais tornou o conjunto mais fiel e confiável, sem alterar os valores de vendas. Em outras palavras, o tratamento deixou os dados mais próximos da realidade, facilitando a aplicação dos modelos de inteligência computacional e garantindo a qualidade dos resultados.

4.4.3 Preparação da base para aplicação dos modelos

Encerrada a fase de imputação, passou-se à preparação da base de dados para aplicação dos modelos de inteligência computacional Floresta Aleatória e XGBoost e do modelo estatístico ARIMA na variante ARIMAX. Essa etapa consistiu em selecionar e organizar as variáveis mais relevantes, assegurando que o conjunto estivesse no formato adequado para os algoritmos, que não aceitam atributos categóricos em formato textual.

Primeiramente, definiu-se que a variável alvo, conhecida como *target*, a ser prevista seria “Vendas”, representando o volume diário de vendas registrado pela loja. Como variáveis explicativas, as *features*, optou-se por incluir a contagem de clientes já ajustada, além de indicadores binários que sinalizam se o dia corresponde a um feriado ou a uma data comemorativa. A variável “Data” foi mantida apenas como chave temporal, de forma a possibilitar a vinculação das previsões aos respectivos dias, mas sem ser utilizada como entrada dos modelos.

Considerando a relevância da sazonalidade semanal e mensal no comportamento do varejo, foram estruturadas três versões distintas da base de treino: compacta, expandida e extra-expandida. A versão compacta manteve apenas os atributos essenciais descritas no texto acima, representando uma configuração mais simples e de fácil interpretação. A versão expandida que inclui, adicionalmente, variáveis *dummies*, que representam valores numéricos geralmente 0 e 1, para os dias da semana, derivadas diretamente da coluna “Data”, e a versão extra-expandida incluiu além de variáveis *dummies* semanais, variáveis *dummies* mensais. Essas expansões tiveram como objetivo permitir que os modelos captassem diferenças sistemáticas entre os dias úteis e os finais de semana, bem como padrões específicos de cada dia da semana e meses correspondentes.

O Quadro 6 apresenta as diferenças entre as versões compacta, expandida e extra-expandida, evidenciando a presença de variáveis essenciais em ambas e a inclusão dos *dummies* de dias da semana e mês.

Quadro 6 - Comparação entre as versões da base de treino

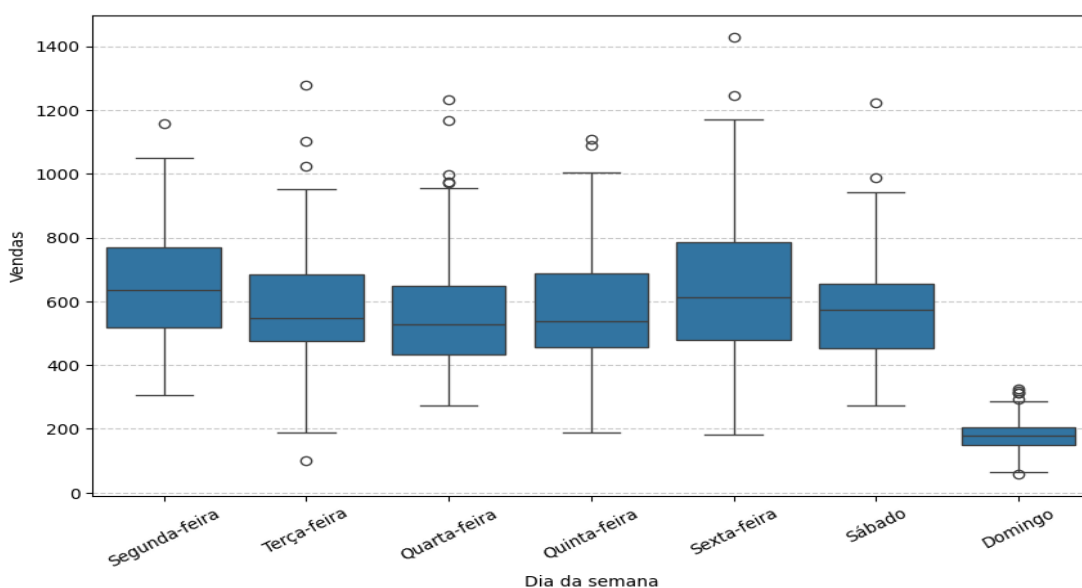
Variável	Compacta	Expandida	Extra-expandida
Contagem de clientes ajustada	SIM	SIM	SIM
Feriado	SIM	SIM	SIM
Comemorativa	SIM	SIM	SIM
<i>Dummies</i> de dia da semana	NÃO	SIM	SIM
<i>Dummies</i> de mês	NÃO	NÃO	SIM
Vendas	SIM	SIM	SIM
Data (chave, não usada)	SIM	SIM	SIM

Fonte: Elaborado pelo autor (2025)

Para justificar empiricamente a necessidade da versão expandida e extra-expandida, analisaram-se as distribuições das variáveis por dia da semana e mês.

A Figura 6 mostra que o volume de vendas apresenta forte variação semanal, com picos às sextas-feiras e valores consistentemente menores aos domingos onde a loja funciona em meio período. Essa evidência indica que o dia da semana exerce influência direta sobre a variável alvo, justificando a inclusão dos *dummies* de dia da semana na base expandida.

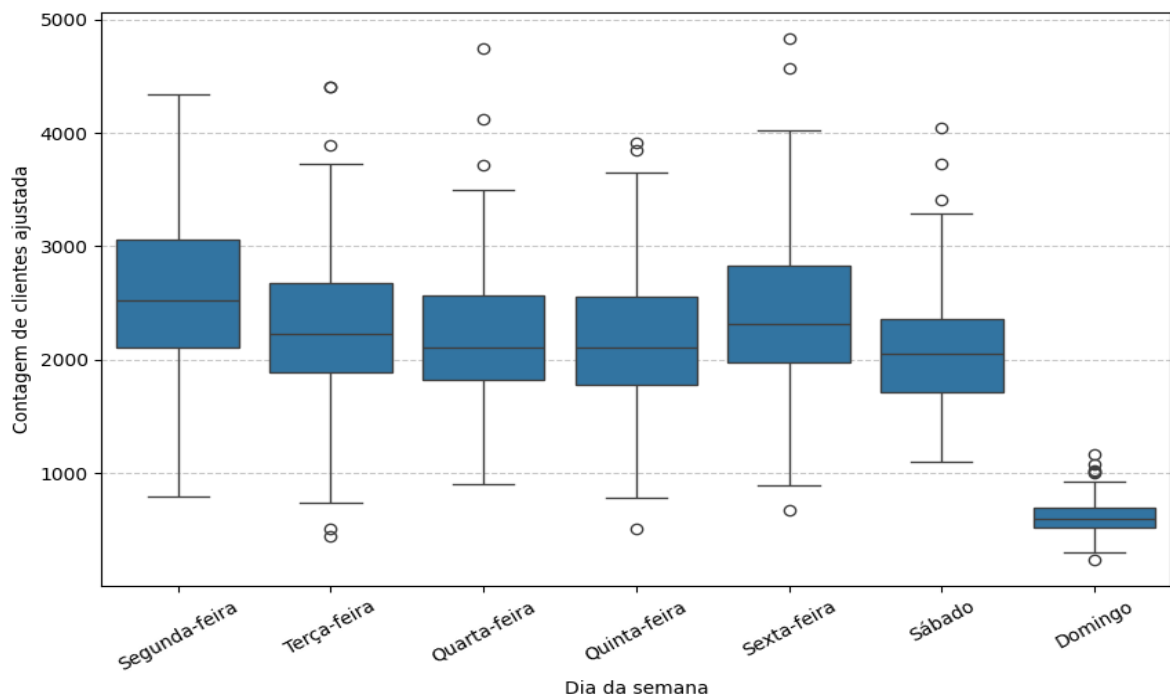
Figura 6 - Boxplot das vendas por dia da semana



Fonte: Elaborado pelo autor (2025)

A Figura 7 apresenta a distribuição da contagem de clientes ajustada, revelando comportamento semelhante ao das vendas, com maior fluxo de clientes nos dias úteis, especialmente segundas e sextas-feiras, e menor nos domingos. Esse padrão reforça a importância de capturar a sazonalidade semanal, uma vez que tanto a variável dependente “Vendas” quanto a principal variável explicativa “Contagem de clientes” são influenciadas por esse fator.

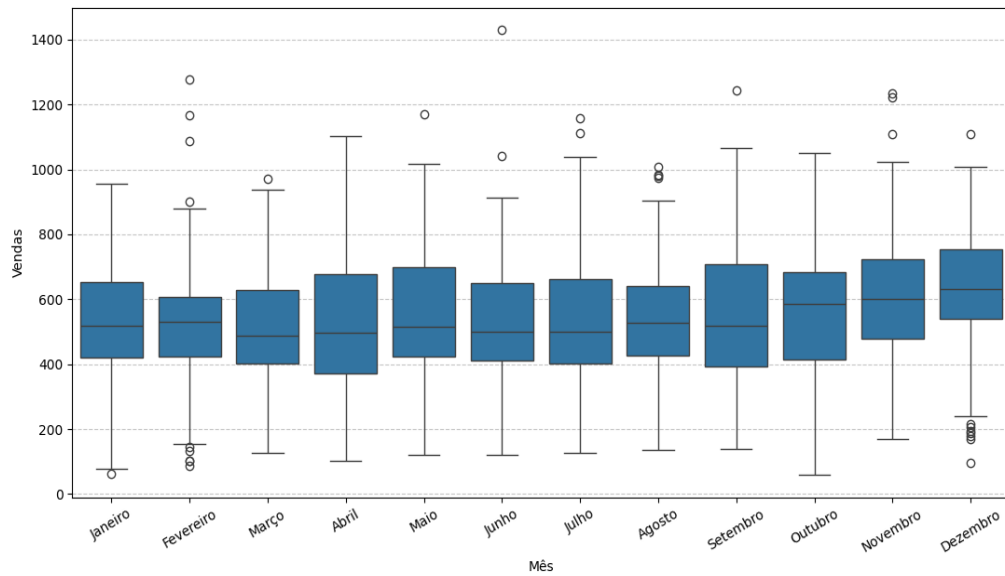
Figura 7 - Boxplot da contagem de clientes por dia da semana



Fonte: Elaborado pelo autor (2025)

A Figura 8 indica que as vendas apresentam variação ao longo dos meses do ano. Embora as diferenças entre os meses não sejam muito acentuadas, a mediana e a faixa central (IQR) mostram mudanças consistentes de mês para mês, com maior dispersão em alguns períodos do último trimestre e níveis um pouco menores em parte do início do ano. Também aparecem valores extremos em determinados meses, possivelmente associados a campanhas ou datas promocionais. Esses indícios sugerem que o mês tem algum grau de influência sobre o desempenho de vendas, o que justifica a inclusão de *dummies* de mês na base.

