



Insper

Business and Economics  
Working Papers

BEWP 208/2014

Modelos de Risco de  
Crédito de Clientes:  
Uma aplicação a  
Dados Reais

Gustavo H. A. Pereira  
Rinaldo Artes

# Modelos de Risco de Crédito de Clientes: Uma aplicação a Dados Reais

Gustavo H. A.Pereira\*  
UFSCar

Rinaldo Artes †  
Insper

## Resumo

Modelos de *behavioural scoring* são geralmente utilizados para estimar a probabilidade de um cliente de uma instituição financeira que já possui um determinado produto de crédito se tornar inadimplente neste produto em um horizonte de tempo pré-fixado. Porém, frequentemente, um mesmo cliente tem diversos produtos de crédito em uma única instituição e os modelos de *behavioural scoring* geralmente tratam cada um deles de forma independente. Para facilitar e tornar mais eficiente o gerenciamento do risco de crédito, é interessante o desenvolvimento de modelos de *customer default scoring*. Esses modelos buscam estimar a probabilidade de um cliente de uma instituição financeira se tornar inadimplente em pelo menos um produto em um horizonte de tempo pré-fixado. Neste trabalho, são descritas três estratégias que podem ser utilizadas para o desenvolvimento de modelos de *customer default scoring*. Uma das estratégias é usualmente utilizada por instituições financeiras e as duas outras são propostas neste trabalho. As performances dessas estratégias são comparadas utilizando um banco de dados real fornecido por uma instituição financeira e um estudo de simulação de Monte Carlo.

*Palavras-chave:* credit scoring; customer scoring; equações de estimação generalizadas; regressão logística; risco de crédito.

## 1 Introdução

A concessão de crédito é uma das principais atividades bancárias. Para que ela seja uma atividade lucrativa para uma instituição financeira é importante que o banco efetue um eficiente gerenciamento do risco de crédito. Uma ferramenta fundamental neste processo são os modelos de *credit scoring*. Esses modelos têm como objetivo mensurar o risco de crédito associado a uma operação de crédito ou a um cliente. Tradicionalmente, os modelos de *credit scoring* são divididos em duas categorias: *application scoring* e *behavioural scoring*. O primeiro é utilizado para a decisão sobre a concessão de um produto para um cliente novo e o último para a avaliação do risco de operações

---

\*Departamento de Estatística da Universidade Federal de São Carlos, Brazil. E-mail: ghpereira@terra.com.br

†Insper Instituto de Ensino e Pesquisa, São Paulo, Brasil. E-mail: RinaldoA@insper.edu.br

já existentes. Thomas (2009), Anderson (2007) e Finlay (2012) descrevem em detalhes diversos aspectos relacionados a esses modelos e um sumário das contribuições nesta área é apresentado em Abdou e Pointon (2011).

Tradicionalmente, os modelos de *behavioural scoring* são utilizados para estimar a probabilidade de um cliente que já possui um determinado produto se tornar inadimplente neste produto em um horizonte de tempo pré-fixado. Esses modelos utilizam principalmente variáveis relacionadas ao comportamento de utilização do produto pelo cliente. Podem ainda ser utilizadas variáveis demográficas e informações do comportamento de crédito do cliente em outras instituições obtidas através de um *bureau* de crédito. Diversas técnicas podem ser usadas para o desenvolvimento desses modelos e a regressão logística é frequentemente utilizada (Thomas, 2010). Podem ainda ser desenvolvidos modelos de *behavioural scoring* para ajustar o tempo até ocorrência de inadimplência através de técnicas de análise de sobrevivência (Stepanova e Thomas, 2001, Cao, Vilar e Devia, 2009 e Sarlija et al., 2009).

Uma instituição financeira possui diversos produtos de crédito. Muitas vezes, porém, os modelos de *behavioural scoring*, tratam cada produto de forma independente dos demais (Thomas et. al., 2001). Porém, pode ser interessante para as instituições ter como foco o cliente e não os produtos do banco. O gerenciamento do risco de crédito baseado no foco no cliente traz inúmeras vantagens. Previne, por exemplo, a concessão de um novo produto ou o aumento de limite em um já existente, para os clientes com atraso ou *behaviour* score de alto risco em um outro produto. Permite ainda um melhor controle dos limites disponíveis totais e valores emprestados ao cliente, evitando que eles atinjam quantias maiores que a pessoa tem condição de pagar. Com o crescimento do foco no cliente, surgiu a preocupação em consolidar o risco de crédito do cliente em cada um dos produtos (dados pelos modelos de *behavioural scoring*) em uma única medida, dando origem aos modelos de *customer default scoring* ou simplesmente *customer scoring* (McNab e Winn, 2003). Tratam-se de modelos que objetivam ordenar os clientes quanto à probabilidade se tornar inadimplente em pelo menos um produto, dentro de um prazo pré-determinado. A grande vantagem dessa ferramenta é permitir uma visão global do risco do cliente, facilitando a criação de políticas de crédito mais adequadas para a instituição. Em um banco que possua, por exemplo, três modelos de *behavioural scoring* de produto, cada cliente possui um vetor de escores com três posições. Dessa forma, a instituição pode ter grande dificuldade em criar estratégias de gerenciamento do risco de crédito para cada um dos possíveis resultados do vetor de escores do cliente. A introdução do modelo de *customer scoring* facilita essa tarefa, pois substitui um vetor de três posições por uma única medida.

Segundo Thomas et. al. (2009, página 214), a maioria dos bancos utiliza atualmente modelos de *customer scoring*. Porém, há poucos trabalhos que tratam de modelos de *customer scoring*. Thomas et al (2001), Thomas (2009), Anderson (2007) e Finlay (2012) apresentam o objetivo desses modelos e os dois últimos comentam adicionalmente que frequentemente há dificuldades de implementação relacionadas ao armazenamento e processamento dos dados. McNab e Winn (2003) discutem rapidamente o conceito, as formas de desenvolvimento e suas componentes, as vantagens e as aplicações dos modelos de *customer scoring*. Já Groom e Gill (1998) discutem diversos aspectos

importantes que devem ser observados no desenvolvimento de um modelo de *customer scoring*. São discutidos os requisitos necessários para o seu ajuste, os tipos de variáveis que devem estar presentes no modelo, o modo de definição da variável resposta e o tamanho do histórico de comportamento de crédito que deve ser utilizado. Além disso, são sugeridas três estratégias de desenvolvimento e apresentadas as situações em que cada uma delas é a mais indicada. Porém, não são abordados aspectos técnicos de desenvolvimento dos modelos.

O fato de modelos de *customer scoring* terem sido pouco abordados na literatura, apesar de serem muito usados por instituições financeiras, parece estar relacionado com os mesmos motivos que os modelos de *behavioural scoring* são bem menos abordados na literatura do que os modelos de *application scoring*. Segundo Kennedy, et al. (2013), há dois motivos para a publicação de poucos artigos sobre *behavioural scoring*. Um deles é que a quantidade de dados necessária para seu desenvolvimento é bem grande e dificilmente uma instituição financeira disponibiliza um grande volume de dados para realização de trabalhos acadêmicos. O outro motivo é o interesse de instituições financeiras em manter sob sigilo modelos baseados em tantas diferentes variáveis. No caso dos modelos de *customer scoring* esses aspectos são ainda mais críticos, já que eles requerem ainda mais informação para serem desenvolvidos do que os modelos de *behavioural scoring*.

A grande dificuldade no desenvolvimento de um modelo de *customer scoring* está no fato da maioria dos indivíduos não possuir todos os produtos de crédito do banco. Mesmo agrupando-se os produtos em poucas famílias, ainda assim, muitos clientes podem não possuir pelo menos um contrato em cada um dos grupos. Dessa forma, o ajuste de um modelo de cliente diretamente a partir de todas as variáveis disponíveis não é possível, já que várias delas podem não ser observadas para um elevado número de clientes. Diante dessa dificuldade, pelo menos três estratégias podem ser utilizadas para contornar o problema. O desenvolvimento de modelo em duas etapas (Estratégia 1) é a solução usualmente utilizada (Groom e Gill, 1998). Neste artigo estarão sendo propostas duas outras: o ajuste de vários modelos simultâneos para o modelo de cliente (Estratégia 2) e a obtenção simultânea não só do modelo de *customer scoring* como também de vários modelos de *behavioural scoring* (Estratégia 3). Nessa última, pelo fato de se observar várias respostas em um mesmo indivíduo, há dependência entre as observações. Evidentemente essa dependência deve ser considerada na análise.

As estratégias requerem a estimação de vários modelos para respostas binárias. Devido a sua popularidade na área em estudo (Thomas, 2010), será utilizada, neste trabalho, a regressão logística para a obtenção desses modelos. Na estratégia 3, os modelos serão estimados por meio de equações de estimação generalizadas (GEE), já que elas permitem o tratamento da dependência entre as observações. A GEE foi introduzida por Liang e Zeger (1986) e Zeger e Liang (1986) para ajustar modelos de regressão longitudinais para variáveis respostas não normais. A técnica é detalhadamente descrita por Hardin e Hilbe (2013) e algumas contribuições e extensões propostas nesta área são apresentadas em Ziegler et. al. (1998) e Song (2007), por exemplo. Trabalhos na área de risco de crédito que utilizam a GEE incluem Hwang (2013) e Ilk et al. (2013).

