Modelos de Risco de Crédito de Clientes: Uma Aplicação a Dados Reais

Gustavo Henrique de Araujo Pereira

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração: **Estatística**

Orientador: **Prof. Dr. Rinaldo Artes**

São Paulo, agosto de 2004

Modelos de Risco de Crédito de Clientes:

Uma Aplicação a Dados Reais

Este exemplar corresponde à redação final da
dissertação devidamente corrigida e defendida
por Gustavo Henrique de Araujo Pereira
e aprovada pela comissão julgadora.

São Paulo, 23 de agosto de 2004.

Banca examinadora:

- Prof. Dr. Rinaldo Artes (orientador) - IME-USP
- Profa. Dra. Lúcia Pereira Barroso - IME-USP
- Profa. Dra. Andrea Maria Accioly Fonseca Minardi - IBMEC

“Nunca ande pelo caminho traçado, pois ele
conduz somente até onde os outros já foram”.
(Alexander Grahan Bell)

À memória de meu pai Adevaldes
e à minha mãe Clélia.

Agradecimentos

Agradeço a Deus por ter me dado força para conciliar minhas atividades profissionais e o mestrado e também às seguintes pessoas:

Ao professor Rinaldo, que durante todas as etapas da elaboração desta dissertação esteve sempre presente e disposto a ler e reler cada uma das versões preliminares escritas, fazendo sempre sugestões que melhoraram significativamente este trabalho.

Aos meus pais Adevaldes e Clélia, que me mostraram desde cedo a importância da educação.

Às minhas irmãs Priscila e Tarsila, que sempre me incentivaram em todas as minhas decisões.

À minha namorada Patrícia, que me apoiou em todos os momentos deste trabalho, mesmo quando isso significava que passaríamos menos tempo juntos.

Ao Fernando e ao Marcelo, que permitiram o uso dos dados utilizados na aplicação deste trabalho.

Ao Alberto, ao Gizelton, ao Juscelino, à Marcela, à Miriam e ao William pelas diversas sugestões que enriqueceram este trabalho.

À Célia e ao Márcio pelas contribuições relacionadas à formatação de texto em Latex.

À Gerusa e à Kelly pelas contribuições relacionadas, respectivamente, à simulação e à GEE.

À Anthea pela disposição que mostrou em me ajudar a encontrar bibliografia sobre *customer scoring*.

Resumo

Modelos de *customer scoring* são utilizados para mensurar o risco de crédito de clientes de instituições financeiras. Neste trabalho, são apresentadas três estratégias que podem ser utilizadas para o desenvolvimento desses modelos. São discutidas as vantagens de cada uma dessas estratégias, bem como os modelos e a teoria estatística associada a elas. Modelos para cada uma das estratégias são ajustados utilizando-se dados reais obtidos de uma instituição financeira. A performance das estratégias para esse conjunto de dados é comparada a partir de medidas usualmente utilizadas na avaliação de modelos de risco de crédito. Uma simulação também é desenvolvida com o propósito de comparar o desempenho das estratégias em condições controladas.

Abstract

Customer scoring models are used to measure the credit risk of financial institution´s customers. In this work, we present three strategies that can be used to develop these models. We discuss the advantages of each of the strategies, as well as the models and statistical theory related with them. We fit models for each of these strategies using real data of a financial institution. We compare the strategies´s performance through some measures that are usually used to validate credit risk models. We still develop a simulation to study the strategies under controlled conditions.

Sumário

Capítulo 1

Introdução

A concessão de crédito tem papel fundamental na economia de um país. Cerca de dois terços do produto interno bruto (PIB) dos Estados Unidos, por exemplo, decorre do consumo[1]. Considerável parte deste, é financiada por instituições interessadas em conceder crédito em troca de um ganho sobre o capital emprestado. No Brasil, a indústria do crédito é, relativamente ao tamanho da economia, bem menor que a dos países desenvolvidos[2]. Porém, o crédito ao consumidor vem apresentando altas taxas de crescimento após a implantação do Plano Real e o controle da inflação[3].

As instituições que concedem crédito necessitam de um procedimento para decidir se emprestarão ou não capital a um proponente. Essa decisão é fundamental para o resultado financeiro da instituição, já que o lucro dos credores está diretamente associado à proporção de candidatos aprovados e ao percentual de clientes que pagam as dívidas contraídas.

A escolha dos proponentes que receberiam crédito era, até o início do século XX, baseada exclusivamente no julgamento de um ou mais analistas (Thomas et al., 2002). Em virtude disso, a aprovação de um pedido de crédito era subjetiva. Em uma mesma instituição, uma solicitação poderia ou não ser aprovada dependendo do analista que julgasse o pedido. Em 1936, Fisher (Fisher, 1936) desenvolveu a análise discriminante, técnica estatística que, a partir de características disponíveis de um

[1]Depois da guerra, confiança cai nos Estados Unidos. Folha de São Paulo, 14 jun. 2003. Caderno Dinheiro, p. B8.

[2]Explosão de crédito é receita para problema, diz Moody´s. Folha de São Paulo, 23 nov. 2003. Caderno Dinheiro, p. B4.

[3]BCB - Séries Temporais. Banco Central do Brasil. http://www4.bcb.gov.br/pec/series/port. Acesso em 10 jun. 2004.

indivíduo, cria uma regra de classificação que permite inferir a que população ele pertence. Isso permitiu o desenvolvimento dos primeiros modelos de credit scoring, que objetivam, a partir das características disponíveis de um proponente a crédito, ordenar os clientes quanto a probabilidade de pagar o empréstimo concedido. Inicialmente, a substituição da experiência dos analistas de crédito pela utilização de ferramentas estatísticas foi recebida com desconfiança. Porém, com o crescimento do número de propostas, percebeu-se que era inviável fazer a análise individual de cada uma delas. Diante da maior agilidade na decisão, menor custo, maior objetividade e até mesmo melhor poder preditivo, os modelos de credit scoring foram aos poucos se popularizando e atualmente são largamente utilizados (Hand e Henley, 1997). No Brasil, os primeiros modelos foram desenvolvidos na década de 1970. Após o Plano Real, o elevado crescimento observado na concessão de crédito contribuiu para a sua difusão, sendo hoje utilizados por praticamente todos os grandes credores do país.

1.1 Modelos de application e behavioural scoring

Os modelos utilizados na concessão de crédito a novos clientes são denominados *application scoring*. O seu principal objetivo é estimar a probabilidade de um indivíduo que está solicitando crédito se tornar inadimplente, antes de completar um período pré-fixado após a abertura de uma conta ou aquisição de um produto. Inúmeras técnicas já foram aplicadas para a construção de modelos desse tipo. A regressão logística talvez seja o método mais freqüentemente usado atualmente. Porém, já foram desenvolvidos modelos utilizando-se uma grande variedade de metodologias como análise discriminante (Hand et al., 1998), regressão linear (Orgler, 1970), modelos probito (Grablowsky and Talley, 1981), árvores de decisão (Arminger et al., 1997), programação matemática (Hand, 1981), sistemas especialistas (Showers and Chakrin, 1981), redes neurais (West, 2000), vizinho mais próximo (Henley e Hand, 1997), entre outras. Inúmeros artigos foram escritos comparando a performance dessas técnicas. Thomas (2000) traz a taxa de classificação correta obtida na utilização de algumas técnicas em diversos desses artigos. Sua conclusão é que, em relação à discriminação entre bons e maus clientes, não há diferença significativa entre as técnicas utilizadas. Os modelos de *application scoring*, em geral, são desenvolvidos utilizando uma amostra de clientes aprovados anteriormente pela in-

stituição. Porém, são utilizados na decisão sobre a concessão ou não de crédito a todos os proponentes, inclusive para aqueles que eram rejeitados anteriormente. Métodos que procuram corrigir esse viés amostral são conhecidos como inferência dos rejeitados. Basicamente eles consistem em inferir qual seria o comportamento dos indivíduos rejeitados caso eles tivessem sido aprovados. Hand (1998) e Feelders (1999) discorrem sobre o tema.

Em Rosa (2000) e Thomas et al. (2002) são descritas todas as etapas necessárias ao desenvolvimento de um modelo de *application scoring*. Já em Li e Hand (2002), a construção indireta de modelos é sugerida e comparada com a metodologia tradicional. Ela consiste na mudança da variável resposta do modelo. Na metodologia tradicional se estima a probabilidade de o cliente se tornar mau e, a partir do valor ajustado, pode-se classificá-lo como bom ou mau cliente. Neste método se prevêem algumas variáveis utilizadas na definição de bom e mau cliente e, a partir dessas estimativas, classifica-se o indivíduo.

Modelos de *behavioural scoring* são aqueles desenvolvidos para estimar a probabilidade de um cliente que já possui um determinado produto ter problema de crédito nos n meses seguintes. A grande vantagem desses modelos sobre os modelos de *application scoring* decorrem do fato deles possuírem um número maior de variáveis disponíveis para o ajuste. Além das variáveis disponíveis no momento da concessão, já se conhece o comportamento de utilização do produto pelo cliente. Dessa forma, é possível obter modelos com poder de discriminação bem superior aos observados em modelos de *application scoring* com a utilização, em geral, das mesmas metodologias. Apesar dessa vantagem, poucos artigos que discorrem especificamente sobre o assunto foram escritos. Hoper e Lewis (1992) descrevem como um modelo de *behavioural scoring* é geralmente utilizado. Blackwell e Sykes (1992) descrevem como esses modelos podem ser utilizados para a decisão de qual o limite de crédito o cliente terá direito. Ohtoshi (2003) compara a performance de várias metodologias utilizadas no desenvolvimento de modelos de *behavioural scoring*. A comparação é feita a partir do ajuste de modelos através de diferentes metodologias, utilizando um conjunto de dados de uma instituição financeira. Thomas et al. (2001) descreve como criar modelos de *behavioural scoring* utilizando cadeias de Markov. Neste último, ao invés de se utilizar as metodologias tradicionais, o cliente é classificado em um estado de acordo com algumas variáveis. A partir daí procura-se estimar a

probabilidade do cliente ir a um estado de *default* (inadimplência). É ressaltada a importância de segmentar a população, de modo a garantir que as probabilidades de transição de estados sejam semelhantes para todos os integrantes de um determinado grupo. Cadeias de segunda ou terceira ordem podem ser necessárias para que o processo se torne markoviano.

1.2 Modelos de customer scoring

Nos últimos anos, as instituições financeiras vêm procurando alterar sua organização interna, de modo a ter como foco o cliente e não os produtos do banco. Com essa mudança, passa-se a adotar estratégias integradas para cada grupo de clientes buscando a maximização do lucro da instituição. Isso evita ainda, dentre outros problemas, que duas áreas de produtos de crédito ou investimento de um banco procurem, ao mesmo tempo, convencer um cliente que o seu produto é o mais indicado no momento, fazendo com que a pessoa tenha uma imagem negativa da instituição.

O gerenciamento do risco de crédito baseado no foco no cliente também traz inúmeras vantagens. Previne, por exemplo, a concessão de um novo produto ou o aumento de limite em um já existente, para os clientes com atraso ou behaviour score de alto risco em um outro produto. Permite ainda um melhor controle dos limites disponíveis totais e valores emprestados ao cliente, evitando que eles atinjam quantias maiores que a pessoa tem condição de pagar. Com o crescimento do foco no cliente, surgiu a preocupação em consolidar o risco de crédito do cliente em cada um dos produtos (dados pelos modelos de *behavioural scoring*) em única medida, dando origem aos modelos de *customer scoring*. Trata-se de um modelo que objetiva ordenar os clientes quanto à probabilidade de ter problema de crédito em pelo menos um produto, dentro de um prazo pré-determinado. A grande vantagem dessa ferramenta é permitir uma visão global do risco do cliente, facilitando a criação de políticas de crédito mais adequadas para a instituição. Um banco que possua, por exemplo, três modelos de *behavioural scoring* de produto pode ter grande dificuldade em criar estratégias de gerenciamento do risco de crédito para cada um dos possíveis resultados do vetor de escores do cliente. A introdução do modelo de *customer scoring* facilita essa tarefa, pois substitui um vetor de três posições

por uma única medida. Isso não significa, porém, que os modelos de *behavioural scoring* de produtos devam ser descartados. Dentre outras utilidades, eles podem indicar se o comportamento em algum produto em particular é responsável pelo alto risco de crédito do cliente dado pelo modelo de *customer scoring*. Podem ainda atuar de forma complementar ao modelo de cliente na criação da política de crédito da instituição. Apesar da sua importância, pelo fato desses modelos serem uma preocupação recente dos grandes credores de crédito, não se tem conhecimento de algum artigo ou livro que discuta com profundidade aspectos técnicos dos modelos de *customer scoring*. A falta de literatura a respeito do assunto também tem relação com o interesse financeiro desses modelos. Uma instituição que desenvolva um eficiente modelo de *customer scoring*, em geral, não tem interesse em divulgá-lo para evitar que os concorrentes o utilizem e se beneficiem de seu bom desempenho. Esta dissertação objetiva o preenchimento desta lacuna. Thomas et al (2001) apenas apresenta o objetivo desses modelos. McNab e Winn (2000) discutem rapidamente o conceito, as componentes utilizadas no desenvolvimento, as vantagens e aplicações dos modelos de *customer scoring*. Já Groom e Gill (1998) discutem diversos aspectos importantes que devem ser observados no desenvolvimento de um modelo de *customer scoring*. São discutidos os requisitos necessários para o seu ajuste, os tipos de variáveis que devem estar presentes no modelo, o modo de definição da variável resposta e o tamanho do histórico de comportamento de crédito que deve ser utilizado. Além disso, são sugeridas três estratégias de desenvolvimento e apresentadas as situações em que cada uma delas é a mais indicada. Porém, não são abordados aspectos técnicos de desenvolvimento dos modelos.

1.3 Outros modelos para o segmento bancário

Uma outra preocupação recente dos pesquisadores e profissionais da área de risco de crédito é o desenvolvimento de um profit scoring (Oliver, 1993). Esses modelos objetivam ordenar os clientes quanto à probabilidade de dar lucro à instituição (ou ao valor desse lucro). Como o interesse final de um credor é o lucro, o profit scoring, se desenvolvido de forma adequada, traz ganhos extraordinários no gerenciamento do risco de crédito. Porém, como discutido por Thomas (2000), a sua construção é mais difícil do que se imagina em princípio e vários avanços ainda são necessários

para que se consiga um modelo adequado. Dentre as dificuldades, destaca-se a necessidade de se ter todas as componentes de receita e despesas presentes no cálculo da variável resposta. Isso deve incluir até mesmo despesas com marketing e recursos humanos e receitas com recuperações feitas pela área de cobrança. Outra dificuldade é que o lucro é diretamente influenciado pelas condições econômicas do país e, portanto, variáveis macroeconômicas devem estar presentes no modelo. Devem ainda ser consideradas questões como cancelamento do produto ou relacionamento pelo cliente e o fato de que, no desenvolvimento do modelo, não se conhece o lucro total dado por um indivíduo, mas apenas o ganho até o momento em que foi coletada a amostra.

Os modelos de análise de sobrevivência também têm sido discutidos na área de crédito. Ao invés de se estimar a probabilidade de um cliente se tornar problema de crédito, eles estudam o tempo necessário para que isso ocorra. Eles podem ser utilizados como uma componente importante de um profit scoring (Thomas et al., 2001), já que o tempo até o cliente se tornar inadimplente está diretamente associado ao lucro que ele dará a instituição. Possuem ainda, entre outras, a vantagem de não ser necessária a fixação de um período após o qual será observada a variável resposta, obtendo assim estimativas da probabilidade de que o cliente se torne problema de crédito após qualquer tempo fixado de interesse. Banasik et al. (1999) e Stepanova e Thomas (2001) indicam que a performance desses modelos, em geral, é semelhante ou apenas ligeiramente inferior aos modelos tradicionais de *scoring*, dependendo do período de observação considerado.

Além desses já discutidos, vêm sendo construídos modelos para estimar a probabilidade do cliente pagar um empréstimo que já está em atraso (*collection scoring*), fraudar a instituição (*fraud scoring*, Henley, 1995), comprar um produto após uma campanha de marketing (*propensity scoring*, Tsai e Yeh, 1999) e cancelar a conta ou um produto (*attrition scoring*). As metodologias são as mesmas utilizadas tradicionalmente nos modelos de *application* e *behavioural scoring* típicos.

1.4 Objetivo e estrutura da dissertação

Este trabalho compara três diferentes estratégias para o desenvolvimento de um modelo de *customer scoring*. As estratégias são desenvolvidas a partir de modelos

de regressão logística, sendo que em uma delas vários modelos são estimados simultaneamente, levando-se em conta a correlação existente entre as respostas (Equações de estimação generalizadas - GEE).

A GEE é uma técnica estatística proposta por Liang e Zeger (1986). Ela permite a estimação de modelos com dependência entre as respostas, quando as variáveis dependentes pertencem à família exponencial. Maiores detalhes serão apresentados no Capítulo 3.

No Capítulo 2 é apresentada a descrição dos dados e do estudo. O Capítulo 3 discute estratégias para o desenvolvimento de um modelo de *customer scoring*, bem como a teoria estatística associada a cada uma delas e medidas de performance utilizadas para compará-las. O Capítulo 4 traz uma aplicação das estratégias discutidas, a partir de um conjunto de dados reais fornecidos por uma instituição financeira. São ajustados os modelos de cada uma das estratégias e comparadas suas performances. Uma simulação com alteração da estrutura de correlação dos dados é apresentada no Capítulo 5. As conclusões do trabalho são discutidas no Capítulo 6.

Capítulo 2

Descrição do estudo

Neste capítulo serão discutidas as vantagens da substituição de modelos de *behavioural scoring* por um modelo de *customer scoring*, para a utilização no gerenciamento de crédito em uma instituição financeira. Além disso, o banco de dados que será utilizado na aplicação do Capítulo 4 será descrito na Seção 2.2.

2.1 Caracterização do problema

Uma instituição financeira possui um elevado número de clientes que solicitam diariamente novos produtos ou alterações em seu limite de crédito. Além disso, os bancos freqüentemente desejam disponibilizar linhas pré-aprovadas para seus clientes de menor risco de crédito, aumentando o número de contratos e conseqüentemente seu lucro. A decisão sobre a concessão ou não de novos produtos, linha pré-aprovada ou alteração de limites deve ser rápida. Não só para garantir a satisfação e conseqüente retenção dos clientes, como para o crescimento dos lucros, já que a partir da aprovação da operação o indivíduo já pode dar receita à instituição. Porém, para que essa decisão seja tão ágil quanto necessário, é preciso que o processo de decisão seja automatizado. Para que isso seja possível é fundamental uma ferramenta que ordene os clientes quanto ao risco de se tornarem inadimplentes em um horizonte de tempo pré-determinado. Para se atingir esse objetivo, freqüentemente são desenvolvidos modelos de *behavioural scoring*. Tais modelos utilizam variáveis que caracterizam o comportamento de crédito do cliente em um determinado produto, bem como características sócio-demográficas do indivíduo. Para o desenvolvimento

de um modelo de *behavioural scoring* utiliza-se o histórico de concessão de crédito da instituição. Uma amostra de clientes que possuía o produto em questão e não tinha nenhum problema de crédito em determinada data (ou em alguma das datas pré-definidas) é selecionada. Posteriormente, esses clientes são classificados quanto ao seu comportamento de crédito nos n meses seguintes. Em geral, os indivíduos são classificados em 4 grupos: bons, maus, indeterminados e cancelados (não possuem mais o produto ao final do período), de acordo com definição pré-fixada. A partir daí, usualmente se ajusta um modelo para resposta binária, descartando-se os clientes classificados como indeterminados ou cancelados. A previsão pontual obtida a partir desse modelo é a medida de risco (escore) do cliente naquele produto. Como, usualmente, se codifica os clientes bons com o valor um e os maus com o valor zero, quanto maior o score, menor o risco de crédito.

Os modelos de *behavioural scoring* trazem inúmeros ganhos à instituição. Em geral, quando o risco de crédito do cliente é alto, essa característica do indivíduo pode ser percebida a partir da análise de qualquer um dos produtos que ele possui. Dessa forma, para a maioria dos casos, um modelo que considera apenas variáveis de um único produto, além das sócio-demográficas, é suficiente para mensurar o risco de crédito do cliente. Para esses casos, se a instituição possuir vários modelos (um para cada produto), todos eles tenderão a indicar que o cliente tem alto risco de crédito. Essa associação entre o escore obtido a partir de produtos diferentes é bastante comum, independentemente do risco do cliente. Porém, existem casos em que o risco do cliente mensurado a partir de produtos diferentes não é semelhante. Nessa situação, caso a instituição possua um único modelo de *behavioural scoring*, ela pode classificar o cliente como sendo de baixo risco, quando variáveis em um outro produto indicam alta probabilidade de inadimplência. Mesmo que a instituição tenha diversos modelos, essa situação traz dificuldades, pois o cliente terá um escore alto em um produto e baixo nos demais, dificultando a sua classificação quanto ao risco de crédito. É comum, nessas situações, classificar o cliente no pior escore disponível. Porém, essa estratégia pode ocasionar perdas de oportunidades de lucro, já que pode-se negar crédito a clientes para os quais a probabilidade de inadimplência não é muito elevada. Suponha, por exemplo, que uma instituição tenha modelo de *behavioural scoring* para 3 produtos. Admita ainda que um cliente tenha um escore baixo em um produto e muito alto nos outros dois. Esse cliente, talvez, não possua

um risco de crédito muito elevado. Porém, se fosse utilizado a estratégia de classificar o cliente no pior escore disponível, ele poderia ter seu pedido negado, caso solicitasse um novo produto ou um aumento de limite.

Atualmente, a maioria dos grandes credores do país possui um elevado número de variáveis demográficas (características pessoais) e relacionadas ao uso de cada um dos produtos de crédito, armazenadas de modo a facilitar a sua utilização no desenvolvimento de modelos. Dessa forma, é possível o ajuste de um modelo de *customer scoring* (modelo de cliente) para estimar a probabilidade do cliente se tornar inadimplente em pelo menos um produto. A grande dificuldade no desenvolvimento de um modelo de *customer scoring* está no fato de que a maioria dos indivíduos não possui todos os produtos de crédito do banco. Mesmo agrupando-se os produtos em poucas famílias, ainda assim, muitos clientes podem não possuir pelo menos um contrato em cada um dos grupos. Dessa forma, o ajuste de um modelo de cliente diretamente a partir de todas as variáveis disponíveis não é possível, já que várias delas não são observadas para um elevado número de clientes . Diante dessa dificuldade, pelo menos três estratégias podem ser utilizadas para contornar o problema. O desenvolvimento de modelo em duas etapas (Estratégia 1) é a solução usualmente utilizada (Groom e Gill, 1998). Nessa dissertação estarão sendo propostas duas outras: o ajuste de vários modelos simultâneos para o modelo de cliente (Estratégia 2) e a obtenção simultânea não só do modelo de *customer scoring* como também de vários modelos de *behavioural scoring* (Estratégia 3). Nessa última, pelo fato de estar observando-se várias respostas em um mesmo indivíduo, há dependência entre as observações. Evidentemente essa dependência deve ser considerada na análise.

As estratégias requerem a estimação de vários modelos para respostas binárias. Devido a sua popularidade na área em estudo, será utilizada, nesse trabalho, a regressão logística para a obtenção desses modelos. Na Estratégia 3, os modelos serão estimados através de equações de estimação generalizadas (GEE), já que elas permitem o tratamento da dependência entre as observações. No Capítulo 3 será apresentada a teoria associada aos modelos das 3 estratégias.

2.2 Descrição dos dados

Um conjunto de dados reais obtido de uma instituição financeira será utilizado para a ilustração e comparação das estratégias de desenvolvimento de um modelo de *customer scoring*. Para esse trabalho, o conjunto de produtos de crédito sem garantia dessa instituição foi dividido em 3 famílias: cheque especial, cartão de crédito e outros produtos sem garantia. A população do estudo engloba todos os clientes que possuíam conta corrente e cheque especial ou cartão de crédito e não tinham nenhum problema de crédito em dezembro de 2001. Dessa população foi extraída uma amostra aleatória simples de 30.000 clientes, gerando assim a base de dados que será usada nesta dissertação.

Cada um dos clientes possuía desde nenhum até vários contratos em cada uma das famílias. Para cada um dos contratos foram obtidas diversas variáveis, caracterizando o comportamento de uso do produto pelo cliente em dezembro de 2001 e nos 5 meses anteriores. Por motivo de sigilo, os nomes e descrição de cada uma das variáveis não poderão ser apresentados. Elas serão representadas pela letra x, um número de 1 algarismo e outro de 2 algarismos. O primeiro número indica a qual família de produtos está associada aquela variável. A identificação da variável dentro de cada família é informada através do segundo número.

A família do cheque especial engloba os produtos de crédito que possuem um limite diretamente associado a uma conta corrente e com juros cobrados mensalmente em uma determinada data fixada. Para cada conta com limite, foram observadas 12 variáveis que serão denotadas nesse trabalho como $x_{101}, \ldots, x_{112}$.

Da família do cartão de crédito foram obtidas 9 variáveis para cada uma das contas cartão, que serão denotadas por $x_{201}, \ldots, x_{209}$. O número de contas cartão por cliente, na amostra de 30.000 indivíduos, varia de 0 a 5.

A família de outros produtos é a mais heterogênea internamente. A maioria dos produtos que a compõem é formada por créditos pagos em prestações mensais. Porém o prazo, e o valor concedido são bastante diferentes de acordo com o produto. Os clientes podem possuir vários contratos em cada um dos produtos que compõem essa família. Para cada um desses contratos de cada um dos clientes, foram obtidas 6 variáveis denotadas como $x_{301}, \ldots, x_{306}$.

