

Projeto em Business Intelligence e Analytics
Fase 2

Nome do estudante: Rafael de Menezes Ehlers

1. Métricas para o negócio proposto (KPIs):

O e-commerce brasileiro movimentou R\$ 204,3 bilhões em 2024 (ABComm, 2025), consolidando-se como um setor maduro em que pequenas variações operacionais geram impactos financeiros substanciais. A Olist opera como um dos principais agregadores de pequenos e médios sellers do país, o que torna seu ecossistema um laboratório natural para o estudo de indicadores de desempenho em marketplaces multi-vendedor. Os KPIs definidos para o ecossistema Olist foram organizados em quatro categorias estratégicas: financeiros, operacionais, de satisfação e diagnósticos. Cada métrica é calculável diretamente a partir do dataset público da Olist e reproduzível por qualquer instância do pipeline de ETL, garantindo consistência analítica entre dashboards e modelo preditivo.

Indicadores Financeiros

1. Faturamento Total e Mensal

- **Definição:** receita bruta gerada por pedidos em um período.
- **Fórmula:** $\sum (\text{price} + \text{freight_value})$, agrupado por mês e ano.
- **Relevância:** principal indicador de saúde financeira do marketplace e base para análise de tendência.
- **Reprodutibilidade:** cálculo determinístico sobre `order_items` e `orders`, sem dependência externa.

2. Ticket Médio

- **Definição:** valor médio capturado por transação.
- **Fórmula:** Faturamento Total / Número de Pedidos Únicos.
- **Relevância:** impacta diretamente o Lifetime Value (LTV) e a sustentabilidade do CAC.
- **Reprodutibilidade:** agregação simples entre `order_items` e `orders`.

Indicadores Operacionais (Logística)

3. Prazo Médio de Entrega

- **Definição:** intervalo médio entre compra e recebimento pelo cliente.

- **Fórmula:** Média(`order_delivered_customer_date` - `order_purchase_timestamp`), em dias.
- **Relevância:** principal driver de satisfação no e-commerce brasileiro, conforme literatura recente (Wangkiat e Polprasert, 2023; Zaghloul et al., 2024).
- **Reprodutibilidade:** cálculo direto entre dois timestamps obrigatórios em `orders`.

4. Taxa de Atraso

- **Definição:** percentual de pedidos entregues após a data estimada comunicada ao cliente.
- **Fórmula:** $\text{COUNT}(\text{order_delivered_customer_date} > \text{order_estimated_delivery_date}) / \text{COUNT}(\text{pedidos entregues}) \times 100$.
- **Relevância:** variável central da Hipótese 1; correlaciona-se diretamente com risco de churn (Luizon, 2019).
- **Reprodutibilidade:** dois campos timestamp já presentes no dataset, sem necessidade de joins externos.

Indicadores de Satisfação

5. NPS Estimado

- **Definição:** balanço entre promotores e detratores, derivado da nota de review.
- **Fórmula:** (% de notas 5) - (% de notas 1, 2 e 3).
- **Relevância:** proxy direto de lealdade e probabilidade de recompra.
- **Reprodutibilidade:** agregação determinística sobre `order_reviews`.

6. Taxa de Avaliações Negativas

- **Definição:** percentual de pedidos com nota baixa; constitui a variável-alvo do modelo preditivo.
- **Fórmula:** $\text{COUNT}(\text{review_score} \leq 3) / \text{COUNT}(\text{pedidos com review}) \times 100$.
- **Relevância:** alvo central da análise preditiva e métrica de monitoramento contínuo da experiência do cliente.
- **Reprodutibilidade:** contagem booleana direta sobre `order_reviews`.

Indicadores Diagnósticos (Suporte às Hipóteses)

7. Razão Frete-Preço Média

- **Definição:** peso percentual do frete em relação ao preço do produto.
- **Fórmula:** Média(`freight_value` / `price`), agrupado por categoria e faixa de ticket (baixo, médio, alto).
- **Relevância:** variável central da Hipótese 2; identifica categorias em que o frete penaliza desproporcionalmente a percepção de valor.

- Reprodutibilidade: cálculo determinístico sobre `order_items` e `products`.

