Jurandir Prazeres Filho

Capacidade preditiva de Modelos *Credit Scoring* em inferência dos rejeitados

São Carlos, março de 2014

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
PPGEst - PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Jurandir Prazeres Filho

Capacidade preditiva de Modelos *Credit Scoring* em inferência dos rejeitados

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Francisco Louzada Neto

São Carlos, março de 2014

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

P921cp

Prazeres Filho, Jurandir.
Capacidade preditiva de Modelos *Credit Scoring* em inferência dos rejeitados / Jurandir Prazeres Filho. -- São Carlos : UFSCar, 2014.
92 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2014.

1. Análise de regressão. 2. Modelos estatísticos. 3. *Credit scoring*. 4. Inferência dos rejeitados. I. Título.

CDD: 519.536 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Jurandir Prazeres Filho

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 28/03/2014
PELA COMISSÃO JULGADORA:

Presidente _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP/Orientador)

1º Examinador _____
Prof. Dr. Gilberto de Araújo Pereira (UFTM)

2º Examinador _____
Prof. Dr. Luis A. Milan (DEs-UFSCar)

"*Combati o combate, completei a carreira,*
guardei a fé"

(II Timóteo 4:6-7)

Agradecimentos

Agradecimentos a Deus, força criadora, que sempre me protege e me ilumina nos momentos mais difíceis.

Aos meus pais, que sempre me incentivaram a estudar e sempre oraram por mim. Aos meus irmãos, Marcos Antônio, Ana Paula, Rita de Cássia, João Paulo (JP) e Bel, onde encontro paz, alegria, harmonia e aconchego.

Ao Prof. Dr. Francisco Louzada Neto e ao Prof. Dr. Gilberto de Araújo Pereira pela dedicação, paciência e apoio.

Ao meu colega de casa, Miro Vigas, pela convivência harmoniosa e pelos ensinamentos. Aos colegas que contribuíram diretamente para a realização deste trabalho, Ricardo Rocha, Alexandre Maiorano, Lorena Janete Cáceres. Aos professores Milan, Adriano Polpo, Galvão e Luís Ernesto. À Isabel, pela educação e receptividade. A Edmundo (Dito), Ronald, tio Pedro, Rogério, Vinícius, Jonatas, Wilson, aos meus primos e amigos.

A minha noiva, Bárbara Ivânia, pelo amor, carinho, dedicação, respeito, por ter palavras de esperança para me oferecer e sempre acreditar em mim.

À Universidade Federal de São Carlos e a CAPES pela oportunidade concedida de estudar e obter mais conhecimento.

Por fim, ao meu irmão JP, esta dissertação, dedico, ofereço e consagro.

Resumo

A concessão de crédito é uma decisão a ser tomada num contexto de incertezas. No momento em que o credor decide conceder um empréstimo, realizar um financiamento ou venda a prazo sempre existe a possibilidade de perda, e, se for atribuída uma probabilidade a esta perda, a decisão de conceder ou não crédito será mais confiável. Com o objetivo de auxiliar a tomada de decisão em relação ao pedido de crédito dos solicitantes são utilizados os modelos *credit scoring*, os quais estimam a probabilidade de perda associada à concessão de crédito. Um dos problemas envolvendo estes modelos é que somente informações a respeito dos proponentes aceitos são utilizadas, o que causa um viés amostral, pois, os solicitantes recusados são descartados no processo de modelagem. Com intuito de solucionar este problema tem-se a inferência dos rejeitados, em que são considerados os indívíduos que tiveram pedido de crédito rejeitado. No entanto, considerar a inferência dos rejeitados e o uso de somente um método de modelagem de dados, muitas vezes, não é suficiente para que se tenha medidas preditivas satisfatórias. Desta forma, foram utilizados resultados combinados de três metodologias, regressão logística, *probit* e árvore de decisão/classificação concomitantemente a utilização dos métodos de inferência dos rejeitados que incluem o uso de variável latente, reclassificação, parcelamento e ponderação. O objetivo dessa combinação foi aumentar a capacidade preditiva e as métricas utilizadas foram a sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia. Através da aplicação em conjuntos de dados concluiu-se que a utilização dos resultados combinados aumentou a capacidade preditiva, principalmente, em relação a sensibilidade.

Palavras-Chave: Modelos, *Credit scoring*, Inferência dos rejeitados.

Abstract

Granting credit to an applicant is a decision made in a context of uncertainty. At the moment the lender decides to grant a loan or credit sale there is always the possibility of loss, and, if it is associated with a probability, the decision to grant or not credit will be more reliable. In order to aid the decision to accept or not the request for applicants are used the *credit scoring* models, which estimate the probability of loss associated with granting credit. But one of the problems involving these models is that only information about the applicants accepted are used, which causes a sampling bias, because the rejected applicants are discarded. With the aim to solve this problem it can use rejected inference, which are considered individuals who have had credit application rejected. However, only considering rejected inference and one method of modeling data, usually, is not sufficient to get satisfactory predictive measures, and thus, were used combined results of three methods, logistic regression, analysis *probit* and decision tree. The purpose of this combination were to increase the predictive perfomance and the metrics used were sensitivity, specificity , positive predictive value, negative predictive value and accuracy. Through the application in data sets we concluded that the use of the combined results increased the predictive performance, specially regarding to sensitivity.

Keywords: Models, Risk, *Credit Scoring*, Inference of rejected.

Lista de Figuras

2.1 Exemplo de árvore de decisão/classificação 30

3.1 Curva ROC criada pelos pontos de corte do conjunto de teste 39
3.2 Área sob a curva ROC 40

4.1 Esquema da distribuição dos dados para um modelo de *credit scoring* . . . 43

8.1 Árvore de decisão/classificação - UVL 78
8.2 Árvore de decisão/classificação - reclassificação 81
8.3 Árvore de decisão/classificação - parcelamento 84
8.4 Árvore de decisão/classificação - ponderação 87

Lista de Tabelas

3.1 Representação da matriz de confusão 36
3.2 Probabilidade conjunta (T,Y) 37
3.3 Observações e probabilidade (valor) 38

4.1 Cálculo dos pesos para aceitos em cada classe de risco - AR 46
4.2 Exemplo numérico do cálculo dos pesos para os aceitos em cada classe de risco - AR . . 46
4.3 Distribuição de risco para aceitos e parcelamento dos rejeitados 48
4.4 Probabilidades de ocorrência do resultado do modelo para cada indivíduo, segundo o comportamento (bom ou mau pagador) do indivíduo 50

6.1 Média das medidas preditivas dos modelos (p = 10%) - uso de variável latente 63
6.2 Média das medidas preditivas dos modelos (p = 25%) - uso de variável latente 63
6.3 Média das medidas preditivas dos modelos (p = 50%) - uso de variável latente 63
6.4 Média das medidas preditivas dos modelos (p = 10%) - reclassificação 64
6.5 Média das medidas preditivas dos modelos (p = 25%) - reclassificação 64
6.6 Média das medidas preditivas dos modelos (p = 50%) - reclassificação 64
6.7 Média das medidas preditivas dos modelos (p = 10%) - parcelamento 65
6.8 Média das medidas preditivas dos modelos (p = 25%) - parcelamento 65
6.9 Média das medidas preditivas dos modelos (p = 50%) - parcelamento 65
6.10 Média das medidas preditivas dos modelos (p = 10%) - ponderação 66
6.11 Média das medidas preditivas dos modelos (p = 25%) - ponderação 66
6.12 Média das medidas preditivas dos modelos (p = 50%) - ponderação 66
6.13 Medidas preditivas dos modelos e do uso combinado - uso de variável latente 68
6.14 Medidas preditivas dos modelos e do uso combinado - reclassificação 69
6.15 Medidas preditivas dos modelos e do uso combinado - parcelamento 69
6.16 Medidas preditivas dos modelos e do uso combinado - ponderação 69

8.1 Distribuição da inadimplência segundo tipo de ocupação 74
8.2 Distribuição da inadimplência segundo escolaridade. 74
8.3 Distribuição da inadimplência segundo o estado civil. 74
8.4 Distribuição da inadimplência segundo canal de entrada. 74
8.5 Distribuição da inadimplência segundo o sexo. 75
8.6 Distribuição da inadimplência segundo tipo de residência. 75
8.7 Distribuição da inadimplência segundo tipo de telefone. 75
8.8 Estatísticas - idade. 75
8.9 Estatísticas - número de dependentes. 75
8.10 Estatísticas - renda. 75
8.11 Estimativas, EP e valores p dos parâmetros do modelo logístico - UVL 76
8.12 Estimativas, EP e valores p dos parâmetros do modelo *probit* - UVL. 77
8.13 Estimativas, EP e valores p dos parâmetros do modelo logístico - reclassificação 79
8.14 Estimativas, EP e valores p dos parâmetros do modelo *probit* - reclassificação. 80
8.15 Estimativas, EP e valores p dos parâmetros do modelo logístico - parcelamento. 82
8.16 Estimativas, EP e valores p dos parâmetros do modelo *probit* - parcelamento. 83
8.17 Estimativas, EP e valores p dos parâmetros do modelo logístico - ponderação. 85
8.18 Estimativas, EP e valores p dos parâmetros do modelo *probit* - ponderação. 86

Sumário

1 Introdução **11**

1.1 Justificativa 14

1.2 Organização dos capítulos 16

2 Métodos utilizados em *credit scoring* **17**

2.1 Regressão logística 17

2.2 Regressão *probit* 25

2.3 Árvore de decisão/classificação 28

2.3.1 Aspectos gerais 28

2.3.2 Um exemplo de árvore de decisão/classificação 29

2.3.3 Algoritmo CART 30

3 Avaliação de capacidade preditiva do modelo **33**

3.1 Medidas Preditivas 33

3.2 Curva ROC 36

3.2.1 Conceitos básicos 36

3.2.2 Gráfico ROC e AUC 38

3.3 Avaliação do modelo 41

4 Inferência dos rejeitados **43**

4.1 O Problema 43

4.2 Principais Métodos 45

4.2.1 Reclassificação 45

4.2.2 Ponderação 45

4.2.3 Parcelamento 47

4.3 Método utilizado 49

4.3.1 Uso de variável latente 49

4.3.2 Algoritmo EM 53

5 Regras de decisão **57**

6 Exemplo de Aplicação **60**

6.1 Banco de dados e simulações 60

6.2 Resultados 62

6.2.1 Aplicação em dados simulados 62

6.2.2 Aplicação em dados reais 68

7 Considerações Finais **71**

7.1 Conclusões 71

7.2 Trabalhos futuros 72

8 Apêndice **73**

8.1 Dados reais 73

8.1.1 Descrição das variáveis 73

8.1.2 Análise descritiva 74

8.1.3 Modelos 76

Referências Bibliográficas **88**

Capítulo 1

Introdução

Os modelos de *credit scoring* prevêm, na data de decisão do crédito, a probabilidade de perda associada a concessão de crédito e tal probabilidade de perda em uma operação de crédito é denominada risco de crédito.

No início do século passado, a concessão de crédito, definida como quantidade de dinheiro emprestado ao solicitante por uma instituição financeira, era baseada na experiência do analista, o que tornava o processo subjetivo. Porém, com o avanços computacionais e o aumento na demanda por crédito, tornou-se necessário transformar a decisão de crédito num processo mais objetivo, e, isto foi realizado por meio de ferramentas estatísticas num processo gradual e lento (Sicsu, 2010).

Os modelos de *credit scoring* surgiram entre 1940 e 1950 orientados por métodos de discriminação produzidos por Fisher (1936). Outros nomes contribuiram com a temática: Henry Markowitz (1952), Fisher Black e Myron Scholes (1973). Em 1984, diretores da *Citicorp* lançaram o livro *Riscos e Recompensa: O Negócio de Crédito ao Consumidor*, em que os modelos de *credit scoring* são mencionados pela primeira vez. Ao longo dos anos, os modelos estatísticos foram sendo aperfeiçoados e as instituições financeiras passaram a utilizar os modelos como auxílio na tomada de decisões de supervisores, gerentes de bancos de investimentos, fundos, etc (Louzada Neto et. al., 2011).

A utilização de um modelo para cálculo de escores é uma das etapas utilizadas pelas empresas concessoras de crédito. Além disso, para operacionalizar o sistema de concessão de créditos são necessários: uma política de crédito bem definida, política de cobranças, formas de pagamento, sistema de informações gerenciais com os dados do cliente, etc (Sicsu, 2010).

Existem duas formas de examinar uma solicitação de crédito: a forma objetiva e

a subjetiva. A primeira envolve metodologia quantitativa, enquanto a segunda envolve a experiência do analista. Entre as vantagens da primeira abordagem estão as seguintes: i) é possível que, de acordo com diferentes perspectivas pessoais dos analistas, tenha-se diferentes opiniões sobre a concessão de crédito, o que não ocorre se forem aplicados modelos de *credit scoring*, uma vez que, no caso de não alteração das características do cliente e da operação, o escore será sempre o mesmo, independente do analista; ii) pode-se tomar decisões de maneira rápida e segura, pois, a tecnologia fornece recursos computacionais que permitem com que, ao serem inseridos os dados dos clientes, as respostas sejam obtidas quase que instantaneamente; iii) não há a necessidade da presença de um analista em cada filial da empresa, ou seja, o vendedor pode inserir os dados do cliente no sistema e receberá a decisão da concessão no computador e a passará ao cliente; iv) interrelações entre covariáveis são consideradas; v) menos erros cometidos; vi) inclusão de novas variáveis para apoiar a tomada de decisão; vii) menor custo (Sicsu, 2010).

Em comparação à decisão subjetiva, o uso de modelo *credit scoring* exige menos informação para decidir se o empréstimo será concedido ou não, pois estes modelos incluem apenas as covariáveis correlacionadas (significativamente) com a variável resposta, enquanto na decisão subjetiva isto não é realizado e não há redução no número de covariáveis. Além disso, os modelos de *credit scoring* consideram as covariáveis que representam as características de bons e maus pagadores e não somente os atributos dos maus pagadores, como é utilizado, geralmente, na decisão subjetiva. Outro fator a se considerar é que os modelos podem ser programados para utilizar somente covariáveis permitidas legalmente, enquanto não se pode ter certeza disto quando a decisão é subjetiva. Exemplos de covariáveis que não podem ser utilizadas para efetuação da análise de crédito seriam cor da pele e religião.

Apesar do grande avanço que a implementação dos modelos de *credit scoring* causou, alguns autores fazem críticas no sentido de que estes modelos, em sua maioria, usam somente informações dos clientes e não incluem fatores econômicos. Clientes podem ser classificados por um modelo como bons pagadores e ter características mais próximas dos maus pagadores. Se este cliente for realmente mau pagador, o que constitui erro de classificação, isto acarretará prejuízo para a instituição financeira. Outra consideração é que os modelos de *credit scoring* não são padronizados e variam de um mercado para outro e treinar um profissional para desenvolvimento destes modelos pode custar caro.

Modelos *credit scoring* são utilizados por 97% dos bancos para decidir se devem conceder ou não cartão de crédito ao solicitante e as instituições financeiras estão constantemente desenvolvendo novos modelos de *credit scoring* para evitar grandes perdas (Adou, 2011). Com isso, o uso de modelos de *credit scoring* constitui uma das mais importantes técnicas em bancos e instiuições financeiras, reduzindo o custo e o risco esperado na realização do empréstimo, economizando tempo e esforço. No entanto, quando se trata de grandes quantias é interessante que dois ou mais analistas experientes sejam consultados, pois a situação exige um nível de detalhamento e cuidado elevado (Sicsu, 2010).

Os modelos *credit scoring* são amplamente utilizados e podem ser aperfeiçoados em termos de aumento na capacidade preditiva. No desenvolvimento destes modelos geralmente são utilizados somente os indivíduos que tiveram solicitação de crédito aprovada, denotados por aceitos. Tal ação causa um viés amostral, uma vez que somente os clientes aceitos não representam satisfatoriamente a população alvo, que inclui os indivíduos aceitos e os indivíduos rejeitados, que tiveram o pedido de crédito negado (Rocha, 2012).

1.1 Justificativa

Com o aumento da demanda por crédito, nos últimos trinta anos, as decisões envolvendo concessões de crédito vem sendo realizadas de maneira objetiva, ou seja, deixaram de ser fundamentadas na experiência do analista de crédito e passaram a utilizar técnicas estatísticas das mais simples as mais sofisticadas.

A utilização de técnicas estatísticas auxilia a resolução de um dos maiores problemas de uma instituição financeira, que é classificar, adequadamente, o cliente como mau ou bom pagador, pois um erro de classificação significa prejuízo para instituição. Se um bom pagador for classificado como mau pagador, o banco deixa de emprestar o dinheiro e não tem lucro; se ocorrer o inverso, ou seja, se o indivíduo mau pagador for rotulado como bom pagador, a instituição emprestará determinada quantia e terá prejuízo pelo seu não reembolso.

Entre os métodos utilizados em *credit scoring* encontram-se a regressão logística, *probit* e a árvore de decisão/classificação e, cada um deles, individualmente, tem seus benífícios e suas limitações. A regressão logística e *probit* são métodos estatísticos bastante utilizados em modelagem *credit scoring*, que fornecem a probabilidade do indivíduo ser bom ou mau pagador em função das covariáveis consideradas na modelagem. A árvore de decisão/classificação consiste, essencialmente, numa hierarquia de testes a algumas das varíaveis envolvidas no problema de decisão/classificação (Vieira et. al., 2005).

Um modelo de *credit scoring* é construído tendo como base clientes utilizados anteriormente pelo credor, isto é, todos os aprovados (aceitos), os quais podem ser classificados como bons ou maus pagadores. Porém, ao utilizar somente indivíduos aceitos, tem-se o problema do viés causado pela não inclusão de muitos indivíduos que não tiveram as suas solicitações de crédito atendidas, denotados por rejeitados, e que poderiam ser, da mesma maneira, bons ou maus pagadores. Desta forma, as amostras usuais não representam devidamente a população de interesse ou mercado potencial, pois há presença de um viés amostral que pode influenciar na classificação do cliente como bom ou mau pagador.

A inferência dos rejeitados é uma alternativa para solucionar este problema, uma vez que, dessa forma, os indivíduos rejeitados são utilizados na modelagem *credit scoring*. Os métodos existentes relacionados à inferência dos rejeitados baseiam-se em suposições sobre o comportamento dos rejeitados, ou seja, através destes métodos busca-se inferir

qual seria o comportamento do solicitante rejeitado se tivesse o pedido de crédito aprovado. A informação fornecida pelo rejeitado, em geral, contribui para a construção de modelo de maior capacidade preditiva. O escore obtido com a inclusão dos rejeitados na formulação do modelo leva em consideração mais informações do que aquele obtido observando apenas os indíviduos aceitos (Rocha, 2012).

Muitas vezes, a utilização da inferência dos rejeitados e de uma única técnica de modelagem de dados não são suficientes para obter resultados satisfatórios. Dessa forma, a fim de melhorar o desempenho preditivo da modelagem, considerou-se a aplicação dos modelos de regressão logística, *probit*, árvore de decisão/classificação e a combinação dos resultados destes modelos para gerar novas regras para concessão de empréstimo. A verdadeira, porém desconhecida condição do indivíduo rejeitado foi gerada aleatoriamente de uma distribuição *Bernoulli* considerando a prevalência de inadimplentes e as medidas de sensibilidade e especificidade, dos três modelos para calcular a probabilidade de cada indivíduo rejeitado ser mau pagador. O objetivo foi avaliar se os resultados combinados de acordo com específicas regras melhoram ou não a capacidade preditiva da modelagem.

1.2 Organização dos capítulos

O texto está organizado da seguinte forma. No Capítulo 2 são abordados os métodos de modelagem de dados comumente utilizados em *credit scoring* como a regressão logística, *probit* e árvore de decisão/classificação. No Capítulo 3 são descritas as medidas preditivas utilizadas e a curva ROC. No Capítulo 4 é abordada a inferência dos rejeitados. No Capítulo 5 tem-se as regras de decisão. No Cápítulo 6, o exemplo de aplicação e no Capítulo 7, as considerações finais são apresentadas. A descrição das variáveis, análise descritiva e os modelos estão no Apêndice.

Capítulo 2

Métodos utilizados em *credit scoring*

Neste capítulo são apresentados os três métodos de modelagem de dados de *credit scoring* utilizados nesta dissertação. Nestes casos, a variável resposta é dicotômica (1 - mau pagador; 0 - bom pagador). Na Seção 2.1 é apresentada a regressão logística. Na Seção 2.2 o *probit* e na Seção 2.3 a árvore de decisão/classificação.

2.1 Regressão logística

O objetivo dos modelos de regressão é descrever a associação existente entre a variável resposta, Y, e a(s) covariável(is) X('s). Na regressão linear, a variável aleatória Y tem natureza contínua. No entanto, muitas vezes, a variável resposta é qualitativa e é expressa por duas ou mais categorias. Tem-se a regressão logística simples, quando a variável resposta tem natureza binária ou dicotômica e existe somente uma covariável envolvida; tem-se a regressão logística múltipla quando, na mesma situação, está envolvida mais de uma covariável. Quando existe uma ordem natural entre as possíveis categorias da variável resposta tem-se a regressão logística ordinal.

