Predição de churn em uma empresa de telecomunicações através de modelos de Machine Learning

Eric Henrique Cestari[1]; Felipe Pinto Da Silva[2]

[1] Graduado em Engenharia de Produção pela Universidade São Francisco – USF e Pós graduado em Administração de Empresas pela Fundação Getúlio Vargas – FGV. Rua Dezesseis, 76 – Jardim São Marcos; 13272-830 Valinhos, São Paulo, Brasil
[2] Aluno de Doutorado do programa de pós-graduação em Economia pela Universidade Estadual de Campinas - UNICAMP. Mestre em Ciências Econômicas. Rua Pitágoras, 353 – Cidade Universitária; 13083-857 Campinas, São Paulo, Brasil
*autor correspondente: ericcestari@hotmail.com

Predição de churn em uma empresa de telecomunicações através de modelos de Machine Learning

Resumo

Este estudo teve como objetivo a predição de churn em uma empresa de telecomunicações utilizando modelos de Machine Learning. A perda de clientes é um problema crítico para as empresas desse setor, já que é mais caro adquirir novos clientes do que manter os existentes. Para realizar o estudo, foi utilizada uma base de dados de clientes, contendo informações como gênero, histórico de contratos e utilização de serviços. Foram aplicados modelos de Machine Learning, árvores de decisão, random forest, xgboost e regressão logística, para prever a probabilidade de um cliente deixar a empresa. Os resultados mostraram que os modelos de Machine Learning são eficazes na predição de churn. O XGBoost e Random Forest foram os modelos mais precisos na previsão de churn. A análise das variáveis mais importantes para a predição mostrou que o TipoContrato: mês-a-mês, MotivoChurn: Concorrente ofereceu velocidade de download mais alta e MotivoChurn: Concorrente ofereceu maior pacote de dados foram os fatores mais relevantes para a previsão de churn. A metodologia utilizada para o estudo incluiu a aplicação de técnicas de pré-processamento de dados, treinamento e teste dos modelos de Machine Learning, bem como a avaliação da precisão dos modelos. Portanto, este trabalho demonstrou que a utilização de modelos de Machine Learning pode ser uma abordagem promissora para a predição de churn em empresas de telecomunicações, possibilitando o desenvolvimento de estratégias preventivas para reduzir a perda de clientes e aumentar a satisfação e fidelização dos mesmos.
Palavras-chave: Clientes; Dados; Análise; Python.

Churn prediction in a telecommunications company through Machine Learning models

Abstract

This study aimed to predict churn in a telecommunications company using Machine Learning models. Customer churn is a critical problem for companies in this sector, as it is more expensive to acquire new customers than to retain existing ones. To conduct the study, a customer database was used, containing information such as gender, contract history, and service usage. Machine Learning models such as random forest decision trees, xgboost, and logistic regression were applied to predict the probability of a customer leaving the company. The results showed that Machine Learning models are effective in predicting churn. The XGBoost and Random Forest models were the most accurate in predicting churn. Analysis of the most important variables for prediction showed that Contract: Month-to-month, Churn Reason: Competitor offer higher download speeds and Churn Reason: Competitor offered more data were the most relevant factors for predicting churn. The methodology used for the study included applying data pre-processing techniques, training and testing Machine Learning models, as well as evaluating the accuracy of the models. Therefore, this study demonstrated that the use of Machine Learning models can be a promising approach for predicting churn in telecommunications companies, enabling the creation of preventive strategies to reduce customer loss and increase customer satisfaction and loyalty.

Keywords: Customers; Data; Analysis; Python.

Introdução

A migração ou abandono do cliente para uma outra empresa que preste um serviço ou venda algum tipo de produto similar é um problema enfrentado por empresas ao redor do mundo, praticamente todos os setores estão sujeitos a isso. Essa migração acontece quando um cliente em um determinado período deixa a base de clientes ativos de uma empresa, por exemplo, quando deixa de assinar os seus serviços de internet e assina um outro plano de internet de outra empresa ou quando deixa de comprar seus produtos para optar por outros da concorrência (Gold, 2020). Esse fenômeno é chamado de “churn”.

A métrica é baseada na quantidade de clientes que uma empresa perde em sua base total de clientes. Manter um controle sobre a taxa de “churn” de uma empresa é estratégico porque pode afetar significativamente a saúde financeira e a sustentabilidade a longo prazo da empresa, estima-se que custa em torno de cinco vezes mais adquirir novos clientes do que retê-los, dependendo do setor (Kurtz e Clow, 1998).

Caso a empresa não tenha um serviço por assinatura, pode-se levar em consideração a quantidade de clientes que deixaram a base, considerando-se os clientes que não compraram nenhum produto nos últimos seis meses por exemplo dependendo do tipo de produto vendido (Gold, 2020).

A habilidade de prever o risco de “churn” é valiosa no mundo dos negócios, permitindo tomar decisões antecipadas para evitar a perda de clientes. A prevenção de “churn” não é uma tarefa fácil, Gold (2020) diz que muitas ações podem ser tomadas como por exemplo, melhorar as características do produto, ajustar o preço ou até mesmo melhorar a interface de usuário, para que essa jornada seja mais assertiva, técnicas de “machine learning” podem ser utilizadas.

A importância dos dados é inegável no mundo atual, uma vez que estão presentes em diversos formatos e lugares, desde números, vídeos, imagens e áudios, estruturados ou não. Dispositivos como “smartphones”, câmeras de segurança, sensores, redes sociais e sites, bem como carros e casas inteligentes e dispositivos esportivos, geram informações que são armazenadas em grandes bancos de dados, permitindo a análise de padrões de uso e consumo.

Nesse contexto, técnicas de “machine learning” têm sido amplamente adotadas por empresas de todo o mundo para identificar semelhanças em meio a essa vasta quantidade de dados, sendo uma intersecção entre estatística, inteligência artificial, ciência da computação e matemática Müller e Guido (2016), portanto, serão abordados especificamente, os modelos supervisionados de “machine learning”, que buscam prever respostas a partir de dados de entradas conhecidas, conforme discutido por Müller e Guido (2016).

