Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Um modelo de risco de crédito bayesiano para classificação de clientes inadimplentes

por

Monique Lohane Xavier Silva

Brasília, 28 de novembro de 2022

Um modelo de risco de crédito bayesiano para classificação de clientes inadimplentes

por

Monique Lohane Xavier Silva

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Brasília, 28 de novembro de 2022

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília como parte dos requisitos para a obtenção do título de Mestre em Estatística.

Texto aprovado por:

Eduardo Yoshio Nakano
Orientador, Departamento de Estatística / UnB

José Augusto Fiorucci
Departamento de Estatística / UnB

Marcelo Angelo Cirillo
Departamento de Estatística / UFLA

We have not got the money, so we have got to think.

(Ernest Rutherford)

Agradecimentos

Só pela vontade de Deus estou aqui, hoje. E é primeiramente a Ele que devo agradecer. Foi Deus que permitiu que o meu mestrado se concretizasse após passar por uma pandemia (Covid-19) e que eu estivesse viva e com saúde para finalizar este estudo.

Sou muito abençoada também pelo imenso apoio que tive para chegar até aqui. Irei me atrever a citar algumas pessoas: minha família (pai, mãe e irmãos) pois foram meu alicerce. Quando tantas vezes estava difícil de continuar, eles me apoiaram, ouviram muitos choros e me consolaram. Me lembro do ensaio para a qualificação, que coloquei todos na sala me assistindo ensaiar, e mesmo sem entender o que explicava, seus olhos estavam fitados em mim e ouviram cada palavra que eu dizia com muita atenção. Meu namorado Wagner que por várias vezes aprendia minhas matérias só para me ensinar, quando eu não entendia alguma coisa. Também me consolou quando chorei e comemorou comigo cada vitória e matéria vencida. O meu orientador Nakano, por todo conhecimento repassado, a paciência em me ensinar e fazer o mesmo cálculo quantas vezes fosse preciso até que eu entendesse. Sou muito grata pelos ensinamentos repassados. Meus amigos que foram fundamentais no início do mestrado, quando iam à noite para UnB para estudar em grupo, fazendo os exercícios no quadro. Foi devido as essas noites de estudo que consegui ingressar no mestrado. Meus amigos do trabalho que acompanharam minha jornada, me apoiaram quando precisei renunciar algumas horas de trabalho para me dedicar aos estudos.

Além de todas essas pessoas que tiveram uma interferência direta na minha jornada do mestrado, sou cercada de muitos amigos e familiares que influenciaram de alguma forma e algum nível na pessoa que sou hoje; amigas da escola, amigas da igreja, da faculdade,

do trabalho... todos contribuíram de alguma para que eu chegasse até aqui, conquistando o meu título de Mestre em Estatística. E tudo isso para dizer que sozinhos não somos nada, e a minha rede de apoio foi a melhor que Deus poderia me dar.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

O objetivo desse trabalho foi propor uma modelagem bayesiana de risco de crédito para a classificação de clientes quanto ao seu risco de inadimplência. O diferencial da metodologia proposta é a possibilidade de incorporar uma informação *a priori* no processo de classificação dos clientes e não apenas na obtenção das estimativas dos parâmetros do modelo que gera o Escore de Risco. A principal vantagem desse procedimento se deve à simplicidade em incorporar a opinião do especialista no processo de classificação, algo que não ocorre na modelagem bayesiana tradicional, cuja informação *a priori* recai sobre os parâmetros dos modelos que, geralmente, são quantidades abstratas e/ou associadas à covariáveis sujeitas a problemas de multicolinearidade. Para a devida ilustração da metodologia proposta, utilizou-se um conjunto de dados na literatura e os resultados obtidos mostraram que o modelo é útil para a classificação de clientes quanto a sua probabilidade de inadimplência.

Abstract

The aim of this work was to propose a bayesian credit risk model for classifying customers in terms of their default risk. The differential of the proposed methodology is the possibility of incorporating *a priori* information in the customer classification process and not just in the estimation of the customers' evaluation parameters. The main advantage of this procedure is due to the simplicity in incorporating the expert's opinion in the classification process, something that does not occur in traditional bayesian modeling, whose a priori information falls on the parameters of the models, which are usually abstract quantities and/or associated with covariates with multicollinearity problems. To illustrate the proposed methodology, a dataset in the literature was used and the results obtained showed that the model is useful for classifying customers in terms of their probability of default.

Sumário

Lista de Tabelas

Lista de Figuras

Abreviaturas e Siglas

C	Categorias de risco
EMV	Estimador da Máxima Verossimilhança
ER	Escore de Risco
E	Grupos de risco
FDP	Função de Densidade de Probabilidade
HPD	*Highest Posterior Density*
IID	Independentes e Identicamente Distribuídas
MCMC	Markov Chain Monte Carlo
MMQ	Método de Mínimos Quadrados
MQO	Mínimos Quadrados Ordinários
MSE	*Mean Squared Errors*
VA	Variável Aleatória
Var	Variância

Capítulo 1

Introdução

1.1 Considarções iniciais

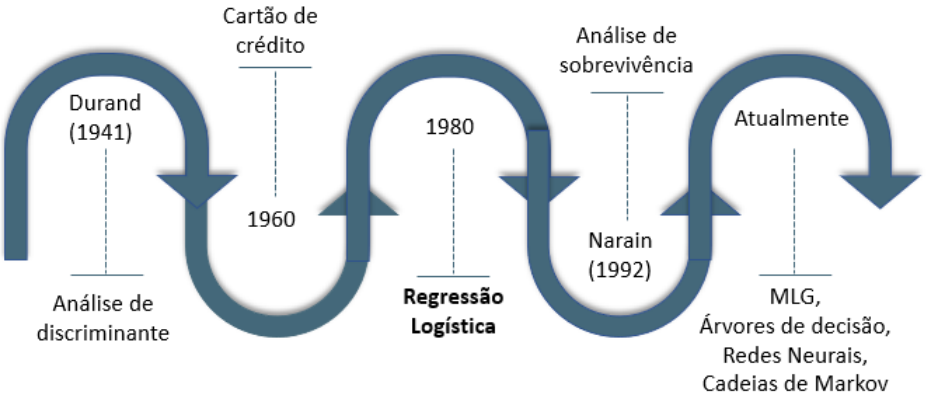
A concessão de crédito é um fenômeno bastante recorrente em economias desenvolvidas. A alocação desse capital em empreendimentos rentáveis estimula o crescimento da economia de um país, uma vez que produtos e serviços são criados para atender as demandas do mercado. Contudo, credores se preocupam com o risco envolvido nessa operação, visto que os clientes podem não conseguir honrar o pagamento de seus empréstimos por diversos motivos. Sendo assim, é de interesse dessas instituições financeiras avaliar o risco associado ao cliente, antes de fazer a concessão do crédito, para evitar casos de inadimplência e, assim, tornar o negócio mais lucrativo. Sob esse cenário, surgiram os modelos de escore de risco de crédito como ferramentas capazes de identificar perfis adimplentes e inadimples, assim como quantificar o risco em disponibilizar crédito a cada um deles. Esses modelos capacitam as instituições financeiras a tomarem decisões de forma rápida, automática, padronizada e objetiva, sendo eles amplamente aceitos e utilizados por inúmeras coorporações.

A elaboração dos modelos de escore de risco de crédito é feita com base em definições e abordagens estatísticas. Além disso, cada vez mais surgem modelos obtidos seguindo o paradigma bayesiano, permitindo a incorporação de uma informação *à priori* no processo de estimação dos parâmetros do modelo. No entanto, uma das dificuldades encontradas é

que, usualmente, é inviável um especialista ter uma opinião *à priori* formada em relação a um parâmetro que, não raramente, está associado a variáveis explicativas sujeitas a problemas de multicolinearidade.

Historicamente, os modelos de escore de risco de crédito foram iniciados pelos estudos de Durand (1941), que reconheceu que a análise discriminante poderia ser utilizada para identificar e categorizar empréstimos com potenciais de serem quitados e empréstimos devidos. O uso de um escore de risco de crédito foi intensificado com a chegada do cartão de crédito na década de 60, que resultou também em um maior interesse no meio acadêmico. Na década de 80, escores de crédito foram desenvolvidos por meio da regressão logística e, em 1992, surgiu o primeiro modelo de crédito utilizando técnicas de análise de sobrevivência (Narain et al., 1992). Os modelos de escore de risco de crédito têm se tornado populares e atualmente são largamente utilizados. Desse modo, várias técnicas vêm sendo aplicadas como ferramentas para novos modelos, sendo as principais árvores de decisão, redes neurais, cadeias de Markov, algoritmos genéticos, modelos lineares generalizados, como mostra a Figura 1.1:

Figura 1.1: Métodos de avaliação de escore de risco de crédito ao longo do tempo



Fonte: Machado, 2015

1.2 Objetivos

O objetivo desse trabalho foi propor um Escore de Risco (ER) fundamentado na regressão logística, que utiliza a metodologia bayesiana para elaborar uma regra de decisão

que incorpore informações *à priori* no critério de classificação dos clientes. A metodologia proposta foi ilustrada por meio de um conjunto de dados obtidos na literatura.

1.3 Dados para ilustração da metodologia

A metodologia proposta neste trabalho será ilustrada utilizando uma base de dados disponibilizada por Dua and Graff (2017) do Center for Machine Learning and Intelligence Systems (CML), o *German Credit Data*. Este conjunto de dados tem sido muito utilizado em estudos sobre risco de credito e também será estudado nesta dissertação. A base de dados apresenta 1000 linhas de solicitantes de crédito e 21 variáveis que permitem uma melhor descrição de cada indivíduo presente no estudo.

1.4 Esboço do trabalho

O trabalho está organizado em cinco capítulos, incluindo esta introdução. O Capítulo 2 apresenta uma breve revisão bibliográfica dos conceitos que serão utilizados na proposição do ER. O primeiro conceito é sobre os modelos de regressão logística binária e o segundo é sobre métricas de performance, no qual é apresentada a matriz de confusão e as medidas de desempenho atreladas e derivadas dela. O próximo assunto é sobre a inferência bayesiana, no qual são discorridos seus principais conceitos e características. A última seção do capítulo de revisões bibliográficas retrata os modelos logísticos bayesianos, em que são combinados ambos conceitos, antes apresentados individualmente, e de suma importância para a aplicação e o desenvolvimento do escore de risco de crédito a ser proposto. O Capítulo 3 contextualiza, inicialmente, modelagens de risco de crédito. Em seguida, é desenvolvida a metodologia do ER a ser proposto, assim como a aplicação das medidas de desempenho que serão utilizadas para verificação do ajuste do modelo. Sua última seção apresenta a regra de decisão para o qual será feita a devida classificação de clientes nos grupos adimplentes e inadimplentes conforme ER apresentado no capítulo anterior. No Capítulo 4 é feita a aplicação da metodologia em dados obtidos na literatura. Neste

capítulo será discorrido uma análise exploratória dos dados, uma aplicação da metodologia assim como avaliação de aderência do modelo. Também será apresentado um estudo de sensibilidade da *priori* em que se pode verificar como o modelo responde a diferentes probabilidades estipuladas *a priori*. Por fim, no Capítulo 5.1 têm-se as considerações do trabalho.

