UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus Presidente Prudente

MARIA ISABELA DE SOUZA POMPEI

ANÁLISE DO PERFIL DE CONSULTORES DE EMPRESA DE VENDAS UTILIZANDO MODELO DE REGRESSÃO LOGÍSTICA

Presidente Prudente
2023

MARIA ISABELA DE SOUZA POMPEI

ANÁLISE DO PERFIL DE CONSULTORES DE EMPRESA DE VENDAS UTILIZANDO MODELO DE REGRESSÃO LOGÍSTICA

Relatório Final do Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp.
Orientador: Prof. Dr. Mário Hissamitsu Tarumoto.

Presidente Prudente
2023

P788a

Pompei, Maria Isabela de Souza
Análise do perfil de consultores de empresa de vendas utilizando modelo de Regressão Logística / Maria Isabela de Souza Pompei. -- Presidente Prudente, 2023
72 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Mário Hissamitsu Tarumoto

1. Regressão Logística. 2. Modelo de Previsão. 3. Estatística. 4. Segmentação de Perfil. 5. Empresa de Vendas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

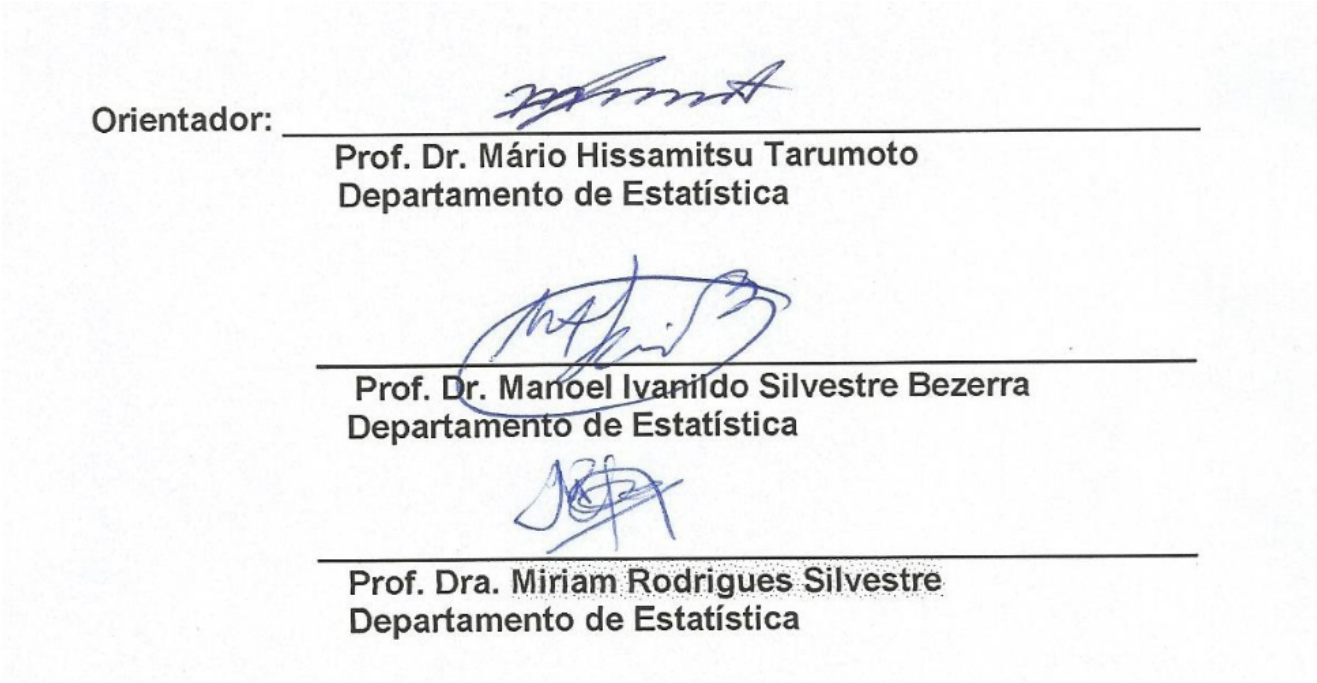
Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

MARIA ISABELA DE SOUZA POMPEI

ANÁLISE DO PERFIL DE CONSULTORES DE EMPRESA DE VENDAS UTILIZANDO MODELO DE REGRESSÃO LOGÍSTICA

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



Presidente Prudente, 28 de Janeiro de 2023

AGRADECIMENTOS

Gostaria de agradecer primeiramente ao Prof. Dr. Mário Hissamitsu Tarumoto, pela orientação, paciência e confiança que me ofereceu nas etapas de desenvolvimento desse trabalho.

Agradeço também aos professores da banca, Miriam Rodrigues Silvestre e Manoel Ivanildo Silvestre Bezerra por toda a dedicação, avaliação e sugestões para que esse trabalho fosse bem desenvolvido.

Um agradecimento especial aos meus pais, a minha irmã e a minha família, pelo apoio e incentivo em todos os momentos.

Aos amigos do curso de Estatística, pelas trocas de conhecimentos ao longo da jornada. A todos os professores do Departamento de Estatística que ao longo desses quatro anos ensinaram o conhecimento que adquiri na graduação.

Muito obrigada a todos!

EPÍGRAFE

"Na vida devemos ter raízes, e não âncoras. Raiz alimenta, âncora imobiliza."
Mário Sergio Cortella

RESUMO

Entre diversas empresas de vendas diretas presentes no Brasil e no mundo, a Mary Kay é uma importante companhia no segmento de vendas de cosméticos. A corporação tem uma estrutura de organização no qual existe um líder que é responsável pelo desempenho de seus liderados, também chamados de consultores, cujo principal objetivo é ajudar seus liderados a terem uma boa performance nas vendas. É importante para o líder que seus subordinados sejam "ativos"na empresa, contudo, esse trabalho aborda a utilização do modelo estatístico preditivo de Regressão Logística para prever a probabilidade dos consultores de serem ativos na companhia, objetivando auxiliar as estratégias que serão desenvolvidas pelo líder a fim de ter a maior quantidade de liderados ativos na equipe. Para a construção do modelo, foi utilizada uma base de dados cedida por um líder da companhia, com o objetivo de aumentar a quantidade de variáveis, foi aplicado um questionário e também construída algumas variáveis. Com a base de dados completa, criou-se modelos preditivos utilizando a técnica de Regressão Logística. Após comparação dos resultados, percebeu-se que o modelo criado utilizando 70% da base de dados para treinar o modelo e 30% da base de dados para testar o modelo, utilizando as variáveis contínuas em formato não categorizado obteve uma melhor performance. Contudo, os modelos com melhores diagnósticos apresentou as métricas: Acurácia, Precisão, Sensibilidade, Especificidade e F1-Score superior a 90% apontou. Os modelos apontaram que as três variáveis significativas e que explicam a probabilidade do consultor(a) de ser ativo(a), são, Idade, Renda Per Capita e o Coeficiente de Atividade.

Palavras-chave: Modelo de Regressão Logística. Diagnóstico do Modelo. Modelos Preditivos. Consultor Ativo.

ABSTRACT

Among several direct sales companies present in Brazil and in the world, Mary Kay is an important company in the cosmetics sales segment. The corporation has an organizational structure in which there is a leader who is responsible for the performance of his subordinates, also called consultants, whose main objective is to help his subordinates to have a good performance in sales. It is important for the leader that his subordinates are "active"in the company, however, this work addresses the use of the predictive statistical model of Logistic Regression to predict the probability of the consultants to be active in the company, aiming to help the strategies that will be developed by the leader in order to have the largest number of active team members. To build the model, a database provided by a company leader was used, with the aim of increasing the number of variables, a questionnaire was applied and some variables were also constructed. With the complete database, predictive models were created using the Logistic Regression technique. After comparing the results, it was noticed that the model created using 70% of the database to train the model and 30% of the database to test the model, using continuous variables in uncategorized format, obtained a better performance. However, the model with the best diagnosis pointed out that three variables are significant and explain the consultant's probability of being active: Age, Per Capita Income and the Activity Coefficient.

Keywords: Logistic Regression Model. Model Diagnostics. Predictive Models. Active Consultant.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Dentre as múltiplas empresas de vendas diretas presentes no Brasil e no mundo, a Mary Kay é uma importante corporação a qual se caracteriza como venda de cosméticos. Em qualquer tipo de empresa, para que haja expansão do alcance de seus produtos, a empresa realiza a implementação de colaboradores treinados para que realizem as vendas. No ramo dos cosméticos, nada se diferencia, afinal, ao tratar-se da empresa estudada, para que se efetive o ingresso de um consultor de beleza, necessita-se da realização de um cadastro simples. Com este modelo de negócios, além dos consultores auferirem lucro sobre os produtos vendidos, a empresa oportuniza o crescimento escalonado através do plano de carreira interno.

A empresa Mary Kay foi fundada por Mary Kay Ash, com o apoio financeiro de U$ 5.000,00 emprestado pelo seu filho Richard, em 1963, na cidade de Dallas no estado americano de Texas (ASH, 2013). Atualmente é uma das maiores companhias no segmento de cuidados com a pele, maquiagem e fragrâncias nos EUA, existente há mais de 55 anos e opera em mais de 40 países, Mary Kay é uma companhia de referência com história e êxito profissional (ASH, 2018).

A entrada desta empresa no Brasil, ocorreu no ano de 1998, com o mesmo modelo utilizado em outros países. Esta modalidade de empresa permite aos seus consultores a opção de obter renda extra ou até mesmo renda única com a venda dos produtos, aplicando a todos os seus consultores o princípio da isonomia, isto é gerar as mesmas oportunidade de desenvolvimento para homens e mulheres.

O plano de carreira da empresa fora desenvolvido para que os consultores de beleza tenham a oportunidade de alcançar cargos maiores de forma meritocrática, cargo esses como por exemplo de diretor de vendas independente. Atingido o cargo de diretor, este obterá lucros oriundos de duas vias: em relação a venda dos produtos e em razão do progresso dos consultores que estão abaixo do diretor.

O diretor de vendas é responsável pelo time de consultores e necessita auxiliá-los em estratégias de vendas, direcionamento de produtos, orientações e nas dificuldades em que os mesmos apresentam ao decorrer do percurso na empresa. Em contrapartida, uma das maiores dificuldades que o diretor de vendas pode encontrar, se baseia na ausência de análise e interpretação estatística dos dados, pois se o mesmo realizasse a análise de dados a fim de obter o perfil de consultor com maiores probabilidades de ser ativo dentro da empresa, o líder poderia direcionar maiores estratégias, dedicação de tempo e estar em busca dos mesmos perfis para trazer ao seu time. Esta análise poderia ainda indicar, quais características são as mais importantes para obter consultores com sucesso nas vendas e

consequentemente ocorrer uma progressão na carreira.

Segundo Bussab e Morettin (2009), necessitamos trabalhar os dados para transformá-los em informações e compará-los a fim de melhor aplicar estratégias. Dessa forma, para a concretização de estudos desta natureza, deve-se definir algumas variáveis, como a título de exemplo, em quais circunstâncias um consultor poderá ser considerado um indivíduo com dificuldades de realizar vendas. Neste estudo, inicialmente será realizado análises mais aprofundadas para definir esta variável e desta forma, existe a possibilidade de aplicação de um modelo de regressão logística, tendo em vista que a variável resposta de interesse será binária.

Em virtude dos aspectos abordados, deseja-se realizar o estudo da análise do perfil dos consultores de beleza no período de onze meses, por meio da aplicação de modelo de regressão logística a fim de obter maiores lucratividades para os consultores, para os líderes e para a corporação.

A tendência atual em nível global, é a obtenção de informações úteis dos dados para serem utilizadas nas tomadas de decisões. Com grandes quantidades de dados que temos disponíveis, as empresas estão focadas em explorá-los para assim obter vantagens (PROVOST e FAWCETT, 2016). Desta forma, a necessidade de obter informações a partir de dados, gerou uma valorização dos mesmos, com isso, a cada dia mais empresas buscam profissionais que saibam manipular, traduzir e trazer informações relevantes para a corporação.

Partindo desse contexto, a área de consultoria de beleza, em específico a constância dos profissionais que desempenham tal função, pode se destacar e obter melhores resultados com o auxílio da análise e interpretação de seus dados.

A partir de trabalhos como esse, os diretores de vendas podem tomar decisões mais precisas, dedicar seu tempo em profissionais específicos, gerar promoções ou incentivo corporativos assertivos e desenvolver todas as estratégias com base nas análises estatísticas. Se realizados corretamente, utilizando trabalhos como esse, os possíveis benefícios para a empresa são: maior produtividade, maior constância dos consultores de beleza, maior precisão no perfil de consultor e por consequência, maior rentabilidade.

Entre as várias possibilidade de realização de análise, a construção de modelos de regressão logística pode identificar as principais características profissionais que levam a maior propensão a terem sucesso na empresa, informação útil que poderá ser utilizada para canalizar esforços a fim de melhorar o rendimento. Por outro lado, estes modelos podem auxiliar na identificação de características que fazem com que a propensão seja baixa, e ajudar estes profissionais a melhorarem o seu desempenho a partir de treinamentos específicos.

Desta forma, este trabalho justifica-se pela possibilidade de solução de uma si-

tuação prática por meio da aplicação de técnicas estatísticas.

O objetivo principal do trabalho foi encontrar a probabilidade de que um consultor esteja ativo, por meio da aplicação do modelo de regressão logística. Para isso, analisou-se individualmente a realização de seus pedidos de produtos pelo site todos os meses, haja visto que, estes produtos comprados significam vendas realizadas ou crescimento de estoque pessoal, e principalmente, identificar quais características aumentam a chance do colaborador de se manter ativo.

Para atingir o objetivo geral, alguns objetivos específicos foram alcançados: conhecer melhor os consultores por meio da análise exploratória de dados, identificando o perfil dos que estão cadastrados; realizar análises de dados para classificar o consultor como ativo ou não; aplicar técnicas de regressão logística para identificar quais consultores apresentarão melhor constância na empresa; aplicar o modelo para segmentar o perfil de consultor.

Este relatório está estruturado da seguinte forma: no capítulo 1 é apresentada a descrição geral do problema, justificativa e objetivos. No capítulo 2 é apresentada a companhia da qual será realizada a análise, no capítulo 3 é apresentada a teoria a ser utilizada para a continuidade do trabalho, sendo ela a Regressão Logística e no capítulo 4 está disposta a aplicação dos dados, bem como a análise exploratória. O capítulo 5 consiste na construção do modelo e por fim o capítulo 6 apresenta a conclusão do presente trabalho. Foram construídos o apêndice A com o objetivo de apresentar as demonstrações e o apêndice B com as variáveis presentes no trabalho.

2 Empresa Mary Kay

Neste capítulo será apresentada a empresa Mary Kay, bem como sua fundação, origem, características, organização e meritocracia.

2.1 Companhia

A Mary Kay Cosmetics foi fundada no dia 13 de setembro de 1963, em uma loja localizada em Dallas Texas (EUA). Segundo Ash (2013), após o falecimento da cosmetologista que fornecia produtos de cuidados com a pele, Mary Kay Ash comprou as fórmulas originais da família e fez algumas modificações nas embalagens e nos produtos, tornando-os vendáveis. Portanto, a partir do produto definido, o principal objetivo era desenvolver uma empresa que contasse com um plano de carreira no qual ofereceria oportunidades de crescimento para todas as pessoas de forma meritocrática.

2.2 Meritocracia Empresarial

A fim de atingir os objetivos empresariais e facilitar a entrada de pessoas na companhia, qualquer pessoa pode ser um consultor cadastrado na empresa Mary Kay, basta entrar em contato com um consultor da empresa e solicitar que ele faça o seu cadastro. Assim que o cadastro é realizado a pessoa recebe o título de consultor de beleza, portanto, está apto a vender os produtos e também cadastrar outras pessoas.

