

Profitability and Risk Segmentation in Retail

Overview

This project analyzes orders from a retail dataset. The goal is to understand what drives profit margin and to separate high-risk orders from more stable ones.

Why it matters

Retail businesses lose money when they apply discounts or shipping policies without knowing the impact. This project shows patterns that help reduce losses and improve margin.

What was done

- Cleaned and structured the dataset
- Created variables related to location, category, shipping, and discounts
- Tested multiple models:
 - Lasso/Ridge (baseline model, poor fit)
 - GAMLSS with t-distribution (better fit, but bimodal residuals)
 - XGBoost / LightGBM (high accuracy, low interpretability)
- - Final model: Mixture of Gaussian Linear Regressions with a concomitant model

Key Findings

- Cluster 1 (High Risk):
 - Share of orders: 31.4%
 - Share of negative margins: 40%
 - Average margin: -13.2%
- Cluster 2 (Stable):
 - Share of orders: 68.6%
 - Average margin: 23.5%

Business Impact

- Orders in Cluster 1 generate a total of \$677 in losses over the period.
- A policy change targeting this group could reduce losses by approximately 67.86%.
- Main cost drivers in Cluster 1: Discount, Tables/Bookcases Subcategory, Region (Mountain).

What to do with this

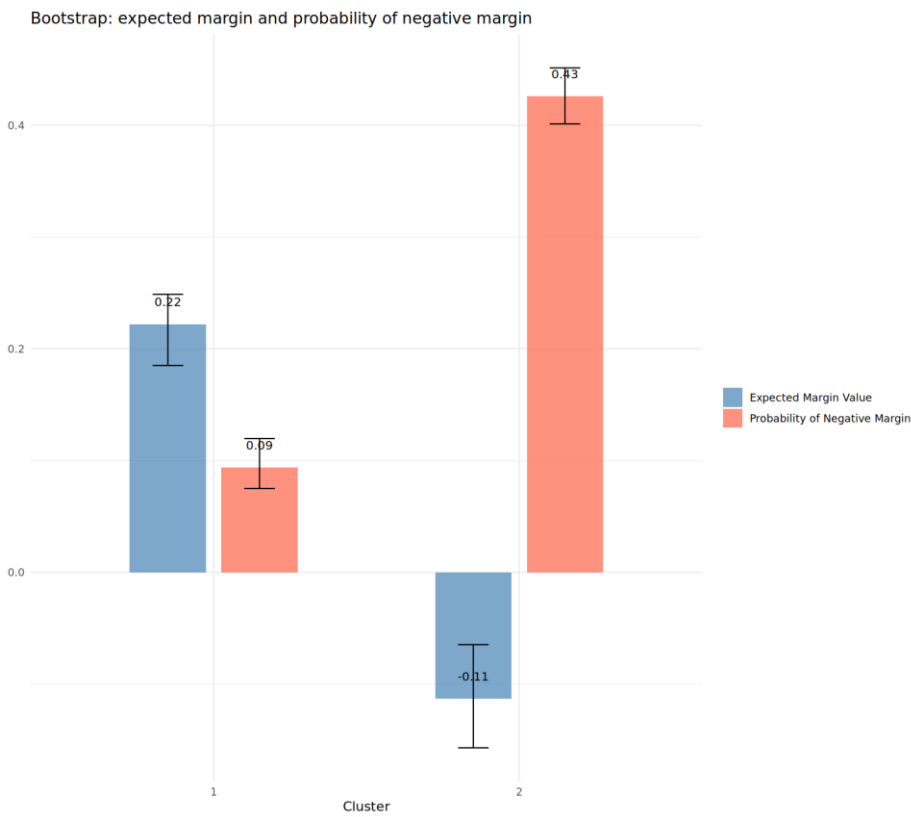
- Review all discount rules
- Reevaluate pricing strategy for Tables, Bookcases, Supplies, and Machines
- Flag risky orders early using model probabilities

