# Segmentação de Rentabilidade e Risco no Varejo

## Visão Geral

Este projeto analisa pedidos de um conjunto de dados de vendas no varejo. O objetivo é entender o que influencia a margem de lucro e separar pedidos de alto risco dos mais estáveis.

## Por que isso importa

Negócios do varejo perdem dinheiro ao aplicar descontos ou políticas de envio sem conhecer o impacto. Este projeto mostra padrões que ajudam a reduzir perdas e melhorar a margem.

## O que foi feito

- Limpeza e estruturação do conjunto de dados
- Criação de variáveis relacionadas a localização, categoria, envio e descontos
- Teste de múltiplos modelos:
- - Lasso/Ridge (modelo base, ajuste ruim)
- - GAMLSS com distribuição t (melhor ajuste, mas resíduos bimodais)
- - XGBoost / LightGBM (alta acurácia, baixa interpretabilidade)
- - Modelo final: Mistura de Regressões Lineares Gaussianas com modelo concomitante

## Principais Resultados

- Cluster 1 (Alto Risco):
- - Participação nos pedidos: 31,4%
- - Participação nas margens negativas: 40%
- - Margem média: -13,2%
- Cluster 2 (Estável):
- - Participação nos pedidos: 68,6%
- - Margem média: 23,5%

## Impacto no Negócio

- Os pedidos do Cluster 1 geram um total de $677 em perdas no período.
- Uma mudança de política direcionada a esse grupo poderia reduzir as perdas em aproximadamente 67,86%.

- Principais fatores de custo no Cluster 1: Desconto, Subcategoria Mesas/Estantes, Região (Mountain).

## O que fazer com isso

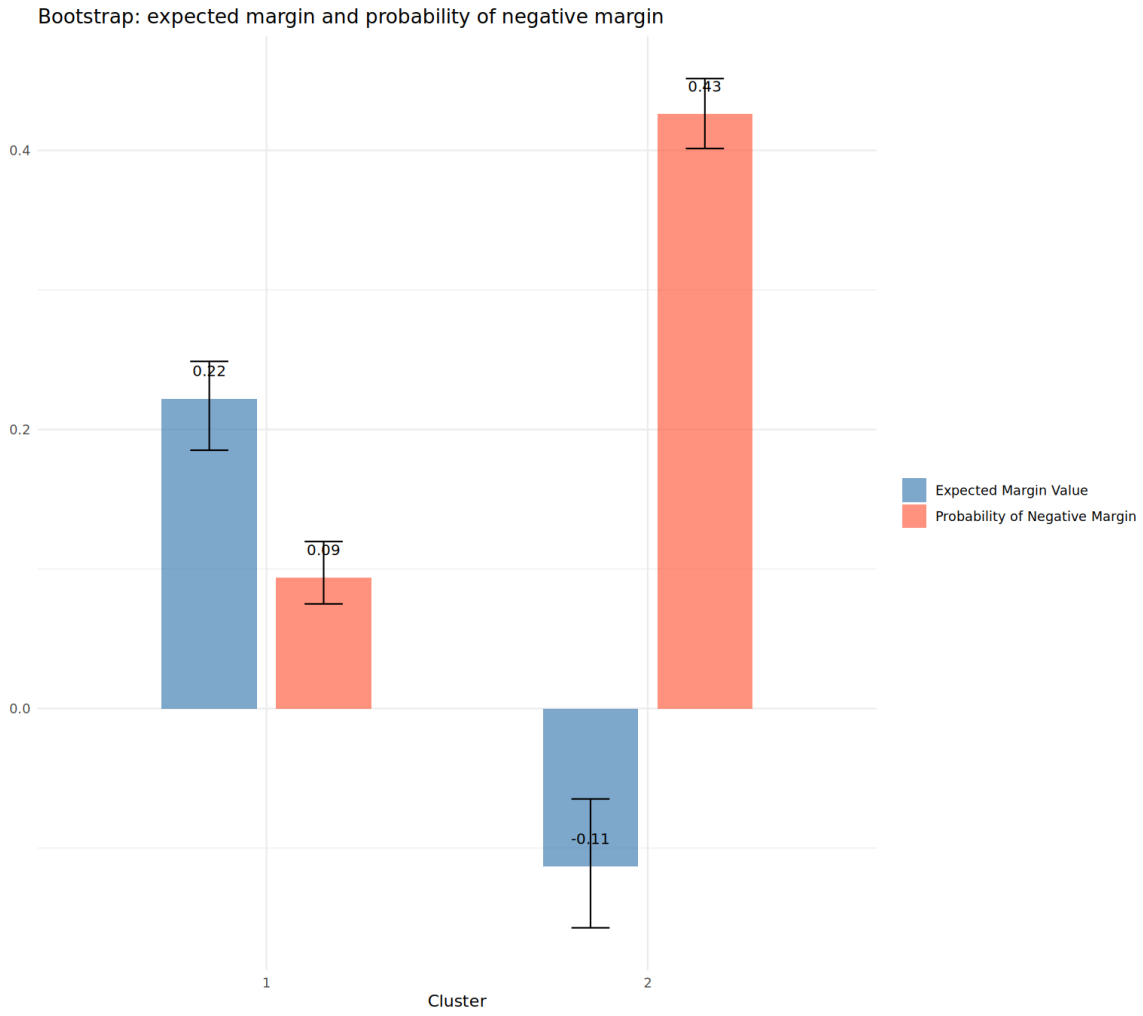
- Revisar todas as regras de desconto
- Reavaliar a estratégia de precificação para Mesas, Estantes, Suprimentos e Máquinas
- Marcar pedidos de risco antecipadamente usando as probabilidades do modelo

