UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Utilizando análise de sobrevivência no crédito: um modelo de *behavioral scoring*

Autor: Jonas de Souza Pacheco
Orientadora: Prof.ª Dra. Lisiane Selau
Coorientador: Prof. Dr. Álvaro Vigo

Porto Alegre, Julho de 2013
Universidade Federal do Rio Grande do Sul

Instituto de Matemática
Departamento de Estatística

Utilizando análise de sobrevivência no crédito: um modelo de *behavioral scoring*

Autor: Jonas de Souza Pacheco

Monografia apresentada como requisito parcial para
obtenção do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Álvaro Vigo
Professora Dra. Lisiane Priscila Roldão Selau
Bel. Mariana Mizutani Ribeiro

Porto Alegre, Julho de 2013.

“*So long, and thanks for all the fish*”
Douglas Adams

Sumário

Este artigo será submetido a uma revista aplicada da área financeira

Resumo

Com o aumento do consumo de bens e serviços e da concessão de crédito, faz-se necessário o controle de risco. Tal medida visa evitar uma possível inadimplência maior que a suportada pelas instituições financeiras e, também, possibilita a geração de lucros. Sendo assim, a utilização de modelos de credit scoring mostra-se benéfica por diversos fatores, como consistência nas decisões, automatização na concessão, aumento no volume de análises e capacidade de monitorar e administrar o risco de uma carteira de crédito. Várias técnicas estatísticas podem ser utilizadas para a construção desses modelos e dentre elas destaca-se a análise de sobrevivência. Essa técnica apresenta a vantagem de poder prever quando ocorrerá o evento de interesse, diferente de outras técnicas tradicionalmente utilizadas, como a regressão logística. Nesse sentido, o objetivo do presente artigo é apresentar um método, composto de cinco etapas, para a construção de modelos comportamentais utilizando a técnica de análise de sobrevivência. A aplicação do método em uma situação real, com 25.858 clientes, proveniente de uma instituição financeira, comprova a sua eficácia e utilidade como suporte à construção de modelos comportamentais, o que permitiria a redução de custos para a instituição. Ao realizar-se a comparação entre o modelo atualmente utilizado pela empresa, proveniente de um bureau externo, e o modelo utilizando análise de sobrevivência, construído com o método proposto, têm-se uma melhora de aproximadamente 20% na discriminação entre bons e maus pagadores.

Palavras-Chave: análise de crédito, behavioral scoring, análise de sobrevivência.

1. Introdução

O risco é inerente às nossas escolhas e estamos sujeitos às consequências decorrentes delas. Quando decidimos, por exemplo, andar na calçada em vez de andarmos em meio aos carros, sabemos que assim temos uma menor probabilidade de sermos atropelados. A probabilidade muitas vezes é, contudo, um conceito de difícil compreensão. Sua teoria foi inicialmente desenvolvida por Fermat & Pascal [1], o que forneceu subsídio para a administração de risco.

Na área financeira, o controle de risco é o que permite a sobrevivência das instituições. Sem o controle do prejuízo e da inadimplência, seria insustentável a concessão de crédito e a obtenção de lucros. Para lidar com isso no âmbito da globalização dos mercados monetários e tratar da instabilidade financeira, foi assinado, em 1988, o 1º acordo de Basiléia [2]. O acordo hoje se encontra na 4ª versão, chamada de Basiléia III[1], e visa orientar e regrar as instituições financeiras a fim de garantir menor risco nas operações, aumentar a capacidade de absorver crises e a transparência no funcionamento e nas operações dessas instituições, estabelecendo regras como reservas mínimas de capital, controle de índice de liquidez e métodos de classificação de risco para clientes [3].

No mundo atual, a concessão de crédito é um dos pilares para o desenvolvimento econômico e, no Brasil, não é diferente. Dentre outras medidas tomadas pelos últimos governos brasileiros, o programa bolsa família e a política de microcrédito têm ajudado a tirar muitas pessoas da linha da pobreza, e, com isso, ajudado a realizar os sonhos de consumo dessas pessoas. Com o aumento do consumo por uma grande parte da população, aumenta também a necessidade da concessão de crédito, porém, por anos, a classe de baixa renda não foi o foco das instituições financeiras, que agora veem grande oportunidade de ganhos ao realizar empréstimos e venda de produtos para essa parcela da população.

[1] Os quatro acordos de Basiléia foram: Basiléia I (1988), Basiléia II (2004), Basiléia 2.5 (2009) e Basiléia III (2010).

O controle de risco associado à concessão de crédito está diretamente ligado ao sucesso das instituições financeiras. Um erro na decisão de concessão de crédito pode levar a, em apenas uma operação, perder o ganho obtido em dezenas de operações bem sucedidas [4]. Assim sendo, a previsão e o controle da inadimplência são fundamentais para que não seja necessária a utilização de altas taxas de juros, a fim de cobrir o prejuízo de créditos mal sucedidos.

A análise de crédito por parte de um analista, apesar de levar em conta a sua experiência, tem duas desvantagens. A primeira é ser subjetiva: se uma mesma solicitação de crédito for submetida a diferentes analistas, pode ser que eles cheguem a determinações diferentes, uns aprovando a solicitação e outros não. A segunda é não quantificar o risco das operações [5]. Sendo assim, a utilização de modelos de *credit scoring* mostra-se benéfica por diversos fatores, como consistência nas decisões, automatização na concessão, aumento no volume de análises e capacidade de monitorar e administrar o risco de uma carteira de crédito.

Os modelos de *credit scoring* dividem-se em dois grandes grupos. No primeiro estão os modelos de *application scoring,* construídos com o propósito de conceder crédito a novos clientes e que utilizam, basicamente, dados cadastrais destes, tais como idade, escolaridade, naturalidade e estado civil. No segundo grupo estão os modelos de *behavioral scoring,* desenvolvidos para administrar o crédito daqueles que já são clientes. Tal modelo utiliza, além dos dados cadastrais, os dados comportamentais destes na instituição ou em outras instituições do sistema financeiro, por exemplo, a apresentação de restritivos de crédito, a pontualidade em pagamentos ou o alto comprometimento da renda. Esse segundo grupo tende a fornecer modelos com maior poder de discriminação entre bons e maus pagadores que os do primeiro grupo, por incorporar um número maior de variáveis e com melhor qualidade de informações [5].

Muitas técnicas de análise estatística são utilizadas para a construção dos modelos de *credit scoring*; entre elas, regressão logística, análise discriminante, redes neurais, algoritmos genéticos e, recentemente, análise de sobrevivência [6]. Na grande maioria, as técnicas utilizadas para a construção dos modelos

apresentam uma resposta simples como resultado final, a probabilidade de o cliente se tornar um mau pagador, mas ignoram o fato de que, apesar de os dados utilizados na modelagem serem discretos, o relacionamento do cliente com a instituição é contínuo, a partir do seu cadastro. A técnica de análise de sobrevivência apresenta, então, uma vantagem sobre as outras técnicas, pois resulta em uma resposta temporal, podendo prever quando ocorrerá o evento de interesse, que, nesse caso, é quando o cliente se tornará, ou não, um mau pagador [7].

Nesse sentido, o objetivo deste trabalho é apresentar um método de construção de modelos de *credit scoring* a pessoas físicas, utilizando a técnica estatística de análise de sobrevivência. Será exemplificada a aplicação do método proposto para a construção de um modelo comportamental de crédito (*behavioral scoring*), visando auxiliar tanto a administração quanto a recuperação de crédito em uma instituição financeira de atuação nacional.