Model Validation

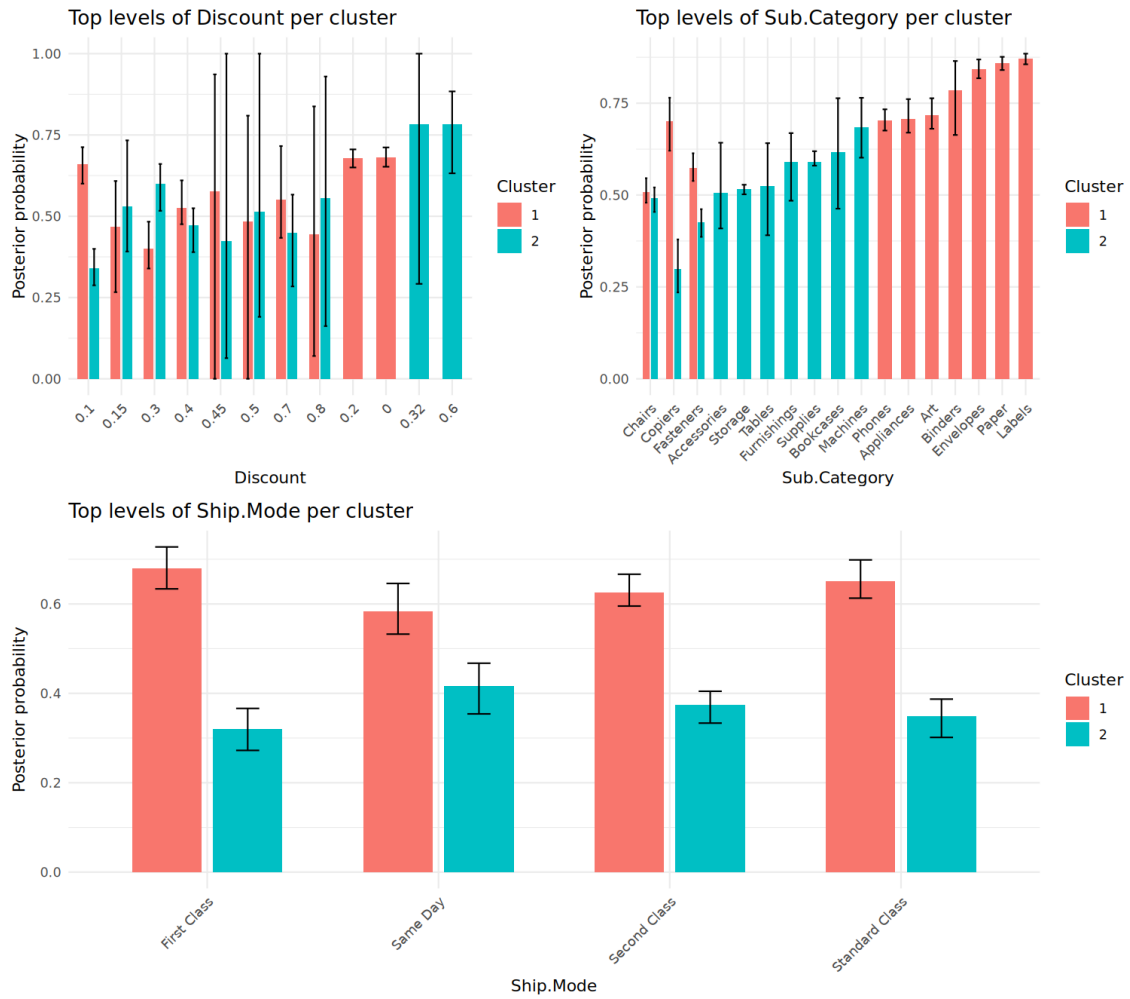
- The EM algorithm was stable after around 10 random restarts
- Bootstrap validation with 1000 resamples
- Performance metrics:
 - - RMSE (out-of-sample): 0.06713472
 - - R^2 (out-of-sample): 0.9786957

Visuals

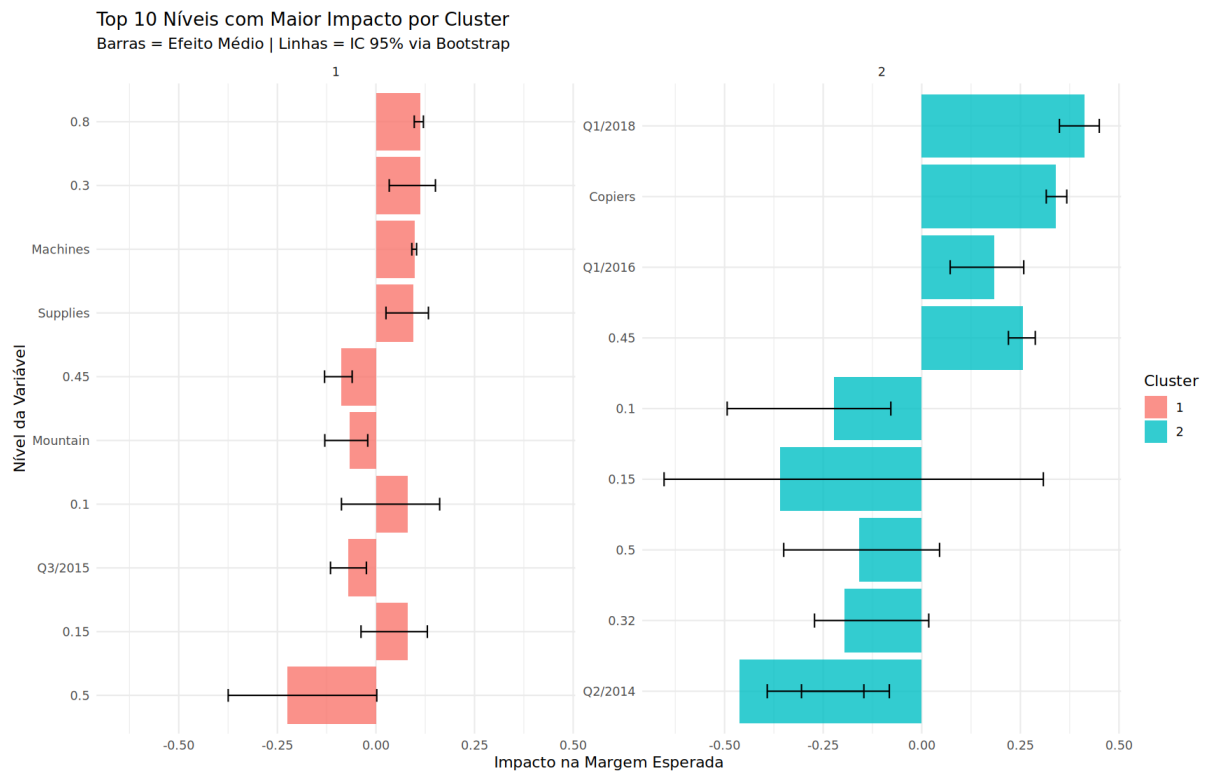
1. Risk-return trade-off between clusters



