2º ESTUDO APLICADO | VALIDAÇÃO BASE

PREVISÃO DA EVASÃO

15 Maio, 2024 | Ingrid Cardoso

AGENDA

Váriáveis ou Parâmetros?

Correlação

O que mudou?

Teste Prático

Thanks

VARIÁVEIS OU PARÂMETROS?

NATUREZA DA QUESTÃO

Algoritmos de Machine Learning como Logistic Regression Classifier, Random Forest e AdaBoostClassifier não levam explicitamente em consideração a correlação temporal dos dados. Esses algoritmos são frequentemente aplicados em problemas onde a ordem dos dados não é relevante e não é considerada diretamente no processo de modelagem.

No entanto, isso não significa necessariamente que esses algoritmos não possam ser aplicados em problemas com dados temporais. Eles podem ser usados, por exemplo, em problemas de classificação onde cada amostra de dados é independente das outras. Por exemplo, se você estiver classificando emails como spam ou não spam, a ordem dos emails pode não importar.

NATUREZA DA QUESTÃO

Se você tem um conjunto de dados onde cada entrada é um vetor de timestamps e cada posição do timestamp está relacionada a informações de vários clientes, isso pode ser interpretado como uma série temporal multivariada, onde cada cliente é uma tabela em diferentes timestamps.

Sendo assim, o modelo aprenderá padrões nos dados de cada cliente e usará esses padrões para fazer previsões individuais sobre se cada cliente em determinado momento está propenso a evadir ou não. A predição seria individual para cada cliente.

Modelos de ML podem ser utilizados com dados de séries temporais, mas é necessário considerar cuidadosamente a natureza dos seus dados e os pressupostos do modelo para garantir resultados confiáveis.

POR QUE A ESTACIONARIDADE É IMPORTANTE?

A estacionaridade dos dados é crucial para modelos de classificação em machine learning, especialmente quando se trata de séries temporais como retornos diários de uma carteira de investimentos. A estacionaridade refere-se à propriedade de que as propriedades estatísticas dos dados, como média e variância, permanecem constantes ao longo do tempo.

Muitos modelos de séries temporais, incluindo aqueles usados em machine learning, assumem estacionaridade.

Se os dados não forem estacionários, os resultados desses modelos podem ser imprecisos ou inválidos.

Modelos treinados em dados estacionários tendem a ter parâmetros mais estáveis.

Isso significa que as relações aprendidas pelo modelo são mais confiáveis e podem ser generalizadas para dados futuros.

A estacionaridade ajuda a identificar padrões temporais consistentes nos dados.

Isso permite que o modelo capture tendências, sazonalidades e ciclos, que são importantes para prever os retornos futuros da carteira de investimentos.

A interpretação dos resultados do modelo é facilitada pela constância.

As relações entre as variáveis tendem a permanecer constantes ao longo do tempo, facilitando a compreensão e a tomada de decisões com base nos resultados do modelo.

SOLUÇÕES PARA ESTACIONARIDADE

MODELAGEM DE DADOS VOLTADA PARA DIFERENCIAÇÃO, SUAVIZAÇÃO E CAPTURA DE SAZONALIDADE E TEMPO.

DIFERENÇA DAS BASES

VALORES NULOS

```
posicao_compromissadas          1.000000
pl_offshore_brl_avenue          1.000000
pl_offshore_usd_avenue          1.000000
pl_wealth                       0.999618
total_contato                   0.976791
total_first_meetings            0.922281
posicao_prt                     0.165141
posicao_previdenciaprivada      0.161383
posicao_produtosestruturados    0.161383
posicao_seguros                 0.161383
posicao_termos                  0.161383
```

