# 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 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 features related to location, category, shipping, and discounts
- Tested multiple models:
  - Lasso/Ridge (baseline, 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 margin: 40%
  - Average margin: -13.2%
- Cluster 2 (Stable):
  - Share of orders: 68.6%
  - Average margin: 23.5%

## Business Impact

- Orders in Cluster 1 generate $677 of total loss over the period.

- A policy change targeting this group could reduce losses by approximately 67.86%.

- Main cost drivers in Cluster 1 were: Discount, Sub-Category Tables/Bookcases, Region (Mountain).

## What to do with this

- Review all discount rules

- Review Tables, Bookcases, Supplies and Machines pricing strategy

- Flag risky orders early using model probabilities

## Model Validation

- The EM algorithm was stable after about 10 random restarts

- Bootstrap validation used 1000 resamples

- RMSE (out-of-sample): 0.0671

- $R^2$ (out-of-sample): 0.9787

## Tools Used

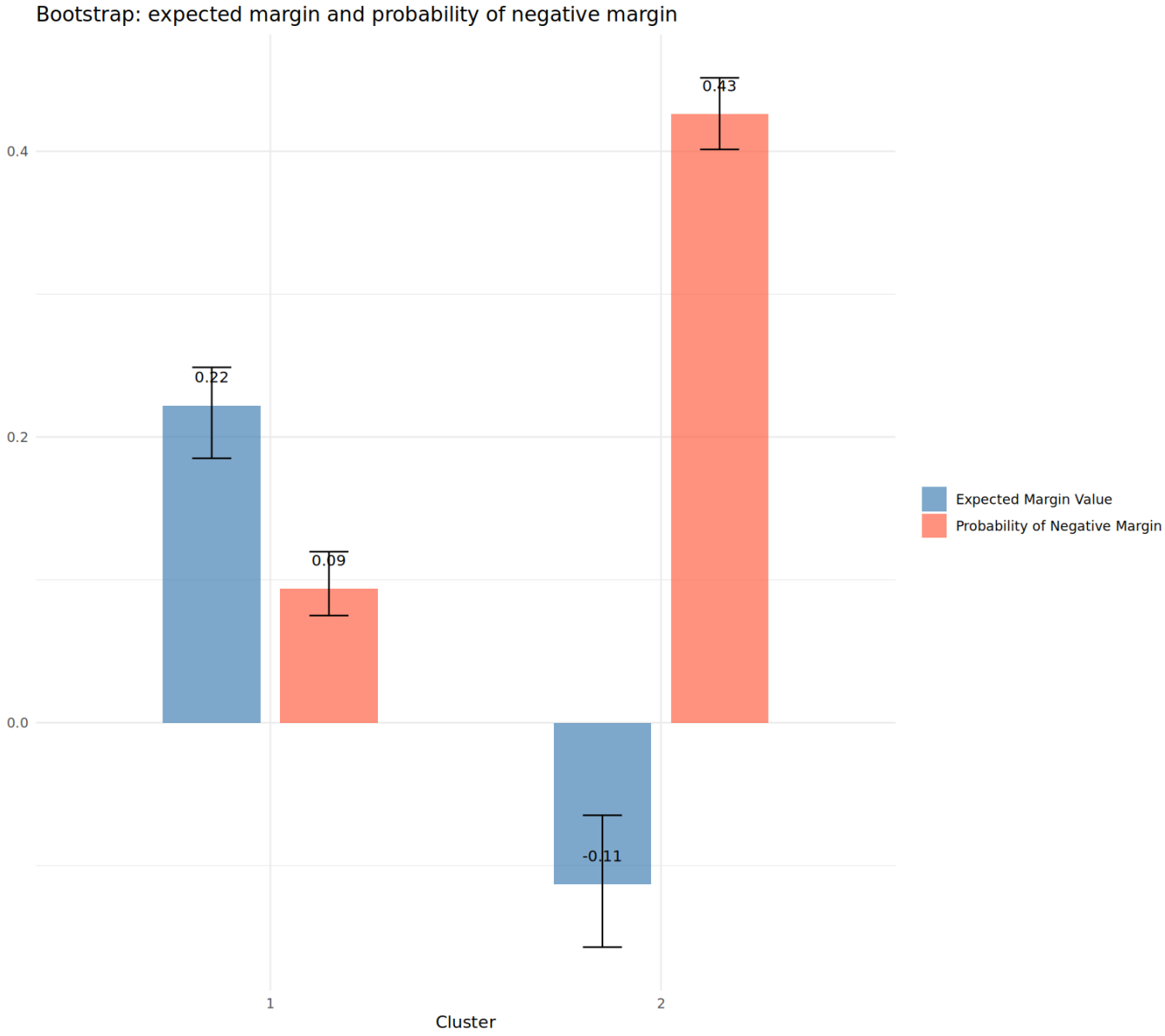
- Language: R

- Modeling: flexmix, gamlss, xgboost, lightgbm, glmnet

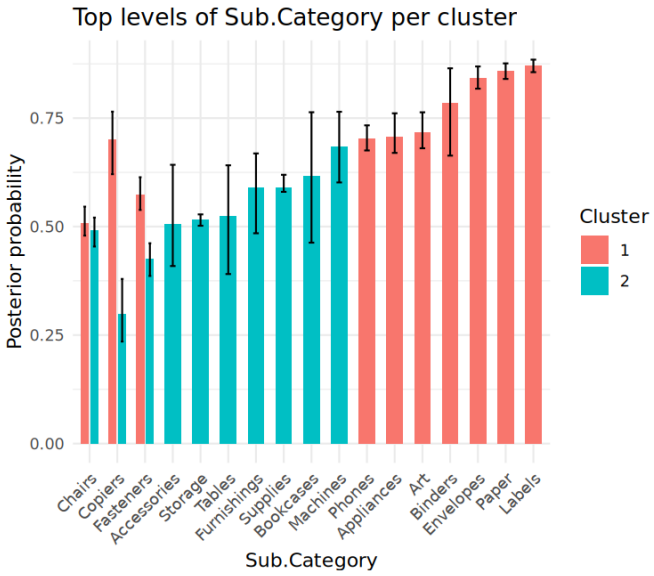
- Visualization: ggplot2

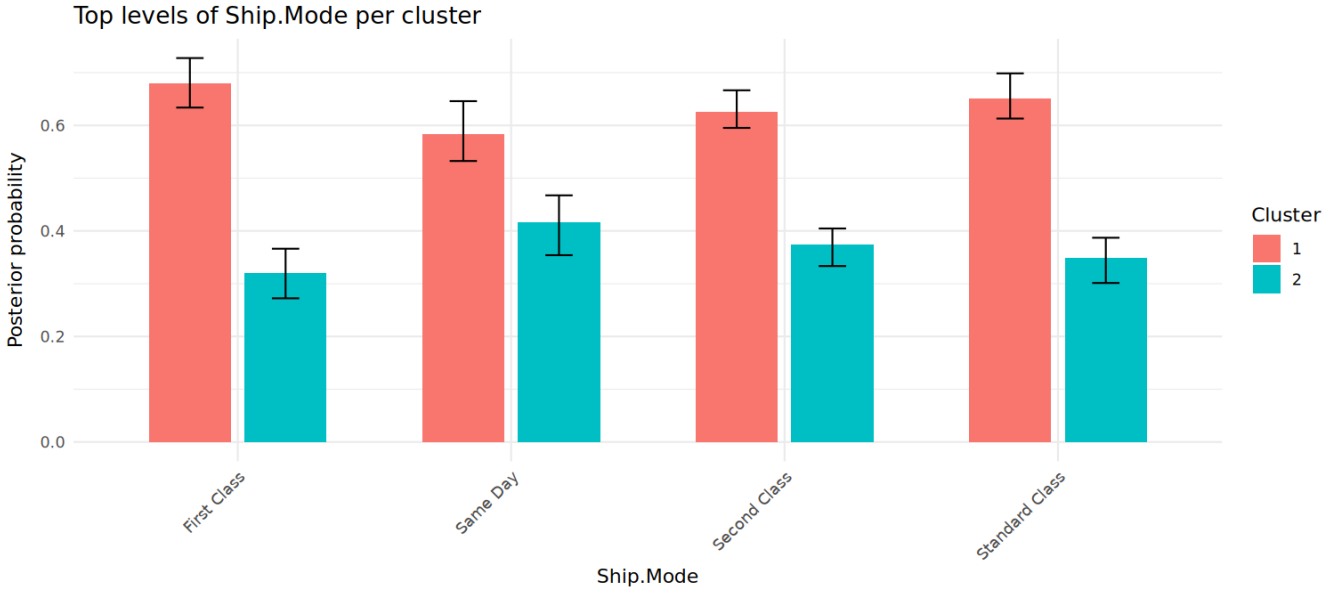
- Parallel computing: foreach, doParallel

