Projeto em Data Science: Um dataset sobre evasão de clientes numa empresa de telecomunicações

Paulo Henrique Moreira Melo
Jean Lucas Almeida Mota
Erick Henrique Campolina da Silva
Evaldo Martins de Souza

12 de julho de 2023

I. Descrição do Problema

O conjunto de dados consiste em informações de 7043 clientes, com 33 variáveis, que fornecem uma ampla gama de detalhes sobre as características dos clientes, como perfil demográfico, localização geográfica, serviços contratados da empresa, valores das cobranças, método de pagamento e status de relacionamento com a empresa. No banco de dados, utiliza-se o termo "churn"para se referir ao cancelamento de contrato por parte dos clientes.

II. Informação Externa

Este projeto possui várias referências disponíveis online. Elas foram divididas em três categorias principais: código do projeto, banco de dados e apresentação em vídeo.

a). **Projeto no GitHub**
O código-fonte deste projeto está hospedado no GitHub. Acessível através do seguinte link: GitHub - Projeto ICD.

b). **Banco de dados**
Disponível na plataforma Kaggle, o dataset de churn de clientes da IBM Telco. Aqui está o link: Kaggle - IBM Telco Customer Churn.

c). **Video Youtube**
Para uma compreensão mais abrangente, um vídeo de apresentação do projeto foi produzido pelo nosso grupo e está disponível no YouTube: YouTube - Apresentação do Projeto.

III. Motivação

Esses dados oferecem uma oportunidade valiosa para análise de evasão de clientes em uma empresa de telecomunicações, permitindo a identificação de fatores que influenciam a retenção de clientes e a adoção de estratégias para melhorar a satisfação e fidelidade dos clientes.

IV. Perguntas de Pesquisa

Este projeto busca responder às seguintes perguntas-chave através da aplicação de modelos de regressão e classificação aos nossos dados.

1. **Modelo de Regressão:**
   (a) É possível montar um modelo de regressão utilizando as variáveis relativas à contratação de serviços para estimar o valor da cobrança mensal dos serviços?
   (b) Quais são as variáveis mais importantes para determinar a cobrança mensal dos serviços?
2. **Modelo de Classificação:**
   (a) É possível montar um modelo de classificação para distinguir entre clientes que cancelaram e clientes que ainda são clientes, usando variáveis relativas à localização geográfica, dados demográficos, serviços contratados e cobrança?
   (b) Quais são as variáveis mais importantes para classificar os clientes?

V. Conhecendo as Variáveis

Nosso conjunto de dados é composto por 7043 observações, cada uma com 33 variáveis distintas. Cada variável fornece informações únicas sobre os clientes, incluindo características demográficas, detalhes do serviço contratado, e informações de faturamento e pagamento. Vamos conhecer as variáveis:

CustomerID: Um ID único que identifica cada cliente.

Count: Um valor usado em relatórios/painéis para somar o número de clientes em um conjunto filtrado.

Country: O país de residência principal do cliente.

State: O estado de residência principal do cliente.

