UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

MARCELO AUGUSTO LIMA SANTOS

COLLECTION SCORING COMO FERRAMENTA DE DEFINIÇÃO DE ESTRATÉGIA DE COBRANÇA DE CLIENTES EM SITUAÇÃO DE INADIMPLÊNCIA BANCÁRIA

Recife
2022

MARCELO AUGUSTO LIMA SANTOS

COLLECTION SCORING COMO FERRAMENTA DE DEFINIÇÃO DE ESTRATÉGIA DE COBRANÇA DE CLIENTES EM SITUAÇÃO DE INADIMPLÊNCIA BANCÁRIA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Produção.

Área de concentração: Gerência da Produção

Orientadora: Profª. Drª. Caroline Maria de Miranda Mota.

Coorientador: Prof. Dr. Raphael Harry Frederico Ribeiro Kramer.

Recife

2022

Catalogação na fonte
Bibliotecário Gabriel Luz CRB-4 / 2222

S237c Santos, Marcelo Augusto Lima.
Collection scoring como ferramenta de definição de estratégia de cobrança de clientes em situação de inadimplência bancária / Marcelo Augusto Lima Santos. 2022.
81 f.

Orientadora: Profa. Dra. Caroline Maria de Miranda Mota.
Coorientador: Prof. Dr. Raphael Harry Frederico Ribeiro Kramer.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-graduação Profissional em Engenharia de Produção, Recife, 2022.
Inclui referências e anexo.

1. Engenharia de produção. 2. Inadimplência bancária. 3. Regressão logística binária. 4. *Collection scoring*. I. Mota, Caroline Maria de Miranda (Orientadora). II. Kramer, Raphael Harry Frederico Ribeiro (Coorientador). III. Título.

UFPE

658.5 CDD (22. ed.) BCTG / 2022 - 335

MARCELO AUGUSTO LIMA SANTOS

COLLECTION SCORING COMO FERRAMENTA DE DEFINIÇÃO DE ESTRATÉGIA DE COBRANÇA DE CLIENTES EM SITUAÇÃO DE INADIMPLÊNCIA BANCÁRIA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Produção. Área de concentração: Gestão da Produção

Aprovada em: 10 / 02 / 2022.

_____

Prof. Dr. Raphael Harry Frederico Ribeiro Kramer
Universidade Federal de Pernambuico

_____

Prof. Dr. Cristiano Alexandre Virgínio Cavalcante
Universidade Federal de Pernambuco

_____

Prof. Dr. Yuri Laio Teixeira Veras Silva
Universidade Federal Campina Grande

AGRADECIMENTOS

Aos meus pais, Tadeu e Socorro, e as minhas tias, Ana e Lourdes, pela dedicação, muitas vezes abnegada, em benefício de minha formação. A minha esposa, Arli, por todo o apoio. E aos meus mestres e orientadores, Profª Carolina Mota e Prof. Raphael Kramer, pelos ensinamentos e compreensão.

Este trabalho não teria sido possível não fosse o privilégio de tê-los comigo nesta jornada.

RESUMO

Este trabalho apresenta um estudo sobre a inadimplência bancária. Para tanto, buscou-se inicialmente realizar uma contextualização histórica do crédito, discorrendo sobre sua importância para o desenvolvimento econômico dos países e apontando os desafios enfrentados pelo Brasil nesta área. Para melhor analisar o tema, também foram esmiuçados os principais atos regulatórios que normatizam o assunto, bem como, foram apresentados alguns dos modelos de análise de clientes, como o *Credit Score* e o *Behavior Score*, que buscam minimizar a ocorrência da inadimplência de crédito.

Como resultado da pesquisa, foi apresentado um modelo de seleção de clientes para ações de cobrança, *Collection Score*, construído por meio da técnica de regressão logística binária, utilizando dados de clientes em situação de inadimplência bancária habitacional de uma centralizadora de cobrança. Também foi detalhado no trabalho toda a fundamentação matemática e estatística do modelo e o nível aceitação dos resultados obtidos.

Além dos resultados considerados satisfatórios no seu propósito de identificar os clientes com maior propensão a pagar, o modelo foi construído utilizando o Excel, um software amplamente utilizado, dotado de muita praticidade de aplicação e adaptabilidade a diferentes disponibilidades de dados.

Palavras-chave: inadimplência bancária; regressão logística binária; *collection scoring.*

ABSTRACT

This research presents a study about bank default. Therefore, we initially sought to carry out a historical contextualization of credit, discussing its importance for the economic development of countries, and pointing out the challenges faced by Brazil in this area. In order to better analyze the topic, we also de detailed the main regulatory acts that regulate the matter, as well as analyze some of the customer analysis models, such as the Credit Score and Behavior Score, which seek to minimize the occurrence of credit default.

As a result of the research, a customer selection model for collection actions was presented, Collection Score, built through the technique of binary logistic regression, using data from customers in a housing bank default situation from a collection center. The entire mathematical and statistical foundation of the model and the level of acceptance of the results obtained were also detailed in the work.

In addition to the results that we considered satisfactory in its purpose of identifying customers with a greater propensity to pay, the model was built using Excel, a widely used software, endowed with very practical application and adaptability to different data availability.

Keywords: bank default; binary logistic regression; collection scoring.

SUMÁRIO

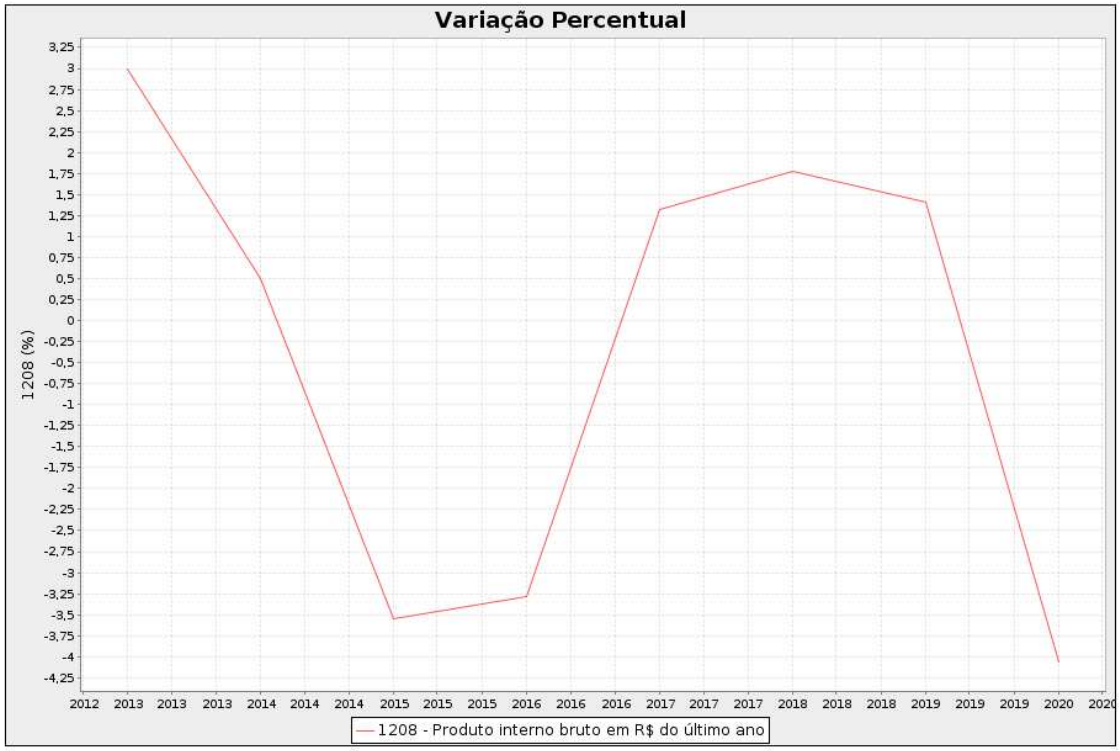
1 INTRODUÇÃO

A experiência internacional, confirmada por diversos estudos já aplicados na área econômica, sugere que há correlação positiva entre o grau de desenvolvimento dos mercados financeiros e as taxas de crescimento de uma economia. Os fundamentos desta relação estão estabelecidos, uma vez que, em mercados desenvolvidos, os intermediários financeiros alocam recursos de forma mais eficiente e, com isso, o crédito se torna mais disponível e ofertado em melhores condições. Esse cenário de crédito mais disponível, incentiva a disponibilidade de informações sobre empresas, aumento da profissionalização, padronização e *compliance* dos relatórios econômico-financeiros; fornecendo transparência e simetria de informação aos agentes econômicos, propiciando a redução dos custos de transação, a viabilização de projetos de longo prazo, e a melhoria da alocação de recursos em relação a diversificação de riscos.

Entre os estudos que buscaram confirmar teórica e empiricamente a existência dessa relação estão: Everton Silva e Sabino Júnior (2006), que aplicaram Regressão Quantílica utilizando dados de 77 países, mapeando as medidas de desenvolvimento financeiro e mensurando seu impacto nas métricas da variável resposta (desenvolvimento econômico), confirmando a hipótese. Já Antônio Carvalho (2002), revisou as evidências empíricas que comprovam o argumento teórico de que em uma economia de mercado há um descasamento natural entre geração de poupança e a capacidade empresarial. A intermediação entre as duas pontas é feita pelo sistema financeiro que se incumbe da alocação de investimentos. Assim, o desenvolvimento financeiro, fomenta a alocação de recursos nos projetos mais produtivos e, consequentemente, acelera o crescimento econômico. No estudo, o autor reforçou através das evidências empíricas internacionais a correlação positiva entre oferta de crédito e crescimento econômico.

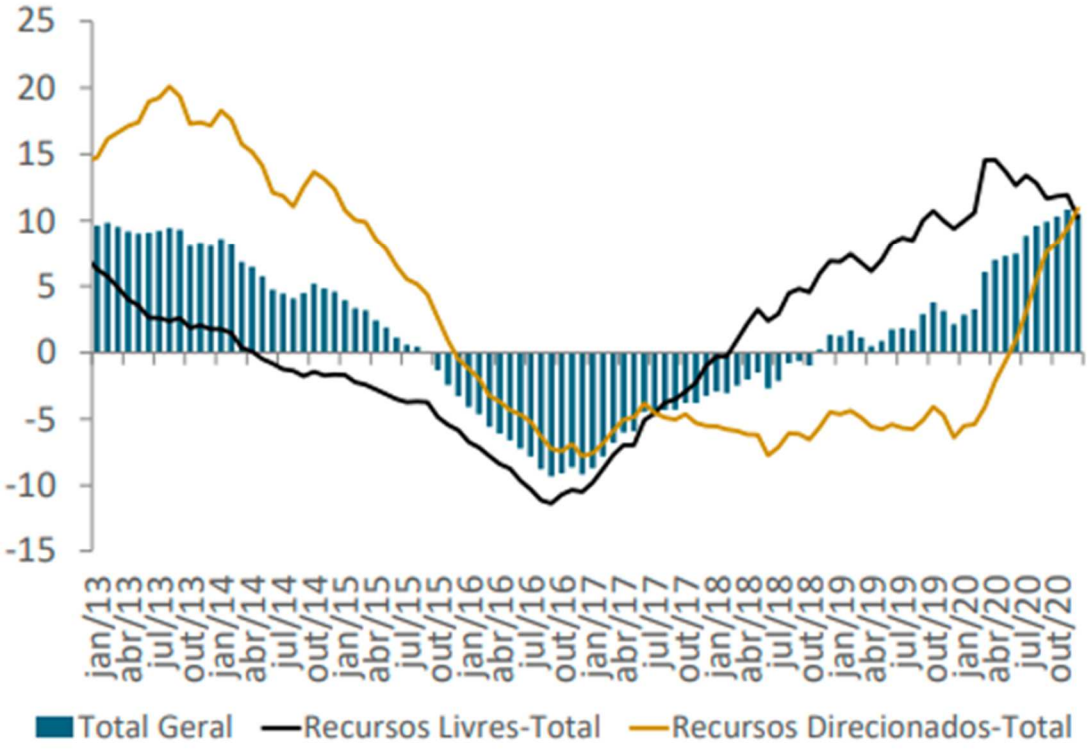
As figuras 1 e 2 a seguir ilustram o comportamento dessas duas variáveis (crescimento econômico e oferta de crédito) na economia brasileira nos últimos anos. Alertando para o período atípico iniciado em 2020 com a Pandemia do Coronavírus:

Figura 1 - Variação percentual do PIB Brasileiro (2012-2020)



Fonte: Banco Central do Brasil: SGS - Sistema Gerenciador de Séries Temporais (bcb.gov.br)

Figura 2 - Variação do saldo das operações de crédito do Sistema Financeiro Nacional (SFN) sobre o mesmo mês do ano anterior (em %)



Fonte: Carta Conjuntura IPEA Nº50 – 1º Trimestre de 2021

Na economia brasileira, o sistema financeiro, conforme passaremos a descrever, possui dois entraves centrais: oferta de crédito insuficiente e taxas de juros cobradas em patamares bem acima da média mundial. Conforme o relatório *Estatísticas do Mercado de Crédito Banco Central* (https://www.bcb.gov.br/estatisticas/estatisticasmonetariascredito) o saldo total de empréstimos no Sistema Financeiro Nacional (SFN) como porcentagem do Produto Interno Bruto (PIB) atingiu 54,2% em dezembro de 2020, o maior valor dos últimos dez anos (o pico anterior foi de 53,9%, em dezembro de 2015). Salientando que crédito também desempenha um papel anticíclico muito importante por estimular o consumo interno e os investimentos estruturantes na economia. A Tabela 1 traz a relação percentual entre oferta de crédito e PIB em diversos outros países, evidenciando o espaço que existe para avançar com a oferta de crédito no país. Destacando os casos do Chile e da Bolívia:

Tabela 1 - Relação percentual entre oferta de crédito e PIB

PAÍS	**(%)**
Estados Unidos	191,8
Japão	174,7
China	164,7
Reino Unido	133,6
Chile	122,5
Franca	107,6
Luxemburgo	107,3
Holanda	100
Finlândia	95,1
Espanha	94,7
Portugal	90,7
Áustria	85,8
Alemanha	80,2
Grécia	79,2
Bolívia	71,2

Fonte: https://data.worldbank.org/

Além da oferta de crédito em níveis inferiores a dos países do bloco europeu, dos Estados Unidos, da China, do Japão e até mesmo a dos vizinhos Chile e Bolívia, o crédito no Brasil é

caro. E embora a taxa básica de juros da economia (SELIC) tenha estado baixa níveis recordes para a economia brasileira desde o fim de 2017, as taxas de juros praticadas na economia brasileira permanece sendo uma das mais altas do mundo. Dos países com dados disponíveis no datacenter do Banco Mundial, apenas Mandagascar aparece a frente do Brasil:

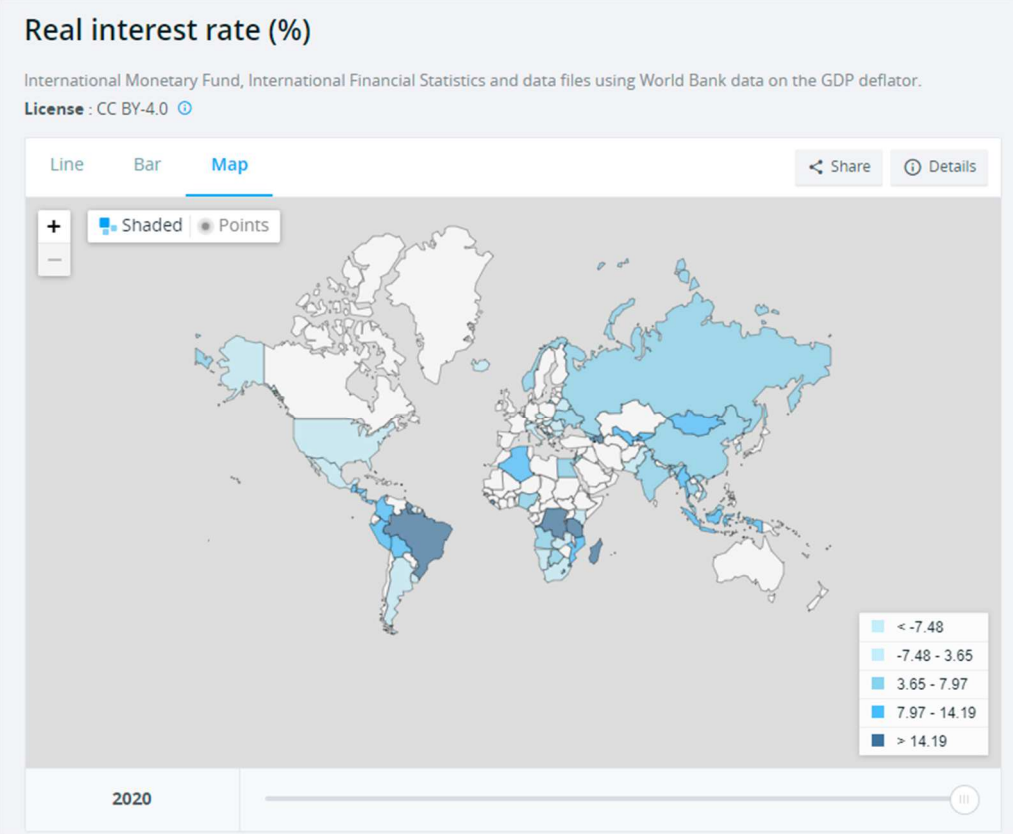
Tabela 2 - Taxa média de juros por país praticada na economia

País	**Dado mais recente**	**%a.a.**
Madagascar	2019	41,30
Brasil	2019	32,00
Iraque	2016	29,80
Líbia	2014	28,20
Malaui	2018	24,00
Zimbábue	2019	21,10
Rep. Dem. Congo	2019	21,10
Gambia	2019	19,50
Tajiquistão	2019	17,70
Azerbaijão	2019	17,60
Irã	2016	16,10
Bahrain	2015	15,90
Libéria	2017	15,80
Paraguai	2017	15,60
Uganda	2018	14,60
Moçambique	2019	13,90
Serra Leoa	2019	13,20

Fonte: https://data.worldbank.org/

O mapa apresentado na Figura 3, apresenta de forma ainda mais visual o quanto as taxas de juros aplicadas no Brasil estão distorcidas em relação às aplicadas no resto do mundo:

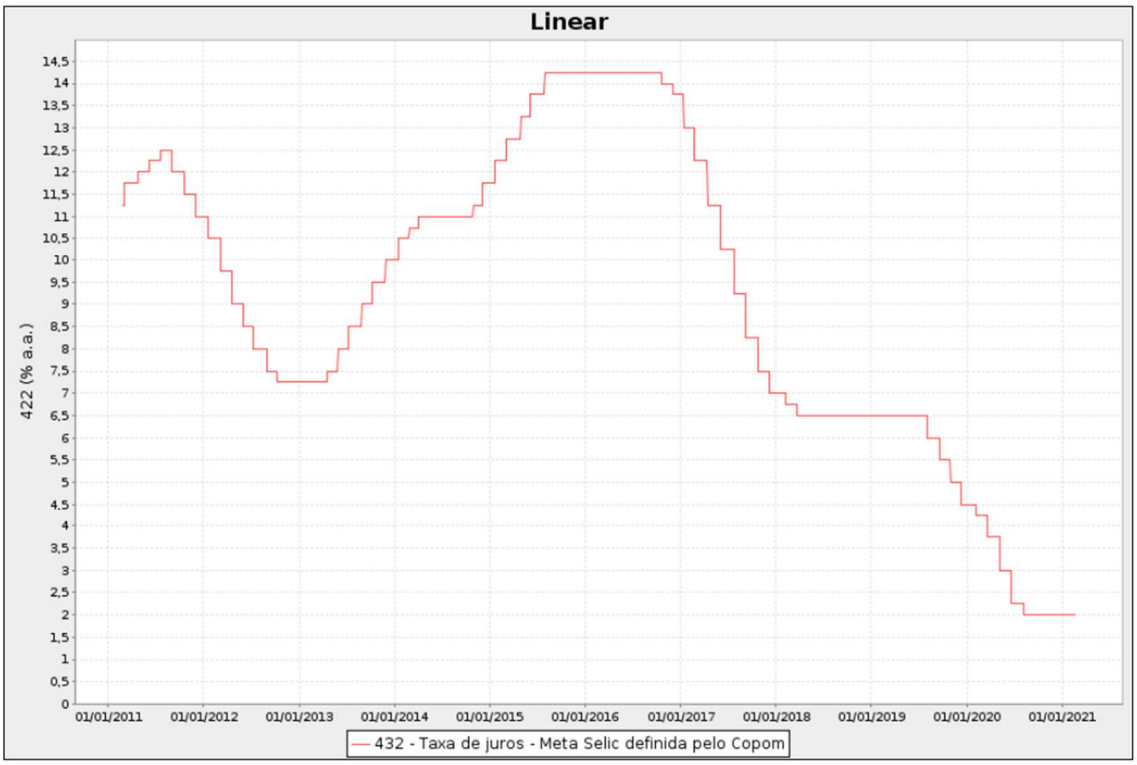
Figura 3 - Taxa de juros real aplicada na economia



