Mariana Barros Ramos

Análise do Perfil dos Usuários Big Brother Brasil: Um estudo de caso dos assinantes do Globloplay

Niterói - RJ, Brasil

20 de julho de 2023

Universidade Federal Fluminense

Mariana Barros Ramos

Análise do Perfil dos Usuários Big Brother Brasil: Um estudo de caso dos assinantes do Globloplay

Trabalho de Conclusão de Curso

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

Orientador(a): Prof[a]. Jessica Quintanilha Kubrusly

Niterói - RJ, Brasil

20 de julho de 2023

Universidade Federal Fluminense

Mariana Barros Ramos

Análise do Perfil dos Usuários Big Brother Brasil: Um estudo de caso dos assinantes do Globloplay

Monografia de Projeto Final de Graduação sob o título *"Análise do Perfil dos Usuários Big Brother Brasil: Um estudo de caso dos assinantes do Globloplay "*, defendida por Mariana Barros Ramos e aprovada em 20 de julho de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Prof. Me. Jony Arrais Pinto Junior
Departamento de Estatística – UFF

Profa. Ma. Ludmilla Da Silva Viana Jacobson
Departamento de Estatística – UFF

Niterói, 20 de julho de 2023

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

R175a Ramos, Mariana Barros
Análise do Perfil dos Usuários Big Brother Brasil: Um estudo de caso dos assinantes do Globloplay / Mariana Barros Ramos. - 2023.
37 f.: il.

Orientador: Jessica Quintanilha Kubrusly.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Modelo Logistico. 2. Analise de Perfil. 3. BBB. 4. Streaming. 5. Produção intelectual. I. Kubrusly, Jessica Quintanilha, orientadora. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O presente estudo teve como objetivo analisar dados fornecidos pela Globo para identificar o perfil do usuário do Big Brother Brasil (BBB). Para isso, foi utilizado um modelo logístico ajustado no software RStudio. O foco da pesquisa foi entender o comportamento dos assinantes do Globoplay que têm o BBB como seu primeiro conteúdo.

Após análise dos dados, o perfil ideal do usuário do BBB foi identificado. Esse perfil é composto, em média, por pessoas de 36 anos, predominantemente do sexo feminino, e provenientes da região Sudeste do Brasil. Os usuários com esse perfil tendem a optar pelo plano Globoplay + Canais Ao Vivo e Premiere, consumindo séries pelo celular e filmes na TV. Além disso, demonstram um interesse significativo por programas do tipo reality, preferindo conteúdos ao vivo em vez de conteúdos on demand. É interessante notar que esses usuários não costumam assistir ao canal da Globo pelo serviço de streaming.

Os resultados obtidos fornecem informações valiosas para explorar novas formas de interação e engajamento com o público-alvo, possibilitando a criação de experiências que atendam às suas expectativas. Compreender o comportamento desses usuários permitirá o desenvolvimento de estratégias mais eficientes e direcionadas, a fim de maximizar a satisfação e o envolvimento com o conteúdo oferecido pelo Globoplay, especialmente no contexto do BBB.

Palavras-chave: Modelo Logístico. Análise de perfil. BBB. Streaming.

Dedicatória

Dedico este trabalho à minha família, que sempre me apoiou e incentivou durante toda essa jornada.

Agradecimentos

Gostaria de expressar meus sinceros agradecimentos a todas as pessoas que contribuíram para a realização deste trabalho. Em especial, agradeço à minha orientadora Jessica Kubrusly, pela orientação, paciência e valiosas contribuições ao longo deste processo. Aos membros da banca, Jony Arrais e Ludmilla Jacobson, agradeço não só pela disponibilidade como também pelos ensinamentos compartilhados durante meus anos acadêmicos. Também sou imensamente grata aos demais professores, cujos ensinamentos e conhecimentos foram fundamentais para o meu desenvolvimento ao longo do curso.

Agradeço de coração à minha família, especialmente à minha mãe, Alanquesia Barros, minha irmã, Fabiana Barros, e minhas avós, Maria do Carmo Ramos e Maria das Dores Ramos, pelo apoio incondicional e encorajamento não apenas durante esse período, mas ao longo de toda a minha vida. Sem vocês, eu não seria quem sou hoje.

Quero expressar minha gratidão ao meu namorado e companheiro de vida, Rodrigo Menchio, que tem sido meu maior apoio e fonte de inspiração ao longo desta jornada. Obrigada por todo o carinho, incentivo, suporte emocional e por estar sempre disposto a me ajudar. Sou imensamente grata por ter você ao meu lado, com certeza essa jornada não teria sido a mesma sem você!

Meus agradecimentos também se estendem aos meus amigos, em especial a Luiza Carolina Brasil, por estar ao meu lado há tantos anos, sempre me ajudando e apoiando nos mais diversos momentos. Agradeço também a Ana Beatriz Santaballa, Lucas Mattos, Iasmyn Lugon, Danielle Ribeiro, Thiago Silva, Desiree Melo, Vander Direito e Laila Mury, por tornarem os anos de faculdade ainda mais especiais na companhia deles. E também às minhas queridas amigas, Bárbara Zanobini, Renata Pinheiro e Susana Alves, por todo apoio e tantos momentos maravilhosos que compartilhamos juntas.

Muito obrigada!

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A era é dos *streamings* e o Globoplay, *stream* brasileiro do Grupo Globo, tornou-se líder nacional nesse meio. Além de ter acesso a um catálogo diverso de filmes, séries, novelas, entre outros para assistir *on demand*, também é possível assistir programas ao vivo, como por exemplo o famoso *reality show* Big Brother Brasil (BBB).

O programa é produzido e transmitido pela TV Globo, com exibições diárias na própria emissora, além do canal pago Multishow e, como já mencionado anteriormente, também no Globoplay para os assinantes do *stream*, onde é possível acompanhar a rotina dos participantes 24h por dia.

Sabe-se que o *reality show* é um sucesso de audiência e prova disso foi o recorde de 'maior quantidade de votos do público conseguidos por um programa de televisão' que entrou para o *Guinness World Records*(LIMITED, 2020), marca antes pertencente ao show de talentos *American Idol*. O impacto de tamanha audiência evidencia a importância de conhecer os gostos e preferências dos assinantes para atender às suas necessidades.

Coletar, organizar e interpretar dados pode gerar informações valiosas que auxiliam no processo de personalização de um serviço. Eke et al. (2019) selecionaram e compararam diferentes artigos sobre análises de perfil de usuário. Neste estudo são comparados os diferentes processos de modelagem, pré-processamento, coleta de recursos e abordagem de desempenho. O resultado mostra que um processo de modelagem eficaz pode potencializar a construção de um perfil mais eficiente para a personalização de um serviço.

Como já comentado, a crescente dos *streams* têm se mostrado forte e constante, e assistir a vídeos *online* tem se tornado um dos passatempos mais populares entre as pessoas, por isso é necessário um servidor de anúncios capaz de direcionar propagandas mais assertivas para cada perfil de usuário. Visto isso, o artigo de Nananukul (2013) teve como objetivo criar um modelo para encaminhar propagandas mais assertivas, através de Regressão Logística, usando dados de visualização do usuário. Para o experimento, foi selecionada uma amostra contendo apenas homens entre 45-54 anos, onde 80% desta

amostra foi utilizada como treinamento do modelo. No experimento foi feita a seleção das variáveis independentes (covariáveis) de acordo com a sua significância estatística. O que se mostrou melhor foi o uso do gênero de vídeo em combinação com os níveis de uso do usuário, aumentando em 26% a acurácia do modelo.

No estudo realizado por RAGHURAM; AKSHAY; CHANDRASEKARAN, a Regressão Logística desempenha um papel importante na categorização de usuários do Twitter com base em três recursos fundamentais: *tweets*, usuários e séries temporais. Essa abordagem demonstra a utilidade da Regressão Logística como um método de aprendizado de máquina para analisar e classificar dados relacionados aos usuários e suas atividades na plataforma. São seis categorias de interesse: política, entretenimento, negócios, jornalismo, ciência e tecnologia e cuidados da saúde. Foram utilizados os seguintes métodos para comparar o conjunto de características proposto: *Support Vector Machines, Naive Bayes, k-Nearest Neighbours*, Árvores de Decisão e Regressão Logística. O método com melhor acurácia foi o *Support Vector Machines*, que apresentou uma acurácia de 89,92%. A Regressão Logística obteve uma acurácia de 83,98% e acima de 87% após o PCA, sendo a segunda maior dentre os testes realizados.

