UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO

FACULDADE DE CIÊNCIAS DE BAURU

CURSO DE BACHARELADO SISTEMAS DE INFORMAÇÃO

Darryê Roberto da Silva Mellin

Vitor Siwerski Aronque

ANÁLISE DE CHURN (CANCELAMENTO DE CLIENTES) - GRUPO 03
REGRESSÃO LOGÍSTICA

Bauru - SP
2025

INTRODUÇÃO

O trabalho tem como intuito realizar uma análise de Churn, que se trata de uma análise de evasão de clientes. O objetivo de uma análise de Churn é identificar, entender e prever quais clientes estão mais propensos a cancelar um serviço ou deixar de usar um produto.

Para realizar a análise serão utilizadas técnicas de estatística descritiva para identificar quais são os principais fatores relacionados ao cancelamento de serviços. Também será treinado um modelo de regressão logística para classificar o nível de risco de cancelamento de usuários baseando-se em suas respectivas características.

DESCRIÇÃO DO CONJUNTO DE DADOS

O conjunto de dados utilizado será o Telco Customer Churn, um dataset da IBM disponibilizado gratuitamente na plataforma Kaggle. O dataset possui 7043 linhas e 21 colunas, das quais, 4 são numéricas e o restante são categóricas.

As colunas são:

- customerID: ID de usuário
- gender: Genêro (Male,Female)
- SeniorCitizen: identifica se a pessoa é maior de 60 (Yes, No)
- Partner: Identifica se o usuário tem parceiro (Yes, No)
- Dependents: Identifica se tem dependentes (Yes, No)
- tenure: Identifica o tempo de contrato (numérico)
- PhoneService: Identifica se o plano de celular é contratado (Yes, No)
- MultipleLines: Identifica se o cliente tem múltiplas linhas (Yes, No, No phone service)
- InternetService: Identifica o contrato de internet (Fiber optic, DSL, No)
- OnlineSecurity: Identifica se o serviço foi contratado (Yes, No, No internet service)
- OnlineBackup: Identifica se o serviço foi contratado (Yes, No, No internet service)
- DeviceProtection: Identifica se o serviço foi contratado (Yes, No, No internet service)
- TechSupport: Identifica se o serviço foi contratado (Yes, No, No internet service)
- StreamingTV: Identifica se o serviço foi contratado (Yes, No, No internet service)
- StreamingMovies: Identifica se o serviço foi contratado (Yes, No, No internet service)
- Contract: Identifica o tipo de contrato (month-to-month, one year, two years)
- PaperlessBilling: Identifica se o serviço foi contratado (Yes, No)
- PaymentMethod: Identifica o método de pagamento (Electronic check, Mailed check, bank transfer(automatic),Credit card (automatic))

- MonthlyCharges: Valor dos pagamentos mensais (numérico)
- TotalCharges: Valor total arrecadado do usuário (numérico)
- Churn: Identifica se o usuário saiu no último mês (Yes, No)

ETAPAS DE TRATAMENTO E LIMPEZA DOS DADOS

O conjunto de dados já é devidamente limpo, mas foram realizados alguns passos para possibilitar a manipulação e análise dos dados. Primeiramente o campo MonthlyCharges foi transformado em numérico, para garantir que não havia nenhum campo vazio, linhas com MonthlyCharges vazio foram eliminadas e o campo SeniorCitizen foi convertido para String.

Para a etapa de treinamento do modelo de regressão logística, todos os campos foram convertidos em valores numéricos utilizando LabelEncoder e SeniorCitizen foi transformado novamente em numérico.

ANÁLISE EXPLORATÓRIA

Primeiramente foi analisada a proporção de Churn no dataset, resultando em 27% de churn e 73% de usuários que permaneceram em contrato. Após isso foram obtidas as medidas de dispersão de todas as variáveis numéricas resultando nos valores:

tenure								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Churn								
No	5163.0	37.650010	24.076940	1.0	15.0	38.0	61.0	72.0
Yes	1869.0	17.979133	19.531123	1.0	2.0	10.0	29.0	72.0

50% dos usuários que cancelaram o serviço têm até 10 meses de contrato, indicando que os usuários têm a tendência a cancelar o serviço antes do término do primeiro ano, em contraste com os usuários que não cancelaram, na casa dos 38 meses, ou três anos. Reforçando ainda mais essa idéia, 75% dos clientes que cancelaram tinham até 29 meses de contrato.