O restante deste artigo está organizado da seguinte forma. A próxima seção discute estratégias para o desenvolvimento de um modelo de *customer scoring*, bem como o modelo estatístico associado

a cada uma delas. Na seção seguinte é apresentada a descrição do conjunto de dados reais fornecidos por uma instituição financeira para a realização deste trabalho. Em seguida, é apresentada uma aplicação das estratégias discutidas a partir desse conjunto de dados. São ajustados os modelos de cada uma das estratégias e comparadas suas performances. Um estudo de simulação de Monte Carlo com alteração da estrutura de correlação dos dados é apresentada na penúltima seção. As conclusões do trabalho são discutidas na seção final.

## 2 Metodologia

O Exemplo 1 será utilizado para facilitar a compreensão dos modelos associados a cada uma das estratégias. Nas seções 2.1 a 2.3 serão apresentados os modelos para esse exemplo. Na Seção 2.4 será feita a generalização dos resultados.

**Exemplo 1.** Suponha que os produtos de uma determinada instituição possam ser divididos em duas famílias. Admita ainda que cada cliente possua no máximo uma conta em cada uma das famílias de produtos em um instante  $t$ . Suponha também que se observe no período entre  $t - \epsilon$  e  $t$ ,  $\epsilon > 0$ , apenas 3 variáveis para cada um dos  $n$  indivíduos com crédito. Define-se então para o indivíduo  $i$ ,  $x_{i11}$  como o valor da primeira variável que está associada à Família 1,  $x_{i21}$  como o valor da segunda variável que está associada à Família 2 e  $x_{ic1}$  como o valor de uma variável de cliente que não está associada à nenhuma das famílias. A partir delas, para um conjunto de  $n$  clientes, define-se  $x_{11} = (x_{111}, x_{211}, \dots, x_{n11})^\top$ ,  $x_{21} = (x_{121}, x_{221}, \dots, x_{n21})^\top$  e  $x_{c1} = (x_{1c1}, x_{2c1}, \dots, x_{nc1})^\top$ . Caso o indivíduo  $i$  não possua conta na Família  $m$ ,  $x_{im1}$  não é observável. Nesse caso, para possibilitar o uso de um artifício algébrico nas estratégias 2 e 3,  $x_{im1}$  será codificado com o valor  $-1$  (admita, sem perda de generalidade, que essas variáveis não assumam valores negativos).

Cada uma das contas de cada cliente é classificada em uma dentre as seguintes categorias: mau, bom, indeterminado e cancelado. A classificação é feita de acordo com o comportamento de crédito da conta entre os instantes  $t + 1$  e  $t + \delta$ ,  $\delta > 1$  e está relacionada principalmente ao comportamento de atraso de pagamento observado durante o período. Essa variável é denominada resposta conta. A partir da resposta conta, pode-se obter a resposta cliente. Ela é definida como a pior situação do indivíduo em todas as contas que ele possui. São consideradas não apenas as contas existentes no instante  $t$  como aquelas contratadas no período entre  $t + 1$  e  $t + \delta$ . A resposta cliente recebe o valor cancelada, se todas as contas do indivíduo foram canceladas no período  $t + 1$  e  $t + \delta$ . Assim, no Exemplo 1, para cada cliente  $i$ , pode-se definir  $y_{i1}$ , como a resposta conta do indivíduo  $i$  na Família de produtos 1,  $y_{i2}$ , como a resposta conta do indivíduo  $i$  na Família de produtos 2 e  $y_{ic}$  como a resposta cliente do indivíduo  $i$ . Em todos os modelos, são desprezadas as observações cujas respostas são classificadas como indeterminada ou cancelada no período entre  $t + 1$  e  $t + \delta$ . Assim, as variáveis mencionadas são codificadas como

$$y_{im}, m = 1, 2, c = \begin{cases} 0 & \text{se a resposta é mau} \\ 1 & \text{se a resposta é bom.} \end{cases}$$

A partir das respostas de cada um dos indivíduos, define-se  $y_1 = (y_{11}, y_{21}, \dots, y_{n1})^\top$ ,  $y_2 = (y_{12}, y_{22}, \dots, y_{n2})^\top$ ,  $y_c = (y_{1c}, y_{2c}, \dots, y_{nc})^\top$  e  $Y = (y_1, y_2, y_c)^\top$ . Caso o indivíduo  $i$  não possua conta na Família  $m$ ,  $y_{im}$  é não observável.

O modelo de customer scoring tem como objetivo mensurar o risco de um cliente que é bom em um instante de origem  $t$  se tornar mau no período entre  $t$  e  $t + \delta$ . Dessa forma, são utilizados apenas clientes que são classificados como bons no instante de origem. Essa condição é válida para todas as estratégias e também para os modelos de *behavioural scoring*.

## 2.1 Estratégia 1

A Estratégia 1 é aquela que geralmente é utilizada no desenvolvimento de modelos de *customer scoring* (Groom e Gill, 1998). Ela possui duas etapas. Inicialmente são ajustados modelos de *behavioural scoring* para cada uma das famílias de produtos e, a partir deles, é obtido o modelo final. Essa estratégia pode ser segmentada em duas outras: 1a e 1b. A Estratégia 1a utiliza no ajuste dos modelos de *behavioural scoring* (modelos de produtos) a resposta conta e a 1b utiliza a cliente como variável dependente. Considera-se apenas as categorias mau e bom dessas variáveis resposta. A regressão logística é utilizada em ambas as estratégias e os modelos de produto podem ser escritos, para o Exemplo 1, como:

$$\begin{cases} g_1(\mu_{i1}) = g_1(E(y_{i1}/x_{i11})) = \beta_{10} + x_{i11}\beta_{11} \\ g_2(\mu_{i2}) = g_2(E(y_{i2}/x_{i21})) = \beta_{20} + x_{i21}\beta_{21} \end{cases} \quad \text{para a Estratégia 1a e}$$

$$\begin{cases} g_3(\mu_{ic}) = g_3(E(y_{ic}/x_{i11})) = \beta_{10}^c + x_{i11}\beta_{11}^c \\ g_4(\mu_{ic}) = g_4(E(y_{ic}/x_{i21})) = \beta_{20}^c + x_{i21}\beta_{21}^c \end{cases} \quad \text{para a Estratégia 1b}$$

no qual

$\beta_{10}$ ,  $\beta_{20}$ ,  $\beta_{10}^c$  e  $\beta_{20}^c$  são parâmetros de intercepto do modelo e

$\beta_{11}$ ,  $\beta_{21}$ ,  $\beta_{11}^c$  e  $\beta_{21}^c$  são parâmetros associados às variáveis preditoras.

Os clientes que não possuem conta em uma das famílias são retirados no momento da estimação do modelo de *behavioural scoring* associado a ela.

Em ambas as estratégias, os valores ajustados para cada uma das famílias de produtos (em geral multiplicados por 100 ou por 1000) são denominados escores de produto. Dessa forma, pode-se definir  $E_{ij}$  como o escore de produto do cliente  $i$  na Família  $j$  e  $E_j = (E_{1j}, E_{2j}, \dots, E_{nj})^\top$ . Pode-se classificar então  $E_{ij}$  em  $e_j$  categorias (classes de escore), através de algum método adequado. Na aplicação deste trabalho foi utilizado o método CHAID (Kass, 1980). Para tratar os indivíduos que não possuem conta na Família  $j$ , cria-se uma categoria adicional. Pode-se assim definir  $E_{ij}^*$ ,  $i = 1, 2, \dots, n$ , como o resultado da categorização de  $E_{ij}$  e com valores variando entre 1 e  $e_j + 1$  e  $E_j^* = (E_{1j}^*, E_{2j}^*, \dots, E_{nj}^*)^\top$ . Variáveis indicadoras relacionadas a  $E_j^*$  são criadas para possibilitar a inclusão dos escores de produto da Família  $j$  no modelo final. Elas serão denotadas pelos vetores

de  $n$  posições  $d_{jl} = (d_{1jl}, d_{2jl}, \dots, d_{njl})^\top, l = 2, 3, \dots, e_j + 1$  nos quais  $d_{ijl}$  é definida como

$$d_{ijl} = \begin{cases} 1 & \text{se } E_{ij}^* = l \\ 0 & \text{caso contrário.} \end{cases}$$

O índice  $l$  se inicia em 2 em virtude de um dos grupos ser tomado como referência e finaliza em  $e_j + 1$  para acomodar os clientes que não possuem conta na família  $j$ .

