Análise de Sobrevivência
com fração de fidelizados:
uma aplicação na área de Marketing

Igor Luiz Quidim

*Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestre em Estatística*

Área de Concentração: Estatística
Orientadora: Lúcia Pereira Barroso

- São Paulo, Dezembro de 2005 -

“Não siga aonde leva a trilha.
Em vez disso, vá aonde não há trilha
e deixe seu rastro.”

(conselho extraído do livro:
Commitment to Excellence,
Celebrating Excellence, Inc.)

“À memória de minhas avós
Eunice e Maria.”

Agradecimentos

Agradeço a Deus por ter me dado saúde e motivação para a realização deste trabalho. Além disso, gostaria de agradecer às seguintes pessoas:

À professora Lúcia, que sempre se mostrou muito predisposta a trazer contribuições para essa dissertação.

Aos meus pais Mário e Dora, que me apoiaram desde cedo e possibilitaram a realização de meus objetivos.

À minha namorada Pamela, que muito me apoiou durante esse período de muita dedicação aos estudos, mostrando-se muito companheira.

Ao Raposo, que acreditou no meu trabalho e permitiu a utilização dos dados para a aplicação nesta dissertação.

À Jacqueline, pelas contribuições fundamentais e pelas dispensas para as reuniões com os professores.

Ao professor Antonio Carlos, que sempre se mostrou muito prestativo para tirar dúvidas e dar excelentes sugestões.

Aos amigos do banco, que através de várias dicas contribuíram muito para o resultado deste trabalho.

Aos professores do Departamento de Estatística do IME-USP que muito contribuíram para minha formação.

Resumo

O mercado de cartões de crédito brasileiro encontra-se em um período de extrema expansão. Há uma forte concorrência entre as instituições financeiras nas quais o cartão de crédito é um de seus principais produtos de crédito. Isso ocasiona a necessidade de identificar o risco de cancelamento dos clientes atuais para tomar ações preventivas. Modelos estatísticos, denominados *attrition score* são utilizados para esta finalidade. Neste trabalho, apresentamos uma alternativa aos modelos usuais, utilizando a Análise de Sobrevida com fração de cura, cuja aplicação tem aumentado nos últimos anos. Em nosso contexto, denominamos esses modelos como modelos de Análise de Sobrevida com fração de fidelizados. A partir de uma base de dados fornecida por uma grande instituição financeira brasileira, aplicamos e comparamos as seguintes técnicas: Regressão Logística, Análise de Sobrevida Paramétrica e Semiparamétrica ambas com fração de fidelizados. A primeira técnica já é muito utilizada por profissionais do mercado e as demais são novas alternativas para serem aplicadas nesse tipo de estudo.

Abstract

The Brazilian market of credit cards is in a period of extreme growth, with heavy competition among financial institutions where credit cards are the most important product. This environment demands the development of tools appropriated to evaluate the attrition risk of customers. Some usual statistical models (e.g. Logistic Regression) have already been used for this purpose. In this study, we consider the use of survival analysis with cure rate models to study the history of credit card cancelations. In our context, we called this technique “survival analysis with fidelity rate”. The new approach is compared with logistic regression models, considering a real dataset.

Sumário

1	Introdução	1
2	Análise de Sobrevivência com fração de fidelizados	4
2.1	Definição	4
2.2	O modelo de mistura Weibull	6
2.2.1	Estimação dos parâmetros	8
2.3	Modelo de Cox com fração de fidelizados	11
2.3.1	Estimação dos parâmetros	13
2.4	Modelo de Cox ponderado	17
2.5	Testando a presença de fidelizados	18
2.6	Testando a suficiência do horizonte de tempo	19
2.7	Medidas de avaliação dos modelos	21
2.7.1	Estatística de Kolmogorov-Smirnov	21
2.7.2	A Curva de Lorenz e o Coeficiente de Gini	22
3	Aplicação	25
3.1	Descrição do estudo	26
3.2	Tratamento das variáveis preditoras	28
3.3	Método de seleção de variáveis	29
3.4	Resultados	31
3.4.1	Testando a presença de clientes fidelizados	32

3.4.2	Testando a suficiência do horizonte de tempo	33
3.4.3	Modelo de mistura Weibull	34
3.4.4	Modelo de mistura de Cox	37
3.4.5	Modelo de mistura de Cox ponderado	40
3.4.6	Modelo final de mistura Weibull	43
3.4.7	Modelo final de mistura de Cox	45
3.4.8	Modelo final de mistura de Cox ponderado	47
3.4.9	Modelo de Regressão Logística	49
4	Comparação dos Modelos	52
5	Conclusão	58
A	Derivadas de 2ª ordem	60
B	Período Analisado	63
C	Variáveis Preditoras	65
	Referências Bibliográficas	75

Capítulo 1

Introdução

As instituições financeiras na entrada do novo século estão se conscientizando da necessidade de intensificar o foco no Marketing de Relacionamento se quiserem manter e ampliar seu mercado. Apesar dos orçamentos das instituições financeiras ainda estarem direcionados para o marketing de massa e programas de conquista de novos clientes, os bancos estão aprendendo que podem reduzir as taxas de cancelamento voluntário e melhorar a rentabilidade vendendo mais produtos e serviços para os clientes atuais, através de técnicas conhecidas como Marketing de Relacionamento, que pode ser definido como um conjunto de estratégias de negócio pelas quais a empresa objetiva construir um relacionamento com seus clientes de maneira personalizada e duradoura. Para isso é essencial que a mesma dedique-se a uma constante melhora nesse relacionamento, para que ambas as partes sejam beneficiadas.

A gestão do relacionamento com o cliente ou *Customer Relationship Management* (CRM) está totalmente ligada à filosofia do Marketing de Relacionamento aliado à utilização constante de informação.

A utilização desse tipo de gestão já era praticada no passado. Um pequeno comércio atendia às pessoas de forma personalizada e quase sempre o próprio dono do negócio conhecia seus clientes pelo nome, bem como seus hábitos e preferências

de compras.

O conceito de CRM parte da premissa de que é de cinco até dez vezes mais caro obter um novo cliente do que reter os existentes, e que o importante não é ter uma imensa carteira de clientes, mas uma boa base de clientes fiéis (Oliveira, 2000).

Cada vez mais as empresas estão se conscientizando de que nada adianta gastar com campanhas publicitárias e de marketing, se não for possível manter os clientes fiéis a seus produtos/serviços.

Em mercados altamente competitivos, a fidelização e retenção de clientes tem papel fundamental nas empresas. Identificar, antecipadamente, clientes com alto potencial de ruptura de relacionamento permitindo ações preventivas, se tornou imprescindível para a manutenção ou acréscimo da lucratividade das empresas.

Este trabalho considera o uso de técnicas de Análise de Sobrevivência na modelagem do tempo até o cancelamento de cartões de crédito de uma grande instituição financeira brasileira. Uma vez que a maioria dos clientes dessa instituição já possui um grande período de relacionamento, não é difícil inferir a presença de clientes fidelizados ao cancelamento do cartão de crédito. Entende-se por fidelizados, os clientes que apresentam baixo risco de cancelar o cartão.

Além disso, estaremos comparando os modelos usualmente utilizados para discriminar os bons e maus clientes com os modelos apresentados nesta dissertação que podem também ser vistos como modelos com poder discriminatório.

O trabalho é desenvolvido na seguinte seqüência: no Capítulo 2, são apresentadas as técnicas de Análise de Sobrevivência com fração de fidelizados; no Capítulo 3, são descritos os dados utilizados, as variáveis consideradas e as aplicações dos modelos descritos no Capítulo 2; além disso, é apresentada uma proposta para seleção de variáveis para obtenção de um modelo final. Avaliação da performance e comparação com modelo usualmente utilizado no tratamento dos dados considerados são apresentados no Capítulo 4. E, por fim, as conclusões finais e sugestões

1. Introdução

para novos estudos estão no Capítulo 5.

Capítulo 2

Análise de Sobrevivência com fração de fidelizados

2.1 Definição

Os modelos de Análise de Sobrevivência incorporando o fato de existirem indivíduos imunes ao evento de interesse possuem a vantagem em relação aos modelos de sobrevivência usuais de incorporarem a heterogeneidade de duas subpopulações (susceptíveis e imunes).

Algumas aplicações desses modelos são apresentados em Boag (1949), Yamaguchi (1992), Berkson and Gage (1952) e Pocock, Gore and Kerr (1982) com aplicação a dados de câncer de mama. Farewell (1977; 1982; 1986) apresenta algumas generalizações do modelo proposto por Boag (1949). Em Maller and Zhou (1996) também são descritas aplicações nas áreas de epidemiologia, criminologia, educação, entre outras. Abreu (2005) em sua dissertação de mestrado utilizou esse tipo de modelagem numa aplicação a dados de credit scoring.

Esses modelos são conhecidos como modelos de Análise de Sobrevivência com fração de imunes (*cure rate models*) e neste trabalho, são denominados mode-

los com fração de fidelizados. Dentre esses, o tipo mais comum é o modelo de mistura (*mixture model*) no qual se considera que a população é dividida em duas subpopulações: imunes (no nosso caso, fidelizados) e susceptíveis (não fidelizados) ao evento de interesse.

Admitimos que os indivíduos podem ser classificados como fidelizados (sem possibilidade de apresentar o evento de interesse) com probabilidade q ou serem não fidelizados com probabilidade $p = 1 - q$.

Assim, como introduzido em Berkson and Gage (1952), a cada indivíduo associamos uma variável aleatória T representando o tempo até a ocorrência do evento de interesse ou até a censura, caracterizado por uma função de sobrevivência imprópria que pode ser escrita como

$$S_1(t) = P(T > t) = q + p \times S^*(t) \quad (2.1)$$

em que $S^*(.)$ é a função de sobrevivência para indivíduos não fidelizados, de tal forma que para $t \rightarrow +\infty$, $S^*(t) \rightarrow 0$ e assim, $\lim_{t \rightarrow \infty} S_1(t) = q (> 0)$.

Nos modelos de Análise de Sobrevivência com fração de fidelizados, os indivíduos fiéis não são identificados a priori, mas a presença dos mesmos pode ser inferida se a quantidade de indivíduos censurados no estudo for suficientemente grande. Então para a utilização desse tipo de modelo é necessário testar:

- a presença de indivíduos fidelizados;
- se o horizonte de tempo considerado no estudo é suficiente para verificar a ocorrência de fidelizados.

Note que é recomendável antes do ajuste do modelo com fração de fidelizados a realização desses testes, pois assim é possível detectar se realmente há evidências de indivíduos fidelizados na base de dados e se o horizonte de tempo considerado para o acompanhamento de cada indivíduo foi suficiente para se inferir a presença desses indivíduos, com um certo grau de confiança.

2.2 O modelo de mistura Weibull

Em (2.1) pode-se adotar que a distribuição dos tempos de cancelamento dos indivíduos não fidelizados ao evento de interesse segue uma distribuição Weibull.

Suponha que $t_1, \dots, t_n$ sejam os tempos até a ocorrência do cancelamento do cartão de crédito para n clientes do estudo e seja c_i o indicador de cancelamento dado por

$$c_i = \begin{cases} 1, & \text{se o } i\text{-ésimo cliente cancela o cartão} \\ 0, & \text{caso contrário.} \end{cases}$$

Considere uma variável aleatória binária b_i , $i=1, \dots, n$ indicando se o i -ésimo indivíduo está em risco ou não em relação ao evento de interesse no final do período de acompanhamento

$$b_i = \begin{cases} 1, & \text{se o } i\text{-ésimo cliente é não fidelizado} \\ 0, & \text{caso contrário.} \end{cases}$$

Assume-se que os b_i 's são independentes e identicamente distribuídos (i.i.d.) com $P[b_i = 1] = p_i$ indicando a probabilidade do indivíduo i ser não fidelizado ao evento e $P[b_i = 0] = 1 - p_i$, a probabilidade do indivíduo ser fidelizado.

Associado ao i -ésimo indivíduo, há um vetor de r^* covariáveis dado por $\mathbf{W}_i = (W_{0i} = 1, W_{1i}, \dots, W_{r^*i}) \in \mathbb{R}^{r^*+1}$. Como proposto em Farewell (1982) pode-se escrever

$$p_i = P[b_i = 1] = \frac{e^{\beta^\top \mathbf{Y}_i}}{1 + e^{\beta^\top \mathbf{Y}_i}} = 1 - P[b_i = 0], \quad (2.2)$$

em que o vetor $\mathbf{Y}_i = (Y_{0i} = 1, Y_{1i}, \dots, Y_{r_2i}) \in \mathbb{R}^{r_2+1}$, $1 \leq r_2 \leq r^*$ é um subvetor do vetor $\mathbf{W}_i$, $i = 1, \dots, n$ e $\beta^\top = (\beta_0, \beta_1, \dots, \beta_{r_2}) \in \mathbb{R}^{r_2+1}$ é um vetor de parâmetros desconhecidos.

Os b_i 's não são observados mas a presença dos clientes fiéis é inferida se a quantidade de indivíduos censurados no estudo for suficientemente grande.

Para os clientes não fidelizados $b_i = 1$, os tempos até a ocorrência do evento são considerados independentes e distribuídos segundo um modelo Weibull com função densidade dada por

$$f_i(t_i, \alpha | b_i = 1) = \delta \lambda_i^\delta (t_i)^{\delta-1} e^{-(\lambda_i t_i)^\delta}, \quad (2.3)$$

em que $t_i \geq 0$, $\lambda_i > 0$, $\delta > 0$.

O parâmetro λ_i está relacionado ao vetor de covariáveis $\mathbf{X}_i = (X_{0i} = 1, X_{1i}, \dots, X_{r_{1i}}) \in \mathbb{R}^{r_1+1}$ através da seguinte relação

$$\lambda_i = e^{\alpha^\top \mathbf{X}_i}, \quad (2.4)$$

com $\alpha^\top = (\alpha_0, \alpha_1, \dots, \alpha_{r_1}) \in \mathbb{R}^{r_1+1}$ sendo o vetor de parâmetros desconhecidos e $\mathbf{X}_i$ subvetor de $\mathbf{W}_i$, $1 \leq r_1 \leq r^*$, $i = 1, \dots, n$.

Observe que (2.4) representa o parâmetro de escala do modelo Weibull escrito na forma de modelos lineares generalizados (McCullagh and Nelder, 1989): A função log-linear é convenientemente utilizada para escrever a relação entre λ_i e o vetor de covariáveis $\mathbf{X}_i$.