Fonte: https://data.worldbank.org/indicator/FR.INR.RINR?view=map

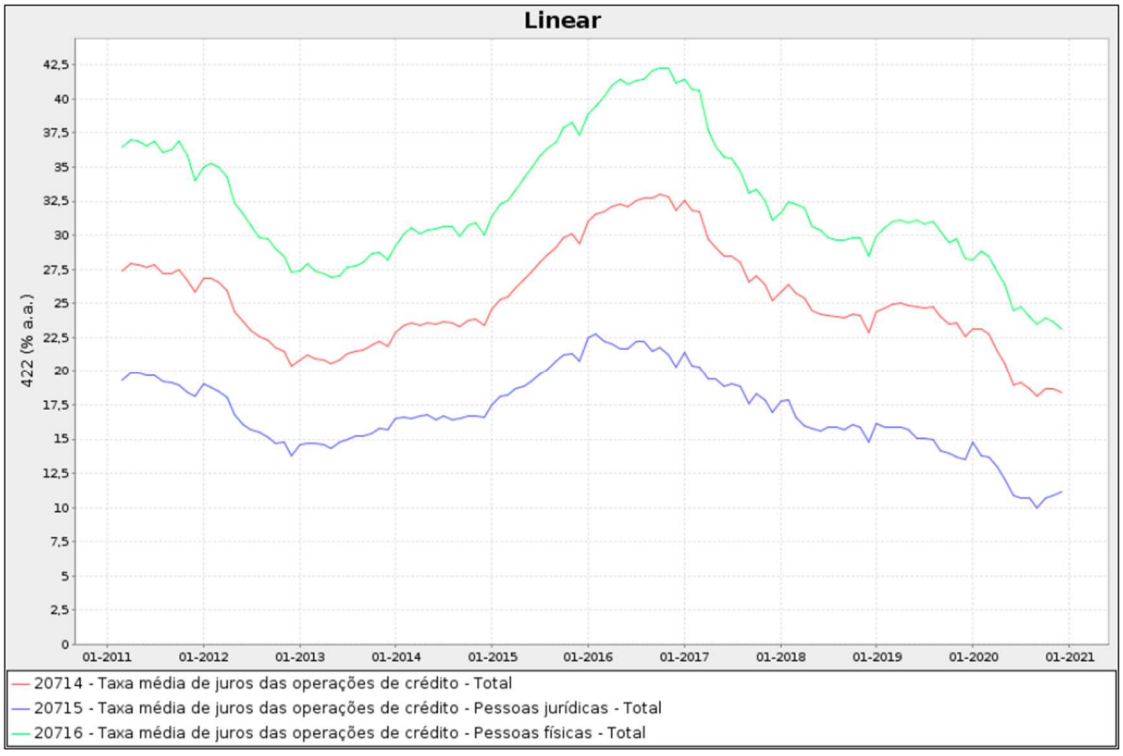
A justificativa mais comumente utilizada para explicar o preço elevado do crédito no Brasil é as altas taxas da Taxa Básica de Juros da Economia brasileira (SELIC). Esse argumento não se sustenta perante os números, especialmente quando se leva em consideração os últimos movimentos de redução desta taxa. Examinando a alegação com mais detalhes, percebemos que a redução da taxa SELIC não foi acompanhada na mesma proporção pelas taxas cobradas nas operações de crédito na economia brasileira no intervalo entre janeiro de 2017 e janeiro de 2021. Enquanto a primeira (a SELIC) reduziu de 14,25% a.a. para a mínima de 2,00% a.a., totalizando uma redução de 85,96%; nas operações de crédito, a taxa média cobrada no mercado financeira reduziu de 32,5% a.a. para a média de 18,5% a.a., ou seja, uma redução de aproximadamente 43,07%. Quase 50% inferior a queda observada na SELIC no mesmo período, conforme podemos visualizar nas Figuras 4 e 5.

Figura 4 - Taxa de Juros – Meta SELIC (2011 – 2021)



Fonte: Banco Central do Brasil: SGS - Sistema Gerenciador de Séries Temporais (bcb.gov.br)

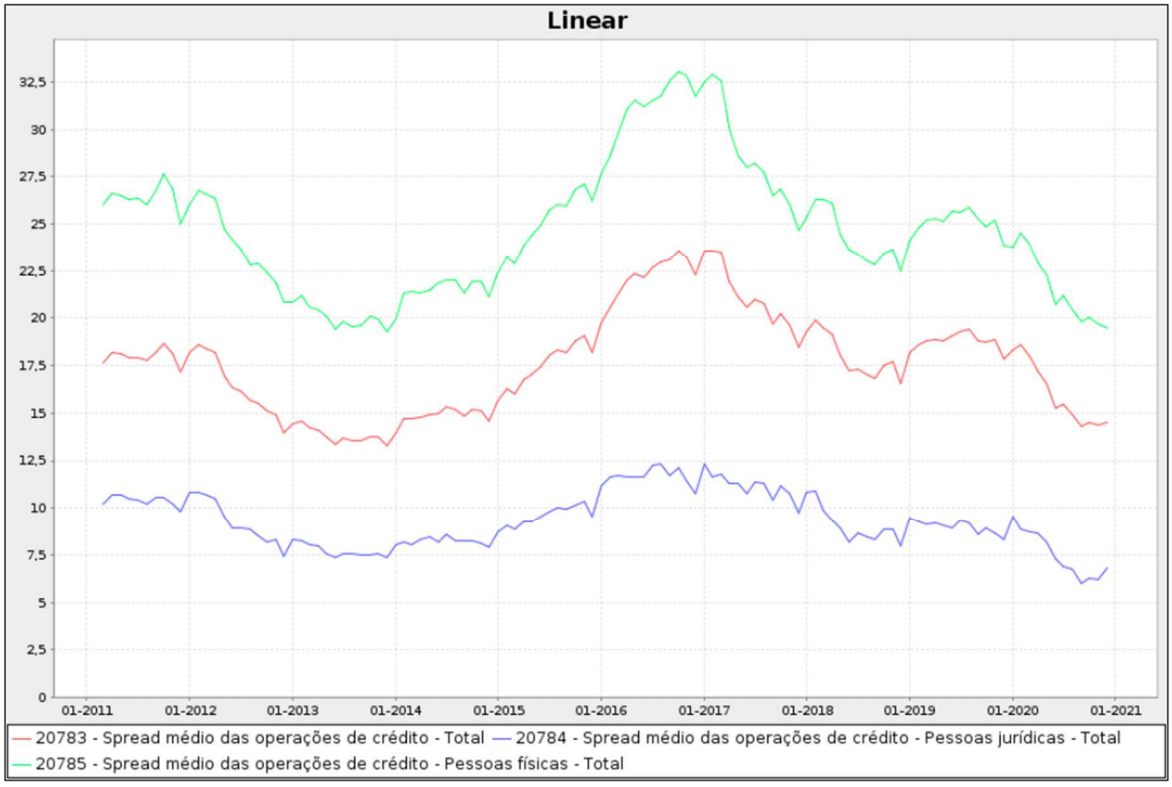
Figura 5 - Taxa de Juros das Operações de Crédito (2011 – 2020)



Fonte: Banco Central do Brasil: SGS - Sistema Gerenciador de Séries Temporais (bcb.gov.br)

Além do custo de captação, muito influenciado pela taxa SELIC, outro componente do custo do crédito é o *spread* bancário, o qual pode ser definido como a diferença entre o preço pago pelos bancos e outros agentes financeiros na captação, e o preço cobrado aos clientes na ponta concessora. Conforme gráfico abaixo, também é perceptível o quanto essa margem vem apresentando resistência à redução na última década.

Figura 6 - Spread bancário – Operações de crédito no Brasil (2011 – 2020)

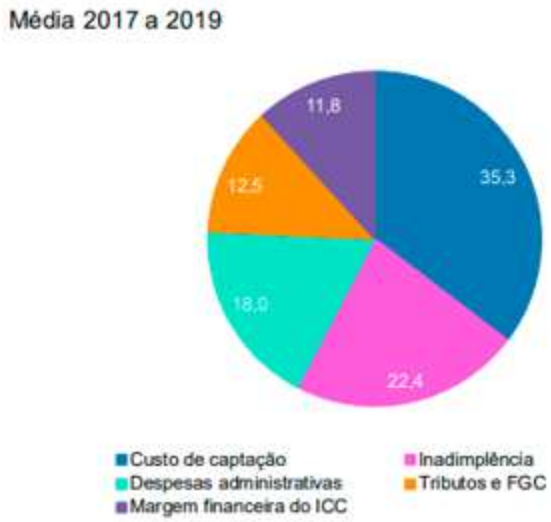


Fonte: Banco Central do Brasil: SGS - Sistema Gerenciador de Séries Temporais (bcb.gov.br)

Conforme demonstrado na Figura 6, o *spread* médio das operações de crédito no Brasil em 2020, apesar de todos os ganhos de produtividade e redução de custos operacionais propiciados pelo incremente tecnológico, bem como, pelo aumento da concorrência ocasionado pela entrada de novos *players* no mercado, especialmente as *Fintechs,* o *spread* praticado pelo setor financeiro apenas se aproximou dos percentuais históricos de 2014/2015.

Para entender o comportamento do custo do crédito no Brasil, é necessário observar com mais detalhes os outros fatores, além do custo de captação, que compõem o *spread* no Brasil, um dos mais elevados do mundo. Neste sentido, o Banco Central possui uma publicação chamada Relatório da Economia Bancária, no qual se detalha a composição do Índice de Custo de Crédito (ver Figura 7):

Figura 7 - Decomposição do Índice de Custo de Crédito (média 2017 – 2019)



Decomposição do ICC

p.p.

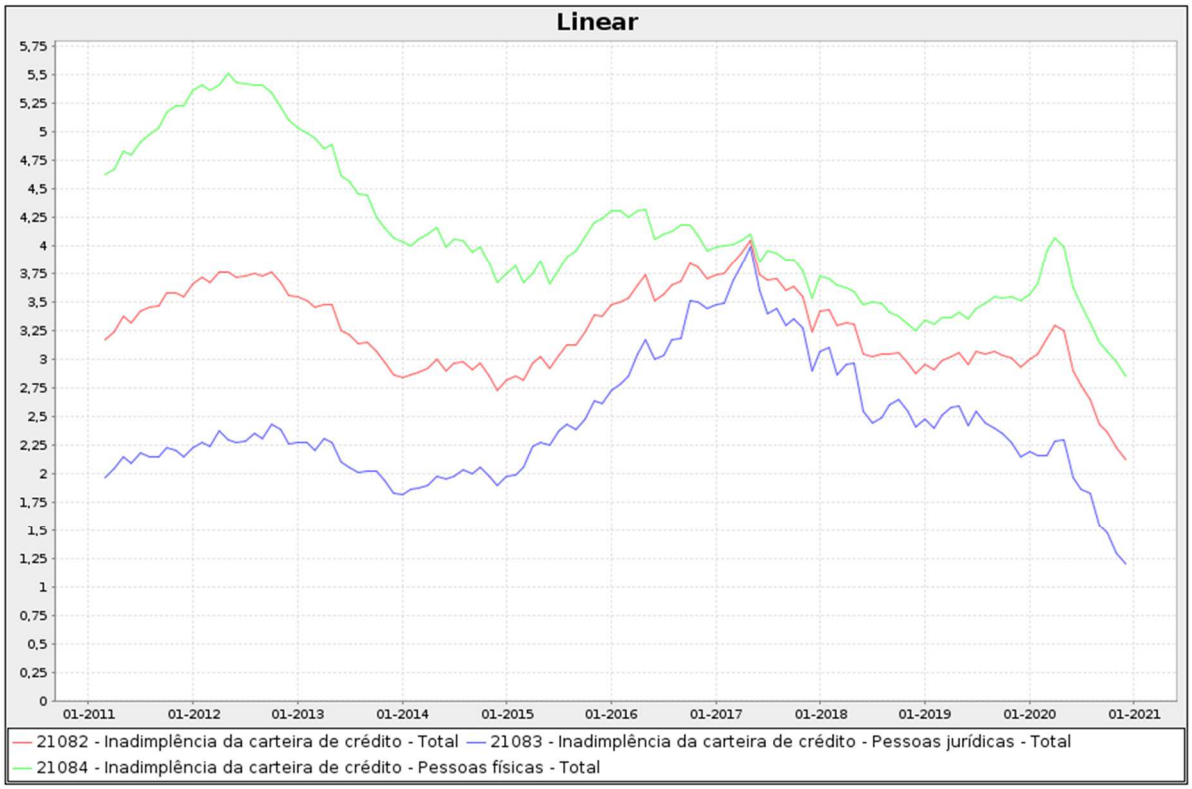
Discriminação	2017	2018	2019	Média
1 – Custo de captação	7,61	6,82	6,28	6,90
2 – Inadimplência	4,92	4,18	4,00	4,37
3 – Despesas administrativas	3,46	3,47	3,59	3,51
4 – Tributos e FGC	2,26	2,51	2,54	2,44
5 – Margem financeira do ICC	1,90	2,20	2,78	2,29
ICC médio ajustado (1 + 2 + 3 + 4 + 5)	20,15	19,18	19,19	19,51

Fonte: Relatório de Economia Bancária Bacen (REB) – Dez/2019

Para nivelar os conceitos, faz-se necessário informar que o Banco Central considera como sendo *Spread* tudo que não é o Custo de Captação, ou seja, os itens: 2, 3, 4 e 5 da tabela mostrada na Figura 7, o maior componente do custo de crédito é o custo de captação, seguido pela Inadimplência que responde por 22,4% do total.

Visto que a inadimplência é um fator relevante na composição do *Spread,* cumpre observar também o comportamento da inadimplência no Brasil na última década. Na Figura 8 é possível perceber o mesmo padrão apresentado nos gráficos da taxa de juros e do *spread* bancário (Figuras 5 e 6).

Figura 8 - Inadimplência das Carteiras de Crédito no Brasil (2011 – 2021)



Fonte: Banco Central do Brasil: SGS - Sistema Gerenciador de Séries Temporais (bcb.gov.br)

A taxa de inadimplência total da carteira de crédito no Brasil está em 2,12%. (referência dez/2020). Tal valor corresponde ao menor nível da série histórica do Banco Central, o qual dispõe de dados a partir de março de 2011. Por ser a inadimplência um fator tão relevante no custo do crédito no Brasil, e pelo crédito ser tão importante para o desenvolvimento econômico de uma nação, o objetivo deste trabalho será o de propor um modelo de seleção de clientes para ações de recuperação de crédito utilizando a análise por Regressão Logística Binária, com o propósito de identificar aqueles com maior propensão ao adimplemento.

1.1 JUSTIFICATIVA E RELEVÂNCIA

Trazendo a análise para o âmbito da Caixa Econômica Federal (CEF), um dos maiores bancos do país em número de clientes, que além das inúmeras atribuições de instituição financeira pública e banco de fomento social, possui também a função de atuar no mercado financeiro brasileiro como banco múltiplo, tendo participação relevante no mercado de crédito brasileiro. A título de exemplo, conforme dados divulgados em março / 2020, o banco detém 69,1% de participação no mercado de crédito imobiliário do país.

A principal fonte de lucro operacional dos bancos é a diferença entre o valor cobrado nas concessões e o pago nas captações, o que contabilmente é denominado de receita de intermediação financeira, a título de exemplo, a CEF em 2020 alcançou o resultado bruto de R$28,6 bilhões. Sendo este o resultado já descontado da Provisão para Créditos de Liquidação Duvidosa (PCLD) de 2020, que totalizou R$11,1 bilhões.

A PCLD é uma provisão contábil imposta às instituições financeiras e normatizada pelo Banco Central através Resolução 2.682 de 1999, que dispõe sobre critérios de classificação das operações de crédito e regras para constituição de provisão para créditos de liquidação duvidosa. Dentre as regras, determinam-se percentuais de provisão crescentes a depender do nível de risco de cada operação, que vai do nível "A" ao nível "H", sendo estes níveis estabelecidos com base em critérios consistentes e verificáveis que analisam a situação econômico-financeira, grau de endividamento, situação patrimonial, setor de atividade, entre outros aspectos socioeconômicos do proponente tomador contemplando, pelo menos, os seguintes aspectos:

I – Em relação ao devedor e seus garantidores:

a) Situação econômico-financeira;
b) Grau de endividamento;
c) Capacidade de geração de resultados;
d) Fluxo de caixa;
e) Administração e qualidade de controles;
f) Pontualidade e atrasos nos pagamentos;
g) Contingências;
h) Setor de atividade econômica;
i) Limite de crédito;

II – Em relação à operação:

a) Natureza e finalidade da transação;
b) Características das garantias, particularmente quanto à suficiência e liquidez;
c) Valor.

A classificação das operações de crédito de titularidade de pessoas físicas deve levar em conta, também, as situações de renda e de patrimônio bem como outras informações cadastrais do devedor.

A norma (Res 2.682 de 1999) determina também que, ao menos uma vez por mês, por ocasião dos balancetes e balanços, os níveis de risco das operações sejam revistos em função do atraso verificado no pagamento das parcelas das obrigações contratuais, iniciando em atrasos de 15 dias conforme Tabela 3:

Tabela 3 - Risco mínimo do crédito por faixa de atraso

Prazo até vencimento		Risco da Operação
Dias de Atraso (de – até)		
15	30	B
31	60	C
61	90	D
91	120	E
121	150	F
151	180	G
Superior a 181		H

Fonte: Brasil, 1999 citado por Fernandes, 2018

Tabela 4 - Percentual de provisionamento por faixa de Risco

Risco da Operação	Percentual de Provisionamento
AA	0,0%
A	0,5%
B	1,0%
C	3,0%
D	10,0%
E	30,0%
F	50,0%
G	70,0%
H	100,0%

Fonte: Brasil, 1999 citado por Fernandes, 2018

A regulação prevê ainda que a operação classificada como de risco nível H deve ser transferida para conta de compensação, com o correspondente débito em provisão, ou seja, lançada em prejuízo, após decorridos seis meses da sua classificação nesse nível de risco, não sendo admitido o registro em período inferior.

Cabe aqui salientar que a implementação desse regramento pelo Banco Central que prevê provisões para créditos em atraso, foi elaborado e implementado na esteira de um processo internacional multilateral denominado Acordo de Capital de Basileia, oficialmente denominado *International Convergence of Capital Measurement and Capital Standards*. Acordo firmado em 1988, na cidade de Basileia (Suíça), por iniciativa do Comitê de Basileia e ratificado por mais de 100 países. Este acordo teve como objetivo criar exigências mínimas de capital, que devem ser respeitadas por bancos comerciais, como precaução contra riscos de mercado, operacional, de crédito, entre outros, visando prover maior solidez ao sistema financeiro mundial.

A definição de índice de inadimplência adotada pelas instituições financeiras no Brasil, também é dada pelo Banco Central do Brasil (BACEN) e consiste no seguinte conceito: percentual da carteira de crédito do Sistema Financeiro Nacional com pelo menos uma parcela com atraso superior a 90 dias. Inclui operações contratadas no segmento de crédito livre e no segmento de crédito direcionado. (Fonte: Banco Central do Brasil – Departamento de Estatísticas).

Existem muitas outras definições para o termo inadimplência que podemos encontrar na literatura, como: “falta de cumprimento de uma obrigação. ” Houaiss, 2001, citado por Annibal, 2009, ou como Westgaard e Wijst, 2001, citado por Annial 2009, afirmam: “...entrar em default é fracassar em pagar uma quantia devida a um banco.”; e ainda Bessis, 1998, citado por Annibal, 2009, apresenta as seguintes definições: “...deixar de pagar uma obrigação, quebrar um acordo, entrar em um procedimento legal ou default econômico”.

Embora não seja consensual, o que se busca ao definir os parâmetros do índice é, por um lado, estabelecer níveis aceitáveis de risco de crédito para as instituições financeiras, e de outro, não inviabilizar a liquidez e oferta de crédito necessária ao dinamismo da economia. Uma vez que, conforme já indicado neste texto, diversos estudos já demonstraram a correlação positiva entre desempenho econômico e acesso ao crédito.

Demonstrada a relevância da adimplência para a solvência das instituições financeiras, e em consequência, para o nível de disponibilidade de crédito e taxas de juros cobradas em um país, fica evidenciada a necessidade de estabelecer formas de reduzir os índices de inadimplência nas instituições financeiras. As linhas de ação mais citadas nesta direção consistem em:

I. Qualificar a sistemática de avaliação dos clientes e de concessão do crédito;
II. Melhorar a efetividade das ações de Cobrança e Recuperação de crédito;

Embora exista farta literatura sobre modelos de *credit score* e *behavior score* como modelos preditivos que visam maior assertividade na concessão do crédito, este trabalho focará na dimensão das ações de cobrança, visando melhor aproveitar a experiência profissional do mestrando e potencializar o nível de detalhe desejado para o estudo.

