UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Lucas Tonegi

MODELO COM FRAÇÃO DE INADIMPLENTES: UMA APLICAÇÃO A DADOS FINANCEIROS

CURITIBA
2017

UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Lucas Tonegi

MODELO COM FRAÇÃO DE INADIMPLENTES: UMA APLICAÇÃO A DADOS FINANCEIROS

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Estatística do Setor de Ciências Exatas da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientadora: Profa. Dra. Suely Ruiz Giolo

CURITIBA
2017

AGRADECIMENTOS

Quero agradecer, em primeiro lugar, a Deus, pela força e coragem durante toda esta longa caminhada.

À minha família, por sua capacidade de acreditar e investir em mim. Minha mãe, pai e irmão. Mãe, seu cuidado e dedicação me deram, em alguns momentos, a esperança para seguir. Pai, sua presença significou segurança e certeza de que não estou sozinho nessa caminhada.

Aos meus tios e tias, que contribuíram muito para o meu crescimento como pessoa, Lourdes S. Tomita, Ismael Albuquerque e Cristine Y. Tomita.

À professora Suely R. Giolo por seus ensinamentos, paciência e confiança ao longo das supervisões das minhas atividades na Universidade Federal do Paraná.

Ao professor José L. Padilha pela disponibilidade em ser membro da banca deste trabalho, foi um prazer tê-lo na banca examinadora.

Ao professor Raul Y. Matsushita da Universidade de Brasília pelo apoio e disponibilidade do *software* estatístico.

À minha namorada Izabelle pelo companheirismo e compreensão nos momentos em que fui ausente durante a elaboração deste trabalho.

À equipe de modelagem da instituição em que trabalho atualmente, em especial ao Henequi e Éder, que sem o apoio deles não teria conseguido realizar este trabalho, em especial pelo banco de dados fornecido.

Aos meus amigos Luis Henrique e Lucas Eduardo, amigos da faculdade e da vida, por me ajudar e apoiar quando eu pensei em desistir da graduação deste curso.

“Não diga nada, apenas faça.

As pessoas se surpreendem com mudanças inesperadas

e sem avisos prévios.”

(*Caio F. de Abreu*)

RESUMO

Este trabalho tem como foco o estudo do tempo até a ocorrência do pagamento de dívidas de clientes que já se encontram em atraso. Buscou-se a identificação de possíveis fatores (covariáveis) que afetam este tempo de pagamento e, também, a identificação e discriminação de bons e maus clientes. O banco de dados, fornecido por uma instituição financeira, era inicialmente composto por aproximadamente um milhão de clientes, com as informações dispostas em oitocentas e noventa e uma variáveis. Deste banco de dados, foi extraída uma amostra aleatória de vinte e seis mil clientes. Os métodos utilizados para a análise dos dados foram: o modelo de regressão logística e o modelo de mistura com fração de inadimplentes, em que apenas três variáveis apresentaram efeito significativo. Para a escolha dos melhores modelos e verificação da adequação dos mesmos aos dados, foram utilizados o critério de informação de Akaike e área sob a curva ROC. Os dois modelos ajustados apresentaram boa adequação aos dados fornecidos pela instituição financeira, porém o modelo de mistura, em relação ao modelo logístico, acabou trazendo um ganho de informação quanto à estimação do tempo até o pagamento das dívidas. Sendo assim, o modelo de mistura se mostrou como uma boa alternativa para a elaboração de um novo *Collection Score*, assim como para a elaboração de estratégias de cobrança mais específicas.

Palavras-chave: Análise de sobrevivência; Cobrança; *Collection Score*; Dados financeiros; Fração de inadimplentes; Modelo de mistura; Regressão Logística

Sumário

1 INTRODUÇÃO

A crise econômica brasileira tem afetado cada vez mais a população no ano de 2017. São vários os fatores que geraram essa crise econômica, sendo a crise política um dos principais motivos para este momento conturbado. No primeiro trimestre de 2017, a taxa de desemprego no Brasil atingiu 13,7%, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE). E também uma alta de 1,3% na comparação com o trimestre anterior, sendo a maior taxa de desocupação no País da série histórica do indicador iniciada em 2012, segundo os dados da Pesquisa Nacional por Amostra de Domicílios (Pnad). No segundo trimestre de 2017, esta taxa caiu, segundo o IBGE. Neste período, o índice de desocupação ficou em 13%, o equivalente a 13,4 milhões de pessoas desempregadas. A redução é de 0,7 ponto percentual em relação ao primeiro trimestre de 2017. Quando comparada com o 2º trimestre de 2016 (11,3%), houve aumento de 1,7 ponto percentual. No terceiro trimestre do ano de 2017 tivemos uma queda de 4,8% em relação ao trimestre anterior, chegando a uma taxa de desemprego de 12,6%.

Apesar da melhora na taxa de desemprego no segundo semestre de 2017, o Brasil ainda tem problemas de estagnação da economia, alta na taxa de inflação, desvalorização da moeda, dentre outros. É de se esperar, cada vez mais, o endividamento das pessoas.

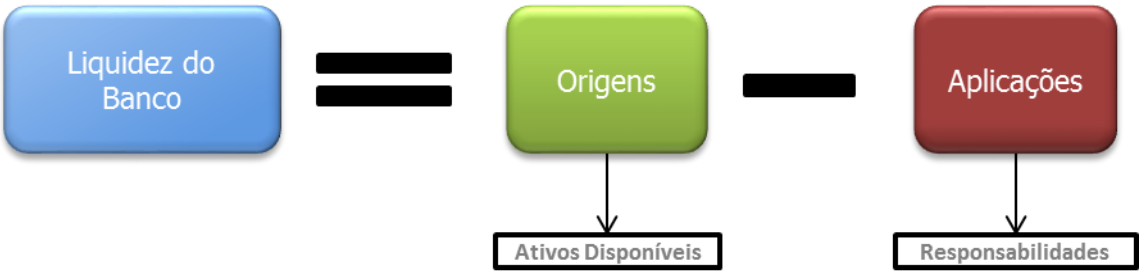
O endividamento da população é algo que afeta diretamente as empresas que hoje atuam no mercado financeiro tais como: bancos, financeiras, seguradoras, operadores de crédito etc. Para entender melhor como o endividamento afeta estas empresas, é necessário entender o funcionamento de um banco e o ciclo de crédito que será explicado a seguir.

Bancos são instituições que trabalham com dinheiro, quer seja de terceiros ou seu próprio investimento, sendo que em grande parte com o dinheiro de terceiros (clientes). Devido a tal fato, os bancos oferecem vários tipos de serviços: cartão de crédito, conta corrente, conta poupança, crédito imobiliário, financiamento de automóveis, dentre outros. Os bancos têm grande importância na economia de um País, pois através dos créditos emprestados gera um aumento no capital circulante e, consequentemente, o aumento da renda da população e o aumento de empregos na sociedade, além de facilitar as transações comerciais.

Uma forma dos bancos terem lucro se dá por meio dos empréstimos (créditos oferecidos), consequência da diferença entre as taxas de juros pagas e cobradas. Outra forma é através dos ativos que se encontram disponíveis nas contas dos clientes, geralmente

aplicados numa operação de curto prazo chamada de *Overnight*, que consiste em aplicar a liquidez de um banco no final de um dia (dado que a liquidez foi positiva) em títulos do governo federal (BACEN), resgatando-a na manhã do próximo dia útil. Com isso, os valores aplicados voltam corrigidos de acordo com os juros equivalente a um dia da taxa básica Selic, como mostra a Figura 1.

Figura 1 – Representação da definição de liquidez de um banco

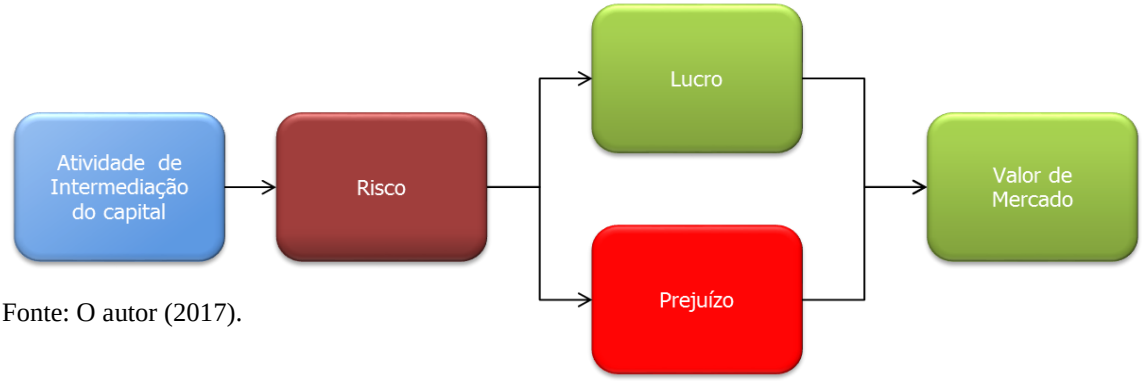


Fonte: O autor (2017).

Porém, os bancos não têm a liberdade de aplicar todo o dinheiro dos correntistas, pois são obrigados a depositar uma parte no Banco Central, a fim de garantir que o banco não fique sem reservas e cumpra com suas obrigações, além de garantir a Provisão para Devedores Duvidosos (PDD), que será discutida adiante.

Nota-se que toda operação tem um risco, definido como a possibilidade da perda resultante da incerteza quanto ao recebimento de valores pactuados com tomadores de empréstimos, contratantes de contratos ou emissão de títulos. E é exatamente por este risco assumido que todo empréstimo está sujeito a juros, um valor a ser pago para a instituição financeira em troca deste risco. Podendo assim, caso o cliente pague seu empréstimo em dia, gerar um lucro, e caso não pague suas dívidas, gerar um prejuízo, conforme mostra a Figura 2.

Figura 2 – Representação do fluxo da atividade de intermediação do capital



Fonte: O autor (2017).

Inicialmente, divide-se o ciclo de crédito em quatro partes:

1. Planejamento do produto
2. Originação ou Concessão de crédito
3. Manutenção dos produtos
4. Cobrança (área que é afetada diretamente pela crise).

O planejamento do produto é o início deste ciclo. Nesta etapa, é necessário definir como serão ofertados os produtos (já citado anteriormente). A segunda etapa deste ciclo, que consiste na concessão de crédito, é a etapa em que entra o chamado *Credit Score*, que consiste em um escore obtido por meio de um modelo estatístico (usualmente o logístico) para cada cliente que chega em uma agência bancária para a solicitação de um produto (esse escore seria a probabilidade de um cliente ser um bom pagador). A terceira etapa, consiste na manutenção destes produtos, em que é utilizado o chamado *Behavior Score*, um escore que descreve o comportamento deste cliente dentro da instituição financeira. Caso este cliente solicite um novo produto, ele possivelmente será avaliado por este escore de comportamento. E por fim, clientes que se tornam maus pagadores (inadimplentes) chegam na última etapa, a cobrança. Quando o cliente já está em cobrança, é gerado um novo score chamado *Collection Score* (também obtido, usualmente, a partir de um modelo de regressão logística), que consiste em uma estimativa da probabilidade do pagamento da dívida já em atraso.

A avaliação da capacidade de crédito das contrapartes atuais e futuras em operações de crédito é fundamental no negócio bancário, em particular a estimativa da propensão dos clientes falharem com suas obrigações financeiras no devido tempo. Esta medida pode ser avaliada em termos probabilísticos sobre um tempo pré-definido, condicional às características observáveis do devedor. A implementação de métodos estatísticos confiáveis para medir e prever essas probabilidades implica a consideração de um período de observação e avaliação de suas características. Em outras palavras, a identificação de bons clientes implica no monitoramento de cada cliente devedor ao longo do tempo e na identificação dos padrões de uma transição de mau para um bom pagador.

Atualmente, grande parte das instituições financeiras no Brasil, utilizam a regressão logística para determinar a probabilidade de pagamento de um determinado cliente. Neste trabalho, utilizaremos a técnica de análise de sobrevivência como uma técnica alternativa à regressão logística.

A análise de sobrevivência tem, em geral, o interesse em estudar o tempo até a ocorrência de um determinado evento. Sendo assim, ela está sendo cada vez mais aplicada em diversas áreas de pesquisa, tal como na Medicina, em que o interesse pode ser o de estudar o tempo até a recidiva de uma doença após o seu tratamento, e na indústria, em que o interesse pode estar no estudo do tempo até um dispositivo eletrônico parar de funcionar.

Na área financeira, há interesse dos bancos em estudar o tempo em que um cliente já devedor venha a pagar sua dívida. Esse estudo é extremamente relevante para se fazer uma previsão de devedores duvidosos (reserva que deve ser depositada para o Banco Central de todos os inadimplentes) e também, em específico, para a venda de carteiras.

Dado o momento de crise e instabilidade econômica atual, este trabalho tem como foco o estudo do tempo até a ocorrência do pagamento de dívidas de clientes que já se encontram em atraso, buscando a identificação de possíveis fatores (covariáveis) que afetam este tempo de pagamento, chegando em um método alternativo para o desenvolvimento do *Collection Score*.

2 REVISÃO DE LITERATURA

Este capítulo faz uma breve descrição sobre inadimplência e cobrança utilizadas nas instituições financeiras que trabalham com crédito ou prestação de serviços. A inadimplência pode ser definida como a incapacidade de uma pessoa física ou empresa quitar suas dívidas no valor, especificidade e data do vencimento. Já a cobrança é um processo para recuperação do crédito que foi tomado. Ocorre quando uma venda é realizada a prazo e o recebimento não ocorre dentro do prazo estabelecido ou tolerável.

2.1 Inadimplência

A inadimplência é um fenômeno em que não é possível efetuar o pagamento de suas dívidas. Segundo Mariani (2008), “o cliente ao realizar uma compra de produtos ou serviços a prazo, a empresa está concedendo crédito no qual pode ocasionar o não recebimento desta compra de serviços ou produto, caso este evento ocorra, este cliente se encontra em inadimplência, também chamado de cliente inadimplente.” Para Hanrejszkow e Stromberg (2013), “a inadimplência prejudica tanto credores quanto tomadores. Quando, por exemplo, uma instituição financeira não recebe o capital emprestado, este valor é pago pelos outros tomadores através de taxas de juros maiores.”

