UNIVERSIDADE PRESBITERIANA MACKENZIE

Cristina Lellis Villanova

Eduardo Pinheiro Canas

Hugo de Moraes Holzer

Luiz Rodrigo Alves Vergino

PADRÕES DE VENDA DE PRODUTOS NA AMAZON

Análise sobre o comportamento de consumo de produtos eletrônicos na Amazon em 2025

São Paulo

2025

Cristina Lellis Villanova

Eduardo Pinheiro Canas

Hugo de Moraes Holzer

Luiz Rodrigo Alves Vergino

PADRÕES DE VENDA DE PRODUTOS NA AMAZON

Análise sobre o comportamento de consumo de produtos eletrônicos na Amazon em 2025

Projeto aplicado II apresentado ao Programa de Graduação da Universidade Presbiteriana Mackenzie como requisito parcial para obtenção do título Tecnólogo em Ciência de Dados.

Orientador: Prof. Anderson Adaime de Borba

São Paulo

2025

Sumário

1. KICK-OFF DO PROJETO

1.1. DEFINIÇÃO DO GRUPO DE TRABALHO

Para a realização desta tarefa e como requisito pré-definido pela orientação do curso, a primeira etapa consiste na definição do grupo de trabalho que irá elaborar um projeto de Ciência de Dados que contemple manipulação de imagens ou textos utilizando os componentes abordados durante a 3ª etapa da graduação em Tecnologia de Ciência de Dados.

Nosso grupo está definido e é composto pelos seguintes integrantes:

- Cristina Lellis Villanova (RA 10319958)
- Eduardo Pinheiro Canas (RA 10184419)
- Hugo de Moraes Holzer (RA 10142961)
- Luiz Rodrigo Alves Vergino (RA 10176038)

Para a realização deste trabalho, iremos utilizar recursos virtuais de reunião como Zoom e Google Meet, além de meios de comunicação instantânea como Whatsapp.

Também faremos a organização e mapeamento de tudo o que for realizado durante as entregas através do Github criado especificamente para este trabalho. O link do repositório está fixado abaixo e poderá ser acessado por todos os integrantes e também pelo corpo docente para atualizações e acompanhamento.

https://github.com/canasep/mackprojeto2

2. PREMISSAS DO PROJETO

2.1. DEFINIÇÃO DA EMPRESA

Para este projeto, escolhemos através do Kaggle, uma base referente ao comportamento de consumo de produtos eletrônicos pela Amazon em 2025.

Essa base pode ser visualizada através do site:
https://www.kaggle.com/datasets/ikramshah512/amazon-products-sales-dataset-42k-items-2025?select=amazon_products_sales_data_uncleaned.csv

2.2. OBJETIVO GERAL

Identificar padrões de consumo de produtos eletrônicos vendidos na Amazon em 2025, a partir da análise de dados de preços, avaliações e promoções.

2.3. OBJETIVO ESPECÍFICO

- Avaliar os dados e realizar as limpezas necessárias ;
- Avaliar a correlação entre preço e volume de vendas;
- Verificar o impacto de cupons e descontos sobre a demanda;
- Identificar produtos frequentemente comprados em conjunto (market basket analysis),
- Desenvolver um modelo preditivo de vendas com base em variáveis como preço, categoria e avaliação do produto.

2.4. METAS

- Identificar pelo menos 3 combinações de produtos com alta frequência de compra conjunta (suporte > 5%).
- Validar a hipótese de que descontos acima de 15% aumentam as vendas em pelo menos 10% no mês seguinte.
- Entrega de insights visual com os principais padrões encontrados.

2.5. ÁREA DE ATUAÇÃO

Atualmente, a Amazon é uma das gigantes de tecnologia e comércio eletrônico no mundo. Seu principal foco atualmente é a atuação no varejo online, onde oferece ampla variedade de produtos, entre eles, o foco de estudo desse projeto.

A Amazon também atua no setor de dispositivos eletrônicos e inovação tecnológica, com produtos como o leitor Kindle, a assistente virtual Alexa e os dispositivos Echo, que integram soluções de casa inteligente. Além disso, investe em inteligência artificial, automação logística e robótica para aprimorar a experiência do cliente e a eficiência operacional.