O modelo de *customer scoring* utiliza como preditoras, além das variáveis  $d_{1l}$  e  $d_{2l}$ ,  $x_{c1}$ , que é a variável de cliente que não foi utilizada nos modelos de *behavioural scoring*. Ele pode ser escrito como

$$g_5(\mu_{ic}) = g_5(E(y_{ic}/x_{ic1}, D_{i1}, D_{i2})) = \beta_0 + x_{ic1}\beta_c + D_{i1}^\top\beta_1 + D_{i2}^\top\beta_2$$

no qual

$$D_{i1} = (d_{i12}, \dots, d_{i1e_1+1})^\top \text{ e } D_{i2} = (d_{i22}, \dots, d_{i2e_2+1})^\top,$$

$\beta_c$  é o parâmetro associado à variável de cliente,

$\beta_1 = (\beta_{12}, \dots, \beta_{1e_1+1})^\top$  e  $\beta_2 = (\beta_{22}, \dots, \beta_{2e_2+1})^\top$  são os vetores de parâmetros associados às variáveis indicadoras dos escores de produto e

$\beta_0$  é o intercepto do modelo.

## 2.2 Estratégia 2

O ajuste de um modelo de *customer scoring* sem a etapa intermediária de desenvolvimento de vários modelos de *behavioural scoring* é outra estratégia possível para a obtenção de um modelo de cliente. Para isso, divide-se a população de clientes da instituição em grupos, de acordo com os produtos que cada um possui. No Exemplo 1, haveriam 3 grupos: o primeiro formado pelos clientes que só têm conta na Família 1, o segundo com os indivíduos que têm conta apenas na Família 2 e o último contendo aqueles que têm contas em ambas as famílias. Ajusta-se então um modelo de *customer scoring* para cada um dos grupos criados, utilizando-se apenas as variáveis preditoras disponíveis em cada um deles. No primeiro grupo, por exemplo, não é utilizada  $x_{i21}$  porque ela não é observável para nenhum dos indivíduos desse grupo. Assim, o modelo de regressão logística de cada um dos grupos é dado por

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, x_{i11})) = \beta_{01} + x_{ic1}\beta_c + x_{i11}\beta_1$$

para o cliente  $i$  que tem conta apenas na Família 1,

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, x_{i21})) = \beta_{02} + x_{ic1}\beta_c + x_{i21}\beta_2$$

para o cliente  $i$  que tem conta apenas na Família 2 e

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, x_{i11}, x_{i21})) = \beta_{03} + x_{ic1}\beta_c + x_{i11}\beta_1 + x_{i21}\beta_2$$

para o cliente  $i$  que tem conta em ambas as famílias, sendo que  $\beta_c$  é o parâmetro associado à variável de cliente,  $\beta_1$  e  $\beta_2$  são os parâmetros relacionadas às variáveis de produto e  $\beta_{01}, \beta_{02}, \beta_{03}$  são os interceptos dos modelos.

Todos esses modelos podem ser estimados simultaneamente através da criação de variáveis que indiquem se o cliente tem ou não conta em determinada família. Dessa forma, define-se:

$$w_{ij} = \begin{cases} 1 & \text{se o cliente } i \text{ possui conta na Família } j \\ 0 & \text{caso contrário,} \end{cases}$$

$w_j = (w_{1j}, w_{2j}, \dots, w_{nj})^\top$ . Observe que se o indivíduo  $i$  não possuir conta, na Família  $j$ , então  $x_{ij1}w_{ij} = 0$ . Assim os modelos podem ser ajustados conjuntamente através da equação

$$g(\mu_{ic}) = \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + x_{i11}w_{i1}\beta_1 + x_{i21}w_{i2}\beta_2 + x_{ic1}\beta_c$$

na qual

$$\mu_{ic} = E(y_{ic}/x_{i11}, x_{i21}, x_{ic1}, w_{i1}, w_{i2})$$

$\alpha_1$  e  $\alpha_2$  são os parâmetros associados, respectivamente, a  $w_{i1}$  e  $w_{i2}$  e  $\beta_0$  é o intercepto do modelo.

Observe que esse modelo é equivalente aos apresentados para cada um dos grupos. Para verificar a igualdade entre eles, é necessário apenas obter  $w_{i1}$  e  $w_{i2}$ , de acordo com as famílias que o cliente possui conta e considerar  $\beta_{01} = \beta_0 + \alpha_1$ ,  $\beta_{02} = \beta_0 + \alpha_2$  e  $\beta_{03} = \beta_0 + \alpha_1 + \alpha_2$ . Pelo fato do modelo apresentar ajustes paralelos de acordo com a família de produtos que o cliente possui conta, ele é semelhante ao de uma análise de covariância (Kutner et al., 2004).

A inclusão do efeito principal de  $w_{i1}$  e  $w_{i2}$  é importante para diferenciar dois grupos de clientes que podem ter comportamentos bastante diferentes. Suponha, por exemplo, dois indivíduos que possuem o mesmo valor de  $x_{ic1}$  e  $x_{i11}$ . A única diferença entre eles está na Família 2. O primeiro cliente não possui conta nessa família. O outro possui, mas, tem  $x_{i21} = 0$ . Nesse caso a não inclusão do efeito principal de  $w_{i2}$  causa a igualdade entre o valor ajustado desses dois indivíduos que podem ter risco de crédito diferentes entre si.

O efeito de  $x_{i11}$ ,  $x_{i21}$  e  $x_{ic1}$  é suposto ser independente de quais as famílias em que o cliente possui conta. Porém, na prática, o efeito de  $x_{i11}$  em um indivíduo que possui conta apenas na Família 1 pode ser diferente em um outro que possui conta nas duas famílias. Pode-se então definir um modelo alternativo para a Estratégia 2 no qual o efeito das variáveis  $x$  varia de acordo com as famílias que o cliente possui conta. No entanto, em situações práticas, isso envolve a criação de um elevado número de variáveis indicadoras. Várias delas podem ter o valor 1 para uma proporção não muito grande de clientes. Assim, permitir que uma variável indicadora tenha efeito diferente no ajuste do modelo, de acordo com as famílias de produtos que o cliente possui pode não ser factível. O motivo é que, provavelmente, para várias variáveis indicadoras, haverá grupos nos quais a quantidade de clientes com valor 1 será muito pequeno. Dessa forma, as estimativas dos parâmetros associados a elas serão pouco robustas.

### 2.3 Estratégia 3

A terceira estratégia sugerida tem similaridades com a segunda. Também são ajustados, simultaneamente, modelos para cada configuração de família de produtos que o cliente possui. A diferença é que, além de um modelo para a resposta cliente, são estimados, simultaneamente, modelos para a resposta conta das famílias de produtos que o cliente possui. Para o Exemplo 1, cada cliente  $i$ , teria na Estratégia 3, o vetor resposta  $Y_i = (y_{i1}, y_{i2}, y_{ic})^\top$ . As duas primeiras posições do vetor são a resposta conta associada, respectivamente, às famílias de produtos 1 e 2, enquanto a última posição é a resposta cliente. Nessa estratégia é introduzida uma estrutura de dependência entre as observações, já que há mais de uma resposta para um mesmo cliente. Dessa forma, as equações de estimação generalizadas (GEE) com ligação logito é uma técnica conveniente para a obtenção das estimativas dos parâmetros do modelo. Como o número de famílias não tende a ser muito grande, sugere-se a adoção de matriz de correlação de trabalho não estruturada. Porém, nem sempre ela pode ser adotada, já que é possível a não convergência dos estimadores dos parâmetros, quando essa estrutura é escolhida.

Para facilitar a compreensão da notação utilizada na Estratégia 3, será feita a comparação das estruturas dos bancos de dados das estratégias 2 e 3. A Estratégia 2 possui uma estrutura do banco de dados semelhante a apresentada na Tabela 1. Nesse exemplo, o cliente 2 não possui conta na Família 2 e o cliente 3 não possui conta na Família 1.

Tabela 1: Estrutura do banco de dados Estratégia 2

Cliente	Família	$y_c$	$x_{11}$	$x_{21}$	$x_{c1}$	$w_1$	$w_2$
1	cliente	$y_{1c}$	$x_{111}$	$x_{121}$	$x_{1c1}$	1	1
2	cliente	$y_{2c}$	$x_{211}$	-1	$x_{2c1}$	1	0
3	cliente	$y_{3c}$	-1	$x_{321}$	$x_{3c1}$	0	1

Na Estratégia 3,  $y_{ic}$ , que contém apenas a resposta cliente do indivíduo  $i$ , é substituído pelo vetor  $Y_i$ , que contém também as respostas conta. Em virtude disso, para o Exemplo 1, o número de linhas do banco de dados é multiplicado por três (a Tabela 2 ilustra esse fato). Os valores  $x_{i11}$ ,  $x_{i21}$ ,  $x_{ic1}$ ,  $w_{i1}$  e  $w_{i2}$  não se alteram para cada uma das ocorrências do cliente  $i$ . Dessa forma,  $v_{11}$ ,  $v_{21}$ ,  $v_{c1}$ ,  $w_1^*$  e  $w_2^*$  são simplesmente  $x_{i11}$ ,  $x_{i21}$ ,  $x_{ic1}$ ,  $w_{i1}$  e  $w_{i2}$  repetido 3 vezes, conforme pode ser visto na Tabela 2. Ela apresenta a estrutura do banco de dados para a Estratégia 3 e os dados são equivalentes aos apresentados na Tabela 1.

