Credit Card Customer: Descoberta de subgrupos utilizando Beam-Search

...

Amanda Mendes Pinho
Gabriel Tonioni Duarte
João Vítor Fernandes Dias
Larissa Duarte Santana

Descrição da base de dados

O conjunto de dados resume o comportamento de uso de cerca de 9.000 titulares ativos de cartão de crédito num período de 6 meses. O arquivo está no nível do cliente, com 18 variáveis comportamentais.

Data Type	Column	Description	Range	# Null
float64	BALANCE	Balance amount left in their account to make purchases	[0, 19043.14]	0
float64	BALANCE_FREQUENCY	How frequently the Balance is updated	[0, 1.00]	0
float64	CASH_ADVANCE	Cash in advance given by the user	[0, 47137.21]	0
float64	CASH_ADVANCE_FREQUENCY	How frequently the cash in advance being paid	[0, 1.50]	0
float64	CREDIT_LIMIT	Limit of Credit Card for user	[0, 30000.00]	1
float64	INSTALLMENTS_PURCHASES	Amount of purchase done in installment	[0, 22500.00]	0
float64	MINIMUM_PAYMENTS	Minimum amount of payments made by user	[0, 76406.21]	313
float64	ONEOFF_PURCHASES	Maximum purchase amount done in one-go	[0, 40761.25]	0
float64	ONEOFF_PURCHASES_FREQUENCY	How frequently Purchases are happening in one-go	[0, 1.00]	0
float64	PAYMENTS	Amount of Payment done by user	[0, 50721.48]	0
float64	PRC_FULL_PAYMENT	Percent of full payment paid by user	[0, 1.00]	0
float64	PURCHASES	Amount of purchases made from account	[0, 49039.57]	0
float64	PURCHASES_FREQUENCY	How frequently the Purchases are being made	[0, 1.00]	0
float64	PURCHASES_INSTALLMENTS_FREQUENCY	How frequently purchases in installments are being done	[0, 1.00]	0
int64	CASH_ADVANCE_TRX	Number of Transactions made with "Cash in Advanced"	[0, 123]	0
int64	PURCHASES_TRX	Numbe of purchase transactions made	[0, 358]	0
int64	TENURE	Tenure of credit card service for user	[6, 12]	0
string	CUST_ID	Identification of Credit Card holder	[C10001, C19190]	0

Descrição da base de dados

- A base de dados permite analisar perfis comportamentais dos clientes, como:
  1. Clientes mais ou menos endividados (BALANCE, CASH_ADVANCE)
  2. Padrões de consumo (PURCHASES, PURCHASES_FREQUENCY)
  3. Capacidade de pagamento e uso do limite (PAYMENTS, CREDIT_LIMIT)
  4. Perfis de risco ou fidelização (TENURE, frequência de uso).
- Oferecendo um contexto real sobre clientes, permitindo extrair insights relevantes em estudos de segmentação e comportamento financeiro.

Descrição da base de dados

- O algoritmo Beam Search permite a identificação de grupos interpretáveis de clientes que compartilham características em comum, e se destacam em relação a um atributo de interesse, o que permitiria a extração de insights relevantes dessas descobertas.
- Com a descoberta, é possível indicar riscos, análise de clientes ativos, dificuldade financeira.
- Isso faz com que a base seja ideal para análises voltadas a marketing segmentado, gestão de risco, definição de limites de crédito, atividade de clientes aplicadas a estratégias de retenção.

Loading e pré-processamento

- Loading
  - `kagglehub.load_dataset(KaggleDatasetAdapter.PANDAS, data_handle, file_path)`
- Pré-processamento
  - Valores ausentes: remoção | substituição por média ou mediana
- Busca de Subgrupos
- Visualização dos subgrupos

Busca de subgrupos: Beam Search

- Beam Search é uma estratégia de busca heurística e controlada para encontrar subgrupos descritivos que se destacam em relação a um atributo-alvo.
- Etapas:
  - **Inicialização** : começa com subgrupos simples e bons (baseados em poucos descritores).
  - **Expansão** : subgrupos são expandidos com novos descritores.
  - **Avaliação** : cada subgrupo expandido é avaliado pela função de qualidade.
  - **Feixe (Beam)** : apenas os melhores k subgrupos continuam na busca.
  - **Iteração** : o processo se repete até atingir profundidade ou parar por qualidade.

