Universidade de Brasília
Departamento de Economia
Mestrado em Economia do Setor Público

JARÇO WIGOR SAMPAIO CHEREGATI

DETERMINANTES DO *SCORE* DE CRÉDITO E TEMPO ATÉ INADIMPLÊNCIA PARA EMPRÉSTIMOS COMERCIAIS A PESSOAS FÍSICAS

BRASÍLIA
2008

JARÇO WIGOR SAMPAIO CHEREGATI

DETERMINANTES DO *SCORE* DE CRÉDITO E TEMPO ATÉ INADIMPLÊNCIA PARA EMPRÉSTIMOS COMERCIAIS A PESSOAS FÍSICAS

Dissertação apresentada ao Departamento de Economia da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Economia do Setor Público.

Orientador: Prof. Donald Matthew Pianto

BRASÍLIA
2008

À memória de meu pai Roberto, minha mãe Jacilene e minha avó Jandira, que me transmitiram as principais características de um vencedor: dedicação, bom caráter e coragem.

AGRADECIMENTOS

Este trabalho, com suas virtudes e defeitos, não seria uma realidade sem o auxílio inestimável de muitas pessoas. Na tentativa de fazer justiça a todos aqueles que contribuíram direta ou indiretamente para a realização desse estudo, passo aos agradecimentos, desculpando-me, antecipadamente, por qualquer omissão.

É de extrema importância a gratidão que tenho pelo prof. Dr. Donald Pianto, orientador dedicado e paciente, pelos ensinamentos econométricos que me apresentou desde o início do curso.

Agradeço à instituição financeira que cedeu os dados necessários à consecução dessa dissertação, e aos colegas de trabalho Luciane, Fabiano, Douglas e Armando, que com seus ensinamentos de ordem prática, forneceram valorosa contribuição à confiabilidade e veracidade do presente estudo.

A todos os professores do curso de Mestrado Profissionalizante em Economia do Setor Público da UnB, deixo o meu muito obrigado pela multiplicação de seus talentos.

Aos meus colegas de mestrado, pelas horas de estudo que juntos passamos, incentivando-nos mutuamente.

Finalmente, gostaria de agradecer a todas as demais pessoas cujo apoio na revisão do texto ou até mesmo um simples abraço foram de vital importância ao bom andamento desse trabalho, a saber: aos meus irmãos Raphael, Gustavo, minha mãe Jacilene, minha avó Jandira e a Lígia.

DETERMINANTES DO *SCORE* DE CRÉDITO E TEMPO ATÉ INADIMPLÊNCIA PARA EMPRÉSTIMOS COMERCIAIS A PESSOAS FÍSICAS

Autor: JARÇO WIGOR SAMPAIO CHEREGATI

Orientador: DONALD MATTHEW PIANTO

RESUMO

O Novo Acordo de Basiléia (Basiléia II) permite aos bancos desenvolverem internamente abordagens para classificação de risco de acordo com as experiências em suas carteiras de crédito como uma das diretrizes que visam dar maior segurança ao sistema financeiro. Neste sentido, o objetivo deste trabalho é o de desenvolver um modelo estatístico de classificação de risco *credit scoring* a partir de informações sobre empréstimos concedidos no passado recente de uma carteira de crédito rotativo, tais como hábitos de pagamentos, variáveis cadastrais de perfil e comportamentais dos tomadores de crédito. Adicionalmente, será abordada a técnica de Análise de Sobrevivência, a qual busca estimar o tempo esperado até a default (inadimplência) de um contrato. As análises partem da definição da qualidade de crédito (bom ou ruim) e tempo observado até a inadimplência, sendo seguida pelo estudo das variáveis dos clientes que influenciam na capacidade destes em honrarem os compromissos de crédito obtidos. As técnicas empregadas para seleção das variáveis com relevância estatística para classificação dos bons e maus pagadores e tempo estimado até default foram a regressão logística (para o modelo de credit scoring) e Modelo de Riscos Proporcionais de Cox (para a Análise de Sobrevivência). Os resultados demonstram que é possível identificar antecipadamente o tempo esperado e a probabilidade de inadimplência dos tomadores de crédito, constituindo instrumentos de considerável relevância nas decisões afetas a políticas de concessão de crédito nas instituições financeiras.

Palavras-Chave: *Credit Scoring*, Regressão Logística, Análise de Sobrevivência, Modelo de Riscos Proporcionais de Cox.

DETERMINATIVE OF CREDIT SCORING AND TIME TILL DEFALUT FOR COMMERCIAL LOANS FOR NATURAL PERSON

Author: JARÇO WIGOR SAMPAIO CHEREGATI

Advisor: DONALD MATTHEW PIANTO

ABSTRACT

Basel II allows banks to develop risk classification techniques which incorporate experiences from their credit bureaus, granting wider security to the financial system. The aim of this work is the development of a statistical model of risk classification. This credit scoring model is based on information about past credit experience, such as payment habit and profile, and the behavioral variables of credit consumers. In addition, Survival Analysis, which attempts to estimate the time until a contract's default occurs, will be used. The analysis starts with the definition of credit quality (good or bad) and the time until default for the contracts studied and is followed by the exploration of which variables are unconditionally related to default. Variables were selected for inclusion in the logistic regression (for the credit scoring model) and Cox Proportional Hazards Model (for the Survival Analysis model) based on their statistical significance in explaining default and time to default, respectively. The results demonstrate that it is possible to identify the time to default and the probability of default for customers who intend to contract credit, hence the models are shown to be important tools for the credit decision policies of financial companies.

Keywords: Credit Scoring, Logistic Regression, Survival Analysis, Cox Proportional Hazard Model.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE EXEMPLOS

LISTA DE SIGLAS

AC	Acre
AL	Alagoas
AM	Amazonas
AP	Amapá
BA	Bahia
BACEN	Banco Central do Brasil
BIS	Bank for International Settlements
CA	Característica Amostral
CDB	Certificado de Depósito Bancário
CE	Ceará
CHAID	Chi-Squared Interaction Detection
CMN	Conselho Monetário Nacional
DEPEP	Departamento de Estudos e Pesquisas do Banco Central
DF	Distrito Federal
ES	Espírito Santo
EUA	Estados Unidos da América
GO	Goiás
IEP	Índice de Estabilidade da População
KM	Método Kaplan-Meier
KS	Estatística Kolgomorov-Smirnov
LR	*Likehood Ratio* ou Razão de Verossimilhança
MA	Maranhão
MG	Minas Gerais
MS	Mato Grosso do Sul
MT	Mato Grosso
PA	Pará
PB	Paraíba
Pc	Ponto de corte
PD	Probabilidade de *default*
PE	Pernambuco
PGD	Perdas geradas pelo *default*
PI	Piauí
PIB	Produto Interno Bruto

PL	Programação Linear
PLE	Patrimônio Líquido Exigido
PR	Paraná
RDE	Relatório de Desempenho da Escoragem
RDM	Relatório de Desempenho do Modelo
REF	Relatório de Escoragem Final
RIE	Relatório de Interferência de Escoragem
RJ	Rio de Janeiro
RN	Rio Grande do Norte
RO	Rondônia
ROC	Receiver Operation Characteristic
RR	Roraima
RS	Rio Grande do Sul

SUMÁRIO

Introdução

Num contexto de acelerada transformação dos mecanismos financeiros oriundos da globalização, um dos principais desafios das instituições financeiras é conceder créditos com segurança e menor risco. Organismos nacionais e internacionais têm implementado medidas regulatórias destinadas a manter a solidez do sistema financeiro, fundamental ao desenvolvimento econômico de qualquer país.

O novo Acordo de Basiléia (Basiléia II), instituído pelos países que integram o G10[1], cujas diretrizes têm sido adotadas não só pelos supervisores bancários dos países que o integram, mas também pelas demais nações, estabelece requisitos analíticos para avaliação de risco baseados em dados coletados pelos bancos durante todo o ciclo de vida de um empréstimo bancário. O mote principal do Basiléia II é introduzir uma estrutura de capital mais sensível a incertezas, incentivando as boas práticas de gerenciamento de risco, para os quais muitos bancos vêm estudando e implementando modelos próprios de análise e mensuração de risco.

Neste sentido, o objetivo deste estudo é o de apresentar uma metodologia de mensuração de risco do tipo *credit scoring* para análise de concessões de crédito, a partir do estudo estatístico de empréstimos concedidos num passado recente de uma linha de crédito de um banco brasileiro, discutindo os principais aspectos práticos que compõem um modelo. Posteriormente, será aplicada a técnica de análise de sobrevivência, cujo objetivo é prever em que momento do ciclo de vida do empréstimo ocorrerá a inadimplência. Esta metodologia tem sua importância sublinhada a partir do momento em que há expectativa de que, em poucos meses, o crédito concedido venha a inadimplir, onde sua contratação não valeria a pena. Adicionalmente, em cumprimento às recomendações Basiléia II, serão abordadas algumas técnicas pragmáticas para acompanhamento e manutenção da qualidade dos modelos após sua implementação.

A nomenclatura *credit scoring* consiste na utilização de métodos estatísticos para classificar candidatos à obtenção de um crédito qualquer em grupos de risco, uma vez que o credor não conhece o perfil de pagamento do contratante. Por meio do histórico de concessões

1 Apesar do nome do G10, onze países integram o grupo: Alemanha, Bélgica, Canadá, Estados Unidos, França, Holanda, Itália, Japão, Reino Unido, Suécia e Suíça.

de crédito efetuado no passado pela instituição financeira é possível, através de técnicas estatísticas, identificar variáveis cadastrais e comportamentais dos clientes que influenciam na qualidade do crédito a ser contratado. O modelo de regressão logística gera notas ou *scores*, cujo objetivo é estimar a probabilidade deste cliente tornar-se inadimplente, incorrendo em prejuízos à instituição credora. Desta forma, tem-se um importante instrumento para tomada de decisões baseada na classificação do risco do tomador. Importante salientar que, caso as características da população com a qual os modelos foram construídos se alterem ao longo do tempo, os mesmos podem tornar-se inadequados, sendo necessário calibrá-los ou reconstruí-los.

O cliente, como um potencial tomador de crédito, fornece seus dados pessoais como idade, tempo de emprego, renda, etc. mediante o preenchimento de uma ficha cadastral que contém suas variáveis de perfil. A instituição financeira, caso tenha informações comportamentais daquele cliente, como saldo médio, quantidade de cheques devolvidos, entre outras, fará uma ponderação de todas estas variáveis gerando um *score* (nota) que varia de 0 a 100 pontos. Modelos deste tipo, desenvolvidos para estimar a probabilidade de um cliente que já possui um determinado produto e tempo de relacionamento, são denominados modelos de *behaviour score*. A grande vantagem desses modelos decorre do fato deles possuírem um número maior de variáveis para ajuste.

Em termos gerais, a decisão de aceitar ou rejeitar o pedido de crédito é tomada comparando a probabilidade do indivíduo não honrar o compromisso com o intervalo de probabilidade aceitável, ou seja, pela comparação do *score* do candidato com um possível *score* de corte. A metodologia de análise de sobrevivência, de forma análoga aos modelos de *credit scoring*, utiliza variáveis explicativas de perfil e comportamentais para estimação da variável dependente, a qual se refere ao tempo esperado até a inadimplência (*default*) do empréstimo concedido.

A aplicação de modelos de *credit scoring* e outras ferramentas para análise de empréstimos iniciou-se nos países desenvolvidos em meados de 1960 e sua utilização por instituições financeiras de crédito vem aumentando rapidamente desde então[2]. No Brasil, o interesse por tais modelos começou a partir de 1994 com a estabilidade da inflação. Em geral,

[2] VASCONCELLOS, Maurício. **Proposta de Método para análise de concessões de crédito a pessoas físicas**. São Paulo: Tese de Mestrado – USP, 2002.

mesmo em países desenvolvidos, as empresas que utilizam modelos de risco de crédito não divulgam abertamente seus modelos em detrimento da necessidade de sigilo, já que boas e sofisticadas técnicas trazem vantagens competitivas. Outro fator crítico na construção de modelos de risco consiste nas informações confidenciais sobre os clientes que realizam operações de crédito, e não podem ser divulgadas a terceiros sem uma série de precauções.

Thomas, Edelman e Crook (2002) afirmam que a escolha dos proponentes a receberem crédito era, até o início do século XX, baseada exclusivamente no julgamento de um ou mais analistas. Com isso, a aprovação de crédito era subjetiva, uma vez que, dependendo do analista que julgasse o pedido de crédito, o pleito poderia ou não ser aprovado. Em 1936, Fisher (1936) desenvolveu a análise discriminante, técnica estatística que, a partir de características de um indivíduo, cria uma regra de classificação que permite inferir a que população ele pertence. O estudo deste autor pode ser considerado um dos primeiros modelos de *credit scoring*.[3]

Inicialmente, a substituição da experiência dos analistas de crédito pela utilização da ferramenta estatística foi recebida com desconfiança. Porém, com o crescimento do número de propostas, percebeu-se que era inviável fazer a análise individual de cada uma delas. Hand e Henley (1997) destacam que as pressões econômicas decorrentes da elevada demanda por crédito, a grande competição comercial do setor e o surgimento de novas tecnologias computacionais levaram ao desenvolvimento de modelos estatísticos sofisticados para decisões de crédito, tornando-as mais objetivas e rápidas, o que diminui as perdas das carteiras de crédito.

Sandroni (2002) define crédito como uma transação comercial em que um comprador recebe imediatamente um bem ou serviço adquirido, mas só fará o pagamento depois de um tempo determinado. O crédito bancário é oferecido por instituições financeiras autorizadas, como bancos múltiplos e comerciais, cooperativas de crédito, cuja operação é regulada e aceita mediante um contrato, com promessa jurídica de pagamento. Essa relação envolve duas noções fundamentais: confiança, expressada na promessa de pagamento, e tempo entre a aquisição e liquidação da dívida. Cartões de crédito, créditos rotativos e

[3] PEREIRA, G. H. **Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais**. São Paulo: Dissertação de Mestrado – USP, 2004.

empréstimos pessoais são as linhas de crédito que têm maior difusão junto aos clientes das instituições financeiras.

Os métodos de *credit scoring* e análise de sobrevivência a serem desenvolvidos ao longo deste trabalho baseiam-se em créditos requisitados para utilização de forma rotativa, a exemplo do cheque especial e cartões de crédito. Empréstimos de outra natureza, pagos em prestações mensais, como financiamento de veículos, imóveis e créditos pessoais, também podem ter modelos com as mesmas bases gerais, sendo necessárias apenas algumas modificações com relação ao estudo do comportamento de pagamento dos clientes do passado recente da carteira.

O gerenciamento de informações no sistema financeiro é de vital importância devido ao dinamismo característico do negócio. As empresas integrantes daquele sistema têm como principais atividades a intermediação financeira e a administração de recursos de terceiros, lidando com informações necessárias nas tomadas de decisões, e que, na maioria dos casos, envolvem expressivos valores monetários.

A globalização trouxe ainda mais dinamismo às atividades desempenhadas pelas instituições financeiras, imprimindo maior velocidade na movimentação de dinheiro entre os países. Como exemplo dessa mobilidade, tem-se as crises financeiras ocorridas em alguns países emergentes no final da década de 1990, que presenciaram a fuga de capital especulativo em velocidade até então desconhecida.

Para se ter uma percepção do papel fundamental que o crédito desempenha na economia, basta citar o fato de que em 2003 cerca de dois terços do Produto Interno Bruto (PIB) dos Estados Unidos decorria do consumo[4]. Parte considerável deste consumo era financiada por instituições interessadas em conceder crédito em troca de um ganho sobre o capital emprestado. No Brasil, a indústria de crédito é relativamente proporcional ao tamanho da economia, bem menor do que em países desenvolvidos. [5] Contudo, a concessão de crédito tem demonstrado franco crescimento nos últimos anos - notadamente no período pós-Plano Real - delineado com o fim do longo período inflacionário.

[4] Idem
[5] Idem

No estudo de Pinheiro e Cabral (1998) foi mencionado que o mercado de crédito brasileiro caracteriza-se por um tamanho (volume) relativamente baixo, com altas taxas de juros e inadimplência elevada. O período pós-Plano Real foi marcado por um forte processo de privatização e de redução de crédito ao setor público – principalmente do Governo Federal – o que ensejou uma contração do percentual de crédito em relação ao PIB. Não obstante, nota-se que os empréstimos ao setor privado no segmento de crédito direto ao consumidor têm mostrado os maiores índices de crescimento desde a implantação do Plano Real. Este período foi caracterizado pela forte queda da inflação e conseqüente redução das receitas não provenientes de juros, as quais encorajaram e facilitaram uma forte expansão das linhas de crédito para o setor privado. O crédito destinado ao consumo das famílias brasileiras, por exemplo, subiu de 2,4% do total de empréstimos nos anos de 1988 a 1993, para 8,4% em 1994, atingindo 13% em 1997, de acordo com os dados do Banco Central do Brasil (BACEN). Outros indicadores, tais como o aumento expressivo do número de cartões de crédito e volume de transações com cartões, mais que dobraram de 1993 a 1997. Em comparação com o mercado de crédito norte-americano, vê-se que ainda há espaço para o crescimento do mercado creditício no Brasil - país que vem demonstrando uma forte expansão de crédito ao consumidor, graças à estabilidade econômica.

Morisson (2005) avalia positivamente a elaboração de um anteprojeto de lei que trata da criação de um banco de dados para proteção ao crédito, fruto de um amadurecimento do mercado de crédito brasileiro e de um ambiente de negócios favorável ao aumento das transações econômicas[6]. O autor constata que embora o projeto não tenha em si uma novidade tão grande, ele visa estabelecer regras claras sobre a oferta de um serviço extremamente necessário ao bom funcionamento do mercado e, portanto, com conseqüências positivas à atividade econômica do País.

A teoria econômica demonstra que há uma perda de bem-estar da sociedade quando os agentes tomam decisões num ambiente com assimetria de informação, ou seja, quando uma das partes possui menos informações que a outra[7]. Isto porque numa negociação de crédito, o tomador sabe exatamente sua real disposição em pagar, enquanto o credor tem informações imprecisas sobre essa disposição. Em um caso limite, isto pode levar a uma situação caótica,

---

[6] MORRISON, J. S. **Preparativos para o Novo acordo da Basiléia**. Disponível em: <http://forecastingsolutions.com/publications/47.pdf>. Acesso em: 11 nov. 2007.
[7] Idem.

na qual a inadimplência se tornaria muito elevada, redundando em altas taxas de juros ao consumidor final. A persistência dessa situação leva a uma perda de dinamismo da economia, com redução de investimentos e qualidade de vida. Os resultados de pesquisas empíricas têm mostrado que, com o uso de informações compartilhadas, há uma melhoria nos processos de previsão de risco de crédito, o que assegura maior estabilidade do mercado, aumento da oferta de crédito e redução nas taxas de juros e inadimplência[8].

A inadimplência tem um papel relevante no alto custo do capital no Brasil, retratado pela elevada taxa de juros cobrada pelos bancos. Como exemplo, pode-se citar que a taxa média de juros bancários durante o período de outubro de 1996 a outubro de 2002 oscilou entre 80% a 60% ao ano, enquanto a taxa média de captação de recursos no mercado (taxa média do Certificado de Depósito Bancário – CDB) manteve-se em cerca de 20% ao ano para o mesmo período[9]. Isto sugere um *spread* (diferença entre a taxa de aplicação e captação) médio de 40%, patamar considerado extremamente alto, mesmo em países com economia emergentes[10]. Estudos do DEPEP (Departamento de Estudos e Pesquisas do Banco Central) demonstram que na composição do *spread* bancário, a inadimplência responde por 35%, constituindo o fator de maior impacto no *spread*[11].

O risco de crédito tem sido um fator determinante do elevado custo dos empréstimos, o que explica a dificuldade ou mesmo a não concessão de empréstimos pelos bancos. Ao realizar concessões, os credores querem ter certeza de receber os recursos emprestados mais os juros pactuados, pois os intermediários financeiros têm obrigações para com seus depositantes e acionistas. Como esta certeza não existe, mesmo para clientes de primeira linha, os bancos sempre cobram um adicional a título de risco de crédito, ou seja, um valor associado à probabilidade de não receber o valor emprestado. Neste contexto, os modelos de risco de crédito têm sua importância ratificada, pois tentam prever o comportamento e risco de inadimplência dos tomadores, reduzindo assim a assimetria de informação. Discriminar bons e maus pagadores de forma eficiente tem como conseqüência

---

[8] Idem.
[9] BRASIL. Banco Central do Brasil. **Avaliação de 3 anos do projeto Juros e *Spread* bancário**. Brasília, 2002. Disponível em <http://www.bcb.gov.br/ftp/juros-spread1.pdf >. Acesso em: 11 nov. 2007.
10 Idem
[11] BRASIL. Banco Central do Brasil. **Economia Bancária e Crédito – Avaliação de 5 anos do Projeto Juros e Spread Bancário**. Brasília, 2004. Disponível em: http://www.bcb.gov.br/Pec/spread/port/economia_bancaria_e_credito.pdf. Acesso em: 11 nov. 2007

lógica a redução dos índices de inadimplência. Esta ação redunda na queda do *spread* bancário, ensejando uma série de benefícios à sociedade.

Em 1998 o Comitê de Basiléia para Supervisão Bancária divulgou o Acordo de Capital (Basiléia I), que propunha um conjunto mínimo de diretrizes para adequação de capital em bancos. O objetivo do Acordo foi fortalecer a solidez e a estabilidade do sistema bancário por meio da recomendação de que os bancos constituíssem um capital mínimo, de forma a minimizar o risco de insolvência das instituições bancárias.

Inicialmente, o Comitê definiu uma medida de solvência que cobria o risco de crédito com adequação de capital igual à pelo menos 8% dos ativos do banco, ponderados pelo risco da relação dos ativos da instituição com as contrapartes envolvidas nas operações. As medidas sugeridas no Acordo foram implementadas nos países membros em 1992, e no Brasil em 1994, por meio da publicação da Resolução n°. 2.099 pelo BACEN, a qual também estabeleceu o índice mínimo de 8% na constituição do Patrimônio Líquido Exigido (PLE) dos ativos das instituições, ponderados pelo risco. Em 1997 esse índice foi alterado para 11%, por meio da Circular BACEN n°. 2.784.

Um fator que dá importância ao estudo de modelos de risco diz respeito à Resolução do CMN nº. 2.682/99, que estabelece percentuais de aprovisionamento de acordo com a classificação de risco das carteiras de operações de crédito. Se os clientes de uma carteira forem classificados como “nulo risco”, estes recebem classificação “AA”, o que significa 0% de aprovisionamento do total de crédito concedido a estes clientes. Clientes classificados como “A” têm 0,5%, enquanto a classificação “B” remete a 1% de aprovisionamento, e assim por diante, até o conceito “H”, que significa um aprovisionamento de 100%. Portanto, um cliente avaliado como “alto risco” (classificado como “H”) fará com que o banco aloque 100% do recurso concedido, para fins de aprovisionamento de crédito. Nestes casos, as instituições financeiras normalmente estabelecem notas ou *scores* de corte com o objetivo de nortear a concessão de crédito, prioritariamente, a clientes que apresentam baixo risco. Esta medida segue as diretrizes do Basiléia II e visa dar maior segurança ao sistema financeiro. Por outro lado, aos bancos significa também um volume menor de recursos disponíveis para circulação no mercado. Neste sentido, observa-se novamente a importância de modelos de risco eficientes, pois à medida em que bons clientes são selecionados, reduz-se o

aprovisionamento de crédito, redundando em maior disponibilidade de capital destinado à sociedade.

Em 2001, o Comitê de Basiléia lançou uma proposição ao Novo Acordo de Basiléia visando desenvolver uma nova estrutura para fortalecer a solidez e estabilidade do sistema bancário internacional, recomendando a adoção de uma política de administração de riscos mais sólida para o setor bancário, não sustentadas simplesmente na determinação de capital. Em 2004 o BIS (*Bank for International Settlements*) publicou o documento *International Convergence of Capital Measurement and Capital Standartds* – conhecido como Basiléia II – adicionando o risco operacional na ponderação dos ativos para efeito de cálculo de capital regulamentar, além dos riscos já mensurados no primeiro Acordo (risco de crédito e de mercado).

A adoção dos critérios do Basiléia II exige uma nova estrutura normativa, para o qual o BACEN e o Conselho Monetário Nacional (CMN) emanaram um conjunto de Resoluções que norteiam os métodos mensuradores dos riscos de crédito, mercado e operacional.[12] Dentre estas estão os Comunicados n°. 12.746/04 e 16.137/07 que informam procedimentos gerais e a cronologia básica de implantação do Novo Tratado no Brasil, "adaptadas às condições, peculiaridades e estágio de desenvolvimento do mercado brasileiro". É estabelecido um cronograma com prazos, ações programadas e normas criadas para guiar o cumprimento das referidas ações.

O Novo Acordo permitiu aos bancos desenvolverem internamente abordagens para classificação de risco de acordo com as experiências em suas carteiras de crédito, onde um de seus pilares principais consiste na exigência de capital mínimo mantido pelos bancos para fazer frente a eventuais insolvências, constituindo como requisito básico a boas práticas para o gerenciamento do risco[13]. Zendersky, Gulias e Silva (2005) destacam que o tratado prevê que as instituições financeiras desenvolvam metodologias proprietárias para cálculo do capital mínimo regulatório, de acordo com a probabilidade de *default* (PD) e perdas geradas pelo

[12] Para melhores detalhes sobre os pontos normativos da passagem do Basiléia I para o II, ver CARVALHO, B. C.; DOS SANTOS, G. M. **Os Acordos de Basiléia – Um roteiro para implementação nas instituições financeiras.** Disponível em http://www.febraban.org.br/Arquivo/Servicos/Imprensa/Artigo_Basileia_6.pdf. Acesso em 01 out. 2008

[13] CORTES, F. P. **Gestão de Risco nas Instituições Financeiras: Uma Análise do Novo Acordo de Basiléia e Apresentação de Conceitos para Desenvolvimento de um Sistema de Informações Gerenciais.** Brasília: Fundação Getúlio Vargas, 2004.

default (PGD)[14]. Isto implica que os modelos de risco, ao classificarem corretamente os potenciais tomadores reduzem não só a inadimplência, mas como também a PD e PGD das carteiras de crédito.

O banco de dados formado para construção de modelo de risco de crédito baseia-se numa amostra de candidatos cujo crédito já foi concedido num passado recente da carteira, denominado base de dados. Esta inclui tipicamente variáveis preditoras, como características do indivíduo e da operação, bem como a classificação real a que pertenceu esta operação – crédito "bom" ou "mau" – de acordo com os atrasos ocorridos no pagamento das prestações. O modelo desenvolvido nos próximos capítulos segue essa divisão de operações passadas em dois grupos qualitativos, mais comum nesse tipo de estudo.

Os estudos sobre risco de crédito recomendam buscar dados que estejam válidos e livres de erros ou viés, ou seja, o mais próximo da realidade. Hand e Thomas (2002) reconhecem que os bancos de dados utilizados na construção de modelos de risco sempre possuem algum tipo de viés, uma vez que a amostra utilizada para construção dos mesmos possui somente informações que foram aprovadas no passado, ou seja, são omitidas informações dos clientes que tiveram acesso ao crédito negado. Nota-se, portanto, que a amostra utilizada não é baseada em toda população de potenciais tomadores de crédito, o que constitui uma limitação do presente estudo, o chamado viés de seleção. Assim, os modelos de risco de crédito normalmente refletem a probabilidade condicional, dado que o consumidor teve acesso ao crédito, conforme descreve Zerbini (2000).

Vasconcellos (2002) argumenta que o problema de viés de seleção não tem se mostrado simples de se solucionar, mesmo com o avançado grau de desenvolvimento atual das técnicas estatísticas relacionadas ao problema de estimação e processos de amostragem. Mesmo assim, em geral, os modelos de risco são considerados capazes de classificar corretamente grande parte das operações de crédito, principalmente quando gerados sobre amostras de grande tamanho, contendo um número considerável de variáveis. A existência do viés de seleção requer que a implementação dos modelos seja feita com a consciência de que os mesmos sofrem daquele problema estatístico, e que, portanto, não refletem uma fotografia perfeita da realidade.

[14] Para melhores detalhes, ver CORTES op. cit.

Além da tradicional classificação de bons e maus pagadores baseada no atraso dos pagamentos, a literatura de risco de crédito também versa sobre uma abordagem alternativa, focada na maximização dos lucros. Esta ótica nos leva a pensar da seguinte forma: clientes que efetuam o pagamento de suas prestações em dia, normalmente têm acesso ao crédito com menores taxas de juros, não pagam multas e juros por atraso e, portanto, não são rentáveis. Da mesma forma, clientes com risco elevado que atrasam o pagamento das prestações podem ser bastante rentáveis desde que as taxas de juros sejam suficientemente altas e que os atrasos não sejam prolongados. Morisson enfatiza que um empréstimo inadimplente pode render essencialmente 100% de retorno ao credor, considerando a premissa de que o tomador pague um mínimo de prestações e multas por atraso suficientes para garantir lucro à operação. Abordagens com este tipo de enfoque, denominadas *profit scoring*, objetivam ordenar os clientes de acordo com a probabilidade de dar lucro à instituição.

Hand e Thomas expõem que uma das vantagens do *profit scoring* é a possibilidade que os credores têm em tomar decisões que maximizem o retorno financeiro dado pelos consumidores, ao invés de apenas estimar o risco de inadimplência. Com isso, as instituições financeiras perceberam que pode-se, inicialmente, escolher o limite de crédito, taxa de juros e outras características da operação, as quais, caso empregadas de forma adequada, podem trazer ganhos extraordinários no gerenciamento de risco de crédito. Porém, como discutido em Pereira (2004), a construção de modelos baseados em *profit scoring* é mais difícil do que se imagina, tendo em vista a necessidade de se ter todas as despesas presentes no cálculo da variável resposta. Segundo o autor, esta contabilização deve envolver até mesmo despesas com marketing e recursos humanos, além de receitas oriundas da recuperação da área de cobrança.

No entanto, os modeladores muitas vezes omitem aspectos mercadológicos de concessão de crédito, inadimplência e lucratividade: na prática, a mensuração de risco por meio do atraso de pagamentos está intimamente ligada com a lucratividade da operação de crédito, sendo observado que contratos com pequeno ou nenhum atraso formam o conjunto lucrativo da carteira de crédito (apesar de menor lucro para os atrasos nulos), enquanto que atrasos maiores formam o conjunto de prejuízo da carteira. Assim, a divisão de “bons” e “maus” créditos por meio do estudo de atraso de pagamentos funciona como uma excelente

aproximação para a lucratividade da carteira de crédito e poderiam ser claramente relacionadas.[15]

Neste sentido, o presente trabalho baseia-se na divisão dicotômica de créditos "bons" e "maus" de acordo com o estudo de risco pelo critério da inadimplência, gerada a partir dos atrasos nos pagamentos, e não da lucratividade. O motivo que levou a esta decisão, além das razões expostas anteriormente, foi a dificuldade em se obter informações econômico-financeiras da carteira de crédito estudada junto à instituição que cedeu os dados para a consecução deste estudo, o que seria essencial caso se deseje agrupar os créditos pelo critério do lucro e não do atraso. Esta situação remete a uma freqüente discussão acerca dos modelos de risco, que consiste na dificuldade e alto custo para obtenção dos dados para confecção dos modelos, basicamente originados pelos empecilhos técnicos de processamento de dados e principalmente do fato das instituições não poderem divulgar abertamente suas informações. Afinal, trata-se de um mercado de elevada competição, com elevados volumes financeiros em questão, e que detém tecnologia, estratégias e conhecimentos vantajosos, o qual não se pretende fornecer indícios de seus procedimentos e resultados financeiros.

Estudos sobre análise de sobrevivência são pouco utilizados na literatura de risco de crédito. O objetivo desta abordagem é a de calcular o tempo esperado até a ocorrência do evento em interesse, que em nosso caso, trata-se da inadimplência. Stepanova *et al* (2005) citam algumas vantagens de se estimar o tempo esperado até que os clientes se tornem inadimplentes:

- A possibilidade de computar a lucratividade de cada cliente, ou seja, o desempenho de *profit scoring*;

- Fornecer aos bancos uma estimativa de níveis de inadimplência, úteis ao aprovisionamento de crédito de acordo com os preceitos do Basiléia II;

- Auxílio na formulação de políticas de crédito que mensurem o prazo máximo das operações a serem concedidas, visto que baixas expectativas de sobrevivência podem não justificar sua concessão.

[15] VASNCONCELLOS op. Cit.

Grande parte dos avanços dos estudos na área de análise de sobrevivência é decorrente de pesquisas militares na Segunda Guerra Mundial, cujo objetivo era prever o tempo que decorreria até que o equipamento militar apresentasse avarias[16]. Vasta é a utilização desta técnica na área da biomedicina, onde se busca estimar a expectativa de vida dos pacientes ou experimentos de acordo com características dos tratamentos a que são submetidos[17]. No âmbito do mercado financeiro existem estudos que buscam prever a inadimplência bancária e falências de empresas, baseado nos indicadores econômico-financeiros da instituição. A maioria destes estudos utiliza as técnicas de análise discriminante, regressão logística, redes neurais e o modelo proporcional de Cox para estimar a probabilidade da insolvência das empresas.

O modelo mais comumente usado na análise de sobrevivência é o modelo proporcional de Cox (também chamado de modelo Cox). Esta técnica permite a inclusão de variáveis explicativas que influenciam na estimativa do tempo de sobrevivência do evento em interesse. Esta será a metodologia utilizada no presente trabalho para estimar o tempo esperado até inadimplência dos contratos da carteira de crédito em estudo. Assim, de forma análoga ao modelo de *credit scoring*, o modelo Cox utilizará variáveis preditoras (comportamentais e cadastrais dos tomadores de crédito) que influenciam no tempo esperado até que ocorra a inadimplência. A variável dependente da análise de sobrevivência é obtida computando-se o prazo (em dias, meses, anos, etc) decorrido entre a data de contratação do crédito e a data em que o contrato tornou-se inadimplente.

Além desta introdução, o estudo está dividido em outros cinco capítulos. O segundo capítulo fala sobre a revisão bibliográfica utilizada como arcabouço teórico, onde são abordados os diversos tipos de regressão existentes. Também são feitas algumas considerações acerca da estabilidade da população e de outros relatórios de acompanhamento do desempenho dos modelos. O terceiro capítulo versa a respeito da base de dados utilizada, mais especificamente sobre a amostra coletada, as variáveis obtidas e sobre o tratamento dos dados. A metodologia está disposta no quarto capítulo, onde são analisados o método para definição da qualidade de crédito, a categorização das variáveis e as técnicas utilizadas na

---

[16] MORRISON op. Cit.

[17] LAMESHOW and HOSMER. **Applied Survival Analysis**. Epidemiology of University of Massachusetts, 1999.

construção dos modelos. O quinto capítulo discute os resultados obtidos, um exemplo prático da aplicação dos modelos, a análise gráfica dos resultados estimados e a estabilidade do modelo. No capítulo seis encontra-se a conclusão.