8. Taxa de Pedidos Multi-Vendedor

- **Definição:** percentual de pedidos com mais de um seller distinto envolvido.
- **Fórmula:** $\text{COUNT}(\text{pedidos com COUNT(DISTINCT seller_id) > 1}) / \text{COUNT}(\text{total de pedidos}) \times 100$.
- **Relevância:** variável central da Hipótese 3; identifica pedidos com coordenação entre múltiplos vendedores, segmento associado a taxa de avaliação negativa significativamente maior.
- Reprodutibilidade: agregação sobre `order_items` com contagem de vendedores únicos por `order_id`.

2. Arquitetura da solução (ferramentas, pipeline com as fases, repositórios, dependências):

Descreva aqui a arquitetura da solução de BI implementada. Adicione diagramas com os componentes, descreva as ferramentas e plataformas selecionadas, repositórios e eventuais dependências.

A arquitetura proposta para a Fase 2 foi projetada para operar inteiramente em ambiente cloud-native gratuito, com foco em reprodutibilidade, baixo custo operacional e separação clara entre a camada descritiva (Business Intelligence) e a camada preditiva (Machine Learning). A solução é orquestrada por três notebooks Python no Google Colab, que se comunicam por meio de um Data Warehouse local em SQLite estruturado em Star Schema (Kimball e Ross, 2013), com versionamento completo no GitHub.

2.1 Componentes da Arquitetura

A Figura 2.1 ilustra o pipeline em seis camadas sequenciais. Os dados brutos do dataset público da Olist (Camada 1) são processados em um notebook de ETL no Google Colab (Camada 2) que produz um Data Warehouse local em SQLite com modelagem dimensional (Camada 3). A partir do Data Warehouse, o fluxo se divide em duas trilhas paralelas: um notebook de Business Intelligence (Camada 4 esquerda) responsável pela geração de quatro dashboards interativos em Plotly, e um notebook de Machine Learning (Camada 4 direita) responsável pelo treinamento e avaliação de modelos de classificação supervisionada. Ambos os artefatos finais (Camada 5) convergem para o repositório GitHub (Camada 6), que centraliza notebooks, dados tratados, dashboards exportados e documentação do projeto.