2.6. APRESENTAÇÃO DOS DADOS

Para este projeto, trabalharemos com a apresentação de texto, onde abordaremos as análises pertinentes ao decorrer do curso. Poderemos também utilizar gráficos para demonstrações assim como a disponibilização dos códigos em python e R utilizados para a realização das análises.

Nosso dataset possui um total de 42.675 linhas e 17 colunas, sendo elas:

- product_title → Nome do produto
- product_rating → Avaliação do produto (nota média)
- total_reviews → Número total de avaliações
- purchased_last_month → Compras realizadas no último mês
- discounted_price → Preço com desconto

- original_price → Preço original
- is_best_seller → Produto mais vendido (sim/não)
- is_sponsored → Produto patrocinado (sim/não)
- has_coupon → Possui cupom de desconto (sim/não)
- buy_box_availability → Disponibilidade na "Buy Box" (sim/não)
- delivery_date → Data estimada de entrega
- sustainability_tags → Selos de sustentabilidade
- product_image_url → Link da imagem do produto
- product_page_url → Link da página do produto
- data_collected_at → Data da coleta das informações
- product_category → Categoria do produto
- discount_percentage → Percentual de desconto aplicado

3. OBJETIVOS E METAS

A presente análise tem como objetivo comparar o faturamento de produtos específicos a partir de variáveis relacionadas a selos de eficiência e certificação, tais como consumo de energia, Amazon Choice, compatibilidade com dispositivos inteligentes (Works with Alexa) e avaliações fornecidas pelos usuários. Para viabilizar esse estudo, torna-se imprescindível a etapa de aquisição e preparação de dados, contemplando o tratamento, limpeza e padronização das informações em um banco de dados relacional. Essa organização possibilita a aplicação de técnicas de análise estatística descritiva, permitindo a identificação de padrões de distribuição, médias, variâncias e outros indicadores que contribuem para compreender o comportamento do volume de vendas e, consequentemente, reconhecer quais produtos se destacam em relação ao desempenho comercial.

Complementarmente, busca-se identificar padrões de consumo a partir da análise de preços, tendências de mercado e comportamento de compra. Nesse sentido, são considerados sistemas de recomendação e a investigação de cestas de compras, com foco na identificação de itens frequentemente adquiridos em conjunto. Para além da análise descritiva, empregam-se técnicas de machine learning, orientadas à construção de modelos preditivos capazes de projetar vendas futuras, estimar demandas e mensurar o impacto de políticas comerciais, como a concessão de descontos. Dessa forma, este trabalho integra metodologias oriundas da estatística preditiva e da aprendizagem de máquina, articulando teoria e prática no apoio à tomada de decisão estratégica voltada à otimização do desempenho comercial.

4. CRONOGRAMA DE ATIVIDADES

Tabela 1: Cronograma com etapas do trabalho e entregas definidas

Prazo	Atividade	Descrição	Responsável	Realizado
31/08/2025	Criação do Repositório	Criação do Github que servirá como repositório de arquivos e entregas realizadas.	Eduardo Pinheiro Canas	*x*
04/09/2025	Aula Inaugural do Curso	O grupo estará presente para a primeira aula com o Professor orientador e tirará dúvidas pertinentes ao trabalho e sua realização.	Todos os integrantes	*x*
04/09/2025	Definição da Empresa e demais atividades da primeira entrega	Primeira reunião do grupo para estruturar o trabalho e realizar a primeira entrega.	Todos os integrantes	*x*
05/09/2025	Envio Final da Etapa 1	Entrega da Primeira Etapa para avaliação do Professor Orientador.	Todos os integrantes	x
13/09/2025	Revisão 1	Revisão do trabalho conforme feedback do professor.	Todos os integrantes	x
18/09/2025	Reunião do Grupo	Discussão de pontos chave e início da confecção da 29 entrega.	Todos os integrantes	x
03/10/2025	Envio Final da Etapa 2	Entrega da Segunda Etapa para avaliação do professor.	Todos os integrantes	
09/10/2025	Revisão 2	Ajustes na Segunda Etapa conforme feedback do professor.	Todos os integrantes	
16/10/2025	Reunião do Grupo Etapa 3	Discussão de pontos chave e início daconfecção da 39 entrega.	Todos os integrantes	
24/10/2025	Envio Final da	Entrega final da Terceira Etapa.	Todos os integrantes	
30/10/2025	Revisão 3	Ajustes na Segunda Etapa conforme feedback do professor.	Todos os integrantes	
21/11/2025	Envio Final da Etapa 4	Entrega final da Quarta Etapa.	Todos os integrantes	