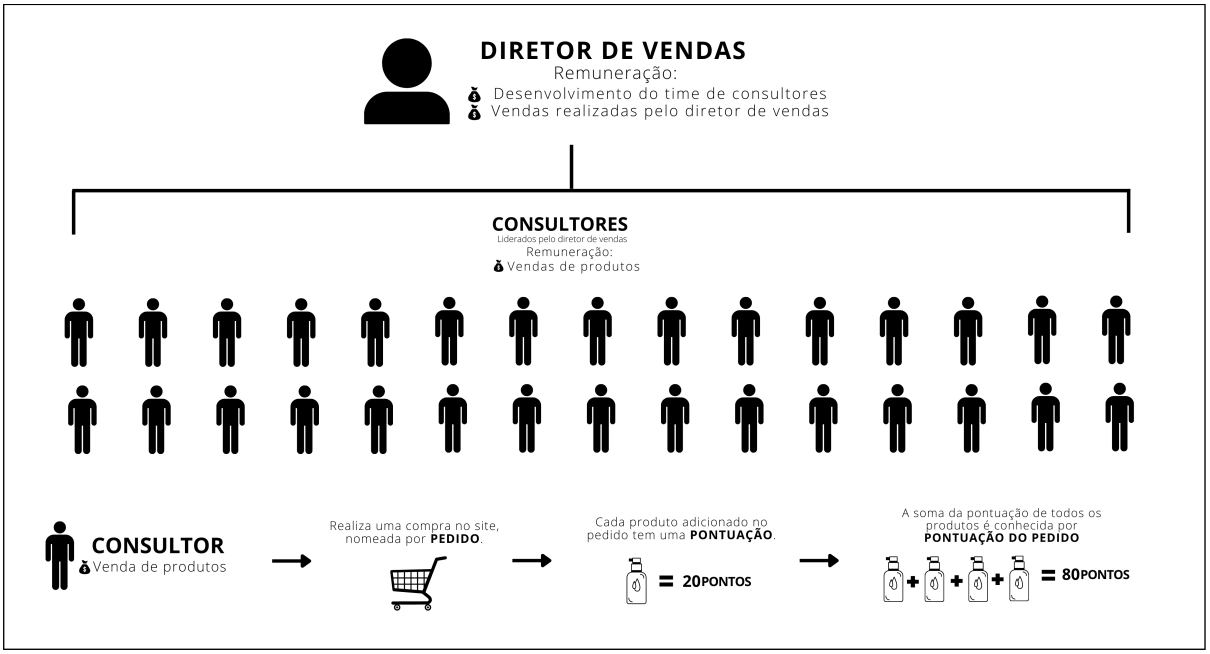
Com o objetivo de vender os produtos, o consultor de beleza, já cadastrado na companhia, compra os produtos com descontos no site e revende os produtos sem o desconto, adquirindo assim uma renda extra ou até mesmo renda fixa, a escolher pelo consultor; a compra é nomeada de pedido. Cada produto tem uma pontuação específica, portanto, cada compra realizada pelo consultor no site gera uma pontuação, que corresponde a soma da pontuação específica dos produtos que o mesmo está adquirindo; é nomeado de pontuação do pedido.

Além do mais, a empresa conta com um plano de carreira meritocrático, no qual todas as pessoas que entram na empresa estão no mesmo nível e têm as mesmas condições e oportunidades para se desenvolver na carreira, bastando apenas realizar os requisitos impostos pela companhia e por consequência sobem de nível.

O próximo nível de carreira é identificado como diretor. O diretor por sua vez também realiza vendas, mas além disso é responsável por liderar um time de consultores que ele desenvolveu, com no mínimo 30 consultores, portanto, seu objetivo é desenvolver

o time de forma com que o time cresça e em contrapartida ele é remunerada de acordo com o desempenho dos consultores presentes no time. Pode ser verificado na Figura 1 a estrutura empresarial.

Figura 1 – Estrutura empresarial.



Fonte: Elaborada pela autora (2022).

Portanto, é essencial que o líder entenda o seu time e compreenda quais são as características que melhor caracterizam o perfil mais adequado para os consultores, para assim conseguir realizar estratégias a fim de aumentar esforços em relação aos consultores que estão mais propensas a não se desenvolverem, além de pautar decisões para melhorar o desempenho dos consultores que apresentam um bom desenvolvimento.

3 Regressão Logística

A modelagem de previsão se refere à tarefa de construir um modelo para a variável dependente em função das variáveis independentes. Conhecida desde os anos 50, a regressão logística é considerada umas das técnicas estatísticas utilizada para a realização de modelos preditivos quando a variável resposta é dicotômica ou politômica, sendo bastante utilizada no mercado de trabalho e com aplicações em diversas áreas.

A regressão logística é uma técnica estatística que tem como objetivo modelar, através de um conjunto de observações, a relação "logística" entre uma variável resposta dicotômica e uma série de variáveis explicativas numéricas e/ou categóricas (CABRAL, 2013). Segundo Batista (2015), a Regressão Logística tem por finalidade a predição de valores tomados por uma variável categórica binária, a partir de um conjunto de variáveis independentes.

Paula (2013) afirma que mesmo quando a resposta de interesse não é originalmente do tipo binário, alguns pesquisadores dicotomizam a mesma, a fim de que a probabilidade de sucesso seja ajustada através da regressão logística.

Assim como em qualquer análise utilizando outras técnicas de modelagem, no presente trabalho o objetivo é obter um modelo que descreva de forma razoável a relação entre a variável de interesse (consultor ativo) e o conjunto de dados presente utilizando a regressão logística, que será retratada neste capítulo desde sua história até sua interpretação.

3.1 História

Segundo Cramer (2002) a regressão logística foi introduzida no século XIX com o objetivo de descrever o crescimento das populações e o curso das reações químicas autocatalíticas.

Contudo, considerando uma quantidade $W(t)$, no qual t é o tempo, tem-se que a variação é dada pela primeira derivada:

$$W'(t) = \frac{dW(t)}{dt}, \tag{3.1}$$

supondo que $W'(t)$ é proporcional a $W(t)$, tem-se:

$$W'(t) = \beta W(t), \tag{3.2}$$

no qual β é a taxa de crescimento constante, logo, leva a um crescimento exponencial:

$$W(t) = Ae^{\beta t}. \tag{3.3}$$

Segundo Cramer (2002), este é um modelo razoável para o crescimento populacional em países jovens. Alphonse Quetelet (1795-1874) estava ciente que a extrapolação do crescimento exponencial implica valores impossíveis, logo, pediu a seu aluno Pierre-François Verhulst (1804-1849), o qual adicionou um termo extra para (equação 3.2):

$$W'(t) = \beta W(t) - \phi(W(t)), \tag{3.4}$$

a fim de representar a crescente resistência ao crescimento adicional. A função logística aparece quando essa é uma quadrática, logo, pode-se reescrever (equação 3.4) como:

$$W'(t) = \beta W(t)(\omega - W(t)), \tag{3.5}$$

no qual ω é o limite superior de W. Se expressar W(t) como uma proporção $P(t) = \frac{W(t)}{\omega}$, tem-se

$$P(t) = \beta P(t)(1 - P(t)), \tag{3.6}$$

e a solução desta equação diferencial é dada por

$$P(t) = \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}}. \tag{3.7}$$

Verhulst (1804-1849) chamou essa função de logística. Portanto, a população $W(t)$ segue:

$$W(t) = \omega \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}}. \tag{3.8}$$

.

Segundo Cramer (2002), Verhulst publicou seus trabalhos entre 1838 e 1847. Ele modelou a curva do crescimento das populações de países como França, Bélgica e Rússia, a curva foi bem ajustada e os parâmetros α, β, ω foram estimados e determinado três pontos em que a curva deveria passar.

Conforme o autor,

> A função logística foi descoberta novamente em 1920 por Pearl e Reed em um estudo sobre o crescimento populacional do Estados Unidos. Mesmo desconhecendo os resultados de Verhulst, chegaram independentemente

à curva logística. Em 1925, Udny Yule (1871-1951) citou Reed e Verhulst ao nomear a função como logística. (CRAMER, 2002).

3.2 Modelo de Regressão Logística

Seja Y uma variável aleatória, no qual

$$Y = \begin{cases} 1, \ evento\ de\ interesse \\ 0, \ caso\ contrário \end{cases},$$

cada Y terá distribuição de Bernoulli, com função de probabilidade dada por:

$$P(y|\theta) = \theta^y(1-\theta)^{1-y}, \quad y = 0, 1, \tag{3.9}$$

sendo:

y = Evento ocorrido

θ = Probabilidade de interesse para a ocorrência do evento

Como tem-se uma sequência dos eventos ocorridos, a soma dos valores observados terão uma distribuição Binomial com n (observações) e θ (probabilidade de sucesso). Cuja função de probabilidade Binomial será

$$P(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{1-y}. \tag{3.10}$$

A transformação logística será o logaritmo da razão de probabilidades. A função de ligação do modelo de regressão logística simples é dada por

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x, \tag{3.11}$$

e para o caso múltiplo será:

$$g(\mathbf{x}) = ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1, + ... + \beta_p x_p. \tag{3.12}$$

A probabilidade do evento Y ocorrer dado x, é dada por:

$$\pi(x) = E(Y|x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \tag{3.13}$$

e a probabilidade do evento Y ocorrer dado o vetor $\mathbf{x}$, será:

$$\pi(\mathbf{x}) = E(Y|\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}. \tag{3.14}$$

O Apêndice A foi construído a fim de apresentar as demonstrações das equações utilizadas. A demonstração para a equação 3.14 é apresentada no Apêndice A.1.

O princípio da máxima verossimilhança afirma que se usa como estimativa de β o valor que maximiza a expressão:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}, \tag{3.15}$$

em que n representa o número de indivíduos na amostra.

No entanto, é mais fácil trabalhar matematicamente com o *log* da equação, definida como:

$$l(\beta) = ln[l(\beta)] = \sum_{i=1}^{n}\{\ y_i ln[\pi(x_i)] + (1-y_i)ln[1-\pi(x_i)]\}. \tag{3.16}$$

Para encontrar o vetor $\boldsymbol{\beta}$ que maximiza $L(\boldsymbol{\beta})$, diferencia-se $L(\boldsymbol{\beta})$ em relação a β_0 e β_1 e define-se as expressões resultantes iguais a zero. Estas equações são respectivamente:

$$\sum [y_i - \pi(x_i)] = 0 \tag{3.17}$$

e

$$\sum x_i[y_i - \pi(x_i)] = 0. \tag{3.18}$$

A demonstração para as equações 3.17 e 3.18 é apresentada no Apêndice A.2.

Sabe-se que essas equações são não-lineares, e para solucioná-las é necessário utilizar métodos numéricos iterativos, como Newton-Raphson (Gourieroux e Monfort, 1995). Para isso, expande-se a função U($\boldsymbol{\beta}$) em torno do ponto inicial $\boldsymbol{\beta}^{(0)}$, de modo que

$$U(\boldsymbol{\beta}) \approx U(\boldsymbol{\beta}^{(0)}) + U'(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}), \tag{3.19}$$

em que U($\boldsymbol{\beta}$) são as derivadas de primeira ordem e $U'(\boldsymbol{\beta})$ são as derivadas de ordem 2 do logaritmo da função de verossimilhança. A repetição do processo (3.19) resulta no seguinte processo iterativo:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + [-U'(\boldsymbol{\beta}^{(k)})]^{-1}U'(\boldsymbol{\beta}^{(n)}), \quad k = 0, 1, \ldots \tag{3.20}$$

Como a matriz $-U'(\boldsymbol{\beta})$ pode não ser positiva definida, não admitindo sua matriz inversa, então, substitui-se pela matriz de informação observada de Fisher.

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + [\text{-I}(\boldsymbol{\beta}^{(k)})^{-1}]U'(\boldsymbol{\beta}^{(k)}), \quad k = 0, 1, \ldots \tag{3.21}$$

A matriz de informação de Fisher com uma variável, para o modelo logístico pode ser escrita como:

$$I(\widehat{\boldsymbol{\beta}}) = -\begin{bmatrix} \frac{\partial^2}{\partial\beta_0^2} \ \ln \ Ł(\beta_0, \beta_1) & \frac{\partial^2}{\partial\beta_0\beta_1} \ \ln \ Ł(\beta_0, \beta_1) \\ \frac{\partial^2}{\partial\beta_0\beta_1} \ \ln \ Ł(\beta_0, \beta_1) & \frac{\partial^2}{\partial\beta_1^2} \ \ln \ Ł(\beta_0, \beta_1) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{k} n_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{k} n_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \\ \sum_{i=1}^{k} n_i x_i \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} & \sum_{i=1}^{k} n_i x_i^2 \frac{e^{\beta_0+\beta_1 x_i}}{(1+e^{\beta_0+\beta_1 x_i})^2} \end{bmatrix} \tag{3.22}$$

Com as estimativas dos parâmetros do modelo é possível calcular as probabilidades estimadas a partir da expressão (3.14).

3.3 Teste da Razão de Verossimilhanças

Na regressão logística, uma das etapas principais é verificar se as variáveis na base de dados são significativas para explicar a variável resposta. Essa verificação é baseada no *log* da verossimilhança. Portanto, o teste de hipótese a ser realizado é:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}, \quad j = 1, 2, ..., \tag{3.23}$$

a estatística do teste será dada por

$$G = -2ln\left[\frac{(verossimilhança\ sem\ as\ variáveis)}{(verossimilhança\ com\ as\ variáveis)}\right], \tag{3.24}$$

ou ainda,

$$G = -2ln(L_s) + 2ln(L_c), \tag{3.25}$$

em que L_s é a verossimilhança do modelo sem a covariável e L_c é a verossimilhança do modelo com a covariável.

O teste funciona da seguinte forma:

i. Calcula-se a diferença entre o valor da base (sem as variáveis explicativas) e o valor do modelo final (com as variáveis explicativas).

ii. Multiplica-se esse valor da diferença por -2, resultando em um valor de χ^2, com graus de liberdade igual ao número de parâmetros β do modelo completo - não incluindo o intercepto β_0.

iii. O valor resultante é comparado com os valores da tabela da distribuição χ^2.

iv. Se o valor do χ^2 calculado for maior que o valor do χ^2 tabelado, para o determinado nível de significância, então rejeita-se H_0.

3.4 Teste Z-Wald

Também usado para determinação da significância dos coeficientes logísticos, o teste Wald é mais uma ferramenta para auxiliar na avaliação do modelo.

Para calcular a estatística do teste Wald, divide-se o coeficiente estimado $\widehat{\beta}_j$ pelo seu desvio padrão (DP), esse resultado terá uma distribuição aproximadamente normal padrão

$$Wald = \frac{\widehat{\beta}_j}{\widehat{DP}(\widehat{\beta}_j)} \sim N(0,1), \tag{3.26}$$

pois o estimador de máximo verossimilhança de $\widehat{\beta}_j \sim N(\beta_j, Var(\beta_j))$. Então, elevando-se ao quadrado essa estatística Wald, tem-se uma aproximação a qui-quadrado com 1 grau de liberdade

$$Wald = \left(\frac{\widehat{\beta}_j}{\widehat{DP}(\widehat{\beta}_j)} \right)^2 \sim \chi^2_1, \tag{3.27}$$

e as hipóteses são:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases},$$

dessa forma, quando $\chi^2_{(Calculado)} > \chi^2_{(Tabelado)}$, rejeita-se H_0 ao nível de significância α.

3.5 Intervalos de Confiança

Nesta seção, serão apresentados os intervalos de confiança que utilizamos para analisar o modelo de regressão logística.

3.5.1 Intervalo de Confiança para os Parâmetros

Os intervalos de confianças são baseados nos testes de Wald. O intervalo de confiança de $100(1-\alpha)\%$ para o parâmetro β_j (j=1,2,..p) é:

$$IC(\beta_j, 1-\alpha) = [\hat{\beta}_j - z_{1-\alpha/2} DP(\hat{\beta}_j); \quad \hat{\beta}_j + z_{1-\alpha/2} DP(\hat{\beta}_j)], \tag{3.28}$$

em que $z_{1-\alpha/2}$ é a ordenada da distribuição normal padrão correspondente a $100(1-\alpha/2)$

3.5.2 Intervalo de Confiança para os Valores Ajustados

Com o estimador da função de ligação e seu intervalo de confiança, pode-se estimar os valores ajustados e calcular os respectivos intervalos.

Temos que, a variância é calculada pelo método delta, logo $var(\hat{g}(x)) = var(\hat{\beta}_0) + x^2 var(\hat{\beta}_1) + 2cov(\beta_0, \beta_1 x)$ desde que $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.