Além das variáveis relacionadas ao comportamento do cliente em cada um dos produtos foram observadas 9 características de cada indivíduo. Essas variáveis não estão associadas a nenhum produto particular, sendo em sua maioria informações sócio-demográficas do indivíduo. Essas variáveis foram denotadas como $x_{601}, \ldots, x_{609}$ e completam a lista de variáveis preditoras.

A situação de cada um dos contratos em cada uma das famílias foi observada em junho de 2002. Cada um deles foi classificado em uma das 4 categorias da variável: mau, indeterminado, bom ou cancelado. A segmentação da situação do contrato em mau, indeterminado e bom está associada principalmente ao número de dias em atraso do cliente. O valor indeterminado é geralmente criado, porque a situação do contrato pode não ser tão ruim para classificá-la como mau nem tão confortável para caracterizá-la como boa. Contratos classificados como cancelados são aqueles que o cliente não possui mais aquele produto em junho de 2002. Apenas os contratos da família cheque especial e cartão de crédito podem assumir esse valor. No caso específico da família de outros produtos, o fato do cliente não possuir mais aquele contrato indica que ele pagou todas as suas prestações. Por isso, para essa família, se o cliente não possui mais aquele contrato, ele é classificado como bom. A situação de cada um dos contratos em junho de 2002 será utilizada na construção da variável dependente em algumas das estratégias e pode ser denominada **resposta conta**.

A partir da resposta conta é criada a variável resposta cliente. Para cada um dos indivíduos foi observada a situação de todos os seus contratos em junho de 2002. Foram avaliados não somente aqueles contratos existentes em dezembro de 2001 como também aqueles criados posteriormente. Também são considerados os produtos que não foram alocados em nenhuma das famílias, como por exemplo, os que possuem garantia. A **resposta cliente** é a pior situação do indivíduo entre todos os contratos que ele possui. A ordem de prioridade para a criação da variável é, da pior para a melhor, a seguinte: mau, indeterminado, bom. A classificação de cancelado ocorre quando todas as contas do cliente foram canceladas. Dessa forma cada contrato possui duas variáveis respostas, uma conta e a outra cliente. Já os indivíduos possuem uma resposta cliente e diversas respostas contas. A quantidade varia em função do número de produtos que ele possuía em junho de 2002.

Capítulo 3

Metodologia

Neste capítulo serão apresentadas as técnicas estatísticas utilizadas no desenvolvimento de modelos de *customer scoring*, a partir das estratégias mencionadas no Capítulo 2.

3.1 Regressão logística

A regressão logística (Hosmer e Lemeshow, 1989 e Collett, 1991) é uma técnica estatística utilizada para estudar a relação entre uma variável categorizada de interesse e um conjunto de outras disponíveis no estudo. Nesta dissertação será apresentada a teoria associada à regressão logística com variável resposta binária. Hosmer e Lemeshow (1989) descrevem a regressão logística com variável dependente com distribuição multinomial.

A regressão logística é um caso particular de modelos lineares generalizados (McCullagh e Nelder, 1989), com função de ligação logito e variável resposta para a unidade amostral i, y_i, com distribuição de Bernoulli de média μ_i. O modelo de regressão logística pode ser expresso como

$$g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^\top \beta \qquad \text{ou} \qquad \mu_i = \frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}$$

sendo que
$x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ é o vetor de dimensão $p+1$ de variáveis preditoras do indivíduo i e

$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ é o vetor de parâmetros do modelo.

Dados binários são freqüentemente analisados através dessa técnica devido a suas vantagens sobre outras. Uma primeira qualidade da regressão logística está no fato da ligação logito produzir valores ajustados que variam entre 0 e 1, propriedade bastante importante no estudo de dados binários. Outra grande vantagem da regressão logística está na interpretação dos parâmetros. Uma das principais estatísticas utilizadas na análise de dados binários é a razão de chances (Agresti, 1990), sendo que a chance pode ser definida como $P(y_i = 1/x_i)/P(y_i = 0/x_i)$. Na regressão logística, e^{β_j} é a razão entre a chance de y_i ser um em relação a ser zero quando $x_{ij} = a + 1$ e quando $x_{ij} = a$. Dessa forma, se $e^{\beta_j} = K$, então, mantidas todas as demais variáveis preditoras constantes, K é o valor pelo qual é multiplicado a chance, quando aumentamos x_{ij} em uma unidade. Além disso, pode-se facilmente ajustar um modelo de regressão logística, já que ela está implementada em todos os principais softwares de análise estatística. Na comparação com outros métodos que podem ser utilizados para a discriminação de dois grupos, a regressão logística possui ainda a vantagem de não possuir suposições fortes. A análise discriminante usual, por exemplo, exige que a distribuição conjunta das variáveis utilizadas na discriminação de dois grupos seja normal multivariada (Johnson e Wichern, 1998).

Os parâmetros da regressão logística são geralmente estimados por máxima verossimilhança. A verossimilhança de uma amostra aleatória de variáveis binárias independentes de média μ_i é dada por

$$L(Y/\beta) = \prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i}$$

na qual $\mu_i = \left(\frac{e^{x_i^\top \beta}}{1+e^{x_i^\top \beta}}\right)$, no caso da regressão logística. O sistema de equações obtidas derivando-se a log verossimilhança em relação a β e igualando a 0 não é linear. Em virtude disso, não é possível obter uma expressão fechada para os estimadores de β. É necessária a utilização de métodos numéricos para a sua solução. O mais utilizado é o mínimos quadrados reponderados (Paula, 2000). A variância assintótica do estimador de máxima verossimilhança é igual ao inverso da informação de Fisher. Utilizando-se essa propriedade, obtém-se que, para a regressão logística, a variância assintótica de $\widehat{\beta}$ é igual a

$$(X^\top V X)^{-1}$$

na qual, $V = \text{diag}\{\mu_1(1-\mu_1), \mu_2(1-\mu_2), \ldots, \mu_n(1-\mu_n)\}$ e $X = (x_1, x_2, \ldots, x_n)^\top$.

Inferências sobre a significância dos parâmetros podem ser feitas, principalmente, a partir de três testes. O mais recomendável deles, segundo Hauck e Donner (1977), é o teste de razão de verossimilhanças. Para testar a hipótese
$H_0 : \beta_k = 0, \beta_{k+1} = 0, \ldots, \beta_{k+j} = 0 \quad \times$
H_1 : pelo menos um dos parâmetros é diferente de 0,
define-se a estatística

$$D_{RV} = -2\{\ln L(Y/\beta, H_0) - \ln L(Y/\beta, H_1)\}$$

em que
$\ln L(Y/\beta, H_0)$ é a log verossimilhança do modelo sob H_0
$\ln L(Y/\beta, H_1)$ é a log verossimilhança do modelo sob H_1.
Sob H_0, para grandes amostras, a estatística D_{RV} tem distribuição $\chi^2(j)$ na qual
j = (número de parâmetros sob H_1 - número de parâmetros sob H_0).
Os outros dois testes bastante utilizados na regressão logística são o teste de Wald (Grizzle et al., 1969) e Escore (Dobson, 1983).

3.1.1 Qualidade do ajuste

A qualidade do ajuste de uma regressão logística, assim como nos demais modelos lineares generalizados, pode ser avaliada a partir da função desvio, que é definida como

$$D = -2(\ln \widehat{L}_m - \ln \widehat{L}_s)$$

sendo que
$\widehat{L}_m$ é a verossimilhança do modelo sob investigação aplicada na estimativa de máxima verossimilhança.
$\widehat{L}_s$ é a verossimilhança do modelo saturado, que contém n parâmetros sendo que n é igual ao número de observações.
Considere
k como sendo o número de combinações diferentes das variáveis preditoras observadas na amostra. Se há, por exemplo, apenas 2 variáveis preditoras dicotômicas,

k vale no máximo 4. Quando há variáveis independentes contínuas no modelo, é comum observarmos $k = n$.
n_i o número de observações da i-ésima combinação de variáveis preditoras.
y_i^* o número de observações dentre as n_i para as quais $y_i = 1$.
Utilizando-se as medidas definidas e desenvolvendo-se a expressão acima, obtém-se que para dados binários

$$D = 2\sum_{i=1}^{k}[y_i^* ln(y_i^*/n_i\widehat{\mu}_i) + (n_i - y_i^*)ln\{(1 - y_i^*/n_i)/(1 - \widehat{\mu}_i)\}]$$

sendo que $\hat{\mu}_i = \left(\frac{e^{x_i^\top \hat{\beta}}}{1+e^{x_i^\top \hat{\beta}}}\right)$. O i-ésimo termo da expressão é substituído por $-2n_i \ln(1-\widehat{\mu}_i)$ quando $y_i = 0$ e por $-2n_i \ln \widehat{\mu}_i$ quando $y_i = n_i$. Valores pequenos de D indicam que o modelo está bem ajustado. Sob a hipótese de que o modelo é adequado, a função D tem distribuição assintótica $\chi^2(k-p)$ na qual p é o número de parâmetros do modelo em investigação. O grande inconveniente da função desvio é que sua distribuição é aproximadamente $\chi^2(k-p)$ apenas quando os n_k´s são grandes (Hosmer e Lemeshow, 1989). Tal situação ocorre somente se $k << n$. Dessa forma, mesmo para estudos nos quais a amostra é grande, a aproximação pode não ser boa, caso haja um elevado número de variáveis preditoras ou pelo menos uma delas seja contínua.

Em situações nas quais $k \approx n$ é mais conveniente utilizar o teste de Hosmer e Lemeshow para verificar o ajuste do modelo (Hosmer and Lemeshow, 1980). Para ajustá-lo, divide-se a população do estudo em g grupos de tamanhos semelhantes, a partir dos percentis das probabilidades ajustadas. Se, por exemplo, deseja-se criar 10 grupos, eles seriam divididos a partir dos decis das probabilidades ajustadas. Compara-se então a diferença entre o número observado e esperado de ocorrências de $y = 1$ em cada um dos g grupos, a partir da estatística dada por

$$\widehat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k\bar{\mu}_k)^2}{n'_k\bar{\mu}_k(1 - \bar{\mu}_k)}$$

na qual
n'_k é o número de combinações de variáveis preditoras observadas no k-ésimo grupo
$o_k = \sum_{j=1}^{n'_k} y_j^*$
$\bar{\mu}_k = \sum_{j=1}^{n'_k} n_j\widehat{\mu}_j/n'_k$
Quando $k \approx n$ e sob a hipótese de que modelo está bem ajustado, $\widehat{C}$ tem distribuição

aproximada $\chi^2(g-2)$. Outras medidas de diagnóstico utilizadas para a regressão logística podem ser vistas em Hosmer e Lemeshow (1989).

3.2 Equações de estimação generalizadas

As equações de estimação generalizadas foram introduzidas por Liang e Zeger (1986). Trata-se de uma extensão dos modelos lineares generalizados para as situações nas quais há dependência entre as observações. Essa técnica pode ser desenvolvida a partir da teoria de funções de estimação (Artes, 1997 e Jørgensen e Labouriau, 1994) que é apresentada a seguir.

3.2.1 Funções de estimação

Uma função de estimação é uma função dos dados e dos parâmetros de interesse. Elas são construídas de modo que suas raízes, quando existirem, sejam estimadores dos parâmetros em estudo. Dessa forma, é importante o estabelecimento de condições que garantam que os estimadores obtidos possuam boas propriedades, como consistência e distribuição assintótica conhecida. As funções de estimação serão denotadas por $\Psi(y;\theta)$ e dado θ, $\Psi(y;\theta)$ é uma variável aleatória[1].

Assumindo-se a existência de uma amostra de n unidades amostrais independentes, sendo que para cada uma delas obtém-se t medidas $(y_i = (y_{i1}, y_{i2}, \ldots, y_{it})^\top, i = 1, 2, \ldots, n)$, pode-se associar a cada um desses vetores aleatórios, uma função de estimação $\psi_i(y_i;\theta) = (\psi_{i1}, \psi_{i2}, \ldots, \psi_{ip})^\top$. Dessa forma, estende-se o conceito de função de estimação para a amostra através de

$$\Psi_n(y;\theta) = \sum_{i=1}^{n} \psi_i(y_i;\theta).$$

O estudo das propriedades de uma função de estimação requer algumas definições que são apresentadas a seguir. Uma função de estimação $\Psi(y;\theta)$ é dita não viciada quando

$$E_\theta(\Psi(y;\theta)) = 0.$$

[1]Formalmente, seja $(\chi, \mathcal{A}, \mathcal{P})$ um espaço de probabilidade, com $\chi \in \mathbf{R}$ e $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbf{R}^p\}$. Por definição, uma função $\Psi : \chi \times \Theta \to \mathbf{R}^p$ é uma função de estimação se para cada $\theta \in \Theta$, $\Psi(.;\theta) = (\psi_1, \ldots, \psi_p)^\top$ é uma função mensurável.

É fácil verificar que se cada uma das $\psi_i(y_i;\theta)$ forem não viciadas, então $\Psi_n(y;\theta)$ também será não viciada.

A matriz de variabilidade de uma função de estimação não viciada é definida como

$$V_\Psi(\theta) = E_\theta(\Psi(y;\theta)\Psi^\top(y;\theta)),$$

enquanto a matriz de sensitividade é dada por

$$S_\Psi(\theta) = E_\theta\left(\frac{\partial}{\partial\theta^\top}\Psi(y;\theta)\right).$$

Ambas possuem dimensão $(p \times p)$.

Uma função de estimação $\Psi(y;\theta)$ é dita regular se, para todo $\theta = (\theta_1, \ldots, \theta_p) \in \Theta$ e para todo $i, j, k, l = 1, 2, \ldots, p$, ela satisfizer as seguintes condições:

1. é não viciada;
2. $\partial\Psi(y;\theta)/\partial\theta_i$ existe quase certamente para todo $y \in \chi$;
3. é possível permutar o sinal de integração e diferenciação da seguinte forma:

$$\frac{\partial}{\partial\theta_i}\int_\chi \Psi(y;\theta)p(y;\theta)dy = \int_\chi \frac{\partial}{\partial\theta_i}\{\Psi(y;\theta)p(y;\theta)\}dy;$$

   sendo que $p(y;\theta)$ é a função densidade de probabilidade de y;
4. $E_\theta(\psi_i(y;\theta)\psi_j(y;\theta)) \in \mathbf{R}$ e $V_\Psi(\theta)$ é positiva definida;
5. $E_\theta\left(\frac{\partial}{\partial\theta_l}\psi_i(y;\theta)\frac{\partial}{\partial\theta_k}\psi_j(y;\theta)\right) \in \mathbf{R}$ e $S_\Psi(\theta)$ é não singular.

Seja $\Psi(y;\theta)$ uma função de estimação regular. A matriz de informação de Godambe de θ associada a $\Psi(y;\theta)$ é definida como

$$J_\Psi(\theta) = S_\Psi^\top(\theta)V_\Psi^{-1}(\theta)S_\Psi(\theta).$$

A informação de Godambe desempenha para as funções de estimação regulares o mesmo papel que a informação de Fisher desempenha para a função escore regular. Para esta função, $S_\Psi(\theta) = -V_\Psi(\theta)$ e portanto as duas matrizes de informação coincidem.

Em Godambe (1960), foi desenvolvido o conceito de otimalidade de uma função de estimação regular. No caso uniparamétrico, uma função de estimação ótima é

aquela cujas raízes possuem variância assintótica mínima. Para definir uma função de estimação ótima para o caso multidimensional, é necessário criar uma ordenação entre matrizes, já que a matriz de variância e covariância possui nesse caso, dimensão $(p \times p)$, sendo que p é a dimensão do vetor de parâmetros. A ordenação entre matrizes introduzida por Godambe é apresentada abaixo.

Seja J_{Ψ}^{-1} a matriz de covariância assintótica de um estimador $\widehat{\theta}$ obtido a partir de uma função de estimação Ψ. Se existir uma função de estimação Ψ^* que gera um estimador de θ com matriz de covariância assintótica $J_{\Psi^*}^{-1}$, tal que $J_{\Psi}^{-1} - J_{\Psi^*}^{-1}$ seja não negativa definida para qualquer Ψ, então Ψ^* é dita ser uma função de estimação ótima para θ.

É possível demonstrar que, no caso regular, a função escore será sempre ótima. Dessa forma, quando a informação de Fisher existir, ela exerce o papel de limite superior para as matrizes de informação de Godambe. Pode-se mostrar ainda que se Ψ^* é ótima, então todas as funções de estimação obtidas do produto entre $\mathrm{C}(\theta)$ e Ψ^* também o serão, sendo que $\mathrm{C}(\theta)$ é uma matriz não singular e não estocástica.

Crowder (1987) estuda a otimalidade em uma classe particular de funções de estimação definidas como

$$\pounds(u) = \left\{ \Psi_n(\theta) \in R : \Psi_n(\theta) = \sum_{i=1}^{n} Q_i(\theta) u_i(y_i; \theta) \right\},$$

sendo que $Q_i(\theta)$ são matrizes não estocásticas de postos completos, $u_i(y_i; \theta)$ são vetores mutuamente independentes com média 0 e R contém todas as funções regulares de θ. Ela é denominada classe das funções de estimação lineares geradas por u_i e tem como função de estimação ótima em R

$$\sum_{i=1}^{n} Q_i^*(\theta) u_i(y_i; \theta)$$

na qual

$$Q_i^* = E_\theta \left(\frac{\partial u_i}{\partial \theta} \right)^\top Cov_\theta^{-1}(u_i).$$

A utilização do resultado de Crowder para diferentes modelos de regressão resulta em estimadores amplamente conhecidos e utilizados conforme pode ser visto nos exemplos abaixo.

Sejam $y_1, ..., y_n$ variáveis aleatórias independentes com $Var(y_i) = \sigma^2$, $E(y_i) = \mu_i = x_i^\top \beta$ com x_i sendo vetores não aleatórios de variáveis preditoras e $X = (x_1, \cdots, x_n)^\top$. A função de estimação ótima na classe $\pounds(y-\mu)$, com $y = (y_1, \cdots, y_n)^\top$ e $\mu = (\mu_1, \cdots, \mu_n)^\top$ é dada por

$$\Psi_n^* = \sigma^{-2} X^\top (y - X\beta).$$

Igualando-se Ψ_n^* a zero, obtém-se as equações normais, ou seja, a partir desta função de estimação obtém-se estimadores equivalentes aos obtidos na aplicação do método de mínimos quadrados aos dados.

Considere agora uma amostra de variáveis aleatórias independentes, y_i, $i = 1, 2, \ldots, n$, de tal modo que $E(y_i) = \mu_i = h_i(x_i^\top \beta) = h_i(\eta_i)$ e $Var(y_i) = \sigma^2 v_i(\mu_i)$, sendo que x_i é um vetor de variáveis preditoras associado a y_i; β é um vetor p-dimensional de parâmetros desconhecidos; $h_i(\cdot)$ é uma função duplamente diferenciável e inversível e $v_i(\cdot)$ é uma função positiva, $i = 1, 2, \ldots, n$. Nesse caso, pode-se provar que a função de estimação ótima em $\pounds(y - \mu)$ é dada por

$$\Psi_n^*(\beta) = -\sigma^{-2} X^\top H V^{-1} (y - \mu),$$

na qual $X = (x_1, \ldots, x_n)^\top$, $H = diag\{\frac{\partial \mu_1}{\partial \eta_1}, \ldots, \frac{\partial \mu_n}{\partial \eta_n}\}$ e $V = diag\{v_1(\mu_1), \ldots, v_n(\mu_n)\}$. Note que $\Psi_n^*(\beta)$ equivale às equações de estimação sugeridas na teoria da quase-verossimilhança, desenvolvida por Wedderburn (1974).

Suponha agora que para cada unidade amostral i sejam avaliadas t variáveis aleatórias, $y_{ij}, j = 1, \ldots, t$. Seja então $y_i = (y_{i1}, y_{i2}, \ldots, y_{it})^\top, i = 1, 2, \ldots, n$ uma amostra de vetores independentes de variáveis aleatórias e ${x_i}^\top = (x_{i1}, x_{i2}, \ldots, x_{it})$ sendo que x_{ij} são vetores não aleatórios p dimensionais de variáveis preditoras. Considere ainda $E(y_{ij}) = \mu_{ij} = h(x_{ij}^\top \beta) = h(\eta_{ij})$, $Var(y_{ij}) = \phi^{-1} v(\mu_{ij})$ e $cor(y_i) = \Gamma(\mu_i)$ sendo que β é um vetor de dimensão p de parâmetros desconhecidos; $h(\cdot)$ é uma função duplamente diferenciável e inversível e $v(\cdot)$ é uma função positiva, $i = 1, 2, \ldots, n$. Nessas condições, pode-se provar que a função de estimação ótima em $\pounds(y_i - \mu_i)$ é dada por

$$\Psi(\beta) = \phi \sum_{i=1}^{n} D_i^\top W_i^{-1} (y_i - \mu_i)$$

na qual $\mu_i = (\mu_{i1}, \ldots, \mu_{it})^\top$, $D_i = x_i^\top \text{diag}\left\{\frac{\partial \text{h}(\eta_\text{i})}{\partial \eta_\text{i}}\right\}$, $W_i = Cov(y_i) = \phi^{-1} A_i^{1/2} \Gamma(\mu_i) A_i^{1/2}$, com $A_i = diag\{v(\mu_{i1}), \ldots, v(\mu_{it})\}$. Esta função é conhecida como quase-escore multivariada e é utilizada na teoria de análise de quase-verossimilhança multivariada.

Na teoria, a análise de quase-verossimilhança multivariada poderia ser utilizada para o ajuste de modelos para situações nas quais há dependência entre as observações. Porém, na prática, ela é uma teoria de difícil utilização por vários motivos. Para que um modelo de quase-verossimilhança multivariado seja ajustado, é necessário modelar a correlação em função da média o que, em geral, não é intuitivo. No entanto, o maior problema é garantir que Γ seja uma matriz de correlação. Isso significa que é necessário que todas os elementos fora da diagonal principal variem entre -1 e 1 e que ela seja positiva definida. Há algumas funções que variam entre -1 e 1, que poderiam ser utilizadas na modelagem da correlação, porém, elas não garantem que a segunda condição será satisfeita. Uma forma de garantir essa condição é trabalhar com correlações parciais. Se t = 3, por exemplo, pode-se estimar a correlação entre as respostas 1 e 2 e 1 e 3 e modelar o outro parâmetro através da correlação parcial entre y_{i2} e y_{i3} eliminado-se o efeito de y_{i1}. No entanto, essa modelagem é difícil e sua dificuldade ainda aumenta bastante à medida que t cresce.

3.2.2 Equações de estimação generalizadas

Os problemas práticos da quase-verossimilhança multivariada foram resolvidos com o desenvolvimento das equações de estimação generalizadas (GEE) por Liang e Zeger (1986). Elas permitem o ajuste de modelos para as situações nas quais mais de uma observação é tomada em uma mesma unidade amostral gerando assim uma dependência entre elas. Observações de unidades amostrais diferentes são supostas independentes e a distribuição marginal da resposta pertence à família exponencial.

Seja t_i o número de observações obtidas para o indivíduo i. Defina $y_i = (y_{i1}, y_{i2}, \ldots, y_{it_i})^\top$, $i = 1, 2, \ldots, n$ vetores independentes de variáveis aleatórias e assuma que y_{ij} pertence à família exponencial. Seja ainda $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^\top$ vetor de variáveis preditoras para a observação j da unidade amostral i e $x_i = (x_{i1}, x_{i2}, \ldots, x_{it_i})^\top$. Admita também que $E(y_{ij}) = \mu_{ij}$, $Var(y_{ij}) = \phi^{-1}v(\mu_{ij})$ e $cor(y_i) = \Gamma(u_i)$ e defina $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{it_i})^\top$. Para facilitar a notação, será assumido, sem perda de generalidade, que $t_i = t$, $i = 1, 2, \ldots, n$.

Para a modelagem de μ_{ij} serão utilizadas as mesmas convenções usadas nos

modelos lineares generalizados. Seja então

$$g(\mu_{ij}) = x_{ij}^{\top}\beta = \eta_{ij},$$

na qual
$g : \mathcal{R} \to \mathcal{R}$ uma função inversível, duplamente diferenciável e denominada função de ligação e
$\beta = (\beta_1, \beta_2, \ldots, \beta_p)^{\top}$ é um vetor de parâmetros.

A função de estimação ótima para β em $\pounds(y_i - \mu_i)$ é dada por $\Psi_n^*(\beta) = \sum_{i=1}^{n} \psi_i^*(\beta)$, na qual

$$\psi_i^*(\beta) = D_i^{\top} W_i^{-1} u_i = X_i^{\top} H_i W_i^{-1} u_i, \tag{3.1}$$

com $W_i = Cov(u_i) = \phi^{-1} A_i^{1/2} \Gamma(u_i) A_i^{1/2}$,
$A_i = diag\{v(\mu_{i1}), \cdots, v(\mu_{it})\}$,
$u_i = y_i - \mu_i$ e
$H_i = diag\left\{\frac{\partial h(\eta_i)}{\partial \eta_i}\right\}, h = g^{-1}$.

Embora a função (3.1) seja ótima entre as lineares geradas por $y_i - \mu_i$, ela tem pouca utilidade prática. A matriz $\Gamma(u_i)$ é a verdadeira matriz de correlação de y_i que, em geral, é desconhecida. A modelagem de $\Gamma(u_i)$ em função de μ_i corresponde à quase-verossimilhança multivariada cujos problemas foram discutidos na Seção 3.2.1.

