PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação Lato Sensu em Ciência de Dados e Big Data

José Nunes da Silva Junior

ANÁLISE CHURN RATE TELECOM

Belo Horizonte

2023

José Nunes da Silva Junior

ANÁLISE DE CHURN RATE TELECOM

Trabalho de Conclusão de Curso apresentado

ao Curso de Especialização em Ciência de

Dados e Big Data como requisito parcial à

obtenção do título de especialista.

Belo Horizonte

2023

Sumário

[1 INTRODUÇÃO 4](#_Toc150200652)

[1.1 Contextualização 4](#_Toc150200653)

[1.2 O problema proposto 4](#_Toc150200654)

[2 Coleta de Dados 5](#_Toc150200655)

[3 Tratamento de Dados 9](#_Toc150200656)

[4 Analise e exploração dos dados 22](#_Toc150200657)

# INTRODUÇÃO

## Contextualização

Nos últimos anos, o setor de telecomunicações tem observado um enorme avanço tecnológico, avanço este, que possibilitou novas oportunidades de negócios para as empresas. A facilidade de contratar determinados serviços com poucos cliques também é encontrada ao se tentar cancelar este mesmo serviço. Dada a crescente competição no mercado bem como o aumento da demanda dos consumidores por serviços mais personalizados e acessíveis, um elemento vem se tornando o foco das atenções: o *churn rate*.

O “*churn rate*”, ou taxa de rotatividade de clientes, é uma métrica crucial para empresas no ramo de telecomunicações, diferentemente de um modelo de compra e venda onde se obtêm o lucro no momento em que o processo se concretiza, o modelo de negócios de telecomunicações se caracterizam por um investimento inicial feito pela empresa, seja em infraestrutura, equipamentos e/ou instalação na residência do cliente. Nesse modelo de negócio, o cliente só passa a dar lucro meses ou até mesmo anos depois de contratar o serviço.

Diante disso, tornou-se fundamental para as empresas o foco nesse indicador, uma vez que a retenção de clientes é crucial para o sucesso e a sustentabilidade dos negócios.

## O problema proposto

Devido as características de investimento inicial alto e lucro futuro, a presente analise priorizara no RECALL, uma vez que devemos buscar minimizar o maximo possivel de falsos negativos, onde o modelo determina que o cliente não vai cancelar porem o cliente cancela

# Coleta de Dados

O conjunto de dados *Telecom Customer Churn Prediction*, objeto de análise deste projeto pode ser encontrado através do sítio da *Maven Analytics*, uma plataforma de desafios relacionados a analises de dados.

Figura 1 - Fonte da base de dados que será utilizada.



Fonte: .

Na Figura 1 é possível observar os nomes e propriedades dos arquivos que serão utilizados. O conjunto em questão consiste em 3 (três) *datasets* que possuem informações e finalidades distintas (Figura 2): *telecom\_customer\_churn*, que contêm a base principal dos dados que serão analisados neste artigo; *telecom\_zipcode\_population*, que possui dados adicionais relacionados a população por código postal e; *telecom\_data\_dictionary*, que como o próprio nome sugere, trata-se de um dicionário que contém as informações do que cada coluna representa em nossa base principal.

Figura 2 - *Datasets* que serão utilizados



Fonte: Autoria própria.

Para uma maior facilidade de interpretação dos leitores deste artigo, os nomes das colunas, originalmente em inglês serão traduzidos para o português na etapa de tratamento de dados. Na Tabela 1 está descrito de forma detalhada as informações que cada coluna representa no arquivo *telecom\_customer\_churn*.