A relevância do trabalho reside na expressividade das carteiras sob gestão da unidade, no tamanho da força de trabalho disponível, quase 100 empregados, e na relevância da adimplência para a sustentabilidade da instituição. Há ainda, a convicção entres os gestores da lacuna existente para otimização do processo de seleção de clientes direcionados ao acionamento proativo, atualmente, por exemplo, a seleção ocorre observando apenas os maiores valores de endividamento dos clientes e suas faixas de dias de atraso (entre 60 e 90 dias, entre 90 e 120 dias, etc.).

A carteira de crédito nas instituições financeiras divide-se em: segmento comercial e segmento habitacional, ambos atendem clientes pessoa física e jurídica. A título de ilustração o segmento comercial é composto, entre outras, por aquelas operações de crédito (cartão de crédito, crédito rotativo, empréstimo consignado, etc.) destinados ao público pessoa física, bem como, produtos comerciais como capital de giro, financiamento de máquinas e equipamentos, etc. que suprem as necessidades das empresas. No segmento habitacional o crédito tem como destinação específica a aquisição de moradia. Seja na forma de aquisição de imóveis prontos ou mesmo para aquisição de terreno e construção. O crédito imobiliários utiliza duas fontes de recursos (*funding*) que são o Sistema Brasileiro de Poupança e Empréstimo (SBPE), ou o Fundo de Garantia por Tempo de Serviço (FGTS), e sempre possuem como garantia a alienação fiduciária do bem adquirido pelo proponente tomador.

Na centralizadora de cobrança utilizada neste estudo, a carteira de crédito inadimplente, ou seja, clientes com mais de 91 dias de atrasos no pagamento das suas prestações contratuais, era de (referência Janeiro de 2020): R$ 1,9 bilhão de reais no segmento habitacional, R$ 664 milhões no segmento Pessoa Física e de R$ 265 milhões no segmento Pessoa Jurídica.

Tendo em vista que a carteira habitacional é a mais relevante, que a concessão atende critérios bem homogêneos: existência de garantia real nos contratos, baixa variação nas taxas de juros cobrada nos contratos (entre 5 e 12% a.a.), e um processo de cobrança com alternativas negociais que contempla quase a totalidade dos contratos inadimplentes, este estudo focará nos clientes dessa carteira.

Além das razões apontadas no parágrafo anterior, a atual ausência de uma modelagem mais cientifica para a atuação da unidade, há um problema prático que precisa ser solucionado, pois, a quantidade de clientes do segmento habitacional da unidade é de, em média, 38 mil contratos mês, sendo a capacidade de acionamento pelo canal telefone da centralizadora limitado a faixa de 6.000 a 7.000 contratos por mês, a depender da quantidade de dias úteis e da quantidade de funcionários disponíveis no período.

Logo, o estudo, além da contribuição na perspectiva financeira, de sustentabilidade e competitividade da instituição, principalmente sob o ponto de vista da redução do risco de crédito e das provisões contábeis já citadas, subsidiará a tomada de decisão da gestão da unidade, permitindo uma melhor alocação dos recursos humanos disponíveis.

1.2 OBJETIVO

O objetivo deste trabalho será o de analisar os dados dos clientes em situação de atraso em seus créditos habitacionais de uma grande instituição financeira nacional, apresentar um modelo de priorização das ações de cobrança, que identifique aqueles clientes com maior propensão à reversão e discorrer sobre os resultados alcançados.

O planejamento do trabalho perpassa os seguintes marcos:

- Objetivo Geral: Criar um processo de priorização de clientes para realização de ações de cobrança.
- Objetivos Específicos:
  - Definir um processo de priorização de clientes;
  - Identificar as variáveis *input* do modelo e coletar os dados necessários;
  - Validar o processo através de testes estatísticos;
  - Apresentar um modelo de priorização de clientes para ações de cobrança;
  - Descrever os resultados alcançados após a implementação do novo método;
  - Descrever os próximos desafios e limitadores encontrados.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado em cinco capítulos e mais dois apêndices. O capítulo 1, como vimos, dedicou-se a enumeração dos fatores que fazem com que a inadimplência seja um fator determinante para o custo do crédito no Brasil, e, por essa razão, acaba influenciando no

nível de contribuição que o Sistema Financeiro Nacional é capaz de fornecer ao desenvolvimento econômico do país. Naquele capítulo, apresentamos dados do Banco Mundial com indicadores, tais como: oferta de crédito em relação ao PIB e taxa média de juros praticada na economia brasileira e em outros países, com o intuito de situar o leitor de como está a situação nacional. Trouxemos ainda o como a inadimplência é normatizada pelo Banco Central e os desdobramentos das regulações nos resultados financeiros dos bancos, por fim, concluindo o capítulo com a apresentação dos objetivos desta pesquisa.

No capítulo 2 foi realizada uma revisão da literatura desde o contexto histórico, citando as origens científicas e legais dos primeiros processos estatísticos que buscaram qualificar a concessão do crédito e, assim, reduzir a chance de inadimplência, bem com, apresentamos a base conceitual do processo estatístico objeto desta pesquisa: o *Collection Score*, detalhando o processo da Regressão Logística e seus testes de significância estatística e de eficiência do modelo.

O capítulo 3 demonstra a aplicação prática do modelo à carteira de clientes com inadimplência habitacional de um banco de varejo país, realizada com intuído de subsidiar a decisão dos gestores sobre a priorização dos clientes para realização de ações de cobrança. Demonstrando desde a preparação da base de dados à realização dos cálculos através da ferramenta Excel. O capítulo demonstra ainda como identificar na ferramenta os testes de significância estatística e de poder preditivo do modelo e como interpretá-los. Finalizamos o capítulo explicando os ganhos obtidos com a implementação como parte do processo de seleção de clientes.

O capítulo 4 traz as considerações finais, elencando os principais ganhos, os limitadores e as possibilidades para futuras pesquisas na área, o capítulo 5 traz as referências bibliográficas utilizadas na pesquisa, e os Apêndices A e B finalizam o trabalho trazendo, respectivamente, o histograma da base de dados média tratada no capítulo 3, e o passo a passo de navegação do suplemento do Excel utilizado para realização dos cálculos da regressão logística.

2 BASE CONCEITUAL E REVISÃO DA LITERATURA

Nesta Seção apresentaremos um breve resumo sobre a história do crédito, demonstrando como a evolução científica influenciou, e permanece influenciando, na criação de métodos de decisão sobre concessão de crédito a clientes individuais (sejam pessoas ou empresas), processos que passaram de julgamentos humanos feitos a partir da experiência do tomador, a modelos estatísticos sofisticados com utilização intensiva de tecnologias computacionais, que tornaram o processo decisório mais objetivo, ágil e minimizaram as perdas das carteiras de crédito.

Surgiram assim vários modelos baseados em técnicas de escore (pontuação). Como o *Credit Score* baseado em dados fornecidos pelo cliente, o *Behavioural Score* que é fundamentado pelo histórico de pagamento junto à empresa concessora, e o *Collection Score,* processo que busca identificar e pontuar os atributos dos clientes de maior propensão ao pagamento. São processos padronizados e impessoais, com ampla fundamentação estatística, utilizado em larga escala por instituições financeiras. Dedicaremos maior atenção à fundamentação teórica do *Collection Score*, que será o objeto desta dissertação.

2.1 CONTEXTO HISTÓRICO

A etimologia da palavra Crédito, do latim *creditu*, como a confiança que se tem em algo, explica a esperança de que se vai receber de volta no futuro, o dinheiro emprestado agora. Aquele que empresta dinheiro a um indivíduo ou a uma instituição, se chama credor, pois ele "crê" que receberá seu dinheiro de volta.

Os bancos, que no Brasil concentram maior parte da concessão de crédito às pessoas físicas e jurídicas, tem sua concepção como instituição associada à Mesopotâmia do período 4000-2000 a.C. onde também nasceu o conceito de emissão de carta de crédito e do empréstimo a juros, antes de Hamurabi e seu Código homônimo (1790-1750 a.C.) quando os templos também serviam como bancos comunitários (CHAHIN, 2001, citado em Rodrigues *et al.* 2014).

Todavia as operações de crédito, em seu verdadeiro caráter, somente foram encontradas na Grécia e em Roma (Chaia, 2003, citado em Lopes 2004). Cerca de 800 a.C. existiam na China notas de banco, bem como, outros meios de troca, como: moeda metálica, títulos de crédito, propiciando o aparecimento de intermediadores que desempenhavam as funções do que hoje definimos como banco (Gonçalves, 2016).

Trata-se de um conceito intrinsicamente associado ao cotidiano de qualquer sociedade moderna, desde a realização de compras em um estabelecimento comercial através de um simples cartão de crédito, às formas de financiamento de projetos produtivos de longo prazo como portos, aeroportos, estradas e ferrovias, passando por plantas industriais e de produção agrícola de pequena e grande escala, até o financiamento de bens de consumo duráveis como: imóveis, veículos, etc.

Numa instituição financeira típica, diariamente são tomadas diversas decisões que dizem respeito à concessão de crédito: aprovar ou rejeitar uma proposta de empréstimo, qual limite aprovar, quando exigir garantia, qual taxa de juros cobrar, qual prazo conceder etc., essa análise originalmente era puramente julgamental.

David Durand (1941) foi o primeiro a reconhecer que a técnica de análise discriminante, inventadas por Fisher em 1936, poderia ser usada para separar bons e maus empréstimos. Em "*Risk Elements in Consumer Installment Financing, 1941 (National Bureau of Economic Research, N.Y.*)", Durand apresentou um modelo que atribuía pesos para cada uma das variáveis usando análise discriminante (KANG e SHIN, 2000, p.2198, citado em Neto *et al.* 2004).

A partir da década de 60 do século passado, os Bancos passaram a adotar a modelos de *Credit Score,* percebendo que sua capacidade preditiva era maior do que a de qualquer sistema julgamental, bem como, permitia que as instituições fizessem frente ao volume crescente da demanda por crédito. Nos Estados Unidos, a aprovação do *Equal Credit Opportunity Acts* (ECOA 1975/1976) forneceu ao modelo a proteção legal necessária a sua implementação e rápida disseminação, uma vez que tornou legítima a recusa de crédito fundamentada por critérios estatísticos.

2.1.1 O Credit Score

O *Credit Score* é um instrumento matemático-estatístico desenvolvido com o intuito de avaliar a probabilidade de que um tomador de crédito venha a ser um bom ou mau pagador no futuro. Essa avaliação considera o conjunto de características do cliente e a experiência de cada instituição financeira evidenciadas como relevantes na previsão do pagamento, atribuindo valores numéricos a cada uma dessas características ou variáveis de risco presentes no perfil do tomador e em cada operação de crédito. Dessa forma, os empréstimos são concedidos ou recusados de forma padronizada a partir de critérios objetivos buscando maximizar a probabilidade de reembolso. Sendo assim, modelos probabilísticos e não determinísticos alimentados por dados históricos.

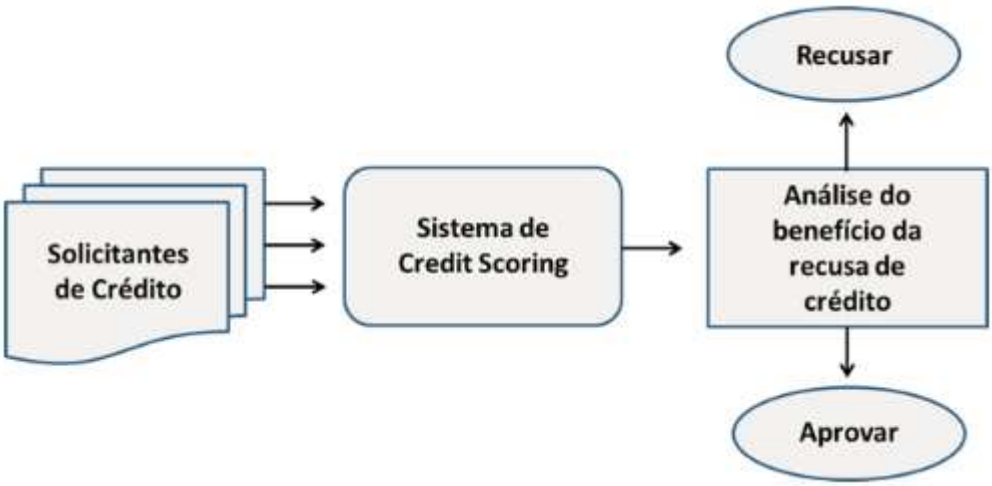
As cinco dimensões conhecidas como os 5 C`s do crédito, que organizam as informações sobre a capacidade de pagamento de cada cliente que alimentam os modelos são:

1. Caráter: Representado pelo histórico de pagamento;
2. Capacidade: Potencial financeiro para honrar os compromissos;
3. Capital: Disponibilidade/patrimônio do solicitante;
4. Colaterais: Garantias adicionais oferecidas;
5. Condição: Diz respeito às condições econômicas vigentes e as características individuais.

Os modelos tradicionais de crédito estabelecem pesos e pontuações para cada um dos atributos dos solicitantes. Dessa forma, um servidor público com vínculo estatutário, alto nível de escolaridade e capacidade financeira elevada, possivelmente obterá uma pontuação mais elevada do que um comerciante com mau histórico de pagamento, que esteja iniciando seu primeiro empreendimento, enquanto ainda cursa a universidade. Ou seja, utiliza-se uma análise do histórico do cliente, rendimentos, categoria profissional, nível de despesas, com o objetivo de obter um escore que informe a política de crédito da instituição.

Importante ressaltar que mesmo sendo o *Credit Score* um processo matemático, não elimina a possibilidade de aceitar um mau pagador, ou de recusar um bom pagador, porque nenhum sistema de avaliação é capaz de capturar todas os dados relevantes necessários à precisa classificação dos tomadores. Importante também informar que mudanças nas circunstâncias econômicas, interfere no comportamento de pagamento de cada cliente durante a vigência do contrato de crédito. Chagas *et al.* (2018), traz o seguinte processo de concessão de crédito com modelos de *credit scoring*:

Figura 9 - Processo de concessão de crédito com modelos de *credit scoring*



Fonte: Souza e Chaia, 2000

Bogges (1980), apresenta uma tabela com o percentual de clientes classificados nas categorias bons ou maus pagadores, citando alguns dos principais fatores utilizados para a avaliação do *Credit Score*:

Tabela 5 - Relação de exemplificativa de fatores *input* do modelo

Características	Bons Pagadores	Maus Pagadores
É casado(a)?	90,50%	86,20%
Tem casa própria?	80,40%	42,30%
Tem carro próprio?	80,70%	68,00%
Tem mais de 35 anos?	97,00%	89,50%
Mora mais de 3 anos no endereço atual?	91,80%	70,30%
Tem referência bancária?	93,60%	71,00%
Tem telefone?	75,40%	70,30%
Tem menos de 3 filhos?	65,80%	49,30%

Fonte: Chaia, 2003

Observando a Tabela 5, é possível concluir que nenhum fator sozinho define os bons ou maus pagadores. Mesmo entre os que possuem casa ou carro próprio, a incidência de maus pagadores é alta. Ou seja, por si só, os fatores não são conclusivos, sua utilidade como ferramenta de apoio à decisão está na utilização do seu conjunto. É desta forma que o *Credit Score* estabelece regras de pontuação (*score*) através da combinação dos fatores.

A Tabela 6 (Boggess, 1980, citada em Chaia 2003) apresenta uma relação hipotética dos fatores e pontos associados a cada um dos fatores:

Tabela 6 - Relação hipotética de fatores e pontuação correspondente

Características	Pontos para resposta sim
É casado (a)?	10
Tem casa própria?	15
Tem carro próprio?	7
Tem mais de 35 anos?	9
Mora mais de 3 anos no endereço atual?	14
Tem referência bancária?	18
Tem telefone?	8
Tem menos de 3 filhos?	19

Fonte: Chaia, 2003

A partir da agregação dos *scores*, é possível calcular os ganhos e perdas em função da rejeição de clientes que não alcançarem determinado índice. O ponto de equilíbrio será aquele a partir do qual a instituição financeira não será capaz de aumentar seus lucros através da aceitação de clientes potenciais. A Tabela 7 (Bogges, 1980, citada em Chaia 2003), ilustra o ganho e perda em função da rejeição de potenciais clientes. No exemplo, considerou-se que cada bom pagador recusado gera uma perda de $100,00, e cada mau pagador recusado um ganho de $50.

Tabela 7 - Resultado do modelo

Credit Score	Bons Pagadores	Maus Pagadores	Ganho na Rejeição
5	5	300	14.500
10	25	600	27.500
15	75	900	37.500
20	125	1.100	42.500
25	200	1.200	40.000
30	400	1.300	25.000
35	800	1.400	(10.000)

Fonte: Chaia, 2003

Conforme a pontuação (*Credit Score*) se eleva, percebe-se um aumento na proporção de bons pagadores em relação aos maus pagadores, até alcançar o ponto em que a recusa dos potenciais clientes se torna economicamente danosa à instituição.

Os fatores considerados nos modelos de avaliação se alteram em função da área de análise (pequenas, médias, grandes empresas), do produto comercializado, de mudanças dos hábitos sociais, com o decorrer do tempo, etc. Como exemplo, segundo Narayanan, *et al.* 1998, um dos modelos *Credit Score* mais utilizados na análise de crédito de clientes corporativos é o método multivariado de pontuação Z de Altman (Altman, 1968), que consiste em combinar índices financeiros dos balanços contábeis que possam diferenciar empresas falidas e saudáveis.

A fórmula a seguir apresenta os pesos indicados para cada fator no estudo do Altman (1968), e a Tabela 8 apresenta os resultados nos diferentes grupos de empresa (saudáveis e com problemas).

$$Z = 0{,}012(X_1) + 0{,}014(X_2) + 0{,}033(X_3) + 0{,}006(X_4) + 0{,}999(X_5)$$

Onde:

- X_1: representa o coeficiente entre capital de giro e ativos totais;
- X_2: representa o coeficiente entre lucros retidos e ativos totais;
- X_3: representa o coeficiente entre lucro antes do importe e juros ativos totais;
- X_4: representa o coeficiente entre valor de marcado do patrimônio líquido e valor escritural do passivo;
- X_5: representa o coeficiente entre vendas e ativos totais.

Tabela 8 - Resultados das variáveis para diferentes grupos de empresa

Variavel	Média do Grupo Quebrado	Média do Grupo Não Quebrado
X_1	-6,1%	41,4%
X_2	-62,6%	35,5%
X_3	-31,8%	15,4%
X_4	40,1%	247,7%
X_5	1,5 vezes	1,9 Vezes

Fonte: Altman, 1968, citado por Chaia, 2003

De forma semelhante ao exemplo anterior, com o método de Altman (1968) é possível determinar um valor crítico de Z abaixo do qual os empréstimos comerciais seriam classificados como ruins e, por isso, recusados.

Posteriormente a esse trabalho de 1968 com cinco variáveis, Altman publicou em 1977 um dos principais trabalhos acadêmicos sobre risco de crédito, neste, ele enfatizou a necessidade do desenvolvimento de um modelo preditivo em função do crescimento das falências e das mudanças no contexto financeiro, e apresentou o clássico modelo de análise discriminante de sete variáveis, fruto da continuação de trabalhos anteriormente apresentados.

Importante citar que os estudos acima apresentados, embora antigos, foram utilizados no presente trabalho pela praticidade com que apresentam o racional que fundamenta a técnica do *Credit Score*. Existe uma infinidade de estudos recentes sobre o tema, como o de Albuquerque *et al.* 2016, que apresenta um modelo de *Credit Score* formulado através da técnica de Regressão Logística Geograficamente Ponderada. Há ainda outros trabalhos que incorporam técnicas de *Machine Learning* ao processo. No entanto, a apresentação destas modelagens mais atualizadas fogem do objetivo de apresentação das técnicas de prevenção de inadimplência desta seção do trabalho.

2.1.2 O Behavior Score

Como a própria denominação já sugere, o *Behavior Score* se fundamenta no comportamento histórico do cliente em seu relacionamento com a instituição de crédito. Esse modelo de avaliação é formado a partir das informações colhidas de cada cliente que possui relacionamento com a instituição. Assim, o credor consegue mensurar com precisão o comportamento de consumo, de poupança, comprometimento da renda, frequência e quantidade de inadimplência, entre outras informações, além da renda líquida média de cada cliente.

Neste modelo, além das informações demográficas, informações comportamentais também são levadas em consideração: histórico de pagamentos em dia, quantidade de empréstimos tomados, quantidade de empréstimos quitados, etc. fornecendo informações instantâneas ao analista e tendo maior poder preditivo do que o modelo *Credit Score*.