A Tabela 2 apresenta ainda  $z_j = (z_{11j}, z_{12j}, z_{1cj}, z_{21j}, z_{22j}, z_{2cj}, \dots, z_{n1j}, z_{n2j}, z_{ncj})^\top$ ,  $j = 1, 2$ , sendo que  $z_{ilj}$  é definida como

$$z_{ilj} = \begin{cases} 1 & \text{se a observação } l \text{ do cliente } i \text{ refere-se à família de produtos } j \\ 0 & \text{caso contrário,} \end{cases}$$

Essas variáveis são criadas para possibilitar a diferenciação entre os valores ajustados para as respostas conta e cliente do indivíduo  $i$ . São criadas ainda interações entre as variáveis preditoras originais e as indicadoras de observações (variáveis  $z$ ) para permitir que o efeito de cada uma das variáveis  $x$  possa ser diferente no ajuste das respostas conta e cliente. Dessa forma, o modelo para a Estratégia 3 pode ser definido como

$$\begin{aligned}
g(\mu_{il}) = & \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + z_{il1}\gamma_1 + z_{il2}\gamma_2 + x_{i11}w_{i1}\beta_{10} + x_{i21}w_{i2}\beta_{20} + \\
& + x_{ic1}\beta_{c0} + x_{i11}w_{i1}z_{il1}\beta_{11} + x_{i21}w_{i2}z_{il1}\beta_{21} + x_{ic1}z_{il1}\beta_{c1} + \\
& + x_{i11}w_{i1}z_{il2}\beta_{12} + x_{i21}w_{i2}z_{il2}\beta_{22} + x_{ic1}z_{il2}\beta_{c2}
\end{aligned} \tag{1}$$

no qual

$$g(\mu_{il}) = g(E(y_{il}/w_{i1}, w_{i2}, z_{il1}, z_{il2}, x_{i11}, x_{i21}, x_{ic1}));$$

$\alpha_j$  e  $\gamma_j, j = 1, 2$  são parâmetros associados, respectivamente, a  $w_{ij}$  e  $z_{ilj}$ ;

$\beta_{ij}, i = 1, 2, c, j = 0, 1, 2$  são parâmetros associados às demais variáveis preditoras;

$\beta_0$  é o intercepto do modelo.

Para o cliente 2 da Tabela 2, que possui conta apenas na Família 1, o modelo para a sua única resposta conta será portanto dado por

$$\begin{aligned}
g(\mu_{i1}) = & \beta_0 + \alpha_1 + \gamma_1 + x_{i11}\beta_{10} + x_{ic1}\beta_{c0} + x_{i11}\beta_{11} + x_{ic1}\beta_{c1} = \\
= & (\beta_0 + \alpha_1 + \gamma_1) + (\beta_{10} + \beta_{11})x_{i11} + (\beta_{c0} + \beta_{c1})x_{ic1}
\end{aligned} \tag{2}$$

e o modelo para a resposta cliente pode ser escrito como

$$g(\mu_{ic}) = \beta_0 + \alpha_1 + x_{i11}\beta_{10} + x_{ic1}\beta_{c0} = (\beta_0 + \alpha_1) + \beta_{10}x_{i11} + \beta_{c0}x_{ic1} \tag{3}$$

Tabela 2: Estrutura do banco de dados Estratégia 3

Cliente	Família	$Y$	$v_{11}$	$v_{21}$	$v_{c1}$	$w_1^*$	$w_2^*$	$z_1$	$z_2$
1	1	$y_{11}$	$x_{111}$	$x_{121}$	$x_{1c1}$	1	1	1	0
1	2	$y_{12}$	$x_{111}$	$x_{121}$	$x_{1c1}$	1	1	0	1
1	cliente	$y_{1c}$	$x_{111}$	$x_{121}$	$x_{1c1}$	1	1	0	0
2	1	$y_{21}$	$x_{211}$	-1	$x_{2c1}$	1	0	1	0
2	2	.	$x_{211}$	-1	$x_{2c1}$	1	0	0	1
2	cliente	$y_{2c}$	$x_{211}$	-1	$x_{2c1}$	1	0	0	0
3	1	.	-1	$x_{321}$	$x_{3c1}$	0	1	1	0
3	2	$y_{32}$	-1	$x_{321}$	$x_{3c1}$	0	1	0	1
3	cliente	$y_{3c}$	-1	$x_{321}$	$x_{3c1}$	0	1	0	0

Já para o cliente 1, que possui conta nas duas famílias, o modelo é dado por

$$\begin{aligned}
g(\mu_{i1}) &= \beta_0 + \alpha_1 + \alpha_2 + \gamma_1 + x_{i11}\beta_{10} + x_{i21}\beta_{20} + x_{ic1}\beta_{c0} + x_{i11}\beta_{11} + x_{i21}\beta_{21} + x_{ic1}\beta_{c1} \\
&= (\beta_0 + \alpha_1 + \alpha_2 + \gamma_1) + (\beta_{10} + \beta_{11})x_{i11} + (\beta_{20} + \beta_{21})x_{i21} + \\
&\quad + (\beta_{c0} + \beta_{c1})x_{ic1}
\end{aligned} \tag{4}$$

para a resposta conta da Família 1,

$$\begin{aligned}
g(\mu_{i2}) &= \beta_0 + \alpha_1 + \alpha_2 + \gamma_2 + x_{i11}\beta_{10} + x_{i21}\beta_{20} + x_{ic1}\beta_{c0} + x_{i11}\beta_{12} + x_{i21}\beta_{22} + x_{ic1}\beta_{c2} \\
&= (\beta_0 + \alpha_1 + \alpha_2 + \gamma_2) + (\beta_{10} + \beta_{12})x_{i11} + (\beta_{20} + \beta_{22})x_{i21} + \\
&\quad + (\beta_{c0} + \beta_{c2})x_{ic1}
\end{aligned} \tag{5}$$

para a resposta conta da Família 2 e

$$\begin{aligned}
g(\mu_{ic}) &= \beta_0 + \alpha_1 + \alpha_2 + x_{i11}\beta_{10} + x_{i21}\beta_{20} + x_{ic1}\beta_{c0} \\
&= (\beta_0 + \alpha_1 + \alpha_2) + \beta_{10}x_{i11} + \beta_{20}x_{i21} + \beta_{c0}x_{ic1}
\end{aligned} \tag{6}$$

para a resposta cliente.

Comparando-se as equações (4), (5) e (6), pode-se ver que o efeito de cada uma das variáveis  $x$  varia em função da resposta que se está modelando para o cliente  $i$ . O coeficiente de  $x_{i11}$ , por exemplo, é  $\beta_{10} + \beta_{11}$ ,  $\beta_{10} + \beta_{12}$  e  $\beta_{10}$ , caso se esteja ajustando, respectivamente, as respostas conta da Família 1, conta da Família 2 e cliente. Assim,  $\beta_{11}$  é a variação no efeito de  $x_{i11}$  quando substitui-se o ajuste da resposta cliente pelo ajuste da resposta conta da Família 1. Porém, assim como na Estratégia 2, o efeito das variáveis  $x$  não se altera de acordo com as famílias de produtos que o cliente possui. Observando-se, por exemplo, as equações (3) e (6), pode-se notar que o efeito de  $x_{i11}$  é o mesmo no ajuste da resposta cliente de um indivíduo que tem conta apenas na Família 1 e de um outro que tem conta nas duas famílias. O modelo pode ser alterado para que o efeito de  $x_{i11}$  na resposta cliente varie de acordo com os produtos que o indivíduo possui. Porém, essa alternativa apresenta os mesmos problemas práticos já discutidos na Seção 2.2.

No momento da estimação do modelo, todas as observações referentes às famílias que os clientes não possuem conta são excluídas. Para o banco de dados da Tabela 2, por exemplo, as linhas 5 e 7 seriam eliminadas. Porém, no ajuste de um modelo de GEE, permite-se que as demais observações dos clientes que não têm contas em todas as famílias sejam utilizadas.

## 2.4 Modelo geral

Em situações práticas, tem-se  $M$  famílias de produtos e  $K_m$  variáveis preditoras relacionadas a elas,  $m = 1, \dots, M$ . No entanto, os modelos das três estratégias são bem semelhantes aos apresentados nas seções anteriores. A única diferença está no maior número de variáveis envolvidas e, por isso, a necessidade de uma amostra suficientemente grande para a obtenção de estimativas robustas dos parâmetros existentes.