1. Revisão de Metodologias

Ao longo dos últimos anos algumas metodologias têm sido criadas e desenvolvidas com o objetivo de gerar modelos melhores para previsão de inadimplência. Apesar deste trabalho não se propor a testar estas novas metodologias, será feita uma breve revisão histórica das tecnologias mais reconhecidas, assim como dos trabalhos que buscaram comparar a eficiência entre estas técnicas.

Posteriormente, em atendimento às recomendações do Novo Acordo de Basiléia, serão abordados indicadores e relatórios de acompanhamento que permitem acompanhar e gerir os riscos de modelagem, após a implantação dos modelos, contribuindo na melhoria do processo de tomada de decisão de crédito.

1.1 Técnicas de Estimação

1.1.1 Análise Discriminante

O primeiro autor a publicar trabalhos sobre a classificação do consumidor em dois grupos em uma população foi Fisher, enquanto Durand (1941) foi o primeiro a aplicar esta metodologia no sistema bancário, diferenciando bons e maus empréstimos. Valorosa contribuição foi dada por Altman (1968) ao incentivar a utilização desta técnica de decisão em larga escala. Nesta época a concessão de crédito norteava-se em regras de bolso escritas por analistas experientes, permitindo, com um certo grau de subjetividade, um razoável padrão de precisão para a época.

Fisher buscou identificar uma combinação linear das variáveis independentes que melhor separasse os dois grupos. Foi criada a Função Discriminante Linear, definida como:

$$Y_i = W_1X_1 + W_2X_2 + ... + W_PX_P$$

Onde W_i são os pesos atribuídos a cada uma das variáveis X_i, de forma a maximizar o poder discriminante de Y.

Estabelecida a função discriminante, é possível atribuir uma pontuação a Y_i para cada indivíduo "i" da amostra, e restará determinar o ponto de corte Y_c, que deverá minimizar os erros Tipo I (aprovar crédito para indivíduos inadimplentes) e tipo II (negar crédito para indivíduos adimplentes).

Desta forma:

- Se $Y_i < Y_c \rightarrow$ crédito será rejeitado
- Se $Y_i > Y_c \rightarrow$ crédito será aprovado

Martell e Fitts (1981) evoluíram para os modelos de análise discriminante quadrática, que ponderam as diferentes variâncias das populações de adimplentes e inadimplentes. Einsembeis (1977, 1978) criticou a utilização da análise discriminante devido à dificuldade de se separar a população em dois grupos distintos, ressaltando a necessidade de considerar aspectos de evolução do relacionamento do cliente com a instituição. Afirmou ainda que a ausência desta dinâmica leva a uma decisão menos eficaz. Capon (1982), particularmente, acreditava que maior peso deveria ser dado ao comportamento de crédito do consumidor, ou seja, mais ênfase ao *Behaviour Score***.**

1.1.2 Programação Linear

A Programação Linear também pode ser usada com o fim de gerar previsões para a inadimplência. Nesta técnica deseja-se construir uma pontuação através de um modelo linear que utiliza as variáveis independentes X de forma a deixar todos os N_g adimplentes acima de um valor predito "c" qualquer, e ao mesmo tempo deixar todos os N_b inadimplentes abaixo deste mesmo valor "c". Em seguida, introduz-se a variável $\left| a_i \right|$, que contém o erro tipo I e tipo II, mas que deve ser minimizada. Desta forma, o modelo segue o formato abaixo:

Mín $\left| a_1 \right| + \left| a_2 \right| + ... + \left| a_{Ng\,Nb} \right|$

s.a. $w_1 x_{i,,1} + w_2 x_{i,,2} + ... + w_m x_{i,,m}, \geq c - a_i$ onde $1 \leq \text{i} \leq N_g$

$w_1 x_{i,,1} + w_2 x_{i,,2} + ... + w_m x_{i,,m}, \leq c - a_i$ onde $N_g + 1 \leq i \leq N_g + N_b$

Nota-se que $1 \leq i \leq N_g + N_b$ e que w_1 = coeficiente da variável independente.

Mangasarian (1965) foi o primeiro a abordar o problema desta forma, fazendo com que esta técnica fosse, de fato, considerada para a previsão da inadimplência. Gehrlein e Wagner (1997) desenvolveram uma formulação de Programação Linear que incorpora na função objetivo o custo de inadimplência e o custo de oportunidade, dando um sentido mais moderno de gestão de carteira, e não simplesmente de previsão da probabilidade de inadimplência.

A novidade deste modelo é incorporar a informação de política de juros e o custo de inadimplência, o que demanda um monitoramento mais freqüente, dado que mudanças de mercado podem modificar rapidamente o ponto "c" ótimo. Scarpel e Milioni (2002) propuseram a utilização conjunta dos modelos de programação linear e de regressão logística, unindo num só modelo a probabilidade de inadimplência e a lucratividade (taxa de juros) do concedente, permitindo determinar o valor de empréstimo ótimo.

1.1.3 Redes Neurais

Uma técnica que tem sido cada vez mais utilizada nos modelos de risco de crédito refere-se às Redes Neurais, metodologia baseada no sistema nervoso central humano, podendo ser classificada segundo suas características principais, ou seja, "topologia da rede neural", "forma de aprendizado" e "algoritmo de aprendizado". [18]

Thomas (2000) descreve a topologia de uma Rede Neural da seguinte forma: numa camada de entrada estão todas as variáveis independentes (X_{tn}) que alimentam vários neurônios, os quais, por sua vez, calculam a função de transferência (soma ponderada dessas variáveis independentes utilizando uma função que pode ser linear, logística ou tangente hiperbólica). Calculada a função de transferência, o resultado final de Rede Neural será expresso numa camada final de apresentação, ou seja, gerará a variável dependente (Y_t). Caso haja uma segunda camada de neurônios entre os iniciais e a camada de saída, haverá um encadeamento dos mesmos, dando origem ao caso *two layer*. Havendo três camadas de neurônios, teremos o caso *three layer*, e assim por diante. O autor destaca, ainda, que a principal vantagem desta técnica é que apesar do modelo ser fixo, o processo de aprendizado realizado constantemente permite modificações freqüentes na fórmula. A desvantagem é que

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

mudanças constantes dificultam a observação dos fenômenos que geram a modificação no modelo.

Paterson (1996) assinala outra característica importante da rede neural que é a forma de aprendizado, cuja metodologia mais utilizada em crédito ao consumidor é o aprendizado supervisionado, o qual utiliza uma amostra de desenvolvimento com variáveis independentes como referência (ou balizadores) para qualquer mudança sugerida pelo algoritmo de aprendizagem.

O algoritmo de aprendizagem refere-se à ponderação dos pesos da função de transferência e é usualmente feito através de um algoritmo que identifica os erros cometidos pelos neurônios de saída e corrige os pesos das funções de transferência das camadas anteriores. Devido ao tempo de processamento, este algoritmo é computacionalmente muito caro, baseando-se em ajustes recursivos dos pesos da função de transferência no sentido de redução do erro na camada de saída. [19]

Hair (1998) destaca que inexistem testes estatísticos que avaliam a significância dos pesos da função de transferência, entretanto, comprova a utilidade da aplicação desta metodologia em modelos de previsão de inadimplência, principalmente em casos onde existe não-linearidade nos dados da amostra.

1.1.4 Algoritmos Genéticos

Albright (1994) foi um dos primeiros autores a descrever a técnica de Algoritmos Genéticos, a qual se baseia em rotinas de sorteios aleatórios, assim como nos processos genéticos, gerando inúmeras fórmulas (ou regras), testadas empiricamente quanto à eficácia na previsão de inadimplência. A fórmula escolhida será a que apresenta as menores pontuações atribuídas aos indivíduos que foram inadimplentes de fato.

Barth (2002) descreve detalhadamente esta metodologia. O autor afirma que o refinamento da capacidade preditiva do modelo de algoritmo genético é feito por meio da fase chamada “reprodução”, onde novas fórmulas são geradas a partir da combinação das fórmulas

[19] PATERSON, Dan W. **Artificial Neural Networks, Theory and Applications**, Pretince Hall Inc USA, 1996

mais eficazes da fase anterior do processo de sorteio. Este é um método recursivo conhecido como seleção de genitores, onde se deseja convergir para regras homogêneas.

Outro processo de refinamento ocorre na fase seguinte, chamada mutação, onde se modifica aleatoriamente uma variável independente, escolhida também ao acaso, causando uma evolução que pode ter efeito positivo ou negativo na previsão da inadimplência. Este mecanismo tem o objetivo de evitar a convergência para regras não ótimas.

O procedimento recursivo destas três fases da técnica (estabelecimento de regras, reprodução e mutação) tende a convergir para fórmulas semelhantes, todas igualmente aptas e com alto poder discriminador, as quais deverão ser estudadas pelo analista para validação. A condição de parada está relacionada à minimização da função objetivo, que se relaciona ao número de erros (ou custo do erro). Entretanto, esta convergência não acontecerá sempre, pois as variáveis independentes podem ser inadequadas para prever o fenômeno desejado, ou podem ocorrer convergências rápidas que podem denotar regras de discriminação fracas.

A metodologia de algoritmos genéticos demanda ao analista três tipos de calibração, que são o número “N” de indivíduos iniciais, a taxa de reprodução e a taxa de mutação. A seleção destes parâmetros envolve a base de dados e a experimentação do analista, mas é importante salientar que os resultados obtidos dependem das escolhas de tais parâmetros, e por isso é recomendável processar as rotinas mais de uma vez para comparar os resultados. Há também que se considerar que esta metodologia gera modelos de fácil interpretação prática, pois o resultado é uma fórmula com variáveis e pesos, que determinarão diretamente a inadimplência esperada de cada indivíduo da amostra.

1.1.5 Análise de Sobrevivência

Nakano e Carrasco (2006) definem Análise de Sobrevivência ou confiabilidade como um conjunto de técnicas e modelos estatísticos usados na análise de experimentos cuja variável resposta é o tempo até a ocorrência de um evento de interesse. Os indivíduos sob estudo podem ser animais, humanos, plantas, equipamentos, etc. Por outro lado, o evento de interesse pode ser: morte, remissão de uma doença, reação de um medicamento, quebra de um equipamento eletrônico, queima de uma lâmpada etc.

No âmbito do mercado financeiro alguns estudos utilizam a análise de sobrevivência com o objetivo de identificar instituições financeiras em dificuldades, na tentativa de obter um sistema de "*early warning*"[20]. Rocha (1999) construiu um modelo de previsão de insolvência utilizando o Modelo de Cox, gerando uma estimativa do tempo decorrido até a falência de um banco. Usando um conjunto de 26 indicadores financeiros, a autora estimou o modelo a partir de uma amostra de 32 bancos (17 solventes e 15 insolventes). Os resultados de seu trabalho indicam que o modelo de risco proporcional pode ser utilizado como um sistema de *early warning*, uma vez que apresenta um alto grau de precisão, identificando com antecedência boa parte das falências verificadas no período em estudo.

Janot (1999) desenvolveu um trabalho semelhante ao construir modelos de previsão de insolvência bancária utilizando regressão logística e o modelo Cox por meio de uma amostra composta por 40 bancos solventes e 21 insolventes (que sofreram intervenção ou liquidação pelo Banco Central entre 1995 e 1998). O autor conclui que tanto a regressão logística como o modelo Cox podem ser usados como *early warning.* Entretanto, os resultados favorecem o modelo de Cox na medida em que este apresenta maior capacidade de previsão, além de estimar o tempo restante até a falência.

A literatura sobre a previsão do momento da inadimplência em operações de crédito é relativamente recente e não muito vasta. Narain (1992) foi um dos primeiros autores a utilizar análise de sobrevivência para métodos de *credit scoring* ao analisar os dados de 1.242 clientes que contrataram empréstimos entre 1986 e 1988. Os dados foram analisados usando o método de Kaplan-Meier, e o modelo de regressão exponencial por um período de 24 meses de observação após a contratação do empréstimo. Narain demonstrou que a análise de sobrevivência adiciona uma nova dimensão à abordagem tradicional, uma vez que as técnicas de concessão de crédito podem ser aprimoradas quando a expectativa de duração do crédito pode ser estimada. O autor assinalou que estes métodos podem ser aplicados a qualquer área de operações de crédito onde existam variáveis preditoras e o tempo de ocorrência dos eventos em interesse.

---

[20] Forma de monitoração de empresas, que por meio de determinados indicadores, permite às autoridades competentes a fiscalizarem e avaliarem riscos envolvidos na operacionalização de seus processos. Para mais detalhes ver MARTINS, M. **A previsão de insolvência pelo Modelo Cox: uma contribuição para a análise das companhias abertas brasileiras**. Porto Alegre: Dissertação de Mestrado. UFRS, 2003.

Stepanova et al. utilizaram uma amostra de 15.000 empréstimos concedidos em uma instituição financeira do Reino Unido entre 1994 e 1997 para estimar o tempo esperado até *default* e o tempo esperado até a quitação antecipada do compromisso. As técnicas de redes neurais, regressão logística e modelo Cox foram usadas como forma de comparar a acurácia destes modelos. Os resultados obtidos demonstraram que as técnicas de redes neurais e modelo Cox tiveram o mesmo desempenho em prever a inadimplência, enquanto no caso de quitação antecipada as três técnicas mostraram-se equiparadas. Os autores finalizam o trabalho recomendando a realização de mais estudos nessa área, utilizando bases de dados provenientes de outras populações.

1.1.6 Comparação de Resultados das técnicas estatísticas

Alguns acadêmicos analisam as diferenças entre as técnicas de previsão de inadimplência. Barth (2002) relata em seu trabalho que:

- Altman (1994) conclui que os resultados da metodologia de Redes Neurais se mostraram piores do que aqueles alcançados com Análise Discriminante e/ou Regressão Logística quando aplicados à amostra de validação;

- Varetto (1998) concluiu que os resultados de Algoritmos Genéticos também se mostraram piores do que aqueles alcançados com Análise Discriminante e/ou Regressão Logística quando aplicados à amostra de validação;

- Adya e Coloppy (1998) estudaram vários trabalhos comparativos da técnica de Redes Neurais com outros métodos. Apesar de não definitivo, alguns dos trabalhos que receberam crédito dos autores posicionavam a metodologia de Redes Neurais como melhor comparativamente à Análise Discriminante/Regressão Logística em algumas situações específicas.

Em Thomas, Oliver e Hand (2005) é feita uma classificação da acurácia em termos percentuais, de acordo com os casos corretamente classificados nas amostras de cada autor, conforme tabela, abaixo:

TABELA 1 - Comparação de acurácia entre diferentes técnicas

Autores	Regressão Linear	Regressão Logística	Programação Linear	Redes Neurais	Algoritmos Genéticos
Henley (1995)	43,4	43,3	-	-	-
Boyle *et al.* (1992)	77,5	-	74,7	-	-
Srinivisan e Kim (1987)	87,5	89,3	86,1	-	-
Yobas *et al.* (1997)	68,4	-	-	62,0	64,5
Desai *et al.* (1997)	66,5	67,3	-	64,0[1]	-

(1) No artigo consta 6,4 – entretanto, acredita-se que se trata de erro de impressão

Ao analisar a tabela, nota-se que em Boyle *et al* e Srinivisan e Kim a regressão linear foi a que teve melhor acurácia, enquanto em Srinivisan e Kim a regressão logística obteve melhor desempenho. Thomas, Oliver e Hand afirmam que a técnica de redes neurais é melhor para trabalhar com relações não-lineares, e que as regressões logística e linear têm a vantagem de gerar modelos mais robustos, visto que testes de significância são preliminarmente executados.

Silva (2006) conclui que na literatura disponível inexistem técnicas unânimes apontadas como a melhor em termos de eficácia na previsão. Ressalta que em muitas vezes, a aplicação de técnicas diferentes leva a resultados semelhantes. Assim, fica evidente que em novos modelos, o aconselhável é testar todas as metodologias, escolhendo aquela que for mais conveniente, pois os melhores resultados dependem do fenômeno e da base de dados estudada.

1.2 Acompanhamento de Modelos

O Novo Acordo de Basiléia exige que os bancos disponham de sistemas robustos para validação da precisão e coerência dos sistemas e processos de escoragem. A validação também é essencial para fins de governança corporativa, contribuindo na identificação de falhas de desempenho do modelo, que podem afetar os limites existentes de tolerância ao risco e a alocação do capital econômico[21].

[21] KARAKOULAS, Grigoris. **Validação empírica de modelos de *credit scoring***. Revista SERASA. Disponível em http://www.serasa.com.br/ingles/i_revista/i_revista1.htm. Acesso em: 01 set. 2008.

As formas de acompanhamento comumente utilizadas em modelos de risco de crédito dizem respeito à:

- Acurácia ou desempenho: indicadores usados para medir o poder discriminatório dos modelos. Como exemplo, cita-se o Teste Kolgomorov-Smirnov, Curva ROC, Teste Hosmer-Lemeshow etc;

- Aderência: índices que mensuram a estabilidade da população com a qual os modelos foram desenvolvidos;

- Utilização de técnicas específicas para monitoramento dos modelos após sua implantação.

Este capítulo traz considerações relevantes que precisam ser levadas em conta sobre metodologias de validação dos modelos, além do problema de viés de seleção, mencionado na introdução deste trabalho e dos indicadores de acurácia Kolgomorov-Smirnov, Curva ROC e Hosmer Lemeshow, os quais serão abordados posteriormente.

1.2.1 Estabilidade da População

Uma questão que deve ser avaliada nos modelos de *credit scoring* diz respeito às flutuações populacionais, a qual descreve a tendência das populações em evoluírem, podendo ter suas características alteradas ao longo do tempo. Esta questão traz impactos, visto haver mudanças na distribuição das variáveis com base nas flutuações econômicas e mudanças no ambiente competitivo. Os relatórios de estabilidade da população visam identificar diferenças significativas entre a base de modelagem e a população atual, sobre a qual os modelos estão sendo aplicados, ou seja, verificar se há evidência estatística de uma grande alteração entre as populações de desenvolvimento e a atual.

Em Lucumberri e Duarte Júnior (2003) são abordadas técnicas de monitoramento de modelos de *credit scoring*, como o IEP (Índice de Estabilidade da População) e CA (Característica Amostral). O primeiro é calculado com base na distribuição de freqüência das classes de pontuação, utilizando dez classes eqüidistantes e busca identificar mudanças na

distribuição das populações de desenvolvimento e de análise. A intuição do IEP é simples: quanto mais diferentes forem as populações em consideração, maior será o IEP calculado. O ideal é que o valor do IEP seja o menor possível, próximo de zero. Segundo os autores, embora haja valores considerados aceitáveis para este índice como (IEP $\leq$ 0,10), e outros considerados muito elevados (como IEP $\geq$ 0,25), sugere-se a observação da tendência do indicador ao longo do tempo: se aumentando, indica maior divergência entre as populações, o que deve ser visto como ruim; se diminuindo, indica menor divergência entre as populações, o que deve ser visto como bom.

O CA compara a distribuição dos atributos da cada variável explicativa na base de desenvolvimento ao longo de sua implantação, buscando identificar em quais variáveis ocorreram mudanças. A principal diferença entre o IEP e CA é que o primeiro capta alterações na distribuição da escoragem e o segundo capta mudanças nos atributos dos modelos. Os autores alertam, ainda, que a equipe de desenvolvimento deverá decidir a necessidade de calibragem ou substituição do modelo, haja vista não haver um valor crítico estabelecido para o CA. Lucumberri e Duarte Júnior finalizam com a recomendação de que o IEP e CA sejam utilizados de forma conjunta no acompanhamento do modelo, visto fornecerem informações diferentes.

Chinelatto Neto, Felício e Campos (2007) apresentam técnicas que substituem o CA, agregando vantagens não contempladas naquele indicador. Os autores argumentam que a Característica Amostral não é capaz de mensurar os impactos das mudanças das características das variáveis sobre a pontuação média, visto que apenas demonstra a existência de mudanças e sinaliza qual sentido elas estão ocorrendo. Outra desvantagem do CA consiste na impossibilidade de identificação das mudanças nos atributos do modelo caso as mesmas se compensem em cada variável ou grupo de variáveis. Como alternativa, os autores propõem outros três indicadores: o Efeito de Variação das Freqüências, a Variação Amostral e o Efeito Compensação. Resumidamente, propõem uma inovação metodológica capaz de mensurar os fatores não captados pelo CA, podendo ser utilizado o IEP como indicador complementar ao monitoramento do modelo. [22]

[22] Para uma leitura mais detalhada sobre o assunto de flutuações populacionais, recomenda-se fortemente o trabalho de CHINELATTO NETO, FELICIO, A.; CAMPOS, D. op. Cit.

1.2.2 Relatórios de Acompanhamento

Embora a literatura cobrindo técnicas sobre o desenvolvimento e implementação de modelos de risco de crédito seja abrangente, não se pode dizer o mesmo a respeito da literatura sobre métodos de gerenciamento destes após sua implementação. A seguir, será apresentado o resumo de algumas das práticas abordadas no estudo de Lucumberri e Duarte Júnior.

1.2.2.1 – Relatório de Inadimplência

O relatório de inadimplência acompanha a proporção de contratos inadimplentes com mais de 30 dias de atraso no pagamento de suas obrigações no decorrer do tempo após a implantação do modelo. Via de regra, é de se esperar que a partir do segundo mês de vigência do modelo a taxa de inadimplência se reduza a patamares inferiores aos observados antes de sua implantação. Se, por exemplo, após a implantação do modelo for observado um crescimento ou, até mesmo, a manutenção da proporção de operações inadimplentes, pode ser um indício de que o modelo *credit scoring* perdeu sua capacidade preditiva e deve ser revisto. Contudo, devem ser observados fatores externos ao modelo que possam causar oscilações nas taxas de inadimplência, a exemplo de mudanças no *score* de corte, campanhas de marketing que elevem a quantidade de concessões em determinado período, períodos de recessão econômica etc. Assim, deve-se também analisar outros relatórios de acompanhamento, de forma a saber exatamente a origem do comportamento da inadimplência.

1.2.2.2 – Relatório de Escoragem Final (REF)

Este relatório monitora a tomada de decisão com base na escoragem, ou seja, após estabelecer as regras de concessão de crédito é feito o acompanhamento dos percentuais de escoragem em cada faixa de aprovação, podendo haver detalhamentos por produto, região geográfica, por agência etc. A instituição financeira pode, por exemplo, decidir que somente concederá crédito a determinadas faixas de *score* ao invés de basear-se no tradicional ponto de corte. Pode-se então acompanhar a evolução do quantitativo de operações em cada faixa de *score*, como forma de diagnosticar e isolar fatores responsáveis por esta oscilação.

1.2.2.3 – Relatório de Interferência de Escoragem (RIE)

A interferência de escoragem pode ocorrer de duas formas:

- Quando um cliente que deveria ter tido sua proposta de crédito reprovada após a escoragem, é, ao final, aceito. Ou seja, seu *score* situa-se abaixo do ponto de corte, mas mesmo assim uma proposta de crédito é feita ao mesmo devido a outros motivos;

- Quando um cliente avaliado com alta pontuação, onde deveria ter seu crédito aprovado, mas que ao final teve sua proposta não efetivada por outros motivos. A quantidade de negócios que deixaram de ser efetuadas para clientes com pontuação acima do ponto de corte deve ser avaliada para identificar possíveis falhas no processo de concessão de crédito.

Os motivos de aprovação ou reprovação citados acima, podem estar relacionados a possíveis filtros praticados pela instituição financeira em sua política de crédito, os quais estejam sendo muito severos no decorrer do processo, ou até mesmo à desistência do cliente na conclusão do negócio. Esse relatório é decorrente do REF, pois visa identificar os principais pontos onde a decisão final foi contrária à decisão do modelo de escoragem, de forma a isolar as eventuais deficiências das políticas de crédito utilizadas.

Consideremos um exemplo onde um banco está interessado em prospectar clientes no setor universitário. É possível, pela natureza do público, que o modelo de *credit scoring* recuse uma boa parte da população avaliada. Para aumento da participação do banco naquele segmento é então necessário calibrar os resultados do modelo, ofertando crédito a universitários que não obtiveram *score* suficiente para aprovação.

O RIE dá então uma visão ao grupo de gestão de risco de crédito das decisões tomadas que contrariam os modelos de escoragem, podendo ser utilizado nas áreas de Auditoria Interna e Revisão de Crédito no Varejo.

1.2.2.4 – Relatório de Desempenho do Modelo (RDM)

O RDM busca comparar, para uma safra de indivíduos, as distribuições de maus clientes entre as amostras de desenvolvimento do modelo e a atual. Como no caso do IEP, é recomendável utilizar uma medida de divergência para acompanhar ao longo do tempo como as distribuições de maus clientes evoluem, observando se estas convergem no período de maturação. A seguir, um exemplo prático para um modelo de *behaviour scoring* cujo período de observação é de seis meses:

TABELA 2 - Relatório de Desempenho

Classe de *Score*	**Maus – base de modelagem**		**Maus - 2 meses**		**Maus - 4 meses**		**Maus - 6 meses**	
	Qtde	**%**	**Qtde**	**%**	**Qtde**	**%**	**Qtde**	**%**
até 80	4.193	19,8%	114	23,0%	469	21,7%	1.967	19,6%
81 a 100	7.672	36,2%	123	24,8%	679	31,4%	3.762	37,5%
101 a 130	6.144	29,0%	142	28,6%	623	28,8%	2.961	29,5%
131 a 165	2.625	12,4%	89	17,9%	324	15,0%	1.183	11,8%
> 166	554	2,6%	28	5,6%	66	3,1%	166	1,7%
Total	21.188	100,0%	496	100,0%	2.161	100,0%	10.039	100,0%

De maneira análoga ao IEP, a comparação é feita por faixas de *score*, muito embora isso não seja imperativo. Nota-se que a divergência total diminui com o passar do tempo, refletindo que as distribuições aproximam-se quando o período de observação embutido no modelo (no caso, seis meses) termina. Entretanto, caso não houvesse uma tendência das duas distribuições citadas aproximarem-se, com as estimativas totais diminuindo, medidas corretivas deveriam ser consideradas.

1.2.2.5 – Relatório de Desempenho da Escoragem (RDE)

O RDE pode ser interpretado da seguinte forma: mede, para cada safra, a quantidade de bons clientes contratada para cada mau cliente contratado. A tabela 3 ilustra a utilização do RDE para um caso de modelo de *behaviour scoring*:

TABELA 3 - Razão de Bons e Maus

Classe de *Score*	Maus Desenvolvimento		Maus - mês 1		Maus = mês 2		Maus - mês 3	
	Qtde	Razão (B/M)	Qtde	Razão (B/M)	Qtde	Razão (B/M)	Qtde	Razão (B/M)
até 80	4.193	4,9	114	5,0	469	3,70	1.967	1,3
81 a 100	7.672	7,7	123	7,7	679	6,40	3.762	6,0
101 a 130	6.144	17,0	142	17,4	623	15,90	2.961	11,9
131 a 165	2.625	74,2	89	74,3	324	77,80	1.183	64,3
> 166	554	568,2	28	568,1	66	564,00	166	425,4
Total	21.188		496		2.161		10.039	

É visível que com o passar do tempo há uma piora generalizada na razão de bons e maus clientes. Por exemplo: analisando as observações de um mês com as observações de três meses, a razão decai para todas as faixas de *score* considerados, o que indica a piora generalizada do modelo de escoragem em questão. Há, portanto, um indício de que o modelo considerado necessita ser revisado.

2. Base de Dados

Na formulação de modelos *credit scoring* normalmente são utilizadas bases de dados de tamanho considerável, contendo mais de 100.000 observações e mais de 100 variáveis explicativas, qualquer que seja a técnica estatística utilizada[23].

No presente estudo serão utilizadas duas bases de dados. A primeira será utilizada para a classificação das operações de crédito em boas ou ruins. De acordo com a forma tradicional de classificação das operações, será utilizado o princípio de que o risco da operação é determinado pelas possibilidades de ocorrerem atrasos nos pagamentos das prestações do empréstimo concedido. Assim, a primeira base de dados requer informações mensais sobre a quantidade de dias em atraso de todos os contratos disponibilizados para análise. Se o princípio do risco for baseado na questão da lucratividade, ou *profit scoring*, serão necessários todos os dados financeiros embutidos na operação como valor total, valor dos juros, valor de encargos por atraso, juros de mora em cada prestação etc., o que pode tornar a base de dados demasiadamente grande, comprometendo a viabilidade técnica da elaboração do modelo.

As maiores dificuldades inerentes ao primeiro grupo de dados refere-se à escolha do critério para classificação das operações e à obtenção dos dados de alguma instituição de crédito. O critério escolhido para o presente trabalho consiste na classificação das operações baseado na quantidade de dias em atraso das prestações, devido à dificuldade de obtenção dos dados necessários para utilização do critério de *profit scoring*, por revelar os resultados financeiros da carteira de crédito em questão. Desta forma, a instituição concordou em ceder somente os dados para classificação de acordo com a inadimplência.

A segunda base de dados é formada por informações de perfil e comportamentais dos indivíduos, as quais constituem variáveis explicativas dos modelos em análise e serão estudadas de acordo com sua influência sobre a variável resposta - a qualidade de crédito e o tempo até a inadimplência. As variáveis mais comumente utilizadas variam conforme a linha de crédito em estudo, podendo englobar não somente as características dos clientes (idade,

[23] HAND, D. J. and HENLEY D. J. Statistical Classifications Methods in Consumer Credit Scoring: a Review**. Journal of the Royal Statistical Society Series**, 1997

quantidade de dependentes, sexo, patrimônio, informações econômico-financeiras, etc), mas também as da operação em si (prazo, valor contratado, forma de pagamento). Operações de curto prazo, destinadas ao público de baixa renda, apresentam variáveis distintas de um financiamento habitacional, caracterizado por longos prazos e prestações mais altas. As variáveis também podem ser diferentes de acordo com o tipo de pessoa a quem se concede – pessoa física ou jurídica (empresa de micro, pequeno, médio e grande porte)[24]. A tabela 4 traz um exemplo de variáveis comumente utilizadas em modelos de *credit scoring* para pessoas físicas:

TABELA 4 - Exemplos de variáveis explicativas

Natureza e tipo	Variáveis	Tipo de resposta
Cadastro	Idade	Anos
	Sexo	Feminino/masculino
	Estado civil	Codificada
	Regime de casamento	Codificada
	Tempo de residência atual	Meses
	Tempo de residência própria	Sim/não
	Escolaridade	Codificada
	Quantidade de dependentes	Números (00, 01, 02, ...)
	Profissão e/ou ocupação	Codificada
	Profissão e/ou ocupação do cônjuge	Codificada
	Tempo de emprego atual	Meses
Renda	Salário líquido	R$
	Outros rendimentos mensais	R$
	Salário líquido do cônjuge	R$
	Renda familiar total	R$
Patrimônio (quantidade e valor com indicação de alienação, comprovação, hipoteca)	Automóveis	Número e R$
	Imóveis	Número e R$
	Outros bens	R$
Informações bancárias do cliente	Data de abertura da conta	dd/mm/aaaa
	Tipo de conta	Codificada
	Saldo médio em conta corrente	R$
	Saldo médio de aplicações	R$
	Cartão de crédito	sim/não
	Bloqueio/restrições	sim/não
Compromissos financeiros (aluguel, educação, financiamento, empréstimo)	Tipo	Codificada
	Periodicidade do pagamento	Codificada (ex: 01 = mensal)
	Valor nominal	R$
	Número de parcelas a vencer	Números (00, 01, 02, ...)
Dados da operação	Valor da operação de	R$

[24] VASCONCELLOS op. cit.

Natureza e tipo	Variáveis	Tipo de resposta
	empréstimos	
	Quantidade de prestações	Números (00, 01, 02, ...)
	Forma de pagamento da prestação	Codificada (ex: carnê)
	Forma de pagamento de impostos	Codificado (ex: 01 = financiado)
	Forma de cobrança de tarifas	Codificado (ex: C = à vista)
	Finalidade da operação	Codificado
	Taxa de juros (1)	dd/mm/aaaa
	Referência monetária	dd/mm/aaaa
Apontamentos negativos (2) - com registros de datas, valores, quantidades e datas de regularização, caso haja	Protestos	Codificado
	Cheques devolvidos	Codificado
	Ações judiciais	Codificado
	Pendências financeiras	Codificado
	Cheques irregulares sem fundo	Codificado

(1): A necessidade de se ter informações sobre a taxa de juros da operação é discutível para métodos que propõem analisar pedidos de crédito, uma vez que podem ser definidas após a aprovação de crédito, podendo depender, inclusive, do *score* obtido pelo cliente. Assim, a taxa de juros não pode ser utilizada como variável explicativa, pois depende da variável resposta do modelo.
(2): também conhecido como "Restrições Cadastrais", são informações obtidas junto às agências reguladoras e de proteção ao crédito como o SPC e SERASA.

Como já mencionado, as variáveis de perfil, como idade, sexo, renda, bens patrimoniais, etc. são computadas mediante o preenchimento de fichas cadastrais pelos clientes no momento em que o crédito é requisitado. Usualmente, os bancos solicitam aos clientes a comprovação da veracidade das informações prestadas na ficha cadastral, o que em muitos casos gera insatisfação por parte dos mesmos devido à burocracia e irritação em preencher longas fichas cadastrais, além de apresentar documentos que comprovem renda e patrimônio. Tais exigências podem fazer com que um cliente de baixo risco, que normalmente possui acesso a várias linhas de crédito em outras instituições, recorra ao crédito fácil e menos burocrático. Se para a instituição credora é de suma importância possuir cadastros completos que forneçam a maior quantidade de informações possíveis sobre seus clientes, para estes torna-se oneroso dispender tempo e paciência em fornecer e comprovar estas informações. Variáveis do tipo comportamentais, como saldo médio de conta corrente, poupança, aplicações, quantidade de cheques devolvidos etc. são computadas pelas próprias instituições com base no relacionamento que têm com seus clientes, sem a necessidade de fichas cadastrais para obtê-las.

Um importante fator a ser analisado no banco de dados refere-se ao tratamento das variáveis que têm como resposta a ausência de valores, os chamados *missing values*. A

ausência de resposta deve-se ao fato do cliente não preencher o questionamento da ficha de entrevista cadastral ou, ainda, devido aos casos em que a resposta inexiste por não ser necessária. Como exemplo, cita-se as situações nas quais o cliente não informa se possui algum tipo de seguro, cartão de crédito, automóvel etc. e os casos em que a resposta é condicional à resposta de outra variável (a renda dos dependentes não é informada devido ao fato do cliente não possuir dependentes). Alguns estudos abordam diferentes formas de se lidar com os *missing values*, os quais redundam em eliminar da base aqueles que apresentam ausência de respostas em qualquer um das variáveis, eliminar as variáveis que apresentam pelo menos um cliente com ausência de resposta ou, ainda, utilizar algoritmos estatísticos na tentativa de estimar valores que possam substituir os *missing values*.

A alternativa adotada no presente estudo para o tratamento dos *missing values* foi a de codificar a ausência de registros como uma resposta válida para cada variável. Desta forma, optou-se por atribuir o valor 0 (zero) à ausência de resposta das variáveis qualitativas. Por exemplo, para a variável “escolaridade” temos as respostas: 0 – missing, 1 – analfabeto, 2 - ensino fundamental incompleto, 3 – ensino fundamental completo, etc. A justificativa para adoção deste procedimento baseia-se no fato de que a ausência de resposta pode funcionar como uma resposta válida e capaz de discriminar bons e maus clientes. Outro tratamento dado às variáveis foi o de eliminar do banco de dados as variáveis com alto grau de *missing*, de forma a não se prover um estudo baseado na ausência de informações.