A saída (*output*) de um modelo *Behavior Score* pode ser interpretada como uma propensão do cliente a honrar com o compromisso financeiro que está pleiteando, e, assim como no *Credit Score*, também é uma pontuação utilizada pelos agentes financeiros para estabelecer pontos de corte para tomada e decisão massificada, ou seja, qual o limite mínimo de escore a partir do qual nenhum crédito será concedido, bem como, quais os escores necessários para

concessão de crédito a uma dada taxa de juros ou de um limite de crédito. Reduzindo o ponto de corte, o gestor está aceitando mais clientes e, com isso, aumentando a exposição da carteira ao risco, e vice-versa.

Kennedy *et al*. (2013) exemplifica dados utilizados como *input* em um típico sistema de análise *Behavior Score*, em sua maioria extraídos do próprio banco de dados da instituição financeira e birôs de crédito. Interessante também a observação de que cada tipo de dado ganhará mais peso a depender do objetivo para o qual o sistema estiver sendo utilizado, por exemplo: características como "Histórico de atendimento à promoções" é menos apropriada para analisar empréstimos de longo prazo, em comparação com a propensão do cliente ao encerramento precoce de contas. A Tabela 9 ilustra o argumento do autor:

Tabela 9 - Exemplos de comportamentos *input*

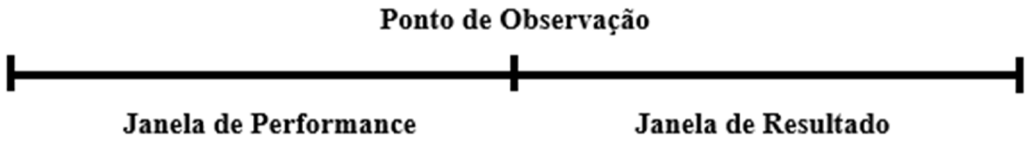
Tipo de Dado	**Exemplo**
Histórico de Atraso	Frequência de atrasos
	Nível máximo de atraso
Histórico de Uso	Relação equilíbrio-limite de Crédito
	Tendência de equilíbrio
Informação Estática	Pontuação de aplicação do cliente
	Idade do cliente
Histórico de Pagamento e Compra	Frequencia de compras
Atividades de Cobrança	Tipos de bens de varejo comprados
	Resultados
Contato de Serviço ao Cliente	Frequência de contatos
	Contato receptivo
Histórico de Promoções	Contato ativo
	Número de ofertas
	Resultado das ofertas
Data Center	*Scores* genéricos de informações compartilhadas

Fonte: Kennedy (2013)

Kennedy *et al*. (2013), também descreveu a dinâmica do processo, relatando que a primeira etapa corresponde à seleção de uma amostra de clientes, garantido que os dados referentes aos seus produtos e padrões de consumo estejam disponíveis em uma determinada

data chamada de ponto de observação. O período antes do ponto de observação é chamado de janela de Performance. Os dados contidos na janela de desempenho são estruturados em atributos que serão usados como entrada. O período posterior ao Ponto de Observação é conhecimento como Janela de Resultado.

Figura 10 - Etapas do *Behavior Score*



Fonte: Kennedy (2013)

A proposta desta última etapa é de acompanhar e distinguir os bons e os maus pagadores baseado em seus níveis de inadimplência. Kennedy *et al.* (2013) chama atenção para a necessidade do cuidadoso dimensionamento do tamanho deste período, informando que não há consenso sobre, sendo o período de 06 a 24 meses o mais comum. Como também é importante considerar na avaliação da qualidade do sistema de predição as condições econômicas, viés político e a mudanças na política de crédito que podem afetar, para melhor ou pior, o comportamento de pagamento dos clientes.

Uma vez que o relacionamento do cliente é dinâmico, novos dados são inseridos constantemente ao modelo, ou seja, mais do que informar se o cliente é bom ou mau pagador, os modelos comportamentais permitem prever o risco associado a cada cliente durante o decorrer do tempo do seu relacionamento com a instituição, permitindo uma atuação mais customizada por parte do credor, definindo taxas, níveis de garantias, limites de crédito e prazos de concessão.

Conhecer o risco do cliente permite também a segmentação do portfólio em conjuntos de desempenho parecidos, estabelecendo estratégias de atuação diferentes para cada segmento, seja do ponto de vista da gestão financeira do risco – minimizar perdas de crédito ou aumentar receitas – seja do ponto de vista de marketing – aumentar o portfólio para públicos selecionados ou reter clientes de alta rentabilidade (Souza, 2000).

2.1.3 Collection Score

Os dois modelos estatísticos discutidos até aqui (*Credit Score e Behavior Score*), visam reduzir o percentual de clientes inadimplentes através da qualificação do processo de concessão do crédito, identificando e qualificando, *a priori*, os riscos de cada pretenso cliente se tornar inadimplente. O *Colletion Score* busca estimar a probabilidade de pagamento de clientes que já estão em situação de inadimplência. Este tipo de modelo viabiliza a mensuração das perdas baseado na probabilidade de pagamento, permitindo que a instituição, entre outras iniciativas, consiga separar os clientes entre aqueles que precisam de ação de cobrança imediata, daqueles que não precisam. Além disso, como o modelo é construído com clientes que já possuem relacionamento com a instituição, existe uma maior e mais abrangente disponibilidade de dados sobre o comportamento histórico do cliente.

Gonçalves, 2016, sugere a que as variáveis de um modelo *Collection Score* podem ser agrupadas nas seguintes categorias:

1. Dados de Registro: Idade dos clientes, gênero, estado civil, endereço, etc. e informações obtidas em birôs de crédito, como os casos do SPC e SERASA no Brasil;
2. Relacionamento do Cliente com a Instituição: atraso de pagamento nos meses anteriores, longevidade do relacionamento do cliente com o credor, montante gasto pelo cliente na instituição em transações anteriores, contatos prévios com o cliente, entre outras informações;

De acordo com Crook *et al.* 2007, citado por Gonçalves, 2016, existem vários modelos de mensuração do risco de inadimplência de uma carteira, informando que a mais utilizada para criação do modelo é a Regressão Logística. O pesquisador cita ainda técnicas como Árvore de Decisão, Redes Neurais, Algoritmos Genéticos e Análise de Sobrevivência. Gouvêa*, et al*. 2012, cita, além desses, a Análise Discriminante e a Regressão Linear ressaltando que não há um método claramente melhor que os demais, tudo depende de como a técnica escolhida se ajusta aos dados. A Regressão Logística apresentou excelente adequação aos dados disponíveis nesta pesquisa, além de uma praticidade de aplicação e de atualização, atributos muito importantes tendo em vista que os dados mudavam a cada mês. Esses fatores em conjunto levaram a seleção desta técnica para o desenvolvimento do trabalho.

A Tabela 10 apresenta os resultados obtidos por estudos similares consultados.

Tabela 10 - Resultados de regressões de outros estudos

Referência	Técnica	Amostra	Porcentagem de acerto
Yap et al. (2011)	Regressão logística	2765 casos de uma instituição local na Malásia	71,52%
Mavri et al. (2008)	Regressão logística	350 casos de um Banco Europeu – solicitação de cartão de crédito	71,87%
Šušteršič et al. (2009)	Algorítmos genéticos	581 casos de um Banco Eslováco – solicitação de empréstimos	76,5% no modelo 1 (seleção das amostras pelo modelo de Kohonen) 72,7% no modelo 2 (seleção aleatória das amostras de treinamento de validação)
Brown e Mues (2012)	Regressão Logística	Cinco conjuntos de dados: - Banco Benelux 1: 2974 casos - Banco Bebelux 2: 7190 casos - Banco Austrália: 547 casos - Banco Alemanha: 1000 casos - Banco Benelux 3: 1197 casos	- Benelux 1: 76,9% - Bebelux 2: 78,7% - Austrália: 90,6% - Alemanha: 76,7% - Benelux 3: 63,4%
Ulises e Carmon (2011)	Regressão Logística	- 200 casos – Fundo Rotativo de Ação e Cidadania Recife/Brasil	80%
Brito e Assaf Neto (2008)	Regressão Logística	- 60 empresas listadas na Bovespa classificadas solventes /insolventes	90%

Fonte: Gouvêa (2012)

2.1.3.1 Regressão Logística

Os modelos de regressão logística, embora bastante úteis e de fácil aplicação, ainda são pouco utilizados em muitas áreas do conhecimento humano. Embora o desenvolvimento de softwares e o incremento da capacidade de processamento dos computadores tenham propiciado a sua aplicação de forma mais direta, muitos pesquisadores ainda desconhecem as suas utilidades e, sobretudo, as condições para que seu uso seja correto (Favero, 2015).

A principal diferença da Regressão Logística em relação as demais técnicas de regressão estimadas por métodos de mínimos quadrados, é que, nestes últimos, a variável dependente apresenta-se na forma quantitativa, enquanto que no primeiro caso a técnica é utilizada quando o fenômeno a ser estudado apresenta-se de forma qualitativa, representado por uma ou mais variáveis *dummy*, que são variáveis categóricas binárias, por isso a técnica é também chamada de Regressão Logística Binária.

Trazendo a técnica para o campo de pesquisa deste trabalho, foi utilizada a regressão logística com o objetivo de estudar a probabilidade de ocorrência do evento pagamento, considerando aqui o pagamento necessário para evitar o impacto no índice Bacen (ou seja, pagamento integral da dívida ou apenas o pagamento da prestação de maior atraso, evitando completar a quarta prestação em atraso), apresentando assim a forma qualitativa dicotômica: $Y = 1$ ocorrência do evento (pagamento) $Y = 0$ ocorrência do não evento. Desta forma, poderemos

definir um vetor de variáveis explicativas, com seus respectivos parâmetros, estimados da seguinte forma:

$$Z_i = \propto + \sum_{j=1}^{k} \beta_j X_{ji} \ , \ \forall\, i = 1, \dots, n$$

Em que Z_i é o denominado logito, $\propto$ representa a constante da regressão, βj (j = 1, 2,..., k) são os parâmetros calculados para cada variável explicativa, X_j são as variáveis explicativas (métricas ou *dummies*) e o subcrito i representa cada observação da amostra (i = *1, 2,..., n,* sendo n o tamanho da amostra). Importante ressaltar que Z não representa uma variável dependente como no caso das regressões lineares, o que se objetiva é encontrar a expressão da probabilidade P_i de ocorrência do evento de interesse (i.e. pagamento) para cada observação i (cada cliente), em função do logito Z_i, ou seja, em função dos parâmetros estimados para cada variável explicativa.

O conceito de chance de ocorrência de um evento, também conhecido na literatura estatística por *odds*, é definido da seguinte forma:

$$Chance\ (odds)_{Y_i=1} = \frac{P_i}{1 - P_i}$$

Assim, tratando do evento "pagamento", se a probabilidade de um determinado cliente regularizar seu débito for de 80%, a sua chance (*odds*) será de 4 para 1. Se a probabilidade de um segundo cliente for de 60%, sua chance de pagar será de 1,5 para 1. Apesar de no cotidiano chance e probabilidade estarem associadas, estatisticamente são conceitos diferentes.

Na regressão logística binária o logito Z é definido como o logaritmo natural da chance, de modo que $\ln(Chance_{Y_i=1}) = Z_i$. De outra maneira; $ln\left(\frac{P_i}{1-P_i}\right) = Z_i$. Portanto, temos que a probabilidade de ocorrência do evento (pagamento) é $p_i = \frac{1}{1+e^{-Z_i}}$; e a probabilidade de ocorrência do não evento (ausência de pagamento) é $1 - p_i = \frac{1}{1+e^{Z_i}}$.

Favero (2015), traz uma tabela apresentando os valores da probabilidade p em função dos valores de Z. Mesmo Z podendo variar de $-\infty$ a $+\infty$, o autor ilustra apenas a variação com números inteiros para o logito Z variando entre -5 a + 5.

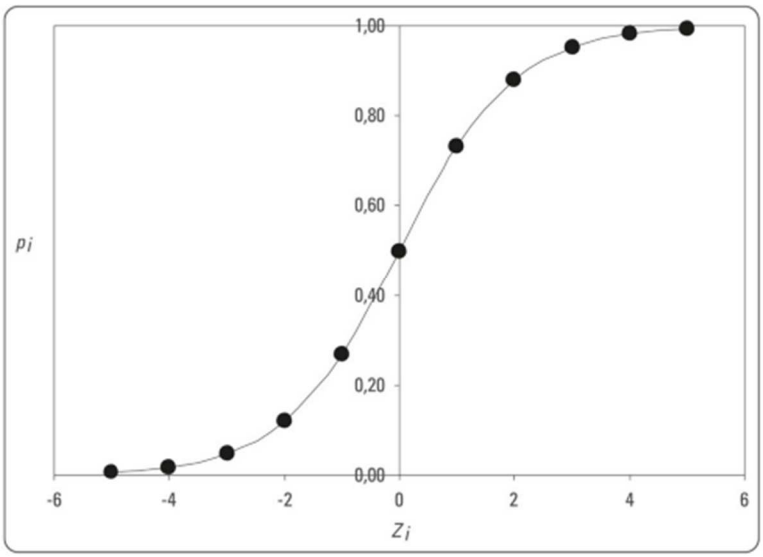
Tabela 11 - Probabilidade de ocorrência de um evento (*p*) em função do logito *Z*

$p_i = \frac{1}{1+e^{-z_i}}$	z_i
0,0067	-5
0,0180	-4
0,0474	-3
0,1192	-2
0,2689	-1
0,5000	0
0,7311	1
0,8808	2
0,9526	3
0,9820	4
0,9933	5

Fonte: Fávero (2015)

A representação gráfica de $p = f(Z)$ das probabilidades estimadas em função dos valores assumidos por *Z*, situam-se no intervalo entre 0 e 1, o que foi garantido quando se impôs que o logito fosse igual ao logaritmo natural da chance. Com isso, será possível estimar a probabilidade de ocorrência do evento em estudo para cada observação (cliente) a partir das variáveis explicativas e do cálculo de *Zi*, representado por meio da curva logística abaixo apresentada na Figura 11.

Figura 11 - Gráfico da função de $p = f(Z)$



Fonte: Fávero (2015)

A partir das equações que definem o logito Z e a probabilidade de ocorrência do evento (Z_i), podemos definir a expressão geral da probabilidade estimada de ocorrência de um evento que se apresenta na forma dicotômica para uma observação *i* da seguinte forma:

$$p_i = \frac{1}{1 + e^{-(\propto + \sum_{j=1}^{k} \beta_j X_{ji})}}$$

Desta forma, fica evidenciado que a regressão logística binária estima não um valor previsto para a variável dependente, mas sim, a probabilidade de ocorrência do evento. Por exemplo, a ocorrência do pagamento de cada cliente previsto numa amostra.

Há, contudo, que se levar em consideração que, como a variável resposta é uma variável qualitativa dicotômica, no caso deste estudo: ocorrência ou não ocorrência do pagamento, não há como minimizar o somatório dos quadrados dos resíduos para se chegar a uma equação de regressão. Favero (2015) aponta que Sharma (1996) indica a estimação por máxima verossimilhança como técnica para estimação dos parâmetros em modelos de regressão logística. Sendo a função de verossimilhança definida por:

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

Sendo o logaritmo da função de verossimilhança definido por:

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

O objetivo central para a elaboração de uma regressão por máxima verossimilhança (ou *maximum likelihood estimation*) é encontrar os valores dos parâmetros do logito que fazem com que o valor do *LL* seja maximizado. Existem muitas ferramentas de programação linear que auxiliam nesse cálculo, além da ferramenta Solver do Excel, a fim de que $\propto, \beta_1, \beta_2, \ldots \beta_k$ sejam estimados visando alcançar a seguinte função-objetivo:

$$LL = \sum_{i=1}^{n} \left\{ \left[(Y_i) . \ln \left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] + \left[(1 - Y_i) . \ln \left(\frac{1}{1 + e^{Z_i}} \right) \right] \right\} = \text{máx}$$

Tendo estimado por máxima verossimilhança os parâmetros da equação de probabilidade de ocorrência do evento, sabemos que a expressão de probabilidade encontrada a partir dos coeficientes das variáveis é a solução ótima.

2.1.3.2 Testes de Ajuste e Significância Estatística

Um dos testes estatísticos a serem utilizados para verificar a significância de cada conjunto de variáveis independentes selecionadas para predizer o comportamento da variável dependente é o teste X^2 que visa verificar a significância do modelo, uma vez que as hipóteses nula e a hipótese alternativa para um modelo geral de regressão logística, são, respectivamente:

$H_0: \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1: existe\ pelo\ menos\ um\ \beta_j \neq 0$

Assim, se o teste X^2 informar que os parâmetros propostos (estimados) no modelo forem estatisticamente iguais a 0, indicará que qualquer alteração nas variáveis dependentes não influenciará em nada a probabilidade de ocorrência do evento, sendo o teste definido pela seguinte expressão:

$$X^2 = -2\ (LL_0 - LL_{máx})$$

Sempre que a X^2 calculado for maior que o X^2 crítico para a mesmo número de graus de liberdade e para o mesmo nível de significância, poderemos rejeitar a hipótese nula de que todos os parâmetros β_j $(j = 1,2,3, \ldots,10)$ sejam estatisticamente iguais a zero, ou seja, pelo menos uma variável independente é estatisticamente significante para explicar a probabilidade de ocorrência do evento e o modelo poderá ser considerado estatisticamente significante.

Para realizar a verificação individual dos parâmetros de cada uma das variáveis utilizadas no modelo, utilizaremos o teste Z de Wald, avaliando sua significância estatística ao nível de 5%, o que permitirá excluir do modelo aquelas variáveis que, com 95% de confiança, não poderão ser consideradas estatisticamente diferentes de zero.

A estatística Z de Wald é importante para fornecer a significância estatística de cada parâmetro a ser considerado no modelo. A nomenclatura Z refere-se ao embasamento desta estatística na distribuição normal padrão. Sendo suas hipóteses para o teste Z de Wald definidas, respectivamente por:

$H_0: \propto = 0$

$H_1: \propto \neq 0$

$H_0: B_j = 0$

$H_1: B_j \neq 0$

As expressões para o cálculo das estatísticas Z de Wald de $\propto$ e cada parâmetro B_j, são dadas, respectivamente, por:

$$Z_{\propto} = \frac{\propto}{s.e.(\propto)}$$

$$Z_{B_j} = \frac{B_j}{s.e.(B_j)}$$

Em que *s.e.* significa o erro-padrão *(Standard Error)* de cada parâmetro em análise.

Calculado o valor das estatísticas *Z* de Wald, é possível utilizar a tabela de distribuição normal padrão para obter os valores críticos de cada nível de significância, verificando se os testes rejeitam, ou não, a hipótese nula. Para o nível de significância de 5%, teremos um *zc* = -1,96 para a cauda inferior (probabilidade na cauda inferior de 0,025 para a distribuição bicaudal) e um *zc* = 1,96 para a cauda superior (probabilidade na cauda superior também de 0,025 para a distribuição bicaudal). Assim, caso o *Z* de Wald de algum dos parâmetros estimados na regressão apresente valor entre -1,96 e 1,96; significará que essa variável independente, ao nível de significância de 5%, não rejeitou a hipótese nula, ou seja, o parâmetro não pode ser considerado estatisticamente diferente de zero. Em outras palavras, o parâmetro não é estatisticamente significante para aumentar ou diminuir a probabilidade de ocorrência do evento na presença das demais variáveis explicativas, portanto, poderá ser excluída do modelo final.

Existe ainda o teste de Hosmer-Lemeshow que aplica um teste qui-quadrado X^2 com g-2 graus de liberdade à base de dados dividida em 10 grupos a partir dos decis das probabilidades estimadas no modelo final, verificando se existem diferenças significativas entre as frequências observada e a esperada em cada grupo, e caso tais diferenças não sejam estatisticamente significativas a um determinado nível de significância, conclui-se que o modelo estimado não apresenta problemas em relação à qualidade do ajuste proposto.

A aplicação do teste é realizada a partir do somatório dos testes qui-quadrado com g-2 graus de liberdade, conforme segue:

$$\sum_{i=1}^{g}\sum_{j=1}^{2}\frac{\left(obs_{ij} - exp_{ij}\right)^2}{exp_{ij}}$$

Onde:

g = número de grupos

O suplemento do Excel já informado em seções anteriores, além da regressão logística, realiza também de forma simultânea o cálculo desses indicadores de ajuste do modelo, ainda assim, realizaremos uma breve revisão teórica de cada um deles nas seções subsequentes.

2.1.3.3 Curva ROC (*Receiver Operating Characteristic*) e Eficiência do Modelo

Favero (2015) reputa à curva ROC a consideração de melhor critério para escolha de um modelo preditivo. No dia a dia das organizações e das academias, é comum a apresentação de relatórios com gráficos que apresentam análises de sensibilidade e, visualmente, transformam dados em informação que subsidiam a tomada de decisão. No que diz respeito às regressões logísticas, a representação gráfica que melhor apresenta os resultados da análise é a Curva ROC, a qual demonstra as variações da sensitividade em função de (1 – especificidade).