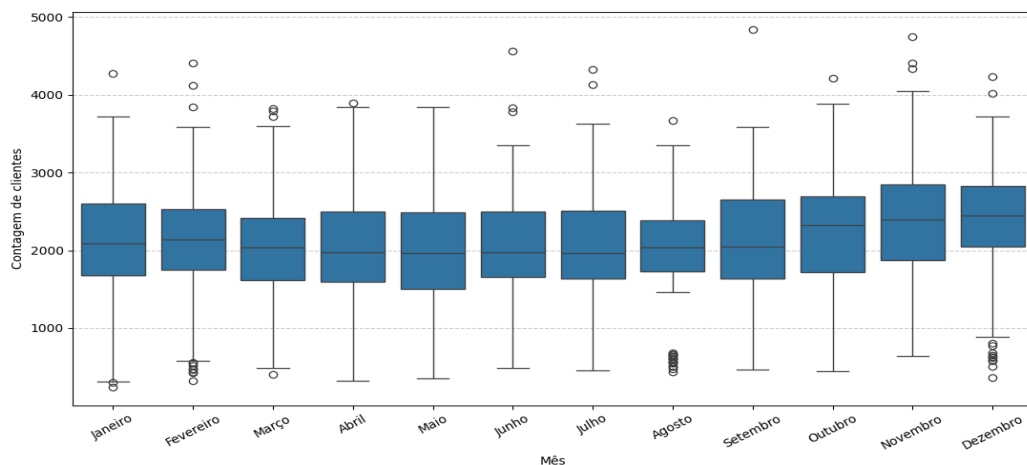
Figura 8 - Boxplot das vendas por mês



Fonte: Elaborado pelo autor (2025)

A Figura 9 mostra que a contagem de clientes ajustada também apresenta variação ao longo dos meses do ano, de forma semelhante ao padrão observado nas vendas. Embora as diferenças entre os meses não sejam muito acentuadas, percebe-se que alguns períodos registram fluxo ligeiramente maior, enquanto outros exibem níveis mais modestos. Em meses comercialmente mais fortes, nota-se uma dispersão um pouco maior, sugerindo picos de movimento associados a campanhas ou datas promocionais. Esses indícios reforçam que o calendário exerce influência sobre o fluxo de clientes e, por consequência, justificam a inclusão das variáveis de mês no conjunto de *features*.

Figura 9 - Boxplot da contagem de clientes por mês



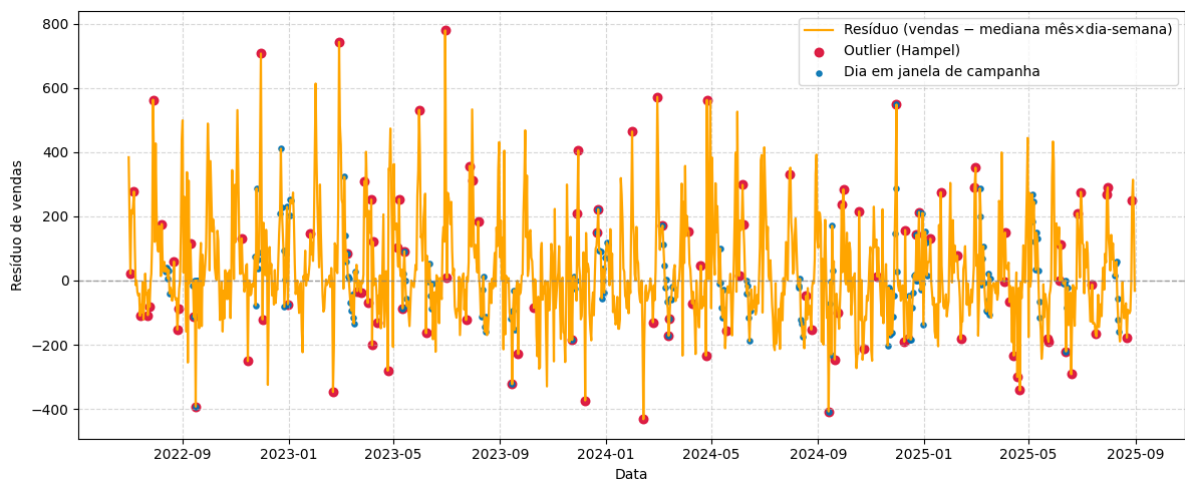
Fonte: Elaborado pelo autor (2025)

4.4.4 Análise de *outliers*

Com as variáveis sazonais criadas e as bases preparadas, seguimos com a análise de *outliers* para não comprometer as etapas de modelagem. Conforme discutido na Seção 2.5, séries temporais de varejo, picos e vales fora do padrão não são, necessariamente, erros. Com frequência refletem efeitos como campanhas, feriados e datas promocionais que precisam ser representados no modelo, e não simplesmente removidos.

Para visualizar onde estão os desvios mais fortes, foi construído um resíduo ajustado a calendário na base expandida. A Figura 10 apresenta a série de resíduos ao longo do tempo, com marcações para os pontos identificados como *outliers* e indicação dos dias que caem em janelas de campanha. Observa-se a concentração dos picos em períodos promocionais e datas especiais, reforçando que parte dos extremos representa o comportamento real do negócio.

Figura 10 - Resíduos de vendas com marcação de *outliers* e janelas de campanha



Fonte: Elaborado pelo autor (2025)

Para verificar se o tratamento de *outliers* influenciaria o desempenho dos modelos, aplicaram-se a Floresta Aleatória e o XGBoost em dois cenários: (A) com os dados íntegros, sem alterar valores extremos; (B) com tratamento apenas os casos classificados como erro provável, substituindo o resíduo pelo valor corrigido.

A avaliação manteve o recorte temporal 80% treino e 20% teste, avaliação usada na aplicação dos modelos no tópico posterior. Os resultados mostraram que

tratar os valores extremos piora a previsão, onde os erros aumentam em todas as bases e modelos. Isso indica que os picos e vales carregam informação útil para prever vendas e, portanto, devem ser preservados. Esses dados podem ser conferidos nos quadros posteriores.

O Quadro 7 mostra o desempenho dos modelos com os dados preservados. Os menores erros aparecem nesse cenário.

Quadro 7 - Cenário A - métricas por base e modelo

Base	Modelo	RMSE	MAE
compacta	Floresta Aleatória	47,97	36,06
compacta	XGBoost	48,44	37,02
expandida	XGBoost	45,84	34,29
expandida	Floresta Aleatória	47,38	34,89
extra-expandida	Floresta Aleatória	46,69	34,21
extra-expandida	XGBoost	47,60	34,95

Fonte: Elaborado pelo autor (2025)

O Quadro 8 reúne as métricas quando apenas os casos classificados como erro provável foram suavizados. Em todas as bases e modelos, os erros aumentaram de forma consistente, mostrando que o tratamento removeu informação relevante do histórico. Nesse cenário, o desempenho é inferior ao observado no Quadro 7.

Quadro 8 - Cenário B - métricas por base e modelo

Base	Modelo	RMSE	MAE
compacta	Floresta Aleatória	65,21	46,44
compacta	XGBoost	64,52	45,86
expandida	Floresta Aleatória	65,35	44,75
expandida	XGBoost	63,91	43,91
extra-expandida	Floresta Aleatória	63,72	43,49
extra-expandida	XGBoost	64,79	43,30

Fonte: Elaborado pelo autor (2025)

Dessa forma, a engenharia de atributos resultou na existência de três bases distintas de dados, ambos consistentes e prontos para análise preditiva. Essa estratégia permitirá, em etapas posteriores, comparar o desempenho dos modelos sob diferentes configurações de variáveis, possibilitando avaliar se a incorporação da sazonalidade semanal e mensal contribuem de maneira significativa para o aumento da acurácia das previsões de vendas.

4.5 APLICAÇÃO DOS MODELOS ARIMA, FLORESTA ALEATÓRIA E XGBOOST

Esta etapa descreve a implementação dos modelos ARIMA, Floresta Aleatória e XGBoost, tomando como base as estruturas de dados preparadas nas etapas anteriores. São apresentados os procedimentos de treinamento, validação e ajuste de cada modelo, bem como os critérios utilizados para garantir comparabilidade entre as abordagens.

4.5.1 Aplicação inicial e diagnóstico de sobreajuste

Neste tópico, passou-se à aplicação inicial dos modelos ARIMA, Floresta Aleatória e XGBoost nas três versões do conjunto de dados: base compacta, base expandida e base extra-expandida.

Vale ressaltar que o ARIMA foi incluído somente como referência comparativa. Esse tipo de modelo costuma funcionar bem quando a série é univariada ou quando há poucas variáveis externas bem sintetizadas. Como este estudo utiliza variáveis exógenas, entendidas como variáveis externas que influenciam a série alvo, optou-se pelo SARIMAX, que pertence à família ARIMA e permite incorporar esses efeitos adicionais. Nas bases deste estudo, especialmente na extra-expandida, com *dummies* de dias da semana e de meses, o número de parâmetros cresce e o SARIMAX tende a perder eficiência. Já os modelos em árvore lidam melhor com muitas variáveis e relações não lineares sem exigir mudanças na estrutura dos dados.

Para medir o desempenho dos modelos foi adotado um *holdout* temporal (desempenho com separação temporal) 80/20. Os dados foram ordenados por data, sendo 80% usados para treino e os 20% finais reservados para teste, onde o corte

ocorreu em 13/01/2025. Dessa forma, o conjunto de treino compreendeu o período de 01/07/2022 a 12/01/2025, enquanto a avaliação final ocorreu entre 13/01/2025 e 31/08/2025.

Foram calculados o RMSE e o R^2 nas etapas de treino e teste. O R^2 no conjunto de teste foi interpretado de forma preditiva, comparando o erro do modelo com o erro de um preditor ingênuo que estima a média dos valores observados no período avaliado, cujo RMSE foi de aproximadamente 193,72. Dessa forma, valores de R^2 próximos de zero indicam desempenho semelhante ao da média, valores positivos indicam ganho preditivo, e valores negativos indicam desempenho inferior a métrica de referência.

E como forma simples e transparente de identificar possíveis indícios de sobreajuste, utilizou-se a razão entre o erro de teste e o erro de treino. Conforme apresentado na Seção 3.1.2, a comparação entre o desempenho em dados vistos e não vistos é uma estratégia fundamental para avaliar a capacidade de generalização dos modelos. Neste estudo, valores próximos de 1 indicam equilíbrio entre as etapas, enquanto valores significativamente superiores sugerem que o modelo pode estar aprendendo padrões excessivamente específicos do conjunto de treino.

Para fins práticos, adotou-se o critério de que valores acima de 2 nessa razão caracterizam sobreajuste, uma vez que representam um caso em que o erro no conjunto não visto mais do que dobra em relação ao erro do treino, deixando pouca dúvida sobre a perda de generalização.

Como referência adicional, utilizou-se uma linha de base simples do tipo modelo ingênuo sazonal, que prevê o valor do dia ttt repetindo o valor observado no mesmo dia da semana anterior. Essa abordagem, também descrita na Seção 3.1.2, funciona como patamar mínimo de comparação, permitindo verificar se os modelos preditivos empregados oferecem ganho real de desempenho.

O Quadro 9 apresenta os resultados para a Floresta Aleatória. Na base compacta se observa bom desempenho no teste, com RMSE de 50,95 e R^2 de 0,9308. Ao incluir os dias da semana na base expandida, o RMSE de teste cai para 47,59 e o R^2 sobe para 0,9396. Com a inclusão dos meses na base extra-expandida, o RMSE atinge 47,47 e o R^2 0,9399. Em todas as configurações a lacuna de generalização permanece próximo de 1, variando entre 0,79 e 0,84, o que indica ausência de sobreajuste.

Quadro 9 - *Holdout* temporal 80/20 para o Floresta Aleatória

Base	RMSE de Treino	R ² de Treino	RMSE de Teste	R ² de Teste	Lacuna de generalização
compacta	64,31	0,9180	50,95	0,9308	0,79
expandida	58,21	0,9328	47,59	0,9396	0,82
extra-expandida	56,69	0,9363	47,47	0,9399	0,84

Fonte: Elaborado pelo autor (2025)

O Quadro 10 mostra o comportamento do XGBoost. Na base compacta, o teste apresenta RMSE de 55,44 e lacuna de generalização de 0,93, o que é aceitável. Com os dias da semana na base expandida, o RMSE de teste piora para 59,76 e a lacuna de generalização sobe para 1,83, sinal de tendência ao sobreajuste. Na base extra-expandida, a lacuna de generalização chega a 4,05. O treino fica quase perfeito, com RMSE de 13,94, enquanto o teste permanece em 56,54 e R² de 0,9148, isso caracteriza sobreajuste claro.