Este trabalho está estruturado em cinco seções. Primeiro, a introdução já apresentada nesta seção. A segunda seção trará a fundamentação teórica, versando sobre os modelos de *behavioral scoring* e sobre a técnica de análise de sobrevivência. Na terceira seção, será especificado o método proposto para a construção do modelo. A quarta seção apresentará os principais resultados da construção do modelo, bem como sua avaliação de desempenho. Na quinta e última seção, serão apresentadas as principais conclusões deste trabalho e a discussão sobre os resultados finais e sobre os possíveis benefícios à instituição da qual provém os dados.

2. Referencial Teórico

A análise de risco é um assunto muito delicado nas instituições financeiras, pois grande parte de seus ganhos são devidos a ela. Não se pode apenas recusar o crédito a todos que oferecem algum tipo de risco, já que é exatamente dos clientes que oferecem mais risco que se extrai um maior lucro, com pagamentos de multa e juros por atrasos. Torna-se, então, um problema decidir qual o ponto em que o cliente para de gerar lucros e começa a ser causador de prejuízos à instituição. Especialmente em se tratando de crédito em grande escala, a chamada concessão massificada, quando não se tem um analista disponível para analisar todas as propostas, é que são aplicados os modelos estatísticos de previsão de risco de crédito, como ferramenta para auxiliar a decisão.

Os modelos de *credit scoring* são aplicados de diversas formas e em diferentes etapas do ciclo financeiro. Além dos dois principais grupos, *application scoring*, utilizado para concessão de crédito, e *behavioral scoring*, utilizado na manutenção de crédito, existem outros modelos. O *collection scoring*, por exemplo, é aplicado na recuperação de crédito já em atraso, e o *attrition scoring*, por sua vez, calcula a probabilidade de um cliente cancelar um produto. Há ainda o *fraud scoring*, cujo objetivo é obter a probabilidade de uma operação ser uma fraude à instituição [8].

2.1. *Behavioral scoring*

Na construção de modelos para risco de crédito, a premissa básica é que o futuro será igual ao passado, ou seja, se um cliente com determinadas características teve um comportamento específico no passado, tanto bom quanto mau, é muito provável que um novo solicitante de crédito, com as mesmas características, apresente o mesmo comportamento no futuro [9].

Modelos de *behavioral scoring* são técnicas estatísticas utilizadas para classificar os solicitantes de crédito da instituição, classificando-os como bons ou maus pagadores [10]. Esses modelos são usados para auxiliar na concessão de crédito e na gestão de crédito e de risco, para os clientes que já tenham um relacionamento com a instituição.

Em se tratando da gestão de crédito, os modelos comportamentais auxiliam na gestão de limites pré-aprovados, na recalibragem de limites já concedidos e no suporte a campanhas de crédito. A grande variedade de aplicações desses modelos se dá pelo fato de serem dinâmicos, de modo que se tem mais informações sobre o cliente e se pode recalcular o risco desse tornar-se mau pagador.

A modelagem de *behavioral scoring* é uma ferramenta essencial, pois a avaliação de risco permite uma melhor classificação dos clientes. Ao recusar-se um maior número de clientes que provavelmente se tornariam maus pagadores e aceitar-se os clientes com maior chance de serem bons pagadores, tem-se uma maior eficiência no negócio. As instituições financeiras que possuem bons modelos comportamentais têm, em razão disso, vantagem em relação às demais, já que tendem a ficar menos vulneráveis aos riscos inerentes à concessão de crédito [11].

Para a construção de um modelo de *behavioral scoring, é* escolhido um ponto de observação e, a partir deste ponto, são coletados dados anteriores e posteriores a ele. O intervalo de tempo desses períodos em torno do ponto escolhido é objeto de discussões, sendo muito utilizados os que se encontram entre 6 e 24 meses. A utilização de períodos mais curtos de tempo resulta em dados atuais e isso minimiza a mudança comportamental dos clientes, natural em mercados instáveis [12]. Períodos mais longos têm a vantagem de englobarem todas as variações de mercado do ano, como mudanças de estações, feriados e recessos [5]. Para obter-se uma melhor exatidão do modelo, é preciso que se tenham dados precisos, que englobem as variações naturais do mercado e sejam atuais, de modo que acabam sendo mais utilizados os períodos intermediários, de 12 a 18 meses [13].

O intervalo de tempo anterior ao ponto de observação é denominado período histórico e serve como base para a classificação do indivíduo entre bom e mau pagador. Esse tempo é utilizado para analisar o comportamento do cliente. Já o intervalo posterior ao ponto de observação é chamado de período de performance. É nele que se define se o cliente tornou-se bom ou mau pagador, dependendo do critério definido no começo do estudo [5]. Como são utilizados

dados do passado dos clientes, é preciso que eles contenham informações não apenas confiáveis, mas também as mais atuais possíveis.

Ao longo do tempo, diversas técnicas foram utilizadas para a construção de modelos comportamentais, e sua utilização sempre esteve muito ligada à evolução computacional, dada a complexidade dos cálculos dos parâmetros dessas técnicas. Algumas das técnicas mais utilizadas são programação linear, algoritmos genéticos, análise discriminante, árvore de decisão, regressão logística e redes neurais [12]. Há ainda a análise de sobrevivência, uma técnica que possui duas grandes vantagens sobre as outras. A primeira, já citada no capítulo anterior, de apresentar uma resposta temporal; a segunda, de extrair informações de dados censurados.

2.2. Análise de Sobrevivência

A análise de sobrevivência, ou análise de sobrevida, é um conjunto de procedimentos para análise de dados que tem como resultado o tempo decorrido, desde um ponto inicial, até o evento de interesse ou censura. Na área financeira, por exemplo, esse evento pode ser o tempo desde a aquisição de um empréstimo até o atraso de uma das parcelas, ou então o tempo de financiamento de um automóvel até o segundo atraso consecutivo de parcelas. Neste trabalho, para a construção de um modelo de *credit scoring*, será utilizado o tempo desde a concessão de um produto bancário específico até a inadimplência do cliente. O método pode ser utilizado não somente para obter-se o resultado temporal até o desfecho especificado, mas também para fazer-se a comparação entre duas ou mais curvas de experiência de sobrevida de diferentes grupos. Pode-se, ainda, utilizá-lo para modelar as variáveis que influenciam na sobrevivência, tempo de adimplência, dos grupos estudados [14].

O método tem, diferentemente dos métodos estatísticos usuais, a capacidade de extrair informações de dados censurados e de ser utilizado em distribuições assimétricas. As censuras são classificadas em três diferentes categorias. A primeira, a censura à direita, ocorre quando o tempo entre o início do estudo e a ocorrência do evento é maior que o tempo de observação. A segunda, a censura à esquerda, existe se o tempo exato de ocorrência do evento é desconhecido, sabendo-se apenas que ele ocorreu antes de um determinado ponto no tempo.

A terceira, a censura intervalar, tem origem quando não se sabe o momento em que ocorreu o evento, apenas que ele aconteceu entre duas datas específicas [15].