Variáveis não presentes

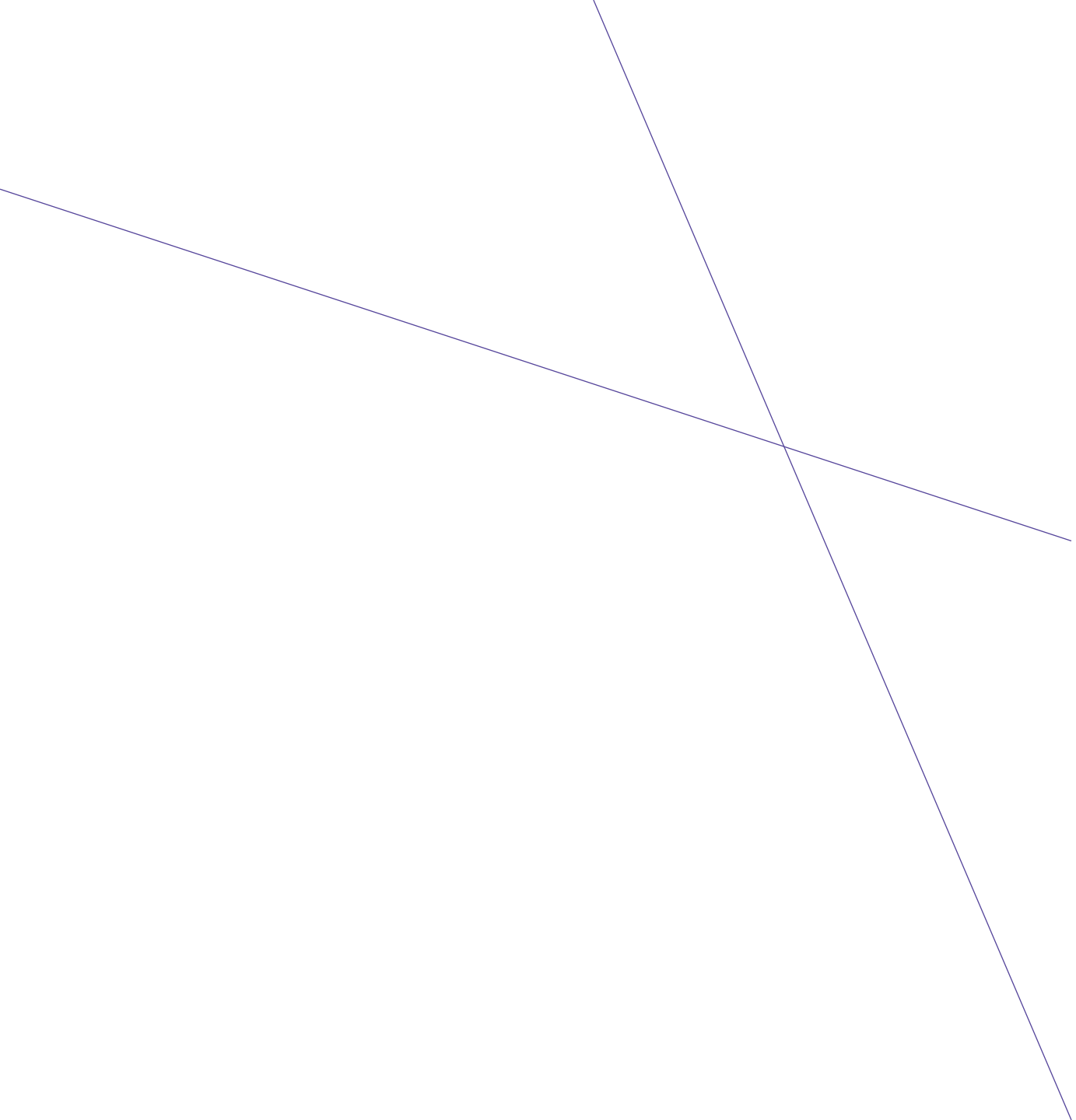
Valor cartão,
Transferência ted (in & out),
De aquisição ou expansão,
Nacionalidade

Ignoradas

45 no total por estarem com
Mais de 90% dos valores nulos.

LISTA DOS IGNORADOS

bmf, bovespa, btc, cambio, categoria, categoria_assessor_nps, cliente_coberto_produtos, cliente_coberto_produtos_e_rv, cliente_coberto_rv, cliente_coberto_total cliente_coberto_total_mensal, cliente_exposto_rv, clube, compromissadas, credito, equipe_investor equipe_prospector, fee_fixo, flag_fonte_online, flag_fundo_exclusive, flag_pj_gestao_de_Caixa, Funcao, fundos, ipo_fee_rf, is_transfer_in, mercado_de_capitais, multa_transfer_in, nonliquidity_funds_allocated_value, nps_promotor, operador_username, periodic_contact_isdone, periodic_contact_was_successfully_done, pl_investimentos_declarado_ajustado Plrendavariavel, posicao_saldo_projetado, prev_xpcs, previous_office, prospector_name, segmento_assessor_nps, status_mes_anterior, transfer_out_date, type