Capítulo 2

Revisão bibliográfica

2.1 Introdução

A análise de regressão visa modelar matematicamente o relacionamento entre uma variável resposta e uma ou mais variáveis explicativas, com o objetivo específico de estimar a variável resposta em função das variáveis explicativas (Barreto, 2011). No caso em que a variável resposta é qualitativa, com dois ou mais valores possíveis, o modelo de regressão logística é frequentemente o mais utilizado para a análise desses dados (Hosmer Jr et al., 2013). Quando a variável resposta apresenta apenas dois valores possíveis denotando "sucesso" ou "fracasso", têm-se o conceito de regressão logística binária; já no caso em que essa variável apresentar três ou mais possibilidades de resposta, têm-se então a regressão logística multinomial.

A regressão logística é usualmente empregada nas mais diversas áreas. Na biomedicina, quando se tem o interesse de modelar a ocorrência de doenças com base em um conjunto de fatores; nas seguradoras é possível identificar fatores de risco em acidentes ou em fraudes de seguros; e também tanto na análise de crédito quanto na formulação de modelos de escore de risco de crédito, foco deste trabalho, que por meio de determinadas características, permite que as instituições financeiras identifiquem possíveis clientes inadimplentes antes de disponibilizarem créditos aos mesmos, ou que permita quantificar o valor ideal de crédito para determinado cliente com base em sua pontuação. Para fins

deste trabalho, o modelo de regressão logística binária será utilizado para a obtenção do ER.

2.1.1 O modelo de regressão logística

Como mencionado acima, existem muitos exemplos do cotidiano em que pode ser feita a aplicação da regressão logística binária. Nesses casos, a resposta é uma variável dicotômica, ou seja, para uma certa característica de interesse, os valores 1 ou 0 representam, respectivamente a presença ou ausência dessa característica. O interesse é poder estimar a probabilidade de que ocorra determinado evento, geralmente o de sucesso.

Seja Y uma variável aleatória que assume valores 1 ou 0, como descrito acima, e um vetor $X = (1, X_1, ..., X_k)$ de k variáveis explicativas. A variável resposta Y segue a distribuição de probabilidade *Bernoulli* com probabilidade de sucesso p e função de probabilidade dada por:

$$P(Y = y) = p^y(1 - p)^{1-y}. \tag{2.1}$$

A média dessa distribuição é definida como $E(Y) = p$ e a variância, $Var(Y) = p(1 - p)$. Na regressão linear, em que a variável resposta é numérica, a associação entre as variáveis explicativas e a variável resposta é feita por meio da média $E(Y)$. Assim, o modelo da regressão linear normal é geralmente $\mu = E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ ou em seu modelo matricial dado por $\mu = X'\beta$, onde $\beta = (\beta_0, \beta_1, ..., \beta_k)$. Note que no modelo de regressão linear a ligação entre o preditor linear $X'\beta$ e o parâmetro μ é obtida por meio da função de ligação identidade. No caso da regressão logística, que tem variável resposta qualitativa binária com distribuição de probabilidade $Bernoulli(p)$, cuja média $E(Y) = p$ se torna limitada, pois nesse caso, $E(Y) = p$ está restrita ao intervalo $[0, 1]$, sendo que o desejado é que o preditor linear continue representando o conjunto dos números reais $\mathbb{R}$. Segundo Hosmer Jr et al. (2013) a quantidade chave é o valor médio da variável resposta dado o valor da variável independente, e essa quantidade é chamada de média condicional expressa como $E(Y|X)$.

Dada a limitação de $E(Y|X) = p(x)$ no modelo *Bernoulli*, é necessário utilizar uma função de ligação $g(.)$ que transforme o intervalo $[0, 1]$ em um intervalo ilimitado. Tendo essa finalidade, existem algumas funções que fazem essa transformação, tais como, logito, probit, complemento log-log, poisson, entre outras. Cada uma dessas funções de ligação corresponde a um tipo de regressão, e no caso da regressão logística, a função $g(.)$ utilizada é a logito:

$$g(p) = log\left(\frac{p(x)}{1 - p(x)}\right) = X'\beta, \tag{2.2}$$

o que resulta em

$$p(x) = \frac{exp(X'\beta)}{1 + exp(X'\beta)}. \tag{2.3}$$

Segundo Hosmer Jr et al. (2013), a importância dessa transformação é que o logito tem muitas das desejáveis propriedades do modelo de regressão linear. Este é linear em seus parâmetros, pode ser contínuo e pode variar de $-\infty$ a ∞, dependendo do intervalo de x. Outra diferença entre o modelo de regressão linear e logístico diz respeito à distribuição condicional da variável resposta. No modelo de regressão linear, a variável resposta pode ser expressa por $y = E(Y|x) + \varepsilon$ sendo o ε o erro associado à media condicional expressando um desvio em relação a mesma. A suposição mais comum é que $\varepsilon \sim Normal(0, \sigma^2)$. Daí, têm-se que a distribuição condicional de $Y|x$ é normal com média $X'\beta$ e variância constante σ^2. No caso de variáveis binárias, a variável resposta é expressa como $y = p(x) + \varepsilon$ e ε pode assumir dois valores (Hosmer Jr et al., 2013):

$$\varepsilon = \begin{cases} 1 - p(x), & Y = 1 \\ -p(x), & Y = 0 \end{cases} \tag{2.4}$$

Na regressão logística, a interpretação dos parâmetros não é feita de forma direta, devido a transformação logarítmica feita no preditor linear. Sendo assim, interpreta-se os parâmetros da seguinte forma: o logito cresce β_1 unidades para cada unidade que a variável explicativa aumenta. Vale ressaltar ainda que essa interpretação só é válida quando a variável explicativa é quantitativa. No caso em que a variável explicativa é

qualitativa binária, o resultado representa o quanto a chance de uma categoria varia em relação a outra (e^{β_1} vezes).

2.1.2 Função de verossimilhança

Seja Y_i a resposta do i-ésimo indivíduo, $i = 1, 2, ..., n$ e x_{ij} a j-ésima variável explicativa do indivíduo i, sendo $j = 1, 2, ..., k$. Assumindo que as observações Y_i são independentes e $Y_i \sim Bernoulli(p_i(x))$, então a função de verossimilhança é dada por:

$$L(\beta) = \prod_{i=1}^{n} [p_i(x)]^{y_i} [1 - p_i(x)]^{1-y_i} . \tag{2.5}$$

A partir das Equações (2.3) e (2.5), temos que a função de verossimilhança para o modelo de regressão logística é:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{exp(X_i'\beta)}{1 + exp(X_i'\beta)}\right]^{y_i} \left[\frac{1}{1 + exp(X_i'\beta)}\right]^{1-y_i}, \tag{2.6}$$

em que $\beta = (\beta_0, \beta_1, ..., \beta_k)$ é o vetor de coeficientes e $X_i = (1, x_{i1}, ..., x_{ik})$ é o vetor de variáveis explicativas do modelo.

2.2 Métricas de performance

Em situações que há proposição de classificação referente a algum modelo, supõe-se que exista uma medida de erro atrelada ao processo, que pode ser quantificada por meio de métodos que mensuram o desempenho de classificação do modelo. Um dos métodos utilizados para essa finalidade é denominado matriz de confusão. Representada por uma tabela de dupla entrada, esta permite conflitar as classificações verdadeiras com as previstas, isto é, as que são realizadas pelo modelo a respeito de um mesmo conjunto de dados. Em um cenário ideal, espera-se que todas as classificações verdadeiras identificadas com valor positivo sejam classificadas pelo modelo também com valor positivo, e que o mesmo ocorra com os valores negativos. Porém, como mencionado acima, existe a possibilidade de ocorrer erros que causam diferenças entre essas classificações. Nesse

contexto, a matriz de confusão visa ser um método prático para visualizar os conflitos entre ambas situações, além de servir como insumo para o cálculo das medidas de desempenho. A Tabela 2.1 apresenta um exemplo de matriz de confusão para uma resposta binária (negativa ou positiva).

Tabela 2.1: Matriz de confusão

		Classificação prevista	
		Negativo	Positivo
Classificação verdadeira	Negativo	A	B
	Positivo	C	D

Nota-se que A é o verdadeiro negativo, B é o falso positivo, C é o falso negativo e D, o verdadeiro positivo. Essa tabela permite identificar quantos acertos e erros a classificação prevista obteve em relação à classificação verdadeira.

Algumas medidas de desempenho podem desenvolvidas, como listado a seguir:

- **Acurácia**

A acurácia, também chamada de taxa de acertos, mede o quanto o modelo acertou na classificação. É definida pela razão de acertos em relação ao total. Em seu resultado final, valores mais altos representam uma maior taxa de acertos. Sua fórmula é dada a seguir:

$$ACUR = \frac{A + D}{A + B + C + D}. \tag{2.7}$$

Porém, nem sempre a acurácia representa a melhor medida de desempenho. Para contornar isso, é necessário complementar com as demais medidas.