Como citado anteriormente, o momento de crise e instabilidade econômica que vivemos em 2017 torna inevitável o aumento da inadimplência por grande parte da população brasileira. Em setembro de 2017, segundo a Confederação Nacional do Comércio de Bens, Serviços e Turismo (CNC), a inadimplência atingiu 10,3% das famílias, o maior patamar da série histórica (iniciada em janeiro de 2010), antes 10,1% em agosto de 2017 e 9,6% em setembro de 2016.

Para grandes bancos e financeiras que têm uma grande concentração do capital alocado em operações de créditos, a inadimplência é claramente inevitável. Segundo a Pesquisa Nacional de Endividamento e Inadimplência do Consumidor (Peic Nacional), o cartão de crédito permanece como a principal forma de endividamento, atingindo 76,4% das famílias que possuem dívidas; seguido dos carnês (16,2%) e crédito pessoal (10,3%).

Segundo a empresa Serasa Experian (2017), o principal motivo declarado pelos inadimplentes para o não pagamento de seus compromissos é a perda do emprego, com 26,3% (percentual que aumenta para 27,4% entre as classes C, D e E), seguido da diminuição da renda (14,2%), da falta de controle financeiro (11,0%) e do empréstimo do

nome para terceiros (5,5%). Ainda, os produtos mais frequentes, dentre os que têm determinada conta a pagar, são: parcelas em cartão de loja; empréstimo em bancos ou financeiras; parcelas do cartão de crédito; e crédito pessoal, todos eles relacionados ao serviço financeiro bancário.

2.2 Cobrança

A cobrança é o ato de cobrar e receber o que é devido, readquirir, recuperar (FERREIRA, 1998). A cobrança se dá devido à concessão de crédito a prazo aos clientes, gerando, assim, os valores a receber. Ou então, os valores a receber com os montantes devidos à firma, provenientes de venda de mercadorias ou serviços no curso ordinário dos negócios (CHERRY, 1976).

Leoni e Leoni (1997), dizem que "a cobrança é uma função importantíssima em qualquer organização empresarial, pois, afinal, é o retorno do dinheiro ou do capital investido". Para Campos Júnior (2003), "não existem milagres na recuperação de crédito. O que temos como aliadas são formas de relacionamentos eficazes que podem se transformar em retorno persistente. Portanto, considere o devedor um potencial cliente, atenda-o, discutindo seus anseios e angústias, facilite sua vida e principalmente, não o penalize. Não o receba no porão da sua pior filial ou trate-o como marginal. Caso insista nessa "estratégia do medo", esteja certo de que, ao se recuperar, a primeira placa que ele irá avistar é do seu concorrente"

A política de cobrança deve ser implementada em conjunto com a política de crédito. A concessão não deve ser facilitada demasiadamente para, posteriormente, ter de aplicar rigidez na cobrança, ou vice-versa. Se já for esperada a dificuldade de cobrança no ato da concessão do crédito a determinados clientes, a avaliação do crédito deverá ser mais rigorosa (HOJI, 2003).

Para Pereira (1998), há três tipos de políticas distintas:

a) política de crédito rígido: é praticada por instituições financeiras e por bancos;
b) política de crédito liberal: é praticada por pequenos estabelecimentos comerciais, em que a compra é anotada em caderno, o comprador não assina documento algum e os pagamentos parcelados não têm valor fixo e não se exigem garantias;

c) política de crédito utilizável: tem suas normas e regras, mas a compra é ajustada ao poder aquisitivo do cliente. É o sistema de crediário mais usado no comércio lojista. Nele, o setor de Crediário é orientado para facilitar a venda, através das seguintes opções: - aumento do plano de pagamento (quantidade de prestações, substituição da mercadoria por uma de menor preço, diminuição do volume de produtos, limitação do crédito com base na renda familiar, exigência de um avalista e exigência de uma entrada, para diminuir o valor da prestação.

As políticas definem ações sequenciais para a área de cobrança, definidas como "régua de cobrança", que através do *Collection Score* são segmentadas em níveis de risco de acordo com a probabilidade de pagamento. Nesta etapa, também temos a elaboração de uma amostra teste e uma amostra controle definidas como estratégias "campeã e desafiantes" em que os clientes são distribuídos de forma aleatória para cada estratégia. Para a estratégia campeã (a que tem o melhor resultado custo/benefício), os tratamentos são mantidos os mesmos. Em contrapartida, para as estratégias desafiantes, testamos tratamentos alternativos e que possamos comparar com a estratégia campeã futuramente, comparando a proporção de bons pagadores de cada estratégia.

Basicamente, os níveis de risco são definidos de acordo com o *Collection Score* e a proporção empírica de clientes bons dividido pelo total de clientes da carteira, conforme a expressão a seguir

$$\%Bom = \frac{Total\ de\ clientes\ bons}{Total\ de\ clientes\ do\ portfólio}.$$

É razoável admitir que quanto melhor o escore, maior deve ser a proporção de bons do portfólio. Após a definição dos níveis de risco, devemos elaborar a régua de cobrança com diferentes períodos de envios aos *call centers*. Nesse contexto, são usualmente aplicados modelos de regressão logística para estimar a probabilidade de um cliente inadimplente vir a realizar um pagamento em um determinado período de tempo. Contudo, uma forma alternativa para a elaboração do "*Collection Score*", o modelo de mistura com fração de cura (QUIDIM, 2005; TOMAZELA et al., 2007; GRANZOTTO et al., 2010), vem ganhando destaque devido ao ganho de informação, pois além da probabilidade de pagamento, se tem também informações relevantes sobre o comportamento dos clientes durante o período de tempo observado.

A cobrança pode ser segmentada em três grandes fases: cobrança interna, cobrança terceirizada e venda de carteiras. Essas fases são tratadas separadamente a seguir.

2.3 Cobrança interna

A cobrança interna preza por um atendimento de excelência. Nesta etapa, ainda há uma relação com o cliente, pois a cobrança interna normalmente ocorre até, em média, sessenta dias em atraso, sendo que a própria empresa se responsabiliza em entrar em contato com os inadimplentes tendo, assim, um maior controle sobre estas operações.

A terceirização do serviço de cobrança acarreta um custo com comissão paga as empresas terceirizadas. Essas comissões variam de acordo com a faixa de atraso, quanto maior a faixa de atraso, maior a comissão paga à empresa para recuperar este contrato. Então, caso a recuperação deste contrato ocorra em até sessenta dias em atraso, não há a necessidade de pagar a comissão, o que implica em redução de custos.

Porém, os custos gerados por um *call center* são altos, variam desde os salários/benefícios dos operadores de cobrança, analistas responsáveis em elaborar a capacidade de um *call center* (também chamado de *"capacity planning"*) até o monitoramento constante, através de relatórios, para medir a performance dos operadores. Portanto, caso a empresa não tenha uma equipe já preparada para todas estas funções, a saída pode ser a terceirização deste serviço.

Todavia, é necessário fazer a relação custo/benefício como, por exemplo, o custo de um *call center* e a comissão paga às assessorias terceirizadas. Caso a comissão paga para assessorias terceirizadas seja mais alta do que se manter um *call center* interno, se mantém a cobrança interna e vice-versa. Outro ponto que deve ser avaliado é a performance dos operadores em relação a recuperação, é necessário avaliar qual dos dois tipos de serviço consegue recuperar mais em até 60 dias em atraso, por exemplo. Modelos de cobrança interna como o chamado "*Collection call model*", que consiste no operador seguir um "*script*" pré-definido, com ações e respostas condicionadas às falas do cliente, tem se mostrado muito efetivo nos últimos anos, elevando a taxa de recuperação da cobrança interna.

2.4 Cobrança terceirizada

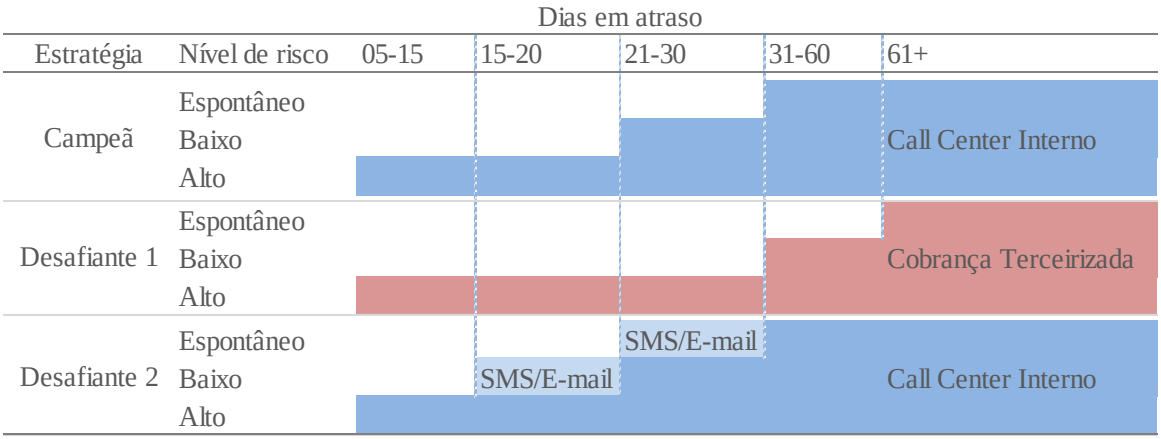
Segundo Giosa (1997), “a terceirização é uma técnica moderna de administração e que se baseia num processo de gestão, que leva a mudanças estruturais da empresa, a mudanças de cultura, procedimentos, sistemas e controles, capitalizando a organização com o objetivo único quando adotada: atingir melhores resultados, concentrando todos os esforços e energia da empresa para sua atividade principal. O principal objetivo da terceirização é a redução de custos com pessoal e equipamentos, entre outros, através da transferência de serviços. O fator que motiva essa mudança de gestão de carteira é o fato de minimizar as perdas financeiras associadas a contratos que já causaram prejuízo à organização”

Como uma forma de desafiar a cobrança interna, a cobrança terceirizada tem como foco principal o repasse de uma carteira de inadimplentes para uma outra empresa especializada em cobrança. Esta carteira é composta por clientes que possuem atraso superior a um período pré-estabelecido de acordo com o nível de risco de cada cliente. Como dito anteriormente, a definição deste período é de suma importância e deve ser realizada levando-se em conta o “C*ollection score*” de cada cliente e a proporção empírica de bons da carteira.

Existem, ainda, aqueles clientes que irão quitar suas dívidas sem mesmo precisarem ser cobrados. Neste caso chamamos estes clientes de pagadores espontâneos. É de se esperar que um pagador espontâneo tenha um “C*ollection Score*” alto e, portanto, o envio para a cobrança terceirizada deve ser retardado o máximo possível para estes clientes, uma vez que existe o pagamento de comissão sobre contratos recuperados pelas assessorias terceirizadas. Também temos aqueles que apenas se esqueceram e necessitam apenas de uma ligação ou uma mensagem de texto para lembrá-los, para estes casos, chamamos de nível de risco baixo, em que o *“Collection Score”* não é tão bom como o dos pagadores espontâneos, porém melhor do que os clientes de nível de risco alto. Para clientes que possuem o nível de risco alto, temos uma baixa propensão a pagamento e, portanto, o esperado é um “*Collection Score”* baixo. Como estes clientes têm grande chance de se tornarem maus pagadores, o envio tanto para o *call center*, quanto para as assessorias terceirizadas de cobrança, deve ser imediato após a entrada deles em inadimplência.

Na Figura 3, a seguir, temos um exemplo de uma “régua de cobrança”.

Figura 3 – Exemplo de uma régua de cobrança



Fonte: O autor (2017).

2.5 Venda de carteira

O processo de venda de carteiras consiste em analisar uma determinada população, já com tempo em atraso elevado e provisionados no Banco Central do Brasil. O processo de provisionamento ocorre a partir de cento e oitenta dias em atraso do cliente devedor. As dívidas dos maus pagadores podem ser vendidas a um preço bem inferior do que valem para uma empresa terceirizada ou outra instituição financeira. Por exemplo, uma carteira com um valor total de dez milhões de reais que podem ser recuperados, pode ter seu valor estipulado em quatro milhões de reais devido ao elevado número de dias em atraso que os clientes já se encontram. Sendo assim, para a empresa ou banco que está vendendo as dívidas dos clientes em atraso, pode-se obter um lucro, dado que a instituição deixará de provisionar estes clientes, sendo contabilizado como lucro direto.

Normalmente, a venda de carteiras ocorre nos casos em que todos os esforços de cobrança já foram aplicados e após todas as etapas anteriores terem sido esgotadas, de modo que o custo de se continuar mantendo a conta ativa em cobrança já não se justifica frente ao benefício da venda e a realização imediata de receita. Porém, antes de qualquer venda é necessário avaliar o público-alvo que será vendido e ver a real necessidade de vender, pois caso ocorra a venda de uma dívida de um cliente que possivelmente a pagaria, a instituição tem um prejuízo.

Para auxiliar na decisão sobre a venda ou não de carteiras, o modelo de mistura com fração de cura, neste caso com fração de inadimplentes, além de estimar a probabilidade de pagamento, também possibilita estimar o tempo até o pagamento e o comportamento de cada perfil de cliente durante todo o período de observação, diferente do modelo logístico, em que se tem apenas a probabilidade de pagamento ao final de uma janela de observação. Sendo assim, o modelo de mistura nos auxilia a definir o tempo em que os clientes com determinado perfil são elegíveis para que suas dívidas sejam vendidas para outras instituições financeiras. Por exemplo, se 9 meses corresponde ao tempo estimado para que 80% dos clientes, com determinado perfil, tenha suas dívidas pagas, então, após o período de 9 meses, os clientes que não pagaram suas dívidas, são elegíveis para que suas dívidas sejam vendidas para outras instituições financeiras.

