




# Análise de Features

Em quais características focar para a  
retenção de capital?

**Matheus Felipe**  
[mathlippo@gmail.com](mailto:mathlippo@gmail.com)

# Sumário

- ◆ Objetivo.
  - ◆ Colunas.
  - ◆ Análise Descritiva.
  - ◆ Análise Temporal.
  - ◆ Análise de Benchmark.
  - ◆ Modelo.
  - ◆ Features mais importantes.
  - ◆ Finalizando.
- 

# Objetivo:

Como analista de dados, meus objetivos nessa pesquisa são:

- Entender o perfil de clientes que efetivam depósito à prazo que o banco possui.
- Analisar métricas e as variáveis de rotina que possam influenciar a produtividade.
- Extrair as variáveis que mais impactam na captação de um depósito à prazo.

Dataset: Bank Marketing.

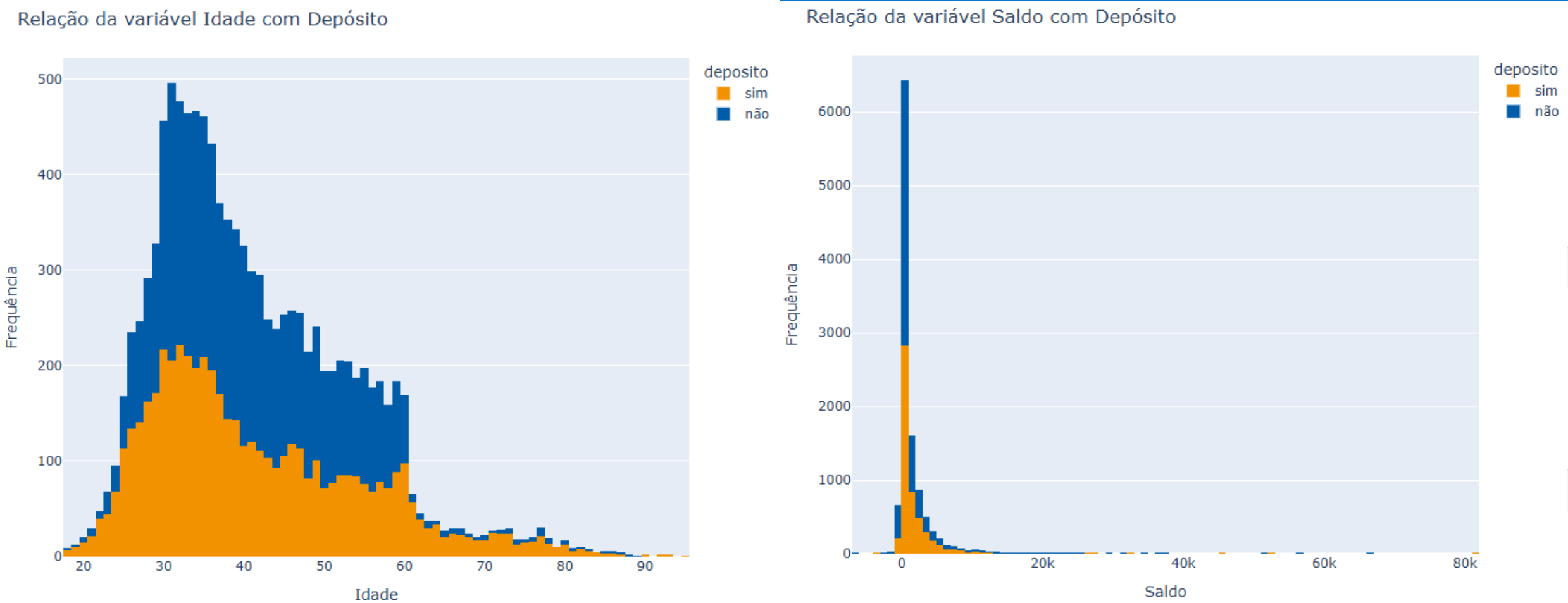
Contexto: Campanhas de marketing direto (chamadas de telefone) de uma instituição bancária portuguesa.

# Colunas :

- idade
- trabalho
- estado\_civil
- escolaridade
- inadimplente
- saldo
- emp\_habitacional
- emp\_pessoal
- tipo\_contato
- dia
- mes
- tempo\_chamada
- campanha\_atual
- dias\_prévios
- campanha\_anterior
- res\_campanha
- deposito

- Idade do cliente.
- Tipo de emprego.
- Estado civil.
- Nível de educação.
- Possui crédito em inadimplência?
- Saldo médio anual (em euros).
- Possui empréstimo habitacional?
- Possui empréstimo pessoal?
- Tipo de comunicação de contato.
- Último dia de contato da semana.
- Último mês de contato do ano.
- Duração do último contato, em segundos (inteiro).
- Número de contatos nesta campanha para o cliente.
- Dias passados desde último contato de campanha anterior.
- Número de contatos antes desta campanha para o cliente (inteiro).
- Resultado da campanha anterior.
- O cliente assinou um depósito a prazo?