Nos modelos definidos nesta seção, foi feita a suposição de que cada cliente tinha zero ou uma conta em cada família de produtos. Porém, é muito comum que vários clientes possuam mais de uma conta em uma ou mais famílias. A inclusão de mais de uma conta de uma mesma família nos modelos apresentados, traz mais uma fonte de dependência entre as observações. Porém, nesse caso, a dependência é difícil de ser tratada, já que cada cliente possui um número aleatório de contas em cada família. A solução é utilizar um procedimento para que cada cliente possua um único valor para a resposta conta e para cada uma das variáveis preditoras da família de produtos.

Em relação às variáveis preditoras, isso pode ser feito pelo menos de duas formas diferentes. A primeira é, para cada variável, consolidar todas as contas de uma mesma família em uma única conta, através de um indicador adequado (soma, média, máximo, mínimo, etc). Em determinadas situações, essa alternativa não pode ser adotada. Isso ocorre, por exemplo, quando determinadas variáveis são resultado da razão de duas outras que não estão disponíveis. Uma outra alternativa é sortear uma das contas para caracterizar o cliente na Família de produtos  $m$  e utilizar suas variáveis independentes. O banco de dados utilizado neste trabalho possui algumas variáveis que não podem ser consolidadas. Em virtude disso, será utilizado o procedimento de sorteio de uma das contas.

Em relação à variável resposta conta também podem ser utilizados pelo menos dois procedimentos. O primeiro é considerá-la como a situação da pior conta daquela família, de acordo com a prioridade apresentada anteriormente. Caso as variáveis preditoras tenham sido escolhidas a partir do sorteio de uma das contas, pode ser mais conveniente adotar a resposta da conta escolhida. Nesse caso, tanto as variáveis preditoras quanto a resposta conta são obtidas a partir da conta sorteada. Para a aplicação deste trabalho, essa opção foi adotada.

### 3 Descrição dos dados

Utiliza-se um conjunto de dados reais obtido de uma instituição financeira para a ilustração e comparação das estratégias de desenvolvimento de um modelo de *customer scoring*. Para este trabalho, o conjunto de produtos de crédito sem garantia dessa instituição foi dividido em 3 famílias: cheque especial, cartão de crédito e outros produtos sem garantia. A população do estudo engloba todos os clientes que possuíam conta corrente e cheque especial ou cartão de crédito e não tinham nenhum problema de crédito em dezembro de 2001. Dessa população foi extraída uma amostra aleatória simples de 30.000 clientes, gerando assim a base de dados que será usada neste artigo.

Cada um dos clientes possuía desde nenhum até vários contratos em cada uma das famílias. Para cada um dos contratos foram obtidas diversas variáveis, caracterizando o comportamento de uso do produto pelo cliente em dezembro de 2001 e nos 5 meses anteriores. Por motivo de sigilo, os nomes e descrição de cada uma das variáveis não poderão ser apresentados.

Da família do cheque especial, cartão de crédito e outros produtos sem garantia, foram obtidas, respectivamente, 12, 9 e 6 variáveis. Além das variáveis relacionadas ao comportamento do cliente em cada um dos produtos foram observadas 9 características de cada indivíduo. Essas variáveis completam a lista de variáveis preditoras e não estão associadas a nenhum produto particular,

sendo em sua maioria informações sócio-demográficas do indivíduo.

A situação de cada um dos contratos em cada uma das famílias foi observada em junho de 2002. Cada um deles foi classificado em uma das 4 categorias da variável: mau, indeterminado, bom ou cancelado. A segmentação da situação do contrato em mau, indeterminado e bom está associada principalmente ao número de dias em atraso do cliente. Contratos classificados como cancelados são aqueles que o cliente não possui mais aquele produto em junho de 2002. Apenas os contratos da família cheque especial e cartão de crédito podem assumir esse valor. No caso específico da família de outros produtos, o fato do cliente não possuir mais aquele contrato indica que ele pagou todas as suas prestações. Por isso, para essa família, se o cliente não possui mais aquele contrato, ele é classificado como bom. A situação de cada um dos contratos em junho de 2002 é a resposta conta da operação de crédito. A partir das respostas conta é criada a variável resposta cliente da forma discutida na seção anterior. Na criação da resposta cliente também são considerados os produtos que não foram alocados em nenhuma das famílias, como, por exemplo, os que possuem garantia. Dessa forma cada contrato possui duas variáveis respostas, uma conta e a outra cliente. Já os indivíduos possuem uma resposta cliente e diversas respostas contas. A quantidade varia em função do número de produtos que ele possuía em dezembro de 2001.

## 4 Aplicação

Os modelos das estratégias descritas anteriormente foram ajustados para os dados apresentados na Seção 3. Para a estimação do modelo todas as variáveis foram categorizadas a partir do procedimento CHAID (Kass, 1980) com nível de significância de 5%. Para as variáveis relacionadas ao comportamento de uso de algum produto, foi selecionada aleatoriamente apenas uma conta por cliente, para evitar dependência entre as observações.

A distribuição da variável resposta em cada uma das famílias de produtos para os 30.000 clientes pode ser observada na Tabela 3. A variável dependente referente a cada uma das famílias é a resposta conta. Pode-se notar que, desconsiderando-se os clientes indeterminados e cancelados, 3,4% dos clientes se tornaram maus após 6 meses de observação. Porém, o risco varia bastante de acordo com a família de produtos. Pode-se observar ainda que o percentual de clientes com produto em determinada família também tem alta variabilidade.

A Tabela 4 apresenta a matriz de correlação entre as variáveis resposta. Para a construção da tabela, desconsiderou-se as observações indeterminadas e canceladas. Pode-se notar que as correlações entre as respostas são muito altas. Dessa forma, se um modelo ajusta todas elas simultaneamente, é fundamental o uso de uma técnica estatística que trata a dependência entre as observações. Por isso, a utilização da GEE com ligação logito é uma alternativa viável na Estratégia 3.

O banco de dados foi dividido aleatoriamente em dois grupos: amostra de desenvolvimento contendo 20.000 clientes e amostra de validação com 10.000 indivíduos. Na amostra de desenvolvimento foram ajustados todos os modelos de cada uma das estratégias. Estes foram então aplicados

na amostra de validação para a comparação da performance de cada uma das estratégias em uma amostra independente da utilizada na estimação dos parâmetros.

Os modelos das Estratégias 1 e 2 foram estimados no software SPSS. A seleção de variáveis nas estratégias 1 e 2 foi feita a partir do procedimento forward stepwise. Para evitar o favorecimento de alguma das estratégias, procurou-se fazer o menor número possível de ajustes nos modelos obtidos a partir do procedimento.

O SAS foi utilizado para a estimação do modelo da Estratégia 3. Ele não permite a execução do procedimento stepwise na estimação de um modelo de GEE. Em virtude disso, a alternativa natural seria estimar o modelo com todas as variáveis, retirando-se, uma a uma, as variáveis não significantes. Porém, em virtude de cada uma das variáveis gerar várias variáveis indicadoras e cada uma delas interagir com cada uma das variáveis  $z_l$ , o número de parâmetros a ser estimado é muito grande (as 36 variáveis originais dão origem a 558 variáveis indicadoras na Estratégia 3). Além disso, o fato da maioria dos clientes não possuir produtos em todas as famílias torna os dados bastante desbalanceados. Em consequência disso, não foi possível ajustar o modelo da Estratégia 3 com todas as variáveis. Contornou-se esse problema através do ajuste no SAS de dois modelos. O

Tabela 3: Distribuição da variável resposta

Variável resposta	Cheque		Cartão		Outros		Cliente	
	#	%	#	%	#	%	#	%
Mau	838	3,3	502	2,7	273	7,3	949	3,4
Bom	24.863	96,7	18.089	97,3	3.459	92,7	26.645	96,6
Total no ajuste	25.701	100,0	18.591	100,0	3.732	100,0	27.594	100,0
Indeterminado	176	0,6	209	1,1	44	1,2	564	1,9
Cancelado	2.462	8,7	1.078	5,4	—	—	1.842	6,1
Total com produto	28.339	100,0	19.878	100,0	3.776	100,0	30.000	100,0
Sem produto	1.661	5,5	10.122	33,7	26.224	87,4	0	0,0
Total	30.000	100,0	30.000	100,0	30.000	100,0	30.000	100,0

Tabela 4: Matriz de correlação entre as respostas conta e produto

	Cheque	Cartão	Outros	Cliente
Cheque	1,000	0,899	0,915	0,963
Cartão		1,000	0,943	0,918
Outros			1,000	0,884
Cliente				1,000

primeiro não considera a interação entre as variáveis  $x$  e as variáveis  $z$ , reduzindo assim em cerca de 75% o número de parâmetros a ser estimado. Dessa forma, foi possível ajustar e obter um modelo final, após a retirada uma a uma das variáveis não significantes (nível de significância de 5%). Esse modelo será denotado como 3r.