Fonte: Elaborado pelos autores

5. REFERENCIAL TEÓRICO

5.1. PANDAS – EXTRAÇÃO E PREPARAÇÃO DE DADOS

A biblioteca Pandas representa uma ferramenta essencial no ambiente Python para análise e o processamento de dados, sendo adotada tanto em ambientes educacionaisquanto em aplicações profissionais. Desenvolvida por Wes McKinney em 2008, elaemergiu para atender à demanda por gerenciamento eficiente de dados variados eorganizados em formatos tabulares, superando limitações das opções disponíveis naépoca (McKINNEY, 2022).

O elemento central do Pandas é o DataFrame, uma estrutura que facilita o manuseio de conjuntos de dados em duas dimensões, suportando tarefas como seleção, filtragem, agrupamento, junção de conjuntos e modificações em grande volume. Adicionalmente, elainclui recursos para leitura e escrita em múltiplos formatos, incluindo CSV, Excel, JSON,SQL e Parquet (PANDAS DEVELOPMENT TEAM, 2024).

No âmbito da ciência da computação, o Pandas se destaca em áreas relacionadas a bancos de dados, extração de padrões em dados e modelagem preditiva, servindo comoponte entre a aquisição e o refinamento de informações. Funções como read_csv(),merge(), groupby() e json_normalize() auxiliam na integração de fontes diversas,preparando o terreno para análises estatísticas avançadas em conjunto com ferramentascomo NumPy e SciPy.

De acordo com McKinney (2022), um dos pontos fortes do Pandas reside em sua interface amigável, que torna o tratamento de dados em Python similar a comandos SQL ou apráticas em ambientes como R.

5.2 NUMPY E SCIPY – FUNDAMENTOS NUMÉRICOS E CÁLCULOS CIENTÍFICOS

NumPy e SciPy constituem as fundações para operações matemáticas e científicas no Python. O NumPy oferece suporte a arrays e matrizes de múltiplas dimensões, juntamente com rotinas otimizadas para cálculos matemáticos elementares e sofisticados. Harris et al.(2020) apontam que o NumPy revolucionou a programação ao introduzir o conceito de vetorização, resultando em melhorias substanciais de performance em relação a loopsconvencionais no Python.

Por sua vez, o SciPy expande as capacidades do NumPy com módulos dedicados a integração numérica, otimização, álgebra linear, análise estatística e tratamento de sinais(VIRTANEN et al., 2020). A versão 1.0 do SciPy estabeleceu-se como um repositório dealgoritmos essenciais para a computação científica, beneficiando tanto iniciativas depesquisa quanto implementações práticas no setor industrial.

Em programas de ciência da computação, essas bibliotecas são indispensáveis para matérias como métodos numéricos, álgebra computacional e estatística, onde elassimplificam a aplicação de técnicas tradicionais para resolver equações diferenciais,realizar decomposições de matrizes e examinar distribuições probabilísticas.

Portanto, NumPy e SciPy emergem como componentes fundamentais da computação científica em Python, fornecendo uma infraestrutura sólida para operações matemáticasque embasam o processamento de dados e a criação de algoritmos em campos comointeligência artificial e desenvolvimento de software.

5.3 MATPLOTLIB E SEABORN – VISUALIZAÇÃO DE DADOS

A representação gráfica de dados é um componente crucial no fluxo analítico, facilitando a compreensão de achados e a disseminação de insights. O Matplotlib,

idealizado por JohnHunter, é reconhecido como a base para gráficos em Python, proporcionando versatilidadena geração de visualizações em duas ou três dimensões (HUNTER, 2007).