A utilização de “machine learning” para a previsão de “churn” é uma estratégia cada vez mais adotada por empresas de diversos setores, pois permite uma análise mais precisa e eficiente do comportamento dos clientes. Com o crescente volume de dados gerados por diversas fontes a análise manual se torna inviável e menos confiável. Além disso, as técnicas de “machine learning” são capazes de identificar padrões e tendências que poderiam passar despercebidos em uma análise humana, permitindo uma tomada de decisão mais precisa e fundamentada. Por essas razões, justifica-se a escolha do uso de “machine learning” para a previsão de “churn” em empresas que buscam manter sua base de clientes e garantir sua sustentabilidade a longo prazo. Um dos benefícios de utilizar modelos de “machine learning” é justamente eliminar a subjetividade na tomada de decisão, com o mínimo de intervenção humana possível, definitivamente as máquinas podem aprender e ajudar os humanos (Géron, 2019).

O objetivo principal é definir o melhor modelo de “machine learning” para prever o risco de “churn”, um problema de classificação binária, para isso, serão analisados dados de uma empresa fictícia de telecomunicações, verificando os formatos em que os dados estão dispostos e entendendo quais variáveis podem ou não influenciar e fidelizar os clientes, após essa análise, os dados serão estruturados e submetidos a quatro diferentes modelos de “machine learning”, em seguida será analisado qual modelo apresenta o melhor desempenho de predição, considerando o risco de “churn”, esta informação será útil para tomada de decisão estratégica da companhia.

Material e Métodos

Para desenvolver as análises e entender o comportamento dos clientes, foi utilizada a linguagem de programação Python, amplamente utilizada, submetendo os dados à análise exploratória e também a quatro modelos supervisionados de “machine learning”, sendo eles: Árvore de decisão, que é um modelo de aprendizado de máquina que representa um conjunto de regras de decisão hierárquica em forma de árvore. Frequentemente utilizada para classificar dados em categorias ou prever valores numéricos (Müller e Guido, 2016). Também o modelo “Random Forest” que é um algoritmo de aprendizado de máquina que cria várias árvores de decisão independentes e combina suas previsões para melhorar a precisão e evitar “overfitting”, ou seja, evitar que o modelo se ajuste demasiadamente aos dados de treinamento e perde a capacidade de generalização (James et al., 2021). O XGBoost, que é um algoritmo de aprendizado de máquina também baseado em árvores de decisão, projetado para melhorar a velocidade e a precisão do modelo. Ele usa um processo de otimização de gradiente estocástico para ajustar os pesos das amostras de treinamento e minimizar a perda (Géron, 2019). E por fim a Regressão logística, que se trata de um modelo de aprendizado de

máquina usado para prever uma variável categórica binária a partir de variáveis explicativas, ele usa uma função logística para mapear as variáveis de entrada para a probabilidade de saída (James et al., 2021).

Uma base de dados fictícia foi utilizada, trazendo informações de serviços de internet e telefone fixo. A procedência da base de dados é da “International Business Machines Corporation [IBM]”, que traz o “churn” de clientes de uma empresa, com 33 variáveis, onde uma delas é a variável alvo “churn”, que indica sim ou não para cada cliente específico no último mês, dentre as 7.043 observações que totalizam o conjunto de dados de clientes da Califórnia, no terceiro quadrimestre do ano.

Cada linha do banco de dados representa um cliente e cada coluna representa uma variável, ou seja, uma característica relacionada à assinatura de determinado cliente. No início desse estudo, os dados foram importados utilizando a linguagem de alto nível Python e em seguida foram tratados, traduzindo cada variável do Inglês para o Português, para facilitar a compreensão e análise, cada variável tem um significado específico, conforme mostrado na Tabela 1.

Tabela 1. Variáveis do conjunto de dados

(continua)

Variável	Tradução	Significado
CustomerID	IDUsuario	Código único que identifica cada cliente
Count	Contagem	Variável de contagem
Country	Pais	País onde reside o cliente
State	Estado	Estado onde reside o cliente
City	Cidade	Cidade onde reside o cliente
Zip Code	CodigoPostal	Código postal onde reside o cliente
Lat Long	Lat Long	Combinação de latitude e longitude, referente a residência do cliente
Latitude	Latitude	Coordenada latitude onde reside o cliente
Longitude	Longitude	Coordenada longitude onde reside o cliente
Gender	Genero	Gênero do cliente, masculino ou feminino
Senior Citizen	Senior	Indica se o cliente tem 65 anos ou mais, sim ou não
Partner	Parceiro	Informa se o cliente possui parceiro ou não.
Dependents	Dependentes	Diz se o cliente possui dependentes, sim ou não
Tenure Months	AssinaturaMeses	Informa o total de meses que o cliente está na empresa
Phone Service	ServicoTelefonia	Cliente assina o serviço telefônico domiciliar da empresa, sim ou não
Multiple Lines	MultiplasLinhas	Diz se o cliente assina várias linhas telefônicas com a empresa, sim ou não
Internet Service	ServicoInternet	Diz se o cliente possui serviço de internet
Online Security	SegurancaOnline	Diz se o cliente assinou segurança adicional
Online Backup	BackupOnline	Diz se o cliente assina um serviço adicional de backup online fornecido pela empresa, sim ou não
Device Protection	ProtecaoDispositivo	Diz se o cliente assina um plano de proteção de dispositivo adicional para o seu equipamento de Internet, sim ou não

Tabela 1. Variáveis do conjunto de dados

(conclusão)

