Determinantes de Satisfação no E-Commerce Brasileiro: Uma Análise sobre o Dataset Olist

Carlos Alberto Mota Da Silva Filho[1], **Sidney Vitor Melo Do Nascimento**[1]

[1] Universidade do Estado do Amazonas - UEA
Manaus – AM – Brasil

camdsf.cid25@uea.edu.br, svmdn.cid25@uea.edu.br

Abstract. *This report presents machine-learning models for customer satisfaction prediction on the Olist dataset. We tackle binary classification (bad $\leq$ 2 vs good $\geq$ 4) and continuous regression (1–5). Stratified cross-validation and holdout evaluation are used, along with feature importance for interpretability. Random Forest and XGBoost achieve higher overall accuracy but low recall for the "bad" class; SVM/LogReg improve recall. Regression performance is modest ($R^2 \approx 0.15$). We outline limitations, future improvements, and a Streamlit app for interactive exploration and prediction.*

Resumo. *Este relatório descreve a construção, avaliação e interpretação de modelos de aprendizado de máquina para prever a satisfação do cliente (*`review_score`*) no conjunto de dados Olist. Implementamos duas frentes: classificação binária (ruim $\leq$ 2 vs bom $\geq$ 4) e regressão contínua (1–5). Utilizamos validação cruzada estratificada, holdout final e interpretabilidade via importância de atributos. Resultados mostram que Random Forest e XGBoost atingem maior acurácia geral, porém com baixo* recall *da classe "ruim"; SVM/LogReg equilibram melhor esse* recall*. Em regressão, o desempenho é limitado ($R^2 \approx 0{,}15$). Discutimos limitações, melhorias futuras e o uso de um aplicativo Streamlit para visualização e predição interativa.*

1. Introdução

A satisfação do cliente é um elemento central para a competitividade no comércio eletrônico. Avaliações negativas podem sinalizar falhas na experiência de compra, afetar a reputação das plataformas e contribuir para a não fidelização do consumidor com as empresas. Nesse cenário, compreender os fatores que influenciam a percepção do consumidor tornou-se fundamental para empresas que buscam aprimorar processos logísticos, operacionais e de atendimento [1].

O dataset público da Olist, amplamente utilizado em pesquisas acadêmicas e competições de ciência de dados, oferece uma oportunidade robusta para investigar determinantes de satisfação no e-commerce brasileiro. Reunindo informações sobre pedidos, entregas, pagamentos, produtos e avaliações, o conjunto de dados permite examinar como características do pedido, do produto e da logística se relacionam com a nota atribuída pelo cliente.

Este trabalho tem como objetivo identificar os principais fatores associados à satisfação do consumidor e desenvolver modelos capazes de prever avaliações com base

em atributos dos pedidos. São combinadas análises exploratórias, técnicas de aprendizado de máquina e métodos de agrupamento para revelar padrões relevantes e compreender como diferentes aspectos influenciam a percepção do usuário.

2. Descrição do Problema

A análise da satisfação do cliente no comércio eletrônico envolve múltiplos fatores que podem influenciar a percepção do consumidor sobre sua experiência de compra. No contexto do dataset Olist, o objetivo central é compreender como diferentes atributos do pedido, do produto e do processo logístico se relacionam com a nota atribuída no *review*.

Dessa forma, esta pesquisa é guiada por quatro perguntas fundamentais:

- **Quais características dos pedidos estão relacionadas a avaliações altas ou baixas (`review_score`)?**
  Busca-se identificar atributos estruturais do pedido (valor, número de itens, tipo de pagamento, entre outros) que possam explicar variações na satisfação.
- **O tempo entre a compra e a entrega (`order_purchase_timestamp`, `order_delivered_customer_date`) afeta a nota do review?**
  Avalia-se se atrasos ou adiantamentos logísticos têm impacto significativo na percepção do cliente.
- **Existem categorias de produtos (`product_category_name`) associadas a maior insatisfação dos clientes?**
  Busca-se identificar grupos de produtos que apresentam maior concentração de avaliações negativas.
- **Clientes de determinadas regiões (`customer_state`, `customer_city`) avaliam de maneira mais positiva ou negativa?**
  Examina-se se aspectos geográficos influenciam padrões de avaliação.