Beam Search: parâmetros

- Alvos: PURCHASES_FREQUENCY|PURCHASES_TRX|ONEOFF_PURCHASES|BALANCE
- Espaço de Busca: Completa|Segmentada
- Função de Qualidade
  - stdQF
    - Centroide
    - Recompensa por tamanho do subgrupo: 0.0|0.3|0.5|1.0
  - stdQFTscore
  - WRAcc
- Tamanho do subgrupo: 10
- Quantidade de descritores: 3|8

Experimentos - PUCHASES_TRX - Quem utiliza muito o cartão?

Beam Search + Diferença da Média Ponderada:

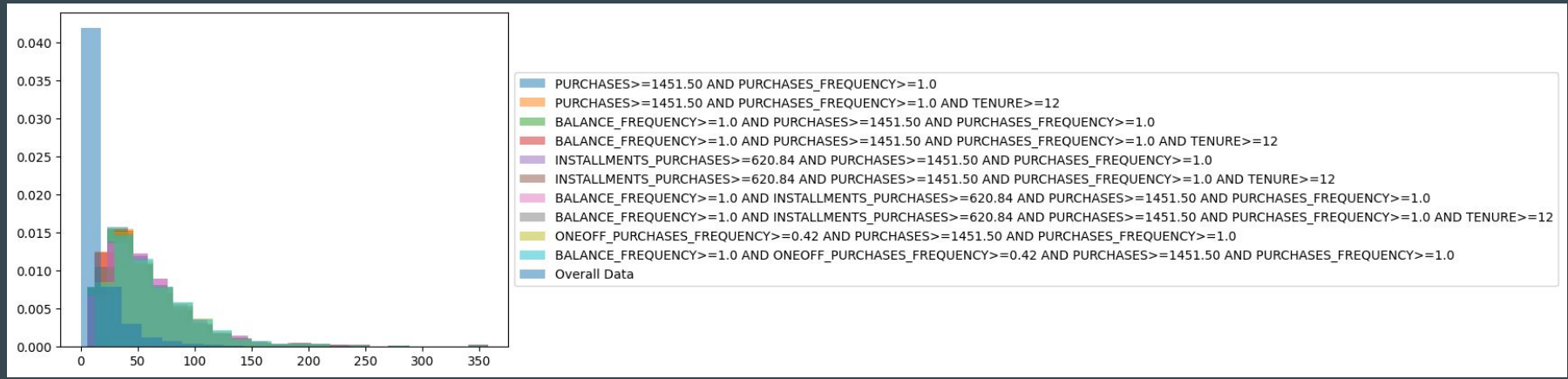
```
return instances_subgroup**a * (mean_sg - mean_dataset)
```

- Subgrupos competem para entrar no feixe.
- Parâmetro “a” permite que subgrupos pequenos (que cobrem poucos casos da variável alvo) consigam ou não competir com subgrupos grandes.
- Parâmetro “a” é o ajuste fino que permite o controle entre encontrar subgrupos mais representativos e menos distintos ou subgrupos menos representativos e mais distintos.

Experimentos - PUCHASES_TRX - Quem utiliza muito o cartão?

Beam Search + Diferença da Média Ponderada:

```
return instances_subgroup**a * (mean_sg - mean_dataset)
```