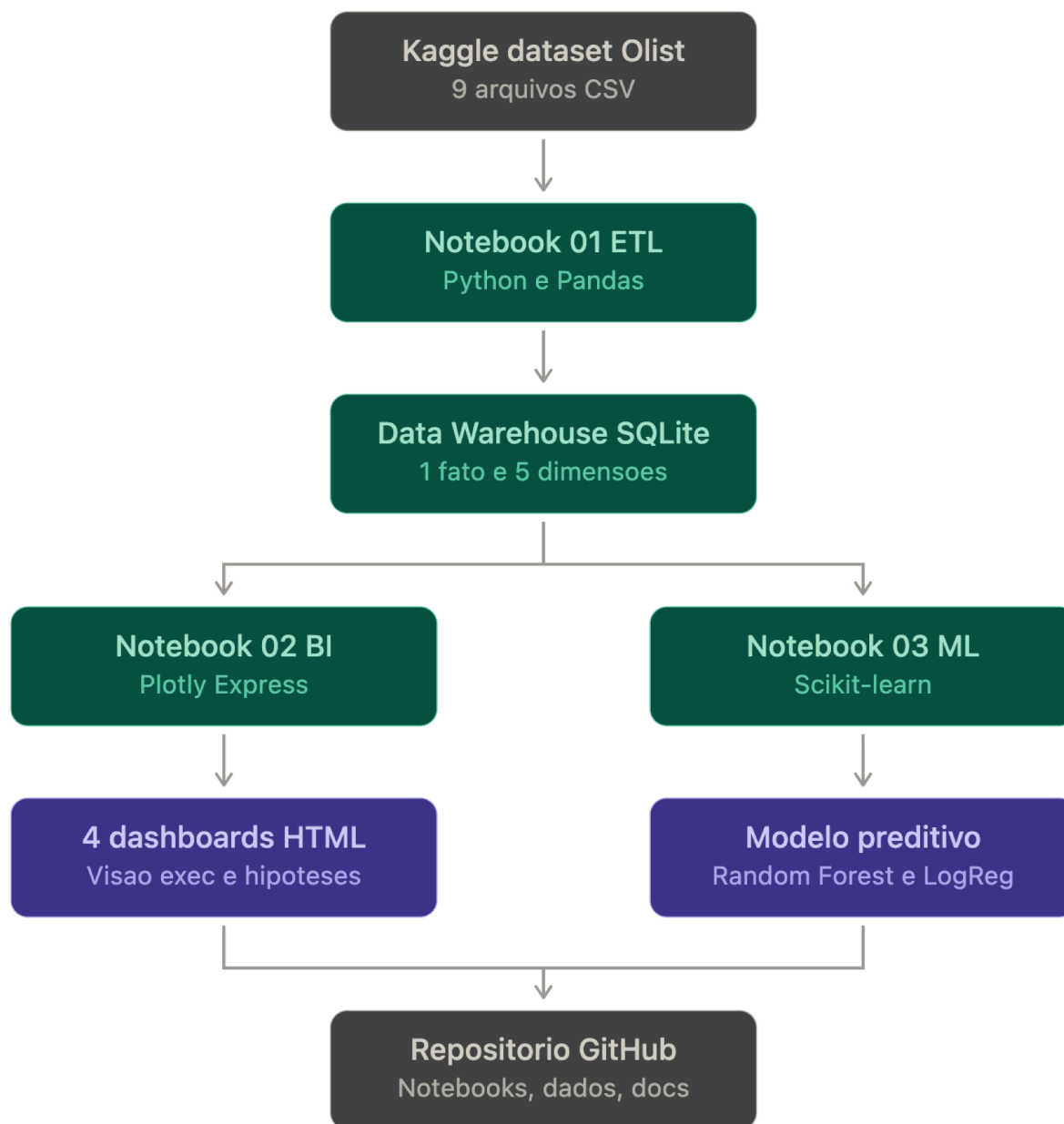


Figura 2.1, Pipeline de seis camadas da arquitetura proposta.
Cor cinza para fontes e repositório (externos), verde para a infraestrutura de processamento (notebooks e DW), e roxo para os artefatos analíticos finais.

2.2 Ferramentas e Plataformas Selecionadas

Tabela 2.1, Ferramentas e plataformas selecionadas

Camada	Ferramenta	Justificativa
Fonte de Dados	Kaggle (Brazilian E-Commerce Public Dataset by Olist)	Dataset benchmark validado pela literatura para estudos de e-commerce brasileiro multi-vendedor (Olist e Sionek, 2026)
Runtime e ETL	Google Colab com Python 3 e Jupyter	Ambiente cloud gratuito, sem instalação local; suporte nativo a Pandas, Scikit-learn e Plotly
Persistência de Sessão	Google Drive	Mantém o arquivo SQLite e os CSVs intermediários entre sessões do Colab
Armazenamento Analítico	SQLite com modelagem Star Schema	Banco SQL embutido, versionável diretamente no GitHub; capacidade suficiente para os 100 mil pedidos do dataset
Manipulação de Dados	Pandas, NumPy	Padrão de fato para ETL em Python; integração nativa com o stack do Scikit-learn
Visualização (BI)	Plotly Express e Plotly Graph Objects	Dashboards interativos renderizados inline no notebook e exportáveis como HTML standalone
Machine Learning	Scikit-learn	Implementações maduras de Regressão Logística e Random Forest, com métricas e pipelines integrados
Versionamento	GitHub	Repositório público centralizando notebooks, dashboards exportados e

		documentação
--	--	--------------

2.3 Pipeline com as Fases

O pipeline segue o framework CRISP-DM definido na Fase 1, organizado em três notebooks executáveis em sequência. Cada notebook implementa fases específicas do framework e produz artefatos persistidos que alimentam o notebook seguinte.