São três as funções básicas utilizadas na análise de sobrevivência, a função densidade de probabilidade, a função de sobrevivência e a função risco, sendo todas elas utilizadas para descrever a variável aleatória tempo de sobrevida, expressa por $T \geq 0$ [16]. Interpretada como a probabilidade de um determinado indivíduo incorrer no evento em um intervalo de tempo instantâneo, a função densidade de probabilidade é expressa por

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}. \quad (1)$$

A função de sobrevivência representa a probabilidade de um determinado indivíduo sobreviver mais do que o tempo *t*, pré-estabelecido, sem ocorrer o desfecho de interesse, partindo do tempo inicial. A função é dada por

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{+\infty} f(u)du. \quad (2)$$

A função risco, ou função taxa de falha, representa a probabilidade instantânea, condicionada à sobrevivência de determinado indivíduo até um tempo *t*, de que tal indivíduo apresente o desfecho de interesse. Ela é muito eficaz para descrever o tempo de vida de indivíduos, isto é, o tempo até que eles incorram no evento de interesse. Essa função pode ser deduzida a partir da probabilidade de um indivíduo apresentar o evento no intervalo $[t_1, t_2)$,

$$P(t_1 \leq T < t_2) = S(t_1) - S(t_2). \quad (3)$$

Da equação (3), obtemos a probabilidade de um indivíduo apresentar o evento no intervalo $[t_1, t_2)$, dado que ele não apresentou esse antes do tempo t_1, fazendo a razão pelo tamanho do intervalo:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \quad (4)$$

A partir da equação (4), chegamos então na função risco. Para isso, reescrevemos o intervalo $[t_1, t_2)$ como $[t, t + \Delta t)$ e consideramos a variação de tempo Δt a menor possível, ou seja,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \quad (5)$$

Uma relação importante entre as funções apresentadas é que a função risco e a função de sobrevida, $\lambda(t)$ e $S(t)$, são inversamente proporcionais. Quando o risco aumenta, a probabilidade de sobrevivência diminui.

É essencial destacar que, para fazer os cálculos das três funções definidas anteriormente, é necessário utilizar algum método de estimação. Quando não se tem conhecimento prévio da distribuição do tempo de sobrevida, é preciso utilizar um método não-paramétrico. O cálculo empírico seria o mais simples, porém ele resulta na perda da informação dos dados censurados. Collett [16] sugere utilizar o modelo de riscos proporcionais proposto por Cox em 1972, cuja única suposição é que os riscos dos grupos a serem comparados são proporcionais.

2.3. Modelo de Regressão de Cox

O modelo de regressão de Cox, também conhecido como modelo de riscos proporcionais, foi construído com base em Kaplan & Meier [17]. A grande vantagem desse modelo de regressão em relação aos demais é não exigir que os dados tenham uma distribuição específica. O modelo é composto por uma parte paramétrica e outra não paramétrica, por isso é classificado como um modelo semi paramétrico [18].

A forma geral do modelo é dada por

$$h_i(t) = h_0(t) exp(\boldsymbol{\beta}_i \boldsymbol{x}_i), \quad (6)$$

onde t é o tempo, $\boldsymbol{x}$ é o vetor das p covariáveis explicativas do modelo $(x_1, x_2, ..., x_p)$, $\boldsymbol{\beta}$ é o vetor de coeficientes associados às p covariáveis explicativas $(\beta_1, \beta_2, ..., \beta_p)$ e $h_i(t)$ é a função taxa de falha para o i-ésimo indivíduo no tempo t. A parte não paramétrica do modelo, $h_0(t)$, é o risco basal. A parte paramétrica do modelo, $exp(\boldsymbol{\beta}_i \boldsymbol{x}_i)$, é uma função dos valores dos vetores $\boldsymbol{x}$ e $\boldsymbol{\beta}$ para o i-ésimo indivíduo.

É importante salientar que o modelo é conhecido como de riscos proporcionais, porque a razão entre a função taxa de falha de dois indivíduos independe do tempo, como pode ser visto a seguir,

$$\frac{h_0(t)exp(\boldsymbol{\beta x}_1)}{h_0(t)exp(\boldsymbol{\beta x}_2)} = exp(\boldsymbol{\beta x}_1 - \boldsymbol{\beta x}_2) = exp\big(\boldsymbol{\beta}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\big). \qquad (7)$$

Para ajustar o modelo de riscos proporcionais dado pela função (6) a um conjunto de dados, é preciso que os componentes da função sejam estimados. O componente paramétrico e o componente não paramétrico podem ser estimados separadamente, no entanto a estimação de $h_0(t)$ depende da estimação do componente linear da função, $exp(\boldsymbol{\beta}_i\boldsymbol{x}_i)$.

Para estimar o componente linear, podem-se utilizar métodos como o de máxima verossimilhança, o de verossimilhança parcial, o de verossimilhança aproximada e o de Cox para dados agrupados. O tempo é contínuo, porém, para a coleta dos dados, ele é discretizado em dias, semanas, meses ou outro intervalo, de acordo com o objetivo do estudo. Dessa forma, podem ocorrer eventos de interesse em um mesmo tempo, os chamados empates. O método de verossimilhança parcial é robusto e útil para a estimação quando existem dados censurados, porém, como se assume que a função risco taxa de falha é contínua, não seria possível a existência de empates [19].

Para tratar-se empates, utiliza-se uma aproximação do método de verossimilhança parcial. O método mais simples e mais amplamente utilizado foi proposto por Breslow [20]. A aproximação utiliza a informação do número de empates ocorridos em cada ponto de tempo e a soma dos coeficientes de cada covariável, para os indivíduos que tiveram o evento de interesse empatado com o de outros, de forma a ponderar os coeficientes quando ocorrerem empates. Para encontrar os estimadores dos coeficientes (β), é então utilizado algum método numérico, em geral o de Newton-Raphson, a fim de maximizar a função do componente linear [16].

A estimação da parte não paramétrica será baseada no método de máxima verossimilhança. Supondo que foram observados *r* tempos de sobrevida distintos para os indivíduos que apresentaram o evento de interesse, ordenam-

se esses tempos ascendentemente e, considerando que existem d_j empates no tempo t_j, a estimativa do risco basal é dada por

$$\hat{h}_0(t_j) = 1 - \hat{\xi}, \tag{8}$$

tal que $\hat{\xi}$ é a solução da seguinte equação,

$$\sum_{l \in D(t_j)} \frac{exp(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_l)}{1 - \hat{\xi}^{\exp(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_l)}} = \sum_{l \in R(t_j)} exp(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_l), \tag{9}$$

sendo $D(t_j)$ o conjunto de todos os indivíduos que tiveram empates no j-ésimo tempo, t_j, e $R(t_j)$ o conjunto de todos os indivíduos que estão sob risco de incorrerem no evento de interesse no j-ésimo tempo. Utilizando as estimativas das partes paramétrica e não paramétrica do modelo de riscos proporcionais, chega-se à estimativa da função risco para o i-ésimo indivíduo,

$$\hat{h}_i(t) = \hat{h}_0(t) exp(\widehat{\boldsymbol{\beta}}_i \boldsymbol{x}_i). \tag{10}$$

A partir dela, é possível chegar às estimativas da função densidade de probabilidade e da função de sobrevivência [21].

3. Método

O método para construção de modelos de *behavioral scoring* que será proposto está dividido em 5 etapas e foi baseado em Sicsú [5] e Selau & Ribeiro [6]. Dentro de cada etapa, estão no total 17 subitens, como pode ser visto na Figura 1.

1	Especificações do Modelo	1.1 Finalidade do modelo 1.2 Definição do público alvo 1.3 Definição do evento de interesse
2	Amostragem	2.1 Reconhecimento das variáveis disponíveis no sistema da instituição 2.2 Definição da data de referência e dos períodos histórico e de performance 2.3 Extração da base de dados 2.4 Validação das variáveis (missings, erros e tempo de informação) 2.5 Separação das amostras de desenvolvimento e de validação
3	Análise preliminar	3.1 Definição de novas variáveis 3.2 Análise bivariada 3.3 Agrupamento de categorias de variáveis 3.4 Geração de variáveis *dummies*
4	Construção da fórmula temporal	4.1 Seleção das variáveis preditoras 4.2 Cálculo da fórmula temporal 4.3 Verificação das suposições da técnica de análise de sobrevivência
5	Validação da fórmula temporal	5.1 Cálculo do poder discriminante 5.2 Comparação de desempenho entre as amostras de desenvolvimento e de validação

Figura 1 - Etapas do método

3.1. Especificações do modelo

Primeiro, é necessário definir qual será a finalidade do modelo a ser construído. Em função disso, deverão ser definidos diferentes públicos alvo, diferentes variáveis preditoras e diferentes eventos de interesse. Alguns dos tipos de modelos foram explanados na seção 2.