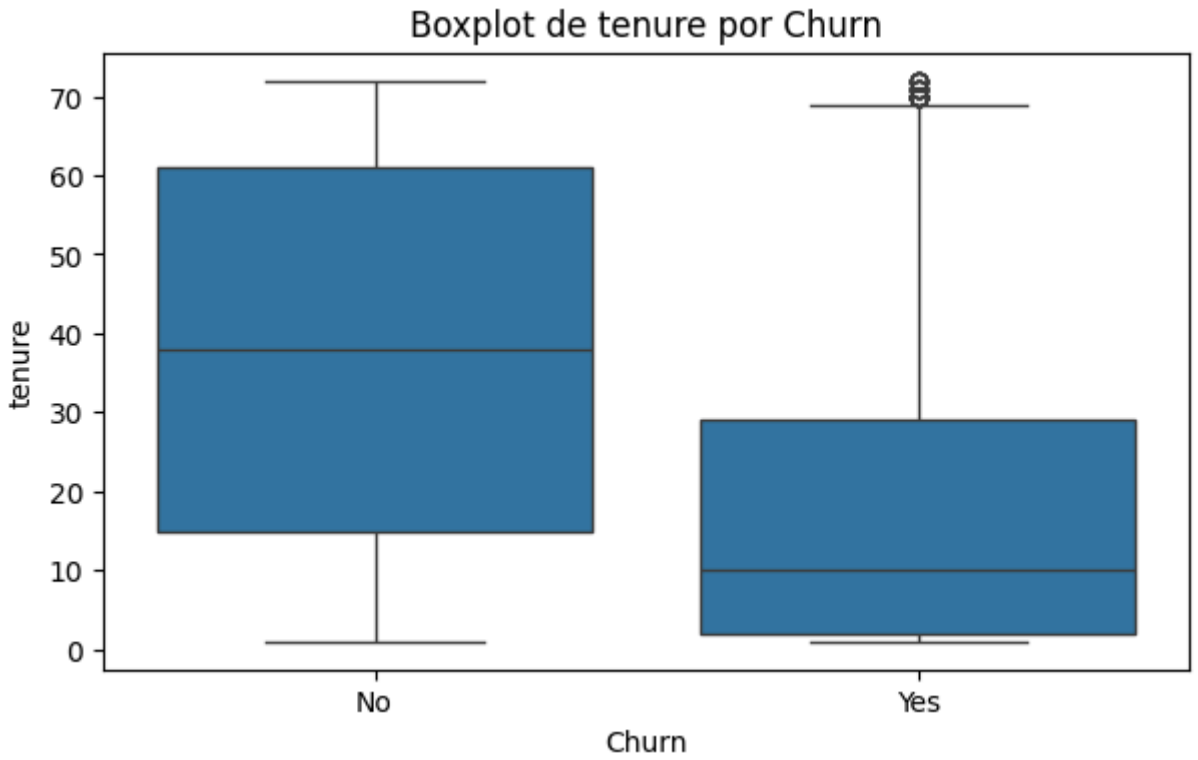
MonthlyCharges								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Churn								
No	5163.0	61.307408	31.094557	18.25	25.10	64.45	88.475	118.75
Yes	1869.0	74.441332	24.666053	18.85	56.15	79.65	94.200	118.35

O valor das cobranças mensais permanece maior em todos os quartis dos usuários que cancelaram o serviço, indicando que o custo dos planos iniciais ou de entrada falha em manter os usuários no serviço.

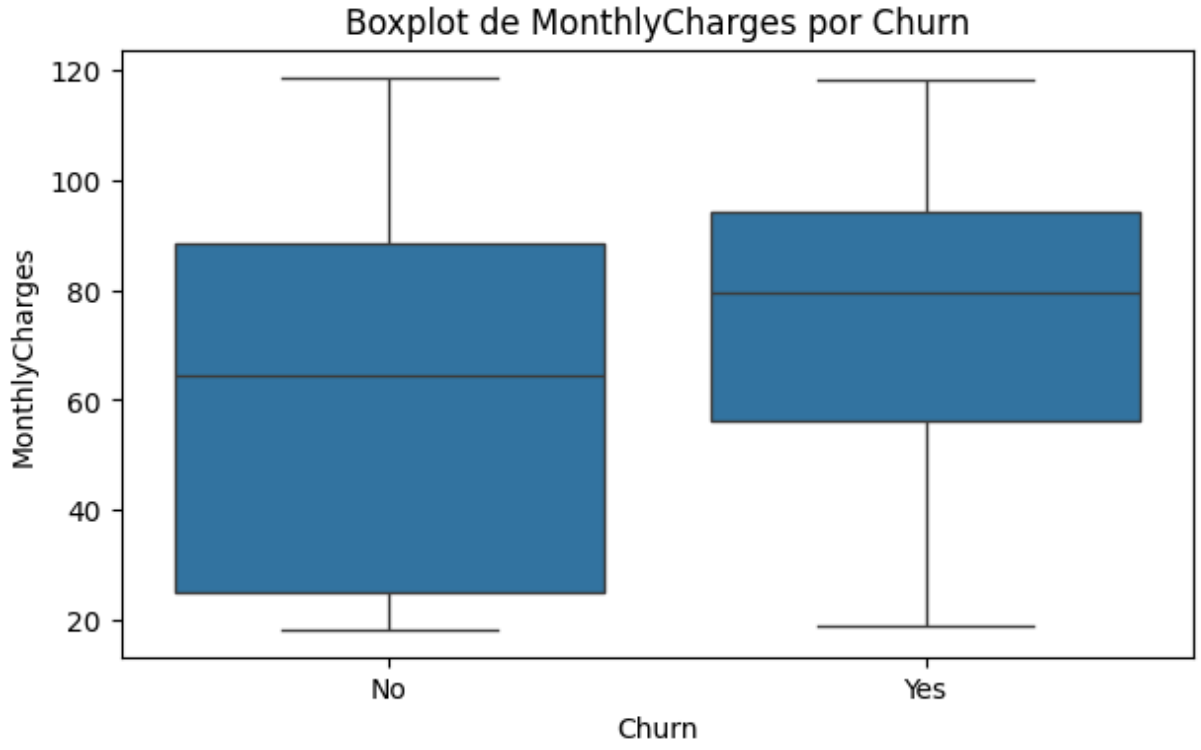
TotalCharges								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Churn								
No	5163.0	2555.344141	2329.456984	18.80	577.825	1683.60	4264.125	8672.45
Yes	1869.0	1531.796094	1890.822994	18.85	134.500	703.55	2331.300	8684.80

O valor das cobranças totais permanece muito inferior nos usuários que cancelaram o serviço, reforçando a ideia de que os usuários que cancelam o serviço são em sua maioria novos clientes.

