UNIVERSIDADE FEDERAL DO PARANÁ

ALINE HANREJSZKOW
EVANDRO STROMBERG

# APLICAÇÃO DE REGRESSÃO LOGÍSTICA E MODELO DE MISTURA EM UM ESTUDO SOBRE CLIENTES INADIMPLENTES DE UMA EMPRESA DE TELECOMUNICAÇÕES

CURITIBA
2013

ALINE HANREJSZKOW
EVANDRO STROMBERG

# APLICAÇÃO DE REGRESSÃO LOGÍSTICA E MODELO DE MISTURA EM UM ESTUDO SOBRE CLIENTES INADIMPLENTES DE UMA EMPRESA DE TELECOMUNICAÇÕES

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório de Estatística, Curso de Estatística, Setor de Ciências Exatas da Universidade Federal do Paraná.

Orientadora: Profª Drª Suely Ruiz Giolo

CURITIBA
2013

## RESUMO

Diversas técnicas estatísticas são aplicadas em dados financeiros a fim de incrementar a rentabilidade da empresa e reduzir custos e perdas financeiras. Nesse trabalho, foram analisados dados de uma empresa de telecomunicações com informações sobre o perfil de 4.535 clientes inadimplentes. Para identificar os fatores que estão associados ao atraso máximo que os clientes atingem no período observado, foram utilizadas duas diferentes técnicas, existindo um particular interesse em compará-las. O primeiro modelo foi construído utilizando regressão logística. Esse modelo tem como objetivo estimar a probabilidade de pagamento. Visando considerar além da ocorrência do pagamento o tempo até a sua ocorrência foi utilizado o modelo de mistura com fração de fidelizados. De modo geral, os dois modelos mostraram um ajuste satisfatório aos dados, sugerindo que ambas as técnicas podem ser utilizadas em bancos de dados reais, sendo que o modelo de mistura com fração de fidelizados mostrou-se mais informativo por levar em conta o tempo até a ocorrência do pagamento.



**Palavras-chave:** Análise de sobrevivência; Dados financeiros; Fração de fidelizados; Modelo de mistura; Regressão logística.

**SUMÁRIO**

**1 INTRODUÇÃO** ..... 1

**2 REVISÃO DE LITERATURA** ..... 3

**2.1 Inadimplência** ..... 3

**2.2 Cobrança** ..... 4

**2.3 Terceirização** ..... 5

**3 MATERIAL E MÉTODOS** ..... 7

**3.1 Material** ..... 7

**3.2 Métodos** ..... 9

3.2.1 Modelo de Regressão Logística ..... 9

3.2.2 Seleção de covariáveis e do modelo final ..... 10

3.2.3 Análise de Sobrevivência ..... 12

3.2.3.1 Estimador não-paramétrico de Kaplan-Meier ..... 13

3.2.3.2 Modelo de Mistura de Cox com Fração de Fidelizados ..... 13

**4 RESULTADOS** ..... 17

**4.1 Análise Descritiva** ..... 17

**4.2 Ajuste do Modelo de Regressão Logística** ..... 18

**4.3 Ajuste do Modelo de Mistura com Fração de Fidelizados ..... 22**

**4.4 Comparação entre os Resultados dos Modelos ..... 23**

**5 CONCLUSÕES ..... 32**

**REFERÊNCIAS ..... 34**

**APÊNDICES ..... 36**

# 1 INTRODUÇÃO

O ciclo de crédito e cobrança está presente em diversos setores da economia, dentre os quais podem ser citados: bancos, varejo, empresas de telecomunicações, entre outros. Este ciclo pode ser dividido em 4 principais etapas. A primeira é a construção do produto, em que são utilizadas pesquisas de mercado e o conhecimento adquirido pelo responsável pelo produto; também é avaliada a rentabilidade do produto, tal como o risco associado a sua comercialização.

A segunda etapa é a venda do produto à crédito feita avaliando o risco de inadimplência associada ao perfil do cliente. Nesta etapa, são utilizados modelos estatísticos denominados *Credit Score* para estimar a probabilidade de não pagamento. Como exemplos, podem ser citados: empréstimo para financiamento de um imóvel ou veículo, concessão de limite de cartão de crédito, vendas a prazo em lojas de varejo e serviços de telecomunicações (internet, telefone e TV a cabo).

Na terceira etapa, é realizada a gestão das contas; neste momento já se tem mais informações comportamentais do cliente. Essas informações são agregadas aos modelos estatísticos denominados *Behaviour Score* fazendo com que se tenha um maior poder de discriminação e maior índice de acertos.

A cobrança é a quarta etapa do ciclo. O foco dos modelos de *Collection Score* é maximizar a receita e minimizar as perdas financeiras, otimizando o custo com cobrança. Essa metodologia direciona qual estratégia deve ser utilizada para o perfil do cliente para que o esforço de cobrança seja empregado quando necessário e no canal de cobrança adequado (e-mail, carta, SMS etc.).

Existem diversas consultorias especializadas apenas na parte de cobrança. Estas possuem uma *expertise* que as empresas convencionais não possuem e, normalmente, não implementam devido ao alto custo de manutenção de uma área com esse foco; por isso é comum a negociação de carteira de dívidas de clientes inadimplentes com essas empresas. Contratar esse serviço é uma alternativa para reduzirem seus custos com cobrança.

Para gerenciamento de carteira de clientes inadimplentes, as empresas que trabalham com concessão de crédito usualmente utilizam modelos de regressão logística para estimar a probabilidade de pagamento.

Com o objetivo de analisar os dados de um grupo de clientes inadimplentes de uma empresa de telecomunicações que passaram a ser monitorados quanto ao

atraso de seus pagamentos por um período de 150 dias, foram utilizados neste trabalho dois modelos estatísticos. Um deles é o modelo de regressão logística, comumente utilizado para estimar a probabilidade de pagamento do cliente. O outro é o modelo de mistura semiparamétrico com fração de cura (ou fração de fidelizados, no contexto dos dados analisados), que permite estimar tanto a probabilidade de pagamento, quanto o tempo até a ocorrência do mesmo.

A comparação dos dois modelos foi realizada em termos dos resultados obtidos para cada um deles a fim de discutir vantagens do modelo de mistura em relação ao de regressão logística, visto o primeiro ser sugerido como mais informativo por considerar simultâneamente nas análises a ocorrência do pagamento (ou o não pagamento) e o tempo até o mesmo.

## 2 REVISÃO DE LITERATURA

Este capítulo faz uma breve descrição sobre inadimplência e as estratégias de cobrança utilizadas nas empresas que trabalham com crédito ou prestação de serviços.

A inadimplência pode ser definida como a incapacidade de uma pessoa física ou empresa quitar suas dívidas no valor, especificidade e data do vencimento. Já a cobrança é um processo para recuperação do crédito que foi tomado. Ocorre quando uma venda é realizada a prazo e o recebimento não ocorre dentro do prazo estabelecido ou tolerável.

### 2.1 Inadimplência

No Brasil, ocorreu uma queda na taxa de juros nos últimos anos, apesar disso, a inadimplência média das operações bancárias com empresas e pessoas físicas no fim de 2012 foi de 5,9%, o que representa o maior valor já registrado até essa data, desde que o indicador teve início em junho de 2000. Este indicador considera inadimplentes as pessoas físicas e jurídicas com atraso superior a 90 dias (IPEADATA).

A inadimplência é um fenômeno que prejudica tanto credores quanto tomadores. Quando uma instituição financeira, por exemplo, não recebe o capital emprestado, este valor é pago pelos outros tomadores através de taxas de juros maiores. Para o caso das empresas de telecomunicações, onde os produtos são pagos mensalmente após o serviço já ter sido fornecido, os valores não recebidos pelas empresas estão incorporados no valor final do produto.

A fim de controlar os índices de inadimplência é de suma importância acompanhar as operações de crédito concedidas evitando, desta forma, custos com cobrança e perdas.

> Há muitas vantagens na utilização de métodos quantitativos em gerenciamento de crédito, destacando-se os benefícios resultantes da otimização no processo de tomada de decisão: fornece-se crédito (ou crédito adicional) aos melhores clientes (mais confiáveis), gerando aumento nos lucros e nega-se (ou diminui-se) o crédito aos piores clientes (menos confiáveis), resultando na diminuição das perdas. Além disso, políticas ótimas de cobrança minimizam os custos de administração e maximizam o montante recuperado do mal pagador (ROSENBERG, GLEIT, 1994, p. 590).

Para as empresas que possuem grande capital alocado em operações de crédito, a inadimplência é inevitável. Passado o prazo para a quitação dos débitos sob a forma acordada, é necessário tomar ações para sua recuperação. Nesse sentido, um ponto importante ao tomar as medidas necessárias para rever o crédito é se preocupar com a fidelização do cliente, efetuando a cobrança da maneira efetiva e adequada.

## 2.2 Cobrança

A cobrança tem assumido um papel cada vez mais importante no ciclo financeiro de uma empresa. Segundo Silva (2006), a gestão de cobrança deve estar focada na maximização, visando melhorar o fluxo de caixa e na minimização de perdas de negócios futuros.

As políticas e perspectivas das empresas devem estar alinhadas com as estratégias de crédito e cobrança. De acordo com Silva (2006), as políticas da empresa devem considerar a perspectiva de risco de crédito pelo esforço de cobrança, categorizando ambos em alto e baixo, desse modo, tem-se as seguintes estratégias:

a. Alto risco de crédito e baixo esforço de cobrança.
b. Alto risco de crédito e alto esforço de cobrança.
c. Baixo risco de crédito e baixo esforço de cobrança.
d. Baixo risco de crédito e alto esforço de cobrança.