Note que a contribuição para a verossimilhança de um indivíduo que falha em t_i é dada por

$$P[b_i = 1]f(t_i, \alpha^\top | b_i = 1) = p_i \delta \lambda_i^\delta (t_i)^{\delta-1} e^{-(\lambda_i t_i)^\delta}, \quad i = 1, \dots, n \quad (2.5)$$

e a contribuição de um indivíduo em risco em t_i pode ser escrita como

$$\begin{aligned} P[T_i > t_i; \alpha_i^\top] &= P[b_i = 0]P[T_i > t_i; \alpha_i^\top | b_i = 0] + \\ &+ P[b_i = 1]P[T_i > t_i; \alpha_i^\top | b_i = 1] = \\ &= 1 - p_i[1 - e^{-(\lambda_i t_i)^\delta}]. \end{aligned} \quad (2.6)$$

Logo, utilizando (2.5) e (2.6), calcula-se a contribuição do i -ésimo indivíduo para a verossimilhança do modelo de mistura Weibull, dada por

$$f_i(t_i; \alpha, \beta) = [p_i \delta \lambda_i^\delta (t_i)^{\delta-1} e^{-(\lambda_i t_i)^\delta}]^{c_i} [1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})]^{1-c_i} \quad (2.7)$$

$i = 1, \dots, n$.

Então, utilizando os resultados apresentados acima, dada uma amostra aleatória de n indivíduos com os respectivos tempos até a ocorrência do evento ou censura $T_1, \dots, T_n$ e os indicadores de cancelamento $c_1, \dots, c_n$, a função de verossimilhança de $\boldsymbol{\theta}^\top = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \delta)$ pode ser escrita como

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^n [p_i \delta \lambda_i^\delta(t_i)^{\delta-1} e^{-(\lambda_i t_i)^\delta}]^{c_i} [1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})]^{1-c_i}. \quad (2.8)$$

Vale ressaltar que o mecanismo de censura é suposto não informativo, uma vez que a distribuição dos tempos de censura não envolve os parâmetros de interesse (λ_i, p_i, δ) .

2.2.1 Estimação dos parâmetros

Nessa seção será considerado o uso do método de estimação de máxima verossimilhança. Tomando o logaritmo da função de verossimilhança dada em (2.8), tem-se que

$$\begin{aligned} l_n(\boldsymbol{\theta}) &= \sum_{i=1}^n \left\{ c_i [\ln(p_i) + \ln(\delta) + \delta \ln(\lambda_i) + (\delta - 1) \ln(t_i) - (\lambda_i t_i)^\delta] + \right. \\ &\quad \left. + (1 - c_i) \ln[1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})] \right\}. \end{aligned} \quad (2.9)$$

Desta forma, o escore é dado pelo vetor

$$S_n(\boldsymbol{\theta}) = \frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} \\ \frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ \frac{\partial l_n(\boldsymbol{\theta})}{\partial \delta} \end{bmatrix}_{r \times 1} = \sum_{i=1}^n U_i s_i(\boldsymbol{\theta}) \quad (2.10)$$

com $r = (r_1 + r_2 + 3)$ e a matriz U_i dada por

$$\begin{pmatrix} X_i & 0 & 0 \\ 0 & Y_i & 0 \\ 0 & 0 & 1 \end{pmatrix}_{r \times 3} \quad (2.11)$$

com $i = 1, \dots, n$.

O estimador de máxima verossimilhança de θ é encontrado maximizando (2.9) que é obtido resolvendo as equações

$$S_n(\theta) = \sum_{i=1}^n s_i(\theta) = 0 \quad (2.12)$$

com $s_i(\theta) = (s_{i1}(\theta), s_{i2}(\theta), s_{i3}(\theta))'$ em que

$$s_{i1}(\theta) = \frac{\partial l_n(\theta)}{\partial \alpha} = \sum_{i=1}^n \left\{ c_i(\delta - \delta(\lambda_i t_i)^\delta) - (1 - c_i) \frac{p_i(\lambda_i t_i)^\delta \delta e^{-(\lambda_i t_i)^\delta}}{1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})} \right\}, \quad (2.13)$$

$$s_{i2}(\theta) = \frac{\partial l_n(\theta)}{\partial \beta} = \sum_{i=1}^n \left\{ c_i(1 - p_i) + (1 - c_i) \frac{-p_i(1 - e^{-(\lambda_i t_i)^\delta}) + p_i^2(1 - e^{-(\lambda_i t_i)^\delta})}{1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})} \right\}, \quad (2.14)$$

e

$$s_{i3}(\theta) = \frac{\partial l_n(\theta)}{\partial \delta} = \sum_{i=1}^n \left\{ c_i \left(\frac{1}{\delta} + \ln(\lambda_i) - (\lambda_i t_i)^\delta \ln(\lambda_i t_i) \right) - c_i \ln(t_i) - (1 - c_i) \frac{p_i(\lambda_i t_i)^\delta \ln(\lambda_i t_i) e^{-(\lambda_i t_i)^\delta}}{1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})} \right\}. \quad (2.15)$$

Para a resolução dessas equações, deve-se utilizar um método numérico como o algoritmo de Newton-Raphson, pois não é possível obter uma expressão fechada para o estimador de θ . Esse método necessita das segundas derivadas do logaritmo da função de verossimilhança para sua implementação. Então, escrevendo

na forma matricial, tem-se que

$$\begin{aligned}
 F_n(\boldsymbol{\theta}) = -\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} &= - \begin{pmatrix} \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \alpha^2} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \alpha \partial \beta} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \alpha \partial \delta} \\ \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \beta \partial \alpha} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \beta^2} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \beta \partial \delta} \\ \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \delta \partial \alpha} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \delta \partial \beta} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \delta^2} \end{pmatrix}_{r \times r} \\
 &= \sum_{i=1}^n U_i \mathcal{F}_i(\boldsymbol{\theta}) U_i', \tag{2.16}
 \end{aligned}$$

em que $\mathcal{F}_i(\boldsymbol{\theta})$ é uma matriz de dimensão 3×3 simétrica com componentes f_i^{jk} , j e $k=1, \dots, 3$, que são detalhadas no Apêndice A.

Pode-se mostrar que a matriz de informação esperada $D_n = E \{F_n(\boldsymbol{\theta})\}$ necessária para o desenvolvimento de inferências sobre os parâmetros de $\boldsymbol{\theta}$, satisfaz a relação $D_n = E \{S_n(\boldsymbol{\theta}) S_n'(\boldsymbol{\theta})\}$ (Maller and Zhou, 1996).

Assim pode-se escrever a esperança de (2.16) como

$$D_n = \sum_{i=1}^n U_i \mathcal{D}_i(\boldsymbol{\theta}) U_i' \tag{2.17}$$

com $\mathcal{D}_i(\boldsymbol{\theta}) = E \{\mathcal{F}_i(\boldsymbol{\theta})\}$, $i = 1, \dots, n$.

A análise inferencial de $\boldsymbol{\theta}$ pode ser desenvolvida através da distribuição assintótica do vetor $\hat{\boldsymbol{\theta}}$, dada por

$$D_n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N_r(0; \mathbf{I}_r).$$

A partir do artigo de Peng, Dear and Denham (1998), o primeiro autor desenvolveu um programa no software S-Plus (Venables and Ripley, 1997) denominado *gfcure* para o ajuste de modelos paramétricos com fração de fidelizados. O programa *gfcure* divide a saída do modelo Weibull com fração de fidelizados em

2 partes: a parte do modelo de sobrevivência Weibull que modela o tempo de sobrevivência dos indivíduos não fidelizados e a parte Logística, que modela a probabilidade p_i do i -ésimo indivíduo ser não fidelizado.

2.3 Modelo de Cox com fração de fidelizados

O modelo de Cox, proposto por Cox (1972), é um dos mais utilizados na análise de dados de sobrevivência. Este modelo, assim como outros na literatura, modela os dados de sobrevivência na presença de covariáveis, utilizando a função de taxa de falha.

Considere b_i como descrito anteriormente, ou seja, $b_i = 0$ se o indivíduo é fidelizado ou $b_i = 1$ se o indivíduo é não fidelizado. Kuk and Chen (1992) propuseram escrever o modelo de Cox com fração de fidelizados condicional a $b_i = 1$ da seguinte forma

$$\alpha(t|b_i = 1; \mathbf{X}) = \alpha_0(t|b_i = 1)e^{\gamma^T \mathbf{X}}, \quad (2.18)$$

com:

$\mathbf{X}$: vetor de covariáveis,

$\alpha_0(t|b_i = 1)$: função de risco basal.

Mais uma vez, vale ressaltar que os b_i 's não são observados mas a presença dos clientes fiéis é inferida se a quantidade de indivíduos censurados no estudo for suficientemente grande.

Pode-se também obter a função de sobrevivência condicional do modelo de Cox, que é dada por

$$S(t|b_i = 1; \mathbf{X}) = [S_0(t|b_i = 1)]^{e^{\gamma^T \mathbf{X}}}, \quad (2.19)$$

em que:

$S_0(t|b_i = 1)$: função de sobrevivência basal.

Suponha que $(T_i, c_i, \mathbf{X}_i, \mathbf{Y}_i)$ são os dados observados para o i -ésimo indivíduo, $i = 1, \dots, n$, em que:

T_i : tempo até a observação do evento ou até a censura,

c_i : indicador de falha,

$\mathbf{X}_i$ e $\mathbf{Y}_i$: vetores de covariáveis, respectivamente, da parte de Cox e da parte logística do modelo com fração de fidelizados.

Portanto, a contribuição para a verossimilhança de um indivíduo que falha em t_i é dada por

$$P[b_i = 1]f(t_i; \mathbf{X}_i|b_i = 1) = p_i f(t_i|b_i = 1; \mathbf{X}_i), \quad (2.20)$$

$$\text{com } p_i = P[b_i = 1] = \frac{e^{\boldsymbol{\beta}^\top \mathbf{Y}_i}}{1 + e^{\boldsymbol{\beta}^\top \mathbf{Y}_i}}.$$

E a contribuição de um indivíduo que está em risco em t_i pode ser escrita como

$$P[T_i > t_i; \mathbf{X}_i] = (1 - p_i) + p_i S(t_i|b_i = 1; \mathbf{X}_i). \quad (2.21)$$

Logo, a verossimilhança pode ser escrita como

$$L_n(\boldsymbol{\beta}, \boldsymbol{\gamma}, \alpha_0) = \prod_{i=1}^n [p_i f(t_i|b_i = 1; \mathbf{X}_i)]^{c_i} [1 - p_i + p_i S(t_i|b_i = 1; \mathbf{X}_i)]^{1-c_i}, \quad (2.22)$$

que depende dos parâmetros: $\boldsymbol{\beta}$ da parte logística; e $\boldsymbol{\gamma}$ e $A_0(t)$ da parte de sobrevivência do modelo de Cox com fração de fidelizados.

Note que $A_0(t)$ é a função de taxa de falha (cancelamento) basal acumulada, definida por $A_0(t) = \int_0^t \alpha_0(s) ds$.

O objetivo é obter as estimativas de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ que maximizem (2.22). No modelo de Cox tradicional, (2.22) é escrito de tal forma que não dependa de $A_0(t)$ através de uma verossimilhança parcial (Cox, 1975).

Breslow (1972) apresenta a construção de uma função de verossimilhança semi-paramétrica completa na qual $A_0(t)$ é trocada por um estimador de máxima verossimilhança não paramétrico, dado $\boldsymbol{\gamma}$. Breslow mostrou que essa função de

verossimilhança, parcialmente maximizada de γ , é proporcional à verossimilhança parcial.

No entanto, essa metodologia não funciona para o modelo de Cox com fração de fidelizados. Diferentemente do modelo de Cox tradicional no qual pouco se perde de informação quando se elimina $S_0(t)$, não se pode desconsiderar $S_0(t|b_i = 1)$ para a estimação dos parâmetros de (2.22) sem perder informação sobre β .

Em Peng (2003) é proposto um método de maximização que utiliza a verossimilhança completa para a estimação dos parâmetros, conforme discutido a seguir.

2.3.1 Estimação dos parâmetros

Note que se $c_i = 1$ então $b_i = 1$, ou seja, se o i -ésimo indivíduo falha então ele é considerado susceptível. No entanto, se $c_i = 0$, não se tem informação sobre b_i , ou seja, b_i pode assumir o valor 0 ou 1.

Nesse caso, b_i é parcialmente observado e será utilizado para compor a verossimilhança completa, que pode ser escrita a partir de (2.22) como

$$\begin{aligned} L_C(\beta, \gamma, A_0; b_i) &= \prod_{i=1}^n \left\{ [p_i f(t_i | \mathbf{X}_i)]^{b_i} \right\}^{c_i} \left\{ (1 - p_i)^{1-b_i} [p_i S(t_i | \mathbf{X}_i)]^{b_i} \right\}^{1-c_i} \\ &= \prod_{i=1}^n \left\{ \frac{[p_i f(t_i | \mathbf{X}_i)]^{b_i}}{(1 - p_i)^{1-b_i} [p_i S(t_i | \mathbf{X}_i)]^{b_i}} \right\}^{c_i} (1 - p_i)^{1-b_i} [p_i S(t_i | \mathbf{X}_i)]^{b_i} \end{aligned}$$

Observe que $\alpha(t_i | \mathbf{X}_i) = \frac{f(t_i | \mathbf{X}_i)}{S(t_i | \mathbf{X}_i)}$ e portanto, tem-se que

$$\begin{aligned} L_C(\beta, \gamma, A_0; b_i) &= \prod_{i=1}^n \left\{ \frac{[\alpha(t_i | \mathbf{X}_i)]^{b_i}}{(1 - p_i)^{1-b_i}} \right\}^{c_i} (1 - p_i)^{1-b_i} [p_i S(t_i | \mathbf{X}_i)]^{b_i} \\ &= \prod_{i=1}^n p_i^{b_i} (1 - p_i)^{1-b_i} \prod_{i=1}^n [\alpha(t_i | \mathbf{X}_i)]^{c_i b_i} [S(t_i | \mathbf{X}_i)]^{b_i} \\ &= \prod_{i=1}^n L_1(\beta; b_i) L_2(\gamma, A_0; b_i). \end{aligned} \tag{2.23}$$

Para realizar a estimação dos vetores de parâmetros β e γ , Peng (2003) sugere a utilização do algoritmo EM da seguinte maneira:

Passo E: Calcula-se a esperança condicional de b_i dados os valores atuais de β , γ e $S_{u_0}(t_i|X_i)$ (função de sobrevivência dos indivíduos não fidelizados) de tal modo que

$$\begin{aligned} a_i &= P[b_i = 1|\beta, \gamma, S_{u_0}(t)] \\ &= \frac{P[b_i = 1; \beta] [S_{u_0}(t_i|X_i)]^{e^{\gamma^\top X}}}{(1 - P[b_i = 1; \beta]) + P[b_i = 1; \beta] [S_{u_0}(t_i|X_i)]^{e^{\gamma^\top X}}} \\ &= \frac{p_i [S_{u_0}(t_i|X_i)]^{e^{\gamma^\top X}}}{(1 - p_i) + p_i [S_{u_0}(t_i|X_i)]^{e^{\gamma^\top X}}} \end{aligned} \quad (2.24)$$

para os indivíduos censurados.