O público alvo precisa ser definido de forma a englobar o mercado que se deseja monitorar. Em modelos comportamentais, por exemplo, pode-se definir o público alvo como os clientes a mais de 12 meses com contas ativas. Já em um modelo de *collection scoring*, o público alvo poderia ser os clientes que apresentam operações em atraso. Uma separação muito indicada para públicos alvo é a de pessoa física e pessoa jurídica, pois tais apresentam, em geral, grandes

diferenças comportamentais, além de diferentes variáveis preditoras para a criação do modelo [5].

O evento de interesse é o que define a variável temporal do modelo. Na análise de sobrevida, o evento de interesse precisa ser definido de modo que se possa medir o tempo entre o início do estudo e a ocorrência do evento. Na concessão de crédito, geralmente se define como evento de interesse o momento em que o cliente se torna um mau pagador. A definição de mau pagador é diferente para cada empresa e depende do perfil de seus clientes [6].

3.2. Amostragem

Uma boa ideia para a identificação das variáveis demográficas é dada por Selau & Ribeiro [6]. Pega-se a proposta de crédito da empresa e observam-se as informações que devem ser preenchidas, sendo assim possível conhecer as variáveis cadastrais que podem ser selecionadas para o modelo. A identificação das variáveis comportamentais é feita observando-se as bases históricas da empresa. Essas variáveis normalmente são utilizadas após algum tratamento, como a média, o máximo ou a soma dos dados de certos períodos [12]. É importante verificar para que período se tem as variáveis, para que seja possível a utilização dessas, dependendo do período de observação que será escolhido. Uma explicação para a escolha da data de referência e para a escolha dos períodos histórico e de performance foi dada na seção 2.1. O processo é escolher um período histórico que seja atual e que capte as variações anuais, e um período de performance suficientemente extenso para identificar a ocorrência do evento de interesse.

A verificação de consistência das variáveis pode ser realizada em conjunto com a extração da base de dados. Evita-se, assim, o trabalho de extrair variáveis com muitas inconsistências, ou de extrair variáveis que, apesar de constarem no banco de dados da empresa, não foram preenchidas. Estatísticas descritivas são importantes ferramentas para a avaliação das variáveis, sendo muito úteis para avaliar a proporção de *missings* nas variáveis e para detectar dados incorretos e valores discrepantes, os *outliers* [5].

Com a coleta dos dados e a definição dos períodos, tem-se a informação de quais indivíduos apresentaram o evento de interesse e quais foram censurados. A censura dos dados pode ser por diversos motivos. O principal é o fim do estudo quando se tem clientes que acabaram de entrar nele, ou que tenham operações mais longas que o final do estudo, cujo desfecho não é conhecido. A censura pode ser também por abandono, o que pode ser considerado aqui como um refinanciamento ou como uma negociação da dívida com outra instituição financeira, o que hoje é uma opção no mercado de crédito brasileiro. Outro motivo para a existência de censura é o término do contrato do cliente sem a ocorrência do evento de interesse.

Após a extração dos dados e a validação das variáveis, é necessário selecionar uma amostra de indivíduos. É indicado por Hair *et al.* [22] que, para modelos multivariados, essa amostra possua ao menos 20 observações para cada variável preditora no modelo. A amostra selecionada é então separada nas amostras de desenvolvimento e de validação. A amostra de desenvolvimento costuma ter entre 70% e 80% das observações da amostra. Com base nela, são definidos os pesos de cada variável do modelo [6]. A amostra de validação, utilizada para testar o valor preditivo do modelo, possui o restante dos dados da amostra total.

3.3. Análise preliminar

As variáveis originais podem não ser as mais significativas, por isso é necessário fazer combinações de variáveis, de modo que essas resumam informações de uma melhor maneira [12]. Nessa etapa, é muito importante a experiência na área de aplicação do modelo, para que a construção dessas variáveis seja lógica e elas tenham significado.

A análise bivariada é realizada avaliando-se a relação de cada variável preditora com a variável que define o evento de interesse; no caso da área de crédito, a identificação de mau pagador. Lewis [23] apresenta uma maneira simples de verificar a capacidade preditora das variáveis, que apresenta bons resultados na modelagem. Deve-se calcular o risco relativo (RR) para cada categoria das variáveis e, conforme a escala apresentada na Figura 2, quanto mais distante de

um for o valor do risco relativo, maior poder discriminante terá a categoria da variável.

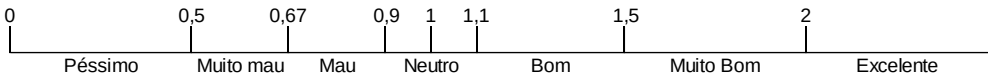

Figura 2- Classes de risco relativo (RR)

O risco relativo é calculado dividindo-se o percentual de bons pagadores pelo percentual de maus pagadores na categoria. Com os valores de RR para as categorias das variáveis pode-se, além de escolher quais variáveis podem vir a fazer parte do modelo final, selecionar categorias de variáveis que podem ser agrupadas. Essa etapa é importante para filtrar as variáveis, descartando as categorias que não acrescentam informação à variável, ou seja, que possuem um RR neutro, conforme a Figura 2. Para o agrupamento, são selecionadas categorias com risco relativo parecido, pois elas apresentam um comportamento semelhante.

Depois da pré-seleção de variáveis e do agrupamento de suas categorias, são geradas as variáveis indicadoras (*dummies)*. Essas variáveis apresentam dois valores, 0 para ausência e 1 para ocorrência, e, para n-1 categorias de cada variável, é criada uma variável *dummy*, onde n é o número total de categorias. A criação de variáveis *dummies* é utilizada para evitar problemas de não-linearidade das categorias das variáveis, porém é importante que não estejam todas as *dummies* de uma mesma variável no modelo [24].

3.4. Construção da fórmula temporal

Nessa etapa, já se tem as variáveis tratadas. Inicia-se, então, a seleção das variáveis que farão parte do modelo. Na maioria dos *softwares* estatísticos, estão implementados métodos de seleção automática. Os mais comuns são *backward elimination, forward selection* e *stepwise*. Segundo Corrar *et al.* [25], o mais indicado desses métodos é o *stepwise*, pois desconsidera variáveis com indícios de multicolinearidade, mantém no modelo apenas variáveis com significância estatística e não tem a limitação de avaliar apenas uma vez a variável, como acontece nos métodos *forward* e *backward*. O método stepwise realiza dois passos iterativamente. Primeiro, inclui no modelo a variável mais

estatisticamente significativa que não está no modelo. Segundo, remove do modelo todas as variáveis que deixaram de ser significativas com a inclusão da variável significativa.

Apenas a utilização de um método automático, por melhor que ele seja, não é o suficiente, é necessário verificar se as variáveis selecionadas fazem sentido para a área de estudo [14]. Outra análise que deve ser feita é a respeito dos sinais dos pesos atribuídos para cada variável. Clientes que possuem restritivos de crédito, por exemplo, são provavelmente maus pagadores, então o sinal dessa variável deve ser positivo.

Após a obtenção da fórmula preliminar, é preciso verificar as suposições do modelo utilizado. No caso do modelo de regressão de Cox, descrito na seção 2, a única suposição é a de riscos proporcionais. O método mais comumente utilizado consiste em uma verificação gráfica [16]. Para essa, é feito um gráfico para cada variável ao longo do tempo, utilizando o logaritmo da função taxa de falha acumulada. O que se espera nessa análise visual é que, em todos os gráficos, as curvas das categorias das variáveis distanciem-se proporcionalmente com o passar do tempo.