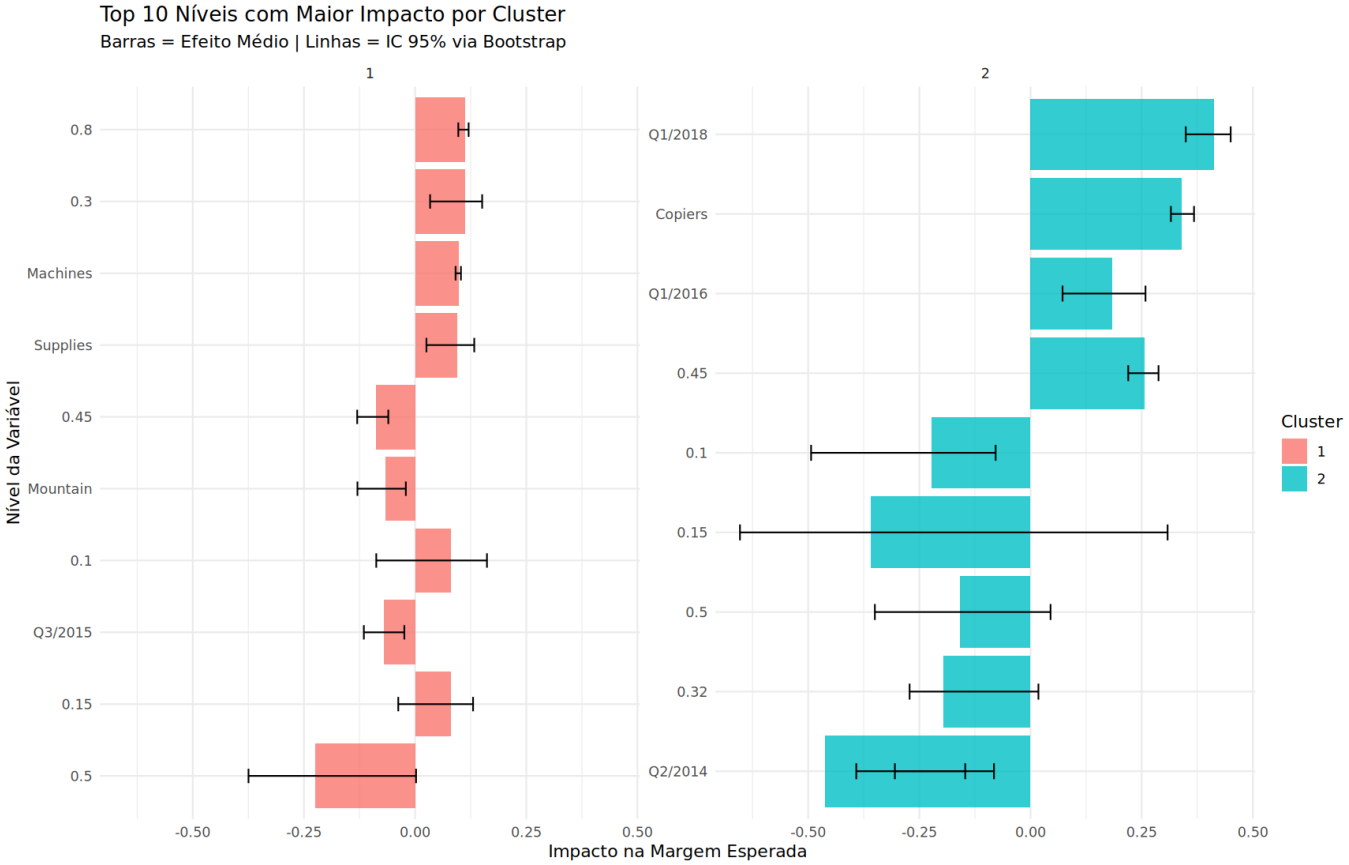
| | Variavel | Nivel | Lucro_Medio | Contagem |
|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <int> |
| 1 | Discount | 0.5 | -340. | 53 |
| 2 | Discount | 0.45 | -206. | 9 |
| 3 | Discount | 0.8 | -111. | 259 |
| 4 | Discount | 0.4 | -108. | 178 |
| 5 | Discount | 0.7 | -92.7 | 353 |
| 6 | Discount | 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 |