Regressão logística simples

O método da regressão logística é conhecido desde os anos 50, porém, tornou-se mais utilizada a partir da década de 80, com Cox e Snell (1989). O objetivo dos modelos de regressão é descrever a associação existente entre a variável resposta, Y, e a covariável X. Na regressão logística simples tem-se apenas uma covariável e a variável resposta é dicotômica, isto é, atribui-se, convencionalmente, o valor 0 a Y, para a não ocorrência do evento de interesse (*fracasso*) e 1 para ocorrência do evento de interesse (*sucesso*) e as probabilidade associadas são $1 - \pi_i = P(Y = 0|X = x_i)$ e $\pi_i = P(Y = 1|X = x_i)$, respectivamente.

Uma das vantagens do uso da regressão logística, aplicada em muitas áreas do conhecimento, é não necessitar atender alguns pressupostos, como normalidade dos erros e igualdade de matrizes de covariância. O modelo final a ser escolhido é aquele que apresenta as melhores medidas de capacidade preditiva e que seja mais razoável de ser interpretado.

Considere n observações de uma variável aleatória $(Y_1, Y_2, ..., Y_n)$ independente com distribuição de Bernoulli, com π_i como probabilidade de sucesso e x_i^T a i-ésima linha da matriz $\mathbf{X}$, $i = 1, 2, 3, ..., n$. Define-se π e $1 - \pi$ da seguinte forma:

$$\pi_i = \pi(x_i) = P(Y = 1|X = x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} \tag{2.1}$$

$$1 - \pi_i = 1 - \pi(x) = P(Y = 0|X = x_i) = \frac{1}{1 + exp(\beta_0 + \beta_1 x_i)} \tag{2.2}$$

em que β_0 e β_1 são parâmetros desconhecidos, que representam o intercepto e o coeficiente angular, respectivamente.

Considerando a regressão linear modela-se a média condicional $E(Y|X = xi)$, que por sua vez está no seguinte intervalo $-\infty < E(Y|X = xi) < +\infty$. No caso da regressão logística tem-se que pela natureza da variável resposta $0 \leq E(Y|X = x_i) \leq 1$ e por definição:

$$E(Y|X = x_i) = 1P(Y_i = 1|X = x_i) + 0P(Y_i = 0|X = x_i) = \pi_i \tag{2.3}$$

Dado x_i, o valor da variável resposta é expresso por: $Y_i = \pi + \epsilon_i$, em que ϵ_i é denominado erro aleatório. No caso da regressão logística é útil a seguinte transformação:

$$ln\left\{\frac{\pi_i}{1 - \pi_i}\right\} = \beta_0 + \beta_1 x_i \tag{2.4}$$

utilizada, por exemplo, para analisar a associação entre o indivíduo ser mau pagador e a presença ou não de determinada característica x_i.

Na estimação dos parâmetros, o método comumente utilizado é o da máxima verossimilhança (Souza, 2006). Para isto, considera-se uma amostra independente com n observações, em que y_i representa o valor da variável resposta e x_i o valor da covariável da i-ésima observação, $i = 1, 2, ..., n$. Como $Y_i \sim$ Bernoulli(π_i) tem-se:

$$\prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^y (1-\pi_i)^{1-y_i}, 0 \leq y_i \leq 1. \tag{2.5}$$

Logo, para estimar o valor de β que maximiza a função de verossimilhança, aplica-se o logaritmo da função de verossimilhança dada por:

$$\begin{aligned} l(\beta) &= ln(L(\beta)) = \mathbf{ln}\left[\prod_{\mathbf{i=1}}^{\mathbf{n}} \pi_{\mathbf{i}}^{\mathbf{y_i}} (\mathbf{1}-\pi_{\mathbf{i}})^{\mathbf{1-y_i}}\right] \\ &= \sum_{i=1}^{n} y_i ln(\pi_i) + ln(1-\pi_i) - y_i ln(1-\pi_i) \\ &= \sum_{i=1}^{n} \left[y_i ln\left(\frac{\pi_i}{1-\pi_i}\right) + ln(1-\pi_i)\right] \\ &= \sum_{i=1}^{n} \left[y_i(\beta_0+\beta_1 x_i) - ln(1+exp(\beta_0+\beta_1 x_i))\right], \end{aligned} \tag{2.6}$$

e deriva-se em relação aos $\beta' s$. Dessa forma:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left[y_i - \frac{1}{1+exp(\beta_0+\beta_1 x_i)exp(\beta_0+\beta_1 x_i)}\right], \tag{2.7}$$

$$\frac{\partial l(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} \left[y_i x_i - \frac{1}{1+exp(\beta_0+\beta_1 x_i)exp(\beta_0+\beta_1 x_i)x_i}\right], \tag{2.8}$$

igualando (2.7) e (2.8) a zero tem-se:

$$\sum_{i=1}^{n} (y_i - \pi_i) = 0, \tag{2.9}$$

$$\sum_{i=1}^{n} x_i (y_i - \pi_i) = 0, \tag{2.10}$$

em que

$$\pi_i = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}, i = 1, 2, ..., n. \tag{2.11}$$

Para resolver (2.9) e (2.10) utilizam-se de procedimentos numéricos, pois estas equações são não lineares nos parâmetros (Souza, 2006).

Regressão logística múltipla

Generalizado por Hosmer e Lemeshow (1989), a regressão logística múltipla envolve uma variável dicotômica Y e mais de uma covariável X. Seja um conjunto com p covariáveis independentes $\boldsymbol{x_i}^{\boldsymbol{T}}=(x_{i0}, x_{i1}, ..., x_{ip})$, o vetor da i-ésima linha da matrix $(\mathbf{X})$ das covariáveis, em que x_{ij} corresponde ao ij-ésimo elemento da matriz $\mathbf{X}$, $i = 1, 2, ..., n$, $j = 0, 1, ..., p$, $x_{i0} = 1$. Seja $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$, o vetor de parâmetros desconhecidos e β_j o j-ésimo parâmetro associado a x_j.

Neste modelo, considera-se $Y_i \sim$ Bernoulli(π_i), em que π_i é a probabilidade de sucesso. Denota-se π_i da seguinte forma:

$$\pi_i = \pi(\boldsymbol{x_i}) = P(Y_i = 1|\boldsymbol{X} = \boldsymbol{x_i}) = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}. \tag{2.12}$$

Pode-se aplicar a mesma transformação da regressão logística simples na múltipla. Assim tem-se o *logit*:

$$logit(\pi_i) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x_i}^T\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{2.13}$$

Estimação pontual

A estimação dos parâmetros é dada de maneira análoga ao caso da regressão logística simples, ou seja, utilizando o método da máxima verossimilhança. Obtendo o logaritmo da função de verossimilhança $l(\beta)$ e derivando-o tem-se a seguinte expressão:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta})}{\beta_j} &= \sum_{i=1}^{n} \left[y_i x_{ij} - \frac{exp(x_i^T\boldsymbol{\beta})}{1 + exp(\beta_0 + x_i^T\boldsymbol{\beta})} x_{ij} \right] \\ &= \sum_{i=1}^{n} \left[y_i - \pi_i \right] x_{ij}. \end{aligned} \tag{2.14}$$

Para resolver esta equação recorre-se a utilização do método númerico. Um método bastante utilizado é o de Newton-Raphson (Souza, 2006), que expande a função $\boldsymbol{U(\beta)}$ em torno do ponto inicial $\beta^{(0)}$. Dessa forma $\boldsymbol{U(\beta)}$ e a matriz de informação de Fisher $\boldsymbol{I(\beta)}$ são dados por:

$$\boldsymbol{U(\beta) = X^T y - X^T \pi = X^T (y - \pi)}, \tag{2.15}$$

$$\boldsymbol{I}(\boldsymbol{\beta}) = E\left[-\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\boldsymbol{T}}}\right] = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{X}, \tag{2.16}$$

em que $\boldsymbol{Q} = \boldsymbol{diag}[\boldsymbol{\pi}_i(\mathbf{1} - \boldsymbol{\pi}_i)]$, $\mathbf{X}$ a matriz dos dados e $[\boldsymbol{I}(\boldsymbol{\beta})]^{-\mathbf{1}}$ a matriz de variâncias e covariâncias das estimativas de máxima verossimilhança dos parâmetros (SOUZA, E. 2006).

Para que seja obtida a resolução das equações de verososimilhança utiliza-se as seguintes equações iterativas:

$$\boldsymbol{\beta}^{\boldsymbol{t+1}} = \boldsymbol{\beta}^{\boldsymbol{t}}[\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Q}^{\boldsymbol{t}}\boldsymbol{X}]^{-1}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{\pi}^{\boldsymbol{t}}), t = 0, 1, 2, ..., \tag{2.17}$$

em que $\boldsymbol{\beta}^{\boldsymbol{t}}$ e $\boldsymbol{\beta}^{\boldsymbol{t+1}}$ são vetores dos parâmetros desconhecidos estimados nos passos t e $t+1$. O procedimento adotado no método iterativo é o seguinte: primeiramente atribui-se um valor inicial considerando todos os coeficientes iguais a zero; após isso, substitui-se na equação imediamente acima e repete-se o processo até que um critério de convergência seja atendido. Como exemplo de critério de parada tem-se $|\frac{\boldsymbol{\beta}^{\boldsymbol{t+1}} - \boldsymbol{\beta}^{\boldsymbol{t}}}{\boldsymbol{\beta}^{\boldsymbol{t}}}| < \epsilon$, em que ϵ é um número suficientemente pequeno.

A distribuição assintótica dos $\boldsymbol{\beta}$ é dada da seguinte forma:

$$\hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \boldsymbol{I}(\boldsymbol{\beta})^{-1}\right). \tag{2.18}$$

Inferência

É necessário, para obter o modelo ajustado, testar a significância dos coeficientes de regressão. Suponha a seguinte hipótese nula: H_0: $\boldsymbol{\beta} = \boldsymbol{\beta_0}$ versus H_1: $\boldsymbol{\beta} \neq \boldsymbol{\beta_0}$, em que $\boldsymbol{\beta_0}$ é um vetor especificado para o vetor de parâmetros $\boldsymbol{\beta}$. Uma das formas de testá-las é através da estatística da razão de verossimilhança, da estatística Wald e da estatística escore, que seguem assintoticamente uma distribuição qui-quadrado com p graus de liberdade. Para hipóteses relacionadas a um único coeficiente de regressão a estatística Wald é mais utilizada, enquanto que para mais de um coeficiente, a estatística de razão de verossimilhança é a mais indicada (Demetrio, 2007). A estatística de razão de verossilhança é dada por:

$$\Lambda = -2ln\left(l(\boldsymbol{\beta_0})/l(\hat{\boldsymbol{\beta}})\right) = -2ln\left(l(\hat{\boldsymbol{\beta}} - l(\boldsymbol{\beta_0}))\right), \tag{2.19}$$

em que $\hat{\boldsymbol{\beta}}$ é o estimador de máxima verossimilhança sob todo o espaço paramétrico.

A estatística Wald é dada por:

$$W = \left((\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta_0}\right)^T I(\hat{\boldsymbol{\beta}}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right), \tag{2.20}$$

em que $\boldsymbol{I(\hat{\beta})}$ é a matriz de informação de Fisher avaliada em $\hat{\boldsymbol{\beta}}$.

Enquanto a estatística escore é dada por:

$$Es = \boldsymbol{U^T(\beta_0) I(\beta_0)^{-1}(U(\beta_0))}, \tag{2.21}$$

em que $\boldsymbol{I(\beta_0)}$ consiste na matriz de informação de Fisher avaliada em $\boldsymbol{\beta_0}$.

Estimação intervalar

Assim como obter estimativas pontuais, é importante calcular os intervalos de confiança dos parâmetros do modelo. A base para construção destes intervalos é a mesma utilizada na formulação dos testes de significância dos parâmetros do modelo. Através da estatística Wald obtém-se os intervalos de confiança. Os limites para o intervalo $100(1-\alpha)\%$ para o intercepto é dado por

$$\hat{\beta}_0 \pm z_{(1-\alpha/2)} \hat{DP}(\hat{\beta}_0), \tag{2.22}$$

e para os demais coeficientes, $(i = 1, 2, ..., p)$

$$\hat{\beta}_i \pm z_{(1-\alpha/2)} \hat{DP}(\hat{\beta}_i), \tag{2.23}$$

em que $\hat{DP}$ é o desvio padrão estimado da estimativa do parâmetro e $z_{(1-\alpha/2)}$ é o quantil $100(1-\alpha)\%$ da distribuição normal padrão. Para obter as estimativas dos desvios padrão utiliza-se a matriz das segundas derivadas parciais do logaritmo da função de verossimilhança, denotada por matriz Hessiana. A forma geral dessas derivadas é a dada por:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_i^2} = -\sum_{k=1}^{n} x_{ki}^2 \pi_k (1-\pi_k), \tag{2.24}$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_i \beta_j} = -\sum_{k=1}^{n} x_{ki} x_{kj} \pi_k (1-\pi_k), \tag{2.25}$$

em que $i, j = 1, 2, ..., p$. $\boldsymbol{I(\beta)}$, é a matriz de informação de Fisher, de dimensão $(p \times 1)(p \times 1)$, que contém os termos negativos das equações acima. A matriz de variâncias e

covariâncias é dada por:

$$\boldsymbol{\Sigma(\beta) = I^{-1}(\beta)}, \tag{2.26}$$

em que na diagonal principal tem-se as variâncias e fora da diagonal principal tem-se as covariâncias. O desvio padrão é dado por: $\hat{DP}(\hat{\beta}_j) = \left[\hat{Var}(\hat{\beta}_j)\right]^{1/2}$. Se o intervalo de confiança para o parâmetro contiver o valor zero então significa que pode-se afirmar com $100(1-\alpha)\%$ de confiança que há evidências de que a covariável não é importante para explicar a variação na variável resposta.

É possível encontrar vários *softwares* estatísticos em que a técnica de regressão logística está disponível tanto para obtenção das estimativas dos coeficientes de regressão quanto para medidas de adequação do modelo, intervalo de confiança para os parâmetros, predição, entre outros.

Adequação do modelo

Com o objetivo de obter o melhor modelo que ajusta os dados, faz-se necessário analisar a adequaçaõ do modelo. Uma forma de realizar este procedimento se dá através da estatística *deviance* D e da estatística χ^2 de Pearson (Souza, 2006). Através do uso da *deviance*, proposta por Nelder e Wedderburn (1972), o processo a ser realizado é o seguinte: se existe um conjunto com n observações, então um modelo com até n parâmetros pode ser ajustado, sendo denotado por modelo saturado.

Por outro lado, o modelo mais simples, com somente β_0, denotado por modelo nulo, indica que toda a variação está relacionada ao componente aleatório. Como um modelo nulo é muito simples e o modelo saturado não resume os dados, ou seja, há a reprodução dos dados; então o modelo saturado pode ser usado para medir a distância de um modelo de p parâmetros no que diz respeito ao adequado ajuste do modelo aos dados.

Seja o valor da verossimilhança do modelo proposto com $p+1$ parâmetros $L(\beta_0, \beta_1, ..., \beta_p)$ e o seu valor no modelo saturado $L(y_1, y_2, ..., y_n)$. Então estes valores são comparados com o conveniente uso de menos duas vezes o logaritmo do quociente destes valores. Logo, a *deviance* é dada por:

$$D = -2ln\left[\frac{L(\beta_0, \beta_1, ..., \beta_p)}{L(y_1, y_2, ..., y_n)}\right]. \tag{2.27}$$

É possível escrever a *deviance* da seguinte forma:

$$\begin{aligned} D &= -2\sum_{i=1}^{n}(y_i - ln(\hat{\pi}) + (1-y_i)ln(1-\hat{\pi}_i) - y_i ln(y_i) + (1-y_i)ln(1-y_i)) \\ &= 2\sum_{i=1}^{n}\left(y_i ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right), \end{aligned} \tag{2.28}$$

O modelo ajusta melhor os dados à medida que o valor da *deviance* diminui, isto é, o modelo que ajusta melhor os dados será aquele com o menor valor da *deviance*. Para um conjunto de dados, considerando réplicas, em que para cada x_k existem m_k elementos amostrais, com $k = 1, 2, ..., K$ e $K \leq n$, em que $\sum_{k=1}^{K} m_k = n$ (Farhart, 2003). Seja a probabilidade estimada de um evento dada por:

$$\hat{\pi_k} = m_k \left[\frac{exp(g(\hat{x_k}))}{1 + exp(g(\hat{x_k}))}\right], \tag{2.29}$$

em que $g(\hat{x_k})$ é o *logit* estimado. E o número de eventos relacionado a covariável x_k é dado da seguinte forma:

$$\hat{y_k} = m_k\hat{\pi_k} = m_k \left[\frac{exp(g(\hat{x_k}))}{1 + exp(g(\hat{x_k}))}\right]. \tag{2.30}$$

Dessa forma, tem-se a componente *deviance* definida por:

$$d(y_k, \hat{\pi_k}) = \pm\left\{2\left[y_k ln\left(\frac{y_k}{m_k\hat{\pi_k}}\right) + (m_k - y_k)ln\left(\frac{(m_k - y_k)}{m_k(1-\hat{\pi_k})}\right)\right]\right\}^{1/2}, \tag{2.31}$$

em que o sinal da *deviance* é o mesmo de $(y_k - m_k\hat{\pi_k})$. A estatística *deviance* fundamentada no resíduo deviance é definida por:

$$D = \sum_{k=1}^{n} d(y_k, \hat{\pi_k})^2. \tag{2.32}$$

A distribuição assintótica da *deviance* é qui-quadrado com $(n-p)$ graus de liberdade em que p é o número de parâmetros estimados do modelo (Demetrio, 2007).

Outra forma de verificar a adequação do modelo é realizada através do uso do

resíduo de Pearson, que pode ser definido da seguinte forma:

$$\begin{aligned} rp(y_k, \pi_k) &= \frac{y_k - \hat{y_k}}{\sqrt{Var(\hat{Y}_k|x_k)}} \\ &= \frac{y_k - m_k\hat{\pi}_k}{\sqrt{m_k\hat{\pi}_k(1-\hat{\pi}_k)}}. \end{aligned} \tag{2.33}$$

De forma análoga, a estatística χ^2 de Pearson é definida da seguinte forma:

$$X^2 = \sum_{k=1}^{n} rp(y_k, \hat{\pi})^2. \tag{2.34}$$

A distribuição assintótica da estatística χ^2 de Pearson é a qui-quadrado com $(n-p)$ graus de liberdade (Demetrio, 2007).

2.2 Regressão *probit*

Outra abordagem muito utilizada em modelos de regressão quando se tem uma variável resposta de natuzera binária ou dicotômica é a regressão *probit*. No entanto, antes de iniciar a discussão sobre a regressão *probit*, convém comentar brevemente sobre modelos lineares generalizados.

Na especificação de um modelo linear generalizado são necessárias três componentes: uma componente aleatória, que identifica à distribuição de probabilidade da variável resposta (pertencente à família exponencial); uma componente sistemática, que determina uma função linear entre as covariáveis e, uma função de ligação, que descreve a relação matemática entre a componente sistemática e o valor esperado da componente aleatória. Esse modelo envolve de uma variável resposta univariada, covariáveis e uma amostra aleatória de n observações independentes (Demetrio, 2007).

A componente aleatória do modelo linear generalizado é composta pelas observações da variável resposta, $\boldsymbol{y} = (y_1, y_2, ..., y_n)$, a componente sistemática é denotada pelo vetor $\eta = (\eta_1, \eta_2, ..., \eta_n)$, o qual está associado as covariáveis através de um função linear $\eta = \boldsymbol{X}\boldsymbol{\beta}$, em que $\boldsymbol{X}$ é uma matriz com as n observações das covariáveis e $\boldsymbol{\beta}$ é um vetor de parâmetros do modelo.

A última componente é a função de ligação entre a componente aleatória e sistemática. O papel da função de ligação é conectar os valores esperados das observações às

covariáveis, para $i \in 1, ..., n$ através da fórmula

$$g(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij}, \tag{2.35}$$

em que $\mu_i = E(Y_i|x_i)$, com $i \in 1, ..., n$, η_i é definida por $\eta_i = g(\mu_i)$, em que g é uma função monotônica e diferenciável.

