UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

VICTOR LUKENCHUKII

ANÁLISE PREDITIVA DE CHURN COM REGRESSÃO LOGÍSTICA EM SERVIÇOS DE PLANEJAMENTO FINANCEIRO

Revisado pelo Orientador

Documento assinado digitalmente
gov.br
KLAUS SCHLUNZEN JUNIOR
Data: 17/12/2025 16:47:05-0300
Verifique em https://validar.iti.gov.br

Assinatura do Orientador

Data ____/ ____ / 2025

PRESIDENTE PRUDENTE
2025

VICTOR LUKENCHUKII

ANÁLISE PREDITIVA DE CHURN COM REGRESSÃO LOGÍSTICA EM SERVIÇOS DE PLANEJAMENTO FINANCEIRO

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientador(a): Prof. Dr. Klaus Schlunzen Junior

PRESIDENTE PRUDENTE
2025

L954a

Lukenchukii, Victor

ANÁLISE PREDITIVA DE CHURN COM REGRESSÃO LOGÍSTICA EM SERVIÇOS DE PLANEJAMENTO FINANCEIRO / Victor Lukenchukii. -- Presidente Prudente, 2025

78 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Klaus Schlunzen Junior

1. Estatística. 2. Análise de regressão. 3. Logistic regression analysis. 4. Serviço ao cliente. 5. Customer relations. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

VICTOR LUKENCHUKII

ANÁLISE PREDITIVA DE CHURN COM REGRESSÃO LOGÍSTICA EM SERVIÇOS DE PLANEJAMENTO FINANCEIRO

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

gov.br Documento assinado digitalmente
KLAUS SCHLUNZEN JUNIOR
Data: 17/12/2025 16:48:25-0300
Verifique em https://validar.iti.gov.br

Orientador: _____

Prof. Dr. Klaus Schlunzen Junior

Departamento de Estatística

_____

Prof. Dr. Sérgio Minoru Oikawa

Departamento de Estatística

Presidente Prudente, 16 de dezembro de 2025.

RESUMO

Este trabalho teve como objetivo a construção de um modelo preditivo para a previsão de churn em clientes de um serviço de mentoria, utilizando técnicas de engenharia de dados, análise de agrupamento (K-means) e regressão logística. Inicialmente, foi realizada uma análise exploratória dos dados originais, esta análise sugeriu um baixo poder discriminativo das variáveis entre clientes ativos e churn. A partir disso, foi realizada a engenharia de dados, criando novas variáveis mais informativas, como tempo de permanência na mentoria, frequência de participação nas reuniões, e percentual de pagamento da mentoria.

Com a base de dados ajustada, a análise de agrupamento foi realizada utilizando o K-means, segmentando os clientes em três clusters, com o objetivo de identificar padrões comportamentais distintos entre eles. O modelo de regressão logística foi ajustado com as variáveis principais e interações entre elas, obtendo resultados significativos, incluindo um AUC de 0.9522 e uma acurácia de 81,82%. A análise de variáveis indicou que o tempo de permanência na mentoria e o nível de engajamento nas reuniões foram os fatores mais relevantes para prever o churn dos clientes.

As implicações práticas deste estudo são significativas, fornecendo à empresa insights valiosos para otimizar suas estratégias de retenção de clientes. Além disso, as contribuições científicas deste trabalho incluem a aplicação de técnicas de engenharia de dados e análise de agrupamento para a análise de churn, uma aplicação pouco documentada na literatura para serviços de mentoria.

Palavras-chave: churn, engenharia de dados, regressão logística, K-means, segmentação de clientes.

ABSTRACT

This study built a predictive model for churn prediction in clients of a mentorship service, using data engineering, clustering analysis (K-means), and logistic regression techniques. Initially, an exploratory data analysis was conducted on the raw data, revealing that the available variables did not provide sufficient discriminative power between active clients and churned ones. Subsequently, data engineering was applied to create more informative variables such as tenure in mentorship, meeting participation frequency, and fraction of mentorship paid.

With the adjusted dataset, clustering analysis was performed using K-means, segmenting clients into three clusters to identify distinct behavioral patterns. The logistic regression model was fitted with the main variables and their interactions, achieving significant results, including an AUC of 0.9522 and accuracy of 81.82%. The variable analysis indicated that tenure in mentorship and engagement in meetings were the most important factors in predicting churn.

The practical implications of this study are significant, providing the company with valuable insights to optimize its customer retention strategies. Furthermore, the scientific contributions of this work include the application of data engineering and clustering analysis for churn prediction, a method still underexplored in mentorship services.

Keywords: churn, data engineering, logistic regression, K-means, customer segmentation.

Sumário

1 INTRODUÇÃO

Com o avanço das tecnologias digitais, aliado ao aumento da competitividade no mercado, organizações têm intensificado estratégias baseadas em análises de dados para aprimorar o atendimento e retenção de clientes (GABRIELLE, 2022). Nesse cenário, a busca pela excelência operacional e pela fidelização torna-se um elemento central para a sustentabilidade dos negócios.

A área de *Customer Success* (CS) desempenha papel estratégico nesse contexto, por orientar práticas que visam maximizar a satisfação, o engajamento e a longevidade do cliente ao longo do ciclo de vida do serviço. Como explica Gabrielle (2022), esse setor é

> "orientado por métricas que visam a saúde do cliente, focado em retenção e fidelização dos mesmos, o que prolonga sua vida útil junto à organização e reduz, por consequência, as taxas de evasão e os custos de aquisição de clientes" (GABRIELLE, 2022).

Assim, uma das funções centrais do Customer Success é mitigar cancelamentos, conhecidos como *churn*, por meio da compreensão aprofundada das necessidades do cliente e da identificação de sinais precoces de risco de evasão.

Nesse contexto, a estatística emerge como ferramenta essencial para a análise sistemática do comportamento do cliente. Métodos estatísticos possibilitam identificar padrões de uso, variáveis críticas associadas ao cancelamento e perfis de risco, subsidiando decisões estratégicas baseadas em evidências. Segundo Steinman,

> "o sucesso do cliente não é uma proposta de tamanho único e está evoluindo no mesmo ritmo acelerado da tecnologia que lhe serve de base. Com efeito, hoje, usamos as tecnologias de *Data Science* – como análise de *big data* e *business intelligence* – para acelerar o *time-to-value* e, em última instância, o sucesso"(STEINMAN, 2017).

O presente trabalho consiste em um estudo de caso desenvolvido na empresa Grupo Santinoni, que oferece serviços de planejamento financeiro voltados ao público médico. O serviço principal é estruturado como um programa de mentoria com duração de um ano, no qual os clientes têm acesso a encontros mensais com especialistas,

materiais de apoio e possíveis produtos adicionais complementares oferecidos ao cliente, como seguros de vida, previdência, responsabilidade civil e entre outros, destinados a ampliar a proteção financeira. Quando o cliente interrompe o serviço antes do término contratado, caracteriza-se um evento de *churn*.

O objetivo central deste estudo é desenvolver um modelo preditivo de churn, entendido como “o número de clientes que uma empresa perde em um determinado período de tempo” (ALVES et al., 2022). A proposta consiste em aplicar técnicas de análise estatística, com destaque para a regressão logística e a análise de agrupamento, para identificar os fatores que influenciam a probabilidade de cancelamento. Busca-se, dessa forma, oferecer subsídios para que a empresa implemente ações preventivas e estratégias mais assertivas de retenção, promovendo maior engajamento ao longo do ciclo de mentoria e otimizando a lucratividade por cliente ativo.

Para atingir esse propósito, será realizado um levantamento detalhado das informações disponíveis na base de dados da empresa, incluindo características demográficas, histórico de engajamento, indicadores financeiros e métricas de satisfação. Em seguida, serão conduzidas etapas de análise exploratória, redução de dimensionalidade e agrupamento de perfis, culminando na modelagem preditiva por meio da regressão logística, conforme preconizado na literatura estatística aplicada (HOSMER; LEMESHOW; STURDIVANT, 2013; MORETTIN; SINGER, 2022).

Espera-se que os resultados contribuam não apenas para fornecer à empresa uma ferramenta preditiva de apoio à tomada de decisão, mas também para demonstrar a relevância da estatística na área de Customer Success, evidenciando seu papel na melhoria da experiência do cliente e no desenvolvimento sustentável das organizações.

1.1 Objetivo geral e específicos

O presente trabalho tem como objetivo geral analisar de forma aprofundada a base de dados dos clientes do Grupo Santinoni, identificando fatores associados ao cancelamento do serviço (*churn*) e desenvolvendo um modelo preditivo capaz de estimar a probabilidade de evasão. A partir dessa modelagem, busca-se oferecer subsídios quantitativos que permitam à empresa antecipar casos potenciais de cancelamento e implementar estratégias preventivas mais eficazes, contribuindo para a redução da taxa de churn e para o aprimoramento das ações de Customer Success.

Além disso, o estudo visa identificar perfis de clientes com comportamentos semelhantes por meio de técnicas de análise de agrupamento, ampliando a compreensão sobre padrões de engajamento, características financeiras e trajetórias dentro do programa de mentoria. A integração entre análise exploratória, clusterização e regressão logística constitui uma abordagem robusta para apoiar a tomada de decisão estratégica e direcionar ações de retenção de forma mais precisa.

Objetivos específicos

- Consolidar, estruturar e preparar a base de dados dos clientes, garantindo a padronização das variáveis e a integridade das informações utilizadas na modelagem.

- Realizar uma análise exploratória detalhada dos dados, de modo a compreender distribuições, padrões, correlações e possíveis inconsistências.

- Agrupar os clientes em clusters utilizando o método *k*-means, com o intuito de identificar perfis e segmentos relevantes para o comportamento de churn.

- Desenvolver um modelo preditivo de churn por meio da regressão logística, estimando a probabilidade de cancelamento a partir das variáveis selecionadas.

- Validar o modelo proposto e avaliar seu desempenho por meio de métricas adequadas, assegurando sua capacidade de generalização e aplicabilidade prática no contexto da empresa.

1.2 Justificativa

Em um mercado cada vez mais competitivo e dinâmico, a capacidade de compreender o comportamento do cliente e antecipar eventuais riscos de cancelamento torna-se essencial para a sustentabilidade das organizações. Empresas que dependem de relacionamentos contínuos, como aquelas que oferecem serviços por assinatura ou programas de mentoria de longo prazo, são particularmente

sensíveis ao fenômeno do *churn*, cuja ocorrência influencia diretamente a previsibilidade do faturamento, a alocação de recursos e a eficiência operacional.

No contexto do Grupo Santinoni, que oferece um programa de mentoria financeira com duração anual e possibilidade de renovação, a retenção de clientes é um elemento central do modelo de negócio. A evasão precoce compromete não apenas a receita recorrente, mas também o retorno sobre o investimento em aquisição (*Customer Acquisition Cost*), treinamento e acompanhamento especializado. Assim, identificar de forma antecipada os clientes com maior probabilidade de cancelar o serviço é estratégico para otimizar esforços, reduzir perdas e fortalecer a experiência do cliente ao longo da jornada. Como afirma Steinman et al. (2017), quando uma empresa não atua sobre os sinais de insatisfação ou desalinhamento do cliente, “[...] o impacto sobre a organização talvez seja fatal, desviando esforços que impulsionam o sucesso”.

O desenvolvimento de um modelo preditivo de churn permite que a empresa adote ações proativas de retenção, tais como: comunicação personalizada, ofertas direcionadas, ajustes no atendimento, intensificação de interações e intervenções específicas para perfis de maior risco. Ao prever potenciais cancelamentos com antecedência, a organização pode concentrar seus recursos nas situações mais críticas, aumentando a eficiência das estratégias de Customer Success e Marketing. Essa abordagem orientada por dados, alinhada às práticas contemporâneas de *Data-Driven Customer Management*, contribui não apenas para a redução da evasão, mas também para o aumento da satisfação, da fidelização e da lucratividade de cada cliente ativo.

Além do impacto empresarial, o estudo possui relevância acadêmica e formativa. A aplicação de técnicas estatísticas, como análise de agrupamento e regressão logística, permite consolidar os conhecimentos adquiridos ao longo da graduação, reforçando a capacidade analítica e a integração entre teoria e prática. O projeto também contribui para a construção de soluções reais que agregam valor ao processo de tomada de decisão dentro de um ambiente corporativo.

Por fim, esta proposta está alinhada ao Objetivo de Desenvolvimento Sustentável (ODS) nº 8, que preconiza o crescimento econômico sustentável e o trabalho decente. Ao desenvolver ferramentas que promovem a eficiência empresarial,

a valorização da relação com o cliente e a compreensão de suas necessidades, o estudo contribui para práticas organizacionais mais responsáveis, sustentáveis e orientadas ao bem-estar do consumidor.

2 REFERENCIAL TEÓRICO

A estatística desempenha um papel crucial na construção de modelos preditivos aplicados a contextos empresariais, especialmente em problemas de retenção de clientes e previsão de churn em serviços recorrentes. Nesses cenários, o volume e a complexidade dos dados exigem tanto fundamentos clássicos de inferência estatística quanto técnicas contemporâneas de ciência de dados, integrando modelagem supervisionada, análise multivariada e procedimentos rigorosos de validação.

Do ponto de vista da análise multivariada, o comportamento dos clientes é descrito por um conjunto de variáveis demográficas, comportamentais e de engajamento com o serviço. Johnson e Wichern (2007) destacam que a análise multivariada permite investigar simultaneamente várias variáveis explicativas, captando estruturas de correlação, padrões latentes e perfis de indivíduos que não seriam identificáveis por análises univariadas. Em particular, técnicas de agrupamento (clustering) e de redução de dimensionalidade constituem instrumentos centrais para segmentar clientes em grupos relativamente homogêneos, o que é fundamental para a definição de estratégias de Customer Success orientadas a dados (JOHNSON; WICHERN, 2007; HAIR et al., 2009).

2.1 Análise de agrupamento e segmentação de clientes

O agrupamento (clustering) é uma técnica de aprendizado não supervisionado com o objetivo de organizar um conjunto de dados em grupos (clusters), baseando-se na similaridade entre as observações. Diversos métodos de agrupamento são aplicados conforme as características dos dados e a finalidade da análise. Os mais conhecidos incluem K-means, Hierarchical Clustering e DBSCAN.

O K-means é um dos algoritmos de agrupamento mais utilizados, particularmente quando se deseja uma segmentação bem definida e não hierárquica dos dados. Este método busca minimizar a soma dos quadrados das distâncias entre

as observações e o centróide de cada cluster, sendo eficaz e simples de implementar (AGGARWAL, 2015). A escolha do K-means deve-se à sua simplicidade computacional e boa performance em bases numéricas, além de sua habilidade em definir grupos de clientes com características semelhantes, como engajamento, pressão financeira e tempo de relacionamento com a mentoria. Esses fatores são essenciais para a análise de churn no contexto de uma empresa de mentoria.

Ademais, o K-means apresenta a vantagem de ser um algoritmo com alto desempenho computacional, adequado para bases de dados com um número razoável de observações e variáveis, como é o caso deste estudo. Outros métodos de agrupamento, como o Hierarchical Clustering, oferecem uma visualização mais detalhada dos dados, mas podem ser mais dispendiosos computacionalmente e menos eficientes quando lidam com grandes bases de dados.

2.1.1 Medidas de Dissimilaridade e Distância

No algoritmo K-means, a principal medida de dissimilaridade utilizada é a distância Euclidiana, que calcula a diferença entre as observações no espaço multidimensional dos dados. A distância Euclidiana é amplamente aplicada em problemas de agrupamento devido à sua simplicidade e intuitividade: quanto maior a distância entre dois pontos, mais distintas são as observações.

Entretanto, em certos casos, outras medidas de distância, como a distância Manhattan ou a similaridade do cosseno, podem ser mais adequadas, dependendo das características dos dados. No presente estudo, a distância Euclidiana foi suficiente para agrupar clientes com características semelhantes, como a participação em reuniões e a fração paga da mentoria, uma vez que essas variáveis possuem uma estrutura numérica contínua.

2.1.2 Determinação do Número Ótimo de Clusters (Método do Cotovelo)

Uma das principais questões ao utilizar o algoritmo K-means é a escolha do número K de clusters. O método do cotovelo é amplamente utilizado para determinar esse número, sendo baseado na análise da soma dos quadrados dentro do cluster (WCSS) para diferentes valores de K. O número ótimo de clusters é geralmente identificado como o ponto em que o WCSS começa a diminuir de forma mais suave, formando um "cotovelo" no gráfico.

2.1.3 Validação de Clusters

Após definir o número de clusters, é crucial validar os resultados para garantir que os agrupamentos formados realmente refletem a estrutura intrínseca dos dados. Diversas métricas de avaliação podem ser utilizadas, sendo as mais comuns o coeficiente de silhueta, o WCSS e a análise visual dos clusters. O coeficiente de silhueta avalia a qualidade do agrupamento, considerando tanto a similaridade intra-cluster quanto a dissimilaridade entre clusters. Valores próximos de 1 indicam uma boa separação entre os clusters, enquanto valores negativos sugerem que os pontos podem ter sido atribuídos ao cluster incorreto.

Adicionalmente, a visualização gráfica dos clusters facilita a identificação de padrões e permite verificar se os grupos formados possuem características distintas. No caso deste estudo, a visualização do K-means com as variáveis de participação em reuniões e fração paga da mentoria demonstrou que os clusters estavam bem separados, com padrões claros de comportamento tanto de clientes ativos quanto de clientes em risco de churn.

2.2 Fundamentos da Regressão Logística

A modelagem estatística é uma ferramenta essencial para a investigação de relações entre variáveis. Enquanto a regressão linear clássica é adequada para variáveis dependentes contínuas, a regressão logística é a técnica padrão para situações em que a variável resposta é dicotômica (binária) ou politômica (HOSMER; LEMESHOW; STURDIVANT, 2013).

Diferentemente da regressão linear, onde a média condicional E(Y|x) pode assumir qualquer valor real, na regressão logística, sendo Y uma variável Bernoulli (assumindo valores 0 ou 1), a média condicional deve ser limitada entre 0 e 1, pois representa uma probabilidade $\pi(x) = P(Y = 1|x)$ (ANDRADE, 2008). Onde, $\pi(x)$ é uma função de distribuição de probabilidade, na qual representam a probabilidade de “sucesso”, dado o valor x de uma variável explicativa qualquer.

Como $\pi(x)$ varia entre zero e uma transformação fundamental nesta técnica é o **logit**, que permite linearizar a relação entre as variáveis independentes e a probabilidade do evento. O logit é o logaritmo natural da chance (*odds*) de sucesso:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

A partir dessa transformação, a probabilidade predita é recuperada pela função logística inversa:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}},$$

Sendo β_0 e β_1 parâmetros desconhecidos. Conforme $x \to \infty$, $\pi(x)$ decresce para 0 quando $\beta_1 < 0$, e $\pi(x)$ cresce para 1 quando $\beta_1 > 1$. Enquanto deve cair no intervalo (0,1), o $logit[\pi(x)]$ pode ser qualquer número real. (AGRESTI, 2002).

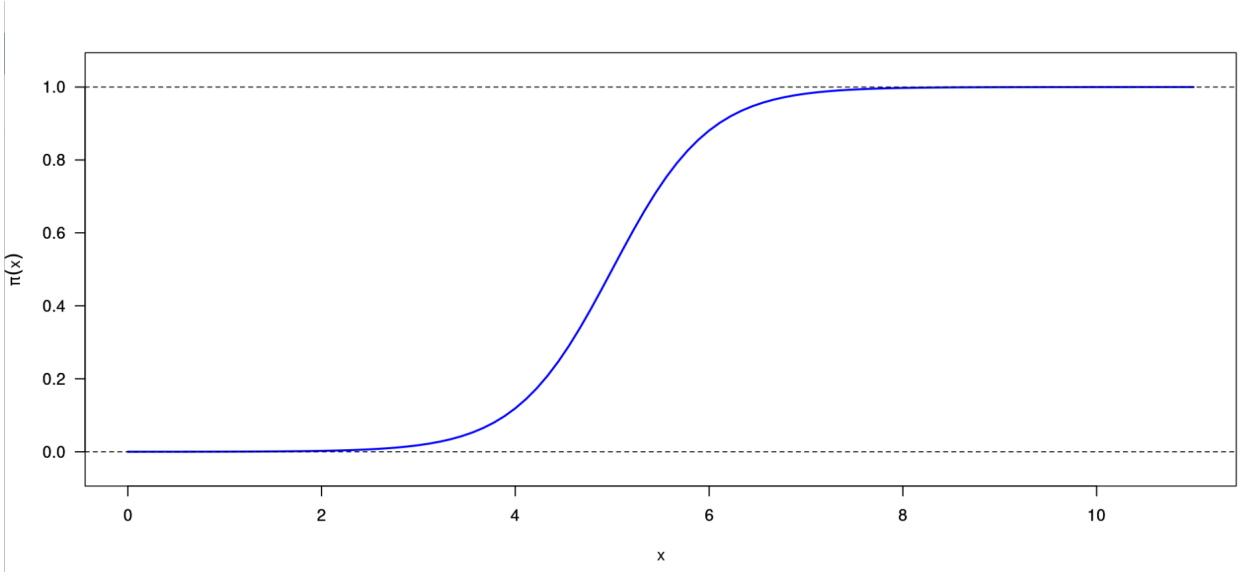
Em qualquer problema de regressão a quantidade a ser modelada é a esperança da variável aleatória depende, dado o valor da variável independente, ou seja, $E(Y|X = x)$, afirma Souza (2006). Na regressão logística devido a natureza da resposta $0 \leq E(Y|X = x) \leq 1$ enquanto que na regressão linear $-\infty \leq E(Y|X = x) \leq \infty$. Para contornar as limitações da linearidade e satisfazer a restrição de probabilidade, utiliza-se a função de distribuição logística, em que, usando a definição de variáveis aleatórias discretas, tem-se que

$$E(Y|X = x) = 1P(Y = 1|X = x) + 0P(Y = 0|X = x) = \pi(x)$$

Outra distinção entre os modelos reside na distribuição condicional de Y. No modelo linear assume-se que uma observação possa ser expressa como $y = E(Y|x) + \epsilon$, com a suposição de que o erro ϵ tenha distribuição Normal com media zero e variância constante. No entanto, no modelo logístico, uma observação pode ser expressa como $y = \pi(x) + \epsilon$, em que ϵ assume duas possibilidades, caso $y = 1$, então $\epsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, e se $y = 0$, então $\epsilon = -\pi(x)$ com probabilidade $1 - \pi(x)$. Portanto, ϵ tem distribuição Bernoulli com média zero e variância $\pi(x)[1 - \pi(x)]$(COSTA, 2013).

