Previsão de Notas de Review (Olist)

Introdução à Ciência de Dados — NES

Tiago Cavalcante Trindade

Julho de 2025

Problema

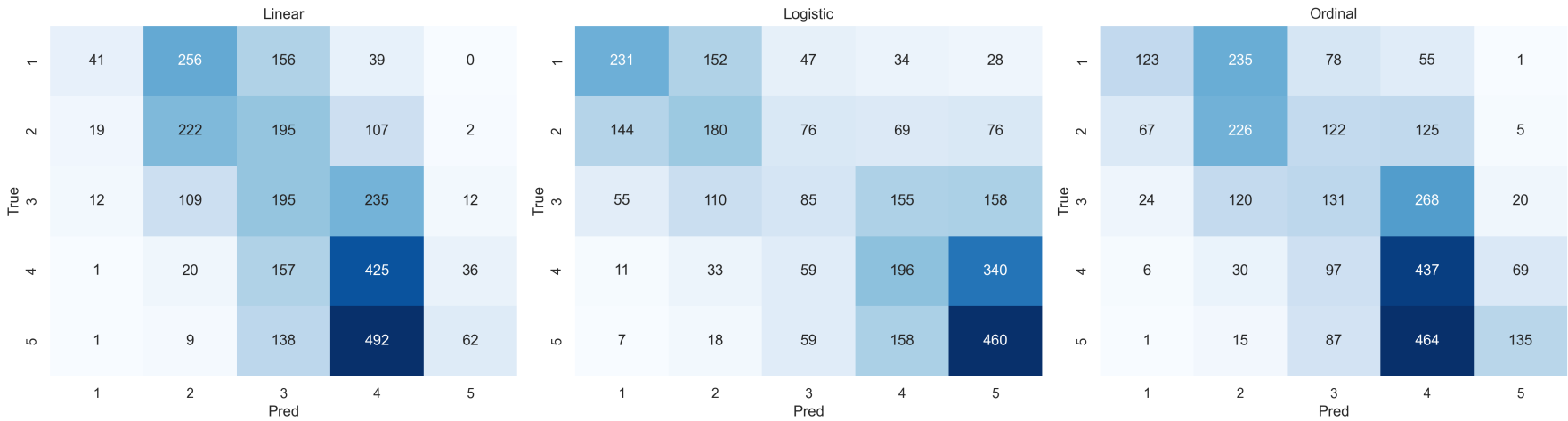
- Prever a **review_score** (1-5 estrelas) antes do cliente avaliar.
- Base de dados pública da **Olist**.
- **Features**
  - Numéricas: tempo de entrega, valor do carrinho, frete, nº de itens, pagamentos, dia da semana, mês.
  - Texto: título + mensagem da review $\rightarrow$ vetor FastText spaCy (300 dim) $\rightarrow$ SVD (120 dim).
- Balanceamento por undersampling; divisão temporal 80/20 (treino/teste).

Modelos Avaliados

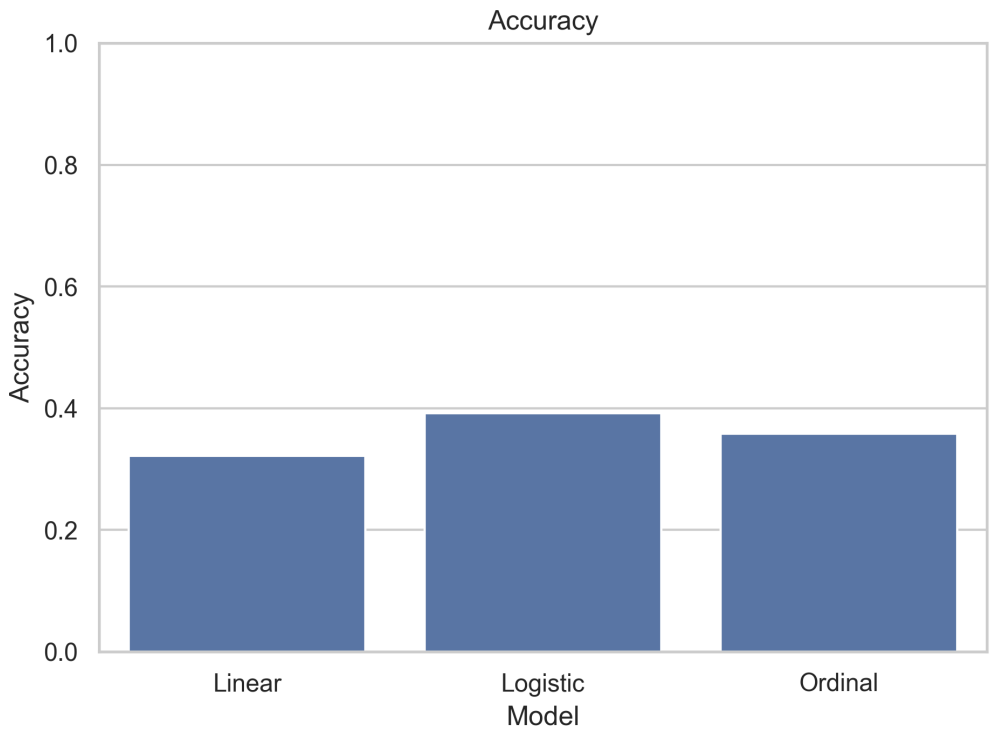
1. Regressão Linear (arredondada & cortada)
2. Regressão Logística Multiclasse
3. Regressão Logística Ordinal

Implementados com `scikit-learn` (+ `mord` para o modelo ordinal).

Matrizes de Confusão



Acurácia por Modelo



Resumo de Métricas

Modelo	Acurácia	MAE	RMSE
Regressão Linear (arred.)	0.321	0.917	**1.105**
Regr. Logística Multiclasse	**0.391**	0.882	1.263
Regr. Logística Ordinal	0.358	**0.823**	1.115

Principais Insights

- Logística Multiclasse: melhor acerto exato (aproximadamente 40%).
- Logística Ordinal: menor MAE e RMSE — erra geralmente só ± 1 estrela.
- Regressão Linear é apenas um baseline simples.

Conclusão

Escolha

- Se o KPI exige acerto exato da nota $\Rightarrow$ **Logística Multiclasse**.
- Se o mais importante é limitar a distância do erro $\Rightarrow$ **Logística Ordinal**.