City: A cidade de residência principal do cliente.
Zip Code: O código postal da residência principal do cliente.
Lat Long: A latitude e longitude combinadas da residência principal do cliente.
Latitude: A latitude da residência principal do cliente.
Longitude: A longitude da residência principal do cliente.
Gender: O gênero do cliente: Masculino, Feminino.
Senior Citizen: Indica se o cliente tem 65 anos ou mais: Sim, Não.
Partner: Indica se o cliente tem um parceiro: Sim, Não.
Dependents: Indica se o cliente vive com quaisquer dependentes: Sim, Não. Dependentes podem ser filhos, pais, avós, etc.
Tenure Months: Indica a quantidade total de meses que o cliente esteve com a empresa até o final do trimestre especificado acima.
Phone Service: Indica se o cliente assina um serviço de telefone residencial com a empresa: Sim, Não.
Multiple Lines: Indica se o cliente assina várias linhas telefônicas com a empresa: Sim, Não.
Internet Service: Indica se o cliente assina um serviço de Internet com a empresa: Não, DSL, Fibra Óptica, Cabo.
Online Security: Indica se o cliente assina um serviço de segurança online adicional fornecido pela empresa: Sim, Não.
Online Backup: Indica se o cliente assina um serviço de backup online adicional fornecido pela empresa: Sim, Não.
Device Protection: Indica se o cliente assina um plano adicional de proteção de equipamentos para seus equipamentos de Internet fornecidos pela empresa: Sim, Não.
Tech Support: Indica se o cliente assina um plano adicional de suporte técnico da empresa com tempos de espera reduzidos: Sim, Não.
Streaming TV: Indica se o cliente usa seu serviço de Internet para transmitir programação de televisão de um provedor terceirizado: Sim, Não. A empresa não cobra uma taxa adicional por este serviço.
Streaming Movies: Indica se o cliente usa seu serviço de Internet para transmitir filmes de um provedor terceirizado: Sim, Não. A empresa não cobra uma taxa adicional por este serviço.
Contract: Indica o tipo de contrato atual do cliente: Mês a Mês, Um Ano, Dois Anos.
Paperless Billing: Indica se o cliente escolheu a faturação sem papel: Sim, Não.
Payment Method: Indica como o cliente paga a sua conta: Débito Bancário, Cartão de Crédito, Cheque Enviado por Correio.
Monthly Charge: Indica a cobrança mensal total atual do cliente por todos os seus serviços da empresa.
Total Charges: Indica as cobranças totais do cliente, calculadas até o final do trimestre especificado acima.
Churn Label: Sim = o cliente deixou a empresa este trimestre. Não = o cliente permaneceu com a empresa. Diretamente relacionado ao Valor de Churn.
Churn Value: 1 = o cliente deixou a empresa este trimestre. 0 = o cliente permaneceu com a empresa. Diretamente relacionado ao Rótulo de Churn.
Churn Score: Um valor de 0-100 que é calculado usando a ferramenta preditiva IBM SPSS Modeler. O modelo incorpora múltiplos fatores conhecidos por causar churn. Quanto maior a pontuação, mais provável é que o cliente deixe a empresa.

CLTV: Valor do Ciclo de Vida do Cliente. Um CLTV previsto é calculado usando fórmulas corporativas e dados existentes. Quanto maior o valor, mais valioso é o cliente. Clientes de alto valor devem ser monitorados para churn.

Churn Reason: A razão específica do cliente para deixar a empresa. Diretamente relacionado à Categoria de Churn.

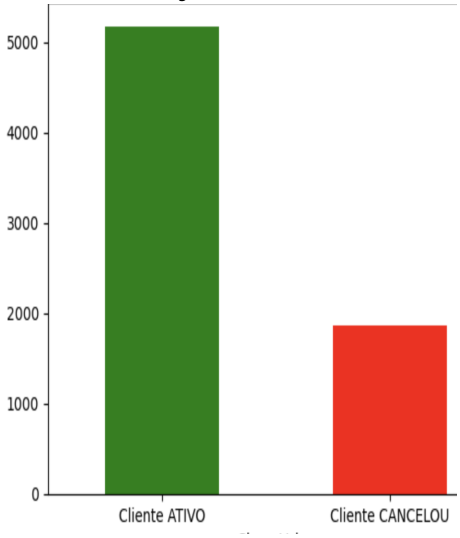
VI. Limpeza dos dados

Para limpeza inicial dos dados removemos a ['CustomerID'], visto que os clientes já podem ser identificado pelo índice que eles ocupam no banco de dados. Ademais, descobrimos que todos os clientes são da Califórnia, portanto removemos as colunas ['Count', 'Country', 'State'] pois continham apenas um valor e não acrescentavam nada à análise. A coluna ['Churn Reason'] possuia mais de 5174 valores nulos! Decidimos não eliminar todas as linhas com valores nulos, pois perderíamos mais da metade das amostras!