Quadro 10 - *Holdout* temporal 80/20 para o XGBoost

Base	RMSE de Treino	R ² de Treino	RMSE de Teste	R ² de Teste	Lacuna de generalização
compacta	59,71	0,9293	55,44	0,9181	0,93
expandida	32,60	0,9789	59,76	0,9048	1,83
extra-expandida	13,94	0,9961	56,54	0,9148	4,05

Fonte: Elaborado pelo autor (2025)

O Quadro 11 apresenta os resultados para o SARIMAX. Na base compacta observa-se desempenho com RMSE de 50,65 e R² de 0,9316, e lacuna de generalização de 0,75, indicando equilíbrio entre treino e avaliação. Ao incluir os *dummies* de dias da semana na base expandida, o desempenho cai, o RMSE de teste sobe para 92,91 e R² recua para 0,7700, com lacuna de generalização de 1,29. Na base extra-expandida, a especificação não se ajusta bem aos dados, resultando em RMSE de 233,73, R² de -0,4558 e lacuna de generalização de 3,34, sinal claro de inadequação do modelo, onde ele foi pior que a métrica de referência.

Quadro 11 - *Holdout* temporal 80/20 para o SARIMAX

Base	RMSE de Treino	R ² de Treino	RMSE de Teste	R ² de Teste	Lacuna de generalização
Compacta	67,33	0,9101	50,65	0,9316	0,75
Expandida	71,78	0,8979	92,91	0,7700	1,29
extra-expandida	69,94	0,9030	233,73	-0,4558	3,34

Fonte: Elaborado pelo autor (2025)

Para verificar se os modelos realmente superam um palpite semanal simples, o Quadro 12 compara o RMSE de teste de cada cenário com o RMSE do *lag-7*, em torno de 195,06 em todos os casos, e calcula o ganho percentual. Em termos práticos, a pergunta é quanto o modelo reduz de erro em relação a repetir o valor de sete dias atrás? O modelo Floresta Aleatória nas bases expandida e extra-expandida reduz o erro do *baseline*, ou seja, do modelo de referência em cerca de 73,88% e 75,60%, respectivamente. O XGBoost também supera o *lag-7* em todas as bases com ganhos de 71,58%, 69,36% e 71,01%, respectivamente. Já o SARIMAX tem dois comportamentos, na compacta fica com ganho de 74,05%, na expandida melhora menos com 52,37% e, na extra-expandida piora em relação ao próprio *lag-7* com -38,67%.

Quadro 12 - Comparação com a linha de base *lag-7*

Modelo	Base	RMSE de Teste (Modelo)	RMSE de Teste (<i>lag-7</i>)	Ganho
Floresta Aleatória	compacta	50,95	195,06	73,88
Floresta Aleatória	expandida	47,59	195,06	75,60
Floresta Aleatória	extra-expandida	47,47	195,06	75,66
XGBoost	compacta	55,44	195,06	71,58
XGBoost	expandida	59,76	195,06	69,36
XGBoost	extra-expandida	56,54	195,06	71,01
SARIMAX	compacta	50,65	195,06	74,05
SARIMAX	expandida	92,91	195,06	52,37
SARIMAX	extra-expandida	233,73	195,06	-38,67

Fonte: Elaborado pelo autor (2025)

Adicionalmente, registram-se os hiperparâmetros usados, para que o experimento possa ser reproduzido. Na Floresta Aleatória, *n_estimators* é a quantidade de árvores, *max_depth* limita a profundidade das árvores e *min_samples_leaf* define o mínimo de registros em cada folha, o que suaviza

decisões. No XGBoost, *n_estimators* e *learning_rate* controlam o ritmo do aprendizado, *max_depth* define a complexidade, *subsample* e *colsample_bytree* fazem amostragem de linhas e colunas para melhorar a generalização, e *reg_lambda* (L2) e *reg_alpha* (L1) aplicam regularização.

No SARIMAX, a notação Ordem (p,d,q) descreve a parte não sazonal do modelo, onde *p* é o número de termos autorregressivos, *d* é o número de diferenças aplicadas na série para estabilizar a tendência, e *q* é o número de termos de média móvel. A notação Sazonal (P,D,Q,s) tem o mesmo significado, mas no ciclo sazonal, na qual *P* é o autorregressivo sazonal, *D* é a diferença sazonal, *Q* é a média móvel sazonal e *s* é o tamanho do período. Os parâmetros usados em cada modelo podem ser vistos no Quadro 13:

Quadro 13- Hiperparâmetros utilizados

Modelo	Bases	Hiperparâmetros
Floresta Aleatória	compacta, expandida e extra-expandida	n_estimators = 400 max_depth = 8 min_samples_leaf=5 random_state=42 n_jobs=-1
XGBoost	compacta, expandida e extra-expandida	n_estimators=500 learning_rate=0,10 max_depth=6 subsample=0,9 colsample_bytree=0,9 reg_lambda=1,0 reg_alpha=0,0 random_state=42 n_jobs=-1
SARIMAX	compacta, expandida e extra-expandida	Ordem: (p,d,q)=(1,1,1) Sazonal (P,D,Q,s)= (1,1,1,7)

Fonte: Elaborado pelo autor (2025)

4.5.2 Regularização do modelo XGBoost na base extra-expandida

Depois de aplicar os modelos, observou-se que o XGBoost apresentou sobreajuste na base extra-expandida, onde o erro no treino ficou muito baixo e o erro no teste permaneceu alto, sinal de que o modelo estava “decorando” o histórico. Para resolver isso, foi feita uma regularização do XGBoost, mantendo a lógica de avaliação por tempo. Os dados foram ordenados por data, os 80% mais antigos serviram para treino e os 20% mais recentes ficaram para teste.

Dentro do próprio conjunto de treino, separou-se aproximadamente os 10% finais para funcionar como validação temporal. Essa divisão permitiu monitorar o desempenho do modelo conforme a data avançava. Com essa parte reservada, tornou-se possível ajustar o número de árvores usando *early stopping*, um mecanismo que interrompe o treinamento quando o desempenho na validação deixa de evoluir. A partir desse processo, foi selecionado um conjunto de hiperparâmetros mais conservador, reduzindo o risco de sobreajuste observado nas primeiras execuções.

A regularização seguiu três frentes complementares. Primeiro, reduziu-se a complexidade das árvores, limitando a profundidade com *max_depth* igual a 4 e exigindo mais dados por divisão com *min_child_weight* igual a 10. Segundo, aplicou-se amostragem por linhas e por colunas em cada árvore com *subsample* de 0,8 e *colsample_bytree* de 0,7, o que força diversidade entre as árvores e reduz a chance de o modelo se ajustar demais a padrões locais. Terceiro, fortaleceu-se a penalização de complexidade com regularização L2 e L1 ambos de 0,5 e um patamar mínimo de ganho para dividir nós com *gamma* igual a 0,1. O passo de aprendizado foi mantido moderado com *learning_rate* igual a 0,03 e o número efetivo de árvores foi definido automaticamente pelo *early stopping*, interrompendo o treinamento quando a validação deixou de melhorar, o melhor ponto encontrado foi por volta de 408 árvores.

Depois de ajustar os hiperparâmetros com validação temporal dentro do treino, o modelo final foi re-treinado reunindo o treino e a validação, sempre sem tocar no conjunto de teste e, em seguida, avaliado no bloco mais recente. O efeito prático foi justamente o que se buscava, o XGBoost saiu do sobreajuste e passou a apresentar desempenho muito próximo ao da Floresta Aleatória, com equilíbrio entre treino e teste. O Quadro 14 mostra a comparação do XGBoost, antes e depois da regularização.

Quadro 14 - XGBoost na base extra-expandida: antes e depois da regularização

Situação	RMSE de Treino	R ² de Treino	RMSE de Teste	R ² de Teste	Lacuna de generalização
Antes	13,94	0,9961	56,54	0,9148	4,05
Depois	56,32	0,9371	47,60	0,9396	0,85

Fonte: Elaborado pelo autor (2025)

Os números mostram que a regularização cumpriu seu papel. O erro em teste caiu de 56,54 para 47,60, o R^2 subiu de 0,9148 para 0,9396 e a lacuna de generalização passou de 4,05, claro sinal de sobreajuste, para 0,85 que indica equilíbrio. Em termos práticos, o XGBoost regularizado ficou empatado tecnicamente com a Floresta Aleatória na mesma base, ambos muito acima do *lag-7*.

Vale a ressaltar porquê essas escolhas funcionarem. Árvores mais rasas e com exigência mínima de amostras em cada divisão evitam regras muito específicas que só valem para o passado. A amostragem de linhas e colunas injeta variedade e diminui a dependência de um subconjunto de variáveis. A regularização L1 e L2 penaliza soluções excessivamente complexas. E o *early stopping* impede que o modelo continue forçando melhorias no treino quando a validação já parou de melhorar. Juntos, esses mecanismos deslocam o foco do modelo do ajuste perfeito para a generalização no período mais recente.

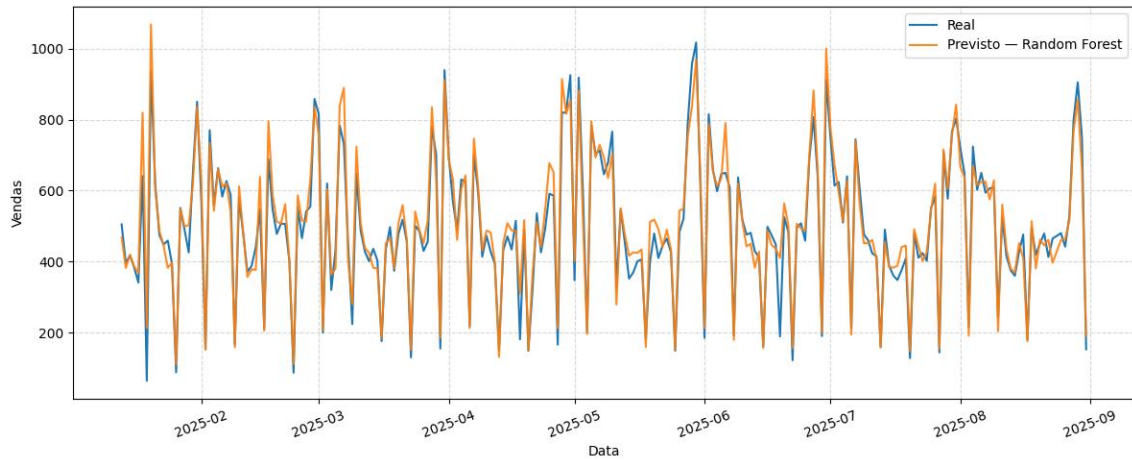
Essas escolhas de regularização só fazem sentido se, na prática, melhorarem a capacidade de generalizar. Para tornar esse efeito visível e não apenas numérico, são apresentadas, no mesmo período de teste e sobre a base extra-expandida, as curvas de vendas reais e previstas para cada modelo. Se os mecanismos discutidos acima realmente funcionaram, as previsões devem acompanhar a série real ao longo do tempo, sem descolar nos picos e vales semanais.

A Figura 11 mostra que a Floresta Aleatória acompanha bem a série real no período reservado ao teste. Os picos e vales semanais são capturados com precisão visual. O desempenho onde RMSE é aproximadamente 47,47 e R^2 é 0,94 confirma a boa aderência observada.

A Figura 12 mostra que após a regularização e o uso de *early stopping*, o XGBoost passa a reproduzir o padrão temporal com qualidade semelhante à Floresta Aleatória, mantendo alinhamento nos principais picos e quedas. O RMSE foi de aproximadamente 47,60, e o R^2 de 0,94, evidenciando o empate técnico entre os modelos e confirmando que o sobreajuste foi mitigado.

