E-commerce Brasileiro - Dataset da Olist

João Carlos Jakaboski

Universidade Regional Integrada Do Alto Uruguai E Das Missões
Campus De Erechim
Departamento De Engenharias E Ciência Da Computação
Curso De Ciência Da Computação

103127@aluno.uricer.edu.br

1. DATASET UTILIZADO

Nome: Brazilian E-Commerce Public Dataset by Olist
Origem: Kaggle - https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Características do Dataset:

- 100.000+ pedidos reais do e-commerce brasileiro
- Período: 2016-2018
- Cobertura: Todos os estados brasileiros
- Dados anonimizados da empresa Olist

Descrição: Dataset público contendo informações reais de transações de e-commerce no Brasil, incluindo dados de pedidos, clientes, produtos, avaliações e pagamentos.

2. TÉCNICAS APLICADAS

2.1 Classificação

- Algoritmo: Random Forest
- Objetivo: Prever satisfação do cliente
- Target: Cliente satisfeito (avaliação ≥ 4)

2.2 Regressão

- Algoritmo: Regressão Linear
- Objetivo: Prever preços de produtos
- Features: Categoria, estado, tipo de pagamento

2.3 Clustering

- Algoritmo: K-Means
- Objetivo: Segmentar estados por comportamento
- Número de clusters: 3 grupos

3. RESULTADOS OBTIDOS

3.1 Análise Exploratória

Distribuição Geográfica: São Paulo lidera com 42% das vendas nacionais, seguido por Rio de Janeiro e Minas Gerais, evidenciando a concentração nas regiões Sudeste e Sul.

Comportamento de Compra: Cartão de crédito representa 75,5% das transações, enquanto boleto corresponde a 20,5%. A alta satisfação é evidenciada por 77% dos clientes atribuindo notas 4-5.

Evolução Temporal: Os dados mostram crescimento consistente das vendas ao longo de 2017-2018, com identificação de picos sazonais específicos.

