Ana Jacinta Cavalcanti Barreto

Modelando *churn* através da regressão logística bayesiana

Niterói - RJ, Brasil

17 de julho de 2023

Universidade Federal Fluminense

Ana Jacinta Cavalcanti Barreto

Modelando *churn* através da regressão logística bayesiana

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Profª. Drª. Patrícia Lusié Velozo da Costa

Niterói - RJ, Brasil

17 de julho de 2023

Universidade Federal Fluminense

Ana Jacinta Cavalcanti Barreto

Modelando churn através da regressão logística bayesiana

Monografia de Projeto Final de Graduação sob o título *"Modelando churn através da regressão logística bayesiana"*, defendida por Ana Jacinta Cavalcanti Barreto e aprovada em 17 de julho de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profª. Drª. Patrícia Lusié Velozo da Costa
Departamento de Estatística – UFF

Documento assinado digitalmente
gov.br
MARIANA ALBI DE OLIVEIRA SOUZA
Data: 19/07/2023 08:26:14-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Mariana Albi de Oliveira Souza
Departamento de Estatística – UFF

Documento assinado digitalmente
gov.br
GUILHERME AUGUSTO VELOSO
Data: 18/07/2023 11:36:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Guilherme Augusto Veloso
Departamento de Estatística – UFF

Niterói, 17 de julho de 2023

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

B273m Barreto, Ana Jacinta Cavalcanti
Modelando churn através da regressão logística bayesiana / Ana Jacinta Cavalcanti Barreto. - 2023.
53 f.

Orientador: Patrícia Lusié Velozo da Costa.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Churn. 2. Regressão Logística. 3. Inferência Bayesiana. 4. Produção intelectual. I. Costa, Patrícia Lusié Velozo da, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O fenômeno do *churn*, caracterizado pela perda de clientes, é uma preocupação crescente para as empresas. A retenção de clientes é fundamental, considerando que o custo de adquirir novos clientes é significativamente maior do que o de mantê-los. Portanto, a retenção de clientes torna-se crucial, exigindo uma compreensão aprofundada do *churn*. Modelos preditivos são ferramentas populares para identificar clientes propensos ao *churn*, mas a escolha adequada do modelo e das variáveis é desafiadora. Este trabalho visa identificar e interpretar as variáveis relevantes para o *churn* em uma empresa de telecomunicações fictícia. Foram construídos modelos de regressão logística bayesiana, utilizando uma amostra de dados de relacionamento com clientes. Os modelos foram comparados com base na relevância estatística das variáveis, e o modelo com melhor desempenho foi selecionado. O modelo escolhido apresentou um ajuste adequado aos dados, com alta acurácia, sensibilidade e área sob a curva (AUC), demonstrando sua eficácia na previsão do *churn*. Os resultados forneceram informações valiosas sobre as variáveis que influenciam o *churn*, permitindo que a empresa tome medidas de retenção de clientes. O estudo alcançou seus objetivos e forneceu um modelo preditivo eficaz para o *churn* na empresa analisada.

Palavras-chave: Regressão Logística. Modelos Lineares Generalizados. Inferência Bayesiana. Predição de *Churn*.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AIC Critério de Informação de Akaike

AUC Área Abaixo da Curva

DIC Critério de Informação da Deviance

HMC Monte Carlo Hamiltoniano

MCMC Monte Carlo via cadeias de Markov

OR *Odds Ratio*

ROC *Receiver Operating Characteristic*

TRV Teste da Razão da Verossimilhança

1 Introdução

Neste Capítulo, na Seção 1.1, serão apresentadas as razões e o contexto que levaram à realização deste estudo, além de discutir a relevância do tema em questão. Em seguida, na Seção 1.2, será realizada uma revisão da literatura sobre o assunto, com o objetivo de contextualizá-lo dentro do campo de estudo. Na Seção 1.3, serão definidos os objetivos a serem alcançados neste trabalho. Por fim, na Seção 1.4, será apresentada uma visão geral da estrutura do trabalho, destacando como ele está organizado em termos de capítulos e ressaltando os principais tópicos abordados em cada um deles.

1.1 Motivação

A perda de clientes, fenômeno conhecido como *churn*, tem sido cada vez mais uma preocupação para as empresas, já que para manter o nível de receita, os clientes perdidos devem ser substituídos por novos. Porém, em um cenário extremamente competitivo como o mercado de produtos e serviços, onde os clientes normalmente são bombardeados por um grande número de empresas com ofertas de serviço ou produtos muito parecidas, a tarefa de conseguir novos clientes torna-se cada vez mais difícil.

Além disso, o custo de aquisição de novos clientes pode ser de cinco a dez vezes superior ao seu custo de retenção (KURTZ; CLOW, 1997), a depender do setor. Na indústria de telecomunicações, por exemplo, o custo de aquisição pode ser até oito vezes maior que o custo de retenção do cliente, segundo Au, Chan e Yao (2003). Dessa forma, uma gestão efetiva da retenção de clientes torna-se indispensável, sendo cada vez mais relevante para as empresas buscar entendimento acerca do fenômeno do *churn*, já que este permite que a empresa trace ações e estratégias a fim de minimizá-lo. No caso de agências bancárias, um estudo empírico mostrou que reduzir em 5% o índice de *churn* é capaz de elevar os lucros em 25% a 95% (REICHHELD; JR, 1990).

Pelas razões citadas anteriormente, os modelos preditivos para a probabilidade de *churn* se tornaram ferramentas populares e bons aliados para as empresas, já que a partir destes, é possível saber de forma antecipada quais clientes estão propensos ao *churn*. Por outro lado, com as particularidades de cada empresa e também com a quantidade de técnicas disponíveis, escolher um modelo para estimar a probabilidade de *churn* pode se tornar uma tarefa difícil, já que não existe um único modelo que seja o mais indicado para todas as situações (KUMAR et al., 2012). Sendo assim, a escolha do melhor modelo é fundamental para a solução do problema, já que a escolha inadequada das covariáveis pode gerar resultados ruins.

Outro fator importante a se destacar é que se apenas a previsão da probabilidade de *churn* for realizada, a empresa não será capaz de entender quais os motivos estão por trás da rotatividade de clientes. Para que um programa de ações de retenção seja mais eficaz, é importante também descobrir as razões que levam o cliente a abandonar o relacionamento com a empresa, pois assim os profissionais de marketing conseguirão agir com mais assertividade, já que além da probabilidade de abandono do cliente, terão também um melhor entendimento acerca dos motivos por trás do fenômeno do *churn*.

1.2 Revisão Bibliográfica

A literatura comprova que a modelagem preditiva para detectar o *churn* é um tema bastante explorado e que possibilita inúmeros resultados. A partir de algumas das técnicas de classificação mais utilizadas em trabalhos de previsão de *churn*: Árvores de Decisão, Regressão Logística, Redes Neurais, Floresta Aleatória e *Support Vector Machine*, Veloso (2013) estudou o fenômeno da perda de clientes para o setor de varejo, com o objetivo de definir e implementar um modelo de predição de *churn*. A avaliação e comparação da performance dos cinco modelos foi feita através das medidas de Acurácia, Sensibilidade, Especificidade, Precisão, *f-measure* e AUC. Os modelos construídos através das técnicas de Árvores de Decisão e Floresta Aleatória obtiveram os melhores resultados quanto à capacidade de predição, com valores muito semelhantes em todas as medidas de avaliação.

Franceschi (2019) também comparou diversos modelos de predição, através das técnicas de Regressão Logística, Regressão Logística com o uso do algoritmo de *stepwise* para a seleção das covariáveis relevantes, Floresta Aleatória e Redes Neurais, para entender qual seria a melhor técnica preditiva para identificar os clientes propensos ao *churn* do Banco do Brasil, uma importante instituição financeira brasileira. Os modelos de

Regressão Logística sem e com o algoritmo de *stepwise* obtiveram melhor performance comparado aos demais.

Com o objetivo de obter um modelo preditivo que possibilitasse não só a classificação, mas também a interpretação dos motivos que levam os clientes ao *churn* em uma *startup* brasileira de plataforma digital para vendedores, Junior (2020) ajustou também um modelo de Regressão Logística com o algoritmo de *stepwise* e fez uso de validação cruzada *K-fold*. O poder preditivo do modelo foi avaliado através da curva *Receiver Operating Characteristic* (ROC). Para seleção de covariáveis usou-se o Critério de Informação de Akaike (AIC) e o Teste da Razão da Verossimilhança (TRV). O modelo proposto teve boa performance no ajuste e na previsão dos dados aplicados.

A fim de fornecer às empresas do setor de telecomunicação uma análise mais completa para o gerenciamento do *churn*, Wu et al. (2021) combinou a previsão do *churn* e a segmentação de clientes, utilizando técnicas de aprendizado de máquinas e o algoritmo de agrupamento *K-means*. Ajustou seis classificadores diferentes para três bases de dados e o estudo indicou que para a primeira base de dados, o AdaBoost obteve melhor desempenho, com precisão de 77,20%. Já para a segunda e a terceira, Floresta Aleatória obteve o melhor desempenho, com precisão de 93,6% e 63,09%, respectivamente. Depois de realizar a previsão, utilizou a Regressão Logística Bayesiana para encontrar os fatores mais importantes que contribuem para o *churn* e assim obter uma segmentação de clientes mais precisa.

Paiva (2022) também propôs um modelo de aplicação do algoritmo de clusterização *K-means* para um estudo de caso a partir de dados de consumo e perfil dos clientes de uma empresa de telecomunicações. Desse modo, foi possível analisar a qualidade da informação obtida ao aplicar o modelo para 2 até 5 agrupamentos, de maneira não supervisionada. De forma geral, o modelo se mostrou capaz de resolver o problema de maneira satisfatória.

1.3 Objetivos

Este trabalho tem como objetivos: i) identificar e interpretar as possíveis variáveis que influenciam no fenômeno do *churn*; ii) apresentar um modelo preditivo para a probabilidade de um cliente abandonar o relacionamento com a empresa, utilizando uma amostra de treino com dados de relacionamento dos clientes de uma empresa de telecomunicações; e iii) avaliar o desempenho do modelo em uma amostra de teste. Para alcançar os objetivos citados, inicialmente é realizada uma análise descritiva dos dados

e, em seguida, ajustado um modelo de regressão logística sob a perspectiva bayesiana. Além disso, é feito um estudo simulado para avaliar o ajuste e a capacidade preditiva do modelo proposto.

1.4 Organização

Esse trabalho está dividido e organizado em quatro capítulos. Este primeiro, introdutório, apresenta a motivação do estudo, uma breve revisão bibliográfica, os objetivos do trabalho e como este está organizado. O Capítulo 2 contém o material e em seguida, os métodos utilizados no estudo. Posteriormente, no Capítulo 3, estão as análises das aplicações dos métodos apresentados na base de dados. Por fim, no Capítulo 4, encontram-se as conclusões sobre o estudo.

2 Materiais e Métodos

Neste Capítulo, serão apresentados os materiais e métodos utilizados para realizar as análises deste trabalho. Na Seção 2.1, serão fornecidas informações detalhadas sobre o banco de dados utilizado, incluindo sua origem e características. Em seguida, na Seção 2.2, será feita uma revisão dos principais conceitos do modelo de regressão logística e serão apresentados os métodos utilizados para estimar os parâmetros do modelo por meio da inferência bayesiana. Essa seção permitirá entender melhor os métodos analíticos usados neste estudo, fornecendo uma base sólida para analisar e interpretar os resultados obtidos.

2.1 Banco de Dados

Os dados usados neste trabalho foram fornecidos pela plataforma de ensino da IBM [1] e contêm informações sobre a rotatividade dos clientes de uma empresa fictícia de telecomunicações que oferece serviços de telefone e internet na Califórnia. O banco de dados consiste em 7.043 linhas, cada uma representando um cliente, e 37 variáveis. A variável principal deste estudo é a "Churn Label", que indica se o cliente cancelou ou não o serviço. As variáveis "País"e "Estado"foram removidas da tabela, pois todos os clientes são residentes da Califórnia, nos Estados Unidos. Além disso, com base na variável "Cidade", foi criada a variável "Região". O detalhamento das variáveis encontra-se na Tabela 1.