Bootstrap: expected margin and probability of negative margin
0.4
0.2
0.0
0.22
0.09
0.43
-0.11
1
2
Cluster
Expected Margin Value
Probability of Negative Margin

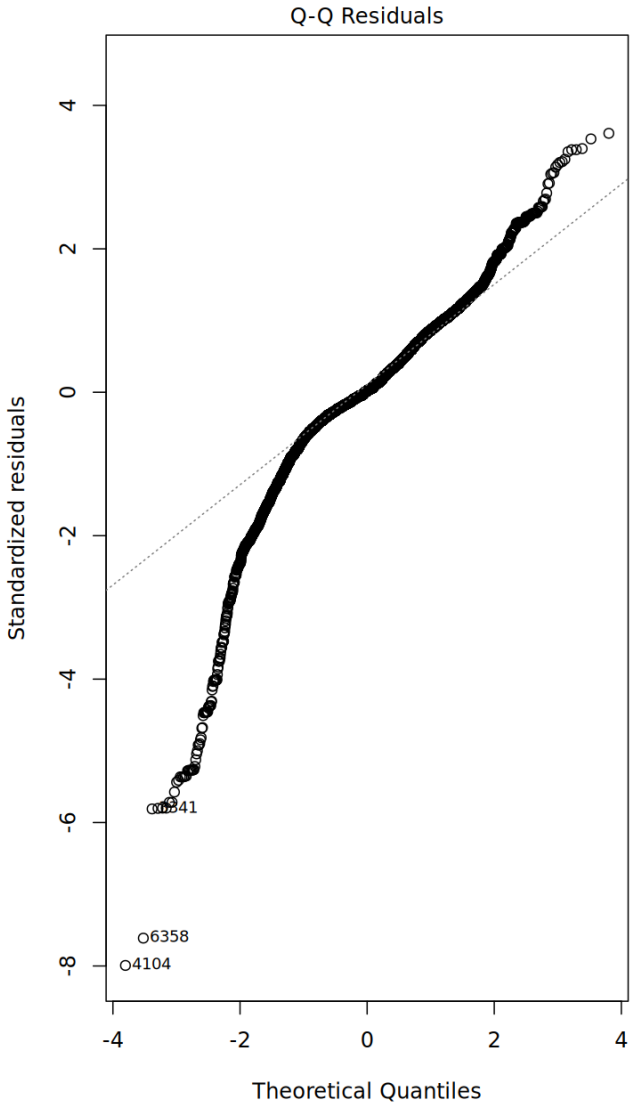

Top levels of Discount per cluster
Posterior probability
1.00
0.75
0.50
0.25
0.00
0.1
0.15
0.3
0.4
0.45
0.5
0.7
0.8
0.2
0
0.32
0.6
Discount
Cluster
1
2