Então, tem-se que

$$g_i = E[b_i|\beta, \gamma, S_{u_0}(t)] = c_i + (1 - c_i) \frac{p_i [S_{u_0}(t_i|X_i)]^{e^{\gamma^\top X}}}{(1 - p_i) + p_i [S_{u_0}(t_i|X_i)]^{e^{\gamma^\top X}}}. \quad (2.25)$$

Logo, o passo E do algoritmo EM atualiza, a cada iteração a probabilidade do i -ésimo indivíduo ser não fidelizado, dado os valores dos parâmetros β , γ e $S_{u_0}(t_i|X_i)$.

Passo M: Maximiza-se L_1 com relação aos parâmetros β e L_2 com relação aos parâmetros γ e A_0 para um dado g_i fixado, L_1 e L_2 dados em (2.23).

A vantagem de se utilizar o algoritmo EM é que os estimadores de máxima verossimilhança podem ser obtidos separadamente, uma vez que em (2.23), L_1 depende apenas do parâmetro β e L_2 apenas da distribuição dos tempos de falha dos indivíduos não fidelizados.

Observe que L_1 é a função de verossimilhança de um modelo de regressão logística usual e pode ser maximizada utilizando Newton-Raphson (Hosmer and Lemeshow, 1989). Com relação a L_2 , note que é a função de verossimilhança do modelo de Cox usual com a inclusão de um termo b_i . Vale ressaltar que, por conta da presença do componente não-paramétrico $A_0(t)$, não é possível utilizar o

método de máxima verossimilhança diretamente. Peng (2003) utiliza neste passo a verossimilhança parcial de Cox (1975).

Reescrevendo L_2 segundo o modelo de Cox, assumindo a aproximação de Efron (1977), ou seja, utilizando uma aproximação para o caso de ocorrência de muitos T_i 's empatados, tem-se que

$$L_2 = \prod_{j=1}^k \frac{e^{S_j^\top \gamma}}{\prod_{l \in R_{t_j}} \left\{ g_i \left[\sum_{l \in R_{t_j}} e^{X_l^\top \gamma} - r \bar{A}(\gamma, t_j) \right] \right\}}, \quad (2.26)$$

em que

$$S_j = \sum_{i=1}^{d_j} X_{j_i} \text{ e } \bar{A}(\gamma, t_j) = d_j^{-1} \sum_{l \in D(t_j)} e^{X_l^\top \gamma},$$

d_j é o número de falhas no instante t_j ,

R_{t_j} é o conjunto de índices de todas as unidades em risco em t_j ,

D_{t_j} é o conjunto de índices das unidades que falham em t_j .

Note que (2.26) é a função de verossimilhança do modelo de taxas de cancelamento proporcionais (Cox, 1975) com a inclusão de uma covariável adicional g_i com coeficiente fixo e igual a 1. Essa é a aproximação de Efron com a inclusão de um termo g_i . Essa aproximação foi utilizada neste trabalho pois a base de dados utilizada para aplicação dos modelos contém muitos indivíduos com tempos de falha (ou de censura) empatados.

Para completar o passo E do algoritmo EM, é necessário calcular $S_{u_0}(t_i)$ para um dado γ .

A função de verossimilhança basal pode ser estimada por um método similar àquele utilizado no modelo de Cox. Kalbfleisch and Prentice (1980) notam que a contribuição para a verossimilhança de um indivíduo que falha em t_j é $[S_{u_0}(t_j)]^{e^{\gamma^\top X_i}} - [S_{u_0}(t_j + 0)]^{e^{\gamma^\top X_i}}$ e a contribuição de um indivíduo censurado em t_j é $[S_{u_0}(t_j)]^{e^{\gamma^\top X_i}}$.

A maximização é encontrada tomando-se $S_{u_0}(t + 0) = S_{u_0}(t_j + 0)$ para $t \in$

$[t_j, t_{j+1})$, o que torna a função basal discreta com probabilidade apenas nos tempos de falha.

Seja

$$\alpha_j = \frac{S_{u_0}(t_{j+1})}{S_{u_0}(t_j)}$$

$j = 1, \dots, k$,

então:

$$\alpha_0 = 1, S_{u_0}(t) = \prod_{j:t_j < t} \alpha_j, S_{u_0}(t_j) = \prod_{l=1}^{j-1} \alpha_l, S_{u_0}(t_j + 0) = \prod_{l=1}^j \alpha_l.$$

A partir desses resultados, pode-se reescrever L_2 como

$$L_2 = \prod_{j=1}^k \left\{ \prod_{i \in D_j} (1 - \alpha_j e^{\gamma^\top x_i}) \prod_{i \in R_j - D_j} \alpha_j^{g_i e^{\gamma^\top x_i}} \right\}, \quad (2.27)$$

e o estimador de máxima verossimilhança de α_j dados os valores de β e γ pode ser escrito como

$$\hat{\alpha}_j \approx \exp \left(\frac{-d_j}{\sum_{i \in R_j} g_i e^{\gamma^\top Y_i}} \right). \quad (2.28)$$

Sob (2.28), a função de sobrevivência basal estimada $\hat{S}_{u_0}(t)$ é dada por:

$$\hat{S}_{u_0}(t) = \exp \left(- \sum_{j:t_j < t} \frac{d_j}{\sum_{i \in R_j} g_i e^{\gamma^\top Y_i}} \right). \quad (2.29)$$

Vale ressaltar que essa é a abordagem sugerida por Peng (2003). Com base nesse artigo, o autor desenvolveu um programa denominado *semicure*, desenvolvido no software S-Plus (Venables and Ripley, 1997) para ajuste do modelo de Cox com fração de fidelizados. Esse programa divide a saída do modelo de Cox com fração de fidelizados em duas partes: a parte do modelo de Cox que modela os tempos de sobrevivência dos indivíduos não fiéis, e a parte Logística, que modela a probabilidade p_i do i -ésimo indivíduo ser não fidelizado.

2.4 Modelo de Cox ponderado

O modelo de Cox ponderado é comumente utilizado em estudos nos quais se observa a presença de variáveis com valores faltantes (“missing values”). Em Binder (1992) e Pugh, Robbins and Harrington (1992) são apresentadas aplicações desse tipo de modelo na área médica.

Uma das alternativas quando se verifica indivíduos para os quais há presença de covariáveis com valores faltantes na base de dados é desconsiderá-los da análise. No entanto, com isso pode-se estar reduzindo a eficiência da técnica adotada e/ou viesando a inferência.

Pugh, Robbins and Harrington (1992) propuseram uma correção ponderando cada indivíduo no modelo de Cox pelo inverso da probabilidade de possuir valores faltantes, ou seja, indivíduos com baixa probabilidade de possuir valores faltantes entram com um peso maior no ajuste do modelo. Usualmente, modela-se essa probabilidade através do modelo de Regressão Logística.

Em Binder (1992) e Therneau and Grambsch (2000) é descrita a teoria para estimação dos parâmetros do modelo de Cox ponderado, bem como a metodologia para obtenção das variâncias dessas estimativas utilizando o método de sanduíche modificado.

Este método consiste na utilização do estimador da variância dos coeficientes dado por $A'[I - B(B'B)^{-1}B']A$, em que I é a matriz identidade, B a matriz $dfbeta$ do modelo logístico e A é a matriz $dfbeta$ do modelo de Cox ponderado, ao invés do estimador usualmente utilizado para o modelo de Cox dado por $A'A$ (Pugh, Robbins and Harrington, 1992). Isso porque no modelo de Cox ponderado deve-se levar em conta a informação do modelo logístico utilizado para obter a probabilidade de cada indivíduo possuir valores faltantes.

Nesse trabalho, o uso da ponderação no modelo de Cox será utilizado como alternativa ao procedimento utilizado em Peng and Dear (2000). No programa

semicure desenvolvido pelo autor para o ajuste do modelo de Cox com fração de fidelizados, na parte de estimação do modelo de Cox, os indivíduos com alta probabilidade de serem fiéis são excluídos do ajuste. Na nova proposta, ao invés da exclusão dos indivíduos, utiliza-se uma ponderação $1/(1 - g_i)$, em que g_i é considerado como em (2.25). Observe que com essa ponderação, um peso maior é atribuído para indivíduos com baixa probabilidade de serem fidelizados em cada uma das iterações do algoritmo EM descrito anteriormente.

2.5 Testando a presença de fidelizados

Ghitany and Maller (1995) e Maller and Zhou (1996) descrevem um teste para verificar a presença ou ausência de indivíduos fidelizados na amostra, ou seja

$$\begin{cases} H_0 : \pi = 1. \\ H_1 : \pi < 1. \end{cases} \quad (2.30)$$

em que π é a fração de indivíduos não fidelizados.

Para testar essas hipóteses, pode-se considerar o teste da Razão de Verossimilhanças, que pode ser descrito como

$$d_n = 2 \left\{ l_n(\hat{\theta}_n) - l_n(\hat{\theta}_{H_0}) \right\}, \quad (2.31)$$

em que:

$l_n(\hat{\theta}_n)$ é a log verossimilhança do modelo aplicada na estimativa de máxima verossimilhança sob H_1 .

$l_n(\hat{\theta}_{H_0})$ é a log verossimilhança do modelo aplicada na estimativa de máxima verossimilhança sob H_0 .

Observe que valores de d_n próximos de 0 indicam que a hipótese sob H_0 é plausível e quando o valor de d_n é alto indica que H_0 não é plausível.

Segundo Peng, Dear and Carriere (2001), pode-se mostrar que, sob o modelo Weibull, a distribuição assintótica de d_n é uma mistura 50%-50% da distribuição qui-quadrado com 1 grau de liberdade e um ponto de massa em 0, de tal modo que

$$P(D \leq t) \rightarrow \frac{1}{2} + \frac{1}{2} \times P(\chi_1^2 \leq t) \quad , \quad n \rightarrow \infty , \quad (2.32)$$

em que χ_1^2 denota uma variável aleatória qui-quadrado com 1 grau de liberdade.

No entanto, note que a hipótese dada em (2.30) envolve testar na fronteira do espaço paramétrico $[0, 1]$ de π . Isso nos leva a um teste da Razão de Verossimilhanças cuja distribuição assintótica não é qui-quadrado.

A idéia da construção desse teste é que se $\pi = 1$, então $\hat{\theta}_n = \hat{\theta}_{H_0}$ e logo $d_n = 0$. Observe que se H_0 é de fato verdadeira, então, espera-se que assintoticamente $d_n = 0$ ocorrerá em aproximadamente 50% das amostras (Maller and Zhou, 1996). E para as outras 50%, o valor de d_n será maior que 0 e portanto a distribuição será qui-quadrado.

Logo, tem-se que a distribuição de d_n é uma mistura 50%-50% da distribuição qui-quadrado com 1 grau de liberdade e um ponto de massa em 0 como visto em (2.32).

A regra de decisão do teste é a seguinte: compara-se o valor da estatística d_n com o valor do percentil da distribuição de D correspondente ao nível de significância fixado, rejeitando a hipótese sob H_0 se d_n exceder esse valor.

2.6 Testando a suficiência do horizonte de tempo

Para a utilização dos modelos com fração de fidelizados deve-se verificar também se o horizonte de tempo no qual os indivíduos foram acompanhados foi suficiente para inferir sobre a presença de fidelizados na população em estudo.

Em Maller and Zhou (1996) é descrito um teste para verificar se o tempo

de observação considerado no estudo é suficiente para constatar a presença de indivíduos fidelizados. Se a hipótese dada em (2.30) não for rejeitada, pode-se concluir que não há evidências da presença de indivíduos fidelizados no estudo e ajustar um modelo usual de Análise de Sobrevida.

No entanto, sua rejeição pode gerar uma indecisão, pois deve-se verificar se o horizonte de tempo é suficiente pra observar a presença desses indivíduos, ou seja, deve-se testar as seguintes hipóteses

$$\begin{cases} H_0 : \text{O horizonte de tempo é suficiente.} \\ H_1 : \text{O horizonte de tempo não é suficiente.} \end{cases} \quad (2.33)$$

A estatística q_n utilizada para o teste é dada por

$$q_n = \frac{N_n}{n}, \quad (2.34)$$

em que:

N_n : número de observações não censuradas em $[2t_{(n)}^* - t_{(n)}, t_{(n)}^*]$ com:

$t_{(n)}$: maior tempo de falha ou de censura e

$t_{(n)}^*$ é o maior tempo de falha dentre os não censurados.

n : número de indivíduos na amostra.

Note que intuitivamente, se a distância $t_{(n)} - t_{(n)}^*$ for “grande”, é um indicativo que no final do período de acompanhamento há uma predominância de indivíduos censurados. E daí, vem a evidência da presença de indivíduos fidelizados na população em estudo.

A partir dessa idéia, Maller and Zhou (1996) criaram esse teste para verificar se o horizonte de tempo é suficiente para inferir a presença de fidelizados na população em estudo. Os autores realizaram simulações e construíram valores tabelados para a estatística q_n , dependendo dos valores de p (proporção de censuras), n (tamanho da amostra) e B (supõe-se que a distribuição dos tempos de falha é $U[0, B]$).

A regra de decisão do teste é a seguinte: se o valor da estatística q_n for inferior ao valor tabelado (Maller and Zhou, 1996) então, rejeita-se H_0 a um determinado nível de significância.

2.7 Medidas de avaliação dos modelos

A performance dos modelos apresentados neste capítulo será comparado à performance do modelo de Regressão Logística.

Existem várias medidas para a avaliação dos modelos cujo objetivo é discriminar entre duas populações. Para avaliação e comparação dos modelos ajustados nesta dissertação serão utilizadas duas das medidas mais populares: a estatística de Kolmogorov-Smirnov (Conover, 1999) e o coeficiente de Gini (ou curva de Lorenz) (Thomas, Edelman and Crook, 2002).

2.7.1 Estatística de Kolmogorov-Smirnov

A estatística de Kolmogorov-Smirnov (KS) é uma estatística não paramétrica que tem como objetivo testar se as funções de distribuição de dois grupos são iguais (Conover, 1999). Logo, essa estatística é utilizada para verificar se a distribuição dos escores é semelhante entre bons e maus clientes. Entende-se por escore, por exemplo, o valor da probabilidade predita de um determinado cliente ser bom.