2.1 Período da amostra e variáveis explicativas coletadas

Conseguir uma base de dados com todas as informações necessárias para o desenvolvimento um modelo de *credit scoring* não é uma tarefa trivial. Conforme exposto, as instituições financeiras têm alta confidencialidade em seus bancos de dados, as quais dispõem de informações sigilosas de seus clientes como renda, restrições cadastrais e endividamento. Outro dificultador consiste no receio que as instituições de crédito têm em divulgar abertamente os critérios e pesos de cada variável na composição da nota final de *score* de seu modelo de risco. Assim, conhecendo os critérios de classificação que determinada entidade utiliza para aprovar seus clientes, poder-se-ia manipular os dados, direcionado respostas às variáveis que sabidamente elevam o *score* do cliente para que o mesmo seja aprovado.

A alternativa encontrada foi uma instituição financeira que fornece diversas linhas de crédito a pessoas físicas e jurídicas e que concordou em ceder os dados necessários para consecução do presente estudo. A instituição, que conta com um sistema de informática robusto e com alta capacidade de armazenamento, forneceu os dados de uma de suas linhas de crédito destinadas exclusivamente a pessoas físicas, desde que houvesse a descaracterização do nome de algumas das variáveis relevantes dos modelos a serem elaborados, que fosse criado um nome fictício para a carteira de crédito e que a identidade da instituição fosse preservada. Esta prática busca não tornar os resultados financeiros diretamente visíveis, bem como manter o sigilo das práticas de concessão de crédito.

A população do estudo engloba os clientes que contrataram operações de crédito do tipo rotativo nos sete primeiros meses de 2005 e que não possuíam restrições cadastrais no momento da contratação. Dessa população foi extraída uma amostra de 171.461 contratações, gerando assim a base de dados a ser utilizada. Para cada um dos contratos foram obtidas diversas variáveis cadastrais e de comportamento do cliente, as quais serão parcialmente codificadas de acordo com as exigências da instituição cedente dos dados. Assim, temos 60 potenciais variáveis preditoras à qualidade de crédito e tempo até a inadimplência, segregadas em 2 naturezas e 6 tipos, dispostas na tabela abaixo:

TABELA 5 - Variáveis explicativas parcialmente codificadas para formulação dos modelos de credit scoring e análise de sobrevivência

Natureza	Tipo	Nome Codificado
PERFIL	Patrimonial: tipo A	Possui Plano de Saúde
		a1
		a2
		a3
		Qtde Antenas Parabólicas
		a4
		Qtde Freezer
		a5
		a6
		a7
		Qtde Máquina Lavar Louça
		Qtde Microondas
		a8
		Qtde Telefone Comum
		a9
		a10
		a11
		Situação do Veículo

Natureza	Tipo	Nome Codificado
		Ano do Veículo
		Valor de Mercado Veículo
		Tipo de Imóvel
		a12
		a13
		a14
		Qtde Computador
	Cadastral (Demográfica): tipo B	Grau de Instrução
		Idade
		Nacionalidade
		Sexo
		Estado Civil
		UF
		b1
		b2
		b3
	Financeira - Renda: tipo C	c1
		c2
		Renda Líquida Formal
		c3
		c4
		Renda Informal
		c5
	Financeira - Informações bancárias: tipo D	d1
		saldo médio das aplicações em outras instituições
		d2
		d3
		d4
		Valor Médio da Fatura do Cartão de Crédito
		Qtde Anos Associado ao Cartão de Crédito
		Qtde Cartões de Crédito
		Qtde Participações Societárias
		d5
		d6
		d7
COMPORTAMENTAL	Apontamentos negativos: tipo E	e1
		e2
		e3
		Valor Médio dos Excessos de Limites
		Valor dos Cheques Devolvidos Motivo 11
		e4
	Cadastral: tipo F	Qtde Dias Abertura Conta

Observa-se que as variáveis de perfil estão segregadas em 6 tipos: as do tipo patrimonial foram codificadas como tipo A; as do tipo demográficas foram codificadas como tipo B; as do tipo C são as relativas à renda, enquanto as do tipo D referem-se às informações bancárias do cliente. As variáveis de natureza comportamental segregam-se em 2 tipos: apontamentos negativos (tipo E) e cadastrais (tipo F). Nota-se que algumas variáveis não foram codificadas, permanecendo com os nomes originais. Nestes casos a instituição que cedeu o banco de dados concordou em não codificá-las, uma vez que são notoriamente utilizadas nos modelos de *credit scoring* e presentes em grande parte dos estudos.

Importante ressaltar que nem todas as informações relativas à renda foram utilizadas. A variável “renda bruta formal” e “renda líquida formal” apresentaram alta correlação entre si e por isso optou-se em utilizar somente a renda líquida. O problema em se ter variáveis explicativas altamente correlacionadas entre si é a ocorrência de multicolinearidade, o que gera grandes covariâncias, variâncias e erros padrões. A conseqüência disto é a dificuldade na obtenção de estimadores precisos, uma vez que os intervalos de confiança tendem a ser maiores, resultando na aceitação da hipótese nula mais prontamente. Além disso, os estimadores e seus erros padrões podem ser sensíveis a pequenas variações nos dados.

O banco de dados coletado possui as variáveis mais importantes para a formação do modelo de análise de sobrevivência e *credit scoring*. Entretanto, algumas variáveis relevantes na modelagem não foram disponibilizadas, dentre as quais devem ser citadas:

- Crédito salário na instituição: por questões de sigilo, também não foi disponibilizado no banco de dados a informação que indica se o cliente recebe seus proventos em conta bancária da instituição que cedeu os dados. A falta de informações completas sobre o perfil financeiro pode reduzir a capacidade de discriminação do modelo de concessão, uma vez que torna o perfil do cliente incompleto;

- Saldo médio em conta corrente, poupança e demais aplicações: a inexistência de informações relativas a saldos médios pode prejudicar a acurácia do modelo, uma vez que as mesmas exprimem o comportamento poupador do potencial tomador de crédito.

- Valor médio de utilização do cheque especial: da mesma forma que saldos médios, a existência variáveis que exprimam o valor da utilização do cheque especial demonstram o comportamento do cliente. Sua omissão pode prejudicar a capacidade preditiva do modelo;

- Apontamentos negativos: também não foram disponibilizadas informações sobre agências de proteção ao crédito, como SERASA e SPC. Consequentemente, não foram disponibilizadas informações sobre protestos, ações judiciais, pendências financeiras, entre outras.

A ausência destas informações diminui a capacidade discriminatória do modelo, o que é atenuado pela presença de diversas outras informações que podem ser usadas como substitutas para grande parte das informações não disponibilizadas.

Para que se possa executar modelos de regressão é necessário gerar um arquivo em formato apropriado para carregá-lo no software estatístico a ser utilizado. Este arquivo deve conter as observações que compõem a amostra para modelagem, além da variável dependente e todas as potenciais variáveis explicativas a serem testadas. Para se chegar a este arquivo apropriado, muitas vezes é necessário trabalhar a base de dados. O anexo A traz considerações sobre o tratamento da base de dados utilizado na geração dos modelos finais.

2.2 Softwares Utilizados

O software utilizado no tratamento da base de dados, sumarizações, cruzamento de tabelas e geração do arquivo apropriado para ser carregado no software estatístico foi o SQL SERVER 2000, devido a sua alta capacidade de armazenamento e robustez em lidar com grande quantidade de registros. Os procedimentos estatísticos para geração do modelo de *credit scoring*, categorização das variáveis e análise de sobrevivência foram executados no SPSS 13.0 (*Statistical Package for Social Sciences*) devido à existência de todas as funções estatísticas necessárias para o trabalho.

3. Metodologia

Abordadas as questões sobre o tratamento da base de dados, etapa indispensável na geração do arquivo a ser utilizado nas regressões, passemos agora à discussão sobre os passos necessários à formulação de um modelo *credit scoring* e análise de sobrevivência.

3.1 Definição da qualidade de crédito

O primeiro passo na elaboração de um modelo de risco de crédito após a obtenção e organização do banco de dados é definir qual a variável resposta do modelo e como ela pode ser obtida. Conforme visto em Einsembeis (1977), faz-se uma divisão dicotômica de dois grupos mutuamente excludentes em bons créditos (aqueles cujas prestações foram pagas em dia ou com poucos atrasos) e maus créditos (aqueles cujas prestações foram pagas com maiores atrasos). Os bons créditos são vistos como baixo risco, enquanto os maus créditos são os de alto risco. Essa definição é feita com base no histórico dos créditos já concedidos pela instituição, na qual são analisados o período de atraso no pagamento de cada prestação e a migração para atrasos ainda maiores. Desta forma é possível gerar os limites de atrasos aceitáveis para classificação da operação como boa ou ruim, de acordo com a inadimplência da carteira de crédito. Determinados os limites de atrasos, gera-se a variável resposta do modelo, que é a qualidade do crédito.

Para se calcular a quantidade de dias em atraso de cada operação, basta calcular a quantidade decorrida de dias entre a data de vencimento e de pagamento de cada prestação, caso a carteira estudada seja baseada em créditos pagos com prestações mensais. No caso de créditos rotativos como cheque especial e cartão de crédito, onde não há prestações, o procedimento é ligeiramente diferenciado:

1º passo: ao final de cada mês é tirada uma “fotografia” de cada contrato, a fim de verificar sua situação de adimplência;

2º passo: caso o contrato esteja em dia, ou seja, o limite de crédito está sendo parcialmente ou não utilizado, é computado 0 (zero) dia de atraso;

3º passo: caso o contrato esteja em atraso, ou seja, o limite de crédito está sendo utilizado além de seu limite, é computada a quantidade de dias de atraso entre a ocorrência do último excesso de limite e a data em que foi tirada a fotografia do contrato.

Cada atraso é então agrupado por faixa de atraso de 30 em 30 dias: atrasos de 0 dia pertencem à faixa de atraso “0”, atrasos de 1 a 30 dias pertencem à faixa de atraso “1-30” e assim por diante, até “180 dias ou mais”, uma vez que contratos com mais de 180 dias são considerados como perda ou processo de liquidação, com pequenas chances de recuperação financeira. A tabela 6 traz um exemplo do cômputo da quantidade de dias de atraso de um crédito rotativo hipotético, e em seguida são feitas algumas considerações a respeito:

TABELA 6 - Exemplo de cômputo da quantidade de dias em atraso

Data da fotografia	Data de ocorrência do último excesso de limite	Quantidade de dias em atraso	Faixa de dias em atraso
31/01/2007	-	0	0
28/02/2007	15/02/2007	13	1 a 30
31/03/2007	15/03/2007	16	1 a 30
30/04/2007	25/04/2007	0	0
31/05/2007	05/05/2007	26	1 a 30
30/06/2007	05/05/2007	56	31 a 60
31/07/2007	05/05/2007	87	61 a 90

- Na primeira fotografia não houve excesso de limite, ou seja, o cliente não utilizava ou utilizava parcialmente seu limite de crédito. Conforme descrito no 2º passo, computa-se 0 (zero) dia de atraso para o mês de jan/07;

- Na segunda fotografia, a partir de 15/02/2007 o cliente incorreu em excesso de limite, passando a utilizar além do seu limite de crédito rotativo disponível. Conforme descrito no 3º passo, é computada a quantidade de dias em atraso entre a data de ocorrência do último excesso de limite (13 dias para o mês de fev/07);

- Na quarta fotografia, apesar do cliente ter incorrido em excesso de limite em 25/04/2007, na data da fotografia o contrato não estava em atraso, o que significa que o cliente efetuou o pagamento do valor que estava em excesso de limite a descoberto (0 dia para o mês de abr/07);

- Na quinta fotografia, o cliente incorreu em excesso de limite por 26 dias de excesso entre a data de excesso de limite e a data de fotografia, computando-se este número para o mês de mai/07;

- Foram computados 56 dias de atraso para o mês de jun/07;

- Raciocínio análogo é utilizado aos demais meses, ou seja, calcula-se a quantidade de dias em atraso da data de ocorrência do último excesso de limite e a data de fotografia do contrato.

Calculados os atrasos existentes em cada contrato analisado, o passo seguinte é gerar tabelas cruzadas, conhecidas como Matrizes de Markov, com a quantidade de contratos que evoluíram para cada uma das faixas de atrasos. A tabela 7 traz um exemplo de tabela cruzada na qual foram relacionados os atrasos do 5º mês para o 6º mês em que iniciou-se a contagem de atrasos da carteira de crédito do presente trabalho.

TABELA 7 - Tabela cruzada para quantidade de contratos em atraso do mês nº. 5 para o mês de nº. 6

		atraso na data de vencimento do mês nº 6 (dias corridos)								
		0	1 a 30	31 a 60	61 a 90	91 a 120	121 a 150	151 a 180	>180	total
atraso na data de vencimento do mês nº 5 (dias corridos)	0	169.014	9.226	2.031	38	.	.	.	.	180.309
	%	93,7%	5,1%	1,1%	0,0%	.	.	.	.	100%
	1 a 30	4.161	2.087	1.899	598		.	.	.	8.745
	%	47,6%	23,9%	21,7%	6,8%	.	.	.	.	100%
	31 a 60	867	199	7	3.353	.	.	.	.	4.426
	%	19,6%	4,5%	0,2%	75,8%	.	.	.	.	100%
	61 a 90	292	10	.	318	3.937	238	.	.	4.795
	%	6,1%	0,2%	.	6,6%	82,1%	5,0%	.	.	100%
	91 a 120	102	.	.	.	17	2.460	623	.	3.202
	%	3,2%	.	.	.	0,5%	76,8%	19,5%	.	100%
	121 a 150	79	.	.	.	.	209	2.631	110	3.029
	%	2,6%	.	.	.	.	6,9%	86,9%	3,6%	100%
	151 a 180	19	.	.	.	.	.	.	1.124	1.143
	%	1,7%	.	.	.	.	.	.	98,3%	100%
	>180	.	.	.	.	.	.	.	8	8
	%	.	.	.	.	.	.	.	100%	100%
	total	174.534	11.522	3.937	4.307	3.954	2.907	3.254	1.242	205.657
		84,9%	5,6%	1,9%	2,1%	1,9%	1,4%	1,6%	0,6%	100,0%

Observa-se, por exemplo, que dos 180.309 contratos que atingiram o 5º mês de vigência com zero dia de atraso, 9.226 (5,1%) evoluíram para um atraso de 1 a 30 dias no 6º mês de vigência. Podemos interpretar que 5,1% é a probabilidade destes contratos migrarem para uma faixa de atraso maior. Já os 8.745 contratos que atingiram o 5º mês de vigência com atraso de 1 a 30 dias, 1.899 (21,7%) evoluíram para um atraso ainda maior, de 31 a 60 dias.

De forma análoga, dos 4.426 contratos com atrasos de 31 a 60 dias, 3.353 (75,8%) evoluíram para uma faixa maior de atraso, de 61 a 90 dias. Dos 4.795 contratos com atraso de 61 a 90 dias, 3.937 (82,1%) evoluíram para uma faixa maior de atraso, de 91 a 120 dias. Nota-se, portanto, que quanto maior for o atraso, maior é a probabilidade de estes contratos evoluírem para atrasos ainda maiores. Pode-se também analisar a probabilidade de reversão de inadimplência, como por exemplo: dos 8.745 contratos que atingiram o 5º mês com atraso de 1 a 30 dias, 4.161 (47,6%) regrediram para 0 dia de atraso. Da mesma forma, dos 4.426 contratos com 31 a 60 dias de atraso, 24,1% (19,6% + 4,5%) regrediram para uma faixa menor de atraso. Já dos 4.795 contratos com 61 a 90 dias de atraso, somente 6,3% regrediram para atrasos inferiores. A conclusão obtida por meio desta última análise é que quanto maior for o atraso no 5º mês, menor é a probabilidade de reversão de inadimplência no mês seguinte.

Aplicando o raciocínio anterior em todos os meses dos contratos da carteira de crédito em estudo, pode-se calcular as probabilidades de evolução de uma faixa de atraso para a faixa seguinte, a qual está expressa na tabela seguinte:

TABELA 8 - Probabilidade de evolução de faixas de atraso mensal

	1° para 2°	2° para 3°	3° para 4°	4° para 5°	5° para 6°	6° para 7°	7° para 8°	8° para 9°	9° para 10°	10° para 11°	11° para 12°	probabilidade média de migração para > inadimplência	probabilidade média de manutenção ou reversão da inadimplência
0 para 31 a 60	1,21%	1,08%	0,88%	1,14%	1,12%	0,75%	0,75%	0,58%	0,42%	0,64%	0,62%	0,83%	99,17%
1 a 7 para 31 a 60	6,75%	12,76%	13,47%	10,86%	11,79%	12,00%	9,26%	7,79%	9,92%	8,75%	8,10%	10,13%	89,87%
8 a 15 para 31 a 60	40,12%	32,42%	28,61%	28,43%	27,33%	24,86%	23,43%	23,89%	20,70%	20,92%	19,94%	26,42%	73,58%
16 a 30 para 31 a 60	46,81%	41,80%	40,10%	36,39%	30,55%	39,11%	30,56%	26,09%	31,28%	26,42%	24,60%	33,97%	66,03%
31 a 37 para 61 a 90	71,72%	72,90%	70,98%	67,72%	67,26%	68,23%	63,57%	55,60%	61,74%	62,68%	71,17%	66,69%	33,31%
38 a 45 para 61 a 90	88,31%	86,80%	86,78%	57,49%	81,71%	87,96%	77,22%	82,07%	81,07%	73,61%	76,38%	79,94%	20,06%
46 a 60 para 61 a 90	93,15%	90,41%	91,94%	88,04%	85,97%	86,17%	85,38%	74,28%	61,39%	83,73%	82,44%	83,90%	16,10%
31 a 60 para 61 a 90	.	77,83%	78,61%	76,83%	75,10%	74,63%	74,02%	75,53%	73,93%	74,65%	71,95%	75,31%	24,69%
61 a 90 para 91 a 120	.	81,05%	73,35%	70,57%	82,02%	73,05%	71,66%	72,83%	69,08%	73,03%	74,18%	74,08%	25,92%
91 a 120 para 121 a 150	.	.	94,92%	89,56%	76,83%	72,39%	76,86%	74,34%	72,41%	74,19%	77,49%	78,78%	21,22%
121 a 150 para 151 a 180	.	.	.	88,19%	86,83%	88,41%	92,58%	91,74%	82,77%	82,23%	82,78%	86,94%	13,06%
151 a 180 para 181 ou mais	.	.	.	.	98,34%	87,28%	86,96%	87,71%	87,13%	81,65%	88,20%	88,18%	11,82%

Verifica-se que, independente dos meses em questão, quanto maior é a faixa de atraso que um contrato alcança, maior é a probabilidade de evolução para um atraso ainda maior na prestação seguinte, logo, menor é a probabilidade de reversão de inadimplência. Contratos que chegam a determinada prestação com 0 dia de atraso têm baixa probabilidade (inferiores a 2%) de migrarem para atrasos de 31 a 60 dias, enquanto que contratos que

atingem uma prestação com atraso de 1 a 7 dias também apresentam baixa probabilidade (10,13%) de atingirem a prestação seguinte com atraso de 31 a 60 dias.

A faixa de atraso de 8 a 15 dias apresenta em média 26,42% de probabilidade de migração para uma faixa maior de atraso. Seguindo este raciocínio, constata-se que contratos que atingem atrasos a partir de 31 dias têm consideravelmente maiores probabilidades de migrarem para uma faixa de atraso ainda maior (66,69% de probabilidade média de evolução para maiores inadimplências). Na faixa de atraso de 61 a 90 dias a probabilidade de evolução tem pouco aumento, com uma média de 79,94%. Já nas faixas seguintes, as probabilidades aumentam ainda substancialmente na medida em que os atrasos migram para faixas maiores de atraso (média de 78,8% de evolução de inadimplência dos contratos com 91 a 120 dias para 121 a 150, e 86,9% de probabilidade de migração dos contratos situados entre 121 e 150 para 151 a 180 dias de atraso). Por fim, 88,2% é a probabilidade média dos contratos com 151 a 180 dias de atraso migrarem para atrasos com mais de 180 dias.

É bastante nítido o comportamento dos clientes da carteira de crédito em estudo: clientes que não apresentam atrasos têm baixa probabilidade de incorrerem em atraso na prestação seguinte, podendo ser considerados como detentores de contratos “bons”. Clientes com atraso de até 30 dias também podem ser classificados como bons, pois apresentam alta probabilidade de se manterem na mesma faixa de atraso ou até mesmo de regredirem para atrasos inferiores. Contudo, o mesmo não ocorre com contratos que apresentam atrasos entre 31 e 60 dias, tendo em vista a alta probabilidade de 66,7% de evolução para atrasos ainda maiores e de 33,3% de regressão ou manutenção da mesma faixa de atraso. Extrai-se, portanto, que o ponto crítico de evolução de inadimplência da carteira de crédito ocorre quando os contratos atingem 31 dias ou mais de inadimplência. A partir desse ponto as probabilidades de evolução de inadimplência são substancialmente maiores quanto maior for o atraso atingido.

Diante do exposto é aceitável considerar atrasos de até 30 dias como delimitador de contratos “bons” (dadas as baixas probabilidades de estas operações atingirem a inadimplência absoluta), enquanto que atrasos de 31 ou mais dias definem as operações de crédito ruins, dada a súbita elevação da probabilidade dessas operações evoluírem para a inadimplência.

Portanto, foi adotada a seguinte classificação para os contratos da carteira de crédito em estudo:

- Contratos com até 30 dias de atraso serão classificados como contratos “bons”, criando-se a variável qualidade de crédito, para a qual foi atribuído valor 1 para as boas operações;

- Contratos com 31 dias ou mais de atraso serão classificados como contratos “maus”, atribuindo-se o valor 0 para as operações ruins.

3.2 Categorização das variáveis

Definida a qualidade de crédito como variável resposta do modelo da regressão logística (0 para créditos ruins e 1 para créditos bons), o passo seguinte é o de encontrar quais as variáveis explicativas (cadastrais, comportamentais) mais relevantes para explicar a qualidade do crédito. Para tal, no desenvolvimento de modelos de *credit scoring* é usual a categorização de todas as variáveis num número não muito grande de classes, com vistas a reduzir a influência de valores discrepantes, que algumas vezes podem ser resultado de erro na obtenção do valor da variável.

O objetivo é o de conhecer os agrupamentos possíveis em cada variável explicativa que têm comportamentos homogêneos em relação à qualidade de crédito. Desta forma, se duas categorias de uma variável apresentam risco de crédito equivalente, é razoável agrupá-las numa única classe. Assim, a cada grupo (categoria, classe) de comportamento semelhante é atribuído um valor que vai de 1 a n (n = número de categorias resultantes), sendo criada uma nova variável (variável categorizada) que é utilizada como variável explicativa. Para tal, a técnica escolhida foi a denominada CHAID (*Chi-Squared Interaction Detection*), que consiste numa estatística χ^2 (qui-quadrado) para detectar comportamento de homogeneidade entre variáveis.

Por exemplo: a variável IDADE tem como resposta a idade do indivíduo no momento em que foi concedida a operação de crédito. Entretanto, o teste estatístico CHAID demonstrou que indivíduos com diferentes faixas etárias possuíam comportamentos diferentes com relação à qualidade de crédito. Assim, a variável IDADE ficou com 7 categorias: 1 – até

25 anos, 2 - de 26 a 36 anos, 3 – de 37 a 40 anos, 4 – de 41 a 48 anos, 5 - de 49 a 53 anos, 6 - de 54 a 60 anos e 7 – 61 ou mais anos. Desta forma, passou-se a considerar a categoria de IDADE, e não mais a idade original.

A principal vantagem na utilização de variáveis categorizadas ao invés das variáveis originais consiste na simplicidade de interpretação e maior poder de previsão do modelo resultante de categorias homogêneas (PEREIRA, 2004). Outra vantagem em se utilizar o CHAID se baseia na categorização dos *outliers* ou pontos de influência, nome dado aos valores discrepantes que podem ocasionar erro na estimação dos parâmetros das variáveis e por isso requerem medidas corretivas para seu tratamento. Assim, por categorizar os *outliers*, o CHAID descarta a necessidade de eliminar da base de dados valores discrepantes.

3.2.1 CHAID

O CHAID é uma estatística para relacionar uma variável dependente categorizada a uma ou mais variáveis preditoras também categorizadas. O propósito é dividir um conjunto de objetos de tal forma que os subgrupos sejam diferentes com relação a determinado critério (qualidade do crédito). As categorias derivadas do CHAID são mutuamente exclusivas e exaustivas, ou seja, cada resposta está contida numa única e exclusiva categoria, além de obrigatoriamente constar em uma das categorias. Por exemplo: se na variável UF (Unidade da Federação de nascimento do detentor do crédito) a resposta for DF, e este estiver na primeira categoria resultante do CHAID, essa mesma resposta não poderá estar contida em nenhuma outra categoria.

O CHAID se baseia na análise dos momentos das variáveis explicativas e da variável resposta, utilizando para tal o teste χ^2, o qual acumula os desvios quadrados padronizados entre as freqüências observadas e esperadas, sendo dado pela seguinte fórmula:

$$\chi^2 = \sum_i \left[\frac{(O_i - E_i)^2}{E_i} \right] \qquad (1)$$

Onde, *i* é a i-ésima variável independente condicional à variável dependente, O_i é a freqüência observada da i-ésima variável e E_i é a freqüência esperada da i-ésima variável.

As hipóteses nula H_0 e alternativa H_1 do teste χ^2 são:

H_0: X e Y são independentes

H_1: X e Y são dependentes

O processamento do algoritmo acima segue as etapas propostas por Lopes (2003) para procurar a melhor tabela de contingência:

1º passo: Para cada variável independente *X* é construída uma tabela de dupla entrada de suas categorias, com as categorias da variável dependente *Y*. A seguir é procurado o par de categorias de *X* menos significante, ou seja, aquela que, calculada a estatística χ^2, apresenta o maior valor de p (nível descritivo do teste de associação). [25]

2º passo: O par de categorias de *X* (variável independente) com o maior valor de p é comparado ao valor crítico α pré-especificado (nesse caso α = 0,05). Se o valor de *p* for maior que 0,05, este par é homogêneo, sendo agrupado em uma única categoria em relação à variável dependente. Um novo bloco de categorias de *X* é formado e o procedimento é repetido para todas as categorias existentes. Se, no entanto, o valor de *p* for menor que α crítico, a variável independente é selecionada e o conjunto de dados é subdividido de acordo com as categorias finais de *X*. [26]

3º passo: Para cada segmento de dados resultantes da etapa anterior, o programa retorna à primeira etapa. [27]

O processo de geração de novos nós termina na ocorrência de uma das seguintes situações:

- Após a análise de todas as variáveis independentes;
- Na ausência de significância estatística nas associações;
- Quando o número de observações for pequeno demais por subgrupo.

[25] O método estatístico utilizado para calcular o valor de p varia de acordo com o tipo da variável dependente (Y). Quando a variável dependente é qualitativa, as inter-relações são estudadas por meio de estatísticas qui-quadrado; quando a variável dependente é quantitativa, as inter-relações são estudadas por meio da razão de verossimilhança

[26] As categorias que apresentam valores pequenos são agrupadas com aquelas de maior semelhança.

[27] Cada subdivisão gerada recebe o nome de nó, ramo, ramificação ou, simplesmente, partição.

O programa apresenta a análise dos dados em forma de árvore. Os últimos ramos da árvore são chamados terminais e definem um subgrupo de indivíduos classificados em um dos níveis da variável dependente: o nível (ou categoria) da variável dependente *Y* que apresentar a maior proporção de indivíduos define sua classificação.

A utilização da metodologia CHAID ocorre, na prática, quando verificados os seguintes componentes:

- Presença de respostas distintas e mutuamente exclusivas da variável dependente, a exemplo da qualidade de crédito;

- Existência de respostas categorizadas ou não para as variáveis independentes. Algumas variáveis são naturalmente categorizadas, como UF, grau de instrução, estado civil, etc. enquanto outras são contínuas, como renda, valor automóvel, valor cartão de crédito, saldo médio conta corrente, etc.

Importante notar que é possível que o CHAID gere um resultado que possui apenas uma categoria contendo todas as possíveis respostas da variável explicativa. Isto ocorre quando o teste atinge uma etapa contendo apenas duas categorias e o *p-value* do teste entre elas é maior que 5%. Se as duas categorias podem ser agrupadas em uma única categoria homogênea, significa que a variável não apresenta correlação com a variável resposta, já que todas as respostas possíveis são consideradas homogêneas em relação à variável resposta e, portanto, não pode ser utilizada para classificação das operações de crédito.

Os parâmetros utilizados no CHAID foram: *i*) nível de significância de 0,05 ($\alpha_{crítico}$); *ii)* 5 para o nível máximo de "níveis" gerados; *iii)* 1.000 para o tamanho mínimo de casos do nó parental e 500 para o tamanho mínimo de casos dos nós derivados.

Finalmente, vale salientar que o CHAID requer uma amostra grande de observações para a obtenção de resultados confiáveis, o que não comprometeu o presente estudo, tendo em vista a existência de 171.461 observações das operações de crédito para análise. Além disso, deve-se considerar que o CHAID foi usado como uma ferramenta intermediária na geração do

modelo final de *credit scoring*, ou seja, não é o resultado final do modelo de concessão de crédito.

3.2.2 Resultados do CHAID

Após a categorização das 60 variáveis em relação à variável dependente "qualidade do crédito" pelo método CHAID, foi gerada a tabela 9, a qual dispõe da razão dos clientes bons em relação aos clientes maus. Chamaremos esta razão de risco relativo, que significa a quantidade de contratos bons para cada contrato mau naquela categoria. Intuitivamente, podemos interpretar o risco relativo como a probabilidade de existência de contratos bons numa determinada amostra de contratos. Por questões de sigilo, demonstraremos a categorização somente das variáveis autorizadas pela instituição que cedeu os dados.

TABELA 9 - Variáveis Explicativas categorizadas pelo método CHAID

Variável Explicativa	Resposta Original	Resposta Categorizada	Créditos bons / Créditos maus (risco relativo)
Possui Plano de Saúde	possui	1	3,88
	missing	2	2,22
	não possui	3	3,46
Situação Veículo	quitado	1	4,00
	missing	2	2,18
	financiado	3	2,34
Tipo de Imóvel	missing/não possui, fazenda	1	2,28
	casa, lote ou loja	2	3,05
	apartamento, terreno, sala comercial	3	3,72
	chácara	4	4,46
Ano do Veículo	até 1993	1	2,95
	de 1994 a 1997	2	3,20
	de 1997 a 2000	3	3,67
	de 2001 a 2002	4	4,29
	2003	5	4,65
	2004 ou mais	6	3,83
Grau de Instrução	ensino fundamental incompleto	1	0,90
	ensino fundamental completo	2	0,84
	ensino médio incompleto	3	0,65
	ensino médio completo e superior incompleto	4	0,84
	superior completo, especialização e mestrado	5	1,70

Variável Explicativa	Resposta Original	Resposta Categorizada	Créditos bons / Créditos maus (risco relativo)
	doutorado	6	5,17
Idade	até 25	1	1,83
	26 a 36	2	2,22
	37 a 40	3	2,38
	41 a 48	4	2,64
	49 a 53	5	3,18
	54 a 60	6	3,90
	61 ou mais	7	5,02
Sexo	feminino ou missing	1	2,77
	masculino	2	2,58
Estado Civil	Solteiro(a)	1	2,00
	divorciado(a), outros	2	2,46
	Separado(a) judicialmente	3	2,73
	casado(a)	4	3,22
	viúvo(a)	5	3,83
UF	PA, RO, MA, AM, RR, AC, AP	1	1,27
	RN, SE, BA, AL, CE, PI	2	1,71
	MT, GO, TO	3	1,87
	PE, MS	4	2,10
	RJ, DF, PB	5	2,31
	RS, missing	6	2,60
	ES	7	2,89
	PR, MG	8	3,13
	SP	9	3,35
	SC	10	3,57
Renda Líquida Formal	até R$581,00	1	1,82
	de R$582,00 a R$815,00	2	2,21
	de R$816,00 a R$999,00	3	2,55
	de R$1.000,00 a R$1.159,00	4	2,34
	de R$1.160,00 a R$1.335,00	5	2,56
	de R$1.336,00 a R$1.999,00	6	3,12
	de R$2.000,00 a R$2.700,00	7	2,70
	de R$2.701,00 a R$4.179,00	8	3,13
	acima de R$4.180,00	9	4,38
Renda Informal	até R$550,00	1	2,33
	R$551,00 a R$800,00	2	2,14
	R$801,00 a R$1.000,00	3	2,37
	R$1.001,00 a R$1.200,00	4	2,06
	R$1.201,00 a R$1.499,00	5	1,76
	R$1.500,00 a R$1.849,00	6	2,50
	R$1.850,00 a R$3.998,00	7	2,85
	R$3.999,00 ou mais	8	3,95
Saldo de Aplicações em Outras	até R$500,01	1	3,50
	de R$501,00 a R1.200,01	2	4,52

Variável Explicativa	Resposta Original	Resposta Categorizada	Créditos bons / Créditos maus (risco relativo)
Instituições	de R$1.201,00 a R$5.760,01	3	6,41
	de R$5.761,00 a R$13.000,01	4	7,85
	acima de R$13.001,01	5	12,33
Valor Médio Fatura Cartão de Crédito	até R$50,00	1	2,60
	R$51,00 a R$149,00	2	2,82
	R$150,00 a R$499,00	3	3,08
	R$500,00 a R$1.100,00	4	3,57
	acima de R$1.101,00	5	4,18
Qtde Anos Associado Cartão	até 1 ano associado cartão de crédito	1	2,04
	2 anos associado cartão credito	2	2,70
	3 anos associado cartão credito	3	3,24
	4 anos associado cartão credito	4	3,90
	4 a 6 anos associado cartão credito	5	4,46
	6 a 8 anos associado cartão credito	6	5,29
	8 ou mais anos associado cartão de crédito	7	7,20
Valor dos Cheques Devolvidos Motivo 11	até R$557,00	1	1,79
	de R$558,00 a R$2.930,00	2	1,22
	acima de R$2.931,00	3	0,97
Qtde Dias Abertura Conta	até 5	1	2,09
	6 a 28	2	2,33
	29 a 130	3	2,50
	131 a 229	4	2,64
	230 a 438	5	3,07
	439 a 1642	6	3,69
	acima de 1643	7	6,63

É importante observar se os resultados da categorização geraram respostas condizentes. Analisando a variável “idade” percebe-se a segregação em sete categorias com comportamento heterogêneo. Em cada uma das categorias a razão de créditos bons em relação à de créditos maus apresentou proporções diferentes entre si e um comportamento crescente: quanto maior a idade, maior é a quantidade de créditos bons (maior o risco relativo) para cada crédito mau. A intuição para este resultado baseia-se no fato de que pessoas mais velhas tendem a possuir situação financeira estabilizada, normalmente oriunda de aposentadorias, pensões e outras fontes de renda. Por exemplo, a categoria 1 (pessoas com idade até 25 anos) apresenta 1,83 de risco relativo, ou seja, a cada 14 contratos, 9 são bons e 5 são maus. Já os

tomadores com idade acima de 61 anos, situadas na categoria 7, possuem 5,02 créditos bons para cada crédito mau como risco relativo, o que significa que temos 5 vezes mais de chances de, aleatoriamente, selecionarmos um contrato bom em relação a um contrato mau nesta categoria.