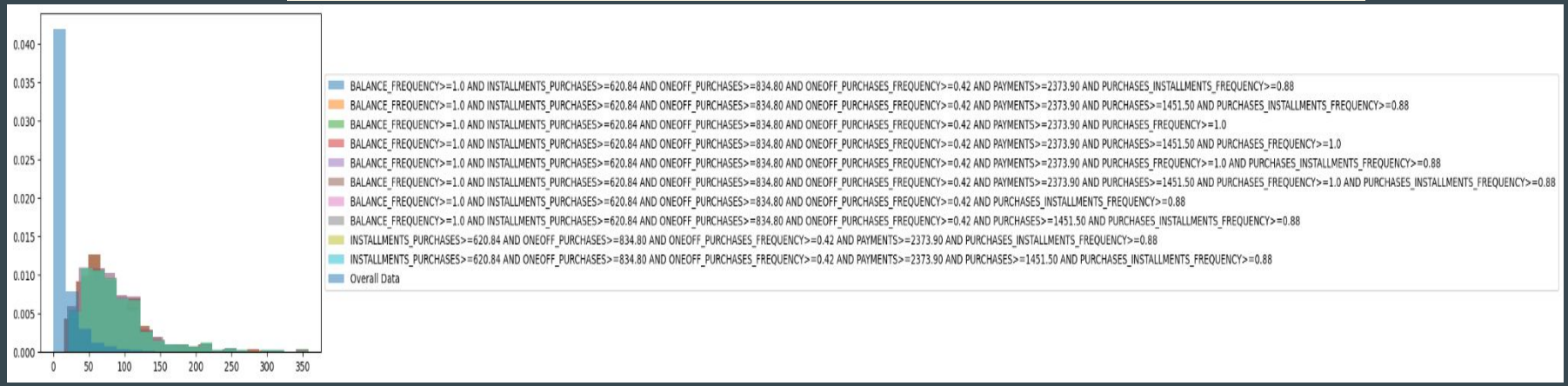
top-10 • depth = 8 • a = 0.5 • competição entre todos os seletores

Subgrupos descritos quase que completamente pelo comportamento de compra.

Experimentos - PUCHASES_TRX - Quem utiliza muito o cartão?

Beam Search + Diferença da Média Ponderada:

```
return instances_subgroup**a * (mean_sg - mean_dataset)
```



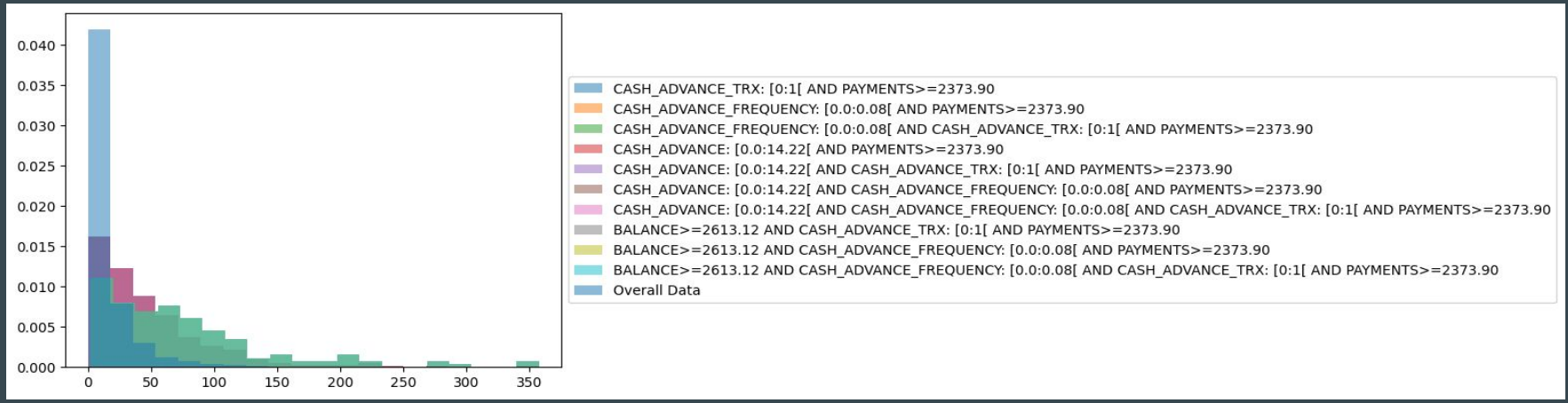
top-10 • depth = 8 • a = 0.3 • competição entre todos os seletores

Subgrupos mais deslocados. Aqui aparecem mais variáveis relacionadas à finanças.

Experimentos - PUCHASES_TRX - Quem utiliza muito o cartão?

