



A3 Data

TESTE TÉCNICO - CIÊNCIA DE DADOS
JOSÉ DE FARIA LEITE NETO

INTRODUÇÃO

O objetivo dessa apresentação é mostrar análises e insights sobre um problema de churns, apontando causas e possíveis caminhos para melhorarmos essa métrica.

ROADMAP

- Avaliação do dataset
- Tratamento dos Dados
- Análise Exploratória de Dados
- Levantamento de hipóteses
- Modelagem
- Simulação
- Melhorias Futuras

Avaliação do Dataset

A primeira etapa visa entender o problema e as variáveis, para isso, é realizada uma análise inicial entendendo o significado de cada variável e como ela se relaciona com as perguntas que procuramos responder.

Após essa avaliação inicial, concluimos que:

- A nossa label se chama "Churn", e indica se houve churn ou não
- Nosso dataset possui uma dimensionalidade de 7043 linhas x 21 colunas
- "customerID" representa o ID do cliente, definimos essa coluna como o ID do nosso dataset
- Nossa label é desbalanceada, com 5174 registros de não churn e 1869 registros de churn
- A grande maioria das variáveis são categóricas
- Uma avaliação descritiva rápida das variáveis numéricas não encontrou a presença de outliers

Churn	
count	7043.000000
mean	0.265370
std	0.441561
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000

Tratamento dos Dados

Nessa etapa, fazemos os tratamentos necessários para prosseguirmos com a análise, como tratamento de outliers, correção de tipos de variáveis, etc.. essa etapa é parte feita em conjunta com a análise exploratória.

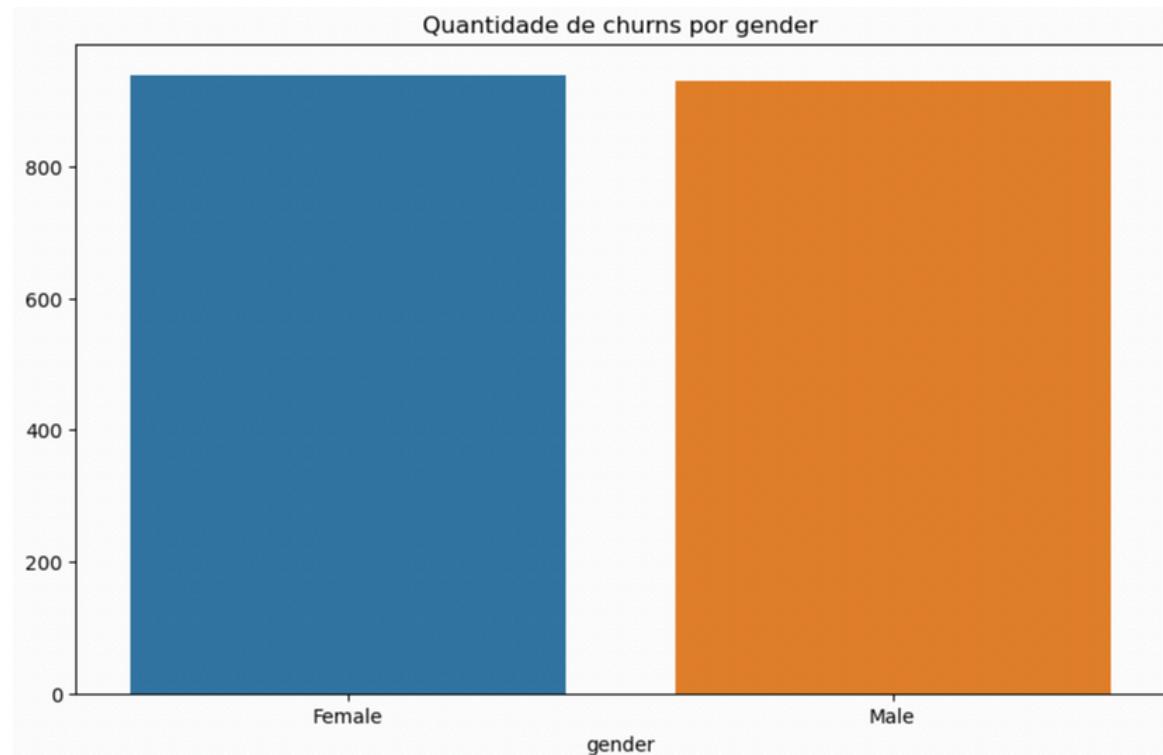
No nosso caso, os tratamentos feitos foram:

- "tenure", que inicialmente era um objeto, foi redefinida como inteiro
- "SeniorCitizen" que inicialmente era um inteiro, foi redefinida como objeto
- "MonthlyCharges" e "TotalCharges" tiveram a vírgula substituída com ponto, e foram redefinidas como float
- "Churn" foi redefinida como inteira, substituindo Yes e No por 1 e 0 respectivamente, essa mudança nos ajudará a visualizar as diferenças entre as features e a label durante a análise exploratória.
- "customerID" foi definido como o ID do dataset

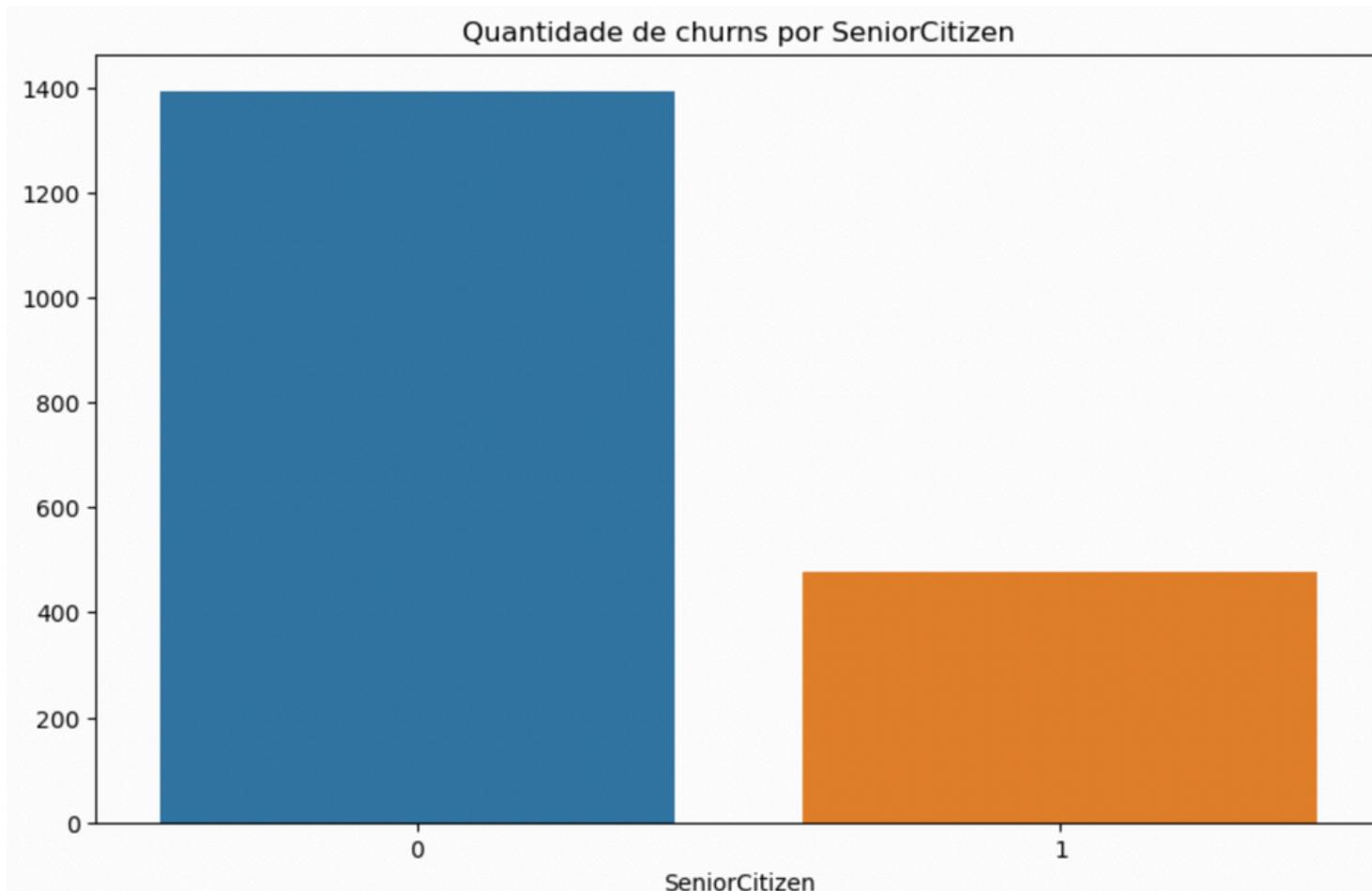
```
df['tenure'] = df['tenure'].astype(int)
df['SeniorCitizen'] = df['SeniorCitizen'].astype(object)
df['MonthlyCharges'] = df['MonthlyCharges'].str.replace(',', '.').astype(float)
df['TotalCharges'] = df['TotalCharges'].str.replace(',', '.').astype(float)
df['Churn'] = df['Churn'].replace({'Yes': 1, 'No': 0})
```