Se a função de ligação g for a função identidade, tem-se o modelo de regressão linear, se a função de ligação é a logito tem-se a regressão logística, se esta função é a *probit* tem-se a regressão *probit*. Ou seja, o modelo logito e *probit* são casos particulares dos modelos lineares generalizados. Na regressão *probit* tem-se π como proporção de sucessos de uma distribuição binomial. A função de ligação *probit* definida por:

$$probit(\pi_i) = \Phi^{-1}(\pi_i) = x_i^T \beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{2.36}$$

em que $\Phi(.)$ é a função de distribuição acumulada da normal padrão. A estimação dos coeficientes do modelo *probit* é realizada, geralmente, através do método da máxima verossimilhança assim como ocorre na regressão logística e os estimadores de máxima verossimilhança dos parâmetros do modelo são aqueles que maximizam a função de verossimilhança, $l(\beta)$.

Devido a dificuldade do cálculo das estimativas no modelo *probit*, era comum a utilização da análise *logit* devido a simplicidade da expressão analítica da sua função de distribuição e a facilidade de cálculo na fase de estimação; porém, com os avanços computacionais, isto não é mais verificado, uma vez que é possível obter as estimativas para ambos os modelos através de *softwares* estatísticos como SAS, R, entre outros. De forma análoga ao modelo de regressão logística é possível testar a significância dos coeficientes do modelo e a adequação do modelo.

Em situações em que as proporções de sucessos e fracassos são próximas, não se espera grandes diferenças entre os resultados entre o modelo *logit* e o *probit*; no caso de dados desbalanceados, ou seja, com elevado número de observações $0's$ ou $1's$, existem modelos mais flexíveis como a distribuição de valor extremo generalizada (Wang e Dey, 2010).

Processo de Seleção de Variáveis

Com o objetivo de selecionar covariáveis para um modelo estatístico, qualquer procedimento utiliza algoritmos que atuam incluindo ou excluindo covariáveis, de acordo com a importância delas para o modelo. Uma covariável é importante se o coeficiente associado a ela tem significância estatística. Entre os métodos mais utilizados estão *stepwise*, *backward* e *forward*, sendo descrito, nessa dissertação, o método *stepwise* por ter sido utilizado na aplicação prática do modelo de regressão logística e *probit*.

As covariáveis que foram incluídas no modelo, através do procedimento *stepwise*, podem ser retiradas na presença de outras covariáveis. Na regressão linear é utilizado um teste F, caso os erros sejam provenientes de uma distribuição normal; no caso da regressão logística, a significância é assegurada via teste da razão de verossimilhança (TRV). Assim, em cada passo do algoritmo, a covariável mais relevante é aquela que produzir maior mudança no logaritmo da função de verossimilhança comparando com o modelo que não possui a covariável.

Suponha que existam k covariáveis a se considerar em um modelo. O primeiro passo é incluir o intercepto e obter L_0, o logaritmo da função de verossimilhança deste modelo. Após isto, ajusta-se k modelos com apenas uma covariável e obtém-se $L_j^{(0)}$, o logaritmo da função de verossimilhança com a covariável X_j. A estatística do teste da razão de verossimilhança do modelo com a covariável X_j versus o modelo com somente o intercepto é dada por $TRV_j^{(0)} = -2(L_0 - L_j^{(0)})$ e o p-valor é dado por $p_j^{(0)} = P(\chi_v^2 > TRV_j^{(0)})$, $v = 1$ se X_j é contínua e $v = c - 1$, se X_j possui c categorias. Se, por exemplo, $p_{e1}^{(0)} = min(p_j^{(0)}) < \alpha_e$, então, inclui-se X_1 no modelo.

A seguir utiliza-se procedimento análogo, ajustando o modelo com X_1 e obtendo $L_{e1}^{(1)}$. Deseja-se, neste momento, avaliar a relevância de $k - 1$ covariáveis e, desta forma, ajusta-se $k - 1$ modelos com X_1 e X_j. Obtém-se $L_{e1;j}^{(1)}$ e $TRV_j^{(1)}$. Se, por exemplo, $p_{e2}^{(2)} = min(p_j^{(2)}) < \alpha_e$, então X_2 é incluída no modelo, caso contrário, somente X_1 é incluída.

O próximo passo envolve o ajuste de X_1 e X_2, porém, pode ocorrer que X_1 passe a ser não relevante na presença de X_2. Então, ajusta-se o modelo com X_1 e X_2. Calcula-se $L_{-ej}^{(2)}$, ou seja, o logaritmo da função de verossimilhança com a covariável X_j removida e calcula-se TRV utilizando $L_{-ej}^{(2)}$ e $L_{e1;e2}$, ou seja, $TRV_{-ej}^{(2)} = -2(L_{-ej}^2 - L_{e1;e2}^{(2)})$. Após isso, calcula-se o p-valor, $p_{-ej}^{(2)}$, associado a cada covariável e decide-se pela remoção ou não

da covariável através do uso do maior p-valor. Por exemplo, supondo X_2 tenha maior p-valor, ou seja, $p_{r2}^{(2)} = max(p_{-e1}^{(2)}, p_{-e2}^{(2)})$. Se $p_{r2}^{(2)} > \alpha_r$, então X_2 é removida do modelo. O algoritmo repete o procedimento até que todas as covariáveis no modelo tenham p-valor de entrada maior que α_e e todas as covariáveis tenham p-valor de saída menor que α_r.

Já o *backward* é o procedimento inverso ao *stepwise*. Este método inclui todas as covariáveis no modelo e vai eliminando-as. Enquanto que no *forward*, em suma, ajusta-se o modelo com a covariável que possua maior correlação com a variável resposta e vai-se adicionando uma covariável por vez.

2.3 Árvore de decisão/classificação

Uma árvore de decisão/classificação é a representação de uma estrutura hierárquica de nós internos e externos, e traduz uma árvore invertida que se desenvolve da raiz para as folhas. Para se utilizar a árvore de decisão para a tomada de decisão, começa-se utilizando o nó raiz da árvore, sendo os ramos seguidos até o nó terminal, e, em uma única árvore, pode-se encontrar várias regras de decisão construídas a partir de cada categoria das variáveis estudadas.

2.3.1 Aspectos gerais

O principal objetivo da árvore é dividir. Em cada nível da árvore, um problema mais elaborado é divido em subproblemas menos complexos e assim sucessivamente, desenvolvendo do geral para a particularidade. A partir do nó raiz, um teste lógico é usado pelo nó e a árvore se ramifica para um dos nós filhos. Este procedimento é repetido em cada nó até chegar num nó terminal. Tal repetição dá o caráter recursivo a árvore de decisão/classificação. O espaço de descrição das árvores de decisão/classificação é dividido em regiões disjuntas, ou seja, cada observação é classificada por apenas um único ramo.

É possível identificar algumas vantagens no uso de árvore de decisão. Segundo Rodrigues (2005) elas são:

1. ausência de pressupostos típicos de modelos paramétricos;

2. evitar dispendiosos tratamentos prévios dos dados, pois pode ser utilizada com qualquer número de covariáveis e em várias escalas de medida;

3. as variáveis não necessitam de transformação;

4. possibilidade de integrar relações complexas entre as variáveis explicativas e a variável resposta e não somente relações lineares;

5. interpretabilidade simples, por observação da árvore.

Entre os principais desvantagens estão a instabilidade, pois perturbações na amostra de treinamento podem provocar alterações no modelo aprendido e pode ocorrer replicações de sub-árvores. Após construir a árvore de decisão/classificação, faz-se necessário avaliar seu desempenho. Primeiramente, separa-se uma parte da amostra, geralmente, 70%, para treinamento e 30% para teste. Isto é feito para avaliar como a árvore generaliza os dados e como ela se adapta a novas informações e também como estima a proporção de erros e acertos envolvidos na construção da árvore.

As árvores de decisão são utilizadas em várias áreas do conhecimento como sáude, análise de mercado, engenharia, ciências sociais, e devem ser usadas de acordo com o problema que se pretende solucionar. Estas soluções podem ser a classificação mais eficiente dos dados referentes a determinada população, compreender e escolher as variáveis que desempenham papel relevante na solução de um problema, descobrir padrões nos dados, etc. Existem alguns algoritmos utilizados para a tarefa de classificação e entre eles destacam-se ID3, C4.5, CART, CHAID e o QUEST. Neste trabalho, foi utilizado o algoritmo CART, a ser descrito na subseção 2.3.3.

2.3.2 Um exemplo de árvore de decisão/classificação

O exemplo da Figura (2.1), representa uma árvore de decisão, em que são apresentadas as condições necessárias para que um indivíduo tenha pedido de crédito aceito. Neste exemplo, considera-se a probabilidade do montante do empréstimo ser alto, médio ou baixo. Algumas características são positivas e indicam que o empréstimo deve ser concedido e outras características indicam que o empréstimo não deve ser concedido. Há a possibilidade de se gerar regras a partir do caminho do nó raiz até as folhas e uma das razões para a utilização de regras no local da árvore é o fato de que, muitas vezes, há um crescimento exagerado da árvore, sendo viável substituí-la por regras. No exemplo a seguir, é possível notar a existência de duas regras advindas da árvore. São as seguintes:

- Se montante = *médio* e salário = *baixo* então empréstimo = *não*

- Se montante = *médio* e salário = *alto* então empréstimo = *sim*

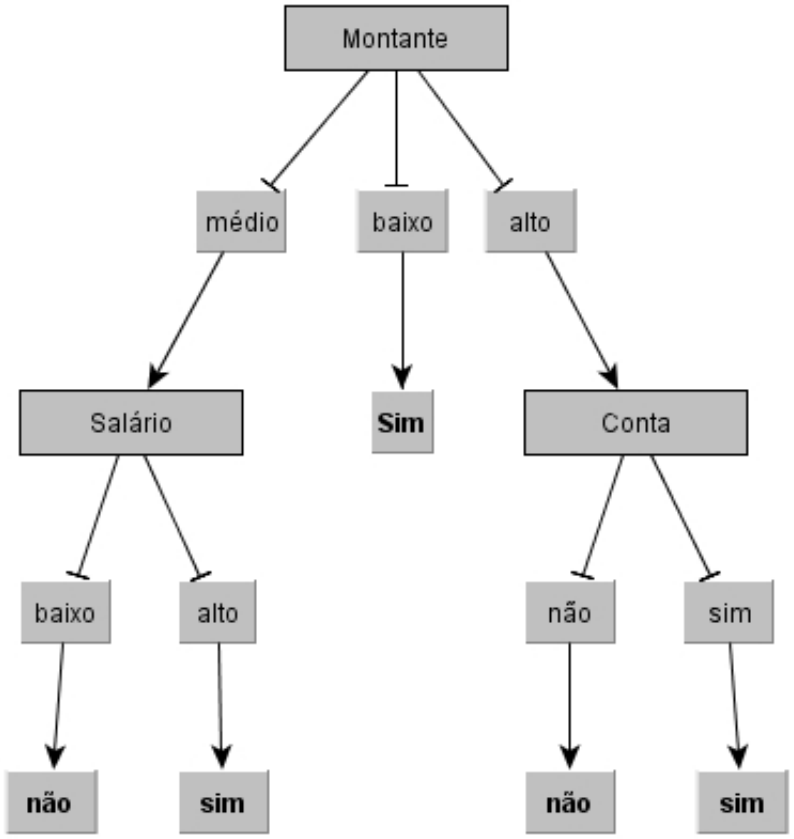

Figura 2.1: Exemplo de árvore de decisão/classificação

2.3.3 Algoritmo CART

A metodologia CART trata da partição recursiva binária dos dados, pois no processo cada nó pai é dividido sempre em dois nós filhos e ocorre a repetição deste processo considerando cada nó filho como nó pai. O algoritmo CART foi proposto por Breiman, et.al. (1984) e significa *Classification and Regression Tree*, sendo definido como um modelo de regressão não-paramétrico usado para relacionar uma única variável resposta com covariáveis.

O objetivo é dividir de forma binária o conjunto de dados até encontrar subconjuntos homogêneos de dados da variável resposta. Para dividir um nó em dois nós filhos, o algoritmo CART sempre questiona de forma a obter respostas do tipo *sim* ou *não*, por exemplo, a idade é maior que 30? A fim de encontrar a melhor partição de dados, o algoritmo busca minimizar as impurezas dos nós folha resultante através de medidas de homogeneidade. Um dos indíces utilizados para selecionar a melhor partição dos dados é o índice de Gini.

O critério de Gini utiliza um índice de dispersão, que é muito utilizado em ciências econômicas e sociais, por exemplo, para quantificar a distribuição de renda de determinado país. Considerando c classes, o $gini_{index}$ é definido por:

$$gini_{index}(no) = 1 - \sum_{i=1}^{c} p(i/no) \tag{2.37}$$

em que $p(i/no)$ é a proporção da classe i no nó. A medida Gini é obtida através do cálculo da diferença entre $gini_{index}$ antes e após a divisão. Logo, a medida Gini é dada por

$$Gini = gini_{index}(pai) - \sum_{j=1}^{n} \left[\frac{N(v_j)}{N} gini_{index}(v_j) \right] \tag{2.38}$$

em que n é o número de nós-filhos, N é o número total de observações do nó-pai e $N(v_j)$ é o número de observações associadas ao nó-filho v_j. No processo de crescimento da árvore, cada variável pode ser usada diversas vezes. A covariável escolhida é aquela que possui maior valor da medida Gini. Quando a melhor divisão é encontrada, o processo é repetido para cada nó filho, até que a divisão seja impossível.

No método CART, a árvore continua crescendo até que não seja mais possível continuar a divisão dos nós, como por exemplo quando se atinge o número mínimo de dados da amostra. Após o encontro de todos os nós terminais, é definida a árvore de tamanho máximo e se inicia o processo de poda. Neste processo elimina-se alguns ramos da árvore de modo a se obter maior poder de generalização. Algumas sub-árvores são testadas e é escolhida aquela que possui menor taxa de erro.

O uso das árvores através do algoritmo CART apresenta vantagens e as principais são as seguintes:

- utilização de covariáveis contínuas, ordinais e nominais.
- utilização da mesma variável em diferentes estágios do modelo, o que permite reconhecer o efeito de covariáveis sobre outras.
- não transformação de covariáveis, pois o método tem bons resultados com qualquer tipo de dado.
- não necessidade de satisfazer qualquer pressuposto como no caso de modelos paramétricos.

Além disso, a análise é efetuada de forma automática e requer intervenção humana mínima. Como desvantagens, as árvores obtidas através do uso do algoritmo CART tem, geralmente, muitos níveis, o que compromete, às vezes, a apresentação dos resultados e para grandes conjuntos de dados, os cálculos podem ser demorados. Com relação a outros algoritmos de árvore de decisão/classificação, alguns problemas devem ser superados como dimensão, erro de classificação, tempo de construção, entre outros.

Capítulo 3

Avaliação de capacidade preditiva do modelo

3.1 Medidas Preditivas

Após obter o modelo ajustado torna-se necessário medir o poder discriminatório do modelo, ou seja, avaliar a capacidade preditiva do modelo. Quando a variável resposta é dicotômica, ou seja, assume os valores zero ou um; a probabilidade da variável assumir o valor um (mau pagador), em geral, multiplicada por 1000, é denotada por escore (S_c) e a partir da definição de um escore limite, é decidido, se o cliente será classificado com bom pagador ou mau pagador.

Então é necessário definir qual o ponto de corte c_0 a ser utilizado. Este ponto de corte pode ser fixado a partir da experiência do analista ou através de alguma técnica. Neste trabalho, o ponto de corte c_0 foi definido com o auxílio da curva ROC.

Na área financeira, defini-se o conjunto de dados e então separa-se em amostra de treinamento e amostra teste. Consta em Hosmer & Lemoshow (1989) a divisão de 70% dos dados para treinamento e 30% para teste. A amostra treinamento é utilizada para construir o modelo e a amostra teste para validar as medidas de desempenho ou de capacidade preditiva. As medidas que avaliam a capacidade preditiva do modelo utilizam a comparação entre as respostas obtidas pelo modelo e as respostas observadas na amostra de teste. As principais métricas utilizadas são:

1. **Sensibilidade** - é a probabilidade de detectar os verdadeiros positivos, $P(T = 1|Y = 1)$, ou seja, a probabilidade do indivíduo ser classificado como positivo dado que ele foi observado como positivo. Neste trabalho, refere-se ao mau pagador como positivo ($Y = 1$) e o indivíduo considerado negativo é referido, neste trabalho, como

bom pagador, $(Y = 0)$.

A sensibilidade avalia a capacidade do modelo em classificar corretamente o indivíduo como mau pagador e, em termos de avaliação de modelo de classificação, é uma das métricas mais úteis, sendo calculada por,

$$SENS = \frac{VP}{VP + FN} \tag{3.1}$$

VP - verdadeiros positivos: é a quantidade de indivíduos classificados como positivos e que são realmente positivos. FN - falsos negativos: é a quantidade de classificados como negativos, mas que realmente são positivos.

2. **Especificidade** - é a probabilidade de detectar os verdadeiros negativos (VN), ou seja, a probabilidade do solicitante ser classificado como negativo (bom pagador) e, realmente, ele foi observado como negativo, $P(T = 0|Y = 0)$. A especificidade é uma das medidas mais importantes para avaliação do desempenho de um modelo de classificação e avalia a capacidade do modelo em classificar corretamente o indivíduo como bom pagador e é dada por,

$$ESPEC = \frac{VN}{VN + FP} \tag{3.2}$$

VN - verdadeiros negativos: é a quantidade de indivíduos classificados como negativos e que realmente são negativos. FP - falsos positivos: é a quantidade de indíviduos classificados como positivos, mas que são realmente negativos.

3. **Valor Preditivo Positivo** - é a probabilidade do indivíduo ser realmente mau pagador $(Y = 1)$ dado que o modelo o classificou como mau pagador $(T = 1)$. Esta medida é importante, pois considera a quantidade de verdadeiros positivos em relação a todos os indivíduos classificados pelo modelo como positivos.

$$VPP = \frac{VP}{VP + FP} \tag{3.3}$$

4. **Valor Preditivo Negativo** - é a probabilidade do indivíduo ser observado como bom pagador ($Y = 0$) uma vez que o modelo o classificou como bom pagador ($T = 0$). Esta métrica considera a quantidade de verdadeiros negativos em relação a todos os indivíduos classificados pelo modelo como negativos.

$$VPN = \frac{VN}{VN + FN} \tag{3.4}$$

5. **Acurácia** - é o acerto total, que não considera os positivos nem os negativos, ou seja, a probabilidade do modelo acertar em suas predições no geral.

$$ACC = \frac{VP + VN}{VP + VN + FN + FP} \tag{3.5}$$

Todas as medidas citadas variam entre zero e um. Quanto mais próximo de 1 mais confiável é o modelo, isto é, tem melhor desempenho. Porém, é necessário cautela ao utilizar algumas métricas, por exemplo, se a amostra fornece um estimador viesado da prevalência, os valores preditivos, tanto positivos quanto negativos, estimados por essas relações, não são tão confiáveis. Por exemplo, num cenário em que tem-se uma prevalência baixa, isso contribui para VPP baixo, mesmo quando a sensibilidade e especificidade são altas (Pereira, G. 2011). Além dos já citados, existem outros indicadores de poder discriminatório do modelo como KS (índice de Kolmogorov-Smirnov); CAP (Perfil de eficiência acumulada); D de Sommers; o coeficiente de Gini (Sicsu, 2010). Após obter o ponto de corte ótimo com o auxílio da curva ROC, as medidas de sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia foram calculadas para avaliação da capacidade preditiva do modelo.

3.2 Curva ROC

3.2.1 Conceitos básicos

A curva ROC pode ser definida como um método gráfico que quantitativamente avalia, organiza e seleciona testes ou modelos de predição. Esta curva foi utilizada primeiramente na detecção de sinais eletrônicos e problemas com radares durante a Segunda Guerra Mundial com o objetivo de quantificar a habilidade dos *receiver operators* em distinguir um sinal de ruído (Zweing & Campbell, 1993). Tal habilidade era denotada por *receiver operating characteristic* (*ROC*) e, quando o radar era acionado, ou seja, detectava algum objeto a se aproximar, o operador deveria decidir se o que acionou o radar era um avião inimigo (o sinal) ou outro objeto voador qualquer (Reiser & Faraggi, 1997).

A curva ROC é bastante utilizada em psicologia experimental para avaliar a capacidade dos indivíduos em distinguirem entre estímulo e não estímulo; em medicina, para analisar a qualidade de um determinado teste clínico; em economia, onde recebe o nome de gráfico de Lorenz, para avaliação de desigualdade de renda; em previsão do tempo, para avaliar a qualidade de predições de eventos raros (Prati et al., 2012). Além disso, ela é utilizada em vários ramos da pesquisa biomédica, pois grande parte destes métodos trata de problemas de classificação de indivíduos em grupos, doentes e não doentes, e, pode ser utilizada na área financeira para auxiliar na classificação de indivíduos como maus ou bons pagadores.

Suponha um modelo que classifique o indivíduo como positivo ou negativo e a observação real (positivo ou negativo). A partir destas informações, as estatísticas de avaliação de um teste podem ser apresentadas por meio de uma tabela cruzada entre os resultados preditos pelo modelo e classe real das observações (positivo ou negativo). Esta tabela é conhecida como matriz de confusão.