- **Sensibilidade**

Também conhecida como a taxa de verdadeiros positivos, a sensibilidade permite observar o seguinte caso: dentre os que tem a classificação verdadeira como positiva, quantos foram classificados como positivos pelo modelo. Assim como a acurácia, valores mais altos de sensibilidade representam melhor qualidade do modelo. A sensibilidade é definida por:

$$SE = \frac{D}{C + D}. \tag{2.8}$$

- **Especificidade**

A especificidade também pode ser chamada de taxa de verdadeiros negativos. E, complementar ao caso anterior, representa a porcentagem de casos verdadeiramente negativos da amostra que foram classificados como negativos pelo modelo. Também, resultados com valores mais altos, evidenciam melhor qualidade do modelo. A especificidade é definida por:

$$ESP = \frac{A}{A + B}. \tag{2.9}$$

- **Taxa de falsos negativos**

A taxa de falsos negativos permite identificar a porcentagem de classificações negativas feitas pelo modelo que, na verdade, são classificadas positivas. Assim como a medida anterior, valores mais baixos melhor indicam melhor qualidade do modelo. A taxa de falsos negativos é definida por:

$$TFN = \frac{C}{A + C}. \tag{2.10}$$

- **Taxa de falsos positivos**

A taxa de falsos positivos representa a porcentagem de classificações positivas feitas pelo modelo que, na verdade, são negativas. Neste caso, valores menores indicam melhor qualidade do modelo. A taxa de falsos positivos é definida por:

$$TFP = \frac{B}{B + D}. \tag{2.11}$$

2.3 Inferência bayesiana

A metodologia bayesiana vem sendo cada vez mais incorporada nos métodos estatísticos, pois sua essência está em poder quantificar sua incerteza inicial e atribuir distribuição de probabilidade para o parâmetro desconhecido θ, dado que neste caso, ao contrário da estatística clássica ou frequentista, assumimos que ele é uma variável aleatória. As áreas de aplicação da abordagem bayesiana são inúmeras, para fins de exemplos associados ao tema deste trabalho, pode-se citar a sua utilização na estimação de riscos operacionais, formulação de modelos para escore de risco de crédito, entre outros.

Na inferência bayesiana têm-se o interesse de estimar o parâmetro desconhecido θ, porém, considerando uma distribuição prévia que possa mensurar a incerteza sobre essa quantidade de interesse. A diferença entre o ponto de vista clássico e o bayesiano consiste em alguns principais pontos, apresentados na Tabela 2.2:

Tabela 2.2: Principais diferenças entre a inferência clássica e bayesiana

	Método clássico	**Método bayesiano**
Parâmetro θ	Fixo	Aleatório
Incerteza inicial	Sabe-se a distribuição dos dados para cada possível valor do parâmetro	Distribuição *a priori* quantifica a incerteza inicial
Estimador de θ	Depende apenas dos dados	A incerteza *a priori* sobre θ é atualizada pelos dados por meio da fórmula de Bayes

Como visto na Tabela 2.2, a principal diferença entre a inferência clássica e bayesiana é a forma de se abordar o parâmetro desconhecido. Na inferência clássica, o parâmetro é considerado fixo (e desconhecido), enquanto que no paradigma bayesiano a incerteza sobre o parâmetro é representado por meio de uma distribuição de probabilidades. Esta abordagem permite que o especialista, no contexto desse trabalho, incorpore uma informação *à priori* ao definir uma distribuição de probabilidades que represente a sua incerteza inicial sobre o parâmetro de interesse.

2.3.1 Distribuição *a priori*

Como citado na subseção anterior, uma das características que difere a abordagem bayesiana da clássica, é a utilização de uma informação ou conhecimento prévio da situação a ser analisada. Esse é um ponto que gera muitas controvérsias e discussões a respeito da metodologia bayesiana, justamente pela *priori* ser algo muito "pessoal", mas há outros que dizem ser a beleza da abordagem bayesiana. De fato, é intuitivo saber que mesmo na inferência clássica ou em investigações científicas de modo geral, é implícito que convicções *à priori* são utilizadas. Algum conhecimento anterior é usado para formular uma função de verossimilhança adequada.

2.3.2 Distribuição *a posteriori*

Em termos mais voltados para este contexto, considere θ uma quantidade desconhecida e $x_1, ..., x_n$ uma amostra que, dado θ, são independentes e identicamente distribuídas (i.i.d.) segundo uma função (densidade) de probabilidades $\pi(x|\theta)$. Deseja-se saber a distribuição de probabilidades de θ dado $x = (x_1, ..., x_n)$. Então, $\pi(\theta|x)$ é a quantidade de interesse nessa análise, porém, antes mesmo de se observar os dados, considera-se a distribuição *à priori* de θ, $\pi(\theta)$ e a função de verossimilhança dos dados, $l(x;\theta)$. Desta forma, pela fórmula de Bayes, a distribuição *à posteriori* de $\theta|X$ é expressa por:

$$\pi(\theta|x) = \frac{l(x;\theta)\pi(\theta)}{\pi(x)} = \frac{l(x;\theta)\pi(\theta)}{\int \pi(\theta;x)d\theta}. \tag{2.12}$$

Note que $1/\pi(x)$, onde $\pi(x) = \int \pi(\theta;x)d\theta$ é a distribuição marginal de x, que não depende de θ e funciona como uma constante normalizadora de $\pi(\theta|x)$. Desta forma, é possível reescrever a Equação (2.12) de forma que a integral sobre os dados possa ser "omitida" e considerando, em seu lugar, o símbolo de proporcional, isto é,

$$\pi(\theta|x) \propto l(x;\theta)\pi(\theta). \tag{2.13}$$

Uma tradução da equação anterior pode ser considerada da seguinte maneira: a distribuição *a posteriori* é proporcional à verossimilhança dos dados multiplicada pela *priori*. Deixando mais claro ainda, essa equação nada mais é do que uma atualização da *priori* utilizando a informação retratada pelos dados (a verossimilhança).

2.3.3 Estimação pontual

A distribuição *a posteriori* é o resumo completo da inferência sobre um parâmetro θ. No entanto, para algumas aplicações, é desejável (ou necessário) resumir essa informação de alguma maneira. Esse resumo é chamado de estimativa pontual e toda a *posteriori* é resumida em um valor para cada parâmetro (Ehlers, 2007).

A estimativa pontual na inferência bayesiana é obtida utilizando o auxílio da função perda, definida como $L(a, \theta)$, onde a denota uma estimativa para θ. A função perda, assim como o nome sugere, representa a "perda" de uma estimativa quando esta tem valor diferente do verdadeiro valor do parâmetro desconhecido, e pode ser definida por diferentes funções. As principais funções de perda são:

- Perda quadrática: $L(a, \theta) = (\theta - a)^2$;
- Perda absoluta: $L(a, \theta) = |\theta - a|$;
- Perda 0-1:

$$L(a, \theta) = \begin{cases} 1, & |a - \theta| > \varepsilon \\ 0, & |a - \theta| < \varepsilon, \end{cases}$$

para $\varepsilon \to 0$.

Segundo Ehlers (2007), o risco de uma regra de decisão, denotado por $R(a)$, é a perda esperada *a posteriori*, isto é,

$$R(a) = E_{\theta|x}[L(a, \theta)] = \int_{\Theta} L(a, \theta)\pi(\theta|x)d\theta. \tag{2.14}$$

Assim, o estimador de Bayes é o valor de a que minimiza $R(a)$. Segue abaixo os estimadores de Bayes considerando as funções de perda quadrática, absoluta e 0-1:

- Perda quadrática: a média *a posteriori*;
- Perda absoluta: a mediana *a posteriori*;
- Perda 0-1: a moda *a posteriori*.

2.3.4 Estimação intervalar

A estimação pontual permite que se resuma toda a informação da distribuição *a posteriori* em um único valor do parâmetro. Porém, isto não define precisão na estimativa. Os intervalos de credibilidade, ou intervalos de confiança bayesianos, têm o objetivo de fornecer essa precisão à estimativa por meio da distribuição *a posteriori*. Segundo Ehlers (2007), os intervalos de credibilidade são definidor por:

Definição 3.1 *C é um intervalo de credibilidade de* $100(1-\alpha)\%$*, ou nível de credibilidade* $1-\alpha$*, para* θ *se* $P(\theta \in C) \geq 1-\alpha$.

De acordo com essa definição e como o parâmetro θ é uma variável aleatória, pode-se atribuir uma probabilidade do parâmetro θ pertencer ao intervalo com $100(1-\alpha)\%$ de credibilidade. Sendo assim, a interpretação é de que existe uma probabilidade de $1-\alpha$, com base na distribuição *a posteriori*, que θ esteja contido na região C ou região de credibilidade $100(1-\alpha)\%$ para θ, como na equação a seguir:

$$\int_C \pi(\theta|x)d\theta = 1-\alpha. \tag{2.15}$$

A Equação (2.15) permite criar uma variedade de intervalos. Qualquer região com probabilidade $1-\alpha$ é um intervalo válido. Sabendo-se que o desejo é encontrar um intervalo com menor amplitude possível, o que define a variabilidade e precisão do parâmetro, houve a necessidade de definir um intervalo mínimo, o que ocorre quando os valores de θ geram a máxima densidade *a posteriori* (Cella, 2013). Esses intervalos são denominados *Highest Poterior Density* (HPD) ou em português Máxima Densidade *a Posteriori*, como

definido por Ehlers (2007):

Definição 3.2 *Um intervalo de credibilidade C de* $100(1-\alpha)\%$ *para* θ *é de máxima densidade a posteriori (HPD) se* $C = \{\theta \in \Theta : \pi(\theta|x) \geq k(\alpha)\}$ *onde* $k(\alpha)$ *é a maior constante tal que* $P(\theta \in C) \geq 1-\alpha$.

Com essa definição, todos os pontos dentro do intervalo HPD terão densidade *a posteriori* maior ou igual do que qualquer ponto fora do intervalo.

2.4 Modelos logísticos bayesianos

A estimação dos parâmetros β do modelo de regressão logística, em geral, é realizada utilizando métodos conhecidos da abordagem clássica, dentre estes a Estimação por Máxima Verossimilhança (EMV) e o Método dos Mínimos Quadrados (MMQ). Porém, como já mencionado na Seção 2.3, ao utilizar a métodos bayesianos para realizar as estimativas, os parâmetros são definidos como variáveis aleatórias e, neste caso, podem lhe ser atribuídas distribuições de probabilidade, ou seja, se β é o vetor de parâmetros de interesse obtido na regressão logística, $\pi(\beta)$ é a sua distribuição de probabilidade *a priori*.

A especificação da distribuição *a priori* para os coeficientes β da regressão logística binária é, de fato, uma tarefa não trivial. Na literatura, há diversas abordagens sobre a escolha de uma *priori* mais adequada. Segundo Pires (2010), a abordagem padrão consiste em assumir uma distribuição normal ou difusa, isto é, $\pi(\beta) \propto 1$. Dessa forma, a distribuição *a posteriori* é simplesmente proporcional a verossimilhança $l(x;\beta)$ (Wakefield, 2013). Adotando a função de verossimilhança especificada na Equação (2.6), e considerando a *priori* $\pi(\beta) \propto 1$, tem-se a distribuição *a posteriori* dada por:

$$\pi(\beta|x) \propto \prod_{i=1}^{n} \left[\frac{exp(X_i'\beta)}{1+exp(X_i'\beta)}\right]^{y_i} \left[\frac{1}{1+exp(X_i'\beta)}\right]^{1-y_i}, \tag{2.16}$$

em que $\beta = (\beta_0, \beta_1, ..., \beta_k)$ é o vetor de coeficientes e $X_i = (1, x_{i1}, ..., x_{ik})$ é o vetor de

variáveis explicativas do modelo.