Portanto, o intervalo de confiança para os valores ajustados pode ser dado por:

$$IC(\pi, 1-\alpha) = \left[\frac{e^{\hat{g}(x) - z_{1-\alpha/2} DP(\hat{g}(x))}}{1 + e^{\hat{g}(x) - z_{1-\alpha/2} DP(\hat{g}(x))}}; \quad \frac{e^{\hat{g}(x) + z_{1-\alpha/2} DP(\hat{g}(x))}}{1 + e^{\hat{g}(x) + z_{1-\alpha/2} DP(\hat{g}(x))}} \right]. \tag{3.29}$$

3.5.3 Intervalo de Confiança para a Razão das Chances (*Odds Ratio*)

Considere os limites do intervalo de confiança para β_j

$$L.I. = \hat{\beta}_j - z_{1-\alpha/2} DP(\hat{\beta}_j) \tag{3.30}$$

e

$$L.S. = \hat{\beta}_j + z_{1-\alpha/2} DP(\hat{\beta}_j), \tag{3.31}$$

intervalo de confiança para a razão das chances será

$$IC(RC, 1-\alpha) = [e^{\beta_I} \ ; \ \ e^{\beta_S}]. \tag{3.32}$$

3.6 Diagnóstico do Modelo

Em qualquer modelo de regressão, é necessário proceder à análise dos resíduos para validação da qualidade do modelo estimado. Assim, pretende-se avaliar quais as "distâncias" entre os valores observados e os valores estimados.

3.6.1 Resíduos de Pearson

Segundo Hosmer e Lemeshow (2003), na regressão logística existem várias maneiras possíveis de medir a diferença entre os valores observados e os valores preditos, uma das formas é calcular os valores ajustados para cada combinação de níveis diferentes das variáveis explicativas, denominada de padrão de covariável.

Então, sendo $\mathbf{X} = (X_1, X_2, ..., X_p)$ o conjunto de variáveis explicativas do modelo, n o número de valores que x assumirá. O número de indivíduos na amostra com valores iguais $\mathbf{x} = \mathbf{x}_j$ é denotado por: m_j. O número de indivíduos y=1 será denotado por y_j em m_j. probabilidade ajustada dos indivíduos em j será $\hat{\pi}_j$. Assim tem-se

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(\mathbf{x}_j)}}{1 + e^{\hat{g}(\mathbf{x}_j)}}, \tag{3.33}$$

em que $\hat{g}(\mathbf{x}_j) = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \hat{\beta}_2 x_{j2} + ... + \hat{\beta}_p x_{jp}$ é o *logit* estimado.

A medida de Pearson para a diferença entre o observado e predito é

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}, \tag{3.34}$$

e assim, a estatística χ^2 de Pearson será:

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2. \tag{3.35}$$

3.6.2 Deviance

Uma das análises realizadas na regressão linear é a ANOVA, onde obtêm-se soma dos quadrados dos resíduos. Já no modelo logístico é denominada *deviance*. (HOSMER, LEMESHOW e STURDIVANT, 2013).

Podendo ser definido da seguinte forma:

$$d(y_j, \widehat{\pi}_j) = \pm \left[2 \left[y_j \ \ln\left(\frac{y_j}{m_j\widehat{\pi}_j}\right) + (m_j - y_j) \ \ln\left(\frac{(m_j - y_j)}{m_j(1-\widehat{\pi}_j)}\right)\right]\right]^{1/2}. \tag{3.36}$$

Os sinais + ou - é o mesmo sinal que $(y_j - m_j\hat{\pi}_j)$. Para covariável padrão com $y_j = 0$ o *deviance* residual é

$$d(y_j, \widehat{\pi}_j) = -\sqrt{2m_j|ln(1-\widehat{\pi}_j)|}, \tag{3.37}$$

e o *deviance* residual quando $y_j = m_j$ é:

$$d(y_j, \widehat{\pi}_j) = -\sqrt{2m_j|ln(\widehat{\pi}_j)|}, \tag{3.38}$$

a estatística do teste baseada no *deviance* residual será:

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2. \tag{3.39}$$

A distribuição das estatísticas de χ^2 e D sob a suposição de que o modelo ajustado está correto em todos os aspectos deve ser χ^2 com $J - (p + 1)$ graus de liberdade.

3.6.3 Teste de Hosmer e Lemeshow

Segundo Paula (2003) o teste de Hosmer e Lemeshow sugerem uma estatística alternativa para avaliação da qualidade do ajuste. Essa estatística é definida comparando o número observado com o número esperado de sucessos de g grupos formados. O primeiro grupo deverá conter n'_1 elementos correspondentes às n'_1 menores probabilidades ajustadas, as quais serão denotadas por $\hat{\pi}_1 \leq \hat{\pi}_2 \leq ... \leq \hat{\pi}_{n'_1}$. O segundo grupo deverá conter os n'_2 elementos correspondentes às seguintes probabilidades ajustadas $\hat{\pi}_{n'_1+1} \leq \hat{\pi}_{n'_1+2} \leq ... \leq \hat{\pi}_{n'_1+n'_2}$. E assim, sucessivamente, até o último grupo que deverá conter as n'_g maiores probabilidades ajustadas $\hat{\pi}_{n'_1+n'_{g-1}+1} \leq \hat{\pi}_{n'_1+n'_{g-1}+2} \leq ... \leq \hat{\pi}_n$.

O número observado de sucessos no primeiro grupo formado será dados por $O_1 = \sum_{k=1}^{n'_1} y_{(j)}$, em que $y_{(j)} = 0$ se o elemento correspondente é fracasso e $y_{(j)} = 1$ se é sucesso. Generalizando, obtemos $O_k = \sum_{j=n'_1+...+n'_{k-1}+1}^{n'_1+...+n'_{k-1}} y_{(j)} y_j$, $2 \leq i \leq g$.

A estatística do teste é dada pela seguinte forma:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}, \tag{3.40}$$

em que

$$\bar{\pi}_1 = \frac{1}{n'_1} \sum_{j=1}^{n'1} \hat{\pi}_j \tag{3.41}$$

e

$$\bar{\pi}_k = \frac{1}{n'_k} \sum_{j=n'_1+...+n'_{k-1}+1}^{n'_k+...+n'_k} \hat{\pi}_j, \tag{3.42}$$

•n_k é o número de indivíduos no k-ésimo grupo.

•C_k: o número total de combinações de níveis.

Hosmer, Lemeshow e Sturdivant (2013) verificaram através de simulações que a distribuição nula assintótica de pode ser bem aproximada por uma distribuição qui-quadrado com $(g-2)$ graus de liberdade. Portanto, a estatística do teste segue aproximadamente uma distribuição χ^2 com $(g-2)$ graus de liberdade.

3.6.4 Matriz de Confusão

Segundo Silvestre (2021), a matriz de confusão mostra os valores atuais dos membros preditos dos grupos, e a partir dela a razão do erro aparente é facilmente calculada. Para a observações n_1 provenientes da população π_1 e n_2 observações da população π_2, a matriz de confusão tem a seguinte forma:

Figura 2 – Matriz de confusão.

		Membro predito		
		π_1	π_2	
Membro Atual	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Fonte: Silvestre (2022, p.78).

tal que:

n_{1C} = número de itens de π_1 corretamente classificados como itens de π_1.

n_{1M} = número de itens de π_1 incorretamente classificados como itens de π_2.

n_{2C} = número de itens de π_2 corretamente classificados como itens de π_2.

n_{2M} = número de itens de π_2 incorretamente classificados como itens de π_1.

Logo, a razão do erro aparente é definida como:

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \tag{3.43}$$

e representa a proporção de itens que foram mal classificados.

A partir da matriz de confusão, é possível calcular a sensibilidade:

$$Sensibilidade = \frac{n_{1C}}{(n_{1C} + n_{2M})} \tag{3.44}$$

enquanto a especificidade é calculada da seguinte forma:

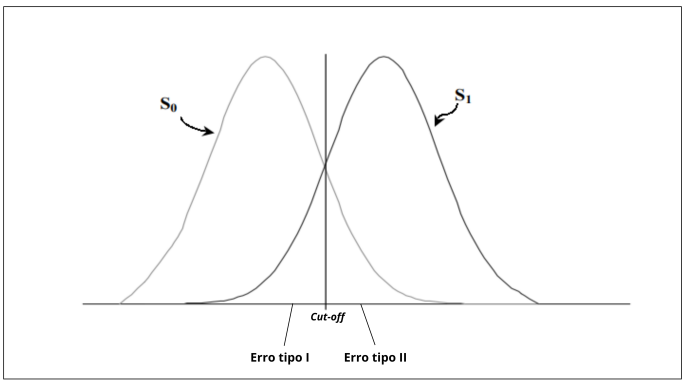
$$Especificidade = \frac{n_{2C}}{(n_{1M} + n_{2C})}. \tag{3.45}$$

3.6.5 Cut-off

Cut-off (ponto de corte) é o valor que se usa para classificar o valor predito em uma ou outra população em relação ao ajuste do modelo. Em termos teóricos,adota-se o ponto de corte igual a 0,5.

Utilizando o ponto de corte que maximiza a taxa de acertos, nunca será possível acertar 100% dos resultados, como pode ser visto na imagem a seguir:

Figura 3 – Erro tipo I (Negativo Falso) e Erro tipo II (Positivo Falso).



Fonte: Braga (2000, p.10).

Erro tipo 1: Classificar o indivíduo como S_0 sendo que ele é pertencente a S_1.

Erro tipo 2: Classificar o indivíduo como S_1 sendo que ele é pertencente a S_0.

O objetivo de achar um cut-off ideal, é de equilibrar os erros dentro dos grupos, para isso, analisa-se as porcentagens da sensibilidade (Verdadeiro positivo) e especificidade (Verdadeiro negativo), quando elas tiverem resultados semelhantes atribui-se o ponto de corte na probabilidade que foi usada para construir a matriz confusão.

3.6.6 *Curva Receiver Operating Characteristic (ROC)*

Segundo Hosmer e Lemeshow (2013), uma descrição completa da precisão da classificação é a área sobre a curva Receiver Operating Characteristic (ROC). Esta curva, originária da teoria de detecção de sinal, mostra como o receptor detecta a existência de sinal na presença de ruído. Ela traça a probabilidade de detectar sinal verdadeiro ($sensibilidade$) e sinal falso ($1 - especificidade$) para toda uma gama de possíveis pontos de corte.

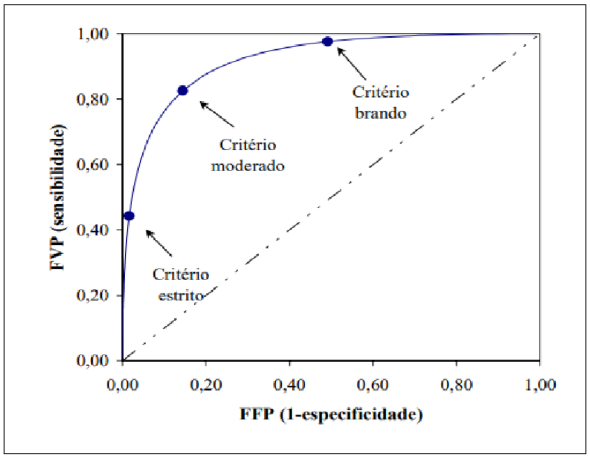
A curva ROC permite a interpretação da classificação 0 ou 1 conforme as variáveis explicativas. Sabe-se que o modelo nunca acertará 100% dos casos, sempre haverá situações em que o modelo previu 1 dado que o verdadeiro valor é 0, e previu 0 dado que era 1.

O valor da área sob a curva (AUC) fornece a capacidade preditiva do modelo classificar corretamente as observações, sendo os seguintes intervalos a serem considerados:

$$\begin{cases} ROC = 0,5 \;\; Não \;\; discrimina \\ 0,5 \leq ROC < 0,7 \;\; Baixa \\ 0,7 \leq ROC < 0,8 \;\; Aceitável \\ 0,8 \leq ROC < 0,9 \;\; Muito \;\; bom \\ ROC \geq 0,9 \;\; Excelente \end{cases} \tag{3.46}$$

Assim, pode-se definir, um critério "estrito"como sendo aquele que conduz a uma pequena fração de falsos positivos e também a uma relativamente pequena fração de verdadeiros positivos, ou seja, gera um ponto na curva ROC que se situa no canto inferior esquerdo do espaço ROC. Progressivamente critérios menos estritos conduzem a maiores frações de ambos os tipos, ou seja, pontos colocados no canto superior direito da curva no espaço ROC. Esta situação pode ser descrita graficamente pela curva ROC apresentada na Figura 4, no qual têm-se no eixo das ordenadas $sensibilidade$ ou FVP (Fração Verdadeiro Positivo) e no eixo das abcissas $1 - especificidade$ ou FFP (Fração Falso Positivo).

Figura 4 – Curva ROC, para uma dada capacidade de discriminação, com a variação do critério de decisão.



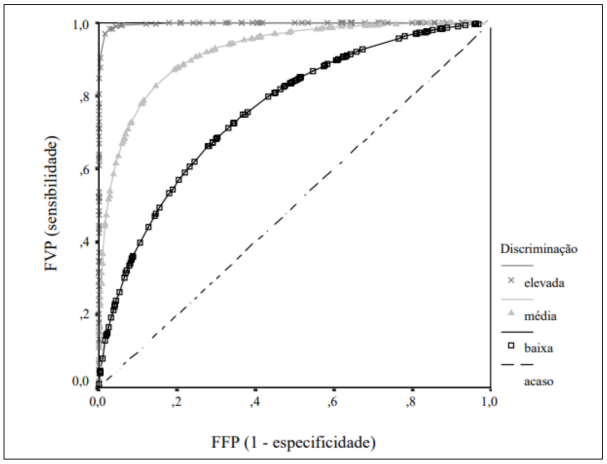
Fonte: Braga (2000, p.30).

Na Figura 5, apresentam-se três graus de discriminação possíveis fornecidos pelas curvas ROC. Quando as curvas ROC se cruzam então podem-se classificar os sistemas para um conjunto de frações de falsos positivos ou verdadeiros positivos de interesse no sentido do diagnóstico, tendo em conta os custos e benefícios de um diagnóstico alternativo, se necessário.

Segundo a autora:

> No que diz respeito ao desempenho de diferentes sistemas de diagnóstico, e considerando a situação em que as curvas ROC associadas a dois sistemas de diagnóstico distintos não se cruzam, o sistema com a curva ROC mais próxima do canto superior esquerdo, fornece um maior poder discriminante. (BRAGA, 2000, p.30).

Figura 5 – Curvas ROC representativas de três graus de capacidade de discriminação.



Fonte: Braga (2000, p.31).

4 Aplicação dos Dados

Neste trabalho os resultados apresentados serão referentes a uma base de dados cedido por uma diretora de vendas da companhia Mary Kay. O modelo a ser desenvolvido neste capítulo será baseado na técnica de Regressão Logística utilizando o software estatístico R e de consultores, passando de um estilo comum, sem nenhum conhecimento do perfil dos consultores, para um desenvolvimento estratégico e baseado em dados.

Mais especificamente, os dados são referentes a uma base de consultores que fazem parte do time da diretora de vendas da companhia, observados de março de 2021 até janeiro de 2022. O objetivo principal é segmentar as principais características que aumentam a probabilidade do consultor ser ativo.

Assim sendo, com o auxílio do modelo a ser desenvolvido, a diretora de vendas diminuirá custos pois poderá investir maior tempo e estratégias para os consultores que tem menores probabilidades de se tornarem ativos, aumentando assim o desenvolvimento do seu time, e como consequência, a sua remuneração também.

4.1 Variáveis

A base de dados utilizada apresenta 70 linhas, ou seja, informações sobre 70 consultoras de beleza. As variáveis são apresentadas no Apêndice B.1, assim como suas categorias e descrição.