3 MATERIAL E MÉTODOS

3.1 Material

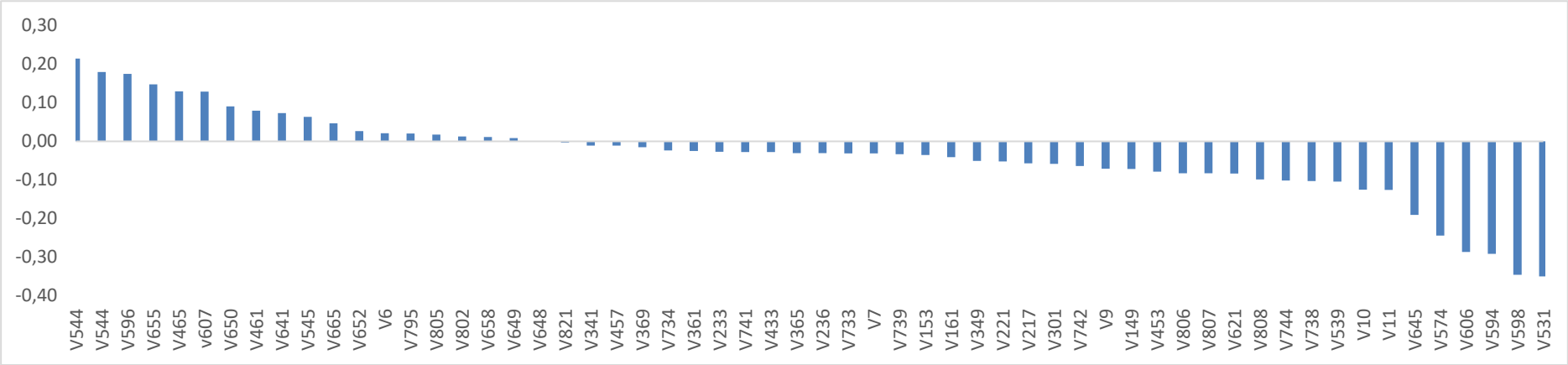
3.1.1 Banco de dados

O banco de dados utilizado neste trabalho foi disponibilizado por uma instituição financeira, ou seja, um banco que oferece aos seus clientes serviços como cartão de crédito, conta corrente, conta poupança, crédito imobiliário, financiamento de automóveis, dentre outros. O mesmo consiste em uma população de um milhão sessenta e nove mil e duzentos e setenta clientes, sendo que a característica comum desses clientes é o fato de serem inadimplentes com mais de sessenta dias em atraso, ou seja, estão em atraso devido ao não pagamento de algum empréstimo, cartão de crédito, dentre outros. Desta população, foi extraída uma amostra aleatória de vinte e seis mil clientes, que foi utilizada para as análises estatísticas.

Os clientes da base de dados mencionada foram monitorados dia a dia desde o atraso do pagamento de suas dívidas. A data de vencimento do primeiro produto em atraso foi considerada como a data do primeiro atraso, sendo que a partir de sessenta dias em atraso, os atrasos dos clientes foram monitorados por vinte e quatro meses.

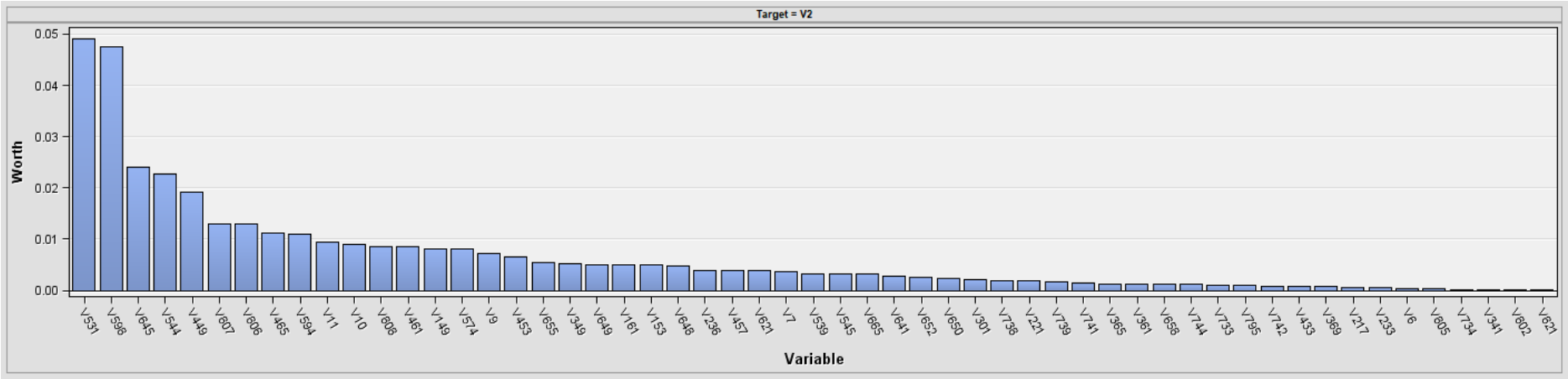
A base de dados era, inicialmente, composta por oitocentas e noventa e uma variáveis. Após a realização de uma análise exploratória dos dados, foi verificada a existência de algumas inconsistências tais como: variáveis com valores faltantes; variáveis com frequência de clientes em suas respectivas categorias demasiadamente pequenas, e variáveis com problemas de discriminação e ordenação. A fim de evitar vieses nas análises, essas variáveis foram excluídas. Após essa análise inicial, restaram cinquenta e seis variáveis. Destas cinquenta e seis variáveis restantes, foi feita mais uma análise de correlação em relação à variável resposta. Também foi medido o valor da informação para explicação da variável resposta (RUSH, 2014), conforme mostrado nos dois gráficos a seguir.

Gráfico 1 – Correlação de Pearson das variáveis explicativas com a variável resposta



Fonte: O autor (2017).

Gráfico 2 – Valor da informação (RUSH, 2014) para as variáveis explicativas em relação a variável resposta



Fonte: O autor (2017).

Com base nessas análises preliminares, as dez variáveis que apresentaram a maior correlação e o maior valor da informação para clientes inadimplentes foram: 1) percentual de restritivos baixados (quantas vezes já foi negativado e retirado por pagar seus compromissos atrasados dos *bureaux* de informação como Serasa, SPC, Boa Vista), 2) grau de severidade máxima do restritivo ativo, 3) percentual de restritivos decursados (decursados são os clientes que expiram cinco anos de atraso e por lei o restritivo precisa ser retirado, ou seja, após cinco anos se o cliente não pagar a dívida é necessário "limpar" o nome nos *bureaux* de informação), 4) grau máximo de severidade do restritivo decursados, 5) atraso inicial, 6) indicativo de renegociação (neste caso, o cliente já fez uma renegociação de dívida e também houve o atraso desta renegociação), 7) utilização do caixa eletrônico nos últimos três meses, 8) percentual de contratos em atraso, 9) percentual de utilização do limite do cartão de crédito, 10) tempo de relacionamento do cliente com a instituição financeira. Algumas informações como sexo, religião, dentre outras, não foram testadas pois não podem ser utilizadas devido à política e leis regulamentares. As dez variáveis citadas foram categorizadas para facilitar a interpretação e explicação no que se refere à regra de negócios e políticas internas da instituição financeira.

A Tabela 1 apresenta uma breve descrição das covariáveis selecionadas para o ajuste dos modelos. De modo geral, elas trazem informações sobre o perfil e comportamento dos clientes.

Tabela 1 – Covariáveis categorizadas selecionadas para modelagem.

Covariável	Categorização
Percentual de restritivos baixados	(00,00%, 55,43%]
	(55,43%, 79,88%]
	(79,88%, 100,00%]
Grau máximo do restritivo ativo	Muito grave e grave
	Remoto, baixo e médio
Percentual de restritivos decursados	(00,00%, 42,76%]
	(42,76%, 60,10%]
	(60,10%, 74,88%]
	(74,88%, 87,50%]
	(87,50%, 100,00%]
	Nunca teve restritivo decursado anterior
Grau máximo restritivo decursados	Muito grave e grave
	Remoto, baixo e médio

Atraso inicial	61 - 180 dias em atraso 181 - 360 dias em atraso 361 - 1440 dias em atraso > 1440 dias
Indicativo de renegociação	Possui renegociação em atraso Não possui uma renegociação em atraso
Utilização do caixa eletrônico nos últimos três meses	Utilizou nos últimos três meses Não utilizou nos últimos três meses
Percentual de contratos em atraso	(00,00%, 48,38%] (48,38%, 65,38%] (65,38%, 100,00%]
Percentual de utilização do limite do cartão de crédito	Sem Uso (00,00%, 61,95%] (61,95%, 100,00%]
Tempo de relacionamento do cliente	> 14,17 anos 9,51 - 14,17 anos 6,59 - 9,50 anos 3,93 - 6,58 anos 0 - 3,92 anos

Fonte: Instituição financeira (2017).

3.1.2 Recursos Computacionais

O *software* estatístico SAS (*Statistical Analysis System*) Enterprise Guide 7.1 foi utilizado para a análise do banco de dados. Os principais procedimentos foram: PROC LIFETEST e PROC LOGISTIC. Também foi utilizada uma macro do SAS, denominada PSPMCM (*parametric and semiparametric mixture cure models*), proposta por Corbière e Joly (2007) para o ajuste do modelo de mistura com fração de inadimplentes.

3.2 Métodos

3.2.1 Regressão Logística

De modo geral, modelos de regressão buscam estabelecer relações entre uma variável resposta e variáveis explicativas. O modelo de regressão logística, em particular, se diferencia do modelo de regressão linear quanto à natureza da variável resposta, caracterizada apenas por valores binários ou dicotômicos, usualmente denotados por 1 e 0,

com o valor 1 denominado evento de interesse (HOSMER; LEMESHOW, 2000). Em outras palavras, o modelo de regressão logística é útil para modelar fenômenos aleatórios com dois desfechos possíveis (sucesso ou fracasso) em função das variáveis explicativas.

Segundo GIOLO (2017), a regressão logística se constitui em um dos principais modelos utilizados quando se deseja analisar dados em que a variável resposta é binária ou dicotômica. Mesmo quando a resposta de interesse não é originalmente binária, é usual que esta seja dicotomizada de modo que a probabilidade de sucesso possa ser estimada por meio de um modelo de regressão logística. Embora existam outros modelos para analisar dados em que a resposta é binária, a regressão logística se tornou popular por ser flexível do ponto de vista matemático, de fácil utilização e por apresentar interpretação simples de seus parâmetros.

O modelo de regressão logística fica definido pelo uso da ligação logito em um modelo linear generalizado binomial. Formalmente, o modelo de regressão logística fica dado por

$$E(Y|\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = p(\boldsymbol{x}) = \frac{\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)}$$

em que $\boldsymbol{x} = (x_1, x_2, ..., x_p)$ denota o vetor de valores observados das variáveis explicativas, β_0 corresponde a uma constante e os componentes β_k são os p parâmetros ou coeficientes de regressão, sendo $\beta = \beta_1, \beta_2, \beta_3 \ldots, \beta_p$. Pode-se, ainda, simplificar $p(\boldsymbol{x})$ por

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

e, também, obter a probabilidade de o indivíduo não apresentar a resposta de interesse, conforme fórmula a seguir

$$1 - p(\boldsymbol{x}) = \frac{1}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x})}.$$

Em termos do logito, o logaritmo da razão entre os termos $p(\boldsymbol{x})$ e $1 - p(\boldsymbol{x})$ fornece um modelo linear (BERKSON,1944), conforme segue

$$ln\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k = \boldsymbol{\beta}'\boldsymbol{x}.$$

.

A razão entre $p(\boldsymbol{x})$ e $1 - p(\boldsymbol{x})$ é chamada de *odds* ou chance, ou seja,

$$chance = \frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = \exp(\boldsymbol{\beta}'\boldsymbol{x}).$$

A interpretação dos parâmetros em um modelo de regressão logística baseia-se em razões de chances (*odds ratios*). Assim, se B e A correspondem às categorias de uma variável explicativa (denotadas por 1 e 0, respectivamente), segue que a razão de chances entre elas é dada por

$$odds\ ratio\ \{B|A\} = \frac{odds\{B\}}{odds\{A\}} = \frac{\exp(\beta_0 + \beta_1 \times 1)}{\exp(\beta_0 + \beta_1 \times 0)} = \exp(\beta_1).$$

Caso a variável explicativa apresente mais de duas categorias, ela é incorporada ao modelo por meio de $(k - 1)$ variáveis *dummy*, sendo $k > 2$ o número de categorias.

Segundo Hosmer e Lemeshow (2000), os estimadores de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\beta}$ são os valores que maximizam a função de verossimilhança $L(\boldsymbol{\beta})$, a qual expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. A estimação de $\boldsymbol{\beta}$ no modelo de regressão logística é usualmente feita pelo método da máxima verossimilhança, em que $L(\boldsymbol{\beta})$ é dada por

$$L(\boldsymbol{\beta}) = \prod_{l=1}^{n} \left(P(Y = y_l|\boldsymbol{x_l})\right) = \prod_{l=1}^{n} \left(p(\boldsymbol{x_l})\right)^{y_l} \left(1 - p(\boldsymbol{x_l})\right)^{1-y_l}$$

em que $l = 1, \ldots, n$ denota o conjunto de *n* indivíduos independentes, $y_l = 1$ se o indivíduo l apresentou a resposta e $y_l = 0$, caso contrário.

3.2.2 Seleção de variáveis e ajuste do modelo de regressão logística

O processo de seleção de covariáveis tem por objetivo a identificação de um modelo parcimonioso (que seja simples e com número reduzido de parâmetros), mas capaz de se ajustar satisfatoriamente aos dados. Em estudos que envolvem um número elevado de covariáveis (ou fatores), pode ser útil usar algum algoritmo de seleção para identificação de um modelo adequado.

Neste trabalho, foi ajustado inicialmente um modelo para cada uma das covariáveis separadamente a fim de avaliar a significância do efeito de cada uma delas. Foram mantidas nos passos subsequentes apenas aquelas que apresentaram um valor p inferior a 0,05 associado ao teste de Wald (WALD, 1943).

Considerando as covariáveis que individualmente apresentaram efeito significativo foi, então, utilizado o método de seleção *forward* para a seleção do modelo final. Esse procedimento começa pelo modelo nulo (apenas com o intercepto), a seguir o método seleciona, dentre todas as covariáveis, aquela que proporciona maior ganho de ajuste (segundo algum critério como, por exemplo, menor AIC). Nos passos seguintes, uma a uma, as demais covariáveis são inseridas ao modelo, sempre selecionando aquela que proporciona maior ganho de ajuste na presença das covariáveis já inseridas ao modelo. O processo se encerra quando nenhuma das covariáveis fora do modelo contribui para um melhor ajuste, segundo o critério adotado.

