UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
Departamento de Economia

TESE DE MESTRADO

"PROPOSTA DE MÉTODO PARA ANÁLISE DE CONCESSÕES DE CRÉDITO A PESSOAS FÍSICAS"

Maurício Sandoval de Vasconcellos

Dissertação apresentado à Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo para obtenção do título de Mestre em Economia

Orientadora: Prof[a]. Dr[a]. Vera Lúcia Fava

São Paulo

2002

Aos meus pais Marco Antonio e Ana pelos esforços dedicados à educação, saúde, felicidade e caráter dos seus filhos. Muito obrigado.

AGRADECIMENTOS

A elaboração dessa tese de mestrado somente foi possível graças à contribuição direta ou indireta de pessoas que me acompanharam durante os últimos anos e que, cientes ou não do apoio que me deram no âmbito pessoal e profissional, permitiram que eu concretizasse mais um sonho entre tantos outros que vislumbrei e lutei para realizar.

É de extrema importância a gratidão que tenho pela Prof[a]. Dr[a]. Vera Lúcia Fava pelos ensinamentos econométricos que me apresentou desde os tempos de graduação, o que fez com que eu criasse curiosidade intelectual e vontade própria de me dedicar e aprofundar nos estudos estatísticos aplicados na prática.

Agradeço a Marcelo Rabbat, João Carlos Prandini e Carla Vital Otero pelo apoio vital que demonstraram na fase final do meu período de mestrado, bem como pelos conhecimentos profissionais que compartilharam comigo ao longo dos últimos quatro anos e pelo respeito, críticas e reconhecimento que demonstram pelo meu trabalho.

Não poderia esquecer de agradecer à instituição financeira que forneceu os dados necessários para elaboração dessa tese e cujos funcionários fizeram diversas sugestões e observações mercadológicas, sem as quais a confiabilidade e veracidade do presente estudo certamente não seriam as mesmas.

Finalmente, gostaria de agradecer a todas às demais pessoas cujo apoio ou mesmo simples abraço também foram de vital importância para o bom andamento desse estudo, a saber: aos meus irmãos Alexandre, Ivan e Rodrigo, às queridas "mães" Maria e Isaura e aos amigos Rafael de A. Medawar, Sandro Raymundo, Rafael P. Rodrigues, Sílvia R. M. Rangel, Renata M. Rangel, Luís C. B. Rueda, Flávia C. de Lima e Rosângela Coelho.

SUMÁRIO

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE EXEMPLOS

RESUMO

Esse trabalho consiste em uma sugestão de metodologia para análise de concessões de crédito a pessoas físicas a partir do estudo matemático e estatístico de informações sobre créditos concedidos no passado recente da carteira de crédito em questão, tais como os hábitos de pagamentos e variáveis cadastrais, financeiras, patrimoniais e de relacionamento com a instituição credora dos clientes analisados. A análise parte da definição da qualidade de crédito (bom ou ruim) dos créditos estudados, sendo seguida pelo estudo das variáveis dos clientes que influenciam na capacidade destes em honrar os compromissos do crédito obtido, a partir de técnicas estatísticas de agrupamento de indivíduos em categorias homogêneas de influência sobre a qualidade de crédito e estimação de coeficientes para as variáveis relevantes estatisticamente, através do método de regressão logística, gerando um modelo de fácil interpretação e implementação. O estudo sugere, também, ferramentas práticas para a verificação da qualidade do modelo estimado e de acompanhamento da performance da utilização do modelo durante os meses seguintes à sua implementação. Assim, o trabalho apresenta todos os aspectos vitais para análise de concessões de crédito, a partir de um enfoque bastante pragmático e tecnicamente viável.

ABSTRACT

This paper comprises a suggested modus operandi to analyze the granting of credits to natural persons on the basis of the mathematical and statistical study of information and data on credits granted in the recent past through the pertinent credit portfolio, such as payment habits and variables involved in the business standing, assets, property and relationship with the institution who provided credit to the clients under analysis. The analysis starts with the definition of credit quality (good or bad) for the credits studied, followed by the study of variables on customers who affected the capacity to honor the granted credit commitments, on the basis of statistical techniques for grouping the individuals into homogeneous influence categories on the quality of the credit and the estimate of coefficients for the variables statistically pertinent, through the logistic regression method, generating a model whose interpretation and implementation is easy. The study suggests, as well, practical tools to verify the quality of the estimated model and for the follow-up of the model usage performance throughout the months following its implementation. Thus, the paper described all vital aspects for the analysis of credit granting, from a rather pragmatic and technically feasible approach.

Introdução

O presente estudo tem o objetivo de apresentar uma sugestão de modelo de *credit scoring* para análise de concessões de crédito, a partir do estudo estatístico de créditos concedidos no passado recente de uma linha de crédito de um banco popular brasileiro, discutindo os principais aspectos práticos e base estatística que compõem um modelo completo, porém simples e de fácil implementação. Além de revelar um método eficiente para a concessão de crédito, o estudo propõe, ainda, algumas ferramentas pragmáticas para acompanhamento e manutenção dos resultados e da qualidade do modelo após sua implementação, ou seja, o monitoramento da carteira de crédito.

O termo *credit scoring* é utilizado para descrever métodos estatísticos adotados para classificar candidatos à obtenção de um crédito qualquer em grupos de risco. A partir do histórico de concessões de crédito efetuadas por uma instituição provedora de crédito é possível, através de técnicas estatísticas, identificar as variáveis cadastrais e financeiras dos clientes e as variáveis da própria operação que influenciam na capacidade de cliente em pagar este crédito, ou seja, na qualidade de crédito do indivíduo. O modelo, baseado em informações do passado recente da carteira de crédito em questão, gera notas (*scores*) para novos candidatos ao crédito que espelham a expectativa de que esses paguem o crédito sem se tornar inadimplentes ou trazer prejuízo para o credor. Assim, o modelo é uma ferramenta valiosa para decisões de aprovação ou não de pedidos de crédito, devendo ser obedecida a hipótese de que o público alvo da carteira de crédito após a implementação do modelo se mantenha o mesmo que no passado recente sobre o qual todo o procedimento estatístico se baseia.

Quando um novo cliente solicitar um crédito, o mesmo deverá fornecer suas variáveis cadastrais e financeiras que, unidas às variáveis da operação, poderão lhe gerar um *score* de 0 a 100 pontos. Esse *score* poderá, então, ser utilizado na decisão de conceder ou não o crédito ao cliente, a partir do momento que se define um *score* de corte, acima do qual o pedido do cliente será aceito.

A aplicação de modelos de *credit scoring* e outras ferramentas para análises de empréstimos se iniciou nos países desenvolvidos em meados de 1960 e sua utilização por instituições financeiras e de crédito vem aumentando rapidamente desde então. No Brasil, o interesse por tais modelos começou a partir de 1994 com a estabilidade da inflação e a literatura que existe sobre o assunto se resume em artigos e reportagens de caráter introdutório ou mencionando técnicas estatísticas de análise discriminante. Também existem trabalhos empíricos comparando diferentes métodos para elaboração de modelos de *credit scoring*, tal como o estudo de Zerbini (2000), que aplica técnicas de análise discriminante, modelos *logit*, *probit* e de redes neurais e cujo resultado principal indica que as técnicas utilizadas geram resultados equivalentes.

Mesmo nos países desenvolvidos muito pouco é conhecido e divulgado sobre o conteúdo dos modelos. A maior razão para isso é a necessidade de sigilo, já que boas e sofisticadas técnicas trazem vantagem competitiva e, portanto, as instituições que as utilizam procuram não divulgá-las. Além disso, as bases de dados com as informações necessárias para a realização de qualquer estudo relacionado à concessão de crédito muitas vezes contêm informações confidenciais sobre os clientes que realizam as operações de créditos, e não podem ser liberadas a terceiros sem uma série de precauções. O que existe em abundância são discussões acerca dos problemas das metodologias estatísticas mas, de fato, dificilmente é encontrado algum estudo

empírico revelando todas as etapas do processo desde a formulação até a aplicação dos modelos, o que motivou a escolha do tema do presente estudo.

O trabalho desenvolvido ao longo desse documento é baseado na definição de crédito utilizada por Hand e Henley (1997), que assumem o termo "crédito" como sendo uma quantidade de dinheiro que é emprestada a um consumidor por uma instituição financeira e que deve ser amortizado em prestações, usualmente em intervalos regulares. Assim, o método de *credit scoring* desenvolvido ao longo desse trabalho somente pode ser aplicado a créditos requisitados para serem amortizados em prestações. Créditos de outra natureza, tais como cheque especial e descontos de cheques e duplicatas, cujas amortizações ocorrem de outra maneira que não através de pagamentos de prestações, também podem ter um modelo de *credit scoring* com as mesmas bases gerais, sendo necessárias apenas algumas modificações com relação ao estudo do comportamento de pagamentos dos clientes do passado recente da carteira.

O método enfocado é baseado na classificação de candidatos a crédito em grupos de acordo com seus prováveis comportamentos de pagamentos (por exemplo, "atraso" ou "não atraso" no pagamento das prestações, "bom" ou "mau" crédito etc.), mas também considera outros problemas pertinentes à indústria de crédito. A probabilidade de um candidato atrasar ou não o pagamento deve ser estimada com base nas informações que o candidato ao crédito fornecer na data do pedido de concessão, e a estimativa servirá como fundamento para a decisão de aprovação ou não do crédito. Um método de classificação eficiente beneficia tanto o credor, através do aumento do lucro ou redução de perdas na carteira de crédito, quanto o candidato ao crédito, por evitar que este assuma mais compromissos financeiros do que é capaz.

A atividade de decidir sobre conceder ou não um crédito é realizada por bancos, sociedades de crédito, empresas de construção civil, comércio varejista (principalmente o

popular), prestadores de serviços e diversas outras organizações. A concessão de crédito é uma atividade que vem apresentado rápido crescimento ao longo dos últimos anos nos mercados desenvolvidos. No Brasil, esse crescimento somente se manifestou com clareza nos últimos 8 anos, após o fim do longo período inflacionário terminado em 1994. O estudo de Pinheiro e Cabral (1998), baseado em informações do Banco Central do Brasil do período de 1988 a 1998 revelou que o mercado de crédito brasileiro é caracterizado por um tamanho (volume) relativamente baixo e altas taxas de juros e inadimplência. Os autores também comprovaram a diminuição do volume total de crédito enquanto percentagem do PIB desde o fim da inflação em 1994, mas como resultado da contração de crédito bancário ao setor público (principalmente ao governo federal) devido ao processo de privatizações que causou a diminuição dos créditos dos bancos às empresas estatais. No entanto, verificaram que os empréstimos ao setor privado no segmento de crédito ao consumidor têm mostrado os maiores índices de crescimento desde a implantação do Plano Real (1994), com a dramática redução da inflação e conseqüente reduções das receitas não provenientes de juros e da incerteza, que encorajaram e facilitaram uma forte expansão de linhas de crédito para o setor privado. No caso de crédito às famílias brasileiras, por exemplo, o crédito ao consumo subiu de 2,4% do total de empréstimos nos anos de 1988 a 1993 para 8,4% em 1994 e crescimento constante desde então, atingindo 13% em 1997, de acordo com os dados do Banco Central do Brasil. Outros indicadores tais como o aumento expressivo do número de cartões de crédito e volume de transações com cartões, que mais que dobraram de 1993 a 1997, e a comparação com o mercado de crédito norte-americano mostram que há muito espaço para o crescimento do mercado de crédito a prestações e cartão de crédito no Brasil, país que vem demonstrando uma forte explosão de crédito ao consumidor originado da estabilidade econômica.

Os métodos tradicionais de decisões sobre concessão de crédito a clientes individuais (sejam eles pessoas ou empresas) são fundamentados em julgamentos humanos a partir de experiências do julgador em decisões anteriores e são, portanto, bastante subjetivos e de agilidade insuficiente para grandes mercados de crédito. 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 concessão de crédito, procurando torná-las mais objetivas e rápidas e diminuir as perdas das carteiras de crédito. A nomenclatura *credit scoring* é usada para descrever o processo formal de determinar as probabilidades de candidatos a crédito atrasarem ou não o pagamento do compromisso financeiro que pretendem assumir, utilizando variáveis preditoras dos cadastros dos candidatos e de quaisquer outras fontes (tais como consultas a sistemas de proteção de crédito). Em termos gerais, a decisão de aceitar ou rejeitar o pedido de crédito é tomada comparando-se a probabilidade do candidato não honrar seu compromisso com o intervalo de probabilidades aceitável, ou seja, pela comparação do *score* (nota) do candidato com um possível *score* de corte.

A literatura existente sobre modelos de *credit scoring* evidencia que a grande maioria dos modelos tem sob foco a classificação dos créditos de acordo com as probabilidades de atraso no pagamento das prestações, levando a uma divisão dicotômica dos créditos do tipo "bom" ou "mau" crédito dependendo dos patamares de atrasos que podem ser aceitos para determinada carteira de crédito. Por exemplo, a instituição credora deseja saber a probabilidade de um tomador de crédito atrasar duas ou mais prestações consecutivas no pagamento de um financiamento automotivo com prazo de 24 meses, ou apresentar dez atrasos não consecutivos ao longo de todas as 180 prestações de um financiamento imobiliário, e assim por diante,

dependendo dos padrões aceitos por cada instituição para cada produto oferecido; padrões esses que, normalmente, refletem a possibilidade de existir lucro ou prejuízo ao final da operação de crédito com base no estudo de concessões passadas.

O banco de dados utilizado para a geração de um modelo de concessão de crédito é formado por uma amostra de candidatos aos quais o crédito já foi concedido no passado recente da carteira, denominada base de dados. Ele inclui as características do indivíduo e da operação de crédito – tipicamente variáveis preditoras – e a classificação real a que pertenceu essa operação – "bom" ou "mau" crédito – de acordo com os atrasos incorridos nos pagamentos das prestações desta. O modelo desenvolvido nos próximos capítulos segue essa divisão de operações passadas em dois grupos qualitativos, mais comum nesse tipo de estudo. Abordagens mais sofisticadas podem dividir a base de dados de formas diferentes, podendo utilizar, inclusive, classificações contínuas multidimensionais de acordo com os hábitos de pagamentos passados da carteira, e esse escalonamento mais refinado de operações pode ser usado como variável de resposta do modelo. Tais abordagens encontram-se em plena discussão de seus problemas estatísticos e dificuldades de implementação e interpretação, sendo que a aprovação técnica e prática dessas abordagens ainda são bastante inferiores à de modelos mais convencionais de divisão bidimensional.

Além dessa discussão sobre a dimensão de modelos de *credit scoring*, existe também a questão da falta de dinamismo de tais modelos, por não levarem em consideração o fato de que os tomadores de crédito são sujeitos a influências (externas, sociais e financeiras) que podem modificar suas propensões a apresentarem atrasos nos pagamentos de seus créditos. No entanto, esse tipo de consideração é um foco de estudo bastante desvinculado do propósito de modelos de concessão, visto que dizem respeito ao monitoramento de créditos já concedidos, ou seja, já

cumprido o objetivo do modelo de *credit scoring*. A literatura apresenta uma ramificação bastante grande e direcionada especificamente para modelos de comportamentos de créditos após a concessão, sob a denominação de *behavioural* ou *performance scoring*.

Conforme citado por Eisenbeis (1978), "com poucas exceções, os modelos de *credit scoring* são essencialmente direcionados a uma dimensão da função de concessão de crédito, embora seja esta a dimensão mais crítica, e que é a avaliação da probabilidade do crédito não ser pago, se tornar um mau crédito ou apresentar dificuldades de pagamento ao longo de sua duração". O autor afirma, ainda, que praticamente nenhum dos modelos existentes esclareceu seu objetivo e que não existem modelos de concessão tratando empiricamente de questões de otimização dinâmica, ou seja, considerando que créditos e empréstimos podem ser tomados e renovados diversas vezes por um cliente e que, portanto, deve ser considerado o relacionamento do cliente com a instituição e as possibilidades desse cliente vir a requisitar mais créditos na instituição durante ou após a quitação de um crédito anterior. Em outras palavras, os modelos de *credit scoring* almejam analisar apenas o crédito que está sendo requisitado por um cliente e fornecer uma regra de decisão única e exclusivamente para esse crédito, sem considerar a potencial evolução de capacidade de crédito e endividamento desse cliente e a trajetória futura que esse cliente pode manifestar no mercado de crédito.

Bierman e Hausman (1978) já haviam constatado que a concessão de crédito em um período é apenas uma parte do relacionamento da instituição credora com o cliente, que se estende por vários períodos. Segundo os autores, a decisão de concessão influencia o valor do relacionamento do cliente com a instituição tanto no período de duração de um determinado empréstimo quanto ao longo da vida útil desse relacionamento, o que inclui períodos anteriores e posteriores à concessão do empréstimo. Nas linhas de crédito ao consumidor, por exemplo, é

bastante nítido que a provisão de crédito está intimamente relacionada com o oferecimento de outros serviços financeiros ou não pelos credores (seguros, títulos de capitalização, financiamentos especiais, pacotes promocionais de tarifas etc.) durante períodos maiores do que o prazo de um único empréstimo.

Outra questão enfatizada na literatura de modelos de concessão de crédito remete novamente à discussão da divisão dos créditos em classes do tipo "bom" e "mau" ou mesmo em planos multidimensionais. O foco da maior parte dos modelos é a divisão de acordo com o risco de atraso no pagamento ou, em outras palavras, de acordo com o comportamento de inadimplência da carteira de crédito, mas diversos autores alertam para o fato de que esse tipo de risco é apenas um dos diversos aspectos do processo de decisão de concessão de créditos. Os críticos dizem que o foco principal normalmente deve ser a maximização de lucros e que essa não necessariamente estaria relacionada de forma monotônica com o risco de atrasos de pagamentos. Por exemplo, candidatos a crédito com risco muito baixo e que pagam suas prestações pontualmente conseguem créditos a taxas de juros mais baixas além de não pagarem juros e multas por atrasos e, portanto, não seriam rentáveis. Analogamente, candidatos com risco muito alto e que atrasam o pagamento de suas prestações podem ser bastante rentáveis desde que as taxas de juros de suas operações sejam suficientemente altas e que os atrasos não sejam prolongados. No entanto, os estudiosos de modelos de concessão de crédito muitas vezes omitem aspectos mercadológicos de concessão de crédito, inadimplência e lucratividade: na prática, a mensuração de risco através de atrasos nos pagamentos é intimamente relacionada com a lucratividade da operação de crédito, sendo observado que as operações de crédito com pequeno ou nenhum atraso formam o conjunto lucrativo da carteira de crédito (apesar do 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 através do estudo dos atrasos nos pagamentos funciona como uma aproximação muito boa para a lucratividade da carteira de crédito e poderiam ser claramente relacionadas: por exemplo, poderiam ser considerados maus créditos de uma carteira hipotética todos aqueles que apresentassem algum atraso de 60 ou mais dias no pagamento de qualquer uma das prestações porque, na prática, seriam exatamente esses créditos que propiciariam maior prejuízo.

O modelo elaborado no presente documento é fundamentado na divisão dicotômica gerada a partir do estudo de risco baseado em atrasos nos pagamentos, ou seja, o modelo é baseado na definição de qualidade de crédito (“bom” ou “mau”) pelo critério de inadimplência e não de lucratividade. O principal fator que levou a essa decisão, além das discussões citadas anteriormente, foi a falta de instituições financeiras dispostas a fornecer os dados financeiros necessários para uma divisão baseada diretamente na lucratividade. Nenhuma instituição se dispôs a fornecer dados completos que pudessem permitir a realização de estudo de lucratividade de suas carteiras de créditos, o que seria essencial caso se desejasse agrupar os créditos pelo critério de lucro e não de atrasos. Esse fato remete a outra discussão bastante presente acerca de modelos de concessão de crédito, que é a dificuldade e o alto custo na obtenção de dados para confecção dos modelos, basicamente originado pelas dificuldades técnicas de processamento de dados e pelo fato das instituições de crédito não poderem divulgar abertamente suas informações. Afinal, o assunto diz respeito a um mercado de elevada competição e volumes financeiros de bilhões de reais que vêm aumentando rapidamente, e quem detém tecnologia, estratégias e conhecimentos vantajosos desse mercado obviamente não pretende fornecer qualquer indício de seus procedimentos e resultados financeiros.

É de vital importância para o presente trabalho apresentar sua principal limitação, o chamado viés de seleção. A obtenção de resultados confiáveis para classificação de operações de crédito requer que a população de operações amostrada para geração do modelo estatístico, para estimação das taxas de acerto do modelo e realização de testes de hipóteses seja equivalente à população de novas operações sobre as quais o modelo será aplicado. No desenvolvimento de um modelo de *credit scoring* o objetivo é desenvolver uma ferramenta para classificar novas operações de crédito de acordo com suas possibilidades de serem operações boas ou ruins conforme o critério escolhido, que usualmente é relacionado às probabilidades de inadimplência no pagamento das prestações. No entanto, o banco de dados a partir do qual o modelo é estimado normalmente se refere a operações de crédito que foram aprovadas no passado, ou seja, são omitidas informações sobre operações que foram requisitadas no passado mas não foram aprovadas. Assim, a amostra utilizada não é baseada em toda a população de potenciais tomadores de crédito.

A ausência de informações sobre todos os potenciais tomadores de crédito acaba forçando os formuladores de sistemas de *credit scoring* a estimar um modelo a partir de uma população truncada de operações de crédito, ou seja, apenas uma parte (a de operações aprovadas) da população verdadeira. As conseqüências mais alarmantes desse procedimento foram estudadas por Avery (1977), que investigou os efeitos do viés de seleção em um estudo no qual os parâmetros verdadeiros de toda a população eram conhecidos. O autor provou que o uso de amostras restritas aos créditos aprovados gera resultados viesados, e o tamanho e direção do viés do modelo nem mesmo podem ser conhecidos. Alertou, também, que não adianta a instituição credora incorrer nos custos de tentar gerar uma amostra não viesada concedendo, por determinado período, créditos a todos os clientes que os requisitassem, na esperança de

conseguir amostrar uma população não viesada e gerar um modelo a partir dela: esse procedimento não assegura que a amostra será representativa de toda a população de potenciais tomadores de crédito. Se essa população conhecer bem as políticas de crédito das instituições, então ainda poderá haver um viés de seleção na população amostrada causado por parte dos próprios tomadores de crédito: eles escolherão pedir crédito na "melhor" instituição de acordo com seus critérios, logo nenhuma das instituições será capaz de obter uma amostra representativa de toda a população.

Os métodos de solução desse problema de viés de seleção ainda não foram capazes de fornecer a resposta ideal à questão. As duas soluções mais típicas investigadas são o uso de operações de crédito boas e ruins (aceitas) em conjunto com as operações rejeitadas, procurando recalibrar o modelo a partir das informações das operações rejeitadas, e o uso do "método de aumento", que é a utilização de pesos em amostras de operações de crédito boas e ruins aceitas baseados na probabilidade de aceitação em cada um dos dois grupos (operações boas aceitas e operações ruins aceitas). Nenhum dos métodos foi capaz de eliminar o viés existente sobre a capacidade preditiva dos modelos.

Existem ainda duas metodologias estatísticas direcionadas a resolver o problema de viés de seleção para dados censurados, ou seja, para quando o viés de seleção ocorre devido à ausência parte de informações que existem na população verdadeira mas não se encontram na amostra (por exemplo, créditos recusados). Essas metodologias, conhecidas respectivamente como Tobit e método de seleção de Heckman, se diferenciam principalmente no enfoque que possuem quanto ao processo de participação ou não de casos na amostra (ou seja, o processo que determina se os indivíduos da população estarão ou não contidos na amostra) e ao processo de decisão (isto é, o processo que estima o comportamento dos indivíduos em relação à variável de

resposta do estudo, uma vez que estejam contidos na amostra). No modelo Tobit esses dois processos são associados, considerados como o mesmo processo, enquanto que no modelo de seleção de Heckman – mais utilizado que o primeiro, do qual é considerado uma generalização - esses processos são distintos. O modelo de Tobit procura corrigir o viés de seleção que causa correlação entre os resíduos e as variáveis explicativas do modelo através do deslocamento das distribuição da variável dependente, baseado nos momentos das variáveis gerados a partir de sua função de densidade e de distribuição cumulativa, calculando uma "nova esperança" dos resíduos. Os coeficientes finais do modelo são estimados por máxima verossimilhança. O modelo de seleção de Heckman estima uma equação para o processo de participação e outra para o processo de decisão, partindo da hipótese de existência de exogeneidade nos dois processos. A variável explicativa encontrada no processo de participação (que é uma variável omitida no processo de decisão) é inclusa no processo de decisão e assim o viés de seleção é eliminado. Os coeficientes finais são estimados em dois estágios por mínimos quadrados ordinários.

O modelo Tobit e o modelo de seleção de Heckman exigem a existência de variáveis explicativas para os casos censurados, isto é, requer que sejam conhecidas as características das operações de crédito recusadas, o que dificilmente é possível em modelos de *credit scoring*, visto que o banco de dados utilizado não contém tais informações (as características das operações recusadas não são armazenadas). Mesmo que sejam fornecidos tais dados, os dois métodos ainda apresentam complicações, tal como a falta de robustez para heterocedasticidade do modelo de Tobit (as estimativas são inconsistentes) e a existência de uma hipótese de normalidade conjunta nos resíduos dos dois processos do modelo de seleção de Heckman, cujas estimativas são consistentes mas não eficientes.

O problema de viés de seleção não tem se mostrado simples de solucionar, dado o grau de desenvolvimento atual das técnicas estatísticas relacionadas ao problema de estimação e processos de amostragem. Mesmo assim, em geral os métodos de *credit scoring* são considerados capazes de classificar corretamente grande parte das operações de crédito, principalmente quando gerados sobre amostras de grande tamanho contendo o maior número possível de variáveis. A existência do viés de seleção requer que a implementação de um modelo de *credit scoring* seja feita com consciência de que o modelo estimado sofre desse problema estatístico e, portanto, não é um reflexo perfeito da realidade.

O presente estudo está dividido em seis capítulos. No capítulo I é discutida a questão do banco de dados necessário para elaboração de um modelo de *credit scoring*, as variáveis obtidas para realização do trabalho e os filtros aplicados sobre o banco de dados. O capítulo II traz a definição do problema, que começa com a definição de qualidade de crédito e compreende o tratamento e transformações estatísticas (técnicas de agrupamento ou categorização) feitas sobre as variáveis obtidas, para trazer uma melhor percepção do problema em estudo. No capítulo III estão discutidas as técnicas estatísticas utilizadas para a geração do modelo, compostas por duas ferramentas básicas, a regressão logística e o método *forward stepwise* para seleção de variáveis explicativas. O capítulo IV apresenta e discute os resultados obtidos, ou seja, o modelo final, um exemplo de sua aplicação, a análise gráfica dos resultados estimados, o estudo de estabilidade do modelo e uma discussão sobre a decisão de crédito, que consiste na definição do *score* de corte para a carteira. No capítulo V são feitas algumas considerações finais acerca de problemas estatísticos e ferramentas para acompanhamento do desempenho do modelo, sendo que no capítulo VI encontra-se a conclusão do estudo.

Capítulo I. Base de Dados

A base de dados necessária para a formulação de modelos de *credit scoring*, qualquer que seja a técnica estatística utilizada, costuma ter tamanho bastante grande, não sendo raras bases de dados contendo mais de 100.000 clientes para estudo e mais de 100 variáveis relativas à esses clientes e às operações que fizeram no passado recente[1].

