unesp

UNIVERSIDADE ESTADUAL PAULISTA

Risco de crédito e a aplicação da modelagem regressão logística

Lourival Vieira

INSTITUTO DE GEOCIÊNCIAS E CIÊNCIAS EXATAS

RIO CLARO - SP

2023

Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

LOURIVAL VIEIRA

RISCO DE CRÉDITO E A APLICAÇÃO DA MODELAGEM REGRESSÃO LOGÍSTICA

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Matemática.

Orientador
Prof. Dr. Wladimir Seixas

Rio Claro - SP
2023

V658r

Vieira, Lourival
Risco de crédito e a aplicação da modelagem regressão logística / Lourival Vieira. -- Rio Claro, 2023
71 p. : il., tabs.

Dissertação (mestrado profissional) - Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Wladimir Seixas

1. Análise de Risco. 2. Regressão Logística. 3. Métodos Matemáticos. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Geociências e Ciências Exatas, Rio Claro. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Impacto potencial desta pesquisa

A dissertação tem como objeto de estudo a promoção do entendimento de uma metodologia que pode ser aplicada para calcular os riscos de crédito de um grupo específico de tomadores. Como base central dessa discussão, tomou-se como parâmetro a implantação da Lei do Cadastro Positivo para abordar a mensuração do score de crédito. Como objetivo secundário busca-se compreender o impacto desses estudos na orientação das práticas de crédito atualmente adotadas no cenário financeiro e creditício brasileiro. A aplicação de dados reais em modelagens estatísticas e econométricas mostra a importância de relacionar a matemática com estatística e economia para compreender as aplicações matemáticas na vida real. Ao avaliar o alcance e a eficácia dessas análises, pode-se compreender como elas contribuem significativamente no aprimoramento das diretrizes utilizadas no processo de concessão de crédito.

Potential impact of this research

The dissertation aims to promote the understanding of a methodology that can be applied to calculate the credit risks of a specific group of borrowers. As a central basis for this discussion, the implementation of the Positive Credit Registry Law was taken as a parameter to address the measurement of the credit score. As a secondary objective, the study seeks to comprehend the impact of these analyses on guiding the credit practices currently adopted in the Brazilian financial and credit scenario. The application of real-world data in statistical and econometric modeling shows the importance of integrating mathematics with statistics and economics to comprehend mathematical applications in real life. By assessing the scope and effectiveness of these analyses, one can understand how they significantly contribute to improving the guidelines used in the credit granting process.

LOURIVAL VIEIRA

RISCO DE CRÉDITO E A APLICAÇÃO DA MODELAGEM REGRESSÃO LOGÍSTICA

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Matemática.

Comissão Examinadora

Prof. Dr. WLADIMIR SEIXAS
DM/UFSCar (SP)

Prof. Dr. LUIS ANTONIO DA SILVA VASCONCELLOS
FC/UNESP/Bauru (SP)

Profa. Dra. SELENE MARIA COELHO LOIBEL
IGCE/UNESP/Rio Claro (SP)

Conceito: Aprovado

Rio Claro/SP, 30 de outubro de 2023

Dedico, aos meus filhos Laura Sousa Vieira e Henrique Sousa Vieira

Agradecimentos

Agradeço,

Ao meu pai (in memoriam) e à minha mãe pelo apoio com suas palavras amigas, conselhos e orações que me deram na vida e especificamente durante o desenvolvimento desta dissertação. Agradeço do fundo do meu coração, pois com certeza, sem eles, não teria chegado onde cheguei.

À minha esposa, pela compreensão e paciência.

Aos colegas de trabalho, especialmente à Professora Juliana Portella Furini, pela paciência, sabedoria, profissionalismo e compreensão, fatores que contribuíram para o amadurecimento das propostas e avanço da pesquisa.

Ao Professor Mário Sérgio Rodrigues Balbino de Oliveira Paschoal, colega do programa e do ambiente de trabalho, por ter proporcionado as amplas discussões sobre o direcionamento das leituras e a organização das ideias. Sem dúvida, as discussões e as trocas de conhecimento foram muito produtivas.

À minha eterna e amada diretora de escola, Alteia Garagnani Turpin (in memoriam), que com suas palavras de incentivos e conselhos proporcionou a minha entrada neste programa.

Aos demais colegas de trabalho que, com suas palavras amigas, proporcionaram incentivos para continuar estudando.

Aos demais colegas discentes do programa, especialmente aos colegas Jefferson David Alves e Alan Gualberto, que com suas sabedorias, humildade e paciência, conduziram as discussões de forma serena e didática nos grupos de estudos.

Ao Professor Dr. Wladimir Seixas, do fundo do meu coração, pela paciência e profissionalismo que guiaram esta orientação com amizade e serenidade.

À Professora Dra. Selene Maria Coelho Loibel, à Professora Dra. Marta Cilene Gadotti e a todos os professores do Departamento de Matemática por terem contribuído para a formação e realização desta dissertação.

Resumo

O presente estudo tem como objetivo principal a realização de um estudo da técnica estatística conhecida como Regressão Logística, também referida como análise logit. Nesse sentido, busca entender a metodologia utilizada e sua aplicação na avaliação e determinação dos fatores de maior relevância que influenciam na pontuação do risco de crédito para indivíduos. Como objetivo secundário busca-se compreender o impacto desses estudos na orientação das práticas de crédito atualmente adotadas no cenário financeiro e creditício brasileiro, considerando especialmente a introdução do sistema de cadastro positivo. Ao avaliar o alcance e a eficácia dessas análises, pretende-se compreender se elas contribuem significativamente para aprimorar as diretrizes utilizadas no processo de concessão de crédito.

Palavras-chave: Análise de Risco. Regressão Logística. Métodos Matemáticos.

Abstract

The main objective of the present study is to conduct an investigation into the statistical technique known as Logistic Regression, also referred to as logit analysis. In this context, it aims to comprehend the methodology employed and its application in assessing and determining the most relevant factors that influence the credit risk scoring for individuals. As a secondary objective, the study seeks to understand the impact of these analyses on guiding the credit practices currently adopted in the Brazilian financial and credit landscape, with special consideration given to the introduction of the positive credit registry system. By evaluating the scope and effectiveness of these analyses, the intention is to ascertain whether they significantly contribute to enhancing the guidelines employed in the credit approval process.

Keywords: Risk Analysis. Logistic Regression. Mathematical Methods.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A concessão de crédito à pessoa física no Brasil vem sendo modificada desde a implantação do processo de cadastro positivo pela Lei n. 12.414/2011, alterada pela Lei Complementar nº 166 (de 8 de abril de 2019)[1], que prevê a inclusão automática de todos os dados das pessoas físicas e jurídicas que demandam empréstimos, financiamentos, compras a prazo ou contas de consumo junto ao mercado. Esta lei proporciona às instituições financeiras acesso aos dados cadastrais da pessoa física no contexto nacional, disponíveis nos registros do Banco Central do Brasil. A Lei do Cadastro Positivo busca ampliar a concessão de crédito ao mercado consumidor, reduzindo o risco de inadimplência que possa ocorrer junto ao sistema financeiro creditício.

A determinação da taxa de risco de crédito é obtida através da utilização de metodologias qualitativas e quantitativas, que, em conjunto, proporcionam um composto de técnicas aplicadas nas tomadas de decisões relacionadas ao crédito pelos credores. Dentre essas técnicas, destacam-se:

1. Técnicas de Redes Neurais, que são regidas por sistemas computacionais empregados para imitar o funcionamento do cérebro humano por meio de emulações de uma rede de neurônios interligados. Essas técnicas modelam otimizações que integram técnicas de programação matemática para descobrir os pesos ideais de atributos do credor e tomador de crédito, visando minimizar os erros do credor e maximizar seus lucros.

2. Técnicas do Sistema Especialista, baseadas em regras, que utilizam sistemas movidos por clonagem de processos empregados para proporcionar análises bem-sucedidas na decisão de crédito.

3. Técnicas dos Sistemas Híbridos, que utilizam comutação, estimativa e simulações diretas, cujos parâmetros são determinados em partes por técnicas de estimativas. Um exemplo dessa aplicação é o modelo KMV, que usa formulação teórica de opções para explicar a inadimplência por meio do relacionamento das estimativas.

[1] Disponível em <http://www.planalto.gov.br/ccivil_03/leis/lcp/Lcp166.htm>. Acesso em 29 mai. 2023.

4. Técnicas dos modelos Econométricos, que são as mais eficientes para mensurar os modelos de risco de crédito. Na prática, as análises discriminatórias lineares e múltiplas, como análise logit e análise probit, são utilizadas para modelar a probabilidade de adimplência ou inadimplência. Devido à praticidade e transparência, são as mais utilizadas na determinação do risco de crédito pelo mercado.

Dentre as várias técnicas abordadas, este trabalho tem como proposta realizar uma abordagem detalhada da Regressão Logística (análise logit), priorizando sua metodologia e aplicação na determinação dos fatores preponderantes da pontuação do risco de crédito à pessoa física. A proposta inicial é identificar se essas análises contribuem com as orientações nas práticas de crédito vigentes no mercado financeiro e creditício no Brasil, a partir da implantação do cadastro positivo. Dessa forma, a estrutura deste trabalho está dividida em três partes: a primeira parte realizará uma abordagem introdutória sobre a teoria dos riscos. A segunda abordará a teoria dos riscos de crédito, através da aplicação do modelo de regressão logística, das técnicas do risco relativo e a mensuração do spread de crédito. A terceira parte será destinada a realizar as aplicações da regressão logística na apuração dos resultados obtidos pelos modelos de risco de crédito do cadastro positivo, bem como a classificação dos riscos de crédito no Brasil destinadas à pessoa física.

2 Teoria dos riscos

O que se entende por crédito é que, na prática, há uma relação entre a disposição de demanda por crédito por parte da pessoa física, denominado Valor Presente, e a disposição de oferta de crédito, denominado Montante, por parte do agente financeiro creditício. O prêmio auferido nas transações por meio da concessão de crédito é composto pela taxa de juros e pela variável tempo incorridos na operação pela qual a oferta de crédito será concedida. A disposição do credor em ofertar crédito está diretamente relacionada com as avaliações do perfil de crédito do tomador, pois acreditam que as análises do risco de crédito garantem o retorno esperado na operação, isto é, os cuidados com as análises de risco de crédito para determinar a taxa de juros se tornam uma base primordial para sustentar a harmonia entre credor e tomador.

Por ser uma das práticas mais antigas da história do mercado financeiro, as análises de risco de crédito se aperfeiçoam na medida em que as instituições financeiras e creditícias se fortalecem. Desse modo, em mercados cuja conjuntura econômica e financeira pode apresentar alto grau de volatilidade em função das incertezas no contexto mercadológico, financeiro e produtivo, os tomadores e credores de créditos estão expostos aos diferentes tipos de riscos, dentre os quais os mais comuns são: riscos de liquidez, risco legal, risco operacional, risco trabalhista, risco de mercado, risco político, risco econômico e risco de crédito. No entanto, os agentes credores baseiam suas análises nos estudos sobre os riscos de crédito para identificar os graus de incerteza da incapacidade do tomador de crédito de cumprir suas obrigações contratuais frente aos credores.

O não cumprimento do contrato por parte do tomador promove um estado de perdas financeiras para a instituição fornecedora de crédito. Portanto, por medidas cautelares, observam também as possíveis perdas da capacidade de pagamentos e a percepção de inadimplência no mercado, situações que também influenciam nas análises dos riscos de crédito à pessoa física.

Em economias de mercado cuja maior parte da renda circulante é oriunda da massa trabalhadora assalariada, o crédito passa a ser uma das práticas mais importantes para proporcionar o desenvolvimento econômico e financeiro do país. Logo, a prática de crédito se torna preponderante na questão do financiamento para aquisições de bens de consumo duráveis, não duráveis e bens de serviços às famílias que, por hora, as receitas mensais

não permitiriam adquiri-los.

Com o avanço produtivo e comercial em conjunto com os incrementos tecnológicos da informação, tem havido implementações de novos modelos, procedimentos e práticas na gestão de análise de risco de crédito, o que visa proporcionar relações mais amistosas e transparentes entre credores e tomadores, situações que se tornam tarefas primordiais na questão financeira e creditícia.

2.1 Conceitos teóricos sobre modelos de Riscos de Crédito

As discussões em relação aos modelos ideais de risco de crédito são fundamentais para orientar as instituições financeiras e creditícias nas decisões políticas sobre as concessões de crédito. Vale (2010) defende que a modelagem, como algoritmos, fórmulas, sistemas ou regras, é fundamental para representar a compreensão de um fenômeno e também contribui para o fortalecimento das decisões em relação às medidas precaucionais em relação ao risco de crédito. Da mesma forma, Caouette, Altman e Narayanan (1999) salientam que, para construir um modelo de risco de crédito, é necessário estabelecer primeiramente estratégias que visem identificar as variáveis que influenciam na ocorrência do não cumprimento de um acordo contratual de crédito. Dessa forma, eles ponderam que é importante basear-se em um conjunto de ferramentas que permita a construção de um modelo formal com base em um conjunto de dados reais que representam a carteira de crédito.

A partir da mensuração e apuração das informações pertinentes, estas devem ser submetidas a testes estatísticos para identificar se o modelo proporciona ou não o desempenho esperado. Por outro lado, Saunders (2000) divide os modelos de risco de crédito em sistemas especialistas, modelos de *Credit Scoring*, modelos de *Credit Rating* e modelos de portfólio. Para ele, os três primeiros modelos caracterizam as abordagens tradicionais de classificação de risco, enquanto o baseado em portfólio considera os retornos e riscos esperados em uma análise da carteira de crédito. Silva (2014) também pondera que os modelos de classificação de risco têm como objetivo analisar o crédito de forma a auxiliar o credor na tomada de decisão a partir da avaliação de várias informações sobre o tomador de crédito. Para ela, as instituições financeiras devem praticar uma gestão de riscos bem definida para garantir que a saúde financeira institucional seja perenemente saudável, uma vez que uma de suas principais atividades é conceder crédito aos tomadores. Da mesma forma, Silva (2014) salienta que a análise quantitativa fundamenta suas informações em modelos estatísticos e econométricos, permitindo, assim, uma mensuração mais precisa do risco do tomador de crédito, proporcionando de forma transparente a modelagem de *Scoring* ou *Rating*. Essa modelagem permite identificar o quão próximo o tomador de

crédito está de dois grupos: o adimplente, que provavelmente cumprirá suas obrigações financeiras, o inadimplente, que apresenta a probabilidade mais alta de não cumprir seus compromissos com as instituições financeiras. Já a análise qualitativa, que tem como parâmetro a análise subjetiva do analista de crédito, mede a capacidade de pagamento do tomador de crédito a partir das práticas da análise fundamentalista.

A classificação como modelo *especialista* parte do pressuposto de que envolve as decisões individuais em relação à concessão ou não de crédito. Esse modelo permite tomar decisões que são expressas nas experiências subjetivas, nas informações disponíveis e nas sensibilidades que cada analista possui em relação ao risco de seus negócios. Caouette, Altman e Narayanan (1999, p. 181) propõem que existem dois modelos de classificação de risco de crédito: o modelo de aprovação e o modelo de escoragem (pontuação) comportamental. Tanto no modelo de aprovação quanto no modelo de escoragem comportamental, as variáveis de escoragem são fundamentais para a tomada de decisões na concessão de novas linhas de crédito, pois a previsão de solvência ou insolvência depende dos resultados apresentados pelas variáveis descritivas das atividades implementadas no modelo de aprovação e no modelo de escoragem comportamental.

O que se pode observar é que o modelo de escoragem comportamental baseia-se na mensuração dos registros históricos, levando em consideração os hábitos de pagamento, o volume de transações, a utilização média das linhas de crédito e as variáveis descritivas das atividades passadas na conta do tomador de crédito. Já as análises subjetivas da capacidade financeira dos tomadores de crédito são tradicionalmente conhecidas como os *5C's do crédito*, que analisam os riscos de crédito tanto para pessoas físicas quanto para pessoas jurídicas. Securato (2002) também analisa o modelo de pontuação de crédito, como estratégias, os parâmetros dos *5C's do crédito* que orientam tanto os tomadores quanto os credores em relação aos riscos de crédito que serão utilizados nas análises quantitativas e qualitativas para obter a pontuação de risco de crédito. No entanto, os *5C's* recebem as seguintes denominações: Caráter, Capacidade, Capital, Colateral e Condições, sendo consideradas variáveis fundamentais na análise de risco de crédito por uma instituição.

A utilização da metodologia dos *5C's* torna-se muito importante para medir o grau de risco incorrido e identificar o valor financeiro que um credor pode emprestar em relação à renda média estabelecida na operação. Para ser mais preciso, é importante levar em consideração a análise do risco total de uma operação de crédito, abrangendo o Risco Conjuntural e o Risco Próprio.

$$\text{Risco Total} = \text{Risco Conjuntural} + \text{Risco Próprio}.$$

Securato (2002) define o Risco Conjuntural (ou Risco Sistemático) como um conjunto de variáveis observadas na conjuntura econômica, estruturas políticas, sociais e ambientais nas quais o tomador de crédito está inserido a curto, médio e longo prazo. Por outro

lado, o Risco Próprio (ou não sistemático) depende exclusivamente das características específicas dos grupos de informações definidos como *5C's do Crédito*. Nesse contexto, ele apresenta o modelo matricial de crédito, no qual cada variável está relacionada com suas características, conforme explicado a seguir:

Caráter: está relacionado ao comportamento ético e responsável de honrar os compromissos de acordo com as cláusulas contratuais;

Capacidade: está relacionada à garantia de liquidez monetária, ou seja, à renda do tomador e sua capacidade de cumprir suas obrigações;

Capital: refere-se à formação patrimonial do tomador, correspondente à capacidade de formação de ativos pessoais do tomador de crédito;

Colateral: são os instrumentos de garantia apresentados pelo tomador no momento da concessão de crédito. Isso envolve processos como avalistas ou fiadores, nos quais garantias adicionais podem ser solicitadas, dependendo do grau de risco do tomador, para facilitar a liberação do crédito;

Condições: são as análises dos impactos dos fatores micro e macroeconômicos que influenciam a concessão de crédito ao tomador, levando em consideração as variabilidades da conjuntura econômica do período.

A organização matricial das variáveis de Risco de Crédito é fundamentada na construção de uma matriz de crédito, na qual as linhas representam os parâmetros de Risco Próprio (5 C's) e as colunas representam os possíveis Riscos Conjunturais, indicados por cenários: C1, C2, ... Cn.