Para cada modelo ajustado no método de seleção descrito, foram comparadas as estimativas deste modelo com aquelas fornecidas pelos modelos que consideraram as covariáveis separadamente com o objetivo de verificar a presença de multicolinearidade entre as covariáveis. Além dos valores p associados ao teste de Wald, também foram monitorados o critério de informação de Akaike (AIC) e a área abaixo da curva ROC dos modelos. Tais critérios adicionais foram utilizados devido à sensibilidade que o teste de Wald apresenta em grandes amostras (GRANZOTTO *et al*., 2010). O critério de informação de Akaike é uma importante medida usada para avaliar a qualidade do ajuste de modelos. De modo geral, pode-se ajustar diferentes modelos e optar por aquele que produzir o menor AIC. Introduzido por Hirotsugu Akaike em 1974, o AIC penaliza os modelos com covariáveis desnecessárias e é calculado da seguinte forma

$$AIC = -2\, log\, L(\boldsymbol{\theta}) + 2\mathrm{p},$$

em que $log\, L(\boldsymbol{\theta})$ corresponde ao logaritmo da função de verossimilhança do modelo com vetor de parâmetros $\boldsymbol{\theta}$ e p ao número de parâmetros do modelo.

Uma forma de analisar o poder preditivo associado ao modelo ajustado é por meio da curva ROC (*receiver operating characteristic*), a qual permite avaliar conjuntamente a sensibilidade (proporção de clientes bons que são classificados corretamente como bons) e a especificidade (proporção de clientes maus que são classificados corretamente como maus pelo modelo). Em geral, deseja-se que o modelo apresente sensibilidade e especificidade elevadas.

Para se ter uma curva ROC é necessário estabelecer pontos de corte, que estão no intervalo $[0,1]$. Estabelecido os pontos de corte, assume-se que $y = 1$, ou seja, que o cliente é um bom pagador, para as probabilidades preditas pelo modelo com valores superior ou

igual ao ponto de corte e $y = 0$ (mau pagador), caso contrário. Em seguida, é construído um gráfico com os pares $(x, y) = (1 - especificidade, sensibilidade)$ para os pontos de corte definidos anteriormente. O modelo com maior poder preditivo será o que apresentar área abaixo da curva ROC mais próxima a um, produzindo, assim, o maior percentual de acertos.

3.2.3 Modelo de mistura de Cox com fração de inadimplentes

A análise de sobrevivência é utilizada quando se deseja estimar a probabilidade de sobrevivência a um evento de interesse (denominado falha) associada a cada instante de tempo durante um período pré-estabelecido de observação. Por isso, no contexto descritivo de análise de sobrevivência, foi utilizado, neste trabalho, o estimador não paramétrico de Kaplan-Meier (KAPLAN; MEIER, 1958) para a estimação da função de sobrevivência. Também pelo fato de termos na base de dados uma parte da população não susceptível ao evento de interesse (EUDES et al., 2012), foi utilizado uma extensão do modelo de Cox (COX,1972), denominado modelo de mistura de Cox com fração de cura (neste caso fração de inadimplentes), que é capaz de acomodar essa fração de indivíduos em que não ocorre o evento, tendo sido proposto por Kuk e Chen (1992).

Segundo EUDES et al. (2012), é usual assumir em análise de sobrevivência que todos os indivíduos sob estudo irão apresentar o evento de interesse se forem acompanhados por um período de tempo suficientemente longo para que isso ocorra. Contudo, existem situações em que uma fração de indivíduos não apresentará o evento de interesse, mesmo se acompanhados por um longo período. Em tais casos, e dependendo da área dos dados sob análise (médica, financeira etc.), essa fração de indivíduos é denominada: imunes, curados, fidelizados, sobreviventes de longa duração ou, ainda, não suscetíveis ao evento de interesse. O modelo de sobrevivência semi-paramétrico considerado neste trabalho levou em conta a fração de cura, ou fração de inadimplentes, no caso dos dados analisados.

Esta fração deve ser explicada de forma que não cause um confundimento com as censuras. Portanto, é importante entender o que é fração de cura e o que é censura. Por exemplo, em um estudo em que os indivíduos foram acompanhados por certo período de tempo (período suficientemente grande, por isso longa duração) e o evento, após esse período, ainda não ocorreu para uma parte deles, então o evento provavelmente não mais acontecerá para a maioria deles. A essa fração (ou proporção) de indivíduos na qual o

evento não ocorrerá, mesmo se observados por mais tempo, denomina-se fração de cura. Por outro lado, a parcela de indivíduos, usualmente muito pequena, na qual o evento não foi observado após um período longo de acompanhamento, mas que possivelmente ocorreria, caso fossem acompanhados por mais tempo, caracterizam as censuras. Os modelos de mistura para dados de longa duração foram propostos para acomodar essas situações. Desse modo, esses modelos foram considerados nesse trabalho, tendo em vista a fração elevada de inadimplentes existente no banco de dados a ser analisado.

Esse modelo considera que existem duas subpopulações distintas (uma suscetível $U = 1$, e outra não suscetível ao evento de interesse $U = 0$, por mais longo que seja o tempo de acompanhamento). Considere T uma variável aleatória não negativa representando o tempo até o evento de interesse e suponha que exista na população em estudo uma proporção $p_0 = 1 - \pi$ de indivíduos imunes ao evento (maus pagadores) e, consequentemente, uma proporção $q_0 = \pi$ suscetíveis ao evento (bons pagadores). O modelo de mistura, na presença de covariáveis, fica expresso por

$$S_p(t|\boldsymbol{x}, \boldsymbol{z}) = P(U = 0|\boldsymbol{z})P(T > t|U = 0, \boldsymbol{x}) + P(U = 1|\boldsymbol{z})P(T > t|U = 1, \boldsymbol{x}) \quad (1)$$

com $\boldsymbol{z}$ o vetor de covariáveis associado à proporção $\pi(\boldsymbol{z})$, que indica a probabilidade de o indivíduo ser suscetível (bom). Para estudar o efeito do vetor de covariáveis $\boldsymbol{z}$ sobre $\pi(\boldsymbol{z})$ são utilizadas com frequência as funções logística, probito e clog-log. Neste estudo, por se tratar de uma maneira alternativa a já utilizada nas instituições financeiras, optou-se pela utilização da função logística, expressa conforme a fórmula a seguir

$$\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}'\boldsymbol{\beta})} = log\left[\frac{\pi(\boldsymbol{z})}{1 - \pi(\boldsymbol{z})}\right] = \boldsymbol{z}'\boldsymbol{\beta}$$

com $\boldsymbol{\beta}$ o vetor de parâmetros associados às covariáveis $\boldsymbol{z}$.

O vetor de covariáveis $\boldsymbol{x}$ associado à função de sobrevivência condicional $S(t|\boldsymbol{x})$, considera apenas os indivíduos suscetíveis ao evento. Em outras palavras, o componente do modelo expresso pela função de sobrevivência condicional $S(t|\boldsymbol{x})$ considera apenas os bons pagadores e, portanto, tal curva de sobrevivência sempre terminará em zero. No contexto semi-paramétrico (utilizado neste trabalho) a expressão para $S(t|\boldsymbol{x})$ é dada por

$$S(t|\boldsymbol{x}) = [S_0(t)]^{\exp(\boldsymbol{x}'\gamma)}$$

com $\boldsymbol{\gamma}$ o vetor de parâmetros associados às covariáveis $\boldsymbol{x}$ e $S_0(t)$ a função de sobrevivência de base associada aos indivíduos suscetíveis ao evento.

Por fim, $S_p(t|\boldsymbol{x}, \boldsymbol{z})$ corresponde à função de sobrevivência para toda a população, isto é, suscetíveis e não suscetíveis (bons e maus pagadores). A função de verossimilhança associada ao modelo representado em (1) é apresentada a seguir

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{\pi(\boldsymbol{z_i}) f(t_i|\boldsymbol{x_i})\}^{\delta_i} \{1 - \pi(\boldsymbol{z_i}) + \pi(\boldsymbol{z_i}) S(t_i|\boldsymbol{x_i})\}^{1-\delta_i}$$

em que $i = 1, \ldots, n$ indexa os n indivíduos; δ_i é o indicador de falha (1 se falha e 0 se censura) e t_i é o tempo de falha (CORBIÈRE; JOLY, 2007).

Estimadores para os vetores de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$, bem como para $S_0(t)$, são obtidos maximizando-se a função de verossimilhança $L(\boldsymbol{\gamma}, \boldsymbol{\beta})$ via o algoritmo EM (do inglês, *estimation and maximization*). Para tal estimação, Corbière e Joly (2007) desenvolveram uma macro no *software* SAS. Nesta macro, $\pi(\boldsymbol{z})$ pode ser modelada por meio dos modelos: logístico, probito e clog-log (neste trabalho foi utilizado o logístico, como mencionado anteriormente), enquanto $S(t|\boldsymbol{x})$ pode ser estimada por meio de modelos paramétricos (exponencial, Weibull, logístico e log-normal) ou, ainda, pelo modelo semi-paramétrico de Cox (utilizado neste trabalho).

x3.2.4 Seleção de variáveis e ajuste do modelo de mistura com fração de inadimplentes

Para a seleção de covariáveis do modelo, foi utilizado o método *forward* e o critério de informação de Akaike (AIC). Para ambos os componentes, logístico e semi-paramétrico de Cox, o nível de significância considerado foi de 0,05. Uma observação quanto ao procedimento de estimação é que: o componente logístico considera toda a população para estimar os parâmetros, enquanto o componente de sobrevivência condicional utiliza apenas as observações que falharam, restringindo assim o número de observações para este último componente. Porém, a amostra utilizada tem tamanho elevado, não tendo sido necessária nenhuma consideração a respeito dos níveis de significância.

Em relação à qualidade de ajuste, foram utilizados o coeficiente de correlação de Pearson e o R^2(coeficiente de correlação de Pearson ao quadrado). Ambos foram utilizados para avaliar a correlação entre as probabilidades de sobrevivência obtidas por meio do estimador de Kaplan-Meier e pelo modelo de fração de inadimplentes ajustado, para todas

as combinações das covariáveis categóricas. Coeficientes próximos a um evidenciam um ajuste satisfatório do modelo.

4 RESULTADOS E DISCUSSÃO

4.1 Análise descritiva

Para começar a análise, foi feita inicialmente uma análise descritiva dos dados a fim de se conhecer a quantidade de clientes por categoria de cada variável e o percentual que cada categoria representa do total. A quantidade de pagamentos em até vinte e quatro meses de observação, assim como o percentual por categoria dos clientes que pagaram em até vinte quatro meses, se encontram na Tabela 2.

No contexto dos dados analisados, notou-se que as variáveis estão fazendo sentido quanto à interpretação. Por exemplo, quanto à variável *percentual de restritivos baixados* tem-se que quanto maior o percentual de restritivos baixados melhor, pois quando um cliente é baixado do restritivo significa que pagou suas dívidas (total ou parcialmente). Para esta variável, observou-se até doze meses antes da entrada do cliente em atraso. Para a variável *grau máximo do restritivo ativo*, graus muito grave e grave são piores do que os graus remoto, baixo e médio. *Percentual de restritivos decursados* significa que quanto maior pior é a sua liquidação, pois um restritivo decursado, conforme descrito anteriormente, é quando chegou no atraso máximo de cinco anos e a instituição financeira precisa necessariamente limpar o nome do cliente. Para a variável *grau máximo de restritivos decursados* tem-se a mesma interpretação da variável *grau máximo do restritivo ativo*. Para o *atraso inicial*, quanto maior for o atraso inicial menor é a proporção de bons pagadores. Quanto ao *indicativo de renegociação*, a interpretação é a de que, se o cliente possui uma renegociação em atraso, ele já é pior que um cliente que não tem, o que pode ser observado na Tabela 2. Para a variável *utilização do caixa eletrônico nos últimos três meses*, nota-se que o cliente que fez alguma transação no caixa eletrônico é melhor do que o cliente que não fez transações nos últimos três meses. *Percentual de contratos em atraso* significa que quanto mais contratos em atraso do cliente, menor é a proporção de bons pagadores. Já quanto ao *percentual de utilização do limite do cartão de crédito*, quanto maior o uso, menor o percentual de bons pagadores. E por fim, *tempo de relacionamento do cliente*, quanto maior o tempo de relacionamento, maior a proporção de bons pagadores.

Tabela 2 - Estatística descritiva associada às covariáveis de vinte e seis mil clientes em atraso.

Covariável	Total (#)	Total (%)	Pagamentos (#)	Pagamentos (%)
Percentual de restritivos baixados				
(00,00%, 55,43%]	8.954	34,44%	608	6,79%
(55,43%, 79,88%]	8.397	32,30%	1.907	22,71%
(79,88%, 100,00%]	8.649	33,26%	3.215	37,17%
Grau máximo do restritivo ativo				
Muito grave e grave	17.009	65,42%	2.207	12,98%
Remoto, baixo e médio	8.991	34,58%	3.523	39,19%
Percentual de restritivos decursados				
(00,00%, 42,76%]	5.113	19,66%	1.828	35,75%
(42,76%, 60,10%]	5.059	19,46%	1.213	23,98%
(60,10%, 74,88%]	5.403	20,78%	941	17,42%
(74,88%, 87,50%]	4.757	18,30%	717	16,09%
(87,50%, 100,00%]	5.081	19,54%	766	14,11%
Nunca teve restritivo decursado anterior	587	2,26%	265	45,21%
Grau máximo de restritivo decursados				
Muito grave e grave	19.672	75,66%	3.380	17,18%
Remoto, baixo e médio	6.328	24,34%	2.350	37,13%
Atraso inicial				
61 - 180 dias em atraso	2.971	11,43%	1.761	59,28%
181 - 360 dias em atraso	2.393	9,20%	836	34,94%
361 - 1440 dias em atraso	9.466	36,41%	1.982	20,94%
> 1440 dias	11.170	42,96%	1.151	10,31%

Continuação Tabela 2 - Estatística descritiva associada às covariáveis de vinte e seis mil clientes em atraso

Covariável	Total (#)	Total (%)	Pagamentos (#)	Pagamentos (%)
Indicativo de renegociação				
Tem uma renegociação em atraso	6.889	26,50%	770	11,18%
Não possui uma renegociação em atraso	19.111	73,50%	4.960	25,96%
Utilização do caixa eletrônico nos últimos três meses				
Utilizou nos últimos três meses	4.581	17,62%	1.826	39,87%
Não utilizou nos últimos três meses	21.419	82,38%	3.904	18,22%
Percentual de contratos em atraso				
(00,00%, 48,38%]	2.572	9,89%	1.301	50,61%
(48,38%, 65,38%]	4.160	16,00%	1.122	26,98%
(65,38%, 100,00%]	19.268	74,11%	3.307	17,16%
Percentual de utilização do limite do cartão de crédito				
Sem Uso	2.495	9,59%	1.191	47,76%
(00,00%, 61,95%]	2.600	10,00%	859	33,05%
(61,95%, 100,00%]	20.905	80,41%	3.680	17,60%
Tempo de relacionamento do cliente				
> 14,17 anos	5.186	19,95%	1.670	32,19%
9,51 - 14,17 anos	5.233	20,13%	1.175	22,45%
6,59 - 9,50 anos	5.216	20,06%	1.060	20,33%
3,93 - 6,58 anos	5.118	19,69%	975	19,05%
0 - 3,92 anos	5.247	20,18%	850	16,19%

Fonte: O autor (2017).