A solução de Liang e Zeger (1986) foi substituir $\Gamma(u_i)$ pela matriz $R(\alpha)$ denominada matriz de correlação de trabalho, sendo que α é um vetor de dimensão s que caracteriza completamente $R(\alpha)$. Admita, por exemplo, um caso em que $t = 3$ e no qual supõe-se que as correlações entre as variáveis sejam iguais. Tem-se então

$$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_1 \\ \alpha_1 & 1 & \alpha_1 \\ \alpha_1 & \alpha_1 & 1 \end{pmatrix} \qquad \text{e} \qquad \alpha = [\alpha_1].$$

Caso seja admitido correlações diferentes entre as variáveis, tem-se

$$R(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_1 & 1 & \alpha_3 \\ \alpha_2 & \alpha_3 & 1 \end{pmatrix} \qquad \text{e} \qquad \alpha = [\alpha_1, \alpha_2, \alpha_3]^{\top}.$$

O termo correlação de trabalho vem do fato de $R(\alpha)$ não precisar, necessariamente, ter a mesma estrutura de correlação de Γ, bastando ter apenas as propriedades de

uma matriz de correlação. O vetor α é tratado como um vetor de parâmetros de perturbação. Liang e Zeger (1986) sugeriram uma alteração em Ψ_n^*, obtendo assim a função de estimação

$$\Psi_n^G(\beta) = \sum_{i=1}^{n} D_i^\top \hat{\Omega}_i^{-1} u_i \tag{3.2}$$

em que $\hat{\Omega}_i = \hat{\phi}^{-1} A_i^{1/2} R(\hat{\alpha}) A_i^{1/2}$ e na qual $\hat{\phi} = \hat{\phi}(\beta)$ e $\hat{\alpha} = \hat{\alpha}(\beta, \hat{\phi}(\beta))$ são estimadores de ϕ e α, respectivamente, que dependem apenas de β. Dessa forma, note que $\Psi_n^G(\beta)$ é função apenas de β. O Teorema (3.1) traz as condições sob as quais a raiz de Ψ_n^G é um estimador consistente e assintoticamente normal de β.

Teorema 3.1 *Seja $\hat{\beta}_n$ a raiz de (3.2). Sob condições gerais de regularidade, com $|\hat{\beta}_n - \beta| = O_p(1)$ e assumindo que*

a. *$\hat{\alpha}(\beta, \phi^{-1})$ é um estimador $\sqrt{n}$-consistente de α dados β e ϕ^{-1};*

b. *$\hat{\phi}^{-1}(\beta)$ é um estimador $\sqrt{n}$-consistente de ϕ^{-1} dado β;*

c. *$\left|\frac{\partial \hat{\alpha}(\beta,\phi^{-1})}{\partial \phi^{-1}}\right| \leq H(y, \beta)$, na qual $H(y, \beta)$ é uma função $O_p(1)$ de β e dos dados;*

então $\hat{\beta}_n$ é um estimador consistente de β e

$$n^{1/2}(\hat{\beta}_n - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \overline{J}_G^{-1}),$$

quando $n \to \infty$, sendo que

$$\overline{J}_G = \lim_{n\to\infty} \frac{J_{nG}}{n},$$

sendo J_{nG} a matriz de informação de Godambe de β associada a $\Psi_n^G(\beta)$ e dada por

$$J_{nG} = \left\{\sum_{i=1}^{n} S_i\right\} \left\{\sum_{i=1}^{n} V_i\right\}^{-1} \left\{\sum_{i=1}^{n} S_i\right\},$$

sendo que $S_i = -{D_i}^\top \Omega_i^{-1} D_i$ e $V_i = D_i^\top \Omega_i^{-1} Cov(u_i) \Omega_i^{-1} D_i$.

A prova do Teorema (3.1) está em Liang e Zeger (1986). Note que o teorema não exige que $R(\alpha)$ seja a verdadeira matriz de correlação de y_i. Quando a estrutura de correlação definida pela matriz de correlação de trabalho coincide com a verdadeira estrutura, os estimadores de β terão um aumento de eficiência (Liang et al., 1992).

Um estimador consistente para J_{nG}^{-1} é dado por

$$\hat{J}_{nG}^{-1} = \left\{\sum_{i=1}^{n} \hat{S}_i\right\}^{-1} \left\{\sum_{i=1}^{n} \hat{D}_i^\top \hat{\Omega}_i^{-1} \hat{u}_i \hat{u}_i^\top \hat{\Omega}_i^{-1} \hat{D}_i\right\} \left\{\sum_{i=1}^{n} \hat{S}_i\right\}^{-1}, \tag{3.3}$$

sendo que todas as quantidades são avaliadas no ponto $\hat{\beta}$. Ele é conhecido como estimador sanduíche.

Estimação dos parâmetros

Os parâmetros α e ϕ são estimados pelo método dos momentos, através do resíduo de Pearson que é definido como

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\{v(\mu_{ij})\}^{1/2}}.$$

Note que $E(r_{ij}) = 0$ e $Var(r_{ij}) = \phi^{-1}$. O estimador do resíduo de Pearson para a observação y_{ij} é dado por

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\{v(\hat{\mu}_{ij})\}^{1/2}}.$$

Assim, se o quarto momento de y_{ij} for finito, pode-se provar que

$$\hat{\phi}^{-1} = \sum_{i=1}^{n} \sum_{j=1}^{t} \hat{r}_{ij}^2/(nt - p)$$

é um estimador $\sqrt{n}$-consistente de ϕ^{-1} dado β. Observe que $\hat{\phi}^{-1}$ é um estimador da variância de r_{ij} que é igual a ϕ^{-1}.

Em Liang e Zeger (1986) são citadas, além do caso de independência, quatro possíveis estruturas para $R(\alpha)$: m-dependente, autoregressiva de ordem 1, uniforme e não estruturada. As duas primeiras só são aplicáveis em estudos longitudinais, já que, para essas estruturas, a correlação entre a observação j e k da unidade amostral i é função da distância entre os instantes j e k. Nesta dissertação, os dados não são longitudinais. Em virtude disso, em princípio, apenas as estruturas uniforme e não estruturada podem ser utilizadas. Por isso, somente para elas serão apresentados estimadores $\sqrt{n}$-consistentes para α.

No caso não estruturado, o elemento (i, j) da matriz $R(\alpha)$ é dado por $R_{ij} = 1$ se $i = j$ e $R_{ij} = \alpha_{ij}$ se $i \neq j$, $R_{ij} = R_{ji}$. Assim, um estimador $\sqrt{n}$-consistente para

$R(\alpha)$ pode ser obtido através de

$$\hat{R}(\alpha) = \frac{\sum_{i=1}^{n} r_i r_i^\top}{n-p},$$

na qual $r_i = (r_{i1}, \cdots, r_{it})^\top$. O número de parâmetros a ser estimado nesse caso é igual a $0,5t(t-1)$ que pode ser muito grande para valores altos de t.

Na estrutura uniforme, supõe-se que a correlação é constante entre todas as observações de uma mesma unidade amostral. Assim, $R_{ij} = \alpha$ se $i \neq j$ e $R_{ij} = 1$ se $i = j$. Nesse caso um estimador $\sqrt{n}$-consistente para α é dado por

$$\hat{\alpha} = \frac{2\sum_{i=1}^{n}\sum_{j>k} r_{ij} r_{ik}}{nt(t-1)}.$$

Processo iterativo

A estimação de β a partir de (3.2) é efetuada através de uma alteração do método scoring de Fisher. No passo $(m+1)$ a estimativa $\hat{\beta}_n$ para β é dada por

$$\begin{aligned}
\hat{\beta}_n^{(m+1)} &= \hat{\beta}_n^{(m)} - \left\{ E_\beta \left[\frac{\partial}{\partial \beta^\top} \Psi_n^G(\hat{\beta}_n^{(m)}) \right] \right\}^{-1} \Psi_n^G(\hat{\beta}_n^{(m)}) = \\
&= \hat{\beta}_n^{(m)} + \left\{ \left[\sum_{i=1}^{n} \hat{D}_i^\top \hat{\Omega}_i^{-1} \hat{D}_i \right]^{-1} \left[\sum_{i=1}^{n} \hat{D}_i^\top \hat{\Omega}_i^{-1} (y_i - \hat{\mu}_i) \right] \right\}^{(m)} . \qquad (3.4)
\end{aligned}$$

Para a obtenção das estimativas das matrizes utilizadas em (3.4), utiliza-se os valores estimados de β, α e ϕ obtidos no passo m.

Testes de hipóteses

Testes de hipóteses referentes a um ou mais parâmetros do vetor β podem ser feitos utilizando o resultado do Teorema (3.1). Suponha que se deseja testar hipóteses do tipo

$$H_0 : C\beta = 0 \qquad \text{versus} \qquad H_1 : C\beta \neq 0,$$

sendo que C é uma matriz de posto completo $q \leq p$. A estatística do teste é dada por

$$Q = \hat{\beta}_n^\top C^\top \left\{ C\widehat{Cov_A(\hat{\beta}_n)}C^\top \right\}^{-1} C\hat{\beta}_n,$$

sendo que $\widehat{Cov_A(\hat{\beta}_n)}$ é o estimador da matriz de covariâncias assintóticas de $\hat{\beta}_n$. Pode-se utilizar o estimador sanduíche dado em (3.3). A partir do teorema (3.1), pode-se mostrar que sob H_0, $Q \xrightarrow{\mathcal{D}} \chi^2_q$ quando $n \to \infty$.

Análise de diagnóstico

Análises de diagnóstico em GEE são feitas a partir de generalizações de técnicas usualmente utilizadas em modelos lineares generalizados. Venezuela (2003) e Hardin e Hilbe (2003) descrevem algumas dessas técnicas.

3.3 Estratégias de desenvolvimento

Nesta seção serão descritas as estratégias de desenvolvimento de um modelo de *customer scoring*, assim como os modelos associados a elas. Neste momento, será apresentada a formulação geral desses modelos para o ajuste em uma instituição financeira qualquer. No Capítulo 4, serão detalhados os procedimentos utilizados no ajuste dos modelos para o banco de dados disponível.

Com o objetivo de facilitar a compreensão dos modelos associados a cada uma das estratégias será utilizado o Exemplo 1. Nas seções 3.3.1 a 3.3.3 serão apresentados os modelos para esse exemplo. Na Seção 3.3.4 será feita a generalização.

Exemplo 1. Suponha que os produtos de uma determinada instituição possam ser divididos em duas famílias. Admita ainda que cada cliente possua zero ou uma conta em cada uma das famílias de produtos em um instante de origem t. No caso de uma delas, por exemplo, ser a família do cartão de crédito, isso significa que cada cliente possui no máximo 1 cartão. Suponha também que pode-se observar no período entre $t-\epsilon$ e $t, \epsilon > 0$, apenas 3 variáveis para cada um dos n indivíduos com crédito. Uma delas está relacionada ao comportamento de uso da conta da Família 1, outra está associada ao comportamento de uso da conta da Família 2 e uma terceira que representa uma característica do cliente e que não está relacionada a nenhuma conta. Esta última é denominada variável de cliente. Define-se então x_{i11} como o valor da variável associada à Família 1 para o indivíduo i, x_{i21} como o valor da variável associada à Família 2 para o indivíduo i e x_{ic1} como o valor da variável de cliente para o indivíduo i. A partir delas, pode-se obter $x_{11} = (x_{111}, x_{211}, \ldots, x_{n11})^\top$,

$x_{21} = (x_{121}, x_{221}, \ldots, x_{n21})^\top$ e $x_{c1} = (x_{1c1}, x_{2c1}, \ldots, x_{nc1})^\top$. Caso o indivíduo i não possua conta na Família 1, x_{i11} não é observável. Nesse caso, para possibilitar o uso de um artifício algébrico nas estratégias 2 e 3, x_{i11} será codificado com o valor -1. De forma equivalente, x_{i21} terá o valor -1 se o indivíduo i não possuir conta na Família 2.

Cada uma das contas de cada cliente é classificada em uma dentre as seguintes categorias: mau, bom, indeterminado e cancelado. A classificação é feita de acordo com o comportamento de crédito da conta entre os instantes $t+1$ e $t+\delta, \delta > 1$ e está relacionada principalmente ao comportamento de atraso de pagamento observado durante o período. Essa variável é denominada resposta conta. A partir da resposta conta, pode-se obter a resposta cliente. Ela é determinada a partir da pior situação do indivíduo em todas as contas que ele possui. São consideradas não apenas as contas existentes no instante t como aquelas contratadas no período entre $t+1$ e $t+\delta$. A resposta cliente recebe o valor cancelada, se todas as contas do indivíduo foram canceladas no período $t+1$ e $t+\delta$. Assim, no Exemplo 1, para cada cliente i, pode-se definir y_{i1}, como a resposta conta do indivíduo i na Família de produtos 1, y_{i2}, como a resposta conta do indivíduo i na Família de produtos 2 e y_{ic} como a resposta cliente do indivíduo i. Elas são codificadas como

$$y_{im}, m = 1, 2, c = \begin{cases} 0 \text{ se a resposta é mau} \\ 1 \text{ se a resposta é bom} \\ 2 \text{ se a resposta é indeterminado} \\ 3 \text{ se a resposta é cancelado.} \end{cases}$$

A partir das respostas para cada um dos indivíduos, pode-se definir $y_1 = (y_{11}, y_{21}, \ldots, y_{n1})^\top$, $y_2 = (y_{12}, y_{22}, \ldots, y_{n2})^\top$, $y_c = (y_{1c}, y_{2c}, \ldots, y_{nc})^\top$ e $Y = ({y_1}^\top, {y_2}^\top, {y_c}^\top)^\top$ Caso o indivíduo i não possua conta na Família m, y_{im} não é observável. Nesse caso, y_{im} também será codificado como -1.

O modelo de customer scoring tem como objetivo mensurar o risco de um cliente que é bom em um instante de origem t se tornar mau no período entre $t+1$ e $t+\delta$. Dessa forma, no seu ajuste são utilizados apenas clientes que são classificados como bons no instante de origem. Essa condição é válida para todas as estratégias e também para os modelos de *behavioural scoring*. Em todos os modelos ajustados de todas as estratégias, também são desprezadas as respostas classificadas como indeterminado ou cancelado no período entre $t+1$ e $t+\delta$.

3.3.1 Estratégia 1

A Estratégia 1 é aquela que geralmente é utilizada no desenvolvimento de modelos de *customer scoring* (Groom e Gill, 1998). Ela possui duas etapas. Inicialmente são ajustados modelos de *behavioural scoring* para cada uma das famílias de produtos e, a partir deles, é obtido o modelo final. Essa estratégia pode ser segmentada em duas outras: 1a e 1b. A única diferença entre elas é na variável resposta de cada um dos modelos de *behavioural scoring* (modelos de produtos). A Estratégia 1a utiliza no ajuste dos modelos de produtos a resposta conta e a 1b utiliza a cliente como variável dependente. Considera-se apenas as situações mau e bom dessas variáveis resposta. A regressão logística é utilizada em ambas as estratégias e os modelos de produto podem ser escritos, para o Exemplo 1, como:

$$\begin{cases} g(\mu_{i1}) = g(E(y_{i1}/x_{i11})) = \beta_{10} + x_{i11}\beta_{11} \\ g(\mu_{i2}) = g(E(y_{i2}/x_{i22})) = \beta_{20} + x_{i21}\beta_{21} \end{cases} \quad \text{para a Estratégia 1a e}$$

$$\begin{cases} g(\mu_{ic}) = g(E(y_{ic}/x_{i11})) = \beta_{10} + x_{i11}\beta_{11} \\ g(\mu_{ic}) = g(E(y_{ic}/x_{i21})) = \beta_{20} + x_{i21}\beta_{21} \end{cases} \quad \text{para a Estratégia 1b}$$

no qual
β_{10} e β_{20} são parâmetros de intercepto do modelo e
β_{11} e β_{21} são parâmetros associados às variáveis preditoras e interpretados conforme explicado na Seção 3.1.
Os clientes que não possuem conta em uma das famílias são retirados no momento da estimação do modelo de *behavioural scoring* associada a ela.

Em ambas as estratégias, os valores ajustados para cada uma das famílias de produtos (em geral multiplicados por 100 ou por 1000) são denominados escores de produto. Dessa forma, pode-se definir E_{i1} como o escore de produto do cliente i na Família 1 e $E_1 = (E_{11}, E_{21}, \ldots, E_{n1})^\top$. Pode-se classificar então E_1 em e_1 categorias (classes de escore). Para isso, costuma-se dividir E_1 em inúmeras faixas de mesma amplitude. Obtém-se então a proporção de clientes maus (bad rate) observada em cada uma delas e agrupam-se as faixas nas quais ela é semelhante. Esse agrupamento pode ser feito de forma subjetiva para, por exemplo, garantir que a proporção de clientes em cada um dos grupos seja próxima de uma distribuição

pré-definida[2]. Na aplicação desta dissertação foi utilizado o método CHAID, que será apresentado no Capítulo 4. Em geral, e_1 varia entre 5 e 15. Para tratar os indivíduos que não possuem conta na Família 1, cria-se uma categoria adicional. Pode-se assim definir $E_{i1}^*, i = 1, 2, \ldots, n$, como o resultado da categorização de E_{i1} e com valores variando entre 1 e $e_1 + 1$ e $E_1^* = (E_{11}^*, E_{21}^*, \ldots, E_{n1}^*)^\top$. Variáveis indicadoras relacionadas a E_1^* são criadas para possibilitar a inclusão dos escores de produto da Família 1 no modelo final. Elas serão denotadas pelos vetores de n posições $d_{1l} = (d_{11l}, d_{21l}, \ldots, d_{n1l})^\top, l = 2, 3, \ldots, e_2 + 1$ nos quais d_{i1l} é definida como

$$d_{i1l} = \begin{cases} 1 \text{ se } E_{i1}^* = l \\ 0 \text{ caso contrário.} \end{cases}$$

O índice l se inicia em 2 em virtude de um dos grupos ser tomado como referência e finaliza em $e_1 + 1$ para acomodar os clientes que não possuem conta naquela família. De forma análoga, são criadas E_2, E_2^* e $d_{2l}, l = 2, 3, \ldots, e_2 + 1$ para possibilitar a inclusão do escore de produto da Família 2 no modelo final.

O modelo de *customer scoring* utiliza como preditoras, além das variáveis d_{1l} e d_{2l}, x_{c1}, que é a variável de cliente que não foi utilizada nos modelos de *behavioural scoring*. Ele pode ser escrito como

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, D_{i1}, D_{i2})) = \beta_0 + x_{ic1}\beta_c + D_{i1}^\top\beta_1 + D_{i2}^\top\beta_2$$

no qual
$D_{i1} = (d_{i12}, \ldots, d_{i1e_1+1})^\top$ e $D_{i2} = (d_{i22}, \ldots, d_{i2e_2+1})^\top$,
β_c é o parâmetro associado à variável de cliente,
$\beta_1 = (\beta_{12}, \ldots, \beta_{1e_1+1})^\top$ e $\beta_2 = (\beta_{22}, \ldots, \beta_{2e_2+1})^\top$ são os vetores de parâmetros associados às variáveis indicadoras dos escores de produto e
β_0 é o intercepto do modelo.

3.3.2 Estratégia 2

O ajuste de um modelo de *customer scoring* sem a etapa intermediária de desenvolvimento de vários modelos de *behavioural scoring* é outra estratégia possível para

[2]Pode-se, por exemplo, definir que $e_1 = 5$ e que E_1 será dividido de forma que o primeiro grupo tenha os 5% clientes de pior escore e o segundo, o terceiro, o quarto e o quinto tenham, respectivamente, 15%, 20%, 25% e 35% dos clientes.

a obtenção de um modelo de cliente. Para isso, divide-se a população de clientes da instituição em grupos, de acordo com os produtos que cada um possui. No Exemplo 1, haveriam 3 grupos: o primeiro formado pelos clientes que só têm conta na Família 1, o segundo com os indivíduos que têm conta apenas na Família 2 e o último contendo aqueles que têm contas em ambas as famílias. Ajusta-se então um modelo de *customer scoring* para cada um dos grupos criados, utilizando-se apenas as variáveis preditoras disponíveis em cada um deles. No Grupo 1, por exemplo, não é utilizada x_{i21} porque ela não é observável para nenhum dos indivíduos desse grupo. Assim, o modelo de regressão logística de cada um dos grupos é dado por

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, x_{i11})) = \beta_{01} + x_{ic1}\beta_c + x_{i11}\beta_1$$

para o cliente i do Grupo 1,

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, x_{i21})) = \beta_{02} + x_{ic1}\beta_c + x_{i21}\beta_2$$

para o cliente i do Grupo 2 e

$$g(\mu_{ic}) = g(E(y_{ic}/x_{ic1}, x_{i11}, x_{i21})) = \beta_{03} + x_{ic1}\beta_c + x_{i11}\beta_1 + x_{i21}\beta_2$$

para o cliente i do Grupo 3, sendo que
β_c é o parâmetro associado à variável de cliente,
β_1 e β_2 são os parâmetros relacionadas às variáveis de produto e
$\beta_{01}, \beta_{02}, \beta_{03}$ são os interceptos dos modelos.

Todos esses modelos podem ser estimados simultaneamente através da criação de variáveis que indiquem se o cliente tem ou não conta em determinada família. Dessa forma, define-se:

$$w_{i1} = \begin{cases} 1 \text{ se o cliente } i \text{ possui conta na Família 1} \\ 0 \text{ caso contrário,} \end{cases}$$

$$w_{i2} = \begin{cases} 1 \text{ se o cliente } i \text{ possui conta na Família 2} \\ 0 \text{ caso contrário,} \end{cases}$$

$w_1 = (w_{11}, w_{21}, \ldots, w_{n1})^\top$ e $w_2 = (w_{12}, w_{22}, \ldots, w_{n2})^\top$. Observe que se o indivíduo i não possuir conta, por exemplo, na Família 1, então $x_{i11}w_{i1} = 0$. Assim os modelos podem ser ajustados conjuntamente através da equação

$$g(\mu_{ic}) = \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + x_{i11}w_{i1}\beta_1 + x_{i21}w_{i2}\beta_2 + x_{ic1}\beta_c$$

na qual
$\mu_{ic} = E(y_{ic}/x_{i11}, x_{i21}, x_{ic1}, w_{i1}, w_{i2})$
α_1 e α_2 são os parâmetros associados, respectivamente, a w_{i1} e w_{i2} e
β_0 é o intercepto do modelo.
Observe que esse modelo é equivalente aos apresentados para cada um dos grupos. Para verificar a igualdade entre eles, é necessário apenas obter w_{i1} e w_{i2}, de acordo com as famílias que o cliente possui conta e considerar $\beta_{01} = \beta_0 + \alpha_1$, $\beta_{02} = \beta_0 + \alpha_2$ e $\beta_{03} = \beta_0 + \alpha_1 + \alpha_2$. Pelo fato do modelo apresentar ajustes paralelos de acordo com a família de produtos que o cliente possui conta, ele é semelhante ao de uma análise de covariância (Neter et al., 1996).

A inclusão do efeito principal de w_{i1} e w_{i2} é importante para diferenciar dois grupos de clientes que podem ter comportamentos bastante diferentes. Suponha, por exemplo, dois indivíduos que possuem o mesmo valor de x_{ic1} e x_{i11}. A única diferença entre eles está na Família 2. O primeiro cliente não possui conta nessa família. O outro possui, mas, tem $x_{i21} = 0$. Nesse caso a não inclusão do efeito principal de w_{i2} causa a igualdade entre o valor ajustado desses dois indivíduos que podem ter risco de crédito diferentes entre si.

O efeito de x_{i11}, x_{i21} e x_{ic1} é suposto ser independente de quais as famílias em que o cliente possui conta. Porém, na prática, o efeito de x_{i11} em um indivíduo que possui conta apenas na Família 1 pode ser diferente em um outro que possui conta nas duas famílias. Assim, assuma que:

$$r_{i1} = \begin{cases} 1 \text{ se o cliente } i \text{ possui conta apenas na Família 1} \\ 0 \text{ caso contrário,} \end{cases}$$

$$r_{i2} = \begin{cases} 1 \text{ se o cliente } i \text{ possui conta apenas na Família 2} \\ 0 \text{ caso contrário.} \end{cases}$$

Dessa forma, pode-se definir um modelo alternativo para a Estratégia 2 no qual o efeito das variáveis x varia de acordo com as famílias que o cliente possui conta. Para isso são criadas interações entre as variáveis x e as variáveis r. Ele será denominado Estratégia 2b e pode ser escrito como

$$\begin{aligned} g(\mu_{ic}) &= \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + x_{i11}w_{i1}\beta_{10} + x_{i11}w_{i1}r_{i1}\beta_{11} + x_{i21}w_{i2}\beta_{20} + \\ &\quad + x_{i21}w_{i2}r_{i2}\beta_{22} + x_{ic1}\beta_{c0} + x_{ic1}r_{i1}\beta_{c1} + x_{ic1}r_{i2}\beta_{c2}. \end{aligned}$$

no qual
$\mu_{ic} = E(y_{ic}/x_{i11}, x_{i21}, x_{ic1}, w_{i1}, w_{i2}, r_{i1}, r_{i2})$;
β_{c0}, β_{c1} e β_{c2} são os parâmetros associado à variável de cliente;
$\beta_{10}, \beta_{11}, \beta_{20}$ e β_{22} são os parâmetros relacionadas às variáveis de produto;
α_1 e α_2 são os parâmetros associados, respectivamente, a w_{i1} e w_{i2} e
β_0 é o intercepto do modelo.