O Matplotlib permite um controle detalhado sobre componentes visuais, abrangendo desde diagramas básicos até representações elaboradas. Sua compatibilidade com Pandas eNumPy o torna uma escolha recorrente em documentos técnicos.

Em complemento, o Seaborn foi projetado para agilizar a produção de gráficos estatísticos com design refinado e forte ligação com os formatos de dados do Pandas (WASKOM,2024). Ele adota uma perspectiva de alto nível, possibilitando a criação rápida de plots dedispersão, histogramas de distribuições e análises categóricas.

Em iniciativas educacionais, a combinação de Matplotlib e Seaborn não só apoia a exploração inicial de dados, mas também aprimora a exposição de resultados de maneiraacessível e instrutiva, auxiliando na interpretação de outcomes de testes e casos práticos.

5.4 SCIKIT-LEARN – APRENDIZADO DE MÁQUINA

O scikit-learn destaca-se como uma biblioteca chave para machine learning em Python, abrangendo uma extensa gama de algoritmos destinados a tarefas de classificação,regressão, agrupamento e redução de dimensões. Pedregosa et al. (2011) ressaltam queseu design prioriza a simplicidade de uso, o reaproveitamento de código e a otimização derecursos computacionais.

Dentre seus recursos principais, sobressaem os componentes para preparação de dados, fluxos de processamento, avaliação cruzada e ajuste de parâmetros. Elementos como aclasse Pipeline e o utilitário GridSearchCV promovem experimentos científicos maisconsistentes e sistemáticos (PEDREGOSA et al., 2011).

No cenário da formação em ciência da computação, o scikit-learn serve como suporteessencial em cursos de inteligência artificial, extração de conhecimento de dados e modelagem de aprendizado, permitindo a concretização de princípios teóricos em práticasreais.

Dessa maneira, o scikit-learn se afirma como um ambiente tanto pedagógico quanto operacional, facilitando a emprego de técnicas de aprendizado de máquina em contextosacadêmicos e demandas comerciais.

5.5 O MÓDULO RE DO PYTHON E AS EXPRESSÕES REGULARES

O módulo re em Python é uma das funcionalidades mais valiosas da linguagem para o processamento avançado de sequências de caracteres. Ele incorpora expressõesregulares (regex), um sistema de padrões projetado para detectar, verificar, isolar emodificar partes específicas de texto em variados cenários. Essa capacidade éamplamente explorada na computação, especialmente em domínios como

processamentode linguagem natural (PLN), análise de dados, verificação de insumos do usuário eextração de informações.

Ao incluir o módulo via import re, o desenvolvedor ganha acesso a um conjunto de procedimentos para gerenciar padrões textuais elaborados. O Python, ao integrar o re nasua biblioteca nativa, assegura alinhamento com a notação padrão de regex emlinguagens como Perl e Java, promovendo portabilidade e aproveitando convençõesestabelecidas na comunidade de programação (ROSSUM; DRAKE, 2009).

Dentre as funções chave disponíveis, incluem-se re.match(), re.search(), re.findall(),re.sub() e re.split(). Elas se aplicam em situações variadas: match() verifica padrões apartir do começo da string; search() varre o texto inteiro e captura a primeira instância;findall() coleta todas as instâncias; sub() executa trocas baseadas em regras; e split()segmenta o texto conforme separadores definidos.

A definição de padrões em regex depende de caracteres especiais, como. para qualquer símbolo, d para números, w para elementos alfanuméricos e s para espaços. Além disso, âncoras como ^ e $ marcam o início e o término da string, enquanto quantificadores como + e {n,m} controlam o número de repetições. Essa notação permite o desenvolvimento de critérios refinados para o tratamento de conteúdo.