Uma vez que a estimação da probabilidade se restringe ao intervalo (0,1), presume-se que o impacto das mudanças na variável independente sobre a variável resposta se torna progressivamente menor à medida que os valores se aproximam dos limites extremos do intervalo. A curva de regressão, caracterizada como monótona (conforme ilustrado na Figura 1 do seu referencial), assemelha-se a uma **função de distribuição acumulada (f.d.a)** de uma variável aleatória contínua quando o coeficiente $\beta 1 > 0$. Isso sugere um modelo para resposta binária tendo a forma $\pi(x) = F_X(x)$, de alguma f.d.a F_X(COSTA, 2013).

Figura 1 - Curva do modelo logístico



Fonte: Elaborado pelo autor.

Quando $\beta 1 > 0$, a curva de regressão logística é uma função de distribuição acumulada da distribuição logística. A f.d.a de uma distribuição logística com média μe parâmetro de dispersão $\tau > 0$ é da seguinte forma:

$$FX\,(x) = \frac{exp[(x-\mu)/\tau\,]}{1+exp[(x-\mu)/\tau\,]}\,, -\infty < x < \infty.$$

A forma padronizada da f.d.a logística tem $\mu = 0$ e $\tau = 1$. Seja $\Phi(.)$ denotando umas f.d.a padrão, então $\Phi(x) = e^{x}/(1+e^{x})$. Para esta função, a curva de regressão logística tem a forma $\pi(x) = \Phi(\beta 0 + \beta 1x)$. A transformação $logit$ é simplesmente a função inversa da função distribuição padrão; isto é, quando $\Phi(x) = e^{x}/(1+e^{x})$, então $x = \Phi^{-1}[\pi(x)] = ln[\pi(x)/(1-\pi(x))]$(AGRESTI, 2002).

2.2.1 Estimação dos Parâmetros: O Método da Máxima Verossimilhança

Identificada a equação que permite calcular a probabilidade relativa à ocorrência de determinado evento, resto estimar os seus coeficientes. Ao contrário da regressão linear, que comumente utiliza o Método dos Mínimos Quadrados (MMQ) para estimar os parâmetros do modelo, a regressão logística baseia-se no **Método da Máxima Verossimilhança** (MMV). Isso se deve ao fato de que, em variáveis binárias, os erros não seguem uma distribuição normal e a variância não é constante (heterocedasticidade), dependendo da média condicional $\pi(x)$ (HOSMER; LEMESHOW; STURDIVANT, 2013).

O princípio da máxima verossimilhança busca encontrar os valores para os coeficientes beta que maximizam a probabilidade de observar os dados amostrais coletados. A função de verossimilhança $L(\beta)$ para uma amostra de n observações independentes é dada pelo produto das probabilidades correspondentes aos valores observados de Y:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_1)^{y_i} \left[1 - \pi(x_1)\right]^{1-y_i}$$

Para facilitar a derivação matemática, trabalha-se com o logaritmo da função de verossimilhança (*log-likelihood*), que é dado por:

$$\ell(\beta) = \sum_{i=n}^{n} [yi\, ln\, \pi(xi) + (1 - yi)\, ln(1 - \pi(xi))],$$

Fazendo as transformações necessárias temos:

$$\ell(\beta) = \sum_{i=n}^{n} \{yi\, (\beta_0 + \beta_1 x_i) - ln[1 + exp(\beta_0 + \beta_1 x_i)]\}.$$

Para encontrar os estimadores de máxima verossimilhança, derivam-se as equações de verossimilhança em relação aos parâmetros e igualam-se a zero. Obtendo-se as equações:

$$\frac{\partial \ell(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}]$$

$$\frac{\partial \ell(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} [y_i x_i - \frac{exp(\beta_0+\beta_1 x_i)x_i}{1+exp(\beta_0+\beta_1 x_i)}].$$

Então o estimador β pelo método da máxima verossimilhança, denotado $\widehat{\beta}$, é a solução das equações de verossimilhança:

$$\sum_{i}^{n} [y_i - \pi(x_i)] = 0$$

$$\sum_{i}^{n} x_i [y_i - \pi(x_i)] = 0.$$

Como essas equações são não-lineares em relação aos parâmetros beta, não existe solução analítica fechada, exigindo métodos numéricos iterativos, como o método de Newton-Raphson ou *Fisher Scoring*, para a obtenção das estimativas (ANDRADE, 2008; HOSMER; LEMESHOW; STURDIVANT, 2013).

2.2.2 Tratamento de Variáveis e Interações no Modelo

A flexibilidade da regressão logística permite a inclusão de variáveis preditoras de diferentes escalas, exigindo, contudo, tratamentos específicos para garantir a interpretação correta.

2.2.2.1 Variáveis Categóricas e Contínuas

Para variáveis independentes **categóricas** (nominais ou ordinais) com mais de dois níveis, é inadequado modelá-las como se fossem contínuas, pois isso imporia uma relação linear e ordenada artificial aos dados. O método correto envolve a criação de variáveis indicadoras (*dummy variables* ou variáveis de planeamento). Se uma variável possui k categorias, são criadas k-1 variáveis binárias, sendo uma categoria escolhida como referência (codificada como 0 em todas as variáveis indicadoras). Os coeficientes resultantes representam a comparação direta entre cada categoria e a categoria de referência (HOSMER; LEMESHOW; STURDIVANT, 2013).

Para variáveis **contínuas**, a premissa básica é a **linearidade no logit**. Isso significa que se assume que a mudança no logaritmo da chance é constante para cada unidade de aumento na variável independente. Hosmer, Lemeshow e Sturdivant (2013) recomendam verificar essa suposição, utilizando métodos como polinômios fracionários ou suavização (*LOWESS*) para detectar não-linearidades, as quais podem exigir transformações na variável ou categorização.

2.2.2.2 Interações

A interação estatística, ou modificação de efeito, ocorre quando a magnitude da associação entre uma variável independente e a variável resposta varia de acordo com o nível de uma terceira variável. No contexto da regressão logística, a interação é modelada pela inclusão de um termo produto entre as duas variáveis (ex: $X_1 \times X_2$). A significância estatística desse termo indica que o efeito de X_1 sobre o logit não é constante, mas depende do valor de X_2. A interpretação dos *Odds Ratios* em modelos com interação torna-se condicional, não podendo ser expressa por um único valor (HOSMER; LEMESHOW; STURDIVANT, 2013).

2.3 Diagnóstico do Modelo

A avaliação da adequação do modelo não se encerra com a estimativa dos parâmetros. É crucial realizar um diagnóstico para identificar observações que não são bem explicadas pelo modelo ou que exercem influência desproporcional nas estimativas.

2.3.1 O likelihood value

Sua função principal é verificar se a regressão como um todo é estatisticamente significante e facilitar comparações entre modelos alternativos, servindo para verificar se o modelo melhora com a inclusão ou exclusão de uma ou mais variáveis independentes. Considerada uma das principais medidas de avaliação geral do modelo de regressão logística, sendo que busca aferir a capacidade do modelo estimar a probabilidade associada à ocorrência de um determinado evento. Ele é representado pela expressão $-2LL$, que é o logaritmo natural do likelihood value multiplicado por -2, seguindo uma distribuição Qui-quadrado. Quanto mais próximo de zero, maior o poder preditivo do modelo (COSTA, 2013).

2.3.2 O teste de Hosmer e Lemeshow

A finalidade desse teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada. A certo nível de significância busca-se aceitar a hipótese de que não existem diferenças entre os valores preditos e observados. Caso existam diferenças significativas entre essas classificações, o modelo não representa a realidade de forma satisfatória.

2.4 Análise de Resíduos

Na análise de diagnóstico um tema fundamental é a detecção de observações influentes, isto é, pontos que exercem um peso desproporcional nas estimativas dos parâmetros do modelo.

Esses resíduos são fundamentais para identificar discrepâncias e padrões não aleatórios no ajuste (HOSMER; LEMESHOW; STURDIVANT, 2013). Segundo Souza (2006), a análise de resíduos e diagnóstico é utilizada para detectar os problemas a seguir: Presença de observações discrepantes; Inadequação das pressuposições para os erros aleatórios ou a médias; Colinearidade entre as colunas da matriz modelo; Forma funcional do modelo inadequado; Presença de observações influentes.

Resumindo as estatísticas de influência determinando o quanto a exclusão de uma observação em particular pode influenciar no ajuste do modelo. A seguir, as medidas geralmente utilizadas para os resíduos e diagnósticos serão abordadas.

2.4.1 Resíduo de Pearson

O **Resíduo de Pearson** atua como um recurso de diagnóstico na classificação de observações que possam ser caracterizadas como outliers no conjunto amostral. O resíduo para cada elemento amostral é definido como a diferença entre os valores observados e os valores preditos e é definido por:

$$r_i = y_i - \widehat{\pi}(x_i)$$

Entretanto, devido à influência da escala de medição, esta formulação do resíduo não é adequada para a detecção de outliers. Desse modo, Souza (2006) preconiza a necessidade de **transformação** desse resíduo para mitigar o efeito das variáveis resposta e preditora. No contexto do modelo de regressão logística, o resíduo de Pearson transformado é especificado pela expressão:

$$(rp)_i = \frac{y_i - \widehat{\pi}(x_i)}{\sqrt{\widehat{\pi}(x_i)\,(1 - \widehat{\pi}(x_i))}},\ i = 1, 2, \ldots, n,$$

2.4.2 Resíduo de Deviance

Os **Resíduos de Deviance** são elementos cruciais da estatística de *deviance* e são empregados no diagnóstico de erros no ajuste do modelo. Sua função é mensurar a discrepância existente entre o modelo saturado e o modelo restrito em relação aos valores observados y_i. A *deviance* constitui uma estatística de bondade de ajuste que, para cada observação individual ($i = 1, 2, \ldots, n$), é derivada do logaritmo da função de verossimilhança, definida por:

$$d_i = -\sqrt{-2ln(1 - \widehat{\pi}(x_i))}, \quad \text{,se } y_i = 0$$

$$d_i = \pm\sqrt{2[y_i ln(y_i/\widehat{\pi}(x_i)) + (-y_i) ln(1 - y_i/1 - \widehat{\pi}(x_i))]} \quad \text{, se } 0 < y_i < 1$$

$$d_i = \sqrt{-2ln(\widehat{\pi}(x_i))}, \quad \text{, se } y_i = 1$$

2.4.3 Detecção de Outliers e Influência

Pontos discrepantes (*outliers*) são observações com resíduos grandes, indicando falha de ajuste. Já a **influência** refere-se ao impacto que uma observação individual tem sobre os coeficientes estimados do modelo. Uma medida chave para

avaliar a influência é a matriz *Hat* (H), cujos elementos da diagonal (h_j) medem a alavancagem (*leverage*) da observação no espaço das covariáveis.

No modelo de regressão linear, a matriz H é definida por:

$$H = X(X^T X)^{-1} X^T$$

Como nos modelos de regressão logística, $Var(\epsilon_i) = \pi(xi)(1 - \pi(xi))$ não é constante, a matriz de projeção para o modelo logístico fica definida como:

$$H = Q^{1/2} X(X^T QX)^{-1} X^T Q^{1/2}$$

Em que a utilização dos elementos da diagonal principal de H é utilizada para detectar a presença de pontos de alavanca. Hosmer e Lemeshow (1989) afirmam, contudo, que o uso da diagonal principal da matriz de projeção H deve ser feito com algum cuidado em regressão logística e que as interpretações são diferentes daquelas no caso do modelo linear.

Para quantificar a influência global de uma observação, utiliza-se frequentemente análogos à Distância de Cook ou a estatística $\Delta\hat{\beta}_j$, que estima o quanto os coeficientes de regressão mudariam se a j-ésima observação fosse excluída da análise. Gráficos de $\Delta\hat{\beta}_j$ versus a probabilidade predita são ferramentas visuais recomendadas para identificar casos que afetam substancialmente o modelo (HOSMER; LEMESHOW; STURDIVANT, 2013).

2.5 Avaliação de Desempenho do Modelo

A qualidade preditiva de um modelo de regressão logística é frequentemente avaliada por sua capacidade de discriminação, isto é, distinguir corretamente entre os sujeitos que experienciam o evento e os que não.

2.5.1 Classificação: Sensibilidade e Especificidade

A partir das probabilidades preditas, pode-se classificar as observações definindo um **ponto de corte** (*cutoff*), usualmente 0,5. A partir disso, constrói-se uma tabela de classificação para calcular:

1. **Sensibilidade:** A proporção de verdadeiros positivos corretamente identificados pelo modelo.
2. **Especificidade:** A proporção de verdadeiros negativos corretamente identificados.

Andrade (2008) ressalta que existe um compromisso (*trade-off*) entre essas medidas: ao alterar o ponto de corte para aumentar a sensibilidade, frequentemente reduz-se a especificidade, e vice-versa.

2.5.2 Curva ROC e AUC

Para avaliar a capacidade discriminatória do modelo independentemente de um ponto de corte fixo, utiliza-se a **Curva ROC** (*Receiver Operating Characteristic*). Este gráfico plota a Sensibilidade (eixo Y) versus 1 - (Especificidade) (eixo X) para todos os pontos de corte possíveis.

A curva de Característica Operatória do Receptor (curva ROC) ilutra a relação entre a sensibilidade e especificidade, e pode ser utilizada para decidir um bom ponto de corte.

A métrica derivada desta curva é a **Área Sob a Curva** (AUC - *Area Under the Curve*). A AUC varia de 0,5 a 1,0, onde:

1. **AUC = 0,5:** Indica ausência de capacidade discriminatória (equivalente ao acaso).
2. **AUC = 1,0:** Indica discriminação perfeita.

Segundo Hosmer, Lemeshow e Sturdivant (2013), uma regra geral para interpretação da AUC é: $0,7 \leq AUC < 0,8$ é considerado aceitável; $0,8 \leq AUC < 0,9$ é excelente; e $AUC \geq 0,9$ é excepcional. A Curva ROC também auxilia na determinação do ponto de corte ótimo, geralmente aquele que maximiza a soma da sensibilidade e especificidade, localizando-se no canto superior esquerdo do gráfico (ANDRADE, 2008).

2.6 Interpretação do modelo ajustado

Um dos principais atrativos da regressão logística é a interpretabilidade dos seus coeficientes em termos de **Razão de Chances** (*Odds Ratio* - OR). A chance é definida como a razão entre a probabilidade de ocorrência do evento e a probabilidade de sua não ocorrência. De acordo com Hosmer, a **chance (odds)** de o resultado estar presente entre indivíduos com $x = 1$ é $\pi(1)/[1 - \pi(1)]$. Da mesma forma, a chance de o resultado estar presente entre indivíduos com $x = 0$ é $\pi(0)/[1 - \pi(0)]$. A **razão de chances (odds ratio)**, denotada por **OR**, é a razão entre as chances para $x = 1$ e as chances para 0, e é dada pela equação:

$$OR = \frac{\frac{\pi(1)}{[1-\pi(1)]}}{\frac{\pi(0)}{[1-\pi(0)]}}$$

Substituindo as expressões pelas probabilidades do modelo de regressão logística, a equação acima pode ser definida por:

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} \Bigg/ \frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0}$$

Ou seja, matematicamente, para uma variável independente dicotômica codificada como 0 e 1, a OR é obtida pela exponenciação do coeficiente beta:

$$OR = e^{\beta}$$

Se OR > 1, a presença do fator está associada a um aumento na chance do evento (fator de risco); se OR < 1, há uma diminuição na chance (fator de proteção); e se OR = 1, não há associação (ANDRADE, 2008; HOSMER; LEMESHOW; STURDIVANT, 2013).

A Razão de Chances (Odds Ratio) é amplamente utilizada como uma medida de associação, pois aproxima o quanto é mais provável ou improvável (em termos de chances) que o desfecho esteja presente entre os sujeitos com x = 1 em comparação com aqueles sujeitos com x = 0.

Em certas configurações, a razão de chances pode se aproximar de outra medida de associação chamada Risco Relativo (Relative Risk), que é a razão das duas probabilidades de desfecho: $RR = \pi(1)/\pi(0)$. Segue-se da equação acima que a razão de chances se aproxima do risco relativo se $[1 - \pi(0)]/[1 - \pi(1)] \approx 1$. Isso ocorre quando $\pi(x)$ é pequeno tanto para x = 0 quanto para x = 1, o que é frequentemente referido em pesquisa médica/epidemiológica como a suposição de doença rara (rare disease assumption),(HOSMER; LEMESHOW; STURDIVANT, 2013).

O referencial teórico detalhado sustenta o desenvolvimento das fases metodológicas da pesquisa. Estas compreendem a análise exploratória dos dados, a

formação de *clusters* de clientes para segmentação, o ajuste e a validação do modelo de regressão logística para a previsão de *churn*, e, subsequentemente, a interpretação dos achados sob a ótica das estratégias de *Customer Success* do Grupo Santinoni

3 METODOLOGIA / MATERIAIS E MÉTODOS

3.1 Dados

A metodologia deste trabalho abrange as etapas de coleta e tratamento de dados, análise exploratória, análise de agrupamento, modelagem preditiva utilizando regressão logística e validação do modelo. Cada fase é crucial para garantir a robustez e a interpretabilidade do modelo de previsão de Churn. Os dados que vão ser utilizados neste estudo foram fornecidos pela empresa a ser estudada e estão armazenados em um arquivo do Google Sheets. A metodologia adotada terá várias etapas descritas a seguir.

3.2 Coleta e estruturação dos dados

A base analítica utilizada neste estudo foi disponibilizada pelo **Grupo Santinoni**, composta por registros históricos de clientes do serviço de mentoria financeira. Os dados foram originalmente armazenados em planilha (Google Sheets) e apresentavam lacunas e inconsistências que demandaram um processo de limpeza e padronização antes da modelagem (tratamento de ausentes, padronização de categorias e transformação de variáveis financeiras assimétricas).

Para enriquecer a base de dados e capturar informações relevantes para a previsão de Churn, foram introduzidas novas variáveis que antes não eram sistematicamente controladas pela empresa. Essas variáveis, essenciais para identificar padrões de comportamento e fatores de risco, incluem:

Renda Mensal do Cliente: Informação sobre a capacidade financeira do cliente no momento da contratação do serviço.

Data de Cancelamento: Registro da data em que o cliente efetivou o cancelamento do serviço.

Motivo do Cancelamento: Categoria que descreve a razão declarada pelo cliente para o cancelamento, fornecendo insights qualitativos sobre a insatisfação ou mudança de necessidade. Quantidade de Reuniões Realizadas: Controle do número de interações do cliente com os especialistas da empresa, indicando o nível de engajamento e suporte recebido.

Quantidade de Produtos Adicionais Contratados: Registro de outros produtos ou serviços adquiridos pelo cliente além da mentoria principal, o que pode indicar o nível de fidelidade e aprofundamento do relacionamento.

Tempo de Mentoria: Duração do período em que o cliente permaneceu ativo na mentoria, crucial para a análise de longevidade do cliente.

Quantidade de tempo: Registro da quantidade em meses que o cliente contratou a mentoria.