# Análise Descritiva: variáveis numéricas



## Variáveis descritivas:

	count	mean	std	min	25%	50%	75%	max
idade	11162.0	41.231948	11.913369	18.0	32.0	39.0	49.0	95.0
saldo	11162.0	1528.538524	3225.413326	-6847.0	122.0	550.0	1708.0	81204.0

## Correlação:

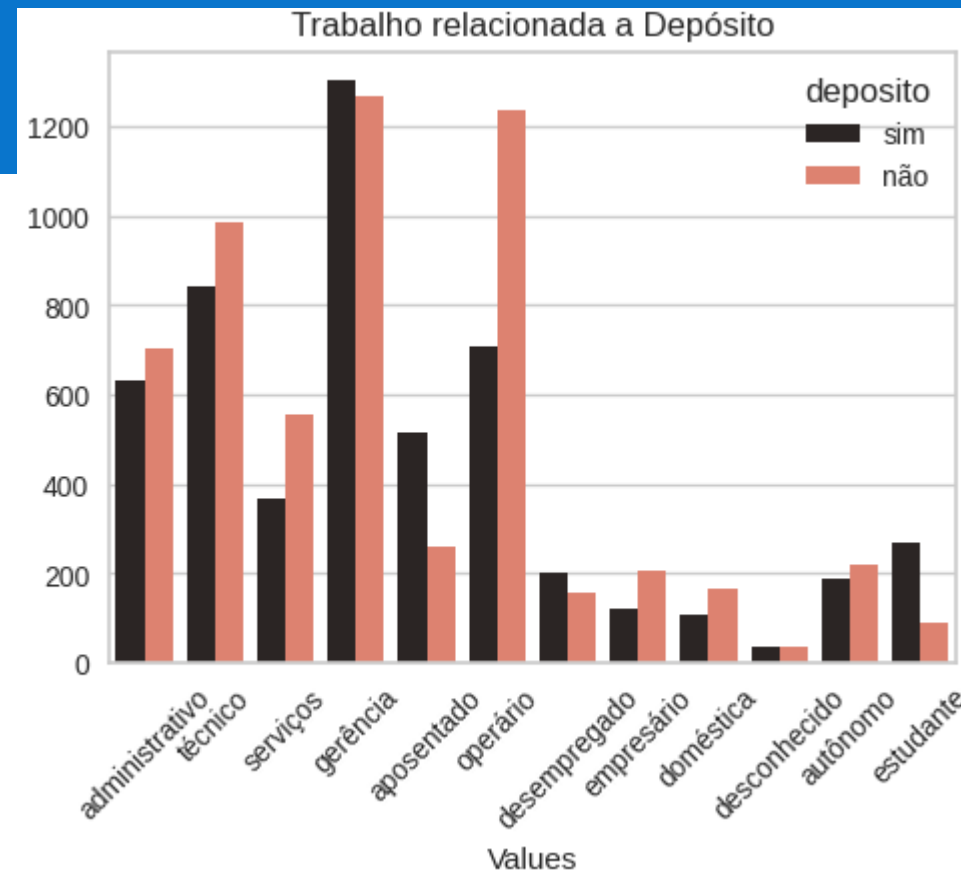
	idade	saldo
idade	1.0000	0.1123
saldo	0.1123	1.0000

Não há correlação aparente entre as idades e os saldos dos clientes,

Chances maiores de depósito para as pessoas com idade menor que 30 anos ou maior que 60 anos. Para as idades entre 30 e 60 anos, parte que contém a maioria dos clientes do banco, as chances de não depósito são levemente maiores.

Para as pessoas que possuem acima de 1000 euros, as chances de depósito aumentam progressivamente. Espera-se, já que a mediana é de 550 euros, que mais pessoas não depositarão parte do seu capital em um depósito à prazo.