Securato (2002) também considera que a etapa final da concessão de crédito está relacionada à conclusão criteriosa das análises da ficha cadastral, que é estabelecida a partir de um sistema de pontuação que visa quantificar os parâmetros analisados em cada caso. Isso, de fato, contribui para a atribuição de pesos correspondentes à relevância de cada informação, culminando na determinação de uma escala classificatória discriminante para cada perfil. A partir da obtenção da pontuação ponderada e das assinaturas do contrato, a concessão de crédito será encaminhada para análise pelo Comitê de Crédito, conforme definido pela política da instituição financeira.

2.2 Modelagem e análise de Risco de Crédito

Quando se pensa em desenvolver uma metodologia para análise de risco de crédito, a primeira coisa que vem à mente é o manejo e uso das variáveis exógenas e endógenas do tomador de crédito. Para apoiar essa prática, a metodologia da Regressão Logística

(análise logit) se torna uma das mais eficazes para distinguir bons pagadores de maus pagadores. O objetivo dessa abordagem é identificar que quanto menor for a sobreposição entre as distribuições de escores, melhor será a distinção entre um bom pagador e um mau pagador. Essa distinção é representada por meio de um processo de pontuação chamado *Credit Scoring* ou *Credit Rating*, geralmente divulgado por agências consolidadas de classificação de riscos, como Moody's e Standard and Poor's internacionalmente, e no Brasil, a principal fornecedora de *ratings* de crédito é a instituição Serviços de Assessoria S.A. ou SERASA. Do ponto de vista metodológico, as instituições financeiras e de crédito também utilizam práticas semelhantes para analisar os riscos de crédito de seus tomadores, a fim de estabelecer seus critérios de avaliação de forma independente.

Devido ao fato de o sistema de avaliação não ser rígido nem fixo, as avaliações de risco são revistas regularmente, acompanhando as variações das variáveis quantitativas e qualitativas no mercado. Duarte (2014) pondera que o *Credit Scoring* e o *Credit Rating* emergiram devido à crescente dificuldade de obter informações sobre o risco financeiro de devedores. Atualmente, o *Credit Rating* é abordado como um instrumento de informação para os investidores, pois pode ser definido como o cálculo da credibilidade de um tomador de crédito.

O principal objetivo do *Credit Rating* é fornecer subsídios para a eficiência do mercado financeiro e de crédito. Os *Credit Ratings* são expressos por meio de letras que variam, por exemplo, de AAA a D, para comunicar a opinião da agência sobre o nível relativo de risco de crédito que foi mensurado.

Para Anderson (2007), as agências de classificação de risco, como a Moody's, S&P e Fitch, são as mais importantes quando se trata da determinação do *rating* de crédito (*Credit Ratings*) de um país e das grandes instituições financeiras e de crédito. Segundo ele, os *ratings* de crédito são fornecidos por meio de notas conceituais em forma de letras, que classificam o grau de investimento ou especulativo, separando o grau de inadimplência e o grau de notas rebaixadas com base nas probabilidades de cumprimento ou não cumprimento dos contratos firmados pelas instituições analisadas perante o mercado em uma escala global.

A Tabela 2.1 representa um exemplo de *rating* das agências internacionais de classificação de risco Moody's, Fitch e Standard & Poor's. Essa classificação é obtida por meio das metodologias indicadas pelos agentes financeiros internacionais e é destinada a fornecer informações ao mercado financeiro e de crédito sobre os graus de solvência dos agentes econômicos na oferta de crédito. As agências mencionadas na Tabela 2.1 utilizam metodologias próprias, mas convergem para a classificação de curto e longo prazo, considerando tanto a solvência nacional quanto a soberana. Assim, o *score* para a pontuação de risco apresentado por Securato (2002) forma a base de informações suficiente para analisar os riscos qualitativos e quantitativos, que, agrupados, constituem um padrão de classificação de inadimplência, sendo o grau de severidade para o crédito crescente em

intervalos de [0 a 100%]. Com base na classificação de risco de crédito apresentada na Tabela 2.1, as instituições financeiras e de crédito estabelecem uma pontuação de risco para cada tomador de crédito, chamada de *score* de classificação de risco.

Tabela 2.1: Escala global de avaliações de risco das principais agências internacionais de *Ratings*

MOODY'S	**FITCH**	**S&P**	**Interpretação**
Aaa	AAA	AAA	Mais alta qualidade; extremamente elevada.
Aa1	AA+	AA+	Capacidade do devedor de honrar os compromissos com crédito é elevado.
Aa2	AA	AA	
Aa3	AA-	AA-	
A1	A+	A+	Qualidade Média Superior com forte capacidade de pagamento.
A2	A	A	
A3	A-	A-	
Baa1	BBB+	BBB+	Qualidade Média Inferior Capacidade de pagamento adequada.
Baa2	BBB	BBB	
Baa3	BBB-	BBB-	
Ba1	BB+	BB+	Provável cumprimento de obrigações; incertezas correntes com operações arriscadas ou Especulativas.
Ba2	BB	BB	
Ba3	BB-	BB-	
B1	B+	B+	Operações de alto risco e altamente especulativa.
B2	B	B	
B3	B-	B-	
Caa1	CCC	CCC+	Vulnerabilidade presente à inadimplência Extremamente Especulativa.
Caa2	CCC	CCC	
Caa3	CCC	CCC-	
Ca	CCC	CC	Estágio extremamente de inadimplência, onde acusa problemas de endividamento com falência e incumprimento de suas obrigações com terceiros.
–	DDD	D	
–	DD	D	
–	D	D	

Fonte: Caouette, Altman e Narayanan (1999, p. 79) e <https://www.moneyland.ch/en/rating-agencies>. Acesso em 30 jul. 2022.

Na Tabela 2.2, é possível observar essa classificação de acordo com os níveis e faixas de severidade em relação à concessão de crédito.

Tabela 2.2: Classificação conforme os níveis e faixas de severidade em relação a concessão de crédito.

Faixa	**Nível**	**Provisão de severidade (%)**
I	AA	0,0
II	A	0,5
III	B	1,0
IV	C	3,0
V	D	10,0
VI	E	30,0
VII	F	50,0
VIII	G	70,0
IX	H	100,0

Fonte: (SECURATO, 2002, p. 195-196)

Para tanto, é necessário estabelecer uma demonstração da modelagem estatística que descreve a metodologia utilizada para determinar os graus de risco utilizados pelas agências mencionadas acima, a fim de fornecer ao mercado informações sobre o comportamento de crédito dos agentes econômicos.

Securato (2002) destaca que o Banco Central do Brasil indica que a responsabilidade de estabelecer o grau de severidade na questão dos níveis deve ser assumida pelas instituições detentoras das carteiras de crédito. Dessa forma, serão elas as responsáveis por efetuar a classificação com base em critérios consistentes e verificáveis, de acordo com as informações internas e externas de cada tomador de crédito. Sendo assim, essas instituições estão autorizadas e têm a responsabilidade de estabelecer demonstrações de modelagens estatísticas que descrevam a metodologia utilizada para determinar os graus de risco, os quais são utilizados como forma de fornecer ao mercado informações sobre o comportamento creditício dos agentes econômicos.

3 Regressão Logística Binária

A regressão logística (*logistic regression* ou análise logit) é um dos modelos estatísticos mais utilizados para prever e explicar a probabilidade de ocorrência de uma variável dependente categórica e binária. Sua utilidade é evidente na determinação da dinâmica das operações de *credit scoring*. Rosa (2000) destaca que o uso de técnicas de regressão logística permite identificar o perfil de cada tomador de crédito, e através da regressão dos parâmetros, é possível localizar e identificar o grupo ao qual o tomador de crédito pertence. De acordo com ela, os parâmetros estudados em um contexto binário permitem identificar a probabilidade de ocorrência encontrada no comportamento das variáveis explicativas de um evento adimplente definido por Y, que se apresenta na forma qualitativa dicotômica $Y = 1$ e a ocorrência de um evento inadimplente quando $Y = 0$. Da mesma forma, Fávero e Belfiore (2017) propõem que essas duas categorias, por definição metodológica, possam ser consideradas como um evento de interesse quando a categoria *Dummy* $Y = 1$ e de não interesse quando a categoria é *Dummy* $Y = 0$. Para eles, caso o estudo apresente mais de duas categorias como possibilidades de ocorrência, com $Y > 1$ e $Y > 0$, será necessário definir a categoria de referência desejada e, a partir daí, desenvolver a técnica de regressão logística multinomial. Devido ao critério de demonstração e interesse metodológico, esta pesquisa não abordará a modelagem logística multinomial. Para obter mais detalhes sobre seu desenvolvimento, consulte Capítulo 13 de (FÁVERO; BELFIORE, 2017).

Devido à importância do modelo de regressão logística binária na aplicação dos estudos das probabilidades de ocorrência de um evento dicotômico de interesse, torna-se necessário focar no desenvolvimento deste estudo para alcançar os objetivos elencados na proposta desta pesquisa. Ao observarmos um modelo de análise de risco de crédito por meio da aplicação da função matemática da regressão logística binária, com um olhar mais categórico sobre os comportamentos das variáveis explicativas pelas quais os tomadores de crédito estão expostos, será possível identificar e compreender as interpretações que corroboram com a determinação classificatória dos perfis em adimplente ou inadimplente de cada agente econômico atuante no mercado de crédito.

3.1 Função matemática da Regressão Logística Binária

Para uma demonstração genérica da função matemática da regressão logística binária, será possível utilizar como base a adequação comportamental das variáveis explicativas à modelagem matemática de probabilidade linear múltipla, conforme demonstrada em (SARTORIS, 2003, p. 252), que possui a seguinte expressão:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \tag{3.1}$$

Em que ϵ_i é a variável aleatória distribuída independentemente com média 0 e o conjunto de variáveis $X_i = \{X_{1i}, X_{2i}, X_{3i}, \ldots, X_{ki}\}$, sendo as variáveis explicativas independentes discriminantes (métricas ou *dummies*), estão associadas à i-ésima variável dependente Y_i, quando $i = 1, \ldots, n$, $\beta_j = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ e $j = 1, 2, 3, \ldots, k$, com um conjunto de vetores de parâmetros desconhecidos associados a cada variável independente X_i. A variável dependente Y_i é uma variável observável composta por valores 0 e 1. Ao supor que a variável resposta tem uma distribuição de Bernoulli para $Y_i = 1$ ou $Y_i = 0$, a definição da função probabilidade para cada valor individual Y_i pode ser representada da seguinte forma:

- $Y_i = 1 \longrightarrow$ Probabilidade $P(Y_1 = 1) = P_i$.
- $Y_i = 0 \longrightarrow$ Probabilidade $P(Y_1 = 1) = 1 - P_i$.

Como P_i é a probabilidade de ocorrência gerada por uma combinação linear $P_i = f(Y_i)$ dos pesos das entradas de X_{ki} na análise discriminante linear, a análise dos valores médios ou resultados esperados de $E(\epsilon_i)$, torna-se fundamental para alcançar a média ponderada dos possíveis resultados dos erros ϵ_i. Por sua vez, para alcançar a média da variável dependente Y_i, é necessário encontrar $E(Y_i)$. Sartoris (2003) e Fávero e Belfiore (2017) salientam que o valor esperado da variável resposta $E(\epsilon_i) = 0$, para $E(Y_i) = P_i$. Assim,

$$\begin{aligned} E(\varepsilon_i) = {} & (1 - \beta_0 - \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})\, P_i \\ & + (-\beta_0 - \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})\, (1 - P_i) = 0 \end{aligned} \tag{3.2}$$

e $E(\epsilon) = 0$ (erros têm média zero),

$$E(Y_i) = 1\,((P_i) + 0\,(1 - P)) = P_i. \tag{3.3}$$

Tanto a expressão (3.3) quanto a expressão (3.2) implicam que, nesta representação, o resultado esperado dado pela função resposta representa a probabilidade da variável resposta assumir o valor 1. Em outras palavras, a regressão logística segue o modelo

de regressão linear, estabelecendo restrições na função resposta para que os valores da função Y_i se restrinjam entre 0 e 1 quando tendem ao infinito. Como exemplo ilustrativo, podemos observar que:

- Quando $X_{ki} \to +\infty$, $P_i(Y_i = 1) \to 1$,
- Quando $X_{ki} \to -\infty$, $P_i(Y_i = 1) \to 0$.

Ou seja, $0 \leqslant E(Y_i) = P_i \leqslant 1$. Tanto P_i quanto $E(Y_i)$ estão entre zero e 1. Ao considerarmos os valores esperados de cada observação da variável dependente Y_i observamos que:

$$P_i = \begin{cases} \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}, \text{ quando } 0 < \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} < 1 \\ 1 \text{ quando } \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \geqslant 1 \\ 0 \text{ quando } \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \leqslant 0 \end{cases}$$

desde que a covariância $Cov(\epsilon_i) = 0$ e que cada variável independente X_{ki} não esteja correlacionada com a combinação linear das demais variáveis. Assim,

$$Y_i = \begin{cases} 1 \text{ se } Prob(Y_i = 1) = P_1, \text{ adimplente} \\ 0 \text{ se } Prob(Y_i = 0) = 1 - P_1, \text{ inadimplente} \end{cases}$$

Para o caso da regressão logística binária acumulada, considera-se que a variável Y_i assume valores no intervalo de $-\infty$ a $+\infty$, pois é a partir desse intervalo que se torna possível encontrar a probabilidade de ocorrência de um evento P_i em função do logito Y_i. Isso permite concluir que a função $f(Y_i)$ é uma função logística com probabilidade estimada da ocorrência de um evento, apresentada na forma dicotômica para uma observação i dada por $P_i = f(Y_i)$.

De acordo com Pyndick e Rubinfeld (2004), usualmente, o modelo logit se baseia na função de probabilidade logística acumulada com a aplicação do logaritmo natural (Logit) da chance na expressão da probabilidade do evento estudado

$$P_i = f(Y_i) = \frac{1}{1 + e^{-Y_i}}.$$

Para estimar o modelo da equação (3.1) segue que:

$$(1 + e^{-Y_i})P_i = 1.$$

Dividindo por P_i e subtraindo 1 temos

$$e^{-Y_i} = \frac{1}{P_i} - 1$$

ou seja,

$$e^{Y_i} = \frac{P_i}{1 - P_i} \tag{3.4}$$

Aplicando o logaritmo em ambos os membros da equação segue que

$$Y_i = \ln \frac{P_i}{1 - P_i}$$

E assim,

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Resolvendo agora a equação (3.4) em termos de P_i temos que

$$P_i = (1 - P_i)e^{Y_i} = e^{Y_i} - P_i e^{Y_i}$$

Assim,

$$P_i + P_i e^{Y_i} = e^{Y_i}$$

$$P_i(1 + e^{Y_i}) = e^{Y_i}$$

$$P_i = \frac{e^{Y_i}}{1 + e^{Y_i}} = \frac{1}{\frac{1 + e^{Y_i}}{e^{Y_i}}}$$

Ou seja,

$$P_i = \frac{1}{1 + e^{-Y_i}}.$$

Para Pyndick e Rubinfeld (2004), a função Logística (Logit) segue a equação:

$$P_i = f(Y_i) = f(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) = \frac{e^{Y_i}}{1 + e^{Y_i}}.$$

E assim,

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}}.$$

Ao aplicar o logaritmo natural de ambos os lados obtemos:

$$P_i = f(+\infty) = \frac{1}{1 + e^{-(+\infty)}} = 1 \quad \text{e} \quad P_i = f(-\infty) = \frac{1}{1 + e^{-(-\infty)}} = 0,$$

uma vez que, P_i's são probabilidades da função de transferência convertida em Y_i. Assim,

$$\lim_{Y_i \to -\infty} f(Y_i) = \lim_{Y_i \to -\infty} \frac{1}{1 + e^{-Y_i}} = 0$$

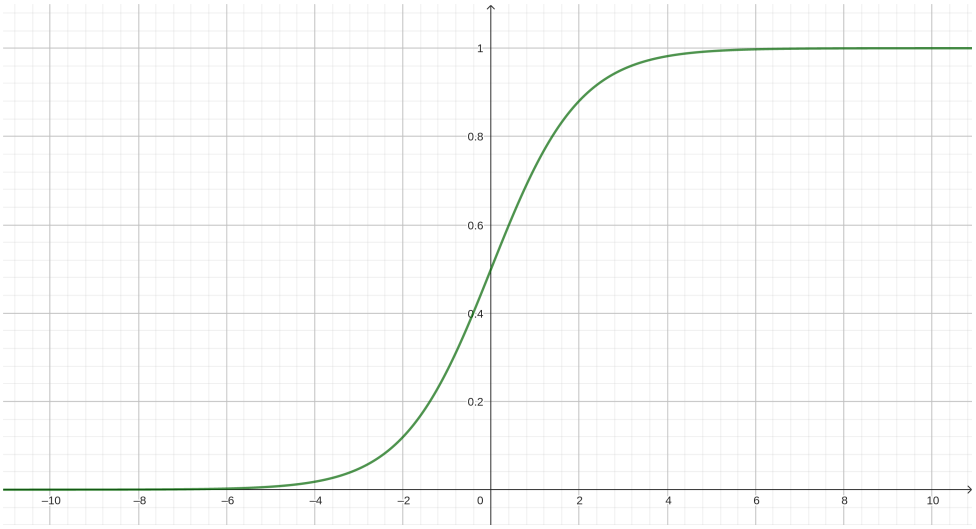
e

$$\lim_{Y_i \to \infty} f(Y_i) = \lim_{Y_i \to \infty} \frac{1}{1 + e^{-Y_i}} = 1.$$

Desse modo, a função geradora da regressão logística binária, ou função sigmoide assume

uma representação igual a representada na Figura 3.1.

Figura 3.1: Curva da regressão logística.



Fonte: Elaborada pelo autor.

De acordo com Fávero e Belfiore (2017) o que a regressão logística binária estima não são os valores previstos da variável dependente, mas a probabilidade de ocorrência do evento em estudo para cada observação.

3.2 Estimação do modelo de Regressão Logística Binária

Para Pyndick e Rubinfeld (2004, p. 379), no estudo da regressão logística para observações individuais de variáveis dicotômicas e binárias, que aborda preferências de escolhas, a técnica de estimação mais adequada é a de máxima verossimilhança, pois, além de sua composição, também possui propriedades estatísticas desejáveis. Segundo eles, todos os estimadores de parâmetros são consistentes e eficientes assintoticamente, o que permite aplicar a regressão quanto o teste de estimação por máxima verossimilhança.