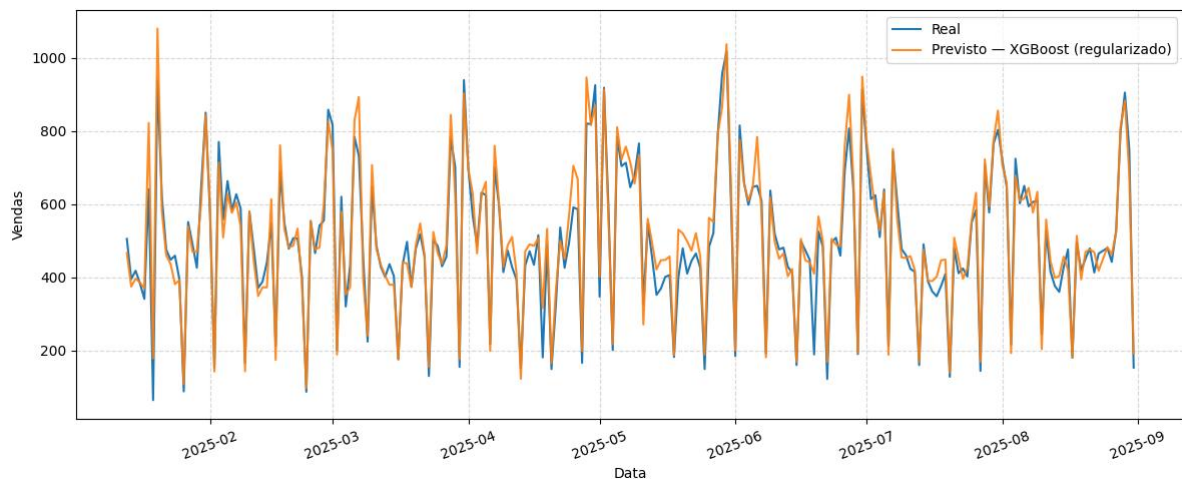
Contudo, observou-se que o SARIMAX não replicou o bom desempenho obtido pelos modelos em árvore na base extra-expandida, possivelmente pela maior quantidade de *dummies* e variáveis externas. O tratamento específico, seleção de ordens, verificação de transformações e demais ajustes é desenvolvido posteriormente, quando o SARIMAX é otimizado e reavaliado no mesmo protocolo.

Figura 11 - Comportamento das vendas reais e previstas pela Floresta Aleatória na base extra-expandida



Fonte: Elaborado pelo autor (2025)

Figura 12 - Relação entre valores reais e previstos pelo modelo XGBoost base extra-expandida



Fonte: Elaborado pelo autor (2025)

4.5.3 Aplicação do XGBoost regularizado e ajuste simétrico da Floresta Aleatória

Como o XGBoost deixou de apresentar sobreajuste após a regularização na base extra-expandida e passou a ter desempenho equivalente ao da Floresta Aleatória, optou-se por aplicar o mesmo protocolo de regularização a todas as bases. Mantiveram-se os dados ordenados por data, dividindo-se 80% para treino e 20% para teste e, dentro do treino, foi reservado os 10% finais como validação temporal para

acionar o *early stopping*. A busca de hiperparâmetros foi deliberadamente conservadora, deixando o *early stopping* fixar o número efetivo de árvores por base. Com isso, o XGBoost regularizado superou a Floresta Aleatória nas bases compacta e expandida e ficou empatado tecnicamente na extra-expandida com uma diferença de RMSE de teste pequena. O Quadro 15 resume esse primeiro contraste, ainda com a Floresta Aleatória na configuração antiga.

Quadro 15 - Comparação entre XGBoost regularizado e Floresta Aleatória

Base	Modelo	RMSE de Treino	R ² treino	RMSE de Teste	R ² de Teste	Lacuna de generalização	Ganho
compacta	Floresta Aleatória	64,31	0,9180	50,95	0,9308	0,79	73,88
compacta	XGBoost regularizado	69,71	0,9037	48,44	0,9375	0,69	75,17
expandida	Floresta Aleatória	58,21	0,9328	47,59	0,9396	0,82	75,60
expandida	XGBoost regularizado	64,03	0,9187	45,84	0,9440	0,72	76,50
extra-expandida	Floresta Aleatória	56,69	0,9363	47,47	0,9399	0,84	75,66
extra-expandida	XGBoost regularizado	56,32	0,9371	47,60	0,9396	0,85	75,60

Fonte: Elaborado pelo autor (2025)

Para que a comparação ficasse justa, aplicou-se à Floresta Aleatória exatamente o mesmo protocolo temporal usado no XGBoost, divisão 80/20 por data, janela de validação temporal com 10% finais do treino para seleção de hiperparâmetros e re-treino no conjunto de treino mais validação, antes de avaliar no teste. Foram ajustados os componentes que efetivamente controlam viés e variância no Floresta Aleatória, o número de árvores (*n_estimators*), profundidade máxima (*max_depth*), tamanho mínimo de folha (*min_samples_leaf*) e fração de variáveis por árvore (*max_features*), mantendo *bootstrap* (amostragem aleatória com reposição usada para treinar cada árvore em um subconjunto diferente dos dados) ativado. O Quadro 16 registra os parâmetros escolhidos por base para ambos os modelos, evidenciando o que mudou e porquê: árvores mais numerosas e um pouco mais

profundas na Floresta Aleatória para estabilizar a variância, XGBoost com árvores rasas, amostragem e regularização para manter a simplicidade.

Para que a comparação fosse justa, a Floresta Aleatória seguiu exatamente o mesmo protocolo temporal aplicado ao XGBoost: divisão 80/20 por data, uso dos 10% finais do conjunto de treino para validação e escolha dos hiperparâmetros, e re-treino final utilizando treino mais validação antes da avaliação no teste.

No caso da Floresta Aleatória, foram ajustados apenas os hiperparâmetros que realmente controlam viés e variância: número de árvores (*n_estimators*), profundidade máxima (*max_depth*), tamanho mínimo de folha (*min_samples_leaf*) e fração de variáveis considerada por árvore (*max_features*). O *bootstrap* permaneceu ativado, garantindo que cada árvore fosse treinada em um subconjunto diferente dos dados.

O Quadro 16 apresenta os parâmetros escolhidos para cada base e mostra claramente o motivo das diferenças: a Floresta Aleatória utilizou mais árvores e um pouco mais de profundidade para estabilizar a variância, enquanto o XGBoost operou com árvores rasas, amostragem e regularização para manter o modelo mais simples e generalizável.

Quadro 16 - Hiperparâmetros escolhidos por base (validação temporal nos 10% finais do treino)

Base	Floresta Aleatória	XGBoost regularizado
compacta	n_estimators=600; max_depth=10; min_samples_leaf=4; max_features=0,7; bootstrap=True	learning_rate=0,05; max_depth=4; min_child_weight=12; subsample=0,8; colsample_bytree=0,7; reg_lambda=8,0; reg_alpha=1,0; gamma=0,1; n_estimators≈130
expandida	n_estimators=600; max_depth=10; min_samples_leaf=4; max_features=0,7; bootstrap=True	learning_rate=0,03; max_depth=4; min_child_weight=10; subsample=0,8; colsample_bytree=0,7; reg_lambda=5,0; reg_alpha=0,5; gamma=0,1; n_estimators≈168
extra-expandida	n_estimators=900; max_depth=12; min_samples_leaf=3; max_features=0,6; bootstrap=True	learning_rate=0,03; max_depth=4; min_child_weight=10; subsample=0,8; colsample_bytree=0,7; reg_lambda=5,0; reg_alpha=0,5; gamma=0,1; n_estimators≈408

Fonte: Elaborado pelo autor (2025)

Com a mesma separação temporal, mesma janela de validação e faixa de busca semelhante, a performance dos modelos ficou equilibrada, onde a Floresta

Aleatória foi melhor nas bases compacta e extra-expandida, enquanto o XGBoost manteve a liderança na expandida. O Quadro 17 apresenta as duas métricas utilizadas na comparação dos modelos: o RMSE de teste e o ganho em relação ao *lag-7*.

Quadro 17 - Desempenho no período de avaliação 80/20 com RMSE e ganho em relação ao *lag-7*

Base	Modelo	RMSE de Teste	Ganho
compacta	Floresta Aleatória	47,97	75,41
compacta	XGBoost	48,44	75,17
expandida	Floresta Aleatória	47,38	75,71
expandida	XGBoost	45,84	76,50
extra-expandida	Floresta Aleatória	46,69	76,07
extra-expandida	XGBoost	47,60	75,60

Fonte: Elaborado pelo autor (2025)

Contudo, quando foi regularizado, o XGBoost superou a Floresta Aleatória em duas bases e empatou tecnicamente na terceira. Ao ajustar também a Floresta Aleatória com o mesmo desenho temporal, os resultados mostram liderança dividida: Floresta Aleatória melhor nas bases compacta e extra-expandida e XGBoost melhor na expandida. Em todas as bases, os ganhos sobre o *lag-7* confirmam que ambos os modelos se generalizam bem. A comparação é considerada justa porque os dois algoritmos foram avaliados sob idênticas regras de separação temporal, mesma janela de validação para escolher hiperparâmetros e faixa equivalente de busca, assim, qualquer diferença de desempenho reflete propriedades dos modelos e das bases, e não vantagens de procedimento.

4.5.4 Otimização do SARIMAX como referência comparativa

Para promover uma comparação justa com os outros modelos e corrigir os resultados para a base extra-expandida, o SARIMAX foi avaliado com seleção de ordens por validação *walk-forward* (deslizante no tempo), dentro do conjunto de treino.

O procedimento escolheu, por base, a combinação de parâmetros com menor erro médio de validação e, em seguida, o modelo foi re-treinado no treino completo e medido no período de avaliação.

O resultado aparece estável nas bases compacta e expandida, com RMSE de 50,71, com R^2 de 0,9315 e 53,11, com R^2 de 0,9248, respectivamente, e lacuna de generalização abaixo de 1, indicando bom equilíbrio. Enquanto na extra-expandida, mesmo após a otimização, o erro ficou em 77,56, com R^2 de 0,8397, mostrando que, com muitos *dummies* sazonais, o modelo tende a perder eficiência frente aos modelos de inteligência computacional. Ainda assim, nas três bases, o SARIMAX fica muito acima da linha de base semanal, com ganhos na faixa de 60% a 74%. O SARIMAX otimizado cumpre bem o papel de referência clássica, aproxima-se dos modelos em árvore nas bases com menos variáveis e se distancia quando a base incorpora mais sinais sazonais.

O Quadro 18 resume, por base, apenas as escolhas de configuração do SARIMAX: a ordem não sazonal, a componente sazonal semanal e a informação de transformação do alvo, neste caso, sem log.

Quadro 18 - Configurações do SARIMAX por base

Base	Ordem (p,d,q)	Sazonal (P,D,Q,s)	Transformação
compacta	(1, 0, 1)	(1, 1, 1, 7)	sem log
expandida	(1, 0, 1)	(0, 1, 1, 7)	sem log
extra-expandida	(1, 0, 1)	(1, 1, 1, 7)	sem log

Fonte: Elaborado pelo autor (2025)

O Quadro 19 apresenta as métricas de desempenho obtidas com essas configurações no recorte temporal de 80% para treino e 20% para avaliação, os RMSE e R^2 , lacuna de generalização e o ganho percentual frente ao *lag-7*.