4.2 Aplicação de Questionário

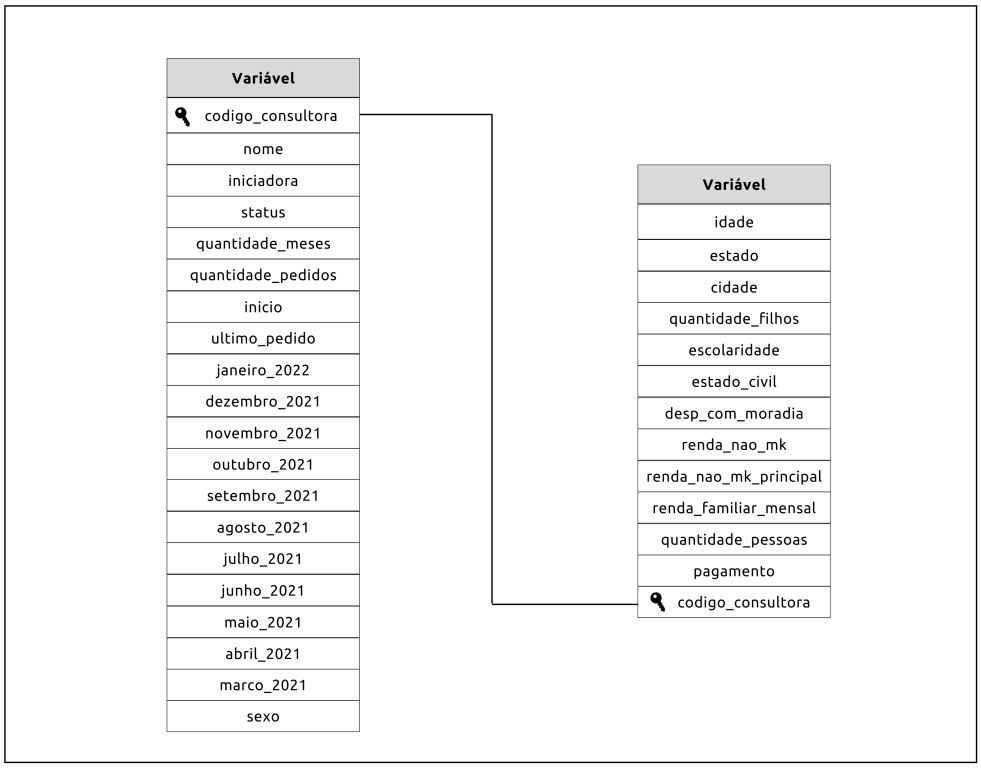
O público alvo deste trabalho são consultores que pertencem ao time da diretora de vendas da companhia Mary kay que disponibilizou os dados. Com a finalidade de aumentar a quantidade de variáveis, aplicou-se um questionário de pesquisa durante os dias 04 de abril de 2022 até o dia 18 de abril de 2022. Foi utilizado o Google Formulários para a aplicação do formulário; o mesmo foi encaminhado via Whatsapp para todos os consultores que estavam presentes na base de dados, com uma breve explicação referente ao estudo, apresentando aos convidados a programação sobre o período de coleta dos dados e a importância da participação de cada participante no estudo.

Após o preenchimento do formulário, foi construído a base de dados coletados utilizando o Excel. Cada consultor preencheu o seu código de cadastro antes de responder as questões, variável que inclusive já estava na base de dados, logo, como o código de cadastro é único essa variável é a chave primária, ou seja, a partir dela as tabelas conse-

guem se relacionar. Obteve-se um total de 70 respostas. O quadro com todas as variáveis coletadas, suas respectivas categoria e descrição estão presentes no Apêndice B.2.

Na Figura 6, podemos verificar que as tabelas se unem a partir do *codigo_consultora* que é único para cada consultor(a).

Figura 6 – Relação entre as tabelas.



Fonte: Elaborada pela autora (2022).

4.3 Construção de Variáveis

A partir de todas as variáveis que têm-se na tabela final (junção da tabela inicial com a tabela dos dados coletados), construiu-se novas variáveis que podem serão utilizadas na construção do modelo.

O Quadro 1 apresenta as variáveis, fórmula das variáveis, descrição da fórmula e o conceito.

Quadro 1 - Variáveis construídas.

Variável	**Fórmula**	**Descrição da Fórmula**	**Conceito**
Pontuação Total	$\sum_{i=1}^{n} p_i$, sendo n a quantidade de pedidos realizados pelo consultor(a)	Soma da pontuação de todos os pedidos feitos pela consultor(a) (até janeiro de 2022).	Calcula a pontuação total de cada consultor(a).
Pontuação Total Mínima	X_{min}	Pontuação total mínima (até janeiro de 2022).	Calcula a pontuação total mínima.
Pontuação Total Máxima	$X_{máx}$	Pontuação total máxima (até janeiro de 2022).	Calcula a pontuação total máxima.
Pontuação Total Padronizada	$\frac{Pontuação\ Total - X_{min}}{X_{máx} - Xmín}$	Pontuação total subtraída pela pontuação mínima entre todos os consultores. Dividido pela pontuação máxima entre todos os consultores subtraída pela pontuação mínima entre todos os consultores	Calcula a Pontuação Total Padronizada de cada consultor(a).
Renda Per Capita	$\frac{Renda\ Familiar\ Total}{Quantidade\ de\ Pessoas}$	Renda familiar mensal total dividido pela quantidade de pessoas que contém na família.	Calcula a Renda Per Capita na família de cada consultor(a).
Pontuação média por tempo (Quantidade de meses)	$\frac{\sum_{i=1}^{t} p_i}{t}, t = 1, 2, ..., 11$	Pontuação total dividido pela quantidade de meses que o consultor(a) está na empresa (até janeiro de 2022).	Calcula a média de pontos por mês do(a) consultor(a)

Continua na próxima página

Quadro 1 – *Variáveis Construídas*

Variável	**Fórmula**	**Descrição da Fórmula**	**Conceito**
Pontuação média por pedido	$\frac{\sum_{i=1}^{n} p_i}{n}$, n: quantidade de pedidos realizados pelo consultor(a)	Pontuação Total dividido pela quantidade de pedidos feito pelo(a) consultor(a) (até janeiro de 2022).	Calcula a média de pontos por pedido do(a) consultor(a)
Pontuação Mínima	$p_{mín}$	Pontuação Mínima entre todos os pedidos que o consultor(a) fez desde que iniciou (até janeiro de 2022).	Calcula a pontuação mínima que o consultor(a) já realizou.
Pontuação Máxima	$p_{máx}$	Pontuação Máxima entre todos os pedidos que o consultor(a) fez desde que iniciou (até janeiro de 2022).	Calcula a pontuação mínima que o consultor(a) já realizou.
Desvio Padrão da Pontuação	$\sqrt{\frac{\sum_{i=1}^{t}(p_i - \bar{p})^2}{t}}$	Desvio Padrão da Pontuação dos pedidos feitos pelo(a) consultor(a) (até janeiro de 2022).	Calcula o desvio padrão da pontuação feita pelo(a) consultor(a).
Coeficiente de Variação (Em Relação a Pontuação média por tempo)	$\frac{Desvio\ padrão\ pontuação}{Pontuação\ média\ tempo} * 100$	Desvio padrão da pontuação dividido pela Pontuação média por tempo (até janeiro de 2022).	Calcula o coeficiente de variação que fornece a variação da pontuação em relação a pontuação média (calculada pela quantidade de meses de empresa).
Coeficiente de Variação (Em relação a Pontuação média por pedido)	$\frac{Desvio\ padrão\ pontuação}{Pontuação\ média\ pedido} * 100$	Desvio padrão da pontuação dividido pela Pontuação média por pedido (até janeiro de 2022).	Calcula o coeficiente de variação que fornece a variação da pontuação em relação é pontuação média (calculada pela quantidade de pedidos feitos)

Continua na próxima página

Quadro 1 – *Variáveis Construídas*

Variável	**Fórmula**	**Descrição da Fórmula**	**Conceito**
Tempo de Atividade	$\frac{Quantidade\ de\ pedidos}{Quantidade\ de\ meses}$	Quantidade de pedidos que o(a) consultor(a) fez dividido pela quantidade de meses que o(a) consultor(a) está na empresa (até janeiro de 2022).	Calcula o Tempo de Atividade do(a) Consultor(a), levando em consideração a quantidade de pedidos que o mesmo realizou.
Coeficiente de Atividade	$\frac{Pontuação\ total\ padronizada}{Quantidade\ de\ meses}$	Pontuação Total Padronizada dividido pela quantidade de meses que o(a) consultor(a) está na empresa (até janeiro de 2022).	Calcula o coeficiente de atividade do(a) consultor(a), levando em consideração a pontuação total padronizada.

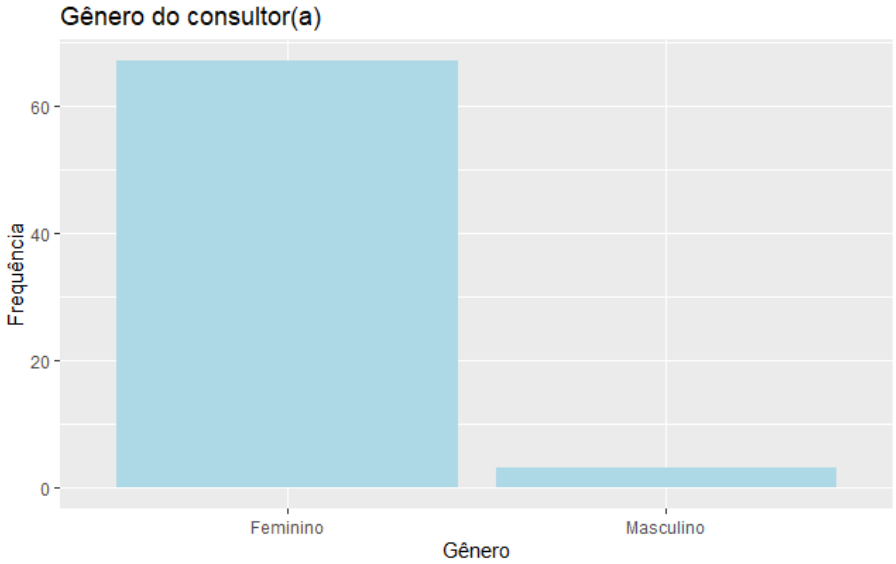
4.4 Análise Exploratória de Dados

A análise exploratória de dados é o processo inicial nas análises estatísticas de conjuntos de dados. Essa fase se faz necessária a fim de conhecer as principais características dos dados.

Tem-se com base nos dados que em relação ao estado dos consultores, dos 70 consultores presentes na pesquisa, o estado que mais reside consultores estudado nesse trabalho é o estado de São Paulo (63 consultores), seguido pelo estado do Paraná (3 consultores), e posteriormente Rio de Janeiro, Maranhão, Mato Grosso do Sul e Minas Gerais com 1 consultor cada.

Com base na Figura 7, conclui-se que entre os 70 consultores presentes no estudo, 67 consultores, ou ainda 96% são do gênero feminino e apenas 3 do gênero masculino.

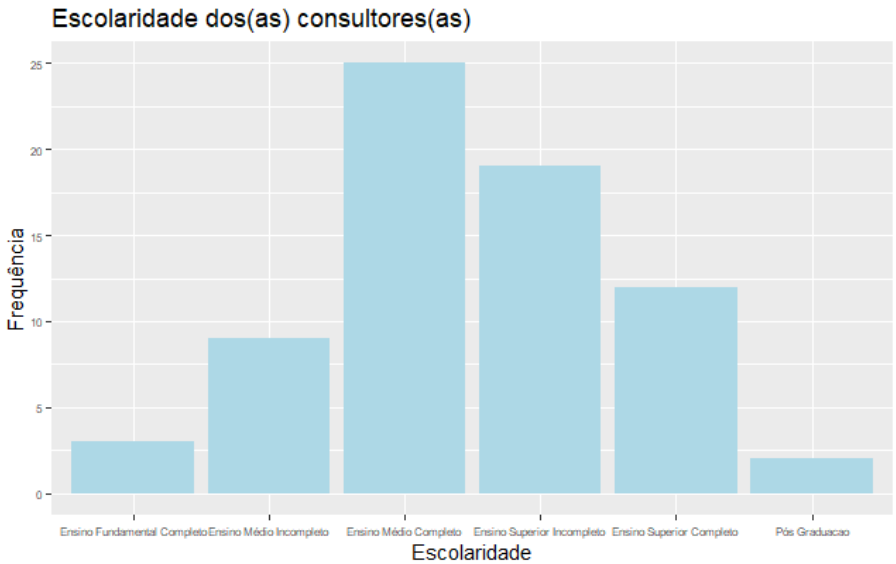
Figura 7 – Gráfico de barras gênero do consultor(a).



Fonte: Elaborada pela autora (2022).

Analisando a Figura 8, podemos perceber que grande parte dos consultores apresentam ensino médio completo (25 consultores - 36%), ensino superior incompleto (19 consultores - 27%), ensino superior completo (12 consultores - 17 %), seguido de ensino médio incompleto (9 consultores - 13%), ensino fundamental completo (2 consultores - 2,5%) e pós graduação (2 consultores - 2,5%).

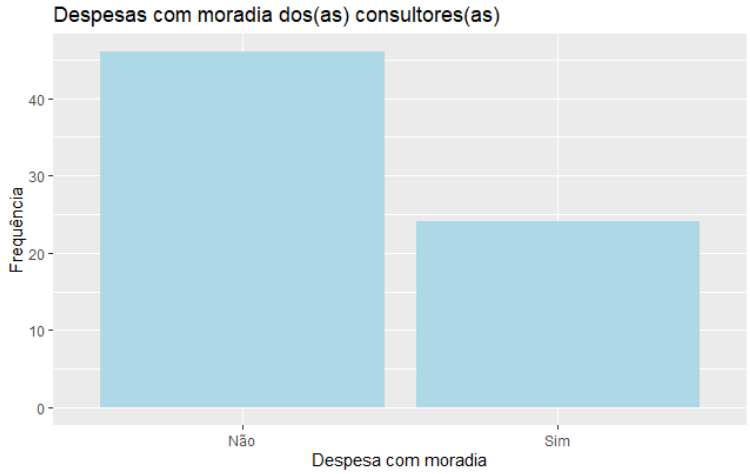
Figura 8 – Gráfico de barras grau de escolar dos(as) consultores(as).



Fonte: Elaborada pela autora (2022).

Analisando o gráfico disposto na Figura 9, despesa com moradia dos consultores, percebemos que 46 consultores, ou ainda 66% não apresentam despesas com moradia.

Figura 9 – Gráfico de barras despesas com moradia dos(as) consultores(as).

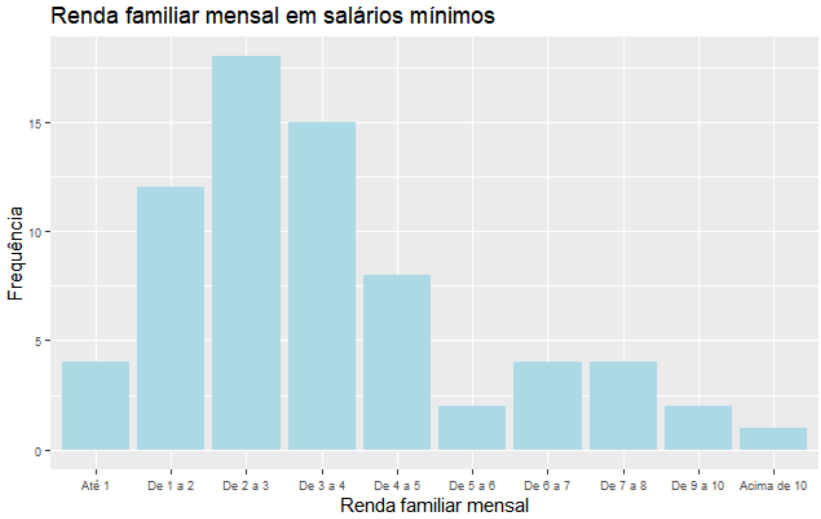


Fonte: Elaborada pela autora (2022).

O gráfico apresentado na Figura 10 apresenta a renda familiar mensal em termos de salários mínimos. No Brasil afirma que o valor do salário mínimo no ano de 2022 é de R$1.212,00. [1]

Assim sendo, analisando o gráfico apresentado na Figura 10, percebemos que a maior parte dos consultores apresentam renda familiar mensal entre 2 e 3 salários mínimos, ou seja entre R$ 2424,00 e R$ 3636,00 (18 consultores).

Figura 10 – Gráfico de barras renda familiar mensal em salários mínimos.

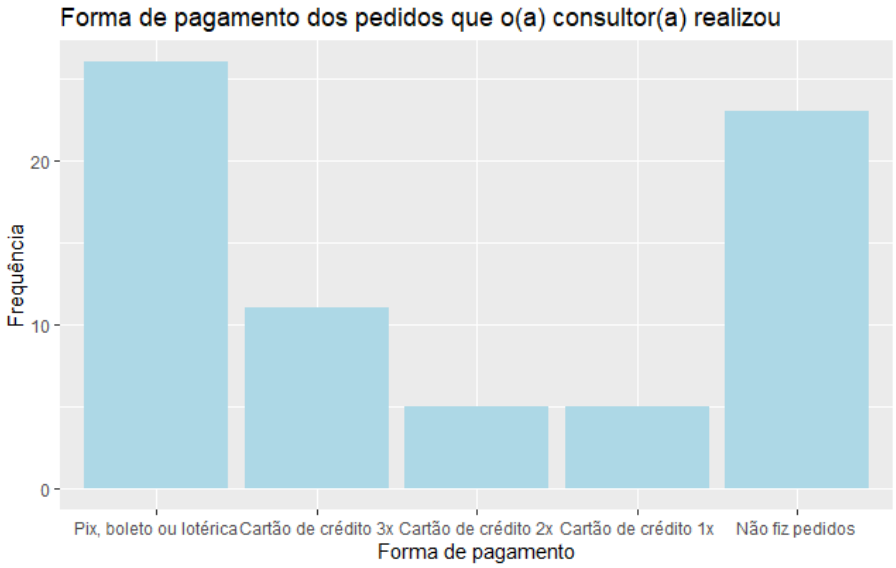


Fonte: Elaborada pela autora (2022).