Nesse caso, para estimar os parâmetros do modelo de regressão logística, será necessário utilizar o seguinte modelo:

$$P_i = f(Y_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}}.$$

Pyndick e Rubinfeld (2004, p. 379) salientam que, como P_i é a probabilidade de um indivíduo fazer uma certa escolha, dado X_{ki} e $f(Y_i)$ é a função de probabilidade logística acumulada. Se o teste da razão da verossimilhança permite observar P_i a fim de obter informações para estabelecer escolhas, então esta técnica é considerada a mais importante para o modelo ajustado dos dados, uma vez que mede a qualidade do ajuste do modelo

de Regressão Logística Binária. No entanto, ela é utilizada para avaliar a qualidade da variável resposta. Brocco (2006) avalia que, após ajustar o modelo a um conjunto de dados, é natural questionar qual a diferença entre os valores ajustados da variável resposta Y_i na razão dos modelos e os valores observados (saturados). Se a diferença entre as observações e os correspondentes valores ajustados for pequena, então o modelo é aceito. Caso contrário, a forma corrente do modelo não será aceita e este precisará ser revisado. Uma maneira de medir a discrepância entre a probabilidade de sucesso observada Y_i e a probabilidade ajustada $\widehat{Y}$ pelo modelo assumido é através da função de verossimilhança, pois esta resume a informação que os dados fornecem sobre um parâmetro desconhecido em um dado modelo. Batista (2015) salienta que, para medir volumes grandes de amostras, será necessário considerar que a diferença entre as estatísticas de dois modelos de log-verossimilhança, designada como razão de verossimilhança, tende a revelar uma aproximação à distribuição assintótica do χ-quadrado, com $n-k$ graus de liberdade, onde n representa o número de observações e k o número de parâmetros do modelo corrente. No caso dos dados binários, onde $n_i = 1$, $i = 1, \ldots, n$, o desvio denominado taxa de verossimilhança L depende apenas das probabilidades de sucesso ajustadas à variável resposta Y_i. Para Batista (2015), a função de máxima verossimilhança na regressão logística será:

$$L(\beta_0, \ldots, \beta_k) = \prod_{i=1}^{n} \left[P_i^{Y_i} (1 - P_i)^{1-Y_i} \right] = \prod_{i=1}^{n} \left[\left(\frac{e^{Y_i}}{1 + e^{Y_i}} \right)^{Y_i} \left(\frac{1}{1 + e^{-Y_i}} \right)^{1-Y_i} \right].$$

De acordo com Brocco (2006), o logaritmo da função de verossimilhança maximizado considerando o modelo corrente é dado por:

$$\ln L_0 = \prod_{i=1}^{n} \left[Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i) \right].$$

Para o modelo completo, $P_i - Y_i$. Como $Y_i \ln Y_i$ e $(1 - Y_i) \ln(1 - P_i)$ são ambos nulos para os únicos dois possíveis valores de Y_i, situados no intervalo $[0, 1]$. Logo, $\ln L_{\text{inicial}} = 0$. Dessa forma, a taxa de verossimilhança (L) para dados binários será:

$$L = -2 \sum_{i=1}^{n} \left(P_i \ln Y_i + \ln(1 - Y_i) \right).$$

Segue que,

$$\ln L = \sum_{i=1}^{n} \ln P_i + \sum_{i=n_i+1}^{n} \ln (1 - P_i)$$

Pyndick e Rubinfeld (2004, p. 380) apontam que para obter as estimações dos parâmetros de inclinação $\widehat{\beta_0}, \ldots, \widehat{\beta_k}$, serão necessários diferenciar $\ln L$ em relação a $\beta_0, \ldots, \beta_k$

igualando os resultados a zero, ou seja,

$$\frac{\partial(\ln L)}{\partial\beta_0} = \sum_{i=1}^{n} \frac{1}{P_i}\frac{\partial P_i}{\partial\beta_0} - \sum_{i=n_i+1}^{n_i} \left(\frac{1}{1-P_i}\right)\frac{\partial P_i}{\partial\beta_0} = 0$$

$$\vdots$$

$$\frac{\partial(\ln L)}{\partial\beta_k} = \sum_{i=1}^{n} \frac{1}{P_i}\frac{\partial P_i}{\partial\beta_k} - \sum_{i=n_i+1}^{n_i} \left(\frac{1}{1-P_i}\right)\frac{\partial P_i}{\partial\beta_k} = 0$$

L é a função da verossimilhança ou da probabilidade, e ln é o logito dos parâmetros que fazem com que os valores da expressão de verossimilhança sejam maximizados.

$L_{(0)}$ é o valor da verossimilhança do modelo ajustado, ou seja, é o valor inicial da função de verossimilhança.

$L_{(\text{máx})}$ é o valor da verossimilhança do modelo saturado, ou o valor mais alto da função de máxima verossimilhança.

A estimação da função de verossimilhança L para n variáveis binárias como função dos parâmetros (β) será determinada de acordo com:

$$\ln L = \left[-2\frac{\ln L_{(0)}}{\ln L_{(\text{máx})}}\right] = -2\left[\ln L_{(0)} - \ln L_{(\text{máx})}\right]$$

De acordo com Pyndick e Rubinfeld (2004, p. 380), a técnica de estimação por máxima verossimilhança possui uma série de propriedades estatísticas desejáveis. Segundo eles, todos os estimadores de parâmetros são consistentes e eficientes assintoticamente. Além disso, como se sabe que todos os estimadores de parâmetros são normais (assintoticamente), eles podem ser aplicados a testes análogos aos testes t de regressão utilizados para determinar o nível de significância para a precisão de classificação (HAIR JR et al., 2009). Caso desejemos testar a significância de todos os coeficientes ou de um conjunto deles nos modelos logit e probit quando a estimação por máxima verossimilhança é usada, podemos aplicar o teste da razão de verossimilhança.

Assim, obtemos a função log-verossimilhança:

$$\ln L = \left\{\left[(Y_i).\ln\left(\frac{e^{Y_i}}{1+e^{Y_i}}\right)\right] + \left[(1-Y_i).\ln\left(\frac{1}{1+e^{-Y_i}}\right)\right]\right\}$$

gerando a função de máxima verossimilhança:

$$L = \prod_{i=1}^{n}\left\{\left[(Y_i).\ln\left(\frac{e^{Y_i}}{1+e^{Y_i}}\right)\right] + \left[(1-Y_i).\ln\left(\frac{1}{1+e^{-Y_i}}\right)\right]\right\}.$$

Pyndick e Rubinfeld (2004) ponderam também que o número de graus de liberdade considerados no teste da razão máxima de verossimilhança (ou teste da razão de verossi-

milhança) é dado por:

$$\ln L = -2 \ln \frac{L_{(0)}}{L_{(\text{máx})}}$$

3.3 Teste da máxima verossimilhança na Regressão Logística Binária

O teste da máxima verossimilhança ($\ln L$) é obtido pela somatória de ln da função Y_i onde

$$Y_i = \left(\frac{e^{Y_i}}{1 + e^{Y_i}} \right) = \frac{\exp^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}}}{1 + \exp^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}}},$$

e ao mesmo tempo, com a aplicação do logaritmo natural da função verossimilhança em ambos os membros da equação, chega-se à seguinte equação:

$$\ln L = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \prod_{i=1}^{n} \left[\left[(Y_i) \cdot \ln \left(\frac{e^{Y_i}}{1 + e^{Y_i}} \right) \right] + \left[(1 - Y_i) \cdot \ln \left(\frac{1}{1 + e^{-Y_i}} \right) \right] \right].$$

Neste caso, a função de verossimilhança será máxima quando os parâmetros ($\beta_0, \beta_1, \beta_2, \ldots, \beta_k$) forem estimados com base na função-objetivo. Isto é aplicado quando se deseja maximizar a equação da verossimilhança.

3.4 Níveis de significância estatística geral do modelo e dos parâmetros

De acordo com Fávero e Belfiore (2017), em modelos de regressão logística não há um coeficiente de ajuste R^2 como nos modelos tradicionais de regressão estimados pelo método de mínimos quadrados ordinários. Muitos pesquisadores apresentam em seus trabalhos um coeficiente conhecido como pseudo R^2 de McFadden, cuja expressão é dada por:

$$\text{pseudo } R^2 = \left[\frac{\ln L_{(0)} - (-2 \ln L(\max))}{-2 \ln L_{(0)}} \right]$$

Sua utilidade é bastante limitada e restringe-se a casos em que o pesquisador tem interesse em comparar dois ou mais modelos distintos. Um dos critérios existentes para a escolha do modelo é o critério de maior pseudo R^2 de McFadden, frequentemente utilizado para medir a qualidade do ajuste dos modelos estimados.

Batista (2015) demonstra que a função de verossimilhança é igual à razão entre o modelo ajustado $L_{(0)}$, que representa o máximo da verossimilhança ajustada, e o modelo de máxima verossimilhança saturado $L_{(\max)}$. Essa relação deve conter tanto os parâmetros β_j quanto observações com grau de liberdade gl para o teste χ-quadrado (χ^2). O modelo maior, representado por $L_{(\max)}$, é designado como modelo completo, enquanto o modelo

menor ou reduzido, $L_{(0)}$, é obtido igualando a zero os parâmetros β_j do modelo completo. Assim, na hipótese nula H_0 a ser testada, os parâmetros β_j do modelo completo serão iguais a zero, e o modelo completo, que mantém os valores dos seus coeficientes, representará a hipótese alternativa H_1. Para representar as taxas de variação da verossimilhança, será utilizado o teste de aderência χ-quadrado sob a hipótese nula de que todos os coeficientes são iguais a zero, o que leva à rejeição da hipótese nula e à interpretação de que ao menos um dos coeficientes seja estatisticamente diferente de zero.

$$\Big\{H_0 : \beta_1, \beta_2, ..., \beta_j = 0 \; H_1 : \text{Existe pelo menos um } \beta_j \neq 0\Big\}$$

Para $\beta_j = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ e $j = 1, 2, 3, ..., k$

Ao ajustar os parâmetros da verossimilhança na forma dicotômica e não observar o termo de erro para cada observação, Fávero e Belfiore (2017) salientam que é necessário adotar o uso da função de verossimilhança. A partir desse ponto, elabora-se a estimação da máxima verossimilhança.

3.5 Teste de Wald na modelagem da Regressão Logística Binária

De acordo com Brocco (2006), o erro padrão (denotado por *s.e.*) depende das observações binárias Y_i obtidas através das probabilidades ajustadas P_i, as quais, por sua vez, não refletem a discrepância entre as probabilidades observadas e suas correspondentes probabilidades ajustadas. Como o erro padrão não deve ser usado como medida de qualidade de ajuste para modelos ajustados a respostas binárias, utiliza-se apenas a diferença dos desvios para comparar os modelos. Essa diferença é utilizada, por exemplo, pelo método de seleção de modelos na escolha do melhor modelo. Costa (1997) argumenta que outra maneira de realizar inferência sobre os parâmetros é por meio do teste estatístico Wald, o qual normalmente é realizado quando há apenas um parâmetro inicialmente testado. De acordo com Fávero e Belfiore (2017), para uma análise mais precisa do modelo, será necessário avaliar cada um dos parâmetros do modelo de regressão logística binária. Neste sentido, a estatística Z de Wald tem a função de fornecer significância estatística para cada parâmetro a ser considerado no modelo. A denominação Z se refere ao fato de que a distribuição desta estatística é a distribuição normal padrão. Batista (2015) e Costa (1997), observam que o teste estatístico Wald é determinado pela razão entre os coeficientes estimados de interesse $\widehat{\beta}_i$ e seu erro padrão $se(\widehat{\beta}_i)$. Para grandes amostras, esse teste estatístico apresenta aproximadamente uma distribuição normal padrão $N(0, 1)$.

$$\widehat{Y}_i = \begin{cases} 1 \text{ se } P(\widehat{Y}_i = 1) = P_i \\ 0 \text{ se } P(\widehat{Y}_i = 0) = 1 - P_i. \end{cases}$$

O valor esperado da variável aleatória de Bernoulli X, pela distribuição binomial com $n = 1$, possui a seguinte característica:

$$E(Y_i) = P_i$$
$$Var(Y_i) = P(1 - P_i)$$
$$Y_i = \begin{cases} 1 \text{ se } P(Y_i = 1) = P_i \rightarrow \text{adimplente} \\ 0 \text{ se } P(Y_i = 0) = 1 - P_i \rightarrow \text{inadimplente.} \end{cases}$$
$$W_j = \frac{\widehat{\beta}_j}{s.e.(\widehat{\beta}_j)}$$

O teste de Wald possui uma distribuição χ-quadrado e é calculado pela relação entre a estimativa de máxima verossimilhança do parâmetro $(\widehat{\beta}_j)$ e a estimativa do seu erro padrão. Portanto, testa a hipótese de que um determinado coeficiente é nulo, seguindo uma distribuição χ-quadrado. Quando a variável dependente possui um único grau de liberdade, a razão entre o coeficiente que está sendo testado e o seu erro padrão pode ser elevada ao quadrado, uma vez que esse teste possui uma distribuição normal padrão $N(0, 1)$ em amostras grandes, conforme observado por (FÁVERO; BELFIORE, 2017).

Segundo Batista (2015) é importante observar que a aplicação do teste de Wald se torna relevante nas operações da Regressão Logística, uma vez que fornece significância estatística para cada parâmetro da função no modelo observado. O autor também destaca que a estatística de Wald deve considerar a distribuição normal padrão. A estatística Y_i de Wald será aplicada para testar as hipóteses de β_0 e para cada β_j $(j = 1, 2, ..., k)$, comparando a estimativa de máxima verossimilhança do parâmetro β_j com a estimativa do seu erro padrão. Sob a hipótese nula $H_0 : \beta_j = 0$, a razão resultante segue uma distribuição normal padrão.

Como veremos a seguir, os testes de hipóteses para $j = 1, 2, ..., k$ são os seguintes:

$$H_0 : \beta_0 = 0$$
$$H_0 : \beta_j = 0$$
$$H_1 : \beta_0 \neq 0$$
$$H_1 : \beta_j \neq 0$$

Moura (2018) enfatiza que $W_j = \dfrac{\widehat{\beta}_j}{s.e.\left(\widehat{\beta}_j\right)}$, onde os coeficientes $(\widehat{\beta}_j)$ estimados são divididos pelos seus respectivos erros padrões. Dessa forma, o teste de Wald (W_j) representa a razão de cada parâmetro $(\widehat{\beta}_j)$ em relação ao seu erro padrão. Ao aplicar este teste, é possível verificar se cada variável independente apresenta uma relação estatisticamente significativa com a variável dependente. Sob a hipótese nula, (W_j) segue uma distribuição

normal padrão.

As notações utilizadas são as seguintes:

- $\widehat{s.e.}$ é o erro padrão da estimativa de máxima verossimilhança;
- $\widehat{\beta}_0$ é o parâmetro estimado com intervalo de confiança dado por $\widehat{\beta}_0 \pm Z_{1-\frac{\alpha}{2}}\,\widehat{s.e.}\left(\widehat{\beta}_0\right)$.
- $\widehat{\beta}_j$, $j = 1, \ldots k$ são estimadores dos coeficientes da regressão com intervalo de confiança dado por $\widehat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}\,\widehat{s.e.}\left(\widehat{\beta}_j\right)$

De acordo com Hosmer e Lemeshow (2004), os pontos finais do intervalo de confiança baseado em Wald quando $Z_{1-\frac{\alpha}{2}}$ for superior a $100\left(1-\frac{\alpha}{2}\right)\%$, denota-se a cauda superior onde é o ponto superior da distribuição normal padrão e denota um estimador baseado em modelo do erro padrão do respectivo estimador de parâmetro e os pontos finais de um intervalo de confiança de 95 são obtidos a partir dos respectivos pontos finais do intervalo de confiança para o logit.

Para os valores $\widehat{Y}_i$ (Y estimado), temos:

$$\widehat{Y} = \left(\frac{e^{Y_i \pm Z_{1-\frac{\alpha}{2}}\left(\widehat{se}\right)Y_i}}{1 + e^{Y_i \pm Z_{1-\frac{\alpha}{2}}\left(\widehat{se}\right)Y_i}}\right)$$

Costa (1997) analisa que a estatística do teste Wald para regressão logística proporciona a capacidade de identificar se uma determinada variável independente apresenta uma relação estatisticamente significativa com a variável dependente. Por possuir uma funcionalidade aceitável, torna-se fundamentalmente importante na avaliação abrangente da regressão logística, uma vez que permite medir o grau de significância de cada coeficiente em uma equação binária.

Assim, o respectivo erro padrão é composto pela distribuição normal padrão e pelas estimativas do seu erro padrão. Fávero e Belfiore (2017) ressaltam que após a obtenção das estatísticas z de Wald, é possível recorrer à tabela de distribuição da curva normal padrão para obter os valores críticos de um determinado nível de significância, o que permite verificar se os testes rejeitam ou não a hipótese nula. No presente caso, considera-se um nível de significância de 5%, onde $z_c = -1,96$ para a cauda inferior (probabilidade de cauda inferior a 0,025 em uma distribuição bicaudal) e $z_c = 1,96$ para a cauda superior (probabilidade de cauda superior também de 0,025 em uma distribuição bicaudal).

Field (2009) destaca que esses testes são empregados para comparar os escores de uma amostra com uma distribuição normal, considerando um modelo com a mesma média e as variâncias dos valores encontrados na amostra. Caso o Teste de Hipótese não seja significativo, ou seja, $H_0 : \beta_j = 0$, e o valor de p seja maior que $0,05$, pode-se concluir que os dados da amostra não diferem significativamente de uma distribuição normal. Em contrapartida, se o Teste de Hipótese for significativo com $H_1 : \beta_j \neq 0$ e p-valor menor

que $0,05$, conclui-se que a distribuição encontrada é significativamente diferente de uma distribuição normal, ou seja, os dados são normalmente distribuídos.

3.6 Construção dos Intervalos de confiança dos parâmetros do modelo de Regressão Logística Binária

De acordo com Fávero e Belfiore (2017), assim como Hosmer e Lemeshow (2004), os intervalos de confiança para os coeficientes da expressão

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}},$$

para os parâmetros β_0 e para cada β_j, $j = 1, 2, ..., k$, ao nível de confiança de 95%, podem ser expressos da seguinte forma:

$$\beta_0 \pm 1,96[s.e.(\beta_0)],$$

e

$$\beta_j \pm 1,96[s.e.(\beta_j)].$$

O valor 1,96 é correspondente ao z_c para um nível de confiança de 95%, com um nível de confiança de 5%. Com isso, é possível calcular os coeficientes estimados dos parâmetros na expressão de probabilidade de ocorrência do evento de interesse, juntamente com os seus respectivos erros padrão, estatísticas z de Wald e intervalos de confiança com um nível de significância de 5%.