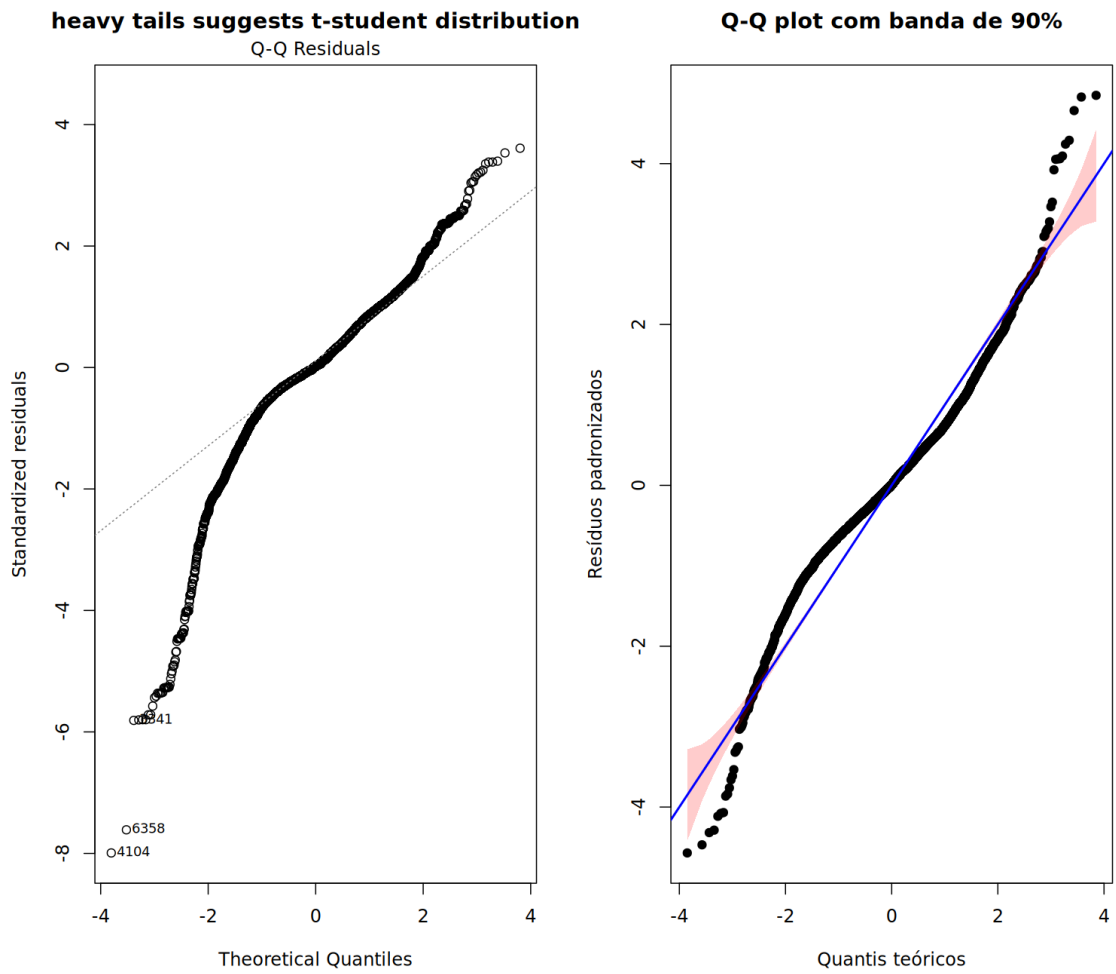
2. Operational drivers of high performance