# Análise Descritiva: variáveis categóricas e binárias

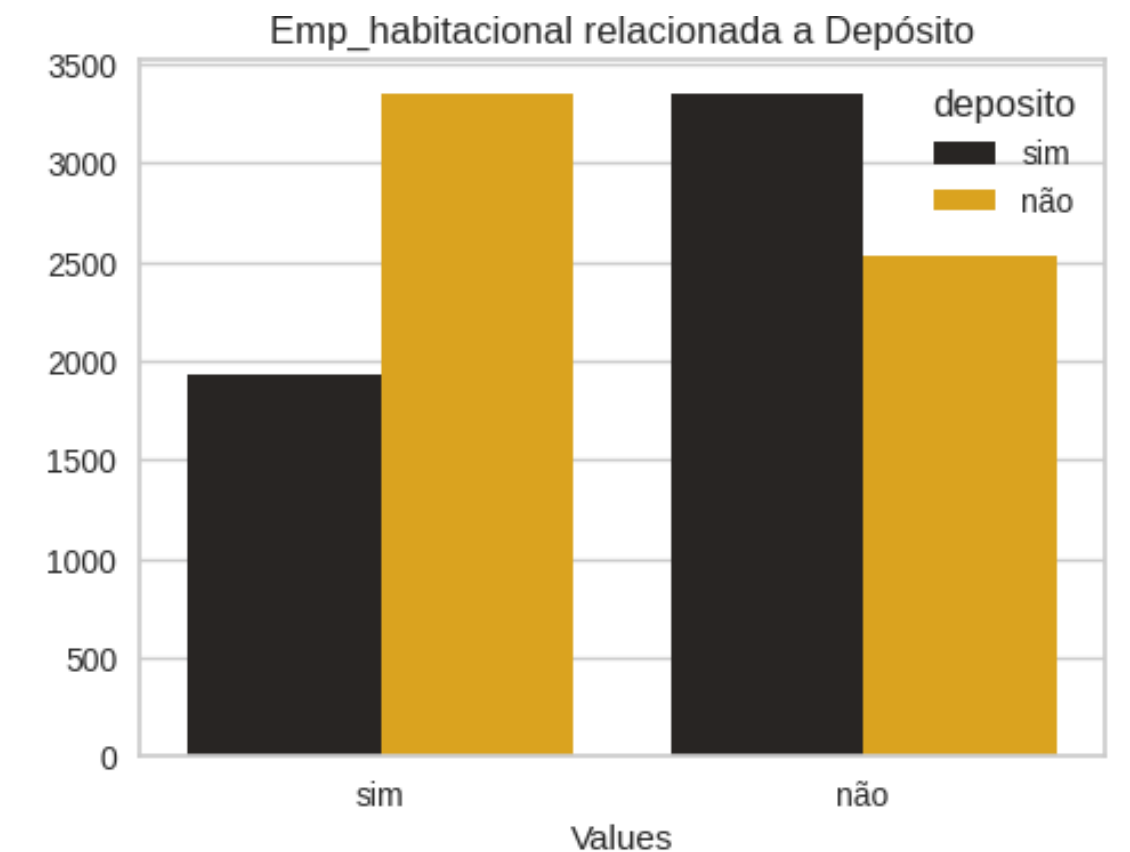
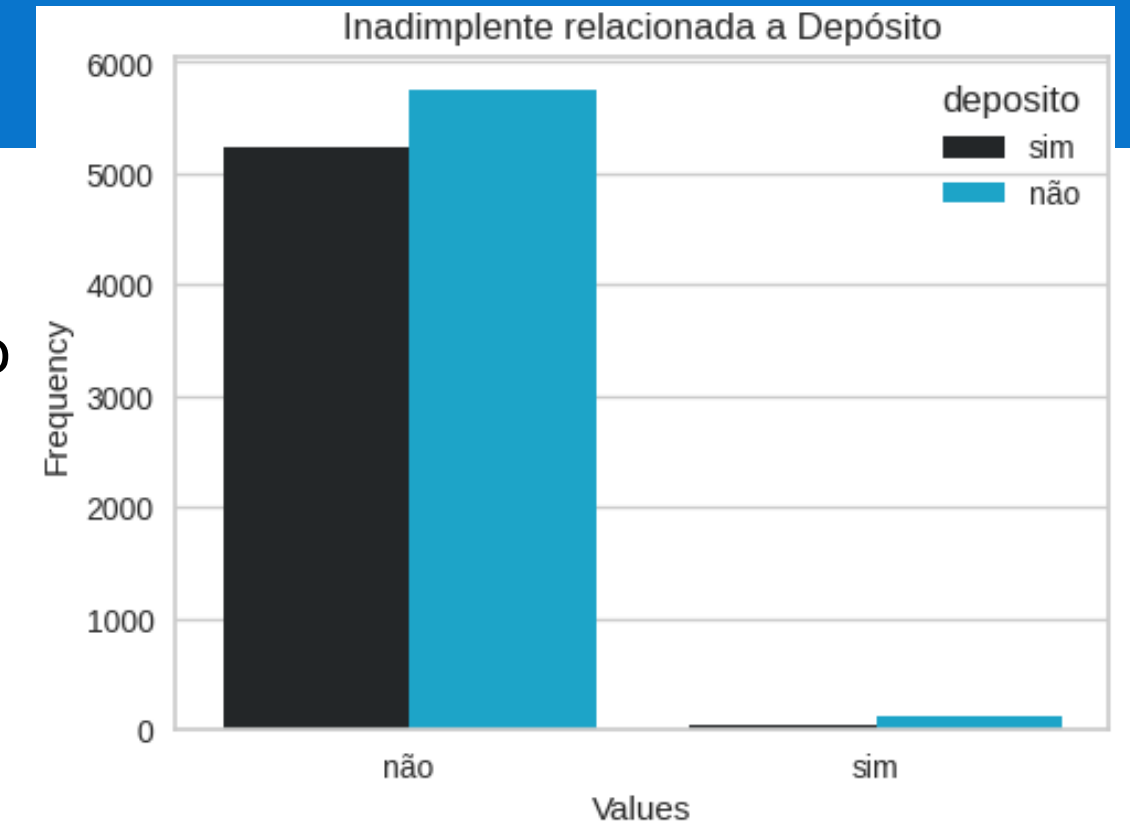
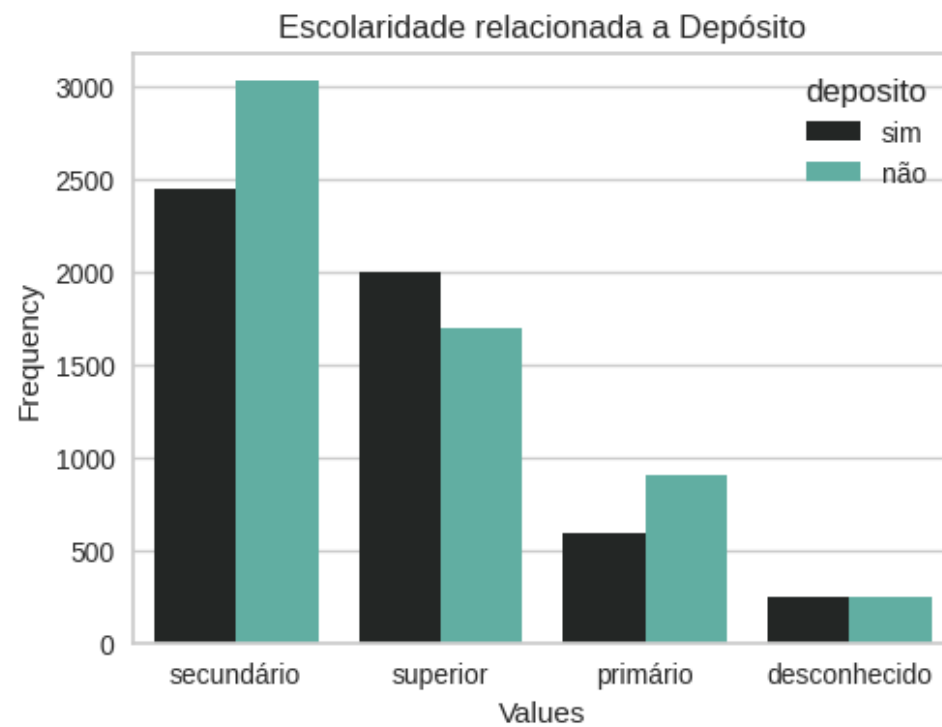