3.7 Avaliação da Modelagem de Regressão Logística Binária

Segundo Batista (2015), a avaliação do modelo logístico permite identificar a qualidade de aderência dos valores produzidos pelo modelo (valores estimados). Essa avaliação é realizada por meio de diversas estatísticas, cuja escolha dependerá daquela que melhor ajuste os valores estimados aos valores observados. Portanto, ele estabelece a comparação entre os valores estimados e os valores observados obtidos através da variância total entre os valores observados e o valor médio das observações. Assim:

$$\text{Variância total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

onde: Y_i são os valores observados e $\overline{Y}$ é o valor médio das observações.

De acordo com Batista (2015), o coeficiente de determinação R^2 é determinado pela razão da variância explicada em relação à variância total nas análises da regressão logística. Portanto, essa determinação requer cuidados especiais em suas apurações. Para ele, uma vez que a avaliação do modelo logístico consiste em verificar a qualidade de aderência dos valores produzidos pelo modelo (valores estimados) através da sua similaridade com os valores observados, nem todos os dados da regressão logística tendem a explicar sua totalidade. Assim, essa relação deve ser calculada pelas diferenças ao quadrado entre os valores observados Y_i e os valores estimados $\widehat{Y_i}$.

Por outro lado, para se obter a variância explicada, é necessário realizar a diferença entre a variância estimada Y_i pela média das variâncias observadas $\overline{Y}$ ao quadrado. Isso ocorre porque no modelo estatístico não é possível explicar a totalidade dos valores observados, devido ao surgimento de erros ou resíduos, que são inerentes a uma estimativa. Como demonstrado por Batista (2015), pode-se verificar que a variância não explicada é dada por $\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2$, e de maneira similar, a variância explicada é calculada como $\sum_{i=1}^{n}\left(\widehat{Y} - \overline{Y}\right)^2$. Assim, é possível concluir que a variância total é igual à variância explicada somada à variância não explicada, e consequentemente, o coeficiente de determinação R^2 é determinado pela razão da variância explicada em relação à variância total, como observado a seguir.

Para obter uma medida de qualidade de ajustamento análoga ao R^2, existem várias opções. Uma delas é calcular $\dfrac{1 - L_{(0)}}{L_{(\text{máx})}}$, onde é considerado que $L_{(0)}$ é o valor inicial da função de verossimilhança e $L_{(\text{máx})}$ é seu valor máximo. Uma segunda opção é calcular os resíduos $\widehat{\varepsilon_i} = Y_i - \widehat{P_i}$. Esses resíduos serão todos positivos para aqueles que escolhem a primeira opção e negativos em caso contrário. Além disso, eles diminuirão em valor absoluto à medida que o modelo melhor explicar as escolhas feitas.

Para tais resíduos, é fácil calcular um teste análogo ao R^2, ou seja, a Soma dos Quadrados dos Resíduos (SQR) e a Soma dos Quadrados Total (SQT), calculando a relação entre a variância explicada e a variância total, conforme veremos a seguir:

$$SQR = \sum_{i=1}^{n}\left(\widehat{\varepsilon_i}\right)^2 = \text{Variância não explicada,}$$

$$SQT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \text{Variância Total.}$$

Assim, o coeficiente de determinação $\widehat{R^2}$ da regressão é definido como:

$$\widehat{R^2} = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

Uma vez que R^2 é a razão entre a variação explicada e a variação total, ele é interpretado como a fração da variação amostral em Y_i que é explicada pela variável aleatória X_i.

O coeficiente R^2 mede quão bem a reta de regressão do método dos mínimos quadrados se ajusta aos dados. O valor de R^2 sempre está entre zero e um. Um valor de R^2 igual a 0 indica um ajuste inadequado da reta de regressão dos mínimos quadrados.

3.8 Função da curva ROC na modelagem da Regressão Logística Binária

Uma das ferramentas utilizadas para representar o comportamento dos tomadores de crédito é a curva ROC, pois é por meio dela que a representação da classificação da Regressão Logística Binária proporciona uma análise mais criteriosa da classificação de risco de crédito. Isso ocorre porque indica a área que o nível de adimplência ou inadimplência ocupa no espaço do plano cartesiano. De acordo com Batista (2015), a curva ROC é uma forma de representação gráfica composta por pontos cujas abscissas são as especificidades, e as ordenadas as sensibilidades, que neste caso representam as medidas de probabilidades variando entre 0 e 1. Batista (2015) e Brocco (2006) reforçam a importância da existência das duas métricas para avaliar a capacidade preditiva de um modelo de Regressão Logística Binária nas análises de risco de crédito. No entanto, eles consideram que a análise de seus comportamentos tende a indicar as porcentagens de acerto ou erro em relação ao que o modelo previu e ao que foi observado. Na análise de risco de crédito, é possível identificar o ponto de corte "cut-off" que permite separar os adimplentes dos inadimplentes. Esse ponto de corte é denominado de ponto de equilíbrio entre os dois comportamentos, pois em sua representação gráfica estabelece a área em torno do ponto 0,5. Crespi Jr, Perera e Kerr (2017) destacam que ao considerar Y_i como o score discriminante para cada candidato a crédito no intervalo de $[0, 1]$, e ao identificar que a pontuação do candidato a crédito está acima de 0,5 na curva ROC, o candidato a crédito será aceito. Do mesmo modo, se a pontuação estiver abaixo de 0,5 pontos, o candidato a crédito será recusado. Para eles, a escolha do ponto de corte na concessão de crédito ao consumidor é influenciada pelas medições dos resultados de um sistema classificador binário para diferentes pontos de corte, que por sua vez é muito utilizado para medir a sensibilidade de um modelo ou técnica para avaliação de análise de risco de crédito. Moura (2018) destaca ainda que a curva ROC é uma das técnicas mais utilizadas para identificar o desempenho da Regressão Logística. Da mesma forma, Hosmer e Lemeshow (2004, p. 162) destacam que a regra geral para avaliar os resultados da área ocupada no plano cartesiano sob a curva ROC de modelos de risco de crédito é determinada pelas seguintes métricas:

Se a área AROC = 0,5: cut-off denominado área AROC ou ponto de equilíbrio da curva ROC.

Se 0,5 < área AROC < 0,7: baixa discriminação.

Se $0,7 \leqslant$ área AROC $<$ 0,8: discriminação aceitável.

Se $0,8 \leqslant$ área AROC $<$ 0,9: discriminação excelente.

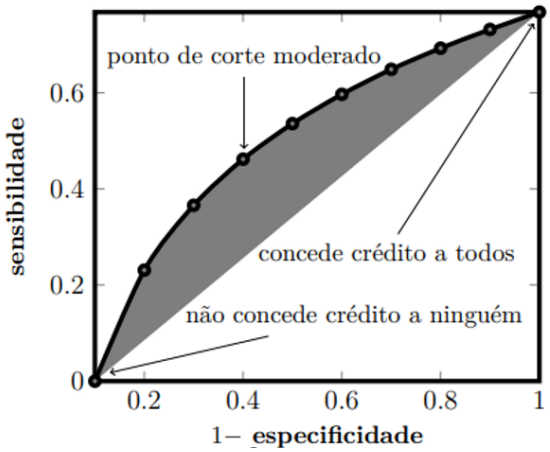
Se área AROC $\geqslant 0,9$: discriminação excepcional.

Hosmer e Lemeshow (2004, p. 162) salientam que

> ... na prática, é extremamente incomum observar a área sob a curva ROC com indicadores de área acima de 0,9, ou seja, acima de 90%, pois quando há separação completa, é impossível estimar os coeficientes de um modelo de regressão logística. No entanto, para viabilizar as análises das amostras, seria necessária uma separação quase completa para que a área sob a curva ROC fosse maior do que 90%.

A área sob a curva AROC é denominada de medida de qualidade do classificador, pois a partir daí, quanto maior for a área, melhor será o desempenho do classificador Oliveira (2015).

Figura 3.2: Área sob a Curva ROC.



Fonte: (RAMIREZ; PETTERINI, 2017).

Ao analisar a curva ROC, observa-se que o ponto de corte moderado (cut-off) indica o equilíbrio entre sensibilidade e especificidade na área representada pelo par ordenado (0,5;0,5). Nesse caso, a concessão de crédito levará em consideração o estágio comportamental do grau de risco que o tomador de crédito apresentará. Ao analisar os indicadores, os fornecedores de crédito verificam a situação posicional gráfica dos indicadores apresentados. As decisões de conceder crédito ou não serão tomadas ao propor os seguintes dilemas: quanto mais distante a área na representação gráfica estiver do ponto de equilíbrio, maior será a propensão a conceder crédito a todos os tomadores. Isso leva em consideração a classificação dos pontos na curva ROC até o limite da área no ponto (1;1). Da mesma forma, qualquer ponto abaixo do ponto de equilíbrio da área na curva ROC

resultará em uma restrição de crédito regressiva, até o limite de não conceder crédito a nenhum tomador no ponto da área (0;0).

3.9 Análises da área sob a curva ROC

De acordo com Moura (2018), a representação gráfica na curva ROC da especificidade e sensibilidade tem como característica indicar a taxa de adimplência que cada tomador de crédito apresenta na avaliação do modelo, levando em consideração os diferentes pontos de cortes. Para ela, a sensibilidade mede a capacidade que o modelo tem de classificar quando um tomador de crédito adimplente é realmente adimplente. Por outro lado, a especificidade tem a capacidade de classificar se o tomador de crédito é adimplente quando realmente for adimplente.

Para Vaz (2009), os estudos de tese importantes que proporcionam obter medidas de desempenho sob o teste de investigação, considerando os valores preditos de sensibilidade e especificidade, serão necessários usar os resultados de testes de diagnóstico das variáveis aleatórias, supondo somente valores (0 ou 1). Neste caso, e adaptando para o nosso modelo, a probabilidade (P_i) de um tomador de crédito ser inadimplente é definida pela proporção de tomadores de crédito inadimplentes observados no modelo. Na Tabela 3.1 serão demonstrados os possíveis resultados de um teste diagnóstico proposto por (VAZ, 2009), denominado como classificação teste padrão ouro. Para ela, este teste também é chamado de teste de referência, justamente porque engloba todos os procedimentos auferidos no modelo. No caso da pesquisa em questão, será adaptado para identificar se um tomador de crédito é adimplente ou se é inadimplente.

Tabela 3.1: Tabela de classificação.

	Padrão ouro		
	Adimplente	**Inadimplente**	**Total**
Teste Adimplente	VA	FA	VA+FA
Teste Inadimplente	FI	VI	FI+VI
Total	VA+FI	FA+VI	VA+FA+FA+VI

Fonte: Adaptado pelo autor a partir da Figura 2.1 de Vaz (2009, p. 8).

sendo:

VA = Verdadeiro Adimplente.

FA = Falso Adimplente.

FI = Falso Inadimplente.

VI = Verdadeiro Inadimplente.

Para determinar o valor das FPR (false positive rate) e TPR (true positive rate)

TPR = Sensibilidade (Taxa de Verdadeiro Adimplente) = 1

FPR = 1- Especificidade (Taxa de Falso Adimplente)

Para encontrar a sensibilidade para avaliar a capacidade de detectar inadimplentes, é necessário supor que

$$\text{Sensibilidade} = \frac{VA}{VA + FI} \quad \text{ou} \quad 1 - \frac{VA}{VA + FI}$$

e a

$$\text{Especificidade} = \frac{FA}{FA + VI}$$

e seu complemento é dado por:

$$1 - \frac{FA}{FA + VI}, \quad \text{ou} \quad 1 - \text{Especificidade}$$

para ocaso individualizado na amostra. A partir de então é que se determina a acurácia para adimplente, para inadimplente e para o total da amostra.

3.9.1 Análise do Score para crédito à pessoa física através da análise da Regressão Logística binária

O Boletim do Banco Central (BRASIL, 2021) apresenta detalhes da regulamentação da Lei nº 12.414/2011, que estabelece a implantação do projeto de lei do cadastro positivo. Esta lei tem como propósito reduzir o *spread* de crédito, baratear os custos dos empréstimos e alavancar o crescimento da indústria de crédito no mercado financeiro e creditício. Observa-se que esta lei leva em consideração a importância da regulamentação da disponibilidade de dados dos credores para auferir as classificações dos riscos de adimplência e/ou inadimplência por parte de cada tomador de crédito. Por isso, o desafio das instituições credoras está diretamente relacionado com a utilização de metodologias eficientes e transparentes para promover as apurações dos indicadores de risco de crédito de forma equitativa. O que se pode dizer é que, através da Lei do Cadastro Positivo, surge a expectativa de uma evolução nas análises das notas de crédito de cada tomador, para atender a uma parcela da população (classificada como demanda reprimida correspondente a 60% dos brasileiros que pertencem às classes C, D e E), como identificado no período da implantação dessa lei.

Ainda de acordo com o Boletim do Banco Central (BRASIL, 2021), o cadastro positivo é um banco de dados que reúne o histórico de pagamentos e obrigações de pagamento em andamento, tanto de pessoas físicas quanto de jurídicas, registrados no Sistema Financeiro Nacional. Essas informações são utilizadas para disponibilização do histórico de crédito, mediante a autorização do tomador, motivando a formação da nota de crédito (*score*) para permitir que a análise de concessão ou extensão de crédito, ou outras transações

com risco financeiro, sejam feitas de forma mais precisa e segura. Essa exposição tende a gerar expectativas de redução das taxas do *spread* de crédito, culminando na queda das taxas de juros cobradas nos empréstimos e financiamentos para consumidores e empresas.

Com a evolução das tecnologias da informação, a estabilidade econômica e financeira e o alto grau de competitividade do mercado financeiro e de crédito, surge a necessidade de proporcionar uma dinâmica equitativa no mercado de crédito brasileiro. Neste caso, a saída foi proporcionar o surgimento da proposta do Cadastro Positivo para modernizar e democratizar o fornecimento de crédito, como forma de aumentar a concorrência no Sistema Financeiro Nacional. A base primordial dessa lei é fornecer benefícios que permitem às empresas creditícias proporcionar mais segurança às pessoas físicas ou jurídicas que concedem e tomam créditos ou realizam operações comerciais a prazo ou não.

O Banco Central do Brasil, por ser um órgão que regulamenta as operações financeiras e creditícias no Brasil, atribuiu a responsabilidade pela gestão do Cadastro Positivo às Gestoras de Banco de Dados, classificadas como (GBDs), denominadas agências como: Boa Vista Serviços S.A, Confederação Nacional de Dirigentes Lojistas (CNDL - SPC Brasil), Gestora de Inteligência de Crédito S.A. (Quod), Serasa S.A e Trans Union Brasil Sistemas em Informática Ltda. Essas agências têm como função administrar as carteiras de tomadores de créditos com o objetivo de promover a inclusão de informações no Cadastro Positivo, tanto oriundas de pessoas físicas quanto de pessoas jurídicas. No Brasil, as informações obtidas em relação aos contratos de crédito são fornecidas pelo processamento de dados do Banco Central e disponibilizadas mensalmente ao Sistema de Informações de Créditos (SCR), e em seguida transferidas aos (GBDs). Estas agências apuram e divulgam mensalmente os scores de adimplência e/ou inadimplência de cada tomador de crédito, a partir da modelagem matemática, estatística e econométrica das seguintes variáveis qualitativas e quantitativas: pagamentos de serviços como contas de água, luz, gás e telefone dentro do prazo de vencimento, histórico de dados cadastrais, histórico de compras a prazo, financiamento e empréstimos, portador de cartão de crédito e correntista de instituições bancárias. A nota de grau de risco (score de crédito) de cada tomador de crédito apurada deverá proporcionar as tomadas de decisões por parte dos credores na questão de oferta e demanda por crédito no mercado financeiro e creditício. Para GONÇALVES, GOUVÊA e MANTOVANI (2013), além do desenvolvimento do modelo de credit scoring, que permite classificar os tomadores de crédito em adimplente e inadimplentes, as instituições financeiras também fundamentam suas decisões nas definições de performances, com análises subjetivas diretamente ligadas às decisões políticas de crédito da própria instituição. O tomador de crédito que honrar em dia seus compromissos financeiros e creditícios terá facilidade para acessar linhas de crédito em melhores condições. Seguindo a lógica conceitual das propostas da Lei do Cadastro Positivo, surgiu a necessidade de apresentar uma prática como exemplo de pesquisa que, em conformidade com as diretrizes das variáveis elencadas no corpo da lei, será de suma importância

para entender a viabilidade da aplicação da metodologia matemática e econométrica da modelagem de Regressão Logística Binária.

3.10 Aplicação do modelo

Será apresentado a seguir um exemplo de aplicação envolvendo 49 pessoas tomadores de crédito acima de 18 anos, com perfis mercadológicos, financeiros e sociais diferentes, com o objetivo de compreender a viabilidade da aplicação da modelagem da Regressão Logística Binária. A fundamentação desta análise baseia se na ideia de identificar o perfil de risco de crédito de cada tomador por intermédio das regras e normas da lei do cadastro positivo com aplicação da modelagem de Regressão Logística Binária. Dentro desta lógica, pretendemos identificar as metodologias que permitem classificar em adimplentes e/ou inadimplentes cada análise observada, para proporcionar uma viabilidade de aplicação nas tomadas de decisões quanto ao fornecimento de crédito a pessoa física. Dessa forma, esta pesquisa foi realizada sob as orientações propostas e contidas na Lei do Cadastro Positivo regulamentada pela Lei nº 12.414/2011, e alterada pela Lei Complementar nº 166 (de 8 de abril de 2019).

3.11 Descrição dos dados para apuração dos resultados

A amostra consistirá em um conjunto de dados financeiros, bancários e pessoais de uma amostra composta por 49 tomadores de crédito. Eles foram categorizados em riscos de adimplência, como uma variável dependente binária, e riscos de inadimplência, também como uma variável dependente binária. As variáveis explicativas independentes contidas na base de pesquisa foram classificadas como qualitativas, categóricas e numéricas, totalizando onze variáveis independentes, além de uma variável dependente de saída que informa se o tomador de crédito é adimplente ou inadimplente.

Para a aplicação experimental do modelo proposto, foi necessário utilizar o software Solver do Excel, empregando exclusivamente o método de Newton. Além disso, como estratégia comparativa para testar a veracidade dos resultados, utilizamos o Software MedCalc, no qual aplicamos a metodologia "enter"como orientação para o formato de entrada e saída proposto. Também usamos as descrições das variáveis descritivas conforme definidas na Tabela 3.2.