Tabela 3.1: Representação da matriz de confusão

	predito		
real	positivo	negativo	
positivo	VP	FN	VP+FN
negativo	FP	VN	FP+VN
	VP+FP	VN+PN	N

Se cada entrada da matriz contida na Tabela (3.1) for dividida pelo tamanho amostral N, a estimativa da probabilidade conjunta da classe real e da predição será

representada por cada casela da matriz de confusão como na Tabela (3.2), na qual Y representa a variável aleatória classe real (observada) do indivíduo e a variável T representa a classe predita pelo modelo de classificação. Suponha $T = 1$ como positivo, $T = 0$ como negativo para o classe predita pelo modelo e $Y = 1$ como a classe real positiva e $Y = 0$ como classe real negativa.

Tabela 3.2: Probabilidade conjunta (T,Y)

	T=1	T=0	
Y=1	p(T=1,Y=1)	p(T=1,Y=0)	p(T=1)
Y=0	p(T=0,Y=1)	p(T=0,Y=0)	p(T=0)
	p(Y=1)	p(Y=0)	1

Nota-se que da Tabela (3.1) é equivalente a Tabela (3.2), uma vez que a última é a Tabela (3.1) reescalada. Pode-se ainda decompor as probabilidade conjuntas em probabilidades condicionais e marginais com a utilização das leis fundamentais de probabilidade: $P(Y,T) = P(Y|T)P(T) = P(T|Y)P(Y)$, em que $P(Y|T)$ é a probabilidade condicional de Y dado T e $P(T|Y)$ é a probabilidade condicional de T dado Y. Calcula-se as probabilidades condicionais da seguinte forma:

$$P(Y|T) = \frac{P(Y,T)}{P(T)}, \tag{3.6}$$

$$P(T|Y) = \frac{P(Y,T)}{P(Y)}. \tag{3.7}$$

As probabilidades marginais de Y não dependem das previsões do modelo, sendo assim a distribuição de $P(Y)$ estimada a partir dos dados de treinamento. Geralmente a $P(Y = 1)$ é denotada por prevalência do evente de interesse, que, neste trabalho, é a inadimplência, ou seja, $P(Y = 1)$ denota a probabilidade do indivíduo ser mau pagador. A partir da Tabela (3.2) é possível calcular as medidas preditivas sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia.

A curva ROC não apresenta alterações quando são realizadas transformações monótonas nas observações, como logaritmo, raiz quadrada ou funções lineares com coeficientes positivos (Louzada Neto et. al., 2003). Pode-se encontrar as instruções para o uso da curva ROC no *software* estatístico *R* em Robin (2011).

3.2.2 Gráfico ROC e AUC

Uma maneira de estabelecer o ponto de corte ótimo é estimar a sensibilidade (S) e especificidade (E) para o modelo considerando diversos pontos de corte, dentro do intervalo de valores possíveis e utilizar aquele que maximiza simultaneamente a sensibilidade e especificidade. Suponha Γ a variável aleatória que representa o resultado do modelo utilizado e, se, $\Gamma > c_0$, então o solicitante é classificado como mau pagador, e no caso contrário, o solicitante é classificado como bom pagador. A curva ROC é, então, uma função contínua de S e $1-E$ para diversos valores de c_0 obtidos dentro do espaço amostral de Γ, ligados por linha reta (Louzada Neto, 2003). Assim, um particular ponto (S, $1-E$) é associado a cada ponte de corte c_0. É importante ressaltar que como S e $1-E$ são calculados separadamente, utilizando o resultado do modelo, e que a curva ROC é independente da prevalência.

Como descrito anteriormente, alguns métodos como árvore de decisão classificam indivíduos em 0 e 1, no caso de resultados binários; outros métodos como modelo de regressão logística, *probit*, *Naive bayes* fornecem probabilidades para cada indivíduo e é necessário estabelecer um ponto de corte ótimo de forma que seja possível transformar o valor contínuo em um resultado binário, 0 ou 1. Pode-se escolher arbitrariamente o ponto de corte, no entanto, dessa forma, o modelo é passível de muitos erros de classificação. A curva ROC pode auxiliar na escolha deste ponto através da *simulação* de vários pontos de corte, e, uma maneira de realizar tal tarefa, consiste na ordenação de todas as observações da amostra teste segundo o valor contínuo predito pelo modelo.

Suponha um exemplo de curva ROC produzida a partir de 20 observações descritas na Tabela (3.3).

Tabela 3.3: Observações e probabilidade (valor)

Observação	Valor	Observação	Valor
1	0.90	11	0.40
2	0.80	12	0.39
3	0.70	13	0.38
4	0.60	14	0.37
5	0.55	15	0.36
6	0.54	16	0.35
7	0.53	17	0.34
8	0.52	18	0.33
9	0.51	19	0.30
10	0.50	20	0.10

Qualquer curva ROC é *função degrau* e é produzida a partir de um número finito de observações. Quanto maior a quantidade de observações mais contínua a curva se torna. Além disso, cada ponto criado no tempo t do processo de geração da curva ROC, depende do ponto $t-1$, caso contrário, a curva gerará uma nuvem de pontos e não pontos de uma curva (Silva, 2006). Na Figura (3.1) tem-se os pontos $(1-E, S)$ criados a partir das 20 observações (pontos de corte) da Tabela (3.3).

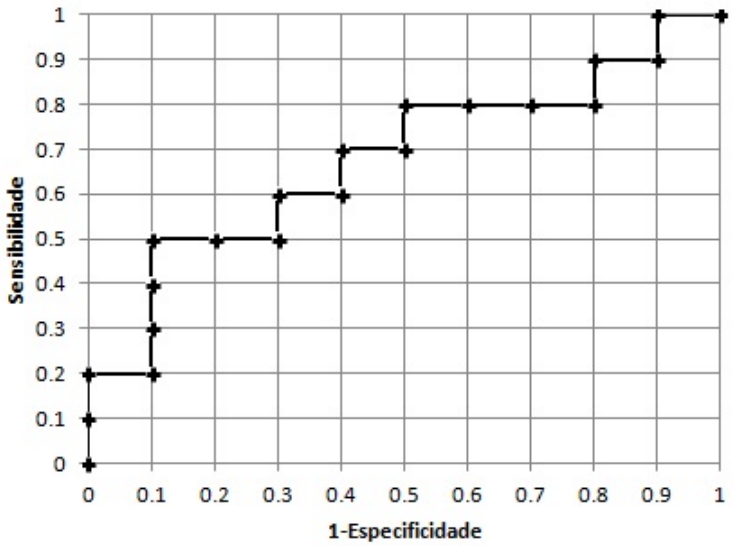

Figura 3.1: Curva ROC criada pelos pontos de corte do conjunto de teste

Alguns pontos específicos são importantes, como por exemplo, o ponto (0,0). Este ponto siginifica que todos os indivíduos são classificados como negativos, ou seja, o modelo não tem nenhum falso positivo, mas falha totalmente em classificar os verdadeiros positivos. O ponto (1,1) é a estratégia inversa, ou seja, o modelo classifica todos os indivíduos como positivo, não apresenta nenhum falso negativo, mas não consegue rotular os verdadeiros negativos. Modelos próximos ao canto inferior esquerdo apresentam poucos falsos positivos, mas têm baixas taxas de verdadeiros positivos, já modelos próximos ao canto superior classifica na maioria das vezes o indivíduo como positivo, mas, usualmente, tem alta taxas de falsos positivos. Um modelo terá melhor desempenho quanto mais afastada a curva ROC estiver da diagonal principal.

Já área sob a curva ROC (AUC) é uma medida escalar utilizada para sumarizar o desempenho de um modelo, uma vez que ela é estimada através do uso de todas as sensibilidade e especificidades associadas a cada um dos pontos de corte utilizados. Os valores de AUC variam entre 0 e 1. Porém, não existem modelos classificadores com

AUC menor que 0,5, pois esta é área de um modelo classificador aleatório. Dentre os métodos utilizados para estimar AUC encontram-se: regra do trapézio; estimação por máxima verossimilhança; a partir do declive e do termo de intercepção da representação dos dados originais em papel de probabilidade *binormal*; a aproximação à estatística U de Wilcoxon-Mann-Whitney (Braga, 2000).

Quanto melhor o modelo discrimina os indivíduos em dois grupos, mais a curva ROC se aproxima do canto superior esquerdo do gráfico, ou seja, quanto mais AUC está próxima de 1, melhor é o desempenho do modelo em classificar corretamente os indivíduos. A Figura (3.2) mostra a área sob a curva ROC.

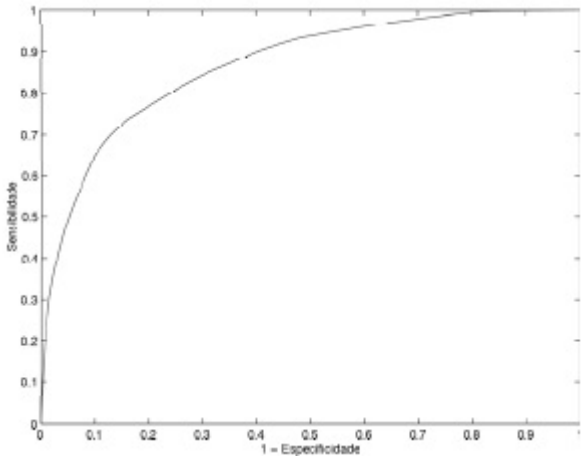

Figura 3.2: Área sob a curva ROC

Um modelo que possui AUC igual a 0.5 não classifica melhor os indivíduos do que um modelo classificador aleatório. Embora a área sob a curva ROC possa sumarizá-la como um todo, em algumas situações, é mais recomendável avaliar o desempenho do teste em específicas partes da curva do que em sua total extensão (Louzada Neto et. al., 2003).

3.3 Avaliação do modelo

Através das medidas de desempenho extraida da matriz de confusão (ou de suas derivações) é possível avaliar se um modelo está classificando bem ou não. Pode-se reduzir a matriz a uma única medida para avaliar quantitativamente o desempenho de um modelo, sendo que tal ação, em primeira análise, apresenta vantagens, uma vez que selecionar um modelo, entre alguns disponíveis, é mais fácil quando tal tarefa é baseada num único valor. No entanto, nos modelos utilizados em *credit scoring*, a utilização de uma única medida não é recomendada, pois um modelo pode perfeitamente ser o melhor em sensibilidade e ter baixo desempenho na especificidade ou acurácia, sendo assim difícil escolher qual modelo será mais adequado a partir do uso de somente uma medida.

Por exemplo, ao utilizar somente modelos de classificação com alta sensibilidade, identifica-se melhor os maus pagadores da base de dados, porém, se o modelo escolhido apresentar baixa especificidade, ele tem um desempenho ruim na identificação dos bons pagadores, o que, em última análise, representa que o credor deixou de ganhar com um possível empréstimo para um bom pagador.

Outro exemplo seria um modelo com acurácia de 95%. Tal modelo não é necessariamente o mais adequado, pois não leva em consideração a prevalência amostral. Se a prevalência do evento de interesse for 5% e o modelo indicar todas as observações como negativas, então a acurácia seria de 95% num modelo que não teria utilidade prática ao classificar todos os solicitantes de crédito como bons pagadores (Rocha, 2012). Em geral, a utilização de apenas uma medida de desempenho para avaliação de um modelo consiste na escolha de um modelo errado, o qual conduzirá a conclusões erradas. Já quando utilizadas em conjunto permitem uma boa interpretação da adequabilidade do modelo.

Portanto, as medidas como sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia devem ser utilizadas conjuntamente. Estas métricas são extraídas da matriz de confusão (ou suas derivações).

Existem modelos que produzem resultados discretos que podem ser organizados na Tabela (3.1) como a árvore de decisão/classificação. Enquanto outros modelos fornecem valores contínuos como a regressão logística e *probit*. Nesse último caso, é necessário utilizar uma regra fundamentada num ponto de corte para classificar o indivíduo em positivo (1 - mau pagador) ou negativo (0 - bom pagador). Adota-se a seguinte posição: o indivíduo com probabilidade de ser mau pagador menor ou igual ao ponto de corte será

classicado como 0 (bom pagador) e o indivíduo com probabilidade maior que o ponto de corte será classificado como 1 (mau pagador).

A relevância da curva ROC, na área financeira, consiste em permitir a obtenção do ponto de corte ótimo. Porém, na prática, o ponto de corte é escolhido de acordo com a política de crédito da instituição financeira e o nível de inadimplência aceitável (Alves, 2008).

Capítulo 4

Inferência dos rejeitados

Neste capítulo são apresentados os métodos usuais de inferência dos rejeitados e o método utilizado nesta dissertação. Na Seção 4.1 é abordado o problema relacionado a modelagem *credit scoring* usual e na Seção 4.2 são apresentadas as principais técnicas utilizadas em inferência dos rejeitados. Na Seção 4.3, o método utilizado nesta dissertação.

4.1 O Problema

Em modelagem *credit scoring* utiliza-se apenas os solicitantes aceitos para realizar a estimação dos parâmetros, sendo os recusados (ou rejeitados) descartados do processo. Porém, uma amostra que represente bem o mercado de clientes deve considerar aqueles que tiveram o pedido de crédito recusado a fim de reduzir o viés amostral causado pela não inclusão destes indivíduos na modelagem. Na Figura (4.1), tem-se um esquema da distribuição dos dados em inferência dos rejeitados.

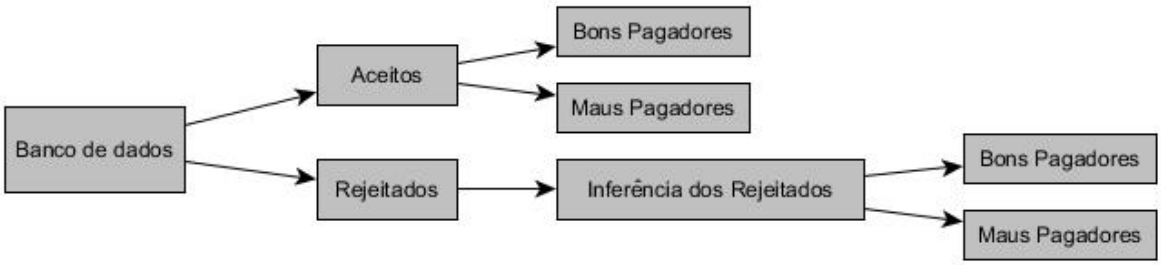

Figura 4.1: Esquema da distribuição dos dados para um modelo de *credit scoring*

No contexto da inferência dos rejeitados associa-se uma resposta ao indivíduo não observado e utiliza-se essa informação no processo de modelagem, ou seja, visa-se inferir qual seria o comportamento dos indivíduos rejeitados caso tivessem decisão favorável na

solicitação de crédito.

Como dito anteriormente, ao considerar os rejeitados na modelagem, reduz-se o viés amostral causado pelo uso somente dos indivíduos aceitos. Dessa forma, espera-se, ao utilizar a inferência dos rejeitados, melhorar o desempenho do modelo de *credit scoring*, no entanto, segundo Sicsu (2010), diversos autores comprovaram empiricamente que existem armadilhas no uso das técnicas considerando a inferência dos rejeitados:

- o fato do modelo ter bom desempenho entre os aprovados não implica necessariamente que isto ocorrerá entre os rejeitados;
- as variáveis que discriminam clientes bons pagadores e maus pagadores entre os aprovados nem sempre são as que melhor discriminam bons pagadores e maus pagadores entre rejeitados;
- as porcentagens de clientes bons pagadores e maus pagadores entre os aprovados não são necessariamente iguais às dos rejeitados.

Segundo Hand (1997) uma forma eficaz de analisar o comportamento dos rejeitados seria selecionar uma amostra aleatória destes rejeitados, conceder crédito a eles e observar seu comportamento e, então, classificá-los como bons pagadores ou maus pagadores. Porém tal ação seria certamente recusada pela diretoria da instituição, mas, a longo prazo, isto traria maiores resultados, pois acarretaria na obtenção de um modelo mais eficiente.

Logo, nos modelos *credit scoring*, uma alternativa é considerar as técnicas de inferência dos rejeitados, que, de acordo com Montrichard (2007), empiricamente, pode proporcionar os seguintes benefícios:

- identificação das características dos clientes associado ao risco de crédito;
- obtenção de estimativas mais precisas da taxa de maus pagadores;
- aumento na capacidade do modelo em distinguir os bons dos maus pagadores;
- facilidade na comparação de modelos candidatos.

4.2 Principais Métodos

Para utilizar a inferência dos rejeitados é necessário adotar algumas técnicas entre elas destacam-se: reclassificação, ponderação e parcelamento.

4.2.1 Reclassificação

A reclassificação consiste em classificar todos os rejeitados como maus pagadores, supondo que, se eles foram rejeitados, é porque *certamente* seriam maus pagadores. O problema é que reclassificando os indivíduos dessa forma classifica-se erroneamente toda a populaçao de bons pagadores que foram rejeitados e os solicitantes de perfis similares são prejudicados (Thomas et. al., 2002). Esta técnica somente é aplicável se for plausível assumir que os recusados são *certamente* maus pagadores. No entanto, tal aplicação não se configura como boa opção e deve ser evitada, tendo, inclusive, presente na literatura a opinião de analistas de *credit scoring* discutindo a validade ética deste método (Sicsu, 2010).

4.2.2 Ponderação

A ponderação consiste em assumir que a probabilidade do cliente ser mau pagador prescinde dele ter sido aceito ou não. Os rejeitados são representados pelos indivíduos com escore semelhante que foram aceitos e esta representação é feita através de pesos, em que indivíduos aceitos com escores altos (representando os rejeitados) tem maior peso. Para um dado valor escore do modelo, por exemplo, s, indivíduos com escores próximos a este valor tem perfis similares, independentes deles terem sido aceitos ou não. Ou seja, considera-se os indivíduos aceitos e rejeitados de forma ponderada e esta ponderação é obtida a partir da classificação de modelo inicial formulado com todos os proponentes (AR - aceita ou rejeita). A próxima etapa é construir um novo modelo ponderado atribuindo-se um peso para cada indivíduo da população de aceitos.

Joanes (1993) foi quem inicialmente propôs o método da ponderação através de análise discriminante, e, posteriormente, Banasik e Crook (2005) utilizaram a regressão logística. Com uma amostra com as mesmas proporções do total de indivíduos que solicitaram crédito, constrói-se um modelo com a variável resposta dicotômica, ou seja, aceita ou rejeita o pedido de crédito. Neste modelo AR a proposta é identificar os clientes aceitos com perfis próximos aos dos rejeitados. Dessa forma, o cliente aceito com este perfil tem

maior influencia no modelo, pois terá mais peso.

A próxima etapa é desenvolver um novo modelo *credit scoring* ponderado. Esta técnica pode ser descrita de acordo com as seguintes etapas:

- desenvolve-se um modelo que discrimine os aceitos e rejeitados (modelo AR).
- calcula-se os escores AR para todos os indivíduos.
- define-se classes de escores, preferencialmente de mesmas frequências.
- classifica-se os indivíduos aceitos e rejeitados nessas classes.
- pondera-se os aceitos.
- constrói-se um novo modelo com os clientes aceitos ponderados.

As duas tabelas a seguir ilustram o método da ponderação.

Tabela 4.1: Cálculo dos pesos para aceitos em cada classe de risco - AR

Classe de Risco (AR)	Aceitos	Rejeitados	Pesos (aceitos)
1	A_1	R_1	$(A_1 + R_1)/A_1$
2	A_2	R_2	$(A_2 + R_2)/A_2$
–	–	–	–
k	A_k	R_k	$(A_k + R_k)/A_k$

Fonte: Adaptado de Sicsu (2010).

Tabela 4.2: Exemplo numérico do cálculo dos pesos para os aceitos em cada classe de risco - AR

Classe de Risco (AR)	Aceitos	Rejeitados	Pesos (aceitos)
0 - 499	233 (16,0%)	1223 (84,0%)	6,25
500 - 549	675 (60,9%)	434 (39,1%)	1,64
550 - 649	897 (70,4%)	378 (29,6%)	1,42
650 - 749	988 (80,9%)	234 (19,1%)	1,24
750 - 899	1432 (88,3%)	190 (11,7%)	1,13
900 - 1000	1220 (100,0%)	0 (0,0%)	1,00

Fonte: Adaptado de Alves (2008).

Em suma, na ponderação são atribuídos pesos para os indivíduos considerando o resultado das classificações do modelo AR, construído com todos os solicitantes. Após isto, constrói-se um modelo com os proponentes aceitos ponderados.

4.2.3 Parcelamento

O parcelamento foi apresentado por Ash e Meester (2002) e consiste em formular um modelo somente com os proponentes aceitos e dividi-los em faixas de escore, verificando a taxa de inadimplência por faixa de escore. Em seguida, escora-se os rejeitados e associa-se a resposta bom ou mau pagador de forma aleatória para os rejeitados em cada faixa de escore de acordo com a respectiva taxa de inadimplência observada nos aceitos. A suposição feita é que os indivíduos com o mesmo escore, devem possuir o mesmo comportamento.