## Análise:

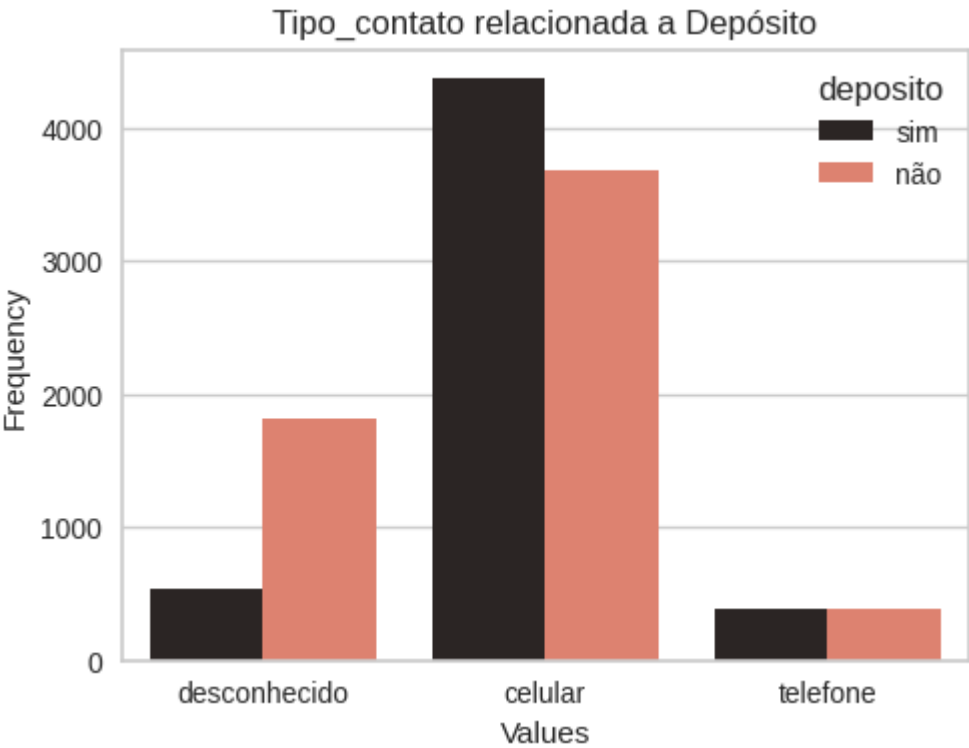
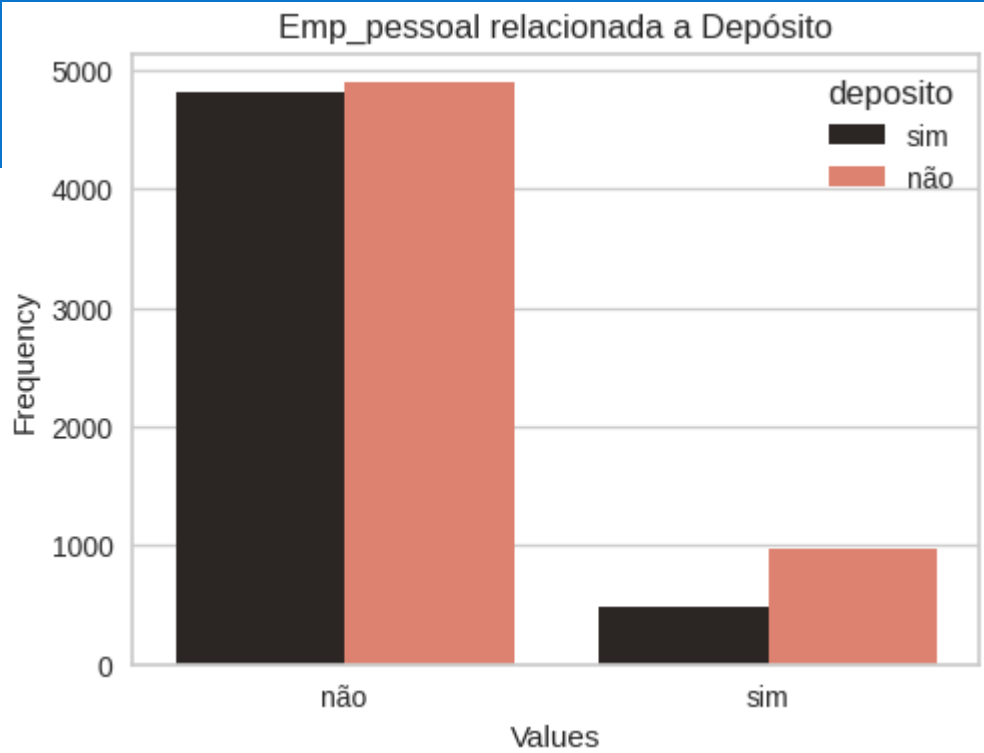
Impacto considerável entre determinadas profissões, escolaridades e caso o cliente possua um empréstimo habitacional na escolha do depósito, em relação à inadimplência, nota-se que não há uma boa amostragem nos clientes inadimplentes para aferir uma escolha.