Considerando o grande interesse pelo Globoplay para assistir ao *reality show* e a competitividade aumentando, é essencial oferecer um serviço de qualidade, que conheça os gostos e preferências dos assinantes, a fim de direcionar conteúdos cada vez mais atrativos e manter os usuários engajados com o produto.

Portanto, o objetivo geral deste presente trabalho é realizar uma análise de perfil dos usuários BBB, através de um modelo logístico. Entende-se como usuário BBB aquele que assinou o Globoplay e seu primeiro consumo foi o *reality*. Destaca-se como objetivos específicos estudar o modelo logístico e realizar uma aplicação em dados reais.

O presente trabalho está organizado em três capítulos principais. O Capítulo 2 discutirá os materiais e métodos utilizados nesta pesquisa, fornecendo uma base sólida para a compreensão do estudo. O Capítulo 3 apresentará o desenvolvimento do trabalho e os resultados obtidos, trazendo uma análise das informações coletadas. Por fim, no Capítulo 4, serão abordadas as conclusões alcançadas a partir desta investigação.

2 Materiais e Métodos

Neste capítulo, será apresentado a base de dados e as variáveis que serão investigadas para identificar o perfil dos usuários BBB. Além disso, serão abordados os principais conceitos envolvidos na utilização de Modelos de Regressão Logística. Este capítulo é fundamental para a fundamentação e compreensão dos resultados que serão discutidos posteriormente, permitindo uma análise detalhada e embasada das características dos usuários envolvidos.

2.1 Materiais

Este estudo utiliza dados extraídos da Globo durante o período de janeiro a abril de 2022. Por motivos de sigilo, o número exato de observações não será divulgado, no entanto, será considerado um valor hipotético de linhas igual a 100.000 e 107 colunas.

O objetivo geral deste trabalho é compreender o perfil dos usuários que consomem o programa BBB. Um usuário BBB é definido como alguém que assinou o Globoplay e cujo primeiro consumo foi o *reality show*. As covariáveis consideradas envolvem informações demográficas, como gênero, idade e estado de residência dos usuários. Além disso, foram analisadas informações sobre o tipo de assinatura, bem como os padrões de consumo, incluindo o tempo gasto em novelas, séries e canais em diferentes dispositivos, como celular, TV e *desktop*, entre outras categorias de consumo.

2.2 O Modelo Linear Normal

O Modelo de Regressão Linear busca estimar o valor médio de uma variável Y a partir da suposição

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{j-1} X_{i,j-1} + \varepsilon_i \tag{2.1}$$

onde Y_i é o valor observado da variável dependente (ou variável resposta) referente a i-ésima observação da amostra; X_i é o valor observado da variável independente (ou variável explicativa) referente a i-ésima observação da amostra; $\beta_0, \beta_1, ..., \beta_{p-i}$ são os parâmetros desconhecidos a serem estimados; ε_i é o erro aleatório do modelo tal que $E(\varepsilon_i) = 0$, VAR$(\varepsilon_i) = \sigma^2$ e $COV(\varepsilon_i, \varepsilon_j) = 0$ para $i = 1, 2, ..., n$,$j = 1, 2, ..., n$ e $i \neq j$.

O modelo mais simples de regressão linear é a Regressão Linear Simples, que envolve apenas uma variável resposta e uma variável independente. Quando são utilizadas mais de uma variável independente, é chamado de Regressão Linear Múltipla.

Esse modelo apresenta diversas vantagens para diferentes objetivos, como a compreensão da relação entre variáveis e, como mencionado anteriormente, a previsão do valor médio de uma variável com base no valor de outra variável.

Para que o modelo de Regressão Linear seja adequado, é necessário que a base de dados atenda a alguns requisitos, além dos mencionados anteriormente, como a presença de relação linear entre a variável resposta e as variáveis explicativas, a adequação à distribuição normal e a ausência de multicolinearidade, que ocorre quando as variáveis independentes possuem relações aproximadamente lineares entre si.

No escopo deste estudo, a variável resposta é a variável BBB, que indica se o usuário assinou o Globoplay e teve seu primeiro consumo relacionado ao *reality* (sucesso) ou não (fracasso), ou seja, assume valores binários, 0 para fracasso e 1 para sucesso. As covariáveis são as demais variáveis, como idade, gênero, estado, quantidade de horas consumidas de um determinado tipo de conteúdo, entre outras.

Como a variável resposta assume valores binários, ou seja, não pressupõe uma relação linear entre as variáveis, o modelo linear normal não é o mais adequado para o problema em questão. Portanto, é necessário adotar uma abordagem mais adequada. A seguir, serão exploradas outras metodologias que atendem a esses requisitos.

2.3 O Modelo Linear Generalizado

Os Modelos Lineares Generalizados (MLGs) foram propostos por Nelder e Wedderburn (1972) como uma ampliação dos modelos lineares normais. Sua utilização é indicada quando a variável resposta não apresenta uma distribuição Normal, isto é, quando Y assume valores na forma de contagem, categóricos, contínuos simétricos e assimétricos ou valores binários, 0 e 1, que é o caso da variável resposta deste trabalho.

A equação do modelo linear generalizado, que supõe que a média da variável resposta pode ser escrita como uma função de uma transformação linear das variáveis independentes é definida por Neter et al. (1996) da seguinte forma:

$$E(Y_i) = g(\beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{j-1} X_{i,j-1}) \tag{2.2}$$

e g é chamada de função de ligação.

As principais distribuições da classe MLGs são a Regressão de Poisson, usada para modelar dados de contagem, a Regressão Gama, utilizada para modelar variáveis contínuas assimétricas e a Regressão Logística, aplicada na modelagem de variáveis binárias, isto é, quando ocorre ou não um evento. Os problemas resolvidos a partir de um modelo de regressão logística são aqueles onde a variável resposta $Y_i \sim Bernoulli(p_i)$, onde p_i é a probabilidade de sucesso para a observação i. O objetivo em geral é encontrar um expressão que descreva $E(Y_i) = p_i$ como função das variáveis dependentes $X_{i,1}$,$X_{i,2}$,...,$X_{i,j-1}$ de acordo com a Equação 2.2

O MLG apresenta algumas exigências que devem ser atendidos para garantir a validade dos resultados obtidos como a linearidade, independência, homogeneidade das variâncias, ausência de multicolinearidade, além disso a exigência de pertencer a uma família exponencial e a escolha da função de ligação estão diretamente relacionadas à especificação correta do MLG, permitindo que as estimativas dos parâmetros sejam válidas e interpretáveis.

2.4 O Modelo Logístico

O modelo de Regressão Logística, que faz parte da classe dos Modelos Lineares Generalizados (MLGs), é uma ferramenta utilizada para analisar variáveis de resposta binária com base em um conjunto de variáveis explicativas. O Modelo Logístico foi selecionado como a abordagem adequada para este trabalho, uma vez que permite investigar a relação entre as variáveis explicativas e a variável resposta que assume valores binários. A escolha desse modelo se deve ao fato de que a variável resposta do estudo assume valores binários: 1 quando o usuário assina o Globoplay e tem seu primeiro consumo relacionado ao *Big Brother*, e 0 caso contrário. Essa seleção se mostra relevante para compreender o perfil dos usuários que consomem o programa BBB e direcionar estratégias de conteúdo mais efetivas.

A partir desse modelo é possível: prever a probabilidade de ocorrência de um evento

dado uma observação aleatória; predizer o efeito do conjunto de variáveis sobre a variável resposta, ou seja, concluir que a inclusão da variável independente X_j no modelo melhora significativamente a capacidade de explicação ou predição do modelo, enquanto a variável X_{j-1} não coopera muito; classificar observações quanto à propensão de ocorrência do evento ou não, estimando a probabilidade de uma dada observação estar em uma determinada categoria.

2.4.1 A Família Exponencial Unidimensional

A aplicação da regressão linear tem como objetivo neste trabalho de investigar a relação entre as variáveis explicativas e a variável resposta BBB. Para isso, é importante garantir que a distribuição da variável resposta seja adequada para a aplicação de técnicas estatísticas. Quando falamos em regressão linear, é preciso garantir algumas propriedades para conseguir estimar parâmetros, fazer testes de hipótese e tirar conclusões sobre os modelos. Portanto, a variável resposta não pode ter qualquer distribuição, mas sim pertença à classe da família exponencial, que garante essas propriedades.