Variável	Tradução	Significado
Tech Support	SuporteTech	Diz se o cliente assina um plano de suporte técnico adicional da empresa com tempos de espera reduzidos, sim ou não
Streaming TV	StreamingTV	Diz se o cliente usa seu serviço de Internet para transmitir programação de televisão de um provedor terceirizado, sim ou não
Streaming Movies	StreamingFilmes	Diz se o cliente usa seu serviço de Internet para transmitir filmes de um provedor terceirizado, sim ou não
Contract	TipoContrato	Informa o tipo de contrato atual do cliente: mensal, um ano ou dois anos
Paperless Billing	FaturaSemPapel	Fatura sem papel, sim ou não
Payment Method	MetodoPagamento	Diz como o cliente paga sua fatura: débito automático, transferência bancária, cartão de crédito ou boleto
Monthly Charges	CobrancasMensais	Indica a cobrança mensal total atual do cliente referente a todos os serviços utilizados
Total Charges	CobrancasTotais	Indica as cobranças totais do cliente, calculadas até o final do trimestre
Churn Label	Churn	Indica se o cliente saiu da empresa ou permanece assinando os serviços, sim ou não
Churn Value	ValorChurn	Variável que indica se o cliente saiu da empresa ou permanece assinando os serviços, 1 para sim ou 0 para não
Churn Score	PontuacaoChurn	Valor de 0 a 100 que é calculado usando a ferramenta preditiva IBM SPSS Modeler. O modelo incorpora vários fatores conhecidos por causar churn. Quanto maior a pontuação, maior a probabilidade de rotatividade do cliente
CLTV	CLTV	Valor vitalício do cliente. Um CLTV previsto é calculado usando fórmulas corporativas e dados existentes. Quanto maior o valor, mais valioso o cliente. Clientes de alto valor, devem ser monitorados quanto ao churn
Churn Reason	MotivoChurn	Motivo específico de um cliente para deixar a empresa. Diretamente relacionado à categoria Churn. Atitude da pessoa de apoio, Concorrente oferece velocidades de download mais altas, Concorrente ofereceu maior pacote de dados, não sabe informar, Concorrente fez oferta melhor, Atitude do prestador de serviços, Concorrente tinha dispositivos melhores, Confiabilidade da rede, Insatisfação com o produto, Preço muito alto, Insatisfação com o serviço, Falta de autoatendimento no site, Taxas extras de dados, Movido, Gama limitada de serviços, Tarifas de longa distância, Falta de velocidade de download/upload acessível, Pouco conhecimento de suporte telefônico, Pouco conhecimento de suporte online ou Falecido

Fonte: Dados originais da pesquisa

Foram aplicadas técnicas de pré-processamento de dados, incluindo a remoção de variáveis desnecessárias, tratamento de valores faltantes por interpolação e de valores discrepantes, bem como a transformação de variáveis categóricas em variáveis numéricas, chamada de “dummies”. Para isso, foram utilizadas as bibliotecas Pandas e Numpy para análise e manipulação, Seaborn e Matplotlib em visualizações dos dados. Utilizando o método info() foram identificados dados faltantes apenas na variável MotivoChurn, conforme mostrado na Tabela 2.

Tabela 2. Índice, nome das variáveis, quantidade de dados faltantes e dtypes

#	Column	Non-Null Count	Dtype
0	IDUsuario	7043 non-null	object
1	Contagem	7043 non-null	int64
2	Pais	7043 non-null	object
3	Estado	7043 non-null	object
4	Cidade	7043 non-null	object
5	CodigoPostal	7043 non-null	int64
6	Lat Long	7043 non-null	object
7	Latitude	7043 non-null	float64
8	Longitude	7043 non-null	float64
9	Genero	7043 non-null	object
10	Senior	7043 non-null	object
11	Parceiro	7043 non-null	object
12	Dependentes	7043 non-null	object
13	AssinaturaMeses	7043 non-null	int64
14	ServicoTelefonia	7043 non-null	object
15	MultiplasLinhas	7043 non-null	object
16	ServicoInternet	7043 non-null	object
17	SegurancaOnline	7043 non-null	object
18	BackupOnline	7043 non-null	object
19	ProtecaoDispositivo	7043 non-null	object
20	SuporteTech	7043 non-null	object
21	StreamingTV	7043 non-null	object
22	StreamingFilmes	7043 non-null	object
23	TipoContrato	7043 non-null	object
24	FaturaSemPapel	7043 non-null	object
25	MetodoPagamento	7043 non-null	object
26	CobrancasMensais	7043 non-null	float64
27	CobrancasTotais	7043 non-null	object
28	Churn	7043 non-null	object
29	ValorChurn	7043 non-null	int64
30	PontuacaoChurn	7043 non-null	int64
31	CLTV	7043 non-null	int64
32	MotivoChurn	1869 non-null	object

Fonte: Dados originais da pesquisa

A variável CobrancasTotais está em formato “object”, foi necessário transformá-la no tipo “float” através de uma função criada para isso, em seguida foram evidenciados mais 11 dados faltantes, para lidar com isso, foi utilizado um método de interpolação através da média

das cobranças totais. A variável PontuacaoChurn foi criada através de cálculos pela IBM e ela traz uma pontuação que informa a tendência de o cliente sair ou não do serviço, quanto maior, mais chances de rotatividade tem o cliente, portanto a variável foi removida para não influenciar o estudo, já que é exatamente essa previsão que está sendo buscada.

A análise exploratória dos dados foi iniciada, com o objetivo de extrair informações sobre a relação de “churn” com cada comportamento dos clientes, a proporção da base de dados está dividida entre 26,54% “Churn” e 73,46% Não “Churn”, conforme indica a Figura 1.