## Validação do Modelo

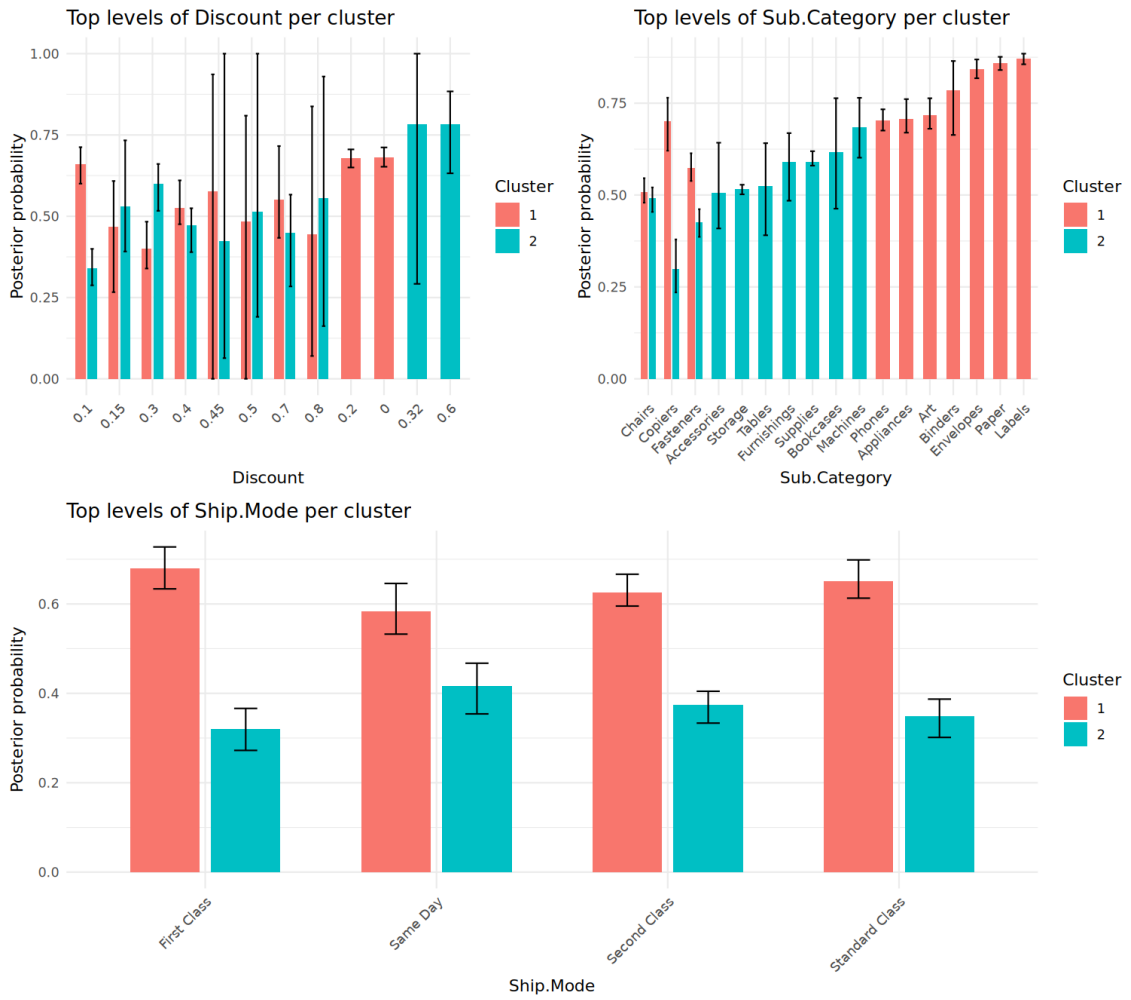
- O algoritmo EM foi estável após cerca de 10 reinicializações aleatórias
- Validação via bootstrap com 1000 reamostragens
- Métricas de desempenho:
- - RMSE (fora da amostra): 0.06713472
- - $R^2$ (fora da amostra): 0.9786957