Para uma empresa que busque um rápido crescimento de sua participação no mercado, a primeira estratégia é uma boa opção, porém ela deve ser estudada cautelozamente antes da implantação, pois ela leva a liberação de crédito à maioria dos proponentes, o que pode contaminar a carteira e trazer grandes perdas. Em contrapartida, esta estratégia economiza recursos com análise de crédito e custos de cobrança. Usualmente, empresas que aderem a esta estratégia agregam uma taxa de juros maior ao produto.

Na segunda estratégia, o baixo custo com análise de crédito é compensado pelo alto custo com cobrança. Esta estratégia possui uma recuperação muito maior

do que a primeira, porém ela pode causar problemas de fidelização de bons clientes que estão inadimplentes apenas momentaneamente, devido a fatores externos, mas que logo quitariam suas dívidas.

A terceira é muito utilizada por empresas em que o foco é o relacionamento com bons clientes. Nesta estratégia, há uma restrição de mercado, porém tem-se uma compensação com a qualidade dos clientes. Como o esforço em cobrança é baixo, alguns clientes podem demorar um pouco mais do que o acordado para honrar suas dívidas, o que requer da empresa um maior volume de capital de giro.

A última estratégia tem maior restrição de mercado. As empresas que a utilizam são aquelas que trabalham com baixa margem de lucro onde as perdas devem ser mínimas para que a empresa se mantenha sustentável. Outros tipos de produto que utilizam esta estratégia são aqueles de alto valor, onde poucos inadimplentes podem trazer prejuízos imensos.

As políticas definem ações sequenciais para a área de cobrança como, por exemplo:

> a) Telefonema de lembrança no segundo dia após o vencimento.
> b) Carta ou e-mail no quinto dia.
> c) Carta ou e-mail com texto mais enérgico no décimo dia (informando que o título está sendo enviado para os advogados da empresa para medidas judiciais cabíveis e/ou que o avalista ou garantidor será acionado).
> d) Acionamento do avalista ou garantidor.
> e) Envio ao Cartório de Protesto de Títulos e comunicação da inadimplência às Agências de Crédito.
> f) Execução da dívida através do encaminhamento do título aos advogados da empresa (LEMES JUNIOR *et al*.,2005, p. 456).

Outras políticas aplicadas para a cobrança de clientes inadimplentes também podem ser utilizadas como a aplicação de juros e multas.

### 2.3 Terceirização

A terceirização dos serviços de cobrança consiste no repasse da carteira de inadimplentes para uma outra empresa especializada no assunto. Esta carteira é composta por clientes que possuem atraso superior a um período pré-estabelecido. A definição deste período é de suma importância e deve ser realizada levando-se em conta os registros históricos de toda a carteira, pois existem aqueles clientes que irão quitar suas dívidas sem mesmo precisarem ser cobrados, ou aqueles que

apenas se esqueceram e necessitam apenas de uma ligação ou uma mensagem de texto para lembrá-los; desta forma a negociação das dívidas destes clientes não é interessante.

> A terceirização é uma técnica moderna de administração e que se baseia num processo de gestão, que leva a mudanças estruturais da empresa, a mudanças de cultura, procedimentos, sistemas e controles, capilarizando a organização com o objetivo único quando adotada: atingir melhores resultados, concentrando todos os esforços e energia da empresa para sua atividade principal (GIOSA, 1993, p.11).

O principal objetivo da terceirização é a redução de custos com pessoal e equipamentos, entre outros, através da transferência de serviços. O fator que motiva essa mudança de gestão de carteira é o fato de minimizar as perdas financeiras associadas a contratos que já causaram prejuízo à organização.

Nesse contexto, são usualmente aplicados modelos de regressão logística para estimar a probabilidade de um cliente inadimplente vir a realizar um pagamento em um determinado período de tempo. Contudo, modelos alternativos vêm sendo utilizados com esse propósito como, por exemplo, o modelo de mistura com fração de fidelizados (QUIDIM, 2005; TOMAZELA et al., 2007; GRANZOTTO *et al*., 2010). Em tal modelo, além da probabilidade de pagamento, se tem também informações relevantes sobre o comportamento dos clientes durante o período de tempo observado.

No presente trabalho, ambos os modelos (regressão logística e mistura com fração de fidelizados) são aplicados aos dados de um grupo de clientes inadimplentes de uma empresa de telecomunicações sendo, então, realizada uma comparação e discussão das possíveis vantagens de um relação ao outro.

# 3 MATERIAL E MÉTODOS

## 3.1 Material

O banco de dados utilizado neste trabalho foi disponibilizado por uma empresa de telecomunicações que oferece aos seus clientes serviços de telefonia fixa e internet, conhecidos popularmente por 'Voz' e 'Dados', respectivamente. O mesmo consiste de uma amostra de 5.000 clientes desta empresa que apresentam em comum a característica de terem se tornado inadimplentes em janeiro de 2009, ou seja, não efetuaram o pagamento da fatura desse mês até a data do seu vencimento, passando, assim, a serem monitorados a cada cinco dias quanto ao atraso do pagamento desta fatura e das demais emitidas nos 150 dias subsequentes. A base de dados foi construída considerando a data de vencimento da fatura como o primeiro ponto de observação, ou seja, todos os clientes apresentavam zero dias de atraso no início do estudo; a partir desse ponto, o atraso destes clientes foram monitorados por 150 dias em intervalos de cinco dias. A base de dados contém informações sobre o perfil e comportamento desses clientes inadimplentes tais como: descrição do serviço contratado (telefonia fixa, internet ou ambos); idade do contratante em anos; data do vencimento da fatura de janeiro de 2009; valor da fatura com vencimento em janeiro de 2009; número de vezes em que o cliente esteve em cobrança nos últimos seis meses e parcelamento quebrado, que informa se o cliente realizou negociação anterior com a empresa para parcelamento de sua(s) fatura(s) atrasadas e não honrou com os pagamentos nas datas acordadas.

Após a realização de uma análise exploratória da base de dados disponibilizada, foi verificada a existência de algumas inconsistências tais como: clientes com valores faltantes em algumas das covariáveis; clientes com perfil atípico da maioria e categorias de algumas das covariáveis com frequência de clientes demasiadamente pequenas. A fim de evitar vieses nas análises, esses clientes foram removidos da base de dados original restando um total de 4.535 clientes para a realização das análises. A Tabela 1 apresenta uma descrição das covariáveis disponíveis no banco de dados analisado neste trabalho.

**Tabela 1.** Descrição das covariáveis disponíveis na base de dados de 4.535 clientes inadimplentes de uma empresa de telecomunicações

| **Covariável** | **Descrição** |
|---|---|
| Produto contratado | 0 = Voz ou 1 = Dados + Voz |
| Idade do cliente | 18 a 80 anos |
| Data de vencimento da fatura de janeiro de 2009 | 2 a 31 |
| Valor em reais da fatura de janeiro de 2009 | R$ 30,00 a R$ 300,00 |
| Número de vezes em cobrança nos últimos 6 meses | 0 a 5 |
| Parcelamento quebrado | 0 = Sim ou 1 = Não |

Fonte: Os Autores (2013).

A empresa de telecomunicações que disponibilizou os dados utilizados nesse trabalho tem como política o encerramento do relacionamento com clientes que ultrapassam um atraso máximo tolerado para pagamento de suas faturas; atraso este que corresponde a 60 dias. Quando isso ocorre, os contratos desses clientes tornam-se escopo de empresas terceirizadas especializadas na cobrança de clientes inadimplentes; ou seja, suas dívidas são negociadas com estas empresas e, nesses casos, o atraso desses clientes deixa de ser monitorado.

Sendo assim, uma resposta de interesse diz respeito à ocorrência ou não do atraso máximo tolerado pela empresa; ou seja, Y = 1 se o cliente atingir atraso superior a 60 dias de atraso (mau pagador) e Y = 0, caso contrário (bom pagador). Nesse estudo, foi observado um total de 638 clientes classificados como maus pagadores, o que corresponde a 14,1% dos 4.535 monitorados por um período total de 150 dias.

Por outro lado, o tempo até o cliente atingir o atraso máximo tolerável e deixar de ser cliente da empresa de telecomunicações, também se caracteriza como outra variável de interesse. Nesse estudo, dos 638 clientes que atingiram atraso superior a 60 dias, foram observados tempos iguais a 65 dias para aqueles que não efetuaram nenhum pagamento após o início do estudo, bem como tempos de no máximo 150 dias para aqueles que efetuaram algum pagamento após o início do estudo. Já para os 3897 clientes que não atingiram atraso superior a 60 dias, seus respectivos tempos correspondem ao período total de monitoramento, isto é, 150 dias.

Em resumo, para os 638 clientes que atingiram o atraso máximo tolerável, seus tempos até tal atraso variaram entre 65 e 150 dias, enquanto para os que não

atingiram, estes foram iguais a 150 dias. No contexto de análise de sobrevivência, os tempos associados aos clientes que atingiram atraso superior a 60 dias são denominados tempos exatos, enquanto os tempos dos que não atingiram tal atraso, de tempos de censura.

## 3.2 Métodos

### 3.2.1 Modelo de Regressão Logística

A regressão logística se constitui em um dos principais modelos utilizados quando se deseja analisar dados em que a variável resposta é binária ou dicotômica. Mesmo quando a resposta de interesse não é originalmente binária, é usual que esta seja dicotomizada de modo que a probabilidade de sucesso possa ser estimada por meio de um modelo de regressão logística. Embora existam outros modelos para analisar dados em que a resposta é binária, a regressão logística se tornou popular por ser flexível do ponto de vista matemático, de fácil utilização e por apresentar interpretação simples de seus parâmetros (GIOLO, 2012).