Tabela 1 – Detalhamento do *dataset* *telecom\_customer\_churn*

|  |  |  |  |
| --- | --- | --- | --- |
| **Nome Original** | **Nome traduzido** | **Descrição** | **Tipo** |
| CustomerID | ID\_cliente | ID única que identifica cada cliente | Texto |
| Gender | Genero | Genero do cliente | Texto |
| Age | Idade | Idade do cliente | Inteiro |
| Married | Casado | Indica se o cliente é casado | Texto |
| Number of Dependents | Qtd\_dependentes | Indica a quantidade de dependentes que moram com o cliente | Inteiro |
| City | Cidade | Cidade em que se localiza a residência do cliente | Texto |
| Zip Code | Codigo\_postal | Código postal da residência do cliente | Inteiro |
| Latitude | Latitude | Latitude da residência do cliente | Real |
| Longitude | Longitude | Longitude da residência do cliente | Real |
| Number of Referrals | Qtd\_indicações | Quantidade de indicações que o cliente realizou a amigos ou parentes | Inteiro |
| Tenure in Months | Meses\_na\_base | Indica o total de meses que o cliente permaneceu na empresa | Inteiro |
| Offer | Ultima\_oferta | Indica a última oferta que o cliente aceitou | Texto |
| Phone Service | Servico\_telefone | Indica se o cliente possui o serviço telefônico da empresa | Texto |
| Avg Monthly Long Distance Charges | Recarga\_longa\_distancia | Indica o custo médio total de recargas de longa distancia | Real |
| Multiple Lines | Multiplas\_linhas | Indica se o cliente possui mais de uma linha telefônica | Texto |
| Internet Service | Servico\_internet | Indica se o cliente possui o serviço de internet da empresa | Texto |
| Internet Type | Tipo\_internet | Indica o tipo de conexão que o cliente utiliza | Texto |
| Avg Monthly GB Download | Media\_mensal\_download\_GB | Indica o volume médio de download do cliente em gigabytes | Real |
| Online Security | Servico\_seguranca | Indica se o cliente possui um serviço adicional de segurança provido pela empresa | Texto |
| Online Backup | Servico\_backup | Indica se o cliente possui um serviço adicional de *backup* provido pela empresa | Texto |
| Device Protection Plan | Servico\_protecao\_dispositivo | Indica se o cliente possui um serviço adicional de proteção de dispositivo | Texto |
| Premium Tech Support | Servico\_suporte\_preferencial | Indica se o cliente possui um serviço adicional de atendimento preferencial | Texto |
| Streaming TV | Servico\_tv | Indica se o cliente possui algum serviço de *streaming* de TV | Texto |
| Streaming Movies | Servico\_filmes | Indica se o cliente possui algum serviço de *streaming* de filmes | Texto |
| Streaming Music | Servico\_musica | Indica se o cliente possui algum serviço de *streaming* de músicas | Texto |
| Unlimited Data | Servico\_ilimitado\_dados | Indica se o cliente possui algum serviço de dados ilimitados | Texto |
| Contract | Tipo\_contrato | Indica o tipo de contrato ao qual o cliente está fidelizado | Texto |
| Paperless Billing | Faturamento\_sem\_papel | Indica se o cliente escolheu pelo tipo de cobrança sem papel | Texto |
| Payment Method | Tipo\_pagamento | Indica o método de pagamento escolhido pelo cliente | Texto |
| Monthly Charge | Cobranca\_mensal | Indica o total mensal pago pelo cliente por todos os serviços | Real |
| Total Charges | Total\_cobranca | Indica a soma total paga pelo cliente em todo o tempo na empresa | Real |
| Total Refunds | Total\_reembolsos | Indica a soma total de reembolso recebida pelo cliente em todo o tempo na empresa | Real |
| Total Extra Data Charges | Qtd\_dados\_extras | Indica a quantidade total de dados extras contratados pelo cliente | Inteiro |
| Total Long Distance Charges | Cobrancas\_longa\_distancia | Indica a soma total paga pelo cliente em todo o tempo na empresa referente a recargas para longas distancias | Real |
| Total Revenue | Total\_cobranca\_geral | Indica a soma total paga pelo cliente de forma geral somando o valor mensal com todas as cobranças de recargas extras | Real |
| Customer Status | Situacao\_cliente | Indica o status do cliente ao final período em que os dados foram coletados: *Churned* (cancelou), *Stayed* (permaneceu cliente) e *Joined* (se tornou cliente) | Texto |
| Churn Category | Categoria\_cancelamento | Indica uma categoria que classifica o motivo de cancelamento do cliente | Texto |
| Churn Reason | Motivo\_cancelamento | Indica o motivo do cancelamento do cliente | Texto |

Fonte:

O mesmo processo será realizado no arquivo *telecom\_zipcode\_population*, como pode ser observado na Tabela 2.