## Variáveis de impacto:

- **Trabalho:** como a alta taxa de depósito entre estudantes e aposentados e baixa taxa entre operários.
- **Escolaridade:** pouca diferença entre as frequências, sendo os formados os que mais efetuam depósito.
- **Empréstimo habitacional:** clientes que não possuem são mais propensos ao depósito de capital.

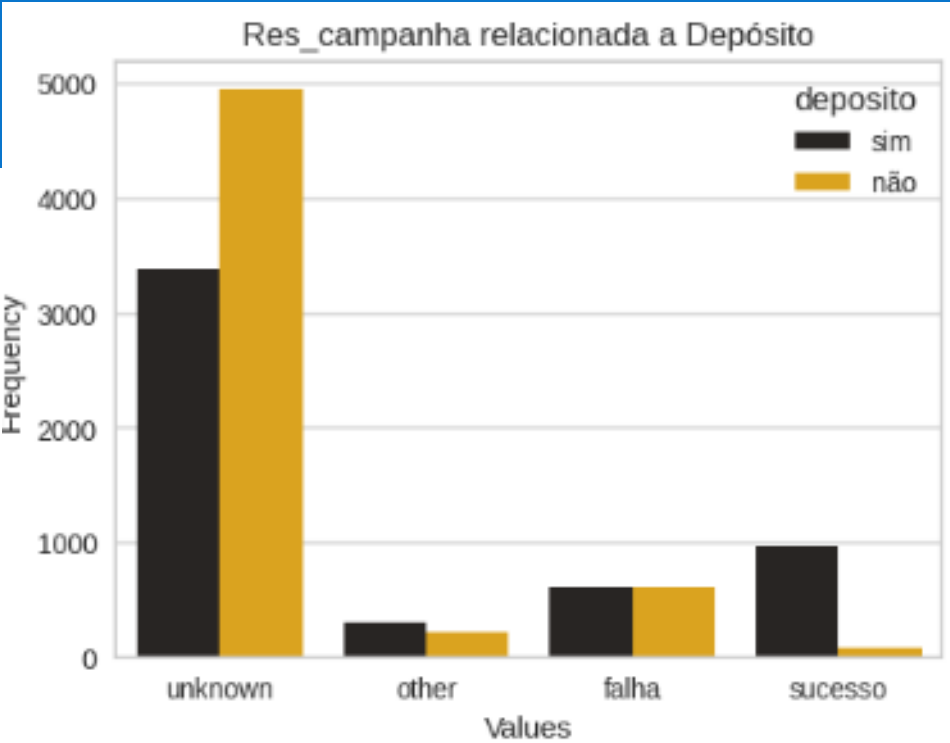


# Análise Descritiva: variáveis categóricas e binárias



## Análise:

Baixa taxa de sucesso para clientes que já possuem um empréstimo.  
Contato principal dos funcionários com o cliente é feito pelo celular.  
Boa parte dos resultados de campanha não são documentados  
Dados entre depositários e não depositários estão balanceados, o que proporciona uma boa análise das melhores variáveis a serem analisadas.



deposito		count
0	não	5873
1	sim	5289

# Análise Temporal:

As proporções de quantos clientes não depositam em relação ao total de um dia ou mês são aleatórias, com alguns meses com maior captação de depósito que outros.