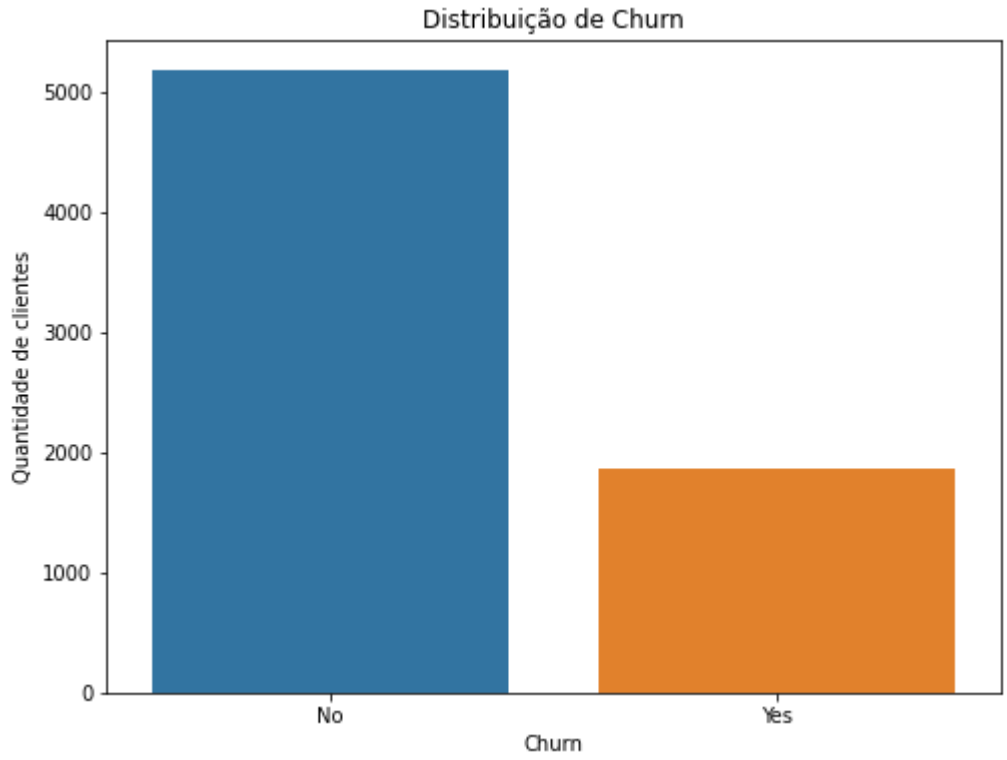

Figura 1. Distribuição de churn e não churn entre os clientes
Fonte: Dados originais da pesquisa

Para os dados que serão utilizados para treinar os modelos de “machine learning”, é importante considerar a presença de valores atípicos, também conhecidos como “outliers”. Segundo Géron (2019), “outliers” podem afetar negativamente a qualidade dos dados, uma vez que esses valores são mais propensos a estarem incorretos e, consequentemente, podem influenciar negativamente os resultados dos modelos preditivos. Para minimizar os efeitos dos “outliers”, é comum utilizar gráficos do tipo “box-plot” para cada variável numérica, a fim de visualizar a distribuição dos valores através de seus quartis. Dessa forma, é possível identificar a presença de valores extremos e avaliar se eles devem ser removidos ou tratados de alguma forma antes de aplicar o modelo de “machine learning”. Portanto, é importante considerar a presença de “outliers” e adotar medidas adequadas para lidar com esses valores, a fim de obter resultados mais precisos e confiáveis nos modelos de machine learning. Após

a análise dos gráficos da Figura 2 abaixo, ficou evidenciado que os dados não possuem outliers.

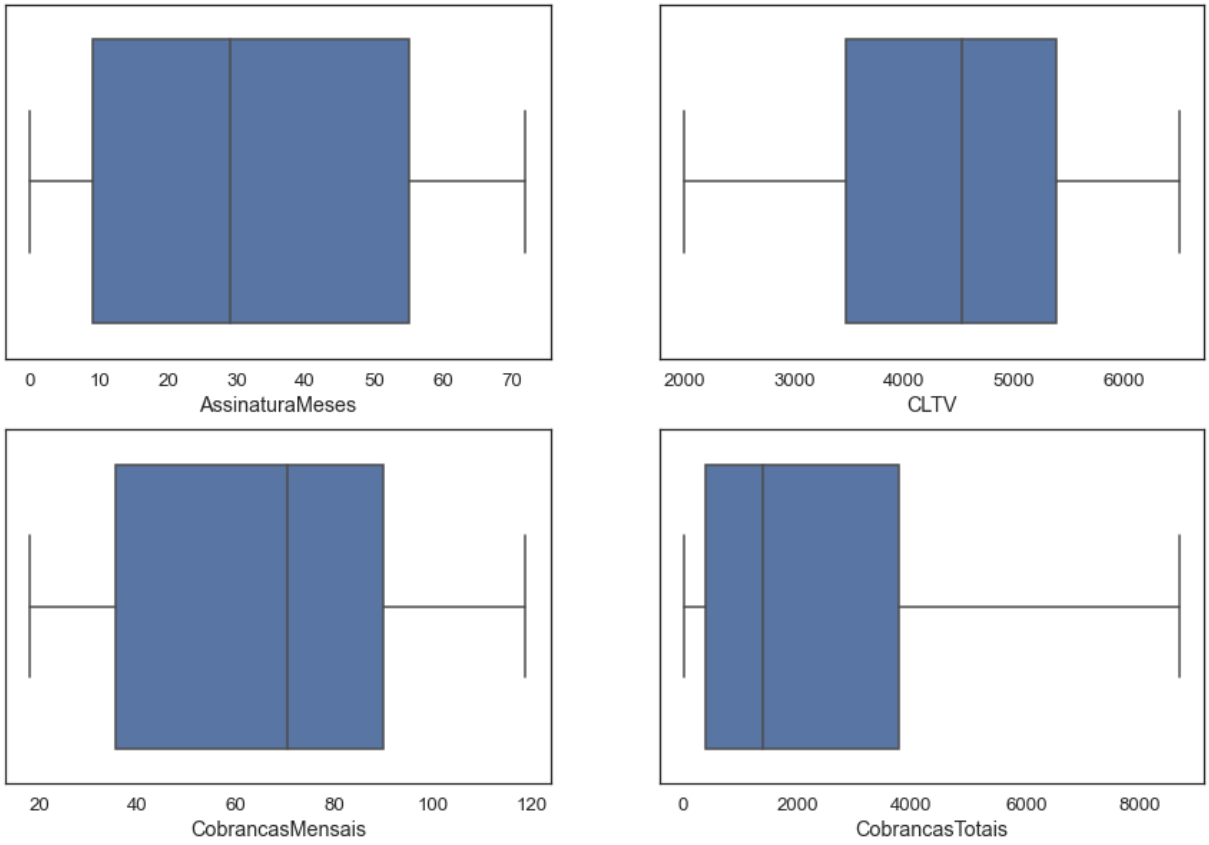

Figura 2. Gráficos Box-Plot que mostram a distribuição das variáveis numéricas para evidenciar a presença de outliers
Fonte: Dados originais da pesquisa

Para entender sobre os motivos que levaram os clientes a deixar de assinar o serviço, um gráfico foi mostrado através da Figura 3. As principais justificativas para os clientes saírem do serviço são Atitude da pessoa de apoio que representa 10,27% do total, seguido por Concorrente oferece velocidades de download mais altas com 10,11% do total, “Concorrente ofereceu maior pacote de dados” representando 8,66% do total de churn, enquanto 8,23% “não souberam informar”, seguido de “Concorrente fez oferta melhor” com 7,49%. A maioria das justificativas envolvem concorrentes que estão oferecendo serviços aprimorados para a captação dos clientes da empresa em questão.