Tabela 1: Descrição das Variáveis

Variável	Descrição	Valores
CustomerID	ID único que identifica cada cliente	
Sexo	Sexo do cliente	Masculino, Feminino

[1] https://community.ibm.com/

Idade	Idade do cliente em anos	
Jovem	Indica se o cliente tem menos de 30 anos	Sim, Não
Idoso	Indica se o cliente tem 65 anos ou mais	Sim, Não
Casado	Indica se o cliente é casado	Sim, Não
Dependentes	Indica se o cliente possui dependentes	Sim, Não
Num de Dependentes	Número de dependentes do cliente	
Região	Região de residência do cliente de acordo com a cidade	Norte, Sul, Bay Area e Central
Indicação	Indica se o cliente já indicou a empresa a alguém	Sim, Não
Num de Indicações	Número de indicações	
Fidelidade	Total de meses que o cliente está na empresa	
Oferta	Última oferta de marketing aceita pelo cliente	Nenhuma, Oferta A, Oferta B, Oferta C, Oferta D, Oferta E
Telefone	Indica se o cliente possui o serviço de telefone residencial	Sim, Não
Longa Distância	Tarifa média de longa distância	
Linhas Múltiplas	Indica se o cliente possui múltiplas linhas telefônicas	Sim, Não
Internet	Indica se o cliente possui serviço de Internet	Sim, Não
Tipo de Internet	Tipo de serviço de internet atual	Cabo, DSL, Fibra Ótica, Nenhuma
Downloads	Volume médio de downloads do cliente em gigabytes	
Segurança Online	Indica se o cliente possui um serviço adicional de segurança online	Sim, Não
Backup Online	Indica se o cliente possui um serviço de backup online adicional	Sim, Não
Proteção de Aparelho	Indica se o cliente assina um plano de proteção de dispositivo adicional	Sim, Não

Suporte Premium	Indica se o cliente assina um plano de suporte técnico adicional para tempos de espera reduzidos	Sim, Não
Streaming TV	Indica se o cliente usa seu serviço de Internet para assistir TV	Sim, Não
Streaming Filme	Indica se o cliente usa seu serviço de Internet para assistir filmes	Sim, Não
Streaming Música	Indica se o cliente usa seu serviço de Internet para escutar música	Sim, Não
Dados Ilimitados	Indica se o cliente pagou uma taxa mensal adicional para ter downloads/uploads ilimitados	Sim, Não
Contrato	Tipo de contrato atual	Mensal, Anual, Bianual.
Conta Digital	Indica se o cliente optou pela fatura sem papel	Sim, Não
Método de Pagamento	Método de pagamento da fatura	Débito automático, Cartão de crédito, Cheque
Cobrança Mensal	Valor total da cobrança mensal atual	
Cobrança Total	Valor total cobrado do cliente até o momento	
Reembolsos	Valor total reembolsado ao cliente até o momento	
Internet Extra	Cobranças totais do cliente para internet extras até o momento	
Longa Distância Total	Total de tarifas extras de longa distância até o momento	
CLTV	Customer Lifetime Value (CLTV) - Valor de vida útil do cliente	
Churn Label	Indica se o cliente deixou o relacionamento com a empresa	Sim, Não

2.2 Regressão Logística

Análise de regressão é uma metodologia que consiste em relacionar duas ou mais variáveis visando explicar o comportamento de dados observados e de realizar previsões de alguma(s) variável(is) de interesse. A Regressão Linear é um método muito utilizado e consiste em relacionar a variável de interesse com covariáveis através de combinação linear somada a um termo que representa um erro. Esse termo tem fortes pressupostos como normalidade, homocedasticidade e independência. Com isso, a variável de interesse é contínua e pode assumir qualquer valor real. Como nem todos os dados possuem essa natureza, surgiu a Regressão Linear Generalizada, que utiliza uma função de ligação para relacionar a combinação linear das covariáveis com a média da variável de interesse. Maiores detalhes sobre os modelos lineares generalizados podem ser vistos em Dobson e Barnett (2018).

A Regressão Logística é um modelo linear generalizado que tem como objetivo produzir um modelo que seja adequado para definir a relação entre a ocorrência de um evento de interesse e um conjunto de variáveis explicativas. O modelo assume a existência de uma variável resposta binária que pode assumir dois valores, normalmente, tratados como "sucesso" (valor 1) e "fracasso" (valor 0). No caso deste trabalho, a variável resposta é a ocorrência de *churn* e pode ser definida da seguinte forma:

$$Y_i \;=\; \begin{cases} 1, \text{ se o } i\text{-ésimo cliente deixar de usar o serviço (ocorrer } \textit{churn}) \\ 0, \text{ se o } i\text{-ésimo cliente não deixar de usar o serviço (não ocorrer } \textit{churn}) \end{cases}$$

com $i = 1, \cdots, n$.

Sendo π_i a probabilidade de ocorrer *churn* para o i-ésimo cliente, temos que $P(Y_i = 1) = \pi_i$ é a probabilidade de sucesso e $P(Y_i = 0) = 1 - \pi_i$, a probabilidade de fracasso. Dessa forma, Y_i tem distribuição Bernoulli com média π_i e função de probabilidade dada por:

$$P(Y_i = y_i|\pi_i) \;=\; \pi_i^{y_i}\,(1-\pi_i)^{1-y_i},\;\; i = 1, \cdots, n. \tag{2.1}$$

Uma das suposições de um modelo linear generalizado é que a função de probabilidade da variável resposta pertença à família exponencial.

Definição 2.1 *Seja Y uma variável aleatória com função de probabilidade ou densidade que depende somente de um parâmetro θ. A distribuição de Y pertencerá à família exponencial de distribuições se puder ser escrita da seguinte forma:*

$$p(y|\theta) = \exp\left\{a(y)b(\theta) + c(\theta)\right\}\ d(y), \tag{2.2}$$

em que a, b, c e d são funções reais e conhecidas, sendo d uma função que retorna valores positivos.

Note que a Equação dada em (2.1), pode ser reescrita da seguinte forma

$$\begin{aligned} P(Y_i = y_i|\pi_i) &= \exp\left\{\ln\Big(P(Y_i = y_i|\pi_i)\Big)\right\}\ I(y_i \in \{0,1\}) \\ &= \exp\left\{y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right\}\ I(y_i \in \{0,1\}) \\ &= \exp\left\{y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right\}\ I(y_i \in \{0,1\}) \end{aligned} \tag{2.3}$$

e, portanto, pode-se observar que a distribuição de Bernoulli pertence à família exponencial com $a(y_i) = y_i$, $b(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, $c(\pi_i) = \ln(1-\pi_i)$ e $d(y_i) = I(y_i \in \{0,1\})$, sendo $I(u)$ uma função indicadora que assume o valor 1 quando a condição u é contemplada e 0 caso contrário.

Nos modelos lineares generalizados, relaciona-se as covariáveis a média da variável resposta através de uma função de ligação $g(\cdot)$ que pode ser a função obtida em $b(\cdot)$, se $a(y) = y$. Sendo assim, usando a função de ligação logística, tem-se a seguinte relação no modelo logístico:

$$g(\pi_i) = b(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}_i^T\boldsymbol{\beta} \tag{2.4}$$

sendo $\mathbf{X}_i^T = (X_{i,0}, X_{i,1}, \cdots, X_{i,k-1})$ o vetor de covariáveis associadas ao i-ésimo indivíduo e $\boldsymbol{\beta}^T = (\beta_0, \cdots, \beta_{k-1})$ o vetor com os coeficientes de regressão. Em geral, considera-se $X_{i,0} = 1\ \forall i$ e então β_0 é chamado de intercepto.

A partir da Equação (2.4), pode-se obter a probabilidade de sucesso através da expressão:

$$\pi_i = \frac{\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}, \tag{2.5}$$

ou de forma equivalente:

$$\pi_i = \frac{1}{1 + \exp\left(-\mathbf{X}_i^T\boldsymbol{\beta}\right)}. \tag{2.6}$$

Ao propor um modelo para um conjunto de dados $\mathbf{Y}^T = (Y_1, \cdots, Y_n)$, atribui-se uma função de distribuição ou de densidade para a variável aleatória de interesse, $p(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta})$. Essa função representa a crença que se tem sobre a distribuição da variável de interesse considerando conhecido o vetor paramétrico $\boldsymbol{\theta}$. Mas, na prática, esse vetor é desconhecido e deseja-se inferir sobre. Sendo assim, quando uma amostra dessa população é aplicada nessa função para inferir sobre $\boldsymbol{\theta}$, essa função passa a ser chamada de função de verossimilhança e passa a ser denotada por $l(\boldsymbol{\theta}; \mathbf{y})$, onde $\mathbf{y}$ representa o vetor de valores amostrados da variável de interesse.

Como o modelo de regressão logística supõe que as variáveis aleatórias são independentes e identicamente distribuídas, o vetor de parâmetros desconhecidos, nesse caso, equivale ao vetor com os efeitos $\boldsymbol{\beta}$. Sendo assim, a função de verossimilhança para o modelo de regressão logística é dada por:

$$l(\boldsymbol{\beta}; \mathbf{y}) = \prod_{i=1}^{n} \left[\left(\frac{\exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right)^{y_i} \left(\frac{1}{1 + \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right)^{1-y_i} \right]. \tag{2.7}$$

2.2.1 Inferindo sobre os parâmetros usando a abordagem Bayesiana

Sob a abordagem Bayesiana, o vetor paramétrico desconhecido $\boldsymbol{\beta}$ é considerado um vetor aleatório. Assim, é possível incorporar uma distribuição de probabilidade a este vetor. A distribuição a *priori*, denotada por $p(\boldsymbol{\beta})$, deve representar toda a crença probabilística sobre esse vetor anterior à amostragem dos dados. Dessa forma, a inferência sobre $\boldsymbol{\beta}$ é dada através da distribuição a *posteriori* $p(\boldsymbol{\beta}|\mathbf{y})$, que pode ser obtida através do Teorema de Bayes, combinando a função de verossimilhança, a distribuição a *priori* e a distribuição marginal dos dados, $p(\mathbf{y})$, obtendo:

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{l(\boldsymbol{\beta}; \mathbf{y}) p(\boldsymbol{\beta})}{p(\mathbf{y})}, \tag{2.8}$$

onde $p(\mathbf{y}) = \int l(\boldsymbol{\beta}; \mathbf{y}) p(\boldsymbol{\beta}) d\boldsymbol{\beta}$.

Note que a distribuição marginal não depende do vetor paramétrico $\boldsymbol{\theta}$. Então, a distribuição a *posteriori* pode ser reescrita como proporcional ao produto da função de verossimilhança e da distribuição a *priori* da seguinte forma:

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto l(\boldsymbol{\beta}; \mathbf{y}) p(\boldsymbol{\beta}). \tag{2.9}$$

Como neste trabalho o vetor paramétrico desconhecido é $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{k-1})^T$, assuma que, a *priori*, cada componente deste vetor é independente e segue uma distribuição Normal, com média μ e variância V. Dessa forma, para essa escolha particular, a distribuição a *priori* associada à este modelo é da seguinte forma:

$$\begin{aligned} p(\boldsymbol{\beta}) &= \prod_{j=0}^{k-1} p(\beta_j) \\ &= \prod_{j=0}^{k-1} \frac{1}{\sqrt{2\pi V}} \exp\left\{-\frac{1}{2V}\left(\beta_j - \mu\right)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi V}}\right)^k \exp\left\{-\frac{1}{2V}\sum_{j=0}^{k-1}\left(\beta_j - \mu\right)^2\right\} \end{aligned} \tag{2.10}$$

A partir da função de verossimilhança e da distribuição a *priori*, apresentadas nas Equações (2.7) e (2.10), respectivamente, a distribuição a *posteriori* do modelo de regressão logística pode ser escrita como:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}) &\propto \prod_{i=1}^{n}\left[\left(\frac{\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}\right)^{y_i}\left(\frac{1}{1+\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}\right)^{1-y_i}\right] \\ &\times \left(\frac{1}{\sqrt{2\pi V}}\right)^k \exp\left\{-\frac{1}{2V}\sum_{j=0}^{k-1}\left(\beta_j - \mu\right)^2\right\} \\ &\propto \prod_{i=1}^{n}\left[\left(\frac{\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}\right)^{y_i}\left(\frac{1}{1+\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)}\right)^{1-y_i}\right] \exp\left\{-\frac{1}{2V}\sum_{j=0}^{k-1}\left(\beta_j - \mu\right)^2\right\} \end{aligned} \tag{2.11}$$

Essa função não possui forma analítica conhecida e, portanto, será necessário recorrer a algum método, como os Métodos de Monte Carlo via cadeias de Markov (MCMC), para obter amostras da distribuição a *posteriori*.

2.2.2 Métodos de Monte Carlo via Cadeias de Markov

Os métodos de MCMC consistem em uma classe de algoritmos para amostrar de uma distribuição de probabilidade de interesse usando cadeias de Markov, sendo uma alternativa aos métodos não iterativos nos problemas em que as soluções analíticas tornam-se inviáveis ou complexas.

Metropolis-Hastings

O algoritmo de Metropolis-Hastings é um dos métodos de MCMC mais utilizados quando uma distribuição $p(\cdot)$ de interesse não possui expressão analítica conhecida. Nesse caso, utiliza-se uma distribuição $q(\cdot)$, denominada distribuição proposta, de modo que o

algoritmo gere um valor candidato de $q(\cdot)$ e que este seja aceito na cadeia a partir de uma condição probabilística que depende de $p(\cdot)$ e $q(\cdot)$. Supondo que o interesse seja obter uma amostra da distribuição a *posteriori*, o método consiste nos seguintes passos:

1. inicialize o contador de iterações j como 0 e atribua um valor inicial para o parâmetro $\boldsymbol{\beta}$ denotado como $\boldsymbol{\beta}^0$;
2. incremente o contador de j para $j+1$;
3. gere um valor proposto $\boldsymbol{\delta}$ usando uma função de distribuição conhecida denotada por $q(\boldsymbol{\delta}|\boldsymbol{\beta}^{j-1})$;
4. aceite o ponto gerado com probabilidade
$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\delta}|\mathbf{y})}{q(\boldsymbol{\delta}|\boldsymbol{\beta}^{j-1})}\frac{q(\boldsymbol{\beta}^{j-1}|\boldsymbol{\delta})}{p(\boldsymbol{\beta}^{j-1}|\mathbf{y})}\right\};$$
5. repita os passos 2, 3 e 4 até obter convergência da cadeia.

Caso o ponto gerado não seja aceito, significa que ele é rejeitado e não faz parte da cadeia de amostragem. Nesse caso, o valor atual do parâmetro $\boldsymbol{\beta}$ permanece o mesmo, e o algoritmo passa para a próxima iteração, repetindo os passos de geração de um novo valor proposto e avaliação da aceitação.

A convergência das cadeias de Markov é esperada após um número suficientemente grande de iterações e um período de aquecimento, que é o intervalo inicial necessário para a cadeia começar a convergir. Uma cadeia de Markov é considerada convergente quando a distribuição de probabilidade das amostras estabiliza e não muda significativamente ao longo das iterações.