No Exemplo 1, está se tratando da situação em que há apenas uma variável preditora por família de produtos. Na prática, há várias variáveis preditoras por família que ainda, em geral, são categorizadas. Dessa forma, associado ao comportamento de uso pelo cliente da conta de cada família existe um elevado número de variáveis indicadoras. Várias delas podem ter o valor 1 para uma proporção não muito grande de clientes. Assim, permitir que uma variável indicadora tenha efeito diferente no ajuste do modelo, de acordo com as famílias de produtos que o cliente possui pode não ser factível. O motivo é que, provavelmente, para várias variáveis indicadoras, haverá grupos nos quais a quantidade de clientes com valor 1 será muito pequeno. Dessa forma, as estimativas dos parâmetros associados a elas serão pouco robustas. Assim, na prática, a utilização da Estratégia 2b pode ter problemas de implementação.

3.3.3 Estratégia 3

A terceira Estratégia sugerida tem similaridades com a segunda. Também são ajustados, simultaneamente, modelos para cada configuração de família de produtos que o cliente possui. A diferença é que, além de um modelo para a resposta cliente, são estimados, simultaneamente, modelos para a resposta conta das famílias de produtos que o cliente possui. Para o Exemplo 1, cada cliente i, teria na Estratégia 3, o vetor resposta $Y_i = (y_{i1}, y_{i2}, y_{ic})^\top$. As duas primeiras posições do vetor são a resposta conta associada, respectivamente, às famílias de produtos 1 e 2, enquanto a última posição é a resposta cliente. Nessa estratégia é introduzida uma estrutura de dependência entre as observações, já que há mais de uma resposta para um mesmo cliente. Dessa forma, as equações de estimação generalizadas (GEE) com ligação logito é uma técnica conveniente para a obtenção das estimativas dos parâmetros do modelo. Como o número de famílias não tende a ser muito grande, sugere-se

a adoção de matriz de correlação de trabalho não estruturada. Porém, nem sempre ela pode ser adotada, já que é possível a não convergência dos estimadores dos parâmetros, quando essa estrutura é escolhida.

Para facilitar a compreensão da notação utilizada na Estratégia 3, será feita a comparação das estruturas dos bancos de dados das estratégias 2 e 3. A Estratégia 2 possui uma estrutura do banco de dados semelhante a apresentada na Tabela 3.1. Nesse exemplo, o cliente 2 não possui conta na Família 2 e o cliente 3 não possui conta na Família 1.

Tabela 3.1: Estrutura do banco de dados Estratégia 2

Cliente	Família	y_c	x_{11}	x_{21}	x_{c1}	w_1	w_2
1	cliente	y_{1c}	x_{111}	x_{121}	x_{1c1}	1	1
2	cliente	y_{2c}	x_{211}	-1	x_{2c1}	1	0
3	cliente	y_{3c}	-1	x_{321}	x_{3c1}	0	1

Na Estratégia 3, y_{ic}, que contém apenas a resposta cliente do indivíduo i, é substituído pelo vetor Y_i, que contém também as respostas conta. Em virtude disso, para o Exemplo 1, o número de linhas do banco de dados é multiplicado por 3. Os valores x_{i11} x_{i21}, x_{ic1}, w_{i1} e w_{i2} não se alteram para cada uma das ocorrências do cliente i. Dessa forma, $v_{11}, v_{21}, v_{c1}, w_1^*$ e w_2^* são simplesmente x_{i11} x_{i21}, x_{ic1}, w_{i1} e w_{i2} repetido 3 vezes, conforme pode ser visto na Tabela 3.2. Ela apresenta a estrutura do banco de dados para a Estratégia 3 e os dados são equivalentes aos apresentados na Tabela 3.1.

Tabela 3.2: Estrutura do banco de dados Estratégia 3

Cliente	Família	Y	v_{11}	v_{21}	v_{c1}	w_1^*	w_2^*	z_1	z_2
1	1	y_{11}	x_{111}	x_{121}	x_{1c1}	1	1	1	0
1	2	y_{12}	x_{111}	x_{121}	x_{1c1}	1	1	0	1
1	cliente	y_{1c}	x_{111}	x_{121}	x_{1c1}	1	1	0	0
2	1	y_{21}	x_{211}	-1	x_{2c1}	1	0	1	0
2	2	-1	x_{211}	-1	x_{2c1}	1	0	0	1
2	cliente	y_{2c}	x_{211}	-1	x_{2c1}	1	0	0	0
3	1	-1	-1	x_{321}	x_{3c1}	0	1	1	0
3	2	y_{32}	-1	x_{321}	x_{3c1}	0	1	0	1
3	cliente	y_{3c}	-1	x_{321}	x_{3c1}	0	1	0	0

A Tabela 3.2 apresenta ainda
$z_1 = (z_{111}, z_{121}, z_{1c1}, z_{211}, z_{221}, z_{2c1}, \ldots, z_{n11}, z_{n21}, z_{nc1})^\top$ e
$z_2 = (z_{112}, z_{122}, z_{1c2}, z_{212}, z_{222}, z_{2c2}, \ldots, z_{n12}, z_{n22}, z_{nc2})^\top$, sendo que z_{il1} e z_{il2} são definidas como

$$z_{il1} = \begin{cases} 1 \text{ se a observação } l \text{ do cliente } i \text{ pertence a Família de produtos 1} \\ 0 \text{ caso contrário,} \end{cases}$$

$$z_{il2} = \begin{cases} 1 \text{ se a observação } l \text{ do cliente } i \text{ pertence a Família de produtos 2} \\ 0 \text{ caso contrário.} \end{cases}$$

Elas são criadas para possibilitar a diferenciação entre os valores ajustados para as respostas conta e cliente do indivíduo i. São criadas ainda interações entre as variáveis preditoras originais e as indicadoras de observações (variáveis z) para permitir que o efeito de cada uma das variáveis x possa ser diferente no ajuste das respostas conta e cliente. Dessa forma, o modelo para a Estratégia 3 pode ser definido como

$$\begin{aligned} g(\mu_{il}) \;=\; & \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + z_{il1}\gamma_1 + z_{il2}\gamma_2 + x_{i11}w_{i1}\beta_{10} + x_{i21}w_{i2}\beta_{20} + \\ & + x_{ic1}\beta_{c0} + x_{i11}w_{i1}z_{il1}\beta_{11} + x_{i21}w_{i2}z_{il1}\beta_{21} + x_{ic1}z_{il1}\beta_{c1} + \\ & + x_{i11}w_{i1}z_{il2}\beta_{12} + x_{i21}w_{i2}z_{il2}\beta_{22} + x_{ic1}z_{il2}\beta_{c2} \end{aligned} \tag{3.5}$$

no qual
$g(\mu_{il}) = g(E(y_{il}/w_{i1}, w_{i2}, z_{il1}, z_{il2}, x_{i11}, x_{i21}, x_{ic1}))$;
α_j e $\gamma_j, j = 1, 2$ são parâmetros associados, respectivamente, a w_{ij} e z_{ilj};
$\beta_{ij}, i = 1, 2, c, j = 0, 1, 2$ são parâmetros associados às demais variáveis preditoras;
β_0 é o intercepto do modelo.

Para o cliente 2 da Tabela 3.2, que possui conta apenas na Família 1, o modelo para a sua única resposta conta será portanto dado por

$$\begin{aligned} g(\mu_{i1}) &= \beta_0 + \alpha_1 + \gamma_1 + x_{i11}\beta_{10} + x_{ic1}\beta_{c0} + x_{i11}\beta_{11} + x_{ic1}\beta_{c1} = \\ &= (\beta_0 + \alpha_1 + \gamma_1) + (\beta_{10} + \beta_{11})x_{i11} + (\beta_{c0} + \beta_{c1})x_{ic1} \end{aligned} \tag{3.6}$$

e o modelo para a resposta cliente pode ser escrito como

$$g(\mu_{ic}) = \beta_0 + \alpha_1 + x_{i11}\beta_{10} + x_{ic1}\beta_{c0} = (\beta_0 + \alpha_1) + \beta_{10}x_{i11} + \beta_{c0}x_{ic1} \tag{3.7}$$

Já para o cliente 1, que possui conta nas duas famílias, o modelo é dado por

$$\begin{aligned} g(\mu_{i1}) &= \beta_0 + \alpha_1 + \alpha_2 + \gamma_1 + x_{i11}\beta_{10} + x_{i21}\beta_{20} + x_{ic1}\beta_{c0} + x_{i11}\beta_{11} + x_{i21}\beta_{21} + x_{ic1}\beta_{c1} \\ &= (\beta_0 + \alpha_1 + \alpha_2 + \gamma_1) + (\beta_{10} + \beta_{11})x_{i11} + (\beta_{20} + \beta_{21})x_{i21} + \\ &\quad +(\beta_{c0} + \beta_{c1})x_{ic1} \end{aligned} \tag{3.8}$$

para a resposta conta da Família 1,

$$\begin{aligned} g(\mu_{i2}) &= \beta_0 + \alpha_1 + \alpha_2 + \gamma_2 + x_{i11}\beta_{10} + x_{i21}\beta_{20} + x_{ic1}\beta_{c0} + x_{i11}\beta_{12} + x_{i21}\beta_{22} + x_{ic1}\beta_{c2} \\ &= (\beta_0 + \alpha_1 + \alpha_2 + \gamma_2) + (\beta_{10} + \beta_{12})x_{i11} + (\beta_{20} + \beta_{22})x_{i21} + \\ &\quad +(\beta_{c0} + \beta_{c2})x_{ic1} \end{aligned} \tag{3.9}$$

para a resposta conta da Família 2 e

$$\begin{aligned} g(\mu_{ic}) &= \beta_0 + \alpha_1 + \alpha_2 + x_{i11}\beta_{10} + x_{i21}\beta_{20} + x_{ic1}\beta_{c0} \\ &= (\beta_0 + \alpha_1 + \alpha_2) + \beta_{10}x_{i11} + \beta_{20}x_{i21} + \beta_{c0}x_{ic1} \end{aligned} \tag{3.10}$$

para a resposta cliente.

Comparando-se as equações (3.8), (3.9) e (3.10), pode-se ver que o efeito de cada uma das variáveis x varia em função da resposta que se está modelando para o cliente i. O coeficiente de x_{i11}, por exemplo, é $\beta_{10} + \beta_{11}, \beta_{10} + \beta_{12}$ e β_{10}, caso se esteja ajustando, respectivamente, as respostas conta da Família 1, conta da Família 2 e cliente. Assim, β_{11} é a variação no efeito de x_{i11} quando substitui-se o ajuste da resposta cliente pelo ajuste da resposta conta da Família 1. Porém, assim como na Estratégia 2, o efeito das variáveis x não se altera de acordo com as famílias de produtos que o cliente possui. Observando-se, por exemplo, as equações (3.7) e

(3.10), pode-se notar que o efeito de x_{i11} é o mesmo no ajuste da resposta cliente de um indivíduo que tem conta apenas na Família 1 e de um outro que tem conta nas duas famílias. O modelo pode ser alterado a exemplo do que foi feito na Estratégia 2 para incluir a interação. Porém, essa alternativa apresenta os mesmos problemas práticos já discutidos na Seção 3.3.2.

No momento da estimação do modelo, todas as observações referentes às famílias que os clientes não possuem conta são excluídas. Para o banco de dados da Tabela 3.2, por exemplo, as linhas 5 e 7 seriam eliminadas. Porém, no ajuste de um modelo de GEE, permite-se que as demais observações dos clientes que não têm contas em todas as famílias sejam utilizadas.

3.3.4 Modelo geral

Admita de agora em diante que os produtos de uma determinada instituição possam ser segmentados em M famílias. Suponha ainda que os n clientes com crédito da instituição possuam zero ou uma conta em cada uma delas em um instante de origem t. Admita também a existência de K_m variáveis relacionadas ao comportamento de uso da conta da Família m no período entre t e $t-\epsilon, \epsilon > 0$. Define-se então x_{imk} como o valor da k-ésima variável da Família m para o indivíduo i. Caso o indivíduo i não possua conta na Família m, x_{imk} recebe o valor -1. Podem-se observar ainda C variáveis de cliente que não estão associadas a nenhuma família de produtos. Seja então x_{ick} o valor da k-ésima variável de cliente para o indivíduo i. A partir das variáveis definidas pode-se obter $X_{im} = (x_{im1}, x_{im2}, \ldots, x_{imK_m})^\top, m = 1, 2, \ldots, M$ e $X_{ic} = (x_{ic1}, x_{ic2}, \ldots, x_{icC})^\top$.

Assim como no Exemplo 1, cada cliente possui uma resposta conta em cada família de produtos e uma resposta cliente. Seja então y_{im} a resposta conta do indivíduo i na Família m e y_{ic} a resposta cliente do indivíduo i. Elas são codificadas como

$$y_{im}, m = 1, 2, \ldots, M, c = \begin{cases} \text{-1 se o indivíduo não possui conta na Família m} \\ \text{0 se a resposta é mau} \\ \text{1 se a resposta é bom} \\ \text{2 se a resposta é indeterminado} \\ \text{3 se a resposta é cancelado.} \end{cases}$$

No momento da estimação dos modelos são excluídas as contas indeterminadas

e canceladas e os clientes que não possuem conta na família em estudo. Assim, em termos práticos, y_{im} só assume os valores zero e um.

Conforme apresentado na Seção 3.3.1, a Estratégia 1 requer o desenvolvimento prévio de modelos de *behavioural scoring*. A partir das variáveis definidas, pode-se obter os modelos de *behavioural scoring* das Estratégias 1a e 1b que são dados por

$$\begin{cases} g(\mu_{i1}) = g(E(y_{i1}/X_{i1})) = \beta_{10} + X_{i1}^{\top}\beta_1 \\ g(\mu_{i2}) = g(E(y_{i2}/X_{i2})) = \beta_{20} + X_{i2}^{\top}\beta_2 \\ \vdots \\ g(\mu_{iM}) = g(E(y_{iM}/X_{iM})) = \beta_{M0} + X_{iM}^{\top}\beta_M \end{cases} \quad \text{para a Estratégia 1a e}$$

$$\begin{cases} g(\mu_{ic}) = g(E(y_{ic}/X_{i1})) = \beta_{10} + X_{i1}^{\top}\beta_1 \\ g(\mu_{ic}) = g(E(y_{ic}/X_{i2})) = \beta_{20} + X_{i2}^{\top}\beta_2 \\ \vdots \\ g(\mu_{ic}) = g(E(y_{ic}/X_{iM})) = \beta_{M0} + X_{iM}^{\top}\beta_M \end{cases} \quad \text{para a Estratégia 1b}$$

sendo que
β_{10},$\beta_{20}, \ldots, \beta_{M0}$ são parâmetros de intercepto do modelo e
$\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jK_j})^{\top}, j = 1, 2, \ldots, M$ são vetores de parâmetros associados às variáveis preditoras.

Para ambas as estratégias, pode-se definir E_{im} como o escore de produto do cliente i na Família m e E_{im}^* como a categorização de E_{im} em $e_m + 1$ classes, sendo que a última classe contém os clientes que não possuem produto na Família m. Os valores assumidos por E_{im}^* variam entre 1 e e_m+1. Para a transformação de E_{im} em E_{im}^*, utiliza-se algum dos métodos citados na Seção 3.3.1. Como E_{im}^* é uma variável categorizada, define-se a partir dela as variáveis indicadoras

$$d_{iml}, l = 2, 3, \ldots, e_m + 1 = \begin{cases} 1 \text{ se } E_{im}^* = l \\ 0 \text{ caso contrário.} \end{cases}$$

O modelo de *customer scoring* para a Estratégia 1 pode então ser escrito como

$$g(\mu_{ic}) = \beta_0 + X_{ic}^{\top}\beta_c + D_{i1}^{\top}\beta_1 + D_{i2}^{\top}\beta_2 + \ldots + D_{iM}^{\top}\beta_M$$

no qual
$\mu_{ic} = E(y_{ic}/X_{ic}, D_{i1}, D_{i2}, \ldots, D_{iM})$;
$D_{ij} = (d_{ij2}, \ldots, d_{ije_j+1})^\top, j = 1, 2, \ldots, M$;
$\beta_c = (\beta_{c1}, \beta_{c2}, \ldots, \beta_{cC})^\top$ é o vetor de parâmetros associado às variáveis de cliente;
$\beta_j = (\beta_{j2}, \ldots, \beta_{je_j+1})^\top, j = 1, 2, \ldots, M$, são os vetores de parâmetros associados às variáveis indicadoras dos escores de produto e
β_0 é o intercepto do modelo.

Na Estratégia 2 são estimados simultaneamente diversos modelos de *customer scoring* conforme descrito na Seção 3.3.2, a partir do Exemplo 1. Para que isso seja possível, é necessário definir

$$w_{im}, m = 1, 2, \ldots, M = \begin{cases} 1 \text{ se o cliente } i \text{ possui conta na Família m} \\ 0 \text{ caso contrário.} \end{cases}$$

Assim, o modelo de *customer scoring* para a Estratégia 2 pode ser escrito como

$$g(\mu_{ic}) = \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + \ldots + w_{iM}\alpha_M + X_{i1}^\top w_{i1}\beta_1 + X_{i2}^\top w_{i2}\beta_2 + \ldots + X_{iM}^\top w_{iM}\beta_M + X_{ic}^\top \beta_c$$

no qual
$\mu_{ic} = E(y_{ic}/X_{i1}, X_{i2}, \ldots, X_{iM}, X_{ic}, w_{i1}, w_{i2}, \ldots, w_{iM})$;
$\beta_c = (\beta_{c1}, \beta_{c2}, \ldots, \beta_{cC})^\top$ é o vetor de parâmetros associados às variáveis de cliente;
$\beta_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jK_j})^\top, j = 1, 2, \ldots, M$, são os vetores de parâmetros relacionadas às variáveis de produto;
$\alpha_1, \alpha_2, \ldots, \alpha_M$ são os parâmetros associados, respectivamente, a $w_{i1}, w_{i2}, \ldots, w_{iM}$ e
β_0 é o intercepto do modelo.

Na Estratégia 3, utilizando-se a GEE, são ajustados simultaneamente modelos para as respostas conta e cliente. Para diferenciar o ajuste de cada uma das respostas de um mesmo cliente, é necessário definir

$$z_{ilm} = \begin{cases} 1 \text{ se a observação } l \text{ do cliente } i \text{ pertence a Família de produtos m} \\ 0 \text{ caso contrário,} \end{cases}$$

Dessa forma, o modelo da Estratégia 3 é dado por

$$\begin{aligned} g(\mu_{il}) \;=\; & \beta_0 + w_{i1}\alpha_1 + w_{i2}\alpha_2 + \ldots + w_{iM}\alpha_M + z_{il1}\gamma_1 + z_{il2}\gamma_2 + \ldots + z_{ilM}\gamma_M + \\ & + X_{i1}^\top w_{i1}\beta_{10} + X_{i2}^\top w_{i2}\beta_{20} + \ldots + X_{iM}^\top w_{iM}\beta_{M0} + X_{ic}^\top \beta_{c0} + \end{aligned}$$

$$+X_{i1}^{\top}w_{i1}z_{il1}\beta_{11} + X_{i2}^{\top}w_{i2}z_{il1}\beta_{21} + \ldots + X_{iM}^{\top}w_{iM}z_{il1}\beta_{M1} + X_{ic}^{\top}z_{il1}\beta_{c1} +$$
$$+X_{i1}^{\top}w_{i1}z_{il2}\beta_{12} + X_{i2}^{\top}w_{i2}z_{il2}\beta_{22} + \ldots + X_{iM}^{\top}w_{iM}z_{il2}\beta_{M2} + X_{ic}^{\top}z_{il2}\beta_{c2} +$$
$$+\ldots + X_{i1}^{\top}w_{i1}z_{ilM}\beta_{1M} + X_{i2}^{\top}w_{i2}z_{ilM}\beta_{2M} + \ldots + X_{iM}^{\top}w_{iM}z_{ilM}\beta_{MM} +$$
$$+X_{ic}^{\top}z_{ilM}\beta_{cM}$$

no qual
$\mu_{il} = E(y_{il}/w_{i1}, w_{i2}, \ldots, w_{iM}, z_{il1}, z_{il2}, \ldots, z_{ilM}, X_{i1}, X_{i2}, \ldots, X_{iM}, X_{ic})$;
α_j e $\gamma_j, j = 1, 2, \ldots, M$ são parâmetros associados, respectivamente, a w_{ij} e z_{ilj};
$\beta_{ij}, i = 1, 2, \ldots, M, c, j = 0, 1, \ldots, M = (\beta_{ij1}, \beta_{ij2}, \ldots, \beta_{ijK_i})^{\top}$ são vetores de parâmetros associados às demais variáveis preditoras;
β_0 é o intercepto do modelo.

Nos modelos definidos nesta seção, foi feita a suposição de que cada cliente tinha zero ou uma conta em cada família de produtos. Porém, é muito comum que vários clientes possuam mais de uma conta em uma ou mais famílias. A inclusão de mais de uma conta de uma mesma família nos modelos apresentados, traz mais uma fonte de dependência entre as observações. Porém, nesse caso, a dependência é difícil de ser tratada, já que cada cliente possui um número aleatório de contas em cada família. A solução é utilizar um procedimento para que cada cliente possua um único valor para a resposta conta e para cada uma das variáveis preditoras da família de produtos, como será visto a seguir.

Em relação às variáveis preditoras, isso pode ser feito pelo menos de duas formas diferentes. A primeira é, para cada variável, consolidar todas as contas de uma mesma família em uma única conta, através de um indicador adequado (soma, média, máximo, mínimo, etc). Suponha, por exemplo, que x_{mk} seja o valor da fatura do cartão de crédito em determinado mês. Nessa situação, é razoável considerar para os clientes com mais de um cartão, x_{imk} como a soma da fatura de todos os cartões. Em determinadas situações, essa alternativa não pode ser adotada. Isso ocorre, por exemplo, quando determinadas variáveis são resultado da razão de duas outras que não estão disponíveis. Suponha que x_{mk} seja a razão entre a fatura do cartão e o limite em determinado mês. Admita que um cliente i possua 2 cartões, um de limite igual a 500 e outro de limite igual a 5.000, com fatura, respectivamente, de 400 e 500. Se utilizarmos a média para consolidar as duas contas em uma única, teríamos um x_{imk} de 0,45 $= ((\frac{400}{500} + \frac{500}{5.000})\frac{1}{2})$, bem diferente do

valor mais razoável de 0,164 = ($\frac{400+500}{500+5.000}$). Uma outra alternativa é sortear uma das contas para caracterizar o cliente na Família de produtos m e utilizar suas variáveis independentes. O banco de dados utilizado no Capítulo 4 possui algumas variáveis que não podem ser consolidadas. Em virtude disso, será utilizado o procedimento de sorteio de uma das contas.

Em relação à variável resposta conta também podem ser utilizados pelo menos dois procedimentos. O primeiro é considerá-la como a situação da pior conta daquela família, de acordo com a prioridade apresentada na Seção 2.2. Caso as variáveis preditoras tenham sido escolhidas a partir do sorteio de uma das contas, pode ser mais conveniente adotar a resposta da conta escolhida. Nesse caso, tanto as variáveis preditoras quanto a resposta conta são obtidas a partir da conta sorteada. Para a aplicação do Capítulo 4, essa opção foi adotada.

3.3.5 Comparação das estratégias

Nesta seção serão apresentadas as vantagens e desvantagens de cada uma das estratégias. Além das características que serão discutidas, é importante numa situação real, comparar a performance das estratégias. Algumas medidas utilizadas para esse propósito serão apresentadas na Seção 3.4.

As Estratégias 1a e 1b possuem pelo menos duas vantagens sobre as demais. A primeira é a simplicidade. Para o desenvolvimento de um modelo de *customer scoring* utilizando essa estratégia, usa-se exatamente a mesma metodologia de ajuste de um modelo de *behavioural scoring*. A outra é a possibilidade de aproveitamento dos modelos já existentes. Caso a instituição possua diversos modelos de *behavioural scoring*, ela pode aproveitar esses modelos no ajuste do modelo de *customer scoring*, diminuindo de forma considerável o tempo de desenvolvimento. A Estratégia 1a, assim como a Estratégia 3, ainda possui a vantagem de produzir uma estimativa da probabilidade de um cliente se tornar mau em determinado produto, que pode ser de interesse da instituição. Porém, essa estimativa não considera todas as variáveis disponíveis como na Estratégia 3. Ela utiliza apenas as variáveis relacionadas à própria família de produtos que está sendo ajustada. A principal desvantagem das Estratégias 1a e 1b é o fato delas não considerarem a dependência existente entre as informações de um mesmo indivíduo em famílias de produtos diferentes. Nessas

estratégias essa dependência é desconsiderada, em virtude do desenvolvimento de forma independente de um modelo para cada família. A Estratégia 1a possui pelo menos mais uma desvantagem. Os parâmetros de variáveis associadas às famílias de produtos são estimados no ajuste da resposta conta. Assim, as estimativas obtidas podem não ser as melhores no propósito de se prever a resposta cliente.

