Estratégias para o desenvolvimento de modelos de *credit score* com inferência de rejeitados

Mauro Correia Alves

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

Programa: Estatística

Orientador: Profa. Dra. Lúcia Pereira Barroso

São Paulo, setembro de 2008

Estratégias para o desenvolvimento de modelos de *credit score* com inferência de rejeitados

Este exemplar corresponde à redação final
da dissertação devidamente corrigida e
defendida por Mauro Correia Alves sendo
aprovada pela comissão julgadora.

São Paulo, 08 de setembro de 2008

Banca Examinadora:

- Profa. Dra. Lúcia Pereira Barroso (orientadora) - IME/USP.
- Prof. Dr. Antonio Carlos Pedroso de Lima - IME/USP.
- Prof. Dr. Francisco José Espósito Aranha Filho - FGV/SP.

"O verdadeiro mestre é aquele que ensina o que aprende e aprende com o que ensina."

Cora Coralina

Agradecimentos

Agradeço primeiramente a Deus por me guiar e poder concluir mais uma importante etapa da minha vida, me concedendo forças e dedicação para conciliar os estudos do mestrado, minha vida profissional e ainda sobrar algum tempo para minha família, agradeço também as seguintes pessoas:

À Professora Lúcia, que em todas as etapas da elaboração desta dissertação esteve sempre disposta a dar sugestões, contribuindo significativamente para a realização desta pesquisa.

Aos meus pais, que me mostraram desde cedo a importância dos estudos na vida de uma pessoa.

À minha esposa Silvia que compartilhou comigo as minhas dificuldades, que está comigo nos momentos mais difíceis e me dá forças para não desistir.

Ao meu filho Gabriel pela paciência e compreensão da espera por um tempo livre para poder bincar com ele, que muitas vezes ficou com a sua vó Maria Anésia, a quem eu tenho eterna gratidão pela tranqüilidade que me passa.

À Tia Eulina ("Nininha") e ao Tio Mansur pelas revisões e sugestões no texto.

À Universidade de São Paulo e ao Instituto de Matemática e Estatística pela oportunidade concedida de aperfeiçoar meus estudos.

À Serasa pela disponibilização dos dados do seu *bureau* de crédito.

Ao Marcelo e à Noemy, que permitiram o uso de parte dos dados nesta pesquisa e pelo apoio de todos colegas de trabalho.

Aos Professores Antonio Carlos e Francisco Aranha pelas sugestões finais, enriquecendo ainda mais este trabalho.

E a todos que de alguma forma contribuíram e me incentivaram pela conclusão deste trabalho.

Resumo

Modelos de *credit score* são usualmente desenvolvidos somente com informações dos proponentes aceitos. Neste trabalho foram consideradas estratégias que podem ser utilizadas para o desenvolvimento de modelos de *credit score* com a inclusão das informações dos rejeitados. Foram avaliadas as seguintes técnicas de *inferência de rejeitados*: classificação dos rejeitados como clientes *Maus*, parcelamento, dados aumentados, uso de informações de mercado e ainda a estratégia de aceitar proponentes rejeitados para acompanhamento e desenvolvimento de novos modelos de risco de crédito. Para a avaliação e comparação dos modelos foram utilizadas as medidas de desempenho: estatística de Kolmogorov-Smirnov (KS), área sob a curva de Lorentz (ROC), área entre as curvas de distribuição acumulada dos escores (AEC), diferença entre as taxas de inadimplência nos intervalos do escore definidos pelos decis e coeficiente de Gini. Concluiu-se que dentre as quatro primeiras técnicas avaliadas, a quarta (uso de informações de mercado) foi a que apresentou melhor desempenho. Quanto à estratégia de aceitar proponentes rejeitados, observou-se que há um ganho em relação ao modelo ajustado só com base nos proponentes aceitos.

Palavras-chave: *credit score*, risco de crédito, *inferência de rejeitados*, regressão logística.

Abstract

Credit scoring models are usually built using only information of accepted applicants. This text considered strategies that can be used to develop credit score models with inclusion of the information of the rejects. We evaluated the techniques of reject inference: classification of rejected customers as bad, parceling, augmentation, use of market information and the strategy of accepting rejected proponents for monitoring and developing new models of credit risk. For the evaluation and comparison between models were used performance measures: Kolmogorov-Smirnov statistics (KS), the area under the Lorentz Curve (ROC), area between cumulative distribution curves of the scores (AEC), difference among the delinquency rate in the score buckets based on deciles (DTI) and the Gini coefficient. We concluded that among the first four techniques evaluated, the fourth (use of market information) had the best performance. For the strategy to accept rejected bidders, it was observed that there is a gain in relation to the model that uses only information of accepted applicants.

Keywords: *credit score, credit risk, reject inference, logistic regression.*

Sumário

1 Introdução **9**

1.1 Apresentação do Problema 9

1.2 Objetivos Principais 11

2 Descrição do Estudo **12**

2.1 Descrição dos Dados 12

2.2 Definição dos Rejeitados 14

2.3 Análise Descritiva 15

3 Metodologia **18**

3.1 Regressão Logística 18

3.2 Métodos de Inferência de Rejeitados 22

3.2.1 Classificação dos Rejeitados como Clientes *Maus* 23

3.2.2 Método de Parcelamento 23

3.2.3 Método de Dados Aumentados 25

3.2.4 Uso de Informações de Mercado 27

4 Avaliação de Modelos de *Credit Score* **29**

4.1 Estatística de Kolmogorov-Smirnov (KS) 29

4.2 Área Entre Curvas - AEC 32

4.3 Curva ROC e Coeficiente de Gini 32

4.4 Diferença entre Taxas de Inadimplência (DTI) 36

5 Resultados para a Aplicação 1 **38**

5.1 Distribuição da Base de Dados 39

5.2 Categorização das Variáveis Explicativas 40

5.3 Processo de Seleção das Variáveis 41

5.4 Ajuste dos Modelos 42

5.5 Comparação dos Resultados 43

6 Resultados para a Aplicação 2 **47**

6.1 Entendendo o Problema 48

6.2 Análise Descritiva 49

6.3 Seleção das Variáveis e Ajuste dos Modelos 52

6.4 Comparação dos Resultados 52

7 Conclusões e Considerações Finais **55**

A Descrição das Variáveis **58**

A.1 Aplicação 1 58

A.2 Aplicação 2 60

B Tabelas Descritivas **61**

C Parâmetros Estimados **69**

D Macros SAS **77**

E Gráficos **82**

Referências Bibliográficas **90**

Capítulo 1

Introdução

1.1 Apresentação do Problema

Instituições financeiras de muitos países, entre eles o Brasil, estão intensificando e aperfeiçoando metodologias de estudos sobre práticas que auxiliam no controle da mitigação de riscos, com o objetivo de se adequarem ao requerimento de capital regulamentar aos níveis de riscos associados às operações financeiras, regras essas instituídas pelo Novo Acordo de Basiléia II. Esse acordo foi assinado pelo Comitê de Supervisão Bancária da Basiléia - *Basel Committee on Banking Supervision*, divulgado em junho de 2004 e estabelece 25 princípios essenciais para uma supervisão bancária eficaz. Essas medidas auxiliam no acompanhamento de padrões internacionais de regulação e fiscalização do sistema financeiro, tendo como resultado o incentivo à adoção de melhores práticas bancárias e o fortalecimento do mercado financeiro brasileiro. Nesse contexto, as instituições financeiras sentem-se cada vez mais obrigadas a desenvolver sistemas eficientes de avaliação de risco, porém muitas não se preocupam com os proponentes rejeitados em uma operação de crédito.

A análise de *inferência de rejeitados* é uma prática presente em instituições financeiras, nas atividades de concessão de crédito. A necessidade dessa análise está ligada ao fato de os modelos de crédito, geralmente conhecidos como modelos de

credit score, serem desenvolvidos apenas sobre o histórico dos proponentes aceitos com base em um modelo inicial (clientes aptos ao crédito), sendo que a amostra utilizada para o desenvolvimento de novos modelos é sistematicamente viesada, pois pode-se deixar de avaliar alguma característica específica, que esteja particularmente presente apenas nos proponentes rejeitados, fazendo com que o novo modelo de *credit score* desenvolvido não consiga predizer de forma adequada o comportamento desses indivíduos.

Modelos estatísticos de *credit score* refletem a qualidade das amostras com as quais foram desenvolvidos, prevendo o comportamento de novos proponentes da população estudada, baseando-se no desempenho dos clientes anteriores utilizados na amostra de desenvolvimento (Thomas et al., 2002).

O viés acontece se alguma das características que fez com que um determinado proponente fosse rejeitado no passado, não estiver presente na base dos clientes aceitos, fazendo assim com que o novo modelo de crédito desenvolvido não represente a população de novos proponentes. Por exemplo, imaginemos que um modelo ou um filtro de crédito de uma determinada instituição rejeite proponentes que possuam títulos protestados em cartório. Nesse caso a base de clientes aceitos, utilizada para extrair futuramente a amostra de desenvolvimento do novo modelo, dificilmente conterá pessoas que possuam essa característica. A amostra de desenvolvimento poderá não refletir uma quantidade suficiente de informações para realizar uma nova predição de como estes proponentes se comportariam, caso tivessem sido aceitos.

Um dos propósitos da *inferência de rejeitados* é a possibilidade de considerar características que foram desprezadas, por pertecerem apenas aos proponentes rejeitados por um processo ou modelo de crédito previamente utilizado na etapa de concessão do crédito, reduzindo assim o viés do novo modelo a ser construído.

Na literatura existem algumas publicações que avaliam de modo empírico as técnicas de *inferência de rejeitados* em *credit score*. As técnicas de *extrapolation e augmentation*, aqui tratada como *dados aumentados*, foram exploradas inicialmente

por Hsai (1978), depois por Hand e Henley (1993) e por Banasik e Crook (2005). Dempster et al. (1977) utilizam o algoritmo EM para a estimação de máxima verossimilhança a partir do tratamento dos rejeitados como dados imcompletos; Reichert et al. (1983) o modelo multinomial; Joanes (1993) propõe a reclassificação iterativa; Ash e Meester (2002) o *parceling* e Feelders (2000) considera a *inferência de rejeitados* como um problema de dados ausentes. Chen e Huang (2003) avaliam algumas técnicas computacionais; Sohn e Shin (2006) utilizam técnica de análise de sobrevivência, apresentando um método de *inferência de rejeitados* baseado no intervalo de confiança mediano do tempo de sobrevida para os clientes inadimplentes.

1.2 Objetivos Principais

Este estudo pretende contribuir para a difusão e maior esclarecimento do uso de técnicas de *inferência de rejeitados* com a construção de modelos de *credit score*, além de estimular a discussão do seu uso neste tipo de modelos e a possibilidade de aprimoramento das mesmas.

O objetivo é ilustrar algumas metodologias sistemáticas de tratamento de rejeitados e avaliar o desempenho das técnicas mais utilizadas no mercado, investigando e apontando possíveis diferenças entre elas.

O trabalho foi desenvolvido na seguinte seqüência. No Capítulo 2 apresentamos a descrição dos dados utilizados e algumas definições. Os métodos de *inferência de rejeitados* são descritos no Capítulo 3 e as medidas de avaliação dos modelos no Capítulo 4. Os Capítulos 5 e 6 mostram os resultados de duas aplicações, seguidos pelas conclusões e considerações finais no Capítulo 7.

Capítulo 2

Descrição do Estudo

2.1 Descrição dos Dados

Neste trabalho são utilizadas duas bases de dados; a primeira, descrita neste capítulo, teve a aplicação das técnicas de *inferência de rejeitados* de parcelamento, dados aumentados e utilização de informações de mercado, cujos resultados são apresentados no Capítulo 5; a segunda é descrita no Capítulo 6, com aplicação da técnica de *inferência de rejeitados* de aceitar uma amostra de rejeitados.

A primeira base de dados utilizada nesta dissertação foi cordialmente cedida pela Serasa, instituição caracterizada principalmente como central de informação de crédito que oferece soluções e auxilia na decisão de crédito no Brasil e na América Latina. Seu banco de dados contém informações sobre pessoas, empresas e grupos econômicos, reunindo dados cadastrais, compromissos e hábitos de pagamento. As centrais de informações de crédito coletam e armazenam informações referentes às pessoas físicas e jurídicas com suas devidas autorizações e também coletam informações de domínios públicos, como protestos, falências, concordatas, ações judiciais executivas, geralmente disponíveis em cartórios e juntas comerciais. Além disso, mantém acesso à base de CCF (serviço de cadastro de cheques sem fundos do Banco Central).

Os indivíduos contidos na base de estudo referem-se apenas às pessoas físicas, sendo a identidade de cada indivíduo preservada pela omissão do número do CPF (Cadastro de Pessoa Física). A população envolvida neste estudo contém indivíduos que tiveram consultas ao crédito e/ou cheques, em instituições bancárias, financeiras e no comércio em geral, em algum tipo de transação comercial no qual o risco de inadimplência esteve envolvido. Em geral, estas transações envolvem operações como: crédito pessoal, crédito direto ao consumidor, financiamento de bens de consumo, aquisição de cartões de crédito, renovação de cheques especiais, entre outras. As consultas foram efetuadas no período de janeiro a dezembro de 2004. Ao todo foram selecionadas aleatoriamente 50.000 consultas dentre todas as consultas efetivadas no período de um ano, 30.000 foram utilizadas para o desenvolvimento dos modelos e 20.000 para a validação dos mesmos.

Para cada indivíduo, observaram-se variáveis comportamentais de mercado, tais como: registro sobre cheques sem fundos, títulos protestados, pendências financeiras adquiridas em instituições financeiras, bancárias e/ou no comércio em geral. Foram utilizadas também as variáveis de referência bancária e *Credit Target Serasa*®, modelo de segmentação de crédito da Serasa que indica o nível e a freqüência da atividade de crédito de pessoas físicas (alta, média, baixa), com entrada recente ou não no mercado; número de consultas crescente ou decrescente, e ainda classifica os inativos em crédito e/ou cheques.

Além das variáveis relacionadas ao comportamento e atividade de crédito, foram observadas também as variáveis sócio-demográficas que estavam disponíveis na base de dados. O detalhamento de todas as variáveis utilizadas neste estudo é integralmente descrito no Apêndice A.1.

A segunda base de dados utilizada foi obtida a partir de uma empresa de TV por assinatura, contendo dados sobre propostas de clientes que solicitaram o serviço de TV em um determinado período. Essa base de dados contém também uma parcela de proponentes que seriam rejeitados por um filtro de crédito, porém, estes foram aceitos, a fim de acompanhar o seu desempenho junto à empresa e com

o objetivo de utilizá-los no desenvolvimento de novos modelos de *credit score*. O detalhamento das variáveis que foram analisadas está descrito no Apêndice A.2.

A primeira base de dados será denominada Aplicação 1, sobre a qual foram avaliadas as técnicas de *inferência de rejeitados*: classificação dos clientes *Maus* como rejeitados; parcelamento; dados aumentados e o uso de informações de mercado. A segunda base de dados será denominada Aplicação 2, sobre a qual foi avaliada a estratégia de *inferência de rejeitados* de aceitar uma amostra de rejeitados.

2.2 Definição dos Rejeitados

Os proponentes rejeitados em crédito são gerados a partir de regras específicas de decisão, estabelecidas pela política de crédito de cada instituição, que é ajustada principalmente em função da taxa de aprovação e de inadimplência esperada. Procura-se sempre a calibração destes parâmetros de forma a maximizar a rentabilidade das carteiras de crédito.

Neste estudo desenvolvemos modelos de *credit score*, utilizando as informações dos clientes rejeitados em duas situações diferentes: proponentes rejeitados a partir de um conjunto de filtros de créditos (situação em que uma instituição ainda não possui um modelo de *credit score*) e clientes rejeitados a partir de um modelo de *credit score* já implantado. Particularmente, usamos as seguintes definições:

1. Sem o uso de modelo de crédito: o cliente é rejeitado caso possua mais que duas restrições de crédito resolvidas nos últimos seis meses e/ou qualquer outra restrição não resolvida até a data de referência da consulta efetuada.

2. Com o modelo de crédito já implantado: a decisão é obtida diretamente da pontuação gerada e atribuída ao ponto de corte. Neste estudo usamos o modelo *Credit Bureau Score*® da Serasa, um modelo de *credit score* de mercado, em que é gerada uma pontuação que varia de 0 a 1000, que fornece a probabilidade

do cliente se tornar inadimplente em um horizonte de 12 meses; quanto menor a sua pontuação, maior o risco associado à inadimplência. Especificamente para este estudo, consideramos um cliente rejeitado se sua pontuação foi inferior a 500.

O grupo de proponentes aceitos foi observado seis meses depois e seu desempenho foi avaliado, sendo classificado em uma de duas categorias da variável resposta, definida como "conceito de inadimplência": *Bom* e *Mau*. A classificação desta variável se dá por conta da capacidade de o indivíduo honrar seus compromissos, ou seja, a classificação se dá pela presença de algum tipo de restrição financeira ativa (dívida não resolvida) de mercado, presente no período de avaliação.

2.3 Análise Descritiva

Cada variável foi analisada individualmente, observando-se a sua freqüência em relação à variável resposta e observadas as suas respectivas freqüências, comparando-as com o total observado na amostra. Para cada variável foi calculado o risco relativo (razão de riscos entre clientes *Bons* e *Maus*) como medida de associação atribuída ao risco de inadimplência. O cálculo do risco relativo está exemplificado na Tabela 2.1.

em que

$$Bons_k(\%) = \frac{b_k}{b} \times 100, \quad Maus_k(\%) = \frac{m_k}{m} \times 100 \quad \text{e} \quad RR_k = \frac{Bons_k(\%)}{Maus_k(\%)}$$

para $k = 1, 2, 3, ..., K$;

b_k: número de clientes *Bons* na k-ésima categoria;

m_k: número de clientes *Maus* na k-ésima categoria;

b : total de clientes *Bons* observados na variável;

m : total de clientes *Maus* observados na variável;

RR_k: risco relativo de um cliente *Bom* presente na k-ésima categoria em relação a um cliente *Mau*.

Tabela 2.1: Exemplo do cálculo do risco relativo.

Variável	*Bons*	*Maus*	*Bons*	*Maus*	RR
Categoria 1	b_1	m_1	b_1/b	m_1/m	$(b_1/b)/(m_1/m)$
Categoria 2	b_2	m_2	b_2/b	m_2/m	$(b_2/b)/(m_2/m)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Categoria k	b_k	m_k	b_k/b	m_k/m	$(b_k/b)/(m_k/m)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Categoria K	b_K	m_K	b_K/b	m_K/m	$(b_K/b)/(m_K/m)$
Total	b	m	1	1	1

Segundo Rosa (2000), o risco relativo é uma medida descritiva que auxilia na identificação de categorias com alto ou baixo poder de discriminação, além de identificar os níveis das variáveis que discriminam melhor os *Maus* dos *Bons* clientes.

A partir do RR, conforme Rosa (2000), podemos avaliar as categorias das variáveis da seguinte maneira:

- $RR_k = 1$: significa que se a variável assumir a k-ésima categoria, não há indícios de o cliente ser de maior ou menor risco comparado à análise desconsiderando essa variável;

- $RR_k < 1$: significa que quanto menor o risco relativo, maior é a probabilidade de o cliente apresentar menores riscos de inadimplência, indicando que a categoria k apresenta algum poder para discriminar clientes *Bons*;

- $RR_k > 1$: significa que quanto maior o risco relativo, maior é a probabilidade de o cliente apresentar maiores riscos de inadimplência, indicando que a categoria k apresenta algum poder para discriminar clientes *Maus*.