**Meses e sua proporção (dividida por 100) de clientes que NÃO efetivaram um depósito:**

mes	deposito	proporção
dez	não	0.090909
mar	não	0.101449
set	não	0.156740
out	não	0.176020
abr	não	0.374865
fev	não	0.431701
ago	não	0.547070
jun	não	0.553191
nov	não	0.572641
jul	não	0.585865
jan	não	0.587209
mai	não	0.672450

**Dias em que as chances de um cliente recusar uma oferta de depósito à termo são as mais baixas, de acordo com sua proporção (dividida por 100):**

dia	deposito	proporção
10	não	0.257669
1	não	0.262295
25	não	0.406250
3	não	0.418301
22	não	0.427509
4	não	0.427861
30	não	0.433054
12	não	0.451685
2	não	0.455090
13	não	0.467991

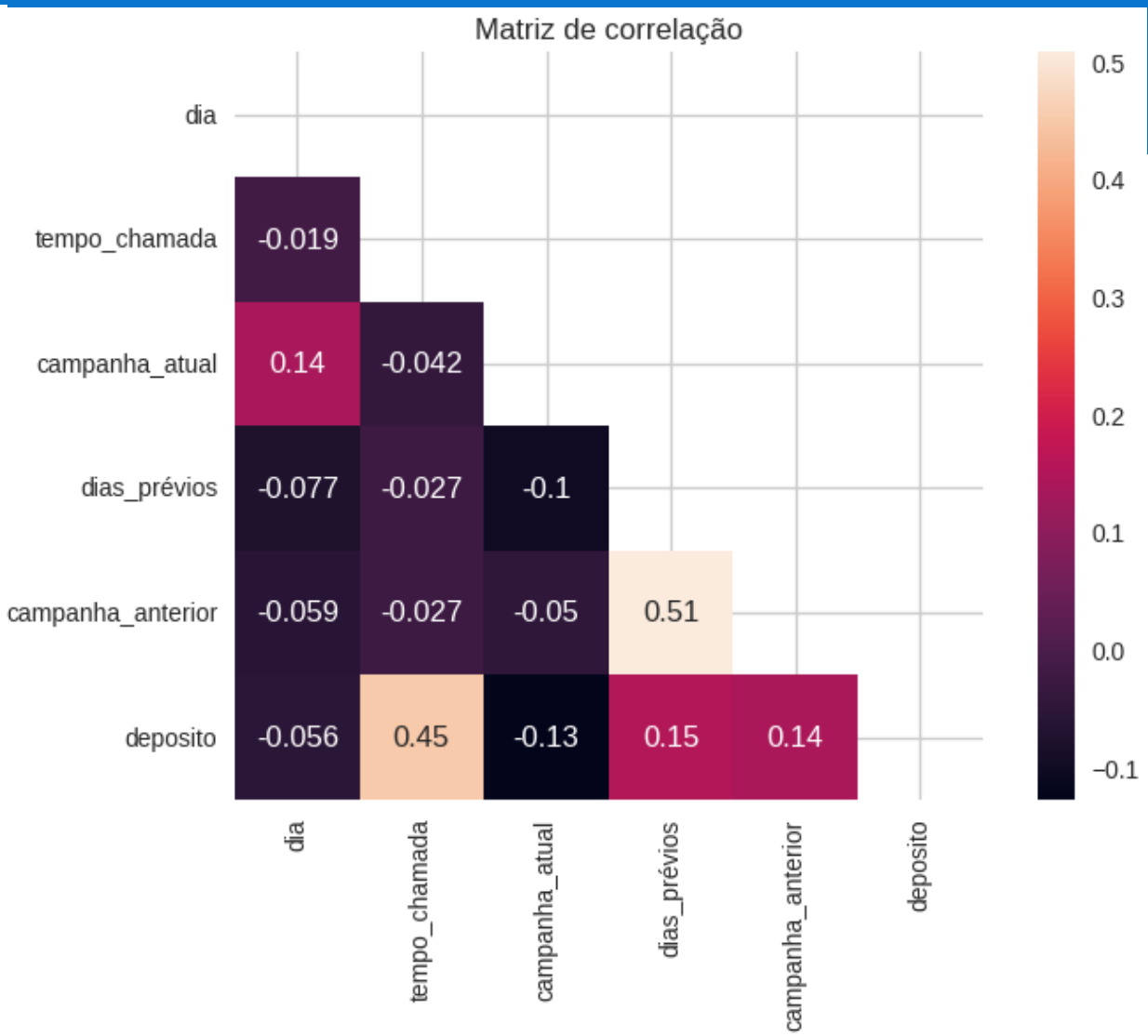
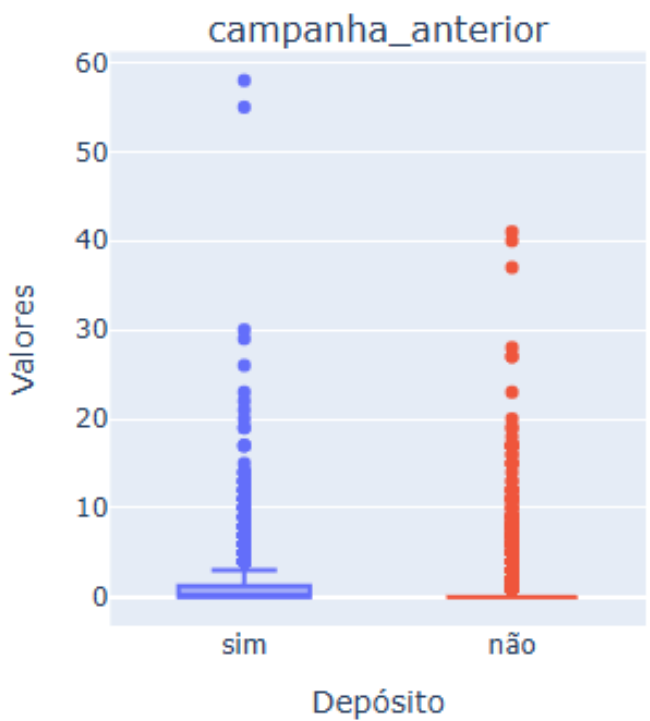
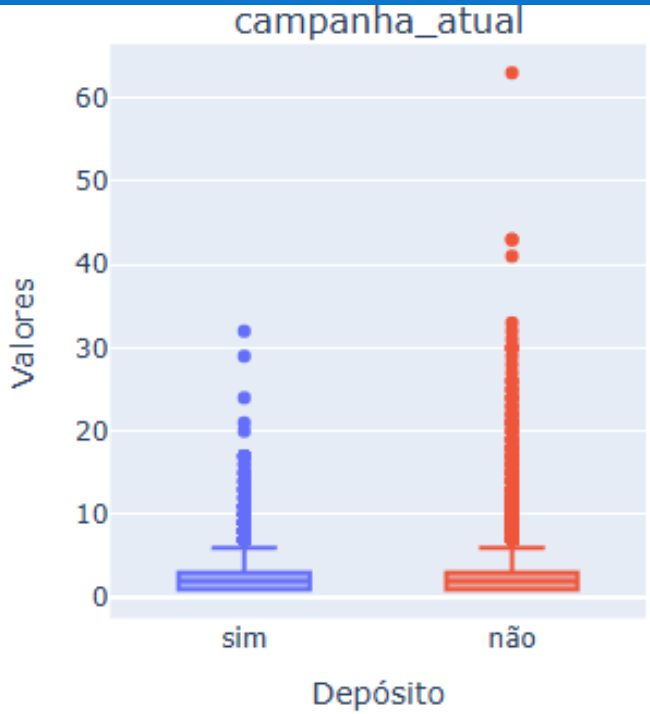
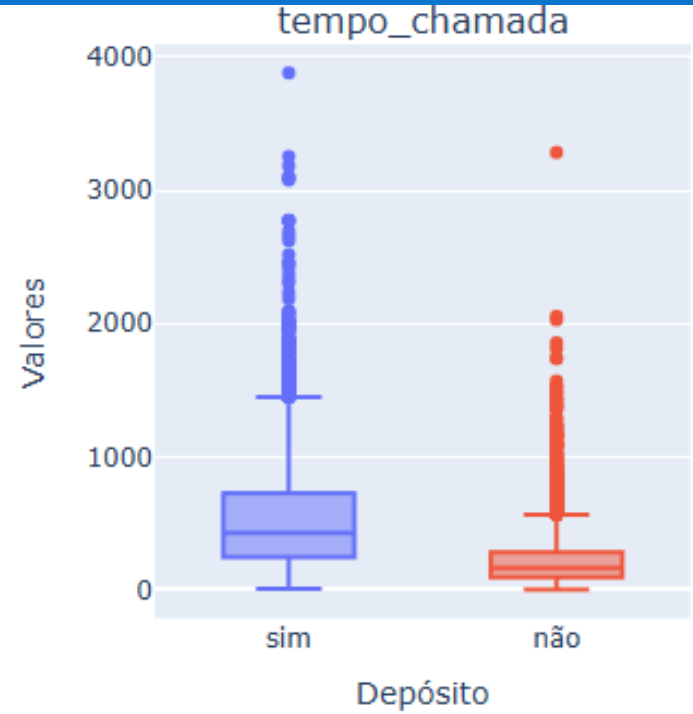