Análise Exploratória de Dados

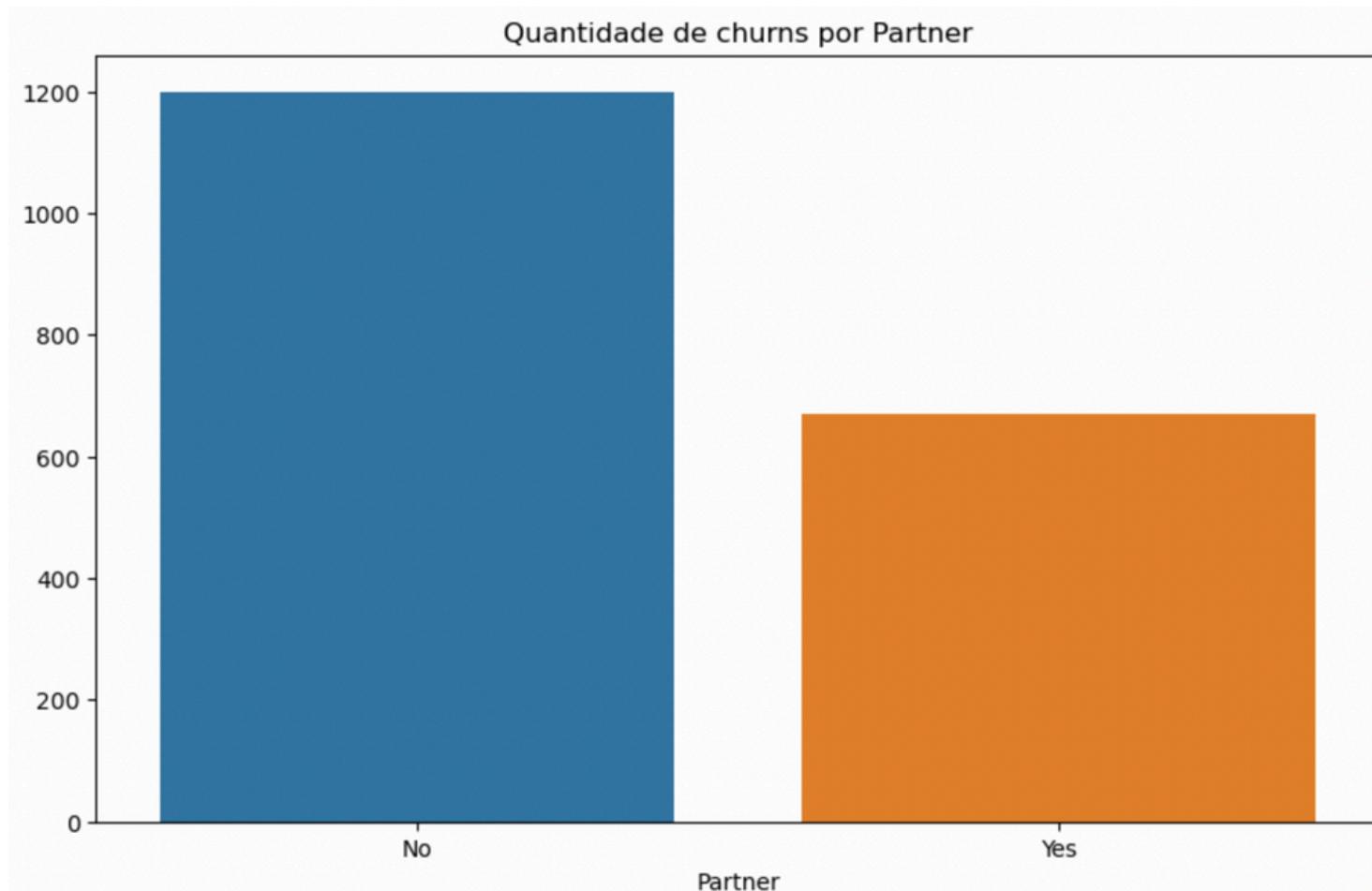
A análise exploratória visa encontrar os primeiros insights, possíveis erros que passaram pelo tratamento inicial, e nos dar "pistas" para o caminho que será seguido posteriormente. A ideia dessa análise foi encontrar as relações entre cada variável (individualmente) e o churn, analisando o impacto de cada uma nesse evento. Alguns destaques dessa análise são mostrados nos slides posteriores.