Tabela 3.2: Descrição das variáveis explicativas de categoria bancária e pessoal dos tomadores de crédito.

Variável Explicativa	**Descrição da Variável**	**Tipo de Variável**	**Categorias**
Conta corrente	Possui conta corrente	Binária	0: não e 1: sim
Cartão de crédito	Possui cartão de crédito	Binária	0: não e 1: sim
CDC	Possui Crédito Direto ao consumidor (CDC)	Binária	0: não e 1: sim
Empréstimo consignado	Possui Crédito Consignado em folha de pagamento	Binária	0: não e 1: sim

Fonte: Elaborada pelo autor.

Tabela 3.3: Descrição das variáveis explicativas de categoria bancária e pessoal dos tomadores de crédito.

Variável Explicativa	**Descrição da Variável**	**Tipo de Variável**	**Número de Categorias**	**Categorias**
Renda mensal	Remuneração mensal individual	Categórica	3	1 : $1 \leq x < 5$ salários-mínimos; 2 : $5 \leq x < 10$ salários-mínimos; 3 : $10 \leq x < 15$ salários-mínimos
Emprego fixo	Registro com carteira de trabalho ou contrato com vínculo empregatício há mais de um ano	Binária	2	1: com registro funcional/funcionário público com comprovante de renda; 0: se autônomo sem registro funcional ou sem contrato temporário, nenhuma comprovação de renda
Imóvel residencial	Casa própria ou alugada	Binária	2	0: aluguel; 1: casa própria.
Dependentes quantidade	Dependentes (números de filhos menores de 18 anos)	Categórica	4	0: nenhum dependente; 1 : $1 \leq x < 3$ dependentes; 2 : $3 \leq x < 5$ dependentes; 3 : $x \geq 5$ dependentes.
Luz	Paga em dia conta de luz residencial	Binária	2	0: Não e 1: Sim
Água	Paga em dia conta de água residencial	Binária	2	0: Não e 1: Sim
Telefone fixo	Paga em dia conta de telefone residencial	Binária	2	0: Não e 1: Sim

Fonte: Elaborada pelo autor.

Para a apuração e organização dos dados da pesquisa foi necessário estabelecer parâ-

metros das variáveis qualitativas (categóricas) binárias e nominais. Neste caso, utilizamos para as variáveis na coluna correspondente como sim igual a 1 e na coluna correspondente como não, consideramos igual a 0. Para as variáveis numéricas discretas que identificam o número de dependentes e as quantidades de salários (renda mensal), utilizamos variáveis numéricas. Para a análise comparativa na questão da renda, utilizamos como base o salário-mínimo nacional vigente desde 1º de maio de 2023, no valor de R$1320,00.

3.12 Análises dos resultados da Regressão Logística Binária

Para a comparação entre duas aplicações, foi necessário tabular e organizar inicialmente os dados em uma planilha eletrônica, considerando as regras estabelecidas para classificar de forma ordenada e por ordem de interesse, em ordem crescente e por grau de importância de cada variável independente (dummy). Após a organização, foi necessário alimentar o software Solver e o software MedCalc para determinar a relevância dos resultados das variáveis observadas no modelo. As tabelas geradas encontram-se no Apêndice A (Tabelas A.1 e A.2).

A aplicação pelo método solver foi importante pois mostrou de forma clara e objetiva as classificações em tabela das variações desejadas. O mesmo foi gerado de forma sintetizada no método MedCalc, como pode ser observado na Tabela 3.4.

Tabela 3.4: Análise da Curva de Regressão Logística - Software MedCalc (Método Enter)

Tamanho da amostra	49	100%
Scores positivos, tomadores de crédito com probabilidade de adimplentes (Suc-Pred)	33	67,35%
Scores negativos, tomadores de crédito com probabilidade de inadimplentes (Fail-Pred)	16	32,65%
Variável dependente Dummy (Y_i) -	Adimplente = 1 Inadimplente = 0	

Fonte: Elaborada pelo autor a partir da aplicação no software Solver Excel e MedCalc.

Na Tabela 3.4 observamos que a aplicação da amostra gerou uma saída tanto pelo método de Newton pelo solver Excel, quanto pelo método enter do software MedCalc, observou-se que os resultados gerados pelos dois modelos das probabilidades de adimplentes e de inadimplentes foram os mesmos nos dois casos da amostra.

Rosa (2000) salienta que a identificação do perfil de cada tomador de crédito pode ser obtida através da realização de regressões dos parâmetros com variáveis categóricas e

binárias, permitindo a localização e identificação do grupo ao qual o tomador de crédito pertence. No caso em questão, ao aplicar a modelagem de regressão logística binária, é possível estabelecer a distinção entre a probabilidade de adimplência, que corresponde a 67,35%, e a probabilidade de inadimplência, que é de 32,65%, apresentada pelo total de tomadores de crédito observados na amostra.

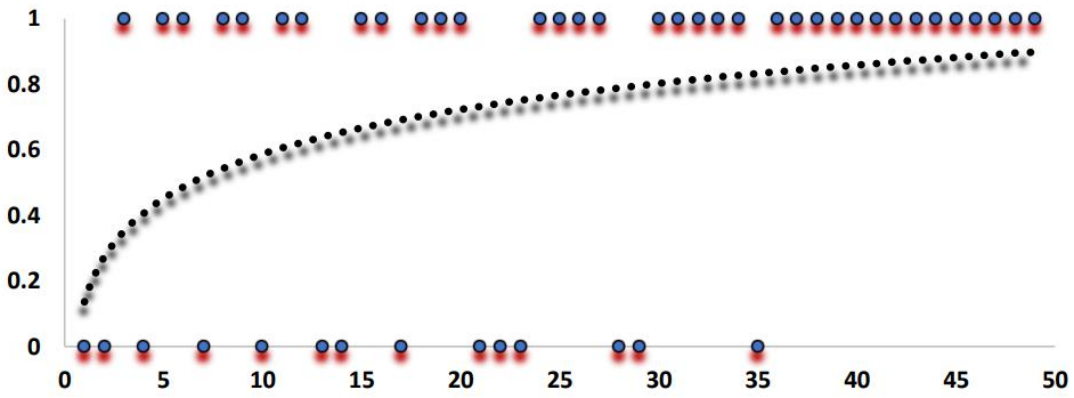
A classificação dos resultados pode ser explicada com base em Fávero e Belfiore (2017), que propõem que as duas categorias, por definições metodológicas, podem ser consideradas como eventos de interesse. Ou seja, "adimplente" quando a categoria ($DummyY = 1$) e não de interesse "inadimplente" quando a categoria ($DummyY = 0$).

A representação da curva de regressão logística binária na Figura 3.3, considerou-se a função inversa

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki})}}$$

para $P(Y_i = 1) > 0,5$ e $Y_i = P_i$, ou seja, $Y_i = 1$ e se $P(Y_i = 1) < 0,5$, então $Y_i = 1 - P_i$, ou seja, $Y_i = 0$. Esta notação será válida para identificar o perfil de cada tomador de crédito. Pela representação obtida, o eixo das ordenadas mede o grau de probabilidade de adimplência e de inadimplência onde formou-se uma representação extrema (superior e inferior). Supondo que somente a variável dependente tem valores 0 e 1, permite estabelecer uma representação indicando os pares relacionando as abscissa, quando $Y_i = 1$ e $Y_i = 0$.

Figura 3.3: Pontos de dispersão desejada da função Logística Binária inversa das probabilidades de Adimplência ou inadimplência.



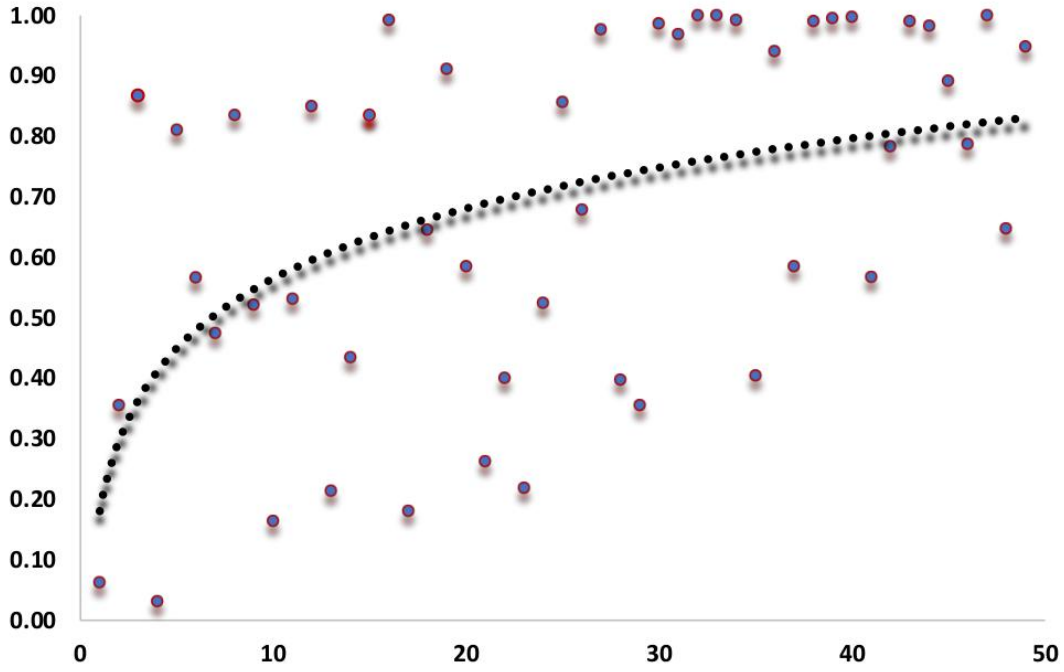
Fonte: Elaborada pelo autor a partir da aplicação do sofware Solver-Excel.

Esta representação separa os grupos de acordo com seu grau de risco, distinguindo entre inadimplentes e adimplentes. Isso permite que se tenha uma visão transparente dos indicadores de risco para cada tomador de crédito. Quanto ao gráfico de dispersão, pode-se dizer que o gráfico fornece uma visão clara das possíveis relações globais entre duas variáveis.

Pela representação gráfica de dispersão, é possível observar que há uma concentração

das probabilidades de adimplência próxima a $Y_i = 1$, o que indica que a qualidade da probabilidade de adimplência oferece uma avaliação positiva do modelo. Por outro lado, pode-se observar que a probabilidade de inadimplência, quando $Y_i = 0$, se dispersa ao longo do eixo vertical representado pela função $P_i = f(Y_i)$. Fávero e Belfiore (2017) salientam que quando $X_{ki} \to \infty$, $P_i(Y_i = 1) \to 1$ e quando $X_{ki} \to -\infty$, $P_i(Y_i = 1) \to 0$. Assim, $E(Y_i)$ está localizada entre zero e um, conforme demonstrado em "$0 \leqslant E(Y_i) = P_i \leqslant 1$", com $\beta_j = (\beta_0, \beta_1, \ldots, \beta_k)$ e $\beta_j \neq 0$.

Figura 3.4: Pontos de dispersão obtida da função Logística Binária inversa das probabilidades de Adimplência ou Inadimplência.



Fonte: Elaborada pelo autor a partir da aplicação do sofware Solver-Excel.

A representação gráfica da probabilidade de risco de crédito (score de crédito) permite identificar e classificar os potenciais credores ou devedores ao analisar o comportamento dos scores de crédito por meio de um gráfico de dispersão. Essa representação permite visualizar os indicadores de risco de crédito com probabilidade acima de 50% para adimplentes e abaixo de 50% para inadimplentes, o que possibilita obter uma visualização rápida e detalhada do comportamento da variável dependente em relação aos dados amostrais $P_i = f(Y_i)$.

De acordo com a lógica dessa representação gráfica, é possível mencionar que, se um tomador de crédito apresentar uma probabilidade de risco de adimplência ou inadimplência, isso não afeta o resultado dos demais tomadores da amostra, pois são eventos independentes.

3.13 Estimação do modelo de regressão logística binária por máxima verossimilhança

Para Pyndick e Rubinfeld (2004, p. 379), todos os estimadores dos parâmetros são consistentes e eficientes assintoticamente, o que permite aplicar não apenas o teste t de regressão, mas também o teste de estimação por máxima verossimilhança. Neste caso, para estimar os parâmetros do modelo de regressão logística, será utilizado

$$P_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})}},$$

sendo P_i a probabilidade de um indivíduo fazer uma escolha específica, dada a variável X_{ki}.

O teste da razão da verossimilhança permite observar que P_i mede a qualidade do ajuste do modelo de regressão logística binária, justamente porque ajusta a qualidade da variável resposta (Y_i).

De acordo com Brocco (2006), se a diferença entre as observações e os valores ajustados correspondentes for pequena, o modelo é aceito. Caso contrário, a forma atual do modelo não será aceita e precisará ser revisada. Uma maneira de medir a discrepância entre a probabilidade de adimplência observada $(Y_i = 1)$ e a probabilidade de inadimplência $(Y_i = 0)$ é utilizar a função de máxima verossimilhança na regressão logística, como demonstrado a seguir:

$$\ln L(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} \left[P_i^{Y_i}(1 - P_i)^{1-Y_i}\right]$$

ou

$$\ln L(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} \left[\left(\frac{e^{Y_i}}{1 + e^{Y_i}}\right)^{Y_i} \left(\frac{1}{1 + e^{-Y_i}}\right)^{1-Y_i}\right]$$

Para tanto, em uma amostra com n observações, define-se a função de verossimilhança como

$$P_i = \frac{e^{Y_i}}{1 + e^{Y_i}} \quad \text{e} \quad f(Y_i) = \frac{1}{1 + e^{-Y_i}}, P_i(Y_i = 1) = \frac{P_i}{1 - P_i}.$$

Para Y_i temos,

$$L = \left[(Y_i) \ln \left(\frac{e^{Y_i}}{1 + e^{Y_i}}\right)\right]$$

e para P_i,

$$L = \left[(1 - Y_i) \ln \left(\frac{1}{1 + e^{-Y_i}}\right)\right].$$

Conforme Brocco (2006), o logaritmo da função de verossimilhança maximizado será

obtido pela seguinte expressão:

$$\ln L_{(0)} = \prod_{i=1}^{n} \left[Y_i \ln P_i + (1 - Y_i) \ln(1 - Y_i) \right],$$

para determinar o modelo completo, $P_i = Y_i$, utiliza-se $[Y_i \ln Y_i]$ e $[(1 - Y_i) \ln(1 - Y_i)]$ considerando zero para os únicos dois possíveis valores de Y_i, situados no intervalo $[0, 1]$. Para realizar os testes de qualidade de ajustes calcula-se a taxa de verossimilhança (L) para dados binários obtidos pela equação

$$L = -2 \sum_{i=1}^{n} \left[P_i \ln Y_i + \ln(1 - Y_i) \right].$$

Na aplicação do modelo proposto para calcular a máxima verossimilhança e estimar a qualidade de ajuste do modelo, será necessário considerar os dois modelos de log-verossimilhança. Para ajustar o modelo, também será necessário considerar que a variável dependente Y_i é igual a 0 quando os parâmetros $(\beta_0, \beta_1, \ldots, \beta_1 1) = 0$. A somatória do logaritmo da função de máxima verossimilhança $\ln L_{(0)} = -30,953$, quando os parâmetros $(\beta_0, \beta_1, \ldots, \beta_1 1) \neq 0$ geraram $\ln L_{(\text{máx})} = -19,7078$. Na aplicação do teste de significância por meio do teste do Qui-Quadrado, utilizaram-se as diferenças entre os dois modelos $-2 \ln L_{(0)}$ e $-2 \ln L_{(\text{máx})}$.

Ao analisar a Tabela 3.4, é possível observar que o modelo possui características aceitáveis para ser utilizado como base de informações de risco de crédito, uma vez que, durante a aplicação dos dados pelo Software Solver, gerou uma saída de $(Y_i) = 33$ para os adimplentes e $(1 - Y_i) = 16$ para os inadimplentes, resultados exatamente iguais aos obtidos pelo MedCal. Para Fávero e Belfiore (2017), determinar a função-objetiva que é a função que deverá ser maximizada, será necessário realizar a somatória do logaritmo da função de verossimilhança. No caso, $\ln L_{(\text{máx})} = -19,7078$.

Os valores obtidos da função de log-verossimilhança são utilizados como base comparativa entre dois modelos, $\ln L_{(0)}$ e $\ln L_{(\text{máx})}$, para gerar os possíveis pseudos R^2, os quais têm como característica a produção de valores que representam o ajuste geral do modelo na explicação da eficácia e da aderência significativa da amostra. A partir de sua determinação, será possível organizar os dados gerados para determinar os preditores que contribuem para a identificação das variações registradas na variável dependente. A partir desse ponto, será possível obter orientações sobre a acurácia correta das análises dos modelos propostos.

Na Tabela 3.5, estão organizados os estimadores de probabilidade que serão utilizados para identificar o escore de crédito a partir dos conjuntos das variáveis independentes e da variável dependente, considerando os valores previstos e os valores observados.

Tabela 3.5: Estimadores das probabilidades de adimplência ou inadimplência

$\ln L_(0)$ Log-verossimilhança	-30,953
$\ln L_{\text{máx}}$ Log-verossimilhança	-19,7078
n (tamanho da amostra (Y_i)	49
Chi-Sq (Qui-Quadrado)	22,4905
GL (Grau de Liberdade)	11
Alpha (α)	0,05
p-valor $<\alpha$	0,020837
s.e. (erro padrão)	sim
Intervalo de Confiança 95%	0,1952 a 0,4578

Fonte: Elaborada pelo autor a partir da aplicação no software Solver.

Aplicando as funções de resultados da Tabela 3.5, nas equações dos modelos propostos, tem-se os seguintes preditos:

- Teste do Qui-Quadrado:

$$\ln L = [-2 \ln L^{(0)}_{\text{máx}}] = [-2 \ln L_{(0)} - (-2 \ln L_{\text{máx}})]$$

- Qui-Quadrado: $(\chi^2) = 2(\ln L_{(\text{máx})} - \ln L_{(0)})$.