4.2 Ajuste do modelo logístico

Para realizar o ajuste do modelo logístico, um modelo foi ajustado, inicialmente, para cada uma das variáveis individualmente. Foram mantidas para a próxima etapa apenas as variáveis explicativas que apresentaram um valor p inferior a 0,05. Sendo assim, restaram apenas cinco variáveis explicativas, sendo elas: 1) *atraso inicial, 2*) *percentual de contratos em atraso, 3*) *percentual de restritivos baixados, 4*) *percentual de utilização do limite do cartão de crédito, e 5) utilização do caixa eletrônico nos últimos três meses*.

Em um próximo passo, foi utilizado o método de seleção de covariáveis *forward*, fazendo a combinação da ordem de entrada das variáveis no modelo (todas as possíveis). Para a seleção do melhor modelo, foram utilizados três critérios: o valor p associado ao teste de Wald, o valor do critério de informação de Akaike (AIC) e a área abaixo da curva ROC (AUC). O modelo final selecionado ficou com a seguintes variáveis explicativas: *atraso inicial, percentual de contratos em atraso,* e *percentual de restritivos baixados.* As estimativas do vetor de parâmetros associado ao modelo de regressão logística estão na Tabela 3.

Tabela 3 - Estimativas e testes associados ao modelo de regressão logística selecionado

Parâmetro	Categoria	GL	Estimativa	Erro padrão	Wald Qui-Quadrado	Valor-p
Intercepto		1	-2,4996	0,0412	2.127,6757	<0,0001
Atraso Inicial	61 - 180	1	2,1304	0,0564	1.426,5380	<0,0001
Atraso Inicial	181 - 360	1	1,4392	0,0571	634,8640	<0,0001
Atraso Inicial	361 - 1440	1	0,8335	0,0419	395,6158	<0,0001
% Contratos em atraso	(00,00%, 48,38%]	1	0,4637	0,0548	71,6892	<0,0001
% Contratos em atraso	(48,38%, 65,38%]	1	0,0629	0,0446	1,9892	0,1584
% Restritivos baixados	(55,43%, 79,88%]	1	0,2971	0,0409	53,8535	<0,0001
% Restritivos baixados	(79,88%, 100%]	1	0,5973	0,0418	204,4850	<0,0001

Fonte: O autor (2017).

Como as variáveis explicativas foram categorizadas, as inclusões das mesmas no modelo ajustado ocorreram por meio de variáveis *dummy*, sempre escolhendo a categoria "menos favorável" como a categoria de referência com o intuito de facilitar a interpretação posteriormente. No caso da variável *atraso inicial*, a categoria de referência foi > 1440 dias, enquanto para *percentual de contratos em atraso* foi (65,38%, 100,00%] e para a *percentual de restritivos baixados* foi (00,00%, 55,43%]. O modelo ajustado, em termos dos logitos ficou dado conforme a expressão a seguir

$$logit\big(\hat{\pi}(\boldsymbol{z})\big) = -2{,}4996 + 2{,}1304z_{i1} + 1{,}4392z_{i2} + 0{,}8335z_{i3} + 0{,}4637z_{i4} + 0{,}0629z_{i5} + 0{,}2971z_{i6} + 0{,}5973z_{i7}$$

em que a variável z_{i1} corresponde à *faixa de atraso inicial* 61 – 180 dias; ou seja, se o cliente tem atraso inicial entre 61 e 180 dias, $z_{i1} = 1$; caso contrário, $z_{i1} = 0$. Esta mesma definição deve ser aplicada às outras covariáveis, sendo que z_{i2} corresponde à *faixa de atraso inicial* entre 181 e 360 dias e z_{i3} à *faixa de atraso inicial* entre 361 e 1440 dias. De forma análoga, z_{i4} corresponde ao *percentual de contratos em atraso* para clientes que possuem até 48,38% dos contratos em atraso; e z_{i5} aos clientes que possuem entre 48,39% e 65,38% dos seus contratos em atraso. Nota-se que para a categoria (48,38%, 65,38%] desta variável, o valor p foi de 0,1584, porém devido ao valor p da categoria (00,00%, 48,38%] ser < 0,0001, a variável foi mantida no modelo. Para a última variável, *percentual dos restritivos baixados*, z_{i6} corresponde aos clientes que têm seus restritos baixado de 55,43% até 79,88%, e z_{i7} aos que possuem os restritivos baixados acima de 79,88%.

Na Tabela 4, tem-se os valores AIC e -2(logaritmo da função de verossimilhança) do modelo final. O AIC do modelo que considera todas as covariáveis é menor do que o do modelo que considera somente o intercepto, o que sugere que essas variáveis ajudam a explicar a resposta com parcimônia, dado que o AIC penaliza a inclusão de covariáveis desnecessárias.

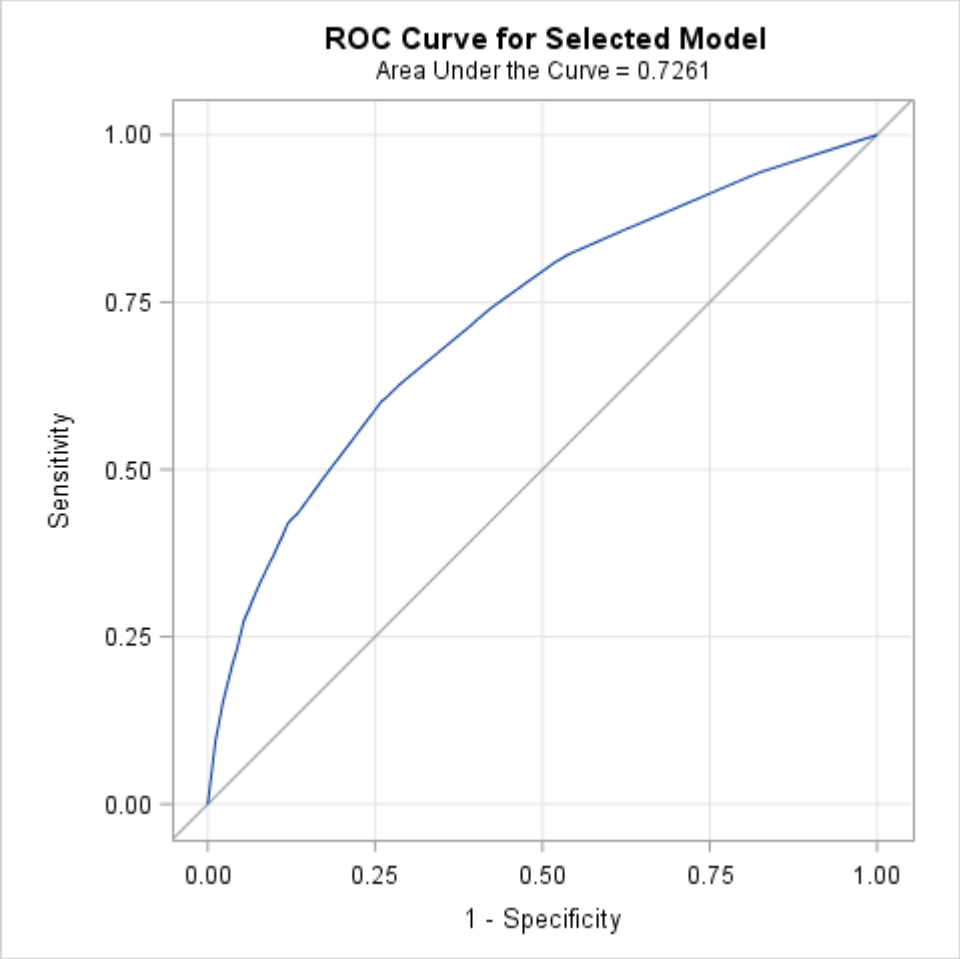
Tabela 4 - Estatísticas associadas ao modelo sem covariáveis e ao modelo selecionado

Critério	Somente intercepto	Intercepto e Covariáveis
AIC	25.499,06	22.525,31
-2 Log L	25.497,06	22.509,31

Fonte: O autor (2017).

Além disso, esse é o modelo que maximiza a área sob a curva ROC. A curva ROC também foi utilizada para verificar a adequação do modelo. A partir do Figura 3, pode-se notar que o modelo escolhido se ajusta aos dados de maneira satisfatória, apresentando bom poder de discriminação com uma área de 0,7261 abaixo da curva.

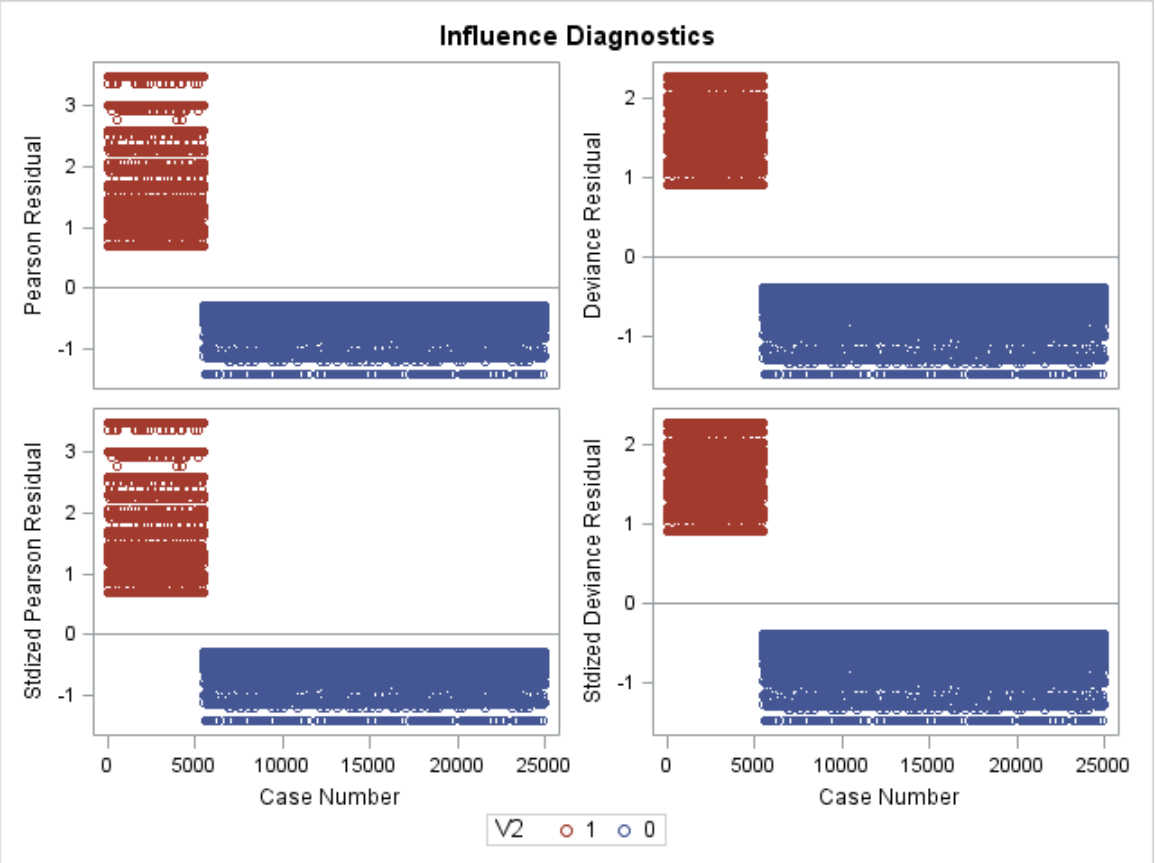
Figura 3 – Curva ROC associada ao modelo de regressão logística ajustado



Fonte: O autor (2017).

Para finalizar a análise da qualidade de ajuste do modelo aos dados, foi analisado os resíduos. Conforme a Figura 4, os resíduos *deviance* e de Pearson estão distribuídos em torno de zero e em um intervalo de variação satisfatório, o que reitera a adequação do modelo ajustado.

Figura 4 - Análise gráfica dos resíduos do modelo de regressão logística ajustado



Fonte: O autor (2017).

A análise dos resultados do modelo de regressão logística ajustado será discutida posteriormente no tópico de discussão sobre os resultados.

4.3 Ajuste do Modelo de Mistura com Fração de Inadimplentes

Anterior ao ajuste do modelo de mistura com fração de inadimplentes, realizou-se uma análise descritiva de cada uma das covariáveis fazendo-se uso do estimador de Kaplan-Meier. Com isso, foi possível analisar o tempo até os clientes inadimplentes pagarem suas dívidas (isto é, a velocidade com que os clientes pagaram suas dívidas) e as possíveis covariáveis que estariam associadas com o tempo mencionado. Ou seja, a análise descritiva serviu de referencial para a escolha das variáveis candidatas a entrarem nos modelos, podendo também auxiliar na interpretação do modelo final.

Em seguida, ajustou-se um modelo de mistura com fração de inadimplentes para cada uma das covariáveis indicadas na análise descritiva. Em cada um desses modelos, a covariável foi incluída simultaneamente no componente logístico, $\pi(\boldsymbol{z})$, e no componente de sobrevivência, $S(t|\boldsymbol{x})$. Foram mantidas nos passos subsequentes apenas aquelas que

apresentaram um valor p inferior a 0,05. Similar ao observado no ajuste do modelo logístico, sobraram apenas cinco variáveis explicativas, sendo elas: 1) a*traso inicial, 2*) *percentual de contratos em atraso, 3*) *percentual de restritivos baixados, 4*) *percentual de utilização do limite do cartão de crédito, e 5*) *utilização do caixa eletrônico nos últimos três meses*. O método de seleção das variáveis explicativas foi o *forward*, sendo excluídas aquelas que tiveram o valor p abaixo de 0,05.