Para implementar esta técnica deve-se segmentar parceladamente a população de rejeitados, divididos entre bons pagadores e maus pagadores, segundo o risco do comportamento dos clientes aprovados, utilizando-se as taxas de inadimplência observadas. Este método é aplicado quando existe um de *credit score* já em operação.

Recomenda-se a utilização de uma amostra de clientes bons pagadores, maus pagadores e rejeitados, na mesma proporção do total de solicitantes de crédito. Para cada faixa de escore é realizada um parcelamento aleatório dos rejeitados, considerando a frequencia observada de clientes bons pagadores e maus pagadores presentes na população de não rejeitados. Segundo Sicsu (2010) nesta técnica consideram-se as seguintes etapas:

- desenvolve-se um modelo para discriminar os clientes bons pagadores e maus pagadores.
- calcula-se os escores para todos os indivíduos, aceitos e rejeitados.
- define-se k classes de risco (faixas de escores), preferencialmente de mesma frequência.
- classifica-se todos os indivíduos, aceitos (A) e rejeitados (R) nessas classes.
- sejam $p(m|j, A)$ e $p(b|j, A)$ as proporções de maus pagadores e de bons pagadores aprovados classificados na classe de risco j ($j = 1, 2, ..., k$).
- dentre os rejeitados classificados na classe j selecionamos aleatoriamente $p(m|j, A)$ rejeitados e os classificamos como maus pagadores. Os demais recusados da classe j são classificados como bons pagadores.

- *roda-se* o novo modelo considerando os aceitos e os rejeitados já devidamente classificados.

Na Tabela (4.3) encontra-se um exemplo de parcelamento. Na faixa de 0 a 200, tem-se 100% de clientes maus pagadores, então os 355 indivíduos rejeitados nesta faixa devem ser classificados como maus pagadores. Na faixa de 201 a 300, tem-se 69% de clientes maus pagadores e 31% de bons pagadores; nesta faixa, os rejeitados (262) devem ser divididos aleatoriamente em 181 maus pagadores e 81 bons pagadores e assim sucessivamente, repetindo-se o processo até que todas as faixas sejam preenchidas.

Tabela 4.3: Distribuição de risco para aceitos e parcelamento dos rejeitados

Classe de risco	Aceitos			Rejeitados		
	$Maus_A$	$Bons_A$	$Total_A$	$Maus_R$	$Bons_R$	$Total_R$
0-200	85 (11%)	685 (89%)	770	14	111	125
201-300	125 (20%)	512 (80%)	637	39	162	201
301-400	112 (24%)	355 (76%)	467	44	141	185
401-500	152 (68%)	68 (31%)	220	181	81	262
500-1000	521 (100%)	0 (0%)	521	355	0	355

Fonte: Adaptado de Alves (2008).

Na aplicação deste método encontra-se um empecilho, uma vez que é necessário que a empresa tenha um modelo *credit scoring* em produção para que seja possível reclassificar os solicitantes rejeitados. Como alternativa para aplicação da técnica de parcelamento poderia-se reclassificar aleatoriamente os rejeitados a partir da taxa de inadimplência total observada para os clientes aceitos.

Outro método a se considerar em inferência dos rejeitados consiste na utilização de informações de mercado (*bureau* de crédito), que, resumidamente, utiliza dados de alguma central de crédito com registros das atividades de crédito dos solicitantes, o que permite verificar como os proponentes se comportam com relação aos outros tipos de compromisso como contas de energia, telefone, seguros, entre outros (Rocha, 2012).

Os rejeitados são avaliados em duas ocasiões, quando solicitam crédito e depois de um tempo pré-determinado, que é considerado como um período de avaliação do comportamento dos solicitantes rejeitados. Na primeira ocasião, é possível haver rejeitado sem nenhum tipo de irregularidade ou pendência e que passe a tê-la ou permaneça sem irregularidade durante o período de avaliação. No entanto, pode ocorrer o caso de indivíduo rejeitado, que em primeira análise, quando ocorre a solicitação de crédito, apresenta irregularidade, e passe a não tê-la ou continue com pendência durante o período de avaliação. A partir desta análise, um novo modelo é construído com os indivíduos aceitos

e os rejeitados com a variável resposta definida segundo as informações do mercado. No entanto, o acesso às informações de mercado pode exigir um investimento financeiro que não deve ser considerado (Rocha, 2012).

4.3 Método utilizado

As principais técnicias utilizadas em inferência dos rejeitados apresentam algumas desvantagens. A reclassificação, por exemplo, não considera que entre os rejeitados possa existir bons pagadores; já na ponderação tem-se a necessidade de um modelo inicial que separe a população de aceitos e rejeitados, e, nota-se, em Alves (2008), que os resultados obtidos pela ponderação são similares ao da regressão logística usual, ou seja, utilizando somente os solicitantes aceitos; já o parcelamento, por sua vez, utiliza as taxas de inadimplência entre os aceitos por faixa de escore, supondo que o comportamento dos rejeitados é similar ao dos aceitos.

Dessa maneira, como alternativa para classificar os rejeitados optou-se pela utilização de variável latente com distribuição conhecida. Foram gerados valores aleatórios desta variável através de uma distribuição *Bernoulli* com probabilidade de sucesso estimada a partir da prevalência de inadimplentes (maus pagadores) e das medidas de desempenho de sensibilidade e especificidade dos modelos de regressão logística, *probit* e árvore de decisão/classificação via algoritmo CART. A escolha da distribuição *Bernoulli* deve-se ao fato dos valores da resposta para os rejeitados assumirem o valor 0 (bom pagador) ou 1 (mau pagador).

4.3.1 Uso de variável latente

Nesta dissertação são utilizados três modelos, regressão logística, *probit* e árvore de decisão/classificação para decidir se o empréstimo será concedido ou não ao solicitante. No entanto, para melhor entendimento, suponha inicialmente apenas um modelo, considerando os aceitos, ou seja, aqueles indivíduos em que é conhecida sua verdadeira condição (bom ou mau pagador). Considere T_1 como a classificação final obtida a partir de determinado modelo, $T_1 = 0$ ou $T_1 = 1$, e a real condição do indivíduo $Y = 1$ ou $Y = 0$ conforme a Tabela (4.4), em que 0 e 1 representam bom pagador e mau pagador, respectivamente.

Tabela 4.4: Probabilidades de ocorrência do resultado do modelo para cada indivíduo, segundo o comportamento (bom ou mau pagador) do indivíduo

T_1	Y	
	1	0
1	$P(T_1 = 1, Y = 1)$	$P(T_1 = 1, Y = 0)$
0	$P(T_1 = 0, Y = 1)$	$P(T_1 = 0, Y = 0)$

Fonte: Adaptado de Pereira, 2011.

Nessa abordagem, tem-se a variável aleatória $T_1|Y$ com função densidade de probabilidade definida por

$$f_{T_1|Y}(t_1|y) = [P(T_1 = t_{i1}, Y = 1)]^{y_i}[P(T_1 = t_{i1}, Y = 0)]^{1-y_i} \tag{4.1}$$

em que $t_{i1} = 0, 1$, $i = 1, 2, ..., n$ é a classificação feita pelo modelo para o i-ésimo solicitante a crédito e y_i a verdadeira condição do indivíduo, bom ou mau pagador.

Considerando uma amostra de n indivíduos independentes e identicamente distribuídos (iid), a função de verossimilhança é dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [P(T_1 = t_{i1}, Y = 1)]^{y_i}[P(T_1 = t_{i1}, Y = 0)]^{(1-y_i)} \tag{4.2}$$

em que $\boldsymbol{\theta} = (p, S_1, E_1)$, $p = P(Y = 1)$, S_1 denota a sensibilidade do modelo, E_1 a especificidade do modelo, p a prevalência de inadimplentes.

É possível escrever a probabilidade conjunta de T e Y em termos de probabilidade condicional da seguinte forma:

$$\begin{aligned}
&P(T_1 = 1, Y = 1) = P(Y = 1)P(T_1 = 1|Y = 1) = pS_1 \\
&P(T_1 = 0, Y = 1) = P(Y = 1)P(T_1 = 0|Y = 1) = p(1 - S_1) \\
&P(T_1 = 1, Y = 0) = P(Y = 0)P(T_1 = 1|Y = 0) = (1 - p)(1 - E_1) \\
&P(T_1 = 0, Y = 0) = P(Y = 0)P(T_1 = 0|Y = 0) = (1 - p)E_1
\end{aligned} \tag{4.3}$$

Utilizando as probabilidades encontradas em (4.3), é possível escrever a função de verossimilhança (4.2) em (4.4):

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [pS_1^{t_{i1}}(1 - S_1)^{(1-t_{i1})}]^{y_i}[E_1^{(1-t_{i1})}(1 - E_1)^{t_{i1}}]^{(1-y_i)} \tag{4.4}$$

em que $\boldsymbol{\theta} = (p, S_1, E_1)$. Como o comportamento do indivíduo foi observado (bom ou mau pagador), ou seja, detém-se esta informação, através da amostra teste composta de aceitos, pode-se estimar diretamente os parâmetros da função de verossimilhança.

No entanto, no caso dos rejeitados, a verdadeira condição do indivíduo, y_i, bom ou mau pagador, é desconhecida. Para encontrar a classificação final do indivíduo, t_{i1}, pode-se utilizar o seguinte procedimento: primeiramente formula-se um modelo, por exemplo, regressão logística; depois escora-se o solicitante rejeitado; obtém-se um ponto de corte ótimo com o auxílio da curva ROC e então classifica-se o indivíduo em bom ou mau pagador, comparando o escore obtido com o ponto de corte. Obtém-se o valor $t_{i1} = 0$, se o indivíduo é classificado como bom pagador, $t_{i1} = 1$, se o indivíduo é classificado mau pagador.

Como a verdadeira condição do rejeitado é desconhecida, trata-se de dados incompletos e a função de verossimilhança para uma amostra de n indivíduos iid pode ser definida em (4.5):

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [pS_1^{t_{i1}}(1 - S_1)^{(1-t_{i1})}][E_1^{(1-t_{i1})}(1 - E_1)^{t_{i1}}]. \tag{4.5}$$

Este modelo não apresenta condição básica de identificabilidade ($gl \geq Qp$), em que gl significa graus de liberdade e Qp denota a quantidade de parâmetros a ser estimada - ($gl = 2^K - 1 = 2^1 - 1 = 1$) e ($Qp = 3$). Dessa maneira, a inclusão de uma variável latente Z com distribuição de probabilidade conhecida pode ser considerada como alternativa viável (Dempster, Laird, Rubin, 1977). A função dessa variável latente é estimar a variável resposta de cada solicitante rejeitado. Como o rejeitado pode ser somente ter dois resultados (bom ou mau pagador), Z_i pode ser estimada de uma distribuição Bernoulli(τ_i).

A função densidade de probabilidade marginal de Z é dada por:

$$f_Z(z) = \tau_i^{z_i}(1 - \tau_i)^{(1-z_i)}, 0 \leq \tau_i \leq 1, z_i = 0, 1, i = 1, 2, ..., n. \tag{4.6}$$

Considerando n indivíduos iid da variável latente Z, a função de verossimilhança é dada por:

$$L(\boldsymbol{\tau}) = \prod_{i=1}^{n} \tau_i^{z_i}(1 - \tau_i)^{(1-z_i)}. \tag{4.7}$$

Seja probabilidade de sucesso τ_i, que representa a probabilidade do indivíduo rejeitado

ser mau pagador, dada por:

$$\begin{aligned}
\tau_i &= P(Y=1|T_1=t_{i1}) \\
&= \frac{P(Y=1,T_1=t_{i1})}{P(T_1=t_{i1})} \\
&= \frac{P(Y=1)P(T_1=t_{i1}|Y=1)}{P(Y=1)P(T_1=t_{i1}|Y=1)+P(Y=0)P(T_1=t_{i1}|Y=0)} \\
&= \frac{pS_1^{t_{i1}}(1-S_1^{1-t_{i1}})}{pS_1^{t_{i1}}(1-S_1^{1-t_{i1}})+(1-p)E_1^{(1-t_{i1})}(1-E_1)^{t_{i1}}}. \qquad (4.8)
\end{aligned}$$

A função de verossimilhança aumentada para um modelo, após a combinação da função de verossimilhança dos dados incompletos (4.5) com a função de verossimilhança da variável latente Z (4.7) é dada por:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[pS_1^{t_{i1}}(1-S_1^{1-t_{i1}})\right]^{z_i} \left[(1-p)E_1^{(1-t_{i1})}(1-E_1)^{t_{i1}}\right]^{(1-z_i)} \qquad (4.9)$$

Com o intuito de estimar os parâmetros do vetor $\boldsymbol{\theta}$ pode-se utilizar métodos como algoritmo EM (*Expectation Maximization*) (Dempster, Liard, Rubin, 1977), *Gibbs Sampling* e *Metropolis Hastings*.

Suponha agora a situação em que são utilizados três modelos condicionalmente independentes. De forma análoga a um modelo, no caso dos rejeitados, após a inclusão da variável latente Z com distribuição de Bernoulli(τ), a função de verossimilhança aumentada para amostra de n indivíduos iid, considerando a prevalência de inadimplentes, as sensibilidades e especificidades dos três modelos, é dada por:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[p\prod_{k=1}^{3} S_k^{t_{ik}}(1-S_k)^{(1-t_{ik})}\right]^{z_i} \left[(1-p)\prod_{k=1}^{3} E_k^{(1-t_{ik})}(1-E_k)^{t_{ik}}\right]^{(1-z_i)} \qquad (4.10)$$

em que $\boldsymbol{\theta}$=(p,**S**,**E**) constitui o vetor de parâmetros, $\boldsymbol{S}=(S_1,S_2,S_3)$ é um vetor composto pelas sensibilidades do modelo 1, 2 e 3; $\boldsymbol{E}=(E_1,E_2,E_3)$, é um vetor composto pela especificidade do modelo 1, 2 e 3; p é a probabilidade de inadimplência, $t_{ik}=0,1$ é o resultado do *k-ésimo* modelo para o *i-ésimo* rejeitado; $z_i=0,1$ é a verdadeira condição do indivíduo rejeitado (bom ou mau pagador, respectivamente). Tal condição deve ser estimada com a utilização de uma variável latente $Z \sim$ Bernoulli(τ) e probabilidade de sucesso definida em (4.11):

$$\tau_i = \frac{p\prod_{k=1}^{3} S_k^{t_{ik}}(1-S_k)^{(1-t_{ik})}}{p\prod_{k=1}^{3} S_k^{t_{ik}}(1-S_k)^{(1-t_{ik})} + (1-p)\prod_{k=1}^{3} E_k^{(1-t_{ik})}(1-E_k)^{t_{ik}}} \tag{4.11}$$

Como são utilizados três modelos, apesar da variável resposta dos rejeitados ser desconhecida, tem-se condição básica de identificabilidade (gl $\geq$ Qp), ou seja, $gl = 7$, $Qp = 7$. Nesta dissertação foi utilizado o algoritmo EM com o objetivo de estimar o valor do vetor $\boldsymbol{\tau}$ e partir disto gerar aleatoriamente a resposta dos rejeitados.

4.3.2 Algoritmo EM

O algoritmo EM é uma ferramenta computacional utilizada para estimação de parâmetros via máxima verossimilhança com dados faltantes. Para isso, é necessário obter o conjunto de dados observados aumentado com o conjunto de dados faltantes e então obter a função de log-verossimilhança associada aos dados aumentados. Métodos semelhantes ao EM vinham sendo utilizados há muito tempo, porém, a partir de 1977, quando Dempster, Laird e Rubin o desenvolveram e mostraram as suas propriedades, este algoritmo passou a ser conhecido com esta denominação e a ser bastante utilizado.

O algoritmo EM é aplicado, geralmente, quando os dados são incompletos, quando a maximização da função de verossimilhança é analiticamente complexa ou em problemas envolvendo variáveis latentes. O algoritmo EM lida com a seguinte ideia: (1) substituir os valores faltantes por valores estimados; (2) estimar os parâmetros; (3) reestimar os valores faltantes adimitindo que os valores estimados estão corretos; (4) reestimar os parâmetros. Repete-se o processo até que um critério de convergência seja atingido (Junger, 2006). O que o algoritmo EM proporciona é a substituição de uma difícil maximização por uma sequência de maximizações mais simples, utilizando dois passos, o passo E (esperança) que obtém o valor esperado do logaritmo da verossimilhança aumentada; e o passo M, que calcula o seu máximo. O processo é então repetido até que se alcance um critério de convergência.

Em outras palavras, seja $\mathbf{t}$ o conjunto de dados observados e $\mathbf{z}$ o conjunto de dados faltantes. O conjunto de dados aumentado $\boldsymbol{t_c}$=($\mathbf{t}$, $\mathbf{z}$) é formado por $\mathbf{t}$ aumentado de $\mathbf{z}$, sua função de densidade é dada por p($\boldsymbol{t_c}|\theta$), $\theta \in \Theta \subset R^p$ e função de log-verossimilhança é denotada por $l_c(\theta|\boldsymbol{t_c})$. O algoritmo EM utiliza um processo iterativo e, a cada iteração deste algoritmo, dois passos estão envolvidos:

- Passo E. Calcular $Q(\theta|\theta^{(k)}) = E[l_c(\theta|\boldsymbol{t_c})|\boldsymbol{t}, \theta^{(k)}]$, em que a esperança é tomada com relação a distribuição condicional p($\boldsymbol{t}|\boldsymbol{z}$,$\theta^{(k)}$).

- Passo M. Encontrar $\theta^{(k+1)}$ que maximiza $Q(\theta|\theta^{(k)})$.

Deve-se repetir os passos até que a convergência seja atingida e como critério de parada pode-se utilizar $|\theta^{(k+1)} - \theta^{(k)}| < \epsilon$, em que ϵ é um número suficientemente pequeno ou um número máximo de iterações K, em que $\ell \leq K$.

Muitas vezes, a maximização é um passo complicado e segundo Zeller (2009) tal fato pode ser amenizado maximizando condicionalmente a alguma função dos parâmetros que estão sendo estimados.

Algoritmo EM - uso de variável latente

Considere a função de verossimilhança dada em (4.10), $\boldsymbol{\theta} = (p, S_k, E_k)$, $k = 1, 2, 3$, $p = P(Y = 1)$. Aplicando-se o logaritmo na função de verossimilhança obtém-se $l_c(\boldsymbol{\theta}|t_c)$ dada por:

$$
\begin{aligned}
l_c(\boldsymbol{\theta}|t_c) &= log(L_c(\theta|t_c)) \\
&= \sum_{k=1}^{3}\sum_{i=1}^{n} z_i log\left[pS_k^{t_{ik}}(1-S_k)^{(1-t_{ik})}\right] + \sum_{k=1}^{3}\sum_{i=1}^{n}(1-z_i)log\left[(1-p)E_k^{(1-t_{ik})}(1-E_k)^{t_{ik}}\right] \\
&= \sum_{k=1}^{3}\sum_{i=1}^{n} z_i\left[logp + t_{ik}logS_k + (1-t_{ik})log(1-S_k)\right] + \\
&\quad \sum_{k=1}^{3}\sum_{i=1}^{n}(1-z_i)\left[log(1-p) + (1-t_{ik})logE_k + t_{ik}log(1-E_k)\right] \\
&= \sum_{i=1}^{n} z_i logp + \sum_{k=1}^{3}\sum_{i=1}^{n} z_i t_{ik} logS_k + \sum_{k=1}^{3}\sum_{i=1}^{n} z_i(1-t_{ik})log(1-S_k) + \\
&\quad \sum_{i=1}^{n}(1-z_i)log(1-p) + \sum_{k=1}^{3}\sum_{i=1}^{n}(1-z_i)(1-t_{ik})logE_k + \\
&\quad \sum_{k=1}^{3}\sum_{i=1}^{n}(1-z_i)t_{ik}log(1-E_k)
\end{aligned} \tag{4.12}
$$

$E(Z_i|t_{ik}, \boldsymbol{\theta}) = \tau_i$, em que τ_i é dado por:

$$\tau_i = \frac{p\prod_{k=1}^{3} S_k^{t_{ik}}(1-S_k)^{(1-t_{ik})}}{p\prod_{k=1}^{3} S_k^{t_{ik}}(1-S_k)^{(1-t_{ik})} + (1-p)\prod_{k=1}^{3} E_k^{(1-t_{ik})}(1-E_k)^{t_{ik}}} \tag{4.13}$$

Passo E

$$\begin{aligned} E_Z[l_c(\boldsymbol{\theta}|t_c)|t, \boldsymbol{\theta}] &= logp\sum_{i=1}^{n}\tau_i + \sum_{k=1}^{3}\sum_{i=1}^{n}\tau_i t_{ik} logS_k + \\ &\sum_{k=1}^{3}\sum_{i=1}^{n}\tau_i(1-t_{ik})log(1-S_k) + log\,(1-p)\left(n - \sum_{i=1}^{n}\tau_i\right) + \\ &\sum_{k=1}^{3}\sum_{i=1}^{n}(1-\tau_i)(1-t_{ik})logE_k + \\ &\sum_{k=1}^{3}\sum_{i=1}^{n}(1-\tau_i)t_{ik}log(1-E_k) \end{aligned} \tag{4.14}$$

Passo M

$$\frac{\partial E_Z[l_c(\boldsymbol{\theta}|t_c, Z)|t, \boldsymbol{\theta}]}{\partial p} = \frac{1}{p}\sum_{i=1}^{n}\tau_i - \frac{1}{(1-p)}(n - \sum_{i=1}^{n}\tau_i) = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n}\tau_i}{n} \tag{4.15}$$

$$\frac{\partial E_Z[l_c(\boldsymbol{\theta}|t_c)|t, \boldsymbol{\theta}]}{\partial S_k} = \frac{1}{S_k}\sum_{i=1}^{n} t_{ik}\tau_i - \frac{1}{(1-S_k)}\sum_{i=1}^{n}(1-t_{ik})\tau_i = 0 \Rightarrow \hat{S}_k = \frac{\sum_{i=1}^{n} t_{ik}\tau_i}{\sum_{i=1}^{n}\tau_i} \tag{4.16}$$

$$\begin{aligned} \frac{\partial E_Z[l_c(\boldsymbol{\theta}|t_c)|t, \boldsymbol{\theta}]}{\partial E_k} &= \frac{1}{E_k}\sum_{i=1}^{n}(1-t_{ik})(1-\tau_i) - \frac{1}{(1-E_k)}\sum_{i=1}^{n} t_{ik}(1-\tau_i) = 0 \\ \hat{E}_k &= \frac{\sum_{i=1}^{n}(1-t_{ik})(1-\tau_i)}{\sum_{i=1}^{n}(1-\tau_i)} \end{aligned} \tag{4.17}$$

em que $k = 1, 2, 3$.