Segundo Neter et al. (1996), uma variável aleatória Y tem distribuição pertencente à família exponencial unidimensional, isto é, cuja distribuição depende de um único parâmetro θ, se sua função de densidade ou de probabilidade puder ser expressa na forma:

$$f_y(y) = e^{c(\theta)T(y)}e^{d(\theta)}e^{S(y)}, y \in A \tag{2.3}$$

onde c e d são funções reais de θ, T e S são funções reais de Y e A=$lm(Y)$ não dependente de θ.

Diversas distribuições conhecidas são pertencentes à família exponencial unidimensional como por exemplo a Binomial, Poisson, Bernoulli, Exponencial, Geométrica, dentre outras. Abaixo, verifica-se que a distribuição de Bernoulli pertence à classe da família exponencial:

$$\mathrm{P}_y(y) = e^{ln(p^y)}e^{ln((1-p)^{1-y})}, y = 0, 1$$

$$\mathrm{P}_y(y) = e^{y \cdot ln(p)} e^{(1-y) \cdot ln(1-p)}, y = 0, 1$$

$$\mathrm{P}_y(y) = e^{y \cdot ln(p) - ln(1-p) - y \cdot ln(1-p)}, y = 0, 1$$

$$\mathrm{P}_y(y) = e^{y(ln(p) - ln(1-p))} e^{-ln(1-p)} e^{0}, y = 0, 1$$

onde $C(p) = ln(p) - ln(1-p)$, $T(y) = y$, $d(p) = -ln(1-p)$, $S(y) = 0$ e $A = 0, 1$.

2.4.2 Função de Ligação

Ainda de acordo com Neter et al. (1996), existem funções de ligação específicas que junto com funções de distribuições linearizam a curva logística. Para problemas deste tipo, é adequado utilizar a função de ligação logística

$$g(w) = \frac{e^w}{1 + e^w} = \frac{1}{1 + e^{-w}}$$

Aplicando esta função de ligação na Equação 2.2 chega-se na expressão

$$p_i = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{j-1} X_{i,j-1})} \tag{2.4}$$

2.4.3 Estimação dos Parâmetros

A distribuição de y_i no modelo de regressão linear permite encontrar os estimadores de máxima verossimilhança dos parâmetros. Para isso, é necessário determinar a função de log-verossimilhança em termos dos parâmetros do modelo, conforme abordado por Neter et al. (1996):

$$\log_e L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\boldsymbol{\beta' X_i}) - \sum_{i=1}^{n} \log_e(1 + exp(\boldsymbol{\beta' X_i})) \tag{2.5}$$

Não existe uma solução analítica para os valores $\beta_0, \beta_1, ..., \beta_{j-1}$ que maximizam a função de verossimilhança, métodos numéricos, como o método de Newton-Raphson, são necessários para encontrar as estimativas. Quando essas são encontradas, substitui-se esses valores para encontrar os valores ajustados para realizar previsões.

2.4.4 Métricas de Qualidade do Ajuste

A avaliação da qualidade do ajuste de um modelo é essencial para compreender o quão bem ele se adequa aos dados observados. Nesta seção, serão apresentadas duas métricas amplamente utilizadas: o Coeficiente de Determinação Ajustado e os Critérios da Informação de Akaike.

Coeficiente de Determinação Ajustado:

Sabe-se que o coeficiente de determinação, mais conhecido como R^2, é uma medida de ajuste do modelo que expressa quanto da variância dos dados é explicada pelo modelo e é apresentada por Neter et al. (1996) como:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.6}$$

onde y_i é o valor observado, $\bar{y}$ é a média das observações e $\hat{y}$ é o valor estimado de y_i.

Assim, quanto maior este coeficiente, mais explicativo é o modelo. O problema desta métrica de qualidade é que sempre que uma variável é acrescentada ao modelo, o R^2 aumenta. Então, para corrigir essa superestimação, utiliza-se o Coeficiente de determinação Ajustado, o R^2_α, definido por:

$$R^2_\alpha = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{2.7}$$

onde k é o número de parâmetros.

Este só aumenta quando as variáveis independentes são significativas, desta forma pode ser um critério de comparação de modelos: quanto maior o R^2_α, melhor é o modelo.

Critérios da Informação de Akaike

O critério de informação é uma medida corretiva pelo acréscimo de variáveis $X's$ ao modelo. A ideia é pegar o resultado ótimo da função de otimização e acrescentar penalidade, que são função do número de parâmetros k e do número de observações.

O critério de informação desenvolvido por Akaike et al. (1998) é definido por:

$$AIC = -2\sum ln(\hat{L}) + 2k \tag{2.8}$$

onde lnL é o log da verossimilhança e k o número de parâmetros do modelo.

Essa medida é considerada um critério de comparação de modelos, onde o melhor modelo é aquele que possuir o menor critério de informação (AIC).

2.4.5 Seleção das Variáveis

A escolha apropriada das variáveis explicativas desempenha um papel fundamental na construção de modelos estatísticos confiáveis e acurados. Nesta seção, vamos explorar duas abordagens muito utilizadas: a comparação de todos os modelos possíveis e os métodos de seleção passo-a-passo.

Comparação entre todos os possíveis modelos:

Neste método, são comparados todos os possíveis modelos a partir da comparação das medidas de bom ajustamento de cada modelo. O que torna esse método mais custoso é que se tivermos p variáveis dependentes, a quantidade de modelos possíveis será de 2^p, tornando inviável de se comparar, visto que o crescimento é exponencial. Ainda assim é um método muito utilizado para a seleção das variáveis.

Métodos de seleção passo-a-passo:

Nesse método, é possível optar por duas maneiras de fazer: com a inclusão progressiva ou pela exclusão progressiva das variáveis explicativas. A inclusão progressiva começa com nenhuma variável preditora no modelo e vai-se acrescentando, uma por vez, começando pelas mais explicativas, ou seja, aquelas com menor p-valor para o teste de Wald e este menor que alfa. Já na exclusão progressiva, tem-se todas as variáveis preditoras no modelo e vai-se excluindo, uma a uma, começando pelas menos explicativas, ou seja, aquelas com menor p-valor para o teste de Wald e este maior que alfa. Assim, segue-se o algoritmo até que todas as variáveis com p-valor maior que alfa não pertençam mais ao modelo.

2.4.6 Razão de Chance

A razão de chances (*odds ratio*) é um conceito fundamental no modelo de regressão logística.

Por definição, a chance de um evento acontecer é representada pela razão entre a probabilidade desse evento ocorrer (p) e a probabilidade de não ocorrer ($1-p$). Essa razão de chances é expressa como $p/(1-p)$. A transformação de probabilidade em chances é uma maneira de medir a relação entre a probabilidade de um evento ocorrer e a probabilidade de não ocorrer.

Considerando dois eventos, um com chance $p/(1-p)$ e outro com chance $q/(1-q)$, a razão entre essas duas chances é conhecida como odds ratio (OR). Essa medida foi descrita por Neter et al. (1996) da seguinte forma:

$$OR = \frac{p/(1-p)}{q/(1-q)} = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n} \tag{2.9}$$

Ao interpretar a razão de chances, considera-se diferentes cenários. Quando tem-se variáveis categóricas na análise de regressão logística, é definida uma categoria como referência para as demais. A categoria de referência é utilizada como ponto de comparação para interpretar as demais categorias. Se a razão de chances for maior do que 1, isso indica que a categoria tem uma maior probabilidade de ocorrência do evento em comparação com a categoria de referência. Por outro lado, se a razão de chances for menor do que 1, a categoria possui uma menor probabilidade de ocorrência do evento em relação à categoria de referência. Quando a razão de chances é igual a 1, indica que o evento tem a mesma probabilidade de ocorrer e de não ocorrer entre as categorias.

Na regressão logística, a razão de chances é utilizada para interpretar o efeito das variáveis preditoras no modelo. O valor $e^{\hat{\beta}_j}$, em que $\hat{\beta}_j$ representa o coeficiente estimado para a variável X_j, é interpretado como a razão de chances entre a variável X_j aumentando em 1 unidade. Isso nos permite entender como a probabilidade de ocorrência do evento muda em relação à alteração na variável preditora, considerando todas as outras variáveis constantes.