Tabela 2 - Detalhamento do *dataset* *telecom\_data\_dictionary*

|  |  |  |  |
| --- | --- | --- | --- |
| **Nome Original** | **Nome traduzido** | **Descrição** | **Tipo** |
| Zip Code | Codigo\_postal |  |  |
| Population |  |  |  |

Fonte:

# Tratamento de Dados

Para a execução do projeto, foi utilizada a linguagem de programação interpretada Python, na sua versão 3.9.13, além do *Jupyter Notebook*, ambiente de desenvolvimento interativo baseado na Web

Nesta etapa do projeto, será utilizada majoritariamente a biblioteca *Pandas* (Figura 3) para manejo dos dados bem como de eventuais ajustes que possam ser necessários para deixar a base de dados melhor estruturada, e com isso, buscar melhores interpretações dos algoritmos de *Machine Learning*.

Figura 3 - Importação do *pandas*



Fonte: Autoria própria

Foi utilizado para a leitura dos dados, uma função própria do *pandas*, a *pd.read\_csv* (Figura 4), através dela, é possível ler um arquivo *csv* e já transforma-lo em um *dataframe* que nada mais é do que uma estrutura de dados composta por duas dimensões: linhas e colunas, similar a uma planilha.

Figura 4 - Leitura do arquivo com o pandas



Fonte: Autoria própria

Em seguida, por meio do comando *display* é possível visualizar o *dataframe* criado (Figura 5)

Figura 5 - Visualização do dataframe



Fonte: Autoria própria

Um detalhe importante sobre a função display é que ela por padrão, exibe no máximo 20 colunas, para que ela exiba todas as colunas e assim conseguirmos ter uma visão geral do nosso dataframe iremos realizar uma configuração através do comando *pd.set\_option* e alterando o parâmetro *display.max\_columns* como mostrado na Figura 6.

Figura 6 - Ajuste no maximo de colunas da função display



Fonte: Autoria própria

Dando continuidade agora realizaremos a leitura e criação do dataframe de população por código postal (Figura 7), esses dados serão agregados ao nosso dataframe principal com o objetivo de enriquecer nossa base de dados, afim de encontrar algum padrão com a quantidade de pessoas de uma determinada região, o código usado para realizar essa junção está representado na Figura 8.

Figura 7 - Criando dataframe da população por código postal



Fonte: Autoria própria

Figura 8 - Junção dos dataframes



Fonte: Autoria própria

Agora temos o nosso *dataframe* consolidado e pronto para ser realizado os devidos ajustes.

Como informado no tópico 2 deste artigo, iremos realizar o ajuste dos nomes das colunas do nosso *dataframe*, traduzindo e alterando para um padrão onde a primeira letra é maiúscula e as palavras são separadas pelo caractere underline (\_). O comando utilizado para realizar tal procedimento pode ser visualizado na Figura 9.

Figura 9 – Tradução dos nomes das colunas



Fonte: Autoria própria

Também será realizada a tradução dos valores contidos em todas as colunas de texto da nossa base de dados com exceção das colunas Cidade e ID\_cliente. No que diz respeito coluna Cidade, não se faz necessário a tradução desses valores por se tratarem de nomes de cidades, já a coluna ID\_cliente, possui apenas valores que identificam o cliente dentro da empresa, para nós, entretanto, trata-se de uma informação sem valor, dessa forma esta coluna será removida da nossa base de dados. Na Figura 10 está exposto o comando utilizado para remover a coluna ID\_cliente.

Figura 10 - Removendo a coluna ID\_cliente



Fonte: Autoria própria

Já na Figura 11 iremos realizar um comando para nos retornar os valores únicos contidos em cada coluna de texto, tornando mais fácil a identificação dos valores que irão necessitar ser traduzidos.

Figura 11 - Identificando os valores que precisarão ser traduzidos



Fonte: Autoria própria

A partir da identificação desses valores, podemos agora realizar a sua tradução com o comando *replace* (Figura 12).

Figura 12 - Renomeando valores da base de dados



Fonte: Autoria própria

Dando continuidade ao nosso processo de tratamento dos dados, iniciaremos agora algumas analises iniciais em cima da nossa base de dados, com o comando *df.info()* (Figura 13) iremos visualizar algumas informações sobre o nosso *dataframe*.