3.3 Conhecimento das variáveis

As variáveis escolhidas para compor o conjunto de dados foram:

Tabela 1 - Tipo das variáveis

Variáveis	Tipo da variável
status	Qualitativa
renda_inicial	Quantitativa continua
divida_total	Quantitativa continua
dívida	Qualitativa
idade	Quantitativa discreta
sexo	Qualitativa
unidade_federativa	Qualitativa
estado_civil	Qualitativa
valor_da_mentoria	Quantitativa continua

quant._de_parcelas_pagas_da_mentoria	Quantitativa discreta
tempo_de_mentoria	Quantitativa continua
motivo_do_cancelamento	Qualitativa
scoreultimo_nps	Quantitativa continua
quantidade_de_reuniões_canceladas	Quantitativa discreta
quantidade_de_reuniões	Quantitativa discreta
quantidade_de_produtos_contratados	Quantitativa discreta

Fonte: Elaborado pelo autor (2025)

3.3.1 Informações sobre as variáveis

Observando o quadro 1, diversas informações relevantes podem ser extraídas, oferecendo uma visão sobre as características dos clientes:

- Status: Descreve o status do cliente na mentoria, onde temos dois possíveis valores, clientes ativos = 0 e clientes cancelados = 1.
- Renda_inicial: Registra a renda mensal do cliente no início da mentoria. Divida_total: Calcula a dívida total do cliente. 12
- Dívida: É uma variável qualitativa descrevendo os clientes que possuem ou não dividas, onde cliente sem dividas = 0, cliente com dívida = 1.
- Idade: Registra a idade do cliente.
- Sexo: Registra o sexo do cliente, onde o sexo feminino= 0 e sexo masculino=1. unidade_federativa: Mostra o estado do cliente.
- estado_civil: Descreve o estado civil do cliente. valor_da_mentoria: Calcula o valor que a mentoria foi vendida na época.
- quant._de_parcelas_pagas_da_mentoria: Descreve a quantidade de parcelas que o cliente já pagou ou pagou durante sua mentoria.
- tempo_de_mentoria: Calcula o tempo em meses que o cliente está ativo na mentoria
- scoreultimo_nps: Registra a nota de satisfação do cliente mais recente.
- quantidade_de_reuniões: Numero total de reuniões feitas com o cliente.
- quantidade_de_reuniões_canceladas: Numero total de reuniões canceladas pelo cliente.

3.4 Análise exploratória dos dados

A Análise Exploratória de Dados (AED) será realizada após o tratamento inicial e a engenharia de features. O objetivo da AED é compreender a estrutura dos dados, identificar padrões, tendências, anomalias e relações entre as variáveis. Esta etapa é crucial para informar as decisões de modelagem e para aprofundar o entendimento do fenômeno de Churn na empresa. As principais atividades da AED incluirão: Análise Descritiva: Cálculo de estatísticas descritivas (média, mediana, desvio padrão, quartis, etc.) para variáveis numéricas e contagem de frequência para variáveis categóricas. Visualizações Gráficas: Criação de gráficos como histogramas, boxplots, identificar outliers e explorar relações entre elas. Gráficos de série temporal serão utilizados para analisar o comportamento de Churn ao longo do tempo. Análise de Correlação: Investigação das correlações entre as variáveis preditoras e a variável resposta (Churn). Isso ajudará a identificar as variáveis mais relevantes para o modelo e a detectar possíveis problemas de multicolinearidade entre as preditoras. Segmentação de Clientes: Embora o objetivo principal seja a regressão logística, a análise de agrupamento (clusterização), como o método K-means, pode ser utilizada para identificar segmentos de clientes com características semelhantes, o que pode fornecer insights adicionais sobre o comportamento de Churn em diferentes grupos.

3.5 Engenharia de variáveis (pós-AED)

Os resultados mostraram que, **nas variáveis originais**, as diferenças entre clientes **ativos** e **churn** nem sempre eram nítidas. Por exemplo: a participação média em reuniões por mês ficou praticamente igual entre os grupos, enquanto **recência da última reunião** e **fração paga** sugeriram maior poder discriminativo (ativos: recência ~118 dias; churn: ~320 dias; fração paga: 0,67 vs. 0,50).

Já **debt-to-income** apresentou assimetria acentuada e muitos zeros, exigindo transformação antes de uso no modelo. A listagem das variáveis do banco original (status, renda, dívida, idade, sexo, UF, estado civil, tempo de mentoria, reuniões, etc.) reforçou a importância de transformar e combinar informações para representar **engajamento** e **capacidade financeira** de modo mais fiel ao processo de cancelamento.

3.5.1 Engajamento e dinâmica temporal

- **tenure_months**: tempo de mentoria em **meses** (base temporal para normalizações).

- **meetings_participated_pm**: **participação por mês** (reuniões/tenure), substituindo contagem total para controlar exposição.

- **recency_last_meeting_days** → **recency_c**: dias desde a última reunião, **winsorizado** (p99 definido no treino) e **padronizado** (z-score), capturando **desengajamento recente**.

- **Interações de negócio** (criadas **após** a padronização):

    - tenure_months × meetings_participated_pm (baixa participação no início do ciclo → **maior risco**);

    - meetings_participated_pm × recency_c (queda recente de participação → **sinal de alerta**).

Essa ordem (padronizar → interagir) reduz **multicolinearidade** e melhora a estabilidade dos coeficientes (HOSMER; LEMESHOW;STURDIVANT, 2013).

3.5.2 Pressão financeira relativa e fase do contrato

- **debt_to_income (DTI = dívida/renda)** com *capping* em p99 (treino) e **transformação** log_dti = log1p(DTI), reduzindo assimetria e aproximando linearidade no logito;

- **fraction_paid** (fração quitada), proxy da **janela de renovação**; próximo ao término, tende a elevar o risco de churn;

- **Flags**: has_debt (possui dívida) e has_cross_sell (outros produtos);

- **near_renewal**: indicador para **tenure ∈ [11,13] meses**, período crítico para a decisão de renovação.

3.5.3 Contexto do cliente e redução de cardinalidade

- **UF_macro**: macrorregiões (N, NE, CO, SE, S) em substituição à UF individual;

- **especialidade_macro**: agrupamento em macroáreas clínicas (Clínico Geral, Cardiologia, Pediatria, Gineco/Obstetrícia, Cirurgia/Ortopedia, Gastro, Urologia, Psiquiatria, Oftalmologia, Nefrologia, Outras);

- **Dummies em k−1** e **lumping** de níveis raros nas demais categóricas, minimizando **quase-separação** e coeficientes instáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.6 Análise de Agrupamento (K-means)

A análise de agrupamento foi realizada utilizando o algoritmo K-means, que é um método de aprendizado não supervisionado para identificar grupos (ou clusters) de dados semelhantes. O K-means organiza as observações em K grupos, de forma que as observações dentro de cada grupo sejam mais semelhantes entre si do que com observações de outros grupos.

3.6.1. Normalização dos Dados

Antes de aplicar o K-means, foi necessário normalizar os dados para evitar que variáveis com escalas diferentes influenciassem os resultados. O K-means depende da distância entre os dados, e variáveis com escalas diferentes (como renda, idade e número de reuniões) podem distorcer a análise. Para isso, utilizou-se a função “scale()” para normalizar as variáveis numéricas, deixando-as com média igual a 0 e desvio padrão igual a 1. Isso garantiu que todas as variáveis tivessem igual importância no processo de agrupamento.

3.6.2 Determinação do Número de Clusters (K)

O número de clusters (K) a ser utilizado no modelo foi determinado por meio do método do cotovelo. Esse método é uma técnica comum para seleção do número ideal de clusters, baseada na soma dos quadrados das distâncias dentro de cada cluster (WCSS - Within-Cluster Sum of Squares). A ideia do método do cotovelo é observar onde ocorre a diminuição acentuada do WCSS no gráfico. Após esse ponto, o WCSS diminui de maneira mais suave, indicando que o número de clusters foi suficientemente grande para representar a variação nos dados sem adicionar complexidade desnecessária.

No caso deste trabalho, o gráfico gerado pelo método do cotovelo indicou que K = 3 era o número ideal de clusters. Esse valor foi escolhido porque, após K=3, a redução do WCSS começou a desacelerar.

3.6.3 Aplicação do Algoritmo K-means

Com o número de clusters determinado, aplicou-se o algoritmo K-means com K = 3 clusters. Para garantir que o algoritmo encontrasse uma solução globalmente boa, foi utilizado o parâmetro nstart = 20, o que significa que o K-means foi executado 20 vezes com diferentes inicializações dos centroides. Isso ajuda a evitar que o algoritmo fique preso em mínimos locais e garante que o cluster ideal seja encontrado.

O K-means atribui a cada cliente um cluster específico, e os resultados são armazenados em dois elementos principais:

- kmeans_resultado$cluster: as atribuições de cluster para cada observação (cliente).

- kmeans_resultado$centers: os centroides de cada cluster, representando a média das características de cada grupo.

3.6.4 Visualização dos Clusters

Uma vez realizados os agrupamentos, foi gerado um gráfico de dispersão para visualizar os clusters de clientes. No gráfico, utilizou-se como variáveis de eixo X e eixo Y as variáveis “meetings_participated_pm” (participação nas reuniões) e “fraction_paid” (fração paga da mentoria). Essas variáveis foram escolhidas por sua importância para o engajamento e comprometimento financeiro do cliente com a mentoria, fatores diretamente relacionados ao risco de churn.

No gráfico, cada cliente foi colorido de acordo com o cluster ao qual foi atribuído, permitindo visualizar como as diferenças de comportamento entre os grupos são representadas nas variáveis selecionadas.

3.6.5 Análise do Desempenho do K-means

Após a execução do K-means, a qualidade do agrupamento foi avaliada com base na soma dos quadrados dentro do cluster (WCSS) e no coeficiente de silhueta, que é uma medida da qualidade do agrupamento. Um valor de silhueta próximo de 1 indica que as observações dentro de um cluster estão bem agrupadas e distantes de outros clusters.

3.6.6 Interpretação dos Resultados

Com a análise dos clusters, foram identificados três grupos de clientes com características bem distintas:

- Cluster 1: clientes com baixo engajamento nas reuniões e baixa fração paga, sugerindo que esses clientes estão em risco de churn.

- Cluster 2: clientes com engajamento intermediário e compromisso financeiro moderado, representando clientes que estão em uma situação intermediária de fidelidade.

- Cluster 3: clientes com alto engajamento e alto comprometimento financeiro, representando clientes com baixo risco de churn e alta probabilidade de renovação.

Esses clusters forneceram insights valiosos sobre os perfis de clientes e podem ser utilizados para direcionar ações de retenção de clientes, marketing segmentado e personalização de ofertas.

3.6.7 Validação dos Clusters

A validade dos clusters foi verificada utilizando o coeficiente de silhueta, que apresentou valores positivos, indicando que a separação dos clusters foi eficaz. Além disso, o gráfico de dispersão e a análise dos centros dos clusters mostraram que os grupos estavam bem separados com base nas características selecionadas, confirmando a relevância do agrupamento para a análise de comportamento de churn.

Considerações Finais do agrupamento

A análise de agrupamento K-means permitiu segmentar os clientes em três grupos distintos com base em suas características de engajamento e comprometimento financeiro. Esse agrupamento revelou padrões importantes para entender o comportamento de churn e retenção. As variáveis selecionadas para o agrupamento refletem bem as dinâmicas de desengajamento, dívida e pagamento da mentoria, fornecendo insights práticos para estratégias de retenção. A metodologia utilizada seguiu boas práticas de pré-processamento e validação, assegurando que os resultados fossem confiáveis e interpretáveis.

3.7 Particionamento dos dados (treinamento e teste)

Para estimar e avaliar o modelo de churn, o conjunto de dados foi particionado em duas amostras mutuamente exclusivas: treinamento (≈80% das observações) e teste (≈20%). O treinamento é utilizado para ajustar o modelo e selecionar especificações; o conjunto de teste permanece resguardado até a etapa final, sendo empregado exclusivamente para mensurar a capacidade de generalização da solução.

Quando disponível, recomenda-se o particionamento temporal, treinamento em períodos pretéritos e teste em períodos posteriores, por refletir o cenário operacional de previsão. Na ausência de datas com qualidade suficiente, realizou-se um sorteio estratificado por classe, preservando a proporção de clientes que cancelaram/renovaram em ambos os conjuntos, de modo a reduzir variações de desempenho devidas a amostragens desbalanceadas.

- A literatura de aprendizado estatístico destaca que a avaliação de classificadores deve ser conduzida em dados não utilizados no ajuste, podendo incluir validação cruzada para comparação entre modelos e definição de hiperparâmetros (e.g., k-fold), a fim de mitigar o otimismo amostral e orientar a escolha do modelo com menor erro de validação (vide discussão de validação/VC e comparação de classificadores) . A etapa subsequente de determinação de pontos de corte ótima (threshold) pode ser guiada por curvas ROC e compromisso entre sensibilidade e especificidade, conforme proposto por Morettin e Singer (2022) .

3.8 Pré-processamento e tratamento das variáveis

O pré-processamento foi implementado por meio de um **pipeline** que garante reprodutibilidade e **ausência de vazamento**: todas as transformações são **ajustadas no conjunto de treinamento** (medianas, modas, identificação de níveis raros, parâmetros de padronização) e **aplicadas no conjunto de teste** sem re-ajuste.

As etapas incluíram:

- Exclusão de identificadores e variáveis sabidamente pós-tratamento;
- **Imputação** de faltantes numéricos por mediana e categóricos por moda (robustez a outliers e manutenção de estrutura de variabilidade);
- **Colapso de níveis raros** em “outros” para reduzir esparsidade e prevenir problemas de separação/quase separação;
- Criação de nível “**novo**” para categorias não observadas no treinamento (evitando falhas na predição);
- **Codificação *one-hot*** (dummies) para variáveis categóricas;
- **Padronização** dos preditores numéricos (média zero e desvio-padrão um), favorecendo estabilidade numérica e interpretação comparativa dos coeficientes.

A estratégia de reduzir a complexidade do espaço de preditores, tratar **esparsidade** e buscar **parcimônia** é coerente com as recomendações de modelagem de regressão logística em Hosmer, Lemeshow e Sturdivant (2013), que discutem

explicitamente a necessidade de **agregar categorias** diante de células com contagem zero e de conduzir a seleção de variáveis de forma criteriosa (abordagem “purposeful”), mitigando sobreajuste e instabilidade dos estimadores . A avaliação posterior do ajuste e da calibração, incluindo testes baseados em **decis de risco** (e.g., Hosmer–Lemeshow), também segue a literatura de referência para verificação global do modelo e identificação de discrepâncias sistemáticas entre probabilidades previstas e observadas.

3.8.1 Diagnóstico de multicolinearidade e ajustes no pré-processamento

Após a etapa de engenharia de variáveis, foi realizado um diagnóstico de **multicolinearidade** por meio da **matriz de correlação** entre preditores contínuos e do **Fator de Inflação da Variância (VIF)** no conjunto de treino. Inicialmente, observou-se **colinearidade severa** entre termos principais e interações criadas a partir deles. Em particular, as duplas “recency_c” ×” inter_recency_eng” e “tenure_months” × “inter_tenure_eng” apresentaram VIFs muito elevados (≈ **573** e **334**, respectivamente), acompanhados do aviso de quase-separação do glm (“probabilidades ajustadas 0 ou 1”). Esse padrão é típico quando **interações são construídas na mesma escala dos termos principais**, gerando dependência quase linear e instabilidade dos coeficientes.

Para mitigar o problema, foram implementados os seguintes **ajustes no pré-processamento**:

1. **Interações após centralização/padronização**
   As interações passaram a ser geradas **dentro do pipeline (recipe)** com step_interact() **depois** de step_normalize() nos contínuos. Essa ordem garante que os termos principais estejam **centrados e escalados** antes do produto, reduzindo a correlação entre “nível” e “inclinação” e, portanto, o VIF.
   – Consequência prática: as interações tenure_months × meetings_participated_pm, recency_c × meetings_participated_pm e fraction_paid × log_dti tornaram-se informativas **sem inflar** a variância dos coeficientes.

2. **Remoção das interações pré-calculadas no dado bruto**
   As colunas inter_tenure_eng, inter_recency_eng e inter_fraction_logdti previamente criadas fora do recipe foram **excluídas**. As mesmas interações passaram a ser geradas **apenas** no recipe, já com as variáveis **centradas**, evitando duplicidade e dependências desnecessárias.

3. **Codificação de categóricas em k−1 dummies (evitando armadilha da dummy)**

A codificação binária foi alterada para **k−1** (em lugar de *one-hot*), preservando o intercepto do modelo e evitando **combinações lineares exatas** entre dummies. Complementarmente, aplicou-se step_lincomb() para remover eventuais colinearidades remanescentes.
– Efeito: o erro “aliased coefficients” no cálculo de VIF foi eliminado.

4. **Tratamento de níveis raros e unificação de categorias “Outras”**
   A agregação de níveis de baixa frequência foi fortalecida com step_other(threshold = 0,08 e 0,10) nas **categóricas**, **exceto** em especialidade_macro (já possui “Outras” por design). Isso reduziu **níveis esparsos** e atenuou sinais de **quase-separação**, frequentes quando há categorias com pouquíssimas observações.

5. **Padronização apenas dos contínuos e limpeza de variáveis pouco informativas**
   Manteve-se a padronização **exclusivamente** nas variáveis contínuas (z-score), preservando as dummies em 0/1 para melhor interpretabilidade dos OR. Preditores com variância zero ou quase zero foram eliminados (step_zv/step_nzv), evitando ruído estrutural.

6. **(Complementar) Mitigação de assimetria/valores extremos**
   Embora não seja a causa direta da colinearidade, o *capping* em p99 de debt_to_income e recency, seguido de log1p(debt_to_income), foi mantido para reduzir a influência de outliers e melhorar a **linearidade do efeito** o que auxilia a **calibração** e diminui a chance de avisos de separação.

3.8.2 Evidência pós-ajustes

Após aplicar os ajustes acima, a matriz de correlação entre contínuos **não apresentou pares acima de 0,90** e os **VIFs** ficaram em patamar **aceitável** (máximo ≈ **7,55** para “meetings_participated_pm”, interações na faixa **5,16–5,55**, e demais abaixo disso). Em termos de regra de bolso, utilizou-se como referência **VIF < 10** (preferencial) e **ideal < 5**. Assim, o conjunto final apresenta **multicolinearidade controlada**, viabilizando a estimação estável dos coeficientes e a interpretação de *odds ratios* com maior confiabilidade.

3.9 Implicações para a modelagem

O modelo logístico pode ser ajustado com os termos principais (p.ex., tenure_months, meetings_participated_pm, recency_c, fraction_paid, log_dti) e as interações justificadas pela análise exploratória, sem inflar indevidamente a variância dos estimadores.O aviso residual de quase-separação pode ocasionalmente surgir por conta de alguma categoria rara, mas não inviabiliza o ajuste; caso necessário, há a alternativa de penalização ridge para robustez adicional, sem alterar a especificação substantiva do modelo.

Síntese. As verificações sistemáticas de correlação e VIF orientaram ajustes estruturais no pré-processamento, centralização antes das interações, dummies em k−1, remoção de combinações lineares e *lumping* de níveis raros, que **reduziram a multicolinearidade** a um nível aceitável. Com isso, o conjunto de dados tornou-se **adequado** para a regressão logística, preservando tanto o **poder preditivo** quanto a **interpretação** dos efeitos em termos de *odds ratio*.

3.10 Modelagem

3.10.1 Formulação do Modelo

A fórmula do modelo logístico foi dada por:

$$status = \beta_0 + \beta_1(Tempo\ de\ Mentoria) + \beta_2(Engajamento\ médio\ mensal) + \beta_3(recency\_c) + \beta_4(fraction\_paid) + \beta_5(log\ dti) + \beta_6(variáveis\_dummies)$$

Explicação da Fórmula:

- **status**: A variável dependente, que representa o **status do cliente** (0 = ativo, 1 = churn).
- β0: O **intercepto** do modelo, que representa o valor de status\text{status}status quando todas as variáveis explicativas são zero.=17.6
- β1,β2,…,β6: São os **coeficientes das variáveis explicativas**, que indicam a **influência** de cada variável na probabilidade de churn.
- **variáveis explicativas**:
    - **tenure_months**: O tempo de permanência do cliente no serviço (em meses).
    - **meetings_participated_pm**: Participação nas reuniões por mês (representando o engajamento).
    - **recency_c**: Recência da última reunião (em dias).
    - **fraction_paid**: Fração paga da mentoria.
    - **log_dti**: Logaritmo da razão entre dívida e renda, representando a pressão financeira.
    - **variáveis_dummies**: Variáveis binárias (dummies), como sexo, estado civil e especialidade médica do cliente.

Essa fórmula expressa a **regressão logística**, onde os coeficientes β representam os **efeitos marginais** de cada variável na probabilidade de churn.