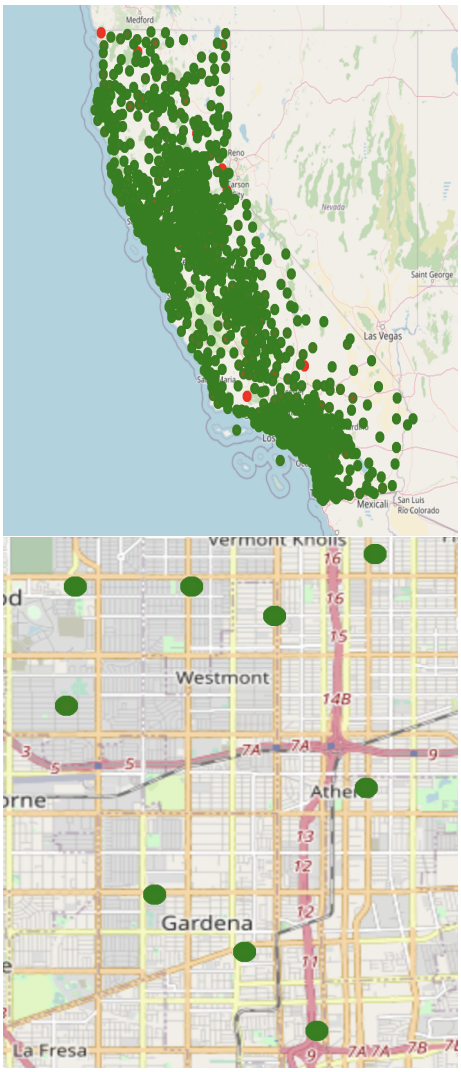
VII. Análise Exploratória

a). Dados Geógraficos

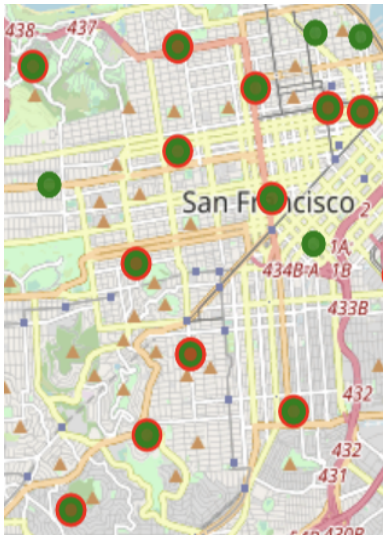
Sabemos que, aproximadamente, 5000 clientes estão ativos e 2000 clientes cancelaram o serviço.



Vamos dar uma olhada mais de perto na Califórnia usando a biblioteca Folium. Com ela, podemos desenhar um mapa 2D para procurar padrões de distribuição geográfica. Quem sabe conseguimos descobrir se há alguma área com problemas específicos de serviço?



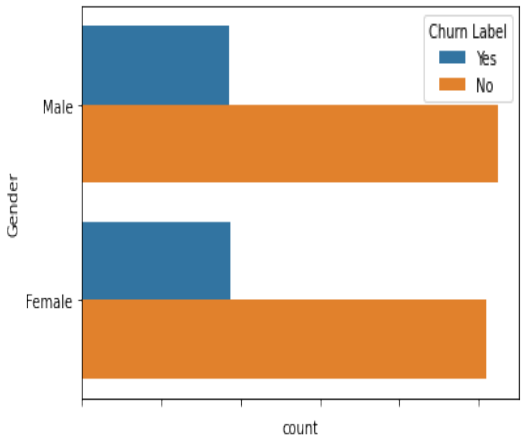
Vemos um excesso inesperado de pontos verdes. Até se esperaria que eles fossem maoria, mas mesmo dando zoom no mapa, quase não se vê pontos vermelhos. Minha hipótese: os pontos devem estar sobrepostos. Criemos, então, um gráfico onde os pontos vermelhos são um pouco maiores que os verdes.



Hipótese confirmada. Vemos vários pontos sobrepostos em quase todo mapa da Califórnia. Será que isso compromete a análise dos dados geográficos? Quantos pontos sobrepostos são?