O segundo modelo ajustado foi construído de forma subjetiva. A partir da análise descritiva e da observação dos ajustes dos modelos das Estratégias 1 e 2, foi feita uma pré-seleção de variáveis, escolhendo-se aquelas que tinham maior associação com a variável resposta. Elas então foram divididas em pequenos grupos de variáveis. Para cada um desses grupos foi possível ajustar o modelo. Assim, obteve-se para cada um deles um modelo final, retirando-se, uma a uma, as variáveis não significantes. Os grupos foram fundidos em outros maiores e o procedimento foi repetido. Isso foi feito até a obtenção de um único grupo no qual todas as variáveis eram significantes. O nível de significância utilizado também foi de 5%. Durante esse processo, algumas variáveis ainda foram excluídas para evitar erro na rotina de estimação pelo SAS. Esse modelo será denotado como 3s.

Ambos os modelos utilizam a estrutura uniforme para a matriz de correlação de trabalho. A Tabela 4 indica que essa estrutura parece ser adequada. As estimativas obtidas para o parâmetro de correlação nos Modelos 3r e 3s foi respectivamente de 0,8317 e 0,8535.

O coeficiente de Gini (Thomas, 2009) foi utilizado para a comparação da performance das estratégias (Tabela 5). Na última linha da tabela, pode-se ver que a diferença do coeficiente de Gini entre as estratégias não é grande. A variação de desempenho entre a estratégia de melhor e pior performance é inferior a 3%. No entanto, as Estratégias 1b e 2 se destacam como as duas que apresentaram melhor performance. Pode-se observar ainda que, mesmo não sendo possível o ajuste do melhor modelo da Estratégia 3 devido a restrições computacionais, o desempenho por ela apresentado não foi muito inferior às demais. Isso é um indício de que essa estratégia poderá vir a se tornar uma boa opção, após o aperfeiçoamento dos algoritmos de ajuste da GEE presente nos principais softwares estatísticos. Nota-se também que a Estratégia 1b apresentou desempenho superior a 1a. Isso sugere que, caso se deseje utilizar a Estratégia 1 e o interesse na obtenção de cada um dos escores de produto seja apenas de utilizá-los como preditora para o modelo principal, é mais interessante utilizar a variação b.

A Tabela 5 permite ainda observar se o número de famílias de produtos que o cliente possui interfere na ordenação de performance entre as estratégias. A ordenação de performance entre as estratégias parece não ter forte associação com o número de famílias de produtos. A Estratégia 2 se destaca nos grupos de clientes com uma e três famílias, enquanto a 1b apresenta melhor desempenho entre os indivíduos com duas famílias. É interessante notar que o desempenho absoluto de todas as estratégias melhora à medida que decresce o número de famílias. Embora o grupo de indivíduos com 3 famílias possua um número maior de variáveis para se estimar o risco, isso parece não ser suficiente para compensar um acréscimo na quantidade de produtos diferentes nos quais o indivíduo pode se tornar mau.

A Tabela 6 apresenta as medidas de performance para os modelos de produto. Ela mostra os resultados apenas das estratégias 1a e 3, porque apenas estas geram um escore de produto que é a estimativa da probabilidade do cliente se manter bom naquela família. Pode-se observar que

o desempenho da Estratégia 3 é superior ao da 1a para todas as famílias. Isso ocorre porque a Estratégia 3 é desenvolvida de forma que todas as variáveis disponíveis participem do ajuste de cada um dos escores de produto. Na Estratégia 1a, apenas as variáveis relacionadas à própria família de produtos para a qual se está estimando o risco são utilizadas. Em virtude dos resultados observados, há indícios de que, caso se deseje obter uma estimativa da probabilidade do cliente se manter bom em determinada família de produtos, é recomendável a utilização da Estratégia 3, mesmo considerando-se os problemas existentes na estimação dos parâmetros e seleção de variáveis.

## 5 Estudo de simulação

Com o objetivo de estudar a performance das estratégias em condições controladas foi feito um estudo de simulação de Monte Carlo. Ele foi desenvolvido para situações nas quais são ajustados modelos de *behavioural scoring* para duas famílias de produtos e todos os clientes possuem conta em ambas. Os dados foram gerados a partir do algoritmo abaixo.

- A partir do banco de dados descrito na Seção 3, foram sorteados 10.000 clientes que possuíam conta tanto na família do cheque especial como na família do cartão de crédito. De cada uma das famílias de produtos, escolheu-se então duas variáveis para participar da simulação, que juntas produziram 16 variáveis indicadoras.

Tabela 5: Coeficiente de Gini das estratégias por número de famílias para a resposta cliente

Número de famílias	Estratégia				
	1a	1b	2	3r	3s
3	0,732	0,748	0,767	0,730	0,741
2	0,809	0,820	0,818	0,802	0,806
1	0,820	0,821	0,843	0,826	0,831
Total	0,817	0,830	0,836	0,814	0,823

Tabela 6: Coeficiente de Gini dos modelos de produtos

Família de produtos	Estratégia		
	1a	3r	3s
Cheque	0,809	0,838	0,850
Cartão	0,679	0,869	0,879
Outros	0,456	0,758	0,772

- Para cada resposta conta, ajustaram-se então modelos de regressão logística tendo como variáveis preditoras as 16 variáveis indicadoras obtidas. Obtiveram-se assim estimativas da probabilidade de cada um dos 10.000 indivíduos se manter bom cliente nas duas famílias de produtos.
- Gerou-se 10000 pares de variáveis com distribuição marginal uniforme no intervalo  $[0, 1]$  correlacionadas a partir de algoritmo descrito em Johnson (1987). A partir dessas uniformes e das probabilidades obtidas no item anterior, obteve-se a resposta conta para cada indivíduo em cada família. As variáveis com distribuição uniforme foram geradas de tal forma que a correlação entre as variáveis resposta conta fosse aproximadamente 0,5 no primeiro grupo de simulações e 0,9 no segundo.
- A resposta cliente foi obtida de duas formas diferentes. Na primeira condição, a resposta cliente era a pior situação entre as duas respostas conta. Na segunda condição, denominada com perturbação, a partir da geração de um vetor de variáveis aleatórias com distribuição uniformes independentes, com probabilidade 0,005 classificou-se o cliente como mau mesmo que ambas as suas repostas contas tivessem sido classificadas como boas. A introdução da perturbação visa simular situações em que os indivíduos se tornam maus clientes em contas que não existiam no instante de origem ou em contas de famílias de produtos para as quais não foram desenvolvidos modelos.

Foram feitas 2.000 repetições desse algoritmo sendo 500 para cada combinação de parâmetro de correlação e ocorrência ou não de perturbação. Para cada repetição, foram ajustados os modelos de cada uma das estratégias conforme descrito na Seção 2 e utilizando as mesmas 16 variáveis indicadoras utilizadas na geração dos dados.

A Tabela 7 apresenta a proporção de repetições em que cada uma das estratégias é superior a todas as demais. Pode-se notar que as estratégias 2 e 3 apresentaram desempenho superior em maior proporção na previsão da resposta cliente. A Estratégia 2 mostrou-se superior às demais quando a correlação entre as resposta conta foi de 0,5 e na condição com correlação de 0,9 e ausência de perturbação. Na condição com correlação 0,9 e presença de perturbação, as estratégias 2 e 3 apresentaram um desempenho muito semelhante e superior às demais.

## 5.1 Comparação entre as estratégias propostas e as usualmente utilizadas

Na Tabela 8, obteve-se a proporção de repetições em que cada uma das estratégias é melhor que cada uma das demais. As linhas 2 a 5 de cada segmento da tabela comparam as estratégias propostas neste trabalho (2 e 3) com as estratégias geralmente utilizadas (1a e 1b). Pode-se ver que, para todas as condições, o intervalo de confiança está acima de 50%. Isso indica que, para todas as condições, as estratégias propostas são superiores às geralmente utilizadas em mais da metade das vezes. Pode-se notar ainda que a ordem de grandeza da proporção de repetições em que uma estratégia é superior à outra não é constante para todas as condições. A proporção de

repetições em que as estratégias 2 e 3 são superiores à Estratégia 1a é maior quando a correlação entre as respostas é 0,5 do que quando ela vale 0,9. Porém, a ocorrência ou não de perturbação parece não afetar as conclusões. Na comparação das estratégias 2 e 3 com a Estratégia 1b ocorre o oposto. A proporção de repetições em que as estratégias 2 e 3 são superiores à Estratégia 1b é maior quando a correlação entre as respostas é 0,9 do que quando ela vale 0,5. Além disso, a proporção de repetições em que as estratégias 2 e 3 são superiores à Estratégia 1b é maior quando há perturbação do que quando não há.