De acordo com Fávero e Belfiore (2017), o teste do Qui-Quadrado fornece a significância do modelo, proporciona verificar a existência ou não do modelo. Neste caso, o modelo proposto com $\beta_0 \neq 0$ e $\beta_j \neq 0$, o comportamento de alteração não influencia a probabilidade de ocorrência do evento.

$$\text{Qui-Quadrado } (\chi^2) = 2(-19,7078 - (-30,953)) = 22,4905.$$

Ainda em Fávero e Belfiore (2017), para 11 graus de liberdade, ou seja, variáveis explicativas consideradas na modelagem, temos o Qui-Quadrado significante igual a 22,4905. O valor crítico obtido na tabela do Qui-quadrado é igual a 19,675 para o nível de significância de 0,05, considerando que β_j sejam iguais a zero.

Tem-se ainda que:

$$\begin{aligned}[\text{McFadden R-Sq(L) do pseudo } R^2] &= \left[\frac{-2 \ln L_{(0)} - (-2 \ln L_{\text{máx}})}{-2 \ln L_{(0)}}\right] \\ &= \left[\frac{-2(-30,953) - (-2(-19,7078)}{-2(-30,953)}\right] = 0,3633.\end{aligned}$$

$$\begin{aligned}[\text{Cox \& Snell } R^2] \quad R - \text{Sq}(CS) &= 1 - e^{\frac{-2}{n}\left(\ln L_{\text{máx}} - \ln L_{(0)}\right)} \\ &= 1 - e^{\frac{-2}{49}(-19,7078-(-30,953)} = 0,368078.\end{aligned}$$

$$[\text{Nagelkerke } R^2] \quad R - \text{Sq}(N) = \frac{R - Sq(CS)}{1 - e^{2\left(\frac{\ln L_{(0)}}{n}\right)}} = \frac{0,368078}{1 - e^{2\left(\frac{-30,953}{49}\right)}} = 0,51314.$$

Tabela 3.6: Teste de Hosmer & Lemeshow

Qui-Quadrado	7,0477
GL (Grau de liberdade)	8
p-valor	0,5315

Fonte: Elaborada pelo autor a partir da aplicação

O teste Hosmer and Lemeshow (Tabela 3.6), que apresenta um Qui-Quadrado de 0,70477, indica a não significância do modelo, uma vez que o seu p-valor resultou ser maior do que 0,05 de nível de significância para um grau de liberdade de 8 variáveis observadas na amostra. A orientação segundo GONÇALVES, GOUVÊA e MANTOVANI (2013, p. 155) é a não rejeição da hipótese nula do teste, pois a não existência de diferenças significativas entre os valores preditivos e observados na amostra, permite afirmar que não há diferenças significativas entre os resultados observados. Conclui-se que os resultados apresentaram um estágio logístico não significativo e, portanto, deve-se considerar que os dados são adequados para realizar uma regressão logística.

Tabela 3.7: Possíveis pseudos R^2, considerados para explicar as variações da variável dependente em relação aos dados amostrais.

Modelo nulo $-2\log$ Verossimilhança $-2\ln L_{(0)}$	61,906
Modelo completo $-2\log$ Verossimilhança $-2\ln L_{(1)}$	39,416
Qui-quadrado (χ^2) Chi-Sq	22,4905
Graus de Liberdade (GL)	11
Nível de significância α	0,05
p-valor	0,020837
Cox & Snell R^2 R-Sq(CS)	0,3681
Nagelkerk e R^2 R-Sq(N)	0,51314
Mc Fadden R-Sq(L) R	0,3633

Fonte: Elaborada pelo autor a partir da aplicação Solver e MedCalc.

Ao submeter os dados da amostragem no Software Solver, foi possível observar que o modelo em questão produziu valores de saída adequados para realizar uma interpretação dos estimadores de adimplência e inadimplência. Com a amostra com df = 11 graus de liberdade e as variações no valor $-2\ln L_{(0)}$ também para o modelo ajustado, com Qui-Quadrado = 22,4905 maior que Qui-Quadrado = 19,675 (obtido na tabela do Qui-Quadrado) para um nível de significância de 5% e um p-valor de 0,020837, ou seja, com p-valor $< \alpha$, deve-se rejeitar H_0 de que todos os parâmetros sejam estatisticamente iguais

a zero conforme mostra a Tabela 3.7. Uma vez que o p-valor é significativo, garante a informação de que o modelo está bem ajustado. No entanto, pelo menos uma variável X é estatisticamente significativa para explicar a probabilidade de ocorrência do evento em estudo. Teremos desta forma, um modelo de regressão logística binária estatisticamente significante para fins de previsão. Neste caso, as relações entre a classificação realizada e a observada não acusam diferenças.

Para Field (2009), se o p-value $< \alpha$, então há evidência de que pelo menos uma das variáveis independentes contribui para a aprovação do resultado. Essa análise é utilizada para obter o grau de acurácia do modelo logístico, situação que tem a finalidade de testar e verificar se existem diferenças significativas entre as duas classificações realizadas pelo modelo em comparação com a realidade observada. Como base complementar da análise, Hair Jr et al. (2009) salientam que os pseudos R^2 de Nagelkerke e Cox & Snell têm a mesma finalidade para explicar a variável dependente, pois produzem um ajuste perfeito para a variável dependente, fornecendo resultados no intervalo entre 0 e 1. De acordo com eles, quanto mais próximo de 1 for o pseudo R^2, melhor será o ajuste para o modelo. Neste caso, pode-se concluir que, nas condições a que foram submetidas, o pseudo R^2 de Nagelkerke é o que melhor explica as variações registradas na variável dependente, pois é capaz de explicar em torno de 51,31% das variações registradas na variável dependente. Este mecanismo é utilizado para comparar o desempenho de modelos concorrentes entre duas equações logísticas igualmente válidas como $\ln L_{(0)}$ e $\ln L_{(1)}$, e seus resultados permitem identificar a validade da aplicação do modelo na regressão logística binária. Pelo que se observa nas técnicas utilizadas, conclui-se que o modelo utilizado para a base da pesquisa com relação aos riscos de crédito apresenta uma acurácia positivamente aceitável.

Outra modelagem também utilizada para avaliar a acurácia do modelo é pela determinação dos indicadores de critérios de informação. Dentre os diversos modelos observados, Fávero e Belfiore (2017) salientam que os critérios de informações Akaike corrigido (AIC) e Bayesiano (Schwarz) Corrigido (BIC) são utilizados com muita frequência na aplicação de modelos com pequenas amostras. Para eles, quanto maior for a quantidade de variáveis no modelo, maior será o indicador, e, portanto, maiores serão também os desajustes dos dados. Emiliano (2009) também destaca que o critério de informação de AIC, assim como BIC, avaliam a verossimilhança do modelo e aplicam uma penalidade por adicionar variáveis ao modelo. Quanto maior for o número de variáveis no modelo, maior tende a ser o valor de AIC. A redução de variáveis pode produzir um modelo com melhor desempenho geral. Assim, a expressão que define os critérios de informação de AIC será

$$AIC = -2 \times (\text{a função suporte maximizada}) + 2 \times (\text{número de parâmetros})$$

ou

$$AIC = -2 \ln L_{\text{máx}} + 2(\theta),$$

sendo $\theta = DF + 1$

$$AIC = -2 \times (-19.7078) + 2 \times (11 + 1) = 63,4156.$$

Para Emiliano (2009), o critério de informação AIC desejável é aquele que apresenta o menor valor possível. Como nem sempre o modelo com o menor valor para um conjunto de preditores necessariamente ajusta bem os dados, será necessário, além disso, usar os testes e os gráficos dos resíduos para avaliar se o modelo ajusta bem os dados ou não. A expressão para a determinação do Critério de Informação Bayesiano é definida pela expressão

$$BIC = -2 \ln L_{\text{máx}} + (\theta) \times \ln n = -2 \times (-19,7078) + (11 + 1) \times \ln(49) = 86,1174436.$$

Ao observar os testes de informações de $BIC = 86,117$ e o critério de informação de $AIC = 63,4156$, é possível afirmar que o teste de informação utilizado no modelo apresentou classificador corretamente aceitável.

Tabela 3.8: Estimativa das probabilidades dos parâmetros e das variáveis independentes, do modelo logístico e avaliação dos riscos de crédito.

Variáveis	**Coeficiente**		**Erro Padrão**	**Teste de Wald**	**p-valor**	**expoente de β**	**Intervalo de Confiança**	
							inferior	**superior**
Constante (intercepto)	β_0	-3,91	2,54	2,37	0,12	0,02	0,00	0
Renda mensal (X1i) Dummy 1	β_1	2,4	1,43	2,82	0,09	11,05	0,67	181,49
Emprego fixo (X2i) Dummy 2	β_2	-1,39	1,22	1,3	0,25	0,25	0,02	2,71
Casa Própria (X3i) Dummy 3	β_3	2,46	1,18	4,32	0,04	11,69	1,15	118,71
Conta Corrente (X4i) Dummy 4	β_4	1,85	1,9	0,95	0,33	6,38	0,15	265,21
Cartão de Crédito (X5i) Dummy 5	β_5	-2,03	1,06	3,69	0,05	0,13	0,02	1,04
CDC (X6i) Dummy 6	β_6	1,9	1,23	2,4	0,12	6,71	0,6	74,52
Empréstimo consignado (X7i) Dummy 7	β_7	1,96	1,25	2,47	0,12	7,12	0,61	82,5
Dependentes (X8i) dummy 8	β_8	0,76	0,56	1,83	0,18	2,14	0,71	6,45
Luz (X9i) Dummy 9	β_9	1,04	1,32	0,62	0,43	2,82	0,21	37,42
Água (X10i) Dummy 10	β_{10}	-2,05	1,32	2,41	0,12	0,13	0,01	1,71
Telefone fixo (X11i) Dummy 11	β_{11}	-1,68	1,33	1,59	0,21	0,19	0,01	2,53

Fonte: Elaborada pelo autor a partir da aplicação do Solver e MedCalc.

O expoente β indica a quantidade de vezes que as chances de ser adimplente aumentam. A renda tem um fator expoente de $\beta = 11,05$, o que significa que o tomador de crédito tem 11 vezes mais chances de se tornar adimplente a cada aumento na renda de 2,40 vezes. No caso da casa própria, o tomador de crédito tem 11,69 vezes mais chances de se tornar adimplente a cada aumento na quantidade de casas próprias em 21,46 vezes. Se expoente de $\beta > 1$ o indicador de inadimplência tende a diminuir. Por outro lado, se expoente de $\beta < 1$, o indicador de inadimplência tende a aumentar. Com relação ao expoente β, se seu valor for maior do que 1, um aumento no p-valor da variável eleva a probabilidade de adimplência. Por outro lado, se seu valor for menor do que 1, um aumento no p-valor da variável promove uma redução na probabilidade de adimplência. Quando seu valor for igual a 1, pode-se dizer que a probabilidade de sucesso permanece inalterada. $\beta_0, \beta_2, \beta_5, \beta_9, \beta_{10}$ e β_{11} são parâmetros cujos expoentes apresentaram valores menores do que 1, e um aumento nos valores do intercepto e das variáveis correspondentes, como emprego fixo, cartão de crédito, luz, água e telefone, provoca uma redução nas probabilidades de adimplência dessas variáveis para o modelo.

O expoente de β é um fator muito significativo para a análise comportamental do modelo. No caso observado, de todas as variáveis submetidas ao teste, a que está demonstrando significância é a variável casa própria e a variável renda, pois ambas produzem p-valores próximos de 0,05, indicando um grau de significância adequado. Para o modelo observado, conclui-se que ter renda e casa própria contribui para a redução do risco de inadimplência do tomador de crédito. Para encontrar a chance real que permita identificar a quantidade de vezes que a variável analisada irá sofrer alterações, será necessário realizar o cálculo $\frac{1}{\text{expoente de } \beta_1}$, pois, fazendo $\frac{1}{11,05}$ para a variável renda e $\frac{1}{\text{expoente de } \beta_3} = \frac{1}{11,69}$ para a variável casa própria, obtemos valores do expoente de β próximos de zero, que permitem dizer que ambas as variáveis analisadas são as mais significativas para a análise de risco de crédito para o modelo considerado.

3.14 Aplicando o Teste Wald

A estatística z de Wald dos parâmetros $\beta_2, \beta_4, \beta_8, \beta_9$ e β_{11} apresentou valores dos intervalos de confiança entre -1,96 e 1,96, o que indica que ao nível de significância de 0,05 para esses casos, não houve rejeição da hipótese nula. Portanto, esses parâmetros não podem ser considerados estatisticamente diferentes de zero, uma vez que os p-valores analisados foram maiores do que 0,05. A não rejeição da hipótese nula para os parâmetros $\beta_2, \beta_4, \beta_8, \beta_9$, β_{10} e β_{11}ao nível de significância de 5%, significa que as correspondentes variáveis, emprego fixo, conta corrente, dependentes, luz e felefone fixo, não são estatisticamente significativas para aumentar ou diminuir a probabilidade de um tomador de crédito apresentar riscos de inadimplência. Em outras palavras, quando comparadas

com as outras variáveis explicativas no modelo, esses parâmetros podem ser excluídos do modelo, pois não causariam nenhuma alteração nas variáveis explicativas.

Por outro lado, a estatística z de Wald para o coeficiente β_0 e para os parâmetros $\beta_3, \beta_5, \beta_6, \beta_7$ e β_{10} indica que a renda mensal gerou p-valor de 0,09, enquanto casa própria teve p-valor de 0,04, o que indica que existe pelo menos um parâmetro $\beta_i \neq 0$ onde há evidência de que pelo menos um dos parâmetros β_i contribui para a validação dos resultados.

Aplicando o teste z de $W_j = \left(\frac{\widehat{\beta_j}}{\widehat{s.e.}(\widehat{\beta_j})}\right)^2$ para a análise da amostra, obtém-se:

$$z \text{ de } W_{\beta_0} = \left(\frac{-3,91}{2,54}\right)^2 = 2,3696603$$
$$z \text{ de } W_{\beta_1} = \left(\frac{2,40}{1,43}\right)^2 = 2,81676366$$
$$\vdots$$
$$z \text{ de } W_{\beta_{11}} = \left(\frac{-1,68}{1,33}\right)^2 = 1,59556788$$

Como observado no teste de Wald, se os parâmetros β's em relação aos seus erros apresentam uma relação estatisticamente significativa entre as variáveis independentes e a variável dependente, seus valores se tornam fundamentalmente importantes para avaliar os comportamentos das variáveis no modelo de Regressão Logística Binária. Isso ocorre porque isso permite medir o grau de significância de cada coeficiente em uma equação binária.

3.15 Aplicação da modelagem como base exploratória da aprendizagem

Para determinar a probabilidade estimada, será necessário utilizar o logito cujo parâmetros obtidos pelo Software solver permitem identificar os graus de riscos de adimplência ou inadimplência de um tomador de crédito.

Dado que

$$P_i = \left(\frac{e^{Y_i}}{1+e^{Y_i}}\right) = \frac{e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}}}{1+e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}}},$$

então o intervalo de confiança do valor esperado de $\widehat{Y}$, para $P(X_{ki}$, aplicando no modelo teremos

$$P(X_{ki}) = \frac{e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}}}{1+e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}}}$$

ou

$$\text{p-pred} = P_i = f(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}) = \frac{e^{Y_i}}{1 + e^{Y_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})}}$$

substituindo os valores dos coeficientes e das variáveis na função geral obtém-se a probabilidade P de um tomador de crédito i, apresentar score credit, substituindo os valores, obtemos

$$\begin{aligned} Y_i &= -3,91 + 2,40X_{1i} + (-1,39)X_{2i} + 2,46X_{3i} + 1,85X_{4i} + (-2,03)X_{5i} + 1,90X_{6i} \\ &\quad + 1,96X_{7i} + 0,76X_{8i} + 1,04X_{9i} + (-2,05)X_{10i} + (-1,68)X_{11i} \\ P_i &= \frac{1}{1 + e^{-Y_i}} \end{aligned}$$

assim determinamos a probabilidade estimada de risco de inadimplência do tomador de crédito 1 (TC1),

$$\text{p-pred} = P(TC1) = 0,063.$$

Os dados gerados tanto pelo Solver do Excel quanto pelo MedCalc foram exatamente iguais para as análises da estatística das estimativas das probabilidades dos parâmetros e das variáveis independentes. A preferência por detalhar com maior importância os dados do Solver ocorre porque os dados gerados e as explicações detalhadas das informações se tornaram mais elucidativos para o entendimento das demonstrações observadas na modelagem. Como pode ser observado, o preditor no caso do tomador de crédito TC1 mostra que a probabilidade média estimada de inadimplência foi de aproximadamente 6,30%.

3.16 Análise e interpretação da Tabela ROC e da Curva ROC (Receiver Operator Characteristic)

De acordo com Batista (2015), Brocco (2006) e Oliveira (2015), uma das técnicas utilizadas para analisar a eficácia do modelo geral de regressão logística binária é a interpretação da curva ROC, gerada a partir da tabela ROC. Nela, a curva ROC é determinada com base nas informações obtidas a partir da saída dos dados gerados. No entanto, a análise da eficiência da Regressão Logística Binária leva em consideração as seguintes observações: a área sob a curva, os pares ordenados (sensibilidade e especificidade), a acurácia e o ponto de corte (cutoff). Neste caso, a leitura e interpretação desses dados se tornam essenciais para identificar os comportamentos dos riscos de inadimplência de um determinado tomador de crédito.

Neste sentido, para determinar a viabilidade da aceitação da amostra em estudo, foi possível, através da metodologia do Solver no Excel, organizar os dados fornecidos na

tabela de classificação geral ROC. A partir daí, foram aplicadas técnicas que permitem identificar a acurácia dos modelos de risco de crédito.