É importante observar que os parâmetros amostrados podem apresentar alta autocorrelação. Uma solução para esse problema é usar um espaçamento de ordem k, que seleciona uma amostra a cada k iterações, corrigindo a autocorrelação da cadeia. O valor de k pode ser determinado pela análise do gráfico de autocorrelação dos parâmetros amostrados. Maiores informações podem ser vistas em Gamerman e Lopes (2006).

Monte Carlo Hamiltoniano (HMC)

O método de amostragem conhecido como Monte Carlo Hamiltoniano (HMC) é uma abordagem avançada que utiliza a dinâmica Hamiltoniana para gerar amostras independentes de distribuições a *posteriori*. A técnica simula o movimento de uma partícula em um espaço de fase expandido, aproveitando informações do gradiente do

log da função de probabilidade para orientar o processo de amostragem (NEAL, 2011). Consequentemente, uma cadeia HMC bem ajustada aceitará propostas com uma taxa muito maior em comparação com o algoritmo Metropolis-Hastings tradicional (GELMAN; GILKS; ROBERTS, 1997).

Durante a simulação, o método integra numericamente a equação de Hamilton para gerar trajetórias na forma de pares de coordenadas e impulsos. As amostras são obtidas por meio da aceitação ou rejeição de propostas de movimento, utilizando a probabilidade de Metropolis-Hastings (BETANCOURT, 2018). Esse método permite explorar o espaço de parâmetros de maneira eficiente e obter amostras de alta qualidade da distribuição a *posteriori* desejada.

Além dos parâmetros de interesse $\boldsymbol{\beta} = (\beta_1, ..., \beta_{k-1})^T$, o HMC inclui um vetor de variáveis de momento $\mathbf{m} = (m_1, ..., m_k)^T$ e realiza as amostragens a partir da distribuição conjunta entre $\boldsymbol{\beta}$ e $\mathbf{m}$:

$$\pi(\mathbf{m}, \boldsymbol{\beta}) = \pi(\mathbf{m}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \tag{2.12}$$

Em geral, atribui-se a $\mathbf{m}$ uma distribuição normal multivariada, independente de $\boldsymbol{\beta}$:

$$\mathbf{m} \sim N_k(\mathbf{0}, \mathbf{M})$$

onde $\mathbf{M}$ é uma matriz de métrica Euclidiana que transforma a distribuição de interesse, tornando a amostragem mais eficiente.

A partir da Equação (2.12), define-se o Hamiltoniano como:

$$\begin{aligned} H(\mathbf{m}, \boldsymbol{\beta}) &= -\ln \pi(\mathbf{m}, \boldsymbol{\beta}) \\ &= -\ln \pi(\mathbf{m}|\boldsymbol{\beta}) - \ln \pi(\boldsymbol{\beta}) \\ &= T(\pi(\mathbf{m}|\boldsymbol{\beta})) + V(\pi(\boldsymbol{\beta})) \end{aligned}$$

em que a função $T(\pi(\mathbf{m}|\boldsymbol{\beta}))$ é a energia cinética e $V(\pi(\boldsymbol{\beta}))$ é a energia potencial.

O algoritmo de atualização dos valores de $\boldsymbol{\beta}$ ocorre em duas etapas, antes de um processo de aceitação semelhante ao do algortimo de Metropolis-Hastings. Primeiro, valores para o vetor $\mathbf{m}$ são gerados de forma independente dos valores atuais de $\boldsymbol{\beta}$, seguindo a distribuição $\pi(\mathbf{m})$. Essa etapa garante que o momento $\mathbf{m}$ não persista ao longo das iterações.

Na segunda etapa, o par $(\boldsymbol{\beta}, \mathbf{m})$, composto pelos valores atuais dos parâmetros $\boldsymbol{\beta}$ e os novos valores gerados para as variáveis de momento, é evoluído pelas equações Hamiltonianas,

$$\begin{aligned} \frac{d\beta_j}{di} &= +\frac{\partial H}{\partial m_j} = +\frac{\partial T}{\partial m_j} \\ \frac{dm_j}{di} &= -\frac{\partial H}{\partial \beta_j} = -\frac{\partial T}{\partial \beta_j} - \frac{\partial V}{\partial \beta_j} \end{aligned}$$

em que $j = 1, 2, ..., k$ e i representa a iteração atual do método.

Para resolver esse sistema de equações diferenciais, utiliza-se o método de integração numérica chamado de composição de Leapfrogs. Primeiro, esse método gera valores para o momento $\mathbf{m}$. Depois, intercala atualizações de meio passo para o momento $\mathbf{m}$ com uma atualização de passo completo para $\boldsymbol{\beta}$.

Ao final do processo de integração, o estado do momento e da posição resultante é denotado por $(\mathbf{m}^*, \boldsymbol{\beta}^*)$. A última etapa do algoritmo é avaliar se o estado $(\mathbf{m}^*, \boldsymbol{\beta}^*)$ será aceito ou não. A probabilidade de aceitação de $(\mathbf{m}^*, \boldsymbol{\beta}^*)$ é determinada por

$$\min(1, \exp\{H(\mathbf{m}, \boldsymbol{\beta}) - H(\mathbf{m}^*, \boldsymbol{\beta}^*)\})$$

Caso a proposta seja rejeitada, os valores atuais dos parâmetros $\boldsymbol{\beta}$ são mantidos e inicia-se a próxima iteração. O processo descrito é repetido até obter-se um número suficiente de amostras representativas da distribuição a *posteriori* desejada.

2.2.3 Interpretação dos Parâmetros

A interpretação dos parâmetros do modelo de regressão logística como uma medida de associação entre a ocorrência de um evento de interesse e uma covariável pode ser obtida através da razão de chances.

A chance de um evento de interesse ocorrer é a razão das probabilidades de ocorrência e de não ocorrência desse mesmo evento. Considerando π_i como a probabilidade de um evento ocorrer, a chance de ocorrência *(odds)* pode ser escrita como:

$$odds = \frac{\pi_i}{1 - \pi_i} \tag{2.13}$$

A partir da chance de ocorrência, pode-se definir a razão de chances, *Odds Ratio* (OR), de um determinado evento entre dois grupos, que consiste na razão entre a chance de

ocorrência do evento para o primeiro grupo ($odds_1$) e a chance de ocorrência para o segundo grupo ($odds_2$).

$$OR = \frac{odds_1}{odds_2} \tag{2.14}$$

Considere um modelo de regressão logística simples, com intercepto e uma covariável, em que $X_i = 1$ indica a presença de um fator e $X_i = 0$, a ausência. Sendo assim, temos que as probabilidades de sucesso e fracasso dada a ocorrência do fator são, respectivamente, $P(Y_i = 1|X_i = 1)$ e $P(Y_i = 0|X_i = 1)$. Através da Equação (2.14), temos que a chance de ocorrência de $X_i = 1$ é dada por:

$$odds(X_i = 1) = \frac{P(Y_i = 1|X_i = 1)}{P(Y_i = 0|X_i = 1)} = \frac{P(Y_i = 1|X_i = 1)}{1 - P(Y_i = 1|X_i = 1)} \tag{2.15}$$

e através da Equação (2.6), pode ser escrita da seguinte forma:

$$odds(X_i = 1) = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{1 - \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}} = e^{\beta_0+\beta_1} \tag{2.16}$$

A chance de ocorrência de $X_i = 0$ pode ser calculada de forma análoga e resulta em $odds(X_i = 0) = e^{\beta_0}$. Logo, a razão de chances é dada por:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \tag{2.17}$$

e pode ser interpretada da seguinte forma para covariáveis qualitativas:

- Se $OR > 1$, a chance de sucesso na presença do fator é OR vezes ou $(OR - 1)100\%$ maior do que a chance de ter sucesso na ausência.
- Se $OR < 1$, a chance de sucesso na presença do fator é OR vezes ou $(1 - OR)100\%$ menor do que a chance de ter sucesso na ausência.
- Se $OR = 1$, a chance de sucesso é a mesma tanto na presença quanto na ausência do fator.

Já para as covariáveis numéricas, a interpretação é dada da seguinte forma:

- Se $OR > 1$, ao aumentar a covariável em 1 unidade, as chances de sucesso aumentam OR vezes ou $(OR - 1)100\%$.
- Se $OR < 1$, ao aumentar a covariável em 1 unidade, as chances de sucesso diminuem OR vezes ou $(1 - OR)100\%$.

- Se $OR = 1$, a mudança nos valores da covariável não está relacionada a mudança nas chances de sucesso.

No modelo de regressão logística múltiplo a interpretação é equivalente a do modelo simples e a razão de chances pode ser generalizada por

$$OR_j = e^{\beta_j} \tag{2.18}$$

com $j = 1, ..., k-1$.

2.2.4 Comparação de modelos

Pode-se querer ajustar diferentes modelos com o objetivo de avaliar o desempenho de cada um. O desempenho dos modelos pode ser medido tanto pela sua bondade de ajuste aos dados quanto pela sua capacidade preditiva. Ao comparar diferentes modelos, é possível identificar quais variáveis explicativas são mais relevantes e fornecem uma compreensão mais completa do fenômeno em análise. Existem diversos critérios para a comparação de modelos, cada um com suas vantagens e limitações. Neste trabalho, foi utilizado o Critério de Informação da Deviance (DIC).

O DIC é um critério de seleção de modelo que leva em consideração tanto o ajuste do modelo aos dados observados quanto a complexidade do modelo. Valores menores de DIC indicam um melhor ajuste e menor complexidade (SPIEGELHALTER et al., 2002). Para calcular o DIC, primeiro é necessário obter a *deviance* para o modelo em análise ($D(\overline{\beta})$) e a *deviance* penalizada (p_D), obtida através da diferença da média das *deviances* das iterações do MCMC, denotada por $\overline{D(\beta)}$, e a *deviance* do modelo, $D(\overline{\beta})$.

Finalmente, o DIC é calculado pela soma da *deviance* penalizada e da *deviance* do modelo em análise:

$$DIC = p_D + D(\overline{\beta}), \tag{2.19}$$

onde $p_D = \overline{D(\beta)} - D(\overline{\beta})$.

2.2.5 Avaliação da capacidade preditiva

Para avaliar a capacidade preditiva do modelo, uma parte específica dos dados é reservada exclusivamente para esse propósito. O ajuste do modelo é realizado sem utilizar essa parte separada, garantindo que ele não tenha conhecimento desses valores, e em

seguida, é feita a previsão para esses dados. Uma abordagem comum para avaliar a capacidade preditiva é construção de uma matriz de confusão. Essa matriz consiste em uma tabela que cruza os valores reais de classificação com os valores previstos pelo modelo, permitindo uma análise detalhada das previsões realizadas.

Nas linhas, tem-se, respectivamente, os elementos que possuem a característica de interesse e os que não possuem essa mesma característica. Já nas colunas, na primeira coluna, os elementos que foram classificados pelo modelo como tendo a característica de interesse e na segunda, os que foram classificados como não tendo a característica de interesse.

Ao cruzar as linhas e as colunas, tem-se os elementos:

- Verdadeiros Positivos (VP): ou seja, os elementos que possuem a característica de interesse e foram corretamente classificados;
- Falsos Negativos (FN): os elementos que possuem a característica de interesse e foram classificados de forma errada.
- Falsos Positivos (FP): elementos que não possuem a característica de interesse e foram classificados de forma errada;
- Verdadeiros Negativos (VN): elementos que não possuem a característica de interesse e foram classificados corretamente.

Tabela 2: Matriz de Confusão

	$\hat{y}_i = 1$	$\hat{y}_i = 0$
$y_i = 1$	Verdadeiro Positivo	Falso Negativo
$y_i = 0$	Falso Positivo	Verdadeiro Negativo

A partir da tabela acima, pode-se extrair diversas métricas para avaliar a capacidade preditiva do modelo, como a Sensibilidade, que mede a capacidade do modelo de classificar corretamente os elementos que possuem a característica de interesse e pode ser calculada como a proporção de Verdadeiros Positivos dentre os elementos que possuem a característica de interesse, possuindo a seguinte expressão

$$S = \frac{VP}{VP + FN}. \tag{2.20}$$

Além disso, pode-se medir também a capacidade do modelo classificar corretamente os elementos que não possuem a característica de interesse, medida conhecida como Especificidade, possuindo a seguinte expressão

$$E = \frac{VN}{VN + FP}. \tag{2.21}$$

Outra medida que pode ser extraída da tabela é a Acurácia. Essa medida avalia a quantidade de acertos nas classificações feitas pelo modelo ajustado de forma geral e pode ser calculada como a proporção das classificações corretas dentre o total de classificações, possuindo a seguinte expressão

$$A = \frac{VP + VN}{VP + FP + FN + VN}. \tag{2.22}$$

Curva ROC

Ao ajustar um modelo de regressão logística, consegue-se obter estimativas para a probabilidade de sucesso da Bernoulli. Essas probabilidades são variáveis contínuas. Pode-se criar uma relação dessas variáveis com a variável resposta binária da seguinte forma: se a probabilidade estimada for maior ou igual do que um ponto de corte escolhido, então a variável binária receberá o valor 1 e caso contrário receberá o valor 0. O melhor ponto de corte é aquele que maximiza as medidas de sensibilidade e especificidade.

Para isso, pode-se utilizar a curva ROC (*Receiver Operating Characteristic*). Segundo Prati, Batista e Monard (2008), é uma forma gráfica de representar a relação entre a sensibilidade e a especificidade do modelo, onde se tem a sensibilidade em função da proporção de falsos positivos (1 - Especificidade), para diferentes pontos de corte. Assim, é possível observar o ponto de corte que otimiza a sensibilidade em função da especificidade, sendo o ponto que se encontra mais próximo do canto superior esquerdo. Veja um exemplo de Curva ROC na Figura 1.