Ao investigarmos a coluna 'Lat Long' em busca de duplicatas, descobrimos que todas as coordenadas possuem mais de um ponto associado. Mais intrigante ainda: esses pontos representam clientes com perfis demográficos e planos de serviços extremamente distintos, tornando improvável que sejam da mesma família. Portanto, concluímos que as variáveis geográficas estão corrompidas e decidimos retirá-las da nossa análise!

b). Dados demográficos



c). Dados de serviço e contratuais

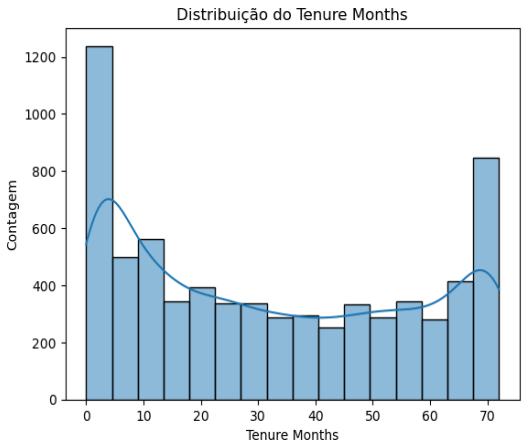


Figura 1. Distruibuição da variável tenure months, representando há quantos meses cada cliente está inscrito

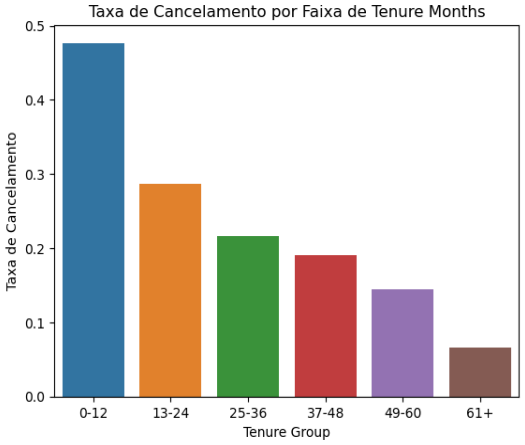
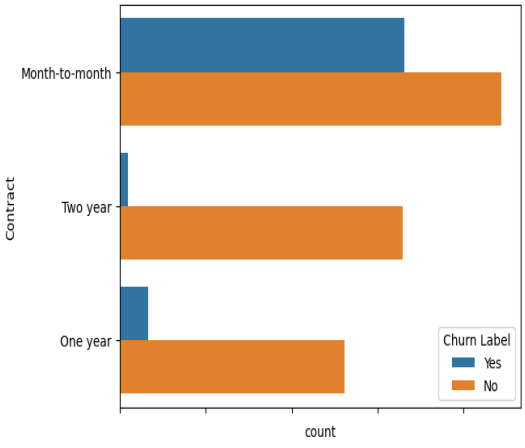
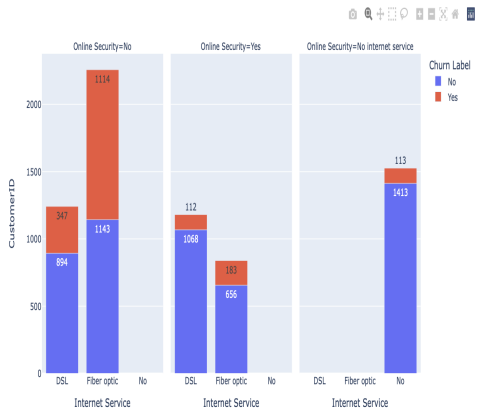
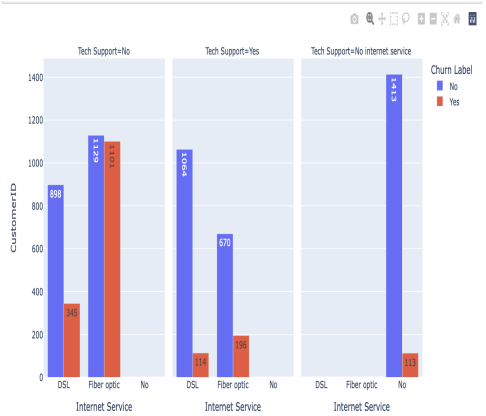


Figura 2. Veja que quanto mais tempo o cliente permance na empresa, menos ele tende a cancelar o serviço