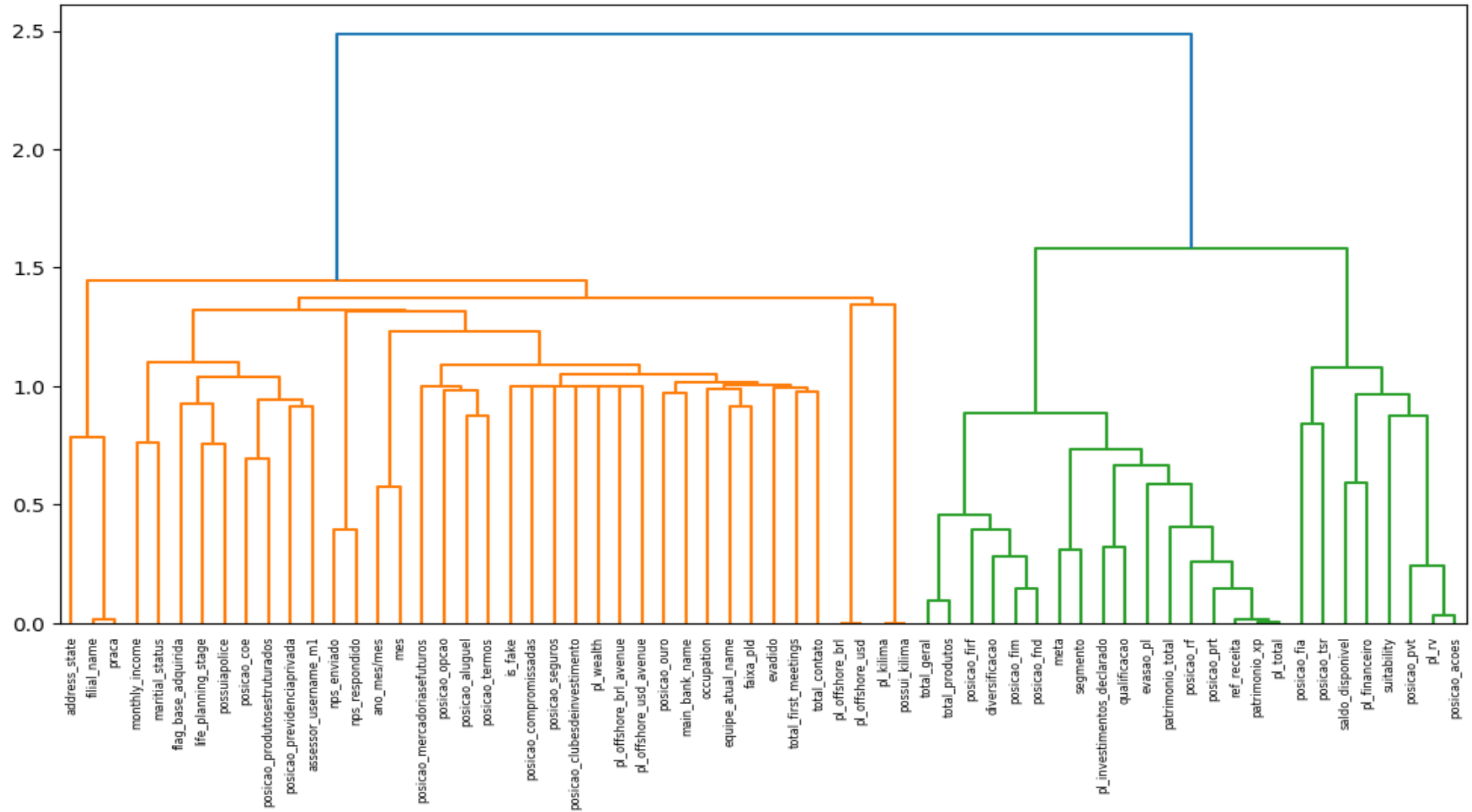


FEATURING DATA

FEATURES TEMPORAIS

- **PL/CDI:** Razão entre o DI e o patrimônio do mês em questão.
- **PL_(PRODUTO)/PATRIMÔNIO:** Razão entre variáveis e patrimonio líquido por mês.
- Total Contato: Soma de todos os contatos feitos ao cliente no mês em questão (assessor, investor e operador).
- **Posições_(produto):** Agrupamento das posições por classe: fundos, opção, portfolio, proventos, renda fixa, tesouro.

FEATURING DATA – CORRELAÇÃO E COLINEARIDADE



COLINEAR E NÃO CORRELACIONADAS

'patrimonio_total’, 'pl_kilima', 'pl_offshore_brl',
'pl_offshore_brl_avenue', 'pl_offshore_usd_avenue', 'pl_rv',
'pl_wealth', 'posicao_aluguel', 'posicao_clubesdeinvestimento',
'posicao_coe', 'posicao_compromissadas', 'posicao_fia',
'posicao_mercadoriasefuturos', 'posicao_ouro',
'posicao_previdenciaprivada', 'posicao_seguros', 'saldo_disponivel',
'assessor_username_m1', 'equipe_atual_name', 'faixa_pld',
'filial_name', 'flag_base_adquirida', 'life_planning_stage',
'monthly_income', 'nps_enviado', 'total_first_meetings', 'evadido',
'main_bank_name', 'occupation', 'posicao_opcao'

SAZONALIDADE

- **Rolling Windows:** Essa função é útil para calcular medidas de volatilidade, como o desvio padrão, em janelas rolantes, ajudando a entender a estabilidade ou instabilidade dos dados ao longo do tempo. Janela de 3 meses.
- **Variação Percentual:** Essa função é útil quando você está interessado nas variações percentuais entre os pontos de dados consecutivos, como a taxa de crescimento ou declínio de uma série temporal. Janela mensal.
- **Diferenciação:** Essa função é útil quando você está interessado na diferençaabsoluta entre os pontos de dados consecutivos, sem considerar a escala ou a direção da mudança. Janela mensal.

LIMITAÇÕES

- **Não foi possível realizar o teste sazional, de tendencia, ou residual de tendencia, pois o período de meses do histórico é menor que 24.**

PARÂMETRO ALVO

QUEM É O ‘EVASOR’?