Em geral, o objetivo ao ajustar um modelo de regressão logística é o de descrever a relação entre uma variável resposta e um conjunto de variáveis explicativas (preditoras ou independentes). Para isso, a regressão logística modela o valor esperado da variável resposta condicionado aos valores de p variáveis explicativas $\mathbf{z} = (z_1, ..., z_p)$, isto é, $E(Y \mid \mathbf{z}) = \pi(\boldsymbol{z})$, em que $E(Y \mid \mathbf{z})$ pertence ao intervalo [0,1] (HOSMER; LEMESHOW, 2000).

O modelo de regressão logística é expresso por:

$$\pi(\boldsymbol{z}) = \frac{\exp\{\boldsymbol{\beta}'\boldsymbol{z}\}}{1 + \exp\{\boldsymbol{\beta}'\boldsymbol{z}\}} = \frac{exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k z_k\right)}{1 + exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k z_k\right)},$$

em que $\mathbf{z} = (z_1, z_2, ..., z_p)$ denota o vetor de valores observados das variáveis explicativas, $\beta_0$ corresponde a uma constante e os componentes $\beta_k$ são os $p$ parâmetros ou coeficientes de regressão.

Observa-se, ainda, que:

$$1 - \pi(\boldsymbol{z}) = \frac{\exp\{\boldsymbol{\beta}'\boldsymbol{z}\}}{1 + \exp\{\boldsymbol{\beta}'\boldsymbol{z}\}} = \frac{exp\left(\beta_0 + \sum_{k=1}^{p}\beta_k z_k\right)}{1 + exp\left(\beta_0 + \sum_{k=1}^{p}\beta_k z_k\right)},$$

fornece a probabilidade de um indivíduo não apresentar a resposta de interesse.

O modelo de regressão logística pode também ser apresentado em termos do logito, em que o logaritmo da razão entre os termos$\pi(\boldsymbol{z})$ e $(1 - \pi(\boldsymbol{z}))$ fornece um modelo linear, isto é,

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

Essa transformação foi proposta por Berkson (1944) e é denominada logito, termo que vem do inglês *logit (logistic probability unit)*. A razão entre $\pi(\boldsymbol{z})$ e $(1 - \pi(\boldsymbol{z}))$, na transformação logito, é chamada de *odds* (chance).

A estimação dos parâmetros em regressão logística é feita, em geral, pelo método de máxima verossimilhança. Segundo Hosmer e Lemeshow (2000), os estimadores de máxima verossimilhança do vetor de parâmetros **β** são os valores que maximizam a função de verossimilhança L(**β**), a qual expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos. Esta função pode ser escrita como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\pi(z_i)^{y_i}\,(1 - \pi(z_i))^{1-y_i},$$

em que *i* = (1, ..., *n*) denota o conjunto de *n* indivíduos independentes.

### 3.2.2 Seleção de covariáveis e do modelo final

Com o intuito de avaliar a significância do efeito de cada uma das covariáveis e, posteriormente, avaliar a presença de multicolinearidade entre elas, foram ajustados, inicialmente, um modelo para cada uma das covariáveis separadamente. Considerando o teste de Wald (WALD, 1943), cuja estatística segue

distribuição qui-quadrado, foram mantidas nos passos subsequentes apenas aquelas que apresentaram um p-valor inferior a 0,05.

Considerando as covariáveis que individualmente apresentaram efeitos significativos foi, então, utilizado o método de seleção *stepwise in both directions* (inclusão e exclusão de covariáveis) para a seleção do modelo final. Esse procedimento inicia ajustando o modelo sem nenhuma covariável sendo, em seguida, incluídas as demais covariáveis uma a uma, iniciando por aquela que possui maior correlação com a resposta. Em cada passo, covariáveis podem ser incluídas ou removidas do modelo, de acordo com os níveis de significância estabelecidos para a inclusão e exclusão das mesmas.

Para cada modelo ajustado no método de seleção descrito, foram comparadas as estimativas deste modelo com aquelas fornecidas pelos modelos que consideraram as covariáveis separadamente com o objetivo de verificar a presença de multicolinearidade entre as covariáveis.

Além dos p-valores associados ao teste de Wald, também foram monitorados o critério de informação de Akaike (AIC) e a área abaixo da curva ROC dos modelos. Tais critérios adicionais foram utilizados devido à sensibilidade que o Teste de Wald apresenta em grandes amostras (GRANZOTTO *et al*., 2010).

O Critério de Akaike (AIC), introduzido em 1974 por Hirotsugu Akaike, penaliza os modelos com covariáveis desnecessárias e é calculado da seguinte forma:

$$AIC = -2logL(\boldsymbol{\theta}_g) + 2p,$$

em que $L(\boldsymbol{\theta}_g)$ é a função de máxima verossimilhança e *p* é o número de parâmetros do modelo. O modelo sugerido por esse critério será o que apresentar o menor AIC.

Quanto à curva ROC (*receiver operating characteristic*), foi utilizada para verificar o poder preditivo e a qualidade do modelo ajustado e é baseada nos conceitos de especificidade e sensibilidade. Sensibilidade é definida como a proporção de clientes bons que são classificados corretamente como bons, ou seja, é a proporção de verdadeiros positivos, sendo obtida por:

$$S = \frac{n_{bb}}{b},$$

em que b é o número de bons clientes de uma determinada população e $n_{bb}$ corresponde ao número de bons clientes classificados como bons pelo modelo.

Já a especificidade corresponde à proporção de clientes maus que são classificados corretamente como maus pelo modelo, ou seja, a proporção de verdadeiros negativos; é obtida de maneira similar à sensibilidade.

Para obtenção da curva ROC é necessário estabelecer pontos de corte, que estão no intervalo [0,1]. Estabelecido os pontos de corte, assume-se que Y=1, ou seja, diz-se que o cliente é mau pagador para as probabilidades preditas pelo modelo com valor superior ou igual ao ponto de corte. Em seguida, é construído um gráfico com os pares (x, y) = (1 − especificidade, sensibilidade) para os pontos de corte definidos anteriormente. O modelo com maior poder preditivo será o que apresentar área abaixo da curva ROC mais próxima a um, produzindo, assim, o maior percentual de acertos.

### 3.2.3 Análise de Sobrevivência

A análise de sobrevivência é utilizada quando se deseja estimar a probabilidade de sobrevivência a um evento de interesse, denominado falha, associada a cada instante de tempo durante um período pré-estabelecido de observação.

No contexto desse trabalho, a falha é definida como:

$$\delta_i = \begin{cases} 1 \; se \; o \; atraso \; máximo \; do \; cliente \; for \; superior \; a \; 60 \; dias; \\ 0 \; se \; o \; atraso \; máximo \; do \; cliente \; for \; inferior \; a \; 60 \; dias. \end{cases}$$

A presença de censuras, que são observações parciais ou incompletas, é a principal característica dos dados de sobrevivência, fornecendo a informação que o tempo de falha é superior ao observado. Podem ser citadas como causas para a ocorrência de censuras:

a. A ausência da falha até o final do período pré-estabelecido;
b. A interrupção do acompanhamento do indivíduo no decorrer do estudo;
c. A ocorrência do evento por motivo diferente do estudado.

Nesse trabalho, censuras referem-se aos clientes que no final do período observado não atingiram o atraso pré-estabelecido de 60 dias para a negociação de suas dívidas às empresas especializadas em cobrança.

*3.2.3.1 Estimador não-paramétrico de Kaplan-Meier*

O modelo de regressão logística é usualmente utilizado para estimar a probabilidade de ocorrência de um evento de interesse, porém esse modelo não leva em consideração observações censuradas. Por isso, no contexto descritivo de análise de sobrevivência foi utilizado, neste trabalho, o estimador não paramétrico de Kaplan-Meier (KAPLAN; MEIER, 1958) para a estimação da função de sobrevivência.

Tal estimador também é conhecido como estimador limite-produto, sendo expresso por:

$$\hat{S}(t) = \prod_{j:\, t_j < t} \left( \frac{n_j - d_j}{n_j} \right) = \prod_{j:\, t_j < t} \left( 1 - \frac{d_j}{n_j} \right),$$

em que, $t_j$ (*j = 1,..., k*) são os *k* tempos distintos e ordenados de falha, $d_j$ é o número de falhas em $t_j$, para *j = 1,..., k, e* $n_j$ é o número de indivíduos sob risco em $t_j$.

*3.2.3.2 Modelo de Mistura de Cox com Fração de Fidelizados*

Em análise de sobrevivência é usual assumir que todos os indivíduos sob estudo irão apresentar o evento de interesse se forem acompanhados por um período de tempo suficientemente longo para que isso ocorra. Contudo, existem situações em que uma fração de indivíduos não apresentará o evento de interesse, mesmo se acompanhados por um longo período. Em tais casos, e dependendo da

área dos dados analisados (médica, financeira etc.), essa fração de indivíduos é denominada: imunes, curados, fidelizados, sobreviventes de longa duração ou, ainda, não suscetíveis ao evento de interesse (EUDES *et al.*, 2012).

Devido à sua versatilidade, o modelo de Cox (COX,1972) tem sido o mais utilizado para a análise de dados de sobrevivência. No entanto, esse modelo não seria adequado às situações em que há uma fração de indivíduos não suscetíveis ao evento, como é o caso dos dados analisados nesse trabalho. Sendo assim, optou-se pela utilização de um modelo capaz de acomodar essa fração de indivíduos não suscetíveis. Tal modelo, denominado modelo de mistura de Cox com fração de cura (ou fração de fidelizados, no contexto dos dados analisados), corresponde a uma extensão do modelo de Cox proposta por Kuk e Chen em 1992.

