

EP FINAL

1.115 LOJAS, 1 RÉGUA INJUSTA

Projeto concluído. Do CSV ao deploy.



Clusterização de 1.115 lojas que termina em **meta por loja e R\$ na mesa.**

A rede aplicava **+5%** para todas as 1.115 lojas.



Loja favorecida

Esquina de alto fluxo. Bate a meta com execução mediana e ainda negocia folga.



Loja pressionada

Concorrente a 300 m. Pede revisão da meta no primeiro trimestre e desiste.

A PERGUNTA QUE MUDOU

Quais lojas são parecidas de verdade, e quem está abaixo dos próprios pares, não da rede inteira?

6 fases do CRISP-DM. Todas fechadas.

1

Negócio

5 critérios de aceite: interpretável, acionável, estável, melhor que o cadastro, em R\$

2

Dados

1.017.209 registros diários · 1.115 lojas · 01/2013 a 07/2015

3

Preparação

1 milhão de linhas → 1 linha por loja, 6 features em log1p + padronização

4

Modelagem

K-Means de $k = 2$ a 10, Ward como segunda opinião

5

Avaliação

Cada critério da Fase 1 cobrado um por vez

6

Deploy

Plano de metas em Excel + classificador de lojas novas com testes

K-Means com $k = 5$

6 features: venda média/dia, ticket médio, efeito da promoção, volatilidade, sazonalidade e distância ao concorrente.

Silhueta média por número de clusters



$k = 5$ tem a maior silhueta e é o maior k que passa nos 3 testes de estabilidade. $k = 6$ e $k = 8$ explicavam mais, mas reprovaram. Tipo e sortimento ficaram fora: são o baseline.

O primeiro modelo **reprovou** no meu próprio teste.

Regra da Fase 1: tirando 20% das lojas, os grupos não podem mudar ($ARI \geq 0,80$). A v1 deu **0,755**.

ARI médio (quanto mais perto de 1, mais estável)

Troca de semente

0,992

Reordenação

0,990

Subamostra 80%

v1 · 0,755

v2 · 0,888

mínimo 0,80

Causa: apliquei log na sazonalidade para "corrigir a assimetria" e apaguei um grupo real de ~65 lojas sazonais. Com log1p nas 6 features, o grupo voltou e a estabilidade também.

Nada de "Grupo 3". Nome que o gestor usa.



Bairro competitivo, cesta pequena

Concorrente a 840 m · ticket R\$ 8,43

292

lojas



Destino de cesta grande

Concorrente a 5,8 km · maior ticket: R\$ 11,35

414

lojas



Sazonal

Sazonalidade 1,7× a da rede

65

lojas



Dependente de promoção

Promoção eleva a venda em +70,8%

165

lojas



Alto fluxo, cesta pequena

R\$ 9.545/dia · 1.235 clientes/dia

179

lojas

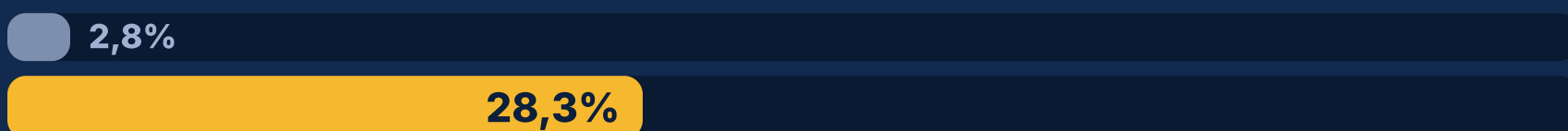
Cada grupo gera uma ação diferente em Comercial, Trade e Supply.
Nenhum par de clusters leva à mesma decisão.

O cadastro explicava **2,8%** da venda. Os clusters, **28,3%**.

■ tipo_loja ■ clusters η^2 = % da variação explicada

Venda média/dia

no modelo



Efeito da promoção

no modelo



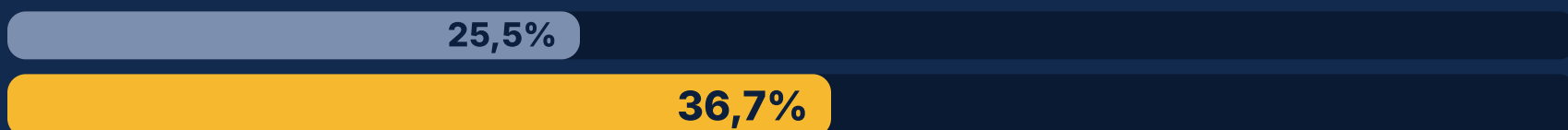
Volatilidade

no modelo



Clientes/dia

fora do modelo



Ganha em **9 de 10** métricas, inclusive 3 que nem entraram no modelo. Perde só em domingos abertos, coluna que tirei de propósito.

Silhueta não paga boleto.

R\$ 51,4 mi

por ano, se as lojas abaixo da mediana do próprio grupo recuperarem **só 25%** da distância

2,2% do faturamento da rede · 556 lojas

R\$ 20,6 mi

cenário conservador (10%)

R\$ 102,8 mi

cenário otimista (50%)

Comparar com a rede inteira daria R\$ 228 mi de gap. Maior, e indefensável: compara lojas que não são comparáveis.

A média dos percentuais mentia.

Regra: 2% de manutenção + 25% da distância até a mediana do grupo.



Média simples: 5,15% → ponderado: 4,19%

Faltariam R\$ 19 mi, mesmo com todo mundo batendo a meta



Calibração com fator único

Preserva quem recebe mais e quem recebe menos



Teto de 20% por loja

Sem ele, 46 lojas chegavam a 37,8%: meta que faz o gerente desistir em março

5,00%

crescimento da rede,
ponderado pelo faturamento

R\$ 117 mi

ganho anual previsto no plano

Loja nova entra no grupo **sem retreinar** o modelo.

```
# a partir dos indicadores da loja  
python -m codigo.classificar_lojas \  
    --indicadores lojas_novas.csv
```

```
# 14 testes; nenhum chama fit  
python -m pytest
```

1.115/1.115

lojas no mesmo cluster do
treino, recalculadas do zero
pelo pacote

96%

de concordância com 24
meses de histórico

- **Menos de 12 meses:** não classifica
- **12 a 23 meses:** classificação provisória
- **Alertas:** fronteira de cluster, perfil atípico, distância imputada

4 lições que levo para o próximo projeto

- 1 Corrigir assimetria por reflexo pode **apagar o padrão** que você quer achar.
- 2 Teste de estabilidade só vale se for **difícil de passar**.
- 3 Média de percentuais **não é crescimento em reais**.
- 4 Meta impossível não é ambiciosa. **É desmotivadora**.

CÓDIGO, NOTEBOOKS E DOCUMENTAÇÃO

github.com/V4lciJr/Auroa-Retail-Store-Clustering

Link no post · Pós em Data Analytics e IA Aplicada a Negócios · FNAT

Valci Júnior · Dados e IA aplicada a negócios