Note que nos modelos em que se têm o objetivo de discriminar duas populações (bons e maus), espera-se que, por exemplo, os escores dos clientes bons estejam concentrados em valores altos enquanto para os maus clientes haja uma predominância em baixos escores. Usualmente, os escores têm uma alta amplitude de variação e para facilitar a interpretação e a construção de KS são criadas faixas de escores.

Seja S o total de faixas de escores e $F_B(s)$ e $F_M(s)$ as distribuições acumuladas

de bons e maus clientes. Tem-se que a estatística KS é dada por

$$KS = \max_s |F_B(s) - F_M(s)|,$$

em que $1 \leq s \leq S$.

Observe que KS pode variar entre 0 e 100% e quanto maior o seu valor, melhor o poder discriminatório do modelo ajustado. Vale ressaltar que os principais softwares estatísticos já possuem rotinas próprias para o cálculo de KS .

Gráfico 2.1: Exemplo de cálculo do valor da estatística KS .

O Gráfico 2.1 apresenta um exemplo de cálculo do valor da estatística KS . Note que o valor do KS para o modelo analisado é igual a 37%.

2.7.2 A Curva de Lorenz e o Coeficiente de Gini

A curva de Lorenz (1905) é uma representação gráfica muito utilizada para visualização de distribuição de renda e foi criada em 1905 pelo estatístico norte-americano M.C. Lorenz (Lorenz, 1905).

Basicamente, a curva de Lorenz consiste num gráfico no plano cartesiano, onde o eixo das abscissas (eixo x) representa o percentual acumulado do total da população, e no eixo das ordenadas (eixo y), o percentual acumulado da renda. Vale destacar que quando todos os indivíduos ganham a mesma parte do total, no caso de perfeita igualdade, o gráfico será representado pela reta de 45 graus.

Assim, “a curva pode ser interpretada da seguinte maneira: $x\%$ das pessoas mais pobres recebem $y\%$ de renda, ou, ao contrário, $y\%$ das rendas mais baixas correspondem a $x\%$ de pessoas. Essa curva serve, portanto, para medir o grau de desigualdade entre os limites opostos de renda” (Sandroni, 1994).

O coeficiente de Gini (1912) foi proposto por Conrado Gini e tem direta associação com a curva de Lorenz. Este coeficiente é representado por duas vezes a área entre a curva de Lorenz traçada e a reta de 45 graus que representa total igualdade de renda. O intervalo de variação do coeficiente está entre 0 e 1.

Note que quanto maior esse valor, maior a desigualdade entre pobres e ricos presente na população em estudo. No contexto de modelos que têm como objetivo discriminar duas populações (bons e maus clientes), a curva de Lorenz e o coeficiente de Gini são comumente utilizados para medir a performance dos modelos ajustados.

Observe que a construção da curva de Lorenz é realizada da seguinte maneira:

1. a partir dos valores de escore calculados para cada cliente, criam-se faixas de escores (de 10 a 20 faixas);
2. para cada faixa obtêm-se o número de clientes bons, maus e total. Além disso, o percentual acumulado de bons, maus e total também são calculados;
3. desenha-se o gráfico:
eixo x : % acumulado total ou % acumulado bons ou % acumulado maus
eixo y : % acumulado bons ou % acumulado maus.

Gráfico 2.2: Curva de Lorenz hipotética.

Pode-se interpretar a curva de Lorenz da seguinte maneira: analisando o Gráfico 2.2, tem-se que, com a utilização do modelo, considerando-se 40% do total da amostra nos escores mais baixos é possível identificar 73% dos maus clientes. Observe que caso não fosse utilizado o modelo (curva de referência) seria esperado que ao invés de 73%, apenas 40% dos maus clientes fossem identificados.

Portanto, conclui-se que quanto mais distante a curva de Lorenz para um determinado modelo estiver da curva de referência, melhor será a discriminação desse modelo.

Também a partir do gráfico, utilizando métodos numéricos, calcula-se o coeficiente de Gini, no qual se observa que quanto maior o seu valor, melhor o poder discriminatório do modelo.

Os principais pacotes estatísticos já fornecem o valor do coeficiente de Gini, no entanto, a construção da curva de Lorenz deve ser feita seguindo os passos descritos anteriormente.

Capítulo 3

Aplicação

Neste capítulo, é apresentada uma breve descrição do estudo e as técnicas apresentadas no Capítulo 2 são aplicadas aos dados. Além disso, um modelo de Regressão Logística usual foi aplicado aos mesmos dados de tal modo a comparar com os resultados das três técnicas descritas no trabalho: modelo de mistura Weibull, de Cox e de Cox ponderado.

Inicialmente, é apresentado como as variáveis preditoras foram preparadas antes dos ajustes dos modelos como também uma proposta de seleção de variáveis para reduzir o número de variáveis e considerar no modelo apenas as mais relevantes.

Em seguida, são apresentados os 3 modelos ajustados com estas variáveis (Modelo Weibull, de Cox e Cox ponderado, todos com fração de fidelizados). E por último, são apresentados os resultados dos modelos finais nos quais as variáveis explicativas foram selecionadas através do método forward realizado manualmente.

As análises foram realizadas nos pacotes SPSS for Windows v 13.0 (Regressão Logística) e R 2.1.0 for Windows no qual foram utilizados os pacotes *gfcure* e *semicure* (desenvolvidos pelo professor Peng - Universidade Memorial de Newfoundland). Para o ajuste do modelo de Cox ponderado foram necessárias algu-

mas modificações no programa *semicure* para a inclusão do ponderador.

3.1 Descrição do estudo

Num modelo no qual se deseja estimar a probabilidade de um determinado cliente cancelar o cartão de crédito, as variáveis históricas são essenciais, pois descrevem o comportamento dos clientes com relação a pagamentos e utilização ao longo do tempo.

Os dados históricos são construídos com base nas variáveis de relacionamento tais como: quantidade de compras nos últimos 6 meses, utilização média nos últimos 12 meses entre outras.

Um particular ponto no tempo é escolhido como o ponto de observação (H_0) e as variáveis históricas são construídas com informações dos clientes num período anterior a esse ponto, por exemplo, 6 ou 12 meses (Thomas, Edelman and Crook, 2002). Para a aplicação nesta dissertação, decidiu-se adotar um ponto no tempo e selecionar uma amostra aleatória de clientes com o cartão de crédito ativo.

Além disso, um outro período, por exemplo, 18 meses após o ponto de observação, é considerado para o acompanhamento da “performance” dos clientes com relação ao cancelamento do cartão de crédito.

Se a Regressão Logística for utilizada para a modelagem dos dados, ao final do período de performance, os clientes são classificados como “bons” ou “maus” dependendo se a conta está ativa ou não no final do período. Já no caso da utilização da Análise de Sobrevivência, os clientes são acompanhados ao longo do período de performance e o tempo até o cancelamento é marcado para os clientes que cancelam o cartão. Para os que não cancelam ou cancelam por outro motivo (por exemplo, não pagamento da conta do cartão), considera-se o tempo de censura. Na Regressão Logística, esses clientes são considerados como “bons”.

3. Aplicação

Neste trabalho foi utilizada uma base com 44.502 clientes observados em 12 diferentes pontos de observação, para evitar a influência da sazonalidade no modelo. Os pontos de observação compreendem os meses de junho de 2002 a maio de 2003.

Essa amostra foi dividida em duas bases: uma para o desenvolvimento do modelo (20.387 clientes) e outra para validação (24.115 clientes). O Apêndice B apresenta os meses que compõem o período histórico (H) e performance (P) para cada ponto de observação.

As amostras de validação são utilizadas para detectar se o modelo ajustado mantém seu poder de discriminação para outras amostras da mesma população. Se esse poder de discriminação mudar muito de uma amostra para a outra, pode significar que o modelo não apresenta estabilidade ou pode estar ocorrendo uma superestimação (“overfitting”).

A maioria das variáveis preditoras consideradas no estudo se referem ao comportamento do cliente na utilização do cartão de crédito. As variáveis comportamentais foram construídas a partir de informações do período histórico para os seguintes horizontes de tempo: 1, 3, 6, 9 e 12 meses.

Outras variáveis preditoras tais como sexo e idade foram observadas no mês que representa o ponto de observação.

A base de dados contém 171 variáveis preditoras, sendo que destas, 8 são variáveis cadastrais e 163 são variáveis comportamentais.

Nas Tabelas C.1 e C.2 do Apêndice C são apresentadas, respectivamente, as variáveis de utilização e as variáveis cadastrais.

Para o período de performance, a partir do qual se constrói a variável resposta, foi utilizado o seguinte conceito dependendo do modelo a ser ajustado:

- Modelo de Regressão Logística : observando os 18 meses de performance, classificou-se como **MAU** os clientes que cancelaram o cartão de crédito nesse período. Caso contrário, o cliente foi classificado como **BOM**.

- Modelo de Análise de Sobrevida: observando os 18 meses de performance, constroem-se duas variáveis que serão utilizadas para o ajuste do modelo, tempo até o cancelamento do cartão e indicador de falha. Por aspectos computacionais, para os clientes que cancelam o cartão no 18º mês, o tempo até o cancelamento (T) deve ser aproximado para $T = 17,99$.

3.2 Tratamento das variáveis preditoras

As variáveis contínuas presentes no estudo não foram categorizadas da forma como usualmente é realizado nos modelos utilizados no mercado como a técnica WOE (Weights of Evidence) introduzida por Good (1950) ou CHAID (Kass, 1980). Isso porque essas técnicas não levam em consideração o tempo até a ocorrência do evento em estudo, informação fundamental nos modelos de Análise de Sobrevida.

Usando algumas dessas técnicas, poderia haver um favorecimento ao modelo de regressão logística que será ajustado para a comparação com os modelos propostos neste trabalho.

Preferiu-se utilizar a categorização das variáveis de modo que fizesse sentido para a instituição, ou seja, as variáveis foram categorizadas como usualmente são utilizadas nas estratégias do negócio.

A partir das variáveis categorizadas foram criadas variáveis dicotômicas (“*dummies*”) para representar as categorias dessas variáveis (Hosmer and Lemeshow, 1989).

Por exemplo, para a variável sexo, com 3 categorias: masculino, feminino e sem informação, a criação de duas novas variáveis “*dummies*” (MASC e FEM) seria suficiente para representar a variável original: MASC=1, se o cliente é do sexo masculino e MASC=0, caso contrário; FEM=1, se o cliente é do sexo feminino e

FEM=0, caso contrário. Não existe a necessidade de criar outra “dummy” para a variável Sem Informação, mas essa categoria pode ser caracterizada da seguinte forma: MASC=0 e FEM=0. Denomina-se essa categoria como **categoria de referência**. Na Tabela 3.1 é apresentado um exemplo da construção das variáveis dicotômicas a partir de variáveis categorizadas.

Tabela 3.1: Exemplo de criação de variáveis “*dummies*” a partir de variáveis categorizadas.

Sexo	MASC	FEM
Masculino	1	0
Feminino	0	1
Sem Informação	0	0

As variáveis a partir das quais foram criadas inicialmente “*dummies*” foram: sexo (masculino, feminino e sem informação), estado civil (solteiro, casado, divorciado e outros), bandeira (mastercard, visa, redeshop e diners), produto (local, internacional, gold e platinum) e variante (emergente, local, internacional e premium).

3.3 Método de seleção de variáveis

A base de dados contém 171 variáveis que a partir da criação das variáveis “*dummies*” geraram 180 variáveis. Usualmente, quando se tem muitas variáveis utilizam-se algoritmos de seleção de variáveis (backward, forward, stepwise, entre outros) para identificar as variáveis que mais explicam a variável resposta.

Neste trabalho foram utilizados os procedimentos PROC LIFETEST do software SAS e o Binary Logistic Regression do SPSS 13.0 com o objetivo de reduzir o número de variáveis preditoras.

3. Aplicação

Como não existe na literatura nenhum tipo de algoritmo de seleção de variáveis para os modelos de Análise de Sobrevivência com fração de fidelizados, decidiu-se fazer a seleção da seguinte maneira:

1. utilizando o procedimento PROC LIFETEST com a opção *test* do software SAS, pode-se verificar a associação de cada variável com o tempo até o cancelamento do cartão de crédito através do output “Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test”. Adotando um nível de significância de 10%, selecionam-se as variáveis que apresentam um p -valor de incremento no qui-quadrado inferior ao valor adotado, de tal modo que esse incremento seja significativo;
2. a partir da informação do indicador de falha, ajusta-se um modelo de Regressão Logística usual, utilizando o método stepwise para a seleção de variáveis;
3. juntar as variáveis selecionadas nos passos anteriores.

A Tabela 3.2 apresenta as variáveis selecionadas pelos dois primeiros passos da metodologia proposta.

Tabela 3.2: Variáveis selecionadas pela metodologia proposta.

Número	Variável	Método
1	BLMAX9	Wilcoxon e Logístico
2	FEM	Wilcoxon
3	FREVOL3	Wilcoxon
4	IDADE	Wilcoxon
5	MOB	Wilcoxon e Logístico
6	NDECBL9	Wilcoxon
7	PARCPEN1	Wilcoxon
8	PURBAL9	Wilcoxon
9	QTDCCF9	Wilcoxon
10	QTMCC3	Wilcoxon
11	PREMIUM	Logístico

Para facilitar a interpretação dos coeficientes e evitar instabilidade no modelo final, decidiu-se categorizar as seguintes variáveis contínuas selecionadas: BLMAX9 e PURBAL9. Depois de categorizar essas variáveis foram construídas variáveis “*dummies*” a partir das mesmas como foi descrito anteriormente antes dos ajustes dos modelos.

3.4 Resultados

Considerando as variáveis apresentadas na Tabela 3.2 descrita na seção anterior, foram ajustados 3 modelos de mistura (ou modelos com fração de fidelizados): Weibull, de Cox e de Cox ponderado. Nesta seção, também serão apresentados os modelos finais (obtidos utilizando o método forward).

O programa *gfcure* foi utilizado para o ajuste do modelo de mistura paramétrico Weibull. Por aspectos computacionais, decidiu-se coletar uma amostra aleatória a partir da base de desenvolvimento original, isso porque o programa apresentou-se muito lento para amostras maiores. A amostra aleatória selecionada para o desenvolvimento foi uma sub-amostra aleatória de tamanho 1.500 da base inicial de desenvolvimento que possuía como visto anteriormente 20.387 clientes.

Para o ajuste do modelo de mistura semiparamétrico de Cox foi utilizado o programa *semicure*. Esse programa ajusta o modelo utilizando o algoritmo EM como visto no Capítulo 2.