Esse modelo considera que existem duas subpopulações distintas (uma suscetível e outra não suscetível ao evento de interesse); daí o termo mistura. Para a formulação do modelo, considere U uma variável que denota se o indivíduo é suscetível (U = 1) ou não suscetível (U = 0) ao evento de interesse, tal que P(U = 1) = $\pi(\boldsymbol{z})$ e P(U = 0) = 1 - $\pi(\boldsymbol{z})$. Assim, para T uma variável aleatória não-negativa denotando o tempo até a falha, a probabilidade de sobreviver ao tempo pode ser expressa por:

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

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

$$S(t|\boldsymbol{x},\boldsymbol{z}) = \pi(\boldsymbol{z})S(t|U = 1,\boldsymbol{x}) + 1 - \pi(\boldsymbol{z}), \qquad (1)$$

sendo $S(t|\boldsymbol{x},\boldsymbol{z})$ a função de sobrevivência para toda a população (suscetíveis e não suscetíveis) e $S(t \mid U = 1,\boldsymbol{x})$ a função de sobrevivência para os indivíduos suscetíveis, com x e z vetores de covariáveis e $\pi(\boldsymbol{z})$ a probabilidade do indivíduo ser suscetível, dado o vetor de covariáveis **z**. As mesmas covariáveis podem ou não estar presentes em ambos os componentes $\pi(\boldsymbol{z})$ e $S(t \mid U = 1,\boldsymbol{x})$ (CORBIÈRE; JOLY, 2007).

A probabilidade $\pi(\boldsymbol{z})$ pode ser modelada utilizando modelos de regressão para dados binários. No contexto desse trabalho, para facilitar a comparação com o modelo de regressão logística, optou-se por utilizar aqui também o logito, cuja expressão é dada por:

$$\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{z})}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{z})}.$$

Quanto à função $S(t \mid U = 1, \boldsymbol{x})$, esta pode ser modelada por meio de modelos paramétricos (exponencial, Weibull, logístico etc.) ou pelo modelo semiparamétrico de Cox. Neste trabalho, foi modelada pelo modelo de Cox, de modo que:

$$S(t \mid U = 1, \boldsymbol{x}) = S_0(t|U = 1)^{\exp(\boldsymbol{\gamma}'\boldsymbol{x})},$$

em que $S_0(t \mid U = 1)$ é denominada função de sobrevivência de base.

A função de verossimilhança associada ao modelo (1) é dada por:

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

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

Estimadores para os vetores de parâmetros γ e β, bem como para $S_0(t|U = 1)$, são obtidos maximizando-se a função de verossimilhança L(**γ**, **β**) via o algoritmo EM (do inglês, *estimation and maximization*). Para tal estimação, Corbière e Joly (2007) desenvolveram uma macro no *software* SAS. Nesta macro, $\pi(z)$ pode ser modelada por meio dos modelos: logístico, probito e complemento log-log, enquanto $S(t \mid U = 1, \boldsymbol{x})$ pelos modelos paramétricos exponencial, Weibull, logístico e log-normal ou, ainda, pelo modelo semiparamétrico de Cox.

Para a seleção de covariáveis do modelo, foi utilizado o método *stepwise in both directions* (inclusão e exclusão de covariáveis) e o Critério de Akaike (AIC), técnicas descritas na Seção 3.2.2, porém, para o *stepwise*, foram reestabelecidos os níveis de significância para inclusão e exclusão das covariáveis. Para o componente logístico o nível permaneceu em 0,05, enquanto que para o componente de sobrevivência fixou-se 0,20. Estes níveis de significância diferem devido ao número

de observações utilizadas na construção de cada componente do modelo; enquanto o componente logístico considera toda a população para estimar os parâmetros, o componente de sobrevivência utiliza apenas as observações que falharam, restringindo bastante o número de observações disponíveis e por isso é considerado para este componente um nível menos conservador.

Para verificar a adequação do modelo, foram utilizados o coeficiente de correlação de Pearson e o $R^2$ (coeficiente de Pearson ao quadrado). O coeficiente de Pearson mede a correlação entre as probabilidades de sobrevivência obtidas por Kaplan-Meier e pelo modelo de fração de fidelizados ajustado para diversas combinações das covariáveis. Coeficientes próximos de 1 (um) evidenciam um ajuste satisfatório do modelo.

# 4 RESULTADOS

Os resultados apresentados a seguir foram obtidos com o auxílio do software SAS. Os principais comandos utilizados nos procedures PROC LIFETEST e PROC LOGISTIC, bem como na macro PSPMCM, são fornecidos no Apêndice A.

## 4.1 Análise Descritiva

Inicialmente foi realizada uma análise descritiva dos dados dos 4.535 clientes em cobrança acompanhados no estudo, em que 638 chegaram ao atraso máximo admissível pela empresa. Considerando essa análise, pode-se observar a partir da Tabela 2 que a maioria dos clientes (73%) possui contrato de serviços de telefonia e internet (dados + voz) e que o percentual de maus clientes na categoria apenas voz é superior (21,6% contra 11,3%). Esses pacotes de serviços oferecidos variam quanto aos minutos disponíveis para realização de ligações e velocidade de internet, sendo que a maior parte custa na faixa de R$80,00 a R$190,00, representando 67% dos clientes. Quanto ao dia do vencimento da fatura, os clientes escolhem o melhor dia para o débito, sendo que usualmente é perto do dia em que recebem seus salários. É sabido que a maior parte das empresas realiza o pagamento no início do mês, o que justifica o acúmulo de clientes na primeira faixa – primeiros dias do mês.

Referente à informação de inadimplência anterior, tem-se as covariáveis: indicativo de parcelamento quebrado e quantidade de vezes em cobrança, com ambas indicando que a maior parte dos clientes inadimplentes nesse estudo não entrou em atraso no histórico recente. Pode-se observar que 63,2% não atrasaram nenhuma fatura nos últimos seis meses e, dos que atrasaram, apenas 8,5% não honraram um acordo firmado com a empresa, ficando inadimplentes inclusive na renegociação (parcelamento) realizada.

Nota-se, também, que a maior parte dos clientes é jovem e que o não pagamento é maior para esses clientes. Na primeira faixa etária, o percentual de maus clientes foi de 16%, enquanto na segunda faixa, caiu para 12%.

**Tabela 2** - Estatística descritiva das covariáveis de 4.535 clientes que atrasaram o pagamento da fatura janeiro de 2009 de uma empresa de telecomunicações

| Covariável | Total | Total(%) | Falha/Mau | Falha/Mau(%)* |
|---|---|---|---|---|
| *Produto* | | | | |
| Dados + Voz | 3.310 | 73,0 | 373 | 11,3% |
| Voz | 1.225 | 27,0 | 265 | 21,6% |
| *Valor da Fatura (em reais R$)* | | | | |
| 30,00 a 80,00 | 825 | 18,2 | 155 | 18,8% |
| 80,00 a 135,00 | 1.833 | 40,4 | 204 | 11,1% |
| 135,00 a 190,00 | 1.221 | 26,9 | 149 | 12,2% |
| 190,00 a 245,00 | 450 | 10,0 | 83 | 18,4% |
| 245,00 a 300,00 | 206 | 4,5 | 47 | 22,8% |
| *Quantidade de vezes em cobrança* | | | | |
| Sem atraso anterior | 2.867 | 63,2 | 231 | 8,1% |
| 1 | 969 | 21,3 | 213 | 22,0% |
| 2 | 342 | 7,5 | 98 | 28,7% |
| 3 | 172 | 3,8 | 46 | 26,7% |
| 4 | 106 | 2,2 | 32 | 30,2% |
| 5 | 79 | 1,7 | 18 | 22,8% |
| *Dia do Vencimento* | | | | |
| 1 a 10 | 2.365 | 52,1 | 354 | 15,0% |
| 11 a 20 | 1.438 | 31,7 | 186 | 12,9% |
| 21 a 31 | 732 | 16,2 | 98 | 13,4% |
| *Indicativo de Parcelamento Quebrado* | | | | |
| Sim | 385 | 8,5 | 254 | 66,0% |
| Não | 4.150 | 91,5 | 384 | 9,3% |
| *Idade (em anos)* | | | | |
| 18 a 35 | 1.676 | 37,0 | 282 | 16,8% |
| 36 a 50 | 1.619 | 35,7 | 202 | 12,5% |
| 51 a 65 | 950 | 20,9 | 111 | 11,7% |
| 66 a 80 | 290 | 6,4 | 43 | 14,8% |

*Percentual de maus pagadores por classe da covariável (Mau/Total).
Fonte: Os Autores (2013).

### 4.2 Ajuste do Modelo de Regressão Logística

Inicialmente foi ajustado o modelo de regressão logística para cada uma das covariáveis apresentadas na Tabela 1, conforme descrito na Seção 3.2.3. Foram mantidas nos passos subsequentes apenas aquelas que apresentaram um p-valor inferior a 0,05. Desse modo, a covariável dia de vencimento da fatura foi removida por não ter apresentado efeito significativo (p-valor de 0,1158), conforme pode ser observado na Tabela 3.

**Tabela 3** – Ajuste do modelo de regressão logística apenas com a covariável idade.

| Teste | Chi-Quadrado | Grau de Liberdade | p-valor |
|---|---|---|---|
| Máxima Verossimilhança | 2.4925 | 1 | 0,1144 |
| Wald | 2.4739 | 1 | 0,1158 |

Fonte: Os Autores (2013)

Utilizando o método de seleção de covariáveis *stepwise (both directions)*, verificou-se uma inversão de sinais nas estimativas dos parâmetros associados à covariável valor da fatura quando esta covariável foi incluída no modelo, evidenciando a presença de multicolinearidade desta com uma ou mais covariáveis no modelo. Optou-se, assim, por remover tal covariável devido ao p-valor associado à mesma ser menos significativo do que o das demais covariáveis e pela dificuldade de interpretar risco associado ao valor devido. O nível de significância estabelecido para inclusão e exclusão das covariáveis foi fixado em 0,05.

Os três critérios utilizados na escolha do modelo final (p-valores associados ao Teste de Wald; valores do AIC e áreas abaixo da curva ROC), convergiram para a escolha das mesmas covariáveis. O modelo selecionado foi o com as seguintes covariáveis: idade, produto, parcelamento quebrado e quantidade de vezes em cobrança. As estimativas dos parâmetros associados a esse modelo estão na Tabela 4, bem como o AIC e logaritmo da verossimilhança na Tabela 5.