De fato, a base de dados pode ser dividida em dois grupos. O primeiro grupo de dados depende da forma escolhida para classificar operações de crédito boas e ruins, ou seja, de determinar a qualidade de crédito de cada operação. Seguindo a linha tradicional de modelos de concessão de crédito, cujo fundamento é a classificação das operações em dois grupos – qualidade de crédito boa ou ruim – baseada no princípio de que o risco de uma operação é determinado pelas possibilidades de ocorrerem atrasos nos pagamentos das prestações da operação de crédito, o primeiro grupo de dados requer, necessariamente, informações completas sobre datas de vencimento e datas de pagamento de cada prestação de todas as operações disponibilizadas para análise. Por exemplo, se um modelo de *credit scoring* é baseado em 100.000 operações de créditos dos últimos três anos até a data de hoje, sendo que todas essas operações têm 24 prestações mensais e já estão quitadas, o primeiro grupo de dados deve ser formado, no mínimo, por 48 datas de vencimento e pagamento das 100.000 operações, ou seja, cerca de 4.800.000 datas. Se o princípio do risco for baseado na questão da lucratividade das

[1] Hand, D. J.; Henley; W.E., 1997, op. cit.

operações, então é necessário uma quantidade de dados ainda maior, sendo necessárias não somente as datas de pagamento e vencimento, mas também todos os valores financeiros envolvidos em cada prestação (entre outros), tais como valor total, valor dos juros e valor da amortização de cada prestação, valor do encargo pago por atraso, valor de juros de mora por atraso em cada prestação etc., o que pode tornar a base de dados muito grande e comprometer a viabilidade técnica da elaboração do modelo. A Tabela 1 resume as variáveis necessárias para o primeiro grupo de dados, relativo ao estudo de qualidade de crédito, de acordo com o critério de percepção de risco e válido somente para créditos assumidos para serem pagos em prestações periódicas. No caso de outros tipos de crédito – desconto de títulos, créditos rotativos, cheque especial etc. – as variáveis são bastante distintas e os cálculos para classificação dicotômica das operações baseada em atrasos ou mesmo lucratividade exigem maior esforço computacional e de montagem de bancos de dados já que, em linhas gerais, é necessário acompanhar o dia-a-dia de cada contrato até o seu vencimento, e não somente as datas de vencimento e pagamento de cada prestação.

Tabela 1. Variáveis da Base de Dados – Grupo I – Classificação de Operações de Crédito por Critério de Percepção de Risco

Critério: atrasos (inadimplência)	
Variáveis:	*Tipo de Resposta:*
Datas de vencimento das prestações	dd/mm/aaaa
Datas de pagamento das prestações	dd/mm/aaaa
Critério: lucratividade	
Variáveis:	*Tipo de Resposta:*
Datas de vencimento das prestações	dd/mm/aaaa
Datas de pagamento das prestações	dd/mm/aaaa
Valores totais das prestações	R$
Valores dos juros das prestações	R$
Valores das amortizações das prestações	R$
Valores dos encargos por atraso das prestações	R$
Valores dos juros de mora das prestações	R$

As duas maiores dificuldades que podem ser destacadas na obtenção do primeiro grupo de dados dizem respeito à escolha do critério de classificação de operações e à dificuldade de obtenção dos dados em alguma instituição de crédito. O critério escolhido para o presente trabalho foi a classificação baseada no estudo de atrasos nos pagamentos das prestações, devido ao fato de que nenhuma instituição se dispôs a fornecer os dados necessários para utilização do outro critério (lucratividade) por revelarem nitidamente os resultados financeiros da carteira de crédito, e pela conclusão de que o critério de atrasos nitidamente satisfaz à questão da lucratividade, conforme discutido anteriormente. O primeiro grupo de dados permite gerar a variável de resposta do modelo de *credit scoring*, que é a qualidade de crédito da operação.

O segundo grupo de dados é formado pelas variáveis disponíveis nos cadastros dos clientes e de suas respectivas operações, ou seja, as potenciais variáveis explicativas do modelo de *credit scoring*, as quais são estudadas de acordo com sua influência sobre a variável de resposta, a qualidade de crédito, para detectar quais delas são realmente relevantes e

significativas para determinar a qualidade de crédito da operação. As variáveis necessárias dependem de cada situação, ou seja, do tipo de linha de crédito que está sendo estudado, e englobam não somente as características dos clientes que realizaram as operações de crédito (cadastro completo, patrimônio, informações de conta bancária etc.), mas também as características das próprias operações (valor da operação, forma de pagamento da prestação etc.). Assim, as variáveis necessárias para um estudo de concessão de crédito popular para compras de baixo valor – tipicamente operações de curto ou médio prazos com altas taxas de juros e prestações de baixo valor - podem ser substancialmente diferentes daquelas necessárias a um estudo relativo a uma linha de financiamento imobiliário, caracterizadas por prazos longos, juros menos elevados e prestações de maiores valores. As variáveis também podem ser bastante diferentes dependendo do tipo de tomador do crédito: uma linha de crédito popular a pessoas físicas exige um cadastro de informações muito menor dos clientes e operações do que uma linha de crédito a pessoas jurídicas (empresas de micro, pequeno, médio e grande portes). A Tabela 2 resume as variáveis mais típicas do grupo de variáveis explicativas em modelos de *credit scoring* direcionados a carteiras de pessoas físicas.

Tabela 2. Variáveis da Base de Dados – Grupo II – Variáveis Explicativas

Tipo de Informação:	*Variáveis:*	*Tipo de Resposta:*
Identificação	Nome do Cliente	Texto
	Código da Operação	Codificada
	Código do Cliente	Codificada
Cadastro	Idade	Anos
	Sexo	Feminino/Masculino
	Estado Civil	Codificada
	Regime de Casamento	Codificada
	Tempo de Residência Atual	Meses
	Residência Própria	Sim/Não
	Escolaridade	Codificada
	Quantidade de Dependentes	Número (00, 01, 02 etc.)
	Profissão e/ou Ocupação	Codificada
	Profissão/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 (quant. e valor, c/ indicação alienação, comprovação, hipoteca etc.)	Automóveis	Número e R$
	Imóveis	Número e RS
	Outros Bens	R$
Informações Bancárias do Cliente (para cada banco com o qual o cliente opera)	Banco e Número de Conta	Codificada
	Data de Abertura de Conta	dd/mm/aaaa
	Tipo de Conta	Codificada
	Saldo Médio em Conta-Corrente	R$
	Saldo Médio de Aplicações	R$
	Cartão de Crédito	Sim/Não
	Bloqueios / Restrições	Sim/Não

Tabela 2. Variáveis da Base de Dados – Grupo II – Variáveis Explicativas (cont.)

Tipo de Informação:	*Variáveis:*	*Tipo de Resposta:*
Compromissos Financeiros (para cada tipo: aluguel, educação, financiamentos, empréstimos etc.)	Tipo	Codificada
	Periodicidade de Pagamentos	Codificada (ex.: 01=mensal)
	Valor Nominal	R$
	Número de Parcelas a Vencer	Número (00, 01, 02 etc.)
Dados da Operação	Valor da Operação de Empréstimo	R$
	Quantidade de Prestações	Número (00, 01, 02 etc.)
	Forma de Pagamento da Prestação	Codificada (ex.: A = carnê)
	Forma de Pagamento de Impostos	Codif. (ex.: 01 = financiado)
	Forma de Cobrança de Tarifas	Codificada (ex.: C = à vista)
	Finalidade da Operação	Codificada
	Data do Contrato	dd/mm/aaaa
	Data de Vencimento da Operação	dd/mm/aaaa
	Taxa de Juros (*)	% a.m. ou % a.a.
	Referência Monetária	Codif. (ex.: 1 = Real, 2 = TR)
Apontamentos Negativos (**) (com registros de datas, valores, quantidades e datas de regularização, caso houver)	Protestos	Codificada
	Cheques Devolvidos	Codificada
	Ações Judiciais	Codificada
	Pendências Financeiras	Codificada
	Cheques Irregulares/Sem Fundo	Codificada

(*) A necessidade de ter a informação sobre taxa de juros da operação é discutível para métodos que propõem analisar pedidos de crédito, já que, nas práticas bancárias, as taxas de juros das operações podem ser definidas somente após a aprovação do crédito, dependendo, inclusive, do score obtido pelo cliente. Logo, a taxa de juros não pode ser usada como variável explicativa já que depende da variável de resposta do modelo.
(**) Também conhecidos como "Restrições Cadastrais", são as informações obtidas junto às agências reguladoras e de proteção ao crédito, tais como Serasa e SPC.

. É importante discutir que a formação de um banco de dados com grande número de informações sobre os clientes e operações pode gerar estatísticas mais confiáveis e robustas para um modelo de *credit scoring*. No entanto, esse fato pode trazer prejuízos financeiros consideráveis para as instituições provedoras de crédito, já que exigiria que os clientes requisitantes de crédito preenchessem um vasto e demorado cadastro e que a instituição fizesse muitas consultas à agências de regulação e proteção de crédito. Portanto, os custos de obter um volume grande de informações são bastante claros: há os custos para os clientes, pela paciência, irritação e perda de tempo ao preencher grandes cadastros; e os custos para a instituição credora,

pelos encargos cobrados pelas agências de regulação e proteção de crédito e, principalmente, pela possível desistência dos clientes mediante toda a burocracia e exigências de preenchimento de dados. Esse último aspecto – custos por desistência – ainda foi pouco explorado pela literatura, mas há evidências de que passarão a ter maior importância à medida que a dificuldade de mensurá-los seja eliminada, o que pode revelar que esses custos podem ser bastante altos para as instituições credoras. A questão mais alarmante é a fuga de bons clientes, que normalmente têm acesso a diversas linhas de crédito de outras instituições concorrentes e muitas vezes não apresentam uma necessidade real de conseguir crédito e, portanto, estão mais sujeitos a desistir de enfrentar um processo burocrático complicado e demorado. Segundo Hand e Henley (1997), "é provável que um processo lento desanime aqueles que não precisam realmente do empréstimo ou têm acesso a outras fontes de crédito, ou seja, os bons riscos podem desistir de requisitar um empréstimo".

Os bancos de dados para estudos de concessão de crédito têm uma característica usual em bases de dados com muitas variáveis, os chamados *missing values* ou ausência de resposta em determinadas variáveis. A ausência pode ser estrutural, nos casos em que a resposta não existe por não ser necessária, tal como em perguntas que somente são feitas condicionalmente à resposta dada em outra variável (por exemplo, o cliente somente precisa responder a profissão de seu cônjuge caso tenha respondido que era casado em uma pergunta sobre estado civil), ou, caso contrário, pode ser aleatória, caso em que normalmente o cliente é quem opta por não responder. As formas de lidar com o problema de *missing values* em modelos de discriminação de características têm sido amplamente discutidas por estatísticos, surgindo alternativas como eliminar do estudo todos os clientes que apresentam ausência de resposta em qualquer uma das variáveis e eliminar também todas as variáveis que apresentam pelo menos um cliente com

ausência de resposta, substituir os *missing values* por valores válidos através de técnicas estatísticas que permitam prever qual seria a resposta real da variável ou simplesmente aplicar o método de classificação em uma população apropriada, ou seja, cujos padrões de resposta sejam os mesmos existentes na base de dados passada utilizada para gerar os cálculos estatísticos.

A alternativa que tem se relevado melhor para trazer informações úteis ao processo de discriminação de clientes é a codificação dos *missing values* como uma classe de respostas adicional para cada variável. Por exemplo, para a variável estado civil, haveria as classes de resposta "solteiro" com código 1, "casado" com código 2, "missing" com código 3, "divorciado" com código 4 e assim por diante. O princípio que tem dado bases à utilização desse procedimento é que a recusa em responder a uma determinada questão apresentada no processo de análise de crédito pode ser indicadora de um risco maior para aquele cliente e sua operação de crédito. Assim, a ausência de resposta é considerada como uma resposta válida e capaz de discriminar bons e maus clientes à obtenção de um crédito qualquer. O presente estudo se baseia nesse princípio, com a diferença de que variáveis ou clientes com alto percentual de respostas do tipo *missing* são eliminadas do banco de dados, de forma que o modelo final não seja uma ferramenta fundamentada em ausências de respostas.

I.1. Nome da Carteira de Crédito, Período Amostrado e Variáveis Coletadas

A dificuldade de obtenção de dados para geração de modelos de *credit scoring*, discutida anteriormente, por questões de necessidade de preservação de informações importantes das instituições, foi visualizada na prática para obtenção do banco de dados a ser utilizado nos próximos capítulos. A origem dessa dificuldade está nas próprias instituições, que participam de

um mercado cujas técnicas evoluem rapidamente e onde a concorrência leva à não divulgação de resultados financeiros e procedimentos burocráticos valiosos.

As diversas instituições visitadas alegam desde a falta de disponibilidade desses dados em seus sistemas de informática ou ao tempo de trabalho que a reprodução dos dados exigiria de alguns de seus funcionários, até o receio de compartilhar informações sobre suas operações de crédito com terceiros e com isso revelar seus resultados financeiros e seus procedimentos burocráticos. A solução foi encontrada em uma instituição bancária brasileira que fornece diversas linhas de crédito (principalmente popular) a pessoas físicas e jurídicas (clientes do banco ou não): o banco – que conta com um sistema de informática bastante desenvolvido e com alta capacidade de armazenamento de dados – aceitou fornecer os dados sobre uma de suas diversas linhas de crédito popular[2] para aquisição de bens de consumo duráveis, exclusivamente direcionada a correntistas, desde que fosse apresentada uma lista de variáveis importantes sobre a qual o próprio banco pudesse escolher as variáveis a fornecer, que não fossem apresentados ou elaborados resultados que pudessem tornar diretamente visíveis seus resultados financeiros ou práticas de concessão de crédito, que fosse criado um nome fictício para a carteira de crédito e que a identidade da instituição fosse preservada.

O resultado desse esforço foi a obtenção de um banco de dados contendo 206.383 operações de crédito uma linha de crédito popular com prazo máximo de 24 meses, assinadas entre janeiro de 1998 e abril de 2001, com as datas de vencimento de todas as prestações e datas de pagamento das prestações pagas até a data de geração do banco de dados e com 43 variáveis referentes às operações e aos clientes que contrataram o crédito. A carteira de crédito foi

[2] A expressão "linha de crédito popular", segundo o banco fornecedor dos dados, se refere operações de empréstimo ou crédito de valor limitado entre R$ 200,00 e R$ 2.000,00 e direcionadas ao público de renda baixa ou média.

denominada *CAB* (de "crédito para aquisição de bens") e as variáveis coletadas encontram-se na Tabela 3, dividida em grupo I (variáveis direcionadas à definição de qualidade de crédito como "boa" ou "ruim" pelo critério de inadimplência) e grupo II (variáveis explicativas).

Tabela 3. Variáveis Coletadas para Formulação do Modelo de Credit Scoring – Carteira CAB

Grupo I – Variáveis para Definição de Qualidade de Crédito	
Datas de Vencimento	Datas de Pagamento
Grupo II – Variáveis Explicativas	
Nome do cliente	Valor comprovado de imóveis
Código da Operação	Quantidade comprovada de imóveis
Código do Cliente	Valor de automóveis
Residência própria (S/N)	Quantidade de automóveis
Idade	Quantidade de seguros de automóvel
Quantidade de dependentes financeiros	Quantidade de seguros de vida
Estado civil	Quantidade total de seguros
UF de nascimento	Cheque
Sexo	Profissão
E-mail (S/N)	Saldo médio em conta-corrente
Salário líquido	Saldo médio de aplicações
Outras rendas mensais	Bloqueio da emissão de cheques
Salário líquido do cônjuge	Tipo de tomador do empréstimo
Idade na data de admissão no atual emprego	Percentual de taxa de juros da operação
Idade na data de admissão (anos) dividida pela idade atual (anos)	Valor dos cheques devolvidos nos 6 meses anteriores à concessão – motivo 12
Tempo de residência (meses) dividido pela idade (meses)	Valor total dos cheques devolvidos nos 6 meses anteriores à concessão
Cartão "Credicard" (S/N)	Valor da operação de empréstimo
Quantidade de cartões de crédito	Quant. de prestações da operação de emprést.
Compromissos financeiros	Referência monetária da operação
Quantidade de imóveis	Forma de pagamento do IOF da operação
Tempo de residência	Valor total de garantias da operação
	Quantidade de avalistas da operação

O banco de dados coletado contém a maioria das variáveis mais importantes para formação de modelos de *credit scoring*, ou seja, as variáveis mais relacionadas com a capacidade

de crédito dos clientes. No entanto, pode ser notada a ausência de algumas variáveis muito importantes que a instituição não concordou em ceder de forma a não revelar totalmente o cadastro de seus clientes e detalhes de suas operações de crédito. Entre as mais relevantes devem ser citadas:

- Informações bancárias do cliente: somente foram disponibilizadas informações sobre contas dos clientes no próprio banco, sem qualquer informação sobre contas em outros bancos. Além disso, para clientes que possuíam mais de uma conta-corrente no banco, somente foi disponibilizada a conta com maior saldo médio nos três meses anteriores à data de concessão do crédito. Informações sobre tipo de conta também foram omitidas, não permitindo saber se a conta é comum, conta poupança, conta de proventos etc. A conseqüência da falta de informações bancárias completas é que o perfil do cliente fica incompleto e, portanto, a capacidade preditiva do modelo de *credit scoring* pode ser prejudicada;
- Apontamentos negativos: as informações de consultas a agências reguladoras e de proteção de crédito, tais como Serasa e SPC, não puderam ser disponibilizadas já que o banco não tem o direito de compartilhar essas informações com terceiros, segundo os contratos firmados entre essas agências e o banco. Assim, não foram recebidas quaisquer informações sobre protestos, ações judiciais, pendências financeiras e cheques irregulares ou sem fundo (entre outros) dos clientes que obtiveram crédito no banco. No entanto, o banco pôde fornecer informações sobre cheques devolvidos de seus clientes (ou seja, somente cheques do próprio banco), visto que essas informações são geradas e armazenadas pelo próprio banco. A falta de informações completas sobre apontamentos negativos pode reduzir a capacidade de discriminação do modelo de concessão, já que torna o perfil dos clientes incompletos.

Entre outros dados que não foram recebidos devem ser citados a escolaridade do cliente, que costuma ser uma informação bastante relevante para concessões de crédito a pessoas físicas, o regime de casamento, a profissão do cônjuge, a renda familiar total, compromissos financeiros individualizados (somente foram recebidas informações totalizadas) e finalidade da operação de crédito. A ausência de todas essas informações diminui o potencial de discriminação no modelo, o que é atenuado pela presença de diversas outras informações que podem ser utilizadas como substitutas para grande parte das informações ausentes.

I.2. Softwares Utilizados

Os *softwares* utilizados para armazenar e organizar o banco de dados e realizar os procedimentos estatísticos para geração do modelo de *credit scoring* da carteira CAB foram o *SPSS 6.0* (Statistical Package for Social Sciences), devido à sua alta capacidade de armazenar uma grande quantidade de dados e a existência de todas funções estatísticas necessárias para o trabalho, e o software *Answer Tree 2.0*, um pacote opcional do *SPSS* utilizado somente no processo de categorização de variáveis.

I.3. Filtros Utilizados na Base de Dados

O banco de dados original recebido da linha de crédito CAB continha, inicialmente, 206.383 concessões de janeiro de 1998 a abril de 2001. Devido ao grande volume de informações contidas em bancos de dados sobre concessões de crédito, é comum encontrar dados inconsistentes (ex.: tempo de residência maior que a idade), dados faltantes (missing), dados

inválidos (ex.: código de estado civil inexistente) e dados extremos (ex.: salário líquido de R$ 100.000,00). Dentre essas 4 categorias de problemas com dados, os dados inconsistentes, dados inválidos e dados extremos – normalmente frutos de erros de digitação e armazenamento de dados – devem ser eliminados integralmente do banco de dados. Integralmente significa que não só aquele dado (por exemplo, estado civil inexistente) da concessão deve ser eliminado, mas sim todas as variáveis daquele cliente (dados cadastrais, financeiros, do banco, da operação de crédito etc.), o que exclui totalmente a concessão de crédito da amostra.

Os dados inconsistentes e inválidos costumam ocorrer por motivos de preenchimento incorreto de cadastro por parte do cliente ou até mesmo por erro do digitador. Os dados extremos não são dados necessariamente incorretos. Por exemplo, havia 94 casos de clientes com salário líquido no valor de R$ 30.000,00 ou mais. Embora esse dado não esteja necessariamente incorreto, a chance dele ser resultado de um erro de digitação ou qualquer outro problema é grande, ainda mais quando se recorda que a linha de crédito que está sendo examinada (CAB) é direcionada a público de renda média ou baixa para aquisição de bens de até R$ 2.000,00. Como o banco de dados do estudo era grande, optou-se por eliminar essas concessões a correr o risco de se trabalhar com dados incorretos. Mesmo que não fosse incorreto, a chance de aparecer outro candidato a crédito com o mesmo tipo de dado extremo é muito baixa, o que garante que o modelo continuará bastante eficaz mesmo com a eliminação do cadastro.

Os dados faltantes (missing) são aqueles em que não há preenchimento para a variável pedida no cadastro da concessão de crédito, por motivos estruturais ou aleatórios, conforme discutido anteriormente, e ocorrem em praticamente todas as variáveis e para todos os clientes. Nesses casos, não há necessidade de eliminar toda a concessão de crédito do estudo por causa da existência de algum dado faltante relacionado aos dados da concessão. A alternativa utilizada foi

a codificação dos *missing values* como uma classe de respostas adicional para cada variável, de forma que a ausência de informação é considerada como uma informação válida e capaz de discriminar bons e maus clientes à obtenção do crédito. Variáveis com excesso de *missing values* – que não foram eliminadas do estudo – têm menos chances de se mostrarem relevantes para discriminar bons e maus clientes de uma linha de crédito quando existem variáveis com melhor preenchimento e maior impacto sobre a classificação de clientes. Por exemplo, a salário líquido do cônjuge apresentou 88,5% de *missing values*, ou seja, das 206.383 operações de crédito analisadas, 179.428 não continham a informação sobre salário líquido do cônjuge do tomador do crédito, o que reduz a possibilidade dessa variável participar de um modelo final para concessão de crédito.

As 206.383 concessões e informações originalmente recebidas foram filtradas em relação a dados inconsistentes, inválidos e extremos, e os resultados foram os seguintes:

1) Eliminação de indivíduos com data de admissão menos data de nascimento maior que 720 meses ou menor que 180 meses (3.557 casos, restaram 202.826);
2) Eliminação de indivíduos com mais de 10 dependentes (15 casos, restaram 202.811);
3) Eliminação de indivíduos com tempo de residência maior que 720 meses (269 casos, restaram 202.542);
4) Eliminação de indivíduos com UF de nascimento "JP – Japão" e "US – Estados Unidos" (17 casos, restaram 202.525);
5) Eliminação de indivíduos com salário líquido maior que R$ 30.000,00 (94 casos, restaram 202.431);

6) Eliminação de indivíduos com outras rendas mensais maiores que R$ 30.000,00 (17 casos, restaram 202.414);
7) Eliminação de indivíduos com salário líquido do cônjuge maior que R$ 30.000,00 (14 casos, restaram 202.400);
8) Eliminação de indivíduos com encargos mensais (compromissos financeiros) maiores que duas vezes o salário líquido (257 casos, restaram 202.143);
9) Eliminação de indivíduos com valor comprovado de imóveis maior que R$ 2.000.000,00 (51 casos, restaram 202.092);
10) Eliminação de indivíduos com idade maior que 80 anos ou menor que 18 anos (980 casos, restaram 201.112);
11) Eliminação de indivíduos com quantidade de imóveis maior ou igual a 11 (37 casos, restaram 201.075).

As 201.075 operações de crédito restantes e todas as variáveis relacionadas a elas foram utilizadas para realização do estudo de qualidade de crédito (que define a divisão das operações em grupos de qualidade de crédito "boa" ou "ruim" pelo critério de inadimplência), a categorização (agrupamento de respostas de cada variável em grupos homogêneos em relação à qualidade de crédito), a geração de tabelas cruzadas das variáveis categorizadas em relação à qualidade de crédito e para extrair a amostra de desenvolvimento do modelo final.

Capítulo II. Definição do Problema

A primeira questão importante para elaboração de um modelo de concessão de crédito após a obtenção e organização do banco de dados é definir qual é a variável de resposta do modelo e como ela pode ser obtida. Conforme observado anteriormente, o procedimento mais utilizado é a criação de dois grupos de crédito que reflitam os comportamentos dos clientes em relação aos atrasos nos pagamentos das prestações do crédito assumido, ou seja, a divisão dos créditos analisados em grupos de "bons" e "ruins" pelo critério de inadimplência.

De acordo com Einsenbeis (1978), a abordagem típica de modelos de *credit scoring* é dividir a amostra de operações de crédito em dois grupos mutuamente exclusivos, os bons créditos – aqueles cujas prestações foram pagas em dia ou houve alguns atrasos de poucos dias em algumas prestações – e os maus créditos – aqueles em que algumas prestações foram pagas com atraso mais prolongado. Os maus créditos são vistos como tendo alto risco, enquanto que os bons créditos são vistos como tendo baixo risco. Assim, o passo seguinte é formular uma regra de classificação (análise discriminante, regressão linear, regressão logística etc.) para discriminar os dois grupos. Ainda segundo o autor, a classificação das operações em dois grupos – boas e ruins – não traz nenhuma complicação para o estudo, contanto que seu objetivo seja perceber o risco das operações de crédito como discreto em vez de contínuo.

Em resumo, as operações de crédito do passado recente da carteira em questão podem ser classificadas de acordo com os atrasos ocorridos nos pagamentos, de forma que seja criada uma nova variável, denominada "qualidade de crédito", que assume valor 0 quando a qualidade de

crédito é considerada ruim e o valor 1 quando é considerada boa. Em outras palavras, são estudadas as operações de crédito do passado, todas essas operações são classificadas em "boas" e "ruins" de acordo com os atrasos ocorridos nos pagamentos das prestações, e essa classificação se torna a variável de resposta do modelo para que, através de técnica estatística, sejam detectadas quais variáveis explicativas (cadastrais, financeiras, patrimoniais etc.) tiveram impacto estatisticamente significante sobre a qualidade de crédito das operações.

O segundo aspecto na definição do problema de elaboração de um método de análise de concessão de crédito é decidir sobre a forma de utilização das variáveis explicativas, além de ter uma visão preliminar das relações entre as variáveis explicativas e a variável de resposta, que é a qualidade de crédito. Pode-se utilizar as variáveis em seu formato original, ou seja, com as respostas originalmente fornecidas sobre as características do cliente e de sua operação de crédito, ou agrupar as respostas de cada variável em classes (categorias) homogêneas em relação à qualidade de crédito, de forma a simplificar a interpretação e mensuração dessas relações e ter uma percepção inicial de quais variáveis parecem ser mais associadas à qualidade de crédito. Esse aspecto da metodologia é essencial para observar e conhecer as características das concessões de crédito analisadas e o comportamento geral da carteira de crédito mas, de fato, não é obrigatório utilizar um processo de agrupamento de respostas das variáveis. Em outras palavras, um modelo de *credit scoring* pode ser perfeitamente estimado a partir das respostas originais das variáveis, sem necessidade de agrupamento de respostas, ou seja, de transformação do banco de dados. Nesse caso, em vez de ser obtido um modelo baseado em variáveis categorizadas – cuja equação final é composta por variáveis *dummies* representativas de cada uma das categorias de cada variável estatisticamente significante para explicar a qualidade de crédito (um modelo com 4 variáveis com 3 categorias cada uma teria 8 coeficientes na equação

final) – o modelo final seria estimado diretamente sobre as variáveis originais, e a equação final seria composta apenas por um coeficiente para cada variável significante (um modelo com 4 variáveis teria 4 coeficientes na equação final). No entanto, a opção de utilizar as respostas originais é inviável quando existem muitas variáveis qualitativas para estudo (estado civil, sexo etc.), já que estas necessariamente precisam ser codificadas e representadas por variáveis *dummies* no processo de estimação de coeficientes. Já que em modelos de concessão de crédito existem muitas variáveis qualitativas relevantes para análise, é praticamente impossível obter um modelo final estimando apenas um coeficiente para cada variável.

A alternativa adotada foi categorizar todas as variáveis explicativas em relação à variável qualidade de crédito. Foram analisadas todas as variáveis explicativas uma a uma, de forma que, para cada variável, as respostas possíveis fossem agrupadas em classes ou categorias homogêneas de acordo com a qualidade de crédito atribuída às operações de crédito. Essa decisão é fundamentada por seu apelo operacional, já que simplifica a interpretação e funcionamento na prática de modelos de *credit scoring*, além de outras vantagens da utilização do processo de categorização, que são o menor número de variáveis no modelo final, o maior poder de previsão de modelos baseados em variáveis com poucas categorias homogêneas internamente e a garantia de que indivíduos com risco de crédito equivalentes têm pesos iguais no modelo.