As covariáveis que foram mais significativas no componente associado à regressão logística do modelo de mistura com fração de inadimplentes apresentou estimativas idênticas às do modelo que considera apenas essa técnica, pois foram selecionadas as mesmas covariáveis (os resultados estão na Seção 4.2).

Quanto ao componente de sobrevivência $S(t|\boldsymbol{x})$, as variáveis selecionadas foram as mesmas do componente logístico. Isso mostra o quanto essas variáveis são fortes na discriminação do perfil do cliente (bom e mau). As estimativas dos parâmetros associadas ao componente de sobrevivência (modelado via o modelo de Cox) estão na Tabela 5.

Tabela 5 - Estimativas e testes associados ao componente $S(t|\boldsymbol{x})$ do modelo de mistura com fração de inadimplentes selecionado

Parâmetro	Categoria	GL	Estimativa	Erro padrão	Wald Qui-Quadrado	Valor-p
Atraso Inicial	61 - 180	1	0,4794	0,0450	113,7356	<0,0001
Atraso Inicial	181 - 360	1	0,1918	0,0488	15,4571	<0,0001
Atraso Inicial	361 - 1440	1	0,0501	0,0387	1,6709	0,1961
% Contratos em atraso	(00,00%, 48,38%]	1	0,3238	0,0402	64,8161	<0,0001
% Contratos em atraso	(48,38%, 65,38%]	1	0,0046	0,0365	0,0160	0,8994
% Restritivos baixados	(55,43%, 79,88%]	1	0,0117	0,0392	0,0886	0,7660
% Restritivos baixados	(79,88%, 100%]	1	0,0880	0,0355	6,1277	0,0133

Fonte: O autor (2017).

Continuando a verificação do componente de sobrevivência $S(t|\boldsymbol{x})$ do modelo, tem-se, na Tabela 6, os valores do AIC e do logaritmo da função de verossimilhança, sendo possível notar que o modelo com as covariáveis apresenta o valor do AIC e do máximo da função de verossimilhança mais baixos do que o do modelo sem covariáveis, o que justifica a permanência das covariáveis no modelo.

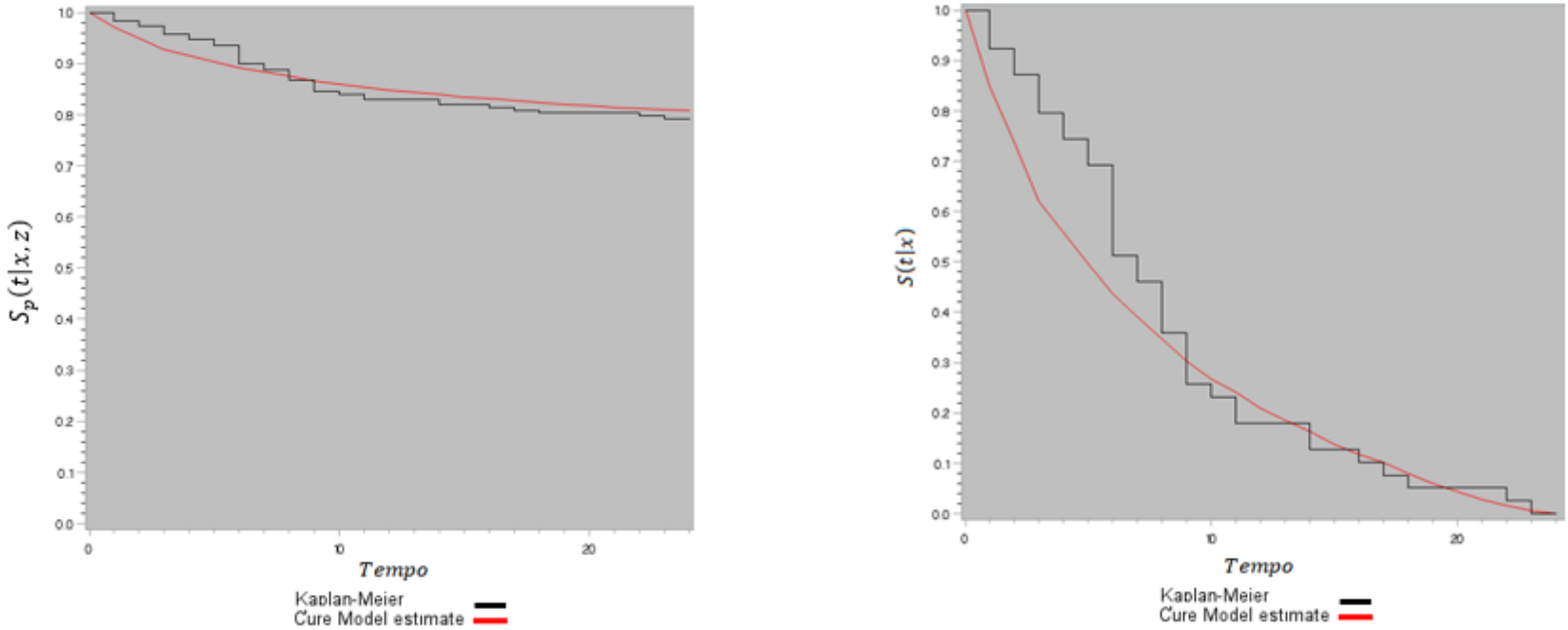
Tabela 6 - Estatísticas associadas ao modelo sem covariáveis e ao modelo selecionado.

Critério	Sem covariáveis	Com Covariáveis
AIC	80.738,932	80.369,622
-2 Log L	80.738,932	80.355,622

Fonte: O autor (2017).

O próximo passo é verificar a adequação do modelo selecionado. Para isso, foram obtidas as curvas de sobrevivência observada (representada pela curva obtida pelo estimador de Kaplan-Meier) e a estimada pelo modelo, tanto para a sobrevivência populacional $S_p(t|\boldsymbol{x}, \boldsymbol{z})$ quanto para a sobrevivência condicional $S(t|\boldsymbol{x})$. Estas curvas, para uma das combinações de $\boldsymbol{x}$ e $\boldsymbol{z}$, podem ser visualizadas na Figura 5 e mostram que as estimativas produzidas pelo modelo são bastante próximas às obtidas por Kaplan-Meier, evidenciando a adequação do modelo aos dados.

Figura 5 – Curva estimada para $S_p(t|\boldsymbol{x},\boldsymbol{z})$ e $S(t|\boldsymbol{x})$ com $\boldsymbol{x}$ *e* $\boldsymbol{z}$ os vetores de covariáveis associados ao cliente com o percentual de contratos em atraso até 48,38%, com o *atraso inicial* >1440 dias e *percentual de restritivos baixados* maiores que 79,88%, comparada com a curva de Kaplan-Meier



Fonte: O autor (2017).

Um critério adicional, que pode ser utilizado para avaliar o bom ajuste do modelo, é o coeficiente de correlação de Pearson, bem como o correspondente R^2, os quais são calculados para medir se cada ponto das curvas estimadas pelo modelo (conforme visto nos gráficos da Figura 5) está próxima das curvas observadas de Kaplan-Meier. Os valores desses coeficientes, para todas as combinações das categorias de todas as variáveis no modelo, podem ser observados na Tabela 7.

Tabela 7 - R^2 e coeficiente de correlação de Pearson para as 36 combinações das categorias das variáveis no modelo

Estrato	Atraso Inicial 61 - 180	Atraso Inicial 181 - 360	Atraso Inicial 361 - 1440	% Contratos em atraso (00,00%, 48,38%]	% Contratos em atraso (48,38%, 65,38%]	% Restritivos baixados (55,43%, 79,88%]	% Restritivos Baixados (79,88%, 100%]	Quantidade de clientes	R^2	Correlação de Pearson
1	0	0	0	0	0	0	0	3.620	0,99736	0,99868
2	0	0	0	0	0	0	1	3.682	0,99697	0,99849
3	0	0	0	0	0	1	0	1.920	0,99827	0,99913
4	0	0	1	0	0	0	0	2.169	0,99942	0,99971
5	0	0	1	0	0	0	1	2.522	0,99841	0,99920
6	0	0	1	0	0	1	0	2.261	0,99612	0,99806
7	0	1	0	0	0	0	0	384	0,99796	0,99898
8	0	1	0	0	0	0	1	301	0,98012	0,99001
9	0	1	0	0	0	1	0	532	0,99764	0,99882
10	1	0	0	0	0	0	0	170	0,99667	0,99833
11	1	0	0	0	0	0	1	183	0,99532	0,99766
12	1	0	0	0	0	1	0	569	0,99357	0,99678
13	0	0	0	0	1	0	0	229	0,99697	0,99848
14	0	0	0	0	1	0	1	304	0,99512	0,99756
15	0	0	0	0	1	1	0	377	0,98240	0,99116
16	0	0	1	0	1	0	0	489	0,99854	0,99927
17	0	0	1	0	1	0	1	528	0,99277	0,99638

Continuação Tabela 7 - R^2 e Coeficiente de Correlação de Pearson para as 36 combinações das categorias das variáveis no modelo

Estrato	Atraso Inicial 61 - 180	Atraso Inicial 181 - 360	Atraso Inicial 361 - 1440	% Contratos em atraso (00,00%, 48,38%]	% Contratos em atraso (48,38%, 65,38%]	% Restritivos baixados (55,43%, 79,88%]	% Restritivos Baixados (79,88%, 100%]	Quantidade de clientes	R^2	Correlação de Pearson
18	0	0	1	0	1	1	0	652	0,99633	0,99816
19	0	1	0	0	1	0	0	265	0,99684	0,99842
20	0	1	0	0	1	0	1	235	0,98545	0,99270
21	0	1	0	0	1	1	0	264	0,99065	0,99531
22	1	0	0	0	1	0	0	197	0,98899	0,99448
23	1	0	0	0	1	0	1	276	0,99399	0,99699
24	1	0	0	0	1	1	0	420	0,98759	0,99377
25	0	0	0	1	0	0	0	250	0,98142	0,99067
26	0	0	0	1	0	0	1	292	0,97355	0,98668
27	0	0	0	1	0	1	0	218	0,96254	0,98109
28	0	0	1	1	0	0	0	224	0,99443	0,99721
29	0	0	1	1	0	0	1	214	0,97934	0,98962
30	0	0	1	1	0	1	0	262	0,99328	0,99663
31	0	1	0	1	0	0	0	206	0,99732	0,99866
32	0	1	0	1	0	0	1	257	0,99431	0,99715
33	0	1	0	1	0	1	0	202	0,98161	0,99076
34	1	0	0	1	0	0	0	205	0,96105	0,98033
35	1	0	0	1	0	0	1	396	0,95776	0,97865
36	1	0	0	1	0	1	0	725	0,97504	0,98744

Fonte: O autor (2017).

A partir da Tabela 7, pode-se notar que aparentemente o modelo está bem ajustado. Ao observar os valores de R^2, tem-se que o menor valor é 0,95776 (quanto mais próximo de um melhor). Para o coeficiente de correlação de Pearson, o menor valor é 0,9786, associado ao estrato de número trinta e cinco (quanto mais próximo de um melhor). Portanto, ambas as estatísticas evidenciaram um bom ajuste do modelo aos dados.

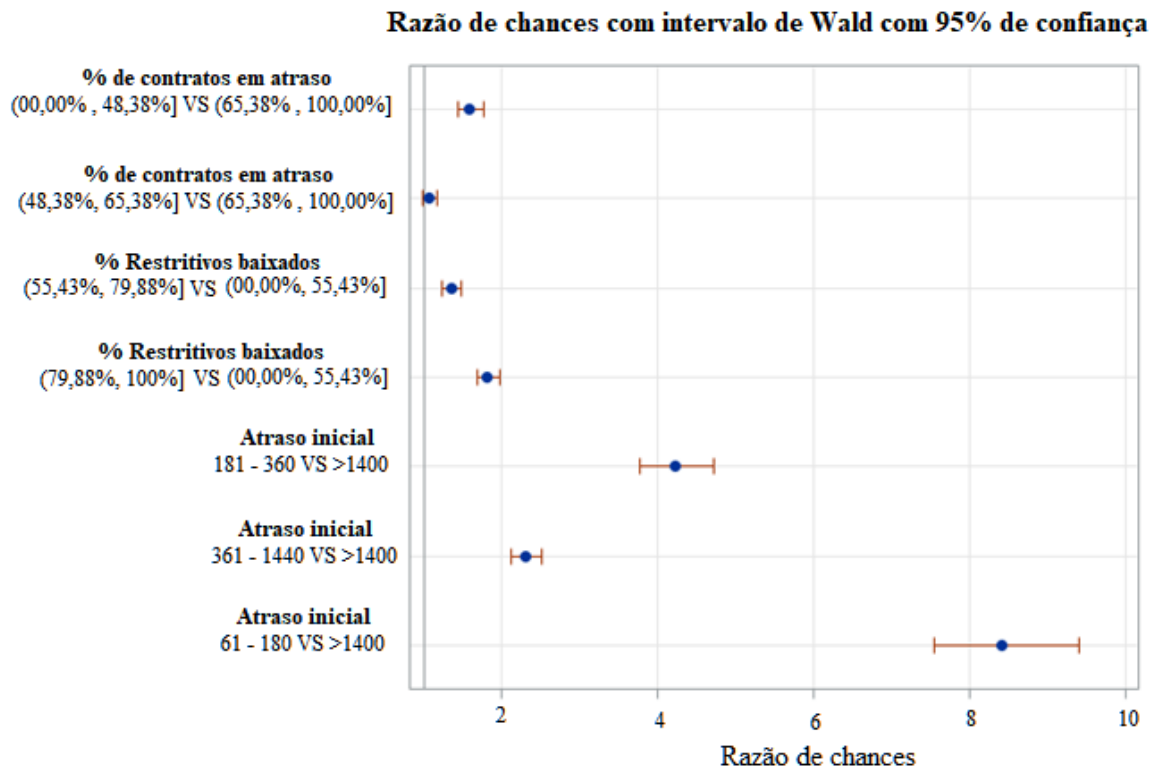
4.4 Interpretação dos resultados

Para interpretação dos resultados, será inicialmente avaliado o modelo de regressão logística ajustado na Seção 4.2, que permaneceu com três variáveis explicativas (*atraso inicial*, *percentual de contratos em atraso e percentual de restritivos baixados*) e ficou expresso da seguinte forma

$$logit\big(\hat{\pi}(\boldsymbol{z})\big) = -2{,}4996 + 2{,}1304z_{i1} + 1{,}4392z_{i2} + 0{,}8335z_{i3} + 0{,}4637z_{i4} + 0{,}0629z_{i5} + 0{,}2971z_{i6} + 0{,}5973z_{i7}.$$

Para proceder às interpretações dos parâmetros desse modelo por meio da razão de chances, foram obtidas as estimativas pontuais e intervalares representadas no gráfico da Figura 6. Na Tabela 8, são apresentados os valores das respectivas estimativas e intervalos de Wald com 95% de confiança (mostrados no gráfico).