**Tabela 4** – Estimativas e testes associados ao modelo de regressão logística selecionado

| Parâmetros | GL | Estimativa | Erro Padrão | Wald Qui-Quadrado | p-valor |
|---|---|---|---|---|---|
| Intercepto | 1 | 1,4316 | 0,2087 | 47,0445 | <0,0001 |
| Idade | 1 | -0,0149 | 0,00358 | 17,2214 | <0,0001 |
| Produto: Dados + Voz | 1 | -0,7992 | 0,1030 | 60,2253 | <0,0001 |
| Parcelamento quebrado: N | 1 | -2,7851 | 0,1243 | 501,8544 | <0,0001 |
| Quantidade de vezes em cobrança | 1 | 0,2975 | 0,0369 | 65,0398 | <0,0001 |

Fonte: Os Autores (2013).

**Tabela 5** - Estatísticas associadas ao modelo sem covariáveis e ao modelo selecionado

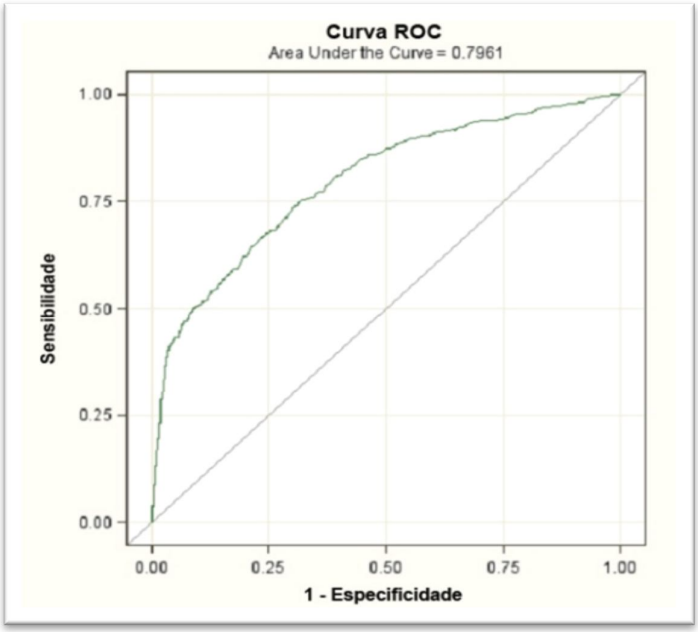
| Critério | Somente Intercepto | Intercepto e Covariáveis |
|---|---|---|
| AIC | 3.686,256 | 2.928,159 |
| -2 Log L | 3.684,256 | 2.918,159 |

Fonte: Os Autores (2013).

A partir da Tabela 4, pode-se notar que todas as covariáveis apresentaram significância estatística ao nível de 0,01; nível este mais rigoroso do que o estabelecido anteriormente. Ainda, o AIC do modelo que considera todas as covariáveis é menor do que o do modelo que considera somente o intercepto (Tabela 5), o que confirma a hipótese de que essas covariáveis ajudam a explicar a resposta com parcimônia, dado que o AIC penaliza a inclusão de covariáveis desnecessárias. Além disso, esse é o modelo que maximiza a área sob a curva ROC.

Os critérios AIC e a curva ROC, citados para a seleção do modelo, também foram utilizados para verificar a adequação do modelo. Pode-se visualizar a partir da Figura 1 que o modelo escolhido se ajusta aos dados de maneira satisfatória, apresentando bom poder de discriminação com uma área de 0,7961 abaixo da curva.

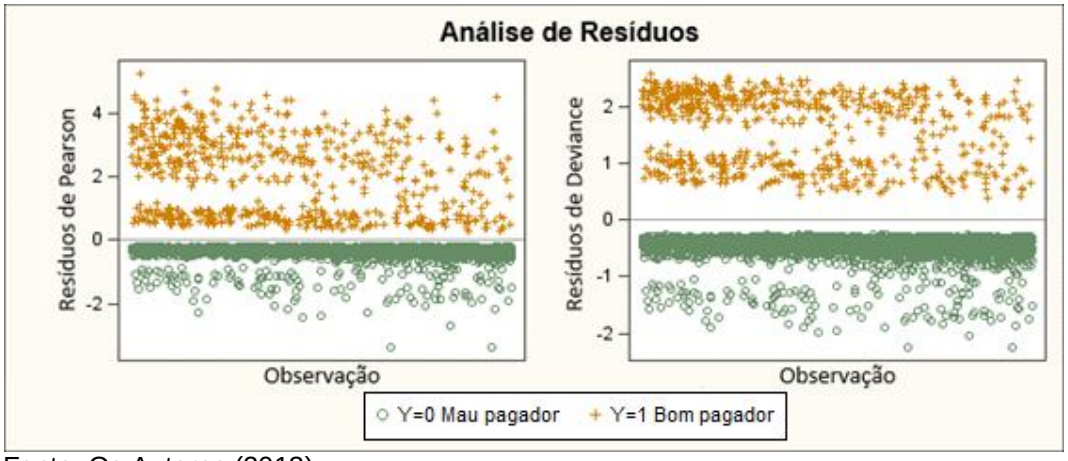
**Figura 1-** Curva ROC associada ao modelo ajustado




Fonte: Os Autores (2013).

Por fim, foi feita a análise de resíduos e, conforme pode-se observar na Figura 2, os resíduos *deviance* e de Pearson estão distribuídos em torno de zero e em um intervalo de variação satisfatório, o que reitera a adequação do modelo ajustado.

**Figura 2 –** Análise gráfica dos resíduos do modelo de regressão logística ajustado




Fonte: Os Autores (2013).

O modelo ajustado, em termos dos logitos, ficou expresso por:

$$logit\left(\hat{\pi}(z_i)\right) = 1{,}4316 + (-0{,}0149)z_{i1} + (-0{,}7992)z_{i2} + (-2{,}7851)z_{i3} + (0{,}2975)z_{i4},$$

sendo $z_{i1}$ a idade do cliente em anos; $z_{i2}$= 1 se produto contratado for dados e voz e 0, se for somente voz; $z_{i3}$= 0 se o cliente possui parcelamento quebrado e 1, caso contrário e; $z_{i4}$ a quantidade de vezes que o cliente esteve em cobrança nos últimos seis meses.

Razões de chances (*odds ratios)* podem ser calculadas com base nas estimativas dos parâmetros, produzindo resultados interessantes a respeito do perfil dos clientes em função dos valores das covariáveis. Ressalta-se, que a *odds* pode ser estimada simplesmente por $exp\{\widehat{\beta}_i\}$ devido à ausência de interações no modelo. Por exemplo, para clientes que possuem somente o produto voz, tem-se:

$$\widehat{OR}_{produto(0|1)} = exp\{-\widehat{\beta_2}\},$$

o que resulta em:

$$\widehat{OR}_{produto(0|1)} = exp\{0{,}7992\} = 2{,}22.$$

Logo, a *odds* (ou chance) de um indivíduo se tornar mau cliente caso ele tenha contratado apenas o pacote de serviços de voz é 2,22 vezes a de um indivíduo que contratou o pacote de serviços dados + voz, mantendo fixas as demais covariáveis.

De maneira similar, utilizando as estimativas da Tabela 4, tem-se que a *odds ratio* entre clientes que possuem e não possuem parcelamento quebrado é dada por

$$\widehat{\mathrm{OR}}_{\mathrm{parc.queb}(0|1)} = \exp\{-(-2{,}7851)\};$$

ou seja, os clientes que possuem parcelamento quebrado apresentam chance 16,2 vezes maior de se tornarem maus pagadores do que aqueles que não possuem.

Analogamente, *odds ratios* podem ser calculadas para as demais variáveis a fim de se caracterizar o perfil dos clientes dessa empresa. De modo geral, foi encontrado que o grupo de clientes que apresentou a maior chance de se tornar mau pagador após os 150 dias observados foi aquele composto por indivíduos com as seguintes características: possuem parcelamento quebrado, têm idade 18 anos, contrataram o produto voz e atrasaram todas as faturas nos últimos seis meses (quantidade de vezes em cobrança igual a 5). Em contrapartida, o grupo com a menor chance foi o composto por indivíduos que não possuem parcelamento quebrado, estão nas maiores faixa-etárias, contrataram os produtos de dados e voz e não atrasaram nenhuma fatura nos últimos seis meses.

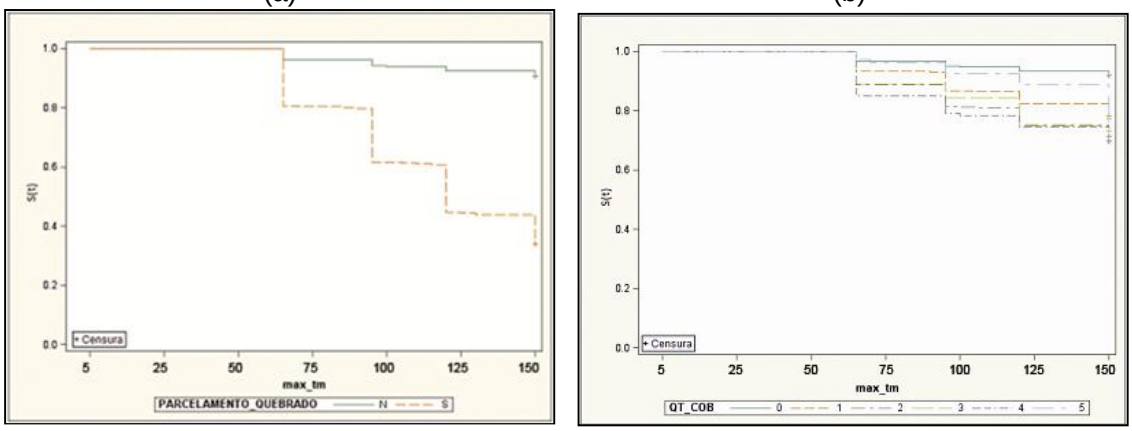
### 4.3 Ajuste do Modelo de Mistura com Fração de Fidelizados

Com o intuito de analisar o tempo até os clientes inadimplentes terem suas dívidas negociadas com empresas especializadas em cobrança em função de atrasos nos pagamentos, realizou-se uma análise descritiva de cada uma das covariáveis fazendo-se uso do estimador de Kaplan-Meier. Essa análise, que indica possíveis associações das covariáveis com o tempo mencionado, serviu de referencial para a escolha das variáveis candidatas a entrarem nos modelos, podendo também auxiliar na interpretação dos modelos finais.