A linearidade das variáveis contínuas é essencial para que essas possam ser incluídas no modelo, de forma que é necessário testá-la. Um teste fácil é a utilização da transformação de Box-Tidwell. Essa transformação é utilizada aplicando-se o logaritmo à variável contínua e, multiplicando-se pela variável original, é então realizada uma nova regressão apenas com as variáveis contínuas e as transformadas, espera-se que nenhuma das variáveis seja significativa assim confirmando a linearidade [26].

3.5. Validação da fórmula temporal

O poder discriminante do modelo é calculado através de diversos indicadores. Na área de crédito, o mais conhecido é o índice KS de Kolmogorov-Smirnov. Esse indicador é calculado com base nas funções de distribuição acumulada dos bons e dos maus pagadores, sendo representado pela maior diferença entre essas funções [5]. Essa medida é representada por um percentual e, quanto maior tal percentual, melhor a discriminação entre os bons e os maus pagadores

obtida pelo modelo construído, conforme a Tabela 1, que apresenta valores de KS para modelos de *behavioral scoring*.

Tabela 1 - Valores referenciais para KS

Valor de KS	Behavioral scoring
KS ≤ 40%	Baixo
40% < KS ≤ 50%	Aceitável
50% < KS ≤ 60%	Bom
60% < KS ≤ 70%	Muito bom
KS > 70%	Excelente, valores pouco usuais

Fonte: Sicsú [5].

Segundo Sicsú [5], outro método muito utilizado é o cálculo da área sob a curva ROC, *receiver operating characteristic*, a qual se baseia na sensibilidade e na especificidade calculadas em relação a um escore X. A sensibilidade é a capacidade de identificar os maus créditos. A especificidade é a capacidade de identificar os bons créditos, ou seja, a proporção de bons créditos classificados corretamente. A vantagem dessa análise é que, diferente do KS, leva em consideração toda a distribuição e não apenas a maior diferença entre as curvas de bons e maus pagadores. É então feito um gráfico da sensibilidade por 1-especificidade para cada valor de X e, se a área sob essa curva for maior que 0,7, o modelo é aceitável.

O problema em avaliar o poder discriminante do modelo utilizando os dados que o geraram é que os indicadores acabam superestimados. Aplica-se, então, a fórmula temporal desenvolvida à amostra de validação, separada no início do estudo, e calculam-se os indicadores de desempenho, para avaliar o poder preditivo do modelo escolhido.

4. Resultados

As definições e os resultados da aplicação serão apresentados na mesma ordem descrita pelo método, de modo a demonstrar sua eficiência. Em todos os passos, foi utilizado o programa estatístico *SAS Enterprise Guide 4.3*, com o pacote *Office Excel 2013* como suporte. Dada a exigência de sigilo do nome da instituição financeira e de seus dados, quando necessário apresentar resultados, tanto os nomes das variáveis, quanto suas categorias, serão codificados.

4.1. Especificações do modelo

4.1.1. Finalidade do modelo

O modelo servirá de suporte à decisão de concessão de crédito para clientes que já tenham um relacionamento com a instituição financeira, sendo possível analisar o seu comportamento de pagamentos. Espera-se que, com a aplicação do método, seja possível construir um modelo que, quando implementado, auxilie na redução da inadimplência, porém sem comprometer a concessão a bons clientes.

4.1.2. Definição do público alvo

Optou-se por delimitar o modelo a um produto específico, o crédito pessoal, o qual é destinado a pessoas físicas. Como o modelo de *behavioral scoring* utiliza dados comportamentais, selecionou-se apenas os clientes que possuíam relacionamento com a instituição em pelo menos dois de três períodos anteriores à solicitação de crédito, definidos como: um a três meses antes da solicitação, quatro a seis meses antes da solicitação e sete a doze meses antes da solicitação. Para clientes que possuíam mais de uma operação de crédito pessoal, foi selecionada apenas a solicitação mais recente, para garantir o uso de um comportamento mais atual.

4.1.3. Definição do evento de interesse

O evento de interesse é o momento em que o cliente se torna um mau pagador. Diferente dos demais métodos comportamentais, na análise de sobrevida não se tem uma faixa intermediária, o evento ocorre ou não. Foi definido então que o

mau pagador é o cliente que fica inadimplente por mais de 90 dias. Dentro da instituição, considera-se que, a partir desse momento, o cliente apresenta baixa probabilidade de honrar a dívida.

4.2. Amostragem

4.2.1. Reconhecimento das variáveis disponíveis no sistema da instituição

Foram mapeadas tanto variáveis cadastrais; como estado civil, quantidade de dependentes, cidade de residência e sexo; quanto variáveis históricas, como saldo utilizado e atrasos em pagamentos. As variáveis históricas necessitam de uma maior atenção, pois nem sempre a variável bruta é a de maior significância para a análise, além de ser necessário observar-se para qual período histórico elas existem.

4.2.2. Definição da data de referência e dos períodos histórico e de performance

Decidiu-se utilizar os períodos histórico e de performance com 12 meses. Foi averiguado, então, qual a data de referência mais atual que se poderia ter considerando os períodos, sendo escolhido o mês de abril de 2012. Definiu-se, enfim, o período histórico, de abril de 2011 a março de 2012, e o período de performance, de maio de 2012 a abril de 2013.

4.2.3. Extração da base de dados

Foram extraídas inicialmente 13 variáveis cadastrais e 25 comportamentais, consolidadas em um banco de dados único. Das variáveis cadastrais mapeadas, três não foram extraídas por não estarem corretamente preenchidas, e uma por apresentar menos de 1% dos dados preenchidos. Considerando os filtros de público alvo e a data de referência, foi selecionado um total de 25.858 clientes, dos quais 976 se tornaram maus pagadores e o restante foi censurado, ou seja, não se tornaram maus pagadores até o final do estudo. Na construção do modelo, não foram incluídos novos clientes durante o estudo, assim como foram

selecionados apenas os clientes que não apresentavam inadimplência na data de referência, de forma que só existiram censuras à direita.

4.2.4. Validação das variáveis

A validação das variáveis históricas não indicou erros, principalmente pelos dados coletados serem de bases atualizadas mensalmente, as quais passam por processos de validação no setor de Tecnologia da Informação da instituição. Das variáveis cadastrais, quatro já haviam sido excluídas na etapa anterior, das demais, uma teve valores transformados em *missings*, pois continha cerca de 200 observações que não eram plausíveis.

4.2.5. Separação das amostras de desenvolvimento e de validação

As amostras foram separadas nas seguintes proporções: 70% para construção do modelo e 30% para validação. Foi utilizada amostragem aleatória simples para separar as amostras, cujo resultado encontra-se na Tabela 2.

Tabela 2 - Número e proporção de clientes por amostra

Amostra	Evento				Total	
	Censurado		Mau pagador			
Desenvolvimento	17.423	70,02%	678	69,47%	18.101	70,00%
Validação	7.459	29,98%	298	30,53%	7.757	30,00%
Total	24.882	100,00%	976	100,00%	25.858	100,00%

4.3. Análise Preliminar

4.3.1. Definição de novas variáveis

Foram construídas variáveis que resumiam as informações dos produtos utilizados pelo cliente, como número de produtos utilizados e exposição total de crédito. Foram ainda agrupadas variáveis por períodos de tempo, como o saldo total tomado nos últimos 3 meses, ou o saldo máximo vencido de 7 a 12 meses antes da data de referência.