Cumpre observar que o processo de categorização é muito usado em modelos de *credit scoring*, existindo, inclusive, uma linha de pesquisa de modelos baseados única e exclusivamente nesse procedimento, com o nome formal de *recursive partitioning*[3], que se resumem em

[3] Srinivasan, V.; Kim, Y. H. Credit granting: a comparative analysis of classification procedures. The Journal of Finance, v. XLII, n.3, July 1987.

processos de geração de árvores de decisão com ramificações significantes quanto ao poder de discriminar bons e maus clientes em operações de crédito. No entanto, os modelos dessa linha não têm encontrado aceitação por parte dos profissionais que trabalham na área de crédito, já que a formulação, interpretação e implementação da árvore de decisão não têm a mesma facilidade que outros métodos posseum (tais como análise discriminante, regressão linear, regressão logística etc.) e que alcançam o mesmo poder de discriminação.

II.1. Definição de Qualidade de Crédito

A definição de qualidade de crédito é o primeiro item da formação de um modelo de *credit scoring* após a obtenção e organização de um banco de dados para o estudo. Essa definição é gerada a partir da análise do histórico de créditos já concedidos pelo banco, na qual são analisados os períodos de atraso no pagamento de cada prestação e a migração para atrasos ainda maiores. Assim, é possível verificar quais são os limites de atrasos que determinam se uma operação de crédito é boa ou ruim em termos de inadimplência da carteira de crédito. Determinados os limites de atrasos, é possível gerar a variável de resposta do modelo, que é a qualidade de crédito, para que o modelo final seja formado por um grupo de variáveis que se mostrar mais relevante para determinar a qualidade de crédito da carteira.

O estudo de migração de faixas de atrasos é feito a partir da análise de atrasos entre uma prestação e a prestação subsequente, utilizando somente as variáveis de datas de pagamento e de vencimento das prestações dos créditos recebidos para análise. Por exemplo, é possível saber as probabilidades de atrasos de 1 a 30 dias no pagamento de determinada prestação evoluírem para atrasos de 31 a 60 dias no pagamento da prestação seguinte, e assim por diante. Dessa forma, é

possível conhecer todas as probabilidades de aumento de atrasos e determinar o nível de atraso que reflete chances elevadas da operação de crédito evoluir para a inadimplência. Esse nível de atraso permite separar as boas operações de crédito das ruins.

A forma de realizar o estudo de qualidade de crédito, na prática, é simples. Cada operação de crédito contém datas de vencimento e de pagamento das prestações. Os pontos de referência para cálculo de atrasos são as datas de vencimento de cada prestação. Em cada data de vencimento (data de referência), é calculado o atraso máximo que esteve em aberto no período entre essa data e a data de referência anterior. Esse atraso é agrupado em faixas de atraso: 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 "181 ou mais", já que contratos que atingem 181 dias de atraso já são considerados contratos em perda ou processo de liquidação, ou seja, com pequenas chances de recuperação financeira. A Tabela 4 contém um exemplo da forma de cálculo de atrasos de um contrato de operação de crédito hipotético, explicado em seguida, sendo que todos os atrasos são contabilizados em dias corridos.

Tabela 4. Exemplo de Forma da Cálculo de Atrasos em Operações de Crédito

Data de Geração do Banco de Dados: 10/01/01		Código de Identificação do contato: 01k27	
Quantidade de Prestações Contratadas: 12		Data de Vencimento 1ª Prestação: 05/03/00	
Datas de Vencimento (Data de Referência)	**Datas de Pagamento**	**Número de Dias do Maior Atraso em Aberto Entre a Data de Referência e a Data de Referência Anterior**	**Faixa de Atraso**
05/03/2000	**11/03/2000**	**-**	**-**
05/04/2000	**14/04/2000**	**6**	**1-30**
05/05/2000	**05/05/2000**	**9**	**1-30**
05/06/2000	**10/08/2000**	**0**	**0**
05/07/2000	**10/08/2000**	**30**	**1-30**
05/08/2000	**10/08/2000**	**60**	**31-60**
05/09/2000	**25/08/2000**	**66**	**61-90**
05/10/2000	**05/12/2000**	**0**	**0**
05/11/2000	**05/12/2000**	**30**	**1-30**
05/12/2000	**05/12/2000**	**60**	**31-60**
05/01/2001	**-**	**5**	**1-30**
05/02/2001	**-**	**-**	**-**

- Na primeira data de referência (05/03/2000) não se pode conhecer nenhum atraso, já que não houve nenhuma prestação vencendo anteriormente;
- Na segunda data de referência, (05/04/2000), houve um atraso de 6 dias corridos, referente à prestação de 05/03 que foi paga somente em 11/03, ou seja, houve atraso de 6 dias entre 05/04 e 05/03;
- Na terceira data de referência (05/05/2000), houve um atraso de 9 dias corridos, referente à prestação de 05/04 que foi paga somente em 14/04, ou seja, houve atraso de 9 dias entre 05/05 e 05/04;
- Na quarta data de referência (05/06/2000), não houve nenhum atraso em aberto entre essa data e a data anterior (05/05), ou seja, entre 05/06 e 05/05 não existiu nenhuma prestação em atraso;

- Na quinta data de referência (05/07/2000), houve um atraso de 30 dias corridos, referente à prestação de 05/06 que ainda não havia sido paga até 05/07, ou seja, houve atraso de 30 dias entre 05/07 e 05/06;
- Na sexta data de referência (05/08/2000), houve um atraso de 60 dias corridos, referente à prestação de 05/06 que ainda não havia sido paga em 05/08. No período pôde ser verificado outro atraso, de 30 dias corridos, referente à prestação de 05/07 que também ainda não havia sido paga em 05/08. O maior atraso existente entre a data de referência (05/08) e a data de referência anterior (05/07) é, portanto, de 60 dias corridos;
- Na sétima data de referência (05/09/2000), houve um atraso de 66 dias corridos, referente à prestação de 05/06 que foi paga somente em 10/08, ou seja, entre a data de referência (05/09) e a data de referência anterior (05/08). Outro atraso existente no período é de atraso, de 36 dias corridos, é referente à prestação de 05/07 que foi paga somente em 10/08. Ainda pôde ser verificado outro atraso, de 5 dias corridos, referente à prestação de 05/08 que foi paga somente em 10/08. O maior atraso existente entre a data de referência (05/09) e a data de referência anterior (05/08) é, portanto, de 66 dias corridos;
- Na oitava data de referência (05/10/2000), todas as prestações estavam em dia, não existindo nenhum atraso em aberto entre essa data e a data de referência anterior (05/09);
- Raciocínio análogo é utilizado nas demais prestações, ou seja, é calculado o maior atraso em aberto entre a data de referência e a data de referência anterior. É preciso estar atento a prestações não pagas até a data de geração do banco de dados. Prestações com datas de vencimento posteriores à data de geração dos dados não tem atraso conhecido, e prestações com datas de vencimento anteriores à data de geração mas não pagas até essa data tem o atraso contabilizado somente entre a data de vencimento de referência e a data de geração do

arquivo, já que, apesar de não se conhecer a data de pagamento, sabe-se que o pagamento não ocorreu até a data de geração dos dados.

Calculados os atrasos existentes em cada contrato de crédito analisado, o passo seguinte é gerar tabelas cruzadas relacionando a quantidade de contratos em cada faixa de atraso na data de vencimento da prestação *n* com a quantidade desses contratos que evoluiu para cada uma das faixas de atraso na data de vencimento da prestação *n + 1*, sendo *n > 2* já que não se conhecem os atrasos na data de vencimento da primeira prestação. A Tabela 5 contém a tabela cruzada relacionando os atrasos na data de vencimento da quarta prestação com os atrasos na data de vencimento da quinta prestação, com os valores reais obtidos para a carteira CAB.

Tabela 5. Transição de Faixa de Atraso na Data de Vencimento da 4ª Prestação para a Data de Vencimento da 5ª Prestação

		Atraso na Data de Vencimento da 5ª Prestação (dias corridos)					
		0	1 a 30	31 a 60	61 a 90	91 a 120	Total
Atraso na Data de Vencimento da 4ª Prestação (dias corridos)	0	108.369	10.774				*119.143*
	%	91,0	9,0				*100,0*
	1 a 30	7.384	9.875	2.485			*19.744*
	%	37,4	50,0	12,6			*100,0*
	31 a 60	398	693	654	436		*2.181*
	%	18,2	31,8	30,0	20,0		*100,0*
	61 a 90	31	38	26	32	116	*243*
	%	12,8	15,6	10,7	13,2	47,7	*100,0*
	Total	*116.182*	*21.380*	*3.165*	*468*	*116*	*141.311*
	%	*82,2*	*15,1*	*2,2*	*0,3*	*0,1*	*100,0*

Conforme observado na Tabela 5, dos 119.143 contratos que atingiram a quarta prestação (quarta data de vencimento) com atraso de 0 dia, 10.774 (9,0%) evoluíram para um atraso de 1 a 30 dias na data de vencimento da quinta prestação. Assim, 9,0% é a probabilidade de evolução

de um atraso nulo na quarta prestação para um atraso de 1 a 30 dias na quinta prestação. Dos 19.744 contratos que atingiram a quarta prestação com atraso de 1 a 30 dias, 2.485 (12,6%) evoluíram para um atraso de 31 a 60 dias na prestação seguinte. Entre os 2.181 contratos que atingiram a quarta prestação com atraso de 31 a 60 dias, 436 (20,0%) evoluíram para um atraso de 61 a 90 dias na prestação seguinte. Além disso, 47,7% dos contratos com atraso de 61 a 90 dias na quarta prestação evoluíram para atraso de 91 a 120 dias na quinta prestação. Observando as probabilidades, é nítido que, quanto maior for o atraso de um contrato na 4ª prestação, maiores são as chances desse contrato evoluir para um atraso ainda maior na prestação seguinte. Também é possível detectar as probabilidades de quitação de atrasos da quarta para a quinta prestação: 37,4% dos contratos que chegaram com atraso de 1 a 30 dias na quarta prestação evoluíram para um atraso nulo na quinta prestação, 18,2% dos contratos com atraso de 31 a 60 dias evoluíram para atraso nulo e 12,8% dos contratos com atraso de 61 a 90 dias quitaram suas prestações em atraso. Assim, quanto maior é o atraso ocorrido na 4ª prestação, menor é a chance desse atraso ser quitado pelo cliente.

A geração de uma tabela de transição de atraso como a Tabela 5 foi feita para cada dupla de prestações subsequentes (2ª para 3ª, 3ª para 4ª, 4ª para 5ª, até a 11ª para 12ª). Todas as probabilidades de evolução de uma faixa de atraso para a faixa de atraso seguinte, de uma prestação para a prestação seguinte, encontram-se na Tabela 6. A realização desse estudo considerou todas as 201.075 concessões de crédito do CAB. Dessas, 180.056 (89,5%) foram contratadas para serem pagas em até 12 prestações, sendo que somente 21.019 (10,5%) tinham prazo de 13 a 24 meses. Assim, apesar do estudo de qualidade de crédito ter considerado todos os 201.075 contratos de crédito, somente foram estudados os atrasos existentes até as datas de vencimento da décima segunda prestação de cada contrato, já que, a partir desse ponto, a

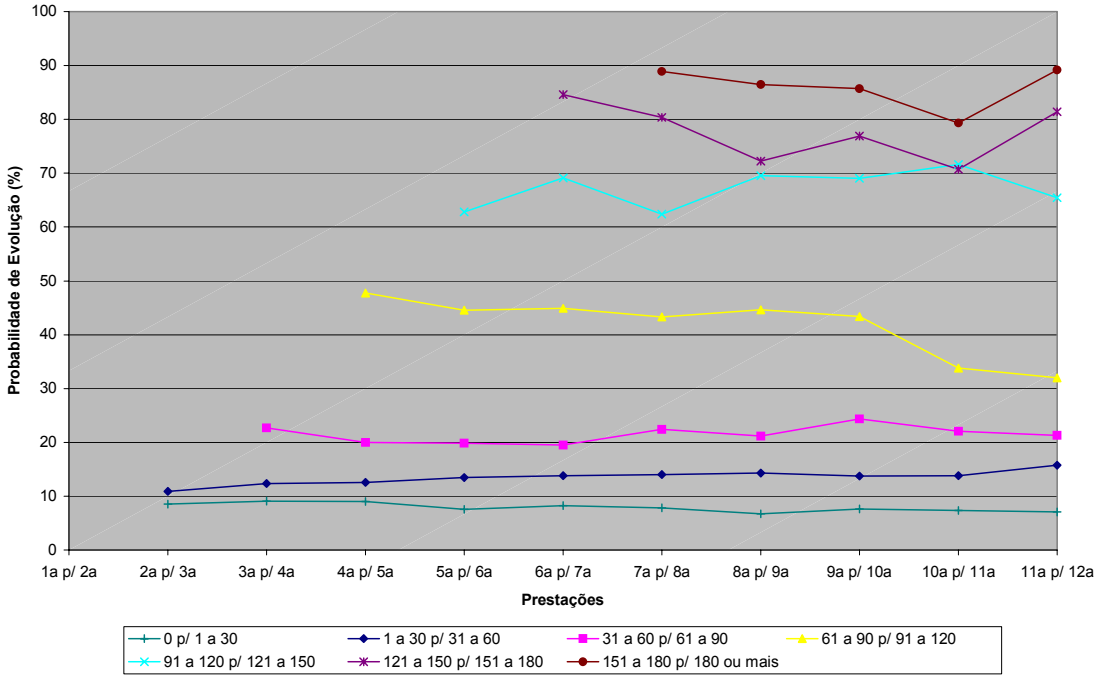
quantidade de operações de crédito cai abruptamente, o que resultaria em oscilações nos cálculos de probabilidades de evolução de atrasos, distorcendo os resultados.

Tabela 6. Probabilidades de Evolução de Faixas de Atraso por Prestações de Referência

Faixas de atraso (em dias corridos)	Probabilidades de evolução por prestações de referência (%)									
	2ª p/ 3ª	3ª p/ 4ª	4ª p/ 5ª	5ª p/ 6ª	6ª p/ 7ª	7ª p/ 8ª	8ª p/ 9ª	9ª p/ 10ª	10ª p/ 11ª	11ª p/ 12ª
0 p/ 1 a 30	8,6	9,1	9,0	7,6	8,2	7,8	6,8	7,7	7,3	7,1
1 a 30 p/ 31 a 60	10,9	12,4	12,6	13,5	13,8	14,0	14,3	13,8	13,8	15,7
31 a 60 p/ 61 a 90	-	22,7	20.0	19,9	19,5	22,4	21,2	19,2	22,0	21,3
61 a 90 p/ 91 a 120	-	-	47,7	44,6	44,9	43,3	44,6	43,4	33,8	32,0
91 a 120 p/ 121 a 150	-	-	-	62,8	69,1	62,4	69,5	69.0	71,6	65,5
121 a 150 p/ 151 a 180	-	-	-	-	84,6	80,4	72,2	76,9	70,7	81,4
151 a 180 p/ 181 ou mais	-	-	-	-	-	88,9	86,5	85,7	79,3	89,2

A observação da Tabela 6 permite uma constatação imediata: independente das prestações de referência, quanto maior é a faixa de atraso que um contrato alcança em determinada prestação, maior é a probabilidade desse atraso evoluir para um atraso maior na prestação seguinte (cada linha da Tabela 6 apresenta probabilidades maiores que da linha anterior). O Gráfico 1 permite uma melhor visualização dos dados da Tabela 6.

Gráfico 1. Probabilidades de Evolução de Atrasos por Prestações de Referência



O gráfico 1 traz em seu eixo horizontal as prestações de referência (por exemplo, "2ª p/ 3ª" significa evolução de atraso da data de vencimento da segunda para a data de vencimento da terceira prestação) e, no eixo vertical, as probabilidades de transição. Por exemplo, "0 p/ 1 a 30" significa evolução de um atraso de 0 dia corrido na prestação *n* para um atraso de 1 a 30 dias corridos na prestação *n+1*, e "1 a 30 p/ 31 a 60" significa evolução de um atraso de 1 a 30 dias corridos na prestação *n* para um atraso de 31 a 60 dias corridos na prestação *n+1*.

Os resultados do Gráfico 1 mostram que o comportamento de evolução de atrasos se manteve razoavelmente estável ao longo das prestações dos contratos da linha de crédito CAB. Contratos que chegam a determinada prestação com atraso nulo têm probabilidades baixas (menores que 10%, com média de 8%) de chegarem à prestação seguinte com atraso de 1 a 30 dias (linha verde), enquanto que contratos que atingem uma prestação com atraso de 1 a 30 dias

também mostram baixas probabilidades (entre 10% e 16%, com média de 13,5%) de atingirem a prestação seguinte com atraso de 31 a 60 dias (linha azul). Atrasos de 31 a 60 dias em uma prestação revelam probabilidades de 20% a 25% (com médias de 21%) de atingirem atrasos de 61 a 90 dias na prestação seguinte (linha rosa). No entanto, quando um contrato atinge atraso de 61 a 90 dias em uma prestação, as probabilidades de evolução para um atraso de 91 a 120 dias são substancialmente maiores, atingindo média de 42%, com pico de 47,7% da quarta para a quinta prestação e uma queda para 32% da 11ª para a 12ª prestação. As probabilidades aumentam mais substancialmente quando os atrasos atingem 91 a 120 dias (evoluindo para 121 a 150 dias com médias de 67% de probabilidade), 121 a 150 dias (evoluindo para 151 a 180 dias com médias de 78% de probabilidade), e assim por diante.

O comportamento de inadimplência da carteira CAB, de acordo com o Gráfico 1, é bastante nítido. Clientes que não têm atraso em determinada prestação têm baixa probabilidade de manifestar algum atraso na prestação seguinte. Então os clientes que não apresentam atrasos durante o vencimento das prestações podem ser considerados "bons clientes", ou seja, as operações de crédito sem nenhum atraso podem ser consideradas "boas", já que a grande maioria (cerca de 92%) se mantém sem atraso nas prestações seguintes. Clientes que apresentam atraso de 1 a 30 dias em uma prestação também têm baixa probabilidade de aumentar o atraso na prestação seguinte, e suas operações também podem ser consideradas "boas", já que a maioria (cerca de 86,5%) se mantém no mesmo patamar de atraso ou reduz o atraso. As operações que atingem 31 a 60 dias de atraso em uma prestação continuam manifestando baixas probabilidades de aumentarem o atraso na prestação seguinte e podem ser consideradas "boas", já que grande parte (cerca de 79%) se mantém no mesmo patamar de atraso ou o reduz.

No entanto, operações que atingem atrasos de 61 a 90 dias em uma prestação passam a apresentar probabilidades muito mais elevadas de ter atraso maior na prestação seguinte, já que, em média, 42% das operações nessa situação seguem para atraso maiores, e apenas pouco mais da metade (cerca de 58%) se mantém no mesmo patamar de atraso ou o reduz. Assim, o ponto crítico de evolução de inadimplência da carteira ocorre quando o atraso no pagamento de qualquer prestação atinge 61 ou mais dias. A partir desse ponto, as probabilidades dos atrasos aumentarem cada vez mais são substancialmente maiores quanto maior for o atraso atingido. Se cerca de 42% dos contratos com 61 a 90 dias de atraso atingem 91 a 120 dias de atraso em seguida, e cerca de 67% dos contratos que alcançam 91 a 120 dias de atraso evoluem para atrasos de 121 a 150 dias, e assim por diante, por encadeamento de probabilidades é simples auferir que cerca de 18,9% dos contratos que manifestam atraso de 61 a 90 dias em uma prestação atingem o atraso máximo (180 ou mais dias corridos) durante as prestações seguintes, ou seja, o atraso evolui até atingir a inadimplência absoluta. Atrasos de 31 a 60 dias manifestam apenas 3,9% de probabilidade de evoluírem para a inadimplência, enquanto que atrasos de 1 a 30 dias têm somente 0,5% de probabilidade de seguir o mesmo comportamento. Já atrasos de 91 a 120 dias apresentam 44,9% de probabilidade de evoluírem para a inadimplência.

Considerando o comportamento das probabilidades analisadas, as probabilidades de evolução para a inadimplência absoluta (180 ou mais dias de atraso) são de 0,5% para operações que atingem 1 a 30 dias de atraso, de 3,9% para aquelas que alcançam 31 a 60 dias, de 18,9% para as que manifestam atrasos de 61 a 90 e de 44,9% para operações com atrasos de 91 a 120 dias. Assim, é tolerável aceitar atrasos de até 60 dias como definidor das boas operações de crédito (dadas as baixas probabilidades dessas operações atingirem a inadimplência absoluta),

enquanto que atrasos de 61 ou mais dias definem as operações de crédito ruins, dado o súbita elevação na probabilidade dessas operações evoluírem para a inadimplência.

Portanto, as definições de qualidade de crédito estipuladas para os créditos analisados da carteira de crédito CAB foram:

- Boa operação de crédito: aquela que apresentou, no máximo, 60 dias corridos no pagamento de qualquer uma das prestações. Foi criada a variável "qualidade de crédito", à qual foi atribuído o valor 1 para as boas operações;
- Operação de crédito ruim: aquela que apresentou atraso de 61 ou mais dias corridos no pagamento de qualquer uma das prestações. Foi atribuído valor 0 para as operações ruins na variável "qualidade de crédito".

II.2. Categorização de Variáveis

A definição de qualidade de crédito realizada na seção anterior permitiu que fosse gerada a variável de resposta do modelo, ou seja, a variável qualidade de crédito (que assume valor 0 para créditos ruins e valor 1 para créditos bons). O objetivo, então, passa a ser encontrar as variáveis explicativas (cadastrais, financeiras, da operação etc.) mais relevantes para explicar a variável qualidade de crédito e conhecer os agrupamentos de respostas possíveis em cada variável explicativa que têm comportamentos homogêneos em relação à qualidade de crédito. Para tal, o procedimento técnico escolhido foi a categorização das variáveis explicativas através de um tipo de teste estatístico bastante atual e pouco conhecido, denominado CHAID (Chi-

Squared Automatic Interaction Detection), que nada mais é do que uma estatística χ^2 (qui-quadrado) para detectar comportamentos de homogeneidade entre variáveis.

Categorizar uma variável consiste em agrupar suas respostas em categorias de comportamento semelhante internamente e diferente das demais categorias de resposta. Nos modelos de *credit scoring*, as variáveis são agrupadas em categorias semelhantes quanto ao risco de crédito, ou seja, com relação à variável de resposta qualidade de crédito. A cada grupo (categoria, classe) de comportamento semelhante é dada uma nota de 1 a n (n = número de categorias resultantes), sendo criada uma nova variável (variável categorizada) que é então utilizada como variável explicativa.

Exemplo: o formato original da variável UF de nascimento no banco de dados era a sigla da UF. Cada cliente possuía uma sigla correspondente. No entanto, o teste estatístico CHAID mostrou que indivíduos de UFs diferentes tinham comportamento semelhante quanto à qualidade de crédito. Assim, dos 27 tipos de resposta da variável original (UF), a variável categorizada (categoria de UF) ficou com 13 grupos, visto que dentro de cada um deles a qualidade de crédito se mostrou bastante semelhante. A partir de então, a variável que passou a ser considerada foi categoria de UF, e não mais UF original.

A utilização de variáveis categorizadas tem algumas vantagens sobre a utilização das variáveis originais: o modelo resultante é mais simples (menor número de coeficientes estimados), maior poder de previsão do modelo resultante de categorias homogêneas internamente e a garantia de que indivíduos com qualidade de crédito equivalentes tenham "pesos" iguais no modelo.

II.2.1. O Método CHAID: Chi-Squared Automatic Interaction Detection

O método de categorização de cada variável, denominado CHAID, é um método explicativo para classificação de variáveis explicativas em grupos relevantes com relação a uma variável de resposta. O propósito do método é dividir um conjunto de objetos de tal forma que os subgrupos sejam significativamente diferentes com relação a um determinado critério. O critério é a variável binária dependente ou de resposta (qualidade de crédito), enquanto que o conjunto de objetos é formado pelas 201.075 operações de crédito do estudo e pelas variáveis explicativas, que são individualmente submetidas ao CHAID. Os segmentos (categorias) derivados pelo CHAID para cada variável são mutuamente exclusivos e exaustivos, o que significa dizer que cada resposta da variável está contida em uma única categoria (ex.: para a variável UF de nascimento, se a resposta "SP" estiver na primeira categoria resultante do CHAID, essa mesma resposta não estará contida em nenhuma outra categoria) e que todas as possibilidades de respostas encontradas na amostra para cada variável estão contidas em alguma categoria resultante no CHAID. De acordo com Magidson (1994), idealizador do método CHAID, a aplicação do método permite a classificação de novos objetos (operações de crédito) através do conhecimento das categorias das variáveis explicativas.

O CHAID se baseia na análise dos momentos das variáveis explicativas e da variável de resposta. A Tabela 7 é uma tabela de contingência em que Y é a variável dependente (de resposta) e X é a variável explicativa a ser categorizada, sendo que possíveis dependências entre as duas variáveis podem ser identificadas através do estudo de suas freqüências cruzadas. Se não há dependência entre as variáveis, então é esperado que a freqüência relativa da variável Y dentro de cada categoria da variável explicativa X corresponda às freqüências marginais de Y. No

exemplo da Tabela 7, é esperada uma freqüência condicional de Y_1 dado X_1 de 58% se considerada a distribuição marginal de Y (110/190 = 58%), valor diferente da freqüência condicional observada de 40% (40/100). Assim, as variáveis Y e X não são independentes.

Tabela 7. Exemplo de Tabela de Contingência

		Y (quantidade de casos)		
		Y_1	Y_2	Total
X (quantidade de casos)	X_1	40	60	100
	X_2	70	20	90
Total		110	80	190

O teste χ^2 compreendido pelo método CHAID acumula os desvios quadrados padronizados entre as freqüências observadas e as freqüências esperadas, sendo calculado pela seguinte fórmula:

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

em que i é a célula da tabela de contingência (na Tabela 7, i =4), O_i é a freqüência observada na célula e E_i é a freqüência esperada da célula.

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

Assim, valores elevados da estatística de teste indicam dependência (interações) entre as variáveis analisadas (variável de resposta vs. variável explicativa). No exemplo da Tabela 7, o valor da estatística do teste seria igual a:

$$\chi^2 = \frac{\left(40 - \frac{110}{190}\cdot 100\right)^2}{\frac{110}{190}\cdot 100} + \frac{\left(60 - \frac{80}{190}\cdot 100\right)^2}{\frac{80}{190}\cdot 100} + \frac{\left(70 - \frac{110}{190}\cdot 90\right)^2}{\frac{110}{190}\cdot 90} + \frac{\left(20 - \frac{80}{190}\cdot 90\right)^2}{\frac{80}{190}\cdot 90}$$

$$\chi^2 = 5{,}53 + 7{,}61 + 6{,}14 + 8{,}45 = 27{,}73$$

A validação dessa dependência é feita pela comparação da estatística de teste com o valor crítico da distribuição χ^2 determinado pelo nível de significância (usualmente 5%) e pelos graus de liberdade da estatística, que são iguais *(n-1)*(m-1)*, em que *n* é o número de colunas da tabela de contingência e *m* é o número de linhas. Se o valor da estatística de teste é maior que o valor crítico, então a hipótese nula de independência entre as variáveis não pode ser aceita. Assim, existe dependência entre as variáveis. No caso do exemplo da Tabela 7, o valor crítico do teste, a 5% de significância e com 1 grau de liberdade ((2-1)*(2-1)) é de 3,841. Já que o valor da estatística (27,73) foi maior que o valor crítico da distribuição χ^2, a hipótese nula de independência não deve ser aceita e, portanto, as variáveis *X* e *Y* podem ser consideradas dependentes, ou seja, tem relações de dependência estatisticamente fortes.