De acordo com a Figura 3a, parece haver diferenças entre as curvas associadas ao grupo que possui parcelamento quebrado e ao que não possui, evidenciando que esta variável pode auxiliar na discriminação da variável resposta.

**Figura 3 -** Curvas de sobrevivência obtidas pelo estimador de Kaplan-Meier para as covariáveis parcelamento quebrado (a) e quantidade de vezes em cobrança (b)

(a) (b)

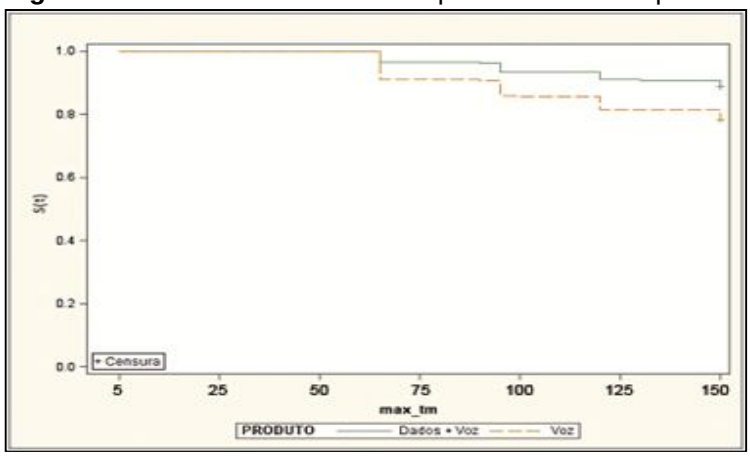



Fonte: Os Autores (2013).

Quanto à covariável quantidade de vezes em cobrança nos últimos seis meses, que é ordinal e possui seis categorias, nota-se, a partir da Figura 3b, que os clientes que estiveram mais vezes em cobrança apresentam, em geral, um número maior de falhas (i.e., atingem o limite máximo de atraso tolerável de 60 dias) em relação aos que estiveram menos vezes em cobrança. Os clientes que estiveram cinco vezes em cobrança apresentam uma queda brusca na probabilidade de sobrevivência (i.e., na probabilidade de não atingir o limite máximo de atraso) no último tempo observado (150 dias), diferente dos outros grupos, nos quais se observam quedas mais homogêneas e menos acentuadas nesta probabilidade ao longo do tempo. Ao final do estudo, pode-se observar que alguns grupos parecem diferir dos demais; por exemplo, o grupo de clientes que não estiveram em cobrança apresenta probabilidade maior de não atingir o limite máximo de atraso.

A partir da Figura 4, pode-se também observar que existem diferenças entre as curvas segmentadas por produto, sendo que os clientes que contrataram apenas o serviço de voz apresentam um percentual superior de falhas em relação aos que contrataram os serviços de dados e voz.

**Figura 4-** Curvas de sobrevivência para a covariável produto




Fonte: Os Autores (2013).

Com base nas curvas de sobrevida estimadas pelo método proposto por Kaplan-Meier, pode-se verificar também um grande número de censuras à direita e, consequentemente, um acúmulo de indivíduos nas faixas mais altas de probabilidade de sobrevivência. Esse fato já era esperado e foi o que motivou a escolha do modelo de mistura com fração de fidelizados para modelagem do tempo de sobrevida dos clientes inadimplentes dessa empresa de telecomunicações.

Para proceder a seleção das covariáveis foi, inicialmente, ajustado um modelo de mistura com fração de fidelizados para cada uma das covariáveis. Em cada um desses modelos, a covariável foi incluída simultaneamente no componente logístico $\pi(\mathbf{z})$ e no componente de sobrevivência $S(t|U=1, \mathbf{x})$. Foram mantidas nos passos subsequentes apenas aquelas que apresentaram um p-valor inferior a 0,05 no componente logístico e 0,20 no componente de sobrevivência. A covariável dia de vencimento da fatura foi removida do componente logístico por não ter apresentado efeito significativo com p-valor 0,12. Do componente de sobrevivência foram removidas as covariáveis: indicativo de parcelamento quebrado (p-valor 0,688); quantidade de vezes em cobrança (p-valor 0,54) e; dia do vencimento da fatura (p-valor 0,64).

Desconsiderando as covariáveis que foram removidas por não terem apresentado significância estatística individualmente, foi utilizado o método de seleção de covariáveis *stepwise*, considerando os níveis de significância para inclusão e exclusão das covariáveis em 0,05 para a parte logística e 0,20 para a parte de sobrevivência. Dessa forma, no componente logístico do modelo

selecionado permaneceram as covariáveis: idade, quantidade de vezes em cobrança, parcelamento quebrado e produto. Já no componente de sobrevivência permaneceram apenas as covariáveis produto e idade.

O componente associado à regressão logística do modelo de mistura com fração de fidelizados apresentou estimativas idênticas às do modelo que considera apenas essa técnica, pois foram selecionadas as mesmas covariáveis[1]. Já, as estimativas dos parâmetros associadas ao componente de sobrevivência (modelado via o modelo de Cox) são apresentadas na Tabela 6, bem como o AIC e logaritmo da verossimilhança na Tabela 7.

**Tabela 6 -** Estimativas e testes associados ao componente S(t|U=1,x) do modelo de mistura com fração de fidelizados selecionado

| Parâmetros | GL | Estimativa | Erro Padrão | Wald Qui-Quadrado | p-valor |
|---|---|---|---|---|---|
| Idade | 1 | -0,00468 | 0,00279 | 2,8289 | 0,0926 |
| Produto: Dados + Voz | 1 | -0,16298 | 0,08212 | 3,9387 | 0,0472 |

Fonte: Os Autores (2013).

**Tabela 7 -** Estatísticas associadas ao modelo sem covariáveis e ao modelo selecionado

| Critério | Somente Intercepto | Intercepto e Covariáveis |
|---|---|---|
| AIC | 7.427,391 | 7.425,731 |
| -2 Log L | 7.427,391 | 7.421,731 |

Fonte: Os Autores (2013).

A partir da Tabela 6, pode-se notar que todas as covariáveis apresentaram significância estatística ao nível de 0,10, obedecendo o critério de 0,20 que foi estabelecido anteriormente. Ainda, da Tabela 7, tem-se que o AIC do modelo que considera todas as covariáveis é menor do que o do modelo que considera somente o intercepto.

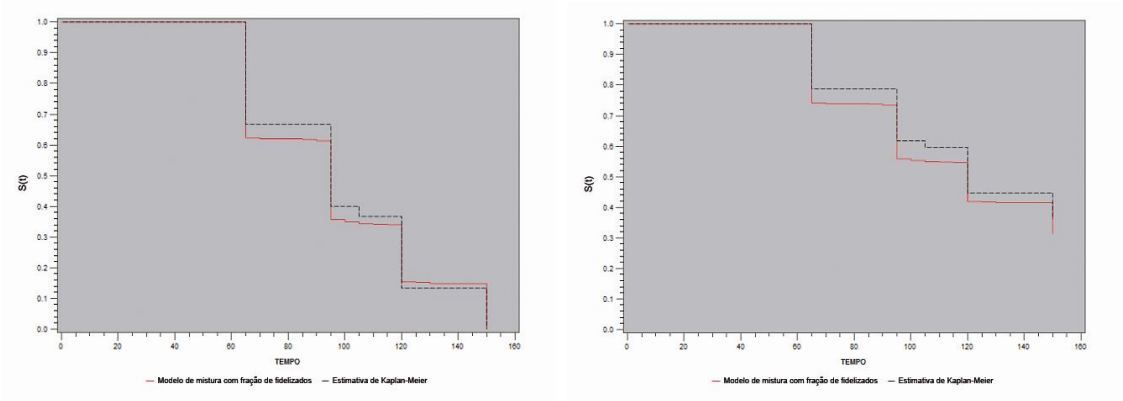
Para verificar a adequação do modelo selecionado, foram, adicionalmente, obtidas as curvas de sobrevivência observada (representada pela curva obtida pelo estimador de Kaplan-Meier) e a estimada pelo modelo, tanto para a sobrevivência populacional S(t | **x**, **z**) quanto para a sobrevivência condicional S(t | U=1, **x**). Estas

[1]Tabelas 4 e 5

curvas, para uma das combinações de **x** e **z**, podem ser visualizadas na Figura 5 e mostram que as estimativas produzidas pelo modelo são bastante próximas às obtidas por Kaplan-Meier, evidenciando a adequação do modelo aos dados.

**Figura 5 -** Curvas de sobrevivência observada e estimada pelo modelo de fração de fidelizados ajustado para (a) S(t | U=1,**x**) e (b) S(t | **x**, **z**) para uma das combinações das covariáveis categóricas e idade média (42 anos)

(a) (b)




Fonte: Os Autores (2013).