Figura 13 – Informações do *dataframe*



Fonte: Autoria própria

É possível observar que em nosso *dataframe*, algumas colunas possuem valores faltantes, em uma breve análise, nota-se que essas colunas são condicionadas ao valor contido em outra coluna, por exemplo, o valor presente na coluna Tipo\_internet está relacionado ao fato do cliente possuir ou não o serviço de internet (coluna Servico\_internet).

O mesmo ocorre com colunas que dependem do valor contido na coluna Servico\_telefone, que tem informações se o cliente possui ou não o serviço de telefone. Na Figura 14 foi utilizado o comando *groupby* para agrupar os dados contidos nas colunas Servico\_internet e Servico\_telefone e o comando *count()* para que o agrupamento seja feito realizando a contagem dos valores.

Figura 14 - Contagem dos valores das colunas



Fonte: Autoria própria

Se visualizarmos os valores gerados na saída do comando na Figura 14, é possível observar que a quantidade de clientes que não possuem o serviço de telefone é a mesma quantidade de clientes com valores faltantes nas colunas que detalham especificamente o serviço de telefone (Recarga\_longa\_distancia e Multiplas\_linhas), o mesmo ocorre com clientes que não possuem o serviço de internet e as colunas que estratificam esse serviço (Tipo\_internet, Media\_mensal\_download\_GB, Servico\_seguranca, Servico\_backup, Servico\_protecao\_dispositivo, Servico\_suporte\_preferencial, Servico\_tv, Servico\_filmes, Servico\_musica e Servico\_ilimitado\_dados), isso pode ser melhor observado na Figura 15, onde o comando realiza um filtro na base de dados que retorna a quantidade de valores presentes em todas as linhas onde o cliente não possui o serviço de telefone, na Figura 16, o mesmo processo é realizado nos clientes que não possuem o serviço de internet. Em ambos os casos, as colunas que detalham o serviço aparecem sem nenhum valor preenchido.

Figura 15 - Filtro de clientes sem o serviço de telefone



Fonte: Autoria própria

Figura 16 - Filtro de clientes sem o serviço de internet



Fonte: Autoria própria

A falta de campos preenchidos se torna um problema para a nossa análise dos dados bem como para a utilização de algoritmos de *machine learning* uma vez que isso pode acarretar em uma perda de contexto e, consequentemente, em uma análise incompleta e enviesada. Além disso, a presença de valores ausentes pode afetar a qualidade dos modelos de aprendizado de máquina, visto que muitos algoritmos não podem lidar diretamente com lacunas nos dados. Em algumas situações, pode ser considerada a remoção completa da linha que contém o dado faltante como uma solução, este não é o nosso caso, além de resultar numa perda quantitativa de dados, estaríamos excluindo da nossa base todos os clientes que não possuem os serviços de telefone e os serviços de internet.

Antes de escolhermos uma estratégia para solucionar nosso problema, é necessário entendermos os dados já preenchidos nessas colunas e seu contexto. Com exceção das colunas Recarga\_longa\_distancia e Media\_mensal\_download\_GB que possuem valores numéricos que informam o custo médio total de recargas de longa distância e o volume médio de download do cliente em gigabytes, respectivamente, as outras colunas possuem os valores binários “Não” e “Sim”

Para resolver esta situação estaremos atribuindo o valor “Não” nas colunas que possuem valores binários e o valor “0” nas duas colunas que possuem valores numéricos. A justificativa para essa escolha é que, uma vez que o cliente não possua o serviço de internet, por consequência ele não irá ter um serviço de filmes ou musica por exemplo, e isso pode ser replicado para as outras colunas que detalham esses dois serviços, a Figura 17 mostra o comando realizado para imputar estes dados nas respectivas colunas.