Parâmetro de perfil

Qualquer cliente que possui-se os últimos 2 meses com o patrimônio_total **(pl_total) zerado** foi selecionado como perfil de evasor (cliente sem intenção de voltar).

```
data.loc[237104][['pl_total', 'evasao_pl', 'evadido']]
```

[6] ✓ 0.0s

account_xp_code	pl_total	evasao_pl	evadido
237104	193.210000	1	False
237104	193.010000	1	False
237104	193.690002	1	False
237104	194.309998	1	True
237104	63.340000	1	False
237104	118.209999	1	False
237104	119.440002	1	False
237104	120.459999	1	False
237104	120.830002	1	False
237104	122.010002	1	False
237104	122.940002	1	True
237104	0.000000	1	False
237104	0.000000	1	False

COMPARAÇÃO

MODELO 1

NÚMERO DE EXEMPLOS (TREINO)

(antigo | Coluna): 16K

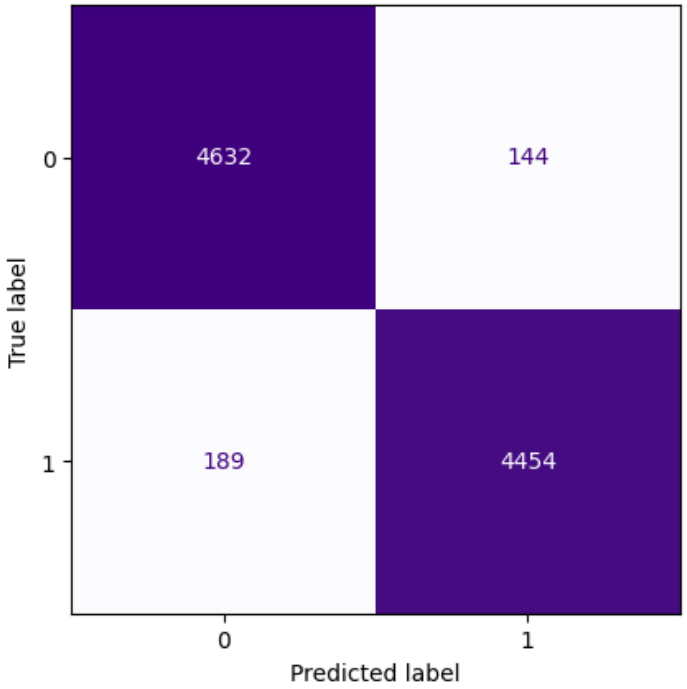
(novo | Linha): 38K

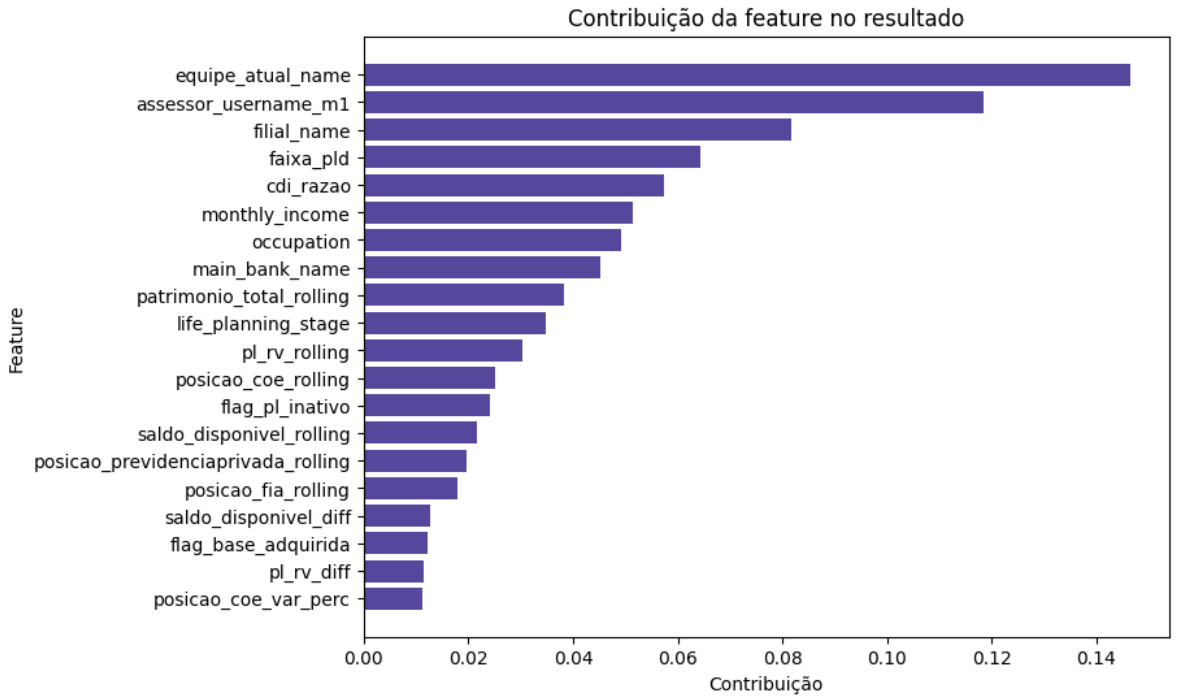
ACURÁCIA MÉDIA (TREINO E TESTE)