Ainda segundo Rosa (2000), outra vantagem de usar o RR é a possibilidade de agrupamentos de categorias com valores de RR semelhantes, desde que haja

coerência nos agrupamentos e que os mesmos façam sentido quanto à dinâmica do negócio de crédito.

Após categorizar cada variável, para construção de variáveis *dummies* (Neter et al., 1996), utilizou-se casela de referência, adotando-se como regra para a indicação da mesma, a categoria em que o risco relativo (RR) ficasse mais próximo de um, indicando que esta categoria tem um efeito neutro na discriminação de *Bons* e *Maus* clientes.

Na Tabela 2.2 temos um exemplo de construção das variáveis *dummies* associadas à variável quantidade de bancos (quantidade de bancos diferentes em que o cliente possui conta).

Tabela 2.2: Exemplo de construção de variáveis *dummies*.

Variável	Variáveis *dummies*			
Quantidade de bancos	$dummy1$	$dummy2$	$dummy3$	$dummy4$
Não possui conta bancária	1	0	0	0
Um banco	0	0	0	0
Dois bancos	0	0	1	0
Acima de dois bancos	0	0	0	1

A análise descritiva das variáveis das Aplicações 1 e 2 encontram-se no Apêndice B.

Capítulo 3

Metodologia

Existem várias técnicas de *inferência de rejeitados* utilizadas no desenvolvimento de modelos de *credit score*. Nesta dissertação são abordadas quatro das mais conhecidas e freqüentemente utilizadas: classificação dos rejeitados como *Maus* clientes, parcelamento, dados aumentados, além do auxílio do uso de informações de mercado, sendo que esta última pode ser apresentada como uma alternativa às demais.

Para o desenvolvimento dos modelos, é utilizada a regressão logística, modelo estatístico que já é amplamente difundido e aceito pelo mercado para o desenvolvimento dos modelos de *credit score* (Pereira, 2004). Mais detalhes sobre o seu processo de desenvolvimento podem ser encontrados em Sicsu (1998a, 1998b).

3.1 Regressão Logística

Desde a década de 50, a regressão logística é conhecida, porém torna-se mais difundida a partir da década de 80, com Cox e Snell (1989) e Hosmer e Lemeshow (2000). Aspectos teóricos do modelo de regressão logística são amplamente discutidos, destacando-se os textos de Agresti (1990) e Kleinbaum (1994).

A regressão logística se enquadra na classe de métodos estatísticos multivariados de estudo de dependência, pois procura relacionar um conjunto de variáveis

independentes com uma variável dependente categórica (Morgan e Griego, 1998). Pode ser usada com objetivo descritivo, quando deseja-se descrever a natureza do relacionamento entre a resposta média (probabilidade de ocorrência de um evento) e uma ou mais variáveis regressoras, ou então, com objetivo preditivo, quando deseja-se prever se um determinado evento ocorrerá em um prazo pré-definido, dado um conjunto de variáveis explicativas.

Uma das maiores vantagens do uso da regressão logística, aplicada em muitos problemas práticos, é a não exigência de algumas suposições, como normalidade dos erros e igualdade de matrizes de covariância, além da possibilidade de interpretação direta dos coeficientes estimados como medidas de associação.

Segundo Hosmer e Lemeshow (2000), o objetivo da regressão logística é modelar a relação entre a variável reposta e um conjunto de variáveis preditivas. O modelo final será aquele que apresentar melhor ajuste, ou seja, melhores medidas de desempenho e for naturalmente mais razoável de se explicar.

Considere o caso em que desejamos classificar os clientes em dois grupos (*Bons* e *Maus*), segundo o seu comportamento de crédito; deste modo, considere a variável resposta $Y \in \{0, 1\}$, definida como:

$$Y_i = \begin{cases} 1 & \text{, se o } i\text{-ésimo cliente é } Mau \text{ (sucesso)} \\ 0 & \text{, se o } i\text{-ésimo cliente é } Bom \text{ (fracasso)} \end{cases}$$

Considere um conjunto com $p-1$ variáveis independentes, representadas por $\boldsymbol{x_i} = (1, x_{i2}, x_{i3}, \ldots, x_{ip})^\top$, o vetor da i-ésima linha da matriz $\boldsymbol{X}$ das variáveis explicativas associado ao i-ésimo indivíduo da amostra. O elemento x_{ij} da matriz $\boldsymbol{X}$ corresponde ao ij-ésimo componente, onde $i = 1, 2, ..., n$, $j = 1, 2, ..., p$ e $x_{i1} = 1$. Denotamos por $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ o vetor de parâmetros desconhecidos, sendo β_j o parâmetro associado a x_{ij}.

No modelo de regressão logística múltipla a probabilidade de sucesso é ex-

pressa por:

$$\begin{aligned} \pi(\boldsymbol{x_i}) = P(Y_i = 1|\boldsymbol{x_i}) &= \frac{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})} \\ &= \frac{\exp(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})} \end{aligned} \tag{3.1.1}$$

e a probabilidade de fracasso é dada por:

$$\begin{aligned} 1 - \pi(\boldsymbol{x_i}) = P(Y_i = 0|\boldsymbol{x_i}) &= \frac{1}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})} \\ &= \frac{1}{1 + \exp(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}, \end{aligned}$$

sendo $\pi(\boldsymbol{x_i})$ a probabilidade condicional da observação $Y_i = 1, i = 1...n$, dados os valores das variáveis explicativas $\boldsymbol{x_i}$.

A função de verossimilhança pode ser escrita da seguinte maneira:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi(\boldsymbol{x_i})]^{y_i} [1 - \pi(\boldsymbol{x_i})]^{1-y_i}.$$

Assumindo que Y_i tem distribuição de Bernoulli(π_i), sendo $\pi_i = \pi(\boldsymbol{x_i})$ e usando a função de ligação *logit* dada por:

$$ln\left[\frac{\pi_i}{1 - \pi_i}\right] = \beta_1 + \sum_{j=2}^{p} \beta_j x_{ij},$$

é mais simples trabalhar com a função de log-verossimilhança, ou seja, o logaritmo da função de verossimilhança, escrita como:

$$l(\boldsymbol{\beta}) = ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i(\boldsymbol{x_i}^\top \boldsymbol{\beta}) - \sum_{i=1}^{n} ln[1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})].$$

Para a estimação dos parâmetros pode-se utilizar o método de máxima verossimilhança, encontrando o valor de $\boldsymbol{\beta}$ que maximiza $l(\boldsymbol{\beta})$. Para este processo pode-se utilizar o método numérico iterativo de Newton-Raphson (Cox, 1975), sendo necessário derivarmos $l(\boldsymbol{\beta})$ em relação a cada parâmetro.

Derivando $l(\boldsymbol{\beta})$ em relação ao j-ésimo parâmetro, temos:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n}[y_i x_{ij} - \frac{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x_i}\boldsymbol{\beta})}x_{ij}] \\ &= \sum_{i=1}^{n}[y_i - \pi_i]x_{ij}\ .\end{aligned}$$

Seja a matriz de informação de Fisher (Bonfarine e Sandoval, 2002) dada por:

$$I_F(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right] = \boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X},$$

em que $\boldsymbol{Q} = \text{diag}[\pi_i(1-\pi_i)]$ e $\boldsymbol{X}$ é a matriz de variáveis independentes. Logo, $[I_F(\boldsymbol{\beta})]^{-1}$ é assintoticamente a matriz de variância e covariância dos estimadores de máxima verossimilhança dos parâmetros, sendo que a distribuição assintótica de $\widehat{\boldsymbol{\beta}}$ é $N_p(\boldsymbol{\beta}, I_F(\boldsymbol{\beta})^{-1})$.

É possível construir intervalos de confiança para os parâmetros estimados através da estatística de Wald, geralmente conhecidos por intervalos de confiança normais. Eles são baseados na normalidade assintótica dos estimadores de máxima verossimilhança (Dobson, 1983).

O intervalo de confiança de Wald $100(1-\alpha)\%$ para β_j é dado por:

$$\widehat{\beta}_j \pm z_{(1-\alpha/2)}\widehat{\sigma}_j,\ \ \forall\ j = 1, 2, ..., p\ ,$$

onde $z_{(1-\alpha/2)}$ é o $\left(1-\frac{\alpha}{2}\right)$-ésimo percentil da distribuição normal padrão, $\widehat{\beta}_j$ é o estimador de máxima verossimilhança de β_j e $\widehat{\sigma}_j$ é o estimador do erro padrão de $\widehat{\beta}_j$.

Existem vários aplicativos estatísticos que disponibilizam a técnica de regressão logística. Nesta dissertação usamos o aplicativo SAS, com o qual é possível ajustar modelos de regressão através dos procedimentos *Catmod*, *Reg Logistic* e *Probit*. Usamos neste trabalho a *Proc Reg Logistic* (Allison, 1999). Após o ajuste do modelo com a estimação dos parâmetros β_j, usando a expressão (3.1.1) estimamos a probabilidade de um indivíduo se tornar inadimplente em determinado período

(definido pelo conceito de inadimplência). Para representar esta probabilidade em forma de um escore S, multiplicamos este número por 1000, sendo que quanto menor este escore, maior será a probabilidade do indivíduo se tornar inadimplente.

A partir do modelo logístico ajustado, é calculado o escore S para cada indivíduo, que corresponde à probabilidade do indivíduo se tornar inadimplente, e que será classificado como *Mau*, se $S < Pc$ e como *Bom*, caso contrário, sendo Pc um número real definido como ponto de corte, geralmente igual a 0,5 nos aplicativos estatísticos. Na prática, porém, o ponto de corte é ajustado de acordo com a política de crédito de cada instituição, sendo basicamente definida a partir da taxa de aprovação e o nível de inadimplência aceitável.

Com o modelo ajustado e com a classificação de cada indivíduo, podemos avaliar a capacidade de acertividade, ou seja, a capacidade de o modelo classificar um indivíduo como *Mau*, dado que realmente ele é *Mau* e classificar como *Bom*, dado que realmente é *Bom*. Estamos interessados em modelos com alta capacidade de acertividade e que possam ser sintéticos, de modo a facilitar a interpretação do modelo final ajustado. Mais detalhes sobre as medidas e as formas de avaliação de modelos de *credit score* são discutidos no Capítulo 4.

3.2 Métodos de Inferência de Rejeitados

A *inferência de rejeitados* é o processo de estimar o risco de inadimplência dos indivíduos que foram rejeitados em uma operação de crédito. Ao desenvolver modelos de *credit score* queremos que este represente o comportamento de todos os proponentes ao crédito, contudo, tipicamente os modelos são desenvolvidos apenas com informações comportamentais históricas dos clientes aceitos, pois o comportamento dos clientes rejeitados é desconhecido.

Existem várias técnicas que utilizam os rejeitados no desenvolvimento de modelos de *credit score*. Entre elas, estão as mais citadas na literatura, como: a classificação dos rejeitados como clientes *Maus*, parcelamento e dados aumentados,

uso de informações de mercado de *bureau de crédito* e a estratégia de aceitar uma amostra de rejeitados.

3.2.1 Classificação dos Rejeitados como Clientes *Maus*

Uma das maneiras mais simples de tratamento dos rejeitados é supor que todos os proponentes negados ao crédito sejam classificados como clientes *Maus*, ou seja, se fossem aceitos, todos se tornariam inadimplentes. Assim, a amostra de desenvolvimento do novo modelo será composta por clientes aceitos (*Bons* e *Maus*) acrescidas dos proponentes rejeitados, todos classificados como clientes *Maus*.

Ao assumir que todos os rejeitados são considerados como inadimplentes, estamos atribuindo a estes proponentes uma "*certeza probabilística*" que, na realidade, não existe e poderá causar um viés no novo modelo a ser desenvolvido. Certamente, o uso desta metodologia penaliza indivíduos que possuam características semelhantes aos dos proponentes rejeitados, classificados a priori como *Maus* clientes, levando muito provavelmente à sua rejeição em modelos ajustados no futuro.

3.2.2 Método de Parcelamento

Ash e Meester (2002) apresentam o método de *inferência de rejeitados* chamado *Parceling*, caracterizado como um processo de reclassificação por risco. Consiste em uma segmentação parcelada da população dos rejeitados, divididos entre *Bons* e *Maus*, segundo o risco do comportamento dos clientes aprovados, utilizando-se as taxas de inadimplência observadas. Este método é aplicado quando existe um modelo de *credit score* já em operação.

Geralmente para o desenvolvimento do modelo, na prática, aconselha-se utilizar uma amostra com clientes *Maus*, *Bons* e proponentes rejeitados, nas mesmas proporções do total de solicitantes de crédito. Para cada faixa de escore, é feito um parcelamento aleatório dos rejeitados, com base na freqüência observada de *Bons* e *Maus*, presentes na população dos aceitos.

Na Tabela 3.1, temos um exemplo da utilização desse método, com a distribuição de risco de um modelo de *credit score* já em operação e aplicado depois de um tempo de acompanhamento sobre uma amostra de clientes aceitos sobre os proponentes rejeitados anteriormente. Na faixa de 0 a 200, temos 100% de *Maus*, então os 355 proponentes rejeitados serão classificados como clientes *Maus*. Já na faixa de 201 a 300, temos 69% de *Maus* e 31% de *Bons*; nesta faixa os rejeitados (262) serão particionados aleatoriamente em 181 *Maus* e 81 *Bons* e assim por diante, repetindo-se o processo para as demais faixas de escore.

Tabela 3.1: Distribuição de risco dos aceitos e particionamento dos rejeitados.

Aceitos			Rejeitados		
Faixas de escore	*Maus*	*Bons*	Rejeitados	$Maus_1$	$Bons_1$
0 - 200	521 (100%)	0 (0%)	355	355	0
201 - 300	152 (69%)	68 (31%)	262	181	81
301 - 400	112 (24%)	355 (76%)	185	44	141
401 - 500	125 (20%)	512 (80%)	201	39	162
500 - 1000	85 (11%)	685 (89%)	125	14	111

O novo modelo de *credit score* é desenvolvido a partir da nova base de dados redistribuída "*parcelada*", ou seja, com todos os proponentes rejeitados particionados como *Bons* e *Maus* clientes e adicionados à base inicial de clientes aceitos.

Este método apresenta uma limitação quanto ao seu uso, pois ele só pode ser aplicado nos casos em que a instituição já tem um modelo de *credit score* em produção, uma vez que para efetuar a reclassificação dos proponentes rejeitados é preciso saber a inadimplência dos clientes aceitos por faixa de escore. Uma alternativa para a utilização deste método, na ausência de um modelo de *credit score* é efetuar a reclassificação dos rejeitados de modo aleatório a partir da taxa de inadimplência total observada para os clientes aceitos.

3.2.3 Método de Dados Aumentados

O método de dados aumentados geralmente é utilizado quando o processo de análise de risco de crédito é feito apenas por filtros estabelecidos pelas regras de negócio, resumidos em uma política de aprovação. Esse método utiliza uma ponderação entre aceitos e rejeitados, obtida a partir da classificação de um modelo construído inicialmente com todos os proponentes - com resposta de interesse *aceita* ou *rejeita*. Em seguida constrói-se um novo modelo ponderado, atribuindo-se um peso para cada indivíduo da população dos aceitos, segundo sua classificação na tabela de distribuição de risco do modelo de aceitação inicialmente construído.

Joanes (1993) foi pioneiro neste método, com a utilização da técnica de análise discriminante para efetuar as classificações. Posteriormente Banasik e Crook (2005) utilizaram a técnica de regressão logística.

Inicialmente, com uma amostra nas mesmas proporções do total de solicitantes de crédito, construímos um modelo cuja variável de interesse é a resposta da decisão de crédito (aceito ou rejeitado), com o intuito de estimar a probabilidade dos proponentes serem aceitos ou rejeitados no processo de crédito, em um determinado período de tempo.

A construção do modelo AR (aceita ou rejeita) tem como objetivo a identificação de proponentes rejeitados freqüentemente e, por seu intermédio, pretende-se detectar proponentes aceitos com perfis semelhantes aos de rejeitados. A identificação desses casos é representada pelo aumento de sua influência no modelo, por intermédio de um peso atribuído para cada indivíduo.

Na Tabela 3.2, tem-se a descrição de um exemplo de como obter a ponderação por faixas de escore, a partir da distribuição de risco dos clientes aceitos e rejeitados, classificados a partir do modelo AR inicialmente construído.

Em seguida desenvolve-se um novo modelo de *credit score* ponderado (no SAS, utiliza-se a opção *weight*) com apenas os proponentes aceitos, utilizando-se como variável de ponderação o peso obtido em cada faixa na tabela de distribuição

Tabela 3.2: Ponderação por faixas de escore - Modelo AR.

Faixas de Escore	Aceitos (A)	Rejeitados (B)	Total	Peso = (A+B)/A
0 - 499	233 (16%)	1223 (84,0%)	1456	6,25
500 - 549	675 (60,9%)	434 (39,1%)	1109	1,64
550 - 649	897 (70,4%)	378 (29,6%)	1275	1,42
650 - 749	988 (80,9%)	234 (19,1%)	1222	1,24
750 - 899	1432 (88,3%)	190 (11,7%)	1622	1,13
900 - 1000	1220 (100,0%)	0 (0%)	1220	1

de risco, conforme apresentado no exemplo da Tabela 3.2. Ao utilizarmos a opção *weight* do SAS, cada observação do conjunto de dados é ponderada pelo valor do peso da classe. A matriz de covariância estimada dos estimadores é invariante à escala do peso. Estamos assumindo que a probabilidade dos clientes *Bons*, dentre os aceitos, é a mesma dos proponentes supostamente *Bons*, dentre os rejeitados, em uma amostra qualquer M_k, na k-ésima faixa de escore, ou seja:

$$P(b_k/M_k, A_k) = P(b_k/M_k, R_k),$$

sendo

A_k: número de clientes aceitos na k-ésima faixa de escore, $k = 1, .., K$;

R_k: número de clientes rejeitados na k-ésima faixa de escore, $k = 1, .., K$;

b_k: número de clientes *Bons* na k-ésima faixa de escore, $k = 1, .., K$;

M_k: amostra de clientes selecionada na k-ésima faixa de escore, $k = 1, .., K$.

Os proponentes aceitos na faixa k são ponderados para representar os casos A_k e R_k. O peso obtido a partir de $(R_k + A_k)/A_k$ é o inverso da proporção dos aceitos na faixa k.

Este método atribui um peso para cada indivíduo a partir da distribuição total de clientes aceitos e rejeitados, como resultado da classificação do modelo AR inicialmente desenvolvido com todos os proponentes, estimando a probabilidade dos proponentes serem aprovados ou rejeitados, segundo as informações de suas

características descritas e analisadas na proposta de concessão de crédito.

3.2.4 Uso de Informações de Mercado

Este método utiliza informações de mercado, obtidas a partir de uma central de informações de crédito de um *bureau*, para inferir sobre o desempenho de crédito dos proponentes rejeitados. Um *bureau* de crédito possui informações sobre a atividade de crédito dos proponentes no mercado de modo amplo, provenientes de vários segmentos de mercado, como varejo, telefonia, seguradoras, bancos, entre outros, refletindo o seu comportamento no mercado em geral.

Este método é uma adaptação do método inicialmente proposto por Rocha e Andrade (2002), sendo que a diferença se dá basicamente pela definição do conceito de inadimplência e pela ausência de estratificação no grupo dos proponentes rejeitados através do uso do modelo genérico de crédito.

Os proponentes rejeitados são avaliados em dois momentos distintos, um no momento de análise do pedido de crédito e outro, no momento do desenvolvimento do novo modelo. Neste estudo, usamos o período de seis meses de avaliação. Avaliamos o desempenho dos indivíduos rejeitados segundo o seu comportamento de crédito, através dos registros sobre restrições de crédito, informados por instituições de mercado em geral.