Beam Search + Diferença da Média Ponderada:

```
return instances_subgroup**a * (mean_sg - mean_dataset)
```



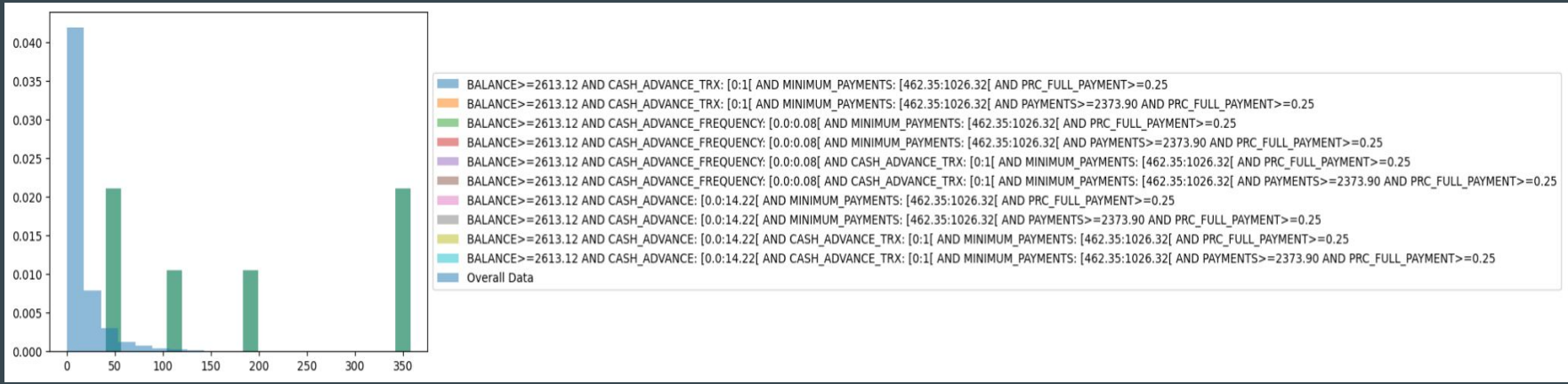
top-10 • depth = 8 • a = 0.5 • apenas seletores de finanças

As descrições financeiras tem uma representatividade bem menor.

Experimentos - PUCHASES_TRX - Quem utiliza muito o cartão?

Beam Search + Diferença da Média Ponderada:

```
return instances_subgroup**a * (mean_sg - mean_dataset)
```



top-10 • depth = 8 • a = 0.3 • apenas seletores de finanças

Diminuir o “a” nesse caso retorna subgrupos que chegam a cobrir apenas um usuário.

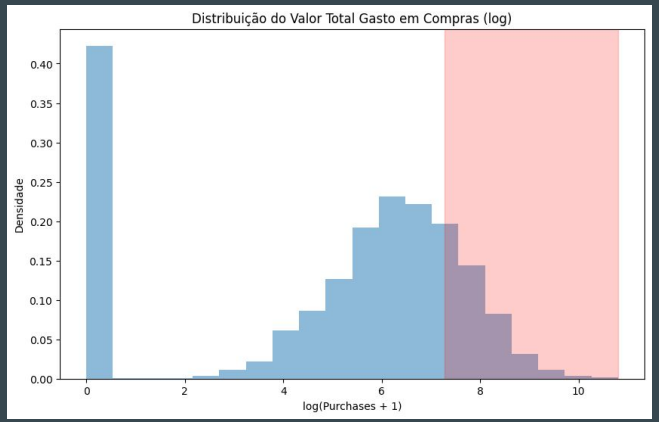
Experimentos - PUCHASES_TRX - Quem utiliza muito o cartão?

Beam Search + Diferença da Média Ponderada:

```
return instances_subgroup**a * (mean_sg - mean_dataset)
```

- O entendimento de como controlar esse parâmetro fundamental permite a colaboração com especialistas, ajustando a descoberta de subgrupos de acordo com sua demanda.

PURCHASES>=1451.50 AND PURCHASES_FREQUENCY>=1.0



Resultados

top-10 • depth = 8 • a = 0.5 • competição entre todos os

	quality	subgroup	size_sg	size_dataset	mean_sg	mean_dataset	std_sg	std_dataset	median_sg	median_dataset	max_sg	max_dataset	min_sg	min_dataset	mean_lift	median_lift
0	1376.919934	PURCHASES>=1451.50 AND PURCHASES_FREQUENCY>=1.0	981	8636	58.994903	15.033233	44.375460	25.17901	48.0	7.0	358	358	6	0	3.924299	6.857143
1	1374.712320	PURCHASES>=1451.50 AND PURCHASES_FREQUENCY>=1....	953	8636	59.564533	15.033233	44.559359	25.17901	48.0	7.0	358	358	12	0	3.962191	6.857143
2	1372.529295	BALANCE_FREQUENCY>=1.0 AND PURCHASES>=1451.50 ...	942	8636	59.752654	15.033233	44.916074	25.17901	49.0	7.0	358	358	6	0	3.974704	7.000000
3	1370.005766	BALANCE_FREQUENCY>=1.0 AND PURCHASES>=1451.50 ...	917	8636	60.274809	15.033233	45.070849	25.17901	49.0	7.0	358	358	12	0	4.009438	7.000000