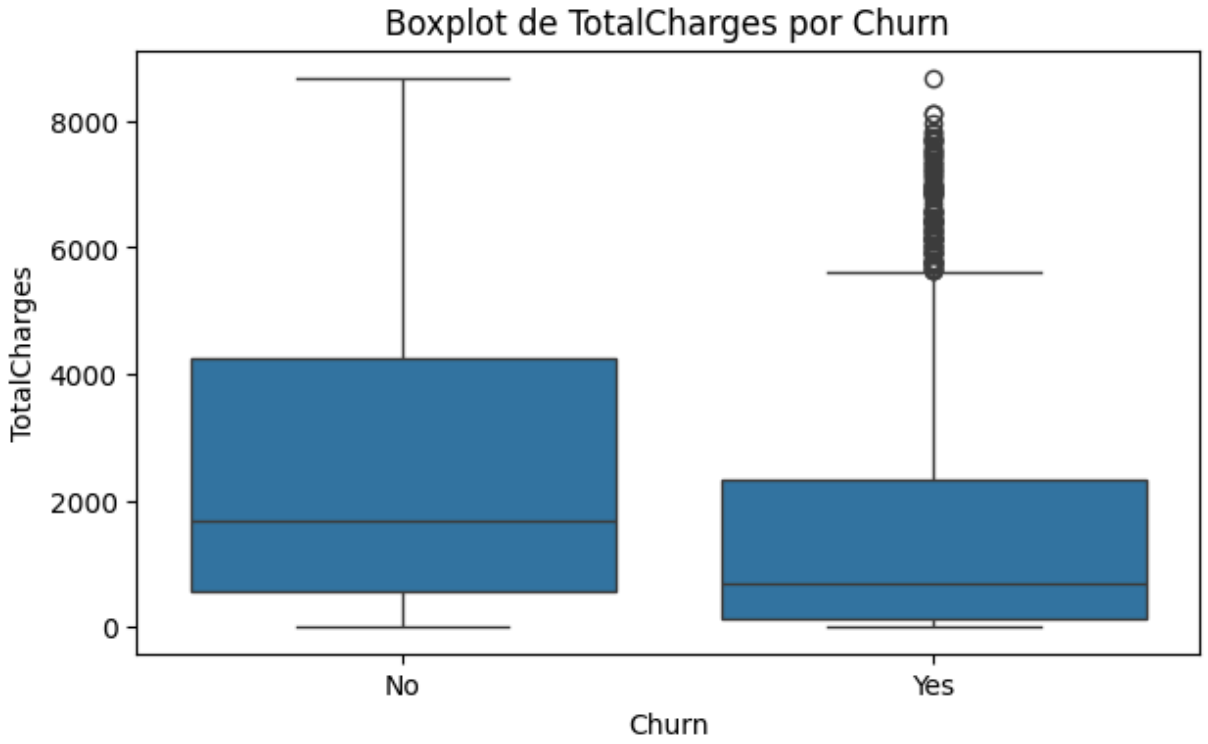
Também foram feitos gráficos de cada variável numérica em relação com churn:



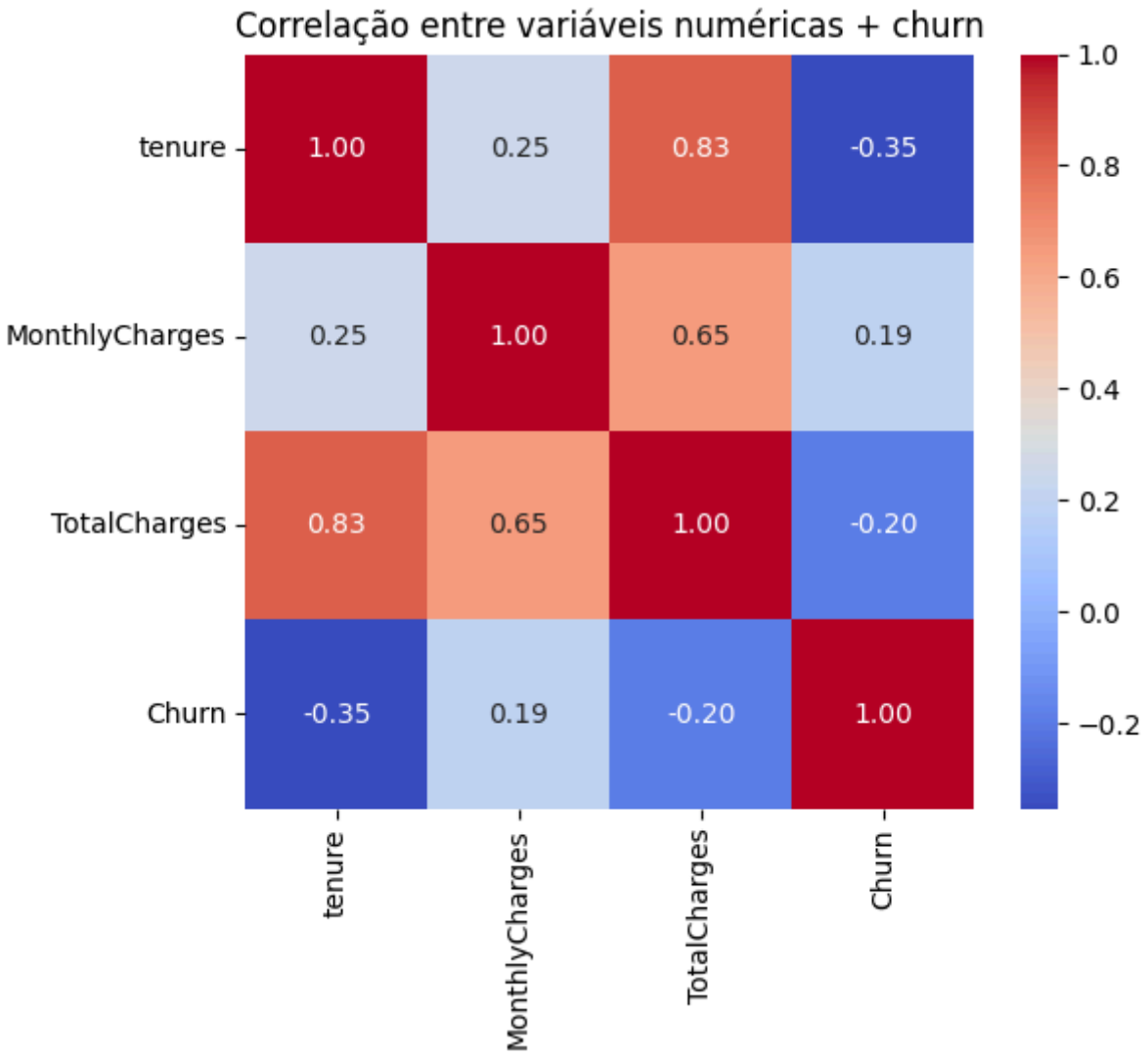
Com o boxplot de tenure fica mais fácil a visualização da dispersão do tempo de contrato dos usuários. Com isso, fica evidente que os usuários que cancelaram o serviço são em sua maioria usuários novos.