Ainda, para avaliar a adequação do modelo selecionado, foi considerado o coeficiente de correlação de Pearson entre as probabilidades de sobrevivência observadas e estimadas pelo modelo para cada uma das combinações das covariáveis categóricas e idade média (42 anos), bem como o $R^2$. Ambas as estatísticas evidenciaram o bom ajuste do modelo, produzindo correlações bastante altas (Tabela 8). O menor valor para o coeficiente de Pearson foi 0,95 e para o $R^2$ 0,902.

A expressão do modelo final ficou dada por:

$$S(t|\boldsymbol{x},\boldsymbol{z}) = \pi(\boldsymbol{z})S(t|U = 1,\boldsymbol{x}) + 1 - \pi(\boldsymbol{z})$$

$$S(t|\boldsymbol{x},\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}'\boldsymbol{\beta})} S_0(t|U = 1)^{\exp(\boldsymbol{\gamma}'\boldsymbol{x})} + 1 - \frac{\exp(\boldsymbol{z}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}'\boldsymbol{\beta})}.$$

As estimativas obtidas para os componentes do vetor **β** são as mesmas obtidas para o modelo de regressão logística ajustado na Seção 4.2 e, por consequência, produzem conclusões análogas para a probabilidade de ocorrência do evento em função das covariáveis **z**.

**Tabela 8** - Correlação estatística entre as funções de sobrevivência estimadas e observadas (Kaplan-Meier) para cada uma das combinações das covariáveis categóricas e idade média

| Estrato | Produto | Parcelamento Quebrado | Quantidade de vezes em cobrança | $R^2$ | Coeficiente de correlação de Pearson |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0,9977 | 0,9989 |
| 2 | 0 | 0 | 1 | 0,9960 | 0,9980 |
| 3 | 0 | 0 | 2 | 0,9601 | 0,9798 |
| 4 | 0 | 0 | 3 | 0,9755 | 0,9877 |
| 5 | 0 | 0 | 4 | 0,9027 | 0,9501 |
| 6 | 0 | 0 | 5 | 0,9879 | 0,9939 |
| 7 | 0 | 1 | 0 | 0,9927 | 0,9963 |
| 8 | 0 | 1 | 1 | 0,9972 | 0,9986 |
| 9 | 0 | 1 | 2 | 0,9480 | 0,9737 |
| 10 | 0 | 1 | 3 | 0,9958 | 0,9979 |
| 11 | 0 | 1 | 4 | 0,9899 | 0,9949 |
| 12 | 0 | 1 | 5 | 0,9870 | 0,9935 |
| 13 | 1 | 0 | 0 | 0,9894 | 0,9947 |
| 14 | 1 | 0 | 1 | 0,9872 | 0,9936 |
| 15 | 1 | 0 | 2 | 0,9619 | 0,9808 |
| 16 | 1 | 0 | 3 | 0,9698 | 0,9848 |
| 17 | 1 | 0 | 4 | 0,9187 | 0,9585 |
| 18 | 1 | 0 | 5 | 0,9698 | 0,9848 |
| 19 | 1 | 1 | 0 | 0,9914 | 0,9957 |
| 20 | 1 | 1 | 1 | 0,9979 | 0,9990 |
| 21 | 1 | 1 | 2 | 0,9993 | 0,9996 |
| 22 | 1 | 1 | 3 | 0,9995 | 0,9998 |
| 23 | 1 | 1 | 4 | 0,9986 | 0,9993 |
| 24 | 1 | 1 | 5 | 0,9024 | 0,9499 |

Fonte: Os Autores (2013).

Quanto à fração de fidelizados, dada uma combinação das covariáveis **z**, esta pode ser obtida por $1 - \pi(\boldsymbol{z})$. Na Tabela 9, estimativas para tais frações são apresentadas em função da quantidade de vezes em cobrança nos últimos 6 meses (0 a 5) para clientes com: 40 anos de idade, que contrataram o produto dados + voz e que possuem parcelamento quebrado. Conforme resultados nesta tabela, pode-se notar que a fração de fidelizados foi maior para clientes que estiveram menos vezes em cobrança. Para os que nunca estiveram em cobrança, por exemplo, estima-se que 49,1% deles terão sua dívidas quitadas dentro do prazo de atraso tolerável pela empresa. Já para os que estiveram três e cinco vezes em cobrança, tal estimativa decresce para 28,3% e 17,9%, respectivamente.

**Tabela 9** – Tabela comparativa da fração de fidelizados de clientes com 40 anos de idade, que contrataram dados e voz e tiveram parcelamento quebrado, em função da quantidade de vezes em cobrança

| Quantidade de vezes em cobrança | $\pi(\boldsymbol{z})$ | $1-\pi(\boldsymbol{z})$ |
|---|---|---|
| 0 | 0,509 | 0,491 |
| 1 | 0,583 | 0,417 |
| 2 | 0,653 | 0,347 |
| 3 | 0,717 | 0,283 |
| 4 | 0,773 | 0,227 |
| 5 | 0,821 | 0,179 |

Fonte: Os Autores (2013).

Para os clientes não fidelizados (isto é, que atingiram o prazo máximo de atraso tolerável pela empresa), o componente de sobrevivência S(t | U=1,**x**) do modelo de fração de fidelizados, permite algumas conclusões específicas sobre o perfil dos mesmos. Por exemplo, dado que as covariáveis idade e produto contratado foram as que apresentaram efeito significativo para este componente (Tabela 6), pode-se dizer que as mesmas influenciaram o tempo *t* em que tais clientes se tornaram não fidelizados.

Assim, considerando o exponencial das estimativas (Tabela 6) associadas ao componente S(t | U=1,**x**), os quais correspondem à razões de risco (ou taxas de falha), tem-se que os clientes mais jovens e que contrataram somente o produto voz foram os que se tornaram não fidelizados em um menor período de tempo.

Por exemplo, para clientes de mesma idade, os que contrataram apenas voz apresentaram risco de se tornarem não fidelizados em um menor período de tempo cerca de 17% maior (exp(0,16) = 1,17) do que os que contrataram dados e voz. De maneira similar, para clientes que contrataram o mesmo produto, tal risco se apresentou 4,8% maior (exp(0,00468*10)) a cada 10 anos de decréscimo na idade.

### 4.4 Comparação entre os Resultados dos Modelos

A partir dos resultados obtidos, tanto para o modelo de regressão logística quanto para o modelo de mistura com fração de fidelizados, foi possível estabelecer perfis bem distintos em relação ao risco de não pagamento dos clientes sob estudo. Dois desses perfis podem ser visualizados na Tabela 10. O perfil 1, composto dos clientes com contrato de serviços de dados + voz, idade mais avançada (80 anos),

sem histórico de cobranças anteriores e de parcelamento de suas dívidas, foi o que apresentou a menor probabilidade (3,43%) de atraso superior a 60 dias nos 150 dias subsequentes. Em contrapartida, o perfil 2, composto dos clientes com contrato de serviços de voz, jovens (18 anos), com histórico de cobranças anteriores e de parcelamento de suas dívidas, foi o que apresentou a maior probabilidade de atraso superior a 60 dias (93,4%).

**Tabela 10** – Perfis dos clientes com maior e menor probabilidade de se tornarem maus pagadores de acordo com ambos os modelos ajustados

| | Perfil (1) | Perfil (2) |
|---|---|---|
| Idade | 80 | 18 |
| Parcelamento quebrado | Não | Sim |
| Quantidade de vezes em cobrança | 0 | 5 |
| Produto contratado | Dados e Voz | Voz |
| Probabilidade de não-pagamento | 3,43% | 93,40% |

Fonte: Os Autores (2013).

Ainda com a finalidade de comparar os modelos de regressão logística e de mistura com fração de fidelizados ajustados aos dados, são mostrados na Tabela 11 resultados os quais são possíveis de serem obtidos para ambos os modelos considerando 24 combinações das covariáveis remanescentes nos mesmos.

A partir desta tabela, fica evidenciado que o modelo de regressão logística é capaz de fornecer informações sobre os clientes sob estudo apenas para o tempo final do estudo (150 dias); enquanto o modelo de mistura fornece informações sobre estes clientes em quaisquer tempos entre 0 e 150 dias (na tabela foram mostrados para os tempos 65, 95 e 150 dias). Além disso, o modelo de mistura fornece também informações sobre os clientes suscetíveis (maus pagadores) ao longo tempo, o que pode auxiliar a empresa a dinamizar suas estratégias de cobrança tomando ações mais enérgicas e rápidas voltadas para clientes com perfis similares aos desses clientes. Pode-se também notar, que as estimativas obtidas a partir de ambos os modelos para $\pi(z)$ no tempo 150 dias são idênticas; isto porque em ambos os modelos tal probabilidade foi modelada considerando o modelo de regressão logística, com as mesmas covariáveis tendo apresentado efeitos significativos em ambos.