[1] BRASIL, 2022. Art. 1º A partir de 1º de janeiro de 2022, o salário-mínimo será de R$ 1.212,00 (mil duzentos e doze reais). Art. 2º Esta Lei entra em vigor na data de sua publicação. (BRASIL, 2022)

Analisando a Figura 11, percebemos que o método de pagamento mais utilizado entre os consultores (26 consultores - 37%) é através de pix, boleto ou até mesmo em lotéricas.

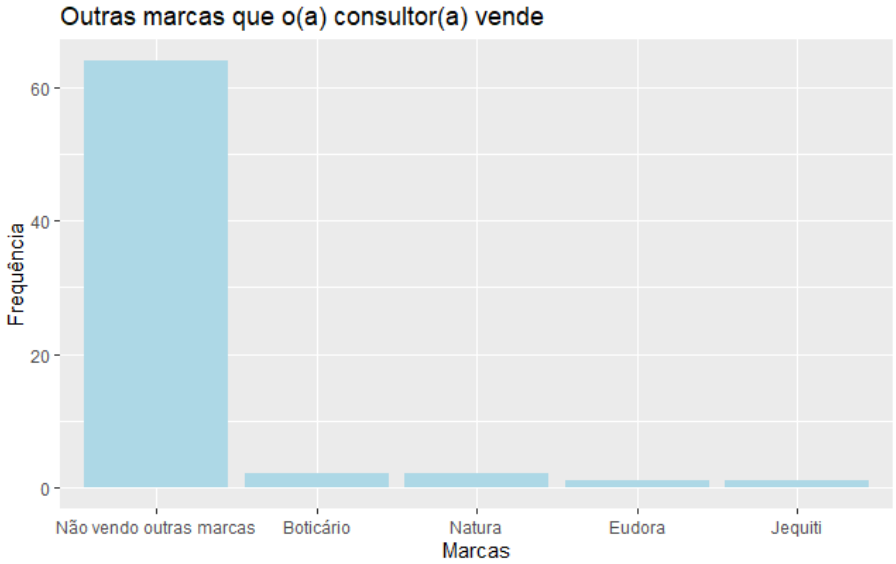
Figura 11 – Gráfico de barras de acordo com o método de pagamento dos pedidos realizado pelos(as) consultores(as).



Fonte: Elaborada pela autora (2022).

A partir da análise do gráfico apresentado na Figura 12, conclui-se que a maior parte, 91% dos consultores, não vendem outras marcas, vendem apenas a marca Mary Kay. Apenas 2 consultores vendem produtos da marca *Boticário*, apenas 2 consultores vendem produtos da marca *Natura*, apenas 1 consultor vede a marca *Jequiti* e apenas 1 consultor vende a marca *Eudora*.

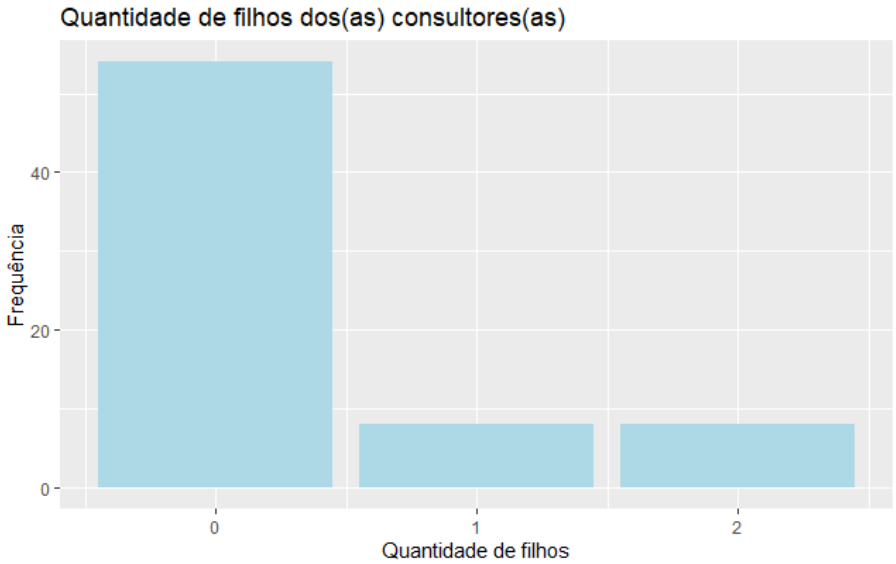
Figura 12 – Gráfico de barras outras marcas que os(as) consultores(as) vendem.



Fonte: Elaborada pela autora (2022).

Analisando o gráfico apresentado na Figura 13, concluímos que a maior parte dos consultores (54 consultores - 77%) não têm filhos.

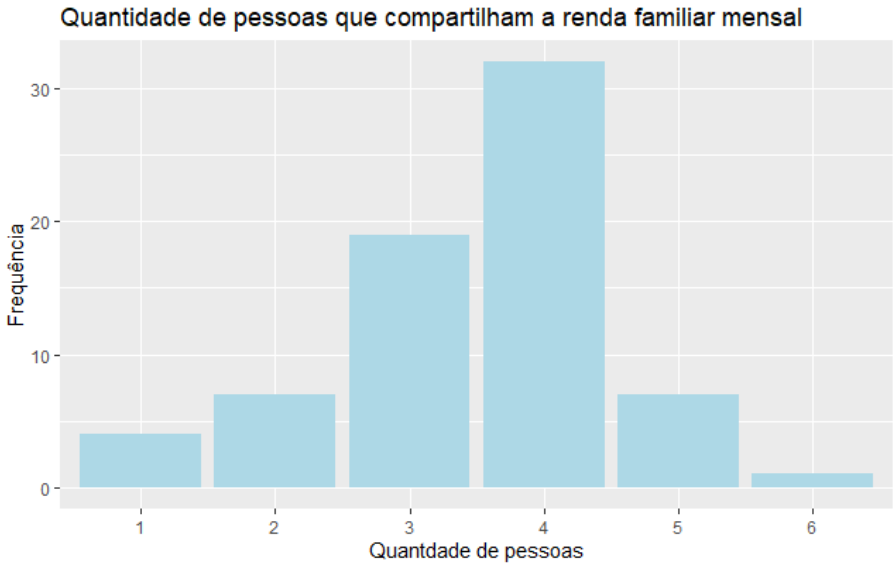
Figura 13 – Gráfico de barras quantidade de filhos dos consultores.



Fonte: Elaborada pela autora (2022).

Analisando o gráfico presente na Figura 14, podemos perceber que a maioria dos consultores (32 consultores - 46%) compartilham a renda familiar mensal com quatro pessoas.

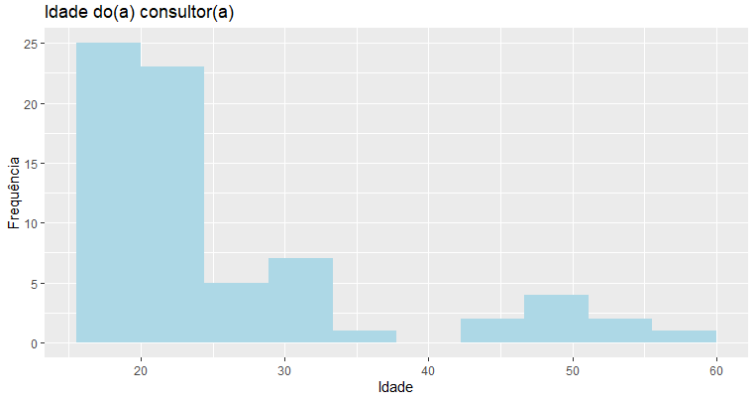
Figura 14 – Gráfico de barras quantidade de pessoas que compartilham da mesma renda familiar mensal que o(a) consultor(a).



Fonte: Elaborada pela autora (2022).

Analisando a idade dos(as) consultores(as), a partir da Figura 15, tem-se que a idade mínima entre os(as) consultores(as) é 16 anos, a média é 25 anos e a idade máxima é de 56 anos.

Figura 15 – Histograma da idade dos(as) consultores(as).



Fonte: Elaborada pela autora (2022).

Além do mais, dentre os 70 consultores presentes no estudo, 49 consultores apresentam renda fixa mensal além da renda Mary Kay, e além disso, de todas os consultores que apresentam renda fixa mensal além da renda Mary Kay, essa é a renda principal mensal.

A Tabela 2 disposta a seguir, apresenta um resumo da análise das variáveis construídas.

Tabela 2 – Resumo análise das variáveis construídas.

Variável	Mínimo	Mediana	Média	Máximo	Desvio Padrão
Pontuação total	0,0000	553,5000	1329,9000	9930,0000	2161,9840
Pontuação total padronizada	0,0000	55,7400	133,9300	1000,0000	217,7225
Pontuação média por tempo	0,0000	59,9600	183,7500	1655,0000	299,7846
Pontuação média por pedido	0,0000	393,4000	412,6000	2216,5000	440,5511
Pontuação mínima	0,0000	223,5000	286,5000	2210,0000	358,1269
Pontuação máxima	0,0000	456,0000	587,2000	2582,0000	672,5654
Desvio padrão da pontuação	0,0000	150,0000	197,8000	946,7000	236,3191
Coeficiente de variação (média por tempo)	0,0000	1,0440	1,1570	3,3170	1,0203
Coeficiente de variação (média por pedido)	0,0000	0,3015	0,2997	0,7278	0,2424
Tempo de atividade	0,0000	0,1429	0,2351	1,0000	0,2642
Coeficiente de atividade	0,0000	6,0380	18,5050	166,6670	30,1898

Fonte: Elaborada pela autora (2022).

Assim sendo, podemos perceber que a pontuação total dos(as) consultores(as) é em média $1329,9$ pontos e no máximo 9930 pontos. Já a pontuação total padronizada é em média $133,93$ pontos e no máximo 1000 pontos.

Além do mais, a pontuação média por tempo, é em média $183,75$ pontos por mês; já a pontuação média por pedido, é em média $412,6$ pontos por pedido, logo, podemos perceber que a pontuação média por pedido é em média maior que a pontuação média por tempo.

Em relação a pontuação mínima do pedido dos(as) consultores(as) é em média $286,5000$ pontos sendo no máximo 2210 pontos, enquanto a pontuação máxima é em média $587,2$ pontos e no máximo 2582 pontos por pedido.

O tempo de atividade dos(as) consultores(as) é em média $0,2351$ e varia até 1, enquanto o coeficiente de atividade é em média $18,5050$ e no máximo $166,6670$.

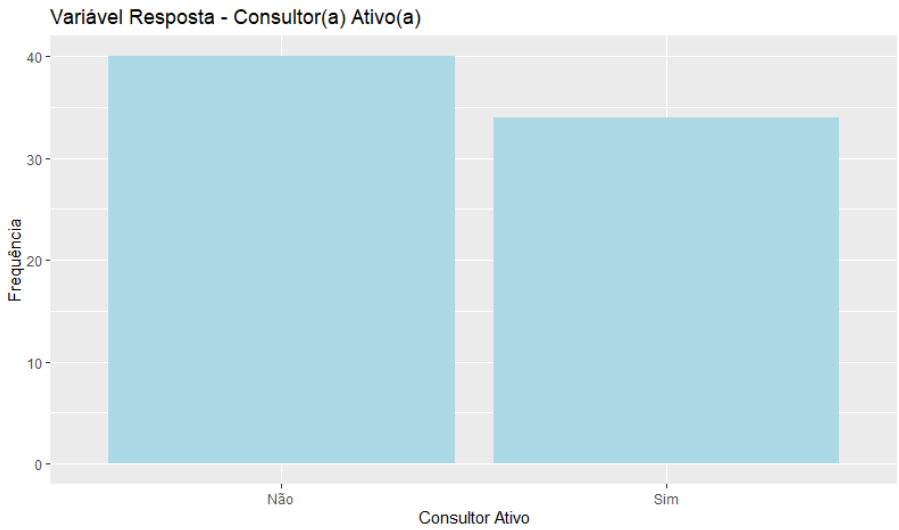
5 Construção dos Modelos

Neste trabalho os resultados apresentados serão referentes ao banco de dados cedido por uma diretora de vendas da companhia estudada, tendo 74 observações. O modelo será construído utilizando regressão logística. Na etapa de préprocessamento, os dados foram tratados no software com interface SQL: DBeaver. Nesse processo foram feitas junções das bases de dados das variáveis coletadas com as variáveis que já estavam presentes. Sarmento (2015), em sua obra sobre Regressão Logística, defende o fato de categorizar as variáveis contínuas, uma vez que esse artifício possa proporcionar um melhor desempenho nos modelos. Para verificar se de fato há eficiência, um estudo comparativo de modelos com variáveis quantitativas originais e variáveis quantitativas categorizadas é apresentado nesta seção.

5.1 Variável Resposta

O objetivo do presente trabalho é prever o consultor(a) de ser ativo, contudo, todos os consultores da base que tinham status A1, A2 e A3 são considerados consultores ativos.

Figura 16 – Gráfico de barras Consultor(a) Ativo(a)



Fonte: Elaborada pela autora (2022).

Na Figura 16 percebe-se que dentre os 74 consultores, 34 são ativos (aproximadamente 46%) e 40 consultores não são ativos.

5.2 Variáveis Categorizadas

O quadro a seguir apresenta as variáveis, a categoria e a descrição das mesmas.

Quadro 2 - Variáveis categorizadas.

Variável	**Categoria**	**Descrição**
Idade	1 a 3	1 = Menor ou igual a 20 anos 2 = Entre 20 e 24 anos 3 = Maior ou igual a 24 anos
Pontuação Total	1 a 3	1 = Menor ou igual a 683 pontos 2 = Entre 683 e 938 pontos 3 = Maior ou igual a 938 pontos
Pontuação Total Padronizada	1 a 3	1 = Menor ou igual a 93 pontos 2 = Entre 93 e 128 pontos 3 = Maior ou igual a 128 pontos
Renda Per Capita	1 a 4	1 = Menor ou igual a R$ 753,00 2 = Entre R$ 753,00 e R$ 999,00 (Inclusive) 3 = Entre R$ 999,00 e R$ 1894,00 4 = Maior ou igual a R$ 1894,00
Pontuação média por tempo (Quantidade de meses)	1 a 4	1 = Menor ou igual a 53 pontos 2 = Entre 53 e 91 (inclusive) pontos 3 = Entre 91 e 333 pontos 4 = Maior ou igual a 333 pontos
Pontuação média por pedido	1 a 4	1 = Menor ou igual a 145 pontos 2 = Entre 145 e 338 (inclusive) pontos 3 = Entre 338 e 655 pontos 4 = Maior ou igual a 655 pontos
Desvio Padrão da Pontuação	1 a 3	1 = Menor ou igual a 244 2 = Entre 244 e 272 3 = Maior ou igual a 272
Coeficiente de Variação (Em Relação a Pontuação média por tempo)	1 a 4	1 = Menor ou igual a 0,51 2 = Entre 0,51 e 0,87 (inclusive) 3 = Entre 0,87 e 1,10 4 = Maior ou igual a 1,10

Continua na próxima página

Quadro 2 – *Variáveis categorizadas*

Variável	**Categoria**	**Descrição**
Coeficiente de Variação (Em relação a Pontuação média por pedido)	1 a 5	1 = Menor ou igual a 1,2 2 = Entre 1,2 e 2,0 (inclusive) 3 = Entre 2,0 e 2,8 (inclusive) 4 = Entre 2,8 e 3,5 5 = Maior ou igual a 3,5
Tempo de Atividade	1 a 4	1 = Menor ou igual a 0,24 2 = Entre 0,24 e 0,35 (inclusive) 3 = Entre 0,35 e 0,61 4 = Maior ou igual a 0,61
Coeficiente de Atividade	1 a 4	1 = Menor ou igual a 7,1 2 = Entre 7,1 e 12,0 (inclusive) 3 = Entre 12,0 e 45,0 4 = Maior ou igual 45,0
Despesa com Moradia	0 e 1	0: Não tem despesa com moradia 1: Tem despesa com moradia
Renda Não Mary Kay	0 e 1	0: Não tem renda diferente da renda Mary Kay 1: Tem renda diferente da renda Mary Kay
Iniciadora	0 e 1	0: Iniciadora não é a diretora 1: Iniciadora é a diretora
Sexo	0 e 1	0: Sexo masculino 1: Sexo feminino

Fonte: Elaborada pela autora (2022).