Desta forma, no primeiro passo, o algoritmo atualiza o valor de τ no passo $\ell + 1$, dados os valores de S, E e p no passo ℓ, ou seja,

$$\tau_{i,\ell+1} = \frac{p_\ell\prod_{k=1}^{3} S_{k,\ell}^{t_{ik}}(1-S_{k,\ell})^{(1-t_{ik})}}{p_\ell\prod_{k=1}^{3} S_{k,\ell}^{t_{ik}}(1-S_{k,\ell})^{(1-t_{ik})} + (1-p_\ell)\prod_{k=1}^{3} E_{k,\ell}^{(1-t_{ik})}(1-E_{k,\ell})^{t_{ik}}}$$

No segundo passo do algoritmo, atualiza-se os valores de S, E e p no passo $\ell+1$, dado o valor de τ no passo $\ell+1$.

$$\begin{aligned} p_{\ell+1} &= \frac{\sum_{i=1}^{n} \tau_{i,\ell+1}}{n}, \\ S_{k,\ell+1} &= \frac{\sum_{i=1}^{n} t_{ik}\tau_{i,\ell+1}}{\sum_{i=1}^{n} \tau_{i,\ell+1}}, \\ E_{k,\ell+1} &= \frac{\sum_{i=1}^{n} (1-t_{ik})(1-\tau_{i,\ell+1})}{\sum_{i=1}^{n} (1-\tau_{i,\ell+1})}. \end{aligned}$$

Sendo assim, segue-se iterativamente até obter a convergência dos valores estimados. Considerou-se como critério de convergência a diferença em valor absoluto de todos os parâmetros do modelo, isto é, $\max\{|\mathbf{S}_\ell - \mathbf{S}_{\ell+\mathbf{1}}|, |\mathbf{E}_\ell - \mathbf{E}_{\ell+\mathbf{1}}|, |\mathbf{p}_\ell - \mathbf{p}_{\ell+\mathbf{1}}|\} < \mathbf{0.1}$ ou número máximo de iterações igual a 100.

Apesar da possibilidade de estimar a probabilidade de inadimplência, sensibilidade e especificidade dos modelos para o caso dos rejeitados, no contexto de *credit scoring*, o interesse é estimar τ_i e, a partir disto, gerar aleatoriamente valores da variável latente Z. O objetivo é incluir os rejeitados na amostra de treinamento, composta inicialmente apenas de aceitos, e novamente construir os três modelos. Dessa forma, espera-se melhorar o desempenho dos modelos, que terão as medidas de sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia calculadas a partir da amostra de teste. A fim de obter ainda maior capacidade preditiva, os resultados dos modelos são combinados segundo regras de decisão, descritas no Capítulo 5.

Capítulo 5

Regras de decisão

É possível especificar algumas combinações de resultados a fim de se melhorar a perfomance dos modelos, ou seja, obter maiores medidas de capacidade preditiva.

Considere a variável T_{ik}, $k = 1, 2, 3$, que define a classicação do i-ésimo indivíduo, $i = 1, 2, ...n$, como bom ou mau pagador para os três modelos: regressão logística, *probit* e árvore de decisão/classificação. Considere também $S^*_{c_{ik}}$, $k = 1, 2$, como o vetor de escoragem composto das probabilidades de inadimplência dos n indivíduos para os modelos de regressão logística, $k = 1$, e *probit*, $k = 2$.

Utilizando a escoragem, a classificação dos indivíduos é feita utilizando o ponto de corte c_0, obtido através do uso da curva ROC, da seguinte forma:

$$T_{ik} = \begin{cases} 1, \text{se } S^*_{c_{ik}} > c_0 \\ 0, \text{caso contrário} \end{cases}$$

em que $T_{ik} = 1$ denota a classificação do i-ésimo indivíduo em mau pagador para o modelo $k = 1, 2$, e, $T_{ik} = 0$, denota a classificação do i-ésimo indivíduo em bom pagador para o modelo $k = 1, 2$.

No caso da árvore de decisão/classificação obtém-se diretamente a classificação do i-ésimo indivíduo, $T_{i3} = 1$, se o i-ésimo indivíduo é classificado como mau pagador; $T_{i3} = 0$, se o i-ésimo indivíduo é classificado como bom pagador.

Nesta dissertação foram utilizadas três combinações de resultados, isto é, três regras de decisão para concessão de crédito. Estas regras são denotadas por:

1. conservadora;
2. combinação via voto majoritário;
3. uso do modelo mais sensível e mais específico.

A regra de decisão **conservadora** consiste em conceder crédito para o cliente se os três modelos indicarem que o solicitante a crédito é bom pagador; caso contrário, o indivíduo é considerado mau pagador. Ou seja,

$$T_i^* = \begin{cases} 0, \text{se } \sum_{k=1}^{3} T_{ik} = 0 \\ 1, \text{caso contrário} \end{cases}$$

em que T_i^* denota a classificação final do i-ésimo indivíduo, sendo $T_i^* = 0$, se o i-ésimo indivíduo é classificado como bom pagador e, $T_i^* = 1$, se o i-ésimo indivíduo é classificado como mau pagador.

Já com a **combinação via voto majoritário** decide-se pela concessão de crédito obedecendo a maioria dos votos obtidos, ou seja, se pelo menos dois dos três modelos utilizados classificarem o solicitante como bom/mau pagador então deve-se conceder/não o crédito. Esta regra pode ser definida da seguinte maneira:

$$T_i^* = \begin{cases} 0, \text{se } \sum_{k=1}^{3} T_{ik} = 0 \text{ ou } \sum_{k=1}^{3} T_{ik} = 1 \\ 1, \text{se } \sum_{k=1}^{3} T_{ik} = 2 \text{ ou } \sum_{k=1}^{3} T_{ik} = 3 \end{cases}$$

para $i = 1, 2, ..., n$.

Com o **uso do modelo mais sensível e mais específico** procede-se da seguinte maneira: **i)** encontra-se o modelo mais sensível e se ele indicar que o indivíduo é mau pagador, então não se concede o crédito independentemente dos resultados dos outros dois modelos; **ii)** encontra-se o modelo mais específico e se este modelo indicar que o indivíduo é bom pagador, então concede-se crédito. No caso do modelo mais sensível indicar que

o indivíduo é bom pagador e do modelo mais específico indicar que o indivíduo é mau pagador, utiliza-se a combinação via voto majoritário para decidir pela concessão ou não de crédito ao solicitante.

Capítulo 6

Exemplo de Aplicação

Neste capítulo são apresentados os exemplos de aplicação. Na Seção 6.1 tem-se a descrição dos bancos de dados. Na Seção 6.2 são apresentados os resultados considerando as medidas preditivas obtidas com os modelos de regressão logística, *probit* e árvore de decisão/classificação individualmente e com a combinação dos resultados destes modelos via regras de decisão considerando o uso de variável latente, reclassificação, parcelamento e ponderação como técnicas de inferência dos rejeitados.

6.1 Banco de dados e simulações

Foram utilizado bancos de dados simulados com 6 variáveis contínuas e prevalência de 10%, 25% e 50% de maus pagadores na amostra. Além disso, foi utilizado um banco de dados real proveniente de uma instituição financeira composto de 8 variáveis categóricas e 7 numéricas, 45.719 observações, com 25.459 aceitos e a prevalência de maus pagadores de 25%, além de 20.260 rejeitados, que corresponde a $44,5\%$ do total de clientes da amostra. Nos bancos de dados utilizados foi modelada a probabilidade do indivíduo ser mau pagador.

Considerou-se uma variável resposta dicotômica indicando bons ou maus pagadores (0 - bom pagador, 1 - mau pagador). Foi utilizada a proposta feita por Breiman (1998), em que a população de bons pagadores é dada por k covariáveis com distribuição normal multivariada com vetor de médias $\mu = (0, ..., 0)$ e matriz de covariâncias kXI_k, em que I é a matriz identidade de ordem k. A população de maus pagadores é gerada de uma normal multivariada com vetor de médias $\mu = (1/k, ..., 1/k)$ e matriz de covariâncias I_k. Na simulação considerou-se $k = 6$ covariáveis contínuas.

Dessa forma, inicialmente foi gerada uma população de aceitos com a seguinte

composição: 1.000.000 bons pagadores e 100.000 maus pagadores. Além disso, foi gerada uma população de rejeitados a partir da distribuição normal multivariada com vetor de média $\mu = (1/2, ..., 1/2)$ e matriz de covariâncias dada por $(1/2)XI_k$ com o objetivo de considerar diferenças entre a população de aceitos e de rejeitados.

Em seguida, em relação aos aceitos, retirou-se uma amostra aleatória estratificada (*full sample*) da população gerada, composta de 10.000 bons pagadores e 1.000 maus pagadores. Foram utilizados três cenários. No cenário 1, as amostras selecionadas foram obtidas mantendo os 1.000 maus pagadores acrescidos de 1.000 bons pagadores, o que equivale a prevalência de inadimplentes (p) de 50%. No cenário 2, utilizou-se 1.000 maus pagadores e 3.000 bons pagadores ($p = 25\%$). No cenário 3, 1.000 maus pagadores e 9.000 bons pagadores ($p = 10\%$). Nesta simulação, os bons pagadores foram retirados aleatoriamente do grupo de bons pagadores da *full sample*. Além disso, foi utilizada uma amostra de rejeitados na proporção de 50% do total de observações para cada cenário.

Separou-se o conjunto de dados dos solicitantes aceitos em amostra de treinamento inicial (70%) e amostra teste (30%) conforme proposto por Hosmer & Lemoshow (1989). Utilizou-se, individualmente, a variaável latente, reclassificação, ponderação, parcelamento para gerar a resposta dos rejeitados.

Para utilizar o método da variável latente, construiu-se então o modelo de regressão logística, *probit* e árvore de decisão/classificação com os solicitantes aceitos da amostra de treinamento inicial. Após isso, foram calculadas as medidas de sensibilidade e especificidade iniciais na amostra de teste. Estes valores de sensibilidade e especificidade, em conjunto com a prevalência de inadimplentes, foram utilizados como valores iniciais no algoritmo EM a fim de se obter a probabilidade de inadimplência (τ_i) de cada rejeitado.

Após calcular as probabilidades de sucesso foram gerados aleatoriamente os valores da variável latente, que estima as respostas dos rejeitados, através de uma distribuição de *Bernoulli*.

A partir disto, os rejeitados foram adicionados aos aceitos da amostra de treinamento inicial a fim de compor a amostra final de treinamento. Com a amostra final de treinamento foi gerado o modelo de regressão logística, *probit*, árvore de decisão/classificação. Combinou-se os resultados destes três modelos segundo as regras de decisão conservadora, combinação por voto majoritário e uso do modelo mais sensível e mais específico. Após esta etapa, as medidas de capacidade preditiva foram calculadas a partir da amostra teste

para os modelos individualmente e para as combinações dos resultados destes modelos via regras de decisão.

No caso da reclassificação atrabui-se a condição mau pagador para todos os solicitantes rejeitados. No parcelamento e na ponderação, a regressão logística foi utilizada para dividir os solicitantes em faixas de escore e obtenção dos pesos utilizados nos aceitos, etapas iniciais necessárias a aplicação ao primeiro e ao segundo método, respectivamente. Após a obtenção dos pesos, a opção *weights* foi utilizada no *software R* a fim de rodar o modelo logístico, *probit* e árvore de decisão/classificação com os aceitos ponderados.

Cada um dos três modelos foi construído a partir da amostra de treinamento composta de aceitos e rejeitados. Após esta etapa, assim como feito no uso de variável latente, as medidas de capacidade preditiva foram calculadas a partir da amostra teste para os modelos individualmente e para as combinações dos resultados.

6.2 Resultados

6.2.1 Aplicação em dados simulados

Foram realizadas 100 simulações obtendo 100 registros de cada uma das medidas de capacidade preditiva para cada cenário. Dessa forma, foram calculadas as médias desses registros e os respectivos erros padrão simulados. Em todos os cenários, os valores dos erros padrão simulados de cada medida de capacidade preditiva foram menores que $0,03$.

Os valores médios para o uso de variável latente podem ser encontrados nas Tabelas (6.1), (6.2), (6.3); reclassificação, nas Tabelas (6.4), (6.5), (6.6); parcelamento, nas Tabelas (6.7), (6.8), (6.9) e ponderação, nas Tabelas (6.10), (6.11), (6.12).

Tabela 6.1: Média das medidas preditivas dos modelos (p = 10%) - uso de variável latente

Modelo	SENS	ESPEC	VPP	VPN	ACC
logit	0,8902	0,8872	0,4712	0,9864	0,8874
probit	0,8912	0,8884	0,4736	0,9866	0,8864
CART	0,8464	0,7131	0,2488	0,9768	0,7264
Conservadora	0,9472	0,6820	0,2503	0,9915	0,7085
Combinação via voto majoritário	0,8905	0,8871	0,4705	0,9865	0,88744
Modelo mais sensível e mais específico	0,9457	0,6833	0,2508	0,9912	0,7096

Tabela 6.2: Média das medidas preditivas dos modelos (p = 25%) - uso de variável latente

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,8799	0,8841	0,7196	0,9567	0,8819
probit	0,8788	0,8870	0,7249	0,9565	0,8839
CART	0,9236	0,5276	0,3966	0,9547	0,6285
Conservadora	0,9606	0,5165	0,4001	0,9765	0,6275
Combinação via voto majoritário	0,8786	0,8816	0,7158	0,9563	0,8808
Modelo mais sensível e mais específico	0,9596	0,5171	0,4002	0,9703	0,6275

Tabela 6.3: Média das medidas preditivas dos modelos (p = 50%) - uso de variável latente

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,8973	0,8713	0,8747	0,8951	0,8787
probit	0,8974	0,8711	0,8744	0,8951	0,8790
CART	0,7961	0,7779	0,7831	0,7936	0,7870
Conservadora	0,9326	0,7354	0,7797	0,9168	0,8340
Combinação via voto majoritário	0,8863	0,8712	0,8742	0,8861	0,8787
Modelo mais sensível e mais específico	0,9298	0,7374	0,7804	0,9137	0,8336

Tabela 6.4: Média das medidas preditivas dos modelos (p = 10%) - reclassificação

Modelo	SENS	ESPEC	VPP	VPN	ACC
logit	0,9166	0,8707	0,4416	0,9894	0,8748
probit	0,9000	0,8864	0,4683	0,9876	0,8878
CART	0,9250	0,6444	0,2242	0,9872	0,6722
Conservadora	0,9600	0,6303	0,2240	0,9930	0,6633
Combinação via voto majoritário	0,9116	0,8718	0,4428	0,9888	0,8758
Modelo mais sensível e mais específico	0,9600	0,6303	0,2240	0,9930	0,6633

Tabela 6.5: Média das medidas preditivas dos modelos (p = 25%) - reclassificação

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,8916	0,8966	0,7422	0,9612	0,8954
probit	0,8983	0,8888	0,7293	0,9632	0,8912
CART	0,6750	0,8988	0,6899	0,8924	0,8429
Conservadora	0,9150	0,8433	0,6606	0,9674	0,8615
Combinação via voto majoritário	0,8916	0,8938	0,7370	0,9611	0,8933
Modelo mais sensível e mais específico	0,9083	0,8433	0,6663	0,9652	0,8633

Tabela 6.6: Média das medidas preditivas dos modelos (p = 50%) - reclassificação

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,8957	0,8913	0,8926	0,8963	0,8902
probit	0,8962	0,8895	0,8912	0,8968	0,8902
CART	0,8363	0,7752	0,7888	0,8264	0,8058
Conservadora	0,9360	0,7449	0,7863	0,9214	0,8404
Combinação via voto majoritário	0,8899	0,8900	0,8913	0,8914	0,8899
Modelo mais sensível e mais específico	0,9330	0,7474	0,7875	0,9183	0,8402

Tabela 6.7: Média das medidas preditivas dos modelos (p = 10%) - parcelamento

Modelo	SENS	ESPEC	VPP	VPN	ACC
logit	0,8870	0,8945	0,4871	0,9861	0,8935
probit	0,8896	0,8903	0,4785	0,9864	0,8902
CART	0,8886	0,6900	0,2424	0,9823	0,7098
Conservadora	0,9450	0,6717	0,2429	0,9909	0,6990
Combinação via voto majoritário	0,8886	0,8914	0,4870	0,9863	0,8911
Modelo mais sensível e mais específico	0,9440	0,6730	0,2435	0,9908	0,7001

Tabela 6.8: Média das medidas preditivas dos modelos (p = 25%) - parcelamento

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,8874	0,8924	0,7375	0,9600	0,8908
probit	0,8869	0,8924	0,7368	0,9597	0,8906
CART	0,6778	0,8820	0,6592	0,8916	0,8309
Conservadora	0,9129	0,8205	0,6306	0,9660	0,8436
Combinação via voto majoritário	0,8831	0,8965	0,7431	0,9586	0,8932
Modelo mais sensível e mais específico	0,9068	0,8282	0,6397	0,9410	0,8479

Tabela 6.9: Média das medidas preditivas dos modelos (p = 50%) - parcelamento

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,9116	0,8716	0,8773	0,9089	0,8883
probit	0,9366	0,8733	0,8782	0,9299	0,8883
CART	0,8233	0,7816	0,7912	0,8156	0,8025
Conservadora	0,9416	0,7400	0,7837	0,9267	0,8408
Combinação via voto majoritário	0,9050	0,8733	0,8788	0,9042	0,8891
Modelo mais sensível e mais específico	0,9400	0,7400	0,7833	0,9248	0,84

Tabela 6.10: Média das medidas preditivas dos modelos (p = 10%) - ponderação

Modelo	SENS	ESPEC	VPP	VPN	ACC
logit	0,8910	0,8827	0,4613	0,9865	0,8835
probit	0,8917	0,8825	0,4610	0,9865	0,8833
CART	0,9251	0,6220	0,2146	0,9868	0,6523
Conservadora	0,9588	0,6088	0,2148	0,9925	0,6438
Combinação via voto majoritário	0,8925	0,8804	0,4568	0,9866	0,8816
Modelo mais sensível e mais específico	0,9585	0,6095	0,2151	0,9925	0,6446

Tabela 6.11: Média das medidas preditivas dos modelos (p = 25%) - ponderação

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,9006	0,8680	0,6981	0,9634	0,8758
probit	0,8893	0,8751	0,7076	0,9600	0,8786
CART	0,8390	0,7780	0,5587	0,9354	0,7932
Conservadora	0,9393	0,7366	0,5434	0,9733	0,7873
Combinação via voto majoritário	0,8946	0,8705	0,7007	0,9616	0,8765
Modelo mais sensível e mais específico	0,9383	0,7384	0,5448	0,9729	0,7884

Tabela 6.12: Média das medidas preditivas dos modelos (p = 50%) - ponderação

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,8673	0,9057	0,9022	0,8724	0,8878
probit	0,8667	0,9043	0,9010	0,8720	0,8837
CART	0,8875	0,6829	0,7376	0,8594	0,7852
Conservadora	0,9342	0,6697	0,7395	0,9110	0,8019
Combinação via voto majoritário	0,8664	0,9031	0,9000	0,8703	0,8838
Modelo mais sensível e mais específico	0,9330	0,6708	0,7399	0,9096	0,8019

Nota-se que, em todos os cenários, para todos os métodos de inferência dos rejeitados utilizados, uso de variável latente, reclassificação, parcelamento e ponderação, as medidas de capacidade preditiva do modelo de regressão logística e *probit* apresentam valores próximos. Nota-se ainda que o desempenho dos modelos e do uso combinado va regras de decisão foi maior quando $p = 50\%$ (cenário 3) para todas as técnicas de inferência dos rejeitados. Em todos os cenários e em todas as técnicas, o uso da combinação de resultados apresentou ligeira melhora nas medidas de desempenho e, tal melhora, tornou-se maior quando desempenho do uso combinado foi comparado ao desempenho da árvore de decisão/classificação, sobretudo, na sensibilidade.