A Estratégia 2 tem pelo menos duas vantagens sobre a Estratégia 1. Uma das vantagens é permitir a obtenção de uma medida de risco para o cliente sem a necessidade do ajuste preliminar de um modelo de *behavioural scoring* para cada um dos produtos. Para aquelas instituições que não possuem modelos para cada uma das famílias de produtos, a utilização dessa estratégia pode poupar um grande período de tempo de desenvolvimento. Outra vantagem está no fato dos parâmetros associados a todas as famílias de produtos serem estimados conjuntamente. Suponha, por exemplo, que duas variáveis de famílias de produtos diferentes tenham uma correlação muito alta. Em virtude disso, o mais adequado é selecionar apenas uma delas para o modelo final. Na Estratégia 2, isso geralmente é feito porque os parâmetros associados a essas duas variáveis são estimados conjuntamente. Porém, na Estratégia 1, as duas variáveis são estimadas de forma independente, dificultando qualquer tipo de tratamento de alta correlação entre variáveis de famílias de produtos diferentes. A desvantagem da Estratégia 2 é a não obtenção de estimativas da probabilidade de um cliente se tornar mau em cada uma das famílias de produtos.

A vantagem da Estratégia 3 em relação à 2 está na obtenção do risco associado a cada um dos produtos, já que ela utiliza uma resposta vetorial. Já em relação às estratégias 1a e 1b há pelo menos duas vantagens. A primeira é que, assim como na Estratégia 2, não é necessário o desenvolvimento prévio de vários modelos de *behavioural scoring*. Além disso, a introdução de uma resposta vetorial e o uso de uma técnica estatística adequada para seu tratamento permitem o controle da dependência existente entre o comportamento dos clientes no uso de cada um dos produtos da instituição. Uma desvantagem da Estratégia 3 é a exclusão de um número maior de observações. Isso ocorre porque todos os clientes que possuem pelo menos uma resposta conta indeterminada ou cancelada são excluídos. Outra desvantagem são as limitações dos algoritmos computacionais utilizados na estimação dos modelos de GEE. Eles não toleram uma grande quantidade de variáveis independentes. Além disso, alguns pacotes estatísticos bastante utilizados ainda não permitem sequer o

ajuste dessa técnica.

3.4 Medidas de performance

Os modelos de credit scoring têm como objetivo principal discriminar os indivíduos que se tornarão maus clientes dos indivíduos que se manterão bons. Existem várias medidas que permitem mensurar e comparar o desempenho de modelos na realização desse propósito. Nesta seção serão apresentadas três das medidas de performance mais populares e que serão utilizadas na comparação dos modelos ajustados nessa dissertação: coeficiente de Gini, estatística de Kolmogorov-Smirnov e distância de Mahalanobis (Thomas et al., 2002 e Oliveira e Andrade, 2002).

3.4.1 Coeficiente de Gini

O coeficiente de Gini é determinado a partir da construção da curva ROC (Thomas et al., 2002). Esta é utilizada não somente em credit scoring como também em diversas outras áreas que trabalham com dados binários. Ela é baseada nos conceitos de sensitividade e especificidade, que são estatísticas que podem ser determinadas a partir da construção de diversas matrizes de confusão (Johnson e Wichern, 1998) e serão explicadas a seguir.

Seja b o número de bons clientes de uma determinada população, m o número de maus e $n = b + m$. A partir de um modelo qualquer, pode-se determinar para cada indivíduo i, um escore s_i. Suponha que um indivíduo seja classificado como bom cliente se $s_i > P_c$ e como mau se $s_i \leq P_c$, sendo que P_c é um número real denominado ponto de corte. Se um bom cliente for classificado como bom ou um mau cliente for classificado como mau, pode-se dizer que ele foi classificado corretamente. Fixando-se P_c, pode-se então construir a matriz de confusão (Tabela 3.3)

Tabela 3.3: Matriz de confusão

Observado	Previsto		Total
	Mau	Bom	
Mau	n_{mm}	n_{mb}	m
Bom	n_{bm}	n_{bb}	b
Total	$n_{.m}$	$n_{.b}$	n

na qual
n_{mm} é o número de clientes maus classificados corretamente como maus;
n_{mb} é o número de clientes maus classificados incorretamente como bons;
n_{bm} é o número de clientes bons classificados incorretamente como maus e
n_{bb} é o número de clientes bons classificados corretamente como bons.
A Estatística $\Pi_{bb} = n_{bb}/b$ e $\Pi_{mm} = n_{mm}/m$ são denominadas, respectivamente, sensitividade e especificidade. Elas podem ser definidas como:

- Sensitividade: corresponde a proporção de clientes bons que são classificados corretamente através de um modelo qualquer por terem um escore superior a P_c.

- Especificidade: corresponde a proporção de clientes maus que são classificados corretamente através de um modelo qualquer por terem um escore menor ou igual a P_c.

Quanto maior a sensitividade e a especificidade melhor o modelo. Evidentemente, ambas as medidas dependem de P_c. À medida que P_c cresce, a sensitividade diminui e a especificidade aumenta. Para a construção da curva ROC, obtém-se a matriz de confusão para v diferentes pontos de corte. Como b geralmente é muito maior que m, o escore costuma ser mais concentrado nos valores altos. Dessa forma, é interessante variar P_c em amplitudes maiores nos baixos valores e em amplitudes menores nos altos. Para cada uma das matrizes de confusão construídas, obtém-se a sensitividade e a especificidade. A curva ROC é formada pela união dos pontos (1 - especifidade, sensitividade) para todos as matrizes de confusão obtidas. A Figura 3.1 traz um exemplo de curva ROC.

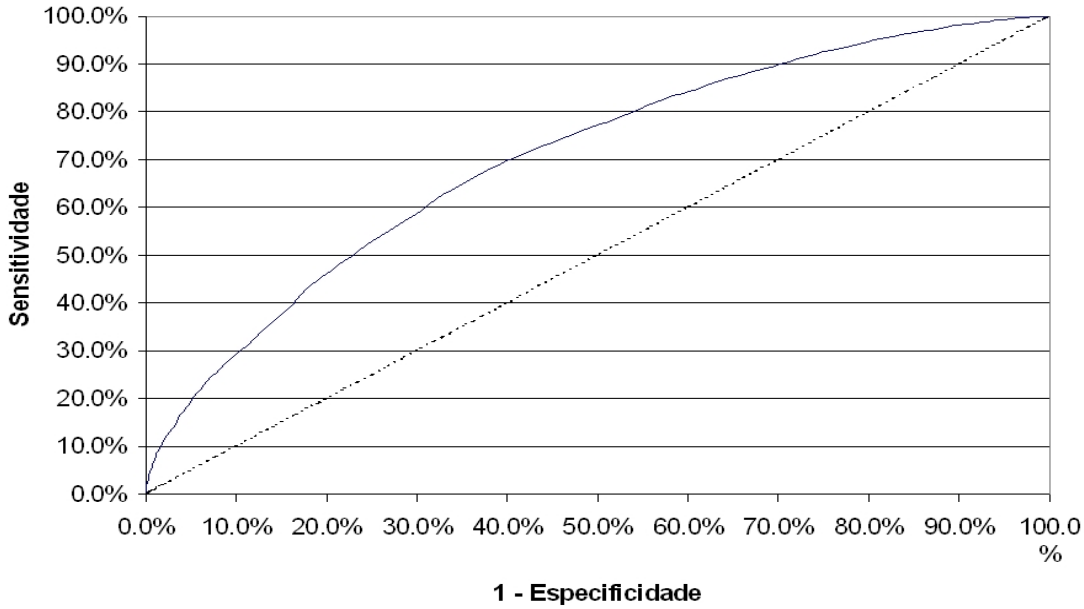

Figura 3.1: Exemplo de curva ROC

Curvas ROC próximas ao eixo da sensitividade e da reta $Y = 1$ apresentam alta sensitividade e especificidade para um grande número de P_c, indicando assim um bom modelo. Se a curva ROC de um modelo A estiver sempre acima da de um outro B, então o modelo A é melhor que o B para todo o intervalo de variação do escore. Essa situação não é muito comum na prática, sendo necessário, a partir da curva ROC, definir uma medida resumo de performance do modelo. Dizer que uma curva está próxima ao eixo da sensitividade e da reta $Y = 1$ é equivalente a dizer que ela está distante da reta $x = 1$ e do eixo 1 - especificidade. Portanto, é natural definir como medida de performance a área da região que se situa entre a curva ROC, a reta $x = 1$ e o eixo 1 - especificidade. Essa medida é conhecida como área sob a curva ROC. Ela é largamente utilizada, mas tem o inconveniente de variar entre 0,5 e 1, já que a curva ROC se situa predominantemente acima da reta $Y = X$. Isso decorre do fato de que a proporção média de clientes classificados corretamente nos grupos dos bons e maus clientes ser maior que 50% para a maioria dos pontos de corte. A explicação dessa afirmação é simples. Suponha que a proporção média de clientes classificados corretamente através de um determinado modelo fosse inferior

a 50% para a maioria dos P_c. Nesse caso, bastaria classificar como maus os clientes classificados anteriormente como bons e vice versa. Assim, essa proporção média se tornaria maior que 50% para a maioria dos pontos de corte. Em virtude da área sob a curva ROC variar entre 0,5 e 1, é mais adequado utilizar o coeficiente de Gini (Thomas et al., 2002), que é dado por duas vezes a área entre a curva ROC e a reta $Y = X$. Dessa forma, tem-se uma medida de performance que varia entre 0 e 1. O modelo de melhor performance, segundo essa medida, é aquele que possui o maior coeficiente de Gini.

3.4.2 Estatística de Kolmogorov-Smirnov

A estatística de Kolmogorov-Smirnov (KS) é usada na teoria estatística não paramétrica para testar se as funções distribuição de uma variável são iguais em dois grupos (Conover, 1999). Em credit scoring, a estatística é utilizada para comparar a distribuição da variável escore, denotada por s, entre os clientes bons e maus.

Em um modelo de bom desempenho, aos bons clientes são atribuídos predominantemente altos escores, enquanto há uma maior concentração de clientes maus entre os baixos escores. Nesse caso, definindo-se $F_b(s)$ como a frequência relativa acumulada do escore entre os clientes bons e $F_m(s)$ a frequência relativa acumulada do escore entre os maus, tem-se que $F_m(s)$ se aproximará rapidamente de 1, enquanto $F_b(s)$ se manterá, para um maior número de valores de s, próximo de 0. Portanto, quanto mais rápido o crescimento de $F_m(s)$ e mais lento o de $F_b(s)$, melhor é o modelo. Em virtude disso, a estatística de Kolmogorov-Smirnov definida como

$$KS = \max_s \{F_m(s) - F_b(s)\}$$

é uma medida de performance de um modelo de credit scoring. Observe que é utilizada a estatística de Kolmogorov-Smirnov para testes unilaterais, já que tem-se interesse exclusivamente nas situações para as quais a função distribuição dos maus clientes é superior a função distribuição dos bons. Assim como o coeficiente de Gini, o KS varia entre 0 e 1 e valores mais altos indicam uma melhor performance. A Figura 3.2 apresenta um exemplo do cálculo do KS.

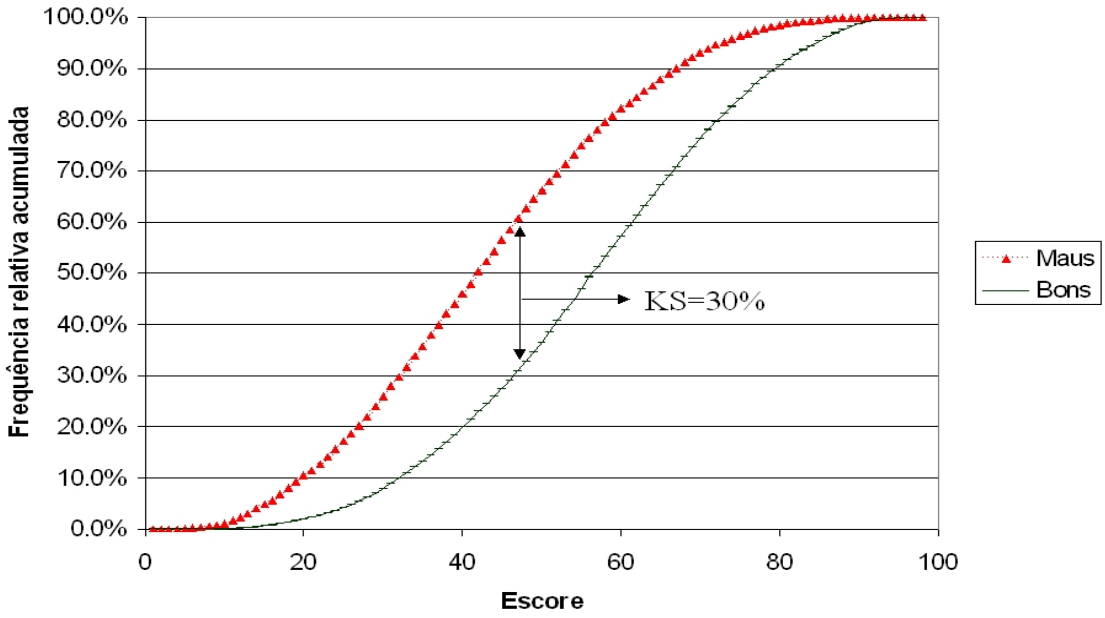

Figura 3.2: Exemplo de cálculo da estatística de Kolmogorov-Smirnov

3.4.3 Distância de Mahalanobis

A distância de Mahalanobis é a mais simples das três. Sabe-se que, quanto mais concentrado os bons estiverem nos altos escores e os maus nos baixos, melhor é o desempenho do modelo. Dessa forma, é natural a comparação do escore médio entre os bons e os maus. O escore, dependendo da técnica utilizada para sua obtenção, pode variar em um intervalo de valores muito diferente. Portanto, na comparação dos escores médios, deve-se levar em consideração a variabilidade dos dados. Definindo-se $\overline{X}_b$ como o escore médio dos bons, $\overline{X}_m$ como o escore médio dos maus, S_b^2 como a variância do escore dos bons e S_m^2 como a variância do escore dos maus, n_b como o número de bons e n_m como o número de maus, a distância de Mahalanobis pode ser definida como

$$DM = \frac{\overline{X}_b - \overline{X}_m}{S_c} \qquad \text{sendo que} \qquad S_c = \sqrt{\frac{n_b S_b^2 + n_m S_m^2}{n_b + n_m}}.$$

Observe que a distância de Mahalanobis é muito semelhante a estatística t, que é utilizada para testar se as médias de uma variável são iguais em duas populações

com mesma variância. Assim como para as demais medidas, quanto maior o valor da distância de Mahalanobis, melhor é a performance do modelo. No entanto, essa medida tem uma desvantagem em relação às anteriores. Ela não tem um intervalo de variação limitado podendo, em princípio, variar entre 0 e infinito.

3.4.4 Comparação das medidas de performance

Os três indicadores de performance fornecem uma medida resumo do desempenho de um modelo. Oliveira e Andrade (2002), através de uma aplicação a dados reais, apresentam indícios de que as três são adequadas para mensurar a performance de um modelo. Dentre as três, o coeficiente de Gini parece ser a medida mais indicada, pois ele resume o desempenho do modelo em toda a amplitude do escore. A estatística de Kolmogorov-Smirnov pode ter um valor alto quando o modelo discrimina eficientemente os bons e os maus em apenas um dos possíveis escores. Isso ocorreria caso as funções distribuições empíricas do escore entre os clientes bons e maus fossem bem semelhantes em todos os pontos com exceção de um. Nesse caso, a estatística de Kolmogorov-Smirmov apresentaria um valor alto, apesar do modelo não ter, provavelmente, boa performance. Já a distância de Mahalanobis pode não ser adequada quando a distribuição do escore apresentar grande assimetria, pois ela compara o escore médio nos grupos dos clientes bons e maus. No entanto, em geral, quando a diferença entre a performance dos modelos que estão sendo comparados é significativa, as três medidas tendem a apresentar resultados equivalentes. Apesar disso, nem sempre o modelo que apresenta maior valor para as três medidas é o mais adequado. Essas medidas fornecem um resumo do desempenho geral do modelo, mas, muitas vezes, uma boa performance é importante em apenas alguns escores fixados (Thomas et al., 2002). Esses pontos são aqueles que dividem a população de clientes em classes, sendo que, em geral, a instituição financeira adota medidas diferentes no gerenciamento do risco de crédito e do relacionamento com cada uma delas.

Capítulo 4

Aplicação

Neste capítulo serão detalhados os procedimentos utilizados no ajuste dos modelos das estratégias descritas na Seção 3.3 aos dados apresentados na Seção 2.2. A performance das estratégias também serão comparadas a partir das medidas descritas na Seção 3.4.

4.1 Categorização das variáveis

A maior parte das variáveis preditoras disponíveis para a estimação dos modelos é quantitativa e algumas delas são qualitativas com um número elevado de categorias. No desenvolvimento de modelos de credit scoring é usual a categorização de todas as variáveis em um número não muito grande de classes em virtude de diversos motivos. O principal deles, para as quantitativas, é que, raramente, a relação entre o logito da esperança da variável resposta e uma preditora qualquer é linear. Dessa forma, para evitar transformação de variáveis que podem tornar o modelo de difícil interpretação, prefere-se efetuar a categorização. Mesmo nos casos em que a relação é linear, ainda assim é comum que ela seja efetuada, para reduzir a influência de valores discrepantes, que algumas vezes podem ser resultado de erro na obtenção do valor da variável. A nova categorização das variáveis qualitativas é feita principalmente por dois motivos. O primeiro é evitar categorias com um número muito pequeno de observações, pois isso pode levar a estimativas pouco robustas dos parâmetros associados a elas. Um segundo motivo é a eliminação de parâmetros desnecessários do modelo. Se duas categorias de uma variável apresentam risco de

crédito equivalente, é razoável agrupá-las em uma única classe.

Toda técnica utilizada para categorizar uma variável necessita de uma categorização inicial. Esta pode ser feita a partir de percentis da variável, mas a experiência do analista pode levar a uma melhor escolha. Suponha, por exemplo, que se deseja dividir uma variável inicialmente em 20 categorias. Caso a escolha seja feita exclusivamente baseada nos percentis, os cortes serão feitos a cada 5%. Porém, não é incomum que observações com valores semelhantes, por exemplo, ao percentil 10 e 25 tenham risco mais próximos do que outras com valores iguais aos percentis 96 e 99. Isso ocorre, freqüentemente, para variáveis com forte assimetria à direita. Nessa situação, é mais adequado construir classes com maior número de clientes para os valores baixos da variável e outras com menos indivíduos no extremo superior. Categorias com um número pequeno de indivíduos devem ser agrupadas com outras semelhantes ou incluídas em uma classe única.

Definida a categorização inicial é importante a realização de uma análise descritiva. Esta pode ser feita a partir da construção, para cada variável, de uma tabela semelhante à Tabela 4.1.

Tabela 4.1: Exemplo de tabela para análise descritiva

Categ	n_{bi}	Π_{bi}	n_{mi}	Π_{mi}	RR_{bmi}	WOE_i
1	n_{b1}	$\Pi_{b1} = \frac{n_{b1}}{b}$	n_{m1}	$\Pi_{m1} = \frac{n_{m1}}{m}$	$RR_{bm1} = \frac{\Pi b1}{\Pi_{m1}}$	$WOE_1 = \ln(RR_{bm1})$
2	n_{b2}	$\Pi_{b2} = \frac{n_{b2}}{b}$	n_{m2}	$\Pi_{m2} = \frac{n_{m2}}{m}$	$RR_{bm2} = \frac{\Pi b2}{\Pi_{m2}}$	$WOE_2 = \ln(RR_{bm2})$
3	n_{b3}	$\Pi_{b3} = \frac{n_{b3}}{b}$	n_{m3}	$\Pi_{m3} = \frac{n_{m3}}{m}$	$RR_{bm3} = \frac{\Pi b3}{\Pi_{m3}}$	$WOE_3 = \ln(RR_{bm3})$
4	n_{b4}	$\Pi_{b4} = \frac{n_{b4}}{b}$	n_{m4}	$\Pi_{m4} = \frac{n_{m4}}{m}$	$RR_{bm4} = \frac{\Pi b4}{\Pi_{m4}}$	$WOE_4 = \ln(RR_{bm4})$
Total	b	100%	m	100%	1	0

na qual

- n_{bi} é o número de clientes bons na categoria i e
- n_{mi} é o número de clientes maus na categoria i.
- RR_{bmi} é o risco relativo de um cliente bom pertencer a categoria i em relação ao risco de um cliente mau ser dessa classe. Categorias com valores de RR_{bmi} maiores que 1 são mais propensas de serem observadas entre os bons clientes do que entre os maus. Portanto, elas possuem um risco de crédito menor do

que a média. O inverso é observado para as classes com RR_{bmi} menores que 1 que possuem um risco de crédito superior à média. Categorias com RR_{bmi} igual a 1 têm risco de crédito semelhante à média. A magnitude de RR_{bmi} também é importante sendo que a medida que essa estatística cresce, o risco de crédito decresce.

- WOE_i (Weights of Evidence, Good, 1950) é obtido a partir do $ln(RR_{bmi})$ e tem a vantagem de ter o valor 0 como ponto de referência. A interpretação é semelhante a de RR_{bmi} com o deslocamento do ponto de igualdade em relação à média de 1 para 0 e o intervalo de variação dos reais positivos para os reais.

Existem inúmeros métodos para categorizar uma variável. Os mais simples e largamente utilizados consistem simplesmente em agrupar as categorias que têm risco de crédito semelhantes, medido através do RR_{bmi} ou do WOE_i. Porém, muitas vezes isso não é fácil sem a utilização de um método formal, pois, freqüentemente, há inúmeras categorizações razoáveis. Para cada uma dessas categorizações, pode-se calcular algum índice de associação entre variáveis categorizadas, como o qui-quadrado ou o valor da informação (Thomas et al., 2002). Escolhe-se então a categorização que apresentar a maior associação entre a variável preditora e a resposta.

O método de categorização utilizado no ajuste dos modelos desta dissertação foi o CHAID (Kass, 1980).

4.1.1 CHAID

O CHAID (Chi-squared Automatic Interaction Detection) é uma técnica estatística utilizada para relacionar uma variável dependente categorizada e uma ou mais variáveis preditoras também categorizadas. Em um modelo de credit scoring, a variável resposta é binária (após a exclusão das observações indeterminadas e canceladas). Além disso, conforme discutido anteriormente, é usual a categorização de todas as variáveis preditoras quantitativas. Assim, o CHAID pode ser utilizado como um modelo de credit scoring. Ele possibilita a divisão dos clientes em grupos que apresentam taxa de maus $(n_{mi}/(n_{bi} + n_{mi}))$ semelhante, a partir das variáveis preditoras disponíveis (Rosa, 2000). A Estratégia 1 poderia utilizar o CHAID em substituição à regressão logística. No entanto, nesta dissertação, a técnica será utilizada exclusivamente para a categorização de variáveis. Em virtude disso, será

descrita em detalhes apenas a parte do algoritmo que cumpre essa tarefa.

O CHAID categoriza variáveis a partir do algoritmo abaixo. Ele supõe que todas as variáveis já possuem uma categorização inicial. Para facilitar a compreensão, cada passo do algoritmo será ilustrado na categorização da variável idade. A distribuição empírica dessa variável é apresentada na Tabela 4.2 (dados fictícios).

Tabela 4.2: Distribuição e WOE_i para a variável idade (em anos)

Idade	Mau		Bom		Total		WOE_i
até 25	14	14,0%	86	9,6%	100	10,0%	-0,38
26 a 35	27	27,0%	173	19,2%	200	20,0%	-0,34
36 a 45	40	40,0%	260	28,9%	300	30,0%	-0,33
46 a 55	14	14,0%	186	20,7%	200	20,0%	0,39
maior que 55	5	5,0%	195	21,7%	200	20,0%	1,47
Total	100	100,0%	900	100,0%	1000	100,0%	0,00

- Passo 1: define-se um nível de significância (α_c). Para o exemplo será utilizado 5%.

- Passo 2: se a variável for nominal, para todos os seus possíveis pares de categoria, constrói-se um teste Qui-quadrado de homogeneidade entre as duas classes em relação à variável resposta. Se ela for ordinal, os testes são obtidos apenas para as categorias adjacentes. A variável idade categorizada é ordinal. Dessa forma, efetua-se um teste de homogeneidade para verificar se as idades até 25 anos e de 26 e 35 são homogêneas em relação à variável resposta. São realizados ainda 3 outros testes para comparar a homogeneidade das demais categorias adjacentes. A Tabela 4.3 apresenta o nível descritivo de cada um deles.

Tabela 4.3: Níveis descritivos do teste de homogeneidade

Idade	Mau	Bom	Total	Idade	Mau	Bom	Total
até 25	14	86	100	26 a 35	27	173	200
26 a 35	27	173	200	36 a 45	40	260	300
Total	41	259	300	Total	67	433	500
nível descritivo = 0,905				nível descritivo = 0,957			
Idade	Mau	Bom	Total	Idade	Mau	Bom	Total
36 a 45	40	260	300	46 a 55	14	186	200
46 a 55	14	186	200	maior que 55	5	195	200
Total	54	446	500	Total	19	381	400
nível descritivo = 0,025				nível descritivo = 0,034			

- Passo 3: identifica-se o par com maior nível descritivo. Se ele for maior que α_c, agrupam-se as duas classes em uma única e repetem-se os passos 2 e 3. No exemplo, agrupam-se as classes 26 a 35 e 36 a 45 com nível descritivo de 0,957, produzindo a Tabela 4.4. Em seguida, são agrupadas as idades até 25 e 26 a 45 com nível descritivo de 0,873 (Tabelas 4.5 e 4.6).