Tendo que os usuários que cancelam o serviço são clientes novos, o gráfico mostra que o serviço falha em prover valor suficiente para manter novos usuários.



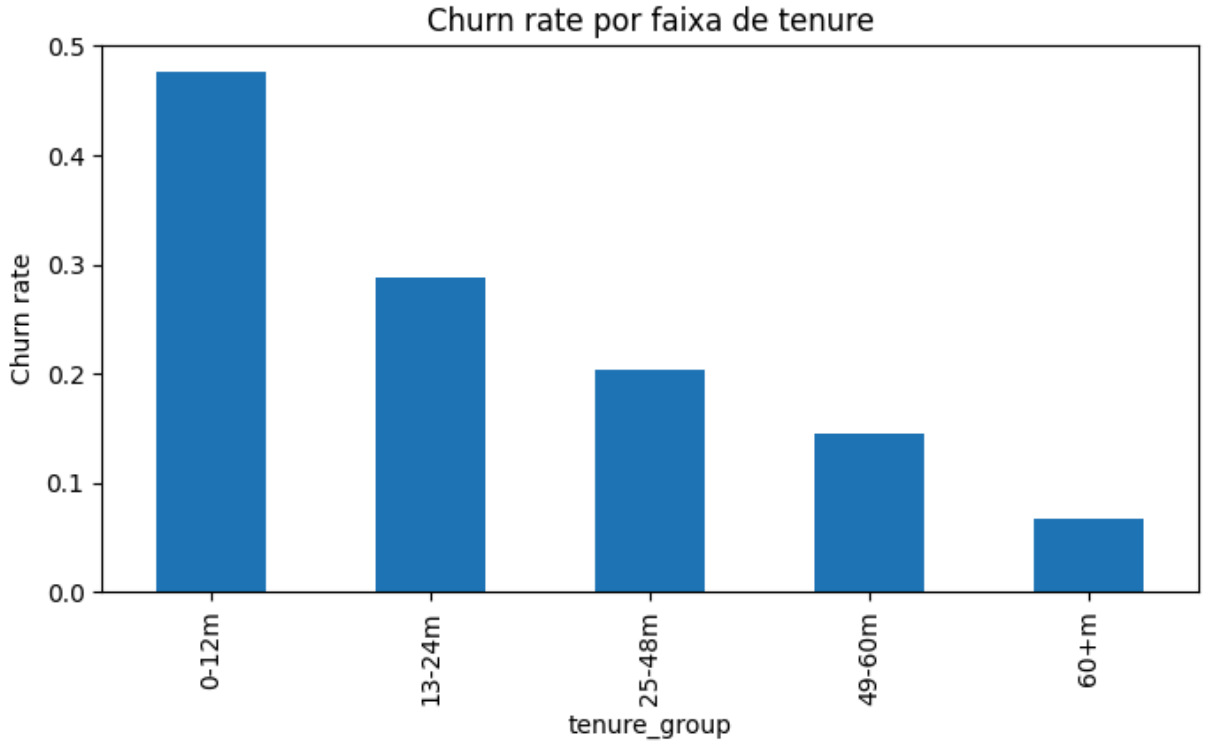
O gráfico de total charges reforça a ideia que a retenção de usuários novos é baixa, dado que o valor arrecadado dos usuários que cancelaram o serviço tem a tendência de ser menor.



Também foi criada uma matriz de correlação entre as variáveis numéricas, que permite confirmar as tendências entre as variáveis. Tenure e churn apresenta o valor -0,35, indicando que tenure e churn tem crescimento oposto, ou seja, quando tenure sobe churn desce, cimentando a alta evasão se usuários novos.