Com base nessas questões, o estudo combina análises exploratórias, técnicas de aprendizado de máquina e métodos de agrupamento para identificar determinantes de satisfação e avaliar a capacidade de modelos preditivos em antecipar avaliações negativas.

3. Base de dados

- **Origem**: conjunto Olist (Kaggle [3]) contendo 112.650 registros e 47 variáveis inicialmente, distribuídos em 9 tabelas relacionais (pedidos, itens, reviews, clientes, produtos, pagamentos, vendedores e geolocalização), reduzidos para 98.068 registros e 34 features após pré-processamento.
- **Consolidação**: a função `preprocess_base` (`src/preprocessing.py`) gera um *dataframe* por pedido, remove cancelados, agrega itens/pagamentos e traduz categorias.
- **Targets**: a função `add_targets` (`src/feature_engineering.py`) cria `review_binary`, `review_positive` e `review_negative`.

Features usadas:

- **Numéricas** (imputação mediana, *clip* 1–99%, Min-Max): `total_items_price`, `total_freight_value`, `payment_installments`, `payment_value`, `n_items`, `delivery_time_days`, `estimated_delivery_days`, `delivery_delay_days`, `review_count`.

- **Categóricas** (imputação moda + OneHot): `payment_type`, `customer_state`, `product_category_name_english`.

Gráficos exploratórios (script `01_data_overview.py`):

- Distribuição de notas (Figura 1).
- Pedidos por mês (Figura 2).
- Atraso versus nota (Figura 3).
- Análises por categoria e estado (Figura 4).

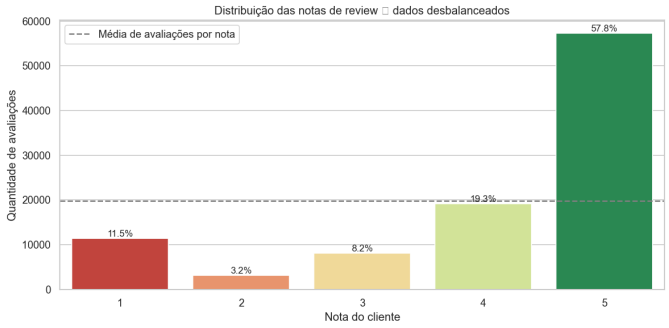

Figura 1. Distribuição das notas de satisfação (`review_score`).

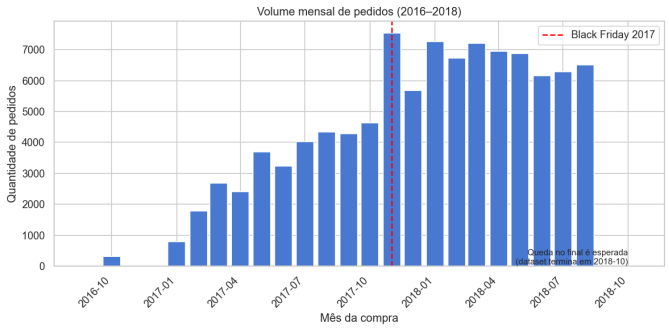

Figura 2. Número de pedidos por mês.

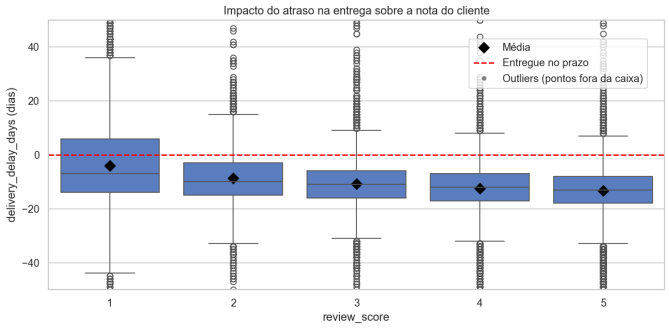

Figura 3. Relação entre atraso de entrega e nota do review.