As seguintes situações podem ocorrer para um proponente rejeitado:

1. O proponente possui restrições de crédito no momento do pedido de crédito e depois do período de avaliação este poderá:
   a) continuar com restrições ativas de crédito;
   b) não possuir mais restrições ativas, apenas resolvidas.

2. O proponente não possui restrições de crédito no momento do pedido de crédito e depois de um período de avaliação este poderá:
   a) possuir restrição de crédito ativa;
   b) continuar sem restrições ativas de crédito.

As etapas propostas para este método são as seguintes:

- seleção de uma amostra de aceitos (*Bons*/*Maus*) e rejeitados nas mesmas proporções do total de proponentes;

- consulta a uma central de informações de crédito e verificação sobre restrições de crédito ativas do grupo dos rejeitados;

- comparação das informações obtidas na central de crédito com as informações da proposta, no momento da rejeição de crédito;

- classificação dos proponentes rejeitados como *Bons* e *Maus*, segundo seu comportamento de crédito no mercado, no período de avaliação;

- desenvolvimento do novo modelo de crédito, incluindo a amostra dos clientes rejeitados, utilizando suas informações de mercado atual.

A base de desenvolvimento do novo modelo de crédito é constituída de clientes aceitos, com o seu real desempenho dentro da instituição credora, mais a amostra dos clientes rejeitados com o seu desempenho estimado pelo seu comportamento de crédito observado no mercado.

Com a utilização do uso de informações de mercado, para avaliação do desempenho de crédito dos clientes rejeitados, é possível reduzir o viés dos mesmos, uma vez que, agora, sabe-se qual o seu comportamento em relação ao mercado.

Apesar de não se ter a certeza de como o cliente rejeitado se comportaria caso fosse aceito na instituição que negou o seu crédito, tem-se a informação de como este cliente se comportou no mercado no período de avaliação observado. Dessa forma, estamos assumindo que este cliente se comportaria de maneira semelhante na instituição.

Quando utilizamos informações de mercado, temos um ganho natural de informação para os novos modelos desenvolvidos, pois temos informações adicionais, além das informações internas disponíveis na instituição credora. Porém a obtenção de informações de mercado junto às centrais de crédito, exige um custo financeiro, que deve ser considerado e avaliado no momento do desenvolvimento de novos modelos.

Capítulo 4

Avaliação de Modelos de *Credit Score*

Neste capítulo apresentamos as medidas mais utilizadas para avaliar o desempenho de modelos de *credit score*. Neste estudo usamos como indicadores de avaliação e comparação dos modelos ajustados, as medidas de Kolmogorov-Smirnov (KS), a curva ROC (*Receiving Operational Characteristic*) o coeficiente de Gini, a diferença entre as taxas de inadimplência nas faixas extremas do escore, definidas pelos decis (DTI), e a área entre as curvas da distribuição acumulada dos escores (AEC), sendo estas duas últimas medidas propostas por Tomazela (2007).

4.1 Estatística de Kolmogorov-Smirnov (KS)

A estatística de Kolmogorov-Smirnov (KS) é um indicador muito utilizado para avaliar o desempenho de modelos de *credit score*. Este indicador é baseado na idéia da distância entre as distribuições de probabilidades dos clientes inadimplentes e adimplentes. O KS mede a máxima separação entre a freqüência relativa acumulada de *Maus* clientes, $F_m(s)$ e a freqüência relativa acumulada de *Bons* clientes, $F_b(s)$. Na estatística não paramétrica, é usado para testar se duas amostras podem ser provenientes de uma mesma função distribuição (Conover, 1999).

A estatística de Kolmogorov-Smirnov é defnida por:

$$KS = \max_{0 \leq s \leq \infty} |F_m(s) - F_b(s)|, \quad 0 \leq KS \leq 100\% \ ,$$

sendo:

$F_m(s)$: freqüência relativa acumulada dos escores dos *Maus* clientes;

$F_b(s)$: freqüência relativa acumulada dos escores dos *Bons* clientes.

Segundo Oliveira e Andrade (2002), modelos de *credit score* com KS acima de 50% não são muito comuns para modelos utilizados na concessão de crédito para clientes novos; são mais freqüentes para modelos de *behaviour score*, aplicados aos que já são clientes, em que as variáveis de comportamento interno produzem um poder maior de discriminação dos modelos. Segundo Alves e Andrade (2004), modelos de *credit score* para concessão de crédito para novos clientes, desenvolvidos utilizando-se dados de mercado, geram maiores valores de KS do que aqueles que utilizam apenas informações internas. Na Tabela 4.1 são descritas algumas faixas de valores de KS, com seus respectivos níveis de discriminação, usualmente adotados pelo mercado. Apesar do KS ser a medida de avaliação mais utilizada, o uso isolado

Tabela 4.1: Valores de referência do KS.

Valor de KS	Níveis de Discriminações
Abaixo de 25	Baixo
De 25 a 35	Aceitável
De 35 a 45	Bom
Acima de 45	Excelente

deste indicador não garante que, para valores altos desta medida, tenhamos modelos bem ajustados, pois pode-se obter valores altos de KS quando o modelo discrimina indivíduos *Bons* e *Maus* em apenas uma faixa de escore. É recomendável o uso desta medida em conjunto com pelo menos outros dois indicadores de desempenho, como ROC e coeficiente de Gini, por exemplo.

A seguir, apresentamos na Tabela 4.2 o cálculo do KS, que pode ser facilmente obtido a partir de uma planilha eletrônica com as freqüências observadas das distribuições acumuladas dos escores dos clientes *Bons* e *Maus*, para cada faixa de risco. No software SAS, o KS é obtido utilizando-se a *Proc Npar1way*, sendo representado pela letra D. A partir da distribuição de risco dos clientes *Bons* e *Maus*

Tabela 4.2: Exemplo de cálculo do KS.

Faixas de Escore	*Bons* (%)	*Maus* (%)	Freq. acumulada *Bons*(%)	Freq. acumulada *Maus*(%)	Diferença absoluta
0-200	1,9	25,5	1,9	25,5	23,5
201-300	3,0	20,5	4,9	46,0	41,1
301-500	15,6	18,5	20,5	64,5	**44,0**
501-600	16,4	14,0	36,9	78,5	41,6
601-700	26,3	12,5	63,2	91,0	27,8
701-1000	36,8	9,0	100	100	0

em função das faixas de escores, apresentadas na Tabela 4.2, podemos representar a medida de KS graficamente, como apresentado na Figura 4.1.

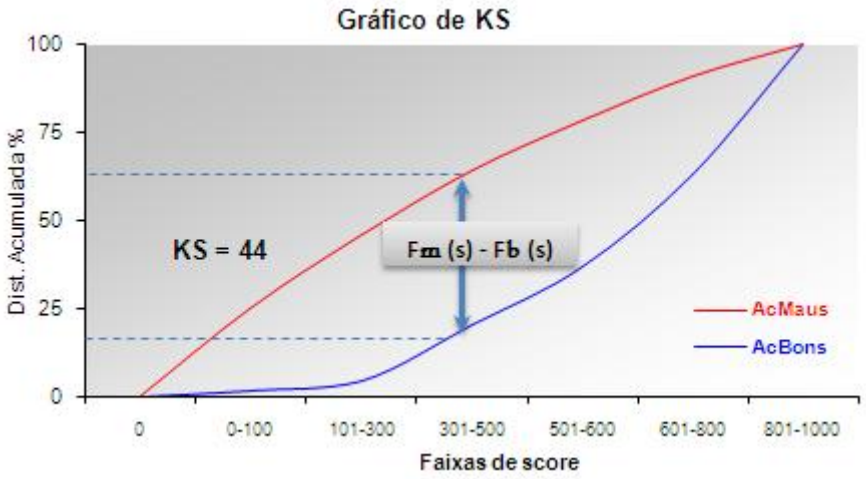

Figura 4.1: Exemplo de gráfico de KS.

4.2 Área Entre Curvas - AEC

A área entre as curvas da distribuição acumulada dos escores (AEC) é uma medida que avalia o desempenho de um modelo de *credit escore*, de maneira similar ao KS. Quanto maior a área, melhor será a capacidade de classificação e discriminação do modelo. Esta medida é obtida a partir da área da distribuição acumulada, utilizada para o cálculo da estatística de Kolmogorov-Smirnov (KS). Enquanto a medida KS mede a máxima diferença entre duas distribuições, em apenas uma faixa de escore, o AEC calcula a área total, gerada entre as curvas da distribuição acumulada dos escores, usadas para representar o KS.

O cálculo da área é feito através da soma das áreas dos retângulos, em que a base de cada retângulo é formada pela diferença entre dois escores consecutivos e a altura, pela função de distribuição acumulada empírica dos clientes inadimplentes até cada escore. A estimativa da área entre as curvas é obtida pela diferença absoluta entre as duas áreas. Para o cálculo da medida AEC, usamos a macro em SAS elaborada inicialmente por Tomazela (2007), sendo feitas algumas pequenas modificações, conforme código apresentado no Apêndice D.

4.3 Curva ROC e Coeficiente de Gini

A curva ROC (*Receiver Operating Characteristic*), também conhecida como curva de Lorenz (Hanley e McNeil, 1982), é bastante utilizada na área médica, para especificar problemas no desempenho de diagnósticos médicos, em que se procura indicar a presença ou a ausência de uma doença, com determinada probabilidade de erro. Na área de risco de crédito é uma técnica bastante útil para validação de desempenho dos modelos de crédito. A curva é baseada nos conceitos de sensibilidade e especificidade estatísticas (medidas de taxas de acertos) que podem ser obtidas a partir da construção de matrizes de confusão (2x2) (Johnson e Wichern, 2002), obtidas do resultado da classificação dos indivíduos, gerada pelo modelo.

Com o modelo ajustado, a partir de uma amostra de n clientes, para cada indivíduo se atribui um escore S. Assim o i-ésimo indivíduo será classificado como *Mau* se $S_i \leq P_c$, (em que P_c é um ponto de corte para o escore S_i, pré-determinado) e como *Bom* caso contrário. Para um determinado P_c, é possível determinar a matriz de confusão, como apresentado na Tabela 4.3.

Tabela 4.3: Matriz de Confusão.

	Classificado como		
Situação real	*Mau*	*Bom*	Total
Mau	n_{11}	n_{12}	$n_{1.}$
Bom	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Os valores representados na Tabela 4.3 são:

n_{11}: número de clientes *Maus* classificados como *Maus* $\Rightarrow$ *Acerto*;

n_{12}: número de clientes *Maus* classificados como *Bons* $\Rightarrow$ *Erro*;

n_{21}: número de clientes *Bons* classificados como *Maus* $\Rightarrow$ *Erro*;

n_{22}: número de clientes *Bons* classificados como *Bons* $\Rightarrow$ *Acerto*.

Através da matriz de confusão é possível determinar as taxas de acertos, que são as medidas de especificidade (proporção de clientes *Maus*, classificados corretamente por terem escore menor que um ponto de corte) e de sensibilidade (proporção de clientes *Bons*, classificados corretamente por terem escore igual ou superior a um ponto de corte), ou seja:

$$Sensibilidade = n_{22}/n_{2.} \quad \text{e} \quad Especificidade = n_{11}/n_{1.} \quad .$$

Pode-se calcular também a acurácia do modelo, que será dada pela proporção de acertos total, ou seja:

$$Acurácia = (n_{11} + n_{22})/n \quad .$$

E ainda construir respectivamente a partir da acurácia, sensibilidade e especificidade, %AT, %AB e %AM, definidos como:

%AT: Percentual de Acertos Total $\Rightarrow$ $\%AT = (n_{11} + n_{22})/n \times 100$;
%AB: Percentual de Acertos entre os *Bons* $\Rightarrow$ $\%AB = n_{22}/n_{2.} \times 100$;
%AM: Percentual de Acertos entre os *Maus* $\Rightarrow$ $\%AM = n_{11}/n_{1.} \times 100$.

Os indicadores %AT, %AB e %AM podem ser utilizados para verificar o desempenho dos modelos ajustados e possibilitar a comparação através das taxas de acertos.

A curva ROC é construída a partir da união dos pontos formados pela sensibilidade e $(1-$*especificidade*$)$, calculadas a partir de todas as matrizes de confusão, geradas pelas observações da amostra, considerando-se diferentes pontos de corte do modelo. A Figura 4.2 mostra um exemplo de curva ROC.

Hosmer e Lemeshow (2000, p.162) apresentam a seguinte regra geral para avaliação do resultado da área sob a curva ROC, aplicada em modelos de *credit score*:
área abaixo de 0,7: baixa discriminação;
área entre 0,7 e 0,8: discriminação aceitável;
área entre 0,8 e 0,9: discriminação excelente;
área acima de 0,9: discriminação excepcional.

Como a área sob a curva ROC varia de 0,5 a 1, é mais adequado utilizar o coeficiente de Gini (Thomas et. al., 2002), que é dado por duas vezes a área entre a curva ROC e a reta diagonal ($X = Y$). Tem-se um indicador de desempenho que varia entre 0 e 1. O cálculo do coeficiente de Gini resulta diretamente da utilização da curva ROC.

Observando-se o exemplo ilustrado na Figura 4.2 da curva ROC, descreve-se a interpretação do cálculo do coeficiente de Gini, que pode ser definido como sendo o quociente da área entre a diagonal e a curva (área A) sobre a área total acima

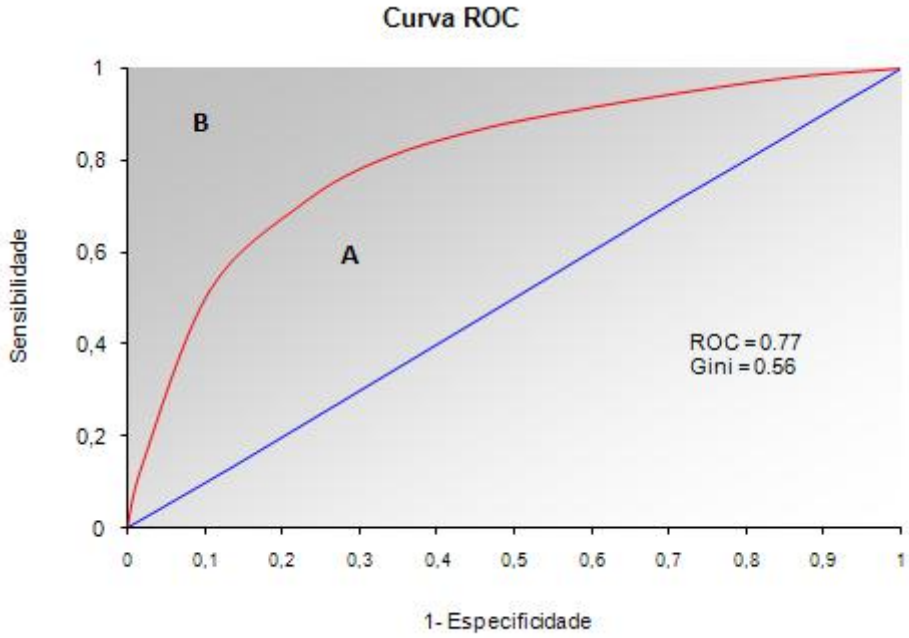

Figura 4.2: Exemplo de curva ROC.

da diagonal (soma da área A com a área B). Quanto mais a curva se distanciar da diagonal, maior será o coeficiente de Gini e maior será a separação entre *Bons* e *Maus*, ou seja,

$$Gini = (\acute{a}rea\ A)/(\acute{a}rea\ A + \acute{a}rea\ B)\ .$$

Como A é a diferença entre a área acima da diagonal e a área acima da curva, e A + B é toda a área acima da diagonal, sendo igual à metade da área de um quadrado, (ou seja, 1/2), podemos obter o coeficiente de Gini da seguinte forma:

$$Gini = 2 \times (\acute{a}rea\ acima\ da\ diagonal\ - \acute{a}rea\ acima\ da\ curva),$$

ou ainda, diretamente do valor obtido da curva ROC, como:

$$Gini = 2\ \times (ROC - 0,5),$$

sendo ROC, neste caso, o valor obtido do cálculo da área sob a curva ROC.

Com o aplicativo SAS, usando-se o procedimento *Proc Logistic* com a opção *OutRoc*, é produzido o conjunto de resultados necessário para construção da curva ROC. A medida ROC e o coeficiente de Gini são calculados a partir dos resultados gerados nesse mesmo procedimento.

4.4 Diferença entre Taxas de Inadimplência (DTI)

Tomazela (2007) usa a diferença entre taxas de inadimplência (DTI) como um indicador do desempenho de modelos de *credit score*, em cada intervalo definido pelos decis dos escores. Esse indicador se constitui pela diferença entre as taxas de inadimplência nas classes, geralmente a primeira e a última (classes extremas), obtendo-se assim uma dimensão da capacidade de separação entre *Bons* e *Maus* clientes, alcançada pelo modelo ajustado.

Tabela 4.4: Exemplo do cálculo da taxa de inadimplência por classes de escore.

Classes	*Bons*	*Maus*	Total	Tx. Inad.
0-421 (c_1)	108	819	927	88%
421-588 (c_2)	177	750	927	81%
588-710 (c_3)	229	691	920	75%
710-797 (c_4)	346	575	921	62%
797-846 (c_5)	492	427	919	46%
846-877 (c_6)	552	376	928	41%
877-898 (c_7)	609	329	938	35%
898-917 (c_8)	655	275	929	29%
917-935 (c_9)	721	225	949	24%
935-1000 (c_{10})	723	146	869	17%

Para se obter a DTI, primeiramente deve-se calcular os decis dos escores. Em seguida calcula-se, para cada classe de decil, a freqüência total de observações e também a freqüência de clientes *Bons* e *Maus*, observada em cada intervalo. A taxa

de inadimplência é calculada a partir do total de clientes classificados como *Maus*, dividido pelo total de clientes em cada faixa de escore.

Na Tabela 4.4 é apresentado um exemplo da obtenção do cálculo das taxas de inadimplência por faixas de escore divididas por decis. Para o cálculo da medida DTI, foi aplicada a macro em SAS, elaborada por Tomazela (2007), cujo código é apresentado no Apêndice D.

A diferença entre as taxas de inadimplência para os intervalos extremos é dada por:

$$DTI(1;10) = Tx.Inad_{(c1)} - Tx.Inad_{(c10)} = 88\% - 17\% = 71\%,$$

ou seja, para o primeiro decil, o modelo classifica 88% como *Maus* pagadores e para o último, 17%, o que resulta em 71% a diferença entre essas taxas de inadimplência. Segundo Tomazela (2007), essa medida expressa em quantos pontos percentuais o escore consegue reduzir a inadimplência e quanto maior o valor detectado, melhor a separação proporcionada pelo modelo.

Capítulo 5

Resultados para a Aplicação 1

Neste capítulo, apresentamos os resultados para a Aplicação 1, das técnicas de *inferência de rejeitados* descritas no Capítulo 3, aplicadas aos dados descritos no Capítulo 2 obtidos de um *bureau* de crédito. Utilizando uma abordagem de um problema de *credit score*, desenvolvido a partir da técnica de regressão logística, avaliamos os resultados do uso da *inferência de rejeitados*, segundo a qualidade de ajuste dos modelos construídos. Analisamos e comparamos o desempenho dos modelos ajustados, apontando as diferenças entre eles e usando como referência o modelo sem a informação dos clientes rejeitados.

Primeiramente apresentamos a distribuição da base de dados em relação aos clientes aceitos e rejeitados, bem como em função do conceito de inadimplência definido no Capítulo 2; em seguida, é descrito o processo de tratamento e seleção das variáveis explicativas, os ajustes dos modelos com a utilização das técnicas de rejeitados e os resultados finais.

Os resultados gerados nesta análise foram obtidos no aplicativo SAS V9.1, módulo SAS/STAT, com o procedimento *Proc Logistic* e o SPSS/Clementine V11.