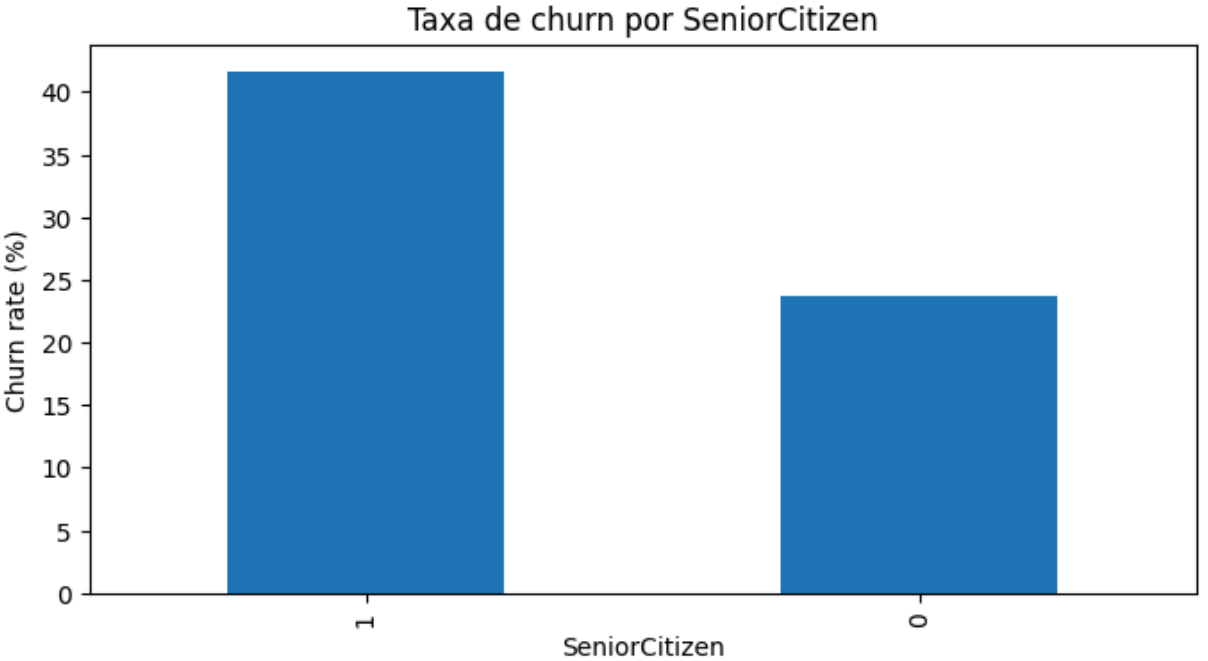
MonthlyCharges e churn apresentam 0.19, indicando que usuários que cancelaram o serviço têm cobranças mensais mais altas.

TotalCharges e churn também tem crescimento oposto, apresentando -0,20, reafirmando as informações obtidas analisando tenure.

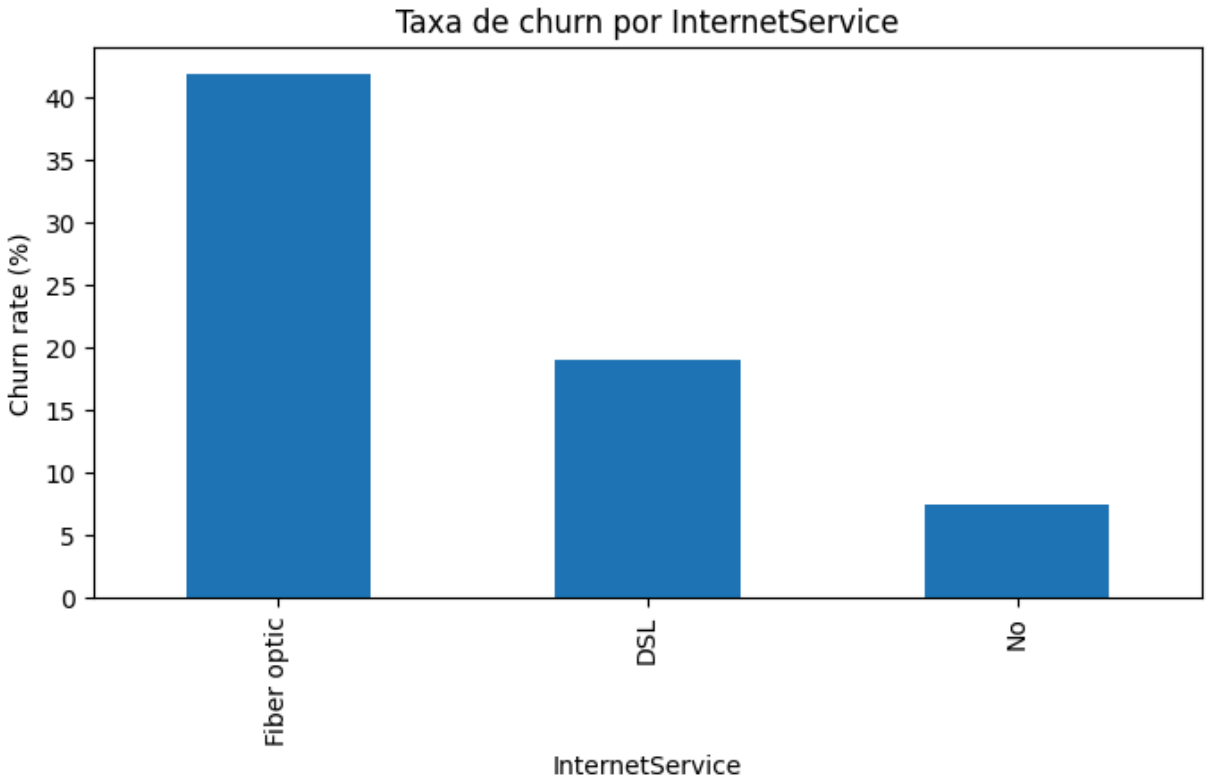


Para verificar com mais clareza a evasão de usuários foi plotado o gráfico de Churn por faixa de tenure. O gráfico mostra que a evasão de usuários ocorre principalmente no primeiro ano de contrato e cai de acordo com o tempo de contrato.