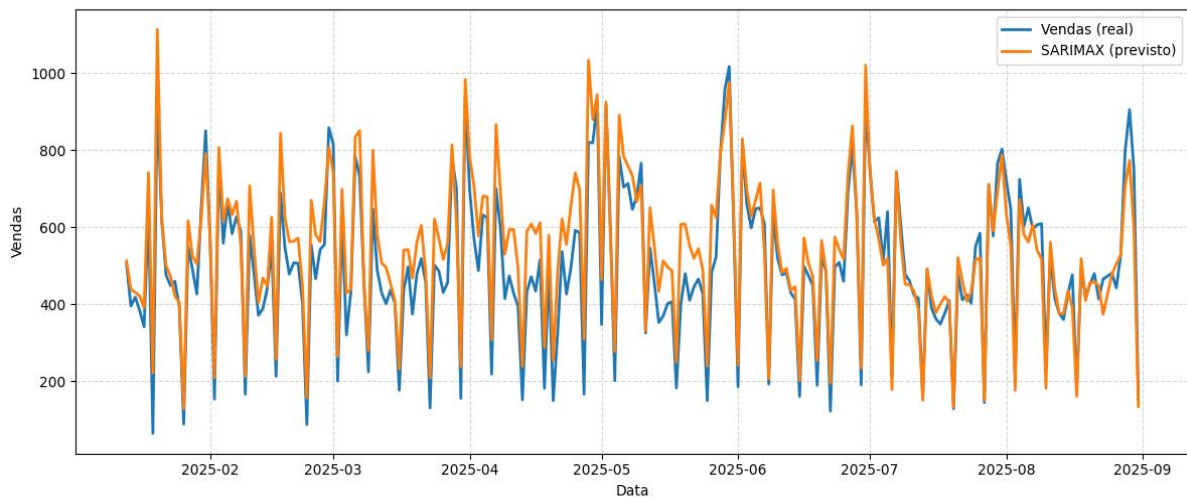
Quadro 19 - Desempenho do SARIMAX por base (80/20)

Base	RMSE de Treino	R^2 de Treino	RMSE da Avaliação	R^2 da Avaliação	Lacuna de generalização	Ganho
compacta	66,61	0,9120	50,71	0,9315	0,76	74,00
expandida	66,38	0,9126	53,11	0,9248	0,80	72,78
extra-expandida	67,10	0,9107	77,56	0,8397	1,16	60,24

Fonte: Elaborado pelo autor (2025)

A Figura 13 mostra que o SARIMAX acompanha a série real no período de teste, capturando a sazonalidade semanal, mas com maior erro nos picos e vales mais acentuados. O desempenho com RMSE de 77,56 e R^2 de 0,84, indica boa coerência geral, embora inferior aos modelos em árvore na mesma base.

Figura 13 - Relação entre valores reais e previstos no teste pelo modelo SARIMAX na base extra-expandida



Fonte: Elaborado pelo autor (2025)

4.6 APLICAÇÃO DA TÉCNICA DE VALIDAÇÃO CRUZADA NOS MODELOS

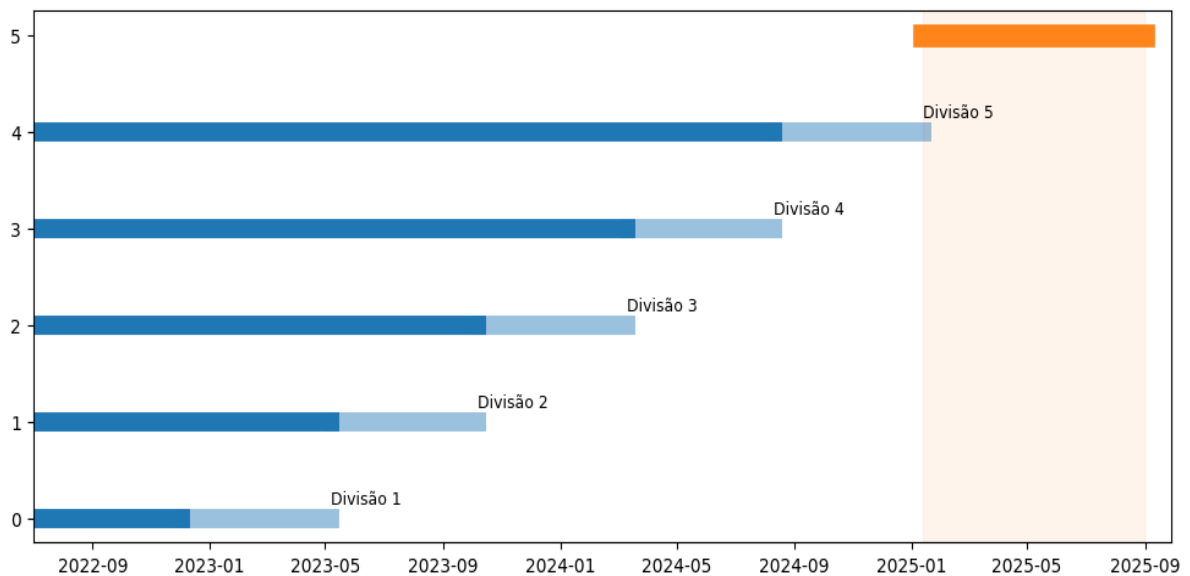
Após a avaliação com divisão temporal de 80% para treino e 20% para teste, aplicou-se validação cruzada temporal do tipo *walk-forward*, implementada com *TimeSeriesSplit*, com janelas crescentes de cinco divisões, sem embaralhamento. Nesse desenho base, os blocos de treino são cumulativos e não há janela fixa de validação, em cada divisão, o modelo aprende com um período inicial e é testado imediatamente adiante, depois o histórico de treino é ampliado e o processo se repete. Como descrito na Seção 3.1.2, esse esquema preserva a ordem cronológica dos dados e evita o vazamento de informações futuras para o passado, característica essencial em séries temporais com dependência temporal e sazonalidade.

Na prática, o modelo aprende com um pedaço inicial do histórico e é testado no período imediatamente posterior, em seguida, ampliamos o histórico e repetimos o processo em cinco rodadas, sempre preservando a ordem das datas.

Na primeira execução dessa técnica, optou-se por não definir janelas específicas de tempo, mantendo apenas cinco divisões. Essa configuração inicial foi utilizada como uma forma de validar a metodologia *walk-forward* em sua estrutura mais simples, permitindo observar como os modelos se comportavam ao longo de diferentes períodos sem ajustes adicionais. Essa configuração permitiu avaliar se a variação dos resultados entre as divisões apontava para a necessidade de ajustar o tamanho das janelas.

A Figura 14 mostra como a validação cruzada temporal foi organizada. Cada linha azul representa uma divisão, o trecho azul mais escuro à esquerda é o período usado para treino, e o segmento azul mais claro à direita é a validação daquele corte. As cinco linhas indicam que o processo foi repetido cinco vezes, sempre avançando no tempo e preservando a ordem das datas, esquema *walk-forward* com janelas crescentes. À direita, a barra laranja destaca o intervalo efetivamente utilizado para mensurar o desempenho final. Para fins de ilustração, a figura foi construída com a base expandida, as demais bases seguem exatamente o mesmo padrão, variando apenas as datas específicas de cada divisão.

Figura 14 - Esquema da validação cruzada temporal *walk-forward*



Fonte: Elaborado pelo autor (2025)

O Quadro 20 apresenta as métricas médias e seus respectivos desvios-padrão obtidos nessa etapa inicial da validação cruzada. Nela, observam-se os valores de RMSE, R^2 , e a razão entre o erro de validação e o erro de treino, funcionando como

uma medida da lacuna de generalização dentro do processo de validação cruzada. O desvio-padrão representa a variação dessas métricas entre as divisões, indicando o quanto o desempenho do modelo oscilou de um período para outro.

Quadro 20 - Resultados da validação cruzada temporal sem definição de janelas

Base	Modelo	RMSE_média	RMSE_dp	R ² _média	R ² _dp	Razão_val/treino_média
Compacta	Floresta Aleatória	80,90	20,88	0,8622	0,0624	0,95
Compacta	XGBoost	73,36	15,50	0,8881	0,0399	0,93
Compacta	SARIMAX	72,32	14,67	0,8913	0,0375	0,95
Expandida	Floresta Aleatória	69,43	14,57	0,8994	0,0363	1,06
Expandida	XGBoost	70,14	15,76	0,8971	0,0397	0,98
Expandida	SARIMAX	85,43	30,88	0,8388	0,1111	1,08
extra-expandida	Floresta Aleatória	67,74	14,98	0,9039	0,0372	1,15
extra-expandida	XGBoost	69,69	17,14	0,8974	0,0449	1,14
extra-expandida	SARIMAX	139,57	52,92	0,5483	0,3376	1,66

Fonte: Elaborado pelo autor (2025)

Os resultados do Quadro 20 mostraram que os modelos apresentaram bom poder explicativo, com valores de R² acima de 0,85 na maioria dos casos. Contudo, o alto desvio-padrão das métricas revelou instabilidade entre as divisões, indicando que os erros variavam consideravelmente conforme o período analisado. Essa oscilação reforçou a necessidade de ajustar o tamanho das janelas de treino e validação, de modo a capturar, de forma mais precisa, as variações temporais nas séries de vendas.

Diante disso, optou-se por testar diferentes tamanhos de janelas móveis e números de divisões, avaliando janelas de 7, 15 e 30 dias, e divisões variando entre 5 e 8 rodadas. Essa exploração permitiu encontrar o melhor equilíbrio entre representatividade temporal e estabilidade das métricas. A configuração com janelas de 15 dias apresentou o melhor comportamento geral, resultando em menor variação dos erros e maior coerência entre as rodadas.

Nessa etapa, foram mantidas 5 divisões para os modelos de Floresta Aleatória e XGBoost, e 8 divisões para o modelo SARIMAX, em razão de sua estrutura autorregressiva e maior sensibilidade a períodos curtos. Os resultados consolidados dessa configuração são apresentados no Quadro 21, que resume as métricas obtidas para cada base e modelo, considerando a média e o desvio-padrão entre as divisões.

Quadro 21 - Resultados médios da validação cruzada temporal com janelas de 15 dias

Base	Modelo	k	RMSE_média	RMSE_dp	MAE_média	MAE_dp	R ² _média	Razão_RMS_E_val/treino	Razão_MAE_val/treino
compacta	Floresta Aleatória	5	46,25	5,95	35,27	7,45	0,9480	0,74	0,80
compacta	XGBoost	5	49,25	8,06	34,19	4,48	0,9445	0,82	0,81
compacta	SARIMAX	8	59,27	15,25	47,60	14,70	0,8987	1,00	1,12
expandida	Floresta Aleatória	5	43,78	11,76	31,10	6,13	0,9533	0,87	0,90
expandida	XGBoost	5	44,95	10,10	33,29	5,64	0,9513	0,79	0,84
expandida	SARIMAX	8	58,70	13,22	47,72	11,83	0,9049	0,99	1,12
extra-expandida	Floresta Aleatória	5	44,05	9,64	31,84	5,42	0,9535	0,96	1,03
extra-expandida	XGBoost	5	45,47	9,96	33,51	6,69	0,9487	0,96	1,03
extra-expandida	SARIMAX	8	62,89	9,46	51,55	8,36	0,8792	1,07	1,20

Fonte: Elaborado pelo autor (2025)

Os resultados do Quadro 21 indicam que o uso de janelas móveis de 15 dias proporcionou maior estabilidade entre as rodadas e uma melhora considerável nas métricas de desempenho. Os modelos baseados em árvores, apresentaram valores de R² superiores a 0,94 em todas as bases, com erros médios reduzidos e coerentes entre as divisões. O modelo SARIMAX, mesmo com comportamento mais variável, também manteve resultados satisfatórios dentro da configuração adotada.

Além disso, as razões entre os erros de validação e treino se mantiveram próximas de 1, o que indica coerência entre as etapas e ausência de sobreajuste relevante. O baixo desvio-padrão entre as divisões reforça a consistência temporal dos modelos, mostrando que o desempenho se manteve estável ao longo dos diferentes períodos avaliados. Esses resultados confirmam que o ajuste das janelas de 15 dias contribuiu para uma validação mais robusta, com métricas equilibradas e boa capacidade explicativa nas três bases de dados.

Por fim, a aplicação da validação cruzada temporal se mostrou essencial para garantir a confiabilidade das estimativas de desempenho e para verificar a estabilidade dos modelos em diferentes recortes do tempo. O método permitiu identificar eventuais variações no comportamento preditivo e demonstrou que os ajustes realizados, tanto na escolha das janelas quanto na definição das divisões, resultaram em métricas consistentes e com boa capacidade de generalização. Dessa

forma, a etapa de validação cumpriu o papel de consolidar o processo de modelagem, fornecendo uma base sólida para a comparação final dos resultados e para a escolha dos modelos mais adequados ao problema de previsão de vendas.

