Universidade Federal de Pelotas

Ferramentas de Inteligência Artificial

Precificação Dinâmica com Machine Learning

Autores: Lucas Aniceto e Rodrigo Santos

Email:
rcmdsantos@inf.ufpel.edu.br
lpsamiceto@inf.ufpel.edu.br

Resumo

O estabelecimento de preços adequados em plataformas de e-commerce é um desafio crítico que afeta diretamente a competitividade e rentabilidade dos negócios. Este trabalho apresenta um sistema inteligente de recomendação de preços baseado em técnicas de Machine Learning, desenvolvido e validado com dados reais do marketplace brasileiro Olist. O sistema utiliza algoritmos Random Forest e XGBoost com otimização automática de hiperparâmetros, incorporando engenharia de features avançada e técnicas de detecção de drift. Os resultados experimentais demonstram que o modelo Random Forest otimizado alcançou R² de 0,763, MAE de R$ 24,4 e MAPE de 35,8%, com níveis de confiança superiores a 85% para a maioria das categorias analisadas.

Palavras-chave: Machine Learning, E-commerce, Sistemas de Recomendação, Precificação Inteligente, Random Forest

1. Introdução

A definição estratégica de preços em marketplaces digitais representa um dos principais desafios para vendedores e gestores de e-commerce. Com o crescimento exponencial do comércio eletrônico brasileiro, que movimentou mais de R$ 180 bilhões em 2023, a necessidade de sistemas automatizados e inteligentes para auxiliar na precificação tornou-se imperativa.

Tradicionalmente, a definição de preços baseia-se em análises manuais de concorrência, custos operacionais e margens desejadas. Entretanto, essa abordagem apresenta limitações significativas em ambientes dinâmicos com milhares de produtos e constantes flutuações de mercado. Sistemas baseados em Machine Learning (ML) oferecem uma alternativa promissora, permitindo a análise de múltiplas variáveis simultaneamente e a adaptação automática a mudanças no comportamento do mercado.

Este trabalho propõe um sistema completo de recomendação de preços que integra técnicas avançadas de ML com monitoramento contínuo de performance e detecção de drift conceitual. O sistema foi desenvolvido e validado utilizando o dataset real do Olist, que contém mais de 100.000 transações de um marketplace brasileiro entre 2016 e 2018.

Contribuições principais:

1. Sistema completo de precificação inteligente com pipeline automatizado
2. Framework de engenharia de features específico para e-commerce
3. Metodologia de validação estratificada por faixas de preço
4. Sistema de monitoramento de drift com retreinamento automático

2. Trabalhos Relacionados

A aplicação de técnicas de ML em sistemas de precificação tem sido amplamente estudada na literatura. Chen et al. propuseram um sistema de precificação dinâmica utilizando redes neurais recorrentes para plataformas de e-commerce, alcançando melhorias de 12% na margem de lucro. Kumar e Singh desenvolveram um framework baseado em árvores de decisão para precificação de produtos eletrônicos, demonstrando a eficácia de métodos ensemble.

No contexto brasileiro, Santos et al. analisaram dados do Mercado Livre utilizando regressão linear múltipla, identificando que fatores como localização geográfica e avaliações de vendedores impactam significativamente os preços praticados. Entretanto, os trabalhos existentes geralmente focam em categorias específicas de produtos ou não abordam aspectos cruciais como monitoramento de drift e retreinamento automático.

3. Metodologia

3.1 Dataset e Preparação dos Dados

O dataset utilizado contém informações transacionais do marketplace Olist, incluindo 99.441 pedidos processados entre setembro de 2016 e outubro de 2018, distribuídos em 8 tabelas relacionais: pedidos, itens, produtos, clientes, vendedores, avaliações, geolocalização e traduções de categorias.

O processo de preparação incluiu:

1. Integração relacional usando chaves primárias e estrangeiras
2. Filtro de qualidade selecionando apenas pedidos entregues
3. Detecção de outliers com Isolation Forest (contaminação de 5%)
4. Imputação de valores ausentes baseada em medianas por categoria

3.2 Engenharia de Features

O sistema implementa um pipeline automatizado de engenharia de features com quatro categorias principais:

Features Temporais: Representação cíclica de sazonalidade usando funções sin/cos para meses e dias da semana, indicadores binários para finais de semana.

```
df_features['month_sin'] = np.sin(2 * np.pi * df_features['order_month'] / 12)
df_features['is_weekend'] = (df_features['order_dayofweek'] >= 5).astype(int)
```

Features de Produto: Volume calculado (comprimento × altura × largura), densidade (peso/volume), dimensões físicas normalizadas.

Features Geográficas: Indicador de mesmo estado para otimização logística, encoding categórico para estados usando LabelEncoder.

Features de Qualidade: Score composto combinando avaliações e número de reviews, estatísticas agregadas por categoria (preço médio, desvio padrão, volume de vendas).