5.3 Partição de Dados: Treino e Teste

Uma das formas de testar o modelo é particionando a base de dados, esse método é usado quando é difícil conseguir mais observações, seja pelo custo ou por ser dados sigilosos.

Suponha que **D** seja uma base de dados com 1000 consultores, dividido em dois grupos: ativos $X^{(1)}$ e não ativos $X^{(2)}$. Particionando cada grupo em dois, obtém-se a matriz 5.1

$$\mathbf{D} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^{(1)} \\ \mathbf{X}_2^{(1)} \\ \cdots \\ \mathbf{X}_1^{(2)} \\ \mathbf{X}_2^{(2)} \end{bmatrix} \tag{5.1}$$

Agora, pode-se formar uma base com um $\mathbf{X}$ pertencente a cada um dos grupos que servirá para encontrar as variáveis potenciais e calcular os respectivos parâmetros, denominada por *Treino*. E pode-se formar uma outra base para testar os resultados do modelos, denominada *Teste*.

$$\mathbf{D}_{Treino} = \begin{bmatrix} \mathbf{X}_1^{(1)} \\ \cdots \\ \mathbf{X}_1^{(2)} \end{bmatrix}; \mathbf{D}_{Teste} = \begin{bmatrix} \mathbf{X}_2^{(1)} \\ \cdots \\ \mathbf{X}_2^{(2)} \end{bmatrix}. \tag{5.2}$$

Dessa forma, a base Treino representará os atuais consultores, e a base Teste os futuros consultores, e assim torna-se uma alternativa de simular e verificar a eficácia do modelo preditivo. Para o presente trabalho, dois dos modelos construídos foram utilizando 70% dos dados para treino e 30% dos dados para teste e os outros dois modelos foram utilizando 80% dos dados para treino e 20% dos dados para teste

5.4 Resultados Utilizando Variáveis Contínuas Não Categorizadas

Nesta seção será apresentado três modelos construídos, todos construídos no software R, utilizando as variáveis contínuas não categorizadas. Para a seleção de variáveis, foi realizada a análise da correlação entre elas e quando houve correlação, foram mantidas as mais importantes para a realização da interpretação dos resultados.

5.4.1 Modelo 1

Para o modelo apresentado nessa subseção, foi utilizado 70% da base dos dados para calcular os parâmetros do modelos e 30% para validação, além do mais, foi fixado o argumento *seed* no software para fins de reprodutibilidade da pesquisa. Tem-se que dentre todas as variáveis presentes na base de dados, 3 variáveis foram significativas para o modelo, conforme segue a Tabela 4 com as variáveis e seus respectivos parâmetros, desvio-padrão, Teste Wald e o P-valor para o teste.

Tabela 4 – Resultados do modelo utilizando 70% da base de dados para treino

Variável	Parâmetro	Estimativa	D.P	Teste Wald	P-Valor
Intercepto	β_0	9,7461	4,3949	2,2180	0,0265
Idade	β_1	-0,7109	0,2636	-2,6960	0,0070
Renda Per Capita	β_2	0,0023	0,0012	1,9360	0,0228
Coeficiente de Atividade	β_3	0,2078	0,0881	2,3570	0,0184

Fonte: Elaborada pela autora (2022).

Na tabela de razão de chances, podemos visualizar as estimativas para cada covariável do modelo.

Tabela 5 – Razão das chances do modelo utilizando 70% da base de dados para treino

Estimativas da Razão de Chance (Odds Ratio)	
Variáveis	**Estimativa**
Idade	$4,9117e^{-01}$
Renda Per capita	$1,0023e^{+00}$
Coeficiente de Atividade	$1,2309e^{+00}$

Fonte: Elaborada pela autora (2022).

Interpretando a covariável idade, temos que $4,9117e^{-01} = 0,4911$, logo, podemos concluir que a cada acréscimo de unidade na idade do consultor(a) diminui em até aproximadamente 51% a chance do consultor(a) de ser ativo desde que não sofra alteração de mais nenhuma covariável.

Para a covariável renda per capita, $1,0023e^{+00} = 1,0023$, logo, a cada acréscimo de unidade na renda per capita do consultor(a) aumenta em até 0,23% as chances do consultor ser ativo, desde que não sofra alteração de mais nenhuma covariável.

Por fim, interpretando o coeficiente de atividade (Desvio padrão da pontuação dividido pela pontuação média por tempo), temos que $1,2309e^{+00} = 1,2309$, ou seja, a cada acréscimo de unidade no coeficiente de atividade aumenta em até 23% as chances do consultor(a) de ser ativo.

No quadro abaixo, pode verificar a matriz de confusão obtida com o modelo.

Quadro 3 - Matriz de Confusão do modelo utilizando 70% da base de dados para treino

	Valor Predito	
Verdadeiro Valor	Evento 1	Evento 0
Evento 1	10	1
Evento 0	1	10

Fonte: Elaborada pela autora (2022).

As métricas de desempenho obtida a partir da matriz de confusão é apresentada na Tabela 6.

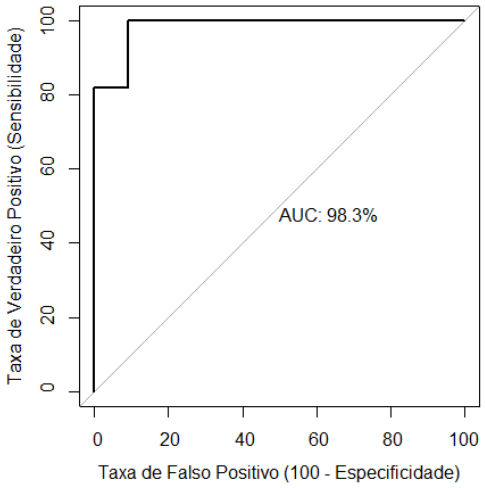
Tabela 6 – Métricas de Desempenho do modelo utilizando 70% da base de dados para treino

Métrica	Valor
Acurácia	0,909
Precisão	0,909
Sensibilidade	0,909
Especificidade	0,909
1-Especificidade	0,091
F1-Score	0,909

Fonte: Elaborada pela autora (2022).

Observando os valores obtidos, percebe-se que o modelo tem um bom desempenho, com todas as métricas de performance com mais de 90%. A imagem abaixo apresenta a curva ROC do modelo.

Figura 17 – Curva ROC do modelo utilizando 70% da base de dados para treino



Fonte: Elaborada pela autora (2022).

A Área Sobre a Curva do modelo construído foi de 98,30%, apresentando uma ótima performance.

5.4.2 Modelo 2

Utilizando para este modelo 80% da base dos dados para calcular os parâmetros do modelo (treino), e 20% para validação, além do mais, foi fixado o argumento *seed* no software para fins de reprodutibilidade da pesquisa. Tem-se que dentre todas as variáveis

presentes na base de dados, 4 variáveis foram significativas para o modelo, conforme segue a Tabela 7, com as variáveis e seus respectivos parâmetros, desvio-padrão, Teste Wald e o P-valor para o teste.

Tabela 7 – Resultados do modelo utilizando 80% da base de dados para treino

Variável	Parâmetro	Estimativa	D.P	Teste Wald	P-Valor
Intercepto	β_0	5,0666	4,0713	1,2440	0,2133
Idade	β_1	-0,7496	0,2461	-3,0450	0,0023
Iniciadora	β_2	0,3455	1,6086	2,1480	0,0317
Renda Per Capita	β_3	0,00371	0,001386	2,67800	0,00741
Coef. de Variação (Tempo)	β_4	0,0187	0,0081	2,3030	0,0212

Fonte: Elaborada pela autora (2022).

Na Tabela 8, podemos visualizar as estimativas para cada covariável do modelo.

Tabela 8 – Razão das chances do modelo utilizando 80% da base de dados para treino

Estimativas da Razão de Chance (Odds Ratio)	
Variáveis	**Estimativa**
Idade	$0,4725$
Iniciadora	$1,4119$
Renda Per capita	$1,0037$
Coef. de Variação (Tempo)	$1,0189$

Fonte: Elaborada pela autora (2022).

Interpretando a covariável idade, temos que $0,4725$, logo, podemos concluir que a cada acréscimo de unidade na idade do consultor(a) diminui em até aproximadamente 53% a chance do consultor(a) de ser ativo desde que não sofra alteração de mais nenhuma covariável.

A covariável inciadora tem estimativa de $1,4119$, ou seja, caso nenhuma covariável tenha alteração, o consultor(a) ser iniciado na companhia pelo diretor de vendas aumenta 42% a chance de ser ativo(a) em relação ao consultor(a) que não foi iniciado pelo diretor de vendas, apresentando um alto impacto.

Para a covariável renda per capita, $1,0037$, logo, a cada acréscimo de unidade na renda per capita do consultor(a) aumenta em até 0,38% as chances do consultor ser ativo, desde que não sofra alteração de mais nenhuma covariável.

Por fim, interpretando o coeficiente de variação em relação ao tempo, temos que $1,0189$, ou seja, a cada acréscimo de unidade no coeficiente de variação de tempo aumenta em até 1,9% as chances do consultor(a) de ser ativo.

No quadro abaixo, pode verificar a matriz de confusão obtida com o modelo.

Quadro 4 - Matriz de Confusão do modelo utilizando 80% da base de dados para treino

	Valor Predito	
Verdadeiro Valor	Evento 1	Evento 0
Evento 1	4	3
Evento 0	1	6

Fonte: Elaborada pela autora (2022).

A partir da matriz de confusão, as métricas de desempenho obtidas são apresentadas na Tabela 9.

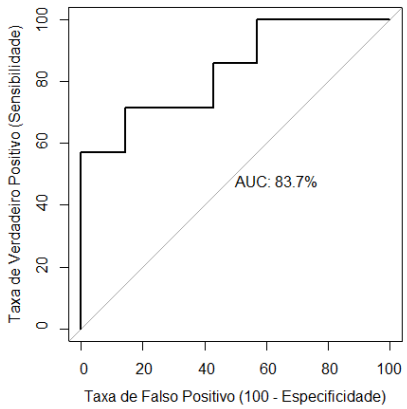
Tabela 9 – Métricas de Desempenho do modelo utilizando 80% da base de dados para treino

Métrica	**Valor**
Acurácia	0,714
Precisão	0,800
Sensibilidade	0,571
Especificidade	0,857
1-Especificidade	0,143
F1-Score	0,667

Fonte: Elaborada pela autora (2022).

Interpretando os resultados das métrica obtidas, é evidente que o modelo construído utilizando 70% dos dados para treino e 30% dos dados para teste obteve melhores resultados nas métricas.

Figura 18 – Curva ROC do modelo utilizando 80% da base de dados para treino



Fonte: Elaborada pela autora (2022).

A Área Sobre a Curva do modelo construído foi de 83,70%, apresentando uma boa performance, porém menor que o modelo utilizando 70% da base de treino e 30% da base de teste.

5.4.3 Modelo 3

Sabendo que a base de dados é consideravelmente pequena, o terceiro modelo construído utiliza toda a base de dados, sem realizar o particionamento entre dados de treino e teste, com o objetivo de construir o modelo sobre uma base com maior quantidade de informações e comparar seu desempenho com os demais modelos.Tem-se que dentre todas as variáveis presentes na base de dados, 3 variáveis foram significativas para o modelo, conforme segue a Tabela 10 com as variáveis e seus respectivos parâmetros, desvio-padrão, Teste Wald e o P-valor para o teste.

Tabela 10 – Resultados do modelo utilizando toda a base de dados

Variável	Parâmetro	Estimativa	D.P	Teste Wald	P-Valor
Intercepto	β_0	6,2807	2,9327	2,1410	0,0322
Idade	β_1	-0,4813	0,1459	-3,2970	0,0009
Renda Per Capita	β_2	0,0013	0,0006	2,0800	0,0375
Tempo de Atividade	β_3	5,220351	3,0902	3,179000	0,0014

Fonte: Elaborada pela autora (2022).

Na tabela de razão de chances, podemos visualizar as estimativas para cada covariável do modelo.

Tabela 11 – Razão das chances do modelo utilizando toda a base de dados

Estimativas da Razão de Chance (Odds Ratio)	
Variáveis	**Estimativa**
Idade	$6,1795e^{-01}$
Renda Per capita	$1,0013e^{+00}$
Tempo de Atividade	$1,8485^{+02}$

Fonte: Elaborada pela autora (2022).

Interpretando a covariável idade, temos que $6,1795e^{-01} = 0.6179$, logo, podemos concluir que a cada acréscimo de unidade na idade do consultor(a) diminui em até 38% a chance do consultor(a) de ser ativo desde que não sofra alteração de mais nenhuma covariável.

Para a covariável renda per capita, $1,0013e^{+00} = 1,0013$, logo, a cada acréscimo de unidade na renda per capita do consultor(a) aumenta em até 0,14% as chances do consultor ser ativo, desde que não sofra alteração de mais nenhuma covariável.

Por fim, interpretando o tempo de atividade, temos que $1,8485e^{+02} = 184,8512$, ou seja, a cada acréscimo de unidade no coeficiente de atividade aumenta em até aproximadamente 185% as chances do consultor(a) de ser ativo.

No quadro abaixo, pode verificar a matriz de confusão obtida com o modelo.

Quadro 5 - Matriz de Confusão do modelo utilizando toda a base de dados

	Valor Predito	
Verdadeiro Valor	Evento 1	Evento 0
Evento 1	32	2
Evento 0	3	37

Fonte: Elaborada pela autora (2022).

As métricas de desempenho obtida a partir da matriz de confusão é apresentada na Tabela 12.

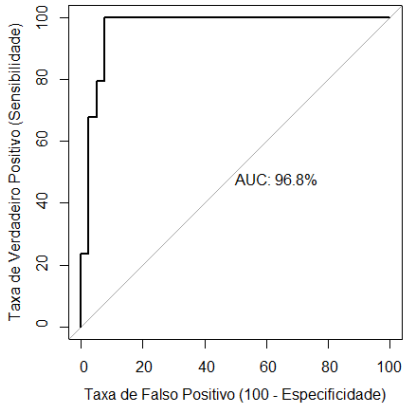
Tabela 12 – Métricas de Desempenho do modelo utilizando toda a base de dados

Métrica	**Valor**
Acurácia	0,932
Precisão	0,914
Sensibilidade	0,942
Especificidade	0,925
1-Especificidade	0,075
F1-Score	0,923

Fonte: Elaborada pela autora (2022).

Observando os valores obtidos, percebe-se que o modelo tem um bom desempenho, com todas as métricas de performance com acima de 90%. A imagem abaixo apresenta a curva ROC do modelo.

Figura 19 – Curva ROC do modelo utilizando toda a base de dados



Fonte: Elaborada pela autora (2022).

A Área Sobre a Curva do modelo construído foi de 96,80%, apresentando uma ótima performance.

5.5 Resultados Utilizando Variáveis Categorizadas

Nesta seção será apresentado três modelos construídos, ambos utilizando as variáveis contínuas categorizadas. Para a seleção de variáveis, foi realizada a análise da correlação entre elas e quando houve correlação, foram mantidas as mais importantes para a realização da interpretação dos resultados.

5.5.1 Modelo 1

Utilizando as variáveis numéricas categorizadas, com 70% da base dos dados para calcular os parâmetros do modelos e 30% para validação, além do mais, foi fixado o argumento *seed* no software para fins de reprodutibilidade da pesquisa. Tem-se, dentre todas as variáveis presentes na base de dados que 3 variáveis foram significativas para o modelo, conforme segue a Tabela 13 com as variáveis e seus respectivos parâmetros, desvio-padrão, Teste Wald e o P-valor para o teste.