Notebook 01, [01_etl_e_star_schema.ipynb](#)

- **Fase CRISP-DM:** Entendimento dos Dados e Preparação dos Dados
- **Entrada:** nove arquivos CSV do dataset público da Olist (Kaggle)
- **Processamento:**
 - Carga, tratamento de nulos, deduplicação e padronização de tipos
 - Engenharia de features para teste das hipóteses: delta de entrega (dias), razão frete/preço, contagem de vendedores únicos por pedido, faixa de ticket
 - Estruturação dimensional em Star Schema (1 fato, 5 dimensões)
- **Saída:** arquivo [olist_dw.sqlite](#) contendo [fato_pedidos](#), [dim_cliente](#), [dim_produto](#), [dim_vendedor](#), [dim_tempo](#), [dim_geolocalizacao](#)

Notebook 02, [02_dashboards_bi.ipynb](#)

- **Fase CRISP-DM:** Avaliação e Implantação
- **Entrada:** Data Warehouse [olist_dw.sqlite](#) (via SQLAlchemy)
- **Processamento:** construção de quatro dashboards interativos em Plotly Express e Graph Objects:
 - Visão Executiva (KPIs gerais)
 - Dashboard H1 (Atrasos vs Avaliações Negativas)
 - Dashboard H2 (Frete vs Preço em Baixo Ticket)
 - Dashboard H3 (Pedidos Multi-Vendedor)
- **Saída:** quatro arquivos HTML standalone exportados para [/dashboards/](#), abriam fora do Colab por qualquer avaliador

Notebook 03, [03_modelo_preditivo.ipynb](#)

- **Fase CRISP-DM:** Modelagem e Avaliação
- **Entrada:** dados pré-processados do [olist_dw.sqlite](#)
- **Processamento:**
 - One-Hot Encoding de variáveis categóricas (categoria de produto, estado do cliente)
 - Divisão treino-teste 80/20 estratificada via [train_test_split](#)
 - Treinamento de Regressão Logística (baseline) e Random Forest (modelo principal) para a variável-alvo Avaliação Negativa

- Avaliação por acurácia, precisão, recall, F1-score, matriz de confusão e ROC-AUC
- Extração de feature importance e inferência exemplo em pedido sintético
- **Saída:** modelo persistido em `modelo_random_forest.pkl` e relatório de métricas

3. Solução de BI:

Descreva a solução implementada. Link para o código fonte, dicionário de dados, dashboards e análises realizadas. No Link de acesso para o repositório na plataforma GitHub apresente:

- Código fonte de preparação, coleta, limpeza, processamento de dados.
- Os Dashboards (visualizações) publicados na ferramenta escolhida.
- Documentação do código no GitHub.

A Seção 3 detalha a implementação da solução proposta, contemplando o código fonte versionado no GitHub, a análise dos quatro dashboards interativos gerados e os resultados da análise preditiva. Os modelos foram treinados sobre o conjunto de pedidos extraído do Star Schema descrito na Seção 2, com divisão estratificada em 80% para treino e 20% para teste.

3.1 Repositório e Código Fonte

Link do projeto: <https://rafaehlers.github.io/projetobi-olist-fase2/>

Código-fonte (GitHub): <https://github.com/rafaehlers/projetobi-olist-fase2>

Link do dataset utilizado: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

O repositório contém três notebooks Jupyter encadeados, cada um responsável por uma fase do pipeline CRISP-DM. A estrutura interna de cada notebook foi detalhada na Seção 2.3. A documentação de uso, incluindo instruções para reprodução completa do pipeline, está no arquivo README.md na raiz do repositório.

3.2 Dashboards Interativos

Os quatro dashboards foram organizados como uma narrativa progressiva: o Dashboard 0 estabelece o contexto macro do negócio, e os Dashboards 1, 2 e 3 testam cada uma das três hipóteses formuladas na Fase 1.

Dashboard 0, Visão Executiva

- **Objetivo:** estabelecer o contexto macro do negócio e a baseline contra a qual as análises subsequentes são comparadas.
- **Componentes:**

- Cartões com os KPIs estratégicos definidos na Seção 1: faturamento total, ticket médio, NPS estimado, taxa de atraso, taxa de avaliações negativas e percentual de pedidos multi-vendedor.
- Evolução temporal mensal do faturamento e do volume de pedidos no período 2016 a 2018.
- Concentração geográfica do faturamento por estado.
- Distribuição completa das notas de review.
- **Função na narrativa:** não testa hipótese; serve como referência comparativa para os três dashboards seguintes.