**Tabela 11** – Estimativas obtidas com o modelo de regressão logística e modelo de mistura com fração de fidelizados ao longo do tempo, para cada combinação das covariáveis

| Estrato | Covariáveis | | | | **Regressão Logística** | **Modelo de mistura com fração de cura** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 - π(z) | 1 - π(z) | S(t\|x,z) | | | S(t\|U=1,x) | | |
| | Produto | Parc. Queb. | Qt. Cob. | Idade Média | t= 150 dias | t= 150 dias | t= 65 dias | t= 95 dias | t= 150 dias | t= 65 dias | t= 95 dias | t= 150 dias |
| 1 | 0 | 0 | 0 | 40 | 30,25% | 30,25% | 73,37% | 54,72% | 30,25% | 61,83% | 35,08% | 0% |
| 2 | 0 | 0 | 1 | 39 | 24,08% | 24,08% | 70,92% | 50,59% | 24,08% | 61,69% | 34,91% | 0% |
| 3 | 0 | 0 | 2 | 39 | 19,07% | 19,07% | 69% | 47% | 19% | 62% | 35% | 0% |
| 4 | 0 | 0 | 3 | 49 | 16,88% | 16,88% | 69% | 47% | 17% | 63% | 37% | 0% |
| 5 | 0 | 0 | 4 | 41 | 11,81% | 11,81% | 66% | 43% | 12% | 62% | 35% | 0% |
| 6 | 0 | 0 | 5 | 38 | 8,68% | 8,68% | 65% | 40% | 9% | 62% | 35% | 0% |
| 7 | 0 | 1 | 0 | 47 | 88,63% | 88,63% | 96% | 93% | 89% | 63% | 36% | 0% |
| 8 | 0 | 1 | 1 | 45 | 84,90% | 84,90% | 94% | 90% | 85% | 63% | 36% | 0% |
| 9 | 0 | 1 | 2 | 42 | 79,97% | 79,97% | 92% | 87% | 80% | 62% | 35% | 0% |
| 10 | 0 | 1 | 3 | 42 | 74,78% | 74,78% | 90% | 84% | 75% | 62% | 35% | 0% |
| 11 | 0 | 1 | 4 | 43 | 69,09% | 69,09% | 88% | 80% | 69% | 62% | 36% | 0% |
| 12 | 0 | 1 | 5 | 48 | 64,13% | 64,13% | 87% | 77% | 64% | 63% | 36% | 0% |
| 13 | 1 | 0 | 0 | 39 | 48,72% | 48,72% | 83% | 70% | 49% | 66% | 41% | 0% |
| 14 | 1 | 0 | 1 | 37 | 40,65% | 40,65% | 80% | 65% | 41% | 66% | 41% | 0% |
| 15 | 1 | 0 | 2 | 40 | 34,72% | 34,72% | 78% | 62% | 35% | 66% | 41% | 0% |
| 16 | 1 | 0 | 3 | 36 | 27,12% | 27,12% | 75% | 57% | 27% | 66% | 40% | 0% |
| 17 | 1 | 0 | 4 | 39 | 22,42% | 22,42% | 74% | 54% | 22% | 66% | 41% | 0% |
| 18 | 1 | 0 | 5 | 43 | 18,55% | 18,55% | 73% | 52% | 19% | 67% | 42% | 0% |
| 19 | 1 | 1 | 0 | 41 | 94,07% | 94,07% | 98% | 97% | 94% | 67% | 41% | 0% |
| 20 | 1 | 1 | 1 | 40 | 92,06% | 92,06% | 97% | 95% | 92% | 66% | 41% | 0% |
| 21 | 1 | 1 | 2 | 38 | 89,32% | 89,32% | 96% | 94% | 89% | 66% | 66% | 0% |
| 22 | 1 | 1 | 3 | 39 | 86,31% | 86,31% | 95% | 92% | 86% | 66% | 41% | 0% |
| 23 | 1 | 1 | 4 | 40 | 82,62% | 82,62% | 94% | 90% | 83% | 66% | 41% | 0% |
| 24 | 1 | 1 | 5 | 43 | 78,68% | 78,68% | 93% | 88% | 79% | 67% | 42% | 0% |

Fonte: Os Autores (2013).

## 5 CONCLUSÕES

A aplicação de técnicas estatísticas mais robustas na área de cobrança pode trazer ganhos financeiros consideráveis para as empresas que trabalham com grande volume de vendas à crédito.

Nesse trabalho, o modelo de regressão logística foi considerado como a técnica que é usualmente aplicada nos bancos de dados da área financeira/cobrança quando se tem interesse em modelar risco associado a pagamentos. Esse modelo serviu de referência para a comparação com o modelo de mistura com fração de fidelizados, considerado aqui como uma alternativa.

Ambos os modelos ajustados apresentaram ajustes satisfatórios aos dados analisados e se mostraram bastante eficientes na discriminação entre clientes bons e maus (falha/não falha). As covariáveis que apresentaram efeitos significativos foram: idade do cliente em anos, produto contratado, indicativo de parcelamento quebrado e quantidade de vezes em cobrança nos últimos seis meses.

No que diz respeito ao modelo de regressão logística, foi possível obter a partir deste estimativas da probabilidade de cada cliente se tornar mau pagador ao final do período de acompanhamento (150 dias), o que auxiliou na definição de alguns perfis de clientes que são de interesse da empresa de telecomunicações. Dentre estes perfis, podem ser citados os dos clientes que apresentaram maior e menor probabilidade de não pagamento, mostrados na Tabela 11.

Observa-se que o conhecimento desses perfis é relevante para a empresa no sentido de direcioná-la ao estabelecimento de estratégias adequadas de cobrança de acordo cada perfil, pois se a cobrança dos clientes inadimplentes for realizada no momento adequado e com a técnica apropriada, a empresa além de recuperar o montante emprestado à credito, consegue também manter seu relacionamento com o cliente, fidelizando-o.

Quanto ao modelo de mistura com fração de fidelizados, este forneceu, além das informações obtidas com o modelo de regressão logística, informações adicionais referente ao risco dos clientes se tornarem não fidelizados (maus pagadores) ao longo do período todo de acompanhamento, de acordo com seus perfis (características). Essas informações ao longo do tempo podem ser bastante relevante e úteis para a definição de estratégias de cobrança mais enérgicas e dinâmicas por parte da empresa em função dos perfis dos clientes. Por exemplo, um

cliente inadimplente com alta probabilidade de se tornar mau pagador em um tempo curto, estaria passível de ações mais enérgicas e rápidas por parte da empresa do que aquele que apresenta alta probabilidade de pagamento.

Nesse contexto, o modelo de mistura com fração de fidelizados mostrou-se eficiente na modelagem de dados na área financeira, fornecendo informações adicionais ao modelo de regressão logística que podem auxiliar a dinamizar as estratégias de cobrança. De modo geral, tal modelo pode ser sugerido como uma alternativa viável para as empresas de cobrança.

Para trabalhos futuros, sugere-se o acompanhamento dos indivíduos inadimplentes por um período de tempo mais extenso para que possam ser feitos testes de estabilidade da população estudada, bem como de suas variáveis. Além disso, pode-se também avaliar possíveis efeitos de sazonalidade ao longo do ano.

## REFERÊNCIAS


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

COLOSIMO, E. A; GIOLO, S. R. *Análise de Sobrevivência Aplicada.* São Paulo: Edgard Blucher, 2006.

CORBIÈRE, F.; JOLY, P. A SAS macro for parametric and semiparametric mixture cure models *Methods and Programs in Biomedicine*, v. 85, n. 2, p.173-80, 2007.

COX, D.R. Regression models and life-tables, *Journal of the Royal Statistical Society*. Series B. v. 34, p. 187-220. 1972. Disponível em: < http://hydra.usc.edu/ pm518b/literature/cox-72.pdf >. Acesso em: 11 Jun 2013.

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

GIOLO, S. R. *Introdução à Análise de Dados Categóricos com Aplicações.* 2012. Disponível em:<http://people.ufpr.br/~giolo/CE073/Material/Suely_Giolo.pdf>. Acesso em: 28 Abr. 2013.

GIOSA, L. A. *Terceirização: uma abordagem estratégica*. São Paulo: Pioneira, 1993.

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

IPEADATA, *Operações de crédito do sistema financeiro aos setores público e privado - recursos livres – inadimplência*. Disponível em < http://www.ipeadata. gov.br/ >. Acesso em: 20 de Jun. 2013.

LEMES JUNIOR, A.B.; CHEROBIM, A.P.M.S.; RIGO, C.M. *Administração Financeira: princípios, fundamentos e práticas brasileiras.* 2.ed., Rio de Janeiro: Campus, 2005.

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

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

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

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

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

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

SAS/STAT© Software: Enterprise Guide, Release 4.3. Copyright, SAS Institute Inc. Cary, NC, USA, 2006.

SILVA, J. P. *Gestão e análise de risco de crédito*. 5. ed., São Paulo: Editora Atlas S.A, 2006.

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

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

# APÊNDICE A

## A1 Comandos SAS - ajuste do modelo de regressão logística.

```
ODS GRAPHICS ON;
title 'Modelo completo';
proc logistic data=TCC.BASE_AGRUP_VLR;
class PARCELAMENTO_QUEBRADO PRODUTO/param=ref;
model MAU(event='1')=  IDADE PRODUTO PARCELAMENTO_QUEBRADO QT_COB
        / selection=stepwise OUTROC=ROC lackfit influence
;
run;
```

## A2 Comandos SAS - obtenção do estimador não paramétrico de Kaplan-Meier.

```
PROC LIFETEST DATA=TCC.BASE_AGRUP_VLR
        ALPHA=0.05
        PLOTS(ONLY)=SURVIVAL( STRATA=UNPACK );
        TIME max_tm * bom (1);
RUN;TITLE;
```

## A3 Comandos SAS - ajuste do modelo de mistura com fração de fidelizados.

Obs: inicialmente se faz necessário carregar a macro PSPMCM

http://www.isped.u-bordeaux2.fr/recherche/biostats/FR-biostats-accueil.htm

```
%pspmcm(DATA=TCC.BASE_FINAL2,
                                ID=EXTERNAL_ID,
                                CENSCOD=MAU,
                                TIME=TEMPO,
                                VAR=  PRODUTO_NUM(I S,0)
                                        QT_COB(I S,0)
                                        PARCELAMENTO_QUEBRADO_NUM(I S,0)
                                        IDADE (I S,42),
                                        DIA_VENC (I S,1)
                                INCPART=logit,
                                SURVPART=Cox,
                                TAIL=zero , SU0MET=pl,
                                FAST=Y,BOOTSTRAP=N,
                                MAXITER=200,CONVCRIT=1e-5, ALPHA=0.05,
                                BASELINE=Y,
                                GESTIMATE=Y,
        SPLOT=Y, PLOTFIT=Y);    run;
```