Tabela 4.4: Distribuição e WOE_i para a variável idade (em anos)

Idade	Mau		Bom		Total		WOE_i
até 25	14	14,0%	86	9,6%	100	10,0%	-0,38
26 a 45	67	67,0%	433	48,1%	500	50,0%	-0,33
46 a 55	14	14,0%	186	20,7%	200	20,0%	0,39
maior que 55	5	5,0%	195	21,7%	200	20,0%	1,47
Total	100	100,0%	900	100,0 %	1000	100,0%	0,00

Tabela 4.5: Níveis descritivos do teste de homogeneidade

Idade	Mau	Bom	Total	Idade	Mau	Bom	Total
até 25	14	86	100	26 a 45	67	433	500
26 a 45	67	433	500	46 a 55	14	186	200
Total	81	519	600	Total	81	619	700
nível descritivo = 0,873				nível descritivo = 0,017			
Idade	Mau	Bom	Total				
46 a 55	14	186	200				
maior que 55	5	195	200				
Total	19	381	400				
nível descritivo = 0,034							

Tabela 4.6: Distribuição e WOE_i para a variável idade (em anos)

Idade	Mau		Bom		Total		WOE_i
até 45	81	81,0%	519	57,7%	600	60,0%	-0,34
46 a 55	14	14,0%	186	20,7%	200	20,0%	0,39
maior que 55	5	5,0%	195	21,7%	200	20,0%	1,47
Total	100	100,0%	900	100,0%	1000	100,0%	0,00

- Passo 4: quando um agrupamento resultar em uma classe que contém 3 ou mais categorias iniciais, verifica-se a necessidade de separação de alguma delas das demais. Isso é feito também a partir do teste de homogeneidade, comparando-se cada categoria isolada com as demais agrupadas. Para garantir que o algoritmo tenha fim, não é permitido o retorno para configurações que já ocorreram. Os passos 2 a 4 são repetidos até que nenhuma categoria possa ser agrupada ou separada das demais. No exemplo, o agrupamento das faixas de idades até 25 e 26 a 45 gerou uma classe com 3 categorias iniciais. Assim foi feito o teste se uma delas podia ser separada das demais (Tabela 4.7). Como todos os níveis descritivos são superiores a 5%, as categorias não foram separadas. Na Tabela 4.8, pode-se ver que nenhum agrupamento adicional pode ser feito, já que todos os níveis descritivos são inferiores a 5%. Assim, a categorização final para a variável idade nesse exemplo é: idade até 45 anos, 46 a 55 e maior que 55. Observando-se a Tabela 4.2, pode-se notar que, nesse exemplo,

se um analista agrupasse as classes com WOE_i semelhantes, provavelmente ele obteria ao final do processo as mesmas categorias. Porém, em várias situações é difícil escolher a divisão de classes a ser utilizada baseado exclusivamente na observação dos WOE_i. Assim, a utilização do CHAID na categorização de variáveis é interessante não somente por produzir resultados satisfatórios, como também por proporcionar uma maior objetividade ao processo.

Tabela 4.7: Níveis descritivos do teste de homogeneidade

Idade	Mau	Bom	Total	Idade	Mau	Bom	Total
até 25	14	86	100	até 35	41	259	300
26 a 45	67	433	500	36 a 45	40	260	300
Total	81	519	600	Total	81	519	600
nível descritivo = 0,873				nível descritivo = 0,905			

Tabela 4.8: Níveis descritivos do teste de homogeneidade

Idade	Mau	Bom	Total	Idade	Mau	Bom	Total
até 45	81	519	600	46 a 55	14	186	200
46 a 55	14	186	200	maior que 55	5	195	200
Total	95	705	800	Total	19	381	400
nível descritivo = 0,014				nível descritivo = 0,034			

No algoritmo completo do CHAID, após essa fase de categorização das variáveis, escolhe-se a melhor variável preditora para a resposta (mais significante pelo teste Qui-quadrado de homogeneidade). Em seguida, segmenta-se o banco de dados em relação a essa variável e repete-se o procedimento de identificação da melhor variável preditora para cada grupo formado. O procedimento é repetido até que não haja nenhuma variável preditora significante ou o número de unidades amostrais por grupo seja inferior a um número pré-fixado. Maiores detalhes podem ser vistos em Kass (1980) e Magidson (1994).

Para a aplicação do CHAID aos dados utilizados nos ajustes dos modelos, foi feita uma categorização inicial para cada uma das variáveis. As quantitativas foram divididas em cerca de 20 categorias e classes com poucas observações foram agrupadas nas qualitativas. Para cada uma delas (com exceção das variáveis que não

se referem a nenhum produto) foram feitas duas categorizações, uma baseada na resposta conta e outra na cliente. O motivo disso é que a resposta prevista pelas variáveis relacionadas aos produtos varia em função da estratégia. O nível de significância utilizado foi de 5% e a categorização foi feita no banco de dados completo, excluindo-se as observações com resposta de valor indeterminado ou cancelado. Para as variáveis relacionadas ao comportamento de uso de algum produto, foi selecionada aleatoriamente apenas um conta por cliente, para evitar dependência entre as observações.

4.2 Ajuste dos modelos

O banco de dados disponível para o ajuste dos modelos é o descrito na Seção 2.2. A distribuição da variável resposta em cada uma das famílias de produtos para os 30.000 clientes pode ser observada na Tabela 4.9. A variável dependente referente a cada uma das famílias é a resposta conta.

Tabela 4.9: Distribuição da variável resposta

Variável	Cheque		Cartão		Outros		Cliente	
resposta	#	%	#	%	#	%	#	%
Mau	838	3,0	502	2,5	273	7,2	949	3,2
Bom	24.863	87,7	18.089	91,0	3.459	91,6	26.645	88,8
Indeterminado	176	0,6	209	1,1	44	1,2	564	1,9
Cancelado	2.462	8,7	1.078	5,4	—	—	1.842	6,1
Total com produto	28.339	100,0	19.878	100,0	3.776	100,0	30.000	100,0
Com produto	28.339	94,5	19.878	66,3	3.776	12,6	30.000	100,0
Sem produto	1.661	5,5	10.122	33,7	26.224	87,4	0	0,0
Total	30.000	100,0	30.000	100,0	30.000	100,0	30.000	100,0

Pode-se ver que 3,2% dos clientes se tornaram maus após 6 meses de observação. Porém, o risco varia bastante de acordo com a família de produtos. No cartão de crédito, apenas 2,5% dos clientes se tornaram maus, enquanto essa taxa foi de 7,2% nos outros produtos sem garantia. A proporção de indeterminados e cancelados também varia bastante, mas é importante lembrar que eles são excluídos no momento de estimação dos parâmetros dos modelos. Pode-se ainda observar que o percentual

de clientes com produto em determinada família também tem alta variabilidade.

A Tabela 4.10 apresenta a matriz de correlação entre as variáveis resposta. Para a construção da tabela, desconsiderou-se as observações indeterminadas e canceladas.

Tabela 4.10: Matriz de correlação entre as respostas conta e produto

	Cheque	Cartão	Outros	Cliente
Cheque	1,000	0,899	0,915	0,963
Cartão		1,000	0,943	0,918
Outros			1,000	0,884
Cliente				1,000

Pode-se notar que as correlações entre as respostas são muito altas. Dessa forma, se um modelo ajusta todas elas simultaneamente, é fundamental o uso de uma técnica estatística que trata a dependência entre as observações. Por isso, a utilização da GEE com ligação logito é adequada na Estratégia 3.

O banco de dados foi dividido aleatoriamente em dois grupos: amostra de desenvolvimento contendo 20.000 clientes e amostra de validação com 10.000 indivíduos. Na amostra de desenvolvimento foram ajustados todos os modelos de cada uma das estratégias. Estes foram então aplicados na amostra de validação para a comparação da performance de cada uma das estratégias. O estudo da performance de um modelo deve ser feito, preferencialmente, em uma amostra de indivíduos não utilizada na estimação de seus parâmetros. Isso deve ser feito para medir qual será o real desempenho do modelo após sua implantação, já que os clientes que serão avaliados, em sua maioria, não integraram a amostra de desenvolvimento.

A partir das variáveis preditoras categorizadas conforme descrito na Seção 4.1, foram criadas variáveis indicadoras. As categorias das variáveis foram ordenadas quanto ao WOE_i e foi tomada como referência a classe de menor risco (maior WOE_i). Dessa forma, espera-se que a estimativa dos parâmetros referentes a cada uma das variáveis indicadoras sejam sempre negativas. O banco de dados descrito na Seção 2.2 possui 36 variáveis preditoras originais. Com a categorização das variáveis e a criação das variáveis de interação entre elas para o ajuste do modelo da Estratégia 3, foram obtidas 558 variáveis indicadoras.

Os modelos das Estratégias 1 e 2 foram estimados no software SPSS versão 10.0. A seleção de variáveis incluídas no modelo foi feita a partir do procedimento forward stepwise (Hosmer e Lemeshow, 1989). Resumidamente, o procedimento se inicia através da estimação de um modelo apenas com o intercepto, inclui, uma a uma, as variáveis mais significantes no modelo e exclui aquelas que, na presença das outras, não são mais importantes. O procedimento termina quando nenhuma variável puder ser incluída ou excluída no modelo de acordo com níveis de significância pré-estabelecidos. Os níveis de significância utilizados para a inclusão e exclusão de variáveis foram, respectivamente, de 0,05 e 0,1. O SPSS utiliza o teste escore para escolher a variável a ser incluída. Para a exclusão, é permitida a escolha do teste a ser utilizado. Foi usado o teste de razão de verossimilhanças. As variáveis indicadoras resultantes de uma mesma variável foram tratadas de forma independente. Assim, é possível a inclusão de apenas algumas das variáveis indicadoras de uma determinada variável. No desenvolvimento de um modelo para a utilização em uma instituição financeira, o procedimento stepwise é o primeiro passo para a obtenção do modelo final. Substituições de variáveis e recategorizações são freqüentemente feitas para diminuir a multicolinearidade, tornar o modelo mais interpretável ou melhorar a performance. Para evitar o favorecimento de alguma das estratégias, procurou-se fazer o menor número possível de ajustes nos modelos obtidos nesta dissertação. Foram feitas pequenas alterações apenas para evitar que os coeficientes das variáveis indicadoras de uma mesma variável indicassem uma ordenação do risco de crédito entre as categorias diferente da sugerida pela análise dos WOE_i. Suponha, por exemplo, que uma categoria a possui um WOE_i superior ao de uma categoria b. Assim, espera-se que a estimativa do parâmetro referente à categoria a seja maior do que a estimativa do parâmetro referente à categoria b. Caso isso não ocorra, um possível ajuste que pode ser feito é juntar as duas categorias em uma única e estimar novamente os parâmetros do modelo. As estimativas dos parâmetros, seus respectivos erros padrão e níveis descritivos do modelo final de cada uma das estratégias estão no Apêndice A. Foram feitos ainda testes de Hosmer e Lemeshow para os modelos finais das estratégias 1a, 1b e 2. Os níveis descritivos obtidos foram, respectivamente, 0,387, 0,243 e 0,289, que indicam um bom ajuste dos 3 modelos.

O SAS versão 8.2 foi utilizado para a estimação do modelo da Estratégia 3. Ele

não permite a execução do procedimento stepwise na estimação de um modelo de GEE. Em virtude disso, a alternativa natural seria estimar o modelo com todas as variáveis, retirando-se, uma a uma, as variáveis não significantes. Porém, em virtude de cada uma das variáveis gerar várias variáveis indicadoras e cada uma delas interagir com cada uma das variáveis z_l, o número de parâmetros a ser estimado é muito grande (superior a 500). Além disso, o fato da maioria dos clientes não possuir produtos em todas as famílias torna os dados bastante desbalanceados. Em conseqüência disso, não foi possível ajustar o modelo da Estratégia 3 com todas as variáveis. O SAS retornou mensagem de "erro na rotina de estimação" quando isso foi feito. Tentativas de ajuste do modelo com todas as variáveis também foram feitas no SPlus 4.5 através da biblioteca Gee e no R 1.8.1 através da biblioteca Geepack. Nenhum dos dois pôde ajustar o modelo.

Contornou-se esse problema através do ajuste no SAS de dois modelos. O primeiro não considera a interação entre as variáveis x e as variáveis z, reduzindo assim em cerca de 75% o número de parâmetros a ser estimado. Dessa forma, foi possível ajustar e obter um modelo final, após a retirada uma a uma das variáveis não significantes (nível de significância de 5%). Esse modelo será denotado como 3r. Muitas variáveis preditoras em um modelo de *customer scoring* possuem correlação relativamente alta. Em virtude disso, variáveis importantes do modelo podem se tornar não significantes devido à multicolinearidade. Esse efeito é reduzido quando utiliza-se o procedimento forward stepwise, já que as variáveis mais importantes são incluídas no início do ajuste. Embora o procedimento permita a exclusão de variáveis, a probabilidade da manutenção das principais variáveis indicadoras até a obtenção do modelo final é bem maior. Portanto, o fato de não haver o procedimento forward stepwise para o ajuste da GEE prejudica, consideravelmente, a seleção de variáveis na Estratégia 3.

O segundo modelo ajustado foi construído de forma subjetiva. A partir da análise descritiva e da observação dos ajustes dos modelos das Estratégias 1 e 2, foi feita uma pré-seleção de variáveis, escolhendo-se aquelas que tinham maior associação com a variável resposta. Elas então foram divididas em pequenos grupos de variáveis. Para cada um desses grupos foi possível ajustar o modelo. Assim, obteve-se para cada um deles um modelo final, retirando-se, uma a uma, as variáveis não significantes. Os grupos foram fundidos em outros maiores e o procedimento foi repetido. Isso foi

feito até a obtenção de um único grupo no qual todas as variáveis eram significantes. O nível de significância utilizado também foi de 5%. Durante esse processo, algumas variáveis ainda foram excluídas para evitar erro na rotina de estimação pelo SAS. Esse modelo será denotado como 3s.

Ambos os modelos utilizaram a estrutura uniforme para a matriz de correlação de trabalho. A Tabela 4.10 indica que essa estrutura parece ser adequada. Para efeito de comparação, procurou-se ajustar os modelos também com a utilização da matriz de correlação não estruturada. Porém, obteve-se erro de estimação do SAS, mesmo com um número não muito grande de variáveis preditoras. As estimativas obtidas para o parâmetro de correlação nos Modelos 3r e 3s foi respectivamente de 0,8317 e 0,8535. As estimativas dos demais parâmetros, seus respectivos erros padrão e níveis descritivos desses modelos estão no Apêndice A.

A GEE é uma técnica relativamente nova e apenas nos últimos anos ela foi incluída nos principais softwares de análise estatística. Certamente, os algoritmos de estimação dos parâmetros ainda serão aperfeiçoados, permitindo o ajuste de modelos com um número maior de variáveis. A inclusão do procedimento stepwise também deve ser feita nos próximos anos, possibilitando uma melhor seleção das variáveis preditoras. Dessa forma, no futuro a Estratégia 3 poderá ser utilizada de forma bem mais eficiente do que é possível atualmente.

As estimativas dos parâmetros dos modelos de cada uma das estratégias, bem como seus respectivos erros padrões e níveis descritivos estão no Apêndice A. Por motivo de sigilo, não serão discutidas questões relacionadas às variáveis selecionadas e aquelas com melhor poder preditivo.

4.3 Comparação da performance

A amostra de validação foi utilizada para comparar a performance do modelo de cada uma das estratégias. Para verificar se uma estratégia tinha desempenho superior para todo o intervalo de variação do escore, construiu-se a curva ROC (Figura 4.1).

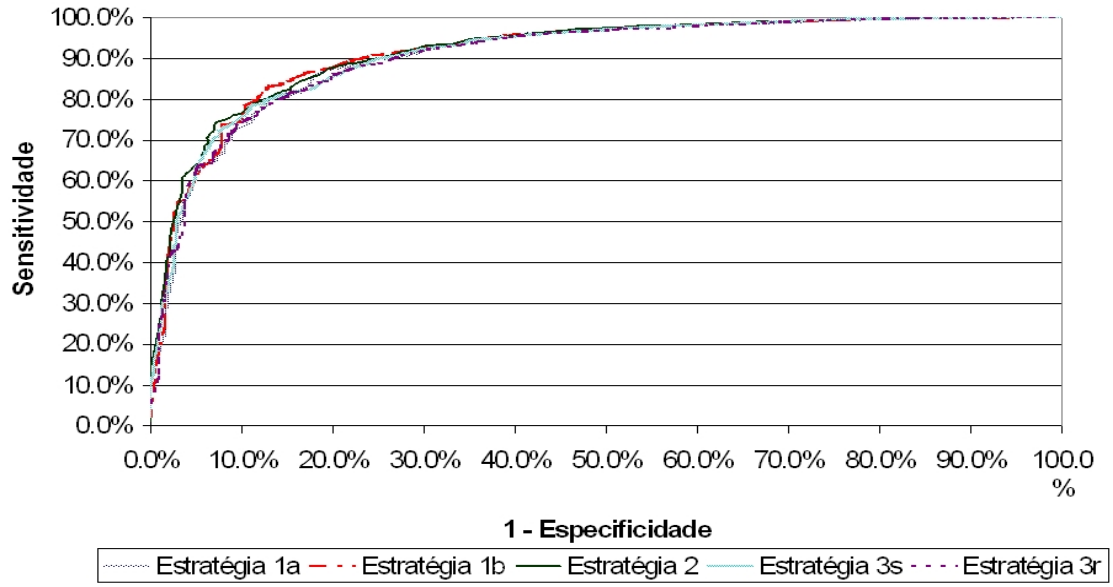

Figura 4.1: Curva ROC para os modelos de customer scoring

Pode-se observar que nenhuma estratégia possui curva ROC com ordenada maior ou igual a das demais para todas as possíveis abscissas. No entanto, pode-se ver que as Estratégias 1b e 2 se destacam. A última possui o melhor desempenho para a maior parte dos pontos de sensibilidade entre 54% e 77%. Isso indica que, se for selecionado, a partir do escore, entre 54% e 77% dos clientes bons, a Estratégia 2 deverá apresentar uma menor quantidade de indivíduos maus que as demais. Já a Estratégia 1b apresenta um performance superior às demais para a maior parte da região de sensibilidade entre 77% e 91%. Como nenhuma estratégia tem desempenho sempre superior às demais, é interessante analisar as medidas descritas na Seção 3.3. A Tabela 4.11 apresenta o resultado de cada uma delas para cada uma das estratégias.

Tabela 4.11: Indicadores de performance das estratégias para a resposta cliente

Medida de performance	Estratégia				
	1a	1b	2	3r	3s
Coeficiente de Gini	0,817	0,830	0,836	0,814	0,823
Estatística de Kolmogorov-Smirnov	0,681	0,704	0,681	0,663	0,672
Distância de Mahalanobis	2,980	2,986	3,086	2,880	3,000

Pode-se ver que nenhuma estratégia apresenta performance superior às demais segundo as três medidas. Além disso, a variação de desempenho entre a estratégia de melhor e pior performance é inferior a 8%. No entanto, conforme já observado na Curva ROC, as Estratégias 1b e 2 se destacam. A última é a estratégia de melhor performance segundo o coeficiente de Gini e a distância de Mahalanobis, enquanto a primeira tem melhor desempenho segundo a estatística de Kolmogorov-Smirnov. Pode-se observar ainda que, mesmo não sendo possível o ajuste do melhor modelo da Estratégia 3 devido a restrições computacionais, o desempenho por ela apresentado não foi muito inferior às demais. O modelo 3s apresentou inclusive distância de Mahalanobis superior a da Estratégia 1b. Isso é um indício de que essa estratégia poderá vir a se tornar uma boa opção, após o aperfeiçoamento dos algoritmos de ajuste da GEE presente nos principais softwares estatísticos. Nota-se também que a Estratégia 1b apresentou desempenho superior a 1a para todas as medidas. Isso sugere que, caso se deseje utilizar a Estratégia 1 e o interesse na obtenção de cada um dos escores de produto seja apenas de utilizá-los como preditora para o modelo principal, é mais interessante utilizar a variação b.

Com o objetivo de verificar se o número de famílias de produtos que o cliente possui interfere na ordenação de performance entre as estratégias, construiu-se a Tabela 4.12.

Tabela 4.12: Indicadores de performance das estratégias por número de famílias para a resposta cliente

Número de famílias	Medida de performance	Estratégia 1a	1b	2	3r	3s
3	Coeficiente de Gini	0,732	0,748	0,767	0,730	0,741
	Kolmogorov-Smirnov	0,613	0,610	0,629	0,621	0,612
	Distância de Mahalanobis	2,017	1,949	2,130	1,932	1,970
2	Coeficiente de Gini	0,809	0,820	0,818	0,802	0,806
	Kolmogorov-Smirnov	0,655	0,698	0,660	0,651	0,653
	Distância de Mahalanobis	2,684	2,744	2,808	2,624	2,767
1	Coeficiente de Gini	0,820	0,821	0,843	0,826	0,831
	Kolmogorov-Smirnov	0,722	0,714	0,727	0,718	0,719
	Distância de Mahalanobis	4,031	4,042	4,067	3,790	4,017

A ordenação de performance entre as estratégias parece não ter associação com o número de famílias de produtos. Nenhuma estratégia altera significativamente sua performance em relação às demais, a medida que o número de famílias decresce. A Estratégia 2 se destaca nos grupos de clientes com uma e três famílias, enquanto a 1b apresenta melhor desempenho entre os indivíduos com duas famílias. É interessante notar que o desempenho absoluto de todas as estratégias melhora à medida que decresce o número de famílias. Embora o grupo de indivíduos com 3 famílias possua um número maior de variáveis para se estimar o risco, isso parece não ser suficiente para compensar um acréscimo na quantidade de produtos diferentes nos quais o indivíduo pode se tornar mau.

A Tabela 4.13 apresenta as medidas de performance para os modelos de produto. Ela mostra os resultados apenas das estratégias 1a e 3, porque apenas estas geram um escore de produto que é a estimativa da probabilidade do cliente se manter bom naquela família.

Tabela 4.13: Indicadores de performance dos modelos de produtos

Família de produtos	Medida de performance	Estratégia 1a	3r	3s
Cheque	Coeficiente de Gini	0,809	0,838	0,850
	Estatística de Kolmogorov-Smirnov	0,676	0,688	0,692
	Distância de Mahalanobis	2,704	3,163	2,981
Cartão	Coeficiente de Gini	0,679	0,869	0,879
	Estatística de Kolmogorov-Smirnov	0,544	0,742	0,751
	Distância de Mahalanobis	2,230	3,424	3,330
Outros	Coeficiente de Gini	0,456	0,758	0,772
	Estatística de Kolmogorov-Smirnov	0,393	0,641	0,646
	Distância de Mahalanobis	0,785	1,913	1,884

Pode-se observar que o desempenho da Estratégia 3 é superior ao da 1a segundo todas as medidas. Isso ocorre porque a Estratégia 3 é desenvolvida de forma que todas as variáveis disponíveis participem do ajuste de cada um dos escores de produto. Na Estratégia 1a, apenas as variáveis relacionadas à própria família de produtos da qual se está estimando o risco são utilizadas. Na família de outros produtos, a diferença de performance entre as Estratégias 1a e 3 é extremamente grande. Isso ocorre porque, nessa família, estão disponíveis apenas uma pequena quantidade de variáveis preditoras com forte associação com a resposta. Em virtude dos resultados observados, há indícios de que, caso se deseje obter uma estimativa da probabilidade do cliente se manter bom em determinada família de produtos, é recomendável a utilização da Estratégia 3, mesmo considerando-se os problemas existentes na estimação dos parâmetros e seleção de variáveis.

Capítulo 5

Simulação

A aplicação apresentada no Capítulo 4 foi baseada em dados reais de uma instituição financeira. Os bancos que atuam no Brasil têm várias características em comum. Um grande número de pessoas, por exemplo, possui conta corrente em mais de uma instituição e os principais produtos de crédito são bem semelhantes. Dessa forma, é razoável esperar que haja alguma semelhança entre a performance relativa das estratégias obtida na instituição financeira estudada e em outros bancos com atuação no Brasil. Porém, a partir dos dados disponíveis, não se pode inferir qual a intensidade dessa semelhança e nem mesmo garantir que ela realmente existe. Com o objetivo de estudar a performance das estratégias em uma situação mais geral foram feitas simulações. Elas permitem ainda uma melhor avaliação da Estratégia 3, já que na aplicação do Capítulo 4, devido aos problemas discutidos anteriormente, não foi possível compará-la com as demais em igualdade de condições.

5.1 Parâmetros da simulação

A simulação foi desenvolvida para situações nas quais são ajustados modelos de *behavioural scoring* para duas famílias de produtos e todos os clientes possuem conta em ambas. Os dados foram gerados a partir do algoritmo abaixo.

- A partir do banco de dados descrito na Seção 2.2, foram sorteados 10.000 clientes que possuíam conta tanto na família do cheque especial como na família do cartão de crédito.

- Para cada uma das famílias foram escolhidas duas variáveis para participar da simulação. Procurou-se selecionar variáveis que possuem alta associação com a variável resposta, para que os modelos ajustados apresentassem boa performance. Além disso, evitou-se a escolha de variáveis com um número muito pequeno de categorias, para que o número de possíveis valores que os escores de produto pudessem assumir fosse relativamente grande. As variáveis selecionadas foram a x_{105}, x_{112}, x_{205} e x_{207}.

- Ajustaram-se então modelos de regressão logística tendo como variável dependente a resposta conta das duas famílias e como preditoras as 16 variáveis indicadoras obtidas a partir de x_{105}, x_{112}, x_{205} e x_{207}. Obtiveram-se assim, estimativas da probabilidade de cada um dos 10.000 indivíduos se manter bom cliente nas famílias de produtos 1 e 2. Define-se então $\hat{p}_{ij}$ como a estimativa da probabilidade do indivíduo i se manter bom cliente na família j.