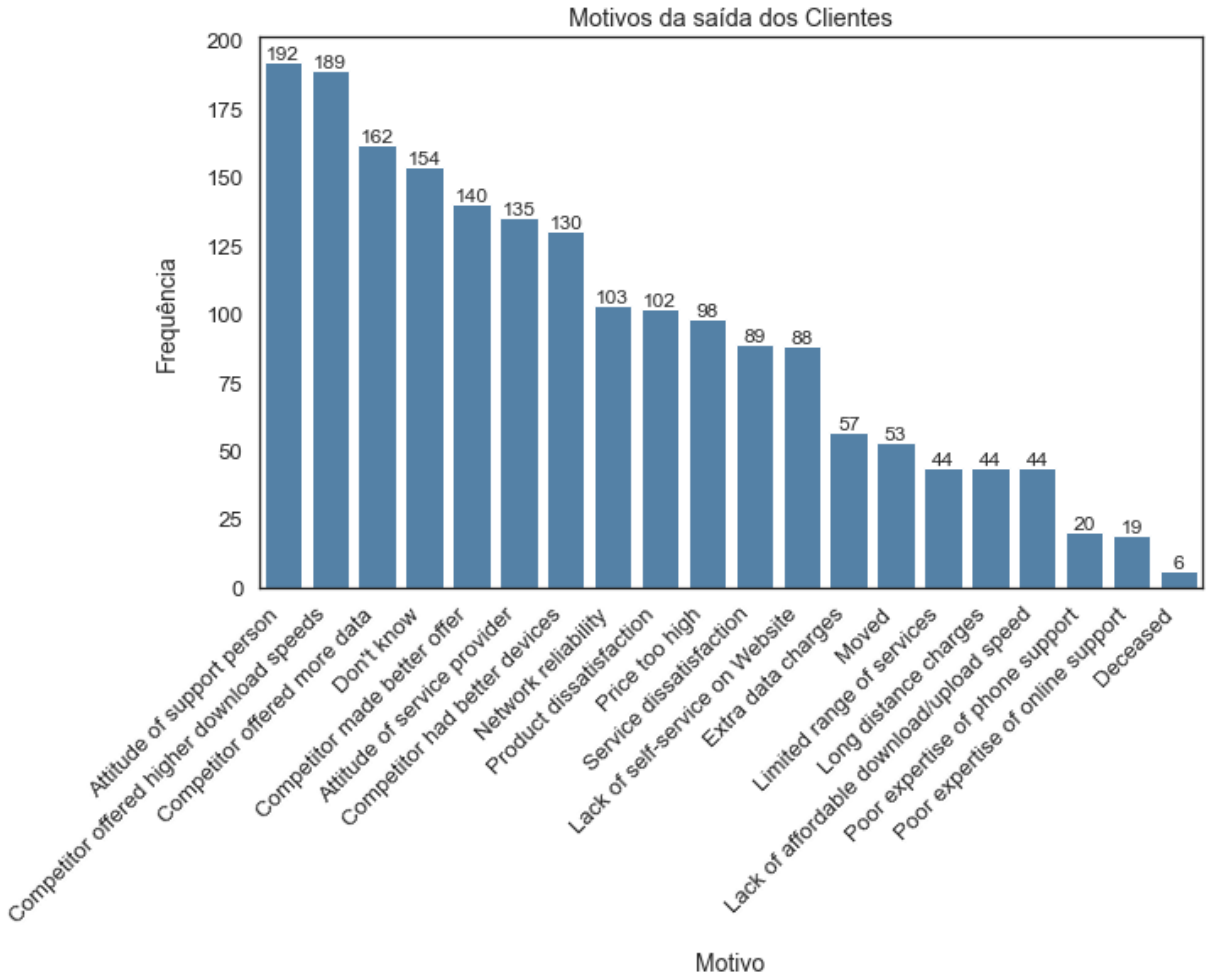

Figura 3. Motivo da saída dos Clientes do serviço
Fonte: Dados originais da pesquisa

Nesta etapa, algumas variáveis categóricas ainda necessitam ser transformadas 'Cidade', 'Genero', 'Senior', 'Parceiro', 'Dependentes', 'ServicoTelefonia', 'MultiplasLinhas', 'ServicoInternet', 'SegurancaOnline', 'BackupOnline', 'ProtecaoDispositivo', 'SuporteTech', 'StreamingTV', 'StreamingFilmes', 'TipoContrato', 'FaturaSemPapel', 'MetodoPagamento' e 'MotivoChurn', a transformação consistiu em aplicar uma técnica chamada de “dumização”, onde são gerados números binários para cada categoria.

Após a análise exploratória e preparação os dados foram separados em 80% para treinamento e 20% teste, uma boa técnica para entender se o modelo será capaz de generalizar previsões para novos dados (Géron, 2019). Após essa divisão através da função train_test_split() da biblioteca scikit-learn, os dados foram submetidos aos quatro modelos de machine learning e, em seguida, foram avaliados utilizando a métrica de acurácia, que é uma medida de desempenho utilizada para avaliar modelos de Machine Learning em tarefas de classificação. Essa métrica representa a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. Em outras palavras, a acurácia mede a habilidade do modelo

em classificar corretamente os dados. Quanto maior a acurácia, melhor é a performance do modelo.

form. (1)

$$Acurácia = \frac{TP + TN}{TP + FN + TN + FP}$$

onde, TP: é o valor “true positive”, ou seja, verdadeiro positivo quando o modelo acerta um valor que era positivo, TN: é o valor “true negative”, ou seja, verdadeiro negativo quando o modelo acerta um valor que era negativo, FP: é o valor “false positive”, ou seja, falso positivo quando o modelo classifica como positivo quando na verdade era negativo e FN: é o valor “false negative”, ou seja, falso negativo quando o modelo classifica como negativo quando na verdade era positivo.