5.1 Distribuição da Base de Dados

A base de modelagem empregada neste estudo, foi distribuída em clientes aceitos e rejeitados, conforme a definição de rejeitados apresentada no Capítulo 2. A distribuição obtida pode ser visualizada na Tabela 5.1.

Tabela 5.1: Distribuição da base de dados segundo clientes aceitos e rejeitados.

Clientes	n	%
Aceitos	40943	82%
Rejeitados	9057	18%

Em seguida, os clientes aceitos e rejeitados são separados aleatoriamente, em base de desenvolvimento (para o ajuste dos modelos) e validação. Pode-se observar na Tabela 5.2 que a distribuição dos clientes aceitos e rejeitados, nas amostras particionadas de desenvolvimento e validação, apresentam-se de modo homogêneo, caracterizando bem o total da base de dados.

Tabela 5.2: Distribuição das amostras - desenvolvimento e validação.

Clientes	Desenvolvimento	Validação	Total
ACEITOS			
Adimplentes (*Bons*)	20060 (67%)	13002 (66%)	33062 (66%)
Inadimplentes (*Maus*)	4612 (15%)	3269 (16%)	7881 (16%)
REJEITADOS	5434 (18%)	3623 (18%)	9057 (18%)
Total	30106 (100%)	19894 (100%)	50000 (100%)

O conjunto de todas das variáveis explicativas é apresentado no Capítulo 2, e o seu detalhamento é descrito no Apêndice A.

5.2 Categorização das Variáveis Explicativas

A grande maioria das variáveis explicativas, presentes na base de dados, é composta por variáveis quantitativas. Segundo Pereira (2004), no desenvolvimento de modelos de *credit score*, é comum a categorização de todas as variáveis em um número não muito excessivo de classes, pois, para as variáveis quantitativas, dificilmente a relação entre o logito da esperança da variável resposta e uma variável preditora qualquer é linear.

É comum efetuar categorizações para as variáveis quantitativas, mesmo quando existe uma relação linear visível, para reduzir a possibilidade de influência de pontos discrepantes, além de possibilitar uma interpretação mais clara do modelo. A categorização das variáveis possibilita, em uma primeira análise, a eliminação de variáveis desprezíveis ao modelo. Por exemplo, variáveis com um número reduzido de observações, podem ocasionar estimativas pouco confiáveis, associadas a estas variáveis, sendo recomendável a não utilização dessa variável no estudo do ajuste do modelo de crédito. Outra situação bastante comum é aquela em que há duas categorias de uma mesma variável que apresentam risco de crédito semelhantes. Neste caso, é indicado que sejam agrupadas em única categoria, desde que esta faça sentido quanto às regras de negócio da empresa.

Existem diversas maneiras para efetuar a categorização de variáveis quantitativas. Uma delas é, por exemplo, através da distribuição dos percentis, sendo possível efetuar agrupamentos a cada 5% ou 10% das observações. Uma outra maneira freqüentemente utilizada é o uso do risco relativo (RR) ou ainda através do WOE (*Weight of Evidence*), proposto inicialmente por Good (1950), e obtido a partir do logaritmo do risco relativo.

Para efetuar as categorizações das variáveis da Aplicação 1 e seus respectivos agrupamentos, seguiu-se a metodologia descrita no Capítulo 2, na Seção 2.3. Os resultados das análises descritivas com os seus respectivos agrupamentos finais, são apresentados nas Tabelas B.1 a B.15, no Apêndice B.

5.3 Processo de Seleção das Variáveis

O processo de seleção de variáveis, utilizado nos ajustes dos modelos de regressão logística, foi a técnica de *forward-stepwise*, (ver Hosmer e Lemeshow, 2000). A técnica de *forward-stepwise* é o processo de inclusão e exclusão de variáveis no modelo. Inicia-se pela estimação de um modelo, incluindo-se primeiramente o intercepto e em seguida uma a uma das variáveis explicativas mais significantes, excluindo-se aquelas que, na presença das demais, não são significativas. O processo é finalizado no momento em que nenhuma variável pode ser mais incluída ou excluída do modelo. Os níveis de significância de inclusão e exclusão são previamente definidos; neste estudo, foram iguais a 0,1.

Quando uma instituição financeira desenvolve um modelo de crédito, além do processo padrão de *forward-stepwise* são feitas recategorizações e agrupamentos entre os níveis das variáveis, para diminuir o efeito de multicolinariedade e deixar o modelo mais estável, buscando possivelmente modelos mais interpretáveis em relação ao negócio de crédito.

Para evitar o favorecimento de alguma técnica de *inferência de rejeitados*, utilizada neste estudo, procurou-se efetuar o menor número de ajustes possível nos modelos. Foram feitos pequenos ajustes somente quando os coeficientes estimados das variáveis indicadoras apresentavam uma ordenação entre as categorias, diferente da sugerida pela análise do RR (risco relativo). Se por exemplo, para uma determinada variável, a categoria 1 possuir RR maior que o da categoria 2, espera-se que o parâmetro estimado, referente à categoria 1, seja maior que o de categoria 2. Caso isso não ocorra, agrupam-se as duas categorias e estimam-se novamente os parâmetros do modelo. Outra verificação freqüentemente utilizada é a avaliação dos parâmetros estimados de cada categoria com o seu respectivo RR calculado. Caso sejam divergentes, é feita a tentativa de reagrupamento.

5.4 Ajuste dos Modelos

A partir da base de dados descrita no Capítulo 2, foram desenvolvidos cinco modelos de *credit score*. Um primeiro modelo para ser utilizado como referência, sem a utilização da informação dos clientes rejeitados e os demais com a inclusão da informação dos clientes rejeitados. Para o desenvolvimento dos modelos com a utilização da informação dos rejeitados, foram aplicadas as técnicas de *inferência de rejeitados*, descritas no Capítulo 3.

Foi criada uma nomenclatura específica para facilitar a referência aos modelos desenvolvidos do seguinte modo: modelo sem o uso de *inferência de rejeitados* (modelo I), que será tomado como referência para comparações entre os demais modelos; modelo com a utilização da técnica de *inferência de rejeitados* de classificação de todos os clientes rejeitados como *Maus* (modelo II); modelo com a utilização da técnica de parcelamento (modelo III); modelo com o uso da técnica de dados aumentados (modelo IV) e modelo com a utilização de informações de mercado (modelo V).

Para a obtenção das estimativas dos parâmetros dos modelos de *credit score* desenvolvidos, as variáveis explicativas foram pré-selecionadas a partir da análise descritiva. Foram selecionadas aquelas que apresentavam maior associação com a variável resposta para, em seguida, ser aplicado o procedimento *proc reg logistic* no programa SAS, com a opção do método de seleção *stepwise*.

As estimativas dos parâmetros dos modelos finais foram obtidas para cada técnica de *inferência de rejeitados*, bem como, os erros padrões e os seus respectivos valores p, que estão no Apêndice C.

Utilizamos o teste de Hosmer-Lemeshow para avaliar o ajuste dos modelos em cada etapa. O teste de Hosmer-Lemeshow associa os dados às suas probabilidades estimadas, da mais baixa à mais alta. Faz-se um teste qui-quadrado para determinar se as freqüências observadas estão próximas das freqüências esperadas (Hosmer e Lemeshow, 2000).

Os níveis descritivos do teste de Hosmer-Lemeshow, para os ajustes finais dos modelos I, II, III, IV e V, foram respectivamente 0,58, 0,10, 0,12, 0,08 e 0,39, indicando para cada caso que não rejeitamos a hipótese nula de que o modelo se ajusta de maneira satisfatória aos dados, ou seja, independentemente do uso ou não das informações dos rejeitados, temos evidências para acreditar que os modelos foram bem ajustados aos dados.

A probabilidade de um cliente classificado nas categorias de referência das variáveis explicativas se tornar inadimplente está relacionada ao valor da constante do modelo logístico. Para o Modelo I temos que, $P(Mau) = e^{(1,019)}/[1 + e^{(1.019)}] = 0,7347$, ou seja, é a probabilidade estimada desse cliente se tornar inadimplente.

Após o ajuste de cada modelo, para cada técnica de *inferência de rejeitados* empregada, foram construídos gráficos de KS, ROC e distribuição de freqüência para os valores de escores, apresentados no Apêncide E. Os Gráficos E.1 a E.4 correspondem aos resultados do ajuste do modelo de referência (modelo I); E.5 a E.8 ao modelo II (classificação dos rejeitados como *Maus*); E.9 a E.12 ao modelo III (parcelamento); E.13 a E.16 ao modelo IV (dados aumentados); E.17 a E.20 ao modelo V (uso de informações de mercado).

5.5 Comparação dos Resultados

Para comparar os modelos de *credit score* desenvolvidos, foram utilizadas as técnicas descritas no Capítulo 4. Na Tabela 5.3 têm-se as medidas de desempenho para a avaliação dos modelos desenvolvidos para a amostra de desenvolvimento e de validação. Pode-se observar que os modelos II e III apresentaram medidas próximas às do modelo I.

Os modelos II e III, que utilizaram as técnicas de classificação e parcelamento respectivamente, apresentaram desempenhos semelhantes quanto às medidas KS, Gini, AEC e ROC, sendo superiores às medidas obtidas para o modelo IV.

Tabela 5.3: Medidas de desempenho para os modelos ajustados.

Medida de Desempenho	Modelo I	Modelo II	Modelo III	Modelo IV	Modelo V
KS (desenvolvimento)	39,4	40,7	41,5	38,4	46,7
KS (validação)	39,2	38,9	39,0	35,4	41,9
AEC (desenvolvimento)	0,26	0,27	0,27	0,25	0,33
AEC (validação)	0,25	0,26	0,24	0,23	0,31
ROC(desenvolvimento)	76,0	77,2	76,3	75,5	78,8
ROC(validação)	75,0	75,5	74,7	73,0	76,8
Gini (desenvolvimento)	0,52	0,54	0,53	0,51	0,57
Gini (validação)	0,50	0,51	0,50	0,46	0,53

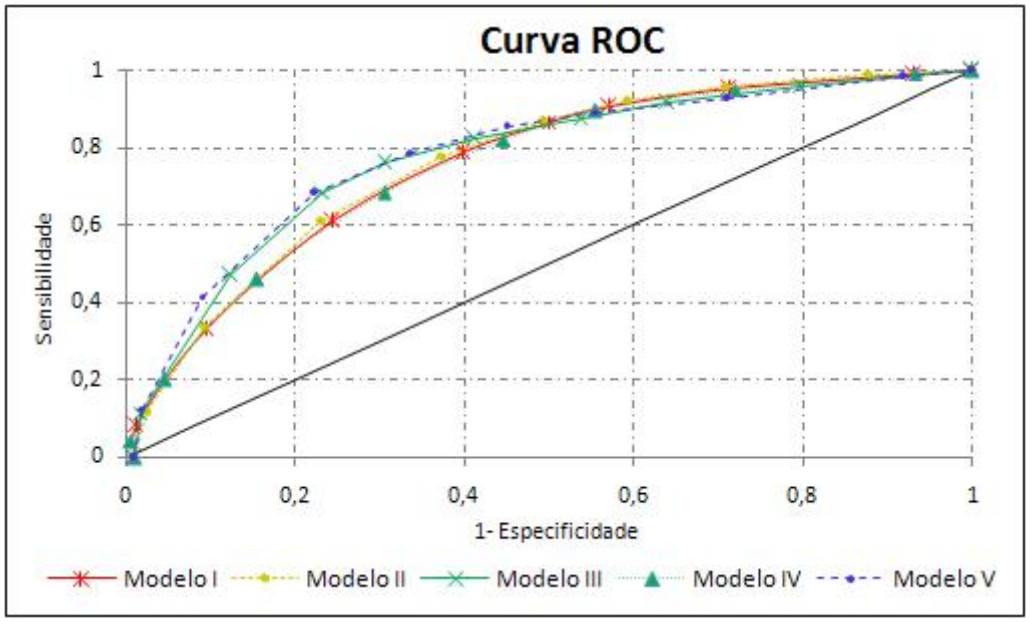

Figura 5.1: Gráfico da curva ROC.

A área sob a curva ROC entre as técnicas de desenvolvimento de melhor e pior desempenho, são de 76,8 (modelo V) e 73,0 (modelo IV) respectivamente, sendo que a diferença entre ambas é de 3,8 pontos. O modelo V, que utilizou dados de mercado para inferir o comportamento dos rejeitados, foi o que apresentou os

melhores resultados para as medidas de desempenho, enquanto que o modelo IV, no qual foi aplicada a estratégia de dados aumentados, apresentou desempenho inferior às demais estratégias.

Na Figura 5.1 temos a curva ROC para os modelos sem o uso das informações de rejeitados (modelo I) e com aplicação das técnicas de *inferência de rejeitados* (modelos II a V). Observa-se que o modelo V (com a utilização de informações de mercado) e o modelo III (estratégia de parcelamento dos rejeitados) apresentam área superior aos demais modelos desenvolvidos.

A Figura 5.2 mostra as taxas de inadimplência para os modelos I a V, por classe de decil. Pode-se observar que para todos os modelos desenvolvidos, a taxa de inadimplência são maiores para os intervalos de escores mais baixos, sendo decrescente à medida em que os intervalos de escore aumentam, demonstrando boa discriminação entre os clientes adimplentes e inadimplentes, pois, nos modelos de *credit score*, espera-se que quanto maior o escore, menor seja a taxa de inadimplência esperada.

Tabela 5.4: Taxas de inadimplência por intervalo de decil (em %).

Classes	ModeloI	ModeloII	ModeloIII	ModeloIV	ModeloV
c1	88,4	88,0	88,0	88,2	90,5
c2	80,6	82,4	80,6	81,7	86,0
c3	70,7	73,7	72,6	70,1	76,7
c4	57,3	61,2	62,9	57,4	64,6
c5	48,8	49,1	54,6	49,6	52,8
c6	42,1	41,3	50,5	42,7	41,2
c7	37,5	33,1	43,7	38,6	32,9
c8	30,1	23,9	38,4	29,4	27,1
c9	26,1	26,1	29,9	26,0	24,8
c10	18,0	14,5	20,9	15,8	15,0
(c1-c10)	70,4	73,5	67,1	72,4	74,5

Na Tabela 5.4 têm-se as taxas de inadimplência por classes de decil e as

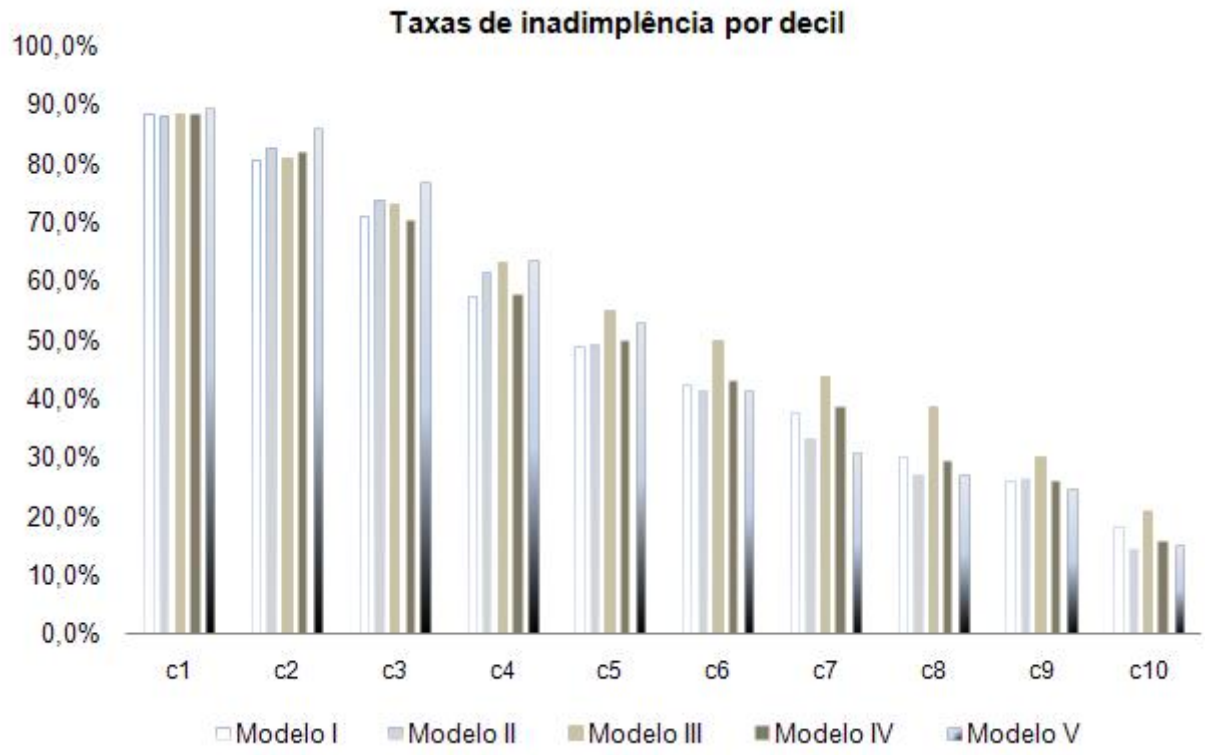

Figura 5.2: Gráfico das taxas de inadimplência por decil.

diferenças das taxas de inadimplências (DTI) para os modelos I a V, obtidos a partir da diferença das classes dos intervalos dos decis extremos (c1-c10). Observa-se que o indicador DTI para os intervalos extremos foi de 74,5% para o modelo V, que utilizou informações de mercado para os rejeitados e apresentou melhor desempenho, enquanto o modelo III foi o que obteve menor DTI, 67,1%. Esse melhor desempenho do modelo V já era esperado, pois ele é o único que conta com informações adicionais às apresentadas na amostra.

Capítulo 6

Resultados para a Aplicação 2

Neste capítulo apresentamos os resultados da Aplicação 2, um estudo de *inferência de rejeitados* aplicados como uma ferramenta de apoio na gestão de relacionamento com o cliente de uma empresa do setor de telecomunicações do segmento de TV por assinatura. Mostramos os resultados de uma aplicação no desenvolvimento de um modelo de crédito utilizado para aprovação de novos clientes, construído com o auxílio do uso de *inferência dos rejeitados*, utilizando a estratégia de análise de aceitar uma amostra de proponentes rejeitados para avaliar e obter o seu real desempenho posterior.

Segundo Ogava (2007), devido ao mercado de TV por assinatura estar cada vez mais competitivo, a aquisição de um novo cliente é muito mais cara que sua manutenção na base. Desta maneira, as empresas se preocupam cada vez mais em desenvolver ferramentas de auxílio na gestão do relacionamento com o cliente. Nesse contexto, surge a necessidade de se obter um modelo de crédito que possa prever a probabilidade de um cliente novo se tornar inadimplente nos próximos meses subseqüentes à data da assinatura, tendo como conseqüência o serviço cancelado por falta de pagamento e possivelmente gerando um cancelamento involuntário.

6.1 Entendendo o Problema

A base de dados da Aplicação 2 é constituída das informações contidas na proposta de aquisição para novos clientes da empresa, além das informações de mercado sobre o comportamento de crédito, adquiridas em fontes externas como ACSP - Associação Comercial de São Paulo e Serasa (ver Tabela 6.1). A descrição detalhada das variáveis utilizadas nessa aplicação são apresentadas no Apêndice A.2.

Tabela 6.1: Variáveis da base de dados da aplicação 2.