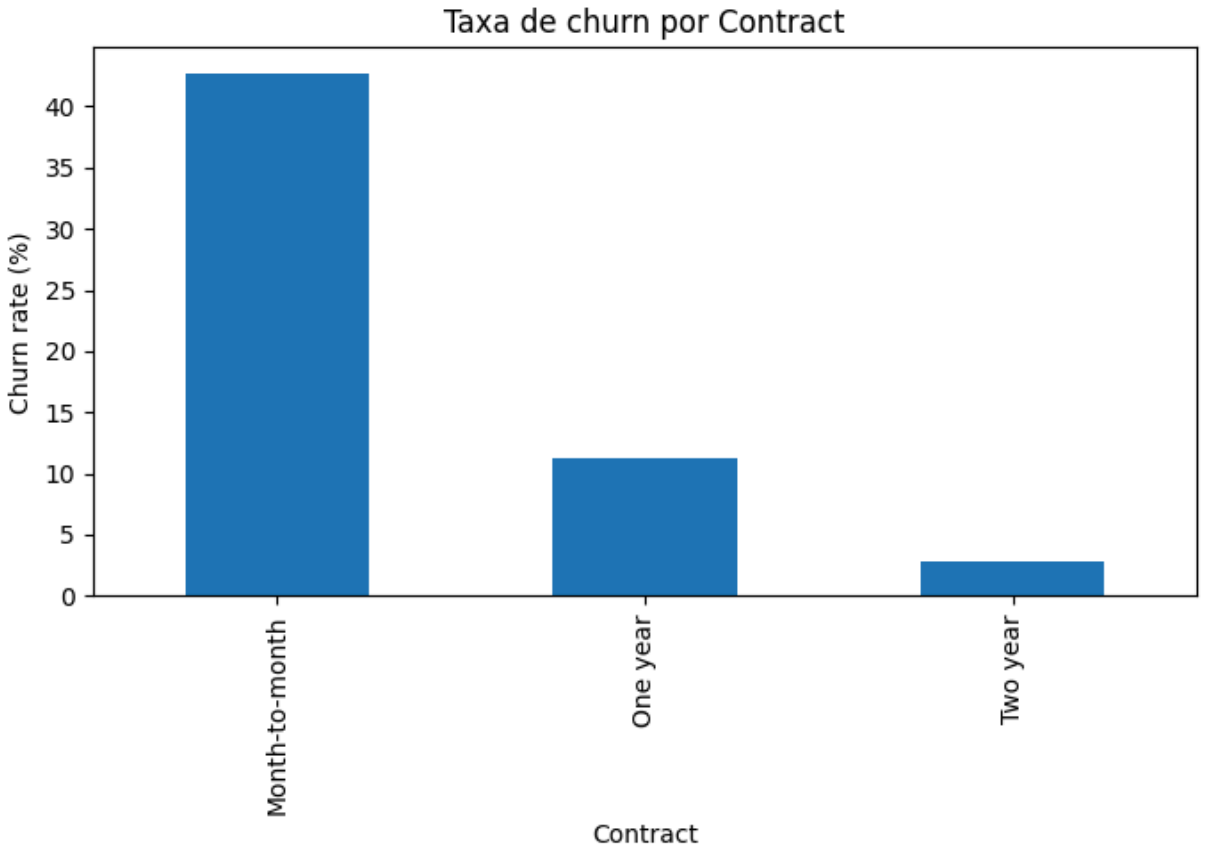
Para as variáveis categóricas, foram plotados gráficos de barra da taxa de churn por variável, mas para melhor visualização foram incluídos no corpo do texto apenas os gráficos que apresentaram resultados relevantes, ou seja, gráficos que realmente apresentaram uma relação entre a variável e a chance do usuário cancelar o serviço.



Usuários classificados como SeniorCitizen tem uma maior probabilidade de cancelar o serviço.



Usuários que assinaram fibra óptica tem a taxa de churn consideravelmente maior em relação ao serviço de internet DSL e usuários que não contrataram serviços de internet.