Dashboard 1, Hipótese 1: Atrasos e Satisfação

- **Objetivo:** testar a relação entre atrasos de entrega e taxa de avaliação negativa.
- **Gráfico-chave:** comparação direta da taxa de avaliação negativa entre pedidos no prazo (17,34%) e pedidos atrasados (73,28%), uma diferença de aproximadamente 56 pontos percentuais.
- **Visualizações de apoio:**
 - Histograma do delta de entrega: a maioria dos pedidos é entregue antes do prazo estimado, mas a cauda direita concentra os piores reviews.
 - Box plot do delta por nota de review: notas baixas correlacionam-se com deltas positivos crescentes (atrasos), notas altas concentram-se em deltas negativos (adiantamentos).
 - Curva segmentada em sete faixas de severidade: crescimento praticamente monotônico, partindo de 16,8% nos pedidos mais adiantados, atingindo o pico de 89,1% na faixa de 8 a 15 dias, com leve queda residual (87,2%) na cauda extrema acima de 15 dias.
- **Conclusão:** o padrão dose-resposta confirma a relação causal proposta em H1.

Dashboard 2, Hipótese 2: Frete vs Preço em Baixo Ticket

- **Objetivo:** verificar se a razão frete/preço, e não o frete absoluto, é o fator que penaliza a percepção de valor em produtos de baixo ticket.
- **Gráfico-chave:** razão frete/preço média por faixa de ticket: 0,57 no ticket baixo (< R\$ 50), 0,22 no médio (R\$ 50 a R\$ 200) e 0,11 no alto (> R\$ 200). Em produtos baratos, o frete pesa em média cinco vezes mais sobre o preço do que em produtos caros.
- **Visualizações de apoio:**
 - Heatmap cruzando faixa de ticket com faixa de razão frete/preço: efeito dose-resposta dentro de cada faixa, com a taxa de avaliação negativa crescendo monotonicamente conforme a razão aumenta.
 - Box plot da razão frete/preço por faixa: em ticket baixo, a razão pode superar 100%, situação em que o cliente paga mais pelo frete do que pelo próprio produto.
 - Ranking das categorias de baixo ticket com maior taxa de avaliação negativa, orientando possíveis ações operacionais.

- **Conclusão:** H2 confirmada; é o frete relativo ao preço, não o absoluto, que pesa na percepção de valor.

Dashboard 3, Hipótese 3: Pedidos Multi-Vendedor

- **Objetivo:** testar se pedidos com múltiplos sellers apresentam taxa de avaliação negativa significativamente maior e investigar o mecanismo causal via consolidação logística.
- **Gráfico-chave:** comparação da taxa de avaliação negativa entre pedidos single-seller (20,54%) e multi-seller (60,51%), uma diferença de aproximadamente 40 pontos percentuais.
- **Visualizações de apoio:**
 - Distribuição em escala logarítmica: pedidos multi-vendedor representam apenas 1,3% do volume total (1.261 de 95.824 pedidos), mas constituem um segmento operacionalmente crítico pela severidade do impacto na satisfação.
 - Box plot do tempo de consolidação logística por tipo de pedido: achado contraintuitivo. Pedidos multi-vendedor apresentam tempo de consolidação médio inferior aos single-seller (1,38 vs 2,30 dias), o que refuta a hipótese de canal indireto via consolidação.
- **Conclusão:** H3 parcialmente confirmada. A associação é forte e real, mas o mecanismo causal exato permanece em aberto. Candidatos prováveis: entregas fragmentadas, não sincronia entre vendedores e diferenças no mix de categoria.

3.3 Análise Preditiva

Foram treinados dois modelos de classificação binária (Provost e Fawcett, 2016) para a variável-alvo `avaliacao_negativa`: Regressão Logística (baseline) e Random Forest (modelo principal) (Wangkiat e Polprasert, 2023; Zaghloul et al., 2024), com divisão estratificada 80/20 e `class_weight='balanced'` para compensar o desbalanceamento das classes.