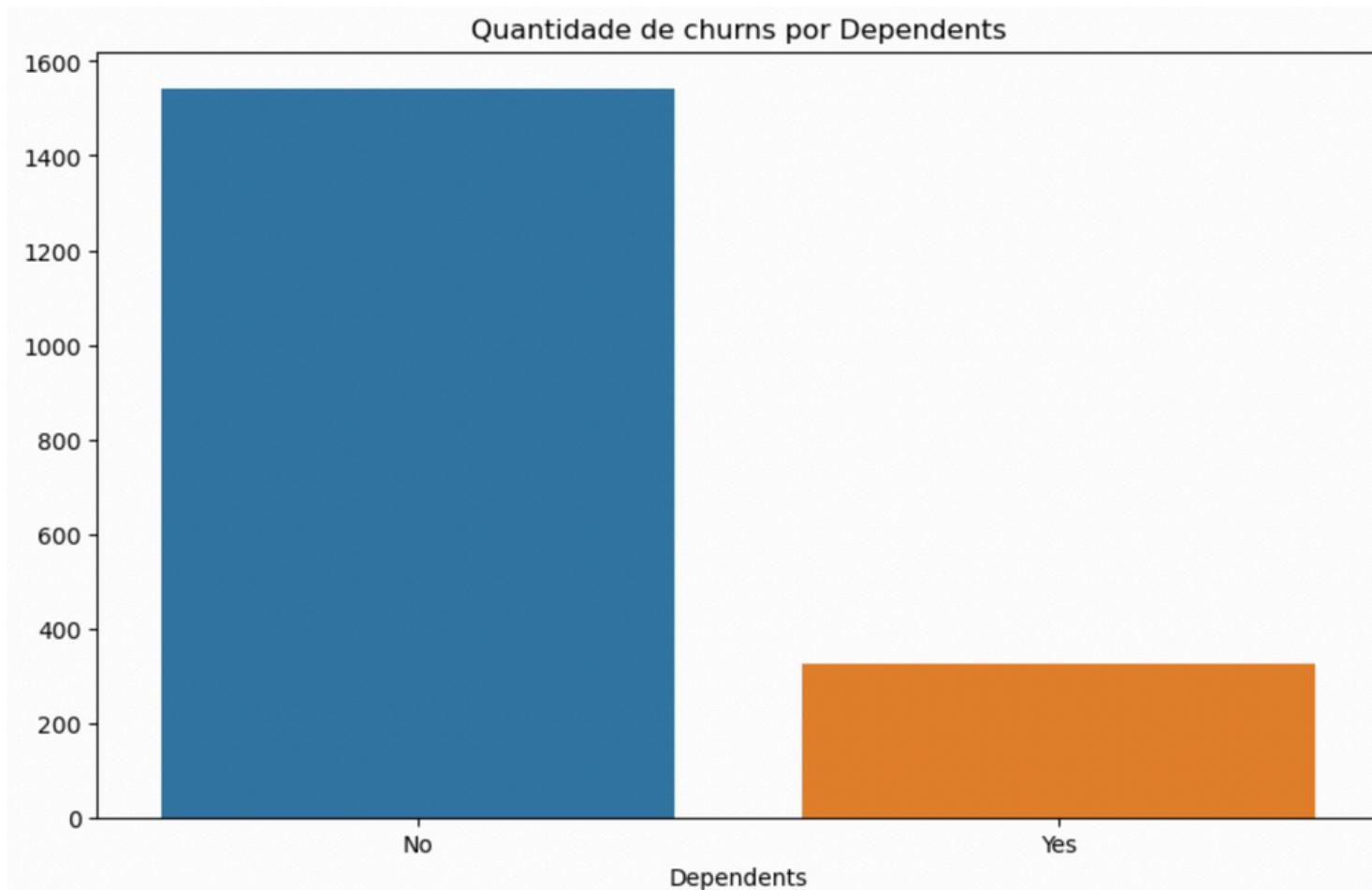
Idosos



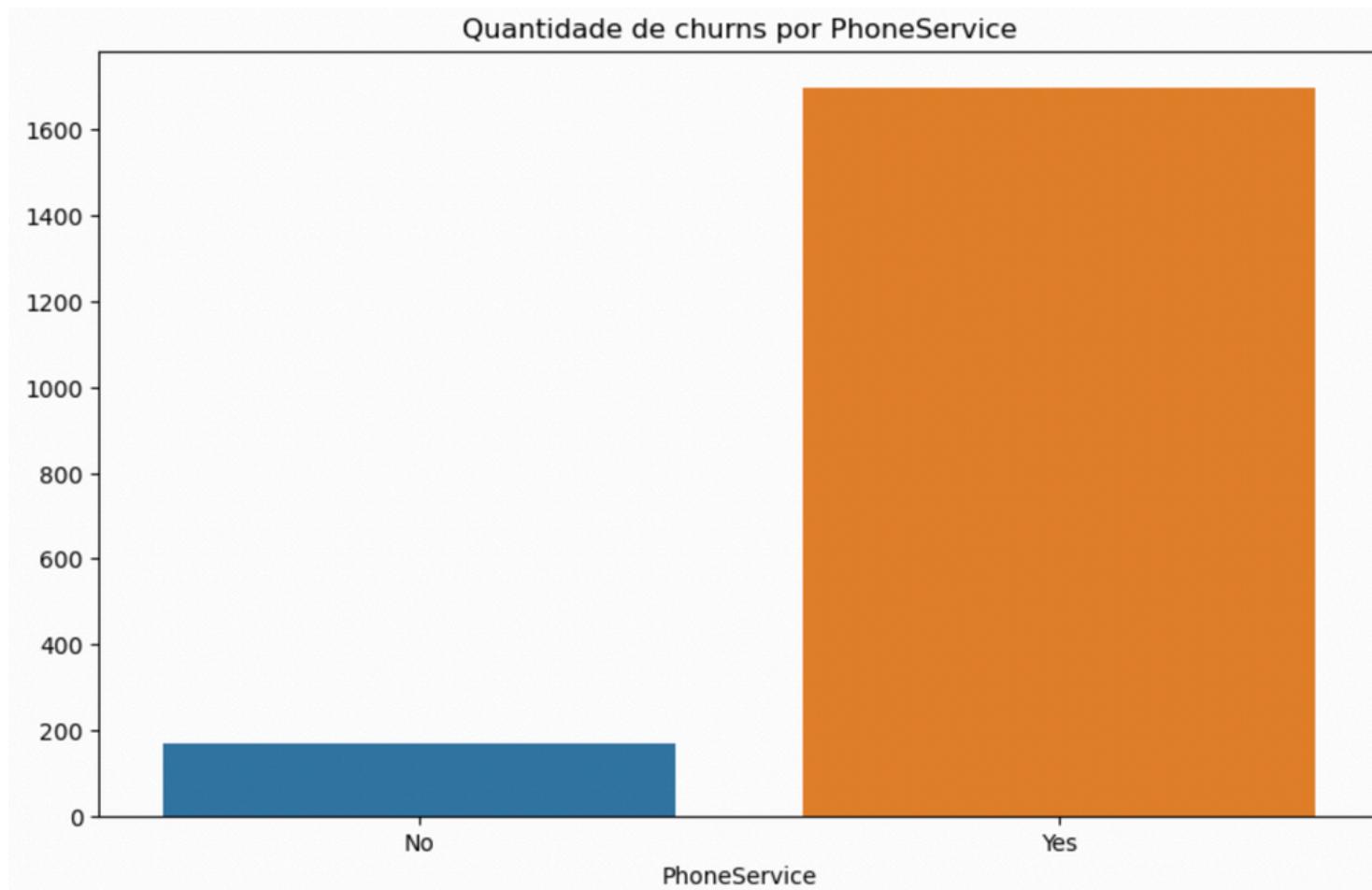
Cônjugue



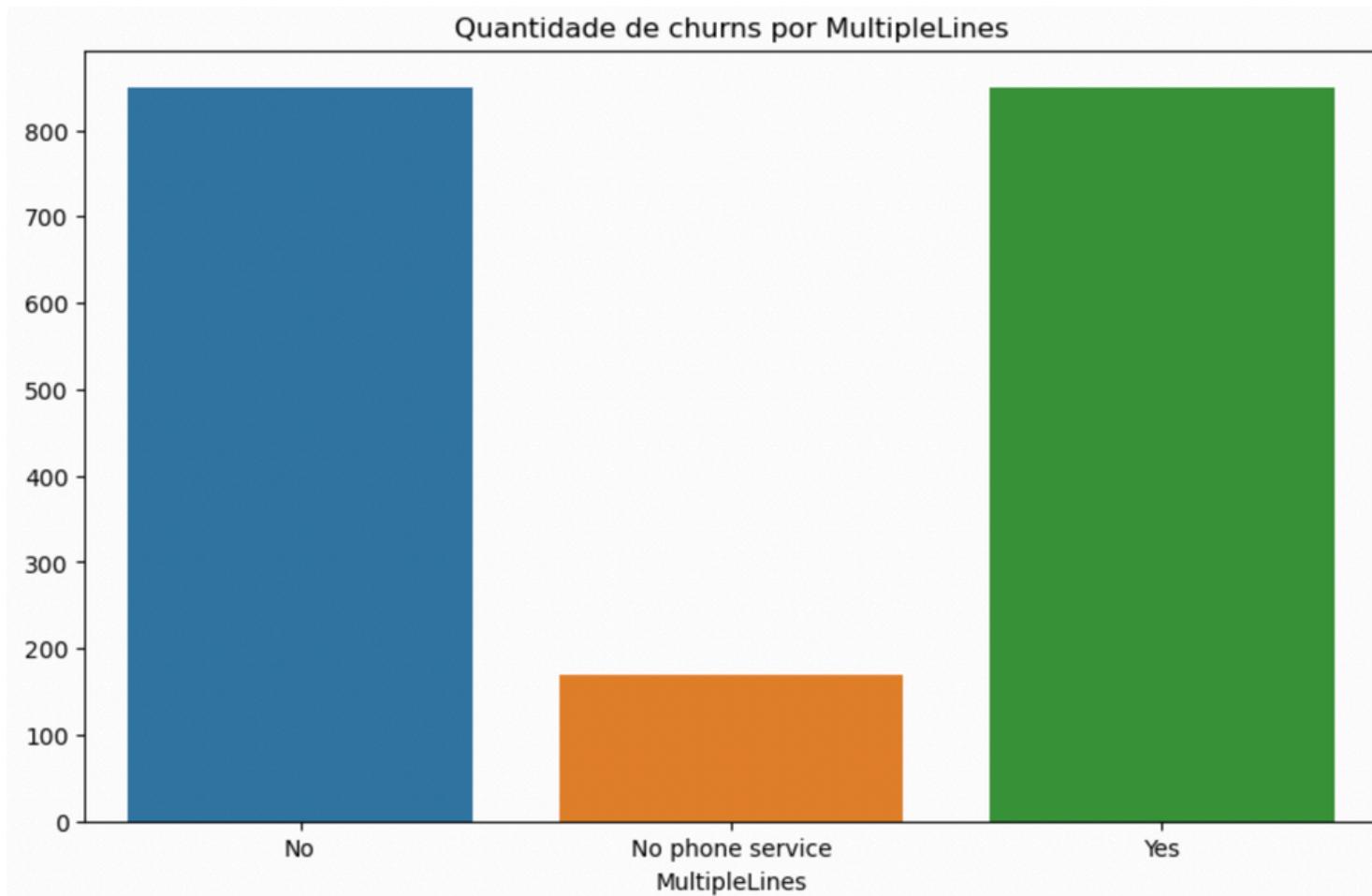
Dependentes



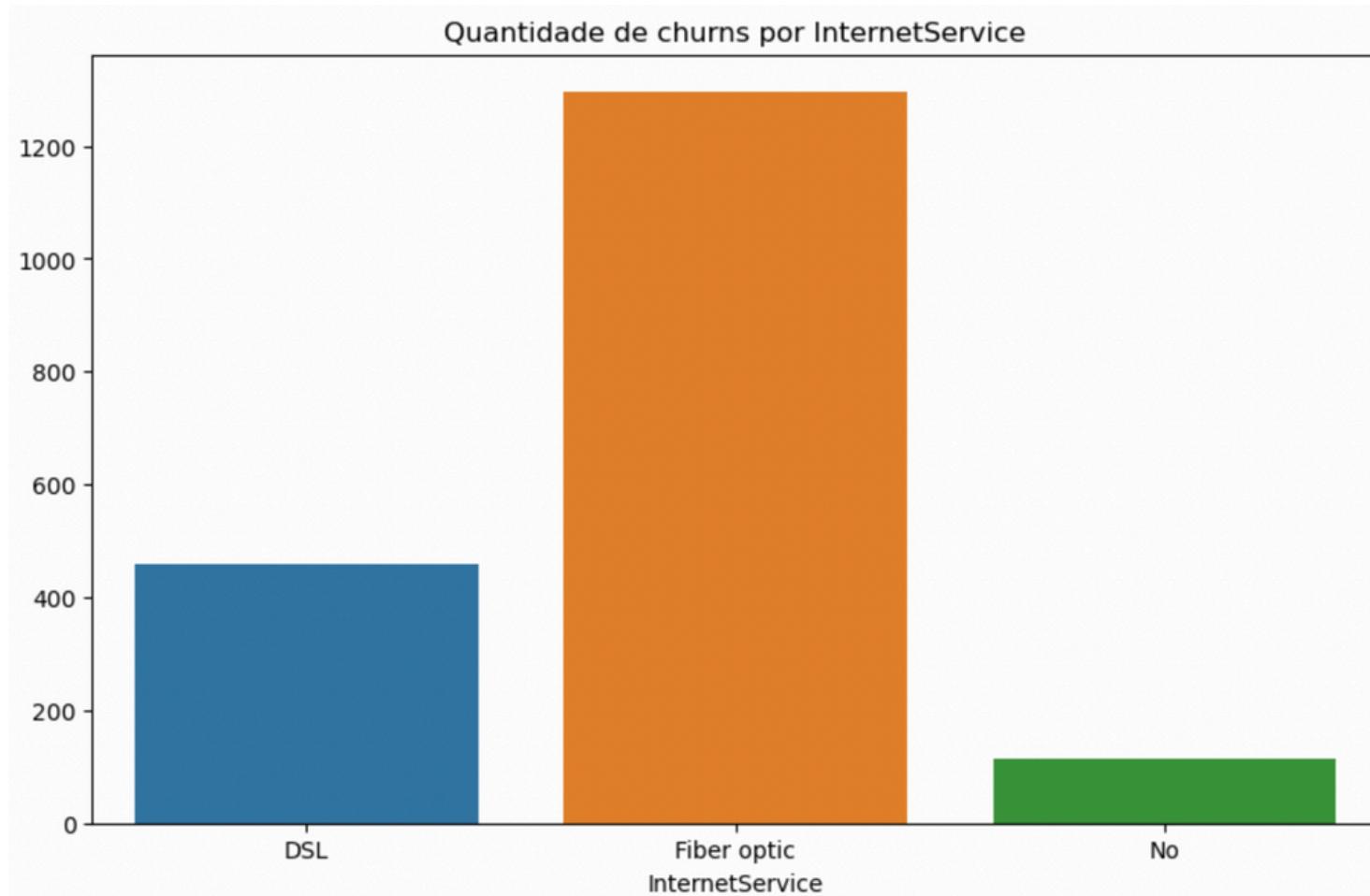
Serviço de Telefone



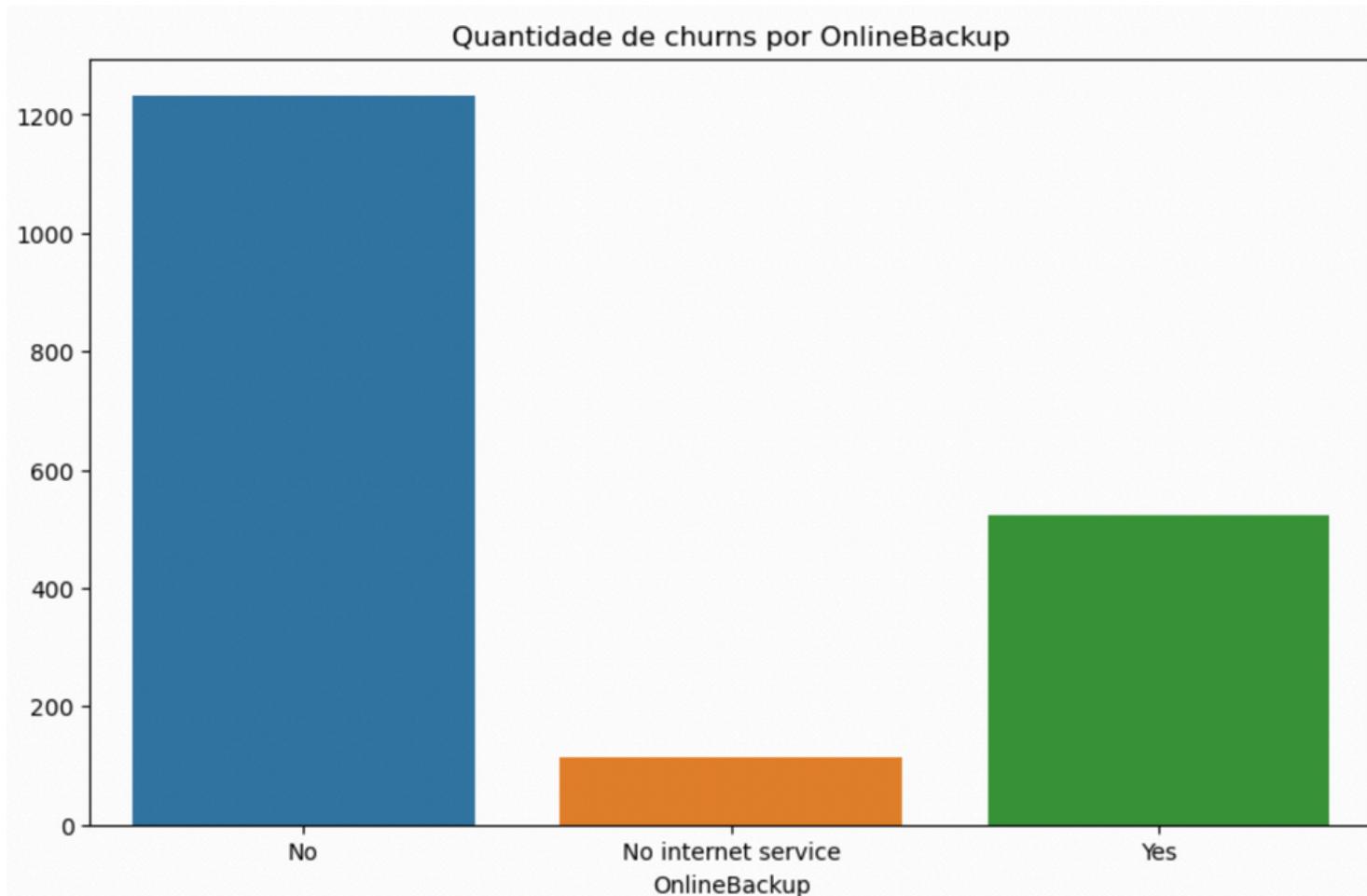
Múltiplas Linhas



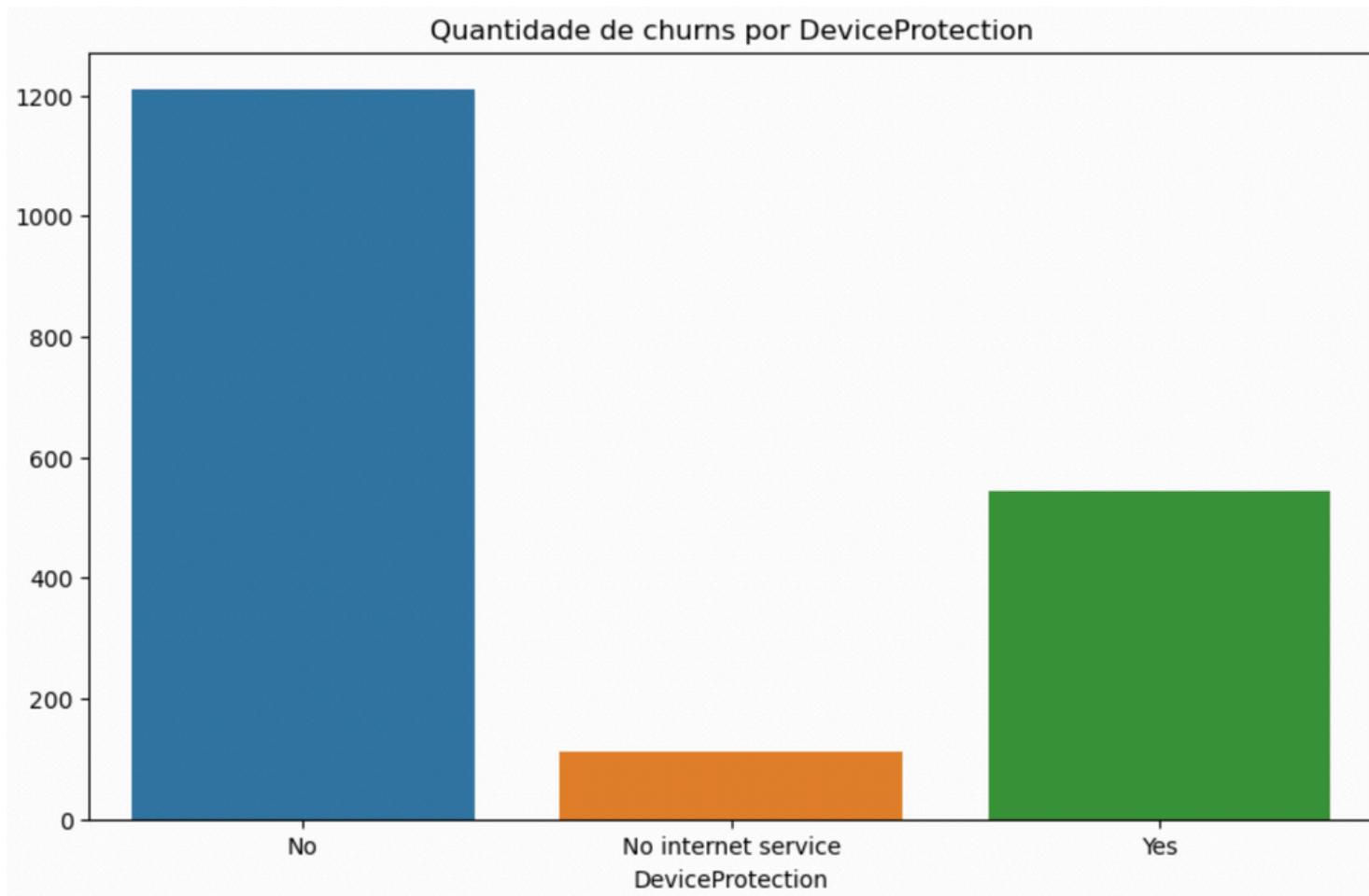
Serviço de Internet