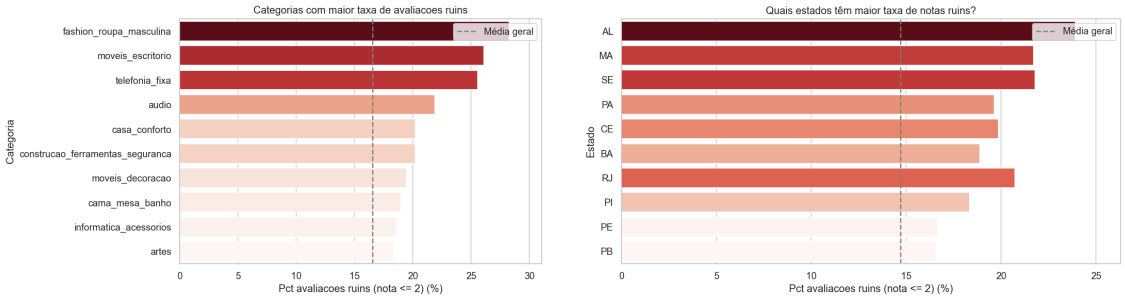

Figura 4. Análises de insatisfação por categoria de produto (esquerda) e por estado (direita).

4. Trabalhos relacionados

- Estudos de previsão de satisfação em e-commerce têm utilizado principalmente modelos como *random forest* e *gradient boosting*, em dados de avaliações e atributos de pedidos, com resultados competitivos em métricas de classificação [5, 4].
- Abordagens para desbalanceamento incluem reamostragem, ajuste de *threshold* e penalização de classe, alinhadas ao que exploramos aqui (pesos de classe e otimização por *recall*). [2]

5. Metodologia

5.1. Pré-processamento

- *Clipping* de outliers nos quantis 1–99%.
- Imputação de faltantes.
- Padronização das variáveis numéricas.
- OneHot denso para variáveis categóricas.
- Semente global fixa (`set_global_seed`) para reprodutibilidade.

5.2. Modelagem

- **Classificação**: Naive Bayes (baseline), Regressão Logística (SGD), SVM linear, Random Forest, XGBoost.
- **Regressão**: modelos Linear e Ridge (`SGDRegressor`).
- **Hiperparâmetros**: `RandomizedSearchCV` opcional para Random Forest e XGBoost com métrica alvo `recall_weighted`.
- **Validação**: validação cruzada estratificada (3 *folds*) para comparação preliminar e *holdout* 80/20 estratificado para métricas finais.

5.3. Avaliação

- **Métricas de classificação**: *accuracy*, *precision*/*recall*/F1 (ponderadas e por classe), matrizes de confusão, importâncias de atributos.
- **Métricas de regressão**: RMSE, R^2; gráficos de paridade e de resíduos para avaliar viés e dispersão das previsões.
- **Visualizações principais**: as Figuras 5 a 8 ilustram o desempenho dos modelos de classificação e os atributos mais relevantes.

6. Resultados

6.1. Validação cruzada (3 folds)

A validação cruzada estratificada em 3 *folds* foi utilizada para comparar modelos de classificação com base em métricas ponderadas (accuracy, precision, recall, F1). O heatmap de métricas (Figura 6) sintetiza o desempenho médio dos modelos, destacando *Random Forest* e XGBoost com melhores resultados globais, enquanto Naive Bayes apresenta desempenho substancialmente inferior.

6.2. Holdout (20%)

No conjunto de teste (holdout 20%), as acurácias aproximadas foram: RF $\approx 0{,}893$, XGB $\approx 0{,}893$, SVM $\approx 0{,}80$, LogReg $\approx 0{,}76$ e NB $\approx 0{,}73$. A Figura 5 resume essa comparação.

Apesar da boa acurácia geral, a classe 0 (avaliações ruins) apresenta *recall* baixo em RF/XGB ($\sim 0{,}32$), enquanto SVM e LogReg alcançam *recall* maior ($\sim 0{,}57$–$0{,}61$), ao custo de mais falsos positivos. A Figura 7 mostra a matriz de confusão do modelo selecionado, evidenciando esse comportamento. A Figura 8 apresenta as importâncias de atributos, destacando variáveis de tempo de entrega, valor do pedido e características de produto como determinantes relevantes.

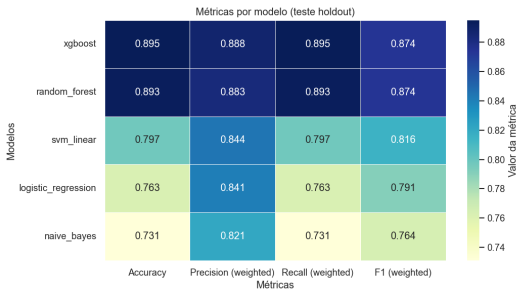

Figura 5. Acurácia dos modelos no conjunto de teste (holdout).