top-10 • depth = 8 • a = 0.3 • competição entre todos os

	quality	subgroup	size_sg	size_dataset	mean_sg	mean_dataset	std_sg	std_dataset	median_sg	median_dataset	max_sg	max_dataset	min_sg	min_dataset	mean_lift	median_lift
0	389.750632	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHA...	240	8636	90.320833	15.033233	53.076450	25.17901	79.5	7.0	358	358	22	0	6.008078	11.357143
1	389.750632	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHA...	240	8636	90.320833	15.033233	53.076450	25.17901	79.5	7.0	358	358	22	0	6.008078	11.357143
2	389.132812	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHA...	299	8636	85.404682	15.033233	53.769732	25.17901	71.0	7.0	358	358	15	0	5.681059	10.142857
3	389.132812	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHA...	299	8636	85.404682	15.033233	53.769732	25.17901	71.0	7.0	358	358	15	0	5.681059	10.142857

Interpretação dos Experimentos

- A partir dos testes podemos entender que:
  - A utilização de descritores apenas numéricos pode dificultar a interpretação dos resultados.
  - A utilização de um alvo binário com uma otimização de subgrupos pela diferença da média ponderada permite uma flexibilização da análise, permitindo a exploração de subgrupos mais representativos ou mais específicos de acordo com as demandas da análise.

Análise - Quem são os clientes com dificuldade de pagar a fatura?

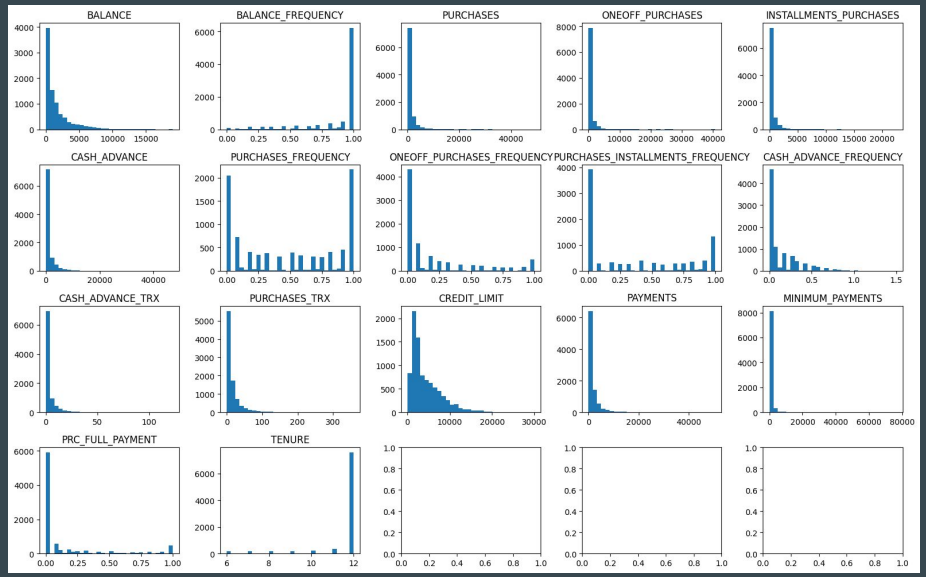
A análise foi realizada nos seguintes passos:

1. Pré-processamento dos dados,
2. Definição de uma variável binária e
3. Descoberta de subgrupos.
4. Visualização de subgrupos.

Análise - Quem são os clientes com dificuldade de pagar a fatura?

A análise foi realizada nos seguintes passos:

1. Pré-processamento dos dados:



- Muitas colunas com assimetria nas distribuições.
- Muitos outliers.