Uma vez estimado o modelo e, assim, já mensurados os valores p_i de probabilidade para cada observação da base, o passo seguinte é o estabelecimento de um ponto de corte (*cutoff*). Esse ponto é utilizado para se elaborarem as previsões de ocorrência do evento para observações futuras, utilizando como subsídio as probabilidades presentes na amostra. Assim, se uma observação apresentar uma probabilidade *p:* superior ao *cutoff* definido, espera-se que haja incidência do evento e, portanto, será classificada como evento. Em resumo:

Se *p:* > *cutoff*, então a observação *i* deverá ser classificada como evento.

Se *p:* < *cutoff*, então a observação *i* deverá ser classificada como não evento.

Segundo Fávero (2015), o *cutoff* serve para avaliar a real incidência do evento para cada observação para compará-la com a expectativa de cada observação incida, de fato, no evento. Com isto feito, será possível avaliar a taxa de acerto do modelo com base nas próprias observações presentes na amostra e, por inferência, assumir que tal taxa de acerto se mantenha quando houver o intuito de avaliar a incidência do evento para outras observações não presentes na amostra (previsão).

Realizando a comparação da quantidade de ocorrências classificadas como “Evento” e “Não Evento” pelo modelo, com a quantidade de ocorrências de fato observadas, realizamos a chamada Análise de Sensibilidade, a qual, informa a Eficiência Global do Modelo que corresponde ao percentual de acerto da classificação para um determinado *cutoff*. A matriz de classificação abaixo contribui para o entendimento do modelo:

Tabela 11 - Modelo da matriz de classificação para um dado *cutoff*

	Incidência Real do Evento	Incidência Real do Não Evento
Classificado como Evento	X_{evento}	Y_{evento}
Classificado como Não Evento	$Y_{não\ evento}$	$X_{não\ evento}$

Onde:

X é a quantidade de ocorrência previstas no modelo e observadas na amostra, tanto para o "evento", quanto para o "não evento";

Y é a quantidade de ocorrência previstas no modelo e não observadas na amostra, tanto para o "evento", quanto para o "não evento";

Assim, a Eficiência Global do Modelo poderá ser calculada da seguinte forma:

$$EGM = \frac{X_{evento} + X_{não\ evento}}{X_{evento} + X_{não\ evento} + Y_{evento} + Y_{não\ evento}}$$

Um *cutoff* de 0,5 implica que todos os elementos da amostra que tenham um $p_i > 0{,}5$ serão classificados como Evento, enquanto os $p_i < 0{,}5$ serão classificados como não evento. Por isso, uma redução no *cutoff* para 0,4 por exemplo, fará com que uma quantidade maior de elementos da amostra sejam classificados como Evento e, por consequência, uma quantidade menor como Não Evento, influenciando assim a taxa de eficiência do modelo.

Outro aspecto relevante a ser considerado na avaliação do modelo, diz respeito ao percentual de acerto para um determinado *cutoff*, a denominada Sensitividade, a qual calcula apenas observações que de fato são evento, tendo sua expressão definida por:

$$Sensitividade = \frac{X_{evento}}{X_{evento} + Y_{não\ evento}}$$

Existe ainda a especificidade, que por outro lado, calcula o percentual de acerto do modelo considerando apenas as observações do não evento e cuja expressão é dada por:

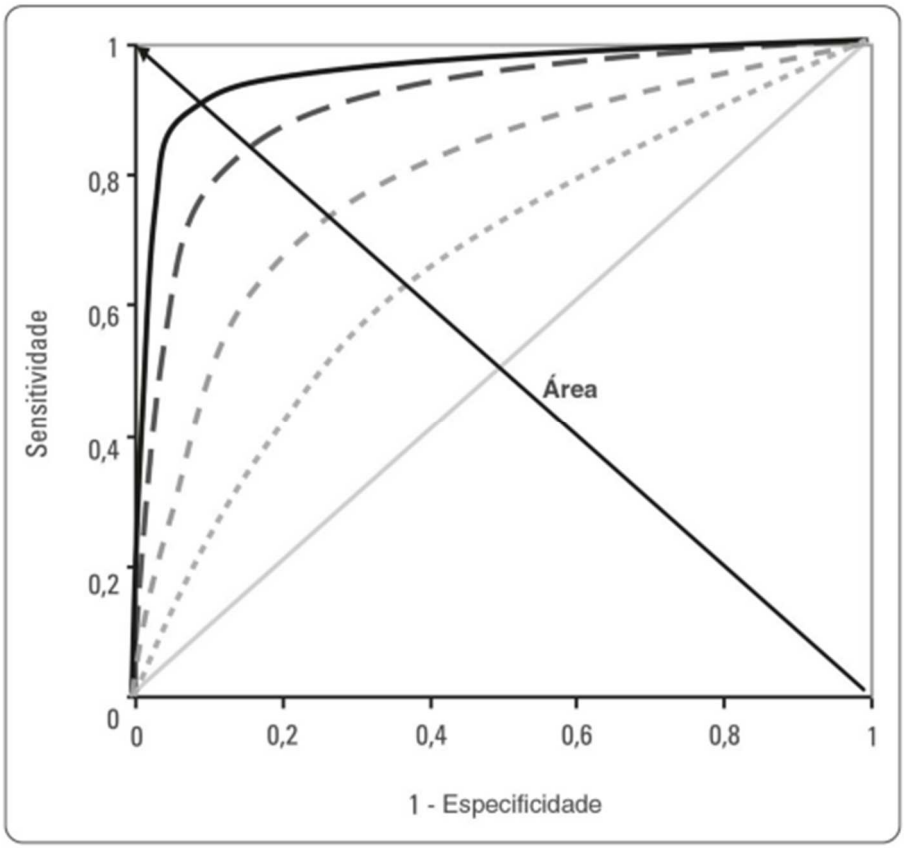
$$Especificidade = \frac{X_{não\ evento}}{X_{não\ evento} + Y_{evento}}$$

Assim como a Eficiência Global do Modelo, a Sensitividade e a Especificidade também são alterados a cada *cutoff*. A Sensitividade tende aumentar quando o *cutoff* é reduzido, pois, uma maior quantidade de elementos serão classificados na categoria Evento (X_{evento}), aumentando com isso a quantidade de previsões Evento com correspondência nas observações e, ainda, por decorrência, reduzindo na mesma proporção, a quantidade de observações classificadas erroneamente como Não Evento ($Y_{não\ evento}$). Esse aumento causa também

impacto na Especificidade, sendo que na direção oposta, um *cutoff* menor faz com que uma quantidade menor de Não Eventos ($X_{não\ evento}$) passem a ser previstos, incrementando a quantidade de Eventos classificados de forma equivocada (Y_{evento}).

A curva *ROC* mostra o comportamento propriamente dito do *trade off* entre a sensitividade e a especificidade, ao trazer no eixo das abscissas os valores de (1 - especifiidade), apresentando formato convexo em relação ao ponto (0,1). Desta forma, um determinado modelo com maior área abaixo da curva *ROC* apresenta maior eficiência global de previsão, combinados todos os níveis de *cutoff* e, sendo assim, preferível quando da comparação com outro modelo com menor área abaixo da curva. Ou seja, na comparação com outros modelos que possuam variáveis explicativas diferentes, aquele que apresentar a maior área abaixo da curva, (maior convexidade), terá maior sensitividade e maior especificidade, sendo assim, considerado o melhor modelo para efeitos de previsão.

Figura 12 - Critério de escolha do modelo de acordo com a maior área abaixo da curva *ROC*



Fonte: Fávero (2015)

Segundo Swets (1996), citado por Favero (2015), a curva ROC (*Receiver Operating Characteristic*) foi primeiramente desenvolvida e utilizada por engenheiros na Segunda Guerra Mundial quando do estudo para detecção de objetos inimigos em batalhas. Na sequência, foi utilizada na Psicologia e é bastante utilizada em campos da Medicina, como a radiologia, e das ciências sociais aplicadas, sendo consideravelmente aplicada em modelos de gestão de risco de crédito e de probabilidade de *default*. Recebe este nome por comparar duas características operacionais do modelo (sensitividade e especificidade).

Segundo Hosmer e Lemeshow (2000), a regra geral para avaliação do resultado da área sob a curva ROC de modelos de *Credit Scoring* é dada por:

área < 0, 7: baixa discriminação

0, 7 _ área < 0, 8: discriminação aceitável

0, 8 _ área < 0, 9: discriminação excelente

área > 0, 9: discriminação excepcional

3 METODOLOGIA

Visando colher subsídios para o desenvolvimento e aplicação do método, foram utilizados o Google Escolar e o portal de periódicos da CAPES, com foco nas bases *Web of Science* e *Scopus* para capturar as contribuições de outras pesquisas sobre o tema objeto deste estudo.

Com o objetivo de estabelecer um processo que permita priorizar a cobrança àqueles clientes com maior chance de realizar o pagamento, foram percorridas as seguintes etapas, conforme os passos para construção de modelos (HILLIER e LIEBERMAN, 2001):

- Definição do problema: Quais clientes priorizar para realização de ações de cobrança;
- Construção do modelo: Seleção da ferramenta ou processo estatístico adequado, bem como, coleta dos dados necessários;
- Solução do modelo: obter da ferramenta a resposta procurada;
- Validação do modelo: Realizar os testes de significância estatística e de poder preditivo do modelo;
- Implementação da solução em uma situação real;

Por ser uma ferramenta bem referenciada na literatura, possuir uma aplicabilidade bastante prática e um ótimo ajuste aos dados e ao objetivo desta pesquisa, a ferramenta utilizada neste trabalho para elaboração do *Collection Score* será a Regressão Logística, ferramenta estatística já pormenorizada na Seção 2.

Existem muitas contribuições com aplicações práticas da técnica de regressão logística, cada qual realizando as adaptações pertinentes ao objeto em estudo. Lopes (2004) realizou uma aplicação da regressão logística multinomial para a definição de um modelo de cobrança de clientes em uma carteira de cartão de crédito. O estudo foi exitoso em encontrar um modelo capaz de prever a probabilidade de pagamento de um cliente inadimplente através do *score* criado usando os coeficientes estimados pelo modelo para cada parâmetro (variáveis explicativas).

Na mesma linha de pesquisa, Oliveira *et al.* (2011), desenvolveram um *Collection Score* com o objetivo de auxiliar nas decisões de cobrança também utilizando a regressão logística binária, associada aos Testes de Hosmer-Lemeshow, Kolmogorov-Smirnov e a Curva ROC para validação do modelo. O modelo gerado apontou as chances de recuperação para cada cliente endividado, classificando-o como “bom” ou “mau” pagador, através das seguintes

variáveis explicativas: valor da fatura, valor de pagamento, dias de atraso, quantidade de atrasos, limite do crédito, entre outras.

Outro trabalho utilizado como referência foi o da Araújo (2016), que também por meio da regressão logística, construiu um modelo de apoio à tomada de decisão inteligente para cobrança de clientes com faturas de consumo de energia elétrica em atraso. Para tanto, a autora utilizou dados da Companhia de Eletricidade do Estado da Bahia (COELBA), incluindo o montante da dívida do cliente, a quantidade de faturas vencidas e o tempo da dívida, obtendo um modelo com 84% de assertividade na série histórica de seis meses.

Este estudo, apresenta um novo processo para acompanhamento e avaliação de clientes a serem priorizados, por meio da análise de *Collection Score*, e implementadas em um banco de grande circulação (Santos, *et al*. 2021).

3.1 COLETA DE DADOS

A obtenção de dados necessários à criação e validação do modelo a ser proposto foi realizada por meio dos registros da própria centralizadora de adimplência. Restringiremos a aplicação da análise a carteira habitacional pelos motivos já expostos na Seção 1, que são: trata-se da carteira mais relevante da unidade, é a carteira que possui a maior homogeneidade nos parâmetros do produto, ou seja, mesma garantia, as taxas de juros vigentes possuem baixa dispersão; 99,5% da carteira são de clientes Pessoa Física e, por fim, possui uma condição negocial para o adimplemento de baixa complexidade e rápida implementação.

Para alimentar o modelo é necessário, primeiramente, discriminar as variáveis explicativas que podem influenciar a probabilidade de ocorrência do evento, ou seja, do cliente adimplir ou ter no máximo 90 dias de atraso. Analisando a base de clientes do segmento habitacional, elencamos inicialmente algumas variáveis que, pelo conhecimento empírico consolidado pela experiência prática do dia a dia, tem influência significativa na probabilidade de reversão da inadimplência pelo cliente, e outras que possuem um racional lógico-econômico que, em tese, também fomenta essa propensão ao pagamento, sendo identificadas inicialmente uma relação de 10 variáveis explicativas contínuas e 2 variáveis *dummy*, as quais encontram-se listadas abaixo:

a) Dias de Atraso: o conhecimento empírico informa que quanto menor o atraso, maior é a probabilidade de reversão/adimplemento do contrato;
b) Saldo devedor do contrato: o valor do saldo devedor evidencia uma capacidade financeira proporcional, ainda que essa condição possa se alterar no decorrer da

vigência do contrato habitacional, que costumam possuir prazos superiores a 10 anos;

c) Valor da garantia;

d) Idade do mutuário;

e) Prestação paga por débito em conta (variável *dummy*: "1" para sim e "0" para não);

f) Relação entre Valor do Atraso e Saldo Devedor $\left(\frac{VA}{SD}\right)$: Essa variável visa capturar os clientes cuja prestação está amortizada (valor de atraso pequeno em relação ao saldo devedor), neste sentido, quanto menor o valor da fração, significa que menor é o valor a ser pago pelo cliente para regularizar seu débito em proporção ao montante do dívida;

g) Relação entre Saldo Devedor e Valor da Garantia $\left(\frac{SD}{VG}\right)$: Essa relação apresenta um raciocínio econômico que é o desdobramento na retomada dos imóveis pelo banco nos casos de inadimplência prolongada. Assim, o cliente que possua um Saldo Devedor baixo em relação ao valor do seu imóvel (Garantia) estará mais propenso a regularizar o débito, ainda que seja vendendo o imóvel, buscando preservar parte do valor já pago;

h) Relação entre Valor do Atraso e Valor da Garantia $\left(\frac{VA}{VG}\right)$: Esta variável tenta capturar o mesmo racional do item anterior utilizando o Valor do Atraso, no lugar do Saldo Devedor;

i) Possui e-mail no cadastro (variável *dummy*: "1" para sim e "0" para não): pela dinâmica de cobrança, são enviados e-mails para todos os clientes da base, inclusive os que serão acionados também por telefone, e essa possibilidade de dupla abordagem, aumenta a probabilidade de reversão do contrato;

j) Prazo restante: quantidade de meses para liquidação do contrato;

k) Total em atraso: montante das prestações em atraso;

l) Taxa de juros do contrato: partindo do racional de que taxas mais elevadas, encarecem a prestação;

Dado o grande volume de dados, em média 38 mil clientes por mês, foi utilizado o Suplemento XRealStats do Excel, abastecendo com dados dos meses de Outubro/2020 a Fevereiro/2021, para calcular os coeficientes da equação de probabilidade. Além de viabilizar o tratamento dos dados, o suplemento realiza também a maximização da função de

verossimilhança, que é a técnica de estimação mais popular de modelos de regressão logística, trabalhando com o logaritmo da função, também conhecido por *log likelihood function.*

$$LL = \sum_{i=1}^{n} \left\{ \left[(Y_i).ln\left(\frac{e^{Z_i}}{1+e^{Z_i}}\right) \right] + \left[(1-Y_i).ln\left(\frac{1}{1+e^{Z_i}}\right) \right] \right\} = \text{máx}$$

O objetivo é calcular e estabelecer a relação de prioridade dos clientes (collection score) a serem encaminhados para o canal acionamento por telefone a cada mês, a partir dos coeficientes estimados considerando os dados históricos, a probabilidade de ocorrência do evento pagamento, repetindo e aprimorando esse processo continuamente.

3.2 ANÁLISE DOS DADOS E CONSTRUÇÃO DO MODELO

O tratamento estatístico dos dados será feito utilizando o Suplemento XRealStats do Excel, aplicando a função de Regressão Logística aos dados dos meses de Outubro/2020 a abril/2021. A seleção desses meses justifica-se pela existência de uma condição negocial específica (i.e., a negociação emergencial) vigente nesse período, onde o cliente que possua uma ou mais prestações em atraso, pode, através do pagamento de uma entrada com valor equivalente ao da prestação mais atual, incorporar todo o valor em atraso e optar pelo pagamento parcial de 75% da prestação pelos 6 meses seguintes, ou 50% do valor da prestação por 3 meses.

Reconhecendo que essa condição negocial influencia sobremaneira a possibilidade de adimplemento por parte do cliente, que, de outro modo, para colocar seu contrato em dia precisaria pagar 3, 4 até 5 prestações de uma só vez, a análise dos dados foi realizada somente nos meses cuja condição negocial era a mesma, no caso a negociação emergencial, visando blindar ao máximo o modelo de variáveis externas que pudessem influenciar nos resultados obtidos.

O resultado esperado da utilização da regressão logística é a escala de priorização de cliente de acordo com a probabilidade de reversão de cada um, no qual os clientes com o *score* mais elevado serão aqueles com maior propensão à reversão (*Collection Score*). A validação dos resultados será feita pelos testes X^2, teste de Hormes-Lemeshow, pela *Receiver Operating Characteristc* (Curva ROC) e pela Estatística *Z* de Wald.

Por último, com a finalidade de avaliar o poder preditivo do modelo encontrado, testaremos a efetividade da previsão avaliando *ex-post* os clientes com predição de pagamentos.

Em outras palavras, utilizando os parâmetros encontrados pelo modelo do mês “X”, avaliaremos qual seria a previsão de clientes pagadores para o mês X+1, e, finalizado o mês (X+1), avaliaremos a acuracia da previsão estimada em X.

4 *COLLECTION SCORING* COMO FERRAMENTA DE DEFINIÇÃO DAS AÇÕES DE COBRANÇA

Este capítulo apresenta as questões decisórias enfrentadas pelos gestores da unidade todo início de mês, ocasião onde ocorre a atualização das novas bases de contratos que impactarão os diferentes índices do período. Considerando que a capacidade de acionamento por telefone da equipe é finita e insuficiente para cobrir toda a base de contratos, a questão a ser respondida é: quais contratos priorizar no canal acionamento por telefone.

4.1 APLICAÇÃO PRÁTICA DO MODELO

A centralizadora de cobrança possui capacidade operacional para acionar, via canal telefone, em média 7.000 clientes do segmento habitacional todos os meses, a depender da quantidade de dias úteis e funcionário disponíveis. Como a base de dados envolve uma média de 38.000 contratos mensais de clientes com diferentes faixas de atraso, valores de dívida, valores de garantia, etc., um recorte dessa base precisa ser realizado.

Seguindo a definição de inadimplência do Banco Central do Brasil, o cliente passa a ser considerado inadimplente assim que completa o 91º dia de atraso, data exata em que, não só o valor em atraso, mas todo o saldo devedor do contrato passa a impactar o índice de inadimplência da instituição. Em outras palavras, caso um cliente que possua um financiamento habitacional com 240 meses de vigência e prestação mensal de R$1.000,00, e complete a 4º prestação em atraso (91 dias), o valor de todo o saldo devedor do contrato (os 240 mil dos meses restantes) impactará no índice de inadimplência.

Desta forma, a base disponibilizada a cada início de mês, traz os contratos que possuem uma quantidade de dias de atraso suficiente para completar o 91º dia de atraso no decorrer do mês. Na Centralizadora de Cobrança, essa base de contratos é chamada de “Pressão do Mês”. Ou seja, como o mês de Março possui 31 dias, aqueles contratos que no primeiro dia do mês possuam a partir de 61 dias de atraso, farão parte do foco da cobrança (Pressão do Mês), cujo objetivo será fazer com que o cliente pague ao menos 1 prestação, livrando o impacto do saldo devedor do contrato no índice de inadimplência da instituição, ou regularize sua situação ficando com atraso de zero dias.