Considerando uma *priori* com distribuição normal multivariada para $\beta = (\beta_0, \beta_1, ..., \beta_k)$, isto é, $\beta \sim N_{k+1}(\mu_0, \Sigma_0)$ em que μ_0 e Σ_0 são hiperparâmetros conhecidos, a distribuição *a posteriori* é dada por:

$$\pi(\beta|x) \propto exp\left(-\frac{1}{2}(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)\right) \times \prod_{i=1}^{n}\left[\frac{exp(X_i'\beta)}{1+exp(X_i'\beta)}\right]^{y_i}\left[\frac{1}{1+exp(X_i'\beta)}\right]^{1-y_i}, \tag{2.17}$$

em que $X_i = (1, x_{i1}, ..., x_{ik})$ é o vetor de variáveis explicativas do modelo.

As *posterioris* acima não podem ser obtidas analiticamente, mas os seus valores podem ser gerados numericamente. Neste trabalho, essas *posterioris* serão obtidas por meio do pacote *rstanarm* no software R (R Core Team, 2022).

Capítulo 3

Metodologia

3.1 Introdução

O conceito de crédito pode ser analisado sob diversas perspectivas. Para uma instituição financeira, crédito refere-se, principalmente, à atividade de colocar um valor à disposição de um tomador de recursos sob a forma de um empréstimo ou financiamento, mediante compromisso de pagamento em uma data futura (Brito and Assaf Neto, 2008).

Segundo Machado (2015), o crédito corresponde a um valor monetário disponibilizado ao tomador de recursos financeiros, em forma de empréstimo ou financiamento, por um período previamente pactuado, com a promessa de pagamento futuro, ao qual é acrescido uma remuneração, denominada juros.

Sob a ótica financeira, a concessão de crédito por parte das instituições financeiras contribui positivamente para a economia local, pois estimula a movimentação de capital e possibilita que o comprador faça maiores aquisições que, sem a disponibilização do crédito, não poderiam ser feitas. Porém, os riscos atrelados a essa concessão de crédito precisam ser quantificados. Sendo assim, as métricas de avaliação de risco associado à disponibilização do crédito são essenciais para avaliar a viabilidade de conceder ou não o crédito à determinado cliente.

Antes da existência de ferramentas tecnológicas que permitissem a classificação dos clientes como sendo de alto ou baixo risco à instituição financeira, a atribuição de confian-

ça era tomada de forma subjetiva, podendo resultar uma avaliação injusta ou tendenciosa. Com o surgimento dos cartões de crédito, computadores, tecnologias e dados de insumo, a evolução do mapeamento de avaliação e, posteriormente, a concessão ou não de crédito, foram evoluindo para modelos mais objetivos.

Sendo assim, a medida em que as trocas e empréstimos foram tomando uma maior proporção e se popularizando, houve também a necessidade de criar modelos que quantificassem o risco das concessões e transações de crédito, ou que fornecessem informações prévias para uma tomada de decisão. Nos dias atuais, os modelos utilizam como insumo dados das instituições financeiras para calcular uma pontuação, ou seja, um escore de risco de crédito para cada cliente, de forma que outras instituições ou demais interessados possam consultar e tomar a decisão de disponibilizar o crédito, tendo a informação prévia a respeito do perfil do cliente.

3.2 O Escore de Risco

A mensuração de risco de crédito é o processo de quantificar a credibilidade de um solicitante de crédito, por meio de variáveis que irão classificar os cliente entre bons ou maus pagadores. O objetivo dessa classificação é poder prever comportamentos que possam indicar padrões de inadimplência, e assim, evitar em maiores prejuízos para a instituição financeira.

O Escore de Risco (ER) é obtido utilizando as estimativas bayesianas realizadas nos parâmetros da regressão logística atribuindo-lhes uma distribuição *a priori* informada pelo especialista, que pode ser, inclusive, não informativa. Como definido no Capítulo 2 deste trabalho, a regressão logística tem como objetivo estimar uma variável resposta qualitativa em função das variáveis explicativas. Sendo assim, a variável resposta em questão é definida como:

$$Y = \begin{cases} 1, & se\ o\ cliente\ é\ inadimplente \\ 0, & se\ o\ cliente\ é\ adimplente. \end{cases} \tag{3.1}$$

Desta forma, o ER pode ser definido como o preditor linear de um modelo de regressão logística (Machado, 2015), isto é,

$$ER = X'\beta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k. \tag{3.2}$$

O valor do ER definido na Equação (3.2) pode ser visto como um Escore de Risco pelo fato do mesmo estar associado à uma probabilidade e o logito ser uma função monótona e crescente. De fato, quanto maior o valor do ER, maior é a probabilidade de um cliente ser considerado inadimplente $(Y = 1)$.

3.3 Classificação dos clientes por meio do ER

Dado o interesse deste trabalho, em propor um modelo de risco de crédito que permita classificar os clientes de acordo com o risco de crédito atrelado a operação, existe a necessidade de avaliar o quão confiável é o modelo. Tendo essa finalidade, será aplicado o método de *cross-validation*, em que serão utilizadas as métricas apresentadas na Seção 2.2. Aplicando os conceitos da matriz de confusão ao contexto do ER, temos as seguintes definições:

Tabela 3.1: Matriz de confusão do modelo de risco de crédito

		Classificação prevista	
		Baixo risco	Alto risco
Classificação verdadeira	Adimplente	A	B
	Inadimplente	C	D

- **Acurácia (total de acertos):** mede o percentual de clientes que obtiveram classificação prevista igual à classificação verdadeira, em relação ao total de clientes (Expressão 2.7);
- **Sensibilidade (de alto risco):** mede o percentual de clientes inadimplentes que obtiveram classificação prevista de alto risco pelo modelo (Expressão 2.8);

- **Especificidade:** mede o percentual de clientes adimplentes que obtiveram classificação prevista de baixo risco pelo modelo (Expressão 2.9);

As métricas de performance que quantificam erros de classificação serão denominados neste trabalho de Erro I e Erro II, sendo estes o foco em que há interesse em controlá-los devido o prejuízo que podem gerar às instituições:

- **Erro I (Taxa de falsos negativos):** mede o percentual de clientes inadimplentes dentre todos aqueles classificados como de baixo risco pelo modelo (Expressão 2.10).

- **Erro II (Taxa de falsos positivos):** mede o percentual de clientes adimplentes dentre todos aqueles classificados como de alto risco pelo modelo (Expressão 2.11).

Na prática, os erros de classificação geram prejuízos as instituições financeiras, e o interesse é que estes sejam minimizados. A consequência do Erro I é conceder crédito a um cliente mal pagador, enquanto que no Erro II, a consequência é deixar de conceder crédito para um cliente bom pagador. Para a finalidade deste trabalho, o Erro I é apontado como sendo mais grave, e por este motivo, objetiva-se conter esse erro. Assim, um ponto de corte pode ser definido de forma controlar a probabilidade (estimada) de ocorrência do Erro I.

Regra de decisão binária: Um cliente será considerado de *alto risco*, se $ER > K$ e será considerado de *baixo risco*, se $ER <= K$, em que K é o ponto cuja classificação desses clientes resulta em uma probabilidade de Erro I de α (α é o erro máximo suportado pela instituição).

Note que a regra de decisão binária pode ser estendida para três ou mais classificações. No caso de três possíveis classificações (baixo, médio e alto risco), têm-se a seguinte regra de decisão:

Regra de decisão para três classificações: Um cliente será considerado de *alto risco* se $ER > K_2$; será classificado como de *médio risco* se $K_1 \leq ER \leq K_2$; e como de *baixo*

risco se $ER < K_1$. Em que K_1 é o ponto cuja classificação desses clientes resulta em uma probabilidade de Erro I de α_1 e K_2 é o ponto cuja classificação desses clientes resulte em uma probabilidade de Erro II de α_2. Aqui, α_1 e α_2 são os erros máximos suportados pela instituição.

3.4 Regra de decisão bayesiana para classificação dos clientes

Nesta seção, a inferência bayesiana é utilizada para atualizar as probabilidades, considerando a *priori* informada pelo especialista, e recategorizar, *a posteriori*, os clientes nos grupos de risco, denominados adimplente e inadimplente.

Segundo a metodologia proposta por Nakano and Pereira (2010), considere uma amostra de n clientes oriundos de k grupos denotados por $E_0, E_1, ..., E_{k-1}$. Para cada cliente, o ER é calculado e, conforme seu resultado, o cliente é classificado em uma das m categorias, $C_1, C_2, ..., C_m$. É importante enfatizar que o número de categorias e grupos não necessariamente precisam ter a mesma quantidade de classes. Inicialmente, definimos as probabilidades *a priori* de um cliente qualquer pertencer a um dos k grupos por

$$P(E_i) = \theta_i, \tag{3.3}$$

em que $0 < \theta_i < 1$ e $\sum_{i=0}^{k-1} \theta_i = 1$.

Uma forma simples de fazer a inferência (pontual) é atribuirmos *a priori*, probabilidades do cliente pertencer a um dos k grupos e atualizá-las pelos resultados dos ER através da fórmula de Bayes (Meyer, 1983), obtendo assim, as probabilidades *a posteriori* de classificação.

Seja $P(C_j|E_i)$ a probabilidade de um cliente apresentar um ER que o classifique na Categoria C_j, dado que ele inicialmente pertence ao Grupo E_i, isto é,

$$P(C_j|E_i) = \frac{número\ de\ clientes\ do\ grupo\ E_i,\ que\ tiveram\ o\ ER\ classificado\ em\ C_j}{número\ de\ clientes\ do\ grupo\ E_i}, \tag{3.4}$$

em que $i = 0, 1, ..., k-1$ representa o grupo e $j = 1, 2, ..., m$, a categoria do ER. As probabilidades resultantes da Equação (3.4) são denominadas verossimilhança dos dados.

Pela fórmula de Bayes, obtemos a probabilidade *(a posteriori)* de um cliente pertencer ao Grupo i, dada a classificação de seu ER.

$$P(E_i|C_j) = \frac{P(E_i \cap C_j)}{P(C_j)} = \frac{P(E_i)P(C_j|E_i)}{\sum_{r=0}^{k-1} P(E_r)P(C_j|E_r)} \tag{3.5}$$

Neste estudo, serão considerados $m = 3$ e $k = 2$, isto é, os clientes são, *a priori*, alocados em três categorias (baixo, médio e alto risco) por meio do ER e, posteriormente, serão classificados, conforme resultado do cálculo da probabilidade *a posteriori*, em dois grupos de risco, adimplente e inadimplente.