Figura 17 - atribuindo dados nos campos com valores nulos



Fonte: Autoria própria

Continuando com nossa análise, agora temos apenas duas colunas com dados ausentes, que são Categoria\_cancelamento e Motivo\_cancelamento. A própria nomenclatura dessas colunas sugere que estão relacionadas à decisão do cliente de cancelar o serviço ou não. Essa informação pode ser encontrada na coluna Situacao\_cliente. No entanto, antes de prosseguirmos com o processo de correção de dados faltantes, é necessário fazer um ajuste. Conforme indicado na Tabela 1, a coluna Situacao\_cliente contém três valores distintos: “Cancelou”, “Continua cliente” e “Se tornou cliente”

Dado que nosso principal objetivo é analisar o churn rate e tentar prever quais clientes têm maior propensão a solicitar o cancelamento, os dados relacionados aos clientes que aderiram ao serviço se tornam um elemento de interferência em nossa base de dados. Isso ocorre porque, quando se trata de análise dos cancelamentos, estamos particularmente interessados em analisar o histórico do cliente e em detectar padrões que o levaram a solicitar o termino da prestação do serviço. No caso dos recém-contratantes, simplesmente não dispomos de um período de observação suficiente para discernir esses padrões, uma vez que não tiveram tempo de estabelecer um histórico de interações significativo com o serviço.

Na Figura 18 podemos observar a quantidade de clientes por sua situação no momento que os dados foram coletados.

Figura 18 - Quantidade de clientes por situação



Fonte: Autoria própria

O comando realizado para remover estes clientes pode ser visualizado na Figura 19.

Figura 19 - Removendo clientes que contrataram o serviço no período analisado



Fonte: Autoria própria

Retornando agora para as colunas Categoria\_cancelamento e Motivo\_cancelamento, podemos realizar um agrupamento contabilizando os valores contidos nessa coluna para entender melhor os seus dados, o comando utilizando está contido na Figura 20.

Figura 20 – Agrupamento e contagem dos valores das colunas



Fonte: Autoria própria

Com base nas informações contidas no arquivo *telecom\_data\_dictionary* é possível inferir que os valores contidos na coluna Motivo\_cancelamento são uma estratificação dos valores contidos na coluna Categoria\_cancelamento, isto pode ser comprovado com a saída do comando mostrado na Figura 21, neste comando também foi realizado o agrupamento por Situacao\_cliente, comprovando que somente os clientes que cancelaram possuem dados nestas colunas.

Figura 21 - Agrupamento em conjunto das colunas



Fonte: Autoria própria

Uma vez que os valores em branco nas colunas mencionadas anteriormente são referentes aos clientes que permaneceram com o serviço, estaremos atribuindo o valor “Permaneceu cliente” em ambas as colunas (Figura 22), pois assim teremos todos os valores preenchidos sem nenhuma perca de contexto ou risco de enviesar a nossa base dados já que o cliente de fato permaneceu com o serviço.

Figura 22 - Atribuição dos valores nas linhas com dados ausentes



Fonte: Autoria própria

Finalizando esta etapa, ao executar o comando df.count() (Figura 23), sua saída nos mostra uma base de dados com 6589 registros em que todas as colunas estão devidamente ajustadas e sem valores faltantes.

Figura 23 - Contagem dos valores por coluna



Fonte: Autoria própria

# Analise e exploração dos dados

Partindo agora para a etapa de análise exploratória dos dados iremos avançar na investigação detalhada do nosso dataframe com o objetivo de identificar padrões, tendências e fatores que moldam a decisão dos clientes de permanecerem ou cancelaram seus serviços. Esta análise aprofundada permitirá a formulação de estratégias de retenção de clientes mais eficazes e aprimoramento da qualidade dos serviços oferecidos pelas empresas de telecomunicações.

Primeiramente, iremos importar a biblioteca *matplotlib*, ela será fundamental nesta etapa pois através dela conseguiremos construir gráficos que facilitarão tanto nossas analises quanto as tomadas de decisões, o comando utilizado para importar esta biblioteca pode ser visualizado na Figura 24.

Figura - Importando o Matplotlib



Fonte: Autoria própria

Com a biblioteca importada, podemos iniciar a construção dos nossos gráficos, para iniciar podemos visualizar o percentual de clientes por gênero em nossa base, o código utilizado está registrado na Figura 25. Já na Figura 26 temos o gráfico gerado pelo código em questão.

Figura - Código para geração do gráfico de percentual por gênero



Fonte: Autoria própria

Figura - Gráfico percentual por gênero



Fonte: Autoria própria

É possível observar que possuímos uma base de dados muito equilibrada nesse aspecto. Se cruzarmos esse dado com outras colunas, podemos estratificar ainda mais essa informação, na Figura 27 por exemplo, utilizaremos um código que irá criar faixas etárias a partir da coluna “Idade”, a fim de agrupar os dados em categorias mais amplas. Esse procedimento simplificará a análise, permitindo-nos identificar padrões de maneira mais eficaz.