Na prática, a utilização da abordagem do CHAID ocorre quando estão presentes os seguintes componentes:

- Uma variável dependente categórica, ou seja, cujas respostas possíveis formem grupos distintos e mutuamente exclusivos, tal como a variável qualidade de crédito;
- Um conjunto de variáveis explicativas categóricas ou não. Algumas das variáveis explicativas são originalmente categóricas por definição (ex.: estado civil, residência própria etc.) e outras, basicamente variáveis contínuas, não são. No caso da variável ser contínua (ex.: salário líquido), o método CHAID realiza uma prévia e arbitrária transformação da variável em categorias, sendo necessário definir aleatoriamente a quantidade de categorias prévias desejada para a variável e a quantidade mínima de casos que deve existir em cada categoria. Por exemplo, no caso da variável explicativa salário líquido, o método divide as 201.075 respostas existentes em 100 categorias prévias, cada uma designando um intervalo de valor do salário líquido, e o método CHAID analisa as relações entre essas 100 categorias, procurando agrupar as categorias mais homogêneas e mantendo-as separadas das categorias heterogêneas.

A partir das categorias prévias (existentes por definição da variável ou arbitrariamente definidas no caso de variáveis contínuas) da variável explicativa submetida ao CHAID com relação à variável de resposta, o método agrupa as categorias homogêneas da variável explicativa submetida ao teste. Já que a variável explicativa pode conter um número grande de categorias prévias, a questão é identificar quais categorias prévias podem ser agrupadas. Para essa identificação, o CHAID gera uma tabela cruzada para cada par de categorias prévias da variável explicativa (ou seja, para cada combinação das categorias prévias duas a duas) em relação à variável de resposta. Essa tabela cruzada é semelhante à Tabela 7, com a diferença de que os totais computados são relativos somente ao par de categorias em questão, e não aos totais de

todas as categorias. É importante ressaltar que o método CHAID tem restrições de combinações dependendo do tipo de variável explicativa. Se for variável contínua (quantitativa) ou ordinal (quantitativa ou qualitativa), o método não testa combinações de categorias não adjacentes; no caso de ser variável nominal (qualitativa ou quantitativa) o método testa todas as combinações.

Em seguida, é calculada a estatística χ^2 e o *p-value* do teste para cada par de categorias em questão, ou seja, é feito um teste para cada tabela cruzada. Calculado o *p-value* de cada par de categorias prévias, o CHAID agrupa o par de categorias prévias que apresentar o maior *p-value* dentro da distribuição χ^2. É importante lembrar que, se a estatística χ^2 de uma tabela cruzada é estatisticamente significante (valor da estatística de teste maior que o valor crítico do teste), isso significa a não aceitação da hipótese nula de independência, o que implica dizer que deve existir uma relação entre as variáveis contidas na tabela cruzada e, portanto, o par de categorias prévias em questão não pode ser agrupado em relação à variável de resposta, já que esse par não revela independência (homogeneidade) com esta variável. Em outras palavras, um *p-value* maior que o nível de significância revela que as duas categorias são homogêneas e podem ser agrupadas em relação à variável de resposta.

Uma vez agrupado o par de categorias mais homogêneas detectado na etapa anterior, ou seja, o par com maior *p-value* da estatística χ^2, o procedimento recomeça com novas categorias (uma a menos que o número inicial de categorias prévias da variável explicativa, já que um par foi unido no passo anterior), sendo geradas as novas combinações de categorias, as novas tabelas cruzadas e os novos testes χ^2, detectando-se um novo par de categorias que podem ser agrupadas e recomeçando novamente, até o ponto que nenhuma das categorias restantes possa ser considerada homogênea, ou seja, nenhuma das categorias resultantes possa ser agrupada. Em outras palavras, a regra de parada do teste é quando nenhum dos *p-values* calculados em

determinada etapa de agrupamento é maior que 5%. É importante notar que o método CHAID permite que o resultado final da categorização seja uma única categoria contendo todas as respostas possíveis da variável explicativa com relação à variável de resposta. Isso acontece quando o teste atinge uma etapa contendo apenas duas categorias e o *p-value* do teste entre elas é maior que 5%. Nesse caso, as duas categorias ainda podem ser agrupadas, formando apenas uma categoria final, podendo-se dizer que a variável explicativa em questão não apresenta relação com a variável de resposta, pois uma única categoria final indica que todas as respostas possíveis da variável explicativa podem ser consideradas homogêneas em relação à variável de resposta.

Finalmente, é preciso observar que o método CHAID não requer a especificação de nenhuma forma funcional de relação entre as variáveis, o que torna o tipo de análise aplicável a uma diversa série de questões, tais como estudos na área médica, social, comercial etc. No entanto, o método requer uma amostra grande de observações de forma a obter resultados confiáveis, fato que não comprometeu sua aplicação no presente estudo, que conta com 201.075 observações de operações de crédito para análise. Além disso, é preciso ter em mente que o método CHAID foi usado no presente trabalho como uma ferramenta intermediária na geração do modelo final, ou seja, a categorização não é o resultado final do modelo de concessão de crédito.

II.2.2. Resultados

A Tabela 8 explicita as categorias resultantes de cada uma das 40 variáveis explicativas[4] recebidas para análise da carteira de crédito CAB, quando submetidas ao método CHAID (através do pacote estatístico *Answer Tree 2.0* do *SPSS 6.0*) com relação à variável qualidade de crédito (variável de resposta), e também explicita o percentual de operações com qualidade de crédito ruim (ou seja, de valor 0) dentro de cada categoria resultante. Por exemplo, para o salário líquido, o método CHAID resultou em 3 categorias: a categoria 1, que inclui as operações de crédito com resposta "missing" para o salário líquido (sendo que 1,70% das operações nessa categoria de salário líquido são ruins, para uma média de 1,40% de todas as operações), a categoria 2, que inclui as operações de clientes com salários de até R$ 500,00 (1,80% das operações são ruins), e a categoria 3, com operações de clientes com salários acima de R$ 500,00 (1,10% das operações são ruins). Assim, para a variável salário líquido, o CHAID detectou 3 categorias de comportamento homogêneo internamente e heterogêneo entre categorias.

É importante observar que o nível de significância adotado foi de 5% e que, portanto, todas as variáveis que apresentam 2 ou mais categorias tiveram *p-values* menores que 5% na última etapa do processo, indicando que as categorias resultantes são dependentes e, portanto, não poderia ser realizado nenhum outro agrupamento. Também deve-se notar que algumas das 40 variáveis explicativas tiveram como resultado final apenas uma categoria, casos em que ao *p-value* da comparação entre as duas categorias anteriores é maior que 5%, indicando que as duas categorias podem ser unidas por serem independentes, resultando em apenas uma categoria final.

[4] Foram recebidas 43 variáveis explicativas, sendo que 3 delas não são variáveis classificatórias mas sim somente para identificação do cliente e sua operação: nome do cliente, código do cliente e código da operação.

Se o resultado é uma única categoria, então qualquer valor que a variável em questão assuma não altera a qualidade de crédito do cliente e, portanto, a variável não pode ser utilizada para classificação das operações de crédito. Por exemplo, se a variável salário líquido tivesse como resultado apenas uma categoria, então um cliente com salário de R$ 200,00 seria classificado exatamente igual que um cliente com salário de R$ 3.000,00, ou seja, não interessaria o salário líquido do cliente para decidir sobre conceder ou não o crédito a ele.

É importante observar o resultado obtido em cada variável para poder verificar se o procedimento de categorização gerou resultados plausíveis. Tomando-se a variável idade, por exemplo, nota-se que todas as respostas possíveis puderam ser agrupadas em 7 categorias homogêneas internamente e diferentes entre si. Cada uma das 7 categorias apresentou um percentual de operações ruins diferente entre si e com comportamento decrescente: quanto maior é a idade, menor é o percentual de operações ruins dentro de cada categoria que contém essa idade, resultado perfeitamente factível visto que, quanto maior é a idade, maior é a renda e a capacidade de pagamento do cliente. A categoria *missing* apresentou o menor percentual de todas as categorias, mas esse resultado não ocorreria obrigatoriamente, pois não se sabe a priori o motivo dos clientes não terem respondido a idade. A variável UF de nascimento, por exemplo, apresenta como resultado 6 categorias finais. As 27 respostas originais possíveis foram testadas e o processo resultou em apenas 6 categorias. Assim, nota-se que pessoas nascidas na região norte e nordeste apresentaram inadimplência homogênea e maior que a de todos os demais estados (2,10%), enquanto que pessoas nascidas em SC, RS (região sul), AL, PI, RN, PB (região nordeste) e ES apresentaram a menor inadimplência (1,20%). Nascidos em CE, SE, BA, MG e RJ também tiveram comportamentos equivalentes (1,50%). Todos esses resultados dependem diretamente do tipo de carteira de crédito, não sendo esperado um determinado comportamento.

Tabela 8. Categorização das Variáveis Explicativas do CAB – Método CHAID

Variáveis Originais	*Categorização*		
	Resposta Original	*Resposta Categorizada (Código)*	*% de Operações Ruins (Total da Carteira: 1,40%)*
Residência Própria	Missing	1	1,10%
	Possui	2	1,20%
	Não Possui	3	1,80%
Idade	Missing	1	0,10%
	Até 24 anos	2	2,40%
	25 a 28 anos	3	1,90%
	29 a 38 anos	4	1,50%
	39 a 51 anos	5	1,20%
	52 a 58 anos	6	0,70%
	59 anos ou mais	7	0,50%
Quantidade de Dependentes Financeiros	Missing (Nenhum)	1	1,30%
	1	2	1,40%
	2 ou 3	3	1,50%
	4 a 6	4	1,60%
	7 ou mais	5	2,30%
Estado Civil	Divorciado ou Missing	1	0,20%
	Solteiro	2	1,70%
	Marital, Desquitado ou Separado Judicialmente (Consensual)	3	1,50%
	Viúvo ou Separado Judicialmente	4	0,60%
	Casado	5	1,30%
UF de Nascimento	Missing	1	0,80%
	SP	2	1,40%
	RR, AC, DF, TO, AM, RO, AP, PA, MA, MT, GO	3	2,10%
	SC, AL, RS, PI, RN, PB, ES	4	1,20%
	PR, MS	5	1,70%
	CE, SE, RJ, MG, BA	6	1,50%
Sexo	Missing	1	-
	Feminino	2	1,20%
	Masculino	3	1,50%
E-Mail	Missing	1	1,10%
	Possui	2	0,90%
	Não Possui	3	1,40%
Salário Líquido	Missing	1	1,70%
	Até R$ 500,00	2	1,80%
	R$ 500,01 ou mais	3	1,10%
Outras Rendas Mensais	Missing	1	1,40%
	R$ 0,01 ou mais	2	1,20%
Salário Líquido do Cônjuge	Missing	1	1,40%
	R$ 0,01 ou mais	2	1,10%

Tabela 8. Categorização das Variáveis Explicativas do CAB – Método CHAID (cont.)

Variáveis Originais	*Categorização*		
	Resposta Original	*Resposta Categorizada (Código)*	*% de Clientes Ruins (Total da Carteira: 1,40%)*
Tempo de Residência (*)	Missing	1	0,90%
	Nulo	2	1,20%
	1 a 8 meses	3	1,70%
	9 a 25 meses	4	1,50%
	26 a 132 meses	5	1,30%
	133 ou mais meses	6	1,10%
Idade na Data de Admissão no Atual Emprego (Data de Admissão Menos Data de Nascimento)	Missing	1	1,00%
	Até 24 anos	2	1,70%
	25 a 32 anos	3	1,50%
	33 a 38 anos	4	1,40%
	39 a 44 anos	5	1,30%
	45 anos ou mais	6	1,10%
Idade na Data de Admissão (Anos) Dividida Pela Idade Atual (Anos)	Missing	1	1,00%
	0,000001 a 0,619241	2	1,00%
	0,619242 a 0,913546	3	1,30%
	0,913547 a 1,000000	4	2,00%
Cartão de Crédito "Credicard"	Não Possui	1	1,40%
	Possui	2	0,90%
Quantidade de Cartões de Crédito	Nenhum	1	1,40%
	1	2	1,10%
	2	3	0,60%
	3 ou mais	4	0,90%
Encargos Mensais (Compromissos Financeiros) (**)	R$ 0,00 ou Missing	1	1,40%
	Até R$ 66,47	2	2,20%
	R$ 66,48 a R$ 209,00	3	1,40%
	R$ 209,01 ou mais	4	1,30%
Quantidade de Imóveis (informada pelo cliente, c/ ou s/ comprovação)	Nenhum ou Missing	1	1,40%
	1	2	1,30%
	2 ou mais	3	1,00%
Valor Comprovado de Imóveis	Nulo ou Missing	1	1,40%
	Até R$ 22.500,00	2	1,90%
	R$ 22.500,01 ou mais	3	0,90%
Quantidade Comprovada de Imóveis	Nenhum ou Missing	1	1,40%
	1	2	1,50%
	2 ou mais	3	0,90%
Valor de Automóveis (informado pelo cliente, c/ ou s/ comprovação)	Nulo ou Missing	1	1,40%
	Até R$ 6.000,00	2	1,60%
	R$ 6.000,01 ou mais	3	1,00%
Quant. de Automóveis (informada pelo cliente, c/ ou s/ comprovação)	Nenhum ou Missing	1	1,40%
	1	2	1,30%
	2 ou mais	3	1,10%
Quantidade Total de Seguros (***)	Nenhum ou Missing	1	1,40%
	1 ou mais	2	0,70%

Tabela 8. Categorização das Variáveis Explicativas do CAB – Método CHAID (cont.)

Variáveis Originais	*Categorização*		
	Resposta Original	*Resposta Categorizada (Código)*	*% de Clientes Ruins (Total da Carteira: 1,40%)*
Quantidade de Seguros de Automóvel	Nenhum ou Missing	1	1,40%
	1 ou mais	2	0,90%
Cheque (de qualquer inst. financeira ou banco)	Não Possui ou Missing	1	1,90%
	Possui	2	1,30%
Valor dos Cheques Devolvidos nos 6 Meses Anteriores à Concessão – Motivo 12 (Cheque s/ fundo - 2ª apresentação)	Nulo (Nenhum Cheque Devolvido)	1	1,30%
	R$ 0,01 a R$ 136,88	2	2,40%
	R$ 136,89 a R$ 257,00	3	3,00%
	R$ 257,01 a R$ 498,51	4	3,30%
	R$ 498,52 ou mais	5	3,50%
Valor Total dos Cheques Devolvidos nos 6 Meses Anteriores à Concessão (****)	Nulo (Nenhum Cheque Devolvido)	1	1,30%
	R$ 0,01 a R$ 139,50	2	2,30%
	R$ 139,51 a R$ 767,20	3	3,10%
	R$ 767,21 ou mais	4	4,20%
Profissão	Missing	1	2,80%
	Aposentado	2	1,20%
	Autônomo	3	1,50%
	Profissional Liberal	4	0,50%
	Funcionário Público	5	1,10%
	Desempregado	6	3,00%
	Funcionário de Empresa Privada	7	0,70%
	Do Lar	8	1,60%
	Outros	9	1,30%
Valor do Saldo Médio em Conta-Corrente (*****)	Missing (Não Correntista)	1	1,70%
	Até R$ 3,10	2	2,90%
	R$ 3,11 a R$ 7,82	3	1,20%
	R$ 7,83 a R$ 66,33	4	1,00%
	R$ 66,34 a R$ 129,47	5	0,80%
	R$ 129,48 a R$ 364,81	6	0,30%
	R$ 364,82 ou mais	7	0,20%
Valor do Saldo Médio de Aplicações (******)	Missing (Não Correntista)	1	1,70%
	Nulo	2	1,80%
	R$ 0,01 a R$ 23,04	3	1,50%
	R$ 23,05 a R$ 81,15	4	0,70%
	R$ 81,16 a R$ 8.591,11	5	0,20%
	R$ 8.591,12 ou mais	6	0,10%
Bloqueio na Emissão de Cheques	Missing	1	1,70%
	Bloqueado pelo Gerente	2	2,40%
	Não Bloqueado	3	0,90%
	Bloq. por Adesão a Pacote de Tarifas	4	1,70%

Tabela 8. Categorização das Variáveis Explicativas do CAB – Método CHAID (cont.)

Variáveis Originais	*Categorização*		
	Resposta Original	*Resposta Categorizada (Código)*	*% de Clientes Ruins (Total da Carteira: 1,40%)*
Percentual de Taxa de Juros da Operação de Empréstimo (taxa mensal nominal do contrato) (*******)	0,95% a 3,30%	1	0,10%
	3,31% a 3,90%	2	0,60%
	3,91% a 3,95%	3	0,00%
	3,96% a 4,50%	4	0,80%
	4,51% a 4,70%	5	1,90%
	4,71% a 5,00%	6	3,10%
	5,00% ou mais	7	8,50%
Valor da Operação de Empréstimo	Até R$ 630,00	1	1,50%
	R$ 630,01 a R$ 850,00	2	2,10%
	R$ 850,01 a R$ 1.500,00	3	1,30%
	R$ 1.500,01 ou mais	4	1,10%
Quantidade de Prestações da Operação de Empréstimo	1 a 5	1	0,60%
	6 ou 7	2	1,10%
	8 a 11	3	1,40%
	12	4	1,60%
	13 a 16	5	0,50%
	17 ou 18	6	1,30%
	19 ou mais	7	0,20%
Referência Monetária da Operação	Real	1	1,30%
	TR	2	3,50%
Forma de Pagamento de IOF da Operação	A Vista Deduzido do Bruto	1	1,40%
	Financiado	2	1,10%
Valor Total de Garantias da Operação (********)	Nulo	1	1,40%
	R$ 0,01 ou mais	2	0,40%
Quantidade de Avalistas da Operação de Empréstimo	Nenhum	1	0,00%
	1	2	1,40%
	2	3	1,30%
	3 ou mais	4	1,00%
Tempo de Residência (Meses) Dividido pela Idade (Meses)	0,00 a 1,00 ou Missing (todas as respostas possíveis)	1	1,40%
Quantidade de Seguros de Vida	1 ou mais ou Missing (todas as respostas)	1	1,40%
Tipo de Tomador do Empréstimo	Pessoa Física Comum ou Pessoa Física Func. Público (todas as respostas possíveis)	1	1,40%

(*) Tempo de residência no último (atual) endereço.
(**) Gastos mensais com educação, aluguel, empréstimos diversos, financiamento imobiliário e de veículos.
(***) Seguros de automóvel, vida e residencial.
(****) Motivos 12 e 13 (cheques sem fundos na 2ª apresentação e conta encerrada, respectivamente).
(*****) Saldo médio em conta nos três meses anteriores à data de concessão.
(******) Saldo médio de aplicações nos três meses anteriores à data de concessão. Somente correntistas podem ter aplicações na instituição.
(*******) Variável não utilizável no modelo, visto que o percentual de juros é definido após a aceitação do pedido de crédito.
(********) Nenhuma das garantias em questão depende da taxa de juros da operação de empréstimo. Não inclui avalistas.

II.2.3 Tabelas Cruzadas

A construção de tabelas cruzadas relacionando as variáveis categorizadas com a variável qualidade de crédito consiste apenas em uma ferramenta de análise descritiva para observar as relações entre as variáveis explicativas categorizadas e a variável de resposta do modelo. Assim, a observação de todas as tabelas cruzadas permite visualizar melhor o resultado da aplicação do método CHAID e, com isso, ter uma percepção mais apurada das características das operações de crédito analisadas.

Conforme observado na Tabela 8, um total de 37 variáveis explicativas foram categorizadas, sendo que outras 3 apresentaram como resultado apenas uma categoria e, portanto, passaram a ser desprezadas do estudo, já que não influenciam a qualidade de crédito das operações. Assim, foram geradas 37 tabelas cruzadas relacionando as variáveis explicativas com a qualidade de crédito. A Tabela 9 apresenta como exemplo a tabela cruzada da variável residência própria. As tabelas cruzadas de todas as 37 variáveis encontram-se no apêndice I.

Tabela 9. Qualidade de Crédito por Residência Própria

		Qualidade de Crédito		
Categoria		Ruim	Bom	Total
1	Missing	396	34995	35391
		1,10%	98,90%	100,00%
2	Sim	1295	106336	107631
		1,20%	98,80%	100,00%
3	Não	1056	56997	58053
		1,80%	98,20%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

O que se pode notar a partir da Tabela 9 é que a variável residência própria parece ser uma boa classificadora de operações em relação à qualidade de crédito, já que o percentual de créditos ruins dentro de cada categoria se mostrou homogêneo internamente (todas as pessoas que responderam "sim" tiveram comportamento semelhante de inadimplência) e estatisticamente diferente do percentual das demais categorias ("sim" ou "missing"). Assim, pode-se dizer que as operações de crédito ruins (ou seja, aquelas que apresentaram atraso de 61 ou mais dias no pagamento de qualquer prestação) se concentraram menos em clientes que não informaram se possuem ou não residência própria (entre as 35.391 operações na categoria 1 denominada missing, apenas 1,10% foram operações ruins, abaixo da média total da carteira de crédito CAB, que foi de 1,40%), enquanto que a maior concentração ocorre com os clientes que declararam não possuir residência própria (1,80% das 58.053 operações na categoria 3 foram operações ruins, acima da média de 1,40%). Também pode-se concluir que 1,20% das operações de clientes que declararam possuir residência própria tiveram qualidade de crédito ruim, índice abaixo da média da carteira (1,40%), mas um pouco mais elevado que os 1,10% de clientes que não declararam a resposta. Note que mais da metade dos clientes da carteira de crédito CAB possui residência própria, um índice notável se for lembrado que o CAB é uma linha de crédito direcionado ao público de rendas baixa ou média.

No entanto, mesmo que as tabelas cruzadas indiquem que alguma variável é boa discriminante entre operações de crédito boas e ruins, é possível que, na etapa de estimação da equação do modelo final de geração de *score* (vide capítulo III deste trabalho), essa variável seja não significativa estatisticamente devido à interação com outras variáveis que participarão da equação e que tenham poder de discriminação semelhante a variável em questão. Portanto, somente é possível dizer que uma variável é realmente boa discriminante entre bons e maus

pagadores quando sua presença na equação final se confirmar estatisticamente, ou seja, na etapa de estimação do modelo final.

Capítulo III. Construção do Modelo

Nos capítulos I e II foram discutidas as questões acerca de banco de dados para estudos de modelos de *credit scoring*, bem como a definição da qualidade de crédito – que gerou a variável de resposta (dependente) do modelo, a variável binária "qualidade de crédito" assumindo valores 0 ou 1 em grupos mutuamente exclusivos – e a categorização das variáveis explicativas (independentes). Assim, o problema em questão passa a ser a estimação de uma equação que reflita o relacionamento entre a variável dependente e as variáveis explicativas, de forma a gerar o *score* (nota) final das operações de crédito analisadas.

Nesse capítulo serão apresentados o tamanho da amostra para montagem da equação (extraída a partir da amostra total de 201.075 casos para estudo) e o procedimento estatístico utilizado, formado por dois tópicos principais: a regressão logística e o método *forward stepwise* para seleção de variáveis relevantes.

É importante notar que não serão discutidas detalhadamente questões como testes de hipóteses de coeficientes (todos os testes necessários foram realizados para obter um modelo mais eficiente, mas não são o foco da discussão), correlação entre variáveis explicativas (problema encarado de forma prática, apenas evitando-se de manter no modelo final variáveis explicativas muito correlacionadas entre si e com a variável dependente) e as estatísticas teóricas de mensuração da capacidade discriminatória do modelo (optou-se por analisar tal capacidade através de ferramentas pragmáticas mais compreensíveis por parte dos potenciais usuários de modelos de *credit scoring*). Todos esses aspectos foram levados em conta na montagem da

equação final do modelo, sendo recomendada, para maiores detalhes, a leitura do livro de Hosmer e Lemeshow (1989), que analisa detalhadamente cada aspecto da geração de um modelo de regressão logística.

III. 1. Tamanho da Amostra de Modelagem

Na construção do modelo, ou seja, da equação que gera o *score* das operações de crédito, não foi utilizada toda a base de dados de 201.075 concessões. Além de uma amostra menor ser suficiente para a construção dos modelos de *credit scoring*, é interessante guardar observações não utilizadas na modelagem para se avaliar a qualidade dos modelos desenvolvidos.

O tamanho da amostra escolhida contém o mesmo número de casos de operações boas e ruins. Dentre os 201.075 casos, observou-se que cerca de 1,37% dos casos eram operações ruins (qualidade de crédito = 0), ou seja, 2.747 casos. Desses casos, cerca de 15% foram sorteados aleatoriamente e mantidos guardados para testes, tendo restado em torno de 85%, ou seja, 2.351 casos, para serem utilizados na montagem da equação. Escolhido o número de casos ruins, foram sorteados, também aleatoriamente, 2.351 casos de bons clientes. Assim, a amostra utilizada para montagem da equação continha 4.702 casos, cada um com todas as 37 variáveis explicativas categorizadas. É importante observar que não existe um "tamanho ideal" para amostra de modelagem, mas deve-se manter em mente que a quantidade de casos deve ser grande o suficiente para incluir operações de crédito representativas de todas as operações, ou seja, com todas a diversidade de respostas existentes no banco de dados inicial. A amostra de 4.702 casos, considerada de bom tamanho, foi suficiente para cumprir tal exigência. A quantidade de casos ruins e bons na amostra de modelagem não necessariamente deve ser igual, opção que foi

adotada no presente trabalho para que os coeficientes estimados tivessem participação equivalente de operações boas e ruins em seus cálculos, não sendo viesados em direção a nenhum dos dois grupos e gerando intervalos de confiança proporcionais para os resultados. A tabela 10 resume as informações de sorteio de amostra.

Tabela 10. Tamanho da Amostra de Modelagem

	Banco de Dados Inicial = 201.075 casos					
	% no Banco de Dados Inicial	Número de Casos no Banco de Dados Inicial	Sorteio da Amostra	Número de Casos na Amostra	% da Amostra	Número de Casos Guardados para Testes
Operações Boas	98,63	198.328	Mesmo número de casos de clientes ruins sorteados	2.351	50,00	195.977
Operações Ruins	1,37	2.747	85 % do número inicial de casos	2.351	50,00	396
Total	100,00	201.075	-	4.702	100,00	196.373

III.2. Método Estatístico Utilizado

As técnicas mais utilizadas para construção de modelos de *score* têm sido, historicamente, a regressão linear e a análise discriminante. Suas vantagens são a simplicidade conceitual e a presença dessas técnicas na maioria dos *softwares* estatísticos. Esses métodos combinam os coeficientes com as respostas das variáveis explicativas de forma que gerem contribuições individuais que, somadas, resultam no *score* final. Outras técnicas também

utilizadas são a regressão logística, modelos *probit*, métodos de programação matemática e redes neurais, entre outros[5].

O método de análise discriminante é o maior alvo de críticas dos estudiosos, basicamente com relação à hipótese de que as variáveis que explicam a variável de resposta seguem uma distribuição normal multivariada. No caso em que todas as variáveis explicativas são variáveis categorizadas, essa hipótese obviamente é descumprida. Os críticos do método ressaltam que o descumprimento da hipótese de normalidade não é uma limitação forte, mas ressaltam que ela é vital para os testes de hipóteses dos coeficientes estimados, que são necessários em qualquer técnica de estimação.

A regressão linear é outro método bastante utilizado na formulação de modelos de *credit scoring* com resposta do tipo binária (bom/ruim). Apesar de gerar estimadores não viesados e consistentes, esse método apresenta problemas de heterocedasticidade (já que a variância dos resíduos depende dos valores das variáveis explicativas, ou seja, não é constante), e a principal limitação é muito clara: os valores estimados para a variável de resposta não pertencem ao intervalo [0,1], podendo assumir valores negativos e até mesmo maiores que 1, ou seja, não condizentes com o problema estudado, em que o intervalo deve ser obedecido.

O método de regressão logística é, por definição, apropriado para estudos em que a variável de resposta assume valores 0 ou 1 (como a qualidade de crédito), e formula uma equação de relação não linear entre as variáveis explicativas e a variável de resposta, devido à forma funcional do método, que apresenta funções exponenciais relacionando as variáveis explicativas com a variável de resposta. Assim, o método parece mais apropriado que a regressão linear. As críticas com relação a esse método não são relacionadas a problemas intrínsecos ao

[5] Hand, D. J.; Henley; W.E., 1997, op. cit.

método (como ocorre com a regressão linear), mas sim aos resultados observados na prática que, em alguns casos, seriam muito próximos daqueles obtidos por regressão linear. No entanto, essa proximidade de resultados somente ocorre quando uma grande proporção dos valores estimados para a variável de resposta pertence a um intervalo razoavelmente distante de 0 e 1 (por exemplo, quando as estimativas assumem valores entre 0,3 e 0,7 aproximadamente), situação na qual a curva logística é bastante próxima a uma linha reta.