- Discretização das colunas numéricas em quartis.
  - Nem todas podem ser discretizadas.
- Preenchimento de valores nulos pela mediana.

Análise - Quem são os clientes com dificuldade de pagar a fatura?

A análise foi realizada nos seguintes passos:

2. Definição de uma variável binária:

MIN_PAY_RATIO_HIGH = (MINIMUM_PAYMENTS / PAYMENTS) > 0.9

- Informações redundantes presentes em colunas correlacionadas são sintetizadas em um rótulo.
- Isso facilita a análise dos dados e a posterior interpretação dos resultados.

Análise - Quem são os clientes com dificuldade de pagar a fatura?

A análise foi realizada nos seguintes passos:

3. Descoberta de subgrupos:

Parâmetros:

- Algoritmo: Beam Search
- Ranking: top-10
- Profundidade: 3
- Função de Qualidade: WRAcc

Composição das descrições:

Variáveis Categóricas: SALDO, COMPRAS, FREQUÊNCIA_COMPRAS, NÚMERO_COMPRAS, LIMITE_CRÉDITO e POSSE.

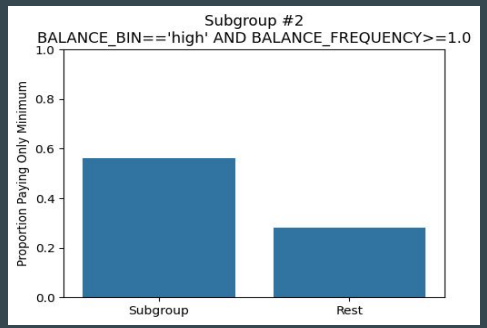
Variáveis Numéricas: FREQUÊNCIA_SALDO, COMPRAS_UMA_PARCELA, COMPRAS_VÁRIAS_PARCELAS, SAQUE, FREQUÊNCIA_COMPRAS_UMA_PARCELA, FREQUÊNCIA_COMPRAS_VÁRIAS_PARCELAS, FREQUÊNCIA_SAQUE, NÚMEROS_SAQUE.

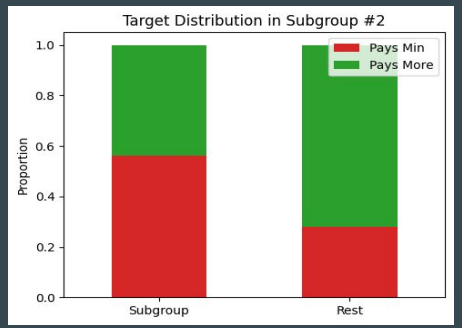
Análise - Quem são os clientes com dificuldade de pagar a fatura?

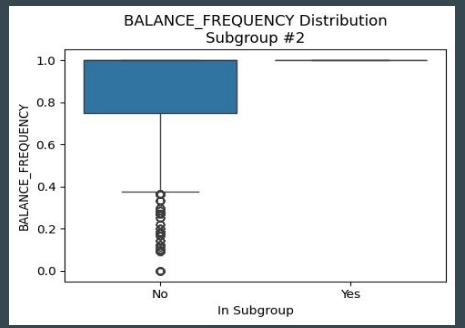
A análise foi realizada nos seguintes passos:

4. Visualização de subgrupos:

	quality	subgroup	size_sg	size_dataset
0	0.053724	BALANCE_BIN=='high'	2238	8950
1	0.050012	BALANCE_BIN=='high' AND BALANCE_FREQUENCY>=1.0	2050	8950
2	0.045886	BALANCE_FREQUENCY>=1.0 AND INSTALLMENTS_PURCHASES: [0.0:1.95[	2607	8950







Conclusões

- A aplicação do Beam-Search para buscar subgrupos foi bem objetiva e direta
- Dificuldades impostas pelos dados:
  - Dificuldade de interpretação dos subgrupos, especialmente quando foram utilizados alvos numéricos com todos os descritores também numéricos.
- A utilização de um alvo binário com uma mistura de seletores categóricos e contínuos para compor a descrição ajudou na geração de subgrupos mais facilmente interpretáveis e interessantes.

Credit Card Customer: Descoberta de subgrupos utilizando Beam-Search

Amanda Mendes Pinho
Gabriel Tonioni Duarte
João Vítor Fernandes Dias
Larissa Duarte Santana