4.3.2. Análise bivariada

A análise bivariada serviu tanto para identificar a relação comportamental entre as variáveis preditoras e o evento de interesse, quanto para fazer a pré-seleção das variáveis. A partir da utilização do RR, retiraram-se do estudo quatro variáveis comportamentais e uma cadastral, as quais não contribuíam na diferenciação dos maus pagadores. Foram, ainda, desconsideradas da análise algumas categorias que não continham um mínimo de 20 observações, como as categorias 3 e 5 da variável V1, mostrada na Tabela 3.

Tabela 3 - Avaliação de RR

V1				
Categoria	**Bom 0**	**Mau 1**	**RR**	**Faixa de RR**
1	11279	286	1,54	Muito Bom
2	688	50	0,54	Muito Mau
3	9	0	-	-
4	5442	342	0,62	Muito Mau
5	1	0	-	-

4.3.3. Agrupamento de categorias de variáveis

Apenas uma variável foi mantida contínua nos valores originais, as demais foram categorizadas ou separadas em faixas de valores de acordo com as faixas de RR. A construção de faixas para variáveis contínuas serve como auxílio para que a interpretação dos níveis da variável seja linear. Por exemplo, a cada aumento de uma faixa de renda, diminui em 2% a chance do cliente tornar-se um mau pagador. A variável 2, por exemplo, possuía um número pequeno de observações em alguns valores, causando uma mudança grande no RR, o que foi solucionado com a criação de faixas, conforme a Figura 3. Das variáveis categóricas, quatro possuíam apenas duas categorias.

4.3.4. Geração de variáveis *dummies*

Foram geradas variáveis *dummies* para todas as variáveis categorizadas que possuíam mais de duas categorias, resultando em 13 variáveis *dummies* de um total de quatro variáveis categóricas.

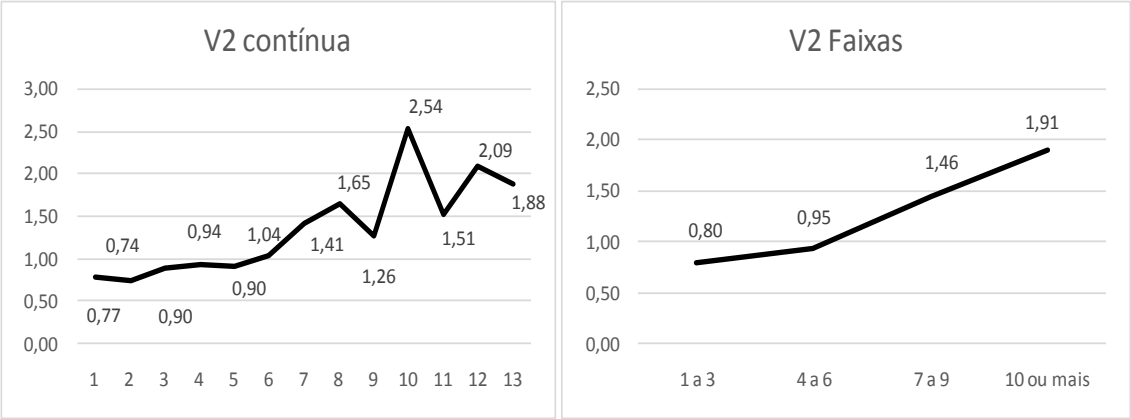

Figura 3 - RR da variável 2 contínua e em faixas

4.4. Construção da fórmula temporal

4.4.1. Seleção das variáveis preditoras

Foram avaliadas, no total, para o modelo de Cox e com o método *stepwise*, uma variável contínua, 24 variáveis contínuas separadas em faixas e 17 variáveis categóricas. Para isso foi utilizada a função PROC PHREG do SAS, levando em consideração a ocorrência, ou não, de evento e o tempo até o desfecho. As variáveis selecionadas, sob um nível de significância de 5%, estão na Tabela 4.

Tabela 4 - Modelo 1 de Análise de sobrevivência selecionado pelo método stepwise

Estimativas da Análise de Máxima Verossimilhança								
Variável	**GL**	**Estimativa da variável**	**Erro Padrão**	**Chi-Quadrado**	**Pr > Chi-Quadrado**	**Razão de Risco**	**IC 95% Razão de Risco**	
V3 (Contínua)	1	-0,008	0,003	5,779	0,016	0,992	0,985	0,998
V4 (Categórica)	1	0,741	0,141	27,780	<,0001	2,097	1,592	2,762
V1 (Categórica)	1	0,201	0,093	4,719	0,030	1,223	1,020	1,466
V5 (Categórica)	1	-0,362	0,132	7,525	0,006	0,697	0,538	0,902
V6 (Categórica)	1	1,704	0,236	52,028	<,0001	5,496	3,459	8,733
V7 (Categórica)	1	0,476	0,116	16,826	<,0001	1,610	1,282	2,021
V8 (Categórica)	1	1,647	0,225	53,364	<,0001	5,191	3,337	8,075
V9 (Faixas)	1	-0,128	0,019	43,291	<,0001	0,880	0,847	0,914
V10 (Faixas)	1	0,099	0,022	20,375	<,0001	1,104	1,058	1,153
V11 (Faixas)	1	-0,102	0,013	57,823	<,0001	0,903	0,879	0,927
V12 (Faixas)	1	-0,037	0,015	5,951	0,015	0,964	0,936	0,993
V13 (Faixas)	1	0,172	0,023	57,774	<,0001	1,188	1,136	1,242
V14 (Categórica)	1	0,675	0,090	56,894	<,0001	1,965	1,649	2,342

4.4.2. Cálculo da fórmula temporal

As variáveis do modelo resultante do método *stepwise*, foram validadas com base no RR da análise bivariada, conforme a Tabela 5. A variável 4 não condiz com o RR, pois com a ocorrência da variável aumenta a Razão de Risco, enquanto o RR indica melhora.

Tabela 5 - Verificação da Razão de Risco com base no RR

Variável	Razão de Risco	Risco Relativo
V3 (Contínua)	0,992	Aumenta
V4 (Categórica)	2,097	2,91
V1 (Categórica)	1,223	0,61
V5 (Categórica)	0,697	2,72
V6 (Categórica)	5,496	0,40
V7 (Categórica)	1,610	0,24
V8 (Categórica)	5,191	0,40
V9 (Faixas)	0,880	Aumenta
V10 (Faixas)	1,104	Diminui
V11 (Faixas)	0,903	Aumenta
V12 (Faixas)	0,964	Aumenta
V13 (Faixas)	1,188	Diminui
V14 (Categórica)	1,965	0,33

Optou-se por rodar novo modelo, sem a inclusão da variável 4 no método de seleção, o que ocasionou a exclusão da variável 12 e inclusão da variável 15. Os resultados desse novo modelo encontram-se na Tabela 6. As comparações com o RR também podem ser vistas na Tabela 6, demonstrando que, para todas as variáveis, a Razão de Risco condiz com o Risco Relativo.