Tabela 3.9: Classificação Geral ROC

Tomadores	***p-Pred***	***Failure***	***Success***	***Fail-Cum***	***Suc-Cum***	***FPR***	***TPR***	***AUC***
TC1	0	0	0	0	0	1	1	0,0625
TC2	0,0308	1	0	1	0	0,9375	1	0,0625
TC3	0,0628	1	0	2	0	0,875	1	0
TC4	0,1639	0	1	2	1	0,875	0,97	0,0606
TC5	0,1808	1	0	3	1	0,8125	0,97	0
TC6	0,2135	0	1	3	2	0,8125	0,939	0,0587
TC7	0,2185	1	0	4	2	0,75	0,939	0,0587
TC8	0,2625	1	0	5	2	0,6875	0,939	0,0587
TC9	0,3557	1	0	6	2	0,625	0,939	0,0587
TC10	0,3557	1	0	7	2	0,5625	0,939	0,0587
TC11	0,3974	1	0	8	2	0,5	0,939	0
TC12	0,4004	0	1	8	3	0,5	0,909	0,0568
TC13	0,4043	1	0	9	3	0,4375	0,909	0
TC14	0,4346	0	1	9	4	0,4375	0,879	0,0549
TC15	0,4743	1	0	10	4	0,375	0,879	0
TC16	0,5218	0	1	10	5	0,375	0,848	0,053
TC17	0,525	1	0	11	5	0,3125	0,848	0
TC18	0,531	0	1	11	6	0,3125	0,818	0,0511
TC19	0,5668	1	0	12	6	0,25	0,818	0
TC20	0,5676	0	1	12	7	0,25	0,788	0,0492
TC21	0,5846	1	0	13	7	0,1875	0,788	0
TC22	0,5846	0	1	13	8	0,1875	0,758	0
TC23	0,6456	0	1	13	9	0,1875	0,727	0,0455
TC24	0,6474	1	0	14	9	0,125	0,727	0,0455
TC25	0,6787	1	0	15	9	0,0625	0,727	0
TC26	0,7833	0	1	15	10	0,0625	0,697	0
TC27	0,7874	0	1	15	11	0,0625	0,667	0
TC28	0,8112	0	1	15	12	0,0625	0,636	0
TC29	0,8354	0	1	15	13	0,0625	0,606	0
TC30	0,8354	0	1	15	14	0,0625	0,576	0
TC31	0,85	0	1	15	15	0,0625	0,545	0,0341
TC32	0,8569	1	0	16	15	0	0,545	0
TC33	0,8675	0	1	16	16	0	0,515	0
TC34	0,8916	0	1	16	17	0	0,485	0
TC35	0,911	0	1	16	18	0	0,455	0
TC36	0,9405	0	1	16	19	0	0,424	0
TC37	0,9481	0	1	16	20	0	0,394	0
TC38	0,9687	0	1	16	21	0	0,364	0
TC39	0,9772	0	1	16	22	0	0,333	0
TC40	0,9826	0	1	16	23	0	0,303	0
TC41	0,9869	0	1	16	24	0	0,273	0
TC42	0,9903	0	1	16	25	0	0,242	0
TC43	0,991	0	1	16	26	0	0,212	0
TC44	0,9921	0	1	16	27	0	0,182	0
TC45	0,9923	0	1	16	28	0	0,152	0
TC46	0,9952	0	1	16	29	0	0,121	0
TC47	0,9974	0	1	16	30	0	0,091	0
TC48	0,9998	0	1	16	31	0	0,061	0
TC49	0,9999	0	1	16	32	0	0,03	0
							TOTAL	0,8693

Fonte: Elaborada pelo autor a partir da aplicação do Solver.

A partir da Tabela 3.9, foi possível organizar e calcular a situação real estimada e a previsão do modelo observado, para então determinar a probabilidade dos casos classificados corretamente da amostra.

Tabela 3.10: Tabela de Confusão do Padrão Ouro – Classificação geral do modelo

	Previsão do Modelo (observado)		
Situação Real (estimado)	**Adimplência (Y=1)**	**Inadimplência (Y=0)**	**Total**
Probabilidade de adimplência (Y=1)	29 (a)	6 (b)	35 (n_3)
Probabilidade de inadimplência (Y=0)	4 (c)	10 (d)	14 (n_4)
Total	33	16	49 (N)
Acurácia (porcentagem de casos classificados corretamente)	0,878788	0,625	0,795918
	87,87%	**37,50%**	**79,59%**

Fonte: Elaborada pelo autor

Sendo assim, para uma compreensão mais clara das previsões de acurácia, será necessário determinar as demonstrações da Sensibilidade, da Especificidade, da precisão e da acurácia geradas no modelo estimado e previsto, conforme contido na tabela de classificação geral da ROC. Após organizar os dados na tabela de classificação geral do modelo (Tabela de confusão do padrão-ouro), será possível demonstrar os resultados das modelagens junto com suas respectivas fórmulas

$$TPR = \text{Sensibilidade } \frac{VA}{VA + FI} = \frac{29}{29 + 4} = 0,878787 - 87,87\%$$

sendo TFP a taxa de verdadeiro adimplente e FPR a taxa de falso adimplente.

$$FPR = \text{Especificidade } = \frac{FA}{FA + VI} = \frac{6}{6 + 10} = 0,375 = 37,5\%$$

$$\text{Precisão } = \frac{VA}{VA + VI} = \frac{29}{29 + 6} = 0,828571 = 82,86\%$$

$$\text{Acurácia } = \frac{VA + VI}{VA + FA + VI + FI} = \frac{29 + 10}{29 + 6 + 10 + 4} = \frac{39}{49} = 0,79591837 = 79,59\%$$

A Tabela 3.10 mostra a classificação geral, permitindo identificar os comportamentos de todas as medidas da amostra, ao ponto de oferecer uma compreensão clara e objetiva dos principais indicadores, tais como: acurácia, sensibilidade, especificidade e precisão. Sendo assim, fica nítido que dos 49 tomadores de crédito observados, 29 foram classificados como verdadeiros adimplentes e 4 como falsos inadimplentes, proporcionando um posicionamento na curva ROC com taxa de sensibilidade (TPR) de 87,87%. Da mesma forma, para os casos de inadimplentes e falsos adimplentes classificados como taxa de especificidade (FPR), 6 tomadores de crédito foram classificados como falsos adimplentes e 10 como verdadeiros inadimplentes, perfazendo um total de 16 tomadores com probabilidade de inadimplência de crédito em relação ao total da amostra de 37,5%. Quando se observa o quadro geral do modelo, é importante notar que o indicador de precisão (que mede a qualidade dos verdadeiros adimplentes, em comparação com totais das probabilidades de adimplência, foi de 82,86%. Esse resultado simboliza a qualidade do modelo por estar bem próximo de 100%. Outra medida também significativa foi a da acurácia de 79,59%, que correspondente a porcentagem geral dos casos classificados corretamente, sinalizando que os casos observados foram corretamente classificados como aceitável. A análise, permitiu identificar o perfil dos grupos dos tomadores de crédito dispostos na amostra e ao mesmo tempo, proporcionou diagnosticar a validade do modelo pela exposição das variações entre adimplentes e inadimplentes dos totais de tomadores de crédito analisados.

Tabela 3.11: Intervalo de Confiança para área da curva ROC 95%

Cutoff (ponto de corte)	0,5
Alpha	0,05
AUC	0,869318 = 86,93%
Intervalo de Confiança 95%	0,7771087 a 0,967549
Qui-Quadrado	13,40076
s.e	0,050118
p-valor $> \alpha$	0,202119881
GL	10
Sensibilidade (Taxa de Verdadeiro Adimplente)	87,87%
Especificidade (Taxa de Falso Adimplente)	62,50%
Acurácia (Total das classificações corretas)	79,59%

Fonte: Elaborada pelo autor.

A análise das informações amostrais da tabela ROC se baseou na extração dos dados dos indicadores que sinalizam a qualidade do modelo, permitindo diagnosticar as probabilidades de adimplência e inadimplência de cada tomador de crédito que compõem a amostra. A partir desse ponto, foi necessário calcular o Qui-Quadrado, o Intervalo de Confiança de 95% para um nível de significância de 5%, e o p-valor, utilizando as informações fornecidas como dados de saída do Software Solver no Excel, tais como o grau de

liberdade $GL = 11$, AUC (Área Sob a Curva) de 86,93%, $n_1 = 33$, $n_2 = 16$,

$$q_0 = AUC(1 - AUC),$$
$$q_1 = \frac{AUC}{2 - AUC} - (AUC)^2,$$
$$q_2 = \frac{2AUC}{1 + AUC} - (AUC)^2.$$

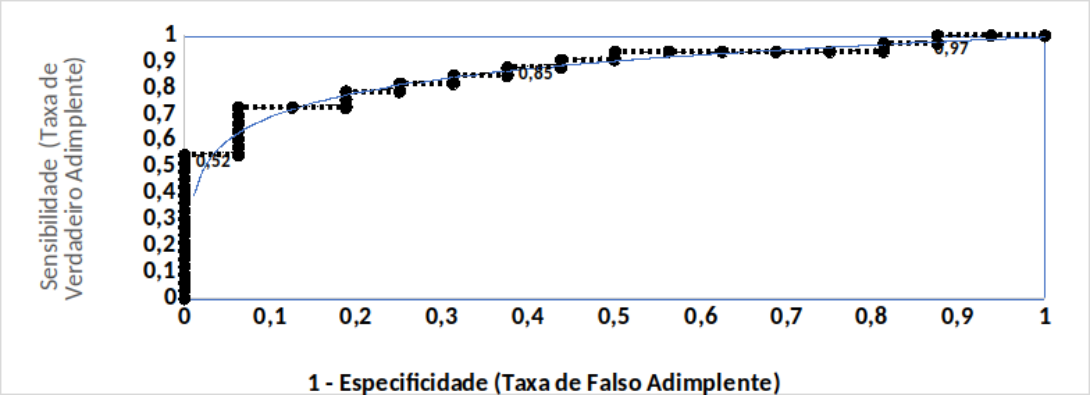
A base das informações para calcular o $s.e$ foi fornecido pelo solver Excel, adaptado pela modelagem proposta por (BRAGA, 2000).

$$s.e = \sqrt{\frac{q_0 + (n-1)q_1 + (n_2 - 1)q_2}{n_1 n_2}}$$
$$\text{Qui-Quadrado } (\chi^2) = \frac{bc - ad)^2 N}{n_1 n_2 n_3 n_4} = \frac{(24 - 290)^2 49}{33 \times 16 \times 35 \times 14} = 13,40075758$$
$$IC_{0,95} = AUC - 1,96 s.e; AUC + 1,96 s.e$$
$$IC_{0,95} = 0,869318 - 1,96 \times 0,050118; 0,869318 + 1,96 \times 0,050118$$
$$IC_{0,95} = (0,77108672; 0,96754928)$$

$IC_{0,95}$ indica que 95% das vezes, a probabilidade média de inadimplência estará contida no intervalo de 0,77108672 e 0,96754928.

Nesta análise de eficiência da regressão logística binária para o risco de crédito, leva-se em consideração os dados dos cálculos da qualidade dos ajustes, munidos com seus valores obtidos a partir da tabela de classificação geral do modelo, e sua representação e interpretação estão relacionadas com as análises dos comportamentos das sensibilidades e das especificidades representadas na curva ROC. A análise destes indicadores é fundamental para diagnosticar a qualidade da probabilidade dos preditores da amostra. Dessa forma, a área sob a curva, conhecida como AUC (área sob a curva), é utilizada como uma medida de qualidade do modelo, sendo nela que se caracterizam os indicadores da Sensibilidade e das Especificidades, demonstrando sob a área da curva os comportamentos dos ratings de crédito do tomador de crédito, dado o modelo e o momento analisado.

Figura 3.5: Curva ROC das probabilidades preditivas para análise de risco de crédito.



Fonte: Elaborada pelo autor a partir da aplicação do sofware Solver-Excel.

3.17 Análise da curva ROC

Ao analisar a Tabela 3.11 comparando com a Figura 3.5, conclui-se que os indicadores de qualidade da acurácia total, da taxa de especificidade e da taxa de sensibilidade, demonstram a sua utilidade para viabilidade dos resultados desejados. Pelas análises adaptados em (BRAGA, 2000), é possível dizer que esses indicadores são importantes para a avaliação do desempenho das taxas de adimplentes e de inadimplentes. Fatores determinantes para proporcionar o acompanhamento das estruturas da saúde creditícia dos tomadores de crédito.

De acordo com Fávero e Belfiore (2017) e Hosmer e Lemeshow (2004), o modelo terá poder discriminatório excelente quando a área da curva estiver entre $0,8 \leqslant$ Área AROC $< 0,9$. Neste caso, os resultados obtidos pela aplicação do modelo geraram um indicador da área sob a curva AROC de 86,93% com seu p-valor = 0,202119881, ou seja um p-valor $> \alpha$, pois para o intervalo de confiança de 95% com nível crítico de 0,05 e -1,96 e +1,96 variando entre 0,7771087 e 0,967549. Isso indica que 95% das vezes, a probabilidade média de inadimplência estará contida no intervalo. Pela análise da hipótese nula H_0, não devemos rejeitá-la, portanto, consideramos que tanto a probabilidade de adimplência quanto a de inadimplência observada na curva ROC indicam que não há variações consistentes entre as duas relações. Dessa forma, rejeitamos H_1 a conclusão de que a área sob a curva do modelo está fora do intervalo citado pelos autores. Essas análises permitem concluir que os dados localizados entre adimplentes e inadimplentes numa estrutura global evidenciam uma acurácia (porcentagem de casos classificados corretamente) com resultados satisfatórios, em torno de 79,59% de adimplência para o caso observado e acurácia de 62,50% para a previsão de inadimplência. Chegamos à conclusão de que os dados gerados são aceitáveis como poder de classificação, pois o nível de desempenho geral (acurácia) está entre 70% e 90%. Isso indica que o risco de crédito para os tomadores analisados na amostra é

moderado, o que pode ser observado pelo indicador da acurácia no sentido estatístico geral, denotado pela proximidade dos cálculos ou estimações do modelo com seus valores reais, permitindo indicar a validade do teste diagnóstico do modelo proposto, como observado na Tabela 3.10 de classificação geral do modelo.

Na representação gráfica da probabilidade de adimplência, dos 33 tomadores de crédito classificados como adimplentes, 29 são classificados como verdadeiros adimplentes, 4 com falsos adimplentes na representação gráfica estão localizados superior ao cutoff $>$ 0,5. Dos 16 analisados como inadimplentes, 10 são classificados como verdadeiros inadimplentes e 6 como falso inadimplentes, caracterizando comportamentos abaixo do cutoff $<$ 0,5. Quando se observa a relação entre verdadeiro adimplente e verdadeiro inadimplente sobre o total da amostra, nota-se que essa relação atingiu uma classificação corretamente aceitável. Vale ressaltar que o cutoff representa o ponto de corte em equilíbrio das probabilidades de adimplência e inadimplência dos tomadores de crédito observados na amostra. A adaptação da classificação de análise de risco de crédito ao modelo de avaliação internacional de rating da Agência Standard & Poor's segue a modelagem mensurada na amostra. A representação dos critérios de classificação de risco foi estabelecida por ordem crescente, tomando como base a menor probabilidade de inadimplência (classificado como AAA) e a maior probabilidade de inadimplência (classificado como DD). Portanto, o cálculo para a classificação das probabilidades de classificação dos ratings dos tomadores de crédito seguiu a seguinte metodologia: TC1 (Tomador de Crédito 1) apresentou a probabilidade de risco de crédito (score credit) no valor de 6,28%, conforme observado anteriormente, e sua classificação, para representar na tabela de rating, será AAA. Este indicador sinaliza o grau de risco (rating) de inadimplência do tomador de Crédito TC1, e sua localização na tabela de classificação de risco fica na faixa vermelha da tabela, indicando que o tomador de crédito está em inadimplência. Como seu grau de risco de inadimplência é de 93,72 ele é classificado como tomador de crédito em inadimplência. Portanto, sua classificação geral fica na faixa vermelha da tabela, indicando a necessidade de uma provisão de severidade de 100% na concessão de crédito.

Por outro lado, o tomador de crédito TC32 apresentou a probabilidade de adimplência de 99,99%, classificado como AAA conforme a Agência Standard & Poor's, e sua probabilidade de inadimplência é de 0,01%. Portanto, sua classificação de rating fica na faixa verde da tabela, indicando que o tomador de crédito tem probabilidade de risco mínimo de inadimplência, e a provisão de severidade para ele é de 0,0%, na composição para análise de risco pelo mercado creditício.

Tabela 3.12: Escala de avaliações de risco de crédito adapta do modelo da agência Standard & Poor's

Standard & Poor's	**Tomadores de Crédito**	**Probabilidade de inadimplência (%)**	**Classificação do Risco (Rating) de inadimplência**
AAA **AA+**	**TC32; TC33; TC47; TC40; TC39**	0,01%; 0,01%; 0,02%; 0,26%; 0,48%	Risco Mínimo
AA – **A +** **A** **A –**	**TC34; TC16; TC38; TC43; TC30; TC44; TC27**	0,77%; 0,79%; 0,90%; 0,97%; 1,31%; 1,74%; 2,28%	Risco baixo
BBB+ **BBB** **BBB –**	**TC31; TC49; TC36; TC19; TC45**	3,13%; 5,19%; 5,95%; 8,90%; 10,84%	Risco Moderado
BB+ **BB** **BB -**	**TC3; TC25; TC12; TC15; TC8; TC5**	13,25%; 14,31%; 15,00%; 16,46%; 16,46%; 18,88%	Risco substancial
B + **B** **B –**	**TC46; TC42; TC26; TC48; TC18; TC20; TC37, TC41; TC6; TC11; TC24; TC9**	21,26%; 21,67%; 32,13%; 35,26%; 35,44%; 41,53%; 41,54%; 43,24%; 43,32%; 46,90%; 47,50%; 47,82%	Alto Risco
CCC + **CCC** **CCC -** **CCC** **CC** **C** **CC** **C**	**TC7; TC14; TC35; TC22; TC28; TC2; TC29; TC21; TC23; TC13; TC17; TC10**	52,57%; 56,54%; 59,56%; 59,96%; 60,26% ;64,43%; 64,43%; 73,75%; 78,15%; 78,65%; 81,92%; 83,61%	Risco muito alto
DDD **DD**	**TC1; TC4**	93,72%; 96,92%	Em inadimplência

Fonte: Elaborada pelo autor.

Como observado na lei do cadastro positivo, o modelo proposto para identificar riscos de crédito foi satisfatório, pois, pela aplicação dos dados gerados, tanto no Software Solver Excel quanto no MedCalc, permitiram realizar uma classificação de score de crédito que se adaptou ao modelo e à avaliação de rating conforme a agência Standard & Poor's. Isso sinalizou uma precisão na classificação dos tomadores de crédito entre adimplentes e inadimplentes. A modelagem da Regressão Logística Binária permitiu obter informações críticas de forma consistente e eficiente para análises de riscos de crédito, sugerindo que sua aplicação para esses casos proporciona uma reflexão sobre a viabilidade na sua implantação no mercado financeiro e creditício.

4 Considerações finais

A dissertação tem como objeto de estudo a promoção do entendimento de uma metodologia que pode ser aplicada para calcular os riscos de crédito de um grupo específico de tomadores. A proposta para o desenvolvimento da pesquisa surgiu da necessidade de realizar um estudo com dados reais para a conclusão do curso de mestrado profissional em matemática. Como base central dessa discussão, tomou-se como parâmetro a implantação da Lei do Cadastro Positivo para abordar a mensuração do score de crédito, pois o acesso ao cálculo de risco de crédito divulgado pelas principais agências de análise de riscos no Brasil e pelo mercado financeiro e creditício ainda é um tabu.