A variável “qtde dias abertura conta” também apresenta resultado bastante intuitivo: à medida que a instituição financeira conhece o comportamento de clientes antigos, a mesma pode negar crédito aos que se mostraram maus pagadores. Caso o tomador não possua relacionamento, seu comportamento creditício é desconhecido, o que enseja maior risco de inadimplência ao se conceder crédito a esses tomadores. Por este motivo, o risco relativo possui uma relação direta com a variável “qtde dias abertura conta”, ou seja, quanto mais antiga for a conta corrente (relacionamento/comportamento junto à instituição financeira), maior será o risco relativo de se encontrar bons pagadores.

A variável “UF” também fornece uma análise interessante: todos os estados da região norte, com exceção de TO, foram agrupados na mesma categoria, o que demonstra um comportamento homogêneo dos tomadores residentes naqueles estados. Da mesma forma, grande parte dos estados da região nordeste foi agrupada na mesma categoria pelo CHAID, com exceção de PE e PB. Analisando o risco relativo dos grupos, observa-se que os estados da região norte e nordeste apresentam o menor risco relativo (1,27 e 1,71 respectivamente). Este resultado pode ser corroborado por meio de uma análise socioeconômica, já que historicamente essas regiões são as que apresentam menores índices de desenvolvimento em comparação com as demais macro-regiões brasileiras.

Os resultados da categorização CHAID também demonstraram a necessidade de se eliminar 9 das 60 potenciais variáveis do estudo por apresentarem somente 1 categoria, não havendo, portanto, heterogeneidade nas respostas daquelas categorias. Desta forma, as variáveis listadas abaixo poderão ser excluídas do estudo, uma vez que qualquer que seja sua resposta, o efeito na variável dependente na predição da qualidade de crédito será o mesmo, restando 51 variáveis:

TABELA 10 - Variáveis excluídas por apresentarem 1 nó

Qtde Antenas Parabólicas	Qtde Telefone Comum
Qtde Freezer	Qtde Computador
Qtde Máquina Lavar Louça	Variável Tipo “e1” (apontamentos negativos)
Qtde de Microondas	Valor Médio dos Excessos de Limite
Nacionalidade	

As explanações da seção 2.1, sobre os problemas relacionados à correlação de variáveis, sublinham a importância de se analisar o grau de correlação entre as variáveis. Assim, é recomendável analisar todas as categorias das 51 potenciais variáveis do estudo, excluindo as que apresentarem alto grau de correlação entre si. Foram consideras variáveis altamente correlacionadas as que apresentaram mais do que 80% de correlação entre si. Desta forma, constatou-se a necessidade de exclusão de outras 6 variáveis, restando 45 potenciais variáveis preditoras à qualidade de crédito:

TABELA 31 - variáveis excluídas por apresentarem alta correlação

variável tipo “d1” (financeira – informações bancárias)	variável tipo “d3” (financeira – informações bancárias)
Bandeira do Cartão de Crédito	Qtde Participações Empresariais
variável tipo “a6” (patrimonial)	Qtde de Cartões de Crédito

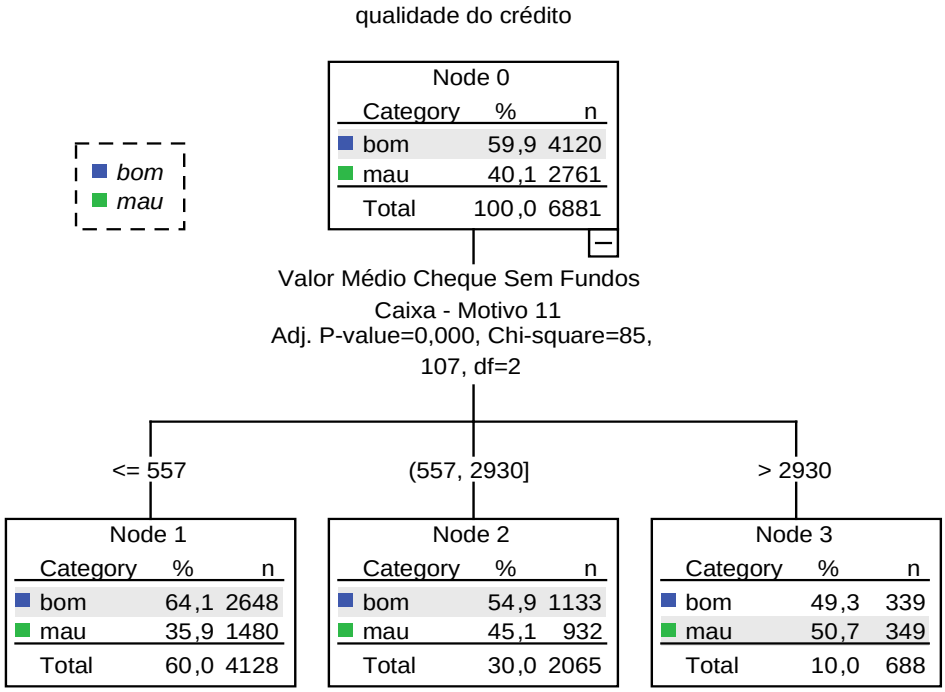
3.2.3 Árvores de classificação

Além dos riscos relativos obtidos na tabela 9, é interessante construir árvores de classificação como uma ferramenta de análise descritiva que relacionam a qualidade do crédito às variáveis explicativas categorizadas. O objetivo é fornecer uma abordagem mais apurada das aplicações do método CHAID quanto às características das operações de crédito analisadas. Árvores de classificação são obtidas como resultado da técnica CHAID executada no SPSS, cujas categorizações são obtidas como forma de nós *(node)*, que nada mais são do que as categorias geradas por esta técnica.

Conforme informado na seção anterior, restaram 45 potenciais variáveis explicativas, uma vez que foram excluídas 9 variáveis que apresentaram somente uma categoria, e outras 6 variáveis por apresentarem alta correlação. A figura 1 abaixo apresenta

como exemplo a árvore de classificação para “Valor Médio dos Cheques Devolvidos Motivo 11”, enquanto as demais variáveis não codificadas encontram-se no anexo B.

FIGURA 1 - Qualidade de crédito por Valor Médio dos Cheques sem Fundos Motivo 11



De acordo com a figura 1, o percentual de créditos ruins dentro de cada categoria mostrou-se homogêneo, interna e estatisticamente, das demais categorias, já que a primeira apresentou 64,1% de operações boas, enquanto as categorias 2 e 3 apresentaram 54,9% e 49,3%, respectivamente. Pode-se dizer que a categoria 3 (clientes que possuem cheques devolvidos por motivo 11 acima de R$2.930,00) possui maior concentração de créditos ruins, já que das 688 operações deste nó, 50,7% foram ruins, percentual superior à média da carteira que é de 40,1%. Clientes que possuem cheques devolvidos até R$557,00 (categoria 1) são os que possuem menor concentração de créditos ruins, pois dentre as 4.128 operações neste nó, 35,9% são de créditos ruins, índice abaixo da média da carteira (40,1%). Já os clientes que possuem cheques devolvidos pelo motivo 11 situados entre R$557,00 e R$2.930,00 apresentaram 45,1% de créditos ruins, índice inferior ao da categoria 3 e ligeiramente superior à média da carteira.

Vale destacar que variáveis quantitativas foram categorizadas considerando somente respostas válidas, ou seja, os *missing values* foram ignorados. Somente após a categorização

foram atribuídos o valor 0 aos clientes com ausência de resposta. Assim, aqueles que não possuem cheques devolvidos pelo motivo 11, portanto, não contemplados nas categorias 1, 2 ou 3, tiveram 0 como resposta atribuída a esta variável.

Contudo, é importante salientar que mesmo as árvores de decisão indiquem que a variável é boa discriminante para a qualidade de crédito, é possível que a mesma seja não significante na estimação do modelo final de geração do *score*, devido à interação com outras variáveis que participarão da equação e que tenham poder de discriminação semelhante ao da variável em questão. Assim, somente podemos afirmar que determinada variável é discriminante entre bons e maus quando sua presença nas equações finais de regressão logística e análise de sobrevivência for estatisticamente significante.

3.3 Amostra utilizada na modelagem

Do universo de 171.461 contratos disponibilizados, nem todos foram utilizados na modelagem, tendo em vista ser interessante guardar observações não utilizadas para se avaliar a acurácia dos modelos desenvolvidos. Assim, dentre o universo de contratos, observou-se que 27% são de operações ruins, ou seja, 46.969 casos. Destes, 15% (7.045 casos) foram selecionados aleatoriamente e guardados para testes, restando aproximadamente 85% (39.614 casos) para montagem das equações de regressão logística e análise de sobrevivência. Também foi selecionado aleatoriamente um número aproximado de contratos bons, de forma que a amostra final para modelagem dispõe de 79.287 observações com todas as 45 variáveis explicativas categorizadas.

A tabela 12 resume a seleção de casos para composição da amostra:

TABELA 42 - Tamanho da amostra para modelagem

	% de contratos no universo de contratos	Qtde de contratos	Seleção da amostra para modelagem	qtde de contratos para modelagem	% da amostra para modelagem
Operações ruins	27%	46.969	85% dos contratos ruins	39.614	50%
Operações boas	73%	124.492	qtde aproximada de contratos bons	39.673	50%
Total	100%	171.461	.	79.287	100%

É recomendável a adoção do mesmo número de contratos bons e maus para que os coeficientes a serem estimados na regressão logística não apresentem viés em relação a nenhum dos grupos, gerando intervalos de confiança proporcionais nos resultados. A tabela 13, obtida por meio da execução da regressão logística que conta com toda base de dados do presente estudo (46.969 contratos maus e 124.492 bons), retrata bem este viés:

TABELA 53 - Tabela de classificação para base de dados com maior quantidade de contratos bons

Observado		Predito		
		Qualidade do crédito		Percentual Correto
		mau	bom	
Qualidade do crédito	mau	4.377	42.592	9,3
	bom	3.117	121.375	97,5
Percentual total				73,3

Os resultados demonstram que dos 121.375 contratos bons, 97,5% foram classificados corretamente, enquanto somente 4.377 (9,3%) dos contratos maus obtiveram o mesmo êxito. Fica evidenciada a existência de viés na regressão logística, no sentido de classificar a maioria dos contratos como bons. De uma forma geral, os estudos sobre análise de sobrevivência não demonstram a necessidade de bases de modelagem com o mesmo número de contratos bons e maus. Entretanto, com o intuito de se manter uma base única para modelagem, usaremos a mesma base de dados na análise de sobrevivência e regressão logística.

3.4 Regressão Logística

A análise Logit ou regressão logística é uma técnica estatística utilizada na separação de dois grupos, que visa obter a probabilidade de que uma observação pertença a um conjunto determinado, em função do comportamento das variáveis independentes. Na aplicação ao risco de crédito, esta técnica é utilizada para avaliação da inadimplência de determinado grupo de clientes em relação à concessão de crédito, assumindo que a probabilidade de inadimplência é logisticamente distribuída, com resultado binomial 0 ou 1. A equação geral da regressão logística pode ser escrita como:

$$\ln Y = \beta_0 + \beta_1 + \beta_2 + ... + \beta_j X_j \qquad (2)$$

ou $$Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j} \quad (3)$$

A esperança condicional E (Y/X) é então representada pela seguinte função:

$$E[Y/X] = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j}} = \pi(X) \quad (4)$$

onde π(X), por comodidade, representa *E[Y/X]; X* representa o conjunto de variáveis explicativas $X_1, X_2, \ldots, X_j$; β representa os parâmetros a serem estimados $\beta_0, \beta_1, \ldots, \beta_j$. Esta função assume valores ajustados no intervalo [0,1], propriedade muito importante no estudo de dados binários, e é não linear em seus parâmetros.

O valor da variável dependente Y é então dado por:

$$Y = \pi(X) + \varepsilon \quad (5)$$

onde o termo ε refere-se ao erro aleatório, representado pela diferença entre o valor observado de Y e a esperança condicional de Y dado X.

Os parâmetros desconhecidos do modelo são geralmente estimados pelo método da máxima verossimilhança, maximizando-se a função de log-verossimilhança L(β) abaixo:

$$L(\beta) = \sum_{i=1}^{n} \{Y_i \cdot \ln[\pi(X_i)] + (1 - Y_i) \cdot \ln[1 - \pi(X_i)]\} \quad (6)$$

Os valores estimados $\beta_0, \beta_2, \ldots, \beta_j$ são os que maximizam *L(β)* e são obtidos derivando-se *L(β)* em relação a cada um dos parâmetros e igualando as expressões resultantes, denominadas equações de verossimilhança, iguais a zero:

$$\sum_{i=1}^{n} [Y_i - \pi(X_i)] = 0$$

e

$$\sum_{i=1}^{n} X_{ij} [Y_i - \pi(X_i)] = 0 \quad (7)$$

onde j = 1, 2, ..., p

Na segunda equação de verossimilhança estão representadas p variáveis explicativas contidas em X^i.

Considerando a utilização da técnica CHAID para categorizar as variáveis explicativas, estas estão representadas por variáveis *dummy* (variáveis que assumem valores 0 ou 1). Assim, se existem 3 variáveis explicativas, cada uma com 3 categorias, teremos 6 dummies criadas (2 para cada variável). A tabela abaixo retrata um exemplo hipotético para o caso de 3 variáveis exemplificadas:

TABELA 14 - Exemplo de variáveis dummies categorizadas

Variável Explicativa	Categoria	variáveis dummy				
		D_{11}	D_{12}	D_{21}	D_{31}	D_{32}
Qtde de imóveis	1 - até 1 imóvel	1	0	.	.	.
	2 - 2 imóveis	0	1	.	.	.
	3 - 3 ou mais imóveis (categoria de referência)	0	0	.	.	.
Sexo	1 - feminino	.	.	1	.	.
	3 –masculino	.	.	0	.	.
Valor do veículo	1 - até R$9.950,00	.	.	.	1	0
	2 - de R$9.950,00 a R$12.450,00	.	.	.	0	1
	3 - R$12.451,00 ou mais (categoria de referência)	.	.	.	0	0

A categoria selecionada como referência (quando as demais *dummies* de sua categoria forem iguais a 0) foi aquela que apresentou o maior número de casos em sua categoria. Por exemplo, se a variável “sexo” possui mais representantes da categoria “masculino”, esta será a *dummy* de referência, sendo representada pela combinação das demais *dummies* de sua categoria.

O *score* da proponente ao crédito é calculado multiplicando-se o resultado da expressão (4) (probabilidade deste proponente ser um cliente bom) por 100. Assim, se o resultado da expressão for 0,90, teremos um *score* de 90, ou seja, há 90% de probabilidade deste tomador ter a qualidade de crédito igual a 1 (crédito bom), dadas as características do tomador.

3.4.1 Método para seleção das variáveis explicativas – foward stepwise

A variável dependente na regressão logística do modelo *credit scoring* é a qualidade de crédito da operação, a qual deverá ser classificada por 45 potenciais variáveis explicativas. É possível que algumas destas potenciais variáveis mostrem-se não significantes estatisticamente, ou seja, forneçam informações não relevantes à qualidade de crédito. Assim, faz-se necessária a escolha de um método para seleção das variáveis explicativas mais relevantes à predição da variável resposta. Alguns dos métodos comumente utilizados são o *foward stepwise, backward stepwise* e *enter*, os quais possuem diferenças sutis entre si.

O método *enter* é utilizado quando se conhece previamente quais as variáveis explicativas são relacionadas à qualidade de crédito, tornando-se indispensáveis ao modelo. Desta forma, o modelo é estimado "forçando" a presença destas variáveis na equação final, testando os coeficientes estimados e o poder de classificação do modelo.[28]

O método *stepwise* computa uma seqüência de equações de regressão, adicionando ou deletando uma variável explicativa em cada passo, de acordo com a significância estatística de entrada e saída desta variável[29]. A rotina de regressão *stepwise* permite que uma variável independente, trazida para dentro do modelo em um estágio anterior, seja removida subseqüentemente se ela não ajudar na conjunção com variáveis adicionadas nos últimos estágios. Esta rotina empregada conduz a um teste para rastrear alguma variável independente que seja altamente correlacionada com variáveis independentes já incluídas no modelo. As principais variantes do método *stepwise* são duas: *foward e backward stepwise*, cujas diferenças são apenas pequenas modificações no seu algoritmo básico. Em suma, o método *backward* parte de um modelo inicial com todas as possíveis variáveis explicativas, que vão sendo testadas e eliminadas caso o nível de significância de exclusão seja inferior ao nível de significância da variável em teste. Este procedimento é executado a cada uma das variáveis até se chegar a um modelo final com as variáveis relevantes. Já o método *foward stepwise* se inicia com um modelo sem nenhuma variável explicativa e a cada passo são incluídas

[28] GAZOLA, Sebastião. **Construção de um modelo de regressão para avaliação de imóveis**. Florianópolis: Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2002.
[29] Parâmetros informados previamente ao software estatístico antes de se executar a regressão.

variáveis relevantes, caso o nível de significância de inclusão seja superior à significância da variável em teste, até a obtenção do modelo final[30].

O método escolhido no presente trabalho para as regressões logística e de análise de sobrevivência foi o *foward stepwise*, enquanto os níveis de significância de entrada e saída de variáveis explicativas foram respectivamente de 0,15 e 020, para que se garanta a presença de variáveis importantes e com coeficientes significativamente diferentes de zero[31]. A utilização do método *backward stepwise* também é possível, mas frequentemente os resultados obtidos são idênticos aos da opção *foward*, conforme constatado em Vasconcellos (2002).

3.5 Análise de Sobrevivência e o Modelo de Cox

O modelo de riscos proporcionais de Cox pertence a uma área da estatística denominada análise de sobrevivência (survival analysis). Este modelo, de larga aplicação na área biomédica, se diferencia dos tradicionais modelos *logit, probit* e análise discriminante por fornecer não apenas a probabilidade de que um determinado evento ocorra no futuro, mas também uma estimativa de tempo até sua ocorrência. Assim, a estimação da probabilidade de sobrevivência para diferentes horizontes de tempo permite obter o "perfil de sobrevivência" dos contratos incluídos na amostra.

Entretanto, o modelo Cox possui a limitação de assumir que as covariáveis[32] não se alterem ao longo do tempo no qual o estudo é desenvolvido. Significa que os dados capturados referentes às variáveis preditivas devem manter-se inalterados ao longo dos 24 meses subseqüentes à contratação do crédito rotativo.

Por definição, o tempo decorrido até o evento de interesse é denominado "tempo de falha", e pode representar o tempo até a morte do indivíduo, o tempo até a cura ou, no caso deste trabalho, o tempo decorrido entre a data de contratação e a data inadimplência do crédito contratado. Nos estudos de análise de sobrevivência o tempo de falha constitui, portanto, a variável dependente do modelo.

---

[30] GAZOLA op. Cit.

[31] Recomenda-se usar o nível de significância de entrada de variáveis superior ao de saída, a fim de se evitar a possibilidade de incluir uma variável em certo passo e eliminá-la no passo subseqüente. Para mais detalhes ver CHARNET *et al*. **Análise de modelos de regressão linear com aplicações**. Campinas: Editora da UNICAMP, 1999

[32] Em análise de sobrevivência as variáveis explicativas são denominadas covariáveis.

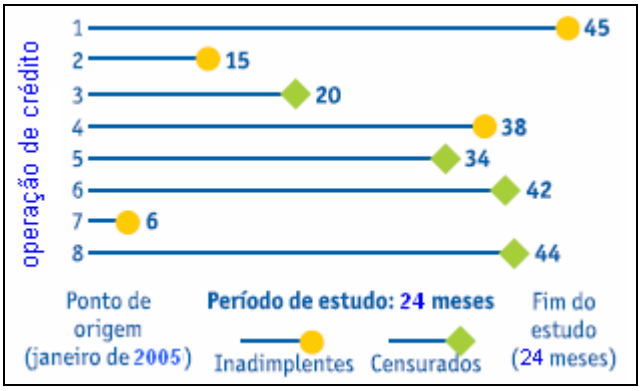
Antes de se passar à descrição do modelo em si, é conveniente abordar alguns conceitos fundamentais sobre análise de sobrevivência, como *dados censurados*, *função de sobrevivência* e *função de risco*.

3.5.1 Dados Censurados e modelagem da base de dados para análise de sobrevivência

A presença da *censura* é uma característica particularmente importante nos estudos da sobrevivência. Compreendida como a observação parcial de resposta, a censura indica que o tempo de falha do evento em interesse é superior ao período de observação do estudo. A presença da censura, entre outras razões, decorre do fato de alguns indivíduos não apresentarem o evento de interesse (falha) até o momento em que se encerra o estudo. Entretanto, tais informações não devem ser descartadas para efeito da análise estatística, pois, mesmo incompletas, fornecem dados importantes sobre o tempo de vida dos indivíduos, além de que, sua omissão acarretaria resultados viesados[33].

Para auxiliar a compreensão de dados censurados, analisemos a figura 2, a qual mostra o monitoramento de contratos por um período de 24 meses após a data de contratação.

FIGURA 2 - Conceito de Censura de Dados



Considerando 24 meses como o período de observação deste trabalho, constata-se que o ponto de origem inicia-se em janeiro de 2005 e o término em julho de 2007 (no caso das

[33] STEPANOVA, M; BAESENS B.; Van GESTEL, T.; Den POEL and VANTHIENEN, D. Neural network survival analysis for personal loan data. **Journal of Operational Research Society**, 2005;

operações contratadas em junho de 2005). Após este ponto não temos como saber se determinado contrato tornou-se inadimplente, devendo ser classificado como censurado. A censura também pode ocorrer antes do fim do período de estudo. A operação 3, por exemplo, pode ter sido encerrada em decorrência do falecimento do titular do contrato, enquanto a conta 5 pode ter sido liquidada antecipadamente a pedido do cliente em decorrência de insatisfação junto à instituição financeira. Neste quadro, se um contrato não se tornou inadimplente seu valor é considerado censurado. Assim, no exemplo da figura 2, todos os empréstimos são considerados censurados, com exceção das operações 2 e 7, ambas inadimplentes, com mais de 30 dias de atraso.

A regressão do modelo Cox requer uma modelagem dos dados um pouco diferente da regressão do *credit score*. Além da variável dependente (tempo até a inadimplência) é necessária a existência de uma variável que discrimine se os dados são censurados ou não. Denominaremos esta variável como "censor", a qual indicará se a observação é censurada ou não, de uma forma binária: as variáveis censuradas recebem o valor de 1, e as não censuradas recebem o valor de 0.

Tendo em mente a figura 2, os dados lá dispostos serão transformados em informações adequadas para a regressão de análise de sobrevivência, de acordo com a tabela 15:

TABELA 65 - Dados para modelagem de análise de sobrevivência

Observação	Variável dependente "tempo de sobrevivência"	Variável de censura (1 = censurado e 0 = não censurado	Variáveis independentes		
			a1	a2	Qtde veículos
1	24	1	2	5	1
2	15	0	3	3	1
3	20	1	5	5	1
4	24	1	7	7	2
5	24	1	1	8	1
6	24	1	2	8	1
7	6	0	2	2	1
8	24	1	1	4	2

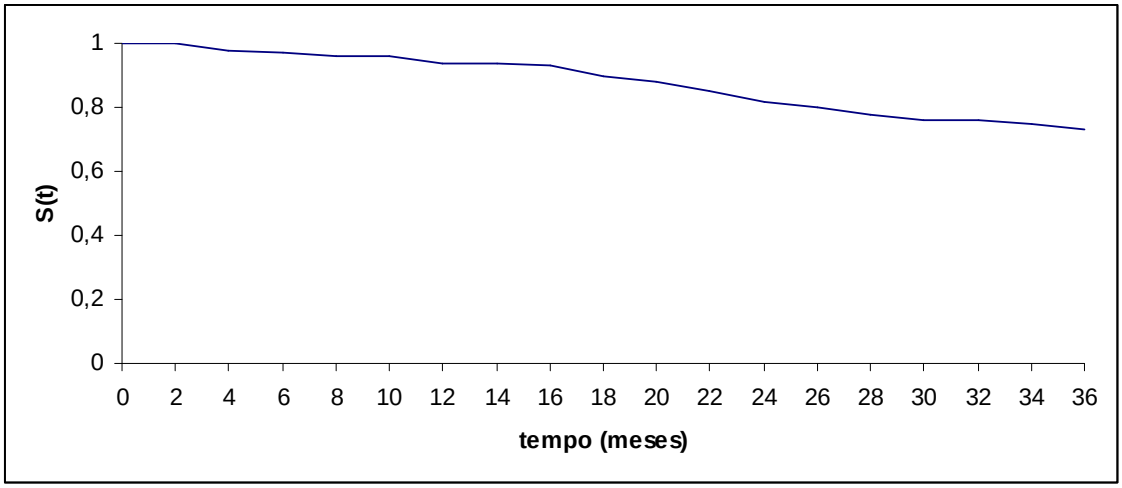
Interessante observar que não há períodos de sobrevivência superiores a 24 meses para a variável dependente "tempo de sobrevivência", pois este é o fim do período do estudo – o limite da censura. As variáveis independentes de perfil e comportamentais estão dispostas

com as respostas categorizadas pela técnica CHAID, enquanto a variável binária de censura indica se as observações são censuradas ou não.

3.5.2 Função de Sobrevivência

Uma das funções mais utilizadas é a função de sobrevivência, cujo objetivo é modelar a probabilidade de uma observação sobreviver além do período *t*. Em termos probabilísticos, isto pode ser descrito como $S(t) = P\ (T \geq t)$. O gráfico abaixo mostra que a função de sobrevivência possui um formato monótono decrescente, para um conjunto de dados fictícios:

GRÁFICO 1 - Exemplo de Função de Sobrevivência



No tempo inicial (quando $t = 0$), a probabilidade de sobrevivência da observação é 1, ou seja $S(0) = 1$. À medida que o tempo passa, decresce a probabilidade de sobrevivência, sendo que a probabilidade de que esta sobreviva no tempo infinito ($t = \infty$) é 0, ou seja, $S(\infty) = 0$.

O Kaplan-Meier (KM) é o estimador de máxima verossimilhança não-paramétrico utilizado na função de sobrevivência. Este estimador incorpora informações de todas as observações disponíveis na base de dados, censuradas ou não, considerando a sobrevivência destas a qualquer ponto do tempo como uma série de passos definidos pelos dados censurados e os que sobreviveram. Em seguida são estimadas as probabilidades condicionais de sobrevivência das observações no tempo seguinte, multiplicando estas probabilidades para

obter uma estimativa de toda a função de sobrevivência. O estimador de máxima verossimilhança KM, então, torna-se:

$$\hat{S}(t) = \prod_{j|t_{(j)} \leqslant t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j|t_{(j)} \leqslant t} \left(1 - \frac{d_j}{n_j} \right) = \hat{S}(t-1)\left(1 - \frac{d_t}{n_t} \right) \quad (8)$$

Onde,

d_j é o número de falhas (inadimplências) no tempo $t_{(j)}$

n_j é o número total de observações em risco no tempo $t_{(j)}$

3.5.3 Função de Risco (ou taxa de falha)

Outra função utilizada nos estudos de análise de sobrevivência é a Função de Risco, cujo objetivo é mensurar a probabilidade de um evento ocorrer no intervalo de tempo entre *t* e *t* + Δ*t*, dado que o evento sobreviveu até o tempo *t*. Para obter a função de risco (ou função de taxa de falha) é preciso expressar a probabilidade de falha ocorrer em um intervalo de tempo (t_1, t_2) por meio da função de sobrevivência:

$$S(t_1) - S(t_2) \quad (9)$$

Em seguida, deve-se definir a taxa de falha no intervalo (t_1, t_2) dividindo a probabilidade de que a falha ocorra no intervalo (uma vez que não ocorreu antes de t_1) pela duração do mesmo. Desta forma, é possível expressar a taxa de falha no intervalo (t_1, t_2) por:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)} \quad (10)$$

Ao redefinir o intervalo (t_1, t_2) como (t, $t+\Delta t$), a expressão assume a seguinte forma:

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t} \quad (11)$$

Quando Δt tende a zero, h(t) representa a taxa de falha instantânea no tempo t. Desta forma, define-se a função de risco T como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t} \tag{12}$$

O formato da função de risco é não especificado, sendo empiricamente determinado pelos dados a serem utilizados no modelo (aspecto não paramétrico do modelo). Dependente dos dados, a função de risco pode direcionar-se para cima, atingir um pico e até voltar para baixo.

3.5.4 Descrição do Modelo de Risco Proporcionais de Cox

Dado que t representa o *tempo até a falha* e T representa a variável aleatória *tempo de falha*, a função de sobrevivência $S(t)$ é definida como a probabilidade de um contrato sobreviver mais do que t períodos, de acordo com a fórmula:

$$S(t) = Prob\ (T > t) = 1 - F(t) \tag{13}$$

onde,

$F(t)$ é a função de distribuição cumulativa para T

Em outras palavras, a função de sobrevivência gera a probabilidade de que o contrato sobreviva além de determinado intervalo de tempo arbitrado pela falha. Considerando que um contrato qualquer não inadimpliu (falhou) dentro do período t, é possível especificar a probabilidade de que isto ocorra no instante $t+1$ pela função risco:

$$h(t) = \lim_{dt \to 0} \frac{P(t < T < t + dt / T > t)}{dt} = \frac{-S'(t)}{S(t)} \tag{14}$$

onde,

$S'(t) = f(t)$ representa a função densidade de probabilidade de t

Considerando que as funções de sobrevivência, densidade de probabilidade e de risco possibilitam a derivação uma das outras, as estimativas de $h(t)$ permitem obter estimativas de $S(t)$ pela fórmula:

$$S(t) = \exp[-\int_0^t h(u)du \,] \qquad (15)$$

A formulação de hipótese sobre a forma de distribuição da variável aleatória *tempo de falha* (*T*) possibilita a especificação de diferentes tipos de funções de risco. No modelo proposto a função risco no tempo *t* é dada pela fórmula:

$$h(t / X\,,\, B) = h_0(t)\,.\, e^{(X'B)} \qquad (16)$$

onde,

$h_0(t)$ representa a função risco de um contrato que é em função do tempo

X' representa um vetor de variáveis explicativas;

B representa um vetor de coeficientes que descreve como cada variável afeta o risco de falha.

Quando *X'* = 0, $e^{(X'B)}$ = 1. Ao centrar as variáveis explicativas, de forma que um contrato com *X'* = 0 representa um tomador de crédito cujo perfil possui as categorias de referência, pode-se interpretar $h_0(t)$ como a função risco de contrato.

Para efeito de tipificação, o modelo de Cox é considerado semi-paramétrico, pois é composto de uma parte paramétrica (vetor de parâmetros *B*) e outra não-paramétrica, chamada função de risco *baseline* $h_0(t)$. Considerando que $h_0(t)$ é arbitrária e só depende do tempo, a estimação de *B* ou de h(*t*) não exige hipóteses sobre a forma de distribuição da variável aleatória *T*. Como o modelo relaciona o risco pela função exponencial, o valor dos coeficientes exponenciados $e^{(X'B)}$ podem ser interpretados como a mudança na taxa de risco decorrente da mudança de 1 unidade na variável preditora X_1, controlando para outros efeitos no modelo (as demais variáveis permanecem constantes). A estimação dos coeficientes *B* é feita por meio dos princípios da máxima verossimilhança.

Dado que *t* representa o intervalo de tempo sobre o qual se deseja inferir a probabilidade de falha de um contrato e *X* representa um vetor de variáveis explicativas comportamentais e de perfil, é possível estimar a função de sobrevivência pela fórmula:

$$S\,(t / X\,,\, B) = S_0(t)^{\exp(X'B)} \tag{17}$$

$$S_0(t) = \exp[-\int_0^t h_0(u)du\,] = e^{-H_0(t)} \tag{18}$$

onde, $S_0(t)$ representa a função de sobrevivência *baseline*, cujo valor apresentado é o mesmo para todos os contratos em cada horizonte de tempo calculado, uma vez que esta função só depende do tempo. O nome Risco Proporcional decorre do fato de que o risco de qualquer contrato é uma proporção fixa do risco de qualquer outro contrato no decorrer do tempo.

Neste tocante, o cálculo da probabilidade de sobrevivência exige que se especifique um horizonte de tempo para a determinação da probabilidade *baseline*. Em seguida, substituem-se os valores das variáveis explicativas na equação 17 obtendo-se a função de sobrevivência para o contrato naquele horizonte de tempo.

O tempo esperado até a inadimplência pode então ser calculado somando-se as probabilidades de sobrevivência em cada horizonte de tempo:

$$E\,(t) = \int_0^{\infty} S(t)dt \rightarrow \text{para dados contínuos ou} \tag{19}$$

$$\sum_{t=1}^{24} S(t) \rightarrow \text{para dados discretos}$$

4. Resultados

Neste capítulo encontram-se os resultados finais das regressões logística e modelo Cox utilizando-se o método foward stepwise e as variáveis “qualidade de crédito” e “tempo de sobrevivência”, respectivamente, como variáveis dependentes dos modelos. Serão apresentados os seguintes resultados:

- Variáveis explicativas participantes dos modelos finais de regressão logística e Modelo Cox, bem como seus coeficientes;
- Testes de hipóteses e medidas de desempenho utilizados;
- Exemplo prático de geração do *score* e o tempo esperado até a inadimplência de uma operação de crédito;
- Análise gráfica dos resultados, abordando a distribuição dos *scores*, funções de risco e sobrevivência acumuladas e tempo médio de duração das operações;
- Considerações sobre a estabilidade do modelo.

4.1 Modelo Final

Os modelos finais foram estimados por meio do software SPSS, selecionando-se a opção “regressão logística – método *foward stepwise* LR (*likehood ratio* ou razão de verossimilhança)” com a variável de resposta “qualidade de crédito” e com as 45 variáveis explicativas categorizadas a serem submetidas à regressão. Da mesma forma, para o Modelo Cox selecionou-se a opção “survival – cox regression – método LR”, com a variável independente “tempo de sobrevivência”, a variável “censor”, indicando as informações censuradas, e as 45 variáveis explicativas a serem regredidas.

A utilização dos métodos *foward stepwise* na escolha das variáveis explicativas relevantes para determinação da qualidade do crédito da regressão logística resultou num modelo com 25 variáveis, escolhidas entre as 45 potenciais variáveis categorizadas. Cada uma destas foi representada por *k*-1 variáveis *dummies* (*k* significa o número de categorias da variável explicativa, sendo uma das categorias a combinação linear das demais, de acordo com a explanação da seção 4.3) e, portanto, cada uma com *k*-1 coeficientes estimados no modelo final. Já a regressão do Modelo Cox resultou na escolha de 36 variáveis explicativas

relevantes entre as 45 potenciais variáveis. O intercepto (β_0) mostrou-se significativamente relevante somente na regressão logística.