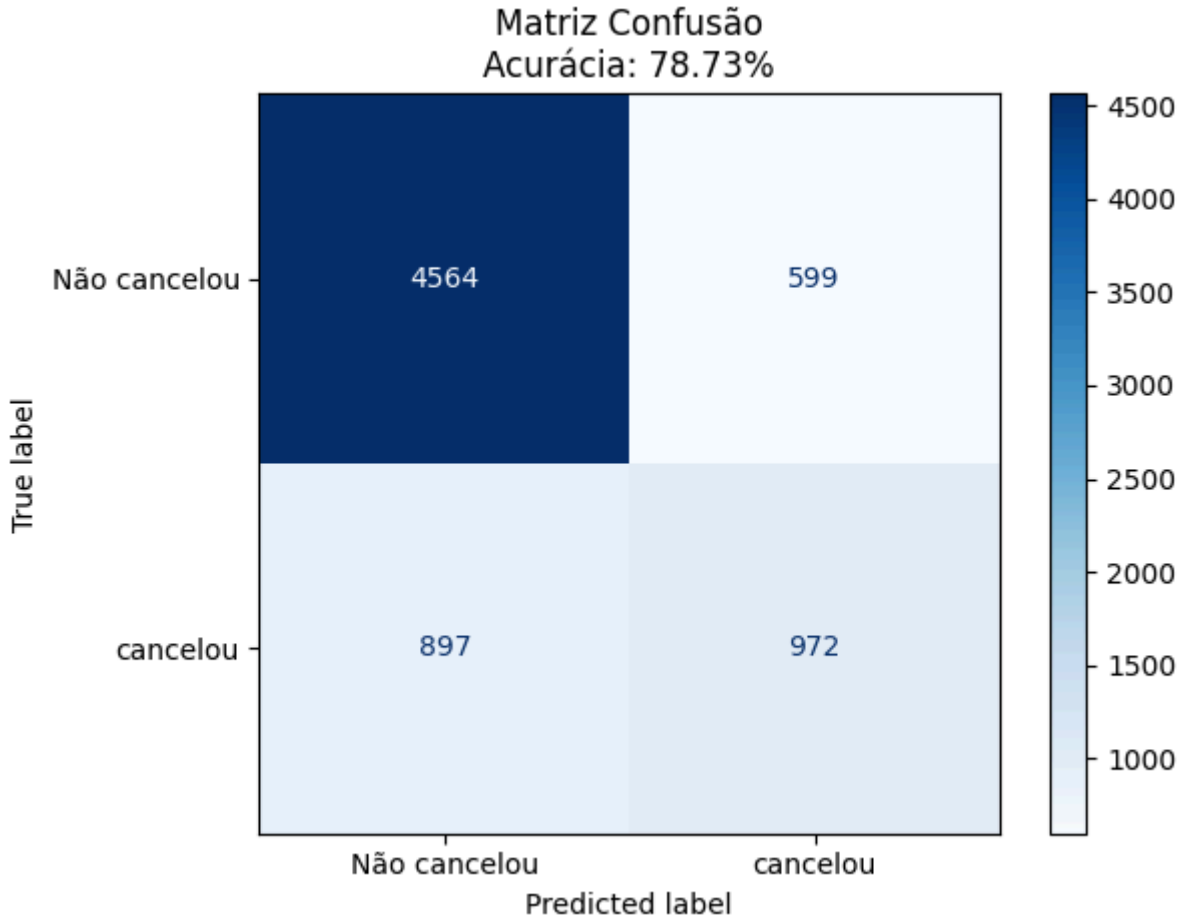


O gráfico de taxa de churn por tipo de contrato enfatiza que a maior parte dos cancelamentos são de usuários que pagam mensalmente. Podemos relacionar esses dados com informações anteriores, anteriormente pudemos verificar que os usuários que cancelaram o serviço são cobrados valores mensais mais altos, com isso podemos afirmar que o valor do contrato mensal está insatisfatório para os clientes.

APLICAÇÃO DA TÉCNICA PREDITIVA

A técnica preditiva usada foi regressão logística e o modelo LogisticRegression foi treinado com todo o dataset, excluindo as colunas de CustomerID e Churn.

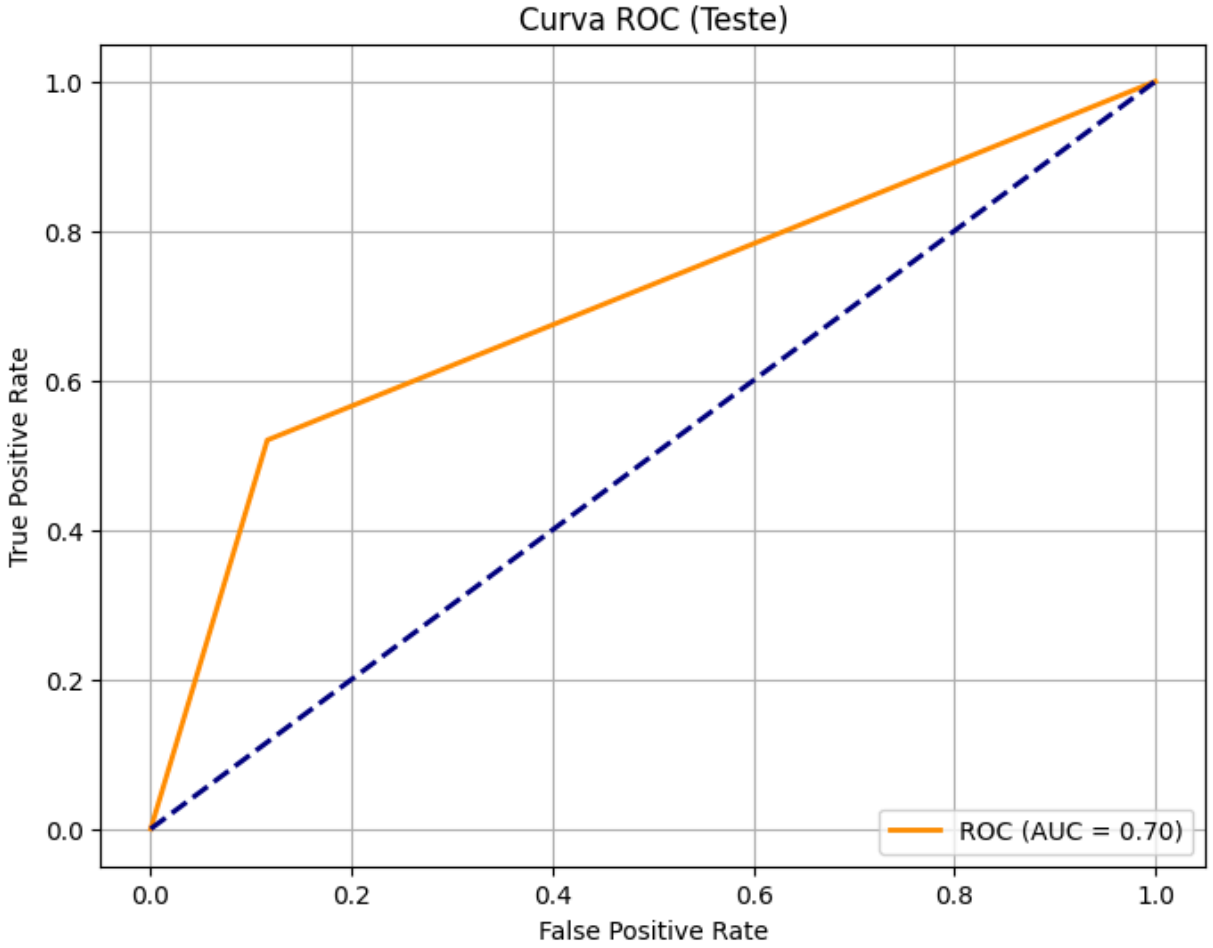
Após o treino foi gerada a matriz de confusão que indicou uma acurácia de 78,73%.