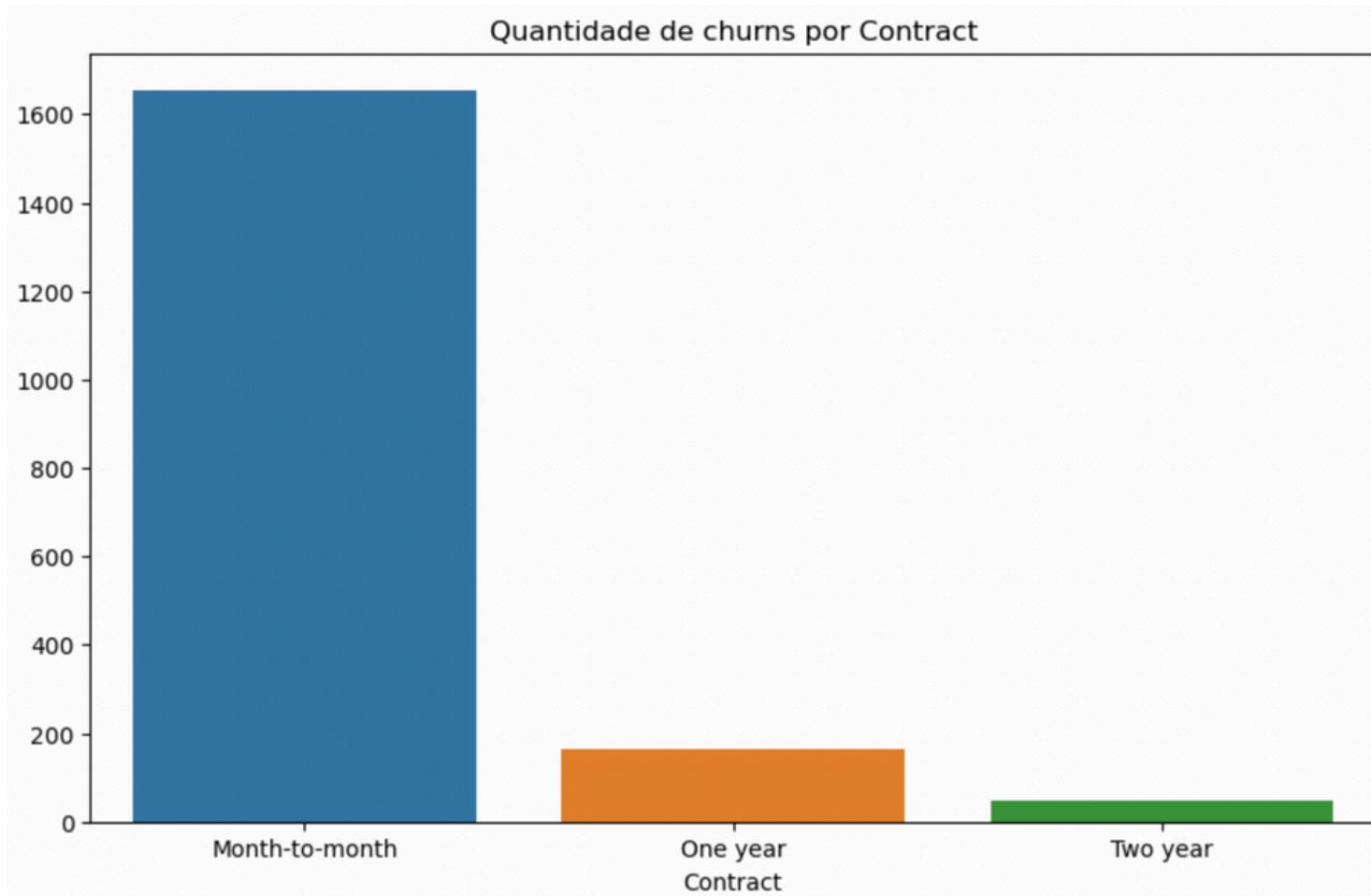
Backup Online



Proteção de Dispositivo

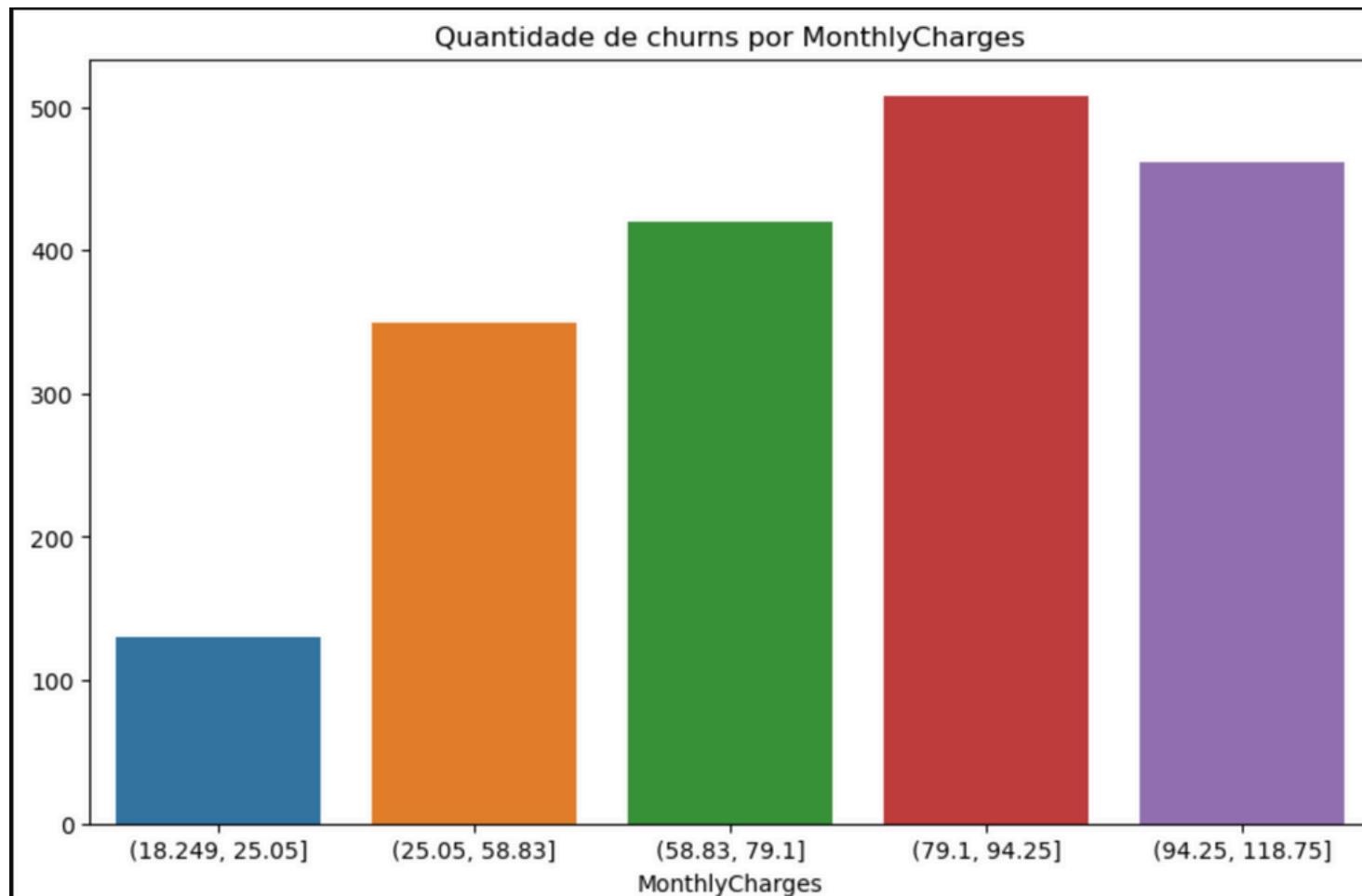


Contrato

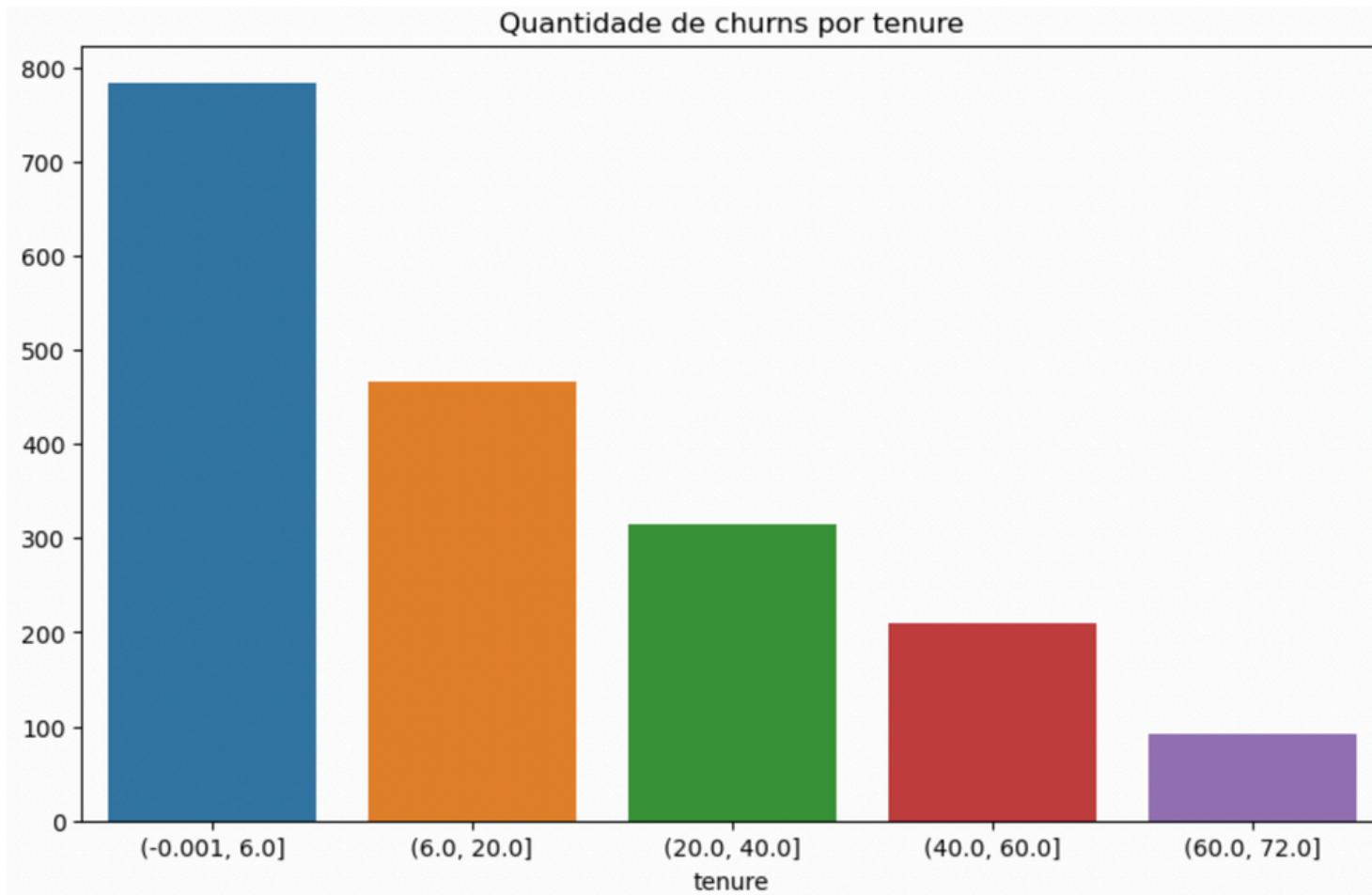


Variáveis Contínuas

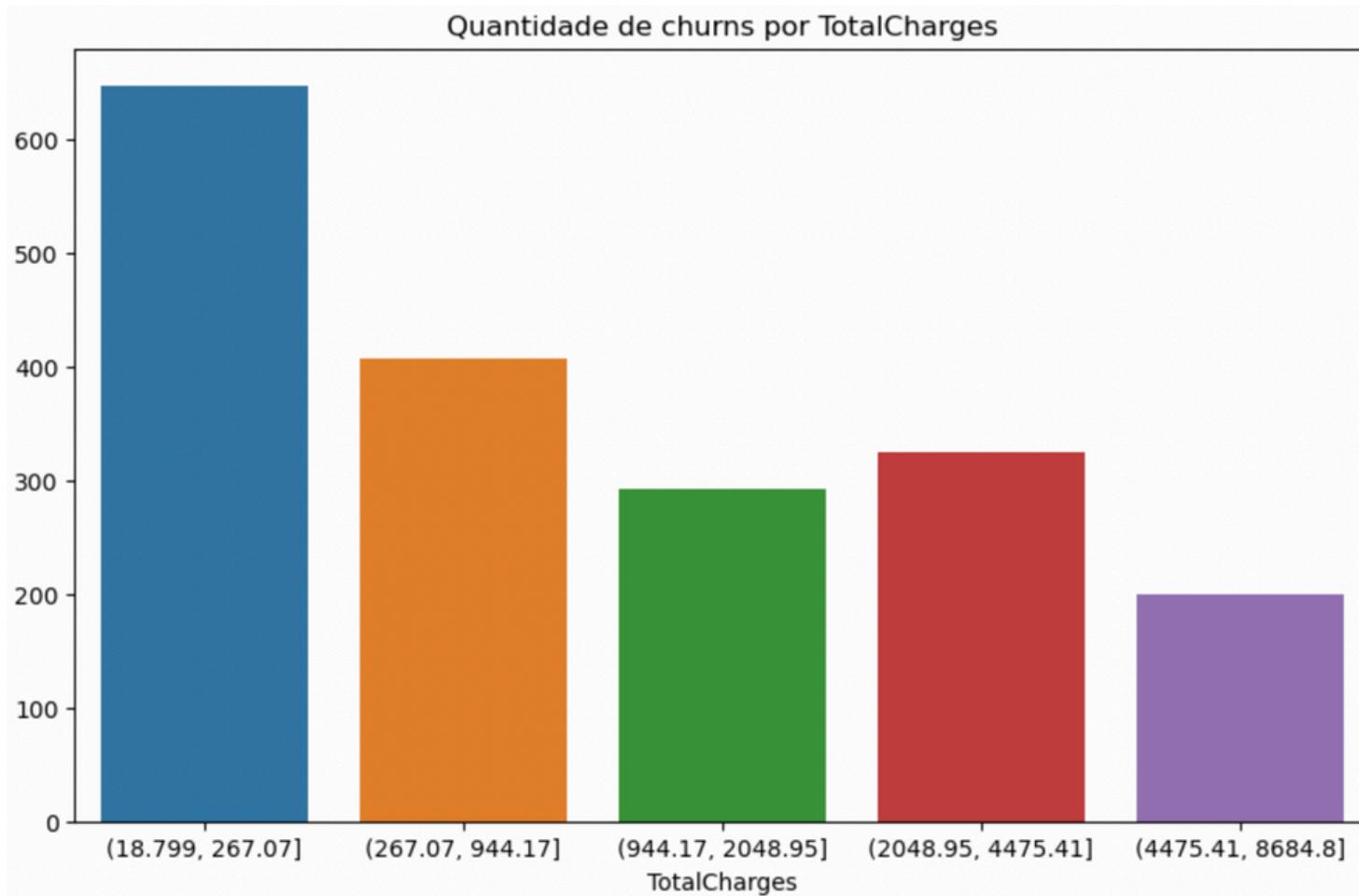
Para a avaliação das variáveis contínuas em relação ao churn, foi realizada uma discretização dessa