O quadro abaixo assinala os resultados cujas variáveis mostraram-se significantes a um nível de 5%:

TABELA 76 - Variáveis explicativas relevantes e significantes

Natureza	Tipo	Nome Codificado	Regressão Logística	Modelo Cox
PERFIL	Patrimonial: tipo A	possui plano de saúde	x	x
		a1	x	x
		a2	-	x
		a3	-	x
		a4	-	-
		a5	-	-
		a7	-	x
		a8	-	x
		a9	x	x
		a10	-	-
		a11	-	-
		situação do veículo	x	x
		tipo de imóvel	x	x
		a12	x	x
		ano veículo	x	x
		valor de mercado veículo	-	-
		a13	-	-
		a14	-	-
	Cadastral (Demográfica): tipo B	grau de instrução	x	x
		idade	x	x
		sexo	-	x
		estado civil	x	x
		UF	x	x
		b1	-	-
		b2	-	x
		b3	x	x
	Financeira - Renda: tipo C	c1	x	x
		c2	-	x
		renda líquida formal	x	x
		c3	x	x
		c4	x	x
		renda informal	x	x
		c5	-	-

Natureza	Tipo	Nome Codificado	Regressão Logística	Modelo Cox
	Financeira - Informações bancárias: tipo D	saldo médio das aplicações em outras instituições	x	x
		d4	x	x
		valor médio da fatura do cartão de crédito	-	x
		qtde anos associado ao cartão de crédito	x	x
		d5	-	x
		d6	x	x
		d7	-	x
COMPORTAMENTAL	Apontamentos negativos: tipo E	e2	x	x
		e3	x	x
		valor dos cheques devolvidos motivo 11	x	x
		e4	-	x
	Cadastral: tipo F	qtde dias abertura conta	x	x

O conjunto de variáveis com significância estatística presentes nos modelos finais mostraram-se complexos, contendo representantes de todos os tipos de variáveis: patrimoniais, demográficas, financeiras de renda, financeira – informações bancárias, apontamentos negativos e cadastrais de comportamento.

Conforme discutido na seção 2.1, algumas variáveis normalmente utilizadas em estudos de risco de crédito não foram disponibilizadas pela instituição financeira, sendo notadas como maiores ausências a indicadora de “crédito salário na instituição” (tipo D); saldos médios em conta corrente, poupança e demais aplicações (tipo F); valor médio de utilização do cheque especial (tipo F) e apontamentos negativo (tipo E).

Observa-se que 7 variáveis patrimoniais (tipo A), dentre elas o valor do veículo, 1 variável demográfica (tipo B) e 1 variável financeira de renda (tipo C) mostraram-se sem significância estatística em ambas as regressões, sendo, portanto, eliminadas dos modelos finais. Nota-se ainda que todas as variáveis relevantes na regressão logística também foram relevantes no modelo Cox.

As variáveis participaram dos modelos em seus formatos categorizados, ou seja, o mesmo formato apresentado na tabela 9. Por exemplo, a variável “tem plano de saúde” possui 3 categorias e 2 coeficientes estimados. A categoria “possui plano de saúde” obteve o coeficiente β igual a 0,248 e a exponencial de β (*exp* β) igual a 1,281; a categoria “não possui plano de saúde” obteve β = 0,255 e *exp* β = 1,291. A resposta “missing”, por ser a categoria de referência, não teve coeficiente, já que representa a combinação linear das demais categorias. Devido ao fato dos coeficientes das categorias possui e não possui plano de saúde serem estritamente próximos entre si, verifica-se que seus efeitos são semelhantes, não havendo diferença se o indivíduo possui ou não plano de saúde. Já aqueles que não responderam a esta variável (*missing*) terão *score* inferior aos que responderam.

O coeficiente β corresponde ao efeito da variação de uma unidade da variável dependente no logaritmo natural da variável dependente *ln* (*Y*). Por exemplo: o fato de o cliente possuir plano de saúde significa um aumento em *ln (Y)* do mesmo ser um bom pagador em 0,248, já que a variável assume valor de 0 ou 1. Com o objetivo de proporcionar uma interpretação mais intuitiva, a *exp* β fornece a o valor exponencial de β. Para a resposta “possui plano de saúde”, este valor corresponde a 1,281, o que equivale a $e^{0,248}$. Este valor agora expressa a taxa de mudança de *Y* do tomador ser um bom pagador por um fator de 1,281, caso o mesmo possua plano de saúde.

No caso do Modelo Cox, o coeficiente β fornece a variação do logaritmo natural do risco em relação à variação de uma unidade da variável preditora, controlando para os outros efeitos do modelo, ou seja, as demais variáveis preditoras permanecem constantes. Usaremos a *exp* β para um melhor entendimento dos resultados, cujo significado é a variação estimada do risco (hazard) associada à variação de 1 unidade da variável dependente, controlando as demais variáveis do modelo. No caso de variáveis categóricas, *exp* β representa a variação no risco ao se mudar de categoria de referência para outra. Também podemos chamar *exp* β como taxa de risco, uma vez que a mesma significa o risco para dois indivíduos que diferem em 1 unidade na variável preditora em questão. A *exp* β para quem possui plano de saúde é 0,866, o que significa que, *ceteris paribus*, o risco dos que possuem plano de saúde têm um decréscimo de 0,866 no risco de se tornarem inadimplentes. Assim, clientes que possuem plano de saúde têm menor risco de se tornarem inadimplentes do que os que não possuem ou que não responderam esta variável (*missing*). O valor (1/*exp* β) representa o inverso da taxa de risco, ou seja, a taxa estimada de sobrevivência daquele contrato não tornar-se inadimplente.

Desta forma, quanto maior for a taxa de risco de um contrato, menor será sua taxa de sobrevivência, e vice-versa.

A tabela a seguir resume os coeficientes de todas as variáveis obtidas nos modelos finais:

TABELA 17 - Variáveis e coeficientes estimados nos modelos finais

Variável	Categoria	Coeficientes								
		Regressão Logística				Modelo Cox				
		β	erro pad.	Sig.	Exp(β)	*β*	erro pad.	Sig.	Exp(*β*)	1/Exp (*β*)
Constante	-	1,612	0,231	0,000	5,011	-	-	-	-	-
TEM PLANO DE SAÚDE	possui plano de saúde	0,248	0,024	0,000	1,281	-0,144	0,015	0,000	0,866	1,155
	missing	-	-	-	-	-	-	-	-	-
	não possui plano de saúde	0,255	0,074	0,001	1,291	-0,104	0,048	0,029	0,901	1,110
a1	a11	-0,286	0,024	0,000	0,751	0,168	0,015	0,000	1,183	0,845
	a12	-	-	-	-	-	-	-	-	-
	a13	0,377	0,116	0,001	1,458	-0,129	0,093	0,169	0,879	1,137
a2	a21	-	-	-	-	-	-	-	-	-
	a22	-	-	-	-	0,082	0,025	0,001	1,085	0,922
a3	a31	-	-	-	-	-	-	-	-	-
	a32	-	-	-	-	-0,150	0,022	0,000	0,861	1,162
	a33	-	-	-	-	-0,321	0,083	0,000	0,726	1,378
a7	a71	-	-	-	-	-	-	-	-	-
	a72	-	-	-	-	0,027	0,014	0,050	1,028	0,973
	a73	-	-	-	-	0,078	0,024	0,001	1,081	0,925
	a74	-	-	-	-	-0,063	0,036	0,074	0,939	1,065
a8	a81	-	-	-	-					
	a82	-	-	-	-	-0,018	0,025	0,463	0,982	1,019
	a83	-	-	-	-	0,118	0,041	0,004	1,126	0,888
a9	a91	-	-	-	-	-	-	-	-	-
	a92	0,158	0,027	0,000	1,171	-0,144	0,023	0,000	0,866	1,155
	a93	-0,043	0,041	0,296	0,958	-0,041	0,032	0,201	0,960	1,041
	a94	-0,022	0,058	0,702	0,978	-0,058	0,044	0,186	0,943	1,060
SITUAÇÃO DO VEÍCULO	quitado	0,170	0,320	0,596	1,185	-0,079	0,204	0,696	0,924	1,083
	missing/não possui	-	-	-	-	-	-	-	-	-
	financiado	-0,301	0,321	0,348	0,740	0,188	0,204	0,356	1,207	0,828
TIPO DE IMÓVEL	missing/não possui, fazenda	-	-	-	-	-	-	-	-	-
	casa, lote ou loja	-0,018	0,022	0,413	0,982	0,012	0,014	0,390	1,012	0,988

Variável	Categoria	Coeficientes								
		Regressão Logística				Modelo Cox				
		β	erro pad.	Sig.	Exp(β)	*β*	erro pad.	Sig.	Exp(*β*)	1/Exp (*β*)
	apartamento, terreno, sala comercial	0,145	0,043	0,001	1,156	-0,083	0,027	0,002	0,920	1,087
	chácara	-0,009	0,039	0,820	0,991	0,006	0,026	0,813	1,006	0,994
a12	a121	0,217	0,046	0,000	1,243	-0,142	0,028	0,000	0,868	1,153
	a122	-	-	-	-	-	-	-	-	-
	a123	0,160	0,069	0,021	1,173	-0,139	0,041	0,001	0,870	1,149
ANO DO VEÍCULO	missing/não possui	-	-	-	-	-	-	-	-	-
	até 1993	-0,171	0,320	0,592	0,843	0,081	0,203	0,689	1,085	0,922
	de 1994 a 1997	-0,143	0,319	0,655	0,867	0,080	0,203	0,694	1,083	0,923
	de 1997 a 2000	-0,048	0,320	0,882	0,954	0,060	0,203	0,768	1,062	0,942
	de 2001 a 2002	0,108	0,320	0,735	1,114	-0,032	0,203	0,875	0,968	1,033
	2003	0,237	0,322	0,462	1,267	-0,087	0,205	0,671	0,917	1,091
	2004 ou mais	0,159	0,321	0,620	1,173	-0,042	0,204	0,835	0,958	1,043
GRAU DE INSTRUÇÃO	ensino fundamental incompleto	-	-	-	-	-	-	-	-	-
	ensino fundamental completo	-0,012	0,036	0,741	0,988	0,011	0,021	0,609	1,011	0,989
	ensino médio incompleto	-0,157	0,042	0,000	0,855	0,059	0,023	0,011	1,061	0,942
	ensino médio completo e superior incompleto	0,143	0,029	0,000	1,153	-0,113	0,017	0,000	0,893	1,120
	superior completo, especialização e mestrado	0,705	0,035	0,000	2,023	-0,450	0,022	0,000	0,638	1,568
	doutorado	0,808	0,054	0,000	2,243	-0,519	0,035	0,000	0,595	1,681
IDADE	até 25 anos	0,149	0,032	0,000	1,161	-0,094	0,018	0,000	0,910	1,099
	25 a 36 anos	-	-	-	-	-	-	-	-	-
	36 a 40 anos	0,014	0,030	0,651	1,014	-0,015	0,017	0,387	0,985	1,015
	40 a 48 anos	0,080	0,026	0,002	1,083	-0,047	0,015	0,002	0,954	1,048
	48 a 53 anos	0,296	0,033	0,000	1,344	-0,182	0,021	0,000	0,834	1,199
	53 a 60 anos	0,523	0,035	0,000	1,688	-0,327	0,022	0,000	0,721	1,387
	60 ou mais anos	0,774	0,039	0,000	2,169	-0,498	0,025	0,000	0,608	1,646
SEXO	feminino	-	-	-	-	-0,046	0,011	0,000	0,955	1,047
	masculino	-	-	-	-	-	-	-	-	-
ESTADO CIVIL	solteiro (a)	-0,146	0,021	0,000	0,865	0,140	0,013	0,000	1,150	0,869
	viúvo(a), missing	0,152	0,043	0,000	1,164	-0,066	0,028	0,020	0,936	1,068
	casado (a)	-	-	-	-	-	-	-	-	-
	Separado(a) judicialmente	-0,101	0,044	0,021	0,904	0,122	0,027	0,000	1,130	0,885
	divorciado(a), outros	-0,206	0,030	0,000	0,814	0,139	0,018	0,000	1,149	0,870
UF	RJ, DF, PB	-0,333	0,030	0,000	0,717	0,199	0,018	0,000	1,221	0,819
	PR, MG	0,101	0,026	0,000	1,106	-0,066	0,016	0,000	0,936	1,068
	SP	-	-	-	-	-	-	-	-	-

Variável	Categoria	Coeficientes								
		Regressão Logística				Modelo Cox				
		β	erro pad.	Sig.	Exp(β)	*β*	erro pad.	Sig.	Exp(*β*)	1/Exp (*β*)
	RN, SE, BA, AL, CE, PI	-0,464	0,034	0,000	0,629	0,244	0,019	0,000	1,276	0,783
	PE, MS	-0,289	0,051	0,000	0,749	0,162	0,030	0,000	1,176	0,850
	RS, missing	-0,069	0,032	0,030	0,933	0,079	0,019	0,000	1,083	0,924
	ES	-0,021	0,061	0,733	0,980	-0,013	0,037	0,716	0,987	1,013
	MT, GO, TO	-0,382	0,042	0,000	0,682	0,228	0,024	0,000	1,256	0,796
	SC	0,330	0,039	0,000	1,391	-0,234	0,025	0,000	0,791	1,264
	PA, RO, MA, AM RR, AC, AP	-0,732	0,052	0,000	0,481	0,389	0,027	0,000	1,476	0,678
b2	b21	-	-	-	-	-	-	-	-	-
	b22	-	-	-	-	-0,026	0,015	0,077	0,974	1,027
	b23	-	-	-	-	0,014	0,032	0,657	1,014	0,986
	b24	-	-	-	-	0,122	0,054	0,025	1,130	0,885
b3	b31	-	-	-	-	-	-	-	-	-
	b32	-0,100	0,052	0,054	0,905	0,073	0,030	0,016	1,075	0,930
	b33	-0,245	0,050	0,000	0,783	0,155	0,029	0,000	1,168	0,856
	b34	-0,245	0,053	0,000	0,782	0,152	0,031	0,000	1,164	0,859
	b35	-0,149	0,052	0,004	0,861	0,102	0,032	0,001	1,108	0,903
	b36	-0,207	0,058	0,000	0,813	0,101	0,036	0,004	1,107	0,904
c1	c11	-	-	-	-	-	-	-	-	-
	c12	0,343	0,066	0,000	1,410	-0,323	0,048	0,000	0,724	1,382
	c13	0,134	0,052	0,009	1,144	-0,249	0,039	0,000	0,779	1,283
	c14	0,243	0,100	0,015	1,274	-0,312	0,066	0,000	0,732	1,366
	c15	0,072	0,093	0,439	1,075	-0,200	0,060	0,001	0,819	1,221
	c16	0,206	0,070	0,003	1,228	-0,261	0,048	0,000	0,770	1,298
	c17	0,365	0,101	0,000	1,440	-0,402	0,070	0,000	0,669	1,494
c2	c21	-	-	-	-	-	-	-	-	-
	c22	-	-	-	-	0,156	0,036	0,000	1,169	0,855
	c23	-	-	-	-	0,217	0,040	0,000	1,243	0,805
	c24	-	-	-	-	0,180	0,067	0,007	1,197	0,835
	c25	-	-	-	-	0,179	0,070	0,010	1,196	0,836
RENDA LÍQUIDA FORMAL	missing/não possui	0,032	0,044	0,473	1,032	0,095	0,027	0,000	1,100	0,909
	até R$581,00	-0,027	0,037	0,456	0,973	0,026	0,021	0,221	1,026	0,975
	de R$582,00 a R$815,00	-0,015	0,036	0,672	0,985	-0,004	0,021	0,834	0,996	1,004
	de R$816,00 a R$999,00	0,080	0,035	0,023	1,083	-0,053	0,021	0,012	0,949	1,054
	de R$1.000,00 a R$1.159,00	-0,090	0,034	0,008	0,914	0,061	0,020	0,002	1,063	0,940
	de R$1.160,00 a R$1.335,00	-0,086	0,035	0,013	0,918	0,056	0,021	0,007	1,058	0,945
	de R$1.336,00 a R$1.999,00	-	-	-	-	-	-	-	-	-

Variável	Categoria	Coeficientes								
		Regressão Logística				Modelo Cox				
		β	erro pad.	Sig.	Exp(β)	*β*	erro pad.	Sig.	Exp(*β*)	1/Exp (*β*)
	de R$2.000,00 a R$2.700,00	-0,211	0,035	0,000	0,810	0,118	0,021	0,000	1,125	0,889
	de R$2.701,00 a R$4.179,00	-0,237	0,037	0,000	0,789	0,130	0,022	0,000	1,138	0,878
	acima de R$4.180,00	-0,236	0,041	0,000	0,790	0,115	0,026	0,000	1,122	0,891
c3	c31	-	-	-	-	-	-	-	-	-
	c32	-2,280	0,020	0,000	0,102	1,278	0,012	0,000	3,590	0,279
c4	c41	-	-	-	-	-	-	-	-	-
	c42	1,140	0,034	0,000	3,128	-0,550	0,019	0,000	0,577	1,733
	c43	1,199	0,037	0,000	3,316	-0,595	0,021	0,000	0,552	1,813
	c44	1,230	0,034	0,000	3,422	-0,617	0,019	0,000	0,539	1,854
	c45	1,300	0,042	0,000	3,668	-0,671	0,025	0,000	0,511	1,956
	c46	1,353	0,040	0,000	3,869	-0,684	0,024	0,000	0,505	1,981
	c47	1,444	0,031	0,000	4,237	-0,732	0,019	0,000	0,481	2,078
	c48	1,257	0,043	0,000	3,515	-0,605	0,026	0,000	0,546	1,831
RENDA INFORMAL	missing/não possui	-	-	-	-	-	-	-	-	-
	até R$550,00	-0,030	0,039	0,437	0,970	0,003	0,023	0,900	1,003	0,997
	R$551,00 a R$800,00	-0,108	0,051	0,033	0,897	0,044	0,029	0,129	1,045	0,957
	R$801,00 a R$1.000,00	-0,078	0,047	0,096	0,925	0,018	0,028	0,525	1,018	0,982
	R$1.001,00 a R$1.200,00	-0,246	0,062	0,000	0,782	0,140	0,036	0,000	1,150	0,869
	R$1.201,00 a R$1.499,00	-0,423	0,080	0,000	0,655	0,216	0,044	0,000	1,242	0,805
	R$1.500,00 a R$1.849,00	-0,165	0,044	0,000	0,848	0,088	0,027	0,001	1,092	0,916
	R$1.850,00 a R$3.998,00	-0,170	0,040	0,000	0,844	0,051	0,025	0,045	1,052	0,950
	R$3.999,00 ou mais	-0,043	0,057	0,454	0,958	-0,036	0,039	0,347	0,964	1,037
SALDO APLICACOES EM OUTRAS INSTITUIÇÕES	missing/não possui	-	-	-	-	-	-	-	-	-
	até R$500,01	0,074	0,131	0,574	1,076	-0,078	0,082	0,337	0,925	1,082
	de R$501,00 a R1.200,01	0,157	0,142	0,269	1,170	-0,087	0,094	0,357	0,917	1,091
	de R$1.201,00 a R$5.760,01	0,596	0,089	0,000	1,814	-0,363	0,064	0,000	0,696	1,437
	de R$5.761,00 a R$13.000,01	0,661	0,114	0,000	1,937	-0,446	0,083	0,000	0,640	1,561
	acima de R$13.001,01	0,961	0,108	0,000	2,616	-0,684	0,087	0,000	0,505	1,981
d4	d41	-	-	-	-	-	-	-	-	-
	d42	-0,603	0,058	0,000	0,547	0,270	0,039	0,000	1,310	0,763
	d43	-0,481	0,054	0,000	0,618	0,240	0,040	0,000	1,271	0,787
	d44	-0,589	0,058	0,000	0,555	0,268	0,041	0,000	1,307	0,765

Variável	Categoria	Coeficientes								
		Regressão Logística				Modelo Cox				
		β	erro pad.	Sig.	Exp(β)	*β*	erro pad.	Sig.	Exp(*β*)	1/Exp (*β*)
	d45	-0,394	0,067	0,000	0,674	0,163	0,046	0,000	1,177	0,850
	d46	-0,434	0,056	0,000	0,648	0,192	0,041	0,000	1,211	0,826
	d47	-0,408	0,061	0,000	0,665	0,145	0,044	0,001	1,156	0,865
	d48	-0,409	0,065	0,000	0,664	0,126	0,048	0,008	1,135	0,881
	d49	-0,345	0,065	0,000	0,708	0,054	0,049	0,276	1,055	0,948
	d4-10	-0,319	0,070	0,000	0,727	-0,024	0,054	0,661	0,976	1,024
	d4-11	-0,285	0,073	0,000	0,752	-0,135	0,060	0,025	0,874	1,145
VALOR MÉDIO FATURA MENSAL	missing/não possui vlr. médio fatura mensal	-	-	-	-	-	-	-	-	-
	até R$50,01	-	-	-	-	-0,031	0,040	0,443	0,970	1,031
	R$51,00 a R$149,01	-	-	-	-	-0,050	0,036	0,171	0,951	1,051
	R$150,00 a R$499,01	-	-	-	-	0,032	0,033	0,324	1,033	0,968
	R$500,00 a R$1.100,01	-	-	-	-	0,150	0,038	0,000	1,162	0,861
	acima de R$1.101,01	-	-	-	-	0,364	0,051	0,000	1,440	0,695
QTDE ANOS ASSOCIADO CARTÃO	missing/não possui	-	-	-	-	-	-	-	-	-
	até 1 ano associado cartão crédito	0,173	0,053	0,001	1,189	-0,063	0,028	0,022	0,939	1,065
	2 anos associado cartão crédito	0,438	0,054	0,000	1,550	-0,221	0,030	0,000	0,801	1,248
	3 anos associado cartão crédito	0,527	0,060	0,000	1,694	-0,284	0,034	0,000	0,753	1,328
	4 anos associado cartão crédito	0,618	0,069	0,000	1,855	-0,346	0,042	0,000	0,707	1,414
	4 a 6 anos associado cartão crédito	0,728	0,057	0,000	2,071	-0,406	0,034	0,000	0,667	1,500
	6 a 8 anos associado cartão crédito	0,830	0,076	0,000	2,294	-0,474	0,048	0,000	0,623	1,606
	8 ou mais anos associado cartão crédito	0,877	0,060	0,000	2,403	-0,538	0,038	0,000	0,584	1,713
d5	d51	-	-	-	-	-	-	-	-	-
	d52	-	-	-	-	0,127	0,057	0,025	1,136	0,880
	d53	-	-	-	-	0,058	0,059	0,327	1,060	0,944
	d54	-	-	-	-	-0,004	0,066	0,952	0,996	1,004
	d55	-	-	-	-	0,079	0,040	0,049	1,082	0,924
	d56	-	-	-	-	-0,039	0,076	0,610	0,962	1,040
	d57	-	-	-	-	-0,174	0,087	0,047	0,840	1,190
d6	d61	-	-	-	-	-	-	-	-	-
	d62	0,798	0,027	0,000	2,222	-0,555	0,019	0,000	0,574	1,742
	d63	-0,597	0,043	0,000	0,550					
d7	d71	-	-	-	-	-	-	-	-	-

Variável	Categoria	Coeficientes								
		Regressão Logística				Modelo Cox				
		β	erro pad.	Sig.	Exp(β)	*β*	erro pad.	Sig.	Exp(*β*)	1/Exp (*β*)
	d72	-	-	-	-	-0,246	0,114	0,031	0,782	1,278
	d73	-	-	-	-	-0,465	0,121	0,000	0,628	1,591
	d74	-	-	-	-	-0,022	0,051	0,671	0,979	1,022
	d75	-	-	-	-	-0,116	0,064	0,068	0,890	1,123
e2	e21	-0,158	0,059	0,007	0,854	0,138	0,036	0,000	1,148	0,871
	e22	-	-	-	-	-	-	-	-	-
	e23	-0,291	0,131	0,027	0,747	0,201	0,073	0,006	1,222	0,818
	e24	-0,490	0,099	0,000	0,613	0,333	0,054	0,000	1,396	0,716
	e25	-0,329	0,089	0,000	0,719	0,252	0,052	0,000	1,287	0,777
e3	e31	-1,107	0,217	0,000	0,330	0,742	0,170	0,000	2,100	0,476
	e32	-	-	-	-	-	-	-	-	-
	e33	-0,498	0,277	0,073	0,608	0,368	0,209	0,078	1,445	0,692
	e34	-0,620	0,270	0,022	0,538	0,449	0,202	0,026	1,566	0,639
	e35	-1,633	0,241	0,000	0,195	1,035	0,179	0,000	2,814	0,355
VALOR MÉDIO CHQ DEVOLVIDO MOT 11	missing/não possui	-	-	-	-	-	-	-	-	-
	até R$557,00	-0,542	0,058	0,000	0,581	0,288	0,033	0,000	1,334	0,750
	de R$558,00 a R$2.930,00	-0,884	0,080	0,000	0,413	0,409	0,043	0,000	1,506	0,664
	acima de R$2.931,00	-1,119	0,138	0,000	0,327	0,440	0,071	0,000	1,553	0,644
e4	e41	-	-	-	-	-	-	-	-	-
	e42	-	-	-	-	0,045	0,061	0,456	1,046	0,956
	e43	-	-	-	-	0,161	0,062	0,010	1,174	0,852
QTDE DIAS ABERTURA CONTA	missing/não possui	-	-	-	-	-	-	-	-	-
	até 5 dias abertura conta	-0,860	0,044	0,000	0,423	0,171	0,016	0,000	1,186	0,843
	6 a 28 dias abertura conta	-0,785	0,044	0,000	0,456	0,154	0,016	0,000	1,167	0,857
	29 a 130 dias abertura conta	-0,606	0,044	0,000	0,546	0,026	0,017	0,117	1,026	0,974
	131 a 229 dias abertura conta	-0,494	0,049	0,000	0,610	-0,041	0,022	0,057	0,960	1,042
	230 a 448 dias abertura conta	-0,360	0,050	0,000	0,698	-0,127	0,023	0,000	0,881	1,136
	439 a 1642 dias abertura conta	-0,372	0,049	0,000	0,690	-0,118	0,024	0,000	0,889	1,125
	acima de 1643 dias abertura conta					-0,364	0,029	0,000	0,695	1,440

Nota (*): os coeficientes das categorias d63 e "acima de 1.643 dias de abertura de conta" foram excluídas pelo SPSS da regressão logística e Modelo Cox, respectivamente, por constituírem a combinação linear de outras variáveis, causando multicolinearidade.

O efeito de cada variável sobre o *score* da operação de crédito pode ser descrito observando os coeficientes estimados em ambas as regressões. Os resultados são os seguintes:

- TEM PLANO DE SAÚDE: os coeficientes daqueles que possuem e não possuem plano de saúde são semelhantes entre si, não constituindo diferença significativa que influencie no probabilidade de ser bom ou mau pagador. Entretanto, clientes que não responderam a esta variável (*missing*) obterão *score* inferiores do que aqueles que responderam;

- GRAU DE INSTRUÇÃO: a relação obtida é perfeitamente plausível, já que pessoas com maior grau de instrução tendem a possuir maior poder aquisitivo, com maior capacidade de pagamento, e por isso obtêm maior *score*. Observa-se que aqueles possuem ensino fundamental completo ou médio incompleto têm os maiores riscos de inadimplência (*exp* β do Modelo Cox iguais 1,011 e 1,061, respectivamente). Já os clientes com doutorado obtêm o menor risco e, consequentemente, seus contratos tendem a demorar mais tempo para tornarem-se inadimplentes;

- IDADE: a relação encontrada mostra que operações de pessoas mais velhas tendem a demorar mais tempo para tornarem-se inadimplentes, logo, apresentam menor risco e maior *score*. Clientes com 60 anos ou mais são os que obtêm maior *score*, enquanto quem têm entre 26 e 36 são os que menos ganham *score*. Intuitivamente, a explicação encontra-se no fato de que pessoas mais velhas normalmente apresentam melhores condições de pagamento por possuírem renda certa, seja por meio de aposentadorias, pensões ou auxílios sociais;

- SEXO: esta variável mostrou-se relevante somente no Modelo Cox, onde contratos de clientes do sexo masculino obtêm chances ligeiramente menores de sobrevivência do que os do sexo oposto;

- ESTADO CIVIL: operações de viúvos(as) apresentam maior *score* e maiores chances de sobrevivência, enquanto que as operações dos divorciados(as) são as que têm menor *score* e maior risco de inadimplência, enquanto clientes casados não sofrem impacto em suas operações;

- UF: clientes que residem nos estados das regiões Sul e Sudeste - especificamente SC, PR, MG e ES - possuem maior *score* em suas operações, e os piores são para os que residem nos estados da região Norte, com exceção dos estados de TO e MA;

- RENDA LÍQUIDA FORMAL: a relação encontrada demonstra que quanto maior a faixa de renda, menor será o *score* do cliente. Este resultado mostra-se contraditório, uma vez que o esperado é que clientes com maior poder aquisitivo tenham melhores condições de pagamento e por isso apresentem menor risco. A explicação para este fato pode estar pautada na existência de dois tipos de viés e de uma justificativa de apelo intuitivo: *i)* existência de um viés na entrada de dados das variáveis de renda, uma vez que a mesma é apurada de acordo com critérios subjetivos de cada funcionário da instituição financeira, diante da existência de vários tipos de contracheques; *ii)* é comum as instituições financeiras mensurarem o limite de crédito a ser concedido de acordo com o nível de renda do tomador, logo, observa-se que há uma correlação direta entre estas duas variáveis. Considerando que o limite de crédito somente é definido após a aprovação do tomador, o mesmo não representa uma variável explicativa do modelo, estando, portanto, omitida e embutida no erro (ε). Desta forma, nota-se a existência de um segundo viés, constituído pela correlação do erro com uma das variáveis explicativas, representada pela "Renda Líquida Formal". Wooldridge (2006) denomina este tipo de viés como viés de variável omitida, ocasionada quando uma variável relevante não está presente no modelo; *iii)* tomadores com alto poder aquisitivo normalmente recorrem a créditos rotativos quando estão com sua vida financeira altamente comprometida, sem outras opções de créditos mais baratos. Nesta ótica, é de se esperar que quanto maior o nível de renda, maior seja o endividamento do tomador e, consequentemente, maior a probabilidade de inadimplência;

- RENDA INFORMAL: não há um comportamento claro desta variável em relação ao *score*, fato que também pode estar relacionado às mesmas explicações atribuídas à variável renda líquida formal;

- SALDO DE APLICACÕES EM OUTRAS INSTITUICÕES: o resultado mostrou-se de acordo com o esperado: quanto maior o saldo de aplicações, mais tempo o contrato

demora para tornar-se inadimplente. Novamente, clientes com maior quantidade de recursos, apresentam melhor capacidade de pagamento e maior *score*. Assim, operações de clientes com saldo de aplicações superior a R$13.000,00 obtêm maior *score*, enquanto os que não possuem não sofrem impacto de suas operações;

- VALOR MÉDIO FATURA MENSAL: a relação encontrada demonstra que quanto maior o valor da fatura, menor é a probabilidade de sobrevivência do contrato ao longo do tempo. Faturas de até R$50,00 possuem o menor risco de inadimplência, enquanto as acima de R$1.101,00 apresentam o maior risco. De forma análoga à variável "tem plano de saúde", a análise econômica destes resultados sugere que clientes com maior dispêndio possuem maior comprometimento de sua renda mensal, portanto, maiores chances de inadimplência. Vale ressaltar que essa variável somente foi relevante no modelo Cox;

- QTDE ANOS ASSOCIADO CARTÃO: a relação obtida é a de que quanto mais tempo o cliente tiver o cartão de crédito, maior será seu *score*. Intuitivamente, o resultado pode ser analisado da seguinte forma: administradoras de cartões de crédito tendem a não manter como clientes aqueles que apresentam mau comportamento de crédito (não pagamento de faturas, pagamentos em atraso etc.), permanecendo somente clientes considerados bons. Assim, espera-se que clientes com mais tempo de associação sejam considerados como bons, obtendo maior *score*;

- VALOR CHEQUE DEVOLVIDO MOTIVO 11: a relação encontrada é perfeitamente plausível, já que quanto menor for o valor de cheques devolvidos, maior será o *score* e tempo de sobrevivência do contrato. Clientes sem cheques devolvidos têm *score* maior do que aqueles que o possuem. Menor *score* é atribuído aos que têm cheques devolvidos por motivo 11 acima de R$2.931,00;

- QTDE DIAS EXISTÊNCIA CONTA: o resultado é similar ao obtido na variável "qtde anos associado cartão". As melhores operações são aquelas de clientes com mais tempo de conta, enquanto que as piores são aquelas de correntistas com até 5 dias de abertura de conta;

- Constante: a presença de uma constante β_0 na regressão logística no valor de 1,612 ou *exp* β_0 = 5,011 indica que, caso o cliente esteja na categoria de referência de todas as variáveis explicativas, o *score* será igual a {*exp* (1,612)/[1+*exp*(1,612)]}.100 = [5,011/(1+5,011)].100 = 83 pontos.

Com as observações dos coeficientes das variáveis é possível informar quais as que causam maior impacto no *score* e no tempo de sobrevivência do contrato, bastando analisar a amplitude do coeficiente entre as categorias (diferença entre o coeficiente máximo e o mínimo), de forma que as que possuírem maior amplitude serão as que causarão maior impacto. Dentre as variáveis com maior impacto no *score* da regressão logística, destacam-se a d6 (variável de perfil, do tipo financeira - informações bancárias), "Saldo de aplicações em outras Instituições", "Grau de Instrução", "Qtde anos associado cartão" e "Idade". Já no Modelo Cox, dentre algumas das variáveis que se destacam com maior impacto no tempo de sobrevivência do contrato estão a e3 (variável comportamental do tipo apontamentos negativos), "UF", "Qtde dias abertura conta", "Valor médio fatura cartão de crédito", d4 (variável de perfil, do tipo financeira - informações bancárias) e "Saldo de Aplicação em outras Instituições".