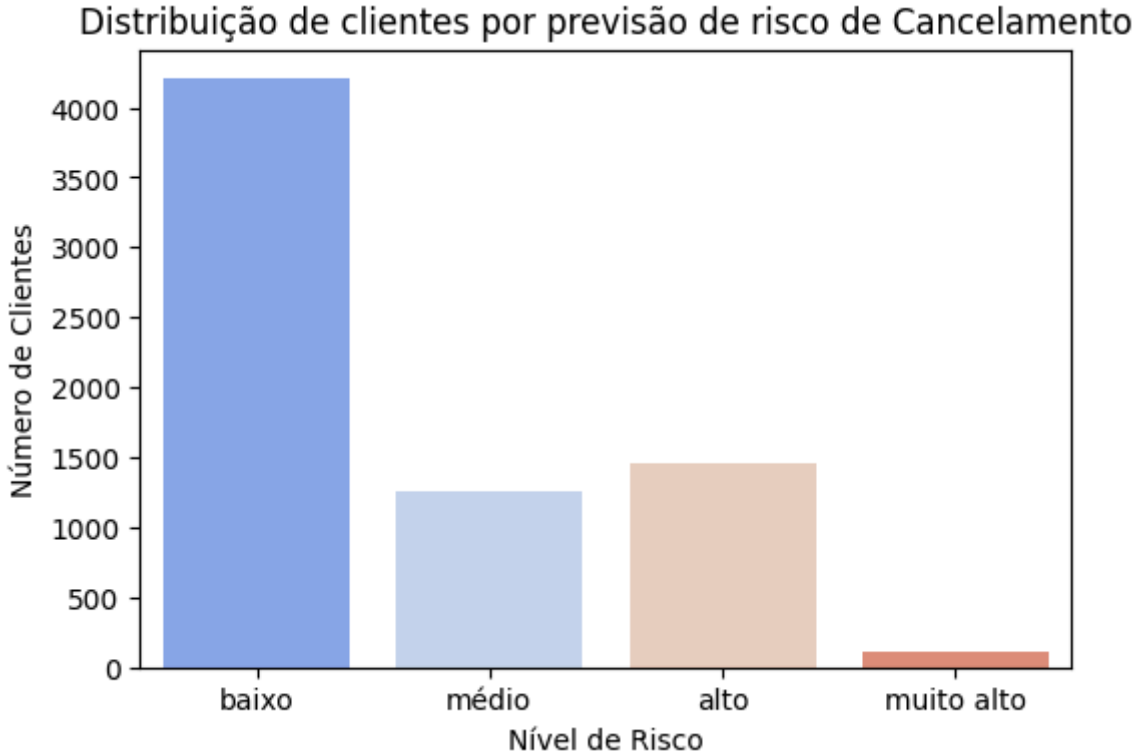
Também foi criada a tabela de classification_report:

	precision	recall	f1-score	support
0	0.835744	0.883982	0.859187	5163.000000
1	0.618714	0.520064	0.565116	1869.000000
accuracy	0.787258	0.787258	0.787258	0.787258
macro avg	0.727229	0.702023	0.712152	7032.000000
weighted avg	0.778061	0.787258	0.781027	7032.000000

Por fim foi plotado o gráfico da curva roc:



Com as probabilidades dos clientes obtidas foi criado um gráfico de risco de cancelamento.



DISCUSSÃO DOS RESULTADOS OBTIDOS

Com a conclusão do modelo preditivo foi possível obter uma métrica de risco de cancelamento para cada usuário, o que pode ser efetivo para tomar providências para evitar perder clientes.

Além disso, com a análise estatística foi possível descobrir quais tipos de usuário tem a maior tendência a cancelar o serviço e quais serviços oferecem maior taxa de evasão de usuários. Com isso é possível reestruturar planos e tomar iniciativa a fim de diminuir a saída de usuários.

CONSIDERAÇÕES FINAIS

A análise de churn realizada ao longo deste trabalho permitiu identificar os principais fatores que influenciam a saída de clientes de uma empresa. Por meio da utilização da Regressão Logística (Logistic Regression), foi possível desenvolver um modelo capaz de estimar a probabilidade de um cliente cancelar o serviço.

Com base nesses resultados, a empresa pode adotar estratégias preventivas mais direcionadas, como o fortalecimento do relacionamento com clientes de risco ou a criação de campanhas personalizadas.