A relação $\frac{1}{e^{\hat{\beta}_j}}$ é útil para interpretar o efeito do coeficiente $\hat{\beta}_j$ quando é negativo. Nesse caso, $\frac{1}{e^{\hat{\beta}_j}}$ representa a razão entre a chance do evento não ocorrer (OR < 1) e a chance do evento ocorrer (OR > 1). É uma forma conveniente de visualizar como o aumento de uma unidade em X_j afeta a razão de chances do evento não ocorrer em relação ao evento ocorrer.

É importante ressaltar que o valor estimado de $\hat{\beta}_j$ é baseado em uma amostra dos dados e, portanto, está sujeito a incerteza. Para avaliar essa incerteza, calcula-se o intervalo de confiança para a razão de chances. O intervalo de confiança é uma faixa de valores em torno da estimativa de $\hat{\beta}_j$, dentro da qual a verdadeira razão de chances provavelmente está contida com um determinado grau de confiança.

O cálculo do intervalo de confiança leva em consideração o erro padrão de $\hat{\beta}_j$, que mede a dispersão da estimativa em relação à sua média. Quando utiliza-se um nível de confiança de 95%, significa que existe uma probabilidade de 95% de que o intervalo de

confiança contenha o valor real da razão de chances.

O intervalo de confiança é então dado por:

$$IC_{95\%} = e^{\hat{\beta}_j \pm z_\alpha \cdot \mathrm{SE}(\hat{\beta}_j)} \tag{2.10}$$

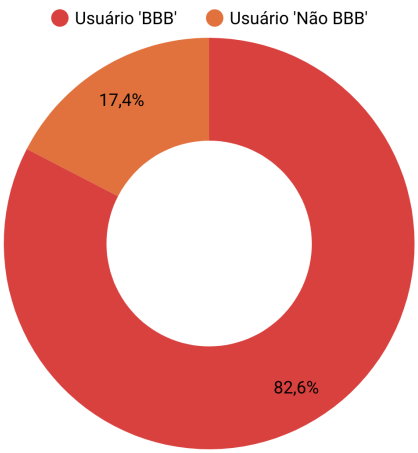
Onde $IC_{95\%}$ é o intervalo de confiança para a razão de chances, $\hat{\beta}_j$ é o coeficiente estimado, $\mathrm{SE}(\hat{\beta}_j)$ é o erro padrão do coeficiente e $z_{1-\alpha/2}$ é o quantil da distribuição normal tal que $P(Z \leq z_\alpha) = \alpha$ correspondente ao nível de confiança desejado.

3 Análise dos Resultados

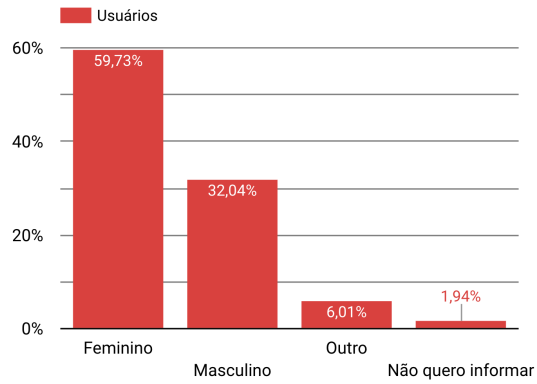
Neste capítulo, será apresentado a análise dos resultados obtidos a partir dos dados coletados para este trabalho.

Para a análise dos dados, utilizou-se o software RStudio(POSIT, 2023), no qual foram identificados alguns valores faltantes. A quantidade de dados faltantes era mínima, portanto, esses registros foram excluídos da base de dados.

Em seguida, foi realizada uma análise descritiva dos dados, que revelou que 82% da base é composta por usuários do tipo BBB. Além disso, constatou-se que a idade média dos usuários é de 36 anos e mais da metade é do sexo feminino.



(a) Distribuição de frequência por tipo de usuário



(b) Distribuição de frequência por gênero

Figura 1: Gráficos de Tipo de usuário e Gênero

Em relação às assinaturas, o plano mais popular é o Globoplay básico, que representa 70% das assinaturas.

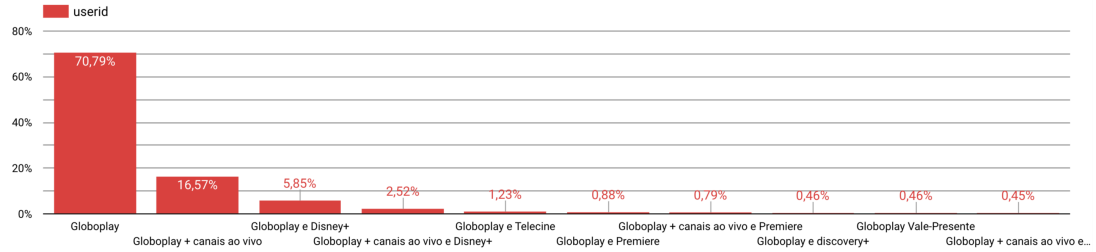
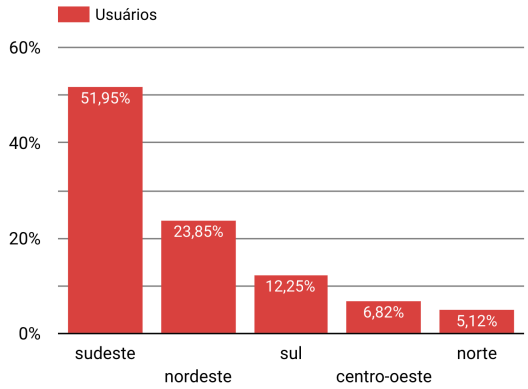
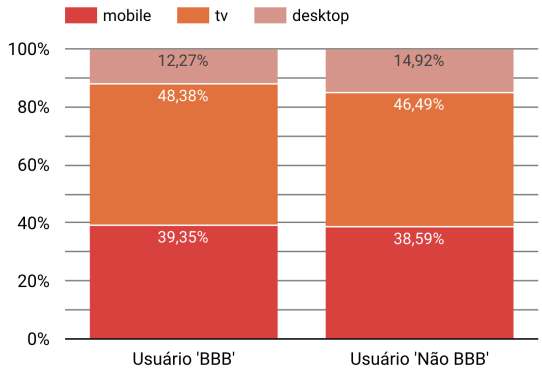


Figura 2: Distribuição de frequência por plano

Também observou-se que 51,9% dos assinantes do *streaming* são da região Sudeste do país. Quanto às plataformas utilizadas pelos usuários "BBB"e "Não BBB", nota-se que há pouca diferença. Ambos consomem mais conteúdo por meio da televisão, seguido dos dispositivos móveis e, por último, desktop.



(a) Distribuição de frequência por região



(b) Distribuição de frequência por plataforma

Figura 3: Gráficos de Região e Plataforma

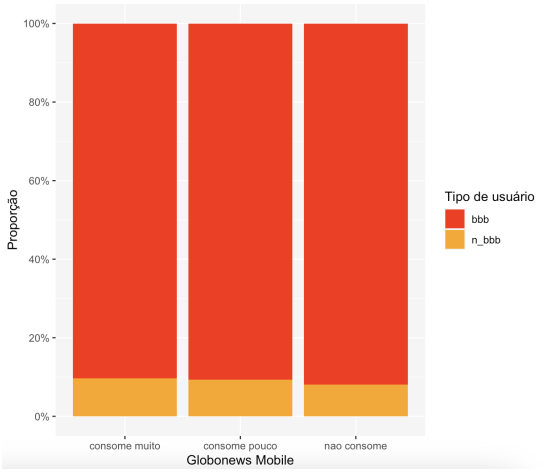
Todas essas informações quando filtradas apenas pelos usuários BBB, obtiveram as mesmas tendências.

Após a etapa de análise descritiva, todas as variáveis relacionadas ao consumo do usuário foram transformadas em variáveis categóricas para facilitar a construção do modelo. Os casos em que o consumo de determinado tipo de conteúdo era igual a zero foram categorizados como "não consome", aqueles com consumo inferior à mediana foram categorizados como "consome pouco"e aqueles com consumo superior à mediana foram categorizados como "consome muito". A variável idade também foi categorizada por faixas etárias.