Tabela 6 – Modelo 2 de Análise de sobrevivência selecionado pelo método stepwise

Estimativas da Análise de Máxima Verossimilhança									
Variável	GL	Estimativa da variável	Erro Padrão	Chi-Quadrado	Pr > Chi-Quadrado	Razão de Risco	IC 95% Razão de Risco		Risco Relativo
V3 (Contínua)	1	-0,009	0,003	6,443	0,011	0,991	0,984	0,998	Aumenta
V1 (Categórica)	1	0,219	0,092	5,637	0,018	1,245	1,039	1,492	0,61
V6 (Categórica)	1	1,738	0,236	54,201	<,0001	5,689	3,581	9,037	0,40
V7 (Categórica)	1	0,439	0,116	14,232	0,000	1,551	1,235	1,948	0,24
V8 (Categórica)	1	1,647	0,225	53,426	<,0001	5,194	3,339	8,079	0,40
V15 (Faixas)	1	-0,093	0,039	5,821	0,016	0,911	0,844	0,983	Aumenta
V9 (Faixas)	1	-0,139	0,019	56,054	<,0001	0,870	0,839	0,902	Aumenta
V10 (Faixas)	1	0,108	0,022	24,929	<,0001	1,114	1,068	1,162	Diminui
V11 (Faixas)	1	-0,107	0,013	68,072	<,0001	0,899	0,876	0,922	Aumenta
V13 (Faixas)	1	0,191	0,023	69,739	<,0001	1,211	1,158	1,266	Diminui
V14 (Categórica)	1	0,665	0,089	55,397	<,0001	1,945	1,632	2,317	0,33

4.4.3. Verificação das suposições da técnica de análise de sobrevivência

Primeiramente, foi verificada a linearidade de todas as variáveis contínuas. Para isso, utilizou-se a transformação de Box-Tidwell. Foi realizada a verificação no modelo 2 utilizando um nível de significância de 0,05, na qual todas as variáveis contínuas demonstraram-se lineares.

É necessário, também, analisar se os riscos mantêm-se proporcionais para todas as variáveis ao longo do tempo. Verificou-se que isso ocorre, pois o logaritmo da função taxa de falha para as categorias das variáveis distanciou-se proporcionalmente ao longo do tempo de estudo, como mostrado para duas variáveis pelos gráficos da Figura 4.

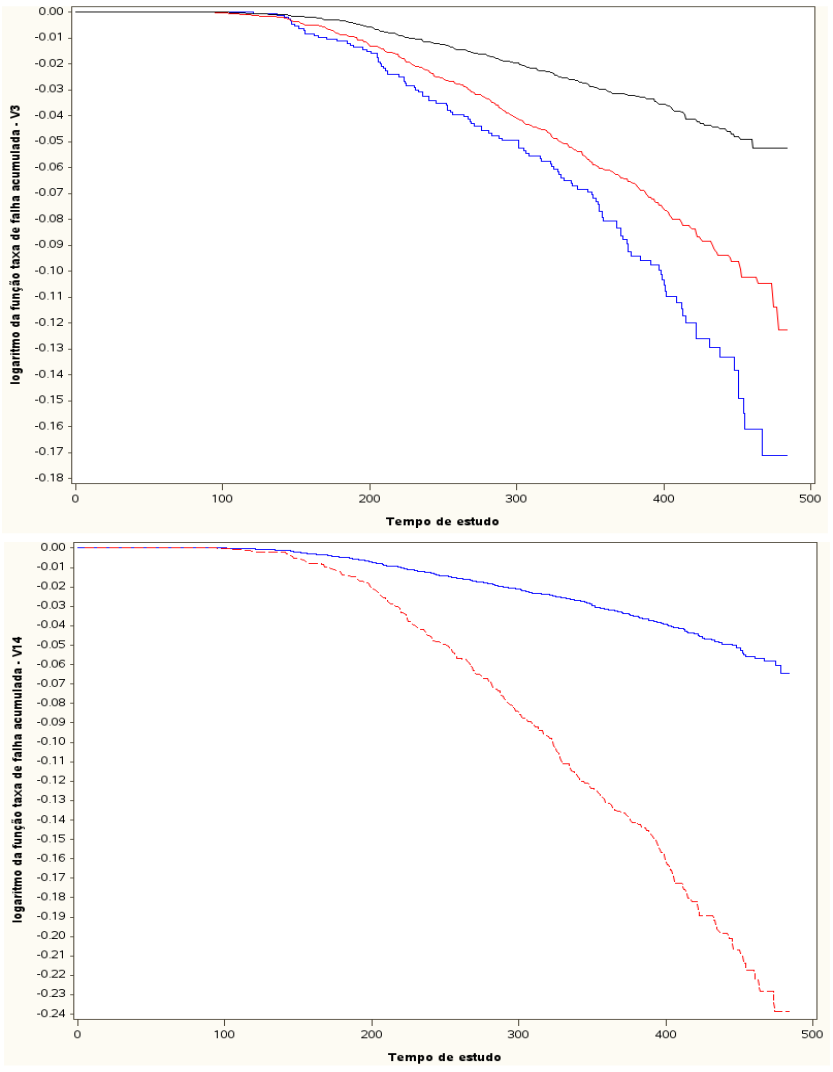

Figura 4 - Avaliação de riscos proporcionais das variáveis 3 (acima) e 14 (abaixo)

4.5. Validação da fórmula temporal

4.5.1. Cálculo do poder discriminante

A utilização do teste de KS resultou nos valores 60,23%, para a amostra de desenvolvimento, e 55,95%, para a amostra de validação. O cálculo da área sob a curva teve coeficientes calculados iguais a 0,8765, para a amostra de desenvolvimento, e 0,8567, para a amostra de validação.

4.5.2. Comparação de desempenho entre as amostras de desenvolvimento e de validação

O poder discriminante foi calculado tanto na amostra de desenvolvimento quanto na amostra de validação. Para a primeira amostra, o valor é superestimado, porém útil para ter-se uma ideia da capacidade de discriminação do modelo. Os resultados indicam que o modelo possui uma boa diferenciação entre os grupos de bons e maus pagadores. Foi importante a utilização da curva ROC, pois assim não se tem uma medida apenas da maior diferença entre as curvas, mas de toda a distribuição. O resultado dos testes na amostra de validação pode ser visto na Figura 5.

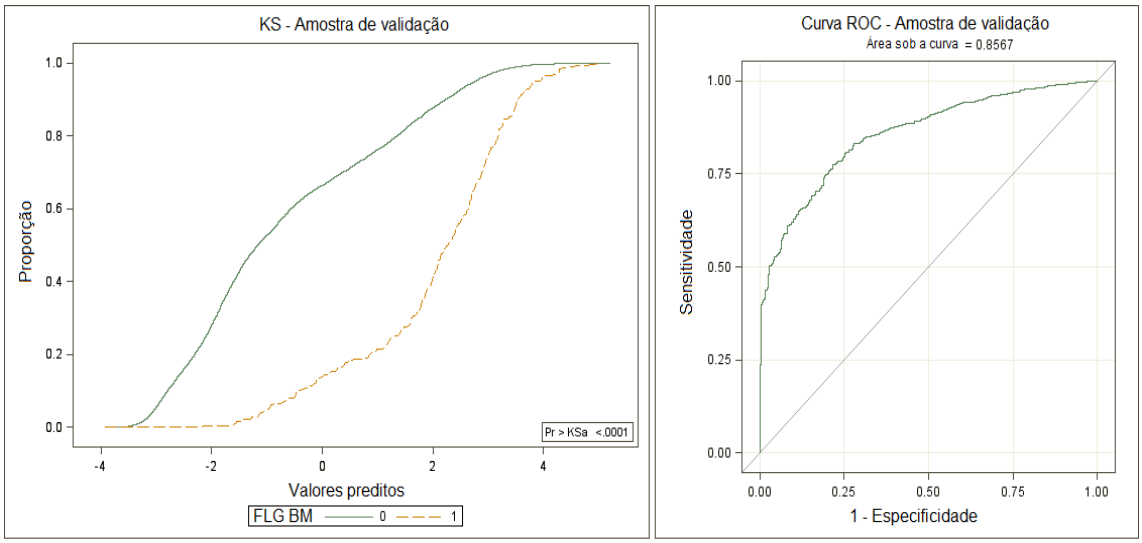

Figura 5 - Avaliação de desempenho

5. Conclusões

O objetivo desse artigo foi apresentar um método simples e de fácil compreensão para construção de modelos de *behavioral scoring*, utilizando para tal a técnica de análise de sobrevivência. De maneira a servir como suporte para analistas de crédito, em especial para aqueles sem uma formação estatística, foram introduzidos os conceitos básicos desse método de análise. Apesar de apresentar resultados interessantes, como a possibilidade de calcular a probabilidade, em todos os períodos de tempo estudados, de terem-se maus pagadores, o método ainda é pouco utilizado na área financeira.