Capítulo 4

Resultados obtidos

4.1 Descrição do problema

Este capítulo tem como intuito aplicar a metologia proposta em um conjunto de dados, a fim de melhor ilustrar os procedimentos que foram desenvolvidos. O ER, descrito no Capítulo 3, é utilizado nessa fase do estudo para ajustar as probabilidades de interesse, isto é, as probabilidades de um cliente ser classificado em uma das m categorias $(C_1, ..., C_m)$ dado que são provenientes em k grupos $E_0, E_1, ..., E_{k-1}$. Considerando inicialmente $m = 3$ categorias e $k = 2$ grupos, têm-se a seguinte distribuição *a priori* $\pi(\theta) = \pi(\theta_0, \theta_1)$, associada aos grupos:

θ_0 = P(cliente ser adimplente)

θ_1 = P(cliente ser inadimplente)

Essas probabilidades representam a opinião do especialista respaldada em sua experiência. Desta forma, a opinião será calibrada pelos conhecimentos adquiridos com o cálculo do ER a partir das informações obtidas na amostra. O ER calculado para cada um dos clientes é, então, adotado para a classificação dos mesmos em $m = 3$ categorias:

- C_1: Baixo risco;
- C_2 Médio risco;
- C_3 Alto risco;

sendo que seus respectivos pontos de corte serão conjecturados de acordo com a regra de decisão definida na Seção 3.3. Após a alocação dos clientes nessas três categorias, as probabilidades *a posteriori* são obtidas a partir das probabilidades *a priori*, informadas pelo especialista, juntamente com a verossimilhança dos dados. Os resultados dessas probabilidades *a posteriori* permitem realocar os clientes nos grupos $E = 0$ (adimplente) ou $E = 1$ (inadimplente).

4.2 Dados para ilustração da metodologia

A metodologia proposta neste trabalho será ilustrada utilizando uma base de dados disponibilizada por Dua and Graff (2017) do Center for Machine Learning and Intelligence Systems (CML). O CML está localizado na Universidade da California, Irvine, e detém de uma grande quantidade de bases de dados tendo como finalidade serem insumos para a realização de estudos e pesquisas em diferentes áreas.

O *German Credit Data* é um importante conjunto de dados, muito utilizado em estudos sobre risco de credito e também será estudado nesta dissertação. A base de dados apresenta 1000 linhas de solicitantes de crédito e 21 variáveis que permitem uma melhor descrição de cada caso. Para a aplicação do modelo proposto, foram selecionadas, dentre as 21 variáveis, a variável resposta além de 12 variáveis explicativas. Estas ultimas foram triadas de acordo com seu respectivo assunto e de forma que fizessem sentido ao modelo.

Nos dados originais, os nomes das variáveis estão codificados, porém, a partir do dicionario de dados é possível identificar o assunto que contempla cada uma. Sendo assim, a renomeação proposta, para maior entendimento, é apresentada na Tabela 4.1.

Tabela 4.1: Correspondência entre os códigos e os nomes das variáveis dos dados

Variáveis originais	**Variáveis renomeadas**	**Códigos das variáveis no modelo**
V3	**cod_historico_credito**	**y**
V1	status_conta_existente	x_1
V9	sexo_estado_civil	x_2
V13	idade_anos	x_3
V15	tipo_casa	x_4
V17	emprego	x_5
V11	tempo_residencia_atual	x_6
V2	duracao_meses	x_7
V5	valor_total_creditos	x_8
V6	valor_poupanca	x_9
V18	qtd_dependentes	x_{10}
V16	qtd_creditos_no_banco	x_{11}
V14	outros_planos_parcelamento	x_{12}

A partir dessa seleção de variáveis, a Tabela 4.3 apresenta a descrição, o tipo da variável e, caso esta seja do tipo categórica, suas respectivas categorias.

Tabela 4.2: Descrição das variáveis selecionadas

Variável	**Descrição**	**Tipo**	**Categorias**
cod_historico_credito	Variável resposta	Categórica	$y = 0$: Adimplente $y = 1$: Inadimplente
status_conta_existente	*Status* da conta existente	Categórica	$x_{1.1}$: Sem conta $x_{1.2}$: DM[1]< 0 $x_{1.3}$: $0 \leq DM < 200$ $x_{1.4}$: $DM \geq 200$
sexo_estado_civil	Sexo e estado civil do solicitante de credito	Categórica	$x_{2.1}$: homem divorciado / separado / casado $x_{2.2}$: homem solteiro $x_{2.3}$: mulher divorciada / separada / casada $x_{2.4}$: mulher solteira
idade_anos	Idade em anos	Numérica	x_3
tipo_casa	Tipo de moradia	Categórica	$x_{4.1}$: Alugada $x_{4.2}$: Própria $x_{4.3}$: Mora de favor
emprego	Tipo de emprego	Categórica	$x_{5.1}$: Desempregado $x_{5.2}$: Empregado informal $x_{5.3}$: Emprego formal $x_{5.4}$: Empresário autônomo
tempo_residencia_atual	Tempo morando na residência atual	Numérica	x_6
duracao_meses	Duração do crédito solicitado em meses	Numérica	x_7
valor_total_creditos	Valor total do crédito solicitado	Numérica	x_8
valor_poupanca	Valor em categorias guardado em poupança	Categórica	$x_{9.1}$: $DM < 100$ $x_{9.2}$: $100 \leq DM < 500$ $x_{9.3}$: $500 \leq DM < 1000$ $x_{9.4}$: $DM \geq 1000$ $x_{9.5}$: Sem conta poupança
qtd_dependentes	Quantidade de dependentes	Numérica	x_{10}
qtd_creditos_no_banco	Quantidade de vezes que solicitou crédito ao banco	Numérico	x_{11}
outros_planos_parcelamento	Outros planos de parcelamento	Categórica	$x_{12.1}$: Banco $x_{12.2}$: Loja $x_{12.3}$: Nenhum

[1] DM é um valor de referência salarial.

A Tabela 4.3 apresenta a distribuição de frequências das variáveis categóricas utilizadas pelo modelo, assim como a Tabela 4.4 apresenta as estatísticas descritivas das variáveis numéricas.

Tabela 4.3: Distribuição de frequências das variáveis categóricas

Variável	**Categorias**	**Frequência**	**Percentual** $(\%)$
Histórico de crédito (Variável resposta)	$y = 0$: Adimplente	707	70,7
	$y = 1$: Inadimplente	293	29,3
Status da conta existente	$x_{1.1}$: Sem conta	394	39,4
	$x_{1.2}$: $DM < 0$	274	27,4
	$x_{1.3}$: $0 \leq DM < 200$	269	26,9
	$x_{1.4}$: $DM \geq 200$	63	6,3
Sexo e estado civil do solicitante de credito	$x_{2.1}$: homem divorciado / separado / casado	142	14,2
	$x_{2.2}$: homem solteiro	548	54,8
	$x_{2.3}$: mulher divorciada / separada / casada	310	31,0
	$x_{2.4}$: mulher solteira	0	0
Tipo de moradia	$x_{4.1}$: Alugada	179	17,9
	$x_{4.2}$: Própria	713	71,3
	$x_{4.3}$: Mora de favor	108	10,8
Tipo de emprego	$x_{5.1}$: Desempregado	22	2,2
	$x_{5.2}$: Empregado informal	200	20,0
	$x_{5.3}$: Emprego formal	630	63,0
	$x_{5.4}$: Empresário autônomo	148	14,8
Valor em categorias guardado em poupança	$x_{9.1}$: $DM < 100$	603	60,3
	$x_{9.2}$: $100 \leq DM < 500$	103	10,3
	$x_{9.3}$: $500 \leq DM < 1000$	63	6,3
	$x_{9.4}$: $DM \geq 1000$	48	4,8
	$x_{9.5}$: Sem conta poupança	183	18,3
Outros planos de parcelamento	$x_{12.1}$: Banco	139	13,9
	$x_{12.2}$: Loja	47	4,7
	$x_{12.3}$: Nenhum	814	81,4

Tabela 4.4: Estatísticas descritivas das variáveis numéricas

Variável	Min	Mediana	Média	Max	DP
Idade em anos	19	33	35,54	75	11,37
Tempo de moradia	1	3	2,84	4	1,10
Duração do credito (meses)	4	18	20,90	72	12,06
Valor do credito	250	2319,50	3271.258	18424	2822,74
Quantidade de dependentes	1	1	1,15	2	0,36
Quantidade de créditos solicitados	1	1	1,41	4	0,58

4.3 Ajustes do modelo logístico bayesiano

A Tabela 4.5 apresenta as estimativas (médias *a posteriori*) dos coeficientes $(\hat{\beta})$ do modelo de regressão logística com suas respectivas estimativas dos desvios padrões $(\hat{\sigma})$, assim como o intervalo com 95% de credibilidade considerando as variáveis da Tabela 4.2. Essas estimativas consideraram *prioris* independentes $\beta \sim N_{k+1}(0;10^4$ I) (Equação 2.17), e foram obtidas por meio de uma amostra MCMC de tamanho 8.000 (4 cadeias de 2.000 réplicas).

Tabela 4.5: Estimativas bayesianas dos coeficientes do modelo logístico múltiplo

Variável	$\hat{\beta}$	$\hat{\sigma}$	**LI**[1]**(IC** 95%**)**	**LS**[2]**(IC** 95%**)**
(Intercept)	-5,674	0,989	-7,658	-3,786
$x_{1.1}$	0	-	-	-
$x_{1.2}$	-0,615	0,225	-1,053	-0,184
$x_{1.3}$	-0,734	0,219	-1,181	-0,325
$x_{1.4}$	-0,165	0,372	-0,909	0,529
$x_{2.1}$	0	-	-	-
$x_{2.2}$	0,295	0,278	-0,212	0,856
$x_{2.3}$	0,228	0,295	-0,324	0,817
x_3	0,019	0,008	0,003	0,036
$x_{4.1}$	0	-	-	-
$x_{4.2}$	0,500	0,265	-0,011	1,025
$x_{4.3}$	0,081	0,397	-0,686	0,860
$x_{5.1}$	0	-	-	-
$x_{5.2}$	0,302	0,700	-1,062	1,688
$x_{5.3}$	0,268	0,688	-1,031	1,652
$x_{5.4}$	0,366	0,716	-0,997	1,819
x_6	0,143	0,087	-0,027	0,315
x_7	-0,014	0,010	-0,033	0,006
x_8	-0,00001	0,00004	-0,0001	0,0001
$x_{9.1}$	0	-	-	-
$x_{9.2}$	-0,622	0,327	-1,286	-0,017
$x_{9.3}$	-0,064	0,368	-0,788	0,627
$x_{9.4}$	0,301	0,382	-0,491	1,026
$x_{9.5}$	-0,064	0,234	-0,527	0,386
x_{10}	-0,341	0,254	-0,868	0,161
x_{11}	2,250	0,170	1,916	2,573
$x_{12.1}$	0	-	-	-
$x_{12.2}$	-0,501	0,513	-1,540	0,457
$x_{12.3}$	0,627	0,273	0,093	1,171

[1] LI: limite inferior do intervalo de credibilidade (quantil 2,5%);
[2] LS: limite superior do intervalo de credibilidade (quantil 97,5%).