variável baseada em quantis, foram usados 5 quantis para cada variável, na imagem abaixo, observamos a relação entre os quantis de churn e a variável MonthlyCharges



Tenure



Total Charges



Hipóteses Levantadas e Modelagem

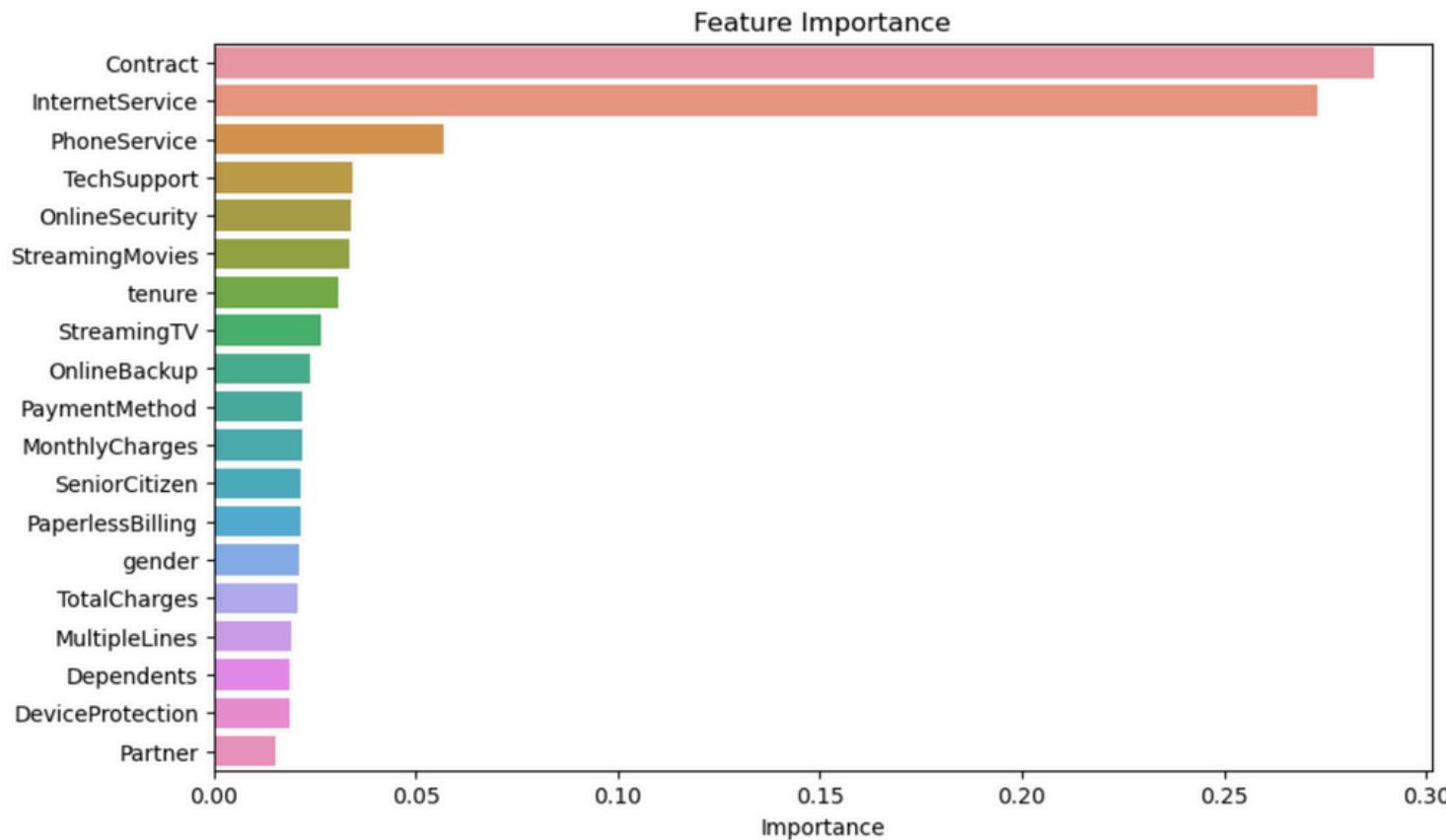
Observamos diversas variáveis onde há um certo padrão de comportamento em relação ao churn, porém, essas variáveis foram analisadas individualmente. Assim, usamos um modelo de machine learning para avaliar as variáveis todas juntas em relação ao churn, respondendo o quão cada variável impacta no churn.

O modelo escolhido foi o xgboost, por ser um modelo consolidado, rápido e que lida bem com variáveis categóricas. A métrica utilizada para avaliar o modelo foi o F1-score devido ao desbalanceamento do dataset e a estratégia utilizada para esse desbalanceamento foi um oversampling simples.

No fim, o modelo apresentou um F1-score de 0.9 para os dados de treino e de 0.85 para os dados de teste, mostrando uma boa estabilidade e métrica para nosso problema e propósito.