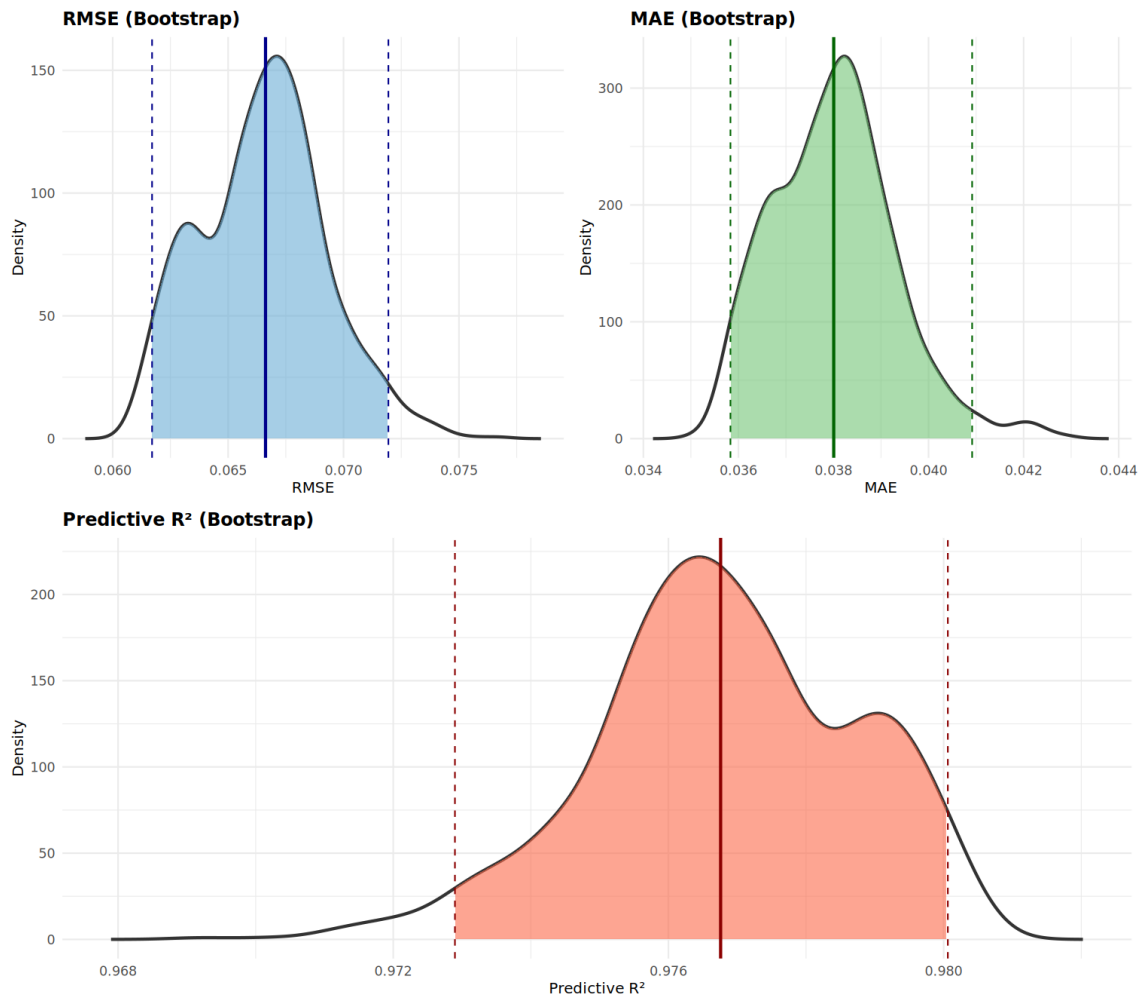
3. Top 10 most impactful levels per cluster



4. Residual comparison between models



5. Bootstrap validation



6. Main cost drivers in Cluster 1

	Variavel	Nivel	Lucro_Medio	Contagem
	<chr>	<chr>	<dbl>	<int>
1	Discount	0.5	-340.	53
2	Discount	0.45	-206.	9
3	Discount	0.8	-111.	259
4	Discount	0.4	-108.	178
5	Discount	0.7	-92.7	353
6	Discount	0.32	-92.6	24
7	Discount	0.3	-44.7	194
8	Sub.Category	Tables	-44.6	264
9	Discount	0.6	-40.2	123
10	division	West South Central	-14.7	977
11	Sub.Category	Bookcases	-13.9	190
12	Sub.Category	Supplies	-8.83	164
13	Sub.Category	Machines	-4.46	100
14	division	Mountain	-3.55	498