Top levels of Sub.Category per cluster
Posterior probability
0.75
0.50
0.25
0.00
Chairs
Copiers
Fasteners
Accessories
Storage
Tables
Furnishings
Supplies
Bookcases
Machines
Phones
Appliances
Art
Binders
Envelopes
Paper
Labels
Sub.Category
Cluster
1
2

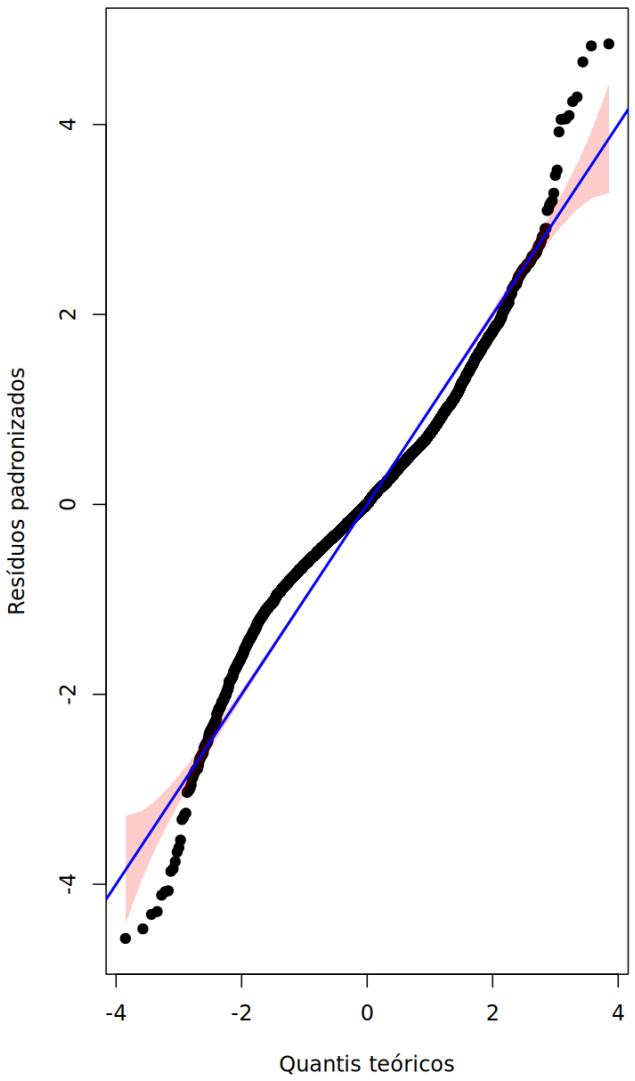

Top levels of Ship.Mode per cluster
Posterior probability
0.6
0.4
0.2
0.0
First Class
Same Day
Second Class
Standard Class
Ship.Mode
Cluster
1
2