3.10.2 Ajustes no Modelo

Durante o ajuste **do modelo**, foi identificado um **aviso de quase-separação**, o que pode ocorrer quando há **categorias raras** no conjunto de dados. Esse tipo de problema pode afetar a estimativa dos coeficientes, mas, no caso deste estudo, não impediu o ajuste do modelo. A solução para esse tipo de problema foi considerar a **penalização ridge**, que ajuda a **evitar overfitting** e **garante maior robustez** ao modelo.

3.10.3 Validação do Modelo

Após o ajuste do modelo, a **validação do modelo** foi avaliada utilizando diversas métricas, incluindo:

- **AUC (Área sob a curva ROC)**: Utilizada para avaliar a **discriminação** do modelo, ou seja, sua capacidade de separar corretamente os clientes que deram churn e os ativos.

- **Matriz de Confusão**: Foi utilizada para verificar **sensibilidade**, **especificidade** e **acurácia** do modelo.

- **Brier Score e Hosmer-Lemeshow**: Mediram a **calibração do modelo**, ou seja, o quanto as **probabilidades previstas** se alinham com as **probabilidades reais**.

O **corte ótimo** foi determinado utilizando o **método de Youden** na **curva ROC**, garantindo o melhor equilíbrio entre **sensibilidade** e **especificidade**.

3.10.4 Interpretação dos Resultados

Os coeficientes do modelo logístico fornecem **odds ratios (OR)**, que indicam como cada variável influencia a **probabilidade de churn**. As variáveis **recency_c** (recência da última reunião) e **tenure_months** (tempo de mentoria) apresentaram **odds ratios altamente significativos**, indicando que o **engajamento** do cliente com o serviço e o **tempo de relacionamento** são os maiores determinantes para a decisão de cancelar ou continuar com a mentoria. Já a **fração paga da mentoria** foi significativa, mas com um efeito menos acentuado, sugerindo que os clientes que pagam mais por suas mentorias são **menos propensos ao churn**.

6 RESULTADOS E DISCUSSÃO CIRCUNSTANCIADA

6.1 Análise exploratória dos dados

A Análise Exploratória de Dados (AED) será realizada após o tratamento inicial e a engenharia de features. O objetivo da AED é compreender a estrutura dos dados, identificar padrões, tendências, anomalias e relações entre as variáveis. Esta etapa é crucial para informar as decisões de modelagem e para aprofundar o entendimento do fenômeno de Churn na empresa.

6.1.1 ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS DISCRETAS

ESTATÍSTICA DAS VARIÁVEIS DISCRETAS

Neste momento, as estatísticas de cada variável serão analisadas, incluindo média, mediana, desvio padrão e os quartis.

Tabela 2 - Estatística das variáveis discretas

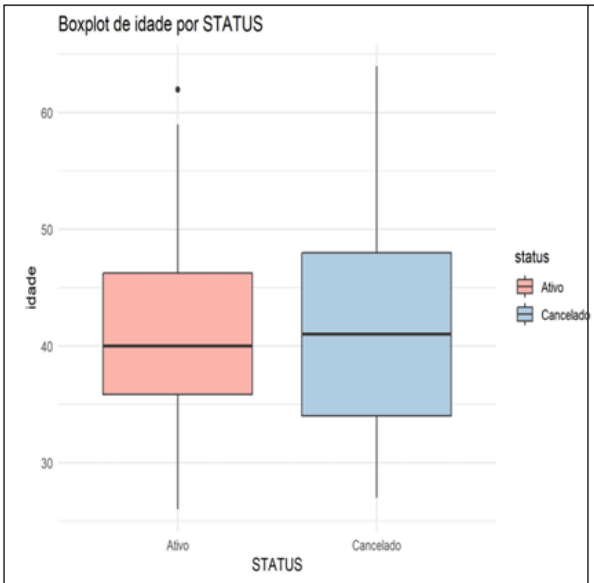
Variável	Média	Desvio Padrão	Mínimo	Q1 25%	Mediana	Q2 75%	Máximo	Valores ausentes
Idade	41,618	8,506	26	35	40	47	64	0
Quantidade de parcelas pagas	8,018	5,461	0	2	9	13	26	0
Quantidade de reuniões	10,530	6,568	0	5	11	15	24	0
Quantidade de reuniões canceladas	0,601	1,021	0	0	0	1	6	0
Quantidade de produtos contratados	3,774	2,475	0	2	4	5	13	0

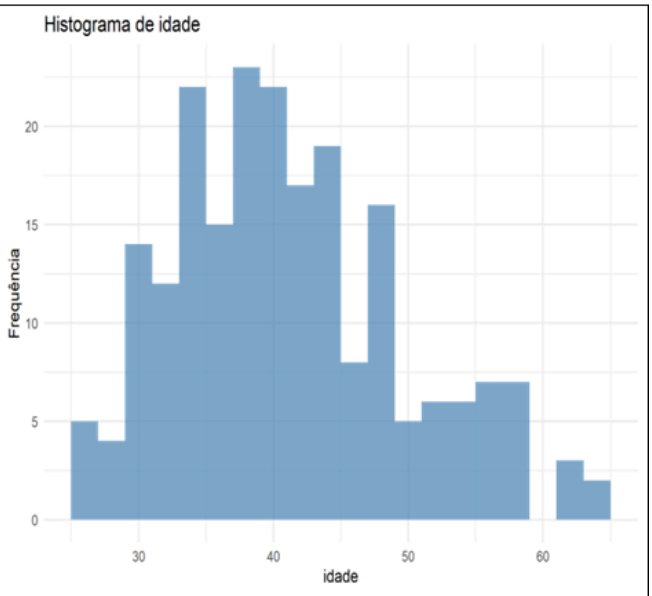
Fonte: Elaborado pelo autor (2025).

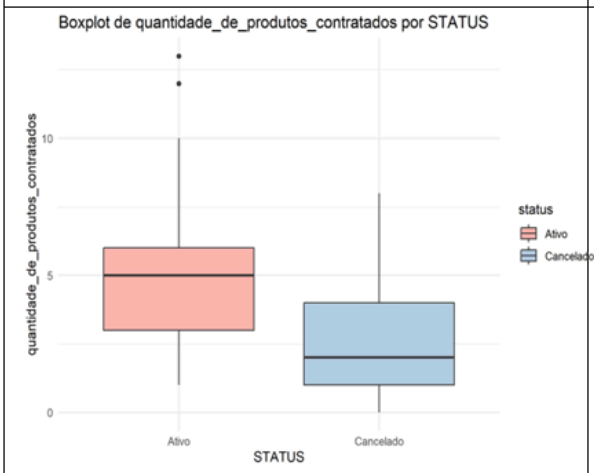
Nesta seção, foram analisadas as variáveis discretas relacionadas às características e ao comportamento dos clientes: idade, quantidade de parcelas pagas da mentoria, quantidade de reuniões, reuniões canceladas e produtos contratados. As

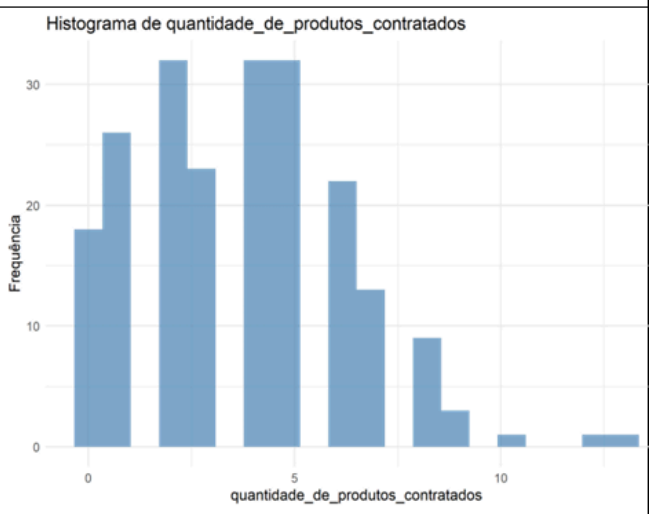
análises foram realizadas por meio de histogramas e boxplots estratificados pela variável resposta ”status”, com as categorias "Ativo" e "Cancelado".

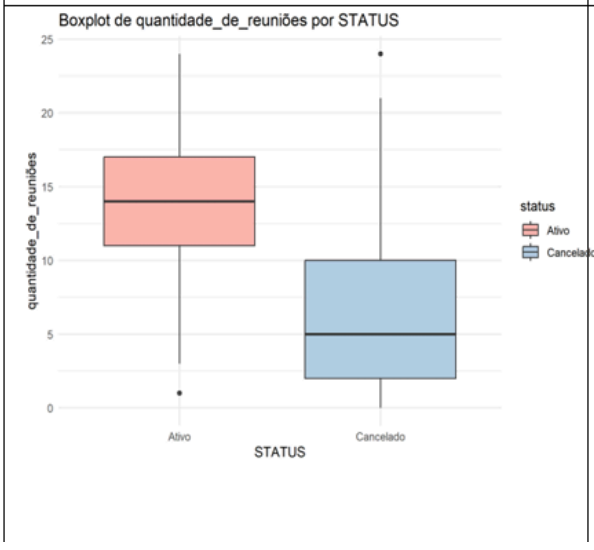
Tabela 3 - EDA das variáveis discretas

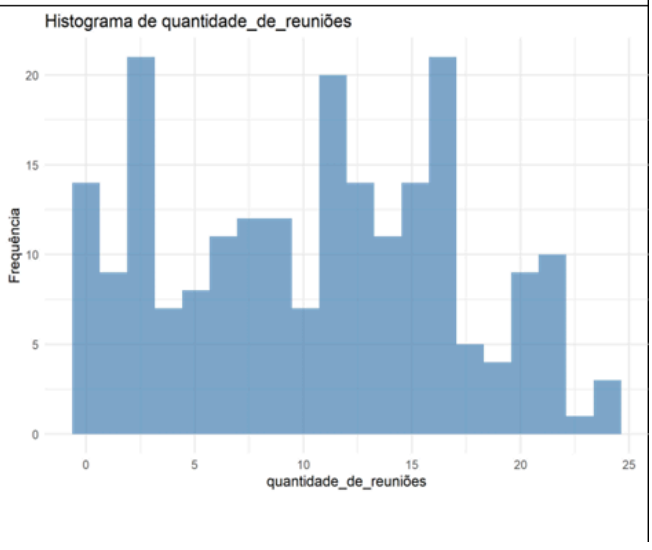

Boxplot de idade por STATUS
idade
60
50
40
30
Ativo
Cancelado
STATUS
status
Ativo
Cancelado

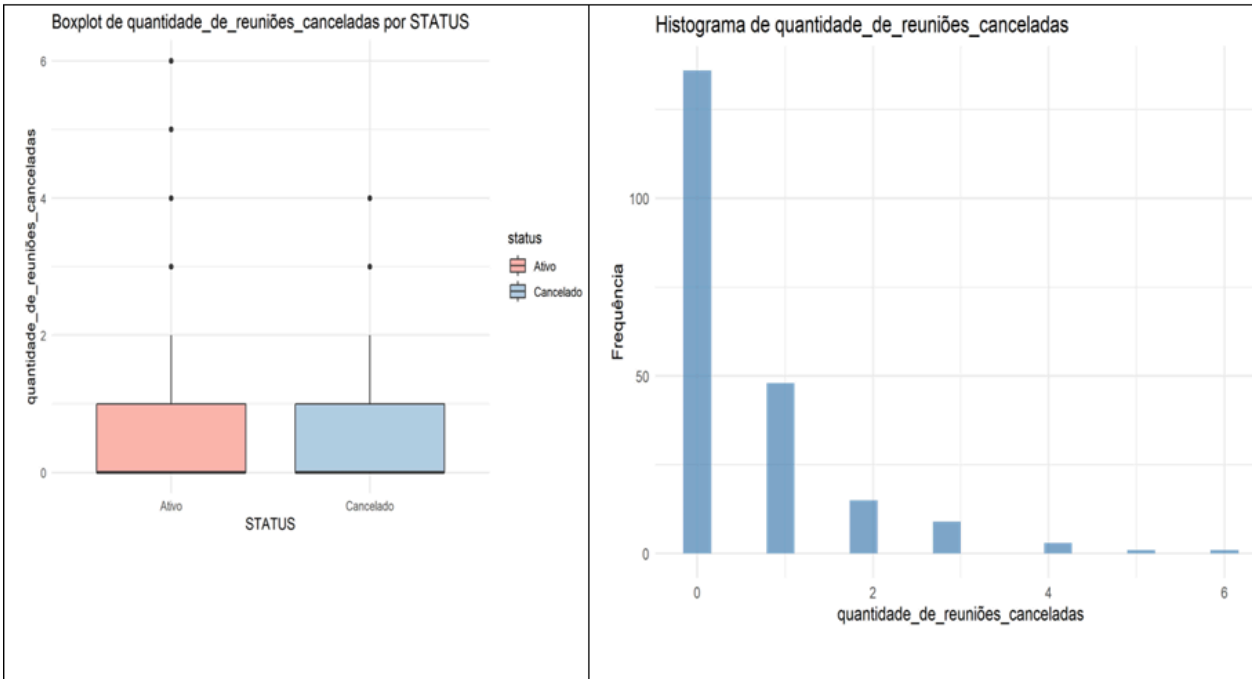

Histograma de idade
Frequência
20
15
10
5
0
30
40
50
60
idade


Boxplot de quantidade_de_produtos_contratados por STATUS
quantidade_de_produtos_contratados
10
5
0
Ativo
Cancelado
STATUS
status
Ativo
Cancelado


Histograma de quantidade_de_produtos_contratados
Frequência
30
20
10
0
0
5
10
quantidade_de_produtos_contratados


Boxplot de quantidade_de_reuniões por STATUS
quantidade_de_reuniões
25
20
15
10
5
0
Ativo
Cancelado
STATUS
status
Ativo
Cancelado


Histograma de quantidade_de_reuniões
Frequência
20
15
10
5
0
0
5
10
15
20
25
quantidade_de_reuniões



Fonte: Elaborado pelo autor (2025).

6.1.2 ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS CONTÍNUAS

ESTATÍSTICAS DAS VARIÁVEIS CONTÍNUAS

As variáveis contínuas incluídas na análise exploratória foram:

- renda_inicial: A renda mensal do cliente no início da mentoria.
- divida_total: Dívida total do cliente.
- valor_da_mentoria: Valor pago pela mentoria.
- tempo_de_mentoria: Tempo em meses que o cliente esta ativo na mentoria
- scoreultimo_nps: A nota de satisfação do cliente mais recente.

As estatísticas de cada variável serão analisadas, incluindo média, mediana, desvio padrão e os quartis, conforme apresentado na Figura 2.

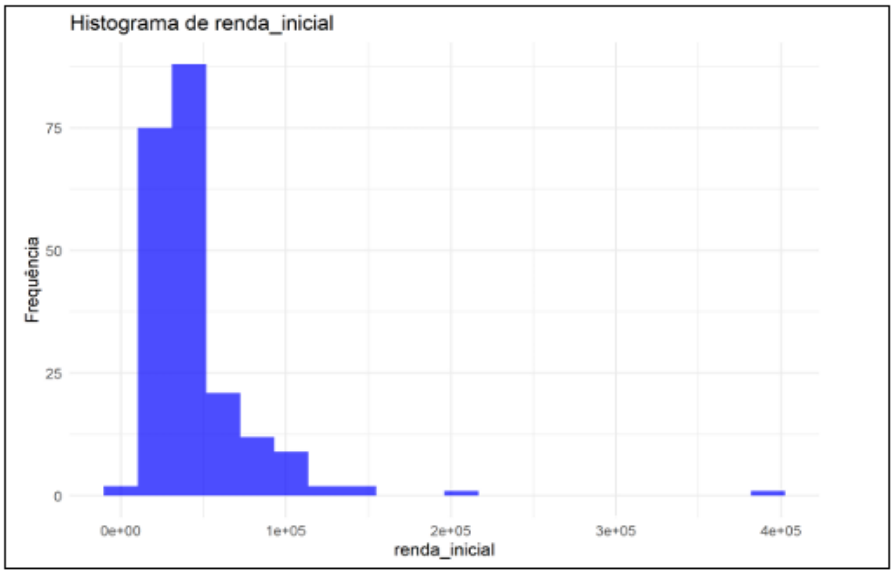
Tabela 4 - Estatística das variáveis contínuas

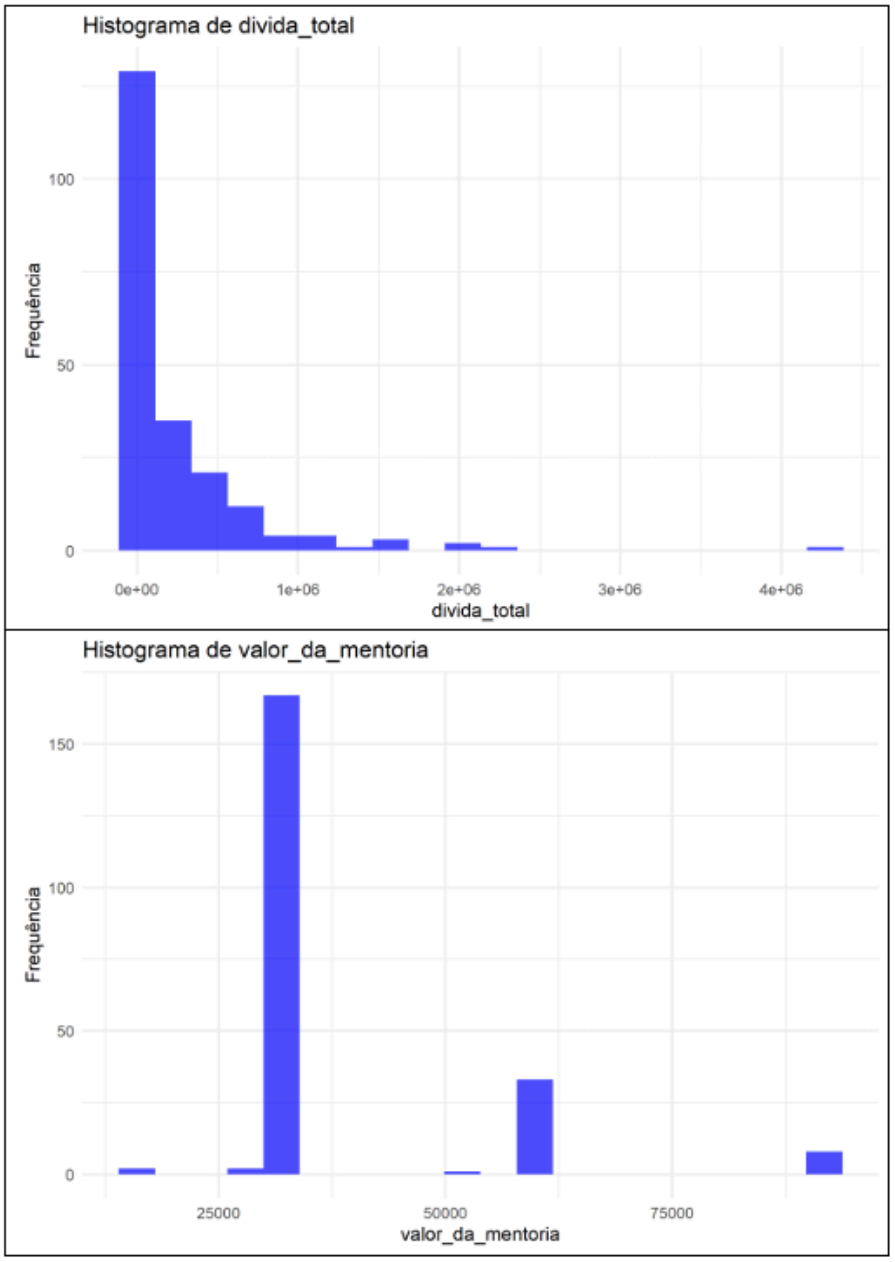
Variável	Média	Desvio Padrão	Mínimo	Q1 25%	Mediana	Q2 75%	Máximo	Valores ausentes
Renda Inicial	46.119,69	36279,78	8.000	28000	37000	50000	400.000	0
Divida Total	234.835,00	479.366,10	0	0	0	299861,93	4.271.719,00	0
Valor da mentoria	38.757,98	14417,07	14200	32500	32500	32500	90.000	0
Tempo de mentoria	11,71	7,015	0	6,067	13	15,475	28,167	1
Último score no NPS	9,38	1,334	5	9,25	10	10	10	195