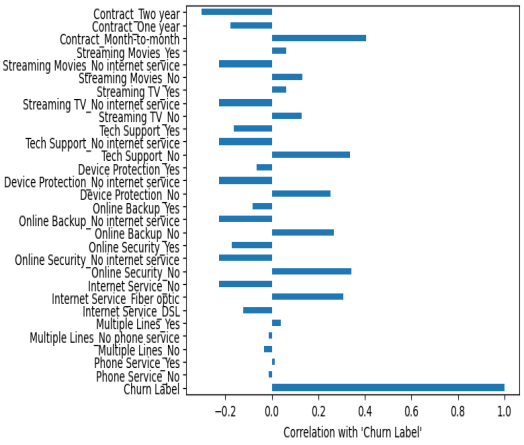


Há uma grande rotatividade em clientes com contrato mês a mês.





Quando internet é oferecida, a presença de suporte técnico parece contribuir para retenção de cliente. É provável que clientes estejam saindo por falta de suporte técnico



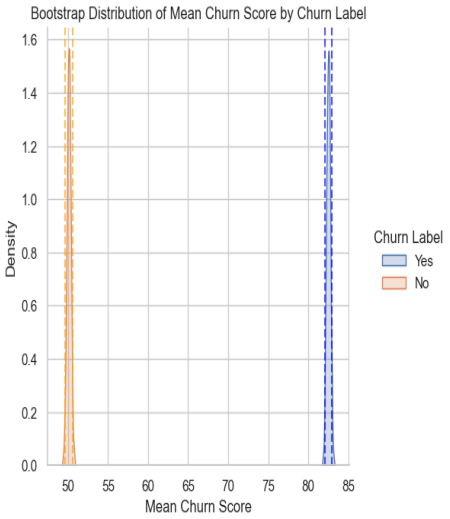
Vê-se como ausência de segurança online, ausência de suporte técnico, contrato mês-a-mês, ausência de proteção a dispositivo e oferta de fibra óptica se correlacionam com cancelamento do serviço com a empresa.

VIII. Teste de Hipótese

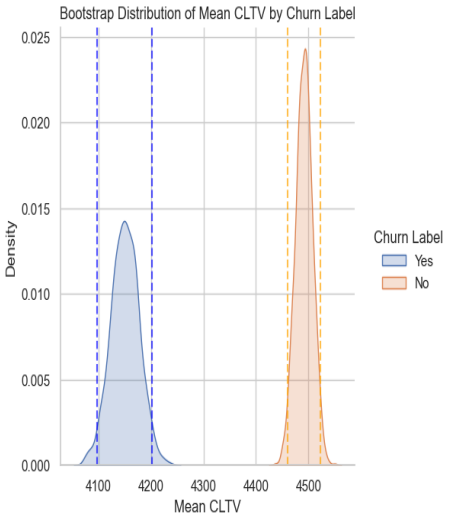
a). Bootstrapping

Realizaremos bootstrapping para averiguarmos se existe diferença estatística entre dos valores das varíaveis do dataset: 'Monthly Charges', 'CLTV', 'Churn Score', 'Total Charges' entre os clientes que cancelaram ou não o serviço com a nossa empresa. A Hipótese Nula (H0) é de que não há diferença estatística entre os valores das variáveis do clientes que cancelaram o serviço com a empresa e os que permanecem nela