Por exemplo, se um modelo de classificação fez 100 previsões e acertou 80 delas, então a acurácia desse modelo seria de 80/100 = 0,8 (ou 80%). A acurácia é uma medida de desempenho comum em problemas de classificação binária. No entanto, a acurácia pode ser enganosa em casos onde as classes são desbalanceadas, ou seja, uma classe possui muito mais exemplos que a outra, pois um modelo pode ter tido uma alta acurácia simplesmente prevendo a classe majoritária em todos os casos.

Adicionalmente foi utilizada outra métrica de avaliação de modelo, denominada Curva “Receiver Operating Characteristic [ROC]”. A curva ROC representa graficamente a relação entre a Taxa de Verdadeiros Positivos ou “True Posivite Rate [TPR]” e a Taxa de Falsos Positivos ou “False Positive Rate [FPR]” em diferentes pontos de corte. A TPR traz a proporção onde o modelo classificou como churn onde de fato o cliente saiu do serviço e FPR as instâncias negativas que foram classificadas incorretamente como positivas. A curva ROC é gerada traçando a TPR em função do FPR para diferentes pontos de corte no modelo de classificação binária. Quanto maior a área sob a curva ROC, melhor o desempenho do modelo, pois indica uma maior capacidade de distinguir entre as classes positiva e negativa, o valor da curva varia de 0 a 1, onde 1 indica um modelo perfeito e 0,5 um modelo aleatório. Na Figura 4 é demonstrado o fluxograma de desenvolvimento do estudo.

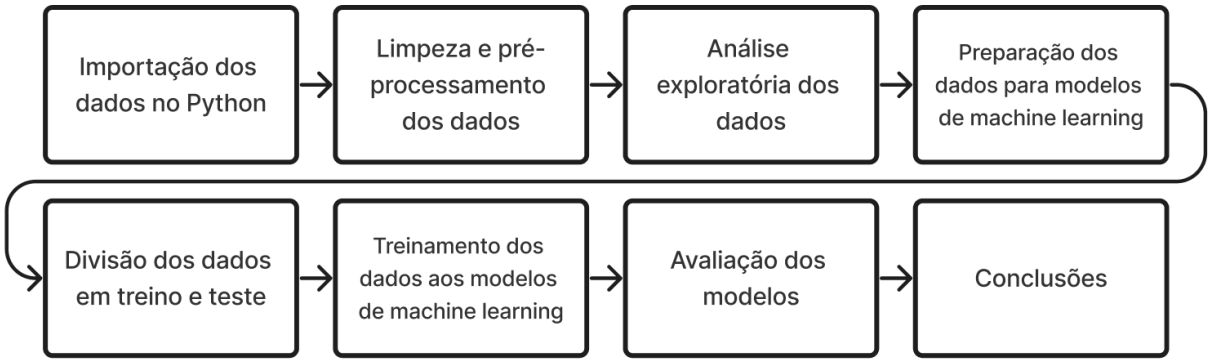

Figura 4. Fluxograma de desenvolvimento do estudo
Fonte: Dados originais da pesquisa

Resultados e Discussão

Analisando os dados, o percentual de churn está distribuído em 73,46% como não churn e 26,54% como churn, a maioria dos clientes do banco de dados não evadiram o contrato. Entre homens e mulheres, o percentual de churn é praticamente o mesmo, sendo 50% para cada gênero. Existe uma maior concentração dos clientes que optam por assinatura mensal, relacionados com a variável contrato, onde 55% dos clientes optam por pagamentos mensais, seguidos de 24% e 21% para pagamentos de dois em dois anos e anuais respectivamente. Dentro dos clientes que optam pela assinatura mensal, a maior concentração de churn vem dessa categoria de contrato. Para os clientes que evadiram, a mediana da variável relacionada a cobranças mensais é de um valor monetário de 79,65, superior ao valor dos clientes que não evadiram que é de 64,43, ou seja, os clientes churn podem estar interpretando a conta da mensalidade como sendo muito alta, e podem estar optando por outros prestadores de serviço no mercado. Isso pode ser evidenciado na Figura 5.

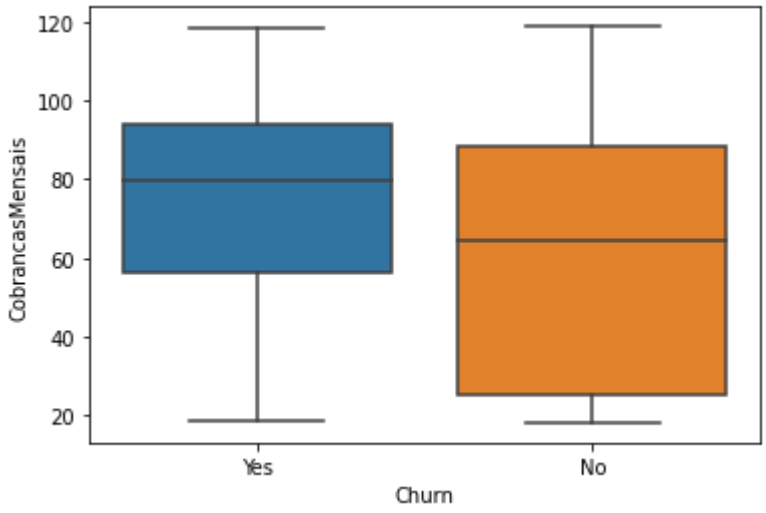

Figura 5. Gráfico boxplot da variável Cobranças Mensais
Fonte: Resultados originais da pesquisa

Os resultados obtidos mostraram que todos os modelos apresentaram acurácia acima de 85% tanto nos dados de treino quanto nos dados de teste. O resumo dos resultados pode ser observado na Tabela 3 e serão explorados logo em seguida.

Tabela 3. Resumo dos resultados

Modelo	Acurácia treino	Acurácia teste	AUC ROC
XGBoost	99,98%	99,93%	100%
Random Forest	98,70%	85,45%	91,47%
Regressão Logística	87,29%	87,37%	91,17%
Árvore de decisão	86,81%	87,30%	87,08%

Fonte: Resultados originais da pesquisa

Também é possível visualizar e comparar todas as curvas ROC para cada modelo na Figura 6 a seguir.