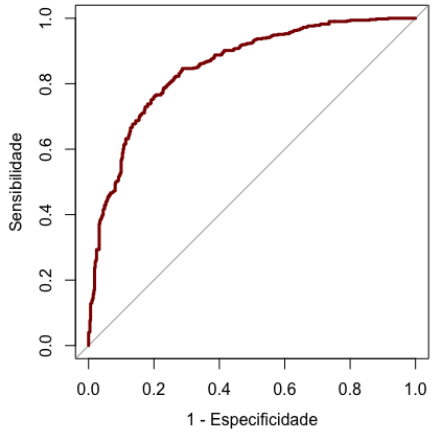

Figura 1: Exemplo - Curva ROC

Além disso, a curva ROC permite avaliar o modelo através da Área Abaixo da Curva (AUC), que resume a curva ROC em um único valor (HUANG; LING, 2005). Esta área será sempre menor que 1 e quanto mais próxima de 1, melhor será a qualidade do ajuste do modelo.

3 Análise dos Resultados

Neste Capítulo, serão apresentados os resultados deste trabalho, juntamente com as análises decorrentes desses resultados. Inicialmente, na Seção 3.1, serão apresentados os resultados e análises de um estudo simulado para o ajuste de um modelo logístico bayesiano. Em seguida, na Seção 3.2, será exposta a análise descritiva realizada da base de dados. Por fim, na Seção 3.3, serão discutidos os resultados e análises dos modelos ajustados utilizando os dados reais, oferecendo uma visão aprofundada sobre a relação entre as variáveis explicativas e o *churn*. Essas análises ajudarão a entender o estudo feito e as implicações práticas dos resultados obtidos. Todas as análises foram realizadas utilizando o software estatístico R (R Core Team, 2014).

3.1 Estudo Simulado

A fim de verificar a capacidade de estimação dos parâmetros, foi aplicado o modelo logístico bayesiano a um conjunto de dados simulados, fixando valores para os parâmetros desconhecidos. Suponha que a matriz $\mathbf{X}$ possui um intercepto e duas variáveis explicativas, com os seguintes valores $\boldsymbol{\beta}^T = (1; -2; 3)$, sendo $X_{i,1} = 1$, $X_{i,2} \sim Normal(0, 1)$ e $X_{i,3} \sim Uniforme(0, 1)$, com $i = 1, ..., 7043$. O banco de dados gerado foi dividido em duas bases distintas: 80% para o ajuste do modelo (base de treino) e os 20% restantes para a realização de previsões e avaliação da capacidade preditiva do modelo (base de teste). Essa divisão foi realizada através do pacote *caret* (KUHN, 2020).

Para o ajuste do modelo utilizou-se o pacote *rstanarm* (GABRY; GOODRICH; VEHTARI, 2022). Ao total, foram realizadas 2.000 iterações, com período de aquecimento de 1.000 e espaçamento 1, o que resultou em amostras a *posteriori* de tamanho 1.000 para 4 cadeias. A Figura 2 apresenta a convergência das cadeias dos parâmetros, juntamente com seus histogramas a *posteriori* e gráficos de autocorrelação. Para os histogramas, as linhas em verdes representam os valores verdadeiros dos parâmetros, enquanto as linhas vermelhas correspondem às estimativas pontuais a *posteriori* dos parâmetros

desconhecidos. Já as linhas azuis indicam os respectivos intervalos de credibilidade de 95%. Observa-se indícios de convergência, com as médias a *posteriori* se aproximando dos valores verdadeiros e os intervalos abrangendo esses mesmos valores.

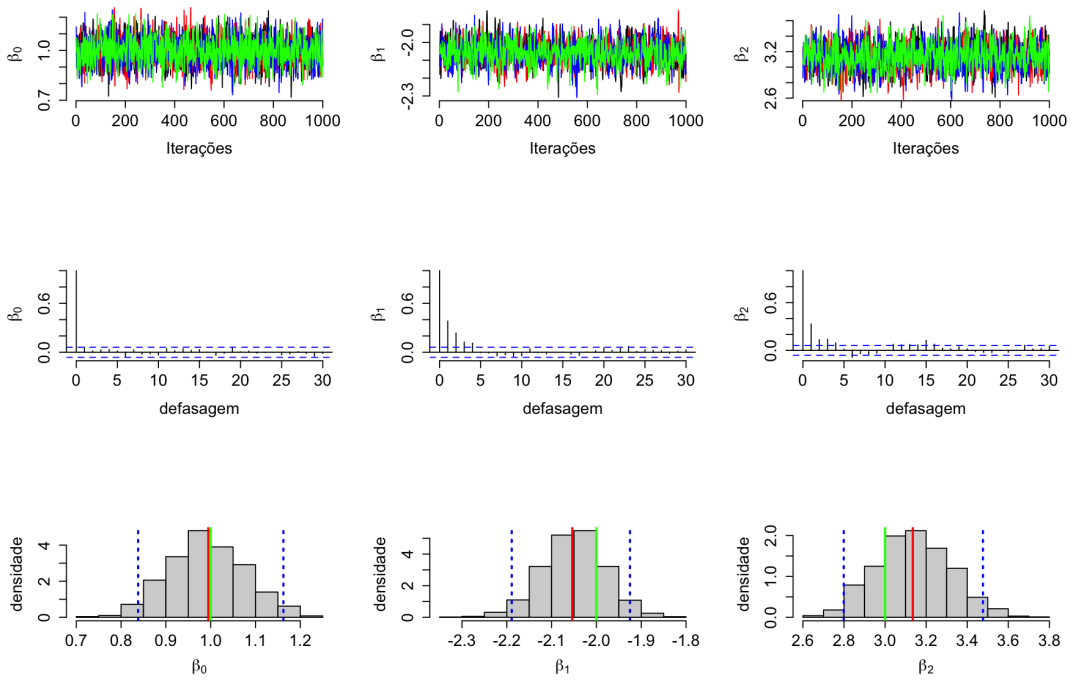

Figura 2: Traços das cadeias, autocorrelação e histogramas das amostras dos parâmetros com dados simulados.

A fim de avaliar a precisão das estimativas do modelo, uma terceira variável explicativa, $X_{i,3} \sim Gama(2, 1)$, foi adicionada ao conjunto de dados e o modelo proposto foi reajustado. A estimativa pontual do coeficiente dessa variável ficou próxima de zero, conforme ilustrado na Figura 3. Isso indica que a influência dessa variável sobre o resultado do modelo é insignificante, como era esperado.

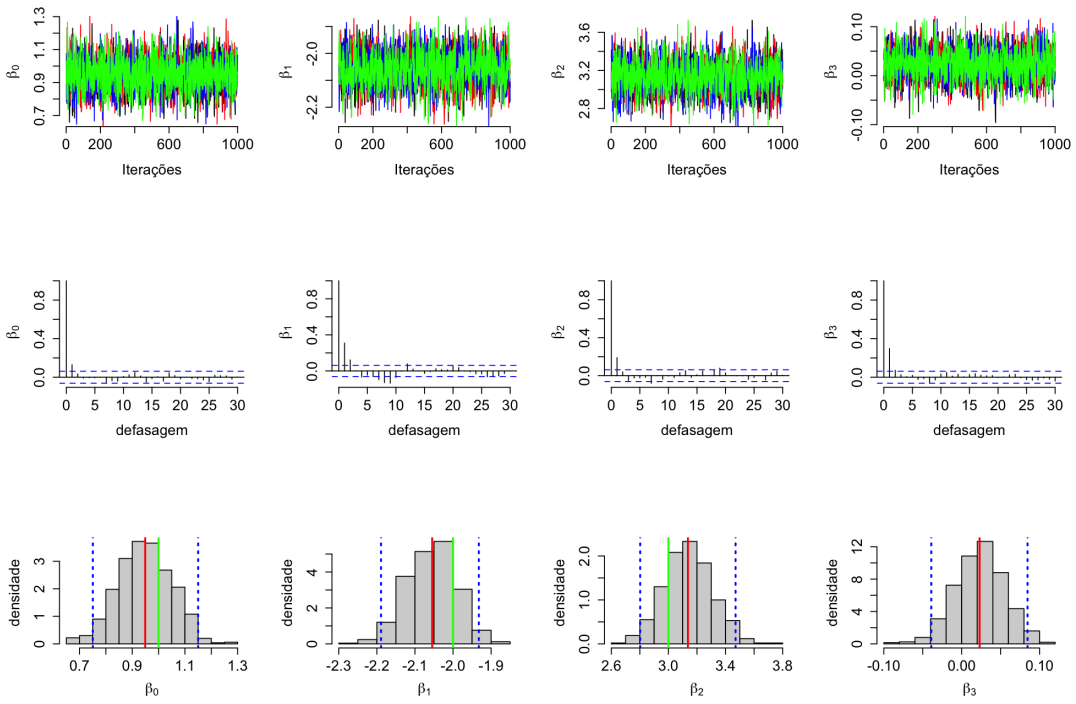

Figura 3: Traços das cadeias, autocorrelação e histogramas das amostras dos parâmetros dos dados simulados com a nova variável explicativa inserida.

Para avaliar a capacidade preditiva do modelo, foram feitas previsões para cada iteração da base de teste. Utilizando o ponto de corte encontrado através da curva ROC para cada iteração, as previsões foram classificadas, em 0 ou 1, e calcularam-se as medidas de avaliação da capacidade preditiva do modelo mencionadas anteriormente para cada iteração. A média a *posteriori* das medidas calculadas foram as seguintes: AUC $= 0,8950$, Acurácia $= 0,8139$, Sensibilidade $= 0,8226$ e Especificidade $= 0,8119$. Essas métricas fornecem uma visão abrangente do desempenho do modelo na classificação dos dados de teste, demonstrando uma precisão geral sólida, bem como uma capacidade de identificar corretamente os casos positivos e negativos.

3.2 Análise Descritiva

A fim de compreender melhor a base de dados descrita na Seção 2.1, é fundamental realizar uma análise descritiva. Neste contexto, a análise descritiva busca identificar se as variáveis em questão desempenham um papel significativo no comportamento de *churn*

dos clientes, contribuindo para uma compreensão mais abrangente desse fenômeno e do perfil dos clientes desta empresa.

A Figura 4 apresenta as proporções de clientes *churn* e não *churn* de acordo com as variáveis que indicam o sexo do cliente, se o cliente é casado ou não, se o cliente é jovem ou não, e se o cliente é idoso ou não. Os gráficos nos permitem observar que as variáveis que indicam o sexo e se o cliente é jovem, não parecem ter influência no fenômeno do *churn*, já que o percentual de clientes *churn* para cada um dos sexos e entre o cliente ser ou não jovem é bem próximo. Por outro lado, pode-se observar que a proporção de clientes *churn* entre os idosos é maior, indicando que um idoso parece ser mais suscetível ao *churn* do que um não idoso. O mesmo acontece para a variável que indica se o cliente é casado ou não, a proporção de clientes *churn* entre os não casados é mais alta, indicando que clientes que não são casados têm maior possibilidade de *churn* do que os clientes casados.

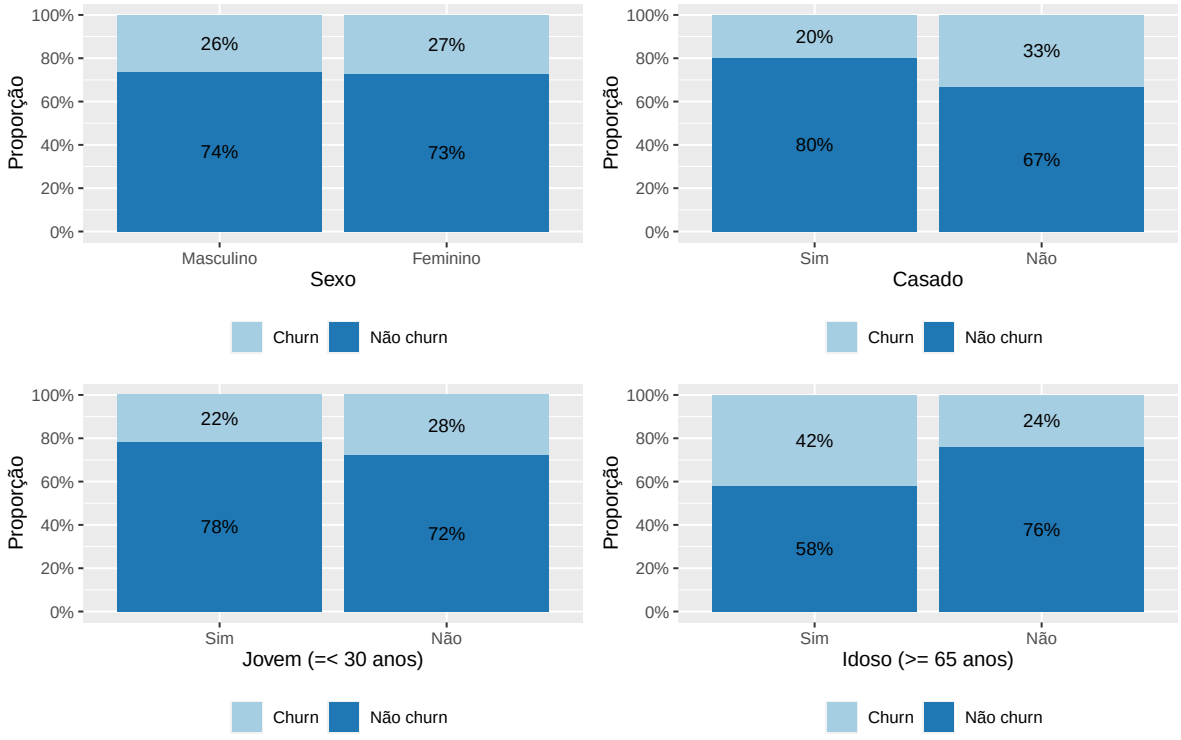

Figura 4: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis Sexo, Casado, Jovem e Idoso.