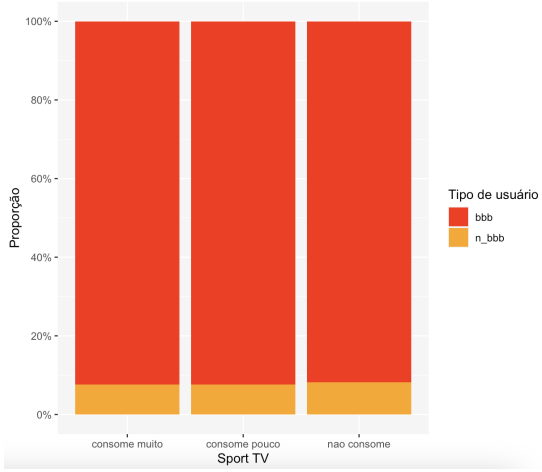
Inicialmente, o plano era aplicar diretamente o modelo completo aos dados e, em seguida, utilizar o método stepwise para selecionar as variáveis relevantes. Entretanto,

devido ao tamanho considerável da base de dados, o RStudio não conseguiu processá-la, o que inviabilizou essa abordagem. Diante dessa limitação, optou-se por realizar uma amostragem estratificada por meio de sorteio. No entanto, novamente pela grande quantidade de dados, o sorteio também não pôde ser concluído, o que exigiu a adoção de uma nova alternativa.

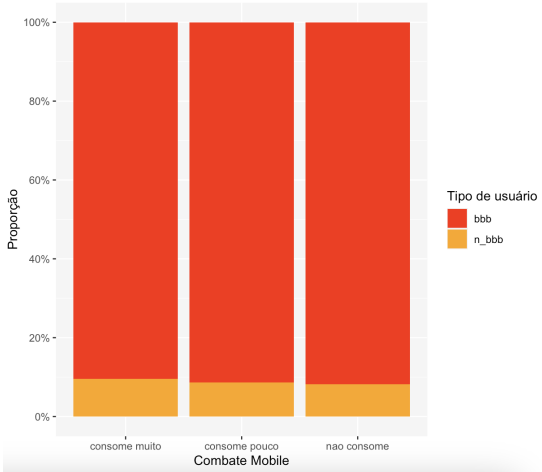
Dessa forma, decidiu-se realizar uma seleção manual das variáveis por meio de análise bivariada entre a variável resposta e as variáveis independentes, utilizando gráficos de barras. Essa etapa resultou na redução de 107 variáveis para 46. Abaixo seguem 3 exemplos de variáveis que foram excluídas:



(a) Distribuição de frequência em proporção de Globonews Mobile por tipo de usuário



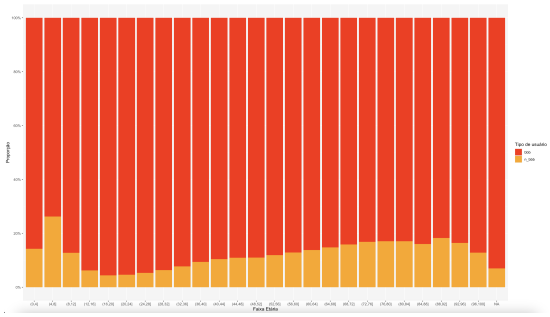
(b) Distribuição de frequência em proporção de Sportv TV por tipo de usuário



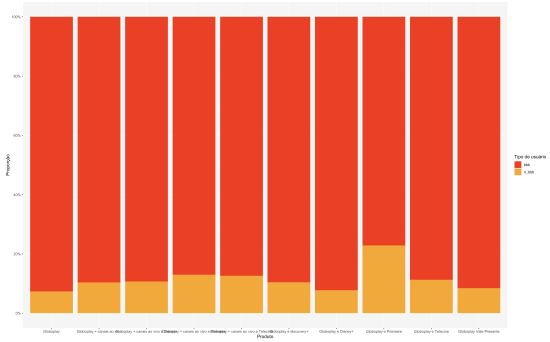
(c) Distribuição de frequência em proporção de Combate Mobile por tipo de usuário

Figura 4: Gráficos das variáveis excluídas manualmente

Na Figura 5 são apresentados 2 exemplos de variáveis que permaneceram na base de dados. Verificou-se que a distribuição de frequência nas categorias das covariáveis diferem quanto a ser ou não ser usuário BBB:



(a) Distribuição de frequência em proporção de Faixa Etária por tipo de usuário



(b) Distribuição de frequência em proporção de Produto por tipo de usuário

Figura 5: Gráficos das variáveis que foram mantidas

Em seguida, foi realizado um sorteio para selecionar aleatoriamente 37.877 usuários, utilizando uma abordagem de estratificação. Os estratos foram feitos a partir das variáveis BBB, gênero, faixa etária, estado e nome do produto assinado.

Após todas essas etapas, finalmente foi realizada a cestimação dos modelos de regressão logística. Para isso, foi usada a função 'glm()' com o objetivo de selecionar as variáveis mais significativas para o modelo.

O modelo completo foi construído contendo todas as variáveis da base de dados criada a partir do sorteio. Em seguida, foi aplicado o método *stepwise* para selecionar as variáveis mais significativas, reduzindo o modelo inicial. O modelo obtido foi o seguinte:

Tabela 1: Modelo 1 - Modelo Aplicando o Stepwise

Variável	**Estimativa**	**IC(β)$_{95\%}$**	**p-valor**
(Intercepto)	114.200	(-36788.700, 37020.900)	0.006
Nome Produto			
Globoplay	–	–	–
Globoplay + canais ao vivo	0.135	(-0.257, 0.528)	0.299
Globoplay + canais ao vivo e Disney+	0.815	(-0.373, 1.603)	0.014
Globoplay + canais ao vivo e Premiere	1.501	(-0.197, 3.200)	0.051
Globoplay + canais ao vivo e Telecine	18.390	(-22818.880, 23028.880)	0.999
Globoplay e discovery+	19.480	(-19764.470, 19784.470)	0.998
Globoplay e Disney+	0.733	(0.295, 1.111)	0.001

Variable	**Estimativa**	**IC(β)$_{95\%}$**	**p-valor**
Globoplay e Premiere	0.530	(-0.588, 1.547)	0.313
Globoplay e Telecine	1.586	(0.591, 2.581)	0.002
Globoplay Vale-Presente	22.630	(-20722.320, 20727.080)	0.998
Variedades Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.397	(-1.514, -0.409)	0.171
Não Consome	0.223	(-0.664, 0.128)	0.312
Series TV			
Consome Muito	–	–	–
Consome Pouco	-0.242	(-0.578, 0.711)	0.169
Não Consome	0.108	(-0.529, 0.187)	0.521
Series Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.011	(-0.517, 1.230)	0.942
Não Consome	0.309	(-0.081, 0.327)	0.023
Reality TV			
Consome Muito	–	–	–
Consome Pouco	0.013	(-0.205, 1.058)	0.958
Não Consome	-2.799	(-3.993, -1.199)	< 0.001
Reality Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.121	(-0.549, 0.750)	0.567
Não Consome	-3.555	(-3.704, -2.889)	< 0.001
Reality Desktop			
Consome Muito	–	–	–
Consome Pouco	-0.470	(-1.144, 0.400)	0.340
Não Consome	-3.567	(-5.309, -4.533)	< 0.001
Multishow Mobile			
Consome Muito	–	–	–
Consome Pouco	1.248	(-0.746, 1.567)	0.021
Não Consome	0.526	(-0.274, 1.218)	0.210
Futura Desktop			
Consome Muito	–	–	–
Consome Pouco	16.670	(-696.881, 1400.592)	0.964
Não Consome	18.360	(-1251.704, 687.481)	0.960
Premiere TV			
Consome Muito	–	–	–
Consome Pouco	-1.754	(-1.382, 0.241)	0.030
Não Consome	-1.196	(-0.800, 0.259)	0.030
BBB TV			
Consome Muito	–	–	–
Consome Pouco	-5.426	(-1924.623, 2300.394)	0.996
Não Consome	-34.020	(-1569.340, 2996.510)	0.966