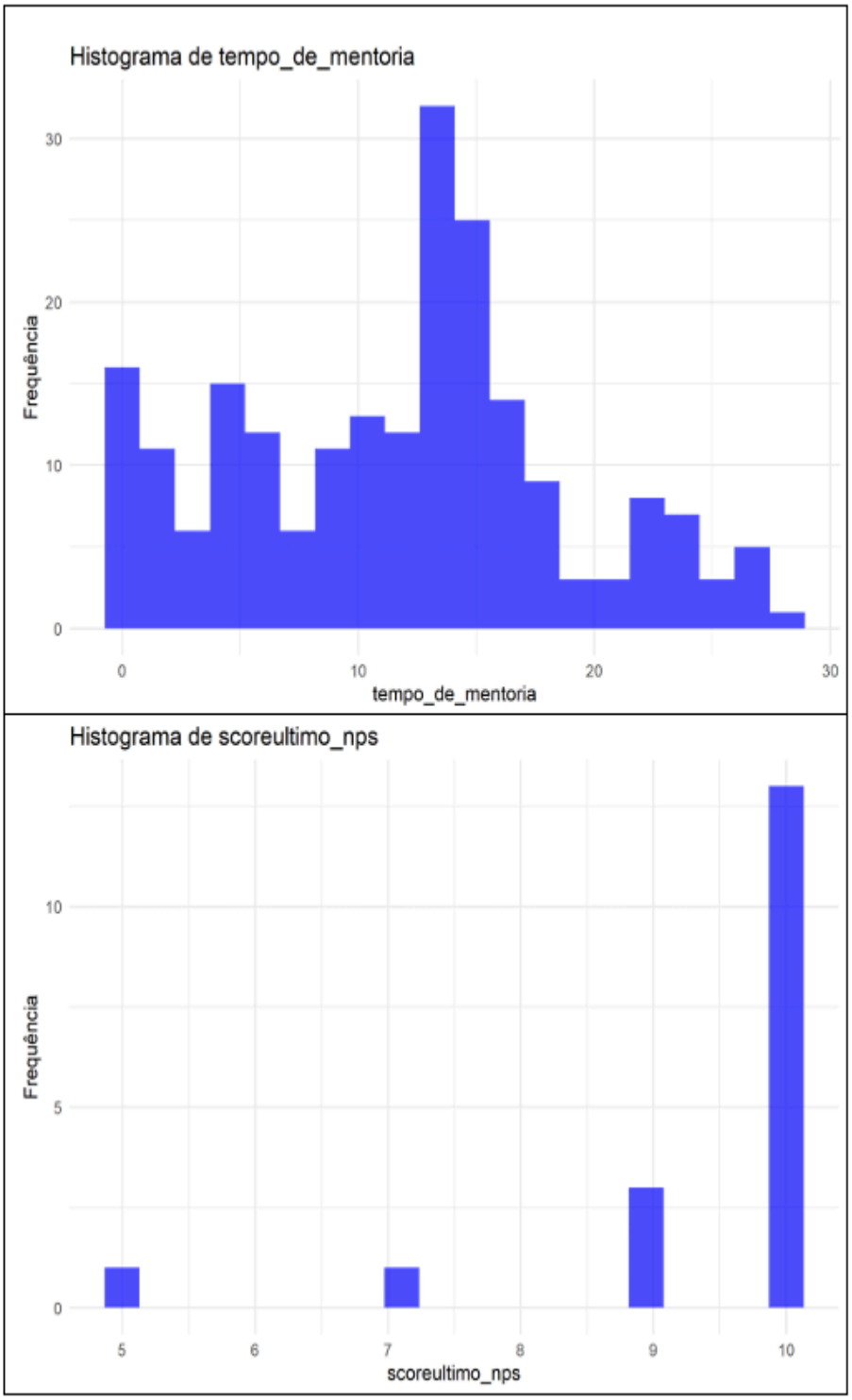
Fonte: Elaborado pelo autor (2025).

Estas variáveis estão diretamente ligadas ao perfil financeiro e ao relacionamento dos clientes com o serviço, sendo avaliadas por meio de histogramas e análise de correlação:

Tabela 5 – EDA das variáveis discretas 1




Histograma de divida_total
Frequência
100
50
0
0e+00
1e+06
2e+06
3e+06
4e+06
divida_total
Histograma de valor_da_mentoria
Frequência
150
100
50
0
25000
50000
75000
valor_da_mentoria

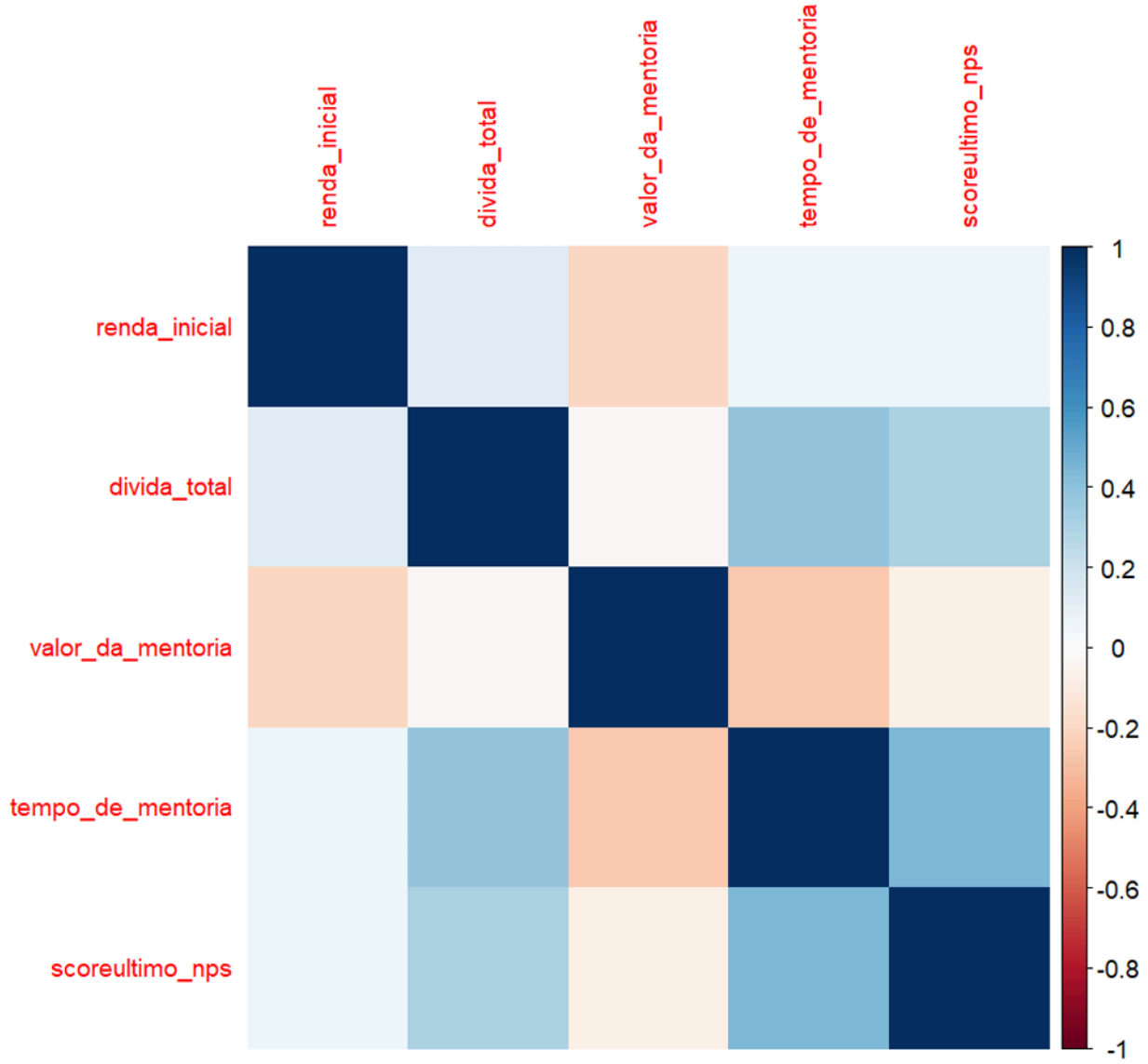


Fonte: Elaborado pelo autor (2025).

6.1.3 Matriz de Correlação

A matriz de correlação entre as variáveis contínuas revelou fracas associações diretas entre elas. Isso sugere que cada variável pode adicionar informação única ao modelo preditivo. A ausência de colinearidade significativa é um bom indicativo para a inclusão simultânea dessas variáveis na regressão logística. Como mostra na Figura 3:

Figura 2 – Matriz de Correlação

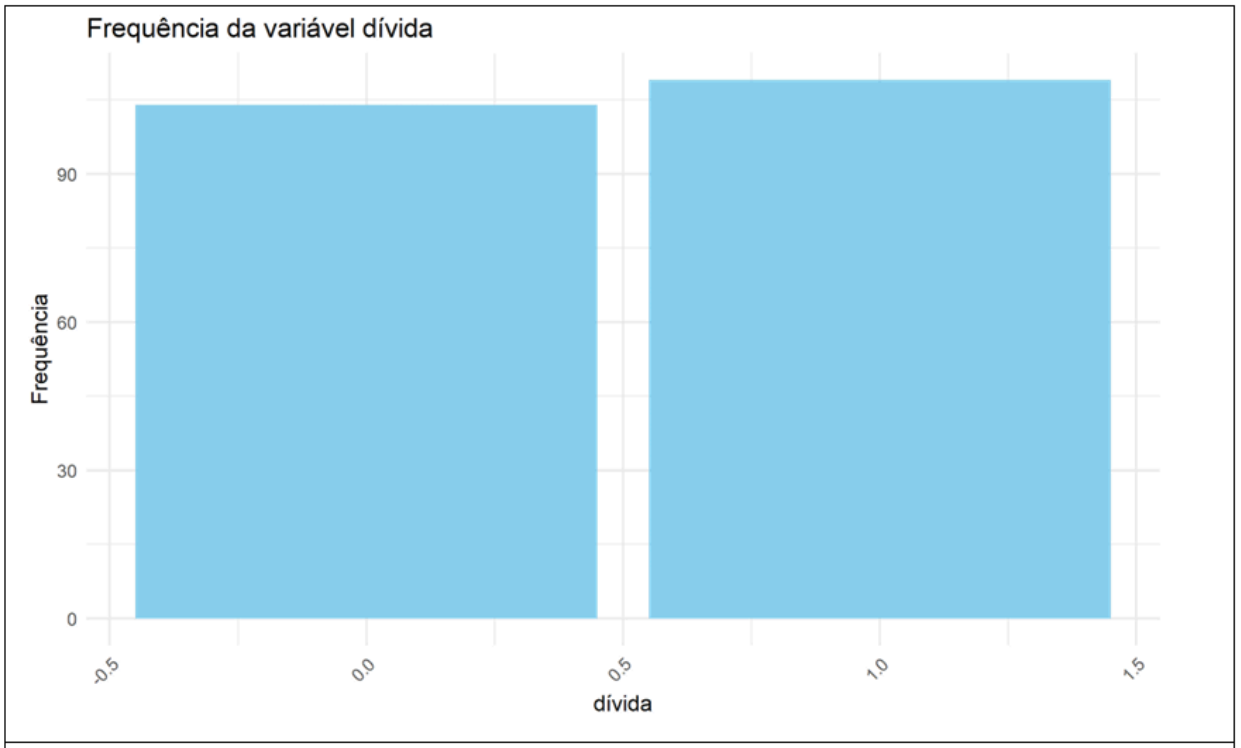


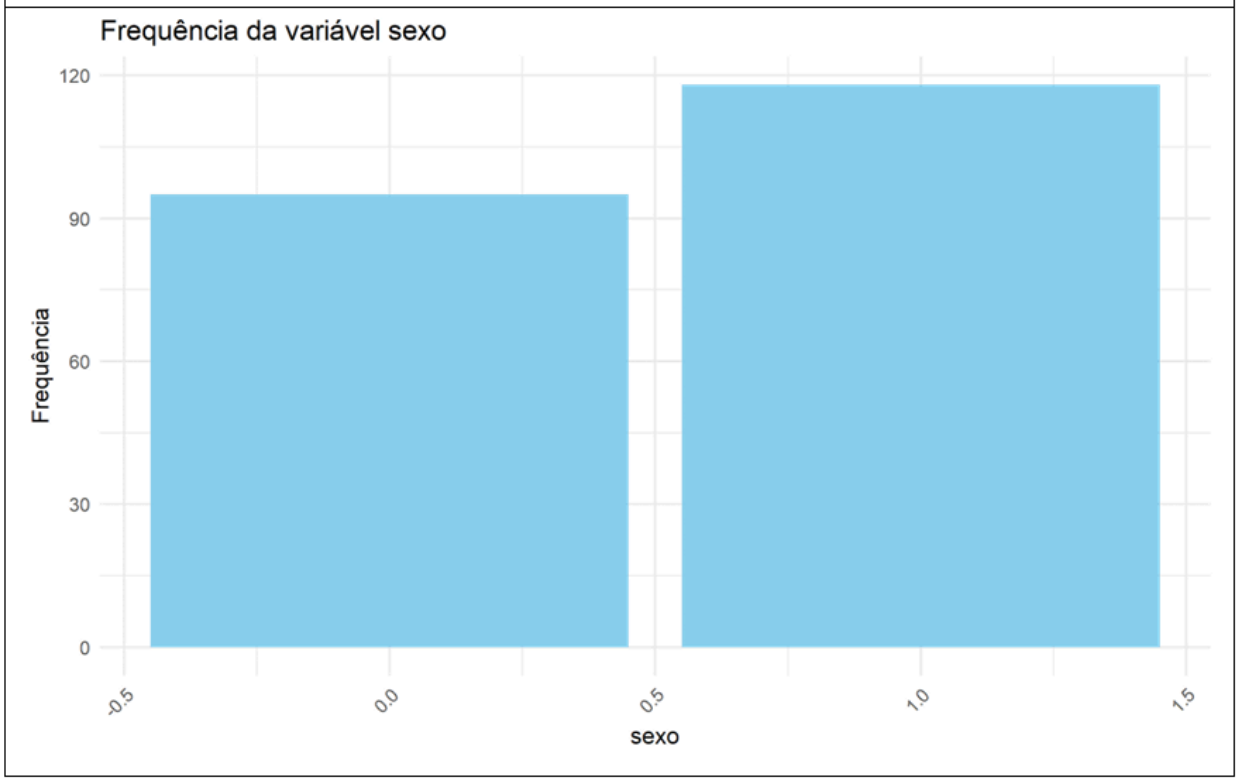
Fonte: Elaborado pelo autor (2025).

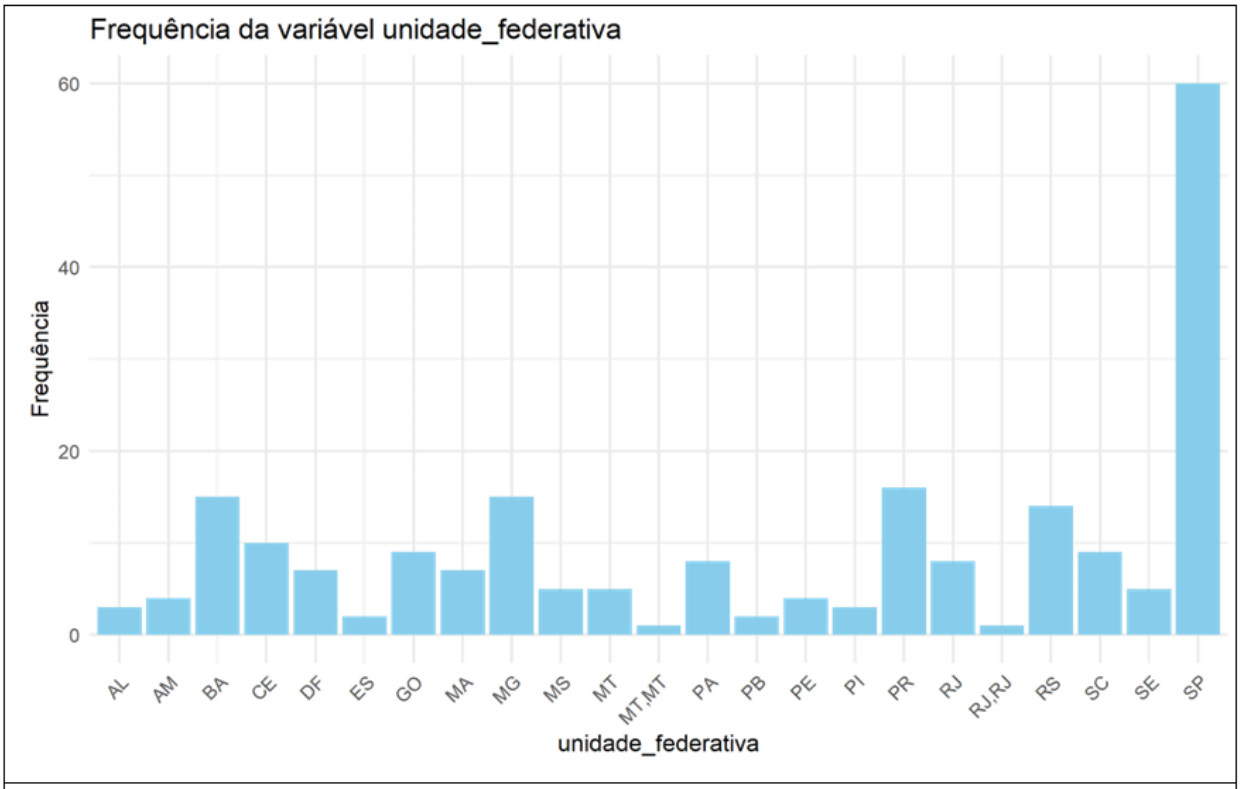
6.1.4 ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS QUALITATIVAS

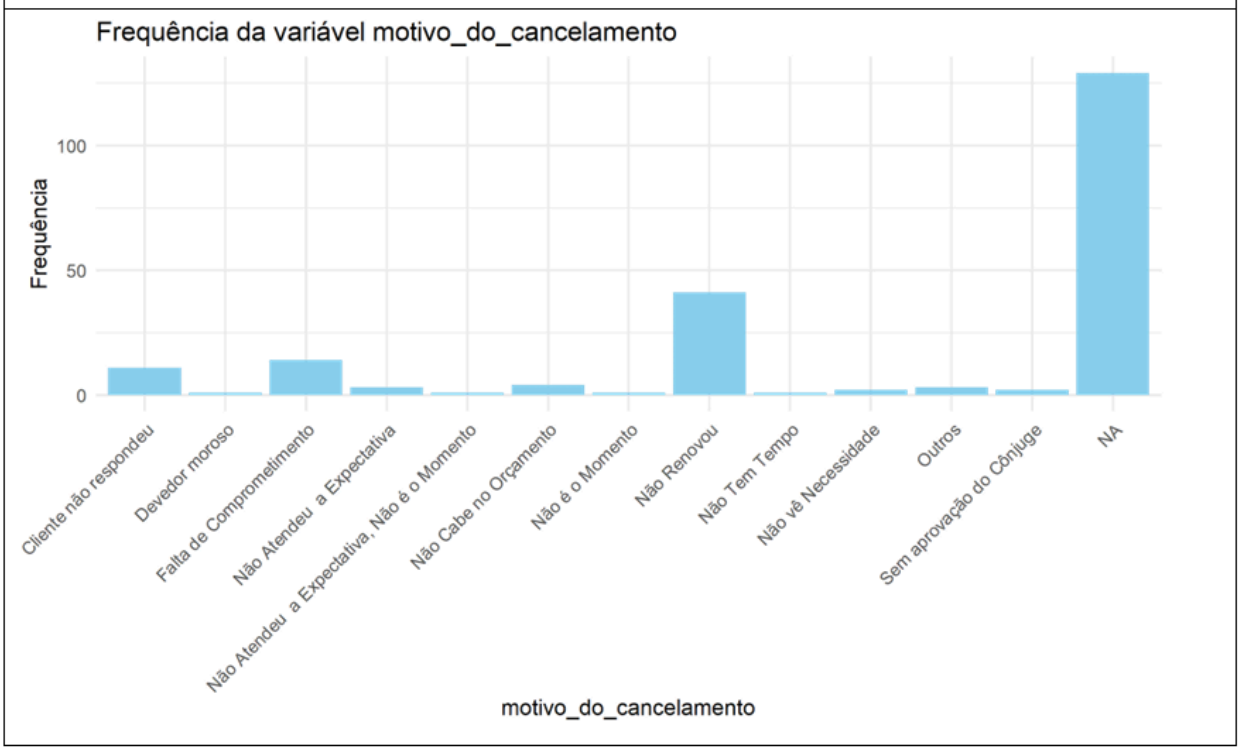
Esta seção apresenta a análise das variáveis qualitativas categóricas do conjunto de dados, com o objetivo de compreender a distribuição das categorias e sua relevância para o entendimento do churn. Foram incluídas as variáveis: dívida, sexo, unidade_federativa, estado_civil e motivo_do_cancelamento. A visualização foi realizada por meio de gráficos de barras (frequência absoluta).