A Figura 5 reflete as proporções de clientes *churn* e não *churn* de acordo com as variáveis Dependentes, que indica se o cliente possui dependentes ou não, e Indicação, que indica se o cliente recomendou a empresa para alguém ou não. Ao analisar os gráficos, pode-se observar que a proporção de clientes *churn* é maior entre os clientes que não

possuem dependentes do que entre os que possuem. Assim como para a variável Indicação, onde a proporção de clientes *churn* que não fizeram uma indicação é maior, indicando que essas variáveis parecem ter influência no fenômeno do *churn*.

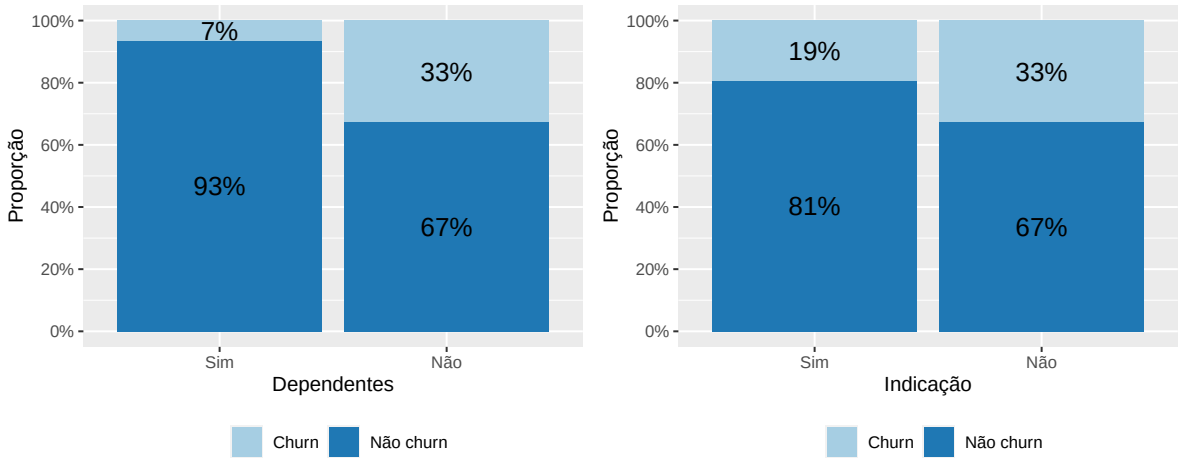

Figura 5: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis Dependentes e Indicação.

A Figura 6 permite analisar os clientes *churn* e não *churn* de acordo com o tipo de internet contratada e região de residência. A variável de região foi criada a partir da variável que continha a cidade de residência do cliente. Ao analisar a figura, pode-se observar que os clientes que possuem internet fibra ótica aparentam ser mais suscetíveis ao *churn* do que os clientes que possuem os demais tipos. Já para a variável de região, a porcentagem de clientes *churn* em cada uma das regiões é bem similar, indicando que essa variável não parece ter influência no *churn*.

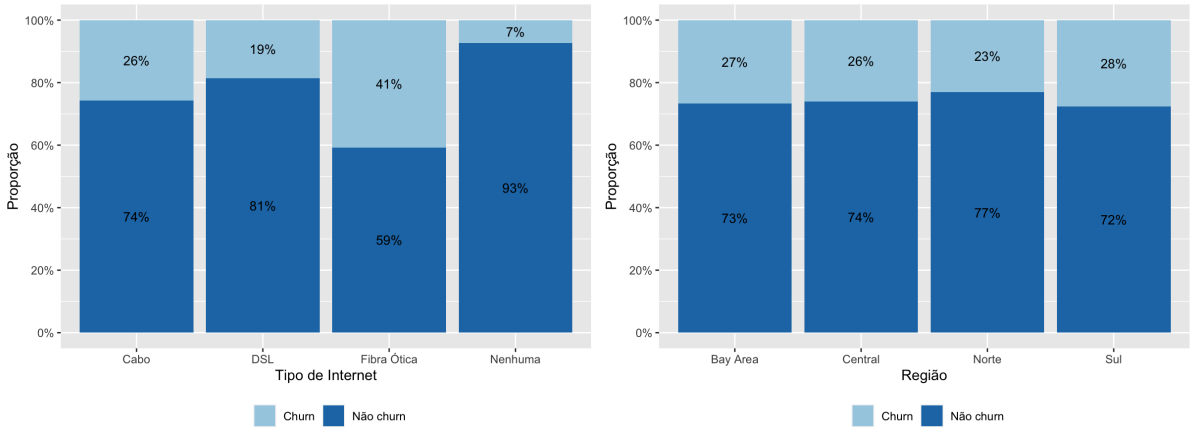

Figura 6: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis Tipo de Internet e Região.

A Figura 7 permite analisar os clientes de acordo com os serviços oferecidos pela empresa. Pode-se observar que clientes que possuem os serviços de internet e dados ilimitados têm maior possibilidade de *churn*, considerando a proporção de clientes que deram *churn* dentre os que possuem esses serviços. Já para telefone e linhas múltiplas, pode-se observar que esses serviços parecem não ter influência no fenômeno, já que a proporção de clientes *churn* dentre os clientes que possuem e os que não possuem esses serviços é bem próxima.

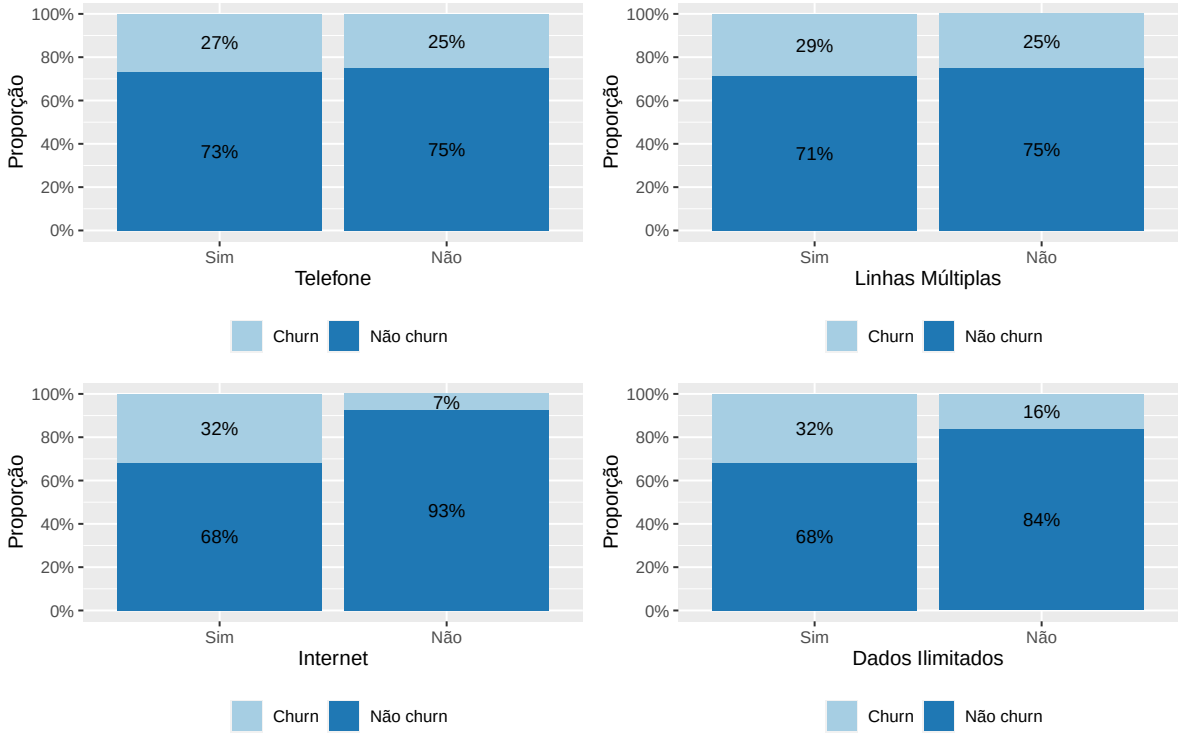

Figura 7: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis de serviço de Telefone, Internet, Linhas Múltiplas e Dados Ilimitados.

A Figura 8 permite analisar os clientes *churn* e não *churn* de acordo com o tipo de contrato, método de pagamento e se o cliente optou pela conta digital ou não. Ao analisar a figura, pode-se observar que os clientes que possuem o tipo de contrato mensal aparentam ser mais propensos ao *churn* que os clientes que optaram pelos tipos de contrato anual e bianual, assim como os clientes que utilizam cartão de débito e cheque como método de pagamento parecem ser mais propensos ao *churn* do que os que utilizam o cartão de crédito. Optar por receber a conta digital, no lugar da fatura em papel, também parece influenciar no fenômeno do *churn*, já que a proporção de clientes *churn* que recebem a fatura de forma digital é maior do que a dos clientes que recebem a fatura em papel.

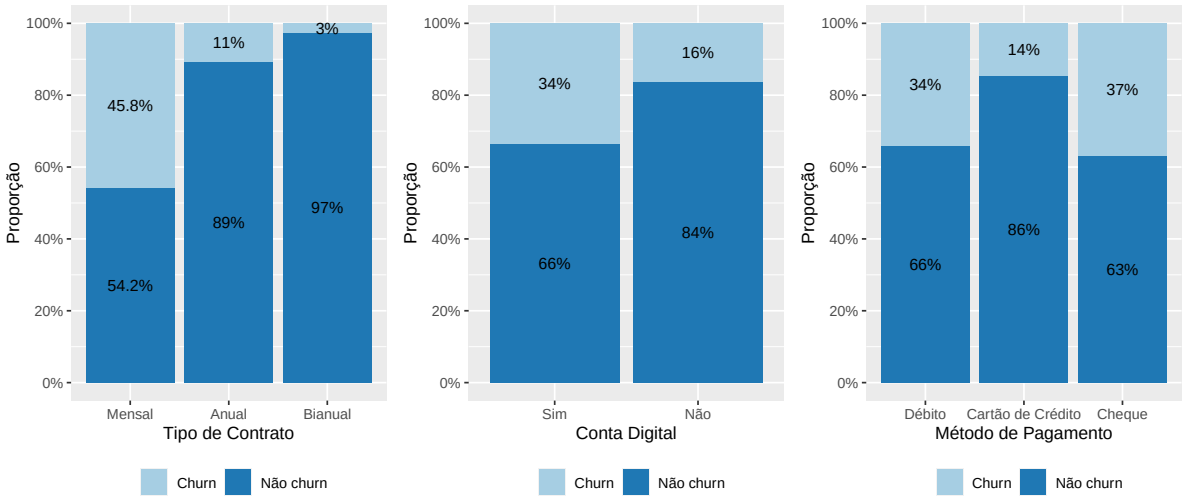

Figura 8: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis de Tipo de Contrato, Conta Digital e Método de Pagamento.

A Figura 9 permite analisar os clientes que utilizam os dados de internet para acessar serviços de *streaming* de TV, música ou filmes. Pode-se perceber que as porcentagens de clientes *churn* e não *churn* é muito parecida para qualquer uma das três categorias de *streaming*. Apesar de haver diferença entre os clientes *churn* que utilizam e os clientes *churn* que não utilizam os dados para acessar conteúdos de *streaming*, essas variáveis não parecem influenciar no fenômeno do *churn*.

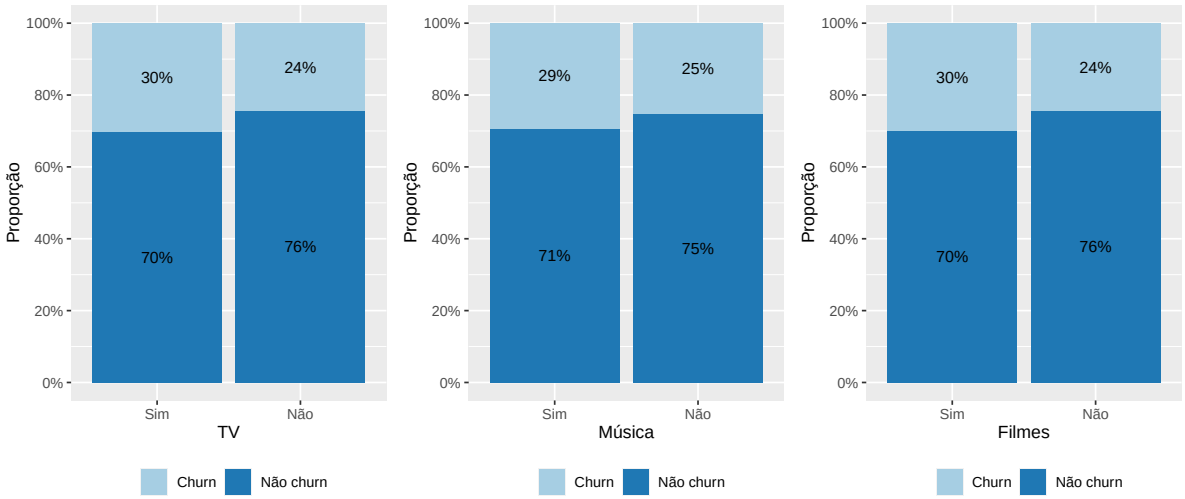

Figura 9: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis que indicam se o cliente utiliza os dados de internet para serviço de *streaming* (TV, Música e Filmes).