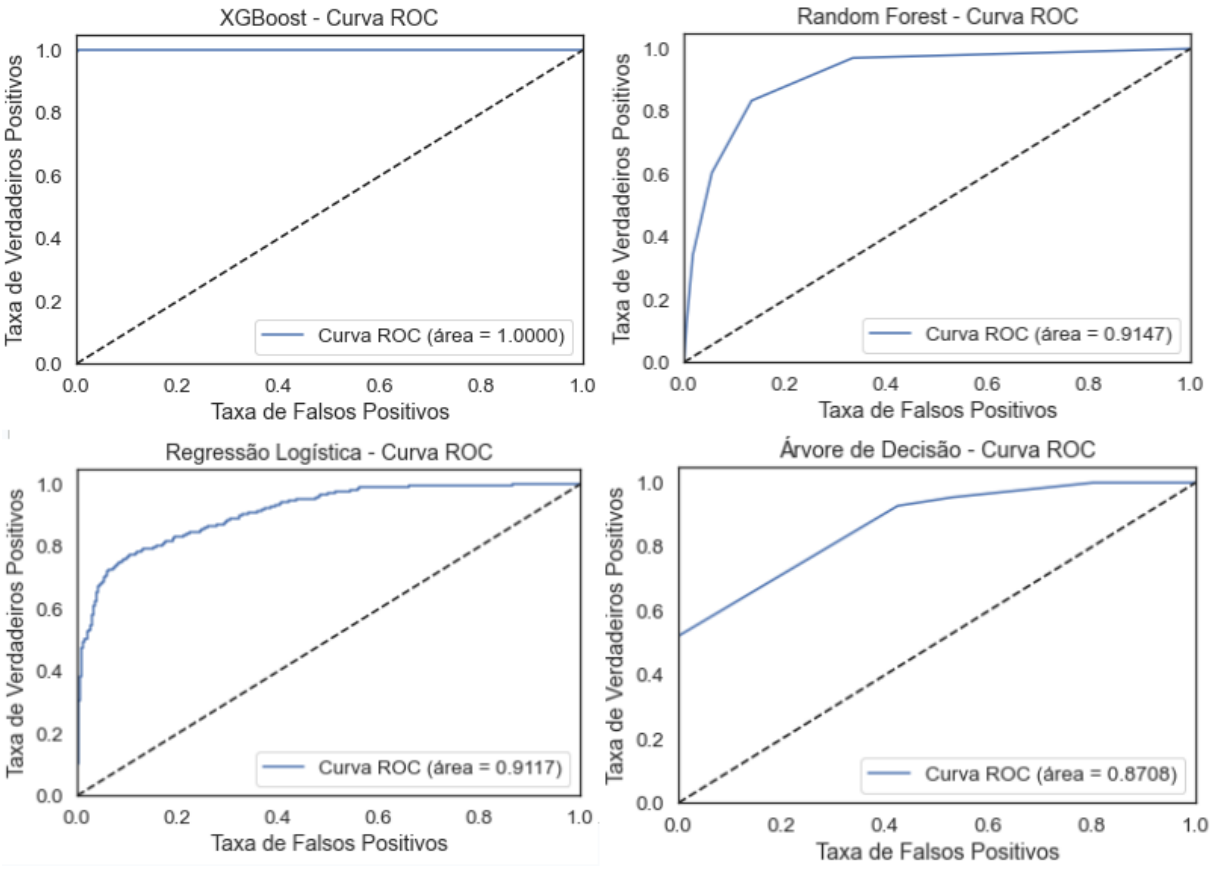

Figura 6 Resultado da métrica curva ROC
Fonte: Resultados originais da pesquisa

Adicionalmente também podemos visualizar o comparativo de todas as curvas ROC em um único gráfico, conforme a Figura 7.

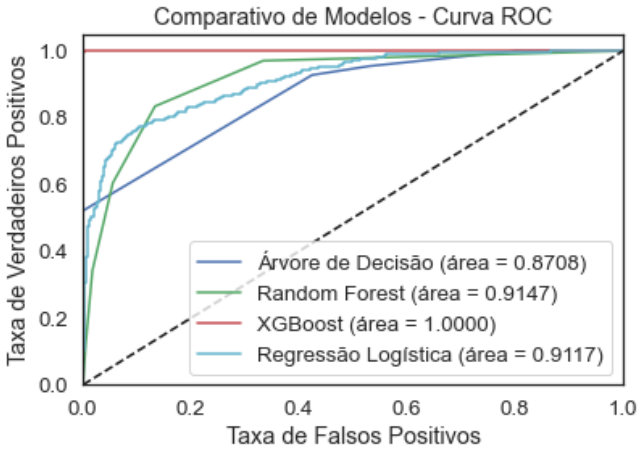

Figura 7. Comparativo entre todas as curvas ROC
Fonte: Resultados originais da pesquisa

O modelo XGBoost obteve a melhor acurácia no treino com 99,98%, 99,93% no teste e curva ROC com 100%. Isso indica que o modelo foi capaz de aprender muito bem com os dados disponíveis e obteve um desempenho excepcional na predição de churn com a base de dados utilizada nesse estudo, esse resultado impressionante pode ser explicado pela capacidade do XGBoost em lidar com grande volume de dados, essa é uma vantagem importante em situações em que há muitas variáveis envolvidas, como é o caso da análise de churn. Porém o alto desempenho pode ser um forte indício de overfitting, ou seja, quando o modelo ao invés de aprender com os dados acaba decorando e pode não acertar predições futuras com dados completamente desconhecidos.

O Modelo Random Forest teve um desempenho inferior em comparação com o modelo XGBoost, com acurácia de 98,70% nos dados de treino e 85,45% nos dados de teste, com uma curva ROC de 91,47%. Apesar de ser o segundo em relação do placar da curva ROC, esse modelo se mostrou ser capaz de realizar boas predições em dados desconhecidos.

A Regressão Logística obteve acurácia de 87,29% nos dados de treino e 87,37% nos dados de teste, seguido de uma curva ROC de 91,17%, indicando que o modelo obteve uma boa generalização, mas não tão bom quanto o Random Forest.

Por fim o modelo Árvore de Decisão apresentou o pior desempenho em relação aos demais modelos, com acurácia de 86,81% no conjunto de treino, 87,30% nos dados de teste e uma curva ROC de 87,08%.