(antigo | Coluna): 99%

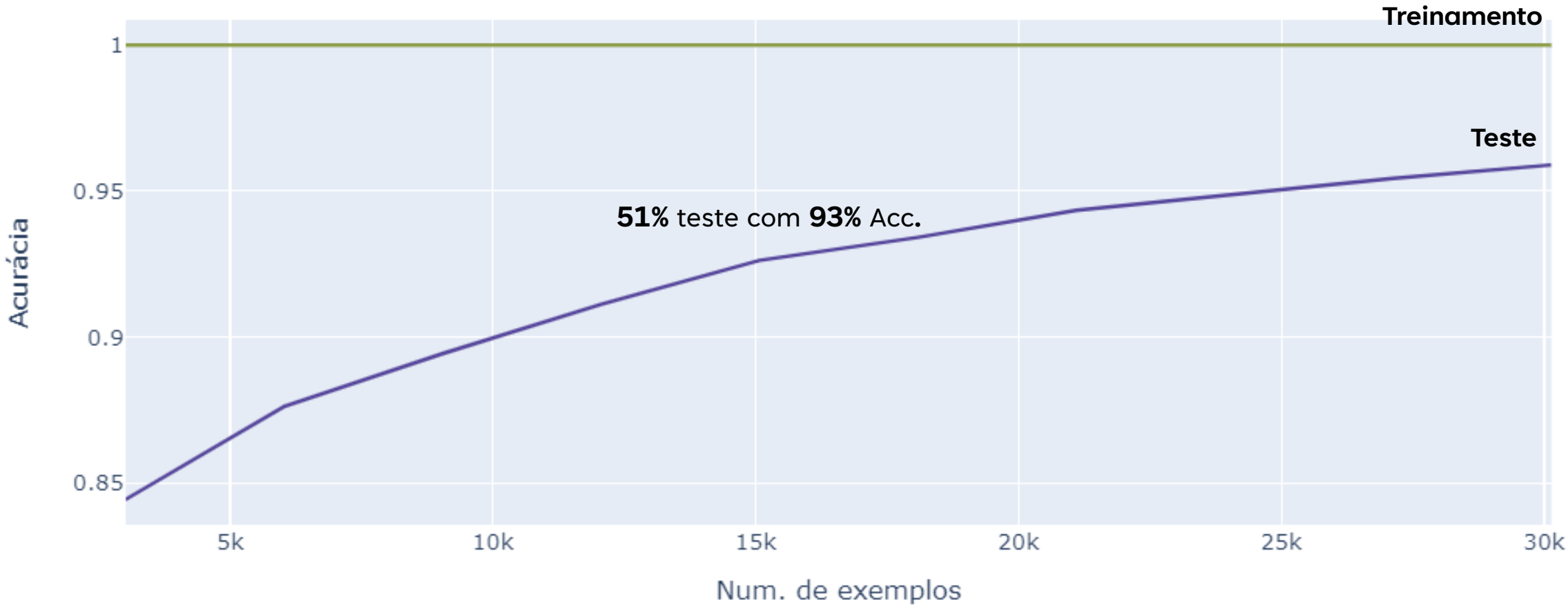
(novo | Linha): 96%

PERFORMANCE (9.4K EXEMPLOS)





CURVA DE APRENDIZADO (TREINAMENTO)



MODELO 2

NÚMERO DE EXEMPLOS (TREINO)