A Figura 10 permite analisar os clientes de acordo com os serviços adicionais oferecidos pela empresa. Pode-se observar que os clientes que não possuem os serviços adicionais contratados têm maior possibilidade de *churn*, se comparado com os clientes que possuem esses serviços, indicando que essas variáveis parecem ter influência no fenômeno do *churn*.

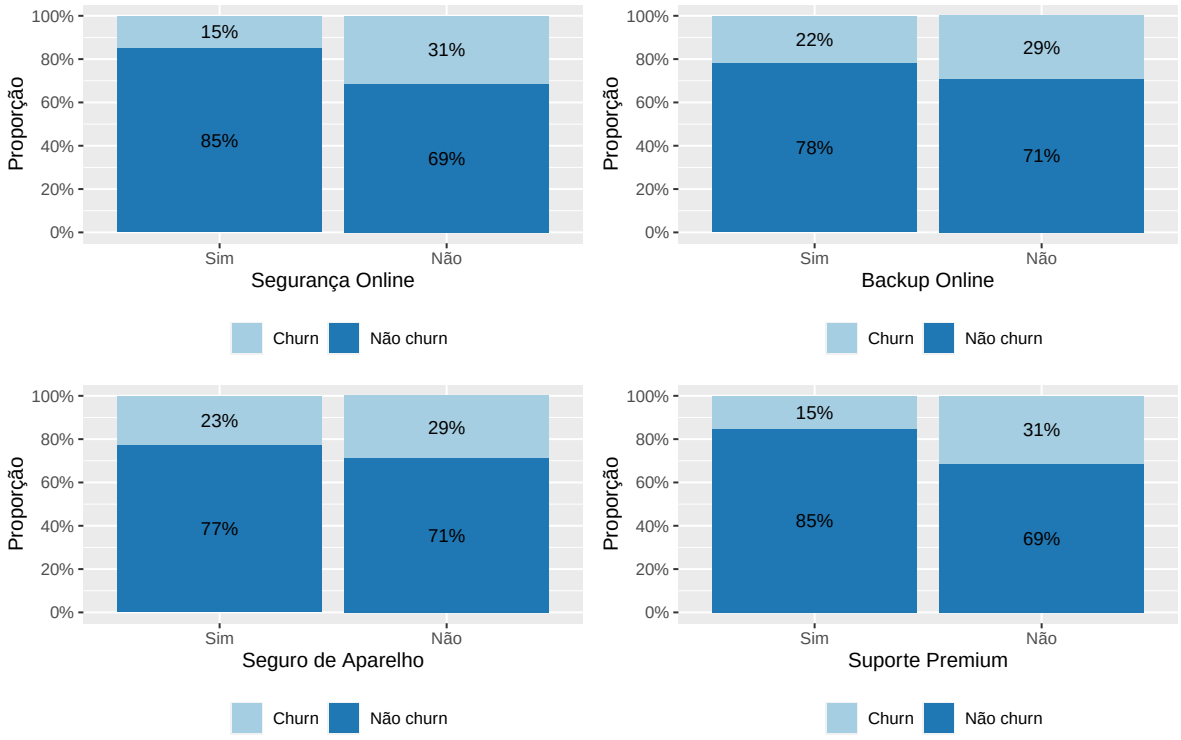

Figura 10: Proporção dos clientes *churn* e não *churn* de acordo com as variáveis de serviço de Segurança Online, Backup Online, Seguro de Aparelho e Suporte Premium.

A Figura 11 permite analisar os clientes *churn* e não *churn* de acordo com a última oferta de marketing. Ao analisar a figura, pode-se observar que diferentes ofertas de marketing têm impactos distintos no *churn* dos clientes. As ofertas A e B registraram proporção de clientes *churn* de 7% e 12%, respectivamente, indicando que essas opções foram relativamente eficazes em manter os clientes engajados. Já para a oferta E, a proporção de clientes *churn* é de 52,9%, indicando que clientes que tiveram essa oferta como última aceita, são mais propensos ao *churn*.

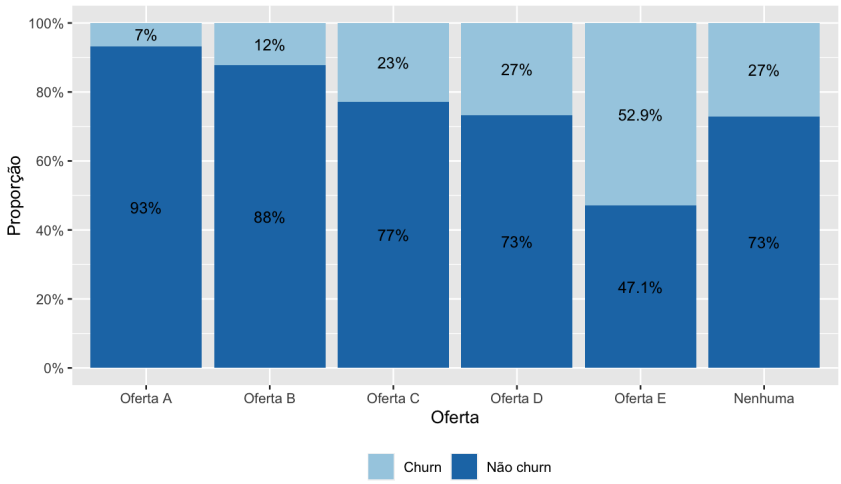

Figura 11: Proporção dos clientes *churn* e não *churn* de acordo com a variável Oferta.

A Figura 12 permite analisar os clientes *churn* e não *churn* de acordo com o valor da mensalidade atual, pelo tempo em meses que o cliente está na empresa representado, pela variável Fidelidade, e pelo custo de vida útil do cliente (CLTV). Ao analisar a figura, pode-se observar que os clientes com valores mais altos de mensalidade estão mais propensos ao *churn* e que os clientes que estão a mais tempo na empresa e possuem CLTV mais altos, são menos propensos. Indicando que essas três variáveis parecem ser significativas para explicar o *churn*.

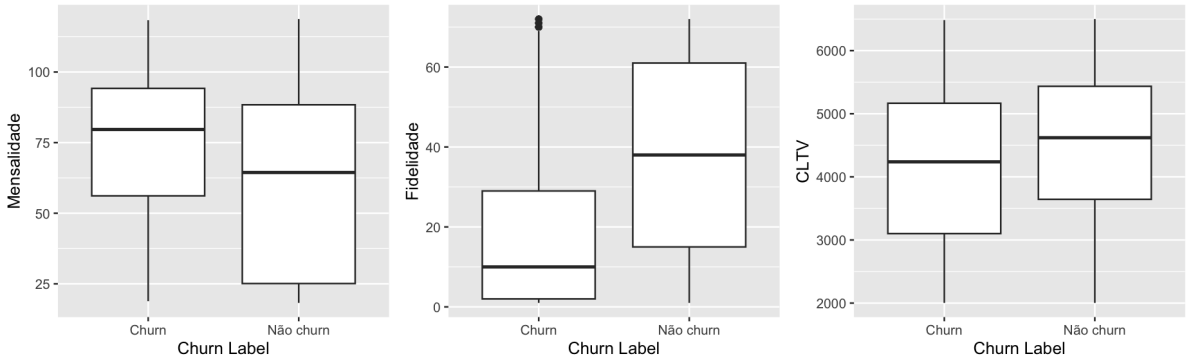

Figura 12: Relação dos clientes *churn* e não *churn* com as variáveis Mensalidade, Fidelidade e CLTV.

A Figura 13 permite analisar os clientes *churn* e não *churn* de acordo com a tarifa média de longa distância e volume médio de downloads dos clientes. Ao analisar a figura, pode-se observar que não parece haver diferença entre os dois tipos de clientes para essas variáveis, logo, não parecem influenciar no *churn*.

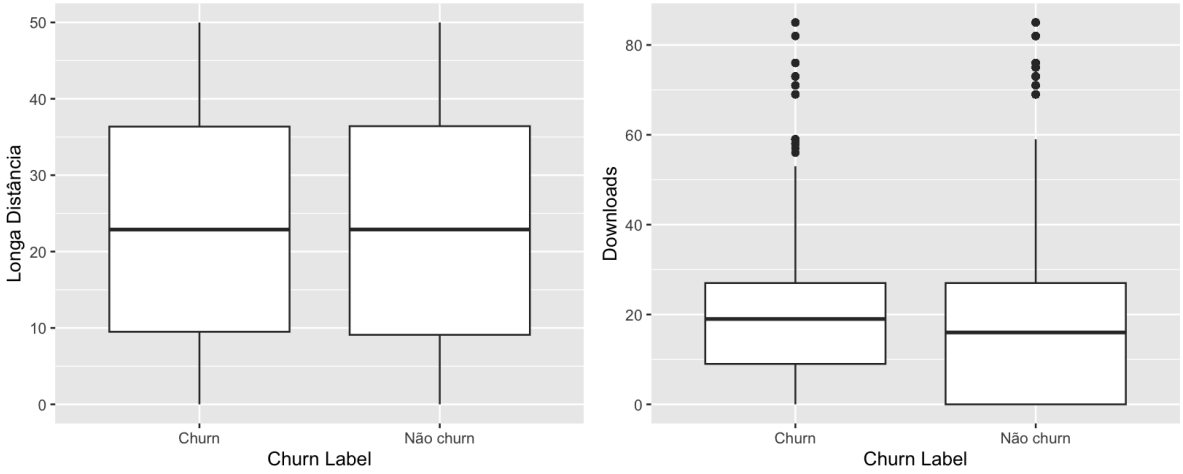

Figura 13: Relação dos clientes *churn* e não *churn* com as variáveis Longa Distância e Downloads.

3.3 Ajuste dos Modelos

Nesta seção, será apresentada a análise do ajuste dos modelos propostos. Para a realização das análises foi adotada a metodologia de divisão dos dados em duas bases distintas: treino e teste. A base de treino possui 80% das observações, contemplando um valor absoluto de 5.363 observações e será utilizada para o ajuste do modelo. Já a base de teste contempla os 20% restantes, sendo 1.407 observações e será utilizada para realizar previsões e avaliar a capacidade preditiva do modelo. Essa divisão foi realizada através do pacote *caret* (KUHN, 2020). O processo de seleção utilizado considera, ao realizar a divisão da base de dados, que as distribuições das observações de ambas as bases possuam comportamentos equivalentes.

Os modelos foram construídos considerando diferentes conjuntos de variáveis explicativas com o objetivo de comparar a performance dos modelos e identificar quais variáveis são relevantes para a previsão do *churn*. A relevância das variáveis é avaliada por meio do intervalo de credibilidade, se o intervalo incluir o zero, isso indica que a variável em questão pode não ser significativa para o modelo, pois existe uma probabilidade razoável de que seu efeito seja nulo. Por outro lado, se o intervalo não incluir o zero, isso sugere que a variável é significativa, pois há evidências de que seu efeito é diferente de zero.

Inicialmente, o primeiro modelo, doravante denominado de Modelo Completo, é desenvolvido considerando todas as variáveis disponíveis descritas na Seção 2.1. Esse modelo tem como objetivo fornecer uma visão geral do ajuste inicial e servir de ponto

de referência para as etapas subsequentes. Em seguida, foi ajustado o segundo modelo, Modelo 2, levando em consideração apenas as variáveis que se mostraram estatisticamente significativas no primeiro modelo. Essa abordagem visa eliminar variáveis que não contribuem de forma significativa para explicar o fenômeno do *churn*. Posteriormente, foi ajustado o terceiro modelo, Modelo 3, utilizando apenas as variáveis que se mantiveram significativas no segundo modelo.

Os modelos foram ajustados utilizando o pacote *rstanarm* (GABRY; GOODRICH; VEHTARI, 2022) e as variáveis categóricas foram transformadas em variáveis *dummies*. Para cada modelo, foram realizadas 2.000 iterações, sendo 1.000 destinadas ao período de aquecimento e as outras 1.000 para obter as distribuições a *posteriori* dos parâmetros de interesse. O ajuste foi realizado com quatro cadeias e espaçamento igual a 1 para os três modelos.

A Figura 14 exibe a evolução dos valores dos parâmetros ao longo das iterações do Método de Monte Carlo Hamiltoniano (HMC) para o Modelo Completo. De maneira similar, as Figuras 15 e 16 apresentam os traços das cadeias para o Modelo 2 e o Modelo 3, respectivamente. Nesses gráficos, pode-se observar a convergência dos parâmetros, uma vez que as cadeias demonstram valores dentro de um intervalo com comportamento estável e sem grandes flutuações. Esse comportamento é um indicativo de que as estimativas dos parâmetros estão convergindo para valores consistentes e confiáveis.

Figura 14: Traço das cadeias dos parâmetros do Modelo Completo.

Figura 15: Traço das cadeias dos parâmetros do Modelo 2.