Figura 6 – Razão de chances associada ao modelo de regressão logística ajustado aos dados



Fonte: O autor (2017).

Tabela 8 – Razão de chances e intervalos de Wald com 95% de confiança

		Intervalo de Wald com 95% confiança	
	Razão de chances	Limite inferior	Limite superior
Atraso inicial 61-180 VS >1440	8,418	7,537	9,402
Atraso Inicial 181-360 VS >1440	4,217	3,771	4,717
Atraso Inicial 361-1440 VS >1440	2,301	2,120	2,498
% Contratos em atraso (0 %, 48,38%] VS (65,38%,100%]	1,590	1,428	1,770
% Contratos em atraso (48,38%, 65,38%] VS (65,38%, 100%]	1,065	0,976	1,162
% Restritivos baixados (79,88%, 100%] VS (0%, 55,43%]	1,817	1,674	1,972
% Restritivos baixados (55,43%, 79,88%] VS (0%, 55,43%]	1,346	1,233	1,469

Fonte: O autor (2017).

A partir do modelo final ajustado, tem-se que a chance de pagamento dos clientes com *atraso inicial* de 61 a 180 dias foi de 8,418 vezes a dos clientes que estão na faixa de atraso > 1440 dias, ou seja, espera-se dos clientes com número menor de dias em atraso, uma maior chance de pagamento em até vinte e quatro meses. Ainda, clientes com *percentual de contratos em atraso* de até 48,38% apresentaram chance de pagamento igual a 1,59 vezes a dos clientes com mais de 65,38% dos seus contratos em atraso. Logo, quanto menor o número de contratos em atraso, aumenta a chance de o cliente pagar e, em consequência, aumenta a chance dele se tornar um bom cliente. Por fim, clientes com *percentual de restritivos baixados* acima de 79,88% apresentaram chance de realizar pagamento igual a 1,817 vezes a dos clientes com até 55,43%. Desse modo, espera-se que quanto mais restritivos baixados os clientes tiverem, melhor seja suas performances.

Para exemplificar, são considerados dois perfis de clientes, um deles muito bom e o outro ruim. As características dos dois perfis considerados estão na Tabela 9.

Tabela 9 – Perfil 1 e Perfil 2 de clientes em inadimplência

Covariáveis	Perfil 1	Perfil 2
Atraso inicial	61 – 180 dias	> 1440 dias
% Contratos em atraso	(00,00%, 48,38%]	(65,38%,100%]
% Restritivos baixados	(79,88%,100%]	(0%, 55,43%]

Fonte: O autor (2017).

Para os clientes com o Perfil 1, tem-se então que

$$logit\big(\hat{\pi}(z_1)\big) = -2{,}4996 + 2{,}1304 + 0{,}4637 + 0{,}5973 = 0{,}6918.$$

Assim, a probabilidade estimada de pagamento em até vinte e quatro meses dos clientes com esse perfil é igual a

$$\hat{\pi}(z_1) = \frac{\exp(0{,}6918)}{\exp(0{,}6918) + 1} = 0{,}6663.$$

Analogamente, para os clientes com o Perfil 2 tem-se

$$logit\big(\hat{\pi}(z_2)\big) = -2{,}4996,$$

e, em consequência, a probabilidade estimada de pagamento desses clientes é de

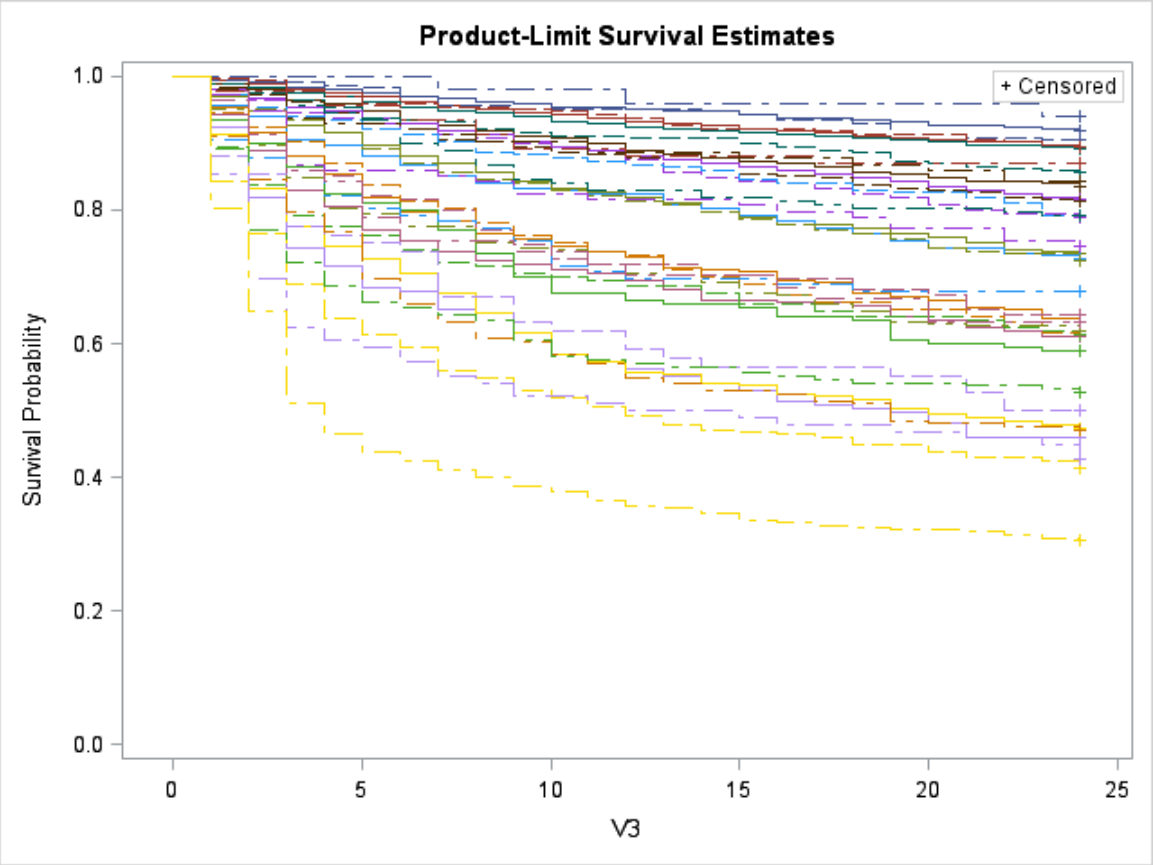
$$\hat{\pi}(z_2) = \frac{\exp(-2{,}4996)}{\exp(-2{,}4996) + 1} = 0{,}0759$$

.

Portanto, nota-se que a probabilidade de pagamento em até vinte e quatro meses dos clientes com o Perfil 1 é de 66,63%, enquanto para os com o Perfil 2 é de 7,59%.

Quanto ao modelo de mistura com fração de inadimplentes, o componente logístico associado a esse modelo apresentou resultados idênticos aos do modelo de regressão logística que acabamos de discutir. Contudo, o ganho com o modelo de mistura com fração de inadimplentes é que além das estimativas das probabilidades de pagamento para cada perfil de cliente, podem ser estimados, a partir do componente de sobrevivência $S(t|\boldsymbol{x})$ e também da função $S_p(t|\boldsymbol{x}, \boldsymbol{z})$, os tempos em que esses pagamentos se concretizaram para cada perfil, ou seja, é possível estimar a velocidade em que os pagamentos foram realizados. No geral, as curvas de sobrevivência $S_p(t|\boldsymbol{x}, \boldsymbol{z})$ estimadas para os 36 perfis de clientes ficaram conforme mostrado na Figura 7, sendo que a última curva em amarelo corresponde ao Perfil 1 de clientes e a primeira curva em azul, ao Perfil 2 de clientes, citados anteriormente.

Figura 7 – Curvas estimadas de $S_p(t|\boldsymbol{x},\boldsymbol{z})$ para os trinta e seis perfis de clientes



Fonte: O autor (2017).

Para aprofundar a análise dos dois perfis de clientes ao longo do tempo de acompanhamento (24 meses), foram obtidas as estimativas das sobrevivências $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ para cada tempo (t = 0 a 24 meses), como mostrado nas Tabelas 10 e 11. Além disso, a representação gráfica das curvas $S(t|\boldsymbol{x})$ de cada perfil também foi analisada com o intuito de se verificar, para cada um deles, a velocidade em que ocorrem os pagamentos. A Figura 8 apresenta as respectivas curvas para os perfis de clientes 1 e 2.

Tabela 10 – $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ para os clientes com Perfil 1

Tempo	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$S(t\|\boldsymbol{x})$	1	0,76	0,61	0,46	0,39	0,32	0,26	0,21	0,18	0,14	0,11	0,10	0,08	0,06	0,05	0,04	0,03	0,02	0,01	0,01	0,006	0,003	0,001	0,0003	0
$S_p(t\|\boldsymbol{x},\boldsymbol{z})$	1	0,84	0,74	0,64	0,59	0,54	0,50	0,47	0,45	0,43	0,41	0,40	0,38	0,37	0,36	0,36	0,35	0,35	0,34	0,34	0,33	0,33	0,33	0,33	0,33

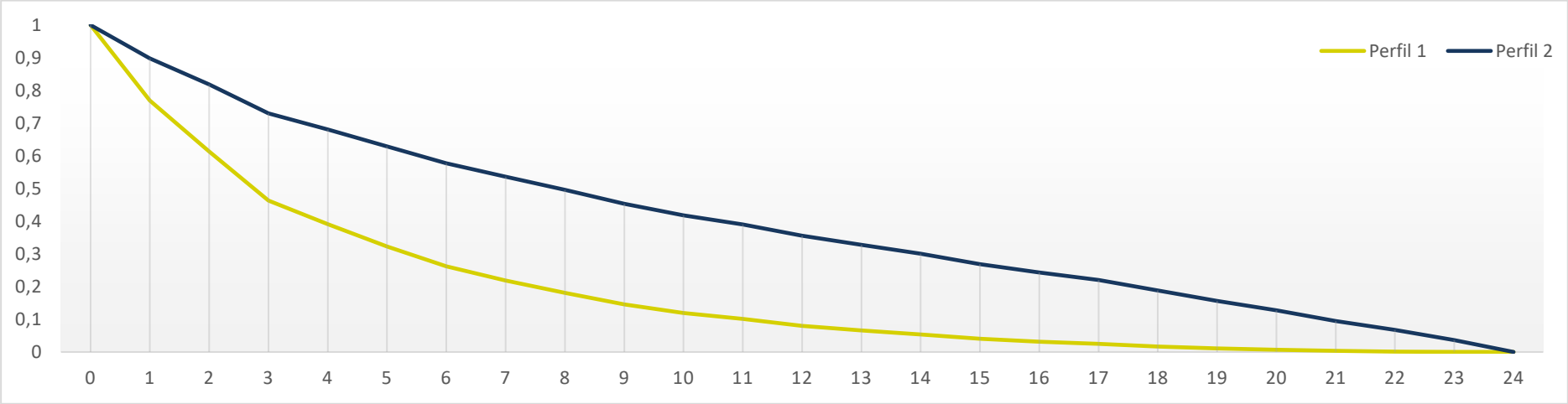
Fonte: O autor (2017).

Tabela 11 – $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ para os clientes com Perfil 2

Tempo	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$S(t\|\boldsymbol{x})$	1	0,90	0,82	0,73	0,68	0,63	0,58	0,54	0,50	0,45	0,42	0,39	0,36	0,33	0,30	0,27	0,24	0,22	0,19	0,16	0,13	0,09	0,07	0,04	0,00
$S_p(t\|\boldsymbol{x},\boldsymbol{z})$	1	0,99	0,99	0,98	0,98	0,97	0,97	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,94	0,94	0,94	0,94	0,94	0,93	0,93	0,93	0,93	0,92

Fonte: O autor (2017).

Figura 8 – $S(t|\boldsymbol{x})$ para os clientes com Perfil 1 e Perfil 2 em função do tempo t, com t entre 0 e 24 meses



Fonte: O autor (2017).

A partir das estimativas mostradas nas Tabelas 10 e 11, é possível notar que os clientes com o Perfil 1 não somente apresentaram probabilidade de pagamento das dívidas ($\hat{\pi}(\boldsymbol{z}_1) = 1 - \hat{S}_p(t = 24\,|\boldsymbol{x}_1, \boldsymbol{z}_1) \approx 0{,}67$) maior do que a dos clientes com Perfil 2 ($\hat{\pi}(\boldsymbol{z}_2) = 1 - \hat{S}_p(t = 24\,|\boldsymbol{x}_2, \boldsymbol{z}_2) \approx 0{,}08$), como também que os pagamentos efetuados pelos clientes com o Perfil 1 ocorreram bem mais rápido do que os do Perfil 2, como mostra as curvas $S(t|\boldsymbol{x})$ na Figura 8.

Uma forma simples de se fazer esta comparação, é fixar um ponto de corte de bons pagadores e observar o tempo em que este percentual é atingido. Esse ponto de corte é, em geral, definido pelas políticas das instituições financeiras, sendo que, neste trabalho, foi fixado em oitenta por cento da população. Para o Perfil 1, nota-se que este ponto de corte já é atingido entre os meses sete e oito (Figura 8). Em compensação, para o Perfil 2, tal ponto de corte é atingido apenas entre os meses dezessete e dezoito.

Por fim, para finalizar a análise, a Tabela 12 mostra todas as estimativas obtidas a partir do modelo de regressão logística e do modelo de mistura com fração de inadimplentes ao longo do tempo, para cada combinação das covariáveis.