O método de regressão logística foi escolhido pelo fato do problema em questão ter uma variável dependente (qualidade de crédito) binária, o que torna o método mais apropriado que os demais, além do fato de ser computacionalmente simples. No entanto, em geral não se pode afirmar qual é o melhor método. Isso depende do problema estudado, da estrutura de dados, das variáveis explicativas disponíveis (inclusive a quantidade de variáveis) e o objetivo da classificação (inadimplência, lucratividade etc.). Também devem ser consideradas a velocidade do processo de classificação (quanto tempo o cliente que pede um empréstimo precisa esperar para ter uma resposta afirmativa ou negativa sobre a concessão) e a facilidade de revisar o modelo periodicamente. Uma boa velocidade de classificação é muito mais atraente para a pessoa que pede crédito do que se esta tiver que esperar vários dias para obter uma resposta (assim, a quantidade de clientes que desistem de tomar o empréstimo na instituição pela demora do processo é menor). A facilidade de revisar o modelo periodicamente é importante porque, visto que o modelo é baseado em concessões passadas, o modelo somente será robusto enquanto as características dos novos tomadores de crédito forem semelhantes às características encontradas no passado, ou seja, se a população da carteira se mantiver estável. No entanto, em algum momento a população certamente se modificará (até mesmo por razões naturais, como ciclos econômicos que afetam os salários das pessoas), então será preciso gerar um novo modelo

baseado na nova população. Além desses fatos, métodos de classificação fáceis de compreender (tais como métodos de regressão) são mais interessantes do que métodos menos transparentes (tal como redes neurais, programação matemática etc.) tanto para os clientes quanto para os profissionais que utilizarão o modelo, e possibilitam explicações imediatas e claras sobre a forma que o resultado final é alcançado.

O fato do problema de análise de concessão de crédito já ter sido bem compreendido e estudado pelos formuladores de métodos de classificação torna improvável o aparecimento de novas metodologias que consigam melhorar ainda mais a eficiência da classificação. Entre os métodos existentes, praticamente não existem diferenças nos resultados obtidos. Segundo Davis, Edelman e Gammerman (1992), "todos os métodos apresentam o mesmo nível de eficiência na classificação, mas os algoritmos de redes neurais levam muito tempo para ser elaborados". É mais provável que os novos estudos sejam baseados na inclusão de variáveis explicativas novas e mais preditivas e em questões ainda pouco exploradas, como por exemplo a montagem de um modelo de *credit scoring* que não tenha o simples objetivo de aprovar ou reprovar um crédito, mas que também estime limites de crédito para os clientes.

III.2.1. Regressão Logística

O objetivo de qualquer técnica de construção de modelos estatísticos é encontrar uma forma funcional adequada e parcimoniosa para descrever o relacionamento entre uma variável de resposta (dependente) e um conjunto de variáveis independentes (explicativas). O modelo de regressão logística assume a existência de uma variável dependente dicotômica Y com esperança condicional $E[Y/X]$, em que Y é o valor variável dependente e X é o conjunto de valores

assumidos pelas variáveis explicativas X_1, X_2, ..., X_j, sendo $E[Y/X]$ interpretada como o valor esperado de Y dados os valores de X e representada pela seguinte função:

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

Na função acima, por comodidade renomeada para π(X), o termo e representa a função exponencial e β_0, β_1, ..., β_j são os parâmetros a serem estimados. Essa função assume valores somente no intervalo [0,1] e é não-linear em seus parâmetros. O valor da variável dependente Y é então dado por:

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

O termo ε é o chamado de erro e representa a diferença entre o valor observado de Y e a esperança condicional de Y dado X. Sendo Y uma variável dicotômica que assume apenas valores 0 ou 1, então ε pode assumir dois valores: se $Y = 1$, então $\varepsilon = 1 - \pi(X)$ com probabilidade igual a $\pi(X)$ e, se $Y = 0$, $\varepsilon = -\pi(X)$ com probabilidade igual a $1 - \pi(X)$. Assim, o erro ε tem distribuição com média zero e variância igual a $\pi(X)[1 - \pi(X)]$, ou seja, a distribuição condicional da variável dependente Y é binomial com probabilidade dada pela esperança condicional $\pi(X)$.

Os parâmetros desconhecidos do modelo e que precisam ser estimados são β_0, β_1, ..., β_j. Esses parâmetros são estimados pelo método de máxima verossimilhança. Se a variável de resposta Y assume valores 0 ou 1, então a expressão $\pi(X)$ é a probabilidade condicional de $Y = 1$ dado X, ou seja, $\pi(X) = P(Y = 1 / X)$ e a probabilidade condicional de $Y = 0$ dado X é igual a $1 - \pi(X)$, ou seja, $1 - \pi(X) = P(Y = 0 / X)$. Então, é formulada uma função de verossimilhança, que

expressa as probabilidades das respostas observadas (que são independentes) em função dos parâmetros desconhecidos, dada por $l(\beta)$:

Na função de verossimilhança $l(\beta)$, n é o número de casos estudados (tamanho da amostra para estimação do modelo), sendo $\pi(X_i)$ a probabilidade condicional de cada observação Y_i (i = 1, ..., n) dados os valores das variáveis explicativas X associados a cada observação, ou seja, X_i.

$$l(\beta) = \prod_{i=1}^{n} \pi(X_i)^{Y_i} \cdot [1 - \pi(X_i)]^{1-Y_i}$$

Pelo princípio do método de máxima verossimilhança, os valores estimados de $\beta_0, \beta_1, ..., \beta_j$ são aqueles que maximizam a função $l(\beta)$. No entanto, matematicamente é mais simples trabalhar com o logaritmo natural de $l(\beta)$, conhecido como função de log-verossimilhança $L(\beta)$, já que os resultados da maximização de $L(\beta)$ são exatamente os mesmos que ocorreriam com a maximização de $l(\beta)$. Assim:

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

Os valores estimados de $\beta_0, \beta_1, ..., \beta_j$ são aqueles que maximizam $L(\beta)$ e são encontrados diferenciando-se $L(\beta)$ em relação a cada um dos parâmetros $\beta_0, \beta_1, ..., \beta_j$ e igualando as expressões resultantes a zero. A diferenciação gera as seguintes expressões, conhecidas como equações de verossimilhança:

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

e

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

em que $j = 1, 2, ..., p$

Na segunda equação de verossimilhança acima, *j* representa cada uma das *p* variáveis explicativas contidas em X_i. Na realidade, quando as variáveis de X_i são todas categorizadas, então elas são representadas por variáveis *dummies* (variáveis que assumem valores 0 ou 1 dependendo da categoria da variável explicativa). Por exemplo, se X_i contém 3 variáveis categorizadas (s=sexo, a=idade e r=renda), cada uma com 3 categorias, então devem ser consideradas na segunda equação de verossimilhança a existência de 6 variáveis *dummies* (2 *dummies* para cada variável, representando duas das categorias das variáveis, sendo que a categoria restante não requer uma variável *dummy* pois é uma combinação linear das 2 outras *dummies*). Nesse caso, p = 6 (apesar de existirem 3 variáveis, elas são representadas por 2 variáveis *dummies* cada), já que precisa ser estimado um coeficiente para cada variável *dummy*. A Tabela 11 mostra um exemplo de variáveis *dummies* para as 3 variáveis exemplificadas.

Tabela 11. Exemplo de Variáveis Dummies para Sexo (s), Idade (a) e Renda (r)

		Variáveis Dummies					
Variável Explicativa	Categoria	D_{1s}	D_{2s}	D_{1a}	D_{2a}	D_{1r}	D_{2r}
Sexo	1 – Masculino	1	0	-	-	-	-
	2 – Feminino	0	1	-	-	-	-
	3 – Missing	0	0	-	-	-	-
Idade	1 – Até 30 anos	-	-	1	0	-	-
	2 – De 31 a 45 anos	-	-	0	1	-	-
	3 – Acima de 45 anos	-	-	0	0	-	-
Renda	1 – Até R$ 500,00	-	-	-	-	1	0
	2 – De R$ 500,01 a R$ 1.000,00	-	-	-	-	0	1
	3 – Acima de R$ 1.000,0	-	-	-	-	0	0

Na Tabela 11, tomando-se a variável sexo, a *dummy* 1 de sexo (D_{1s}) assume valor 1 quando o sexo é masculino e 0 caso contrário, a *dummy* 2 de sexo (D_{2s}) assume valor 1 quando o sexo é feminino e 0 caso contrário. Quando o sexo é "missing", as *dummies* D_{1s} e D_{2s} assumem valor 0. Assim, cada variável tem suas categorias devidamente identificadas por variáveis *dummies* e o método de máxima verossimilhança estimará um coeficiente para cada *dummy* (β_{1s}, β_{2s}, β_{1a}, β_{2a}, β_{1r} *e* β_{2r}) além do coeficiente β_0. No processo de criação de *dummies* utilizado, sempre foi deixada a última categoria de cada variável para ser representada pela combinação linear das *dummies* das demais categorias da variável. É importante salientar que existem diferentes procedimentos estatísticos sobre o processo de criação de variáveis *dummies*, que não fazem parte do escopo do presente trabalho, e maiores detalhes podem ser encontrados no livro de Hosmer e Lemeshow (1989). Note que quando as variáveis de X_i são categorizadas, então $\pi(X_i)$ contido nas equações de verossimilhança deve ser escrito como:

$$\pi(X_i) = \frac{e^{\beta_0 + (\beta_{1,X1} D_{1,X1i} + \ldots + \beta_{k,X1} D_{K,X1i}) + (\beta_{1,X2} D_{1,X2i} + \ldots + \beta_{k,X2} D_{k,X2i}) + \ldots + (\beta_{1,Xp} D_{1,Xpi} + \ldots + \beta_{k,Xp} D_{k,Xpi})}}{1 + e^{\beta_0 + (\beta_{1,X1} D_{1,X1i} + \ldots + \beta_{k,X1} D_{K,X1i}) + (\beta_{1,X2} D_{1,X2i} + \ldots + \beta_{k,X2} D_{k,X2i}) + \ldots + (\beta_{1,Xp} D_{1,Xpi} + \ldots + \beta_{k,Xp} D_{k,Xpi})}}$$

em que $\beta_{k,Xp}$ representa o coeficiente associado à *k*-ésima categoria da variável explicativa *p* e $D_{k,Xpi}$ representa o valor da variável *dummy* do indivíduo *i* para a *k*-ésima categoria da variável explicativa *p*.

As soluções das equações de verossimilhança geram as estimativas de todos os coeficientes β, e requerem um grande esforço matemático e computacional, já que tais equações são não lineares em relação ao conjunto de coeficientes β a serem estimados. Os principais

softwares estatísticos solucionam essas equações e fornecem as estimativas finais dos coeficientes β, sendo que tais estimativas são chamadas de estimadores de máxima verossimilhança e podem ser e são representadas por β^*. O símbolo * é usado para representar os estimadores de um determinado termo. Por exemplo, $\pi^*(X_i)$ é o estimador de máxima verossimilhança de $\pi(X_i)$, obtido substituindo-se os valores estimados de β (ou seja, β^*) na expressão de $\pi(X_i)$. O valor de $\pi(X_i)$ fornece a estimação da probabilidade condicional de Y ser igual a 1, dados os valores das variáveis explicativas X_i, e, portanto, representa o resultado final do modelo.

Assim, para cada operação de crédito i contida na amostra de modelagem, Y^*_i é a probabilidade da variável de resposta (qualidade de crédito) ser igual a 1 para essa operação i, dados os valores das variáveis explicativas X_i. O *score* da operação de crédito nada mais é do que a multiplicação de Y^*_i por 100. Assim, se uma operação i tiver $Y^*_i = 0{,}75$, o *score* dessa operação é igual a 75, ou seja, a probabilidade dessa operação ter uma qualidade de crédito igual a 1 (boa qualidade de crédito) é de 75% dadas as características dessa operação.

III.2.2. Método de Escolha de Variáveis Explicativas – Forward Stepwise

A variável de resposta (dependente) do modelo de *credit scoring* é a qualidade de crédito da operação de empréstimo, a qual se pretende classificar a partir do conjunto de 37 variáveis explicativas categorizadas. Na realidade, essas 37 variáveis explicativas são apenas potenciais variáveis independentes do modelo, já que nem todas necessariamente fornecem informações relevantes para obter a qualidade de crédito. Assim, é necessário escolher, dentro desse grupo de

37 variáveis, somente aquelas mais significativas para explicar a variável de resposta. Quando se conhece exatamente as relações teóricas entre as variáveis do problema em questão (por exemplo, quando é sabido previamente quais determinadas variáveis explicativas são relacionadas com a qualidade de crédito), é possível simplesmente estimar o modelo "forçando" a participação apenas dessas determinadas variáveis e testando os coeficientes estimados e o poder de classificação do modelo. No entanto, quando o problema é relativamente desconhecido (não se conhecem seguramente as variáveis explicativas relevantes), é preciso adotar algum procedimento técnico de escolha de variáveis. O método de escolha de variáveis adotado no presente trabalho é denominado *forward stepwise*.

O método *stepwise* é usado mais freqüentemente em situações nas quais as variáveis independentes importantes não são conhecidas e suas associações com a variável de resposta não são bem compreendidas. Tais situações são muito comuns em estudos em que a variável de resposta é um assunto relativamente novo. Nessas situações, são colhidas muitas possíveis variáveis explicativas sobre as quais tenta-se examinar as associações mais significativas com a variável de resposta. A utilização de métodos *stepwise* possibilita uma forma rápida e efetiva para examinar um grande número de variáveis e, simultaneamente, examinar diversas equações de regressão logística possíveis a partir dessas variáveis. Existem algumas variantes entre os métodos *stepwise*, mas basicamente as possibilidades são duas: *backward stepwise* ou *forward stepwise*, cujas diferenças são apenas simples modificações de um algoritmo básico. Resumidamente, o método *backward stepwise* parte de um modelo inicial com todas as possíveis variáveis, que vão sendo eliminadas a cada passo até atingir um modelo final, sendo que no método *forward stepwise* se inicia com um modelo sem nenhuma variável explicativa e a cada passo são incluídas as variáveis relevantes, até a obtenção do modelo final.

A inclusão ou exclusão de variáveis explicativas relevantes é baseada num algoritmo estatístico que detecta a importância das variáveis baseado em medidas de significância estatística dos coeficientes dessas. No caso de regressão logística, em que os erros têm distribuição binomial, então a significância de cada variável é analisada com base em testes de razão de verossimilhança qui-quadrado (χ^2). Assim, em qualquer passo do procedimento, a variável explicativa mais "importante" é aquela que produz a maior variação na função de log-verossimilhança $L(\beta)$ em relação a um modelo que não contenha essa variável.

Os passos do procedimento *forward stepwise* estão resumidos a seguir. O *software* SPSS tem a opção "regressão logística, método forward stepwise", que foi utilizada no estudo de *credit scoring*. A utilização do método *backward* também é possível, mas freqüentemente os resultados obtidos são idênticos aos da opção *forward*.

Passo 0: Suponha a existência de p possíveis variáveis explicativas. O passo inicial computa um modelo contendo somente uma constante (β_0) e sua função de log-verossimilhança L_0. Em seguida, são computados p modelos univariados contendo a constante mais uma variável explicativa (p modelos para p variáveis explicativas). Para cada um dos p modelos contendo a variável X_j, é computada a função de log-verossimilhança L_j^0 e o teste de razão de verossimilhança determinado por $G_j^0 = 2(L_j^0 - L_0)$, com distribuição $\chi 2$ com $k-1$ graus de liberdade (K = número de categorias da variável em questão). A variável mais importante é aquela cujo *p-value* do teste é o menor, e é denominada X_j^{0*}. No entanto, o fato de uma variável ter o menor *p-value* do teste não garante que ela é estatisticamente significante. Por exemplo, se o menor *p-value* for de 0,75 (maior que o nível de significância de inclusão de variáveis, P_i), o estudo não prosseguiria porque a variável detectada como a mais importante não seria

relacionada com a variável de resposta. Se o menor *p-value* estiver abaixo do nível de significância de inclusão (P_i), então o processo segue para o passo 1.

Passo 1: Esse passo começa computando um modelo contendo a constante mais a variável escolhida no passo anterior, X_j^{0*}, e sua função de log-verossimilhança L_j^{0*}. Agora existem p-1 possíveis variáveis explicativas. Então são computados p-1 modelos contendo, cada um, a constante, a variável X_j^{0*} e mais uma variável explicativa (p-1 modelos para p-1 variáveis explicativas), e suas respectivas funções de log-verossimilhança. São realizados os testes de razão de verossimilhança para os p-1 modelos, dados por $G_j^1 = 2(L_j^1 - L_j^{0*})$. A variável mais importante do passo 1 é aquela cujo *p-value* do teste é o menor, e é denominada X_j^{1*}. Se esse menor *p-value* for inferior P_i, então o procedimento segue para o passo 2 contendo a constante, a variável X_j^{0*} e a variável X_j^{1*}. Se o menor *p-value* for superior a P_i, então o procedimento termina com um modelo contendo apenas constante e X_j^{0*}.

Passo 2: Essa etapa começa computando um modelo contendo a constante, X_j^{0*} e X_j^{1*}, e sua função de log-verossimilhança L_j^{1*}. Uma vez que X_j^{1*} foi adicionada no passo 1, então é possível que X_j^{0*} não seja mais importante. Assim, esse passo inclui testes para eliminação de variáveis. Esses testes são feitos computando-se modelos excluindo cada uma das variáveis adicionadas em passos anteriores. Assim, são calculadas as funções de log-verossimilhança de um modelo removendo X_j^{0*} e de um modelo removendo X_j^{1*}, denominadas L_{-j}^{2*}. São então realizados os mesmos testes de razão de verossimilhança, comparando-se os modelos sem cada uma das variáveis X_j^{0*} e X_j^{1*} ao modelo completo (contendo as duas variáveis mais a constante), ou seja, os testes são $G_{-j}^2 = 2(L_j^{1*} - L_{-j}^{2*})$. A variável que deve ser eliminada nesse passo é aquela cujo

teste fornecer o maior *p-value*. Esse maior *p-value* deve ser comparado ao nível de significância de exclusão (P_e) e, se for maior que P_e, então essa variável deve ser excluída. Se o maior *p-value* for menor que P_e, então nenhuma variável deve ser excluída. Após decidir sobre exclusão (ou não) de variáveis no passo 2, o procedimento continua, no mesmo passo, com o processo de inclusão de variáveis, computando os modelos adicionando-se cada uma das variáveis explicativas ainda não adicionadas no modelo e realizando os mesmos testes de razão de verossimilhança. Se mais nenhuma variável puder ser incluída (o *p-value* mínimo dos testes for maior que P_i), então o procedimento termina. Caso contrário, prossegue-se com a variável incluída no passo 0 (X_j^{0*}), no passo 1 (X_j^{1*}), sem a variável excluída no passo 2 e com a variável incluída no passo 2 (X_j^{2*}).

Passo 3: Esse passo é idêntico ao passo 2, ou seja, o procedimento testa exclusão de variáveis seguida de testes para inclusão de variáveis. O processo continua até o passo terminal, o passo *T*.

Passo T: Esse passo terminal ocorre quando: a) todas as *p* variáveis estiverem no modelo ou b) todas as variáveis presentes no modelo tiverem *p-value* para exclusão menor que P_e e as variáveis não presentes no modelo tiverem *p-value* para inclusão maior que P_i.

Conforme pôde ser constatado, a regra de decisão do processo é baseada em P_i, a probabilidade de inclusão, e P_e, a probabilidade de exclusão. Conforme Hosmer e Lemeshow[6], muitos estudos sobre procedimentos *stepwise* mostraram que adotar $P_i = 0{,}05$ (o valor padrão em

[6] op. cit.

estudos estatísticos) é uma opção muito restritiva que freqüentemente causa a ausência de variáveis explicativas importantes. A recomendação é que seja adotado P_i de 0,15 a 0,20, o que garantiria a presença de variáveis importantes e garantiria a seleção de variáveis com coeficientes significativamente diferentes de zero. O mesmo vale para a probabilidade de exclusão, P_e. No entanto, o valor escolhido de P_e deve ser maior que o valor de P_i para evitar a possibilidade de incluir uma determinada variável em certo passo e eliminá-la no passo subsequente. Tendo em vista essas observações, o modelo de *credit scoring* elaborado utilizou $P_i = 0,15$ e $P_e = 0,20$.

As principais críticas sobre métodos *stepwise* se concentram na observação de que seu uso significa assumir ignorância em relação ao fenômeno que está sendo estudado e ao fato de exigir grande esforço computacional. No entanto, o procedimento é muito bom quando o fenômeno em estudo é relativamente desconhecido (estudos estatísticos sobre concessão de crédito no Brasil ainda são recentes, apesar dos provedores de crédito afirmarem conhecer as variáveis importantes mas com base em julgamentos pessoais) e permite que sejam inclusas no modelo inicial do passo 0 as variáveis cuja relação com a variável de resposta é reconhecidamente relevante. Além disso, o fato do procedimento ser condicional por construção (por exemplo, a segunda variável a entrar no modelo é a segunda mais "importante", dado que a primeira já está inclusa no modelo) as correlações entre as variáveis explicativas são levadas em conta, evitando a existência de problemas de multicolinearidade no modelo estimado.

Capítulo IV. Resultados

Nesse capítulo encontram-se apresentados os resultados finais da estimação do modelo de *credit scoring* através da regressão logística pelo método *forward stepwise* com a variável qualidade de crédito sendo a variável dependente do modelo, conforme discutido no capítulo anterior. Esses resultados compreendem a apresentação das variáveis explicativas participantes do modelo final, os coeficientes estimados dessas variáveis e os testes de hipóteses realizados, um exemplo de aplicação do modelo para gerar o *score* da operação de crédito, uma análise gráfica das distribuições dos *scores* estimados em todo o banco de dados com 201.075 operações de crédito, o estudo de estabilidade do modelo e uma discussão sobre a escolha do *score* de corte para a carteira de crédito CAB.

IV.1. Modelo Final

A utilização da regressão logística através do método *forward stepwise* de seleção de variáveis explicativas relevantes para determinar a qualidade de crédito (variável de resposta) resultou num modelo de *credit scoring* para a linha de crédito CAB contendo 13 variáveis explicativas (escolhidas entre as 37 possíveis variáveis explicativas categorizadas), cada uma representada por $k - 1$ variáveis *dummies* (k = número de categorias da variável explicativa, sendo a última categoria representada pela combinação linear das demais *dummies*) e, portanto,

cada uma com k - 1 coeficientes estimados no modelo final. O modelo final conteve as seguintes variáveis explicativas para a variável de resposta qualidade de crédito, além da constante que também se mostrou significante estatisticamente:

1) Profissão;
2) Valor dos Cheques Devolvidos - Motivo 12;
3) Saldo Médio em Conta-Corrente;
4) Saldo Médio em Aplicações;
5) Estado Civil;
6) Tempo de Residência;
7) Quantidade de Prestações da Operação;
8) Valor da Operação de Empréstimo;
9) Bloqueio na Emissão de Cheques;
10) Idade na Admissão Dividida pela Idade Atual;
11) Salário Líquido;
12) Residência Própria;
13) Quantidade de Dependentes Financeiros.

O conjunto de variáveis contidas no modelo final é bastante completo, visto que todas as possíveis características de uma operação de crédito estão inclusas: profissão, residência própria, estado civil, quantidade de dependentes financeiros, idade na admissão dividida pela idade atual e tempo de residência definem o perfil geral do cliente (variáveis cadastrais), sendo notada como maior falta a variável escolaridade (que não foi disponibilizada para o estudo); salário líquido,

saldo em conta-corrente e saldo de aplicações definem o perfil de renda do cliente (variáveis financeiras), valor da operação e quantidade de prestações definem as características da própria operação de crédito (podendo ser lidas como o valor em risco e o prazo em que o valor está em risco) e as variáveis bloqueio na emissão de cheques e valor de cheques devolvidos caracterizam o perfil negativo do cliente (variáveis indicadoras de restrições dos clientes). As variáveis sexo, UF de nascimento, comuns em sistemas de *score*, não se mostraram relevantes para a qualidade de crédito, bem como variáveis patrimoniais (tais como valor de imóveis e automóveis), sendo estas últimas estão indiretamente representadas pelas variáveis financeiras dos clientes.

O modelo foi estimado através do software *SPSS*, selecionando-se a opção "regressão logística – método *forward stepwise* LR (*likelihood ratio* ou razão de verossimilhança)" com a variável de resposta "qualidade de crédito" e com as 37 variáveis explicativas categorizadas para serem submetidas ao processo de seleção de variáveis. A Tabela 12 resume os passos exigidos para alcançar o modelo final, revelando qual variável foi incluída ou excluída a cada passo e os *p-values* dos testes de razão de verossimilhança para inclusão e exclusão de variáveis (representados pelos $p\text{-}value_i$ e $p\text{-}value_e$), sendo adotadas $P_i = 0{,}15$ *e* $P_e = 0{,}20$ (valores críticos da probabilidade de inclusão e da probabilidade de exclusão) conforme discutido anteriormente.

Tabela 12. Regressão Logística – Forward Stepwise – Passo-a-Passo

Passo	Variável Incluída	Motivo de Inclusão ou Não	Variável Excluída	Motivo de Exclusão ou Não
0	Profissão	$p\text{-}value_i = 0{,}000 < P_i = 0{,}15$	-	-
1	Valor dos Cheques Devolvidos – Motivo 12	$p\text{-}value_i = 0{,}000 < P_i = 0{,}15$	-	-
2	Saldo Médio em Conta-Corrente	$p\text{-}value_i = 0{,}000 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
3	Saldo Médio de Aplicações	$p\text{-}value_i = 0{,}000 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
4	Estado Civil	$p\text{-}value_i = 0{,}012 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
5	Tempo de Residência	$p\text{-}value_i = 0{,}011 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
6	Quantidade de Prestações da Operação	$p\text{-}value_i = 0{,}000 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
7	Valor da Operação de Empréstimo	$p\text{-}value_i = 0{,}000 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
8	Bloqueio na Emissão de Cheques	$p\text{-}value_i = 0{,}036 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
9	Idade na Admissão / Idade Atual	$p\text{-}value_i = 0{,}021 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
10	Salário Líquido	$p\text{-}value_i = 0{,}006 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
11	Residência Própria	$p\text{-}value_i = 0{,}004 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
12	Quantidade de Dependentes Financeiros	$p\text{-}value_i = 0{,}024 < P_i = 0{,}15$	Nenhuma	Todos $p\text{-}values_e$ menores que P_e
13	(Nenhuma – Passo Terminal)	Todos $p\text{-}values_i$ maiores que P_i	Nenhuma	Todos $p\text{-}values_e$ menores que P_e

Conforme pode ser observado na Tabela 12, o procedimento não detectou a necessidade de excluir nenhuma variável ao longo dos passos. Assim, uma variável incluída em determinado passo não foi excluída em nenhum dos passos seguintes. Portanto, a cada passo foi incluída uma nova variável até a obtenção do modelo final no passo 12 (com todas as variáveis incluídas até esse passo), já que no passo 13 nenhuma outra variável pôde ser incluída ou excluída do modelo.

As variáveis participaram do modelo em seu formato categorizado, ou seja, o mesmo formato apresentado na Tabela 8 (ou pelas tabelas do apêndice I). Por exemplo, a variável residência própria, que continha 3 categorias, teve dois coeficientes estimados na equação: -0,259 para a primeira categoria e 0,758 para a segunda categoria. A terceira categoria não teve nenhum coeficiente, já que essa categoria é representada por uma combinação linear das demais categorias. Assim, se o cliente da operação de crédito está na segunda categoria de residência própria (possui casa própria), a operação de crédito terá um *score* maior que *scores* de operações de clientes de outras categorias da variável (*ceteris paribus*). Operações de crédito na primeira categoria de residência própria (missing – cliente não informou) terão *score* menor que as demais operações e operações na terceira categoria (cliente não possui casa própria) terão *score* maior que operações daqueles que não informaram e menor que operações daqueles que informaram possuir residência própria.