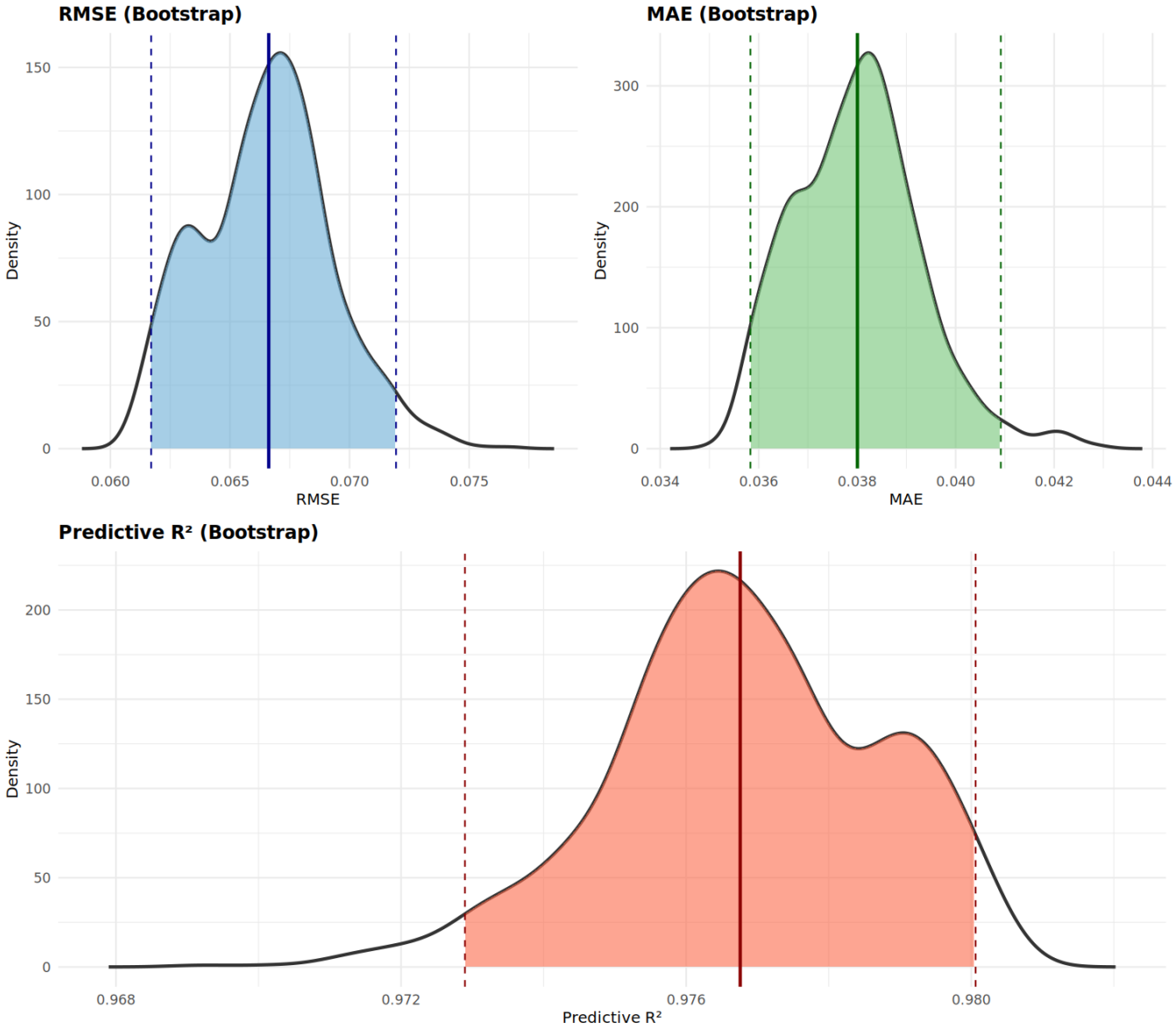

Top 10 Níveis com Maior Impacto por Cluster
Barras = Efeito Médio | Linhas = IC 95% via Bootstrap
1
2
0.8
0.3
Machines
Supplies
0.45
Mountain
0.1
Q3/2015
0.15
0.5
Q1/2018
Copiers
Q1/2016
0.45
0.1
0.15
0.5
0.32
Q2/2014
Nível da Variável
-0.50
-0.25
0.00
0.25
0.50
Impacto na Margem Esperada
Cluster
1
2

**heavy tails suggests t-student distribution**




**Q-Q plot com banda de 90%**




RMSE (Bootstrap)
Density
RMSE
0.060
0.065
0.070
0.075
0
50
100
150
MAE (Bootstrap)
Density
MAE
0.034
0.036
0.038
0.040
0.042
0.044
0
100
200
300
Predictive R² (Bootstrap)
Density
Predictive R²
0.968
0.972
0.976
0.980
0
50
100
150
200