Em resumo, o módulo re do Python, com sua interface robusta para expressões regulares, revela-se um instrumento vital para profissionais de computação, engenheiros acadêmicos. Sua importância abrange o plano educacional, sustentando tanto teoria quanto prática, e o contexto laboral, ao oferecer respostas ágeis para desafios de processamento e exame de textos

5.6 INTEGRAÇÃO NO CONTEXTO ACADÊMICO

A adoção dessas bibliotecas no meio universitário é respaldada pelas diretrizes curriculares e recursos da graduação em Ciência da Computação da UniversidadePresbiteriana Mackenzie.

Matérias como gerenciamento de bancos de dados, estatísticaprática, análise de padrões em dados e sistemas inteligentes rotineiramente indicam essasferramentas, seja em leituras adicionais ou em atividades hands-on.

O arquivo institucional da Mackenzie abriga inúmeras monografias de final de curso queempregam Pandas, NumPy, SciPy, Matplotlib, Seaborn e scikit-learn como pilares metodológicos, demonstrando sua importância no progresso científico dos discentes.

A sinergia entre essas bibliotecas capacita o aluno a navegar desde a obtenção de dados,p assando pelo processamento matemático, a geração de gráficos e a construção demodelos preditivos, espelhando processos profissionais em tecnologia e pesquis aaplicada.

Essa conexão reforça a capacitação do profissional de computação, preparando-o

para enfrentar questões intricadas no manejo e exame de informações.

6. LIMPEZA DE DADOS

A etapa inicial consistiu na aquisição e organização da base de dados. O dataset foi obtido a partir de sua fonte original, estruturado em diretório próprio e convertido em DataFrame para permitir a manipulação dentro do ambiente de análise. Esse processo assegurou a rastreabilidade da origem dos insumos e o controle de versões necessários para o desenvolvimento subsequente do pipeline.

Foram aplicados procedimentos de limpeza e padronização de atributos. Variáveis categóricas relacionadas a certificados e selos foram normalizadas em classes consistentes, reduzindo o ruído causado por rótulos heterogêneos e facilitando a construção de métricas confiáveis. Os atributos operacionais também passaram por transformações: o campo de compra direta foi convertido em variável binária, enquanto o cupom foi decomposto em múltiplas dimensões (presença, percentual e valor absoluto), preservando granularidade para análises comparativas e modelagem preditiva.

Definiu-se como variável alvo o volume de unidades vendidas, enquanto o conjunto de preditores incluiu preço, avaliações, rating, patrocínio, selos normalizados, cupons e compra direta, entre outros atributos. A divisão entre treino e teste foi realizada de forma controlada para garantir reprodutibilidade, e os valores ausentes foram tratados por imputação estatística. Na etapa final, um modelo de regressão linear foi ajustado e avaliado com métricas adequadas, assegurando que a performance refletisse exclusivamente a capacidade de generalização sobre dados de teste, sem risco de leakage.

7. ANÁLISE EXPLORATÓRIA DE DADOS

O conjunto de dados foi submetido a um processo de padronização prévio à análise exploratória. As variáveis numéricas foram normalizadas e convertidas para formatos consistentes, garantindo comparabilidade entre registros. Procedimentos específicos, como a extração do preço principal por meio de expressões regulares e a decomposição do atributo de cupom em múltiplas representações (binária, percentual e absoluta), permitiram preservar granularidade e viabilizar análises posteriores mais robustas. Tais transformações reduziram ruído informacional e estabeleceram a base para a construção de medidas e visualizações coerentes.

Na sequência, o conjunto foi segmentado por categorias, originando subconjuntos destinados à análise comparativa entre grupos de produtos e marcas. Essa estratégia possibilitou identificar diferenças estruturais, como padrões de presença de selos de promoção ou patrocínio, revelando particularidades que podem não ser captadas em análises globais.

A análise exploratória incluiu a investigação de atributos binários — como patrocínio, cupons e selos de destaque — e variáveis contínuas, a exemplo de preços, avaliações e unidades vendidas. A partir de gráficos de barras, histogramas e distribuições, foi possível observar padrões de concentração e assimetrias características do varejo on-line, em que poucos itens concentram elevados volumes de vendas e avaliações, enquanto a maioria apresenta baixa representatividade.