Dessa forma, os dois melhores modelos foram o XGBoost e o Random Forest, que apresentaram as maiores acurácias nos dados de teste e também as melhores curvas ROC e podemos concluir que os modelos foram os mais eficazes na previsão de churn. As 12 principais variáveis que explicam as predições podem ser observadas na Figura 8 abaixo.

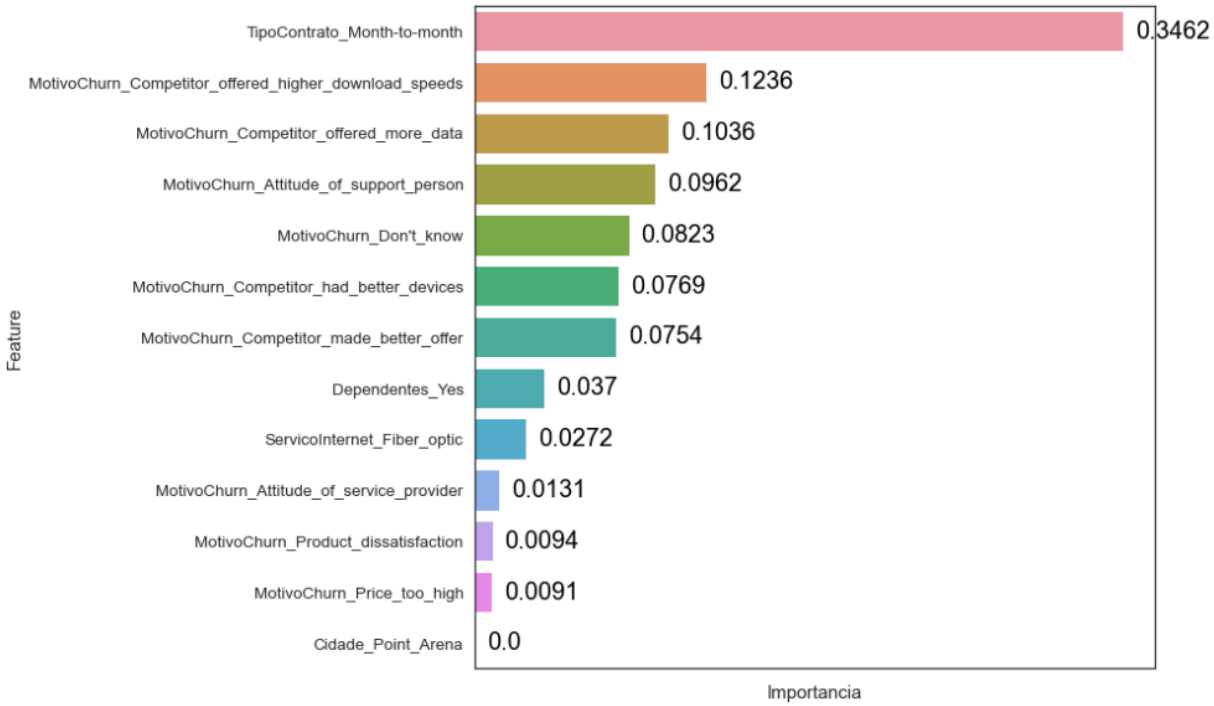

Figura 8. A importância de cada variável para a construção dos modelos de machine learning
Fonte: Resultados originais da pesquisa

Baseado nos pesos de cada importância das variáveis, é possível que a empresa se antecipe para criar ações internas com o objetivo de reter os clientes, bem como antecipar os possíveis clientes que tendem a abandonar o serviço através do uso do modelo de machine learning.

Em resumo, a utilização de modelos de machine learning mostrou-se promissora na previsão de churn em uma empresa de telecomunicações. Os modelos XGBoost e Random Forest apresentaram as melhores performances na tarefa de classificação, sendo que deve ser levado em consideração a possibilidade de overfitting do primeiro modelo. A escolha do modelo mais adequado dependerá das necessidades específicas da empresa, levando em consideração não apenas a acurácia e curva ROC, mas também outras métricas de desempenho.

Considerações Finais

É importante ressaltar que este estudo tem algumas limitações. A amostra utilizada foi relativamente pequena e a previsão do churn pode ser afetada por fatores externos, como mudanças na economia ou novos concorrentes. A base de dados é fictícia, portanto, alguns resultados podem ter sido causados meramente por falta de aleatoriedade que seria encontrada em um caso real. Além disso, o modelo não leva em consideração outras variáveis que podem afetar o churn, como a satisfação do cliente, o suporte ao cliente e economia.

Também devem ser levados em consideração os ajustes de parâmetros dos modelos, que podem melhorar consideravelmente os resultados.

Sendo assim, futuros estudos podem ser realizados para aprimorar a previsão do churn, levando em consideração outras variáveis que podem afetar o comportamento do cliente. A análise de dados em tempo real também pode ajudar a empresa a identificar os clientes mais propensos a cancelar o serviço e tomar medidas preventivas para mantê-los satisfeitos e engajados.

Os resultados deste estudo mostram que os modelos de Machine Learning são uma ferramenta valiosa para a previsão do churn em uma empresa de telecomunicações, e podem ser usados para melhorar a eficiência operacional e a satisfação do cliente.

Referências

Géron, A. 2019. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2ed. O'Reilly Media. Sebastopol, CA, USA.

Gold, C. 2020. Fighting Churn with Data: The Science and Strategy of Customer Retention. 1ed. Manning. Shelter Island, NY, USA.

Grus J. 2019. Data Science from Scratch: First Principles with Python. 2ed. O'Reilly Media. Sebastopol, CA, USA.

James, G., Witten, D., Hastie, T., Tibshirani, R. 2021. An Introduction to Statistical Learning: with Applications in R. 2ed. Springer Science & Business Media. New York, NY, USA.

Kurtz, D. L.; Clow, K. E. 1998. Services Marketing. 1ed. John Willey & Sons. New York, NY, USA.

Müller, C. A.; Guido, S. 2016. Introduction to Machine Learning with Python: A Guide for Data Scientists. 1ed. O'Reilly Media. Sebastopol, CA, USA.