A Tabela 13 resume os coeficientes estimados de todas as variáveis do modelo final, sendo que as categorias estão ordenadas de acordo com a Tabela 8 (ou tabelas dessas variáveis apresentadas no apêndice I).

Tabela 13. Variáveis e Coeficientes Estimados no Modelo Final

Variável	*Categoria*	*Coeficiente*	*Variável*	*Categoria*	*Coeficiente*
Profissão	1	-4,003	Valor da Operação de Empréstimo	1	-0,644
	2	0,183		2	-1,290
	3	-0,984		3	-0,409
	4	1,650		4	-
	5	0,426	Quantidade de Prestações da Operação	1	-0,240
	6	-4,721		2	-0,542
	7	1,267		3	-1,009
	8	-1,311		4	-1,329
	9	-		5	-0,141
Valor dos Cheques Devolvidos pelo Motivo 12	1	3,529		6	-0,866
	2	0,937		7	-
	3	0,443	Quantidade de Dependentes Financeiros	1	0,659
	4	0,121		2	0,512
	5	-		3	0,327
Saldo Médio em Conta-Corrente	1	-1,876		4	0,111
	2	-3,111		5	-
	3	-0,844	Idade na Admissão Dividida pela Idade Atual	1	0,559
	4	-0,599		2	0,927
	5	-0,421		3	0,525
	6	-0,108		4	-
	7	-	Bloqueio na Emissão de Cheques	1	-0,655
Saldo Médio de Aplicações	1	-1,614		2	-1,057
	2	-2,300		3	0,882
	3	-1,434		4	-
	4	-0,623	Tempo de Residência	1	0,085
	5	-0,097		2	-0,129
	6	-		3	-1,477
Residência Própria	1	-0,259		4	-0,982
	2	0,758		5	-0,366
	3	-		6	-
Estado Civil	1	1,722	Salário Líquido	1	-0,847
	2	-0,701		2	-0,800
	3	-0,350		3	-
	4	1,258	CONSTANTE	-	2,013
	5	-			

O efeito de cada variável sobre o *score* da operação de crédito pode ser descrito observando-se os coeficientes estimados associados às categorias, sendo possível verificar a plausibilidade dos resultados obtidos. Os resultados são os seguintes:

- Profissão: operações de clientes desempregados ou com profissão *missing* (categorias 6 e 1, respectivamente) têm os menores coeficientes (-4,721 e –4,003) e, portanto, obtêm *scores* inferiores a clientes de outras profissões. Profissionais liberais e funcionários de empresas privadas (categorias 4 e 7, respectivamente) têm os maiores coeficientes (1,650 e 1,267) e atingem *scores* superiores a clientes de outras profissões. Operações de clientes "do lar" ou autônomos (categorias 8 e 3) têm coeficientes negativos e seus *scores* são reduzidos devido às suas profissões, enquanto que operações de funcionários públicos e aposentados (categorias 5 e 2) têm coeficientes positivos e seus *scores* são elevados devido às suas profissões. Clientes de "outras profissões" (categoria 9) não sofrem impacto de suas profissões sobre o *score* de suas operações;
- Valor dos Cheques Devolvidos pelo Motivo 12: a relação encontrada é perfeitamente plausível, já que quanto menor valor de cheques devolvidos tiver o cliente da operação de crédito, maior será o *score* da operação pois maior será o coeficiente de sua categoria. Operações de clientes sem nenhum valor de cheques devolvidos (categoria 1) têm maior *score* que operações de clientes com valores positivos de cheques devolvidos, sendo que operações na categoria 5 (R$ 498 ou mais em cheques devolvidos) têm o menor *score*;
- Saldo Médio em Conta-Corrente: as piores operações são aquelas da categoria 2, cujo coeficiente de –3,111 indica que operações de clientes correntistas com saldo de até R$ 3,10 em conta-corrente obtêm os menores *scores*, enquanto que operações de clientes da categoria 7 (mais de R$ 364 em conta-corente) obtêm os maiores *scores.* Clientes não correntistas do banco (categoria 1) são considerados melhores que correntistas com até R$ 3,10 em conta, mas são piores que todos os demais clientes;

- Saldo Médio de Aplicações: o resultado é similar ao obtido no saldo médio em conta-corrente. As melhores operações são aquelas de clientes com mais aplicações, enquanto que as piores são aquelas de correntistas sem nenhuma aplicação (categoria 2), seguidas pelas operações de não correntistas (categoria 1);
- Residência Própria: as piores operações são as dos clientes que não informam se possuem residência própria (missing – categoria 1), enquanto que as melhores são as de clientes que possuem residência própria. Quem não possui residência (categoria 3) tem *score* melhor do que quem não informa, mas pior do que quem possui;
- Estado Civil: operações de divorciados ou que não informaram estado civil (categoria 1) e de viúvos ou separados judicialmente obtêm os melhores *scores*, enquanto que operações de solteiros (categoria 2) resultam nos piores *scores* e clientes casados (categoria 5) não sofrem impacto do estado civil no *score* de suas operações;
- Valor da Operação de Empréstimo: a relação obtida indica que a relação entre *score* e valor da operação não é sempre crescente ou decrescente, sofrendo inversão: quanto maior for o valor da operação, pior será o *score*, mas aumentando ainda mais o valor da operação o *score* passa a melhorar. Assim, operações de clientes da categoria 4 (operações de mais de R$ 1.500) são as que obtêm melhor *score,* seguidas por operações de clientes da categoria 3 (operações de R$ 850 a R$ 1.500). Na categoria 2 (R$ 630 a R$ 850) encontram-se as piores operações, enquanto que na categoria 1 (até R$ 630) encontram-se operações melhores que operações da categoria 2, mas piores que as demais;
- As demais variáveis apresentam resultados análogos, devendo-se observar que a variável quantidade de prestações, ou seja, o prazo no qual o valor emprestado está em risco, tem resultado similar ao obtido pelo valor da operação de empréstimo: a relação entre quantidade

de prestações e *score* não é sempre crescente ou decrescente, já que quanto maior é o número de prestações pior é o *score*, mas a partir de certo ponto o aumento do número de prestações passa a ter efeito melhor sobre o *score*;

- Constante: a presença de uma constante no valor de 2,013 indica que, se o cliente estiver na última categoria de todas as variáveis explicativas, o *score* de sua operação será igual a exp(2,013)/[1+exp(2,013)] = 88 pontos.

Além dos resultados analisados sob o foco de cada variável, também é possível informar quais são, em geral, as variáveis que mais impactam sobre o *score*. As variáveis que mais afetam o *score* são aquelas com maior amplitude de coeficientes entre as categorias (diferença entre o coeficiente máximo e o mínimo), já que, quanto maior é a amplitude, maior é a oscilação que a variável pode trazer ao *score* da operação. Em ordem decrescente de impacto sobre o *score* encontram-se as variáveis profissão, valor de cheques devolvidos pelo motivo 12, saldo médio em conta-corrente, estado civil, saldo médio de aplicações, bloqueio na emissão de cheques, tempo de residência, quantidade de prestações, valor da operação de empréstimo, residência própria, idade na admissão dividida pela idade atual, salário líquido e quantidade de dependentes.

A partir dos coeficientes estimados mostrados na Tabela 13, a equação final para determinar Y_i^* (o valor estimado da variável qualidade de crédito para a operação de crédito *i*), ou seja, a probabilidade condicional de *Yi* ser igual a 1 dadas as respostas das variáveis explicativas do indivíduo *i* pode ser escrita como:

$$Y_i^* = \pi^*(X_i) = \frac{e^{g^*(X_i)}}{1 + e^{g^*(X_i)}}$$

em que

$$g^*(X_i) = 2{,}013 - 4{,}003 \cdot D_{prof1} + \ldots - 1{,}311 \cdot D_{prof8} + 3{,}529 \cdot D_{cheq1} + \ldots + 0{,}121 \cdot D_{cheq4} +$$
$$-1{,}876 \cdot D_{saldocc1} + \ldots - 0{,}108 \cdot D_{saldocc6} + \ldots + \ldots + 0{,}659 \cdot D_{depfin1} + \ldots + 0{,}111 \cdot D_{depfin4}$$

A equação de $g^*(X_i)$ contém um total de 56 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, D_{prof1} é o valor da *dummy* da variável profissão, em que $D_{prof1} = 1$ se a operação estiver categoria 1 de profissão ou $D_{prof1} = 0$ se a operação não estiver na categoria 1 de profissão. O *score* da operação de crédito *i* é dado por:

$$Score_i = 100 \; . \; Y_i^*$$

Note que, já que o *score* nada mais é do que 100 vezes o valor estimado da variável de resposta e esse valor estimado tem relação positiva (mas não linear) com os coeficientes das variáveis explicativas, então coeficientes positivos aumentam o *score* da operação de crédito, coeficientes negativos abaixam o *score* e coeficientes nulos não o afetam. Além disso, os coeficientes mais positivos da Tabela 13 representam as características que mais aumentam o *score* e os coeficientes mais negativos representam as características que mais diminuem o *score*.

Uma vez estimado o modelo final, foi feita uma avaliação do modelo através do teste de razão de verossimilhança para verificar a significância estatística conjunta de todos os 56 coeficientes estimados no modelo final apresentado na Tabela 13. Esse teste é dado por:

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

Em que G segue uma distribuição χ^2 com p graus de liberdade, sendo p o número de coeficientes estimados no modelo final ($p = 56$) e hipótese nula de que todos os coeficientes do modelo final são estatisticamente iguais a zero. L(modelo final) representa o valor da função de log-verossimilhança do modelo final e L(modelo sem nenhuma variável nem constante) representa o valor da função de log-verossimilhança sob a hipótese de que todos os coeficientes seriam nulos. Substituindo-se os valores estimados pelo software *SPSS* para as funções de log-verossimilhança o resultado é:

$$G = -2\,[-3259{,}178 - (-2597{,}820)] = 1322{,}716$$

A probabilidade do valor da distribuição $\chi 2$ com 56 graus de liberdade ser maior que 1322,716 (ou seja, o *p-value* do teste) é de 0,000, indicando a rejeição da hipótese nula a qualquer nível de significância. Assim, pode-se dizer que o conjunto de coeficientes estimados no modelo final é estatisticamente significante, ou seja, não nulo.

Os coeficientes estimados também foram testados individualmente, bem como os conjuntos de coeficientes de cada variável. Assim, foram realizados testes para cada um dos 56 coeficientes e também para cada grupo de coeficientes de determinada variável. Individualmente,

o teste tem a hipótese nula de que o coeficiente *i* (i = 1, 2, ..., 56) é estatisticamente nulo e apenas 1 grau de liberdade já que é testado cada coeficiente individualmente, enquanto que para cada grupo de coeficientes de determinada variável o teste tem hipótese nula de que o conjunto de coeficientes das k –1 categorias da variável são estatisticamente nulos, com k – 1 graus de liberdade. Por exemplo, para a variável residência própria, que teve dois coeficientes estimados no modelo final (-0,259 e 0,758), foi feito um teste individual para a hipótese nula de que -0,259 é estatisticamente nulo, outro teste individual para a hipótese nula de que 0,758 é estatisticamente nulo, e um teste conjunto para a hipótese nula de que ambos os coeficientes (-0,259 e 0,758) são estatisticamente nulos. Esse procedimento foi feito para cada variável. Em todos os casos foi elaborado o teste Wald, que é um teste similar ao teste *G* de razão de verossimilhança (que também poderia ser aplicado), baseado nas estimativas dos coeficientes e em seus erros-padrão estimados. O teste Wald segue uma distribuição normal com *m* graus de liberdade, em que m é o número de coeficientes que estão sendo testados. Para os testes individuais, $m = 1$, e para os testes do conjunto de coeficientes de cada variável, $m = k - 1$, em que k é o número de categorias (coeficientes estimados) da variável. Os resultados dos testes Wald, com base no nível de significância de 5%, revelaram que:

- Os 13 testes conjuntos para o grupo de coeficientes de cada variável levaram à rejeição da hipótese nula de que o grupo de coeficientes seria estatisticamente nulo, qualquer que seja a variável testada já que, para todas elas, os *p-values* dos testes Wald foram inferiores a 0,05 (5%). Assim, pode-se dizer que, para cada uma das 13 variáveis em questão, o grupo de coeficientes estimados para essa variável foram estatisticamente significantes, ou seja, não nulos;

- Os 56 testes individuais para cada um dos coeficientes estimados do modelo final levaram rejeição da hipótese nula de que o coeficiente seria estatisticamente nulo, para 50 dos 56 coeficientes testados (inclusive a constante) já que, nesses testes, os *p-values* dos testes Wald foram inferiores a 0,05 (5%). Os outros 6 coeficientes apresentaram *p-values* superiores a 5% e, portanto, seus testes levariam à aceitação da hipótese nula de que tais coeficientes seriam, individualmente, iguais a zero, tendo sido os seguintes casos:

  +0,121: Coeficiente da 4ª categoria de valor de cheques dev. mot.12 (*p-value* = 0,072);

  -0,108: Coeficiente da 6ª categoria de saldo médio em conta-corrente (*p-value* = 0,131);

  -0,097: Coeficiente da 5ª categoria de saldo médio de aplicações (*p-value* = 0,144);

  -0,129: Coeficiente da 2ª categoria de tempo de residência (*p-value* = 0,086);

  -0,141: Coeficiente da 5ª categoria de quantidade de prestações (*p-value* = 0,077);

  +0,111: Coeficiente da 4ª categoria de quant. de dependentes financ. (*p-value* = 0,157).

Apesar dos testes individuais dos coeficientes terem revelado 6 coeficientes estatisticamente nulos, nenhum dos *p-values* esteve exageradamente acima do nível de significância e, somando-se o fato de que todos os testes conjuntos revelaram que cada grupo de coeficientes não é estatisticamente nulo e que o modelo como um todo é significante estatisticamente (conforme visto pelo teste *G*), o modelo final apresentado na Tabela 13 não foi modificado, isto é, não foi eliminada nenhuma das 6 variáveis *dummies* representativas das 6 categorias com coeficientes individualmente não significantes. Em outras palavras, o modelo apresentado na Tabela 13 foi considerado estatisticamente válido em termos de significância dos coeficientes estimados.

Uma vez considerado estatisticamente significante o modelo final estimado, foi testado o poder explicativo desse modelo, isto é, com que eficiência o modelo conseguiu descrever a variável de resposta qualidade de crédito. Em linhas gerais, uma medida para verificar o poder explicativo analisa as diferenças entre os valores estimados de Y_i (Y_i^*) e os valores observados de Y_i. Quando essas diferenças são consideradas pequenas, então é dito que o modelo tem um poder explicativo elevado, ou seja, consegue estimar Y_i com boa precisão. No caso de regressão logística, em que Y_i^* pode assumir qualquer valor entre 0 e 1 mas Y_i assume somente valor 0 ou 1, existem diferentes formas de se verificar o poder explicativo do modelo. No entanto, as principais formas são alvos de críticas devido aos resultados indesejáveis gerados. É o caso, por exemplo, da medida denominada χ^2 de Pearson e também da medida denominada desvio residual. Ambas as medidas são baseadas no somatório das diferenças ao quadrado entre os valores estimados e observados de Y_i, com a diferença de que a medida de Pearson utiliza uma forma funcional quadrática e o desvio residual utiliza uma forma funcional logarítimica. O resultado de ambas é basicamente o mesmo: um teste de hipótese baseado na distribuição χ^2 cujo resultado apenas indica se o modelo é eficiente ou não (dependendo do nível de significância adotado), mas não indica o quanto o modelo é eficiente. O maior problema desses testes, no entanto, é que a distribuição χ^2 utilizada nos testes é incorreta quando o número de configurações possíveis das variáveis explicativas na amostra de modelagem é aproximadamente igual ao número de casos contidos na amostra de modelagem, ou seja, quando cada caso da amostra de modelagem é único e não repetido, situação que ocorre na maioria dos estudos[17].

O teste utilizado no estudo de *credit scoring* para verificar o poder explicativo do modelo de regressão logística estimado foi o teste de Hosmer-Lemeshow, calculado automaticamente

[17] Hosmer, D. W.; Lemeshow, S., 1989, op. cit., p.135-175.

pelo *software SPSS*. Esse teste consiste em uma técnica de agrupamento dos casos da amostra de modelagem em percentis formados em ordem crescente de valor estimado de Y_i, a partir dos quais é gerada uma estatística C comprovadamente distribuída por χ^2 com g-2 graus de liberdade, em que g é o número de percentis adotado para o teste (usualmente g = 10), cada percentil contendo o mesmo número de casos. Por exemplo, no presente estudo, cuja amostra contém 4.702 operações de crédito, foram calculados os valores estimados Y_i^* para todas as i = 4.702 operações. Essas operações foram ordenadas em ordem crescente de Y_i^* e divididas em 10 grupos de cerca de 470 operações (4.702/10). O teste C calculou, dentro de cada grupo, a soma das diferenças entre valores estimados e observados nos grupos, ponderada pela quantidade de casos em cada grupo e pelos percentuais de operações boas e ruins dentro de cada grupo. O resultado obtido para cada grupo foi somado resultando na estatística C, que segue uma distribuição χ^2 com 10 - 2 = 8 graus de liberdade sob a hipótese nula de que o modelo estimado é eficiente para estimar Y_i (já que, quanto maior é C, maiores são os desvios entre os valores estimados e observados de Y_i).

O resultado do teste C de Hosmer-Lemeshow a partir do modelo estimado apresentado na Tabela 13 foi de C = 8,424. Ao nível de significância de 5% e com 8 graus de liberdade, o valor crítico da distribuição χ^2 é de 15,507. Visto que o valor da estatística C foi menor que o valor crítico, a hipótese nula do teste foi aceita e, portanto, a conclusão foi que o modelo estimado é eficiente para estimar Y_i (as diferenças entre os valores estimados e observados de Y_i somente seriam consideradas estatisticamente grandes se a estatística C fosse maior que 15,507).

Apesar de ser conceitualmente correto e não ter problemas conceituais na definição da distribuição da estatística C do teste, o teste de Hosmer-Lemeshow apresenta um problema igual aos testes de Pearson e de desvios residuais mencionados anteriormente, que é indicar apenas se

o modelo é ou não eficiente para estimar Y_i, mas sem informar o quanto a estimação é eficiente. Logo, para complementar o resultado do teste de Hosmer-Lemeshow, existe uma maneira com forte apelo intuitivo para quantificar o poder de estimação do modelo, a denominada tabela de classificação. Essa tabela nada mais é do que uma comparação cruzada entre a quantidade de casos com Y_i observado igual a 1 ou 0 *versus* a quantidade de casos com Y_i estimado igual a 1 ou 0. Visto que os Y_i estimados assumem valores de 0 a 1 e não valores 0 ou 1, é preciso gerar uma variável dicotômica derivada dos valores estimados de *Yi*. Para derivar uma variável dicotômica a partir desses valores estimados, é necessário definir arbitrariamente um ponto de corte c ($0 \leq c \leq 1$) para os valores estimados de Y_i. Se o valor estimado de Y_i for superior a c, então a variável derivada assume valor 1 (ou seja, o valor estimado de Y_i é 1 e a variável de resposta estimada é classificada como "boa") e, se o valor estimado for inferior ou igual a c a variável estimada assume valor 0 (ou seja, o valor estimado de Y_i é 0 e a variável de resposta estimada é classificada como "ruim"). Quando já se sabe que o modelo estimado apresenta coeficientes estatisticamente significantes e que o modelo é eficiente em estimar Y_i – mas não se sabe o quanto é eficiente – o ponto de corte c mais usual é de 0,5. A Tabela 14 contém os resultados da tabela de classificação para o modelo estimado, baseado na amostra de 4.702 operações de crédito e no ponto de corte $c = 0,5$.

Tabela 14. Tabela de Classificação do Modelo Estimado

	Valores Estimados		Total
Valores Observados	Y_i^* =0 (qualidade de crédito ruim)	Y_i^* =1 (qualidade de crédito boa)	
Y_i =0 (qualidade de crédito ruim)	2171	180	2351
Y_i =1 (qualidade de crédito boa)	200	2151	2351
Total	2371	2331	4702

A partir da Tabela 14 é possível calcular as taxas de acerto do modelo. A taxa total de acerto foi igual a (2171 + 2151)/4702 = 91,92%, enquanto que a taxa de acerto para boas operações (qualidade de crédito boa) foi de 2151/2351 = 91,49% e a taxa de acerto para operações de crédito ruins foi de 2171/2351 = 92,34%. Em outras palavras, espera-se que o modelo seja capaz de classificar corretamente 91,92% de todas as operações (independente de ser boa ou ruim), sendo classificadas corretamente 91,49% das operações boas e 92,34% das operações ruins.

A conclusão que pode ser feita a partir de todos os resultados observados é que o modelo de *credit scoring* gerado através da regressão logística para a linha de crédito CAB é um modelo estatisticamente válido, com coeficientes estimados considerados significantes individual ou conjuntamente, e que o poder de classificação resultante indica que a aplicação do modelo na prática será capaz de classificar corretamente uma porção bastante elevada de todas as operações de crédito submetidas à análise.

IV.2. Exemplo

O modelo final elaborado para a linha de crédito popular CAB apresentou alto desempenho quanto ao grau de acerto e à significância dos coeficientes estimados. Do total de 13 variáveis, todas são variáveis categorizadas, ou seja, representadas por coeficientes que são aplicáveis quando a operação de crédito pertence a determinada categoria da variável em questão. O exemplo abaixo clarifica como é feito o cálculo da equação e do *score* da operação de crédito. Note que, ao inserir os dados no cadastro do exemplo, tais dados são associados às

categorias pré-definidas pelo teste CHAID, determinadas nas tabelas do apêndice I. 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.

Exemplo 1. Cálculo do Score da Operação Crédito

		Categoria	Coeficiente
Nome:	Nair de Castro Mello	-	-
Profissão:	Autônoma	3	-0,984
Valor dos Cheques Devolvidos – Mot. 12:	Nulo	1	3,529
Saldo Médio em Conta-Corrente:	R$ 527,32	7	-
Saldo Médio de Aplicações:	Nulo	2	-2,300
Residência Própria:	Possui	2	0,758
Estado Civil:	Solteira	2	-0,701
Valor da Operação de Empréstimo:	R$ 1.500,00	3	-0,409
Quantidade de Prestações da Operação:	12	4	-1,329
Quantidade de Dependentes Financeiros:	Nenhum	1	0,659
Idade na Admissão / Idade Atual:	23 / 32 = 0,71875	3	0,525
Bloqueio na Emissão de Cheques:	Não Bloqueado	3	0,882
Tempo de Residência:	18 Meses	4	-0,982
Salário Líquido:	R$ 700,00	3	-
CONSTANTE	-	-	2,013
(A) Soma dos Coeficientes:			***1,661***
SCORE = 100 * exp(A) / 1+exp(A)			***84***

IV.3. Distribuição dos Scores Estimados

Os resultados de poder de classificação calculados na seção anterior foram baseados na aplicação do modelo estimado somente sobre a amostra de modelagem de 4.702 operações de crédito. De forma a visualizar melhor o desempenho discriminatório do modelo, é conveniente aplicar o modelo em toda a base de dados disponível (com 201.075 operações de crédito) e verificar graficamente se os resultados de classificação são realmente bons e se mantêm quando comparados aos resultados obtidos na amostra de modelagem. Uma variação grande na capacidade de classificação pode ser um indício de que a amostra de modelagem sorteada estava viesada em relação a todas as operações de crédito disponíveis.

Os Gráficos 2 e 3 mostram a distribuição observada do *score* estimado das operações de crédito quando o modelo é aplicado sobre toda a base de dados disponível de operações boas e ruins (201.075 casos). Conforme esperado, a distribuição do *score* das operações boas (Gráfico 2) está muito mais concentrada a direita do *score* 50, enquanto que a distribuição do *score* das operações ruins (Gráfico 3) está muito mais concentrada à esquerda do *score* 50. Esse fato mostra que a maior parte das operações boas têm *score* alto e que a grande maioria das operações ruins tem *score* baixo, o que garante um forte poder de discriminação do modelo entre bons e maus pagadores. As linhas pontilhadas representam as funções acumuladas, ou seja, o percentual de operações boas (Gráfico 2) ou ruins (Gráfico 3) que têm *score* menor ou igual ao *score* determinado no eixo horizontal.

Gráfico 2. Score das Operações de Crédito Boas

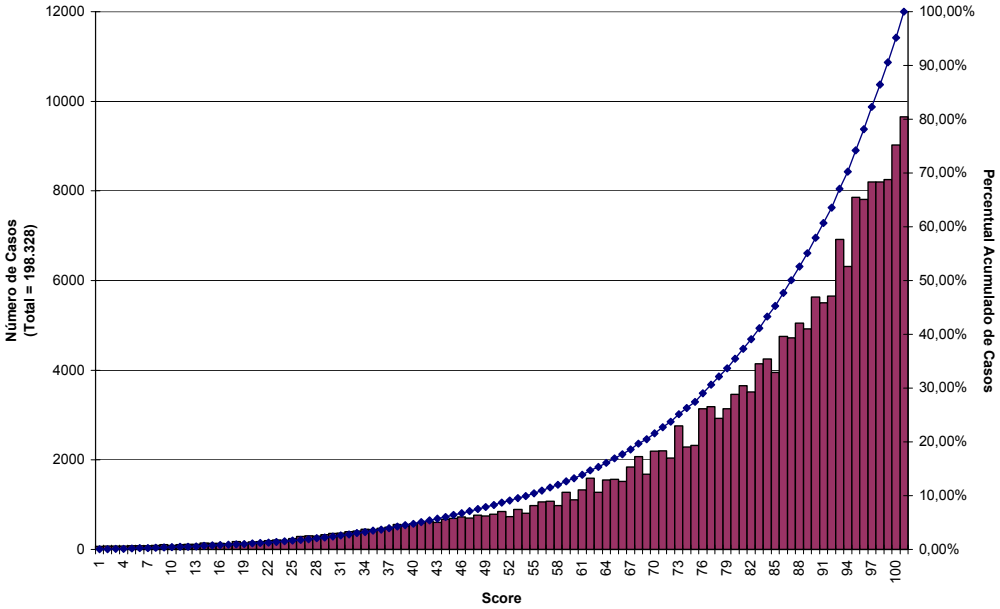
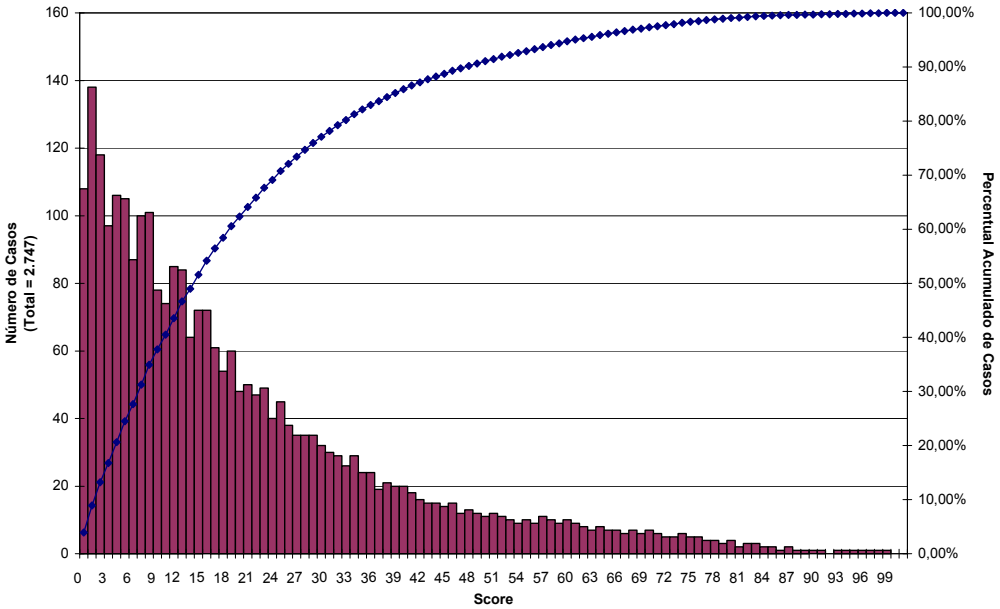
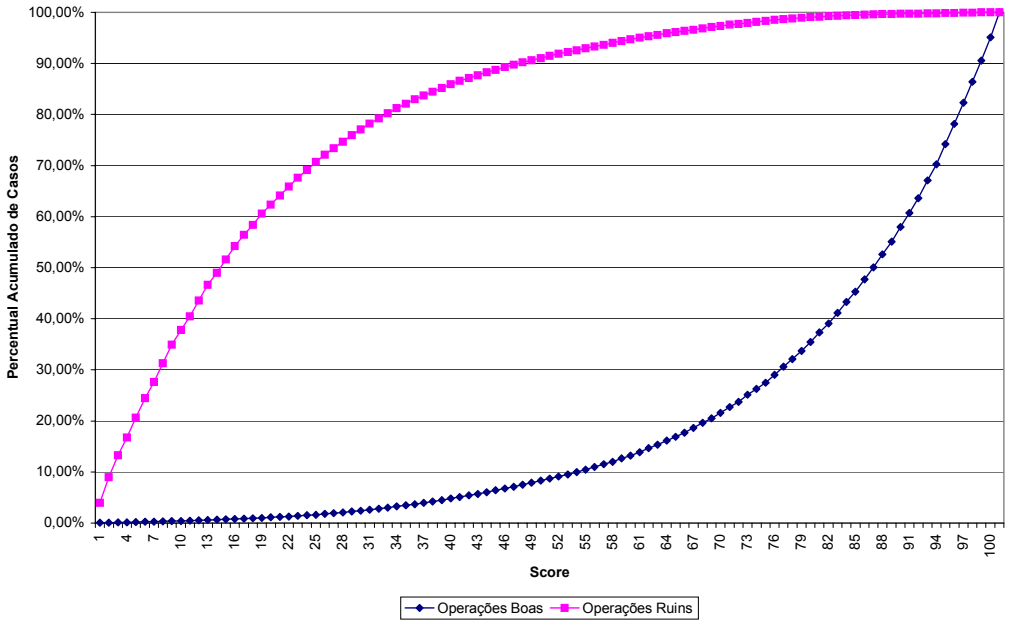


Gráfico 3. Score das Operações de Crédito Ruins



O Gráfico 4 mostra as funções acumuladas das operações boas e ruins. Quanto mais acima da função das boas operações estiver a função das operações ruins, maior é a capacidade de discriminação do modelo. Um modelo com poder de classificação perfeito teria uma função acumulada de boas operações seguindo exatamente o eixo horizontal inferior e o eixo vertical direito (todas as boas operações teriam *score* igual a 100) e uma função acumulada de operações ruins seguindo exatamente o eixo vertical esquerdo e o eixo horizontal superior (todas as operações ruins teriam *score* igual a 0). O modelo elaborado para a linha de crédito CAB apresenta um distanciamento nítido bastante acentuado, comprovando uma diferenciada qualidade do modelo.