A Tabela 9 compara as estratégias duas a duas em relação à média das medidas de performance. Para todas as condições, as estratégias 2 e 3 também apresentam desempenho superior às estratégias 1a e 1b. Embora as estratégias 2 e 3 tenham uma melhor performance que as estratégias 1a e 1b, tanto em relação à proporção de vezes em que elas são superiores quanto em relação à média do coeficiente de Gini, a diferença nos valores médios não é grande. A diferença entre as médias do coeficiente de Gini nunca é superior a 0,005. Além disso, para nenhuma comparação é comum a ocorrência de grandes diferenças entre as estratégias. A Tabela 10 apresenta estatísticas das diferenças das medidas na comparação das estratégias duas a duas. O terceiro quartil da diferença das estratégias 2 e 3 em relação às estratégias 1a e 1b, por exemplo, nunca é superior a 0,008. Até mesmo as diferenças mínimas e máximas não são muito elevadas sendo, em módulo, inferiores a 0,03.

## 5.2 Demais comparações

Em relação à proporção de vezes em que uma é superior a outra (Tabela 8), para todas as condições, a Estratégia 3 ou tem desempenho semelhante à Estratégia 2 ou esta última tem performance ligeiramente superior. Em relação a média do coeficiente de Gini (Tabela 9), a Estratégia 2 também é ligeiramente superior ou equivalente a Estratégia 3 para todas as condições. Nas condições em que a Estratégia 2 é superior a Estratégia 3, a diferença média para o coeficiente de Gini nunca é superior a 0,0005. Pode-se observar ainda que a diferença entre as estratégias 2 e 3 nunca excede,

Tabela 7: Proporção de vezes que a estratégia é a de melhor performance

Estratégia	Perturbação			
	Sim		Não	
	Correlação 0,5	Correlação 0,9	Correlação 0,5	Correlação 0,9
1a	10%	23%	9%	22%
1b	22%	10%	29%	16%
2	38%	33%	35%	34%
3	30%	34%	27%	28%

em módulo, 0,02 (Tabela 10).

O desempenho comparativo das estratégias 1a e 1b se altera de acordo com a condição. Em relação à proporção de vezes em que uma é superior a outra e à média, para as duas condições com correlação de 0,5 (tabelas 8 e 9), a Estratégia 1b tem performance superior. Já nas condições com correlação 0,9, a Estratégia 1a tem performance superior tanto em relação à proporção (Tabela 8) como em relação à média (Tabela 9). No entanto, pode-se observar que as diferenças médias (Tabela 9) e máximas (Tabela 10) não são grandes em todas as condições.

### 5.3 Comentários gerais

Nas subseções anteriores comparou-se a performance das estratégias a partir de dados simulados e na seção anterior elas foram comparadas utilizando-se modelos ajustados a partir de dados reais. Apesar da comparação da performance ser importante, é interessante também comparar as estratégias em relação à outros aspectos.

As Estratégias 1a e 1b possuem pelo menos duas vantagens sobre as demais. A primeira é a simplicidade. Para o desenvolvimento de um modelo de *customer scoring* utilizando essa estratégia, usa-se exatamente a mesma metodologia de ajuste de um modelo de *behavioural scoring*. A outra é a possibilidade de aproveitamento dos modelos já existentes. Caso a instituição possua diversos modelos de *behavioural scoring*, ela pode aproveitar esses modelos no ajuste do modelo de *customer scoring*, diminuindo de forma considerável o tempo de desenvolvimento. A Estratégia 1a, assim como a Estratégia 3, ainda possui a vantagem de produzir uma estimativa da probabilidade de um cliente se tornar mau em determinado produto, que pode ser de interesse da instituição. Porém, conforme discutido na Seção 4, essa estimativa não considera todas as variáveis disponíveis como na Estratégia 3. A principal desvantagem das Estratégias 1a e 1b é o fato delas não considerarem a dependência existente entre as informações de um mesmo indivíduo em famílias de produtos diferentes. Nessas estratégias essa dependência é desconsiderada, em virtude do desenvolvimento de forma independente de um modelo para cada família. A Estratégia 1a possui pelo menos mais uma desvantagem. Os parâmetros de variáveis associadas às famílias de produtos são estimados no ajuste da resposta conta. Assim, as estimativas obtidas podem não ser as melhores no propósito de se prever a resposta cliente.

A Estratégia 2 tem pelo menos duas vantagens sobre a Estratégia 1. Uma das vantagens é permitir a obtenção de uma medida de risco para o cliente sem a necessidade do ajuste preliminar de um modelo de *behavioural scoring* para cada um dos produtos. Para aquelas instituições que não possuem modelos para cada uma das famílias de produtos, a utilização dessa estratégia pode poupar um grande período de tempo de desenvolvimento. Outra vantagem está no fato dos parâmetros associados a todas as famílias de produtos serem estimados conjuntamente. Suponha, por exemplo, que duas variáveis de famílias de produtos diferentes tenham uma correlação muito alta. Em virtude disso, o mais adequado é selecionar apenas uma delas para o modelo final. Na Estratégia 2, isso geralmente é feito porque os parâmetros associados a essas duas variáveis são estimados conjuntamente. Porém, na Estratégia 1, as duas variáveis são estimadas de forma independente,

dificultando qualquer tipo de tratamento de alta correlação entre variáveis de famílias de produtos diferentes. A desvantagem da Estratégia 2 é a não obtenção de estimativas da probabilidade de um cliente se tornar mau em cada uma das famílias de produtos.

A vantagem da Estratégia 3 em relação à 2 está na obtenção do risco associado a cada um dos produtos, já que ela utiliza uma resposta vetorial. Já em relação às estratégias 1a e 1b há pelo menos duas vantagens. A primeira é que, assim como na Estratégia 2, não é necessário o desenvolvimento prévio de vários modelos de *behavioural scoring*. Além disso, a introdução de uma resposta vetorial e o uso de uma técnica estatística adequada para seu tratamento permitem o controle da dependência existente entre o comportamento dos clientes no uso de cada um dos produtos da instituição. Uma desvantagem da Estratégia 3 é a exclusão de um número maior de observações. Isso ocorre porque todos os clientes que possuem pelo menos uma resposta conta indeterminada ou cancelada são excluídos. Outra desvantagem são as limitações dos algoritmos computacionais utilizados na estimação dos modelos de GEE. Eles não toleram uma grande quantidade de variáveis independentes.

Uma questão adicional importante está relacionada com o cálculo periódico do score para todos os clientes da instituição financeira. Na prática, um modelo de *customer scoring* é desenvolvido utilizando uma amostra de clientes e posteriormente este modelo é implantado no sistema do banco para que, mensalmente, o score de cada cliente do banco seja calculado. Nesse cálculo mensal, se o cliente tiver mais uma conta em uma mesma família de produto, recomenda-se que o score de cliente seja calculado utilizando a conta que dê origem ao menor score. Isso deve ser feito mesmo que no desenvolvimento do modelo tenha sido sorteada uma conta para representar o cliente em determinada família de produto. O objetivo desse procedimento é evitar que clientes que apresentem comportamento de alto risco de crédito em uma determinada conta possam receber um bom score de cliente.

## 6 Conclusão

Neste trabalho foram estudados os modelos de *customer scoring*. Esses modelos são utilizados para estimar a probabilidade de um cliente de uma instituição financeira ter problema de crédito em pelo menos um produto, em um horizonte de tempo pré-fixado. Foram apresentadas três estratégias para o desenvolvimento de modelos de *customer scoring*. A primeira, que possui duas variações, é a geralmente utilizada. As demais foram propostas neste trabalho. Foram discutidas as técnicas estatísticas e os modelos relacionados a cada uma das estratégias. Seus desempenhos foram comparados através de uma aplicação a dados reais, utilizando-se algumas medidas de performance que foram definidas. Uma simulação foi ainda desenvolvida para a comparação das estratégias em condições controladas.

Observando-se as características discutidas e os resultados da aplicação e da simulação, a Estratégia 2 parece ser a mais indicada para o desenvolvimento de modelos de *customer scoring*. Considerando-se o coeficiente de Gini, a Estratégia 2 apresentou, em geral, performance ligeira-

mente superior às demais. Além disso, o tempo de desenvolvimento do modelo dessa estratégia é inferior ao observado nas estratégias geralmente utilizadas, já que ela não exige o desenvolvimento prévio de modelos para cada uma das famílias de produtos da instituição.

A Estratégia 3 apresenta alguns problemas práticos, em virtude de limitações dos algoritmos computacionais utilizados para o ajuste de modelos de GEE. Isso prejudicou sua performance na aplicação. No entanto, na simulação, a performance da Estratégia 3 foi superior a das estratégias usualmente utilizadas e apenas ligeiramente inferior ao desempenho da Estratégia 2. Assim, com o aperfeiçoamento dos algoritmos computacionais, essa estratégia pode se tornar uma boa opção. No futuro, ela tende a se tornar a estratégia mais indicada em pelo menos uma situação: quando se deseja também mensurar o risco associado a cada família de produtos, já que isso não pode ser obtido a partir da Estratégia 2.

Deve-se ressaltar que a simulação foi feita em condições bem simplificadas em relação ao que ocorre na prática. O número de variáveis, por exemplo, é geralmente muito maior. Também costuma ser maior o número de famílias de produtos. Além disso, em situações reais, a maioria dos clientes não possuem contas em todas as famílias de produtos. Na prática, também há clientes classificados como indeterminados ou cancelados. Embora, eles não sejam utilizados na estimação dos modelos, eles podem afetar a performance relativa das estratégias, já que a Estratégia 3 descarta um número maior de observações por esse motivo. Assim, para estudos futuros, sugere-se a comparação da performance das estratégias através da simulação de um número maior de condições.