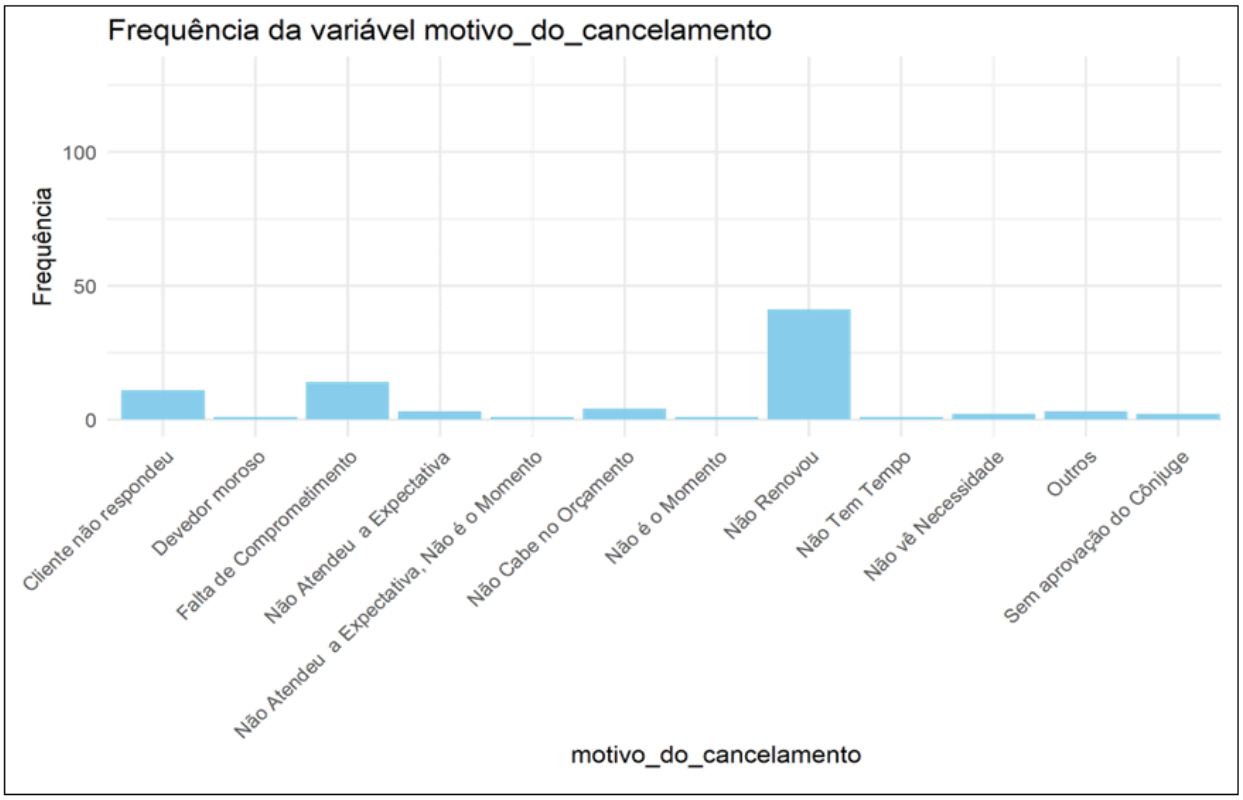
Tabela 6 - Análise das variáveis






Frequência da variável unidade_federativa
Frequência
60
40
20
0
AL
AM
BA
CE
DF
ES
GO
MA
MG
MS
MT
MT,MT
PA
PB
PE
PI
PR
RJ
RJ,RJ
RS
SC
SE
SP
unidade_federativa


Frequência da variável motivo_do_cancelamento
Frequência
100
50
0
Cliente não respondeu
Devedor moroso
Falta de Comprometimento
Não Atendeu a Expectativa
Não Atendeu a Expectativa, Não é o Momento
Não Cabe no Orçamento
Não é o Momento
Não Renovou
Não Tem Tempo
Não vê Necessidade
Outros
Sem aprovação do Cônjuge
NA
motivo_do_cancelamento



Fonte: Elaborado pelo autor (2025).

A análise exploratória inicial revelou que as variáveis originais disponibilizadas pela empresa, como **renda inicial, quantidade de reuniões, tempo de mentoria**, informações demográficas e financeiras, apresentavam **pouca capacidade discriminativa** para diferenciar clientes **ativos** e **churn**. Em diversas variáveis analisadas, os histogramas, boxplots e medidas de tendência central mostraram **sobreposição intensa** entre os grupos, indicando que os preditores brutos não capturavam adequadamente os padrões comportamentais associados ao cancelamento.

Diante desse diagnóstico, tornou-se necessária uma etapa aprofundada de **engenharia de dados**, com o objetivo de construir variáveis com **significado estatístico mais forte** e melhor alinhadas às hipóteses de churn na literatura.

6.2 Variáveis pós processamento

Tabela 7 - Nome da variável x transformação realizada x função.

Variável no modelo	Transformação realizada	Informação que a variável traz
status	Variável alvo binária construída a partir do campo *status (churn = 1)* da base original. Não é padronizada.	Indica se o cliente **cancelou a mentoria (1)** ou se permanece **ativo (0)**.
idade	Copiada da idade original e depois **centralizada e padronizada** pelo recipe.	Idade do cliente no momento da contratação da mentoria. Relação com churn pode indicar perfis etários mais ou menos propensos a cancelar.
tenure_months	Calculada a partir do **tempo, em dias, desde o início da mentoria**, convertida para meses. Depois **padronizada**.	Tempo de relacionamento do cliente com a mentoria (em meses). Captura efeitos de “curva de aprendizagem” e de maturidade do contrato sobre o churn.
meetings_participated_pm	A partir de **quantidade de reuniões realizadas** e do **tempo de mentoria**, foi calculada a média de reuniões **por mês**; depois padronizada.	Mede o **engajamento médio mensal** do cliente nas reuniões de mentoria. Valores maiores indicam maior participação.

recency_c	A partir da **data da última reunião com o cliente** foi calculado o número de **dias desde a última reunião**; depois a variável foi **centrada (e padronizada)**, gerando recency_c.	Mede a **recência de contato** com o cliente: quanto maior, mais tempo o cliente está sem se reunir com o mentor, o que se associa a desengajamento e maior risco de churn.
fraction_paid	Construída como **(nº de parcelas pagas da mentoria) / (nº total de parcelas contratadas)**. Posteriormente padronizada.	Representa a **fração do contrato já quitada**. Valores próximos de 1 indicam clientes próximos do fim da mentoria; valores baixos indicam contratos ainda no início.
log_dti	A partir de **Divida_total** e **Renda_inicial**, foi calculado o *debt-to-income ratio* (debt_to_income = divida_total / renda_inicial). Em seguida foi aplicado **logaritmo** para reduzir assimetria (**log_dti**) e depois padronizado.	Pressão financeira do cliente: razão entre dívida total e renda. Valores altos indicam maior comprometimento da renda com dívidas, podendo aumentar a probabilidade de cancelamento.
has_debt	Indicador criado a partir de Divida_total: **1 se divida_total > 0**, **0 se divida_total = 0**. Depois foi centralizado/padronizado, mas semanticamente continua sendo um dummy.	Indica se o cliente possui **alguma dívida registrada** no momento da contratação. Diferencia clientes endividados de clientes sem dívidas.

has_cross_sell	Criada a partir da **Quantidade_produtos** além da mentoria: **1 se o cliente contratou ≥ 1 produto adicional**, **0 caso contrário**. Depois padronizada.	Mede se o cliente possui **produtos adicionais** com a empresa (cross-sell). Sinaliza maior profundidade de relacionamento e potencialmente menor risco de churn.
sexo_Masculino	A partir da variável categórica “sexo” foi criada uma dummy: **1 = masculino**, **0 = feminino**. Em seguida a dummy foi padronizada.	Indica o sexo do cliente. Permite avaliar diferenças de churn entre clientes homens e mulheres.
estado_civil_Solteiro.a.	Do estado civil original foi criada uma variável categórica; o recipe gerou dummies. Essa coluna representa **1 = cliente solteiro(a)**, 0 caso contrário, depois padronizada.	Identifica clientes solteiros, permitindo avaliar se o estado civil está associado à probabilidade de cancelamento.
estado_civil_other	Dummy criada para agrupar **demais categorias de estado civil** que não são “solteiro(a)” nem a categoria de referência (casado). Padronizada.	Agrupa estados civis menos frequentes (divorciado, separado etc.), capturando efeitos desses perfis sem gerar muitas dummies com baixa frequência.
UF_macro_CentroOeste	A partir da UF original, os estados foram agregados em **macrorregiões** (Norte, Nordeste, Centro-Oeste, Sudeste, Sul). O recipe criou dummies; esta é **1 = cliente da região Centro-Oeste**, 0 caso contrário, padronizada.	Indica se o cliente é da **região Centro-Oeste**. Permite captar diferenças regionais de comportamento e churn.
UF_macro_Sudeste	Dummy gerada para a **região Sudeste**; padronizada.	Indica se o cliente reside na região

		Sudeste, principal mercado da empresa.
UF_macro_Sul	Dummy gerada para a **região Sul**; padronizada.	Indica se o cliente é da região **Sul** do país.
especialidade_macro_Cardiologia	A especialidade médica original foi recodificada em grupos macro (ClinicoGeral, Cardiologia, Pediatria, GinecoObst, CirOrtop, Gastro, Urologia, Psiquiatria, Oftalmo, Nefrologia, Outras). O recipe gerou dummies; esta é **1 = cardiologista**, 0 caso contrário, padronizada.	Identifica clientes da área de **Cardiologia**, grupo com perfil de renda/agenda específico e possível padrão próprio de churn.
especialidade_macro_Pediatria	Dummy para a macroespecialidade **Pediatria**; padronizada.	Indica se o cliente é **pediatra**.
especialidade_macro_GinecoObst	Dummy para a macroespecialidade **Ginecologia/Obstetrícia**; padronizada.	Identifica médicas(os) de **ginecologia e obstetrícia**, permitindo verificar se esse grupo se comporta de forma distinta em relação ao churn.
especialidade_macro_CirOrtop	Dummy para o grupo **Cirurgia/Ortopedia** (cirurgia geral, ortopedia e traumatologia, mastologia etc.); padronizada.	Representa especialidades predominantemente **cirúrgicas**, que podem ter padrão de renda e agenda diferente dos clínicos.
especialidade_macro_Nutrologia	Dummy para **Nutrologia**; padronizada.	Indica se o cliente atua em **Nutrologia**, especialidade mais

		voltada à gestão nutricional/metabólica.
especialidade_macro_Intensivista	Dummy para **Medicina Intensiva/UTI**; padronizada.	Identifica médicos **intensivistas**, com rotina e renda potencialmente distintas, o que pode influenciar o risco de churn.
especialidade_macro_Outras	Dummy que agrega as **demais especialidades** não incluídas nas categorias explícitas; padronizada.	Agrupa especialidades menos frequentes, garantindo representatividade sem inflar o número de variáveis.
tenure_months_x_meetings_participated_pm	Variável de **interação** criada como o produto: tenure_months * meetings_participated_pm (antes da padronização); depois foi padronizada junto com as demais.	Captura o efeito conjunto de **tempo de mentoria** e **engajamento**. Distingue, por exemplo, clientes antigos pouco engajados de clientes novos muito engajados.
meetings_participated_pm_x_recency_c	Interação calculada como meetings_participated_pm * recency_c e, em seguida, padronizada.	Mede como a **combinação entre participação média** e **recência da última reunião** afeta o risco de churn (por exemplo, cliente que sempre participou muito, mas parou recentemente).

fraction_paid_x_log_dti	Interação calculada como fraction_paid * log_dti e depois padronizada.	Relaciona o **estágio do contrato** (fração paga) com a **pressão financeira**. Permite avaliar, por exemplo, se clientes muito endividados próximos ao fim da mentoria estão mais propensos a cancelar.

Fonte: Elaborado pelo autor (2025).

6.3 Segunda Análise Exploratória de Dados (AED)

A análise exploratória foi conduzida em duas etapas: primeiro, com as **variáveis originais da empresa**, e posteriormente com as **variáveis ajustadas** após a **engenharia de dados**. O objetivo dessa análise foi **entender as características das variáveis** e identificar quais delas seriam mais informativas para a modelagem de churn. A seguir, apresentamos os resultados dessa análise, incluindo as **estatísticas descritivas** e a **distribuição** das variáveis.

6.3.1 Estatísticas Descritivas das Variáveis Numéricas

Para entender melhor as variáveis numéricas, calculamos as **médias**, **medianas** e **distribuições** dessas variáveis para todo o conjunto de dados. A tabela a seguir apresenta as **estatísticas descritivas** das variáveis numéricas importantes para a modelagem.

Tabela 8 – Estatísticas Descritivas das Variáveis Numéricas

Variável	Média	Mediana
Idade	41,89	41
Tempo de Mentoria (tenure_months)	11,56	12,78

Participação_em_Reuniões (meetings_participated_pm)	0,93	1
Recência_da_Última_Reunião (recency_last_meeting_days)	215,89	147,50
Fração Paga (fraction_paid)	0,59	0,67
Dívida sobre Renda (debt_to_income)	5,53	0

As variáveis mostram uma distribuição de dados razoavelmente **simétrica**, com exceção de **recency_last_meeting_days** (dias desde a última reunião), que possui um valor médio relativamente alto, o que pode indicar a **presença de clientes com longos períodos de desengajamento**.

6.3.2 Comparação de Médias entre Clientes Ativos e Churn

Em seguida, calculamos a **média** das variáveis numéricas para os dois grupos: **clientes ativos** e **clientes churn**. Os resultados mostraram algumas diferenças importantes entre os dois grupos, que foram visualizadas nas **tabelas** e **gráficos a seguir**.

Tabela 9 – Comparação entre Ativos e Churn

Variável	**Ativo (média)**	**Churn (média)**
Idade	40,5	43,3
Tempo de Mentoria (tenure_months)	14,0	8,97
Participação em Reuniões (meetings_participated_pm)	0,93	0,93
Recência da Última Reunião (recency_last_meeting_days)	118,0	320,0
Fração Paga (fraction_paid)	0,67	0,50

Dívida sobre Renda (debt_to_income)	0,00	0,12

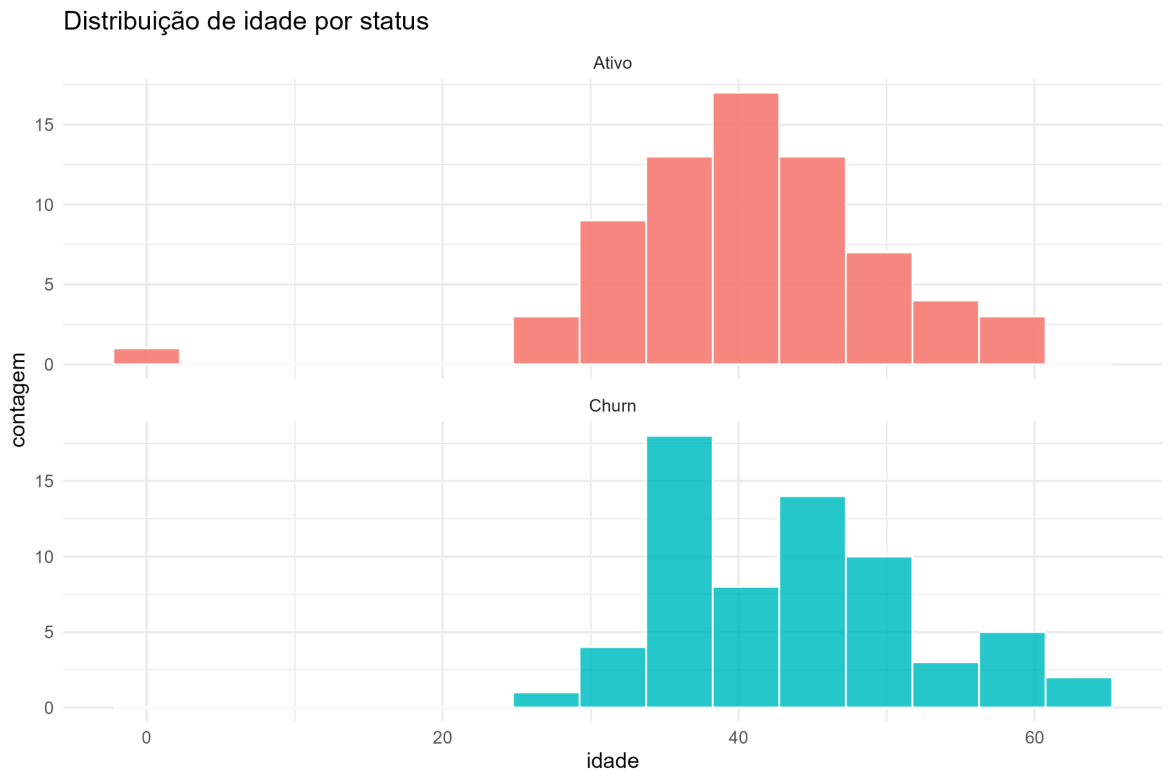
Fonte: Elaborado pelo autor (2025)

Análise dos Resultados: A principal diferença entre **clientes ativos** e **clientes churn** foi observada na **recência** e no **tempo de mentoria**. Os **clientes ativos** têm **maior tempo de mentoria** (média de 14 meses contra 8,97 meses dos churns), e uma **menor recência** (média de 118 dias contra 320 dias de desengajamento dos churns). Essas variáveis mostram que os clientes **mais engajados e com maior tempo de permanência** têm menos chances de churn. Além disso, a **fração paga da mentoria** é **mais alta entre os clientes ativos**, indicando que o comprometimento financeiro é um forte indicador de **retenção**.

6.4 Gráficos da Análise Exploratória

A seguir, apresentamos os gráficos gerados durante a análise exploratória para **visualizar a distribuição das variáveis** e a **comparação entre clientes ativos e churn**.

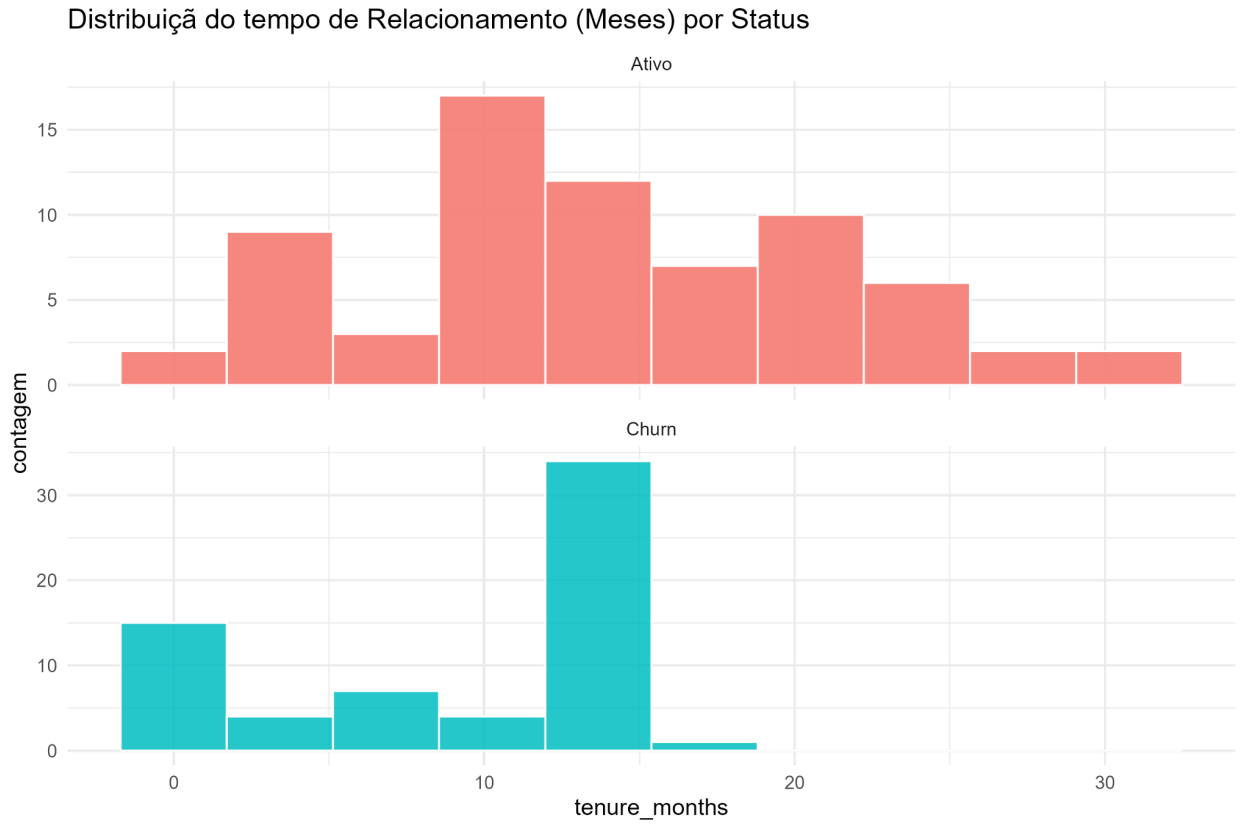
Figura 3 – Distribuição de Idade entre Clientes Ativos e Churn



Fonte: Elaborado pelo autor (2025)

Comentário: A variável **idade** não apresenta grandes diferenças entre os grupos, mas a **média** foi ligeiramente mais alta entre os clientes churn (43,3 anos) em comparação aos ativos (40,5 anos), o que sugere que **clientes mais velhos** podem ter um comportamento mais propenso ao **cancelamento**.