Depois de superada a marca dos 91 dias de atraso, o contrato passa a fazer parte da base já impactada no índice BACEN, permanecendo ali até que o cliente regularize o contrato, ou que o contrato complete o mínimo de 6 meses em rating H e seja lançado em prejuízo nos

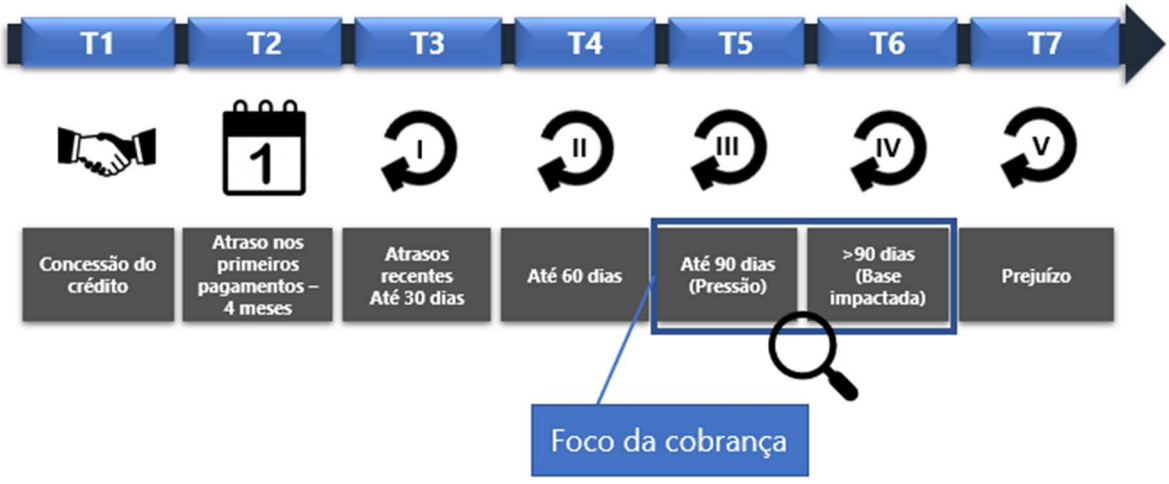
balanços financeiros da instituição financeira, conforme preconizado na resolução BACEN 2.682/99.

Sobre o lançamento em prejuízo, a Carta-Circular Bacen 2.899, item 12, inciso VII, 2000, preconiza que os créditos baixados como prejuízo devem ser registrados em contas próprias do sistema de compensação, em subtítulos adequados a identificação do período em que ocorreu o registro, devendo ser mantido controle analítico desses créditos, com identificação das características da operação, devedor, valores recuperados, garantias e respectivas providencias administrativas e judiciais, visando a sua recuperação.

Esse conjunto de normas traz, entre outras, uma consequência muito importante acerca do processo de cobrança que precisa ficar suficientemente clara para o melhor entendimento das decisões que são tomadas. Os contratos da base impactada, ou seja, aqueles que já superaram os 90 dias de atraso, também fazem parte da ação de cobrança, uma vez que recuperados, compensam eventual contrato da Pressão não revertido. Porém, diferente dos contratos da pressão, esses clientes cujos contratos já impactaram o índice, precisarão realizar o pagamento de mais de uma prestação dentro do mês, a depender do período de atraso, de forma a acabar o mês com menos de 90 dias de atraso. Exemplo: um contrato que iniciou o mês com 150 dias de atraso, precisará pagar 4 prestações para finalizar o mês com até 90 dias de atraso.

A Figura 13 apresenta a linha do tempo do processo de cobrança acima descrito.

Figura 13 - Linha do tempo da inadimplência bancária



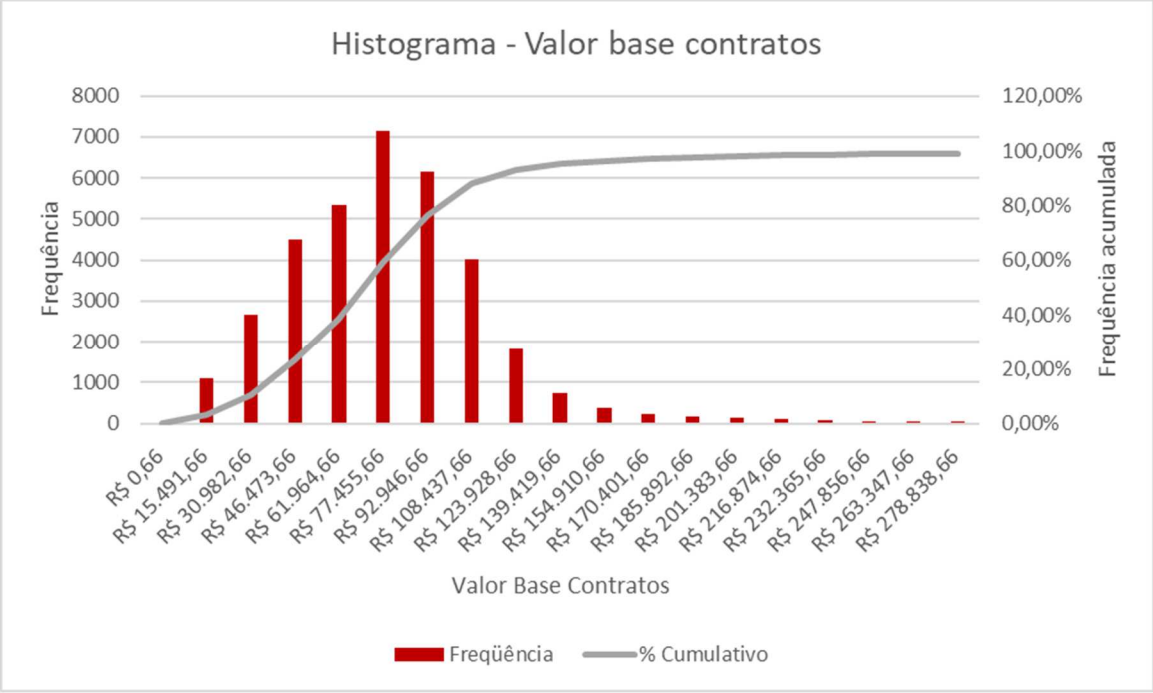
Fonte: O autor (2022)

Para subsidiar a compreensão do problema a ser decidido, enumeramos abaixo algumas informações sobre a base de contratos de clientes com inadimplência habitacional em um mês típico:

- 98,00% da base de contratos possui valor base inferior a R$201.383,66;
- 77,31% da base de contratos possui valor base (saldo devedor) contido no intervalo entre R$46.473,66 e R$108.437,66. Totalizando cerca 27.197 de um total de 35.175 contratos.

No Apêndice A encaminhamos a tabela de frequência completa da base cujas informações acima foram retiradas. Na Figura 14 apresentamos o gráfico do histograma com a distribuição dos contratos em suas respectivas frequências, correspondendo a uma curva de distribuição normal.

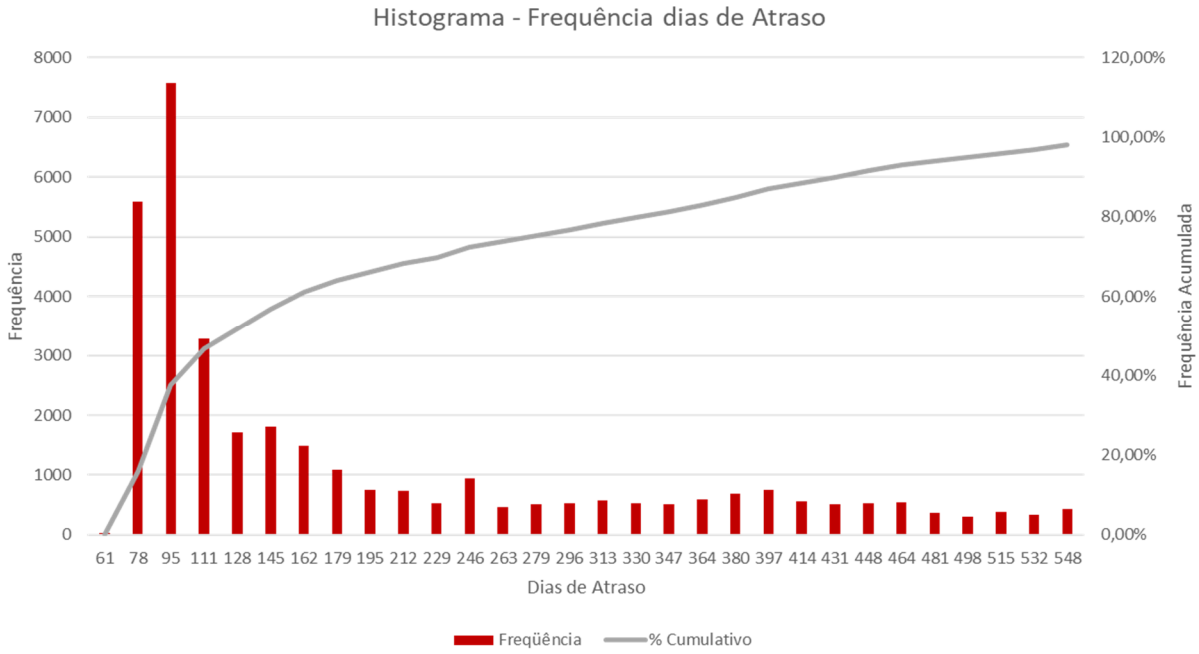
Figura 14 - Histograma com a frequência do valor base dos contratos em atraso



Fonte: O autor (2022)

O Apêndice 2 apresenta a tabela de distribuição acumulada dos dias de atraso da mesma base. Destacamos a maior concentração na faixa até 163 dias de atraso, a qual responde por 61,13% da quantidade de contratos, ou por 21.501 contrato de um total de 35.175, sendo que a faixa até 364 dias corresponde a 82,90%. A Figura 15 mostra o gráfico de distribuição dos dias de atraso, tendo um formato mais assemelhado ao de uma curva de distribuição Qui-quadrado:

Figura 15 - Histograma com a frequência dos contratos por dias de atraso



Fonte: O autor (2022)

4.2 EXECUÇÃO OPERACIONAL

Uma vez selecionados os contratos alvo da ação de cobrança, cada integrante da equipe recebe sua base fixa de acionamento para o decorrer do mês. A base é fixada com o intuito de viabilizar o acompanhamento individualizado do desempenho de cada membro, e permitir que esse possa realizar o acompanhamento dos agendamentos dos clientes da sua relação, e assim oportunizar o retorno a alguma solicitação do cliente durante a vigência do mês.

O fluxograma do processo de cobrança é apresentado na Figura 16:

Figura 16 - Fluxograma de cobrança ativa

Fonte: O autor (2022)

A relação de clientes, priorizada de acordo com o seu escore, visa discriminar a relação de clientes que abastecerá esse processo que hoje conta com cerca de 30 profissionais, sendo 7 profissionais com jornada de 8 horas e os demais com jornada de 6 horas diárias.

4.2.1 Aplicação do Modelo e Elaboração do Collection Score

Esta Seção descreve como foi realizada a aplicação da Regressão Logística à base de clientes da centralizadora de cobrança, detalhando desde a preparação da base à utilização do suplemento XRealStats do Excel, assim como os resultados encontrados na regressão e os testes estatísticos que asseveraram a significância estatística do modelo encontrado.

4.2.1.1 Preparação dos Dados

O primeiro desafio da pesquisa foi o de coletar os dados necessários à utilização da técnica, resgatando a informação de meses anteriores com a finalidade de obter uma massa de dados suficiente à validação do modelo, dispondo de todas as variáveis explicativas selecionadas para iniciar os testes. Tendo em conta sempre que a finalidade era calcular a probabilidade de pagamento de cada cliente de forma a priorizar aqueles com maior chance de pagamento. Em outras palavras, busca-se no modelo a resposta para a pergunta: se a centralizadora possui capacidade operacional média mensal de acionamento de 7.000 clientes por mês, e possui uma base média de 38.000 contratos, quais clientes/contratos selecionar para a abordagem?

Durante os meses de pesquisa, foram realizadas diversas regressões simulando diferentes combinações das variáveis independentes listadas na Seção 3. Também testamos o comportamento das mesmas variáveis independentes para diferentes faixas de valor base (valor do saldo devedor do contrato) utilizando os seguintes parâmetros:

1. Contratos com saldo devedor acima de R$300.000,00
2. Contratos com saldo devedor entre R$120.000,00 e R$300.000,00
3. Contratos com saldo devedor entre R$80.000,00 e R$120.000,00
4. Contratos com saldo devedor abaixo de R$80.0000,00

Com base nos testes realizados, constatou-se que o modelo que apresentou maior estabilidade na comparação mês a mês, com menor dispersão entre os coeficientes calculados,

aprovação nos testes de significância estatística e bom desempenho preditivo (avaliado pela EGM e pela Área Sob a Curva ROC), utilizando as variáveis abaixo relacionadas:

Figura 17 – Relação das variáveis independentes selecionadas

Variáveis Independentes Selecionadas	Unidade
Quantidade de dias de atraso (dias_atraso)	Unidade
Saldo devedor do contrato (vr_base)	Monetária
Relação entre valor do atraso de saldo devedor $\left(\frac{VA}{SD}\right)$	Percentual
Relação entre Saldo Devedor e Valor da Garantia $\left(\frac{SD}{VG}\right)$	Percentual
Disponibilidade de e-mail (e-mail_cliente)	Sim ou Não (binária)
Prazo restante do contrato (prazo_restante)	Unidade
Encargo em atraso (encargo_atraso)	Monetária
Total em atraso i.e., encargo em atraso somado aos juros e multa	Monetária

Fonte: O autor (2022)

Através dessas variáveis, relacionamos o Valor do Atraso e o Saldo Devedor $\left(\frac{VA}{SD}\right)$, Saldo Devedor e o Valor da Garantia $\left(\frac{SD}{VG}\right)$, variáveis que, conforme já informado na Seção 3.1. possuem um racional econômico e/ou empírico que se buscou aproveitar na aplicação da técnica. Neste sentido, Santos *et al.* (2021) apresentou análises nas quais o valor do coeficiente foi condizente com o racional econômico, citando o exemplo do coeficiente calculado para a variável $\left(\frac{SD}{VG}\right)$ ou $\left(\frac{Saldo\ Devdor}{Valor\ da\ Garantia}\right)$ em -0,53336, que pode ser traduzido como quanto menor a relação, ou seja, quão menor o saldo devedor for em relação ao valor da garantia, maior será o incentivo do cliente a regularização do débito.

Para realizar o cálculo dos coeficientes de cada variável, utilizando os dados já citados anteriormente, organizando-os no formato especificado no suplemento, sendo: todas as variáveis explicativas a esquerda e a variável resposta (representada por 1 para o evento pagamento e 0 para o não evento) na última coluna à direita. Observe ainda que a variável e-mail do cliente também é uma variável *dummy*, (apresentando 0 para não possui e-mail, e 1 para possui e-mail). Ao fim, a base de dados irá apresentar a configuração demonstrada na Figura 18:

Figura 18 - Base de contratos com a configuração especificada

dias_atraso	vr_base	VA/SD	SD/VG	email_cliente	prazo_restante	encargo_atraso	total_atraso	divida_total	Pagou
167	1575280,86	0,074811694	0,814157735	1	301	109768,24	117849,43	1698319,41	0
127	1184436,39	0,066973896	0,66974011	1	322	73444,27	79326,32	1237234,5	1
155	1013513,15	0,071418027	0,76878753	1	315	67070,47	72383,11	1053103,07	1
73	957762,75	0,03620928	0,811663347	1	345	31894,46	34679,9	969194,76	0
81	956128,59	0,034863731	0,330676524	1	336	31069,18	33334,21	967297,73	1
102	917998,4	0,04407109	0,871451699	1	333	38419,82	40457,19	937241,73	1
124	860325,5	0,090673146	0,682515368	1	153	73771,71	78008,42	922149,53	0
167	855482,33	0,077889581	0,831330589	1	225	60566,93	66633,16	897340,32	1
66	847380,81	0,03180616	0,848647841	0	402	25844,54	26951,93	857143,05	1
545	798716,06	0,264553601	0,786435128	1	243	176787,32	211303,21	949096,74	0

Fonte: O autor (2022)

Entre os diversos softwares especializados no tratamento estatístico de bases de dados, como: o SPSS, R Studio, Stata entre outros, a opção para realização deste trabalho foi pelo Microsoft Excel, por se tratar de um software que possui um uso extremamente difundido nas organizações e centros acadêmicos. Todos os cálculos realizados para estimar os coeficientes da regressão logística, bem como, os testes estatísticos de validação do modelo puderam ser feitos apenas com a utilização das funções já pré-formatadas na própria ferramenta e com a habilitação do suplemento XRealStats.

Após habilitado no Excel, de posse da base de dados já preparada, a utilização da ferramenta ocorre através do seguinte procedimento descrito no Apêndice 3.

4.3 INTERPRETAÇÃO DOS RESULTADOS ENCONTRADOS

O mês de Outubro de 2020 foi o primeiro mês a ter os dados da cobrança avaliado pelo modelo de regressão logística já descrito aqui e com as variáveis explicativas já mencionadas. A escolha desse mês para início dos testes foi devido ao fato de a ter sido o primeiro mês de uma condição negocial vigente até o presente momento, exclusiva para os clientes do segmento habitacional, a qual facilita o adimplemento pelo cliente, a chamada Negociação Emergencial já descrita anteriormente na Seção 3.

Após realizar a etapa descrita no Apêndice 3, os resultados do modelo são obtidos em formato ilustrado na Figura 19.

Figura 19 - Cálculos do Suplemento

	J	K	L	M	N	O	P	Q	R	S	T	U
1												
2												
3	*Success*	*Failure*	*Total*	*p-Obs*	*p-Pred*	*Suc-Pred*	*Fail-Pred*	*LL*	*% Correct*	*HL Stat*		Coeff
4	0	1	1	0	0,462682	0,462682	0,537318	-0,62117	100	0,861097		
5	0	1	1	0	0,472184	0,472184	0,527816	-0,63901	100	0,894599		1,024912
6	1	0	1	1	0,572506	0,572506	0,427494	-0,55773	100	0,746708		-0,00759
7	1	0	1	1	0,534344	0,534344	0,465656	-0,62672	100	0,871453		3,06E-05
8	1	0	1	1	0,544613	0,544613	0,455387	-0,60768	100	0,836166		-3,02068
9	0	1	1	0	0,599179	0,599179	0,400821	-0,91424	0	1,494878		-0,53336
10	1	0	1	1	0,587529	0,587529	0,412471	-0,53183	100	0,702043		0,283656
11	0	1	1	0	0,577016	0,577016	0,422984	-0,86042	0	1,364153		-0,00117
12	0	1	1	0	0,535487	0,535487	0,464513	-0,76676	0	1,152791		-0,00038
13	1	0	1	1	0,539536	0,539536	0,460464	-0,61705	100	0,853443		0,000348
14	1	0	1	1	0,549967	0,549967	0,450033	-0,5979	100	0,818291		-2,9E-05
15	0	1	1	0	0,548101	0,548101	0,451899	-0,7943	0	1,212882		
16	0	1	1	0	0,494336	0,494336	0,505664	-0,68188	100	0,977598		
17	0	1	1	0	0,556884	0,556884	0,443116	-0,81392	0	1,256745		
18	1	0	1	1	0,487587	0,487587	0,512413	-0,71829	0	1,050917		
19	0	1	1	0	0,536642	0,536642	0,463358	-0,76926	0	1,15816		

Fonte: O autor (2022)

As colunas *Success* e *failure* tratam da variável *input* do modelo tratada na Seção anterior que discriminava se aquele cliente tinha, ou não, realizado o pagamento no mês de observação, discriminando na mesma coluna com a resposta “0” aquele que não pagou e com “1” aquele que pagou. Conforme podemos observar na Figura 18, no *output* do modelo essa variável é tratada em colunas separadas (*Success* e *Failure*), sendo a coluna *p-Obs* destinada a realizar o somatório dos sucessos (clientes que pagaram).

A coluna “Q” apresenta o *p-Pred*, que é a probabilidade de ocorrência do evento (pagamento) estimada pelo modelo para cada um dos clientes (elementos) da amostra através dos coeficientes tanto para o intercepto, célula X5, quanto para cada uma das nove variáveis explicativas: X6 a X14. Cabe destacar ainda que a coluna LL (coluna T), trata de um dos parâmetros mais importantes da regressão, que é a maximização da função de verossimilhança ou *log likelihood function* (Ver Seção 2.1.3.1). Já a coluna “U” identifica com a informação 100 os acertos do modelo, ou seja, aqueles elementos que a previsão de pagamento se confirmou considerando um dado *cutoff*. Por fim, a coluna V apresenta o *HL Stat* que se refere ao teste de Hosmer-Lemeshow já explicado na Seção 2.1.3.1. e será novamente detalhado adiante.

Para atestar a capacidade preditiva dos modelos encontrados, cada um fruto de um conjunto diferente de variáveis independentes, e assim poder comparar diversos cenários, utilizamos como referência a Tábua de Classificação e a Área Sob a Curva *ROC,* sendo ambas calculadas também pelo suplemento.

A Figura 20 apresenta a tábua de classificação gerada pelo modelo acrescida dos conceitos atinentes a cada um dos parâmetros calculados.