Tabela 13 – Resultados do modelo com variáveis categorizadas utilizando 70% da base de dados para treino

Variável	**Parâmetro**	**Estimativa**	**D.P**	**Teste Wald**	**P-Valor**
Intercepto	β_0	1,4343	0,9383	1,5290	0,1263
Despesa com Moradia	β_1	1,4830	0,8649	1,7150	0,0364
Renda Per Capita 3	β_2	-2,8775	0,9106	-3,1600	0,0015
Coef. de Atividade 1	β_3	-2,5376	1,0572	-2,4000	0,0164

Fonte: Elaborada pela autora (2022).

A Tabela 14 apresenta as estimativas para cada covariável do modelo.

Tabela 14 – Razão das chances do modelo com variáveis categorizadas utilizando 70% da base de dados para treino

Estimativas da Razão de Chance (Odds Ratio)	
Variáveis	**Estimativa**
Despesa com Moradia	$4,4060$
Renda Per Capita 3	$0,0562$
Coef. de Atividade 1	$0,0790$

Fonte: Elaborada pela autora (2022).

Interpretando a covariável despesa com moradia, temos que o consultor(a) ter despesa com moradia aumenta aproximadamente 341% a chance do consultor(a) de ser ativo em relação ao consultor(a) que não tem despesa com moradia, desde que não haja mudança de nenhuma covariável.

O consultor(a) que apresenta a renda per capita entre R$ 999,00 e R$ 1894,00 (inclusive), diminui as chances de ser ativo em 94% em relação ao consultor(a) que não apresenta a renda per capita entre R$ 999,00 e R$ 1894,00 (inclusive).

Interpretando a estimativa do coeficiente de atividade 1, tem-se que o consultor(a) ter coeficiente de atividade menor ou igual a 0,51 diminui as chances de se tornar ativo em 92% em relação ao consultor(a) que não tem coeficiente de atividade menor ou igual a 0,51.

No quadro 6, pode verificar a matriz de confusão obtida com o modelo.

Quadro 6 - Matriz de Confusão do modelo com variáveis categorizadas utilizando 70% da base de dados para treino

	Valor Predito	
Verdadeiro Valor	Evento 1	Evento 0
Evento 1	7	3
Evento 0	3	7

Fonte: Elaborada pela autora (2022).

As métricas de desempenhos obtidas através da matriz de confusão é apresentada na Tabela 15.

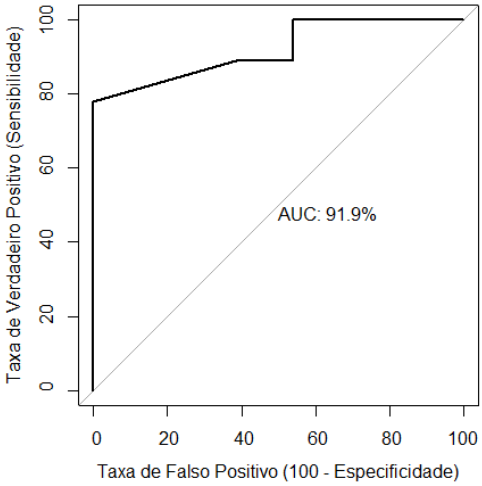
Tabela 15 – Métricas de Desempenho do modelo com variáveis categóricas utilizando 70% da base de dados para treino

Métrica	**Valor**
Acurácia	0,727
Precisão	0,700
Sensibilidade	0,700
Especificidade	0,750
1-Especificidade	0,250
F1-Score	0,700

Fonte: Elaborada pela autora (2022).

As métricas de desempenho obtidas apresentam performance aceitável, porém menor quando comparado com os resultados obtidos no modelo 1 para dados não categorizados.

Figura 20 – Curva ROC do modelo com variáveis categorizadas utilizando 70% da base de dados para treino



Fonte: Elaborada pela autora (2022).

A Área Sobre a Curva do modelo construído foi de 91,90%, apresentando uma boa performance.

5.5.2 Modelo 2

O modelo foi construído utilizando as variáveis numéricas categorizadas, porém com 80% da base dos dados para treinar o modelo e 20% para testá-lo. Além do mais, foi fixado o argumento *seed* no software para fins de reprodutibilidade da pesquisa. Conforme apresentado na tabela 16, 3 variáveis foram significativas para o modelo.

Tabela 16 – Resultados do modelo com variáveis categorizadas utilizando 80% da base de dados para treino

Variável	Parâmetro	Estimativa	D.P	Teste Wald	P-Valor
Intercepto	β_0	5,1963	1,5900	3,2680	0,0010
Idade 3	β_1	-2,7686	1,2287	-2,2530	0,0242
Renda Per Capita 3	β_2	-3,6833	0,9769	-3,7710	0,0002
Coef. de Atividade 1	β_3	-2,4216	0,9690	-2,4990	0,0124

Fonte: Elaborada pela autora (2022).

Na tabela de Razão de Chances, podemos visualizar as estimativas para cada covariável do modelo.

Interpretando a covariável Idade 3, temos que o consultor(a) ter idade maior ou igual a 24 anos diminui aproximadamente 95% a chance do consultor(a) de ser ativo em relação ao consultor(a) que não tem idade maior ou igual a 24 anos, desde que não haja mudança de nenhuma covariável.

Tabela 17 – Razão das chances do modelo com variáveis categorizadas utilizando 80% da base de dados para treino

Estimativas da Razão de Chance (Odds Ratio)	
Variáveis	**Estimativa**
Idade 3	0,0627
Renda Per Capita 3	0,0251
Coef. de Atividade 1	0,0887

Fonte: Elaborada pela autora (2022).

O consultor(a) que apresenta a renda per capita entre R$ 999,00 e R$ 1894,00 (inclusive), diminui as chances de ser ativo em 97% em relação ao consultor(a) que não apresenta a renda per capita entre R$ 999,00 e R$ 1894,00 (inclusive), desde que não haja mudança de nenhuma covariável.

Interpretando a estimativa do coeficiente de atividade 1, tem-se que o consultor(a) ter coeficiente de atividade menor ou igual a 0,51 diminui as chances de se tornar ativo em 91% em relação ao consultor(a) que não tem coeficiente de atividade menor ou igual a 0,51, desde que não haja mudança de nenhuma covariável.

No Quadro 7, pode verificar a matriz de confusão obtida com o modelo.

Quadro 7 - Matriz de Confusão do modelo com variáveis categorizadas utilizando 80% da base de dados para treino

	Valor Predito	
Verdadeiro Valor	Evento 1	Evento 0
Evento 1	5	3
Evento 0	1	5

Fonte: Elaborada pela autora (2022).

As métricas de desempenho construídas a partir da matriz de confusão pode ser observada na Tabela 18.

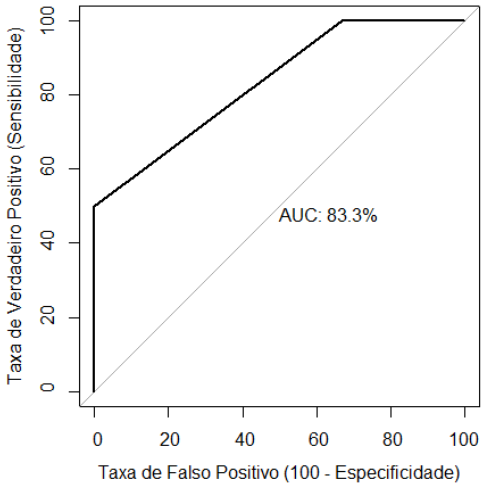
Tabela 18 – Métricas de Desempenho modelo com variáveis categorizadas utilizando 80% da base de dados para treino

Métrica	Valor
Acurácia	0,714
Precisão	0,833
Sensibilidade	0,625
Especificidade	0,833
1-Especificidade	0,167
F1-Score	0,714

Fonte: Elaborada pela autora (2022).

É evidente que os resultados obtidos estão próximos quando comparado aos resultados do modelo construído utilizando as variáveis categorizadas com 70% da base de dados para treino.

Figura 21 – Curva ROC do modelo com variáveis categorizadas utilizando 80% da base de dados para treino



Fonte: Elaborada pela autora (2022).

A Área Sobre a Curva do modelo construído foi de 83,30%, apresentando uma boa performance, porém menor que o modelo covariáveis contínuas categorizadas utilizando 70% da base de treino e 30% da base de teste.

5.5.3 Modelo 3

Utilizando as variáveis numéricas categorizadas, o modelo apresentado nessa subseção foi construído utilizando toda a base de dados, sem realizar particionamento. Conforme segue na Tabela 19, pode-se observar as covariáveis, seus respectivos parâmetros, desvio-padrão, Teste Wald e o P-valor para o teste.

Tabela 19 – Resultados do modelo com variáveis categorizadas utilizando toda a base de dados

Variável	Parâmetro	Estimativa	D.P	Teste Wald	P-Valor
Intercepto	β_0	-0,3101	0,8363	-0,3710	0,7108
Idade 3	β_1	-2,0204	0,8610	-2,3470	0,0189
Despesa com Moradia	β_2	1,3527	0,6764	2,0000	0,0455
Renda Per Capita 3	β_2	1,2500	0,7535	4,1710	$3{,}03e^{-05}$

Fonte: Elaborada pela autora (2022).

A Tabela 20 apresenta as estimativas para cada covariável do modelo.

Tabela 20 – Razão das chances do modelo com variáveis categorizadas utilizando toda a base de dados

Estimativas da Razão de Chance (Odds Ratio)	
Variáveis	**Estimativa**
Idade 3	$0,1325$
Despesa com Moradia	$3,8679$
Renda Per Capita 3	$3,4904$

Fonte: Elaborada pela autora (2022).

O consultor(a) que apresenta a idade maior ou igual a 24 anos diminui as chances de ser ativo em 87% em relação ao consultor(a) que não apresenta a idade maior ou igual a 24 anos.

Interpretando a covariável despesa com moradia, temos que o consultor(a) ter despesa com moradia aumenta aproximadamente 286% a chance do consultor(a) de ser ativo em relação ao consultor(a) que não tem despesa com moradia, desde que não haja mudança em covariável alguma.

O consultor(a) que apresenta a renda per capita entre R$ 999,00 e R$ 1894,00 (inclusive), aumenta as chances de ser ativo em 249% em relação ao consultor(a) que não apresenta a renda per capita entre R$ 999,00 e R$ 1894,00 (inclusive).

No quadro 8, pode verificar a matriz de confusão obtida com o modelo.

Quadro 8 - Matriz de Confusão do modelo com variáveis categorizadas utilizando base de dados total

	Valor Predito	
Verdadeiro Valor	Evento 1	Evento 0
Evento 1	27	7
Evento 0	3	37

Fonte: Elaborada pela autora (2022).

As métricas de desempenhos obtidas através da matriz de confusão é apresentada na Tabela 21.

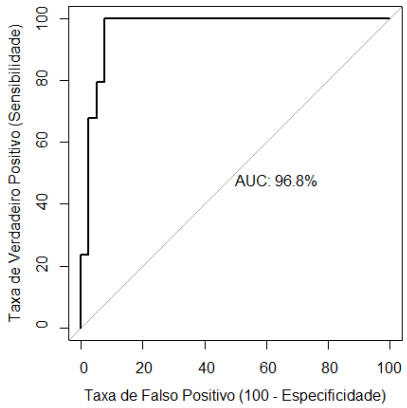
Tabela 21 – Métricas de Desempenho do modelo com variáveis categóricas utilizando toda a base de dados

Métrica	**Valor**
Acurácia	0,864
Precisão	0,900
Sensibilidade	0,794
Especificidade	0,925
1-Especificidade	0,075
F1-Score	0,844

Fonte: Elaborada pela autora (2022).

As métricas de desempenho obtidas apresentam performance aceitável, porém menor quando comparado com os resultados obtidos no modelo 3 para dados não categorizados.

Figura 22 – Curva ROC do modelo com variáveis categorizadas utilizando toda a base de dados



Fonte: Elaborada pela autora (2022).

A Área Sobre a Curva do modelo construído foi de 96,80%, apresentando uma boa performance e exatamente a mesma área que o Modelo 3 para dados não categorizados.

5.6 Resumo dos Modelos Construídos

Nesta seção, será apresentado o resumo dos seis modelos construídos.

Tabela 22 – Resumo das Métricas dos Modelos Construídos Utilizando Variáveis Contínuas Não Categorizadas

Modelos	Acurácia	Precisão	Sensibilidade	Especificidade	1-Especificidade	F1-Score
Modelo 1	0,909	0,909	0,909	0,909	0,091	0,909
Modelo 2	0,714	0,800	0,571	0,857	0,143	0,667
Modelo 3	0,932	0,914	0,942	0,925	0,075	0,923

Fonte: Elaborada pela autora (2022).

Observando os modelos utilizando variáveis contínuas não categorizadas, nós temos que o Modelo 1 e o Modelo 3 tiveram melhor performance em todas as métricas.

Tabela 23 – Resumo das Métricas dos Modelos Construídos Utilizando Variáveis Contínuas Categorizadas

Modelos	Acurácia	Precisão	Sensibilidade	Especificidade	1-Especificidade	F1-Score
Modelo 1	0,727	0,700	0,700	0,750	0,250	0,700
Modelo 2	0,714	0,833	0,625	0,833	0,167	0,714
Modelo 3	0,864	0,900	0,794	0,925	0,075	0,844

Fonte: Elaborada pela autora (2022).

Contudo, observando o diagnóstico dos 6 modelos construídos, podemo concluir que os modelos utilizando variáveis contínuas não categorizadas obtiveram um melhor desempenho. Além do mais os modelos 1 (Utilizando 70 % da base de dados para treino) e 3 (Utilizando toda a base de dados para treino) apresentados na Tabela 22 foram os modelos com melhor performance. Podemos concluir no entanto que um diretor de vendas que deseja facilitar a tomada de decisões utilizando modelos preditivos estatísticos, podem utilizar o Modelo 1 e o Modelo 3 como ferramentas para a criação de estratégias.

6 Considerações Finais

Os modelos construídos e apresentados no trabalho usando diferentes partição dos dados entre treino, teste e a base total sem particionamento, tiveram efeitos semelhantes em relação às sensibilidades, especificidades e acurácia. Quando comparado os modelos construídos com variáveis contínuas não categorizadas e com variáveis categorizadas, observa-se uma leve semelhança entre os resultados, porém, os modelos com variáveis contínuas não categorizadas obteve melhores resultados em relação a acurácia, precisão, sensibilidade, especificidade, f1-score e AUC.

Contudo, após a construção, análise e interpretação dos modelos, é possível afirmar que o terceiro modelo apresentado, com variáveis contínuas não categorizadas e sem o particionamento da base de dados obteve melhores métricas quando comparado aos demais modelos criados. Porém, vale ressaltar que como não houve separação da base de dados, a validação do modelo foi realizada na própria base de dados utilizada para criar o modelo. Em decorrência da performance dele, podemos considerá-lo útil para desenvolvimento de estratégias para o time de consultores(as).

Interpretando os parâmetros do modelo sob ponto de vista de negócio, tem-se que consultores(as) com idades mais avançadas tem menores probabilidades de serem ativos(as). Outro fator interessante é que consultores que têm renda per capita maiores obtém também maiores probabilidades de serem ativos. Por fim, o tempo de atividade do consultor tem um forte impacto, quanto maior essa métrica, maiores as probabilidades do consultor de ser ativo na companhia. Contudo, com essas informações, o líder de vendas tem possibilidade e ferramentas para criar estratégias com o foco em atingir esses consultores.

Além desse, o primeiro modelo apresentado, com variáveis contínuas não categorizadas e com partição de 70% da base de dados para treino e 30% da base de dados para teste obteve uma ótima performance e pode ser também o modelo utilizado para facilitar a tomada de decisões da diretora de vendas.

Objetivando auxiliar o dia a dia do diretor(a) de vendas da equipe, temos que por interpretação dos parâmetros do modelo 1 (utilizando variáveis contínuas não categorizadas), consultores(as) com idades não muito avançadas apresentam maiores chances de serem ativos(as), logo, o diretor(a) de vendas da companhia pode concentrar esforços em convidar pessoas jovens para entrarem na empresa.

Além disso, o coeficiente de atividade é uma variável importante para prever a probabilidade do consultor(a) ser ativo(a), logo, o diretor(a) de vendas tem a oportunidade de criar estratégias para incentivar a equipe a realizarem pedidos constantemente, para

assim aumentar a probabilidade dos consultores(as) de serem ativos(as).

É importante ressaltar que essas conclusões são válidas para os dados utilizados neste trabalho, em específico para a equipe de vendas estudada da companhia.

Referências

ASH, M. K.**The Mary Kay Way:** O estilo de liderança de uma das maiores empreendedoras norte-americanas. São Paulo: CLA Cultural, 2013.