Figura - Criação de faixas etárias



Fonte: Autoria própria

Posteriormente, exploraremos a composição de gênero dos clientes dentro de cada faixa etária, permitindo-nos visualizar essas informações de maneira mais representativa por meio de um gráfico de pirâmide. Na Figura 28 podemos visualizar o código utilizado para criação gráfico e na Figura 29 o gráfico em si.

Figura - Código para criação do gráfico de pirâmide do gênero do cliente



Fonte: Autoria própria

Figura - Gráfico de gênero por idade



Fonte: Autoria própria

Neste nível de estratificação conseguimos observar que as faixas etárias de 66 até 80 anos possuem uma quantidade de clientes menor em relação as outras. Iremos agora utilizar esta mesma estrutura de faixas etárias e estratificar a coluna “Situacao\_cliente”, o código para criar este novo gráfico pode ser visualizado na Figura 30.

Figura - Código para criação do gráfico de pirâmide da situação do cliente



Fonte: Autoria própria

Figura - Gráfico de pirâmide da situação do cliente



Fonte: Autoria própria

Na Figura 31 temos o gráfico gerado, nele é possível observar que temos uma quantidade bem menor de clientes que cancelaram em relação aos que continuam cliente, algo que já tínhamos observado anteriormente (Figura 18), porém, importante nos atentarmos que o volume referente as faixas de 66 a 80 anos diminuíram pouco em relação as outras, o que pode indicar uma maior tendência desse público em cancelar o serviço.

Para melhor visualizarmos essa hipótese, iremos realizar verificar a proporção de clientes que cancelam por faixa etária. Primeiramente iremos identificar qual o valor que corresponde a proporção geral da nossa base (Figura 32).

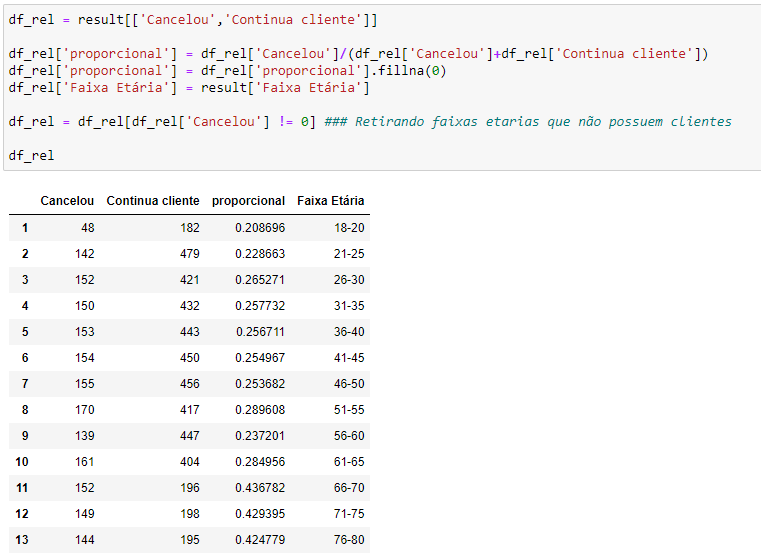
Figura - Encontrando o valor proporcional de clientes que cancelaram



Fonte: Autoria própria

Em seguida realizaremos o mesmo procedimento para todos os valores no dataframe com as faixas etárias (Figura 33).

Figura - Calculando valor proporcional por faixa etária



Fonte: Autoria própria

Com isso podemos então criar um gráfico comparando o proporcional de cancelamento de cada faixa etária com o proporcional de cancelamento geral da nossa base de dados (Figura 34).

Figura - Gráfico de proporção de cancelamento por faixa etária



Fonte: Autoria própria

Através do gráfico, fica bem perceptivo o quanto o proporcional das últimas três faixas etárias são bem mais altas em relação as outras, chegando a valores acima de 40%

BIBLIOGRAFIA

O que são dataframes? <https://www.databricks.com/br/glossary/what-are-dataframes>. Acessado 08/10/2023