Figura 20 - Tábua de classificação com acréscimo dos conceitos de cada indicador

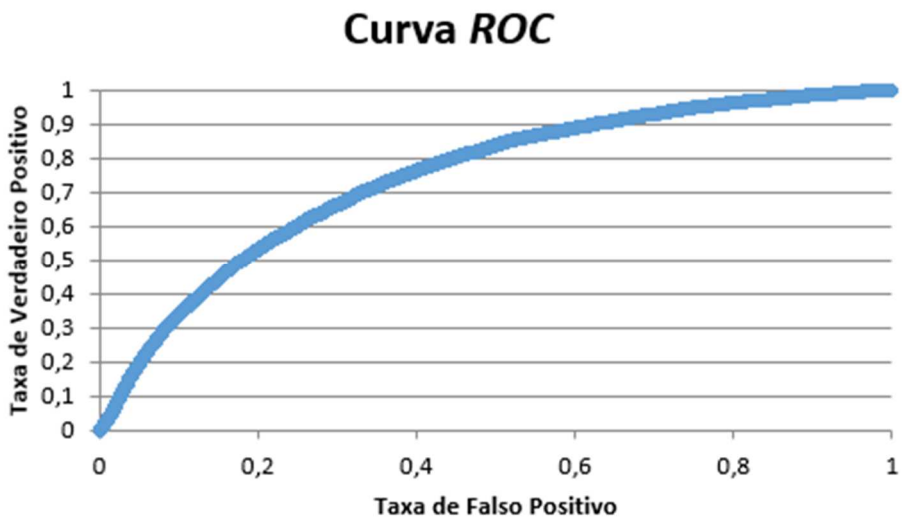
	AS	AT	AU	AV
1	Classification Table			
2				
3		Incidência Real do Evento	Incidência Real do Não-Evento	
4	Classificado como Evento	38	26	64
5	Classificado como Não-evento	417	1990	2407
6		455	2016	2471
7		Sentividade	Especificidade	EGM
8	Accuracy	0,083516484	0,987103175	0,820720356
9				
10	Cutoff	0,5		

Fonte: O autor (2022)

Como podemos perceber, a regressão que resultou na Tábua de Classificação apresentada na Figura 18, obteve uma Eficiência Geral do Modelo (EGM) de 0,8207 (ou 82,07%), para um *Cutoff* de 0,5, ou seja, classificando como evento apenas aqueles elementos da amostra que obtiveram um *p-Pred* superior a 0,5. Assim, analisando isoladamente a Tábua de classificação, se a regressão "A" possuir uma EFG superior a regressão "B", significa que "A" é preferível a "B".

Já tratamos em seções anteriores sobre a curva ROC, a qual mostra o comportamento propriamente dito do trade off entre a sensitividade e a especificidade, ao trazer no eixo das abscissas os valores de (1 - especificidade), apresentando formato convexo em relação ao ponto (0,1). Desta forma, um determinado modelo com maior área abaixo da curva ROC apresenta maior poder preditivo, combinados todos os possíveis níveis de *cutoff* de 0 a 1, sendo assim, preferível quando da comparação com outro modelo com menor área abaixo da curva.

Figura 21 - Gráfico da Curva ROC



Fonte: O autor (2022)

O suplemento também estima a Área Sob a Curva ROC (AUC – Area Under the Curve) e apresenta o gráfico da curva ROC, tornando bastante visual e intuitivo o entendimento acerca da validade do modelo.

A Figura 22 mostra a tábua ROC e os conceitos estatísticos relacionados a cada coluna.

Figura 22 - Tábua da Curva ROC

	A	B	C	D	E	F	G	H
1	ROC Table							
2								
3	*p-Pred*	*Falha*	*Acerto*	*Soma das Falhas*	*Soma dos Acertos*	*Taxa de Falso Positivo*	*Taxa de Verdadeiro Positivo*	*Área Sob a Curva*
4				0	0	1	1	0,000496032
5	1,83954E-10	1	0	1	0	0,999503968	1	0,000496032
6	3,91141E-10	1	0	2	0	0,999007937	1	0,000496032
7	2,9691E-05	1	0	3	0	0,998511905	1	0,000496032
2473	...	...	...	...	...	...	...	...
2474	0,619605848	1	0	2016	453	0	0,004395604	0
2475	0,648654309	0	1	2016	454	0	0,002197802	0
2476	0,676372801	0	1	2016	455	0	0	0
2477								0,77815825

Fonte: O autor (2022)

No modelo cuja Área Sob a Curva calculada foi de 0,77815825, conforme apresentado na Figura 21. Esse valor indica que o modelo apresentou poder discriminatório aceitável. Segundo Hosmer e Lemeshow (2000) a regra geral para avaliação do resultado da área sob a curva ROC de modelos de *Credit Scoring* é dada por:

área < 0, 7: baixa discriminação

0, 7 < área < 0, 8: discriminação aceitável

0, 8 < área < 0, 9: discriminação excelente

área > 0, 9: discriminação excepcional

Conforme já citado na Seção 3. Existem diversos testes estatísticos que visam atestar a significância do modelo. Entres eles estão o Teste X^2, o Hosmer-Lemeshow e ainda o Estatística *Z* de Wald. São realizados os cálculos desses parâmetros a cada regressão estimada, permitindo que se possa asseverar que, além da validade preditiva, vista na Seção anterior, o modelo e as variáveis selecionadas possuam significância estatística.

O conjunto de células formado a partir da célula W3 até a coluna X14, apresentam os dados do Teste X^2 para o resultado da regressão. Conforme Figura 23, constata-se que as variáveis selecionadas e o consequente resultado encontrado são relevantes ao nível de significância de 5%:

Figura 23 - Resultado do Teste X^2 calculado no modelo

LL0	-14007,6
LL1	-12432

Chi-Sq	3151,309
df	9
p-value	0
alpha	0,05
sig	yes

Fonte: O autor (2022)

O conjunto de células formado a partir da célula W16 até a coluna X20, apresentam os dados do teste de Hosmer-Lemeshow, calculados para os dados da regressão. Conforme Figura 24, para a regressão em questão, os dados foram considerados relevantes ao nível de 5%.

Figura 24 - Resultado do Hosmer-Lemeshow calculado no modelo:

Hosmer	27007,24
df	22564
p-value	1,45E-86
alpha	0,05
sig	yes

Fonte: O autor (2022)

Os resultados da estatística Z de Wald para cada uma das variáveis utilizadas na regressão são demonstrados na coluna D ao final dos resultados da regressão. Na tabela apresentada na Figura 25 é possível verificar o valor para cada um dos parâmetros (variáveis) utilizados na regressão.

Figura 25 - Resultado da estatística Z de Wald:

	A	B	C	D	E	F	G	H
22569	948	63987,09	0,264359733	0,672244	1	304	13745,92	16915,61
22570								
22571								
22572		*coeff b*	*erro padrão*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
22573	Intercept	1,024912	0,145304392	49,75255	1,74E-12	2,786849		
22574	dias_atraso	-0,00759	0,000383666	391,4342	4,03E-87	0,992438	0,991692	0,993185
22575	vr_base	3,06E-05	3,88484E-06	62,03966	3,37E-15	1,000031	1,000023	1,000038
22576	VA/SD	-3,02068	1,372162267	4,846155	0,027708	0,048768	0,003312	0,717999
22577	SD/VG	-0,53336	0,163987747	10,57853	0,001144	0,586628	0,425379	0,809002
22578	email_cliente	0,283656	0,033532574	71,55682	2,69E-17	1,327976	1,243505	1,418186
22579	prazo_restante	-0,00117	0,000423181	7,638552	0,005713	0,998831	0,998003	0,99966
22580	encargo_atraso	-0,00038	5,08821E-05	54,53201	1,53E-13	0,999624	0,999525	0,999724
22581	total_atraso	0,000348	4,28739E-05	65,77294	5,06E-16	1,000348	1,000264	1,000432
22582	divida_total	-2,9E-05	3,948E-06	53,6165	2,44E-13	0,999971	0,999963	0,999979

Fonte: O autor (2022)

Conforme dados destacados na coluna D da Figura 24, para o nível de significância de 5%, nenhum dos parâmetros estimados apresentou valor contido no intervalo entre -1,96 e 1,96; significando que todas variáveis independentes, ao nível de significância de 5%, rejeitaram a hipótese nula, ou seja, todos os parâmetros podem ser considerado estatisticamente diferentes de zero, possuindo alguma influência no comportamento da variável resposta.

Por todo o exposto nesta Seção, fica demonstrado o poder preditivo do modelo de regressão na predição dos contratos com maior chance de reversão. Para construção do modelo foi usado o suplemento do Microsoft Excel, considerado uma ferramenta poderosa para realização dos cálculos, elaboração dos testes estatísticos de validação do resultado encontrado, bem como, dos testes de performance do poder preditivo do modelo encontrado.

4.4 IMPLANTAÇÃO DO MODELO NA ORGANIZAÇÃO

Quando o modelo apresentou as primeiras respostas estatisticamente consistentes, ou seja, resultados cuja Eficiência Global do Modelo e Área Sob a Curva Roc apresentaram índices superiores a 0,7; e uma Sensitividade superior a 0,8; a ferramenta passou a ser utilizada na Centralizadora de Adimplência no processo de definição mensal das bases de clientes para acionamento pelo canal telefone no período de Março a Agosto de 2021.

A cada fechamento de mês, os coeficientes das variáveis explicativas eram atualizados através de uma nova regressão, sendo esses parâmetros utilizados para cálculo das probabilidades de pagamento de cada um dos clientes presente na base do mês seguinte, e, a partir disso, selecionava-se para a base de acionamento os clientes com maior probabilidade (*P-pred*) de recebimento. Um processo simples que era aprimorado a cada mês, comparando se os indicadores de performance estavam apresentando índices superiores aos dos meses anteriores.

Não obstante a confiabilidade nos parâmetros de Sensitividade, Especificidade, Eficiência Global do Modelo e Área Sob a Curva ROC, foi realizada uma verificação *ex-post* da validade preditiva do modelo calculando a quantidade de contratos, cujo *P-Pred* estimado era superior a 0,5 (pagamento previsto), e que efetivamente realizaram o pagamento ao fim do período da predição. Os resultados descritos a seguir foram encontrados em uma amostra de 3.130 contratos pertencentes a base do mês Junho/2021, com valor de dívida superior a R$100.000,00.

O resultado medido pelo indicador de Eficiência Global do Modelo, para um *cutoff* de 0,5 foi de 94,82%, sendo a Sensitividade encontrada de 90,09% e a Especificidade de 97,91%, conforme Tabela 12:

Tabela 12 - Tabela de Classificação e*x-post*

	Incidência Real do Evento	Incidência Real do Não Evento	Total
Classificado como Evento	1.310	35	1.345
Classificado como Não-evento	144	1.641	1.785
Total	1.454	1.676	3.130
	Sensitividade	Especificidade	EGM
	0,900962861	0,979116945	0,942812

Fonte: O autor (2022)

Em outras palavras, conforme dados da Tabela 12, dos 1.345 contratos que efetivamente realizarem seus pagamentos, 1.310 foram previstos pelo modelo, e dos 1.785 clientes finalizaram o mês inadimplentes, 1.641 foram corretamente previstos.

Embora não tenhamos conseguido isolar os benefícios do novo processo e saber com precisão o quanto a nova modelagem contribuiu para o ganho de eficiência e efetividade da unidade, a contribuição do modelo, cuja comprovação da significância estatística está detalhado nas seções anteriores, se concentrou em fornecer elementos probabilísticos e uma resposta objetiva que pôde subsidiar as decisões mais importantes dos gestores da unidade, decisões que até então eram tomadas por um processo *ad hoc*, sem formalização, usando apenas o conhecimento empírico.

Para ilustrar a importância da ferramenta, tomemos como exemplo o mês de Junho/21 já mencionado, cuja a base de contratos que possuía o benefício da Negociação Emergencial totalizava 22.584 contratos, perfazendo um valor total de R$ 1,8 bilhão, sendo que naquele mês, a capacidade operacional da unidade para acionamento dos clientes enquadrados nesta modalidade de negociação correspondia a apenas 3.300 contratos.

Antes de adoção do *Collection Score*, o critério mais comum de definição da base de acionamentos era o de selecionar os clientes com base unicamente no valor da dívida, escolhendo aqueles de maior saldo devedor. Esse procedimento possuía ineficiências tais como a de ignorar a quantidade de dias de atraso do contrato, a qual possui correlação negativa com a probabilidade de reversão do cliente. Uma análise feita com dados do mês de Dezembro/2020 da centralizadora apresentou os seguintes resultados:

- Clientes com menos de 90 dias de atraso: 78% de reversão;
- Clientes com atraso entre 91 e 250 dias de atraso: 31,90% de reversão;
- Clientes acima de 250 dias de atraso: 13,09% de reversão.

Ressaltando que a quantidade de dias de atraso é proporcional a quantidade de prestações que o cliente precisa desembolsar para colocar seu contrato em dia.

Neste sentido, se o critério de seleção da base de acionamento do mês de Junho/21 fosse unicamente o valor da dívida dos clientes, poderíamos incorrer em escolhas sem justificativas conforme Tabela 13 abaixo:

Tabela 13 - Valor base e dias de atraso

Nº contrato	Dias de atraso	Saldo Devedor
3.796	**602**	**R$ 99.983,75**
3.797	**269**	**R$ 99.977,86**
3.798	**78**	**R$ 99.975,00**
3.799	**453**	**R$ 99.971,49**
3.800	**106**	**R$ 99.969,11**
3.801	**256**	**R$ 99.967,69**
3.802	**80**	**R$ 99.964,07**
3.803	**105**	**R$ 99.956,92**
3.804	**83**	**R$ 99.953,11**
3.805	**534**	**R$ 99.946,33**
3.806	**273**	**R$ 99.945,96**
3.807	**230**	**R$ 99.938,05**
3.808	**88**	**R$ 99.934,11**
3.809	**238**	**R$ 99.927,83**
3.810	**178**	**R$ 99.926,33**
3.811	**467**	**R$ 99.923,19**
3.812	**542**	**R$ 99.911,33**
3.813	**496**	**R$ 99.910,33**
3.814	**173**	**R$ 99.893,11**
3.815	**116**	**R$ 99.878,67**

A Tabela 13 contém as informações da quantidade de dias de atraso e do saldo devedor de uma amostra 20 contratos retirados da base de 22.584 organizada por ordem decrescente de

valor do saldo devedor, os contratos apresentados na tabela ocupam o intervalo entre as posições: 3.796 e 3.815 da base total. Conforme a tabela, em um intervalo de variação do saldo devedor de pouco mais de R$100,00, (irrelevante para índice de inadimplência habitacional), verifica-se uma variação de 524 unidades na quantidade de dias de atraso (contrato posição 3.796 com 608 dias de atraso, e contrato posição 3.798 com 78 dias). Ou seja, ao se selecionar a base olhando unicamente para o valor da base, incorria-se no risco de desprezar contratos com maior probabilidade de reversão (sob a perspectiva dos dias de atraso) motivados por diferenças insignificantes do saldo devedor.

Um outro racional utilizado para definição da base de acionamento foi o de limitar a quantidade de dias de atraso, por exemplo a 250 dias, e acionar os maiores devedores cujos contratos atendessem a esse critério temporal. As questões permaneciam as mesmas, afinal, por que 250 dias e não 260, ou 240, ou 300 dias de atraso? Com isso, surgia o mesmo risco de desprezar clientes com débitos muito superiores por pequenas diferenças na quantidade de dias de atraso.

O *Collection Score* permitiu uma abordagem mais holística da base de clientes, viabilizando analisar e combinar diferentes critérios para além do saldo devedor e da quantidade de dias de atraso, como por exemplo: valor da garantia, percentual da dívida coberto pela garantia, valor da prestação, disponibilidade de e-mail, entre outros. Permitindo ao gestor selecionar os contratos com base na probabilidade de ocorrência do evento pagamento calculada para cada cliente da base, viabilizando a definição de uma base de acionamento cuja com probabilidade de reversão mais elevada.

Entre os resultados absolutos conhecidos, temos, por exemplo, que no mês de Junho de 2021, a unidade conseguiu alcançar um resultado positivo de R$68 milhões, o primeiro resultado positivo até então. Ou seja, naquele mês, a centralizadora conseguiu reverter (fazer pagar) um valor que superou em R$68 milhões a pressão do período que foi de R$1,35 bilhão.

O processo hoje está incorporado às rotinas da centralizadora. A volatilidade das condições negociais que deixam de viger e ressurgem em decorrência das fases de maior ou menor restrição social provocadas pela pandemia representam um desafio à sensibilidade do gestor, que precisa levar em consideração cenários de condição negociais semelhantes no momento de estimar o *Collection Score* de cada base. O momento atual é de aprimoramento contínuo do processo expandindo-o para outros segmentos, tais como: o segmento de recuperação de prejuízo e inadimplência do produto cartão de crédito.

5 CONSIDERAÇÕES FINAIS

A proposta desse trabalho foi a de elaborar um modelo de *collection Scoring* usando a regressão logística através do Excel, visando subsidiar o estabelecimento de ações de cobrança. Como visto, a inadimplência é um indicador central para qualquer instituição de crédito por ter correlação direta com sua solvência e, no agregado, afeta também o nível de desenvolvimento e abrangência do mercado financeiro e, como consequência, do crescimento de uma economia.

Neste sentido, visando aumentar a eficiência das ações de cobrança por parte das instituições, a maior contribuição do trabalho é a praticidade da sua aplicação, e adaptabilidade a diferentes disponibilidades de dados, além da automatização dos cálculos dos parâmetros estatísticos, tanto de criação, quanto de validação do modelo. Por outro lado, o estudo não dispôs de dados relacionados ao perfil social do cliente (eg., vínculo empregatício, nível de escolaridade, estado civil, etc.) que poderiam aumentar seu poder preditivo.

Além da indicação de agregar dados financeiros dos clientes ao estudo, será de grande contribuição agregar ao modelo algum processo de aprendizagem com algoritmos ou redes neurais que permitam simulações mais robustas, combinando os diferentes tipos de variáveis dependentes disponíveis e que permita otimizar o conjunto de variáveis independentes que apresentem o melhor poder preditivo.

Uma outra simplificação realizada neste trabalho, ainda que acidentalmente, foi assumir como restrição absoluta a quantidade de empregados disponíveis para execução da cobrança, algo que não é a realidade na maioria das empresas. Assim, não se considerou questões como: qual o tamanho ótimo da equipe, ou seja, qual a quantidade de funcionários maximiza o retorno financeiro da empresa com as ações de cobrança

REFERÊNCIAS

ANNIBAL, C. A. (2009), Inadimplência do Setor Bancário Brasileiro: Uma Avaliação de suas Medidas. Trabalhos para Discussão Brasília, n° 192, setembro 2009, pp. 1-36.

ARAÚJO, R. V., MARTINEZ, L., MOREIRA, F. A. Seleção de Clientes Para Ações de Cobrança Através de Regressão Logística na Inadimplência do Consumo de Energia Elétrica, XXI Congresso Brasileiro de Automática - CBA2016, UFES, Vitória - ES, 3 a 7 de outubro.

BALASSIANO, M., VIDAL, V. (2019), *A parcimônia com o mercado de crédito.* FGV.

BANCO CENTRAL DO BRASIL (1999), *Resolução Nº 2682.* Disponível em: https://www.bcb.gov.br/pre/normativos/res/1999/pdf/res_2682_v2_L.pdf Acesso em 08 jul 2021.

BANCO CENTRAL DO BRASIL (2019), *Relatório de Economia Bancária e Crédito*. Disponível em: https://www.bcb.gov.br/publicacoes/relatorioeconomiabancaria/REB_2019. Acesso em 08 jul. 2021.

BANCO CENTRAL DO BRASIL (2021), Estatísticas do Mercado de Crédito do Banco Central, https://www.bcb.gov.br/estatísticas/estatísticasmonetariascredito. Acesso em Março de 2021.

BANCO CENTRAL DO BRASIL (2021), Inadimplência da carteira de crédito – Total. Disponível em: https://dadosabertos.bcb.gov.br/dataset/21082-inadimplencia-da-carteira-de-credito---total. Acesso em 08 jul. 2021.

BANCO CENTRAL DO BRASIL (2021), SGS – Sistema Gerenciador de Sérias Temporais (bcb.gov.br) acesso em Março de 2021

Caixa Econômica Federal (2021), Apresentação de Resultados, https://www.caixa.gov.br/sobre-a-caixa/relacoes-com-investidores/central-resultados/Paginas/default.aspx acesso em Abril de 2021.

CARTA CONJUNTURA IPEA, nº 50, Nota Conjuntura 11, 1º Trimestre de 2021

CARVALHO, A. G. Desenvolvimento Financeiro e Crescimento Econômico. Revista Econômica do Nordeste, Fortaleza, v. 33, n. 4, out-dez. 2002.

CHAIA, A. J,. Modelos de Gestão do Risco de Crédito e Sua Aplicabilidade ao Mercado Brasileiro, São Paulo, 2003.

FÁVERO, L. P. (2015), *Análise de Dados:* Estatística e Modelagem Multivariada com Excel, SPSS e Stata, LTC., 2015.