5 ANÁLISE DOS MODELOS

Com a etapa de modelagem concluída e os resultados da validação cruzada consolidados, esta seção apresenta a análise detalhada do desempenho dos modelos aplicados à previsão de vendas. O objetivo é interpretar os resultados obtidos, destacando como cada abordagem se comportou diante das diferentes bases de dados e configurações temporais utilizadas. A análise busca compreender não apenas qual modelo apresentou as melhores métricas, mas também quais deles demonstraram maior estabilidade e capacidade de generalização ao longo do tempo.

Para isso, são discutidos separadamente o comportamento dos modelos Floresta Aleatória, XGBoost e SARIMAX, considerando os resultados observados tanto na etapa de avaliação 80/20 quanto nas validações cruzadas temporais. Em seguida, realiza-se uma comparação direta entre os modelos, a fim de identificar padrões de desempenho e evidenciar as vantagens e limitações de cada abordagem no contexto proposto. Por fim, os resultados são discutidos de forma crítica, relacionando-os às características das séries de vendas.

5.1 AVALIAÇÃO DOS MODELOS

Nesta etapa, são apresentados os resultados dos três modelos aplicados, Floresta Aleatória, XGBoost e SARIMAX, considerando o comportamento observado nas diferentes bases e nas duas etapas principais de validação: a avaliação com divisão 80/20 e a validação cruzada temporal com janelas de 15 dias. Essa análise permite compreender a capacidade preditiva e a estabilidade temporal de cada modelo, além de verificar como se adaptaram às características das séries de vendas.

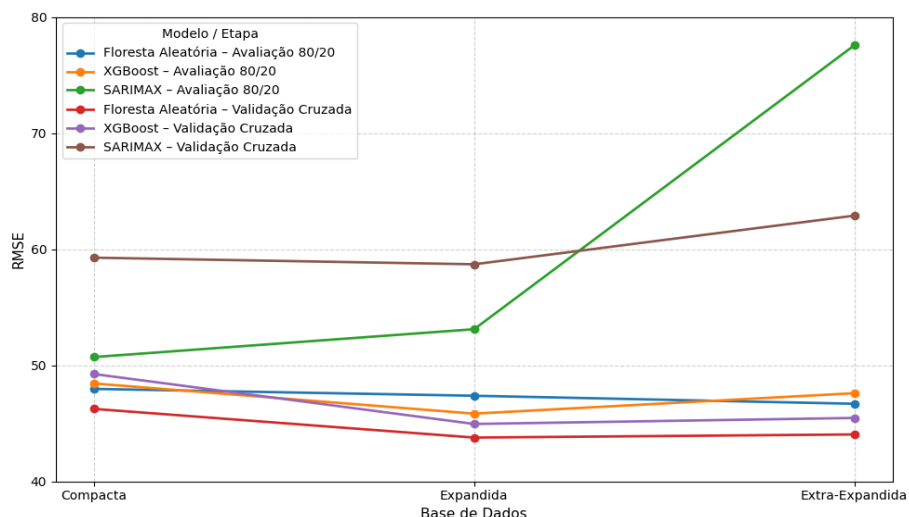
Na avaliação inicial 80/20, a Floresta Aleatória apresentou resultados consistentes, com RMSE entre 47 e 48, e ganhos superiores a 75% em relação ao modelo de referência *lag-7*, (Quadro 17). Em seguida, na validação cruzada temporal, o modelo manteve desempenho estável, como mostra o Quadro 21, com valores de R^2 acima de 0,94, baixa variação entre as divisões, e RMSE médio entre 43 e 46 nas bases expandida e extra-expandida. Essa estabilidade reforça sua capacidade de generalização, mesmo em cenários com sazonalidade e ruído.

No XGBoost, o padrão foi semelhante. Na avaliação 80/20, o algoritmo registrou RMSE entre 45 e 48, confirmando sua boa adaptação aos diferentes níveis de detalhe das bases de dados. Já na validação cruzada temporal, o Quadro 21 mostra R^2 superiores a 0,94 nas três bases e RMSE entre 44 e 49, com baixa dispersão entre as divisões. A regularização aplicada nas etapas anteriores reduziu o sobreajuste observado nas primeiras execuções, garantindo desempenho equilibrado e previsões mais robustas.

O SARIMAX, embora mais simples que os modelos de inteligência computacional, apresentou comportamento coerente. Na avaliação 80/20 (Quadro 19), os valores de R^2 ficaram entre 0,84 e 0,93, enquanto na validação cruzada temporal (Quadro 21) variaram entre 0,88 e 0,90. Contudo, houve maior variabilidade entre as divisões e aumento dos erros na base extra-expandida, com RMSE chegando a 77,56. Isso sugere que, embora capture bem padrões sazonais semanais, o modelo perde eficiência quando enfrenta muitas variáveis exógenas e relações não lineares.

A Figura 15 apresenta, de forma comparativa, os valores de RMSE obtidos na avaliação 80/20 e na validação cruzada com janelas de 15 dias. Observa-se que os modelos em árvore, mantêm desempenho estável entre as duas etapas, com pequenas variações nas três bases. Já o SARIMAX apresenta melhora moderada após a validação cruzada, embora continue com erros mais elevados, especialmente na base extra-expandida. O gráfico evidencia a consistência dos modelos de inteligência computacional e destaca a sensibilidade maior do SARIMAX a bases com maior complexidade.

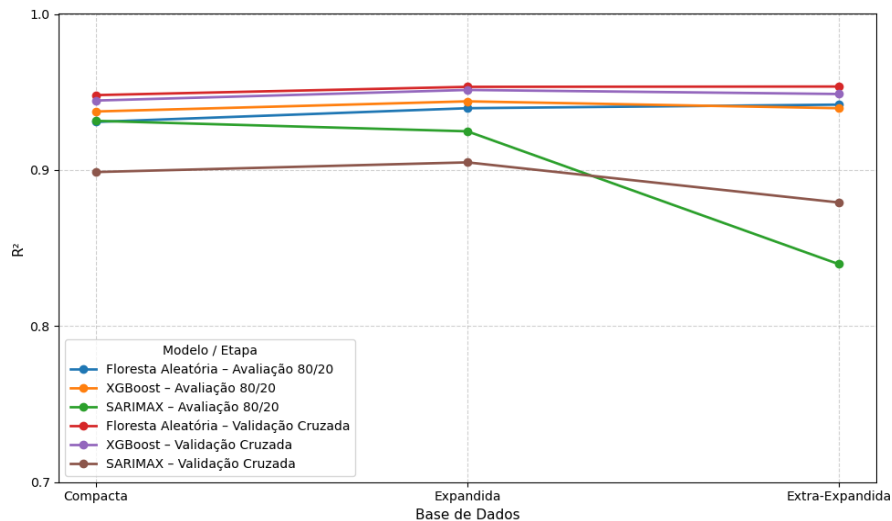
Figura 15 - RMSE dos modelos antes e depois da validação cruzada



Fonte: Elaborado pelo autor (2025)

A Figura 16 mostra a evolução dos valores de R^2 na avaliação 80/20 e na validação cruzada com janelas de 15 dias. Nota-se que Floresta Aleatória e XGBoost mantêm coeficientes elevados em todas as bases, indicando forte capacidade explicativa e estabilidade temporal. O SARIMAX apresenta desempenho competitivo nas bases compacta e expandida, mas perde aderência na extra-expandida. O gráfico reforça a superioridade dos modelos em árvore na captura dos comportamentos não lineares da série.

Figura 16 - R^2 dos modelos antes e depois da validação cruzada



Fonte: Elaborado pelo autor (2025)

Além dos resultados obtidos com RMSE e R^2 , o MAE também reforça o comportamento observado entre os modelos tanto no teste 80/20 quanto na validação cruzada, essas métricas podem ser vistas no Quadro 22. No conjunto 80/20, a Floresta Aleatória manteve erros absolutos baixos, variando entre 34 e 36 unidades nas três bases, enquanto o XGBoost apresentou valores muito próximos, entre 34 e 37. Na validação cruzada com janelas de 15 dias, esse padrão se repete, onde a Floresta Aleatória permanece com os menores erros absolutos, com MAE entre 31 e 35, seguida de perto pelo XGBoost, que ficou na faixa de 33 a 34. Enquanto o SARIMAX apresentou oscilações mais acentuadas nas duas etapas, com MAE acima de 38 nas bases compacta e expandida e valores superiores a 50 na validação da extra-expandida, além de alcançar 62,67 no teste dessa mesma base. Esses resultados mostram que, mesmo nos cenários mais desafiadores, os modelos de inteligência computacional baseados em árvores conseguiram manter erros menores e mais

estáveis, reforçando sua vantagem na hora de capturar as relações não lineares presentes na série de vendas.

Quadro 22 - Valores de MAE nos testes 80/20 e na validação cruzada com janelas de 15 dias

Base	Modelo	MAE Teste 80/20	MAE Validação
Compacta	Floresta Aleatória	36,06	35,27
Compacta	XGBoost	37,02	34,19
Compacta	SARIMAX	38,82	47,60
Expandida	Floresta Aleatória	34,89	31,10
Expandida	XGBoost	34,29	33,29
Expandida	SARIMAX	40,49	47,72
Extra-expandida	Floresta Aleatória	34,21	31,84
Extra-expandida	XGBoost	34,95	33,51
Extra-expandida	SARIMAX	62,67	51,55

Fonte: Elaborado pelo autor (2025)

5.2 COMPARAÇÃO DOS MODELOS APLICADOS

A comparação entre os modelos Floresta Aleatória, XGBoost e SARIMAX revela diferenças importantes não só nos números, mas também na forma como cada modelo se comporta ao longo do tempo e se adapta às características das bases utilizadas. As métricas principais RMSE, MAE e R^2 oferecem uma avaliação direta da qualidade das previsões, enquanto indicadores complementares, como a lacuna de generalização ajudam a entender melhor a estabilidade dos modelos e a identificar possíveis sinais de sobreajuste.

De maneira geral, os modelos de inteligência computacional baseados em árvores, apresentaram desempenho superior ao SARIMAX, mas a disputa entre Floresta Aleatória e XGBoost revelou nuances que merecem atenção.

No que se refere a base compacta, a Floresta Aleatória apresentou o melhor desempenho. No teste 80/20, alcançou um RMSE de 47,97, superando levemente o XGBoost, que atingiu 48,44. Na validação cruzada, a diferença se manteve, com a Floresta Aleatória registrando um RMSE médio de 46,25, o menor entre todos os modelos. O mesmo comportamento aparece no MAE, onde o modelo obteve um erro absoluto de 36,06, enquanto o XGBoost permaneceu próximo, com 37,02. O

SARIMAX apresentou um desempenho intermediário, no teste 80/20, obteve um RMSE de 50,71 e MAE de 38,82, valores próximos aos modelos em árvore, entretanto, na validação cruzada, mostrou um RMSE médio mais alto, de 59,27, e um MAE médio de 47,60, revelando maior variabilidade entre as divisões e menor consistência temporal quando comparado aos demais modelos.

Já na base expandida, o equilíbrio entre os modelos em árvore tornou-se mais evidente. No teste 80/20, o XGBoost alcançou o menor RMSE, com 45,84, além de apresentar o maior R^2 da base. Entretanto, ao observar a validação cruzada, a Floresta Aleatória retomou os melhores resultados, com um RMSE médio de 43,78 e o R^2 médio mais elevado. Essa diferença mostra que, enquanto o XGBoost entrega o melhor resultado pontual no teste, a Floresta Aleatória sustenta um desempenho mais estável ao longo das janelas temporais. O MAE confirma esse comportamento, onde o XGBoost é levemente superior no teste único, mas a Floresta Aleatória mostra maior consistência nas diferentes divisões da série. Já o SARIMAX apresenta perda de eficiência conforme a base incorpora mais variáveis, sugerindo que sua estrutura linear não acompanha bem o aumento de complexidade.