No entanto, quando são combinados os resultados dos três modelos, ocorre uma inversão entre sensibilidade e especificidade, fazendo com que o modelo final, obtido a partir das regras de decisão, apresente sensibilidade aumentada e especificidade diminuída. Com relação as demais medidas, deve-se ter cautela, sobretudo com o uso do valor preditivo positivo, pois, tal medida é influenciada pela prevalência amostral (Pereira, G. 2011).

O uso de variável latente, por sua vez, apresentou desempenho satisfatório e resultados bem próximos aos dos outros métodos usuais em inferência dos rejeitados.

6.2.2 Aplicação em dados reais

O conjunto de dados reais utilizado foi proveniente de uma instituição financeira com 45719 observações no total, sendo 25459 observações de solicitantes aceitos e 25% de prevalência de inadimplência. Além dos aceitos, constam 20260 solicitantes rejeitados na amostra, que correspondem a $44,3\%$ do total de indivíduos. Inicialmente, a amostra dos aceitos foi dividida em 70% para treinamento e 30% para a teste, sendo esta última utilizada para validação das medidas de capacidade preditiva. Após a geração dos valores para os rejeitados utilizando variável latente, reclassificação, parcelamento, ponderação, a amostra de treinamento inicial de aceitos foi acrescida dos rejeitados formando a amostra de treinamento final. Então foi gerado novamente o modelo de regressão logística, *probit* e árvore de decisão/classificação e seus resultados foram combinados via regras de decisão.

Os valores das medidas de capacidade preditiva podem ser encontrados nas Tabelas (6.13), (6.14), (6.15), (6.16) para o uso de variável latente, reclassificação, parcelamento e ponderação, respectivamente. A descrição das variáveis, análise descritiva e os modelos estão no Apêndice.

Tabela 6.13: Medidas preditivas dos modelos e do uso combinado - uso de variável latente

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,7059	0,6910	0,4336	0,8761	0,6947
probit	0,7001	0,6969	0,4365	0,8749	0,6977
CART	0,6641	0,7195	0,4425	0,8655	0,7057
Conservadora	0,7313	0,6585	0,4175	0,8805	0,6767
Combinação via voto majoritário	0,7013	0,6958	0,4358	0,8751	0,6972
Modelo mais sensível e mais específico	0,7290	0,6608	0,4184	0,8800	0,6779

Tabela 6.14: Medidas preditivas dos modelos e do uso combinado - reclassificação

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,6979	0,6815	0,4223	0,8712	0,6856
probit	0,6973	0,6834	0,4235	0,8713	0,6869
CART	0,6382	0,7306	0,4413	0,8582	0,7075
Conservadora	0,7225	0,6499	0,4076	0,8753	0,6681
Combinação via voto majoritário	0,6973	0,6820	0,4224	0,8716	0,6851
Modelo mais sensível e mais específico	0,7219	0,6501	0,4076	0,8752	0,6681

Tabela 6.15: Medidas preditivas dos modelos e do uso combinado - parcelamento

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,6555	0,7397	0,4564	0,8655	0,7186
probit	0,6591	0,7369	0,4551	0,8663	0,7174
CART	0,3607	0,9462	0,6910	0,8161	0,7998
Conservadora	0,6589	0,7360	0,4517	0,8663	0,7174
Combinação via voto majoritário	0,6554	0,7396	0,4564	0,8665	0,7186
Modelo mais sensível e mais específico	0,6549	0,7451	0,4590	0,8634	0,7184

Tabela 6.16: Medidas preditivas dos modelos e do uso combinado - ponderação

Método	SENS	ESPEC	VPP	VPN	ACC
logit	0,7078	0,6785	0,4234	0,8744	0,6859
probit	0,7109	0,6751	0,4218	0,8750	0,6840
CART	0,3947	0,9261	0,6406	0,8210	0,7932
Conservadora	0,7136	0,6726	0,4209	0,8756	0,6829
Combinação via voto majoritário	0,7052	0,6810	0,4243	0,8738	0,6870
Modelo mais sensível e mais específico	0,7078	0,6785	0,4243	0,8744	0,6859

Nota-se que os três métodos, individualmente, apresentam valores próximos nas medidas de capacidade preditiva, exceto o parcelamento, que teve desempenho inferior aos demais. A árvore de decisão/classificação apresenta menor sensibilidade e maior especificidade do que os modelos logístico, *probit* e as combinações via regras de decisão, sobretudo, quando utilizado o parcelamento e a ponderação. Ao combinar os resultados dos modelos, obtém-se maior sensibilidade como o uso da regra (1) - conservadora.

De um modo geral, a variável latente pode ser utilizada como alternativa em inferência dos rejeitados uma vez que apresenta valores satisfatórios em relação as todas as medidas de capacidade preditiva conjuntamente utilizadas.

Capítulo 7

Considerações Finais

7.1 Conclusões

As técnicas de inferência dos rejeitados tem por objetivo incluir os solicitantes que tiveram pedido de crédito negado na formulação do modelo. Muitas vezes, somente as técnicas de inferência dos rejeitados e o uso de uma única técnica de modelagem de dados não são suficientes para que se tenha medidas de capacidade preditiva satisfatórias. Dessa forma, optou-se por combinar o resultado dos três modelos regressão logística, *probit* e árvore de decisão/classificação via algoritmo CART.

A performance da combinação em que é utilizada regra conservadora e a utilização do modelo mais sensível e mais específico alcançaram melhor performance, sobretudo, quanto a sensibilidade. O uso da regra da combinação via voto majoritário e os modelos individualmente apresentaram maior especificidade.

De maneira geral, melhorias nas medidas de capacidade preditiva foram alcançadas quando foi utilizada das combinações de resultados dos três modelos, regressão logística, *probit*, árvore de decisão/classificação via regras de decisão para todas as técnicas utilizadas, uso de variável latente, reclassificação, parcelamento e ponderação.

No conjunto de dados simulados, os valores médios das medidas de capacidade preditiva obtidas foram similares para os três modelos. Houve um ligeiro aumento na sensibilidade quando utilizada a regra de decisão conservadora em todas as técnicas. Além disso, no cenário 3, em que a prevalência foi de 50%, os modelos apresentaram melhor desempenho individualmente e com os uso da combinação via regras de decisão em todas as técnicas de inferência dos rejeitados.

O conjunto de dados reais apresentou o mesmo comportamento dos dados simulados. A sensibilidade foi maior quando foram utilizadas as combinações de resultados

via regras de decisão. Dessa forma, se o analista busca um modelo que identifique melhor os *maus* pagadores é indicada a utilização de combinação através da regra de decisão conservadora ou uso do modelo mais sensível e mais específico.

A estratégia do uso da variável latente mostrou-se interessante, não somente pelo aumento na sensibilidade, mas por apresentar boas medidas de capacidade preditiva de maneira geral, sobretudo quando foi utilizado conjunto de dados reais. Quanto a escolha do método de inferência dos rejeitados, o pesquisador deve optar por aquele que apresentar melhor desempenho para o conjunto de dados investigado uma vez que todas as técnicas apresentaram desempenho satisfatório.

7.2 Trabalhos futuros

Para estudos futuros pode-se utilizar outros métodos de modelagem estatística e simular os valores da variável resposta dos rejeitados através da distribuição *Bernoulli* com probabilidade de sucesso definida a partir de uma relação das medidas de desempenho destes métodos. É possível, também, modificar a regra de decisão utilizada para concessão de crédito, além de considerar interações entre covariáveis.

É interessante utilizar outras técnicas de simulação da resposta dos rejeitados concomitantemente o uso de novos métodos de modelagem estatística ou realizar uma abordagem bayesiana. Poderia-se utilizar distribuições de probabilidade que levassem em consideração a assimetria nos dados, uma vez que, no conjunto de dados abordados, notou-se maior presença de bons pagadores.

Além disso, pode-se utilizar as técnicas empregadas nesta dissertação em modelagem na área médica e biológica.

Capítulo 8

Apêndice

8.1 Dados reais

8.1.1 Descrição das variáveis

- x1: tipo de ocupação.
- x2: escolaridade.
- x3: estado civil.
- x4: canal entrada.
- x5: sexo.
- x6: tipo de residência.
- x7: tempo de residência.
- x8: tipo de telefone.
- x9: idade em anos completos.
- x10: número de dependentes.
- x11: renda.
- x12: escore proveniente de instituição financeira externa I.
- x13: escore proveniente de instituição financeira externa II.
- conceito: variável resposta (0 - bom pagador, 1 - mau pagador).

8.1.2 Análise descritiva

Análise descritiva das covariáveis[1] candidatas aos modelos.

Tabela 8.1: Distribuição da inadimplência segundo tipo de ocupação

Tipo de ocupação	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
1	9678	50,6	3464	54,6	13142	51,6
2	747	3,9	288	4,5	1035	4,1
3	0	0,0	0	0,0	0	0,0
4	2083	10,9	571	9,0	2654	10,4
5	3731	19,5	1313	20,7	5044	19,8
6	208	1,1	61	1,0	269	1,1
7	2603	13,6	637	10,0	3240	12,7
8	0	0,0	0	0,0	0	0,0
9	62	0,3	13	0,2	75	0,3
Total	19112	100,0	6347	100,0	25459	100,0

Tabela 8.2: Distribuição da inadimplência segundo escolaridade.

Escolaridade	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
101	61	3,0	20	3,0	81	3,0
102	2473	12,9	783	12,3	3256	12,8
103	1811	9,5	601	9,5	2412	9,5
104	1288	6,7	501	7,9	1789	7,0
105	8432	44,1	2873	44,5	11305	44,4
106	1487	7,8	635	10,0	2122	8,3
107	2988	15,6	790	12,4	3778	14,8
108	73	0,4	19	0,3	92	0,4
109	499	2,6	125	0,5	624	2,5
Total	19112	100,0	6347	100,0	25459	100,0

Tabela 8.3: Distribuição da inadimplência segundo o estado civil.

Estado civil	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
1	5356	28,0	2004	31,6	7360	28,9
2	9047	47,3	2596	40,9	11643	45,7
3	442	2,3	167	2,6	609	2,4
4	782	4,1	262	4,1	1044	4,1
5	810	4,2	236	3,7	1046	4,1
6	2675	14,1	1082	17,1	3757	14,8
Total	19112	100,0	6347	100,0	25459	100,0

Tabela 8.4: Distribuição da inadimplência segundo canal de entrada.

Canal de entrada	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
1	13823	72,3	4648	73,2	18471	72,6
2	4637	24,3	1501	23,6	6138	24,1
3	443	2,3	137	2,2	580	2,3
4	20	0,1	7	0,1	27	0,1
5	189	1,0	54	0,9	243	33,1
Total	19112	100,0	6347	100,0	25459	100,0

[1] A instituição financeira não divulgou as categorias.

Tabela 8.5: Distribuição da inadimplência segundo o sexo.

Sexo	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
Feminino	9385	49,1	3051	48,1	12436	48,8
Masculino	9727	50,9	3296	51,9	13023	51,2
Total	19112	100,0	6347	100,0	25459	100,0

Tabela 8.6: Distribuição da inadimplência segundo tipo de residência.

Tipo de residência	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
1	5727	30,0	1640	25,8	7367	28,9
2	661	3,5	191	3,0	852	3,3
3	3831	20,0	1141	20,0	4972	19,5
4	2403	12,6	924	14,6	3327	13,1
5	6120	32,0	2315	36,5	8435	33,1
6	185	1,0	59	0,9	244	1,0
7	185	1,0	77	1,2	262	1,0
Total	19112	100,0	6347	100,0	25459	100,0

Tabela 8.7: Distribuição da inadimplência segundo tipo de telefone.

Tipo de telefone	Bons pagadores	(%)	Maus pagadores	(%)	Total	(%)
0	27	0,1	6	0,1	33	0,1
1	841	4,4	324	5,1	1165	4,6
2	6853	35,9	2143	33,8	8996	35,3
3	3167	16,6	1207	19,0	4372	17,2
4	2652	13,9	825	13,0	3477	13,7
5	2600	13,6	697	11,0	3297	13,0
6	0	0,0	0	0,0	0	0,0
7	2972	15,6	1145	18,0	4117	16,2
Total	19112	100,0	6347	100,0	25459	100,0

Tabela 8.8: Estatísticas - idade.

	Média	Mediana	Desvio padrão
Bons pagadores	42,8	41	14,06
Maus pagadores	39,9	38	12,81

Tabela 8.9: Estatísticas - número de dependentes.

	Média	Mediana	Desvio padrão
Bons pagadores	0,73	0,00	1,01
Maus pagadores	0,86	1,00	1,07

Tabela 8.10: Estatísticas - renda.

	Média	Mediana	Desvio padrão
Bons pagadores	2314,23	1500,00	8242,92
Maus pagadores	2222,02	1400,00	12620,80

8.1.3 Modelos

- **Uso de variável latente.**

Tabela 8.11: Estimativas, EP e valores p dos parâmetros do modelo logístico - UVL

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	0,02999	0,3133	0,92374
x3(2)	-0,02482	0,0312	0,42625
x3(3)	0,06204	0,07858	0,42976
x3(4)	0,08303	0,06874	0,22706
x3(5)	-0,0331	0,07585	0,66253
x3(6)	0,06982	0,03576	<0,001
x4(2)	-0,07641	0,02712	<0,001
x4(3)	-0,09242	0,07352	0,2087
x4(4)	0,05244	0,3334	0,8750
x4(5)	-0,3194	0,09771	0,0010
x8(1)	0,3977	0,3165	0,2089
x8(2)	0,5096	0,3126	0,1030
x8(3)	0,5785	0,3127	0,0643
x8(4)	0,3907	0,3135	0,2126
x8(5)	0,4087	0,3143	0,1934
x8(7)	0,5822	0,3129	0,0627
x10	0,04879	0,01171	<0,001
x12	-0,0005123	0,00006	<0,001
x13	-0,002926	0,000055	<0,001

Tabela 8.12: Estimativas, EP e valores p dos parâmetros do modelo *probit* - UVL.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	0,05132	0,1807	0,776364
x4(2)	-0,04594	0,01616	<0,001
x4(3)	-0,05684	0,04379	0,1943
x4(4)	0,02648	0,1996	0,8945
x4(5)	-0,1971	0,05825	<0,001
x8(1)	0,2175	0,1827	0,2338
x8(2)	0,2893	0,1803	0,1085
x8(3)	0,3295	0,1804	0,0678
x8(4)	0,2116	0,1808	0,2418
x8(5)	0,2256	0,1812	0,2131
x8(7)	0,3307	0,1805	0,0669
x10	0,02976	0,006708	<0,001
x12	-0,0003279	0,00004092	<0,001
x13	-0,001781	0,0000321	<0,001

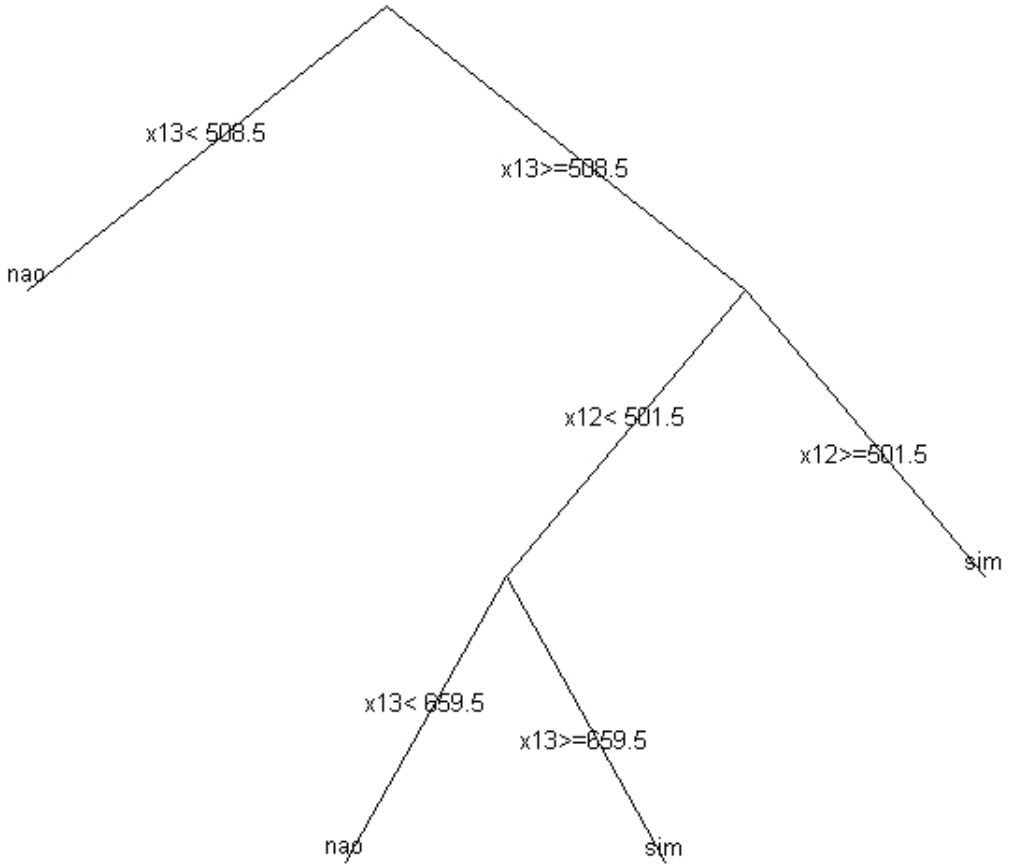

Figura 8.1: Árvore de decisão/classificação - UVL

- **Reclassificação.**

Tabela 8.13: Estimativas, EP e valores p dos parâmetros do modelo logístico - reclassificação

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	4,737	0,3972	<0,001
x1(2)	0,1174	0,07211	0,1036
x1(4)	-0,02706	0,0513	0,5980
x1(5)	0,5032	0,03346	<0,001
x1(6)	-0,211	0,1568	0,1784
x1(7)	0,09749	0,05868	0,0966
x1(9)	0,5101	0,2564	0,0466
x3(2)	-0,1929	0,03502	<0,001
x3(3)	0,06058	0,09006	0,5011
x3(4)	-0,05401	0,07717	0,4839
x3(5)	0,0007602	0,08637	0,9929
x3(6)	-0,001155	0,04108	0,9775
x4(2)	0,1747	0,03026	<0,001
x4(3)	0,2579	0,08184	<0,001
x4(4)	0,1747	0,3695	0,6362
x4(5)	0,2999	0,1169	0,0103
x7	-0,004033	0,001271	0,0015
x8(1)	-0,6547	0,3153	0,0378
x8(2)	-0,7622	0,3105	0,0140
x8(3)	-0,537	0,3111	0,0843
x8(4)	-0,8524	0,3118	0,0062
x8(5)	-0,8335	0,3122	0,0076
x8(7)	-0,5611	0,3111	0,0713
x9	0,004889	0,001485	<0,001
x12	-0,004411	0,0001062	<0,001
x13	-0,003074	0,00006536	<0,001

Tabela 8.14: Estimativas, EP e valores p dos parâmetros do modelo *probit* - reclassificação.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	2,803	0,2315	<0,001
x1(2)	0,06817	0,04199	0,1044
x1(4)	-0,0174	0,03001	0,5620
x1(5)	0,2932	0,01943	<0,001
x1(6)	-0,1283	0,09141	0,1604
x1(7)	0,05974	0,03418	0,0805
x1(9)	0,2758	0,1501	0,06606
x3(2)	-0,1145	0,02049	<0,001
x3(3)	0,03267	0,05246	0,5334
x3(4)	-0,0302	0,04508	0,5028
x3(5)	-0,0004345	0,05049	0,9931
x3(6)	0,003137	0,02393	0,8956
x4(2)	0,1006	0,01766	<0,0001
x4(3)	0,1537	0,0477	0,0012
x4(4)	0,07188	0,2166	0,7400
x4(5)	0,1726	0,06756	0,0106
x7	-0,002251	0,0007424	<0,001
x8(1)	-0,3802	0,1831	0,0378
x8(2)	-0,4416	0,1802	0,0142
x8(3)	-0,3092	0,1806	0,0869
x8(4)	-0,4966	0,181	0,00609
x8(5)	-0,4797	0,1813	0,0081
x8(7)	-0,3262	0,1806	0,0709
x9	0,002732	0,0008652	0,0015
x12	-0,00256	0,00005848	<0,001
x13	-0,001837	0,00003664	<0,001