## Visualizações

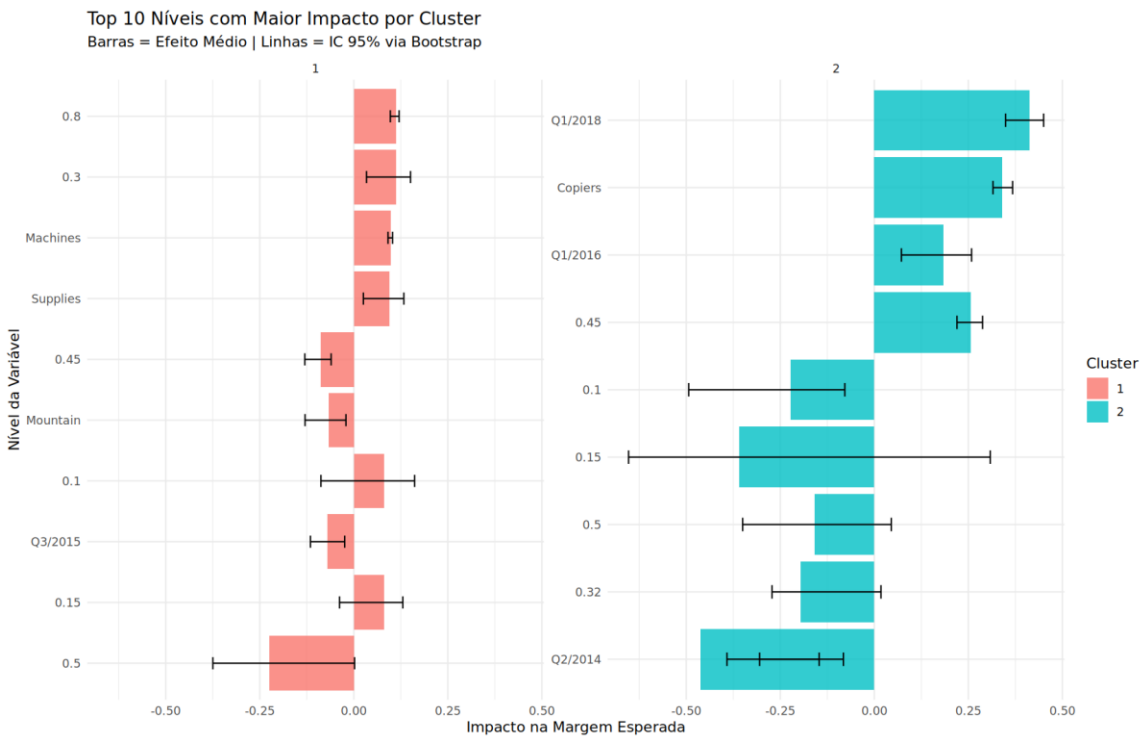
1. Trade-off risco vs. retorno entre clusters