Foram também conduzidas análises bivariadas por meio de matrizes de correlação, avaliando relações entre volume de vendas e variáveis como preço, patrocínio, cupons e selos de destaque. Observou-se que fatores ligados ao reconhecimento social, como número de avaliações e classificações atribuídas pelos consumidores, apresentam maior associação com vendas, ainda que sujeitos a possíveis efeitos de causalidade reversa. Já os efeitos de patrocínio e selos mostraram-se heterogêneos, variando de acordo com a categoria de produto analisada.

Apesar dessas tendências, a aplicação do coeficiente de correlação de Spearman indicou a inexistência de correlações fortes entre as variáveis de interesse e o volume de vendas. Ainda assim, procedeu-se à realização de regressões em cenários hipotéticos, com o objetivo de explorar possíveis relações latentes e avaliar a utilidade dos atributos como preditores em modelos futuros. Essa etapa buscou sustentar hipóteses de trabalho, mesmo diante da ausência de correlações robustas, fornecendo subsídios para análises subsequentes e para a seleção de variáveis em modelos preditivos.

8. MÉTODO E BASES TEÓRICAS

A Regressão Linear Múltipla é uma técnica usada para estudar como uma variável, como o número de vendas, pode ser explicada por várias outras variáveis ao mesmo tempo, como preço, avaliações e selos de destaque. A ideia principal é que cada variável tem um coeficiente que mostra, em média, quanto a variável resposta pode mudar quando aquela preditora aumenta em uma unidade, mantendo as outras constantes. Para que o modelo funcione bem, existem algumas condições esperadas, como a relação linear entre variáveis, erros independentes, variância constante e resíduos próximos de uma distribuição normal.

Um recurso comum é calcular correlações, como o coeficiente de Pearson, que mostra se existe relação linear entre duas variáveis. No entanto, a correlação não significa necessariamente causalidade e pode ser influenciada por valores extremos ou relações não lineares. Quando duas ou mais variáveis preditoras estão muito relacionadas entre si, ocorre a chamada multicolinearidade, que atrapalha a interpretação dos coeficientes da regressão.

A a forma como as variáveis entram no modelo: as que possuem apenas dois valores, como “patrocinado” (sim/não), são transformadas em indicadores binários (0 e 1). Já variáveis categóricas com várias classes precisam ser transformadas em múltiplas colunas (one-hot encoding), para que o modelo consiga interpretá-las corretamente. Também é comum aplicar transformações em variáveis muito distorcidas, como preço ou número de avaliações, para que se ajustem melhor às hipóteses do modelo.

Nós analisamos como algumas características dos produtos se relacionam com as vendas. Para isso, usamos um método estatístico chamado regressão linear, que basicamente “aprende” a partir dos dados qual é o peso de cada fator na quantidade vendida. Os fatores usados foram: preço líquido do produto, quantidade de avaliações, nota média, e alguns sinais de destaque (se o anúncio é patrocinado, se tem cupom/desconto, se permite compra direta, e os selos “Mais vendido” e “Amazon’s Choice”).

Antes de rodar o modelo, organizamos a base: separamos parte dos dados para treinar o modelo e outra parte para testar se ele faz previsões razoáveis; tratamos valores ausentes

com um preenchimento simples e deixamos as informações em formato padronizado. Depois disso, ajustamos a regressão e medimos a qualidade das previsões com métricas de erro. O objetivo aqui não foi "acertar o número perfeito" de vendas, mas ter um instrumento confiável o suficiente para comparar cenários.

A utilidade prática para o negócio vem das simulações (os "e se…"). Com o modelo pronto, conseguimos estimar como a previsão de vendas mudaria se, por exemplo, ativarmos um cupom, reduzirmos um pouco o preço ou melhorarmos o rating. Em termos simples: mantemos o produto igual e ligamos/desligamos um fator por vez para ver o que tende a acontecer. Isso ajuda a priorizar ações: quais SKUs devem receber cupom primeiro, onde um pequeno ajuste de preço tem mais chance de retorno, ou em que categorias faz sentido investir em patrocínio.