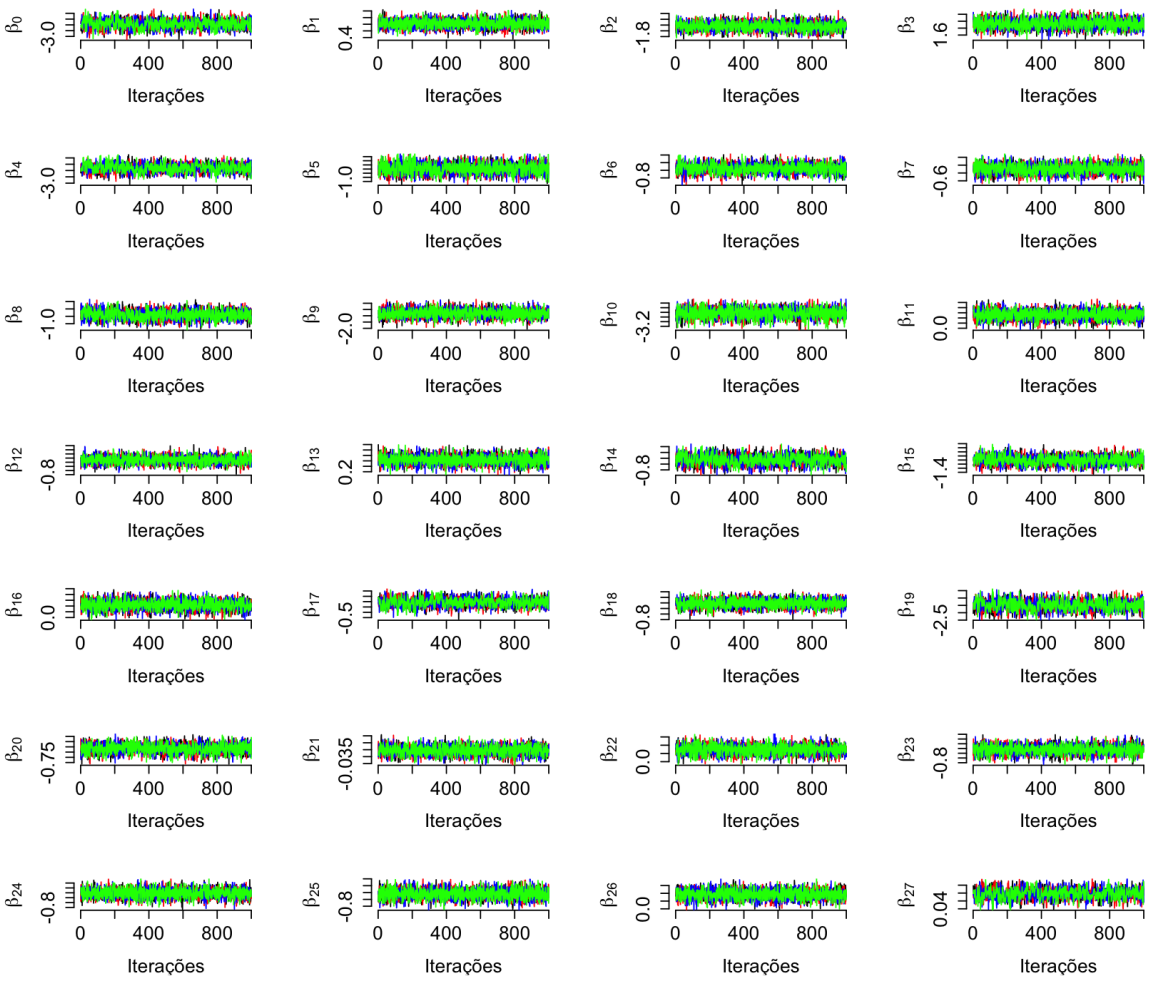

Figura 16: Traço das cadeias dos parâmetros do Modelo 3.

Calculou-se como medida de qualidade do ajuste o DIC para cada modelo e esses resultados podem ser vistos na Tabela 3. Além disso, para cada modelo proposto foram realizados os seguintes passos: para cada iteração do HMC, utilizando a base de treino, foi calculada a AUC e encontrado o ponto de corte daquela iteração. Em seguida, foram estimadas as probabilidades $\hat{\pi}_i$ de cada iteração para as observações da base de teste e essas probabilidades estimadas foram classificadas, em 0 ou 1, através do ponto de corte da iteração correspondente encontrado anteriormente. A partir da classificação, calculou-se para cada iteração as medidas de Acurácia, Sensibilidade e Especificidade. Os resultados para as métricas calculadas para cada iteração de cada modelo estão resumidos na Figura 17 e as médias a *posteriori* de cada medida para cada modelo proposto estão na Tabela 3. O melhor valor para cada métrica está destacado em vermelho.

Tabela 3: DIC e média a *posteriori* das medidas de avaliação para os modelos propostos.

Modelo	DIC	AUC	Acurácia	Sensibilidade	Especificidade
Completo	3833,1	0,9057	0,8060	0,8073	0,8022
2	3826,8	0,9049	0,8075	0,8103	0,7998
3	3829,3	0,9048	0,8065	0,8084	0,8014

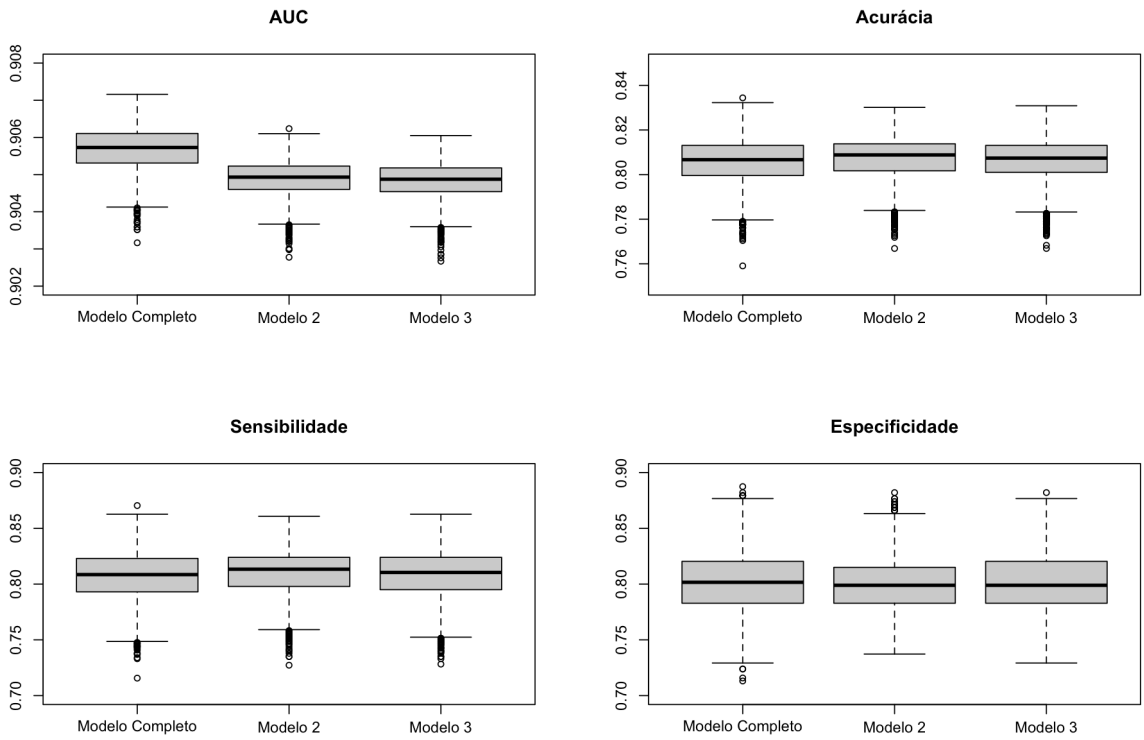

Figura 17: Bloxplots das medidas de avaliação para os modelos propostos.

Ao comparar os resultados obtidos na Tabela 3 e na Figura 17, observa-se que o Modelo 2 apresenta um valor de DIC ligeiramente menor em relação aos outros dois modelos, indicando uma melhor adequação aos dados observados e uma menor complexidade. Além disso, obteve média a *posteriori* da acurácia igual a 80,75%, e sensibilidade igual 81,03%, superando os outros dois modelos. Embora as médias a *posteriori* das métricas de sensibilidade e AUC para este modelo tenham sido ligeiramente inferiores aos melhores resultados observados, ainda foram bastante próximas e satisfatórias, com valores de 79,98% e 0,9049, respectivamente. Portanto, considerando o conjunto geral de métricas

avaliadas, o Modelo 2 apresentou o desempenho mais equilibrado e, portanto, é o mais indicado para as previsões e interpretação dos dados.

Os resultados para as estimativas pontuais, intervalos de credibilidade de 95% e razão de chances para o Modelo 2 estão representados na Tabela 5 e para os Modelos 1 e 3, encontram-se no Apêndice 1.

Tabela 5: Resultados para o Modelo 2.

Variável	Estimativa	IC 95%	OR
Intercepto	-2,57	[-3,86; -1,52]	-
Idoso (X_1)	0,82	[0,57; 1,06]	2,2676
Dependentes (X_2)	-1,43	[-1,72; -1,15]	0,2391
Indicação (X_3)	1,92	[1,66; 2,19]	6,8401
Telefone (X_4)	-2,24	[-3,23; -1,41]	0,1061
Linhas Múltiplas (X_5)	-0,25	[-0,56; 0,01]	0,7758
Segurança Online (X_6)	-0,81	[-1,12; -0,53]	0,4455
Backup Online (X_7)	-0,67	[-0,97; -0,40]	0,5099
Proteção de Aparelho (X_8)	-0,42	[-0,72; -0,14]	0,658
Suporte Premium (X_9)	-0,89	[-1,19; -0,60]	0,4121
Contrato - Anual (X_{10})	-1,53	[-1,80; -1,26]	0,2175
Contrato - Bianual (X_{11})	-2,63	[-3,04; -2,25]	0,0722
Conta Digital (X_{12})	0,26	[0,09; 0,44]	1,2968
Método de Pagamento - Crédito (X_{13})	-0,44	[-0,63; -0,26]	0,6413
Método de Pagamento - Cheque (X_{14})	0,64	[0,30; 0,98]	1,89
Streaming TV (X_{15})	-0,69	[-1,20; -0,26]	0,5021
Streaming Filme (X_{16})	-1,04	[-1,61; -0,53]	0,3548
Streaming Música (X_{17})	0,43	[0,10; 0,74]	1,5354
Tipo de Internet - DSL (X_{18})	1,56	[0,51; 2,82]	4,7636
Tipo de Internet - Fibra (X_{19})	-0,20	[-0,49; 0,10]	0,8223
Tipo de Internet - Nenhum (X_{20})	-2,02	[-3,22; -1,10]	0,1327
Número de Indicações (X_{21})	-0,61	[-0,70; -0,53]	0,5423
Fidelidade (X_{22})	-0,02	[-0,03; -0,02]	0,9759
Oferta - A (X_{23})	0,74	[0,17; 1,30]	2,1062
Oferta - B (X_{24})	-0,25	[-0,62; 0,12]	0,7803
Oferta - C (X_{25})	-0,20	[-0,56; 0,16]	0,8182

Oferta - D (X_{26})	-0,42	[-0,72; -0,14]	0,6559
Oferta - E (X_{27})	0,38	[0,14; 0,61]	1,4604
Cobrança Mensal (X_{28})	0,10	[0,06; 0,15]	1,1078

Dessa forma, o modelo estimado escolhido pode ser escrito como:

$$\begin{aligned}\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) &= -2,57 + 0,82X_{i,1} - 1,43X_{i,2} + 1,92X_{i,3} - 2,24X_{i,4} - 0,25X_{i,5} \\ &\quad -0,81X_{i,6} - 0,67X_{i,7} - 0,42X_{i,8} - 0,89X_{i,9} - 1,53X_{i,10} \\ &\quad -2,63X_{i,11} + 0,26X_{i,12} - 0,44X_{i,13} + 0,64X_{i,14} - 0,69X_{i,15} \\ &\quad -1,04X_{i,16} + 0,43X_{i,17} + 1,56X_{i,18} - 0,20X_{i,19} - 2,02X_{i,20} \\ &\quad -0,61X_{i,21} - 0,02X_{i,22} + 0,74X_{i,23} - 0,25X_{i,24} - 0,20X_{i,25} \\ &\quad -0,42X_{i,26} + 0,38X_{i,27} + 0,10X_{i,28}\end{aligned}$$

Por meio do cálculo das razões de chance, é possível analisar as relações entre as variáveis explicativas e a variável resposta. Observou-se que as chances de *churn* aumentam em 126,76% quando o cliente é idoso e mais de 5 vezes quando o cliente indica alguém para a empresa. No entanto, a cada indicação feita, as chances de *churn* reduzem em 45,77%. Além disso, caso o cliente possua dependentes, a chance de *churn* é reduzida em aproximadamente 76%.

Quanto ao tipo de contrato, optar pelo plano anual reduz a chance de *churn* em 78,25% em comparação ao plano mensal. Para os clientes que escolhem o plano bianual, a chance de *churn* é reduzida em 92,78% em relação ao plano mensal. No caso de receber a fatura digitalmente, as chances de *churn* aumentam em 29,68%.

Os clientes que pagam as faturas com cartão de crédito têm uma redução de 35,87% nas chances de *churn* em comparação aos clientes que optam pelo débito automático. Por outro lado, aqueles que pagam em cheque têm um aumento de 89% nas chances de *churn* em comparação aos que pagam via débito automático. A cada dólar de aumento na mensalidade paga pelo cliente, as chances de *churn* aumentam em 10,78%, enquanto a cada mês adicional de permanência do cliente, as chances de *churn* são reduzidas em 2,41%.

A contratação de serviços como telefone, linhas múltiplas, segurança online, backup, proteção de aparelho e suporte premium reduz as chances de *churn*. Clientes que possuem o tipo de internet DSL têm as chances de *churn* aumentadas em quase 4 vezes, em

comparação com aqueles que possuem internet a cabo. Por outro lado, possuir internet de fibra óptica ou não ter o serviço de internet contratado reduz as chances de *churn* em 17,77% e 86,73%, respectivamente, quando comparado com possuir internet a cabo.

A última oferta de marketing aceita pelo cliente também influencia nas chances de *churn*. Caso seja a Oferta A ou Oferta E, as chances de *churn* aumentam em comparação com nenhuma oferta aceita. Por outro lado, se for alguma das ofertas B, C ou D, as chances de *churn* diminuem em comparação com nenhuma oferta aceita. Os clientes que utilizam o pacote de internet para assistir filmes têm suas chances de *churn* diminuídas em aproximadamente 65%, enquanto aqueles que o utilizam para assistir TV têm suas chances de *churn* diminuídas em cerca de 50%. Já o uso do pacote de internet para ouvir música aumenta as chances em aproximadamente 54%.