Variáveis Explicativas	Cadastral	Crédito	Interna	Externa
Sexo	x		x	
Idade	x		x	
Estado Civil	x		x	
Nome do Banco	x		x	
Cartão de Crédito	x		x	
Instalação	x		x	
Tipo de Imóvel	x		x	
Forma de Pagamento	x		x	
PDV - Ponto de Venda	x		x	
Pontos Adicionais	x		x	
Presença de Telefone	x		x	
Região de Venda	x		x	
Cancelamento por Região	x		x	
Cheques Devolvidos		x		x
CCF - Banco Central		x		x
PEFIN (Pendências Financeiras)		x		x
Valor de PEFIN		x		x

Durante um período pré-definido, foi liberada a aprovação de todos os proponentes, ou seja, as pessoas que solicitaram o pedido de prestação de serviço da empresa tiveram a aprovação sem nenhum tipo de restrição, com exceção dos ex-clientes que tiveram suas assinaturas canceladas por falta de pagamento; para esses clientes o serviço foi negado.

Três meses depois, todos os clientes selecionados no período de liberação do filtro de crédito para a concessão do serviço foram avaliados segundo o seu comportamento de inadimplência dentro da empresa. Verificou-se quais deles se tornaram inadimplentes e quantos dos que seriam rejeitados permaneceram adimplentes. Nessa empresa, para que um cliente seja considerado inadimplente, é preciso que este esteja em atraso de pagamento de fatura acima de 35 dias.

6.2 Análise Descritiva

Para o desenvolvimento do modelo de *credit score* utilizou-se a amostra controle gerada pela liberação do filtro de crédito. Para efeito de comparação foi desenvolvido também um outro modelo de *credit score*, somente com a utilização de clientes aprovados em períodos próximos da obtenção da amostra de controle. Na Tabela 6.2, tem-se a distribuição da amostra de desenvolvimento para o modelo sem o uso dos proponentes rejeitados.

Tabela 6.2: Distribuição da amostra utilizada para o desenvolvimento do modelo sem a utilização dos rejeitados.

Cliente	Amostra Utilizada
Bom (atraso até 34 dias)	8453 (54,8%)
Mau (atraso superior a 34 dias)	6964 (45,2%)
Total	15417 (100,0%)

Na Tabela 6.3 temos a distribuição da amostra final utilizada para o desenvolvimento do modelo de *credit score* com a utilização dos rejeitados. Foram selecionados todos os clientes *Maus* e o mesmo número de *Bons*, a fim de não termos nenhum desequilíbrio amostral, pois temos na base de clientes da empresa muito mais clientes *Bons* do que *Maus*.

Tabela 6.3: Distribuição da amostra utilizada para o desenvolvimento do modelo com a informação dos clientes rejeitados.

Cliente	Amostra Utilizada
Bom (atraso até 34 dias)	4309 (50,1%)
Mau (atraso superior a 34 dias)	4297 (49,9%)
Total	8606 (100,0%)

As categorizações das variáveis da Aplicação 2 e seus respectivos agrupamentos seguiram a metodologia descrita no Capítulo 2. As tabelas descritivas com os agrupamentos finais são apresentadas nas Tabelas B.16 a B.23.

Para o grupo de clientes aceitos sem nenhum filtro de crédito, chamado grupo controle, foram simulados dois filtros de crédito para verificar de modo exploratório quais clientes seriam inicialmente rejeitados caso o modelo de crédito fosse aplicado e como tais clientes se comportaram depois de aceitos.

Foram aplicados dois filtros de crédito, um mais rigoroso e outro menos rigoroso; para o filtro mais rigoroso um cliente era rejeitado se tivesse pendências financeiras ativas e/ou resolvidas nos últimos seis meses maior que um determinado valor. Já para o filtro menos rigoroso um cliente era rejeitado se tivesse pendências financeiras ativas e/ou resolvidas nos últimos doze meses pelo menos duas vezes maior do que o valor estabelecido para o filtro mais rigoroso, sendo o segundo filtro bem mais flexível do que o primeiro.

Nas Tabelas 6.4 e 6.5 temos, respectivamente, a distribuição das propostas selecionadas e a distribuição da inadimplência observada das propostas negadas e aceitas com a simulação e aplicação dos filtros de crédito.

Observa-se que se os filtros de créditos tivessem sido utilizados, a inadimplência seria menor, porém esta diferença não segue a mesma proporção dos clientes que seriam rejeitados, ou seja, caso o Filtro 1 (filtro mais rigoroso) fosse aplicado, teríamos uma taxa de inadimplência reduzida em 3% (de 16,1% para 13%), com uma taxa de clientes rejeitados em torno de 11%. Já para o Filtro 2 (menos rigoroso), a diferença

Tabela 6.4: Distribuição das propostas selecionadas e aplicação dos filtros de crédito.

Propostas	Filtro 1	Sem Filtro	Filtro 2
Aceitas	23644 (88,7%)	26649 (100)%	25125 (94,3)%
Rejeitadas	3005 (11,3%)	0 (0)%	1524 (5,7)%
Total	26649 (100%)	26649 (100)%	26649 (100)%

Tabela 6.5: Distribuição das propostas aceitas segundo a aplicação dos filtros de crédito.

Clientes	Filtro 1	Sem Filtro	Filtro 2
Adimplentes	20571 (87,0%)	22352 (83,9)%	21608 (86,0)%
Inadimplentes	3074 (13,0%)	4297 (16,1)%	3518 (14,0)%
Total	23644 (88,7%)	26649 (100)%	25125 (94,3)%

da taxa de inadimplência é de 2% (de 16,1% para 14%) enquanto que os clientes rejeitados chegariam a 5,7%.

Na Tabela 6.6 temos a distribuição de inadimplência da amostra utilizada para o desenvolvimento do modelo com informação dos clientes rejeitados, simuladas para o filtro 2.

Tabela 6.6: Distribuição da amostra aceita segundo a aplicação do filtro 2 de crédito.

Filtro de Crédito	*Bom*	*Mau*	Total
Aceitos	3979 (51,7%)	3719 (48,3)%	7698 (100)%
Rejeitados	330 (36,3%)	578 (63,7)%	908 (100)%

Pode-se observar na Tabela 6.6 que, dos 908 clientes rejeitados pela simulação do filtro 2, 330 (36,3%) permaneceram adimplentes, representando uma oportunidade de ganho para a empresa, pois, caso a política de crédito tivesse sido aplicada para esses clientes, eles teriam sido descartados.

Quando existe simplesmente a aplicação de alguns filtros como política de de crédito, a liberação de uma amostra de rejeitados, pode ser interessante, para

avaliar e recalibrar a política de crédito em vigência de uma empresa.

6.3 Seleção das Variáveis e Ajuste dos Modelos

Para a seleção das variáveis foram utilizados os mesmos critérios descritos no Capítulo 2, isto é, da análise descritiva utilizou-se o risco relativo para selecionar as variáveis candidatas à seleção dos modelos e também para efetuar os agrupamentos das categorias semelhantes, segundo o risco relativo para uma mesma variável explicativa. Para avaliar o uso das informações dos rejeitados, como na primeira parte desta dissertação, foi desenvolvido um modelo para que fosse empregado como referência, sem a utilização das informações dos clientes rejeitados (modelo VI). Esse modelo foi comparado com o modelo ajustado (modelo VII) a partir da base controle (amostra sem filtros), a qual contém todos os clientes, aceitos e possíveis rejeitados, caso o modelo de crédito tivesse sido aplicado.

As estimativas dos parâmetros são apresentados nas Tabelas C.6 e C.7 e os gráficos com as medidas resumos (KS e ROC) são apresentados nas Figuras E.21 a E.24.

6.4 Comparação dos Resultados

Para comparar os modelos de *credit score* desenvolvidos, foram utilizadas as ténicas descritas no Capítulo 4. Na Tabela 6.7 temos as medidas de desempenho para a avaliação dos modelos, para a amostra de desenvolvimento e de validação a partir da base de dados reais, para os modelos VI e VII.

Observa-se que, no modelo desenvolvido a partir da amostra controle, os valores das medidas de desempenho são superiores para todas as medidas avaliadas. Os ganhos mais expressivos são observados na medida de KS, equivalente a 6 pontos e no coeficiente de Gini, sendo superior em 4 pontos.

Na Figura 6.1 tem-se a curva ROC para os dois modelos avaliados. Observa-se

Tabela 6.7: Medidas de desempenho para os modelos ajustados - Aplicação 2.

Medidas de desempenho	Modelo VI	Modelo VII
KS (desenvolvimento)	28,8	35,8
KS (validação)	28,5	34,7
AEC (desenvolvimento)	0,19	0,23
AEC (validação)	0,19	0,23
ROC (desenvolvimento)	65,0	67,3
ROC (validação)	64,9	67,2
Gini (desenvolvimento)	30,0	34,0
Gini (validação)	30,0	34,0

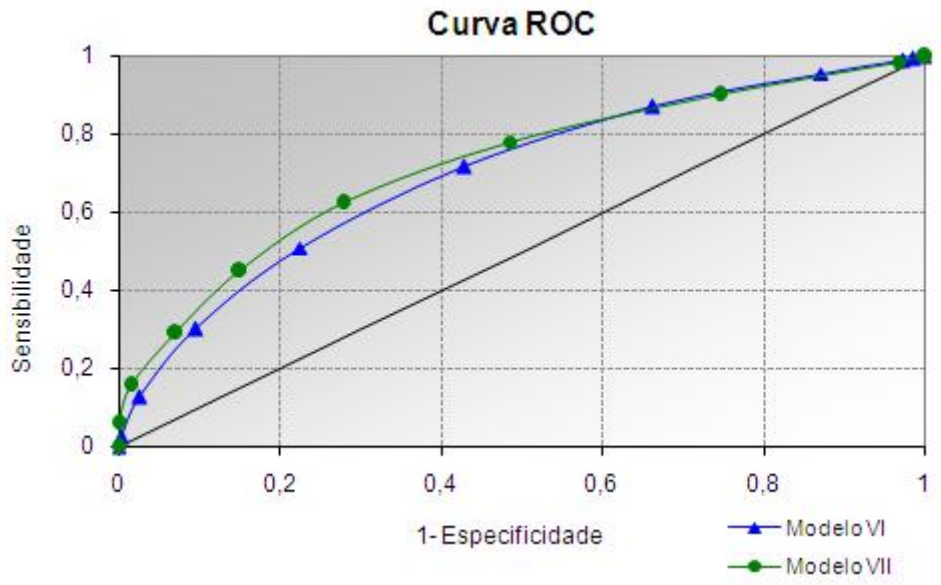

Figura 6.1: Curva ROC com e sem o uso das informações dos rejeitados.

que o modelo que utiliza informações dos rejeitados apresenta maior área, indicando melhor ajuste e conseqüentemente melhor acurária no poder classificatório.

Nas Tabelas 6.8 e 6.9 temos, respectivamente, as matrizes de confusão para

os modelos desenvolvidos sem e com o uso dos clientes rejeitados.

Tabela 6.8: Matriz de confusão - modelo VI.

	Classificado como		
Situação real	*Bom*(%)	*Mau*(%)	Total(%)
Bom	6096 (72)	2357 (28)	8453 (100)
Mau	3033 (44)	3931 (56)	6964 (100)

Tabela 6.9: Matriz de confusão - modelo VII.

	Classificado como		
Situação real	*Bom*(%)	*Mau*(%)	Total(%)
Bom	2694 (62)	1615 (38)	4309 (100)
Mau	1197 (28)	3100 (72)	4297 (100)

A partir das matrizes de confusão é possivel calcularmos as porcentagens de acertos dos clientes *Bons*, *Maus* e total (%AB, %AM, e %AT), conforme definiu-se no Capítulo 4.

Tabela 6.10: Taxas de acertos para os modelos desenvolvidos.

Modelos	%AB	%AM	%AT
Modelo VI	72,1	56,4	65,0
Modelo VII	62,5	65,0	67,3

Pode-se observar na Tabela 6.10, que o modelo sem o uso das informações dos rejeitados, apresenta maiores acertos para os *Bons* clientes (72,1% contra 62,5%). O modelo que usa as informações dos rejeitados apresentou maiores acertos para os *Maus* clientes (65% contra 56,4%) e acertos total de 67,3% contra 65,0%, apresentando melhor acurácia.

Capítulo 7

Conclusões e Considerações Finais

Nesta dissertação apresentamos algumas estratégias de desenvolvimento de modelos de *credit score* com a utilização das informações dos proponentes rejeitados. Geralmente para a construção de modelos de crédito, utilizam-se somente as informações dos proponentes aceitos, uma vez que para os proponentes rejeitados não se pode avaliar o seu desempenho de crédito.

Em uma primeira aplicação, foram apresentadas e aplicadas quatro estratégias de desenvolvimento de modelos de *credit score* com a utilização das informações de rejeitados. Primeiramente ajustamos um modelo sem o uso das informações dos rejeitados. Este modelo serviu de referência para comparação com os demais modelos. Para os outros modelos foram aplicadas técnicas de *inferência de rejeitados*, conhecidas como *classificação*, *parcelamento*, *dados aumentados* e *informações de mercado*. Nesta última é necessário o auxílio de uma central de crédito para o fornecimento das informações de comportamento de mercado.

Seus desempenhos foram avaliados e comparados através de uma aplicação em uma base de dados do mercado de crédito, utilizando indicadores de desempenhos mais difundidos pelo mercado, como KS, AEC, ROC e Gini. O modelo desenvolvido com informações de mercado apresentou, maiores valores para os indicadores de desempenho, sugerindo que o uso desta metodologia apresenta ganhos superiores em relação a acertividade do modelo sem os rejeitados e com as demais técnicas

utilizadas.

Uma outra estratégia de desenvolvimento foi avaliada em uma segunda aplicação, utilizando-se uma *amostra controle*, em que todas as propostas em um determinado período são aceitas e uma amostra de rejeitados é avaliada. Depois de um período, esses clientes são acompanhados, segundo o seu comportamento de crédito. Quando aceitamos os rejeitados, isto é, propostas com forte potencial de inadimplência, pode haver um acréscimo do risco, porém, alguns clientes que seriam rejeitados podem não se tornar indadimplentes, ocasionando ganho para a instituição. Os organismos de crédito escolhem o grupo controle tentando limitar este risco. Segundo Weldon (2000) essa técnica de tratamento de rejeitados é a única que, na prática, controla o risco, olhando o benefício atingido.

Na Aplicação 1, o modelo com informações de mercado apresentou maiores valores para os indicadores de desempenho do que as demais estratégias.

Com o uso da estratégia de utilização de uma amostra contole na Aplicação 2, com a prática de aceitar uma amostra de proponentes rejeitados no desenvolvimento de modelos de *credit score*, observou-se que também há um ganho no desempenho do novo modelo construído sobre o modelo que utiliza apenas os clientes aceitos. Este método possibilita o controle do risco olhando o benefício atingido, pois ao aceitarmos uma amostra de rejeitados podemos avaliar o seu real desempenho.

Deve-se resaltar que o uso da técnica a ser utilizada depende de como foi tomada a decisão anterior, de como se deu o processo de crédito, dependendo também da disponibilidade da amostra e do custo de aquisição de informações obtidas em fontes externas.

Como sugestões para trabalhos futuros recomenda-se a avaliação do custo de obtenção das informações dos clientes rejeitados, pois como em algumas situações práticas, as informações sobre os rejeitados podem não estar disponíveis na base de dados, uma vez que a instituição pode não acreditar que tais informações sejam úteis para a melhoria do seu porcesso de crédito.

Outros possíveis tópicos são o estudo da avaliação de técnicas de tratamento

de rejeitados que utilizem a análise de sobrevivência, com o intuito de avaliar o limite inferior do intervalo de confiança da vida mediana dos proponentes rejeitados (Sohn e Shin, 2006), a avaliação do uso combinado de diferentes técnicas aplicadas simultaneamente e a investigação do ganho com o melhor desempenho de classificação do modelo *versus* o custo financeiro para a obtenção de informação.

Apêndice A

Descrição das Variáveis

A.1 Aplicação 1

Cadastro

- Idade: em anos completos.
- Nível Escolar: 1grau, 2grau, superior completo, superior incompleto, pós-graduação.
- Sexo: masculino, feminino.
- Estado Civil: solteiro, casado, viúvo, divorciado, separado, outros.
- Dependentes: número de dependentes.
- Natureza de ocupação: empresário/sócio/proprietário, vive de renda, funcionário público, autônomo, profissional liberal, aposentado/pensionista, outros.

Comportamento de Mercado

- Quantidade de consultas a crédito: número de consultas efetuadas de origem a crédito.

- Quantidade de consultas a cheques: número de consultas efetuadas de origem a cheques.
- *Credit Target Serasa*® - segmentação de atividade de crédito:
  - A1: Alta atividade de crédito e entrada no mercado recente;
  - A2: Alta atividade de crédito com tendência crescente;
  - A3: Alta atividade de crédito com tendência decrescente;
  - M1: Atividade de crédito média e entrada no mercado recente;
  - M2: Atividade de crédito média com tendência crescente;
  - M3: Atividade de crédito média com tendência decrescente;
  - B1: Baixa atividade de crédito;
  - I1: Inativos em crédito e cheques;
  - I2: Inativos em crédito e ativos em cheques.
- Quantidade de cheques pré-datados: número de cheques a vencer/vencidos.
- Quantidade de cheques sustados: número de cheques sustados (furto/roubo e/ou desacordo comercial).
- Quantidade de contas bancárias: número de contas bancárias informadas pelo cliente.

Restrições de Crédito

- Quantidade de cheques devolvidos (CCF): número de cheques devolvidos por Alíneas 12 (segunda devolução).
- Quantidade de PEFIN: número de pendências financeiras resolvidas.
- Protestos: número de títulos protestados resolvidos.

A.2 Aplicação 2

Cadastro

- Sexo: masculino, feminino.
- Idade: em anos completos.
- Estado civil: solteiro, casado, viúvo, outros.
- Tipo de imóvel: casa, apartamento.
- Forma de pagamento: débito automático em conta bancária; cartão de crédito e boleto.
- Tipo de venda: parceiro, televendas.
- Pontos adicionais: número de pontos adicionais.

Restrições de Crédito

- Total de negativações nos últimos três meses: número total de restrições de cheques sem fundos, protestos e pendências financeiras de mercado nos últimos três meses.
- Total de negativações antes dos últimos três meses: número total de restrições de cheques sem fundos, protestos e pendências financeiras de mercado antes dos últimos três meses.

Apêndice B

Tabelas Descritivas

Aplicação 1

Tabela B.1: Distribuição da inadimplência segundo o sexo*.

Sexo	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Feminino	2189 (48)	1843 (41)	4032 (44)	0,84
Masculino	2405 (52)	2616 (59)	5021 (56)	1,09
Total	4594 (100)	4459 (100)	9053 (100)	1,00

* 171 casos foram omitidos para essa variável por falta de informação.

Tabela B.2: Distribuição da inadimplência segundo a idade.

Idade(anos)	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
≤ 26	897 (19)	1081 (23)	1978 (21)	1,21
27 a 35	1186 (26)	867 (19)	2053 (22)	0,73
36 a 45	1066 (23)	887 (19)	1953 (21)	0,83
46 a 55	825 (18)	753 (16)	1578 (17)	0,91
56 a 60	263 (7)	333 (7)	596 (6)	1,27
61 a 65	163 (3)	267 (6)	430 (5)	1,64
66 a 70	108 (2)	168 (4)	276 (4)	1,56
≥ 71	104 (2)	256 (6)	360 (4)	2,46
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.3: Distribuição da inadimplência segundo o nível escolar.

Nível escolar	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Não informado	4 (0)	2 (0)	6 (1)	0,5
1grau	641 (14)	465 (10)	1106 (11)	0,73
2grau incompleto	187 (4)	129 (3)	316 (3)	0,69
2grau completo	1055(23)	697 (15)	1752 (20)	0,66
Superior completo	488 (11)	453 (10)	941 (10)	0,93
Pós-graduação	2237(48)	2866 (62)	5103 (55)	1,28
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.4: Distribuição da inadimplência segundo o estado civil.