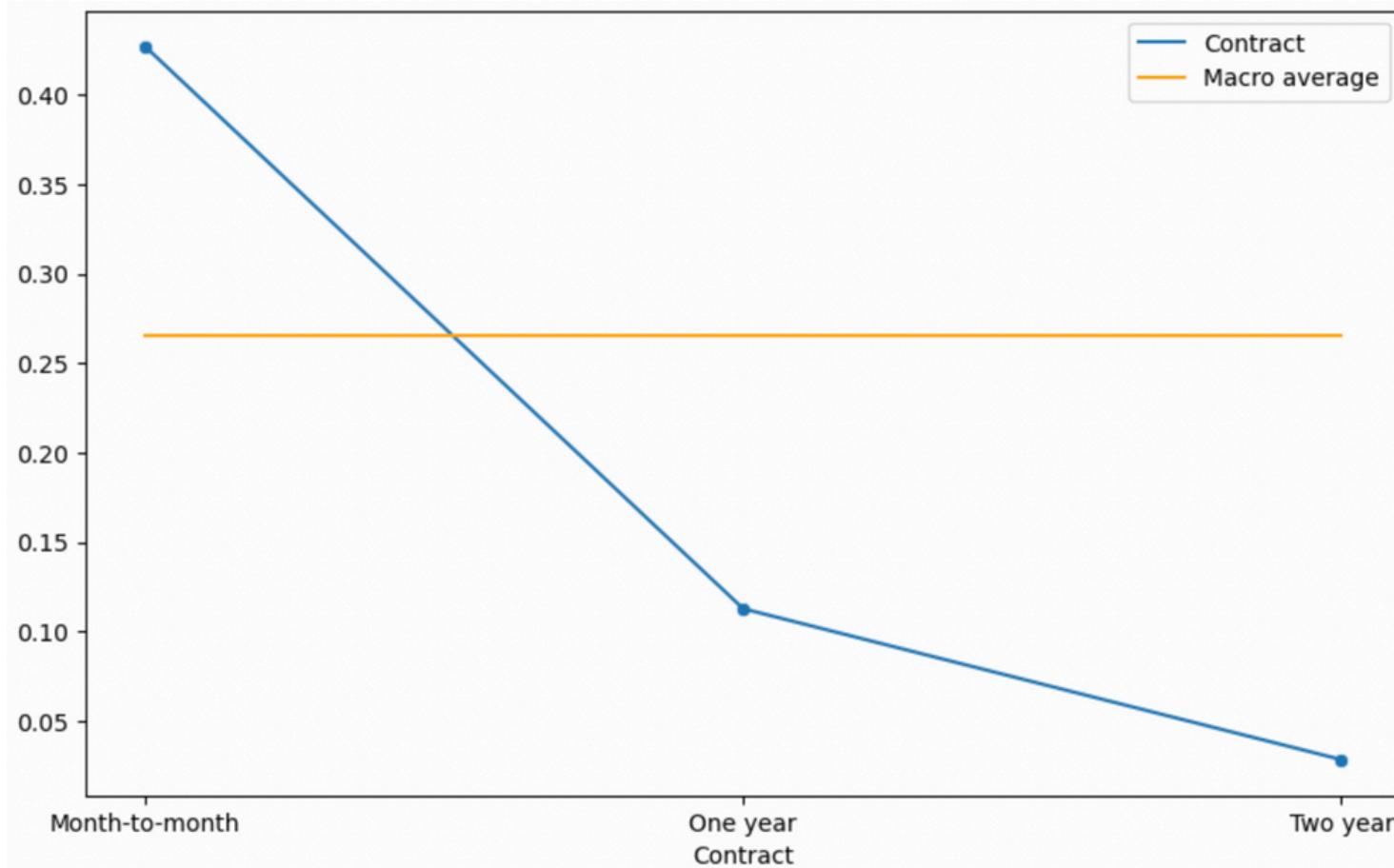
Importância de Features

Com o nosso modelo treinado, é possível avaliarmos a importância de cada feature, agora em contexto com as outras variáveis, o resultado foi o seguinte:



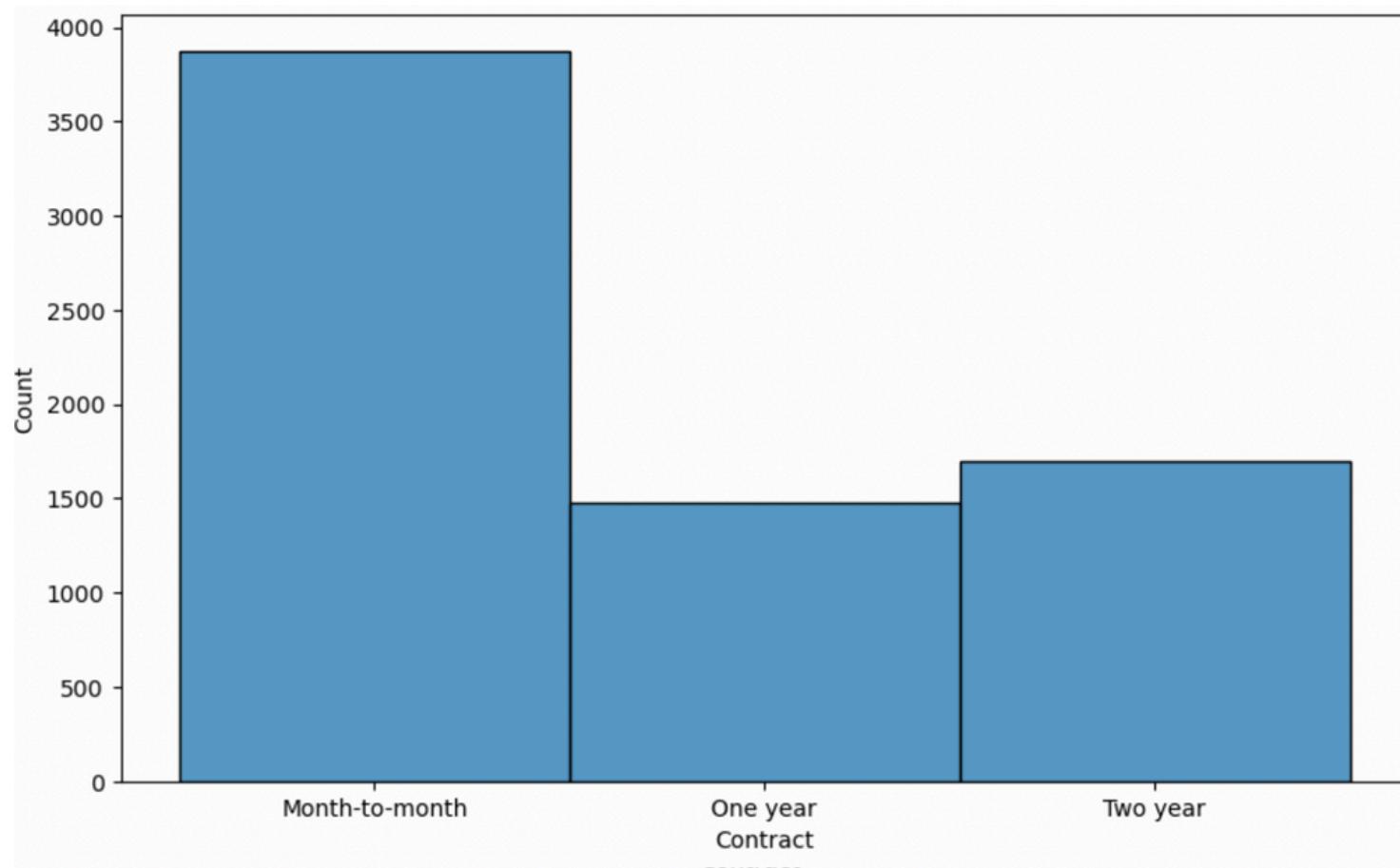
Contract

Percebemos que o tipo do contrato é a feature mais impactante para o churn, o gráfico abaixo mostra a média de churn de cada classe dessa feature em relação à média geral de chuns (excluindo essa feature).



Contract

Percebe-se que a classe Month-to-month possui uma média de churn muito maior que a média do dataset geral e muito maior que suas duas outras classes, e é o tipo de contrato que mais aparece no dataset em relação aos outros:



Contract

Agora vamos simular que a empresa consiga igualar as vendas de tipos de contrato.

Metodologia: Para fins de simulação, igualamos a quantidade de cada classe de contratos através de um oversampling simples e retiramos a média de churn por cliente desse novo dataset.

Resultado:

```
Média de churn por pessoa (dados originais) 0.2654
Média de churn por pessoa (números de contratos iguais) 0.1892
Uma redução de -28.72000000000002%
```

Conclusões

- Existem diversos caminhos diferentes possíveis para atacar o problema de churns, vários deles são mostrados nessa análise.
- Existem duas variáveis que destoam das outras em relação a importância, é interessante focar nelas para um resultado mais rápido

Melhorias Futuras

- Feedback do time de negócios
- Fazer simulações mais assertivas de acordo com o feedback
- Análise de correlação e possibilidade de "pacotes" de ações
- Melhoria do modelo preditivo (tunning, seleção e criação de variável, etc..)
- Análise das outras variáveis
- Testes de hipótese para checar as amostras de variáveis e churn.