O modelo de Cox ponderado com fração de fidelizados foi ajustado utilizando o programa *semicure*, com uma modificação para incorporar o ponderador.

A amostra para o ajuste dos 3 modelos foi a mesma, com um total de 1.500 clientes.

3.4.1 Testando a presença de clientes fidelizados

Antes de aplicar o modelo de Análise de Sobrevida deve-se testar a presença de clientes fidelizados na população, como descrito na Seção 2.3. Para a base de dados, considerando o percentil 95% da distribuição qui-quadrado, tem-se que

$$\frac{1}{2} + \frac{1}{2} \times P(\chi^2_1 \leq t_{0,95}) = 0,95. \quad (3.1)$$

De (3.1), tem-se que $t_{0,95}$ satisfaz $P(\chi^2_1 \leq t_{0,95}) = 0,90$. Da tabela qui-quadrado com 1 grau de liberdade, obtém-se $t_{0,95}=2,71$.

Logo, como

$$d_n = 2 \times (-734,78 + 1149,93) = 830,30$$

e $830,30 > 2,71$, rejeita-se H_0 a um nível de significância de 5%, ou seja, há forte evidência de que existam clientes fidelizados na base de dados.

3.4.2 Testando a suficiência do horizonte de tempo

Nesta seção, será realizado o teste para verificar se o horizonte de tempo considerado no estudo é suficiente para a identificação de indivíduos fidelizados.

Como descrito na Seção 2.4 do Capítulo 2, a estatística q_n é utilizada para realizar o teste (2.33).

Para a amostra de desenvolvimento dos modelos, com $n = 1.500$ clientes, tem-se que:

$N_n = 16$ que representa o número de indivíduos não censurados em $(16; 17]$ e $n = 1500$ que é o tamanho da amostra.

Portanto

$$q_n = \frac{N_n}{n} = \frac{16}{1500} = 0,011. \quad (3.2)$$

Da tabela apresentada em Maller and Zhou (1996) considerando $p = 0,90$ (note que o valor original de p na amostra é igual a 0,866), $B = 10$ e $n = 1.000$, e fixando um nível de significância de 10%, o valor crítico do teste é dado por 0,652. Observe que, os valores de B e n na amostra de desenvolvimento são, respectivamente, iguais a 18 e 1.500. No entanto, os valores máximos tabelados são $B = 10$ e $n = 1.000$.

Portanto, como o valor da estatística $q_n = 0,011$ é menor que 0,652, há evidências para rejeitar H_0 , ou seja, o horizonte de tempo não é suficiente.

Observe que o teste para verificar a presença de clientes fidelizados, nos leva a uma forte evidência da presença desse tipo de clientes na população em estudo. No entanto, o teste para verificar se o horizonte de tempo considerado foi suficiente para inferir dos mesmos, nos mostra que os 18 meses de acompanhamento não foi suficiente.

Os testes apresentados levam a resultados contraditórios, no entanto, analisando a base de dados é bastante viável assumir que existam clientes fidelizados na

população em estudo, uma vez que grande parte dos clientes já possui um grande tempo de relacionamento com o banco. Por isso, decidiu-se ajustar os modelos com fração de fidelizados aos dados do estudo.

3.4.3 Modelo de mistura Weibull

Nesta seção, o modelo de mistura Weibull foi ajustado com as 11 variáveis pré-selecionadas na Seção 3.3. As variáveis foram consideradas tanto na parte de Sobrevivência como na parte Logística do modelo de mistura.

As variáveis, as respectivas estimativas dos parâmetros, os riscos relativos (RR), os erros padrões, as estatísticas Z e os níveis descritivos (ou p -valores) encontram-se dispostos nas Tabelas 3.3 e 3.4. As categorias com estimativa igual a zero representam as categorias das variáveis que não são representadas por uma única variável “*dummy*” (categoria de referência). Os valores foram acrescentados para facilitar a visualização da variável por inteiro.

Tabela 3.3: Modelo de mistura Weibull - parte Sobrevivência

	Coefficiente	RR	Erro Padrão	Z	p-valor
Log(escala)	-0,503	-	0,074	-6,846	< 0,001
Intercepto	1,888	-	0,370	5,090	< 0,001
BLMAX9					
<= 0%	0,000	1,000	-	-	-
De 0% a 50%	0,706	0,311	0,282	2,502	0,012
Acima de 50%	0,510	0,430	0,289	1,761	0,078
FEM	-0,066	1,116	0,143	-0,466	0,642
FREVOL3	-0,273	1,572	0,175	-1,565	0,117
IDADE	-0,001	1,002	0,008	-0,143	0,886
MOB	0,000	1,001	0,001	0,226	0,821
NDECBL9	0,025	0,959	0,056	0,451	0,652
PARCPEN1	0,052	0,917	0,036	1,456	0,145
PURBAL9					
<= 0%	0,000	1,000	-	-	-
De 0% a 50%	-0,129	1,238	0,225	-0,574	0,566
Acima de 50%	0,558	0,397	0,286	1,948	0,051
QTDCCF9	-0,030	1,051	0,116	-0,261	0,794
QTMCC3	-0,050	1,086	0,116	-0,433	0,665
PREMIUM	0,306	0,603	0,338	0,907	0,364

Tabela 3.4: Modelo de mistura Weibull - parte Logística

	Coefficiente	Erro Padrão	Z	p-valor
Intercepto	0,519	0,519	0,999	0,318
BLMAX9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,131	0,420	-0,312	0,755
Acima de 50%	-0,489	0,380	-1,286	0,198
FEM	-0,322	0,207	-1,554	0,120
FREVOL3	-0,263	0,281	-0,936	0,349
IDADE	-0,010	0,010	-1,039	0,299
MOB	-0,002	0,002	-1,317	0,188
NDECBL9	0,081	0,076	1,069	0,285
PARCPEN1	0,019	0,055	0,345	0,730
PURBAL9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,923	0,345	-2,673	0,008
Acima de 50%	-0,638	0,418	-1,527	0,127
QTDCCF9	-0,318	0,189	-1,686	0,092
QTDMCC3	-0,199	0,137	-1,455	0,146
PREMIUM	-0,211	0,450	-0,470	0,639

Note que fixando-se um nível de significância de 5%, verifica-se que apenas as variáveis BLMAX9 da parte Logística (Tabela 3.3) e PURBAL9 da parte de Sobrevivência (Tabela 3.4) são significativas para o modelo ajustado. Como algumas variáveis desse modelo são correlacionadas, decidiu-se ajustar um modelo com menos variáveis, chegando a um modelo final. Na Seção 3.4.6 serão

apresentados os resultados do modelo final Weibull com fração de fidelizados e as interpretações dos parâmetros desse modelo.

3.4.4 Modelo de mistura de Cox

O modelo de mistura de Cox, utilizando a proposta de Peng (2003) foi ajustado com as 11 variáveis pré-selecionadas na Seção 3.3. As variáveis foram novamente consideradas tanto na parte de Sobrevivência como na parte Logística do modelo de mistura.

As variáveis, suas respectivas estimativas dos parâmetros, os erros padrões, as estatísticas Z e os níveis descritivos (ou p -valores) encontram-se dispostos nas Tabelas 3.5 e 3.6. Além disso, na Tabela 3.5 são apresentados os valores exponenciais das estimativas dos parâmetros (riscos relativos) para a parte do modelo de Cox.

Tabela 3.5: Modelo de mistura de Cox - parte Sobrevivência

	Coefficiente	RR	EP	Z	p-valor
BLMAX9					
<= 0%	0,000	1,000	-	-	-
De 0% a 50%	-0,425	0,653	0,252	-1,691	0,091
Acima de 50%	-0,369	0,692	0,252	-1,466	0,140
FEM	0,058	1,060	0,150	0,390	0,700
FREVOL3	0,191	1,211	0,163	1,173	0,240
IDADE	0,014	1,014	0,007	1,909	0,056
MOB	0,000	1,000	0,001	0,108	0,910
NDECBL9	-0,060	0,942	0,056	-1,076	0,280
PARCPEN1	-0,035	0,966	0,027	-1,268	0,200
PURBAL9					
<= 0%	0,000	1,000	-	-	-
De 0% a 50%	-0,024	0,976	0,253	-0,096	0,920
Acima de 50%	-0,292	0,747	0,252	-1,159	0,250
QTDCCF9	-0,177	0,838	0,130	-1,367	0,170
QTDGCC3	0,014	1,014	0,092	0,153	0,880
PREMIUM	0,024	1,024	0,266	0,090	0,930

Tabela 3.6: Modelo de mistura de Cox - parte Logística

	Coefficiente	Erro Padrão	Z	p-valor
Intercepto	0,590	0,376	1,570	0,116
BLMAX9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,613	0,263	-2,334	0,020
Acima de 50%	-0,809	0,256	-3,164	0,002
FEM	-0,277	0,156	-1,779	0,075
FREVOL3	-0,004	0,179	-0,022	0,982
IDADE	-0,011	0,007	-1,533	0,125
MOB	-0,002	0,001	-1,401	0,161
NDECBL9	0,064	0,055	1,153	0,249
PARCPEN1	-0,033	0,026	-1,253	0,210
PURBAL9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,852	0,280	-3,047	0,002
Acima de 50%	-1,009	0,269	-3,750	< 0,001
QTDCCF9	-0,261	0,142	-1,836	0,066
QTDGCC3	-0,152	0,079	-1,931	0,053
PREMIUM	-0,472	0,259	-1,822	0,069

Observe que diferentemente do modelo de mistura Weibull, fixando-se um nível de significância de 5%, a variável BLMAX9 da parte de Sobrevida (Tabela 3.5) não é significativa e da parte Logística (Tabela 3.6) é significativa.

3.4.5 Modelo de mistura de Cox ponderado

Nesta seção será ajustado o modelo de mistura de Cox ponderado, utilizando uma modificação realizada na proposta inicial de Peng (2003). As variáveis foram novamente consideradas tanto na parte de Sobrevivência como na parte Logística do modelo de mistura.

As variáveis, suas respectivas estimativas dos parâmetros, os erros padrões, as estatísticas Z e os níveis descritivos (ou p -valores) encontram-se dispostos nas Tabelas 3.7 e 3.8. Além disso, na Tabela 3.7 são apresentados os valores exponenciais das estimativas dos parâmetros (riscos relativos) para a parte do modelo de Cox.

3. Aplicação

Tabela 3.7: Modelo de mistura de Cox ponderado - parte Sobrevivência

	Coeficiente	RR	EP	EP Robusto	Z	p-valor
BLMAX9						
<= 0%	0,000	1,000	-	-	-	
De 0% a 50%	-0,454	0,635	0,079	0,231	-1,963	0,050
Acima de 50%	-0,353	0,702	0,080	0,238	-1,485	0,140
FEM	0,067	1,069	0,047	0,139	0,482	0,630
FREVOL3	0,213	1,237	0,051	0,153	1,393	0,160
IDADE	0,012	1,012	0,002	0,007	1,755	0,079
MOB	0,000	1,000	0,000	0,001	0,239	0,810
NDECBL9	-0,063	0,939	0,018	0,053	-1,193	0,230
PARCPEN1	-0,036	0,965	0,009	0,023	-1,539	0,120
PURBAL9						
<= 0%	0,000	1,000	-	-	-	
De 0% a 50%	-0,023	0,977	0,080	0,258	-0,088	0,930
Acima de 50%	-0,352	0,703	0,080	0,277	-1,269	0,200
QTDCCF9	-0,147	0,863	0,040	0,040	-3,661	< 0,001
QTDMMC3	0,004	1,004	0,029	0,088	0,043	0,970
PREMIUM	-0,032	0,968	0,085	0,216	-0,148	0,880

Vale destacar que com a inclusão do ponderador, deve-se considerar uma correção nas estimativas dos erros padrões dos parâmetros, que pode ser vista em Pugh, Robbins and Harrington (1992). Logo, para a análise do modelo de Cox ponderado, considera-se o erro padrão robusto para o qual essa correção já é realizada.

Tabela 3.8: Modelo de mistura de Cox ponderado - parte Logística

	Coefficiente	Erro Padrão	Z	p-valor
Intercepto	0,328	0,388	0,845	0,398
BLMAX9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,580	0,272	-2,133	0,033
Acima de 50%	-0,780	0,265	-2,947	0,003
FEM	-0,272	0,163	-1,670	0,095
FREVOL3	-0,009	0,187	-0,050	0,960
IDADE	-0,010	0,007	-1,387	0,166
MOB	-0,002	0,001	-1,274	0,203
NDECBL9	0,064	0,058	1,099	0,272
PARCPEN1	-0,032	0,028	-1,151	0,250
PURBAL9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,799	0,291	-2,748	0,006
Acima de 50%	-0,938	0,279	-3,362	0,001
QTDCCF9	-0,263	0,152	-1,724	0,085
QTD MCC3	-0,148	0,082	-1,804	0,071
PREMIUM	-0,451	0,270	-1,671	0,095

Fixando-se um nível de significância de 5%, tem-se que as variáveis BLMAX9 e QTDCCF9 na parte de Sobrevivência (Tabela 3.7) e BLMAX9 E PURBAL9 na parte Logística (Tabela 3.8) são significativas.

Observe que na parte de Sobrevivência dos modelos de mistura ajustados, os riscos relativos das variáveis produzem as mesmas conclusões com relação às categorias das mesmas, no entanto, para algumas variáveis observam-se valores bas-

tante distintos. Por exemplo, a variável PREMIUM na parte do modelo Weibull, mostra que o risco de cancelamento do cartão por um cliente classificado como PREMIUM é 0,603 vezes o risco de clientes com outras classificações. No entanto, no modelo de mistura de Cox, na parte de Sobrevivência, esse risco é igual 1,024 vezes.

Na parte Logística dos modelos, verifica-se que a variabilidade das estimativas dos parâmetros do modelo de mistura Weibull (Tabela 3.4) é maior que as dos modelos de mistura de Cox (Tabela 3.6) e Cox ponderado (Tabela 3.8).

Comparando os modelos de Cox e Cox ponderado ajustados pode-se verificar que os valores estimados dos coeficientes são bem próximos. Note também que o modelo de Cox ponderado produz erros padrões menores (erro padrão robusto).

3.4.6 Modelo final de mistura Weibull

Após os modelos ajustados anteriormente, foi utilizado um método de seleção forward (manual) no qual se testou a significância de cada variável para os modelos de misturas. A entrada das variáveis foi testada tanto na parte Logística como na parte de Sobrevivência.

Adotando-se um nível de significância de 5%, as variáveis foram testadas uma a uma e as que apresentaram p -valor superior a esse valor foram retiradas do modelo em cada passo.