Este estudo permitiu concluir que a viabilidade da Lei do Cadastro Positivo depende da análise de variáveis dicotômicas e binárias, assim como dos coeficientes β que compõem os resultados das probabilidades de adimplência e/ou inadimplência. A aplicabilidade desta modelagem torna-se útil no que diz respeito à metodologia que será desenvolvida. No caso em questão, utilizou-se a modelagem matemática e estatística da Regressão Logística Binária, fundamentada em um estudo econométrico que, quando aplicado, possibilitou a análise de riscos de crédito de uma amostra composta por 49 tomadores de crédito.

Os resultados obtidos foram analisados por meio do uso de técnicas estatísticas e econométricas para demonstrar as características de riscos de cada um dos tomadores de crédito que compõem a amostra. Os eventos probabilísticos analisados fundamentaram a demonstração inicial da teoria dos riscos e suas características, sendo um ponto chave para o entendimento das metodologias e aplicações da modelagem.

Após as análises, foi possível realizar um estudo detalhado das variáveis dependentes e independentes, que, submetidas aos testes de viabilidade, apresentaram seus poderes discriminantes e independentes. A aplicação do teste de verossimilhança e dos testes de Wald, de Hosmer & Lemeshow, Cox & Snell R^2, Nagelkerke R^2, permitiu demonstrar que os estudos da Regressão Logística Binária para avaliações individuais das variáveis binárias e das técnicas da máxima verossimilhança são as mais adequadas para explicar os resultados que permitem identificar a validade da aplicação do modelo na Regressão Logística Binária.

As técnicas utilizadas para a análise e conclusão deste estudo demonstraram que o modelo utilizado para a base da pesquisa, com relação aos riscos de crédito, apresenta

acurácia positivamente aceitável. Na avaliação das variáveis pelo método "enter"do software MedCalc, foi possível identificar as variáveis com maior poder discriminante entre o grupo de tomadores de créditos adimplentes e inadimplentes. Os coeficientes observados pela estruturação do modelo e das variáveis dependentes e independentes permitiram identificar as relações entre os preditores que proporcionam as probabilidades de adimplentes e de inadimplentes, apresentando uma acurácia favorável ao modelo representada pela curva ROC e pela área AUC definida pelos indicadores de Sensibilidade e de Especificidade.

O modelo permitiu observar também o bom desempenho das expectativas de adimplência, onde foi observada, na Tabela 3.10, uma taxa de acerto no valor de 87,87%, e na contrapartida, apresentou 62,50% de indício de inadimplência, totalizando uma acurácia (porcentagem de casos classificados corretamente) no valor de 79,59%. A mesma demonstração pode ser observada na curva ROC, apresentando uma área AUC de 86,93%, erro padrão de 0,050118 e intervalo de confiança de 95% entre 0,7771087 para a cauda inferior e 0,967549 para a cauda superior, indicando uma classificação equitativa do modelo estudado.

A pesquisa de campo realizada possibilitou obter um panorama de análise de risco de crédito do mundo real, justamente porque a média frequência de perfis inadimplentes pode ser caracterizada pelo tamanho da amostra e pela qualidade das informações que compõem as variáveis. Isso permitiu observar que, mesmo com as expectativas de adimplência, o modelo apresentou uma prevalência de inadimplência de 63,3%, sinalizando para o credor ser parcimonioso na avaliação da propensão ao risco de crédito.

Devido ao modelo apresentar uma taxa de acerto geral de 87,76%, é possível afirmar que é um bom modelo na questão da avaliação geral. O mesmo se confirma quando os holofotes apontam para a avaliação da curva ROC, sendo uma das práticas mais importantes na modelagem da regressão logística. Chega-se à conclusão de que os objetivos, que são demonstrar a importância da modelagem da Regressão Logística Binária para a aplicação da análise de riscos de crédito, foram favoravelmente adequados quando se consideram as propostas previstas na Lei do Cadastro Positivo.

A aplicação de dados reais em modelagens estatísticas e econométricas formalizou o aprendizado e a maturidade em relação à importância de relacionar a matemática com estatística e economia para compreender as aplicações matemáticas na vida real. Como previsto inicialmente, a proposta deste estudo e a construção de uma modelagem de análise de risco de crédito (Credit Scoring) foram desenvolvidas para medir o risco de inadimplência, com a intenção de proporcionar a viabilidade de identificação de risco de crédito de pessoas físicas e, assim, possibilitar uma política equitativa de fornecimento de empréstimos e financiamentos por parte dos agentes credores.

Referências

ANDERSON, R. **The Credit Scoring Tookit. Theory and Practice for Retail Credit Risk Management and Decision Automation**. 1. ed. New York: Oxford University Inc., 2007. Citado na página 20.

BATISTA, A. S. **Regressão Logística: uma introdução ao modelo estatístico – Exemplo de aplicação**. Lisboa: Editora Vida Econômica, 2015. Citado 8 vezes nas páginas 28, 30, 31, 32, 34, 35, 36 e 56.

BRAGA, A. C. da S. **Curvas ROC: Aspectos funcionais e aplicações**. Tese (Doutorado) — Universidade do Minho, Braga, 2000. Citado 2 vezes nas páginas 61 e 62.

BRASIL, B. B. C. do. **Análise dos efeitos do Cadastro Positivo**. [s.n.], 2021. Disponível em: <https://www.bcb.gov.br/content/publicacoes/Documents/outras_pub_alfa/analise_dos_efeitos_do_cadastro_positivo.pdf>. Acesso em: 20 set. 2023. Citado na página 39.

BROCCO, J. B. **Ponderação de Modelos com Aplicação em Regressão Logística Binária**. Dissertação (Mestrado) — Universidade Federal de São Carlos, São Carlos, 2006. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/4599>. Acesso em: 16 jul. 2023. Citado 5 vezes nas páginas 28, 31, 36, 47 e 56.

CAOUETTE, J.; ALTMAN, B.; NARAYANAN, P. **Gestão do Risco de Crédito: O próximo grande desafio financeiro**. Rio de Janeiro: Ed. Qualitymark, 1999. Citado 3 vezes nas páginas 17, 18 e 21.

COSTA, S. C. da. **Regressão logística aplicada na identificação de fatores de risco para doenças em animais domésticos**. Dissertação (Mestrado) — Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 1997. Disponível em: <https://doi.org/10.11606/D.11.2019.tde-20191218-164446>. Acesso em: 17 jul. 2023. Citado 2 vezes nas páginas 31 e 33.

CRESPI JR, H.; PERERA, L. C. J.; KERR, R. B. Gerenciamento do ponto de corte na concessão do crédito direto ao consumidor. **RAC Ampad**, Rio de Janeiro, v. 21, n. 2, p. 269–285, mar./abr. 2017. Citado na página 36.

DUARTE, I. M. R. **Modelo de Avaliação de Risco de Crédito**. Dissertação (Mestrado) — Instituto Superior de Contabilidade e Administração do Porto. Instituto Politécnico de Porto, Porto, 2014. Citado na página 20.

EMILIANO, P. C. **Fundamentos e aplicações dos critérios de informação: Akaike e Bayesiano**. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras, 2009. Citado 2 vezes nas páginas 51 e 52.

FÁVERO, L. P.; BELFIORE, P. **Manual de Análise de dados Estatística e modelagem multivariada com Excel, SPSS e Stata**. 1. ed. Rio de Janeiro: Elsevier, 2017. Citado 14 vezes nas páginas 23, 24, 27, 30, 31, 32, 33, 34, 45, 46, 48, 49, 51 e 62.

FIELD, A. **Descobrindo a estatística usando o SPSS**. 2. ed. Porto Alegre: Artmed, 2009. Citado 2 vezes nas páginas 33 e 51.

GONÇALVES, E. B.; GOUVÊA, M. A.; MANTOVANI, D. M. N. Análise de risco de crédito com o uso de regressão logística. **Revista Contemporânea de Contabilidade**, v. 10, n. 20, p. 139–160, 2013. Citado 2 vezes nas páginas 40 e 50.

HAIR JR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. Citado 2 vezes nas páginas 29 e 51.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. New York: Wiley, 2004. (Applied Logistic Regression). Citado 5 vezes nas páginas 33, 34, 36, 37 e 62.

MOURA, G. M. **Regressão logística aplicada a risco de crédito**. [S.l.], 2018. Disponível em: <https://imef.furg.br/images/stories/Monografias/Matematica_aplicada/2018/2018_Gabriela.pdf>. Acesso em: 15 ago. 2023. Citado 3 vezes nas páginas 32, 36 e 38.

OLIVEIRA, P. H. M. A. **Detecção de fraudes em cartões: um classificador baseado em regras de associação e regressão logística**. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulo, São Paulo, 2015. Disponível em: <doi:10.11606/D.45.2016.tde-01022016-204144>. Acesso em: 21 ago. 2023. Citado 2 vezes nas páginas 37 e 56.

PYNDICK, R. S.; RUBINFELD, D. L. **Econometria: Modelos & Previsões**. Rio de Janeiro: Elsevier, 2004. Citado 6 vezes nas páginas 25, 26, 27, 28, 29 e 47.

RAMIREZ, B.; PETTERINI, F. C. O risco visto a posteriori e o risco imputado a priori nos contratos de um banco de desenvolvimento. **Revista Brasileira de Finanças (online)**, Rio de Janeiro, v. 15, n. 1, p. 135–166, 2017. Citado na página 37.

ROSA, P. de T. M. **Modelos de "Credit Scoring": Regressão Logística, CHAID e REAL**. Dissertação (Mestrado) — Instituto de Matemática e Estatística - USP, São Paulo, 2000. Citado 2 vezes nas páginas 23 e 44.

SARTORIS, A. **Estatística e introdução à econometria**. São Paulo: Saraiva, 2003. Citado na página 24.

SAUNDERS, A. **Medindo o Risco de Crédito: Novas abordagens para *value at risk* e outros paradigmas**. Rio de Janeiro: [s.n.], 2000. Citado na página 17.

SECURATO, J. R. **Crédito Análise e Avaliação do Risco – Pessoas Físicas e Jurídicas**. São Paulo: Ed. Saint Paul, 2002. Citado 4 vezes nas páginas 18, 19, 20 e 22.

SILVA, A. F. A. V. **Modelação do Risco de Crédito numa Carteira de Crédito ao Consumo**. Dissertação (Mestrado) — Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa, Lisboa, 2014. Citado na página 17.

VALE, C. A. L. do. **Modelação e estimação do risco de crédito: estudo de uma carteira**. Dissertação (Mestrado) — Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa, Lisboa, 2010. Citado na página 17.

VAZ, J. C. L. **Regiões de incerteza para a curva ROC em testes diagnósticos**. 166 p. Dissertação (Mestrado em Estatística) — Universidade Federal de São Carlos, São Carlos, 2009. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/4538>. Acesso em: 21 ago. 2023. Citado na página 38.

A Tabelas obtidas para uma dada amostra

Tabela A.1: Organização dos resultados na planilha eletrônica - parte 1

Tomador	***X1i***	***X2i***	***X3i***	***X4i***	***X5i***	***X6i***	***X7i***	***X8i***	***X9i***	***X10i***	***X11i***
TC1	1	0	0	1	1	0	0	0	1	1	0
TC2	1	1	1	1	1	0	1	1	1	1	1
TC3	1	0	1	1	0	0	0	1	0	0	1
TC4	1	0	0	0	0	0	0	1	1	1	1
TC5	1	1	1	1	1	0	1	1	1	0	1
TC6	1	1	1	1	1	1	0	0	1	1	0
TC7	1	1	0	1	0	0	1	0	1	1	0
TC8	1	0	1	1	0	0	0	2	1	1	1
TC9	1	0	0	1	0	0	0	1	1	1	0
TC10	1	1	1	1	1	0	1	1	0	1	1
TC11	1	0	1	1	1	0	0	0	1	0	1
TC12	1	0	1	0	0	0	1	2	1	1	1
TC13	1	1	0	1	0	0	0	1	1	1	0
TC14	1	1	0	1	0	0	1	2	1	1	1
TC15	1	0	1	1	0	0	0	2	1	1	1
TC16	1	1	1	1	0	1	0	2	0	0	0
TC17	1	0	0	0	0	0	0	0	0	0	0
TC18	1	1	0	1	0	1	0	1	1	1	0
TC19	1	1	1	1	1	1	1	1	0	0	1
TC20	1	0	0	1	0	0	0	0	0	0	0
TC21	1	1	1	1	1	0	0	3	1	1	1
TC22	1	0	1	1	1	0	0	2	1	1	1
TC23	1	1	0	1	1	0	1	2	0	0	1
TC24	1	0	1	1	0	0	0	0	1	1	1
TC25	1	0	1	1	0	0	1	1	0	1	1
TC26	1	1	0	1	0	0	0	1	1	0	0
TC27	1	1	1	1	1	1	1	2	1	1	0
TC28	1	0	0	1	1	0	0	3	1	1	0
TC29	1	1	1	1	1	0	1	1	1	1	1
TC30	2	1	1	1	0	0	0	2	1	1	0
TC31	2	1	1	1	1	1	0	1	1	1	0
TC32	2	0	1	1	0	1	1	2	1	1	0
TC33	3	1	1	1	1	1	1	2	0	0	0
TC34	2	1	1	1	0	0	1	1	0	0	1
TC35	1	1	0	1	1	1	1	2	1	1	1
TC36	2	0	1	1	1	0	0	3	1	1	1
TC37	1	0	0	1	0	0	0	0	0	0	0
TC38	2	1	1	1	0	1	0	0	1	1	0
TC39	3	0	1	1	1	0	0	1	1	1	0
TC40	3	1	1	1	0	0	0	1	1	1	0
TC41	2	1	0	1	1	0	1	0	1	1	0
TC42	1	1	1	1	1	1	0	0	0	0	0
TC43	2	1	1	0	0	1	0	1	0	0	0
TC44	2	1	1	1	0	1	0	0	0	0	1
TC45	1	1	1	1	1	0	1	1	0	0	0
TC46	1	1	1	1	1	1	1	1	1	1	1
TC47	3	1	1	1	1	1	1	2	1	1	0
TC48	2	1	1	1	1	0	0	2	1	1	1
TC49	2	1	1	0	0	0	1	0	1	1	0

Fonte: Elaborada pelo autor.

Tabela A.2: Organização dos resultados na planilha eletrônica - parte 2

Tomador	***Success***	***Failure***	***Total***	***p-Obs***	***p-Pred***	***Suc Pred***	***Fail Pred***	$\ln L$	***% Correct***
TC1	0	1	1	0	6,28%	0,063	0,937	-0,065	100
TC2	0	1	1	0	35,57%	0,356	0,644	-0,44	100
TC3	1	0	1	1	86,75%	0,867	0,133	-0,142	100
TC4	0	1	1	0	3,08%	0,031	0,969	-0,031	100
TC5	1	0	1	1	81,12%	0,811	0,189	-0,209	100
TC6	0	1	1	0	56,68%	0,567	0,433	-0,837	0
TC7	0	1	1	0	47,43%	0,474	0,526	-0,643	100
TC8	1	0	1	1	83,54%	0,835	0,165	-0,18	100
TC9	1	0	1	1	52,18%	0,522	0,478	-0,65	100
TC10	1	0	1	1	16,39%	0,164	0,836	-1,808	0
TC11	1	0	1	1	53,10%	0,531	0,469	-0,633	100
T12	1	0	1	1	85,00%	0,85	0,15	-0,163	100
TC13	1	0	1	1	21,35%	0,213	0,787	-1,544	0
T14	1	0	1	1	43,46%	0,435	0,565	-0,833	0
TC15	1	0	1	1	83,54%	0,835	0,165	-0,18	100
TC16	1	0	1	1	99,21%	0,992	0,008	-0,008	100
TC17	0	1	1	0	18,08%	0,181	0,819	-0,199	100
TC18	1	0	1	1	64,56%	0,646	0,354	-0,438	100
T19	1	0	1	1	91,10%	0,911	0,089	-0,093	100
TC20	1	0	1	1	58,46%	0,585	0,415	-0,537	100
TC21	0	1	1	0	26,25%	0,262	0,738	-0,304	100
TC22	1	0	1	1	40,04%	0,4	0,6	-0,915	0
T23	0	1	1	0	21,85%	0,219	0,781	-0,247	100
TC24	0	1	1	0	52,50%	0,525	0,475	-0,744	0
TC25	0	1	1	0	85,69%	0,857	0,143	-1,944	0
TC26	0	1	1	0	67,87%	0,679	0,321	-1,135	0
T27	1	0	1	1	97,72%	0,977	0,023	-0,023	100
TC28	0	1	1	0	39,74%	0,397	0,603	-0,506	100
T29	0	1	1	0	35,57%	0,356	0,644	-0,44	100
TC30	1	0	1	1	98,69%	0,987	0,013	-0,013	100
T31	1	0	1	1	96,87%	0,969	0,031	-0,032	100
TC32	1	0	1	1	99,99%	1	0	0	100
TC33	1	0	1	1	99,99%	1	0	0	100
TC34	1	0	1	1	99,23%	0,992	0,008	-0,008	100
TC35	0	1	1	0	40,43%	0,404	0,596	-0,518	100
TC36	1	0	1	1	94,05%	0,941	0,059	-0,061	100
TC37	0	1	1	0	58,46%	0,585	0,415	-0,879	0
T38	1	0	1	1	99,10%	0,991	0,009	-0,009	100
TC39	1	0	1	1	99,52%	0,995	0,005	-0,005	100
TC40	1	0	1	1	99,74%	0,997	0,003	-0,003	100
T41	1	0	1	1	56,76%	0,568	0,432	-0,566	100
TC42	1	0	1	1	78,33%	0,783	0,217	-0,244	100
T43	1	0	1	1	99,03%	0,99	0,01	-0,01	100
TC44	1	0	1	1	98,26%	0,983	0,017	-0,018	100
TC45	1	0	1	1	89,16%	0,892	0,108	-0,115	100
TC46	1	0	1	1	78,74%	0,787	0,213	-0,239	100
TC47	1	0	1	1	99,98%	1	0	0	100
TC48	0	1	1	0	64,74%	0,647	0,353	-1,043	0
TC49	1	0	1	1	94,81%	0,948	0,052	-0,053	100
Total	**33**	**16**	**49**	–	–	**33**	**16**	**-19,708**	**79,592**

Fonte: Elaborada pelo autor.