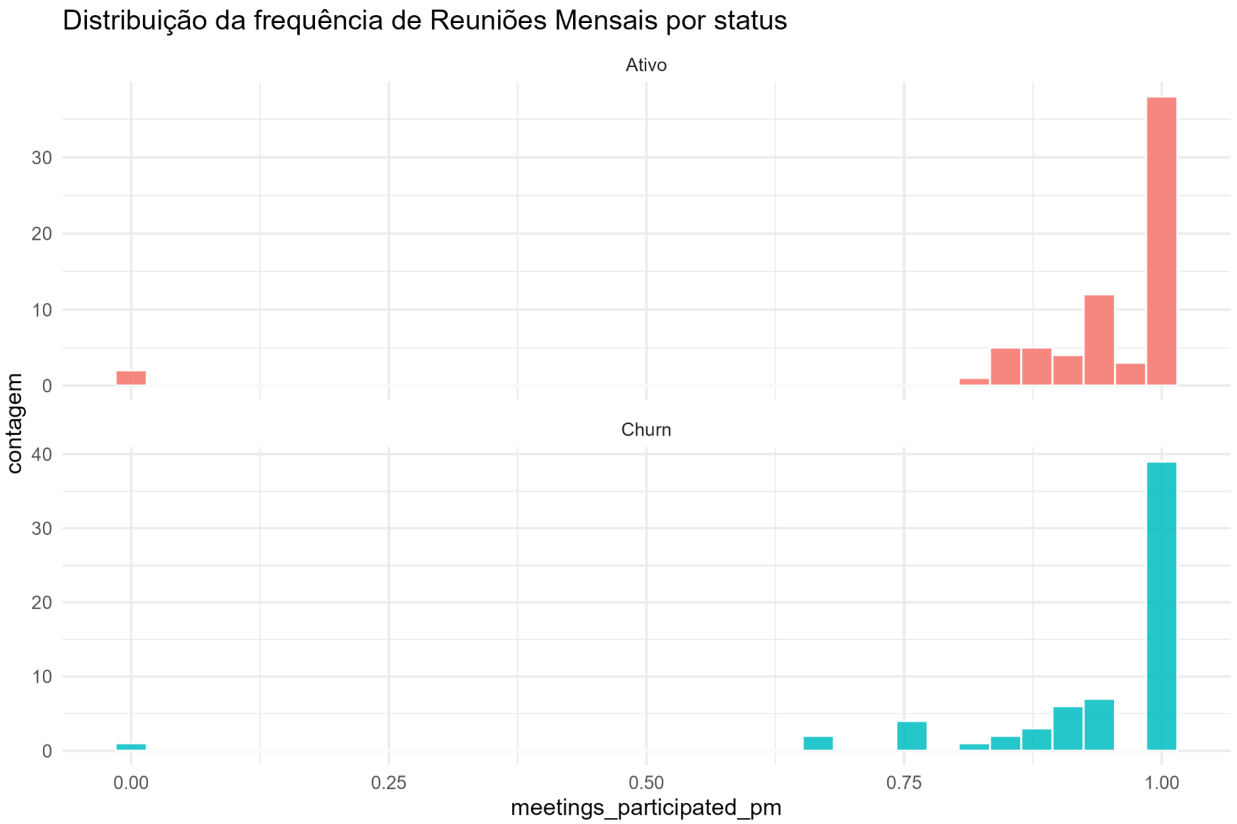
Figura 5 – Distribuição do Tempo de Mentoria entre Clientes Ativos e Churn



Fonte: Elaborado pelo autor (2025)

Comentário: O gráfico confirma que **clientes ativos** possuem um tempo de mentoria **mais longo** em comparação aos churns, o que é esperado, pois o **tempo de relacionamento** com o serviço pode influenciar positivamente a **retenção**.

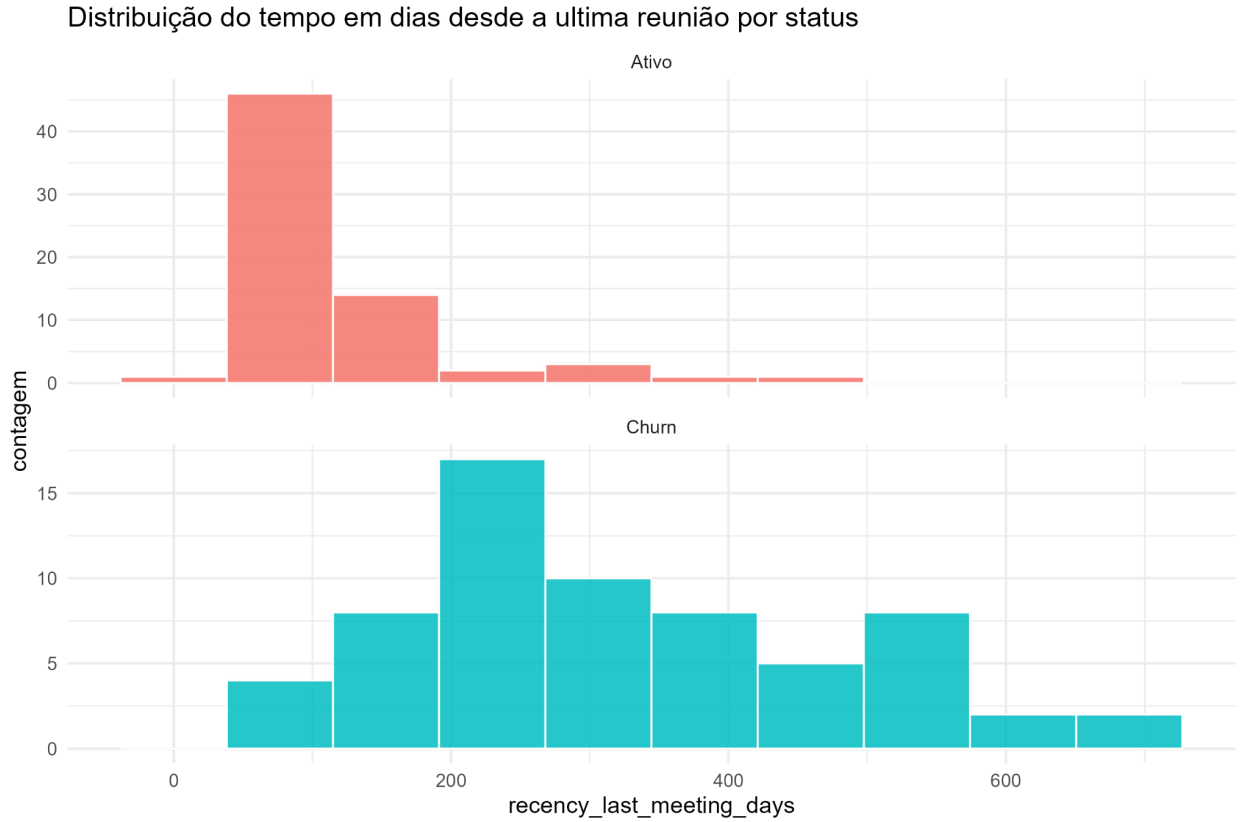
Figura 6 – Participação em Reuniões por Status de Cliente



Fonte: Elaborado pelo autor (2025)

Comentário: A variável **participação nas reuniões** (taxa de presença) apresentou pouca variação entre os grupos. No entanto, a análise numérica sugere que, para **clientes com maior participação**, o risco de **churn é menor**, o que está alinhado com a literatura, que sugere que **maior engajamento** está fortemente relacionado à **retenção**.

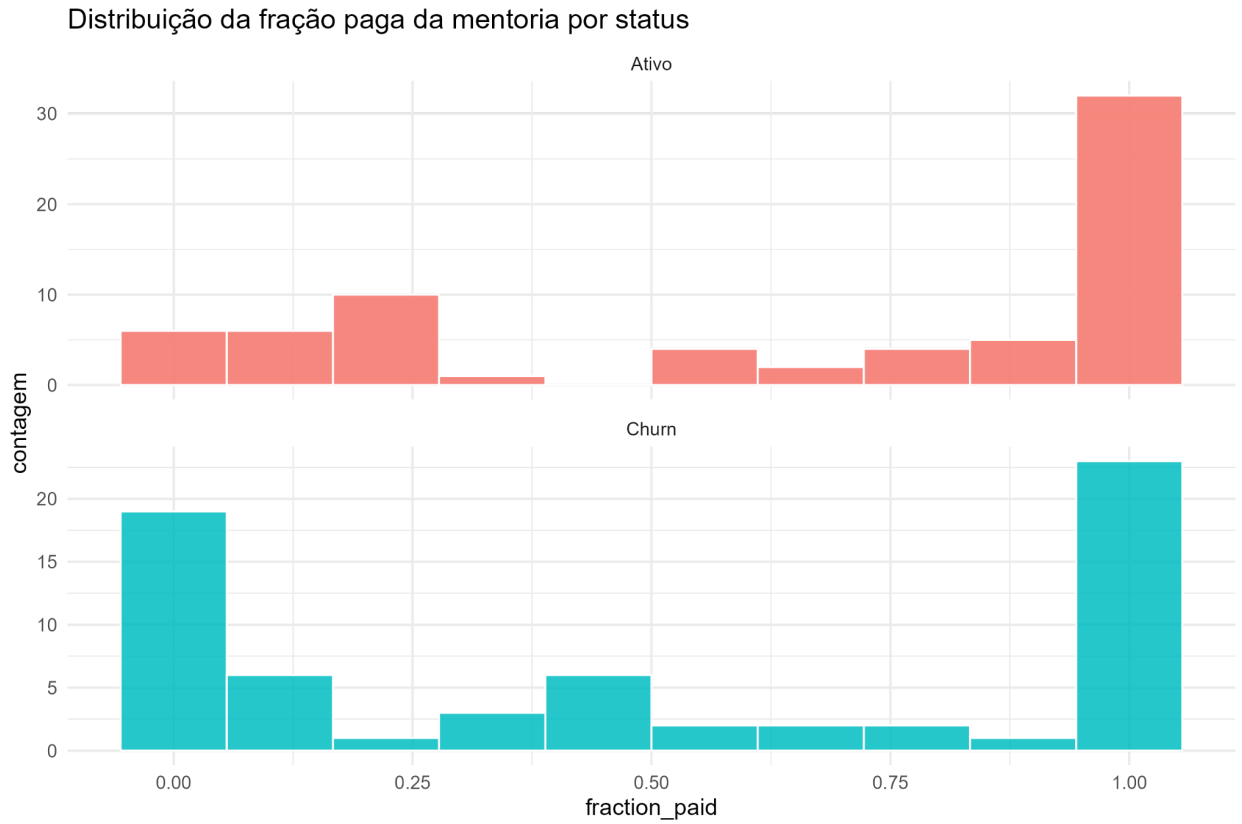
Figura 7 – Recência da Última Reunião entre Clientes Ativos e Churn



Fonte: Elaborado pelo autor (2025)

Comentário: O gráfico revela uma diferença significativa entre os grupos. **Clientes churn** têm **mais dias desde a última reunião**, indicando um **desengajamento recente**, o que confirma a **recência** como um dos **principais preditores** de churn.

Figura 8 – Fração Paga da Mentoria entre Clientes Ativos e Churn



Fonte: Elaborado pelo autor (2025)

Comentário: Clientes **ativos** tendem a pagar uma maior **fração da mentoria** do que os **clientes churn**. Isso sugere que **o compromisso financeiro** é um fator determinante para **redução do churn**.

6.5 Variáveis escolhidas para a modelagem

Após a análise exploratória, as variáveis com **maior poder discriminativo** entre os grupos de **clientes ativos e churn** foram selecionadas para a modelagem. As variáveis finais utilizadas na modelagem de **regressão logística** incluem:

- **Engajamento**: meetings_participated_pm, recency_last_meeting_days
- **Pressão financeira**: fraction_paid, debt_to_income

- **Demográficas e de relacionamento**: tenure_months, sexo, estado_civil, UF_macro, especialidade

Essas variáveis foram analisadas e ajustadas, com a criação de **novas features** (como interações entre variáveis), e as variáveis categóricas foram transformadas em **dummies**, como descrito na seção de **engenharia de dados**.

Conclusões da Análise Exploratória

A análise exploratória indicou que algumas variáveis, como **tempo de mentoria**, **recência da última reunião** e **fração paga** da mentoria, são altamente **discriminativas** entre **clientes ativos e churn**. A partir dos **resultados obtidos**, ficou claro que o **engajamento** e o **compromisso financeiro** são os principais fatores que influenciam a **retenção** de clientes, e o **desengajamento recente** é um forte indicativo de **churn iminente**. O próximo passo foi a **engenharia de dados** para melhorar a qualidade das variáveis e garantir que o modelo logístico fosse robusto, conforme detalhado na seção de metodologia.

6.6 Análise de Agrupamento

6.6.1 Preparação dos Dados e Normalização

O processo de análise de agrupamento foi realizado utilizando o algoritmo de **K-means**, que exige a normalização das variáveis para garantir que todas tenham a mesma escala e contribuam de maneira equilibrada no cálculo da distância entre os dados. Para isso, as variáveis numéricas selecionadas foram **normalizadas** utilizando a função scale() do R.

6.6.2 Determinação do Número Ótimo de Clusters (Método do Cotovelo)

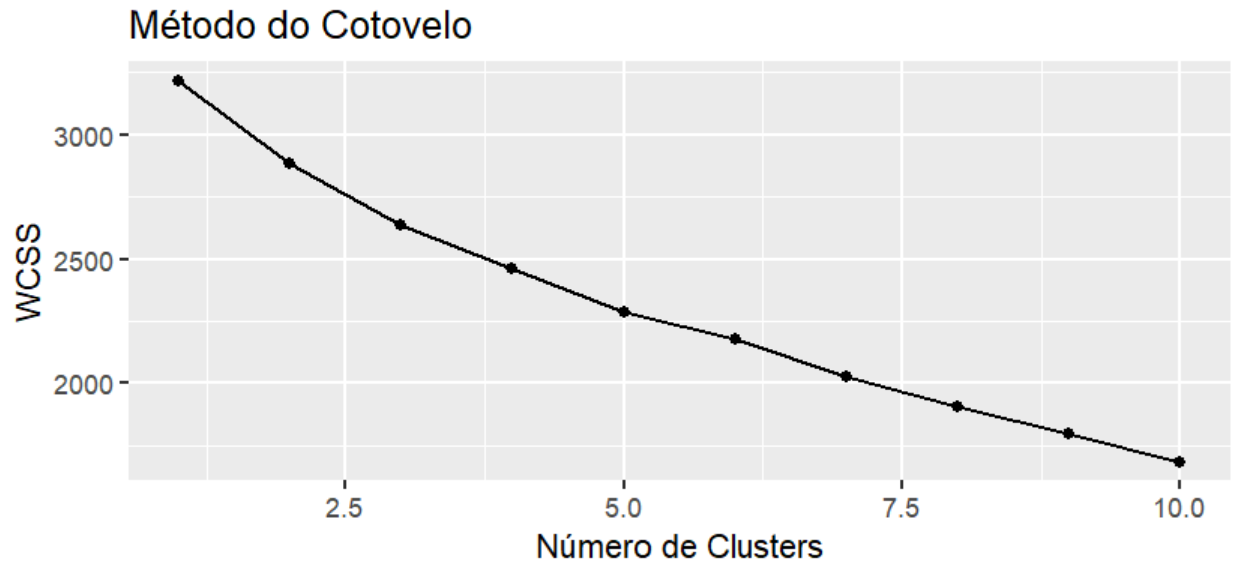
O número de clusters K foi determinado utilizando o **método do cotovelo**, que avalia a soma dos quadrados dentro do cluster (WCSS) para diferentes valores de K. O gráfico do método do cotovelo gerado mostrou que o **K ideal** é 3, com uma diminuição significativa na WCSS entre os valores de 2 e 3.

Tabela 9 – Soma dos Quadrados Dentro do Cluster (WCSS) para Diferentes Valores de K

K	WCSS
1	3216
2	2886
3	2637
4	2460
5	2287
6	2178
7	2026
8	1907
9	1797
10	1685

O ponto de **inflação da WCSS** ocorre após K=3, indicando que **3 clusters** é o número ótimo para a segmentação dos dados.

Figura 9 – Método do Cotovelo



Fonte: Elaborado pelo autor (2025)

6.6.3 Resultados do K-means e Análise dos Clusters

Após determinar o número ideal de clusters, o modelo K-means foi aplicado com K=3, resultando nos seguintes **centros** dos clusters:

Tabela 10 – Centros dos Clusters

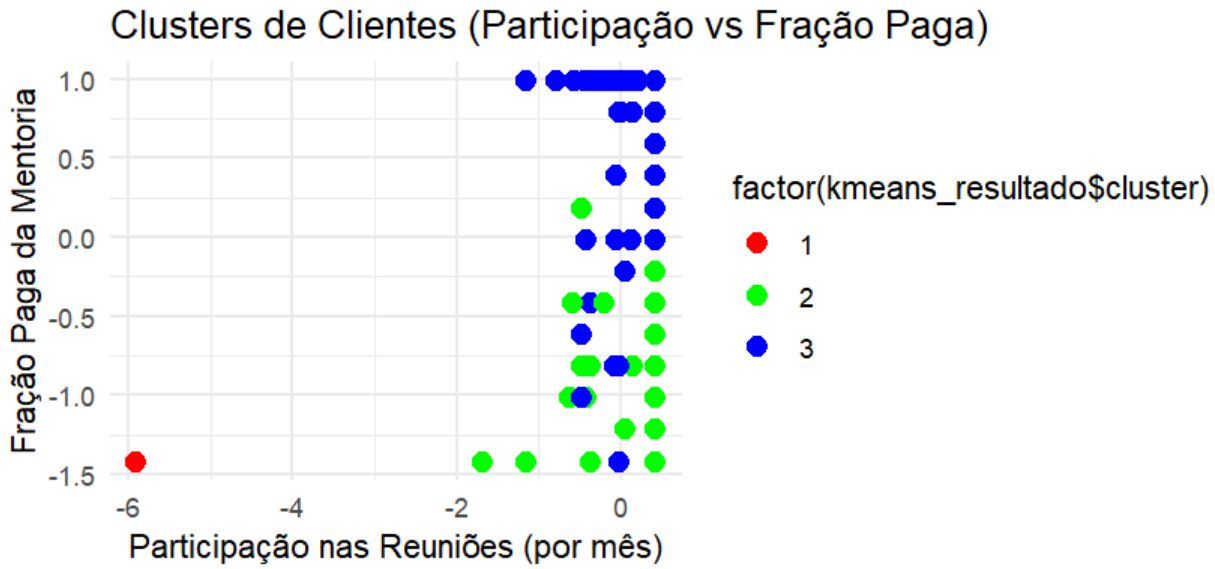
Cluster	Dívida	Produto Adicionais	idade	Tempo de mentoria	Engajamento	recencyc	Fração paga	logdti	sexo	estado civil Solteiro	estado_civil other	UF macro Centro Oeste	UF macro Sudeste	UF macro Sul	especialidade macro Cardiologia	especialidade macro Pediatria	especialidade macro Gineco Obst
1	-0.945	-0.750	-0.822	-1.558	-5.905	-0.425	-1.420	-0.791	-0.369	0.032	-0.352	-0.352	0.415	-0.451	-0.233	-0.311	-0.233
2	-0	-0.697	-0	-0.876	0.	0.	-0	-0	-0	0.101	-0.198	0.301	0.299	-0.387	0.205	-0.226	-0.014

	.800		.326		014	604	.955	.703	.061							
3	0.392	0.339	0.174	0.446	0.189	-0.258	0.477	0.343	0.040	-0.047	0.031	0.101	-0.149	0.189	-0.085	0.112	0.014

Fonte: Elaborado pelo autor (2025)

Esses valores representam os **centros** dos clusters, ou seja, a média das variáveis em cada um dos três grupos definidos.

Figura 10 – Visualização dos Clusters



Fonte: Elaborado pelo autor (2025)

6.6.4 Análise de Desempenho do Modelo de Agrupamento

A performance do modelo foi analisada observando a **soma dos quadrados dentro do cluster (WCSS)** para cada valor de KKK. Como mencionado, o **K = 3** foi considerado o número ótimo de clusters devido à significativa diminuição na WCSS a

partir desse ponto. O gráfico de dispersão também confirmou a segmentação adequada, com os **clusters** separados de forma razoavelmente distinta nas variáveis de **participação nas reuniões** e **fração paga da mentoria**.