(antigo | Coluna): 16K

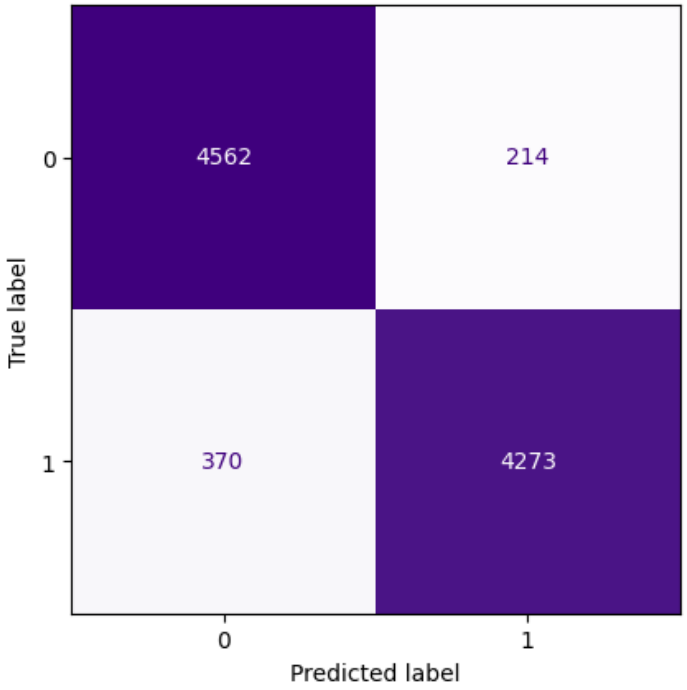
(novo | Linha): 38K

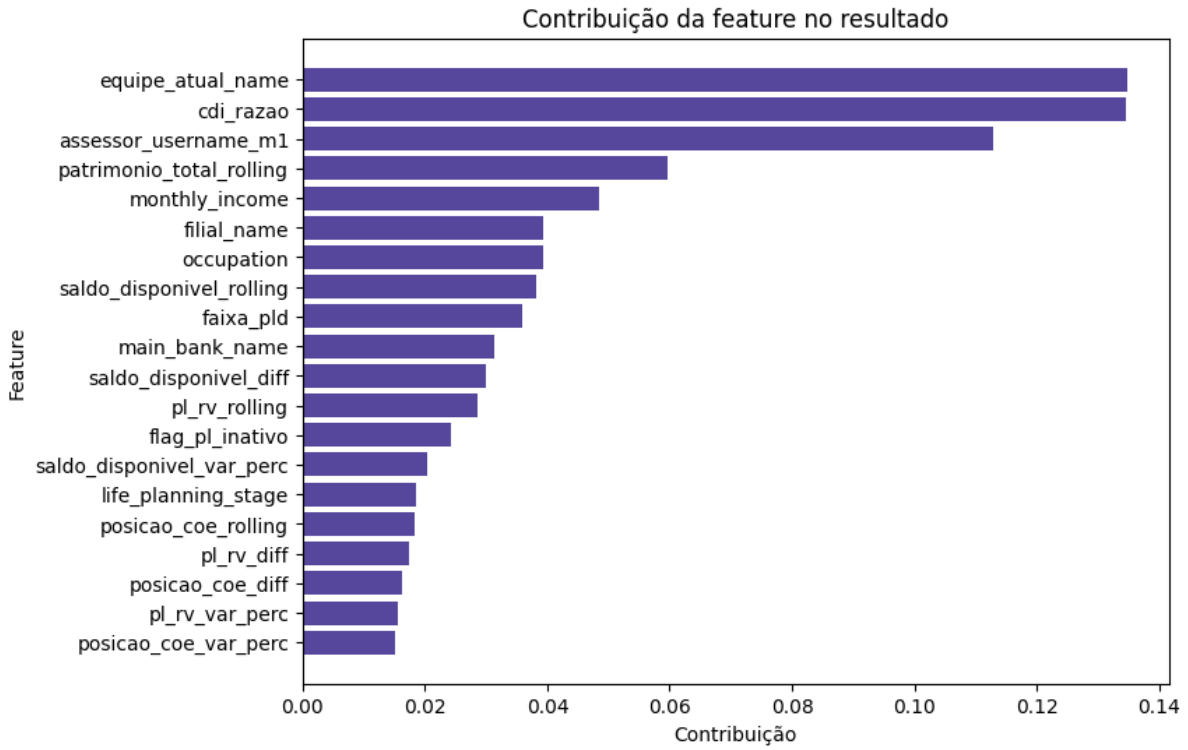
ACURÁCIA MÉDIA (TREINO E TESTE)