- Geraram-se em seguida normais bivariadas para cada um dos 10.000 indivíduos, com correlação ρ, vetor de médias 0 e parâmetros de variância 1, a partir do algoritmo descrito em Johnson (1987). O parâmetro ρ foi obtido de forma que a correlação entre as variáveis resposta conta obtida no passo seguinte fosse aproximadamente 0,5 no primeiro grupo de simulações e 0,9 no segundo. Seja $x_i = (x_{i1}, x_{i2})^\top$ o vetor normal obtido para o indivíduo i. A partir de x_i foi obtido u_i através da expressão $u_i = (F_z(x_{i1}), F_z(x_{i2}))^\top$, no qual F_z é a função distribuição da normal univariada com média 0 e variância 1. É fácil perceber que $u_i = (u_{ii}, u_{i2})^\top$ tem distribuição marginal uniforme[0,1].

- Obteve-se então y_{i1} e y_{i2} que são, respectivamente, a resposta conta da família de produtos 1 e 2 para o indivíduo i fazendo

$$y_{ij} = \begin{cases} 1 \text{ se } u_{ij} \leq \hat{p}_{ij} \\ 0 \text{ caso contrário.} \end{cases}$$

- A resposta cliente para o indivíduo i, y_{ic}, foi obtida de duas formas diferentes. Na primeira condição, denominada simulação sem perturbação, $y_{ic} = \min\{y_{i1}, y_{i2}\}$. Na outra, $y_{ic} = \min\{y_{i1}, y_{i2}, A_i\}$, na qual

$$A_i = \begin{cases} 1 \text{ com probabilidade } 0{,}995 \\ 0 \text{ com probabilidade } 0{,}005. \end{cases}$$

A probabilidade de 0,005 para a ocorrência da perturbação é um pouco superior à proporção de casos em que ela foi observada nos dados utilizados na aplicação do Capítulo 4. Obteve-se A_i a partir da geração de uniformes no intervalo [0,1]. A introdução de A_i visa permitir que os indivíduos se tornem maus clientes em contas que não existiam no instante de origem ou em contas de famílias de produtos para as quais não foram desenvolvidos modelos. Essa segunda condição será denominada **simulação com perturbação**.

Foram feitas 2.000 repetições desse algoritmo sendo 500 para cada combinação de parâmetro de correlação e ocorrência ou não de perturbação. Para cada repetição, foram ajustados os modelos de cada uma das estratégias conforme descrito na Seção 3.3 e utilizando como variáveis preditoras as variáveis indicadoras obtidas a partir de x_{105}, x_{112}, x_{205} e x_{207}. A performance dos modelos foram então comparadas através das medidas de performance descritas na Seção 3.4.

5.2 Comparação da performance

Os resultados da simulação estão resumidos nas tabelas do Apêndice B. Cada uma das combinações de parâmetro de correlação e ocorrência ou não de perturbação será chamada de **condição da simulação**. Inicialmente, obteve-se a proporção de repetições em que cada uma das estratégias é superior a todas as demais (Tabela B.1). Percebe-se que em nenhuma condição, uma das estratégias tem desempenho superior às demais para todas as medidas de performance. De um modo geral, pode-se concluir que:

- Baseado no coeficiente de Gini, as estratégias 2 e 3 apresentaram desempenho superior em maior proporção na previsão da resposta cliente. A Estratégia 2 mostrou-se superior às demais quando a correlação entre as resposta conta foi de 0,5 e na condição com correlação de 0,9 e ausência de perturbação. Na condição com correlação 0,9 e presença de perturbação, as estratégias 2 e 3 apresentaram um desempenho muito semelhante e superior às demais.

- Resultados semelhantes foram obtidos ao se utilizar a estatística de Kolmogorov-Smirnov, ou seja, as estratégias 2 e 3 apresentaram-se superiores às demais em

maior proporção. Nas condições com presença de perturbação a Estratégia 2 teve um desempenho muito melhor do que as outras.

- Pela distância de Mahalanobis, a Estratégia 3 não apresentou um bom desempenho. Segundo essa medida, as estratégias 2 e 1b tiveram melhor performance.

Não é difícil imaginar uma situação em que uma estratégia seja a melhor dentre todas em proporção pequena de repetições, mas apresenta-se superior às demais na comparação duas a duas. Assim, além da tabela B.1, é importante também comparar o desempenho de cada uma das estratégias em relação à cada uma das demais. Nas seções seguintes são feitas essas comparações.

5.2.1 Comparação entre as estratégias propostas e as usualmente utilizadas

Nas tabelas B.2 a B.5, obteve-se a proporção de repetições em que cada uma das estratégias é melhor que cada uma das demais. As tabelas apresentam a proporção obtida e um intervalo de confiança para a proporção na população ($\gamma = 95\%$, Magalhães e Lima, 2001). Quando o intervalo contém o valor de 50%, as estratégias i e j são equivalentes em relação à proporção de vezes em que uma é superior a outra. Caso o intervalo esteja acima do valor de 50%, a Estratégia i é superior a Estratégia j. O oposto ocorre se o intervalo estiver abaixo do valor de 50%. Para cada medida de desempenho, as linhas 2 a 5 nas tabelas B.2 a B.5 comparam as estratégias propostas nessa dissertação (2 e 3) com as estratégias geralmente utilizadas (1a e 1b). Pode-se ver que, para todas as condições e considerando-se as medidas de Gini e Kolmogorov-Smirnov, o intervalo de confiança está acima de 50%. Isso indica que, para todas as condições, as estratégias propostas são superiores às geralmente utilizadas em mais da metade das vezes. Ainda considerando-se as medidas de Gini e Kolmogorov-Smirnov, pode-se notar que a ordem de grandeza da proporção de repetições em que uma estratégia é superior à outra não é constante para todas as condições. A proporção de repetições em que as estratégias 2 e 3 são superiores à Estratégia 1a é maior quando a correlação entre as respostas é 0,5 (tabelas B.2 e B.4) do que quando ela vale 0,9 (tabelas B.3 e B.5). Porém, a ocorrência ou não de perturbação parece não afetar as conclusões. Na comparação das estratégias 2

e 3 com a Estratégia 1b ocorre o oposto. A correlação entre as respostas parece não ter relação com a performance relativa dessas estratégias. No entanto, a proporção de repetições em que as estratégias 2 e 3 são superiores à Estratégia 1a é maior quando há perturbação (tabelas B.2 e B.3) do que quando não há (tabelas B.4 e B.5). Os resultados para a distância de Mahalanobis foram diferentes dos obtidos para as demais medidas. Para as condições com correlação 0,5 (tabelas B.2 e B.4), os intervalos de confiança indicam superioridade das estratégias 2 e 3 com relação às estratégias 1a e 1b. No entanto, para as demais condições, em várias comparações (tabelas B.3 e B.5), não há diferença significante entre as estratégias 1a e 1b e as estratégias propostas. Para a correlação 0,9 sem perturbação (Tabela B.5), a Estratégia 1b é até mesmo superior à Estratégia 3.

As tabelas B.6 a B.9 comparam as estratégias duas a duas em relação à média das medidas de performance. Elas trazem a estimativa pontual da diferença média das medidas, assim como um intervalo de confiança para essa média ($\gamma = 95\%$, Bussab e Morettin, 2002). Caso o intervalo contenha o valor 0 para uma determinada medida, não há diferença entre a média da medida nas estratégias i e j. Intervalos acima do valor 0 indicam que a Estratégia i é, em média, superior à Estratégia j. Intervalos abaixo do valor 0 indicam superioridade média da Estratégia j em relação à Estratégia i. Em relação a média do coeficiente de Gini e da estatística de Kolmogorov-Smirnov e para todas as condições, as estratégias 2 e 3 também apresentam desempenho superior às estratégias 1a e 1b (tabelas B.6 a B.9). Já em relação à distância de Mahalanobis, assim como ocorreu para a comparação das proporções, não se pode dizer que as estratégias propostas são em média melhores que as demais para todas as condições. Nas condições com correlação de 0,5 (tabelas B.6 e B.8), as estratégias 2 e 3 são superiores às estratégias 1a e 1b. No entanto, nas condições com ausência de perturbação e correlação de 0,9 (Tabela B.9), a Estratégia 1b é ligeiramente superior às estratégias propostas. Embora as estratégias 2 e 3 tenham uma melhor performance que as estratégias 1a e 1b, tanto em relação à proporção de vezes em que elas são superiores quanto em relação à média das medidas (Gini e Kolmogorov-Smirnov), a diferença nos valores dos indicadores não é grande (tabelas B.6 a B.9). A diferença entre as médias do coeficiente de Gini e da estatística de Kolmogorov-Smirnov nunca é superior a 0,006. Além disso, para nenhuma medida é comum a ocorrência de grandes diferenças entre as estratégias.

As tabelas B.10 a B.13 apresentam estatísticas das diferenças das medidas na comparação das estratégias duas a duas. O terceiro quartil da diferença das estratégias 2 e 3 em relação às estratégias 1a e 1b, por exemplo, nunca é superior a 0,01 para o coeficiente de Gini e a estatística de Kolmogorov-Smirnov. Até mesmo as diferenças mínimas e máximas não são muito elevadas. Em relação às mesmas medidas, essas estatísticas nunca são superiores, em módulo, a 0,03.

5.2.2 Comparação entre as estratégias 2 e 3

Comparando-se as estratégias 2 e 3 em relação à proporção de vezes em que uma é superior a outra, pode-se notar que, para a distância de Mahalanobis e correlação 0,9, a Estratégia 2 é muito superior à Estratégia 3 (tabelas B.3 e B.5). Porém, nas demais condições e medidas, a Estratégia 3 ou tem desempenho semelhante à Estratégia 2 ou esta última tem performance ligeiramente superior (tabelas B.2 a B.5). Em relação a média das medidas (tabelas B.6 a B.9), a Estratégia 2 também é ligeiramente superior ou equivalente a Estratégia 3 para todas as condições e medidas. Nas condições em que a Estratégia 2 é superior a Estratégia 3, a diferença média para o coeficiente de Gini e a estatística de Kolmogorov-Smirnov nunca é superior a 0,0005. A diferença entre as estratégias 2 e 3 para essas mesmas medidas nunca excede, em módulo, 0,007 (tabelas B.10 a B.13) .

5.2.3 Comparação entre as estratégias 1a e 1b

O desempenho comparativo das estratégias 1a e 1b se altera de acordo com a condição. Para a correlação de 0,5 entre as respostas, a Estratégia 1b só tem desempenho equivalente a Estratégia 1a em relação a proporção de vezes em que uma é superior a outra e para uma medida e condição: estatística de Kolmogorov-Smirnov e presença de perturbação (Tabela B.2). Para todas as demais medidas e condições com correlação de 0,5 (tabelas B.2 e B.4), a Estratégia 1b tem performance superior. Em relação à média, para todas as medidas e condições com correlação de 0,5 (tabela B.6 e B.8), a Estratégia 1b é melhor que a Estratégia 1a. No entanto, pode-se observar que as diferenças médias (tabelas B.6 e B.8) e máximas (tabelas B.10 e B.12) não são grandes. Para a correlação 0,9 com perturbação, a Estratégia 1a tem performance superior tanto em relação à proporção (Tabela B.3) como em relação

à média (Tabela B.7). Na outra condição (tabelas B.5 e B.9), a melhor estratégia entre as duas varia de acordo com a medida, em relação à proporção e à média.

5.3 Conclusões da simulação

Nas condições simuladas, a conclusão é que, em geral, as estratégias propostas têm performance superior às usualmente utilizadas. A diferença entre elas é significativa, em relação à proporção de vezes em que as estratégias propostas são melhores que as geralmente utilizadas. No entanto, a intensidade da diferença entre essas estratégias não é muito grande, conforme pôde ser visto a partir das tabelas de médias e de estatísticas descritivas. Embora os resultados tenham tido alguma variação em função da condição da simulação, em geral, a correlação entre as respostas conta e a presença ou não de perturbação não parece exercer forte influência na performance relativa entre as estratégias 2 e 3 e as estratégias 1a e 1b.

Dentre as duas estratégias propostas, a Estratégia 2 apresenta, nas condições simuladas, desempenho, em geral, ligeiramente superior. Já a estratégia de melhor performance dentre as geralmente utilizadas varia em função da condição da simulação.

A simulação apresentada foi feita para um número reduzido de condições. Dessa forma, é interessante em estudos futuros, a simulação de diversas outras condições que não foram tratadas nesta dissertação. No Capítulo 6 serão citadas algumas das possíveis condições que podem ser simuladas posteriormente.

Capítulo 6

Conclusão

Nesta dissertação foram estudados os modelos de *customer scoring*. Esses modelos são utilizados para estimar a probabilidade de um cliente de uma instituição financeira ter problema de crédito em pelo menos um produto, em um horizonte de tempo pré-fixado. Foram apresentadas três estratégias para o desenvolvimento de modelos de *customer scoring*. A primeira, que possui duas variações, é a geralmente utilizada. As demais foram propostas neste trabalho. Foram discutidas as técnicas estatísticas e os modelos relacionados a cada uma das estratégias. Seus desempenhos foram comparados através de uma aplicação a dados reais, utilizando-se algumas medidas de performance que foram definidas. Uma simulação foi ainda desenvolvida para a comparação das estratégias em condições controladas.

Observando-se as características discutidas e os resultados da aplicação e da simulação, a Estratégia 2 parece ser a mais indicada para o desenvolvimento de modelos de *customer scoring*. Considerando-se o coeficiente de Gini, que é a medida mais indicada dentre as discutidas, a Estratégia 2 apresentou, em geral, performance ligeiramente superior às demais. Além disso, o tempo de desenvolvimento do modelo dessa estratégia é inferior ao observado nas estratégias geralmente utilizadas, já que ela não exige o desenvolvimento prévio de modelos para cada uma das famílias de produtos da instituição.

A Estratégia 3 apresenta alguns problemas práticos, em virtude de limitações dos algoritmos computacionais utilizados para o ajuste de modelos de GEE. Isso prejudicou sua performance na aplicação. No entanto, na simulação, a performance da Estratégia 3 foi superior a das estratégias usualmente utilizadas e apenas ligeira-

mente inferior ao desempenho da Estratégia 2. Assim, com o aperfeiçoamento dos algoritmos computacionais, essa estratégia pode se tornar uma boa opção. No futuro, ela tende a se tornar a estratégia mais indicada em pelo menos uma situação: quando se deseja também mensurar o risco associado a cada família de produtos, já que isso não pode ser obtido a partir da Estratégia 2.

Deve-se ressaltar que a simulação foi feita em condições bem simplificadas em relação ao que ocorre na prática. O número de variáveis, por exemplo, é geralmente muito maior. Também costuma ser maior o número de famílias de produtos. Além disso, em situações reais, a maioria dos clientes não possuem contas em todas as famílias de produtos. Na prática, também há clientes classificados como indeterminados ou cancelados. Embora, eles não sejam utilizados na estimação dos modelos, eles podem afetar a performance relativa das estratégias, já que a Estratégia 3 descarta um número maior de observações por esse motivo. Assim, para estudos futuros, sugere-se a comparação da performance das estratégias através da simulação de outras condições que são citadas abaixo.

- Simulação com um número maior de possíveis valores para a correlação entre as respostas em cada uma das famílias de produtos.
- Simulação variando o percentual de clientes que não possuem contas em todas as famílias de produto.
- Simulação variando a quantidade de variáveis por família de produtos.
- Simulação alterando a correlação entre as variáveis preditoras dos modelos.
- Simulação variando a quantidade de observações classificadas como indeterminadas ou canceladas. Nesse caso, pode-se ajustar modelos com variável resposta com distribuição multinomial. Uma outra possibilidade é proceder conforme foi feito na aplicação do Capítulo 4 e excluir as observações classificadas como indeterminadas ou canceladas.

Apêndice A

Estimativas dos parâmetros dos modelos

Tabela A.1: Estimativas, erros padrão e níveis descritivos dos parâmetros do customer scoring da Estratégia 1a

Efeito	Estimativa	Erro padrão	Nível descritivo
constante	7,398	0,519	$< 0,001$
x_{6032}	-2,088	0,228	$< 0,001$
x_{6045}	-0,290	0,120	0,016
x_{6046}	-1,203	0,503	0,017
x_{6052}	-1,244	0,480	0,010
x_{6053}	-1,506	0,494	0,002
x_{6062}	-0,349	0,122	0,004
x_{6073}	-0,242	0,119	0,043
x_{6074}	-0,604	0,138	$< 0,001$
x_{6075}	-0,972	0,222	$< 0,001$
x_{6082}	-0,406	0,195	0,037
x_{6083}	-0,496	0,153	0,001
x_{6084}	-0,610	0,172	$< 0,001$
x_{6093}	-0,548	0,293	0,062
d_{12}	-0,506	0,217	0,020
d_{13}	-0,820	0,300	0,006
d_{14}	-1,651	0,200	$< 0,001$
d_{15}	-1,831	0,301	$< 0,001$
d_{16}	-3,028	0,201	$< 0,001$
d_{17}	-3,473	0,196	$< 0,001$
d_{18}	-4,156	0,209	$< 0,001$

Tabela A.1: (continuação) Estimativas, erros padrão e níveis descritivos dos parâmetros do customer scoring da Estratégia 1a

Efeito	Estimativa	Erro padrão	Nível descritivo
d_{25}	-0,458	0,201	0,022
d_{26}	-0,742	0,188	$< 0,001$
d_{27}	-1,235	0,160	$< 0,001$
d_{28}	-1,455	0,236	$< 0,001$
d_{29}	-0,407	0,121	0,001
d_{34}	-0,569	0,127	$< 0,001$
d_{35}	-0,734	0,285	0,010

Tabela A.2: Estimativas, erros padrão e níveis descritivos dos parâmetros do customer scoring da Estratégia 1b

Efeito	Estimativa	Erro padrão	Níveis descritivos
constante	7,340	0,514	$< 0,001$
x_{6032}	-2,112	0,228	$< 0,001$
x_{6045}	-0,277	0,119	0,020
x_{6046}	-1,156	0,512	0,024
x_{6052}	-1,320	0,484	0,006
x_{6053}	-1,593	0,497	0,001
x_{6062}	-0,282	0,121	0,019
x_{6073}	-0,252	0,119	0,034
x_{6074}	-0,587	0,138	$< 0,001$
x_{6075}	-0,983	0,221	$< 0,001$
d_{12}	-0,927	0,221	$< 0,001$
d_{13}	-1,232	0,295	$< 0,001$
d_{14}	-1,639	0,218	$< 0,001$
d_{15}	-2,532	0,257	$< 0,001$
d_{16}	-3,249	0,207	$< 0,001$
d_{17}	-3,504	0,235	$< 0,001$
d_{18}	-3,854	0,219	$< 0,001$
d_{19}	-4,568	0,222	$< 0,001$
d_{25}	-0,671	0,148	$< 0,001$
d_{26}	-1,293	0,189	$< 0,001$
d_{27}	-1,366	0,162	$< 0,001$
d_{28}	-0,521	0,121	$< 0,001$
d_{34}	-0,552	0,128	$< 0,001$
d_{35}	-0,745	0,283	0,009

Tabela A.3: Estimativas, erros padrão e níveis descritivos dos parâmetros do customer scoring da Estratégia 2

Efeito	Estimativa	Erro padrão	Nível descritivo
constante	7,388	0,517	$< 0,001$
w_3	-0,728	0,288	0,011
x_{1019}	-0,633	0,191	0,001
x_{1052}	-0,655	0,165	$< 0,001$
x_{1053}	-1,245	0,227	$< 0,001$
x_{1063}	-0,404	0,206	0,050
x_{1073}	-0,436	0,160	0,007
x_{1074}	-0,597	0,245	0,015
x_{1124}	-0,676	0,200	0,001
x_{1125}	-1,043	0,224	$< 0,001$
x_{1126}	-1,379	0,172	$< 0,001$
x_{1127}	-1,751	0,193	$< 0,001$
x_{1128}	-2,059	0,170	$< 0,001$
x_{2044}	-0,577	0,200	0,004
x_{2045}	-0,948	0,143	$< 0,001$
x_{2053}	-0,289	0,134	0,032
x_{3024}	-1,957	0,897	0,029
x_{3062}	-0,546	0,117	$< 0,001$
x_{6032}	-2,251	0,231	$< 0,001$
x_{6045}	-0,297	0,120	0,013
x_{6046}	-1,203	0,500	0,016
x_{6052}	-1,292	0,488	0,008
x_{6053}	-1,565	0,501	0,002
x_{6062}	-0,316	0,122	0,009
x_{6073}	-0,237	0,120	0,048
x_{6074}	-0,604	0,138	$< 0,001$
x_{6075}	-0,963	0,220	$< 0,001$
x_{6082}	-0,426	0,194	0,028
x_{6083}	-0,563	0,153	$< 0,001$
x_{6084}	-0,803	0,169	$< 0,001$

Tabela A.4: Estimativas, erros padrão e níveis descritivos dos parâmetros do customer scoring da Estratégia 3r

Efeito	Estimativa	Erro padrão	Nível descritivo
constante	7,632	0,677	$< 0,001$
z_1	0,100	0,016	$< 0,001$
z_2	0,224	0,030	$< 0,001$
z_3	0,269	0,038	$< 0,001$
w_2	0,637	0,137	$< 0,001$
w_3	-0,658	0,306	0,032
x_{10190}	-0,655	0,195	$< 0,001$
x_{10340}	-0,784	0,157	$< 0,001$
x_{10350}	-1,000	0,236	$< 0,001$
x_{10360}	-1,717	0,178	$< 0,001$
x_{10370}	-2,431	0,207	$< 0,001$
x_{10480}	-0,399	0,132	$< 0,001$
x_{10490}	-0,705	0,210	$< 0,001$
x_{10520}	-0,523	0,194	0,007
x_{10530}	-1,253	0,206	$< 0,001$
x_{11030}	-0,639	0,150	$< 0,001$
x_{11040}	-1,139	0,170	$< 0,001$
x_{11050}	-1,743	0,192	$< 0,001$
x_{20270}	-0,673	0,319	0,035
x_{20460}	-0,970	0,166	$< 0,001$
x_{20520}	-0,503	0,154	$< 0,001$
x_{20640}	-0,416	0,149	0,005
x_{30630}	-0,509	0,130	$< 0,001$
x_{60320}	-2,642	0,248	$< 0,001$
x_{60440}	-0,298	0,119	0,012
x_{60450}	-0,456	0,135	$< 0,001$
x_{60520}	-1,677	0,640	0,009
x_{60530}	-1,927	0,651	$< 0,001$
x_{60620}	-0,345	0,133	0,010
x_{60730}	-0,275	0,131	0,036
x_{60740}	-0,497	0,156	$< 0,001$
x_{60750}	-0,930	0,228	$< 0,001$
x_{60830}	-0,397	0,129	$< 0,001$
x_{60840}	-0,457	0,163	0,005

Tabela A.5: Estimativas, erros padrão e níveis descritivos dos parâmetros do customer scoring da Estratégia 3s

Efeito	Estimativa	Erro padrão	Nível descritivo
constante	7,883	0,610	$< 0,001$
z_1	0,104	0,016	$< 0,001$
z_2	0,385	0,045	$< 0,001$
z_3	0,284	0,042	$< 0,001$
w_3	-0,648	0,295	0,028
x_{10522}	0,296	0,084	$< 0,001$
x_{10520}	-0,784	0,193	$< 0,001$
x_{10530}	-1,369	0,276	$< 0,001$
x_{10630}	-0,626	0,239	0,0088
x_{10731}	-0,172	0,046	$< 0,001$
x_{10732}	-0,220	0,074	$< 0,001$
x_{10733}	-0,283	0,068	$< 0,001$
x_{10740}	-0,758	0,282	0,0072
x_{11232}	-0,267	0,109	0,014
x_{11240}	-0,617	0,221	0,0051
x_{11250}	-1,171	0,230	$< 0,001$
x_{11260}	-1,485	0,182	$< 0,001$
x_{11270}	-1,781	0,205	$< 0,001$
x_{11280}	-2,280	0,180	$< 0,001$
x_{11290}	-2,904	0,187	$< 0,001$
x_{20273}	-0,240	0,070	$< 0,001$
x_{20450}	-0,660	0,222	$< 0,001$
x_{20460}	-1,001	0,151	$< 0,001$
x_{20522}	-0,187	0,059	$< 0,001$
x_{20642}	-0,184	0,052	$< 0,001$
x_{20742}	-0,163	0,061	0,0077
x_{20752}	-0,180	0,063	$< 0,001$
x_{30630}	-0,480	0,130	$< 0,001$
x_{60322}	-0,263	0,087	$< 0,001$
x_{60320}	-2,497	0,255	$< 0,001$
x_{60450}	-0,319	0,124	0,0102
x_{60520}	-1,648	0,576	$< 0,001$
x_{60530}	-1,899	0,588	$< 0,001$
x_{60620}	-0,344	0,131	0,009
x_{60730}	-0,310	0,129	0,016
x_{60740}	-0,614	0,158	$< 0,001$
x_{60750}	-0,943	0,223	$< 0,001$
x_{60830}	-0,490	0,122	$< 0,001$

Apêndice B

Tabelas da simulação

Tabela B.1: Proporção de vezes que a estratégia é a de melhor performance

Medida	Estratégia	Perturbação			
		Sim		Não	
		Correlação		Correlação	
		0,5	0,9	0,5	0,9
Coeficiente de Gini	1a	10%	23%	9%	22%
	1b	22%	10%	29%	16%
	2	38%	33%	35%	34%
	3	30%	34%	27%	28%
Estatística de Kolmogorov-Smirnov	1a	16%	22%	14%	20%
	1b	17%	11%	27%	20%
	2	40%	40%	29%	30%
	3	27%	27%	30%	30%
Distância de Mahalanobis	1a	12%	32%	6%	17%
	1b	36%	27%	41%	42%
	2	27%	37%	31%	34%
	3	25%	4%	22%	7%