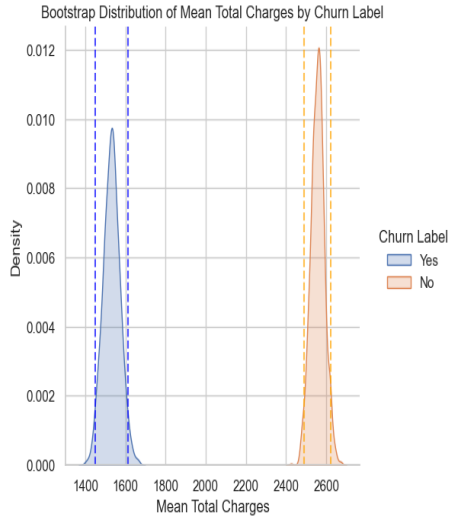
i. Bootstrapping Churn Score



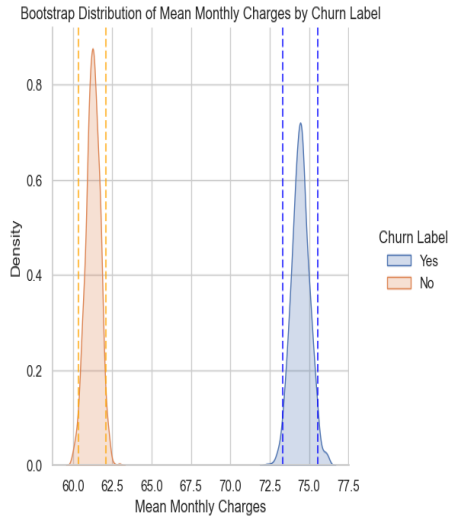
ii. Bootstrapping CLTV



iii. Bootstrapping Total Charges



iv. Bootstrapping Monthly Charges



Legenda das figuras i, ii, iii,iv:

As linhas verticais tracejadas represetam os intervalos de confiança (IC)

v. Conclusão do Bootstrapping

Houve diferença significativa entre os grupos de clientes inativos e ativos de todas as variáveis analisadas. Podemos rejeitar as hipótese nula (H0).

Ademais, vê-se que clientes que cancelam em média possuem um CTLV menor (menos importantes para empresa) o que é um sinal de que não estão cancelando os melhores perfis de cliente

Igualmente, vê-se que clientes que pagam uma mensalidade maior tendem a cancelar o serviço, entretando, cliente que pagaram um maior valor total tendem a permanecer e a não cancelar o relacinamento com a empresa. Muito possivelmente, quanto maior tempo um cliente permanece na empresa, menos a chance de ele encerrar o serviço.

b). **T test**

Variable	T-statistic	p-value
Churn Score	74.6	0.000
CLTV	-10.78	0.000
Total Charges	-17.0	0.0000
Monthly Charges	16.53	0.000

IX. Regressão

a). **Regressão Monthly Charges**

Queremos montar uma regressão múltipla pegando APENAS as variáveis relativa aos serviços prestados para prever o valor cobrado mensalmente a cada cliente. Portanto, queremos as variáveis ['Partner', 'Dependents', 'Phone Service', 'Multiple Lines', 'Internet Service', 'Online Security', 'Online Backup', 'Device Protection', 'Tech Support', 'Streaming TV', 'Streaming Movies', 'Contract', 'Paperless Billing', 'Payment Method'] para prever a variável 'Monthly Charges'. Acrescentamos também as variáveis relativas ao metódo de pagamento, pois seria um hipótese de que a empresa cobre diferente pelos métodos de pagamento que ela fornece, e queremos saber se for o caso.

i. One Hot Encoding

Foi-se realizada um one hot encoding pois todas as variáveis preditoras eram categóricas, muitas delas com múltiplos valores categóricos diferentes.

ii. Random Forest Regressor

Fizemos uso do RandomForestRegressor, um algoritmo de aprendizado de máquina fornecido pela biblioteca sklearn.ensemble. Este algoritmo gera várias árvores de decisão e combina suas previsões. Um recurso valioso deste algoritmo é a capacidade de calcular a

importância de cada característica no modelo, medindo o impacto de cada característica na redução da impureza dos nós da árvore. Estas informações nos ajudam a entender quais características são mais significativas nas previsões do modelo e auxiliam na seleção de características, potencialmente permitindo a remoção de características menos importantes para simplificar o modelo. Ao final, descobrimos que as variáveis 'Internet Service', 'Streaming Movies', 'Device Protection', 'Streaming TV' eram as variáveis com maior importância na cobrança mensal 'Monthly Charges'.

iii. Regressão com modelos mais simples

A fim de explorar a possibilidade de criar modelos mais eficientes e simplificados, realizamos um experimento adicional. Neste experimento, criamos dois modelos de regressão adicionais: o primeiro usando apenas as quatro variáveis mais importantes identificadas pelo RandomForestRegressor ('Internet Service', 'Streaming Movies', 'Device Protection', 'Streaming TV'), e o segundo utilizando todas as outras variáveis, excluindo as quatro primeiras. Para essa comparação, selecionamos três métricas de avaliação: o Erro Absoluto Médio (MAE), o Erro Quadrático Médio (MSE) e o coeficiente de determinação (R^2).