Tabela 8: Proporção de vezes que a Estratégia  $i$  é melhor que a Estratégia  $j$

Condição	Estratégia		Proporção observada	Intervalo de confiança	
	i	j		Lim. inf.	Lim. sup.
Com perturbação e correlação de 0,5	3	2	43%	38%	47%
	3	1b	72%	68%	76%
	3	1a	82%	79%	85%
	2	1b	73%	69%	77%
	2	1a	84%	81%	87%
	1b	1a	65%	61%	69%
Com perturbação e correlação de 0,9	3	2	46%	42%	51%
	3	1b	81%	77%	84%
	3	1a	69%	65%	73%
	2	1b	80%	76%	83%
	2	1a	70%	66%	74%
	1b	1a	32%	28%	36%
Sem perturbação e correlação de 0,5	3	2	44%	40%	49%
	3	1b	65%	61%	69%
	3	1a	83%	79%	86%
	2	1b	66%	62%	70%
	2	1a	84%	81%	87%
	1b	1a	75%	71%	79%
Sem perturbação e correlação de 0,9	3	2	43%	39%	47%
	3	1b	71%	67%	75%
	3	1a	63%	59%	67%
	2	1b	74%	70%	78%
	2	1a	69%	65%	73%
	1b	1a	42%	37%	46%

Tabela 9: Comparação do coeficiente de Gini médio na Estratégia  $i$  e na Estratégia  $j$ 

Condição	Estratégia		Média na Estrat.		Diferença média	I. C. para a dif. média	
	i	j	i	j		Lim. inf.	Lim. sup.
Com perturbação e correlação de 0,5	3	2	0,7629	0,7631	-0,0001	-0,0002	-0,0001
	3	1b	0,7629	0,7611	0,0019	0,0016	0,0022
	3	1a	0,7629	0,7596	0,0033	0,0030	0,0037
	2	1b	0,7631	0,7611	0,0020	0,0017	0,0023
	2	1a	0,7631	0,7596	0,0035	0,0031	0,0038
	1b	1a	0,7611	0,7596	0,0015	0,0011	0,0018
Com perturbação e correlação de 0,9	3	2	0,7416	0,7416	0,0000	-0,0001	0,0001
	3	1b	0,7416	0,7374	0,0042	0,0037	0,0046
	3	1a	0,7416	0,7396	0,0020	0,0016	0,0024
	2	1b	0,7416	0,7374	0,0042	0,0037	0,0046
	2	1a	0,7416	0,7396	0,0020	0,0016	0,0024
	1b	1a	0,7374	0,7396	-0,0022	-0,0026	-0,0017
Sem perturbação e correlação de 0,5	3	2	0,8423	0,8423	0,0000	-0,0001	0,0000
	3	1b	0,8423	0,8415	0,0008	0,0006	0,0010
	3	1a	0,8423	0,8394	0,0029	0,0026	0,0031
	2	1b	0,8423	0,8415	0,0008	0,0006	0,0010
	2	1a	0,8423	0,8394	0,0029	0,0026	0,0032
	1b	1a	0,8415	0,8394	0,0021	0,0018	0,0024
Sem perturbação e correlação de 0,9	3	2	0,8473	0,8477	-0,0005	-0,0006	-0,0003
	3	1b	0,8473	0,8451	0,0022	0,0018	0,0025
	3	1a	0,8473	0,8462	0,0011	0,0008	0,0014
	2	1b	0,8477	0,8451	0,0026	0,0023	0,0030
	2	1a	0,8477	0,8462	0,0015	0,0013	0,0018
	1b	1a	0,8451	0,8462	-0,0011	-0,0014	-0,0007

Tabela 10: Medidas descritivas para a diferença entre a Estratégia  $i$  e a Estratégia  $j$

Condição	Estratégia		Estatísticas da diferença entre as estratégias $i$ e $j$				
	$i$	$j$	Mínimo	Q1	Mediana	Q3	Máximo
Com perturbação e correlação de 0,5	3	2	-0,0035	-0,0003	-0,0001	0,0002	0,0032
	3	1b	-0,0124	-0,0003	0,0018	0,0042	0,0153
	3	1a	-0,0095	0,0011	0,0032	0,0057	0,0166
	2	1b	-0,0118	-0,0002	0,0021	0,0043	0,0142
	2	1a	-0,0096	0,0012	0,0034	0,0059	0,0174
	1b	1a	-0,0132	-0,0011	0,0012	0,0040	0,0145
Com perturbação e correlação de 0,9	3	2	-0,0062	-0,0007	-0,0001	0,0007	0,0062
	3	1b	-0,0141	0,0011	0,0040	0,0075	0,0245
	3	1a	-0,0187	-0,0008	0,0023	0,0051	0,0258
	2	1b	-0,0156	0,0008	0,0042	0,0078	0,0262
	2	1a	-0,0191	-0,0009	0,0021	0,0048	0,0230
	1b	1a	-0,0316	-0,0049	-0,0020	0,0008	0,0232
Sem perturbação e correlação de 0,5	3	2	-0,0025	-0,0003	0,0000	0,0001	0,0023
	3	1b	-0,0068	-0,0009	0,0008	0,0023	0,0085
	3	1a	-0,0046	0,0009	0,0026	0,0046	0,0176
	2	1b	-0,0070	-0,0006	0,0008	0,0023	0,0087
	2	1a	-0,0054	0,0010	0,0028	0,0046	0,0172
	1b	1a	-0,0070	0,0001	0,0018	0,0039	0,0135
Sem perturbação e correlação de 0,9	3	2	-0,0124	-0,0009	-0,0001	0,0004	0,0039
	3	1b	-0,0110	-0,0007	0,0020	0,0046	0,0204
	3	1a	-0,0121	-0,0010	0,0009	0,0031	0,0132
	2	1b	-0,0090	-0,0001	0,0021	0,0049	0,0195
	2	1a	-0,0085	-0,0005	0,0013	0,0033	0,0125
	1b	1a	-0,0181	-0,0029	-0,0005	0,0012	0,0129

## Referências

- Abdou, H. A. and Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management* , **18**, 59-88.
- Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*, Palgrave Macmillan: Basingstoke.

- Cao, R., Vilar, J.M. and Devia, A. (2009). Modelling consumer credit risk via survival analysis. *SORT*, **33**, 3-30.
- Finlay, S. (2012). *Credit scoring, response modeling, and insurance rating: A practical guide to forecasting consumer behavior*, 2 Ed. Palgrave Macmillan: Basingstoke.
- Groom, G. and Gill, L. (1998). *Customer Scoring - Practical Issues for Development Success*. In *InterAct98 Conference*, Fair, Isaac and Company Inc., San Francisco.
- Hardin, J. W. and Hilbe, J. M. (2013). *Generalized estimating equations*. 2ed, Chapman and Hall: Boca Raton.
- Hwang, R. (2013). Predicting issuer credit ratings using generalized estimating equations. *Quantitative Finance*, **13**, 383-398.
- Ilk, O., Pekkurnaz, D. and Cinko, M. (2013). Modeling company failure: a longitudinal study of Turkish banks. *Optimization: A Journal of Mathematical Programming and Operations Research*, in press.
- Johnson Johnson, M. E. (1987). *Multivariate statistical simulation*, John Wiley and Sons: New York.
- Kass, G. V. (1980). An explanatory technique for investigating large quantiles of categorical data. *Applied Statistics*, **29**, 119-127.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W. (2004). *Applied Linear Statistical Models*, 5 Ed. McGraw-Hill: Columbus.
- Liang, K. Y. and Zeger, S. L. (1986). Longitudinal analysis using generalized linear models. *Biometrika*, **73**, 13-22.
- McNab, H. and Wynn, A. (2003). *Principles and Practice of Consumer Credit Risk Management*, 2 Ed. Institute of Financial Services: Kent.
- Sarlija, N., Bencic, M. and Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications*, **36**, 8778-8788.
- Song, P.X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer: New York.
- Stepanova, M. and Thomas, L. C. (2001). PHAB scores: proportional hazards analysis behavioural scores. *Journal of the Operational Research Society*, **52**, 1007-1016.
- Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16**, 149-172.

- Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios*, Oxford University Press: New York.
- Thomas, L. C. (2010). Consumer finance: challenges for operational research. *Journal of the Operational Research Society*, **61**, 41-52
- Thomas, L. C., Ho, J and Scherer, W. T. (2001). Time will tell: behaviour scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*, **12**, 89-103.
- Thomas, L. C., Oliver, R. W. and Hand, D. J. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society*, **56**, 1006-1015.
- Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-130.
- Ziegler, A., Kastner, C and Blettner, M. (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal*, **40**, 115-139.