Na base extra-expandida, a Floresta Aleatória assume vantagem de forma clara. No teste 80/20, alcançou um RMSE de 46,69 e, na validação cruzada, manteve o melhor desempenho com um RMSE médio de 44,05, ambos os menores entre todos os modelos. Seus valores de MAE também foram os mais baixos, tanto no teste quanto na validação, reforçando um ajuste firme e estável. O XGBoost apresentou desempenho próximo, mas ainda inferior, mesmo após a regularização, o que indica que o ganho adicional de complexidade não se converteu em melhoria significativa nessa base. Por outro lado, o SARIMAX mostrou uma deterioração expressiva, com RMSE de teste acima de 77 e MAE superior a 62, revelando grande dificuldade em lidar com o volume adicional de variáveis e com as interações não lineares presentes. Nesse cenário mais rico em efeitos exógenos, os modelos de inteligência computacional se mostraram muito mais capazes de capturar os padrões complexos e generalizar de forma eficiente.

Outro ponto relevante na comparação é o comportamento dos modelos ao longo das etapas de treino, teste e validação cruzada. A Floresta Aleatória se mostrou o modelo mais estável, mantendo relações equilibradas entre as etapas e variações pequenas entre as janelas temporais, o que reforça sua capacidade de generalização. O XGBoost, por sua vez, apresentou oscilações maiores nas primeiras execuções,

sobretudo nas bases mais amplas, mas após a regularização com *early stopping* passou a registrar relações mais consistentes entre erro de treino e erro de teste, além de menor divergência entre as divisões da validação cruzada. O SARIMAX, embora tenha se comportado bem nas bases mais simples, perdeu coerência conforme o número de variáveis aumentou, apresentando diferenças maiores entre as etapas de avaliação e maior sensibilidade à complexidade com o crescimento das bases.

Além dos resultados numéricos, a forma como cada modelo se comportou ao longo do estudo também ajuda a entender melhor seus desempenhos. A Floresta Aleatória se mostrou consistente desde o início, mesmo sem muitos ajustes finos, entregou erros baixos e boa estabilidade ao longo do tempo. Isso indica que sua lógica de combinar várias árvores independentes conseguiu capturar bem o padrão das vendas da empresa.

O XGBoost, por sua vez, se mostrou mais sensível às configurações. Nas bases mais complexas, ele deu sinais claros de sobreajuste antes da aplicação das técnicas de regularização. Só depois de limitar a profundidade das árvores, reduzir a taxa de aprendizado, ajustar a amostragem e aplicar o *early stopping* é que o modelo passou a se comportar de forma mais estável. A partir daí, conseguiu competir diretamente com a Floresta Aleatória, chegando inclusive a superá-la em alguns cenários na base expandida. Isso mostra que o XGBoost tem um potencial muito alto, mas exige um cuidado maior na configuração para evitar ajustes excessivos ao histórico.

De modo geral, enquanto o XGBoost precisou de um processo mais rigoroso de regularização para encontrar um ponto de equilíbrio, a Floresta Aleatória entregou um desempenho estável de maneira mais consistente em todas as bases, indicando que sua arquitetura combina com o tipo de dado utilizado neste estudo. Já o SARIMAX, embora seja um modelo clássico e bastante interpretável, não se mostrou adequado como solução principal em contextos com muitas variáveis e relações não lineares, servindo melhor como referência de base do que como modelo final.

5.3 DISCUSSÃO DOS RESULTADOS

A análise conjunta dos resultados permite observar, como cada modelo reagiu às diferentes bases e métodos de validação aplicados ao longo do estudo. Um dos

pontos que mais se destacou foi o papel da validação cruzada temporal. Ao adotar janelas de 15 dias, o comportamento dos modelos tornou-se mais estável e coerente ao longo dos períodos, reduzindo oscilações e oferecendo uma visão mais realista do desempenho em cenários futuros. Esse tipo de validação se mostrou fundamental para revelar padrões que não aparecem quando se observa apenas o recorte 80/20, contribuindo para decisões mais seguras sobre qual modelo realmente generaliza melhor no tempo.

Outro aspecto importante diz respeito às próprias características dos modelos frente ao formato dos dados. A Floresta Aleatória se mostrou forte desde as primeiras execuções, mesmo antes de ajustes mais finos. Seus resultados foram consistentes nas três bases, sugerindo que sua lógica de construção baseada em várias árvores independentes que se complementam se adaptou muito bem ao padrão das vendas da empresa, marcado por relações não lineares e por variáveis exógenas com efeitos diretos. Em outras palavras, o modelo conversou bem com os dados desde o início, o que ajudou a mantê-lo estável ao longo de todo o processo.

O XGBoost, por outro lado, exigiu mais cuidado. Nas primeiras execuções, dava sinais claros de sobreajuste, principalmente nas bases com maior número de variáveis. Foi só depois da aplicação do *early stopping* e de configurações mais conservadoras de regularização que ele alcançou o nível de estabilidade necessário. Uma vez ajustado, passou a apresentar resultados bastante competitivos e em alguns momentos até superando a Floresta Aleatória, mas seu caminho até a estabilidade foi menos direto. Isso mostra que, apesar de poderoso, é um modelo que demanda maior atenção ao processo de ajuste.

A relação entre erro de validação e erro de treino ajudou a reforçar essa interpretação. A Floresta Aleatória manteve valores próximos de 1, indicando equilíbrio entre aprender com o passado e prever o futuro sem exagerar no ajuste. O XGBoost, após regularizado, também passou a apresentar esse equilíbrio. Já o SARIMAX foi o que mostrou mais instabilidade. Apesar de capturar bem a sazonalidade semanal, seu desempenho variou bastante conforme aumentava a quantidade de variáveis, sugerindo que sua estrutura linear encontrou limitações diante de padrões mais complexos.

A diferença entre modelos ficou ainda mais clara na base extra-expandida. Enquanto Floresta Aleatória e XGBoost conseguiram lidar bem com o acréscimo de variáveis e mantiveram métricas baixas e estáveis, o SARIMAX teve uma queda

significativa de desempenho, indicando que a complexidade adicional não se encaixou bem em sua lógica autorregressiva. Isso não invalida o modelo que segue sendo útil e interpretável, mas mostra que, para problemas deste tipo, com vários efeitos combinados e relações não lineares, os modelos em árvore tendem a responder melhor.

No que tange as bases de dados, com o conjunto das análises feitas, a base expandida se mostrou a mais equilibrada para o problema estudado. Ela apresentou métricas estáveis, baixo nível de sobreajuste e respostas consistentes tanto na validação 80/20 quanto na validação cruzada temporal. A base compacta, embora simples e eficiente, perdeu capacidade preditiva ao não incorporar efeitos relevantes presentes nas variáveis adicionais. Já a base extra-expandida, apesar de mais completa, elevou demais a complexidade e afetou especialmente o desempenho do SARIMAX, trazendo ganhos apenas marginais em relação aos modelos de inteligência computacional. Assim, a base expandida atingiu o melhor ponto de equilíbrio entre informação, complexidade e desempenho, tornando-se a escolha mais adequada para futuras aplicações e refinamentos dos modelos.

Por fim, vale destacar um ponto essencial: mesmo trabalhando com um conjunto limitado de informações, imposto pela política de segurança da empresa, os modelos conseguiram capturar padrões relevantes das vendas. Várias variáveis importantes, como valores de pagamento, categorias de produtos, tipos de transação, informações sobre contratos, informações de campanhas e muito mais, não puderam ser incluídas. Ainda assim, o desempenho foi consistente e mostrou que, mesmo com dados restritos, é possível extrair previsões úteis e estruturar uma análise sólida do comportamento da empresa. Isso reforça que, com acesso a dados mais completos, o potencial de precisão e detalhamento tende a ser ainda maior.

6 CONSIDERAÇÕES FINAIS

Este estudo avaliou modelos de inteligência computacional (Floresta Aleatória, XGBoost aplicados à previsão de vendas, comparando-os a um modelo estatístico de referência (SARIMAX), utilizando dados históricos de fluxo de clientes, vendas e informações temporais, apoiando a tomada de decisões gerenciais.

Inicialmente, revisou-se a literatura para identificar abordagens adequadas para séries temporais. Em seguida, os modelos foram aplicados em três bases de dados com diferentes níveis de informação, permitindo analisar como cada algoritmo se comporta diante de estruturas mais simples ou mais complexas.

Os testes mostraram que os modelos em árvore apresentaram melhor desempenho na maior parte dos cenários, com destaque para a Floresta Aleatória, que manteve resultados robustos tanto no teste 80/20 quanto na validação cruzada com janelas de 15 dias. A validação temporal teve papel fundamental ao revelar a estabilidade das previsões ao longo do tempo e ao permitir comparar os modelos de forma mais realista, aproximando a avaliação das condições de uso futuro pela empresa. Mesmo com uma quantidade reduzida de variáveis devido às restrições de acesso impostas pela organização, os modelos foram capazes de captar o padrão da série e produzir estimativas consistentes.

Entre as dificuldades encontradas no desenvolvimento, destaca-se o acesso limitado a dados sensíveis, o que restringiu o uso de informações potencialmente úteis. Além disso, o processo de ajuste dos modelos exigiu atenção especial ao controle de sobreajuste, sobretudo no XGBoost, que precisou de regularização para estabilizar os resultados. Essas limitações, porém, não comprometeram o estudo, que ainda assim apresentou ganhos significativos de desempenho em relação ao modelo de referência.

Destaca-se também que, embora este trabalho tenha se concentrado no treinamento de modelos para prever as vendas diárias a partir de informações de fluxo de clientes e variáveis de calendário, sua aplicação prática é direta. Em um cenário real de implantação, os modelos poderiam ser integrados a um sistema que diariamente recebesse informações operacionais e do calendário da loja, gerando previsões automáticas para apoiar a tomada de decisão. Tal integração permitiria ao varejo utilizar as estimativas em atividades como planejamento de estoque, alocação

de equipes e gerenciamento de promoções, ampliando o impacto das previsões no ambiente empresarial.

Para pesquisas futuras, recomenda-se ampliar o conjunto de variáveis, incluindo dados comerciais, comportamentais e financeiros, que podem enriquecer ainda mais as previsões. Sugere-se também explorar modelos mais avançados, como redes neurais recorrentes e arquiteturas baseadas em aprendizado profundo, do inglês *deep learning*, além de integrar informações externas. Esses avanços podem tornar o sistema preditivo mais preciso e aplicável em diferentes frentes operacionais da empresa.

Como continuidade deste trabalho, recomenda-se o desenvolvimento de modelos específicos para previsão do fluxo diário de clientes, permitindo a construção de sistemas mais completos e capazes de antecipar tanto a demanda quanto o volume de visitantes na loja. A integração entre previsões de fluxo e de vendas ampliaria significativamente a capacidade de planejamento e fortaleceria as decisões estratégicas da organização.

REFERÊNCIAS

- ABEPRO - Associação Brasileira de Engenharia de Produção. A profissão. Disponível em: <<https://portal.abepro.org.br/abepro2025/profissao/>>. Acesso em: 01 jun. 2025.
- ABURASS, Sanad; RUMMAN, Maha Abu. Quantifying overfitting: introducing the overfitting index. In: 2024 International Conference on Electrical, Computer and Energy Technologies - ICECET. IEEE, 2024.
- AHN, Hyung-il; SONG, Young Chol; OLIVAR, Santiago; MEHTA, Hershel; TEWARI, Naveen. GNN-based Probabilistic Supply and Inventory Predictions in Supply Chain Networks. arXiv preprint arXiv:2404.07523, 2024.
- ALHARBI, Fahad Radhi; CSALA, Denes. A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach. *Inventions*, v. 7, n. 4, p. 94, 2022.
- AVINASH, G.; PACHORI, H.; SHARMA, A.; MISHRA, S. Time series forecasting of bed occupancy in mental health facilities in India using machine learning. *Scientific Reports*, v. 15, p. 2686, 2025.
- BI, Xuan; ADOMAVICIUS, Gediminas; LI, William; QU, Annie. Improving Sales Forecasting Accuracy: a tensor factorization approach with demand awareness. *INFORMS Journal on Computing*, v. 34, n. 3, p. 1644-1660, 2022.
- BIANCHESSI, Bernardo Cardozo. Análise de séries temporais: comparação entre modelos preditivos em uma microcervejaria do Rio Grande do Sul. 2023.
- BLÁZQUEZ-GARCÍA, Ane et al. A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, v. 54, n. 3, p. 1-33, 2021.
- BOX, George EP et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- BREIMAN, Leo. Random forests. *Machine learning*, v. 45, p. 5-32, 2001.
- BURINSKIENE, Aurelija. Forecasting model: The case of the pharmaceutical retail. *Frontiers in Medicine*, v. 9, p. 582186, 2022.
- CERQUEIRA, Vitor; TORGO, Luis; MOZETIČ, Igor. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, v. 109, n. 11, p. 1997-2028, 2020.
- CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. p. 785-794, 2016.