Métricas de desempenho:

Tabela 3.1, Métricas de desempenho dos modelos

Métrica	Regressão Logística	Random Forest	Diferença (RF – LR)
Acurácia	0,7550	0,7887	+0,0337
Precisão	0,4257	0,4975	+0,0718
Recall	0,4704	0,3975	-0,0730
F1-score	0,4469	0,4419	-0,0050

ROC-AUC	0,6983	0,7024	+0,0041
---------	--------	---------------	---------

Trade-off entre os modelos. O Random Forest entrega maior acurácia e precisão; a Regressão Logística captura mais avaliações negativas (recall superior). A escolha do modelo depende do uso pretendido: para priorização de inspeções (precisão), usar Random Forest; para sistemas de alerta (recall), usar Regressão Logística.

Análise da Importância das Features.

O ranking de feature importance do Random Forest evidencia três clusters distintos, alinhados às três hipóteses da Fase 1:

- **Cluster H1 (atrasos):** `delta_entrega_dias` (0,24), `tempo_total_entrega_dias` (0,17) e `atraso` (0,12) somam ~52% da importância total, confirmando H1 com larga margem.
- **Cluster H2 (frete):** `valor_total_frete` (0,07) e `razao_frete_preco` (0,06) em 4º e 5º lugares, validando que é o frete relativo ao preço, não o absoluto, que pesa na percepção de valor.
- **Cluster H3 (multi-vendedor):** `tempo_consolidacao_dias` em 6º lugar (0,06); `multi_vendedor` (0,02) e `num_sellers_unicos` (0,02) atrás. A leitura inicial sugeria efeito via tempo de consolidação, mas a análise descritiva refutou (multi-vendor consolida mais rápido, 1,38 vs 2,30 dias). H3 fica parcialmente confirmada: associação forte (60,5% vs 20,5%), mas mecanismo causal em aberto.

Variáveis de controle (estados, categorias individuais) aparecem nas posições finais do ranking, indicando que o modelo se apoia em variáveis operacionais e logísticas acionáveis pelo negócio, não em viés geográfico ou de categoria.

4. Considerações Finais

Esta seção consolida os achados da pesquisa, demonstra a aplicação prática do modelo treinado e discute as limitações da abordagem com seus respectivos trabalhos futuros.

4.1 Síntese dos Achados

A pergunta de pesquisa formulada na Fase 1 questiona quais fatores operacionais, logísticos e financeiros possuem maior peso estatístico na determinação de uma avaliação negativa no e-commerce brasileiro. Os resultados da análise descritiva combinada com a análise preditiva permitem responder essa pergunta em três blocos, alinhados às três hipóteses originais.

Tempo de entrega como determinante principal. O cluster de variáveis ligadas a atrasos (delta de entrega, tempo total de entrega e flag de atraso) responde por

aproximadamente 52% da capacidade preditiva do Random Forest. A taxa de avaliação negativa salta de 17,34% nos pedidos no prazo para 73,28% nos atrasados, com a curva de severidade atingindo 89,1% na faixa de 8 a 15 dias. A Hipótese 1 está confirmada com larga margem, e o achado é convergente com a literatura recente (Wangkiat e Polprasert, 2023; Zaghloul et al., 2024).

Razão frete/preço como driver secundário. Em produtos de baixo ticket (< R\$ 50), a razão frete/preço média é de 0,57, podendo superar 100% em situações em que o cliente paga mais pelo frete do que pelo próprio produto. Valor de frete e razão frete/preço ocupam o quarto e o quinto lugar no ranking de feature importance, validando o mecanismo proposto: não é o frete absoluto, e sim o frete relativo ao preço, que penaliza a percepção de valor. Hipótese 2 confirmada.

Pedidos multi-vendedor com mecanismo causal em aberto. Pedidos com múltiplos sellers apresentam taxa de avaliação negativa três vezes maior que pedidos single-seller (60,51% vs 20,54%), constituindo o segundo achado mais expressivo em termos de magnitude. No entanto, a análise descritiva refutou o canal causal proposto originalmente: o tempo de consolidação logística é menor em pedidos multi-vendedor (1,38 vs 2,30 dias), não maior. A Hipótese 3 fica parcialmente confirmada, com a associação validada estatisticamente e o mecanismo causal exato em aberto para investigação futura.