Com base nos coeficientes obtidos na tabela 17, é possível determinar as equações finais das variáveis dependentes das regressões Logística ($Y^*_{RL,i}$) e Modelo Cox ($Y^*_{MC,i}$) para a operação *i*, ou seja, a probabilidade condicional de $Y^*_{RL,i}$ ser igual a 1 e o tempo de sobrevivência do contrato *i*, dadas as respostas das variáveis explicativas do indivíduo. No caso da regressão logística, teremos a seguinte equação:

$$Y^*_{RL,i} = \pi^*(X_i) = \frac{e^{g^*(X_i)}}{1 + e^{g^*(X_i)}} \qquad (20)$$

em que:

$g^*(Xi)$ = 1,612 + 0,248.*Dplano_saúde1* + 0,255.*Dplano_saúde2* – 0,286.*Dseg_pessoal1* + 0,377.*Dseg_pessoal2* – 0,012.*Dgrau_instru1* - ... + 0,808.*Dgrau_instru5* + 0,149.*Didade1* + ... + 0,774 *Didade6* – 0,542.*Dchq_m11_1* - ... – 1,119.*Dchq_m11_3*

A equação de $g^*(Xi)$ para a regressão logística contém 122 coeficientes, sendo D_{Xk} o valor da variável *dummy* da k-ésima categoria da variável explicativa *X*, ou seja, $D_{Xk} = 1$ se a operação *i* estiver na categoria *k* da variável *X* ou $D_{Xk} = 0$ se a operação não estiver na categoria *k* da variável *X*. Por exemplo, *Dgrau_instru1* é o valor da *dummy* para a primeira categoria do grau de instrução, em que *Dgrau_instru1* = 1 se a operação estiver na categoria 1 desta variável, ou seja, caso o tomador possua o ensino fundamental. *Dgrau_instru1* será igual a 0 caso a operação não esteja na categoria 1 do grau de instrução. O *score* da operação de crédito será, então, dado por:

$$Score = 100.\ Y^*_{RL,i}$$

Observa-se que o *score* reflete a probabilidade do tomador ser um bom pagador ($Y^*_{RL,i} = 1$) multiplicado por 100, de forma que coeficientes positivos aumentam o *score*, coeficientes negativos diminuem e coeficientes nulos não causam impacto. Além disso, os coeficientes mais positivos da tabela 17 são as características que mais aumentam o *score* e os coeficientes mais negativos são os que mais diminuem o *score*.

A partir dos coeficientes estimados na tabela 17, também é possível obter a função de risco do Modelo Cox apresentada na equação 16:

$$h(t / X, B) = h_0(t) \,.\, e^{(X'B)}$$

A função de risco baseline $h_0(t)$ é calculada pelo software SPSS. Com base na equação 18 é possível calcular a função de sobrevivência baseline $S_0(t)$, possibilitando-nos estimar a função de sobrevivência dada pela equação 17, abaixo:

$$S(t / X, B) = S_0(t)^{e(X'B)}$$

A expressão $e^{(X'B)}$ possui 155 coeficientes, sendo representada por:

$$e^{(X'B)} = \exp [-0{,}144.Dplano\_saúde1 - 1{,}104.Dplano\_saúde2 + 0{,}011.Dgrau\_instru1 + ... - 0{,}519.Dgrau\_instru5 - 0{,}094.Didade1 - ... - 0{,}498\ Didade6 + 0{,}288.Dchq\_m11\_1 - ... +0{,}44.Dchq\_m11\_3]$$

Após estimação dos modelos finais, foram feitas avaliações por meio do teste de razão de verossimilhança para verificar a significância estatística conjunta de todos os coeficientes das regressões. O teste é dado por:

$$G = -2\ [L(\text{modelo final}) - L\ (\text{modelo sem nenhuma variável})]$$

onde G segue uma distribuição χ^2 com p graus de liberdade, sendo p o número de coeficientes estimados nos modelo finais (p = 122 para regressão logística e 155 para Modelo Cox) e hipótese nula de que todos os coeficientes dos modelos finais são estatisticamente iguais a zero. Substituindo os valores estimados pelo SPSS para as funções de log-verossimilhança o resultado é:

$$G_{RL} = 97.789{,}02 - 82.971{,}63 = 14.817{,}39$$
$$G_{MC} = 868.875{,}22 - 845.944{,}76 = 22.930{,}46$$

Os valores da estatística G para a regressão logística e modelo Cox indicam a rejeição da hipótese nula a qualquer nível de significância. Desta forma, pode-se afirmar que o conjunto de coeficientes estimados para ambas as regressões é estatisticamente significante, ou seja, não nulo.

Os coeficientes também foram testados individualmente, bem como o conjunto de coeficientes de cada variável. Em todos os casos foi elaborado o teste Wald (teste similar ao teste *G*) baseado nas estimativas dos coeficientes e em seus erros-padrões, cujos resultados com nível de significância de 5%, estão dispostos na tabela 17. Os testes mostraram que 98 dos 122 coeficientes da regressão logística e 116 dos 155 coeficientes do Modelo Cox foram considerados estatisticamente diferentes de zero, rejeitando-se, portanto, a hipótese nula de que os mesmos são iguais a zero. Os demais 24 coeficientes para a regressão logística e 39 do Modelo Cox apresentaram *p-values* superiores a 5%, levando à aceitação da hipótese nula de que tais coeficientes, individualmente, são iguais a zero.

Apesar dos testes individuais terem revelado a existência de coeficientes estatisticamente nulos, todos os testes conjuntos mostraram que cada grupo de coeficientes é estatisticamente não nulo a 5% de significância, e que o modelo como um todo possui significância estatística, corroborado pelo teste *G*. Desta forma, os modelos finais apresentados na tabela 17 não foram modificados, sendo mantidas as variáveis *dummy* com coeficientes individualmente não significantes.

4.2 Medidas de Desempenho

Verificada a significância estatística dos modelos, o próximo passo é testar o grau de eficiência, isto é, a eficácia com que se consegue prever as variáveis dependentes. Em linhas gerais, uma medida para verificar o poder explicativo do modelo é comparar a diferença entre os valores estimados de Y_i^* com os valores observados de Y_i. Quando esta diferença é pequena, diz-se que o modelo possui poder explicativo elevado, ou seja, estima Y_i com boa precisão. Existem indicadores que mensuram o grau de eficiência dos modelos de risco de crédito, de acordo com o tipo de regressão aplicado. A seguir serão apresentados alguns destes indicadores para as regressões abordadas.

4.2.1 Medidas de Eficiência *(goodness of fit)* para a regressão logística

O teste utilizado para verificar o poder explicativo do modelo de regressão logística foi o teste Hosmer-Lemeshow, calculado automaticamente pelo SPSS. Este teste é calculado dividindo a amostra, por ordem crescente de *score*, em 10 centis de tamanhos aproximados (7.929 observações em cada grupo), para, posteriormente, somar o quadrado das diferenças entre os valores estimados e observados, ponderada pela quantidade de casos em cada grupo e pelos percentuais de operações boas e ruins dentro de cada centil. A partir de então, é gerada uma estatística *C* comprovadamente distribuída por χ^2 com g-2 graus de liberdade, em que g é o número de centis adotado. A tabela 18 traz a tabela de contingência utilizada no teste Hosmer-Lemeshow, a qual contém a quantidade observada e esperada de Y_i e Y_i^*:

TABELA 18 - Teste Hosmer-Lemeshow para a regressão logística

Grupo	Qualidade do crédito = mau		Qualidade do crédito = bom		Total
	Observado	Esperado	Observado	Esperado	
1	6.668	7.063	1.261	866	7.929
2	6.461	6.409	1.468	1.520	7.929
3	6.027	5.898	1.902	2.031	7.929
4	5.461	5.342	2.468	2.587	7.929
5	4.917	4.700	3.012	3.229	7.929
6	4.095	3.862	3.834	4.067	7.929
7	2.842	2.765	5.087	5.164	7.929
8	1.635	1.862	6.294	6.067	7.929
9	1.018	1.186	6.911	6.743	7.929
10	490	527	7.436	7.399	7.926
TOTAL	39.614	39.614	39.673	39.673	79.287

Como era de se esperar, os primeiros grupos são os que possuem *score* inferior, e, portanto, dispõem de grande quantidade de contratos maus e poucos bons, enquanto os últimos grupos, por possuírem alto *score,* têm muitos contratos bons e poucos maus.

O resultado para o teste Hosmer-Lemeshow calculado pelo SPSS a partir do modelo estimado na tabela 18 foi de C = 346,09. Ao nível de significância de 5% e com 8 graus de liberdade, o valor crítico da distribuição χ^2 é 15,507, o que nos leva a rejeitar a hipótese nula de que o modelo estimado é eficiente para estimar Y_i. Entretanto, existem dois potenciais problemas com o teste Hosmer-Lemeshow, alertado por alguns autores[34]: *i)* são necessárias amostras suficientemente grandes para que o número de observações em cada grupo não seja inferior a 5 observações; *ii)* em amostras muito grandes, como o caso em questão, o qual conta com uma base de modelagem de mais de 70.000 observações, é fácil rejeitar a hipótese nula, uma vez que o valor da estatística χ^2 é proporcional ao tamanho da amostra.

Tendo em vista que o teste de Hosmer-Lemeshow mostrou-se inadequado, é aconselhável aplicar outros testes que comprovem a eficácia do modelo. Assim, foram aplicados os testes de Kolgomorov-Smirnov (estatística KS) e a curva ROC (*Receiver Operation Characteristic*) ou Diagrama de Lorenz[35].

A estatística KS é usada na teoria estatística não paramétrica para testar se as funções de distribuição de uma variável são iguais em dois grupos, sendo largamente utilizada nos modelos de *credit scoring*. A hipótese deste teste é supor que indivíduos com alta chance

[34] ADVANCED Statistical Analysis Using SPSS. V14.0, pág 3-11

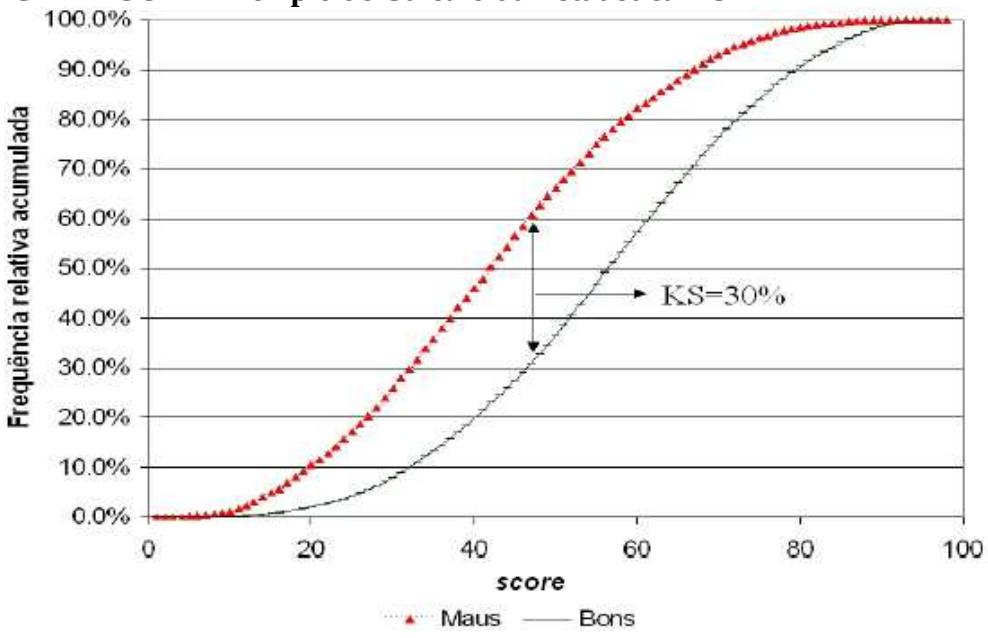
[35] A denominação Diagrama de Lorenz ocorre em função da semelhança com a Curva de Lorenz, desenvolvida por Max O. Lorenz para descrever a desigualdade social (DRISLANE e PARKINSON, 2001).

de serem inadimplentes devem estar concentrados no baixo valor predito, e indivíduos com alta chance de se tornarem adimplentes devem estar concentrados no alto valor predito. Desta forma, um bom modelo deverá prover a maior separação entre os bons e maus tomadores ao longo do valor predito, de modo que a curva de distribuição acumulada da população de maus pagadores cresça antes da curva de distribuição acumulada dos bons pagadores. Em virtude disso, a estatística KS pode ser definida como:

$$KS = \max [F_{maus}(score) - F_{bons}(score)]$$

onde F_{maus} = distribuição acumulada dos inadimplentes e F_{bons} é a distribuição acumulada dos adimplentes. A seguir, um exemplo gráfico do KS:

GRÁFICO 2 - Exemplo de Cálculo da Estatística KS



A medida de desempenho varia entre 0 e 1, de forma que quanto mais próximo de 1 for o KS, melhor é o modelo. O resultado do teste KS obtido para o modelo de regressão logística foi de 0,503 a um nível de significância de 0,000, indicando a rejeição da hipótese nula de que as distribuições acumuladas dos bons e maus clientes são iguais. Significa dizer que o modelo foi considerado eficiente para discriminar bons e maus tomadores.

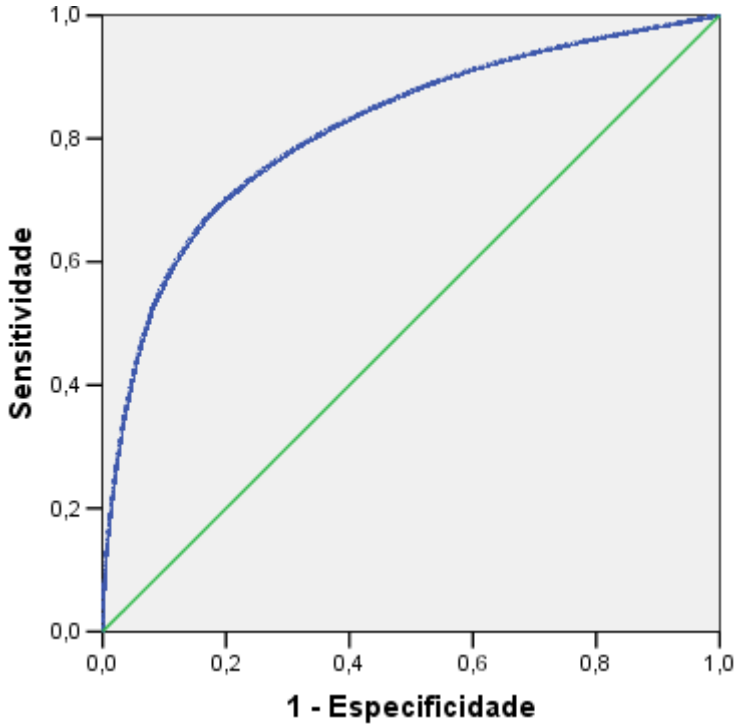
Outro teste comumente utilizado para medir a eficiência dos modelos de *credit scoring* é a curva ROC. Para o cálculo desta medida é necessário apresentar as seguintes definições:

- Sensitividade: corresponde à proporção de clientes bons classificados corretamente, por meio de um modelo qualquer, por terem um *score* superior a um determinado ponto de corte (*Pc*);

- Especificidade: corresponde à proporção de clientes maus classificados corretamente por terem um *score* inferior a determinado ponto de corte;

Quanto maior a sensitividade e especificidade, melhor o modelo. À medida que cresce o ponto de corte, a sensitividade diminui e a especificidade aumenta, ou seja, ambas as medidas dependem de *Pc*. Para a construção da curva ROC deve-se variar *Pc*, obtendo diferentes pontos de sensitividades e especificidades. O gráfico abaixo mostra a curva ROC do modelo em estudo, representada pela linha azul, em que a sensitividade da predição é observada no eixo das ordenadas e o índice (1-especificidade da predição) é observado no eixo das abscissas para cada um dos possíveis valores preditos do modelo:

GRÁFICO 3 - Curva ROC



A estatística é obtida a partir do cálculo da área abaixo da curva ROC, que pode variar de 0 a 1. Quanto mais próximo de 1 for a área abaixo da curva ROC, melhor será o modelo. A diagonal representa uma curva ROC = 0,50, a qual não possui poder

discriminatório, visto que a sensitividade e especificidade são iguais em todos os pontos de valores preditos, ou seja, o percentual de adimplentes é sempre igual ao de inadimplentes. A estatística ROC foi calculada utilizando o SPSS, cujo valor obtido foi de 0,816, indicando que o modelo possui excelente grau de eficiência em discriminar bons e maus pagadores, sendo a classificação dos autores Hosmer e Lemeshow (1989).

Para complementar o resultado do teste KS e curva ROC foi utilizada uma tabela de classificação fornecida pelo SPSS, que consiste na comparação cruzada entre a quantidade de casos com *Yi* observado igual a 1 ou 0 *versus* a quantidade de casos com *Yi* estimado igual a 1 ou 0. A tabela 19 contém os resultados da tabela de classificação para o modelo de regressão logística estimado, baseado na amostra de 79.287 operações de crédito e ponto de corte igual a 0,5:

TABELA 19 - Tabela de Classificação para amostra de modelagem

		Estimado			Percentual correto de classificação
		mau	bom	TOTAL	
Observado	mau	31.352	8.262	39.614	79,1%
	bom	11.588	28.085	39.673	70,8%
	TOTAL	42.940	36.347	79.287	75,0%

Os resultados acima mostram que o modelo classificou corretamente 75% do total das operações, com uma taxa de acerto de 79,1% para as operações ruins e 70,8% para as operações boas. Apesar de o modelo ter classificado com mais eficiência as operações boas, podemos afirmar que, de uma forma geral, espera-se que o modelo seja capaz de classificar corretamente 75% de todas as operações, independente de serem boas ou ruins.

4.2.2 Medidas de Eficiência para o modelo Cox

No estudo realizado por Hosmer e Lemeshow sobre análise de sobrevivência é abordada uma metodologia de *goodness fit* por eles elaborada. O procedimento é bastante similar ao aplicado no modelo de regressão logística, com algumas pequenas diferenças:

- 1º passo: inicialmente deve-se fixar um horizonte de tempo para o qual o teste será aplicado;

- 2º passo: de forma semelhante ao teste de Hosmer-Lemeshow na regressão logística, a amostra é segregada em dez centis de tamanhos aproximados de acordo com o risco de falha;

- 3º passo: computa-se a diferença entre a quantidade de operações inadimplentes observadas e estimadas;

- 4º passo: o resultado desta diferença é então dividido pela raiz quadrada da quantidade de operações inadimplentes estimadas naquele horizonte de tempo, obtendo-se o valor crítico "*z*" e o *p-value* de cada centil.

A hipótese nula do teste é a de que os valores observados e esperados são iguais. Desta forma, para que o modelo seja considerado eficiente, espera-se não rejeitar a hipótese nula em cada centil.

A tabela 20 resume os resultados obtidos para este teste:

TABELA 20 -Teste de Hosmer-Lemeshow para o modelo Cox para *t* = 12 meses

grupo	Inadimplentes		Total	*z*	*p-value*
	Observado	Estimado			
1	3.625	3.115	7.927	-9,14	0,000
2	2.802	2.973	7.928	3,13	0,002
3	2.894	2.790	7.928	-1,97	0,049
4	3.026	2.940	7.928	-1,59	0,112
5	3.149	2.739	7.928	-7,84	0,000
6	1.320	2.261	7.928	19,80	0,000
7	3.431	2.890	7.928	-10,07	0,000
8	1.934	2.465	7.928	10,69	0,000
9	2.293	2.381	7.928	1,80	0,072
10	1.848	1.974	7.936	2,84	0,005
Total	**26.322**	**26.526**	**79.287**	-	-

Com *p-values* iguais a 0,112 e 0,072 somente o quarto e nono decis indicam a aceitação da hipótese nula a 5% de significância. Implica dizer que, em sua maioria, há diferenças significantes entre as quantidades observadas e estimadas, sinalizando que o modelo Cox não possui bom ajuste. Contudo, como já identificado no teste de Hosmer-Lemeshow da regressão logística, esta metodologia sofre de um potencial problema ao lidar

com amostras grandes. Com efeito, esta situação pôde ser novamente constatada na aplicação desta técnica numa base com um número excessivo de observações.

Diante do exposto, decidiu-se aplicar um novo teste de *goodness-fit*, desta vez baseado na minimização dos erros tipo I e tipo II, realizado por Martins (2003). De acordo com o autor, um bom modelo deve exibir percentuais baixos de erro tipo I (aprovar crédito para indivíduos inadimplentes), pois erros de classificação redundam em altos custos às instituições financeiras. Por outro lado, um modelo razoavelmente preciso também deverá minimizar o erro tipo II (negar créditos a bons pagadores). Contudo, o erro tipo II deve ser criteriosamente analisado, pois pode representar um contrato que, efetivamente, venha a falhar no futuro. Nessa situação, o erro tipo II representa um sucesso, pois indica que o modelo sinalizava, antecipadamente, uma falha futura.

O teste aplicado baseia-se na comparação das probabilidades de sobrevivência previstas pelo modelo para os horizontes de 12 e 24 meses com valores de corte específicos para cada horizonte de tempo. Estes foram obtidos a partir do percentual de contratos bons existentes na amostra, de acordo o horizonte de tempo, obtendo-se os seguintes resultados: 0,668 para 12 meses e 0,50 para 24 meses. Assim, sempre que a probabilidade de sobrevivência for inferior ao valor de corte, a mesma é classificada como uma possível inadimplência (ruim). Caso contrário, o contrato é classificado como adimplente (bom).

A tabela 21 apresenta a classificação dos contratos de toda base de dados segundo o Modelo Cox para 12 e 24 meses:

TABELA 21 - Classificação do Modelo Cox

Meses	Bem classificados	Erro tipo I	Erro tipo II	TOTAL
12	116.364 (68%)	6.950 (4%)	48.147 (28%)	171.461 (100%)
24	120.984 (71%)	9.262 (5%)	41.215 (24%)	171.461 (100%)

Os resultados demonstram que o modelo classificou corretamente 68% dos contratos bons e ruins no horizonte de 12 meses. O percentual de erro tipo I foi de apenas 4%, enquanto que para o erro tipo II foi de 28%. No horizonte de tempo de 24 meses há uma sensível melhoria na eficiência do modelo, onde 71% dos contratos foram corretamente classificados.

Entretanto, considerando os erros de classificação associados a contratos que efetivamente falharam em determinado momento no futuro, tem-se que o erro tipo II reduz para 7%, elevando o percentual de acerto do modelo para 89% no horizonte de 12 meses. A tabela 22 sumariza os resultados:

TABELA 22 - Nível de acerto global para o Modelo Cox

Meses	Bem classificados	Erro tipo I	Erro tipo II	TOTAL
12	152.965 (89%)	6.950 (4%)	11.546 (7%)	171461 (100%)
24	120.984 (71%)	9.262 (5%)	41.215 (24%)	171461 (100%)

Note que não foram observadas variações nos percentuais de acerto e erro para o horizonte de tempo de 24 meses, visto que o período de observação do estudo limita-se a este horizonte de tempo. Assim, não foi possível identificar se as operações censuradas incorreram em inadimplência em determinado momento futuro.

A conclusão que pode ser feita a partir da análise dos resultados observados é que os modelos gerados por meio da regressão logística e modelo Cox para a linha de crédito em estudo são válidos estatisticamente, com coeficientes considerados significantes individual e conjuntamente, e que o poder de classificação resultante indica que a aplicação dos modelos poderá ser capaz de classificar corretamente e estimar o tempo esperado de inadimplência de uma porção considerável de todas as operações de crédito submetidas às análises.

4.3 Exemplo

Uma vez verificada a significância estatística dos modelos e seu grau de eficiência, é interessante aplicar um exemplo prático de mensuração do *score* e do tempo esperado até a inadimplência, clarificando a forma de como os cálculos são realizados. Cumpre lembrar que o *score* do indivíduo é representado pelo valor arredondado da seguinte função:

Score = 100 * exponencial (equação) / 1 + exponencial (equação)

em que “equação” é o valor resultante da soma dos coeficientes estimados aplicáveis à operação de crédito. A função de sobrevivência no horizonte de tempo *t* pode ser calculada da seguinte forma:

Função de sobrevivência = (função de sobrevivência baseline para o tempo t) $^{\text{exponencial (equação)}}$

O tempo esperado até a inadimplência do contrato é, então, calculado somando-se as funções de sobrevivência de cada horizonte de tempo.

EXEMPLO 1 - Cálculo do *Score* e Função de Sobrevivência da Operação de Crédito

		Regressão Logística	Modelo Cox
		Coeficiente *(β)*	
Nome:	José Antônio de Melo		
Tem plano de saúde:	sim	0,248	-0,144
a1	codificado	-0,286	0,168
a2	codificado	não significante	-
a3	codificado	não significante	-0,150
a7	codificado	não significante	-
a8	codificado	não significante	-0,018
a9	codificado	-0,022	-0,058
Situação veículo	quitado	0,170	-0,079
Tipo de imóvel:	casa	-0,018	0,012
a12	codificado	0,217	-0,142
Ano veículo	2004	0,159	-0,042
Grau de instrução:	superior completo	0,705	-0,450
Idade:	30	-	-
Sexo:	M	não significante	-
Estado civil:	solteiro	-0,146	0,140
UF:	DF	-0,333	0,199
b2	codificado	não significante	-
b3	codificado	-	-
c1	codificado	0,134	-0,249
c2	codificado	não significante	0,217
Renda líquida formal:	R$4.800,00	-0,236	0,115
c3	codificado	-2,280	1,278
c4	codificado	1,353	-0,684
Renda informal:	R$0,00	-	-
Saldo de aplicações em outras instituições:	R$0,00	-	-
d4	codificado	-0,285	-0,135
Valor médio fatura mensal:	R$1.500,00	não significante	0,364
Qtde anos associado cartão:	6	0,728	-0,474
d5	codificado	não significante	-
d6	codificado	-	-
d7	codificado	não significante	-
e2	codificado	-0,158	0,138
e3	codificado	-1,107	0,742
Valor médio cheque	0	-	-

EXEMPLO 1 - Cálculo do *Score* e Função de Sobrevivência da Operação de Crédito			
devolvido mot. 11:			
e4	codificado	não significante	-
Qtde dias abertura conta:	2.160	-	-0,364
CONSTANTE		1,612	
soma dos coeficientes (A)		0,456	0,383
exp (A)		1,577	1,467
score = 100 * exp(A)/1+exp (A)		**61,201**	
função de sobrevivência baseline para o tempo t = 12 meses (B)			0,848
função de sobrevivência para t = 12 meses			**0,786**
função de sobrevivência baseline para o tempo t = 24 meses (C)			0,731
função de sobrevivência para t = 24 meses			**0,631**
tempo esperado até inadimplência			**18,92 meses**

Nota: Os valores da função de risco e sobrevivência *baseline* para os horizontes de 1 a 24 meses estão dispostos no anexo C.

Observa-se que caso um indivíduo possua o perfil dado no exemplo, este terá sua operação de crédito avaliada como boa, uma vez que o *score* arredondado calculado pelo modelo de regressão logística foi de 61, valor acima do ponto de corte que é 50. Os resultados para o Modelo Cox mostram que a probabilidade de que o contrato sobreviva além dos 12 meses após sua contratação é de 78,6%, enquanto que a probabilidade de sobrevivência para além dos 24 meses é de 63,1%. Tais resultados são perfeitamente plausíveis, já que para um tomador avaliado como bom, não é de se esperar que seu contrato incorra em inadimplência num breve horizonte de tempo, fato corroborado pelo Modelo Cox, que mesmo após 24 meses de contratação da operação de crédito, prevê boas as chances de sobrevivência. A esperança de vida até que ocorra a inadimplência da operação de crédito do indivíduo com o perfil em análise é obtida somando-se as probabilidades de sobrevivência em cada horizonte de tempo, cujo resultado é de 18,92 meses.

A próxima seção traz análises gráficas que auxiliam na interpretação dos resultados apresentados.

4.4 Análises gráficas dos resultados

Estimados os coeficientes para os modelos de regressão logística e Modelo Cox por meio da base de modelagem, formada por 79.287 observações, é interessante aplicar o modelo em toda a base de dados do estudo, formada por 171.461 observações. Uma variação grande na capacidade de classificação pode ser um indício de que a amostra de modelagem sorteada estava viesada em relação à população de todas as operações de crédito disponíveis. A tabela 23 apresenta a classificação de todas as observações do estudo:

TABELA 23 - Tabela de Classificação para toda amostra

		Estimado			Percentual correto de classificação
		mau	bom	TOTAL	
Observado	mau	38.527	8.442	46.969	82,0%
	bom	43.106	81.386	124.492	65,4%
	TOTAL	81.633	89.828	171.461	73,7%

Analisando as taxas de acerto obtidas na tabela 22, gerada sobre a amostra de modelagem de 79.287 observações, percebe-se que houve pouca flutuação dos percentuais de classificação. Na amostra foram classificadas corretamente 79,1% das operações boas, enquanto que em todo o banco este percentual foi de 82%, sinalizando uma diferença sutil de 2,9%. Já as operações com *score* inferior a 50 foram classificadas corretamente em 70,8% da amostra e 65,4% em toda a base, uma variação aceitável. Ao todo, o modelo classificou corretamente 75% da amostra e 73,7% de toda a base de dados, apontando com uma pequena diferença de 2,3%, variação também aceitável. A ausência de grandes variações é indício de que a amostra sorteada para modelagem não estava enviesada em relação ao banco de dados integral e, portanto, o forte poder discriminatório verificado para os testes de poder explicativo baseados na amostra deve se manter em todo o banco de dados.

Os gráficos 4 e 5 demonstram a distribuição do *score* estimado das operações de crédito boas e ruins quando o modelo é aplicado em toda a base de dados.

GRÁFICO 4 - *Score* das operações de crédito boas

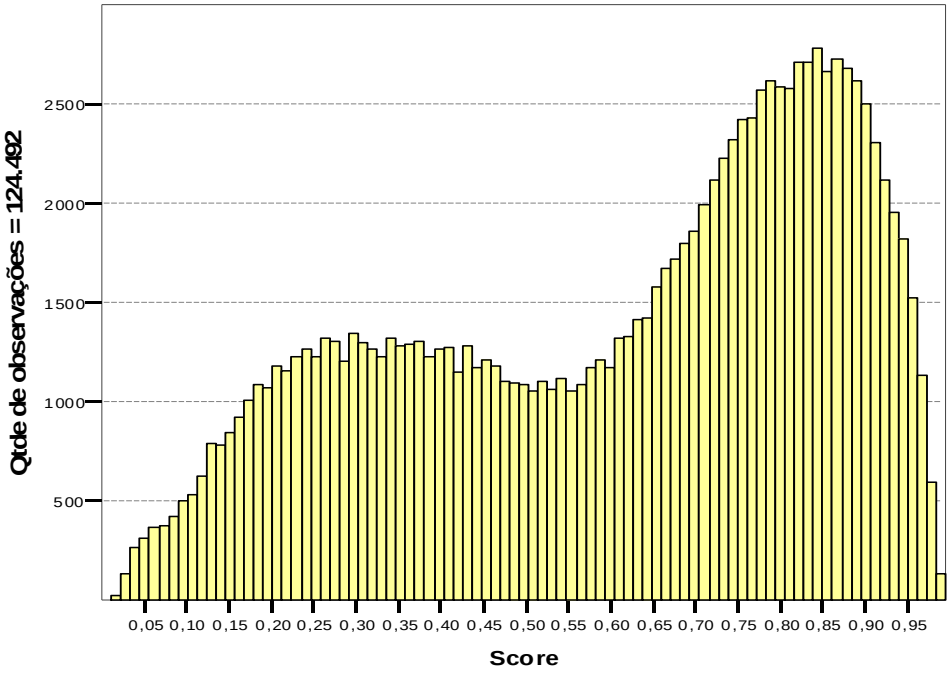
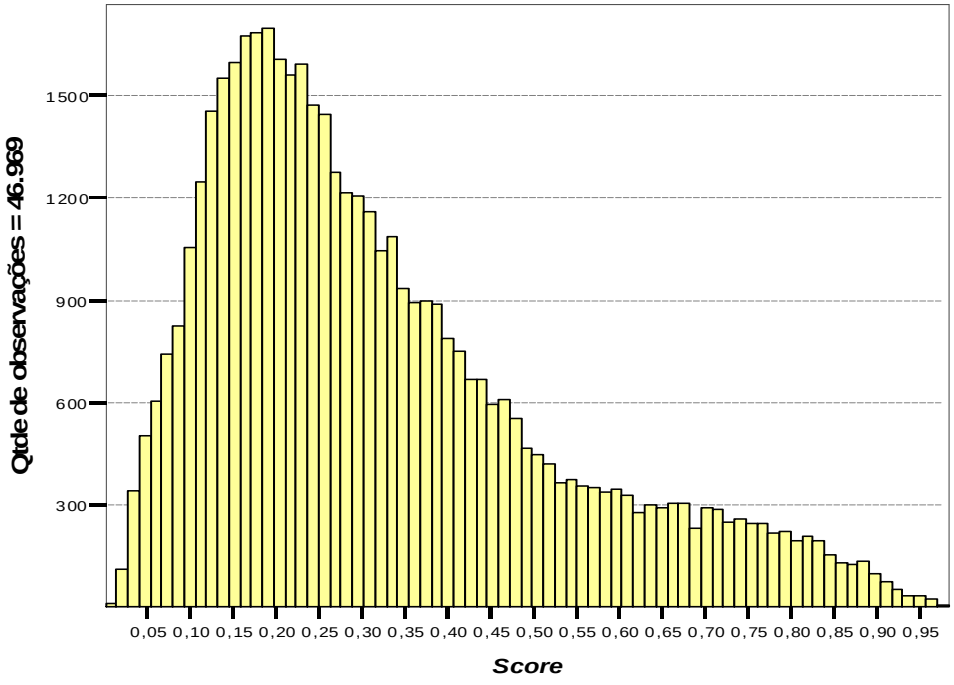


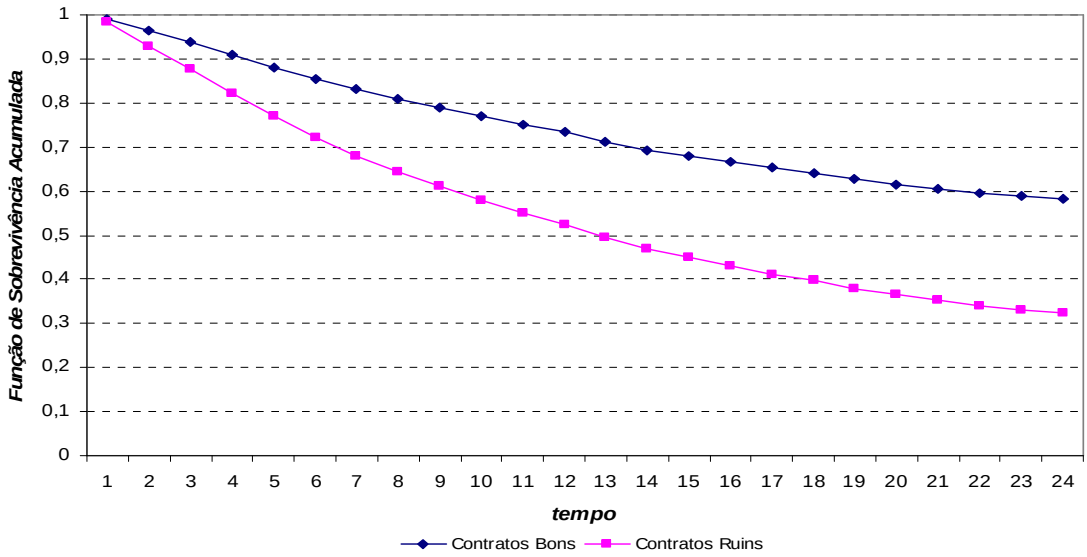
GRÁFICO 5 - *Score* das operações de crédito ruins



Conforme esperado, a distribuição de operações boas está concentrada mais à direita do *score* 50, com um valor médio de 61, enquanto que a distribuição do *score* das operações ruins está à esquerda do *score* 50, com um valor médio de 32.