(antigo | Coluna): 98%

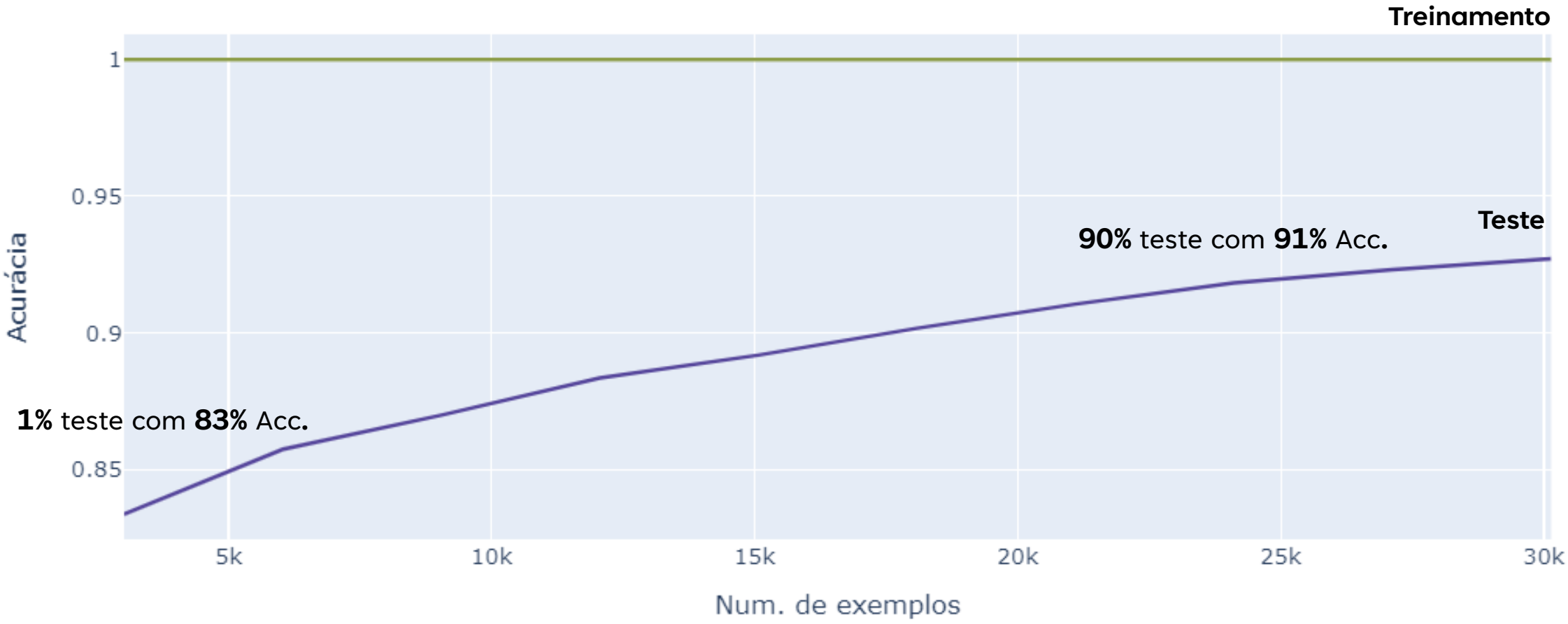
(novo | Linha): 93%

PERFORMANCE (9.4K EXEMPLOS)





CURVA DE APRENDIZADO (TREINAMENTO)



MODELO 3

NÚMERO DE EXEMPLOS (TREINO)

(antigo | Coluna): 16K

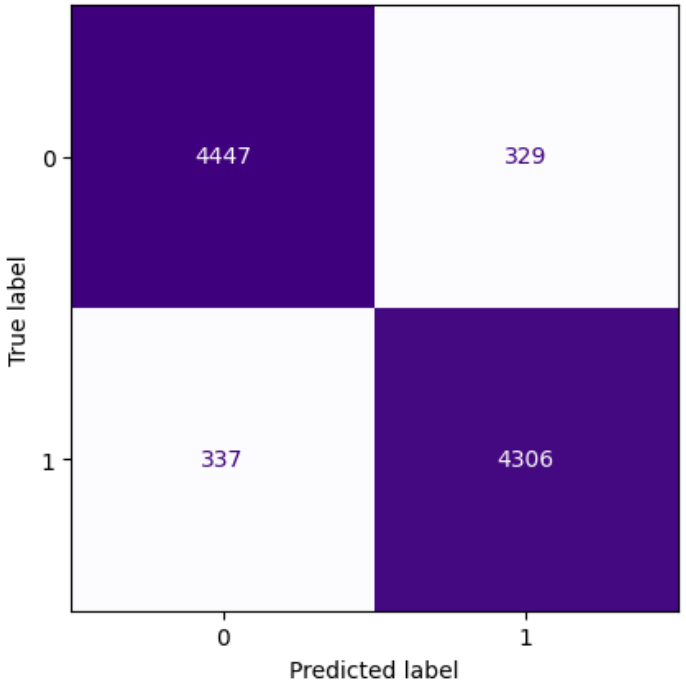
(novo | Linha): 38K

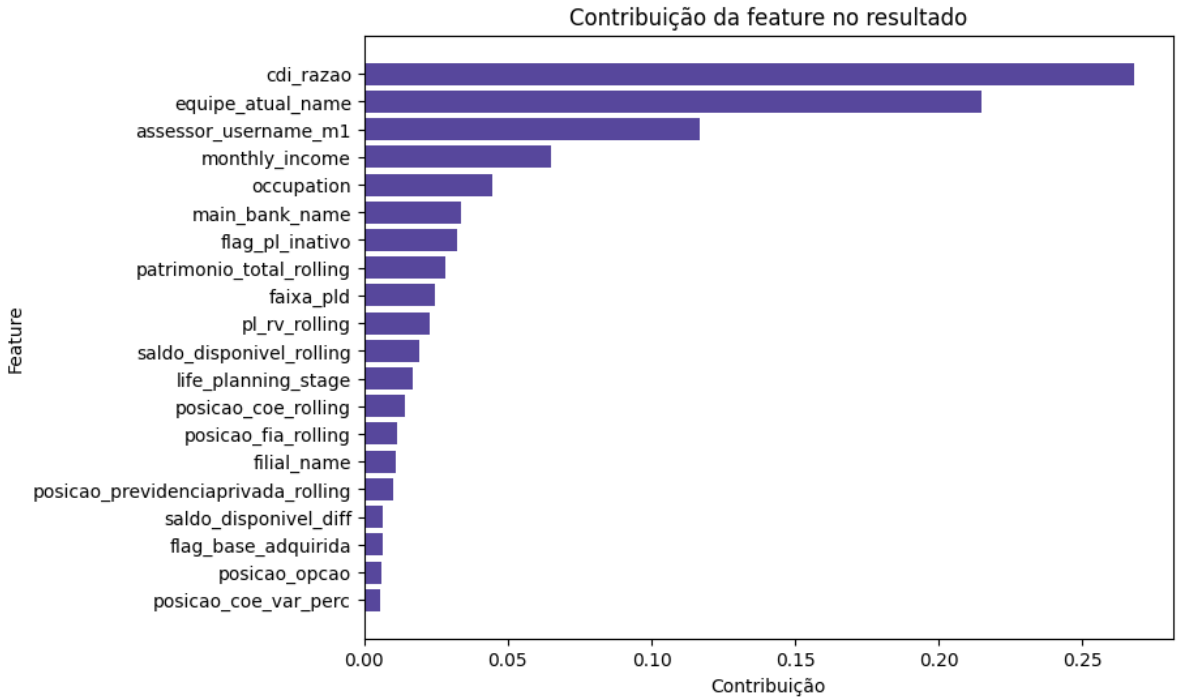
ACURÁCIA MÉDIA (TREINO E TESTE)