Nas Tabelas 3.9 e 3.10 são apresentadas as estimativas dos parâmetros das variáveis selecionadas para o modelo final, respectivamente, da parte de Análise de Sobrevivência e da parte Logística. As mesmas, também contêm os riscos relativos, os erros padrões das estimativas, valor da estatística Z e o respectivos p -valores do teste Z .

Tabela 3.9: Modelo final de mistura Weibull - parte Sobrevivência

	Coeficiente	RR	Erro Padrão	Z	p-valor
Log(escala)	-0,431	-	0,073	-5,937	< 0,001
Intercepto	2,057	-	0,097	21,149	< 0,001
PURBAL9					
<= 50%	0,000	1,000	-	-	-
Acima de 50%	0,379	0,558	0,199	1,906	0,057
QTDGCC3	0,232	0,699	0,083	2,813	0,005

Tabela 3.10: Modelo final de mistura Weibull - parte Logística

	Coeficiente	Erro Padrão	Z	p-valor
Intercepto	-0,217	0,227	-0,958	0,338
BLMAX9				
<= 0%	0,000	1,000	-	-
De 0% a 50%	-1,047	0,266	-3,935	< 0,001
Acima de 50%	-1,319	0,258	-5,106	< 0,001
PREMIUM	-0,576	0,269	-2,144	0,032

A variável PURBAL9 (Tabela 3.9) foi recategorizada, juntando-se as duas primeiras categorias da variável (<= 0% e De 0% a 50%).

Pode-se observar através da Tabela 3.9 que quanto maior o valor assumido pela variável PURBAL9, que representa a utilização do cartão de crédito, maior o tempo até o cancelamento no cartão. Do mesmo modo, isso ocorre para a variável QTDGCC3 (quantidade de meses com compras nos últimos 3 meses) quando desconsidera-se a variação das demais variáveis.

3. Aplicação

Para facilitar a interpretação dos coeficientes do modelo de mistura Weibull, pode-se escrever o modelo na forma de riscos proporcionais de tal modo que com a parametrização adotada, o risco de uma determinada categoria em relação a categoria de referência é dado por $e^{-\frac{\beta}{\alpha}}$. Desse modo, o risco de cancelar o cartão para um cliente com PURBAL9 acima de 50% é 0,558 vezes o risco de um cliente com PURBAL9 abaixo de 50% (grupo de referência). Note que apesar dessa categoria da variável ser marginalmente não significativa, foi decidido deixá-la no modelo final. Da mesma forma, que o risco de cancelamento para um cliente com QTDMCC3 igual a 1 é 0,699 vezes o risco de um cliente que não possui compras nos últimos 3 meses. Logo, a cada incremento de uma unidade em QTDMCC3, tem-se que o risco de cancelamento é multiplicado por 0,699.

Analizando a Tabela 3.10 que apresenta os resultados da parte logística do modelo de mistura paramétrico, tem-se que:

- BLMAX9: quanto maior o valor da variável BLMAX9, ou seja, quanto mais o cliente utiliza o cartão, maior a probabilidade dele ser um cliente fidelizado;
- PREMIUM: clientes classificados como PREMIUM tendem a ser mais fiéis do que os outros tipos de clientes;

Os resultados da performance desse modelo podem ser vistos no Capítulo 4.

3.4.7 Modelo final de mistura de Cox

Nessa seção são apresentados os resultados do ajuste do modelo final de mistura de Cox.

Nas Tabelas 3.11 e 3.12 são apresentados, respectivamente, os resultados do programa *semicure* para a parte de Sobrevivência e para a parte Logística do modelo de mistura de Cox.

Tabela 3.11: Modelo final de mistura de Cox - parte Sobrevivência

	Coeficiente	RR	EP	Z	p-valor
BLMAX9					
<= 0%	0,000	1,000	-	-	-
Acima de 0%	-0,464	0,629	0,174	-2,660	0,008
PURBAL9					
<= 50%	0,000	1,000	-	-	-
Acima de 50%	-0,391	0,677	0,147	-2,660	0,008

Tabela 3.12: Modelo final de mistura de Cox - parte Logística

	Coeficiente	Erro Padrão	Z	p-valor
Intercepto	0,226	0,219	1,032	0,302
BLMAX9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,706	0,258	-2,738	0,006
Acima de 50%	-1,025	0,241	-4,262	< 0,001
MOB	-0,003	0,001	-2,117	0,034
PURBAL9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,957	0,244	-3,920	< 0,001
Acima de 50%	-1,190	0,224	-5,309	< 0,001
PREMIUM	-0,519	0,254	-2,048	0,041

Observe que as variáveis BLMAX9 e PURBAL9 foram recategorizadas apenas para a parte de Sobrevivência (Tabela 3.11) do modelo de Cox com fração de fidelizados.

Pode-se notar que através da parte de Sobrevivência do modelo ajustado, o risco de cancelamento para um cliente com valor da variável PURBAL9 acima de 50% é 0,677 vezes o risco de cancelamento para um cliente com PURBAL9 menor ou igual a 50%.

Do mesmo modo, analisando a variável BLMAX9 que pode ser interpretada como a utilização do cartão pelo cliente, observa-se que o risco de cancelar o cartão para um cliente com valor positivo ($BLMAX9 > 0\%$) é 0,629 vezes o risco de um cliente com valor da variável BLMAX9 menor ou igual a 0% (Tabela 3.11).

Observe também que da Tabela 3.12, tem-se que:

- BLMAX9 e PURBAL9: quanto maior os valores das variáveis BLMAX9 e PURBAL9, maior a probabilidade dele ser um cliente fidelizado;
- MOB: clientes mais antigos, apresentam uma maior probabilidade de serem fiéis;
- PREMIUM: clientes classificados como PREMIUM tendem a ser mais fiéis do que os outros tipos de clientes.

Os principais indicadores de performance desse modelo podem ser vistos no Capítulo 4.

3.4.8 Modelo final de mistura de Cox ponderado

Utilizando uma modificação no programa *semicure* original proposto por Peng (2003), para a inclusão do ponderador, foi obtido um modelo final de Cox com fração de fidelizados.

Nas Tabelas 3.13 e 3.14 são apresentados, respectivamente, os resultados do programa *semicure* para a parte de Sobrevivência e para a parte Logística do modelo de mistura de Cox ponderado.

3. Aplicação

Tabela 3.13: Modelo final de mistura de Cox ponderado - parte Sobrevivência

	Coefficiente	RR	EP	EP Robusto	Z	p-valor
BLMAX9						
<= 0%	0,000	1,000	-	-	-	
Acima de 0%	-0,458	0,632	0,0553	0,168	-2,72	0,007
PURBAL9						
<= 50%	0,000	1,000	-	-	-	
Acima de 50%	-0,453	0,636	0,0464	0,155	-2,93	0,003

Tabela 3.14: Modelo final de mistura de Cox ponderado - parte Logística

	Coefficiente	Erro Padrão	Z	p-valor
Intercepto	-0,008	0,222	-0,035	0,972
BLMAX9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,669	0,267	-2,507	0,012
Acima de 50%	-0,988	0,249	-3,967	< 0,001
PURBAL9				
<= 0%	0,000	-	-	-
De 0% a 50%	-0,908	0,254	-3,568	< 0,001
Acima de 50%	-1,118	0,233	-4,802	< 0,001
PREMIUM	-0,499	0,264	-1,890	0,059
MOB	-0,003	0,001	-1,933	0,053

Novamente as variáveis BLMAX9 e PURBAL9 foram recategorizadas apenas para a parte de Sobrevivência (Tabela 3.13) do modelo de Cox ponderado com fração de fidelizados.

Observa-se que através da parte de Sobrevivência do modelo ajustado (Tabela 3.13), o risco de cancelamento para um cliente com valor da variável PURBAL9 acima de 50% é 0,636 vezes o risco de cancelamento para um cliente com PURBAL9 menor ou igual a 50%.

Do mesmo modo, analisando a variável BLMAX9, pode-se observar que o risco de cancelar o cartão para um cliente com valor positivo dessa variável é 0,632 vezes o risco de um cliente com valor da variável BLMAX9 menor ou igual a 0%.

Note que as variáveis PREMIUM e MOB são marginalmente não significativas mas mesmo assim decidiu-se deixá-las no modelo final. As interpretações das variáveis da parte Logística do modelo de mistura de Cox ponderado (Tabela 3.14) são semelhantes as já descritas na seção anterior.

A performance desse modelo será apresentada no Capítulo 4.

3.4.9 Modelo de Regressão Logística

Foi gerado um Modelo de Regressão Logística considerando as 171 variáveis preditoras convertidas em variáveis “*dummies*” equivalentes da amostra de desenvolvimento com 1.500 casos. O modelo foi ajustado no Pacote Estatístico SPSS 13.0 utilizando o método “Forward:LR” para a seleção das variáveis. A probabilidade de entrada e de saída das variáveis foi fixada em 5%.

Vale ressaltar que o modelo foi ajustado considerando a variável resposta como sendo o valor do indicador de cancelamento, ou seja, os indivíduos que cancelaram o cartão até o 18º mês de acompanhamento foram considerados MAUS CLIENTES e receberam o valor 1, caso contrário, foram considerando BONS CLIENTES e a variável resposta assumiu valor 0.

Na Tabela 3.15 são apresentadas as variáveis selecionadas para o modelo final com as respectivas estimativas dos coeficientes, dos erros padrões, das estatísticas de Wald, os graus de liberdade e os níveis descritivos.

3. Aplicação

Tabela 3.15: Resultados do modelo final de Regressão Logística

	Coeficiente	Erro Padrão	Wald	g.l.	p-valor
Intercepto	-0,131	0,212	0,381	1	0,537
PREMIUM	-0,593	0,254	5,456	1	0,019
BLMAX9	-	-	24,179	2	-
<= 0%	0,000	-	-	-	-
De 0% a 50%	-0,810	0,260	9,708	1	0,002
Acima de 50%	-1,188	0,245	23,542	1	< 0,001
PURBAL9	-	-	15,920	2	-
<= 0%	0	-	-	-	-
De 0% a 50%	-0,796	0,259	9,474	1	0,002
Acima de 50%	-0,929	0,233	15,847	1	< 0,001

A partir da análise de cada variável sem considerar a variação das demais, pode-se destacar os seguintes pontos:

- **BLMAX9**: essa variável diz respeito a utilização do cartão, ou seja, quanto mais o cliente utiliza o cartão, menor a probabilidade dele cancelá-lo;
- **PURBAL9**: também representa uma variável de utilização, logo, quanto maior, menor a probabilidade do cliente cancelar o cartão;
- **PREMIUM**: novamente como esperado, clientes classificados como PREMIUM apresentam menor risco de cancelamento do cartão de crédito.

Na Tabela 3.16 são apresentadas as 11 variáveis selecionadas pelo método de seleção visto na Seção 3.2 e quais delas são significativas para cada um dos modelos finais ajustados. Note que por exemplo, as variáveis BLMAX9, PURBAL9 e PREMIUM estão presentes em todos os modelos ajustados.

Tabela 3.16: Variáveis de cada modelo final.

#	Variável	Descrição	Weibull	Cox	Cox ponderado	Logístico
1	BLMAX9	Último saldo sobre o saldo máximo nos últimos 9 meses.	L	SL	SL	L
2	FEM	Se o indivíduo é do sexo feminino.	-	-	-	-
3	FREVOL3	Financiou pelo menos 1 vez nos últimos 3 meses?	-	-	-	-
4	IDADE	Idade do associado.	-	-	-	-
5	MOB	Idade da conta.	-	L	L	-
6	NDECB9	Número de vezes que o saldo foi reduzido nos últimos 9 meses.	-	-	-	-
7	PARCPEN1	Quantidade de parceladas pendentes no último mês.	-	-	-	-
8	PURBAL9	Valor de compras sobre o saldo nos últimos 9 meses.	S	SL	SL	L
9	QTDCCF9	Quantidade de meses com Cadastro de Cheque sem Fundo (CCF) nos últimos 9 meses.	-	-	-	-
10	QTMCC3	Quantidade de meses com compras	S	-	-	-
11	PREMIUM	Indicador de cliente PREMIUM.	L	L	L	L

Legenda - L: Logístico, S: Sobrevivência, SL: ambos.

Capítulo 4

Comparação dos Modelos

A comparação das performances dos modelos finais ajustados tem como finalidade identificar qual das técnicas apresenta o melhor poder discriminatório, ou seja, melhor discrimina entre bons e maus clientes. Para tal propósito foram utilizados: a estatística de Kolmogorov-Smirnov (KS) e coeficiente de Gini, descritos na Seção 2.6.

As bases de desenvolvimento (1.500 clientes) e de validação (24.115 clientes) foram escoradas, ou seja, aplicou-se a fórmula obtida a cada um dos modelos ajustados nessas bases, obtendo-se o **escore** para cada cliente. Para cada um dos modelos ajustados foram calculados o valor de KS e o coeficiente de Gini. Para uma melhor visualização dos resultados, também foram construídas as curvas de Lorenz.

Vale ressaltar que os modelos de Análise de Sobrevida dependem do valor de t para o cálculo do valor de $S(t)$, que é a probabilidade de um cliente não cancelar o cartão até o tempo t , representando, nesse caso, escore. Portanto, para os modelos com fração de fidelizados Weibull, de Cox e de Cox ponderado, considerou-se 3 tempos de previsão diferentes para o cálculo de $S(t)$: 6, 12 e 18 meses.

No ajuste do modelo logístico considerou-se os maus clientes com o valor da

4. Comparação dos Modelos

variável resposta igual a 1, modelando-se assim a probabilidade do i -ésimo indivíduo cancelar o cartão (p_i). No entanto, o valor do escore a ser considerado deve ser a probabilidade do i -ésimo cliente não cancelar o cartão ($1 - p_i$).

Para a comparação dos modelos ajustados em mais de um horizonte de tempo de previsão, decidiu-se identificar os clientes com BONS ou MAUS em 3 diferentes tempos: 6, 12 e 18 meses.

Na Tabela 4.1 são apresentados os percentuais de indivíduos censurados, ou seja, que não cancelam o cartão de crédito para diferentes tempos de previsão.

Tabela 4.1: Percentual de censura para cada tempo de performance.

Tempo	Desenvolvimento $n = 1500$	Desenvolvimento	Validação
6	94,5%	95,1%	96,0%
12	90,2%	91,0%	92,6%
18	86,6%	87,3%	90,3%

Observa-se na amostra de Validação uma quantidade um pouco maior de clientes que cancelam o cartão quando comparado com as amostras de Desenvolvimento (Tabela 4.1).