# Análise de Benchmark:

Variáveis de rotina que mensuram o tempo de contato ou não com o cliente e a quantidade de ofertas a ele ofertadas:

A maioria dos clientes decidem a contratação do depósito à termo no primeiro contato com o vendedor, representada no boxplot por -1 (nota-se os boxplots de ‘dias prévios’ e ‘campanha anterior’ como sendo somente o número 1).

No geral, essas variáveis apresentam igual equilíbrio entre as taxas de depósito, excetuando o tempo de chamada, que, naturalmente, as inclusões de depósito vão possuir mais duração (características como mais detalhes do produto e ofertas de outros serviços serão mencionados).



No geral, infere-se baixa relação com as chances de depósito, sendo a variável tempo\_chamada a com maior taxa de correlação.

# Modelo:

Escolhi o CatBoost para análise da importância das variáveis para o cliente efetivar o depósito pelas seguintes razões:

1. Muitas variáveis categóricas nominais e ordinais (não necessito fazer o encoding manual).
2. Outliers presentes na variável 'saldo' (algoritmos de boosting são robustos contra outliers).
3. Facilidade de conseguir os coeficientes.

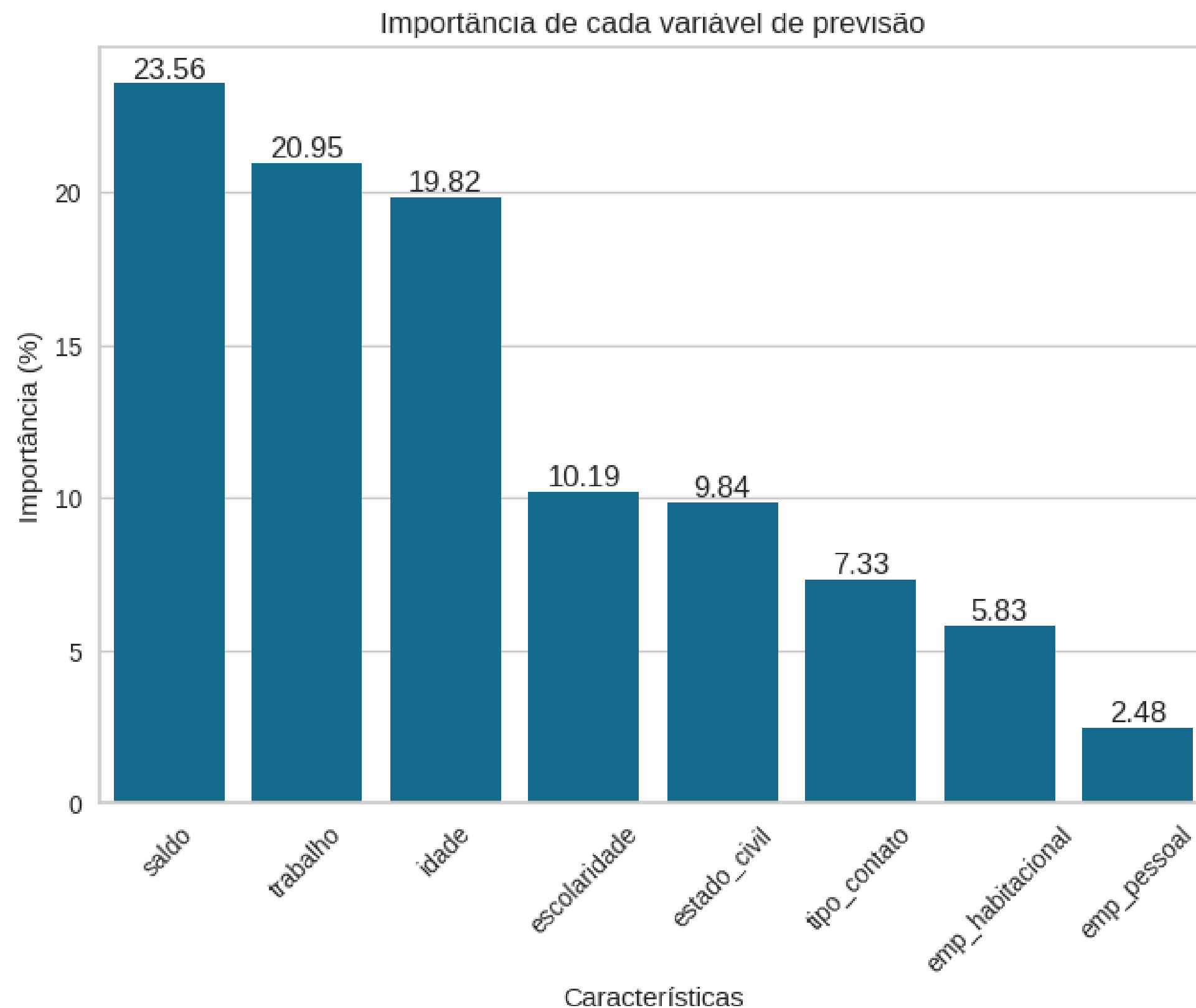
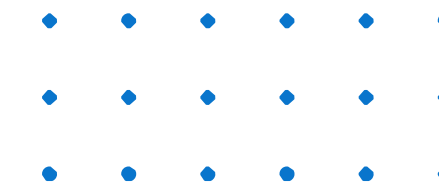
```
1 X_train, X_test, y_train, y_test = train_test_split(X, y,
2                                                    test_size=0.25,
3                                                    random_state=42)
4
5 model = CatBoostClassifier(iterations=600,
6                             learning_rate=1.0,
7                             depth=8,
8                             eval_metric='Accuracy',
9                             verbose=200)
10
11 model.fit(X_train, y_train, cat_features=cat_vars)
```

## Escolha das variáveis:

- As features escolhidas para compor o modelo são, para as variáveis independentes: idade, trabalho, estado\_civil, escolaridade, saldo, emp\_habitacional, emp\_pessoal e tipo\_contato. A variável independente é o depósito.

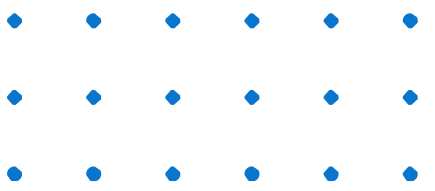
Outras variáveis foram removidas por serem desbalanceadas (como inadimplente), apresentarem baixa correlação com a variável 'depósito' ou por conterem vieses, podendo causar overfitting no modelo.

# Features mais importantes:



**Variáveis mais importantes:**

- SALDO
- TRABALHO
- IDADE



# Finalizando

(de acordo com os objetivos mencionados):

- Os perfis de clientes mais amplos que o banco possui são de jovens ou idosos, com um bom capital guardado e sem empréstimo habitacional.
- Efetiva-se ou não boa parte dos contratos através do primeiro contato com o cliente.
- Descobrindo a idade, trabalho e saldo financeiro do cliente, conseguimos obter uma boa estimativa acerca da possibilidade do cliente investir através do banco.

## OBRIGADO!

Matheus Felipe, 27/08/2025