DAS, Sangita; MAJI, Subhrajyoti. Impact of Comprehensive Data Preprocessing on Predictive Modelling of COVID-19 Mortality. arXiv preprint arXiv:2408.08142, 2024.

DE BAETS, Shari; HARVEY, Nigel. Incorporating external factors into time series forecasts. In: Judgment in predictive analytics. Cham: Springer International Publishing, 2023.

DENG, Jhonatan Zhang; DE OLIVEIRA, Rogério. Análise e previsão de receitas no varejo: comparação entre modelos SARIMA e XGBoost. São Paulo: Universidade Presbiteriana Mackenzie, 2024.

DORING, Lina; GRUMBACH, Felix; REUSCH, Pascal. Optimizing Sales Forecasts through Automated Integration of Market Indicators. arXiv preprint arXiv:2406.07564, 2024.

DOS SANTOS, Ana Lucia; BARBOSA, Elizabete; SILVA, Fernando; MENDES, Valderlaine. Três tipos de estudos de revisão nas pesquisas educacionais: caracterização e análise. Revista Tópicos Educacionais, Pernambuco, v. 18, n. 47, p. 1–22, 2022.

DU, Wenjie; CÔTÉ, David; LIU, Yan. Saits: Self-attention-based imputation for time series. Expert Systems with Applications, v. 219, p. 119619, 2023.

FATIMA, Syeda Sitara Wishal; RAHIMI, Afshin. A review of time-series forecasting algorithms for industrial manufacturing systems. Machines, v. 12, n. 6, p. 380, 2024.

FÁVERO, Luiz Paulo et al. Count data regression analysis: Concepts, overdispersion detection, zero-inflation identification, and applications with R. Practical Assessment, Research & Evaluation, v. 26, p. 13, 2021.

GANGULY, Priyam; MUKHERJEE, Isha. Enhancing Retail Sales Forecasting with Optimized Machine Learning Models. In: 2024 4th International Conference on Sustainable Expert Systems (ICSES). IEEE. p. 884-889, 2024.

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc., 2022.

GHAFFARIASL, Parviz; ZEINALNEZHAD, Masoomah; AHMADISHOKOOH, Amir. Optimizing PM2.5 Forecasting Accuracy with Hybrid Meta-Heuristic and Machine Learning Models. arXiv preprint arXiv:2407.01647, 2024.

HAQUE, Md Sabbirul; AMIN, Md Shahedul; MIAH, Jonayet. Retail demand forecasting: a comparative study for multivariate time series. arXiv preprint arXiv:2308.11939, 2023.

HEWAGE, Harsha Chamara; PERERA, H. Niles; BANDARA, Kasun. Enhancing Demand Forecasting in Retail: A Comprehensive Analysis of Sales Promotional Effects on the Entire Demand Life Cycle. Journal of Forecasting, 2025.

HEWAMALAGE, Hansika; BERGMEIR, Christoph; BANDARA, Kasun. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, v. 37, n. 1, p. 388-427, 2021.

HWANG, S.; LEE, Y.; JEON, B. K.; OH, S. H. Sales forecasting for new products using homogeneity-based clustering and ensemble method. *Electronics*, v. 14, n. 520, 2025.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. *Forecasting: Principles and Practice*. 3. ed. Melbourne: OTexts, 2021.

JAHIN, Md Abrar; SHAHRIAR, Asef; AL AMIN, Md. MCDFN: Supply Chain Demand Forecasting via an Explainable Multi-Channel Data Fusion Network Model. arXiv preprint arXiv:2405.15598, 2024.

JIANG, Haichen; RUAN, Jiatong; SUN, Jianmin. Application of machine learning model and hybrid model in retail sales forecast. In: *IEEE 6th International Conference on Big Data Analytics (ICBDA)*. IEEE, 2021.

JO, H.; HAN, S.; HOU, L.; MOON, S.; KIM, J. Developing and evaluating a classification model for construction defect control: A text mining and ensemble approach. *Journal of Management in Engineering*, v. 41, n. 2, p. 04024071, 2025.

KÁDÁROVÁ, Jaroslava; LACHVAJDEROVÁ, Laura; SUKOPOVÁ, Dominika. Impact of digitalization on SME performance of the EU27: Panel data analysis. *Sustainability*, v. 15, n. 13, p. 9973, 2023.

KAMBLE, Torana; VARDHAN, Harsh; GHUGE, Madhuri Sahebrao; SHELAR, Yash; RANA, Ronit; MACHALE, Tushar. Ensemble machine learning models to forecast sales. In: *International Conference on Innovative Mechanisms for Industry Applications - ICIMIA*. IEEE. p. 1056–1062, 2024.

KIM, SeungHyun et al. Probabilistic imputation for time-series classification with missing data. In: *International Conference on Machine Learning*. PMLR, 2023.

KONTOPOULOU, Vaia I. et al. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, v. 15, n. 8, p. 255, 2023.

KRISHNA, M. Ramya; FARHEEN, Mohammad Alifa Firdhos; KALYANI, D. Leela. Utilizing Attention Based Machine Learning Models for Improved Demand Forecasting in Supply Chains. In: *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*. IEEE, 2025.

LAINDER, A. D.; WOLFINGER, R. D. Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies. *International Journal of Forecasting*, v. 38, n. 4, p. 1426–1433, 2022.

LEARNING, Machine; NITHINRAJ, N.; JAISACHIN, B. Integrating Data Mining and Predictive Modeling Techniques for Enhanced Retail Optimization. *International Journal of Computer Science and Information Security (IJCSIS)*, v. 22, n. 5, 2024.

LIM, Bryan; ZOHREN, Stefan. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, v. 379, n. 2194, p. 20200209, 2021.

MA, Zhipeng; JØRGENSEN, Bo Nørregaard; MA, Zheng Grace. A Novel Hybrid Feature Importance and Feature Interaction Detection Framework for Predictive Optimization in Industry 4.0 Applications. *arXiv preprint arXiv:2403.02368*, 2024.

MAHIN, Md. Parvezur Rahman; SHAHRIAR, Munem; DAS, Ritu Rani; ROY, Anuradha; REZA, Ahmed Wasif. Enhancing Sustainable Supply Chain Forecasting Using Machine Learning for Sales Prediction. *Procedia Computer Science*, v. 252, p. 470-479, 2025.

MAKRIDAKIS, Spyros; HYNDMAN, Rob J.; PETROPOULOS, Fotios. Forecasting in social settings: The state of the art. *International Journal of Forecasting*, v. 36, n. 1, p. 15-28, 2020.

MISHRA, Arun Kumar; SINHA, Megha. Data analytics for product segmentation and demand forecasting of a local retail store using Python. *International Journal of Advanced Computer Science and Applications*, v. 16, n. 2, 2025.

MONTERO-MANSO, Pablo; HYNDMAN, Rob J. Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, v. 37, n. 4, p. 1632-1653, 2021.

MUTHUKALYANI, Ananth Raja. Unlocking accurate demand forecasting in retail supply chains with AI-driven predictive analytics. *Information Technology and Management*, v. 14, n. 2, p. 48-57, 2023.

NOSEDA, Fernando Daniel Duarte. Previsão de vendas do comércio de varejo com técnicas clássicas e de aprendizado de máquina. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Universidade Tecnológica Federal do Paraná, Curitiba, 2021.

OYEWOLE, Adedoyin Tolulope; OKOYE, Chinwe Chinazo; OFODILE, Onyeka Chrisanctus; EJAIRU, Emuesiri. Reviewing predictive analytics in supply chain management: Applications and benefits. *World Journal of Advanced Research and Reviews*, v. 21, n. 3, p. 568–574, 2024.

PANAY, Belisario; BALOIAN, Nelson; PINO, José A.; PEÑAFIEL, Sergio; FREZ, Jonathan; FUENZALIDA, Cristóbal; SANSON, Horacio; ZURITA, Gustavo. Forecasting key retail performance indicators using interpretable regression. *Sensors*, Basel, v. 21, n. 5, 2021.

PASUPULETI, Vikram; THURAKA, Bharadwaj; KODETE, Chandra Shikhi; MALISETTY, Saiteja. Enhancing supply chain agility and sustainability through

machine learning: Optimization techniques for logistics and inventory management. *Logistics*, Basel, v. 8, n. 3, p. 73, 2024.

PUNIA, Sushil; SHANKAR, Sonali. Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, v. 258, p. 109956, 2022.

ROOS-HOEFGEEST TORIBIO, Mario et al. A Novel Approach to Speed Up Hampel Filter for Outlier Detection. *Sensors*, v. 25, n. 11, p. 3319, 2025.

SAPUTRA, Jeffri Prayitno Bangkit; KUMAR, Aayush. Modeling the impact of holidays and events on retail demand forecasting in online marketing campaigns using intervention analysis. *Journal of Digital Market and Digital Currency*, v. 1, n. 2, p. 144-164, 2024.

SEKEROGLU, B; EVER, Y; DIMILLILER, K. AL-TURJMAN, F. Comparative evaluation and comprehensive analysis of machine learning models for regression problems. *Data intelligence*, v. 4, n. 3, p. 620-652, 2022.

SHAIK, Nagoor Basha; ALURU, Vamsi; JONGKITTINARUKORN, Kittiphong; ALURU, Prasad. Optimizing structural integrity of a pressure vessel via finite element analysis and machine learning based XGBoost approaches. *Scientific Reports*, v. 15, p. 11485, 2025.

STONE, Mervyn. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, v. 36, n. 2, p. 111-133, 1974.

SULLIVAN, Brian et al. Comparing conventional and Bayesian workflows for clinical outcome prediction modelling with an exemplar cohort study of severe COVID-19 infection incorporating clinical biomarker test results. *BMC Medical Informatics and Decision Making*, v. 25, n. 1, p. 123, 2025.

TISSOT, Patrícia Tais; VIDOR, Gabriel; CHIWIACOWSKY, Leonardo Dagnino. Implementação de um modelo de previsão de vendas em uma empresa de distribuição de aços especiais. *Revista Gestão e Secretariado (GeSec)*, São Paulo, v. 13, n. 4, ed. esp., p. 2499–2513, 2022.

TONY, Aneesh; KUMAR, Pradeep; JEFFERSON, Rohith; SUBRAMANIAN. A study of demand and sales forecasting model using machine learning algorithm. *Psychology and Education*, v. 58, n. 2, p. 10182-10194, 2021.

XU, Yuanhui et al. Applications of artificial intelligence and computational intelligence in hydraulic optimization of centrifugal pumps: a comprehensive review. *Engineering Applications of Computational Fluid Mechanics*, v. 19, n. 1, p. 2474675, 2025.

YANG, Bin; CHEN, Min; ZHOU, Jianjun. Forecasting the monthly retail sales of electricity based on the semi-functional linear model with autoregressive errors. *AIMS Mathematics*, v. 10, n. 1, p. 1602–1627, 2025.

ZUBAIR, Muhammad; WALEED, Aashir; REHMAN, Ans; AHMAD, Farhan; ISLAM, Mohammad; JAVED, Saqib. Machine learning insights into retail sales prediction: a comparative analysis of algorithms. In: Horizons of Information Technology and Engineering (HITE). IEEE, 2024.