Note que o modelo de Regressão Logística não possibilita essa flexibilidade e portanto, o valor do escore é considerado sempre o mesmo para os diferentes tempos de previsão.

Nas Tabelas 4.2 a 4.4 estão dispostos os valores da estatística KS para cada um dos tempos de previsão fixados.

4. Comparação dos Modelos

Tabela 4.2: Valores de KS para cada um dos modelos (6 meses).

	Desenvolvimento	Validação
Regressão Logística	30,2%	29,6%
Modelo de mistura Weibull	35,9%	32,2%
Modelo de mistura de Cox	37,5%	32,3%
Modelo de mistura de Cox ponderado	37,8%	32,3%

Comparando-se os valores de KS para os quatro modelos na Tabela 4.2, nota-se uma ligeira vantagem dos métodos de Análise de Sobrevivência com fração de fidelizados (ou de mistura) sobre o modelo de Regressão Logística, tanto na base de Desenvolvimento quanto na base de Validação.

Gráfico 4.1: Curva de Lorenz dos modelos ajustados (6 meses).

4. Comparação dos Modelos

Tabela 4.3: Valores de KS para cada um dos modelos (12 meses).

	Desenvolvimento	Validação
Regressão Logística	27,3%	24,3%
Modelo de mistura Weibull	31,9%	24,3%
Modelo de mistura de Cox	31,6%	24,6%
Modelo de mistura de Cox ponderado	32,2%	24,8%

Gráfico 4.2: Curva de Lorenz dos modelos ajustados (12 meses).

Tabela 4.4: Valores de KS para cada um dos modelos (18 meses).

	Desenvolvimento	Validação
Regressão Logística	21,6%	21,0%
Modelo de mistura Weibull	22,7%	20,7%
Modelo de mistura de Cox	26,0%	21,9%
Modelo de mistura de Cox ponderado	26,1%	21,9%

Gráfico 4.3: Curva de Lorenz dos modelos ajustados (18 meses).

Analisando as Tabelas 4.3 e 4.4, é possível identificar uma ligeira vantagem dos métodos de Análise de Sobrevivência com fração de fidelizados (ou de mistura) sobre o modelo de Regressão Logística, apenas para a base de Desenvolvimento. Além disso, é possível notar que a diferença entre a performance dos modelos ajustados vai se reduzindo a medida que o período de previsão vai aumentando.

Na Tabela 4.5 são apresentados os valores dos coeficientes de Gini para os modelos ajustados para os diferentes tempos de previsão.

Tabela 4.5: Coeficientes de Gini.

	6 meses	12 meses	18 meses
Regressão Logística	0,71	0,70	0,66
Modelo de mistura Weibull	0,73	0,70	0,65
Modelo de mistura de Cox	0,74	0,71	0,66
Modelo de mistura de Cox ponderado	0,74	0,71	0,66

4. Comparação dos Modelos

Calculando os coeficientes dos modelos ajustados (Tabela 4.5), observa-se que como já notado nas curvas de Lorenz (Gráficos 4.1 a 4.3), os resultados das performances do modelos são bem próximos.

Capítulo 5

Conclusão

Neste trabalho foram desenvolvidos modelos denominados *attrition score* para prever o cancelamento voluntário do cartão de crédito de clientes de uma grande instituição financeira brasileira. Os modelos foram ajustados por meio de quatro técnicas: Regressão Logística, Modelo Weibull, Modelo de Cox e Modelo de Cox ponderado, sendo que esses três últimos de Análise de Sobrevida foram ajustados considerando que existe uma fração de fidelizados na base de dados. O objetivo principal desta dissertação foi realizar uma comparação entre a performance desses modelos.

Na aplicação a dados reais, os modelos de Análise de Sobrevida com fração de fidelizados se mostraram muito competitivos em relação aos modelos de Regressão Logística que são comumente utilizados no mercado para discriminar os bons e maus clientes.

Uma vantagem dos modelos de Análise de Sobrevida é que com um só modelo, é possível ter uma visão do comportamento dos clientes ao longo do tempo e não apenas em um ponto, como acontece nos modelos de Regressão Logística. As previsões feitas com base nos modelos de Análise de Sobrevida são especialmente melhores quando feitas para curtos espaços de tempo.

Entre os modelos ajustados com fração de fidelizados, todos apresentaram

5. Conclusão

resultados muito próximos. A proposta de utilização do modelo de mistura de Cox ponderado feita neste trabalho, apresentou resultados muito próximos aos do proposto por Peng (2003). Vale destacar que a maioria dos erros padrões dos parâmetros na parte de Sobrevivência do modelo de mistura de Cox ponderado foram inferiores ao da parte de Sobrevivência do modelo de mistura de Cox.

Como sugestão para trabalhos futuros, a inclusão de variáveis dependentes do tempo pode ser considerada no modelo de mistura (paramétrico ou semi-paramétrico) bem como novas propostas para o teste para verificar se o horizonte de tempo foi suficiente para inferir a presença de indivíduos fidelizados na população em estudo. Além disso, toda a parte de resíduos dos modelos de Análise de Sobrevivência com fração de fidelizados ainda não está muito desenvolvida na literatura.

Apêndice A

Derivadas de 2ª ordem

De (2.16), tem-se que a matriz $\mathcal{F}_i(\boldsymbol{\theta})$ é uma matriz de dimensão (3×3) simétrica com componentes f_i^{jk} , j e $k = 1, \dots, 3$, em que

$$\begin{aligned} f_i^{11}(\boldsymbol{\theta}) = & \sum_{i=1}^n \left\{ -c_i \delta^2 (\lambda_i t_i)^\delta - (1 - c_i) \frac{p_i (\lambda_i t_i)^\delta \delta^2 e^{-(\lambda_i t_i)^\delta}}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \right. \\ & + (1 - c_i) \frac{p_i [(\lambda_i t_i)^\delta \delta]^2 e^{-(\lambda_i t_i)^\delta}}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\ & \left. - (1 - c_i) \frac{[p_i (\lambda_i t_i)^\delta \delta e^{-(\lambda_i t_i)^\delta}]^2}{[1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})]^2} \right\} \end{aligned}$$

$$\begin{aligned} f_i^{22}(\boldsymbol{\theta}) = & \sum_{i=1}^n \left\{ c_i (-p_i + p_i^2) \right. \\ & + (1 - c_i) \frac{(p_i (\lambda_i t_i)^\delta - 3p_i^2 (\lambda_i t_i)^\delta + 2p_i^3 (\lambda_i t_i)^\delta)}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\ & \left. - (1 - c_i) \frac{[-p_i (\lambda_i t_i)^\delta + p_i^2 (\lambda_i t_i)^\delta]^2}{[1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})]^2} \right\} \end{aligned}$$

$$\begin{aligned}
 f_i^{33}(\boldsymbol{\theta}) = & \sum_{i=1}^n \left\{ c_i \left(-\frac{1}{\delta^2} - [\ln(\lambda_i t_i)]^2 (\lambda_i t_i)^\delta \right) \right. \\
 & - (1 - c_i) \frac{p_i (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta} [\ln(\lambda_i t_i)]^2}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\
 & + (1 - c_i) \frac{p_i [(\lambda_i t_i)^\delta]^2 e^{-(\lambda_i t_i)^\delta} [\ln(\lambda_i t_i)]^2}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\
 & \left. - (1 - c_i) \frac{[p_i (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta} \ln(\lambda_i t_i)]^2}{[1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})]^2} \right\}
 \end{aligned}$$

$$\begin{aligned}
 f_i^{12}(\boldsymbol{\theta}) = & \sum_{i=1}^n \left\{ (1 - c_i) \left[\frac{(-p_i \delta (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta})}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \right. \right. \\
 & + \left. \frac{(p_i^2 \delta (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta})}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \right] \\
 & + (1 - c_i) \frac{-p_i (1 - e^{-(\lambda_i t_i)^\delta}) + p_i^2 (1 - e^{-(\lambda_i t_i)^\delta})}{[1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})]^2} \\
 & \times p_i (\lambda_i t_i)^\delta \delta e^{-(\lambda_i t_i)^\delta} \left. \right\}
 \end{aligned}$$

$$\begin{aligned}
 f_i^{13}(\boldsymbol{\theta}) = & \sum_{i=1}^n \left\{ c_i (1 - \delta (\lambda_i t_i)^\delta \ln(\lambda_i t_i)) \right. \\
 & - (1 - c_i) \frac{p_i (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta} \ln(\lambda_i t_i) \delta}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\
 & - (1 - c_i) \frac{p_i (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta}}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\
 & + (1 - c_i) \frac{p_i \ln(\lambda_i t_i) e^{-(\lambda_i t_i)^\delta} \delta [(\lambda_i t_i)^\delta]^2}{1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})} \\
 & \left. - (1 - c_i) \frac{\delta \ln(\lambda_i t_i) [p_i (\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta}]^2}{[1 - p_i (1 - e^{-(\lambda_i t_i)^\delta})]^2} \right\}
 \end{aligned}$$

$$\begin{aligned}
 f_i^{23}(\boldsymbol{\theta}) &= \sum_{i=1}^n \left\{ (1 - c_i) \left[\frac{p_i(\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta} \ln(\lambda_i t_i)}{1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})} \right. \right. \\
 &\quad \left. \left. - \frac{p_i^2(\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta} \ln(\lambda_i t_i)}{1 - p_i(1 - e^{-(\lambda_i t_i)^\delta})} \right] \right. \\
 &\quad \left. - (1 - c_i) \left[\frac{p_i(1 - e^{-(\lambda_i t_i)^\delta}) + p_i^2(1 - e^{-(\lambda_i t_i)^\delta})}{(1 - p_i(1 - e^{-(\lambda_i t_i)^\delta}))^2} \right. \right. \\
 &\quad \left. \left. \times p_i(\lambda_i t_i)^\delta e^{-(\lambda_i t_i)^\delta} \ln(\lambda_i t_i) \right] \right\}
 \end{aligned}$$

Apêndice B

Período Analisado

B. Período Analisado

[illegible]

Apêndice C

Variáveis Preditoras

Tabela C.1: Variáveis Cadastrais.

Num	Variável	Descrição
1	MOB	Idade da Conta (em anos completos).
2	IDADE	Idade do associado (em anos completos).
3	SEXO	Sexo do associado.
4	ESTCIV	Estado Civil do associado.
5	BANDEIRA	Bandeira (Mastercard, Visa ou Diners).
6	LIM	Limite de Crédito (em reais).
7	VARIANTE	Variante do Associado (Emergente, Local, Internacional ou Premium).
8	QTDADIC	Quantidade de cartões adicionais.

Tabela C.2: Variáveis Históricas.

Num	Variável	Descrição
1	BLMAX3	Último saldo sobre o saldo máximo nos últimos 3 meses.
2	BLMAX6	Último saldo sobre o saldo máximo nos últimos 6 meses.
3	BLMAX9	Último saldo sobre o saldo máximo nos últimos 9 meses.
4	BLMAX12	Último saldo sobre o saldo máximo nos últimos 12 meses.
5	frevol3	Financiou pelo menos 1 vez nos últimos 3 meses?
6	frevol6	Financiou pelo menos 1 vez nos últimos 6 meses?
7	frevol9	Financiou pelo menos 1 vez nos últimos 9 meses?
8	frevol12	Financiou pelo menos 1 vez nos últimos 12 meses?
9	NADATR1	Número médio de dias de atraso no último mês.
10	NADATR3	Número médio de dias de atraso nos últimos 3 meses.
11	NADATR6	Número médio de dias de atraso nos últimos 6 meses.
12	NADATR9	Número médio de dias de atraso nos últimos 9 meses.
13	NADATR12	Número médio de dias de atraso nos últimos 12 meses.
14	NDECBL3	Número de vezes em que o saldo foi reduzido nos últimos 3 meses.
15	NDECBL6	Número de vezes em que o saldo foi reduzido nos últimos 6 meses.
16	NDECBL9	Número de vezes em que o saldo foi reduzido nos últimos 9 meses.
17	NDECBL12	Número de vezes em que o saldo foi reduzido nos últimos 12 meses.
18	NINBL3	Número de vezes em que o saldo foi aumentado nos últimos 3 meses.
19	NINBL6	Número de vezes em que o saldo foi aumentado nos últimos 6 meses.

Tabela C.2: Variáveis Históricas (continuação).

Num	Variável	Descrição
20	NINBL9	Número de vezes em que o saldo foi aumentado nos últimos 9 meses.
21	NINBL12	Número de vezes em que o saldo foi aumentado nos últimos 12 meses.
22	NINCRE12	Número de vezes em que o saldo de financiamento foi aumentado nos últimos 12 meses.
23	NINCRE3	Número de vezes em que o saldo de financiamento foi aumentado nos últimos 3 meses.
24	NINCRE6	Número de vezes em que o saldo de financiamento foi aumentado nos últimos 6 meses.
25	NINCRE9	Número de vezes em que o saldo de financiamento foi aumentado nos últimos 9 meses.
26	NMCASH3	Número de meses que efetuou saque em dinheiro nos últimos 3 meses.
27	NMCASH6	Número de meses que efetuou saque em dinheiro nos últimos 6 meses.
28	NMCASH9	Número de meses que efetuou saque em dinheiro nos últimos 9 meses.
29	NMCASH12	Número de meses que efetuou saque em dinheiro nos últimos 12 meses.
30	NMDATR1	Número máximo de dias de atraso no último mês.
31	NMDATR3	Número máximo de dias de atraso nos últimos 3 meses.
32	NMDATR6	Número máximo de dias de atraso nos últimos 6 meses.

Tabela C.2: Variáveis Históricas (continuação).

33	NMDATR9	Número máximo de dias de atraso nos últimos 9 meses.
34	NMDATR12	Número máximo de dias de atraso nos últimos 12 meses.
35	NMFIN3	Número de meses com financiamento nos últimos 3 meses.
36	NMFIN6	Número de meses com financiamento nos últimos 6 meses.
37	NMFIN9	Número de meses com financiamento nos últimos 9 meses.
38	NMFIN12	Número de meses com financiamento nos últimos 12 meses.
39	NMPAY3	Número de meses com pagamento nos últimos 3 meses.
40	NMPAY6	Número de meses com pagamento nos últimos 6 meses.
41	NMPAY9	Número de meses com pagamento nos últimos 9 meses.
42	NMPAY12	Número de meses com pagamento nos últimos 12 meses.
43	NOVER3	Número de meses com utilização acima do limite de crédito nos últimos 3 meses.
44	NOVER6	Número de meses com utilização acima do limite de crédito nos últimos 6 meses.
45	NOVER9	Número de meses com utilização acima do limite de crédito nos últimos 9 meses.
46	NOVER12	Número de meses com utilização acima do limite de crédito nos últimos 12 meses.
47	NPARATIVSJ	Número de planos parcelados sem juros ativos no mês.