Estado civil	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Solteiro	786 (17)	597 (12)	1383 (15)	0,76
Casado	1447 (31)	1327 (30)	2774 (30)	0,92
Separado/divorciado	142 (3)	93 (2)	235 (2)	0,65
Viúvo	84 (2)	78 (2)	162 (2)	0,93
Outros	2153 (47)	2517 (54)	4670 (51)	1,17
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.5: Distribuição da inadimplência segundo o número de dependentes.

Dependentes	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Sem dependentes/não informado	4196 (91)	4399 (95)	8595 (93)	1,05
1 a 2	303 (7)	167 (4)	470 (5)	0,55
≥ 3	113 (2)	46 (1)	159 (2)	0,41
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.6: Distribuição da inadimplência segundo a natureza de ocupação.

Natureza	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Autônomo	121 (3)	71 (1)	192 (2)	0,59
Vive de renda	117 (2)	79 (2)	196 (2)	0,68
Empregado	1742 (38)	1264 (27)	3006 (32)	0,73
Funcionário público	495 (11)	395 (9)	890 (10)	0,8
Profissional liberal	110 (2)	121 (3)	231 (2)	1,1
Sócio/proprietário/Aposentado	464 (10)	589 (13)	1053 (11)	1,27
Outros/não informado	1563 (34)	2093 (45)	3656 (39)	1,34
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.7: Distribuição da inadimplência segundo o *Credit Target Serasa*.

Classes	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
A1	124 (3)	76 (1)	200 (3)	0,61
A2	505 (11)	233 (5)	738 (8)	0,46
A3	1404 (30)	1044 (23)	2448 (26)	0,74
M1	217 (5)	413 (9)	630 (7)	1,9
M2	505 (11)	563 (12)	1068 (11)	1,11
M3	1343 (29)	1319 (29)	2662 (29)	0,98
B1	436 (9)	732 (16)	1168 (13)	1,68
I1	20 (1)	65 (1)	85 (1)	3,25
I2	58 (1)	167 (4)	225 (2)	2,88
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.8: Distribuição da inadimplência segundo a quantidade de contas bancárias.

Quantidade de contas	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	1301 (28)	1717 (37)	3018 (33)	1,32
1	1772 (38)	1723 (37)	3495 (38)	0,97
2	962 (21)	790 (17)	1752 (19)	0,82
≥ 3	577 (12)	382 (9)	959 (10)	0,66
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.9: Distribuição da inadimplência segundo a quantidade de cheques sustados.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	3617 (78)	4019 (87)	7636 (82)	1,11
1	457 (11)	345 (8)	802 (9)	0,75
2 a 3	293 (6)	157 (3)	450 (5)	0,54
≥ 4	245 (5)	91 (2)	336 (4)	0,37
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.10: Distribuição da inadimplência segundo a quantidade de cheques pré-datados.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	3948 (86)	4193 (91)	8141 (89)	1,06
1	294 (6)	209 (4)	503 (5)	0,71
≥ 2	370 (8)	210 (5)	580 (6)	0,57
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.11: Distribuição da inadimplência segundo a quantidade de CCF resolvidos.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	3984 (87)	4428 (96)	8412 (90)	1,11
1	260 (6)	88 (2)	348 (4)	0,34
2	110 (2)	32 (1)	142 (2)	0,29
3	61 (1)	21 (1)	82 (1)	0,34
≥ 4	197 (4)	43 (1)	240 (3)	0,22
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.12: Distribuição da inadimplência segundo a quantidade de protestos resolvidos.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	4282 (93)	4503 (97)	8785 (95)	1,05
1 a 2	254 (6)	90 (3)	344 (5)	0,35
3	26 (2)	6 (0)	32 (1)	0,23
≥ 4	50 (1)	13 (0)	63 (1)	0,26
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.13: Distribuição da inadimplência segundo a quantidade de pendências financeiras resolvidas.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	3048 (66)	4158 (90)	7206 (78)	1,36
1 a 2	873 (19)	316 (7)	1189 (13)	0,36
≥ 3	691 (15)	138 (3)	829 (9)	0,2
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.14: Distribuição da inadimplência segundo a quantidade de consultas a cheques.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	1324 (29)	1851 (40,1)	3175 (34,4)	1,4
1	592 (13)	614 (13,3)	1206 (13,1)	1,04
2 a 3	809 (17)	703 (15,2)	1512 (16,4)	0,87
4 a 6	675 (15)	545 (11,8)	1220 (13,2)	0,81
≥ 7	1212 (26)	899 (19,5)	2111 (22,9)	0,74
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Tabela B.15: Distribuição da inadimplência segundo a quantidade de consultas a crédito.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	990 (21)	1596 (35)	2586 (28)	1,61
1	820 (18)	1022 (22)	1842 (20)	1,25
2	638 (14)	642 (14)	1280 (14)	1,01
3	479 (10)	421 (9)	900 (11)	0,88
4	364 (9)	259 (5)	623 (7)	0,71
5	259 (5)	170 (4)	429 (5)	0,66
6 a 8	510 (11)	273 (6)	783 (9)	0,54
≥ 9	552 (12)	229 (5)	781 (9)	0,41
Total	4612 (100)	4612 (100)	9224 (100)	1,00

Aplicação 2

Tabela B.16: Distribuição da inadimplência segundo o sexo.

Sexo	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Feminino	1427 (33)	1344 (31)	2771 (32)	1,03
Masculino	2870 (67)	2965 (69)	5835 (68)	0,94
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.17: Distribuição da inadimplência segundo a idade.

Idade(anos)	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
18 a 25	998 (23)	526 (12)	1522 (18)	0,53
26 a 30	829 (19)	646 (15)	1475 (17)	0,78
31 a 40	1229(29)	1176(27)	2405 (28)	0,95
41 a 50	718 (17)	978 (23)	1696 (20)	1,36
≥ 51	533 (12)	982 (23)	1519 (17)	1,84
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.18: Distribuição da inadimplência segundo o estado civil.

Estado Civil	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Solteiro	821 (19)	685 (16)	1506 (17)	0,83
Casado	2557(60)	2797(65)	5354 (63)	1,09
Outros	920 (21)	827 (19)	1747 (20)	0,90
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.19: Distribuição da inadimplência segundo o tipo de venda.

Venda	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Parceiro	3218 (75)	2585 (60,0)	5803 (67)	0,80
Televendas	1079 (25)	1724 (40,0)	2803 (33)	1,60
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.20: Distribuição da inadimplência segundo o tipo de imóvel.

Imóvel	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Casa	3807 (89)	3447 (80,0)	7254 (84)	0,90
Apartamento	490 (11)	862 (20,0)	1352 (16)	1,75
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.21: Distribuição da inadimplência segundo a aquisição de pontos adicionais.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
Principal	4142 (96)	3645 (85)	7787 (91)	0,88
1	120 (3)	414 (10)	534 (6)	3,43
2	21 (1)	138 (2)	159 (1)	8,33
3	14 (1)	112 (3)	126 (2)	6,4
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.22: Distribuição da inadimplência total de negativações nos últimos três meses.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	4073 (95)	4210 (97)	8283 (96)	1,03
1 a 2	129 (3)	60 (2)	189 (2)	0,47
≥ 3	95 (2)	39 (1)	134 (2)	0,41
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Tabela B.23: Distribuição da inadimplência total de negativações antes dos últimos três meses.

Quantidade	*Maus*(%)	*Bons*(%)	Total	*Bons*(%)/*Maus*(%)
0	3820 (89)	4020 (93)	7840 (91)	1,05
1 a 2	116 (3)	69 (2)	185 (2)	0,59
≥ 3	361 (8)	220 (5)	581 (7)	0,61
Total	4297 (100)	4309 (100)	8606 (100)	1,0

Apêndice C

Parâmetros Estimados

Tabela C.1: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, sem o uso dos clientes rejeitados - Modelo I.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,0194	0,0806	<0,0001
ESTCIVIL1	-0,2676	0,1537	0,0817
ESTCIVIL2	-0,1717	0,0708	0,0153
ESTCIVIL3	-0,1243	0,0588	0,0345
DEPEND1	-0,2814	0,1133	0,013
DEPEND2	-0,6293	0,1944	0,0012
RENDA1	-0,2445	0,071	0,0006
RENDA4	0,2738	0,0906	0,0025
RENDA5	0,4076	0,1382	0,0032
RENDA6	0,4319	0,206	0,036
IDADE1	-0,3191	0,0681	<0,0001
IDADE2	-0,1798	0,0696	0,0098
IDADE3	-0,198	0,0735	0,007
IDADE5	0,271	0,1193	0,0231
IDADE7	0,3904	0,1354	0,0039
UF1	-0,9748	0,2572	0,0002
UF2	-0,5899	0,0864	<0,0001
UF3	-0,2963	0,0679	<0,0001
UF4	-0,222	0,112	0,0475
UF6	0,376	0,1348	0,0053
SEXOF	-0,2589	0,0487	<0,0001
SEGM1	-0,3236	0,094	0,0006
SEGM6	0,4894	0,1027	<0,0001
SEGM7	0,184	0,0852	0,0308
SEGM8	0,5728	0,1729	0,0009
SEGM9	0,5258	0,2735	0,0546
NATUREZA1	-0,3811	0,1696	0,0247
NATUREZA2	-0,5693	0,167	0,0007
NATUREZA3	-0,3297	0,0589	<0,0001
NATUREZA4	-0,2576	0,0853	0,0025
QTBANCOS1	-0,2218	0,0654	0,0007
CHPREVENCIDOS1	-0,3307	0,0801	<0,0001
CHPREVENCIDOS2	-0,3436	0,0765	<0,0001
CHSUSTADOS2	-0,2622	0,1154	0,0231
CHSUSTADOS3	-0,4105	0,1419	0,0038
TOTNEGU6M1	-1,4835	0,0756	<0,0001
TOTNEGU6M	-1,7767	0,0873	<0,0001

Tabela C.2: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, assumindo os rejeitados como clientes *Maus* - Modelo II.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,0296	0,0443	< 0,0001
ESTCIVIL3	0,0864	0,0394	0,0283
DEPEND1	-0,3456	0,0823	< 0,0001
DEPEND2	-0,6326	0,1333	< 0,0001
RENDA1	-0,3698	0,0435	< 0,0001
RENDA4	0,2768	0,0693	< 0,0001
IDADE1	-0,3481	0,0397	< 0,0001
IDADE4	0,4138	0,072	< 0,0001
IDADE5	0,6233	0,0847	< 0,0001
IDADE7	0,732	0,0961	< 0,0001
UF1	-1,22	0,1745	< 0,0001
UF2	-0,6886	0,0551	< 0,0001
UF3	-0,5296	0,045	< 0,0001
SEXOF	-0,1644	0,0332	< 0,0001
SEGM1	-0,4011	0,0684	< 0,0001
SEGM5	-0,3634	0,0498	< 0,0001
SEGM6	0,5875	0,0722	< 0,0001
SEGM7	-0,1797	0,0534	0,0008
QTBANCOS1	-0,1662	0,0412	< 0,0001
CCF1	-0,593	0,0705	< 0,0001
CCF2	-1,2338	0,1083	< 0,0001
CCF3	-1,4822	0,1066	< 0,0001
CCF4	-1,9952	0,0781	< 0,0001
REFIN1	-1,072	0,0585	< 0,0001
REFIN3	-1,693	0,0789	< 0,0001
REFIN4	-2,0034	0,1229	< 0,0001
PROTESTO1	-0,6234	0,0977	< 0,0001
PROTESTO3	-0,8588	0,2668	0,0013
CONSULTASCHEQUES1	-0,3834	0,0385	< 0,0001
CHSUSTADOS2	-0,2424	0,0825	0,0033
CHSUSTADOS3	-0,5551	0,0988	< 0,0001

Tabela C.3: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, utilizando a técnica de *inferência de rejeitados* de *parcelamento* - Modelo III.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,0249	0,0525	<0,0001
ESTCIVIL1	-0,9371	0,3459	0,0067
ESTCIVIL3	-0,1156	0,0477	0,0154
ESTCIVIL4	-0,312	0,1335	0,0194
DEPEND1	-0,2463	0,0695	0,0004
DEPEND2	-0,3798	0,2192	0,0832
RENDA3	0,2814	0,0638	<0,0001
IDADE2	-0,1042	0,0404	0,0098
IDADE4	0,087	0,0504	0,0844
IDADE5	0,4798	0,0582	<0,0001
UF1	-0,6796	0,146	<0,0001
UF2	-0,3718	0,0698	<0,0001
UF3	-0,2692	0,0428	<0,0001
UF4	-0,2501	0,0986	0,0112
UF6	0,0952	0,0508	0,0611
SEXOF	-0,0726	0,0334	0,0296
SEGM1	-0,1865	0,0386	<0,0001
SEGM2	0,2056	0,1232	0,0953
SEGM3	-0,2605	0,0676	0,0001
SEGM7	0,2498	0,1152	0,0302
SEGM8	0,5657	0,0778	<0,0001
NIVESCO2	-0,3585	0,0426	<0,0001
NIVESCO3	-0,3665	0,0495	<0,0001
QTBANCOS1	0,3791	0,0428	<0,0001
QTBANCOS2	-0,0935	0,0427	0,0287
QTBANCOS3	-0,3186	0,0967	0,001
CONSULTASCHEQUES1	-0,2148	0,0378	<0,0001
CONSULTASCREDITO4	-0,2806	0,0548	<0,0001
CONSULTASCREDITO5	-0,2841	0,0592	<0,0001
CONSULTASCREDITO6	-0,1742	0,088	0,0477
CHPREVENCIDOS1	-0,0833	0,0489	0,0884
CHSUSTADOS2	-0,26	0,0789	0,001
CHSUSTADOS3	-0,469	0,1385	0,0007
CHSUSTADOS4	-0,6219	0,121	<0,0001
CHPREAVENCER1	0,1947	0,0797	0,0145
TOT2U6M1	-1,3065	0,0466	<0,0001
TOT2U6M2	-1,6086	0,0534	<0,0001
TOT2U6M3	-1,8014	0,0599	<0,0001
TOT2U6M4	-1,9348	0,0886	<0,0001

Tabela C.4: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, utilizando a técnica de *inferência de rejeitados* de *dados aumentados* - Modelo IV.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	0,7026	0,0743	$<$ 0,0001
FONE1	-0,3656	0,0472	$<$ 0,0001
ESTCIVIL1	-0,2740	0,1000	0,0061
ESTCIVIL2	-0,1283	0,0478	0,0073
ESTCIVIL3	-0,1080	0,0403	0,0073
ESTCIVIL4	-0,3745	0,1392	0,0071
DEPEND1	-0,5608	0,0806	$<$ 0,0001
DEPEND2	-0,6117	0,1247	$<$ 0,0001
RENDA1	-0,2384	0,0442	$<$ 0,0001
RENDA3	0,2102	0,0445	$<$ 0,0001
RENDA4	0,4155	0,0673	$<$ 0,0001
RENDA5	0,6460	0,1091	$<$ 0,0001
IDADE1	-0,4778	0,0486	$<$ 0,0001
IDADE2	-0,4466	0,0503	$<$ 0,0001
IDADE3	-0,2834	0,0543	$<$ 0,0001
IDADE4	-0,1525	0,0739	0,0391
IDADE5	0,1844	0,0918	0,0446
IDADE7	0,3862	0,1087	0,0004
UF1	-0,7755	0,1327	$<$ 0,0001
UF2	-0,393	0,0628	$<$ 0,0001
UF3	-0,1355	0,0543	0,0125
UF4	0,3281	0,0595	$<$ 0,0001
UF5	-0,2199	0,0657	0,0008
UF8	0,2746	0,0472	$<$ 0,0001
SEGM1	-0,4379	0,0558	$<$ 0,0001
SEGM6	0,4828	0,0789	$<$ 0,0001
NIVESCO1	-0,404	0,0947	$<$ 0,0001
NIVESCO2	-0,359	0,0418	$<$ 0,0001
GBANCO1	-0,1481	0,0566	0,0089
QTBANCOS2	0,1175	0,0443	0,0079
QTBANCOS3	-0,1314	0,0621	0,0344
CONSULTASCREDITO	-0,1355	0,0655	0,0386
CHPREVENCIDOS1	-0,2721	0,0569	$<$ 0,0001
CHPREVENCIDOS3	-0,2811	0,0937	0,0027
CHPREAVENCER3	-0,4283	0,1839	0,0198
CHSUSTADOS2	-0,1954	0,0771	0,0113
CHSUSTADOS4	-0,2027	0,0846	0,0166
TOTNEGU6M1	-1,2962	0,0499	$<$ 0,0001
TOTNEGU6M2	-1,4017	0,0529	$<$ 0,0001
TOTNEGU6M3	-1,7422	0,0731	$<$ 0,0001

Tabela C.5: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, utilizando a técnica de *inferência de rejeitados* de *uso de informações de mercado* - Modelo V.