Tabela B.2: Proporção de vezes que a Estratégia i é melhor que a Estratégia j - condição com perturbação e correlação de 0,5

Medida	Estratégia		Proporção	Intervalo de confiança	
	i	j	observada	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	43%	38%	47%
	3	1b	72%	68%	76%
	3	1a	82%	79%	85%
	2	1b	73%	69%	77%
	2	1a	84%	81%	87%
	1b	1a	65%	61%	69%
Estatística de Kolmogorov-Smirnov	3	2	43%	39%	48%
	3	1b	76%	73%	80%
	3	1a	76%	73%	80%
	2	1b	77%	74%	81%
	2	1a	77%	73%	80%
	1b	1a	53%	48%	57%
Distância de Mahalanobis	3	2	50%	46%	54%
	3	1b	58%	54%	62%
	3	1a	79%	75%	82%
	2	1b	58%	53%	62%
	2	1a	79%	76%	83%
	1b	1a	73%	69%	77%

Tabela B.3: Proporção de vezes que a Estratégia i é melhor que a Estratégia j - condição com perturbação e correlação de 0,9

Medida	Estratégia		Proporção	Intervalo de confiança	
	i	j	observada	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	46%	42%	51%
	3	1b	81%	77%	84%
	3	1a	69%	65%	73%
	2	1b	80%	76%	83%
	2	1a	70%	66%	74%
	1b	1a	32%	28%	36%
Estatística de Kolmogorov-Smirnov	3	2	41%	36%	45%
	3	1b	80%	76%	83%
	3	1a	68%	64%	72%
	2	1b	81%	78%	85%
	2	1a	71%	67%	75%
	1b	1a	38%	34%	42%
Distância de Mahalanobis	3	2	10%	8%	13%
	3	1b	54%	49%	58%
	3	1a	52%	47%	56%
	2	1b	61%	57%	65%
	2	1a	58%	53%	62%
	1b	1a	47%	43%	51%

Tabela B.4: Proporção de vezes que a Estratégia i é melhor que a Estratégia j - condição sem perturbação e correlação de 0,5

Medida	Estratégia		Proporção	Intervalo de confiança	
	i	j	observada	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	44%	40%	49%
	3	1b	65%	61%	69%
	3	1a	83%	79%	86%
	2	1b	66%	62%	70%
	2	1a	84%	81%	87%
	1b	1a	75%	71%	79%
Estatística de Kolmogorov-Smirnov	3	2	49%	45%	54%
	3	1b	65%	61%	70%
	3	1a	77%	74%	81%
	2	1b	65%	61%	69%
	2	1a	76%	72%	80%
	1b	1a	62%	58%	67%
Distância de Mahalanobis	3	2	42%	38%	47%
	3	1b	55%	51%	60%
	3	1a	85%	82%	88%
	2	1b	56%	52%	61%
	2	1a	86%	83%	89%
	1b	1a	84%	81%	87%

Tabela B.5: Proporção de vezes que a Estratégia i é melhor que a Estratégia j - condição sem perturbação e correlação de 0,9

Medida	Estratégia		Proporção	Intervalo de confiança	
	i	j	observada	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	43%	39%	47%
	3	1b	71%	67%	75%
	3	1a	63%	59%	67%
	2	1b	74%	70%	78%
	2	1a	69%	65%	73%
	1b	1a	42%	37%	46%
Estatística de Kolmogorov-Smirnov	3	2	49%	45%	54%
	3	1b	66%	62%	70%
	3	1a	66%	62%	70%
	2	1b	69%	65%	73%
	2	1a	68%	64%	72%
	1b	1a	53%	49%	58%
Distância de Mahalanobis	3	2	13%	10%	16%
	3	1b	38%	33%	42%
	3	1a	54%	50%	59%
	2	1b	47%	43%	52%
	2	1a	64%	60%	69%
	1b	1a	69%	65%	73%

Tabela B.6: Comparação da Média na Estratégia i e na Estratégia j - Condição com perturbação e correlação de 0,5

Medida	Estratégia		Média na Estrat.		Diferença	I. C. para a dif. média	
	i	j	i	j	média	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	0,7629	0,7631	-0,0001	-0,0002	-0,0001
	3	1b	0,7629	0,7611	0,0019	0,0016	0,0022
	3	1a	0,7629	0,7596	0,0033	0,0030	0,0037
	2	1b	0,7631	0,7611	0,0020	0,0017	0,0023
	2	1a	0,7631	0,7596	0,0035	0,0031	0,0038
	1b	1a	0,7611	0,7596	0,0015	0,0011	0,0018
Estatística de Kolmogorov-Smirnov	3	2	0,6566	0,6567	-0,0001	-0,0002	0,0000
	3	1b	0,6566	0,6536	0,0031	0,0027	0,0034
	3	1a	0,6566	0,6531	0,0035	0,0031	0,0039
	2	1b	0,6567	0,6536	0,0032	0,0028	0,0035
	2	1a	0,6567	0,6531	0,0036	0,0032	0,0040
	1b	1a	0,6536	0,6531	0,0004	0,0000	0,0009
Distância de Mahalanobis	3	2	2,7722	2,7722	-0,0001	-0,0001	0,0000
	3	1b	2,7722	2,7692	0,0030	0,0018	0,0042
	3	1a	2,7722	2,7579	0,0143	0,0128	0,0158
	2	1b	2,7722	2,7692	0,0030	0,0018	0,0043
	2	1a	2,7722	2,7579	0,0143	0,0128	0,0159
	1b	1a	2,7692	2,7579	0,0113	0,0097	0,0129

Tabela B.7: Comparação da Média na Estratégia i e na Estratégia j - Condição com perturbação e correlação de 0,9

Medida	Estratégia		Média na Estrat.		Diferença	I. C. para a dif. média	
	i	j	i	j	média	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	0,7416	0,7416	0,0000	-0,0001	0,0001
	3	1b	0,7416	0,7374	0,0042	0,0037	0,0046
	3	1a	0,7416	0,7396	0,0020	0,0016	0,0024
	2	1b	0,7416	0,7374	0,0042	0,0037	0,0046
	2	1a	0,7416	0,7396	0,0020	0,0016	0,0024
	1b	1a	0,7374	0,7396	-0,0022	-0,0026	-0,0017
Estatística de Kolmogorov-Smirnov	3	2	0,6401	0,6405	-0,0003	-0,0005	-0,0002
	3	1b	0,6401	0,6352	0,0050	0,0045	0,0055
	3	1a	0,6401	0,6376	0,0026	0,0021	0,0030
	2	1b	0,6405	0,6352	0,0053	0,0048	0,0058
	2	1a	0,6405	0,6376	0,0029	0,0024	0,0034
	1b	1a	0,6352	0,6376	-0,0024	-0,0029	-0,0018
Distância de Mahalanobis	3	2	2,7245	2,7280	-0,0034	-0,0037	-0,0032
	3	1b	2,7245	2,7233	0,0012	-0,0004	0,0028
	3	1a	2,7245	2,7244	0,0001	-0,0017	0,0018
	2	1b	2,7280	2,7233	0,0047	0,0031	0,0063
	2	1a	2,7280	2,7244	0,0035	0,0018	0,0052
	1b	1a	2,7233	2,7244	-0,0012	-0,0030	0,0007

Tabela B.8: Comparação da Média na Estratégia i e na Estratégia j - Condição sem perturbação e correlação de 0,5

Medida	Estratégia		Média na Estrat.		Diferença	I. C. para a dif. média	
	i	j	i	j	média	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	0,8423	0,8423	0,0000	-0,0001	0,0000
	3	1b	0,8423	0,8415	0,0008	0,0006	0,0010
	3	1a	0,8423	0,8394	0,0029	0,0026	0,0031
	2	1b	0,8423	0,8415	0,0008	0,0006	0,0010
	2	1a	0,8423	0,8394	0,0029	0,0026	0,0032
	1b	1a	0,8415	0,8394	0,0021	0,0018	0,0024
Estatística de Kolmogorov-Smirnov	3	2	0,7249	0,7249	0,0000	-0,0001	0,0001
	3	1b	0,7249	0,7232	0,0017	0,0013	0,0021
	3	1a	0,7249	0,7213	0,0036	0,0032	0,0040
	2	1b	0,7249	0,7232	0,0017	0,0014	0,0021
	2	1a	0,7249	0,7213	0,0036	0,0032	0,0040
	1b	1a	0,7232	0,7213	0,0019	0,0014	0,0024
Distância de Mahalanobis	3	2	3,1359	3,1362	-0,0002	-0,0003	-0,0002
	3	1b	3,1359	3,1334	0,0025	0,0010	0,0041
	3	1a	3,1359	3,1153	0,0207	0,0189	0,0225
	2	1b	3,1362	3,1334	0,0028	0,0012	0,0043
	2	1a	3,1362	3,1153	0,0209	0,0191	0,0227
	1b	1a	3,1334	3,1153	0,0181	0,0164	0,0199

Tabela B.9: Comparação da Média na Estratégia i e na Estratégia j - Condição sem perturbação e correlação de 0,9

Medida	Estratégia		Média na Estrat.		Diferença	I. C. para a dif. média	
	i	j	i	j	média	Lim. inf.	Lim. sup.
Coeficiente de Gini	3	2	0,8473	0,8477	-0,0005	-0,0006	-0,0003
	3	1b	0,8473	0,8451	0,0022	0,0018	0,0025
	3	1a	0,8473	0,8462	0,0011	0,0008	0,0014
	2	1b	0,8477	0,8451	0,0026	0,0023	0,0030
	2	1a	0,8477	0,8462	0,0015	0,0013	0,0018
	1b	1a	0,8451	0,8462	-0,0011	-0,0014	-0,0007
Estatística de Kolmogorov-Smirnov	3	2	0,7349	0,7351	-0,0002	-0,0004	0,0000
	3	1b	0,7349	0,7329	0,0020	0,0015	0,0024
	3	1a	0,7349	0,7324	0,0025	0,0020	0,0030
	2	1b	0,7351	0,7329	0,0021	0,0017	0,0026
	2	1a	0,7351	0,7324	0,0027	0,0022	0,0031
	1b	1a	0,7329	0,7324	0,0005	0,0001	0,0010
Distância de Mahalanobis	3	2	3,1986	3,2072	-0,0086	-0,0104	-0,0067
	3	1b	3,1986	3,2098	-0,0112	-0,0138	-0,0085
	3	1a	3,1986	3,1983	0,0003	-0,0025	0,0030
	2	1b	3,2072	3,2098	-0,0026	-0,0045	-0,0007
	2	1a	3,2072	3,1983	0,0088	0,0067	0,0109
	1b	1a	3,2098	3,1983	0,0114	0,0095	0,0134

Tabela B.10: Medidas descritivas para a diferença entre a Estratégia i e a Estratégia j - Condição com perturbação e correlação de 0,5

Medida	Estratégia		Estatísticas da diferença entre as estratégias i e j				
	i	j	Mínimo	Q1	Mediana	Q3	Máximo
Coeficiente de Gini	3	2	-0,0035	-0,0003	-0,0001	0,0002	0,0032
	3	1b	-0,0124	-0,0003	0,0018	0,0042	0,0153
	3	1a	-0,0095	0,0011	0,0032	0,0057	0,0166
	2	1b	-0,0118	-0,0002	0,0021	0,0043	0,0142
	2	1a	-0,0096	0,0012	0,0034	0,0059	0,0174
	1b	1a	-0,0132	-0,0011	0,0012	0,0040	0,0145
Estatística de Kolmogorov-Smirnov	3	2	-0,0043	-0,0006	0,0000	0,0004	0,0047
	3	1b	-0,0077	0,0001	0,0030	0,0058	0,0241
	3	1a	-0,0099	0,0002	0,0033	0,0066	0,0220
	2	1b	-0,0070	0,0002	0,0029	0,0059	0,0245
	2	1a	-0,0090	0,0002	0,0033	0,0065	0,0223
	1b	1a	-0,0157	-0,0029	0,0003	0,0038	0,0167
Distância de Mahalanobis	3	2	-0,0056	-0,0006	0,0000	0,0005	0,0026
	3	1b	-0,0513	-0,0064	0,0026	0,0122	0,0405
	3	1a	-0,0351	0,0020	0,0145	0,0263	0,0589
	2	1b	-0,0518	-0,0062	0,0027	0,0124	0,0417
	2	1a	-0,0344	0,0018	0,0143	0,0262	0,0596
	1b	1a	-0,0497	-0,0009	0,0093	0,0234	0,0788

Tabela B.11: Medidas descritivas para a diferença entre a Estratégia i e a Estratégia j - Condição com perturbação e correlação de 0,9

Medida	Estratégia		Estatísticas da diferença entre as estratégias i e j				
	i	j	Mínimo	Q1	Mediana	Q3	Máximo
Coeficiente de Gini	3	2	-0,0062	-0,0007	-0,0001	0,0007	0,0062
	3	1b	-0,0141	0,0011	0,0040	0,0075	0,0245
	3	1a	-0,0187	-0,0008	0,0023	0,0051	0,0258
	2	1b	-0,0156	0,0008	0,0042	0,0078	0,0262
	2	1a	-0,0191	-0,0009	0,0021	0,0048	0,0230
	1b	1a	-0,0316	-0,0049	-0,0020	0,0008	0,0232
Estatística de Kolmogorov-Smirnov	3	2	-0,0061	-0,0014	-0,0003	0,0006	0,0051
	3	1b	-0,0098	0,0006	0,0047	0,0090	0,0249
	3	1a	-0,0169	-0,0010	0,0025	0,0062	0,0179
	2	1b	-0,0092	0,0013	0,0050	0,0094	0,0241
	2	1a	-0,0164	-0,0009	0,0028	0,0067	0,0180
	1b	1a	-0,0257	-0,0062	-0,0021	0,0019	0,0200
Distância de Mahalanobis	3	2	-0,0155	-0,0052	-0,0029	-0,0012	0,0056
	3	1b	-0,0711	-0,0091	0,0011	0,0129	0,0702
	3	1a	-0,0801	-0,0122	0,0008	0,0132	0,0695
	2	1b	-0,0579	-0,0055	0,0061	0,0164	0,0693
	2	1a	-0,0792	-0,0076	0,0037	0,0157	0,0718
	1b	1a	-0,0972	-0,0140	-0,0016	0,0119	0,0659

Tabela B.12: Medidas descritivas para a diferença entre a Estratégia i e a Estratégia j - Condição sem perturbação e correlação de 0,5

Medida	Estratégia		Estatísticas da diferença entre as estratégias i e j				
	i	j	Mínimo	Q1	Mediana	Q3	Máximo
Coeficiente de Gini	3	2	-0,0025	-0,0003	0,0000	0,0001	0,0023
	3	1b	-0,0068	-0,0009	0,0008	0,0023	0,0085
	3	1a	-0,0046	0,0009	0,0026	0,0046	0,0176
	2	1b	-0,0070	-0,0006	0,0008	0,0023	0,0087
	2	1a	-0,0054	0,0010	0,0028	0,0046	0,0172
	1b	1a	-0,0070	0,0001	0,0018	0,0039	0,0135
Estatística de Kolmogorov-Smirnov	3	2	-0,0065	-0,0006	0,0000	0,0005	0,0042
	3	1b	-0,0123	-0,0011	0,0016	0,0048	0,0152
	3	1a	-0,0089	0,0004	0,0034	0,0070	0,0216
	2	1b	-0,0106	-0,0012	0,0019	0,0048	0,0151
	2	1a	-0,0080	0,0002	0,0035	0,0070	0,0223
	1b	1a	-0,0136	-0,0017	0,0016	0,0053	0,0199
Distância de Mahalanobis	3	2	-0,0049	-0,0009	-0,0002	0,0005	0,0031
	3	1b	-0,1064	-0,0085	0,0027	0,0139	0,0596
	3	1a	-0,0595	0,0071	0,0213	0,0340	0,0965
	2	1b	-0,1050	-0,0079	0,0029	0,0138	0,0624
	2	1a	-0,0601	0,0075	0,0211	0,0347	0,0972
	1b	1a	-0,0528	0,0049	0,0170	0,0290	0,1106

Tabela B.13: Medidas descritivas para a diferença entre a Estratégia i e a Estratégia j - Condição sem perturbação e correlação de 0,9

Medida	Estratégia		Estatísticas da diferença entre as estratégias i e j				
	i	j	Mínimo	Q1	Mediana	Q3	Máximo
Coeficiente de Gini	3	2	-0,0124	-0,0009	-0,0001	0,0004	0,0039
	3	1b	-0,0110	-0,0007	0,0020	0,0046	0,0204
	3	1a	-0,0121	-0,0010	0,0009	0,0031	0,0132
	2	1b	-0,0090	-0,0001	0,0021	0,0049	0,0195
	2	1a	-0,0085	-0,0005	0,0013	0,0033	0,0125
	1b	1a	-0,0181	-0,0029	-0,0005	0,0012	0,0129
Estatística de Kolmogorov-Smirnov	3	2	-0,0138	-0,0012	0,0000	0,0011	0,0099
	3	1b	-0,0154	-0,0011	0,0019	0,0051	0,0186
	3	1a	-0,0139	-0,0013	0,0022	0,0061	0,0266
	2	1b	-0,0123	-0,0009	0,0020	0,0051	0,0166
	2	1a	-0,0121	-0,0009	0,0026	0,0059	0,0245
	1b	1a	-0,0159	-0,0026	0,0005	0,0035	0,0240
Distância de Mahalanobis	3	2	-0,1658	-0,0069	-0,0037	-0,0013	0,0059
	3	1b	-0,1513	-0,0227	-0,0071	0,0068	0,0555
	3	1a	-0,1519	-0,0153	0,0039	0,0200	0,0778
	2	1b	-0,0834	-0,0160	-0,0009	0,0111	0,0530
	2	1a	-0,0801	-0,0068	0,0095	0,0246	0,0896
	1b	1a	-0,0565	-0,0041	0,0108	0,0238	0,0966

Referências Bibliográficas

[1] Agresti, A. (1990). Categorical data analysis, John Wiley and Sons: New York.

[2] Arminger, G., Enache, D. and Bonne, T (1997). Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward neural networks. Computational Statistics, 12, 293-310.

[3] Artes, R (1997). Extensão da teoria das equações de estimação generalizadas a dados circulares e modelos de dispersão. Tese de Doutorado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[4] Banasik, J., Crook, J. N. and Thomas, L. C. (1999). Not if but when will borrowers default. Journal of the operational research society , 50, 1185-1190.

[5] Blackwell, M. and Sykes, C. (1992). The assignment of credit limits with a behaviour-scoring system. IMA Journal of Mathematics Applied in Business and Industry, 4, 73-80.

[6] Bussab, W. O. e Morettin P. A. (2002). Estatística básica, 5 Ed. Saraiva: São Paulo.

[7] Collett, D. (1991). Modelling binary data, Chapman and Hall: New York.

[8] Conover, W. J. (1999). Practical nonparametric statistics, 3 Ed. John Wiley and Sons: New York.

[9] Crowder, M. (1987). On linear and quadratic estimating function. Biometrika, 74, 591-597.

[10] Dobson, A. (1983). An Introduction to Statistical Modelling, Chapman and Hall: London.

[11] Feelders, A. J. (1999). Credit scoring and reject inference with mixture models. International Journal of Inteligent Systems in Accounting Finance and Management, 8(4), 271-279.

[12] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

[13] Grablowsky, B. J. and Talley, W. K. (1981). Probit and discriminat functions for classifying credit applicants: a comparison. Journal of Economics and Business, 33, 254-261.

[14] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. Annals of Mathematical Statistics, 31, 1208-1211.

[15] Good, I. J. (1950). Probability and the weighting of evidence, Charles Griffin: London.

[16] Grizzle, J., Starmer, F. and Koch, G. (1969). Analysis of categorical data by linear models. Biometrics, 25, 489-504.

[17] Groom, G. and Gill, L. (1998). Customer Scoring - Practical Issues for Development Success. In InterAct98 Conference, Fair, Isaac and Company Inc., San Francisco.

[18] Hand, D. J. (1981). Discrimination and Classification, Chichester: Wiley.

[19] Hand, D. J. (1998). Reject Inference in credit operations. In Credit Risk Modelling Design and Application, ed E. Mays, Glenlake Publishinh: Chicago, 181-190.

[20] Hand, D. J. and Henley, D. J. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society, Series A, 160, part 3, 523-541.

[21] Hand, D. J., Oliver, J. J. and Lunn, A. D. (1998). Discriminant analysis when the classes arise from a continuum. Pattern Recognition, 31, 641-650.

[22] Hardin, J. W. and Hilbe J. M. (2003). Generalized Estimating Equations, Chapman and Hall: New York.

[23] Hauck, W. W. and Donner, A. (1977). Wald´s test as applied to hypotheses in logit analysis. Journal of the American Statistical Association , 72, 851-853.

[24] Henley, W. E. (1995). Statistical aspets of credit scoring. PhD Thesis. The Open University, Milton Keynes.

[25] Henley, W. E. and Hand, D. J. (1997). Construction of a K-nearest neighbour credit scoring system. IMA Journal of Mathematics Applied in Business and Industry, 8, 143-151.

[26] Hoper, M. A. and Lewis, E. M. (1992). Behaviour scoring and adptive control systems. In Credit Scoring and Credit Control, ed L. C. Thomas, J. N. Crook, D. B. Edelman, Claredon Press: Oxford.

[27] Hosmer, D. W. and Lemeshow S. (1980). A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics, A10, 1043-1069.

[28] Hosmer, D. W. and Lemeshow S. (1989). Applied logistic regression, John Wiley and Sons: New York.

[29] Johnson, M. E. (1987). Multivariate statistical simulation, John Wiley and Sons: New York.

[30] Johnson, R. A. and Wichern D. W. (1998). Applied Multivariate Statistical Analysis, 4 Ed. Prentice-Hall: New York.

[31] Jørgensen, B. and Labouriau, R. S. (1994). Exponential families and theoretical inference. Lecture Notes. Departament of Statistical. University of British Columbia.

[32] Kass, G. V. (1980). An explanatory technique for investigating large quantiles of categorical data. Applied Statistics, 29, 119-127.

[33] Li, H. G. and Hand, D. J. (2002). Direct versus indirect credit scoring classifications. Journal of Operational Research Society, 53, 647-654.

[34] Liang, K. Y. and Zeger, S. L. (1986). Longitunidal data analysis using generalized linear models. Biometrika, 73, 13-22.

[35] Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data. Journal of the Royal Statistical Society, 54, 3-40.

[36] Magalhaes, M. N. e Lima A. C. P. (2001). Noções de probabilidade e estatística, 3 Ed. Adusp: São Paulo.

[37] Magidson, J. (1994). The CHAID approach to segmentation modelling: Chi-squared Automatic Interaction Detection. In Advanced Methods of Marketing Research, ed Bagozzi, R. P., Blackwell Publishers: Cambridge.

[38] McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2 Ed. Chapman and Hall: London.

[39] McNab, H. and Wynn, A. (2000). Principles and Practice of Consumer Credit Risk Management, Financial World Publishing: Kent.

[40] Neter, J., Kutner, M. K., Nachtsheim, C. J., Wasserman, W. (1996). Applied Linear Statistical Models, 4 Ed. Irwin: Chicago.

[41] Oliveira, J. G. C. e Andrade, F. W. M. (2002). Comparação de medidas de performance de modelos de credit scoring. Tecnologia de Crédito, 33, 1-11.

[42] Oliver, R. M. (1993). Effects of calibrations and discrination on profitability scoring. In Proceedings of Credit Scoring and Credit Control III , Credit Research Centre, University of Edinburgh.

[43] Orgler, Y. E. (1970). A credit scoring for comercial loans. Journal of Money, Credit and Banking, november, 31-37.

[44] Ohtoshi, C. (2003). Uma comparação de regressão logística, árvores de classificação e redes neurais: analisando dados de crédito. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[45] Paula, G. A. (2000). Modelos de Regressão com apoio computacional, Instituto de Matemática e Estatítica, Universidade de São Paulo.

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

[47] Showers, J. L. and Chakrin, L. M. (1981). Reducing uncollectable revenue from residential telephone customers. Interfaces, 11, 21-31.

[48] Stepanova, M. and Thomas, L. C. (2001). PHAB scores: proportional hazards analysis behavioural scores. Journal of the Operational Research Society, 52, 1007-1016.

[49] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 16, 149-172.

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

[51] Thomas, L. C., Ho, J and Scherer, W. T. (2001). Time will tell: behaviour scoring and the dynamics of consumer credit assessment . IMA Journal of Management Mathematics, 12, 89-103.

[52] Tsai, H. T. and Yeh, H. C. (1999). A Two-stage screening procedure for mailing credit assessment. IMA Journal of Mathematics Applied in Business and Industry, 10, 317-329.

[53] Venezuela, M. K. (2003). Modelos lineares generalizados para análise de dados com medidas repetidas. Dissertação de Mestrado. Instituto de Matemática e Estatística, Universidade de São Paulo.

[54] Wedderburn, R. W. M. (1974). Quasi-likelihood function, generalized linear models, and the Gauss-Newton method. Biometrika, 61, 439-447.

[55] West, D. (2000). Neural network credit scoring problems. Computers and Operational Research, 27, 1131-1152.