iv. Resultado das regressões

Modelo	Mean Absolute Error	Mean Squared Error	R^2
Todas variáveis (14)	0.79	1.01	0.99
As 4 mais importantes	6.25	74.0	0.92
As varaveis restantes (10)	10.2	166.1	0.82

v. Conclusão da regressão

Conclui-se que Serviço de Internet, o Serviço de Streaming, e Proteção de Dispositivo são os serviços contratados que mais influenciam a cobrança mensal, sendo possivelmente os serviços mais caros oferecidos pela empresa. Usando somente as 4 variáveis mais importantes conseguimos montar um modelo melhor que as outras 10 restantes, porém ambas deixaram a desejar em relação ao modelo com as 14 variáveis, que teve um R2 excelente (>0.99). Concluímos que a mais de 99 % da variância em Monthly Charges é explicada pelas variância das 14 variáveis relativas ao serviço prestado aos clientes. Concluímos, sobretudo, que: É possivel montar um excelente modelo de predição do preço usando as variáveis dos serviços prestados ao clientes.

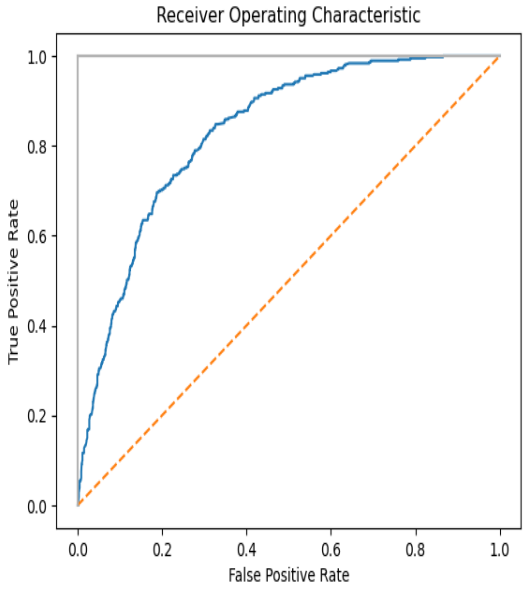
X. Classificação

a). Regressão Logística

A regressão logística pode ser usada como classificador ao estebelecermos um ponto de corte. Foi-se estabelido ponto de corte p = 0.5 . Assim resultados acima de 0.5 serão classificados como 1 e resultados abaixo de 0.5 serão classificados como 0. Usamos dados demográficos e do serviço prestado e da cobrança para montar nosso modelo. Excluímos CLTV, Churn Score e Churn Reason, pois queremos saber se é possível prever apenas com base nas variáveis mais concretas sobre o cliente.

b). **Resultados**

	predicted 0	predicted 1
actual 0	906	142
actual 1	158	203
precision		0.79



c). **Conclusão da Classificação**

As variáveis "Contract-Month-to-month", "Dependents-No", "Internet Service-Fiber optic", "Payment Method-Electronic check", "Online Security-No", e "Tech Support-No"se destacam com as maiores odds ratios, o que sugere que elas têm um papel significativo na previsão do churn dos clientes.

- **Contract Month-to-month** (Odds Ratio = 3.3): Esta é a variável com a maior odds ratio, o que significa que os clientes com contratos de mês a mês são cerca de 3,31 vezes mais propensos a cancelar seu serviço do que aqueles com outros tipos de contrato.
- **Dependents-No** (Odds Ratio = 2.19): Isso indica que os clientes que não têm dependentes são aproximadamente 2,20 vezes mais propensos a cancelar seu serviço.
- **Internet Service-Fiber optic** (Odds Ratio = 1.46): Os clientes que usam o serviço de internet por fibra óptica têm 1,47 vezes mais chances de cancelar seu serviço.
- **Contract two-year** (Odds Ratio = 0.31): Com odds ratio bem abaixo de 1, sabemos que contract two-year é um fator importante de retenção de cliente, sendo o fator de retenção mais importante de todas as variáveis analisadas!