Gráfico 4. Distribuição Acumulada dos Scores por Qualidade de Crédito



A Tabela 15 apresenta alguns resultados numéricos (medidas de dispersão) gerados a partir dos 3 gráficos anteriores. Esses resultados confirmam as conclusões de que o modelo apresentou acentuado poder de discriminação entre bons e maus pagadores.

Tabela 15. Medidas de Dispersão das Distribuições de Scores de Operações Boas e Ruins

	Score Médio	*Score Mediano*	*% de Clientes com Score Maior que 50*	*% de Clientes com Score Menor que 50*	*Moda*
Operações Boas	81	86	91,27	8,73	100
Operações Ruins	20	13	8,52	91,48	1

Lembrando das taxas de acerto obtidas através da tabela de classificação (Tabela 14) gerada sobre a amostra de modelagem de 4.702 operações de crédito, e comparando tais resultados com os valores expressos na Tabela 15, baseada em todo o banco de dados com 201.075 operações, nota-se que houve pouca flutuação dos resultados. Enquanto que na amostra o percentual de operações boas com *score* superior a 50 foi de 91,49%, em todo o banco de dados esse percentual foi de 91,27%, uma variação perfeitamente aceitável. Ainda na amostra, o percentual de operações ruins com *score* inferior a 50 foi de 92,34%, enquanto que em todo o banco de dados o resultado foi de 91,48%, uma variação também aceitável. A ausência de grandes variações é um indício de que a amostra sorteada para montagem do modelo (4.702 operações) não estava viesada em relação ao banco de dados integral (201.075 operações) e, portanto, o forte poder classificatório verificado pelos testes de poder explicativo baseados na amostra deve se manter em todo o banco de dados.

É importante ressaltar que a maior parte das operações classificadas incorretamente são aquelas que têm um *score* intermediário (entre 30 e 70). Visto que no banco de dados há um total

de 40.790 operações nessa faixa de *score* – cerca de 20% do total de operações – o poder classificatório do modelo é confiável em pelo menos 80% de todas as operações de crédito.

IV.4. Estabilidade do Modelo

Em um modelo de *credit scoring* eficiente não é desejável que uma variável sozinha seja responsável por aceitar ou rejeitar o pedido de crédito de um indivíduo. Nenhuma variável deve ter força suficiente, independente dos valores que as outras variáveis assumam, para se chegar à conclusão sobre conceder ou não o crédito. Por exemplo, não se pode negar o pedido de crédito de um cliente pelo fato dele ser divorciado, sem nem mesmo conhecer suas outras características. Em países desenvolvidos como os Estados Unidos, existem restrições jurídicas quanto à discriminação em processos de concessão de crédito, tal como o *Consumer Credit Act* de 1974, um conjunto de normas regulatórias que prevê punições legais para casos de discriminação em função de características como sexo e raça[18]. Apesar de no Brasil não existir nenhuma norma similar totalmente voltada à questão de concessão de crédito, é possível que a evolução das instituições brasileiras levem ao aparecimento de normas regulamentando o processo de concessão de crédito, motivo que torna o estudo de estabilidade uma ferramenta valiosa para não incorrer em desrespeito a alguma lei que possa surgir. Deve-se ressaltar que o estudo proposto nessa seção não se encontra na literatura de crédito, sendo apenas uma sugestão de análise original do presente trabalho.

[18] Hand, D. J.; Henley, W. E. 1997, op. cit.

É necessário que se faça um estudo para saber se existe alguma variável que responda sozinha pela concessão. No estudo de estabilidade é determinado o *score* mínimo que uma operação de crédito pode ter se tiver todas as piores características nas variáveis exceto em uma, na qual estaria no melhor nível. Isso resultaria no que se chama de *score* mínimo no melhor nível da variável. Da mesma forma, é determinado o *score* máximo que uma operação pode ter se tiver todas as melhores características nas variáveis exceto em uma, na qual estaria no pior nível, o que resultaria no *score* máximo no pior nível da variável. Para que o modelo não seja muito influenciado por uma única variável é necessário que o *score* mínimo no melhor nível seja baixo (até 5 pontos maior que o mínimo global que a equação do modelo possa gerar) e que o *score* máximo no pior nível seja alto (até 5 pontos menor que o máximo global gerado pelo modelo). Em outras palavras, o *score* de uma operação não pode ser muito elevado ou muito rebaixado devido a exclusivamente uma variável explicativa.

Por exemplo, suponha que uma determinada categoria de estado civil tenha o maior coeficiente estimado entre todas as categorias de estado civil. É estudado, então, qual é o mínimo *score* que um operação de crédito que pertença a essa "melhor" categoria de estado civil pode alcançar. Para que o estado civil não seja uma variável que determine sozinha a decisão sobre concessão de crédito, esse *score* mínimo deve ser baixo e insuficiente para o crédito ser aprovado simplesmente pela resposta dada à variável estado civil.

Esse estudo foi feito para todas as variáveis incluídas no modelo, maximizando-se uma a uma e mantendo as demais variáveis minimizadas (*score* mínimo no melhor nível) e minimizando-se uma a uma mantendo as demais variáveis maximizadas (*score* máximo no pior nível). Conforme foi observado anteriormente, os *scores* máximos devem ser altos (mais próximos possíveis do maior *score* que o modelo pode fornecer) e os *scores* mínimos devem ser

baixos (mais próximos possíveis do pior *score* que o modelo pode fornecer). De acordo com a Tabela 16, pode-se perceber que o modelo é extremamente estável, não sendo afetado exclusivamente por nenhuma das variáveis em questão: o pior *score* possível gerado pelo modelo é 0 e o melhor é 100, sendo que apenas a variável profissão causa uma pequena queda no *score* máximo (para 98 pontos) e todos os demais *scores* máximos no pior nível não caíram abaixo de 100, bem como nenhum dos *scores* mínimos no melhor nível subiu acima de 0. Assim, nenhuma das variáveis isoladamente tem força suficiente para gerar uma decisão de conceder ou não o crédito.

Tabela 16. Estabilidade do Modelo de Credit Scoring

Score Máximo Possível (Máximo Global):	***0***	
Score Mínimo Possível (Mínimo Global):	***100***	
Variável	Score Mínimo no Melhor Nível	Score Máximo no Pior Nível
Profissão	0	98
Valor dos Cheques Devolvidos - Motivo 12	0	100
Saldo Médio em Conta-Corrente	0	100
Saldo Médio de Aplicações	0	100
Residência Própria	0	100
Estado Civil	0	100
Valor da Operação de Empréstimo	0	100
Quantidade de Prestações da Operação	0	100
Quantidade de Dependentes Financeiros	0	100
Idade na Admissão / Idade Atual	0	100
Bloqueio na Emissão de Cheques	0	100
Tempo de Residência	0	100
Salário Líquido	0	100

Caso o modelo de *credit scoring* gerado tivesse relevado problemas de instabilidade em alguma das variáveis, seria necessário reformular o modelo desde o processo de categorização da variável com problema até a etapa de estimação do modelo, de forma a gerar um modelo estável.

IV.5. Decisão de Crédito – Score de Corte

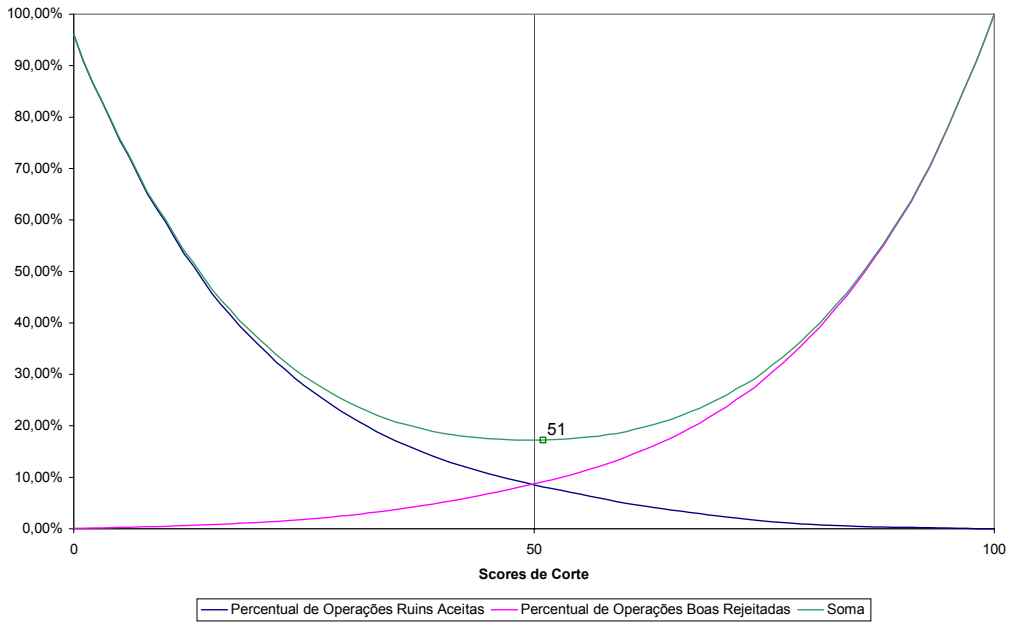
Os resultados obtidos pelo modelo de *credit scoring* estimado mostraram um poder de discriminação acentuado entre operações de crédito boas e ruins. Foi possível notar, através dos gráficos de distribuição de *score*, que as operações de crédito boas ficam concentradas em *scores* bastante elevados e que as operações de crédito ruins ficam concentrados em *scores* baixos. No entanto, para obter a decisão final sobre conceder ou não o crédito, é preciso determinar um *score* de corte para a carteira de crédito, ou seja, um valor de *score* até o qual a operação de crédito é rejeitada e acima do qual a operação de crédito é aceita.

É importante lembrar qual foi o critério fundamental de determinação de operações boas e ruins. Conforme discutido no capítulo II, o critério adotado foi a inadimplência, ou seja, o estudo de atrasos no pagamento das prestações do crédito. Assim, foram definidas como boas operações aquelas em que não houve atraso de mais de 60 dias corridos no pagamento de qualquer prestação, enquanto que as operações ruins foram aquelas com algum atraso de 61 ou mais dias no pagamento de qualquer prestação. Portanto, qualquer que seja o *score* de corte adotado, esse *score* de corte representará o nível de *score* a partir do qual a instituição provedora do crédito considerará a operação de crédito boa devido a baixa probabilidade de ter atrasos de 61 ou mais dias no pagamento de qualquer prestação (e portanto a aprovará), e até o qual a

instituição considerará a operação ruim devido a alta probabilidade de ter atrasos de 61 ou mais dias (e portanto a reprovará).

Uma ferramenta muito utilizada para determinar o *score* de corte é analisar o percentual de operações boas rejeitadas (ou seja, o percentual de operações boas classificadas pela instituição como ruins) e o percentual de operações ruins aceitas (isto é, o percentual de operações ruins classificadas pela instituição como boas) em relação os possíveis *scores* de corte. O princípio dessa análise é minimizar a soma desses percentuais, ou seja, minimizar as chances de classificação incorreta das operações de crédito. O Gráfico 5 contém os resultados da análise do *score* de corte baseado baseada nesse princípio.

Gráfico 5. Análise do Score de Corte



O Gráfico 5 foi gerado baseado nas 201.075 operações de crédito do banco de dados e é composto por três curvas. A curva azul representa o percentual de operações ruins que seriam

aceitas nos possíveis *scores* de corte, ou seja, o percentual de operações ruins que seriam classificadas como boas caso fosse adotado determinado *score* de corte. A curva rosa representa o percentual de operações boas que seriam rejeitadas, isto é, o percentual de operações boas que seriam classificadas como ruins. A curva verde representa a soma desses percentuais, ou seja, a soma do percentual operações que seriam classificadas incorretamente. O ponto mínimo da curva de soma é o ponto que minimiza a probabilidade de classificação incorreta das operações de crédito.

Obviamente a adoção de *score* de corte próximos de 100 ou de 0 não é recomendada. Por exemplo, se fosse adotado o *score* de corte 100, somente seriam aceitas operações com *score* de 101 ou mais pontos e, visto que não existem *scores* acima de 100, então a carteira de crédito CAB não aceitaria nenhuma operação de crédito pois classificaria todas elas como ruins. *Scores* de corte muito próximos de 100 resultariam na classificação incorreta de muitas operações boas, apesar de classificarem incorretamente quase nenhuma das operações ruins. *Scores* de corte muito próximos de 0 resultariam na classificação incorreta de muitas operações ruins, apesar de classificarem incorretamente quase nenhuma das operações boas. O resultado da minimização indica o *score* 51 (destacado no Gráfico 5 o qual, apesar de não ser nítido visualmente, fornece um mínimo ligeiramente menor que no *score* 50) como o melhor *score* de corte para evitar classificação incorreta de operações. Com o *score* de corte 51, ao mesmo tempo uma pequena proporção de operações ruins (mais exatamente 8,12%) seriam classificadas como boas e poucas operações boas (9,10%) seriam classificadas como ruins.

A ferramenta baseada na probabilidade de classificação incorreta de operações é apenas uma das diversas formas de decidir o melhor *score* de corte para a carteira de crédito. Na realidade, diversos outros fatores podem (e devem) ser analisados, dependendo das metas que

vigoram na instituição financeira que utilizar um modelo de *credit scoring*. Muitas instituições têm, por exemplo, metas mínimas de rentabilidade para suas carteiras de crédito que, obviamente, dependerão do *score* de corte adotado. Essas metas também incluem patamares máximos de rejeição de operações de crédito (por exemplo, não podem ser rejeitadas mais que 20% das operações de crédito) normalmente vinculadas à necessidade de manter o volume financeiro da carteira de crédito em níveis seguros, que também dependerão do *score* de corte adotado. Também deve ser considerado o caráter da linha de crédito: uma linha de crédito popular, por exemplo, muitas vezes não têm preocupações de inadimplência ou de lucratividade em primeiro plano, mas sim a exigência de prover crédito a pessoas com dificuldades de acesso a outras linhas de crédito. É o caso, por exemplo, de linhas de crédito oferecidas por bancos públicos para pessoas de baixa renda com o único objetivo de promover o bem-estar dessa população. Nesse caso, o *score* de corte pode ser flexibilizado, permitindo que a maior quantidade possível de pessoas consigam obter o crédito e o programa de aumento do bem-estar atinja seus objetivos.

A conclusão é que, antes de tomar a decisão final sobre o *score* de corte de uma linha de crédito, é recomendado realizar uma série de outros estudos verificando o impacto do *score* de corte sobre os pontos de interesse que regem a carteira de crédito. Esses estudos não foram realizados no presente trabalho principalmente pelo fato de não terem sido disponibilizados os dados necessários para análises dessa dimensão.

Capítulo V. Considerações Finais

A utilização de modelos estatísticos para criação de métodos de *credit scoring* por pessoal especializado requer que este conheça todos os pontos críticos da abordagem utilizada, bem como a criação de uma percepção segura sobre a capacidade do modelo em classificar operações de crédito, não somente baseada nos testes estatísticos do modelo, mas também sobre as flutuações que os resultados podem sofrer quando o modelo for utilizado na prática. Esse capítulo apresenta alguns dos tópicos relevantes que precisam ser considerados além do problema de viés de seleção apresentado na introdução deste trabalho.

V.1. Flutuações Populacionais

As flutuações populacionais podem ser um problema na aplicação de modelos de *credit scoring*. Essa questão descreve a tendência das populações em evoluir com o tempo, com conseqüentes mudanças nas distribuições das variáveis e se baseia nas flutuações econômicas e mudanças no ambiente competitivo que acometem a população. Por exemplo, em um prazo de 5 anos ocorrem flutuações na renda e no patrimônio das pessoas, a proporção de pessoas casadas se modifica, a idade média da população evolui e os prazos e valores das operações de crédito se modificam. O efeito de flutuações populacionais é a perda no desempenho classificatório dos modelos de *score* de crédito, que ocorre gradualmente a medida que as populações se

modificam, alterando as distribuições dos *scores* estimados das operações. A solução desse problema é a reformulação periódica do modelo de *credit scoring*, sendo necessário repetir o processo desde o início (geração de um banco de dados atualizado) até o final (a estimação da equação final e a verificação dos resultados do modelo). Para se saber o momento em que o modelo deve ser recalibrado, é necessário formular ferramentas para verificação das flutuações populacionais, sendo que a abordagem mais comum é comparar as propriedades estatísticas das operações de crédito da carteira (em vários pontos no tempo após a elaboração do modelo original) com as propriedades estatísticas da amostra utilizada para elaboração do modelo original. Grandes diferenças nas variáveis são indícios de que a população mudou e que o modelo precisa ser reestruturado.

V.2. Relatórios de Acompanhamento

A classificação de operações de crédito é o foco dos modelos de *credit scoring*. É desejável que, uma vez implementado um modelo gerado para determinada linha de crédito, os resultados possam ser verificáveis na prática. A literatura simplesmente alerta para essa necessidade, raramente mencionando formas de acompanhamento dos resultados. Considerando-se que os resultados esperados são totalmente vinculados ao objetivo da classificação, que no presente estudo foi o critério de inadimplência (atrasos nos pagamentos de prestações), a sugestão é que sejam feitos mensalmente (após a implementação do modelo) relatórios de acompanhamento que permitam detectar os efeitos do modelo com relação a esse objetivo, bem como verificar aspectos que podem revelar perda na eficiência do modelo ao longo do tempo.

Não existe um método fechado para elaborar os relatórios necessários, mas pode-se sugerir o formato de alguns deles, conforme resumido a seguir.

Relatório de Acompanhamento de Variáveis

O relatório de acompanhamento das variáveis consiste na construção de uma tabela de freqüências envolvendo todos os pedidos de crédito de determinado período, para cada uma das variáveis inclusas no modelo de *credit scoring*, divididas em categorias de acordo com o modelo. Esse relatório permite identificar flutuações na população que requer crédito e, caso exista uma flutuação grande ao longo do tempo após a implementação do modelo em relação à amostra que foi utilizada para geração da equação do modelo, é um indício de que o público alvo da carteira de crédito está se modificando ou que a população verdadeira não é equivalente à população da amostra. Nesse caso, o modelo perde sua eficiência e não é mais capaz classificar corretamente as operações de crédito, ou seja, o modelo precisa ser reconstruído.

É importante observar que esse relatório não expressa nenhum efeito da implementação do modelo, visto que apenas descreve as características de todos os pedidos de crédito, e não somente daqueles cujo crédito foi realmente concedido. Também deve-se salientar que, apesar do esforço computacional ser muito maior, o recomendado é acompanhar a evolução não somente das variáveis inclusas no modelo (apesar dessas serem as mais importantes de acompanhar), mas sim do maior número possível de variáveis disponíveis, para se obter uma percepção mais completa sobre as flutuações ocorridas.

Relatório de Taxas de Rejeição

Consiste no acompanhamento, após a implementação do modelo, da proporção de pedidos de crédito rejeitados (ou seja, com *score* inferior ou igual ao de corte) em determinado período em relação ao total de pedidos de crédito elaborados no mesmo período. Se o relatório de acompanhamento de variáveis não estiver indicando flutuações ao longo do tempo, então a taxa de rejeição não deve sofrer variações expressivas. Oscilações pequenas e sem direção definida (para mais e para menos em períodos subsequentes) na taxa de rejeição são perfeitamente aceitáveis. Já oscilações grandes e com direção definida (sempre para mais ou sempre para menos) são indícios da necessidade de reconstrução do modelo e, caso não seja possível reconstruí-lo imediatamente, recomenda-se uma reavaliação do *score* de corte da carteira de crédito.

Relatório de Inadimplência

O relatório de inadimplência obtém a proporção – em volumes financeiros – de operações de crédito que estão com mais de 60 dias de atraso no pagamento de qualquer prestação (número de dias equivalente ao da definição de qualidade de crédito), de acordo com o mês de concessão da operação, ao longo dos meses. Isso permite o constante monitoramento da qualidade da concessão de crédito e o cálculo da taxa de inadimplência de acordo com a definição de qualidade de crédito. Se, por exemplo, nos últimos meses há uma elevada proporção de operações que se tornaram inadimplentes (com atrasos de 61 ou mais dias) poucos meses após a concessão, isso indica que o modelo de *credit scoring* perdeu sua capacidade de discriminação.

Quando isso ocorre um novo modelo deve ser criado. É importante observar que o relatório de inadimplência só manifesta resultados do modelo a partir de 3 meses após a implementação do modelo (já que atrasos de 61 ou mais dias levam pelo menos 2 meses e 1 dia para serem verificados). A Tabela 17 mostra um exemplo hipotético do relatório de inadimplência.

Tabela 17. Exemplo de Relatório de Inadimplência

Relatório de Inadimplência: Operações com Mais de 60 dias de Atraso por Mês de Concessão ***Data de Implementação do modelo: 01/01/2001***						
Mês de Concessão	***Jan/01 (%)***	***Fev/01 (%)***	***Mar/01 (%)***	***Abr/01 (%)***	***Mai/01 (%)***	***Jun/01 (%)***
Out/00	9,0%	14,8%	14,4%	15,7%	16,0%	14,8%
Nov/00	----	8,9%	14,7%	15,8%	16,3%	15,0%
Dez/00	-----	----	8,8%	15,6%	16,2%	14,9%
Jan/01	-----	-----	----	8,0%	13,7%	12,5%
Fev/01	-----	-----	-----	----	7,2%	10,7%
Mar/01	-----	-----	-----	-----	----	6,6%
Abr/01	-----	-----	-----	-----	-----	----

No exemplo da Tabela 17, foi implementado um modelo de *credit scoring* em 01/01/2001. O relatório de safras expressa a proporção de operações, em termos de valor total emprestado, com mais de 60 dias de atraso por mês de concessão. Assim, o primeiro dado referente ao mês de concessão de janeiro somente aparece com o vencimento da terceira parcela do crédito, ou seja, abril de 2001. Em abril de 2001 (3 meses após a concessão), pôde-se verificar

que 8% do volume financeiro concedido em janeiro estavam com 61 ou mais dias de atraso (por exemplo, se um valor total de R$ 100.000 foram concedidos em janeiro, então operações de valor total de R$ 8.000 estariam com 61 ou mais dias de atraso). Em maio de 2001, (4 meses após a concessão), 13,7% do volume de crédito concedido em janeiro estavam com 61 ou mais dias de atraso.

Antes da implementação do modelo, por exemplo, no mês de concessão de dezembro de 2000, a taxa de inadimplência três meses após a concessão do crédito (março) era de 8,8%. Quatro meses após a concessão (abril), era de 942,00 / 10.854,00 = 15,6%. Assim, os efeitos do modelo podem ser verificados comparando-se as taxas de inadimplência tomando-se os meses de concessão antes e depois da implementação do modelo. Por exemplo, a taxa de inadimplência três meses após a concessão (em dezembro) era de 8,8% (três meses depois = março) e, com a implementação do modelo, caiu para 8,0% (concessão em janeiro, três meses depois = abril), depois para 7,2% (concessões de fevereiro) e para 6,6% (concessões de março). Da mesma forma, a taxa de inadimplência quatro meses após a concessão caiu de 15,6% para 13,7% e depois para 10,7%. O mesmo pode ser feito para 5, 6 ou *n* meses após a concessão. Essa queda comparativa (diagonal) nos percentuais de inadimplência é o efeito esperado da implementação de um modelo de *credit scoring*.

Caso ao longo dos meses de análise (determinados nas colunas do relatório) as quedas nos percentuais de inadimplência deixem de ser observadas, passando a apresentar um aumento ou manutenção, é um indício de que o modelo está perdendo sua eficiência e precisa ser reestruturado ou que o *score* de corte da carteira precisa ser redefinido. No entanto, devem ser observados também os demais relatórios, de forma a saber exatamente a origem do comportamento das taxas de inadimplência. Por exemplo, se o relatório de acompanhamento de

variáveis e de taxa de rejeição não mostrarem flutuações relevantes e o relatório de inadimplência apresentar aumentos em suas taxas, então é mais recomendada uma reavaliação do *score* de corte do modelo do que a reconstrução do mesmo.

Relatório de Distribuição de Score

O relatório de distribuição de *score* obtém a proporção de pedidos de crédito (aceitos ou não) em cada um dos *scores* de 0 a 100, acompanhado periodicamente. Ele permite verificar se sua população de operações de crédito está mudando com o passar do tempo. Caso isso ocorra, deve-se estudar cuidadosamente se o motivo da mudança foi resultado, por exemplo, de uma campanha de marketing ou se ocorreu perda de discriminação do modelo.

Esse relatório consiste na construção de uma tabela que relaciona cada *score* com a proporção de operações de crédito que possuem *score* menor ou igual a este valor. Assim, permite identificar qual seria o efeito de uma mudança no *score* de corte sobre a taxa de rejeição da mesma carteira de crédito.

Capítulo VI. Conclusões

O método de análise de concessão de crédito proposto neste estudo mostrou que a implementação do modelo de *credit scoring* baseado na regressão logística seria capaz de classificar corretamente a grande maioria das operações de crédito de uma carteira de crédito específica para a qual o modelo foi gerado.