Todas as etapas necessárias para a construção de um modelo foram explicadas e apresentadas com a aplicação do método a um banco de dados real. Algumas explicações teóricas foram necessárias tanto para o entendimento da técnica, como para expor a sua flexibilidade. A utilização do método produziu um modelo considerado bom, de acordo com os testes de desempenho realizados. Se comparado ao modelo para pessoa física de um *bureau* externo, com atuação nacional, o modelo produzido possui um desempenho superior em aproximadamente 20%, se comparadas as estatísticas de KS. Há indícios que essa melhora seja devida não apenas ao método utilizado, mas também à maneira com que a técnica de análise de sobrevivência lida com a variável resposta, considerando o tempo até a ocorrência e não apenas o status final do cliente. É importante levar em consideração que não se tem conhecimento das variáveis que fazem parte do modelo externo. Sendo assim, uma comparação entre regressão logística e análise de sobrevivência, ambas utilizando os dados internos da instituição, seria interessante para averiguar a diferença entre a capacidade de discriminação das técnicas utilizando o método proposto. Para a instituição, a utilização desse modelo pode significar uma economia de pelo menos vinte centavos mensais por cliente, dado que não seria necessária a compra das informações do modelo externo.

Uma medida pouco explorada no trabalho foi a função taxa de falha acumulada, pois, devido à baixa inadimplência do produto escolhido para aplicação e aos períodos muito longos para pagamento, a probabilidade de os clientes se tornarem maus pagadores foi de apenas 0,0140 em até 474 dias. A análise de

resíduos não foi considerada, pois o tamanho de amostra, em geral, é grande na área de crédito. Seria interessante a inclusão dessa análise em trabalhos futuros, utilizando os resíduos de Schoenfeld para dados censurados e Deviance para não censurados, de forma a auxiliar na utilização da técnica por instituições financeiras que não tenham um grande banco de dados à sua disposição. Para estudos futuros, a aplicação de outro método de pré-seleção e agrupamento, que não o RR, pode trazer melhoras, pois, apesar desse método ter resultado em um bom modelo, ele não leva em consideração o tempo, o que é uma característica importante da análise de sobrevivência.

Espera-se que com a aplicação do método para a análise de concessão de crédito, haja uma redução na inadimplência para produtos com alta inadimplência dentro da instituição. Com a capacidade da análise de sobrevivência de fazer previsão, do tempo até os clientes se tornarem maus pagadores, planejar um método de intervenção para evitar a inadimplência de possíveis maus pagadores, com integração com a área de cobrança.

Referências Bibliográficas

[1] FERMAT, PIERRE DE; PASCAL, BLAISE. “**The complete extant 1654 correspondence. Fermat and Pascal on Probability.”** Traduzido para o inglês. Disponível em: www.york.ac.uk/depts/maths/histstat/pascal.pdf. Acesso em 16 de Março de 2013.

[2] BB. Banco do Brasil. **Acordo de Basiléia**. Disponível em: http://www.bb.com.br/portalbb/page51,136,3696,0,0,1,8.bb?codigoNoticia=7724. Acesso em 29 de Março de 2013.

[3] BASEL COMMITTEE ON BANKING SUPERVISION (2010). **Basel III: A global regulatory framework for more resilient banks and banking systems**, BIS. Disponível em: http://www.bis.org/publ/bcbs189.pdf. Acesso em 22 de Março de 2013.

[4] STEINER, M. T. A., CARNIERI, C., KOPITTKE, B. H., STEINER NETO, P. J. **Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário**. Revista de Administração, São Paulo. p. 34. 1999.

[5] SICSÚ, A. L. **Credit scoring: desenvolvimento, implantação, acompanhamento**. São Paulo, Blucher. 2010.

[6] SELAU, L. P. R., RIBEIRO, J. L. D. **Uma sistemática para construção e escolha de modelos de previsão de risco de crédito**. Gestão Produção, v. 16, n. 3, p. 398-413, jul.-set. 2009.

[7] DINIZ, C.; LOUZADA, F. **Modelagem estatística para risco de crédito**. Minicurso no 20º SINAPE – Simpósio Nacional de Probabilidade e Estatística, João Pessoa – PB. 2012.

[8] THOMAS, C. L., EDELMAN, D. B., CROOK, J. N. **Credit scoring and its applications**. Society for Industrial and Applied Mathematics. 1987.

[9] SABATO, G. **Modelos de scoring de risco de crédito**. Revista Tecnologia de Crédito, v.68, p. 29-47, abril 2009.

[10] HAND, D. J., HENLEY, W. E. **Statistical classification methods in consumer credit scoring: a review.** Journal Royal Statistical Society. V. 160, p 523-541. 1997.

[11] GOUVÊA, M. A., GONÇALVES, E. B., MANTOVANI, D. M. N. **Aplicação de regressão logística e algoritmos genéticos na análise de risco de crédito**. Revista Universo Contábil, Blumenau, v. 8, n. 2, p. 84-102, abr./ jun., 2012.

[12] THOMAS, C. L. **A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers**. International Journal of Forecasting, v.16, n.2, p.149-172, 2000.

[13] LAWRENCE, D. **Handbook of consumer lending**. New Jersey: Prentice Hall. 1992.

[14] HOSMER JR, D. W., LEMESHOW. **Applied survival analysis: regression modeling of time to event data**. Wiley-Interscience, 1998.

[15] LEE, E. T., WANG, J. **Statistical methods for survival data analysis**. Wiley-Interscience, v. 476, 2003.

[16] COLLETT, D. **Modelling survival data in medical research**. 2ª ed. Boca Raton: Chapman & Hall/CRC, 2003.

[17] KAPLAN, EDWARD L., MEIER P. **Nonparametric estimation from incomplete observations**. Journal of the American Statistical Association, v. 53, p. 457-481, 1958.

[18] COX, DAVID R. **Regression models and life-tables** (com discussões). Journal of the Royal Statistical Society, Series B (Methodological), p. 187-220, 1972.

[19] COX, DAVID R. **Partial likelihood**. Biometrika, v. 62(2), p. 269-276, 1975.

[20] BRESLOW, N. **Covariance analysis of censored survival data**. Biometrics. v. 30, p. 89-100, 1974.

[21] KALBFLEISCH, J. D., PRENTICE, R. L. **Marginal likelihoods based on Cox's regression and life model**. Biometrika, v. 60, p. 267-278, 1973.

[22] HAIR, J. F., ANDERSON, R. E., TATHAM, R. L. **Análise multivariada de dados**. Bookman, 2007.

[23] LEWIS, E. M. **An introduction to credit scoring**. San Rafael: Fair, Isaac and Co., Inc. 1992.

[24] SUITS, DANIEL B. **Use of dummy variables in regression equations**. Journal of the American Statistical Association, v. 52, n. 280, p. 548-551, 1957.

[25] CORRAR, L. J., PAULO, E., DIAS FILHO, J. M. **Análise Multivariada: para cursos de Administração**, Ciências Contábeis e Economia. São Paulo: Atlas, 2007.

[26] BOX, G. E. P., TIDWELL, P. W. **Transformation of the independent variable**s. *Technometrics, v.* 4, n. 4, p. 531-550, 1962.