Variable	Estimativa	$IC(\beta)_{95\%}$	p-valor
BBB Mobile			
Consome Muito	–	–	–
Consome Pouco	-10.880	(-2109.344, 1688.629)	0.992
Não Consome	-34.470	(-1634.424, 2266.136)	0.962
BBB Desktop			
Consome Muito	–	–	–
Consome Pouco	-0.093	(-2578.558, 3051.671)	1.000
Não Consome	-35.990	(-2400.281, 2851.220)	0.972
Mobile P			
Consome Muito	–	–	–
Consome Pouco	0.035	(-0.207, 0.291)	0.755
Não Consome	-0.305	(-0.568, -0.004)	0.046
TV P			
Consome Muito	–	–	–
Consome Pouco	-0.327	(-0.817, 0.168)	0.019
Não Consome	-0.043	(-0.371, 0.282)	0.783
Faixa Etária			
(16,20]	–	–	–
(20,24]	-19.720	(-38638.420, 38598.200)	0.999
(24,28]	-19.930	(-38735.530, 38671.840)	0.999
(28,32]	-19.650	(-40163.070, 19991.030)	0.999
(32,36]	-19.830	(-40173.070, 19991.230)	0.999
(36,40]	-19.960	(-40174.070, 19991.260)	0.999
(40,44]	-20.270	(-40174.070, 19991.440)	0.999
(44,48]	-20.330	(-40178.070, 19991.380)	0.999
(48,52]	-20.630	(-40181.070, 19991.250)	0.999
(52,56]	-19.940	(-40172.070, 19991.380)	0.999
(56,60]	-20.470	(-40180.070, 19991.210)	0.999
(60,64]	-20.260	(-40178.070, 19991.440)	0.999
(64,68]	-19.550	(-40170.070, 19991.520)	0.999
(68,72]	-19.980	(-40172.070, 19991.530)	0.999
(72,76]	-20.400	(-40174.070, 19991.350)	0.999
(76,80]	-20.190	(-40173.070, 19991.440)	0.999
(80,84]	-18.890	(-40167.070, 19991.800)	0.999
(84,88]	-26.910	(-41164.070, 19981.150)	0.999
(88,92]	-18.000	(-42872.070, 19976.870)	0.999

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 18978.5 on 37876 degrees of freedom

Residual deviance: 2830.1 on 37820 degrees of freedom

AIC: 2944.1

Number of Fisher Scoring iterations: 23

Apesar do uso do método *stepwise*, o modelo resultou em um número considerável de variáveis não significativas, com p-valores muito altos. Portanto, as variáveis que apresentaram todas as categorias com p-valor acima de 0.9 foram removidas.

Ainda assim, algumas variáveis se mostraram não significativas para o modelo, então mais uma vez houve a exclusão destas variáveis e o terceiro modelo foi estimado e os resultados estão apresentados na Tabela 2:

Tabela 2: Modelo 3 - Removendo variáveis com todas as categorias com p-valor > 0.20

Variable	**Estimativa**	**IC(β)$_{95\%}$**	**p-valor**
(Intercept)	9.867	(7.620, 12.114)	< 0.001 ***
Gênero			
Feminino	–	–	–
Masculino	-0.161	(-0.259, -0.064)	0.001 **
Não quero informar	-0.066	(-0.398, 0.266)	0.693
Outro	0.135	(-0.071, 0.340)	0.198
Nome Produto			
Globoplay	–	–	–
Globoplay + canais ao vivo	-0.474	(-0.605, -0.344)	< 0.001 ***
Globoplay + canais ao vivo e Disney+	-0.052	(-0.438, 0.333)	0.790
Globoplay + canais ao vivo e Premiere	1.126	(0.081, 2.172)	0.035 *
Globoplay + canais ao vivo e Telecine	13.429	(-521.330, 548.187)	0.960
Globoplay e discovery+	13.899	(-450.450, 478.249)	0.953
Globoplay e Disney+	0.282	(0.055, 0.510)	0.015 *
Globoplay e Premiere	-0.314	(-0.839, 0.211)	0.219
Globoplay e Telecine	1.058	(0.419, 1.697)	0.001 **
Globoplay Vale-Presente	14.044	(-489.603, 517.691)	0.955
Variedades Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.115	(-0.254, 0.024)	0.107
Não Consome	-0.175	(-0.305, -0.044)	0.008 **
Series Mobile			
Consome Muito	–	–	–
Consome Pouco	0.194	(-0.029, 0.416)	0.089
Não Consome	0.067	(-0.145, 0.279)	0.533

Variable	Estimativa	$IC(\beta)_{95\%}$	p-valor
Reality TV			
Consome Muito	–	–	–
Consome Pouco	-0.070	(-0.466, 0.325)	0.725
Não Consome	-2.802	(-3.092, -2.511)	< 0.001 ***
Reality Mobile			
Consome Muito	–	–	–
Consome Pouco	0.049	(-0.298, 0.396)	0.790
Não Consome	-3.605	(-3.854, -3.355)	< 0.001 ***
Reality Desktop			
Consome Muito	–	–	–
Consome Pouco	-0.150	(-1.072, 0.773)	0.720
Não Consome	-2.790	(-3.393, -2.188)	< 0.001 ***
Filmes TV			
Consome Muito	–	–	–
Consome Pouco	0.011	(-0.188, 0.210)	0.913
Não Consome	-0.339	(-0.501, -0.177)	< 0.001 ***
Globo Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.050	(-0.181, 0.081)	0.420
Não Consome	0.506	(0.384, 0.629)	< 0.001 ***
Globo TV			
Consome Muito	–	–	–
Consome Pouco	0.036	(-0.132, 0.204)	0.679
Não Consome	0.728	(0.578, 0.878)	< 0.001 ***
Multishow TV			
Consome Muito	–	–	–
Consome Pouco	-0.033	(-0.414, 0.348)	0.912
Não Consome	-0.860	(-1.271, -0.450)	< 0.001 ***
Multishow Mobile			
Consome Muito	–	–	–
Consome Pouco	0.464	(-0.061, 0.989)	0.086
Não Consome	-0.551	(-0.922, -0.187)	< 0.001 ***
MaisNaTela Desktop			
Consome Muito	–	–	–
Consome Pouco	0.751	(-0.295, 1.798)	0.431
Não Consome	1.128	(-0.607, 2.864)	0.068
Combate Desktop			
Consome Muito	–	–	–
Consome Pouco	13.183	(-1208.879, 1235.245)	0.983
Não Consome	1.697	(-0.177, 3.571)	0.070

Variable	**Estimativa**	**IC(β)$_{95\%}$**	**p-valor**
CBN TV			
Consome Muito	–	–	–
Consome Pouco	0.091	(-0.217, 0.399)	0.828
Não Consome	-0.632	(-1.182, -0.081)	0.024 *
Consumo Live			
Consome Muito	–	–	–
Consome Pouco	-0.812	(-0.919, -0.704)	< 0.001 ***
Não Consome	-0.012	(-0.105, 0.082)	0.796
Consumo VOD			
Consome Muito	–	–	–
Consome Pouco	-0.009	(-0.145, 0.127)	0.842
Não Consome	0.935	(0.583, 1.286)	< 0.001 ***
Mobile P			
Consome Muito	–	–	–
Consome Pouco	-0.003	(-0.130, 0.125)	0.976
Não Consome	-0.343	(-0.491, -0.195)	0.000 ***
TV P			
Consome Muito	–	–	–
Consome Pouco	-0.089	(-0.215, 0.037)	0.170
Não Consome	-0.432	(-0.573, -0.290)	< 0.001 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
(Dispersion parameter for binomial family taken to be 1)			
Null deviance: 18979 on 37876 degrees of freedom			
Residual deviance: 11809 on 37835 degrees of freedom			
AIC: 11892			

Desta vez, todas as variáveis se mostraram significativas, ainda assim foi feito um novo modelo excluindo as variáveis que tinham p-valor maior que 0.05 para ver o que acontecia com o modelo e obteve-se o seguinte:

Tabela 3: Modelo 4 - Removendo variáveis com todas as categorias com p-valor>0.05