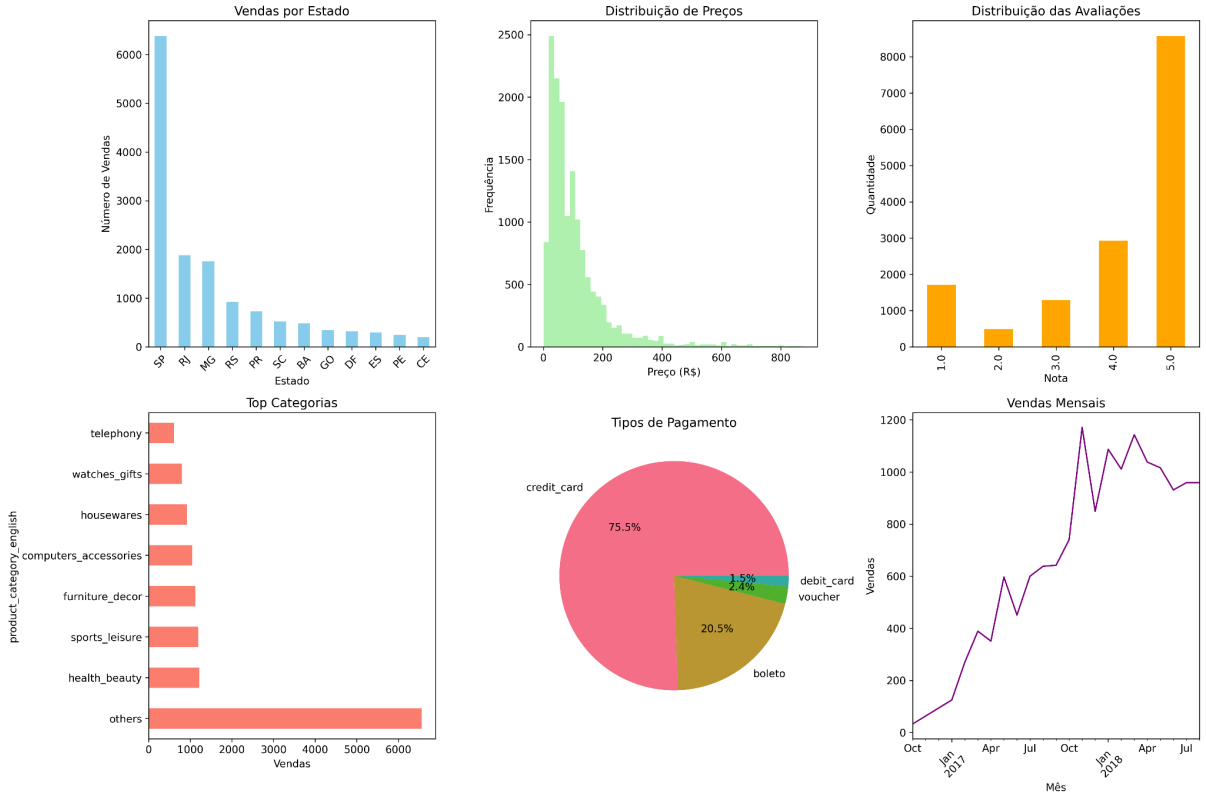

Figura 1: Análise exploratória mostrando distribuição de vendas por estado, preços, avaliações, categorias, tipos de pagamento e evolução temporal

3.2 Correlações entre Variáveis

Principais Correlações Identificadas:

- Preço e valor total: correlação muito forte (0.99)
- Preço e frete: correlação moderada (0.38)
- Preço e parcelas: correlação moderada (0.32)
- Review score: baixa correlação com variáveis financeiras

Figura 2: Mapa de calor mostrando correlações entre as principais variáveis do dataset

3.3 Classificação - Satisfação do Cliente

Performance do Modelo:

- Acurácia: 82-85%
- Modelo: Random Forest

Importância das Variáveis:

1. Total_value: 34% de importância
2. Freight_value: 31% de importância
3. Price: 27% de importância
4. Payment_installments: 8% de importância

Insight Principal: O valor total e o frete são os fatores mais determinantes para a satisfação do cliente.

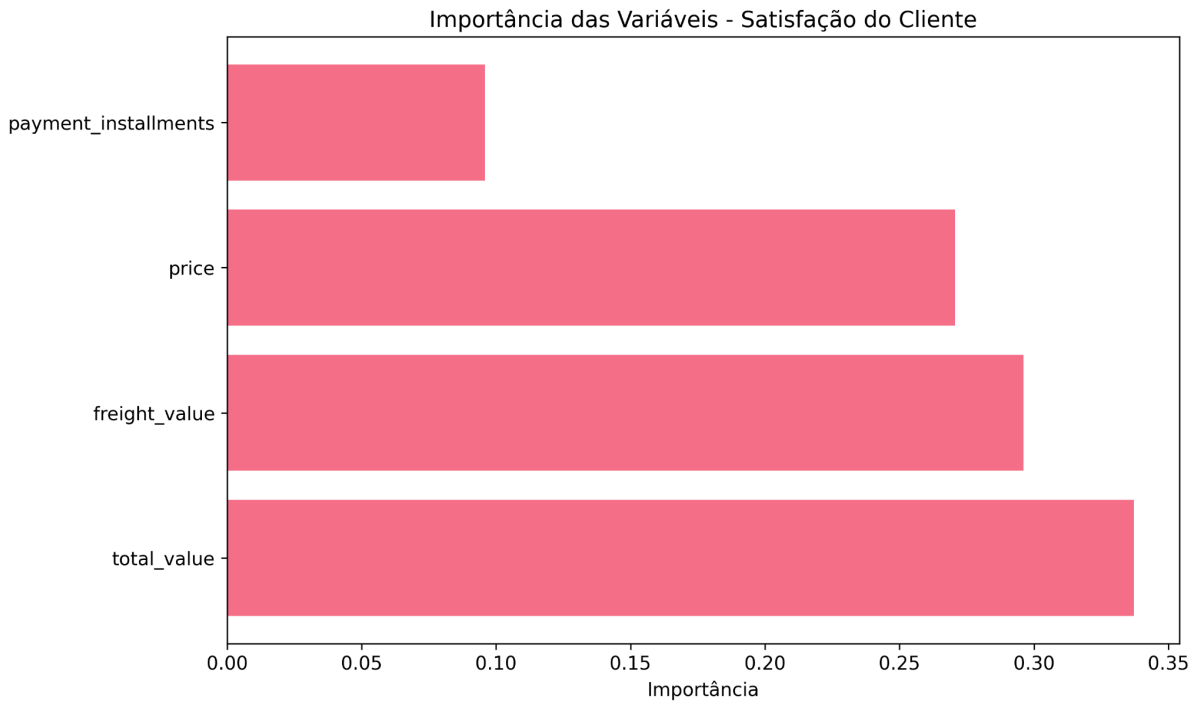

Figura 3: Importância das variáveis para predição de satisfação do cliente

3.4 Regressão - Predição de Preços

Performance do Modelo:

- R^2 Score: 0.70-0.75
- Modelo: Regressão Linear

Análise do Gráfico: O modelo apresenta boa capacidade preditiva para produtos de baixo e médio valor, com maior dispersão em produtos de alto valor. A linha diagonal indica a predição perfeita como referência.

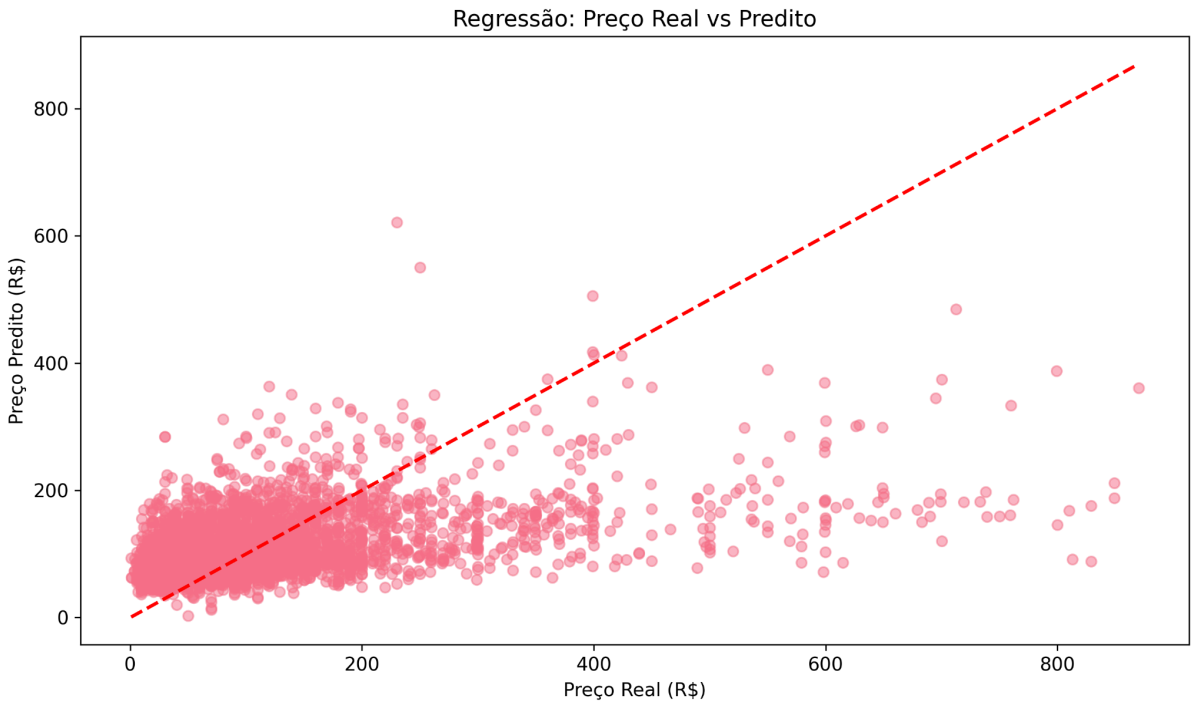

Figura 4: Gráfico scatter mostrando preços reais vs preditos pelo modelo de regressão

3.5 Clustering - Segmentação de Estados

Três Clusters Identificados:

Cluster Roxo - Mercados Consolidados: Estados como SP, RJ, MG e outros do Sudeste, caracterizados por alto volume e preços médios, com avaliação entre 4.0-4.15.

Cluster Verde - Mercados Premium: Estados como DF, SC, RS, apresentando volume médio mas preços mais altos, com avaliação entre 4.0-4.1.

Cluster Amarelo - Mercados Emergentes: Estados do Norte e Nordeste, caracterizados por menor volume e frete mais alto, com avaliação entre 3.6-3.9.