(antigo | Coluna): 98%

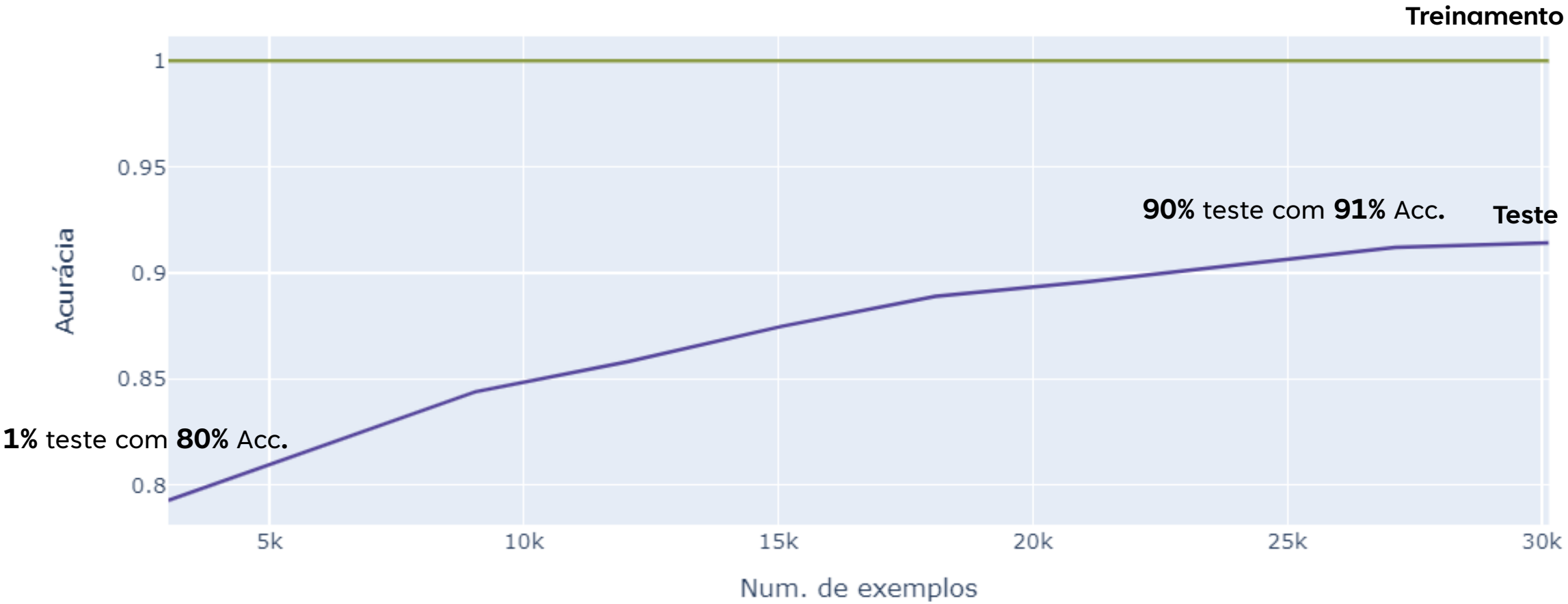
(novo | Linha): 92%

PERFORMANCE (9.4K EXEMPLOS)





CURVA DE APRENDIZADO (TREINAMENTO)



2º ESTUDO APLICADO | TESTE DOS 5

VALIDAÇÃO MODELO

15 Maio, 2024 | Ingrid Cardoso

OS 5 VALIDADORES

```
[2]  cinco_evadido.index
     ✓ 0.0s
```

```
Index([428068, 2028974, 3512051, 223499, 50935], dtype='int64', name='account_xp_code')
```

```
[3]  cinco_ativo.index
     ✓ 0.0s
```

```
Index([11446828, 247312, 2439622, 12193571, 2745124], dtype='int64', name='account_xp_code')
```

RESULTADO DO PREVISTO VS. CLASSE REAL

Clientes de valor

```
[8]  models_ativo
     ✓ 0.0s
```

account_xp_code	0	1	previsao	label
11446828	100.0	0.0	0	0
247312	93.0	7.0	0	0
2439622	100.0	0.0	0	0
12193571	100.0	0.0	0	0
2745124	80.0	20.0	0	0

Clientes de não-valor

```
[10] models_evadido
     ✓ 0.0s
```

account_xp_code	0	1	previsao	label
428068	0.0	100.0	1	1
2028974	0.0	100.0	1	1
3512051	0.0	100.0	1	1
223499	0.0	100.0	1	1
50935	0.0	100.0	1	1

THANKS!

Beatriz - Líder da equipe e dados que dá super apoio aos estudos

Leo - Por me passar a base super-rápido XD

Luiz - Por me acompanhar e trazer insights INCRÍVEIS!

Codes, modelos e validação no github

https://github.com/Ingrid-0906/Evasao_Modelagem-validacao2/