Assim, conforme as estimativas apresentadas na Tabela 4.5 e partindo da Equação (3.2), pode-se calcular o ER de um cliente. Para efeito de ilustração, considere, por exemplo, um cliente sem conta ($x_{1.1} = 1$), sendo este um homem solteiro ($x_{2.2} = 1$), com 30 anos de idade ($x_3 = 30$), que mora de favor ($x_{4.3} = 1$), com emprego formal ($x_{5.3} = 1$), morando na residência atual há 2 anos ($x_6 = 2$), solicitando credito com duração de 6

meses ($x_7 = 6$). Suponha também que o cliente tenha solicitado crédito no valor de 20.000 unidades ($x_8 = 20.000$), que não tenha conta poupança ($x_{9.5} = 1$), que apresente dois dependentes ($x_{10} = 2$), que nunca havia solicitado crédito ao banco anteriormente ($x_{11} = 0$) e sem outros planos de parcelamento ($x_{12.3} = 1$). O ER desse cliente será dado por:

$$\begin{aligned} ER = -5,674 + (0 \times 1) + 0,295 + (30 \times 0,019) + 0,081 + 0,268 \\ +(2 \times 0,143) - (6 \times 0,014) - (20.000 \times 0,00001) \\ -0,064 - (2 \times 0,341) + (0 \times 2,250) + 0,627 = -4,577. \end{aligned} \tag{4.1}$$

A partir do ER, é necessário aplicar a regra de decisão que permite a alocação desses clientes nas $m = 3$ categorias (alto, médio ou baixo risco). A designação é feita pelos pontos de corte que podem estar atrelados a quaisquer métricas associadas à matriz de confusão (Seção 3.3), porém, neste caso, utilizou-se os Erros I e II para tal definição de corte. Para fins de análise, a Tabela 4.7 apresenta as possíveis alocações dos clientes nas três categorias definindo os pontos de corte que apresentam probabilidade do Erro I ser menor que 10% e que controle a probabilidade do Erro II em menos que 20%. Para a obtenção desses resultados, os pontos de corte são determinados da seguinte forma:

Cortes que minimizam a probabilidade do Erro I em 10$\%$ e controlam a probabilidade do Erro II em 20$\%$:

- C_1: Baixo risco, se $ER <$ -1,10 ;
- C_2: Médio risco, se -1,10 $\leq ER \leq$ 1,21;
- C_3: Alto risco, se $ER >$ 1,21;

Os resultados mostrados na Tabela 4.6 representam a verossimilhança dos dados. Tomando-a como referência, verifica-se que a probabilidade de um cliente ser alocado na categoria de baixo risco (C_1), dado que ele é adimplente (E_0), é 0,7737. Da mesma

forma, a probabilidade deste ser alocado na categoria de alto risco (C_3) dado que ele é inadimplente (E_1) é 0,7474.

Tabela 4.6: Classificação do ER minimizando a probabilidade do Erro I em 10% e controlando a probabilidade do Erro II em 20% com ênfase na verossimilhança

Categorias	**Grupos**		**Total**
	Adimplente (E_0)	**Inadimplente** (E_1)	
Baixo risco (C_1)	547 (77,37$\%$)	60 (20,48$\%$)	607 (60,7$\%$)
Médio risco (C_2)	151 (4,52$\%$)	194 (4,78$\%$)	345 (34,5$\%$)
Alto risco (C_3)	9 (18,10$\%$)	39 (74,74$\%$)	48 (4,8$\%$)
Total	707 (100$\%$)	293 (100$\%$)	1000 (100$\%$)

Nota-se que a Tabela 4.7 apresenta as mesmas frequências ta Tabela 4.6, porém, ao calcular os percentuais colunas, facilmente pode-se encontrar as probabilidades dos Erros I e II definidos para gerar os cortes das categorias. Sendo assim, ao verificar a probabilidade de uma cliente ser alocado na categoria de baixo risco (C_1) dado que ele é inadimplente(E_1), configura o Erro I, com probabilidade $0,0988$, sendo, de fato, limitado superiormente a 10%. Da mesma maneira, pode-se identificar a probabilidade do Erro II ao alocar um cliente na categoria de alto risco (C_3) dado que ele é adimplente (E_0), com probabilidade de $0,1875$. Este resultado também corrobora com a definição de controlar a probabilidade do Erro II em 20%. Observe que ambas as situações estão grifadas na Tabela 4.7:

Tabela 4.7: Classificação do ER minimizando a probabilidade do Erro I em 10% e controlando a probabilidade do Erro II em 20% com ênfase na identificação dos Erros I e II

Categorias	**Grupos**		**Total**
	Adimplente (E_0)	**Inadimplente** (E_1)	
Baixo risco (C_1)	547 (90,12$\%$)	**60 (9,88$\%$)**	607 (100$\%$)
Médio risco (C_2)	151 (43,77$\%$)	194 (56,23$\%$)	345 (100$\%$)
Alto risco (C_3)	**9 (18,75$\%$)**	39 (81,25$\%$)	48 (100$\%$)
Total	707 (70,7$\%$)	293 (29,3$\%$)	1000 (100$\%$)

Após ter-se efetuado o cálculo da verossimilhança, necessita-se da definição da *priori* a ser utilizada para concretizar o cômputo das probabilidades *a posteriori* de um cliente ser inadimplente (ou adimplente) conforme sua classificação dada pelo ER. Neste estudo, as

probabilidades *à priori* retratam o conhecimento do especialista a cerca da inadimplência de um cliente.

Inicialmente, serão considerados dois cenários, fixando no primeiro a probabilidade de 0,5 para ambos os grupos (inadimplentes e adimplentes); e, por conseguinte, 0,707 e 0,293 para os grupos adimplente e inadimplente, respectivamente. O primeiro exercício, em que os dois grupos apresentam a mesma probabilidade, caracteriza a *priori* não informativa. Já o segundo representa o percentual observado de cada grupo na amostra, de acordo com a classificação verdadeira. Conforme os procedimentos descritos na Seção 3.4, é possível obter as probabilidades *a posteriori* dos eventos de interesse. As Tabelas 4.8 e 4.9 apresentam esses resultados:

Tabela 4.8: Probabilidades *a posteriori* da classificação dos clientes considerando uma *priori* não informativa

Categorias	**Grupos**	
	Adimplente (E_0)	**Inadimplente** (E_1)
priori	0,5	0,5
Baixo risco (C_1)	0,7907	0,2093
Médio risco (C_2)	0,4860	0,5140
Alto risco (C_3)	0,1950	0,8050

Tabela 4.9: Probabilidades *a posteriori* da classificação dos clientes considerando como *priori* o percentual observado na amostra

Categorias	**Grupos**	
	Adimplente (E_0)	**Inadimplente** (E_1)
priori	0,707	0,293
Baixo risco (C_1)	0,9011	0,0989
Médio risco (C_2)	0,6953	0,3047
Alto risco (C_3)	0,3688	0,6312

É possível observar a partir das Tabelas 4.8 e 4.9 que, apesar das probabilidades *a posteriori* se modificarem conforme a escolha da *priori*, ambas apresentam uma alta probabilidade *a posteriori* de um cliente ser adimplente quando o ER o classifica como baixo risco (C_1) e uma alta probabilidade *a posteriori* de inadimplência quando o ER classifica um cliente como alto risco (C_3). Esses resultados indicam um bom desempenho do modelo de risco proposto.

Devido as disparidades observadas entre as duas *prioris* adotadas, nota-se a importância de uma opinião devidamente advinda de um especialista que possa utilizar esse recurso para gerar resultados mais fidedignos a realidade dos clientes, e consequentemente, gerando previsões mais precisas.

Para melhor compreender o efeito da escolha da *priori* nas probabilidades *a posteriori*, a Tabela 4.10 e a Figura 4.1 apresentam as probabilidades *a posteriori* segundo a classificação dos clientes pelo ER, variando-se a *priori*.

Tabela 4.10: Probabilidades *a posteriori* de inadimplência para cada categoria de risco conforme escolha da *priori*

Priori	Baixo risco (C1)	Médio risco (C2)	Alto risco (C3)
0,05	0,0137	0,0527	0,1785
0,10	0,0286	0,1051	0,3145
0,15	0,0446	0,1573	0,4215
0,20	0,0621	0,2091	0,5080
0,25	0,0811	0,2606	0,5792
0,30	0,1019	0,3119	0,6389
0,35	0,1248	0,3628	0,6898
0,40	0,1500	0,4135	0,7335
0,45	0,1780	0,4639	0,7716
0,50	**0,2093**	**0,5140**	**0,8050**
0,55	0,2444	0,5638	0,8346
0,60	0,2842	0,6133	0,8610
0,65	0,3296	0,6626	0,8846
0,70	0,3818	0,7116	0,9060
0,75	0,4426	0,7603	0,9253
0,80	0,5143	0,8088	0,9429
0,85	0,6000	0,8570	0,9590
0,90	0,7043	0,9049	0,9738
0,95	0,8341	0,9526	0,9874