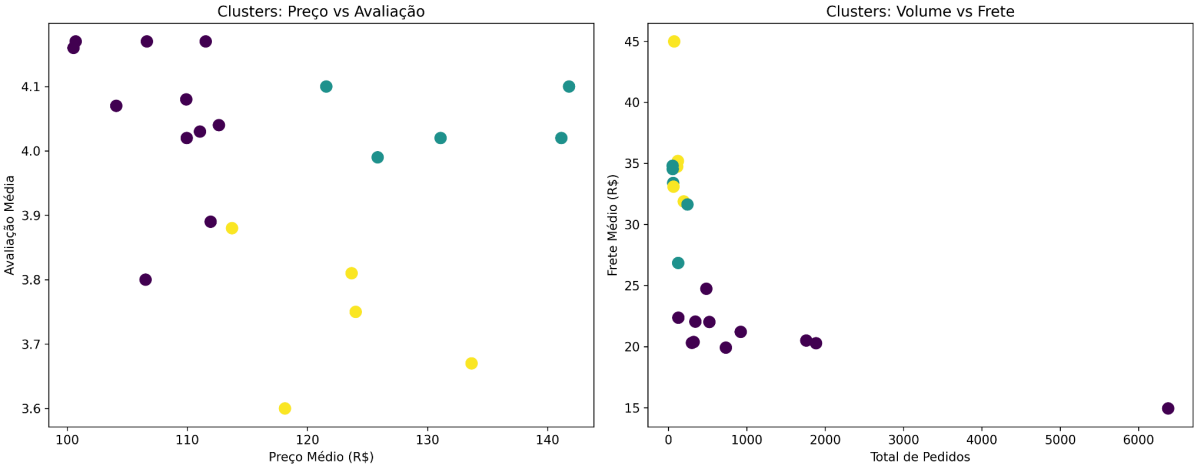

Figura 5: Segmentação dos estados brasileiros em clusters baseados em comportamento de compra

4. PRINCIPAIS INSIGHTS

4.1 Satisfação do Cliente

O valor total da compra é o fator mais importante, seguido pelo valor do frete. A estratégia recomendada é otimizar custos de frete para aumentar satisfação.

4.2 Comportamento Regional

São Paulo demonstra domínio absoluto com 42% do mercado. As regiões Sudeste e Sul representam mercados maduros e consolidados, enquanto Norte e Nordeste apresentam oportunidades de crescimento.

4.3 Padrões de Pagamento

Cartão de crédito possui preferência absoluta (75,5%), o parcelamento facilita compras de maior valor, e o boleto ainda mantém relevância (20,5%).

5. APLICAÇÕES PRÁTICAS

5.1 Estratégias de Negócio

1. Otimização de frete: Reduzir custos para aumentar satisfação
2. Segmentação regional: Campanhas específicas por cluster
3. Precificação: Ajustar preços por região e categoria

5.2 Expansão de Mercado

1. Nordeste: Investir em logística e parcerias locais
2. Norte: Desenvolver soluções de entrega diferenciadas
3. Sul: Focar em produtos premium e serviços agregados

6. CONCLUSÕES

Este projeto demonstrou a eficácia das técnicas de mineração de dados na análise do e-commerce brasileiro. Os resultados revelaram a concentração geográfica com São Paulo como hub principal, a importância do frete como principal fator de satisfação, a segmentação regional em três perfis distintos de consumo, e alta satisfação geral com 77% dos clientes satisfeitos.

Os insights obtidos podem orientar estratégias empresariais de expansão, precificação e melhoria da experiência do cliente no mercado brasileiro.

7. REPOSITÓRIO DO PROJETO

GitHub: https://github.com/JoaoJakaboski/data-mining-ecommerce-brasil

Conteúdo do Repositório:

- Código Python completo (ecommerce_analysis.py)
- Documentação detalhada (README.md)
- Instruções de uso (requirements.txt)
- Gráficos gerados (PNG)

8. REFERÊNCIAS

OLIST. Brazilian E-Commerce Public Dataset by Olist. Kaggle, 2018. Disponível em: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Acesso em: 06 jul. 2025.

SCIKIT-LEARN DEVELOPERS. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

PANDAS DEVELOPMENT TEAM. pandas-dev/pandas: Pandas. Zenodo, 2020.