Os gráficos 6 e 7 foram construídos por meio dos valores médios das covariáveis de todos os contratos bons e maus, discriminadamente, e mostram as funções de sobrevivência e risco acumuladas:

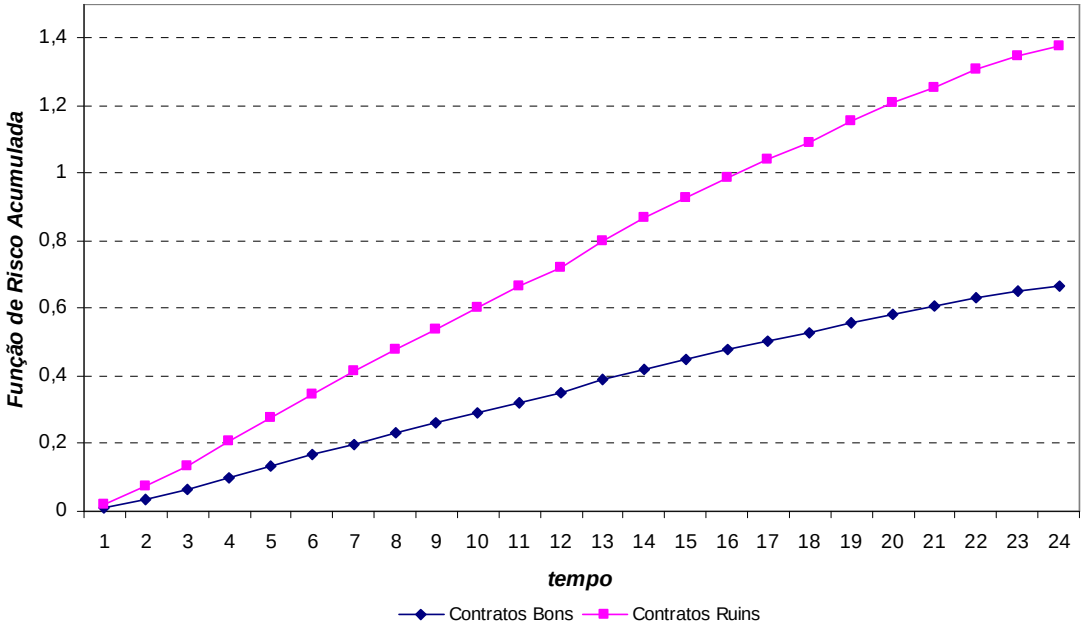
GRÁFICO 6 - Função de Sobrevivência Acumulada



Como esperado, as curvas de sobrevivência demonstram comportamento monotônico decrescente, visto que com o passar do tempo, menores são as chances de sobrevivência de um contrato. Vê-se claramente que a curva de contratos ruins é mais inclinada, apresentando menores chances de sobrevivência do que a curva de contratos com *score* superior a 50. Analisando o ponto onde t = 13 meses, por exemplo, as chances de sobrevivência dos contratos maus são em torno de 50%, enquanto que para os contratos bons são pouco acima de 70%. No ponto onde t = 24 meses os contratos bons mantêm chances de sobrevivência em torno de 60% e os contratos ruins pouco mais de 30%.

O gráfico 8 corrobora o fato de que contratos bons têm maiores chances de sobrevivência do que os contratos maus, por meio das funções de risco acumuladas:

GRÁFICO 7 - Função de Risco Acumulada

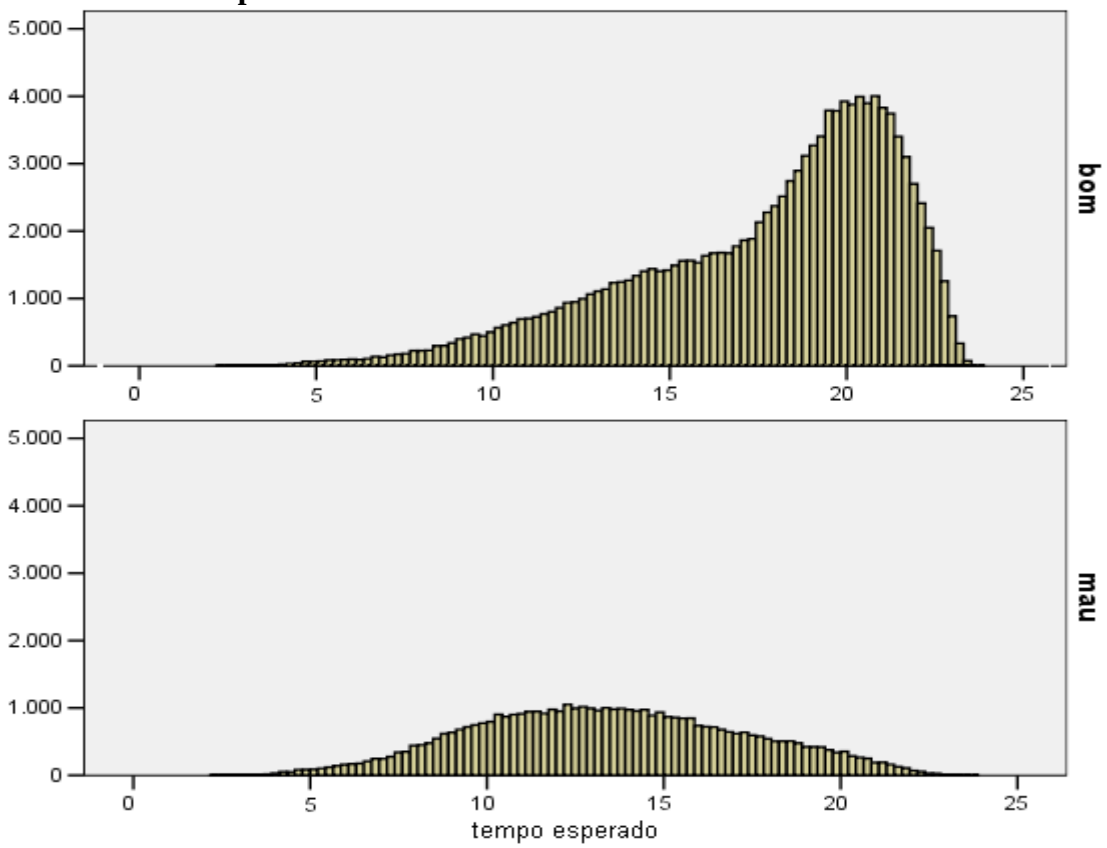


Nota-se que as curvas possuem comportamento monotônico crescente ao longo do tempo, indicando que, com o passar do tempo, maiores são as expectativas de falha (inadimplência) dos contratos. Novamente, como esperado, as curvas de risco indicam que a inadimplência ocorre mais rapidamente para as operações ruins ao longo do tempo. Importante lembrar que a função de risco representada neste gráfico não representa a probabilidade, e sim a taxa de risco, representada pela probabilidade condicional de inadimplência divida pelo período em particular. Ela indica a expectativa de inadimplência de uma operação em um determinado período, podendo variar de zero a infinito. Valores elevados indicam maiores chances de falha instantânea. [36]

O gráfico 8 foi obtido mediante a soma das funções de sobrevivência nos 24 horizontes de tempo, conforme demonstrado na equação 19. É notória a concentração de bons contratos na porção direita do histograma, sinalizando maior expectativa de sobrevivência do que os contratos ruins.

[36] ADVANCED Statistical Analysis Using SPSS. V14.0, pág 5-13

GRÁFICO 8 - Tempo de sobrevivência médio



Em média, espera-se que as operações boas sobrevivam 18 meses, enquanto que para as operações classificadas como ruins a expectativa é de 13 meses.

4.5 Estabilidade do Modelo

Alguns autores alertam que em modelos de *credit scoring* não é desejável que uma variável sozinha tenha poder para aprovar ou reprovar uma operação de crédito, quaisquer que sejam as respostas das outras variáveis[37]. Não se pode, por exemplo, negar o pedido de crédito devido ao fato do indivíduo possuir nível superior, ser solteiro ou ter casa própria. Nos Estados Unidos o *Consumer Act* de 1974 dispõe de uma série de medidas regulatórias que prevêem punições legais para os casos de discriminação de cor e raça. No Brasil, o CMN editou a Resolução nº. 2.682/99, que dispõe sobre os critérios de classificação das operações de crédito e as regras para constituição de provisão para créditos de liquidação duvidosa.[38] Estes critérios norteiam, inclusive, a avaliação de risco de crédito. Contudo, inexiste

[37] HAND e HENLEY, op. cit
[38] CHINELATTO NETO, A.; FELICIO, R. S.; CAMPOS, D. Métodos de Monitoramento para o Gerenciamento de Modelo de *Credit Scoring*. **Tecnologia de Crédito, SERASA**, n. 61, 2007.

legislação brasileira específica voltada à concessão de crédito, porém, é possível que a evolução das instituições brasileiras leve ao aparecimento de normas que regulem o processo de concessão de crédito.

Assim, é interessante analisar se há alguma variável que, sozinha, responda pela concessão do crédito. Com este intuito, aplicaremos o mesmo teste realizado por Vasconcellos (2002), denominado Estabilidade do Modelo, no qual é calculado o *score* mínimo que uma operação pode obter se tiver as piores características nas variáveis explicativas, exceto uma, na qual estaria em seu melhor nível. Isso resultaria no que se chama de *score* mínimo no melhor nível da variável. De forma análoga, é calculado o maior *score* de uma operação com as melhores características das varáveis, exceto uma, a qual estaria no pior nível, resultando no *score* máximo no pior nível da variável. Para que o modelo não seja muito influenciado por uma única variável, o *score* máximo no pior nível da variável e o *score* mínimo no melhor nível não devem ser muito rebaixados e nem elevados devido a, exclusivamente, uma única variável.

O mesmo teste aplicado ao modelo de regressão logística será aplicado ao Modelo Cox, onde a probabilidade de sobrevivência de um contrato não deve sofrer grande variação. A tabela 24 sintetiza o resultado dos exercícios aplicados, considerando um horizonte de tempo de 12 meses:

TABELA 24 - Teste de Estabilidade do Modelo

Variável	Regressão Logística		Modelo Cox para t = 12 meses	
	***Score* máximo no pior nível**	***Score* mínimo no melhor nível**	**Função de Sobrevivência mais elevada no pior nível**	**Função de Sobrevivência mais baixa no melhor nível**
Tem plano de saúde	100	0	1	0
a1	100	0	1	0
a2	não significante	não significante	1	0
a3	não significante	não significante	1	0
a7	não significante	não significante	1	0
a8	não significante	não significante	1	0
a9	100	0	1	0
Situação veículo	100	0	1	0
Tipo de imóvel	100	0	1	0
a12	100	0	1	0

Ano veículo	100	0	1	0
Grau de instrução	100	0	1	0
Idade	100	0	1	0
Sexo	não significante	não significante	1	0
Estado civil	100	0	1	0
UF	100	0	1	0
b2	não significante	não significante	1	0
b3	100	0	1	0
c1	100	0	1	0
c2	não significante	não significante	1	0
Renda líquida formal	100	0	1	0
c3	100	0	1	0
c4	100	0	1	0
Renda informal	100	0	1	0
Saldo de aplicações em outras instituições	100	0	1	0
d4	100	0	1	0
Valor médio fatura mensal	não significante	não significante	1	0
Qtde anos associado cartão	100	0	1	0
d5	não significante	não significante	1	0
d6	100	0	1	0
d7	não significante	não significante	1	0
e2	100	0	1	0
e3	100	0	1	0
Valor médio cheque devolvido mot. 11	100	0	1	0
e4	não significante	0	1	0
Qtde dias abertura conta	100	0	1	0

De acordo com a tabela, percebe-se que o modelo é extremamente estável, não sendo influenciado exclusivamente por nenhuma variável: o pior *score* gerado foi 0 e o melhor foi 100, enquanto a função de sobrevivência mais elevada foi 1 e a mais baixa foi 0. Ainda que se atribua o melhor ou pior nível de qualquer variável, tanto o *score* como a probabilidade de sobrevivência não sofrem alterações, implicando que nenhuma variável tem poder o suficiente para aprovar ou reprovar uma operação de crédito. Caso os modelos apresentem problemas relacionados à instabilidade das variáveis, é necessária a reformulação destes, desde o processo de categorização das variáveis até a etapa de estimação, de forma a gerar modelos mais estáveis.

5. Conclusão

Nesta dissertação foram abordados estudos sobre modelos de *credit scoring* e de análise de sobrevivência, utilizados na classificação de risco de operações de crédito e para estimação do tempo esperado até a inadimplência do empréstimo. As técnicas utilizadas foram a regressão logística e o modelo Cox, as quais se mostraram capazes de classificar corretamente a maioria dos contratos e de estimar a expectativa de sobrevivência das operações de uma carteira de crédito do tipo rotativo.

A dificuldade inicial na formulação de estudos sobre *credit scoring* e análise de sobrevivência consiste na obtenção de um banco de dados com as informações necessárias a esta finalidade. A instituição financeira que cedeu os dados somente concordou em fazê-lo mediante a codificação de grande parte das variáveis explicativas utilizadas nos modelos, a fim de resguardar a confidencialidade de seus critérios de aprovação de crédito.

A base de dados utilizada no presente estudo continha 171.461 observações de operações de créditos contratadas entre Janeiro e Julho de 2005, das quais 79.287 foram selecionadas aleatoriamente para composição da base de modelagem, enquanto as demais observações foram reservadas para testes.

A definição da qualidade de crédito foi realizada pelo critério de inadimplência, que analisa o comportamento de pagamento a partir dos atrasos apresentados nas operações da carteira de crédito. No modelo de regressão logística foram considerados como créditos ruins aqueles com 31 ou mais dias de atraso, sendo os créditos bons aqueles com menos de 30 dias. Para o modelo Cox foi computada a quantidade de meses até a inadimplência de cada operação de crédito como variável de resposta. As contratações classificadas como boas e que, portanto, não incorreram em inadimplência, foram classificadas como censuradas. Todo o estudo estatístico foi, então, elaborado com os objetivos de classificar operações de crédito de acordo com suas chances de apresentarem inadimplência e de estimar o tempo de sobrevivência de cada uma destas operações. Cabe ressaltar que não foram levados em consideração os critérios de lucratividade comumente utilizados em estudos de *profit scoring*, uma vez que os dados necessários para esta análise não foram disponibilizados pela instituição financeira. Recomenda-se que sejam realizados estudos futuros com esta

abordagem que, conforme discutido, oferecem uma ótica alternativa a modelos baseados somente no critério de inadimplência.

As variáveis obtidas para a realização deste trabalho foram coletadas pela instituição financeira por meio do preenchimento de fichas cadastrais pelos potenciais tomadores de crédito, as quais foram codificadas como sendo de perfil ou comportamental, possuindo outras subdivisões de acordo com o tipo. Algumas variáveis importantes não foram disponibilizadas, tais como apontamentos negativos (informações sobre agências de proteção ao crédito como SPC e SERASA), informações comportamentais como valor médio de utilização do cheque especial, saldo médio conta-corrente, poupança e demais aplicações. A ausência destas informações não prejudicou os modelos elaborados, mas recomenda-se sempre que possível a utilização das mesmas.

Todas as variáveis disponibilizadas foram individualmente categorizadas pelo método CHAID, que detectou grupos homogêneos em relação à qualidade do crédito. Das 60 variáveis originalmente obtidas, 9 foram excluídas por apresentarem somente uma categoria. Outras 6 também foram excluídas por possuírem alto grau de correlação com outras variáveis explicativas, restando 45 potenciais variáveis preditoras aos modelos, as quais foram representadas por variáveis *dummy*.

O método para seleção das variáveis explicativas aplicado foi o *foward stepwise*, resultando na seleção de 25 variáveis para o modelo de regressão logística e 35 para o modelo Cox, com representantes de todos os tipos: variáveis patrimoniais, cadastrais, financeira de renda e informações bancárias, além de apontamentos negativos. Os coeficientes estimados para as categorias das variáveis se mostraram estatisticamente significantes, assim como a capacidade preditiva dos modelos, com taxas de acerto aceitáveis a ambos os modelos abordados. Também foram aplicados testes de eficiência (*Goodness of Fit*) baseados na Curva ROC e estatística Kolgomorov-Smirnov para o modelo de regressão logística e minimização dos erros tipo I e II ao Modelo Cox, cujos resultados foram considerados satisfatórios.

Os gráficos dos *scores* estimados de clientes bons e ruins, gerados a partir das 171.461 observações, mostraram que houve uma maior concentração de clientes bons nos *scores* mais altos e de clientes ruins nos *scores* mais baixos. Também foi possível constatar que não houve flutuações substanciais na comparação dos resultados obtidos na amostra de

modelagem e no banco de dados total, indicando que a amostra selecionada não estava viesada em relação à totalidade dos dados. As curvas das funções de sobrevivência e de risco para o modelo Cox também mostraram resultados bastante coerentes: contratos bons possuem curvas de sobrevivência mais elevadas e curvas de risco menos inclinadas do que os contratos ruins. Estes resultados vão ao encontro do que foi constatado pela regressão logística: contratos ruins possuem menores chances de sobrevivência, maior risco e probabilidade de inadimplência do que os contratos classificados como bons.

As curvas de sobrevivência permitem observar que, em média, a chance de sobrevivência para o horizonte de tempo de 12 meses é de 52,5% para os contratos ruins e 73,3% para os contratos bons. Já para o horizonte de 24 meses, a probabilidade de sobrevivência é de 32% para os contratos ruins e 58% para os bons. Um estudo de estabilidade dos modelos foi realizado para verificar se alguma das variáveis significantes seria capaz de, sozinha, determinar a qualidade ou a expectativa de vida da operação de crédito, o que não é desejável. Os modelos elaborados mostram-se estáveis.

Após a implantação dos modelos na prática, seguindo os preceitos do Novo Acordo de Basiléia, devem ser criados mecanismos periódicos de acompanhamento do desempenho dos modelos ao longo do tempo, a fim de verificar quando o mesmo não é mais apropriado e necessita ser reformulado. Neste ínterim foram abordados relatórios de acompanhamento fundamentais à melhoria do processo de tomada de decisão de crédito.

Em suma, os modelos de regressão logística e Cox apresentados neste trabalho mostraram-se capazes de classificar corretamente grande parte dos contratos de empréstimos de determinada instituição financeira, identificando antecipadamente o tempo esperado até inadimplência das operações de crédito. Por esta razão, acredita-se que estes modelos possam ser utilizados como instrumento de *early warning* na identificação de futuros problemas de insolvência. Assim, espera-se contribuir de forma relevante na redução de custos operacionais atrelados às altas taxas de inadimplência, proporcionando sustentabilidade financeira à instituição, além da conseqüente redução dos juros bancários ofertados à população.

ANEXO A – TRATAMENTO DA BASE DE DADOS

Os dados cedidos para este estudo foram agrupados em 53 arquivos, de acordo com o tipo de variável. Por exemplo: o arquivo de dados pessoais possui informações do tipo cadastral (idade, escolaridade, sexo, etc.), enquanto o arquivo de veículos possui informações afetas ao patrimônio de bens móveis do cliente (automóvel, motocicleta, embarcação, valor de cada um dos bens, etc.). Cada bem móvel, imóvel, renda ou qualquer outra variável que o cliente possui é representada por um registro no arquivo ou tabela correspondente. Se o cliente possui 2 veículos, a tabela de veículos receberá 2 registros para aquele indivíduo. Desta forma, uma tabela pode conter vários registros para um único cliente, caso este possua mais de um registro para aquela variável. Ocorre que para se gerar um arquivo no formato apropriado para executar a regressão, é necessário "cruzar" todas as tabelas, relacionando-as para que seja gerado um único arquivo. Este arquivo é o que será utilizado no cômputo da regressão.

Contudo, antes de cruzar os arquivos, faz-se necessário eliminar as repetições em cada uma das tabelas, sumarizando-as. Este procedimento é feito de acordo com o tipo de variável. Vejamos o exemplo para o caso dos veículos: a tabela abaixo possui registros repetidos e não sumarizados para os clientes A e B:

TABELA 25 - Registros não sumarizados para veículo

Cód. identificador cliente	Cód. veículo	Valor (R$)
A	1	40.000,00
A	2	15.000,00
A	3	20.000,00
B	1	8.000,00
B	2	10.000,00
C	1	32.000,00

Para eliminar os registros repetidos para o mesmo cliente, a tabela foi sumarizada, criando-se as variáveis "qtde veículos" e "valor total", assim:

TABELA 26 - Registros sumarizados para veículo

Cód. identificador cliente	Qtde veículos	Valor total (R$)
A	3	75.000,00
B	2	18.000,00
C	1	32.000,00

Desta forma, todas as tabelas que compõem a base de dados foram tratadas, tornando possível relacioná-las para obtenção do arquivo final a ser utilizado no software estatístico.