Variáveis	**Estimativa**	**IC(β)$_{95\%}$**	**p-valor**
(Intercept)	10.914	(8.802, 13.026)	< 0.001 ***
Genero			
Feminino	–	–	–
Masculino	-0.159	(-0.257, -0.061)	0.002 **
Não quero informar	-0.062	(-0.391, 0.266)	0.709
Outro	0.136	(-0.070, 0.341)	0.195
nome produto			
Globoplay	–	–	–
Globoplay + canais ao vivo	-0.486	(-0.615, -0.355)	< 0.001 ***
Globoplay + canais ao vivo e Disney+	-0.066	(-0.429, 0.298)	0.739
Globoplay + canais ao vivo e Premiere	1.132	(0.074, 2.188)	0.034 *
Globoplay + canais ao vivo e Telecine	13.441	(-526.481, 553.363)	0.960
Globoplay e discovery+	13.892	(-445.686, 473.469)	0.953
Globoplay e Disney+	0.284	(0.056, 0.512)	0.014 *
Globoplay e Premiere	-0.323	(-0.822, 0.177)	0.206
Globoplay e Telecine	1.063	(0.423, 1.703)	0.001 **
Globoplay Vale-Presente	14.039	(-471.879, 499.957)	0.955
Series Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.113	(-0.252, 0.026)	0.113
Não Consome	-0.174	(-0.303, -0.046)	0.009 **
Reality TV			
Consome Muito	–	–	–
Consome Pouco	-0.069	(-0.461, 0.322)	0.731
Não Consome	-2.800	(-3.092, -2.507)	< 0.001 ***
Reality Mobile			
Consome Muito	–	–	–
Consome Pouco	0.049	(-0.313, 0.411)	0.789
Não Consome	-3.580	(-3.836, -3.323)	< 0.001 ***
Reality Desktop			
Consome Muito	–	–	–
Consome Pouco	-0.132	(-0.957, 0.692)	0.751
Não Consome	-2.771	(-3.071, -2.471)	< 0.001 ***
Filmes TV			
Consome Muito	–	–	–
Consome Pouco	0.005	(-0.191, 0.201)	0.956
Não Consome	-0.338	(-0.165, 0.069)	< 0.001 ***
Globo Mobile			
Consome Muito	–	–	–
Consome Pouco	-0.048	(-0.170, 0.074)	0.439
Não Consome	0.510	(0.389, 0.631)	< 0.001 ***

Variable	Estimativa	$IC(\beta)_{95\%}$	p-valor
Globo TV			
Consome Muito	–	–	–
Consome Pouco	0.042	(-0.128, 0.213)	0.629
Não Consome	0.631	(0.580, 0.880)	< 0.001 ***
Multishow TV			
Consome Muito	–	–	–
Consome Pouco	-0.039	(-0.363, 0.285)	0.894
Não Consome	-0.870	(-1.285, -0.455)	< 0.001 ***
Multishow Mobile			
Consome Muito	–	–	–
Consome Pouco	0.476	(0.046, 0.904)	0.078
Não Consome	-0.530	(-0.797, -0.261)	0.005 **
Combate Desktop			
Consome Muito	–	–	–
Consome Pouco	13.194	(-986.204, 1012.592)	0.983
Não Consome	1.681	(-3.153, 6.514)	0.072
Cbn TV			
Consome Muito	–	–	–
Consome Pouco	0.094	(-0.743, 0.931)	0.823
Não Consome	-0.635	(-1.183, -0.086)	0.023 *
Consumo Live			
Consome Muito	–	–	–
Consome Pouco	-0.807	(-0.915, -0.699)	< 0.001 ***
Consumo VOD			
Consome Muito	–	–	–
Consome Pouco	-0.009	(-0.101, 0.082)	0.842
Não Consome	0.949	(0.697, 1.201)	< 0.001 ***
Mobile p			
Consome Muito	–	–	–
Consome Pouco	-0.002	(-0.108, 0.103)	0.976
Não Consome	-0.333	(-0.479, -0.187)	< 0.001 ***
TV p			
Consome Muito	–	–	–
Consome Pouco	-0.090	(-0.218, 0.038)	0.165
Não Consome	-0.439	(-0.579, -0.299)	< 0.001 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
(Dispersion parameter for binomial family taken to be 1)			
Null deviance: 18979 on 37876 degrees of freedom			
Residual deviance: 11809 on 37835 degrees of freedom			
AIC: 11893			

Comparado os dois últimos modelos, é possível observar que a maioria das variáveis possui estimativas semelhantes em ambos os modelos. Isso indica que as variáveis selecionadas para o modelo 4 têm um efeito semelhante no resultado em comparação com o modelo 3. Além disso, considerando o critério do AIC, o modelo 4 apresenta um valor de AIC quase idêntico ao modelo 3.

Portanto, considerando a similaridade das estimativas dos coeficientes e o fato de que o AIC do modelo 4 quase não se altera, este se mostra ser a escolha mais apropriada. Esse modelo simplificado tem a vantagem de manter uma boa capacidade de explicação dos dados, mas com menos variáveis, o que facilita sua interpretação e reduz a complexidade do modelo.

A seguir é apresentada a tabela com as razões de chance (odds ratio) bem como o intervalo de confiança das razões de chances do modelo selecionado:

Tabela 4: Variáveis e Odds Ratio com Intervalos de Confiança (IC$(OR)_{95\%}$)

Variáveis	**Odds Ratios**	**IC$(OR)_{95\%}$**
Genero		
Feminino	–	–
Masculino	0.853	(0.773, 0.942)
Não quero informar	0.940	(0.675, 1.305)
Outro	1.145	(0.933, 1.373)
nome produto		
Globoplay	–	–
Globoplay + canais ao vivo	0.615	(0.540, 0.695)
Globoplay + canais ao vivo e Disney+	0.936	(0.650, 1.346)
Globoplay + canais ao vivo e Premiere	3.098	(1.077, 7.929)
Globoplay + canais ao vivo e Telecine	548555.411	(0.000, inf)
Globoplay e discovery+	894307.228	(0.000, inf)
Globoplay e Disney+	1.327	(1.057, 1.648)
Globoplay e Premiere	0.724	(0.440, 1.195)
Globoplay e Telecine	2.892	(1.527, 5.466)
Globoplay Vale-Presente	1204606.390	(0.000, inf)
Series Mobile		
Consome Muito	–	–
Consome Pouco	0.893	(0.777, 1.029)
Não Consome	0.841	(0.725, 0.967)

Variáveis	**Odds Ratios**	**IC(OR)$_{95\%}$**
Reality TV		
Consome Muito	–	–
Consome Pouco	0.934	(0.631, 1.358)
Não Consome	0.061	(0.078, 0.081)
Reality Mobile		
Consome Muito	–	–
Consome Pouco	1.050	(0.730, 1.525)
Não Consome	0.028	(0.021, 0.034)
Reality Desktop		
Consome Muito	–	–
Consome Pouco	0.876	(0.384, 1.997)
Não Consome	0.062	(0.049, 0.078)
Filmes TV		
Consome Muito	–	–
Consome Pouco	1.005	(0.827, 1.250)
Não Consome	0.713	(0.848, 0.935)
Globo Mobile		
Consome Muito	–	–
Consome Pouco	0.953	(0.845, 1.082)
Não Consome	1.665	(0.497, 2.442)
Globo TV		
Consome Muito	–	–
Consome Pouco	1.043	(0.880, 1.238)
Não Consome	1.880	(1.787, 2.394)
Multishow TV		
Consome Muito	–	–
Consome Pouco	0.962	(0.696, 1.327)
Não Consome	0.419	(0.278, 0.321)
Multishow Mobile		
Consome Muito	–	–
Consome Pouco	1.608	(1.048, 2.470)
Não Consome	0.588	(0.459, 0.771)
Combate Desktop		
Consome Muito	–	–
Consome Pouco	607573.390	(0.000, inf)
Não Consome	5.375	(0.042, inf)

Variáveis	**Odds Ratios**	**IC(OR)$_{95\%}$**
Cbn TV		
Consome Muito	–	–
Consome Pouco	1.099	(0.477, 2.540)
Não Consome	0.531	(0.308, 0.725)
Consumo Live		
Consome Muito	–	–
Consome Pouco	0.166	(0.153, 0.180)
Consumo VOD		
Consome Muito	–	–
Consome Pouco	0.988	(0.812, 1.212)
Não Consome	2.570	(1.886, 3.508)
Mobile p		
Consome Muito	–	–
Consome Pouco	0.998	(0.898, 1.108)
Não Consome	0.714	(0.623, 0.817)
TV p		
Consome Muito	–	–
Consome Pouco	0.915	(0.814, 1.030)
Não Consome	0.647	(0.575, 0.727)

Com base nos resultados obtidos pelo modelo, pôde-se observar algumas tendências interessantes. O sexo do usuário demonstrou ter um impacto significativo nas chances de ser um usuário BBB, ou seja, ser uma pessoa que assinou o *stream* e teve seu primeiro conteúdo sendo algo relacionado ao *reality show*. Os assinantes do sexo feminino apresentaram um aumento de 17% nas chances de serem assinantes BBB em comparação aos assinantes do sexo masculino.