Parâmetro	Estimativa	Erro Padrão	Valor p
Intercepto	1,0463	0,1541	<0,0001
SEXOF	-0,2646	0,0455	<0,0001
FONE1	-0,4435	0,0675	<0,0001
ESTCIVIL2	-0,2459	0,0669	0,0002
ESTCIVIL3	-0,6179	0,1828	0,0007
DEPEND1	-0,3681	0,1073	0,0006
DEPEND2	-0,6328	0,1598	<0,0001
RENDA1	0,1906	0,0867	0,0279
RENDA2	0,2177	0,0682	0,0014
RENDA3	0,4209	0,0669	<0,0001
RENDA4	0,3941	0,0997	<0,0001
RENDA5	0,5696	0,1382	<0,0001
IDADE2	-0,2364	0,0707	0,0008
IDADE3	-0,1832	0,0595	0,0021
IDADE8	0,5114	0,2055	0,0128
UF2	-0,6687	0,1326	<0,0001
UF3	-0,4087	0,0756	<0,0001
UF4	-0,475	0,0797	<0,0001
UF5	-0,1892	0,0708	0,0075
UF6	-0,2962	0,0734	<0,0001
UF8	0,4263	0,0677	<0,0001
SEGM1	-0,338	0,0855	<0,0001
SEGM3	0,2884	0,0695	<0,0001
SEGM6	0,23	0,1224	0,0603
SEGM7	0,3045	0,1047	0,0036
NIVESCO1	-0,3193	0,0802	<0,0001
NIVESCO2	-0,3462	0,0734	<0,0001
NIVESCO4	-0,2916	0,1307	0,0257
NATUREZA1	-0,1961	0,0858	0,0223
NATUREZA2	-0,2319	0,0716	0,0012
NATUREZA3	-0,2977	0,1533	0,0521
NATUREZA6	-0,3055	0,132	0,0206
GBANCO1	0,2255	0,0528	<0,0001
QTBANCOS2	-0,1216	0,0575	0,0345
REFIN1	0,6257	0,0574	<0,0001
REFIN2	-0,4365	0,1695	0,01
PROTESTO	0,7685	0,113	¡,0001
CONSULTASCHEQUES2	-0,1862	0,076	0,0143
CONSULTASCHEQUES3	-0,1552	0,0876	0,0763
CONSULTASCHEQUES4	-0,2012	0,0568	0,0004
CHPREVENCIDOS1	-0,1571	0,0888	0,077
CHPREVENCIDOS3	-0,3561	0,1167	0,0023
CHPREVENCIDOS4	-0,202	0,0844	0,0167
CHSUSTADOS1	0,1673	0,06	0,0053
TOTU6M1	-1,0041	0,063	<0,0001
TOTU6M2	-1,3996	0,0922	<0,0001
TOTU6M3	-1,6214	0,0985	<0,0001
TOTU6M4	-1,5864	0,1076	<0,0001

Tabela C.6: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, sem o uso das informações dos rejeitados - Modelo VI.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,944	0,306	<0,001
Cartaocredito1	1,168	0,182	<0,001
Cartaocredito2	0,334	0,044	<0,001
Cartaocredito3	0,316	0,159	0,048
Banco1	0,225	0,064	<0,001
Banco2	-0,269	0,063	<0,001
Banco3	0,028	0,102	0,783
Banco4	0,462	0,088	<0,001
Banco5	0,127	0,061	0,036
Mirror1	0,949	0,086	<0,001
Mirror2	1,081	0,201	<0,001
Mirror3	1,614	0,206	<0,001
FoneCel	-0,24	0,053	<0,001
FoneCom	-0,442	0,053	<0,001
Idade1	-1,635	0,232	<0,001
Idade2	0,153	0,099	0,123
Idade3	-1,033	0,08	<0,001
Idade4	-0,623	0,079	<0,001
Idade5	-0,408	0,072	<0,001
Idade6	-0,093	0,077	0,223
PDV	-0,662	0,046	<0,001
PEFIN	-1,08	0,395	0,006
SexoF	-0,135	0,044	0,002
TipoInstalacao1	-0,492	0,113	<0,001
TOTNEGA3M1	-0,476	0,194	0,014
TOTNEGA3M2	0,117	0,176	0,505
TOTNEGU3M1	1,414	0,52	0,006
TOTNEGU3M2	0,651	0,307	0,034

Tabela C.7: Estimativas, erros padrões e valores p dos parâmetros do modelo de *credit score*, com o uso das informações dos rejeitados, obtido a partir da amostra de controle - Modelo VII.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	-0,223	0,299	0,456
CCF0	0,577	0,215	0,007
ChurnGrupo01	0,110	0,279	0,697
ChurnGrupo02	0,080	0,264	0,763
ChurnGrupo03	-0,010	0,242	0,958
ChurnGrupo04	-0,520	0,248	0,037
ChurnGrupo05	-0,810	0,316	0,011
Banco1	0,201	0,085	0,019
Banco2	-0,059	0,085	0,485
Banco3	0,044	0,128	0,728
Banco4	0,479	0,117	<0,001
Banco5	0,259	0,085	0,002
Cartao1	0,689	0,000	<0,001
Cartao2	0,342	0,06	<0,001
Cartao3	1,574	0,324	<0,001
Mirror1	1,104	0,133	<0,001
Mirror2	1,951	0,334	<0,001
Mirror3	1,778	0,292	<0,001
Idade1	-1,949	0,268	<0,001
Idade2	-0,037	0,143	0,799
Idade3	-1,137	0,118	<0,001
Idade4	-0,828	0,116	<0,001
Idade5	-0,68	0,107	<0,001
Idade6	-0,268	0,113	0,017
PDV1	-0,800	0,062	<0,001
UF1	0,520	0,292	0,075
UF2	0,138	0,275	0,618
UF3	0,755	0,217	0,001
UF4	1,231	0,247	<0,001
TipoImovel1	0,631	0,080	<0,001
TOTNEG1	-0,435	0,228	0,057
TOTNEG0	0,264	0,201	0,188

Apêndice D

Macros SAS

Macro DTI - Diferença entre Taxas de Inadimplência;

Tomazela (2007).

```
**************************************************************************/
%MACRO DTI (ARQENT, ARQSAI, NUM);
**************************************************************************/
Esta macro serve para calcular as medidas de desempenho sugerida
como diferença entre as taxas de inadimplentes para os decis extremos
Para utilizar a macro será necessário renomear as variáveis para:

OBS: Identificador de cada observação;
RESP: Variável resposta, com 0 = Bom E 1 = Mau;
SCORE: escore final atribuido a cada observação em inteiros;

Parâmetros da Macro:
ARQUENT: é o nome do arquivo SAS que deve conter:
ARQSAI: é o nome do arquivo SAS que deverá gravar a saida já com as medidas calculadas.
NUM: é a variável indicadora de reamostragem. Para utilizar em apenas uma amostra colocar
o número 1, para amostras maiores colocar o número de reamostragem.
**************************************************************************/

PROC UNIVARIATE DATA = &ARQENT.;
VAR SCORE;
OUTPUT OUT = SAIDA PCTLPTS = 10 TO 100 BY 10 PCTLPRE = POP_;
PROC TRANSPOSE DATA = SAIDA OUT=DATA;
PROC PRINT;
RUN;
```

```
DATA APP_MIX0;
SET &ARQENT.;
RUN;

PROC SORT DATA=APP_MIX1; BY SCORE RESP; RUN;

PROC TRANSPOSE DATA=APP_MIX1 OUT=APP_MIX2(DROP= _NAME_ );
BY SCORE;
VAR N;
ID RESP;
RUN;

DATA APP MIX2 ; SET APP MIX2;
IF _0 = . THEN _0 = 0;
IF _1 = . THEN _1 = 0;
RUN;

PROC SQL;
CREATE TABLE APP_MIX3 AS SELECT A.* , B.COL1
FROM APP_MIX2 A, DATA B;
QUIT;

DATA APP_MIX3 ; SET APP_MIX3;
IF SCORE GT COL1 THEN DELETE;
RUN;

PROC SORT DATA=APP_MIX3 ; BY COL1 ; RUN;
PROC MEANS DATA = APP_MIX3 NOPRINT;
BY COL1 ;
VAR _0 _1 ;
OUTPUT OUT = APP_MIX4
SUM = SUM_0 SUM_1 ;
RUN;

DATA APP_MIX4 ; SET APP_MIX4;
DEC_1= SUM_1 - LAG(SUM_1 );
DEC_0= SUM_0 - LAG(SUM_0 );
IF _N_ EQ 1 THEN DO; DEC_0 = SUM_0; DEC_1 = SUM; END;

BAD_RATE = (DEC_1/(SUM(DEC_1 , DEC_0 ) ) ) * 100;
RUN;

PROC TRANSPOSE DATA=APP_MIX4 OUT=APP_MIX5(DROP= _NAME_ );
VAR BAD_RATE;
RUN;
```

```
PROC PRINT DATA = APP_MIX4;
RUN;

DATA &ARQSAI. ; SET APP_MIX5;
DIF_DEC = ABS(COL1 - COL10);
RUN;

PROC PRINT DATA = &ARQSAI.;
RUN;

%MEND DTI;
```

Macro AEC - Área Entre Curvas;

Tomazela (2007).

```
**************************************************************************/
%MACRO AEC (ARQENT, ARQSAI, NUM);
**************************************************************************/
Esta macro serve para calcular as medidas de desempenho sugerida
como diferença entre as taxas de inadimplentes para os decis extremos
Para utilizar a macro será necessário renomear as variáveis para:

OBS: Identificador de cada observação;
RESP: Variável resposta, com 0 = Bom E 1 = Mau;
SCORE: escore final atribuido a cada observação em inteiros;

Parâmetros da Macro:
ARQUENT: é o nome do arquivo SAS que deve conter:
ARQSAI: é o nome do arquivo SAS que deverá gravar a saida já com as medidas calculadas.
NUM: é a variável indicadora de reamostragem. Para utilizar em apenas uma amostra colocar
o número 1, para amostras maiores colocar o número de reamostragem.
**************************************************************************/

DATA APP_MIX0; SET &ARQENT.; RUN;

PROC TABULATE DATA = APP_MIX0 MISSING OUT = APP_MIX1(DROP= _TYPE_ _PAGE_ _TABLE_);
CLASS RESP SCORE;
TABLE SCORE, RESP*N;
RUN;

PROC SORT DATA=APP_MIX1; BY SCORE RESP; RUN;

PROC TRANSPOSE DATA=APP_MIX1 OUT = APP_MIX2(DROP = _NAME_ );
BY SCORE;
VAR N;
ID RESP;
RUN;
```

```
DATA APP_MIX2; SET APP_MIX2;
IF _0 = . THEN _0 = 0;
IF _1 = . THEN _1 = 0;
RUN;

PROC SORT DATA=APP_MIX2; BY DESCENDING SCORE; RUN;

DATA APP_MIX2; SET APP_MIX2;
RETAIN ACUM_0;
ACUM_0 = SUM(ACUM_0 , _0 );
RETAIN ACUM_1;
ACUM_1 = SUM(ACUM_1 , _1 );
TOT = SUM( _0 , _1 );
RUN;

DATA APP_MIX2; SET APP_MIX2;
RETAIN ACUM_T;
ACUM_T = SUM(ACUM_T,TOT);
B= LAG(SCORE);
T= LAG(ACUM_T);
IF _N_ EQ 1 THEN DO;
B = SCORE;
T= ACUM_T;
END;
RUN;

DATA APP_MIX2; SET APP_MIX2;
BASE_S = SCORE - B;
BASE_T = ACUM_T - T;
RUN;

PROC SQL; CREATE TABLE MAXIMO AS
SELECT MAX(ACUM_0) AS MAX_0 ,
MAX(ACUM_1) AS MAX_1 ,
MAX(ACUM_T) AS MAX_T
FROM APP_MIX2;
QUIT;

PROC SQL;
CREATE TABLE APP_MIX3 AS SELECT *
FROM APP_MIX2, MAXIMO;
QUIT;

DATA APP_MIX3; SET APP_MIX3;
PCT_AC_0 = (ACUM_0/MAX_0);
```

```
PCT_AC_1 = (ACUM_1/MAX_1);
PCT_AC_T = (ACUM_T/MAX_T);
PCT_T = (BASE_T/MAX_T);
RUN;

DATA TEST1; SET APP_MIX3;
IF _N_ EQ 1 THEN AREAM = PCT_T*(PCT_AC_1/ 2 );
ELSE AREAM = PCT_T*( (PCT_AC_1 + LAG(PCT_AC_1) ) / 2 );
IF _N_ EQ 1 THEN AREAB = PCT_T*(PCT_AC_0/ 2 );
ELSE AREAB= PCT_T*( (PCT_AC_0+LAG(PCT_AC_0) ) / 2 );
RUN;

PROC SQL;
CREATE TABLE AREA AS
SELECT SUM(AREAM) AS AREAM,
SUM(AREAB) AS AREAB
FROM TEST1;
QUIT;
DATA &ARQSAI.; SET AREA;
DIF = ABS(AREAM - AREAB);
RUN;

PROC PRINT DATA = &ARQSAI.;
RUN;

%MEND AEC;
```

Apêndice E

Gráficos

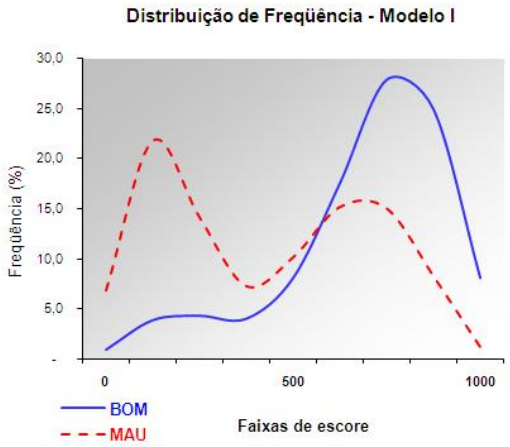

Figura E.1: Distribuição de Freqüência.

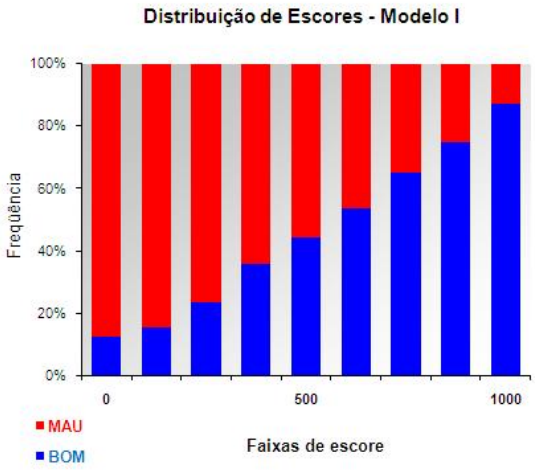

Figura E.2: Distribuição de Escores.

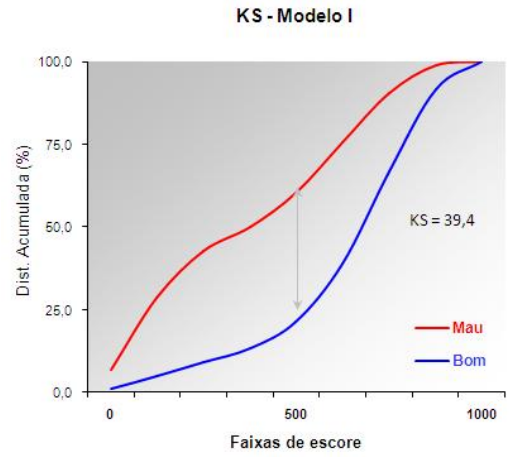

Figura E.3: Curva de KS.

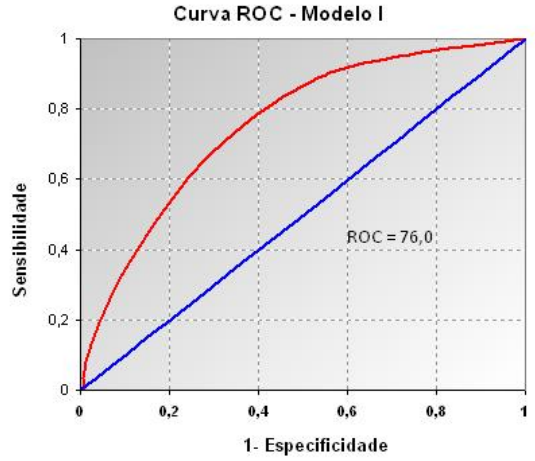

Figura E.4: Curva ROC.

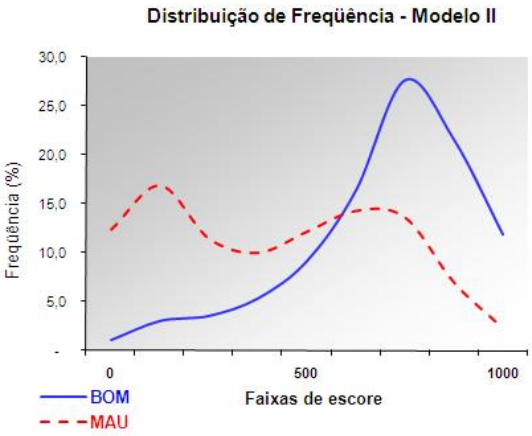

Figura E.5: Distribuição de Freqüência.

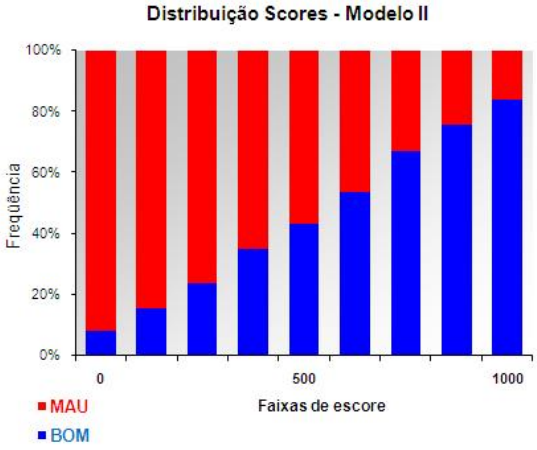

Figura E.6: Distribuição de Escores.

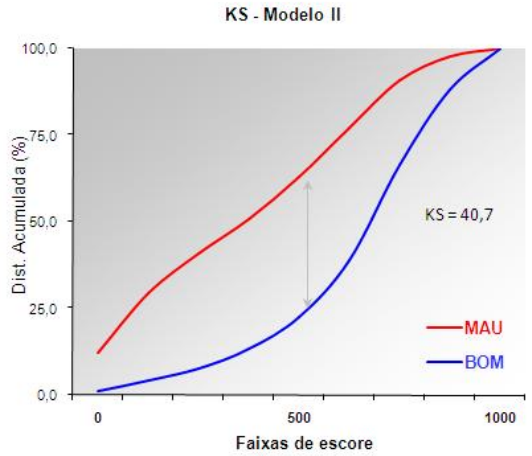

Figura E.7: Curva de KS.

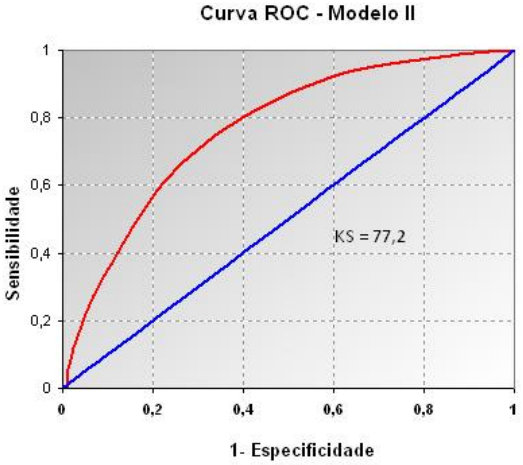

Figura E.8: Curva ROC.

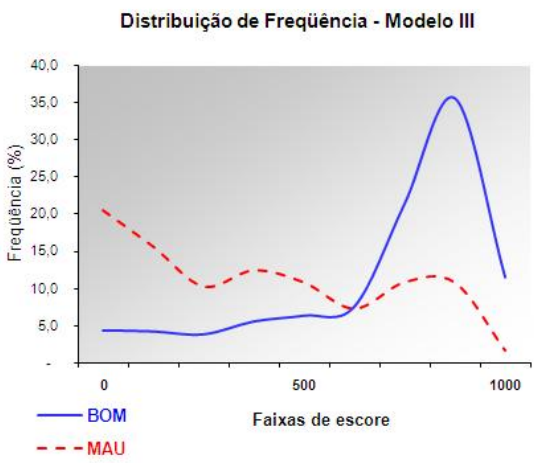

Figura E.9: Distribuição de Freqüência.

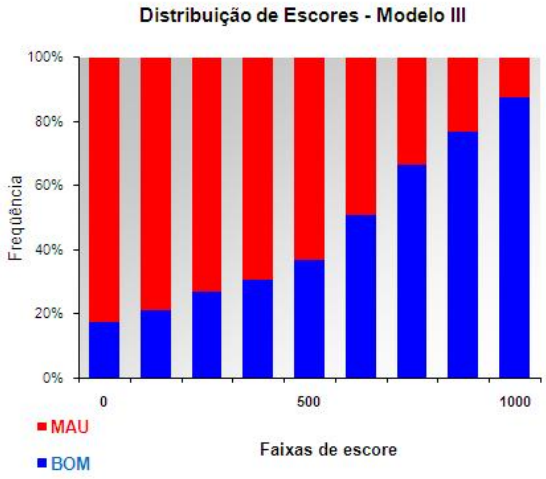

Figura E.10: Distribuição de Escores.

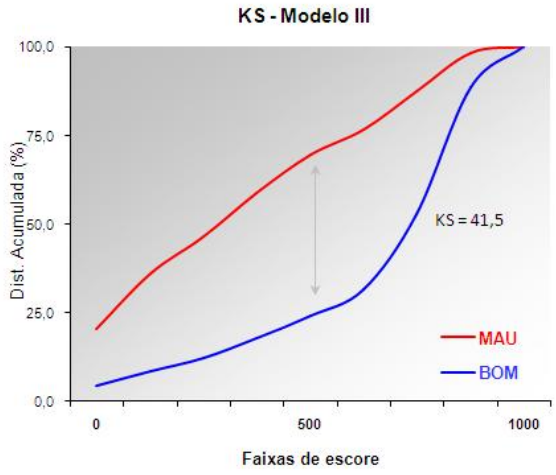

Figura E.11: Curva de KS.