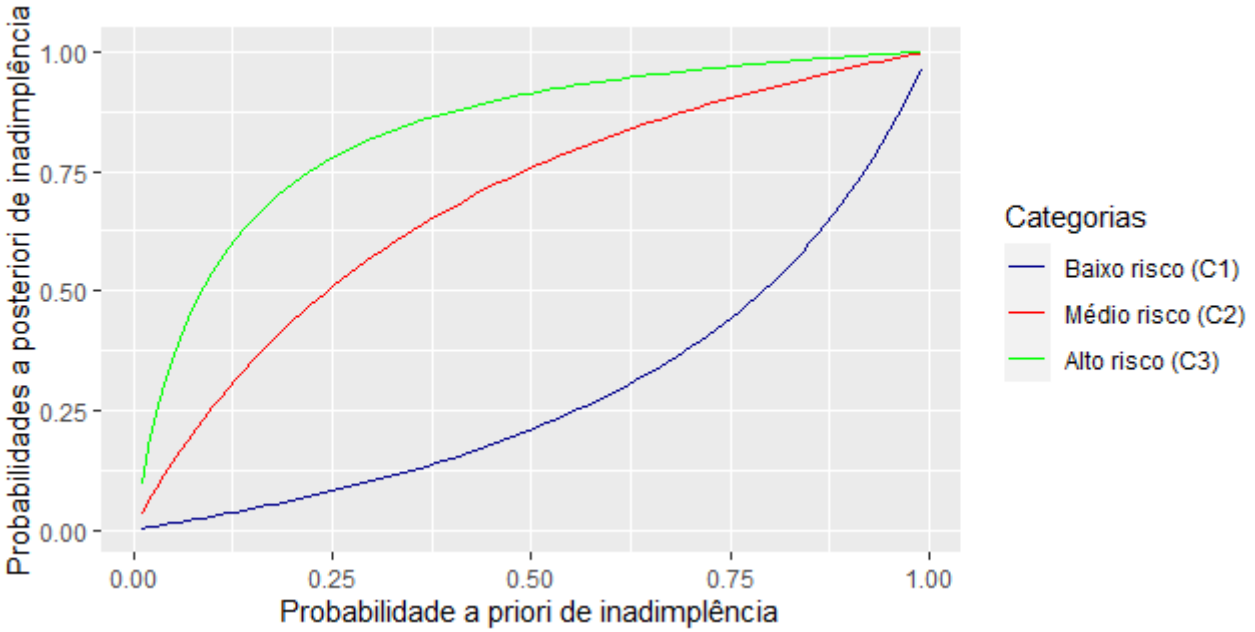


Figura 4.1: Probabilidades *a posteriori* de inadimplência para as três categorias de risco conforme escolha da *priori*

Os resultados apresentados pela Tabela 4.10 e pela Figura 4.1 mostram a robustez dessa modelagem para o conjunto de dados adotado nessa ilustração. Note que mesmo com *prioris* extremas (probabilidade *a priori* de inadimplência de 20% ou 80%) este modelo ainda estima altas probabilidades de adimplência para clientes classificados como baixo risco pelo ER e altas probabilidades de inadimplência para clientes classificados como alto risco pelo ER.

Capítulo 5

Conclusões

5.1 Considerações finais

Este trabalho propôs uma modelagem bayesiana de risco de crédito para a classificação de clientes quanto ao seu risco de inadimplência. Neste trabalho, a modelagem teve como referência um Escore de Risco fundamentado em um modelo de regressão logística, e a metodologia apresentada propôs também a incorporação de uma informação *a priori* no processo de classificação dos clientes, e não apenas na obtenção das estimativas dos parâmetros do modelo logístico. Cabe destacar que a metodologia proposta neste trabalho pode ser utilizada para qualquer outro tipo de modelagem que gere um Escore de Risco e não somente com modelos de regressão logística.

Ademais, a principal vantagem de se considerar a informação *a priori* no processo de classificação, se deve a sua simplicidade em incorporar a opinião do especialista na modelagem que apenas deverá emitir opinião sobre sua crença da probabilidade de um cliente ser ou não inadimplente. Essa mesma simplicidade não é encontrada nos modelos tradicionais, cuja informação *a priori* recai sobre os parâmetros dos modelos que não raramente, estão associados a covariáveis sujeitas a problemas de multicolinearidade.

A metodologia proposta neste trabalho foi ilustrada por meio de um conjunto de dados obtidos na literatura e os resultados obtidos mostraram que o modelo é útil para a classificação de clientes quanto a sua probabilidade de inadimplência.

Por fim, poder dispor de mais uma opção de modelagem de risco intuitivamente simples, teoricamente coerente e facilmente implementável pode popularizar ainda mais a aplicação de métodos bayesianos na área de risco de crédito.