Do ponto de vista de gestão, o ganho é enxergar trade-offs com dados. Em vez de aplicar cupom "no escuro", podemos direcionar para itens com maior sensibilidade estimada; em vez de mexer no preço de todo o portfólio, começamos onde o impacto previsto é maior; e podemos avaliar se selos e patrocínio parecem mover a agulha em determinadas categorias. Isso não substitui teste de mercado, mas reduz o custo de experimentação ao indicar onde testar primeiro.

Também deixamos claras duas limitações para leitura responsável. Primeiro, vendas são contagens e tendem a ser desiguais (poucos campeões, muitos caudas longas), o que pode dificultar previsões perfeitas. Segundo, fatores como nota e número de avaliações podem ser consequência do próprio sucesso do produto (causalidade reversa). Por isso, tratamos os resultados como guias de priorização, não como garantias.

9. ACURÁCIA

Avaliamos o modelo no conjunto de teste para entender, em termos práticos, "o quanto ele erra" e "o quanto ele explica". O MAE (erro absoluto médio) mostra, em média, quantas unidades o modelo se afasta do valor real, é a leitura mais direta para gestão porque indica o desvio esperado por item. O RMSE (raiz do erro quadrático médio) dá mais peso aos grandes erros e, por isso, revela riscos concentrados em poucos produtos únicos; se o RMSE ficar muito acima do MAE, há outliers que merecem revisão. O MSE é a base do RMSE e serve como controle técnico. Já o R^2 indica a parcela da variação de vendas que o modelo consegue explicar; quanto mais próximo de 1, maior a capacidade de diferenciar produtos de alta e baixa tração.

O significado dos números aparece quando comparamos com um baseline simples (por exemplo, "prever a média da categoria"). Se nosso MAE/RMSE forem menores que os do baseline e o R^2 for positivo e razoável, o modelo já agrega valor: ele reduz o erro médio e melhora a separação entre itens que tendem a vender mais ou menos. Como vendas em marketplace têm cauda longa (poucos campeões, muitos de baixa saída), é importante ler as métricas por categoria e vigiar casos com erro alto. Dois sinais práticos: (i) RMSE muito maior que MAE sugere revisar dados e regras para produtos únicos extremos; (ii) R^2 baixo em uma categoria específica pode indicar que faltam variáveis relevantes (ex.: sazonalidade ou competição) para aquele segmento.

Do ponto de vista do negócio, usamos essas métricas para priorizar decisões e

controlar risco. Um MAE menor que o baseline dá segurança para testar cupom e ajuste fino de preço de produtos únicos com maior sensibilidade estimada; um RMSE sob controle reduz as chances de "estourar meta" por erro grande em poucos itens; e um R^2 estável por categoria aumenta a confiança na alocação de mídia/patrocínio. Operacionalmente, recomenda-se: reportar MAE/RMSE e R^2 por categoria a cada ciclo, acompanhar os top-20 maiores erros para correção de dados e refinar o portfólio de variáveis onde o R^2 for fraco. Assim, a acurácia deixa de ser um número isolado e vira um instrumento de gestão, orientando onde investir esforço para obter mais retorno com menos tentativa-e-erro.

10. RESULTADOS PRELIMINARES

A hipótese inicial de que descontos, promoções, selos e rating explicariam diretamente as vendas não se confirmou no agregado. A base apresenta distribuição em cauda longa e forte heterogeneidade por categoria e marca, o que dilui correlações lineares simples. Em outras palavras, não há uma regra única que valha para todo o portfólio. A relação entre atributos e desempenho existe, mas é local e dependente de contexto.

Mesmo com correlações fracas, as regressões e os cenários "what-if" foram úteis para comparar alternativas de ação mantendo as demais variáveis constantes. Observamos sensibilidade heterogênea a cupom, preço e patrocínio: algumas categorias respondem de forma clara a incentivos, enquanto outras quase não se movem. Marcas mais fortes tendem a depender menos de desconto contínuo, e sinais como rating e número de avaliações se associam a vendas, mas devem ser lidos com cautela por possível causalidade reversa.