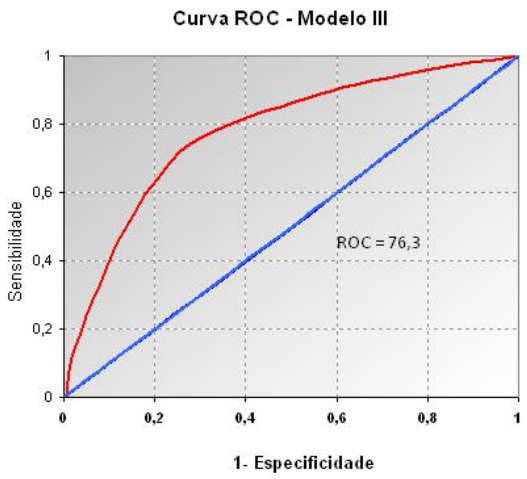

Figura E.12: Curva ROC.

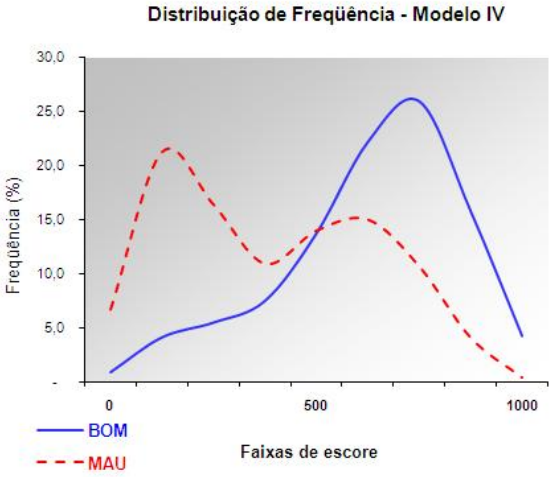

Figura E.13: Distribuição de Freqüência.

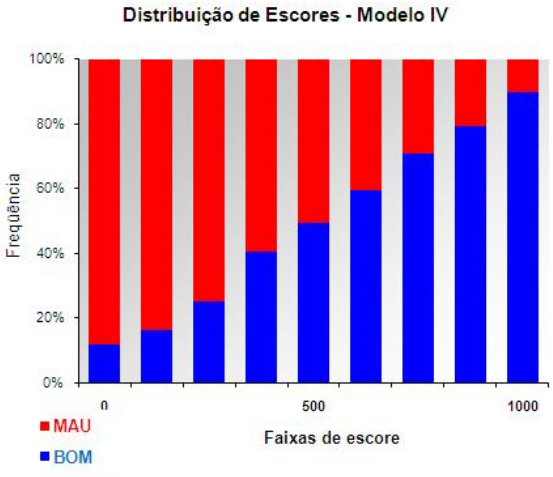

Figura E.14: Distribuição de Escores.

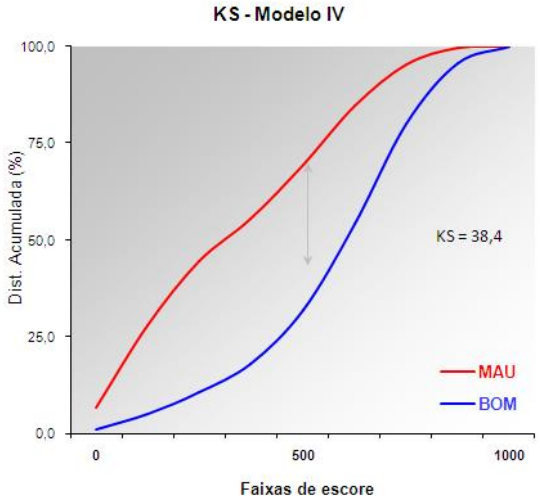

Figura E.15: Curva de KS.

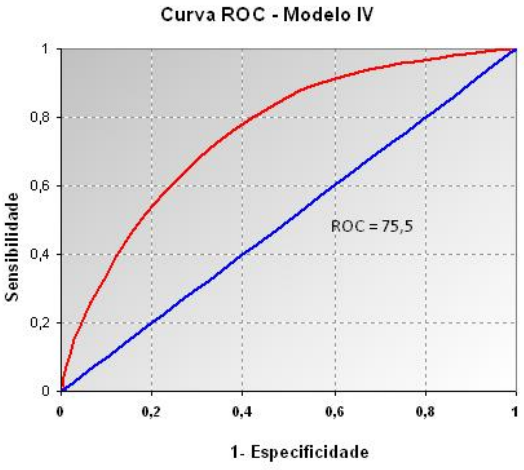

Figura E.16: Curva ROC.

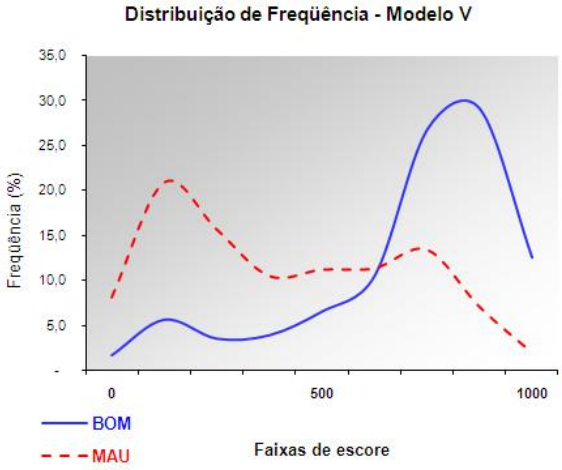

Figura E.17: Distribuição de Freqüência.

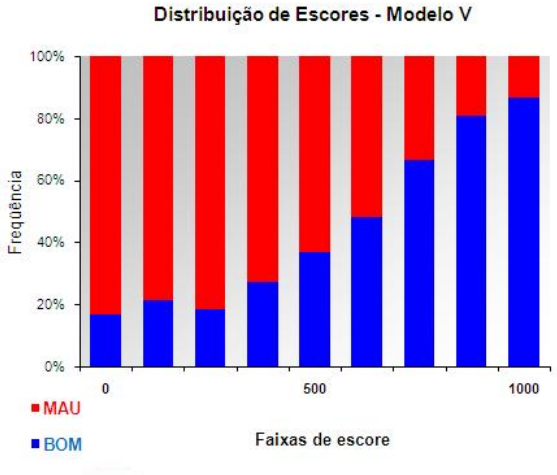

Figura E.18: Distribuição de Escores.

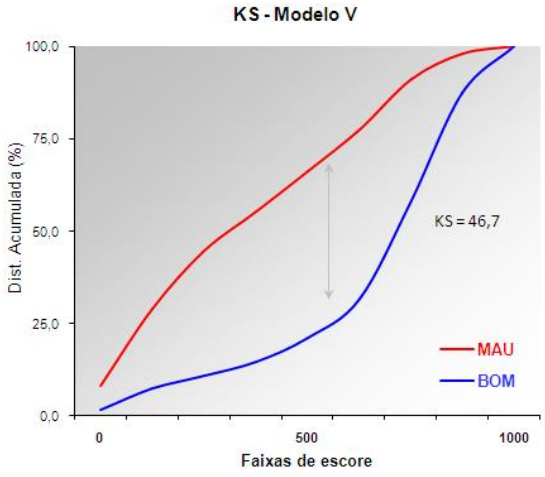

Figura E.19: Curva de KS.

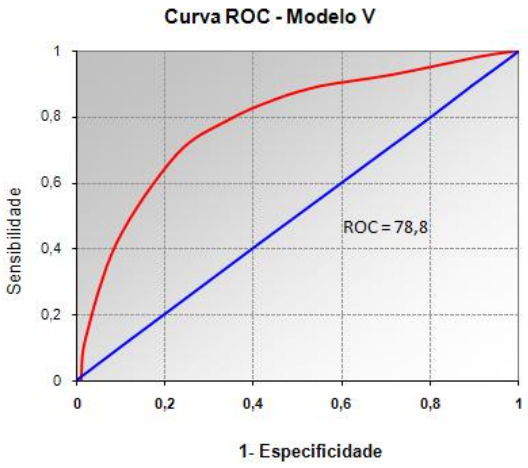

Figura E.20: Curva ROC.

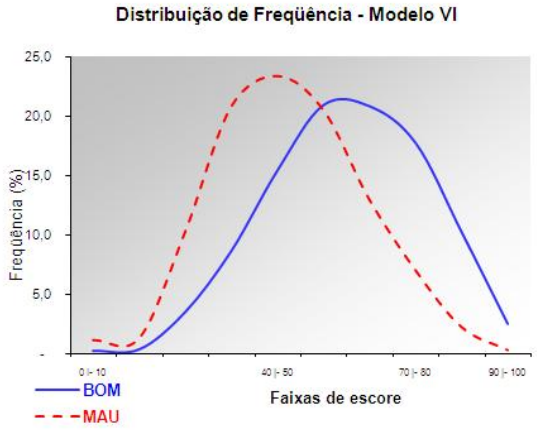

Figura E.21: Distribuição de Freqüência.

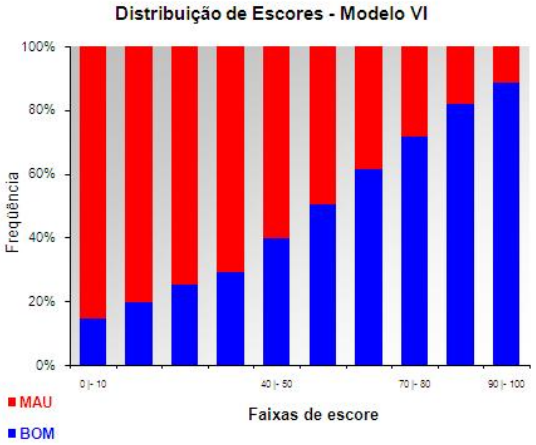

Figura E.22: Distribuição de Escores.

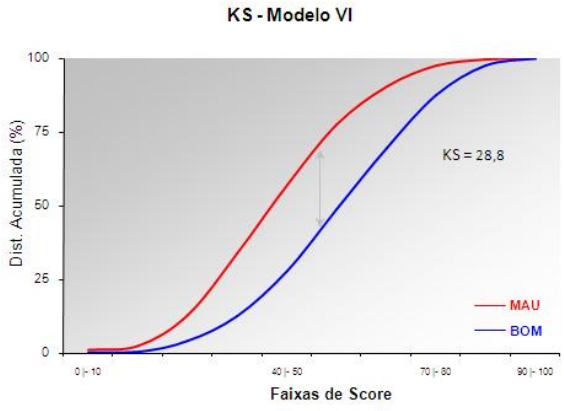

Figura E.23: Curva de KS.

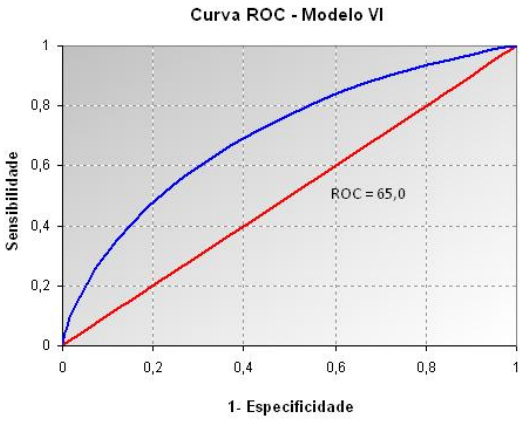

Figura E.24: Curva ROC.

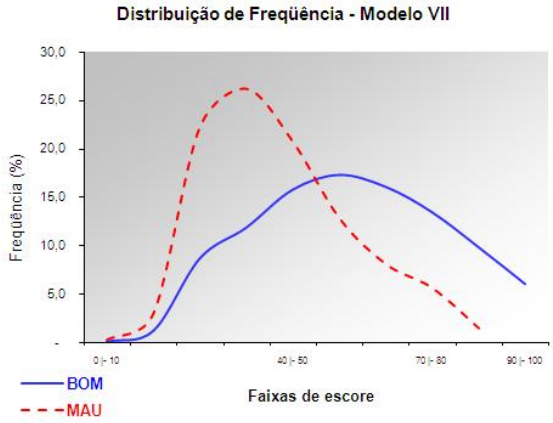

Figura E.25: Distribuição de Freqüência.

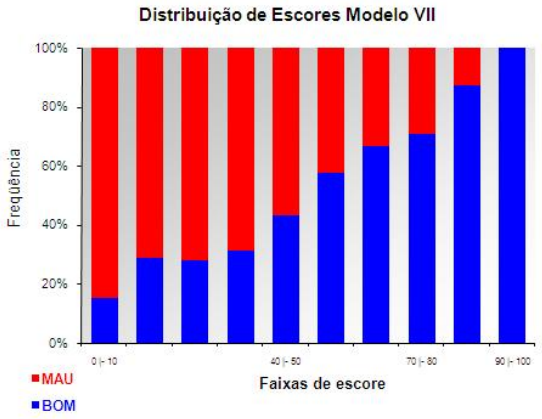

Figura E.26: Distribuição de Escores.

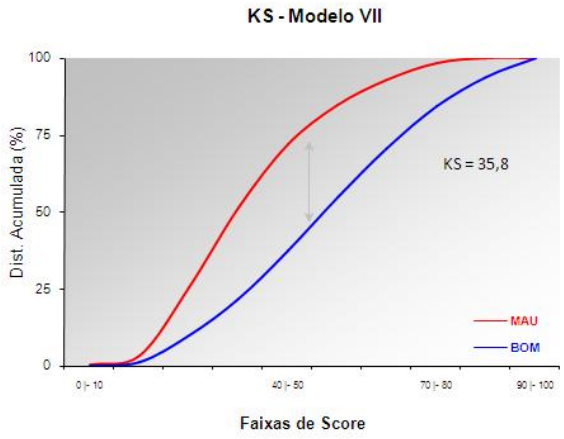

Figura E.27: Curva de KS.

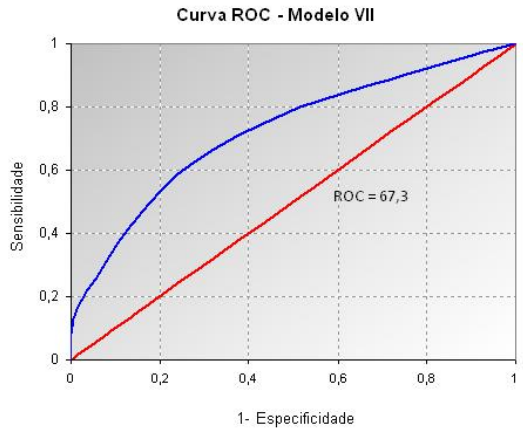

Figura E.28: Curva ROC.

Referências Bibliográficas

[1] Agresti, A. (1990). *Categorical Data Analisys*. New York: Wiley.

[2] Allison, P. D. (1999). *Logistic Regression Using the SAS System: Theory and Application*. Cary, NC: SAS Institute Inc.

[3] Alves, M. C., Andrade, F. W. M. (2004). Contribuição de Informações de Mercado no Poder Preditivo de Modelos de Credit Scoring. *Tecnologia de Crédito*, **42**, 21-30.

[4] Ash, D., Meester, S. (2002). *Best Practices in Reject Inferencing*. Presentation at Credit Risk Modeling and Decisioning Conference, Wharton FIC, University of Pensylvania.

[5] Banasik, J. L., Crook, J. N. (2005). Credit Scoring, Augmentation and Lean Models. *Journal of the Operational Research Society*, **56**, 1072-1091.

[6] Bolfarine, H., Sandoval, M. C. (2002). *Introdução à Inferência Estatística*. São Paulo: Sociedade Brasileira de Matemática.

[7] Chen, M. C., Huang, S. H. (2003). Credit Scoring and Rejected Instances Reassigning through Evolutionary Computation Techniques. *Expert Systems with Applications*, **24**, 433-441.

[8] Conover, W. J. (1999). *Practical Nonparametric Statistics*, 3rd ed. New York: John Wiley and Sons.

[9] Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269-276.

[10] Cox, D. R., Snell, E. J. (1989). *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

[11] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data. *Journal of the Royal Statistical Society*, **39**, 1-38.

[12] Dobson, A. (1983). A Introducition to Statistical Modelling. London: Chapman and Hall.

[13] Feelders, A. J. (2000). Credit Score and Reject Inference with Mixture Models. *International Journal of Intelligent System in Accounting, Finance and Management*, **9**, 1-8.

[14] Good, I. J. (1950). *Probability and the Weighting of Evidence*. London: Charles Griffin.

[15] Hand, D. J., Henley W. E. (1993). Can Reject Inference Ever Work? *IMA Journal of Mathematics Applied in Business and lndustry*, **5**, 45-55.

[16] Hanley, J. A., McNeil, B. J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteistic (ROC) Curve. *Radiology*, **143**, 29-36.

[17] Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: John Wiley & Sons.

[18] Hsai, D. C. (1978). Credit Score and the Equal Credit Opportunity Act. *The Hastings Law Journal*, **30**, 371-448.

[19] Joanes, D. N. (1993). Reject Inference Applied to Logistic Regression for Credit Score. *IMA Journal of Mathematics Applied in Business and Industry*, **5**, 35-43.

[20] Johnson, R. A., Wichern, D. W. (2002). *Multivariate Statistical Analysis*, 5th ed. New York: Prentice Hall.

[21] Kleinbaum, D. G.(1994). *Logistic Regression: A self-learning Text*. New York: Springer.

[22] Morgan, G. A., Griego, V. (1998). *Easy use and Interpretation of SPSS for Windows: Answering Research Questions with Statistics*. New Jersey: Lawrence Erlbaum Associates, Publishers.

[23] Neter, J., Kutner, M. K., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th ed. New York: McGraw-Hill.

[24] Ogava, M. H. (2007). *Redes Neurais em Análise de Sobrevivência: Uma Aplicação na Área de Relacionamento com Clientes*. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[25] Oliveira, J. G. C., Andrade, F. W. M. (2002). Comparação entre Medidas de Performance de Modelos de Credit Scoring. *Tecnologia de Crédito*, **33**, 35-47.

[26] Pereira, G. H. A. (2004). *Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais*. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[27] Reichert, A. K., Cho, C. C., Wagner, G.M. (1983). An Examination of the Conceptual Issues Involved in Developing Credit Score Models. *Journal of Business and Economic Statistics*, **1**, 101-114.

[28] Rocha, C. A., Andrade, F. W. M. (2002). Metodologia de Inferência de Rejeitados no Desenvolvimento de Credit Scoring. *Tecnologia de Crédito*, **31**, 46-55.

[29] Rosa, P. T. M. (2000). *Modelos de Credit Scoring: Regressão Logística, Chaid e Real*. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[30] Sicsu, A. L. (1998a). Desenvolvimento de um Sistema de Credit Scoring Parte I. *Tecnologia de Crédito*. **4**, 63-76.

[31] Sicsu, A. L. (1998b). Desenvolvimento de um Sistema de Credit Scoring Parte II. *Tecnologia de Crédito*. **5**, 57-68.

[32] Sohn, S. Y., Shin, H. W. (2006). Reject Inference in Credit Operations Based on Survival Analysis. *Expert Syst. Appl*, **31**, 26-29.

[33] Thomas, L. C., Edelman, D. B., Crook, J. N. (2002). *Credit Scoring and Its Applications*. Philadelphia: Siam.

[34] Tomazela, S. M. O. (2007). *Avaliação de Desempenho de Modelos de Credit Score Ajustados por Análise de Sobrevivência*. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[35] Weldon, G. (2000). Infering Behavior on Rejected Credit Applicants-Three Approaches. *Users Group International Conference - Statistics, Data Analysis, and Modeling*, Sigma Analytics & Consulting, Inc., Atlanta, GA, **paper 257**.