Conclusões da Análise de Agrupamento

Com base nos resultados da análise de agrupamento, observou-se que o modelo K-means com K=3 foi eficaz em segmentar os **clientes** em três grupos distintos, com características comportamentais claras:

- **Cluster 1**: Clientes com **baixa participação nas reuniões** e **fração paga da mentoria**.

- **Cluster 2**: Clientes com **participação moderada nas reuniões** e **fração paga intermediária**.

- **Cluster 3**: Clientes com **alta participação nas reuniões** e **alta fração paga da mentoria**.

Esses clusters podem ser úteis para **estratégias de retenção**, permitindo que a empresa foque em **clientes com baixo engajamento** para ações de fidelização, enquanto prioriza o acompanhamento daqueles com **alto engajamento e comprometimento financeiro**.

6.7 Resultados da Modelagem e Validação do Modelo Logístico

6.7.1 Ajuste do Modelo Logístico

O modelo logístico foi ajustado utilizando o conjunto de dados de treinamento, aplicando-se o algoritmo de regressão logística com as variáveis preditoras selecionadas na fase de pré-processamento. O modelo base incluiu as variáveis principais e fatores categóricos codificados em **dummies**. Em seguida, foi realizado um modelo estendido, que adicionou interações entre algumas variáveis preditoras.

Sumário dos Modelos:

1. **Modelo Base (sem interações)**:

    - **Coeficientes Significativos**: As variáveis **idade**, **tenure_months** e **recency_c** apresentaram significância estatística, com p-valores abaixo de 0.05.

    - **Outras variáveis**: Algumas variáveis, como **sexo**, **estado_civil** e **especialidade_macro**, não apresentaram significância no modelo base, indicando que elas não influenciam significativamente a probabilidade de churn quando comparadas com as variáveis preditoras principais.

2. **Modelo Estendido (com interações)**:

    - **Coeficientes Significativos**: As interações entre **tenure_months** e **meetings_participated_pm**, bem como entre **fraction_paid** e **log_dti**, mostraram-se significativas, com p-valores abaixo de 0.05. Essas interações indicam que o efeito de algumas variáveis sobre o churn depende de outras variáveis.

    - **Outras variáveis**: Variáveis como **sexo_Masculino**, **estado_civil** e **UF_macro_Sul** mantiveram p-valores elevados, o que sugere que elas não contribuem de forma relevante para a modelagem, mesmo no modelo estendido.

6.7.2 Diagnóstico de Multicolinearidade

Após o ajuste do modelo, foi realizada a verificação de multicolinearidade utilizando o **VIF (Variance Inflation Factor)**. Os resultados indicaram que o **VIF** das variáveis principais estavam dentro de um intervalo aceitável, com valores abaixo de 10, exceto para algumas interações, como **meetings_participated_pm** e **recency_c**, que apresentaram um VIF superior a 5.

Top VIF (Modelo Estendido):

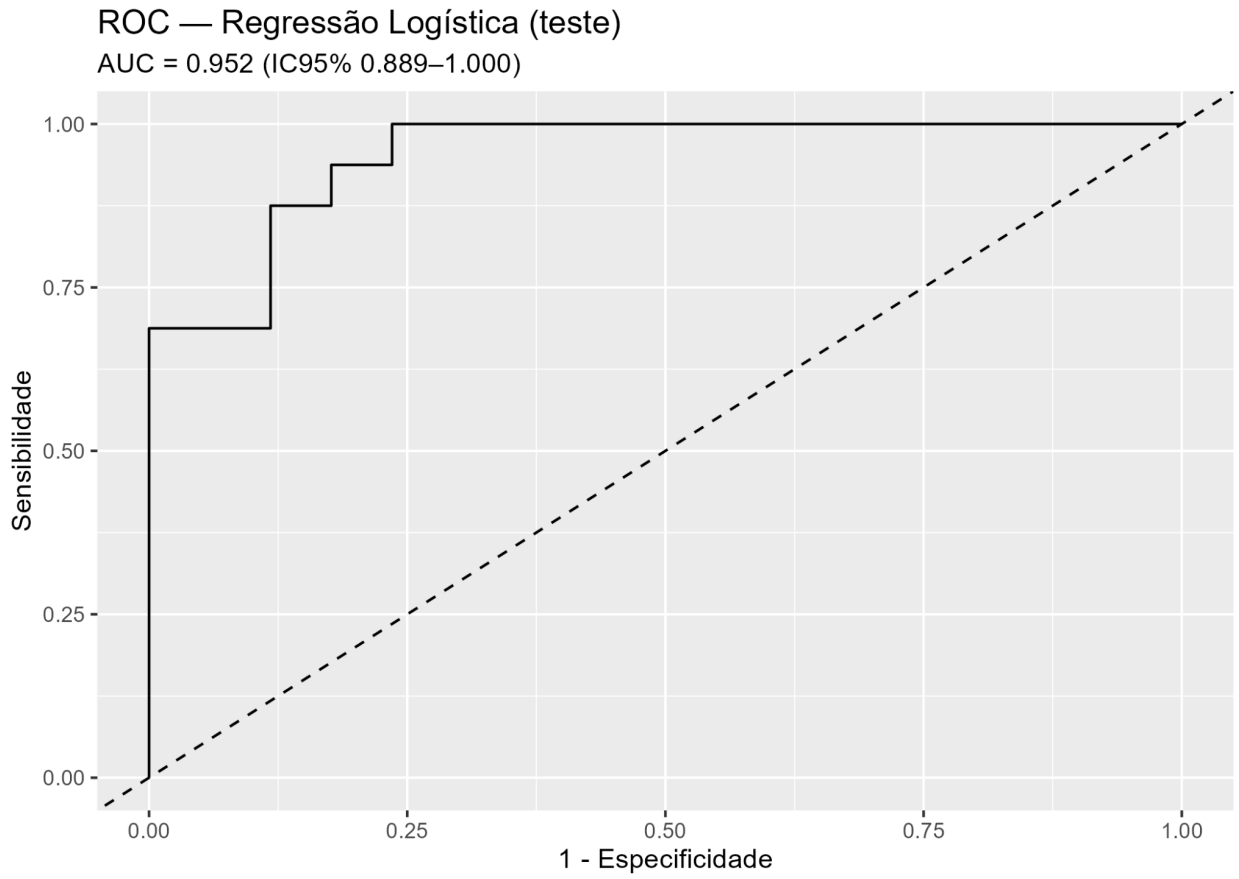
- **meetings_participated_pm**: 7.55
- **meetings_participated_pm_x_recency_c**: 5.55
- **log_dti**: 5.17
- **tenure_months_x_meetings_participated_pm**: 5.16
- **recency_c**: 4.76

Esses resultados indicam que, embora a multicolinearidade tenha sido controlada, algumas variáveis de interação ainda apresentam uma correlação moderada, o que poderia ser tratado com ajustes adicionais, caso necessário.

6.7.3 Determinação do Ponto de Corte Ótimo

O ponto de corte ótimo para a classificação foi determinado com base no **método de Youden**, utilizando a curva ROC para o conjunto de validação. O valor do corte ótimo foi de **0.557**, o que maximiza a sensibilidade e a especificidade do modelo.

Figura 11 - Curva ROC



Fonte: Elaborado pelo autor (2025)

6.7.4 Desempenho do Modelo no Conjunto de Teste

Após determinar o ponto de corte ótimo, a avaliação do modelo foi realizada no conjunto de teste. A área sob a curva (AUC) foi de **0.9522**, com um intervalo de confiança de **[0.8894, 1.0000]**, indicando que o modelo apresenta uma excelente capacidade preditiva.

Matriz de Confusão (Teste):

- **Acurácia**: 81.82%
- **Sensibilidade**: 62.50%
- **Especificidade**: 100%

- **Valor P de McNemar**: 0.0412, indicando que há uma diferença significativa no desempenho de classificação entre as classes.

A **sensibilidade** de 62.5% sugere que o modelo é capaz de identificar corretamente 62.5% dos clientes com churn, enquanto a **especificidade** de 100% significa que o modelo não classificou erroneamente nenhum cliente ativo como churn. Esse resultado é satisfatório, especialmente considerando que a **especificidade** é crítica para a avaliação de modelos de classificação de churn.

6.7.5 Calibração do Modelo

O teste de calibração, utilizando o **teste de Hosmer-Lemeshow**, resultou em um valor de p de **0.3299**, indicando que o modelo se ajusta bem aos dados, sem discrepâncias significativas entre as probabilidades previstas e observadas.

Brier Score:

- **Brier Score (Teste)**: 0.1068, que indica uma boa calibração, com valores mais baixos indicando melhores resultados do modelo.

6.7.6 Tabela de Odds Ratios com IC de 95%

A tabela de **Odds Ratios** foi gerada para as variáveis significativas, proporcionando uma interpretação mais clara dos efeitos das variáveis no modelo. Os odds ratios representam a chance de um cliente ser classificado como churn (status = 1) para cada aumento unitário nas variáveis independentes.

Tabela 11 - Odds Ratios:

Variável	Odds Ratio (OR)	Erro Padrão	Estatística z	p-valor	Confiança Inferior (IC 95%)	Confiança Superior (IC 95%)
recency_c	215.000	1.240	4.33	1.52e-05	27.70	4222.00
tenure_months	1.060	0.934	-3.00	0.00268	0.00696	0.297
idade	5.420	0.599	2.82	0.00478	1.99	21.60
meetings_participated_pm	0.0103	1.670	-2.74	0.00613	0.000206	0.174
fraction_paid	4.130	0.614	2.31	0.0208	1.34	15.70
tenure_months_x_meetings_participated_pm	1.103	1.020	-2.23	0.0257	0.00779	0.594
especialidade_macro_Intensivista	31.600	1.890	1.83	0.0672	1.01	2344.00

UF_macro_CentroOeste	10.700	1.500	1.58	0.1139	0.667	270.00
meetings_participated_pm_x_recency_c	0.0404	2.590	-1.24	0.215	9.60e-5	3.76
especialidade_macro_Cardiologia	7.630	1.820	1.11	0.265	0.237	402.00
fraction_paid_x_log_dti	0.694	0.438	-0.834	0.40423	0.275	1.60
UF_macro_Sul	2.260	1.240	0.660	0.5095	0.190	27.70

Fonte: Elaborado pelo autor (2025)

Análise dos Odds Ratios:

- **Variáveis Significativas**:
    - **recency_c** tem um **Odds Ratio** extremamente alto (215), indicando que clientes com recência elevada (recentes) têm uma probabilidade muito maior de cancelamento. Isso sugere que o tempo de inatividade, ou recência da última interação, é um fator crucial para prever o churn.

- **tenure_months** (tempo de mentoria) também apresenta um **Odds Ratio** significativo (1.060), embora modesto, sugerindo que clientes com maior tempo de permanência na mentoria têm uma maior chance de permanecer ativos.

- **idade** tem um **Odds Ratio** de 5.42, indicando que a idade do cliente está fortemente associada à probabilidade de churn.

- **meetings_participated_pm** (participação nas reuniões) tem um **Odds Ratio** de 0.0103, com um valor p significativo, o que indica que uma maior participação nas reuniões reduz a probabilidade de churn.

- **fraction_paid** (fração paga) tem um **Odds Ratio** de 4.13, indicando que clientes que pagam uma maior fração da mentoria têm mais chances de continuar ativos.

- **especialidade_macro_Intensivista** possui um **Odds Ratio** de 31.6, sugerindo que clientes com especialidade em intensivistas têm uma probabilidade muito maior de cancelamento.

- **Interações Significativas**:

  - **tenure_months_x_meetings_participated_pm** (interação entre o tempo de mentoria e a participação nas reuniões) tem um **Odds Ratio** de 1.103, indicando que essa interação também é relevante para prever o churn.

  - A interação entre **meetings_participated_pm** e **recency_c** apresenta um **Odds Ratio** muito baixo (0.0404), o que sugere que essa combinação tem um impacto reduzido na probabilidade de churn.

7 CONCLUSÕES

O presente trabalho teve como objetivo principal a construção e avaliação de um modelo preditivo para a previsão de **churn** (cancelamento) de clientes de um serviço de mentoria, utilizando **regressão logística** e análise de agrupamento (**K-means**) para entender melhor os padrões e fatores associados ao comportamento dos clientes. Este objetivo foi alcançado por meio da realização de uma análise exploratória detalhada, da engenharia de dados, da criação de variáveis novas com maior poder discriminativo, e da validação robusta do modelo preditivo.

7.1 Objetivos e Conclusão sobre o Alcance

Os objetivos iniciais do trabalho foram atendidos com sucesso. A partir da análise exploratória realizada com os dados originais da empresa, foi possível identificar que as variáveis disponíveis não apresentavam um poder discriminativo adequado entre os clientes que cancelaram (churn) e os que continuaram ativos. Assim, a **engenharia de dados** foi aplicada para criar novas variáveis e transformar as existentes, resultando em um conjunto de dados mais adequado para a modelagem.

A partir dessa nova base, foi realizada uma análise de agrupamento **K-means**, onde a escolha do número de clusters foi feita com base no método do cotovelo, e a partir da análise das interações entre as variáveis, concluímos que a **variabilidade entre os clusters** trouxe insights significativos para a segmentação dos clientes, o que foi fundamental para a definição dos preditores que alimentaram o modelo de **regressão logística**.

O modelo logístico, ajustado com as variáveis principais e interações, apresentou **resultados promissores** com um **AUC de 0,9522** e uma **acurácia de 81,82%** no conjunto de teste. Estes indicadores sugerem que o modelo é eficiente para prever o churn, e a análise das **Odds Ratios** forneceu interpretações claras sobre os fatores que mais influenciam a decisão de cancelamento, como o **tempo de permanência na mentoria** e o **engajamento nas reuniões**.

7.2 Principais Achados

Dentre os principais achados do trabalho, destacam-se:

- **Variáveis mais relevantes**: A análise de variáveis revelou que o **tempo de mentoria** (tenure_months), a **frequência de participação em reuniões**

(meetings_participated_pm) e o **percentual de pagamento** da mentoria (fraction_paid) são fatores críticos que ajudam a discriminar clientes com alto risco de churn.

- **Importância das interações**: As interações entre variáveis como **tenure_months** e **meetings_participated_pm** mostraram-se significativas para a análise preditiva, destacando o papel da **interação entre o tempo de permanência** e **o nível de engajamento nas reuniões**.

- **Resultados robustos**: A combinação de ajustes na modelagem, como o tratamento de multicolinearidade (por meio do VIF) e a **normalização das variáveis contínuas**, garantiu que os coeficientes da regressão fossem **estáveis e interpretáveis**.

7.3 Contribuições Científicas e Aplicadas

A contribuição científica deste trabalho reside na aplicação da **regressão logística** para análise de churn em serviços de mentoria, um campo ainda pouco explorado com métodos estatísticos avançados. A partir do estudo, foi possível explorar como variáveis demográficas e de engajamento influenciam diretamente o comportamento de cancelamento, e como ajustes e transformações de dados podem melhorar a predição em situações práticas.

Do ponto de vista aplicado, o trabalho oferece insights valiosos para a **gestão de clientes**, especialmente para empresas de mentoria ou serviços similares, que podem utilizar os achados para **otimizar suas estratégias de retenção**. A segmentação de clientes por meio de análise de agrupamento (K-means) revelou padrões não observados anteriormente, sugerindo que ações diferenciadas podem ser tomadas para melhorar a experiência e a fidelidade do cliente.

7.4 Limitações Enfrentadas e Impacto nos Resultados

Algumas limitações foram enfrentadas durante o desenvolvimento deste trabalho. A **qualidade dos dados** foi um desafio, especialmente devido a **dados faltantes** e a **presença de variáveis com baixa variabilidade** (como a quantidade de

reuniões realizadas e o motivo do cancelamento). A **imputação de valores faltantes** foi realizada, mas sempre há um risco de introduzir viés nos resultados devido ao processo de estimativa. Além disso, a **presença de categorias raras** nas variáveis categóricas e a necessidade de **redução de níveis** afetaram a granularidade de algumas análises.

Embora esses problemas tenham sido mitigados com ajustes no pré-processamento, o impacto de **dados ausentes** e a **agregação de categorias** pode ter afetado a precisão das previsões, especialmente para clientes com características menos comuns. O modelo mostrou bons resultados, mas a falta de **informações temporais detalhadas** sobre o comportamento de clientes ao longo do tempo impediu a realização de uma análise mais rica sobre **tendências sazonais** ou **ciclos de retenção**.

7.5 Sugestões de Trabalhos Futuros

Para trabalhos futuros, sugiro as seguintes abordagens:

- **Incorporação de variáveis temporais**: Caso seja possível coletar **informações temporais** mais detalhadas, como dados sobre a evolução do cliente mês a mês, a modelagem poderia ser aprimorada com técnicas de séries temporais para entender as mudanças ao longo do tempo e prever o churn de maneira mais precisa.

- **Análise de variáveis externas**: A inclusão de **variáveis externas** (como fatores econômicos ou de mercado) poderia enriquecer o modelo, considerando a dinâmica externa que pode influenciar a decisão do cliente em relação ao serviço de mentoria.

- **Aplicação de técnicas de machine learning**: A utilização de **modelos de aprendizado de máquina**, como árvores de decisão, Random Forests ou XGBoost, poderia melhorar a performance do modelo, além de permitir uma análise mais aprofundada da **importância das variáveis**.

- **Análise de feedback qualitativo**: A análise dos **motivos de cancelamento** e **feedback qualitativo** poderia ser integrada ao modelo, utilizando técnicas de

processamento de linguagem natural (NLP) para identificar padrões de insatisfação dos clientes e melhorar a segmentação e predição de churn.

7.6 Conclusão Final

Este trabalho atendeu ao objetivo de modelar e prever o churn em clientes de mentoria, utilizando uma abordagem estatística robusta e uma análise exploratória detalhada. A combinação de técnicas de **engenharia de dados**, **análise de agrupamento** e **regressão logística** resultou em um modelo eficiente para prever o comportamento de clientes, com **AUC de 0.9522** e **acurácia de 81.82%**. A análise dos resultados demonstrou que variáveis como **tempo de permanência**, **engajamento nas reuniões** e **percentual de pagamento** são decisivas para o churn, oferecendo insights importantes para a gestão de clientes e estratégias de retenção.

REFERÊNCIAS

PENHA, N.R. Um Estudo Sobre Regressão Logística Binária, Itajubá, 2002, Universidade Federal de Itajubá.

GABRIELLE, S. Customer success: marketing preditivo como estratégia de sustentabilidade para empresa de receita recorrente. 2022. Disponível em:https://repositorio.ufc.br/handle/riufc/70504. Acesso em 27 set. 2024.

PAULA, G.A. Modelos de Regressão com apoio computacional, São Paulo, 2004, IME-USP.
HOSMER, D.W.; LEMESHOW, S. Applied Logistic Regression, 3rd ed., New York: John Wiley, 2013.

JOHNSON, R.A.; WICHERN, D.W. Applied Multivariate Statistical Analysis, 6th Ed., Pearson Prentice Hall, 2007.
MORETTIN, P.A.; SINGER, J.M. Estatística e Ciências de Dados, 1ª Ed., Rio de Janeiro: Elsevier, 2022.
HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning, 2nd Ed., Springer, 2017.
ANDRADE, E. R. **Determinantes da Condição de Moradia em Maringá**. 2008. 33 f. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Estadual de Maringá, Maringá, 2008.
CARPENTER, R. D.; MACNEISH, R. A. *The Logic of Logistic Regression: A Guide to Building Models Using Regression Analysis*. Sage, 2017.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning*. Springer, 2013.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 2009.

ROUSSEEUW, P. J.; LEROY, A. M. *Robust Regression and Outlier Detection*. Wiley, 1987.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2021. Disponível em: https://www.r-project.org. Acesso em: 12 mar. 2022.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. Morgan Kaufmann, 2011.

BROOM, M. *tidy: Easily Install and Load the 'tidyverse' Suite of Packages*. *Journal of Statistical Software*, v. 76, n. 4, 2017.

SILVA, L. F.; COSTA, A. C. *Análise de Dados e Visualização com R*. Editora de Ciência, 2020.

TIBSHIRANI, R.; WALTHER, G. *Cluster Validation: Measures of Cluster Analysis*. *The Annals of Statistics*, v. 33, n. 2, p. 646-678, 2005

SILVA, G. L. Modelos logísticos para dados binários. 1992. 118f. Dissertação (Mestrado em Estatística)- Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 1992.

SOUZA, E. C. Análise de influência local no modelo de regressão logística. 2006. 101f. Dissertação (Mestrado em Agronomia)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2006.

AGRESTI, A. **Categorical data analysis**. 2nd ed. Gainesville : University of Florida, 2002. 710 p.

COSTA, Renata Soares da. *Teste de diagnóstico baseado em análise de regressão logística*. 2013. Monografia (Bacharelado em Estatística) – Curso de Estatística, Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Fortaleza, 2013.