FERNANDES, ALESSANDRO. PCLD e seus Efeitos no Resultado Contábil Final do Banco do Brasil no Exercício De 2016. Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano 03, Ed. 09, Vol. 01, pp. 90-100, Setembro de 2018.

FORTI, M. Técnicas de Machine Learning Aplicadas na Recuperação de Crédito do Mercado Brasileiro, São Paulo, 2018.

GONÇALVES, E. B. GOUVÊA, M. A., Collection Score and the opportunities for nonperforming loans Market, 2016.

GONÇALVES, M. Em Busca das Origens e Evolução da Contabilidade, Revista Mineira de Contabilidade, p. 23 – 30, Jun 2016.

GOUVÊA, M. A.; GONÇALVES, E. B.; MANTOVANI, D. M. N. (2012) Aplicação de regressão logística e algoritmos genéticos na análise de risco de crédito. *Revista Universo Contábil*, 8 (2), pp. 84-102.

HILLIER, Frederick S.; LIEBERMAN, Gerald J. Introduction to operations research. 7th. ed. New York, NY: McGraw-Hill, 2001.

HOSMER, D, W., LEMESHOW, S. (2000), *Applied Logistic Regression*, 2nd ed. New York.

JONATHAN N. CROOK, DAVID B. EDELMAN, LYN C. THOMAS (2007) Recent developments in consumer credit risk assessment. *European Journal of Operational Research*,183 (3), pp. 1447-1465. https://doi.org/10.1016/j.ejor.2006.09.100.

KENNEDY, K., MAC NAMEE, B., DELANY, S. J. O`SULLIVAN, M., WATSON, N., A Window of Opportunity: Assessing Behavioural Scoring, School of Computing, Dublin Institute of Technology, Ireland, Julho, 2012.

LOPES, L. S., Definição de um Modelo de Cobrança (Collection Score) Utilizando Regressão Logística Multinomial, Porto Alegre, Julho de 2004.

LOPES, M. G., CIRIBELI, J. P., MASSARDI, W. O., MENDES, W. A., Análise dos Indicadores de Inadimplência nas Linhas de Crédito para Pessoa Física: Um Estudo Utilizando Modelo de Regressão Logística, Revista do CEPE. Santa Cruz do Sul, n. 46, p. 75-90, jul./dez. 2017.

MOURA, G. M., Regressão Logística aplicada a análise de risco de crédito, 2018.

NETO, A. A. A., CARMONA, C. U. M., Modelagem do Risco de Crédito: Um Estudo do Segmento de Pessoas Físicas em um Banco de Varejo, Revista Eletrônica de Administração, Edição 40 - jul/ago 2004

OLIVEIRA, M. S., MENEZES, M. F. R., FARIAS, F. F., SILVA, J. R. S., SOUZA, R. P. Ferramenta de Recuperação de Clientes: Collection Score. Desenvolvida com a Regressão Logística Binária, XIV Escola de Séries Temporais e Econometria, Gramado-RS, 01-05 Agosto 2011

RODRIGUES, P. H. C. NETO, W. J. FERREIRA, R. M. Da História do Crédito: Da Mesopotâmia aos Médici e a Expansão do Modelo de Negócio Bancário, Revista Jurídica, Ano XIV, n. 23, v2, Anápolis/GO, Unievangélica Jan. – jun, 2014

SANTOS, M, MOTA, C.M., KRAMER, R., LIMA, S. Collection Scoring como ferramenta de definição de estratégia de cobrança de clientes em situação de inadimplência bancária. In: INSID, 2021

SILVA, E. N., JÚNIOR, S. S. P. Sistema Financeiro e Crescimento Econômico: Uma Aplicação de Regressão Quantílica, Econ. Aplic., São Paulo, V. 10, N. 3, P. 425-442, Julho-Setembro 2006

SOUSA, A. F., CHAIA, A. J., Política de Crédito: Uma Análise Qualitativa dos Processos Em Empresas, Caderno de Pesquisas em Administração, São Paulo, v. 07, nº 3, julho/setembro 2000.

THOMAS, L. C., MUES, C., MATUSZYK, A., Modelling LGD For Unsecured Personal Loans: Decision Tree Approach, The Journal of the Operational Research Society, Vol. 61, No. 3, Consumer CreditRisk Modelling; Transportation, Logistics and the Environment, Mar., 2010.

World Bank (2021), https://data.worldbank.org/.

APÊNDICE A - FREQUÊNCIA POR SALDO DEVEDOR DOS CONTRATOS DA BASE DE UM MÊS TÍPICO

Bloco	*Freqüência*	*% Cumulativo*
R$ 0,66	1	0,00%
R$ 15.491,66	1090	3,10%
R$ 30.982,66	2657	10,66%
R$ 46.473,66	4495	23,43%
R$ 61.964,66	5355	38,66%
R$ 77.455,66	7156	59,00%
R$ 92.946,66	6165	76,53%
R$ 108.437,66	4026	87,97%
R$ 123.928,66	1840	93,21%
R$ 139.419,66	732	95,29%
R$ 154.910,66	384	96,38%
R$ 170.401,66	242	97,07%
R$ 185.892,66	179	97,57%
R$ 201.383,66	148	98,00%
R$ 216.874,66	119	98,33%
R$ 232.365,66	95	98,60%
R$ 247.856,66	49	98,74%
R$ 263.347,66	60	98,91%
R$ 278.838,66	56	99,07%
R$ 294.329,66	44	99,20%
R$ 309.820,66	45	99,33%
R$ 325.311,66	30	99,41%
R$ 340.802,66	21	99,47%
R$ 356.293,66	17	99,52%
R$ 371.784,66	14	99,56%
R$ 387.275,66	18	99,61%
R$ 402.766,66	19	99,66%
R$ 418.257,66	11	99,70%

R$	433.748,66	12	99,73%
R$	449.239,66	14	99,77%
R$	464.730,66	7	99,79%
R$	480.221,66	4	99,80%
R$	495.712,66	4	99,81%
R$	511.203,66	9	99,84%
R$	526.694,66	3	99,85%
R$	542.185,66	1	99,85%
R$	557.676,66	7	99,87%
R$	573.167,66	2	99,87%
R$	588.658,66	5	99,89%
R$	604.149,66	3	99,90%
R$	619.640,66	4	99,91%
R$	635.131,66	2	99,91%
R$	650.622,66	0	99,91%
R$	666.113,66	2	99,92%
R$	681.604,66	2	99,93%
R$	697.095,66	1	99,93%
R$	712.586,66	1	99,93%
R$	728.077,67	1	99,93%
R$	743.568,67	2	99,94%
R$	759.059,67	0	99,94%
R$	774.550,67	0	99,94%
R$	790.041,67	1	99,94%
R$	805.532,67	2	99,95%
R$	821.023,67	0	99,95%
R$	836.514,67	0	99,95%
R$	852.005,67	0	99,95%
R$	867.496,67	1	99,95%
R$	882.987,67	1	99,95%
R$	898.478,67	0	99,95%
R$	913.969,67	3	99,96%
R$	929.460,67	2	99,97%

R$ 944.951,67	1	99,97%
R$ 960.442,67	0	99,97%
R$ 975.933,67	0	99,97%
R$ 991.424,67	0	99,97%
R$ 1.006.915,67	0	99,97%
R$ 1.022.406,67	1	99,97%
R$ 1.037.897,67	1	99,98%
R$ 1.053.388,67	1	99,98%
R$ 1.068.879,67	0	99,98%
R$ 1.084.370,67	0	99,98%
R$ 1.099.861,67	0	99,98%
R$ 1.115.352,67	0	99,98%
R$ 1.130.843,67	0	99,98%
R$ 1.146.334,67	0	99,98%
R$ 1.161.825,67	0	99,98%
R$ 1.177.316,67	0	99,98%
R$ 1.192.807,67	1	99,98%
R$ 1.208.298,67	0	99,98%
R$ 1.223.789,67	0	99,98%
R$ 1.239.280,67	0	99,98%
R$ 1.254.771,67	0	99,98%
R$ 1.270.262,67	1	99,99%
R$ 1.285.753,67	0	99,99%
R$ 1.301.244,67	0	99,99%
R$ 1.316.735,67	0	99,99%
R$ 1.332.226,67	0	99,99%
R$ 1.347.717,67	0	99,99%
R$ 1.363.208,67	0	99,99%
R$ 1.378.699,67	0	99,99%
R$ 1.394.190,67	0	99,99%
R$ 1.409.681,67	0	99,99%
R$ 1.425.172,67	1	99,99%
R$ 1.440.663,67	0	99,99%

R$ 1.456.154,67	0	99,99%
R$ 1.471.645,67	0	99,99%
R$ 1.487.136,67	0	99,99%
R$ 1.502.627,67	0	99,99%
R$ 1.518.118,67	0	99,99%
R$ 1.533.609,67	0	99,99%
R$ 1.549.100,67	0	99,99%
R$ 1.564.591,67	1	99,99%
R$ 1.580.082,67	0	99,99%
R$ 1.595.573,67	1	99,99%
R$ 1.611.064,67	0	99,99%
R$ 1.626.555,67	0	99,99%
R$ 1.642.046,67	0	99,99%
R$ 1.657.537,67	0	99,99%
R$ 1.673.028,67	0	99,99%
R$ 1.688.519,67	0	99,99%
R$ 1.704.010,67	0	99,99%
R$ 1.719.501,67	0	99,99%
R$ 1.734.992,67	0	99,99%
R$ 1.750.483,67	0	99,99%
R$ 1.765.974,67	0	99,99%
R$ 1.781.465,67	0	99,99%
R$ 1.796.956,67	0	99,99%
R$ 1.812.447,67	0	99,99%
R$ 1.827.938,67	0	99,99%
R$ 1.843.429,67	0	99,99%
R$ 1.858.920,67	0	99,99%
R$ 1.874.411,67	0	99,99%
R$ 1.889.902,67	0	99,99%
R$ 1.905.393,67	0	99,99%
R$ 1.920.884,67	0	99,99%
R$ 1.936.375,67	0	99,99%
R$ 1.951.866,67	0	99,99%

R$ 1.967.357,67	0	99,99%
R$ 1.982.848,67	0	99,99%
R$ 1.998.339,67	0	99,99%
R$ 2.013.830,67	0	99,99%
R$ 2.029.321,67	0	99,99%
R$ 2.044.812,67	0	99,99%
R$ 2.060.303,67	0	99,99%
R$ 2.075.794,67	0	99,99%
R$ 2.091.285,67	0	99,99%
R$ 2.106.776,67	0	99,99%
R$ 2.122.267,67	0	99,99%
R$ 2.137.758,67	0	99,99%
R$ 2.153.249,67	0	99,99%
R$ 2.168.740,67	0	99,99%
R$ 2.184.231,68	0	99,99%
R$ 2.199.722,68	0	99,99%
R$ 2.215.213,68	0	99,99%
R$ 2.230.704,68	0	99,99%
R$ 2.246.195,68	0	99,99%
R$ 2.261.686,68	0	99,99%
R$ 2.277.177,68	0	99,99%
R$ 2.292.668,68	0	99,99%
R$ 2.308.159,68	0	99,99%
R$ 2.323.650,68	0	99,99%
R$ 2.339.141,68	0	99,99%
R$ 2.354.632,68	0	99,99%
R$ 2.370.123,68	0	99,99%
R$ 2.385.614,68	0	99,99%
R$ 2.401.105,68	0	99,99%
R$ 2.416.596,68	0	99,99%
R$ 2.432.087,68	0	99,99%
R$ 2.447.578,68	0	99,99%
R$ 2.463.069,68	0	99,99%

R$ 2.478.560,68	0	99,99%
R$ 2.494.051,68	0	99,99%
R$ 2.509.542,68	0	99,99%
R$ 2.525.033,68	0	99,99%
R$ 2.540.524,68	0	99,99%
R$ 2.556.015,68	0	99,99%
R$ 2.571.506,68	0	99,99%
R$ 2.586.997,68	0	99,99%
R$ 2.602.488,68	0	99,99%
R$ 2.617.979,68	0	99,99%
R$ 2.633.470,68	0	99,99%
R$ 2.648.961,68	0	99,99%
R$ 2.664.452,68	0	99,99%
R$ 2.679.943,68	0	99,99%
R$ 2.695.434,68	0	99,99%
R$ 2.710.925,68	0	99,99%
R$ 2.726.416,68	0	99,99%
R$ 2.741.907,68	0	99,99%
R$ 2.757.398,68	0	99,99%
R$ 2.772.889,68	0	99,99%
R$ 2.788.380,68	0	99,99%
R$ 2.803.871,68	0	99,99%
R$ 2.819.362,68	0	99,99%
R$ 2.834.853,68	1	100,00%
R$ 2.850.344,68	0	100,00%
R$ 2.865.835,68	0	100,00%
R$ 2.881.326,68	0	100,00%
Mais	1	100,00%

APÊNDICE B - FREQUÊNCIA DOS CONTRATOS POR DIAS DE ATRASO EM UM MÊS TÍPICO

Bloco	*Freqüência*	*%Cumulativo*
61	32	0,09%
78	5583	15,95%
95	7580	37,48%
111	3279	46,79%
128	1711	51,65%
145	1814	56,80%
162	1481	61,01%
179	1081	64,08%
195	751	66,21%
212	730	68,29%
229	521	69,77%
246	942	72,44%
263	457	73,74%
279	511	75,19%
296	522	76,68%
313	572	78,30%
330	518	79,77%
347	513	81,23%
364	590	82,90%
380	693	84,87%
397	745	86,99%
414	552	88,56%
431	513	90,01%
448	522	91,50%
464	543	93,04%
481	368	94,08%
498	306	94,95%
515	374	96,01%
532	334	96,96%
548	437	98,20%

565	301	99,06%
582	300	99,91%
599	1	99,91%
616	0	99,91%
632	2	99,92%
649	0	99,92%
666	1	99,92%
683	0	99,92%
700	0	99,92%
716	1	99,93%
733	0	99,93%
750	1	99,93%
767	0	99,93%
784	0	99,93%
801	1	99,93%
817	0	99,93%
834	0	99,93%
851	0	99,93%
868	0	99,93%
885	1	99,93%
901	0	99,93%
918	0	99,93%
935	0	99,93%
952	0	99,93%
969	0	99,93%
985	0	99,93%
1002	0	99,93%
1019	0	99,93%
1036	0	99,93%
1053	0	99,93%
1069	0	99,93%
1086	0	99,93%
1103	0	99,93%

1120	1	99,94%
1137	0	99,94%
1153	0	99,94%
1170	0	99,94%
1187	0	99,94%
1204	1	99,94%
1221	0	99,94%
1238	0	99,94%
1254	0	99,94%
1271	0	99,94%
1288	0	99,94%
1305	0	99,94%
1322	0	99,94%
1338	0	99,94%
1355	0	99,94%
1372	0	99,94%
1389	0	99,94%
1406	0	99,94%
1422	0	99,94%
1439	0	99,94%
1456	0	99,94%
1473	0	99,94%
1490	0	99,94%
1506	0	99,94%
1523	1	99,94%
1540	0	99,94%
1557	0	99,94%
1574	0	99,94%
1590	0	99,94%
1607	1	99,95%
1624	1	99,95%
1641	1	99,95%
1658	0	99,95%

1675	0	99,95%
1691	0	99,95%
1708	0	99,95%
1725	0	99,95%
1742	0	99,95%
1759	0	99,95%
1775	0	99,95%
1792	0	99,95%
1809	0	99,95%
1826	0	99,95%
1843	0	99,95%
1859	0	99,95%
1876	0	99,95%
1893	0	99,95%
1910	0	99,95%
1927	0	99,95%
1943	0	99,95%
1960	0	99,95%
1977	0	99,95%
1994	0	99,95%
2011	0	99,95%
2027	1	99,95%
2044	1	99,96%
2061	0	99,96%
2078	0	99,96%
2095	0	99,96%
2112	0	99,96%
2128	0	99,96%
2145	1	99,96%
2162	1	99,96%
2179	0	99,96%
2196	1	99,97%
2212	1	99,97%

2229	0	99,97%
2246	0	99,97%
2263	4	99,98%
2280	0	99,98%
2296	2	99,99%
2313	0	99,99%
2330	1	99,99%
2347	1	99,99%
2364	0	99,99%
2380	0	99,99%
2397	0	99,99%
2414	0	99,99%
2431	0	99,99%
2448	0	99,99%
2464	0	99,99%
2481	0	99,99%
2498	0	99,99%
2515	0	99,99%
2532	1	99,99%
2549	0	99,99%
2565	0	99,99%
2582	0	99,99%
2599	1	100,00%
2616	0	100,00%
2633	0	100,00%
2649	0	100,00%
2666	0	100,00%
2683	0	100,00%
2700	0	100,00%
2717	0	100,00%
2733	0	100,00%
2750	0	100,00%
2767	0	100,00%

2784	0	100,00%
2801	0	100,00%
2817	0	100,00%
2834	0	100,00%
2851	0	100,00%
2868	0	100,00%
2885	0	100,00%
2901	0	100,00%
2918	0	100,00%
2935	0	100,00%
2952	0	100,00%
2969	0	100,00%
2986	0	100,00%
3002	0	100,00%
3019	0	100,00%
3036	0	100,00%
3053	0	100,00%
3070	0	100,00%
3086	0	100,00%
3103	0	100,00%
3120	0	100,00%
3137	0	100,00%
3154	0	100,00%
3170	0	100,00%
3187	0	100,00%
Mais	1	100,00%

ANEXO A - PASSO A PASSO DE NAVEGAÇÃO DO SUPLEMENTO DO EXCEL

Figura A1 - Acionamento do suplemento

Figura A2 - Seleção da Ferramenta de Regressão Logística:

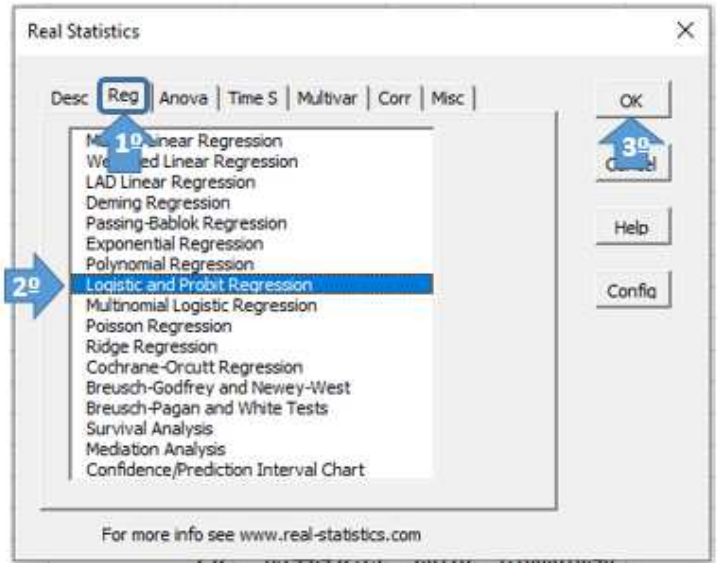
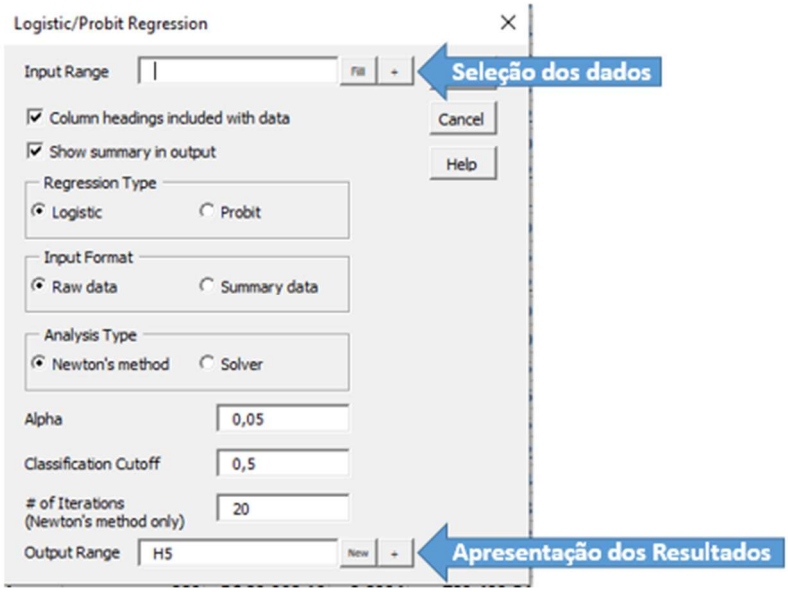


Figura A3 - Seleção dos dados:



Os demais campos da figura 2 já vêm preenchidos por *default* com os dados acima indicados. Merecem destaque especial os itens *Alpha*, cujo preenchimento default é 0,05 e que implica em um nível de significância de 5% para os testes estatísticos.