Tabela 12 - Estimativas obtidas a partir do modelo de regressão logística e do modelo de mistura com fração de inadimplentes ao longo do tempo, para as 36 combinações (estratos) das categorias das covariáveis

Estrato	$S(t\|\boldsymbol{x})$ $t = 8$	$t = 16$	$t = 24$	$S_p(t\|\boldsymbol{x},\boldsymbol{z})$ $t = 8$	$t = 16$	$t = 24$	$\hat{\pi}(\boldsymbol{z})$	Tempo t (em meses) estimado para se ter 80% de pagadores
1	50%	24%	0%	96%	94%	92%	8%	18
2	49%	24%	0%	95%	92%	90%	10%	18
3	46%	21%	0%	93%	90%	87%	13%	17
4	48%	23%	0%	92%	88%	84%	16%	18
5	47%	22%	0%	89%	84%	80%	20%	17
6	45%	20%	0%	86%	79%	74%	26%	16
7	43%	18%	0%	85%	79%	74%	26%	16
8	42%	18%	0%	82%	74%	68%	32%	16
9	40%	15%	0%	77%	67%	61%	39%	15
10	32%	10%	0%	72%	63%	59%	41%	12
11	32%	10%	0%	67%	57%	52%	48%	12
12	29%	8%	0%	60%	49%	44%	56%	11
13	45%	24%	0%	96%	94%	92%	8%	18
14	45%	18%	0%	96%	94%	92%	8%	18
15	46%	21%	0%	93%	89%	86%	14%	17
16	48%	22%	0%	91%	87%	83%	17%	18
17	47%	22%	0%	89%	83%	79%	21%	17
18	45%	20%	0%	85%	78%	73%	27%	16
19	43%	18%	0%	85%	78%	73%	27%	16
20	42%	18%	0%	81%	73%	67%	33%	16
21	39%	15%	0%	76%	66%	60%	40%	15
22	28%	7%	0%	71%	61%	58%	42%	12

Continuação Tabela 12 - Estimativas obtidas a partir do modelo de regressão logística e do modelo de mistura com fração de inadimplentes ao longo do tempo, para as 36 combinações (estratos) das categorias das covariáveis

Estrato	$S(t\|\boldsymbol{x})$ $t = 8$	$t = 16$	$t = 24$	$S_p(t\|\boldsymbol{x},\boldsymbol{z})$ $t = 8$	$t = 16$	$t = 24$	$\hat{\pi}(\boldsymbol{z})$	Tempo t (em meses) estimado para se ter 80% de pagadores
23	27%	5%	0%	68%	53%	50%	50%	12
24	29%	8%	0%	59%	47%	43%	57%	11
25	42%	24%	0%	93%	91%	88%	12%	23
26	37%	12%	0%	91%	87%	85%	15%	17
27	35%	12%	0%	87%	83%	81%	19%	14
28	36%	15%	0%	85%	79%	77%	23%	15
29	36%	12%	0%	81%	75%	71%	29%	15
30	33%	11%	0%	76%	68%	65%	35%	12
31	31%	9%	0%	75%	68%	64%	36%	12
32	26%	9%	0%	69%	61%	57%	43%	12
33	28%	8%	0%	64%	54%	50%	50%	11
34	21%	4%	0%	59%	50%	48%	52%	9
35	20%	4%	0%	53%	43%	40%	60%	9
36	18%	3%	0%	45%	35%	33%	67%	8

Fonte: O autor (2017).

5 CONSIDERAÇÕES FINAIS

Técnicas estatísticas na área de cobrança trazem ganhos financeiros consideráveis para as empresas que trabalham com grande volume de empréstimos. Nesse trabalho, o modelo de regressão logística foi considerado como a técnica que é usualmente aplicada nos bancos de dados da área financeira/cobrança quando se tem interesse em modelar risco associado a pagamentos. Esse modelo serviu de referência para a comparação com o modelo de mistura com fração de inadimplentes, considerado aqui como uma alternativa. Ambos os modelos ajustados apresentaram ajustes satisfatórios aos dados analisados e se mostraram bastante eficientes na discriminação entre clientes bons e maus.

As variáveis explicativas que apresentaram efeito significativo foram: *atraso inicial, percentual de contratos em atraso, e percentual de restritivos baixados.* Em relação ao modelo de regressão logística, foi possível estimar a probabilidade de cada cliente se tornar bom pagador ao final do período de acompanhamento (vinte e quatro meses), o que auxiliou na definição de alguns perfis de clientes que são de interesse da instituição financeira. Dentre esses perfis, podem ser citados os dos clientes que apresentaram maior e menor probabilidade de pagamento. O conhecimento desses perfis é extremamente importante para a instituição financeira no sentido de elaborar diferentes estratégias de cobrança de acordo cada perfil, pois se a cobrança dos clientes inadimplentes for realizada no momento adequado e com a técnica apropriada, a empresa além de recuperar o montante emprestado à credito, consegue também manter seu relacionamento com o cliente.

Em relação ao modelo de mistura com fração de inadimplentes, além das informações obtidas a partir do modelo de regressão logística (dado que o componente logístico do modelo de mistura com fração de inadimplentes foi igual à regressão logística ajustada), obteve-se um ganho de informação em relação ao tempo em que os pagamentos ocorreram de acordo com seus perfis (características). Por exemplo, os clientes que têm uma baixa probabilidade de pagamento e, ainda, os que pagam após certo tempo, delimitado por um ponto de corte (neste trabalho definido como 80%), poderão estar elegíveis para a venda de carteiras, trazendo grandes benefícios à instituição, tendo em vista que à medida que o atraso aumenta, o valor de venda diminui. A partir do modelo de mistura com fração de inadimplentes, foi possível identificar antes estes clientes. Em média, após o período observado neste trabalho (vinte e quatro meses), os contratos passam a valer cinco por cento para uma possível venda. Se detectarmos antes os contratos para a

venda, não é necessário esperar vinte e quatro meses para se efetuar uma venda, esta pode acontecer antes havendo, consequentemente, um aumento no valor total dos contratos. Essas informações ao longo do tempo podem ser bastante relevantes e úteis para a definição de estratégias de cobrança mais enérgicas e dinâmicas por parte da empresa em função dos perfis dos clientes, pois a medida que se consegue identificar o tempo de pagamento dos clientes, estratégias e vendas de carteiras mais específicas podem ser feitas devido ao ganho de informação.

Portanto, pode-se concluir que o modelo com fração de inadimplentes se caracteriza como uma boa alternativa ao modelo de regressão logística, seja para a identificação de bons e maus clientes com uma janela de observação de vinte e quatro meses, seja para a identificação do tempo em que os bons pagadores possivelmente irão pagar suas dívidas, sugerindo, assim, uma possível venda de carteira antecipada, trazendo grandes benefícios às instituições financeiras.

REFERÊNCIAS

BERKSON, J. Application to the logistic function to bio-assay. *Journal of the American Statistical Association*, v. 39, n. 227, p. 357-365, 1944.

BERKSON, J.; GAGE, R Survival cure for cancer patients following treatment. *Journal of the American Statistical Association*, v. 47, p. 501-515, (1952).

CAMPOS JÚNIOR, N. (2003). *Sua Excelência*, o *devedor*. Disponível em: < https://www.equifax.com.br >. Acesso em: 06 de nov. 2017.

CHERRY, Richard T. *Introdução a Administração Financeira*: tradução de Vera Maria Conti Nogueira e Danilo A. Nogueira. São Paulo: Atlas, 1796.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Editora Blucher, 2006. 392 p.

CORBIÈRE, F.; JOLY, P. A SAS macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, v. 83, n.2, p. 173- 180, 2007.

COX, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society*. Series B. v. 34, p. 187-220. 1972. Disponível em: < http://hydra.usc.edu/ pm518b/literature/cox-72.pdf >. Acesso em: 12 jun. 2017.

DIXON, S. N.; DARLINGTON, G. A.; DESMOND, A. F. A competing risks model for correlated data based on the sub distribution hazard, *Lifetime Data Analysis*. Boston, v. 17, p. 473-495, 2011.

EUDES, A.M.; TOMAZELLA, V.L.D.; CALSAVARA, V.F. Modelagem de sobrevivência com fração de cura para dados de tempo de vida weibull modificada. *Rev. Bras. Biom*., São Paulo, v. 30, n. 3, p. 326-342, 2012.

EXPERIAN, S. *Análise de inadimplentes no Brasil em 2017*. Agosto de 2017. Disponível em: < https://www.serasaexperian.com.br/>. Acesso em: 07 de nov. 2017.

FERREIRA, Aurélio Buarque de Holanda. *Dicionário Aurélio Básico da Língua Portuguesa*. Rio de Janeiro: Nova Fronteira, 1998.

GIOSA, Livio A. *Terceirização*: uma abordagem estratégica. 5ª Ed. São Paulo Editora Pioneira,1997.

GRANZOTTO, D.C.T; LOUZADA-NETO, F; PERDONÁ, G.S.C. Modelos de sobrevivência com longa duração: uma aplicação a grandes bancos de dados financeiros. *Revista Brasileira de Biometria*, v. 24, n. 4, p.102-116, 2010.

HANREJSZKOW. A.; STROMBERG. E. *Regressão logística e modelo de mistura em um estudo sobre clientes inadimplentes de uma empresa de telecomunicações*. Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Federal do Paraná, 2013.

HOJI, M. *Administração Financeira*. 4º edição. Editora Atlas. São Paulo – 2003.

HOSMER, D. W; LEMESHOW, S. *Applied Logistic Regression*. New York: John Wiley & Sons, Inc., 2000.

KAPLAN, E.L; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, p. 457-81, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival analysis: techniques for censored and truncated data*. 2. ed. New York: Springer, 2003. 536 p.

KUK, A.Y.C.; CHEN, C.H. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, Oxford, v. 79, p. 531-541, 1992.

LEONI, Geraldo; LEONI, Evandro Geraldo. Cadastro. *Crédito e Cobrança*. 2. ed. São Paulo: Atlas, 1997.

MARIANI, F. *Análise e implementação de estratégias de cobrança como forma de redução e controle da inadimplência de uma operadora de planos de saúde da cidade de Caçador/SC*. 2008. Monografia (Administração). Universidade do Contestado, Caçador, 2008.

QUIDIM, I. L. *Análise de sobrevivência com fração de fidelizados*: uma aplicação na área de marketing. Dissertação (Mestrado em Estatística) São Paulo: IME - Instituto de Matemática e Estatística, Universidade de São Paulo, 2005.

R CORE TEAM. *R*: A language and environment for statistical computing. Viena, Áustria, 2015. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

ROCHA F. C. *A inadimplência de crédito no setor bancário brasileiro*: um estudo de caso. Monografia (Graduação em Economia). Florianópolis: Universidade Federal de Santa Catarina, 2010.

ROSENBERG, E.; GLEIT, A. Quantitative Methods in Credit Management: A Survey. *Operations Research*, v. 42, n. 4, p. 589-613, 1994.

RUSH, M. How to select the best predictor variables. Using *SAS enterprise guide.* <http://www.sascommunity.org/mwiki/images/2/20/How_to_Select_the_Best_Variables.pdf > Acesso em: 08 jun. 2017.

SAS/STAT© Software: Enterprise Guide, 7.1 Copyright, SAS Institute Inc. Cary, NC, USA, 2016.

SILVA, Jose Pereira da. *Gestão e análise de risco de crédito*. São Paulo: Atlas, 1998.

SIMONSEN, Mário H. Cinqüenta anos de Teoria Geral do Emprego. *Revista Brasileira de Economia*. v.40, n.4, p.301-34, out.-dez, 1986.

TOMAZELA, S.M.O. *Avaliação de desempenho de modelos de Credit Score ajustados por Análise de Sobrevivência*. Dissertação de Mestrado. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo, 2007.

WALD, A. Tests of Statistical Hypotheses concerning Several Parameters when the number of Observations is Large, *Trans. Amer. Math. Soc.*, v. 54, p. 426-482, 1943.

APÊNDICES

APÊNDICE A – Ajuste do modelo de mistura com fração de inadimplente como auxílio da macro SAS: "%PSPMCM"

```
%pspmcm(DATA=PF_VAR_2_SAMPLE_MODEL_dummy,ID=V821,CENSCOD=V2,TIME=
V3, VAR= D_1(IS, 1) D_2(IS, 0)  FAIXA1_0(IS, 0)  FAIXA1_1(IS, 0)
FAIXA1_2(IS, 0) PERC_BAIXA_121_0(IS, 0) PERC_BAIXA_121_1(IS, 0),
                        INCPART=logit,
                        SURVPART=cox,
                        TAIL=zero , SU0MET=pl,
                        FAST=Y,BOOTSTRAP=N,
                        NSAMPLE=2000, STRATA=,
                        MAXITER=200,CONVCRIT=1e-5, ALPHA=0.05,
                        BASELINE=Y,
                        BOOTMET=ALL,
                        JACKDATA=,
                        GESTIMATE=Y,
                    SPLOT=Y,
                    PLOTFIT=Y);
run;
```

APÊNDICE A2 – Criação de variáveis *dummy* com o auxílio de macro SAS;

```
%macro dummy(
   data=_last_ ,     /* name of input dataset                  */
   out=&data,        /* name of output dataset                 */
   var= ,            /* variable(s) to be dummied              */
   base=_last_,      /* base category                          */
   prefix = D_,      /* prefix for dummy variable names        */
   format =,         /* format used to categorize variable     */
   name  = VAL,      /* VAL: variable names are D_value        */
   fullrank=1        /* Eliminate dummy for baseline category? */
   );

%dummy (data = tcc.EDS_VC_PF_VAR_2_SAMPLE_MODEL,
out=PF_VAR_2_SAMPLE_MODEL_dummy, var = FX_CONTR_ATR1 FAIXA1
PERC_BAIXA_12 V655 temp_rel);
```

APÊNDICE A3 – Código para o ajuste do modelo de regressão logística;

```
proc logistic data=PF_VAR_2_SAMPLE_MODEL_DUMMY2 plots=all
PLOTS(MAXPOINTS=NONE) OUT= LOGISTICO2 ;
   class FAIXA2 FX_CONTR_ATR  PERC_BAIXA_12(ref='0') ;
   model V2(event='1') =  FX_CONTR_ATR  PERC_BAIXA_12 FAIXA2
         /selection=forward expb ;
run;
```