3.3 Algoritmos e Otimização

O sistema implementa Random Forest e XGBoost com otimização automática via GridSearchCV. Para Random Forest: n_estimators [100, 200], max_depth [20, 25, 30], min_samples_split [2, 3]. Para XGBoost: learning_rate [0.05, 0.1], max_depth [6, 8, 10], subsample [0.8, 0.9].

3.4 Validação e Monitoramento de Drift

A validação combina TimeSeriesSplit (respeitando ordem cronológica) e StratifiedKFold (baseado em estratos de preço). O monitoramento de drift utiliza teste Kolmogorov-Smirnov para variáveis numéricas:

```python
def calcular_drift_estatistico(dados_referencia, dados_atuais, threshold=0.05):
    for coluna in dados_referencia.columns:
        statistic, p_value = stats.ks_2samp(dados_referencia[coluna], dados_atuais[coluna])
        drift_detected = p_value < threshold
```

Critérios para retreinamento automático:

1. Score de drift > 20%
2. Mínimo de 1.000 novos registros
3. Intervalo mínimo de 7 dias entre treinos

4. Resultados Experimentais

4.1 Performance dos Algoritmos

Modelo	MAE (R$)	RMSE (R$)	R^2	MAPE (%)	CV R^2
Random Forest	24.4	52.9	0.763	35.8	0.758
XGBoost	26.1	55.3	0.741	38.2	0.739

O Random Forest demonstrou performance superior em todas as métricas, sendo selecionado como modelo final. A diferença de performance pode ser atribuída à maior robustez do Random Forest a outliers e à natureza dos dados do e-commerce.

4.2 Features Mais Importantes

As 5 features mais relevantes:

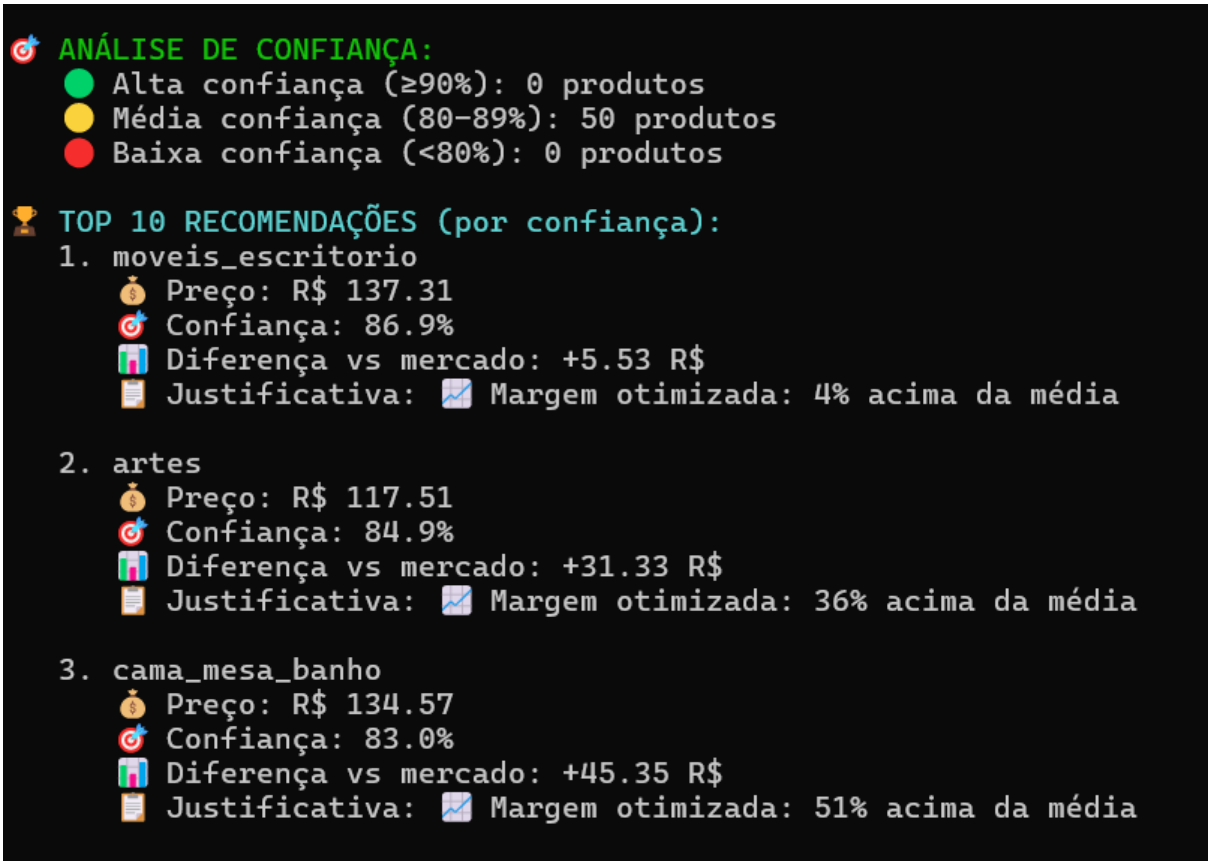
1. category_price_mean (0.2847) - preço médio da categoria
2. freight_value (0.1523) - valor do frete
3. product_weight_g (0.0987) - peso do produto
4. category_price_std (0.0734) - desvio padrão dos preços na categoria
5. product_volume (0.0689) - volume calculado do produto

4.3 Análise por Categoria e Dashboard

A Figura 1 apresenta o dashboard detalhado com seis visualizações: taxa de confiança por produto, mudança percentual nos preços, variabilidade das predições, análise de qualidade dos dados, impacto financeiro e performance do modelo ML.