Do ponto de vista do negócio, a implicação é operar de forma seletiva e por segmento. Promoções devem priorizar SKUs "quase lá", nos quais um pequeno ajuste de preço ou um cupom leve tem maior retorno esperado. O patrocínio deve ser alocado nas categorias onde o efeito marginal foi mensurável e reduzido onde se mostrou inelástico. Para marcas consolidadas, faz mais sentido investir em sortimento, posição e experiência do que em descontos amplos, reduzindo desperdício de verba e aumentando a taxa de acerto dos testes.

Como leitura responsável, tratamos os resultados como um guia de priorização e não como garantias. A acurácia em teste indica utilidade para ordenar decisões, mas com variação entre categorias e atenção a outliers. Os próximos passos recomendados são aprofundar a segmentação, testar modelos mais aderentes a contagens ou reformular casos como classificação de top quartil, além de rodar pilotos A/B de cupom e preço nos SKUs priorizados e ajustar o patrocínio conforme o uplift observado. Com isso, transformamos uma hipótese inicial refutada em um processo de decisão mais eficiente e com menor risco.

11. STORYTELLING

Entramos neste trabalho com uma hipótese direta: descontos, promoções, selos e rating deveriam se refletir em mais vendas. Padronizamos a base, exploramos categorias e marcas e, logo no início, apareceu o "plot twist": o comportamento é bem mais complexo e as correlações lineares no agregado são fracas. A distribuição tem cauda longa (poucos campeões, muitos itens com baixa venda) e os efeitos mudam conforme o contexto — categoria, marca, posicionamento. Conclusão: não existe "regra única" que valha para todo

o portfólio.

Mesmo assim, decidimos avançar com regressões e simulações “what-if” para comparar cenários de forma controlada (ceteris paribus). O objetivo não foi acertar um número perfeito de vendas, mas ordenar decisões: o que tende a responder melhor a cupom, ajuste fino de preço, patrocínio ou selos. As simulações mostraram sensibilidade heterogênea: há categorias onde cupom realmente move a agulha e outras onde o efeito é pequeno; patrocínio aparece como acelerador em alguns nichos, mas não em todos; e métricas sociais (rating/avaliações) pedem leitura cautelosa por possível causalidade reversa.

O principal insight de negócio é que alavancas são locais. Marcas mais fortes tendem a depender menos de desconto contínuo; já marcas menos conhecidas ganham com promoção seletiva e patrocínio bem alocado. Em vez de políticas amplas, a recomendação é operar por segmento/categoria, priorizando produtos únicos “quase lá” (onde pequeno ajuste de preço ou um cupom leve traz maior retorno) e realocando mídia para os grupos onde o ganho marginal estimado é maior. Isso reduz desperdício de verba e aumenta a taxa de acertos dos testes.

Como próximos passos, tratamos o modelo como bússola de priorização, não como garantia. Vamos aprofundar a leitura por categoria, testar modelos mais aderentes a contagem (Poisson/NegBin) ou reformular alguns casos como classificação (probabilidade de entrar no top quartil). Em paralelo, propomos uma rodada curta de A/B com cupom e preço nos produtos priorizados e um ajuste tático de patrocínio nas categorias mais promissoras. Com isso, transformamos uma hipótese inicial refutada em um processo de decisão mais inteligente, com menor risco e melhor uso do orçamento comercial.

12.REFERÊNCIAS

HARRIS, C. R. et al. Array programming with NumPy. Nature, v. 585, p. 357–362, 2020.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, v. 9, n. 3, p. 90–95, 2007.

McKINNEY, W. Python for Data Analysis. 3. ed. O’Reilly Media, 2022.

PANDAS DEVELOPMENT TEAM. Pandas Documentation. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 20 set. 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, v. 17, p. 261–272, 2020.

WASKOM, M. L. Seaborn: Statistical Data Visualization. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 set. 2025.

ROSSUM, G. van; DRAKE, F. L. Python 3 Reference Manual. Scotts Valley: CreateSpace, 2009.

ZANELLA, M. Expressões Regulares: uma abordagem prática. São Paulo: Novatec, 2018.