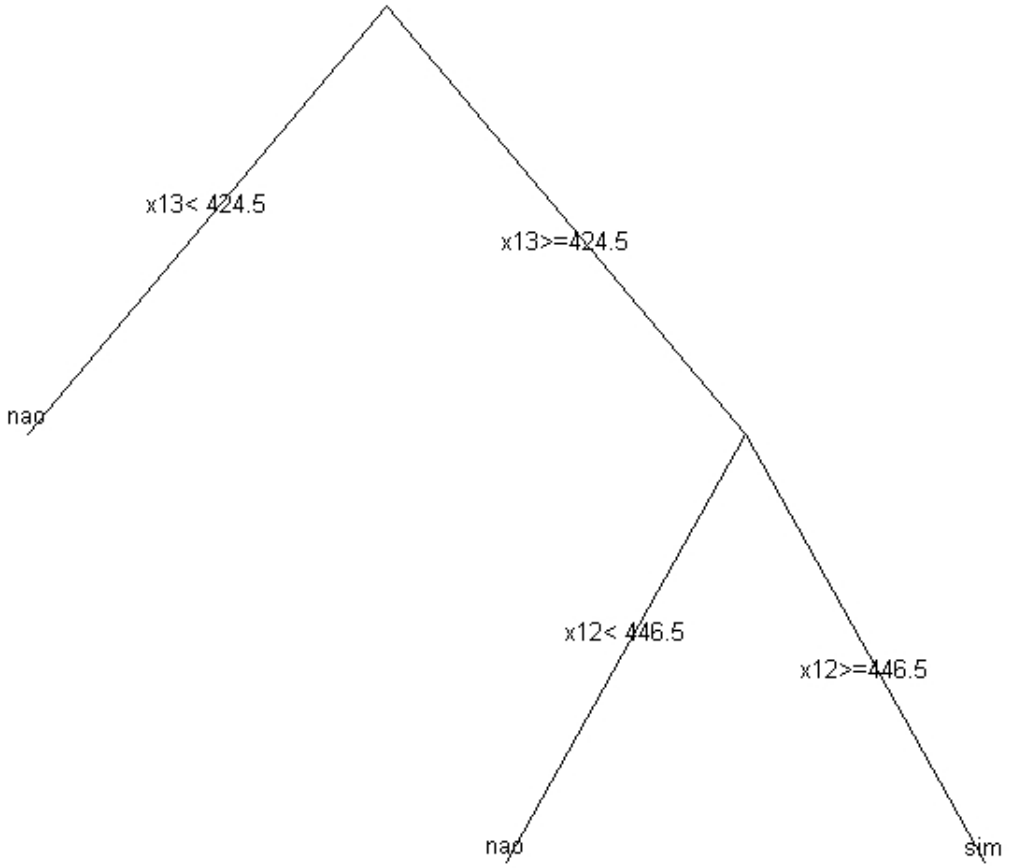

Figura 8.2: Árvore de decisão/classificação - reclassificação

- **Parcelamento.**

Tabela 8.15: Estimativas, EP e valores p dos parâmetros do modelo logístico - parcelamento.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,723	0,3872	<0,001
x1(2)	0,312	0.0,6826	<0,001
x1(4)	0,05385	0,05296	0,3091
x1(5)	0,1455	0,0309	<0,001
x1(6)	0,1206	0,1612	0,4542
x1(7)	0,1431	0,06012	0,01729
x1(9)	-1,367	0,3815	<0,001
x2(102)	-0,4564	0,2266	0,0439
x2(103)	-0,3616	0,2277	0,1123
x2(104)	-0,4587	0,2285	0,0446
x2(105)	-0,4959	0,2255	0,0278
x2(106)	-0,2758	0,2291	0,2285
x2(107)	-0,2309	0,2285	0,3121
x2(108)	-0,4438	0,3233	0,1698
x2(109)	-0,2611	0,2482	0,2928
x3(2)	-0,08353	0,03514	0,0174
x3(3)	0,0533	0,08573	0,5340
x3(4)	0,01148	0,07679	0,8811
x3(5)	0,3275	0,08921	<0,001
x3(6)	-0,04849	0,03855	0,2084
x4(2)	-0,01348	0,02894	0,6414
x4(3)	0,2175	0,07637	0,0044
x4(4)	0,3951	0,3489	0,2574
x4(5)	-0,1661	0,1033	0,10763
x5(M)	0,1309	0,02606	<0,001
x9	0.008985	0,001412	<0,001
x10	0.0445	0,01268	<0,001
x12	-0.002045	0,00007509	<0,001
x13	-0.004165	0,00005928	<0,001

Tabela 8.16: Estimativas, EP e valores p dos parâmetros do modelo *probit* - parcelamento.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,015	0,2257	<0,001
x1(2)	0,1791	0,03991	<0,001
x1(4)	0,03276	0,03051	0,2828
x1(5)	0,08345	0,01814	<0,001
x1(6)	0,0753	0,09269	0,4165
x1(7)	0,08747	0,03451	0,0112
x1(9)	-0,7388	0,2073	<0,001
x2(102)	-0,2531	0,1335	0,0578
x2(103)	-0,1975	0,1341	0,1409
x2(104)	-0,2517	0,1346	0,0615
x2(105)	-0,2712	0,1329	0,0411
x2(106)	-0,1437	0,135	0,2869
x2(107)	-0,1159	0,1345	0,3888
x2(108)	-0,2202	0,1879	0,2412
x2(109)	-0,119	0,1454	0,4129
x3(2)	-0,04665	0,02054	0,0231
x3(3)	0,02713	0,04998	0,5872
x3(4)	0,007631	0,04435	0,8633
x3(5)	0,185	0,05135	<0,001
x3(6)	-0,02744	0,02269	0,2266
x4(2)	-0,008568	0,01689	0,6119
x4(3)	0,1217	0,04483	<0,001
x4(4)	0,2462	0,2026	0,2243
x4(5)	-0,1034	0,06018	0,0857
x5(M)	0,07823	0,01523	<0,001
x9	0,005344	0,0008202	<0,001
x10	0,02586	0,007442	0.00051
x12	-0,00118	0,00004369	<0,001
x13	-0,002453	0,00003407	<0,001

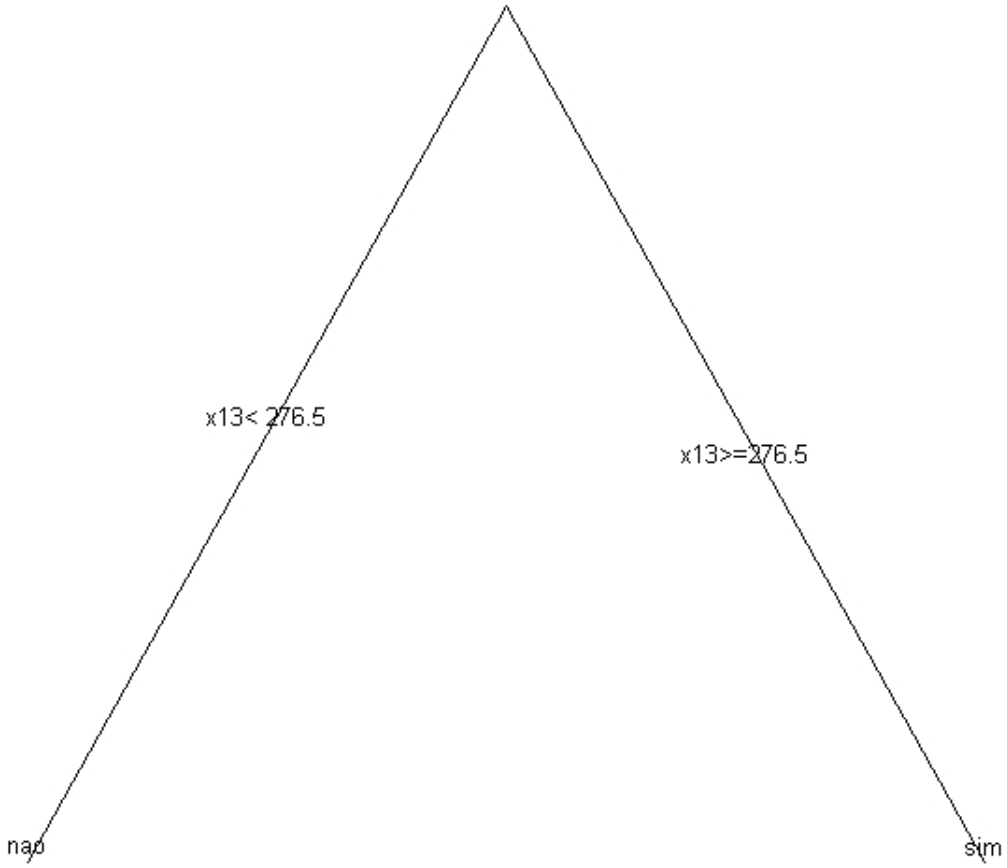

Figura 8.3: Árvore de decisão/classificação - parcelamento

- **Ponderação.**

Tabela 8.17: Estimativas, EP e valores p dos parâmetros do modelo logístico - ponderação.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	1,681	0,4636	<0,001
x1(2)	0,3942	0,06385	<0,001
x1(4)	0,1132	0,04804	0,0184
x1(5)	0,1306	0,03407	<0,001
x1(6)	-0,1282	0,1407	0,3622
x1(7)	0,07846	0,05531	0,1560
x1(9)	0,2759	0,2873	0,3368
x3(2)	-0,06278	0,03463	0,0698
x3(3)	-0,07893	0,08982	0,3795
x3(4)	0,2176	0,07177	<0,001
x3(5)	0,3416	0,07892	<0,001
x3(6)	0,02741	0,04045	0,4980
x4(2)	0,02521	0,02913	0,3869
x4(3)	-0,3335	0,08737	<0,001
x4(4)	-0,4137	0,5499	0,4519
x4(5)	-0,1077	0,1374	0,4329
x9	0,01387	0,001394	<0,001
x12	-0,001578	0,0001172	<0,001
x13	-0,004452	0,00006466	<0,001

Tabela 8.18: Estimativas, EP e valores p dos parâmetros do modelo *probit* - ponderação.

Parâmetro	Estimativa	Erro padrão	Valor p
Intercepto	0,9467	0,2697	<0,001
x1(2)	0,232	0,03776	<0,001
x1(4)	0,06448	0,02785	0,0206
x1(5)	0,07906	0,02011	<0,001
x1(6)	-0,08759	0,08083	0,2785
x1(7)	0,0524	0,03198	0,1013
x1(9)	0,1763	0,1619	0,2763
x3(2)	-0,03485	0,02031	0,0862
x3(3)	-0,04043	0,05241	0,4404
x3(4)	0,1317	0,04205	0,0017
x3(5)	0,1996	0,04633	<0,001
x3(6)	0,01719	0,02383	0,4705
x4(2)	0,01169	0,01705	0,4929
x4(3)	-0,191	0,05018	<0,001
x4(4)	-0,2143	0,3107	0,4903
x4(5)	-0,06245	0,07981	0,4339
x9	0,008047	0,000814	<0,001
x12	-0,000847	0,00006821	<0,001
x13	-0,002611	0,00003669	<0,001

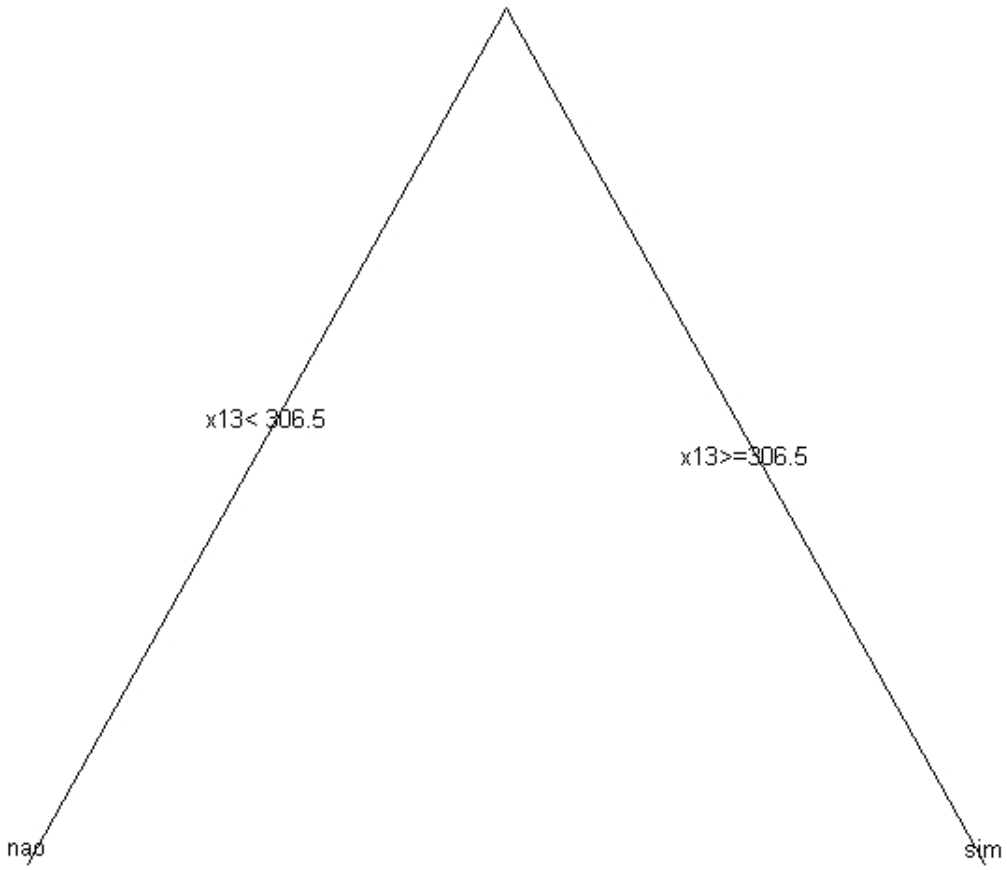

Figura 8.4: Árvore de decisão/classificação - ponderação

Referências Bibliográficas

[1] Adou, H.A., Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of a literature. *John Wiley and Sons*, (18):59-88.

[2] Alves, M.C. (2008). *Estratégia para o desenvolvimento de modelos credit score com inferência dos rejeitados*. Dissertação de Mestrado. IME, Universidade de São Paulo, São Paulo.

[3] Breiman, L. Bagging Predictors. Machine Learning. 24(2):123-140p. 1996.

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

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

[6] Banasik, J., Crook, J. (2007). Reject inferences, augmentation, and sample slection. *European Journal Operation Research*, (183):1582-1594.

[7] Banasik, J., Crook, J. Does Reject Inference Really Improve the Performance of Application Scoring Models? *Credit Research Center, The School of Management*. University of Edinburgh.

[8] Braga, A.C.S. (2000). *Curvas ROC: aspectos funcionais e aplicações*. Dissertação de Mestrado, Universidade do Minho, Portugal.

[9] Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, (26):801-849.

[10] BREIMAN, L. Bagging Predictors. Machine Learning. 24(2):123-140p. 1996.

[11] Breiman, L. et. al. (1984). *Classification and Regression Tree*. Wadsworth International: California, USA.

[12] BUHLMANN, P. YU, B. Explaining Bagging. University of California at Berkeley, 2000.

[13] Demetrio,C., Cordeiro, G. (2007). *Modelos lineares generalizados e extensões*. São Paulo.

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

[15] Diniz, C.A., Louzada Neto, F. Modelagem estatística para risco de crédito. *Apresentação no 20º SINAPE*. ABE.

[16] EstatCamp. (2012). Seleção Stepwise. Disponível em: *http://www.portalaction.com.br/954-seleção-stepwise*

[17] Farhart, C.A.V. (2003) *Análise de diagnóstico em regressão logística*. Dissertação de Mestrado. IME, Universidade de São Paulo, São Paulo.

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

[19] Hand, D.J., Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal Royal Statistic Society*, (160):523-541.

[20] Homer, D., Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

[21] Junger, W. L. (2006). Parameter estimation in probabilistic record linkage: an application of the EM algorithm. *Caderno de Saúde Coletiva*, Rio de Janeiro, 14 (2):225-232.

[22] Keramati, A., Yousefi, N. (2011). A proposed classification of data mining techniques in credit scoring. *International Conference on Industrial Engineering and Operations Management*. Malaysia.

[23] Louzada Neto, et. al. (2003). Anaylisis of diagnostic tests using ROC curves. *Caderno de Saúde Coletiva*, (1):7-31. Rio de Janeiro.

[24] Louzada Neto, et. al. A general latent class model for performance evaluation of Diagnostic tests in the abscence of a gold standard: an application to Chagas Disease. Computational and Mathematical Methods in Medicine. Article ID 487502. Vol. 2012.

[25] Louzada Neto, F. e.a. (2011). Inferência dos rejeitados: aumentando a capacidade preditiva da modelagem via combinação de modelos. *Revista de Tecnologia de Crédito*, (77):6-17.

[26] Montrichard, D. (2007). Reject Inference Methodologies In Credit Risk Modeling. *Presentation in CIBIC*.

[27] Mum, L. (2012). *Reject Inference in online purchase*. Dissertação de Mestrado. Royal Institute of Technology, KTH, Estocolmo, Suécia.

[28] Oliveira, M.M. (1998). Modelos de escolha binária. Disponível em: *http://www.fep.up.pt/disciplinas/2E103/modelos_escolha_binaria.pdf*

[29] Pereira, G.A. (2011). *Avaliação de testes diagnósticos na ausência de padrão ouro considerando relaxamento da suposição de independência condicional, covariáveis e estratificação da população: uma abordagem Baeysiana*. Tese de Doutorado. Departamento de Estatística, Universidade Federal de São Carlos, 234 f.

[30] Prati, R.C., Batista, G.E.A.P.A, Monard, M.C.(2013). Curvas ROC para avaliação de classificadores. IME, Universidade de São Paulo, São Paulo.

[31] Reiser, B., Faraggi, D. (1997). Confindence intervals for the generalized ROC criterion. *Biometrics*, (53):644-652.

[32] Robin, X. e.a. (2011). Proc: an open-source package for r and s++ to analyze and compare roc curves. *BMC Bioinformatics*, (12):77.

[33] Rocha, R.F. (2012). *Combinação de classificadores para Inferência dos Rejeitados*. Dissertação de Mestrado. Departamento de Estatística, Universidade Federal de São Carlos, São Carlos.

[34] Rodrigues, M.A.S. (2005). *Árvore de Classificação*. Monografia. Universidade dos Açores, Portugal.

[35] Semedo, D.P.V. (2009). *Credit Scoring: aplicação da regressão logística vs redes neurais artificiais na avaliação de risco de crédito no mercado cabo-verdiano*. Dissertação de Mestrado. Universidade Nova de Lisboa.

[36] Sicsu, A.L. (2010). *Credit Scoring: desenvolvimento, implantação, acompanhamento*. Blucher, São Paulo.

[37] Silva, F. (2006). Análise ROC. São José do Rio Preto, SP.

[38] Silva, P.H.F. (2008). *Medidas do Valor Preditivo de Modelos de Classificação Aplicados a Dados de Crédito*. Projeto Científico. Universidade Federal de São Carlos, São Carlos.

[39] Souza, E. (2006). *Análise de influência local no modelo de regressão logística*. ESALQ, Universidade de São Paulo, Piracicaba, SP.

[40] Thomas, L.C. Edelman, D.E. Crook,J.N. Credit Scoring and its Applications. Monographs on Mathematical Modelling and Computation. Philadelphia: Society for Industrial and Applied Mathematics. 2002.

[41] Thomas, L.C., Crook, J.N. (2002). *Credit Scoring and its Applications - Monographs on Mathematical Modelling and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, USA.

[42] Vieira, H.A.S. e. a. (2005). *Data Mining em R*. Universidade do Minho, Azurem, Portugal.

[43] Von Zubenn, F. J., Atuux, R.R.F. (2013). *Árvore de Classificação*. UNICAMP, Campinas, SP.

[44] Wang, X., Dey, D. K. (2010). Generalized Extreme Value Regression for Binary Response data: An Application to B2B Eletronic Payments System Adoption. *The Annals of Applied Statistics*, (4): 2000-2023.

[45] Zeller, C.B. (2009). *Distribuições Mistura de Escala Skew Normal: Estimação e Diagnóstico em Modelos Lineares*. Tese de Doutorado. Departamento de Estatística, IMECC-UNICAMP.

[46] Zweig, M.H., Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39 (4):561-577.