4.2 Aplicação Prática

Para demonstrar o uso operacional do modelo, foram construídos dois cenários sintéticos contrastantes, submetidos à inferência do Random Forest treinado.

Cenário A, baixo risco (probabilidade de avaliação negativa: 37,3%)

- Entrega 5 dias adiantada em relação ao prazo estimado
- Ticket alto (R\$ 350)
- Frete leve (R\$ 15)
- Single-vendedor
- Cliente em São Paulo

Cenário B, alto risco (probabilidade de avaliação negativa: 74,7%)

- Atraso de 12 dias em relação ao prazo estimado
- Ticket baixo (R\$ 30)
- Frete pesado (R\$ 25, razão frete/preço de 0,83)
- Multi-vendedor
- Cliente no Amazonas

A diferença de 37,5 pontos percentuais entre os dois cenários demonstra a capacidade do modelo de distinguir perfis operacionais a partir de combinações realistas de variáveis. Em produção, isso viabiliza monitoramento preventivo de risco de churn em tempo real: pedidos classificados como de alto risco podem ser direcionados para fluxos de

inspeção logística, comunicação proativa com o cliente ou compensações antecipadas (cupons, frete grátis em compra futura), antes que a avaliação negativa seja registrada.

4.3 Limitações e Trabalhos Futuros

Limitações da abordagem atual

- ROC-AUC de aproximadamente 0,70, indicando que há sinal preditivo não capturado pelas features estruturadas disponíveis.
- Recall do Random Forest de 39,75%, o que significa que o modelo deixa passar mais da metade dos pedidos efetivamente classificados como negativos.
- Ausência de análise textual sobre os comentários das reviews, descartando potencial sinal qualitativo relevante.
- Dataset estático cobrindo o período 2016 a 2018, sem capturar tendências de comportamento pós-pandemia nem mudanças no perfil logístico do e-commerce brasileiro.

Direções para trabalhos futuros

- Sentiment analysis sobre o texto das reviews via técnicas de Processamento de Linguagem Natural, ampliando o conjunto de features preditivas.
- Incorporação de variáveis geoespaciais, em especial a distância entre vendedor e cliente, como preditor adicional do tempo de entrega e da satisfação.
- Modelagem em janelas temporais móveis, permitindo o monitoramento dinâmico de risco de churn ao longo do tempo, em vez do snapshot estático atual.
- Investigação dirigida do mecanismo causal por trás do efeito multi-vendedor (Hipótese 3), explorando candidatos como entregas fragmentadas, dessincronia entre sellers e diferenças de mix de categoria.
- Generalização da abordagem para outros marketplaces brasileiros, validando a transferibilidade dos resultados encontrados no ecossistema Olist.

Referências Bibliográficas

Associação Brasileira de Comércio Eletrônico. “E-commerce fatura R\$ 204,3 bilhões no Brasil em 2024”. Disponível em: <https://www.ecommercebrasil.com.br/noticias/e-commerce-resultados-2024-brasil-abcomm>, Maio 2026.

Kimball, R.; Ross, M. “The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling”. Indianapolis: John Wiley & Sons, 2013, 3 ed., 600.

Luizon, B.F.G. “Análise preditiva de churn em um e-commerce”, Trabalho de conclusão de curso (graduação em gestão da informação), Universidade Federal do Paraná, Curitiba, Brasil, 2019, disponível em: <https://hdl.handle.net/1884/64983>. Acesso em: maio 2026.

Olist; Sionek, A. "Brazilian E-Commerce Public Dataset by Olist". Disponível em: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>, Maio 2026.

Provost, F.; Fawcett, T. "Data Science para Negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados". Rio de Janeiro: Alta Books, 2016, 408.

Wangkiat, P.; Polprasert, C. "Machine Learning Approach to Predict E-commerce Customer Satisfaction Score". In:: 2023 8th International Conference on Business and Industrial Research (ICBIR), 2023, p. 1176–1181.

Zaghloul, M.; Barakat, S.; Rezk, A. "Predicting e-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches", Journal of Retailing and Consumer Services, v. 79, 2024, p. 103865.