BATISTA, A. M. Sarmento. **Regressão Logística:** uma introdução ao modelo estatístico: exemplo de aplicação ao revolving credit. Porto: Vida Económica, 2015.

BOLFARINE, H.; BUSSAB, W.O. **Elementos de Amostragem**. 1 ed. São Paulo: Blucher, 2004.

BRAGA, A. C. **Curva ROC:** Aspectos funcionais e aplicações, 2000. Tese (Doutorado em Estatística) - Universidade do Minho, Braga, 2000.

BRASIL. Lei 14.358/22, de 01 de junho de 2022. **Dispõe sobre o valor do salário-mínimo a vigorar a partir de 1º de janeiro de 2022**. Diário Oficial da União, Brasília, DF. Disponível em: $https://www.planalto.gov.br/ccivil_03/_{ato}2019-2022/2022/lei/l14358.htm$. Acesso em: 10 de junho de 2022

BUSSAB, W.; MORETTIN P. **Estatística básica**. 6th ed. Pinheiros: Editora Saraiva, 2009.

CABRAL, C. I. S. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado.** Tese (Mestrado em Matemática Aplicada à Economia e à Gestão) - Universidade de Lisboa Faculdade de Ciências, Lisboa, 2013.

COX, D. R. **The Analysis of Binary Data.** [S. l.]: Methuen, London.1970.

CRAMER, J. S. **The Origins of the Logistic Regression.** Tinbergen Institute Working Paper, Amsterdã, Nº 2002-119/4, p. 3-8, 2002. Disponível em: $https://papers.ssrn.com/sol3/papers.cfm?abstract_id = 360300$. Acesso em 02 de jun. de 2022.

FONSECA, J. S.; MARTINS, G. A. **Curso de Estatística.** 6th ed. São Paulo: Atlas, 1996.

FOREMAN, E. K. **Survey Sampling Principles.** 1th ed. Nova York: M. Dekker, 1991.

GOURIEROUX, C; MONFORT, A.**Statistics and econometric models**. Cambridge: University Press, 1995.

HOSMER, D.; LEMESHOW, S.; STURDIVANT, X. **Applied Logistic Regression.** 3th ed. Canada: John Wiley Sons, 2013.

MARY KAY. **Mary Kay comemora 55 anos de empoderamento e inspiração para empreendedores ao redor do mundo.** São Paulo, 13 set. 2018. Disponível em: https://www.marykay.com.br/pt-br/about-mary-kay/press-room/press-releases/2018/55-anos-mary-kay. Acesso em: 04 jan. 2022.

PAULA, G. A.; **Modelos de Regressão:** com apoio computacional. São Paulo: [s.n.], 2013.

PROVOST, F.; FAWCETT, T. **Data Science para negócios:** O que você precisa saber sobre mineração de dados e pensamento analítico dos dados. Rio de Janeiro: Alta Books, 2016.

SARMENTO, A. M. B.;**Regressão Logística Uma introdução ao modelo**. 1 ed. Portugal: Laureata, 2015.

SILVESTRE, M. R.**Notas de aula de Análise Multivariada II**. 2022. Presidente Prudente: [s.n.], 2022. Apostila disponível no ambiente virtual de apoio disciplina Análise Multivariada II da FCT/Unesp.

THOMPSON, S. K. Sampling. 2th ed. **United States of America**: Wiley-Interscience, 2002.

APÊNDICE A – Demonstrações

Nessa seção serão apresentadas as demonstrações das equações que foram apresentadas no trabalho.

A.1 Demonstração da equação (3.13)

Tem-se o seguinte *logit*:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x,$$

aplica-se exponencial em ambos os lados

$$e^{ln\left(\frac{\pi(x)}{1-\pi(x)}\right)} = e^{ln(\beta_0+\beta_1 x)},$$

tem-se por propriedade de logarítimo que $a^{log_a(b)} = b$, assim sendo:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x}.$$

Contudo, isola-se $\pi(x)$

$$\begin{aligned}
\frac{\pi(x)}{1-\pi(x)} &= e^{\beta_0+\beta_1 x} \\
\pi(x) &= e^{\beta_0+\beta_1 x}(1-\pi(x)) \\
\pi(x) &= e^{\beta_0+\beta_1 x} - e^{\beta_0+\beta_1 x}\pi(x) \\
\pi(x) + e^{\beta_0+\beta_1 x}\pi(x) &= e^{\beta_0+\beta_1 x} \\
\pi(x)(1+e^{\beta_0+\beta_1 x}) &= e^{\beta_0+\beta_1 x} \\
\pi(x) &= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}
\end{aligned}$$

Portando, a probabilidade do evento Y ocorrer dado x, é dada por

$$\pi(x) = E(Y|x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \tag{A.1}$$

A.2 Demonstração das equações (3.17 e 3.18)

Cada observação terá uma distribuição Bernoulli com parâmetro $(\pi(x))$:

$$f(y) = \pi(x_i)^y[1-\pi(x_i)]^{1-y},$$

assim, a função de verossimilhança será

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}.$$

O logaritmo natural da função de verossimilhança de $\boldsymbol{\beta}$ é denotado por

$$\begin{aligned}
L(\boldsymbol{\beta}|\mathbf{y}) &= ln(\prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}) \\
&= \sum_{i=1}^{n} \{\ y_i ln[\pi(x_i)] + (1-y_i) ln[1-\pi(x_i)]\} \ ^{1} \\
&= \sum_{i=1}^{n} \{\ y_i\ ln[\pi(x_i)] + ln[1-\pi(x_i)] - y_i ln[1-\pi(x_i)]\} \\
&= \sum_{i=1}^{n} \{\ y_i\ ln[\pi(x_i)] - ln[1-\pi(x_i)] + ln[1-\pi(x_i)]\} \\
&= \sum_{i=1}^{n} \{\ y_i\ (ln[\frac{\pi(x_i)}{1-\pi(x_i)}]) + ln[1-\pi(x_i)]\} \ ^{2} \\
&= \sum_{i=1}^{n} \ \{y_i\ (\beta_0 + \beta_1 x_i) + ln[1 - \frac{e^{(\beta_0+\beta_1 xi)}}{1+e^{(\beta_0+\beta_1 xi)}}]\} \ ^{3} \\
&= \sum_{i=1}^{n} \ \{y_i\ (\beta_0 + \beta_1 x_i) + ln[\frac{1}{1+e^{(\beta_0+\beta_1 xi)}}]\} \\
&= \sum_{i=1}^{n} \ \{y_i(\beta_0 + \beta_1 x_i)\ + ln(1) - ln(1+e^{(\beta_0+\beta_1 xi)})\} \ ^{4\,5} \\
&= \sum_{i=1}^{n} \ \{y_i(\beta_0 + \beta_1 x_i)\ - ln(1+e^{(\beta_0+\beta_1 xi)})\}
\end{aligned}$$

[1] Propriedades de logaritmo: $log_a b^n = n\ log_a b$
[2] Propriedades de logaritmo: $log(\frac{a}{b}) = log(a) - log(b)$
[3] Sabe-se que $ln[\frac{\pi(x_i)}{1-\pi(x_i)}] = \beta_0 + \beta_1 x_i$ e também que $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$
[4] Propriedades de logaritmo: $log(\frac{a}{b}) = log(a) - log(b)$
[5] Propriedades de logaritmo: $ln(1) = 0$

Os estimadores de máxima verossimilhança para os parâmetros β_0 eβ_1 são os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que maximizam o logaritmo da função de verossimilhança, para isso basta fazer a derivada em relação aos parâmetros do modelo, da seguinte forma

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_0} = \frac{\partial \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_i) \ - ln(1 + e^{(\beta_0+\beta_1 xi)})\}}{\partial \beta_0}$$

$$= \sum_{i=1}^{n} \{y_i \ \{\frac{\partial(\beta_0)}{\partial \beta_0} + \frac{\partial(\beta_1 x_i)}{\partial \beta_0}\} - \frac{\partial ln(1 + e^{(\beta_0+\beta_1 xi)})}{\partial \beta_0}\} \ [6]$$

$$= \sum_{i=1}^{n} \{y_i \ - \frac{1}{1 + e^{\beta_0+\beta_1 x}} \ \frac{\partial(1 + e^{(\beta_0+\beta_1 x)})}{\partial \beta_0}\}$$

$$= \sum_{i=1}^{n} \{y_i \ - \frac{1}{1 + e^{\beta_0+\beta_1 x}} \ e^{(\beta_0+\beta_1 x)}\}$$

$$= \sum_{i=1}^{n} \{y_i \ - \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{\beta_0+\beta_1 x}}\}$$

$$= \sum_{i=1}^{n} \{y_i - \frac{e^{\beta_0+\beta_1 xi}}{1 + e^{\beta_0+\beta_1 xi}}\}$$

$$= \sum_{i=1}^{n} \{y_i - \pi(x)\} \ [7]$$

analogamente, para β_1 :

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_1} = \frac{\partial \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_i) \ - ln(1 + e^{(\beta_0+\beta_1 xi)})\}}{\partial \beta_1}$$

$$= \sum_{i=1}^{n} \{y_i \ \{\frac{\partial(\beta_0)}{\partial \beta_1} + \frac{\partial(\beta_1 x_i)}{\partial \beta_1}\} - \frac{\partial ln(1 + e^{(\beta_0+\beta_1 xi)})}{\partial \beta_1}\} \ [8]$$

$$= \sum_{i=1}^{n} \ x_i\{y_i \ - \frac{1}{1 + e^{\beta_0+\beta_1 x}} \ \frac{\partial(1 + e^{(\beta_0+\beta_1 x)})}{\partial \beta_1}\}$$

$$= \sum_{i=1}^{n} \ x_i\{y_i \ - \frac{1}{1 + e^{\beta_0+\beta_1 x}} \ e^{(\beta_0+\beta_1 x)}\}$$

$$= \sum_{i=1}^{n} \ x_i\{y_i \ - \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{\beta_0+\beta_1 x}}\}$$

$$= \sum_{i=1}^{n} \ x_i\{y_i - \frac{e^{\beta_0+\beta_1 xi}}{1 + e^{\beta_0+\beta_1 xi}}\}$$

[6] Propriedades de derivadas: $(f - g)' = f' - g'$
[7] Substituindo (equação 3.13)
[8] Propriedades de derivadas: $(f - g)' = f' - g'$

$$= \sum_{i=1}^{n} x_i\{y_i - \pi(x)\} \text{ [9]}$$

Igualando estas derivadas a zero e substituindo os parâmetros β_0 e β_1 pelos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ obtem-se as equações apresentadas em (3.17) e (3.18):

$$\sum_{i=1}^{n} \{y_i - \frac{e^{(\hat{\beta}_0+\hat{\beta}_1 xi)}}{1+e^{(\hat{\beta}_0+\hat{\beta}_1 xi)}}\} = \sum_{i=1}^{n} \{y_i - \hat{\pi}(x_i)\} = 0.$$

$$\sum_{i=1}^{n} x_i\{y_i - \frac{e^{(\hat{\beta}_0+\hat{\beta}_1 xi)}}{1+e^{(\hat{\beta}_0+\hat{\beta}_1 xi)}}\} = \sum_{i=1}^{n} x_i\{y_i - \hat{\pi}(x_i)\} = 0$$

[9] Substituindo (equação 3.13)

APÊNDICE B – Variáveis

Nesta seção serão apresentadas as variáveis presentes no estudo.

B.1 Variáveis do Conjunto de Dados

O quadro a seguir apresenta as variáveis presente na base de dados a serem utilizadas no estudo, disponibilizada pela diretora de vendas da companhia Mary Kay.

Quadro 7 - Variáveis do conjunto de dados.

Variável	**Tipo de Variável**	**Categoria**	**Descrição**
codigo_consultora	ID - Identificação	-	Código do(a) consultor(a).
nome	Nominal	-	Nome do(a) consultor(a).
iniciadora	Nominal	-	Nome da inciadora da(o) consultor(a).
status	Ordinal	A1, A2, A3, N1, N2, N3, P1, P2, P3, T1, T2, T3, T4, T5, T6	Status do(a) consultor(a). A: ativo(a), N: não realizou, P: pendente, T: terminado(a).
quantidade_meses	Discreta	-	Quantidade de meses que o(a) consultor(a) está na empresa.
quantidade_pedidos	Discreta	-	Quantidade de pedidos que o(a) consultor(a) fez desde que entrou na empresa até janeiro de 2022.
inicio	Data	-	Data de início, ou seja, data de cadastro, do(a) consultor(a).
ultimo_pedido	Data	-	Data do último pedido do(a) consultor(a).
janeiro_2022	Discreta	-	Quantidade de pontuação do(a) consultor(a) em janeiro de 2022.
dezembro_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em dezembro de 2021.

Continua na próxima página

Quadro 7 – *Variáveis do conjunto de dados (Continuação)*

Variável	**Tipo de Variável**	**Categoria**	**Descrição**
novembro_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em novembro de 2021.
outubro_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em outubro de 2021.
setembro_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em setembro de 2021.
agosto_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em agosto de 2021.
julho_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em julho de 2021.
junho_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em junho de 2021.
maio_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em maio de 2021.
abril_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em abril de 2021.
março_2021	Discreta	-	Quantidade de pontuação do(a) consultor(a) em março de 2021.
sexo	Nominal	F: Feminino e M: Masculino	Sexo do(a) consultor(a).

B.2 Variáveis Coletadas

O quadro a seguir apresenta as variáveis oriundas do questionário aplicado.

Quadro 8 - Variáveis coletadas.

Variável	**Tipo de Variável**	**Categoria**	**Descrição**
codigo_consultora	ID - Identificação	-	Código do(a) consultor(a).
idade	Contínua	-	Idade do(a) consultor(a).

Continua na próxima página

Quadro 8 – *Variáveis coletadas (Continuação)*

Variável	**Tipo de Variável**	**Categoria**	**Descrição**
estado	Nominal	-	Nome do(a) consultor(a).
iniciadora	Nominal	-	Sigla da Unidade Federativa (UF) do estado do(a) consultor(a).
quantidade_filhos	Discreta	-	Quantidade de filhos que o(a) consultor(a) tem.
escolaridade	Ordinal	Sem Escolaridade, Ensino Fundamental Incompleto, Ensino Fundamental Completo, Ensino Médio Incompleto, Ensino Fundamental Completo, Superior Incompleto, Superior Completo, Pós Graduação	Nível de escolaridade do(a) consultor(a).
estado_civil	Nominal	Solteira, Casada, Uniao Estável, Viúva, Separação Legal	Estado civil do(a) consultor(a).
desp_com_moradia	Nominal	Sim: apresenta despesas com moradia e Não: não apresenta despesas com moradia	Indica se o(a) consultora(a) apresenta despesas com moradia.
renda_nao_mk	Nominal	Sim: apresenta renda financeira mensal com origem diferente da Mary Kay e Não: não apresenta renda financeira mensal com origem diferente da Mary Kay	Indica se o(a) consultor(a) tem renda financeira mensal que não seja oriunda da Mary Kay.

Continua na próxima página

Quadro 8 – *Variáveis coletadas (Continuação)*

Variável	**Tipo de Variável**	**Categoria**	**Descrição**
renda_nao_mk_principal	Nominal	Sim: Renda não Mary Kay é a principal Não: Renda não Mary Kay não é a principal	Caso o(a) consultor(a) tenha renda financeira que não seja da Mary Kay, a variável indica se é a principal.
renda_familiar_mensal	Ordinal	Menor que 1 s.m., entre 1 e 2 s.m., entre 2 e 3 s.m., entre 3 e 4 s.m., entre 4 e 5 s.m., entre 5 e 6 s.m., entre 6 e 7 s.m., entre 7 e 8 s.m., entre 9 e 10 s.m., maior que 10 s.m.	A renda familiar mensal total em termos de salários mínimos (s.m.).
quantidade_pessoas	Discreta	-	Quantidade de pessoas que compartilha da renda familiar mensal incluindo o(a) consultor(a).
pagamento	Nominal	Pix, boleto ou lotérica, 1 parcela cartão de rédito, 2 parcelas no cartão de crédto, 3 parcelas no cartão de crédito, Não fiz pedidos	O método de pagamento utilizado pela consultora para realizar o pagamento dos pedidos.
outras_marcas	Nominal	Avon, Eudora, Jequiti, Hinode, Natura, Outras, Não vendo outras marcas	Indica se o(a) consultor(a) vende outras marcas.