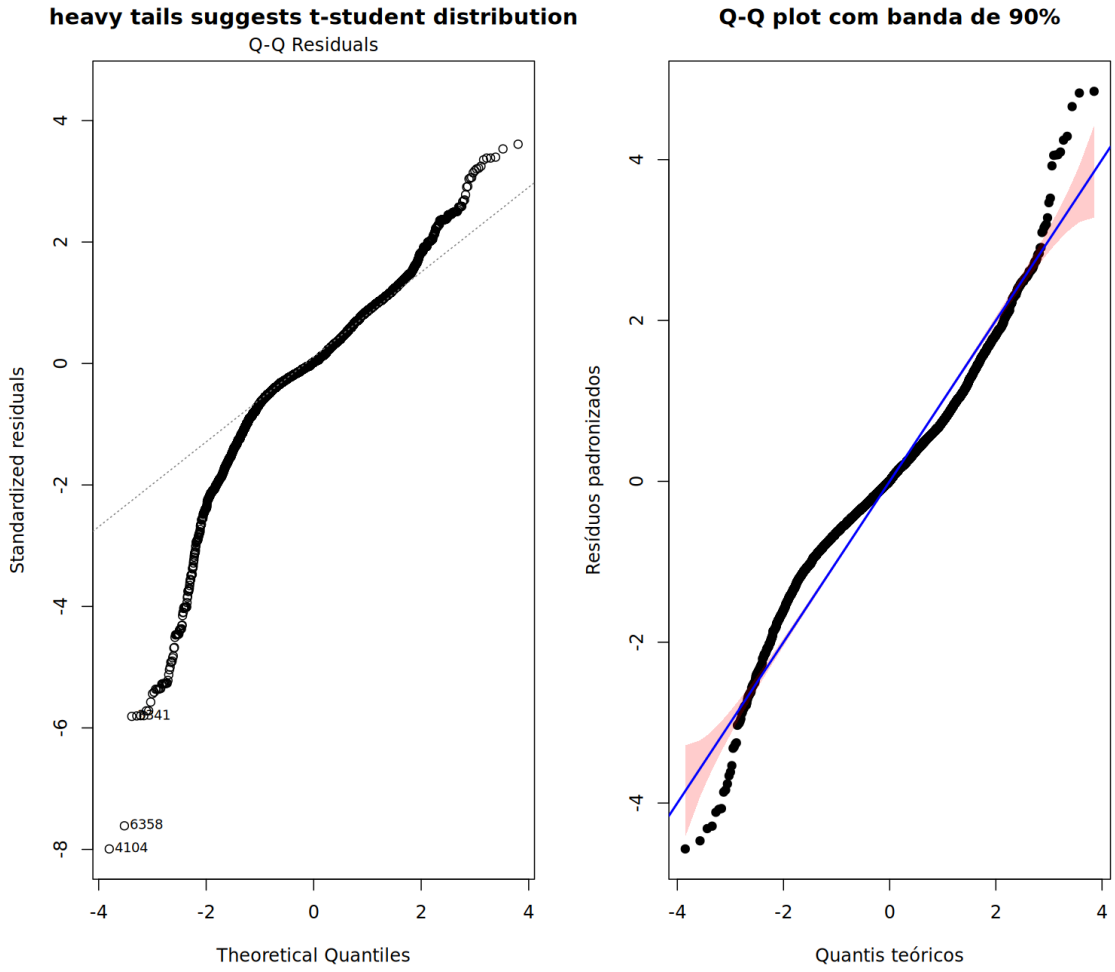

2. Fatores operacionais de alta performance



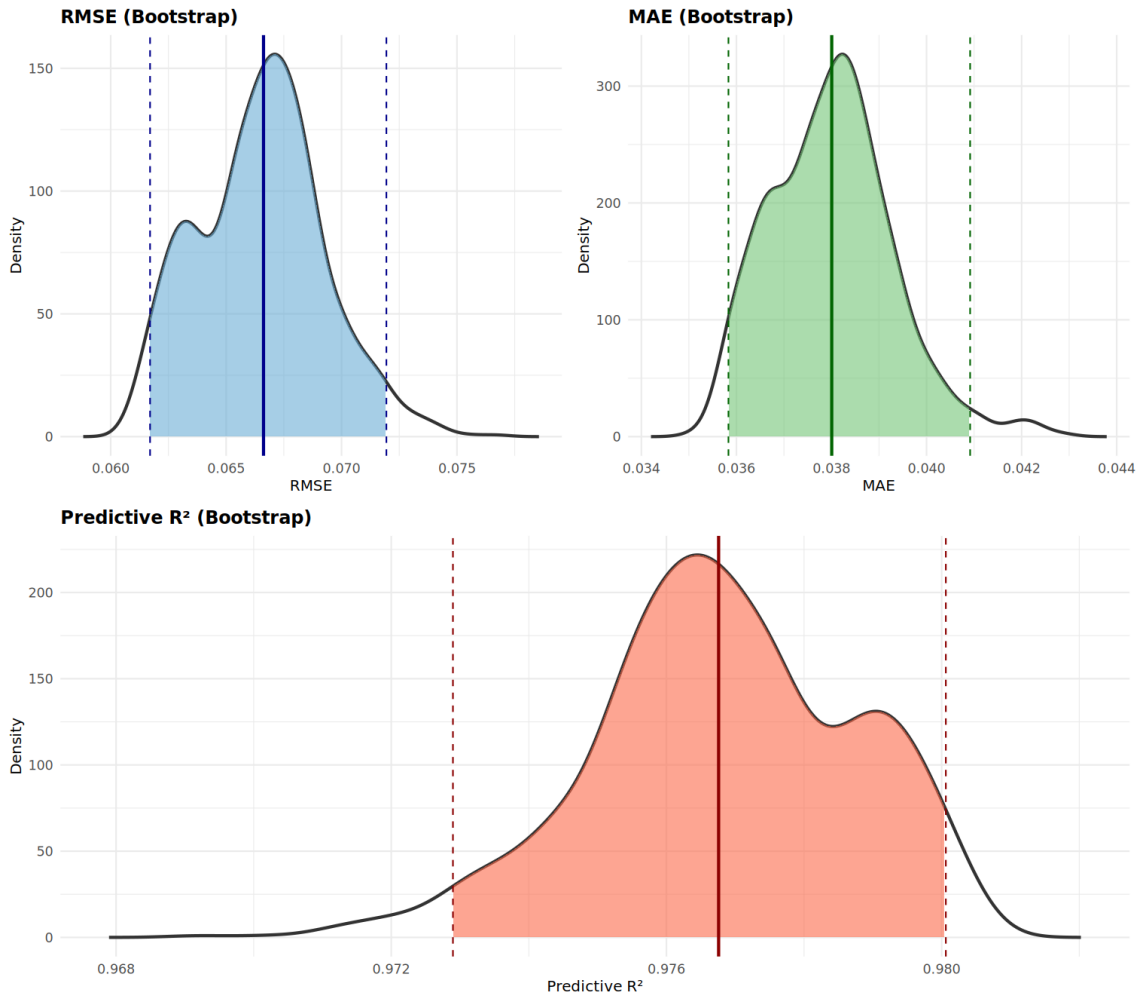

3. 10 níveis mais impactantes por cluster




4. Comparação de resíduos entre modelos




5. Validação via bootstrap




6. Principais fatores de custo no 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 |