A escolha do plano de assinatura no Globoplay também influencia as chances de ser um usuário BBB. Os assinantes do Globoplay básico têm 62% mais chances de serem assinantes BBB em comparação com aqueles que optam pelo Globoplay + Canais Ao Vivo. Já os usuários que assinam o Globoplay + Canais Ao Vivo e Premiere apresentam um aumento de 3.1 vezes mais chances de serem assinantes BBB em relação aos que escolhem o Globoplay básico. Além disso, os assinantes do Globoplay e Disney e do Globoplay e Telecine têm, respectivamente, 1.32 vezes e 2.8 vezes mais de chances de serem assinantes BBB em comparação com os assinantes básicos da plataforma.

O comportamento de consumo de conteúdo também desempenha um papel importante nas chances de um usuário ser um espectador BBB. Aqueles que consomem muitas séries pelo celular apresentam 18% mais chances de serem usuários BBB em comparação com os

que não consomem séries pelo celular. Da mesma forma, os usuários que consomem muitos filmes na tv têm uma vantagem de 40% em relação aos que não consomem, aumentando suas chances de serem usuários BBB.

O gênero de programas do tipo *reality* parece ter um forte apelo para os usuários BBB, o que é bastante compreensível, uma vez que o BBB é um *reality show*. Aqueles que consomem muito conteúdo de *realities*, seja por meio de dispositivos móveis, web ou TV, apresentam um aumento significativo nas chances de serem usuários BBB. Essas chances aumentam em 35.86, 15.98 e 16.43 vezes, respectivamente, em comparação com aqueles que não consomem nada relacionado a *realities*.

Aqueles que consomem muito conteúdo do Multishow pelo celular e tv apresentam um aumento de 1.69 e 2.38 nas chances de serem usuários BBB em comparação com aqueles que não consomem o canal. No entanto, é interessante notar que o consumo do canal Globo na plataforma de *streaming* tem um efeito oposto nas chances de ser um espectadador BBB. Tanto para aqueles que consomem o canal na TV quanto para aqueles que o consomem por meio de dispositivo móvel, há uma redução nas chances em 87% e 66%, respectivamente.

Assistir ao canal combate também tem um impacto significativo. Os usuários que não assistem ao canal pelo *desktop* tem 5.37 vezes mais chance de ser usuario BBB em comparação aos que consomem muito esse conteúdo. Ainda sobre canais, o efeito do canal Cbn também se mostra relevante. Os usuários que consomem o canal têm 88% mais chance de serem assinantes BBB do que aqueles que não consomem.

Além disso, o hábito de assistir a conteúdos ao vivo também influencia consideravelmente a probabilidade de ser um espectador BBB. Os consumidores assíduos de conteúdos ao vivo têm um aumento de 2.24 vezes nas chances em comparação com aqueles que assistem pouco este tipo de conteúdo. Nota-se também que aqueles que assistem pouco conteúdo *on demand* têm 2.58 vezes mais chances de serem usuários BBB. Uma vez que, se não consomem tanto conteúdo *on demand*, é provável que consumam mais conteúdo ao vivo.

Além das variáveis mencionadas anteriormente, também se constatou que o tipo de plataforma utilizada desempenha um papel relevante. Aqueles que consomem conteúdo tanto pela TV quanto pelo celular têm um aumento de 55% e 39%, respectivamente, nas chances de serem usuários BBB em comparação com aqueles que não consomem conteúdo nessas plataformas.

4 Conclusões

Este capítulo revela os resultados do estudo teve como objetivo ajustar um modelo logístico utilizando o software Rstudio, utilizando dados fornecidos pela Globo no período de janeiro a abril de 2022. O objetivo foi analisar e identificar o perfil dos usuários do BBB. Entende-se como usuário BBB aquele que assinou o Globoplay e teve seu primeiro consumo algo relacionado ao *reality show*.

Um perfil ideal do usuário do programa "Big Brother Brasil"pode ser elaborado com base nas informações fornecidas pelo modelo.

Primeiramente, observou-se que a faixa etária média dos usuários BBB é de 36 anos. Isso sugere que o programa atrai principalmente pessoas que estão na fase adulta e possuem uma certa maturidade para apreciar e se envolver nas dinâmicas do *reality show*.

Em relação ao gênero, mais da metade dos usuários BBB são do sexo feminino, além disso, esse fato aumenta em 17% as chances de ser um assinante BBB. Isso indica uma forte presença e identificação das mulheres com o programa, o que pode ser atribuído a fatores como o envolvimento emocional, as relações interpessoais e os dramas que são apresentados no contexto do *reality*.

A região Sudeste do Brasil se destaca como a área geográfica com a maior representatividade de assinantes do *streaming* que consomem o BBB, isso evidencia a popularidade do programa nessa região do país.

Quanto aos planos de assinatura, o Globoplay básico é o mais popular, representando cerca de 70% das assinaturas. No entanto, o perfil ideal de usuário BBB opta pelo plano Globoplay + Canais Ao Vivo e Premiere, pois essa escolha aumenta significativamente as chances de ser um espectador assíduo do programa. Essa preferência pelo plano pode indicar um maior interesse em acompanhar outras transmissões ao vivo além das relacionados ao BBB.

No que diz respeito ao comportamento de consumo, tanto os usuários do BBB quanto

os não usuários apresentam padrões semelhantes, com maior consumo de conteúdo por meio da televisão, seguido de dispositivos móveis e, por último, *desktop*. Isso sugere que a plataforma de *streaming* é uma alternativa conveniente e acessível, permitindo que os usuários acompanhem os conteúdos em diferentes dispositivos de acordo com sua conveniência.

Além disso, o perfil ideal consome séries, filmes e também demonstra interesse por outros *realities* já que consome muitos programas da categoria *reality*.

Em resumo, o perfil ideal de usuário do BBB é composto por pessoas com uma idade média de 36 anos, predominantemente do sexo feminino e provenientes da região Sudeste do Brasil. Esses usuários tendem a optar pelo plano Globoplay + Canais Ao Vivo e Premiere, além de consumirem séries pelo celular e filmes na TV. Mostram um interesse significativo por programas do tipo *reality* e têm preferência por conteúdos ao vivo em detrimento dos conteúdos *on demand*. É interessante notar também que esses usuários não têm o hábito de assistir ao canal da Globo pelo serviço de *streaming*.

Essas observações fornecem informações valiosas, permitindo uma compreensão mais precisa dos dos usuários que consomem o *Big Brother Brasil*. Essas informações podem ser úteis para direcionar estratégias de marketing e segmentar o público-alvo com maior precisão, garantindo a entrega de conteúdos cada vez mais atrativos e relevantes, o que contribui para manter esses usuários engajados com o produto.

Referências

AKAIKE, H. et al. *Selected papers of hirotugu akaike*. [S.l.]: Springer Science & Business Media, 1998.

EKE, C. I. et al. A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access*, IEEE, v. 7, p. 144907–144924, 2019.

LIMITED, G. W. R. *Guinness Book*. 2020. Disponível em: ⟨https://www.guinnessworldrecords.com.br/about-us/our-story⟩.

NANANUKUL, N. An inference model for online media users. *Journal of Data Science*, , v. 11, n. 1, p. 143–155, 2013.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

NETER, J. et al. Applied linear statistical models. Irwin Chicago, 1996.

POSIT. *RStudio*. 2023. Disponível em: ⟨https://docs.posit.co/ide/user/ide/get-started/⟩.

RAGHURAM, M. A.; AKSHAY, K.; CHANDRASEKARAN, K. Efficient user profiling in twitter social network using traditional classifiers. In: *Intelligent systems technologies and applications*. [S.l.]: Springer, 2016. p. 399–411.