Apêndice A

Código fonte em R

```
require(tidyverse)
require(rstan)
require(rstanarm)
require(stargazer)

options(scipen = 10000)
setwd('C:/Users/moniq/Disserta  o/Risco de credito/02-Arquivos_
    disserta  o/00-Dados')
# 1. Leitura dos dados ----
dados <- read.table("german.data", fileEncoding="Latin1", dec=",",
    header = F)
# 2. Selecionando potenciais variaveis dados1 <- dados %>% select(V1,V2
    ,V3,V5,V6,V9,V11,                    V13,V14,V15,V16,V17,V18)
names(dados1) <- c('conta_existente',
                   'duracao_meses',
                   'cod_historico_credito',
                   'valor_total_creditos',
                   'valor_poupanca',
                   'sexo_estado_civil',
                   'tempo_residencia_atual',
                   'idade_anos',
                   'outros_planos_parcelamento',
```

```
                    'tipo_casa',
                    'qtd_creditos_no_banco',
                    'emprego',
                    'qtd_dependentes')

# 3. Recategorizando as variaveis ----
dados2 <-
  dados1 %>%
  mutate(conta_existente =
  case_when(
  conta_existente == 'A14' ~ 1, # SEM CONTA
  conta_existente == 'A11' ~ 2, # <0 DM
  conta_existente == 'A12' ~ 3, # 0<DM<200
  conta_existente == 'A13' ~ 4), # DM>=200
  cod_historico_credito =
  case_when(
  cod_historico_credito == 'A30' ~ 0, # adimplente
  cod_historico_credito == 'A31' ~ 0,
  cod_historico_credito == 'A32' ~ 0,
  cod_historico_credito == 'A33' ~ 0,
  cod_historico_credito == 'A34' ~ 1), # inadimplente
  valor_poupanca =
  case_when(
  valor_poupanca == 'A61' ~ 1, # < 100 DM
  valor_poupanca == 'A62' ~ 2, # 100 <= DM < 500
  valor_poupanca == 'A63' ~ 3, # 500 <= DM < 1000
  valor_poupanca == 'A64' ~ 4, # >= 1000 DM
  valor_poupanca == 'A65' ~ 5),# SEM CONTA POUPAN A
  sexo_estado_civil =
  case_when(
  sexo_estado_civil == 'A91' ~ 1, # homem divorciado/separado/ casado
  sexo_estado_civil == 'A92' ~ 3, # mulher divorciada/ separada/ casada
  sexo_estado_civil == 'A93' ~ 2, # homem solteiro
  sexo_estado_civil == 'A94' ~ 1, # homem divorciado/ separado/ casado
```

```
  sexo_estado_civil == 'A95' ~ 4), # mulher solteira
  outros_planos_parcelamento = case_when(outros_planos_parcelamento ==
     'A141' ~ 1, #BANCO
           outros_planos_parcelamento == 'A142' ~ 2, #LOJA
           outros_planos_parcelamento == 'A143' ~ 3),#NENHUM
  tipo_casa = case_when(
  tipo_casa == 'A151' ~ 1, # ALUGADA
  tipo_casa == 'A152' ~ 2, # PROPRIA
  tipo_casa == 'A153' ~ 3), # MORA DE FAVOR
  emprego = case_when(
  emprego == 'A171' ~ 1, # DESEMPREGADO
  emprego == 'A172' ~ 2, # nao qualificado (sem carteira)
  emprego == 'A173' ~ 3, # qualificado/oficial (com carteira ou
     concursado)
  emprego == 'A174' ~ 4 ) # empresario autonomo
  )
head(dados2)

res1 <- dados %>% select(V1,V2,V3,V5,V6,V9,V11,
                         V13,V14,V15,V16,V17,V18)
a <- names(res1)
res2 <- dados2 %>% select(c('conta_existente',
                            'duracao_meses',
                            'cod_historico_credito',
                            'valor_total_creditos',
                            'valor_poupanca',
                            'sexo_estado_civil',
                            'tempo_residencia_atual',
                            'idade_anos',
                            'outros_planos_parcelamento',
                            'tipo_casa',
                            'qtd_creditos_no_banco',
                            'emprego',
                            'qtd_dependentes'))
```

```
b <- names(res2)
df <- data.frame(a,b)

# Tabela 4.1
names(df) <- c('Vari veis originais','Vari veis renomeadas')
stargazer(df, type = 'latex', summary = F, rownames = F)

# 5. Categorizando as variaveis do modelo STAN ----
y <- as.factor(dados2$cod_historico_credito)
x1 <- as.factor(dados2$conta_existente)
x2 <- as.factor(dados2$sexo_estado_civil)
x3 <- dados2$idade_anos
x4 <- as.factor(dados2$tipo_casa)
x5 <- as.factor(dados2$emprego)
x6 <- dados2$tempo_residencia_atual
x7 <- dados2$duracao_meses
x8 <- dados2$valor_total_creditos
x9 <- as.factor(dados2$valor_poupanca)
x10 <- dados2$qtd_dependentes
x11 <- dados2$qtd_creditos_no_banco
x12 <- as.factor(dados2$outros_planos_parcelamento)

# 6. Gerando modelo STAN ----

dados_stan <- data.frame(y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12)
eststst <- stan_glm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12,
                    data = dados_stan,
                    family = binomial(link = "logit"),
                    prior = normal(0,100),
                    prior_intercept = normal(0,100),
                    cores = 16, seed = 12345)

# 7. Carregando modelo STAN ----
summary(eststst)
```

```
120
121 # 8. Estatisticas descritivas ----
122
123 posterior_interval(eststst, prob = 0.95)
124 round(eststst$coefficients,3)
125
126 p_eststst <- eststst$linear.predictors
127
128 # 9. Construção da figura do ponto de corte ----
129 cortes <- matrix(0,length((-401:150)),4)
130 j <- 1
131 for (i in -401:150){
132   K<-i/100
133   classificacao<-as.numeric(p_eststst>=K)
134   x00 <- sum(as.numeric(dados_stan$y == 0)*as.numeric(classificacao==0)
        )
135   x01 <- sum(as.numeric(dados_stan$y == 0)*as.numeric(classificacao==1)
        )
136   x10 <- sum(as.numeric(dados_stan$y == 1)*as.numeric(classificacao==0)
        )
137   x11 <- sum(as.numeric(dados_stan$y == 1)*as.numeric(classificacao==1)
        )
138   #XX<-table(dados_stan$y,classificacao)
139   # Erro1<-XX[2,1]/(XX[1,1]+XX[2,1])
140   # Erro2<-XX[1,2]/(XX[1,2]+XX[2,2])
141   # Acuracia<-(XX[1,1]+XX[2,2])/sum(XX)
142
143   Erro1<-x10/(x00+x10)
144   Erro2<-x01/(x01+x11)
145   Acuracia<-(x00+x11)/length(p_eststst)
146
147   cortes[j,]<-c(K,Acuracia,Erro1,Erro2)
148   j<-j+1
149   cat(K,"\n")
```

```
}
head(cortes)
cortes <- as.data.frame(cortes)
names(cortes) <- c("corte","Acuracia","Erro I","Erro II")

ggplot(cortes) +
  geom_line(aes(x = corte, y = Acuracia, colour = "Acuracia")) +
  geom_line(aes(x = corte, y = `Erro I`, colour = "Erro I")) + geom_
      line(aes(x = corte, y = `Erro II`, colour = "Erro II")) + ggtitle
      ("") + labs(x="Escores de Risco",y="") +
  scale_color_manual(name = "ver",
                     values = c("Acuracia" = "darkblue",
                                "Erro I" = "red",
                                "Erro II" = "dodgerblue")) +
  geom_vline(xintercept = -0.409)

# 10. Corte para 3 categorias (2 cortes) ----

# Corte que minimiza o erro1
cont <- NULL
for (i in 110:130) {
  for (j in (i+1):100) {
    corte1 <- i/100
    corte2 <- j/100
    classificacao <- 1*as.numeric(p_eststst <= corte1) +
      2*as.numeric(p_eststst > corte1)*as.numeric(p_eststst <= corte2) +
      3*as.numeric(p_eststst > corte2)
    # xx <- table(y, classificacao)
    # erro1 <- xx[2,1]/sum(xx[,1])
    # erro2 <- xx[1,3]/sum(xx[,3])
    x01 <- sum(as.numeric(dados_stan$y == 0)*as.numeric(classificacao
       ==1))
    x02 <- sum(as.numeric(dados_stan$y == 0)*as.numeric(classificacao
       ==2))
```

```
180     x03 <- sum(as.numeric(dados_stan$y == 0)*as.numeric(classificacao
            ==3))
181     x11 <- sum(as.numeric(dados_stan$y == 1)*as.numeric(classificacao
            ==1))
182     x12 <- sum(as.numeric(dados_stan$y == 1)*as.numeric(classificacao
            ==2))
183     x13 <- sum(as.numeric(dados_stan$y == 1)*as.numeric(classificacao
            ==3))
184     erro1 <- x11/(x01+x11)
185     erro2 <- x03/(x03+x13)
186
187     cat(corte1, corte2, erro1, erro2, "\n")
188     # cortes[cont, ] <- c(corte1, corte2, erro1, erro2)
189     # cont <- cont + 1
190   }
191 }
192
193 #Corte que controla o erro1 em 10% e erro2 em 20%
194 corte1 <- -1.1
195 corte2 <- 1.21
196 classificacao <- 1*as.numeric(p_estst <= corte1) +
197   2*as.numeric(p_estst > corte1)*as.numeric(p_estst <= corte2) +
198   3*as.numeric(p_estst > corte2)
199 xx <- table(y, classificacao)
200 xx # matriz de confusao utilizada (Tabela 5.4)
201
202 # aperfeioando a tabela
203 xxt <- t(xx)
204 pxt <- prop.table(xxt,margin = 1)
205 tab <- round(pxt*100,2)
206
207 tab <- as.data.frame(tab)
208 tab <- tab %>% spread(y, Freq) %>% janitor::adorn_totals(c('row','col')
    )
```

```
209
210
211 # 11. Probabilidades a posteriori ----
212  P(E_0|C_1) = P(E0, C1)/P(C1) = P(E0,C1)/[P(E0,C1)+P(E1,C1)]
213  = P(E0)*P(C1|E0)/[P(E0)*P(C1|E0)+P(E1)*P(C1|E1)]
214  = 0.5*0.7737/(0.5*0.7737+0.5*0.2048) = 0.7907
215
216 ## 1. Verossimilhan a dos dados ----
217 ## 1.1. Priori nao informativa ----
218
219 PE0 <- 0.5
220 PE1 <- 0.5
221
222 PC1E0 <- 0.7737
223 PC2E0 <- 0.0452
224 PC3E0 <- 0.1810
225
226 PC1E1 <- 0.2048
227 PC2E1 <- 0.0478
228 PC3E1 <- 0.7474
229
230 PE0C1 <- PE0*PC1E0 / (PE0*PC1E0 + PE1*PC1E1)
231 PE0C2 <- PE0*PC2E0 / (PE0*PC2E0 + PE1*PC2E1)
232 PE0C3 <- PE0*PC3E0 / (PE0*PC3E0 + PE1*PC3E1)
233
234 PE1C1 <- PE1*PC1E1 / (PE0*PC1E0 + PE1*PC1E1)
235 PE1C2 <- PE1*PC2E1 / (PE0*PC2E0 + PE1*PC2E1)
236 PE1C3 <- PE1*PC3E1 / (PE0*PC3E0 + PE1*PC3E1)
237
238 # Posteriori
239 dados1.1 <- data.frame('Adimplente' = c(PE0C1,PE0C2,PE0C3),
240                        'Inadimplente' = c(PE1C1,PE1C2,PE1C3))
241 dados1.1$Adimplente <- round(dados1.1$Adimplente,4)
242 dados1.1$Inadimplente <- round(dados1.1$Inadimplente,4)
```

```
243 ## ESTUDO DA SENSIBILIDADE A PRIORI - CONTROLANDO O ERRO I EM 10% E
      ERRO II EM 20%
244
245 PC1E0 <- 0.7737
246 PC2E0 <- 0.2136
247 PC3E0 <- 0.0127
248
249 PC1E1 <- 0.2048
250 PC2E1 <- 0.6621
251 PC3E1 <- 0.1331
252
253 M_PROB <- matrix(0,99,4)
254
255 for (i in 1:99) {
256   PE1 <- i/100
257   PE0 <- 1 - PE1
258   PE0C1 <- PE0*PC1E0 / (PE0*PC1E0 + PE1*PC1E1)
259   PE0C2 <- PE0*PC2E0 / (PE0*PC2E0 + PE1*PC2E1)
260   PE0C3 <- PE0*PC3E0 / (PE0*PC3E0 + PE1*PC3E1)
261
262   PE1C1 <- PE1*PC1E1 / (PE0*PC1E0 + PE1*PC1E1)
263   PE1C2 <- PE1*PC2E1 / (PE0*PC2E0 + PE1*PC2E1)
264   PE1C3 <- PE1*PC3E1 / (PE0*PC3E0 + PE1*PC3E1)
265
266   M_PROB[i,] <- c(PE1, PE1C1, PE1C2, PE1C3)
267 }
268
269 plot(M_PROB[,1], M_PROB[,2], ylim = c(0,1), type = "l", main = "
      Probabilidades a posteriori de inadimpl ncia",
270      xlab = "Probabilidade a priori de inadimpl ncia", ylab = "
          Probabilidades a posteriori de inadimpl ncia")
271 points(M_PROB[,1], M_PROB[,3], type = "l", col =2)
272 points(M_PROB[,1], M_PROB[,4], type = "l", col =3)
273 legend(0.7,0.3, c("Baixo risco (C1)","M dio risco (C2)","Alto risco (
```

```
    C3)"), text.col = c(1,2,3), bty = "n")
274 # Probabilidades a posteriori de inadimpl ncia considerando a
    classifica  o do ER que controla o Erro I em 10%
275
276 M_PROB <- as.data.frame(M_PROB)
277 names(M_PROB) <- c("Priori", "Baixo risco (C1)", "M dio risco (C2)", "
    Alto risco (C3)")
278
279 ggplot(M_PROB) +
280   geom_line(aes(x = Priori, y = `Baixo risco (C1)`, colour = "Baixo
      risco (C1)")) +
281   geom_line(aes(x = Priori, y = `M dio risco (C2)`, colour = "M dio
      risco (C2)")) +
282   geom_line(aes(x = Priori, y = `Alto risco (C3)`, colour = "Alto risco
       (C3)")) +
283   scale_color_manual(name = "Categorias",
284                      values = c(
285                        "Baixo risco (C1)" = "darkblue",
286                        "M dio risco (C2)" = "red",
287                        "Alto risco (C3)" = "green")) +
288   ggtitle("") +
289   labs(x="Probabilidade a priori de inadimpl ncia",
290        y="Probabilidades a posteriori de inadimpl ncia")
291
292 valores <- seq(5, 95, 5)
293 M_PROB2 <- M_PROB %>% filter(index %in% valores)
294 M_PROB2 <- M_PROB2 %>% select(-index)
295 M_PROB2$Priori <- round(M_PROB2$Priori,2)
296
297 stargazer::stargazer(M_PROB2, summary = F, rownames = F, decimal.mark =
    ",", digits = 4)
```

Referências Bibliográficas

A. Barreto. Teoria e aplicações com o programa estatístico r. *Brasília: Ed. do Autor*, 2011.

G. A. S. Brito and A. Assaf Neto. Modelo de classificação de risco de crédito de empresas. *Revista Contabilidade & Finanças*, 19:18–29, 2008.

L. O. G. Cella. Regressão ordinal bayesiana. *Dissertação (Mestrado em Estatística), Universidade de Brasília, Brasília*, 2013.

D. Dua and C. Graff. UCI machine learning repository. *http://archive.ics.uci.edu/ml. University of California, Irvine, School of Information and Computer Sciences*, 2017.

D. Durand. Credit-rating formulae. In *Risk Elements in Consumer Instalment Financing*, pages 83–91. NBER, 1941.

R. Ehlers. Introdução à inferência bayesiana [online]. *Available from: http://www.icmc.usp.br/pessoas/ehlers/bayes/bayes.pdf. Viewed Octuber 2022*, 2007.

D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

A. R. Machado. Collection scoring via regressão logística e modelo de riscos proporcionais de cox. *Dissertação (Mestrado em Estatística) - Universidade de Brasília, Brasília*, 2015.

P. L. Meyer. *Probabilidade: aplicações à estatística*. Livros Técnicos e Científicos Rio de Janeiro, 1983.

E. Y. Nakano and C. A. d. B. Pereira. Soluções bayesianas para alguns problemas clássicos com dados discretos. *Tese (Doutorado em Ciências) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo*, 2010.

J. P. Narain, M. C. Raviglione, and A. Kochi. Hiv-associated tuberculosis in developing countries: epidemiology and strategies for prevention. *Tubercle and lung disease*, 73 (6):311–321, 1992.

M. C. Pires. Abordagem bayesiana para modelos de regressão logística com erros e classificações repetidas. 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

J. Wakefield. *Bayesian and frequentist regression methods*, volume 23. Springer, 2013.