Tabela C.2: Variáveis Históricas (continuação).

48	NPARFIMSJ	Número de planos parcelados sem juros finalizados no mês.
49	NPARINISJ1	Número de planos parcelados sem juros iniciados no último mês.
50	NPARINISJ3	Número de planos parcelados sem juros iniciados nos últimos 3 meses.
51	NPARINISJ6	Número de planos parcelados sem juros iniciados nos últimos 6 meses.
52	NPARINISJ9	Número de planos parcelados sem juros iniciados nos últimos 9 meses.
53	NPARINISJ12	Número de planos parcelados sem juros iniciados nos últimos 12 meses.
54	NPURCARD1	Número médio de compras por cartão no último mês.
55	NPURCARD3	Número médio de compras por cartão nos últimos 3 meses.
56	NPURCARD6	Número médio de compras por cartão nos últimos 6 meses.
57	NPURCARD9	Número médio de compras por cartão nos últimos 9 meses.
58	NPURCARD12	Número médio de compras por cartão nos últimos 12 meses.
59	PARCPEN1	Quantidade de parcelas pendentes total no último mês.
60	PARCPEN3	Quantidade de parcelas pendentes total nos últimos 3 meses.
61	PARCPEN6	Quantidade de parcelas pendentes total nos últimos 6 meses.

Tabela C.2: Variáveis Históricas (continuação).

62	PARCPEN9	Quantidade de parcelas pendentes total nos últimos 9 meses.
63	PARCPEN12	Quantidade de parcelas pendentes total nos últimos 12 meses.
64	PARCTOTAL1	Quantidade de compras parceladas total no último mês.
65	PARCTOTAL3	Quantidade de compras parceladas total nos últimos 3 meses.
66	PARCTOTAL6	Quantidade de compras parceladas total nos últimos 6 meses.
67	PARCTOTAL9	Quantidade de compras parceladas total nos últimos 9 meses.
68	PARCTOTAL12	Quantidade de compras parceladas total nos últimos 12 meses.
69	PURBAL1	Valor das compras sobre o saldo no último mês.
70	PURBAL3	Valor das compras sobre o saldo nos últimos 3 meses.
71	PURBAL6	Valor das compras sobre o saldo nos últimos 6 meses.
72	PURBAL9	Valor das compras sobre o saldo nos últimos 9 meses.
73	PURBAL12	Valor das compras sobre o saldo nos últimos 12 meses.
74	QCASH1	Quantidade de saques em dinheiro no último mês.
75	QCASH3	Quantidade de Cash nos últimos 3 meses.
76	QCASH6	Quantidade de saques em dinheiro nos últimos 6 meses.
77	QCASH9	Quantidade de saques em dinheiro nos últimos 9 meses.
78	QCASH12	Quantidade de saques em dinheiro nos últimos 12 meses.
79	QTDCCF3	Quantidade de meses com CCF (Cadastro de Cheques sem Fundo) nos últimos 3 meses.

Tabela C.2: Variáveis Históricas (continuação).

80	QTDCCF6	Quantidade de meses com CCF (Cadastro de Cheques sem Fundo) nos últimos 6 meses.
81	QTDCCF9	Quantidade de meses com CCF (Cadastro de Cheques sem Fundo) nos últimos 9 meses.
82	QTDCCF12	Quantidade de meses com CCF (Cadastro de Cheques sem Fundo) nos últimos 12 meses.
83	QTDCOM1	Quantidade de compras no último mês.
84	QTDCOM3	Quantidade de compras nos últimos 3 meses.
85	QTDCOM6	Quantidade de compras nos últimos 6 meses.
86	QTDCOM9	Quantidade de compras nos últimos 9 meses.
87	QTDCOM12	Quantidade de compras nos últimos 12 meses.
88	QTDFAT3	Quantidade de meses com faturamento nos últimos 3 meses.
89	QTDFAT6	Quantidade de meses com faturamento nos últimos 6 meses.
90	QTDFAT9	Quantidade de meses com faturamento nos últimos 9 meses.
91	QTDFAT12	Quantidade de meses com faturamento nos últimos 12 meses.
92	QTDMAF3	Quantidade de meses com acordo em fatura nos últimos 3 meses.
93	QTDMAF6	Quantidade de meses com acordo em fatura nos últimos 6 meses.
94	QTDMAF9	Quantidade de meses com acordo em fatura nos últimos 9 meses.

Tabela C.2: Variáveis Históricas (continuação).

95	QTDMAF12	Quantidade de meses com acordo em fatura nos últimos 12 meses.
96	QTMCCC3	Quantidade de meses com compras (últimos 3 meses).
97	QTMCCC6	Quantidade de meses com compras (últimos 6 meses).
98	QTMCCC9	Quantidade de meses com compras (últimos 9 meses).
99	QTMCCC12	Quantidade de meses com compras (últimos 12 meses).
100	QTDPGT1	Quantidade de Pagamentos no último mês.
101	QTDPGT3	Quantidade de Pagamentos nos últimos 3 meses.
102	QTDPGT6	Quantidade de Pagamentos nos últimos 6 meses.
103	QTDPGT9	Quantidade de Pagamentos nos últimos 9 meses.
104	QTDPGT12	Quantidade de Pagamentos nos últimos 12 meses.
105	SALACU1	Saldo acumulado no último mês.
106	SALACU3	Saldo acumulado nos últimos 3 meses.
107	SALACU6	Saldo acumulado nos últimos 6 meses.
108	SALACU9	Saldo acumulado nos últimos 9 meses.
109	SALACU12	Saldo acumulado nos últimos 12 meses.
110	SDPAR1	Valor do parcelado acumulado no último mês.
111	SDPAR3	Valor do parcelado acumulado últimos 3 meses.
112	SDPAR6	Valor do parcelado acumulado últimos 6 meses.
113	SDPAR9	Valor do parcelado acumulado últimos 9 meses.
114	SDPAR12	Valor do parcelado acumulado últimos 12 meses.
115	TPCJ1	Total de parcelado com juros no último mês.
116	TPCJ3	Total de parcelado com juros nos últimos 3 meses.
117	TPCJ6	Total de parcelado com juros nos últimos 6 meses.
118	TPCJ9	Total de parcelado com juros nos últimos 9 meses.
119	TPCJ12	Total de parcelado com juros nos últimos 12 meses.

Tabela C.2: Variáveis Históricas (continuação).

120	UTMAX3	Saldo máximo sobre o limite nos últimos 3 meses.
121	UTMAX6	Saldo máximo sobre o limite nos últimos 6 meses.
122	UTMAX9	Saldo máximo sobre o limite nos últimos 9 meses.
123	UTMAX12	Saldo máximo sobre o limite nos últimos 12 meses.
124	UTMEAN1	Utilização média no último mês.
125	UTMEAN3	Utilização média últimos 3 meses.
126	UTMEAN6	Utilização média últimos 6 meses.
127	UTMEAN9	Utilização média últimos 9 meses.
128	UTMEAN12	Utilização média últimos 12 meses.
129	VLCASH1	Soma do valor de saques em dinheiro no último mês.
130	VLCASH3	Soma do valor de saques em dinheiro nos últimos 3 meses.
131	VLCASH6	Soma do valor de saques em dinheiro nos últimos 6 meses.
132	VLCASH9	Soma do valor de saques em dinheiro nos últimos 9 meses.
133	VLCASH12	Soma do valor de saques em dinheiro nos últimos 12 meses.
134	VLCOMP1	Soma do valor das compras no último mês.
135	VLCOMP3	Soma do valor das compras nos últimos 3 meses.
136	VLCOMP6	Soma do valor das compras nos últimos 6 meses.
137	VLCOMP9	Soma do valor das compras nos últimos 9 meses.
138	VLCOMP12	Soma do valor das compras nos últimos 12 meses.
139	VLJUCO1	Soma dos juros cobrados no último mês.
140	VLJUCO3	Soma dos juros cobrados nos últimos 3 meses.
141	VLJUCO6	Soma dos juros cobrados nos últimos 6 meses.

Tabela C.2: Variáveis Históricas (continuação).

142	VLJUCO9	Soma dos juros cobrados nos últimos 9 meses.
143	VLJUCO12	Soma dos juros cobrados nos últimos 12 meses.
144	VLJUFIN1	Soma dos juros de financiamento no último mês.
145	VLJUFIN3	Soma dos juros de financiamento nos últimos 3 meses.
146	VLJUFIN6	Soma dos juros de financiamento nos últimos 6 meses.
147	VLJUFIN9	Soma dos juros de financiamento nos últimos 9 meses.
148	VLJUFIN12	Soma dos juros de financiamento nos últimos 12 meses.
149	VLPAG1	Soma do valor dos pagamentos no último mês.
150	VLPAG3	Soma do valor dos pagamentos nos últimos 3 meses.
151	VLPAG6	Soma do valor dos pagamentos nos últimos 6 meses.
152	VLPAG9	Soma do valor dos pagamentos nos últimos 9 meses.
153	VLPAG12	Soma do valor dos pagamentos nos últimos 12 meses.
154	VLSAFIN1	Valor acumulado saldo financiado no último mês.
155	VLSAFIN3	Valor acumulado saldo financiado nos últimos 3 meses.
156	VLSAFIN6	Valor acumulado saldo financiado nos últimos 6 meses.
157	VLSAFIN9	Valor acumulado saldo financiado nos últimos 9 meses.
158	VLSAFIN12	Valor acumulado saldo financiado nos últimos 12 meses.
159	VPARCPT1	Montante de parcelas pendentes total no último mês.
160	VPARCPT3	Montante de parcelas pendentes total nos últimos 3 meses.
161	VPARCPT6	Montante de parcelas pendentes total nos últimos 6 meses.
162	VPARCPT9	Montante de parcelas pendentes total nos últimos 9 meses.
163	VPARCPT12	Montante de parcelas pendentes total nos últimos 12 meses.

Referências Bibliográficas

- Abreu, H. (2005), *Aplicação da análise de sobrevivência em um problema de credit scoring e comparação com a regressão logística*, Dissertação de Mestrado. Universidade Federal de São Carlos.
- Berkson, J. and Gage, R. (1952), 'Survival cure for cancer patients following treatment', *Journal of the American Statistical Association* **47**, 501–515.
- Binder, D. (1992), 'Fitting Cox's proportional hazards models from survey data', *Biometrika* **79**, 139–147.
- Boag, J. (1949), 'Maximum likelihood estimates of the proportion of patients cured by cancer therapy', *Journal of the Royal Statistical Society, Ser. B* **11**, 15–44.
- Breslow, N. (1972), 'Contribution to the discussion of D.R. Cox', *Journal of the Royal Statistical Society, Ser. B* **34**, 216–217.
- Conover, W. (1999), *Practical nonparametric statistics*, 3 edn, New York: John Wiley and Sons.
- Cox, D. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society, Ser. B* **39**, 1–38.
- Cox, D. (1975), 'Partial likelihood', *Biometrika* **62**, 269–276.

REFERÊNCIAS BIBLIOGRÁFICAS

- Efron, B. (1977), 'The efficiency of Cox's likelihood function for censored data', *Journal of the American Statistical Association* **72**, 557–565.
- Farewell, V. (1977), 'A model for a binary variable with time censored observations', *Biometrika* **64**, 43–46.
- Farewell, V. (1982), 'The use of mixture models for the analysis of survival data with long-term survivors', *Biometrics* **38**, 1041–1046.
- Farewell, V. (1986), 'The use of mixture models in survival analysis: Are they worth the risk?', *Canadian Journal of Statistics* **14**, 257–262.
- Ghitany, M. and Maller, R. (1995), 'Asymptotic results for exponential mixture models with long-term survivors and covariates', *Journal of Multivariate Analysis* **49**, 218–241.
- Gini, C. (1912), Variabilità e mutabilità, Technical report, Rome: Libreria Eredi Virgilio Veschi.
- Good, I. (1950), Probability and the weighting of evidence, Master's thesis, Charles Griffin, London.
- Hosmer, J. D. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley and Sons.
- Kalbfleisch, J. and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.
- Kass, G. (1980), 'An explanatory technique for investigating large quantiles of categorical data', *Applied Statistics* **29**, 119–127.
- Kuk, A. and Chen, C. (1992), 'A mixture model combining logistic regression with proportional hazards regression', *Biometrika* **79**, 531–541.

REFERÊNCIAS BIBLIOGRÁFICAS

- Lorenz, M. O. (1905), 'Methods for measuring the concentration of wealth', *Journal of the American Statistical Association* **1**, 209–219.
- Maller, R. and Zhou, X. (1996), *Survival Analysis with Long-Term Survivors*, London: Wiley Series in Probability and Statistics.
- McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, 2 edn, London: Chapman and Hall.
- Oliveira, W. (2000), *CRM & e-business*, Florianópolis: Visual Books, p. 154.
- Peng, Y. (2003), 'Fitting semiparametric cure models', *Computational Statistics and Data Analysis* **41**, 481–490.
- Peng, Y. and Dear, K. (2000), 'A nonparametric mixture model for cure rate estimation', *Biometrics* **56**, 237–243.
- Peng, Y., Dear, K. and Carriere, K. (2001), 'Testing for the presence of cured patients: a simulation study', *Statistics in Medicine* **20**, 1783–1796.
- Peng, Y., Dear, K. and Denham, J. W. (1998), 'A generalized F mixture model for cure rate estimation', *Statistics in Medicine* **17**, 813–830.
- Pocock, S., Gore, S. and Kerr, G. (1982), 'Long-term survival analysis: the curability of breast cancer', *Statistics in Medicine* **1**, 93–104.
- Pugh, M., J. Robbins, S. L. and Harrington, D. (1992), Inference in the Cox proportional hazards model with missing covariates, Technical Report 758Z, Department of Biostatistics, Harvard School of Public Health, Boston.
- Sandroni, P. (1994), *Novo Dicionário de Economia*, 5 edn, São Paulo: Editora Best Seller.

REFERÊNCIAS BIBLIOGRÁFICAS

Therneau, T. and Grambsch, P. (2000), *Modeling Survival Data*, New York: Springer.

Thomas, L., Edelman, D. and Crook, N. (2002), *Credit Scoring and Its Applications*, Philadelphia: Siam.

Venables, W. and Ripley, B. (1997), *Modern Applied with S-Plus*, 2 edn, New York: Springer.

Yamaguchi, K. (1992), 'Accelerated failure-time regression models with a regression model of survival fraction: An application to the analysis of permanent employment in Japan', *Journal of the American Statistical Association* **87**, 284–292.