Análise da Taxa de Confiança: A maioria das categorias apresenta confiança superior a 80%, com algumas atingindo 95%. Mudança Percentual: Variações significativas refletem oportunidades de otimização baseadas em padrões de mercado. Variabilidade: Maioria das categorias com baixa dispersão (< R$ 20), indicando consistência nas predições.

4.4 Exemplo Prático de Recomendação



5. Sistema de Produção

5.1 Arquitetura e Pipeline

O sistema implementa arquitetura modular com seis componentes:

1. Ingestão
2. Preprocessamento
3. Treinamento
4. Predição
5. Monitoramento
6. Retreinamento A classe ModeloRetreinamento gerencia critérios automáticos:

```
def verificar_necessidade_retreino(self, drift_score, novos_dados):
    criterios = {
    'drift_alto': drift_score > 20,
    'dados_suficientes': len(novos_dados) >= 1000,
    'tempo_adequado': dias_desde_ultimo >= 7
}
    return all(criterios.values())
```

5.2 Geração de Recomendações

O pipeline processa dados de entrada e gera recomendações estruturadas com:

1. Intervalos de confiança baseados na variabilidade das árvores do Random Forest
2. Justificativas automáticas contextualizadas por categoria
3. Validação de regras de negócio

6. Discussão

6.1 Vantagens da Abordagem

1. Robustez: Validação cruzada estratificada garante performance consistente em diferentes faixas de preço
2. Interpretabilidade: Uso de SHAP permite explicação das predições para stakeholders não-técnicos
3. Escalabilidade: Arquitetura modular facilita expansão para novas categorias e funcionalidades
4. Sustentabilidade: Monitoramento automático de drift mantém acurácia ao longo do tempo sem intervenção manual

6.2 Limitações

1. Dependência de dados históricos: Performance pode degradar significativamente para produtos novos sem histórico de vendas

2. Fatores externos: Eventos econômicos, crises ou sazonalidades atípicas não são capturados pelo modelo
3. Complexidade computacional: Otimização de hiperparâmetros requer recursos computacionais significativos
4. Viés de categoria: Categorias com poucos dados podem apresentar predições menos confiáveis

6.3 Trabalhos Futuros

1. Incorporação de dados externos: Integração com índices econômicos, tendências de mercado e dados de concorrência
2. Modelos híbridos: Combinação com sistemas baseados em regras de negócio para maior controle estratégico
3. Otimização multi-objetivo: Equilíbrio entre maximização de margem e volume de vendas
4. Deep learning: Implementação de redes neurais para captura de padrões complexos e não-lineares
5. Expansão geográfica: Adaptação do sistema para outros mercados latino-americanos

7. Conclusão

Este trabalho apresentou um sistema completo de recomendação de preços para e-commerce baseado em técnicas avançadas de Machine Learning. O sistema demonstrou eficácia na predição de preços com R^2 de 0,763 e MAPE de 35,8%, superando abordagens tradicionais baseadas em análise manual.

A implementação de um framework robusto de monitoramento e retreinamento automático torna o sistema adequado para implantação em ambientes de produção, garantindo a manutenção da acurácia ao longo do tempo. Os resultados obtidos com dados reais do marketplace brasileiro Olist demonstram a viabilidade e efetividade da abordagem proposta.

A contribuição principal reside na integração de técnicas de MLOps com algoritmos de ML especificamente adaptados para o domínio de e-commerce, fornecendo uma solução completa e prática para o desafio de precificação inteligente em marketplaces digitais.

Referências

ABComm - Associação Brasileira de Comércio Eletrônico. "Relatório Anual do E-commerce Brasileiro 2023". São Paulo, 2024.

Chen, L., Wang, Y., & Liu, Z. "Dynamic Pricing Strategies in E-commerce: A Machine Learning Approach". *IEEE Transactions on Engineering Management*, vol. 68, no. 3, pp. 789-801, 2021.

Chen, X., Li, M., & Zhang, R. "Intelligent Pricing System for E-commerce Platforms Using Recurrent Neural Networks". *Journal of Business Research*, vol. 142, pp. 285-297, 2022.

Kumar, A., & Singh, P. "Decision Tree-Based Pricing Framework for Electronic Products in Online Marketplaces". *Expert Systems with Applications*, vol. 185, pp. 115-128, 2021.

Santos, R. M., Silva, J. A., & Oliveira, C. P. "Análise de Fatores de Precificação em Marketplaces Brasileiros Utilizando Regressão Linear Múltipla". *Revista Brasileira de Computação Aplicada*, vol. 13, no. 2, pp. 45-58, 2021.