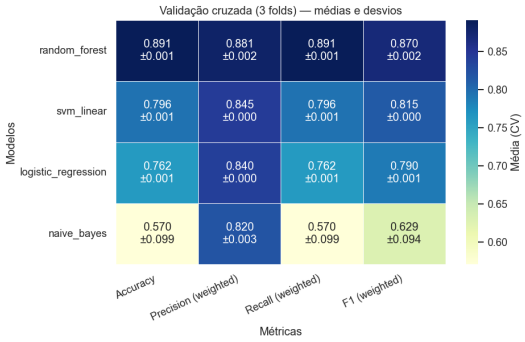

Figura 6. Métricas em validação cruzada (heatmap).

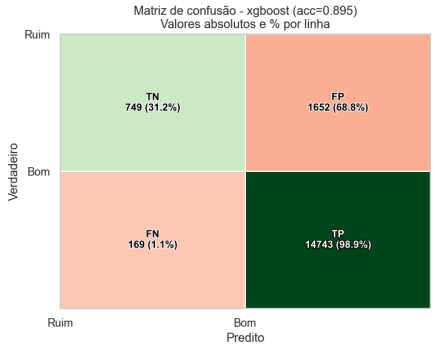

Figura 7. Matriz de confusão do modelo selecionado (XGBoost).

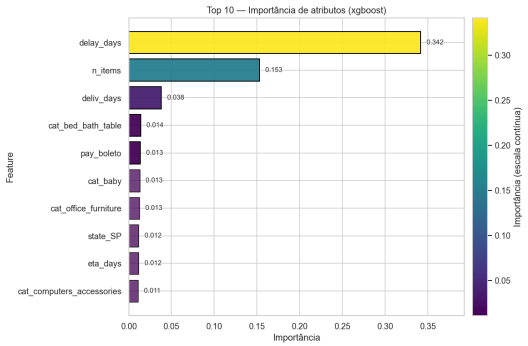

Figura 8. Importância das *features* no melhor modelo.

6.3. Regressão

Na tarefa de regressão para prever `review_score` na escala de 1 a 5, os modelos Linear e Ridge apresentaram RMSE em torno de 1,23 e $R^2 \approx 0{,}15$, indicando capacidade limitada de explicar a variabilidade das notas. As análises de paridade e de resíduos mostram concentração em torno da média, com achatamento nas pontas e viés nas extremidades (tendência a superestimar notas muito baixas e subestimar notas muito altas), sugerindo que as *features* tabulares disponíveis não são suficientes para capturar nuances da percepção de satisfação.

6.4. Clustering (diagnóstico)

Como análise complementar, foi aplicado K-Means sobre *features* monetárias e de tempo, após transformações com `log1p` e *winsorização* 1–99%. Três *clusters* principais foram identificados: (i) pedidos de baixo valor, com entrega adiantada e maior média de *review*; (ii) pedidos de valor médio–alto, com um produto e entrega adiantada; e (iii) pedidos com múltiplos produtos e valores intermediários, associados a avaliações ligeiramente piores. Esses grupos fornecem indícios de segmentação útil, ainda que sem validação externa formal.

7. Discussão

- **Trade-off entre *recall* e acurácia**: RF/XGB otimizam acurácia geral, mas perdem avaliações "ruins"; SVM/LogReg recuperam mais casos críticos ao custo de mais falsos positivos.
- **Desbalanceamento**: métricas ponderadas refletem a dominância da classe positiva. Ajuste de pesos e *threshold* é essencial para o objetivo de negócio.
- **Regressão limitada**: o baixo R^2 indica que as *features* atuais não capturam bem a variabilidade das notas; é útil como diagnóstico, não como predição fina.
- **Interpretabilidade**: importâncias apontam que atraso de entrega, valores monetários e categoria influenciam a predição; matrizes de confusão mostram onde o modelo erra (FP/FN).

8. Comparação com resultados existentes

- Trabalhos públicos com Olist (Kaggle/competições) reportam bons resultados com modelos de árvore/boosting para classificação; nossos resultados são consistentes (acurácia $\sim 0{,}89$), mas evidenciam a mesma dificuldade em *recall* da classe minoritária.
- Em regressão, valores baixos de R^2 também aparecem em estudos correlatos, reforçando a dificuldade de prever a nota exata sem *features* adicionais (como texto de reviews ou histórico detalhado).