Em geral, muitas das percepções vistas na análise descritiva foram confirmadas pelo ajuste dos modelos, já que variáveis que inicialmente pareciam ser influentes no fenômeno do *churn*, como ser idoso, casado, possuir dependentes, mensalidade atual paga e última oferta de marketing aceita, demonstraram aumentos ou reduções nas chances de *churn*. Outras variáveis, como sexo, ser jovem, região de residência, tarifa de longa distância e downloads, que não pareciam ter influência significativa, de acordo com a análise descritiva, se mantiveram assim após o ajuste dos modelos.

Por meio das relações identificadas através da razão de chances, é possível identificar oportunidades de melhoria para reduzir as chances de *churn*. No caso dos clientes idosos, cujas chances de *churn* aumentam significativamente, é importante desenvolver estratégias de retenção específicas, como suporte técnico especializado, atendimento ao cliente personalizado e ofertas exclusivas adaptadas às suas preferências.

Embora a indicação por parte dos clientes possa aumentar as chances de *churn*, é fundamental reconhecer a importância das indicações para o crescimento do negócio. Nesse sentido, a empresa pode implementar um programa de indicações mais eficaz, oferecendo incentivos aos clientes que indicarem novos clientes e fornecendo benefícios adicionais.

Outra sugestão é aprimorar a experiência dos clientes com dependentes, oferecendo pacotes familiares e recursos adicionais que atendam às necessidades específicas das famílias. Além disso, a empresa pode oferecer promoções atrativas para os planos de longo prazo, como descontos e benefícios exclusivos, considerando que esses planos têm um impacto positivo na redução do *churn*.

Diversificar os serviços oferecidos pode ser uma estratégia eficaz, levando em consideração que a contratação de serviços adicionais reduz as chances de *churn*. Por fim, é essencial investir na melhoria dos serviços de internet, expandindo a cobertura de internet a cabo e fibra óptica, para fornecer uma conexão estável e de alta velocidade aos clientes. Essas medidas podem ajudar a aumentar a satisfação dos clientes e reduzir as chances de *churn* na empresa.

4 Conclusões

Diante dos resultados obtidos nesta pesquisa, pode-se concluir que o objetivo proposto foi alcançado com sucesso. O estudo identificou e interpretou as variáveis que influenciam o fenômeno do *churn* em uma empresa de telecomunicações fictícia da Califórnia, apresentando um modelo preditivo baseado em regressão logística bayesiana. Através da análise dos parâmetros estimados, foi possível quantificar os efeitos dessas variáveis no *churn* e compreender suas relações.

As descobertas revelaram importantes percepções sobre o comportamento dos clientes e suas interações com a empresa. Foi observado que características como ser idoso, ter feito indicação de novos clientes, ter a presença de dependentes no plano, o tipo de contrato, o método de pagamento, o valor mensal pago, o tempo de permanência, a contratação de serviços adicionais e o tipo de internet têm influência significativa nas chances de *churn*. Esses resultados fornecem um panorama abrangente e detalhado sobre os fatores que impactam a rotatividade de clientes na empresa analisada.

Além disso, a análise realizada dos modelos propostos revelou que o Modelo 2 demonstrou um melhor ajuste aos dados observados e um desempenho mais equilibrado em relação às métricas avaliadas.

Com o conhecimento adquirido sobre os principais fatores que influenciam o *churn*, a empresa pode direcionar seus esforços para aprimorar a experiência do cliente, personalizar ofertas e desenvolver políticas de retenção mais eficazes.

No entanto, é importante ressaltar que este estudo possui limitações, como a utilização de dados de uma empresa fictícia e a restrição geográfica à Califórnia. Portanto, recomenda-se que pesquisas futuras sejam realizadas com dados reais e em diferentes contextos, a fim de validar e expandir os resultados encontrados nesta investigação.

Referências

AU, W.-H.; CHAN, K.; YAO, X. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, v. 7, n. 6, p. 532–545, 2003.

BETANCOURT, M. *A Conceptual Introduction to Hamiltonian Monte Carlo*. [S.l.]: University of Warwick, 2018.

DOBSON, A.; BARNETT, A. *An Introduction to Generalized Linear Models*. CRC Press, 2018. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781351726221. Disponível em: ⟨https://books.google.com.br/books?id=YOFstgEACAAJ⟩.

FRANCESCHI, P. R. de. *Modelagens Preditivas de Churn: o caso do Banco do Brasil*. Dissertação (Mestrado) — Programa de Pós-Graduação em Gestão e Negócios - Universidade do Vale do Rio dos Sinos, 2019.

GABRY, J.; GOODRICH, B.; VEHTARI, A. *rstanarm: Bayesian Applied Regression Modeling via Stan*. [S.l.], 2022. R package version 2.21.2. Disponível em: ⟨https://mc-stan.org/rstanarm/⟩.

GAMERMAN, D.; LOPES, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd. ed. [S.l.]: Chapman and Hall/CRC, 2006.

GELMAN, A.; GILKS, W. R.; ROBERTS, G. O. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, Institute of Mathematical Statistics, v. 7, n. 1, p. 110 – 120, 1997. Disponível em: ⟨https://doi.org/10.1214/aoap/1034625254⟩.

HUANG, J.; LING, C. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, v. 17, n. 3, p. 299–310, 2005.

JUNIOR, A. C. da S. *Classificação de Churn Utilizando um Modelo de Regressão Logística*. Monografia (Especialização) — Programa de Especialização em Data Science e Big Data - Universidade Federal do Paraná, 2020.

KUHN, M. *caret: Classification and Regression Training*. [S.l.], 2020. R package version 6.0-86. Disponível em: ⟨https://CRAN.R-project.org/package=caret⟩.

KUMAR, P. et al. A benchmark to select data mining based classification algorithms for business intelligence and decision support systems. *International Journal of Data Mining & Knowledge Management Process*, v. 2, 10 2012.

KURTZ, D. L.; CLOW, K. E. *Services Marketing*. 1. ed. [S.l.]: Wiley, 1997.

NEAL, R. M. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. Disponível em: ⟨https://doi.org/10.1201%2Fb10905⟩.

PAIVA, P. dos S. *Aplicação de algorimto de clusterização na análise de churn: estudo de caso no setor de telecomunicações*. Monografia (Graduação) — Universidade Federal Fluminense, 2022.

PRATI, R. C.; BATISTA, G. E. d. A. P. A.; MONARD, M. C. Curvas roc para avaliação de classificadores. *IEEE Latin America Transactions*, 2008.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: ⟨http://www.R-project.org/⟩.

REICHHELD, F.; JR, W. S. Zero defections: quality comes to services. *Harvard Business Review*, v. 68, n. 5, p. 105–111, 1990.

SPIEGELHALTER, D. J. et al. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 64, n. 4, p. 583–639, 10 2002. ISSN 1369-7412. Disponível em: ⟨https://doi.org/10.1111/1467-9868.00353⟩.

VELOSO, F. J. M. *Um modelo para previsão de churn na área do retalho*. Dissertação (Mestrado) — Universidade do Minho, 2013.

WU, S. et al. Integrated churn prediction and customer segmentation framework for telco business. *IEEE Access*, v. 9, p. 62118–62136, 2021.

APÊNDICE 1

Neste apêndice encontram-se as tabelas com os resultados para as estimativas pontuais e intervalos de credibilidade de 95% para os Modelos 1 e 3.

Tabela 6: Resultados para o Modelo 1.

Variável	Estimativa	IC 95%
Intercepto	-1,486	[-10,0506; 7,1793]
Sexo (X_1)	-0,0817	[-0,2474; 0,0856]
Jovem (X_2)	-0,2788	[-0,6382; 0,0881]
Idoso (X_3)	0,7615	[0,4029; 1,1134]
Casado (X_4)	0,4953	[-0,0253; 1,0248]
Dependentes (X_5)	-1,7632	[-2,3683; -1,1611]
Indicação (X_6)	1,472	[0,9214; 2,0278]
Telefone (X_7)	-2,3388	[-3,2838; -1,4844]
Linhas Múltiplas (X_8)	-0,2971	[-0,6008; -0,022]
Internet (X_9)	-0,9699	[-9,6261; 7,5507]
Dados Ilimitados (X_{10})	0,279	[-0,0832; 0,6429]
Segurança Online (X_{11})	-0,8501	[-1,1526; -0,5654]
Backup Online (X_{12})	-0,7039	[-0,9989; -0,4356]
Proteção de Aparelho (X_{13})	-0,4518	[-0,7637; -0,1724]
Suporte Premium (X_{14})	-0,9181	[-1,2155; -0,6386]
Contrato - Anual (X_{15})	-1,542	[-1,8147; -1,291]
Contrato - Bianual (X_{16})	-2,6466	[-3,0687; -2,2596]
Conta Digital (X_{17})	0,2586	[0,0732; 0,4385]
Método de Pagamento - Crédito (X_{18})	-0,4632	[-0,6385; -0,2827]
Método de Pagamento - Cheque (X_{19})	0,607	[0,2678; 0,9538]
Streaming TV (X_{20})	-0,7324	[-1,2223; -0,3008]

Streaming Filme (X_{21})	-1,1692	[-1,72; -0,6517]
Streaming Música (X_{22})	0,5348	[0,1918; 0,8797]
Tipo de Internet - DSL (X_{23})	0,9125	[-7,67; 9,414]
Tipo de Internet - Fibra (X_{24})	-0,2201	[-0,526; 0,0923]
Tipo de Internet - Nenhum (X_{25})	-2,1629	[-3,2799; -1,1936]
Região - Central (X_{26})	-0,1077	[-0,3751; 0,1496]
Região - Norte (X_{27})	-0,2332	[-0,5334; 0,0769]
Região - Sul (X_{28})	-0,0707	[-0,2956; 0,1459]
Idade (X_{29})	0,0014	[-0,0083; 0,0112]
Número de Dependentes (X_{30})	0,1226	[-0,1389; 0,3808]
Número de Indicações (X_{31})	-0,6171	[-0,7062; -0,5371]
Fidelidade (X_{32})	-0,0436	[-0,0602; -0,0278]
Oferta - A (X_{33})	0,6486	[0,036; 1,2125]
Oferta - B (X_{34})	-0,2426	[-0,6036; 0,1181]
Oferta - C (X_{35})	-0,1443	[-0,5291; 0,2178]
Oferta - D (X_{36})	-0,4008	[-0,6841; -0,1245]
Oferta - E (X_{37})	0,3391	[0,0905; 0,5813]
Longa Distância (X_{38})	-0,0005	[-0,0086; 0,0078]
Downloads (X_{39})	0,0024	[-0,0043; 0,0088]
Cobrança Mensal (X_{40})	0,1017	[0,0634; 0,1472]
Cobrança Total (X_{41})	0,0002	[0; 0,0004]
Reembolsos (X_{42})	-0,0081	[-0,0188; 0,003]
Internet Extra (X_{43})	0,0031	[-0,0017; 0,008]
Longa Distância Total (X_{44})	0	[-0,0002; 0,0003]
CLTV (X_{45})	0	[-0,0001; 0,0001]

Tabela 7: Resultados para o Modelo 3.

Variável	Estimativa	IC 95%
Intercepto	-1,9088	[-2,6836; -1,1778]
Idoso (X_1)	0,8178	[0,575; 1,0643]
Dependentes (X_2)	-1,4388	[-1,7265; -1,1582]
Indicação (X_3)	1,9252	[1,6682; 2,199]
Telefone (X_4)	-1,7905	[-2,4514; -1,1284]

Segurança Online (X_5)	-0,6826	[-0,9275; -0,4292]
Backup Online (X_6)	-0,55	[-0,7889; -0,3141]
Proteção de Aparelho (X_7)	-0,2935	[-0,5376; -0,0578]
Suporte Premium (X_8)	-0,7595	[-1,0001; -0,5146]
Contrato - Anual (X_9)	-1,5214	[-1,7887; -1,2505]
Contrato - Bianual (X_{10})	-2,6321	[-3,0569; -2,2416]
Conta Digital (X_{11})	0,2586	[0,0817; 0,4346]
Método de Pagamento - Crédito (X_{12})	-0,444	[-0,6263; -0,2614]
Método de Pagamento - Cheque (X_{13})	0,6424	[0,3019; 0,9873]
Streaming TV (X_{14})	-0,4473	[-0,7975; -0,1225]
Streaming Filme (X_{15})	-0,7995	[-1,2321; -0,3788]
Streaming Música (X_{16})	0,4367	[0,1152; 0,7513]
Tipo de Internet - DSL (X_{17})	0,9341	[0,1466; 1,7551]
Tipo de Internet - Fibra (X_{18})	-0,2214	[-0,5123; 0,0719]
Tipo de Internet - Nenhum (X_{19})	-1,4785	[-2,1602; -0,8042]
Número de Indicações (X_{20})	-0,6142	[-0,7053; -0,5316]
Fidelidade (X_{21})	-0,0255	[-0,0322; -0,019]
Oferta - A (X_{22})	0,7473	[0,1708; 1,2906]
Oferta - B (X_{23})	-0,2449	[-0,6315; 0,1308]
Oferta - C (X_{24})	-0,2177	[-0,5816; 0,1434]
Oferta - D (X_{25})	-0,4393	[-0,7163; -0,1523]
Oferta - E (X_{26})	0,3713	[0,13; 0,6084]
Cobrança Mensal (X_{27})	0,0781	[0,0527; 0,1046]