A primeira dificuldade que surge em qualquer método de análise de concessão de crédito, ou seja, em métodos de *credit scoring*, diz respeito à elaboração de um banco de dados em condições apropriadas para o estudo. É preciso obter e organizar um grande número de informações, com o maior número possível de operações de crédito e variáveis relativas às operações, sendo necessário observar as condições de preenchimento das variáveis e, caso necessário, realizar cortes nos dados sobre os quais desconfia-se da veracidade. Para gerar um modelo de *credit scoring*, é necessário que a instituição que pretende formular e utilizar o modelo conte com um sistema computacional de grande capacidade de armazenamento e processamento de dados, principalmente quando existirem diversas linhas de crédito para as quais devem ser elaborados modelos individuais que considerem as características de cada linha específica. O banco de dados utilizado no presente estudo continha um grande número de operações de crédito (201.075), mas algumas variáveis certamente muito importantes não foram disponibilizadas, tais como dados de escolaridade e apontamentos negativos (informações provenientes de agências de proteção ao crédito como SPC e Serasa). A ausência dessas

variáveis não prejudicou o modelo elaborado, mas recomenda-se que sempre que possível esses dados sejam utilizados.

A partir da definição de qualidade de crédito pelo critério de inadimplência, que analisa o comportamento de pagamentos das prestações a partir dos atrasos apresentados em operações de crédito da carteira realizadas no passado recente, foram considerados créditos ruins aqueles que apresentaram atraso de 61 ou mais dias no pagamento de prestação, sendo os créditos bons aqueles com atrasos de no máximo 60 dias. Todo o modelo estatístico, então, foi elaborado tendo como objetivo classificar operações de crédito de acordo com suas chances de apresentar inadimplência, tendo sido desprezados critérios de lucratividade que, conforme discutido, constituiriam uma ótica alternativa que difeririam do modelo sugerido apenas no estudo para a definição de qualidade de crédito. Mesmo desprezando tais critérios, é recomendado que seja feito paralelamente algum estudo verificando se o modelo baseado na inadimplência não trará resultados negativos quanto à lucratividade das operações de crédito: o modelo pode ser ao mesmo tempo muito bom para reduzir a inadimplência da carteira de crédito mas ruim para aumentar (ou manter) sua lucratividade. No entanto, dadas as estritas (e inversas) relações entre inadimplência e lucratividade, é esperado que o modelo seja condizente com os dois critérios, mesmo sendo gerado a partir de um deles somente.

Todas as variáveis disponibilizadas foram individualmente submetidas ao processo de categorização, que detectou grupos (categorias) de resposta homogêneos com relação à qualidade de crédito das operações baseado em testes estatísticos de agrupamento de dados. Das 43 variáveis explicativas originalmente recebidas, 37 puderam ser categorizadas (as demais não apresentaram padrões de homogeneidade) e passaram a ser representadas por variáveis *dummies* indicadoras de cada categoria de cada variável.

O modelo foi estimado a partir de uma amostra de 4.907 operações de crédito (sorteadas aleatoriamente entre as 201.075 operações disponibilizadas no banco de dados total) pelo método de regressão logística através do processo de escolha de variáveis explicativas denominado *forward stepwise*, que captou o efeito de uma quantidade satisfatória de variáveis (13 no total) e levou em consideração variáveis bastante representativas das operações de crédito, já que entre as 13 variáveis do modelo final houve variáveis de todas as modalidades: variáveis cadastrais dos clientes (profissão, residência própria, estado civil, quantidade de dependentes financeiros, idade na data de admissão no emprego e idade na data de concessão do crédito e tempo de residência), variáveis financeiras dos clientes (saldo em conta-corrente, saldo de aplicações e salário líquido), variáveis da operação de crédito (valor da operação e quantidade de prestações) e também variáveis indicadoras de restrições dos clientes (valor de cheques devolvidos e bloqueio na emissão de cheques). Os coeficientes estimados para as diversas categorias das 13 variáveis se mostraram estatisticamente significantes, bem como a capacidade classificatória do modelo, cujas taxas de acerto de classificação giraram em torno de 91% para operações de crédito boas e ruins. Assim, o método apresentado se mostrou estatisticamente confiável.

Os gráficos de distribuição dos *scores* estimados de clientes bons e ruins – gerados a partir do banco de dados total 201.075 operações e não somente pela amostra de modelagem de 4.907 operações – mostraram que tanto o número de clientes bons com *score* baixo como o número de clientes ruins com *score* alto é muito pequeno. Em outras palavras, os clientes bons ficaram concentrados nos *scores* mais altos e os clientes ruins nos *scores* mais baixos. Também pode-se notar que não houve flutuações graves na comparação dos resultados obtidos na amostra de modelagem e no banco de dados total, indício de que a amostra selecionada não estava viesada em relação ao banco de dados total. Também foi sugerido um estudo de estabilidade do

modelo, com o objetivo de verificar se alguma das 13 variáveis do modelo final seria capaz de determinar sozinha a concessão ou não do empréstimo, o que não é desejado. O modelo elaborado se mostrou bastante estável.

O ponto final do modelo foi a análise do *score* de corte para a linha de crédito, ou seja, o *score* a partir do qual uma operação de crédito deve ser aceita e até o qual deve ser rejeitada. A análise, baseada no princípio de minimização da probabilidade de classificação incorreta, sugeriu o *score* de corte de 51 pontos para a carteira de crédito CAB. No entanto, outras dimensões do problema devem ser consideradas para definir um *score* de corte mais apurado (caso sejam disponibilizados os dados necessários), tais como os impactos desse *score* sobre a rentabilidade da carteira e a concordância desse *score* com outros critérios e objetivos da linha de crédito em questão.

É importante ressaltar que modelos de *credit scoring* são baseados em amostras de concessões de crédito do passado recente que foram aprovadas, não levando em consideração aquelas que foram rejeitadas, já que para essas não se conhecem as respostas das variáveis e muito menos a qualidade de crédito da operação. Portanto, os bancos de dados geralmente sofrem de viés de seleção, já que não consideram todo o universo (população) de pessoas que requisitam crédito, e o efeito desse viés pode se manifestar em todas as estimativas do modelo (coeficientes, testes de hipóteses e poder classificatório viesados) sem que se saiba a magnitude e direção do viés. As tentativas de soluções apresentadas pelos estatísticos para o problema ainda não conseguiram resolvê-lo, sendo que linhas de pesquisa nesse sentido ainda devem sofrer muitos aprimoramentos. No entanto, mesmo com a existência do viés, é sabido que a maioria dos tipos de modelos de *credit scoring* (tais como análise discriminante, regressão linear, regressão logística, redes neurais etc.) mantém sua capacidade classificatória (que costuma ser equivalente

entre os diversos tipos de modelo) para a maioria das operações de crédito analisadas. Também devem ser consideradas as flutuações populacionais, ou seja, a evolução natural ou fruto de ciclos econômicos sobre as variáveis das operações de crédito, que tornam qualquer modelo estimado menos eficiente ao longo do tempo, exigindo sua reformulação periodicamente.

Assim que o modelo é implementado na prática pelas instituições credoras, é recomendável criar mecanismos periódicos (tais como os relatórios de acompanhamento) para que seja acompanhado o desempenho do modelo ao longo, enfocando o critério fundamental do modelo (evolução da inadimplência da carteira após a implementação do modelo) e também formas de verificar quando o modelo não é mais apropriado e precisa ser reformulado.

Em suma, apesar de todas as dificuldades (práticas ou técnicas) dos modelos de *credit scoring*, esses modelos consistem em ferramentas bastante válidas para avaliar a concessão de crédito de uma forma objetiva, racional e prática, tendo em vista que seu desempenho é certamente superior aos métodos de julgamento humano puro – que predominam em muitas instituições, onde os gerentes avaliam se concedem ou não o crédito baseados muitas vezes em critérios subjetivos – também são formas muito válidas para analisar créditos massificados (que tem muito potencial de crescimento no Brasil), pois estes exigem processamento de muitas informações com a maior velocidade possível e obviamente não podem ser prontamente avaliados por um pequeno grupo de pessoas. A estabilidade da economia é uma hipótese fundamental para o bom funcionamento dos modelos de *credit scoring*.

APÊNDICE I

Tabelas Cruzadas de Análise Descritiva

Tabela 1. Residência Própria

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	396	34995	35391
		1,10%	98,90%	100,00%
2	Possui	1295	106336	107631
		1,20%	98,80%	100,00%
3	Não Possui	1056	56997	58053
		1,80%	98,20%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 2. Idade (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	1	856	857
		0,10%	99,90%	100,00%
2	Até 24 anos	475	19555	20030
		2,40%	97,60%	100,00%
3	25 a 28 anos	389	19638	20027
		1,90%	98,10%	100,00%
4	29 a 38 anos	925	59179	60104
		1,50%	98,50%	100,00%
5	39 a 51 anos	716	59373	60089
		1,20%	98,80%	100,00%
6	52 a 58 anos	139	19855	19994
		0,70%	99,30%	100,00%
7	59 anos ou mais	102	19872	19974
		0,50%	99,50%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Idade na data de preenchimento do pedido de crédito, denominada "Idade Atual".

Tabela 3. Quantidade de Dependentes Financeiros

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing (Nenhum)	1789	132641	134430
		1,30%	98,70%	100,00%
2	1	319	23215	23534
		1,40%	98,60%	100,00%
3	2 ou 3	510	34620	35130
		1,50%	98,50%	100,00%
4	4 a 6	96	5753	5849
		1,60%	98,40%	100,00%
5	7 ou mais	49	2083	2132
		2,30%	97,70%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 4. Estado Civil

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Divorciado ou Missing	2	1214	1216
		0,20%	99,80%	100,00%
2	Solteiro	994	57278	58272
		1,70%	98,30%	100,00%
3	Marital, Desquitado ou	299	20072	20371
	Sep. Judicialmente (Consensual)	1,50%	98,50%	100,00%
4	Viúvo ou Separado Judicialmente	73	11944	12017
		0,60%	99,40%	100,00%
5	Casado	1379	107820	109199
		1,30%	98,70%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 5. UF de Nascimento

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	33	3971	4004
		0,80%	99,20%	100,00%
2	SP	2214	161768	163982
		1,40%	98,60%	100,00%
3	RR, AC, DF, TO, AM, RO, AP, PA, MA, MT, GO	36	1657	1693
		2,10%	97,90%	100,00%
4	SC, AL, RS, PI, RN, PB, ES	48	3902	3950
		1,20%	98,80%	100,00%
5	PR, MS	150	8590	8740
		1,70%	98,30%	100,00%
6	CE, SE, RJ, MG, BA	286	18420	18706
		1,50%	98,50%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 6. Sexo

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing		839	839
			100,00%	100,00%
2	Feminino	1073	89143	90216
		1,20%	98,80%	100,00%
3	Masculino	1674	108346	110020
		1,50%	98,50%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 7. E-Mail

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	174	16234	16408
		1,10%	98,90%	100,00%
2	Possui	12	1329	1341
		0,90%	99,10%	100,00%
3	Não Possui	2561	180765	183326
		1,40%	98,60%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 8. Salário Líquido

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	170	9699	9869
		1,70%	98,30%	100,00%
2	Até R$ 500,00	1334	73119	74453
		1,80%	98,20%	100,00%
3	R$ 500,01 ou mais	1243	115510	116753
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 9. Outras Rendas Mensais

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	2481	176947	179428
		1,40%	98,60%	100,00%
2	R$ 0,01 ou mais	266	21381	21647
		1,20%	98,80%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 10. Salário Líquido do Cônjuge

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	2537	180151	182688
		1,40%	98,60%	100,00%
2	R$ 0,01 ou mais	210	18177	18387
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 11. Tempo de Residência (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	25	2608	2633
		0,90%	99,10%	100,00%
2	Nulo	461	39368	39829
		1,20%	98,80%	100,00%
3	1 a 8 Meses	734	42477	43211
		1,70%	98,30%	100,00%
4	9 a 25 Meses	603	39831	40434
		1,50%	98,50%	100,00%
5	26 a 132 Meses	569	42187	42756
		1,30%	98,70%	100,00%
6	133 ou mais Meses	355	31857	32212
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Tempo de residência no último (atual) endereço.

Tabela 12. Idade na Data de Admissão no Atual Emprego (Data de Admissão Menos Data de Nascimento)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	530	52589	53119
		1,00%	99,00%	100,00%
2	Até 24 anos	770	43768	44538
		1,70%	98,30%	100,00%
3	25 a 32 anos	680	43755	44435
		1,50%	98,50%	100,00%
4	33 a 38 anos	416	29165	29581
		1,40%	98,60%	100,00%
5	39 a 44 anos	190	14564	14754
		1,30%	98,70%	100,00%
6	45 anos ou mais	161	14487	14648
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 13. Idade na Data de Admissão (Anos) Dividida Pela Idade Atual (Anos)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	530	52665	53195
		1,00%	99,00%	100,00%
2	0,000001 a 0,619241	178	17889	18067
		1,00%	99,00%	100,00%
3	0,619242 a 0,913546	1044	77785	78829
		1,30%	98,70%	100,00%
4	0,913547 a 1,000000	995	49989	50984
		2,00%	98,00%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 14. Cartão de Crédito "Credicard"

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Não Possui	2691	192084	194775
		1,40%	98,60%	100,00%
2	Possui	56	6244	6300
		0,90%	99,10%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 15. Quantidade de Cartões de Crédito

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum	2691	192084	194775
		1,40%	98,60%	100,00%
2	1	32	2755	2787
		1,10%	98,90%	100,00%
3	2	12	2161	2173
		0,60%	99,40%	100,00%
4	3 ou mais	12	1328	1340
		0,90%	99,10%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 16. Encargos Mensais (Compromissos Financeiros) (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	R$ 0,00 ou Missing	2586	188059	190645
		1,40%	98,60%	100,00%
2	Até R$ 66,47	48	2096	2144
		2,20%	97,80%	100,00%
3	R$ 66,48 a R$ 209,00	60	4217	4277
		1,40%	98,60%	100,00%
4	R$ 209,01 ou mais	53	3956	4009
		1,30%	98,70%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Gastos mensais com educação, aluguel, empréstimos diversos, financiamento imobiliário e de veículos.

Tabela 17. Quantidade de Imóveis (informada pelo cliente, com ou sem comprovação)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum ou Missing	2356	167945	170301
		1,40%	98,60%	100,00%
2	1	324	23938	24262
		1,30%	98,70%	100,00%
3	2 ou mais	67	6445	6512
		1,00%	99,00%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 18. Valor Comprovado de Imóveis

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nulo ou Missing	2652	190757	193409
		1,40%	98,60%	100,00%
2	Até R$ 22.500,00	46	2433	2479
		1,90%	98,10%	100,00%
3	R$ 22.500,01 ou mais	49	5138	5187
		0,90%	99,10%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 19. Quantidade Comprovada de Imóveis

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum ou Missing	2652	190757	193409
		1,40%	98,60%	100,00%
2	1	71	4792	4863
		1,50%	98,50%	100,00%
3	2 ou mais	24	2779	2803
		0,90%	99,10%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 20. Valor de Automóveis (informado pelo cliente, com ou sem comprovação)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nulo ou Missing	2493	178599	181092
		1,40%	98,60%	100,00%
2	Até R$ 6.000,00	157	9720	9877
		1,60%	98,40%	100,00%
3	R$ 6.000,01 ou mais	97	10009	10106
		1,00%	99,00%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 21. Quantidade de Automóveis (informada pelo cliente, com ou sem comprovação)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum ou Missing	2493	178599	181092
		1,40%	98,60%	100,00%
2	1	222	16716	16938
		1,30%	98,70%	100,00%
3	2 ou mais	32	3013	3045
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 22. Quantidade Total de Seguros (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum ou Missing	2730	196049	198779
		1,40%	98,60%	100,00%
2	1 ou mais	17	2279	2296
		0,70%	99,30%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Seguros de automóvel, vida e residencial.

Tabela 23. Quantidade de Seguros de Automóvel

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum ou Missing	2733	196746	199479
		1,40%	98,60%	100,00%
2	1 ou mais	14	1582	1596
		0,90%	99,10%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 24. Cheque (de qualquer instituição financeira ou banco)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Não Possui ou Missing	621	32532	33153
		1,90%	98,10%	100,00%
2	Possui	2126	165796	167922
		1,30%	98,70%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 25. Valor dos Cheques Devolvidos nos 6 Meses Anteriores à Concessão - Motivo 12 (Cheque Sem Fundo - 2a. Apresentação)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nulo (Nenhum cheque devolvido)	2541	191782	194323
		1,30%	98,70%	100,00%
2	R$ 0,01 a R$ 136,88	49	1980	2029
		2,40%	97,60%	100,00%
3	R$ 136,89 a R$ 257,00	41	1312	1353
		3,00%	97,00%	100,00%
4	R$ 257,01 a R$ 498,51	45	1311	1356
		3,30%	96,70%	100,00%
5	R$ 498,52 ou mais	71	1943	2014
		3,50%	96,50%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 26. Valor Total dos Cheques Devolvidos nos 6 Meses Anteriores à Concessão (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nulo (Nenhum cheque devolvido)	2539	191768	194307
		1,30%	98,70%	100,00%
2	R$ 0,01 a R$ 139,50	48	2002	2050
		2,30%	97,70%	100,00%
3	R$ 139,51 a R$ 767,20	104	3288	3392
		3,10%	96,90%	100,00%
4	R$ 767,21 ou mais	56	1270	1326
		4,20%	95,80%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Motivos 12 e 13 (cheques sem fundos na 2a. apresentação e conta encerrada, respectivamente).

Tabela 27. Profissão

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	98	3449	3547
		2,80%	97,20%	100,00%
2	Aposentado	320	25444	25764
		1,20%	98,80%	100,00%
3	Autônomo	720	48794	49514
		1,50%	98,50%	100,00%
4	Profissional Liberal	40	8872	8912
		0,50%	99,50%	100,00%
5	Funcionário Público	226	21109	21335
		1,10%	98,90%	100,00%
6	Desempregado	250	8177	8427
		3,00%	97,00%	100,00%
7	Funcionário de Empresa Privada	182	24250	24432
		0,70%	99,30%	100,00%
8	Do Lar	888	56524	57412
		1,60%	98,40%	100,00%
9	Outros	23	1709	1732
		1,30%	98,70%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 28. Saldo Médio em Conta-Corrente (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing (Não Correntista)	144	8299	8443
		1,70%	98,30%	100,00%
2	Até R$ 3,10	1409	46808	48217
		2,90%	97,10%	100,00%
3	R$ 3,11 a R$ 7,82	223	19086	19309
		1,20%	98,80%	100,00%
4	R$ 7,83 a R$ 66,33	745	76284	77029
		1,00%	99,00%	100,00%
5	R$ 66,34 a R$ 129,47	145	19070	19215
		0,80%	99,20%	100,00%
6	R$ 129,48 a R$ 364,81	57	19194	19251
		0,30%	99,70%	100,00%
7	R$ 364,82 ou mais	24	9587	9611
		0,20%	99,80%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Saldo médio em conta nos 3 meses anteriores à data de concessão.

Tabela 29. Saldo Médio de Aplicações (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing (Não Correntista)	144	8299	8443
		1,70%	98,30%	100,00%
2	Nulo	2097	112262	114359
		1,80%	98,20%	100,00%
3	R$ 0,01 a R$ 23,04	296	19026	19322
		1,50%	98,50%	100,00%
4	R$ 23,05 a R$ 81,15	130	19145	19275
		0,70%	99,30%	100,00%
5	R$ 81,16 a R$ 8.591,11	79	38470	38549
		0,20%	99,80%	100,00%
6	R$ 8591,12 ou mais	1	1126	1127
		0,10%	99,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) Saldo médio de aplicações nos 3 meses anteriores à data de concessão. Somente correntistas podem ter aplicações na instituição..

Tabela 30. Bloqueio da Emissão de Cheques

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Missing	144	8299	8443
		1,70%	98,30%	100,00%
2	Bloqueado pelo Gerente	1150	47748	48898
		2,40%	97,60%	100,00%
3	Não Bloqueado	1078	120525	121603
		0,90%	99,10%	100,00%
4	Bloqueado por Adesão a Pacote de Tarifas	375	21756	22131
		1,70%	98,30%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 31. Percentual de Taxa de Juros da Operação de Empréstimo (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	0,95% a 3,30%	20	20324	20344
		0,10%	99,90%	100,00%
2	3,31% a 3,90%	123	20863	20986
		0,60%	99,40%	100,00%
3	3,91% a 3,95%	14	40602	40616
		0,00%	100,00%	100,00%
4	3,96% a 4,50%	460	58932	59392
		0,80%	99,20%	100,00%
5	4,51% a 4,70%	543	27652	28195
		1,90%	98,10%	100,00%
6	4,71% a 5,00%	632	19728	20360
		3,10%	96,90%	100,00%
7	5,01% ou mais	955	10227	11182
		8,50%	91,50%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*)Variável não utilizável no modelo, visto que o percentual de juros é definido após aceitação do pedido de crédito ao crédito.

Tabela 32. Valor da Operação de Empréstimo

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Até R$ 630,00	672	44751	45423
		1,50%	98,50%	100,00%
2	R$ 630,01 a R$ 850,00	421	20101	20522
		2,10%	97,90%	100,00%
3	R$ 850,01 a R$ 1.500,00	944	69789	70733
		1,30%	98,70%	100,00%
4	R$ 1.500,01 ou mais	710	63678	64388
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 33. Quantidade de Prestações da Operação de Empréstimo

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	1 a 5	47	7274	7321
		0,60%	99,40%	100,00%
2	6 ou 7	325	30185	30510
		1,10%	98,90%	100,00%
3	8 a 11	346	23865	24211
		1,40%	98,60%	100,00%
4	12	1877	116137	118014
		1,60%	98,40%	100,00%
5	13 a 16	19	3586	3605
		0,50%	99,50%	100,00%
6	17 ou 18	116	8900	9016
		1,30%	98,70%	100,00%
7	19 ou mais	17	8381	8398
		0,20%	99,80%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 34. Referência Monetária da Operação

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Real	2605	194399	197004
		1,30%	98,70%	100,00%
2	TR	142	3929	4071
		3,50%	96,50%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 35. Forma de Pagamento do IOF da Operação

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	A vista - Deduz. Bruto	2266	155013	157279
		1,40%	98,60%	100,00%
2	Financiado	481	43315	43796
		1,10%	98,90%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

Tabela 36. Valor Total de Garantias da Operação (*)

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nulo	2736	195555	198291
		1,40%	98,60%	100,00%
2	R$ 0,01 ou mais	11	2773	2784
		0,40%	99,60%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

(*) A forma de cálculo do valor necessário de garantias e os tipos de garantia não foram fornecidos pela instituição para esse estudo. Nenhuma das garantias em questão depende da taxa de juros da operação de empréstimo.Não inclui avalistas.

Tabela 37. Quantidade de Avalistas da Operação de Empréstimo

Categoria		Qualidade de Crédito		
		Ruim	Bom	Total
1	Nenhum		151	151
			100,00%	100,00%
2	1	2163	152203	154366
		1,40%	98,60%	100,00%
3	2	499	37914	38413
		1,30%	98,70%	100,00%
4	3 ou mais	85	8060	8145
		1,00%	99,00%	100,00%
	Total	2747	198328	201075
		1,40%	98,60%	100,00%

BIBLIOGRAFIA

1. AVERY, R. B. *Credit scoring models with discriminant analysis and truncated samples*, 1977.

2. BENDEL, R. B.; AFIFI, A. A. *Comparison of stopping rules in forward regression*. Journal of the American Statistical Association, n.72, p. 46-53, 1977.

3. BIBOROSCH, R. A. *Numerical credit scoring*. Credit World, June, 1967.

4. BIERMAN, H.; HAUSMAN, W.H. *The credit granting decision*. Management Science, April, 1978.

5. BLUNDELL, R.; DUNCAN, A.; MEGUIR, C. *Estimating labor supply responses using tax reforms.* Econometrica, v.6, n.4, p.827-861, January 1998.

6. BORGES, L. F. X.; JUNIOR, S. B. *O risco legal na análise de crédito*. Revista do BNDES, Rio de Janeiro, v.8, n.16, p. 215-260, Dezembro 2001.

7. BROWN, C. C. *On a goodness-of-fit test for the logistic model based on score statistics*. Communications in Statistics, n.11, p. 1087-1105, 1982.

8. COX, D. R. *The Analysis of Binary Data*. Methuen, London, 1970.

9. BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica*. Métodos Quantitativos. Ed. 4, Editora Atual, São Paulo, 1987.

10. CAPON, N. *Credit scoring systems: a critical analysis*. Journal of Marketing, n.46, p. 82-91, 1982.

11. CHANDLER, G. G.; COFFMAN, J. Y. *A comparative analysis of empirical versus judgemental credit evaluation*. Journal of Retail Banking, v.1, n.2, p. 15-26, 1979.

12. CHANG, P. C., AFIFI, A. A. *Classification based on dichotomous and continuous variables*. Journal of the American Statistical Association, v.69, n.346, June 1974.

13. DAVIS, R. H.; EDELMAN, D. B.; GAMMERMAN, A. J. *Machine-Learning algorithms for credit-card applications*. IMA Journal of Mathematics Applied to Business Industry, n. 4, p.43-51, 1992.

14. DURAND, D. *Risk elements in consumer instalment financing*. National Bureau of Economic Research, New York, 1941.

15. EINSENBEIS, R. A. *Problems in applying discriminant analysis in credit scoring models*. Journal of Banking and Finance, n.2, p.211, North-Holland Publishing Company, 1978.

16. EWERT, D. C.; CHANDLER, G. G. *Credit formulas for loan extension*. Atlanta Economic Review, Jul-Aug 1974.

17. GREER, C. C. *The optimal credit acceptance scheme*. Journal of Financial Quantitative Analysis, n.3, p. 399-415, 1967.

18. HAND, D. J.; HENLEY; W. E. *Statistical classification methods in consumer credit ccoring: a review*. Journal of the Royal Statistical Society, v.160, part 3, p.523-541, 1997, Series A.

19. HAUCK, W. W.; DONNER, A. *Wald´s test as applied to hypothesis in logit analysis*. Journal of the American Statistical Association, n.72, p. 851-853, 1977.

20. HECKMAN, J. J. *Sample selection bias as a specification error.* Econometrica, v.47, n.1, p.153-161, January 1979.

21. HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*, Wiley Publication, USA, 1989.

22. JOANES, D. N. *Reject inference applied to logistic regression for credit scoring*. IMA Journal of Mathematics Applied to Business Industry, n.5, p. 35-43, 1993.

23. KOLESAR, P.; SHOWERS, J. L. *A robust credit screening model using categorical data*. Management Science, n.31, p. 123-133, 1985.

24. LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data*. Wiley, New York, 1987.

25. MAGIDSON, J. *The CHAID approach to segmentation modelling*, In: BAGOZZI, R. P. (ed.), *Advanced Methods of Marketing Research*, Cambridge (Massachusetts), p. 118-159, 1994.

26. MEHTA, D. *The formulation of credit policy models*. Management Science, n.15, p. 30-50, 1968.

27. MORAES, R. G. DE; BRANDI, V. R. *Mercado de crédito brasileiro: financiamento à exportação*. Cadernos Discentes COPPEAD, Rio de Janeiro, n.8, p. 112-149, 2001.

28. ORGLER, Y. E. *A credit scoring model for commercial loans*. Journal of Money, Credit and Banking, p. 435-445, Nov. 1970.

29. 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.

30. PINHEIRO, A. C.; MOURA, A. *Segmentação e uso de informações nos mercados de crédito brasileiros*. Rio de Janeiro, Fevereiro de 2001.

31. PREGIBON, D. *Logistic regression diagnostics*. Annals of Statistics, n.9, p. 705-724, 1981.

32. ROY, H. J. H; LEWIS, E. M. *Overcoming obstacles in using credit scoring systems*, Credit World, June 1970.

33. SRINIVASAN, V.; KIM, Y. H. *Credit granting: a comparative analysis of classification procedures*. The Journal of Finance, v. XLII, n.3, July 1987.

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

35. ZOCCO, D. P. *A framework for expert systems in bank loan management*. Journal of Commercial Bank Lending, n.6, p. 47-54.