9. Limitações

- Desbalanceamento forte, mitigado apenas parcialmente com pesos/*threshold*; não exploramos reamostragem nem custos customizados no *holdout* final.
- *Features* limitadas à base tabular; ausência de texto dos reviews e histórico detalhado de clientes/produtos.
- Regressão restrita a modelos lineares; não avaliamos GBDT regressivos ou redes neurais.
- Clustering exploratório sem validação externa (apenas interpretação descritiva).

10. Conclusões

- Modelos de classificação atingem alta acurácia, mas o *recall* da classe "ruim" permanece o principal desafio.
- SVM/LogReg são preferíveis quando o custo de perder insatisfeitos é alto; RF/XGB são mais interessantes quando se prioriza acerto geral.
- Regressão não captura bem a variabilidade das notas ($R^2 \approx 0{,}15$); é mais adequada para sinalizar tendência central do que para predição exata.
- Ferramentas de interpretação (importâncias, matrizes de confusão) e o app Streamlit facilitam a comunicação dos resultados para stakeholders.

11. Respostas às perguntas de negócio

- **Quais características se relacionam a avaliações altas/baixas?** As importâncias do melhor modelo (Figura 8) destacam atraso de entrega, valores monetários e características de produto como principais *drivers* de `review_binary`, em linha com os padrões observados nos gráficos exploratórios.
- **Tempo entre compra e entrega afeta a nota?** Sim. A relação entre atraso e nota (Figura 3) indica queda significativa na média de `review_score` quando há atraso na entrega, e variáveis de tempo aparecem entre as mais importantes no modelo.
- **Categorias com maior insatisfação?** A Figura 4 (painel de categorias) e as estatísticas associadas evidenciam grupos de produtos com maior concentração de notas ≤ 2, sugerindo segmentos críticos para intervenção.
- **Regiões mais positivas/negativas?** A Figura 4 (painel de estados) mostra diferenças entre unidades federativas, com alguns estados apresentando maior proporção de avaliações negativas, o que reforça o papel de `customer_state` como *feature* relevante.

12. Melhorias futuras

- **Balanceamento avançado**: SMOTE/undersampling e custo customizado por classe; escolha de *threshold* via curva *precision-recall*.
- **Otimização de Hiperparâmetros**: Utilizar Optuna para otimização de hiperparâmetros.
- **Novas features**: texto do review, histórico de compras/entregas, sazonalidade, interação produto–categoria.
- **Seleção de features**: Aplicar técnicas de seleção de atributos para eliminar variáveis irrelevantes e redundantes.
- **Redução de dimensionalidade (PCA)**: Avaliar o uso de PCA para redução de dimensionalidade e ruído, assim como seu impacto no desempenho dos modelos.
- **Modelos**: GBDT para regressão, calibração de probabilidades, modelos hierárquicos por categoria/região.
- **Clustering**: validar com métricas internas (por exemplo, *silhouette*) e cruzar *clusters* com NPS/retorno.

13. Disponibilidade de Códigos e dados

- O códig-fonte e os scripts utilizados nesse trabalho estão disponíveis em: `https://github.com/SidneyMelo/olistbr-ml2-final`

Referências

[1] S. Asawawibul, K. Na-Nan, K. Pinkajay, N. Jaturat, Y. Kittichotsatsawat, and B. Hu. The influence of cost on customer satisfaction in e-commerce logistics: The mediating role of service quality, technology use, transit time, and production conditions. *Journal of Open Innovation: Technology, Market, and Complexity*, 11(1):100482, 2025.

[2] H. M. Della-Justina. Avaliação de técnicas de classificação para dados desbalanceados, 2023. Orientador: Prof. Dr. Luiz Eduardo Soares de Oliveira.

[3] Olist. Brazilian e-commerce public dataset by olist. `https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce`. Acessado em 2025-12-01.

[4] P. Wangkiat and C. Polprasert. Machine learning approach to predict e-commerce customer satisfaction score. In *2023 8th International Conference on Business and Industrial Research (ICBIR)*, pages 1176–1181, 2023.

[5] M. Zaghloul, S. Barakat, and A. Rezk. Predicting e-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches. *Journal of Retailing and Consumer Services*, 79:103865, 2024.