ANEXO B – ÁRVORES DE CLASSIFICAÇÃO PARA ANÁLISE DESCRITIVA

FIGURA 3 – Qualidade de crédito por possui plano de saúde

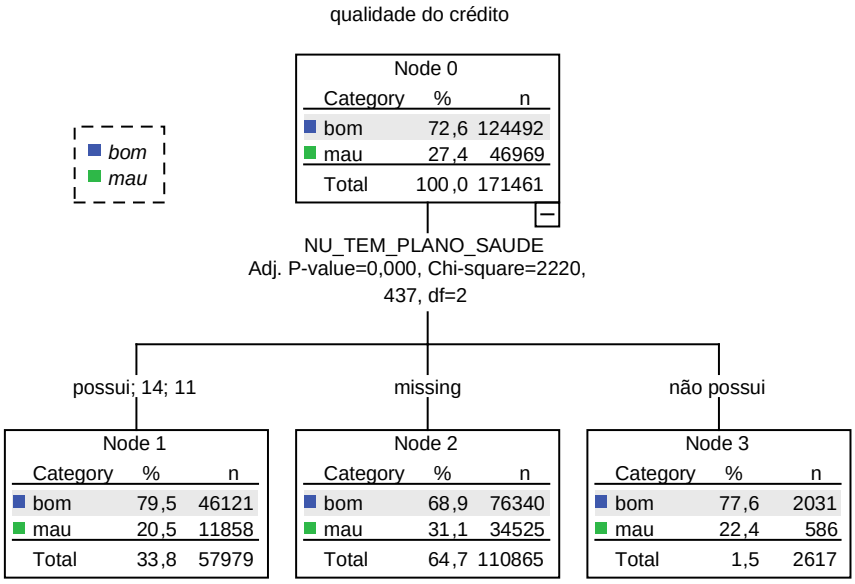


FIGURA 4 – Qualidade de crédito por situação veículo

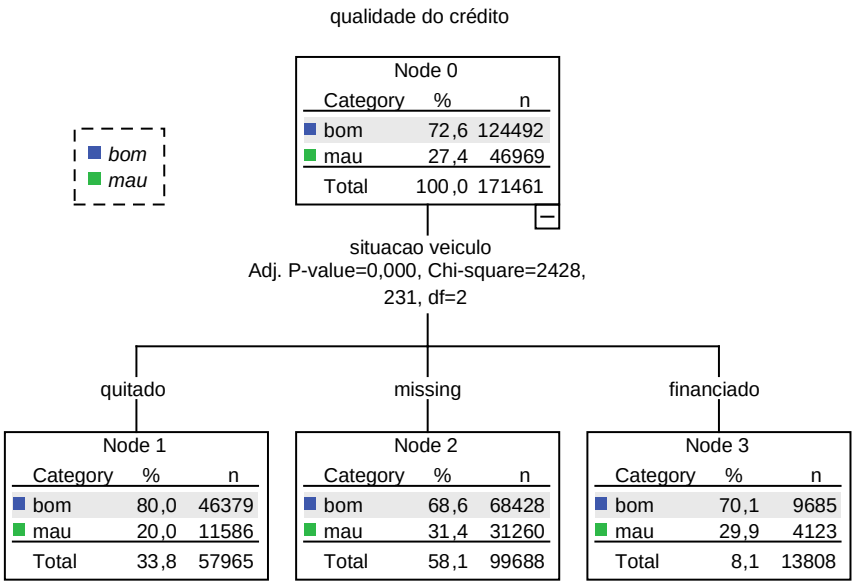


FIGURA 5 – Qualidade de crédito por tipo de imóvel

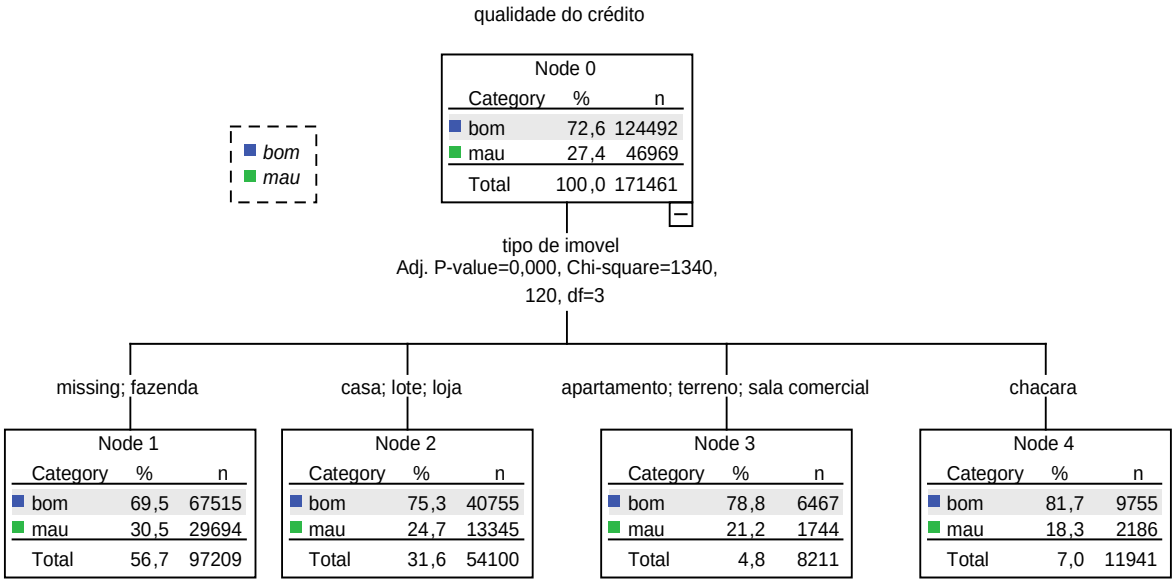


FIGURA 6 – Qualidade de crédito por ano do veículo

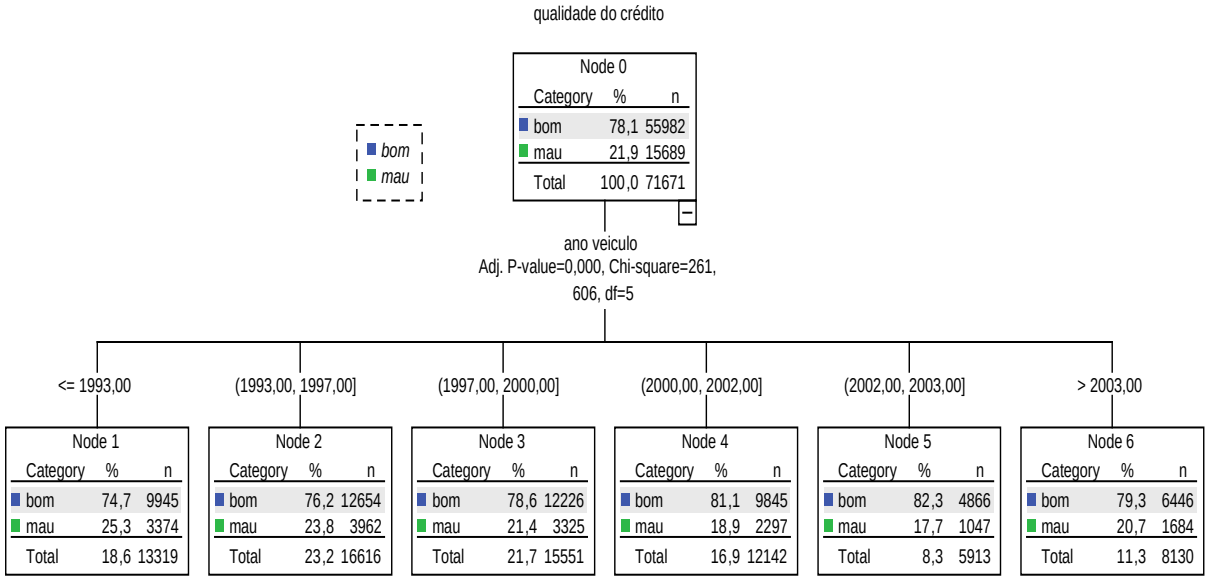


FIGURA 7 – Qualidade de crédito por valor de mercado do veículo

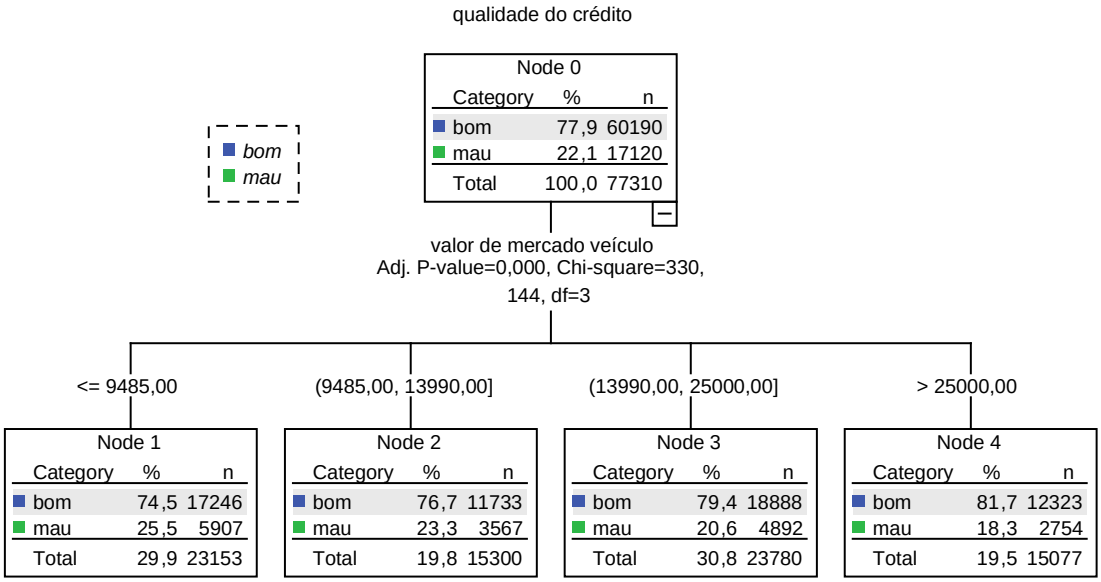


FIGURA 8 – Qualidade de crédito por sexo

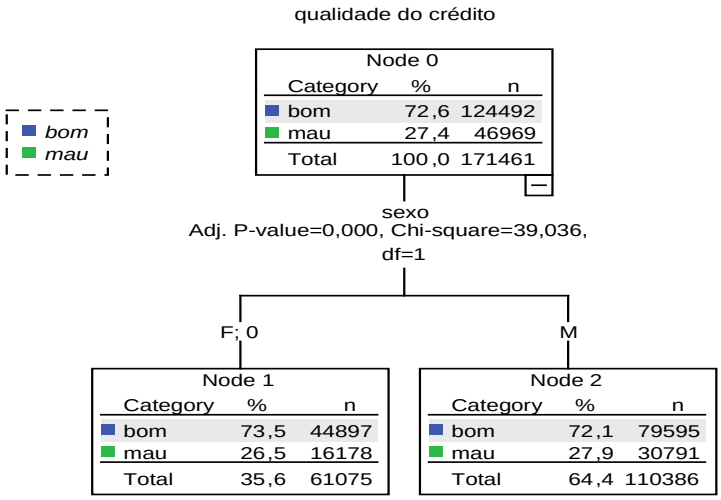


FIGURA 9 – Qualidade de crédito por grau de instrução

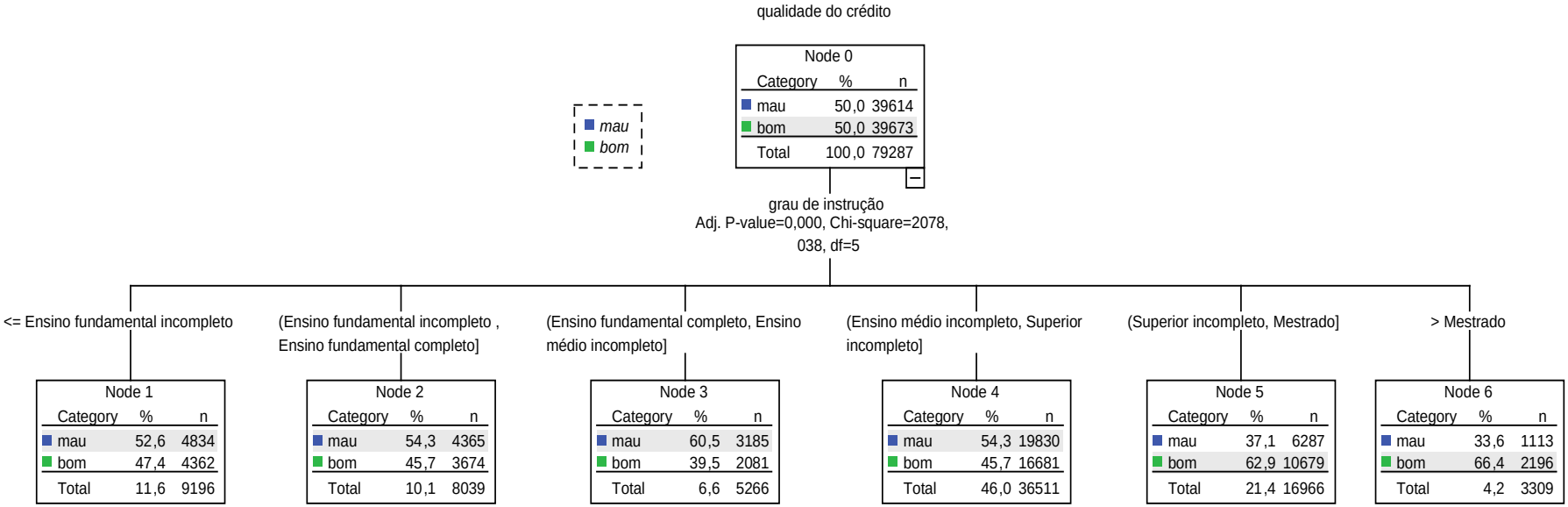


FIGURA 10 – Qualidade de crédito por idade

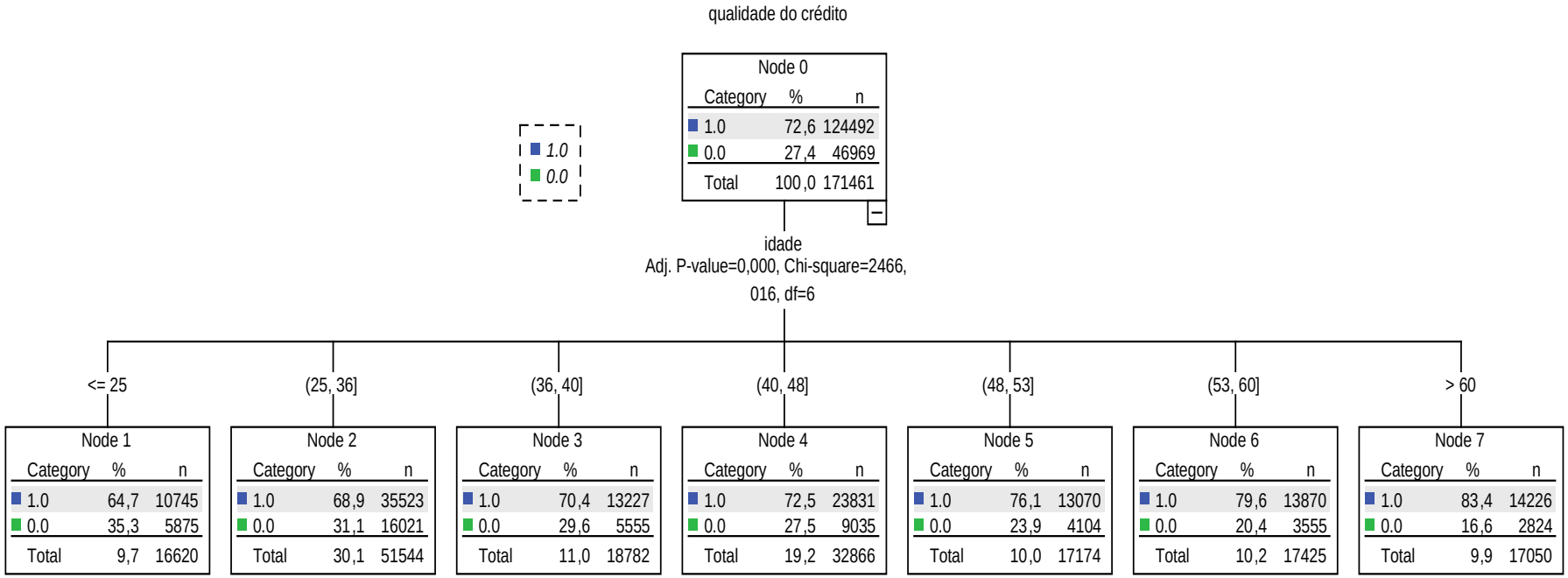


FIGURA 11 – Qualidade de crédito por estado civil

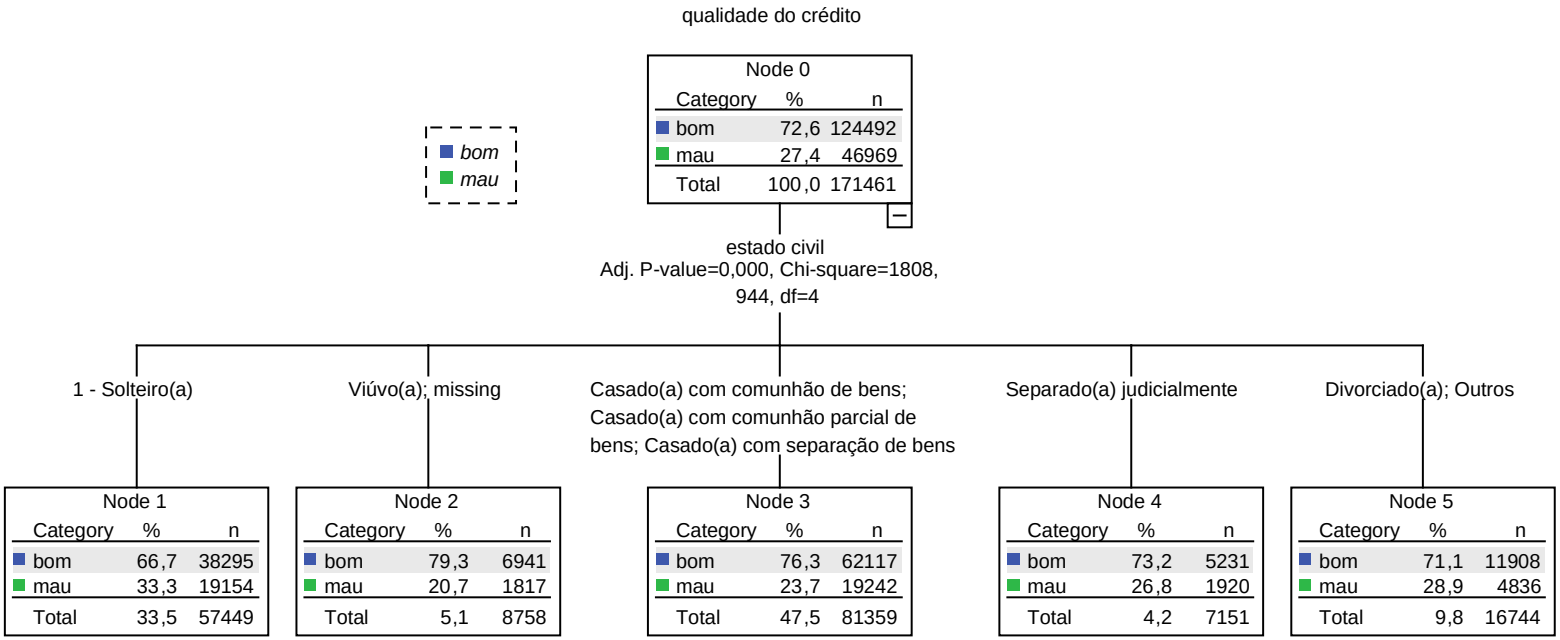


FIGURA 12 – Qualidade de crédito por UF

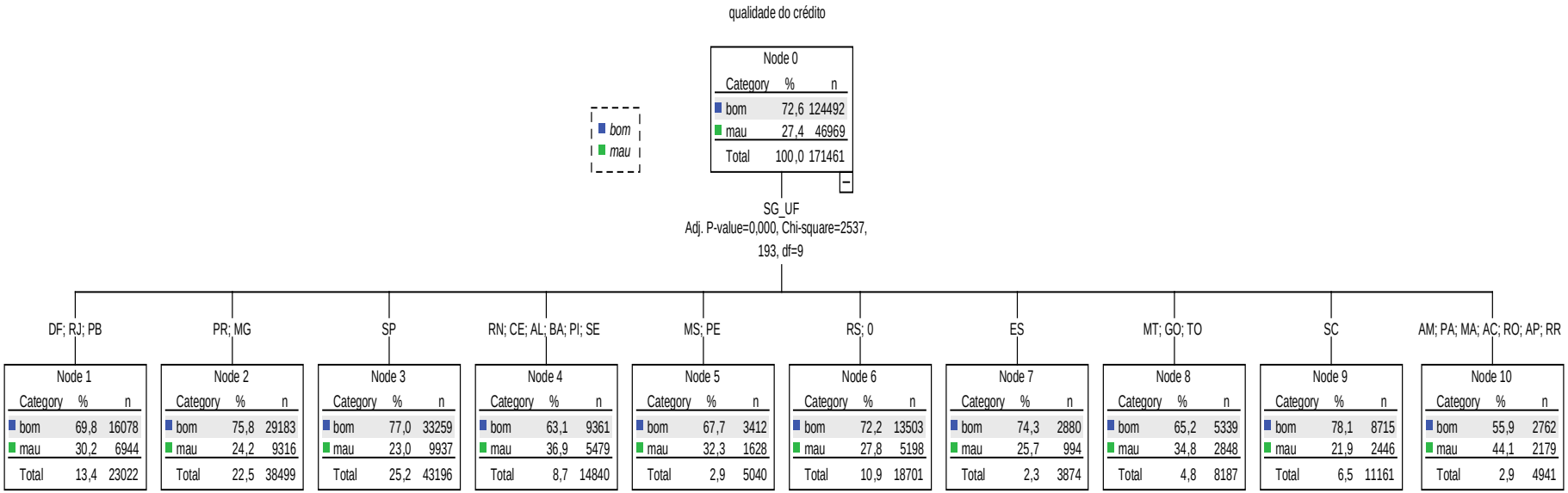


FIGURA 13 – Qualidade de crédito por renda líquida formal

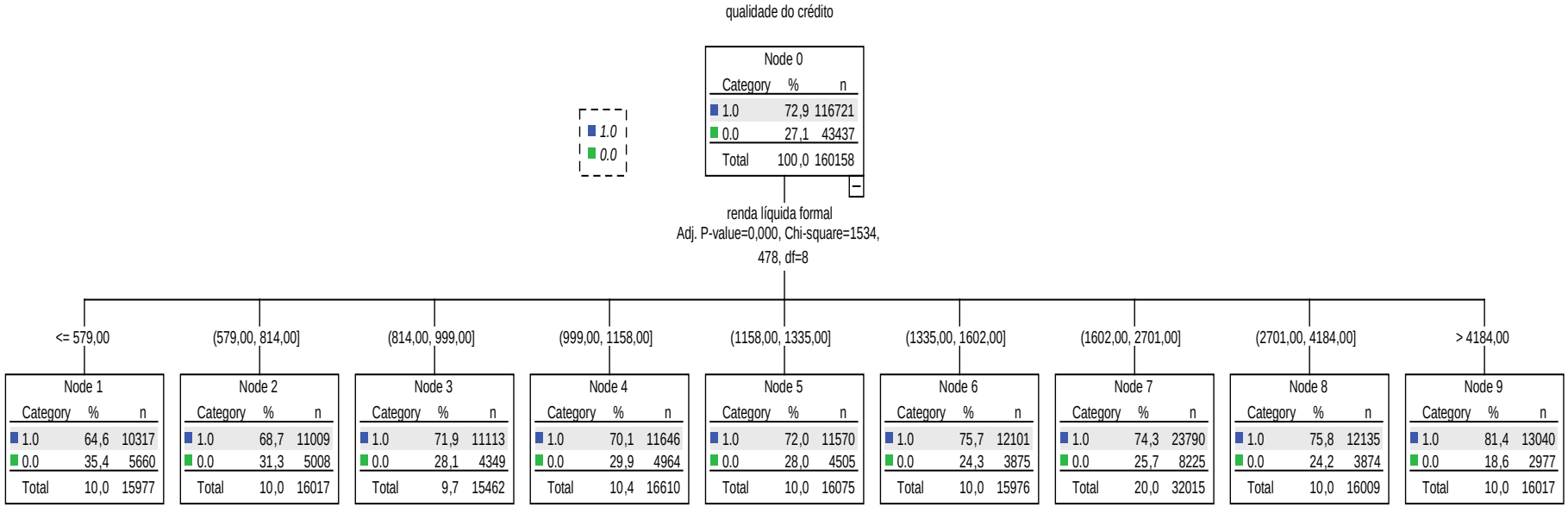


FIGURA 14 – Qualidade de crédito por renda informal

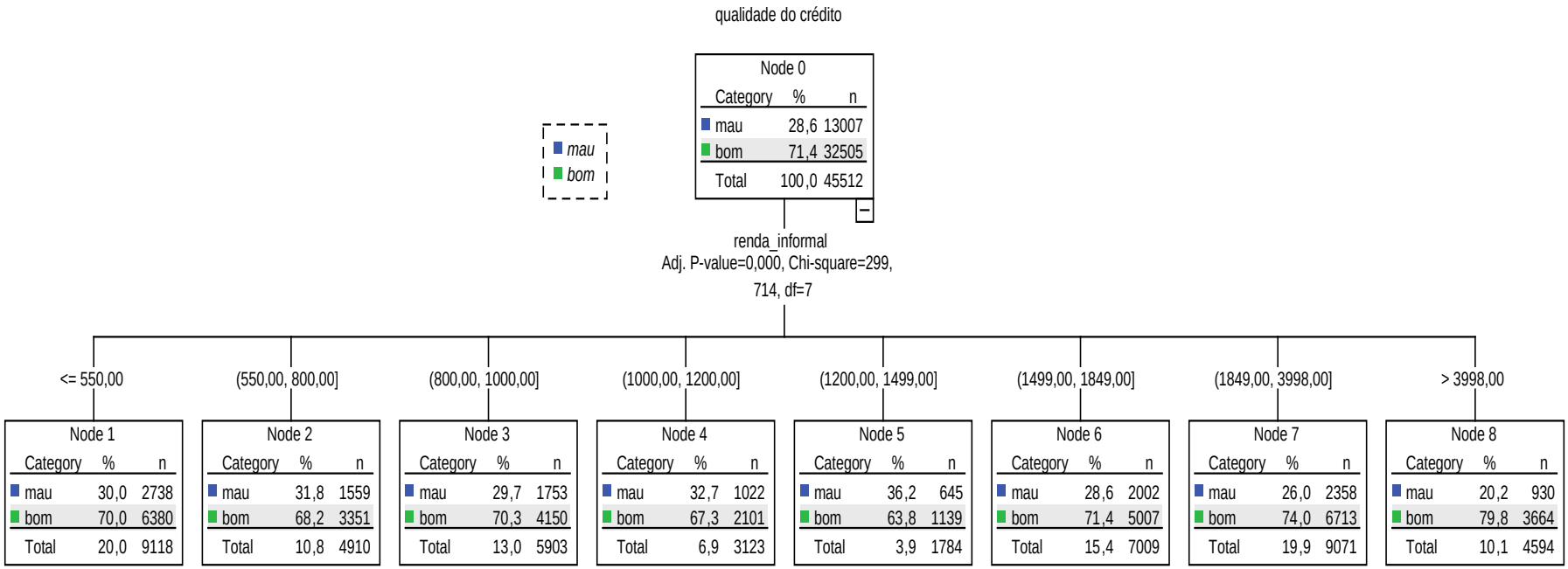


FIGURA 15 – Qualidade de crédito por Saldo médio de aplicações em outras instituições

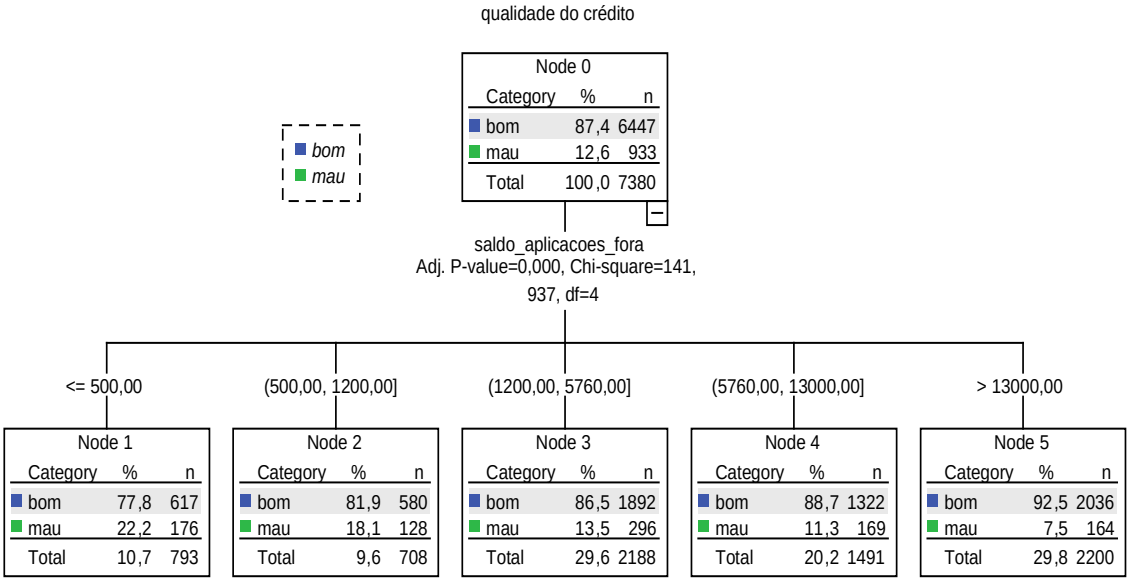


FIGURA 16 – Qualidade do crédito por Valor médio fatura cartão de crédito

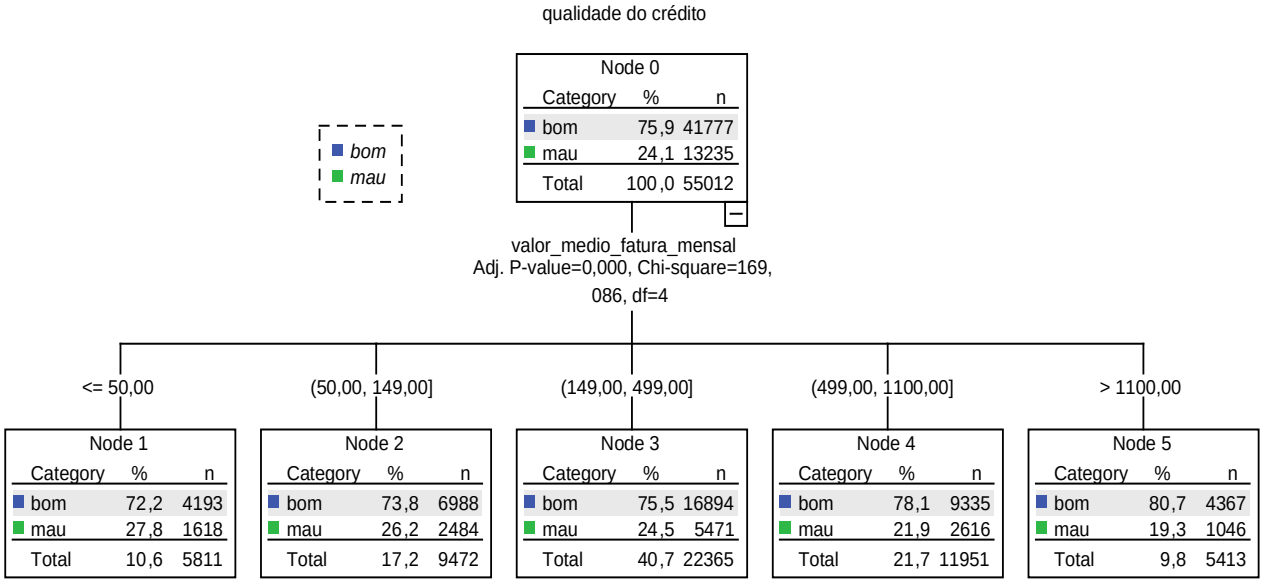


FIGURA 17 – Qualidade do crédito por Qtde anos associado cartão de crédito

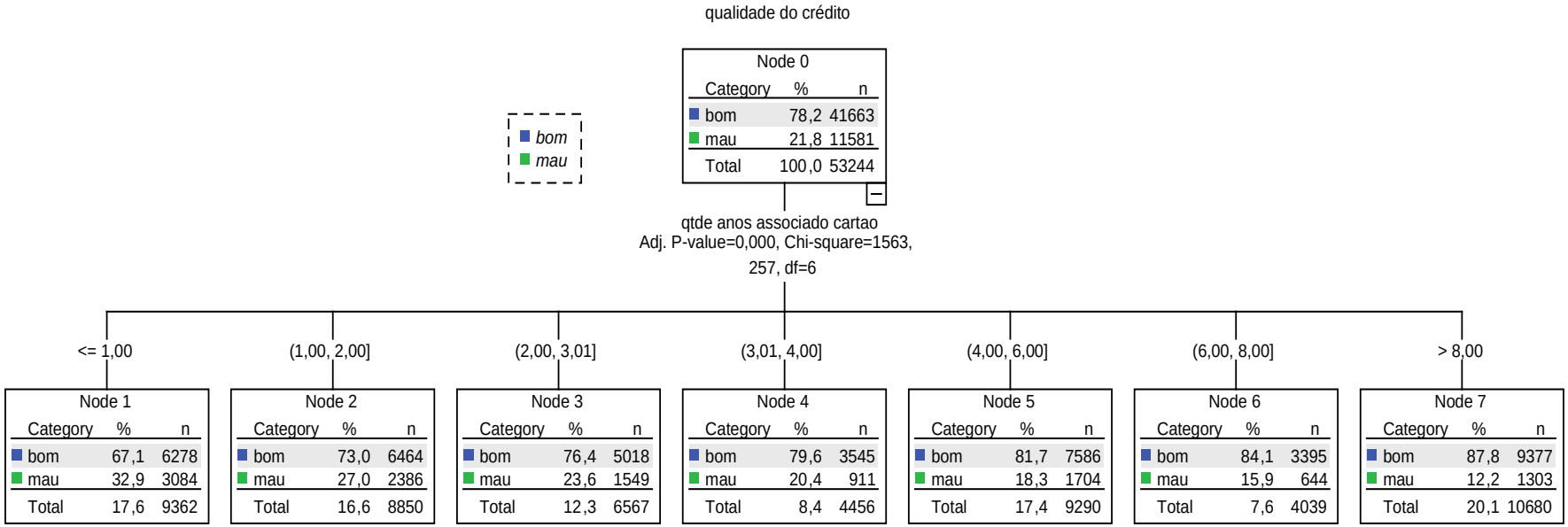
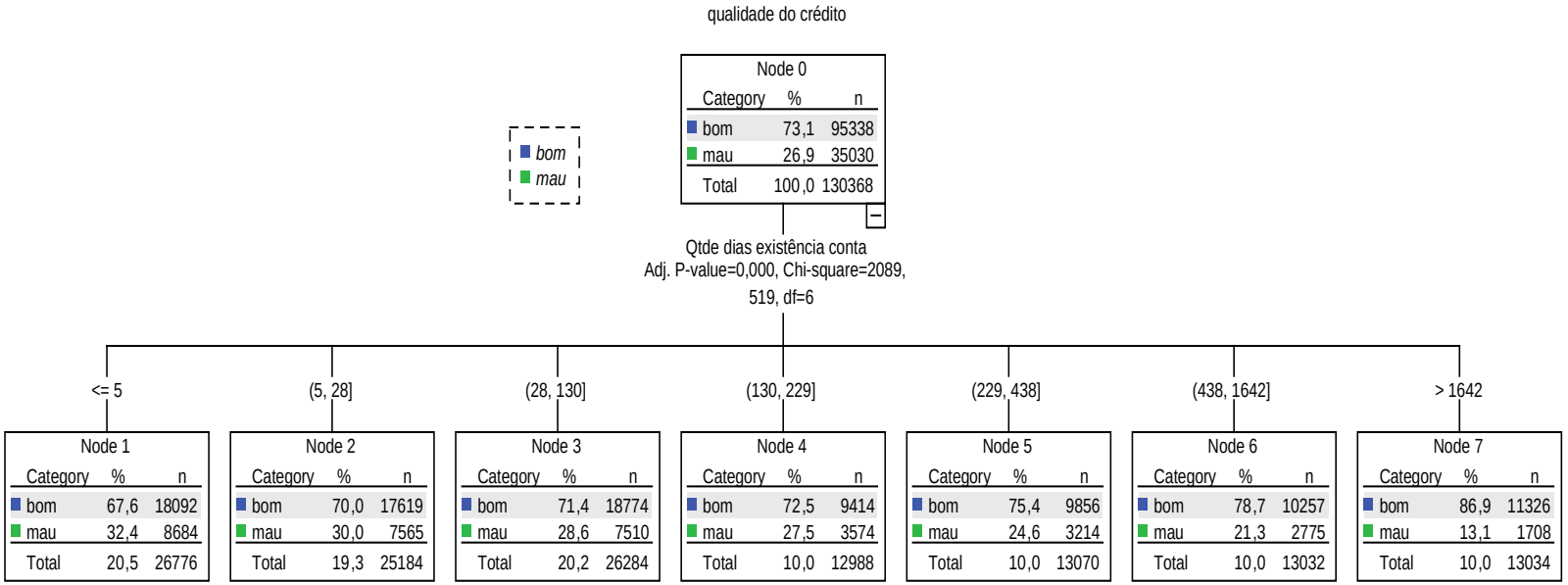


FIGURA 18 – Qualidade de crédito por qtde dias abertura conta



ANEXO C - FUNÇÃO DE RISCO E SOBREVIVÊNCIA BASELINE

Tempo *(t)*	Função Risco Baseline: $h_0(t)$	Função de Sobrevivência Baseline: $S_0(t) = e^{-h_0(t)}$
1	0,004007361	0,996000658
2	0,016850091	0,983291078
3	0,030857025	0,969614194
4	0,04660958	0,954459965
5	0,062734833	0,939192483
6	0,07828329	0,92470243
7	0,093959721	0,910319428
8	0,10858515	0,897102504
9	0,122029957	0,885121853
10	0,136861857	0,872090691
11	0,15120181	0,85967419
12	0,164482656	0,84833247
13	0,182409157	0,833260337
14	0,198094599	0,820292251
15	0,210635146	0,81006957
16	0,224103763	0,799232199
17	0,23691572	0,789057787
18	0,247886045	0,780448875
19	0,262417662	0,769189695
20	0,274637572	0,759847463
21	0,285445767	0,751679106
22	0,297873498	0,742395248
23	0,306548316	0,735982958
24	0,313449942	0,730920966

REFERÊNCIAS BIBLIOGRÁFICAS

ADVANCED Statistical Analysis Using SPSS. V14.0, pág. 3-11.

ADYA, M., COLOPPY, F. How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation. **Journal of Forecasting**, 1998.

ALBRIGHT H. T. **Construction of polynomial classifier for consumer loan application using genetic algorithms**. Virginia: Working Paper, Department of System Engineering. University, 1994.

ALTMAN E. I. Financial ration, discriminant analysis and the prediction of corporate bankruptcy. **Journal of Finance**, p. 589-529, 1994.

BRASIL. Banco Central do Brasil. **Avaliação de 3 anos do projeto Juros e *Spread* bancário**. Brasília, 2002. Disponível em <http://www.bcb.gov.br/ftp/juros-spread1.pdf >. Acesso em: 11 nov. 2007.

_____. **Economia Bancária e Crédito – Avaliação de 5 anos do Projeto Juros e Spread Bancário**. Brasília, 2004. Disponível em: <http://www.bcb.gov.br/Pec/spread/port/economia_bancaria_e_credito.pdf.> Acesso em: 11 nov. 2007.

BARTH, N. **Método de discriminação entre grupos – aplicação aos problemas de concessão de crédito**. São Paulo: Dissertação de Mestrado – FGV, 2002.

BOYLE, M.; CROOK, J.N.; HAMILTON, R.; THOMAS, L.C. **Methods for credit scoring applied to slow payers**. Oxford: Oxford University Press, p. 75-90, 1992.

CARVALHO, B. C.; DOS SANTOS, G. M. **Os Acordos de Basiléia – Um roteiro para implementação nas instituições financeiras**. Disponível em <http://www.febraban.org.br/Arquivo/Servicos/Imprensa/Artigo_Basileia_6.pdf>. Acesso em 01 out. 2008

CAPON N. Credit Scoring Systems: a critical analysis. **Journal of Marketing**, 46, 82-91,1982.

CHARNET, R.; BONVINO, H.; DE LUNA, C. A. **Análise de modelos de regressão linear com aplicações**. Campinas: Editora da UNICAMP, 1999.

CHINELATTO Neto, A.; FELICIO, R. S.; CAMPOS, D. Métodos de Monitoramento para o Gerenciamento de Modelo de *Credit Scoring*. **Tecnologia de Crédito, SERASA**, n. 61, 2007.

CORTES, F. P. **Gestão de Risco nas Instituições Financeiras: Uma Análise do Novo Acordo de Basiléia e Apresentação de Conceitos para Desenvolvimento de um Sistema de Informações Gerenciais**. Brasília: Fundação Getúlio Vargas, 2004.

DESAI,V.S; CONVAY,D.G.;CROOK, J.N.;OVERSTREET,G.A. **Credit scoring models in the credit union environment using neural networks and genetic algorithms**. Mathematics applied in Business and Industry, p. 323-346, 1997.

DRISLANE, R. e PARKINSON, G. **On line Dictionary of Social Sciences**. Canada's Open University. Disponível em: <http://datadump.icaap.org/cgibin/glossary/SocialDict/SocialDict?term=LORENZ%20CURVE>. Acesso em: 09 ago. 2007.

DURAND, D. Risk Elements in Consumer Installment Financing. **National Bureau of Economic Research**, 1941.

EISENBEIS R.A. Pitfalls in the application of discriminat Analysis in business, finance and economics. **Journal of Finance** 32, 975-900,1977.

EISENBEIS R.A. Problems in applying discriminat analysis in credit scoring models. **Journal of Banking and Finance** 2, p. 205-219,1978.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics,** 1936.

GAZOLA, Sebastião. **Construção de um modelo de regressão para avaliação de imóveis**. Florianópolis: Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2002.

GEHRLEIN, W.V.; WAGNER, B.J. A two-stage least cost credit scoring model. **Annals of Operations Research** 74, p.159-171, 1997.

HAIR Jr. J. F. **Multivariate Data Analysis**. Edition nº 5, Pretince Hall, USA, 1998.

HAND, D. J. and HENLEY D. J. Statistical Classifications Methods in Consumer Credit Scoring**:** a Review**. Journal of the Royal Statistical Society Series**, 1997.

HAND D.J. e THOMAS L.C. A survey of the issues in consumer credit modeling research. **Journal of the Operational Research Society**, 2002.

HARDY, W.E.; ADRIAN, J.L. **A linear programming alternative to discriminant analysis**. Abribus1, p. 285-292, 1985.

HENLEY, W.E. **Statistical aspects of Credit Scoring**. Ph.D. thesis, Open University, 1995.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. Massachusetts: Wiley Publication, 1989.

_____. **Applied Survival Analysis**. Epidemiology of University of Massachusetts, 1999

JANOT, Márcio M. **Previsão de insolvência bancária no Brasil: aplicação de diferentes modelos entre 1995 e 1998**. Rio de Janeiro: Dissertação de Mestrado, Departamento de Economia. PUC-RJ, 1999.

KARAKOULAS, Grigoris. **Validação empírica de modelos de *credit scoring***. Revista SERASA. Disponível em: <http://www.serasa.com.br/ingles/i_revista/i_revista1.htm>. Acesso em: 01 set. 2008.

LOPES, Fernanda **Mulheres negras e não negras vivendo com HIV/AIDS no estado de São Paulo: um estudo sobre vulnerabilidades**. São Paulo: Tese de Doutorado. USP, 2003.

LUCUMBERRI, L. F. L; DUARTE JÚNIOR, A. M. Uma metodologia para o gerenciamento de Modelos de Escoragem em Operações de Crédito de Varejo no Brasil. **Revista de Economia Aplicada**, São Paulo: v.7, n.4, p. 795-818, 2003.

MANGASARIAN, O. L. **Linear and non-linear separation of patterns by linear programming. Operations Research** 13, p. 444-452, 1965.

MARTELL, T. F.; FITTS, R. L. A quadratic discriminant analysis of bank credit card user characteristics. **Journal of Economics and Business** 33, p. 153-159, 1981.

MARTINS, M. **A previsão de insolvência pelo Modelo Cox: uma contribuição para a análise das companhias abertas brasileiras**. Porto Alegre: Dissertação de Mestrado. UFRS, 2003.

MORRISON, J. S. **Preparativos para o Novo acordo da Basiléia**. Disponível em: <http://forecastingsolutions.com/publications/47.pdf>. Acesso em: 11 nov. 2007.

NAKANO, E. Y. e CARRASCO, C. G. Uma avaliação do Uso de um Modelo Contínuo na Análise de Dados Discretos de Sobrevivência. **Publicação da Sociedade Brasileira de Matemática Aplicada e Computacional**, 2006.

NATH R., JACKSON W.M., JONES T.W. A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis.
Statistical Computation and Simulation 41, p.73-93, 1992.

NAIRAN, B. **Survival Analysis and the credit granting decision**. In Credit Scoring and Credit Control. Oxford: Oxford University Press, pp. 109-122, 1992.

OLIVER R.M. e KEENEY R. L. Designing win-win financial loan products for consumer and businesses. **Journal of the Operational Research Society**. New York: 01 June 2005.

PATERSON, Dan W. **Artificial Neural Networks, Theory and Applications**. Pretince Hall Inc. USA, 1996.

PEREIRA, G. H. **Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais**. São Paulo: Dissertação de Mestrado – USP, 2004.

PINHEIRO, A. C.; CABRAL , C. Mercado de Crédito no Brasil: o papel do judiciário e de outras instituições. **Ensaios BNDES**, Rio de Janeiro: n.9, 1998.

ROCHA, Fabiana. **Previsão de falência bancária: um modelo de risco proporcional**. Disponível em: <http://www.ppe.ipea.gov.br/index.php/ppe/article/viewFile/194/128> Rio de Janeiro, 1999. Acesso em: 01 set. 08.

SANDRONI, Paulo **Novíssimo Dicionário de Economia**. São Paulo: Editora Best Seller, 2002.

SCARPEL, R.A. e MILIONI, A.Z. **Utilização Conjunta de Modelagem Econométrica e Otimização em Decisões de Concessão de Crédito**. Disponível em: <http://www.scielo.br/pdf/pope/v22n1/a04v22n1.pdf>. Acesso em: 01 dez. 2007.

SILVA, Edson R. **Aplicação de Metodologia de dados de painel em modelos de *behaviour score* do varejo**. São Paulo: Dissertação de Mestrado – IBMEC, 2006.

STEPANOVA, M; BAESENS B.; Van GESTEL, T.; Den POEL and VANTHIENEN, D. Neural network survival analysis for personal loan data. **Journal of Operational Research Society,** 2005.

SRINIVASAN,V.; KIM,Y.H. Credit granting: a comparative analysis of classification procedures. **Journal of Finance**, 42, p. 665-683, 1987.

THOMAS L. C. A survey of credit and behavior scoring: forecasting financial risk of lending to consumers. **International Journal of Forecasting**. 16, p. 149-172, 2000.

THOMAS, L. C., EDELMAN D. B. and CROOK, J. N. ***Credit Score* and its application**, Philadelphia, 2002.

THOMAS, L. C.; OLIVER, R. W.; HAND, D. J. A survey of the issues in consumer credit modeling research. **Journal of the Operation Research Society**, 2005.

VARETTO, Franco. Genetic Algorithms Applications in the Analysis of Insolvency Risk, **Journal of Banking and Finance**, 22, 1998.

VASCONCELLOS, Maurício. **Proposta de Método para análise de concessões de crédito a pessoas físicas**. São Paulo: Tese de Mestrado – USP, 2002.

WOOLDRIDGE, Jeffrey. "**Introdução à Econometria: Uma abordagem moderna**" Michigan State University. Tradução: Rogério Cezar de Souza e José Antônio Ferreira. São Paulo: Editora Thomson, 2006.

YOBAS, M.B; CROOK, J.N.; ROSS, P. Credit scoring using neural and evolutionary techniques, **Working Paper 97/2, Credit Research Centre**, University of Edinburgh, 1997.

ZENDERSKY, H. C.; GULIAS Jr. S.; SILVA E. **Reflexos do Novo Acordo de Capital da Basiléia e do modelo de requerimento de capital do Sistema Financeiro Nacional sobre as operações de crédito**. Disponível em <http://www.febraban.org.br/Arquivo/Servicos/Biblioteca/ShowBib.asp?codassunto=548&mesano.> Acesso em: 01 out. 2008.

ZERBINI, M. B. do A. A. **Três ensaios sobre o crédito**. São Paulo: Tese de Doutorado. FEA-USP, 2000.